BUSINESS ANALYTICS SKILLS FOR THE FUTURE-PROOF SUPPLY CHAINS -

STATYSTYCZNE METODY ANALIZY DANYCH LOGISTYCZNYCH

Autorzy:
Maja Fošner
Roman Gumzej
Rebeka Kovačič Lukman
Marinko Maslarić
Dejan Mirčetić
Kristijan Brglez
Benjamin Marcen

Poznan 2025

Wydawca:
Wyższa Szkoła Logistyki
Estkowskiego 6
61-755 Poznań, POLSKA
www.wsl.com.pl

Recenzenci:
Prof. dr Ajda Fošner, Uniwersytet Primorski, Wydział Zarządzania
Prof. dr Nikša Alfirević, Uniwersytet w Splicie, Wydział Ekonomii

Redaktor techniczny:
Kristijan Brglez

Autorzy:

Prof. Dr Maja Fošner
Prof. Dr Roman Gumzej
Prof. Dr Rebeka Kovačič Lukman
Prof. Dr Marinko Maslarić
Dr Dejan Mirčetić
Kristijan Brglez
Asyst. prof. Dr. Benjamin Marcen

Tłumaczenie na język polski:

Katarzyna Siemieniak
Konsultacja: dr Tomasz Brzęczek

Poznań 2024, Wydanie I

Wersja drukowana

Wydanie online

ISBN: 978-83-62285-71-6

Monografia została napisana w ramach projektu Business Analytics Skills for the Future-proof Supply Chains (BAS4SC) [2022-1-PL01-KA220-HED-000088856], finansowanego z programu Erasmus+. Projekt jest finansowany przy wsparciu Komisji Europejskiej. Niniejsza publikacja odzwierciedla jedynie poglądy autora, a Komisja nie ponosi odpowiedzialności za jakiekolwiek wykorzystanie informacji w niej zawartych.

Monografia jest dostępna bezpłatnie pod adresem:

Przedmowa

Obszar logistyki staje się coraz bardziej zależny od wiedzy opartej na danych wykorzystywanych w celu optymalizacji operacji, obniżenia kosztów i zapewnienia wydajności w łańcuchach dostaw. Niniejszy podręcznik, *Statystyczne metody analizy danych logistycznych*, służy jako krytyczne źródło wiedzy zarówno dla studentów, jak i profesjonalistów, wyposażając ich w umiejętności niezbędne do poruszania się w złożoności nowoczesnego zarządzania łańcuchem dostaw. Opracowana w ramach projektu Business Analytics Skills for Future-proof Supply Chains (BAS4SC) książka oferuje kompleksowy przegląd metod statystycznych, zarządzania danymi i zaawansowanych technik analitycznych wyraźnie dostosowanych do branży logistycznej

Dzięki starannym badaniom podręcznik ten wypełnia luki w obecnej ofercie edukacyjnej, łącząc wiedzę teoretyczną z praktycznymi zastosowaniami. Dziesięć rozdziałów zawiera szczegółową eksplorację kluczowych tematów, takich jak prognozowanie popytu, modelowanie symulacyjne, analiza regresji oraz integracja sztucznej inteligencji i uczenia maszynowego w operacjach logistycznych. Dzięki wykorzystaniu powszechnie uznanych narzędzi, takich jak SPSS, R i SQL, treść tego podręcznika została zaprojektowana w taki sposób, aby wypełnić lukę między nauczaniem akademickim a potrzebami przemysłu.

Mamy nadzieję, że ten podręcznik dostarczy podstawową wiedzę z zakresu analityki biznesowej dla logistyki i zainspiruje innowacje w tej dziedzinie, wspierając przyszłych liderów przygotowanych do stawienia czoła wyzwaniom szybko zmieniającego się u łańcucha dostaw.

Prof. dr Sanja Bojić
Kristijan Brglez
Prof. dr Maja Fošner
Prof. dr Roman Gumzej
Prof. dr Rebeka Kovačič Lukman
Asystent Prof. dr Benjamin Marcen
Prof. dr Marinko Maslarić
Asystent Prof. dr Boško Matović
Dr Dejan Mirčetić

SPIS TREŚCI

WPROWADZENIE

Niniejszy podręcznik, zatytułowany *Statystyczne metody analizy danych logistycznych*, jest trzecim z serii opracowanej w ramach projektu Business Analytics Skills for Future-proof Supply Chains (BAS4SC). W celu określenia zawartości tego podręcznika przeprowadzono kilka wstępnych działań badawczych. Po pierwsze, przeprowadzono kompleksowe badanie kursów analityki biznesowej, ich treści i umiejętności, które przekazują studentom logistyki w Unii Europejskiej, Stanach Zjednoczonych i Wielkiej Brytanii. Analiza ta ujawniła lukę między wiedzą logistyczną i umiejętnościami statystycznymi wymaganymi w tej dziedzinie a tymi, które są obecnie oferowane studentom. W oparciu o pogłębione wywiady z kadrą dydaktyczną uniwersytetów, studentami i specjalistami z branży, zidentyfikowano ponad 100 umiejętności z zakresu analityki biznesowej. Korzystając z metody klasyfikacji rankingowej ABC, wybrano 33 umiejętności do uwzględnienia w tej książce, koncentrując się głównie na matematyce, informatyce, zarządzaniu, matematyce stosowanej i statystyce. Połączenie tych zidentyfikowanych potrzeb i umiejętności doprowadziło do opracowania dziesięciu rozdziałów treści, które odnoszą się do najbardziej krytycznych umiejętności wymaganych w tej dziedzinie.

Pierwszy rozdział obejmuje wprowadzenie do statystyki i zapewnia kompleksowy przegląd koncepcji statystycznych i ich zastosowań, w szczególności w analizie łańcucha dostaw. Rozpoczyna się od podkreślenia kluczowej roli, jaką statystyka odgrywają w optymalizacji łańcuchów dostaw. Wykorzystuje statystyki opisowe, takie jak średnia, mediana i odchylenie standardowe do analizy czasów dostaw, poziomów zapasów i kosztów. Rozdział ten wprowadza techniki predykcyjne, takie jak regresja i analiza szeregów czasowych do prognozowania popytu i zapasów. W dalszej części zbadano znaczenie zmiennych, rozróżniając typy jakościowe i ilościowe, a także zagłębiono się w podstawowe miary statystyczne, takie jak średnia, mediana, moda, wariancja i odchylenie standardowe.

Dodatkowo rozdział ten omawia metody graficznej reprezentacji danych, takie jak histogramy i wykresy punktowe korelacyjne, a także podkreśla różnicę między statystyką opisową a wnioskowaniem. Wreszcie, wprowadza kluczowe koncepcje korelacji, regresji i rozkładów prawdopodobieństwa, oferując narzędzia do zrozumienia relacji między zmiennymi i modelowania zjawisk losowych w danych. Te techniki statystyczne

pomagają poprawić podejmowanie decyzji w łańcuchu dostaw, wydajność i zarządzanie ryzykiem.

Drugi rozdział, *Statystyka dla analityki biznesowej*, bada podstawowe koncepcje i techniki statystyczne w celu uzyskania wglądu w dane biznesowe. Rozpoczyna się od przedstawienia znaczenia analizy danych w podejmowaniu decyzji biznesowych. Wyjaśnia fundamentalną rolę rozkładu normalnego, który służy jako podstawa wielu metod statystycznych. Rozdział zagłębia się w odchylenie standardowe, podkreślając jego znaczenie w pomiarze zmienności danych. Obejmuje również rozkłady statystyk z próbkowania i Centralne Twierdzenie Graniczne, wyjaśniając, w jaki sposób wnioskuje się o parametrach populacji na podstawie danych próbnych. Zagadnienia takie jak testowanie hipotez, statystyki standaryzowane *Z* i *t* są badane w celu ułatwienia podejmowania decyzji i obliczeń prawdopodobieństwa. Rozdział kończy się omówieniem błędu standardowego i przedziałów ufności, które pomagają określić niepewność związaną z szacunkami. Ostatecznie rozdział ten wyposaża czytelników w narzędzia statystyczne niezbędne do analityki biznesowej, umożliwiając im podejmowanie świadomych i opartych na danych decyzji.

Rozdział poświęcony *zarządzaniu danymi* analizuje różne aspekty zarządzania danymi w logistyce, koncentrując się na formatach danych, organizacji i technologiach. Rozpoczyna się on od roli Elektronicznej Wymiany Danych (EDI) w wymianie informacji w łańcuchach dostaw przy użyciu standardowych formatów alfanumerycznych. Wyjaśnia pojęcie informacji, danych i wiedzy, omawiając, w jaki sposób dane są digitalizowane i organizowane w bazach danych, magazynach i bazach wiedzy. Rozdział zagłębia się w dane logistyczne, w szczególności wykorzystanie kodów kreskowych i znaczników RFID do identyfikacji i śledzenia czegoś w logistyce. Wprowadza również techniki organizacji danych, od arkuszy kalkulacyjnych po relacyjne bazy danych (RDB), wyjaśniając kluczowe pojęcia, takie jak klucze podstawowe i obce, normalizacja i języki zapytań, takie jak SQL. Ponadto w rozdziale omówiono najlepsze praktyki filtrowania danych i zapobiegania błędom podczas wprowadzania danych. Na koniec omówiono różnice między hurtowniami danych a bazami wiedzy, podkreślając ich rolę w analizie biznesowej i podejmowaniu decyzji.

Rozdział poświęcony *Modelowaniu i Analizie Symulacyjnej* (SMA) koncentruje się na tworzeniu modeli cyfrowych do symulacji rzeczywistych systemów w celu optymalizacji i podejmowania decyzji. Rozpoczyna się od wyjaśnienia twierdzenia Conanta-Ashby'ego, które sugeruje, że model symulacyjny musi odpowiadać złożoności swojego odpowiednika

w świecie rzeczywistym, aby skutecznie go regulować. SMA optymalizuje łańcuchy dostaw i sieci ruchu w logistyce, umożliwiając menedżerom symulację i ocenę różnych scenariuszy. W rozdziale przedstawiono najważniejsze metodologie symulacji, takie jak Symulacja Zdarzeń Dyskretnych (DES) do analizy zorientowanej na proces, Dynamika Systemu (SD) do wysokopoziomowej wydajności systemu, Symulacja Agentowa (ABS) do modelowania zachowania poszczególnych podmiotów oraz Symulacja Sieci (NS) do analizy przepływów sieciowych. Każda metoda zapewnia wgląd w różne aspekty logistyki, od cykli produkcyjnych po optymalizację ruchu. Rozdział kończy się przeglądem projektów symulacji logistycznych, zorganizowanych wokół cyklu doskonalenia Design for Six Sigma i Deminga, które pomagają w planowaniu, wykonywaniu i udoskonalaniu złożonych systemów logistycznych.

Rozdział *Regresja Liniowa z Pojedynczym i Wieloma Regresorami* wprowadza analizę regresji w celu zrozumienia związku między zmiennymi zależnymi i niezależnymi. Rozdział rozpoczyna się od prostej regresji liniowej, w której pojedyncza zmienna niezależna, taka jak wydatki na reklamę, przewiduje wynik, taki jak sprzedaż. Rozdział wyjaśnia budowę modelu regresji i równania regresji używanego do prognozowania zmiennej zależnej na podstawie przykładowych danych. Omówiono również metodę najmniejszych kwadratów, podstawową technikę szacowania linii regresji poprzez minimalizację błędów dopasowania do danych. Następnie wprowadzono współczynnik determinacji (R^2), aby zmierzyć, jak dobrze model regresji pasuje do danych. Rozdział ten zagłębia się następnie w regresję wieloraką, w której dwie lub więcej zmiennych niezależnych przewiduje zmienną zależną, oferując bardziej kompleksową analizę. Przykłady obejmują przewidywanie czasu podróży na podstawie odległości i liczby dostaw.

Wprowadzenie do Badań Operacyjnych koncentruje się na wykorzystaniu metod analitycznych w celu usprawnienia procesu podejmowania decyzji, w szczególności w logistyce i zarządzaniu łańcuchem dostaw. Badania operacyjne wykorzystują modelowanie, statystykę i techniki optymalizacji w celu znalezienia optymalnych rozwiązań złożonych problemów, umożliwiając efektywne zarządzanie zasobami, kontrolę zapasów i optymalizację procesów. Rozdział ten zwraca uwagę na strategiczne planowanie logistyczne, które obejmuje metody takie jak Six Sigma i produkcja Just-in-Time w celu zwiększenia wydajności operacyjnej. Business Intelligence (BI) i Business Analytics (BA) mają kluczowe znaczenie w analizie danych, umożliwiając firmom podejmowanie świadomych decyzji przy użyciu prognozowania, analizy predykcyjnej i wizualizacji danych. Rozdział ten wprowadza również wielokryterialne podejmowanie decyzji (MCDM), które pomaga w ocenie i wyborze

optymalnych rozwiązań w oparciu o wiele kryteriów oceny. Wreszcie, systemy wspomagania decyzji (DSS) i inżynieria oparta na wiedzy (KBE) pomagają zintegrować wiedzę i dane z procesami decyzyjnymi, dodatkowo zwiększając wydajność operacyjną i planowanie strategiczne.

Rozdział poświęcony *Statystycznemu Przetwarzaniu Danych za Pomocą SPSS* przedstawia oprogramowanie IBM SPSS jako potężne narzędzie do automatyzacji złożonych analiz statystycznych, zwiększania wiarygodności i ułatwiania podejmowania decyzji. Wyjaśnia, w jaki sposób SPSS umożliwia importowanie danych, manipulowanie nimi i przygotowywanie ich za pomocą przyjaznego dla użytkownika interfejsu. Rozdział obejmuje kluczowe funkcje, takie jak statystyki opisowe, tworzenie wykresów i wizualizacja danych. Wprowadza również podstawowe testy statystyczne – testy parametryczne, korelację, Chi-kwadrat i ANOVA - prowadząc czytelników przez konfigurację i interpretację każdego testu. Dodatkowo analizuje narzędzia do zarządzania danymi, takie jak łączenie, dzielenie i obliczanie zmiennych w zestawie danych, pokazując, w jaki sposób SPSS usprawnia analizę statystyczną w logistyce i innych dziedzinach.

Rozdział 8 omawia Analitykę Biznesową (BA) i jej zastosowanie za pomocą narzędzi takich jak R i SQL do rozwiązywania problemów biznesowych. BA ma na celu poprawę podejmowania decyzji i wydajności firmy przy użyciu metod opartych na danych. Obejmuje ona opisowe, predykcyjne i preskryptywne środowisko analityczne wspomagania podejmowania świadomych decyzji. R został przedstawiony jako solidne, otwarte narzędzie do analizy statystycznej i wizualizacji, podczas gdy SQL jest niezbędny do zarządzania dużymi bazami danych i wysyłania do nich zapytań. Rozdział szczegółowo opisuje integrację R i SQL w celu wydajnej analizy biznesowej, podkreślając, w jaki sposób dane przechowywane w SQL mogą być analizowane za pomocą skryptów R w celu automatyzacji zadań. Praktyczne przykłady, takie jak kwerenda w bazie danych Chinook Database, ilustrują współpracę R i SQL w celu generowania spostrzeżeń, do których należy identyfikacja najlepiej sprzedających się albumów. Ta synergia między BA, R i SQL zwiększa możliwości zarządzania i analizowania dynamicznych danych biznesowych.

Rozdział 9 koncentruje się na prognozowaniu popytu w łańcuchu dostaw, wizualizacji i analizie funkcji. Prognozowanie popytu przewiduje potrzeby klientów, zmieniając cały łańcuch dostaw i zmniejszając koszty logistyczne. Rozdział ten przedstawia kluczowe etapy prognozowania popytu, w tym definiowanie problemu, gromadzenie danych, analizę

trendów, wybór modeli i ich ocenę. Wizualizacja pomaga zidentyfikować wzorce, takie jak sezonowość i trendy, które mogą informować o wyborze modelu. S-ARIMA (ang. Seasonal Autoregressive Integrated Moving Average) jest podkreślana jako skuteczny model złożonych szeregów czasowych danych, zwłaszcza w przypadku sezonowych wzorców popytu. Przykład z branży spożywczej demonstruje zdolność modelu S-ARIMA do przewidywania popytu i kierowania procesem decyzyjnym. Wreszcie, rozdział obejmuje sposób testowania i walidacji modeli prognostycznych w celu zapewnienia ich skuteczności w rzeczywistych zastosowaniach, przy użyciu wskaźników takich jak RMSE i MAPE do oceny trafności.

Ostatni rozdział bada rolę Sztucznej Inteligencji (AI) i Uczenia Maszynowego (ML) w łańcuchach dostaw, zaczynając od przeglądu rozwoju AI od systemów symbolicznych po nowoczesne podejścia ML. AI odnosi się do automatyzacji zadań, które zazwyczaj wymagają ludzkiej inteligencji, przy czym ML jest podzbiorem AI, który koncentruje się na uczeniu się wzorców zaszytych w danych. Rozdział podkreśla, w jaki sposób modele AI/ML, takie jak uczenie nadzorowane i nienadzorowane, są stosowane do rozwiązywania problemów biznesowych, w tym prognozowania popytu, zarządzania zapasami i optymalizacji. Istotne studium przypadku obejmuje zastosowanie algorytmów AI i ML w magazynie centralnym fabryki produkującej żywność, optymalizując wykorzystanie wózków widłowych. Poprzez włączenie systemów wspomagania decyzji (DSS), modele AI/ML pomagają menedżerom w wyborze optymalnej liczby wózków widłowych, zwiększając wydajność operacyjną. Rozdział podkreśla, w jaki sposób AI/ML może przechwytywać wiedzę ekspercką, obniżać koszty i usprawniać procesy decyzyjne w zarządzaniu łańcuchem dostaw.

1. Statystyka

1.1 Rola i znaczenie statystyki w analizie danych w łańcuchach dostaw

Statystyka odgrywa kluczową rolę w nowoczesnych łańcuchach dostaw, gdzie efektywne zarządzanie, planowanie i kontrola są niezbędne. Metody statystyczne są wykorzystywane do gromadzenia, analizowania i interpretowania danych, umożliwiając firmom lepsze zrozumienie i optymalizację ich łańcuchów dostaw.

Przedstawmy niektóre z ważnych ról statystyki w analizie łańcucha dostaw.

Statystyki opisowe są kluczem do opisania podstawowych cech serii danych łańcucha dostaw, takich jak średnia, odchylenie standardowe, mediana, kwartyle i inne miary. Narzędzia te pomagają nam zrozumieć rozkład i charakterystykę danych, takich jak średni czas dostawy, ilości w magazynie i średnie koszty, co przyczynia się do lepszego zrozumienia i zarządzania łańcuchem dostaw.

Ponadto techniki statystyczne, takie jak regresja, analiza szeregów czasowych i wzorce analityczne danych, są wykorzystywane do przewidywania przyszłych zdarzeń i trendów w łańcuchach dostaw. Obejmuje to prognozowanie popytu, zapasów i czasu dostawy, umożliwiając lepsze planowanie i dostosowanie dostaw.

Statystyki odgrywają kluczową rolę w identyfikowaniu wzorców w danych, umożliwiając lepsze zrozumienie zachowania łańcucha dostaw, w tym wzorców sezonowych, trendów i cykli popytu.

Optymalizacja zapasów to kolejny kluczowy obszar, w którym statystyki pomagają określić optymalne ilości zamówień, które minimalizują koszty magazynowania i zamawiania, przy użyciu metod takich jak EOQ (Economic Order Quantity).

Ponadto, statystyki są również wykorzystywane do oceny ryzyka związanego z łańcuchem dostaw, takiego jak prawdopodobieństwo opóźnień w dostawach, uszkodzeń podczas transportu i innych potencjalnych problemów.

Dzięki monitorowaniu statystycznemu i kontroli procesów identyfikujemy odchylenia od standardów, co pozwala nam poprawić jakość i wydajność procesów łańcucha dostaw.

Ponadto statystyki są wykorzystywane do monitorowania i poprawy jakości produktów i usług w łańcuchu dostaw, w tym kontroli jakości u dostawców.

Wreszcie, statystyki są kluczowym narzędziem do podejmowania bardziej świadomych decyzji dotyczących zaopatrzenia, zapasów, wyboru dostawców i innych aspektów zarządzania dostawami, przyczyniając się do wydajnego i skutecznego działania całego łańcucha dostaw.

W analizie łańcucha dostaw statystyki są wykorzystywane do optymalizacji procesów, redukcji kosztów, zwiększenia wydajności i poprawy zadowolenia klientów. Umożliwia lepsze zrozumienie dynamiki łańcucha dostaw i lepsze zarządzanie ryzykiem, co ma kluczowe znaczenie dla pomyślnego funkcjonowania firm i organizacji w dzisiejszym globalnym środowisku.

1.2 Podstawowe pojęcia statystyki

Zmienne

Zmienne są podstawowymi elementami składowymi w statystyce, ponieważ reprezentują właściwości lub cechy, które są mierzone lub obserwowane w badaniu, eksperymencie lub próbie danych. Zmienne są niezbędne do zrozumienia i analizy danych, ponieważ pozwalają badaczom, analitykom i statystykom opisywać, analizować i rozumieć zjawiska.

Ważne jest, aby zrozumieć różne rodzaje zmiennych i ich znaczenie w statystyce.

Zmienne jakościowe (nominalna, kategorialna/porządkowa) to zmienne reprezentujące cechy jakościowe, czyli niemierzalne na skali liczbowej lub zmienne, które klasyfikują rozłącznie jednostki danych. Przykłady obejmują płeć (mężczyzna, kobieta), kolor oczu (niebieski, brązowy, zielony) lub typ samochodu (sedan, kombi, SUV). Zmienne jakościowe są często przydatne do opisywania cech demograficznych lub innych cech.

Zmienne ilościowe (liczbowe) to zmienne reprezentujące wartości liczbowe, które można policzyć lub zmierzyć i które można posortować w pewnym porządku matematycznym. Przykłady obejmują wiek w latach, wzrost, temperaturę, dochód lub wyniki ankiet. Zmienne ilościowe są często wykorzystywane do analizy i ilościowego badania zjawisk.

Zmienne zależne i niezależne. Zmienna zależna to ta, którą chcemy zbadać, zmierzyć lub przewidzieć, podczas gdy zmienna niezależna to ta, która ma wpływać na zmienną zależną. Na przykład, jeśli chcemy zbadać, czy poziom wykształcenia wpływa na dochód, dochód jest zmienną zależną, a poziom wykształcenia zmienną niezależną.

Zmienne dyskretne i ciągłe. Zmienne można również podzielić na dyskretne i ciągłe. Zmienne dyskretne mają ograniczony zestaw możliwych wartości i są zwykle reprezentowane przez liczby całkowite. Przykładem może być liczba dzieci w rodzinie, gdzie możliwe wartości to 0, 1, 2 itd. Z drugiej strony zmienne ciągłe mają nieskończoną liczbę możliwych wartości i są zwykle mierzone za pomocą liczb dziesiętnych. Przykładem jest wzrost osób, gdzie możliwa jest nieskończona liczba wartości w danym zakresie.

Zmienne są podstawowymi narzędziami badań i analizy danych. Zrozumienie i prawidłowe zdefiniowanie zmiennych ma kluczowe znaczenie dla przeprowadzania analiz statystycznych i badania zjawisk w badaniach. Zmienne pozwalają badaczom wyrażać i kwantyfikować różne aspekty rzeczywistości, umożliwiając lepsze zrozumienie zjawisk, podejmowanie decyzji i przewidywanie przyszłych zdarzeń. Pozwalają również na wykorzystanie różnych technik statystycznych do testowania hipotez, tworzenia prognoz i lepszego zrozumienia związków przyczynowych między zmiennymi.

1.3 Podstawowe pojęcia statystyczne z przykładami

Średnia arytmetyczna (średnia)

Średnia, znana również jako **średnia arytmetyczna**, jest jedną z podstawowych miar statystycznych. Średnia jest średnią arytmetyczną wszystkich wartości w zestawie danych. Oblicza się ją poprzez zsumowanie wszystkich danych, a następnie podzielenie przez liczbę danych.

Obliczanie średniej arytmetycznej:

- Zsumowanie wszystkich wartości w zbiorze danych.
- Podzielenie sumy przez liczbę danych w zbiorze.
- Równanie do obliczania średniej $(\bar{x})$ to: $\bar{x} = (x_1 + x_2 + x_3 + \ldots + x_n) / n$

Gdzie: $\bar{x}$ jest średnią arytmetyczną. $x_1, x_2, x_3, \ldots, x_n$ są wartościami w zbiorze danych. n to liczba wartości w zbiorze danych.

Przykład:

Wyobraźmy sobie zbiór danych reprezentujący oceny uczniów z egzaminu z matematyki: 80, 85, 90, 75, 95. Aby obliczyć średnią, należy dodać wszystkie te wartości i podzielić przez liczbę ocen, która w tym przypadku wynosi 5:

Średnia arytmetyczna = (80 + 85 + 90 + 75 + 95) / 5 = 425 / 5 = 85

Tak więc średni wynik ucznia wynosi 85. Średnia jest przydatna do pomiaru tendencji centralnej danych i daje nam przybliżony obraz tego, czego możemy się spodziewać jako „typowej" wartości w zestawie danych. Jednak średnia może się znacznie zmienić, jeśli w danych występują wartości odstające. Dlatego ważne jest, aby znać inne miary statystyczne, takie jak mediana i moda, aby lepiej zrozumieć rozkład danych.

Mediana

Mediana to pojęcie statystyczne używane do pomiaru środkowej wartości cechy. W przypadku parzystej liczby danych o różnych wielkościach liczbowych połowa danych ma wartości mniejsze lub równe medianie, a druga połowa ma wartości większe lub równe medianie. Mediana jest jedną z podstawowych miar tendencji centralnej w statystyce i służy do opisu rozkładu danych, zwłaszcza gdy dane są skośne lub zawierają wartości odstające

Jak obliczyć medianę:

- Najpierw należy posortować zestaw danych od najmniejszej do największej wartości.
- Jeśli liczba danych jest parzysta (n), to mediana jest średnią dwóch środkowych wartości. Oznacza to, że mediana jest równa średniej wartości na pozycji $n/2$ i $(n/2 + 1)$ gdy dane są posortowane w porządku rosnącym.
- Jeśli liczba danych jest nieparzysta, wartość mediany znajduje się na środkowej pozycji.

Przykład:

Wyobraźmy sobie następujący zestaw danych przedstawiający liczbę godzin snu w danym okresie: 7, 6, 5, 8, 6, 9, 7

Najpierw uporządkujmy dane w kolejności rosnącej: 5, 6, 6, 7, 7, 8, 9

Ponieważ liczba danych jest nieparzysta (7), medianą będzie wartość na środkowej pozycji, która jest czwartą wartością w uporządkowanym zestawie danych. Zatem mediana w tym

przypadku wynosi 7 godzin. Oznacza to, że połowa osób w tym zbiorze danych śpi 7 lub mniej godzin, podczas gdy druga połowa śpi 7 lub więcej godzin.

Moda (dominanta)

Moda (dominanta) jest jednym z podstawowych wskaźników statystycznych używanych do pomiaru tendencji centralnej zbioru danych. Moda (dominanta) reprezentuje wartość, która występuje najczęściej w zbiorze danych. Jest to wartość, która ma najwyższą częstotliwość występowania spośród wszystkich wartości w zbiorze danych.

Moda (dominanta) jest przydatna do identyfikowania najczęstszych wartości w zbiorze danych i jest szczególnie przydatna podczas analizowania zmiennych jakościowych (kategorialnych), w których wartości są nienumeryczne.

Jeśli w zestawie danych występuje wiele mód (wiele wartości występujących z podobną maksymalną częstotliwością), mówimy o rozkładzie wielomodalnym. Jeśli wszystkie dane mają taką samą częstotliwość występowania, zestaw danych nie ma mody.

Przykład: wyobraźmy sobie zbiór danych reprezentujący kolory samochodów na parkingu:

Czerwony, Niebieski, Czerwony, Zielony, Niebieski, Niebieski, Czerwony

W tym przypadku modą jest „Czerwony", ponieważ ta wartość występuje najczęściej (trzy razy), podczas gdy „Niebieski" i „Zielony" występują rzadziej.

Moda (dominanta) jest prosta do obliczenia, ponieważ po prostu identyfikuje wartość o najwyższej częstotliwości występowania w zbiorze danych. Moda jest używana do opisywania charakterystycznych wartości w danych i może być przydatna w zrozumieniu, która wartość jest najbardziej charakterystyczna dla określonej sytuacji lub grupy.

Zakres zmienności (VR, zakres, interwał, amplituda)

Różnica między wartościami maksymalnymi i minimalnymi w zestawie danych jest pojęciem statystycznym zwanym zakresem. Mierzy on, jak duża jest różnica między wartościami maksymalnymi (maksimum) i minimalnymi (minimum) w zestawie danych. Zakres to prosty sposób na oszacowanie zakresu wartości w zestawie danych i zmierzenie zmienności między wartościami minimalnymi i maksymalnymi.

Obliczenie marginesu zmienności jest proste:

- Najpierw należy znaleźć wartość minimalną (min) i maksymalną (max) w zbiorze danych.
- Następnie obliczyć różnicę między wartością maksymalną i minimalną (max - min).

Przykład: wyobraźmy sobie zbiór danych reprezentujący wiek uczestników wydarzenia: 20, 25, 30, 35, 40. Aby obliczyć margines zmienności, należy najpierw znaleźć wartość minimalną (20) i maksymalną (40) w zbiorze danych. Następnie obliczamy różnicę między wartością maksymalną i minimalną: VR = 40 - 20 = 20

Zatem margines zmienności w tym przypadku wynosi 20 lat. Oznacza to, że różnica między najstarszym a najmłodszym uczestnikiem wynosi 20 lat.

Rozkład wariancji jest przydatny do oszacowania zakresu wartości w zbiorze danych, ale jest dość prosty i nie uwzględnia wszystkich wartości w zbiorze danych. Do bardziej szczegółowej analizy zmienności i rozproszenia danych powszechnie stosuje się inne miary statystyczne, takie jak wariancja lub kwartyle.

Wariancja i odchylenie standardowe

Wariancja to średnia kwadratów odchyleń od średniej. Jest to kwadrat odchylenia standardowego**. Odchylenie standardowe** jest miarą statystyczną używaną do pomiaru rozproszenia lub zmienności w zestawie danych. Informuje, jak daleko wartości są od średniej (przeciętnej) w zestawie. Odchylenie standardowe jest jedną z najczęściej używanych miar rozproszenia w statystyce i jest obliczane poprzez obliczenie pierwiastka kwadratowego z wariancji (wariancji)**.**

Obliczanie odchylenia standardowego:

- Najpierw należy obliczyć zmienność (wariancję). Zmienność (wariancję) oblicza się, biorąc średnią z kwadratów różnic od średniej.
- Po uzyskaniu wartości wariancji (σ^2), pierwiastkując ją, należy obliczyć odchylenie standardowe. Odbywa się to poprzez wyznaczenie pierwiastka kwadratowego z σ^2:

Odchylenie standardowe $\sigma = \sqrt{\sigma^2}$.

Odchylenie standardowe mierzy, jak rozproszone są wartości wokół średniej w zestawie danych. Wyższa wartość odchylenia standardowego oznacza, że wartości są bardziej rozproszone i bardziej różnią się od średniej, podczas

gdy niższa wartość odchylenia standardowego wskazuje na mniejsze rozproszenie.

Przykład: wyobraźmy sobie zbiór danych reprezentujący oceny uczniów z egzaminu z matematyki: 80, 85, 90, 75, 95. Wzór, który zostanie przedstawiony poniżej, jest ważny tylko wtedy, gdy pięć wartości, od których zaczęliśmy, tworzy całą populację. Najpierw obliczamy średnią, która wynosi 85. Następnie obliczamy margines zmienności, który wynosi 50.

Najpierw należy obliczyć odchylenia każdego punktu danych od średniej i podnieść wynik każdego z nich do kwadratu:

$(80-85)^2=(-5)^2=25,\quad (85-85)^2=(0)^2=0,\quad (90-85)^2=(5)^2=25,\quad (75-85)^2=(-10)^2=100,\ (95-85)^2=(10)^2=100$

Wariancja jest średnią tych wartości:

$$\sigma^2=\frac{25+0+25+100+100}{5}=\frac{250}{5}=50$$

Na koniec oblicza się odchylenie standardowe, biorąc pierwiastek kwadratowy z marginesu zmienności:

Odchylenie standardowe = $\sqrt{(50)}\ \approx\ 7{,}07$

Zatem odchylenie standardowe w tym przypadku wynosi około 7,07. Oznacza to, że średnio wyniki uczniów są oddalone od średniej o około 7,07 jednostki. Odchylenie standardowe jest często wykorzystywane do analizy rozkładu danych i oceny zmienności wartości w zbiorze.

Kwantyle

Kwantyle to wartości, które dzielą uporządkowane rosnąco dane na określone części. Na przykład kwartyle dzielą dane na cztery równe części. Pierwszy kwartyl (Q1) dzieli dolne 25% danych, drugi kwartyl (Q2) jest równy medianie, a trzeci kwartyl (Q3) oddziela górne 25% danych.

Przykład: w zbiorze danych 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, pierwszy kwartyl (Q1) jest równy 6, drugi kwartyl (Q2) jest równy 11, a trzeci kwartyl (Q3) jest równy 16.

1.4 Wyświetlanie statystyk

Prezentacja statystyk obejmuje wykorzystanie różnych metod i narzędzi w celu przedstawienia danych w jasny, przejrzysty i pouczający sposób.

Oto kilka popularnych sposobów wyświetlania statystyk:

Tabele

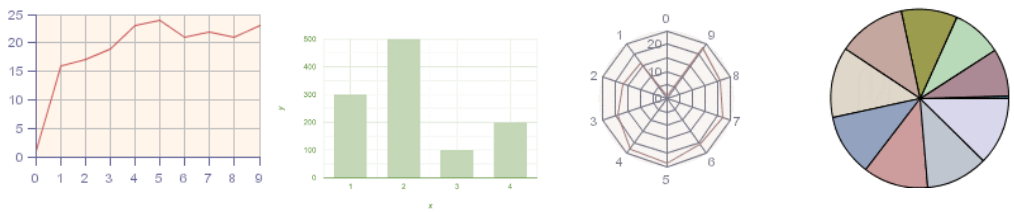
Tabele są podstawową metodą wyświetlania danych. Przykłady obejmują tabele częstotliwości, które pokazują liczbę wystąpień dla różnych wartości, oraz tabele danych, które pokazują więcej informacji o danych.

Oceny uzyskane przez studentów	Liczniki	Częstotliwość
41 - 49	𝍸	3
50 - 58	𝍸 |	6
59 - 67	𝍸	5
68 - 76	𝍸 |	6
77 - 85	||	2
		Razem=22

Rysunek 1.1 Przykładowa tabela częstości

Reprezentacje graficzne

Prezentacje graficzne są skutecznym narzędziem do wizualizacji danych. Obejmują one różne rodzaje wykresów, takie jak wykresy słupkowe, wykresy liniowe, wykresy kołowe, histogramy, wykresy pudełkowe itp.



Rysunek 1.2 Przykłady graficznej reprezentacji danych

Wykresy liniowe służą do wizualizacji trendów i zmian w czasie, dzięki czemu idealnie nadają się do śledzenia danych, które ewoluują w sposób ciągły. Są one szczególnie

skuteczne w pokazywaniu relacji między zmiennymi i podkreślaniu wzorców, takich jak wzrosty, spadki lub wahania. Wykresy liniowe są powszechnie stosowane w dziedzinach takich jak finanse, nauka i biznes do analizy danych szeregów czasowych, porównywania trendów w różnych kategoriach lub prognozowania przyszłych zmian na podstawie danych historycznych.

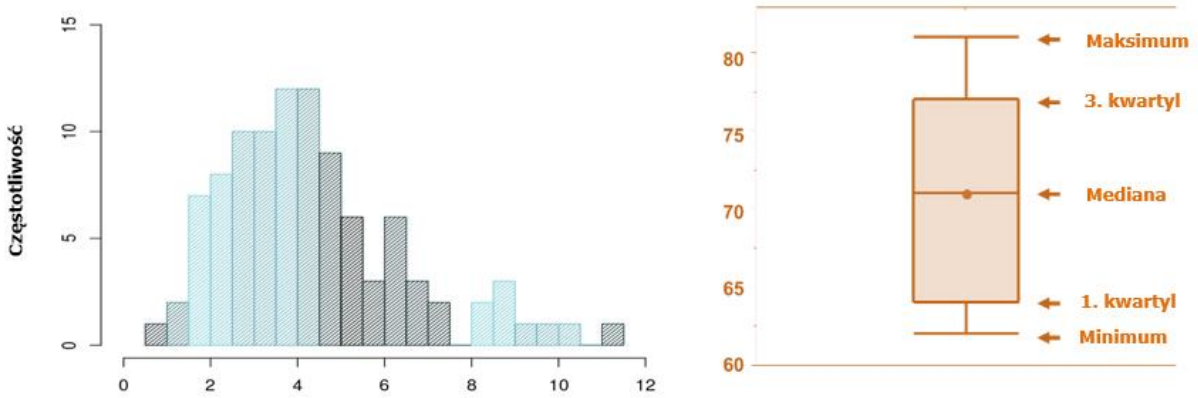
Wykresy słupkowe służą do porównywania ilości w różnych kategoriach, dzięki czemu idealnie nadają się do prezentacji danych dyskretnych. Są one szczególnie skuteczne w podkreślaniu różnic, podobieństw i trendów między grupami. Wykresy słupkowe są powszechnie używane, gdy trzeba pokazać częstotliwości, wartości procentowe lub inne miary liczbowe w jasny i prosty wizualnie sposób. Są szeroko stosowane w biznesie, edukacji i badaniach w celu analizowania i przekazywania kategorialnych.

Wykresy radarowe, znane również jako wykresy pająka, służą do wyświetlania danych wielowymiarowych w wielu wymiarach w formacie kołowym. Są one idealne do porównywania kilku zmiennych lub jednostek pod kątem tych samych kryteriów, podkreślając mocne i słabe strony w jasny, wizualny sposób. Wykresy radarowe są często wykorzystywane w analizie wydajności, podejmowaniu decyzji i porównaniach konkurencji, takich jak ocena cech produktu, umiejętności zespołu lub wyników ankiet w różnych kategoriach.

Wykresy kołowe służą do przedstawiania proporcji lub procentów całości, dzięki czemu idealnie nadają się do wizualizacji względnych rozmiarów różnych kategorii. Są szczególnie skuteczne, gdy chcesz pokazać, w jaki sposób części przyczyniają się do całości lub porównać proporcje na pierwszy rzut oka. Wykresy kołowe są powszechnie używane w raportach, prezentacjach i ankietach do wyświetlania danych, takich jak udział w rynku, alokacja budżetu lub rozkład demograficzny.

Histogramy

Histogramy to graficzne reprezentacje rozkładu danych. Są one używane do pokazania częstotliwości wartości zmiennej w różnych zakresach liczbowych.



Rysunek 1.3 Histogram i wykres kwantylowy (wykres pudełkowy)

Wykres kwantylowy (wykres pudełkowy)

Wykres kwantylowy to rodzaj wykresu stosowanego w statystyce opisowej jako wygodny sposób graficznego przedstawienia grup danych liczbowych poprzez podsumowanie ich pięcioma liczbami: minimum, pierwszy kwartyl, mediana, trzeci kwartyl i maksimum.

Wybór metody wyświetlania statystyk zależy od charakteru danych, celów analizy i grupy docelowej. Ważne jest, aby wybrać metodę, która najlepiej pasuje do przekazu i ułatwia zrozumienie danych.

1.5 Rozkład częstości

Rozkład częstości, znany również jako tabela częstości lub histogram, to sposób na pokazanie liczby wystąpień różnych wartości zmiennej w zestawie danych. Korzystając z rozkładu częstości, można zidentyfikować wzorce, rozkłady i częstotliwości wartości w danych. Jest on powszechnie używany do analizy zmiennych jakościowych (kategorialnych), ale może być również używany do wyświetlania dyskretnych wartości zmiennych ilościowych (numerycznych).

Proces tworzenia rozkładu częstości obejmuje następujące kroki:

- gromadzenie danych: najpierw należy zebrać dane, dla których ma zostać utworzony rozkład częstotliwości,

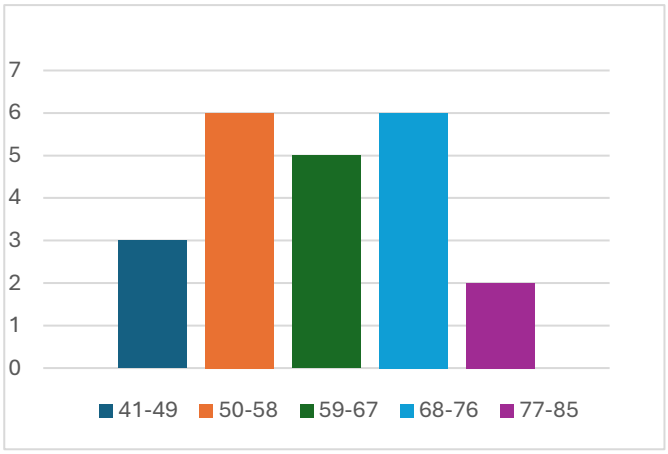
- zidentyfikowanie różnych wartości: należy zidentyfikować różne wartości, które pojawiają się w danych. Są to kategorie lub wartości dyskretne, które należy przeanalizować,
- zliczanie wystąpień: zliczanie, ile razy każda wartość pojawia się w zbiorze danych,
- utworzenie tabeli częstotliwości: należy utworzyć tabelę pokazującą wszystkie różne wartości zmiennej i liczbę wystąpień dla każdej wartości,
- rysowanie histogramu: w przypadku dużej liczby różnych wartości można utworzyć histogram przedstawiający rozkład częstotliwości. Jest to graficzna reprezentacja, która pokazuje liczbę wystąpień dla każdej wartości w postaci słupków.

Przykład rozkładu częstości: Wyobraźmy sobie, że analizowany jest rozkład częstości kolorów samochodów w salonie samochodowym. Zebrane zostały dane dotyczące kolorów 100 samochodów, należy więc sprawdzić, ile samochodów występuje w każdym kolorze?

Tabela 1.1 Nr rozkładu częstości

Punkty uzyskane przez studentów	Liczniki	Częstość
41 - 49	𝍸	3
50 - 58	𝍸 |	6
59 - 67	𝍸	5
68 - 76	𝍸 |	6
77 - 85	||	2
		Razem=22

Wykres rozkładu częstości (histogram) pokazywałby słupki dla każdego koloru z wysokością reprezentującą liczbę samochodów tego koloru. W ten sposób można wyraźnie zobaczyć, który kolor występuje najczęściej i jak rozłożone są inne kolory w zbiorze danych. Rozkłady częstości są przydatnym narzędziem do wizualizacji i analizy danych jakościowych oraz do szybkiego identyfikowania wzorców.



Rysunek 1.4 Wykres rozkładu częstości

1.6 Statystyka opisowa a wnioskowanie

Statystyka opisowa: Statystyka opisowa zajmuje się opisywaniem i podsumowywaniem danych z badanej próby lub populacji. Służy do analizy i zrozumienia danych, ale nie do wnioskowania na temat całej populacji. Głównym celem statystyki opisowej jest opisanie charakterystyki danych, na przykład obliczenie średniej, mediany, zakresu, odchylenia standardowego i stworzenie reprezentacji graficznych, takich jak histogramy lub wykresy. Służy do tworzenia podsumowań i wykresów, które pomagają wizualizować dane.

Statystyka z próby: statystyka inferencyjna zajmuje się wnioskowaniem o populacji na podstawie próby, do której wszystkie jednostki danych miały takie samo prawdopodobieństwo trafienia. Oznacza to, że statystyka wnioskowania pozwala na wyciąganie wniosków na temat całej populacji na podstawie analizy próby reprezentatywnej. Wykorzystuje różne metody statystyczne, takie jak testowanie hipotez, przedziały ufności i analiza regresji, aby zrozumieć, czy zaobserwowane wyniki próby można uogólnić na populację. Na przykład, jeśli chcemy dowiedzieć się, czy średnia wieku w próbie jest reprezentatywna dla całej populacji, użyjemy statystyk inferencyjnych.

Statystyka inferencyjna/matematyczna

Statystyka inferencyjna to gałąź statystyki, która koncentruje się na wnioskach i konkluzjach, jakie możemy wyciągnąć z zebranych danych. Jej głównym zadaniem jest wyciąganie ogólnych wniosków na temat populacji lub próby na podstawie analizy próbki danych.

Główne cele statystyki inferencyjnej to:

Szacowanie parametrów populacji: statystyka inferencyjna pozwala nam oszacować parametry populacji, takie jak średnia, wariancja, proporcje i inne cechy na podstawie próby.

Testowanie hipotez: statystyka inferencyjna może być wykorzystywana do testowania hipotez dotyczących populacji w oparciu o dane z próby. Obejmuje to testy statystyczne, w których porównujemy próbę z założeniami dotyczącymi populacji.

Tworzenie przedziałów ufności: statystyka inferencyjna pozwala nam obliczyć przedziały zawierające szacowane wartości parametrów populacji z określonym poziomem ufności.

Przykład statystyki inferencyjnej: załóżmy, że ma zostać oszacowany średni wzrost wszystkich studentów na uniwersytecie. Ponieważ niemożliwe jest sprawdzenie wszystkich studentów, do rozważań wzięta zostaje próba 100 studentów i to ich wzrost jest mierzony.

Następnie wykorzystujemy statystykę inferencyjną do obliczenia przedziału ufności dla średniego wzrostu wszystkich uczniów. Nasza próba charakteryzuje się średnim wzrostem równym 170 cm i odchyleniem standardowym wynoszącym 5 cm.

Zakładając, że wzrost uczniów w populacji ma w przybliżeniu **rozkład normalny**, można użyć błędu standardowego średniej do obliczenia przedziału ufności. Na przykład, jeśli ma zostać uzyskany 95% przedział ufności, należy użyć błędu standardowego i kwantyli rozkładu normalnego.

Przybliżony 95% przedział ufności dla średniego wzrostu wszystkich studentów na uniwersytecie wynosiłby:

$$170\ cm\ \pm\ 1{,}96\ \times\ (\frac{5\ cm}{\sqrt{100}})\ =\ 170\ cm\ \pm\ 0{,}98\ cm$$

Oznacza to, że można powiedzieć z 95%-ową pewnością, że średni wzrost wszystkich studentów wynosi od około 169,02 cm do 170,98 cm. Ten przedział ufności pozwala nam wywnioskować średni wzrost wszystkich studentów na uniwersytecie z ogólnej próby.

Łącznie opisane powyżej metody statystyczne pozwalają firmom logistycznym lepiej zrozumieć ich procesy, przewidywać przyszłe zdarzenia i podejmować bardziej świadome decyzje w celu poprawy wydajności i konkurencyjności.

1.7 Korelacja i regresja

Są to metody statystyczne wykorzystywane do badania współzależności między wielkościami zmiennymi i przewidywania warunkowej wartości zmiennej zależnej. Obie metody pomagają zrozumieć, na ile jedna zmienna wpływa na drugą i czy pierwsza zmienna może być wykorzystana do przewidywania? Poniżej znajduje się wyjaśnienie każdej z tych dwóch metod:

Korelacja

Korelacja służy do pomiaru stopnia powiązania między dwiema zmiennymi ilościowymi (liczbowymi). Określa, czy istnieje liniowa zależność między dwiema zmiennymi i jak silna jest ta zależność. Korelacja jest mierzona za pomocą współczynnika korelacji, który przyjmuje **wartości od -1** do **1.**

Współczynnik korelacji równy 1 stanowi doskonałą korelację dodatnią, co oznacza, że zmienne są doskonale skorelowane i poruszają się w tym samym kierunku.

Współczynnik korelacji równy -1 oznacza idealną korelację ujemną, co oznacza, że dwie zmienne są całkowicie odwrotnie skorelowane i poruszają się w przeciwnych kierunkach.

Współczynnik korelacji równy 0 oznacza, że nie ma liniowej zależności między zmiennymi.

Przykład: korelacja między liczbą godzin nauki a ocenami uzyskiwanymi przez studentów będzie dodatnia, jeśli wzrost liczby godzin nauki zwykle odpowiada wyższym ocenom.

Regresja

Regresja służy do modelowania i przewidywania wartości jednej zmiennej ilościowej (zmiennej zależnej) na podstawie wartości innej zmiennej ilościowej (zmiennej niezależnej). Istnieją różne rodzaje regresji, w tym **prosta regresja liniowa, wieloraka regresja liniowa**, regresja logistyczna itp.

Regresja liniowa prosta: służy do modelowania związku między jedną zmienną niezależną a jedną zmienną zależną. Model jest liniowy i jest zwykle reprezentowany przez równanie linii prostej ($y = a + bx$), gdzie a jest punktem przecięcia z osią y a b jest nachyleniem linii.

Wieloraka regresja liniowa: używana, gdy należy modelować związek między kilkoma zmiennymi niezależnymi a jedną zmienną zależną.

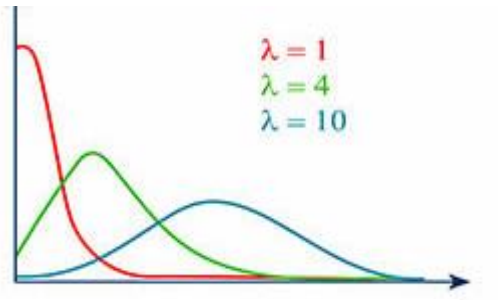
Rysunek 1.5 Wykresy prostej regresji liniowej

Przykład: regresja liniowa prosta może być wykorzystana do modelowania związku między liczbą ukończonych zadań edukacyjnych (zmienna niezależna) a oceną z egzaminu końcowego (zmienna zależna).

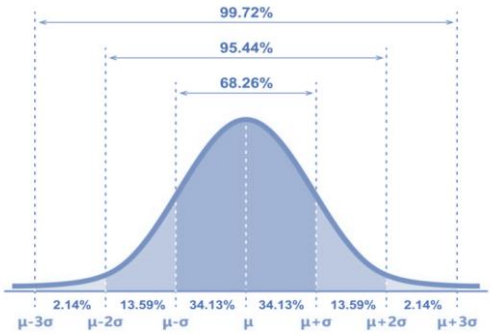
1.8 Rozkłady prawdopodobieństwa

W statystyce rozkład prawdopodobieństwa opisuje prawdopodobieństwa różnych wartości, jakie może przyjąć zmienna. Jest to model matematyczny, który pomaga nam zrozumieć i analizować zjawiska losowe oraz przewidywać rozkład wartości w określonych okolicznościach. Istnieje kilka różnych rozkładów prawdopodobieństwa, z których każdy ma swoją własną charakterystykę i zastosowanie w różnych sytuacjach. Oto niektóre z najbardziej znanych rozkładów prawdopodobieństwa w statystyce:

Rozkład normalny (Gaussa): rozkład normalny jest jednym z najważniejszych i najszerzej stosowanych rozkładów. Opisuje on symetryczny rozkład w kształcie dzwonu ze znanymi parametrami: średnią (μ) i odchyleniem standardowym (σ). Wiele naturalnych zjawisk jest zbliżonych do rozkładu normalnego.



Rysunek 1.6 Wykres rozkładu Poissona



Rysunek 1.7 Wykres rozkładu normalnego

Rozkład dwumianowy: rozkład dwumianowy jest używany do modelowania liczby sukcesów (np. liczby wyrzuconych „orzełków") w danej liczbie niezależnych prób Bernoulliego. Ma on dwa parametry: liczbę prób (n) i prawdopodobieństwo sukcesu(p).

Rozkład Poissona: rozkład Poissona służy do modelowania liczby zdarzeń występujących w danym okresie czasu lub przestrzeni. Jest on zwykle używany do modelowania rzadkich zdarzeń, takich jak wypadki, wezwania służb ratunkowych itp. Parametrem rozkładu jest średnia liczba wystąpień zdarzenia(λ).

Rozkład wykładniczy: rozkład wykładniczy jest szczególnym przypadkiem rozkładu gamma i jest używany do modelowania czasów do pierwszego zdarzenia w procesie Poissona. Parametrem rozkładu jest średnia (λ).

Rozkład t-Studenta: rozkład t-Studenta jest używany do szacowania przedziałów ufności i testowania hipotez, gdy masz małą próbkę i nie znasz odchylenia standardowego populacji. Jest to ważne podczas analizy próbek, w przypadku których założenie o rozkładzie normalnym nie ma uzasadnienia.

Rozkład chi-kwadrat: rozkład chi-kwadrat jest używany do analizy rozkładu częstotliwości w tabelach, testowania niezależności i testowania hipotez. Jest on często wykorzystywany w testach statystycznych, takich jak test chi-kwadrat.

Rozkład F: rozkład F jest używany do porównywania zmienności między dwiema próbkami. Jest używany w analizie wariancji (ANOVA) i innych testach statystycznych.

Te rozkłady prawdopodobieństwa są podstawowymi elementami składowymi w statystyce i są wykorzystywane do modelowania i analizowania różnych typów danych w różnych kontekstach. Wybór właściwego rozkładu prawdopodobieństwa ma kluczowe znaczenie podczas przeprowadzania analiz statystycznych i przewidywania wyników.

BIBLIOGRAFIA DO ROZDZIAŁU 1.

- Introductory Statistics. Bentham Science Publishers, Kahl, A. (Publish 2023). DOI:10.2174/9789815123135123010l
- Introductory Statistics 2e, Openstax, Rice University, Houston, Texas 77005, Jun 23, senior contributing authors: Barbara Illowsky and Susan dean, De anza college, Publish Date: Dec 13, 2023, (https://openstax.org/details/books/introductory-statistics-2e);
- Introductory Statistics 4th Edition, Susan Dean and Barbara Illowsky, Adapted by Riyanti Boyd & Natalia Casper (Published 2013 by OpenStax College) July 2021, (http://dept.clcillinois.edu/mth/oer/IntroductoryStatistics.pdf)
- Introductory Statistics 7th Edition, Prem S. Mann, eastern Connecticut state university with the help of Christopher Jay Lacke, Rowan university, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030-5774, 2011
- Journal of the Royal Statistical Society 2024**,** A reputable journal publishing cutting-edge research and articles on various aspects of statistics, including theoretical advancements and practical applications. Recent issues have featured studies on sampling and hypothesis testing
- Introduction to statistics, made easy second edition, Prof. Dr. Hamid Al-Oklah Dr. Said Titi Mr. Tareq Alodat, March 2014
- Statistics for Business and Economics, Thirteenth Edition, David R. Anderson, Dennis J. Sweeney, Thomas A. Williams, Jeffrey D. Camm, James J. Cochran, 2017, 2015 Cengage Learning®
- Statistics for Business, First edition, Derek L Waller, 2008 Copyright © 2008, Derek L Waller, Published by Elsevier Inc. All rights reserved

Dodatkowe linki do literatury i filmów na Youtube do Rozdziału 1

https://open.umn.edu/opentextbooks/textbooks/196

https://www.scribbr.com/category/statistics/

https://stats.libretexts.org/Bookshelves/Introductory_Statistics

https://assets.openstax.org/oscms-prodcms/media/documents/IntroductoryStatistics-OP_i6tAI7e.pdf

https://saylordotorg.github.io/text_introductory-statistics/

https://drive.uqu.edu.sa/_/mskhayat/files/MySubjects/20178FS%20Elementary%20Statistics/Introductory%20Statistics%20(7th%20Ed).pdf

https://dept.clcillinois.edu/mth/oer/IntroductoryStatistics.pdf

https://www.geeksforgeeks.org/introduction-of-statistics-and-its-types/

https://onlinestatbook.com/Online_Statistics_Education.pdf

https://www.researchgate.net/profile/Tareq-Alodat-2/publication/340511098_INTRODUCTION_TO_STATISTICS_MADE_EASY/links/5e8de3dc4585150839c7b58a/INTRODUCTION-TO-STATISTICS-MADE-EASY.pdf

https://byjus.com/maths/statistics/

https://www.khanacademy.org/math/statistics-probability

https://www.youtube.com/watch?v=XZo4xyJXCak

https://www.youtube.com/watch?v=LMSyiAJm99g

https://www.youtube.com/watch?v=VPZD_aij8H0

https://www.youtube.com/watch?v=TLwp5DwcqD4

https://www.youtube.com/watch?v=fpFj1Re1l84

https://youtube.com/playlist?list=PLqzoL9-eJTNAB5st3mtP_bmXafGSH1Dtz&si=z-IXQ1iKbw2-ieJW

https://www.youtube.com/watch?v=44MJyNTxaP8

2. Statystyka dla analityki biznesowej

Witamy w świecie statystyki biznesowej, w którym dane przekształcają się w znaczące spostrzeżenia, kierując podejmowaniem decyzji i odkrywając ukryte prawdy. W tej kompleksowej eksploracji wyruszamy w podróż, aby zdemistyfikować podstawowe koncepcje i techniki statystyczne, które stanowią podstawę rygorystycznej analizy danych biznesowych. Od zrozumienia zawiłości rozkładów do stosowania testowania hipotez i konstruowania przedziałów ufności, każdy rozdział rozwija nowy aspekt umiejętności statystycznych.

W sercu analizy statystycznej leży **rozkład normalny**, krzywa w kształcie dzwonu, która przenika niezliczone zjawiska w przyrodzie i ludzkim zachowaniu. W tej części zagłębiamy się w istotę rozkładu normalnego, odkrywając jego właściwości i znaczenie we wnioskowaniu statystycznym. Za pomocą wizualizacji i rzeczywistych przykładów wyjaśniamy wszechobecność tego fundamentalnego rozkładu i jego rolę jako kamienia węgielnego teorii statystycznej.

Odchylenie standardowe służy jako kompas w krajobrazie statystycznym, prowadząc nas przez zmienność nieodłącznie związaną ze zbiorami danych. W tym rozdziale zostanie dokonana analiza pojęcia odchylenia standardowego, ujawniając jego znaczenie w ilościowym określaniu dyspersji i ocenie rozrzutu punktów danych. Uzbrojony w głębsze zrozumienie odchyleń standardowych, będziesz poruszać się po danych z pewnością siebie, precyzyjnie rozpoznając wzorce i wartości odstające.

Zmienne stanowią elementy składowe analizy statystycznej, a każda z nich posiada odrębne cechy i implikacje. Niniejszy rozdział wyjaśnia dychotomię między zmiennymi ciągłymi i dyskretnymi, rzucając światło na ich rolę w modelowaniu i interpretacji danych. Dzięki zrozumieniu niuansów związanych z typami zmiennych, będziesz mógł w pełni wykorzystać potencjał technik statystycznych dostosowanych do różnych struktur danych.

Rozkład statystyki próbkowania służy jako podstawa wnioskowania statystycznego, wypełniając lukę między obserwacjami z próby a parametrami populacji. W tym rozdziale rozwikłamy koncepcję rozkładu próbkowania, wyjaśniając jego znaczenie w formułowaniu probabilistycznych stwierdzeń dotyczących charakterystyki populacji. Dzięki konkretnym

przykładom rozwiniesz intuicyjne zrozumienie roli rozkładu próbkowania w solidnej analizie statystycznej.

Centralne Twierdzenie Graniczne jest jak latarnia morska statystycznego oświecenia, oświetlająca drogę do wiarygodnego wnioskowania w obliczu niepewności. Niniejszy rozdział demistyfikuje Centralne Twierdzenie Graniczne, ujawniając jego transformacyjną moc w stabilizowaniu średnich prób i ułatwianiu testowania hipotez. Uzbrojony w spostrzeżenia zebrane w tym rozdziale, będziesz mógł wykorzystać Centralne Twierdzenie Graniczne do wyciągnięcia znaczących wniosków z danych empirycznych.

Zrozumienie **testowania hipotez** jest niezbędne do podejmowania decyzji opartych na danych. Pozwala nam określić, czy zaobserwowane wzorce w danych są znaczące, czy po prostu wynikają z przypadku. Stosując testowanie hipotez, możemy oceniać założenia, porównywać grupy i oceniać istotność statystyczną wyników, dzięki czemu jest to niezbędne narzędzie w badaniach naukowych, analizach biznesowych i wielu innych dziedzinach

Wyniki *Z* i tabele *Z* służą jako pomoce nawigacyjne w morzu standardowego rozkładu normalnego, ułatwiając znormalizowane porównania i obliczenia prawdopodobieństwa. Niniejszy rozdział wyjaśnia zawiłości współczynników *Z*, umożliwiając interpretację znormalizowanych wyników i wykorzystanie tabel Z do analizy statystycznej. Dzięki biegłej znajomości współczynników Z będziesz poruszać się po rozległym obszarze rozkładu normalnego z pewnością i precyzją.

W sytuacjach, w których liczebność próby jest niewielka lub nieznane jest odchylenie standardowe populacji, **współczynniki *t*** i **tabele *t*** stają się niezbędnymi narzędziami analizy statystycznej. Niniejszy rozdział odkrywa tajemnice wyników *t*, prowadząc przez ich obliczanie i interpretację za pomocą tabel *t*. Uzbrojony w tę wiedzę, będziesz poruszać się po niuansach rozkładów *t* z finezją, zapewniając solidne wnioskowanie w różnych scenariuszach statystycznych.

Rozkład normalny i rozkład *t* są filarami teorii prawdopodobieństwa, a każdy z nich posiada unikalne cechy i zastosowania. W tym rozdziale wyjaśniamy różnice między tymi rozkładami, umożliwiając rozeznanie, kiedy zastosować każdy z nich w analizie statystycznej. Dzięki praktycznym przykładom i analizom porównawczym wybierzesz spośród rozkładów normalnego i *t*, wzbogacając swój zestaw narzędzi statystycznych.

Przedziały ufności zapewniają wgląd w niepewność otaczającą parametry populacji, umożliwiając nam ilościowe określenie precyzji naszych szacunków. W tym rozdziale zbadamy konstrukcję przedziałów ufności dla średnich i częstotliwości/odsetka/frakcji, odkrywając metodologię i interpretację tych podstawowych narzędzi statystycznych. Dzięki opanowaniu przedziałów ufności będziesz w stanie przekazać niepewność nieodłącznie związaną z wynikami w sposób przejrzysty i rygorystyczny.

Podczas gdy **wartości *p*** stanowią bramę do wnioskowania statystycznego, ich błędna interpretacja może prowadzić do błędnych wniosków i błędnych decyzji.
Niniejszy rozdział analizuje potencjalne pułapki związane z nadmiernym poleganiem na wartościach *p*, podkreślając znaczenie kontekstu i wielkości efektu w analizie statystycznej. Dzięki krytycznej analizie i praktycznym spostrzeżeniom będziesz poruszać się po złożoności wartości *p* z zachowaniem czujności, zapewniając integralność wniosków statystycznych.

Na tych stronach znajdują się klucze do odblokowania tajemnic analizy statystycznej, umożliwiając poruszanie się po złożoności danych z pewnością i precyzją. Wyruszając w tę podróż razem, niech ciekawość będzie naszym kompasem, a dociekanie naszym światłem przewodnim, oświetlającym drogę do głębszego zrozumienia i praktycznych spostrzeżeń.

2.1 Rozkład normalny

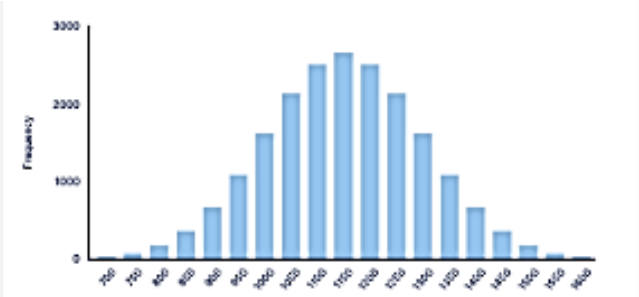
W sercu analizy statystycznej leży rozkład normalny, wszechobecny rozkład prawdopodobieństwa, który służy jako punkt odniesienia dla wielu technik statystycznych. Zagłębimy się w jego charakterystykę, symetryczną krzywą w kształcie dzwonu i jego znaczenie w zrozumieniu rozkładu danych.

Rozkład normalny znajduje zastosowanie w różnych dziedzinach, w tym w finansach, psychologii, inżynierii i biologii. Od modelowania cen akcji po zrozumienie rozkładu ludzkiego wzrostu, rozkład normalny służy jako wszechstronne narzędzie do analizy i interpretacji danych.

W tym rozdziale zagłębimy się w matematyczne właściwości rozkładu normalnego, badając, jak obliczać prawdopodobieństwa, percentyle i wskaźniki *z*-score. Ponadto omówimy praktyczne techniki wizualizacji i interpretacji rozkładów normalnych za pomocą histogramów, wykresów gęstości i funkcji rozkładu skumulowanego.

Pod koniec tego rozdziału będziesz kierował się głębokim uznaniem dla rozkładu normalnego i jego znaczenia w analizie statystycznej. Uzbrojony w tę wiedzę, będziesz dobrze przygotowany do radzenia sobie z bardziej zaawansowanymi koncepcjami statystycznymi i stosowaniem ich w rzeczywistych zestawach danych. Wyruszmy razem w tę podróż, aby rozwikłać tajemnice rozkładu normalnego.

Rozkład normalny, znany również jako rozkład Gaussa lub krzywa dzwonowa, wykazuje symetryczny rozkład danych bez asymetrii. Na wykresie dane tworzą krzywą w kształcie dzwonu, przy czym większość wartości gromadzi się wokół środka i maleje w miarę oddalania się od niego.



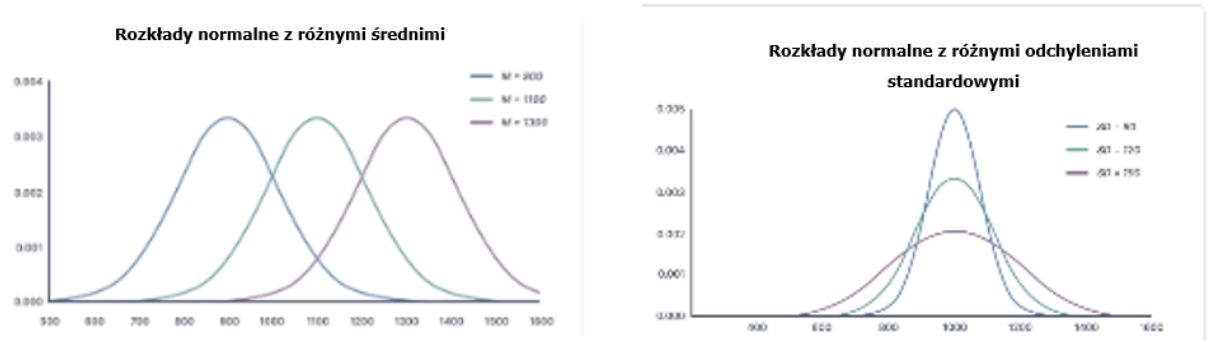
Rysunek 2.1 Przykład rozkładu gaussowskiego lub krzywej dzwonowej

Różne zmienne w naukach przyrodniczych i społecznych zazwyczaj wykazują rozkład normalny lub jego przybliżenie. Przykłady obejmują wzrost, wagę urodzeniową, umiejętność czytania, satysfakcję z pracy i wyniki SAT. Ze względu na powszechność zmiennych o rozkładzie normalnym, liczne testy statystyczne są dostosowane do takich populacji. Biegłość w zrozumieniu charakterystyki rozkładów normalnych upoważnia osoby do stosowania statystyk wnioskowania do porównywania grup i generowania szacunków populacji z próbek.

Rozkłady normalne mają kluczowe cechy, które łatwo zauważyć na wykresach:

- średnia, mediana i dominanta są dokładnie takie same,
- rozkład jest symetryczny względem średniej - połowa wartości spada poniżej średniej, a połowa powyżej średniej,
- rozkład można opisać za pomocą dwóch wartości: średniej i odchylenia standardowego.

Średnia służy jako parametr lokalizacji, dyktując środek szczytu krzywej. Dostosowanie średniej odpowiednio przesuwa krzywą: zwiększenie jej przesuwa krzywą w prawo, a zmniejszenie przesuwa krzywą w lewo. Tymczasem odchylenie standardowe działa jako parametr skali, wpływając na rozrzut lub szerokość krzywej.



Rysunek 2.2 Rozkład normalny z różnymi średnimi i różnymi odchyleniami

Odchylenie standardowe rozciąga lub ściska krzywą. Małe odchylenie standardowe skutkuje wąską krzywą, podczas gdy duże odchylenie standardowe prowadzi do szerokiej krzywej.

2.2 Reguła empiryczna

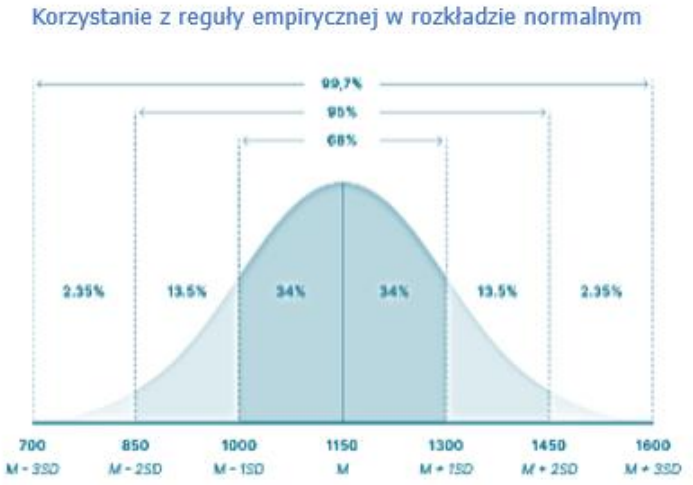
Reguła empiryczna, znana również jako reguła 68-95-99,7, zapewnia wgląd w rozkład wartości w rozkładzie normalnym:

- około 68% wartości mieści się w zakresie jednego odchylenia standardowego od średniej,
- około 95% wartości mieści się w zakresie dwóch odchyleń standardowych od średniej,
- około 99,7% wartości mieści się w zakresie trzech odchyleń standardowych od średniej.

Na przykład, rozważmy scenariusz, w którym zbierane są wyniki SAT od studentów na nowym kursie przygotowującym do testu, a dane są zgodne z rozkładem normalnym ze średnim wynikiem (M) 1150 i odchyleniem standardowym (SD) 150.

Zastosowanie reguły empirycznej daje następujące wnioski:

- około 68% wyników mieści się w zakresie od 1000 do 1300, co odpowiada 1 odchyleniu standardowemu powyżej i poniżej średniej,
- około 95% wyników mieści się w zakresie od 850 do 1450, co stanowi 2 odchylenia standardowe powyżej i poniżej średniej,
- prawie wszystkie wyniki, około 99,7%, mieszczą się w zakresie od 700 do 1600, obejmując 3 odchylenia standardowe powyżej i poniżej średniej.



Rysunek 2.3 Reguła empiryczna w rozkładzie normalnym

Reguła empiryczna oferuje szybką metodę oceny danych, umożliwiając wykrywanie wartości odstających lub wyjątkowych, które odbiegają od oczekiwanego wzorca. W przypadkach, w których dane z małych próbek znacznie odbiegają od tego wzorca, bardziej odpowiednie mogą być alternatywne rozkłady, takie jak rozkład *t*. Identyfikacja rozkładu zmiennej pozwala na zastosowanie odpowiednich testów statystycznych.

2.3 Wzór krzywej normalnej

Rysunek 2.4 Wyniki SAT wykazują rozkład zbliżony do normalnego

W funkcji gęstości prawdopodobieństwa obszar pod krzywą reprezentuje prawdopodobieństwo. Biorąc pod uwagę, że rozkład normalny służy jako rozkład prawdopodobieństwa, skumulowany obszar pod krzywą niezmiennie sumuje się do 1 lub 100%. Chociaż wzór na normalną funkcję gęstości prawdopodobieństwa może wydawać się skomplikowany, jego wykorzystanie wymaga jedynie znajomości średniej populacji i odchylenia standardowego. Podstawiając te parametry do wzoru, można określić gęstość prawdopodobieństwa związaną z dowolną wartością x.

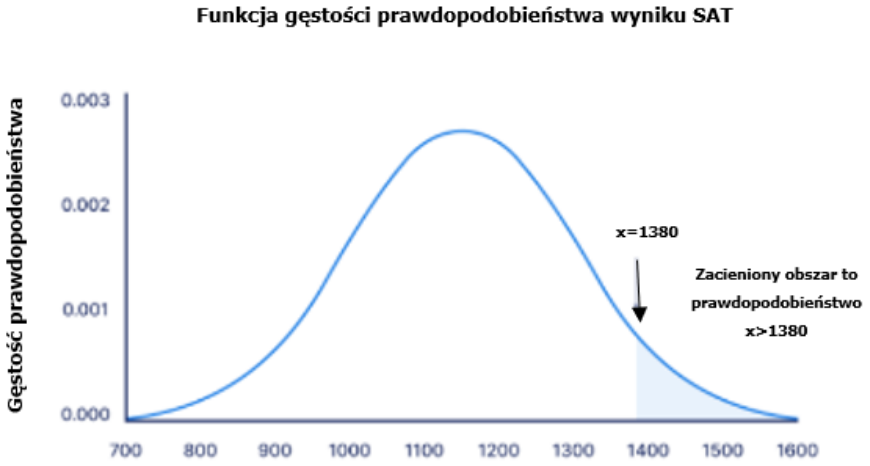
- $f(x)$ = prawdopodobieństwo,
- x = wartość zmiennej,
- μ = średnia arytmetyczna,
- σ = odchylenie standardowe,
- σ^2 = wariancja.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Przykład:

Korzystając z funkcji gęstości prawdopodobieństwa, chcesz poznać prawdopodobieństwo, że wyniki egzaminu SAT w próbie przekroczą 1380 punktów.

Na wykresie funkcji gęstości prawdopodobieństwa, prawdopodobieństwo jest zacieniowanym obszarem pod krzywą, który leży na prawo od miejsca, w którym wyniki SAT są równe 1380.



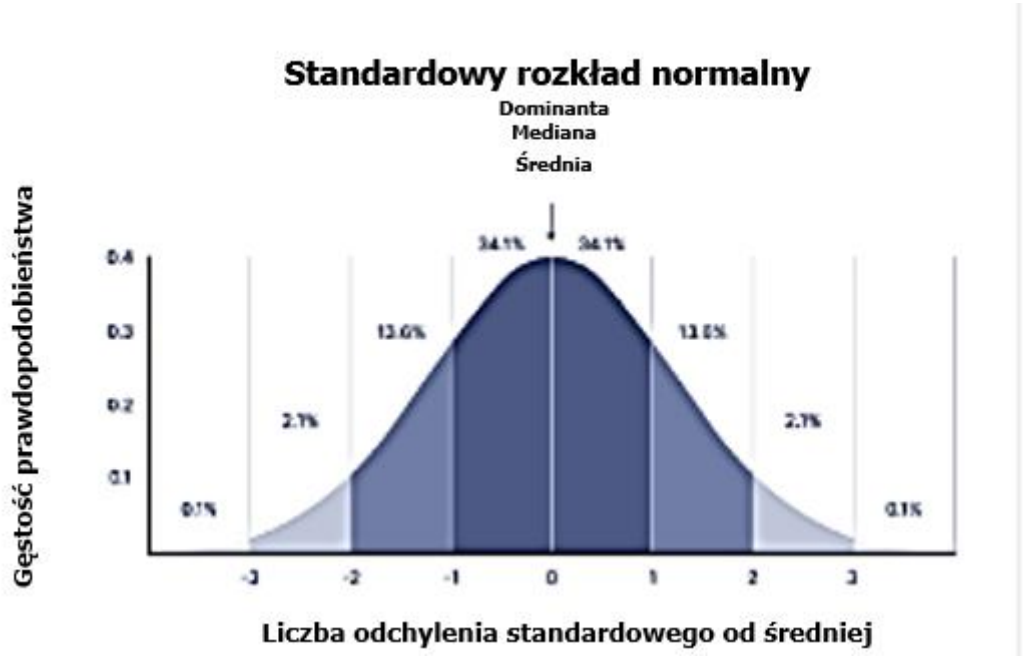
Rysunek 2.5 Wykres funkcji gęstości prawdopodobieństwa

Wartość prawdopodobieństwa tego wyniku można znaleźć przy użyciu standardowego rozkładu normalnego.

2.4 Standardowy rozkład normalny

Standardowy rozkład normalny, znany jako **rozkład *z*** wyróżnia się tym, że ma średnią równą 0 i odchylenie standardowe równe 1. Każdy rozkład normalny może być postrzegany jako transformacja standardowego rozkładu normalnego, podlegająca dostosowaniom skali, położenia lub obu tych elementów.

W kontekście rozkładu *z*, poszczególne obserwacje, które są zwykle oznaczane jako *x* w rozkładach normalnych, są określane jako wskaźnik *z*. Reprezentują one liczbę odchyleń standardowych, o jaką każda wartość odbiega od średniej. W związku z tym przekształcenie wartości z dowolnego rozkładu normalnego we wskaźniki *z* ułatwia porównanie i analizę w ramach standardowego rozkładu normalnego.



Rysunek 2.6 Wykres standardowego rozkładu normalnego

Wystarczy znać średnią i odchylenie standardowe rozkładu, aby znaleźć wskaźnik *z* dla danej wartości.

Objaśnienie wzoru na wskaźnik *z*:

- x = wartość indywidualna,
- μ = średnia arytmetyczna,
- σ = odchylenie standardowe.

$$z = \frac{x - \mu}{\sigma}$$

Rozkład normalny jest przekształcany w standardowy rozkład normalny z kilku powodów:

- aby znaleźć prawdopodobieństwo, że obserwacje w rozkładzie znajdą się powyżej lub poniżej danej wartości,
- aby znaleźć prawdopodobieństwo, że średnia z próby istotnie różni się od znanej średniej z populacji,
- aby porównać wyniki z rozkładów z różnymi średnimi i odchyleniami standardowymi.

2.5 Znajdowanie prawdopodobieństwa przy użyciu rozkładu *z*

Każdy wskaźnik *z* odpowiada prawdopodobieństwu, często określanemu jako wartość *p*, wskazując prawdopodobieństwo zaobserwowania wartości poniżej określonego wskaźnika *z*.

Przekształcając pojedynczą wartość we wskaźnik *z*, można określić prawdopodobieństwo wystąpienia wszystkich wartości do tego punktu w rozkładzie normalnym.

Rozważmy na przykład scenariusz, w którym chcemy ustalić prawdopodobieństwo, że wyniki SAT w rozważanej próbie przekroczą 1380. Początkowo oblicza się wynik z przy użyciu średniej i odchylenia standardowego rozkładu. Przy średniej 1150 i odchyleniu standardowym 150, wskaźnik *z* ujawnia liczbę odchyleń standardowych, o które 1380 odbiega od średniej.

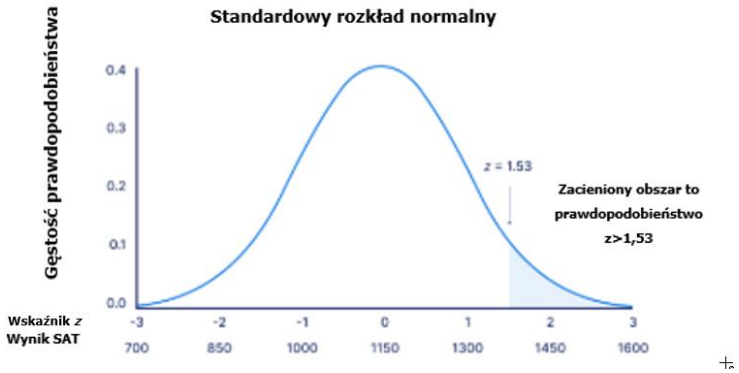
Wzór	**Obliczenie**
$z = \frac{x - \mu}{\sigma}$	$z = \frac{1380 - 1150}{150} = 1{,}53$

Dla wyniku z równego 1,53 wartość *p* wynosi 0,937. Jest to prawdopodobieństwo, że wynik SAT wynosi 1380 lub mniej (93,7%) i jest to obszar pod krzywą na lewo od zacieniowanego obszaru.

Aby znaleźć zacieniony obszar, należy odjąć 0,937 od 1, co stanowi całkowity obszar pod krzywą.

Prawdopodobieństwo x > 1380 = 1 – 0,937 = **0,063**

Oznacza to, że prawdopodobnie tylko 6,3% wyników SAT w próbie przekracza 1380.



Rysunek 2.7 Standardowy rozkład normalny ze wskazanym wynikiem

2.6 Rozkład próbkowania (rozkład próby)

Rozkłady próbkowania stanowią podstawę wnioskowania statystycznego, umożliwiając nam wyciąganie wniosków na temat populacji na podstawie danych z próby. Zagłębimy się w zawiłości rozkładów próbkowania, rozumiejąc, w jaki sposób odzwierciedlają one zmienność statystyk próby i ich kluczową rolę w testowaniu hipotez.

Rozkład próby odnosi się do rozkładu statystyki, takiej jak średnia próby lub odsetek z próby, uzyskanej z wielu prób o tej samej wielkości pobranych z populacji. Zapewnia wgląd w zachowanie statystyk próby i ich zmienność w różnych próbach.

2.7 Centralne Twierdzenie Graniczne i Rozkład Próbkowania

Centralne Twierdzenie Graniczne (CLT) to podstawowa koncepcja w statystyce, która leży u podstaw zachowania rozkładów próbkowania. Stwierdza ono, że rozkład średniej z próby zbliża się do rozkładu normalnego wraz ze wzrostem wielkości próby, niezależnie od kształtu rozkładu populacji. Twierdzenie to umożliwia nam wyciąganie wiarygodnych wniosków na temat parametrów populacji na podstawie danych z próby.

Centralne twierdzenie graniczne służy jako kamień węgielny zrozumienia rozkładów normalnych w statystyce. W warunkach badawczych uzyskanie dokładnego oszacowania średniej populacji często wiąże się z gromadzeniem danych z wielu losowych próbek w populacji. Te indywidualne średnie z próbek wspólnie tworzą tak zwany rozkład próbkowania średniej.

Centralne twierdzenie graniczne określa dwie kluczowe zasady:

1. **Prawo wielkich liczb:** Wraz ze wzrostem wielkości próby lub liczby prób, średnia z próby ma tendencję do zbliżania się do średniej z populacji.
2. **Normalność rozkładu próbkowania:** Pomimo oryginalnego rozkładu zmiennej, podczas pracy z wieloma dużymi próbkami, rozkład próbkowania średniej ma tendencję do zbliżania się do rozkładu normalnego.

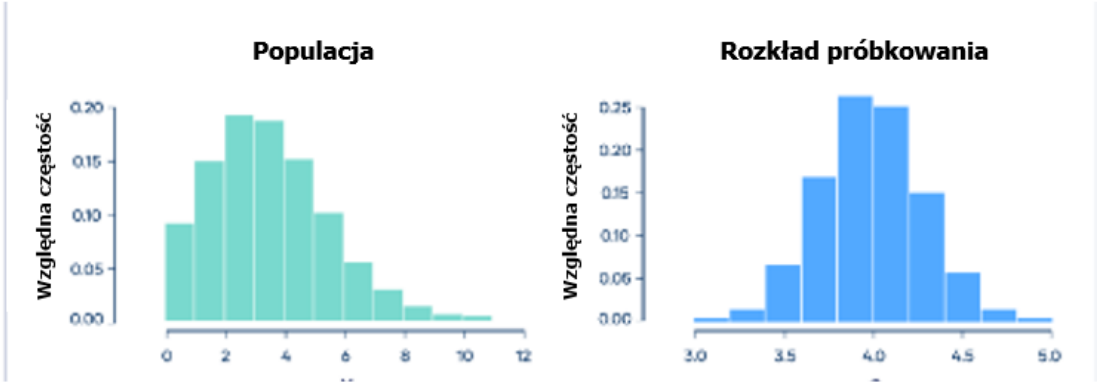
Parametryczne testy statystyczne konwencjonalnie zakładają, że próbki pochodzą z populacji o rozkładzie normalnym. Jednak centralne twierdzenie graniczne eliminuje konieczność tego założenia dla wystarczająco dużych prób. Przy dużych próbach testy parametryczne mogą być stosowane niezależnie od rozkładu populacji, pod warunkiem spełnienia innych istotnych

założeń. Wielkość próby wynosząca 30 lub więcej jest powszechnie uważana za wystarczająco dużą.

Z drugiej strony, w przypadku małych prób, zapewnienie założenia normalności ma kluczowe znaczenie ze względu na niepewność związaną z rozkładem próbkowania średniej. Dokładne wyniki wymagają potwierdzenia, że populacja jest zgodna z rozkładem normalnym przed zastosowaniem testów parametrycznych z małymi próbkami.

Przykładowo, centralne twierdzenie graniczne zakłada, że uzyskując wystarczająco duże próbki z populacji, średnie tych próbek będą wykazywać rozkład normalny, nawet jeśli podstawowy rozkład populacji odbiega od normalności.

Przykład: Rozważmy populację o rozkładzie Poissona (przedstawioną na lewym obrazku). Po pobraniu 10 000 próbek z tej populacji, z których każda składa się z 50 obserwacji, rozkład średnich próbek jest ściśle zgodny z rozkładem normalnym, zgodnie z centralnym twierdzeniem granicznym (jak pokazano na prawym obrazku).



Rysunek 2.8 Przykład populacji o rozkładzie Poissona i rozkładzie średnich z dużych próbek

Centralne twierdzenie graniczne opiera się na pojęciu rozkładu próbkowania, który reprezentuje rozkład prawdopodobieństwa statystyki obliczonej z wielu próbek pobranych z populacji.

Konceptualizacja eksperymentu może pomóc w zrozumieniu rozkładów próbkowania:

- Wyobraźmy sobie losowanie próbki z populacji i obliczanie statystyki, takiej jak średnia.
- Następnie losowana jest kolejna próba o identycznej wielkości, a średnia jest obliczana ponownie.

- Proces ten jest powtarzany wiele razy, w wyniku czego powstaje mnóstwo średnich, z których każda odpowiada próbce.

Agregacja tych średnich z próby stanowi przykład rozkładu próbkowania. Zgodnie z centralnym twierdzeniem granicznym, rozkład próbkowania średniej dąży do rozkładu normalnego, gdy wielkość próby jest wystarczająco duża. Co ciekawe, niezależnie od rozkładu populacji - normalnego, Poissona, dwumianowego lub innego - rozkład próbkowania średniej wykazuje normalność.

Na szczęście nie trzeba wielokrotnie próbkować populacji, aby określić kształt rozkładu próbkowania. Zamiast tego parametry rozkładu próbkowania średniej są zależne od parametrów samej populacji.

- Średnia rozkładu próbkowania jest średnią populacji.

$$\mu_{\bar{x}} = \mu$$

- Odchylenie standardowe rozkładu próbkowania to odchylenie standardowe populacji podzielone przez pierwiastek kwadratowy z wielkości próby.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Można opisać rozkład próbkowania średniej za pomocą tej notacji:

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

gdzie:

- $\bar{X}$ to rozkład próbkowania średnich z próby,
- $\sim$ oznacza „podąża za rozkładem”,
- N jest rozkładem normalnym,
- μ jest średnią arytmetyczną populacji,
- σ jest odchyleniem standardowym populacji,
- n to wielkość próby.

Wielkość próby, oznaczana jako *n*, reprezentuje liczbę obserwacji pobranych z populacji dla każdej próby, zachowując jednorodność we wszystkich próbach. Wielkość próby znacząco wpływa na rozkład średniej w dwóch kluczowych aspektach.

1. Wielkość próby i normalność:
    - Większe rozmiary próbek zwykle dają rozkłady próbkowania, które ściśle przylegają do rozkładu normalnego.
    - I odwrotnie, przy małej liczebności próby rozkład próbkowania średniej może odbiegać od normalności. Ta rozbieżność powstaje, ponieważ ważność centralnego twierdzenia granicznego zależy od posiadania „wystarczająco dużej" wielkości próby.
    - Konwencjonalnie, próba o wielkości 30 lub więcej jest uważana za „wystarczająco dużą".
    - Gdy $n<30$, centralne twierdzenie graniczne nie ma zastosowania, a rozkład próbkowania odzwierciedla rozkład populacji. W związku z tym rozkład próbkowania jest normalny tylko wtedy, gdy rozkład populacji jest normalny.
    - I odwrotnie, gdy $n\geq30$, centralne twierdzenie graniczne jest prawdziwe, a rozkład próbkowania jest zbliżony do rozkładu normalnego.

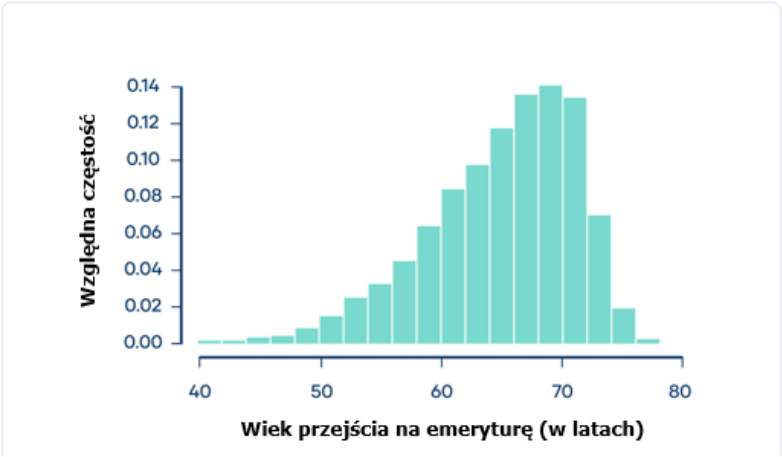
2. Wielkość próby i odchylenia standardowe:
    - Wielkość próby bezpośrednio wpływa na odchylenie standardowe rozkładu próbkowania, odzwierciedlając zmienność lub rozrzut rozkładu.
    - Przy mniejszej liczebności próby odchylenie standardowe jest zazwyczaj wyższe, co wskazuje na większą zmienność między średnimi z próby ze względu na ich niedokładne oszacowanie średniej populacji.
    - I odwrotnie, większe liczebności próby odpowiadają niższym odchyleniom standardowym, wskazując na mniejszą zmienność między średnimi z próby ze względu na ich dokładniejsze oszacowanie średniej populacji.

Znaczenie Centralnego Twierdzenia Granicznego:

Testy parametryczne, takie jak testy t, ANOVA i regresja liniowa, mają większą moc statystyczną w porównaniu z większością testów nieparametrycznych. Ta zwiększona moc statystyczna wynika z założeń dotyczących rozkładu populacji, które opierają się na centralnym twierdzeniu granicznym.

Rozkład ciągły

Rozważmy wiek emerytalny osób w Stanach Zjednoczonych. Populacja składa się ze wszystkich emerytowanych Amerykanów, a rozkład tej populacji można przedstawić w następujący sposób:



Rysunek 2.9 Wykres rozkładu ciągłego

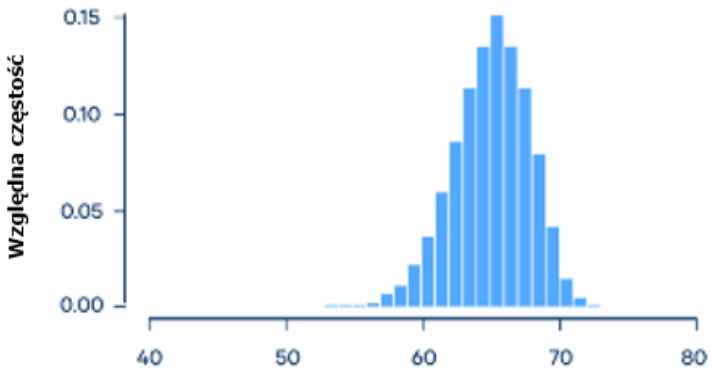
Rozkład wieku emerytalnego jest pochyla się w lewą stronę, a większość osób przechodzi na emeryturę w ciągu około pięciu lat od średniego wieku emerytalnego wynoszącego 65 lat. Istnieje jednak wydłużony ogon osób przechodzących na emeryturę znacznie wcześniej, na przykład w wieku 50 lub nawet 40 lat. Odchylenie standardowe w populacji wynosi 6 lat.

Wyobraźmy sobie, że przeprowadzamy próbę na małą skalę z tej populacji. Pięciu emerytów jest wybieranych losowo, a ich wiek emerytalny jest rejestrowany. Na przykład: 68, 73, 70, 62, 63.

Średnia z tej próby służy jako przybliżenie średniej populacji, aczkolwiek z ograniczoną precyzją ze względu na małą wielkość próby wynoszącą 5. Na przykład: Średnia = (68 + 73 + 70 + 62 + 63) / 5. Średnia = 67,2 lat.

Załóżmy teraz, że ten proces próbkowania jest powtarzany 10 razy, a każda próbka obejmuje pięciu emerytów. Obliczana jest średnia z każdej próby, co daje rozkład znany jako rozkład próbkowania średniej. Na przykład: 60,8; 57,8; 62,2; 68,6; 67,4; 67,8; 68,3; 65,6; 66,5; 62,1.

Ponieważ proces ten jest powtarzany wiele razy, histogram przedstawiający średnie tych próbek będzie zbliżony do rozkładu normalnego.



Rysunek 2.10 Normalny rozkład średnich

Pomimo tego, że rozkład próbkowania wykazuje nieco bardziej normalny kształt w porównaniu z populacją, nadal zachowuje lekkie odchylenie w lewo. Ponadto widać, że zmienność w rozkładzie próbkowania jest węższa niż w populacji.

Choć rozkład próbkowania wykazuje nieco bardziej normalny kształt w porównaniu z populacją, nadal zachowuje lekkie odchylenie w lewo. Ponadto widać, że zmienność w rozkładzie próbkowania jest węższa niż w populacji.

2.8 Statystyka testowa

Statystyka testowa reprezentuje wartość liczbową pochodzącą z testu hipotezy statystycznej, wskazującej stopień dopasowania obserwowanych danych do rozkładu oczekiwanego przy hipotezie zerowej tego testu.

Statystyka te odgrywa kluczową rolę w obliczaniu wartości *p* wyników, ułatwiając określenie, czy należy przyjąć, czy odrzucić hipotezę zerową?

Ale co dokładnie stanowi statystykę testową?

Statystyka testowa określa podobieństwo między rozkładem danych a rozkładem przewidywanym przy hipotezie zerowej zastosowanego testu statystycznego. Rozkład danych wyjaśnia częstotliwość każdej obserwacji, charakteryzując się tendencją centralną

i zmiennością wokół niej. Ponieważ różne testy statystyczne przewidują różne typy rozkładów, wybór odpowiedniego testu jest zgodny z postawioną hipotezą.

Statystyka testowa kondensuje obserwowane dane w pojedynczą liczbę, wykorzystując takie miary, jak tendencja centralna, zmienność, wielkość próby i liczba zmiennych predykcyjnych w modelu statystycznym.

Zazwyczaj statystykę testową wyłania się z dostrzegalnych wzorców w danych (np. korelacji między zmiennymi lub rozbieżności między grupami), podzielonych przez wariancję danych (tj. odchylenie standardowe).

Rozważmy następujący przykład:

Badany jest związek między temperaturą a datami kwitnienia określonego rodzaju jabłoni. Analizując kompleksowy zbiór danych obejmujący 25 lat, śledząc temperaturę i daty kwitnienia poprzez losowe pobieranie próbek 100 drzew rocznie z pola eksperymentalnego.

- Hipoteza zerowa (H_0): Nie istnieje korelacja między temperaturą a terminem kwitnienia.
- Hipoteza alternatywna (H_A lub H_1): Istnieje korelacja między temperaturą a terminem kwitnienia.

Aby sprawdzić tę hipotezę, należy przeprowadzić test regresji, uzyskując wartość *t* jako statystykę testową. Ta wartość *t* zestawia zaobserwowaną korelację między zmiennymi z hipotezą zerową o zerowej korelacji.

2.9 Rodzaje statystyk testowych

Poniżej, w tabeli 2.1, znajduje się podsumowanie najpopularniejszych statystyk testowych wraz z odpowiadającymi im hipotezami i kategoriami testów statystycznych, w których są one stosowane. Podczas gdy różne testy statystyczne mogą wykorzystywać różne metodologie do obliczania tych statystyk, podstawowe hipotezy i interpretacje statystyk testu pozostają spójne.

Tabela 2.1 Zestawienie najpopularniejszych statystyk testowych

Statystyka testowa	Hipotezy zerowe i alternatywne	Testy statystyczne, które ją wykorzystują
Wartość *t*	**Hipoteza zerowa:** Średnie dwóch grup są równe **Hipoteza alternatywna:** Średnie dwóch grup nie są równe.	• Test *T* • Testy regresji
Wartość *z*	**Hipoteza zerowa:** Średnie dwóch grup są równe. **Hipoteza alternatywna:** Średnie dwóch grup nie są równe.	• Test *Z*
Wartość *F*	**Hipoteza zerowa:** Zmienność między dwiema lub więcej grupami jest większa lub równa zmienności wewnątrz grup. **Hipoteza alternatywna:** Zmienność między dwiema lub więcej grupami jest mniejsza niż zmienność wewnątrz grup.	• ANOVA • ANCOVA • MANOVA
Wartość X^2	**Hipoteza zerowa:** Dwie próby są niezależne. **Hipoteza alternatywna:** Dwie próby nie są niezależne (tj. są skorelowane)	• Test chi-kwadrat • Nieparametryczne testy korelacji

W rzeczywistych scenariuszach zazwyczaj oblicza się statystykę testową za pomocą pakietu oprogramowania statystycznego, takiego jak R, SPSS lub Excel, który dostarczy również wartość *p* z nią powiązaną. Niemniej jednak, wzory do ręcznego obliczania tych statystyk można znaleźć w Internecie.

Na przykład, testując hipotezę dotyczącą temperatury i daty kwitnienia, przeprowadzasz analizę regresji. Test regresji daje następujące wyniki:

- współczynnik regresji 0,36,
- wartość t porównującą ten współczynnik z przewidywanym zakresem współczynników regresji przy hipotezie zerowej o braku związku.

Wynikowa wartość *t* z testu regresji 0,36, reprezentuje statystykę testową.

2.10 Błąd standardowy

Błąd standardowy średniej (SE lub SEM) służy jako wskaźnik prawdopodobnej rozbieżności między średnią populacji a średnią próby. Zapewnia wgląd w stopień zmienności, jaki można by przewidzieć w średniej próbie, gdyby badanie zostało powtórzone przy użyciu świeżych próbek pobranych z tej samej populacji.

Podczas gdy błąd standardowy średniej jest najczęściej wymienianą formą błędu standardowego, podobne miary istnieją dla innych parametrów statystycznych, takich jak mediany lub proporcje. Błąd standardowy funkcjonuje jako powszechny miernik błędu próbkowania, przedstawiając rozbieżność między parametrem populacji a statystyką próby.

Aby zmniejszyć błąd standardowy, zaleca się zwiększenie wielkości próby. Zatrudnienie dużej, randomizowanej próby służy jako najskuteczniejsza strategia minimalizowania stronniczości próbkowania i zwiększania wiarygodności ustaleń.

Błąd standardowy i **odchylenie standardowe** są miarami zmienności:

- **Odchylenie standardowe** opisuje zmienność **w obrębie pojedynczej próbki**.
- **Błąd standardowy** szacuje zmienność **w wielu próbkach** populacji.

Odchylenie standardowe służy jako statystyka opisowa pochodząca bezpośrednio z danych próbki, podczas gdy błąd standardowy reprezentuje statystykę wnioskowania, zwykle szacowaną, chyba że znany jest dokładny parametr populacji.

2.11 Wzór na błąd standardowy

Błąd standardowy średniej jest określany poprzez zastosowanie odchylenia standardowego wraz z wielkością próby. Ze wzoru wynika, że wielkość próby i błąd standardowy mają odwrotną zależność. Mówiąc prościej, wraz ze wzrostem wielkości próby, błąd standardowy maleje. Zjawisko to występuje, ponieważ większa próba zwykle daje statystykę próby bliższą parametrowi populacji.

Stosowane są różne wzory w zależności od tego, czy znane jest odchylenie standardowe populacji. Wzory te mają zastosowanie do próbek zawierających więcej niż 20 elementów (n > 20).

Gdy znane są parametry populacji

Gdy znane jest odchylenie standardowe populacji, można użyć go w poniższym wzorze, aby dokładnie obliczyć błąd standardowy.

Wzór	Objaśnienie
$SE = \frac{\sigma}{\sqrt{n}}$	• SE to błąd standardowy, • σ to odchylenie standardowe populacji, • n to liczba elementów w próbie.

Gdy parametry populacji są nieznane

Gdy odchylenie standardowe populacji jest nieznane, można użyć poniższego wzoru do oszacowania błędu standardowego. Wzór ten przyjmuje odchylenie standardowe z próby jako oszacowanie punktowe dla odchylenia standardowego populacji.

Wzór	Objaśnienie
$SE = \frac{s}{\sqrt{n}}$	• SE to błąd standardowy, • s to odchylenie standardowe próby, • n to liczba elementów w próbie.

Przykład: Korzystanie z formuły błędu standardowego. Aby oszacować błąd standardowy dla wyników SAT z matematyki, należy wykonać dwa kroki.
Najpierw należy znaleźć pierwiastek kwadratowy z wielkości próby (*n*).

Wzór	Objaśnienie
$n = 200$	$\sqrt{n} = \sqrt{200} = 14{,}1$

Następnie należy podzielić odchylenie standardowe próbki przez liczbę znalezioną w kroku pierwszym.

Wzór	Objaśnienie
$SE = \frac{s}{\sqrt{n}}$	$s = 180 \quad \sqrt{n} = 14{,}1 \quad \frac{s}{\sqrt{n}} = \frac{180}{14{,}1} = 12{,}8$

Błąd standardowy wyników SAT z matematyki wynosi 12,8.
Można przedstawić błąd standardowy obok średniej lub włączyć go do przedziału ufności, aby przekazać niepewność związaną ze średnią.

Przykład: Przedstawienie średniej i błędu standardowego. Średni wynik egzaminu SAT z matematyki dla losowej próby zdających wynosi 550 ± 12,8 (SE).

Podawanie błędu standardowego w przedziale ufności jest preferowane, ponieważ eliminuje potrzebę wykonywania przez czytelników dodatkowych obliczeń w celu uzyskania znaczącego zakresu.

Przedział ufności oznacza zakres wartości, w których nieznany parametr populacji będzie znajdował się najczęściej, jeśli badanie zostanie powtórzone z nowymi losowymi próbkami.

Przy poziomie ufności 95% oczekuje się, że 95% wszystkich średnich z próby będzie mieścić się w przedziale ufności obejmującym ± 1,96 błędu standardowego średniej z próby. Przedział ten służy jako oszacowanie, w którym prawdziwy parametr populacji znajduje się z 95% pewnością.

Przykład: Konstruujesz 95% przedział ufności (CI), aby oszacować średni wynik SAT z matematyki w populacji. Biorąc pod uwagę charakterystykę o rozkładzie normalnym, taką jak wyniki SAT, około 95% wszystkich średnich z próby mieści się w zakresie około 4 błędów standardowych średniej z próby.

Wzór na przedział ufności

$CI = \bar{x} \pm (1{,}96 \times SE)$

$\bar{x}$ = średnia z próby = 550
SE =błąd standardowy = 12,8

Dolna wartość graniczna	**Górna wartość graniczna**
$\bar{x} - (1{,}96 \times SE)$	$\bar{x} + (1{,}96 \times SE)$
$550 - (1{,}96 \times 12{,}8) =$ **525**	$550 + (1{,}96 \times 12{,}8) =$ **575**

W przypadku losowego doboru próby 95% przedział ufności (CI) [525 575] mówi, że istnieje prawdopodobieństwo 0,95, dla którego średni wynik SAT z matematyki w populacji wynosi od 525 do 575.

BIBLIOGRAFIA DO ROZDZIAŁU 2.

- Introductory Statistics. Bentham Science Publishers, Kahl, A. (Publish 2023). DOI:10.2174/9789815123135123010 1
- Introductory Statistics 2e, Openstax, Rice University, Houston, Texas 77005, Jun 23, senior contributing authors: Barbara Illowsky and Susan dean, De anza college, Publish Date: Dec 13, 2023, (https://openstax.org/details/books/introductory-statistics-2e);
- Introductory Statistics 4th Edition, Susan Dean and Barbara Illowsky, Adapted by Riyanti Boyd & Natalia Casper Published 2013 by OpenStax College) July 2021, (http://dept.clcillinois.edu/mth/oer/IntroductoryStatistics.pdf);
- Journal of the Royal Statistical Society 2024, A reputable journal publishing cutting-edge research and articles on various aspects of statistics, including theoretical advancements and practical applications. Recent issues have featured studies on sampling and hypothesis testing.
- Introductory Statistics 7th Edition, Prem S. Mann, eastern Connecticut state university with the help of Christopher Jay Lacke, Rowan university, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030-5774, 2011
- Introduction to statistics, made easy second edition, Prof. Dr. Hamid Al-Oklah Dr. Said Titi Mr. Tareq Alodat, March 2014
- Statistics for Business and Economics, Thirteenth Edition, David R. Anderson, Dennis J. Sweeney, Thomas A. Williams, Jeffrey D. Camm, James J. Cochran, 2017, 2015 Cengage Learning®

- Statistics for Business, First edition, Derek L Waller, 2008 Copyright © 2008, Derek L Waller, Published by Elsevier Inc. All rights reserved

Dodatkowe linki do literatury i filmów na Youtube do rozdziału 2

https://open.umn.edu/opentextbooks/textbooks/196

https://www.scribbr.com/category/statistics/

https://stats.libretexts.org/Bookshelves/Introductory_Statistics

https://assets.openstax.org/oscms-prodcms/media/documents/IntroductoryStatistics-OP_i6tAI7e.pdf

https://saylordotorg.github.io/text_introductory-statistics/

https://drive.uqu.edu.sa/_/mskhayat/files/MySubjects/20178FS%20Elementary%20Statistics/Introductory%20Statistics%20(7th%20Ed).pdf

https://dept.clcillinois.edu/mth/oer/IntroductoryStatistics.pdf

https://www.geeksforgeeks.org/introduction-of-statistics-and-its-types/

https://onlinestatbook.com/Online_Statistics_Education.pdf

https://www.researchgate.net/profile/Tareq-Alodat-2/publication/340511098_INTRODUCTION_TO_STATISTICS_MADE_EASY/links/5e8de3dc4585150839c7b58a/INTRODUCTION-TO-STATISTICS-MADE-EASY.pdf

https://byjus.com/maths/statistics/

https://www.khanacademy.org/math/statistics-probability

3. Zarządzanie danymi

W elektronicznej wymianie danych B2B (EDI) komunikaty zawierające kody produktów lub usług, identyfikatory jednostek transportowych, a także dokumenty prawne są wymieniane między partnerami w łańcuchu dostaw. Zazwyczaj mają one postać ciągów alfanumerycznych.

3.1 Informacja-Dane-Wiedza

W komputerowych systemach informatycznych reprezentacja informacji różni się w zależności od ich przeznaczenia i zastosowania. Może być alfanumeryczna lub binarna, biorąc pod uwagę fakt, czy reprezentuje tekst, rysunek, dźwięk lub program wykonywalny. Aby móc przechowywać, wyszukiwać, przetwarzać i przesyłać dane różnych typów (np. liczby, znaki, daty, waluty np.), muszą one być odpowiednio zakodowane. Alfanumeryczne formaty reprezentacji danych mają swoje korzenie w alfabecie ASCII (ask-key), ewoluowały w sensie krajowych zestawów znaków (np. standardy 8859-1, Latin 1 i 8859-2, Latin 2) i ostatecznie przekształciły się w międzynarodowe formaty UTF-8 i UTF-16. Umożliwiają one wspólną interpretację danych przez partnerów biznesowych należących do różnych grup etnicznych i środowisk geograficznych. Podczas gdy ciągi alfanumeryczne zależą od ich kodowania, dane numeryczne różnią się głównie rozmiarem i/lub precyzją.

Proces przekształcania danych z ich oryginalnej postaci analogowej do postaci cyfrowej jest popularnie nazywany digitalizacją. Odpowiednio zaprojektowane programy użytkowe i aplikacyjne akceptujące dane z różnych źródeł (np. skanery optyczne, czujniki elektryczne, EDI np.) umożliwiają organizacjom automatyzację pozyskiwania, przechowywania, przetwarzania i przesyłania danych w ramach i pomiędzy komputerowymi systemami informatycznymi.

Po zebraniu, dane różnych typów mogą być łączone i organizowane w tabele danych, bazy danych, hurtownie danych (porządek chronologiczny) lub bazy wiedzy (porządek koncepcyjny). Wyższe poziomy organizacji danych pozwalają na zautomatyzowaną

klasyfikację, wnioskowanie i reprezentację zgromadzonej w ten sposób wiedzy do celów analitycznych.

3.2 Dane logistyczne

W logistyce EDI służy do przesyłania danych transakcyjnych między partnerami biznesowymi. Ponieważ mogą oni używać różnych języków i aplikacji, ich konwersja do wspólnego formatu (np. XML, JSON) jest konieczna, aby mogły być interpretowane przez systemy informatyczne różnych partnerów (W3schools, 2023). W celu szybkiej identyfikacji i manipulacji opracowano kody kreskowe i znaczniki RFID.

Aby umożliwić współpracę międzynarodową, konieczne było zdefiniowanie globalnie akceptowanych formatów danych dla celów logistycznych. Formaty danych logistycznych odpowiadają etykietom usług i produktów, identyfikatorom jednostek transportowych i kodom transakcji, zwykle mającym postać ciągów alfanumerycznych ze znacznikiem czasu. Dla uproszczenia manipulacji i szybkości przetwarzania, kody te zostały ustandaryzowane i zakodowane jako optycznie czytelne kody kreskowe lub elektromagnetycznie czytelne kody identyfikacji radiowej (RFID).

Kody kreskowe (EAN/UCC) to wielobranżowa i międzynarodowa forma numerowania przedmiotów (POS EAN-8 i EAN-13, zmienny EAN-128, pasek danych, opakowanie ITF-14, QR, matryca danych np.) Są one wykorzystywane do identyfikacji produktów, partii produktów lub przesyłek (kody 1D), a także usług (kody 2D). Spośród kodów kreskowych 2D kod QR jest najbardziej popularny i może być odczytywany również przez smartfony, co zwiększa jego użyteczność w różnych obszarach zastosowań.

Kody RFID są przede wszystkim używane w taki sam sposób jak kody kreskowe. Umożliwiają one jednoznaczną identyfikację przedmiotów lub usług. Zazwyczaj, oprócz opcjonalnego kodu kreskowego, etykiety RFID zawierają jeszcze więcej informacji na chipie wielkości główki od szpilki. Oprócz identyfikacji, etykiety RFID umożliwiają rejestrowanie informacji o śledzeniu, często wymaganych w zastosowaniach logistycznych. W przeciwieństwie do kodów kreskowych, RFID umożliwia ich skanowanie bez bezpośredniej linii wzroku, a także wielu etykiet jednocześnie.

Dzięki standardowi GS1 EPC Gen2 (ISO/IEC 18000-6:2013) ustanowiono standard technologiczny określający komunikację między tagami RFID a czytnikami. Podobnie jak kody kreskowe, standardy EPCglobal łączą technologię RFID ze znakowaniem EPC produktów, logistycznych jednostek transportowych, lokalizacji, zapasów, przedmiotów zwrotnych, dokumentów np. W celu bezpośredniej, zautomatyzowanej identyfikacji i śledzenia jednostek logistycznych w łańcuchach dostaw.

Standardy EPCglobal stanowią również podstawę GDSN (ang. Global Data Synchronization Network). Umożliwia ona zautomatyzowane pozyskiwanie i wymianę danych specyfikacji produktów i ich opakowań, co pozwala przedsiębiorstwom na centralne zarządzanie tymi danymi, które mogą być zamiennie wykorzystywane przez nie i ich partnerów.

Tabela 3.1 podsumowuje różne technologie identyfikacji wraz z ich zastosowaniami. Przedstawia ona różne jedno- i dwuwymiarowe kody kreskowe, a także różne klasy kodów RFID wraz z ich możliwościami.

Tabela 3.1 Technologie znakowania

Technologia	Zastosowanie
Pasek jednowymiarowy	Artykuły detaliczne i komponenty produktów
Pasek dwuwymiarowy	Usługi (np. UPS, bilety lotnicze), produkty hurtowe wymagające śledzenia
RFID klasy 1 (pasywne, znaczniki R)	Produkty wymagające masowej identyfikacji, kontroli dostępu
RFID klasy 2 (pasywne, znaczniki RW)	Produkty wymagające śledzenia
RFID klasy 3 (półaktywne, znaczniki RW)	Kontrola dostępu z dodatkowymi informacjami o śledzeniu
RFID klasy 4 (aktywne, znaczniki RW)	Śledzenie i namierzanie w przestrzeni zamkniętej
RFID klasy 5 (aktywne znaczniki/interrogatory)	Śledzenie i namierzanie otwartej przestrzeni, usługi zbliżeniowe z włączonymi urządzeniami, usługi oparte na lokalizacji

Przyszłe trendy w znakowaniu, śledzeniu i namierzaniu podążają w dwóch głównych kierunkach: miniaturyzacji i różnorodności. Kody kreskowe (1D) umożliwią również znakowanie miniaturowych przedmiotów (np. kapsułek medycznych). Nowatorskie kody matrycowe (2D) nie tylko umożliwią korekcję błędów podczas skanowania, ale także szyfrowanie danych.

RFID nadal rozprzestrzenia się na inne obszary zastosowań, takie jak identyfikacja przez dostawców usług (np. karty kolejowe, rejestracja w pracy np.), płatności zbliżeniowe (np. bezprzewodowe przelewy pieniężne, płatności w automatach sprzedających) i inteligentne rozwiązania (np. inteligentne zarządzanie domem, zdalna obsługa inteligentnych urządzeń np.), a także e-waluty.

3.3 Organizacja danych

Oprócz posiadania określonego formatu, dane mogą być zorganizowane na różne sposoby, aby ułatwić zarządzanie nimi, ich przetwarzanie i prezentację. Chociaż dane wejściowe są w większości nieustrukturyzowane, ich przechowywanie, przesyłanie i przetwarzanie zwiększa ich organizację. W dalszej części przedstawiono typowe formy organizacji danych o rosnącej złożoności, od formatów częściowo ustrukturyzowanych (np. CSV) do ustrukturyzowanych (np. Arkusze kalkulacyjne, bazy danych itp.).

Arkusze kalkulacyjne

Pierwszą formą organizacji danych są dwuwymiarowe tablice pól, zwane również tabelami lub arkuszami kalkulacyjnymi. Zazwyczaj pierwszy wiersz arkusza kalkulacyjnego oznacza znaczenie wartości przechowywanych w kolumnach, po których następują wiersze danych.

Pole tabeli lub komórka to najmniejsza jednostka danych. Ma określony typ (ciąg znaków, liczba, data, waluta np.). Jej zawartość można adresować za pomocą oznaczeń wierszy i kolumn (np. A1, reprezentujące pierwszy wiersz kolumny A).

Każdy wiersz tabeli jest grupą powiązanych pól, reprezentujących rekord (np. transakcję, rekord studenta, dane produktu np.). Ponieważ wszystkie wiersze tabeli mają tę samą strukturę, możemy zdefiniować typ rekordu jako listę atrybutów (np. dane ucznia (imię, nazwisko, data urodzenia, miejsce urodzenia, identyfikator...) odpowiednich typów danych.

Bazy danych

Plik lub tabela bazy danych to zbiór rekordów tego samego typu. Baza danych (DB) składa się z wielu wzajemnie połączonych tabel. Stąd definicja bazy danych według ANSI:

- Dane z bazy danych są ze sobą połączone i posortowane,
- Dane z bazy danych mogą być jednocześnie wykorzystywane przez wielu użytkowników,
- Dane w bazie danych składają się z unikalnych rekordów,,
- Baza danych jest przechowywana w komputerze.

Z powyższej definicji można wyciągnąć pewne wnioski na temat architektury klient – serwer, w której serwer przechowuje bazę danych, do której dostęp mają jego klienci. Oczywiście, aby uzyskać dostęp do bazy danych, należy ustanowić sieć komunikacyjną między serwerem a jego klientami. Serwer DB jest zwykle określany jako jego „back-end", podczas gdy klienci reprezentują „front-end". System zarządzania bazą danych (DBMS) na serwerze umożliwia swoim klientom dostęp do danych przechowywanych w bazie danych za pośrednictwem interfejsu aplikacji (API) i funkcji DBM. Funkcje DBM to mechanizmy umożliwiające wprowadzanie, pobieranie, przetwarzanie i prezentację danych DB. Aby wywołać te funkcje, zdefiniowano standardowe języki zapytań (SQL).

Model relacyjnej bazy danych

Istnieją różne formy organizacji DB, przy czym najbardziej powszechny jest model relacyjny (RDB). Podstawową ideą tego modelu jest fakt, że użytkownik nie może z góry znać wszystkich możliwych zastosowań danych przechowywanych w bazie danych. Ponieważ zwykle nie ma ustalonych ścieżek przeszukiwania plików bazy danych, opracowano różne języki zapytań do pobierania danych i manipulowania nimi. Model RDB opiera się na koncepcji encji i relacji:

- Encja to osoba/rzecz/koncepcja, którą można jednoznacznie zidentyfikować i która posiada atrybuty,
- Relacja reprezentuje sposób powiązania dwóch lub więcej jednostek danych.

Tabele RDB, reprezentujące encje lub relacje, są połączone za pomocą kluczy. Zestaw atrybutów, które jednoznacznie identyfikują encję, nazywany jest kluczem podstawowym. Gdy klucz podstawowy pojawia się jako pole w innej tabeli w celu wypełnienia relacji z oryginalną tabelą, nazywany jest kluczem drugorzędnym lub obcym. Podczas gdy tabela

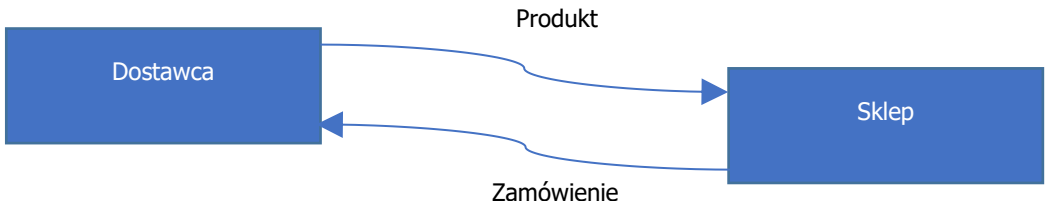
może zawierać tylko jeden klucz główny, aby jednoznacznie identyfikować swoje rekordy, może zawierać wiele kluczy drugorzędnych.

Ogólnie rzecz biorąc, istnieją dwa podejścia do konstruowania RDB: analityczne i syntetyczne.

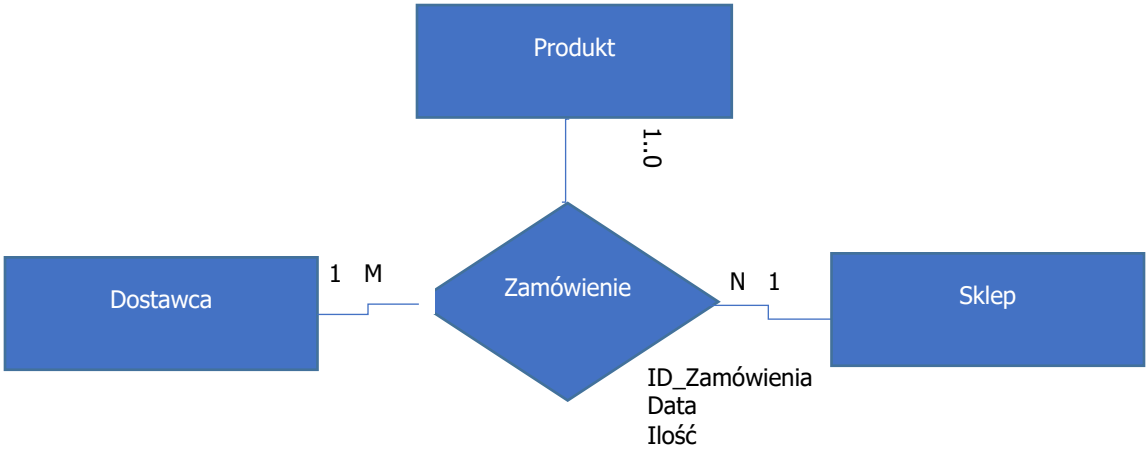
Analityczne podejście do budowy RDB obejmuje następujące cztery kroki:

1. Analiza świata rzeczywistego – model globalny.
2. Określenie encji i relacji – model konceptualny (np. diagram E-R).
3. Określenie modelu logicznego – schematu relacyjnego.
4. Budowa bazy danych (DBMS) – model fizyczny.

Aby zilustrować to podejście, rozważmy przykład sieci sklepów ogólnospożywczych i ich dostawców (rysunek 3.1). Każdy sklep ma wielu dostawców. Każdy dostawca może zaopatrywać różne sklepy. Cykl uzupełniania zapasów rozpoczyna się od zamówienia ze sklepu. W zamian dostawca dostarcza towary do sklepu. Zamówienia to transakcje, w których łączone są dane dotyczące sklepu, dostawcy i dostarczanych towarów (rysunek 3.2).



Rysunek 3.1 Model globalny



Rysunek 3.2 Model koncepcyjny

DOSTAWCA (ID_Dostawcy#, Nazwisko_dostawcy, Kontakt_z_dostawcą)

SKLEP (ID_Sklepu#, Nazwa_sklepu, Adres_sklepu)

ZAMÓWIENIE (ID_Dostawcy #, ID_Sklepu #, ID_Zamówienia#, Data, Ilość, EPC)

PRODUKT (EPC#, Nazwa_produktu, Cena_produktu)

Rysunek 3.3 Model logiczny

Model logiczny reprezentuje tabele RDB poprzez typy ich rekordów. W modelu logicznym (rysunek 3.3.) niektóre atrybuty zawierają symbol hashtagu (#). Oznacza to, że reprezentowane przez nie pole jest lub należy do (złożonego) klucza głównego. Niektóre pola są podkreślone. Są one nazywane kluczami drugorzędnymi lub obcymi, ponieważ odwołują się do kluczy głównych powiązanych tabel.

Syntetyczne podejście do RDB obejmuje następujące trzy kroki:

1. Analiza danych - lista wszystkich istotnych atrybutów.
2. Określenie modelu logicznego poprzez normalizację - schemat relacyjny.
3. Model fizyczny (DBMS).

Normalizacja lub synteza kanoniczna (Kent, 1983) zapewnia, że poprzez inżynierię odwrotną z odpowiednich atrybutów tworzona jest baza danych spełniająca warunki RBD (por. tabela 3.2). Podczas gdy początkowa postać normalna (0PN) atrybutów reprezentuje tabelę nieuporządkowanych atrybutów, kolejne postacie normalne, zgodnie z definicją (Codd, 1970), reprezentują wyższe poziomy organizacji danych. Można twierdzić, że osiągając

trzecią postać normalną, osiągnięto schemat spełniający wymagania dla modelu logicznego RDB.

Tabela jest w Pierwszej postaci normalnej (1PN), jeśli reprezentuje relację. Zapewnia to, że wszystkie powtarzające się grupy danych są przechowywane oddzielnie, a zatem nie powtarzają się.

Tabela w Drugiej postaci normalnej (2PN) jest w 1PN. Dodatkowo, żaden klucz-atrybut nie może być częściowo funkcjonalnie zależny od klucza głównego. W ten sposób wszystkie klucze, które jednoznacznie identyfikują określone atrybuty, są przechowywane oddzielnie. Głównie ma to na celu zapewnienie, że tylko atrybuty zależne od wszystkich (części) klucza głównego są przechowywane w jednej tabeli.

Tabela w Trzeciej postaci normalnej (3PN) jest w 2PN. Dodatkowo, żadne atrybuty niekluczowe nie są przejściowo zależne od klucza głównego. Oznacza to, że wszystkie atrybuty niekluczowe, które mogą reprezentować klucz dla pewnych innych atrybutów niekluczowych, są przechowywane w oddzielnej tabeli, przy czym tylko ich klucz jest przechowywany w oryginalnej tabeli jako klucz obcy.

Postępując zgodnie z krokami normalizacji od 1PN do 3PN, uzyskuje się model logiczny, który odpowiada przepisom relacyjnej bazy danych (RBD). Wyższe formy normalizacji służą głównie do optymalizacji modelu RBD.

Tabela w Boyce-Codd PN (BCPN) jest w 3PN; dodatkowo każdy wyznacznik jest kluczem. Powoduje to usunięcie wszystkich relacji nieobjętych istniejącą kolejnością kluczy z oryginalnej tabeli, tworząc w ten sposób dodatkowe tabele dla każdego klucza kandydującego. Tabela w 4PN jest w 3NF i BCPN. Dodatkowo, każdy atrybut wielowartościowy, który jest częściowo zależny od klucza, znajduje się we własnej tabeli. 4NF ma na celu usunięcie wszystkich możliwych pozostałych atrybutów wielowartościowych z oryginalnej tabeli. Tabela w 5PN jest w 4PN; dodatkowo każda operacja JOIN jest przewidziana przez klucze. Tabela w 6PN jest w 5PN; dodatkowo, wszystkie nietrywialne zależności JOIN są brane pod uwagę.

Można zauważyć, że przy wyższych formach organizacji liczba tabel rośnie z każdym krokiem. Dlatego rozsądne jest obserwowanie fragmentacji danych, aby zapobiec tworzeniu niepotrzebnych, rzadko używanych tabel.

Tabela 3.2 Przykład normalizacji Relacyjnej Bazy Danych

<table>
<tr><td>Początkowa Postać Normalna

ZAMÓWIENIE

• ID_Dostawcy*
• Nazwa_Dostawcy
• Kontakt_Dostawcy
• ID_Sklepu*
• Nazwa_sklepu
• Adres_sklepu
• ID_Zamówienia*
• Data
• EPC
• Ilość
• Nazwa_produktu
• Cena_produktu</td><td>Pierwsza Postać Normalna

ZAMÓWIENIE

• ID_Dostawcy#
• Nazwa_Dostawcy
• Kontakt_Dostawcy
• ID_Sklepu #
• Nazwa_sklepu
• Adres_sklepu
• ID_Zamówienia #
• Data
• EPC
• Ilość
• Nazwa_produktu
• Cena_produktu</td></tr>
<tr><td colspan="2">* Klucze kandydujące powtarzającej się grupy atrybutów, jednoznacznie identyfikujące zamówienie</td></tr>
<tr><td>Druga Postać Normalna

DOSTAWCY

• ID_Dostawcy #
• Nazwa_Dostawcy
• Kontakt_Dostawcy

SKLEP

• ID_Sklepu #
• Nazwa_sklepu
• Adres_sklepu</td><td>

ZAMÓWIENIE

• ID_Zamówienia #
• ID_Dostawcy #
• ID_Sklepu #
• EPC
• Nazwa_produktu
• Cena_produktu
• Data
• Ilość</td></tr>
<tr><td>Trzecia Postać Normalna

ZAMÓWIENIE

• ID_Dostawcy #
• ID_Sklepu #
• ID_Zamówienia#
• Data
• Ilość
• ID_Produktu</td><td>

PRODUKT

• EPC#
• Nazwa_produktu
• Cena_produktu</td></tr>
</table>

Języki zapytań

Są one głównie dwojakiego rodzaju: Structured Query Language (SQL) i Query by Example (QBE). Podczas gdy SQL jest uważany za język programowania, QBE jest używany głównie z DBMS bezpośrednio do zarządzania DB i hurtowniami danych.

Standard SQL (ISO/IEC 9075, 1986-2016) to język programowania czwartej generacji służący do manipulowania bazami danych. Umożliwia wyszukiwanie, dodawanie, modyfikowanie, a także usuwanie rekordów danych. Pomimo standaryzacji istnieją niewielkie różnice w jego implementacji w różnych systemach zarządzania bazami danych (DBMS).

W dalszej części język jest krótko przedstawiony z najczęściej używanymi opcjami. Zgodnie z konwencją słowa kluczowe SQL są pisane wielkimi literami, a każde zdanie kończy się średnikiem. Zdania są prezentowane z linkami do materiałów bibliografii oferujących dalsze informacje.

Każda manipulacja bazą danych zaczyna się od jej utworzenia. Komenda:

CREATE DATABASE [*UTWÓRZ_BAZĘ_DANYCH*] *nazwa_bazy_danych*;

tworzy nową pustą bazę danych o określonej nazwie.

Jak wspomniano powyżej, dane w bazach danych są zorganizowane w tabelach rekordów danych określonego typu, w których wszystkie wiersze mają wspólną strukturę. Aby utworzyć tabelę, używane jest następujące polecenie:

CREATE TABLE [*UTWÓRZ_TABELĘ*] *nazwa_tabeli* (
kolumna1 typ_danych1,
kolumna 2 typ_danych 2,
kolumna 3 typ_danych 3,
….);

Każda nazwana kolumna reprezentuje atrybut o określonym typie danych. Na przykład w:

CREATE TABLE *Sklep* (*ID_Sklepu* int *[liczba całkowita]* NOT NULL PRIMARY KEY *[KLUCZ PODSTAWOWY NIEPUSTY]* (,…);

CREATE TABLE *Zamówienie* (*ID_Zamówienia* int *(liczba całkowita)* NOT NULL PRIMARY KEY *[KLUCZ PODSTAWOWY NIEPUSTY]*, …, *Id_Produktu* int *[liczba całkowita]* FOREIGN KEY REFERENCES *Produkt* (*EPC*) *[KLUCZ OBCY ODWOŁANIA Produkt(EPC)]*);

tworzone są dwie tabele. Pierwsza z nich zawiera dane sklepów, podczas gdy druga zawiera dane ich zamówień, odwołując się do pierwszej tabeli poprzez numer produktu jako klucz obcy.

Podczas gdy dane w tabeli są już posortowane według klucza głównego, można je dodatkowo posortować według innych atrybutów, pod warunkiem, że są one indeksowane.

Można je indeksować, tworząc indeks na podanym atrybucie (atrybutach) za pomocą następującego polecenia:

CREATE INDEX *nazwa_indeksu* ON *nazwa_tabeli* (*nazwa_kolumny*);

Każda manipulacja danymi na indeksowanej tabeli trwa nieco dłużej, ponieważ dla jej spójności należy sprawdzić nie tylko dane dostarczane przez klucze i odpowiednio je uporządkować, ale także inne atrybuty z określonego indeksu.

Najczęstszą operacją na bazie danych jest zapytanie o dane za pomocą instrukcji SELECT:

SELECT *kolumna1, kolumna2*, …
FROM *nazwa_tabeli*;

To zapytanie o dane zwraca dane w kolumnie1, kolumnie2 itd. z tabeli. Zdania zapytania są zwykle tworzone poprzez podanie dodatkowych opcji, odfiltrowanie danych, spełnienie określonego warunku (warunków):

WHERE definiuje warunek, który określa kryteria wyboru rekordów;

GROUP BY łączy rekordy, mające wspólną właściwość umożliwiającą funkcje agregacji;

HAVING określa funkcje agregujące dla grup zdefiniowanych przez instrukcję GROUP BY;

ORDER BY określa atrybuty, według których uporządkowane są zwracane rekordy.

Na przykład:

```
SELECT "Sklep"."Nazwa_Sklepu", "Produkt"."Nazwa_produktu", "Zamówienie"."Ilość" FROM "Zamówienie", "Produkt", "Dostawca", "Sklep" WHERE "Zamówienie"."ID_Produktu" = "Produkt"."EPC" AND "Zamówienie"."ID_Dostawcy" = "Dostawca"."ID_Dostawcy" AND "Zamówienie"."ID_Sklepu" = "Sklep"."ID_Sklepu" ORDER BY "Sklep"."Nazwa_Sklepu" ASC
```

zwraca listę sklepów z zamówionymi produktami i ilościami, uporządkowaną rosnąco według nazwy sklepu.

Najważniejszą operacją w procesie selekcji jest operacja JOIN. Często zastępuje ona warunek WHERE jako JOIN ON, po którym następuje warunek. Porównuje wartości kolumn i na podstawie porównania określa, czy powinny one zostać uwzględnione w wyniku. W LEFT JOIN rekord jest zwracany, jeśli kryteria są spełnione w lewej tabeli i odwrotnie w operacji RIGHT JOIN. Jak wspomniano powyżej, warunek musi być spełniony w obu tabelach, aby był zgodny z operacją INNER JOIN lub FULL JOIN. Ponieważ ta ostatnia jest najczęściej

używana, można użyć JOIN jako synonimu. Odnosząc się do warunków piątej i szóstej postaci normalnej, jest to ta sama operacja JOIN, która musi zostać spełniona, aby spełnić warunki odpowiedniej postaci normalnej.

Do wprowadzania nowych danych do tabeli używana jest operacja INSERT INTO:

INSERT INTO *[WPROWADŹ DO] nazwa_tabeli* (*kolumna1*, [*kolumna2*, ...])
VALUES *[WARTOŚCI]* (*wartość1*, [*wartość2*, ...]);

Aby operacja zakończyła się powodzeniem, wartości muszą spełniać wszystkie warunki atrybutów określonych przez nazwy kolumn. Nie trzeba określać nazw kolumn, jeśli wszystkie wartości są wymienione. W przypadku, gdy w tabeli przewidziane są pewne wartości DEFAULT, nie trzeba ich określać, chyba że są inne.

Po wprowadzeniu danych można je zmodyfikować za pomocą instrukcji UPDATE:

UPDATE *nazwa_tabeli*
SET *kolumna1=wartość1, kolumna2=wartość2*,...
WHERE *pewna_kolumna=jakaś_wartość*;

W instrukcji podawane są nowe wartości dla pól w wymienionych kolumnach. Kryteria wyboru wiersza są oznaczone specyfikatorem WHERE, który określa wszystkie wartości kolumn, do których odnosi się instrukcja UPDATE. Aby zapobiec niepożądanym zmianom, należy zachować szczególną ostrożność przy formułowaniu kryteriów wyboru.

Rekord lub wiele rekordów można usunąć z tabeli za pomocą operacji DELETE:

DELETE FROM *[USUŃ Z] nazwa_tabeli*
WHERE *pewna_kolumna=jakaś_wartość*;

Podobnie jak w przypadku instrukcji UPDATE, specyfikator WHERE jest używany do określenia wszystkich wierszy, które powinny zostać usunięte.

Oczywiście zarządzanie bazą danych na tym się nie kończy. Każdy element bazy danych można również usunąć, zmienić i/lub zastąpić nowym. W przypadku, gdy indeks, tabela lub baza danych ma zostać usunięta, można zastosować następujące instrukcje:

DROP INDEX *nazwa_indeksu* ON *nazwa_tabeli*;

DROP TABLE *nazwa_tabeli*;

DROP DATABASE *nazwa_bazy_danech*;

Jeśli chcemy tylko usunąć dane z tabeli, można użyć instrukcji TRUNCATE:

TRUNCATE TABLE *nazwa_tabeli*;

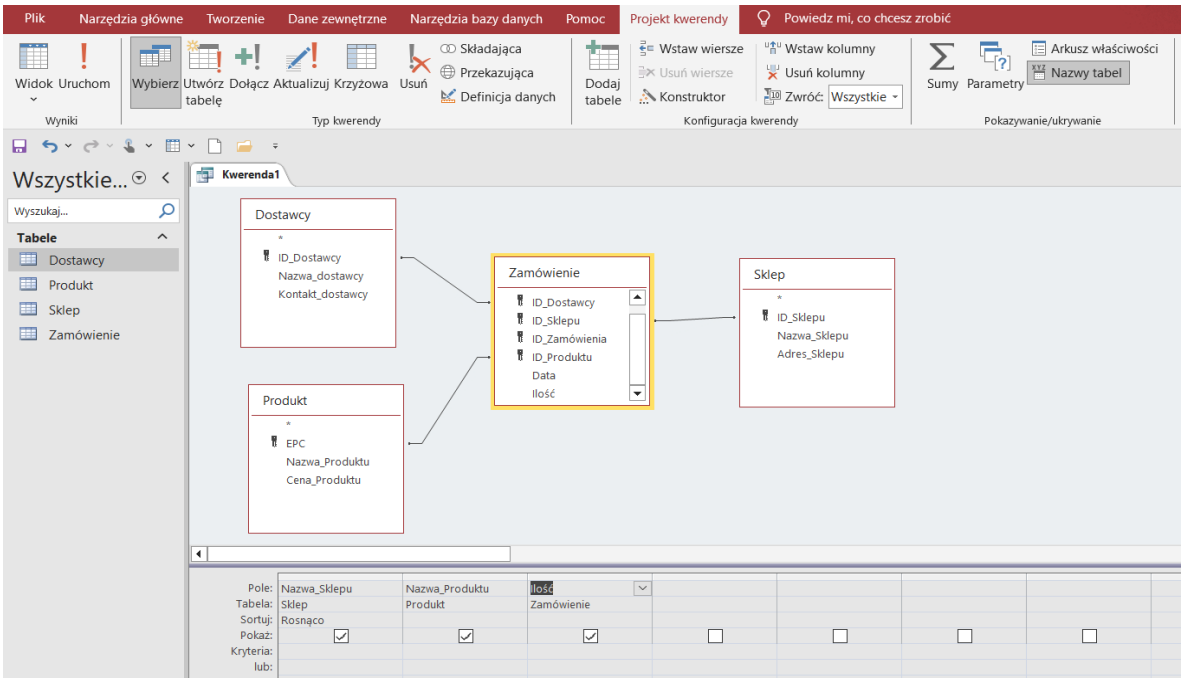
W przypadku, gdy chcemy dodać lub usunąć atrybut (kolumnę) do/z tabeli, możemy to zrobić za pomocą instrukcji ALTER:

ALTER TABLE *nazwa_tabeli* ADD *nazwa_kolumny typ_danych*;

ALTER TABLE *nazwa_tabeli* DROP COLUMN *nazwa_kolumny*;

Na tym kończy się ten krótki przegląd języka SQL i jego najczęstszych scenariuszy użycia. SQL jest powszechnie używany w architekturach klient-serwer z systemem DBMS hostowanym przez serwer. Aby uzyskać do niego dostęp, instrukcje SQL są wydawane przez program aplikacji klienckiej lub interfejs sieciowy DBMS serwera.

Z drugiej strony, QBE jest również często używany z relacyjnymi systemami DBMS z graficznym interfejsem użytkownika (GUI), takimi jak MS Access lub LibreOffice Base. Dzięki QBE baza danych i jej tabele są tworzone znacznie bardziej interaktywnie, a ich struktura jest łatwiejsza w utrzymaniu. Jak sama nazwa wskazuje, oferuje również prostszą formę wprowadzania i wyszukiwania danych. Aby przeprowadzić wyszukiwanie, należy zebrać wszystkie tabele, które są używane w wyszukiwaniu, a następnie ustalić warunki jako wzorce w polach kolumn, aby odfiltrować odpowiednie dane (rysunek 3.4). Sformułowanie zapytania jest uzupełniane przez istniejące relacje między tabelami. Jak zwykle, wynikiem takiego zapytania jest kolejna tabela z wynikowymi danymi, które mogą być później dalej przetwarzane. W ten sposób można również tworzyć zapytania kaskadowe lub wielofazowe.



Rysunek 3.4 Zapytanie według przykładu odpowiadające powyższemu zapytaniu SQL

Filtry i maski danych

Aby zapobiec błędnym lub niekompletnym danym, które mogłyby zakłócić ich przetwarzanie i interpretację, należy zastosować dodatkowe środki ostrożności:

1. Filtrowanie pustych wierszy i kolumn.
2. Stosowanie silnego typowania danych w celu zapobiegania błędom obliczeniowym.
3. Definiowanie masek wejściowych w celu zapobiegania wprowadzaniu nieprawidłowych danych.
4. Zniekształcanie danych w celu zapobiegania błędnym wynikom.

Puste wiersze i kolumny są częstym źródłem błędów, które wynikają głównie ze słabych interfejsów w aplikacjach do gromadzenia danych. Anulowane lub niekompletne transakcje zwykle skutkują pustymi wierszami lub brakującymi danymi w dziennikach transakcji. Tylko częściowo można sobie z nimi poradzić w aplikacjach arkuszy kalkulacyjnych, w których puste wiersze i brakujące dane można wykryć, sprawdzając całkowitą liczbę w stosunku do liczby niezerowych danych w wierszach/kolumnach. Można je odfiltrować, usuwając puste wiersze/kolumny, jednak nie zawsze jest to pożądane działanie, ponieważ możemy również stracić cenne dane. Najlepszym sposobem, aby temu zapobiec, jest użycie systemu DBMS,

który z jednej strony pozwalałby na wprowadzanie tylko pełnych danych transakcyjnych, z drugiej zaś zapobiegałby pustym wierszom/kolumnom, ponieważ w bazach danych ich nie ma.

Niewłaściwe wpisywanie danych jest kolejnym częstym źródłem błędów. Jeśli dane alfanumeryczne, takie jak data lub kwota waluty, zostaną wprowadzone zamiast danych numerycznych, spowoduje to błędy podczas przetwarzania tych danych. Zarówno w arkuszach kalkulacyjnych, jak i w bazach danych poszczególnym komórkom danych, reprezentującym wartości atrybutów w rekordach danych, można przypisać typy danych, które zaalarmują nas, gdy wprowadzimy dane w niewłaściwym formacie. W ten sposób można zapobiec sytuacji, w której błędne dane przeszkadzają w ich przetwarzaniu.

Podczas konstruowania baz danych można zastosować ograniczenia do pól danych reprezentujących atrybuty encji lub relacji. Oprócz przypisania im odpowiedniego typu, można zdefiniować maski wejściowe pozwalające na wprowadzanie danych, takich jak daty, waluty, kody EAN itp. tylko w określonym formacie. Zwykle rozwiązuje to wiele nieporozumień, które w przeciwnym razie mogłyby wystąpić podczas przetwarzania danych.

Innym częstym źródłem błędów są nieobiektywne dane, reprezentujące dane, które są o rzędy wielkości większe lub mniejsze niż oczekiwano. Ponownie, mogą one zakłócać nasze przetwarzanie, dając błędne wyniki. Są one trudniejsze do wykrycia i można je odfiltrować jedynie poprzez analizę danych. W aplikacjach arkusza kalkulacyjnego dobrą powszechną praktyką byłoby określenie minimalnych, maksymalnych i średnich wartości danych w odpowiednich kolumnach w celu wykrycia możliwych odchyleń. Jeśli zostaną wykryte, można je zaznaczyć i zająć się nimi ręcznie, jeśli jest ich kilka, lub odfiltrować i zmodyfikować za pomocą zapytania w tabeli bazy danych, jeśli jest ich wiele. Tak czy inaczej, należy je dokładnie ocenić, aby nie pogorszyć sytuacji, w którym to przypadku lepiej byłoby usunąć te dane.

Hurtownie danych i bazy wiedzy

Dane w hurtowniach danych są pobierane z RDB i katalogowane w porządku chronologicznym. Zwykle na danych przeprowadzane są analizy biznesowe, a wyniki są przechowywane w celu późniejszego wykorzystania przez dział analityczny. Po zapisaniu w hurtowni dane te zazwyczaj nie są zmieniane, aby zachować ich spójność.

Oprócz porządku chronologicznego, podczas tworzenia baz wiedzy brane są pod uwagę zamówienia kontekstowe. W tym przypadku jednostki są reprezentowane jako pochodne

jednostki najwyższego poziomu lub jej jednostki zależnej. Ich relacje są ustalane bardziej swobodnie, ponieważ mają być aktualizowane i ulepszane w miarę ich używania. Są one ustanawiane w formie reguł opartych na właściwościach encji. W związku z tym forma, w jakiej są przechowywane, jest nieco inna. Często są one przechowywane w formie ontologii zawierających głębszą wiedzę na temat zebranych danych. Podobnie jak w przypadku przechowywania wyników zapytań w hurtowniach danych, zapytania w bazach wiedzy są również przechowywane do późniejszego wykorzystania w celu renderowania bieżących wyników, ponieważ jednostki, relacje i instancje danych ulegają zmianie.

W przeciwieństwie do baz danych i hurtowni, które są specyficzne dla aplikacji, bazy wiedzy mogą być niezależne od aplikacji i często są używane w różnych dziedzinach przez różne aplikacje. Przykład został przedstawiony w (Gumzej i in., 2023).

3.4 Wnioski

W tym rozdziale omówiono różne aspekty zarządzania danymi w logistyce. Oprócz reprezentacji danych i standardów ich przechowywania, przedstawiono mechanizmy organizacji i wyszukiwania danych. Na koniec omówiono niektóre typowe błędy w zautomatyzowanym przetwarzaniu danych, aby zachować czujność świadomego czytelnika. Oprócz wymienionych przykładów, więcej można znaleźć w powiązanych materiałach edukacyjnych.

BIBLIOGRAFIA DO ROZDZIAŁU 3.

- Codd E.F. (1970). A relational model of data for large shared data banks. Communications. ACM 13, 6, pp. 377–387.
- Gumzej, R., Kramberger, T., Dujak, D. (2023). A knowledge base for strategic logistics planning, Proceedings of the 23rd International Scientific Conference Business Logistics in Modern Management: October 5-6, 2023, Osijek, Croatia, Dujak, Davor (ed.) Osijek: Josip Juraj Strossmayer University of Osijek, Faculty of Economics and Business, pp. 317-330. [available at: https://blmm-conference.com/past-issues/, Dostęp November 3rd, 2023]
- GS1 (2023). GS1 Standards. [available at: https://www.gs1.org/standards, Dostęp October 27th, 2023]

- Kent, W. (1983). A Simple Guide to Five Normal Forms in Relational Database Theory, Communications of the ACM, vol. 26, pp. 120-125.
- W3schools (2023). XML Tutorial [Dostępne: https://www.w3schools.com/xml/, access November 3rd, 2023]
- W3schools (2023). JSON - Introduction Dostępne: https://www.w3schools.com/js/js_json_intro.asp, access November 3rd, 2023]
- W3schools (2023). Database Normalization [Dostępne: https://www.w3schools.in/DBMS/database-normalization/, access December 7th, 2023]

4. Modelowanie symulacyjne i analiza

Poprzez modelowanie i analizę symulacyjną (SMA) dąży się do spełnienia wymagań twierdzenia Conanta-Ashby'ego (Conant i Ashby, 1970), definiując model systemu jako dobry regulator, posiadający tyle samo uchwytów, części i stanów, co jego oryginalny fizyczny odpowiednik; zapewniając w ten sposób możliwość zbudowania jego cyfrowego modelu i stworzenia cyfrowego laboratorium, które umożliwi jego eksplorację, adaptację i optymalizację. Wynikowe modele symulacyjne są abstrakcyjne, dynamiczne i w większości przypadków stochastyczne, ponieważ ich zmienne systemowe są modelowane przez rozkłady prawdopodobieństwa.

4.1 Symulacja w logistyce

W logistyce SMA może dostarczyć cennych danych wejściowych do optymalizacji łańcucha dostaw (SC) i sieci ruchu (TN). Modelowanie symulacyjne może być wykorzystywane do graficznej wizualizacji czasowych przepływów przez złożone procesy i zasoby SC i TN, umożliwiając przewidywanie i kwantyfikację możliwych wyników różnych scenariuszy. Pomaga to podmiotom SC i TN uzyskać cenny wgląd i zrozumieć wpływ ich potencjalnych decyzji na wydajność SC i TN, w tym czasy realizacji SC, czasy podróży TN i koszty. W związku z tym SMA w modelowaniu SC i TN może przyczynić się do analizy SC i TN oraz do ulepszenia ich projektów w celu osiągnięcia wyższej wydajności i zrównoważonego rozwoju.

Istnieje wiele aspektów SC, reprezentujących różne perspektywy zarządzania SC. Spojrzenie kierownika produkcji na SC różni się od spojrzenia kierownika ds. marketingu, które z kolei różni się od spojrzenia kierownika ds. zaopatrzenia itd. Dlatego też stosowane modele są różne, nawet dla tej samej firmy, nie mówiąc już o całym SC.

Podczas rozwiązywania problemów SMA w logistyce, menedżerowie muszą podejmować decyzje na poziomie strategicznym, taktycznym i operacyjnym, w zależności od ich wpływu

na SC lub TN jako całość. Ze względu na ich współzależności, menedżerowie często nie są w stanie rozwiązywać problemów na pojedynczym poziomie. Jednocześnie trudno jest obserwować wszystkie trzy poziomy z perspektywy pojedynczego podmiotu. Z perspektywy SMA można obserwować SC lub TN na dwóch poziomach:

1. Poziom makro:

- samoorganizacja,
- współewolucja podmiotów,
- zależność od połączeń/tras transportowych.

2. Poziom mikro:

- liczne i niejednorodne podmioty,
- lokalne interakcje między podmiotami,
- podmioty strukturyzowane,
- podmioty adaptacyjne.

Chociaż są one wykonywane w czasie rzeczywistym, czasowy aspekt operacji SC jest nieco niejednoznaczny. W zależności od poziomu i perspektywy, czas trwania operacji może być mierzony w dniach, tygodniach lub nawet miesiącach w przypadku działań międzyorganizacyjnych, podczas gdy operacje wewnątrzorganizacyjne są mierzone w godzinach lub nawet sekundach. W zależności od charakteru modelowanego problemu, czas trwania najkrótszej operacji lub maksymalna częstotliwość przychodzących/wychodzących żądań determinuje nie tylko reprezentację czasu w modelu SMA, ale także jego ziarnistość. Im krótszy jest minimalny czas trwania najkrótszej operacji lub im wyższa jest najwyższa częstotliwość żądań, tym drobniejsza jest ziarnistość czasu lub innymi słowy precyzja utrzymywania czasu w modelu. Jest to ważne dla modelującego, ponieważ czas reakcji modelu nie może być krótszy niż wstępnie zdefiniowana ziarnistość czasu. W związku z tym należy z wyprzedzeniem oszacować czas trwania wszystkich operacji i czasy między kolejnymi sygnałami przychodzącymi/wychodzącymi, aby móc poprawnie określić jednostki czasu modelu systemu.

W modelu symulacyjnym czas może płynąć od transakcji do transakcji poprzez zdarzenia krytyczne lub w sposób ciągły. W tym drugim przypadku progresja czasu w modelu jest niezależna od częstotliwości operacji. W przypadku przepływu czasu wyzwalanego zdarzeniami krytycznymi, operacje są wywoływane zgodnie z czasem ich wystąpienia, czyli zdarzeniami krytycznymi. Zaletą SMA jest to, że podczas symulacji można przyspieszyć

postęp czasu w modelu, dzięki czemu procesy działają szybciej niż w czasie rzeczywistym. W ten sposób można wcześnie przewidzieć następujące zdarzenia.

4.2 Symulacja zdarzeń dyskretnych

Analiza DES oferuje kierownikowi produkcji najbardziej szczegółowy wgląd w proces logistyczny (produkcyjny) dzięki spójnemu i konsekwentnemu modelowi. Dlatego DES jest wysoko cenionym narzędziem do określania zachowania w czasie rzeczywistym i wykorzystania zasobów w przemyśle przetwórczym, w tym w logistyce.

Konstrukcja:

- Jednostki przepływu reprezentują jednostki symulacji (np. zamówienia, materiały itp.), które wchodzą do systemu na wejściu (wejściach) i przechodzą przez model systemu.
- Procesory reprezentują mobilne (np. ludzie, wózki widłowe itp.) i stałe (np. maszyny, linie produkcyjne itp.) zasoby, które przetwarzają jednostki symulacji.
- Kolejki przechowują jednostki przepływu do momentu ich przejścia do następnego dostępnego procesora.
- Łączniki definiują promocję jednostek poprzez model systemu.

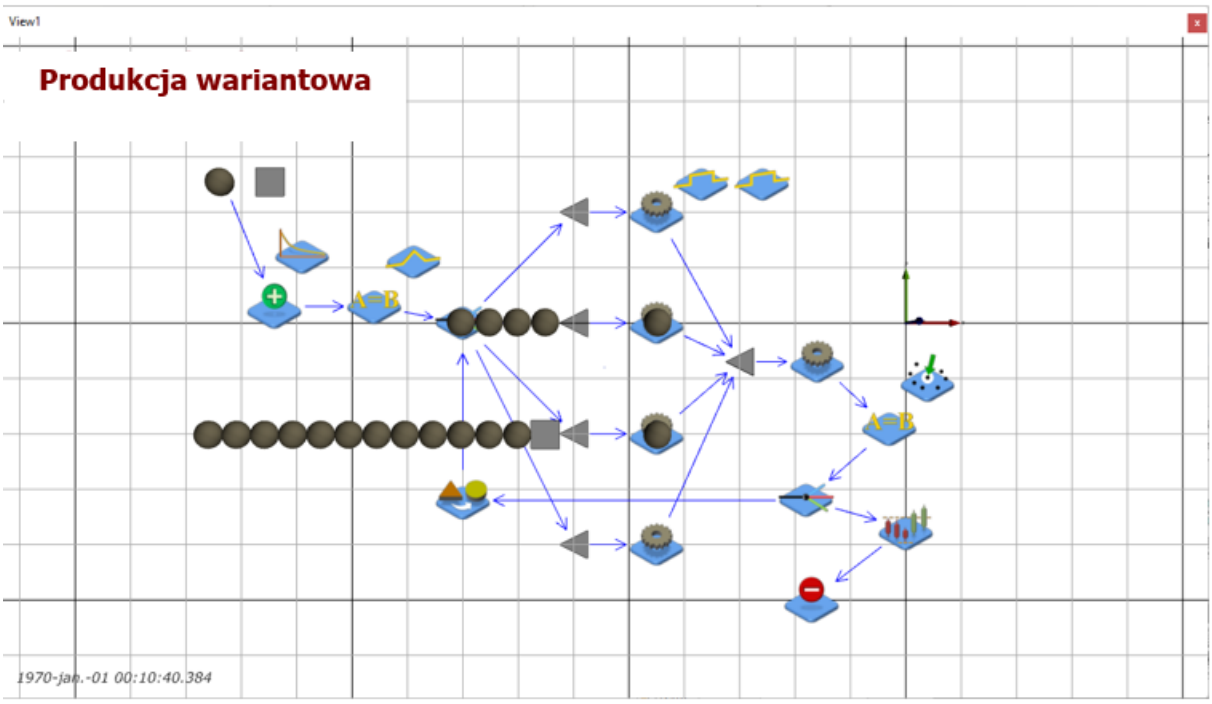
Właściwości:

- zorientowanie na proces,
- koncentruje się na szczegółowym modelowaniu procesów,
- niejednorodne podmioty,
- mikropodmioty są obiektami pasywnymi,
- zdarzenia wprowadzają dynamikę do systemu,
- dyskretna progresja czasowa; od jednego (czasowego) zdarzenia do następnego,
- elastyczność osiąga się poprzez zmianę struktury modelu; struktura systemu podczas symulacji jest stała.

Przykład

Przykład DES (rysunek 4.1, zaczerpnięty ze środowiska symulacyjnego JaamSim (JaamSim Development Team, 2023)) obejmuje model produkcji wariantowej, w którym wytwarzane są

cztery różne produkty (Gumzej i Rakovska, 2020). Zgodnie z planem produkcji produkowanych jest odpowiednio około 10, 30, 40 i 20% typów produktów 1, 2, 3 i 4. Wybór typu produktu jest indukowany przez trójkątny rozkład między 1 a 4 z modulo na 3. Każdy typ produktu ma dedykowaną linię produkcyjną. Zlecenia produkcyjne są realizowane zgodnie z rozkładem wykładniczym wokół średniej wartości czasu 30 s. Produkcja każdego pojedynczego produktu trwa 100-120 s zgodnie z rozkładem jednostajnym dyskretnym. Po zakończeniu produkcji produkty są sprawdzane pod kątem jakości w dedykowanym miejscu testowym. Kontrola jakości trwa 10 s. Z doświadczenia firmy wynika, że średnio 1 na 10 produktów nie przechodzi kontroli. Produkty o niewystarczającej jakości są transportowane z powrotem na oryginalną linię produkcyjną. Ich ponowne przetworzenie zajmuje 120-130 s, zgodnie z jednostajnym dyskretnym rozkładem. Czas trwania produkcji, kontroli jakości i ponownego przetwarzania nie zależy od rodzaju produktu. Po pomyślnym przejściu kontroli jakości gotowe produkty są transportowane z zakładu produkcyjnego do magazynu produktów gotowych. Ponowne wytwarzanie wadliwych produktów jeszcze w trakcie produkcji jest skutecznym sposobem na zmniejszenie zarówno wpływu na środowisko, jak i kosztów produkcji.



Rysunek 4.1 Produkcja wariantowa z kontrolą jakości

Podsumowanie

Następujące parametry procesu mogą być analizowane i optymalizowane przez DES:

- czas cyklu produkcji i wydajność,
- wykorzystanie komórek i przestrzeni produkcyjnych,
- pojemność przestrzeni magazynowej oraz czas przebywania w niej jednostki magazynowej,
- wykorzystanie zasobów mobilnych (np. operatorów, przenośników, wózków widłowych).

Analiza DES oferuje kierownikowi produkcji najbardziej szczegółowy wgląd w proces logistyczny (produkcyjny) dzięki spójnemu i konsekwentnemu modelowi. Jego czas jest zgodny ze światem rzeczywistym. Dlatego DES jest wysoko cenionym narzędziem do określania zachowania w czasie rzeczywistym i wykorzystania zasobów w przemyśle procesowym, w tym w logistyce.

4.3 Dynamika systemu

Analiza SD reprezentuje spojrzenie menedżera SC na proces produkcyjny za pomocą spójnego i konsekwentnego modelu. SD jest uważana za narzędzie najlepiej nadające się do określenia struktury, a także optymalnych ilości (kiedy i ile poszczególnych wejść, zapasów i wyjść). Dlatego też pozwala na efektywne wykorzystanie obiektów produkcyjnych i magazynowych.

Konstrukcja:

- Zapasy reprezentują bufory, które mogą przechowywać elementy dostawy w łańcuchu dostaw.
- Przepływy reprezentują kanały dostaw.
- Pętle sprzężenia zwrotnego reprezentują parametry dostrajania uzupełniania zapasów.

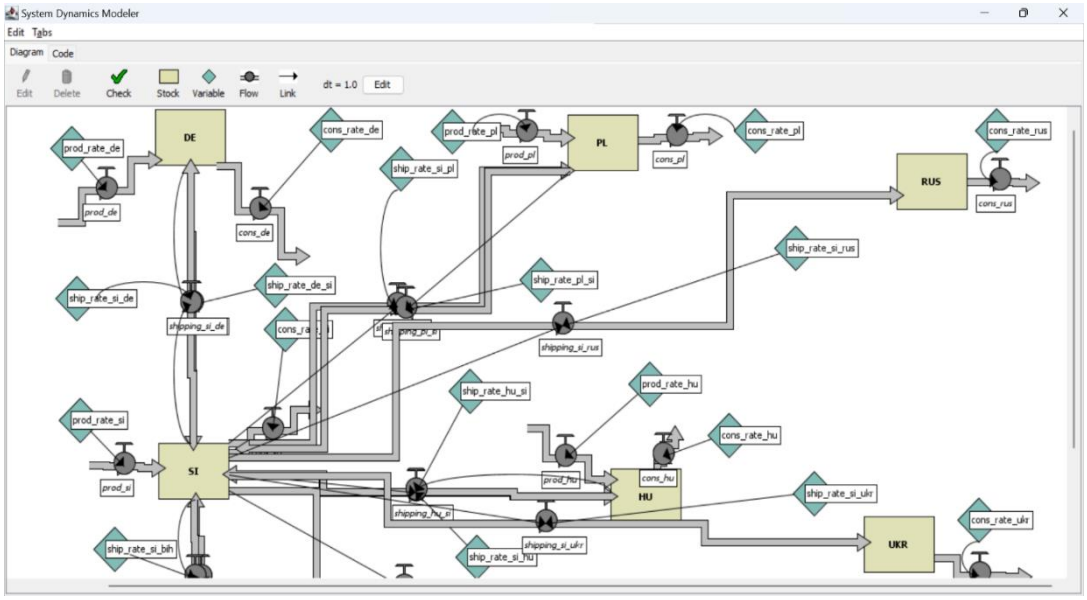
Właściwości:

- zorientowanie na system,
- zorientowane na kluczowe wskaźniki wydajności modelowanie zmiennych systemowych,

- jednostki jednorodne,
- podmioty na poziomie mikro są pomijane,
- dynamika jest wprowadzana przez sprzężenie zwrotne,
- ciągła progresja czasowa; czas postępuje synchronicznie dla wszystkich komponentów modelu systemu,
- elastyczność jest osiągana poprzez zmianę struktury modelu
- struktura systemu podczas symulacji jest stała.

Przykład

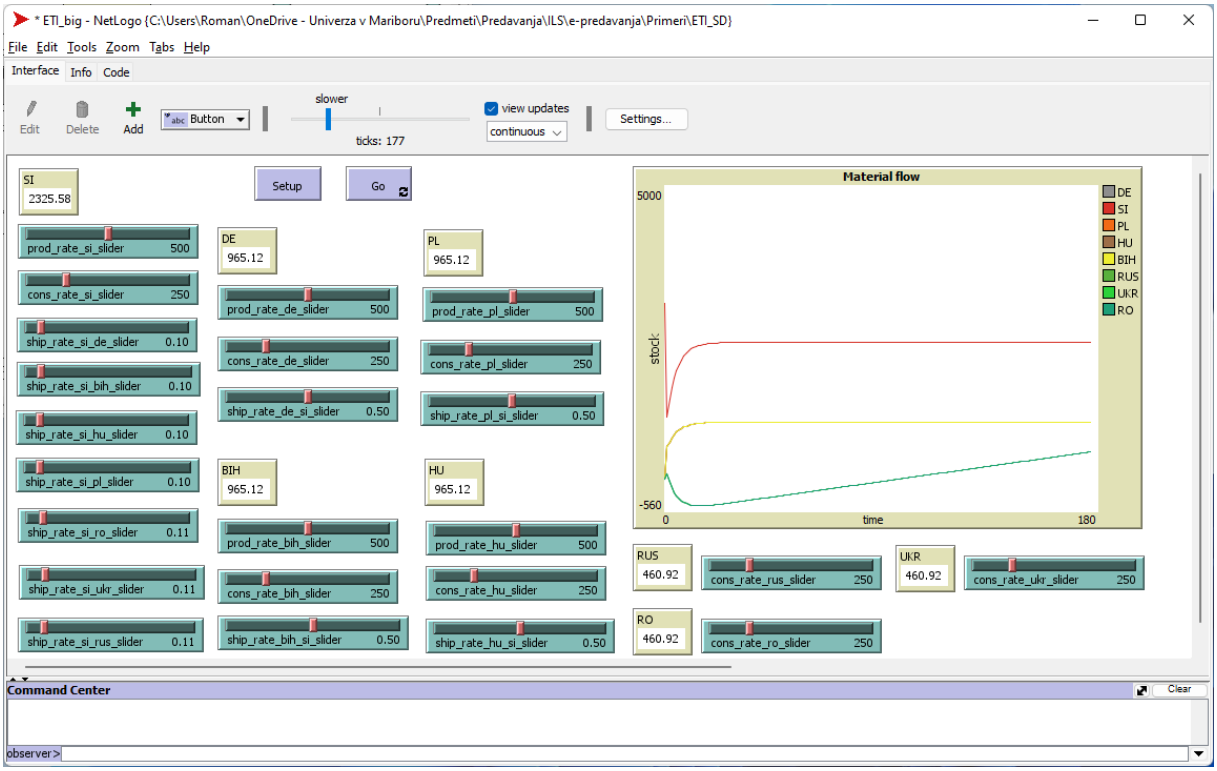
Przykład SD (rysunek 4.2, wyodrębniony ze środowiska symulacyjnego NetLogo (Wilensky, 1999) obejmuje SC firmy produkującej sprzęt AGD i opisuje przepływy materiałów między jej spółkami zależnymi (Gumzej i Rakovska, 2020). Firma posiada wiele zakładów produkcyjnych: główny zakład w Słowenii (SI), a także firmy partnerskie w Niemczech (DE), Polsce (PL), na Węgrzech (H) i w Bośni i Hercegowinie (BIH). Oprócz zakładów produkcyjnych, zakłady sprzedaży brutto znajdują się w Rosji (RUS), na Ukrainie (UKR) i w Rumunii (RU). Zakłady produkcyjne zaopatrują własne rynki w gotowe produkty, a siebie nawzajem w komponenty produktów.



Rysunek 4.2 Układ SC

Powiązany pulpit nawigacyjny NetLogo (rysunek 4.3) służy jako narzędzie wspomagające podejmowanie decyzji (DST) w celu powiązania wielkości produkcji i zapasów z predyspozycjami i ich fizyczną dystrybucją. Przepływ czasu jest ciągły w transakcjach

każdego dnia, tj. każdego dnia pewna liczba komponentów jest wysyłana między zakładami produkcyjnymi, a pewna liczba gotowych produktów jest zużywana na miejscu lub wysyłana do miejsc dystrybucji. W oparciu o początkowy zapas 300 jednostek w lokalizacji SI i 0 zapasów w innych lokalizacjach oraz model dystrybucji, ilości zapasów w poszczególnych lokalizacjach reprezentują średnie zapasy zgodnie z danymi wskaźnikami produkcji (szt.), zużycia (%) i wysyłki (%).



Rysunek 4.3 Pulpit nawigacyjny S.C

Podsumowanie

Symulacja dynamiki systemów pozwala na:

- planowanie układu S.C.,
- optymalizację zdolności produkcyjnych i dystrybucyjnych,
- oszacowanie obciążeń kanałów dystrybucji i powiązanych kosztów.

4.4 Symulacja agentowa

Analiza ABS oferuje strategiczne spojrzenie menedżera lub regulatora rynku na rynek. W związku z tym ABS jest uważany za narzędzie najlepiej nadające się do określenia

optymalnej struktury i układu/asortymentu danego rynku i/lub SC poprzez uwzględnienie ich globalnych cech (np. demografii, klimatu, PKB, jakości, świadomości itp.).

Konstrukcja:

Agenci reprezentujący węzły łańcucha dostaw (np. dostawców, sprzedawców detalicznych i inspektorów) wraz z ich właściwościami, relacjami i zachowaniem.

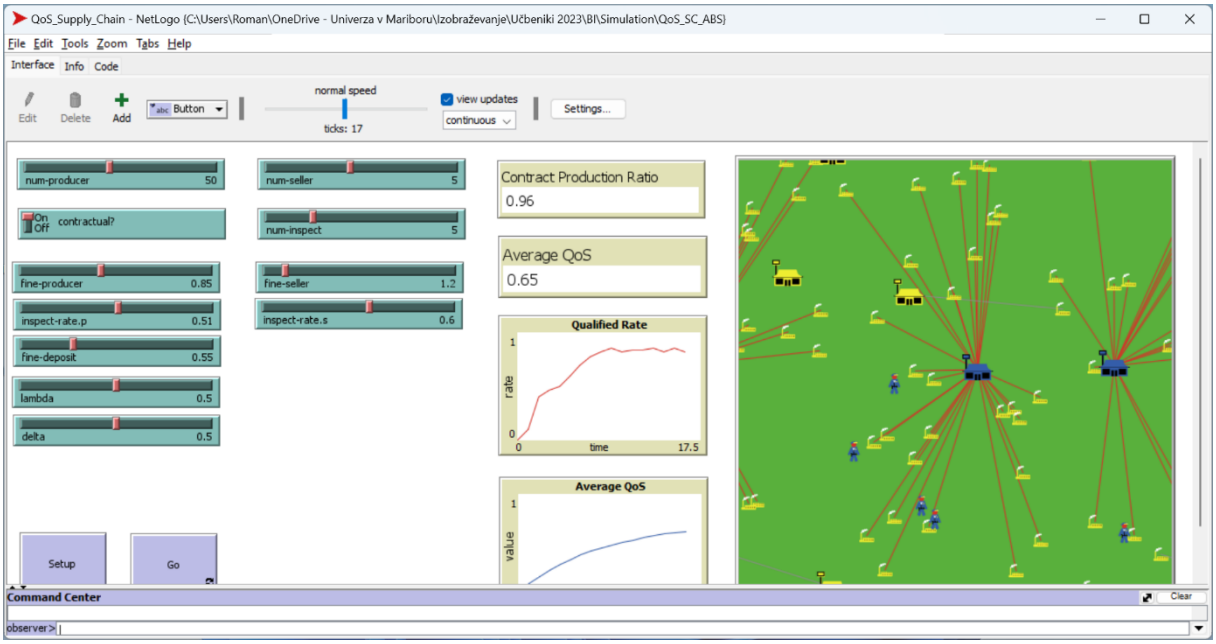
Właściwości:

- skoncentrowanie na jednostce,
- zorientowanie na problem modelowanie podmiotów i ich interakcji,
- niejednorodność podmiotów,
- mikropodmioty są aktywnymi obiektami, które działają w swoich środowiskach, komunikują się między sobą i autonomicznie podejmują decyzje,

- decyzje i interakcje między agentami wprowadzają dynamikę do systemów,
- agenci i ich środowiska stanowią modele formalne,
- przepływ czasu jest dyskretny i uniwersalny na poziomie modelu; czas modelu jest zgodny z częstotliwością transakcji SC i cyklami życia węzłów SC,
- elastyczność modelu jest osiągana dzięki zmieniającej się strukturze systemu i zachowaniu agentów,
- struktura systemu podczas symulacji jest zmienna.

Przykład

Przykład ABS (rysunek 4.4, wyodrębniony ze środowiska symulacyjnego NetLogo (Wilensky, 1999) został wykorzystany do analizy zachowania szczebli SC na otwartym rynku (Gumzej i Rakovska, 2020), w odniesieniu do ich jakości usług (QoS). W tym przykładzie różne polityki dotyczące całkowitego zarządzania jakością firmy zostały zbadane przez model obejmujący jej dostawców, klientów i regulatorów rynku.



Rysunek 4.4 Regulacje dotyczące rynku

Podsumowanie

Symulacja agentowa pozwala na:

- Planowanie układu S.C.,
- Modelowanie dynamicznego wzrostu S.C.,
- Modelowanie zachowań partnerów w ramach S.C,
- Optymalizację wskaźników globalnych.

4.5 Symulacja sieciowa

Analiza NS oferuje widok sieci z punktu widzenia regulatora sieci. W związku z tym NS jest uważana za narzędzie najlepiej nadające się do określenia optymalnej struktury, układu i asortymentu sieci, biorąc pod uwagę jej globalne cechy (np. przepustowość, emisje, wskaźniki QoS itp.)

Konstrukcja:

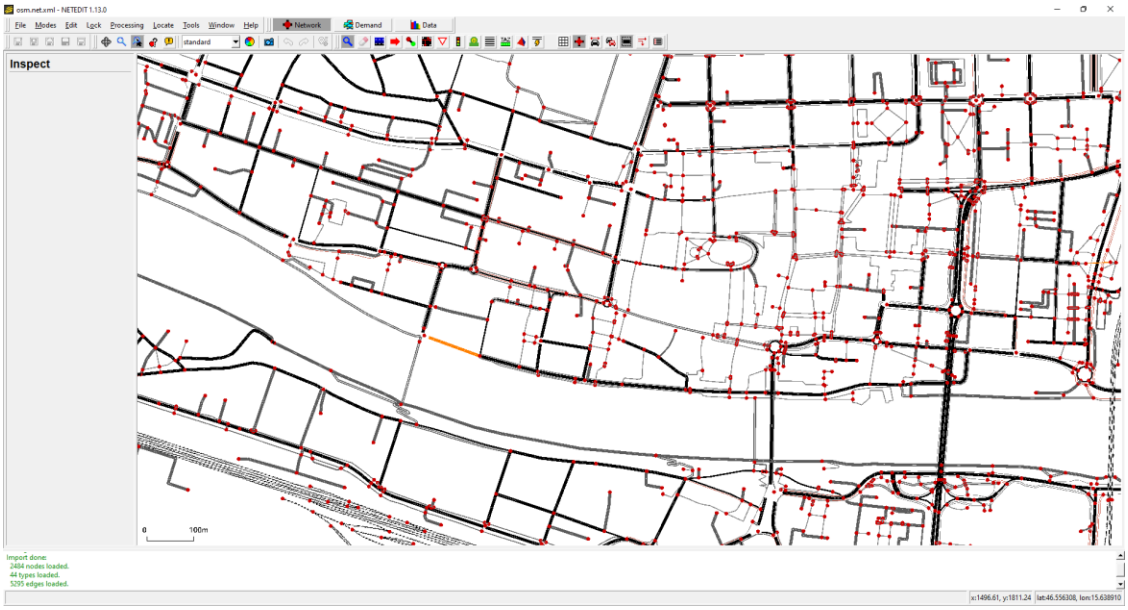
- Agenci reprezentujący obiekty przepływu wraz z ich właściwościami, relacjami i zachowaniem.
- Sieć reprezentująca sieć nakładkową (np. sieć ruchu), w której obiekty przepływu dojeżdżają.

Właściwości:

- zorientowanie na system,
- zorientowanie na problem modelowanie podmiotów i ich interakcji,
- niejednorodność podmiotów,
- mikropodmioty są aktywnymi obiektami, które działają w swoich środowiskach, komunikują się między sobą i autonomicznie podejmują decyzje,
- decyzje i interakcje między agentami wprowadzają dynamikę do systemów,
- agenci i ich środowiska stanowią modele formalne,
- przepływ czasu jest dyskretny i uniwersalny na poziomie modelu; czas jest zgodny ze względnymi prędkościami obiektów przepływu,
- elastyczność modelu jest osiągana poprzez zmianę struktury sieci, która jest stała podczas symulacji, oraz zachowania agentów, które różnią się w zależności od stanu sieci (ruchu) i ich celów.

Przykład

Przedstawiony przykład (rysunek 4.5, wyodrębniony ze środowiska symulacyjnego SUMO (Pablo i in., 2018)) został wykorzystany do określenia przepływów ruchu i przepustowości ulic w centrum miasta dotkniętych planowaną blokadą dróg (Šinko i Gumzej, 2021). Ponadto zmierzono wskaźniki związane z ruchem, takie jak czas podróży, zużycie paliwa i emisje spalin.



Rysunek 4.5 Sytuacja na drogach i sieć

Podsumowanie

Symulacja sieci pozwala na:

- Planowanie układu sieci,
- Modelowanie dynamicznego zachowania sieci w celu określenia wąskich gardeł i słabych ogniw,
- Modelowanie przepływu elementów sieci,
- Optymalizację globalnych wskaźników sieciowych.

4.6 Projekty symulacji logistycznych

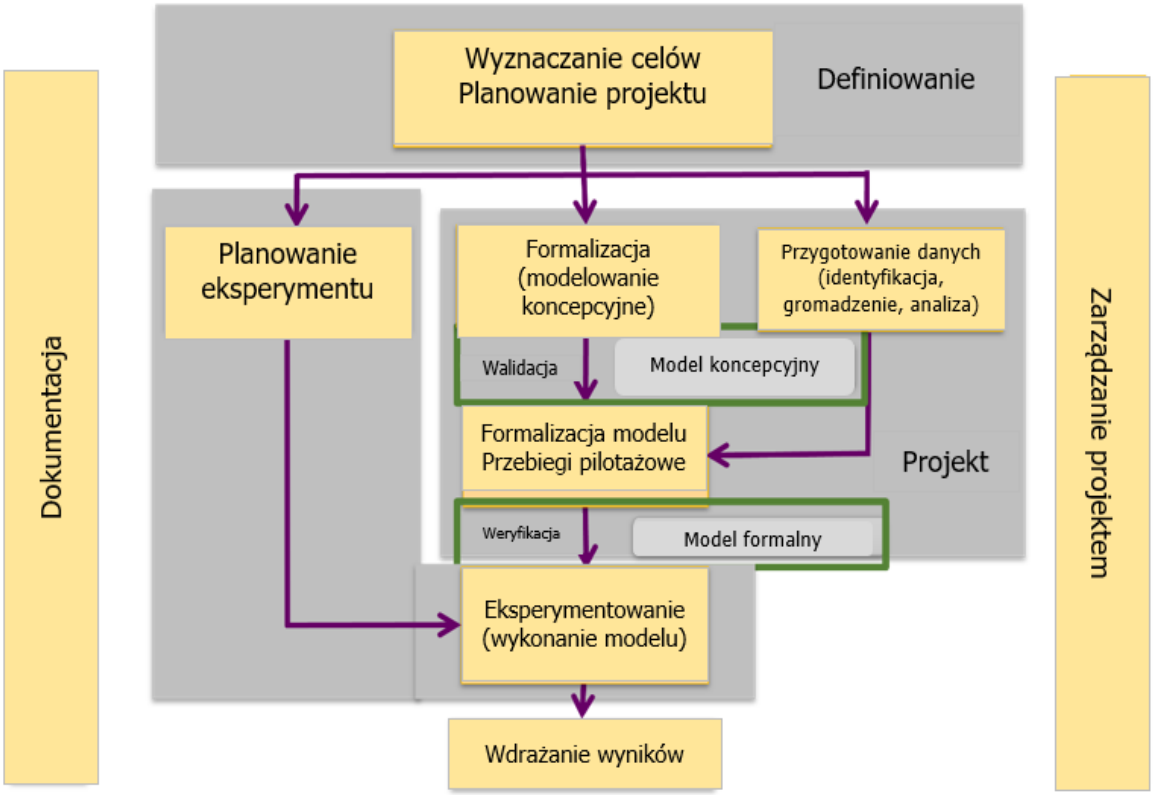
Projekty symulacji logistycznych są projektowane zgodnie z paradygmatem Projektowanie na potrzeby Six Sigma (DFSS) i opierają się na cyklu doskonalenia Deminga:

- Planowanie: definicja systemu i celów,
- Wykonanie: projekt modelu symulacyjnego,
- Analiza: eksperymentowanie z modelem symulacyjnym i ocena alternatywnych rozwiązań,
- Działanie: wykorzystanie wyników symulacji do wdrożenia

ulepszeń.

Każdy projekt symulacji logistycznej (rysunek 4.6) składa się z siedmiu faz:

1. Plan strategiczny: analiza istniejących i proponowanych zasobów i procesów.
2. Model koncepcyjny: abstrakcyjny model systemu i definicja predyspozycji, gromadzenie danych.
3. Model logiczny: przepływ obiektów, pętla sprzężenia zwrotnego lub diagram sieciowy modelu systemu
4. Model symulacyjny: opracowanie odpowiedniego modelu symulacyjnego.
5. Weryfikacja i walidacja modelu symulacyjnego: sprawdzenie zgodności i spójności modelu.
6. Analiza oparta na modelu symulacyjnym: projektowanie i przeprowadzanie eksperymentów.
7. Wykorzystanie wyników symulacji do opracowania planu działania: projekcja ulepszeń procesu.



Rysunek 4.6 Proces modelowania i analizy symulacyjnej (SMA)

4.7 Wnioski

W logistyce SMA jest ważnym elementem badań operacyjnych (OR), umożliwiającym optymalizację procesów.

Dynamika systemowa (SD) to metodologia analizy złożonych, dynamicznych i nieliniowych interakcji w systemach, w wyniku której powstają nowe struktury i polityki mające na celu poprawę zachowania systemu. W tym przypadku przepływy fizyczne i informacyjne są uwzględniane w celu zmniejszenia ich opóźnień i ostatecznie inwentaryzacji SC.

Inną popularną metodologią zorientowaną na proces jest symulacja zdarzeń dyskretnych (DES). Jest to jedno z najczęściej stosowanych i najbardziej elastycznych narzędzi analitycznych w SMA systemów produkcyjnych. Z powodzeniem radzi sobie z niepewnością i zapewnia możliwości porównania alternatywnych sposobów skrócenia czasu realizacji, a także optymalizacji wykorzystania maszyn i zasobów.

Metodologią pomocną w zrozumieniu zachowań organizacji i ich interakcji (np. SC i ich podmiotów) jest symulacja agentowa (ABS). Symulacja sieci (NS), będąca specjalnym rodzajem ABS, pozwala na modelowanie i optymalizację sieci nakładkowych (ruchu).

Prowadzi to do wniosku, że holistyczne podejście do stosowania SMA w logistyce w znacznym stopniu przyczynia się do podejmowania złożonych decyzji dotyczących projektowania procesów, w których występuje wiele zmiennych wchodzących ze sobą w interakcje. Przydatne zintegrowane podejście, obejmujące metodologie SD, DES i ABS, które może kwantyfikować przepływy pracy na różnych poziomach łańcucha dostaw, zostało przedstawione w (Gumzej & Rakovska, 2020). W analizie i optymalizacji przepływu ruchu metodologia NS zapewnia niezbędne ramy w zakresie monitorowania i dostrajania kluczowych wskaźników wydajności (Šinko i Gumzej, 2021).

BIBLIOGRAFIA DO ROZDZIAŁU 4.

- Conant R.C. and Ashby W.R. (1970). Every good regulator of a system must be a model of that system, Int. J. Systems Sci., 1(2), pp. 89-97.
- Gumzej, R. and Rakovska, M. (2020). Simulation modeling and analysis for sustainable supply chains. In Ecoproduction – Sustainable logistics and production in

industry 4.0 : new opportunities and challenges, Grzybowska, K., Awasthi, A., Sawhney, R. (ed.). Springer Nature, pp. 145-160.

- JaamSim Development Team (2023). JaamSim: Discrete-Event Simulation Software. Version 2023-08. [Available at: https://jaamsim.com, dostęp November 8th, 2023]
- Šinko, S. and Gumzej, R. (2021). Towards smart traffic planning by traffic simulation on microscopic level. International journal of applied logistics, 11(1), pp. 1-17.
- Wilensky, U. (1999). NetLogo. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL. [Dostępne: http://ccl.northwestern.edu/netlogo/, access November 8th, 2023]
- Pablo A.L., Behrisch, M., Bieker-Walz, L., Erdmann, J., Flötteröd, Y.-P., Hilbrich, R., Lücken, L., Rummel, J., Wagner, P. and Wießner, E. (2018). Microscopic Traffic Simulation using SUMO. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC), IEEE. The 21st IEEE International Conference on Intelligent Transportation Systems, 4.-7. Nov. 2018, Maui, USA, pp. 2575-2582.

5. Regresja liniowa z pojedynczym i wieloma regresorami

Decyzje zarządcze często opierają się na związku między dwiema lub więcej zmiennymi. Na przykład menedżer ds. marketingu może próbować prognozować sprzedaż przy określonym poziomie wydatków na reklamę po zbadaniu związku między tymi wydatkami a sprzedażą.

W drugim przypadku przedsiębiorstwo publiczne może wykorzystać stosunek między dzienną maksymalną wartością temperatury a zapotrzebowaniem na energię elektryczną, aby przewidzieć zużycie energii elektrycznej. Czasami menedżer polega na intuicji. Intuicyjnie ocenia, w jaki sposób te dwie zmienne są ze sobą powiązane. Jeśli jednak możliwe jest uzyskanie danych, warto zastosować procedurę statystyczną zwaną analizą regresji, aby pokazać, w jaki sposób te dwie zmienne są ze sobą powiązane.

W terminologii regresji zmienna przewidywana nazywana jest zmienną zależną.

Zmienna lub zmienne używane do przewidywania wartości zmiennej zależnej nazywane są zmiennymi niezależnymi.

Analizując wpływ wydatków na reklamę na sprzedaż, sprzedaż byłaby zatem zmienną zależną. Wydatki na reklamę byłyby zmienną niezależną. W notacji statystycznej y oznacza zmienną zależną, a x oznacza zmienną niezależną.

W tej sekcji przyjrzymy się najprostszemu typowi analizy regresji, który obejmuje jedną zmienną niezależną i jedną zmienną zależną. Związek między tymi dwiema zmiennymi będzie przybliżony linią prostą. Nazywa się to prostą regresją liniową. Analiza regresji obejmująca dwie lub więcej zmiennych niezależnych nazywana jest analizą regresji wielorakiej.

5.1 Model prostej regresji liniowej

Best Burger to sieć restauracji szybkiej obsługi zlokalizowanych na obszarze wielu stanów. Lokalizacje Best Burger znajdują się w pobliżu kampusów uniwersyteckich. Menedżerowie uważają, że kwartalna sprzedaż tych restauracji (oznaczona przez y) jest dodatnio skorelowana z wielkością populacji studentów (oznaczonej przez x). Restauracje w pobliżu kampusów z dużą liczbą

studentów mają tendencję do generowania większej sprzedaży niż te w pobliżu kampusów z małą liczbą studentów. Korzystając z analizy regresji, możemy opracować równanie, które pokazuje, w jaki sposób zmienna zależna y jest powiązana ze zmienną niezależną x.

5.2 Model regresji i równanie regresji

W przypadku Best Burger populację stanowią wszystkie restauracje Best Burger. Dla każdej restauracji w populacji istnieje wartość x (populacja studentów) i odpowiadająca jej wartość y (kwartalna sprzedaż). Równanie opisujące, w jaki sposób y jest powiązane z x, nazywane jest **modelem regresji**.

$$y = \beta_0 + \beta_1 x + \epsilon$$

β_0 i β_1 są nazywane parametrami modelu, ϵ (grecka litera epsilon) jest zmienną losową nazywaną błędem modelu. Błąd reprezentuje zmienność y, której nie można wyjaśnić liniową zależnością między x oraz y.

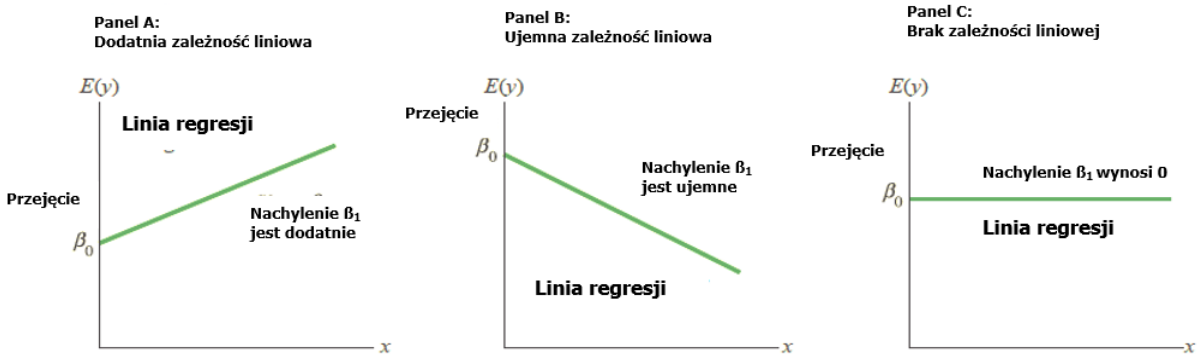
Populacja wszystkich restauracji Best Burger może być również postrzegana jako zbiór subpopulacji, po jednej dla każdej oddzielnej wartości x. Na przykład jedna subpopulacja składa się ze wszystkich restauracji Best Burger w pobliżu kampusów uniwersyteckich z 8000 studentów. Druga subpopulacja składa się ze wszystkich restauracji Best Burger zlokalizowanych w pobliżu kampusów uniwersyteckich z 9000 studentów i tak dalej. Każda subpopulacja ma odpowiadający jej rozkład wartości y. Każdy rozkład wartości y ma swoją średnią lub wartość oczekiwaną. Równanie opisujące wartość oczekiwaną y, oznaczaną jako E(y), która jest powiązana z x, nazywane jest równaniem regresji. Równanie prostej regresji liniowej jest następujące:

$$E(y) = \beta_0 + \beta_1 x$$

Wykresem równania prostej regresji liniowej jest linia prosta. β_0 reprezentuje wartość początkową linii regresji, β_1 to współczynnik kierunkowy linii, a $E(y)$ to wartość średnia lub wartość oczekiwana y dla danej wartości x.

Przykłady możliwych linii regresji pokazano na poniższym rysunku 5.1. Linia regresji w przypadku A pokazuje, że wartość y jest dodatnio skorelowana z x. Wraz ze wzrostem wartości x rosną również wartości $E(y)$. Linia regresji na rysunku B pokazuje wartość y, która jest ujemnie skorelowana z x. Mniejsze wartości $E(y)$ są powiązane z wyższymi wartościami x. Linia regresji na rysunku C pokazuje przypadek, w którym wartość y nie jest powiązana

z x. Oznacza to, że oczekiwana wartość y jest taka sama dla każdej wartości x.



Rysunek 5.1 Przykłady wykresów zależności liniowej

5.3 Szacowane równanie regresji

Gdyby wartości parametrów populacji były znane β_0 i β_1, moglibyśmy użyć powyższego równania do obliczenia wartości y dla danej wartości x. W praktyce parametry te są trudno dostępne, więc są po prostu szacowane na podstawie danych z próby. Statystyki próby (oznaczone jako b_0 i b_1) są obliczane jako oszacowania parametrów populacji β_0 i β_1. Zastąpienie wartości statystyk z próby b_0 i b_1 zamiast β_0 i β_1 w równaniu regresji daje nam nowe, oszacowane równanie regresji. Oszacowane równanie regresji dla prostej regresji liniowej jest następujące:

$$\hat{y} = b_0 + b_1 x$$

Wykres oszacowanej prostej regresji liniowej nazywany jest oszacowaną linią regresji. b_0 reprezentuje wartość początkową linii regresji, b_1 to współczynnik kierunkowy linii.

Poniżej pokazane jest, jak użyć metody najmniejszych kwadratów do obliczenia wartości b_0 i b_1 w oszacowanym równaniu regresji.

Ogólnie $\hat{y}$ (wynik dla $E(y)$) to średnia wartość y dla danej wartości x. Gdybyśmy teraz chcieli oszacować oczekiwaną wartość kwartalnej sprzedaży dla wszystkich restauracji Best Burger zlokalizowanych w pobliżu kampusów z 10000 studentów, wartość x zostałaby zastąpiona wartością 10000 w ostatnim równaniu. W niektórych przypadkach możemy być jednak bardziej zainteresowani prognozowaniem sprzedaży tylko dla jednej konkretnej restauracji. Załóżmy na przykład, że chcesz prognozować kwartalną sprzedaż dla restauracji, którą planujesz zbudować w pobliżu uczelni z 10000 studentów. Jak się okazuje, nawet w tym

przypadku najlepszym predyktorem (źródłem prognozy) wartości y dla danej wartości x jest wartość $\hat{y}$.

5.4 Metoda najmniejszych kwadratów

Metoda najmniejszych kwadratów to procedura, w której wykorzystując przykładowe dane, znajdujemy równanie szacowanej linii regresji. Aby zilustrować metodę najmniejszych kwadratów, załóżmy, że dane zostały zebrane z próby 10 restauracji Best Burger w pobliżu kampusów uniwersyteckich. Przez x_i oznaczymy wielkość populacji studentów (w tysiącach), a przez y_i wielkość kwartalnej sprzedaży (w tysiącach EUR). x_i w y_i dla 10 przykładowych restauracji podsumowano

w poniższej tabeli (rysunek 5.2). Widzimy, że restauracja 1, z x_1 = 2 i y_1 = 58, znajduje się w pobliżu kampusu z 2000 studentów i osiąga kwartalną sprzedaż w wysokości 58 000 EUR. Restauracja 2, z x_2= 6 i y_2 = 105, znajduje się w pobliżu kampusu z 6000 studentów i osiąga kwartalną sprzedaż w wysokości 105 000 €. Restauracja o najwyższej wartości sprzedaży to restauracja 10, która znajduje się w pobliżu kampusu z 26 000 studentów i osiąga kwartalną sprzedaż w wysokości 202 000 €.

Poniżej na rysunku 5.2. przedstawiono wykres punktowy danych nazywany też wykresem rozrzutu. Populacja studentów jest pokazana na osi poziomej, a kwartalna sprzedaż na osi pionowej. Wykresy rozrzutu dla analizy regresji są konstruowane ze zmienną niezależną x na osi poziomej i zmienną zależną y na osi pionowej. Wykres rozrzutu pozwala nam zatem wyciągnąć wstępne wnioski na temat możliwego związku między zmiennymi.

Restauracja	Populacja studentów (tys.)	Sprzedaż kwartalna (tys. €)
i	x_i	y_i
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

Rysunek 5.2 Wykres punktowy danych

Jakie wstępne wnioski można wyciągnąć z poniższego rysunku 5.3? Wyższa sprzedaż kwartalna występuje w kampusach z większą populacją studentów. Ponadto istnieje stała zależność między wielkością populacji studentów a kwartalną sprzedażą, którą można opisać linią prostą. Pomiędzy x i y istnieje dodatnia zależność liniowa. Dlatego wybrany został model prostej regresji liniowej, aby przedstawić związek między kwartalną sprzedażą a populacją studentów. Biorąc pod uwagę ten wybór, naszym następnym zadaniem jest wykorzystanie przykładowej tabeli danych do określenia wartości b_0 i b_1, które są ważnymi parametrami w estymacji równania prostej regresji liniowej. Dla i-tej restauracji oszacowane równanie regresji to:

$$\hat{y}_i = b_0 + b_1 x_i$$

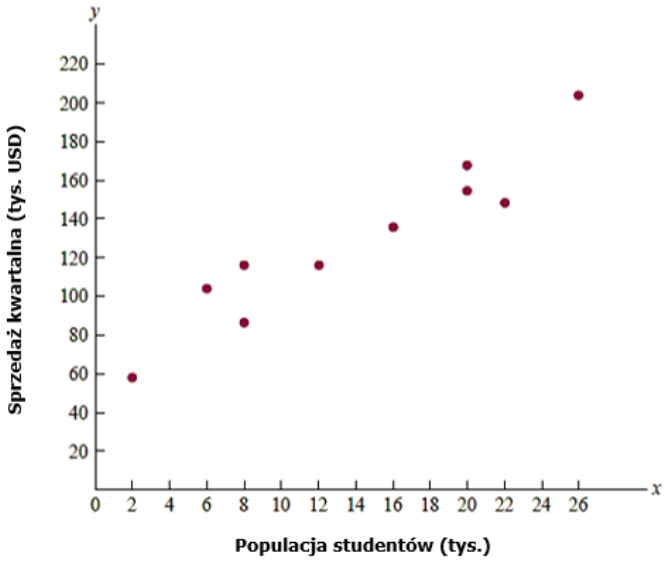
gdzie:

$\hat{y}_i$ - modelowa (teoretyczna) wartość sprzedaży kwartalnej (1000 EUR) dla i-tej restauracji,

b_0 - wyraz wolny oszacowanej linii regresji, inaczej przecięcie (z osią y),

b_1 - współczynnik kierunkowy oszacowanej linii regresji, inaczej nachylenie

x_i _ wielkość populacji studentów (1000) dla i-tej restauracji.



Rysunek 5.3 Wykres punktowy

y_i oznacza zaobserwowaną (rzeczywistą) sprzedaż dla restauracji i, a $\hat{y}_i$, teoretyczna wartość sprzedaży dla restauracji i, każda restauracja w próbie będzie miała zaobserwowaną wartość sprzedaży y_i i teoretyczną wartość sprzedaży $\hat{y}_i$. Aby oszacowana linia regresji zapewniała dobre dopasowanie do danych, chcemy, aby różnice między obserwowanymi wartościami sprzedaży a teoretycznymi wartościami sprzedaży były jak najmniejsze.

Metoda najmniejszych kwadratów wykorzystuje przykładowe dane do uzyskania wartości b_0 i b_1.

Minimalizacja sumy kwadratów odchyleń między obserwowanymi wartościami zmiennej zależnej y_i a przewidywaną wartością zmiennej zależnej $\hat{y}$. Punktem wyjścia do obliczenia minimalnej sumy metodą najmniejszych kwadratów jest wyrażenie:

Kryterium sumy minimalnej: $\min \sum(y_i - \hat{y}_i)^2$,

gdzie

y_i - zaobserwowana wartość zmiennej zależnej dla i-tej obserwacji,

$\hat{y}_i$ – teoretyczna wartość zmiennej zależnej dla i-tej obserwacji.

Współczynnik kierunkowy linii regresji i wyraz wolny:

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

x_i _ wartość zmiennej niezależnej dla i-tej obserwacji,

y_i _ wartość zmiennej zależnej dla i-tej obserwacji,

$\bar{x}$ _ średnia wartość dla zmiennej niezależnej,

$\bar{y}$ _ średnia wartość dla zmiennej zależnej,

n _ całkowita liczba obserwacji.

Poniżej przedstawiono niektóre obliczenia potrzebne do opracowania szacunkowej linii regresji metodą najmniejszych kwadratów. Przy próbie 10 restauracji mamy n=10 obserwacji. Powyższe równania wymagają najpierw obliczenia średniej wartości x i średniej wartości y.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{140}{10} = 14, \qquad \bar{y} = \frac{\sum y_i}{n} = \frac{1300}{10} = 130$$

Alternatywne równanie obliczeniowe b_1:

$$b_1 = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\sum x_i)^2}$$

Korzystając z ostatnich równań i informacji na rysunku 5.4, można obliczyć współczynnik kierunkowy linii regresji dla przykładu restauracji Best Burger. Obliczenie nachylenia (b_1) przedstawiono poniżej.

Rysunek 5.4 przedstawia wykres tego równania na wykresie rozrzutu.

Nachylenie oszacowanego równania regresji lub współczynnik kierunkowy równania ($b_1 = 5$) jest dodatnie.

Restauracja *i*	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	2	58	−12	−72	864	144
2	6	105	−8	−25	200	64
3	8	88	−6	−42	252	36
4	8	118	−6	−12	72	36
5	12	117	−2	−13	26	4
6	16	137	2	7	14	4
7	20	157	6	27	162	36
8	20	169	6	39	234	36
9	22	149	8	19	152	64
10	26	202	12	72	864	144
Suma:	140	1300			2840	568
	Σx_i	Σy_i			$\Sigma(x_i - \bar{x})(y_i - \bar{y})$	$\Sigma(x_i - \bar{x})^2$

Rysunek 5.4 Wykres równania na wykresie rozrzutu.

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{2840}{568} = 5$$

Następnie obliczany jest wyraz wolny (b_0).

$$b_0 = \bar{y} - b_1\bar{x} = 130 - 5(14) = 60$$

W ten sposób szacowane jest równanie regresji:

$$\hat{y} = 60 + 5x$$

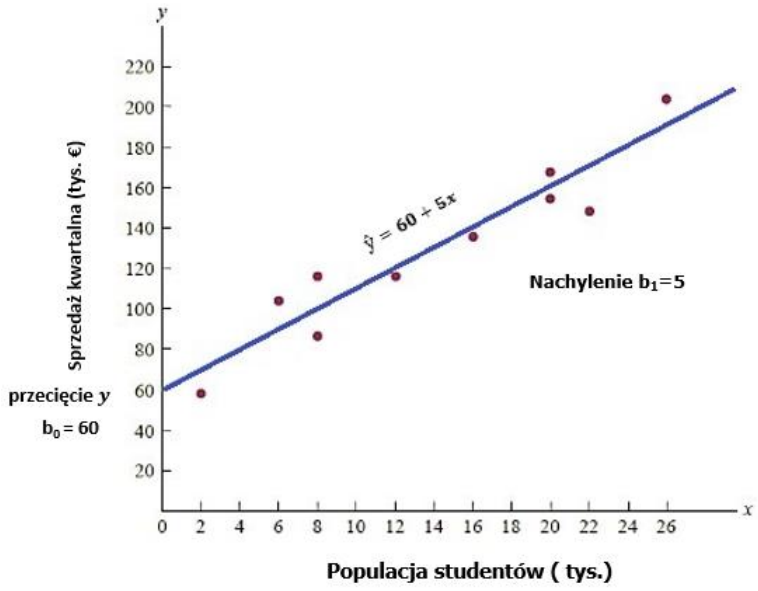
Rysunek 5.5 przedstawia wykres tego równania na wykresie punktowym.

Nachylenie oszacowanego równania regresji (b_1 = 5) jest dodatnie, co oznacza, że wraz ze wzrostem populacji studentów wzrasta sprzedaż. W rzeczywistości możemy wywnioskować (w oparciu o zmierzoną sprzedaż w tysiącach i populację studentów w tysiącach), co oznacza, że wzrost populacji studentów o 1000 wiąże się ze wzrostem oczekiwanej sprzedaży o 5000; tj. oczekuje się, że kwartalna sprzedaż wzrośnie o 5 USD na studenta.

Jeśli uważamy, że równanie regresji, oszacowane metodą najmniejszych kwadratów, odpowiednio opisuje związek między x i y, rozsądne wydaje się użycie oszacowanego równania regresji do przewidzenia wartości y dla danej wartości x. Na przykład, jeśli chcemy przewidzieć kwartalną sprzedaż dla restauracji znajdującej się w pobliżu kampusu z 16 000 studentów, obliczona byłaby ona na podstawie wzoru:

$$\hat{y} = 60 + 5 * 16 = 140$$

W związku z tym należy założyć, że kwartalna sprzedaż tej restauracji wyniesie 140 000. W kolejnych sekcjach omawiane są metody oceny stosowności wykorzystania oszacowanego równania regresji do estymacji i prognozowania.



Rysunek 5.5 Wykres regresji prostej kwartalnej sprzedaży

5.5 Współczynnik determinacji

Dla przykładu restauracji Best Burger opracowane zostało szacunkowe równanie regresji $y = 60 + 5x$ dla w przybliżeniu liniowej zależności między wielkością populacji studentów x a kwartalną sprzedażą y. Teraz pytanie brzmi: jak dobrze oszacowane równanie regresji pasuje do danych? W tej sekcji jest pokazane, że współczynnik determinacji zapewnia miarę dobroci (stopnia) dopasowania dla oszacowanego równania regresji. Dla i-tej obserwacji różnica między zaobserwowaną wartością zmiennej zależnej y_i a przewidywaną wartością zmiennej zależnej nazywana jest i - tą wartością rezydualną. Suma kwadratów tych reszt lub błędów jest funkcją nachylenia oraz wyrazu wolnego. Metodą najmniejszych kwadratów wyznaczamy wartości b_1 i b_0, dla których powyższa funkcja osiąga minimum. Wielkość ta oznaczana jest jako SSE.

$$SSE = \sum (y_i - \hat{y}_i)^2$$

Wartość SSE jest miarą błędu w wykorzystaniu oszacowanego równania regresji do opisu wartości zmiennej zależnej w próbie. Rysunek 5.6 pokazuje obliczenia potrzebne do obliczenia sumy kwadratów błędów dla przypadku Best Burger.

Restauracja i	x_i=Populacja studentów (tys.)	y_i=Sprzedaż kwartalna (tys. €)	Prognozowana sprzedaż $\hat{y}_i = 60 + 5x$	Błąd $y_i - \hat{y}_i$	Kwadrat błędu $(y_i - \hat{y}_i)^2$
1	2	58	70	−12	144
2	6	105	90	15	225
3	8	88	100	−12	144
4	8	118	100	18	324
5	12	117	120	−3	9
6	16	137	140	−3	9
7	20	157	160	−3	9
8	20	169	160	9	81
9	22	149	170	−21	441
10	26	202	190	12	144
					SSE = 1530

Rysunek 5.6 Kwadraty błędów dla przypadku Best Burger

Załóżmy, że poproszono nas o oszacowanie kwartalnej sprzedaży bez znajomości wielkości populacji studentów. Nie znając żadnych powiązanych zmiennych, użylibyśmy średniej z próby jako oszacowania kwartalnej sprzedaży w dowolnej restauracji. Tabela na rysunku 5.6 pokazuje, że dla danych sprzedaży $\sum y_i$ = 1300. Dlatego średnia kwartalna wartość sprzedaży dla próby 10 restauracji Best Burger wynosi $\sum y_i$/n = 1300/10 = 130. Na rysunku 5.7 pokazana jest suma kwadratów odchyleń uzyskanych przy użyciu średniej próby 130 do przewidywania wartości kwartalnej sprzedaży dla każdej restauracji w próbie. Dla i-tej restauracji w próbie różnica $y_i - \hat{y}_i$ stanowi miarę błędu, który jest uwzględniony w aplikacji modelu do wielkości sprzedaży. Odpowiednia suma kwadratów, zwana całkowitą sumą kwadratów, jest oznaczana przez SST.

$$SST = \sum (y_i - \bar{y})^2$$

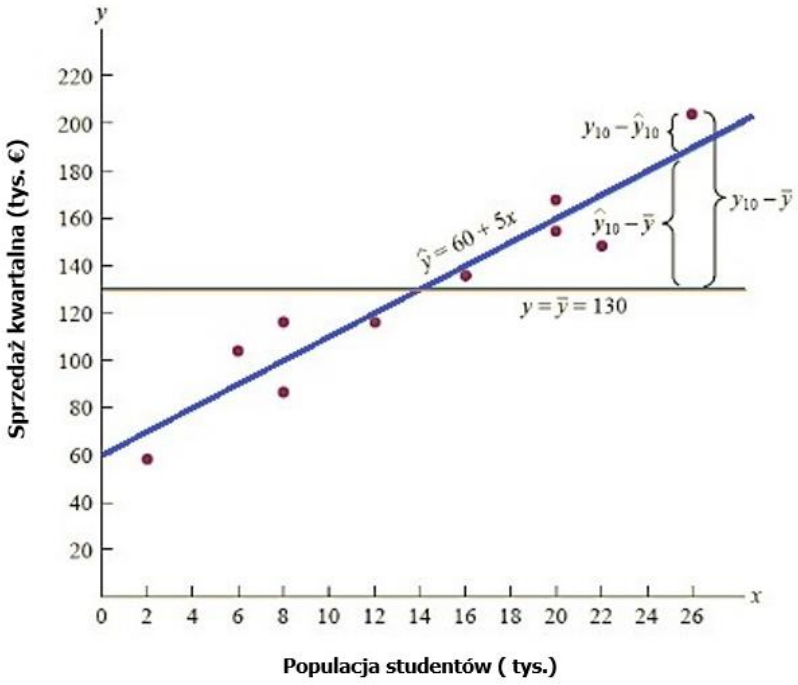
Restauracja i	x_i=Populacja studentów (tys.)	y_i=Sprzedaż kwartalna (tys. €)	Odchylenie $y_i - \bar{y}$	Kwadrat odchylenia $(y_i - \bar{y})^2$
1	2	58	−72	5184
2	6	105	−25	625
3	8	88	−42	1764
4	8	118	−12	144
5	12	117	−13	169
6	16	137	7	49
7	20	157	27	729
8	20	169	39	1521
9	22	149	19	361
10	26	202	72	5184
				SST = 15,730

Rysunek 5.7 Suma kwadratów

Suma na dole ostatniej kolumny na rysunku 5.7 to całkowita suma kwadratów dla restauracji BestBurger SST = 15 730. Na rysunku 5.8 pokazana jest szacowana linia regresji $y = 60 + 5x$ i linia odpowiadająca y = 130. Należy zauważyć, że punkty skupiają się bardziej wokół szacowanej linii regresji niż wokół linii y = 130. Na przykład dla 10. restauracji w próbie widzimy, że błąd jest znacznie większy, gdy 130 jest używana jako wielkość y, niż gdy użyjemy równania $y = 60 + 5x$, dla którego wynosi on 190. Można myśleć o SST jako o pomiarze tego, jak dobrze obserwacje skupiają się wokół linii o stałej wartości y wynoszącej 130, a SSE jako o pomiarze tego, jak dobrze obserwacje skupiają się wokół prostej regresji.

Aby zmierzyć, jak bardzo wartości na oszacowanej linii regresji odbiegają od średniej, obliczana jest kolejna suma kwadratów. Ta suma kwadratów, zwana sumą kwadratów z powodu regresji, jest oznaczana jako SSR.

$$SSR = \sum (\hat{y}_i - \bar{y})^2$$



Rysunek 5.8 Linia regresji dla przypadku Best Burger

Z poprzedniej dyskusji powinniśmy się spodziewać, że SST, SSR i SSE są ze sobą powiązane. W rzeczywistości związek między tymi trzema sumami kwadratów jest jednym z najważniejszych wyników w statystyce.

5.6 Związek między SST, SSR i SSE

$$SST = SSR + SSE$$

gdzie:

SST = całkowita suma kwadratów,

SSR = suma kwadratów wynikająca z regresji,

SSE = suma kwadratów spowodowana błędem.

Równanie $(SST = SSR + SSE)$ pokazuje, że całkowitą sumę kwadratów można podzielić na dwa składniki, sumę kwadratów wielkości teoretycznych regresji i sumę kwadratów błędów. Jeśli więc znane są wartości dowolnych dwóch z tych sum kwadratów, trzecią sumę kwadratów można łatwo obliczyć. Na przykład w przypadku restauracji Best Burger wiadomo

już, że SSE = 1530 i SST = 15730; dlatego rozwiązując SSR w powyższym równaniu, okazuje się, że suma kwadratów wynikająca z regresji wynosi:

$$SSR = SST - SSE = 15730 - 1530 = 14200$$

Zobaczmy teraz, jak możemy wykorzystać trzy sumy kwadratów, SST, SSR i SSE, aby zapewnić kryterium dobrego dopasowania dla oszacowanego równania regresji. Oszacowane równanie regresji zapewniłoby idealne dopasowanie, gdyby każda wartość zmiennej zależnej $yi\_$ leżała losowo na oszacowanej linii regresji. W tym przypadku wynosiłaby ona zero dla każdej obserwacji, co skutkowałoby SSE =0. Ponieważ SST = SSR + SSE, widzimy, że dla idealnego dopasowania SSR musi być równe SST, a stosunek (SSR/SST) musi być równy jeden. Gorsze dopasowanie spowoduje większe wartości SSE. Rozwiązując równanie dla SSE, widzimy, że SSE = SST - SSR. Dlatego największa wartość SSE (a tym samym najgorsze dopasowanie) występuje, gdy SSR = 0 i SSE = SST.

Stosunek SSR/SST jest wykorzystywany do estymacji, która ma wartości od zera do jednego pasujące do szacowanego równania regresji.

Współczynnik ten nazywany jest współczynnikiem determinacji i jest oznaczany przez r^2.

$$r^2 = \frac{SSR}{SST}$$

Dla przykładu restauracji Best Burger wartość współczynnika determinacji wynosi:

$$r^2 = \frac{SSR}{SST} = \frac{14200}{15730} = 0{,}9027$$

Gdy współczynnik determinacji jest wyrażony w procentach, r^2 można go interpretować jako procent całkowitej sumy kwadratów wyjaśniony za pomocą oszacowanego równania regresji. W przypadku Best Burger Restaurants możemy stwierdzić, że 90,27% całkowitej sumy kwadratów można wyjaśnić za pomocą oszacowanego równania regresji $y = 60 + 5x$, aby przewidzieć kwartalną sprzedaż. Innymi słowy, 90,27% zmienności sprzedaży można wyjaśnić liniową zależnością między wielkością populacji studentów a sprzedażą. Powinniśmy być zadowoleni, że tak dobrze pasuje do oszacowanego równania regresji.

5.7 Współczynnik korelacji

Współczynnik korelacji można traktować jako opisową miarę siły współzależności liniowej między dwiema zmiennymi, x i y. Wartości współczynnika korelacji zawsze mieszczą się

w przedziale od -1 do +1. Wartość +1 oznacza, że dwie zmienne x i y są doskonale skorelowane. Oznacza to, że linia prosta o dodatnim nachyleniu stanowi wykres punktowy obserwacji. Dlatego wykres rozrzutu nazywa się też wykresem korelacji. Wartość -1 oznacza, że zmienne x i y są doskonale skorelowane. Wszystkie punkty danych leżą na prostej o ujemnym nachyleniu. Wartości współczynnika korelacji bliskie zeru oznaczają, że x i y nie współzależą liniowo.

Jeśli przeprowadzono już analizę regresji i obliczono współczynnik determinacji r^2, współczynnik korelacji próby można obliczyć w następujący sposób.

$$r_{xy} = (znak\ b_1)\sqrt{współczynnik\ determinacji} \quad (13)$$

$$r_{xy} = (znak\ b_1)\sqrt{r^2} \quad (14)$$

WSPÓŁCZYNNIK KORELACJI PEARSONA: PRZYKŁADOWE DANE

$$r_{xy} = \frac{V_{xy}}{V_x V_y} = \frac{\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n-1}}{\sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}\sqrt{\frac{\sum(y_i-\bar{y})^2}{n-1}}} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}}$$

$$r_{xy} = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

gdzie:

$V_{xy} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n-1}$, $V_x = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}$, $V_y = \sqrt{\frac{\sum(y_i-\bar{y})^2}{n-1}}$

Znak współczynnika korelacji próby jest dodatni, jeśli oszacowane równanie regresji ma dodatnie nachylenie ($b_1 > 0$) i ujemny, jeśli oszacowane równanie regresji ma ujemne nachylenie ($b_1 < 0$).

W przypadku Best Burger wartość współczynnika determinacji odpowiadającego oszacowanemu równaniu regresji $y = 60 + 5x$ wynosi 0,9027. Ponieważ nachylenie oszacowanego równania regresji jest dodatnie, równanie (13-14) pokazuje, że próbny współczynnik korelacji wynosi r_{xy}=0,9501. Na podstawie próbnego współczynnika korelacji można wywnioskować, że istnieje silna dodatnia zależność liniowa między x a y.

W przypadku liniowej zależności między dwiema zmiennymi, zarówno współczynniki determinacji, jak i współczynnik korelacji stanowią miarę siły związku w próbce.

Współczynnik determinacji zapewnia miarę między zero a jeden, podczas gdy współczynnik korelacji zapewnia miarę między -1 a +1. Chociaż współczynnik korelacji jest ograniczony do liniowej zależności między dwiema zmiennymi, współczynnik determinacji może być stosowany do zależności nieliniowych i do zależności, które mają dwie lub więcej niezależnych zmiennych. W ten sposób współczynnik determinacji zapewnia szerszy zakres zastosowań.

5.8 Model regresji wielorakiej

W tej sekcji kontynuujemy nasze badanie analizy regresji, rozważając sytuacje obejmujące dwie lub więcej zmiennych niezależnych. Ten obszar tematyczny, zwany analizą regresji wielorakiej, pozwala nam wziąć pod uwagę więcej czynników, a tym samym uzyskać lepsze prognozy niż jest to możliwe w przypadku jednorakiej regresji liniowej.

Analiza regresji wielorakiej to badanie, w jaki sposób zmienna zależna y jest powiązana z dwiema lub więcej zmiennymi niezależnymi. W ogólnym przypadku oznaczymy przez p liczbę zmiennych niezależnych.

5.9 Model i równanie regresji

Pojęcia modelu regresji i równania regresji wprowadzone w poprzedniej sekcji mają zastosowanie w przypadku regresji wielorakiej. Równanie opisujące, w jaki sposób zmienna zależna y jest powiązana ze zmiennymi niezależnymi $x_1, x_2, \ldots, x_p$ i wielkością błędu, nazywane jest modelem regresji wielorakiej. Zaczynamy od założenia, że model regresji wielorakiej ma następującą postać.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

W modelu regresji wielorakiej $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ są parametrami, a składnik błędu (ϵ) jest zmienną losową. Dokładna analiza tego modelu ujawnia, że y jest liniową funkcją zmiennych $x_1, x_2, \ldots, x_p$ plus składnik błędu ϵ epsilon. Wielkość błędu uwzględnia zmienność y, której nie można wyjaśnić liniowym efektem p niezależnych zmiennych.

W sekcji 5.10 omawiane są założenia dla modelu regresji wielorakiej i epsilon. Jednym z założeń jest to, że średnia lub wartość oczekiwana (ϵ) wynosi zero. Konsekwencją tego założenia jest to, że średnia lub wartość oczekiwana y, oznaczana jako $E(y)$, jest równa $\beta_0 +$

$\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$. Równanie opisujące zależność wartości średniej y od $x_1, x_2, \ldots, x_p$ nazywane jest równaniem regresji wielorakiej.

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p (5.9)$$

5.10 Oszacowane równanie regresji wielorakiej

Jeśli znane są wartości $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$, równanie (5.9), można wykorzystać do obliczenia średniej wartości *y* przy danych wartościach $x_1, x_2, \ldots, x_p$. Niestety, te wartości parametrów zazwyczaj nie są znane i muszą zostać oszacowane na podstawie danych z próby. Prosta próba losowa jest używana do obliczenia statystyk próby $b_0, b_1, b_2, \ldots, b_p$, które będą używane jako estymatory punktowe parametrów $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$. Te przykładowe statystyki zapewniają następujące oszacowanie równania regresji wielokrotnej:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

gdzie:

$b_0, b_1, b_2, \ldots, b_p$ są oszacowaniami $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$,

$\hat{y} =$ przewidywana wartość zmiennej zależnej.

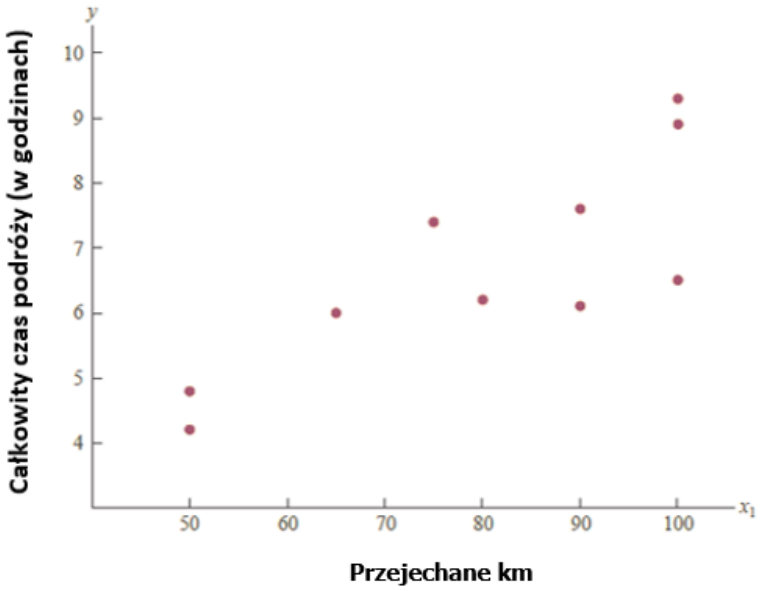
Przykład: Frigo Transport Company

Jako ilustrację analizy regresji wielorakiej rozważymy problem, przed którym stoi Frigo Trucking Company, niezależna firma transportowa w południowych Włoszech. Największa część działalności Frigo obejmuje dostawy na całym obszarze lokalnym. Aby lepiej opracować harmonogramy pracy, menedżerowie chcą zaplanować wspólny dzienny czas podróży dla swoich kierowców.

Początkowo menedżerowie uważali, że całkowity dzienny czas podróży będzie ściśle powiązany z liczbą kilometrów przejechanych w codziennych dostawach. Prosta losowa próba 10 przydziałów kierowców dostarczyła danych pokazanych w tabeli 5.1 i wykresu rozrzutu (rys.5.9). Po zapoznaniu się z tym wykresem rozrzutu menedżerowie założyli, że model jednorakiej regresji liniowej może być użyty do opisania zależności $y = \beta_0 + \beta_1 x_1 + \epsilon$ między całkowitym czasem podróży (y) a liczbą przejechanych kilometrów (x_1).

Tabela 5.1 Dane dla przykładowej firmy Frigo Transport Company

Zadanie kierowcy	x_i =Przejechane km	y=Czas podróży (w godzinach)
1	100	9.3
2	50	4.8
3	100	8.9
4	100	6.5
5	50	4.2
6	80	6.2
7	75	7.4
8	65	6.0
9	90	7.6
10	90	6.1



Rysunek 5.9 Wykres rozrzutu dla Frigo Transport Company.

Do oszacowania parametrów β_0 i β_1 wykorzystano metodę najmniejszych kwadratów w celu opracowania szacunkowej regresji:

$$\hat{y} = b_0 + b_1 x_1$$

Powyższy rysunek przedstawia wyniki programu Minitab przy użyciu prostej regresji liniowej dla danych z powyższej tabeli. Oszacowane równanie regresji to:

$$\hat{y} = 1{,}27 + 0{,}0678x_1$$

Wartość *F* równa 15,81 i odpowiadająca jej wartość *p-value* równa 0,004 wskazują, że związek jest istotny na poziomie istotności 0,05. Oznacza to, że można odrzucić $H_0: \beta_1 = 0$, ponieważ wartość *p-value* jest mniejsza niż poziom istotności alfa (α=0,05). Należy zauważyć, że ten sam wniosek wynika z wartości t = 3,98 i związanej z nią wartości *p-value* równej 0,004. Możemy zatem stwierdzić, że związek między całkowitym czasem podróży a liczbą przejechanych kilometrów jest znaczący. Dłuższy czas podróży wiąże się z większą liczbą przejechanych kilometrów i odwrotnie. Przy współczynniku determinacji (wyrażonym w procentach) $r^2 = 66{,}4\,\%$ widać, że 66,4% zmienności czasu podróży można wyjaśnić liniowym efektem liczby przejechanych kilometrów.

Wynik ten jest całkiem dobry, ale menedżerowie mogą rozważyć dodanie drugiej zmiennej niezależnej, aby zmniejszyć błąd regresji (błąd szczątkowy) dla zmiennej zależnej.

DANE WYJŚCIOWE MINITAB DLA BUTLER TRUCKING
Z JEDNĄ ZMIENNĄ NIEZALEŻNĄ

Równanie regresji ma postać:
Czas=1,27+0,0678 km

Predyktor	Współczynnik	Współczynnik SE	T	p-value
Stała	1,274	1,401	0,91	0,39
Nachylenie (b_1)	0,06783	0,01706	3,98	0,004

S=1,00179 R-Sq=66,4% R-SQ(adj)=62,2%

Analiza wariancji

Źródło	DF	SS	MS	F	p-value
Regresja	1	15,871	15,871	15,81	0,004
Błąd szczątkowy	8	8,029	1,004		
Suma	9	23,9			

Rysunek 5.10 Wyniki z jedną zmienną niezależną

Próbując zidentyfikować drugą zmienną niezależną, menedżerowie uznali, że liczba dostaw może również przyczynić się do całkowitego czasu podróży. Dane Frigo Trucking: liczbę

przejechanych kilometrów (x_1) i liczbę dostaw (x_2), jako zmienne niezależne, przedstawiono na rysunku 5.11. Oszacowane równanie regresji to:

$$\hat{y} = -0{,}869 + 0{,}0611x_1 + 0{,}923x_2$$

DANE DLA FRIGO TRUCKING Z PRZEJECHANYMI KILOMETRAMI (x_1) I DOSTAWAMI (x_2) JAKO ZMIENNYMI NIEZALEŻNYMI

Zadanie kierowcy	x_1 =Przejechane km	x_2 =Liczba dostaw	y=Czas podróży (w godzinach)
1	100	4	9.3
2	50	3	4.8
3	100	4	8.9
4	100	2	6.5
5	50	2	4.2
6	80	2	6.2
7	75	3	7.4
8	65	4	6.0
9	90	3	7.6
10	90	2	6.1

Rysunek 5.11 Dane Frigo Trucing i zmienne niezależne

Przyjrzyjmy się bliżej wartościom b_1= 0,0611 i b_2= 0,923 w ostatnim równaniu.

Uwaga dotycząca interpretacji współczynników

W tym momencie można poczynić jedną uwagę na temat związku między oszacowanym równaniem regresji prostej, z tylko przejechanymi kilometrami jako zmienną niezależną, a równaniem uwzględniającym również liczbę dostaw jako drugą zmienną niezależną. Wartość b_1 nie jest taka sama w obu przypadkach. W prostej regresji liniowej interpretujemy b_1 jako oszacowanie zmiany y dla zmiany zmiennej niezależnej o jedną jednostkę. W analizie regresji wielorakiej interpretacja ta musi zostać nieco zmodyfikowana. Oznacza to, że w analizie regresji wielorakiej każdy współczynnik regresji jest interpretowany w sposób, który reprezentowałby szacunkową zmianę y odpowiadającą zmianie x_i o jedną jednostkę, gdy wszystkie inne zmienne niezależne są utrzymywane na stałym poziomie.

W przypadku Frigo Trucking obejmuje to dwie niezależne zmienne, b_1=0,0611 i b_2= 0,923.

DANE WYJŚCIOWE MINITAB DLA FRIGO TRUCKING Z DWIEMA ZMIENNYMI NIEZALEŻNYMI

Równanie regresji ma postać:
Czas=-0,869+0,0611 km + 0,923 Dostaw

Predyktor	Współczynnik	Współczynnik SE	T	p-value
Stała	-0,8687	0,9515	-0,91	0,392
Nachylenie (b_1)	0,061135	0,009888	6,18	0,000
Dostawy	0,9234	0,2211	4,18	0,004

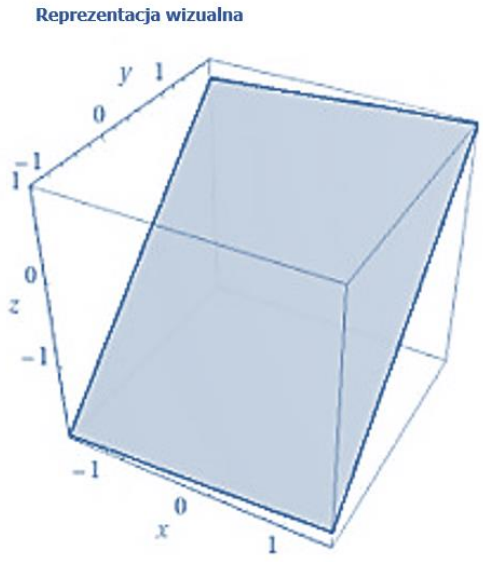
S=0,573142 R-Sq=90,4% R-SQ(adj)=87,6%

Analiza wariancji

Źrodło	DF	SS	MS	F	p-value
Regresja	2	21,601	10,8	32,88	0,000
Błąd szczątkowy	7	2,299	0,328		
Suma	9	23,900			

Rysunek 5.12 Wyniki dla Frigo Trucking z dwiema zmiennymi niezależnymi

Zatem 0,0611 godziny to szacunkowy oczekiwany wzrost czasu podróży odpowiadający wzrostowi o jeden km na przejechaną odległość, gdy liczba dostaw jest stała. Podobnie, ponieważ b_2 =0,923, szacowany oczekiwany wzrost czasu podróży odpowiadający wzrostowi o jedną dostawę, gdy liczba przejechanych kilometrów jest stała, wynosi 0,923 godziny.



Rysunek 5.13 Wizualna reprezentacja wyników dla przypadku Frigo Trucking

BIBLIOGRAFIA DO ROZDZIAŁU 5.

- Introductory Statistics. Bentham Science Publishers, Kahl, A. (Publish 2023). DOI:10.2174/9789815123135123010l
- Introductory Statistics 2e, Openstax, Rice University, Houston, Texas 77005, Jun 23, senior contributing authors: Barbara Illowsky and Susan dean, De anza college, Publish Date: Dec 13, 2023, (https://openstax.org/details/books/introductory-statistics-2e);
- Introductory Statistics 4th Edition, Susan Dean and Barbara Illowsky, Adapted by Riyanti Boyd & Natalia Casper (Published 2013 by OpenStax College) July 2021, (http://dept.clcillinois.edu/mth/oer/IntroductoryStatistics.pdf);
- Journal of the Royal Statistical Society 2024**,** A reputable journal publishing cutting-edge research and articles on various aspects of statistics, including theoretical advancements and practical applications. Recent issues have featured studies on sampling and hypothesis testing.
- Introductory Statistics 7th Edition, Prem S. Mann, eastern Connecticut state university with the help of Christopher Jay Lacke, Rowan university, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030-5774, 2011
- Introduction to statistics, made easy second edition, Prof. Dr. Hamid Al-Oklah Dr. Said Titi Mr. Tareq Alodat, March 2014
- Statistics for Business and Economics, Thirteenth Edition, David R. Anderson, Dennis J. Sweeney, Thomas A. Williams, Jeffrey D. Camm, James J. Cochran, 2017, 2015 Cengage Learning®
- Statistics for Business, First edition, Derek L Waller, 2008 Copyright © 2008, Derek L Waller, Published by Elsevier Inc. All rights reserved

6. Wprowadzenie do badań operacyjnych

Badania operacyjne (British English: operational research, U.S. Air Force Specialty Code: Operations Analysis), często skracane do skrótowca OR, to dyscyplina zajmująca się opracowywaniem i stosowaniem metod analitycznych w celu usprawnienia procesu podejmowania decyzji. Termin nauki o zarządzaniu jest czasami używany jako synonim.

Metody OR są wykorzystywane do analizy: wydajności zasobów, ich rozkroju lub mieszanki,, wąskich gardeł produkcji, czasów realizacji i harmonogramowania, wzorców popytu, zapasów, kosztów transportu i alokacji zasobów, asortymentu, niezawodności, przepustowości procesu, planowania trasy, itp. (Ueda, 2010).

Wykorzystując techniki z innych nauk matematycznych, takich jak modelowanie decyzji, statystyka i optymalizacja, badania operacyjne dochodzą do optymalnych lub prawie optymalnych rozwiązań problemów decyzyjnych. Ze względu na nacisk na praktyczne zastosowania, badania operacyjne pokrywają się z wieloma innymi dyscyplinami, w szczególności inżynierią przemysłową i logistyką, czyniąc je integralną częścią ich systemów zarządzania wiedzą (KMS).

6.1 Strategiczne planowanie logistyczne

Według (Robinson, 2004) badania operacyjne dotyczą głównie interakcji między zaplanowanymi systemami technicznymi i ludzkimi. Charakteryzują się one zmiennością, współzależnością między komponentami oraz złożonością strukturalną i behawioralną. Aby zarządzać tymi cechami, opracowano szeroki wachlarz metod, aby odpowiednio zająć się nimi jako całością. Są one wykorzystywane w planowaniu strategicznym do:

- umożliwienia złożonej symulacji, inaczej analizy typu „co jeśli?”,
- zarządzania złożonością przedsiębiorstwa (np. produktową): współzależność + zmienność + dynamika procesów,

- wiąże się z mniejszymi kosztami i mniejszą ingerencją w proces niż eksperymentowanie z rzeczywistym systemem,
- operacjonalizacji działań,
- poprawienia zrozumienia systemu,
- poprawy komunikacji między kierownictwem a ekspertami.

Strategiczne planowanie logistyczne (rysunek 6.1) obejmuje wszystkie działania, które należy wykonać na poziomie strategicznym, taktycznym i operacyjnym w celu zapewnienia kompleksowego zarządzania jakością (TQM) (Ciampa, 1992) i produkcji Just-in-time (JIT) (Britannica, 2023).

Rysunek 6.1 Strategiczne planowanie logistyczne

6.2 Six Sigma

W strategicznym planowaniu logistycznym model referencyjny SCOR (AIMS, 2021) pomaga firmom oceniać i doskonalić zarządzanie łańcuchem dostaw pod kątem niezawodności, spójności i wydajności. Rozpoznaje 6 głównych procesów biznesowych - Planowanie, Pozyskiwanie, Wytwarzanie, Dostarczanie, Zwrot i Udostępnianie.

Proces planowania SCOR obejmuje wszystkie działania związane z opracowywaniem planów zarządzania i doskonalenia łańcucha dostaw. Ciągłe wysiłki w celu osiągnięcia stabilnych i przewidywalnych wyników procesu poprzez redukcję zmienności procesu (Six-sigma) mają kluczowe znaczenie dla sukcesu biznesowego.

Wyjaśnienie DMAIC i DMADV

Procesy produkcyjne (pozyskiwanie, wytwarzanie, dostarczanie, udostępnianie, a także obsługa zwrotów) mają cechy, które można zdefiniować, zmierzyć, przeanalizować, ulepszyć i kontrolować. Dlatego też fazy te stanowią metodologię zarządzania procesami produkcyjnymi, w skrócie DMAIC.

Niektórzy praktycy połączyli ideę 6-sigma ze szczupłą produkcją, tworząc metodologię o nazwie Lean Six Sigma (Wheat & Mills & Carnell, 2003). Metodologia Lean Six Sigma obejmuje Lean Manufacturing (produkcję „dokładnie na czas"), która odnosi się do wydajności procesu, oraz Six-sigma, skupiającą się na redukcji zmienności i marnotrawstwa, jako uzupełniające się dyscypliny, które promują doskonałość biznesową i operacyjną.

Metodologia DMADV (ang. define, measure, analyze, design and verify), znana również jako DFSS (ang. „Design for Six Sigma"), jest zgodna z KBE (ang. „Knowledge Based Engineering"). Fazy metodologii DFSS (Chowdhury, 2002) są następujące:

1. Zdefiniowanie celów projektowych, które są zgodne z wymaganiami klientów i strategią przedsiębiorstwa.
2. Pomiar i identyfikacja cech krytycznych dla jakości (CTQ), pomiar możliwości produktu, zdolności procesu produkcyjnego i pomiar ryzyka.
3. Analiza w celu opracowania i zaprojektowania alternatywnych rozwiązań.
4. Zaprojektowanie ulepszonej alternatywy, najlepiej dostosowanej do analizy przeprowadzonej w poprzednim kroku.

5. Weryfikacja projektu, konfiguracja serii pilotażowych, wdrożenie procesu produkcyjnego i przekazanie go właścicielowi (właścicielom) procesu.

Projekty doskonalenia biznesowego Six Sigma (Tennant, 2001), zainspirowane cyklem W. Edwardsa Deminga „Plan-Do-Study-Act" (Tague, 2005), w zależności od ich charakteru, opierają się na jednej z wyżej wymienionych metodologii, z których każda składa się z pięciu faz:

1. DMAIC jest stosowany w projektach mających na celu usprawnienie istniejącego procesu biznesowego.
2. DMADV jest używany w projektach mających na celu tworzenie nowych produktów lub projektów procesów.

6.3 Business Intelligence (Analiza biznesowa)

Business Intelligence (BI) obejmuje wszystkie strategie i technologie wykorzystywane przez przedsiębiorstwa do analizy danych dotyczących przeszłych i bieżących informacji biznesowych (Tableau, 2023). Jest on wspierany przez systemy zarządzania wiedzą (KMS), które stanowią część systemów informacji logistycznej (LIS), umożliwiając ekspertom z różnych dziedzin doradzanie i zapewnianie wsparcia różnym szczeblom zarządzania.

Analityka biznesowa

Analityka biznesowa (BA) to proces oparty na BI, umożliwiający nowy wgląd w proces biznesowy i lepsze podejmowanie strategicznych decyzji na przyszłość. Wywodzi się z Data Mining (DM), czyli procesu znajdowania anomalii, wzorców i korelacji w większych zbiorach danych w celu przewidywania wyników.

Proces BA składa się z następujących elementów:

1. **Organizowanie danych**: przed analizą dane muszą być najpierw zebrane, uporządkowane i przefiltrowane, albo poprzez dobrowolne dane, albo zapisy transakcji.
2. **Data Mining**: eksploracja danych sortuje duże zbiory danych przy użyciu baz danych, statystyk i uczenia maszynowego w celu zidentyfikowania trendów i ustalenia relacji.

3. **Identyfikacji powiązań i sekwencji**: identyfikacja przewidywalnych działań, które są wykonywane w powiązaniu z innymi działaniami lub sekwencyjnie.
4. **Text Mining**: eksploruje i organizuje duże, nieustrukturyzowane zbiory danych tekstowych w celu analizy jakościowej i ilościowej.
5. **Prognozowania**: analizuje dane historyczne z określonego okresu w celu dokonania predykcji, czyli procesu wypracowywania prognozy przyszłych, ale też nieznanych zdarzeń lub zachowań.
6. **Analityki predykcyjnej**: predykcyjna analityka biznesowa wykorzystuje różnorodne techniki statystyczne do tworzenia modeli predykcyjnych, które wyodrębniają informacje ze zbiorów danych, identyfikują wzorce i zapewniają wynik predykcyjny dla szeregu wyników organizacyjnych.
7. **Optymalizacji**: po zidentyfikowaniu trendów i dokonaniu prognoz, firmy mogą zaangażować techniki symulacyjne w celu przetestowania najlepszych scenariuszy.
8. **Wizualizacji danych**: zapewnia wizualne reprezentacje, takie jak wykresy i grafy dla łatwej i szybkiej analizy danych.

Planowanie sprzedaży i operacji

Planowanie sprzedaży i operacji (SOP) to elastyczne narzędzie do prognozowania i planowania działań produkcyjnych. Etapy SOP:

1. Plan sprzedaży.
2. Plan produkcji.
3. Planowanie zdolności produkcyjnych.

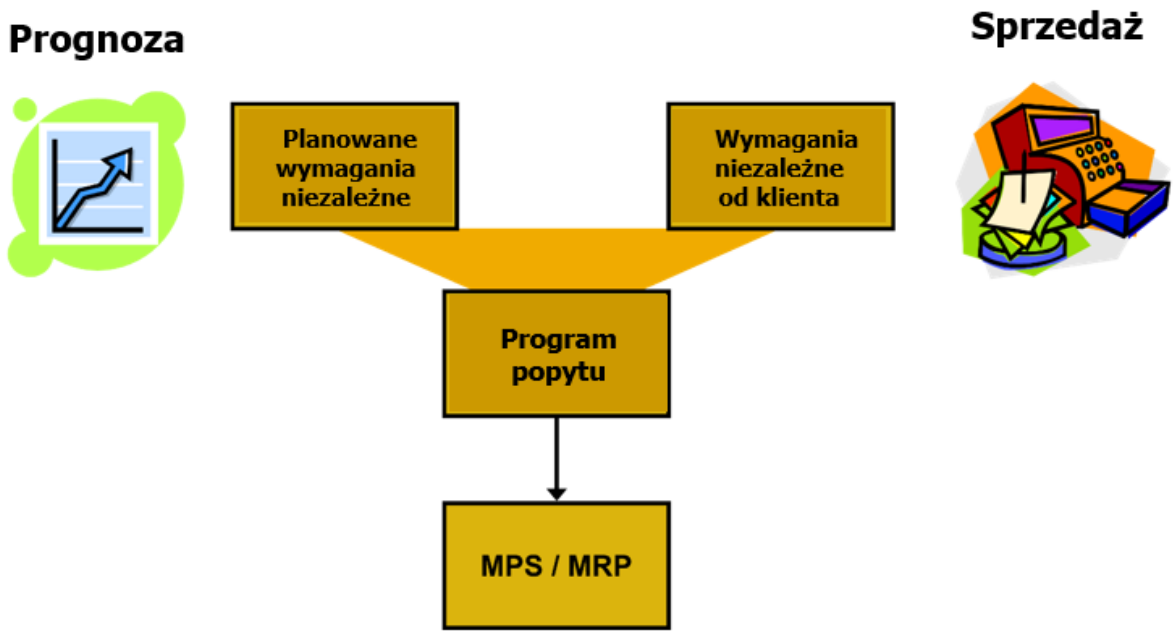
SOP działa na danych pochodzących z różnych źródeł informacji w całej firmie: Sprzedaż, Marketing, Produkcja, Księgowość, Zasoby Ludzkie i Zapotrzebowanie. Zazwyczaj są one dostarczane przez odpowiednie działy za pośrednictwem systemu planowania zasobów przedsiębiorstwa (ERP).

Podczas gdy SOP działa na poziomie strategicznym, ERP działa na poziomie taktycznym systemu informacji logistycznej firmy. Łączy je program zarządzania popytem, łączący strategiczne planowanie sprzedaży i operacji (SOP) ze szczegółowym planowaniem produkcji (Główny Harmonogram Produkcji / Planowanie Wymagań Materiałowych) na poziomie

operacyjnym. Tutaj do gry wkraczają wspomniane wcześniej harmonogramowanie i symulacja celem stworzenia wykonalnego i optymalnego planu produkcji.

Program zarządzania popytem (rysunek 6.2) składa się z dwóch rodzajów prognoz:

1. Planowanych niezależnych potrzeb (PIR) z prognozowanych wielkości sprzedaży w oparciu o działania marketingowe.
2. Wymagań niezależnych od klienta (CIR) z danych opartych na istniejących i planowanych zamówieniach sprzedaży.



Rysunek 6.2 Zarządzanie popytem

Przykład SOP

Firma handluje kilkoma rodzajami produktów w swojej sieci sprzedaży. Transakcje sprzedaży są przechowywane centralnie, aby móc monitorować stan zapasów, a także przeprowadzać analizy sprzedaży w celu zarządzania popytem. Są one rejestrowane w formacie CSV (Comma Separated Values), który jest łatwy do przetworzenia przez system ERP firmy, jak również przez dział analityczny, który korzysta z arkuszy kalkulacyjnych.

W analizie sprzedaży transakcje sprzedaży są początkowo filtrowane w celu ustalenia, czy są kompletne i poprawnie sformatowane. Dopiero wtedy są one gotowe do oceny statystycznej,

ponieważ w przeciwnym razie wszelkie brakujące lub źle sformatowane dane mogą spowodować błędną interpretację wyników.

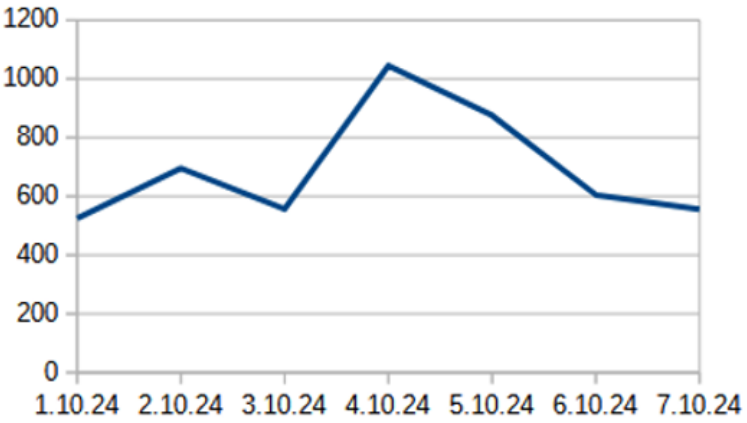
Data	ID Sprzedawcy	ID Klienta	ID Transakcji	ID Produktu	Cena Produktu
1.10.24	1	12	1	101	195,00 €
1.10.24	1	12	1	102	45,00 €
1.10.24	1	12	1	103	35,00 €
1.10.24	2	14	2	104	55,00 €
1.10.24	2	14	3	101	195,00 €
2.10.24	3	15	4	105	85,00 €
2.10.24	3	15	4	101	195,00 €
2.10.24	3	15	4	103	35,00 €
2.10.24	3	16	5	104	55,00 €
2.10.24	1	17	6	101	195,00 €
2.10.24	1	17	6	102	45,00 €
2.10.24	1	17	6	105	85,00 €
3.10.24	2	18	7	106	35,00 €
3.10.24	2	18	7	107	65,00 €
3.10.24	2	18	7	108	86,00 €
3.10.24	4	19	8	105	85,00 €
3.10.24	4	19	8	101	195,00 €
3.10.24	4	19	8	103	35,00 €
3.10.24	4	19	9	104	55,00 €
4.10.24	5	20	10	105	110,00 €
4.10.24	5	20	10	106	125,00 €
4.10.24	5	20	10	104	55,00 €
4.10.24	5	20	10	101	195,00 €
4.10.24	1	21	11	102	45,00 €
4.10.24	1	21	11	105	85,00 €
4.10.24	1	21	12	106	35,00 €
4.10.24	3	12	13	103	35,00 €
4.10.24	3	12	13	104	55,00 €
4.10.24	3	12	13	105	110,00 €
4.10.24	3	12	13	101	195,00 €
5.10.24	1	22	14	107	35,00 €
5.10.24	1	22	14	108	25,00 €

Rysunek 6.3 Tygodniowe dane dotyczące sprzedaży

Zwykle dzięki analityce możliwe jest uzyskanie różnych sensownych wglądów w zebrane „surowe dane” (rysunek 6.3). Można to osiągnąć za pomocą tabel przestawnych, które umożliwiają grupowanie danych według wybranych atrybutów i przeprowadzanie analiz statystycznych. Zasadniczo każdy atrybut (kolumnę) danych wejściowych można uznać za tabelę przestawną. Dlatego często mówimy o „kostce danych” o wielu wymiarach. Ponieważ nie możemy graficznie przedstawić więcej niż dwóch lub trzech wymiarów, najprostsze, ale zwykle najbardziej przydatne reprezentacje danych wejściowych są tworzone

z dwóch lub trzech atrybutów przestawnych. W oparciu o podaną próbkę danych, w dalszej części przedstawiono kilka przykładów tabel przestawnych.

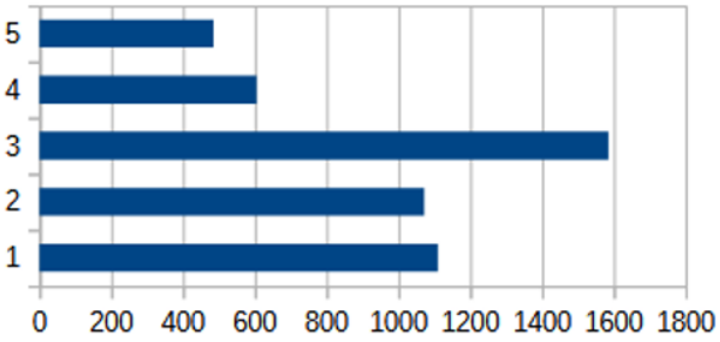
Data	Wartość sprzedaży produktów
1.10.24	525,00 €
2.10.24	695,00 €
3.10.24	556,00 €
4.10.24	1.045,00 €
5.10.24	875,00 €
6.10.24	605,00 €
7.10.24	556,00 €
Razem:	**4.857,00 €**



Rysunek 6.4 Statystyki sprzedaży według tygodnia

Statystyki sprzedaży według dnia tygodnia lub miesiąca (rysunek 6.4.) umożliwiają wgląd w trendy sezonowe.

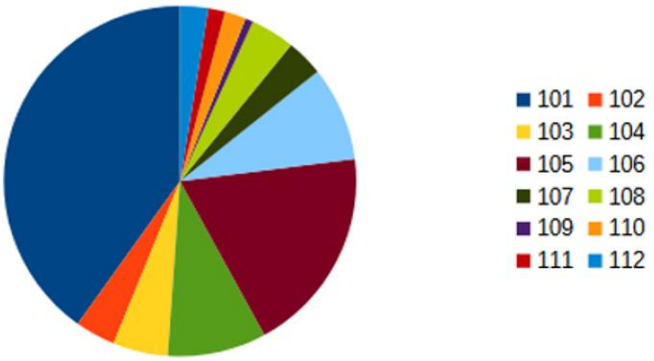
ID Sprzedawcy	Wartość sprzedaży produktów
1	1.110,00 €
2	1.072,00 €
3	1.585,00 €
4	605,00 €
5	485,00 €
Razem:	**4.857,00 €**



Rysunek 6.5 Statystyki sprzedaży według biura sprzedaży.

Statystyki sprzedaży według biura sprzedaży (rysunek 6.5) określają, które biura są najbardziej oblegane i/lub generują największe przychody.

ID Produktu	Wartość sprzedaży produktów
101	1.950,00 €
102	180,00 €
103	245,00 €
104	440,00 €
105	925,00 €
106	425,00 €
107	165,00 €
108	197,00 €
109	35,00 €
110	95,00 €
111	75,00 €
112	125,00 €
Razem:	**4.857,00 €**



Rysunek 6.6 Statystyki sprzedaży według produktów

Statystyki sprzedaży według produktów (rysunek 6.6) określają produkty, które są najbardziej poszukiwane lub stanowią znaczący udział w portfolio.

Tabela 6.1 Przegląd transakcji

Data	ID Sprzedawcy	ID Klienta	ID Transakcji	ID Produktu	Wartość sprzedaży
1.10.24	1	12	1	101	195,00 €
				102	45,00 €
				103	35,00 €
	2	14	2	104	55,00 €
			3	101	195,00 €
2.10.24	1	17	6	101	195,00 €
				102	45,00 €
				105	85,00 €
	3	15	4	101	195,00 €
				103	35,00 €
				105	85,00 €
		16	5	104	55,00 €
3.10.24	2	18	7	106	35,00 €
				107	65,00 €
				108	86,00 €
	4	19	8	101	195,00 €
				103	35,00 €
				105	85,00 €
			9	104	55,00 €

Przegląd transakcji (tabela 6.1) według dnia, biura sprzedaży, transakcji i kupującego oferuje uporządkowany wgląd w dane, który jest przydatny przy rozwiązywaniu zapytań dotyczących określonego produktu lub transakcji sprzedaży.

6.4 Systemy wspomagania decyzji

System wspomagania decyzji (DSS) to interaktywny system, który pomaga decydentom wykorzystywać dane i modele do rozwiązywania nieustrukturyzowanych lub częściowo ustrukturyzowanych problemów. System ekspercki (ES) to program aplikacyjny lub środowisko, które skutecznie wspiera rozwiązywanie problemów w wyspecjalizowanym obszarze problemowym, wymagającym wiedzy i umiejętności eksperckich.

Oprócz metod statystycznych i symulacyjnych stosowanych w BA, DSS często obejmuje modele umożliwiające podejmowanie decyzji w oparciu o zestaw charakterystycznych kryteriów. Modele te mogą być proste (np. tabele decyzyjne, drzewa itp.) prowadzące do jednego poprawnego rozwiązania. Z drugiej strony, gdy mamy do czynienia z wieloma (potencjalnie sprzecznymi) kryteriami i wieloma rozwiązaniami, konieczny jest bardziej rozbudowany model. Odpowiednim modelem często stosowanym w życiu zawodowym i osobistym jest wielokryterialny model decyzyjny.

Wielokryterialne podejmowanie decyzji

Wielokryterialne podejmowanie decyzji (MCDM) lub wielokryterialna analiza decyzyjna (MCDA) to subdyscyplina OR, która wyraźnie ocenia wiele (potencjalnie) sprzecznych kryteriów w procesie podejmowania decyzji w odniesieniu do zestawu potencjalnych rozwiązań lub wariantów.

Model decyzyjny MCDM składa się z:

- kryteriów – są to parametry wariantów wejściowych, krytyczne dla naszego projektu,
- wag – jest to względne znaczenie wybranych kryteriów,
- funkcji użyteczności – jest to funkcja, która łączy ważone parametry wariantów w wartość zgodności,
- dane – dane reprezentujące warianty; dane wejściowe do naszego modelu MCDM.

Typy danych w MCDM mogą być następujące:

- ilościowe – reprezentujące wartości, które mogą być porównywane jako takie,
- jakościowe – reprezentujące względne wartości porównawcze (np. wysoka, komfortowa, niska temperatura itp.), które należy określić ilościowo, aby uzyskać unikalne wartości,
- binarne – reprezentuje kryteria binarne; własność jest spełniona (1) lub nie (0).

Procedura wielokryterialnego podejmowania decyzji:

1. Reprezentacja wariantów (V) przez ich charakterystyczne parametry (P): $\{V_i (P_{i,1}; P_{i,2}; \dots P_{i,n}); i=1...m\}$.
2. Normalizacja parametrów poprzez obliczenie względnego lokalnego stopnia $p_{i,j}$ dla każdego $P_{i,j}$ (j=1...n), w odniesieniu do j-tego parametru maksymalnej wartości $P_{i,j}$ ze wszystkich *i* próbek:
   a) $p_{i,j}=P_{i,j}/\max \{P_{i,j}\}$ jeśli większa wartość $P_{i,j}$ jest bardziej korzystna,
   b) $p_{i,j}=1 - P_{i,j}/\max \{P_{i,j}\}$ jeśli mniejsza wartość $P_{i,j}$ jest bardziej korzystna.
3. Oceny są ważone zgodnie z preferencjami: $x_{i,j}=p_{i,j}*U_j$ dla każdego j=1...n, przez wagi U_j, które muszą sumować się do 1, tj. 100%.
4. Ważone oceny wszystkich wariantów są sumowane: $X_i=\sum x_{i,j},j$ dla każdego i=1...m w celu uzyskania ocen złożonych zgodnie z naszą funkcją użyteczności.
5. Wybierany jest najlepszy wariant $Y=\max \{X_i\}$.

Przykład MCDM

Wybierając nowy sprzęt w firmie, często musimy podejmować decyzje wielokryterialne. Rozważmy przykład wyboru najbardziej opłacalnej (poniżej 300 EUR) platformy mobilnej z systemem operacyjnym Android dla naszej firmy. W tabeli 6.2 wymieniono modele, które zostały wybrane na podstawie zapytania wśród pracowników w celu zawężenia naszego asortymentu. Dla każdego z nich podano parametry, które zostały wybrane jako najbardziej istotne. W dalszej części wybrane dane parametrów są normalizowane w celu uzyskania porównywalnych wartości, ważone w celu podkreślenia wartości, które są bardziej znaczące i sumowane w celu uzyskania ocen dla wybranych okazów.

Tabela 6.2 MCDM dla ekonomicznej platformy mobilnej Android

PARAMETRY									
	Cena (€)	Stopień *	Wydajność			Właściwości			Aparat
Model		(1-10)	prędk.proc. (GHz)	RAM (GB)	pamięć wew.. (GB)	waga (g)	rozmiar (mm³)	pojem.bat. (mAh)	(MP)
Honor Magic Lite 5	263 €	2	2,2	6	128	175	94344	5100	64
Honor X7a	210 €	2,5	2,3	4	128	196	106442	6000	50
Samsung A34	297 €	1,8	2,6	8	256	199	103300	5000	48
Redmi Note 12 Pro	240 €	1,9	2,6	6	128	187	99104	5000	50
Redmi Note 12 S	224 €	1,7	2,05	8	256	176	95715	5000	108

*Vir: www.testberichte.de

WAGI PARAMETRÓW									
	Cena (€)	Stopień *	Wydajność			Właściwości			Aparat
			prędk.proc. (GHz)	RAM (GB)	pamięć wew.. (GB)	waga (g)	rozmiar (mm³)	pojem.bat. (mAh)	(MP)
			10 %	5 %	5 %	10 %	10 %	10 %	
Uteż	20 %	10 %	20 %			30 %			20 %

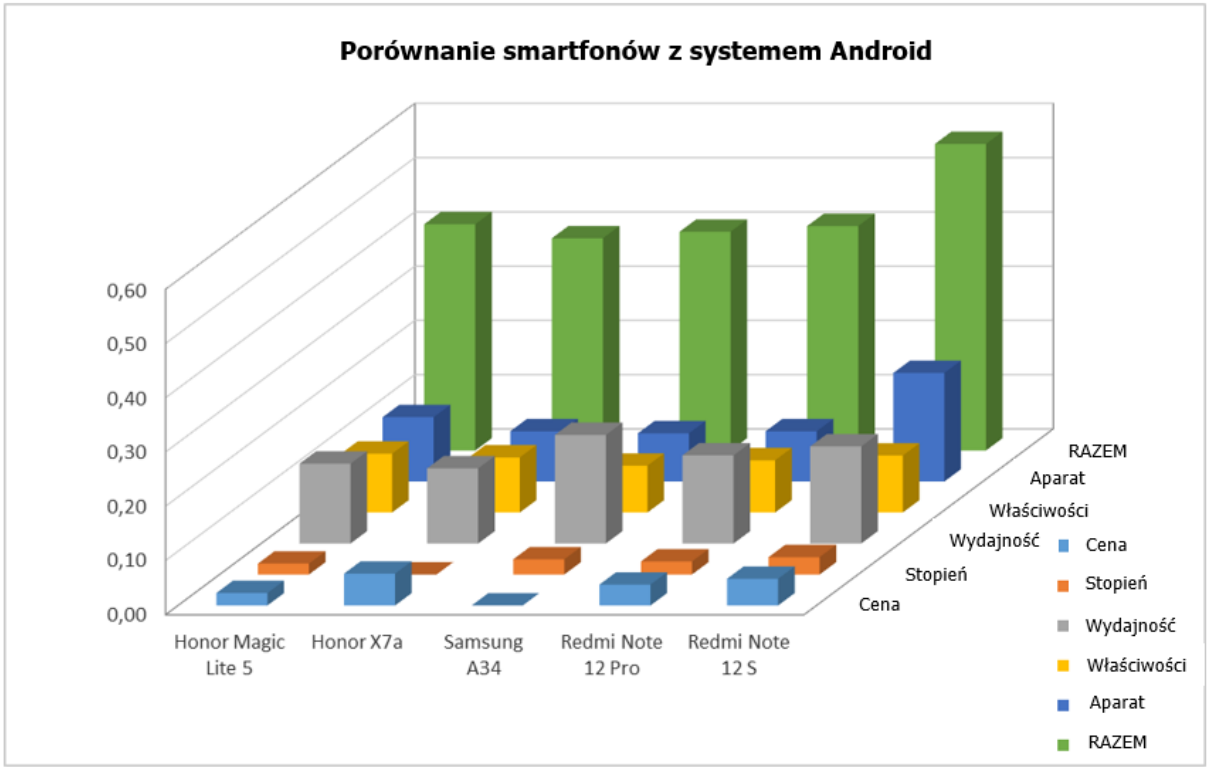
PARAMETRY ZNORMALIZOWANE									
	Cena (€)	Stopień *	Wydajność			Właściwości			Aparat
Model		(1-10)	prędk.proc. (GHz)	RAM (GB)	pamięć wew.. (GB)	waga (g)	rozmiar (mm³)	pojem.bat. (mAh)	(MP)
Honor Magic Lite 5	0,11	0,20	0,85	0,75	0,50	0,12	0,11	0,85	0,59
Honor X7a	0,29	0,00	0,88	0,50	0,50	0,02	0,00	1,00	0,46
Samsung A34	0,00	0,28	1,00	1,00	1,00	0,00	0,03	0,83	0,44
Redmi Note 12 Pro	0,19	0,24	1,00	0,75	0,50	0,06	0,07	0,83	0,46
Redmi Note 12 S	0,25	0,32	0,79	1,00	1,00	0,12	0,10	0,83	1,00

OCENA PARAMETRÓW KOŃCOWYCH									
	Cena (€)	Stopień *	Wydajność			Właściwości			Aparat
Model		(1-10)	prędk.proc. (GHz)	RAM (GB)	pamięć wew.. (GB)	waga (g)	rozmiar (mm³)	pojem.bat. (mAh)	(MP)
Honor Magic Lite 5	0,02	0,02	0,08	0,04	0,03	0,01	0,01	0,09	0,12
Honor X7a	0,06	0,00	0,09	0,03	0,03	0,00	0,00	0,10	0,09
Samsung A34	0,00	0,03	0,10	0,05	0,05	0,00	0,00	0,08	0,09
Redmi Note 12 Pro	0,04	0,02	0,10	0,04	0,03	0,01	0,01	0,08	0,09
Redmi Note 12 S	0,05	0,03	0,08	0,05	0,05	0,01	0,01	0,08	0,20

Końcowym wynikiem naszej analizy jest tabela podsumowująca (tabela 6.3) i ewentualnie wykres (rysunek 6.7), który podsumowuje proces decyzyjny i przedstawia najlepszy wybór, a także mocne i słabe strony poszczególnych wariantów. Często jednostka z najlepszą oceną nie jest tą, która wypadła najlepiej we wszystkich kategoriach, ale tą, która średnio najlepiej pasuje do naszych kryteriów wyboru, a także ważenia parametrów. Jest to również główna siła metody MCDM, ponieważ wybór według dowolnego pojedynczego parametru może wprowadzić nas w błąd.

Tabela 6.3 Podsumowanie MCDM dla naszego przykładu wyboru platformy mobilnej

OCENA PARAMETRÓW KOŃCOWYCH						
Model	Cena	Stopień	Wydajność	Właściwości	Aparat	RAZEM:
Honor Magic Lite 5	0,02	0,02	0,15	0,11	0,12	0,42
Honor X7a	0,06	0,00	0,14	0,10	0,09	0,39
Samsung A34	0,00	0,03	0,20	0,09	0,09	0,40
Redmi Note 12 Pro	0,04	0,02	0,16	0,10	0,09	0,41
Redmi Note 12 S	0,05	0,03	0,18	0,10	0,20	0,56



Rysunek 6.7 Wybór najlepszej platformy mobilnej

Najlepszy wybór: 0,56 (Redmi Note 12 S).

6.5 Inżynieria oparta na wiedzy

Inżynieria oparta na wiedzy (Knowledge Based Engineering - KBE) to metodologia inżynieryjna służąca do systematycznej integracji wiedzy inżynieryjnej z systemem projektowania (Andersson & Larsson & Ola, 2011).

Cykl życia zarządzania doświadczeniem

Potrzeba przechwytywania, zarządzania i wykorzystywania wiedzy projektowej oraz automatyzacji procesów unikalnych dla doświadczenia producenta w zakresie rozwoju produktu doprowadziła do rozwoju technologii inżynierii opartej na wiedzy (KBE) (Prasad, 2005). KBE ma na celu wzbogacenie wiedzy instytucjonalnej poprzez zarządzanie doświadczeniem. Fazy zarządzania doświadczeniem, według (Andersson & Larsson & Ola, 2011) są następujące:

1. Identyfikacja: wybrano niezgodność z pożądanym stanem, która pojawia się w procesie produkcyjnym z powodu źle zdefiniowanego produktu lub procesu.
2. Przechwytywanie: doświadczenie wraz z jego właściwościami jest przechwytywane.
3. Analizowanie: przeprowadzana jest analiza przyczyn źródłowych przechwyconego doświadczenia w celu zidentyfikowania odpowiedniej strategii naprawczej i jej ponownego wykorzystania w celu zapobiegania powtarzającym się anomaliom.
4. Przechowywanie: spostrzeżenia z analizy są archiwizowane wraz z doświadczeniem.
5. Szukanie i Pobieranie: doświadczenie jest wyszukiwane i pobierane.
6. Użycie: element doświadczenia jest używany.
7. Ponowne wykorzystanie: kończy cykl zarządzania wiedzą i rozpoczyna nowy.

Inżynieria oparta na wiedzy jest ogólnie uzupełniana przez inne dyscypliny, których bliższe rozważenie wykracza poza zakres tego rozdziału:

- Komputerowe wspomaganie zarządzania projektami (PS),
- Projektowanie wspomagane komputerowo (CAD), produkcja (CAM) i robotyka (CIM),
- Modelowanie i analiza symulacji komputerowej (SMA),
- Wspomagane komputerowo szczegółowe planowanie produkcji (MPS / MRP).

6.6 Wnioski

Jak przedstawiono w tym rozdziale, główne zastosowania BI w zarządzaniu przedsiębiorstwem odnoszą się do analityki biznesowej (BA) i systemów wspomagania decyzji (DSS). Są one powszechnie określane jako badania operacyjne (OR). Oprócz podstawowych technik BI, opisanych w tym rozdziale, w rozdziałach 3 i 4 dotyczących zarządzania danymi oraz modelowania i analizy symulacyjnej (SMA) podano kilka dodatkowych rozważań na temat gromadzenia danych, manipulacji i prezentacji, które również wspierają podejmowanie decyzji. Podsumowując, aplikacje BI można znaleźć w systemach zarządzania wiedzą (KMS), obejmujących:

- systemy wspomagania decyzji (DSS),
- analitykę biznesową (BA) jako rozszerzenie Data Mining (DM) i
- inżynierię opartą na wiedzy (KBE) jako unowocześnienie inżynierii wspomaganej komputerowo (CAE).

W oparciu o wyniki BA, DSS i SMA opracowywane są doświadczenia, które poszerzają wiedzę instytucjonalną i stanowią systemy eksperckie oparte na wiedzy (KBS). Jak wykazano w (Gumzej i in., 2023), mogą one być wykorzystane przez KBE do wprowadzenia zasady „wyciągniętych wniosków" do zarządzania usprawnieniami przedsiębiorstwa poprzez strategiczne planowanie logistyczne.

BIBLIOGRAFIA DO ROZDZIAŁU 6.

- AIMS (2021). SCOR - Supply Chain Operations Model. [Dostępne: https://aims.education/study-online/supply-chain-operations-reference-model-scor/, dostęp June 20, 2023]
- Ciampa, D. (1992). Total Quality: A User's Guide for Implementation, Addison-Wesley.
- Britannica, T. Editors of Encyclopaedia (2023). Just-in-time manufacturing, Encyclopedia Britannica. [Dostępne: https://www.britannica.com/topic/just-in-time-manufacturing, dostęp June 20, 2023]
- Tennant, G. (2001). SIX SIGMA: SPC and TQM in Manufacturing and Services, Gower Publishing, Ltd.
- Wheat B. & Mills C. & Carnell M. (2003). Leaning into Six Sigma: a parable of the journey to Six Sigma and a lean enterprise, McGraw-Hill.
- Tague, N.R. (2005). Plan–Do–Study–Act cycle, The quality toolbox (2nd ed.), ASQ Quality Press, pp. 390–392.
- Chowdhury, S. (2002). Design for Six Sigma: The revolutionary process for achieving extraordinary profits, Prentice Hall,
- Ueda, M. (2010). How to Market OR/MS Decision Support. International Journal of Applied Logistics (IJAL), 1(2), 23-36.

- Tableau (2023). Comparing Business Intelligence, Business Analytics and Data Analytics. [Dostępne: https://www.tableau.com/learn/articles/business-intelligence/bi-business-analytics, Dostęp June 20, 2023]
- Prasad, B. (2005). Knowledge Technology, What Distinguishes KBE From Automation. COE NewsNet – June 2005. [Dostępne: https://web.archive.org/web/20120324223130/http://legacy.coe.org/newsnet/Jun05/knowledge.cfm, dostęp June 20, 2023]
- Robinson, S. (2004). Simulation: The Practice of Model Development and Use, Wiley.
- Gumzej, R., Kramberger, T., Dujak, D. (2023). A knowledge base for strategic logistics planning. In: Dujak, Davor (edt.). Proceedings of the 23rd International Scientific Conference Business Logistics in Modern Management: October 5-6, 2023, Osijek, Croatia. Osijek: Josip Juraj Strossmayer University of Osijek, Faculty of Economics and Business, pp. 317-330, illustr. Business logistics in modern management (Online). ISSN 1849-6148. https://blmm-conference.com/past-issues/.
- Andersson, P. & Larsson, T. & Ola, I. (2011). A case study of how knowledge based engineering tools support experience re-use. In Research into Design – Supporting Sustainable Product Development, Chakrabarti, A. (Edt.), Research Publishing, Indian Institute of Science, Bangalore, India.

7.Przetwarzanie danych statystycznych SPSS

Do tej pory zdobyłeś już podstawową wiedzę na temat statystyki, manipulacji danymi, tworzenia symulacji, modelowania i analizy w ramach logistycznych łańcuchów dostaw, a także prostych metod regresji liniowej. Podczas gdy statystyka oferuje różnorodny zakres

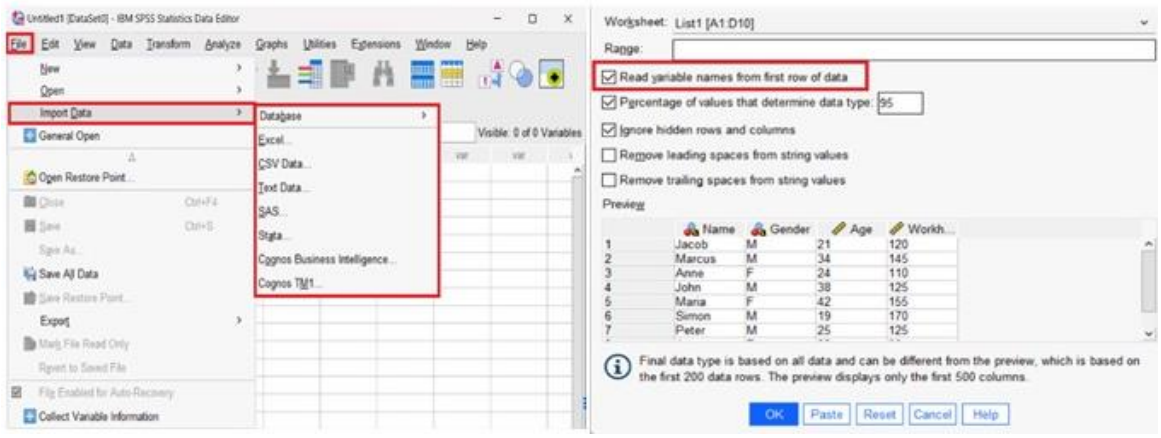
modeli i technik w celu zwiększenia wysiłków optymalizacyjnych, przeprowadzania analiz i identyfikowania potencjalnych ulepszeń, być może zauważyłeś, że wraz ze wzrostem złożoności analizowanych danych i obliczeń, tradycyjne podejścia mogą stać się coraz bardziej skomplikowane i trudne do obliczenia. Wraz ze wzrostem złożoności danych i obliczeń, konwencjonalne metody mogą stać się przestarzałe, a w niektórych przypadkach mogą zagrozić wiarygodności wyników. Aby temu zaradzić, statystyka wykorzystuje różne programy, które automatyzują analizę i interpretację zebranych danych, zapewniając jednocześnie wiele modeli i funkcji zapewniających wiarygodne wyniki. Jednym z takich programów jest IBM SPSS, który będzie kluczowym narzędziem w tym rozdziale. W tym rozdziale przedstawione zostanie zwięzłe wprowadzenie do podstawowego wykorzystania oprogramowania SPSS, badając jego funkcje i praktyczne zastosowania. Po wstępnym wprowadzeniu nastąpi praktyczne zastosowanie programu poprzez cztery podstawowe testy do obliczania wyników: test T, korelacje, Chi - kwadrat i ANOVA. Aby ułatwić naukę, przedstawione zostaną proste problemy i ich rozwiązania, które pomogą w zapoznaniu się z tymi testami.

7.1 Podstawy programu IBM SPSS

Być może masz już pewne doświadczenie z oprogramowaniem SPSS. Jeśli jednak nie, to pozwól nam zaoferować krótkie wprowadzenie. SPSS, podobnie jak jego bardziej znany odpowiednik, Excel, ułatwia manipulowanie danymi, ich analizę i wizualizację. Niemniej jednak, w przeciwieństwie do Excela, który czasami może być pracochłonny i skomplikowany w programowaniu funkcji, SPSS oferuje przyjazny dla użytkownika interfejs do analizy statystycznej (IBM, 2021). Oferuje różnorodne funkcje i metodologie do efektywnej obsługi dostarczonych danych. Chociaż oprogramowanie SPSS wyróżnia się szerokimi możliwościami

analizy statystycznej, poruszanie się po danych i konfigurowanie ustawień początkowych do analizy może czasami stanowić wyzwanie (IBM, 2021). W związku z tym zagłębimy się w podstawy importu danych i przygotowania danych do późniejszych testów statystycznych.

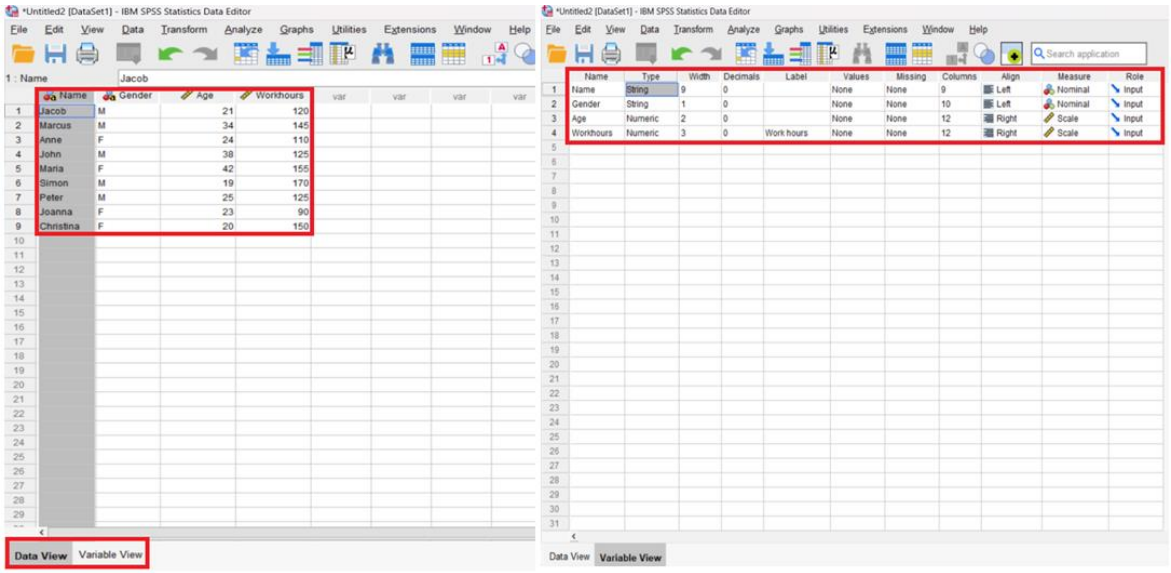
Biorąc pod uwagę powszechne wykorzystanie programu Excel do obsługi danych numerycznych, dane mogą być uzyskane lub przygotowane w arkuszu kalkulacyjnym Excel. Na szczęście SPSS może importować dane z różnych formatów plików do swoich arkuszy kalkulacyjnych. Po przygotowaniu gotowych arkuszy kalkulacyjnych Excel, otwórz oprogramowanie SPSS. Na ekranie początkowym przejdź do zakładki „File" (*Plik*) i wybierz „Import Data" (*Importuj dane*). W kolejnym oknie możesz wybrać format danych, który chcesz zaimportować (patrz rysunek 7.1). Po wykonaniu tego kroku należy zlokalizować przygotowany plik, wybrać go i przejść do następnego okna. Okno to wyświetli monit o skonfigurowanie dodatkowych ustawień. Jeśli nazwy kolumn zostały już uwzględnione w pierwszym wierszu danych, wybierz opcję „Read variable names from first row of data" (*pl. Odczytaj nazwy zmiennych z pierwszego wiersza danych*) (patrz rysunek 7.1), a następnie kliknij przycisk „Finish" (*Zakończ*), aby dane pojawiły się w arkuszu kalkulacyjnym (IBM, 2021).



Rysunek 7.1 Ustawienia importowania danych w programie SPSS.

Teraz, gdy mamy już dane w naszym arkuszu kalkulacyjnym, zauważysz wyraźną różnicę w ich prezentacji w porównaniu do Excela. SPSS kategoryzuje dane na dwa podstawowe typy, każdy z dwoma dodatkowymi podtypami. Jak pokazano na rysunku 7.2, dane można sklasyfikować jako numeryczne lub kategoryczne. Dane numeryczne składają się z liczb i można je

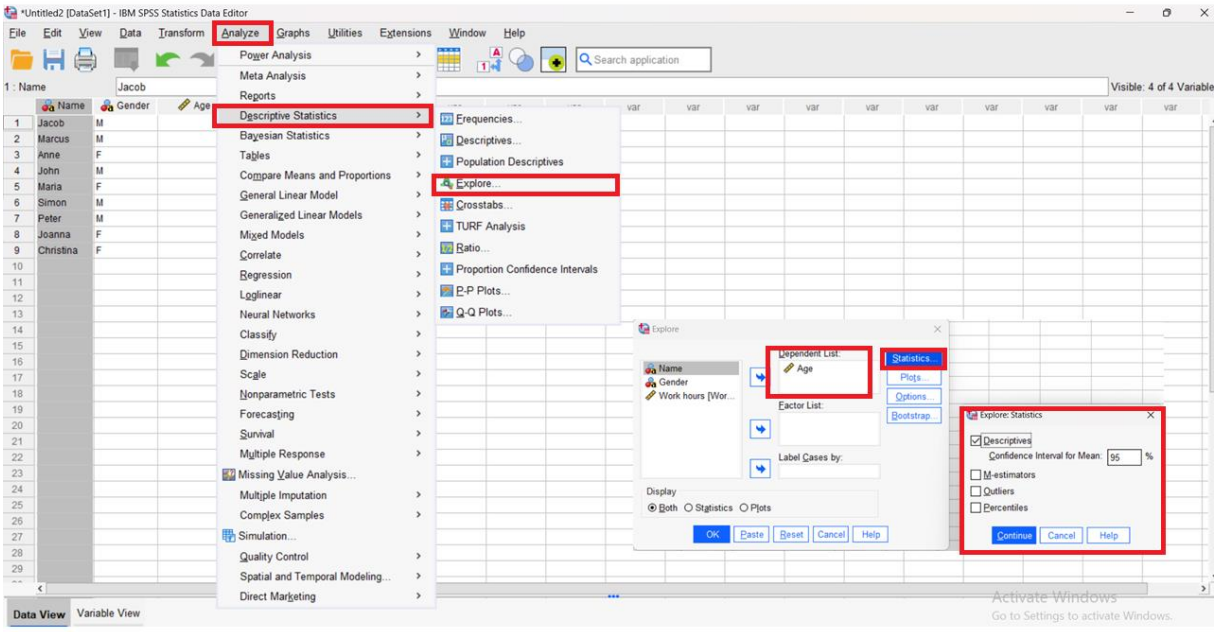
sklasyfikować jako dyskretne (ze skończonymi opcjami) lub ciągłe (oferujące nieskończone opcje). Z drugiej strony, dane kategoryczne obejmują słowa i mogą być dalej rozróżniane jako porządkowe (posiadające hierarchię) lub nominalne (bez hierarchii). W zależności od charakteru danych może być konieczne skonfigurowanie zmiennych w celu dostosowania ich do pożądanej analizy. W większości przypadków SPSS automatycznie odpowiednio skategoryzuje zmienne. Załóżmy, że chcesz dokonać dalszej manipulacji typami danych. W takim przypadku można uzyskać dostęp do opcji „View" (*Widok*) i w sekcji „Variable view" (*Widok zmiennej*) dostosować informacje o zmiennej, takie jak Name (*Nazwa*), Type (*Typ*), Width (*Szerokość*), Measure (*Miara*) i inne (IBM, 2021).



Rysunek 7.2 Okna widoku danych i zmiennych.

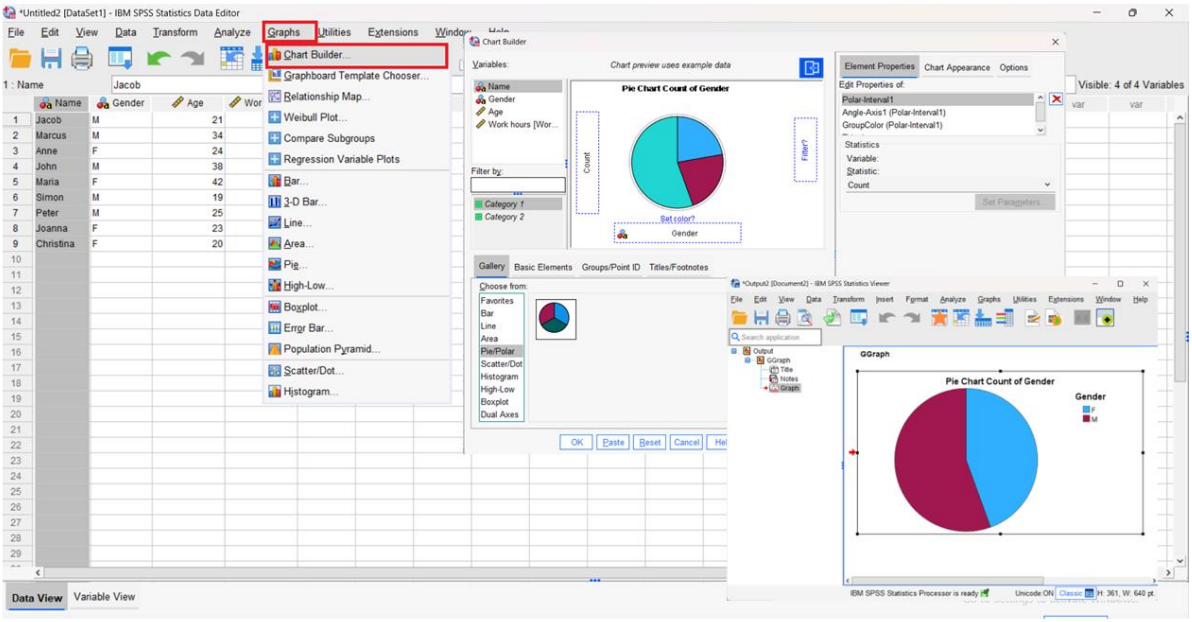
Po prawidłowym skonfigurowaniu danych można je eksplorować w SPSS. SPSS umożliwia użytkownikom wykonywanie podstawowych analiz statystycznych bez polegania na predefiniowanych funkcjach. Na ekranie początkowym (patrz rysunek 7.3), przejdź do „Analyze" (*pl. Analizuj*), a potem „Descriptive Statistics" (*pl. Statystyki opisowe*), a następnie wybierz „Explore" (*Eksploruj*). W sekcji „Explore" (*Eksploruj*) można znaleźć różne opcje w zależności od charakterystyki dostarczonych danych. W tym trybie SPSS dostarczy informacje o danych w postaci „statystyk opisowych". Chociaż jest to cenne dla wstępnej analizy danych, oferuje jedynie podstawowe spostrzeżenia i nie zagłębia się w bardziej szczegółową analizę statystyczną, która zostanie omówiona w kolejnych

rozdziałach. Zanim przejdziemy dalej, zbadamy również inną funkcję w SPSS - wizualizację wykresów (IBM, 2021).



Rysunek 7.3 Ustawienia statystyk opisowych

SPSS oferuje szereg opcji wizualizacji danych, w tym histogramy, wykresy pudełkowe, wykresy słupkowe, wykresy punktowe, wykresy liniowe, wykresy kołowe i inne. Do tego momentu powinieneś mieć podstawową wiedzę na temat tego, co reprezentuje każdy typ wykresu i jak interpretować przedstawione na nich wyniki. Dlatego też skupimy się na sposobie tworzenia tych wykresów w oprogramowaniu SPSS. Aby utworzyć wykresy, wybierz zakładkę „Graphs" (*Wykresy*) na ekranie początkowym, a następnie "Chart Builder" (*Kreator wykresów*). W nowym oknie możesz wybrać typ wykresu, który chcesz utworzyć i wybrać zmienne, które mają zostać uwzględnione. Po wybraniu opcji "Finish" (*Zakończ*) pojawi się nowe okno z wynikami zwizualizowanymi w wybranym formacie wykresu. W tym nowym oknie można aktywnie wchodzić w interakcje z wykresem, umożliwiając modyfikowanie kolorów i czcionek zmiennych, eksplorowanie rozkładów zmiennych na wykresie i nie tylko (IBM, 2021).



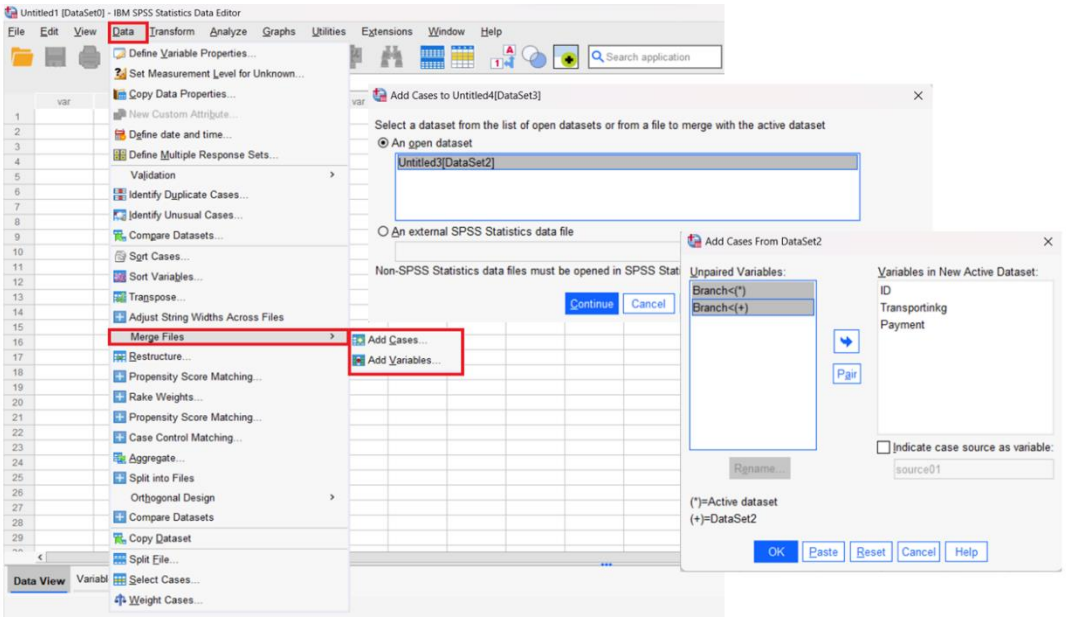
Rysunek 7.4 Ustawienia kreatora wykresów w programie SPSS

Do tego momentu zostały omówione trzy z czterech zasad „Eksploracji danych", które obejmują przeglądanie danych (eksploracja danych surowych), identyfikację danych (określanie typów danych) oraz, do pewnego stopnia, tworzenie wykresów i opisywanie danych za pomocą statystyk opisowych. Ostatnią zasadą jest „Formułowanie pytań", gdzie zadajemy sobie pytanie, co chcemy osiągnąć poprzez analizę danych i odpowiednio konfigurujemy wykresy i statystyki opisowe, aby uzyskać odpowiedzi na nasze konkretne pytania. Na przykład w naszym obecnym przykładzie pytanie może brzmieć: „Czy w analizowanej populacji przeważają kobiety?". Wykorzystując zarówno wykresy, jak i statystyki opisowe, możemy stwierdzić, że nasza populacja składa się głównie z mężczyzn. Podczas formułowania pytań należy zawsze brać pod uwagę dostępne dane izidentyfikowane zmienne (Garth, 2008). Na tym kończy się pierwsza część analizy danych SPSS, a teraz przejdziemy do przeprowadzania testów.

7.2 Zarządzanie danymi

Podczas pracy z kluczowymi danymi w oprogramowaniu SPSS istotne staje się zrozumienie technik manipulowania informacjami w poszczególnych aktywnych zbiorach danych. SPSS zapewnia funkcje ułatwiające manipulowanie istniejącymi danymi zawartymi w aktywnych

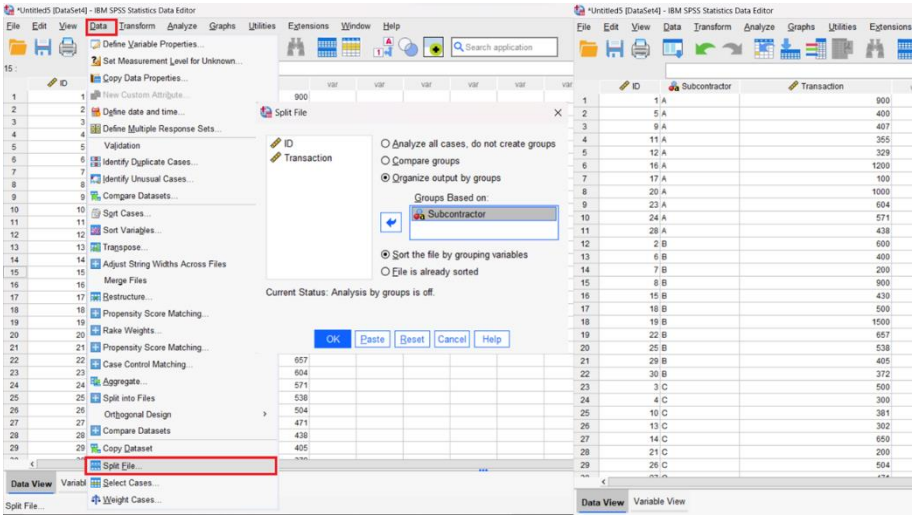
zestawach danych. Czasami można napotkać dwie bazy danych oddzielnie zaimportowane do dwóch plików, zestawów danych, ale preferowane jest ich połączenie w celu lepszej analizy. Rozważmy firmę logistyczną z dwoma oddziałami, z których każdy dostarcza dane dotyczące kosztów i przewożonych ładunków w kilogramach. Celem kierownictwa jest analiza ogólnej wydajności firmy. W programie SPSS osiągnięcie tego celu wymaga przejścia do sekcji "Data" (*pl. Dane*), wybrania opcji "Merge files" (*pl. Scal pliki*) i skorzystania z dwóch różnych opcji. Jedna z nich obejmuje wybranie "Cases" (*pl. Przypadki*) i określenie zmiennej do scalenia, usunięcie tej zmiennej podczas scalania innych. Alternatywnie, wybranie opcji "Variable" (*pl. Zmienna*) zachowuje zmienną w nowym zestawie danych. Praktyczne zastosowanie jest widoczne w naszym scenariuszu logistycznym, w którym scalanie zestawów danych upraszcza kompleksową analizę wydajności firmy.



Rysunek 7.5 Okno scalania plików

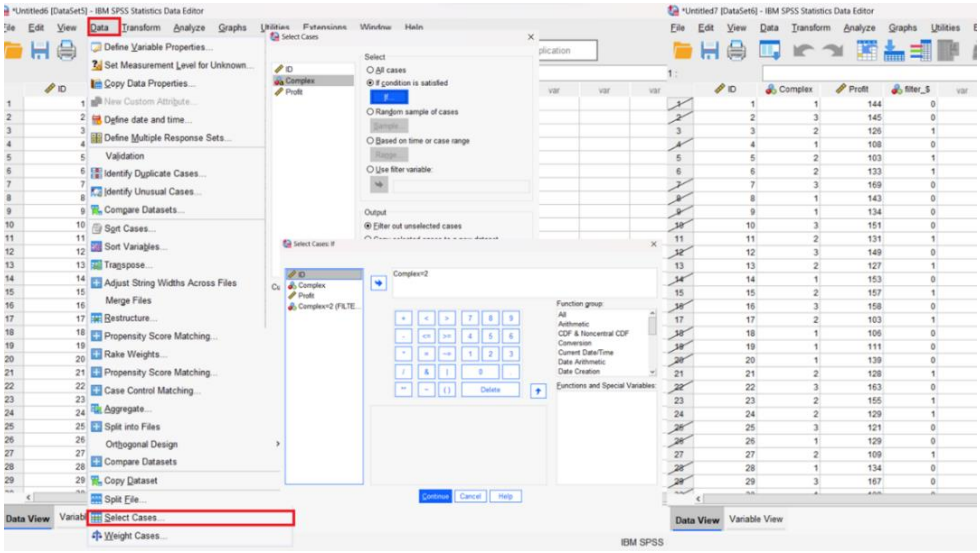
Podczas gdy funkcje scalania i dzielenia umożliwiają określone manipulacje danymi, opcja "Select cases" (*pl. Wybierz przypadki*) oferuje wyraźne korzyści. Wyobraź sobie, że masz dane dla sklepów B, C i D w jednej bazie danych, a skupiasz się wyłącznie na porównaniu sklepu A i sklepu C. Wybierając "Data" (*pl. Dane*) i "Select cases" (*pl. Wybierz przypadki*), można określić interesujące zmienne, skutecznie odfiltrowując niechciane dane. Na przykład ustawienie sklepu C, czyli wartości 2 nakazuje oprogramowaniu skoncentrowanie się wyłącznie na sklepie C, generując dane wyjściowe, które są następnie dostępne

do późniejszych analiz, takich jak statystyki opisowe, koncentrując się wyłącznie na wybranych przypadkach. Takie podejście umożliwia również analizę porównawczą tylko wartości sklepu A i sklepu C.



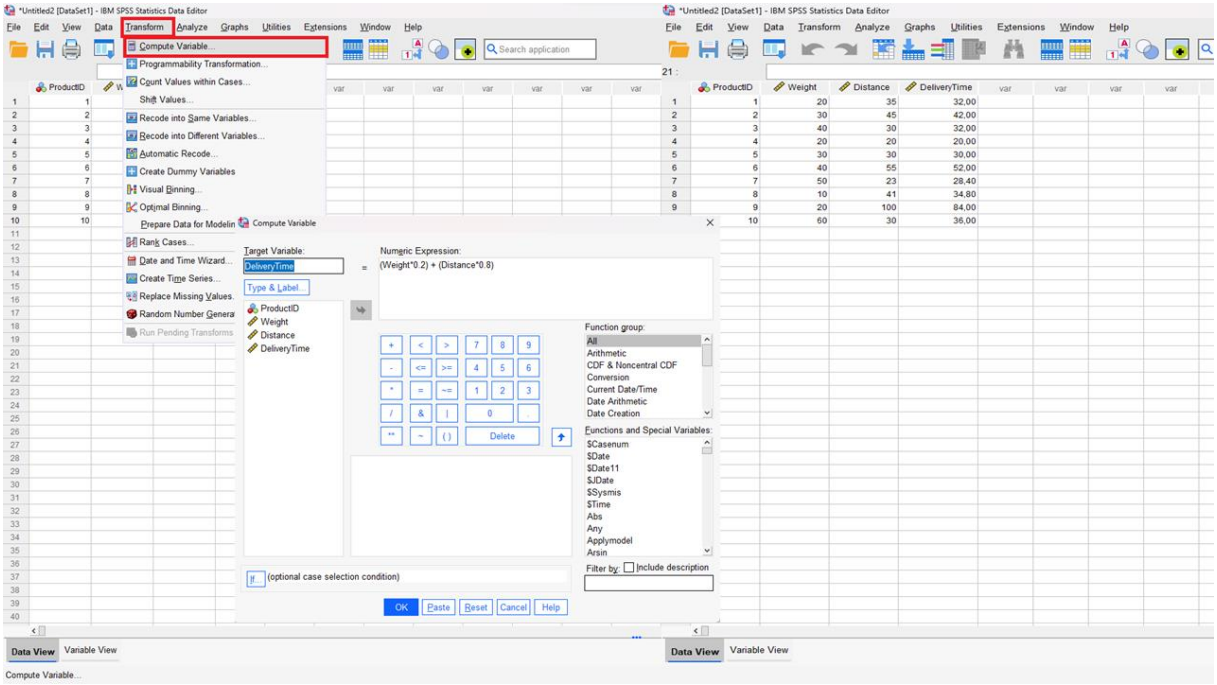
Rysunek 7.6 Okno podziału pliku

Podczas gdy funkcje scalania i dzielenia umożliwiają pewną manipulację danymi, istnieje również opcja "Select cases" (*pl. Wybierz przypadki*). Wyobraźmy sobie, że wiemy na pewno, że sklep A ma średnio 120 € zysku i chcemy porównać to ze sklepem C. Niestety w naszej bazie danych mamy dane dla sklepów B, C i D w jednej bazie danych, a analiza obejmowałaby dane ze wszystkich trzech sklepów. Klikając "Data" (*pl. Dane*) i "Select cases" (*pl. Wybierz przypadki*) można wybrać przypadki o interesującej nas wartość zmiennej, na której chcemy się skupić. Dlatego ustawiliśmy, że zmienna sklep ma wynosić 2, a następnie utworzyliśmy funkcję dla oprogramowania, aby skupić się tylko na sklepie C, który ma kod 2. Dane wyjściowe można następnie wykorzystać do późniejszej analizy, wybierając tę nową kolumnę (np. statystyki opisowe).



Rysunek 7.7 Wybór procedury postępowania

Czasami zbiory danych mogą już zawierać zmienne, ale istnieje potrzeba wprowadzenia nowych zmiennych w oparciu o istniejące. Weźmy na przykład menedżera firmy logistycznej, który posiada dane dotyczące wagi i przebytej odległości dla różnych produktów, ale potrzebuje danych o czasie dostawy w celu optymalizacji tras. W SPSS osiągnięcie tego wymaga kliknięcia "Transform" (*Przekształć*), a następnie "Compute Variables" (*Oblicz zmienne*). Nowa zmienna, DeliveryTime (*CzasDostawy*), jest tworzona w nowym oknie poprzez ustawienie wyrażeń numerycznych. W tym przypadku przypisanie skali 0,8 do odległości i 0,2 do wagi skutkuje nową zmienną reprezentującą czas dostawy, co jest kluczowym dodatkiem do zbioru danych. Istnieje również elastyczność obliczania dodatkowych zmiennych, stworzonych na potrzeby testów statystycznych.



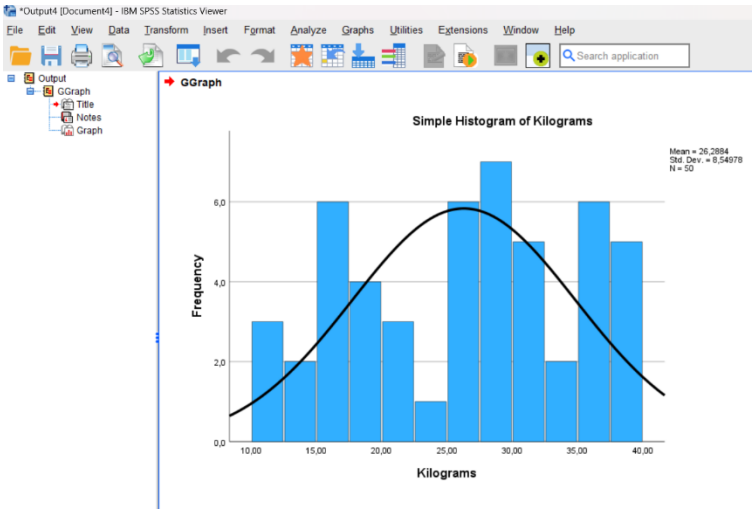
Rysunek 7.8 Procedura obliczania zmiennych

Na tym kończy się mały przegląd funkcji zarządzania danymi, które obejmuje SPSS, a które mogą być przydatne podczas kolejnych testów modeli omówionych w tym rozdziale. Będziemy kontynuować etapy wymagane przed przeprowadzeniem testu statystycznego w oprogramowaniu SPSS.

7.3 Przygotowanie testu

Przed przystąpieniem do testów statystycznych konieczne jest przestrzeganie standardowego procesu analizy danych, który obejmuje eksplorację danych (jak opisano w rozdziałach 7.1 i 7.2), analizę danych i interpretację wyników (Garth, 2008; George & Mallery, 2022). W tym rozdziale skupiamy się na analizie danych przy użyciu oprogramowania SPSS. Ponieważ hipotezy zostały już omówione w poprzednich rozdziałach, skupimy się przede wszystkim na przeprowadzeniu testów normalności w SPSS. Istnieją trzy metody oceny normalności: histogram, wykres QQ i test normalności. Wskazane jest zastosowanie co najmniej dwóch, jeśli nie wszystkich trzech z tych opcji, ponieważ każda z nich dostarcza odrębnych informacji (Ghasemi & Zadesiasl, 2012). Aby utworzyć histogram, przejdź do "Graphs" (*Wykresy*), a następnie "Chart Builder" (*Kreator wykresów*). W nowym oknie wybierz „Histogram”. Jeśli masz wiele zmiennych, musisz powtórzyć ten proces dla każdej z nich, aby uzyskać wyniki.

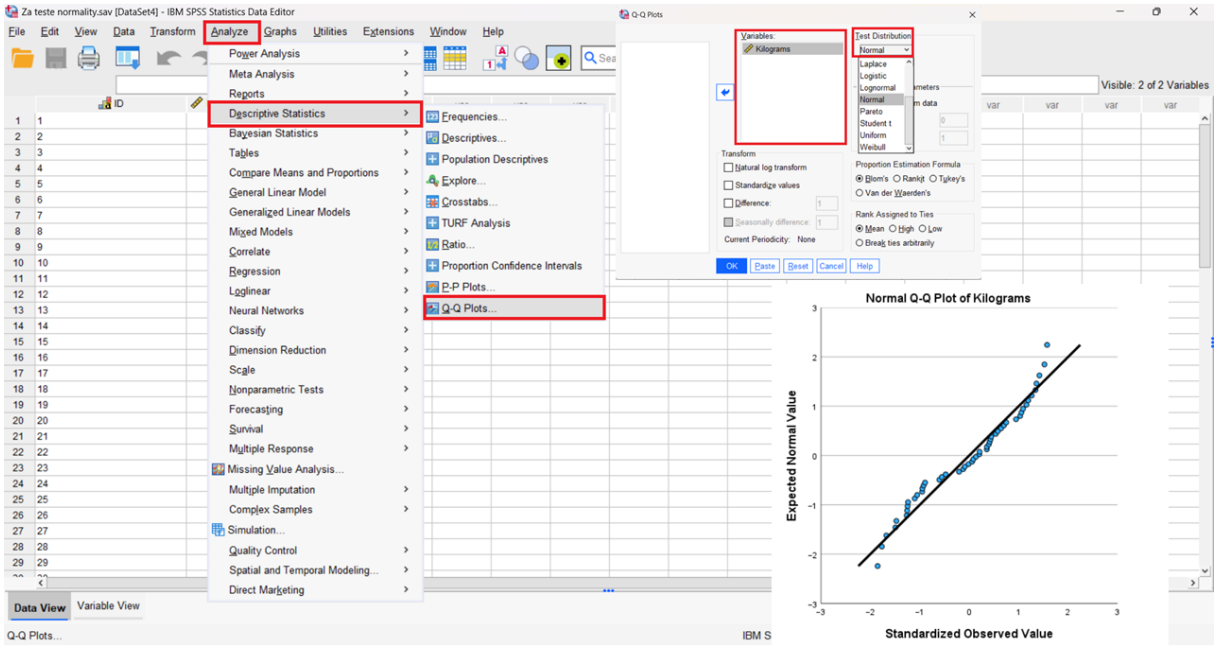
Histogram potwierdza test rozkładu normalnego, jeśli słupki reprezentujące wartości zmiennych przypominają krzywą dzwonową. Jeśli słupki przechylają się bardziej w lewą lub prawą stronę, może to wskazywać na rozkład wykładniczy. Na przykład została wygenerowana baza danych zawierająca 100 rekordów, obserwacji, z których każdy miał zmienną reprezentującą wagę w kilogramach. Postępując zgodnie z instrukcjami, utworzony został histogram, jak pokazano na rysunku 7.9. Jak widać na rysunku, słupki są rozmieszczone na wykresie i chociaż mogą nie odzwierciedlać idealnie krzywej, niemniej jednak sugerują rozkład normalny i pozytywny wynik testu (George & Mallery, 2022; Goeman & Solari, 2021).



Rysunek 7.9 Histogram wyników testu normalności

Inną opcją przeprowadzania testów normalności jest wykres Q-Q, który można zainicjować, klikając "Analyze" (*pl. Analizuj*), następnie "Descriptive Statistics" (*pl. Statystyki opisowe*), a następnie wybierając "Q-Q Plots" (*pl. Wykresy Q-Q*). Zaletą tego podejścia jest to, że pozwala ono na ocenę wielu zmiennych jednocześnie (Williamson, b.d). Test uznaje się za udany, gdy punkty na wykresie skupiają się blisko linii prostej, reprezentującej rozkład normalny. Jeśli punkty tworzą „ogony", oznacza to niepowodzenie testu normalności (Andersen & Dennison, 2018). Korzystając z tej samej bazy danych z testu wykresu histogramu, przeprowadzony został test Q-Q Plot. Na rysunku 7.10 poniżej można zauważyć, że większość punktów skupienia dla naszej zmiennej pokrywa się z linią prostą, co wskazuje na normalny rozkład naszych danych. Chociaż już na tym etapie można było stwierdzić,

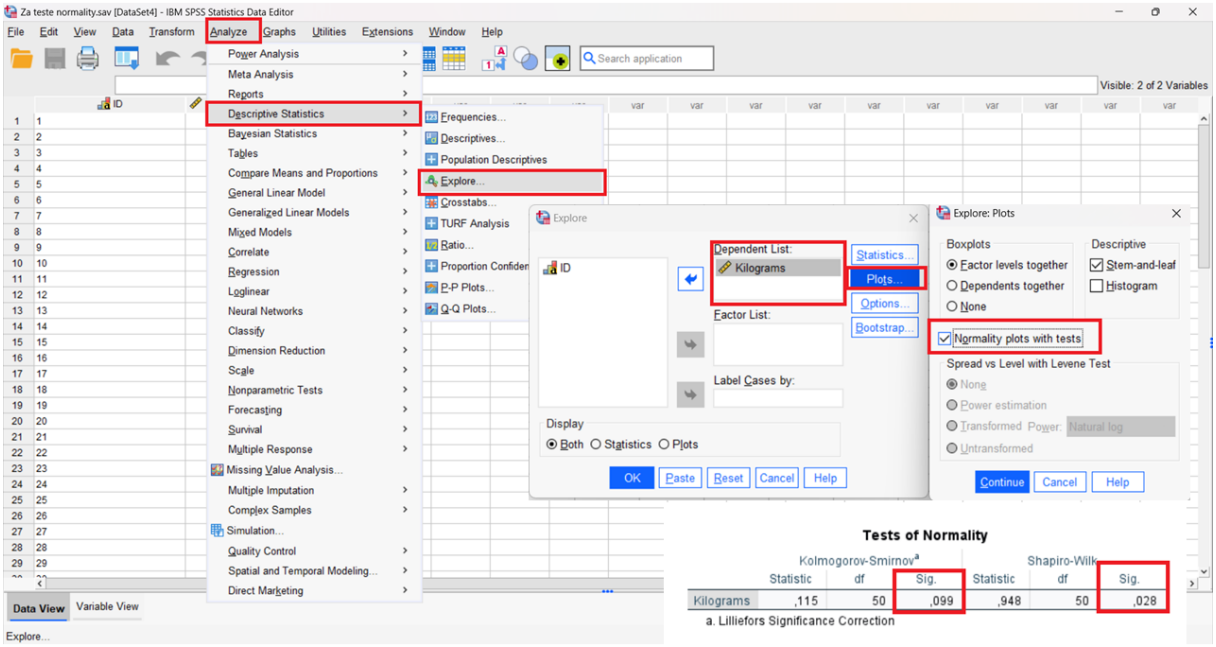
że test normalności jest pozytywny, zdecydowaliśmy się poszukać potwierdzenia we wszystkich trzech testach.



Rysunek 7.10 Ustawienia i wyniki testu normalności wykresu Q-Q.

Ostatnią opcją weryfikowania normalności danych jest tak zwany Test Normalności, uważany za test statystyczny. Zazwyczaj wykorzystuje on test Kołmogorowa-Smirnowa, ale w przypadku małych rozmiarów, do 20 obserwacji, próby należy zastosować test Shapiro-Wilka (Goeaman & Solari, 2021). W programie SPSS można wykonać ten test, klikając "Analyze" (*Analizuj*), następnie "Descriptive Statistics" (*Statystyki opisowe*) a później "Explore" (*Eksploruj*). Należy ustawić zmienne, które mają być sprawdzone w polu "Dependent List" (*Lista zależna*). Następnie w sekcji "Plots" (*pl. Wykresy*) trzeba wybrać "Normality Plots with Tests" (*Wykresy normalności z testami).* Wyjątkowo tylko w tym teście wynik testu hipotezy zerowej (H0) uznaje się za pozytywny (potwierdzający), jeśli kolumna „Sig" (p-value) w wynikach jest większa niż 0,05, co wskazuje na rozkład normalny. Vice versa, jeśli *p-value* jest mniejsze niż 0,05. Sugeruje to rozkład inny niż normalny, a wynik testu uznaje się za negatywny dla hipotezy zerowej. Test ten został przeprowadzony ponownie, korzystając z tej samej bazy danych, co w poprzednich testach. Na podstawie wyników można stwierdzić, że zgodnie ze standardem Kołmogorowa-Smirnowa test jest pozytywny, ponieważ wartość p jest wyższa niż 0,05.

Jednak w przypadku testu Shapiro-Wilka wartość *p* jest niższa, co wskazuje na negatywny wynik testu. Te różne wyniki występują, ponieważ oba podejścia mają różne ustawienia czułości i mocy w wykrywaniu odchyleń (Ghasemi & Zahediasl, 2012). Ponieważ zostały przeprowadzone już zarówno testy wykresów Q-Q, jak i histogramów, test normalności można ogólnie uznać za pozytywny. Po potwierdzeniu testów normalności można przeprowadzić główne testy, takie jak Test dla Jednej Próby.



Rysunek 7.11 Ustawienia i wyniki testu normalności

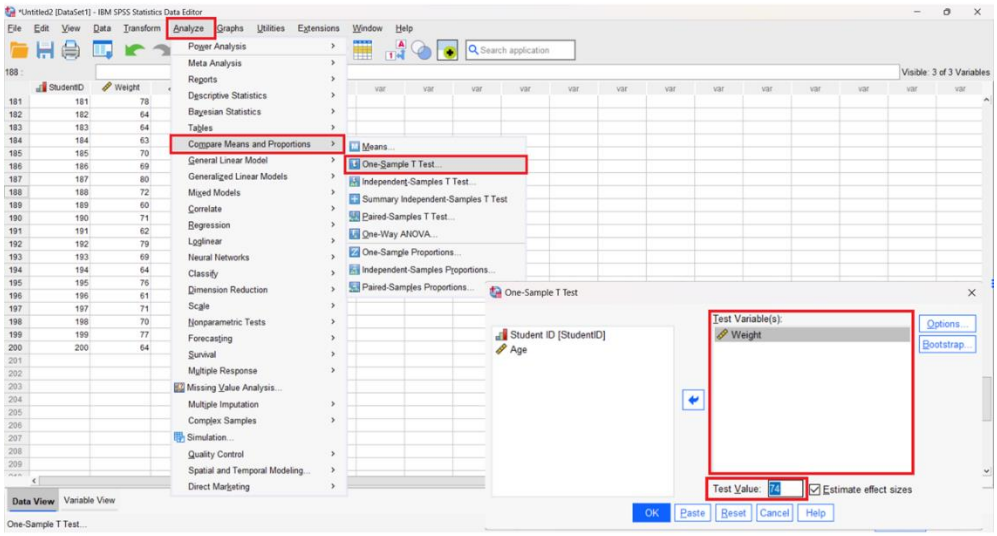
7.4 Test T dla jednej próby

W poprzednich rozdziałach omówiono już teorię stojącą za testem T dla jednej próby; skupimy się zatem przede wszystkim na przeprowadzeniu testu za pomocą oprogramowania SPSS. Na potrzeby naszego testu T dla jednej próby przygotowana została baza danych z próbą 200 danych wejściowych, która obejmuje 1 zmienną kategorialną (Student ID) (*Identyfikator studenta*) i 2 zmienne numeryczne (Weight and Age) (*Waga i Wiek*) (Kim, 2015). Postępując zgodnie ze wskazówkami z poprzednich podrozdziałów, przeprowadzamy następujące kroki:

- Zbadaj dane, a mianowicie nasze **zmienne** i **statystyki opisowe** oraz postaw **pytanie hipotezę**,

- Sprawdź **normalność**, ponieważ powinien wystarczyć tylko jeden histogram zmiennej i wykres Q-Q,
- Postaw hipotezę, w której dla **Wartości zerowej** zmienna nie różni się od określonej wartości i **Alternatywnej**, w której się różni,
- Przeprowadź **test t-Studenta**,
- Zinterpretuj wyniki, koncentrując się na **odrzuceniu** lub **nieodrzuceniu wartości zerowej**, odpowiedz na pytanie i napisz raport z naszego testu.

W naszym przypadku zdecydowaliśmy, że pytanie powinno brzmieć: „Czy średnia waga studentów przekracza 74 kilogramy?". Po zadaniu pytania ustalamy naszą hipotezę, która brzmi: „Zerowa = nie ma różnicy" i „Alternatywna = istnieje różnica". Przeprowadziliśmy histogram i wykresy Q-Q, aby sprawdzić testy normalności, a po ich zakończeniu przeprowadziliśmy test T. Aby uruchomić test T, klikamy "Analyze" (*Analizuj*), następnie "Compare Means" (*Porównaj średnie*) i "One-Sample T-test" (*test T dla jednej próby*). W polu zmiennej testowej umieszczamy „weight" (waga), ustawiamy wartość testowaną na 74 i rozpoczynamy test (patrz rysunek 7.12).



Rysunek 7.12 Ustawienia testu T dla jednej próby

Po potwierdzeniu testu pojawi się kolejne okno z wynikami naszej analizy (patrz rysunek 7.13). Okno to zawiera kilka informacji dotyczących naszej analizy. W tym przypadku obie wartości *p-value* są niższe niż 0,05, co wskazuje na istotność testu, a więc wyniki

negatywny dla H0. Dodatkowo sprawdzamy wartości *t* i *df*, które w naszym przypadku wynoszą odpowiednio -9,806 i 199. Na podstawie tych wyników możemy stwierdzić, że nasza hipoteza zerowa została odrzucona. W związku z tym pełny raport wyników jest następujący: „Średnia waga studenta jest znacząco niższa niż wartość 74 kg. Stwierdzamy poziom średniej wagi niższy niż w hipotezie (był 74 kg), gdyż zauważmy, że średnia = 69,63, a przede wszystkim wartość testu *t*, tzw. sprawdzian, okazała się ujemna: 1-próbkowy test t, *t* = -9,806, *df* = 199, wartość *p*< 0,001)".

T-Test

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Weight	200	69,63	6,303	,446

One-Sample Test

Test Value = 74

	t	df	Significance One-Sided p	Significance Two-Sided p	Mean Difference	95% Confidence Interval of the Difference Lower	95% Confidence Interval of the Difference Upper
Weight	-9,806	199	<,001	<,001	-4,370	-5,25	-3,49

One-Sample Effect Sizes

		Standardizer[a]	Point Estimate	95% Confidence Interval Lower	95% Confidence Interval Upper
Weight	Cohen's d	6,303	-,693	-,847	-,538
	Hedges' correction	6,326	-,691	-,844	-,536

a. The denominator used in estimating the effect sizes.
Cohen's d uses the sample standard deviation.
Hedges' correction uses the sample standard deviation, plus a correction factor.

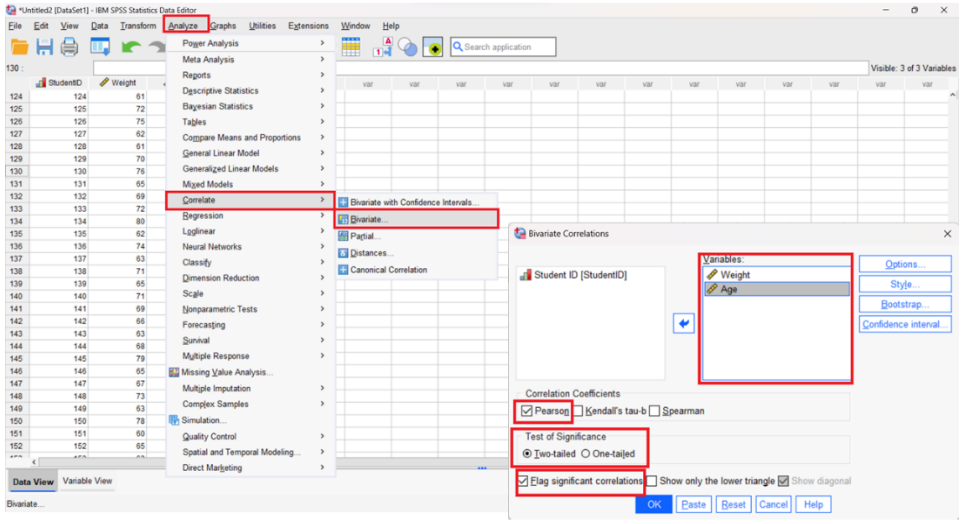
Rysunek 7.13 Wyniki testu T dla jednej próby.

7.5 Korelacja

Przejdźmy teraz do drugiego testu, którym jest test korelacji. Przeprowadzimy go przy użyciu tej samej bazy danych, co w przykładzie testu t dla jednej próby. Podobnie jak w przypadku testu t dla jednej próby, będziemy postępować zgodnie z procedurą z kilkoma modyfikacjami. Podczas przeprowadzania korelacji między dwiema zmiennymi ważne jest, aby określić, która z nich jest zmienną zależną, a która zmienną niezależną (Janse i in., 2021; Mishra i in., 2019). Wyboru tego można dokonać na podstawie pytania badawczego.

W naszym przypadku chcemy zbadać „Czy istnieje korelacja między wiekiem ucznia a jego wagą?". Zgodnie z tym pytaniem uważamy wagę za zmienną zależną, a wiek za zmienną niezależną, ponieważ chcemy zbadać, czy zmiany wieku są związane ze zmianami wagi. Definiujemy hipotezę zerową i alternatywną (patrz rysunek 7.13 i 7.14), a następnie uruchamiamy test, klikając "Analyze"

(*Analizuj*), a następnie "Correlate" (*Koreluj*) i "Bivariate" (*Współzależność prosta*). Obie zmienne powinny zostać umieszczone w polu "Variable" (*Zmienna*). Upewnij się, że „Pearson", „Two-Tailed" (*Dwustopniowy*) i „Flag Significant" (*Oznacz znaczące*) są wybrane lub ustawione (patrz rysunek 7.14). W tym przypadku wybraliśmy „Pearson", ponieważ nasze dane wskazywały na rozkład normalny i mogły być analizowane przy użyciu metod parametrycznych. Jeśli rozkład normalny nie został potwierdzony, należy zastosować metody nieparametryczne (w tym przypadku należy wybrać Spearmana zamiast Pearsona) (George & Mallery, 2022; McClure, 2005).



Rysunek 7.14 Ustawienia testu korelacji

Ponownie otrzymujemy wyniki w nowym oknie (patrz rysunek 7.15). Na podstawie wyników możemy zauważyć, że nasza korelacja Pearsona wynosi -0,038, a nasze *p-value* wynosi 0,596. W analizie korelacji im wartość korelacji jest bliższa zeru, tym słabsza jest korelacja między zmiennymi. W naszym przypadku korelacja jest bardzo bliska zeru, co wskazuje na brak istotnej korelacji między dwiema zmiennymi (McClure, 2005). Ponadto duże *p-value* (0,596) sugeruje, że nie ma istotnych dowodów na to, że istnieje znacząca korelacja między dwiema wybranymi zmiennymi (Williamson, b.d.). W rezultacie nasza hipoteza zerowa nie została odrzucona. Na tej podstawie możemy stwierdzić, że „Nie było korelacji między wiekiem a wagą uczniów".

Correlations

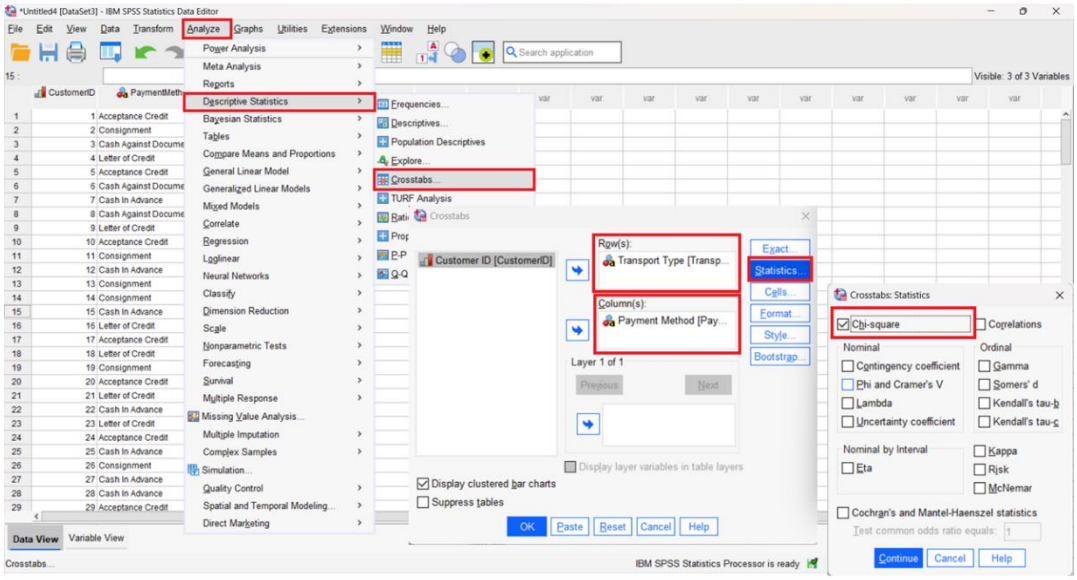
Correlations

		Weight	Age
Weight	Pearson Correlation	1	-,038
	Sig. (2-tailed)		,596
	N	200	200
Age	Pearson Correlation	-,038	1
	Sig. (2-tailed)	,596	
	N	200	200

Rysunek 7.15 Wyniki testu korelacji

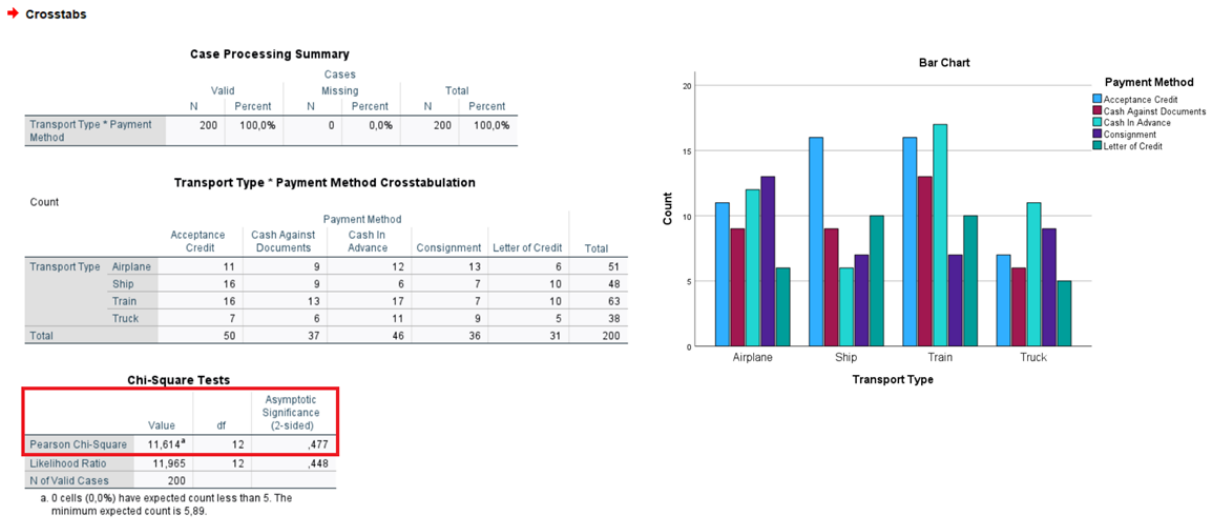
7.6 Chi-Kwadrat

Trzecim testem, który przeprowadzimy w oprogramowaniu SPSS, jest test Chi-kwadrat. W przeciwieństwie do poprzednich dwóch testów, test Chi-kwadrat porównuje dwie zmienne kategorialne, a nie zmienne liczbowe (Turhan, 2020). Podobnie jak w sekcjach 7.4 i 7.5, zaczynamy od zbadania danych i sformułowania pytania badawczego. W naszym przykładzie mamy firmę logistyczną z 200 klientami i dysponujemy danymi na temat rodzaju płatności i rodzaju transportu wybranego przez każdego klienta. Pytanie, na które chcemy odpowiedzieć, brzmi: „Czy różne rodzaje płatności wykazują różne preferencje dotyczące rodzajów transportu?". Ponieważ mamy do czynienia tylko ze zmiennymi kategorycznymi, nie ma potrzeby przeprowadzania testu normalności. Ustalamy hipotezę zerową (H0: preferencje dotyczące rodzajów transportu są takie same wśród wszystkich rodzajów płatności) i hipotezę alternatywną. Aby przeprowadzić analizę Chi-kwadrat, zgodnie z H0 nazywaną też testem niezależności, kliknij "Analyze" (*Analizuj*), następnie "Descriptive Statistics" (*Statystyki opisowe*) i wybierz „Crosstabs" (*Tabele krzyżowe*). Kluczowe jest umieszczenie zmiennych opartych na pytaniu badawczym w polu kolumny lub wiersza (patrz rysunek 7.16) (Garth, 2008).



Rysunek 7.16 Ustawienia testu Chi-kwadrat

Po zakończeniu analizy w nowym oknie wyświetlone zostaną wyniki (patrz rysunek 7.17). W tym oknie można zauważyć, że wartość Chi-kwadrat Pearsona wynosi 11,614, wartość *df* wynosi 12, a *p-value* (istotność empiryczna) wynosi 0,477. Na podstawie tych wyników możemy stwierdzić, że nie ma istotnego związku między dwiema zmiennymi, a hipoteza zerowa nie została odrzucona. W związku z tym raport jest następujący: „Nie wykryto znaczącej preferencji między różnymi rodzajami płatności dla różnych rodzajów transportu (2- stopniowy test Chi-kwadrat, chi-sq = 11,614, *df* = 12, *p-value* = 0,477)."



Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Transport Type * Payment Method	200	100,0%	0	0,0%	200	100,0%

Transport Type * Payment Method Crosstabulation

Count

		Payment Method					
		Acceptance Credit	Cash Against Documents	Cash In Advance	Consignment	Letter of Credit	Total
Transport Type	Airplane	11	9	12	13	6	51
	Ship	16	9	6	7	10	48
	Train	16	13	17	7	10	63
	Truck	7	6	11	9	5	38
Total		50	37	46	36	31	200

Chi-Square Tests

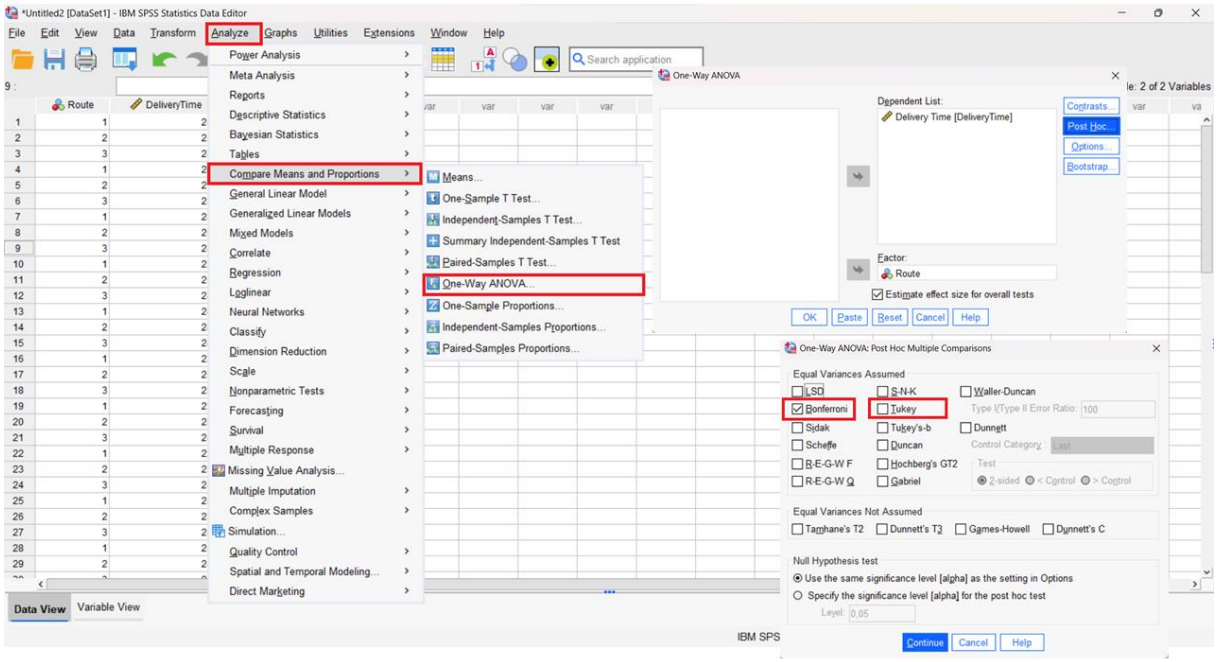
	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	11,614[a]	12	,477
Likelihood Ratio	11,965	12	,448
N of Valid Cases	200		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 5,89.

Rysunek 7.17 Wyniki testu Chi-kwadrat

7.7 ANOVA

Ostatnim testem, który zostanie omówiony, jest test ANOVA, w szczególności wariant testu oparty na prostym modelu znanym jako jednoczynnikowa ANOVA, który obejmuje zmienną niezależną kategorialną (to jest czynnik potencjalnie oddziałujący na wartość zmiennej zależnej) i zmienną zależną liczbową (Goeman & Solari, 2021). Podobnie jak w przypadku testu T, będziemy postępować zgodnie z tą samą procedurą: zbadać dane, sformułować pytanie badawcze, przeprowadzić test i postawić hipotezy. Rozważmy studium przypadku dyspozytora transportu pracującego dla firmy logistycznej. Dyspozytor ściśle współpracuje z firmą partnerską i regularnie planuje trzy różne trasy dla ciężarówek w celu dostarczenia towarów. Ze względu na politykę „Just-in-time" kładącą nacisk na szybsze dostawy, pojawia się pytanie: „Czy wybór trasy dostawy wpływa na czas dostawy dla firmy?". Aby przeprowadzić test ANOVA w SPSS, przejdź do "Analyze" (*Analizuj*), następnie „Compare Means..." (*Porównaj średnie…*, a następnie „One-way ANOVA" (*Jednoczynnikowa ANOVA*). Umieść zmienną zależną w polu "Dependent List" (*Lista zależnych*), a zmienną czynnikową w polu "Factor" (*Czynnik)* (patrz rysunek 7.18). W celu dokładnej analizy uwzględnione zostało również ustawienie „Post Hoc", czyli kontrastów (różnic) wartości czasowych dla dróg. Ważne jest, aby pamiętać, że analiza Post Hoc powinna być przeprowadzana tylko wtedy, gdy wynik testu ANOVA jest negatywny dla H0. Stosując analizę Post Hoc, można zidentyfikować najbardziej optymalny wybór kategorii czynnika (w naszym przypadku trasę). Najbardziej niezawodnymi metodami analizy post hoc są korekta Bonferroniego lub metoda HSD Tukeya (Goeman i Solari, 2021).



Rysunek 7.18 Ustawienia ANOVA

Wyniki naszej analizy wskazują, że nasza wartość statystyki *F* wynosi 11,173 (wyższe wartości wskazują na większe różnice między grupami), a *p-value* <0,001, co oznacza, że nasza hipoteza zerowa została odrzucona (patrz rysunek 7.19). Ponieważ istnieje znacząca różnica między trzema trasami (<0,001), test post hoc jest również ważny w naszym przypadku (George & Mallery, 2022). Po przeprowadzeniu testu kontrastów Bonferroniego widzimy, że parę najmniejszych wartości *p-value* odnotowano w przypadku trasy 2 (patrz rysunek 7.19) w porównaniu z obiema pozostałymi trasami. W raporcie możemy stwierdzić, że „wystąpiła znacząca różnica w wyborze trasy dostawy w korelacji z czasem dostawy (1-kierunkowa ANOVA, *F*=11,173, *df* = 47, *p* = <0,001). Trasa 2 miała najdłuższy czas dostawy".

ANOVA

Delivery Time

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	16,527	2	8,263	11,173	<,001
Within Groups	33,280	45	,740		
Total	49,807	47			

Post Hoc Tests

Multiple Comparisons

Dependent Variable: Delivery Time

Bonferroni

(I) Route	(J) Route	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
1	2	-1,4250*	,3040	<,001	-2,181	-,669
	3	-,5500	,3040	,231	-1,306	,206
2	1	1,4250*	,3040	<,001	,669	2,181
	3	,8750*	,3040	,018	,119	1,631
3	1	,5500	,3040	,231	-,206	1,306
	2	-,8750*	,3040	,018	-1,631	-,119

*. The mean difference is significant at the 0.05 level.

Rysunek 7.19 Wstępne wyniki ANOVA i wyniki testu post-hoc

Kończymy ten rozdział książki ze świadomością, że omówiliśmy w nim niektóre z bardziej powszechnych testów. Istnieją jeszcze inne testy, takie jak ANOVA z powtarzanymi pomiarami, testy wiarygodności i testy wrażliwości, które również można modelować i analizować za pomocą oprogramowania SPSS. Te dodatkowe testy zapewniają szerszy zakres narzędzi do analizy danych i wyciągania znaczących wniosków na temat różnych badań i praktycznych zastosowań.

BIBLIOGRAFIA DO ROZDZIAŁU 7.

- Andersen, A.J. & Dennison, J.R. (2018). An Introduction to Quantile-Quantile Plots for the Experimental Physicist. Journal Articles, 51.
- Garth, A. (2008). Analysing data using SPSS [Dostępne: https://students.shu.ac.uk/lits/it/documents/pdf/analysing_data_using_spss.pdf, dostęp October 26, 2023]
- George, D. & Mallery, P. (2022). IBM SPSS Statistics 27 Stet by Step: A Simple Guide and Reference, 17TH edition, Abingdon: Routledge
- Ghasemi, A. & Zahediasl, S. (2012). Normality Tests for Statistical Analysis: A Guide for Non-Statisticians. International Journal of Endocrinology and Metabolism, 10(2), pp. 486-489.
- Goeman, J.J. & Solari, A. (2021). Comparing Three Groups. The American Statistician, 76(2), pp. 168-176
- IBM (2021). IBM SPSS Statistics 28 Brief [Dostępne: https://www.ibm.com/docs/en/SSLVMB_28.0.0/pdf/IBM_SPSS_Statistics_Brief_Guide.pdf, dostęp October 26, 2023]

- Janse, R.J., Hoekstra, T., Jager, K.J., Zoccali, C., Tripepi, G., Dekker, F.W. & van Diepen, M. (2021). Conducting correlation analysis: important limitations and pitfalls, 14(11), pp. 2332-2337.
- Kim, T.K. (2015). T test as a parametric statistic. Korean Journal of Anesthesiology, 68(6), pp. 540-546
- Landau, S. & Everitt, B.S. (2004). A Handbook of Statistical Analyses using SPSS, 1st edition, London: Chapman & Hall/CRC
- McClure, P. (2005). Correlation Statistics Review of the Basics and Some Common Pitfalls. Journal of Hand Therapy, 18(3), pp. 378-380
- Mishra, P., Singh, U., Pandey, C.M., Mishra, P. & Pandey, G. (2019). Application of Student's t-test, Analysis of Variance, and Covariance. Annals of Cardiac Anesthesia, 22(4), pp. 407-411
- Turhan, N.S. (2020). Karl Pearson's chi-square tests. Educational Research and Reviews, 15(9), pp. 575-580
- Williamson, M. (b.d.). Data Analysis using SPSS [Dostępne: https://med.und.edu/research/daccota/_files/pdfs/berdc_resource_pdfs/data_analysis_using_spss.pdf, dostęp October 26, 2023]

8.Podstawy analityki biznesowej, w tym R i SQL

Czym jest analityka biznesowa (BA)? Jakie problemy rozwiązuje i jakich narzędzi używa? Czym są R i SQL? W jaki sposób BA jest powiązana z R i SQL? Czy istnieją przykłady dobrych praktyk, w których te programy są wykorzystywane do rozwiązywania logistycznych problemów biznesowych?

Na te i podobne pytania postaramy się udzielić odpowiedzi w poniższym rozdziale.

8.1 Czym jest analityka biznesowa?

BA reprezentuje holistyczne podejście do analizy danych i podejmowania decyzji biznesowych. Jest to środowisko oparte na danych, którego celem jest poprawa wyników biznesowych firmy poprzez zapewnienie podstaw do bardziej świadomego podejmowania decyzji. Jest to systematyczny proces myślenia, który stosuje jakościowe, ilościowe i statystyczne narzędzia i metody obliczeniowe do analizy danych, uzyskiwania wglądu, informowania i wspierania podejmowania decyzji. Każda konkretna analiza może wykorzystywać różne techniki, w tym modele diagnostyczne, predykcyjne, preskryptywne i optymalizacyjne (Power i in., 2018). Mikalef i in., (2019) zapewnia mapę drogową zarówno dla eksploracji akademickiej, jak i praktycznego wdrożenia, podkreślając transformacyjny potencjał analityki, gdy jest ona odpowiednio zintegrowana z procesami organizacyjnymi. W związku z tym autorzy stwierdzają, że BA wymaga od organizacji radykalnego przeprojektowania sposobu podejścia do takich inicjatyw, ich projektowania i udoskonalania, sposobu planowania i orkiestracji zasobów oraz ich strategicznego dostosowania, a także ponownej oceny oczekiwanych wyników, ich powiązania z celami strategicznymi, a w rezultacie opracowania odpowiednich wskaźników KPI (Mikalef i in., 2019).

Głównym zadaniem BA jest zapewnienie potoku wiedzy w celu zagwarantowania spójnego powiązania między surowymi danymi a decyzjami biznesowymi. Ogólnym celem jest efektywność biznesowa poprzez „wertykalizację", użyteczność i integrację z systemami operacyjnymi (Kohavi, Rothleder, & Simoudis, 2002). BA ma wiele obszarów zastosowań

i powiązanych: Analityka finansowa, Analityka łańcucha dostaw, Analityka kryzysowa, Analityka wiedzy, Analityka marketingowa, Analityka klienta, Analityka usług, Analityka zasobów ludzkich, Analityka talentów, Analityka procesów, Analityka ryzyka (Holsapple, Lee-Post, & Pakath, 2014).

Istnieją trzy rodzaje platform analityki biznesowej:

- **opisowe** – analizują one istniejące dane i dostarczają statystyk podsumowujących oraz podstawowych wizualizacji,
- **przewidujące** – wykorzystują istniejące dane do oszacowania najbardziej prawdopodobnych przyszłych scenariuszy,
- **nakazowe** – automatycznie przetwarzają duże zbiory danych, reguły biznesowe, warunki rynkowe itp. Platformy te wykorzystują metody uczenia maszynowego i sztucznej inteligencji. Celem jest w pełni zautomatyzowane podejmowanie decyzji o tym, jakie działania powinna podjąć firma, biorąc pod uwagę obecną sytuację, aby osiągnąć pożądane cele biznesowe.

Zainteresowanie Big data i analityką biznesową wzrosło wykładniczo w ciągu ostatniej dekady (Mikalef i in., 2019). Współczesna BA jest zakorzeniona w ciągłym rozwoju systemów wspierających podejmowanie decyzji. Postępy te obejmują coraz potężniejsze mechanizmy pozyskiwania, generowania, przyswajania, wybierania i emitowania wiedzy istotnej dla podejmowania decyzji. Biorąc pod uwagę dziedzictwo wspomagania decyzji, analityka biznesowa z konieczności bierze udział w tych mechanizmach i je wykorzystuje. Wiedza, która musi być przetwarzana, waha się od jakościowej do ilościowej, a BA dotyczy działania na obu typach wiedzy, odpowiednio do podejmowanej decyzji (Kohavi, Rothleder i Simoudis, 2002). Powodem, dla którego niektóre organizacje powinny stosować BA są problemy, które rozwiązują. Problemy, na które BA kładzie nacisk, są problemami ograniczającymi efektywne zarządzanie firmą. W związku z tym istnieje kilka powodów stosowania BA (Holsapple, Lee-Post, & Pakath, 2014):

- Osiągnięcie przewagi konkurencyjnej,
- Wspieranie strategicznych i taktycznych celów organizacji,
- Lepsza wydajność organizacyjna,
- Lepsze wyniki decyzyjne,

- Lepsze lub bardziej świadome procesy decyzyjne,
- Tworzenie wiedzy,
- Uzyskiwanie wartości z danych.

Niezależnie od rodzaju platformy wykorzystywanej w BA do rozwiązywania i wspierania procesu podejmowania decyzji w każdej firmie, istnieją trzy kluczowe filary każdego rozwiązania BA (rysunek 8.1). Ogólnie rzecz biorąc, BA zajmuje miejsce w spektrum między informatyką/matematyką/nauką o danych (z jednej strony) a biznesem i zarządzaniem (z drugiej). Analityka biznesowa wymaga zarówno wiedzy technicznej, jak i biznesowej. Głównym problemem w projektowaniu BA jest to, że granice te nie są wyraźne (Power i in., 2018). W związku z tym do przeprowadzenia BA potrzebne są narzędzia matematyczne do identyfikacji, wyodrębniania i reprezentowania spostrzeżeń we właściwy sposób za pomocą tabel, grafik, formuł itp. Dodatkowo, narzędzia programistyczne służą jako wsparcie dla tego rodzaju działań i umożliwiają szybkie i bezbłędne obliczenia, w porównaniu do tradycyjnego podejścia opartego na papierze i długopisie. Wreszcie, co nie mniej ważne, potrzebna jest wiedza logistyczna w danym obszarze lub problemie biznesowym, aby wskazać kluczowe czynniki wpływające i powiązany ekosystem.

Rysunek 8.1 Kluczowe filary BA w kontekście łańcucha dostaw i logistyki

Kluczowym konsumentem jest użytkownik biznesowy, którego praca, być może w merchandisingu, marketingu lub sprzedaży, nie jest bezpośrednio związana z analityką per se, ale który zazwyczaj wykorzystuje narzędzia analityczne do poprawy wyników niektórych procesów biznesowych w jednym lub kilku wymiarach (takich jak zysk i czas wprowadzenia na rynek). Użytkownicy biznesowi nie chcą zajmować się zaawansowanymi koncepcjami statystycznymi; chcą prostych wizualizacji i istotnych dla zadania wyników (Kohavi, Rothleder i Simoudis, 2002).

8.2 Czym jest R?

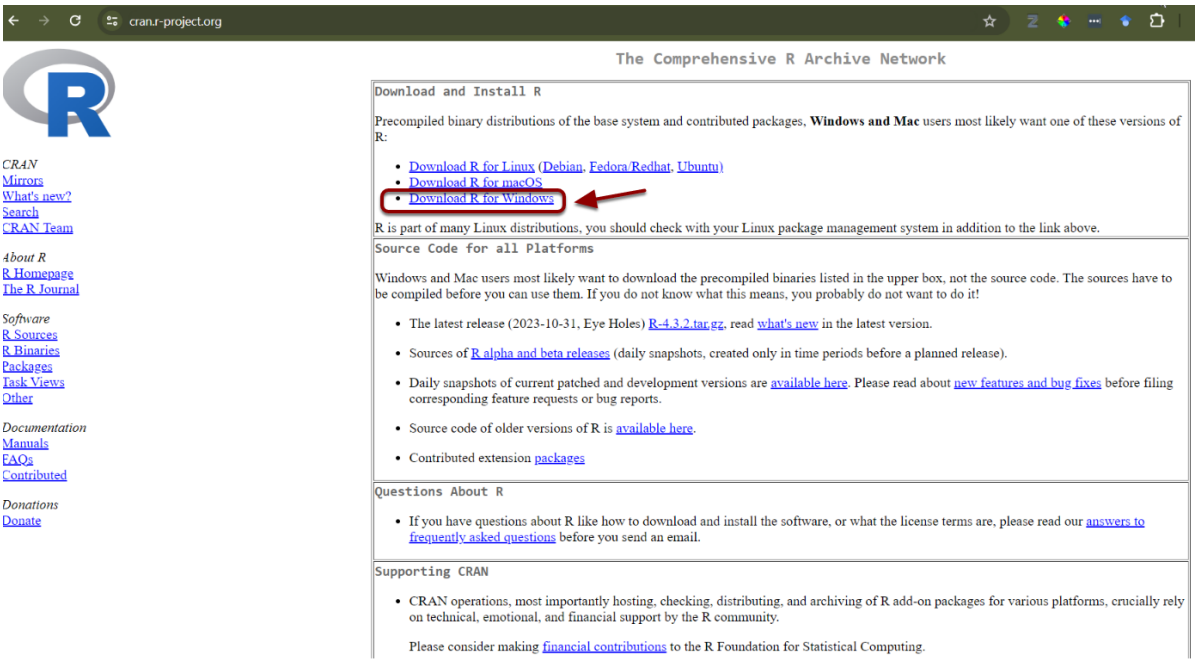
R to zintegrowany pakiet oprogramowania do manipulacji danymi, obliczeń i wyświetlania graficznego (R Core Team, 2019). Posiada on między innymi:

- skuteczny system obsługi i przechowywania danych,
- zestaw operatorów do obliczeń na tablicach, w szczególności macierzach,
- duży, spójny, zintegrowany zbiór narzędzi pośrednich do analizy danych,
- graficzne funkcje do analizy i wyświetlania danych bezpośrednio na komputerze lub na papierze, a także dobrze rozwinięty, prosty i skuteczny język programowania (zwany „S"), który zawiera funkcje warunkowe, pętle, funkcje rekurencyjne zdefiniowane przez użytkownika oraz funkcje wejścia i wyjścia. (W rzeczywistości większość funkcji dostarczanych przez system jest napisana w języku S).

Głównymi zaletami R jest fakt, że R jest darmowy i że istnieje wiele pomocy w zakresie obsługi dostępnych online. Jest dość podobny do innych pakietów programistycznych, takich jak MatLab (nie darmowy), ale bardziej przyjazny dla użytkownika niż języki programowania, takie jak C++ lub Fortran (Torfs & Brauer, 2014). R jest w dużej mierze narzędziem do opracowywania nowych metod interaktywnej analizy danych. Szybko się rozwinął i został rozszerzony o dużą kolekcję pakietów. Jednak większość programów napisanych w R jest zasadniczo efemeryczna, napisana dla pojedynczego fragmentu analizy danych (R Core Team, 2019).

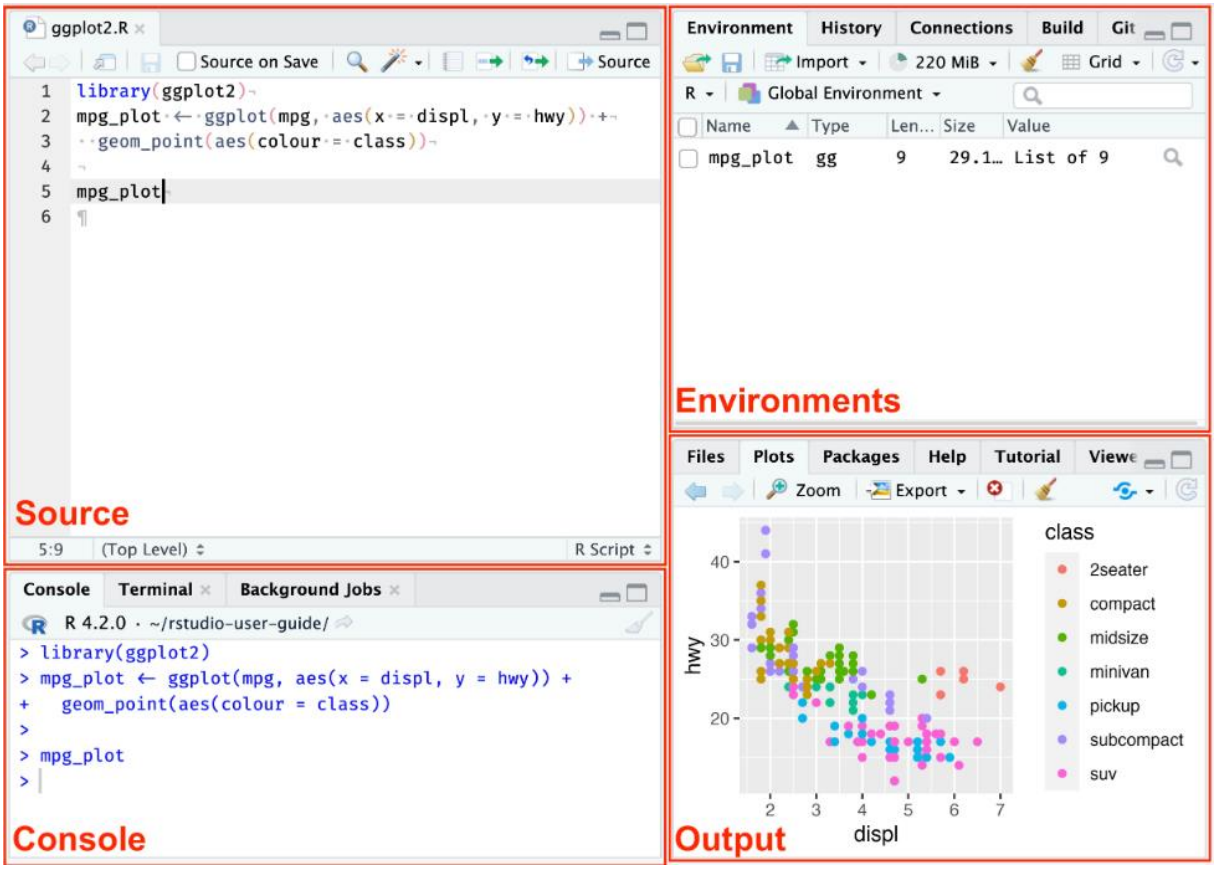
Instalowanie R i R Studio

Aby zainstalować R, należy przejść do strony cran.r-project.org i kliknąć przycisk R dla określonego systemu operacyjnego na komputerze (zazwyczaj Windows) (rysunek 8.2).



Rysunek 8.2 Strona do pobrania oprogramowania R

Spowoduje to pobranie oprogramowania R, a procedura instalacji jest taka sama, jak w przypadku każdego innego oprogramowania. Po zainstalowaniu oprogramowania R będzie ono dostarczane bez zaawansowanego zintegrowanego środowiska programistycznego (IDE), które pomaga użytkownikom w przeprowadzaniu różnych analiz. Chociaż możliwe jest wykonywanie wszelkiego rodzaju analiz z zainstalowanym tylko R, lepiej jest sparować go z jakimś nowoczesnym IDE, takim jak RStudio, które jest jednym z najpopularniejszych IDE. Procedura instalacji RStudio jest podobna do instalacji podstawowego oprogramowania R. Przejdź do https://posit.co/download/rstudio-desktop/, wyszukaj licencję open source RStudio Desktop, pobierz ją i zainstaluj. Po zainstalowaniu R i RStudio użytkownik będzie miał następujący ekran interfejsu użytkownika (rysunek 8.3).



Rysunek 8.3 Interfejs użytkownika oraz R i Rstudio (RStudio, 2024)

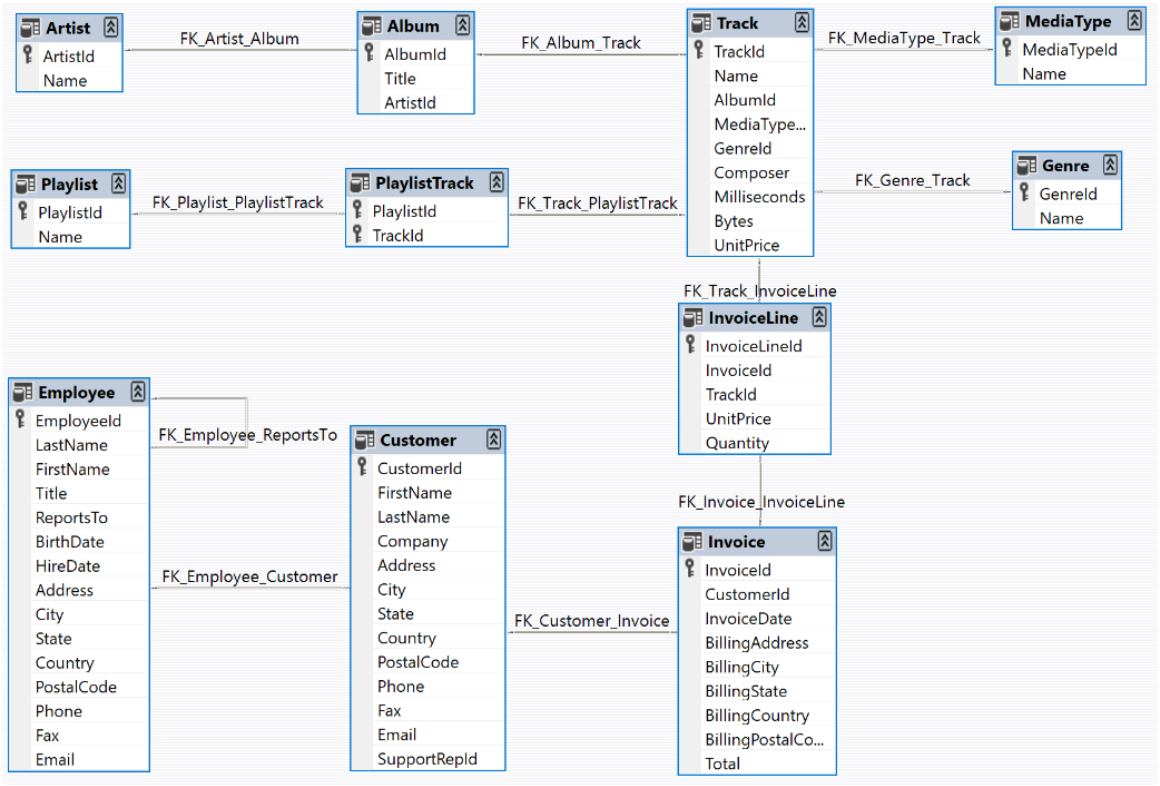
Interfejs użytkownika RStudio składa się z 4 podstawowych paneli (RStudio, 2024):

- Source (Panel źródła),
- Console (Panel konsoli),
- Environment (panel Środowisko), zawierający zakładki Environment (Środowisko) , History (Historia), Connections (Połączenia), Build (Budowanie), VCS , oraz Tutorial (Przewodnik, Instruktaż),
- Output (Panel wyjściowy), zawierający zakładki Files (Pliki), Plots (Wykresy), Packages (Pakiety), Help (Pomoc), Viewer (Przeglądarka) oraz Presentation (Prezentacja).

Każdy z paneli oraz znajdujące się w nim podsekcje i opcje pozwalają użytkownikom wykonywać różne operacje, mieć kontrolę nad niektórymi analizami danych lub mieć bardziej uporządkowany i przejrzysty widok na proces analizy danych w toku.

8.3 Czym jest SQL i jak jest powiązany z BA i R?

Chinook Database to przykładowa baza danych używana do nauki i demonstracji systemów zarządzania bazami danych (DBMS) i zapytań SQL. Została zaprojektowana jako sklep z mediami cyfrowymi wykorzystujący rzeczywiste dane z biblioteki iTunes; fikcyjne nazwy / adresy klientów i pracowników; oraz losowe dane dla informacji o sprzedaży. Baza danych zawiera różne tabele, które reprezentują dane sklepu muzycznego, w tym informacje o artystach, albumach, utworach, klientach, fakturach i nie tylko (rysunek 8.4).



Rysunek 8.4 Model danych bazy danych Chinook

Rysunek 8.4 przedstawia model danych otaczający bazę danych Chinook z różnymi tabelami danych i ich kluczami (unikalnymi identyfikatorami) oraz tabelami wspólnymi, takimi jak tabela PlaylistTrack. Tabele przekazują różne informacje o danym sklepie cyfrowym (Tabela 8.1).

Tabela 8.1 Informacje zawarte w każdej z tabel bazy danych Chinook

Nazwa tabeli	Opis
Artist (Artysta)	Zawiera informacje o artystach muzycznych.

Nazwa tabeli	Opis
Album	Zawiera informacje o albumach muzycznych, każdy powiązany z artystą.
Track (Utwór)	Zawiera informacje o poszczególnych utworach muzycznych, w tym odniesienia do albumów, typów multimediów i gatunków.
Genre (Gatunek)	Zawiera informacje o gatunkach muzycznych.
MediaType (TypNośnika)	Zawiera informacje o różnych typach multimediów (np. audio, wideo).
Customer (Klient)	Zawiera informacje o klientach, w tym dane kontaktowe i informacje o przedstawicielach pomocy technicznej.
Employee (Pracownik)	Zawiera informacje o pracownikach, w tym ich role, relacje podwładnych i dane kontaktowe.
Invoice (Faktura)	Zawiera informacje o fakturach, w tym dane klienta, informacje rozliczeniowe i łączne kwoty.
InvoiceLine (WierszFaktury)	Zawiera szczegółowe informacje o każdej pozycji na fakturze, w tym odniesienia do ścieżek i ilości.
Playlist (ListaOdtwarzania)	Zawiera informacje dotyczące list odtwarzania.
PlaylistTrack (ŚcieżkaListy Odtwarzania)	Łączy utwory z listami odtwarzania, wskazując, które utwory znajdują się na poszczególnych listach odtwarzania.

8.4 W jaki sposób analityka biznesowa, SQL i R są ze sobą powiązane?

Połączenie między BA, SQL i R jest naturalne, ponieważ wszystkie dane biznesowe powinny być przechowywane w bazach danych SQL. Jest to nadal idealistyczny cel, ponieważ w niektórych frakcjach małych i średnich przedsiębiorstw, które nadal nie w pełni rozumieją moc danych, zarządzanie danymi jest nadal słabe. W dużych firmach zostało to rozpoznane dawno temu, a dane są odpowiednio ustrukturyzowane w bazach danych (SQL lub innych, ale zwykle w SQL). Z drugiej strony analizę danych można przeprowadzić w SQL, ale w tym celu lepiej jest użyć oprogramowania zorientowanego statystycznie, w którym R znajduje się w centrum uwagi, jako jedna z najpopularniejszych platform statystycznych do przeprowadzania analizy danych.

W związku z tym SQL i R mogą być postrzegane jako idealne narzędzie do współpracy, gdy problem dotyczy obszaru BA. Istnieje kilka głównych powodów, a jednym z nich jest to, że dane BA są codziennie zmieniane i aktualizowane zgodnie z rzeczywistością rynkową i działaniami firmy: sprzedażą, pracownikami, przychodami itp. Bazy danych SQL są idealne do przechwytywania tych zmian i aktualizowania istniejących danych, podczas gdy skrypty

R są bardzo dobre w automatyzacji zadań, a także projektowaniu nowych pakietów do analizy danych. Powodem tego jest to, że rdzeń R jest bardziej zbudowany wokół koncepcji analizy danych, a nie raczej na ogólnym programowaniu, takim jak na przykład Python.

Zapytanie bazy danych SQL za pomocą R

Język programowania R i bazy danych SQL mają naturalny związek, ponieważ R jest przeznaczony głównie do analizy danych statystycznych, a większość danych transakcyjnych znajduje się w bazach danych. „Sposób", w jaki R działa w celu zarządzania manipulacją danymi z baz danych SQL, odbywa się za pośrednictwem pakietów programistycznych DBI i RSQLite. Pakiet DBI zapewnia znormalizowany interfejs do interakcji z różnymi systemami DBMS, umożliwiając użytkownikom łączenie się, wysyłanie zapytań i zarządzanie transakcjami w sposób spójny w różnych bazach danych. Pakiet RSQLite, który jest zgodny z interfejsem DBI, w szczególności ułatwia interakcję z bazami danych SQLite, umożliwiając użytkownikom wykonywanie zapytań SQL, pobieranie danych i wykonywanie operacji na bazach danych bezpośrednio z R. Razem pakiety te usprawniają proces pracy z bazami danych w R, oferując spójny i wydajny przepływ pracy.

Aby skutecznie zademonstrować wykonywanie operacji SQL z R i generowanie pożądanych wniosków z danych, w odniesieniu do omawianego problemu, przedstawionych zostało kilka fragmentów kodu na rysunkach 8.5 i 8.6.

```
---
title: "BUSINES ANALYTICS FOUNDATINS INCLUDING THE R AND SQL"
format: html
editor: visual
---

# R & SQL

## Loading libraries

```{r setup, warning=FALSE, message=FALSE}
library(DBI)
library(RSQLite)
```

Connect to the Chinook SQLite database

```{r}
con <- dbConnect(RSQLite::SQLite(), dbname = "Chinook_Sqlite.sqlite")
```

List all tables in the database

```{r}
tables <- dbListTables(con)
print(tables)
```
```

Rysunek 8.5 Fragment kodu służący do nawiązywania połączenia między SQL i R oraz eksplorowania tabel danych zawartych w SQL

Pierwszym krokiem w odpytywaniu bazy danych SQL przez R jest ustanowienie połączenia (rysunek 8.5). Rysunek demonstruje użycie pakietów DBI i RSQlite, które umożliwiają nawiązanie połączenia za pomocą funkcji dbConect(). Wynik połączenia i tabele danych, które są ujawniane za pośrednictwem wspomnianego połączenia, są następnie eksportowane za pomocą funkcji dbListTables(), która drukuje listę wszystkich danych znalezionych za pośrednictwem połączenia: Album, Artist (Artysta), Customer (Klient), Employee (Pracownik), Genre (Gatunek), Invoice (Faktura), InvoiceLine (WierszFaktury), MediaType (TypNośnika), Playlist (ListaOdtwarzania), PlaylistTrack (ŚcieżkaListyOdtwarzania), Track (Utwór).

Po nawiązaniu połączenia istnieje szereg możliwych analiz, które można przeprowadzić w zależności od celu biznesowego i przyszłego wykorzystania danych wyników. Tutaj, ze względu na ograniczenia przestrzenne, zademonstrujemy tylko ułamek możliwej analizy danych, z małym fragmentem kodu i zestawem reguł kodu potrzebnych do wyodrębnienia informacji z SQL. Kod wykonuje zapytanie do bazy danych za pośrednictwem R i wyświetla

najlepiej sprzedające się albumy, ich autorów i liczbę sprzedanych egzemplarzy (rysunek 8.6). Tabela 8.2 przedstawia wyniki kwerendy danych za pomocą fragmentu kodu na rysunku 8.6.

```
## Choose a Album table from the database
```{r}
query_album <- "SELECT * FROM Album LIMIT 10"
data_album <- dbGetQuery(con, query_album)
print(data_album)
```

Query to get album details along with artist names
```{r}
query_album_artist <- "
SELECT Album.AlbumId, Album.Title AS AlbumTitle, Artist.Name AS ArtistName
FROM Album
JOIN Artist ON Album.ArtistId = Artist.ArtistId
LIMIT 10"

data_album_artist <- dbGetQuery(con, query_album_artist)
print(data_album_artist)
```

Query to get the top-selling albums along with artist names
```{r}
query_top_selling_albums <- "
SELECT
    Album.Title AS AlbumTitle,
    Artist.Name AS ArtistName,
    SUM(InvoiceLine.Quantity) AS TotalQuantitySold
FROM
    InvoiceLine
JOIN
    Track ON InvoiceLine.TrackId = Track.TrackId
JOIN
    Album ON Track.AlbumId = Album.AlbumId
JOIN
    Artist ON Album.ArtistId = Artist.ArtistId
GROUP BY
    Album.AlbumId, Album.Title, Artist.Name
ORDER BY
    TotalQuantitySold DESC
LIMIT 10"

# Execute the query
top_selling_albums <- dbGetQuery(con, query_top_selling_albums)
knitr::kable(top_selling_albums)
```
```

Rysunek 8.6 Fragment kodu do zapytania SQL za pośrednictwem R i określenia 10 najlepiej sprzedających się albumów

Tabela 8.2 10 najlepiej sprzedających się albumów w cyfrowym sklepie Chinook

Tytuł albumu	Nazwa artysty	Liczba sprzedan
Minha Historia	Chico Buarque	27
Greatest Hits	Lenny Kravitz	26

Tytuł albumu	Nazwa artysty	Liczba sprzedan
Unplugged	Eric Clapton	25
Acústico	Titãs	22
Greatest Kiss	Kiss	20
Prenda Minha	Caetano Veloso	19
Chronicle, Vol. 2	Creedence Clearwater Revival	19
My Generation - The Very Best Of The Who	The Who	19
International Superhits	Green Day	18
Chronicle, Vol. 1	Creedence Clearwater Revival	18

BIBLIOGRAFIA DO ROZDZIAŁU 8.

- Holsapple, C., Lee-Post, A., & Pakath, R. (2014). A unified foundation for business analytics. Decision Support Systems, 64, 130-141.
- Kohavi, R., Rothleder, N. J., & Simoudis, E. (2002). Emerging trends in business analytics. Communications of the ACM, 45(8), 45-48.
- R Core Team. (2019). An introduction to R: Notes on R, a programming environment for data analysis and graphics. The R Foundation.
- RStudio. (2024). RStudio IDE cheatsheet: UI panes. Posit. https://docs.posit.co/ide/user/ide/guide/ui/ui-panes.html
- Torfs, P., & Brauer, C. (2014). A (very) short introduction to R. Hydrology and Quantitative Water Management Group, Wageningen University, The Netherlands, 1-12.

9. Prognozowanie popytu, wizualizacja i inżynieria cech szeregów czasowych w łańcuchach dostaw

Czym jest prognozowanie popytu? Jak skutecznie wizualizować dane o klientach i wyciągać z nich wnioski? Jak przeprowadzić inżynierię cech szeregów czasowych?

Na te i podobne pytania postaramy się udzielić odpowiedzi w poniższym rozdziale.

9.1 Czym jest popyt i prognozowanie popytu?

Ostateczny popyt klienta wprawia w ruch cały łańcuch dostaw (Syntetos i in., 2016). W związku z tym popyt klienta jest kluczowym elementem planowania wszystkich procesów logistycznych w łańcuchu dostaw, a zatem określenie poziomów popytu klienta jest bardzo interesujące dla menedżerów łańcucha dostaw. Prognozowanie popytu jest niezbędnym działaniem do planowania i harmonogramowania działań logistycznych w ramach obserwowanego łańcucha dostaw (Mircetic i in., 2017). Dokładne modele prognozowania popytu bezpośrednio wpływają na obniżenie kosztów logistycznych, ponieważ zapewniają ocenę popytu klientów z wyprzedzeniem czasowym potrzebnym do planowania i działania(Mircetic i in., 2016). Prognozowanie w łańcuchach dostaw wykracza poza operacyjne zadanie ekstrapolacji wymagań popytu na jednym poziomie. Obejmuje ono złożone kwestie, takie jak koordynacja łańcucha dostaw i dzielenie się informacjami między wieloma zainteresowanymi stronami (Syntetos i in., 2016).

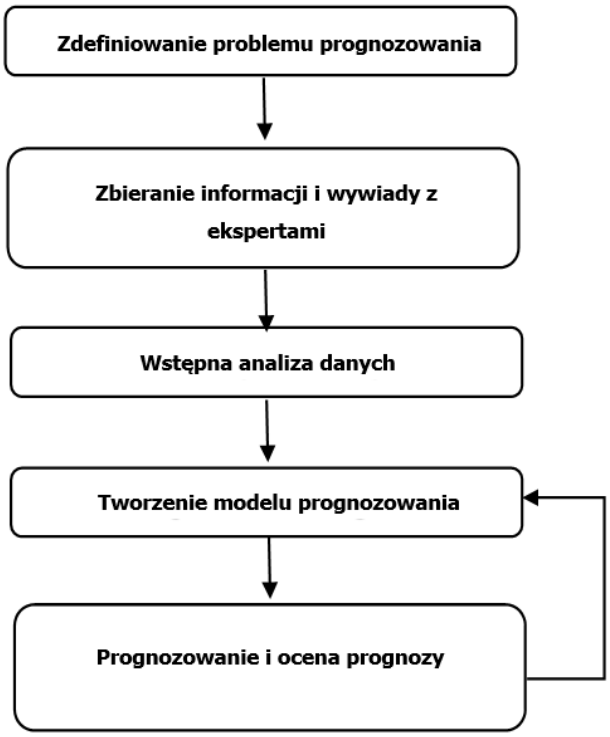
Popyt klientów i towarzyszące mu prognozy mają kluczowe znaczenie dla łańcuchów dostaw (SC), ponieważ zapewniają podstawowe dane wejściowe do planowania i kontroli wszystkich obszarów funkcjonalnych, w tym logistyki, marketingu, produkcji itp. Gdyby ostateczny popyt konsumentów był stały lub znany z dużym wyprzedzeniem, wówczas działanie łańcucha dostaw byłoby prostym (wstecznym) planowaniem. Jednak popyt nie jest znany i dlatego musi być prognozowany. To właśnie niepewność związana z tym popytem sprawia, że zarządzanie łańcuchem dostaw jest bardzo trudne (Syntetos i in., 2016). Na skuteczność

prognozowania popytu wpływa nieodłączna niepewność w szeregach czasowych popytu, które występują w łańcuchach dostaw (Rostami-Tabar, 2013). W związku z tym zajęcie się i zrozumienie tych niepewności jest głównym wyzwaniem dla menedżerów podczas koordynowania i planowania operacji w ramach łańcuchów dostaw (Mircetic, 2018).

Niepewność popytu jest jednym z najważniejszych wyzwań dla współczesnych łańcuchów dostaw. Niedawna pandemia COVID-19 jeszcze bardziej podkreśliła tę kwestię, powodując powszechne zakłócenia, które skomplikowały planowanie i kontrolę łańcucha dostaw (Nikolopoulos i in., 2020;). Prognozowanie popytu w łańcuchach dostaw często wiąże się z przewidywaniem popytu na wiele produktów. Osoby zajmujące się prognozowaniem w łańcuchach dostaw zazwyczaj ekstrapolują dane szeregów czasowych dla każdej jednostki magazynowej indywidualnie. Na przykład sprzedawca detaliczny może wykorzystywać dane z punktów sprzedaży do generowania prognoz na poziomie poszczególnych sklepów (Mircetic i in., 2022).

9.2 Etapy prognozowania popytu w łańcuchach dostaw

Zgodnie z powyższymi stwierdzeniami i wnioskami dotyczącymi znaczenia popytu i prognozowania popytu dla łańcuchów dostaw, ważne jest, aby postępować zgodnie z określonymi krokami podczas opracowywania prognoz w łańcuchach dostaw (rysunek 9.1).



Rysunek 9.1 Podstawowe kroki dla prawidłowego wdrożenia prognoz w firmie

Źródło: Makridakis i in., 1998; Makridakis i in., 1983

Każdy z etapów przedstawionych na rysunku 9.1 ma swoje zalety i przyczynia się do tworzenia wiarygodnych i użytecznych (zorientowanych biznesowo) prognoz. W związku z tym zdefiniowanie problemu jest często najtrudniejszą częścią prognozowania i wymaga zrozumienia, w jaki sposób prognozy będą wykorzystywane, a także roli funkcji prognozowania w obserwowanej firmie. Prognozujący powinien poświęcić dużo czasu na komunikację z wszystkimi osobami zaangażowanymi w gromadzenie danych, utrzymanie bazy danych i wykorzystywanie prognoz do planowania przyszłości. Jednym z głównych czynników obciążających definicję problemu jest to, w jaki sposób ostateczna prognoza będzie wykorzystywana w codziennych operacjach logistycznych (jaka platforma, projekt oprogramowania, interfejs użytkownika itp.).

Na etapie gromadzenia informacji zawsze potrzebne są co najmniej dwa rodzaje informacji: dane statystyczne i zgromadzona wiedza osób, które gromadzą dane i wykorzystują prognozy. W praktyce często trudno jest uzyskać dane historyczne, aby stworzyć dobry model statystyczny. Ponadto istnieje duże nieporozumienie co do tego, czym są dane dotyczące popytu i co można wykorzystać jako ich zamiennik. Istnieje zła praktyka wykorzystywania danych o wysyłkach i dostawach jako zamiennika danych o popycie,

co tylko pogorszy proces podejmowania decyzji w oparciu o prognozy sporządzone na podstawie łatwego rodzaju danych. Dane o sprzedaży są jedynym wiarygodnym zamiennikiem danych o popycie (Syntetos i in., 2016), chociaż to uproszczenie nie jest idealne, szczególnie w łańcuchach dostaw, w których występuje wiele sytuacji braku zapasów.

Na etapie wstępnej analizy danych zaleca się, aby zawsze rozpoczynać analizę danych od reprezentacji graficznych, aby odpowiedzieć na następujące pytania. Czy istnieją spójne wzorce? Czy istnieje znaczący trend? Czy istnieje zauważalna sezonowość? Czy istnieją dowody na cykle koniunkturalne? Jak silne są zależności między zmiennymi? Są to pytania, na które prosta grafika może dostarczyć odpowiedzi i umożliwić dalszą analizę danych poprzez zawężenie zakresu modeli, które należy zastosować do odkrytych cech popytu. Zwykle prosty model, określony w ten sposób, może pokonać bardziej wyrafinowane i skomplikowane (Rostami-Tabar & Mircetic, 2023).

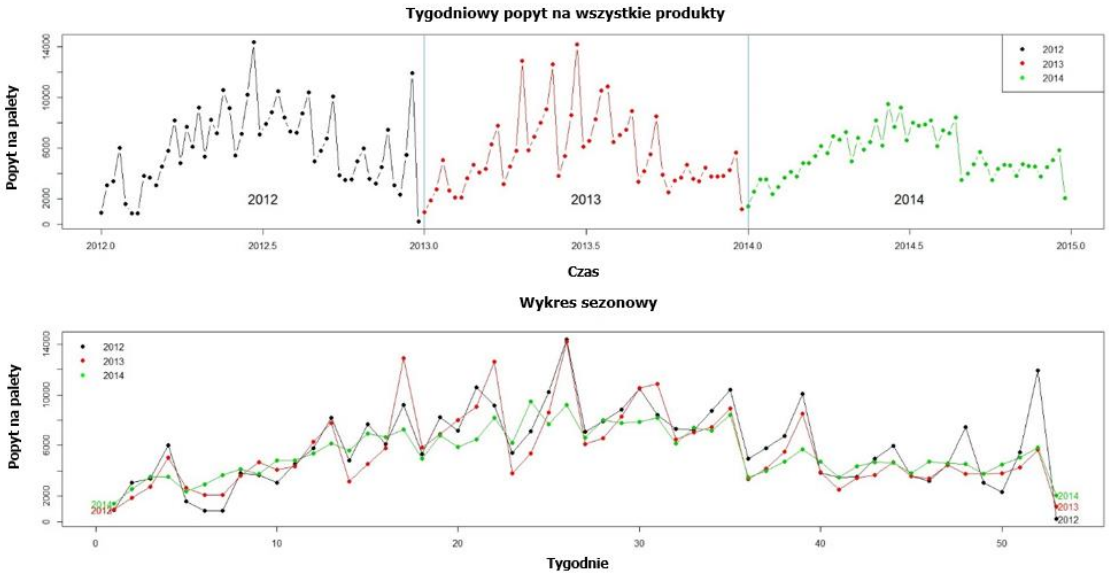
Wybór i tworzenie modeli prognostycznych jest najważniejszym krokiem podczas budowy modelu prognostycznego. Wybór modelu zależy od kilku czynników, z których najważniejsze to dostępność danych historycznych oraz korelacja między zmiennymi zależnymi i niezależnymi. Podczas wyboru modelu często porównuje się dwa lub trzy potencjalne modele. Każdy model jest sztuczną konstrukcją opartą na zestawie założeń (jawnych i ukrytych) i generalnie obejmuje jeden lub więcej parametrów, które muszą zostać utworzone przy użyciu znanych danych historycznych. W poniższym rozdziale przedstawiony zostanie proces rozwoju i zastosowania modelu ARIMA, jako jednego z najlepszych i najpopularniejszych modeli.

Ocena modelu prognostycznego jest krokiem, który mierzy użyteczność stworzonego modelu. Po wybraniu modelu prognostycznego i oszacowaniu jego parametrów, model jest wykorzystywany do tworzenia prognoz. Dokładność modelu jest oceniana za pomocą różnych statystyk podobnie jak trafność prognoz, ale ważne jest również testowanie prognoz za pomocą miar implikacji biznesowych (tj. wskaźników użyteczności).

9.3 Prognozowanie popytu w przemyśle spożywczym

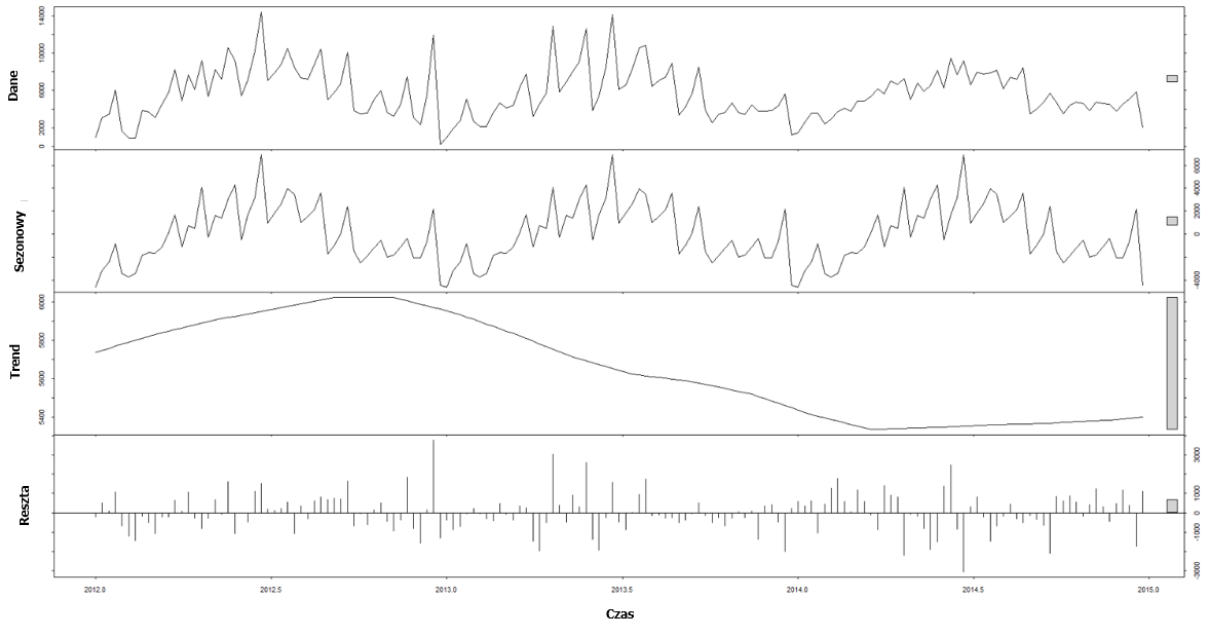
Dane dotyczące konsumpcji wszystkich produktów obserwowanego przedsiębiorstwa z branży spożywczej przedstawiono jako tygodniowy popyt obejmujący okres od stycznia 2012 r. do grudnia 2014 r. (Wykres na rysunku 9.2). Oś x reprezentuje czas, podczas gdy

wartości popytu są przedstawione na osi y. Dane te są przedstawione w odstępach tygodniowych, ponieważ odpowiada to okresowi, w którym prowadzona jest dostawa do końcowych punktów sprzedaży. W związku z tym kierownictwo firmy koncentruje się na prognozowaniu tygodniowej konsumpcji na rynku.



Rysunek 9.2 Dane dotyczące popytu dla obserwowanej firmy spożywczej

Rysunek 9.2 pokazuje kilka ważnych cech i właściwości danych. Po pierwsze, dane mają silny charakter sezonowy, a szczyt sprzedaży przypada na połowę roku (miesiące letnie). Po drugie, wykres sezonowy (dolny wykres na rysunku 9.2) pokazuje znaczącą zmianę wzorca trendu spadkowego w 2014 roku! Jest to bardzo istotna charakterystyka dla wyboru właściwego modelu prognozowania i dla kierowników wyższego szczebla w firmie, ponieważ ujawnia znaczny spadek konsumpcji i utratę rynku. Aby dokładniej zbadać zaobserwowane cechy, zastosowano metodologię dekompozycji trendu i sezonowości (STL) (rysunek 9.3). Dekompozycja STL dzieli pierwotne wzorce popytu na trzy komponenty: sezonowy, trend i komponent pozostały



Rysunek 9.3 Dekompozycja STL danych dotyczących popytu

STL ujawnił, że obserwowane szeregi czasowe mają charakter addytywny co oznacza, że wahania wokół krzywej trendu-cyklu nie zwiększają się znacząco w czasie. W rezultacie transformacje Box-Cox nie były konieczne dla surowych szeregów czasowych. Dekompozycja ujawniła, że składnik sezonowy jest dominujący w obserwowanej serii, wykazując duże wahania w ciągu jednego roku. Biorąc pod uwagę ograniczoną liczbę lat obserwacji, trudno jest zidentyfikować cykl koniunkturalny. Dekompozycja wykazała również, że trend w szeregu jest minimalny, z malejącym wzorcem rozpoczynającym się od połowy 2013 roku.

Te zidentyfikowane cechy stanowiły ważne dane wejściowe podczas procesu projektowania odpowiedniego modelu prognostycznego. W tym celu wybrano model S-ARIMA (ang. Seasonal Autoregressive Integrated Moving Average). Model S-ARIMA jest bardzo skuteczny w prognozowaniu, ponieważ łączy w sobie zarówno komponenty autoregresyjne, jak i średnie ruchome, wraz z różnicowaniem w celu uczynienia danych stacjonarnymi. Model ten jest szczególnie skuteczny w wychwytywaniu i modelowaniu wzorców sezonowych w danych szeregów czasowych, co czyni go idealnym dla branż o cyklicznych wzorcach popytu, takich jak przemysł spożywczy. Dodatkowo, zdolność S-ARIMA do radzenia sobie ze złożonymi strukturami sezonowymi i trendami pozwala na dokładniejsze i bardziej wiarygodne prognozy, które są kluczowe dla skutecznego planowania łańcucha dostaw

i zarządzania zapasami. Strukturalna postać modelu S-ARIMA została przedstawiona w ównaniu (9.1).

$$\Phi(B^m)\phi(B)(1-B^m)^D(1-B)^d y_t = c + \Theta(B^m)\theta(B)\varepsilon_t, \tag{9.1}$$

gdzie $\Phi(z)$ i $\Theta(z)$ są wielomianami odpowiednio rzędu *P* i *Q*, z których każdy nie zawiera pierwiastków wewnątrz okręgu jednostkowego. *B* jest operatorem przesunięcia wstecznego używanym do opisu procesu różniczkowania, tj. $By_t=y_{t-1}$. Jeśli $c\neq 0$, istnieje implikowany wielomian rzędu *d+D* w funkcji prognozy. Ponieważ model S-ARIMA jest modelem wysoce sparametryzowanym, kluczową kwestią podczas korzystania z modelu S-ARIMA jest wybór odpowiedniego rzędu modelu, co wiąże się z określeniem wartości of *p, q, P, Q, D*, i *d*. Jeżeli *d* oraz *D* są nieznane, rzędy *p, q, P*, i *Q* można wybrać za pomocą kryterium informacyjnego, takiego jak Kryterium Informacyjne Akaike (AIC) lub Bayesowskie Kryterium Informacyjne (BIC). Wzory na AIC i BIC są następujące:

$$AIC = -2\log(L) + 2(k);$$

$$BIC = N\log\left(\frac{SSE}{N}\right) + (k+2)\log(N) \tag{9.2}$$

gdzie $k=p+q+P+Q+1$, jeśli uwzględniono stałą, a 0 w przeciwnym razie, *L* jest zmaksymalizowanym prawdopodobieństwem modelu dopasowanego do zróżnicowanych danych, *SSE* jest sumą błędów podniesionych do kwadratu, *N* jest liczbą obserwacji wykorzystanych do estymacji, a *k* jest liczbą predyktorów w modelu

W celu określenia optymalnego zestawu parametrów Hyndman i Khandakar (2007) zaproponowali test pierwiastka jednostkowego Canova-Hansena i KPSS w następujących krokach:

- Użyj testu Canova-Hansena do określenia *D* w ramach ARIMA,
- Wybierz *d*, stosując kolejny test pierwiastka jednostkowego KPSS na danych zróżnicowanych sezonowo (jeśli $D = 1$) lub na danych oryginalnych (jeśli $D = 0$),
- Wybierz optymalne wartości dla *p, q, P* i *Q* poprzez minimalizację AIC.

9.4 Opracowanie modelu prognostycznego S-ARIMA

Opracowanie modelu prognostycznego S-ARIMA

Zgodnie z powyższą procedurą przetestowano kilka ustawień parametrów, a ich wydajność przedstawiono w tabeli 9.1. Do testowania wydajności różnych modeli zastosowano kilka miar: średni bezwzględny błąd procentowy (MAE), średni błąd kwadratowy (RMSE), średni bezwzględny błąd skalowany (MASE), AIC i BIC (równania 2 i 3).

$$MAPE = \frac{100}{L}\sum_{l=1}^{L}\left|\frac{e_l}{y_l}\right|;$$

;

$$RMSE = \sqrt{\frac{1}{L}\sum_{l=1}^{L} e_l^2}; \tag{9.3}$$

,

gdzie e_l są wartościami rezydualnymi, ale liczonymi dla L okresów następnych po N okresach stanowiących próbkę, z której danych otrzymano szacunek modelu: $e_l = y_l - y_l^*$, Tak liczone reszty nazywa się błędami prognoz y_l^*, czyli wartości teoretycznych z predykcji poza próbkę. Do miar trafności prognoz zaliczamy także średni absolutny błąd skalowany:

.

$$MASE = \frac{\frac{1}{L}\sum_{l=1}^{L}|e_l|}{z}; \tag{9.4}$$

Gdzie z stanowi współczynnik skalowania danych sezonowych o długości cyklu S

$$z = \frac{1}{N-m}\sum_{t=m+1}^{N}|y_t - y_{t-S}|. \tag{9.5}$$

Im niższa wartość MASE tym trafniej prognozuje model. Najpopularniejszymi miarami dla menedżerów w łańcuchach dostaw są RMSE i MAPE, ponieważ zapewniają one „poczucie" tego, jak dobrze model działa w liczbach rzeczywistych (RMSE) i procentach (MAPE).

Tabela 9.1 Wydajność modeli S-ARIMA z różnymi ustawieniami parametrów

Modele[a]	RMSE	MAPE	MASE	AIC	BIC
S-ARIMA $(5,0,1)(1,0,0)_{52}$[b]	1023	14,18 %	0,60	86,62	110,50
S-ARIMA $(4,0,0)(1,0,0)_{52}$[c]	1041	14,53 %	0,62	86,73	105,31
S-ARIMA $(4,0,0)(0,1,1)_{52}$[d]	1882	22,12 %	1,05	41,74	53,560
S-ARIMA $(4,0,1)(1,0,0)_{52}$[e]	1050	14,18 %	0,60	88,23	109,47
S-ARIMA $(0,0,1)(0,1,0)_{52}$[f]	1797	21,42 %	1,02	36,52	40,460

[a] Błędy modelu są obliczane na zestawie danych testowych.

[b] Szczegóły dotyczące modelu S-ARIMA $(5,0,1)(1,0,0)_{52}$ są podane poniżej.

[c] $(1-0{,}39B+0{,}06B+0{,}04B-0{,}46B)(1-0{,}65B^{52})y_t = 8{,}52.$

[d] $(1-0{,}29B+0{,}19B-0{,}05B-0{,}19B)(1-B^{52})y_t = (1+0{,}12B^{52})e_t.$

[e] $(1-0{,}29B+0{,}01B+0{,}05B-0{,}48B)(1-0{,}65B^{52})y_t = 8{,}52+(1+0{,}13B)e_t.$

[f] $(1-B^{52})y_t = (1+0{,}3B)e_t.$

Tabela 9.1 pokazuje, że model S-ARIMA $(5,0,1)(1,0,0)_{52}$, gdzie cykl sezonowy trwa $S = 52$ tygodnie, przewyższył konkurencyjne modele, osiągając najniższe błędy RMSE, MAPE i MASE. Postać modelu S-ARIMA $(5,0,1)(1,0,0)_{52}$ przedstawiono w równaniu (9.4), gdzie $\phi_1 = -0{,}5421$, $\phi_2 = 0{,}2962$, $\phi_3 = -0{,}099$, $\phi_4 = 0{,}3974$, $\phi_5 = 0{,}4994$, $\Phi_1 = 0{,}9558$, $c = 8{,}523$, i $\Theta_1 = 0{,}6345$.

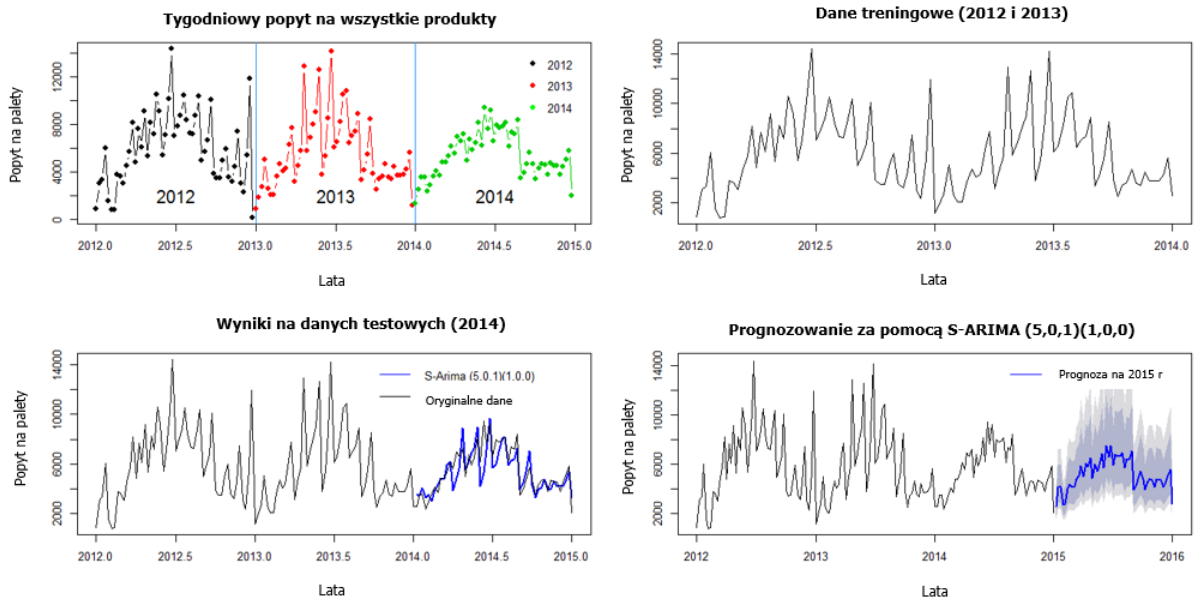
$$\left(1-\phi_1 B-\phi_2 B-\phi_3 B-\phi_4 B-\phi_5 B\right)\left(1-\Phi_1 B^{52}\right)y_t = c+\left(1+\theta_1 B\right)e_t\,, \tag{9.6}$$

Podobne wyniki zaobserwowano w przypadku modeli S-ARIMA $(4,0,0)(1,0,0)_{52}$ i S-ARIMA $(4,0,1)(1,0,0)_{52}$. Podczas wykonywania prognoz model S-ARIMA $(5,0,1)(1,0,0)_{52}$ generuje średnio błąd 1023 sztuk, co przekłada się na 14,18%. Wyniki te podkreślają znaczenie starannego doboru warunków do uwzględnienia w modelu S-ARIMA, ponieważ nie ma znaczących różnic między wynikami modelu z różnymi parametrami. Ocena wykazała, że modele zawierające składniki autoregresyjne (*p, P*) i średniej ruchomej (*q, Q*) były bardziej skuteczne w prognozowaniu spożycia napojów niż te zawierające sezonowe lub niesezonowe różnicowanie. Ponadto przegląd porównawczy ujawnił pewne nieoczekiwane ustalenia, takie jak modele uwzględniające sezonowe różnicowanie S-ARIMA $(4,0,0)(0,1,1)_{52}$

i S-ARIMA $(0,0,1)(0,1,0)_{52}$, które wypadły gorzej niż prosty średni naiwny model prognostyczny, o czym świadczą ich błędy MASE przekraczające jeden.

9.5 Prognozy dotyczące przyszłego popytu.

Na rysunku 9.4 przedstawiono wydajność metody S-ARIMA $(5,0,1)(1,0,0)_{52}$. Lewy górny panel przedstawia początkowe dane popytu, pokolorowane w celu łatwiejszego rozróżnienia różnych lat marketingowych. Prawy górny panel to dane wejściowe do modelu S-ARIMA $(5,0,1)(1,0,0)_{52}$, reprezentujące dane treningowe, na podstawie których parametry w modelu S-ARIMA są określane zgodnie z procedurą opisaną w podrozdziale 9.3. Jest to bardzo ważne, aby zrozumieć, jak trudne jest zadanie modelu prognostycznego, ponieważ w tym przypadku dysponuje on dwoma latami danych jako danymi wejściowymi i musi prognozować przyszły popyt z rocznym wyprzedzeniem! Jest to dość powszechny scenariusz w łańcuchach dostaw i logistyce, ponieważ istnieje niepisana zasada, że firmy przechowują historię swoich danych przez trzy lata, po czym je odrzucają.



Rysunek 9.4 Trening, test i prognozowany popyt modelu S-ARIMA $(5,0,1)(1,0,0)_{52}$

Lewy dolny panel na rysunku 9.4 przedstawia wydajność obserwowanego modelu. Wydajność modelu jest już przedstawiona w tabeli 9.1 za pomocą różnych miar statystycznych, ale dla menedżerów zwykle trudno jest uzyskać opinię, jak dobry lub zły jest model. W tym celu panel graficznie przedstawia wydajność modelu. Można argumentować, że model dość dobrze podąża za danymi testowymi i w większości okresów wykazuje

doskonałą wydajność. W celu wygenerowania przyszłych prognoz, model S-ARIMA $(5,0,1)(1,0,0)_{52}$ został ponownie dopasowany poprzez dodanie danych z 2014 roku do danych treningowych. Następnie model wygenerował 52-tygodniowe prognozy na rok 2015. Prognozy na rok 2015 przedstawiono w lewym dolnym panelu. Prognozom towarzyszą 80% i 95% przedziały predykcyjne, które pokazują, że możliwe przyszłe prognozy różnią się od średnich przewidywanych wartości. Model przewiduje dalszy spadek popytu, który rozpoczął się w 2014 roku. Przyczyny tego spadku mogą być różne i powinny być dalej badane przez menedżerów na strategicznym poziomie firmy.

BIBLIOGRAFIA DO ROZDZIAŁU 9.

- Hyndman, R., & Khandakar, Y. (2007). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software, 27*(3). http://www.jstatsoft.org/v27/i03
- Makridakis, S., Wheelwright, S., & Hyndman, R. J. (1998). *Forecasting: Methods and applications* (3rd ed.). New York: John Wiley & Sons.
- Makridakis, S., Wheelwright, S., & McGee, V. (1983). *Forecasting: Methods and applications* (2nd ed.). New York: John Wiley & Sons.
- Mircetic, D. (2018). *Boosting the performance of top-down methodology for forecasting in supply chains via a new approach for determining disaggregating proportions*. University of Novi Sad.
- Mircetic, D., Nikolicic, S., Maslaric, M., Ralevic, N., & Debelic, B. (2016). Development of S-ARIMA model for forecasting demand in a beverage supply chain. *Open Engineering, 6*(1).
- Mircetic, D., Rostami-Tabar, B., Nikolicic, S., & Maslaric, M. (2022). Forecasting hierarchical time series in supply chains: An empirical investigation. *International Journal of Production Research, 60*(8), 2514-2533.
- Mircetic, D., Nikolicic, S., Stojanovic, D., & Maslaric, M. (2017). Modified top-down approach for hierarchical forecasting in a beverage supply chain. *Transportation Research Procedia, 22*, 193–202.
- Nikolopoulos, K., Punia, S., Schäfers, A., Tsinopoulos, C., & Vasilakis, C. (2020). Forecasting and planning during a pandemic: COVID-19 growth rates, supply chain disruptions, and governmental decisions. *European Journal of Operational Research, 290*(1), 99–115.
- Rostami-Tabar, B. (2013). *ARIMA demand forecasting by aggregation*. Université Sciences et Technologies Bordeaux I.
- Rostami-Tabar, B., & Mircetic, D. (2023). Exploring the association between time series features and forecasting by temporal aggregation using machine learning. *Neurocomputing, 548*, 126376.
- Syntetos, A. A., Babai, Z., Boylan, J. E., Kolassa, S., & Nikolopoulos, K. (2016). Supply chain forecasting: Theory, practice, their gap and the future. *European Journal of Operational Research, 252*(1), 1-26.

10. Sztuczna inteligencja i uczenie maszynowe w łańcuchu dostaw

Czym jest sztuczna inteligencja (AI)? Czy naprawdę jest to „żywe" stworzenie zdolne do myślenia i podejmowania własnych decyzji w oparciu o swój umysł, przeszłe doświadczenia, etykę, przekonania itp. W jaki sposób jest ona powiązana z uczeniem maszynowym (ML)? Czy AI i ML to to samo?

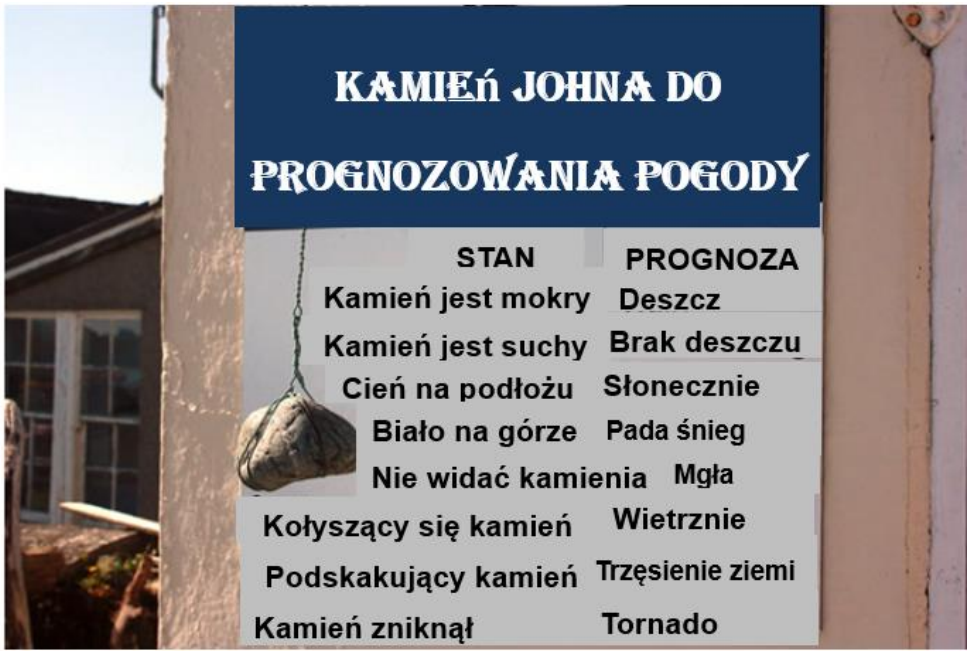
Jakie narzędzia są wykorzystywane w AI i ML? Jaką rolę odgrywają AI i ML w kontekście biznesowym i jak można je wykorzystać w codziennych operacjach biznesowych i optymalizacjach? Czy istnieją konkretne architektury i przykłady zastosowania AI i ML w łańcuchach dostaw?

Na te i podobne pytania postaramy się udzielić odpowiedzi w następnym rozdziale, zamykając go rzeczywistym przykładem zastosowania algorytmów AI i ML w magazynie dystrybucyjnym.

10.1 Czym jest sztuczna inteligencja (AI)?

Dziedzina sztucznej inteligencji rozpoczęła się w latach pięćdziesiątych XX wieku, kiedy to informatycy zapytali: „Czy komputery mogą myśleć jak ludzie"? Naukowcy w tamtym czasie byli entuzjastycznie nastawieni do możliwości uczenia komputerów wykonywania złożonych zadań i w tym celu opracowali zestaw różnych algorytmów. Definicję tej dziedziny można określić jako wysiłek na rzecz automatyzacji zadań intelektualnych normalnie wykonywanych przez ludzi (Chollet, 2021). Najbardziej znaczące algorytmy w dziedzinie sztucznej inteligencji pochodzą obecnie z uczenia maszynowego i głębokiego uczenia, które są podzbiorami sztucznej inteligencji. Oprócz uczenia maszynowego i głębokiego uczenia, sztuczna inteligencja obejmuje wiele algorytmów nieuczenia się. Co więcej, możemy argumentować, że tego rodzaju algorytmy były bardziej dominujące we wczesnej fazie rozwoju sztucznej inteligencji. W związku z tym jest to część sztucznej inteligencji znana jako symboliczna sztuczna inteligencja, która opiera się na założeniu, że ludzki poziom wydajności i inteligencji można osiągnąć poprzez programowanie komputerów za pomocą dużego

zestawu wyraźnych reguł rozwiązywania obserwowanego problemu. Podejście to zapewniało doskonałe wyniki w przypadku problemów logicznych, które były dobrze zdefiniowane, takich jak komputer grający w szachy, ale okazało się skomplikowane w przypadku bardziej złożonych problemów. Rzeczywisty świat okazał się znacznie bardziej skomplikowany niż wszystkie jawne reguły programowania, które można było wprowadzić do komputera. Podejście to koncentruje się wokół idei dla danej sytuacji zrób to lub tamto (reguły if-then). Jest to łatwe do zrozumienia podejście, ale z drugiej strony bardzo czasochłonne i czasami bardzo trudne do określenia wszystkich możliwych scenariuszy, które należy wprowadzić do programu. Jako zabawny przykład, ale dobrze ilustrujący dany temat, przedstawiony na rysunku 10.1, który pokazuje kilka scenariuszy określania prognozy na podstawie stanu kamienia.



Rysunek 10.1 Reguły programowania Jeżeli – To

Źródło: Gibbs, M. 2019

Dziedzina symbolicznej sztucznej inteligencji zyskała największą popularność w latach 80. wraz z pojawieniem się systemów eksperckich (ES). ES reprezentują podzbiór systemów wspomagania decyzji (DSS) (Turban, 1998), koncentrując się na dostarczaniu skomputeryzowanych zdolności decyzyjnych podobnych do tych, które posiada ludzki specjalista w danej dziedzinie. Systemy te są stworzone do rozwiązywania skomplikowanych problemów poprzez zastosowanie szeregu reguł lub algorytmów, które symulują ludzkie

procesy rozumowania. Olson i Courtney opisują systemy eksperckie (ES) jako programy komputerowe, które symulują ludzkie procesy myślowe w celu podejmowania decyzji w określonej dziedzinie, włączając pewien stopień sztucznej inteligencji, aby dopasować wnioski, do których doszedłby ludzki ekspert (Olson & Courtney, 1992). Komponent ES jest szczególnie przydatny do wspierania decydentów w obszarach wymagających specjalistycznej wiedzy (Turban i in., 2005). Zasadniczo ES przechwytuje wiedzę specjalistyczną od ludzkiego eksperta (lub innego źródła) i przekazuje ją do komputera. Technologia ta może pomóc decydentom lub w pełni ich zastąpić, co czyni ją jedną z najczęściej stosowanych i odnoszących sukcesy komercyjne form sztucznej inteligencji (Turban i in., 2005). Jednym z kluczowych powodów rozwoju ES jest rozpowszechnianie wiedzy eksperckiej wśród szerszego grona odbiorców (Jackson, 1999). W kolejnych podrozdziałach zademonstrujemy zastosowanie ES opartego na algorytmach AI i ML w centralnym magazynie jako części ogólnego systemu DSS dla menedżerów.

Obecnie dziedzina sztucznej inteligencji składa się z różnych podejść i algorytmów, ale najczęściej używane z nich opisano na rysunku 10.2. Odeszła ona od systemów ES, a ponowne pojawienie się tej dziedziny przypisuje się głównie algorytmom głębokiego uczenia, które odniosły znaczący sukces w ciągu ostatnich 12 lat w problemach rozpoznawania obrazów, rozpoznawania mowy, segmentacji obrazów, rozpoznawania twarzy itp. Głębokie uczenie wykorzystuje wiele warstw abstrakcji do identyfikacji złożonych wzorców w danych wielowymiarowych. Podejście to pozwoliło osiągnąć znaczący postęp w takich dziedzinach jak rozpoznawanie mowy i obrazów, odkrywanie leków i przetwarzanie języka naturalnego. Głębokie uczenie ma zdolność do automatycznego odkrywania odpowiednich cech i zmniejsza potrzebę interwencji człowieka w projektowanie cech, dzięki czemu jest bardzo wydajne w wykorzystywaniu dużych zbiorów danych i mocy obliczeniowej (LeCun, Bengio i Hinton, 2015.

Rysunek 10.2 Główne dziedziny i poddziedziny sztucznej inteligencji

Źródło: Athanasopoulou i in., 2022

Dużym krokiem w kierunku dzisiejszej sytuacji były badania nad klasyfikacją cyfr przeprowadzone przez Hintona, Osindero i Teha (2006), którym udało się osiągnąć ponad 98% dokładności w klasyfikacji bazy danych Modified National Institute of Standards and Technology (MNIST). Jednym ze sposobów myślenia o tym, jak sztuczna inteligencja przeszła od symbolicznej sztucznej inteligencji do uczenia maszynowego i jaka jest istota uczenia maszynowego, jest wyobrażenie sobie algorytmów uczenia maszynowego jako amorficznej masy, która kształtuje się zgodnie z pożądanymi wynikami. System reguł, który zmienia dane wejściowe w wyjściowe, zmienia się z problemu na problem i dostosowuje się do istniejącej sytuacji. Jego celem jest znalezienie reguł, które zautomatyzują zadanie poprzez wyszukiwanie wzorców statystycznych w danych. Takie podejście do rozwiązywania różnych problemów znacznie skraca czas konfiguracji systemu (w porównaniu do reguł if

(*jeżeli)* – then (*wtedy)*) i sprawia, że jest to bardziej uniwersalne podejście do rozwiązywania różnych problemów.

Bengio, Lecun i Hinton (2021) podkreślili, że przyszłość sztucznej inteligencji leży w głębokim uczeniu się i rewolucyjnym wpływie miękkiej uwagi i architektur transformatorowych w sztucznej inteligencji. Innowacje te pozwalają sieciom neuronowym dynamicznie koncentrować się na ważnych danych wejściowych i przechowywać informacje w zróżnicowanych pamięciach, znacznie poprawiając przetwarzanie sekwencyjne.

10.2 Czym jest ekosystem sztucznej inteligencji i uczenia maszynowego?

Ekosystem algorytmów AI i ML składa się z trzech kluczowych filarów:

- danych wejściowych,
- danych wyjściowych,
- funkcji kosztów.

Dane wejściowe reprezentują nagrania danych określonej cechy lub cech (w zależności od zaobserwowanego problemu). Dokładność danych wejściowych jest kluczowa dla budowania dokładnych algorytmów. Często nie ma to miejsca w rzeczywistych zastosowaniach i zwykle poświęca się dużo czasu i wysiłku na zbieranie danych, czyszczenie, porządkowanie, ujednolicanie, sprawdzanie fałszywych danych wejściowych itp. Oprócz dokładnych danych, innym ważnym aspektem charakterystyki danych wejściowych jest ich reprezentacja i kodowanie. Różne sposoby kodowania danych mogą ujawnić różne cechy danych i znacząco „pomóc" modelom ML w ujawnianiu ukrytych wzorców w danych. Tutaj widzimy piętę achillesową algorytmów ML. Często zbyt wiele uwagi poświęca się tworzeniu inteligentnych metod wydobywania informacji z danych (tj. samych algorytmów), podczas gdy niewystarczająca uwaga jest poświęcana danym wejściowym i ich związkom z danymi wyjściowymi. Ogólnie przyjmuje się za pewnik, że dane wejściowe mają związek przyczynowy z danymi wyjściowymi, co czasami wcale nie ma miejsca. Dlatego kolejnym dużym krokiem w rozwoju ML powinno być znalezienie lepszych sposobów gromadzenia, reprezentowania i kodowania danych.

Dane wyjściowe reprezentują pomiary konkretnego problemu, który próbujemy rozwiązać. W problemie klasyfikacji będzie to etykieta klasy. W regresji będzie to liczba rzeczywista,

którą próbujemy przewidzieć. W kontekście konkretnego problemu, w przypadku problemów z rozpoznawaniem mowy, danymi wyjściowymi mogą być wygenerowane przez człowieka transkrypcje plików dźwiękowych. W przypadku rozpoznawania obrazów, danymi wyjściowymi mogą być etykiety klas obrazów itp.

Funkcja kosztu reprezentuje sposób pomiaru wydajności AI i ML. Zasadniczo chcielibyśmy, aby odpowiedzi algorytmów pasowały do danych wyjściowych dla danych wejściowych. Funkcja kosztu jest również sygnałem zwrotnym do zestawu parametrów, które kierują pracą algorytmu, tj. pozwala na optymalizację ogólnej wydajności algorytmu poprzez proces uczenia się (znalezienie optymalnego zestawu parametrów). Proces uczenia się zazwyczaj obejmuje uczenie nadzorowane, w którym model jest szkolony na oznaczonych danych w celu zminimalizowania błędów przewidywania za pomocą technik takich jak stochastyczne opadanie gradientu i propagacja wsteczna. Umożliwia to modelowi skuteczne dostosowanie jego wewnętrznych parametrów, co prowadzi do poprawy wydajności w zadaniach takich jak wykrywanie i klasyfikacja obiektów (LeCun, Bengio, & Hinton, 2015).

10.3 Jakie narzędzia są wykorzystywane w uczeniu maszynowym?

Ogólnie rzecz biorąc, algorytmy uczenia maszynowego można podzielić na dwie główne kategorie: uczenie nadzorowane i nienadzorowane.

Uczenie nadzorowane obejmuje szkolenie algorytmów na etykietowanym zbiorze danych, w którym każdy punkt danych wejściowych jest sparowany z prawidłowym wynikiem. Ten jasny „obraz" tego, jaka powinna być prawidłowa odpowiedź dla danego wejścia, pozwala algorytmowi nauczyć się funkcji mapowania z wejść na wyjścia. W związku z tym znane są zarówno dane wejściowe, jak i wyjściowe (Athanasopoulou i in., 2022). Typowe zastosowania uczenia nadzorowanego obejmują zadania klasyfikacji (np. określanie, czy wiadomość e-mail jest spamem, czy nie) i zadania regresji (np. przewidywanie cen domów na podstawie różnych cech). Niektóre z najpopularniejszych algorytmów, które zostały sprawdzone w licznych zastosowaniach, to uogólnione modele addytywne, Random Forest, boosting, drzewa klasyfikacyjne i regresyjne, maszyny wektorów nośnych, rozszerzona regresja liniowa, regresja logistyczna, algorytm k-najbliższych sąsiadów, liniowa analiza dyskryminacyjna, lasso, sieci neuronowe, adaptacyjny system wnioskowania neurorozmytego itp. (Rostami-Tabar & Mircetic, 2023). Uczenie nadzorowane jest potężne, ponieważ

wykorzystuje dane z adnotacjami człowieka, aby osiągnąć wysoką dokładność przewidywań. Jednak jego skuteczność zależy w dużej mierze od jakości i ilości etykietowanych danych.

Z kolei uczenie bez nadzoru dotyczy zbiorów danych, w których brakuje etykietowanych odpowiedzi. W związku z tym algorytmy uczenia maszynowego bez nadzoru wykorzystują nieoznakowane zbiory danych, które zawierają tylko dane wejściowe (Athanasopoulou i in., 2022). W tym przypadku algorytm otrzymuje tylko dane wejściowe, a jego celem jest znalezienie podstawowych wzorców, struktur lub relacji w danych. Typowe techniki uczenia bez nadzoru obejmują grupowanie (np. grupowanie klientów według zachowań zakupowych) i redukcję wymiarowości (np. zmniejszenie liczby zmiennych w zbiorze danych przy jednoczesnym zachowaniu ważnych informacji). Uczenie bez nadzoru jest cenne w przypadku analizy danych eksploracyjnych i odkrywania ukrytych struktur w danych. Jest często używane, gdy etykietowane dane są rzadkie lub niedostępne.

Inną ważną kategorią, choć różniącą się od uczenia nadzorowanego i nienadzorowanego, jest uczenie ze wzmocnieniem. W tym przypadku algorytm uczy się poprzez interakcję ze środowiskiem i otrzymywanie informacji zwrotnych w postaci nagród lub kar. To podejście oparte na metodzie prób i błędów pomaga algorytmowi nauczyć się optymalnych działań w celu maksymalizacji skumulowanych nagród. Uczenie ze wzmocnieniem jest szeroko stosowane w takich dziedzinach jak robotyka, gry i systemy autonomiczne.

Jednym z najbardziej popularnych i skutecznych algorytmów ML jest gałąź sieci neuronowych. Sieci neuronowe istnieją od lat 50. ubiegłego wieku, ale zyskały popularność w latach 80. i w ciągu ostatnich 12 lat. Są one zbudowane na wzór biologicznych neuronów i sposobu, w jaki dzielą się one informacjami w mózgu, ale poza tym nie ma między nimi znaczących powiązań. Obecnie najczęściej stosowaną formą sieci neuronowych są sieci głębokiego uczenia, które reprezentują kilka warstw ukrytych między cechami wejściowymi i wyjściowymi, które wykonują kilka nieliniowych transformacji cech wejściowych. Ponieważ okazało się to bardzo skuteczne, uczenie głębokie jest obecnie jedną z najważniejszych poddziedzin ML (Chollet, 2021).

10.4 Studium przypadku

Ustalenia Wenzel, Smit i Sardesai (2019) dotyczące ML w zarządzaniu łańcuchem dostaw wskazują na rosnącą integrację aplikacji ML w różnych zadaniach SC. W związku z tym obserwowane studium przypadku przedstawia zastosowanie sztucznej inteligencji i uczenia

maszynowego w centralnym magazynie fabryki żywności (Mirčetić i in., 2016; Mircetic i in., 2014). W kompleksie fabrycznym znajduje się 30 wózków widłowych. Wózki widłowe są zaangażowane w różne operacje wewnątrz kompleksu, które są kluczowe dla operacji logistycznych w produkcji, magazynowaniu i wysyłce produktów. Centralny magazyn ma pojemność 11 100 miejsc paletowych i roczną produkcję od 300 000 do 350 000 palet. Obecnie fabryka zaopatruje około 20 000 supermarketów za pośrednictwem dostaw bezpośrednich.

Problem zaangażowania wózków widłowych jest związany z faktem, że nadmierne lub niedostateczne zaangażowanie wózków widłowych w różne procesy fabryczne prowadzi do znacznych strat finansowych i rynkowych. Obecnie proces podejmowania decyzji, gdzie i co będzie robił każdy wózek widłowy, opiera się na decyzjach ekspertów (menedżerów). Decyzje ekspertów opierają się na ich doświadczeniu, bez pomocy jakiegokolwiek systemu wspomagania decyzji (DSS). Liczne dowody empiryczne sugerują, że ludzka intuicyjna ocena i podejmowanie decyzji często nie są optymalne, szczególnie w warunkach złożoności i stresu (Druzdzel i Flynn, 2002). Podkreśla to znaczenie włączenia systemów wspomagania decyzji (DSS) celem wspomagania ekspertów w procesie podejmowania decyzji.

W tej aplikacji wybraliśmy kilka algorytmów ML, aby pomóc w optymalizacji pracy magazynu załadunkowego. Algorytmy ML zostały zebrane w unikalną strukturę decyzyjną, która służy jako DSS dla menedżerów i ekspertów w danej firmie. Co więcej, cały system decyzyjny DSS można postrzegać jako platformę sztucznej inteligencji, ponieważ stale przelicza on sugestie z kilku modeli ML (ile wózków widłowych należy użyć i które z nich) i automatycznie wybiera najlepsze z nich, biorąc pod uwagę dostarczone dane wejściowe operatorów.

Opis problemu

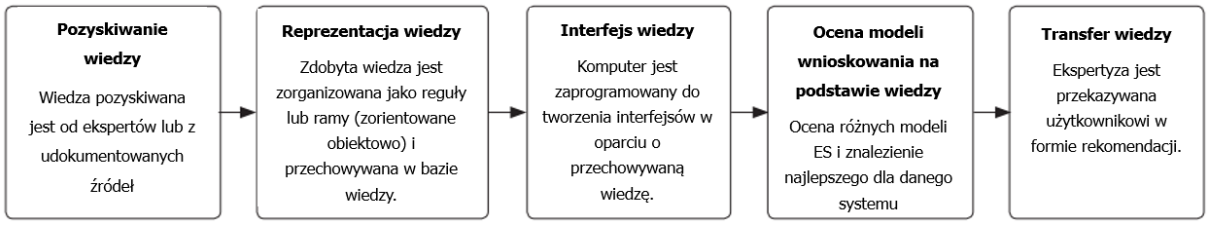
Proces załadunku ma kluczowe znaczenie dla logistyki magazynowej, wpływając na poziom obsługi rynku. Podczas wysyłek, ekspert magazynowy określa liczbę i wybór wózków widłowych do załadunku, kierując się trzema czynnikami: (1) ukończenie załadunku w określonych ramach czasowych, (2) zminimalizowanie zakłóceń w innych zadaniach wózków widłowych oraz (3) dostosowanie wykorzystania wózków widłowych do możliwości konserwacyjnych, które mogą obsługiwać dwa przeglądy jednocześnie. Każdy wózek widłowy przechodzi od czterech do pięciu przeglądów konserwacyjnych rocznie.

Wózki widłowe mają kluczowe znaczenie dla operacji załadunku, które muszą wspierać strategię marketingową firmy, zapewniając jednocześnie płynne działanie innych czynności.

Nieprawidłowe przydzielanie wózków widłowych może prowadzić do niepełnego wykorzystania zasobów lub zaszkodzić reputacji firmy i poziomowi świadczonych usług. Opóźnienia
w załadunku wiążą się z karami. Menedżer musi koordynować wykorzystanie wózków widłowych we wszystkich działaniach, aby uniknąć jednoczesnych przeglądów i zarządzać różnymi potrzebami konserwacyjnymi. Chociaż menedżerowie zazwyczaj podejmują trafne decyzje, praca w środowisku charakteryzującym się wysokim poziomem stresu może prowadzić do błędów. Dlatego potrzebne jest wykorzystanie systemu DSS, aby zwiększyć pewność
i niezawodność podejmowania decyzji.

AI i ML jako DSS dla centralnego magazynu

Pierwszym krokiem do wygenerowania systemów AI i ML jest pozyskanie stabilnego i poprawnego źródła wiedzy (bazy danych) oraz rzucenie światła na role biznesowe, które musi wspierać. Dlatego też rysunek 10.3 przedstawia metodologię budowy systemu AI i ML DSS.

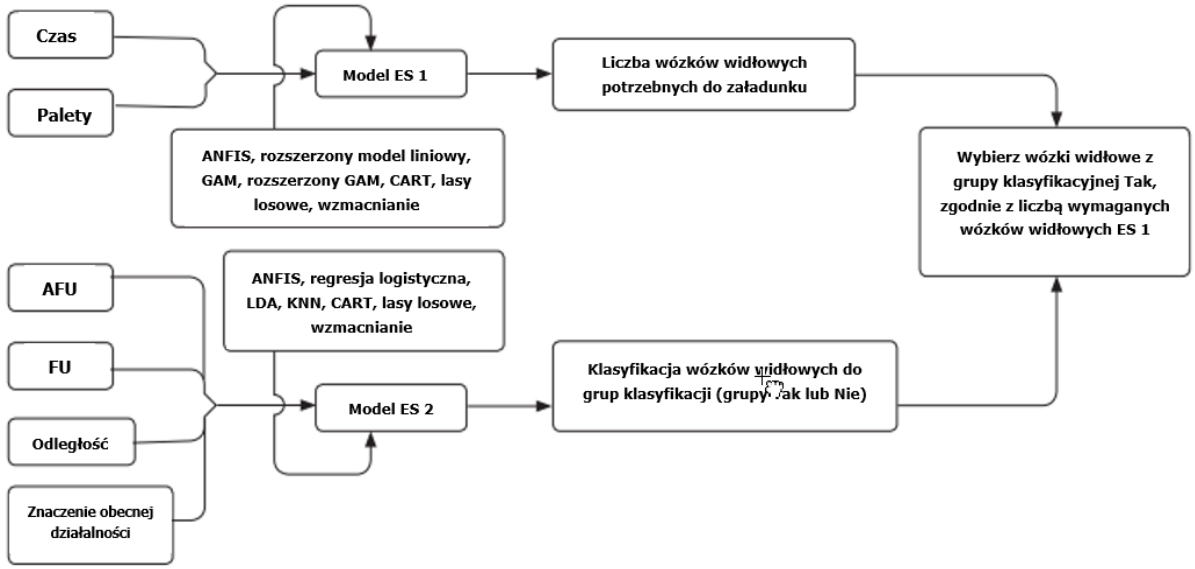


Rysunek 10.3 Etapy metodologii tworzenia systemu AI i ML DSS

Źródło: Rainer i Turban, 2008; Turban, Aronson oraz Liang, 2005

Pozyskanie wiedzy zostało osiągnięte na drodze wywiadów z menedżerami, obserwację ich procesów decyzyjnych oraz przegląd dokumentacji magazynowej. Aby opracować system DSS dla danego problemu, utworzono dwie bazy wiedzy. Pierwsza baza wiedzy obejmuje decyzje dotyczące liczby wózków widłowych rozmieszczonych w strefie załadunku (434 decyzje eksperckie), podczas gdy druga obejmuje decyzje dotyczące tego, które wózki widłowe były używane (368 decyzji eksperckich) w różnych scenariuszach operacyjnych. Na etapie wnioskowania o wiedzy zastosowano kilka algorytmów ML przy użyciu oprogramowania Matlab: Adaptacyjny neuro-rozmyty system wnioskowania (ANFIS), uogólnione modele addytywne (GAM), Random Forest, boosting, drzewa klasyfikacji i regresji

(CART), rozszerzona regresja liniowa, regresja logistyczna, algorytm k-najbliższych sąsiadów (KNN) i liniowa analiza dyskryminacyjna (LDA). Oceniono różne modele ML i zidentyfikowano te o najlepszej wydajności. ANFIS i CART wykazały lepsze wyniki i zostały wybrane jako ostateczne systemy DSS do praktycznego zastosowania w firmie. Transfer wiedzy został ułatwiony poprzez interfejs użytkownika ostatecznych modeli DSS. Strukturę i logikę systemu DSS zilustrowano na rysunku 10.4.



Rysunek 10.4 Budowanie struktury magazynu DSS w oparciu o algorytmy AI i ML

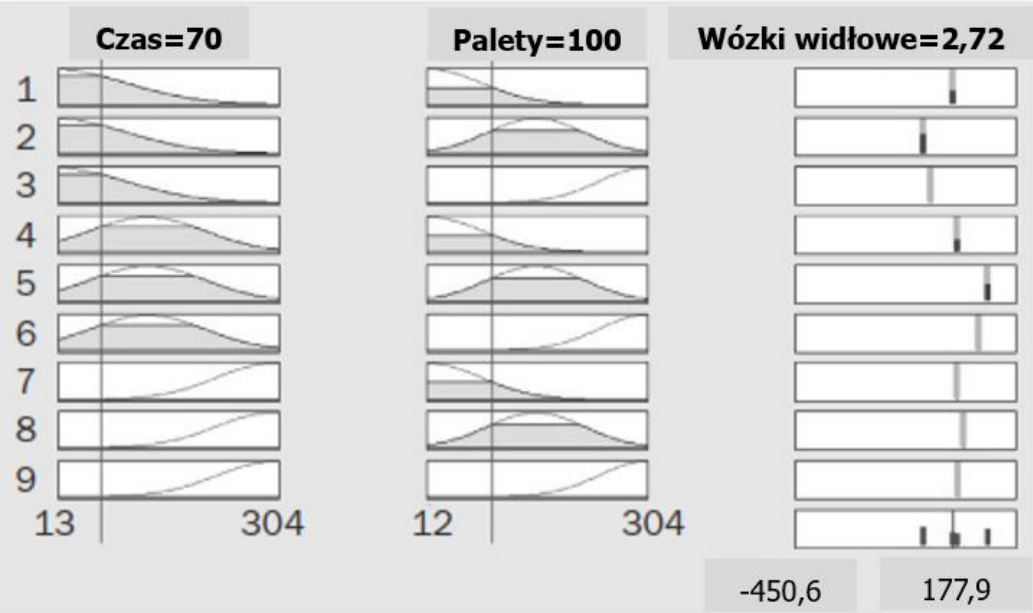
Struktura DSS składa się z warstwy wejściowej, zbudowanej z kilku kluczowych czynników, które wpływają na zaangażowanie wózków widłowych. Warstwa ML zawiera modele ML, które ponownie obliczają sugestie dotyczące liczby i wyboru wózków widłowych do wykorzystania w danym scenariuszu wejściowym. Najlepiej działające modele są wybierane jako modele systemu eksperckiego (modele ES), ponieważ baza wiedzy, na podstawie której tworzone są modele ML, jest pozyskiwana od ekspertów. Pierwszy model koncentruje się na określeniu liczby wózków widłowych wymaganych w strefie załadunku (model ES 1). Drugi model odnosi się do problemu wyboru konkretnych wózków widłowych, które powinny zostać zaangażowane (model ES 2). Oba modele zostały opracowane przy użyciu nadzorowanych technik uczenia maszynowego. Według Turban, Aronson i Liang (2005), uczenie maszynowe wykazało doskonałe wyniki w projektowaniu inteligentnych systemów wspomagania decyzji (DSS). Modele ES wysyłają sygnały (sugestie i propozycje

ML) dalej do operacji sortowania, gdzie każdy wózek widłowy, który jest sklasyfikowany w grupie sortowania „Tak", może być zaangażowany w daną operację załadunku.

Czynniki wpływające na decyzje kierownika zostały zidentyfikowane w drodze konsultacji. W celu określenia liczby wózków widłowych do rozmieszczenia w strefie załadunku, kluczowymi czynnikami są określony czas załadunku (od 15 do 135 minut) i wielkość ładunku (od 15 do 225 palet). Wybierając wózki widłowe do zaangażowania w operacje, menedżer bierze pod uwagę znaczenie bieżącej aktywności (ocenianej od 1 do 9 zgodnie z polityką firmy), wskaźnik wykorzystania wózka widłowego, jego odległość od doku załadunkowego oraz średni wskaźnik wykorzystania wszystkich wózków widłowych. Każdy wózek widłowy ma określoną liczbę godzin pracy, zanim konieczny będzie przegląd, a jego użytkowanie jest ograniczone po przekroczeniu tego limitu. Wykorzystanie wózka widłowego (FU) to procent godzin pracy zrealizowanych przez pojedynczy wózek widłowy, podczas gdy średnie wykorzystanie wózka widłowego (AFU) to średnia godzin pracy zrealizowanych przez wszystkie wózki widłowe. Wyższy wskaźnik AFU sugeruje, że większość wózków widłowych będzie wkrótce wymagać przeglądu.

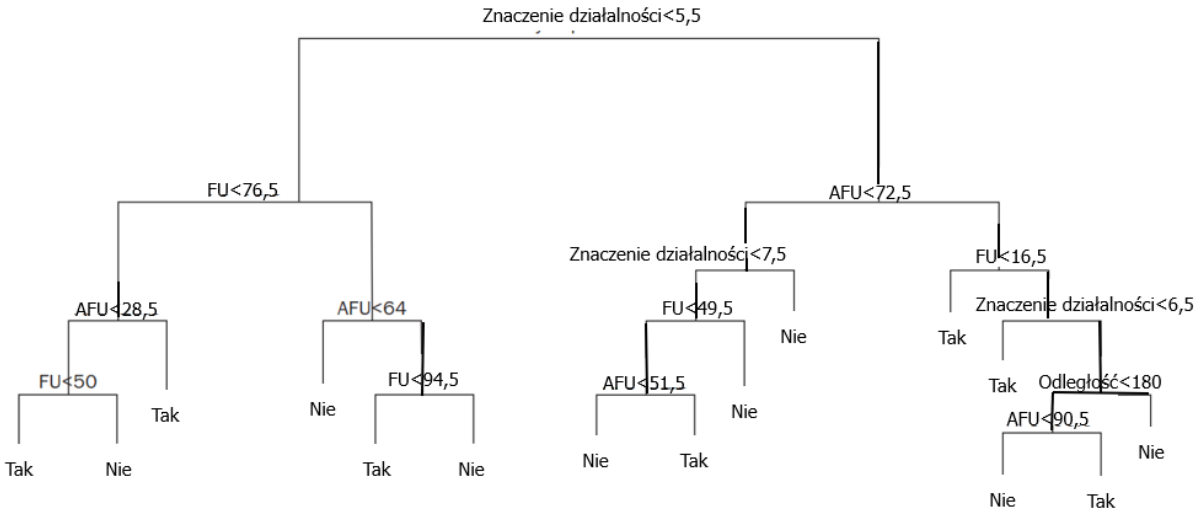
Interfejs użytkownika DSS

W większości sytuacji wejściowych najlepszą wydajność wykazały ANFIS i CART. W związku z tym zostały one wybrane jako silniki danego DSS i jego ES. Interfejs użytkownika modelu ES 1 jest przedstawiony na rysunku 10.5 i umożliwia operatorom szybkie i łatwe podejmowanie decyzji dotyczących liczby wózków widłowych do rozmieszczenia, po prostu przesuwając pionową linię przez domenę zmiennych wejściowych, w oparciu o określony czas załadunku i ilość ładunku.



Rysunek 10.5 System wnioskowania rozmytego modelu ES 1

Model ES 2 służy jako narzędzie uzupełniające do Modelu ES 1, usprawniając podejmowanie decyzji poprzez dostarczanie informacji o tym, czy dany wózek widłowy powinien zostać rozmieszczony w strefie załadunku (rysunek 10.6). Biorąc pod uwagę pozycję wózka widłowego (odległość od strefy załadunku), jego bieżącą aktywność (znaczenie aktywności), zrealizowane godziny pracy (FU) i średnie wykorzystanie wszystkich wózków widłowych (AFU), użytkownicy mogą łatwo określić, czy dany wózek widłowy nadaje się do załadunku, czy też należy wybrać inny. Drzewo decyzyjne CART jest proste w interpretacji, eliminując potrzebę wprowadzania wartości do oprogramowania. Zamiast tego drzewo z rysunku 10.6 można wydrukować i wyświetlić w widocznym miejscu w magazynie w celu szybkiego sprawdzenia.



Rysunek 10.6 Drzewo decyzyjne modelu ES 2 dotyczące zaangażowania wózka widłowego

Menedżerowie mogą codziennie korzystać z prezentowanego systemu DSS, pomagając w osiągnięciu wyższej responsywności łańcucha dostaw na potrzeby klientów i zapewniając wysokie prawdopodobieństwo terminowej dostawy. Zaproponowany AI i ML DSS wykazał skuteczne wyniki w pozyskiwaniu wiedzy eksperta „know-how" i przechwytywaniu jego „logiki wnioskowania". Korzystając z tej metody, wiedza menedżerów może być wyodrębniona i zastosowana do innych operacji magazynowych. Jest to szczególnie cenne dla praktyków, ponieważ zatrudnianie ekspertów magazynowych jest często kosztowne. Ponadto DSS może również służyć jako narzędzie szkoleniowe dla początkujących menedżerów, pomagając im zdobyć doświadczenie i poprawić umiejętności podejmowania decyzji w miarę upływu czasu. Dlatego systemy AI i ML, które mogą symulować decyzje menedżera, są niezbędnymi narzędziami, oferującymi znaczne oszczędności kosztów i zwiększoną wydajność operacji magazynowych.

BIBLIOGRAFIA DO ROZDZIAŁU 10.

- Athanasopoulou, K., Daneva, G. N., Adamopoulos, P. G., & Scorilas, A. (2022). Artificial intelligence: The milestone in modern biomedical research. BioMedInformatics, 2(4), 727-744. https://doi.org/10.3390/biomedinformatics2040049
- Chollet, F. (2021). Deep learning with Python. Simon and Schuster.
- Druzdzel, M. J., & Flynn, R. R. (2002). Decision support systems. In A. Kent (Ed.), Encyclopedia of library and information science (2nd ed.). Marcel Dekker, Inc.
- Gibbs, M. (2019, December 17). Table-driven programming and the weather forecasting stone. Global Nerdy. https://www.globalnerdy.com/2019/12/17/table-driven-programming-and-the-weather-forecasting-stone/
- Jackson, P. (1999). Introduction to expert systems. Addison-Wesley.
- Mirčetić, D., Ralević, N., Nikoličić, S., Maslarić, M., & Stojanović, Đ. (2016). Expert system models for forecasting forklifts engagement in a warehouse loading operation: A case study. Promet-Traffic & Transportation, 28(4), 393-401.
- Mircetic, D., Lalwani, C., Lirn, T., Maslaric, M., & Nikolicic, S. (2014, July). ANFIS expert system for cargo loading as part of decision support system in warehouse. In 19th International Symposium on Logistics (ISL 2014).
- Rostami-Tabar, B., & Mircetic, D. (2023). Exploring the association between time series features and forecasting by temporal aggregation using machine learning. Neurocomputing, 548, 126376.
- Olson, D. L., & Courtney, J. F. (1992). Decision support models and expert systems. Macmillan.
- Rainer, R. K., & Turban, E. (2008). Introduction to information systems: Supporting and transforming business. John Wiley & Sons.
- Turban, E. (1998). Decision support and expert systems (2nd ed.). Macmillan.
- Turban, E., Aronson, J., & Liang, T.-P. (2005). Decision support systems and intelligent systems (7th ed.). Pearson Prentice Hall.

11.SPIS RYSUNKÓW

12.SPIS TABEL

BUSINESS ANALYTICS SKILLS FOR THE FUTURE-PROOF SUPPLY CHAINS -

BUSINESS ANALYTICS SKILLS FOR THE FUTURE-PROOF SUPPLY CHAINS -