



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 „wertyzację”, 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,
- **przewidyujące** – 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?

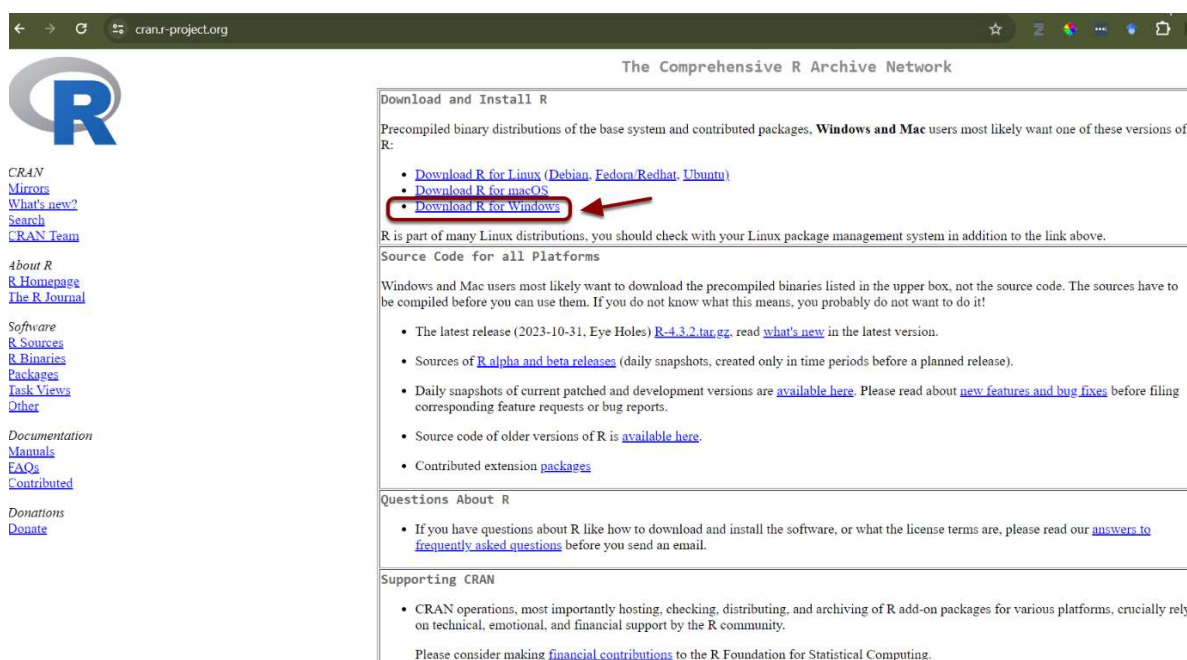
R to zintegrowany pakiet oprogramowania do manipulacji danymi, obliczeń i wyświetlania graficznego (R Core Team, 2019). Posiada on między innymi:

- 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).

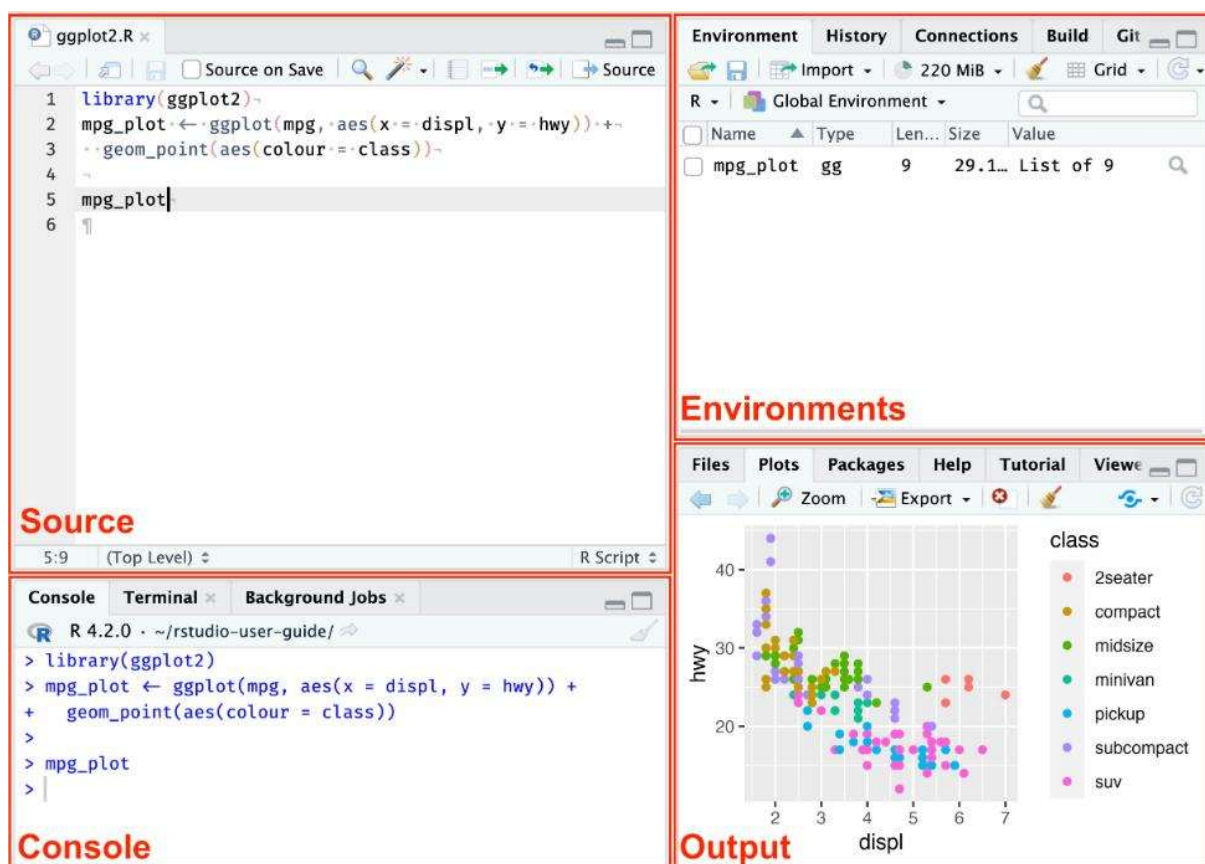
Instalowanie R i R Studio

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):

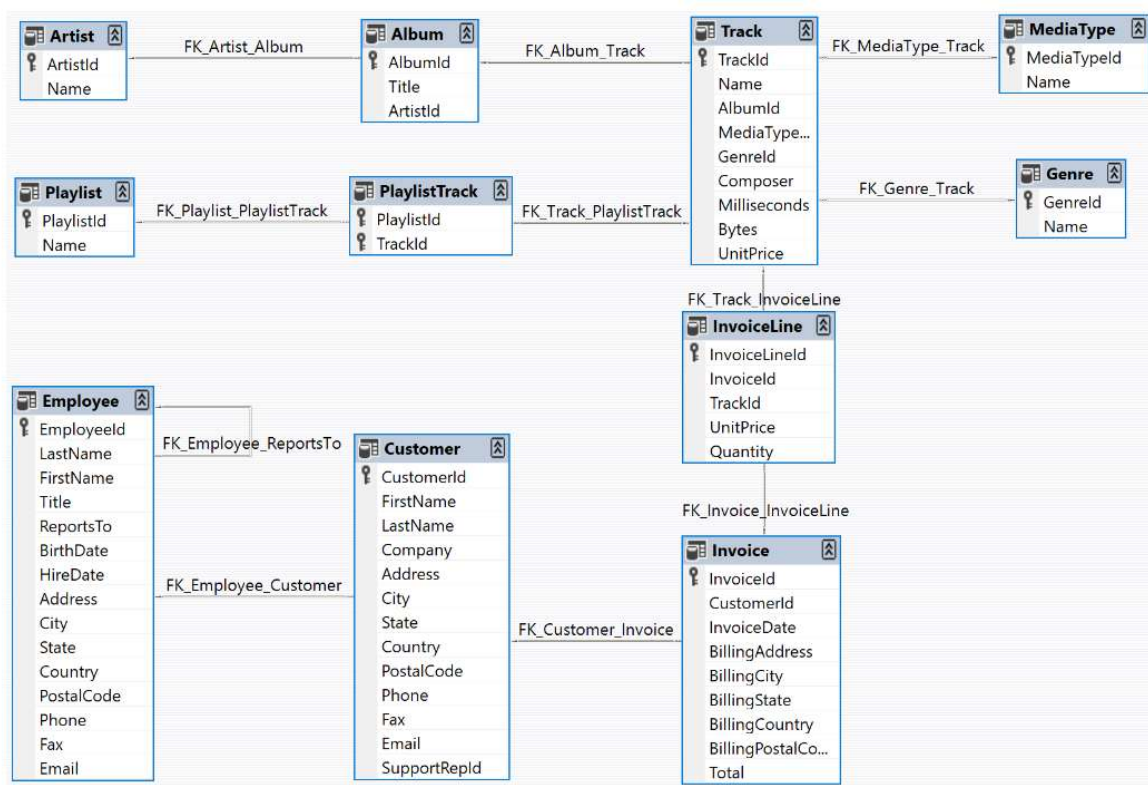
- Source (Panel źródła),
- Console (Panel konsoli),
- Environment (panel Środowisko), zawierający zakładki Environment (Środowisko), History (Historia), Connections (Połączenia), Build (Budowanie), VCS, oraz Tutorial (Przewodnik, Instruktaż),
- Output (Panel wyjściowy), zawierający zakładki Files (Pliki), 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.
Album	Zawiera informacje o albumach muzycznych, każdy powiązany z artystą.
Track (Utwór)	Zawiera informacje o poszczególnych utworach muzycznych, w tym odniesienia do albumów, typów multimediiów i gatunków.
Genre (Gatunek)	Zawiera informacje o gatunkach muzycznych.
MediaType (TypNośnika)	Zawiera informacje o różnych typach multimediiów (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 przedsiębiorstwach, 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.

```
29 ▾ ## Choose a Album table from the database
30 ▾ ```{r}
31 query_album <- "SELECT * FROM Album LIMIT 10"
32 data_album <- dbGetQuery(con, query_album)
33 print(data_album)
34 ▾ ```
35
36 ▾ ## Query to get album details along with artist names
37 ▾ ```{r}
38 query_album_artist <- "
39 SELECT Album.AlbumId, Album.Title AS AlbumTitle, Artist.Name AS ArtistName
40 FROM Album
41 JOIN Artist ON Album.ArtistId = Artist.ArtistId
42 LIMIT 10"
43
44 data_album_artist <- dbGetQuery(con, query_album_artist)
45 print(data_album_artist)
46 ▾ ```
47
48 ▾ ## Query to get the top-selling albums along with artist names
49 ▾ ```{r}
50 query_top_selling_albums <- "
51 SELECT
52     Album.Title AS AlbumTitle,
53     Artist.Name AS ArtistName,
54     SUM(InvoiceLine.Quantity) AS TotalQuantitySold
55 FROM
56     InvoiceLine
57 JOIN
58     Track ON InvoiceLine.TrackId = Track.TrackId
59 JOIN
60     Album ON Track.AlbumId = Album.AlbumId
61 JOIN
62     Artist ON Album.ArtistId = Artist.ArtistId
63 GROUP BY
64     Album.AlbumId, Album.Title, Artist.Name
65 ORDER BY
66     TotalQuantitySold DESC
67 LIMIT 10"
68
69 # Execute the query
70 top_selling_albums <- dbGetQuery(con, query_top_selling_albums)
71 knitr::kable(top_selling_albums)
72 ▾ ```
```

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 sprzedanych albumów |
|--|------------------------------|-----------------------------------|
| Minha Historia | Chico Buarque | 27 |
| Greatest Hits | Lenny Kravitz | 26 |
| 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.