Polona Tominc, Vesna Čančer in Maja Rožman

Research Methods – Data Analysis Techniques

University of Maribor Press

University of Maribor

Faculty of Economics and Business

Research Methods – Data Analysis Techniques

Authors

Polona Tominc

Vesna Čančer

Maja Rožman

January 2026

Title **Research Methods – Data Analysis Techniques**

Authors Polona Tominc
(University of Maribor, Faculty of Economics and Business)

Vesna Čančer
(University of Maribor, Faculty of Economics and Business)

Maja Rožman
(University of Maribor, Faculty of Economics and Business)

Review Blaž Frešer
(University of Maribor, Faculty of Economics and Business)

Dijana Oreški
(University of Zagreb, Faculty of Organization and Informatics)

Language editing Janja Šmid
(University of Maribor, Faculty of Economics and Business)

Technical editors Marina Bajić
(University of Maribor, University of Maribor Press)

Jan Perša
(University of Maribor, University of Maribor Press)

Graphics material Sources are own unless otherwise noted. Tominc, Čančer, Rožman, (authors), 2026

Cover designer Jan Perša
(University of Maribor, University of Maribor Press)

Cover graphic Hand, Type, Keyboard, author: nattanan23, pixabay.com, 2017

Published by **University of Maribor**
Založnik **University of Maribor Press**
Slomškov trg 15, 2000 Maribor, Slovenia
https://press.um.si, zalozba@um.si

Issued by **University of Maribor**
Izdajatelj **Faculty of Economics and Business**
Razlagova ulica 14, 2000 Maribor, Slovenia
https://www.epf.um.si, epf@um.si

Edition First edition

Published at Maribor, Slovenia, January 2026

Publication type E-book

Available at https://press.um.si/index.php/ump/catalog/book/1079

```
CIP - Kataložni zapis o publikaciji
Univerzitetna knjižnica Maribor

001.891:311(0.034.2)

TOMINC, Polona
    Research methods - data analysing techniques [Elektronski vir] / authors Polona
Tominc, Vesna Čančer, Maja Rožman. - 1st ed. - E-publikacija. - Maribor : University of
Maribor, University of Maribor Press, 2026

Način dostopa (URL): https://press.um.si/index.php/ump/catalog/book/1079

ISBN 978-961-299-098-5 (PDF)
doi: 10.18690/um.epf.1.2026
COBISS.SI-ID 264425987
```

ISBN 978-961-299-098-5 (pdf)

DOI https://doi.org/10.18690/um.epf.1.2026

Price Free copy

For publisher Prof. Dr. Zdravko Kačič,
Rector of University of Maribor

Attribution Tominc, P., Čančer, V. Rožman, M. (2026). *Research Methods – Data Analysis Techniques*. University of Maribor, University of Maribor Press. doi: 10.18690/um.epf.1.2026

Table of Contents

University of Maribor Press

Preface

POLONA TOMINC, VESNA ČANČER, MAJA ROŽMAN
monograph authors

In the age of data-driven decision-making, the ability to understand, analyse, and interpret data is no longer reserved for statisticians or data scientists alone. It has become a vital skill for all professionals in business and economics. This tutorial is designed to support master's students in developing solid foundations in applied statistical methods and data analysis techniques using one of the most widely used software tools in academic and business research.

The content is built around a practical, example-based approach that reflects real-world data, contemporary challenges, and hands-on exercises to support experiential learning. The tutorials cover a wide range of topics, from descriptive statistics and sampling strategies to multivariate analysis techniques, including factor analysis, regression analysis, discriminant analysis, and Monte Carlo simulations. Each chapter builds on core concepts, gradually guiding students toward more advanced applications, while fostering critical thinking and analytical skills.

This collection of exercises is primarily intended for use with the statistical software package SPSS in selected thematic areas that fall within the framework of the master's study programme in Economic and Business Sciences, with a specialization in Data Science in Business (conducted in English) at the University of Maribor, Faculty of Economics and Business.

The material supports the course Research Methods – Data Analysis Techniques, and serves as a practical supplement to the core study literature and lecture content. It offers students a structured, hands-on approach to using the SPSS statistical software for analysing real-world business and economic data.

The exercises are carefully designed to guide students through key concepts in descriptive statistics, sampling, and inference, to foster critical thinking and statistical reasoning. In addition to understanding the classification of quantitative methods based on data characteristics and research hypotheses, students will learn to interpret SPSS outputs and apply them in a decision-making context.

Through this material, students develop practical skills in data analysis, strengthen their data literacy, and gain confidence in applying statistical techniques to support business insights and evidence-based management. The tasks cover topics relevant to modern data-driven environments, such as digital transformation, screen time behaviours, and investment trends, ensuring both academic depth and practical relevance. We believe that students best acquire methodological knowledge when statistical theory is accompanied by meaningful, data-rich context and a clear research narrative. Therefore, the exercises in this tutorial are based on realistic datasets addressing themes such as digital transformation, AI adoption, employee productivity, and market behaviour.

We encourage students to engage actively with the material, discuss their interpretations, and connect statistical results to broader business challenges. We wish all students a productive learning experience and successful application of these methods in their academic and professional journeys.

This tutorial builds upon the earlier editions of the exercise collection, particularly the one introduced in the 2023–2024 academic year for the Research Methods course[1]. While the core structure and pedagogical approach remain consistent, this updated version has been significantly enriched to reflect recent developments in data science and digital transformation. New examples and datasets have been introduced to enhance relevance, and expanded guidance is provided to support students specializing in Data Science in Business. The revision aims to provide a more comprehensive and practice-oriented experience aligned with current research and industry challenges.

[1] Tominc, P., Čančer, V., Rožman, M. (2018). *Zbirka vaj za predmet Metode raziskovanja.* 1. izd. Maribor: Univerzitetna založba Univerze v Mariboru.
Tominc, P., Čančer, V., Rožman, M. (2024). *Research Methods.* University of Maribor, Faculty of Economics and Business.

University of Maribor Press

1 Descriptive Statistics and Sampling Approach

In research methodology, descriptive statistics and sampling represent the foundation for understanding data structures, summarising key characteristics, and drawing initial insights from collected information. These methods are essential for analysing empirical data in both academic and business contexts, especially when working with statistical software such as SPSS (Gravetter and Wallnau, 2017).

Descriptive statistics refer to the methods used to organise, summarise, and present data in an informative way (Goos and Meintrup, 2015). They provide a concise overview of the dataset and allow researchers to identify patterns, central tendencies, and variability. Commonly used measures include:

- **Measures of central tendency:** mean, median, and mode;
- **Measures of variability:** standard deviation, variance, and range;
- **Shape of distribution:** skewness and kurtosis;
- **Graphical representation:** frequency tables and histograms.

These statistics do not allow conclusions beyond the analysed data, but they provide a solid foundation for further inferential techniques (Triola, 2022). Descriptive statistics are particularly useful in exploratory stages of research, when the goal is to get an initial understanding of the dataset before applying more complex methods (Weiss, 2021). On the other hand, inferential or mathematical statistics use a random sampling approach and focus on testing assumptions about statistical parameters of a statistical population based on collected data from a random sample (Tabachnick and Fidell, 2013).

1.1 Statistical Unit, Statistical Population and Random Sample

A statistical unit is an individual element of a statistical population that is the subject of observation (e.g., a company, household, worker, student, etc.) at a given point in time or interval. Clearly defining the statistical unit is crucial because it influences the structure of the dataset, the validity of statistical conclusions, and the generalizability of the results to the broader population (Frost, 2019).

A statistical population is a set of statistical units that meet certain defining criteria or characteristics. A statistical population is also called a population. A sample is a part of the entire population from which inferences about the entire population are made. The fundamental principle here is that the sample must be random. A characteristic of a random sample is that each element in the population has a known and non-zero probability of being included or selected in the random sample. This probability is known in advance (Tominc and Kramberger, 2007).

1.2 Statistical Variables

A statistical variable describes the characteristic of a statistical unit. Variables can be classified in several ways, depending on their nature and permissible values (Gravetter and Wallnau, 2017).

Types of variables (Artenjak, 2003):

- Descriptive (attribute) variables are those variables whose values can only be described in words (e.g., gender: male, female).
- Numeric variables are those variables whose values can be expressed with numbers (e.g., age, company profit). Among them, we distinguish:
    - continuous variables, which are numeric variables that can take any value within an interval (e.g., length, time, weight), and
    - discrete variables (or non-continuous variables), which are numeric variables that can take only certain finite, usually integer, values (e.g., the number of household members).

Types of variables based on the type of measurement (measurement scales) (Bastič, 2006):

- Descriptive variables are measured on a nominal and ordinal measurement scale, while numeric variables are measured on an interval and ratio measurement scale.
- Nominal variables are measured on a nominal measurement scale, which allows the classification of units based on a specific characteristic. Statistical units are classified into groups so that units classified in the same group have the same characteristic (e.g., gender: 1 – male, 2 – female; response: 1 – yes, 2 – no).
- Ordinal variables are measured on an ordinal measurement scale, which allows groups to be classified based on a criterion, meaning that values can be ordered from smallest to largest (e.g., education level, success, company size: 1 – small company, 2 – medium-sized, 3 – large company).
- Interval variables are measured on an interval measurement scale, which uses a unit of measure. It is divided into equally sized intervals between its initial and final points (e.g., temperature in °C).
- Ratio variables are measured on a ratio measurement scale, which has a starting point of 0 and does not change. This measurement scale is absolute, and the difference is always measured from point zero (e.g., the number of tourists, age, income).

Table 1 presents a summary of the main types of statistical variables based on their measurement scale, along with typical examples and recommended statistical procedures that can be performed in SPSS. Proper identification of variable types is crucial for valid data analysis and interpretation of results.

Table 1: Types of Statistical Variables and SPSS Applications

Variable Type	Measurement Scale	Example	Statistical Procedures in SPSS
Nominal	Nominal	Gender (1 – male, 2 – female); yes/no questions	Frequencies, mode, chi-square test
Ordinal	Ordinal	Education level, satisfaction scale	Median, frequencies, non-parametric tests (e.g., Mann-Whitney U)
Interval	Interval	Temperature in °C	Mean, standard deviation, correlation, regression
Ratio	Ratio	Income, age, number of employees	Mean, standard deviation, variance, correlation, regression, ANOVA
Discrete	Usually ratio	Number of children, number of contracts	Frequencies, Poisson regression, descriptive statistics
Continuous	Interval or ratio	Weight, height, revenue	All parametric tests, descriptive statistics, histograms, confidence intervals

1.3 Basic Statistical Parameters and Statistics

A parameter is a numerical or descriptive value that characterizes a specific feature of a statistical population. Parameters refer to population-level metrics, such as the population mean, population variance, or population proportion. These values are typically fixed but unknown, as it is often impractical or impossible to measure the entire population directly (Triola, 2022; Gravetter & Wallnau, 2017). Therefore, a parameter is a numerical or descriptive value that describes a characteristic of a statistical population. Statistics is a numerical or descriptive value that estimates a characteristic of a statistical population and is obtained from a sample (Artenjak, 2003).

Measures of central tendency are values commonly used in basic descriptive statistical analysis. Measures of central tendency represent all observed values. Among the most important measures of central tendency are the mean, median, and mode.

- Mean (arithmetic mean) is the average value obtained by dividing the sum of all variable values by the number of units in the collected data (n).
- Median is the middle value, with half of the units having smaller or equal values and half having larger values. It is denoted as Me.
- Mode or modal value is the value of the variable that occurs most frequently. It is denoted as Mo (ibid).

Among the most common measures of variability are the range, variance, and standard deviation.

- Range is the difference between the maximum and minimum values of the variable.
- Variance measures the deviations of individual variable values from the mean. It is defined as the average of the squares of the deviations of individual values from the mean.
- Standard deviation is defined as the square root of the variance. With standard deviation, we can measure how the values are spread around the mean of the collected data; it is expressed in the same units of measurement as the observed variable. If we observe multiple groups of units for the same variable, a higher standard deviation indicates greater dispersion of units in the sample, whereas a lower value indicates less dispersion of units and a greater concentration of units around the mean (ibid).

Skewness and kurtosis are measured with skewness and kurtosis measures. Asymmetric distributions (skewness) can be skewed to the right (positive skewness), characterised by greater density at smaller variable values, or skewed to the left (negative skewness), characterised by greater density at larger variable values. The skewness coefficient is less than zero if the distribution of the variable is skewed to the left; for skewness to the right, the skewness coefficient is greater than 0. The more the skewness coefficient differs from the value of 0, the greater the strength of the skewness. For most empirical distributions, the skewness coefficient typically ranges between −3 and +3 (Artenjak, 2003). Kurtosis of the distribution (kurtosis) is compared to a normal distribution, which is said to be normally kurtotic. If the distribution is more peaked than the normal distribution, it is said to be leptokurtic (has longer tails and a narrower central part). If the distribution is flatter than the normal distribution, it is said to be platykurtic. The kurtosis coefficient indicates that when it is greater than 0, it suggests a leptokurtic distribution, and when the kurtosis coefficient is less than 0, it suggests a platykurtic distribution. In a normal distribution, both skewness and kurtosis coefficients are equal to 0. The values of both coefficients (among other things) show how the studied distribution of variable values differs from a normal distribution (Bastič, 2006).

1.4 Data Analysis

1.4.1 Data Entry Into the SPSS Programme

Task 1

One of the key factors linked to digital behaviour is the amount of time spent daily in front of screens. In January 2025, researchers from the Faculty of Economics and Business at the University of Maribor conducted a short study to explore daily screen time among students enrolled in the master's programme Economic and Business Sciences. A sample of 15 students was surveyed to determine their average daily screen time during the exam preparation period.

Table 2 presents the self-reported average daily screen time (in hours) for 15 students enrolled in the master's programme at the Faculty of Economics and Business, University of Maribor.

Table 2: Average Daily Screen Time (In Hours) for 15 Students Enrolled in the Master's Programme at the Faculty of Economics and Business

Student ID	Screen Time (Hours per Day)
1	3.5
2	4.0
3	5.2
4	6.8
5	4.9
6	7.3
7	6.1
8	5.5
9	3.9
10	4.4
11	6.5
12	7.0
13	4.7
14	5.9
15	3.8

a) Define the statistical unit, statistical variable, and sample size.

b) Enter the data on the average daily screen time (in hours) for the 15 students into the SPSS programme. Calculate and interpret the results of descriptive statistics for the variable *screen time.*

c) Based on skewness and kurtosis measures, explain whether the values of the variable *screen time* appear to be normally distributed.

d) For the variable *screen time*, create a frequency table.

e) Draw a frequency histogram with a fitted normal curve for the variable *screen time.*

Data Entry Procedure

Enter the data on the average daily screen time (in hours) for the 15 students into SPSS by clicking on the bottom-right box, *Variable View*. In the *Name* box, enter the variable name *screen time* specify the variable type (numeric) in the *Type* box, determine the number of decimal places in the *Decimals* box, and provide a full variable name (*screen time* (hours per day)) in the *Label* box. Then click on the bottom-left box, *Data View*, and enter the variable values into the column. Calculation of Descriptive Statistics: Click on *Analyze*, then *Descriptive Statistics*, and select *Descriptives* (or *Frequencies*). In the opened dialog box, click on the variable *screen time* and transfer it to the right box under *Variable(s)*. Click on *Options* and choose the statistics you want to display in the results.

Answers and Display of Results

a)

Statistical unit: 1 student enrolled in the master's programme in January 2025.

Statistical variable: Average daily screen time (numeric, continuous variable).

Sample size: 15 students.

b) and c)

Table 3: Descriptive Statistics for the Variable Average Daily Screen Time (In Hours) for 15 Students Enrolled in the Master's Programme at the Faculty of Economics and Business

N	Valid	15
	Missing	0
Mean		5.300
Std. Error of Mean		0.3244
Median		5.200
Mode		3.5[a]
Std. Deviation		1.2564
Variance		1.579
Skewness		0.149
Std. Error of Skewness		0.580
Kurtosis		−1.339
Std. Error of Kurtosis		1.121
Range		3.8
Minimum		3.5
Maximum		7.3

[a] Multiple modes exist. The smallest value is shown.

Table 3 presents the descriptive statistics for the variable *screen time* (average daily screen use in hours) among the 15 master's students from the Faculty of Economics and Business. We can observe that 15 students were included in the sample (*n = 15*), and there are no missing values (*Missing = 0*). The **mean** screen time is 5.300 hours per day, indicating the average amount of time spent in front of screens per student. The **standard error of the mean** is 0.3244 hours, which reflects the variability of the sample mean and indicates how well this sample represents the population of students (the smaller the value, the less variability exists between sample mean values, and the better the sample represents the statistical population). The **median** is 5.200 hours, which means that 50 % of the students spend less than or equal to 5.2 hours in front of screens daily, while the other 50 % spend more than 5.2 hours in front of screens daily. The **mode** is reported as 3.5 hours. In this case, no true mode exists because all values occur only once. SPSS automatically displays the smallest value among the multiple values with the same (lowest) frequency. The **standard deviation** is 1.2564 hours, which is the square root of the variance (1.579 hours^2) and shows how much the individual values deviate from the mean. The skewness coefficient **(skewness)** is 0.149 hours, indicating a right-skewed distribution (positive skewness). The kurtosis coefficient **(kurtosis)** is -1.339 hours, suggesting a flattened distribution (negative value). The **range** is 3.8 hours, calculated as the difference between the **maximum** value (7.3 hours) and the **minimum** value (3.5 hours) of daily screen time.

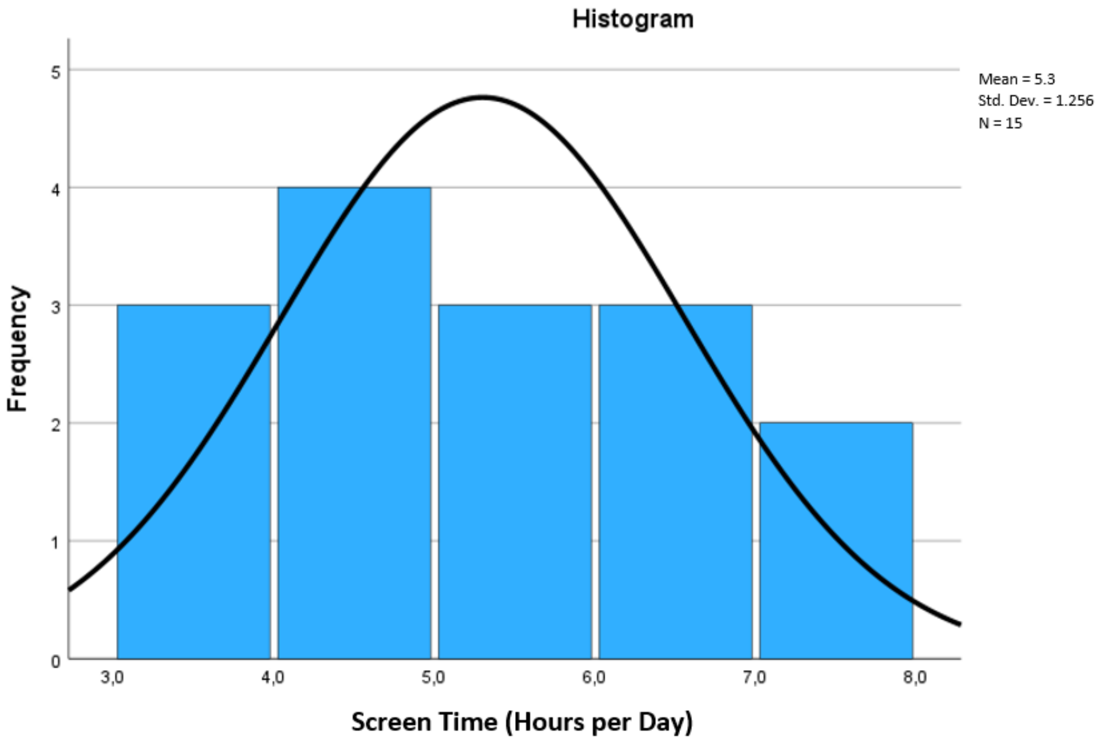
d)

Table 4: The Frequency Distribution (Data Array – Each Variable Value Is Observed Only Once) of the Variable Screen Time (In Hours) for the 15 Students

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	3.5	1	6.7	6.7	6.7
	3.8	1	6.7	6.7	13.3
	3.9	1	6.7	6.7	20.0
	4.0	1	6.7	6.7	26.7
	4.4	1	6.7	6.7	33.3
	4.7	1	6.7	6.7	40.0
	4.9	1	6.7	6.7	46.7
	5.2	1	6.7	6.7	53.3
	5.5	1	6.7	6.7	60.0
	5.9	1	6.7	6.7	66.7
	6.1	1	6.7	6.7	73.3
	6.5	1	6.7	6.7	80.0
	6.8	1	6.7	6.7	86.7
	7.0	1	6.7	6.7	93.3
	7.3	1	6.7	6.7	100.0
	Total	15	100.0	100.0	

Table 4 shows that each of the 15 students in the sample reported a unique value for their average daily screen time. For example, one student reported spending 3.5 hours per day in front of screens (or 6.7 %), another 4.0 hours, and yet another 6.8 hours. Each screen time value occurred exactly once, which is why the frequency for all values is 1 (or 6.7 % of the sample).

e)



Histogram 1: Frequency Histogram With a Fitted Normal Distribution Curve

Histogram 1 displays the frequency distribution of the variable *screen time*, including a fitted normal curve. The histogram provides a visual representation of the distribution of screen time among students. The normal curve overlay helps to assess whether the data follow a bell-shaped (normal) distribution. This visual impression is consistent with the values of skewness and kurtosis discussed above.

Task 2

Digital transformation has become a key focus for companies across Europe. In 2024, a research team collected data from 11 medium-sized Slovenian companies to explore how much of their total annual investment was allocated to digital technologies (e.g., software, automation tools, digital infrastructure). Table 4 shows the data for the variable *digital investment* (in %) for the 11 Slovenian medium-sized companies.

Table 5: Data for the Variable "Digital Investment" (In %) for the 11 Medium-Sized Companies

Company ID	Digital Investment (% of Total Annual Investment)
1	15.0
2	22.5
3	18.0
4	25.0
5	12.5
6	20.0
7	10.0
8	30.0
9	14.5
10	16.0
11	22.5

a) Define the statistical unit, statistical variable, and sample size.

b) Enter the data on the share of annual investment in digital technologies (in %) for the 11 medium-sized companies into the SPSS programme. Calculate and interpret the results of descriptive statistics for the variable *digital investment*.

c) Based on skewness and kurtosis measures, explain whether the values of the variable *digital investment* appear to be normally distributed.

d) For the variable *digital investment*, create a frequency table.

e) Draw a frequency histogram with a fitted normal curve for the variable *digital investment*.

1.4.2 Sampling and Interval Estimation of Parameters

Sampling

For the effective design of random samples, various techniques of probability sampling are used in practice. In simple random sampling, one formerly common method for selecting units was the use of a table of random numbers – now considered archaic – while today, random number generators are the standard tool for ensuring randomness. Among the most widespread are systematic sampling, multistage sampling, and probability proportional to size sampling, which is proportional to the size of strata (subsets) of the statistical population. A random sample ensures that each statistical unit in the population has an equal probability of being selected into the sample. The probability of selecting an individual unit into the sample is known or can be calculated (Agresti and Finlay, 2009).

Statistical sampling theory is based on random samples, characterised by the ability to assess the quality of sample estimates of statistical parameters with appropriate indicators and defining the probability of an incorrect conclusion based on data from a random sample. Statistical theory is based on random samples to ensure objectivity in the selection of units for observation, representativeness of the sample, and the ability to determine the quality of estimates for parameters obtained from sample data (ibid).

Confidence Interval Estimation of the Mean

The sample estimate of the mean of the statistical population is a point estimate. This is an estimate of the value of a statistical parameter given by a single value calculated from a random sample. In addition to point estimation, we also estimate the value of the statistical parameter with interval estimation, which is based on the point estimate of the sample statistical parameter (the result from the sample is generalized to the statistical population). Based on sample data, we determine an interval within which the value of the statistical parameter is expected with a certain level of probability. The latter is called the confidence level (usually 90 %, 95 %, or 99 %), and the interval is called the confidence interval. The probability (100 % minus the confidence level) is called the risk level (%). If we determine the limits of the interval for the value of the statistical parameter from sample data by specifying both upper and lower limits of the interval at a given confidence level, we speak of a two-sided estimation of the statistical parameter. If we set only one of the two limits, either upper or lower, at a given confidence level, we speak of a one-sided estimation of the statistical parameter (Artenjak, 2003).

Task 3

A research team analysed the monthly revenue (in EUR) of a random sample of 12 digital marketing agencies in Slovenia. The goal was to estimate the average monthly revenue for such companies in 2024, using a confidence interval approach.

Calculate and explain the 95 % and 80 % confidence intervals for the variable *monthly revenue (in EUR)*.

Table 6: Monthly Revenue (In EUR)

Agency ID	Monthly Revenue (EUR)
1	12,500
2	14,200
3	13,000
4	15,300
5	12,800
6	13,500
7	14,700
8	11,900
9	13,800
10	12,400
11	14,300
12	13,200

Procedure: Click on *Analyze*, then *Descriptive Statistics*, and choose the *Explore* function. Transfer the variable *monthly revenue* from the left window to the *Dependent List* window on the right and click on the *Statistics* box. Check *Descriptives* and enter 95 % (for an 80 % interval, enter 80 %) for the *Confidence Interval for Mean*.

Table 7: 95 % Confidence Interval for the Variable Monthly Revenue

			Statistic	Std. Error
Monthly Revenue (EUR)	Mean		13,466.67	294.992
	95 % Confidence Interval for Mean	Lower Bound	12,817.39	
		Upper Bound	14,115.94	
	5 % Trimmed Mean		13,451.85	
	Median		13,350.00	
	Variance		1,044,242.42	
	Std. Deviation		1,021.882	
	Minimum		11,900	
	Maximum		15,300	
	Range		3,400	
	Interquartile Range		1,700	
	Skewness		.273	.637
	Kurtosis		−733	1.232

Table 7 displays the descriptive statistics and the 95 % confidence interval for the variable *monthly revenue (in EUR)* based on a sample of the 12 digital marketing agencies. The mean value of monthly revenue is 13,466.67 EUR, with a standard error of approximately 294.99 EUR. This indicates the extent to which the sample mean would vary if different samples were taken from the same population. In Table 7, we can see that the lower bound of the confidence interval is 12,817.39 EUR, and the upper bound is 14,115.94 EUR. This means that with 95 % probability, we estimate that the average monthly revenue of digital marketing agencies in the statistical population is between 12,817.39 EUR and 14,115.94 EUR.

Table 8: 80 % Confidence Interval for the Variable Monthly Revenue

			Statistic	Std. Error
Monthly Revenue (EUR)	Mean		13,466.67	294.992
	80 % Confidence Interval for Mean	Lower Bound	13,064.47	
		Upper Bound	13,868.87	
	5 % Trimmed Mean		13,451.85	
	Median		13,350.00	
	Variance		1,044,242.42	
	Std. Deviation		1,021.882	
	Minimum		11,900	
	Maximum		15,300	
	Range		3,400	
	Interquartile Range		1,700	
	Skewness		.273	.637
	Kurtosis		−.733	1.232

With an 80 % probability, we estimate that the average monthly revenue of digital marketing agencies in the statistical population is between 13,064.47 EUR and 13,868.87 EUR (Table 8).

Task 4.

Table 9 presents the data for n = 15 surveyed individuals who answered the question: "Would fear of failure prevent you from starting your own business?" Respondents answered the question with the following possible responses: 0 – no and 1 – yes. The gender of the respondent is indicated by 1 – male and 2 – female. Age is measured in completed years.

Table 9: Data on Responses From Surveyed Individuals

Gender	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2
Age	45	45	58	62	64	55	57	20	23	20	21	33	25	33	30
Fear of Failure	0	0	0	0	0	1	1	0	0	1	1	1	1	1	1

a) Enter the data into the SPSS programme. Split the file into two parts according to the respondents' gender, and perform a separate analysis for men and women to determine the percentage of those who would be deterred by fear of failure from starting a new business.

Procedure for Splitting the File Into Two Parts (According to Gender)

Click on *Data*, then *Split File* (this command will split the file into two parts according to gender), click on *Organize output by groups*, and move the gender variable to the right box (where it says *Groups Based On*), then click *OK*.

Next, perform a separate analysis for men and women: Click on *Analyze*, then choose the *Descriptive Statistics* command, and click on *Descriptives* (you can also choose *Frequencies* or *Explore*).

b) Create a new variable that will have a negative value if a person is under 40 years old and a positive value if they are older than 40. Convert this variable into a new variable that will have a value of 0 if the person is up to 40 years old and a value of 1 if the person is older than 40.

Procedure for Creating a New Variable Based on Age

Click on *Transform* and choose the *Compute Variable* command and a new window will open. In this window, enter the name of the new variable you want to create, for example, Age1, under *Target Variable* at the top left. In the *Numeric Expression* box, enter the mathematical expression to calculate the new variable (in this case, enter: age – 40 to obtain the new variable – *Age1*) and click *OK*.

In the *Data View* of the data file, you can find the values of the newly formed variable named Age1 in the last column. Also, in *Variable View*, you can see the newly created variable *Age1*.

Procedure for Converting It Into a New Variable With a Value of 0 if the Person Is up to 40 Years Old and a Value of 1 if the Person Is Older Than 40

Click on *Transform* and choose the *Recode into Different Variables* command. Move the variable you want to transform – *Age1* to the *Numeric Variable* box. In the *Output Variable* box below the line *Name*, enter the name of the new variable (e.g., *Age2*), click on the *Change* button and then on the *Old and New Values* button. A window will open, showing the old values (*Old Value*) on the left, and the new values (*New Value*) on the right. All values of the variable that are lower than or equal to 0 in the variable *Age1* must be set to 0 in the new variable *Age2*. Therefore, on the left side (*Old Value*), choose *Range*, *lowest through value*: type *0*; on the right side (*New Value*), choose 0, then click *Add*. Repeat the process for all values greater than 0: on the left side (*Old Value*), choose *All other values*, and on the right side, select *Value*: 1, then click *Add*, *Continue*, and *OK*.

c) Calculate an 80 % confidence interval for the average age of individuals in the statistical population.

Answers and Result Outputs

a)

Table 10: Descriptive Statistics for Surveyed Individuals of Male Gender

	N	Range	Minimum	Maximum	Mean		Std. Deviation	Variance
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic
Fear of Failure	7	1	0	1	0.29	0.184	0.488	0.238
Valid N (Listwise)	7							

Table 11: Descriptive Statistics for Surveyed Individuals of Female Gender

	N	Range	Minimum	Maximum	Mean		Std. Deviation	Variance
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic
Fear of Failure	8	1	0	1	0.75	0.164	0.463	0.214
Valid N (Listwise)	8							

Among men, 29 % would be deterred from starting a business by the fear of failure, whereas among women, the corresponding percentage is 75 %.

c)

Table 12: 80 % Confidence Interval for the Average Age of Surveyed Individuals

			Statistic	Std. Error
Age	Mean		39.40	4.263
	80 % Confidence Interval for Mean	Lower Bound	33.67	
		Upper Bound	45.13	
	5 % Trimmed Mean		39.11	
	Median		33.00	
	Variance		272.543	
	Std. Deviation		16.509	
	Minimum		20	
	Maximum		64	
	Range		44	
	Interquartile Range		34	
	Skewness		0.236	0.580
	Kurtosis		−1.656	1.121

2 Normal Distribution

2.1 Definition and Characteristics of Normal Distribution

Normal distribution represents one of the most essential and extensively applied probability distributions in both theoretical and applied statistics. Its significance arises from its well-defined mathematical characteristics and its broad applicability to various natural and social phenomena, in which measured variables typically exhibit symmetric dispersion around a central tendency (Pham-Gia, 2022).

In inferential statistics, which is concerned with making generalizations about population parameters based on sample data, the normal distribution occupies a central role. Numerous statistical procedures, such as hypothesis testing, constructing confidence intervals, regression analysis, and analysis of variance (ANOVA), rely on the assumption that the data or the sampling distribution of the relevant statistic follows a normal distribution (Lehenbauer, 2022).

One of the fundamental reasons for the central role of the normal distribution in statistical inference lies in the Central Limit Theorem. This theorem postulates that, irrespective of the shape of the original population distribution, the sampling distribution of the sample mean (or sum) will approximate a normal distribution as the sample size increases. The convergence toward normality occurs even when the population data are skewed or non-normal, provided the sample size is sufficiently large (Agresti and Finlay, 2009; Frost, 2019).

The bell-shaped curve of the normal distribution is symmetric around the mean, and the standard deviation determines its spread. Approximately (Figure 1):

- 68 % of the data fall within ± 1 standard deviation from the mean,
- 95 % within ± 2 standard deviations,
- 99.7 % within ± 3 standard deviations.

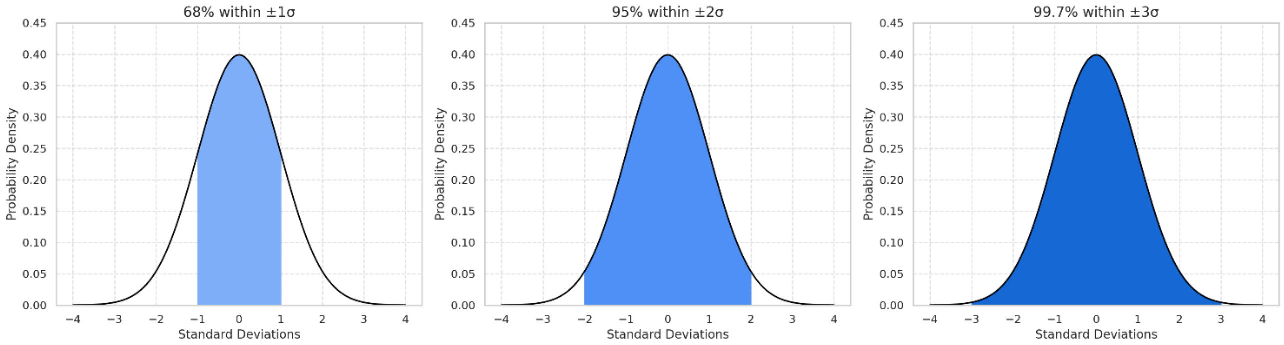


Figure 1: Normal Distribution and Standard Deviations

Due to its mathematical characteristics and symmetry, the normal distribution serves as a reference model in statistical analysis. Its properties are frequently used to identify potential outliers, evaluate data quality, and assess whether the assumptions of various parametric statistical methods are met. In empirical research, statistical software such as SPSS offers multiple approaches for testing the normality of data. These include descriptive measures (e.g., skewness and kurtosis), graphical methods (e.g., histograms and Q–Q plots), and formal statistical tests (e.g., the Shapiro–Wilk and Kolmogorov–Smirnov tests). Determining whether the distribution of a variable approximates normality is essential for the correct selection of statistical procedures and for ensuring the validity and reliability of analytical results.

Normal distribution is a probability distribution of values of statistical units in a statistical population, and it is schematically represented in Figure 1. Figure 1 illustrates the normal distribution curve, a key concept in probability and statistics. In a perfectly normal distribution, the mean, median, and mode are all equal and located at the centre of the distribution. This property is visually represented in the diagram, where the blue dashed line marks the central value shared by all three measures of central tendency. In a normal distribution, the mode is equal to the mean, and due to symmetry, the median is also equal to both of them. Values to the left and right (up and down) of the mean have decreasing probability density. For normal distribution, skewness and kurtosis coefficients are both equal to 0.

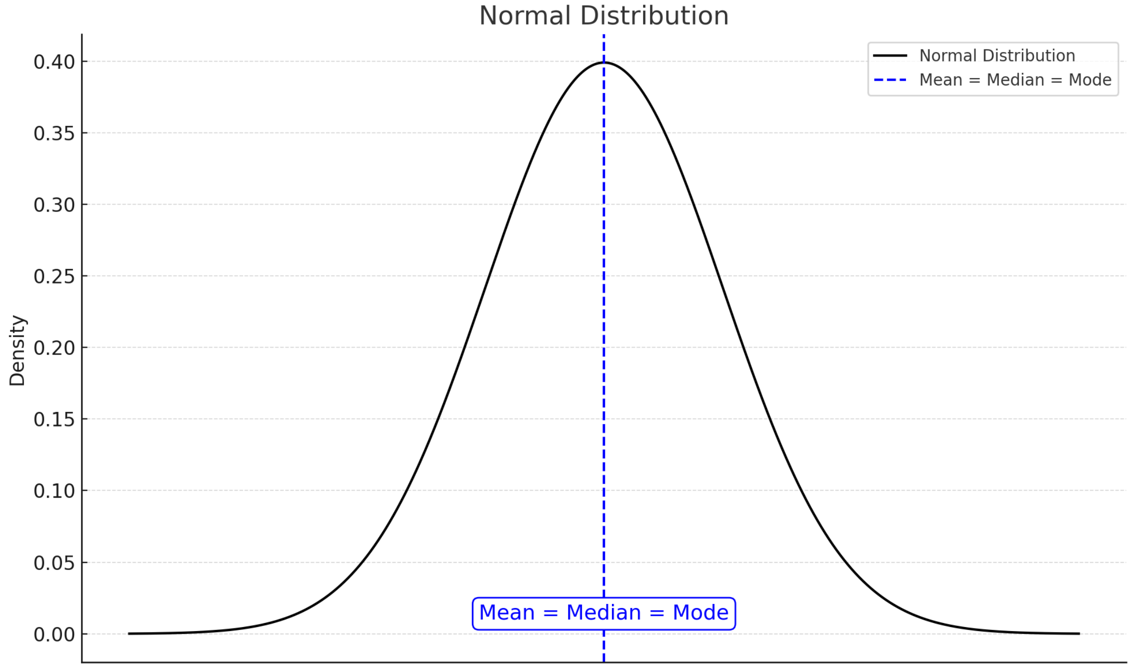


Figure 2: Representation of Normal Distribution

The normal distribution is a fundamental probability distribution characterised by a bell-shaped, symmetric, and unimodal curve that extends infinitely in both directions, approaching but never touching the x-axis (i.e., it is asymptotic). Often referred to as the Gaussian distribution, it plays a central role in both theoretical and applied statistics due to its mathematical properties and widespread occurrence in empirical data. The normal distribution has its peak at the mean, and the density of relative frequency is highest around this point, decreasing with distance from the mean. The standard deviation alters the spread of the normal distribution—increasing the standard deviation widens and lowers the curve. The standard deviation measures the dispersion of values around the mean, influencing the flatness of the curve (the larger the standard deviation, the flatter the curve) (Lehenbauer, 2022; Pham-Gia, 2022).

2.2 Kolmogorov-Smirnov Test and Shapiro-Wilk W Test

The Kolmogorov-Smirnov test and Shapiro-Wilk W test are used to verify whether the examined variable in a statistical set is normally distributed. The Kolmogorov-Smirnov test and the Shapiro-Wilk test are commonly used statistical procedures to assess whether a given variable follows a normal distribution. These tests are part of formal normality tests that compare the empirical distribution of the sample data with a theoretical normal distribution (Uhm and Yi, 2023; Wagner, 2019).

More specifically, the Kolmogorov-Smirnov test evaluates the maximum distance between the empirical cumulative distribution function of the sample and the cumulative distribution function of a normal distribution with the same mean and standard deviation. It is suitable for larger samples but is generally less powerful for small sample sizes. The Shapiro-Wilk test, on the other hand, is specifically designed for small to moderate sample sizes and is considered one of the most sensitive tests for normality. It tests the null hypothesis that a sample comes from a normally distributed population by comparing the ordered sample values with the expected values under normality (Burns and Burns, 2008; Field, 2017).

In both tests, a non-significant result ($p > 0.05$) indicates that the assumption of normality is not rejected, suggesting that the sample distribution does not significantly differ from a normal distribution. Conversely, a significant result ($p < 0.05$) suggests a deviation from normality, which may require the use of non-parametric methods in further analysis. Thus, in the case when the test is non-significant (test significance level or risk level, $p > 0.05$), we do not reject the null hypothesis (H0: The considered variable can be reasonably adjusted to a normal distribution.), and we can conclude that the distribution of the studied variable in the statistical set does not differ from a normal distribution. We conclude that the studied variable is not normally distributed when the test is statistically significant ($p < 0.05$).

Task 1

Artificial intelligence (AI) has become a transformative force in modern business environments, driving innovation, automation, and data-driven decision-making. To explore current trends in AI readiness, this study included 300 employers from large companies. Participants were asked to self-assess their level of AI knowledge on a scale from 1 (very limited understanding) to 10 (advanced understanding and practical experience).

The data is in the file *Normal Distribution_ Artificial Intelligence.sav.*

a) Write the null hypothesis about the normal distribution of the value of the variable under consideration.

b) Check whether the variable is distributed according to a normal distribution and plot a frequency histogram for the variable *AI knowledge.*
Procedure

Click on the *Analyze* tab, then select *Descriptive Statistics*, and *Explore*. In the *Dependent List* box, transfer the variable *AI knowledge*. Click the *Plots* button and select *Normality plots with tests*, then click *Continue*.

Answers and Result Outputs

a)

H0: The variable *AI knowledge* is normally distributed.

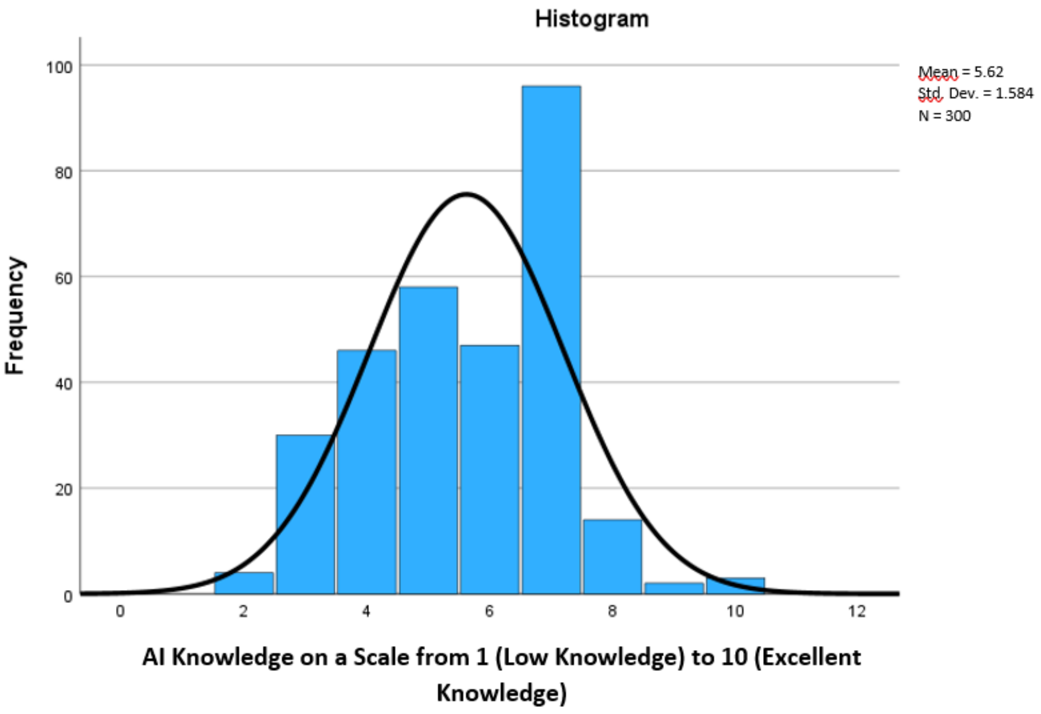
H1: The variable *AI knowledge* is not normally distributed.

b)

Table 13: Kolmogorov-Smirnov Test and Shapiro-Wilk W Test (Tests of Normality)

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
AI Knowledge on a Scale From 1 (Low Knowledge) to 10 (Excellent Knowledge)	.191	300	**< .001**	.931	300	**< .001**

[a] Lilliefors Significance Correction.



Histogram 2: Frequency Histogram for the Variable AI Knowledge With a Curve of the Fitted Normal Distribution

The result of both tests leads to the conclusion that the variable *AI knowledge* is not distributed according to a normal distribution. This is inferred because the significance level is less than 0.05, meaning that we reject the null hypothesis (H0) – we cannot confirm that the variable is normally distributed (Table 13). Deviations from the shape of a normal distribution, as observed in the frequency histogram, are therefore statistically significant (Histogram 2).

Task 2

A company has introduced a new product to the market. It wants to determine how many new products were sold in the first week, based on a random sample of 30 different stores. Sales data for the new product in the first week are in the file *Normal Distribution_Product Sales.sav*.

a) Check if the variable in the statistical set is distributed according to a normal distribution and draw a frequency histogram.

Answers and Result Outputs

a)

H0: The variable *product sales* is normally distributed.

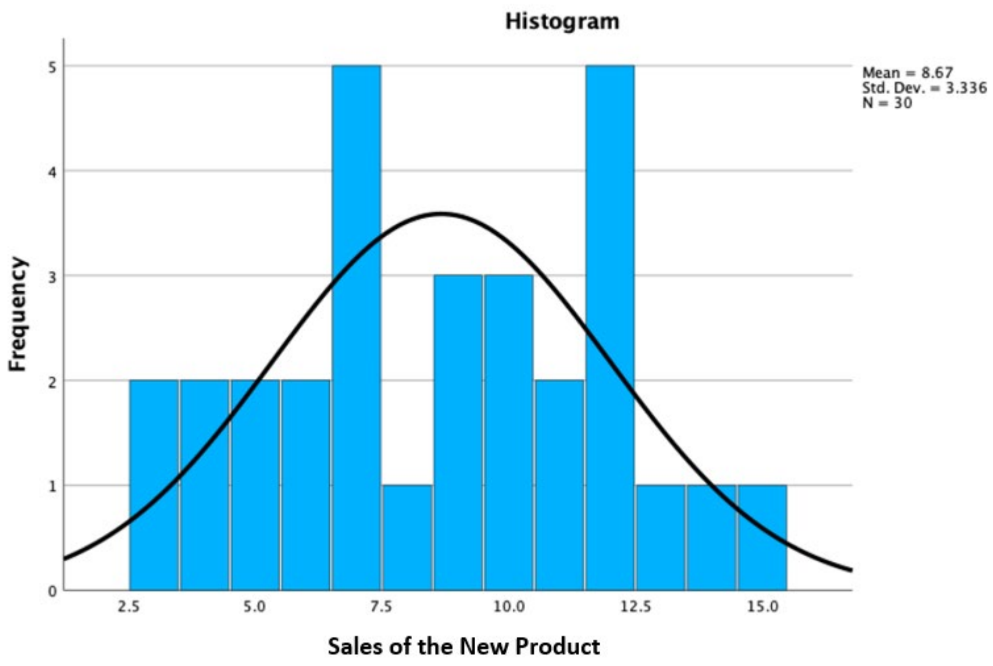
H1: The variable *product sales* is not normally distributed.

Table 14: Kolmogorov-Smirnov Test and Shapiro-Wilk W Test (Tests of Normality)

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Sales of the New Product	.125	30	**.200***	.964	30	**.384**

* This is a lower bound of the true significance.
[a] Lilliefors Significance Correction.

The Kolmogorov-Smirnov test and Shapiro-Wilk W test (Table 14) show that we do not reject the null hypothesis presupposing the variable in the statistical set is distributed according to a normal distribution. The significance level for the Kolmogorov-Smirnov test is 0.200, and for the Shapiro-Wilk test, it is 0.384. Both values are greater than 0.05 (Kolmogorov-Smirnov 0.200 and Shapiro-Wilk 0.384).



Histogram 3: Frequency Histogram for the Variable Product Sales

Task 3

Employee productivity is a fundamental indicator of operational efficiency and overall business performance. In this study, a sample of 73 employees was selected, comprising two groups from different departments (IT and Sales) within a company operating in the field of digital solutions and business services. Each participant reported the number of tasks they completed during a workweek. The variable *employee productivity* is expressed as the weekly number of completed tasks. This task aims to determine whether variable *employee productivity* is normally distributed.

The data is in the file *Normal Distribution_Employee Productivity.sav.*

Answers and Result Outputs:

H0: The variable *employee productivity* is normally distributed.

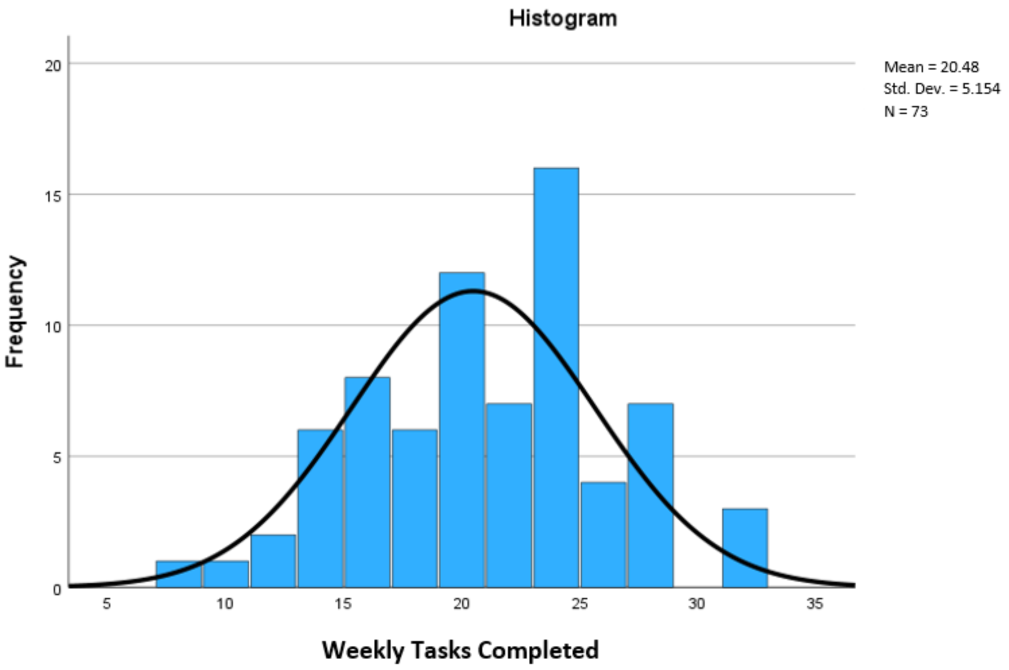
H1: The variable *employee productivity* is not normally distributed.

Kolmogorov-Smirnov test and Shapiro-Wilk W test (Table 15) show that we accept H0: The variable *employee productivity* is normally distributed, because both values are greater than 0.05 (Kolmogorov-Smirnov 0.076 and Shapiro-Wilk 0.382).

Table 15: Kolmogorov-Smirnov Test and Shapiro-Wilk W Test (Tests of Normality)

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Weekly Tasks Completed	.099	73	**.076**	.982	73	**.382**

[a] Lilliefors significance correction.



Histogram 4: Frequency Histogram for the Variable Employee Productivity

University of Maribor Press

3 Parametric and Nonparametric Univariate Statistical Tests

Parametric tests for detecting statistically significant differences between the mean values of variables in samples are used when it is permissible to fit the data for the dependent variable on an interval or ratio scale to a normal distribution (Corder and Foreman, 2014). When the units of samples belong to the same statistical set, we are dealing with dependent samples. When the units of the sample belong to different statistical sets, we are dealing with independent samples. For analysing significant differences between the mean values of a variable in two dependent samples, we will use the t-test for dependent samples, and between the mean values of a variable in two independent samples, we will use the t-test for independent samples. For analysing significant differences between the mean values in more than two independent samples, we will use analysis of variance (ANOVA).

Nonparametric equivalents to parametric tests are used when it is not possible to fit the data for the dependent numerical variable, which is based on an interval or ratio scale, to a normal distribution, or when the data for the dependent variable is based on an ordinal scale (Corder and Foreman, 2014). To compare two independent samples, we will use the Mann-Whitney U test, and to compare two dependent samples, we will use the Wilcoxon signed-rank test.

In the case that the two studied variables are nominal, we can use the χ^2-test for testing the association between the two.

3.1 Parametric Test for Dependent Samples: T-Test for Dependent Samples

Task 1

Company X introduced automation technologies to improve product quality and reduce the number of defective units produced. In the first monitoring period (before automation), quality control teams assessed 300 randomly selected products 35 times and recorded the number of defective items in each sample. After automation measures were implemented, the process was repeated under the same conditions. The dataset includes the number of defective products in each of the 35 subgroups for both the pre-automation and post-automation periods. We aim to determine whether automation has significantly reduced the number of defects.

For this purpose, we will check the following hypotheses:

H0: There is no statistically significant difference in the average number of defective products before and after automation (or H0: $\bar{y}_1 = \bar{y}_2$).

H1: There is a statistically significant difference in the average number of defective products before and after automation (or H1: $\bar{y}_1 \neq \bar{y}_2$).

The data is in the file *t-test for paired samples_Quality control.sav.*

Using the Kolmogorov-Smirnov and Shapiro-Wilk tests, we first determined that it is permissible to fit the data to a normal distribution (the procedure is described in chapter 2). Therefore, to test the stated hypotheses, we will use the parametric test for dependent samples, namely the t-test for dependent samples.

Procedure: T-Test for Paired Samples

In the *Analyze*, select *Compare Means*, then *Paired-Samples T Test*. Mark the variable *number of defective products before AI was used* and transfer it to the *Paired Variables* box by clicking on the arrow button, specifically to the cell under *Variable 1*. Transfer the variable *number of defective products before AI was used* to the cell under *Variable 2*. Click *OK* to perform the t-test for two dependent samples.

Answers and Result Outputs

Table 16: Normality Test Results Before and After Automation in Company X

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Number of Defective Products Before AI Was Used	.146	35	**.057**	.940	35	**.056**
Number of Defective Products After AI Was Used	.106	35	**.200***	.947	35	**.090**

* This is a lower bound of the true significance.

[a] Lilliefors significance correction.

From Table 16, it is evident that for both variables, $p > 0.05$, therefore, we do not reject the null hypotheses of the Kolmogorov-Smirnov test and the Shapiro-Wilk test. The number of defective products in the first period can be fitted to a normal distribution. Similarly, the normal distribution can be fitted to the number of defective products in the second period.

Table 17: Average Number of Defective Products in the First and Second Period

Paired Samples Statistics					
		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Number of Defective Products Before AI Was Used	15.74	35	5.570	.941
	Number of Defective Products After AI Was Used	12.49	35	5.559	.940

Table 17 presents the descriptive statistics for the paired samples. The average number of defective products before the implementation of automation in Company X was 15.74 (SD = 5.570), while after the implementation, it decreased to 12.49 (SD = 5.559). This suggests a potential improvement in product quality following the adoption of automation processes. A paired samples t-test will be conducted to determine whether this observed difference is statistically significant.

The two-tailed significance level in Table 18 is $p < 0.05$; therefore, we reject the null hypothesis. The average number of defective products before the implementation of automation is not equal to the average number of defective products after the implementation of automation in Company X (Table 18). Considering the results presented in Table 18, we reject the H0 and conclude that there is a statistically significant difference in the average number of defective products before and after automation (i.e., H1: $\bar{y}_1 \neq \bar{y}_2$).

Table 18: Results of the Paired-Samples T-Test in Company X

Paired Samples Test										
		Paired Differences					t	df	Significance	
		Mean	Std. Deviation	Std. Error Mean	95 % Confidence Interval of the Difference				One-Sided p	Two-Sided p
					Lower	Upper				
Pair 1	Number of Defective Products Before AI Was Used – Number of Defective Products After AI Was Used	3.257	3.147	.532	2.176	4.338	6.123	34	< .001	**< .001**

Task 2

In a logistics company, a new AI-supported order management system was introduced to replace the old software. The goal was to improve operational efficiency by reducing the average time required to process customer orders. To assess the impact of the new system, the company measured the average order processing time (in minutes) before and after the software implementation. A random sample of 90 orders was selected for each period (before and after the implementation), and the processing times were recorded.

We aim to determine whether there is a statistically significant difference in the average processing time before and after the introduction of the AI-supported system.

For this purpose, we will check the following hypotheses:

H0: There is no statistically significant difference in the average order processing time before and after the implementation of the AI-supported system (or H0: $\bar{y}_1 = \bar{y}_2$).

H1: There is a statistically significant difference in the average order processing time before and after the implementation of the AI-supported system (or H1: $\bar{y}_1 \neq \bar{y}_2$).

The data is in the file *t-test for paired samples_ TimeAI.sav.*

Table 19: Normality Test Results Before and After the Implementation of AI-Supported System

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Processing Time After AI (In Min)	.064	90	**.200***	.973	90	**.055**
Processing Time Before AI (In Min)	.057	90	**.200***	.982	90	**.233**

* This is a lower bound of the true significance.

[a] Lilliefors significance correction.

Table 19 presents the results of the Kolmogorov-Smirnov and Shapiro-Wilk tests for the two variables, *processing time before AI* and *processing time after AI*. Since the significance values for both tests are greater than 0.05 ($p > 0.05$), we do not reject the null hypothesis of normality. Therefore, both variables can be considered normally distributed.

Table 20: Average Processing Time Before and After the Implementation of AI

Paired Samples Statistics					
		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Processing Time Before AI (In Min)	24.02	90	1.059	.112
	Processing Time After AI (In Min)	20.95	90	.514	.054

Table 20 displays the descriptive statistics for the average processing time of customer orders before and after the implementation of an AI-supported system in a logistics company. The mean processing time before the introduction of AI was 24.02 minutes. After the implementation of the AI-based system, the mean processing time decreased to 20.95 minutes, accompanied by a lower standard deviation of 0.514 minutes. This suggests that the process became not only faster but also more consistent. The results indicate a potential improvement in operational efficiency following the adoption of AI technologies. However, a more detailed paired-samples t-test is required to confirm whether this difference is statistically significant.

Table 21 shows that the two-tailed significance level is $p < 0.05$; therefore, we reject the null hypothesis. The average processing time before the implementation of the AI-supported system is not equal to the average processing time after its implementation in the logistics company. Considering the direction of the change, the results suggest that the implementation of AI significantly reduced the average processing time of orders. Based on the results in Table 20, we reject H0 and conclude that there is a statistically significant difference in the average order processing time before and after the implementation of the AI-supported system (i.e., H1: $\bar{y}_1 \neq \bar{y}_2$).

Table 21: Results of the Paired-Samples T-Test in the Logistics Company

Paired Samples Test										
		Paired Differences					t	df	Significance	
		Mean	Std. Deviation	Std. Error Mean	95 % Confidence Interval of the Difference				One-Sided p	Two-Sided p
					Lower	Upper				
Pair 1	Processing Time Before AI (In Min) - Processing Time After AI (In Min)	3.063	.883	.093	2.878	3.248	32.903	89	< .001	**< .001**

3.2 Parametric Test for Independent Samples: Independent Samples T-Test

Task 1

Employee productivity is a key metric in business performance evaluation. This study investigates differences in the weekly productivity of employees from two departments: IT (coded as 1) and Sales (coded as 2). A sample of 73 employees reported the number of tasks they complete during an average week. The primary aim of this task is to analyse whether there are statistically significant differences in employee productivity between the two departments in the company.

As established in Task 3, subsection 2.2. The Kolmogorov-Smirnov test and Shapiro-Wilk W test, the variable *employee productivity* is normally distributed (Kolmogorov-Smirnov test is 0.076 and Shapiro-Wilk test is 0.382). Therefore, the conditions for conducting a parametric test for independent samples are met.

The data is in the file *Normal Distribution_Employee Productivity.sav.*

For this purpose, the following hypotheses will be tested:

H0: There is no statistically significant difference in average employee productivity between the IT and Sales departments (or H0: $\bar{y}1 = \bar{y}2$).

H1: There is a statistically significant difference in average employee productivity between the IT and Sales departments (or H1: $\bar{y}1 \neq \bar{y}2$).

Procedure: Parametric Test for Independent Samples

Navigate to *Analyze*, select *Compare Means*, and then *Independent-Samples T Test*. Transfer the dependent variable, *weekly tasks completed*, to the right window *Test Variable(s)*. Then, transfer the variable indicating the department label to the window *Grouping Variable*. Select *Define Groups* and specify group values: enter *1* for Group 1 (IT) and *2* for Group 2 (Sales), then click *Continue*. Click *OK* to perform the t-test for two independent samples.

Answers and Result Outputs

Table 22: Average Number of Weekly Tasks Completed by Department (IT vs. Sales)

	Department	N	Mean	Std. Deviation	Std. Error Mean
Weekly Tasks Completed	IT	38	20.08	4.576	.742
	Sales	35	20.91	5,.52	.972

Table 22 illustrates the descriptive statistics for weekly tasks completed by employees in the IT and Sales departments. While the mean productivity appears slightly higher in the Sales department (mean = 20.91) than in the IT department (mean = 20.08), an independent samples t-test will be performed to assess whether this difference is statistically significant.

In Table 23, we first use Levene's test for equality of variances to check whether we can assume equal variances in both independent samples.

H0: Equal variances assumed.

H1: Equal variances not assumed.

From the output of Levene's test for equality of variances, we observe that p = 0.127, which is greater than 0.05; therefore, we do not reject the null hypothesis and assume equal variances.

Table 23: Results of the Independent Samples T-Test in the IT and Sales Departments

Independent Samples Test											
		Levene's Test for Equality of Variances		T-Test for Equality of Means							
		F	Sig.	t	df	Significance		Mean Difference	Std. Error Difference	95 % Confidence Interval of the Difference	
						One-Sided p	Two-Sided p			Lower	Upper
Weekly Tasks Completed	Equal Variances Assumed	2.382	**.127**	−.689	71	.246	**.493**	−.835	1.212	−3.252	1.581
	Equal Variances Not Assumed			−.683	64.923	.249	.497	−.835	1.223	−3.278	1.608

In the output of the independent samples t-test for comparing the mean values of independent samples, we consider the first row of the results. It shows that p = 0.493, which is greater than 0.05. Hence, we do not reject the null hypothesis of equal mean values. This means that the average number of weekly tasks completed by employees in the IT department is not statistically different from that of employees in the Sales department. Consequently, we conclude that there is no statistically significant difference in employee productivity between the two departments.

Task 2

In Company X, they want to compare the quality of their products with the quality of products from a competing Company Y. In both companies, quality control assesses whether a product meets the required quality standards. During the same time period as in Company X, Company Y checked 30 batches of 300 products each for quality. In the dataset *t-test for independent samples_Quality control.sav*, the first column contains data on the number of defective products, and the second column contains data on the Company X or Y. The goal is to determine whether the quality of products in Company X significantly differs from the quality of products in the competing Company Y.

For this purpose, the following hypotheses will be tested:

H0: The average number of defective products in Company X is equal to the average number of defective products in Company Y (H0: $\bar{y}1 = \bar{y}2$).

H1: The average number of defective products in Company X significantly differs from the average number of defective products in Company Y (H1: $\bar{y}1 \neq \bar{y}2$).

Procedure

After confirming with the Kolmogorov-Smirnov and Shapiro-Wilk tests that the data can be adjusted to a normal distribution (Table 23), a parametric test for independent samples, specifically the independent samples t-test, will be used to verify the stated hypotheses.

Navigate to *Analyze*, select *Compare Means*, and then *Independent-Samples T Test*. Transfer the dependent variable, *number of defective products*, to the *Test Variable(s)* window on the right. Then, transfer the variable indicating the company label to the *Grouping Variable* window. Click *Define Groups* and specify group values: enter *1* for Group 1 (Company X) and *2* for Group 2 (Company Y). Click *Continue*, then *OK* to perform the t-test for two independent samples.

Answers and Result Outputs

Table 24: Results of Tests to Check Whether the Data on the Number of Defective Products Can Be Adjusted to a Normal Distribution.

	Company	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Number of Defective Products	X	0.115	25	0.200*	0.958	25	0.371
	Y	0.097	30	0.200*	0.981	30	0.850

* This is a lower bound of the true significance.
[a] Lilliefors significance correction.

Table 25: Average Number of Defective Products in Company X and Company Y

	Company	N	Mean	Std. Deviation	Std. Error Mean
Number of Defective Products	X	25	11.80	5.172	1.034
	Y	30	11.43	4.974	0.908

Table 26: Results of the Independent Samples T-Test for Companies X and Y

		Levene's Test for Equality of Variances		T-Test for Equality of Means						
		F	Sig.	t	df	Sig. (2-Tailed)	Mean Difference	Std. Error Difference	95 % Confidence Interval of the Difference	
									Lower	Upper
Number of Defective Products	Equal Variances Assumed	0.000	**0.991**	0.267	53	**0.790**	0.367	1.371	−2.384	3.117
	Equal Variances Not Assumed			0.266	50.448	0.791	0,.67	1.376	−2.397	3.131

In Table 26, we first use Levene's test for equality of variances to check whether we can assume equal variances in both independent samples.

H0: Equal variances assumed.

H1: Equal variances not assumed.

From the output of Levene's test for equality of variances, we observe that p = 0.991, which is greater than 0.05. Therefore, we do not reject the null hypothesis and assume equal variances.

In the output of the independent samples t-test for comparing the mean values of independent samples, we consider the first row of results. It shows that p = 0.790, which is greater than 0.05. Hence, we do not reject the null hypothesis of equal mean values: "The average number of defective products in Company X is statistically not different from the average number of defective products in Company Y." Therefore, the quality in Company X does not statistically differ from the quality in its competing Company Y.

Task 3

A company operating in the field of digital services provides customer support through two independent support teams: Team Alpha (coded as 1) and Team Beta (coded as 2). In order to evaluate the effectiveness of these two teams, the company collected data on the number of resolved customer support requests per day over a 30-day period for each team. The objective of the task is to determine whether there is a statistically significant difference in the average number of completed requests per day between Team Alpha and Team Beta.

The data is in the file *t-test for independent samples_SupportTeams.sav.*

For this purpose, the following hypotheses will be tested:

H0: The average number of completed requests is equal between the two teams (H0: ȳ1 = ȳ2).

H1: The average number of completed requests is significantly different between the two teams (H1: ȳ1 ≠ ȳ2).

Table 27: Results of the Independent Samples T-Test: Teams Alpha and Beta

Independent Samples Test											
		Levene's Test for Equality of Variances		T-Test for Equality of Means							
		F	Sig.	t	df	Significance		Mean Difference	Std. Error Difference	95 % Confidence Interval of the Difference	
						One-Sided p	Two-Sided p			Lower	Upper
Number of Completed Customer Support Requests	Equal Variances Assumed	4.385	**.041**	−1.638	58	.053	.107	−1.640	1.001	−3.645	.364
	Equal Variances Not Assumed			−1.757	56.18	.042	**.084**	−1.640	.933	−3.510	.229

In Table 27, we first use Levene's test for equality of variances to check whether we can assume equal variances in both independent samples.

H0: Equal variances assumed.

H1: Equal variances not assumed.

From the output of Levene's test for equality of variances, we observe that p = 0.041, which is less than 0.05. Therefore, we reject the null hypothesis and conclude that equal variances are not assumed.

In the output of the independent samples t-test for comparing the mean values of independent samples, we consider the second row of results. It shows that p = 0.084, which is greater than 0.05. Therefore, we do not reject the null hypothesis of equal mean values and accept H0: "The average number of completed requests is equal between the two teams (H0: ȳ1 = ȳ2)."

Task 4

A company has launched a new product not only in the domestic market A but also in the foreign market B. In the file t-*test for independent samples Product sales.sav*, there is data on the first-week sales from 30 randomly selected stores in the domestic market A and in other 30 randomly selected stores in the foreign market B. The company wants to determine whether the average sales of the new product in the domestic market A significantly differ from the average sales in the foreign market B.

a) Determine if it is permissible to adjust the sales data of the new product to a normal distribution.

b) Justify the use of the independent samples t-test to examine whether the average sales of the new product in the domestic market A significantly differ from the average sales in the foreign market B.

c) Write the null and research hypotheses.

d) Determine if the average sales of the new product in the domestic market A significantly differ from the average sales in the foreign market B.

Answers and Presentation of Results

a)

Table 28: Results of Tests for Checking Whether Sales Data Can Be Adjusted to a Normal Distribution.

	Market	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Sales of the New Product	Market A	0.125	30	0.200*	0.964	30	0.384
	Market B	0.109	30	0.200*	0.939	30	0.085

* This is a lower bound of the true significance.

[a] Lilliefors significance correction.

b)

Stores belong to different statistical populations: stores from the domestic market A and stores from the foreign market B. Sales data are numerical, and it is permissible to adjust them to a normal distribution.

c)

H0: The average sales of the new product in the domestic market A are equal to the average sales of the new product in the foreign market B (or H0: ȳ1 = ȳ2).

H1: The average sales of the new product in the domestic market A significantly differ from the average sales of the new product in the foreign market B (or H1: ȳ1 ≠ ȳ2).
d)

Table 29: *Descriptive Statistics of Independent Samples: Domestic Market A and Foreign Market B*

	Market	N	Mean	Std. Deviation	Std. Error Mean
Sales of the New Product	Market A	30	8.67	3.336	0.609
	Market A	30	10.40	4.116	0.751

Table 30: Results of the Independent Samples T-Test: Domestic Market A and Foreign Market B

		Levene's Test for Equality of Variances		T-Test for Equality of Means						
		F	Sig.	t	df	Sig. (2-Tailed)	Mean Difference	Std. Error Difference	95 % Confidence Interval of the Difference	
									Lower	Upper
Sales of the New Product	Equal Variances Assumed	2.195	**0.144**	−1.792	58	**0.078**	−1.733	0.967	−3.669	0.203
	Equal Variances Not Assumed			−1.792	55.615	0.079	−1.733	0.967	−3.671	0,.05

From the output of the Levene's test for equality of variances in Table 30, we observe that p > 0.05 (significance level is 0.144); therefore, we assume equal variances. In the results of the t-test for comparing the mean values of independent samples, we consider the first row of the output. We read that p > 0.05, meaning that we do not reject the null hypothesis of equal mean values: "The average sales of the new product in the domestic market A are the same as the average sales of the new product in the foreign market B (H0: ȳ1 = ȳ2)." We conclude that the average sales of the new product in the domestic market A do not statistically significantly differ from the average sales of the new product in the foreign market B.

3.3 Nonparametric Test for Paired Samples: Wilcoxon Signed Ranks Test

Task 1

Time-management skills are crucial for improving productivity and reducing stress among employees. In a company, a time-management training was conducted for 150 employees. The objective was to determine whether the training helped reduce the number of hours spent on non-priority tasks per week. Each participant reported the number of hours spent on non-priority tasks before and after attending the training. Although the variables are numerical, they do not follow a normal distribution.

The data is in the file *Wilcoxon Signed Rank_Effectiveness of Time Management Training.sav.*

The Wilcoxon signed-ranks test is used to test the following hypotheses:

H0: There is no statistically significant difference in the number of hours spent on non-priority tasks before and after the training among employees (or H0: ȳ1 = ȳ2).

H1: There is a statistically significant difference in the number of hours spent on non-priority tasks before and after the training among employees (or H1: ȳ1 ≠ ȳ2).

Procedure: Nonparametric Test for Paired Samples

After confirming with the Kolmogorov-Smirnov and Shapiro-Wilk tests that the data cannot be adjusted to a normal distribution (Table 31), a nonparametric test for paired samples will be used to verify the set hypotheses.

Select *Analyze*, then choose *Nonparametric Tests*, followed by *Legacy Dialogs* and *2 Related Samples*. Move the variable *number of hours spent on non-priority tasks before attending the training (hours_before)* to *Variable 1* and the variable *number of hours spent on non-priority tasks after attending the training (hours_after)* to *Variable 2* to *create Pair 1*. In the *Test Type* section, mark *Wilcoxon*. Click *OK* to perform the Wilcoxon signed-rank test for two dependent samples, the results of which are displayed in Tables 32 and 33.

Answers and Result Outputs

Table 31: Results of Tests for Checking Whether Tasks Before and After Attending the Training Can Be Adjusted to a Normal Distribution

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Hours_Before	.123	150	**< .001**	.950	150	**< .001**
Hours_After	.195	150	**< .001**	.939	150	**< .001**

[a] Lilliefors significance correction.

Table 32: Ranks in the Wilcoxon Signed-Rank Test Execution

Ranks				
		N	Mean Rank	Sum of Ranks
Hours_After – Hours_Before	Negative Ranks	150[a]	75.50	11,325.00
	Positive Ranks	0[b]	.00	.00
	Ties	0[c]		
	Total	150		

[a] Hours_after < hours_before.
[b] Hours_after > hours_before.
[c] Hours_after = hours_before.

Table 32 presents the ranking results from the Wilcoxon signed-rank test used to evaluate changes in the number of hours spent on non-priority tasks before and after the time-management training. The results show that all 150 employees (N = 150) reported fewer hours spent on non-priority tasks after the training compared to before, which is reflected in the 150 negative ranks (hours_after < hours_before). There were no positive ranks (hours_after > hours_before) and no ties (hours_after = hours_before), indicating a consistent reduction in non-priority work across all participants. This uniform direction of change suggests that the training was effective for the entire group.

Table 33: Results of the Wilcoxon Signed-Rank Test: Hours Spent on Non-Priority Tasks Before and After Time Management Training

Test Statistics[a]	
	Hours_After – Hours_Before
Z	**−10.873[b]**
Asymp. Sig. (2-Tailed)	**< .001**

[a] Wilcoxon signed ranks test.

[b] Based on positive ranks.

Based on the results presented in Table 33, we observe that $|Z| > 1.96$ and $p < 0.05$. We reject the null hypothesis (H0: $\bar{y}_1 = \bar{y}_2$), and conclude that there is a statistically significant difference in the number of hours spent on non-priority tasks before and after the training among employees (H1: $\bar{y}_1 \neq \bar{y}_2$).

Task 2

In an organization, 80 employees completed a questionnaire before and after using AI chatbots in their daily work. They rated their perceived stress levels on a 5-point ordinal scale (1 – no stress, 5 – very high stress). The objective is to determine whether there is a statistically significant difference in the perceived stress levels before and after the use of AI chatbots among employees in an organization.

The data is in the file *Wilcoxon Signed Rank_Before and after the use of AI chatbots.sav.*

We will test the following null hypothesis:

H0: There is no statistically significant difference in perceived stress levels before and after using AI chatbots among employees (H0: $\bar{y}_1 = \bar{y}_2$).

We have also formulated the following research hypothesis:

H1: There is a statistically significant difference in perceived stress levels before and after using AI chatbots among employees (H1: $\bar{y}_1 \neq \bar{y}_2$).

Answers and Result Outputs

Table 34: Ranks in the Wilcoxon Signed-Rank Test for Perceived Stress Levels Before and After Chatbot Use

Ranks				
		N	Mean Rank	Sum of Ranks
After Chatbot – Before Chatbot	Negative Ranks	50[a]	25.50	1,275.00
	Positive Ranks	0[b]	.00	.00
	Ties	30[c]		
	Total	80		

[a] After chatbot < before chatbot.
[b] After chatbot > before chatbot.
[c] After chatbot = before chatbot.

Table 34 presents the results of the Wilcoxon signed-rank test, which was used to assess whether there is a statistically significant difference in perceived stress levels before and after the use of AI chatbots in daily work. The results show that out of 80 employees, 50 reported lower stress levels after using the chatbot compared to before (negative ranks), none reported higher stress levels (positive ranks), and 30 rated their stress levels the same before and after (ties). The consistent direction of change (with no positive ranks and a substantial number of negative ranks) indicates that most employees experienced reduced stress levels after implementing AI chatbot support.

Table 35: Results of the Wilcoxon Signed-Rank Test for Changes in Employee Stress Levels Following AI Chatbot Implementation

Test Statistics[a]	
	After Chatbot – Before Chatbot
Z	**−7.071[b]**
Asymp. Sig. (2-Tailed)	**< .001**

[a] Wilcoxon signed ranks test.

[b] Based on positive ranks.

[b] Based on positive ranks.

From the outputs in Table 35, it is evident that $|Z| > 1.96$ and $p < 0.05$. We reject the null hypothesis (H0: $\bar{y}_1 = \bar{y}$), and conclude that there is a statistically significant difference in perceived stress levels before and after using AI chatbots among employees (H1: $\bar{y}_1 \neq \bar{y}_2$).

Task 3

Company X operating in a specific market recorded the value of services performed in two consecutive years (Year 1 and Year 2). Data were collected for 210 selected business units operating in the same market and under comparable conditions. The goal is to determine whether there is a statistically significant difference in the average value of services performed between the first and the second observed year.

In the file *Wilcoxon Signed Rank Value of services.sav*, the following variables are included (1) *Services_year1*: Value of performed services in the first observed year (in EUR), and (2) *Services_year2:* Value of performed services in the second observed year (in EUR).

We want to check if the average rank of the value of performed services in the first observed year statistically significantly differs from the average rank of the value of performed services in the second observed year.

a) Write the null and research hypotheses.

b) Test the null hypothesis and write the outcome of the test. Take into account that ranks have been assigned to the variable values.

Answers and Result Outputs

a)

H0: There is no statistically significant difference in the average value of services performed between the first and the second year (H0: $\bar{y}_1 = \bar{y}_2$).

H1: There is a statistically significant difference in the average value of services performed between the first and the second year (H1: $\bar{y}_1 \neq \bar{y}_2$).

b)

Table 36: Ranks for the Value of Services for Two Dependent Samples

Ranks				
		N	Mean Rank	Sum of Ranks
Service Values at the First Market Year 2 – Service Values at the First Market Year 1	Negative Ranks	117[a]	101.08	11,826.00
	Positive Ranks	93[b]	111.06	10,329.00
	Ties	0[c]		
	Total	210		

[a] Service values at the first market year 2 < service values at the first market year 1.
[b] Service values at the first market year 2 > service values at the first market year 1.
[c] Service values at the first market year 2 = service values at the first market year 1.

Table 36 shows the results of the Wilcoxon signed-rank test applied to compare the values of performed services in two consecutive years for the same market. Among 210 observations, 117 cases had negative ranks, indicating that the value of services in the second year was lower than in the first year. On the other hand, 93 cases had positive ranks, meaning that the value of services increased in the second year compared to the first. Although the number of negative ranks slightly exceeds the number of positive ranks, the presence of both directions suggests variability in service performance across the market. The test results (presented in the following table with Z and significance values) will confirm whether this difference is statistically significant.

Table 37: Results of the Wilcoxon Signed-Rank Test for the Value of Services

Test Statistics[a]	
	Service Values at the First Market Year 2 – Service Values at the First Market Year 1
Z	−.849[b]
Asymp. Sig. (2-Tailed)	.396

[a] Wilcoxon signed-ranks test.
[b] Based on positive ranks.

From the outputs in Table 37, it is evident that $|Z| < 1.96$ and $p > 0.05$. We accept the null hypothesis: “H0: There is no statistically significant difference in the average value of services performed between the first and the second year (H0: $\bar{y}_1 = \bar{y}_2$).”

3.4 Nonparametric Test for Independent Samples: Mann-Whitney U Test

Task 1

To examine potential differences in employee well-being related to the use of AI chatbot support, a company introduced an internal AI assistant to assist employees with administrative and technical tasks. To evaluate the perceived effect of the assistant, randomly selected employees rated their stress levels associated with daily work tasks on a 5-point ordinal scale (1 – very low stress, 5 – very high stress). Group 1 consists of employees who use the AI chatbot regularly, while Group 2 includes those who do not use the AI chatbot. The goal is to determine whether there is a statistically significant difference in perceived stress levels between the two groups.

The data is in the file *Mann-Whitney U Test_ AI chatbot.sav.*

For this purpose, we tested the following null hypothesis:

H0: There is no statistically significant difference in the perceived stress levels between employees who regularly use the AI chatbot and those who do not use it (H0: $\bar{y}_1 = \bar{y}_2$).

We also formulated an alternative hypothesis:

H1: There is a statistically significant difference in the perceived stress levels between employees who regularly use the AI chatbot and those who do not use it (H1: $\bar{y}_1 \neq \bar{y}_2$).

Procedure: Nonparametric Test for Independent Sample

In the *Analyze*, select *Nonparametric Tests*, *Legacy Dialogs*, and then *2 Independent Samples*. Move the dependent variable *stress level* to the *Test Variable List* on the right. Move the variable *Chatbot Use* to the *Grouping Variable* window. Click *Define Groups* and define the values for groups: enter *1* for Group 1 and *2* for Group 2, then select *Continue*. In the *Test Type* section, mark *Mann-Whitney U*. Click *OK* to perform the Mann-Whitney U test for two independent samples and obtain the results in Tables 38 and 39.

Answers and Result Outputs

Table 38: Ranks for Perceived Stress Levels by Chatbot Usage Group

Ranks				
	Chatbot Use	N	Mean Rank	Sum of Ranks
Stress Level	Use	20	10.88	217.50
	No Use	20	30.13	602.50
	Total	40		

Table 38 presents the results of the Mann-Whitney U test comparing perceived stress levels between employees who use AI chatbots and those who do not. The group of employees who use the chatbot has a lower mean rank (10.88) compared to the group that does not use the chatbot (mean rank = 30.13). This substantial difference in mean ranks suggests that employees who regularly use the chatbot report significantly lower levels of stress related to daily work tasks than those who do not use it. The direction of the ranks indicates that chatbot support could enhance employee well-being by reducing perceived stress levels. For final confirmation, statistical significance should be verified by examining the U-value and the corresponding p-value in the test statistics table (which follows the ranks table).

Table 39: Results of the Mann-Whitney U Test for Perceived Stress Levels by Chatbot Usage Group

Test Statistics[a]	
	Stress Level
Mann-Whitney U	7.500
Wilcoxon W	217.500
Z	**−5.323**
Asymp. Sig. (2-Tailed)	**< .001**
Exact Sig. [2*(1-Tailed Sig.)]	< .001[b]

[a] Grouping variable: chatbot use.
[b] Not corrected for ties.

From the result output in Table 39, it is evident that $|Z| > 1.96$, and $p < 0.05$. Based on the results, we reject the null hypothesis H0: $\bar{y}_1 = \bar{y}_2$ and conclude that there is a statistically significant difference in the perceived stress levels between employees who regularly use the AI chatbot and those who do not use it (H1: $\bar{y}_1 \neq \bar{y}_2$).

Task 2

In the study on the use of quantitative methods in companies, we examined the frequency of using statistical methods for sales forecasting (1 – never, 2 – twice a year, 3 – monthly, 4 – weekly, 5 – daily). Data on the frequency of using statistical methods for sales forecasting obtained from randomly selected 62 domestic and 137 foreign companies can be found in the file *Mann-Whitney U test_Frequency of use.sav*. We aim to determine whether the average ranks of statistical method usage for sales forecasting differ significantly between domestic and foreign companies. For this purpose, we tested the following null hypothesis:

H0: There is no statistically significant difference in the average ranks of the frequency of the statistical method usage for sales forecasting between domestic and foreign companies (H0: $\bar{y}_1 = \bar{y}_2$).

We have also formulated the hypothesis:

H1: There is a statistically significant difference in the average ranks of the frequency of the statistical method usage for sales forecasting between domestic and foreign companies (H1: $\bar{y}_1 \neq \bar{y}_2$).

Procedure

In the *Analyze*, select *Nonparametric Tests*, then *Legacy Dialogs*, and lastly, *2 Independent Samples*. Move the dependent variable *frequency of using statistical methods for sales forecasting* to the *Test Variable List* on the right. After, move the variable *company* to the *Grouping Variable* window. Click *Define Groups* and define the values for groups: enter *1* for Group 1 and *2* for Group 2. Click *Continue*. In the *Test Type* section, mark *Mann-Whitney U*. Click *OK* to perform the Mann-Whitney U test for two independent samples and obtain the results in Tables 37 and 38.

Answers and Result Outputs

Table 40: Ranks for the Frequency of Statistical Method Usage for Sales Forecasting for Two Independent Samples

	Company	N	Mean Rank	Sum of Ranks
The Frequency of Statistical Method Usage for Sales Forecasting	Domestic	62	86.83	5,383.50
	Foreign	137	105.96	14,516.50
	Total	199		

Table 40 shows that the average rank of the frequency of the statistical method usage for sales forecasting is higher in foreign companies than in domestic ones. The same applies to the sum of ranks.

Table 41: Results of the Mann-Whitney U Test for the Frequency of Use

	The Frequency of Statistical Method Usage for Sales Forecasting
Mann-Whitney U	3,430.500
Wilcoxon W	5,383.500
Z	**−2.231**
Asymp. Sig. (2-Tailed)	**0.026**

[a] Grouping variable: company.

Table 41 shows that $|Z| > 1.96$ and $p < 0.05$. Based on the results, we reject the null hypothesis H0: $\bar{y}_1 = \bar{y}_2$ and there indeed is a statistically significant difference in the average ranks of the frequency of the statistical method usage for sales forecasting between domestic and foreign companies (H1: $\bar{y}_1 \neq \bar{y}_2$).

Task 3

Artificial Intelligence (AI) is reshaping innovation across various industries, especially within the startup ecosystem. Investors are often attracted to startups with cutting-edge technology, including those implementing AI. The goal of this task is to determine whether there is a statistically significant difference in the average amount of investment received between AI and non-AI startups. In this analysis, 50 startups were included: 25 that use AI technologies (coded as Group 1) and 25 that do not use AI technologies (coded as Group 2). Each startup reported the amount of funding received (in thousands of EUR). The dependent variable *investment amount* (in EUR) is numerical, but not normally distributed.

The data is in the file *Mann-Whitney U Test_Investment_AI_Startups.sav.*

a) Write the appropriate hypotheses to test whether there are statistically significant differences in the average amount of investment received between AI and non-AI startups.

b) Test the null hypothesis stated in task a).

Answers and Result Outputs

a)

Hypotheses

H0: There is no statistically significant difference in the average amount of investment received between AI and non-AI startups (H0: $\bar{y}_1 = \bar{y}_2$).

H1: There is a statistically significant difference in the average amount of investment received between AI and non-AI startups (H1: $\bar{y}_1 \neq \bar{y}_2$).

b)

To test the null hypotheses stated in task a), we use the nonparametric test for independent samples, namely the Mann-Whitney U test. In this process, ranks have been assigned to the variable values.

In the *Analyze*, select *Nonparametric Tests*, *Legacy Dialogs*, and then *2 Independent Samples*. Move the dependent variable, *investment amount,* to the *Test Variable List* on the right. Afterward, move the variable *AI use* to the *Grouping Variable* window. Click *Define Groups* and define the values for groups: enter *1* for Group 1 and *2* for Group 2, and click *Continue*. In the *Test Type* section, mark *Mann-Whitney U*. Click *OK* to perform the Mann-Whitney U test for two independent samples and obtain the results in Tables 42 and 43.

Table 42: Results of Tests to Verify Whether the Data on the Amount of Investment Could Be Fitted to a Normal Distribution

Tests of Normality						
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Investment Amount (EUR)	.119	60	.034	.946	60	.010

[a] Lilliefors significance correction.

b)

Table 43: Display of Ranks: The Mann-Whitney U Test for the Amount of Investment

Ranks				
	AI Use	N	Mean Rank	Sum of Ranks
Investment Amount (EUR)	Startups That Use AI Technologies	30	45.35	1,360.50
	Startups That Do Not Use AI Technologies	30	15.65	469.50
	Total	60		

Table 44: Results of the Mann-Whitney U Test for the Amount of Investment

Test Statistics[a]	
	Investment Amount (EUR)
Mann-Whitney U	4.500
Wilcoxon W	469.500
Z	−6.589
Asymp. Sig. (2-Tailed)	< .001

[a] Grouping variable: AI use.

Table 44 shows that |Z| > 1.96, p < 0.01, therefore, we reject the null hypothesis and conclude that there is a statistically significant difference in the average amount of investment received between AI and non-AI startups (H1).

3.5 Nonparametric Test: χ2 Test for Analysis of Association (Relationship) Between Two Nominal Variables

Task 1

As artificial intelligence (AI) rapidly transforms various sectors, students are increasingly seeking ways to acquire AI-related knowledge. Understanding whether students from different fields of study (1 – Business and Economics, 2 – Computer Science and Engineering, 3 – Social Sciences and Humanities) prefer different learning methods (1 – online courses, 2 – formal university courses, 3 – industry-led workshops or hackathons) can help universities and educators tailor their AI literacy initiatives more effectively. In a survey, 250 randomly selected students were asked to indicate both their field of study and their preferred method of learning about AI. We aim to investigate whether there is a statistically significant relationship between a student's field of study and their preferred method of learning about artificial intelligence.

The data is in the file *Chi_square_AI Learning Preferences by Field of Study*.

We tested the following null hypothesis:

H0: There is no statistically significant association between a student's field of study and their preferred method of AI learning.

H1: There is a statistically significant association between a student's field of study and their preferred method of AI learning.

Procedure: χ2 Test

Select *Analyze*, *Descriptive Statistics* and *Crosstabs*. From the *Variables* window, move the *fields of study* variable to *Rows*, and the *learning methods* variable to *Columns* on the right side. Under *Statistics*, select *Chi-square*, and continue with clicking *Continue* and *OK*.

Answers and Result Outputs

Table 45: Results of the χ2 Test for Checking the Association Between the Variables, Students' Field of Study and their Preferred Method of AI Learning

Chi-Square Tests			
	Value	df	Asymptotic Significance (2-Sided)
Pearson Chi-Square	81.619[a]	4	**< .001**
Likelihood Ratio	77.920	4	< .001
Linear-by-Linear Association	57.606	1	< .001
N of Valid Cases	250		

[a] No cells (0.0 %) have an expected count less than 5. The minimum expected count is 15.55.

From Table 45, it is evident from the row *Pearson Chi-Square* that p < 0.05. Therefore, we reject the null hypothesis and conclude that (H1) there is a statistically significant association between a student's field of study and their preferred method of AI learning.

Task 2

In times of economic uncertainty, consumer behaviour often shifts, particularly in terms of spending and shopping habits. With the rise of e-commerce, it is important to understand how different demographic groups adapt their purchasing strategies. This knowledge can help businesses tailor their digital marketing strategies more effectively. In a survey, 200 randomly selected consumers were asked about their gender (1 – male, 2 – female) and how their online shopping habits changed during recent periods of economic uncertainty (online shopping behaviour: 1 – only essential items, 2 – essential items and some non-essential items, 3 – no change in shopping habits). We aim to investigate whether there is a statistically significant relationship between a consumer's gender and their preferred online shopping behaviour during economic uncertainty.

The data is in the file *Chi-Square_Gender_Online Shopping.sav.*

a) Write down the relevant hypotheses.

b) Check the null hypothesis with the appropriate test.

Answers and Result Outputs

a)

We tested the following null hypothesis:

H0: There is no statistically significant association between a consumer’s gender and their preferred online shopping behaviour during economic uncertainty.

We also defined the research hypothesis:

H1: There is a statistically significant relationship between a consumer’s gender and their preferred online shopping behaviour during economic uncertainty.

b)

Table 46: Results of the χ2 Test for Checking the Association Between the Variables Consumer’s Gender and Their Preferred Online Shopping Behaviour During Economic Uncertainty

Chi-Square Tests			
	Value	df	Asymptotic Significance (2-Sided)
Pearson Chi-Square	6.234[a]	2	**.044**
Likelihood Ratio	6.244	2	.044
Linear-by-Linear Association	5.991	1	.014
N of Valid Cases	200		

[a] No cells (0.0 %) have an expected count less than 5. The minimum expected count is 25.23.

From Table 46, it is evident from the row *Pearson Chi-Square* that $p < 0.05$. Therefore, we reject the null hypothesis and conclude that (H1) there is a statistically significant relationship between a consumer’s gender and their preferred online shopping behaviour during economic uncertainty.

Task 3

As global sustainability efforts gain momentum, understanding individuals’ environmental behaviours in various professional sectors is increasingly important. In this study, we aim to examine whether there is a statistically significant relationship between the occupation sector of individuals (1 – private sector, 2 – public sector) and their methods of plastic waste disposal (1 – recycle properly: separated and disposed of according to guidelines, 2 – throw in regular waste bin, 3 – irregular or mixed methods: sometimes recycle, sometimes not). In a survey, 360 randomly selected individuals were asked to indicate their occupation sector and how they typically dispose of plastic waste. The objective is to investigate whether individuals employed in the private sector dispose of plastic waste differently than those employed in the public sector.

The data is in the file *Chi-Square_Plastic disposal by sector.sav.*
a) Write down the relevant hypotheses.

b) Check the null hypothesis with the appropriate test.

Answers and Result Outputs

a)

Hypotheses:

H0: There is no statistically significant relationship between occupation sector and plastic waste disposal method.

H1: There is a statistically significant relationship between occupation sector and plastic waste disposal method.

b)

Table 47: Results of the χ2 Test for Checking the Association Between the Variables "Occupation Sector" and "Plastic Waste Disposal Method"

Chi-Square Tests			
	Value	df	Asymptotic Significance (2-Sided)
Pearson Chi-Square	.498[a]	2	**.780**
Likelihood Ratio	.497	2	.780
Linear-by-Linear Association	.427	1	.514
N of Valid Cases	460		

[a] No cells (0.0 %) have an expected count less than 5. The minimum expected count is 55.22.

Based on the results in Table 47, we accept the null hypothesis: "H0: There is no statistically significant relationship between occupation sector and plastic waste disposal method" (Pearson Chi-Square: $p > 0.05$).

4 Factor Analysis

Factor analysis is a multivariate statistical method used to reduce a large set of observed variables or items into a smaller number of latent dimensions, known as factors, which represent the shared variance among the items. The main objective of factor analysis is to uncover the underlying structure within a set of interrelated variables and to identify patterns that explain the observed correlations in a more parsimonious and interpretable way (Tavakol et al., 2020). Through the application of factor analysis, researchers can statistically identify common dimensions that influence responses to multiple items. This process simplifies complex data structures by grouping variables that measure similar underlying constructs, thereby reducing redundancy and enhancing interpretability. The analysis results in a factor solution in which each factor represents a cluster of items with high intercorrelations, suggesting that they tap into the same underlying concept (Kline, 2023).

Thus, factor analysis is used when it is possible to form a smaller number of mutually independent factors (variables) from a larger number of interconnected measured variables (when the conditions for performing factor analysis are met). Factors represent a linear combination of measured variables (Tabachnick and Fidell, 2013).

In applied research and business analytics, factor analysis plays a critical role in simplifying complex datasets, improving construct validity, and supporting theoretical model development. By examining the shared variance among variables, this method reveals hidden structures that may represent psychological traits, behavioural tendencies, organizational attributes, or market dynamics (Gorsuch, 2014).

Factor analysis is extensively used across various disciplines, including economics, finance, marketing, organizational behaviour and the social sciences. In business settings, it is frequently applied in the development and validation of measurement instruments (e.g., customer satisfaction scales and employee engagement surveys), segmentation analysis, and strategic decision-making. The resulting factor structure provides researchers and practitioners with an empirical basis for identifying dimensions of interest and for refining constructs used in further statistical modelling, such as regression or structural equation modelling (Fabrigar and Wegener, 2011; Kline, 2023)

Steps in Factor Analysis

- Analysis of correlation among the measured variables forming several dimensions of the multidimensional variable (this is a condition for using factor analysis).
- Determining the (smaller) number of new variables – factors, that describe the measured variables well (explaining their variance in a high %).
- Defining the meaning of new variables – factors.

Task 1

In order to explore which dimensions of business performance are deemed most important by executives, the sample consists of company owners. The objective is to explore whether the observed indicators related to performance can be reduced to a smaller number of underlying latent dimensions using factor analysis. This can help organizations identify key performance drivers more effectively. The sample includes 420 randomly selected owners of medium-sized and large companies. The objective is to perform the factor analysis on a set of items (variables) assessing the performance of companies *(variables BP1 to BP10).*

Based on the data in the file *Factor analysis_Company_AI.sav,* perform a factor analysis for the multidimensional variable *performance of the company* and explain the results.

Procedures: Factor Analysis

We click *Analyze* and choose *Dimension Reduction,* then *Factor.* To the right windows, we transfer those variables for which we want to carry out factor analysis. In our example, variables *BP1 to BP10* are transferred to the right window. With a click on the *Descriptives* button, a new dialog window appears in which, under *Statistics,* we choose *Univariate descriptives* and *Initial Solution.* In the *Correlation Matrix* window, we select *Coefficients*, *Significance levels,* as well as *KMO and Bartlett's Test of sphericity*, then click *Continue.*

We continue by clicking the *Extraction* button. Among the offered methods, we choose *Principal components.* In the *Analyze* window, we select *Correlation matrix* and in the *Extract* window, we type *1* at *Eigenvalues greater than.* With a click on *Continue,* we return to the initial window.

We click the *Rotation* button. Among the available methods, we chose *Varimax*, the most familiar method. This method ensures that each measured variable has a high factor loading on only one factor, while the loadings on other factors are minimized. Higher factor loadings enable easier interpretation. In the *Display* box, we select *Rotated Solution.* With a click on *Continue,* we return to the initial window. We select *Scores* and then *Save as variables.* With a click on *Continue,* we return to the initial window and select *OK.*

Answers and Result Outputs

Table 48: Correlation Matrix for Items Related to Business Performance

Correlation Matrix								
		1. Through AI, companies can obtain accurate results.	2. Through AI, the chance of employees' errors at work is reduced.	3. AI enhances the effectiveness of decisions and actions.	4. AI accelerates quick and better decision making to achieve successful results.	5. AI provides accurate data and information.	6. Products or services meet customers' expectations.	7. The delivery of goods or services is conducted in a timely fashion.
Correlation	1. Through AI, companies can obtain accurate results.	**1.000**	.789	.836	.804	.717	.750	.792
	2. Through AI, the chance of employees' errors at work is reduced.	.789	**1.000**	.885	.846	.846	.852	.810

Correlation Matrix								
	3. AI enhances the effectiveness of decisions and actions.	.836	.885	**1.000**	.886	.830	.851	.892
	4. AI accelerates quick and better decision making to achieve successful results.	.804	.846	.886	**1.000**	.855	.866	.887
	5. AI provides accurate data and information.	.717	.846	.830	.855	**1.000**	.929	.812
	6. Products or services meet customers' expectations.	.750	.852	.851	.866	.929	**1.000**	.831
	7.The delivery of goods or services is conducted in a timely fashion.	.792	.810	.892	.887	.812	.831	**1.000**
Sig. (1-Tailed)	1. Through AI, companies can obtain accurate results.		< .001	< .001	< .001	< .001	< .001	< .001
	2. Through AI, the chance of employees' errors at work is reduced.	.000		.000	.000	.000	.000	.000
	3. AI enhances the effectiveness of decisions and actions.	.000	.000		.000	.000	.000	.000
	4. AI accelerates quick and better decision making to achieve successful results.	.000	.000	.000		.000	.000	.000
	5. AI provides accurate data and information.	.000	.000	.000	.000		.000	.000
	6. Products or services meet customers' expectations.	.000	.000	.000	.000	.000		.000
	7. The delivery of goods or services is conducted in a timely fashion.	.000	.000	.000	.000	.000	.000	

Table 48 shows that all correlation coefficients between items are positive and range from 0.717 to 0.929, indicating moderate to strong linear relationships between the items. For instance, the strongest correlation is between Item 5 (AI provides accurate data and information.) and Item 6 (Products or services meet customers' expectations.), with a coefficient of 0.929. Other high correlations include Item 3 (AI enhances the effectiveness of decisions and actions.) and Item 7 (The delivery of goods or services is conducted in a timely fashion.) with 0.892, as well as Item 2 and Item 3 with 0.885. All p-values are below 0.001, indicating that these correlations are statistically significant at the 0.05 level. Such high and significant correlations suggest that the items are sufficiently interrelated, supporting the suitability for factor analysis.

Table 49: Kaiser-Meyer-Olkin Statistics and Bartlett's Test of Sphericity for the Multidimensional Variable Business Performance

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy		**.928**
Bartlett's Test of Sphericity	Approx. Chi-Square	4,174.879
	df	21
	Sig.	**< .001**

Bartlett's test of sphericity is used to check if conditions for using factor analysis are met. With it, the H0 that the correlation matrix is the unit matrix, is tested, which means that no relationship between observed (measured) variables exists. We tested the following hypotheses:

H0: The correlation coefficients between the observed variables are equal to 0 (i.e., the correlation matrix is an identity matrix).

H1: At least some correlation coefficients between the observed variables are significantly different from 0 (i.e., the correlation matrix differs from the identity matrix).

Table 50: Communalities for the Multidimensional Variable Business Performance

Communalities		
	Initial	Extraction
1. Through AI, companies can obtain accurate results.	1.000	.764
2. Through AI, the chance of employees' errors at work is reduced.	1.000	.863
3. AI enhances the effectiveness of decisions and actions.	1.000	.907
4. AI accelerates quick and better decision making to achieve successful results.	1.000	.897
5. AI provides accurate data and information.	1.000	.853
6. Products or services meet customers' expectations.	1.000	.878
7.The delivery of goods or services is conducted in a timely fashion.	1.000	.862

Besides Bartlett's test of sphericity, Kaiser-Meyer-Olkin statistics (KMO) is also used, indicating that the use of factor analysis is meaningful at values greater than 0.5 (Tabachnick and Fidell, 2013). In our case, KMO > 0.5 (KMO = 0.928) and Bartlett's test, $p < 0.05$; therefore, we accept H1: "At least some correlation coefficients between the observed variables are significantly different from 0, indicating that the use of factor analysis is justified (Table 49).

Extraction method: Principal component analysis.

Communalities explain the proportion of each measured variable that is explained by factors with eigenvalues greater than 1 – the proportion of variance of each measured variable, explained by factors, should be higher than 0.40 (Tabachnick and Fidell, 2013).

Communality presents the proportion of the total variance of a given variable that is explained by all extracted factors combined. For example, the variable *Through AI, companies can obtain accurate results* has a communality value of 0.764, indicating that 76.4 % of the variance in this item is explained by the extracted factor. All communalities in Table 50 for the multidimensional variable *business performance* are greater than 0.40; therefore, no variable was excluded from analysis.

Table 51: Complete Explanation of Variance for the Multidimensional Variable Business Performance

Total Variance Explained						
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	**6.022**	**86.035**	86.035	6.022	86.035	86.035
2	.346	4.939	90.973			
3	.207	2.956	93.930			
4	.166	2.365	96.295			
5	.109	1.552	97.847			
6	.082	1.177	99.024			
7	.068	.976	100.000			

Extraction method: Principal component analysis.

The sum of the squares of the factor loadings for the factor j for all k basic variables (items) is the eigenvalue of the factor j. If the eigenvalue > 1, the factor explains a substantial portion of the variance of all measured variables.

Table 50 shows that seven measured variables resulted in seven principal components, but only the first component has an eigenvalue greater than 1 (total: 6.022). This factor accounts for 86 % of the total variance of the measured variables.

Table 52: Factor Loadings for the Multidimensional Variable Business Performance

Component Matrix[a]	
	Component
	1
1. Through AI, companies can obtain accurate results.	.874
2. Through AI, the chance of employees' errors at work is reduced.	.929
3. AI enhances the effectiveness of decisions and actions.	**.952**
4. AI accelerates quick and better decision making to achieve successful results.	.947
5. AI provides accurate data and information.	.923
6. Products or services meet customers' expectations.	.937
7.The delivery of goods or services is conducted in a timely fashion.	.928

Extraction method: Principal component analysis.
[a] 1 components extracted.

Table 52 shows that all factor loadings, a_{ij}, are higher than 0.60. A square value of the factor loading at *i*-th variable and *j*-th factor denotes the share of explained variance of *i*-th variable with the *j*-th factor. Within the context of *business performance*, the most important role is held by the variable *AI enhances the effectiveness of decisions and actions*, at which the value of the factor loadings is the highest.

Meaning of Selected Factors

Once factors are identified through factor analysis, it is essential to interpret and assign meaningful names to them. When a large proportion of the variance of all measured variables can be explained by a single factor, the factor can be named after the multidimensional latent construct it represents. In such cases, the interpretation is relatively straightforward, as all variables contribute to a single underlying concept. In our case, we can name the obtained factor *business performance* (Table 52).

In situations where multiple factors are extracted, the naming process becomes more complex. Each factor must be interpreted based on the content of the variables that load most highly on it. To enhance the clarity and interpretability of the factor structure, factor rotation is commonly applied. The Varimax rotation method is frequently used, as it facilitates a simpler and more distinct allocation of variables to specific factors.

Once clearly defined, the extracted factors can be used in subsequent stages of the research process (for example, regression analysis, which is presented in the fifth section).

Task 2

Following the analysis of key dimensions related to company performance in Task 1, this part of the research focuses on another important aspect of modern business practices, namely the implementation of artificial intelligence technology in the work environment (IAIT).

The objective of Task 2 is to examine whether the observed variables related to the implementation of AI technology (variables *IAIT1* to *IAIT5*) can be reduced to a smaller number of factors. This statistical approach enables the identification of core areas through which AI is integrated into organizational processes and provides valuable insights for guiding digital transformation strategies.

The analysis is conducted using the same dataset as in Task 1. The sample includes 420 randomly selected owners of medium-sized and large companies. This consistency ensures the comparability of results and contributes to a more comprehensive understanding of business development patterns across different functional domains.

Based on the data in the file *Factor analysis_Company_AI.sav.*

a) Conduct a factor analysis for the multidimensional variable *implementation of AI technology in the work environment* and interpret the results.
b) Calculate Cronbach's alpha and describe the reliability of the scale.

Answers and Result Outputs

a) Factor analysis for the multidimensional variable *implementation of AI technology in the work environment* and interpret the results

Table 53: Kaiser-Meyer-Olkin Statistics and Bartlett's Test of Sphericity for Multidimensional Variable Implementation of AI Technology in the Work Environment

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy		**.838**
Bartlett's Test of Sphericity	Approx. Chi-Square	1,629.320
	df	10
	Sig.	**< .001**

In accordance with Table 53, KMO (0.838) and Bartlett's test, $p < 0.05$, we accept H1: “At least some correlation coefficients between the observed variables are significantly different from 0”, showing that the use of factor analysis is justified.

Table 54: Communalities for Multidimensional Variable Implementation of AI Technology in the Work Environment

Communalities		
	Initial	Extraction
1. Our company uses programme and portfolio structures for managing projects.	1.000	.788
2. Our company has a digital transformation strategy, including AI adoption.	1.000	.664
3. Our company uses AI technologies for work design.	1.000	.664
4. Our company uses AI technologies to better plan new tasks.	1.000	.840
5. Our company uses AI technologies in projects to create teams.	1.000	.747

Extraction method: Principal component analysis.

Communalities in Table 54 for the multidimensional variable *implementation of AI technology in the work environment* are higher than 0.40; therefore, no variable was excluded from analysis. For example, the first variable *Our company uses programme and portfolio structures for managing projects* has a communality value of 0.788, indicating that 78.8 % of the variance in this item is explained by the extracted factor.

Table 55: Complete Explained Variance for Multidimensional Variable Implementation of AI Technology in the Work Environment

Total Variance Explained						
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	**3.703**	**74.055**	74.055	3.703	74.055	74.055
2	.659	13.186	87.242			
3	.242	4.842	92.084			
4	.205	4.104	96.189			
5	.191	3.811	100.000			

Extraction method: Principal component analysis.

Table 55 shows that five measured variables resulted in five principal components, but only the first component has an eigenvalue greater than 1 (total: 3.703). This factor explains 74.1 % of the variability of the measured variables together.

Table 56: Factor Loadings for Multidimensional Variable Implementation of AI Technology in the Work Environment

Component Matrix[a]	
	Component
	1
1. Our company uses programme and portfolio structures for managing projects.	.888
2. Our company has a digital transformation strategy, including AI adoption.	.815
3. Our company uses AI technologies for work design.	.815
4. Our company uses AI technologies to better plan new tasks.	**.916**
5. Our company uses AI technologies in projects to create teams.	.864

Extraction method: Principal component analysis.
[a] 1 component extracted.

Table 56 shows that all factor loadings are higher than 0.60. Within the context of the *implementation of AI technology in the work environment* variable, the most important role is held by *Our company uses AI technologies to better plan new tasks* where the value of the factor loadings is the highest.

b) Cronbach's alpha for the multidimensional variable *implementation of AI technology in the work environment:*

Table 57: Cronbach's Alpha for the Multidimensional Variable Implementation of AI Technology in the Work Environment

Reliability Statistics	
Cronbach's Alpha	N of Items
.912	5

The reliability of an individual measurement scale used to analyse attitudes toward a particular topic, object, etc., refers to how well all individual items or statements measure the attitude being analysed. It is measured using an appropriate reliability indicator (reliability analysis), such as Cronbach's alpha coefficient (α). The reliability of measurement, with a coefficient $\alpha \geq 0.80$, is considered excellent. If the coefficient is in the range of $0.70 \leq \alpha < 0.80$, it is considered very good; in the range of $0.60 \leq \alpha < 0.70$, it is considered moderate, and if the coefficient α is less than 0.60, it is considered barely acceptable.

Table 57 indicates Cronbach's alpha value of 0.912, which falls within the excellent reliability range. Therefore, we can assert that the *implementation of AI technology in the work environment* factor exhibits excellent reliability.

Task 3.

The growing incorporation of artificial intelligence (AI) into higher education has prompted a scholarly interest in examining students' perceptions regarding its utility in academic contexts. A deeper understanding of these perceptions provides valuable insights for academic institutions, enabling them to design more effective digital learning environments and to develop AI-based support systems tailored to students' educational needs. The sample included 197 randomly selected master's students from the Faculty of Economics and Business, University of Maribor. Students rated their level of agreement with a set of statements on a 5-point Likert scale (1 – strongly disagree, 5 – strongly agree), each reflecting a specific aspect of AI usefulness in their study process.

The objective of this study is to determine whether the items determining students' perceptions of the usefulness of AI in their study *(variables Q1a to Q1h)* can be reduced to a smaller number of factors. Using the data in the file *Factor analysis_Students_AI.sav.*

Answers and Result Outputs

Table 58: Kaiser-Meyer-Olkin Statistics and Bartlett's Test of Sphericity for the Multidimensional Variable Students' Perception of the Usefulness of AI in Their Study

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.745
Bartlett's Test of Sphericity	Approx. Chi-Square	1,560.091
	df	28
	Sig.	< .001

Table 59: Communalities for Multidimensional Variable Students' Perception of the Usefulness of AI in Their Study

Communalities		
	Initial	Extraction
1. In my opinion, using AI in education improves the educational environment for learning.	1.000	.698
2. In my opinion, AI clarifies many points that the teacher cannot cover in their explanation.	1.000	.831
3. In my opinion, AI fulfils and complements all students' learning needs.	1.000	.541
4. In my opinion, AI enables students to obtain additional educational support that complements the teacher's work in the classroom.	1.000	.637
5. In my opinion. learning through artificial intelligence will make learning less terrifying than learning in the traditional way.	1.000	.565
6. In my opinion. AI changes the way students acquire skills in certain subjects.	1.000	.681
7. In my opinion, the teacher's role will diminish when the student uses artificial intelligence to learn certain subjects.	1.000	.677
8. In my opinion, using AI affects the ability to communicate with the teacher.	1.000	.825

Extraction method: Principal component analysis.

In Table 58, KMO > 0.5 (KMO = 0.745) and Bartlett's tests ($p < 0.05$) show that the use of factor analysis is justified.

Communalities in Table 59 for the multidimensional variable *students' perception of the usefulness of AI in their study* are higher than 0.40.

Table 60: Complete Explained Variance for the Multidimensional Variable Students' Perception of the Usefulness of AI in Their Study

Total Variance Explained									
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	**4.444**	**55.549**	55.549	4.444	55.549	55.549	3.232	40.398	40.398
2	**1.013**	**12.663**	**68.213**	1.013	12.663	68.213	2.225	27.814	68.213
3	.927	11.581	79.794						
4	.674	8.424	88.218						
5	.501	6.256	94.475						
6	.393	4.913	99.387						
7	.034	.430	99.818						
8	.015	.182	100,000						

Extraction method: Principal component analysis.

Table 60 shows that, from eight measured variables, two factors are extracted, as both have an eigenvalue greater than 1. Together, these factors explain 68.213 % of the total variance of all eight measured variables together, namely the first factor accounts for 55.549 % and the second factor for 12.663 %. In the rotated solution, the first factor explains 40.398 % and the second factor 27.814 % of the total variance of all eight measured variables together.

To improve the structure of extracted factors, the factor loadings were rotated using the Varimax method. Based on the loadings presented in Table 61, two different factors emerged. The first factor, which includes items 1, 2, 3, and 8, can be named *pedagogical support and learning enhancement*, reflecting the students' perception of AI as a tool that enhances the educational environment, supports teachers, and facilitates learning. The second factor, which includes items 4, 5, 6, and 7, can be named *independent learning and skill development through AI*, as it captures how students perceive AI as a means of gaining additional educational support, changing learning approaches, and reducing dependency on traditional teaching.

Table 61: Rotated Factor Loadings for Multidimensional Variable Students' Perception of the Usefulness of AI in Their Study

Rotated Component Matrix[a]		
	Component	
	1	**2**
1. In my opinion, using AI in education improves the educational environment for learning.	.753	.361
2. In my opinion, AI clarifies many points that the teacher cannot cover in his/her explanation.	.895	.173
3. In my opinion, AI fulfils and complements all students learning needs.	.572	.463
4. In my opinion, AI enables students to obtain additional educational support that complements the teacher's work in the classroom.	.300	.740
5. In my opinion. learning through artificial intelligence will make learning less terrifying than learning in the traditional way.	.132	.740
6. In my opinion, AI changes the way students acquire skills in certain subjects.	.278	.777
7. In my opinion, the teacher's role will diminish when the student uses artificial intelligence to learn certain subjects.	.745	.350
8. In my opinion, using AI affects the ability to communicate with the teacher.	.892	.170

Extraction method: Principal component analysis.

Rotation method: Varimax with Kaiser normalization.

[a] Rotation converged in 3 iterations.

5 Regression Analysis

5.1 Simple Linear Regression

Using simple linear regression, we analyse the relationship between one dependent variable (y) and one independent or explanatory variable (x_1) (assuming the conditions for conducting regression analysis are met).

The regression model assumes that the variation in the dependent variable can be explained by the variation in the independent variable using a linear function. (Tabachnick and Fidell, 2019). The equation for the simple linear regression model is as follows:

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

Where:

- y is the dependent variable,
- x is the independent variable,
- β_0 is the constant (constant estimate: average value of y when all explanatory variables equal 0),
- β_1 is the partial regression coefficient (average change in y associated with a one-unit increase in x),
- ε is the error term (unexplained variability).

Indicators of simple linear regression are as follows (Kline, 2023; Bastič, 2006):

- Estimated values of both regression coefficients (the regression constant and the explanatory variable).
- Coefficient of determination (r^2_{xy}): indicates the percentage of total variance of variable y (dependent variable) explained by the regression function or variable x_1 (independent variable). It defines the strength of the linear relationship between the variables. The value of the coefficient of determination ranges from 0 to 1 ($0 \leq r^2_{xy} \leq 1$).
- Correlation coefficient (r_{xy}): defines the strength and direction of the linear relationship between the dependent and independent variables. The value of the correlation coefficient ranges from -1 to 1 ($-1 \leq r_{xy} \leq 1$).
- Standard error of the estimate of the dependent variable (σ_{ey}): indicates whether variables other than x1 and random effects affect the variability of variable y.

The strength of the linear relationship between variables based on the values of the correlation and determination coefficients is as presented in Table 62 (Artenjak, 2003).

Table 62: The Strength of the Linear Relationship Between Variables

Correlation Coefficient (r_{xy})	Coefficient of Determination ($r^2{}_{xy}$)	Strength of Linear Relationship
0	0	No correlations
0–0.5	0–0.25	Weak correlation
0.51–0.79	0.26–0.64	Medium strong correlation
0.80–0.99	0.65–0.99	Strong correlation
1	1	Perfect correlation

We assess the overall quality of the regression model with an F-test, and the statistically significant impact of explanatory variables with a t-test (or a single independent variable x_1, in the case of simple regression).

The F-test tests the hypotheses:

H0: Coefficient of determination is equal to 0 ($r^2_{xy} = 0$).

H1: Coefficient of determination is greater than 0 ($r^2_{xy} > 0$).

The statistically significant impact of explanatory variable x1 is tested with a t-test, where we test the following hypotheses:

H0: Regression coefficient β1 is equal to 0 ($\beta_1 = 0$).

H1: Regression coefficient β1 is not equal to 0 ($\beta_1 \neq 0$).

Task 1

Sustainability has become a key strategic priority for modern companies, not only in terms of environmental and social responsibility but also as a driver of employee satisfaction. In order to explore the relationship between employees' perceptions of their company's sustainability commitment and their work satisfaction, a sample of 130 employees from Company X was included in the study. Employees evaluated their agreement with a series of statements on a 5-point Likert scale (1 – strongly disagree, 5 – strongly agree). Company X wants to examine whether the *perceived level of sustainability practices in the company* (independent variable or x_1) impacts *employee satisfaction* (dependent variable or y).

Using the data in the file *Regression analysis_Company's sustainability practices.sav,* factor analysis has previously been conducted, resulting in one factor – *employee satisfaction (y)* and one factor – *perceived sustainability commitment of the company (x_1).*

Check the results of the factor analysis for both obtained factors (for *employee satisfaction* and *perceived sustainability commitment of the company*) and explain them in terms of content. Save the values of both obtained factors in a data file. Next, perform a simple linear regression and explain the results.

Procedure: Factor Analysis (Saving the Values of Both Obtained Factors)

Click *Analyze*, then *Dimension Reduction*, and choose *Factor*. A dialog box opens; transfer the variables belonging to the multidimensional variable *employee satisfaction*. To save the values of the obtained factors in the data file, click *Scores*, check *Save as variables*, and under *Method*, select *Regression*. Repeat the same procedure for the multidimensional variable *perceived sustainability commitment of the company.*

Procedure: Simple Linear Regression

Click on *Analyze*, then Regression, and select *Linear*. A dialog box opens where you transfer the variable *employee satisfaction* to the right box under *Dependent* and the variable *perceived sustainability commitment of the company* to the right box under *Independent*, then click *Continue* and *OK*.

Answers and Output of Results

Table 63: Results: Correlation and Determination Coefficient for Sustainability Commitment and Work Satisfaction

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	**.960**[a]	**.922**	.921	**.8218458**

[a] Predictors: (constant), FAC1_Perceived sustainability commitment of the company.

The correlation coefficient value is 0.960, indicating a strong linear relationship between the variables *perceived sustainability commitment of the company* and *employee satisfaction*. The determination coefficient value is 0.922, which explains that 92.2 % of the total variance in *employee satisfaction* is successfully explained by the variability of the independent variable *perceived sustainability commitment of the company*. The standard error of the estimate of the dependent variable is different from zero, which indicates that the variability of the variable y is influenced by other variables and random effects besides the independent variable *perceived sustainability commitment of the company* (Table 62).

Table 64: Results: F-Test for Sustainability Commitment and Work Satisfaction

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	118.795	1	118.795	1491.873	**< .001**[b]
	Residual	10.113	127	.080		
	Total	128.908	128			

[a] Dependent variable: FAC1_Employee satisfaction.

[b] Predictors: (constant), FAC1_Perceived sustainability commitment of the company.

We assess the overall quality of the regression model with an F-test that tests the hypotheses:

H0: Coefficient of determination is equal to 0 ($r^2_{xy} = 0$).

H1: Coefficient of determination is greater than 0 ($r^2_{xy} > 0$).

Based on the p-value ($p < 0.05$), we conclude that the model as a whole is "of good quality" and reject the null hypothesis (H0: $r^2_{xy} = 0$). We conclude that the coefficient of determination is greater than 0 (H1: $r^2_{xy} > 0$), which indicates that there is a linear dependence between the independent variable and the dependent variable.

Table 65: Results: Regression Coefficients and T-Test for Sustainability Commitment and Work Satisfaction

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	−.002	.025		−.094	.925
	FAC1_Perceived sustainability commitment of the company	.963	.025	.960	38.625	**< .001**

[a] Dependent variable: FAC1_Employee satisfaction.

The statistically significant effect of the independent variable x_1 is tested using a t-test, with the following hypotheses:

H0: Regression coefficient β1 is equal to 0 ($\beta_1 = 0$).

H1: Regression coefficient β1 is not equal to 0 ($\beta_1 \neq 0$).

The t-test value and the significance level ($p < 0.05$) indicate that the regression coefficient β_1 is different from zero, therefore, we accept the research hypothesis H1: $\beta_1 \neq 0$. This indicates that the independent variable *perceived sustainability commitment of the company* has a statistically significant effect on the dependent variable *employee satisfaction.* This is also supported by the regression model results, as only this independent variable is included in the model (see Table 64)

The regression function (Table 65) is written as:

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

Where:

- $\hat{y}_i$ is the estimated value of variable *y* for the i-th observed value of variable *x*;
- b_0 in b_1 are estimated regression coefficients.

$$\hat{y} = -0.002 + 0.963x_1$$

The regression coefficient b_0 presents the estimated value of the dependent variable (*employee satisfaction)* when the independent variable *(perceived sustainability commitment of the company)* is equal to zero. This means that if the company does not invest in sustainable development, the predicted level of employee satisfaction is on average –0.002 units (in this this is not a meaningful result).

The regression coefficient b_1 indicates the expected change in the dependent variable *(employee satisfaction)* for a one-unit increase in the independent variable *(perceived sustainability commitment of the company)*. For example, the regression coefficient b_1 indicates that for every one-unit increase in the perceived sustainability commitment of the company, the predicted level of employee satisfaction increases by 0.963 units on average.

Task 2

Sustainability-oriented innovation has become a crucial part of modern business strategies. In this task, we are interested in whether the level of green innovation adoption in a company (measured by the number of implemented green initiatives) impacts its annual revenue (measured in thousands of euros).

Perform a simple linear regression analysis and interpret the results. Use the data from the file *Regression analysis_Sustainability_Innovation_Revenue.sav* (in this case, the company's annual revenue is the dependent variable, and the number of green initiatives adopted in a company is the independent variable).

a) Explain the correlation coefficient and coefficient of determination.
b) Check the appropriateness of the model as a whole and write the corresponding hypothesis. Explain the results.
c) State the appropriately formulated hypotheses for testing the statistical significance of regression coefficients in the regression function and explain the results.

Answers and Output of Results

Table 66: Results: Correlation and Determination Coefficient for Green Innovation and Revenue

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	**.880**[a]	**.775**	.769	595.602

[a] Predictors: (constant), number of green initiatives adopted in a company (x).

Table 66 shows that the correlation coefficient value is 0.880, indicating a strong linear relationship between the variables *number of green initiatives adopted in a company (x)* and *company's annual revenue in thousands of euros (y)*. The determination coefficient explains that

77.5 % of the total variance in the dependent variable is explained by the variability of the independent variable (*number of green initiatives adopted in a company*).

Table 67: Results: F-test for Green Innovation and Revenue.

ANOVA[a]						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	46,345,100.888	1	46,345,100.888	130.644	**< .001[b]**
	Residual	13,480,201.487	38	354,742.144		
	Total	59,825,302.375	39			

[a] Dependent variable: company's annual revenue in thousands of euros (y).

[b] Predictors: (constant) number of green initiatives adopted in a company (x).

Based on the p-value ($p < 0.05$), we conclude that the coefficient of determination is greater than 0 (H1: $r^2_{xy} > 0$), and that the model as a whole is "of good quality" (Table 67).

Table 68: Results: Regression Coefficients and T-Test for Green Innovation and Revenue

Coefficients[a]						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1,354.336	259.065		5.228	< .001
	Number of Green Initiatives Adopted in a Company (x)	253.077	22.142	.880	11.430	**< .001**

[a] Dependent variable: company's annual revenue in thousands of euros (y).

The t-test value and the significance level ($p < 0.05$) indicate that the regression coefficient β_1 is different from zero; therefore, we reject the hypothesis H0: $\beta1 = 0$. This means that the independent variable *number of green initiatives adopted in a company* has a statistically significant impact on the dependent variable *company's annual revenue* (Table 67).

Task 3

In the age of digital transformation, companies increasingly invest in digital tools to improve internal efficiency and workforce performance. Understanding how the level of digital technology adoption affects employee productivity can help organizations make more informed decisions about technological investments. In this task, we examine whether the number of digital tools adopted by a company (e.g., project management software, internal communication platforms and data analytics systems) impacts the average productivity per employee (measured in output per employee per month in EUR).

Use the data from the file *Regression analysis_Digital Tools and Employee Productivity.sav.*

a) Explain the correlation coefficient and coefficient of determination.

b) Check the appropriateness of the model as a whole, write down the appropriately formulated hypothesis and explain the results.

c) Form appropriately formulated hypotheses for testing the statistical significance of regression coefficients in the regression function and explain the results.

d) Write down the equation of the regression function with the estimated regression coefficients.

Answers and Output of Results

Table 69: Results: Correlation and Determination Coefficient for the Number of Digital Tools and the Average Productivity per Employee

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	**.959**[a]	**.919**	.917	196.078

[a] Predictors: (constant), digital_tools.

Table 70: Results: F-Test for the Number of Digital Tools and the Average Productivity per Employee

ANOVA[a]						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	20,864,381.963	1	20,864,381.963	542.686	**< .001**[b]
	Residual	1,845,431.717	48	38,446.494		
	Total	22,709,813.680	49			

[a] Dependent variable: productivity.

[b] Predictors: (constant), digital_tools.

Table 71: Results: Regression Coefficients and T-Test for the Number of Digital Tools and the Average Productivity per Employee

Coefficients[a]						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2,013.855	92.923		21.672	< .001
	Digital_Tools	301.176	12.928	.959	23.296	**< .001**

[a] Dependent variable: productivity.

5.2 Multiple Regression Analysis

Multiple regression is used when the dependent variable *(y)* is influenced by more than one independent variable x_i ($i = 1, 2, \ldots, k$), and when the assumptions for conducting multiple regression analysis are met (Tabachnick & Fidell, 2013).

The multiple correlation coefficient *R* indicates the strength of the association between the dependent variable and the set of *k* independent variables and is expressed in absolute terms.

The adjusted coefficient of determination R^2 presents the proportion of the variance in the dependent variable that is explained by the variability in the independent variables included in the model (ibid).

Task 1

Amid growing environmental awareness and climate-related concerns, it has become increasingly important to understand the psychological and informational drivers behind sustainable consumer behaviour. Individuals are increasingly exposed to alarming news about environmental degradation, climate change, and biodiversity loss, which can trigger emotional responses such as anxiety, but also a greater willingness to act responsibly.

In this task, we aim to investigate whether two psychological and informational factors, *climate-related anxiety* (x_1) and *exposure to sustainability topics in the media* (x_2), can statistically significantly impact an individual's willingness to buy environmentally-friendly products (y).

Use the data from the file *Multiple regression analysis_Green Purchasing Behavior.sav.*

a) Conduct a factor analysis and save the values of the obtained factor (single-factor solutions) for the regression analysis.

b) Explain the multiple correlation coefficient and adjusted coefficient of determination.

c) Check the appropriateness of the model as a whole and write down the appropriately formulated hypothesis. Explain the results.

d) State appropriately formulated hypotheses for testing the statistical significance of regression coefficients in the regression function and explain the results.

e) Write the equation of the regression function with the estimated regression coefficients.
Answers and Output of Results

Table 72: Results: Multiple Correlation and Adjusted Determination Coefficients for Green Purchasing Behaviour

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	**.931**[a]	.867	**.865**	.36927049

[a] Predictors: (constant), FAC1_Exposure to Sustainability Topics in Media (x_2), FAC1_Level of Climate-Related Anxiety (x_1).

Table 72 shows that the value of the multiple correlation coefficient is 0.931, indicating a strong correlation between the dependent variable *willingness to buy environmentally friendly products,* and the independent variables *level of climate-related anxiety* and *exposure to sustainability topics in media*. The value of the adjusted multiple determination coefficient is 0.865, which means that 86.5 % of the total variance in the dependent variable *willingness to buy environmentally friendly products* is explained by the regression model (with the independent variables being *level of climate-related anxiety* and *exposure to sustainability topics in the media*).

Table 73: Results: F-Test for Green Purchasing Behaviour

ANOVA[a]						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	111.743	2	55.871	409.732	**< .001**[b]
	Residual	17.181	126	.136		
	Total	128.924	128			

[a] Dependent variable: FAC1_Willingness to Buy Environmentally Friendly Products (y).

[b] Predictors: (constant), FAC1_Exposure to Sustainability Topics in Media (x_2), FAC1_Level of Climate-Related Anxiety (x_1).

The quality of the regression model as a whole was assessed with an F-test, and based on the p-value ($p < 0.05$), we can conclude that the model is of high quality. This implies that there is a dependency between the dependent variable (*willingness to buy environmentally friendly products*) and at least one independent variable.

The F-test is used to test the following hypotheses:

H0: The adjusted determination coefficient is equal to 0 ($R^2 = 0$).

H1: The adjusted determination coefficient is greater than 0 ($R^2 > 0$).

Based on the results, we can reject the null hypothesis that $R^2 = 0$ (Table 72).

Table 74: Results: Regression Coefficients and T-Test for Green Purchasing Behaviour

Coefficients[a]						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	**.000**	.033		−.009	.993
	FAC1_Level of Climate-Related Anxiety (x_1)	**.566**	.093	.566	6.097	**< .001**
	FAC1_Exposure to Sustainability Topics in Media (x_2)	**.381**	.093	.380	4.091	**< .001**

[a] Dependent variable: FAC1_Willingness to Buy Environmentally Friendly Products (y).

For Table 74, to test the statistical significance of regression coefficients in the regression function, the following hypotheses are formulated:

x_1 (level of climate-related anxiety):

H0: $\beta_1 = 0$

H1: $\beta_1 \neq 0$

The t-test and significance levels for regression coefficients ($p < 0.05$) indicate that the independent variable *level of climate-related anxiety* statistically significantly impacts the dependent variable; therefore, we accept H1.

x_2 (exposure to sustainability topics in media):

H0: $\beta_2 = 0$

H1: $\beta_2 \neq 0$

The t-test and significance levels for regression coefficients ($p < 0.05$) indicate that the independent variable *exposure to sustainability topics in media* statistically significantly impacts the dependent variable; therefore, we accept H1.

The obtained equation for the regression function, with estimated values for the regression coefficients based on the sample data used is as follows:

$$\hat{y} = 0.000 + 0.566\ x_1 + 0.381\ x_2$$

Here, x_1 presents the *level of climate-related anxiety* and x_2 presents the *exposure to sustainability topics in media.*

Estimated values of regression coefficients indicate by how many units, on average, the dependent variable changes when the value of the individual independent variable increases by 1 unit, while the value of the other independent variable remains unchanged (assuming there is no multicollinearity between independent variables).

Multicollinearity is a statistical phenomenon that arises in multiple regression analysis when two or more independent variables are highly linearly correlated. This high intercorrelation reduces the ability of the model to estimate the unique contribution of each independent variable to the explanation of the dependent variable. As a result, regression coefficients may become unstable, standard errors inflated, and statistical significance tests unreliable. To detect multicollinearity, we use the Variance Inflation Factor (VIF). A VIF value greater than 10 (VIF > 10) is generally considered to indicate the presence of problematic multicollinearity (Kutner et al., 2004).

Task 2

In the context of the digital economy, companies increasingly depend on strategic marketing efforts and brand positioning to enhance their business performance. In this task, we aim to investigate whether two independent variables, *number of digital marketing campaigns* and *brand awareness score* (the level of brand recognition among consumers, measured on a 5-point Likert scale, where: 1 – very low brand awareness, 5 – very high brand awareness), have a statistically significant impact on the *company's sales revenue* (in EUR).

Conduct a multiple linear regression analysis and provide a substantive explanation of the obtained results. Use the data from *Multiple Regression_Company's sales revenue.sav.*

Use multiple regression analysis to test the significance of the predictors and interpret the results in context.

a) Explain the multiple correlation coefficient and adjusted determination coefficient.

b) Check the appropriateness of the model as a whole and write the appropriately formulated hypothesis. Explain the results.

c) Write the appropriately formulated hypotheses for testing the statistical significance of regression coefficients in the regression function and explain the results.

d) Write and explain the equation of the regression function.

e) Use appropriate indicators to check for the possible presence of multicollinearity. Provide a substantive explanation of the results.

Procedure for the multicollinearity: Click on *Analyze*, then *Regression*, and select *Linear*. A dialog box opens where you transfer the variable *company's sales revenue* to the right box under *Dependent* and the variables *number of digital marketing campaigns* and *brand awareness score* to the right box under *Independent*, then click *Statistics*, and select the *Collinearity diagnostics* box. Click *Continue* and *OK*.

Answers and Output of Results

Table 75: Results: Multiple Correlation and Adjusted Determination Coefficients for Sales Revenue

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	**.894**[a]	.800	**.791**	583.558

[a] Predictors: (constant), brand awareness score, number of sales campaigns.

Table 75 shows that the value of the multiple correlation coefficient is 0.894, indicating a strong correlation between the dependent variable *company's sales revenue,* and the independent variables *number of digital marketing campaigns* and *brand awareness score*. The value of the adjusted multiple determination coefficient is 0.791, which means that 79.1 % of the total variance in the dependent variable *company's sales revenue* is explained by the regression model.

Table 76: Results: F-Test for Sales Revenue

ANOVA[a]						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	63,976,848.723	2	31,988,424.361	93.935	**< .001**[b]
	Residual	16,005,359.277	47	340,539.559		
	Total	79,982,208.000	49			

[a] Dependent variable: sales revenue.

[b] Predictors: (constant), brand awareness score, number of sales campaigns.

The quality of the regression model as a whole was assessed with an F-test, and based on the p-value ($p < 0.05$), we can conclude that the model is of high quality. We accept hypothesis H1: "The adjusted determination coefficient is greater than 0 ($R^2 > 0$)" (Table 76).

Table 77: Results: Regression Coefficients, T-Tests, and Collinearity Indicators for Sales Revenue

Coefficients[a]								
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	**VIF**
1	(Constant)	**424.617**	438.218		.969	.338		
	Number of Sales Campaigns	**206.586**	25.838	.719	7.995	**< .001**	.526	**1.902**
	Brand Awareness Score	**340.246**	132.513	.231	2.568	**.013**	.526	**1.902**

[a] Dependent variable: sales revenue.

For Table 77, to test the statistical significance of regression coefficients in the regression function, we state the following hypotheses:

x_1 (number of sales campaigns):

H0: $\beta_1 = 0$

H1: $\beta_1 \neq 0$

x_2 (brand awareness score):

H0: $\beta_2 = 0$

H1: $\beta_2 \neq 0$

The t-test and significance levels for both regression coefficients ($p < 0.05$) indicate that both independent variables *number of sales campaigns (x_1)* and *brand awareness score (x_2),* statistically significantly impact the dependent variable *company's sales revenue*, therefore, we reject H0 in both cases.

The obtained equation for the regression function, with estimated values for the regression coefficients based on the sample data used, is as follows:

$$\hat{y} = 424.617 + 206.586\ x_1 + 340.246\ x_2$$

(In this case, x_1 presents the *number of sales campaigns* and x_2 presents the *brand awareness score*).

The estimated regression coefficient for x_1 indicates that the dependent variable *company's sales revenue*, on average, increases by 206.586 units when the *number of sales campaigns* increases by 1 unit, with the value of the variable *brand awareness score,* remaining unchanged.

The Tolerance and VIF (VIF < 10) indicators show that there is no significant degree of dependence between the variables (multicollinearity is not present).

Task 3

In the context of rapid technological advancement and global competition, innovation and strategic investment have become key drivers of business growth and competitive advantage. Companies that prioritize innovation and allocate sufficient financial resources often achieve superior performance and long-term sustainability.

In this task, we aim to explore whether *innovation activity level* (measured using a 10-point numeric rating scale, ranging from 1 (very low innovation activity) to 10 (very high innovation activity) and *annual R&D investment* (in thousands of euros) have a statistically significant impact on the *company's annual business output* (in thousands of euros).

The sample includes 125 randomly selected companies operating in the technology and manufacturing sectors. Use the data from the file *Multiple Regression_Company's annual business output.sav.*

a) Explain the multiple correlation coefficient and adjusted determination coefficient.

b) Check the appropriateness of the model as a whole and write the appropriately formulated hypothesis. Explain the results.

c) Write the appropriately formulated hypotheses for testing the statistical significance of regression coefficients in the regression function and explain the results.

d) Write and explain the equation of the regression function.

Answers and Output of Results

Table 78: Results: Multiple Correlation and Adjusted Determination Coefficients for the Company's Annual Business Output

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	**.833**[a]	.694	**.681**	93.320

[a] Predictors: (constant), innovation activity level, R&D investment.

Table 79: Results: F-Test for the Company's Annual Business Output

ANOVA[a]						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	928,803.387	2	464,401.694	53.326	**< .001**[b]
	Residual	409,309.033	47	8,708.703		
	Total	1,338,112.420	49			

[a] Dependent variable: company's annual business output.

[b] Predictors: (constant), innovation activity level, R&D investment.

Table 80: Results: Regression Coefficients, T-Tests, and Collinearity Indicators for the Company's Annual Business Output

Coefficients[a]						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	17.616	85.167		.207	.837
	R&D Investment	.494	.072	.559	6.895	**< .001**
	Innovation Activity Level	30.440	4.348	.567	7.001	**< .001**

[a] Dependent variable: company's annual business output.

6 Logistic Regression

For logistic regression analysis in this tutorial, the description follows the IBM SPSS instructions (IBM, 2024).

Logistic regression is useful for situations in which we want to be able to predict the presence or absence of a characteristic or outcome based on values of a set of predictor variables. It is similar to a linear regression model but it is suited to models where the dependent variable is dichotomous (i.e., it has only two possible values such as 0 and 1 or yes and no). Logistic regression coefficients can be used to estimate odds ratios for each independent variable in the model. Moreover, logistic regression is applicable to a broader range of research situations than discriminant analysis (described in Chapter 7).

Example (UCLA, 2024):

A researcher is interested in how variables such as GRE (Graduate Record Exam) scores, GPA (grade point average), and the prestige of the undergraduate institution (bachelor's studies) effect admission into graduate school (master's studies). The response variable, whether an individual is admitted or not, is binary.

Statistics – Indicators

For each analysis: total cases, selected cases, valid cases.

For each categorical variable: parameter coding.

For each step: variables entered or removed, iteration history, −2 log-likelihood, goodness of fit, Hosmer-Lemeshow goodness-of-fit statistic, model chi-square, improvement chi-square, classification table, correlations between variables, observed groups and predicted probabilities chart, residual chi-square.

For each variable in the equation: coefficient (*B*), standard error of *B*, Wald statistic, estimated odds ratio (exp(*B*)), confidence interval for exp(*B*), log-likelihood if term removed from model.

For each variable not in the equation: score statistic.

For each case: observed group, predicted probability, predicted group, residual, standardized residual.

Methods

Models can be estimated using block entry of variables or any of the following stepwise methods: forward conditional, forward LR, forward Wald, backward conditional, backward LR, or backward Wald.

Data Considerations

The dependent variable should be dichotomous. Independent variables can be interval level or categorical; if categorical, they should be dummy or indicator coded (there is an option in the procedure to recode categorical variables automatically).

Assumptions

Logistic regression does not rely on distributional assumptions in the same sense as for example the discriminant analysis does. However, the solution may be more stable if predictors have a multivariate normal distribution. Additionally, as with other forms of regression, multicollinearity among the predictors can lead to biased estimates and inflated standard errors. The procedure is most effective when group membership is a truly categorical variable; if group membership is based on values of a continuous variable (for example, *high IQ* versus *low IQ*), we should consider using linear regression to take advantage of the richer information offered by the continuous variable itself.

The scatterplot procedure can be used to screen the data for multicollinearity. If assumptions of multivariate normality and equal variance-covariance matrices are met, the discriminant analysis procedure may provide a quicker solution. If all predictor variables are categorical, the loglinear procedure can also be used. If the dependent variable is continuous, we can use the linear regression procedure.

Task 1 (UCLA, 2024)

We have generated hypothetical data, which can be obtained by choosing the data file *Binary_logistic_regression.sav*. This dataset has a binary response (outcome, dependent) variable called *admit*, which equals 1 if the individual was admitted to graduate school (master's studies), and 0 otherwise. There are three predictor variables: *GRE, GPA*, and *rank*. We will treat the variables *GRE* and *GPA* as continuous. The variable *rank* takes on values 1 through 4. Institutions with a rank of 1 have the highest prestige, while those with a rank of 4 have the lowest. We start out by opening the dataset and looking at some descriptive statistics results in Tables 80 and 81.

Table 81: Descriptive Statistics for Variables GRE and GPA

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
GRE	400	220	800	587.70	115.517
GPA	400	2.26	4.00	3.3899	.38057
Valid N (Listwise)	400				

Table 81 presents the basic descriptive statistics for the two continuous predictor variables *GRE* (Graduate Record Examination scores) and *GPA* (grade point average). The sample includes 400 observations. The average GRE score is 587.70, with values ranging from 220 to 800, and a relatively high standard deviation of 115.52, indicating considerable variation in the scores. The average GPA is 3.39, with values ranging from 2.26 to 4.00, and a standard deviation of 0.38, suggesting that GPA scores are more tightly clustered around the mean compared to GRE scores. These descriptive statistics provide a useful initial overview of the scale and dispersion of the continuous predictors used in the logistic regression analysis.

Table 82 shows the frequency distribution of the categorical variable *rank*, which presents the prestige of the undergraduate institution attended by the applicants. The *rank* values range from 1 (highest prestige) to 4 (lowest prestige). Most applicants in the sample come from institutions ranked 2 (37.8 %), followed by rank 3 (30.3 %), rank 4 (16.8 %), and the least from rank 1 (15.3 %).

Table 82: Frequency Distribution for Rank

Rank					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	61	15.3	15.3	15.3
	2	151	37.8	37.8	53.0
	3	121	30.3	30.3	83.3
	4	67	16.8	16.8	100.0
	Total	400	100.0	100.0	

This distribution indicates that the majority of students in the dataset graduated from mid-prestige institutions. Since rank is included as a categorical predictor in the logistic regression model, understanding its distribution helps in interpreting the reference group and comparisons across institutional prestige levels.

Procedure for logistic regression: Click on *Analyze*, then *Regression*, and select *Binary Logistic*. A dialog box opens; transfer the variable *admit* into the *Dependent* field. Transfer all three independent variables into the *Covariates* field. Select the *Categorical* button and choose *rank* for the *Categorical Covariates*. Click *Continue* and *OK*.

Answers and Output of Results

Table 83: Case Processing Summary

Case Processing Summary			
Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	400	100.0
	Missing Cases	0	.0
	Total	400	100.0
Unselected Cases		0	.0
Total		400	100.0

[a] If weight is in effect, see classification table for the total number of cases.

Table 83 shows the case processing summary for the logistic regression model. All 400 cases (100 %) were included in the analysis, with no missing data.

Table 84: Dependent Variable Encoding

Dependent Variable Encoding	
Original Value	Internal Value
0	0
1	1

Table 84 displays the coding scheme for the dependent variable (*ADMIT*), which is binary (original values in the dataset were 0 and 1). These were internally coded the same way in the logistic regression model (0 and 1), which is important because logistic regression models the log-odds of the event coded as 1 (in this case, being admitted).

Table 85: Categorical Variables Codings

Categorical Variables Codings					
		Frequency	Parameter coding		
			(1)	(2)	(3)
Rank	1	61	.000	.000	.000
	2	151	1.000	.000	.000
	3	121	.000	1.000	.000
	4	67	.000	.000	1.000

Table 85 shows how the values of the categorical variable *rank* were handled. The model includes terms (essentially dummy variables) for *rank – 1*, *rank – 2*, and *rank – 3, with rank – 4* being the omitted (reference) category.

The first model in the output is a null model (Table 83–85), meaning it contains no predictors. The constant in the table labelled *Variables in the Equation* gives the unconditional log-odds of admission (i.e., *admit – 1*).

The table labelled *Variables Not in the Equation* gives the results of a score test, also known as a Lagrange multiplier test (follows after Table 85). The column labelled *Score* gives the estimated change in model fit if the term is added to the model; the other two columns give the degrees of freedom and p-value (labelled *Sig.*) for the estimated change. Based on the table above, all three of the predictors, *GRE*, *GPA*, and *rank*, are expected to improve the fit of the model.

Tables 86, 87 and 88 present the null model.

Table 86: Classification Table

Classification Table[a,b]					
	Observed		Predicted		
			admit		Percentage Correct
			0	1	
Step 0	Admit	0	273	0	100.0
		1	127	0	.0
	Overall Percentage				68.3

[a] Constant is included in the model.

[b] The cut value is .500.

Table 86 presents the classification table for the null model, which includes only the constant and no predictor variables. The model predicts that all cases fall into the most frequent category — in this case, not admitted (admit – 0). As a result, all 273 individuals who were not admitted were correctly classified, while none of the 127 admitted individuals were correctly predicted. The overall classification accuracy is 68.3 %, which reflects the base rate of the majority class, not the predictive power of the model.

Table 87: Variables in the Equation

Variables in the Equation							
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	−.765	.107	50.764	1	< .001	.465

Table 87 provides the results for the null model's intercept (constant). The coefficient for the constant is −0.765, which corresponds to the log-odds of being admitted when no predictors are included. The associated p-value is p < 0.001, indicating that the constant is significantly different from zero. The Exp(B) value, 0.465, represents the odds of being admitted without considering any explanatory variables – that is, an individual has approximately a 46.5 % chance of being admitted in the absence of predictor information.

Table 88: Variables Not in the Equation

Variables Not in the Equation					
			Score	df	Sig.
Step 0	Variables	GRE	13.606	1	< .001
		GPA	12.704	1	< .001
		Rank	25.242	3	< .001
		Rank(1)	1.801	1	.180
		Rank(2)	5.934	1	.015
		Rank(3)	7.114	1	.008
	Overall Statistics		40.160	5	< .001

Table 88 displays the results of the score test (also known as the Lagrange multiplier test) for each predictor variable. This test examines whether each variable would significantly improve the model if it were added individually. All three main predictors *GRE*, *GPA*, and *rank*, have statistically significant score values (p < 0.001), suggesting that each would significantly contribute to the model. Among the dummy-coded rank variables, *rank(2)* and *rank(3)* are significant, while *rank(1)* is not (p = 0.180), implying that certain institutional prestige levels have a stronger association with admission outcomes than others.

Tables 89, 90, and 91 present Block 1 results.

Table 89: Omnibus Tests of Model Coefficients

Omnibus Tests of Model Coefficients				
		Chi-Square	df	Sig.
Step 1	Step	41.459	5	< .001
	Block	41.459	5	< .001
	Model	41.459	5	< .001

Table 89 gives the overall test for the model that includes the predictors. The chi-square value of 41.46 with a p-value of less than 0.05 tells us that our model as a whole, fits significantly better than an empty model (i.e., a model with no predictors).

Table 90: Model Summary

Model Summary			
Step	−2 Log-Likelihood	Cox & Snell R Square	Nagelkerke R Square
1	458.517[a]	.098	.138

[a] Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

The −2 log-likelihood (458.517) in the *Model Summary* table can be used to compare nested models, but we won't show an example of that here. This table also gives two measures of pseudo R-square. Many different measures of pseudo-R-squared exist. They all attempt to provide information similar to that provided by Rsquared in OLS regression; however, none of them can be interpreted exactly as R-squared in OLS regression is interpreted (Table 90).

Table 91: Classification Table

Classification Table[a]					
	Observed		Predicted		
			admit		Percentage Correct
			0	1	
Step 1	Admit	0	254	19	93.0
		1	97	30	23.6
	Overall Percentage				71.0

[a] The cut value is .500.

Table 92: Variables in Equation Results

Variables in the Equation							
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	GRE	.002	.001	4.284	1	.038	1.002
	GPA	.804	.332	5.872	1	.015	2.235
	Rank			20.895	3	< .001	
	Rank(1)	−.675	.316	4.555	1	.033	.509
	Rank(2)	−1.340	.345	15.064	1	< .001	.262
	Rank(3)	−1.551	.418	13.787	1	< .001	.212
	Constant	−3.990	1.140	12.251	1	< .001	.019

[a] Variable(s) entered on step 1: GRE, GPA, rank.

In Table 92, we see the coefficients, their standard errors, the Wald test statistic with associated degrees of freedom and p-values, and the exponentiated coefficient (also known as an odds ratio).

Both *GRE* and *GPA* are statistically significant.

The overall (i.e., multiple degree of freedom) test for rank is given first, followed by the terms for *rank* – 1, *rank* – 2, and *rank* – 3. The overall effect of rank is statistically significant, as are the terms for *rank* – 1 and *rank* – 2.

The logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the predictor variable. For every one unit change in *GRE,* the log odds of admission (versus non-admission) increase by 0.002. According to Exp(B), by increasing the *GRE* results, it becomes more likely to be admitted.

For a one unit increase in *GPA,* the log odds of being admitted to graduate school increase by 0.804, meaning that unit increase in *GPA* increases the likelihood of admission.

The indicator variables for rank have a slightly different interpretation. For example, having attended an undergraduate institution with a rank of 1 versus an institution with a rank of 4, increases the log odds of admission by 1.551. The likelihood that an individual having attended an undergraduate institution with a rank of 1 is admitted for master's studies is 4.7 times higher than that of an individual who attended an undergraduate institution with a rank of 4 (Exp(*B*) = 1.4718).

Task 2

Remote work has become an integral part of modern employment models, offering flexibility to both employees and organizations. Understanding which factors influence the likelihood of working remotely can help employers design better work policies. This task explores whether job satisfaction and employment sector are associated with an employee's likelihood to work from home.

In this task, we want to investigate whether the independent variables job satisfaction (rated from 1 to 5) and employment sector (1 – public, 2 – private) significantly impact the likelihood that an employee works remotely. The dependent variable *remote work* is a binary variable (1 – works remotely, 0 – does not work remotely).

Open data file *Binary logistic regression_Remote work.sav.*

Answers and Output of Results

Table 93: Case Processing Summary

Case Processing Summary			
Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	60	100.0
	Missing Cases	0	.0
	Total	60	100.0
Unselected Cases		0	.0
Total		60	100.0

[a] If weight is in effect, see classification table for the total number of cases.

Table 93 provides an overview of the cases included in the binary logistic regression analysis. The dataset contains 60 valid cases, with no missing data. All cases were included in the analysis, which ensures the stability and completeness of the model estimation.

Table 94: Dependent Variable Encoding

Dependent Variable Encoding	
Original Value	Internal Value
Does Not Work Remotely	0
Works Remotely	1

Table 94 shows how the dependent variable *remote work* was encoded for the logistic regression model. The original category *does not work remotely* was assigned a value of 0, and *works remotely* was assigned a value of 1. This binary coding is essential, as logistic regression estimates the log-odds of the event coded as 1; in this case, the likelihood that an employee works remotely. All predictor effects will be interpreted in terms of increasing or decreasing the probability of working remotely (coded 1).

Tables 95, 96 and 97 present the null model.

Table 95: Classification Table

Classification Table[a,b]					
	Observed		Predicted		
			Remote_work		Percentage Correct
			Does Not Work Remotely	Works Remotely	
Step 0	Remote Work	Does Not Work Remotely	0	24	.0
		Works Remotely	0	36	100.0
	Overall Percentage				60.0

[a] Constant is included in the model.
[b] The cut value is .500.

Table 95 presents the classification accuracy of the model at Step 0, which includes only the constant (intercept) and no predictor variables. At this stage, the model makes predictions based solely on the majority class in the dataset. In this case, the model predicts that all individuals work remotely.

As shown in Table 95, the model correctly classified all 36 individuals who work remotely, but it failed to correctly classify any of the 24 individuals who do not work remotely. Consequently, the classification accuracy for the *does not work remotely* group is 0 %, while the accuracy for the *works remotely* group is 100 %. The overall prediction accuracy of the model is 60.0 %, which corresponds with the proportion of the majority class.

Table 96: Variables in the Equation

Variables in the Equation							
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.405	.264	2.367	1	.124	1.500

Table 96 displays the regression coefficient (B), standard error (S.E.), Wald statistic, degrees of freedom (df), significance level (Sig.), and odds ratio (Exp(B)) for the constant in the model.

Since this is Step 0, no predictor variables have been entered yet (only the intercept (constant) is included). The value of the constant (B = 0.405) represents the log odds of an employee working remotely when no independent variables are considered. The corresponding odds ratio Exp(B) = 1.500 suggests that the odds of working remotely are 1.5 to 1 under the baseline model. However, the p-value (Sig. = 0.124) indicates that the intercept is not statistically significant at the 0.05 level. This is typical for Step 0, where the model does not yet include explanatory variables. In the next step, predictor variables such as *job satisfaction* and *employment sector* will be added to test whether they significantly improve the model's explanatory power.

Table 97: Variables Not in the Equation

Variables Not in the Equation					
			Score	df	Sig.
Step 0	Variables	Job_Satisfaction	18.129	1	< .001
		Sector	23.616	1	< .001
	Overall Statistics		24.815	2	< .001

Table 97 shows that the score statistic for job satisfaction is 18.129, and for sector, it is 23.616, both with p-values < 0.001, indicating that each predictor, if included individually, would significantly contribute to the model. The overall model score statistic is 24.815

with 2 degrees of freedom and a significance level < 0.001, showing that the model with both variables included would significantly improve prediction accuracy compared to the null model (which includes only the constant).

Tables 98, 99, and 100 present Block 1 results.

Table 98: Omnibus Tests of Model Coefficients

Omnibus Tests of Model Coefficients				
		Chi-Square	df	Sig.
Step 1	Step	28.905	2	< .001
	Block	28.905	2	< .001
	Model	28.905	2	< .001

Table 98 shows the performance of the model using only the constant (null model). It predicted that all cases would fall into the category *works remotely*. While the model correctly predicted all 36 cases where employees worked remotely (100 % accuracy), it failed to correctly classify the 24 cases where employees did not work remotely (0 % accuracy). Overall, the model correctly classified 60.0 % of the cases. This shows the baseline accuracy without including any independent variables.

Table 99: Model Summary

Model Summary			
Step	−2 Log-Likelihood	Cox & Snell R Square	Nagelkerke R Square
1	51.857[a]	.382	.517

[a] Estimation terminated at iteration number 5 because parameter estimates changed by less than.001.

The model summary provides information on the fit of the logistic regression model. The −2 log-likelihood value is 51.857, which indicates how well the model fits the data – the lower the value, the better the model. The pseudo R-squared statistics, Cox & Snell R^2 = 0.382 and Nagelkerke R^2 = 0.517, suggest that the model explains between 38.2 % and 51.7 % of the variance in the outcome variable. This implies a moderate to strong explanatory power of the model (Table 99).

Table 100: Classification Table

Classification Table[a]					
	Observed		Predicted		
			Remote_work		Percentage Correct
			Does Not Work Remotely	Works Remotely	
Step 1	Remote Work	Does Not Work Remotely	22	2	91.7
		Works Remotely	10	26	72.2
	Overall Percentage				80.0

[a] The cut value is .500.

Among employees who do not work remotely, 91.7 % were correctly classified by the model.

For those who work remotely, the model correctly predicted 72.2 % of cases. The overall classification accuracy of the model is 80.0 %, meaning that the model correctly classified the majority of all cases. These results indicate that the model performs well in distinguishing between employees who work remotely and those who do not (Table 100).

Table 101: Variables in the Equation

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Job Satisfaction	.638	.430	2.204	1	.138	1.893
	Sector	2.415	.970	6.195	1	.013	11.189
	Constant	−4.522	1.164	15.094	1	< .001	.011

[a] Variable(s) entered on step 1: job satisfaction, sector.

Table 101 shows the contribution of each predictor variable (*job satisfaction* and *sector*) in the logistic regression model. Job satisfaction has a coefficient (B) of 0.638, suggesting that for each one-unit increase in job satisfaction, the odds of working remotely increase by a factor of 1.893 (Exp(B)). However, the p-value (Sig. = 0.138) is greater than 0.05, meaning this predictor is not statistically significant at the 5 % level. The second independent variable sector has a coefficient of 2.415, with an odds ratio (Exp(B)) of 11.189. The p-value (0.013) is below 0.05, indicating this result is statistically significant. The constant (intercept) is −4.522, with a p-value of < 0.001, which is also significant. This represents the log odds of working remotely when all predictors are set to zero. In summary, sector is a significant predictor of remote work, while job satisfaction does not show a statistically significant effect in this model.

7 Discriminant Analysis

Discriminant analysis is used when we want to evaluate the likelihood of a statistical unit belonging to a specific group of statistical units (mutually exclusive groups), based on the values of statistical independent variables (numerical variables, known as discriminating variables), or when we are interested in whether it is possible to classify statistical units into two groups – discriminant analysis with two groups – or into multiple groups – multiple discriminant analysis (provided the conditions for conducting discriminant analysis are met).

Indicators Within Two-Group Discriminant Analysis

- **Eigenvalue of the discriminant function**: The proportion of explained variance (sum of squared differences between groups divided by the sum of squared differences within groups). A higher eigenvalue indicates a higher quality of the discriminant function.
- **Canonical correlation coefficient**: The correlation coefficient between estimated values and actual values of the dependent variable. A value close to 1 indicates a higher quality of the discriminant function.
- **Wilk's lambda and chi-square test**: If Wilk's lambda is close to 1, it suggests that the mean values of the discriminant function are equal in both groups (indicating poor quality of the discriminant function). A value close to 0 indicates low variability within both groups compared to the total variability of the variable (suggesting good quality of the discriminant function). The Chi-square test is used to test the hypothesis that

the mean values of the discriminant function are equal in both groups. If we can reject the hypothesis (at a low risk), it indicates the appropriate quality of the discriminant function.

- **Centroid**: The average value of the discriminant function for statistical units within each group.
- **Classification matrix**: Shows the number and percentage of correctly and incorrectly classified statistical units.

Task 1

We have data from a random sample of export-oriented companies in a certain economic sector. Open data file *Discriminantanalysis.sav*. The companies are divided into two groups (variable *EU*): those for which the majority of their exports are generated in the European Union (EU – 1) and those that create most of their exports in global markets outside the European Union (EU – 0).

We are interested in whether the following variables are significant for classifying companies into these two groups:

- What percentage of products sold in foreign markets are sold under your own brand? (Variable *vp2*: companies provided a percentage greater than 0 and up to 100 %).
- For the following variables, companies indicated their agreement with statements regarding market and technological changes in the market where the company operates (on a scale from 1 – strongly disagree, to 7 – strongly agree), variables *vp8a*, *vp8b*, vp8g, *vp8h*, *vp8i*:
  - The rate of technological change in this market is rapid.
  - Technological changes in this market represent significant opportunities.
  - Customers in this market are highly receptive to new products (services).
  - New customers in this market have needs for products (services) that differ from those of existing customers.
  - We operate in a market where customer preferences change very slowly.

Using discriminant analysis, examine whether the independent variables described above are significant for classifying companies into the two groups: those for which the majority of exports are generated in the EU markets (EU – 1) and those for which the majority of exports are generated in global markets outside the EU (EU – 0).
Procedure: Discriminant Analysis

Click on *Analyze*, then *Classify*, and select *Discriminant.* A dialog box opens; transfer the variable *EU* into the *Grouping variable* field and define the range values 0 and 1 using the *Define Range* button. Transfer all six independent variables into the *Independents* field. Select the *Statistics* button and choose *Means* under *Descriptive.* To display centroids graphically, select *Classify*, then choose *Separate-Groups* under the *Plots* section. For the classification table or matrix, select *Summary table* under the *Display section*, and click *OK.*

Answers and Output of Results

Table 102: Eigenvalues

Eigenvalues				
Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	**2.729**[a]	100.0	100.0	**.855**

[a] First 1 canonical discriminant functions were used in the analysis.

The high eigenvalue of 2.729 and the canonical correlation coefficient of 0.855 indicate that the quality of the discriminant function is high and that the included independent variables are significant for distinguishing between companies in both groups (Table 102).

This is also confirmed by the value of Wilks' lambda and the chi-square test in Table 103, which tests the hypothesis that the mean value of the discriminant function is equal in both groups. This would imply that the included independent variables are not significant for distinguishing between the groups.

Table 103: Wilks' Lambda

Wilks' Lambda				
Test of Function(s)	Wilks' Lambda	Chi-Square	df	Sig.
1	.268	35.533	6	**< .001**

Here, we test the following hypotheses:

H0: The mean value of the discriminant function is equal in both groups.

H1: The mean value of the discriminant function is not equal in both groups.

With a risk level of less than 5 %, we can reject the hypothesis that the mean value of the discriminant function is equal in both groups, confirming the quality of the discriminant function (Table 103).

Table 104: Standardized Canonical Discriminant – Function Coefficients

Standardized Canonical Discriminant Function Coefficients	
	Function
	1
vp2: What percentage of products sold in foreign markets are sold under your own brand?	**−.829**
vp8a: The rate of technological change in this market is rapid.	.014
vp8b: Technological changes in this market represent significant opportunities.	.507
vp8g: Customers in this market are highly receptive to new products (services).	.446
vp8h: New customers in this market have needs for products (services) that differ from those of existing customers.	−.611
vp8i: We operate in a market where customer preferences change very slowly.	.597

The greater the standardized discriminant coefficient of a variable, the more it contributes to distinguishing between the groups. The variable *vp2* contributes the most to distinguishing between the two groups. We can also write the discriminant function equation (Table 104):

$$\mathrm{D} = -0.829\ vp2 + 0.014\ vp8a + 0.507\ vp8b + 0.446\ vp8g - 0.611\ vp8h + 0.597\ vp8i$$

Table 105: Descriptive Statistics

Group Statistics					
European Union		Mean	Std. Deviation	Valid N (listwise)	
				Unweighted	Weighted
No	vp2: What percentage of products sold in foreign markets are sold under your own brand?	63.86	17.284	14	14.000
	vp8a: The rate of technological change in this market is rapid.	4.93	1.817	14	14.000
	vp8b: Technological changes in this market represent significant opportunities.	4.21	1.805	14	14.000
	vp8g: Customers in this market are highly receptive to new products (services).	4.50	1.743	14	14.000
	vp8h: New customers in this market have needs for products (services) that differ from those of existing customers.	4.29	1.204	14	14.000
	vp8i: We operate in a market where customer preferences change very slowly.	2.50	.760	14	14.000
Yes	vp2: What percentage of products sold in foreign markets are sold under your own brand?	32.43	13.375	18	18.000

Group Statistics					
	vp8a: The rate of technological change in this market is rapid.	4.56	1.790	18	18.000
	vp8b: Technological changes in this market represent significant opportunities.	4.78	1.517	18	18.000
	vp8g: Customers in this market are highly receptive to new products (services).	4.33	1.414	18	18.000
	vp8h: New customers in this market have needs for products (services) that differ from those of existing customers.	3.22	1.353	18	18.000
	vp8i: We operate in a market where customer preferences change very slowly.	4.44	1.617	18	18.000
Total	vp2: What percentage of products sold in foreign markets are sold under your own brand?	46.18	21.779	32	32.000
	vp8a: The rate of technological change in this market is rapid.	4.72	1.782	32	32.000
	vp8b: Technological changes in this market represent significant opportunities.	4.53	1.646	32	32.000
	vp8g: Customers in this market are highly receptive to new products (services).	4.41	1.542	32	32.000
	vp8h: New customers in this market have needs for products (services) that differ from those of existing customers.	3.69	1.378	32	32.000
	vp8i: We operate in a market where customer preferences change very slowly.	3.59	1.624	32	32.000

The results in Table 105 of the descriptive statistics also show that the average value of this variable in the group of companies with the majority of exports in EU markets is 32.4 %, while in the group of companies with the majority of exports in global markets outside the EU, the average is 63.9 %.

Table 106: Classification Results

Classification Results[a]					
		European Union	Predicted Group Membership		Total
			No	Yes	
Original	Count	No	14	0	14
		Yes	2	16	18
	%	No	100.0	.0	100.0
		Yes	11.1	88.9	100.0

[a] 93.8 % of original grouped cases correctly classified.

Table 106 presents the classification results based on the discriminant function. Among companies whose majority of exports are outside the EU ("No"), all 14 were correctly classified (100.0 %). Among companies that primarily export within the EU ("Yes"), 16 out of 18 were correctly classified, corresponding to 88.9 %. In total, 30 out of 32 companies (93.8 %) were correctly classified, which demonstrates that the discriminant function successfully distinguishes between the two groups (which also confirms the quality of the discriminant function).

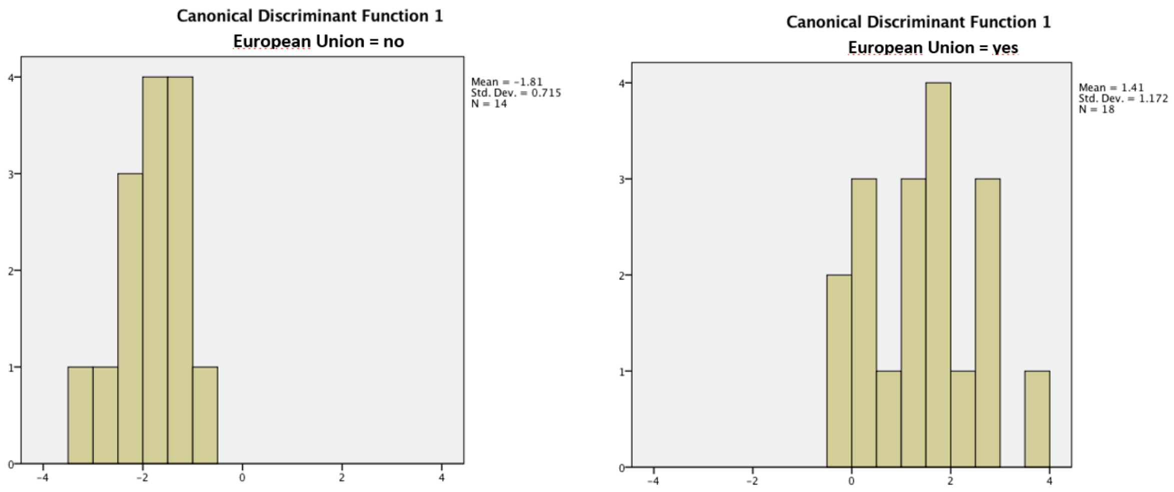


Figure 3: Frequency Distribution of Companies in Each Group

Table 107: Functions at Group Centroids

Functions at Group Centroids	
European Union	Function
	1
No	−1.814
Yes	1.411

Unstandardized canonical discriminant functions evaluated at group means.

In Table 107, centroids present the average value of the discriminant function for all cases within a particular group. In this case, the group centroid for respondents who answered *no* regarding the European Union is −1.814, whereas the centroid for those who answered *yes* is 1.411. These centroids indicate how far apart the groups are on the discriminant function. The greater the distance between centroids, the better the function discriminates between the groups. Classification is typically based on how close a case's discriminant score is to one of these group centroids.

Task 2.

We have data from a random sample of 180 companies operating in various industries. Open data file *Discriminant analysis_Sustainability.sav*. Companies are divided into two groups based on their strategic sustainability orientation (*SustainStrategy*): *SustainStrategy* – 1 (Companies that have included sustainability in their long-term business strategy), and *SustainStrategy* – 0 (Companies that have not incorporated sustainability into their strategic plans). We are interested in whether the following variables are significant for classifying companies into these two groups:

- *InnovationScore:* Innovation intensity measured by the number of new products/services launched in the last 3 years (scale from 0 to 10).
- *GreenInvest:* The company has invested in environmentally friendly technologies. *(1 – strongly disagree to 7 – strongly agree).*
- *CSRPolicy:* The company has a formal corporate social responsibility (CSR) policy. *(*from 1 – strongly disagree to 7 – strongly agree).
- *ConsumerPressure:* The company perceives growing pressure from consumers to operate more sustainably (from 1 – strongly disagree to 7 – strongly agree).
- *EmployeeInvolvement*: Employees are actively involved in sustainability initiatives (from 1 – strongly disagree to 7 – strongly agree).

Using discriminant analysis, examine whether the independent variables described above significantly differentiate between companies that have adopted a sustainability strategy (SustainStrategy – 1) and those that have not (SustainStrategy – 0). Identify which variables contribute most to explaining this distinction.

Answers and Output of Results

Table 108: Eigenvalues

Eigenvalues				
Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	41.430[a]	100.0	100.0	.988

[a] First 1 canonical discriminant functions were used in the analysis.

The eigenvalue of 41.430 indicates that the canonical discriminant function explains a large proportion of the variance between the two groups – companies with and without a sustainability strategy. A high eigenvalue suggests that the function has strong discriminative power. The canonical correlation coefficient of 0.988 shows a very strong relationship between the discriminant function and the grouping variable. Since this value is close to 1, it implies that the discriminant function is highly effective at distinguishing between the two types of companies based on the independent variables included in the analysis. Therefore, both the high eigenvalue and the high canonical correlation confirm the quality and significance of the discriminant function in this analysis (Table 108).

Table 109: Wilks' Lambda

Wilks' Lambda				
Test of Function(s)	Wilks' Lambda	Chi-Square	df	Sig.
1	.024	657.751	5	< .001

Based on the value of Wilks' lambda and the chi-square test, we can reject the null hypothesis and accept the alternative hypothesis (H1): "The mean value of the discriminant function is not equal in both groups." The very low value of Wilks' lambda ($\lambda = 0.024$) and the statistically significant chi-square value ($p < 0.001$) indicate that the discriminant function significantly differentiates between the groups. Therefore, we confirm the high quality and discriminative power of the model (Table 109).

Table 110: Standardized Canonical Discriminant Function Coefficients

Standardized Canonical Discriminant Function Coefficients	
	Function
	1
Innovation intensity measured by the number of new products/services.	−.924
The company has invested in environmentally friendly technologies.	.700
The company has a formal corporate social responsibility (CSR) policy.	1.206
The company perceives growing pressure from consumers to operate more sustainably.	.691
Employees are actively involved in sustainability initiatives.	−.521

The standardized canonical discriminant function coefficients indicate the relative contribution of each variable to the discriminant function. A higher absolute value of the coefficient suggests a greater contribution to distinguishing between the two groups of companies. Based on the results in Table 110, the variable that contributes most to distinguishing between the two groups is *The company has a formal corporate social responsibility (CSR) policy*. It has the highest standardized coefficient (1.206), indicating that this variable has the strongest discriminatory power between the two groups.

We can also write the discriminant function equation (Table 110):

$$\mathrm{D} = -0.924\ \mathit{InnovationScore} + 0.700\ \mathit{GreenInvest} + 1.206\ \mathit{CSRPolicy} + 0.691\ \mathit{ConsumerPressure} - 0.521\ \mathit{EmployeeInvolvement}$$

Table 111: Functions at Group Centroids

Functions at Group Centroids	
Strategic Sustainability Orientation	Function
	1
No	−6.843
Yes	5.987

Unstandardized canonical discriminant functions evaluated at group means.

Table 111 presents the group centroids, which are the average discriminant function scores for each group. The centroid for companies without a strategic sustainability orientation is −6.843, while for those with a strategic orientation, it is 5.987. The large distance between the centroids indicates a strong discrimination between the two groups. In practice, new cases can be classified into one of the two groups based on their discriminant score and proximity to one of these centroid values.

8 Time Series Analysis

Time series analysis is used when observations are made repeatedly over time periods. The most important factor is time, and the most important components of the time series are as follows (Render et al., 2018):

- **Trend (T)**: The gradual upward or downward movement of the data over time.
- **Seasonality (S)**: A pattern of the variable's fluctuation above or below the trend line that repeats at regular intervals.
- **Cycles (C)**: Patterns in annual data that occur every several years.
- **Random variations (R)**: "Blips" in the data caused by chance and unusual situations.

Time-series models attempt to predict the future by using historical data (Render et al., 2018). The assumption is that what happens in the future is a function of what had happened in the past. The goals of the analysis are to identify patterns in the sequence of numbers over time, to test the impact of one or more interventions and to forecast future patterns or to compare series of different kinds of events (Tabachnick and Fidell, 2014).

Time-series models use a series of past data to make a forecast. Forecasting methods include moving averages, exponential smoothing, trend projections and seasonal variations, together with auto-regressive integrated moving average (ARIMA). In this tutorial, special attention is given to the following forecasting methods based on time series analysis:

- **Simple Moving Average (SMA)**: The SMA forecast is the average of the variable's observations (data, actual values) over *m* time periods. With each passing period, the most recent time period's data are added to the sum of the previous *m* time periods' data, and the earliest time period is dropped. This tends to smooth out short-term irregularities in the data series.
- **Weighted Moving Average (WMA)**: The WMA forecast is the sum of the products between weights and the variable's observations in m periods.
- **Simple Exponential Smoothing (SES)**: The SES forecast is based not only on the last period's observations, but also on the last period's forecast and a smoothing constant.
- **Trend Projections**: Trend projections fit a trend line to a series of historical data points and extend it into the future. When the regressor (predictor, independent variable) is time in the simple regression model, a linear trend equation is obtained. Trend projections can use also quadratic, cubic, and exponential functions.
- **ARIMA**: The ARIMA (*p, d, q*) model estimates three types of terms, where *p, d* and *q* are integer, usually very small values (0 or 1 or 2). They are as follows (Tabachnick and Fidell, 2014):
  - Auto-regressive terms (*p*) describe the dependency among successive observations. p = 0 means that the element is not needed in the model, 1 means that an observation depends on one previous observation, and 2 means that an observation depends on two previous observations.
  - Trend terms (*d*) are needed to make a nonstationary times series stationary. For example, if d = 2, a model must be differenced twice. The first difference removes linear trend, and the second one removes quadratic trend.
  - Moving average terms (*q*) describe the persistence of a random shock from one observation to the next. If q = 2, an observation depends on two preceding random shocks.

The most appropriate forecasting method for a particular time series is selected regarding the accuracy measures:

- Mean Absolute Error (MAE) is the average of absolute errors for time periods for which these errors can be calculated.
- Mean Squared Error (MSE) is the average of the squared errors.
- Root Mean Squared Error (RMSE) is the standard deviation of the residuals.

– Mean Absolute Percent Error (MAPE) is the average of the absolute values of the errors expressed as percentages of the actual values.

The lower the values of the above-written accuracy measures, the more accurate the forecast.

Task 1

A pharmaceutical company has started selling a new drug. The file *TSA_sales* contains sales data in 1000 EUR for the first 100 days. Use SPSS to predict future sales.

a) Using simple regression analysis, form a linear trend equation and predict the sales value for the 101st day.
b) Predict the sales value for the 101st day by exponential smoothing.
c) Predict the sales value for the 101st day by ARIMA (1, 1, 0).
d) Predict the sales value for the 101st day by ARIMA (0, 1, 1) and answer the following questions:
   d1) Explain what the obtained values of the model fit parameters R-squared, RMSE, MAPE and MAE mean for this example.
   d2) What is the predicted sales value for the 101st day, obtained by ARIMA (0, 1, 1)?

Answers and Results

a) Simple Linear Regression Analysis

Procedure

Open the file *TSA_sales*. Click on *Analyze*, then *Regression*, and select *Linear*. A dialog box opens where you transfer the variable *sales in 1000 EUR* in the right box under *Dependent*, and the variable *time in days* to the right box under *Independent*. Click on *Method* and choose *Enter*. Click on *Statistics* and select *Estimates*, and *Model fit* under the *Regression Coefficient* window. Then click *Continue* and *OK*.

Results

Table 112: Linear Trend: Model Summary for Sales and Time

Model	R	R Square	Adjusted R Square	Standard Error of the Estimate
1	0.992[a]	0.985	0.985	1.966257

[a] Predictors: (Constant), Time in days.

When evaluating the accuracy of the obtained linear model with the results in Table 112, it can be concluded that the correlation coefficient R, which is 0.992, shows a strong correlation between sales and time, and the value of R^2 indicates that 98.5 % of the variability in sales is explained with the variability in time.

Table 113: Linear Trend: Analysis of Variance for Sales and Time

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	24603.835	1	24603.835	6363.887	< 0.001[b]
	Residual	378.884	98	3.866		
	Total	24982.720	99			

[a] Dependent variable: sales in 1000 EUR.

[b] Predictors: (constant), time in days.

The results of the F-test ($p < 0.001$, therefore also $p < 0.05$) in Table 113 show that the model is reliable. We can namely reject the null hypothesis: H0: $r^2_{xy} = 0$ and conclude that coefficient of determination is greater than 0 (H1: $r^2_{xy} > 0$). This indicates that there is a linear dependence between the independent variable and the dependent variable.

Table 114: Linear Trend: Coefficients for Sales Forecasting

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.109	0.396		2.799	0.006
	Time in Days	0.543	0.007	0.992	79.774	< 0.001

[a] Dependent variable: sales in 1000 EUR.

The results of the t-test presented in Table 114 let us report the following. As $p < 0.05$ at constant and the regression coefficient at the variable *time in days*, both parameters are reliable: the constant β_0 and the regression coefficient β_1 are different from zero; we accept the research hypotheses H1: $\beta_0 \neq 0$, H1: $\beta_1 \neq 0$. Therefore, the following linear function is written:

$$\hat{y}_t = 1.109 + 0.543t,$$

where $\hat{y}_t$ is sales in 1000 EUR, and *t* is time in days.

The sales forecast for the 101st day of sales ($t = 101$) is 55.952, i.e., 55,952 EUR.

b) Exponential Smoothing

Procedure

Click on *Analyze*, then *Forecasting*, and *Create traditional models*. Before we use the dialog box, we should define the starting time and time interval for our time series. Click *Define date and time*, in left box *Cases Are* select *Days*, and in the right box *First Case* write *1*, then click *OK*. Again, click on *Analyze*, then *Forecasting*, and *Create traditional models*. The window *Time Series Modeler* is open, and the button *Variables* is activated. In the right-hand box *Dependent Variables* write *sales in 1000 EUR*. We should not put the variables in the box *Independent Variables* if we want to conduct exponential smoothing. Under *Method*, select *Exponential Smoothing*. Click on *Criteria* and under *Model Type* select *Nonseasonal: Holt's linear trend* and click *Continue*. After clicking *Statistics*, choose *Display fit measures*, *Ljung-Box statistic*, and *number of outliers by model*. Under *Fit Measures*, select R *square*, *Root mean square error*, *Mean absolute percentage error* and *Mean absolute error*. Under *Statistics for comparing models*, select *Goodness of fit*, and under *Statistics for individual models* select *Parameter estimates*. Click on *Save* and in the table *Variables,* indicate *Predicted values*. Click on *Options*, and for *Forecast period* select *First case after end of estimation period through a specified date* and write *101* under *Day*. Then click *OK*.

Results

Table 115 shows the forecast for the 101st day obtained by exponential smoothing.

Table 115: Forecast: Exponential Smoothing

Model		101
Sales in 1000 EUR (Model_1)	Forecast	55.508
	UCL	59.176
	LCL	51.839

The predicted sales for the 101st day are 55.508 monetary units, that is 55,508 EUR (Table 115).

Table 116 that is included in the output file shows the selected model fit statistics and the results of the Ljung-Box test.

Table 116: Model Fit: Exponential Smoothing for Sales Forecasting

Model	Model Fit Statistics				Ljung-Box Q(18)		
	R-Squared	RMSE	MAPE	MAE	Statistics	DF	Sig.
Sales in 1000 EUR	0.987	1.849	13.744	1.092	15.724	16	0.472

In Table 116, the R^2 value shows that 98.7 % of variations in sales are explained by variations in time. The RMSE value, i.e., the standard deviation of the residuals, is 1.849. MAPE means that, on average, the absolute values of the errors present 13.7 % of the actual values. The MAE value 1.092 means that the average absolute error is 1.092. The results of the Ljung-Box test show that the null hypothesis stating that the data are independently distributed cannot be rejected ($p > 0.05$).

c) ARIMA (1, 1, 0)

Procedure and Results

Click on *Analyze*, then *Forecasting*, and *Create traditional models*. Before we use the dialog, we should define the starting time and time interval for our time series. Click *Define date and time*, in left box *Cases Are* select *Days*, and in the right box *First Case* write *1*, then click *OK*.

Afterward, draw a sequence plot for non-stationary data. Select *Analyze*, *Forecasting*, and *Sequence charts*. Transfer the variable *sales in 1000 EUR* to the window *Variables*, and newly created variable *day* to the window under *Time axis labels*. In addition, indicate *One chart per variable*. Click *OK*. The sequence chart is presented in Figure 4.

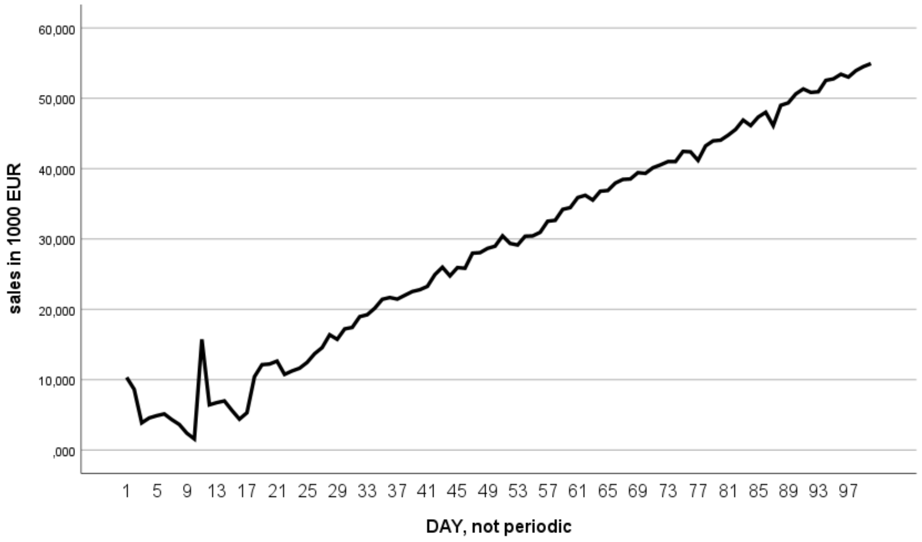


Figure 4: Sequence Plot for Non-Stationary Data of Sales

As trend is observed in Figure 4, return to the *Sequence charts* dialog and at *Difference*, put *d = 1* (the series is differenced once, and linear trend is removed). Click *OK*. The obtained sequence plot for stationary data is presented in Figure 5.

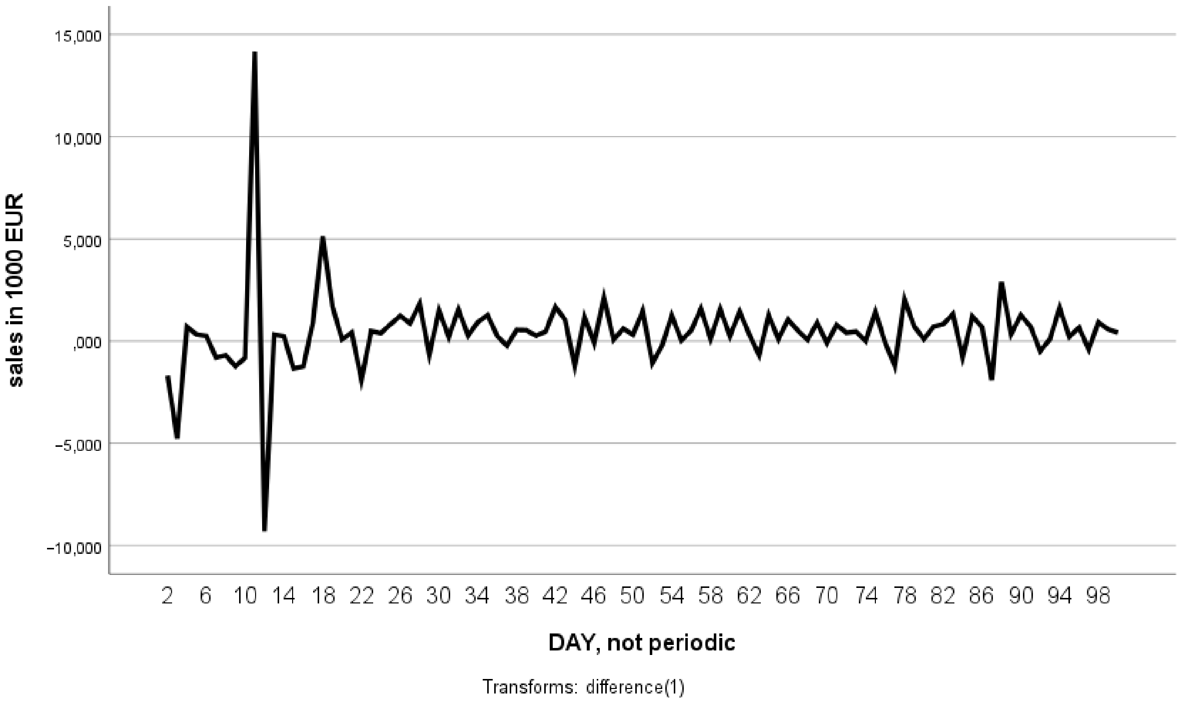


Figure 5: Sequence Plot for Stationary Data of Sales

Figure 5 shows that linear trend is removed.

Then examine autocorrelation and partial autocorrelation functions. Select *Analyze*, then *Forecasting*, and *Autocorrelations*. Transfer the variable *sales in 1000 EUR* to the window *Variables*, under *Transform* write *1* at *Difference*, and under *Display* indicate *Autocorrelations*. Table 117 presents the results of autocorrelation analysis for *sales*.

A statistical test to check if a time series contains autocorrelations is called Ljung-Box test. To examine if a time series contains autocorrelations, we check the following hypotheses:

H0: The data are independently distributed.

H1: The data are not independently distributed.

Table 117 shows that till lag 13, $p < 0.05$, the null hypothesis that the data are independently distributed, is rejected, and the research hypothesis that the data are not independently distributed is confirmed. They exhibit serial correlation.

Table 117: Autocorrelations for Sales

Series: Sales in 1000 EUR					
Lag	Autocorrelation	Std. Error[a]	Box-Ljung Statistic		
			Value	df	Sig.[b]
1	−0.354	0.099	12.782	1	< 0.001
2	−0.033	0.098	12.896	2	0.002
3	0.019	0.098	12.933	3	0.005
4	−0.059	0.097	13.298	4	0.010
5	−0.015	0.097	13.323	5	0.021
6	−0.062	0.096	13.737	6	0.033
7	0.206	0.096	18.337	7	0.011
8	−0.111	0.095	19.698	8	0.012
9	0.015	0.095	19.722	9	0.020
10	0.115	0.094	21.215	10	0.020
11	−0.116	0.094	22.742	11	0.019
12	0.028	0.093	22.834	12	0.029
13	0.023	0.093	22.897	13	0.043
14	0.016	0.092	22.926	14	0.061
15	−0.024	0.092	22.992	15	0.084
16	−0.061	0.091	23.439	16	0.102

[a] The underlying process assumed is independence (white noise).

[b] Based on the asymptotic chi-square approximation.

Again, click on *Analyze*, *Forecasting*, and *Create traditional models*, in the right box *Dependent Variables* put *sales in 1000 EUR*. Select *Method ARIMA* and click on *Criteria*. In the *Time Series Modeler: ARIMA Criteria*, write the following in the *Nonseasonal column: 1* at *Autoregressive (p)*, and *1* at *Difference (d)* and then *0* at *Moving average (q)*. Click *Continue* and *OK*.

Now click *Statistics* and select *Fit measures*, *statistics for comparing models* and *statistics for individual models* (indicate *Parameter estimates*), then indicate *Display forecast*. Click on *Plots* and select *Plots for individual models (Series, Residual autocorrelation function, Residual partial autocorrelation function)*, and at *Each plot displays* indicate *Observed values* and *Forecasts*. Click on *Save* and indicate *Save at predicted values*. Finally choose *Options* and under *Forecast period,* indicate *First case after end of estimation period through a specified date*, then click *OK*.

The model fit is presented in Table 118.

Table 118: Model Fit: ARIMA (1, 1, 0) for Sales and Time

Model	Model Fit statistics					Ljung-Box Q(18)		
	Stationary R-Squared	R-Squared	RMSE	MAPE	MAE	Statistics	DF	Sig.
Sales in 1000 EUR	0.125	0.985	1.926	10.705	1.003	13.124	17	0.728

Compare the results in Table 118 with the ones obtained by exponential smoothing in Table 116.

Table 119: ARIMA (1, 1, 0) Model Parameters for Sales and Time

		Estimate	SE	t	Sig.
Sales in 1000 EUR (Model_1)	Constant	0.456	0.143	3.187	0.002
	AR Lag1	−0.354	0.095	−3.738	< 0.001
	Difference	1			

Table 119 shows that the ARIMA model parameters are statistically significant ($p < 0.05$).

Table 120: Forecast: ARIMA (1, 1, 0) for Sales

Model		101
Sales in 1000 EUR (Model_1)	Forecast	55.397
	UCL	59.217
	LCL	51.578

With the ARIMA method predicted sales value for the 101st day, presented in Table 120, is 55.397 monetary units, i.e., 55,397 EUR.

d)

Procedure

Click on the file *TSA-sales*. Then click *Analyze*, *Forecasting*, and *Create traditional models*. Before we use the dialog, we should define the starting time and time interval for our time series. In the *Time series modeler*, click *Define date and time*, in left box *Cases Are* select *Days*, and in the right box *First Case* write *1*, then click *OK*.

Click on *Analyze*, *Forecasting*, and *Create traditional models*, in the right box *Dependent Variables* put *sales in 1000 EUR*. Select *Method ARIMA* and click on *Criteria*. In the *Time Series Modeler: ARIMA Criteria*, write the following in the *Nonseasonal column: 0* at *Autoregressive (p)*, and *1* at *Difference (d)* and then *1* at *Moving average (q)*. Click *Continue* and *OK*.

Now click *Statistics* and select *Fit measures* (indicate *R square*, *Root mean square error*, *Mean absolute percentage error*, and *Mean absolute value*), *Statistics for comparing models* (indicate *Goodness of fit*) and *Statistics for individual models* (indicate *Parameter estimates*), then indicate *Display forecast*. Click on *Plots* and select *Plots for individual models* (*Series*, *Residual autocorrelation function*, *Residual partial autocorrelation function*), and at *Each plot displays* indicate *Observed values* and *Forecasts*. Click on *Save* and indicate *Save at predicted values*. Finally, click *Options*, under

Forecast period indicate *First case after end of estimation period through last case in active dataset*, then click *OK*.

Results

d1)

Table 121: Model Fit: ARIMA (0, 1, 1) for Sales and Time

Model	Model Fit Statistics			
	R-Squared	RMSE	MAPE	MAE
Sales in 1000 EUR (Model_1)	0.986	1.867	13.722	1.097

d2)

Table 122: Forecast: ARIMA (0, 1, 1) for Sales

Model		101
Sales in 1000 EUR (Model_1)	Forecast	55.416
	UCL	59.094
	LCL	51.737

Task 2

In the last 6 months of the year, the following transshipment volumes, expressed in million tons, were recorded in a port: 1.9; 2.0; 2.2; 2.1; 2.3; 2.5.

a) Use simple moving averages (m = 3), weighted moving averages ($w_{t-1} = 0.5$, $w_{t-2} = 0.3$, $w_{t-3} = 0.2$) to predict the transshipment volumes for the period from the 4th to the 7th month. After, use simple exponential smoothing ($\alpha = 0.3$) to predict the transshipment volumes for the period from the 2nd to the 7th month.

b) Calculate the MAE, MSE and MAPE to select the most appropriate forecasting method and interpret the obtained results. Which of the forecasting methods should be used?

c) By using SPSS, perform a linear trend analysis to predict the transshipment volumes for the 7th month.

a) Procedure

Simple Moving Averages

$$F_{43} = (1.9 + 2 + 2.2) / 3 = 2.03 \doteq 2.0$$

$$F_{53} = (2 + 2.2 + 2.1) / 3 = 2.1$$

By sliding along the time series, calculate F_{63} and F_{73}.

Weighted Moving Averages

$$t = 4\text{: } w_{4-1} = w_3 = 0.5, w_{4-2} = w_2 = 0.3, w_{4-3} = w_1 = 0.2$$

$$F_{43} = 0.2 * 1.9 + 0.3 * 2 + 0.5 * 2.2 = 2.08 \doteq 2.1$$

$$t = 5\text{: } w_{5-1} = w_4 = 0.5, w_{5-2} = w_3 = 0.3, w_{5-3} = w_2 = 0.2$$

$$F_{53} = 0.2 * 2 + 0.3 * 2.2 + 0.5 * 2.1 = 2.11 \doteq 2.1$$

By sliding along the time series, calculate F_{63} and F_{73}.

Simple Exponential Smoothing

$$F_2 = 0.3 * y_1 + (1 - 0.3) * F_1$$

To calculate F_2, let us assume that $F_1 = y_1 = 1.9$, therefore:

$$F_2 = 0.3 * y_1 + (1 - 0.3) * y_1 = y_1 = 1.9$$

$$F_3 = 0.3 * y_2 + (1 - 0.3) * F_2 = 0.3 * 2 + 0.7 * 1.9 = 1.93 \doteq 1.9$$

Use the unrounded value F_3 for calculating F_4.

$$F_4 = 0.3 * y_3 + (1 - 0.3) * F_3 = 0.3 * 2.2 + 0.7 * 1.93 = 2.011 \doteq 2.0$$

By sliding along the time series, calculate F_5, F_6 and F_7.

Results

Table 123: Forecasts Obtained by Simple Moving Averages, Weighted Moving Averages, and Simple Exponential Smoothing

Month	Actual Value	Predicted Value – SMA	Predicted Value – WMA	Predicted Value – SES
1	1.9	-	-	-
2	2	-	-	1.9
3	2.2	-	-	1.9
4	2.1	2.0	2.1	2.0
5	2.3	2.1	2.1	2.0
6	2.5	2.2	2.2	2.1
7		2.3	2.4	2.2

b) Procedure and Results

First calculate the errors in predicted values of transshipment volumes, forecasted by simple moving average. The results are presented in Table 124.

Table 124: Errors in Predicted Values of Transshipment Volumes, Forecasted by Simple Moving Average

Month	y_t	F_t	$y_t - F_t$	$\lvert y_t - F_t \rvert$	$(y_t - F_t)^2$
1	1.9	-	-	-	-
2	2	-	-	-	-
3	2.2	-	-	-	-
4	2.1	2.0	0.1	0.1	0.01
5	2.3	2.1	0.2	0.2	0.04
6	2.5	2.2	0.3	0.3	0.09
7		2.3	-	-	
Sum	13.0		0.6	0.6	0.14

Similarly, calculate the errors of transshipment volumes, forecasted by weighted moving average and simple exponential smoothing. The sum of absolute errors of predictions by weighted moving average is 0.5, and the sum of absolute errors of predictions by simple exponential smoothing is 1.2.

Then calculate the values of accuracy measures for predicted values of transshipment volumes, forecasted by simple moving average.

$$MAE_{SMA} = 0.6 / 3 = 0.2$$

Similarly, calculate the value of MAE for predicted values of transshipment volumes, forecasted by weighted moving average (0.167) and by simple exponential smoothing (1.2 / 5 = 0.24).

$$\min (0.2, 0.167, 0.24) = 0.167$$

Considering MAE, the most accurate method for forecasting transshipment volumes is weighted moving average. The average of absolute errors is 0.167 million tons.

$$\mathrm{MSE}_{\mathrm{SMA}} = 0.14 / 3 = 0.047$$

Similarly, calculate the value of MSE for predicted values of transshipment volumes, forecasted by weighted moving average (0.043) and by simple exponential smoothing (0.072).

Determine the minimal value of MSE.

Considering MSE, the most accurate method for forecasting transshipment volumes is weighted moving average. The average of squared errors is 0.043 (million tons)^2.

$$\mathrm{MAPE}_{\mathrm{SMA}} = ((0.6 / 13) / 3) * 100 = 1.54\ \%$$

Similarly, calculate the value of MAPE for predicted values of transshipment volumes, forecasted by weighted moving average (1.28 %) and simple exponential smoothing (1.85 %).

Determine the minimal value of MAPE.

Considering MAPE, the most accurate method for forecasting transshipment volumes is weighted moving average. The average of absolute errors represents 1.28 % of the actual transshipment volumes.

c) Procedure and Results

For the data entry procedure, follow guidelines in the first chapter, and for the regression analysis procedure, follow guidelines in the fifth chapter and in Task 1a) of this chapter. The results of regression analysis obtained by SPSS are presented in Tables 125–127.

Table 125: Linear Trend: Model Summary for Transshipment Volume and Time

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	0.940[a]	0.884	0.855	0.0822

[a] Predictors: (constant), time.

Table 126: Linear Trend: Analysis of Variance for Transshipment Volume and Time

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	0.206	1	0.206	30.507	0.005[b]
	Residual	0.027	4	0.007		
	Total	0.233	5			

[a] Dependent variable: transshipment volume.

[b] Predictors: (constant), time.

Table 127: Linear Trend: Coefficients for Transshipment Volume Forecasting

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.787	0.077		23.339	< 0.001
	Time	0.109	0.020	0.940	5.523	0.005

[a] Dependent variable: transshipment volume.

The results in Table 125 let us report that the obtained regression model is acceptably accurate. Support this finding by explaining the meaning of the correlation coefficient, regression coefficient, and standard error of the estimate (Table 125). This finding is reliable, as supported by the results of the F-test (Table 126). The constant and the regression coefficient at time are also reliable, as supported by the results of t-test (Table 127). Write down the linear trend equation, explain the meaning of dependent and independent variables, insert the sequential number of the month for which you want to forecast transshipment volume, calculate it, and interpret the obtained result.

9 Monte Carlo Simulation

In many real-world applications, values of inputs to predictive models are uncertain. Simulation allows us to account for uncertainty in the inputs to predictive models and evaluate the likelihood of various outcomes of the model in the presence of that uncertainty.

The Monte Carlo simulation aims to model the probability of different outcomes in a process that cannot easily be predicted due to the intervention of random variables. Using this method, values for the variables are generated. The procedure is based on the probabilistic elements through random sampling (Render et al., 2018). The procedure starts with establishing probability distributions for important input variables, continues with building cumulative probability distributions, establishing an interval of random numbers, generating random numbers, and finishes with simulating a series of trials. Probability distributions can be either empirical or based on several patterns, e.g., the normal, uniform, gamma, lognormal, Weibull, exponential, or beta ones.

Table 128 summarizes those fundamental characteristics of probability distributions that are essential for defining input variables for simulation in SPSS.

Table 128: Characteristics of Probability Distributions for Defining Input Variables for Simulation in SPSS

Distribution	Values	Parameters	Additional Specifics
Normal	Real	Location parameter: μ, scale parameter: σ	$\sigma > 0$ The distribution has mean μ and standard deviation σ.
Exponential	$x \geq 0$	Scale parameter: λ	$\lambda > 0$ The mean of the distribution is $1/\lambda$.
Uniform	$a < x < b$	Minimal value: a, maximal value: b	The mean of the distribution is (b – a) / 2.
Lognormal	$x \geq 0$	μ, σ	$\mu > 0, \sigma > 0$
Beta	$0 < x < 1$	Shape parameters: α_1 and α_2	$\alpha_1 > 0, \alpha_2 > 0$ The mean of the distribution is $\alpha_1/(\alpha_1 + \alpha_2)$.
Gamma	$x \geq 0$	Shape parameter: α, scale parameter: λ	$\alpha > 0, \lambda > 0$ The mean of the distribution is α/λ.
Weibull	$x \geq 0$	Shape parameter: α, scale parameter: λ	$\alpha > 0, \lambda > 0$

Source: Summarized and adapted from IBM SPSS Statistics (2022).

Task 1

A large manufacturing company wants to assess its weekly productivity. Productivity is defined as the quotient between output and input:

$$productivity = \frac{output}{input}.$$

The input is 15,000 working hours, and the output is normally distributed random variable with a mean 20,000 pieces of product X, standard deviation 500 pieces, min 11,000 pieces and max 25,000 pieces. They want to conduct the Monte Carlo simulation.

a) Design the Monte Carlo simulation for measuring the weekly productivity with SPSS. Determine the average weekly productivity.

b) Make sensitivity analysis so that the input is 18,000 working hours. Create cumulative distribution function for productivity.

a) Procedure and Results: Type in the Equations

To run a Monte Carlo simulation in SPSS, any SPSS input file must be open. Click on *Analyze*, then select *Simulation*. In the box *Simulation: Model source,* select *Type in the equations*, then click *Continue*. In the *Simulation builder*, select *Model*, indicate *Type in the equations for the model*, then select *New equations*. In the *Equation editor*, write *productivity* under *Target*. Then click *New* and define variables that are not defined in the file, i.e., in the *Defined inputs*, write word *income*, and indicate *Fixed value input*, Type is *Numeric*, whereas *Default value* is *15000*. Also, write name *output*, indicate *Input to be simulated*; *Measurement* should be *Continuous*. In the *Equation editor*, use the arrows to transfer the input variables and appropriate symbols to obtain *output / input* in the *Numeric expression* box, and click *Continue*. In the *Simulation builder*, select *Simulation*, select an item *Simulated Fields*, and then determine *Normal distribution* for output, with mean 20000, standard deviation 500, min 11000 and max 25000. In *Select an item*, select advanced options, then *Output* and determine display formats. In the dialog box *Save* choose both *Save the simulation plan* and *Save the simulated data*, *Browse*, determine the file name and click *Run*.

Table 129 presents the simulated values for the first 10 out of 1,000 cases.

Table 129: Part of the Completed Input File for Productivity Simulation

	Output	Input	Productivity
1	19345.30	15000.00	1.29
2	20495.81	15000.00	1.37
3	19209.58	15000.00	1.28
4	20321.43	15000.00	1.35
5	20475.04	15000.00	1.37
6	20358.91	15000.00	1.36
7	19120.33	15000.00	1.27
8	19308.08	15000.00	1.29
9	20454.55	15000.00	1.36
10	19704.55	15000.00	1.31

The values for the first case written in the first row in Table 129 can be interpreted as follows: when the simulated value of output is 19,345.30 and the fixed value of input is 15,000.00, the resulting productivity is 1.29. The obtained values in the completed input table thus help us understand productivity depending on various simulated output values.

The results in Table 130 from the output file show the basic descriptive statistics.

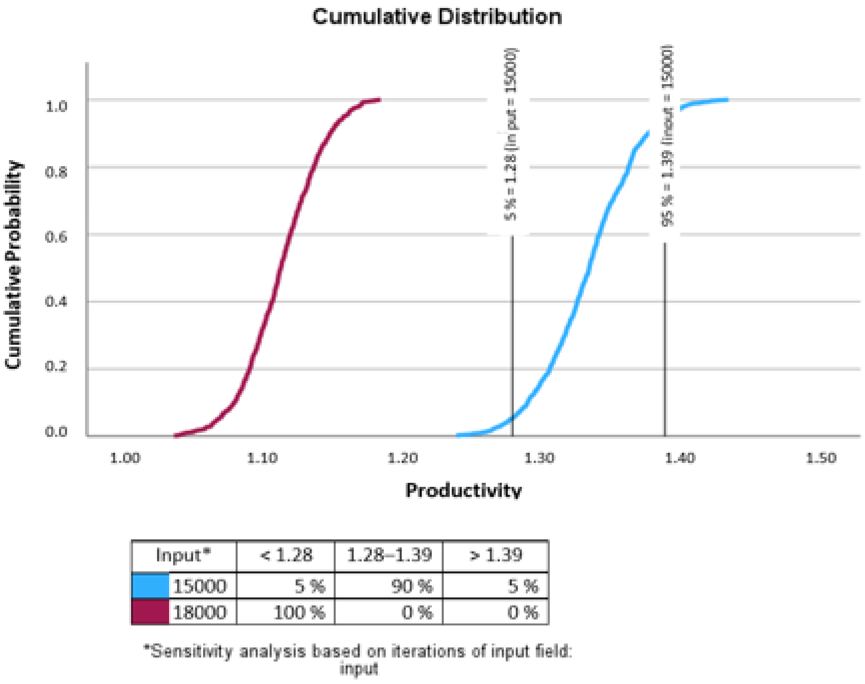
Table 130: Descriptive Statistics About Productivity

	Mean	Standard Deviation	Median	Minimum	Maximum
Productivity	1.333	0.034	1.333	1.24	1.44

Table 130 shows that the mean weekly productivity is 1.33, ranging from 1.24 to 1.44.

b) Procedure and Results: Sensitivity Analysis

For sensitivity analysis, in *Simulation: Model source*, open an existing simulation plan. Check that *Open* in *Simulation builder* is NOT indicated. Select the *Simulation tab*, click on *input field to simulate* on the left side, and click *Sensitivity analysis* on the right side. In the sensitivity analysis dialog box, indicate *Iterate* and write another value for income: *18000*. Then click *Continue*. Select the *output tab*, under *Density Functions* select the *cumulative distribution function*. It is presented in Figure 6.



Input*	< 1.28	1.28–1.39	> 1.39
🟦 15000	5 %	90 %	5 %
🟥 18000	100 %	0 %	0 %

*Sensitivity analysis based on iterations of input field: input

Figure 6: Sensitivity Analysis, Cumulative Function for Productivity

Task 2

A pharmaceutical company uses a group of machines to produce medicine. The annual fixed costs of this group of machines amount to 200,000 EUR, the price of the medicine is 0.8 EUR per piece, and the variable costs are 0.12 EUR per piece. To determine the capacity of this group of machines, a cost-profit model is used. This model enables the estimation of the quantity of medicine the company must produce to cover the target profit as well as the variable and fixed costs:

$$production\ quantity = \frac{target\ profit + fixed\ costs}{price - variable\ costs\ per\ unit}.$$

The annual profit is a random variable, uniformly distributed, with a minimum value of 1 million EUR and a maximum value of 4 million EUR.

a) Design a Monte Carlo simulation to determine the production quantity of medicine.
b) Explain the results obtained in the input table for the 3,762nd case and determine the average production quantity of medicine.

a) Procedure

Consider that any SPSS input file must be open if we want to run a Monte Carlo simulation in it. Click on *Analyze*, then select *Simulation*. In the box *Simulation: Model source,* select *Type in the equations*, then click *Continue*. In the *Simulation builder*, select *Model*, indicate *Type in the equations for the model*, then select *New equations*. In the *Equation editor*, write *production_quantity* under *Target*. Then click *New* and define variables that are not defined in the file, i.e., in the *Defined inputs*, write *target_profit*, indicate *Input to be simulated*; *Measurement* should be *Continuous*. Again, click *New* and in the *Defined inputs* write name *fixed_costs*, and indicate *Fixed value input*, Type is *Numeric*, whereas *Default value* is *200000*. Similarly, we determine the variables *price* and *variable_costs_per_unit*. As *production_quantity* is already written in the window under *Target*, use the arrows to transfer the input variables and appropriate symbols to obtain (*target_profit* + *fixed costs*) / (*price* − *varible_costs_per_unit*) in *Numeric expression* (i.e., in the right hand-side of the equation), and click *Continue*. In the *Simulation builder*, select *Simulation*, select an item *Simulated Fields*, and then determine *Uniform distribution* for target profit, with minimal value 1000000 and maximal value 4000000. In *Select an item*, select advanced options, then *Output* and determine display formats. In the dialog box *Save* choose both *Save the simulation plan* and *Save the simulated data*, *Browse*, determine the file name and click *Run*.

Table 131 presents the characteristics of input variables.

Table 131: Input Variables for Cost-Profit Model

Input Distributions			Parameter Value
Fixed_Costs	Fixed input	Value	200,000.00
Price	Fixed input	Value	0.80
Target_Profit	Uniform	min	1,000,000.00
		max	4,000,000.00
Variable_Costs_per_Unit	Fixed input	Value	0.12

b) Results

Table 132 shows the production quantity obtained for 10 simulated values (from 3,755 to 3,764) out of 3,894, namely of the target profit, which is distributed according to a uniform distribution, as well as the values of fixed costs, prices, and variable costs per unit, as specified in the task description and evident from table 131.

Table 132: Part of the Input File for the Simulation With the Cost-Profit Model

	Target Profit	Fixed Costs	Price	Variable Costs per Unit	Production Quantity
3755	2986336.77	200000.00	0.80	0.12	4685789
3756	2837036.60	200000.00	0.80	0.12	4466230
3757	3074820.74	200000.00	0.80	0.12	4815913
3758	2097348.41	200000.00	0.80	0.12	3378454
3759	2929109.11	200000.00	0.80	0.12	4601631
3760	1396122.10	200000.00	0.80	0.12	2347238
3761	1762283.38	200000.00	0.80	0.12	2885711
3762	3592961.29	200000.00	0.80	0.12	5577884
3763	2599029.66	200000.00	0.80	0.12	4116220
3764	1913729.35	200000.00	0.80	0.12	3108426

If the pharmaceutical company wishes to achieve a target profit of 3,592,961 EUR, with fixed costs 200,000 EUR, a price of 0.80 EUR per piece, and variable costs of 0.12 EUR per piece, they must produce 5,577,884 pieces of medicine annually (Table 132).

The results in Table 133 from the output file show the basic descriptive statistics for production quantity.

Table 133: Descriptive Statistics About Production Quantity

	Mean	Standard Deviation	Median	Minimum	Maximum
Production_Quantity	3973121	1264778	3963757	1764866	6176363

Table 133 shows that as a result of simulations based on the cost-profit model, the average production quantity of medicine per year is 3,973,121 pieces.

Task 3

In a medium-sized Slovenian municipality, there are 3000 companies. Using systematic sampling, select a sample of 150 companies.

a) Calculate and explain the sample percentage.
b) Determine the first and all subsequent companies in sample.

Procedure and Results

a)

N = 3000

n = 150

Sample proportion: $n / N = 150 / 3000 = 1 / 20 = 0.05, 0.05 * 100 = 5\,\%$, which means that every 20th company (i.e., 5 % of companies) in this municipality is included in the sample.

b)

The first selected number is a randomly selected number from 1 to 20 (e.g., we selected number 7, which is the 7th company in the database of companies in this municipality). First selected number: 7, followed by $27\ (7 + 20), 47, 67, 87, 107, 127, 147, \ldots, 2987\ (7 + 149 * 7)$.

References

Agresti, A., Finlay, B. (2009). *Statistical Methods for the Social Sciences*. Pearson: Prentice Hall.

Artenjak, J. (2003). *Poslovna statistika, Prenovljena in dopolnjena izdaja*. Maribor: UM Ekonomsko-poslovna fakulteta.

Bastič, M. (2006). *Metode raziskovanja*. Maribor: UM Ekonomsko-poslovna fakulteta.

Burns, R. B., Burns, R. A. (2008). *Business research methods and statistics using SPSS*. SAGE Publications.

Corder, G. W., Foreman, D. I. (2014). *Nonparametric statistics for non-statisticians: A step-by-step approach*. New Jersey, CA: Wiley.

Fabrigar, L. R., Wegener, D. T. (2011). *Exploratory factor analysis*. Oxford University Press.

Field, A. (2017). *Discovering statistics using IBM SPSS Statistics: North American edition*. (5th ed.). SAGE Publications Ltd.

Frost, J. (2019). *Introduction to Statistics: An Intuitive Guide for Analyzing Data and Unlocking Discoveries*. Statistics By Jim Publishing, USA

Goos, P., Meintrup, D. (2015). *Statistics with JMP: Graphs, Descriptive Statistics and Probability*. New Jersey, CA: Wiley.

Gorsuch, R. L. (2014). *Factor analysis: Classic edition*. Psychology Press.

Gravetter, F. J., Wallnau, L. B. (2017). *Statistics for the Behavioral Sciences*. Cengage Learning.

IBM SPSS Statistics (2022). Random variable and distribution function – IBM Documentation. Available August 7, 2025, at: https://www.ibm.com/docs/en/spss-statistics/cd?topic=expressions-random-variable-distribution-functions

IBM, (2024). Binary Logistic Regression, Available November 9, 2024 at: https://www.ibm.com/docs/en/spss-statistics/beta?topic=regression-binary-logistic

Kline, R. B. (2023). *Principles and practice of structural equation modeling*. Guilford Press.

Kutner, M. H., Nachtsheim, C. J., Neter, J. (2004). *Applied Linear Regression Models*. McGraw-Hill Irwin.

Lehenbauer, K. D. (2022). *Introduction to Business Statistics: A Simple Stepwise Approach to Basic Statistics*. Publisher: Analytics TX, LLC

Pham-Gia, T. (2022). *The multivariate normal distribution: Theory and applications*. New Jersey: World Scientific.

Render, B., Stair, R. M., Hanna, M. E., & Hale, T. S. (2018). *Quantitative Analysis For Management* (13th ed.). Harlow: Pearson.

Tabachnick, B. G., Fidell, L. S. (2019). *Using multivariate statistics*. Pearson: Boston

Tabachnick, B.G., & Fidell, L.S. (2014). *Using Multivariate Statistics* (6th Edition). Harlow: Pearson Education.

Tabachnick, B.G., Fidell, L.S. (2013). *Using multivariate statistics*. Pearson: Boston, Columbus etc.

Tavakol M, Wetzel A. (2020). Factor Analysis: a means for theory and instrument development in support of construct validity. *International Journal of Medical Education*, 6(11), pp. 245-247. doi: 10.5116/ijme.5f96.0f4a.

Tominc, P., Kramberger, T. (2007). *Statistične metode v logistiki*. Celje: UM Fakulteta za logistiko.

Triola, M. F. (2022). *Elementary Statistics*. Pearson.

UCLA. (2024). Statistical Methods and Data Analytics, Available November 8, 2024 at: https://stats.oarc.ucla.edu

Uhm, T., Yi, S. (2023). *A comparison of normality testing methods by empirical power and distribution of P-values*. Communications in Statistics - Simulation and Computation, 52(9), pp. 4445–4458. https://doi.org/10.1080/03610918.2021.1963450

Wagner, W. E. (2019). *Using IBM® SPSS® Statistics for research methods and social science statistics*. SAGE Publications.

Weiss, N. A. (2021). *Introductory Statistics*. Pearson.

Research Methods – Data Analysis Techniques

Polona Tominc, Vesna Čančer, Maja Rožman
University of Maribor, Faculty of Economics and Business, Maribor, Slovenia
maja.rozman1@um.si, polona.tominc@um.si, vesna.cancer@um.si

The tutorial *Research Methods – Data Analysis Techniques* is intended for master's students of the programme Economic and Business Sciences, specialization Data Science in Business. It provides a systematic and practice-oriented overview of quantitative methods and statistical analyses using the SPSS software. The material connects theoretical foundations with real-world business examples, encouraging the development of data literacy and analytical reasoning. It covers topics such as descriptive statistics, sampling, normal distribution, parametric and nonparametric tests, regression and factor analysis, time series analysis, discriminant analysis, and Monte Carlo simulation. The publication equips students with practical research and analytical skills essential for understanding contemporary data-driven challenges and for making evidence-based decisions in business environments.

DOI
https://doi.org/
10.18690/um.epf.1.2026

ISBN
978-961-299-098-5

Ključne besede:
statistical methods,
descriptive statistics,
univariate statistics,
multivariate statistics,
data science

Univerzitetna založba
Univerze v Mariboru

University of Maribor

Faculty of Economics and Business

Gradivo je vsebinsko in zahtevnostno povsem primerno za raven podiplomskega študija. Obravnava napredne statistične metode in analitične pristope, ki presegajo osnovno raven ter zahtevajo kritično razmišljanje in samostojno delo študentov.

Blaž Frešer
Univerza v Mariboru

Vsebina publikacije je napisana jasno, razumljivo in primerno ciljni skupini. Pojmi in statistične metode so razloženi s strokovno natančnostjo, hkrati pa so razlage dovolj enostavne, da jih lahko razumejo študenti brez poglobljenega predznanja. Posebej pomembno je, da je teoretični del vedno dopolnjen s praktičnimi primeri in nalogami v programu SPSS, kar bistveno olajša razumevanje in uporabo metod v praksi.

Dijana Oreški
Univerza v Zagrebu