REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

الجمهورية الجزائرية الديمقراطية الشعبية

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR
ET DE LA RECHERCHE SCIENTIFIQUE

وزارة التعليم العالي والبحث العلمي

ECOLE SUPERIEURE EN SCIENCES APPLIQUEES
-TLEMCEN-

المدرسة العليا في العلوم التطبيقية
-تلمسان-

Higher School of Applied Sciences of Tlemcen

Second-cycle department

LECTURE HANDOUT

Data Analysis

Prepared by:
Dr Imane NEDJAR

Preface

In today's data-driven world, the ability to harness the power of data analysis is a vital skill for professionals across various industries. This lecture handout is designed to serve as a comprehensive guide to data analysis, with a particular focus on its applications in industrial contexts.

Throughout this lecture handout, we strive to offer a balanced blend of theory and practical applications. Real-world examples and case studies drawn from industrial settings will illustrate how data analysis techniques can be applied to solve complex problems and drive informed decision-making.

We hope that this lecture handout will serve as a valuable resource for students aiming to become proficient data analysts in the industrial landscape.

Contents

List of Figures

List of Tables

INTRODUCTION

In today's industrial landscape, the mode of production is characterized by larger firm sizes, substantial investments, mass production, and assembly-line manufacturing. This phase is also marked by the evolution of networked technologies, including railways, pipelines, electricity, and telephones. The increase in positive externalities resulting from advancements in transportation and communication, coupled with the growing integration of science into business operations, akin to the industrial revolution of the previous century.

In the industrial sphere, there is an unprecedented generation of vast amounts of data. This raw data emanates from various domains, encompassing product lifecycle management, design, assembly processes, and quality control. It can also be sourced from equipment such as engines, compressors, or conveyors, as well as from external partners, suppliers, or customers.

In the face of this deluge of data, the recognition of the importance of extracting insights from it has taken center stage in the industry. This realization has propelled the development and adoption of data analysis techniques within the industrial landscape.

This lecture handout, designed for third-year industrial engineering students, aims to provide a comprehensive introduction to data analysis, with a focus on essential concepts and methods.

The structure of this handbook is designed to provide a comprehensive journey through the world of data analysis, catering to both beginners and seasoned practitioners:

Chapter 1: Fundamental Concepts of Data Analysis

In this opening chapter, we lay the foundation by introducing the fundamental concepts and principles of data analysis in an industrial context.

Chapter 2: Univariate Analysis

The second chapter takes a deep dive into univariate analysis. Students will explore various unidimensional measures and learn how to leverage them for effective numerical operations.

Chapter 3: Bivariate Analysis

Our third course focuses on bivariate analysis. Here, students will gain valuable insights into utilizing and computing different types of correlations between variables.

Chapter 4: Principal Component Analysis (PCA)

In this fourth chapter, students will delve into the world of Principal Component Analysis (PCA), a powerful method for examining relationships between multiple variables and reducing dimensionality.

Chapter 5: Multiple Correspondence Analysis (MCA)

The fifth course introduces Correspondence Analysis and Multiple Correspondence Analysis (MCA). Students will discover how these techniques unveil the underlying structures within multidimensional data.

Chapter 6: Classification in Machine Learning

Chapter 6 delves into the foundational approaches used in machine learning classification. Students will gain a solid understanding of classification techniques and their practical applications.

Chapter 7: Clustering in Machine Learning

Our final chapter, Chapter 7, serves as a comprehensive introduction to clustering in machine learning. It covers the core concepts of clustering, explores different clustering tasks, and provides insights into various types of clustering algorithms.

CHAPTER 1

FUNDAMENTAL CONCEPTS OF DATA ANALYSIS

Data analysis encompasses a variety of methods and techniques designed to extract actionable knowledge from raw data. Whether in the realms of industry, commerce, scientific research, or other domains, data analysis plays a pivotal role in transforming data into actionable information. This equips stakeholders with the capacity to make informed decisions and address complex challenges.

In this chapter, we will introduce the fundamental concepts and principles of data analysis in an industrial context.

1.1 Definition of Data Analysis

Data analysis originates from the realm of statistics and focuses on jointly describing data. Its purpose is to provide a more concise description of the fundamental information within the data, facilitating classification, description, and analysis.

The primary goal of data analysis is to extract valuable information from raw data and convert it into actionable knowledge, ultimately improving comprehension and supporting decision-making.

Data analysis plays a crucial role in various fields, including business, science, and technology. It involves the systematic examination of data sets, whether they are numerical, categorical, or textual, to identify patterns, relationships, and trends. This process often requires the use of statistical techniques, data visualization tools, and advanced software to uncover meaningful insights from the data.

1.2 Understanding the Relationship Between Data Analysis (DA), Artificial Intelligence (AI), Machine Learning (ML), and Data Mining (DM)

According to the American mathematician and statistician John Wilder Tukey, data analysis procedures include techniques for interpreting their results, methods for planning data collection to enhance analysis precision or accuracy, and all the tools and findings of mathematical statistics relevant to analysis (Tukey, 1962). Data analysis shares connections with Data Mining, ML, and AI. In this section, We will provide a summary of each concept separately and then explore the relationships between them.

Data Mining as a subdomain of AI, involves the process of analyzing vast amounts of information to uncover trends and patterns. In this context, generating knowledge means discovering new and non-trivial patterns, relations, and trends in data that are valuable to the user. The Data Mining process includes data collection and selection, data pre-processing, data analysis with result visualization, interpretation of findings, and the application of knowledge (Schuha et al., 2019).

Machine Learning is a natural outgrowth of the intersection of Computer Science and Statistics. It focuses on getting computers to program themselves from experience and initial structure, rather than manual programming. Unlike Statistics, which primarily deals with drawing conclusions from data, Machine Learning incorporates questions about computational architectures and algorithms to effectively capture, store, retrieve, and merge data. It also considers how multiple learning subtasks can be orchestrated in a larger system, as well as issues of computational tractability (Mitchell, 2006).

Artificial Intelligence Turing proposed a 'definition of psychological phenomena in terms of behavioral patterns,' suggesting that a machine can be considered to possess intelligence if its behavior is 'indistinguishable from that of a human being.' According to Turing, a computer can be said to have artificial intelligence if it can replicate human responses under specific conditions (Moor, 2003).

DA, AI, ML, and DM often intersect in practical applications. For instance, in the development of AI-driven recommendation systems, DA is used to understand user behavior, DM identifies patterns in user preferences, and ML algorithms power personalized recommendations.

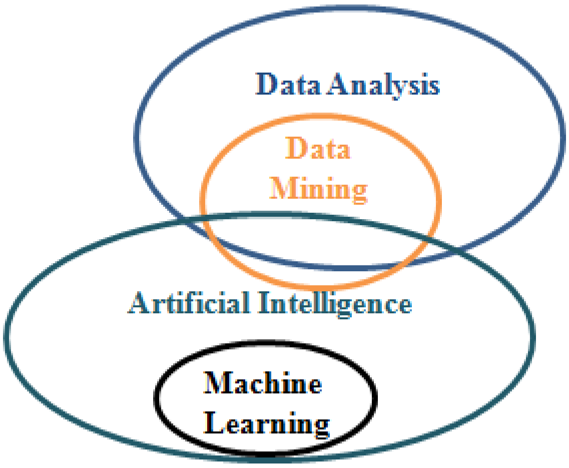


Figure 1.1: Relationship Between DA, AI, ML, and DM

1.3 Data Analysis Models

Data analysis models are mathematical or computational frameworks used to analyze and interpret data. These models help extract meaningful insights, patterns, and information from raw data. There are various types of data analysis models that can be grouped into the following main categories :

- **Statistical models:** use mathematical equations to encode the information extracted from data, make predictions, and infer relationships between variables. In some cases, statistical modeling techniques can quickly provide suitable models. Examples include linear regression and logistic regression.

- **Factorial models:** are mathematical models designed to assess the simultaneous impact of multiple factors. Their objective is to streamline the variables by condensing them into a limited set of synthetic components, primarily utilizing linear algebra tools.
  Factorial models prove especially valuable when investigating intricate systems in which multiple variables can produce interrelated and interactive effects. These models find widespread application in industrial design, allowing exploration of the connections between different factors and their influence on the studied phenomena.

- **Classification models:** are a type of machine learning model used in the field of supervised learning. These models are designed to categorize or classify input data into one or more predefined classes or categories based on their characteristics or features.
  Classification models have a wide range of applications across various industries. For example to optimize transportation logistics, these models can used to classify routes based on factors such as traffic, weather, and cost-effectiveness.

- **Clustering models:** are machine learning or statistical techniques used in data analysis to group similar data points or objects together based on certain features or characteristics. The primary goal of clustering

is to find hidden patterns, structures, or natural groupings within a dataset without prior knowledge of class labels or categories.
Clustering models find applications in various industries due to their ability to uncover patterns and group similar data points or objects together. For example, in customer segmentation, they are used to cluster customers with similar purchase behaviors.

1.4 Types of Data Analysis

Data analysis encompasses a variety of methodologies and techniques employed to investigate, interpret, and derive insights from data. The selection of these methodologies is driven by the particular objectives of the analysis. Below are several prevalent categories of data analysis:

- **Descriptive analysis:** revolves around the task of summarizing and presenting data in a manner that provides a comprehensive view of key characteristics, such as mean, median, mode, and standard deviation.
  Its primary objective is to facilitate a grasp of the fundamental attributes of the data, offering insights into accomplishments and providing an understanding of past events.

- **Predictive Analysis:** involves building models that can make predictions about future outcomes based on historical data.
  Data Mining, Machine learning algorithms and regression models are commonly used for predictive analysis.

- **Prescriptive analysis:** helps in selecting the best solution among several possible actions to guide what will be achieved.
  It is often employed in decision optimization and provides recommendations for achieving desired outcomes.

- **Diagnostic analysis:** focuses on identifying the causes of specific events or outcomes. It is particularly useful in troubleshooting issues and understanding why certain events occurred.

1.5 Nature of Data

In data analysis, two primary types of data or variables are commonly distinguished: quantitative data and qualitative data.

1.5.1 Quantitative or Numeric Data

In data analysis, quantitative or numeric data refers to information or characteristics that can be quantified and are expressed as numerical values.
In quantitative data, two primary types are recognized: continuous quantitative data and discrete quantitative data.

1.5.1.1 Continuous Quantitative Data

Quantitative data is considered continuous when it can take an infinite number of real values within a given interval. The height of a person is an example of continuous quantitative data.

1.5.1.2 Discrete Quantitative Data

Discrete quantitative data is limited to a finite set of real values within a specified interval. For instance, the number of employees in a company

1.5.2 Qualitative or Categorical Data

Qualitative or categorical data refers to a non-quantifiable characteristic, often stemming from a count.
Qualitative data can be divided into two major categories: nominal qualitative data and ordinal qualitative data.

1.5.2.1 Nominal Qualitative Data

Refers to descriptions of names or categories without any order. These data are primarily used for labeling variables. For instance, color.

1.5.2.2 Ordinal Qualitative Data

Is data that exhibits values defined by an ordering relationship among the different possible categories. Customer ratings of a company's service quality are an example of ordinal qualitative data. It includes categories like 'Good,' 'Very Good,' 'Excellent'.

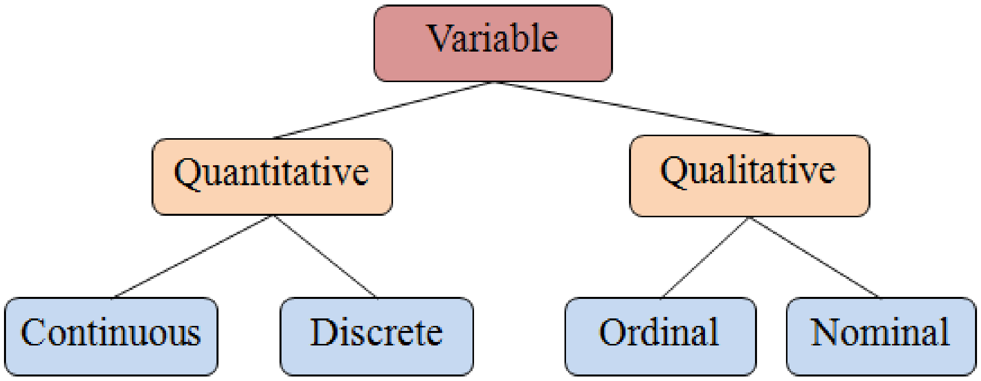


Figure 1.2: Data types

1.6 Impact of Data Analysis in the Industry

The process of analyzing information is invaluable, not only for decision-making but also for business development. Indeed, data analysis has brought about a revolution in the way businesses operate.

Here are some key aspects of its impact:

- **Improved decision-making:** the decision can be improved by analyzing historical and real-time data, businesses can identify trends, patterns, and insights that guide strategic choices.
- **Supply chain optimization:** data analysis optimizes supply chain management by forecasting demand, managing inventory efficiently, and minimizing disruptions. This ensures that products are delivered to customers on time.

Figure 1.3: Data Analysis and the Industry

- **Quality control:** industries like pharmaceuticals and automotive use data analysis for quality control. It helps identify defects and deviations early in the production process.

- **Energy efficiency:** energy-intensive industries utilize data analysis to optimize energy consumption. This reduces energy costs and environmental impact.

- **Predictive maintenance:** in manufacturing and heavy industries, data analysis is used for predictive maintenance. By analyzing sensor data and equipment performance, organizations can predict when machinery needs maintenance, reducing downtime and maintenance costs.

- **Personalized customer:** industries such as e-commerce, marketing, and retail leverage data analysis to understand customer behavior and preferences. This enables the delivery of personalized products and services, improving customer satisfaction and loyalty.

- **Competitive advantage:** companies that effectively harness data analysis gain a competitive advantage. They can respond quickly to market changes, adapt strategies, and innovate based on data-driven insights.

1.7 Conclusion

In this chapter, we've outlined the core principles of data analysis. We commenced by defining data analysis and its associations with artificial intelligence, machine learning, and data mining. Following this, we delved into the models of data analysis and its various types. We also introduced the different data types, encompassing quantitative data with its subsets of continuous and discrete categories, as well as qualitative data with nominal and ordinal classifications. Finally, we explored the significance of data analysis within industry.

In the subsequent chapter, we will introduce the unidimensional measures employed in data analysis.

CHAPTER 2
UNIDIMENSIONAL MEASURES

In this chapter, we will introduce the most frequently employed unidimensional measures in descriptive statistics. Specifically, Section 2.2.2 covers Measures of Central Tendency. Section 2.2.3 delves into Measures of Variation. Measures of Position are elucidated in Section 2.2.4, and Section 2.2.5 outlines Measures of Skewness and Kurtosis. We conclude this chapter with a set of exercises and their respective solutions.

2.1 Definition of Descriptive Statistics

Descriptive statistics, within the realm of statistics, entails techniques for summarizing and presenting data in a meaningful manner, facilitating a comprehensive grasp of its fundamental characteristics. Descriptive statistics empower us to draw inferences beyond the analyzed data and arrive at conclusions concerning any hypotheses we may have formulated.
The most common unidimensional descriptive statistical measures will be presented in the following section.

2.2 Measures in Descriptive Statistics

2.2.1 Frequency Distribution

In statistics, frequency relates to the frequency of occurrence of specific values within a dataset, which can consist of categorical or numeric variables. It is common to use class intervals for continuous variables. Relative frequency distributions, meanwhile, convey frequency information as percentages.

2.2.2 Measures of Central Tendency

2.2.2.1 Mean (Average)

Calculated by summing all values and dividing by the number of data points. It represents the central or typical value.

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{2.1}$$

where:
$\bar{X}$: represents the sample mean.
N: represents the total number of values in the sample.
denotes the summation of all values from i=1 to i=N.
x_i : represents each individual value in the sample.

We can also compute the mean using the frequencies or relative frequencies.

$$\bar{X} = \frac{\sum_{i=1}^{N} x_i \times f_i}{\sum_{i=1}^{N} f_i} \tag{2.2}$$

Where,

f_i : represents the frequency of individual x_i.

2.2.2.2 Median

The middle value when data is ordered from least to greatest, It is less influenced by outliers in comparison to the mean.

$$\text{Median} = \begin{cases} \frac{(x_{\frac{N}{2}} + x_{(\frac{N}{2}+1)})}{2} & \text{if } N \text{ is even} \\ x_{\frac{N+1}{2}} & \text{if } N \text{ is odd} \end{cases} \tag{2.3}$$

2.2.2.3 Mode

The most frequently occurring value in a data.

$$\text{Mode} = \text{value with the highest frequency} \tag{2.4}$$

To compute the mode, another formula can be used:

$$Mode = L + h \times \frac{(f_m - f_1)}{(f_m - f_1) + (f_m - f_2)} \tag{2.5}$$

Where:

L denotes the lower limit of the modal class.

h denotes the size of the class interval.

f_m denotes the frequency of the modal class.

f_1 denotes the frequency of the class preceding the modal class.

f_2 denotes the frequency of the class succeeding the modal class.

	Advantages	**Disadvantages**
Mean	-Easy to calculate	-Strongly influenced by extreme values -Poor representation of a heterogeneous population
Median	-Not influenced by extreme values -Calculable for cyclical data where the mean has little significance.	-Represents only the value that divides the sample into two equal parts
Mode	-Not influenced by extreme values -Calculable for cyclical data -A good indicator of heterogeneous population	-Very sensitive to variations in class amplitudes

Table 2.1: Comparison between Mean, Median, and Mode

2.2.3 Measures of Variation (Spread)

2.2.3.1 Range

Represent the difference between the maximum and minimum values in a data.

$$Range = max(x_i) - min(x_i) \tag{2.6}$$

2.2.3.2 Absolute Deviation

Is a statistical term that refers to the absolute value of the difference between a data point and a measure of central tendency, such as the mean, median, or mode. It quantifies the extent to which individual data points in a data set vary from the chosen measure of central tendency.

$$AbsoluteDeviation = |x - C| \tag{2.7}$$

Where:
x : is data point
C : is central value (mean, median, or mode)

2.2.3.3 Variance

A measure of how far the set of data is dispersed from their mean value. Represented by σ^2 or S^2 the variance is the average of the squared difference from the mean.
The formula for population variance is defined as follows:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \tag{2.8}$$

Where:
N : is the number of observation in the population
x_i : is the ith observation in the population
μ : is the mean of the population.

When using frequencies, the formula becomes:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} f_i x_i^2 - \mu^2 \tag{2.9}$$

The formula for Sample Variance is defined as follows:

$$S^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{X})^2}{N-1} \tag{2.10}$$

Where
$\bar{X}$: is the mean of x_i .

2.2.3.4 Standard Deviation

Standard Deviation, also called root mean square deviation, is the square root of the variance of the given data set. It is a measure of the average amount of variation.
The formula for population standard deviation is defined as follows:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}} \tag{2.11}$$

The formula for sample standard deviation is defined as follows:

$$S = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{X})^2}{N-1}} \tag{2.12}$$

	Advantages	**Disadvantages**
Range	-Easy to calculate	-Sensitivity to outliers
Absolute Deviation	-Simplicity -Used to measure error or deviation from a desired value	-Limited for complex situation
Variance	-Quantifiable measure of how spread out the data points are around the mean value -Sensitive to deviations	-Lack of direction -It less intuitive to understand -Sensitive to outliers
Standard Deviation	-Measured in the same units as the original data so it clear and interpretable -Useful for comparisons	-May not reflect data distribution shape -Sensitive to outliers

Table 2.2: Comparison between Range, Absolute deviation, Variance and Standard Deviation

2.2.4 Measures of Position

Quantiles represent a natural extension of the concept of the median, as they are values that partition a dataset into equal segments. These quantiles are commonly classified as quartiles, deciles, and percentiles.

2.2.4.1 Quartiles

Divide the distribution into four equal parts, and they are labeled as Q1 (the first quartile), Q2 (the second quartile or median), and Q3 (the third quartile).
The significance of each quartile is as follows:
Q1: 25% of the data lies below Q1, while 75% lies above it.
Q2: Marks the midpoint, with 50% below and 50% above.
Q3: 75% of the data falls below Q3, leaving 25% above.

2.2.4.2 Deciles

It split the distribution into ten equal segments, represented as D1, D2, D3, D4, D5, D6, D7, D8, and D9. Each 'D' value divides the lower 10% of the data from the upper 90%, and so forth.

2.2.4.3 Percentiles

Percentiles divide the data into 100 equal parts, yielding 99 percentiles labeled as P1, P2, P3, P4, and so on.

2.2.5 Measures of Skewness and Kurtosis

2.2.5.1 Skewness

Skewness is a statistical measure that quantifies the departure from symmetry or the presence of asymmetry in a dataset. It becomes evident on a bell curve when data points are not evenly distributed to the left and right of the median.

The distribution of skewness values is as below:

Skewness = 0 when the distribution is normal.

Skewness > 0 when more weight is on the left side of the distribution.

Skewness < 0 when more weight is on the right side of the distribution.

$$Skewness = \frac{\sum_{i=1}^{N}(x_i - \bar{X})^3}{(N-1) \times \sigma^3)} \tag{2.13}$$

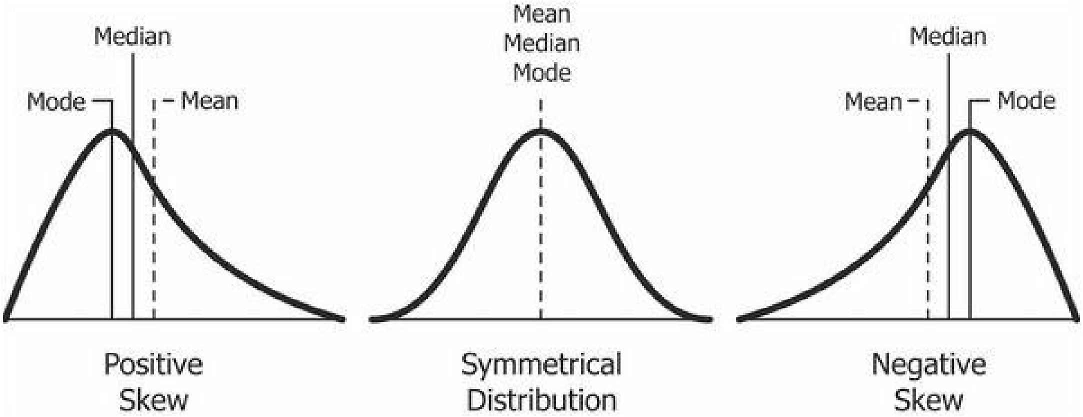


Figure 2.1: Positive and negative skew

2.2.5.2 Kurtosis

Kurtosis is a statistical term that characterizes frequency distribution. It measures the peakedness or flatness of the data distribution.

kurtosis =3, indicates that the distribution has the same tailedness as a normal distribution

kurtosis < 3, is referred to as platykurtic, which means the distribution has lighter tails and is flatter compared to a normal distribution.
kurtosis > 3, is termed leptokurtic, signifying that the distribution has heavier tails and is more peaked compared to a normal distribution.

$$Kurtosis = \frac{\sum_{i=1}^{N}(x_i - \bar{X})^4}{N \times S^4} \tag{2.14}$$

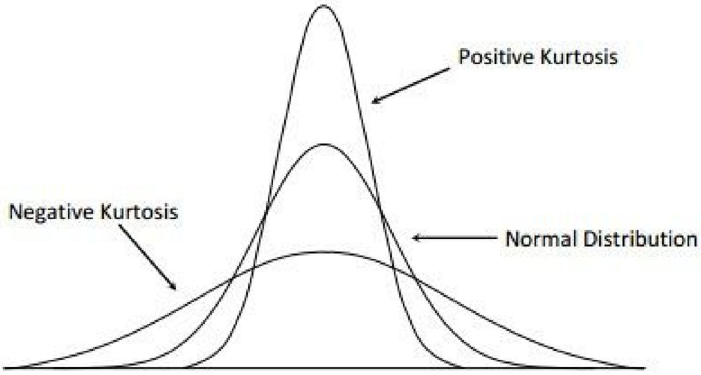


Figure 2.2: Positive and negative Kurtosis

2.3 Exercices

Exercice 1

The following table shows the distribution of daily wages in dinars in a company:

Daily Wage (xi)	**# Employees (fi)**
450	6
500	10
550	24
600	18
700	5

Table 2.3: Daily Wage

1- Calculate the mean and the standard deviation of this series.

2- If the daily wage of each employee is increased by 200 DA, what will be the new mean and standard deviation of the series ?

Exercice 2

The table below displays the quantity of pieces manufactured by a machine over a span of 10 years.

1- What was the average number of pieces produced during this period ?

2- How many pieces were manufactured each day within the same timeframe ?

Years	**Piece**
2009	623
2010	583
2011	959
2012	1 037
2013	960
2014	797
2015	663
2016	652
2017	560
2018	619
Total	**7 453**

Table 2.4: Quantity of pieces manufactured

2.4 Solutions

Solution exercice 1

1-The calculation of the mean and standard deviation:

$$\mu = \frac{\sum_{i=1}^{N} x_i \times f_i}{\sum_{i=1}^{N} f_i} \tag{2.15}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}} \tag{2.16}$$

OR

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} f_i x_i^2 - \mu^2} \tag{2.17}$$

μ=(450*6 + 500*10 + 550*24 + 600*18 + 700*5)/63
μ=558.73

σ^{2} = (6*(450-μ)^{2} + 10*(500-μ)^{2} +24*(550-μ)^2+18*(600-μ)^2+5*(700-μ)^2)/63
OR

σ^2 = ((6*(450)^{2} + 10*(500)^{2} +24*(550)^2+18*(600)^2+5*(700)^2)/63)-μ^2

σ=61.43

2- If the daily wage of each employee is increased by 200 DA, the new mean will be:

μ = Old (μ) + 200
μ =558.73+200=758.73

The standard deviation will remain the same in this case because adding a constant value to each data point does not change the spread or variability of the data.

Solution exercice 2

The total production of pieces, as shown in the table, is 7,453.

1- We obtain the yearly average by dividing the total by 10 since the table spans 10 years, resulting in an annual average of approximately 745 pieces.

2- To calculate the daily average, divide the total by 365, which yields approximately 2 pieces per day.

2.5 Conclusion

In this second chapter, we have introduced essential unidimensional measures crucial for data analysis. We discussed descriptive statistical measures, covering frequency distribution and measures of central tendency like mean, median, and mode. Additionally, we explored measures of variance or spread, including range, absolute deviation, variance, and standard deviation. Furthermore, we presented measures of position, such as quartiles, deciles, and percentiles, along with measures of skewness and kurtosis. Lastly, we provided a series of exercises for practice.

CHAPTER 3
BIDIMENSIONAL MEASURES

In this chapter, we will explore commonly used bidimensional measures in descriptive statistics. We will begin by introducing bidimensional descriptive statistics in Section 3.1. Then, we will present the measure of covariance and its properties in Section 3.2. Section 3.3, we will focus on various types of correlation measures, including the Pearson correlation coefficient, Spearman correlation coefficient (rho), Kendall's tau correlation coefficient, and point biserial correlation. In Section 3.4, we will delve into the chi-square test of independence. To wrap up, we will provide exercises and their solutions in Sections 3.5 and 3.6.

3.1 Bidimensional Descriptive Statistics

Bidimensional descriptive statistics, often referred to as bivariate or two-variable statistics, enable us to explore and quantify the relationships and associations between two variables. This offers valuable insights into how changes in one variable correspond to changes in another. Such insights are beyond the scope of univariate statistics. Bidimensional descriptive statistics serve as a foundation for informed decision-making and the derivation of meaningful conclusions.

In the following section, we will present the most commonly used measures and methods for analyzing the relationship or association between two variables.

3.2 Covariance

In statistics, covariance is a mathematical technique used to gauge the direction of variation between two variables, helping to describe the extent to which these variables are independent of each other.

For two datasets with N data points (xi and yi), the population covariance is calculated as the average of the products of the deviations of each pair of data points from their respective means:

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_{(X)})(y_i - \mu_{(X)}) \tag{3.1}$$

Where

$\mu_{(X)}$ and $\mu_{(Y)}$ represent the means of variables X and Y, respectively.

The formula for sample covariance is as following:

$$Cov(X,Y) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{X}(y_i - \overline{Y}) \tag{3.2}$$

Where

$\overline{X}$ and $\overline{Y}$ represent the means of variables X and Y, respectively.

3.2.1 Covariance vs Variance

Covariance and variance are distinct statistical measures that serve different purposes in data analysis. Variance helps quantify the spread of a single variable's data, while covariance assesses the relationship between two variables and whether they move together or in opposite directions.

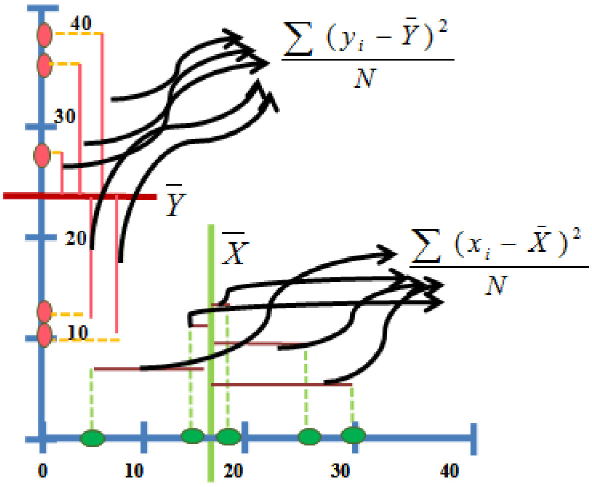

$\frac{\sum (y_i - \bar{Y})^2}{N}$
$\bar{Y}$
$\bar{X}$
$\frac{\sum (x_i - \bar{X})^2}{N}$
40
30
20
10
0
10
20
30
40

Figure 3.1: Variance of X and Y

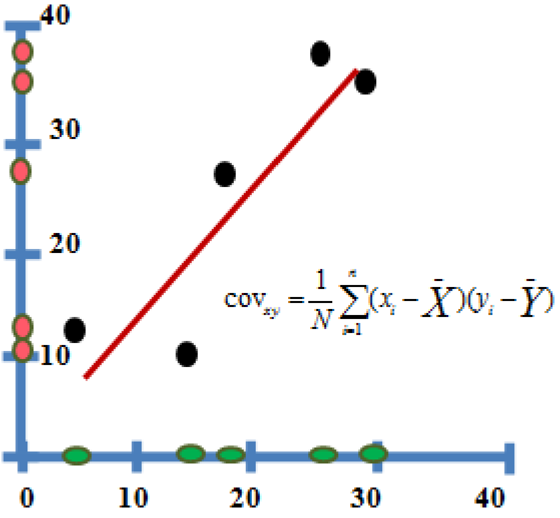


Figure 3.2: Covariance(x,y)

3.2.2 Properties of Covariance

If X and Y are random variables with real values, and c is a constant, then the following formulas are direct consequences of the definition of covariance:

$$Cov(X, X) = Var(X) \tag{3.3}$$

$$Cov(X, Y) = Cov(Y, X) \tag{3.4}$$

$$Cov(cX, Y) = cCov(Y, X) \tag{3.5}$$

The sign of the covariance indicates the direction of the association between the two variables:
Positive Covariance (Cov(X, Y) $>$ 0): indicates that when one variable is above its mean, the other tends to be above its mean as well, and vice versa. This suggests a positive linear association.
Negative Covariance (Cov(X, Y) $<$ 0): indicates that when one variable is above its mean, the other tends to be below its mean, and vice versa. This suggests a negative linear association.
Zero Covariance (Cov(X, Y) $=$ 0): indicates no linear association between the variables. However, it doesn't necessarily imply independence.

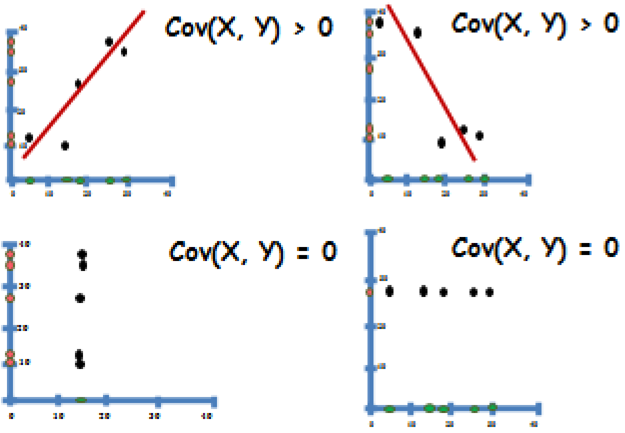


Figure 3.3: Direction of Association

3.3 Correlation

Correlation is a statistical measure that quantifies the extent to which two variables are linearly related and assesses the strength and direction of their relationship.

Correlation is typically measured using the sample correlation coefficient, denoted as **'r'**.
Unlike covariance, which has units that are the product of the units of the two variables, correlation is a dimensionless measure, making it easier to compare and interpret the strength and direction of relationships between variables.

The formula for the sample correlation coefficient, which measures the linear relationship between two variables X and Y, is as follows:

$$r = \frac{\sum(x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum(x_i - \bar{X})^2(y_i - \bar{Y})^2}} \tag{3.6}$$

The formula for calculating the sample correlation coefficient (r) using the covariance (Cov) and standard deviations (σ) of two variables X and Y is as follows:

$$r = \frac{Cov(X, Y)}{\sigma(X) \times \sigma(Y)} \tag{3.7}$$

3.3.1 Properties of Correlation

The correlation coefficient 'r' is a dimensionless measure that falls within the range of -1 to +1.

The sign of the correlation coefficient indicates the direction of the association between the two variables:
Positive Correlation (r> 0): suggesting that both variables tend to increase together.
Negative Correlation (r< 0): indicating that when one variable's values increase, the other variable's values tend to decrease
When **'r' is closer to zero**, it indicates a weaker linear relationship between the variables.

Common correlation coefficients include Pearson correlation coefficient (for linear relationships), Spearman rank correlation coefficient (for ordinal data or non-linear relationships), and Kendall's tau correlation coefficient (for rank-ordered data).

3.3.2 Pearson Correlation Coefficient

The Pearson correlation coefficient, often symbolized as **'r'**, serves as a statistical metric that gauges the strength and direction of the linear relationship between two variables.

The Pearson correlation is the preferred choice when the following conditions are true:

- Both variables under consideration are quantitative.
- The variables exhibit normal (or nearly normal) distributions.
- Outliers are not present in the dataset.
- The relationship between the variables is fundamentally linear in nature. It's important to note that the Pearson correlation may not accurately capture non-linear associations.

The formula for computing the Pearson correlation coefficient between two variables X and Y is as follows:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2} \times \sqrt{n\sum y^2 - (\sum y)^2}} \tag{3.8}$$

Where,
n is the sample size.

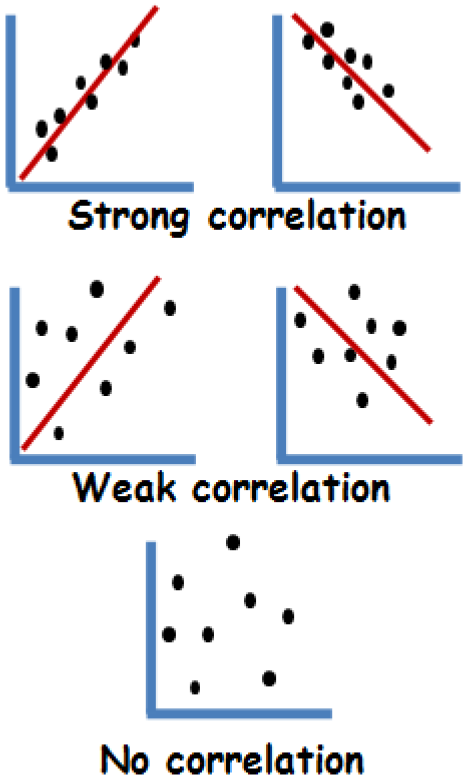


Figure 3.4: Visualizing the Pearson correlation coefficient

Testing for the significance of the Pearson correlation coefficient

To draw inferences about the population correlation ρ from the sample correlation 'r' statistical hypothesis tests are commonly employed. These methods assist in assessing whether the observed correlation in the sample is statistically significant and provide a confidence interval within which the population correlation ρ is likely to lie.

The Pearson correlation of the sample is r. It is an estimate of rho (ρ), the Pearson correlation of the population. Knowing r and n, we can infer whether ρ is significantly different from 0.

Null hypothesis (H_0): $\rho = 0$
Alternative hypothesis (H_1): $\rho \neq 0$

We have three steps to test the hypotheses:

- Calculate the t value using this formula

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \tag{3.9}$$

- Find the critical value of t^* in "t table". This value depends on the degrees of freedom (df) and the significance level (α). For Pearson correlation tests, df $=$ n $-$ 2, the significance level is usually .05 and, two-tailed is an appropriate choice.

- Accept or reject the null hypothesis:
  If the t value $> t^*$ value, then the relationship is statistically significant. The data allows you to reject the null hypothesis and provides support for the alternative hypothesis.
  If the the t value $< t^*$ value, then the relationship is not statistically significant. The data doesn't allow you to reject the null hypothesis.

3.3.3 Spearman Correlation Coefficient (Rho)

Spearman's rho, or Spearman's rank correlation coefficient **'r_s'**, is a non-parametric statistical measure employed to evaluate the strength and direction of the monotonic relationship between two variables.

The Spearman's rank correlation coefficient is the preferred choice when the following conditions are true

- The variables are on an ordinal level of measurement.
- The relationship between variables is not expected to be linear.

In a linear relationship, both variables change in the same direction at a consistent rate across the entire data range. Conversely, in a monotonic relationship, both variables change in only one direction, though not necessarily at the same rate.

- **Positive Monotonic:** as one variable increases, the other also increases.
- **Negative Monotonic:** As one variable increases, the other decreases.

As with the previous correction, the value of Spearman's rho falls within the range of -1 to +1. A value of +1 signifies a perfect positive monotonic relationship. A value of -1 signifies a perfect negative monotonic relationship. A value of 0 signifies no monotonic relationship.

The first step in computing Spearman's rank correlation coefficient is to rank the values of the variables X and Y separately. Arrange them from lowest to highest and assign ranks, starting with 1 for the smallest value, 2 for the next smallest, and so forth. In cases where values are equal, handle them by assigning the average rank.

The second step involves calculating the differences (d) between the ranks of corresponding data points for X and Y.

$$d_i = rank(X_i) - rank(Y_i) \tag{3.10}$$

Finaly we used the formula of spearman's rank correlation (3.11)

$$r_s = 1 - \frac{6\sum d_i^2}{(n^2 - n)} \tag{3.11}$$

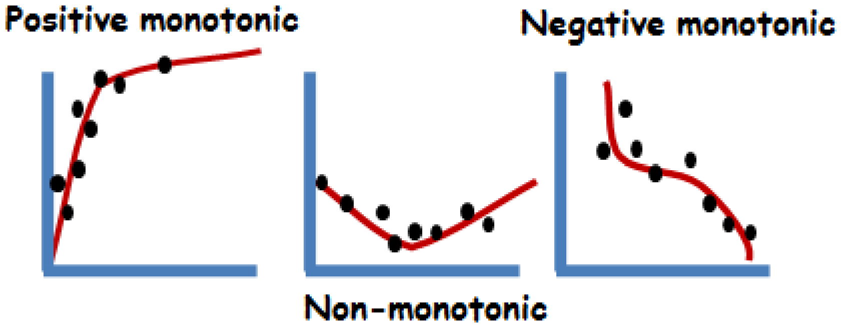


Figure 3.5: Visualizing the Spearman correlation coefficient

3.3.4 Kendall's tau Correlation Coefficient

Kendall rank correlation coefficient or Kendall's Tau, often denoted as (τ), measures the relationship between two variables.
Kendall's Tau is the preferred choice when the following condition is true:

- The two variables must only have an ordinal scale level.

The Kendall's tau is very similar to Spearman's rank correlation coefficient. However, Kendall's Tau should be preferred over Spearman's correlation when there is very little data and many rank ties.

The formula for computing Kendall's Tau between two variables X and Y is as follows:

$$\tau = \frac{C - D}{\frac{1}{2}n(n-1)} \tag{3.12}$$

Where:
C represents the number of concordant pairs, which are data points that have the same order in both sets of rankings.
D represents the number of discordant pairs, which are pairs that have different orders in the two sets of rankings.
n represents the total number of data points.
An alternate formula for the Kendall's Tau is as follows:

$$\tau = \frac{C - D}{C + D} \tag{3.13}$$

Kendall's Tau can take on values between -1 and 1, where:

- τ = **1**, it indicates perfect positive concordance, meaning that the two sets of rankings are identical.
- τ = **-1**, it indicates perfect negative concordance, meaning that the two sets of rankings are in reverse order.
- τ = **0**, it suggests no association or correlation between the rankings.

3.3.5 Point Biserial Correlation

The point-biserial correlation coefficient,denoted as r_{pb}, is a statistical measure used to quantify the strength and direction of the linear relationship between a dichotomous variable and a continuous variable. The dichotomous variable is a binary variable with two expressions, for example (male and female) or (yes and no).

The formula for calculating the point-biserial correlation coefficient is expressed as:

$$r_{pb} = \frac{\bar{X}_1 - \bar{X}_0}{S} \sqrt{\frac{n_1 n_0}{N(N-1)}} \tag{3.14}$$

Where:
$\bar{X}_1, \bar{X}_0$ are the means of the continuous variable for the two binary categories (1 and 0, typically).
S is the standard deviation of the continuous variable.
n_1, n_1 are the sample sizes for the two binary categories.
N is the total sample size.
Like other correlations, the point-biserial correlation coefficient can range from -1 to 1, where:

- r_{pb} =1, it indicates a strong positive relationship between the binary and continuous variables.
- r_{pb} =-1 it indicates a strong negative relationship, meaning that as the binary variable increases, the continuous variable tends to decrease.
- r_{pb} close to 0, it suggests little to no linear relationship between the binary and continuous variables

3.3.6 Correlation Coefficient Types: A Comprehensive Comparison

Correlation coefficient	Type of relationship	Levels of measurement
Pearson's r	Linear	Two quantitative variables
Spearman's rho	Non-linear	Two ordinal variables
Kendall's tau	Non-linear	Two ordinal variables
Point-biserial	Linear	One dichotomous (binary) variable and one quantitative variable

Table 3.1: Different types of correlation coefficients comparison

3.4 Chi-Square Test

Chi-square test of independence, also known as the χ^2 test, is a statistical method used to determine if there is a significant association or relationship between two categorical variables.
Chi-square test of independence (χ^2) is the preferred choice when the following conditions are true:

- The two variables are categorical variables (nominal or ordinal).
- There are a minimum of five observations expected in each group.

The chi-square test methodology comprises the following sequential steps:

1. **Define the hypothesis**
   H_0: there is no link between the two variables.
   H_1: there is a link between the two variables.
2. **Construct an observed data table**
   Organize the data into a contingency table or cross-tabulation table, depicting the counts for each combination of categories related to the two variables.
3. **Construct an expected data table**
   Organize the data into a contingency table, where each cells represent the expected value.
   The expected value is computed according to the following formula:

$$ExpectedValue = \frac{RowTotal \times ColumnTotal}{TotalNumberofObservation} \tag{3.15}$$

4. **Calculate Chi-Square**
   The formula is :

$$\chi^2 = \frac{\sum(O-E)^2}{E} \tag{3.16}$$

   Where:
   O is the observed value in each cell of observed data table.
   E is the expected value in each cell of expected data table.

5. **Compare the obtained χ^2 to the critical statistic found in the chi-square table**
Degrees of freedom: are calculated based on the dimensions of the contingency table.

$$df = (NumberofRows - 1) * (NumberofColumns - 1) \tag{3.17}$$

Significance level (α): can be 1%,5%,10%.

6. **Determine if there is an association**
If $\chi^2 > \chi^2$ critical, we reject the null hypothesis. This implies that H_1 is accepted, concluding that there is a relationship between the two variables.
If $\chi^2 < \chi^2$ critical, we accept the null hypothesis, indicating that there is no relationship between the two variables.

3.5 Exercises

Exercise 1

A study was undertaken to examine the relationship between the weight (kg) and length (cm) of 10 babies. The results are displayed in the following table.

1- Utilizing the Pearson correlation coefficient, is there a correlation between newborns' weight and length ?

2- Determine the statistical significance of the Pearson correlation coefficient by using the t-test.

Weight (X)	**Length (Y)**
3.63	53.1
3.02	49.7
3.82	48.4
3.42	54.2
3.59	54.9
2.87	43.7
3.03	47.2
3.46	45.2
3.36	54.4
3.3	50.4

Table 3.2: The weight (kg) and length (cm)

Exercise 2

The table below displays the ranking of players in two different competitions.
- Can we infer that those who achieved high rankings in the first competition also attained good rankings in the second competition ?

Players	Competition 1 (X)	Competition 2 (Y)
P1	7	10
P2	10	12
P3	1	4
P4	6	7
P5	9	11
P6	13	9
P7	3	2
P8	5	4
P9	11	5
P10	9	11
P11	6	6
P12	4	1

Table 3.3: The ranking of players

Exercise 3

Consider a scenario where two employers are ranking six candidates for a job, ranging from worst to best. The table below displays the rankings assigned by each employer to the employees.

- Is there a correlation between the two ranks ?

Employee	Rank 1 (X)	Rank 2 (Y)
E1	1	3
E2	2	1
E3	3	4
E4	4	2
E5	5	6
E6	6	5

Table 3.4: Recruitment Rankings

Exercise 4

A company has analyzed the number of purchases of three decorative products in relation to the gender of their customers (male/female). The results obtained are as follows: men purchased 100, 70, and 30 units of products 1, 2, and 3, respectively, while women purchased 140, 60, and 20 units of products 1, 2, and 3, respectively.

- Determine whether there is a correlation between the gender of the clientele and the products they bought ($\alpha = 0.05$).

3.6 Solutions

Solution exercise 1

1- Utilizing the Pearson correlation coefficient The formula for computing the Pearson correlation coefficient between two variables X and Y is as follows:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2} \times \sqrt{n\sum y^2 - (\sum y)^2}}$$

n=10
$\sum x = 33.5$
$\sum y = 501.2$
$\sum x^2 = 113.05$
$\sum y^2 = 25264$
$\sum xy = 1684.2$

$$r = \frac{10 * 1684.2 - (33.5)(501.2)}{\sqrt{10 * 113.05 - (33.5)^2} \times \sqrt{10 * 25264 - (501.2)^2}}$$

$r = 0.47$.
With a correlation coefficient of r = 0.47, it represents a very weak correlation, indicating a minimal relationship between newborns' weight and length.

2- The statistical significance of the Pearson correlation coefficient by using the t-test :

"r" is an estimate of rho (ρ), the Pearson correlation of the population. Knowing r and n, we can infer whether ρ is significantly different from 0.

Null hypothesis (H$_0$): $\rho = 0$
Alternative hypothesis (H$_1$): $\rho \neq 0$

- We calculate the t value using this formula

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

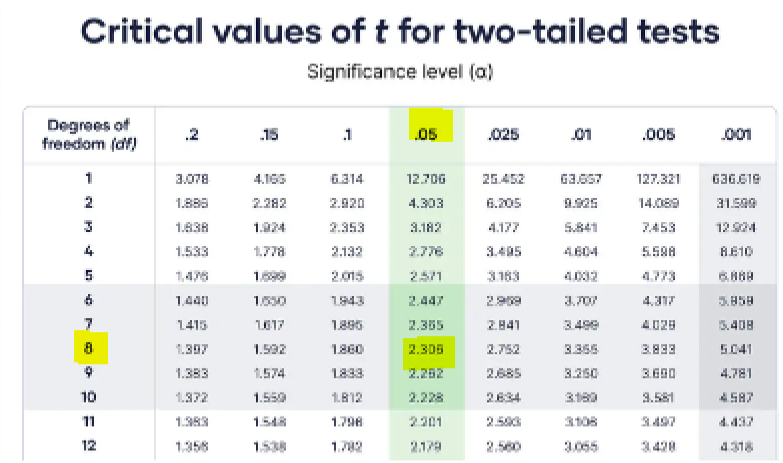
$$t = \frac{0.47}{\sqrt{\frac{1-(0.47))^2}{10-2}}} = 1.506$$

- The critical value of t^*

  The degrees of freedom (df) is n-2, so df=8.

  The significance level is α =05.

  Accorrding to the "t table" t^*=2.305



Critical values of *t* for two-tailed tests

Degrees of freedom *(df)*	Significance level (α) .2	.15	.1	.05	.025	.01	.005	.001
1	3.078	4.165	6.314	12.706	25.452	63.657	127.321	636.619
2	1.886	2.282	2.920	4.303	6.205	9.925	14.089	31.599
3	1.638	1.924	2.353	3.182	4.177	5.841	7.453	12.924
4	1.533	1.778	2.132	2.776	3.495	4.604	5.598	8.610
5	1.476	1.699	2.015	2.571	3.163	4.032	4.773	6.869
6	1.440	1.650	1.943	2.447	2.969	3.707	4.317	5.959
7	1.415	1.617	1.895	2.365	2.841	3.499	4.029	5.408
8	1.397	1.592	1.860	2.306	2.752	3.355	3.833	5.041
9	1.383	1.574	1.833	2.262	2.685	3.250	3.690	4.781
10	1.372	1.559	1.812	2.228	2.634	3.169	3.581	4.587
11	1.363	1.548	1.796	2.201	2.593	3.106	3.497	4.437
12	1.356	1.538	1.782	2.179	2.560	3.055	3.428	4.318

Figure 3.6: Critical values of t

- Accept or reject the null hypothesis ?

  The t value (1.506) $<$ t^* value (2.305), it signifies that the relationship lacks statistical significance, and thus, we do not reject the null hypothesis. This outcome can be attributed to the limited sample size of 10. Increasing the sample size may reveal a significant relationship.

Solution exercise 2

We use Spearman's correlation because the variables are on an ordinal scale level.

First, we rank the data according to the variable (X). We can observe from Table 3.5 that player 4 and 11 share the same ranking in competition 1, as do players 5 and 10 . Therefore, their final rank is the average of their ranks.

Players	**Competition 1 (X)**	**Rank**	**Final Rank (X)**
P3	1	1	1
P7	3	2	2
P12	4	3	3
P8	5	4	4
P4	6	5	5.5
P11	6	6	5.5
P1	7	7	7
P5	9	8	8.5
P10	9	9	8.5
P2	10	10	10
P9	11	11	11
P6	13	12	12

Table 3.5: Competition 1 (X) Rankings

Secondly, we rank the data according to the variable (Y). We can observe from Table 3.6 that players 8 and 3 share the same ranking in competition 2, as do players 5 and 10. Consequently, their final rank is calculated as the average of their ranks

Table 3.7, presente the final rang of the variable (X) and the variable (Y) with their difference ranks (d). Finaly we used the formula of spearman's rank correlation (equation 3.11) to compute r_s :

$$r_s = 1 - \frac{6\sum d_i^2}{(n^3 - n)}$$

$$r_s = 1 - \frac{6 \times 81}{(12^3 - 12)}$$

Players	**Competition 2 (Y)**	**Rank**	**Final Rank (Y)**
P12	1	1	1
P7	2	2	2
P8	4	3	3.5
P3	4	4	3.5
P9	5	5	5
P11	6	6	6
P4	7	7	7
P6	9	8	8
P1	10	9	9
P5	11	10	10.5
P10	11	11	10.5
P2	12	12	12

Table 3.6: Competition 2 (Y) Rankings

r_s=0.72,
With an r_s value of 0.72, it indicates a strong correlation between variables X and Y. Hence, we can deduce that individuals who performed well in the first competition also achieved favorable rankings in the second competition.

Players	Rank (X)	Rank (Y)	d	d*d
P1	7	9	-2	4
P2	10	12	-2	4
P3	1	3.5	-2.5	6.25
P4	5.5	7	-1.5	2.25
P5	8.5	10.5	-2	4
P6	12	8	-4	16
P7	2	2	0	0
P8	4	3.5	-0.5	0.25
P9	11	5	6	36
P10	8.5	10.5	-2	4
P11	5.5	6	-0.5	0.25
P12	3	1	2	4

Table 3.7: Final Rankings

Solution exercise 3

The variables are on an ordinal scale level, so we can use Kendall's Tau or Spearman's correlation. In this case, we prefer to use Kendall's Tau as there is very little data.
We choose rank 1 as the reference, and then we sort the employees from 1 to 6. Afterward, we compute the number of concordant and discordant pairs.

Rank 1 (X)	**Rank 2 (Y)**	**Concordant**	**Discordant**
1	3	3	2
2	1	4	0
3	4	2	1
4	2	2	0
5	6	0	1
6	5	-	-

Table 3.8: Concordant and Discordant pairs

Computing the concordant pairs:

y=3, we have 4,6 and 5 -> C=3
y=1, we have 4,2,6 and 5 -> C=4
y=4, we have 6 and 5 -> C=2
y=2, we have 6 and 5 -> C=2
y=6, − -> C=0

The total number of concordant pair C=11

Computing the discordant pairs:

y=3, we have 1 and 2 -> D=2
y=1, − -> D=0
y=4, we have 2 -> D=1
y=2, − -> D=0
y=6, we have 5 -> D=1

The total number of discordant pair D=4

Replacing in

$$\tau = \frac{C - D}{C + D} \tag{3.18}$$

$\tau = \frac{11-4}{11+4} = 0.47$
There is a medium and positive correlation between the two variables with τ= 0.47.

Solution exercise 4

To analyze the dependence between the gender of the clientele and the products they bought, we used the chi-square test of independence.

1. **Define the hypothesis**
   H_0: there is no link between the gender of the clientele and the products.
   H_1: there is a link between the gender of the clientele and the products.

2. **The observed data table**

	Product 1	**Product 2**	**Product 3**	**Total**
Male	100	70	30	200
Female	140	60	20	220
Total	240	130	50	440

Table 3.9: The observed data table

3. **The expected data table**

   The expected value for each cell in the table is computed according to equation 3.15

	Product 1	**Product 2**	**Product 3**	**Total**
Male	114	62	24	200
Female	126	68	26	220
Total	240	130	50	420

Table 3.10: The expected data table

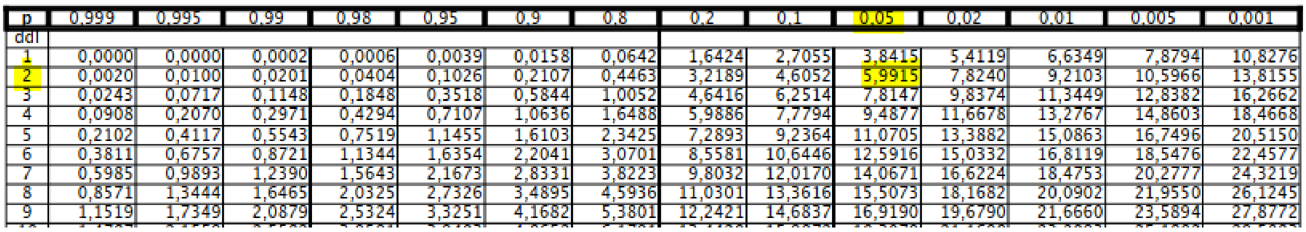
4. **Calculate Chi-Square**

   We calculate χ^2 according to the equation 3.16

   $\chi^2 = \frac{(100-114)^2}{114} + \frac{(70-62)^2}{62} + \frac{(30-24)^2}{24} + \frac{(140-126)^2}{126} + \frac{(60-68)^2}{68} + \frac{(20-26)^2}{26}$
   χ^2=8.13

5. **Comparison of the obtained χ^2 to the critical statistic found in the chi-square table**

   Degrees of freedom according to the equation 3.17 :
   df = (2 - 1) * (3 - 1)=2
   Significance level (α)=0.05.
   For an alpha level of 0.05 and two degrees of freedom, the criticalχ^2 **value is 5.9915**



p	0,999	0,995	0,99	0,98	0,95	0,9	0,8	0,2	0,1	0,05	0,02	0,01	0,005	0,001
ddl														
1	0,0000	0,0000	0,0002	0,0006	0,0039	0,0158	0,0642	1,6424	2,7055	3,8415	5,4119	6,6349	7,8794	10,8276
2	0,0020	0,0100	0,0201	0,0404	0,1026	0,2107	0,4463	3,2189	4,6052	5,9915	7,8240	9,2103	10,5966	13,8155
3	0,0243	0,0717	0,1148	0,1848	0,3518	0,5844	1,0052	4,6416	6,2514	7,8147	9,8374	11,3449	12,8382	16,2662
4	0,0908	0,2070	0,2971	0,4294	0,7107	1,0636	1,6488	5,9886	7,7794	9,4877	11,6678	13,2767	14,8603	18,4668
5	0,2102	0,4117	0,5543	0,7519	1,1455	1,6103	2,3425	7,2893	9,2364	11,0705	13,3882	15,0863	16,7496	20,5150
6	0,3811	0,6757	0,8721	1,1344	1,6354	2,2041	3,0701	8,5581	10,6446	12,5916	15,0332	16,8119	18,5476	22,4577
7	0,5985	0,9893	1,2390	1,5643	2,1673	2,8331	3,8223	9,8032	12,0170	14,0671	16,6224	18,4753	20,2777	24,3219
8	0,8571	1,3444	1,6465	2,0325	2,7326	3,4895	4,5936	11,0301	13,3616	15,5073	18,1682	20,0902	21,9550	26,1245
9	1,1519	1,7349	2,0879	2,5324	3,3251	4,1682	5,3801	12,2421	14,6837	16,9190	19,6790	21,6660	23,5894	27,8772

Figure 3.7: Chi-square Table

6. **Decision**
   With a chi-squared value of 8.13, which exceeds the critical chi-squared value of 5.991 at an alpha level of 0.05, we reject the null hypothesis (H_0) and accept H_1. This leads us to the conclusion that there is indeed a significant relationship between the gender of the clientele and the products.

3.7 Conclusion

The analysis of relationships between two variables is often a key aspect of data analysis. In this chapter, we introduced bidimensional descriptive statistics, starting with covariance and its properties, highlighting its distinctions from variance. Additionally, we discussed correlation and its properties. We then delved into various types of correlations, such as the Pearson correlation coefficient, Spearman correlation coefficient, Kendall's tau, and point biserial correlation, providing a comprehensive comparison among them. Furthermore, we introduced the chi-square test and supplemented this chapter with practical exercises.

CHAPTER 4

PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis (PCA) is a valuable technique in multidimensional analysis for extracting meaningful insights from complex datasets. PCA aids in identifying the primary variables contributing to process variations and reducing dimensionality. In this chapter, we will introduce Principal Component Analysis in Section 4.1. Section 4.2 will delve into the steps of this method. Section 4.3 will cover the most commonly used software and libraries for performing PCA. Lastly, in Sections 4.4 and 4.5, we will present some practical examples along with their solutions.

4.1 Definition of Principal Component Analysis (PCA)

Analyzing datasets with a large number of variables presents challenges in visualization and interpretation. Additionally, selecting the relevant variables for two or three dimensional plots is not straightforward.
The Principal Component Analysis (PCA) technique stands as one of the most renowned unsupervised dimensionality reduction methods. It works by transforming a large set of variables into a smaller one that retains most of the information from the original set. The objective of PCA is to identify the PCA space, which represents the direction of maximum variance within the given data (Tharwat,2016).

4.2 Understanding the Inner Workings of PCA

The primary objective of PCA is to transform a dataset with a potentially large number of correlated variables into a smaller set of uncorrelated variables known as **principal components**.

The PCA method is based on three key steps:

1. Data standardization
2. Defining the new multidimensional space,
3. Representing the data the new multidimensional space.

4.2.1 Data Standardization

In PCA, data preparation plays a critical role, involving two distinct processes: data centering and data normalization.

4.2.1.1 Data Centering

Data centering refers to the procedure of adjusting the values of a variable so that its mean becomes zero. When variables are centered, their values are modified by subtracting the mean of each variable from each individual data point associated with that variable. This adjustment does not alter the shape or spread of the data but rather relocates the data in relation to its central point. Centering proves particularly beneficial when the initial variables are directly comparable, sharing the same nature and exhibiting similar ranges of variation.

4.2.1.2 Data Normalization

Data normalization becomes necessary when different variables are measured using different metrics. For instance, one variable may represent the length of an object in meters, while another variable represents the width of the same object in centimeters. Normalization involves dividing each variable by its standard deviation.

4.2.2 Defining the new multidimensional space

The goal of Principal Component Analysis (PCA) is to project the data onto a lower-dimensional subspace while preserving the majority of the relevant information. This newly defined lower-dimensional space is characterized by fresh axes referred to as **Principal Axes**, which represent the new variables known as **Principal Components**, as illustrated in Figure 4.4. This is donne by transforming data with a potentially large number of correlated variables into a smaller set of uncorrelated principal components.

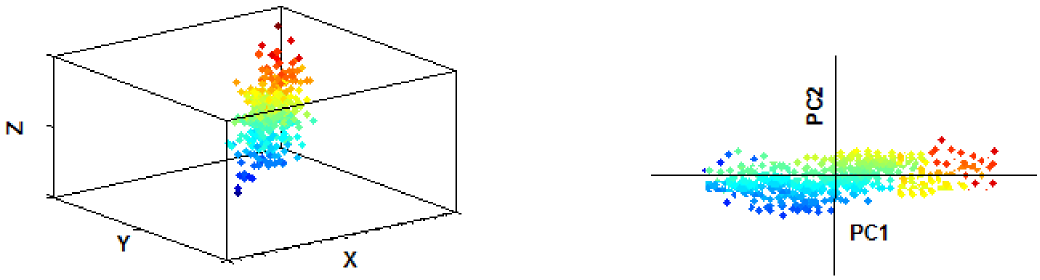


Figure 4.1: Projection onto a lower-dimensional subspace

To understand this transformation, let's consider an example with two variables, X and Y. As we've previously discussed in Chapter 3 Section 3.2, covariance is employed to analyze the variation between two variables.
Figure4.2 demonstrates how the overall shape of the data determines the covariance matrix and its associated eigenvectors and eigenvalues. Essentially, the covariance matrix captures both the data's spread (variance) and its orientation (covariance). The eigenvectors indicate the directions of maximum data spread, while their corresponding eigenvalues quantify the extent of spread (variance) in those directions.

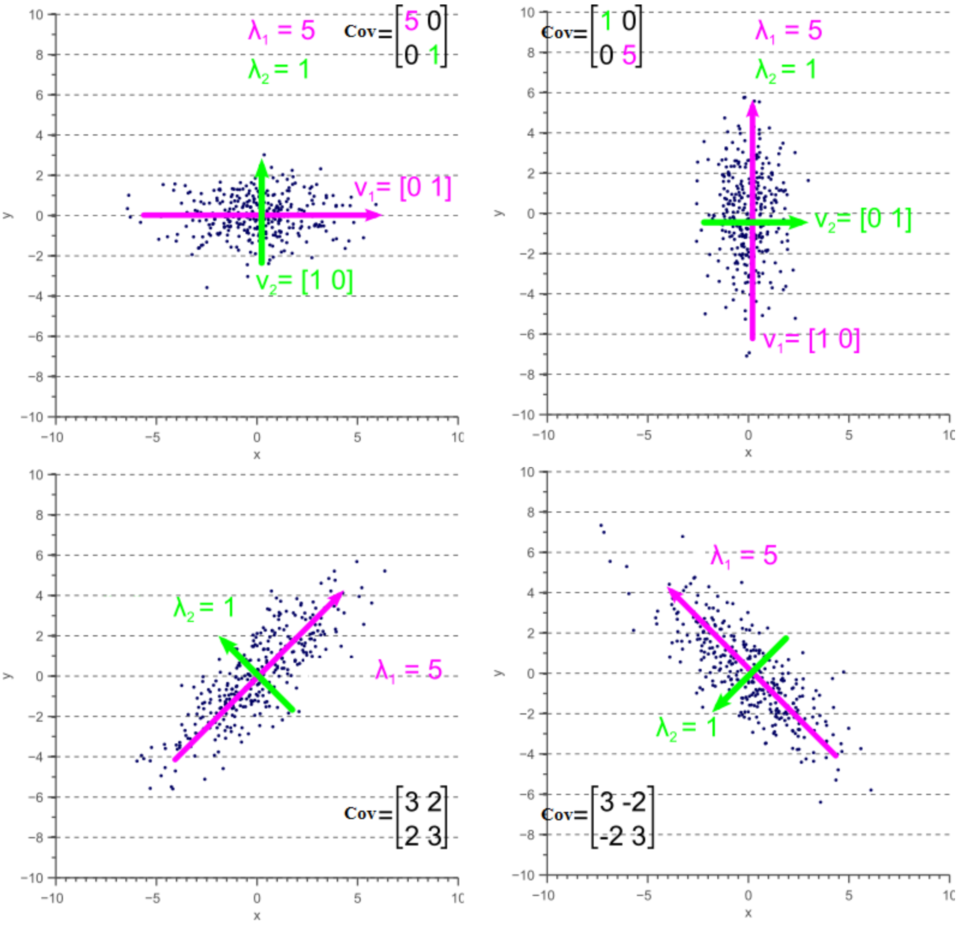


Figure 4.2: Covariance matrix with their corresponding eigenvectors and eigenvalues

Recalling that the covariance matrix for the two variables X and Y is as follows:

$$Cov(x, y) = \begin{bmatrix} Cov(x, x) & Cov(x, y) \\ \\ Cov(y, x) & Cov(y, y) \end{bmatrix}$$

This is equal to :

$$Cov(x, y) = \begin{bmatrix} var(x) & Cov(x, y) \\ \\ Cov(y, x) & var(y) \end{bmatrix}$$

Upon observing the covariance matrices in the top two plots of Figure 4.2, it becomes apparent that cov(x, y) and cov(y, x) are both equal to zero. These covariance matrices are diagonal matrices, in which case the variances are equal to the eigenvalues.

Conversely, when we examine the two bottom plots of Figure 4.2, the covariance matrices are not diagonal. It is noteworthy that the four covariance matrices are not the same, even though the data exhibit the same spread. This is because covariance matrices represent the magnitude of variance along the x-axis and y-axis.

Nonetheless, the eigenvalues still represent the magnitude of variance in the direction of the largest spread of the data. On the other hand, the eigenvectors indicate the direction of the largest spread of the data while maintaining the same magnitude in different orientations.

Now, *what is the relationship between PCA and covariance matrix?*

We will answer this question in the next sections.

4.2.2.1 Principal Axes

PCA is founded on the identification of axes that explain data variance, with the goal of capturing maximum information. While the number of axes matches the number of variables in the dataset, PCA concentrates information within the initial axes. This allows for retaining only the first two or three axes, resulting in dimensionality reduction while preserving more valuable information and enhancing data visualization.

These axes are the **Principal Axes**, and their unit vectors are *the eigenvectors of the covariance matrix of the data.* Principal axes represent the directions of maximum variance in a dataset. The first principal axis corresponds to unit vectors associated with the largest eigenvalue, the second principal axis corresponds to the second-largest eigenvalue, and so forth.
The measure of the proportion of variance explained by axis$_i$ is determined by equation 4.1, and it is often represented as a percentage.

$$PV = \frac{\lambda_i}{\lambda_1 + \lambda_2 + ...\lambda_N} \tag{4.1}$$

Where:
N is the total number of variables.
λ_i is eigenvalue of axis$_i$.
You can also compute the proportion of variance explained by a set of axes of size q (where q < N):

$$PV = \frac{\lambda_1 + ... + \lambda_q}{\lambda_1 + \lambda_2 + ... + \lambda_N} \tag{4.2}$$

4.2.2.2 Principal Components

Principal components are new, uncorrelated variables that correspond to the principal axes. The majority of the information from the original variables is condensed into the first components, which capture the most crucial information in the original dataset, leaving the maximum remaining information for the subsequent components. Figure 4.3 illustrates the explained variance percentages for 7 principal components. It's important to mention that the original dataset consists of 7 variables.

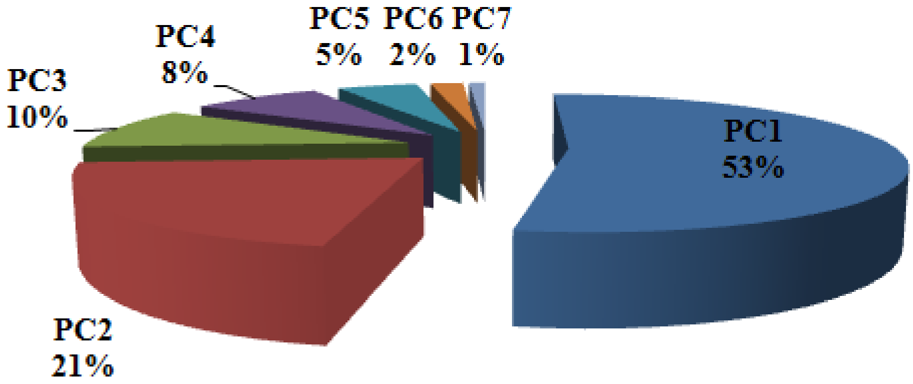


Figure 4.3: The explained variance percentages for 10 principal components

Selecting the Optimal Number of Principal Components

In dimensionality reduction, there are two ways to choose the number of components that you want to retain:

- *Explained Variance Threshold*: You can choose a threshold for the amount of variance you want to retain in your data (e.g., retaining 95% of the variance). You then select the number of principal components that collectively explain at least that much variance.

- *Scree Plot*: Plot the explained variance for each principal component and visually inspect the "Elbow" point on the plot. The number of components before the elbow is often selected.

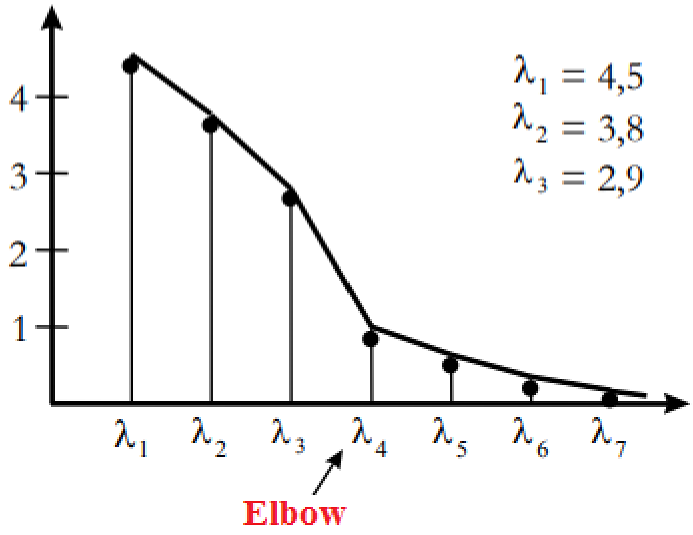


Figure 4.4: The Elbow point on the plot

4.2.3 Representing Data in the New Multidimensional Space

This is the final step, in which we represent data within the new multidimensional space based on the principal axes. To realign the data from its original axes to those represented by the principal components, we must:

1. **Construct the projection matrix:** by stacking the eigenvectors associated with the selected principal components as columns.

2. **Transform the data:** multiply the standardized data by the projection matrix to obtain the final dataset in the reduced-dimensional space. Each row of the transformed dataset represents an observation, and each column represents a principal component.

$$D' = D * PM \tag{4.3}$$

Where:
D' is the new data.

D is the original standardized data.

PM is the projection matrix.

The final dataset obtained after this transformation contains fewer columns (dimensions) than the original dataset but retains most of the essential information.

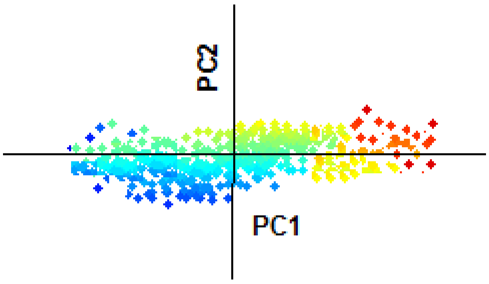


Figure 4.5: Data in the New Space

4.3 Software Tools and Libraries for Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a widely used technique for dimensionality reduction and data analysis. Various software tools and libraries are available to perform PCA.

Among the popular software tools are IBM SPSS (Statistical Package for the Social Sciences link), SAS (Statistical Analysis System link), and Orange, which is an open-source data visualization tool link.

As for libraries, there are Weka link, a Java-based machine learning library; the Eigen C++ library link; FactoMineR, an R package link; and the scikit-learn Python library link.

4.4 Exercises

Exercise 1

The following table presents data for 4 variables. To visualize this data, apply PCA to reduce the variables to two dimensions while retaining more than 80% of the information.

X1	**X2**	**X3**	**X4**
1	2	3	4
5	5	6	7
1	4	2	3
5	3	2	1
8	1	2	2

Table 4.1: Original data with four variables

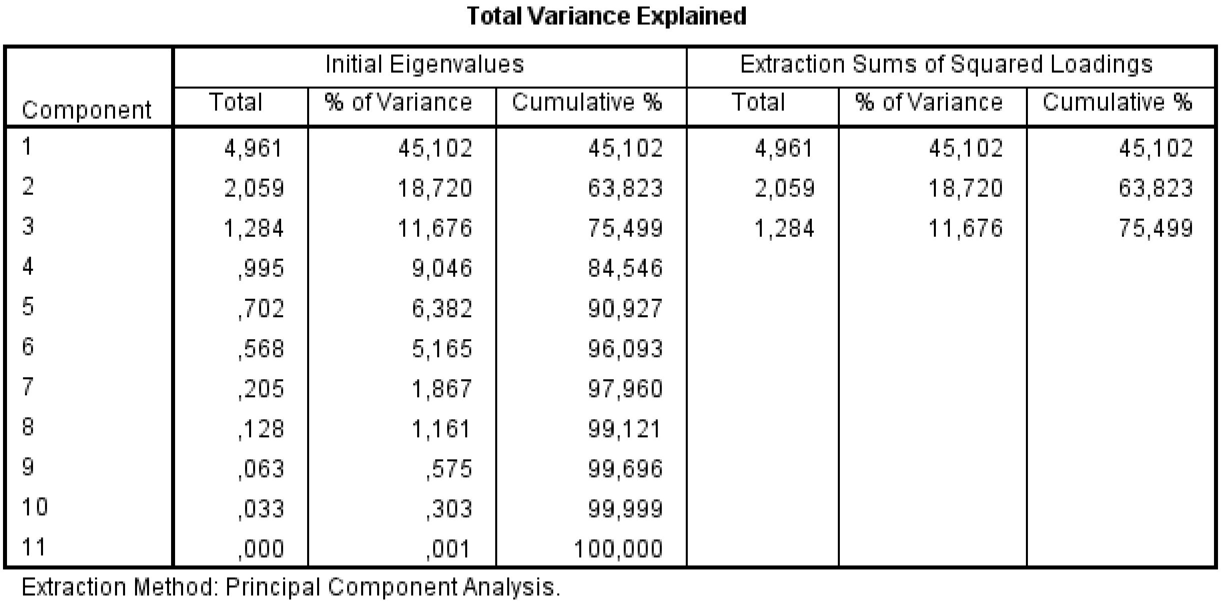
Exercise 2

In a data analysis using SPSS, we have obtained the following values:

1- According to the table, how many variables are present in the original dataset?

After applying PCA, the dataset was reduced to three principal components.

2- What is the percentage of variation explained by each component, and what is the rate of information loss?



Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4,961	45,102	45,102	4,961	45,102	45,102
2	2,059	18,720	63,823	2,059	18,720	63,823
3	1,284	11,676	75,499	1,284	11,676	75,499
4	,995	9,046	84,546			
5	,702	6,382	90,927			
6	,568	5,165	96,093			
7	,205	1,867	97,960			
8	,128	1,161	99,121			
9	,063	,575	99,696			
10	,033	,303	99,999			
11	,000	,001	100,000			

Extraction Method: Principal Component Analysis.

4.5 Solutions

Solution exercise 1

X1	**X2**	**X3**	**X4**
1	2	3	4
5	5	6	7
1	4	2	3
5	3	2	1
8	1	2	2

Table 4.2: Original data with four variables

1. **Data Standardization**
   We implement centering and data normalization using the following formula:
   $$X_{new} = \frac{X - \bar{X}}{S} \tag{4.4}$$
   Where:
   $\bar{X}$ is the mean of variable X.
   S is the standard deviation.

	X1	**X2**	**X3**	**X4**
$\bar{X}_i$	4	3	3	3.40
S_i	3	1.5811	1.7321	2.3022

Table 4.3: Mean and standard deviation of the 4 variables

X1	**X2**	**X3**	**X4**
-1.0000	-0.6325	0	0.2606
0.3333	1.2649	1.7320	1.5637
-1.0000	0.6325	-0.5773	-0.1737
0.3333	0	-0.5773	-1.0425
1.3333	-1.2649	-0.5773	-0.6081

Table 4.4: Original data standardized

2. **Defining the new multidimensional space**
 - We compute the covariance matrix of the original data standardized.

X1	**X2**	**X3**	**X4**
1.0000	-0.3162	0.0481	-0.1810
-0.3162	1.0000	0.6390	0.6181
0.0481	0.6390	1.0000	0.9404
-0.1810	0.6181	0.9404	1.0000

Table 4.5: Covariance Matrix

 - From the covariance matrix, we compute the eigenvalues and the eigenvectors :
 $\lambda_1 = 2.5158$
 $\lambda_2 = 1.0653$
 $\lambda_3 = 0.3939$
 $\lambda_4 = 0.0251$

Their corresponding eigenvectors are:

$$\overrightarrow{v_1} = \begin{pmatrix} -0,1620 \\ 0,5241 \\ 0,5859 \\ 0,5965 \end{pmatrix}$$

$$\overrightarrow{v_2} = \begin{pmatrix} 0.9171 \\ -0.2069 \\ 0.3205 \\ 0.1159 \end{pmatrix}$$

$$\overrightarrow{v_3} = \begin{pmatrix} -0.3071 \\ -0.8173 \\ 0.1882 \\ 0.4497 \end{pmatrix}$$

$$\overrightarrow{v_4} = \begin{pmatrix} -0.1962 \\ -0.1206 \\ 0.7201 \\ -0.6545 \end{pmatrix}$$

So we have 4 principal components:
The first principal component (PC1) explains 62.9% of the variation.
The second principal component (PC2) explains 26.63% of the variation.
The third principal component (PC3) explains 9.85% of the variation.

The fourth principal component (PC4) explains 0.62% of the variation.

3. **Representing data in the new multidimensional space**
   Using only PC1 and PC2, we retain more than 80% of the information (89.53%).
   - The projection matrix (PM) is composed of PC1 and PC2

$$PM = \begin{bmatrix} -0,1620 & 0.9171 \\ 0,5241 & -0.2069 \\ 0,5859 & 0.3205 \\ 0,5965 & 0.1159 \end{bmatrix}$$

   - Transform the data

$$D' = \begin{bmatrix} -1.0000 & -0.6325 & 0 & 0.2606 \\ 0.3333 & 1.2649 & 1.7320 & 1.5637 \\ -1.0000 & 0.6325 & -0.5773 & -0.1737 \\ 0.3333 & 0 & -0.5773 & -1.0425 \\ 1.3333 & -1.2649 & -0.5773 & -0.6081 \end{bmatrix} * \begin{bmatrix} -0,1620 & 0.9171 \\ 0,5241 & -0.2069 \\ 0,5859 & 0.3205 \\ 0,5965 & 0.1159 \end{bmatrix}$$

$$D' = \begin{bmatrix} -0.0140 & -0.7560 \\ 2.5565 & 0.7804 \\ 0.0515 & -1.2531 \\ -1.0141 & -0.0002 \\ -1.5798 & 1.2289 \end{bmatrix}$$

Solution exercise 2

1-From the table, we can see that the number of principal components is 11, which means that the number of variables in the original dataset is also 11.
2- The sum of $\sum \lambda_i$ is 10.998.
PC1, the first principal component, accounts for 45.10% (4.961/10.998) of the variation.
PC2, the second principal component, explains 18.72% (2.059/10.998) of the variation.
PC3, the third principal component, accounts for 11.68% (1.284/10.998) of the variation.
Collectively, these three principal components explain 75.5% of the total variation, resulting in an information loss of 24.5% (100 - 75.5).

4.6 Conclusion

Principal Component Analysis (PCA) is a method used to represent a dataset in a reduced space while minimizing distortion. This space is defined by eigenvectors and eigenvalues. Essentially, PCA is a special case of singular value decomposition where the eigenvectors have unit magnitude. In this chapter, we've explored the foundational concepts of PCA, detailing the method's steps, which include data standardization, determining the new multidimensional space defined by principal axes and principal components, and representing data within this new space. Furthermore, we've introduced software tools and libraries for implementing PCA. The chapter concludes with a series of exercises.

CHAPTER 5

MULTIPLE CORRESPONDENCE ANALYSIS

Multiple Correspondence Analysis (MCA) is multidimensional analysis approaches used to explore and visualize relationships between categorical variables.

MCA aims to simplify the complexity of a dataset by transforming the original variables into a new coordinate system called factor scores

This chapter provides an overview of Multiple Correspondence Analysis (MCA). We will begin by introducing Correspondence Analysis in Section 5.1, followed by an examination of its inner workings in Section 5.2. An Example of Correspondence Analysis (CA) Application is given in Section 5.3. Subsequently, we will explore the connection between Correspondence Analysis and Multiple Correspondence Analysis, including computational specifics, in Section 5.4. To conclude the chapter, we will introduce the software tools and libraries for Multiple Correspondence Analysis (MCA) in Section 5.5.

5.1 Definition of Correspondence Analysis (CA)

Correspondence analysis belongs to a broad class of methods based on singular value decomposition. It is essentially a generalized version of principal component analysis, specifically designed for qualitative data analysis. While initially developed for contingency tables, Correspondence Analysis has demonstrated its versatility and is now frequently applied to various other types of data tables (Abdi and Béra, 2014).
CA operates by computing a set of orthogonal axes, referred to as dimensions or factors, in a way that maximizes the variance of the data. It then maps the categories of the two categorical variables onto these dimensions.

5.2 Understanding the Inner Workings of CA

The main goal of CA is to map the dataset onto the row and column profiles, thus aiding interpretation.
The CA method is based on three key steps:
1. Data preparation
2. Defining the new multidimensional space
3. Representing the data in the new multidimensional space

5.2.1 Data preparation

The initial step involves constructing the cross-table, or contingency table, that summarizes the joint distribution of the two categorical variables.

5.2.2 Defining the new multidimensional space

The axes that define the new multidimensional space are determined using singular value decomposition.
Considering the following notation:
N: Data matrix (I×J) where $N_{ij} \geq 0$.
n: is the total of N, $\sum^i \sum^j N_{ij}$
P: is correspondence matrix $P = n^{-1}N$, where each element of N is devided

by the total of N.
r: is the row masses r=P1 (the notation 1 is used for a vector of ones of length) **ie** $r_i = \sum_{j=1}^{J} P_{ij}$, it is a vector that counts the total number for each row.
c: is the column masses $c = P^T 1$ **ie** $c_j = \sum_{i=1}^{I} P_{ij}$, it is a vector that counts the total number for each column.
Diagonal matrices of row and column masses: $D_r = diag(r)$ and $D_c = diag(c)$.

5.2.2.1 Computation Profiles Matrices

CA is particularly interested in the marginal matrices of N, computed from row and column sums, called profiles. There are matrix row profiles (X) and matrix column profiles (Y) that aid in the interpretation of N.

$$X_{ij} = \frac{P_{ij}}{ri} \tag{5.1}$$

$$Y_{ij} = \frac{P_{ij}}{cj} \tag{5.2}$$

5.2.3 Representing the data in the new multidimensional space

The computational algorithm to obtain coordinates of the row and column profiles with respect to principal axes, using the singular value decomposition (SVD), is as follows:

- Calculate the matrix S of standarized residuales:

$$S = D_r^{-\frac{1}{2}}(P - rc^T)D_c^{-\frac{1}{2}} \tag{5.3}$$

- Calculate the SVD of S:

$$SVD(S) = UD_\alpha V^T \tag{5.4}$$

Where
$U^TU = V^TV = I$
D_α is the diagonal matrix of singular values in descending order $\alpha_1 \geq \alpha_2 \geq ..$

- Principal inertias λ_K are defnied as

$$\lambda_K = \alpha_K^2 \tag{5.5}$$

Where
k=1,2,...K and K = min $\{I-1, J-1\}$

- Standard coordinates of rows ϕ :

$$\phi = D_r^{-\frac{1}{2}}U \tag{5.6}$$

- Standard coordinates of columns γ:

$$\gamma = D_c^{-\frac{1}{2}}V \tag{5.7}$$

- Principal coordinates F of rows are defined according to principal inertias:

$$F = D_r^{-\frac{1}{2}}UD_\alpha \tag{5.8}$$

- Principal coordinates G of columns are defined according to principal inertias:

$$G = D_c^{-\frac{1}{2}}VD_\alpha \tag{5.9}$$

The SVD is matrix decomposition expresses any rectangular matrix as a product of three matrices of simple structure, as equation 5.4. The columns of the matrices U and V are the left and right singular vectors respectively, and the positive values α_k down the diagonal of D_{α_k}, in descending order, are the singular values.

The rows of the coordinate matrices from equation 5.6 to equation 5.9 refer to the rows or columns, while the columns of these matrices refer to the principal axes, or dimensions, of which there are min $\{I-1, J-1\}$.

5.2.3.1 Computation of Total Inertia

The inertia is the sum of squares of the singular values, i.e., the sum of the eigenvalues:

$$inertia = \sum_{k=1}^{K} \alpha_K^2 = \sum_{k=1}^{K} \lambda_K \tag{5.10}$$

The total inertia is also defined by the sum of squares of the matrix S:

$$inertia = trace(SS^T) = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(p_{ij} - r_i c_j)^2}{r_i c_j} = n^{-1}\chi^2 \tag{5.11}$$

5.2.3.2 Contributions of points to principal inertias

The contributions of the row and columns points to the inertia on the k-th dimension are the inertia components:
For row i:

$$\frac{r_i f_{ik}^2}{\lambda_k} = r_i \phi_{jk}^2 \tag{5.12}$$

For column j:

$$\frac{c_j g_{jk}^2}{\lambda_k} = c_j \gamma_{jk}^2 \tag{5.13}$$

5.2.3.3 Contributions of principal axes to point inertias (squared correlations)

The contributions of the dimensions to the inertia of the i-th row and j-th column points is the squared cosines or squared correlations :
For row i:

$$\frac{f_{ik}^2}{\sum_k f_{ik}^2} \tag{5.14}$$

For column j:

$$\frac{g_{ik}^2}{\sum_k g_{ik}^2} \tag{5.15}$$

5.3 An Example of Correspondence Analysis (CA) Application

We consider the following example (see Table 5.1) of a statistical study on the factors that prevent having many children, conducted with a different number of individuals with varying levels of education.

The factors (conditions) can be: money, future, unemployment, decision, difficult, economic, selfishness, occupation, finances, war, housing, fear, health, work.

The levels of education: no degree, elementary school, trade school, high school and college.

The results of the analysis are provided using SPSS software. The initial step involves examining the row and column profiles.

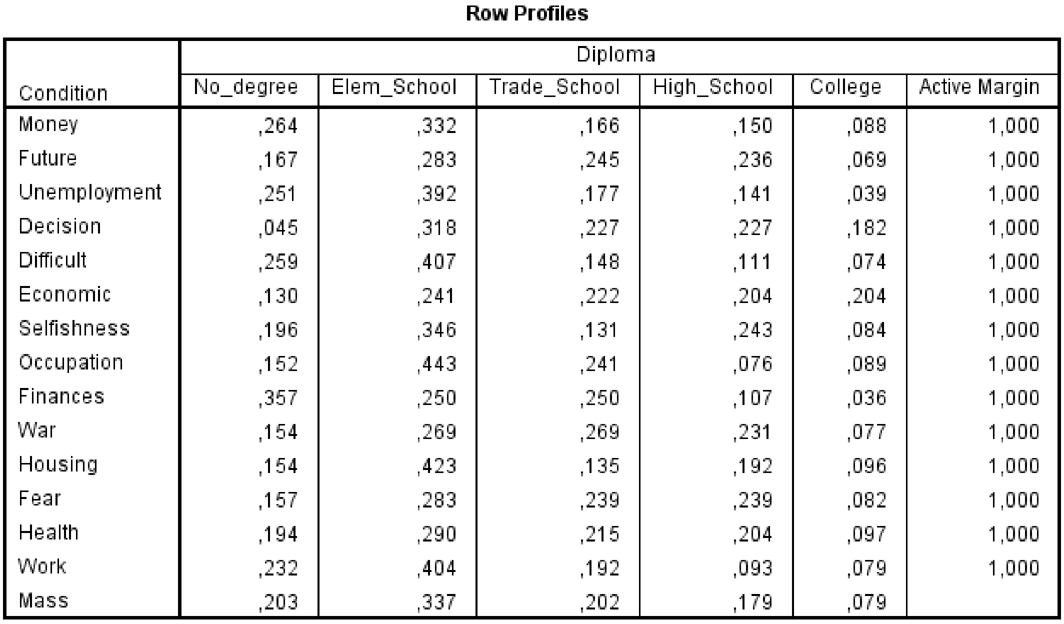
Row profiles describe the information related to the variable in the row, i.e., the conditions (money, future, etc.).
Comparing two row profiles will help us understand how the respective patterns are associated with different levels of education.
For example, in row profiles presented in Figure 5.1, the factor 'money' was considered by individuals without a diploma at a rate of 26.4%, compared to individuals with only an elementary school level at 33.2%, 16.6% for those with a trade school education, 15% for those with a high school education, and finally, 8.8% for individuals with a college degree. As we can observe, this information cannot be extracted directly from the correspondence table.

Column profiles describe the information related to the variable in the column, i.e., levels of education (no degree, elementary school, etc.). Comparing two column profiles informs us about the proximities existing between different diploma categories.
For instance, in Column profiles presented in Figure 5.2, among individuals without a diploma, motivations for not having many children vary: 15.8% cite financial concerns, 16.4% express worries about the future, 22% fear unemployment, 0.3% attribute it to decision, 2.2% mention difficulty, 2.2% point to economic factors, 6.5% cite selfishness, 3.7% link it to occupation,

3.1% mention financial issues, 1.2% attribute it to war, 2.5% blame housing, 7.7% connect it to fear, 5.6% relate it to health, and 10.8% attribute it to work.

	No degree	**Elem School**	**Trade School**	**High School**	**College**	**Active Margin**
Money	51	64	32	29	17	**193**
Future	53	90	78	75	22	**318**
Unemployment	71	111	50	40	11	**283**
Decision	1	7	5	5	4	**22**
Difficult	7	11	4	3	2	**27**
Economic	7	13	12	11	11	**54**
Selfishness	21	37	14	26	9	**107**
Occupation	12	35	19	6	7	**79**
Finances	10	7	7	3	1	**28**
War	4	7	7	6	2	**26**
Housing	8	22	7	10	5	**52**
Fear	25	45	38	38	13	**159**
Health	18	27	20	19	9	**93**
Work	35	61	29	14	12	**151**
Active Margin	**323**	**537**	**322**	**285**	**125**	**1592**

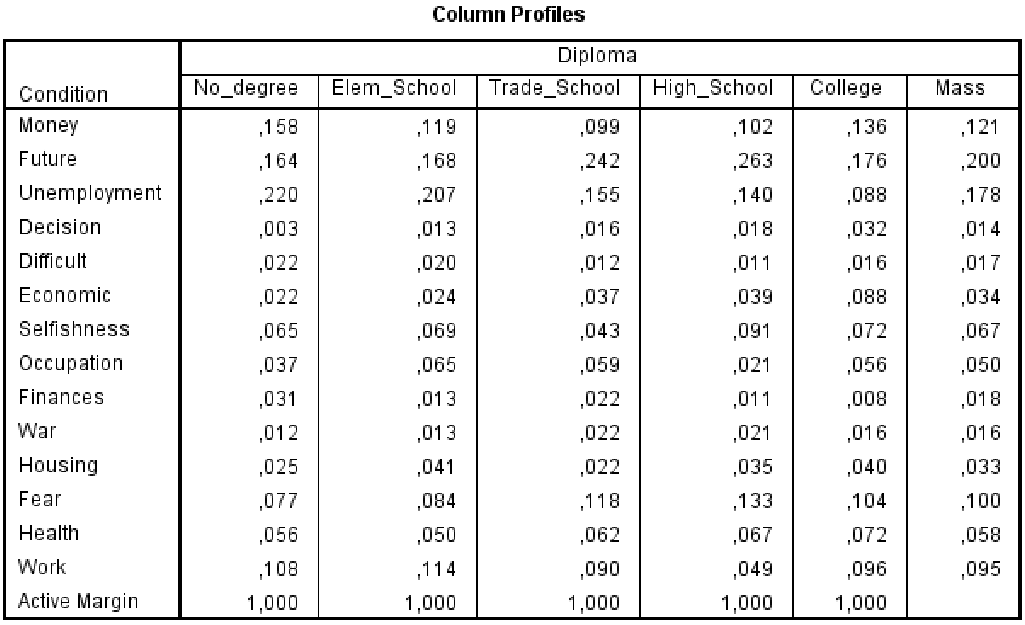
Table 5.1: Correspondence Table



Row Profiles

Condition	Diploma					
	No_degree	Elem_School	Trade_School	High_School	College	Active Margin
Money	,264	,332	,166	,150	,088	1,000
Future	,167	,283	,245	,236	,069	1,000
Unemployment	,251	,392	,177	,141	,039	1,000
Decision	,045	,318	,227	,227	,182	1,000
Difficult	,259	,407	,148	,111	,074	1,000
Economic	,130	,241	,222	,204	,204	1,000
Selfishness	,196	,346	,131	,243	,084	1,000
Occupation	,152	,443	,241	,076	,089	1,000
Finances	,357	,250	,250	,107	,036	1,000
War	,154	,269	,269	,231	,077	1,000
Housing	,154	,423	,135	,192	,096	1,000
Fear	,157	,283	,239	,239	,082	1,000
Health	,194	,290	,215	,204	,097	1,000
Work	,232	,404	,192	,093	,079	1,000
Mass	,203	,337	,202	,179	,079	

Figure 5.1: Row Profiles



Column Profiles

Condition	Diploma					
	No_degree	Elem_School	Trade_School	High_School	College	Mass
Money	,158	,119	,099	,102	,136	,121
Future	,164	,168	,242	,263	,176	,200
Unemployment	,220	,207	,155	,140	,088	,178
Decision	,003	,013	,016	,018	,032	,014
Difficult	,022	,020	,012	,011	,016	,017
Economic	,022	,024	,037	,039	,088	,034
Selfishness	,065	,069	,043	,091	,072	,067
Occupation	,037	,065	,059	,021	,056	,050
Finances	,031	,013	,022	,011	,008	,018
War	,012	,013	,022	,021	,016	,016
Housing	,025	,041	,022	,035	,040	,033
Fear	,077	,084	,118	,133	,104	,100
Health	,056	,050	,062	,067	,072	,058
Work	,108	,114	,090	,049	,096	,095
Active Margin	1,000	1,000	1,000	1,000	1,000	

Figure 5.2: Column Profiles

Representing the data in the new multidimensional space

The principal inertias λ_k and the singular value α_K^2 are présented in Figure 5.3

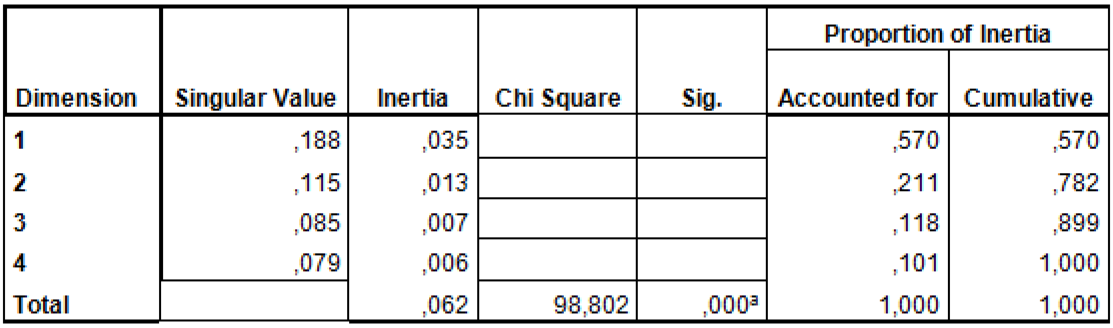
$$\lambda_K = \alpha_K^2$$

Where
k=1,2,...K and K = min $\{I-1, J-1\}$
In this case, $I = 14$ and $J = 5$
So K =min $\{14-1, 5-1\}$ that mean K=4. The proportion of inertia for λ_k is defined as

$$\frac{\lambda_k}{\sum^K \lambda_k}$$

The diagonal matrix of singular values $D_\alpha$is:

$$D_\alpha = \begin{bmatrix} 0.188 & 0 & 0 & 0 \\ 0 & 0.115 & 0 & 0 \\ 0 & 0 & 0.085 & 0 \\ 0 & 0 & 0 & 0.079 \end{bmatrix}$$

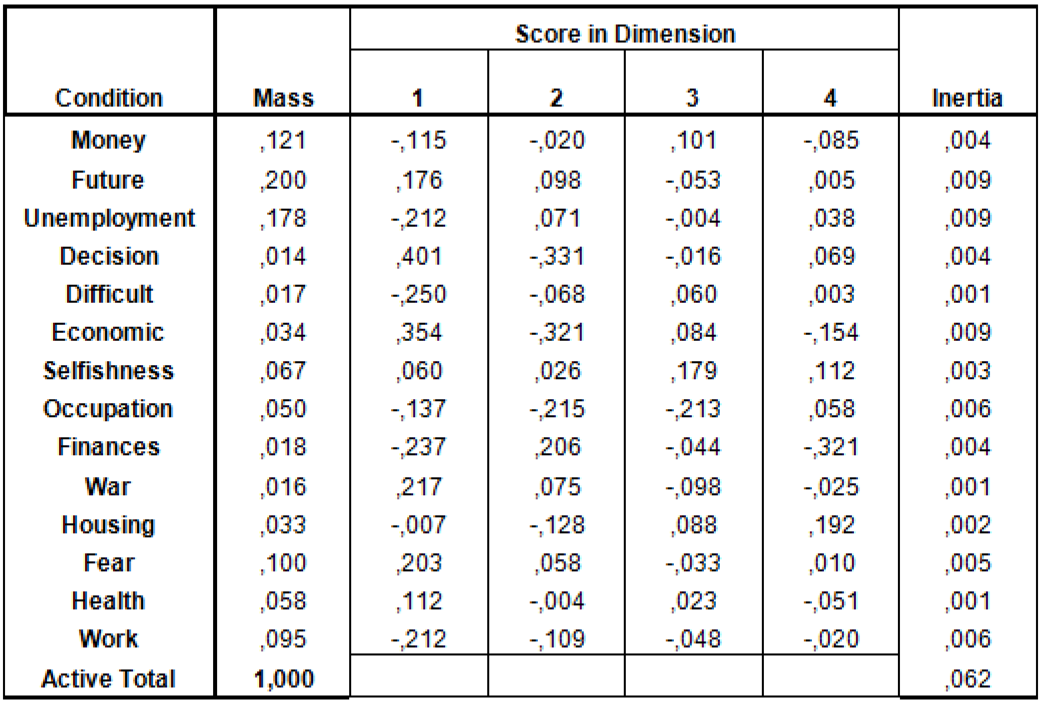


Dimension	Singular Value	Inertia	Chi Square	Sig.	Proportion of Inertia	
					Accounted for	Cumulative
1	,188	,035			,570	,570
2	,115	,013			,211	,782
3	,085	,007			,118	,899
4	,079	,006			,101	1,000
Total		,062	98,802	,000a	1,000	1,000

Figure 5.3: Inertias and Singular Value

Principal coordinates F of rows are defined according to principal inertias (see equation 5.3 and 5.4):

$$F = D_r^{-\frac{1}{2}} U D_\alpha$$

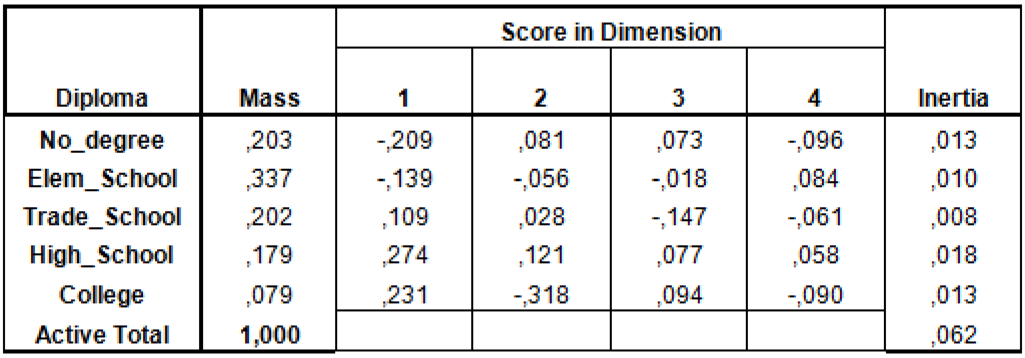


Condition	Mass	Score in Dimension				Inertia
		1	2	3	4	
Money	,121	-,115	-,020	,101	-,085	,004
Future	,200	,176	,098	-,053	,005	,009
Unemployment	,178	-,212	,071	-,004	,038	,009
Decision	,014	,401	-,331	-,016	,069	,004
Difficult	,017	-,250	-,068	,060	,003	,001
Economic	,034	,354	-,321	,084	-,154	,009
Selfishness	,067	,060	,026	,179	,112	,003
Occupation	,050	-,137	-,215	-,213	,058	,006
Finances	,018	-,237	,206	-,044	-,321	,004
War	,016	,217	,075	-,098	-,025	,001
Housing	,033	-,007	-,128	,088	,192	,002
Fear	,100	,203	,058	-,033	,010	,005
Health	,058	,112	-,004	,023	-,051	,001
Work	,095	-,212	-,109	-,048	-,020	,006
Active Total	1,000					,062

Figure 5.4: Principal coordinates F of rows

Principal coordinates G of columns are defined according to principal inertias (see equation 5.3 and 5.4):

$$G = D_c^{-\frac{1}{2}} V D_\alpha$$



Diploma	Mass	Score in Dimension				Inertia
		1	2	3	4	
No_degree	,203	-,209	,081	,073	-,096	,013
Elem_School	,337	-,139	-,056	-,018	,084	,010
Trade_School	,202	,109	,028	-,147	-,061	,008
High_School	,179	,274	,121	,077	,058	,018
College	,079	,231	-,318	,094	-,090	,013
Active Total	**1,000**					,062

Figure 5.5: Principal coordinates G of columns

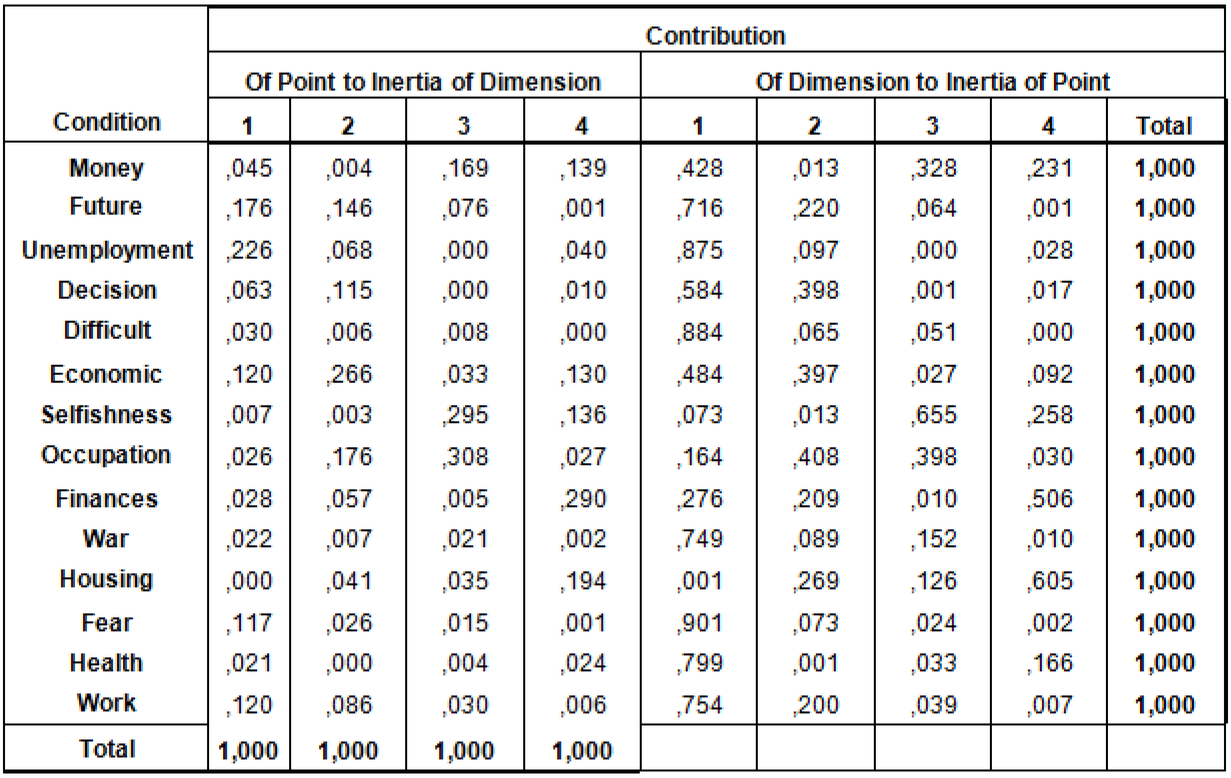
Contributions of row points to principal inertias and contributions of principal axes to row point inertias

The contributions of the i-th row points to the inertia:

$$\frac{r_i f_{ik}^2}{\lambda_k} = r_i \phi_{jk}^2$$

The contributions of the dimensions to the inertia of the i-th row points:

$$\frac{f_{ik}^2}{\sum_k f_{ik}^2}$$



Condition	Contribution								
	Of Point to Inertia of Dimension				Of Dimension to Inertia of Point				
	1	2	3	4	1	2	3	4	Total
Money	,045	,004	,169	,139	,428	,013	,328	,231	1,000
Future	,176	,146	,076	,001	,716	,220	,064	,001	1,000
Unemployment	,226	,068	,000	,040	,875	,097	,000	,028	1,000
Decision	,063	,115	,000	,010	,584	,398	,001	,017	1,000
Difficult	,030	,006	,008	,000	,884	,065	,051	,000	1,000
Economic	,120	,266	,033	,130	,484	,397	,027	,092	1,000
Selfishness	,007	,003	,295	,136	,073	,013	,655	,258	1,000
Occupation	,026	,176	,308	,027	,164	,408	,398	,030	1,000
Finances	,028	,057	,005	,290	,276	,209	,010	,506	1,000
War	,022	,007	,021	,002	,749	,089	,152	,010	1,000
Housing	,000	,041	,035	,194	,001	,269	,126	,605	1,000
Fear	,117	,026	,015	,001	,901	,073	,024	,002	1,000
Health	,021	,000	,004	,024	,799	,001	,033	,166	1,000
Work	,120	,086	,030	,006	,754	,200	,039	,007	1,000
Total	1,000	1,000	1,000	1,000					

Figure 5.6: Contributions According to Row Points

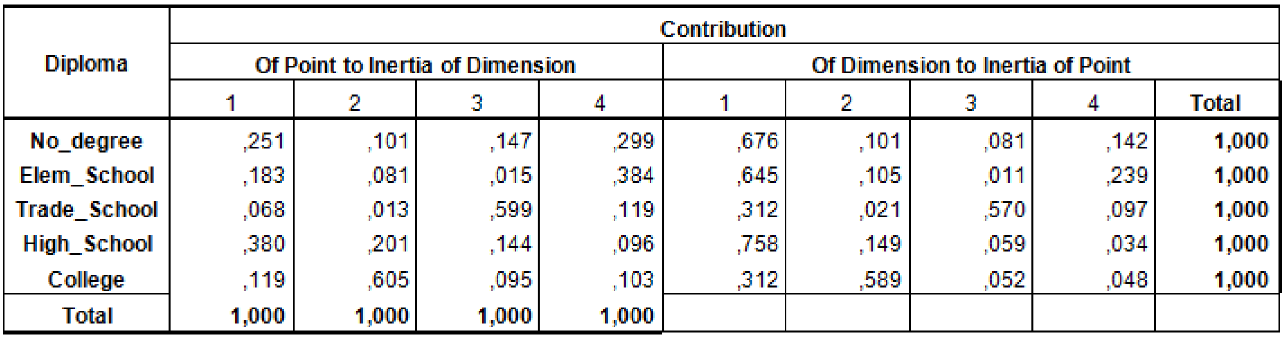
Contributions of column points to principal inertias and contributions of principal axes to column point inertias

The contributions of the j-th column points to the inertia:

$$\frac{c_j g_{jk}^2}{\lambda_k} = c_j \gamma_{jk}^2$$

The contributions of the dimensions to the inertia of the j-th column points:

$$\frac{g_{ik}^2}{\sum_k g_{ik}^2}$$



Diploma	Contribution								
	Of Point to Inertia of Dimension				Of Dimension to Inertia of Point				
	1	2	3	4	1	2	3	4	Total
No_degree	,251	,101	,147	,299	,676	,101	,081	,142	**1,000**
Elem_School	,183	,081	,015	,384	,645	,105	,011	,239	**1,000**
Trade_School	,068	,013	,599	,119	,312	,021	,570	,097	**1,000**
High_School	,380	,201	,144	,096	,758	,149	,059	,034	**1,000**
College	,119	,605	,095	,103	,312	,589	,052	,048	**1,000**
Total	**1,000**	**1,000**	**1,000**	**1,000**					

Figure 5.7: Contributions According to Column Points

In order to gain a better understanding of the relationship between the condition variable and the education level variable, we have represented the data in a new space, focusing on just two dimensions.

Figure 5.8 illustrates that individuals without a diploma tend to have fewer children due to factors like unemployment and financial constraints. Those with an elementary school education cite reasons such as money, difficulties, work, housing, and occupation. Individuals who have completed trade school or high school express nearly identical concerns, including health, future prospects, war, fear, and selfishness. However, individuals holding a college diploma are predominantly influenced by economic and decision-related conditions.

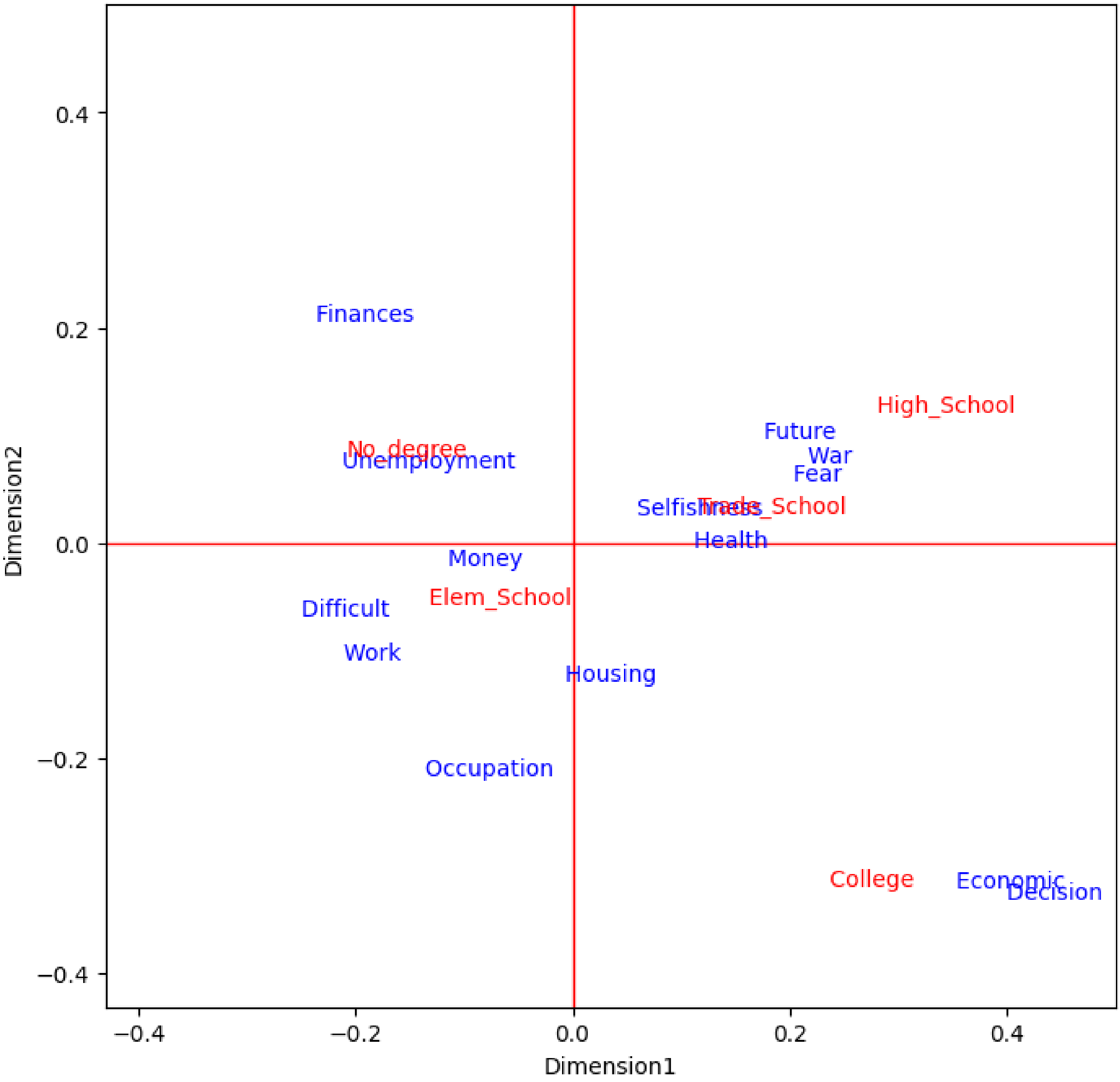


Figure 5.8: 2D Data Visualization in the New Dimensional Space

5.4 From Correspondence Analysis (CA) to Multiple Correspondence Analysis (MCA)

Multiple Correspondence Analysis (MCA) is an advanced statistical technique that extends the capabilities of Correspondence Analysis (CA). It is specifically designed to analyze the complex patterns of relationships that exist among several categorical dependent variables. MCA can also be used for data reduction and dimensionality reduction, which can be helpful in simplifying complex data sets.

While CA is suitable for analyzing two-way contingency tables, MCA can analyze multi-dimensional contingency tables involving three or more categorical variables.

To apply MCA, the data must be presented in a specific matrix format that allows us to subsequently apply CA. These matrices are the indicator matrix and the Burt matrix.
The standard coordinates of the categories are identical in the two versions of MCA.

5.4.1 Indicator Matrix

The indicator matrix **Z** is generated by transforming the data, organized as cases-by-variables, into binary variables.

In the context of a data matrix with N cases and Q categorical variables, if the q-th variable has J_q categories, this variable will be presented by
$J = \sum^{q} J_q$ columns.
Next, the indicator matrix Z, which consists of N cases and J categories, serves as the input for the CA algorithm.

Exemple:

	Gender	**Nationality**	**Eye Color**
1	Male	Algerian	Blue
2	Female	Foreigner	Brown
3	Female	Foreigner	Black
4	Male	Foreigner	Blue
5	Female	Algerian	Brown
6	Male	Algerian	Black

Table 5.2: Initial Table

Gender	**Nationality**	**Eye Color**
1	1	1
2	2	2
2	2	3
1	2	1
2	1	2
1	1	3

Table 5.3: Modality coding

Male	**Female**	**Algerian**	**Foreigner**	**Bleu**	**Brown**	**Black**
1	0	1	0	1	0	0
0	1	0	1	0	1	0
0	1	0	1	0	0	1
1	0	0	1	1	0	0
0	1	1	0	0	1	0
1	0	1	0	0	0	1

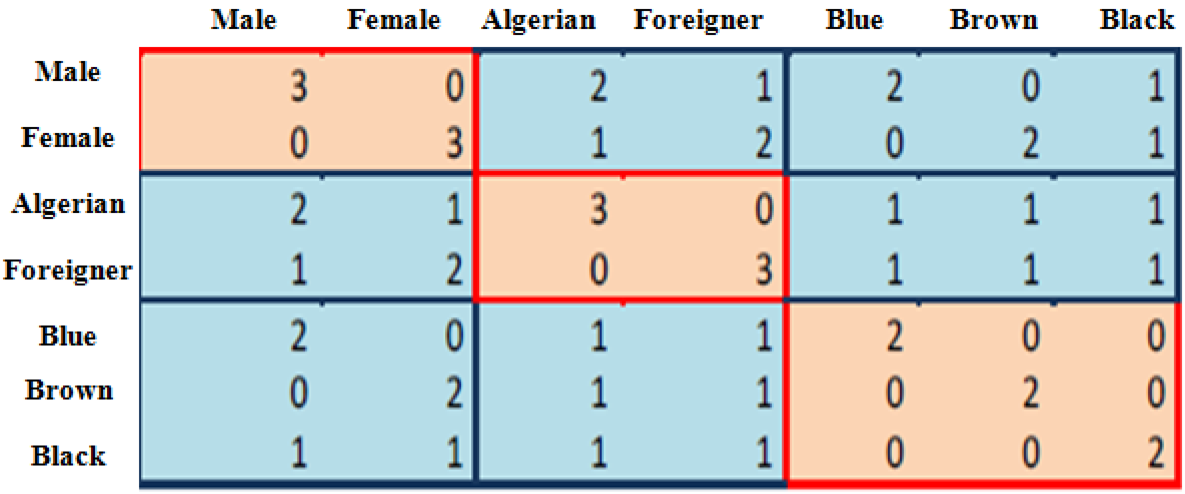
Table 5.4: The indicator Matrix Z

5.4.2 Burt Matrix

The Burt matrix $B = Z^T Z$ of all two-way cross-tabulations of the Q variables. Then the Burt version of MCA is the application of the basic CA algorithm to the matrix B, resulting in coordinates for the J categories (B is a symmetric matrix).

Male	1	0	0	1	0	1
Female	0	1	1	0	1	0
Algerian	1	0	0	0	1	1
Foreigner	0	1	1	1	0	0
Bleu	1	0	0	1	0	0
Brown	0	1	0	0	1	0
Black	0	0	1	0	0	1

Table 5.5: The indicator Matrix Z'



	Male	Female	Algerian	Foreigner	Blue	Brown	Black
Male	3	0	2	1	2	0	1
Female	0	3	1	2	0	2	1
Algerian	2	1	3	0	1	1	1
Foreigner	1	2	0	3	1	1	1
Blue	2	0	1	1	2	0	0
Brown	0	2	1	1	0	2	0
Black	1	1	1	1	0	0	2

Figure 5.9: Burt Matrix

5.5 Software Tools and Libraries for Multiple Correspondence Analysis (MCA)

MCA is a frequently employed method for reducing dimensionality and analyzing data. Numerous software tools and libraries are accessible for conducting MCA.

Popular software tools for this purpose include IBM SPSS (Statistical Package for the Social Sciences), SAS (Statistical Analysis System), Jamovi, which is an open-source statistical software, and Factoshiny, a web-based application.

When it comes to libraries, there are FactoMineR, ca, and MASS as R packages, as well as the scikit-learn and Prince Python libraries.

5.6 Conclusion

Correspondence Analysis (CA) and Multiple Correspondence Analysis (MCA) are statistical techniques employed for investigating and visualizing relationships among categorical variables. In this chapter, we've introduced both methods. We commenced with Correspondence Analysis (CA), delineating its procedural steps, encompassing data preparation and the establishment of a new multidimensional space via the computation of profile matrices. Furthermore, we elaborated on the data representation within this new space, covering the computation of total inertia, points to principal inertias, and the contribution of principal axes to point inertias. An example illustrating CA application was provided. Subsequently, we elucidated the transition from Correspondence Analysis (CA) to Multiple Correspondence Analysis, introducing the indicator matrix and Burt matrix. The chapter concluded with an overview of the software tools and libraries utilized for MCA.

CHAPTER 6

CLASSIFICATION IN MACHINE LEARNING

Industries generate vast and diverse datasets, which can be a daunting task to process manually. As a result, machine learning classification techniques have found increasing application in enhancing data management, ultimately leading to improved return on investment.

In this chapter, we will delve into the realm of machine learning, exploring the fundamental concepts in Section 6.1. Subsequently, in Section 6.2, we will introduce the concept of classification within the domain of machine learning, emphasizing its significance. We will then pivot to Section 6.3, where we will illuminate the practical utilization of classification techniques in various industrial applications.

Our journey will also encompass an exploration of the different types of classification tasks in machine learning, providing valuable insights in Section 6.4. Next, we will provide an overview of the different classification algorithms in Section 6.5 and discuss the performance metrics used to assess the quality of classification algorithms in Section 6.6. Finally, we will conclude the chapter with a set of exercises in Section 6.7.

6.1 What is Machine Learning ?

Machine learning (ML), a subfield of artificial intelligence (AI) and computer science, focuses on the exploration of data and algorithms that mimic human learning processes. In essence, it entails machines progressively enhancing their accuracy through the learning process.

Machine learning has become an integral part of diverse industries. Its significance lies in its ability to furnish organizations with valuable insights into customer behavior trends and operational patterns.

6.1.1 Types of Machine Learning

There are four types of machine learning:

- Supervised Learning
- Unsupervised Learning
- Semi-Supervised Learning
- Reinforcement Learning

6.1.1.1 Supervised Learning

In this scenario, machine learning operates on labeled data, meaning that each data point is associated with a class. During the training process, the algorithm strives to discern patterns and relationships among the inputs in order to predict their corresponding outputs. As input data is fed into the model, it refines its weights iteratively until the model is appropriately fitted.

6.1.1.2 Unsupervised Learning

Machine learning operates on unlabeled data, meaning that each data point lacks a class label. In such cases, the algorithm analyzes the available data to uncover correlations and connections without requiring human intervention. The output of the model is the organization of the input data into clusters.

6.1.1.3 Semi-Supervised Learning

Semi-supervised learning bridges supervised learning and unsupervised learning. It employs both labelled and unlabelled data. The model can learn to categorise unlabelled data using this combination.

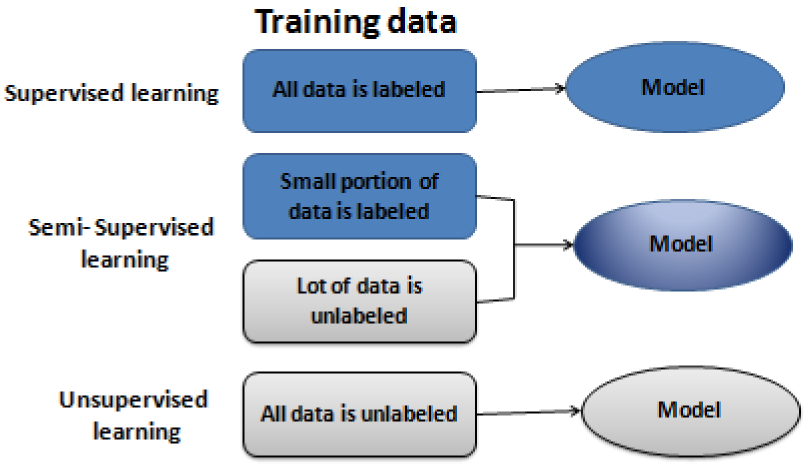


Figure 6.1: Supervised Learning VS Unsupervised Learning VS Semi-Supervised Learning

6.1.1.4 Reinforcement Learning

Is a machine learning training method based on rewarding desired behaviors and punishing undesired ones. Following the definition of the rules, the method attempts to explore several options and prospects, monitoring and assessing each output to determine which is ideal. Reinforcement learning instructs the machine through trial and error. It learns from previous experiences and begins to change its approach to the situation to reach the best possible outcome.

6.2 Understanding Classification in Machine Learning: A Definition

Classification, in the context of machine learning, is a technique in which a model categorizes samples into specific classes or categories. This process involves thorough training of the model using training data to capture patterns and relationships within the dataset. Once trained, the model can effectively assign the correct class labels to new, unseen data points.

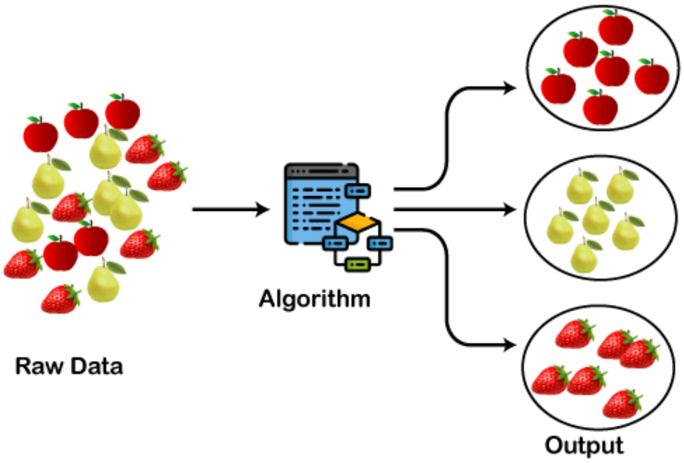


Figure 6.2: Classification

6.3 Classification in Machine Learning with Industrial Applications

The Industrial applications involves the use of classification methods to categorize and assign items, components, or data points to specific classes or categories. This enables automated decision-making, quality control, fault detection, and optimization in various industrial sectors, such as manufacturing, supply chain management, quality assurance, and predictive maintenance.

6.4 Different Types of Classification Tasks in Machine Learning

There are four types of classification tasks :

- Binary Classification
- Multi-Class Classification
- Multi-Label Classification
- Imbalanced Classification

6.4.1 Binary Classification

Binary classification stands as a fundamental task in the realm of machine learning. Its objective is to classify data into one of two possible classes, guided by input features.

Binary classification finds applications in industries like manufacturing, where it can be utilized to monitor the health of equipment and machinery. Models can classify equipment as either "healthy" or "faulty".

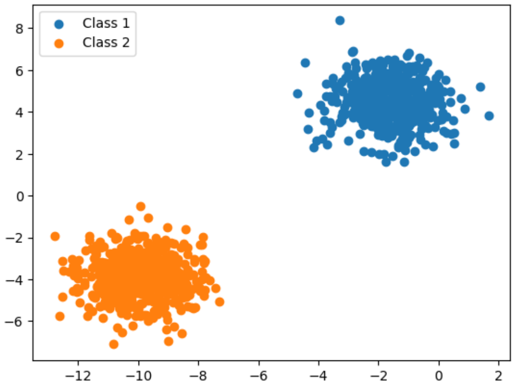


Figure 6.3: Binary Classification

6.4.2 Multi-Class Classification

Multi-class classification is a machine learning modeling task where the goal is to categorize data into one of more than two possible classes.
Industries like retail and marketing use multi-class classification to segment customers into different groups based on purchasing behavior, demographics, or preferences.

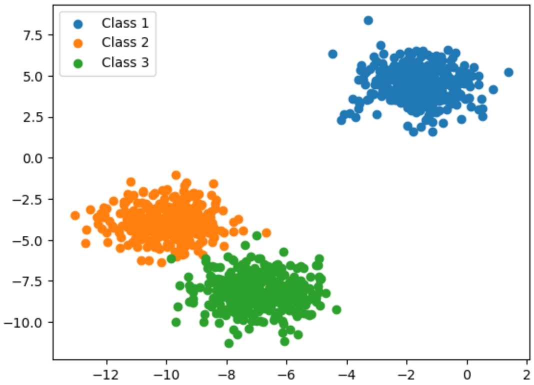


Figure 6.4: Multi-Class Classification: three classes

6.4.3 Multi-Label Classification

When data can be associated with multiple classes, multi-label classification is employed.
This differs from binary classification and multi-class classification, where a single class label is assigned to each sample.
For example multi-label classification can categorize products based on several quality parameters, such as "size," "color," "shape," and "material".

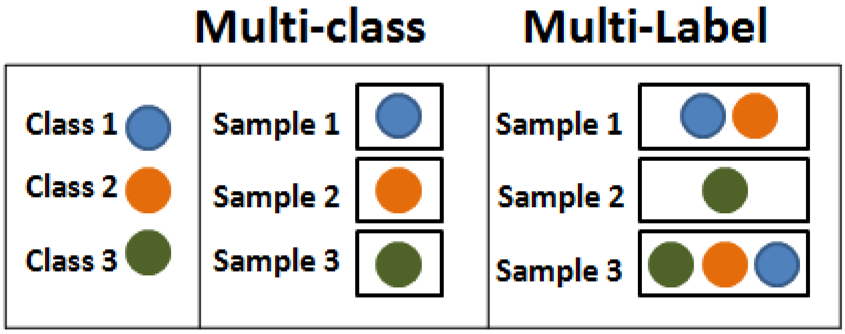


Figure 6.5: Multi-Class Classification VS Multi-Label Classification

6.4.4 Imbalanced Classification

Imbalanced classification refers to a particular challenge in machine learning, characterized by an unequal distribution of classes within the dataset. In this scenario, the number of examples in each class is unevenly distributed.
For example, in the context of manufacturing, the majority of products are typically free from defects. Therefore, the crucial task is to detect and identify rare defects or anomalies to ensure and uphold product quality.

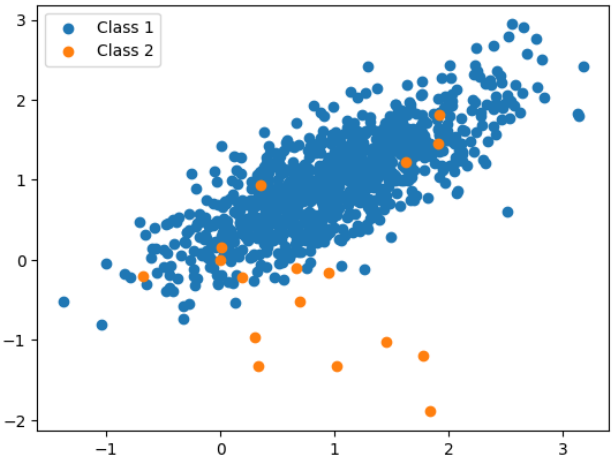


Figure 6.6: Imbalanced Classification

6.5 Instances of Classification Algorithms

There are several instances of classification algorithms, each with its own characteristics and suitability for different types of problems. Here are some common types of classification algorithms:

- Logistic Regression
- Decision Trees
- Random Forest
- Support Vector Machines (SVM)
- k-Nearest Neighbors (k-NN)
- Naive Bayes

6.5.1 Logistic Regression

Logistic Regression is a statistical method used in machine learning to model the relationship between a binary dependent variable and one or more independent variables.

The logistic regression model estimates the probability that a given input instance belongs to a specific class. The result is a value between 0 and 1. The logistic regression model uses the logistic function (also called the sigmoid function) to transform a linear combination of the predictor variables into a probability. The formula for logistic regression can be expressed as:

$$Logit(p_i = 1/(1 + exp(-pi))) \tag{6.1}$$

$$ln(p_i/1 - p_i) = Beta\_0 + Beta\_1 * X\_1 + ... + Beta\_K * X\_K \tag{6.2}$$

In this logistic regression equation, logit(pi) is the dependent or response variable and x is the independent variable. The beta parameter, or coefficient, in this model is commonly estimated via maximum likelihood estimation (MLE). This method tests different values of Beta through multiple iterations to optimize for the best fit of log odds. All of these iterations produce the log likelihood function, and logistic regression seeks to maximize this function to find the best parameter estimate. Once the optimal coefficient (or coefficients if there is more than one independent variable) is found, the conditional probabilities for each observation can be calculated, logged, and summed together to yield a predicted probability. For binary classification, a probability less than .5 will predict 0 while a probability greater than 0 will predict 1.

6.5.2 Decision Trees

A decision tree is a non-parametric supervised learning algorithm. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes.

Decision tree learning employs a divide and conquer strategy by conducting a greedy search to identify the optimal split points within a tree. This process of splitting is then repeated in a top-down, recursive manner until all, or the majority of records have been classified under specific class labels.

There are multiple ways to select the best attribute at each node, two methods, information gain and Gini impurity, act as popular splitting criterion for decision tree models.

$$Entropy(S) = -\sum_{c \in C} p_c log_2 p_c \tag{6.3}$$

Where

S represents the data set that entropy is calculated.

c represents the classes in set S.

p_c represents the proportion of data points that belong to class c to the number of total data points in set S.

$$Information - Gain(S, a) = Entropy(S) - \sum_{v \in Values(a)} \frac{|S_v|}{|S|} Entropy(S_v) \tag{6.4}$$

Where

a represents a specific attribute or class label

Entropy(S) is the entropy of dataset S

$\frac{|S_v|}{|S|}$ represents the proportion of the values in S_v to the number of values in dataset

Entropy(S_v) is the entropy of dataset S_v.

As Entropy(S) is fixed for a given S, independent of the splitting attribute a, maximising $Information - Gain(S, a)$ is equivalent to minimising

$$E = \sum_{v \in Values(a)} \frac{|S_v|}{|S|} Entropy(S_v) \tag{6.5}$$

Gini impurity is defined as:

$$Gini - Impurty = 1 - \sum_{c} (p_c^2) \tag{6.6}$$

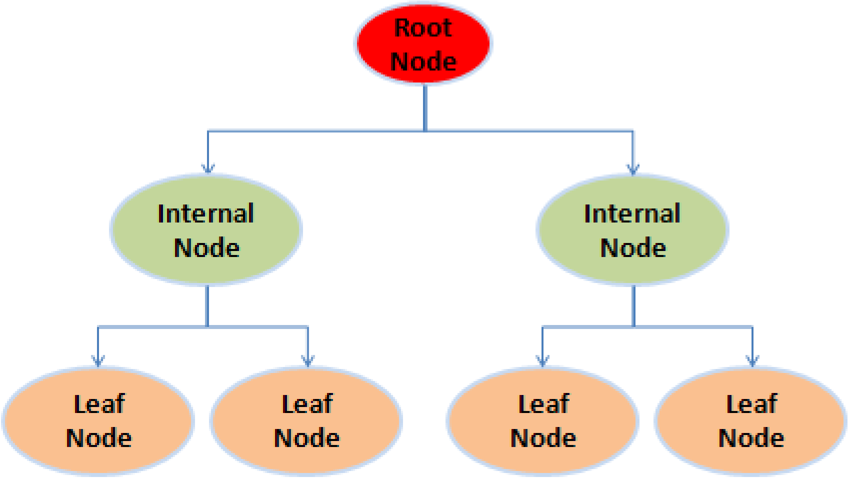


Figure 6.7: Decision tree

6.5.3 Random Forest

The random forest algorithm consists of an ensemble of decision trees, with each tree in the ensemble trained on a data sample drawn from the training set with replacement, a process known as bootstrapping. Within this training sample, one-third of the data is reserved as test data, referred to as the out-of-bag (oob) sample. Additional randomness is introduced through feature bagging, enhancing dataset diversity and reducing correlations among individual decision trees.

For classification tasks, the predicted class is determined by a majority vote, which selects the most frequently occurring categorical variable among the trees.

Subsequently, the oob sample is used for cross-validation, finalizing the prediction process.

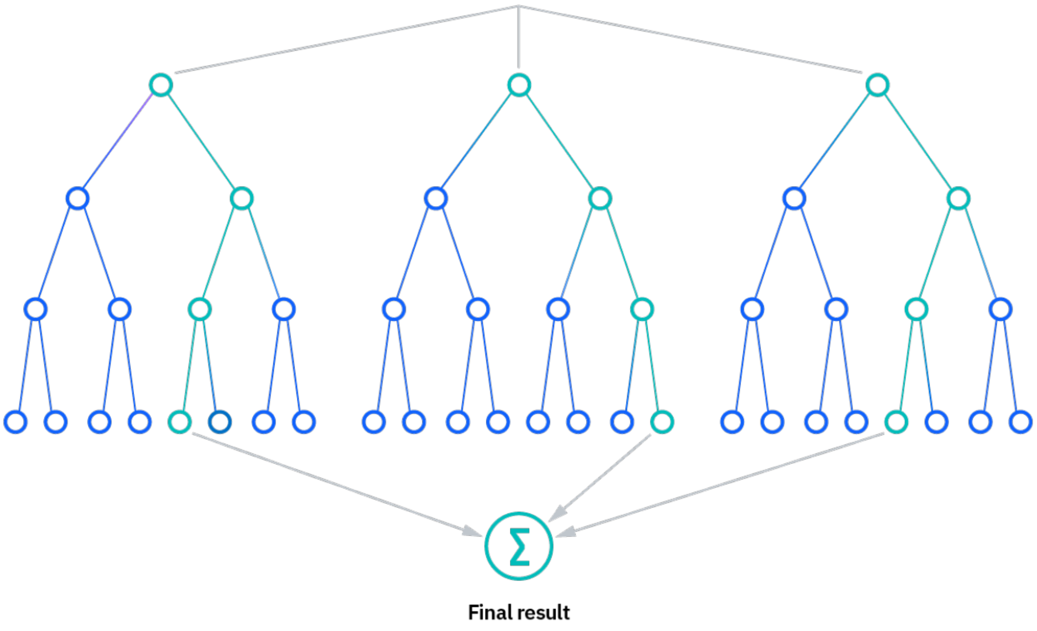


Figure 6.8: Random Forest

6.5.4 Support Vector Machines (SVM)

Support Vector Machines is a supervised machine learning algorithm that aims to find the optimal hyperplane (decision boundary) that maximally separates data points from different classes in a feature space. It does this by identifying the support vectors, which are the data points closest to the decision boundary, and finding the hyperplane that maximizes the margin between these support vectors.

SVM can handle non-linear classification problems by using kernel functions (e.g., polynomial, radial basis function) to map the data into a higher-dimensional space, where a linear hyperplane can separate the classes.

SVM can be extended to multi-class classification problems using various techniques, including one-vs-one and one-vs-all strategies.

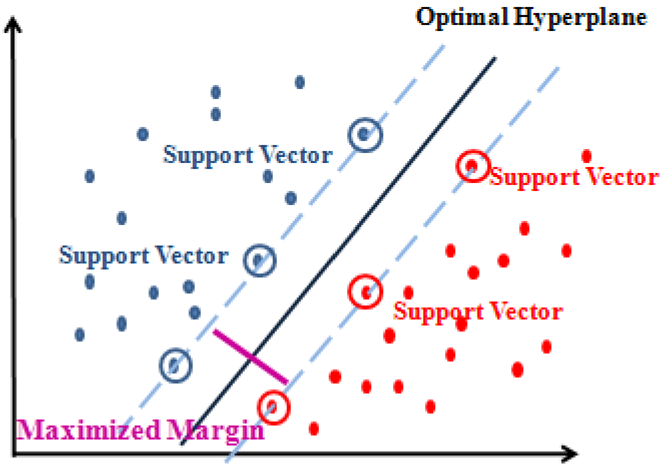


Figure 6.9: Support Vector Machines (SVM)

6.5.5 Naive Bayes

A naive Bayes classifier is an algorithm that uses Bayes' theorem to classify objects. Naive Bayes classifiers assume strong, or naive, independence between attributes of data points. A naive Bayes classifier uses probability theory to classify data. The key insight of Bayes' theorem is that the probability of an event can be adjusted as new data is introduced.

6.5.6 k-Nearest Neighbors (k-NN)

The k-nearest neighbors algorithm, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. It is typically used as a classification algorithm, working off the assumption that similar points can be found near one another.

For classification problems, a class label is assigned on the basis of a majority vote—i.e. the label that is most frequently represented around a given data point is used.

In order to determine which data points are closest to a given query point, the distance between the query point and the other data points will need to be calculated.

This distance can be Euclidean distance, Manhattan distance, Minkowski distance, Hamming distance.

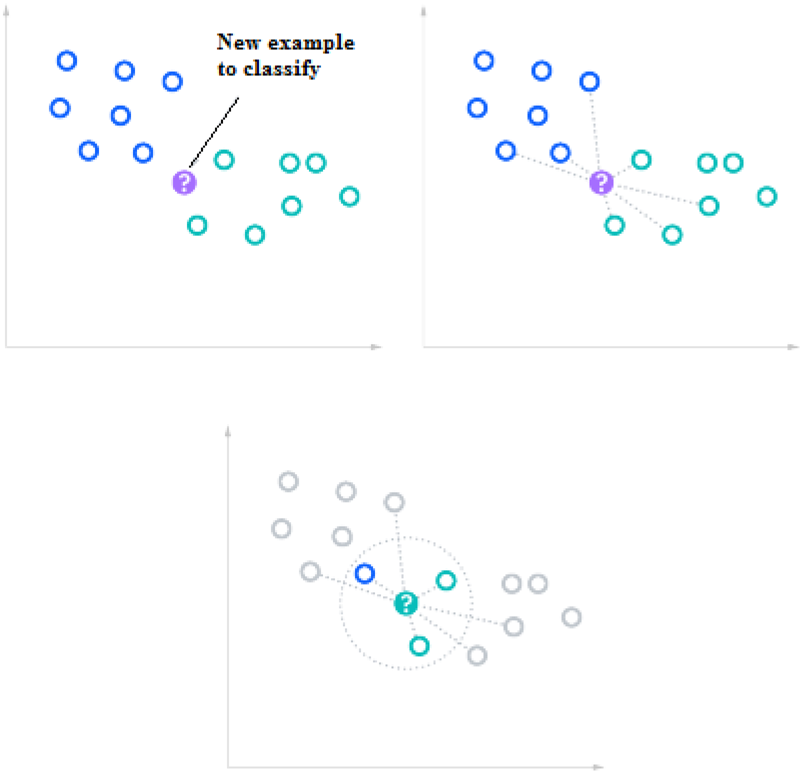


Figure 6.10: k-Nearest Neighbors (k-NN)

6.6 Performance Metrics for Assessing Classification Algorithm Quality

There are several metrics commonly used to evaluate the quality of classification algorithms.
Here are some widely used classification evaluation metrics:

- Accuracy
- Precision
- Recall
- F1 Score
- Specificity
- Cohen's Kappa
- Receiver Operating Characteristic (ROC) Curve
- Area Under the ROC Curve (AUC-ROC)

6.6.1 Accuracy

This is one of the most straightforward metrics. It measures the ratio of correctly predicted instances to the total number of instances in the dataset.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6.7}$$

Where:
TP refers to True Positive
TN refers to True Negative
FP refers to False Positive
FN refers to False Negative

6.6.2 Precision

Precision is the ratio of true positive predictions to the total positive predictions. It indicates how many of the predicted positive instances are actually correct.

$$Precision = \frac{TP}{TP + FP} \tag{6.8}$$

6.6.3 Recall

Recall measures the ratio of true positive predictions to the total actual positive instances. It tells you how many of the actual positive instances were correctly predicted.

$$Recall = \frac{TP}{TP + FN} \tag{6.9}$$

6.6.4 F1 Score

The F1 Score is the harmonic mean of precision and recall. It is useful when you want to balance precision and recall, especially when there is an imbalance between the classes.

$$F1score = 2 \times \frac{recall \times precision}{recall + precision} \tag{6.10}$$

6.6.5 Specificity

Specificity measures the ratio of true negative predictions to the total actual negative instances.

$$Specificity = \frac{TN}{TN + FP} \tag{6.11}$$

6.6.6 Cohen's Kappa

Kappa Statistic is based on the difference between how much agreement is actually present ("observed"
agreement-Po) compared to how much agreement would be expected to be

present by chance alone (“expected”
agreement-Pe).

$$Kappa = \frac{(Po - Pe)}{(1 - Pe)} \tag{6.12}$$

6.6.7 Receiver Operating Characteristic (ROC) Curve

The ROC curve is a graphical representation of the trade-off between sensitivity and specificity. It helps you choose an appropriate threshold for classification.

6.6.8 Area Under the ROC Curve (AUC-ROC)

AUC-ROC quantifies the overall performance of a classification model. A higher AUC-ROC indicates a better model.

6.7 Exercises

Exercise 1

Given the dataset distribution presented in the following figure:

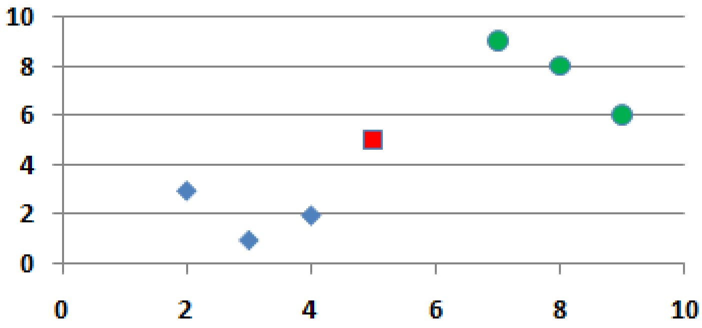


Figure 6.11: Dataset distribution

-Perform k-Nearest Neighbors classification using Euclidean distance with k=3 for the red point (5, 5) to determine its class.

Exercise 2

Provided is a dataset illustrating the probability of machine breakdown (High or Low), considering variables such as the duration since the last repair (1-5 years, 5-10 years, >10 years), the gender of the maintenance expert involved (Male or Female), and the installation zone (Zone 1 or Zone 2).

-Using information gain, build a decision tree from this data.

ID	Duration	Gender	Zone	Risk
1	>10	M	1	L
2	5−10	M	2	H
3	1−5	F	2	L
4	>10	F	2	H
5	1−5	M	2	H
6	>10	M	2	H
7	5−10	F	1	L
8	5−10	M	1	L

Table 6.1: Dataset

6.8 Solutions

Solution exercise 1

From the figure we obtained the coordinate of each point.

X-coordinate	Y-coordinate	Class
2	3	Class 1
3	1	Class 1
4	2	Class 1
7	9	Class 2
9	6	Class 2
8	8	Class 2

Table 6.2: Data points coordinates

Distance from (5, 5) to (2, 3): $sqrt((5-2)^2 + (5-3)^2) \approx 3.61$

Distance from (5, 5) to (3, 1): $sqrt((5-3)^2 + (5-1)^2) \approx 4.47$

Distance from (5, 5) to (4, 2): $sqrt((5-4)^2 + (5-2)^2) \approx 3.16$

Distance from (5, 5) to (7, 9): $sqrt((5-7)^2 + (5-9)^2) \approx 4.47$

Distance from (5, 5) to (9, 6): $sqrt((5-9)^2 + (5-6)^2) \approx 4.12$

Distance from (5, 5) to (8, 8): $sqrt((5-8)^2 + (5-8)^2) \approx 4.24$

The three nearest neighbors are: (2, 3), (4, 2), and (9, 6).

Count the number of neighbors for each class:

Class 1: 2 neighbors

Class 2: 1 neighbor

Since Class 1 has a higher count, we classify the point (5, 5) as belonging to Class 1. Therefore, the predicted class for the point (5, 5) using k-Nearest Neighbors with k=3 is Class 1.

Solution exercise 2

We compute E according to 6.5 for the variable Duration

$$E = \frac{3}{8}.0.918 + \frac{3}{8}0.918 + \frac{2}{8}.1 \simeq 0.94$$

Duration	$\lvert S_v \rvert$	**Risk**	p_v	**Entropy**(S_v)
>10	3	L	1/3	0.918
		H	2/3	
5−10	3	L	2/3	0.918
		H	1/3	
1−5	2	L	1/2	1
		H	1/2	

Table 6.3: Entropy Computation for Variable Duration

We compute E according to 6.5 for the variable Gender

$$E = \frac{5}{8}.0.971 + \frac{3}{8}.0.918 \simeq 0.95$$

Gender	$\lvert S_v \rvert$	**Risk**	p_v	**Entropy**(S_v)
M	5	L	2/5	0.971
		H	3/5	
F	3	L	2/3	0.918
		H	1/3	

Table 6.4: Entropy Computation for Variable Gender

We compute E according to 6.5 for the variable Zone

$$E = \frac{3}{8}.0 + \frac{5}{8}.0.722 \simeq 0.45$$

Zone	$\lvert S_v \rvert$	**Risk**	p_v	**Entropy**(S_v)
1	3	L	3/3	0
		H	0/3	
2	5	L	1/5	0.722
		H	4/5	

Table 6.5: Entropy Computation for Variable Zone

As Zone yields the lowest entropy, resulting in the highest information gain, it is selected for splitting. The branch for Zone 1 is already pure, so no further processing is required. The branch of Zone 2 contains the following data:

ID	**Duration**	**Gender**	**Risk**
2	5–10	M	H
3	1–5	F	L
4	>10	F	H
5	1–5	M	H
6	>10	M	H

We compute again E according to 6.5 for the variable Duration

$$E = \frac{2}{5}.0 + \frac{1}{5}.0 + \frac{2}{5}.1 \simeq 0.4$$

We compute again E according to 6.5 for the variable Gender

$$E = \frac{3}{5}.0 + \frac{2}{5}.1 \simeq 0.4$$

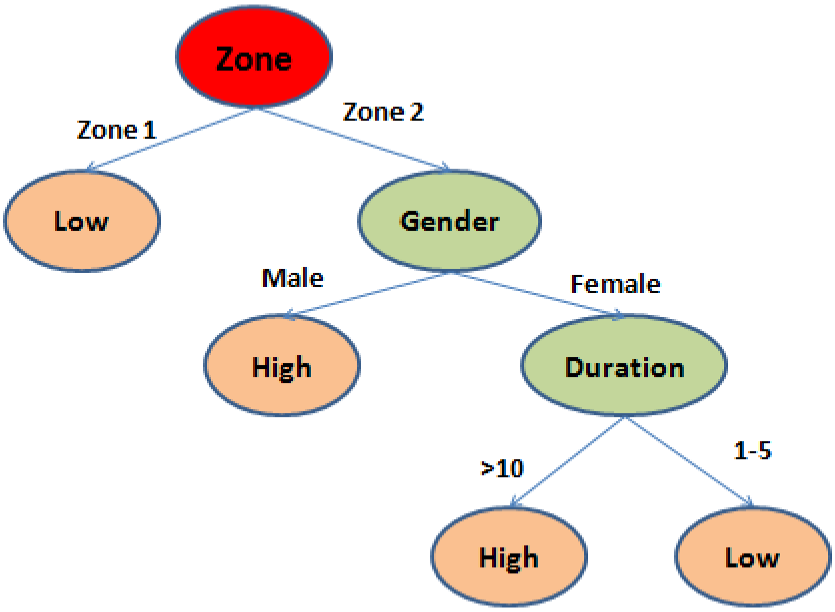
Duration	$\lvert S_v \rvert$	**Risk**	p_v	**Entropy**(S_v)
>10	2	L	0/2	0
		H	2/2	
5−10	1	L	0/1	0
		H	1/1	
1−5	2	L	1/2	1
		H	1/2	

Table 6.6: Entropy Calculation for Variable Duration - Second Iteration

Gender	$\lvert S_v \rvert$	**Risk**	p_v	**Entropy**(S_v)
M	3	L	0/3	0
		H	3/3	
F	2	L	1/2	1
		H	1/2	

Table 6.7: Entropy Computation for Variable Gender - Second Iteration

Finally, We'll opt for Gender as an arbitrary choice. Now, we're left with only one non-pure branch, which is 'female.' We can further split it using Duration. The resulting final tree is shown below:



6.9 Conclusion

Machine learning plays a crucial role in data analysis within industries. In this chapter, we've outlined the fundamental concepts of machine learning. We began by introducing the most common types of machine learning: supervised, unsupervised, semi-supervised, and reinforcement learning. Following this, we provided illustrative diagrams to explain classification in machine learning and offered examples of classification in industrial applications. We then delved into various classification scenarios including binary classification, multi-class classification, multi-label classification, and imbalanced classification. Moreover, we discussed several classification algorithms such as logistic regression, decision trees, random forest, support vector machines, naive Bayes, and k-nearest neighbors. Furthermore, we covered metrics for evaluating classification algorithms including accuracy, precision, recall, F1 score, kappa, ROC, and AUC-ROC. Finally, we wrapped up this chapter with a series of exercises.

CHAPTER 7

CLUSTERING IN MACHINE LEARNING

In our final chapter, we will delve into the intriguing world of clustering within the realm of machine learning. We embark on this journey by building a solid understanding of this concept in Section 7.1. Following that, in Section 7.2, we illuminate the practical significance of clustering through industrial applications.

Section 7.3 is dedicated to unraveling the different types of clustering tasks that machine learning encompasses, providing clarity on their distinctions. Moving forward, both Section 7.4 and Section 7.5 offer valuable insights, with the former presenting an array of clustering algorithms and the latter focusing on the metrics employed to assess the quality of these algorithms.

As we approach the conclusion of this chapter, we turn our attention to Section 7.6, where we introduce some popular libraries tailored for implementing clustering algorithms, equipping you with the tools necessary to embark on your own clustering adventures. Section 7.7 provides a set of exercises.

7.1 Understanding Clustering in Machine Learning: A Definition

Cluster analysis, a machine learning technique falling under unsupervised learning, is concerned with grouping unlabeled datasets. Its primary objective is to reveal underlying patterns, structures, or relationships within a dataset, all without prior knowledge of predefined groups or categories. This process involves assigning a cluster-ID to each cluster or group formed. This cluster-ID can then be utilized by machine learning systems to streamline the processing of large and intricate datasets.

7.2 Clustering in Machine Learning with Industrial Applications

Cluster analysis, a vital technique in the realm of machine learning, finds wide-ranging industrial applications that significantly benefit organizations. One of the most prominent applications is customer segmentation for businesses. By employing clustering algorithms, companies can categorize their customers into groups with similar purchasing behaviors. This segmentation proves invaluable for targeted marketing efforts, personalized product recommendations, and enhancing overall customer satisfaction. For instance, a retail company can utilize cluster analysis to divide its customer base into distinct clusters based on their shopping habits, allowing the company to tailor its marketing strategies accordingly.

Another compelling application of clustering is in optimizing supply chain management. Clustering can be employed to group products or suppliers based on various factors such as demand patterns, lead times, or transportation costs. This strategic approach leads to more efficient inventory management and substantial cost reduction. By identifying clusters of products or suppliers that share similar characteristics or requirements, organizations can streamline their logistics and procurement processes, ultimately improving their bottom line.

7.3 Different Types of Clustering Tasks in Machine Learning

There are several types of clustering tasks in machine learning, each with its own specific objectives and characteristics. Here are some of the main types:

- Partitioning Clustering (Exclusive)
- Density-Based Clustering
- Distribution Model-Based Clustering
- Hierarchical Clustering
- Fuzzy Clustering (Probabilistic)
- Spectral Clustering

7.3.1 Partitioning Clustering (Exclusive)

Partitioning clustering, also known as exclusive clustering, is a type of data clustering method where each data point belongs exclusively to one and only one cluster. The goal of partitioning clustering algorithms is to divide a dataset into non-overlapping clusters in such a way that the data points within each cluster are more similar to each other.

Partitioning clustering algorithms work to create cluster centers in such a manner that the intra-cluster distances (i.e., the distances between data points within the same cluster) are minimized, while inter-cluster distances (i.e., the distances between data points from different clusters) are maximized. This process results in a partitioning of the data into non-overlapping clusters, where each data point exclusively belongs to one cluster, facilitating the identification of meaningful patterns and groupings within the dataset.

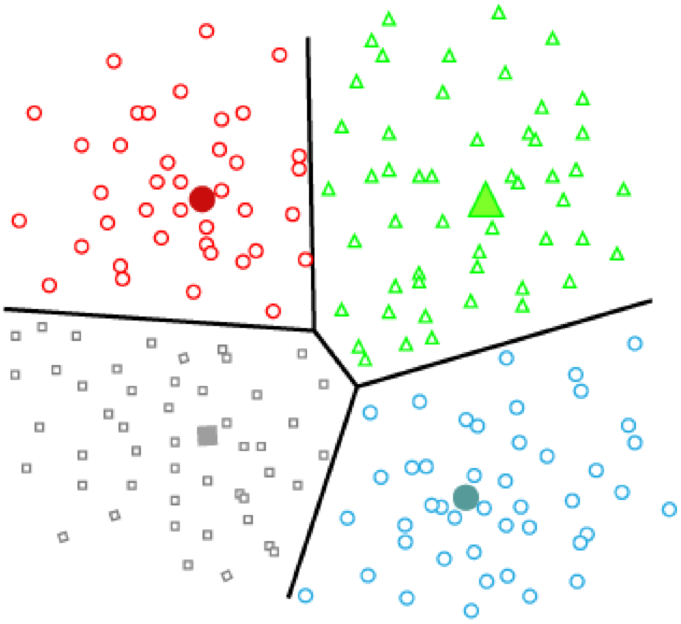


Figure 7.1: Partitioning Clustering (Exclusive)

7.3.2 Density-Based Clustering

Density-based clustering is a method that aims to group data points by identifying highly dense regions and connecting them into clusters. This approach allows for the formation of clusters with arbitrarily shaped distributions, as long as dense regions can be linked together. The algorithm achieves this by discerning distinct clusters within the dataset and connecting areas of high density.

These algorithms may encounter challenges when dealing with datasets that exhibit varying densities across the data space and when working with high-dimensional data, as the concept of density becomes more complex in higher dimensions.

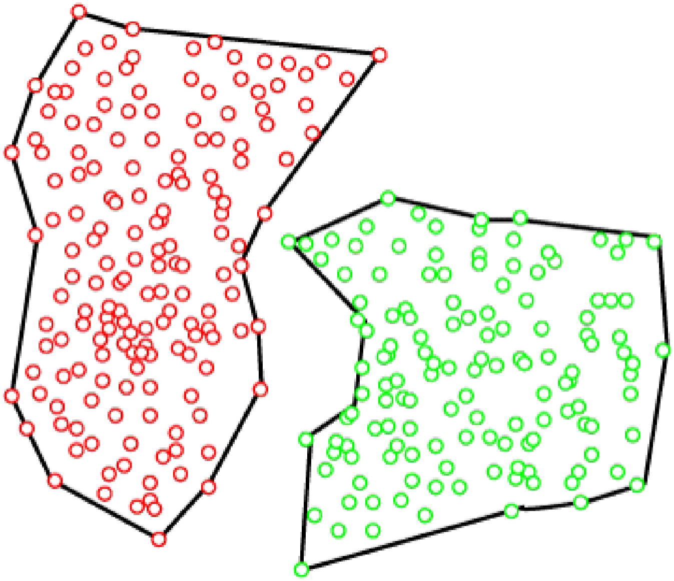


Figure 7.2: Density-Based Clustering

7.3.3 Distribution Model-Based Clustering

Distribution model-based clustering is an approach in data analysis that operates under the assumption that data points within each cluster adhere to a specific probability distribution or statistical model. In this method, clusters are established by assessing how closely the observed data conforms to the presumed distribution models. The primary objective is to determine the distribution parameters, such as mean and variance, that offer the best fit for the data within each cluster.

A couple of noteworthy examples of distribution model-based clustering encompass Gaussian Mixture Models (GMMs) and Multinomial Mixture Models. GMMs posit that each cluster adheres to a Gaussian (normal) distribution, making them suitable for continuous data. In contrast, Multinomial Mixture Models find application in clustering categorical data and operate on the assumption of a multinomial distribution within each cluster.

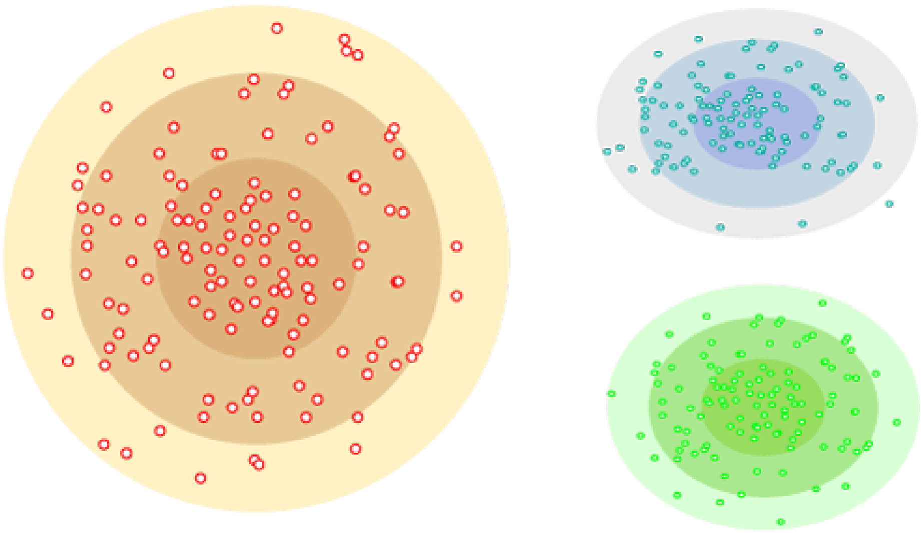


Figure 7.3: Distribution Model-Based Clustering

7.3.4 Hierarchical Clustering

Hierarchical clustering offers an alternative to partitioned clustering methods by eliminating the need to predefine the number of clusters in advance. This technique organically divides the dataset into clusters, forming a hierarchical tree-like structure known as a dendrogram. The beauty of hierarchical clustering lies in its flexibility, as it allows you to choose the desired number of clusters by slicing the dendrogram at an appropriate level.

In hierarchical clustering, the algorithm starts with each data point as its own cluster and then iteratively merges clusters based on their similarity, creating a hierarchy of nested clusters. This process continues until all data points belong to a single, overarching cluster. At any point in the hierarchy, you can cut the dendrogram to obtain clusters at different levels of granularity. This adaptability makes hierarchical clustering a valuable tool for exploring data without the need for a priori knowledge about the number of clusters, making it particularly useful in exploratory data analysis and data mining tasks.

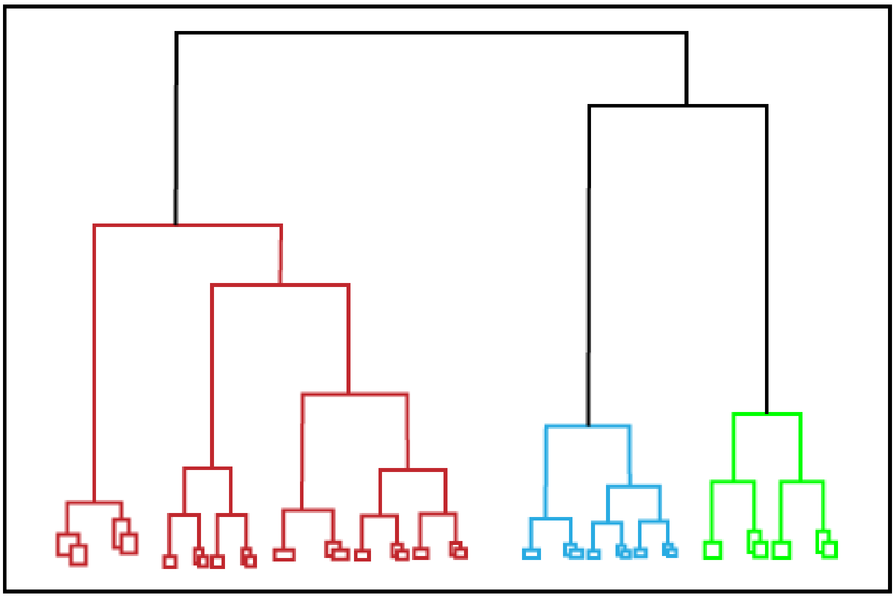


Figure 7.4: Distribution Model-Based Clustering

7.3.5 Fuzzy Clustering

Fuzzy clustering represents a soft clustering approach where a data object is allowed to have partial membership in multiple clusters rather than being strictly assigned to just one cluster. In this method, each data point is associated with a set of membership coefficients that reflect the extent to which it belongs to each cluster. These membership coefficients indicate the degree of membership, allowing for a more nuanced representation of the data's relationship with multiple clusters.

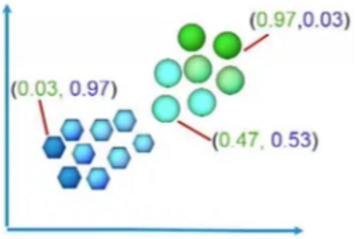


Figure 7.5: Fuzzy Clustering

7.4 Types of Clustering Algorithms

There are several types of clustering algorithms, each with its own characteristics and suitability for different types of problems. Here are some common types of classification algorithms:

- K-Means Clustering
- Fuzzy C-Means
- Gaussian Mixture Model (GMM)
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
- Mean Shift
- Hierarchical Clustering

7.4.1 K-Means Clustering

K-means clustering stands out as one of the most prevalent clustering algorithms in use today. This centroid-based method represents the simplest form of unsupervised learning.

The primary objective of the K-means algorithm is to minimize the variance of data points within each cluster. It accomplishes this by iteratively adjusting cluster centroids until convergence.

However, it's worth noting that K-means has its limitations. It performs best when applied to smaller datasets due to its iterative nature, which entails processing all data points. Consequently, on large datasets, K-means can be computationally intensive and may require more time to classify data points, making it less suitable for such scenarios.

7.4.2 Fuzzy C-Means

This unsupervised clustering algorithm enables the creation of a fuzzy partition from input data. Its operation hinges on the allocation of membership values to each data point with respect to every cluster center. This membership assignment is contingent on the distance between a data point and the cluster center. In simpler terms, the closer a data point is to a specific cluster center, the higher its membership score for that particular cluster.

7.4.3 Gaussian Mixture Model (GMM)

The distance metric used in K-means calculations assumes a circular path, which can lead to inaccurate clustering results when dealing with non-circular or elongated data shapes.

Gaussian Mixture Models (GMMs) address this limitation effectively by offering more flexibility. GMMs are capable of handling data with diverse shapes, not limited to circular or spherical structures.

In a GMM, the model employs multiple Gaussian distributions, often referred to as components or clusters, to fit the data. These individual Gaussian distributions can adapt to different shapes within the data. The model calculates the probability that a data point belongs to each of these Gaussian distributions, and the data point is assigned to the cluster associated with the highest probability.

7.4.4 DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

DBSCAN, which stands for Density-Based Spatial Clustering of Applications with Noise, is a clustering algorithm that operates based on the density of data points in a dataset. Its primary purpose is to identify outliers and discover clusters of arbitrary shapes within the data. DBSCAN effectively distinguishes different clusters by identifying regions of low density, which helps it detect outliers situated between high-density clusters.

This algorithm surpasses K-means when it comes to handling datasets with irregular or non-circular shapes. DBSCAN relies on two key parameters to define clusters:

minPts: This parameter specifies the minimum number of data points required to form a cluster.

eps (epsilon): If the distance between two data points is less than or equal to epsilon, they are considered to belong to the same cluster.

7.4.5 Mean Shift

Mean-shift is a versatile clustering algorithm that doesn't require specifying the number of clusters in advance, making it a valuable tool in exploratory data analysis.

Mean-shift lies in its mode-seeking behavior. It operates by traversing each data point and shifting them toward the mode, which corresponds to the region of high data point density. Hence, it's often referred to as a mode-seeking algorithm. During this iterative process, each data point gradually gravitates toward the nearest high-density area, ultimately leading to all data points being assigned to a cluster that corresponds to a local density peak.

However, one limitation of Mean-shift is its scalability issue with large datasets, as it involves iterating over all data points, which can be computationally intensive.

7.4.6 Hierarchical Clustering

Hierarchical clustering doesn't require specifying the number of clusters in advance, making it a versatile tool for exploratory data analysis.

Hierarchical clustering algorithms are of 2 types: Agglomerative and Divisive.

7.4.6.1 Agglomerative Hierarchy clustering algorithm

Its purpose is to organize objects into clusters by assessing their degree of similarity. This method follows a bottom-up clustering approach, where each individual data point initially forms its own cluster. Subsequently, these clusters are systematically merged together.

During each step of this process, clusters that exhibit a high degree of similarity are combined, and this merging operation continues iteratively until all data points are consolidated into a single, overarching root cluster.

7.4.6.2 Divisive clustering algorithm

In contrast to agglomerative clustering, divisive hierarchical clustering begins with all data points in a single cluster and then recursively divides clusters into smaller subclusters. This process continues until each data point is in its own cluster, yielding a dendrogram in the reverse order of agglomerative clustering.

7.5 Metrics Used to Evaluate the Quality of Clustering Algorithms

There are several metrics commonly used to evaluate the quality of clustering algorithms.
Here are some widely used clustering evaluation metrics:

- Silhouette Score
- Davies-Bouldin Index
- Calinski-Harabasz Index (Variance Ratio Criterion)
- Rand Index
- Dunn Index

7.5.1 Silhouette Score

The silhouette score measures how similar an object is to its own cluster compared to other clusters. It ranges from -1 to 1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.

$$S(i) = \frac{b(i) - a(i)}{max\left\{b(i), a(i)\right\}} \tag{7.1}$$

Where:
$a(i)$ is the average distance from data point "i" to the other data points within the same cluster "A." It measures the cohesion of the data point to its own cluster.
$b(i)$ is the minimum average distance from data point "i" to the data points in any other cluster, where "i" does not belong to that cluster. It measures the separation from other clusters.
$b(i)$-$a(i)$ quantifies how much better the data point is clustered with its own cluster compared to the neighboring clusters.
$max\left\{b(i), a(i)\right\}$ is used to normalize the score to fall within the range of -1 to +1.

7.5.2 Davies-Bouldin Index

This index measures the average similarity ratio of each cluster with the cluster that is most similar to it. A lower value indicates better clustering.

$$DB = \frac{1}{n}\sum_{i=1}^{n} max_{j\neq i}(\frac{S_i + S_j}{d(C_i + C_j)}) \tag{7.2}$$

Where
n is the number of clusters.
C_i and C_j are two different clusters.
S_i is the average distance from each point in cluster $C_i$to the centroid of cluster C_i. It measures the intra-cluster similarity.
S_j is the average distance from each point in cluster $C_j$to the centroid of cluster C_j .
$d(C_i + C_j)$ is the distance between the centroids of cluster C_i and cluster C_j. It measures the inter-cluster dissimilarity.

7.5.3 Calinski-Harabasz Index (Variance Ratio Criterion)

This index calculates the ratio of the between-cluster variance to the within-cluster variance. Higher values suggest better-defined clusters.

$$CH = \frac{B}{W} \times \frac{N-K}{K-1} \tag{7.3}$$

Where
B is the between-cluster variance, which is the sum of the variances between each cluster's centroid and the overall centroid.
W is the within-cluster variance, which is the sum of the variances within each cluster.
N is the total number of data points.
k is the number of clusters.

7.5.4 Dunn Index

The Dunn Index evaluates the ratio of the minimum inter-cluster distance to the maximum intra-cluster distance. A higher Dunn Index indicates better clustering.

$$D = \frac{min_{i \neq j}(d(C_i, C_j))}{max_k(d_{intra}(C_k))} \tag{7.4}$$

Where
$d(C_i, C_j)$ represents the distance between cluster C_i and C_j, which is the minimum distance between any two data points from different clusters.
$d_{intra}(C_k)$ represents the intra-cluster distance for cluster C_k, which is the maximum distance between any two data points within the same cluster.

7.5.5 Rand Index

The Rand index measures the similarity between the true clustering and the clustering produced by the algorithm. It provides a measure of the accuracy of the clustering.

$$RI = \frac{TP + TN}{TP + TN + FP + FN} \tag{7.5}$$

Where:
TP is the number of true positive pairs, i.e., pairs of data points that are correctly grouped together in both the predicted clustering and the reference clustering.
TN is the number of true negative pairs, i.e., pairs of data points that are correctly placed in separate clusters in both the predicted clustering and the reference clustering.
FP is the number of false positive pairs, i.e., pairs of data points that are grouped together in the predicted clustering but not in the reference clustering.
FN is the number of false negative pairs, i.e., pairs of data points that are placed in separate clusters in the predicted clustering but not in the reference clustering.

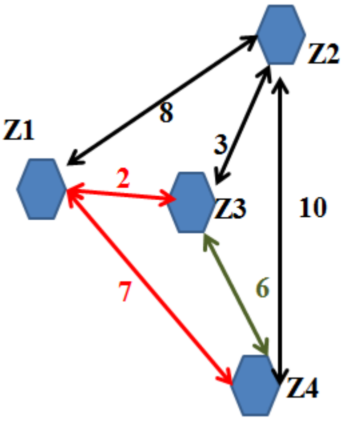
7.6 Libraries for Implementing Clustering Algorithms

There are several libraries available for implementing and applying clustering algorithms in various programming languages. Here are some popular options:
Scikit-learn, SciPy and PyClustering are Python libraries. Cluster and fpc are R packages. Weka is Java's library. Dlib and MLPack are C++ libraries. MATLAB Statistics and Machine Learning Toolbox it is toolbox for various clustering algorithms and data analysis tasks.

7.7 Exercises

Exercise 1

A logistics report provides the distance (in kilometers) between different zones (Z1, Z2, Z3, Z4). To organize the circuit for visiting these zones, create a dendrogram using agglomerative hierarchical clustering following the maximum linkage strategy.



Exercise 2

Table 7.1 displays a set of two-dimensional points.

-Apply K-Means clustering with two centers to assign each point to a class.

Centroid 1: (4, 5)

Centroid 2: (12, 13)

-Provide the coordinates of the centers after the first iteration.

X-coordinate	**Y-coordinate**	**Point**
2	3	Point 1
4	5	Point 2
6	7	Point 3
8	9	Point 4
10	11	Point 5
12	13	Point 6
14	15	Point 7
16	17	Point 8

Table 7.1: Data points coordinates

7.7.1 Exercise 3

Given the data points and their membership for the two centers (Table 7.2), apply fuzzy C-means clustering to determine the new coordinates of the centers and the membership of each point to the two centers after the first iteration.

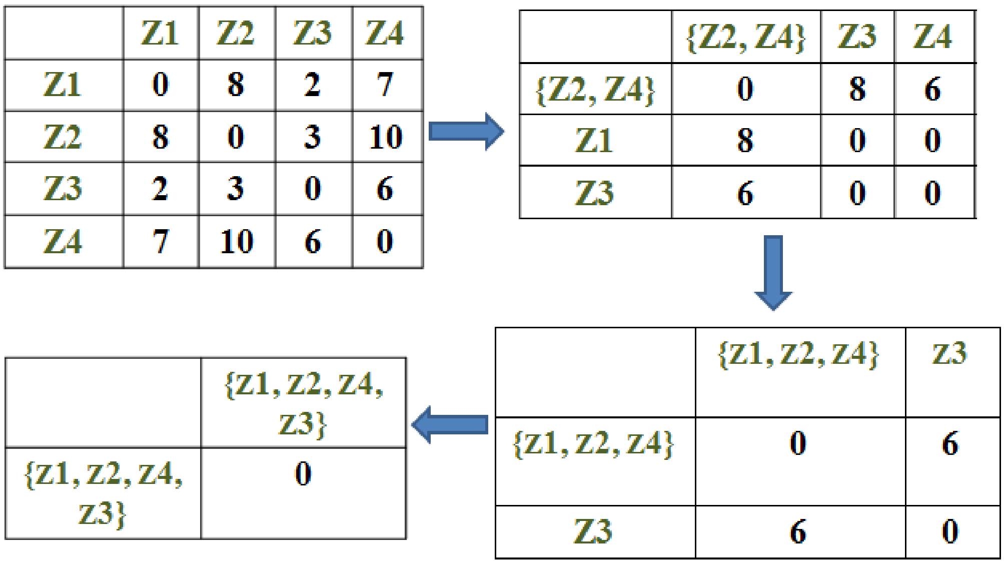
Cluster	**(1,3)**	**(2,5)**	**(4,8)**	**(7,9)**
1	0.8	0.7	0.2	0.1
2	0.2	0.3	0.8	0.9

Table 7.2: Data points membership

7.8 Solutions

Solution exercise 1

The agglomerative hierarchical clustering, employing the maximum linkage strategy, entails grouping zones with the maximum distance between them. This process is illustrated in Figure 7.6, and the resulting dendrogram is depicted in Figure 7.7.



	Z1	Z2	Z3	Z4
Z1	0	8	2	7
Z2	8	0	3	10
Z3	2	3	0	6
Z4	7	10	6	0

	{Z2, Z4}	Z3	Z4
{Z2, Z4}	0	8	6
Z1	8	0	0
Z3	6	0	0

	{Z1, Z2, Z4}	Z3
{Z1, Z2, Z4}	0	6
Z3	6	0

	{Z1, Z2, Z4, Z3}
{Z1, Z2, Z4, Z3}	0

Figure 7.6: Maximum linkage strategy for grouping the zones

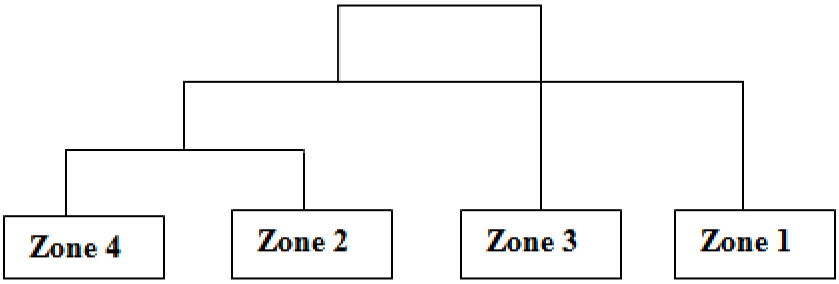


Figure 7.7: Dendrogram

Solution exercise 2

We compute the Euclidean distance between each point and the two centers.
For point (2, 3):

Distance to Centroid 1: $sqrt((2-4)^2 + (3-5)^2) = 2.83$

Distance to Centroid 2: $sqrt((2-12)^2 + (3-13)^2) = 14.14$

Assign point (2, 3) to Centroid 1

After applying the same distance for all other points, we obtained the following results:

Assign point (2, 3) to Centroid 1
Assign point (4, 5) to Centroid 1
Assign point (6, 7) to Centroid 1
Assign point (8, 9) to Centroid 1
Assign point (10, 11) to Centroid 1
Assign point (12, 13) to Centroid 2
Assign point (14, 15) to Centroid 2
Assign point (16, 17) to Centroid 2

-We calculate the mean of the points assigned to each centroid

Centroid 1:
(2+4+6+8+10) / 5 = 6
(3+5+7+9+11) / 5 = 7

Centroid 2:
(12+14+16) / 3 = 14
(13+15+17) / 3 = 15

So the new coordinates of the centers are:
Centroid 1: (6, 7)
Centroid 2: (14, 15)

Solution exercise 3

We compute the new center coordinates according to the following formula:

$$C_{i,j} = \frac{\sum_{k=1}^{n} \gamma_{ik}^{m} * x_k}{\sum_{k=1}^{n} \gamma_{ik}^{m}} \tag{7.6}$$

where

γ: fuzzy membership value

m: fussiness parameter generally taken as 2

x_k: is the data point

n: number of data points

$$C_{11} = \frac{0.8^2 * 1 + 0.7^2 * 2 + 0.2^2 * 4 + 0.1^2 * 7}{0.8^2 + 0.7^2 + 0.2^2 + 0.1^2} = 1.568$$

$$C_{12} = \frac{0.8^2 * 3 + 0.7^2 * 5 + 0.2^2 * 8 + 0.1^2 * 9}{0.8^2 + 0.7^2 + 0.2^2 + 0.1^2} = 4.051$$

So the coordinates of Center 1 are: (1.568,4.051).

$$C_{21} = \frac{0.2^2 * 1 + 0.3^2 * 2 + 0.8^2 * 4 + 0.9^2 * 7}{0.2^2 + 0.3^2 + 0.8^2 + 0.9^2} = 5.35$$

$$C_{22} = \frac{0.2^2 * 3 + 0.3^2 * 5 + 0.8^2 * 8 + 0.9^2 * 9}{0.2^2 + 0.3^2 + 0.8^2 + 0.9^2} = 8.215$$

So the coordinates of Center 2 are: (5.35,8.215).

For point (1,3):

Distance to Centroid 1: $sqrt((1 - 1.568)^2 + (3 - 4.051)^2) = 1.2$

Distance to Centroid 2: $sqrt((1 - 5.35)^2 + (3 - 8.215)^2) = 6.79$

For point (2,5):

Distance to Centroid 1: $sqrt((2 - 1.568)^2 + (5 - 4.051)^2) = 1.04$

Distance to Centroid 2: $sqrt((2 - 5.35)^2 + (5 - 8.215)^2) = 4.64$

For point (4,8):

Distance to Centroid 1: $sqrt((4 - 1.568)^2 + (8 - 4.051)^2) = 4.63$

Distance to Centroid 2: $sqrt((4 - 5.35)^2 + (8 - 8.215)^2) = 1.36$

For point (7,9):

Distance to Centroid 1: $sqrt((7-1.568)^2+(9-4.051)^2)=7.34$

Distance to Centroid 2: $sqrt((7-5.35)^2+(9-8.215)^2)=1.82$

We compute the new membership of all points using the following formula:

$$\gamma_{ki}=\left(\sum_{j=1}^{n}\frac{d_{ki}^2}{d_{kj}^2}^{(\frac{1}{m-1})}\right)^{-1}$$

where

d represent the distance between the point and the centroid.

For point (1,3):

Distance to Centroid 1:

$$\gamma_{11}=\left(\left\{\frac{(1.2)^2}{(1.2)^2}+\frac{(1.2)^2}{(6.79)^2}\right\}^{\left(\frac{1}{(2-1)}\right)}\right)^{-1}=0.97$$

Distance to Centroid 2:

$$\gamma_{12}=\left(\left\{\frac{(6.79)^2}{(1.2)^2}+\frac{(6.79)^2}{(6.79)^2}\right\}^{\left(\frac{1}{(2-1)}\right)}\right)^{-1}=0.03$$

For point (2,5):

Distance to Centroid 1:

$$\gamma_{21}=\left(\left\{\frac{(1.04)^2}{(1.04)^2}+\frac{(1.04)^2}{(4.64)^2}\right\}^{\left(\frac{1}{(2-1)}\right)}\right)^{-1}=0.95$$

Distance to Centroid 2:

$$\gamma_{22}=\left(\left\{\frac{(4.64)^2}{(1.04)^2}+\frac{(4.64)^2}{(4.64)^2}\right\}^{\left(\frac{1}{(2-1)}\right)}\right)^{-1}=0.05$$

For point (4,8):

Distance to Centroid 1:

$$\gamma_{31} = \left(\left\{ \frac{(4.63)^2}{(4.63)^2} + \frac{(4.63)^2}{(1.36)^2} \right\}^{\left(\frac{1}{(2-1)}\right)} \right)^{-1} = 0.08$$

Distance to Centroid 2:

$$\gamma_{32} = \left(\left\{ \frac{(1.36)^2}{(4.63)^2} + \frac{(1.36)^2}{(1.36)^2} \right\}^{\left(\frac{1}{(2-1)}\right)} \right)^{-1} = 0.92$$

For point (7,9):

Distance to Centroid 1:

$$\gamma_{41} = \left(\left\{ \frac{(7.34)^2}{(7.34)^2} + \frac{(7.34)^2}{(1.82)^2} \right\}^{\left(\frac{1}{(2-1)}\right)} \right)^{-1} = 0.06$$

Distance to Centroid 2:

$$\gamma_{42} = \left(\left\{ \frac{(1.82)^2}{(7.34)^2} + \frac{(1.82)^2}{(1.82)^2} \right\}^{\left(\frac{1}{(2-1)}\right)} \right)^{-1} = 0.94$$

Cluster	**(1,3)**	**(2,5)**	**(4,8)**	**(7,9)**
1	0.97	0.95	0.08	0.06
2	0.03	0.05	0.92	0.64

Table 7.3: The new data points membership

7.9 Conclusion

Clustering analysis serves as a vital technique in industrial applications, facilitating organizational improvements and informed decision-making processes. Throughout this chapter, we've explored different types of clustering tasks in machine learning, encompassing partitioning clustering, density-based clustering, distribution model-based clustering, hierarchical clustering, and fuzzy clustering. Additionally, we've examined various clustering algorithms such as K-means, fuzzy C-means, Gaussian mixture models, and density-based spatial clustering of applications with noise (DBSCAN), mean-shift, and hierarchical clustering. Furthermore, we've delved into the metrics utilized to assess clustering algorithm quality, along with the libraries available for implementing these algorithms. This chapter concludes with a series of exercises.

CONCLUSION

This course on data analysis is designed for third-year engineering students specializing in industrial engineering. It consolidates key concepts in data analysis.

Spanning 7 chapters, we have introduced diverse models and approaches employed in univariate, bivariate, and multivariate analysis. We have also presented classification and clustering models, extensively utilized in today's industry for making informed decisions.

By offering this educational resource, our aim is to enrich the learning journey for students. We believe that a strong foundation in data analysis and its practical applications will empower them to excel in their future endeavors, both academically and professionally.

BIBLIOGRAPHY

[1] Hervé Abdi and Michel Béra. Correspondence analysis., 2014.

[2] Ravinder Ahuja, Aakarsha Chug, Shaurya Gupta, Pratyush Ahuja, and Shruti Kohli. Classification and clustering algorithms of machine learning with their applications. *Nature-inspired computation in data mining and machine learning*, pages 225–248, 2020.

[3] Jafar Alzubi, Anand Nayyar, and Akshi Kumar. Machine learning from theory to algorithms: an overview. In *Journal of physics: conference series*, volume 1142, page 012012. IOP Publishing, 2018.

[4] R Artusi, P Verderio, and EJTIjobm Marubini. Bravais-pearson and spearman correlation coefficients: meaning, test of hypothesis and confidence interval. *The International journal of biological markers*, 17(2):148–151, 2002.

[5] Taiwo Oladipupo Ayodele. Machine learning overview. *New Advances in Machine Learning*, 2(9-18):16, 2010.

[6] James C Bezdek and Nikhil R Pal. Cluster validation with generalized dunn's indices. In *Proceedings 1995 second New Zealand international two-stream conference on artificial neural networks and expert systems*, pages 190–193. IEEE, 1995.

[7] FM Bi, WK Wang, and L Chen. Dbscan: density-based spatial clustering of applications with noise. *J. Nanjing Univ*, 48(4):491–498, 2012.

[8] Gérard Biau and Erwan Scornet. A random forest guided tour. *Test*, 25:197–227, 2016.

[9] Sorana-Daniela Bolboaca and Lorentz Jäntschi. Pearson versus spearman, kendall's tau correlation analysis on structure-activity relationships of biologic active compounds. *Leonardo Journal of Sciences*, 5(9):179–200, 2006.

[10] Rasmus Bro and Age K Smilde. Principal component analysis. *Analytical methods*, 6(9):2812–2831, 2014.

[11] Marielle Brunette, Robin Bourke, Marc Hanewinkel, and Rasoul Yousefpour. Adaptation to climate change in forestry: A multiple correspondence analysis (mca). *Forests*, 9(1):20, 2018.

[12] Dorin Comaniciu and Peter Meer. Mean shift analysis and applications. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1197–1203. IEEE, 1999.

[13] Giovanni Di Franco. Multiple correspondence analysis: one only or several techniques? *Quality & Quantity*, 50(3):1299–1315, 2016.

[14] Michael Greenacre, Patrick JF Groenen, Trevor Hastie, Alfonso Iodice d'Enza, Angelos Markos, and Elena Tuzhilina. Principal component analysis. *Nature Reviews Methods Primers*, 2(1):100, 2022.

[15] Brendon Hall. Facies classification using machine learning. *The Leading Edge*, 35(10):906–909, 2016.

[16] Donna L Hoffman and Jan De Leeuw. Interpreting multiple correspondence analysis as a multidimensional scaling method. *Marketing letters*, 3:259–272, 1992.

[17] Carl Kingsford and Steven L Salzberg. What are decision trees? *Nature biotechnology*, 26(9):1011–1013, 2008.

[18] David G Kleinbaum, K Dietz, M Gail, Mitchel Klein, and Mitchell Klein. *Logistic regression.* Springer, 2002.

[19] Sotiris B Kotsiantis, Ioannis Zaharakis, P Pintelas, et al. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160(1):3–24, 2007.

[20] Oliver Kramer and Oliver Kramer. K-nearest neighbors. *Dimensionality reduction with unsupervised nearest neighbors*, pages 13–23, 2013.

[21] Anthony McGregor, Mark Hall, Perry Lorier, and James Brunskill. Flow clustering using machine learning techniques. In *Passive and Active Network Measurement: 5th International Workshop, PAM 2004, Antibes Juan-les-Pins, France, April 19-20, 2004. Proceedings 5*, pages 205–214. Springer, 2004.

[22] Tom Michael Mitchell. *The discipline of machine learning*, volume 9. Carnegie Mellon University, School of Computer Science, Machine Learning . . . , 2006.

[23] James Moor. *The Turing test: the elusive standard of artificial intelligence*, volume 30. Springer Science & Business Media, 2003.

[24] Fionn Murtagh and Pedro Contreras. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1):86–97, 2012.

[25] Leann Myers and Maria J Sirois. Spearman correlation coefficients, differences between. *Encyclopedia of statistical sciences*, 12, 2004.

[26] Shi Na, Liu Xumin, and Guan Yong. Research on k-means clustering algorithm: An improved k-means clustering algorithm. In *2010 Third International Symposium on intelligent information technology and security informatics*, pages 63–67. Ieee, 2010.

[27] William S Noble. What is a support vector machine? *Nature biotechnology*, 24(12):1565–1567, 2006.

[28] Nikolaos Pandis. The chi-square test. *American journal of orthodontics and dentofacial orthopedics*, 150(5):898–899, 2016.

[29] Maja Popović. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395, 2015.

[30] Douglas A Reynolds et al. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663), 2009.

[31] Frédéric Ros, Rabia Riad, and Serge Guillaume. Pdbi: A partitioning davies-bouldin index for clustering evaluation. *Neurocomputing*, 528:178–199, 2023.

[32] Günther Schuh, Gunther Reinhart, Jan-Philipp Prote, Frederick Sauermann, Julia Horsthofer, Florian Oppolzer, and Dino Knoll. Data mining definitions and applications for the management of production complexity. *Procedia Cirp*, 81:874–879, 2019.

[33] Ketan Rajshekhar Shahapure and Charles Nicholas. Cluster quality analysis using silhouette score. In *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*, pages 747–748. IEEE, 2020.

[34] Douglas Steinley. Properties of the hubert-arable adjusted rand index. *Psychological methods*, 9(3):386, 2004.

[35] R Suganya and R Shanthi. Fuzzy c-means algorithm-a review. *International Journal of Scientific and Research Publications*, 2(11):1, 2012.

[36] Ronald J Tallarida, Rodney B Murray, Ronald J Tallarida, and Rodney B Murray. Chi-square test. *Manual of pharmacologic calculations: With computer programs*, pages 140–142, 1987.

[37] Alaa Tharwat. Principal component analysis-a tutorial. *International Journal of Applied Pattern Recognition*, 3(3):197–240, 2016.

[38] John W Tukey. The future of data analysis. *The annals of mathematical statistics*, 33(1):1–67, 1962.

[39] Matthijs J Warrens. Five ways to look at cohen's kappa. *Journal of Psychology & Psychotherapy*, 5, 2015.

[40] Geoffrey I Webb, Eamonn Keogh, and Risto Miikkulainen. Naïve bayes. *Encyclopedia of machine learning*, 15(1):713–714, 2010.

[41] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[42] Emine Yilmaz, Javed A Aslam, and Stephen Robertson. A new rank correlation coefficient for information retrieval. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 587–594, 2008.

[43] Wenbo Zhang, Zixian Yue, Jinmei Ye, Hengying Xu, Yuxiang Wang, Xiaoguang Zhang, and Lixia Xi. Modulation format identification using the calinski–harabasz index. *Applied Optics*, 61(3):851–857, 2022.