7. Statistical data processing SPSS

By now, you have already acquired a fundamental understanding of statistics, data manipulation, simulation establishment, modelling, and analysis within logistics supply chains, along with straightforward linear regression methods. While statistics offers a diverse range of models and techniques to enhance your optimization efforts, conduct analysis, and identify potential enhancements, you may have observed that as the complexity of the analysed data and calculations grows, traditional approaches can become increasingly intricate and challenging to compute. As the intricacy of your data and computations escalates, conventional methods may become outdated and, in some cases, compromise the reliability of your results. To address this, statistics uses various software programs that automate the analysis and interpretation of collected data while also providing a multitude of models and functions to ensure reliable outcomes. One such software is IBM's SPSS, which will be a key tool in this chapter. In this chapter, we will provide a concise introduction to the primary usage of SPSS software, exploring its functionalities and practical applications. The initial introduction will be followed by the practical application of the program through four fundamental tests for result calculation: the T-test, correlations, Chi-Square, and ANOVA. To facilitate your learning, we will present simple problems and their solutions to help you become acquainted with these tests.

7.1 Basics of IBM`s SPSS

You may have already had some experience with SPSS software. However, if you still need to, let us offer a brief introduction to it. SPSS, much like its more widely recognized counterpart, Excel, facilitates data manipulation, analysis, and visualization. Nevertheless, unlike Excel, which can sometimes be laborious and complex in function programming, SPSS offers a user-friendly interface for statistical analysis (IBM, 2021). It offers a variety of functions and methodologies to handle your data provided efficiently. While the SPSS software excels in providing extensive statistical analysis capabilities, navigating data manipulation and configuring initial settings for analysis can sometimes be challenging (IBM, 2021). Therefore,

we will delve into the fundamentals of data importation and data preparation for subsequent statistical tests.

Given the widespread use of Excel for handling numerical data, your data may be obtained or prepared in an Excel spreadsheet. Fortunately, SPSS can import data from various file formats into its spreadsheets. Once you have your finalized Excel spreadsheets ready, open the SPSS software. On the initial screen, navigate to the "File" tab and select "Import Data". In the subsequent window, you can choose the data format you intend to import (refer to the figure 7.1). Following this step, locate your prepared file, select it, and proceed to the next window. This window will prompt you to configure additional settings. If you have already included column names in the first row of your data, opt for the "Read variable names from the first row of data" (refer to figure 7.1), and then click "Finish" to have the data appear in the spreadsheet (IBM, 2021).

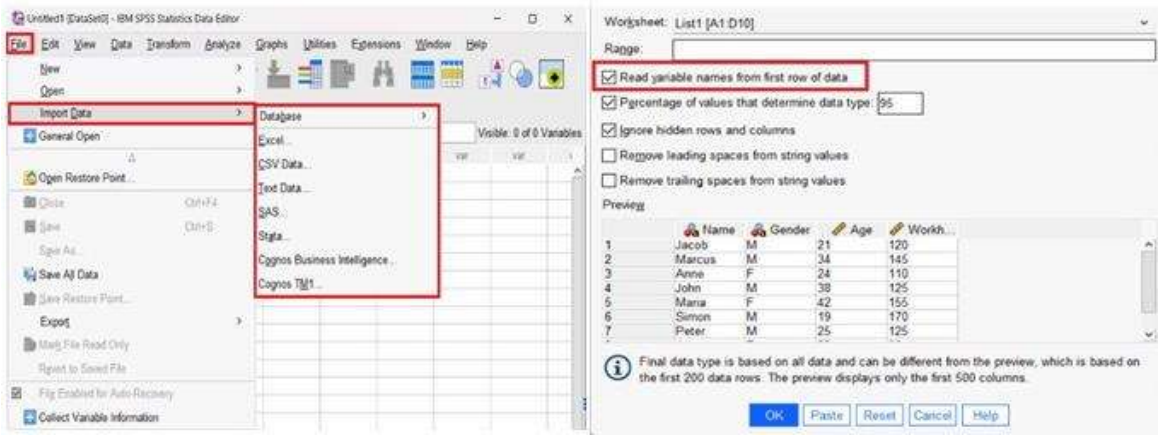


Figure 7.1 SPSS Import settings.

Now that we have the data in our spreadsheet, you will notice a distinct difference in presentation compared to Excel. SPSS categorizes data into two primary types, each with two additional sub-types. As illustrated in figure 7.2, data can be classified as numerical or categorical. Numerical data consists of numbers and can be categorized as discrete (with finite options) or continuous (offering infinite options). On the other hand, categorical data comprises words and can be further distinguished as ordinal (having a hierarchy) or nominal (lacking a hierarchy). Depending on the nature of your data, you may need to configure the variables to align with your desired analysis. In most cases, SPSS will automatically categorize variables appropriately. Suppose you wish to perform further manipulation of data types. In that case,

you can access the "View" option and, under "Variable View", adjust variable information such as Name, Type, Width, Measure, and more (IBM, 2021).

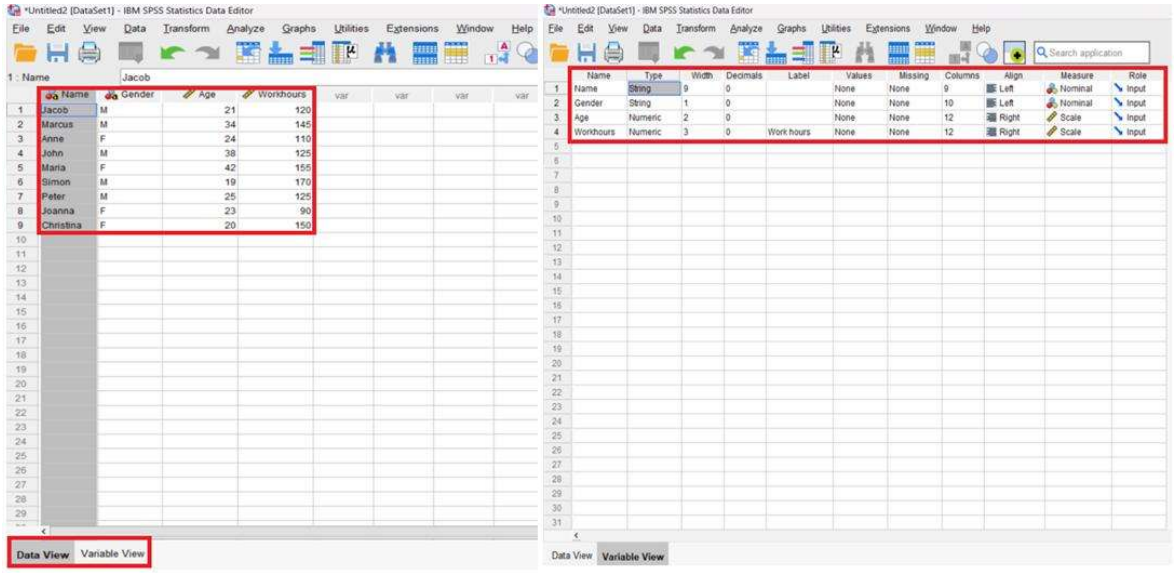


Figure 7.2 Data and variable view windows.

Once you have correctly set up your data, you can explore it within SPSS. SPSS allows users to perform fundamental statistical analysis without relying on predefined functions. On the initial screen (refer to figure 7.3), navigate to " Analyse ", followed by "Descriptive Statistics", and then select "Explore". In the "Explore" section, you will find various options depending on the characteristics of the data you provided. In this mode, SPSS will provide you with "descriptive statistics" information about your data. While this is valuable for initial data analysis, it offers only fundamental insights and does not delve into more detailed statistical analysis, which will be covered in subsequent chapters. Before proceeding further, we will also explore another function in SPSS—graph visualization (IBM, 2021).

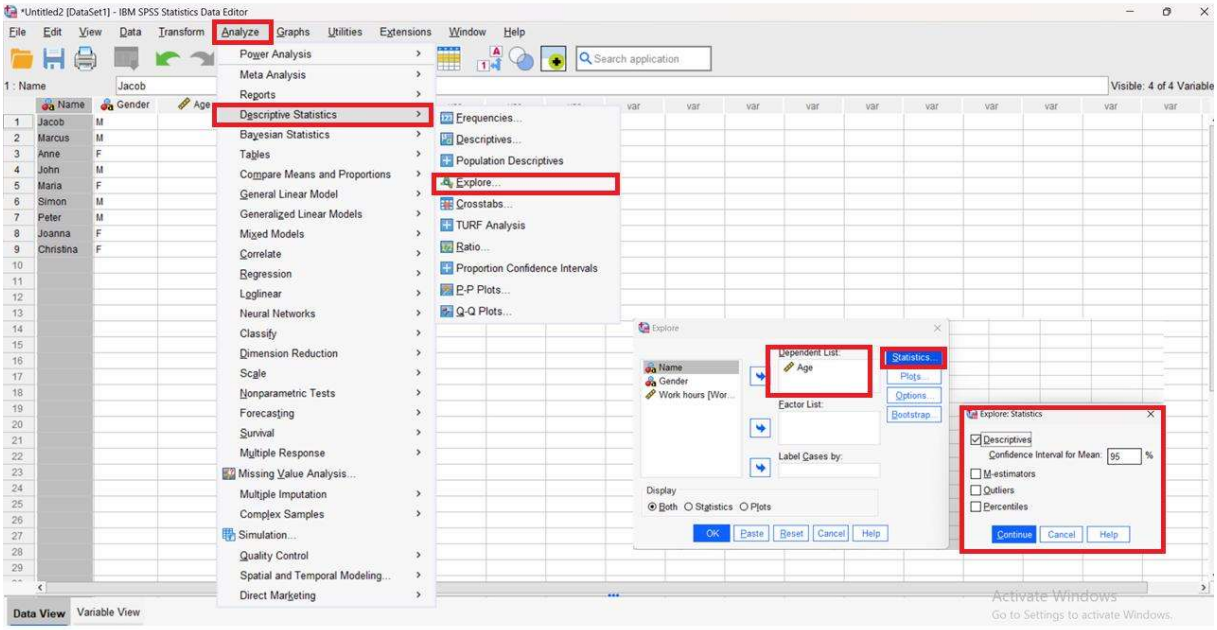


Figure 7.3 Descriptive statistics settings.

SPSS offers a range of data visualization options, including histograms, box plots, bar charts, scatterplots, line charts, pie graphs, and more. By this point, you should have a basic understanding of what each type of graph represents and how to interpret the results they provide. Therefore, we will focus on how to create these graphs within the SPSS software. To create graphs, select the "Graphs" tab on the initial screen, followed by "Chart Builder". In the new window, you can choose the type of graph you want to create and select the variables to be included. After selecting "Finish", a new window will appear with the results visualized in the chosen graph format. In this new window, you can interact with the graph actively, allowing you to modify variable colors and fonts, explore variable distributions across the graph, and more (IBM, 2021).

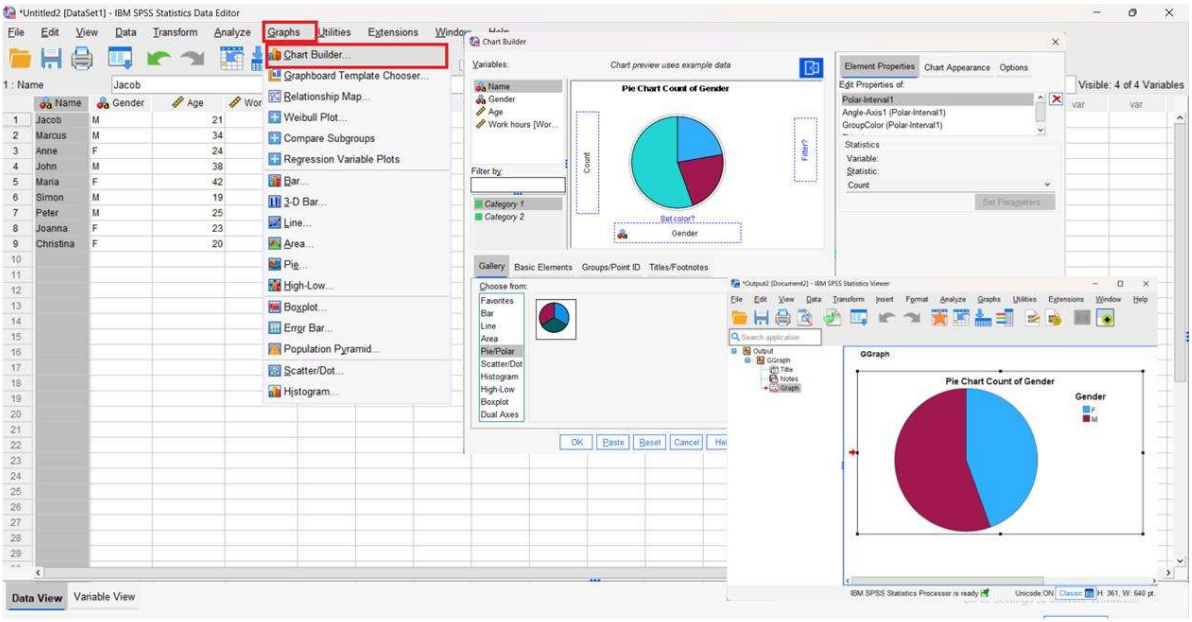


Figure 7.4 Chart Builder settings in SPSS.

Up to this point, we have covered three of the four rules for "Exploring data", which include looking at data (raw data exploration), identifying data (determining data types), and, to some extent, graphing and describing data through descriptive statistics and graph creation. The final rule is "Question Formulation", where we ask ourselves what we aim to achieve through data analysis and set up graphs and descriptive statistics accordingly to obtain answers to our specific questions. For instance, in our current example, the question could be, "Is our analysed population predominantly female?" By employing both graphs and descriptive statistics, we can conclude that our population consists mainly of male individuals. When formulating your questions, always consider the available data and the variables you have identified (Garth, 2008). This concludes the first part of the SPSS data analysis, and we will now proceed with test preparation.

7.2 Data management

When engaging with crucial data in the SPSS software, it becomes crucial to comprehend the techniques for manipulating information across individual active datasets. SPSS provides functionalities facilitating the manipulation of existing data contained within active datasets. Occasionally, you may encounter two databases separately imported into datasets, yet

the preference is to merge them for enhanced analysis. Consider a logistic company with two branches, each contributing data on costs and transporting cargo in kilograms. The managerial objective is to analyse the overall efficiency of the company. In SPSS, accomplishing this involves navigating to "Data," selecting "Merge files," and having two distinct options. One involves selecting "Cases" and specifying the variable for merging, removing that variable while merging the others. Alternatively, opting for "Variable" retains the variable in the new dataset. A practical application is evident in our logistics scenario, where merging datasets simplifies the company's comprehensive performance analysis.

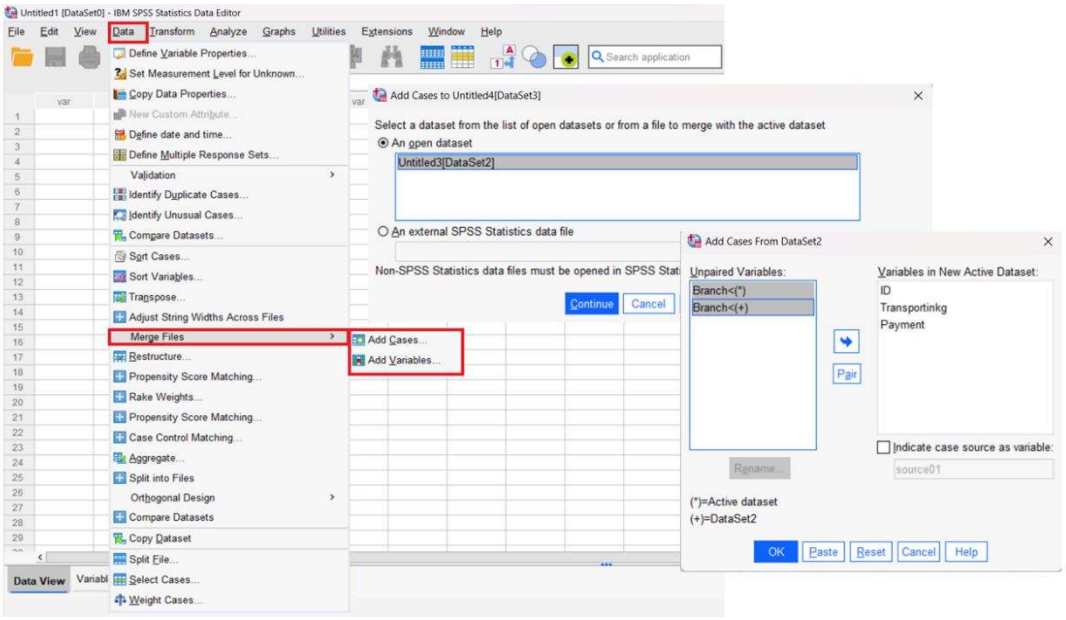


Figure 7.5 Merging file window.

While the merge and split functions enable specific data manipulation, the "Select cases" option offers distinct advantages. Imagine having data for shops B, C, and D in a single database, and the focus is solely on comparing Shop A and Shop C. By selecting "Data" and "Select Cases," one can specify the variables of interest, effectively filtering out unwanted data. For instance, setting Shop C as 2 instructs the software to concentrate solely on Shop C, generating output that is then available for subsequent analyses, such as descriptive statistics, focusing exclusively on the selected cases. Such an approach also enables comparative analysis between only Shop A and Shop C values.

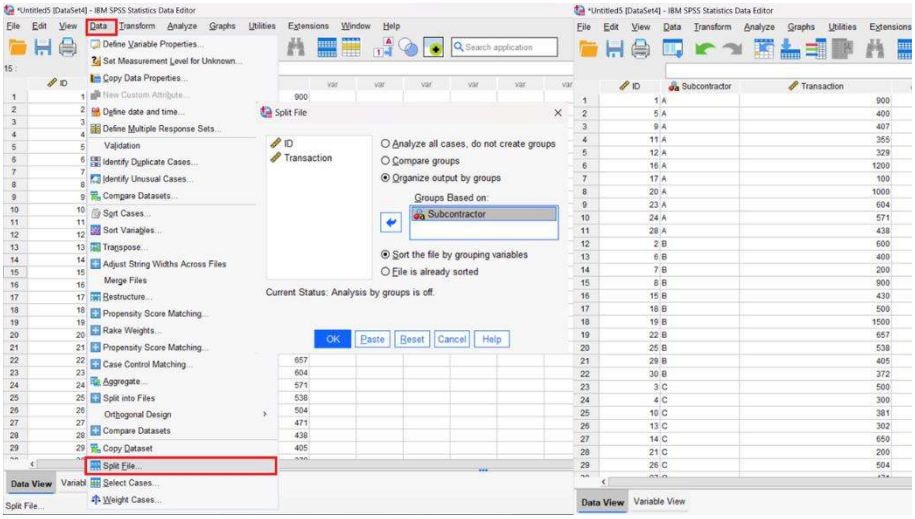


Figure 7.6 Splitting file window.

While merging and split functions enable certain data manipulation, there is also the option to "Select cases". Imagine that we know for certain that Shop A has on average 120 € profit and we want to compare that to Shop C. Unfortunately, in our database we have data for shops B, C and D in a single database and the analysis would include data from all three shops. Byclicking "Data" and "Select Cases" we can select which variable we want to focus. In our cases we set that shop C should be set as 2 and then created the function for the software to focus only Shop C. The output can be then used for subsequent analysis by choosing this new column (e.g. descriptive statistics).

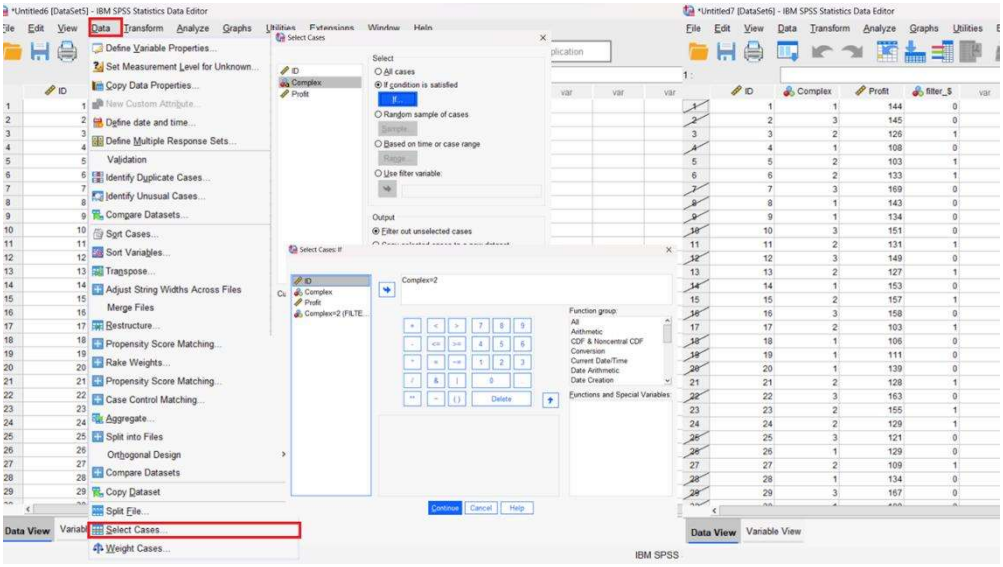


Figure 7.7 Selecting case procedure.

Occasionally, datasets may already contain variables, yet there is a need to introduce new variables based on existing ones. Take, for example, a logistic company manager who possesses data on the weight and distance travel for various products but requires delivery time for optimizing routes. In SPSS, achieving this involves clicking "Transform" and then "Compute Variables." A new variable, DeliveryTime, is created within the new window by setting numeric expressions. In this case, assigning a scale of 0.8 to distance and 0.2 to weight results in a new variable representing delivery time, a crucial addition to the dataset. The flexibility of computing additional variables exists, created for the needs of statistical tests.

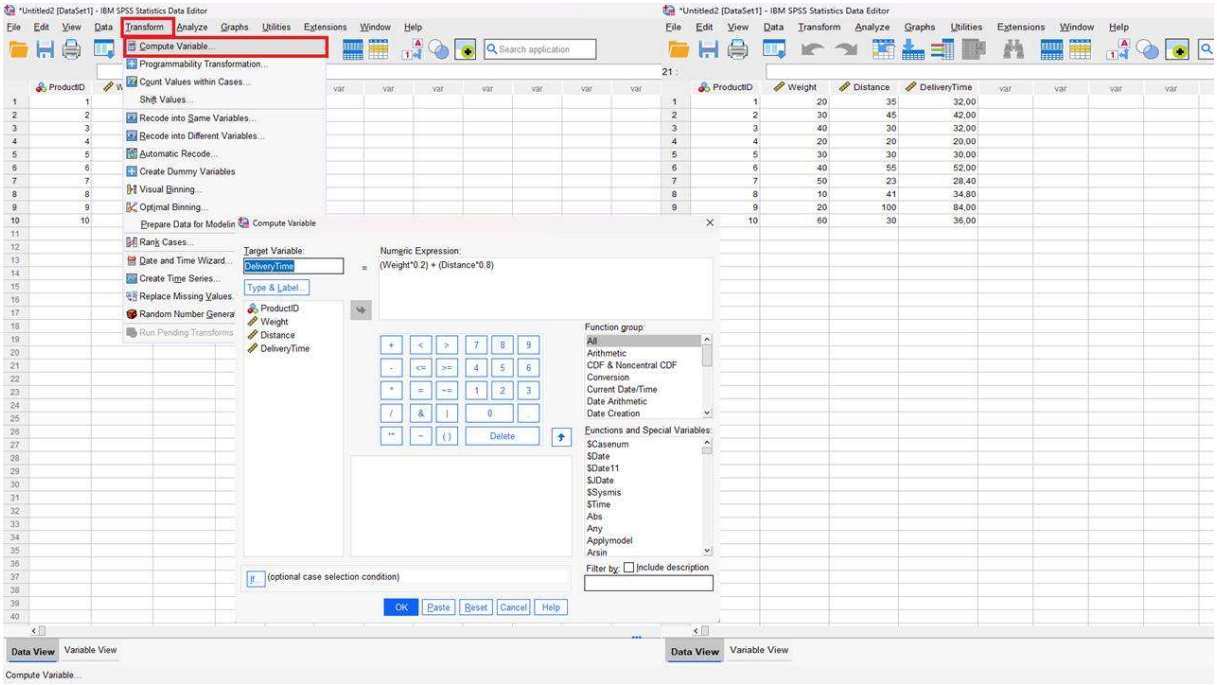


Figure 7.8 Computing variables procedure.

This concludes a small overview of data management functions that SPSS covers, which might be useful during the subsequent model tests covered in this chapter. We will continue with the phases needed before we can conduct a statistical test in the SPSS software.

7.3 Test preparation

Before proceeding with statistical tests, it is essential to adhere to a standard data analysis process flow, which includes data exploration (as covered in Chapter 7.1 and 7.2), data analysis, and results interpretation (Garth, 2008; George & Mallery, 2022). In this chapter, our

focus is on data analysis using the SPSS software. Since hypotheses have already been addressed in previous chapters, our primary focus will be on conducting normality tests within SPSS. There are three methods to assess normality: the histogram, QQ-plot, and the normality test. It is advisable to employ at least two, if not all three of these options, as they each provide distinct information (Ghasemi & Zadesiasl, 2012). To create a histogram, go to "Graphs", followed by "Chart Builder". In the new window, select "Histogram". If you have multiple variables, you must repeat this process for each one to obtain the results. A histogram validates the test for normal distribution if the bars representing variable values resemble a bell curve. If the bars lean more to the left or right side, it may indicate an exponential distribution. For example, we generated a database of 100 IDs, each with a variable representing weight in kilograms. Following the instructions, we created a histogram, as shown in the figure 7.9. As evident from the figure, the bars are distributed across the graph, and while they may not perfectly mirror the curve, they nonetheless suggest a normal distribution and a positive test result (George & Mallery, 2022; Goeman & Solari, 2021).

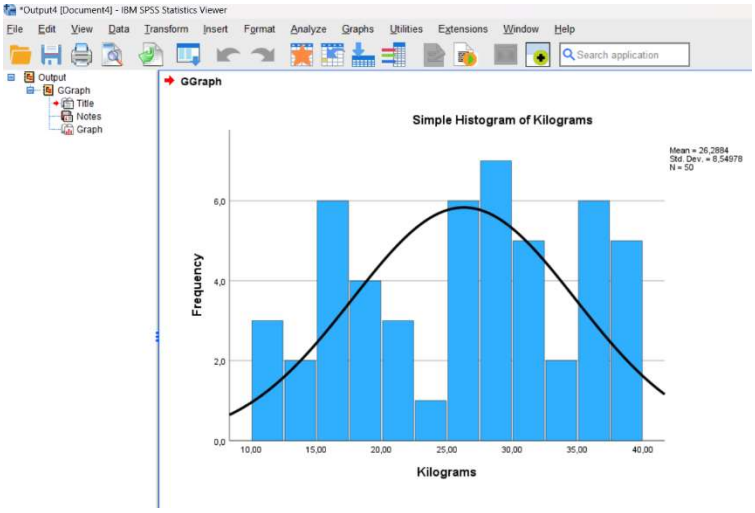


Figure 7.9 Histogram of normality test results.

Another option for conducting normality tests is the QQ-plot, which can be initiated by clicking " Analyse ", followed by "Descriptive Statistics", and then selecting "Q-Q Plots". The advantage of this approach is that it allows for the assessment of multiple variables simultaneously (Williamson, b.d). The test is considered successful when the points on the plot cluster closely around a straight line, representing a normal distribution. If the points form "tails", it indicates a failed normality test (Andersen & Dennison, 2018). Using the same database from the histogram graph test, we conducted a Q-Q Plot test. In the figure 7.10 below, you can

observe that most cluster points for our variable align with the straight line, indicating a normal distribution of our data. While we could already conclude that the normality test is positive at this stage, we decided to seek confirmation from all three tests.

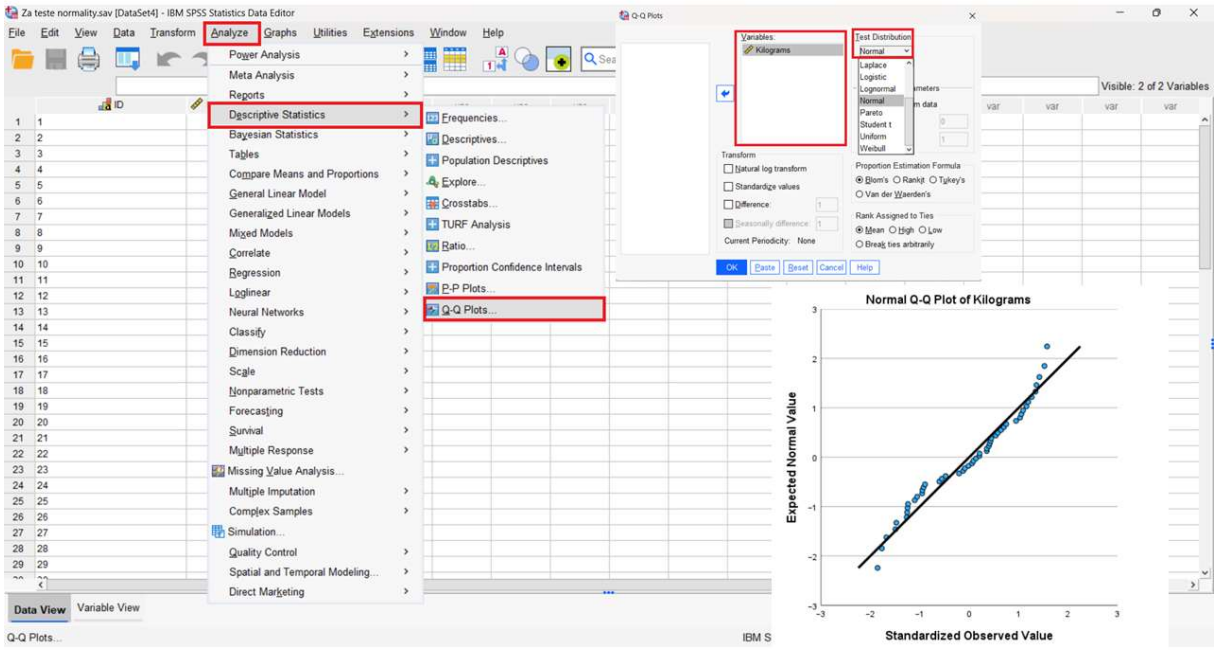


Figure 7.10 Q-Q plot normality test settings and results.

The final option for conducting a normality test is the so-called Test of Normality, considered a statistical test. Typically, it uses the Kolmogorov-Smirnov test, but for small sample sizes, the Shapiro-Wilk test can be employed (Goeaman & Solari, 2021). In SPSS, you can perform this test by clicking " Analyse ", followed by "Descriptive Statistics", and then "Explore". You must set the variables you want to check under the "Dependent List" box. Then, under "Plots", select "Normality Plots with Tests". The test is considered successful if the Sig column (*p*-value) in the results is greater than 0.05, indicating a normal distribution. If the *p*-value is less than 0.05, it suggests a non-normal distribution, and the test is considered unsuccessful. We conducted this test once again using the same database as in the previous tests. From the results, we can conclude that according to the Kolmogorov-Smirnov standard, the test is positive as the *p*-value is higher than 0.05. However, for the Shapiro-Wilk test, the *p*-value is lower, indicating a negative test result. These differing results occur because both approaches have different sensitivity settings and power in detecting deviations (Ghasemi & Zahediasl, 2012). Since we have already conducted both

Q-Q Plots and the histogram graph tests, the Test of Normality can be considered positive overall. With the normality tests confirmed, we can conduct the main tests, such as the One-Sample Test.

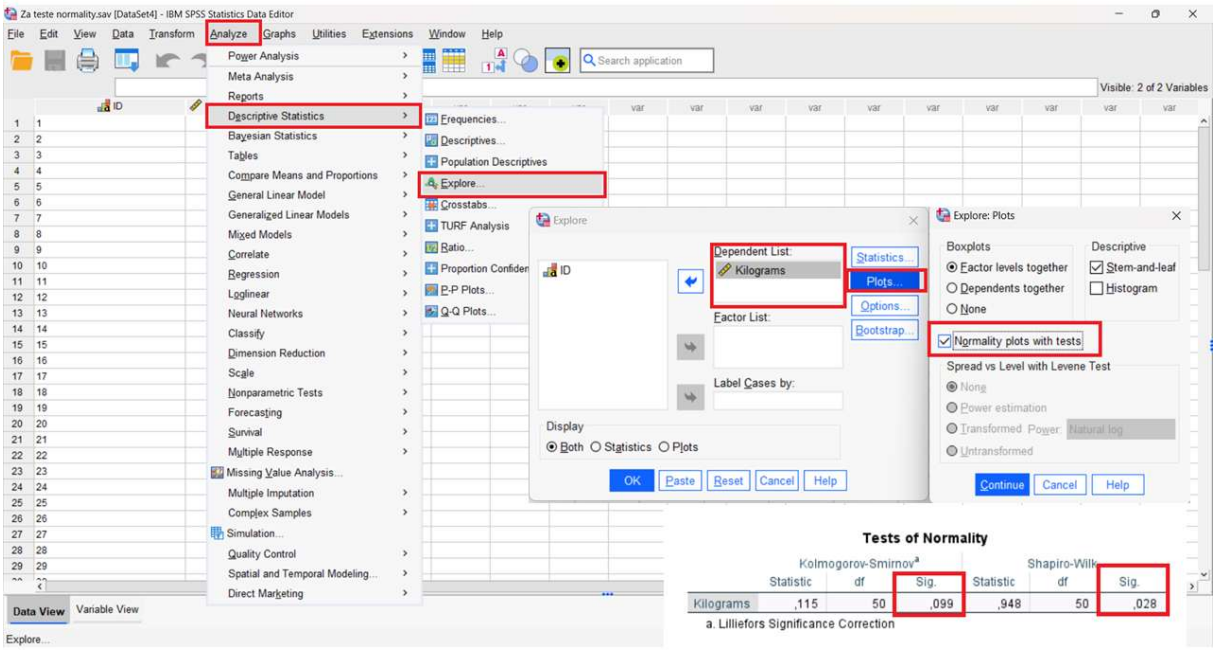


Figure 7.11 Test of Normality settings and results.

7.4 One Sample T-test

You have already covered the theory behind the One Sample T-test in the previous chapters; we will thus focus primarily on conducting a test with the SPSS software. For our One Sample T-test we have prepared a database with a sample of 200 inputs, which includes 1 categorical variable (Student ID) and 2 numerical variables (Weight and Age) (Kim, 2015). Following the guides from the previous subchapters we conduct:

- Explore the data, namely our **variables** and **descriptive statistics** and establish our **question**.
- Check the **normality**, since only one variable histogram and Q-Q plot should suffice.
- Set up hypothesis, where for **Null** the variable is no different from a certain value and **Alternative** where it is different.
- Conduct the **Students T-test**.

- Interpret results, focusing on **Null rejected** or **not**, answer the question and write a report on our test.

In our case, we decided that our question should be, "Is the average weight of students greater than 74 kilograms?". Following the question, we establish our hypothesis to the question, which is "Null = there is no difference" and "Alternative = there is a difference". We conducted the histogram and Q-Q plots to check for normality tests, and after their conclusion, we followed with the T-test. To run the T-test, we click "Analyse" and follow up with "Compare Means" and "One-Sample T-test". In the test variable box, we put our student ID, set the test value to 74 and start the test (refer to figure 7.12).

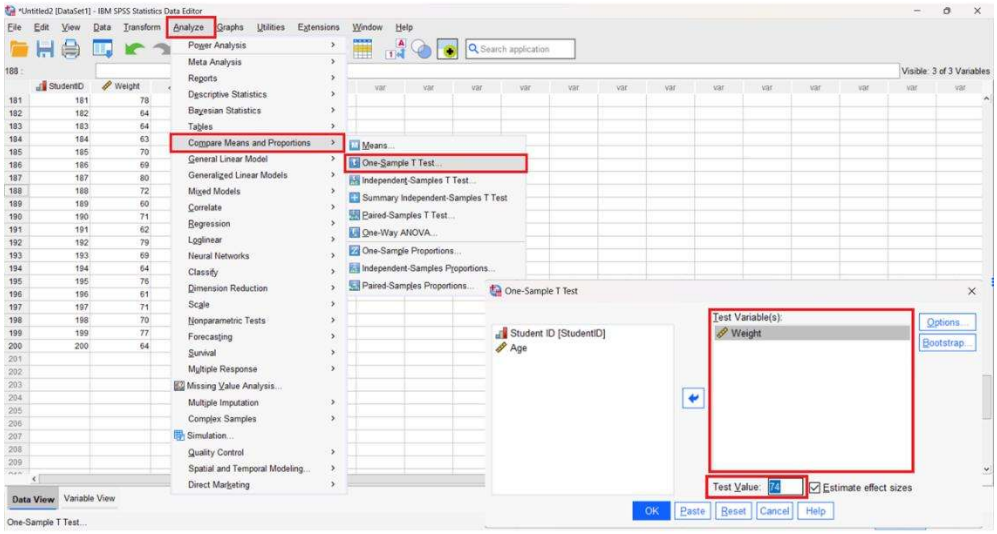


Figure 7.12 One Sample T-Test settings.

Upon confirming the test, another window will appear with the results of our analysis (refer to figure 7.13). This window provides several pieces of information regarding our analysis. In this case, both *p*-values are lower than 0.05, indicating the test's significance. Additionally, we check the *t* and *df* values, which, in our case, are -9.806 and 199, respectively. From these results, we can conclude that our null hypothesis is rejected. Therefore, the complete result report is as follows: "The average student weight is significantly lower (mean = 69.63) than the value of 74 kg (1-sample t-test, $t = -9.806$, $df = 199$, p-value < 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.

Figure 7.13 One Sample T-Test results.

7.5 Correlation

Let us now move on to the second test, which is the correlation test. We will conduct it using the same database as in the one-sample t-test example. Similar to the one-sample t-test, we will follow the procedure with a few modifications. When performing a correlation between two variables, it is important to specify which one is the dependent variable, and which is the independent variable (Janse *et al.*, 2021; Mishra *et al.*, 2019). This selection can be made based on your research question. In our case, we want to investigate "Whether there is a correlation between a student's age and their weight?". Following the question, we consider weight as the dependent variable and age as the independent variable, as we want to explore if variations in age are related to variations in weight. We define our null and alternative hypotheses (see 7.3 and 7.4) and then run the test by clicking " Analyse ", followed by "Correlate" and "Bivariate." Both variables should be placed in the "Variable" box. Ensure that "Pearson", "Two-Tailed", and "Flag Significant" are selected or set (refer to figure 7.14). In this case, we chose "Pearson" because our data indicated a normal distribution and could be analysed using parametric methods. If normal distribution is not indicated, non-parametric methods should be used (in this case, you would select Spearman instead of Pearson) (George & Mallery, 2022; McClure, 2005).

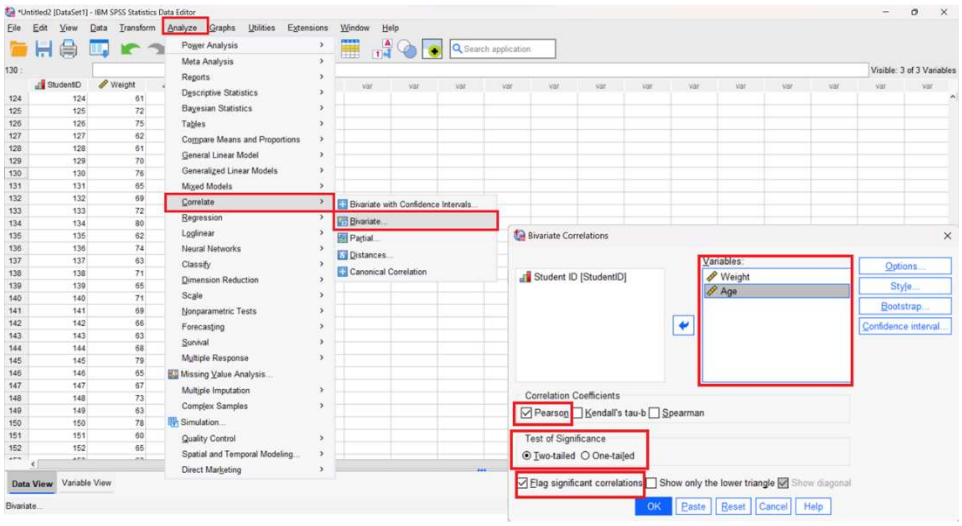
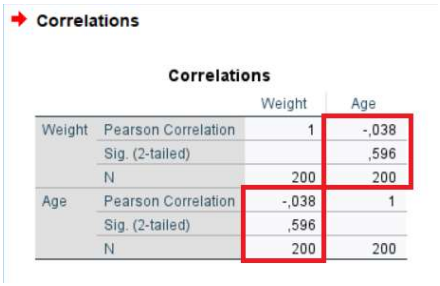


Figure 7.14 Correlation test settings.

Once again, we obtain the results in a new window (refer to figure 7.15). From the results, we can observe that our Pearson Correlation is -0.038, and our *p*-value is 0.596. In correlation analysis, the closer the correlation value is to zero, the weaker the correlation between the variables. In our case, the correlation is very close to zero, indicating no significant correlation between the two variables (McClure, 2005). Additionally, the high *p*-value (0.596) suggests that there is no substantial evidence to conclude that there is a meaningful correlation between the two selected variables (Williamson, b.d.). As a result, our null hypothesis is not rejected. Based on this, we can report that "There was no correlation between the students' age and weight".



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

Figure 7.15 Correlation test results.

7.6 Chi-Square

The third test we will perform in SPSS software is the Chi-Square test. Unlike the previous two tests, the Chi-Square test compares two categorical variables rather than numerical variables

(Turhan, 2020). Like the process in sections 7.4 and 7.5, we begin by exploring the data and formulating a research question. In our example, we have a logistics company with 200 customers, and we have data on the type of payment and the type of transportation chosen by each customer. The question we aim to answer is, "Do different payment types exhibit different preferences for transport types?" Since we are dealing with only categorical variables, there is no need for a normality test. We establish our Null hypothesis (The preferences for transport types are the same among all payment types) and the Alternative hypothesis. To conduct the Chi-Square analysis, click "Analyse", followed by "Descriptive Statistics", and select "Crosstabs". It is crucial to place the variables based on your research question in either the column or row box (refer to figure 7.16) (Garth, 2008).

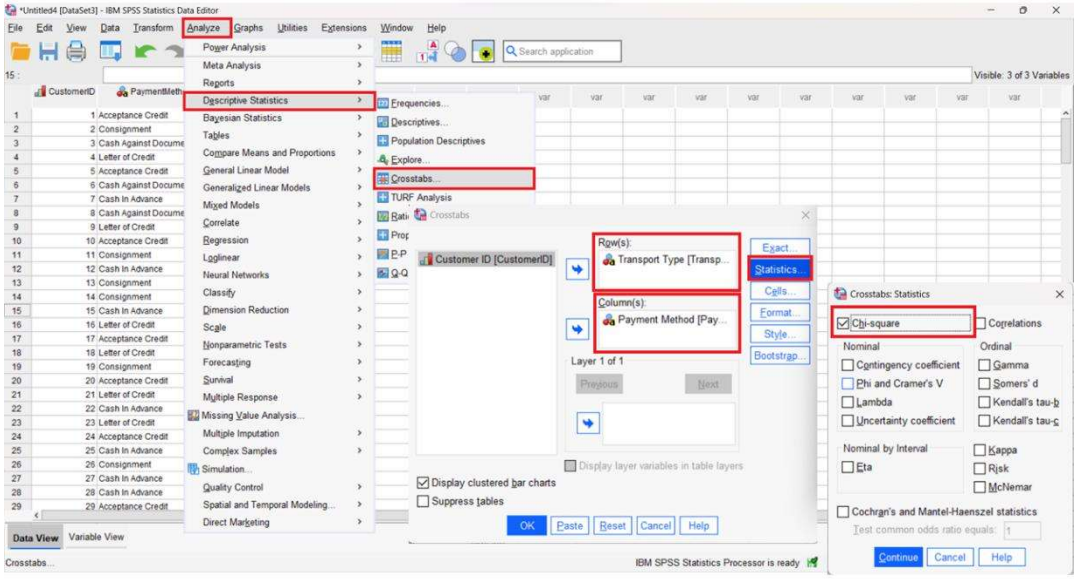


Figure 7.16 Chi-Square test settings.

After the analysis, a new window displays the results (refer to figure 7.17). In this window, you can observe that the Pearson Chi-Square value is 11.614, *df* value is 12, and the *p*-value (asymptotic significance) is 0.477. Based on these results, we can conclude that there is no significant association between the two variables, and the null hypothesis is not rejected. Therefore, the report follows: "There is no significant preference detected between different payment types for different transport types (2-tailed Chi-Square test, 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.

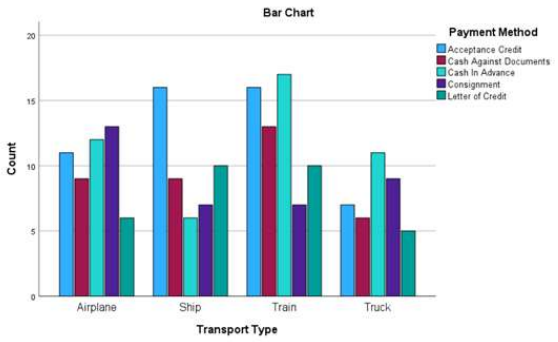


Figure 7.17 Chi-Square test results.

7.7 ANOVA

The final test we will cover is the ANOVA test, specifically focusing on the simpler model known as one-way ANOVA, which involves a categorical variable and a numerical variable (Goeman & Solari, 2021). As with the T-test, we will follow the same procedure: explore the data, formulate a research question, conduct a normality test, and set hypotheses. Let us consider a case study of a transport dispatcher working for a logistics company. The dispatcher closely collaborates with a partner company and regularly plans three different routes for the trucks to deliver their goods. Due to a "Just-in-time" policy emphasizing faster deliveries, the question arises: "Does the choice of delivery route impact on the delivery time for the company?" To run the ANOVA test in SPSS, go to "Analyse" followed by "Compare Means..." and then "One-way ANOVA". Place the dependent variable in the "Dependent List" box and the Factor variable in the "Factor" box (refer to figure 7.18). For thorough analysis, we have also included the Post Hoc setting. It is important to note that Post Hoc analysis should only be conducted if the initial ANOVA test is positive. By employing Post Hoc analysis, we can identify the most optimal choice (in our case, the route). The most reliable methods to use for Post Hoc analysis are either the Bonferroni correction or the Tukey HSD method (Goeman & Solari, 2021).

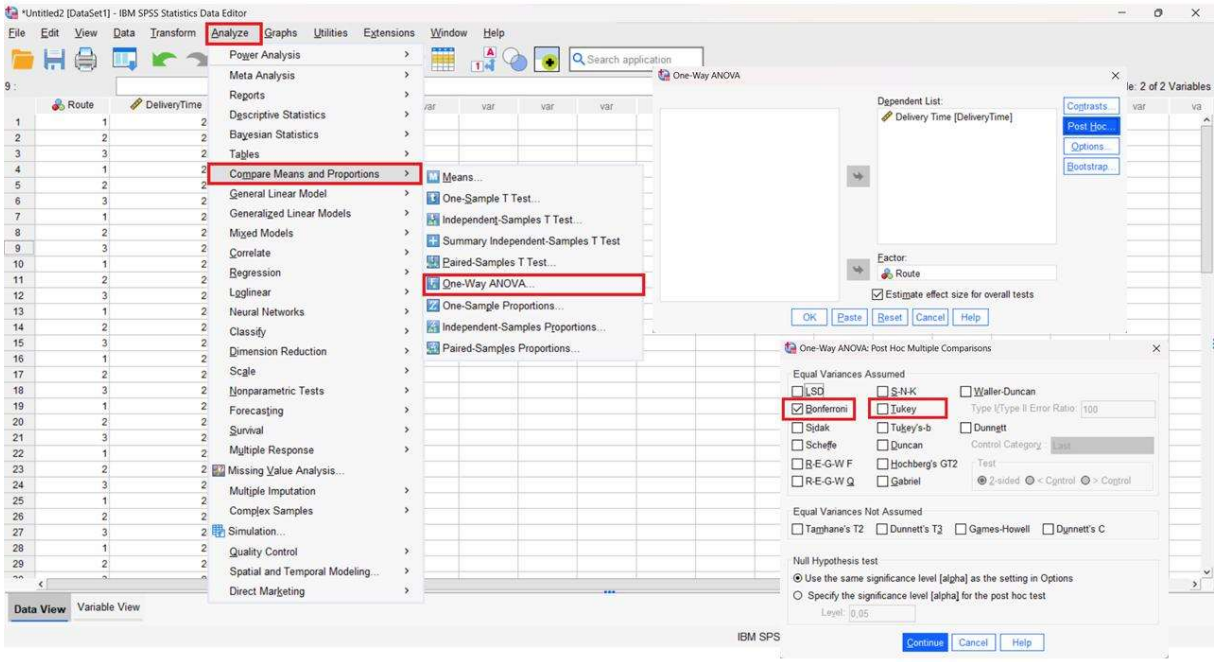


Figure 7.18 ANOVA settings.

The results of our analysis indicate that our *F*-statistic value is 11.173 (higher values indicate more variations between groups) and *p*-value <0.001, which means that our Null hypothesis is rejected (see figure 7.19). Since there is a significant difference between the three routes (<0.001), a post hoc test is also valid in our case (George & Mallery, 2022). After conducting the Bonferroni correction test, we can see that the best *p*-values are noted in the case of route 2 (refer to figure 7.19). In the report, we can conclude that "There was a significant difference in choosing a delivery route in correlation to delivery times (1-way ANOVA, *F*=11.173, *df* = 47, *p*-value = <0.001). Route 2 had the best delivery time results."

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.

Figure 7.19 ANOVA initial results and Post Hoc Test results.

We conclude this chapter of the book with the understanding that we have covered some of the more common tests in this chapter. There are still other tests, such as Repeated Measures ANOVA, reliability tests, and sensitivity tests, which can also be modelled and Analysed using SPSS software. These additional tests provide a broader range of tools for data analysis and draw meaningful insights into various research and practical applications.

REFERENCES CHAPTER 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 [available at: https://students.shu.ac.uk/lits/it/documents/pdf/analysing_data_using_spss.pdf, access 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 [available at: https://www.ibm.com/docs/en/SSLVMB_28.0.0/pdf/IBM_SPSS_Statistics_Brief_Guide.pdf, access 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 [available at: https://med.und.edu/research/daccota/_files/pdfs/berdc_resource_pdfs/data_analysis_using_spss.pdf, access October 26, 2023]