Statistical Study & Data Analysis on

Mental Health of Medical Students in Switzerland

Project work Submitted to

The Department of Statistics

Pondicherry University

By

Ms. Niyati Sharma

Regd. No.: 24MSSTAPY0029

As an Assignment for the

Partial Fulfilment of the Course work

STAT-414 Programming in R (Lab Based)

Under the Guidance of

Dr. P. Tirupathi Rao

Professor & Course Teacher

November – 2024

DR.TIRUPATHI RAO PADI

M.Sc., M.Phil., Ph.D., M.B.A., D.C.A., CP-SAS

PROFESSOR, Department of Statistics

Ramanujan School of Math. Sciences

PONDICHERRY UNIVERSITY

(A Central University)

R.V. Nagar, Kalapet,

Puducherry (UT) – 605014, India

Email: drtrpadi@pondiuni.ac.in;

traopadipu@gmail.com,

Phone: 9486492241 (Mobile),

9629862241 (WAN), 0413-2654-829(Office),

Certificate

This is to certifiy that Ms. Niyati Sharma has completed her project work as partial fulfilment fo the course work of Stat 414: R-Programming (Lab based) submitted to the department of Statistics, Pondicherry University in November 2024. She has carried out all the stages of project work right from the data collection to repot making, on her own. The data that they have collected is from a real time context. No part of this work was carried out earlier in any format by any one for M.Sc./ MBA/ M.Tech/ etc. dissertation works or Ph.D. thesis.

Date: 30.11.2024

(P. Tirupathi Rao)

TABLE OF CONTENTS

ABSTRACT

This study examines the mental health, empathy, and burnout among 886 medical students in Switzerland. Utilizing a comprehensive dataset capturing demographic, psychological, and academic factors, the research aims to identify predictors of mental health challenges, investigate the role of empathy in medical training, and explore the interplay of academic stress and personal attributes. The analysis includes descriptive and inferential statistics, correlational studies, and advanced regression techniques. Key findings are intended to inform policy recommendations for enhancing mental health support systems and fostering resilience in medical education. This work contributes to the broader societal goal of promoting the well-being of healthcare professionals and improving patient care outcomes.

CHAPTER-1

Introduction

The dataset captures the mental health parameters of 886 medical students in Switzerland. The data includes both categorical and continuous variables that capture aspects like age, sex, mental health assessment scores, and various psychological test results. It explores medical students' empathy, mental health, and burnout in Switzerland. A brief about the features included in the dataset are:

a. Id: Participants ID number
b. Age: Age of the participant age at questionnaire 20-21
c. Year: curriculum year
d. Sex: Gender
e. Glang: mother tongue
f. Part: partnership status
g. Job: having a job
h. stud_h: hours of study per week
i. health: satisfaction with health
j. psyt: psychotherapy last year
k. jspe: JSPE total empathy score
l. qcae_cog: QCAE Cognitive empathy score
m. qcae_aff: QCAE Affective empathy score
n. amsp: AMSP total score
o. erec_mean: GERT
p. cesd: CES-D total score
q. stai_t: STAI score
r. mbi_ex: MBI Emotional Exhaustion
s. mbi_cy: MBI Cynicism
t. mbi_ea: MBI Academic Efficacy

The goal is to analyse the problems faced by medical students during their time of study and understand what factors actually affect the mental health of medical students in order to promote better policies by understanding relationships between the features involved. After looking at the individual factors that may contribute to different outcomes, one can work on improvement of the educational systems for the benefits of both students and their eventual patients

1.1 Introduction and description on the problem

The dataset comprises 886 observations and 20 variables, representing demographic, psychological, and academic factors. Each row corresponds to an individual medical student, and each column provides information about a specific aspect of their health, academic performance, or personal background. By analysing this dataset, one can aim to shed light on the factors impacting the mental health of medical students, contributing to better policy-making and student well-being programs. The key features include:

a. Demographic Information such as – id, age, sex, glang
b. Academic Information such as – year, stud_h
c. Menta Health Scores such as – jspe, qcae_cog, qcae_aff, cesd, stai_t, mbi_ex, mbi_cy, mbi_ea
d. Health and Well-being indicators such as – health, psyt, amsp, erec_mean
e. Personal and Social Attributes – part, job

The primary objective of analyzing this problem dataset is –

- To identify significant predictors of mental health conditions among medical students
- To understand the relationship between academic stress, personal attributes, and mental health outcomes.
- To assess the impact of psychological scores (e.g. empathy, anxiety) on students' overall well-being.
- To suggest interventions to improve mental health support systems in educational insitutions.

1.2 Significance of the problem under study

The mental health of medical students is a critical issue due to their exposure to intense academic pressure, long study hours, and emotional challenges inherent in their training. This population faces a higher risk of anxiety, depression, and burnout compared to peers in other disciplines. Understanding the factors contributing to these mental health challenges is vital for ensuring their well-being and professional sustainability. Addressing these problems is also essential to improving their academic performance and the quality of care they will provide as future healthcare professionals.

1.3 Relevance/ importance of the study to the society

This study relates with students' mental health and specifically for medical students which eventually affects the wellbeing of their patients too. Following are some aspects were this study can cause a good impact:

a. Improved Healthcare Outcomes

   Understanding the predictors of poor mental health allows for timely interventions, helping at-risk students manage stress, build resilience, and prevent severe mental health outcomes.

b. Educational Policy Enhancements

   Findings from the study can inform the development of better academic and mental health support systems, reducing stress and enhancing learning environments in medical institutions.

c. Early Identification of Risk Factors

   Understanding the predictors of poor mental health allows for timely interventions, helping at-risk students manage stress, build resilience, and prevent severe mental health outcomes.

d. Reduction in Burnout Rates

   Mitigating burnout among medical students ensures a more committed, compassionate, and effective healthcare workforce, addressing a critical issue in the medical profession.

e. Societal Awareness

   The study raises awareness about the importance of mental health in high-stress professions and contributes to destigmatizing seeking help for mental health issues.

The study aligns with the broader societal goals of fostering healthier educational practices, promotes mental health awareness and ensuring the well-being of future healthcare providers.

1.4 Glossary (description about some important terminology and key words)

A brief description about the attributes (variables) in the dataset:

a. Demographic Information

- **id**: Participants ID number; string
- **age**: Age of the student at questionnaire 20-21 ; numeric
- **sex**: Gender of the student; To which gender do you identify the most ? (1 = Male, 2 = Female, 3 = Non-Binary).
- **glang**: Mother tongue of the student; What is your mother tongue? (1=French; 15=German; 20=English; 37=Arab; 51=Basque; 52=Bulgarian; 53=Catalan; 54=Chinese; 59=Korean; 60=Croatian; 62=Danish; 63=Spanish; 82=Estonian; 83=Finnish; 84=Galician; 85=Greek; 86=Hebrew; 87=Hindi; 88=Hungarian; 89=Indonesian; 90=Italian; 92=Japanese; 93=Kazakh; 94=Latvian; 95=Lithuanian; 96=Malay; 98=Dutch; 100=Norwegian; 101=Polish; 102=Portuguese; 104=Romanian; 106=Russian; 108=Serbian; 112=Slovak; 113=Slovenian;

114=Swedish; 116=Czech; 117=Thai; 118=Turkish; 119=Ukrainian; 120=Vietnamese; 121=Other)

b. Academic Information

- **year**: CURICULUM YEAR :In which curriculum year are you? (1=Bmed1; 2=Bmed2; 3=Bmed3; 4=Mmed1; 5=Mmed2; 6=Mmed3)
- **stud_h**: Weekly study hours reported by the student.

c. Mental Health Scores (numeric)

- **jspe**: Jefferson Scale of Physician Empathy score
  A measure designed to assess the empathy levels of medical professionals or students, focusing on understanding and responding to patients' feelings and perspectives. A higher score indicates greater empathy.
- **qcae_cog**: Cognitive empathy score from the QCAE test
  Reflects the ability to understand and infer the thoughts, feelings, and intentions of others. It measures the intellectual component of empathy, focusing on perspective-taking and comprehension.
- **qcae_aff**: Affective empathy score from the QCAE test.
  Captures the emotional component of empathy, reflecting the ability to feel and respond emotionally to another person's experiences or distress.
- **cesd**: Center for Epidemiologic Studies Depression Scale score.
  A widely used scale that measures the frequency and severity of depressive symptoms in individuals over the past week. Higher scores indicate greater depressive symptoms, with specific thresholds for clinical concern.
- **stai_t**: State-Trait Anxiety Inventory score.
  A comprehensive tool to measure two types of anxiety:
  - State Anxiety: Temporary feelings of stress or tension in specific situations.
  - Trait Anxiety: An individual's general tendency to perceive situations as stressful.
  Higher scores indicate greater levels of anxiety.
- **mbi_ex**: Maslach Burnout Inventory exhaustion score.
  Measures feelings of being emotionally overextended and exhausted due to work or academic demands. A key dimension of burnout, with higher scores indicating greater exhaustion.
- **mbi_cy**: MBI cynicism score.

Reflects a detached or negative attitude toward work or academic activities, often arising as a coping mechanism for burnout. Higher scores indicate greater cynicism.

- **mbi_ea**: MBI emotional achievement score.
  Assesses feelings of competence, productivity, and achievement in academic or professional settings. Lower scores suggest reduced efficacy, a characteristic of burnout.

d. Health and Well-being Indicators
   - **health**: Self-reported satisfaction with health; How satisfied are you with your health? (scale of 1–5 : 1=Verydissatisfied; 2=Dissatisfied; 3=Neithersatisfiednordissatisfied; 4=Satisfied; 5=Verysatisfied)
   - **psyt**: History of psychotherapy in the past year; During the last 12 months, have you ever consulted a psychotherapist or a psychiatrist for your health? (0=No; 1=Yes).
   - **amsp**: Affective, Mental Stress, and Psychological Score
     A composite measure evaluating mental stress, emotional responses, and psychological well-being. A higher score might indicate greater stress or mental health challenges.
   - **erec_mean**: mean value of correct responses

e. Personal and Social Attributes
   - **part**: Partnership status; Do you have a partner? (0=No; 1=Yes).
   - **job**: Employment status; Do you have a paid job? (0=No; 1=Yes).

1.5 Motivation for selecting the specific topic

a. Rising Mental Health Concerns in Medical Education:

- Medical students face intense academic pressure, long hours of study, and emotionally demanding clinical experiences, which often lead to higher rates of anxiety, depression, and burnout compared to their peers in other disciplines.
- Understanding the root causes of these issues and addressing them is essential for safeguarding their well-being and future professional effectiveness.

b. Importance of Empathy in Healthcare:

- Empathy is a critical skill for medical professionals, directly influencing patient care and satisfaction.
- Investigating factors like empathy levels (measured by JSPE and QCAE) in medical students can help ensure that training programs cultivate compassionate healthcare providers.

c. Contribution to Educational Policy

- Identifying factors influencing mental health can guide educational institutions to develop better support systems, such as mental health resources, peer counseling, and curriculum adjustments, creating a healthier learning environment.

By addressing these motivations, the study seeks to highlight the significance of mental health in medical education and its far-reaching implications for individuals and society.

1.6 Data collection sources

The data is collected from Kaggle.com. It has a large number of datasets that provides researches, data scientists, students, etc. to use them for learning purpose and carry out results that may be helpful for the society. This dataset contains information on the mental health, empathy and burnout of medical students in Switzerland. The dataset is originally taken from zenodo.org.

1.7 Data preparation steps

The steps involved are:

a. Understanding the Metadata: Review the metadata and accompanying documentation to understand the variables, data types, and dataset structure.

b. Data Cleaning:
   - Replace missing values with appropriate substitutes (e.g., mean/median for numerical data, mode for categorical data).
   - Remove Duplicates
   - Check for Outliers: Identify unusually high or low values using summary statistics or visualization and decide whether to retain or adjust them based on the context.

c. Standardize Variable Names: Rename columns to ensure consistent, descriptive, and meaningful names

d. Categorise the variables into - Demographic Information, Academic Information, Mental Health Scores and Personal and Social Attributes, so that analysis becomes easier. Along with this, categorise the variables into nominal, ordinal and scalar variables.

e. Check Data Types- Verify the data types of each variable (e.g., integers, strings, dates) and convert them as necessary to ensure compatibility with analysis.

f. Data Transformation:

   Categorical Variables - Assign meaningful labels to numerical codes (e.g., 1 = Male, 2 = Female).

   Numerical Variables - Scale values if necessary to bring them to a comparable range.

Derived Variables - Create new variables, if needed, based on existing ones (e.g., age groups from raw age data).

g. Address Data Imbalances: Examine class distributions in categorical variables and ensure sufficient representation for meaningful analysis.\

h. Save Prepared Data: Save the cleaned and prepared data in a widely used format (e.g., CSV or Excel) with clear documentation on any changes made.

These steps ensure the dataset is well-structured, clean, and ready for descriptive statistical analysis, allowing for meaningful insights to be drawn.

1.8 Specimen data with first five rows and last five rows

First five rows :

id	age	year	sex	glang	part	job	stud_h	health	psyt	jspe	qcae_cog	qcae_aff	amsp	erec_mea	cesd	stai_t	mbi_ex	mbi_cy	mbi_ea
2	18	1	1	120	1	0	56	3	0	88	62	27	17	0.738095	34	61	17	13	20
4	26	4	1	1	1	0	20	4	0	109	55	37	22	0.690476	7	33	14	11	26
9	21	3	2	1	0	0	36	3	0	106	64	39	17	0.690476	25	73	24	7	23
10	21	2	2	1	0	1	51	5	0	101	52	33	18	0.833333	17	48	16	10	21
13	21	3	1	1	1	0	22	4	0	102	58	28	21	0.690476	14	46	22	14	23
14	26	5	2	1	1	1	10	2	0	102	48	37	17	0.690476	14	56	18	15	18
17	23	5	2	1	1	0	15	3	0	117	58	38	23	0.714286	45	56	28	17	16

Last five rows:

id	age	year	sex	glang	part	job	stud_h	health	psyt	jspe	qcae_c	qcae_a	amsp	erec_m	cesd	stai_t	mbi_ex	mbi_cy	mbi_ea
1780	20	2	2	1	1	0	20	5	0	109	61	37	16	0.738095	3	32	13	5	29
1781	21	2	1	1	1	0	45	3	0	106	63	39	28	0.619048	41	39	23	4	34
1785	20	2	2	1	0	0	13	3	0	113	67	40	21	0.809524	26	41	17	5	24
1787	19	1	1	1	0	0	50	5	0	100	50	31	24	0.547619	14	45	15	8	31
1789	24	5	2	1	0	0	20	2	1	120	64	39	21	0.785714	33	58	22	15	19
1790	22	3	1	1	0	1	20	5	0	102	54	26	25	0.571429	5	27	11	9	30

1.9 Preparation of Data matrix

A data matrix can be created by organizing the variables or attributes in the dataset into rows and columns, where each row represents a student that has been interviewed on his/her mental health status, and each column represents the different types of variables like – age, sex, glang (mother tongue), jspe, amsp, etc.

1.10 Data formatting on Excel sheet

The data has been formatted using an excel sheet in order to have ease during analysis. This data had many categorical variables which are set to different classes and ordinal variables were also identified using excel tools. The dataset can be exported as an excel sheet for further analysis.

CHAPTER-2

Descriptive Statistics / Exploratory Data Analysis

2.1 Importance of exploratory data analysis:

Exploratory Data Analysis (EDA) is a critical step in any data analysis project. It involves summarizing, visualizing, and examining data to uncover patterns, anomalies, and relationships before applying statistical methods or models. Below are the key reasons why EDA is essential:

a) Understanding the Dataset
   - Data Familiarity: Helps analysts and researchers understand the structure, variables, and characteristics of the dataset.
   - Variable Identification: Identifies types of variables (categorical, numerical, ordinal) and their distributions.

b) Detecting Data Issues :
   - Missing Values: Identifies the extent and patterns of missing data, which can influence the choice of imputation methods or exclusion criteria.
   - Outliers: Detects unusual data points that may distort statistical results or models.
   - Inconsistencies: Reveals data entry errors, duplicate entries, or incorrect data formats.

c) Summarizing Data :
   - **Descriptive Statistics**: Provides summary measures like mean, median, variance, and frequency distributions, offering a snapshot of the dataset.
   - **Data Visualization**: Visual tools such as histograms, box plots, and scatter plots make patterns and relationships easier to interpret. Various libraries for visualisation can be used like – ggplot2, plotly, etc.

d) Guiding Analysis Decisions
   - Hypothesis Generation: Helps in forming hypotheses by identifying trends or anomalies in the data.
   - Variable Relationships: Highlights correlations or associations between variables to focus further analysis.
   - Model Selection: Aids in determining suitable statistical models or machine learning algorithms based on data behavior.

e) Guiding Analysis Decisions :

- Hypothesis Generation: Helps in forming hypotheses by identifying trends or anomalies in the data.
- Variable Relationships: Highlights correlations or associations between variables to focus further analysis.
- Model Selection: Aids in determining suitable statistical models or machine learning algorithms based on data behavior.

f) Reducing Bias and Errors

- By identifying irregularities or limitations in the dataset, EDA minimizes the risk of bias, errors, and misinterpretation in subsequent analysis.

g) Communicating Insights

- **Clear Presentation**: Visualization and summary statistics from EDA are often used to communicate initial findings to stakeholders.
- **Data Storytelling**: EDA provides the foundation for presenting a coherent narrative about the data and its potential implications.

EDA is a foundational step that ensures data readiness, helps uncover valuable insights, and paves the way for accurate and meaningful analysis. Skipping or neglecting EDA can lead to flawed conclusions and ineffective decision-making.

2.2 Meaning and Scope of Descriptive statistics:

2.2.1 Meaning

Descriptive statistics refers to the branch of statistics that focuses on summarizing, organizing, and presenting data in a meaningful way. It involves using numerical and graphical methods to describe the main features of a dataset, providing an overview of the data without making any inferences or predictions.

Key points:

- **Descriptive in Nature**: It does not infer or predict but merely describes the dataset as it is.
- **Data Summarization**: Uses measures such as averages, variability, and distribution to condense and present data effectively.
- **Foundation for Analysis**: Serves as the first step before conducting advanced statistical analyses.

2.2.2 Scope of Descriptive Statistics

The scope of descriptive statistics covers a wide range of methods and applications, which are essential in various fields such as business, healthcare, education, and social sciences. Below are its key aspects:

a. **Measures of Central Tendency-** Central tendency refers to measures that summarize a dataset by identifying the central point around which the data tends to cluster.
   Examples –
   Mean: The average of all data points.
   Median: The middle value when data is ordered.
   Mode: The most frequently occurring value.

b. **Measures of Dispersion -** Dispersion measures the spread or variability in a dataset.
   Examples –
   Range: Difference between the highest and lowest values.
   Variance: Average of squared deviations from the mean.
   Standard Deviation: Square root of variance, indicating the spread of data.
   Interquartile Range (IQR): Range of the middle 50% of data.

c. **Frequency Distribution -** Summarizes data by showing how often each value or range of values occurs. Tools are –
   Frequency Tables: Lists the number of occurrences for each category or range.
   Relative Frequencies: Express frequencies as percentages.

d. **Data Visualization:** Uses graphical methods to present data visually for easier understanding and pattern recognition. Common Methods used for data visualisation are,
   Histograms: Show data distribution.
   Boxplots: Highlight spread and outliers.
   Pie Charts: Represent proportions of categorical data.
   Bar Graphs: Compare different categories.

e. **Shape and Distribution:** Describes the overall pattern of data, including its symmetry and skewness.

2.3 Need & Significance descriptive Statistics:

Descriptive statistics play a fundamental role in data analysis by providing a clear and organized summary of data. They are indispensable in research, business, healthcare, education, and many other fields for understanding and interpreting data effectively. Following are some examples for the respective domains:

- **Business**: Summarizing sales performance, customer demographics, or financial data.
- **Healthcare**: Analyzing patient health indicators or treatment outcomes.
- **Education**: Summarizing test scores or student demographics.
- **Social Sciences**: Understanding population characteristics through surveys.

The need and significance of descriptive statistics lie in their ability to transform raw data into meaningful information. They help researchers, analysts, and decision-makers understand the "what" of the data, paving the way for deeper analysis and informed decision-making. By organizing and summarizing data, descriptive statistics are essential for interpreting real-world phenomena and communicating insights effectively.

2.3.1 Need for Descriptive Statistics:

a. **Data Summarization**: Large datasets can be overwhelming and difficult to interpret in their raw form. Descriptive statistics simplify the data by summarizing it into key metrics like mean, median, standard deviation, or percentages.
   Example: Summarizing the test scores of a class to find the average performance.

b. **Foundation for Advanced Analysis:** Acts as the starting point for any statistical or data-driven analysis. Helps identify patterns, trends, and anomalies, which are crucial for further inferential or predictive statistical methods.
   Example: Detecting outliers in customer purchase data before building a sales forecast model.

c. **Decision-Making:** Provides actionable insights to support decision-making processes across industries. Helps identify strengths, weaknesses, opportunities, and threats in a dataset.
   Example: A business may use descriptive statistics to identify its best-selling product categories.

d. **Quality Control and Validation**: Validates the integrity and quality of the dataset by identifying inconsistencies, missing values, or errors. Example: Checking if data is symmetrically distributed before applying parametric tests.

e.

f. **Communication and Presentation**: Facilitates the effective communication of data findings to stakeholders through tables, graphs, and summary statistics.Example: Presenting monthly sales growth using a bar chart and average values.

2.3.2 Significance for Descriptive Statistics:

a. Provides Insights
   Gives a comprehensive overview of the data by describing its key characteristics.
   Example: Identifying whether customer satisfaction scores are generally high, low, or balanced.

b. Enables Comparisons

Allows for the comparison of different datasets or groups within the same dataset.

Example: Comparing average test scores of students across multiple schools.

c. Identifies Patterns and Trends

Highlights recurring patterns, relationships, or distributions within the dataset.

Example: Understanding seasonal trends in sales through monthly averages.

d. Ensures Data Readiness

Prepares data for inferential statistics by highlighting its structure and distribution.

Example: Ensuring a variable follows a normal distribution before applying a t-test.

e. Aids in Resource Allocation

Helps allocate resources efficiently by identifying areas requiring focus based on data insights.

Example: Allocating medical supplies based on the frequency of certain diseases in a region.

f. Supports Evidence-Based Research

Provides the foundation for drawing meaningful conclusions from research data.

Example: Describing demographic statistics of participants in a clinical trial to set the context.

2.4 Brief statistical description of descriptive statistics

In R, descriptive statistics summarize and describe the characteristics of a dataset through measures of central tendency, dispersion, and distribution. These measures are fundamental for understanding data and preparing it for further analysis.

a. **Central Tendency:** Central tendency summarizes the dataset by identifying a central value that represents the data.

Syntax:

```
mean(data$column_name) # Replace 'column_name' with the variable name
```

Median: The middle value in the ordered dataset.

Syntax:

```
median(data$column_name) # data will be the data used and column_name is the variable whose median has to be calculated
```

Mode: The most frequently occurring value.

Syntax:

```
#using R library
library(modeest) #load the library
mfv(data$column_name)  # mfv = Most Frequent Value
```

b. Dispersion: Dispersion measures the spread or variability of data around the central tendency.

Range: The difference between the maximum and minimum values.

```
range(data$column_name)  #range() is the function used
```

or it can be calculated using :

```
#calculate minimum and maximum values
v1 <- max(variable)  # max() function
v2<- min(variable)   #min() function
range_v <- v1 - v2  # by definition of range
```

Variance: The average of the squared differences from the mean.

```
var(data$column_name)  #var() is the function  used
```

Standard Deviation: The square root of the variance.

```
sd(data$column_name)  #incorporate the function sd
```

Interquartile Range (IQR): The range of the middle 50% of the data.

```
IQR(data$column_name)  #incorporate function iqr
```

c. Frequency Distribution: Frequency distribution summarizes categorical or grouped numerical data.

Table of Frequencies:

Syntax -

```
table(data$column_name)
```

Proportions/Relative Frequencies:

Syntax –

```
prop.table(table(data$column_name))
```

d. Shape of Data: Understanding the shape of data through skewness and kurtosis.

Skewness: Indicates asymmetry in the data.

```
library(e1071)  #load the library
skewness(data$column_name)
```

Kurtosis: Measures the "tailed ness" of the data.

```
kurtosis(data$column_name)
```

e. **Summary Statistics**: R provides built-in functions to generate quick summaries of data.

Summary Function's syntax:

```
summary(data)
```

f. **Visualization for Descriptive Statistics:** Visual tools in R complement numerical summaries by presenting data graphically.

Histogram: Shows the distribution of numerical data.

```
hist(data$column_name, main = "Histogram", xlab = "Values", col = "skyblue")
```

Boxplot: Highlights data spread and outliers.

```
boxplot(data$column_name, main = "Boxplot", ylab = "Values")
```

Bar Chart: Displays frequencies for categorical data.

```
barplot(table(data$column_name), main = "Bar Chart", col = "pink")
```

This analysis provides a comprehensive view of the dataset, helping to understand its characteristics and prepare for further statistical exploration.

2.5 Frequency tables

A **frequency table** is a method of summarizing and organizing data to display the number of occurrences (frequencies) of each unique value or category in a dataset. It provides a clear and concise view of the data distribution, helping to identify patterns, trends, and relationships.

2.5.1 Frequency Tables in R

Frequency tables are used to show how often each value or category appears in a dataset. They are especially useful for categorical data analysis, as they help in identifying patterns, distribution, and frequency distribution of variables.

2.5.2 Creating Frequency Tables

For Categorical Data: The **table()** function is used to create frequency tables for categorical variables.

For Continuous Data: To categorize continuous data into bins, use the **cut()** function and then apply **table()**.

Relative Frequencies: The **prop.table()** function calculates the proportion of each category in the dataset.

R code :

```
Data<-read.csv(choose.files(),header=TRUE)
View(Data)
str(Data)

## 'data.frame': 886 obs. of 13 variables:
## $ Id : int 2 4 9 10 13 14 17 21 23 24 ... ##
$ Age : int 18 26 21 21 21 26 23 23 23 22 ...
## $ Year : int 1 4 3 2 3 5 5 4 4 2 ...
## $ sex : int 1 1 2 2 1 2 2 1 2 2 ...
## $ Mother.Tongue: int 120 1 1 1 1 1 1 1 1 1 ...
## $ Part_Status : int 1 1 0 0 1 1 1 1 1 1 ...
## $ Job : int 0 0 0 1 0 1 0 1 1 0 ...
## $ Study_hrs : int 56 20 36 51 22 10 15 8 20
20 ... ## $ Health : int 3 4 3 5 4 2 3 4 2 5
...
## $ psyt : int 0 0 0 0 0 0 0 0 0 0 ...
   ## $ jspe : int 88 109 106 101 102 102 117 118 118 108 ...
## $       qcae_cog : int 62 55 64 52 58 48 58 65 69 56 ... ##
$ qcae_aff : int 27 37 39 33 28 37 38 40 46 36 ...

head(Data,5)

## Id Age Year sex Mother.Tongue Part_Status Job Study_hrs
Health psyt jspe
                ## 1 2 18 1 1 120 1 0 56 3 0
88
## 2 4 26 4 1 1 1 0 20 4 0 109
## 3 9 21 3 2 1 0 0 36 3 0 106
## 4 10 21 2 2 1 0 1 51 5 0 101
## 5 13 21 3 1 1 1 0 22 4 0 102
## qcae_cog qcae_aff
## 1 62 27
```

```
## 2 55 37
## 3 64 39
## 4 52 33
## 5 58 28

tail(Data,5)

## Id Age Year sex Mother.Tongue Part_Status Job Study_hrs
Health psyt jspe
## 882 1781 21 2 1 1 1 0 45 3 0 106
## 883 1785 20 2 2 1 0 0 13 3 0 113
## 884 1787 19 1 1 1 0 0 50 5 0 100
## 885 1789 24 5 2 1 0 0 20 2 1 120
## 886 1790 22 3 1 1 0 1 20 5 0 102
## qcae_cog qcae_aff
## 882 63 39
## 883 67 40
## 884 50 31
## 885 64 39
## 886 54 26

library(dplyr)
## Attaching package: 'dplyr'
#Frequency tables
table1<-table(Data$id)
print(table1)

## < table of extent 0 >

table2<-table(Data$sex)
print(table2)

##
## 1 2 3
```

```
## 275 606 5

table3<-table(Data$glang)
print(table3)

##
## 1 15 20 37 54 60 63 90 92 95 98 102 104 106 108 114 118
120 121
## 717 31 22 3 1 3 5 45 1 1 1 27 4 6 1 1 2 2 13

table4<-table(Data$part)
print(table4)

##
## 0 1
## 387 499

table5<-table(Data$job)
print(table5)

##
## 0 1
## 577 309

table6<-table(Data$psyt)
print(table6)

##
## 0 1
## 687 199
```

Importance of Frequency Tables in Analysis:

- Helps summarize the distribution of categorical and continuous data.
- Provides insights into the data's general pattern, concentration, and the presence of any outliers or rare categories.

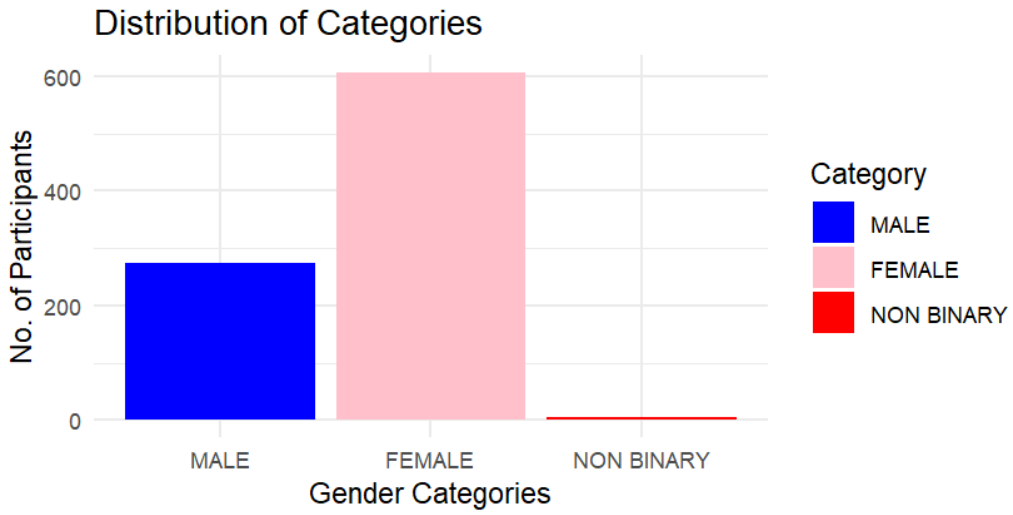
2.6 Statistical Diagrams in R

Statistical diagrams (or charts/plots) are visual representations of data that help in identifying patterns, trends, and distributions. R provides several ways to create statistical diagrams.

2.6.1 Bar Diagrams

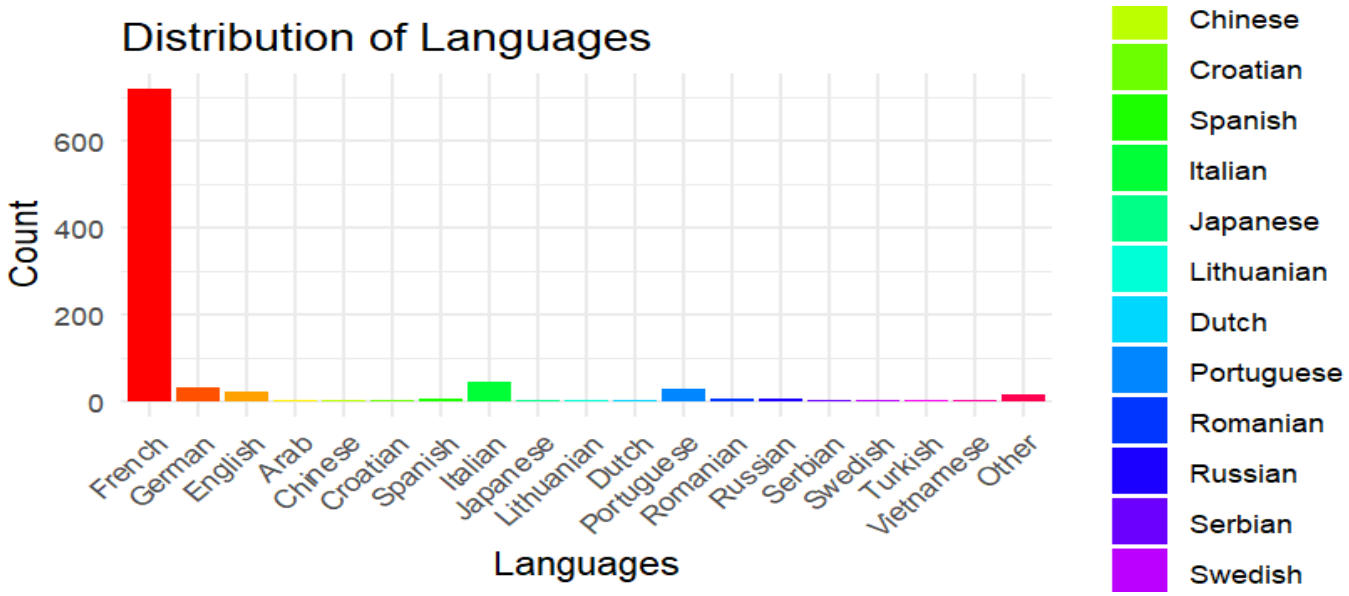
Bar charts are used to compare frequencies or values for categorical variables. They can be vertical or horizontal.

```
#Bar diagrams of each nominal variables
df <- as.data.frame(table2)
      colnames(df) <- c("Category", "Count")
df$Category <- factor(df$Category,
                      levels = c(1, 2, 3),
                      labels = c("MALE", "FEMALE", "NON
BINARY"))
ggplot(df, aes(x = Category, y = Count, fill = Category)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  labs(title = "Distribution of Categories",
       x = "Gender Categories",
       y = "No. of Participants") +
  scale_fill_manual(values = c("blue", "pink", "red"))
```



```
df <- as.data.frame(table3)
colnames(df) <- c("Code", "Count")
# Replace numeric values with language labels
df$Language <- factor(df$Code,
                      levels = c(1, 15, 20, 37, 51, 52,
53, 54, 59, 60, 62, 63, 82,
                                 83, 84, 85, 86, 87, 88,
89, 90, 92, 93, 94, 95,
                                 96, 98, 100, 101, 102,
104, 106, 108, 112, 113,
                                 114, 116, 117, 118, 119,
120, 121),
labels = c("French", "German", "English", "Arab",
"Basque",
                                 "Bulgarian", "Catalan",
"Chinese", "Korean",
                                 "Croatian", "Danish",
"Spanish", "Estonian",
                                 "Finnish", "Galician",
"Greek", "Hebrew", "Hindi",
                                 "Hungarian",
"Indonesian", "Italian", "Japanese",
                                 "Kazakh", "Latvian",
"Lithuanian", "Malay",
                                 "Dutch", "Norwegian",
"Polish", "Portuguese",
                                 "Romanian", "Russian",
"Serbian", "Slovak",
                                 "Slovenian", "Swedish",
"Czech", "Thai",
                                 "Turkish", "Ukrainian",
"Vietnamese", "Other"))
```

```
ggplot(df, aes(x = Language, y = Count, fill = Language))
+
  geom_bar(stat = "identity") +
  theme_minimal() +
  labs(title = "Distribution of Languages",
       x = "Languages",
       y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust =
1)) +
  scale_fill_manual(values =
rainbow(length(unique(df$Language))))
```



Interpretation:

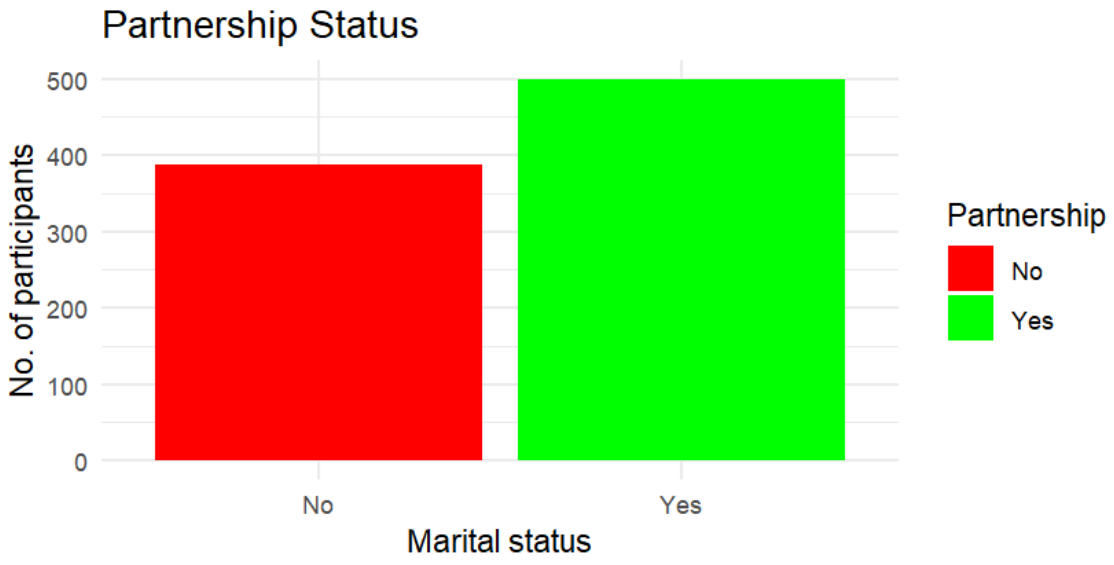
The bar graph titled **"Distribution of Languages"** visualizes the frequency of participants' mother tongues within the dataset.

- **Dominance of French**: French is overwhelmingly the most common mother tongue among the participants, with a count exceeding 600.
- **Low Representation of Other Languages**: Most other languages, such as German, English, Arabic, and others, have significantly fewer participants, with counts close to or below 50.
- Many languages (e.g., Lithuanian, Japanese, Turkish) show minimal representation, suggesting a lack of diversity in linguistic backgrounds.

- **Color-Coded Language Groups :** Each language is represented with a unique color, making it easy to distinguish between categories. However, languages with lower counts might not be as visually prominent.\

Bar plot for partnership status:

```
df <- as.data.frame(table4)
colnames(df) <- c("Code", "Count")
# Replace numeric values with labels
df$Partnership <- factor(df$Code,
                         levels = c(0, 1),
                         labels = c("No", "Yes"))
ggplot(df, aes(x = Partnership, y = Count, fill =
Partnership)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  labs(title = "Partnership Status",
       x = "Marital status",
       y = "No. of participants") +
scale_fill_manual(values = c("red", "green")
```



Interpretation:

- The bar chart depicts the distribution of partnership status among the participants
- Approximately 400 people reported not having a partner
- around 500 people indicated they do have a partner.
- A higher proportion of individuals in a relationship compared to those without.

Bar plot for Employment Status:

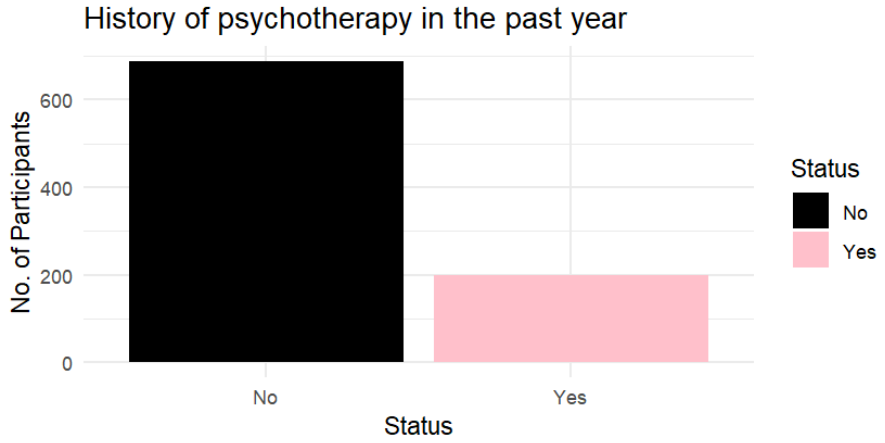
```
df <- as.data.frame(table5)
colnames(df) <- c("Code", "Count")
df$Employment <- factor(df$Code,
                        levels = c(0, 1),
                        labels = c("Unemployed",
"Employed"))
ggplot(df, aes(x = Employment, y = Count, fill =
Employment)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  labs(title = "Employment Status",
       x = "Status",
       y = "No. of Employees") +
  scale_fill_manual(values = c("blue", "orange"))
```

Interpretation:

- The bar chart illustrates the employment status of individuals.
- Around 600 people are unemployed
- Approximately 400 people are employed. The number of unemployed individuals exceeds the number of employed individuals in the dataset.

Bar plot for Psychotherapy taken by participants:

```
df <- as.data.frame(table6)
# Rename the columns for better readability
colnames(df) <- c("Code", "Count")
df$Partnership <- factor(df$Code,
                         levels = c(0, 1),
                         labels = c("No", "Yes"))
ggplot(df, aes(x = Partnership, y = Count, fill =
Partnership))+
  geom_bar(stat = "identity") +
  theme_minimal() +
  labs(title = "Pyschotherapy",
       x = "No. of Pyschotherapy",
       y = "Frequency") +
  scale_fill_manual(values = c("black", "pink"))
```



Interpretation:

- The bar chart illustrates the frequency of individuals who have or have not participated in psychotherapy.
- The majority of individuals (represented by the black bar) have not participated in psychotherapy, with a frequency exceeding 600.
- A smaller proportion (represented by the pink bar) have participated in psychotherapy, with a frequency under 300.

2.6.2 Pie Charts

Pie charts are used to show the proportion of each category as slices of a circle.

Syntax:

```
pie(x, labels = names(x), edges = 200, radius = 0.8,
    clockwise = FALSE, init.angle = if(clockwise) 90 else 0,
    density = NULL, angle = 45, col = NULL, border = NULL,
    lty = NULL, main = NULL, ...)
```

Arguments

x : a vector of non-negative numerical quantities. The values in x are displayed as the areas of pie slices.

Labels : one or more expressions or character strings giving names for the slices. Other objects are coerced by as.graphicsAnnot. For empty or NA (after coercion to character) labels, no label nor pointing line is drawn.

edges : the circular outline of the pie is approximated by a polygon with this many edges.

radius: the pie is drawn centered in a square box whose sides range from −1 to 1

init.angle: number specifying the starting angle (in degrees) for the slices. Defaults to 0 (i.e., ‘3 o'clock’) unless clockwise is true where init.angle defaults to 90 (degrees), (i.e., ‘12 o'clock’).

density : the density of shading lines, in lines per inch. The default value of NULL means that no shading lines are drawn. Non-positive values of density also inhibit the drawing of shading lines.

angle: the slope of shading lines, given as an angle in degrees (counter-clockwise).

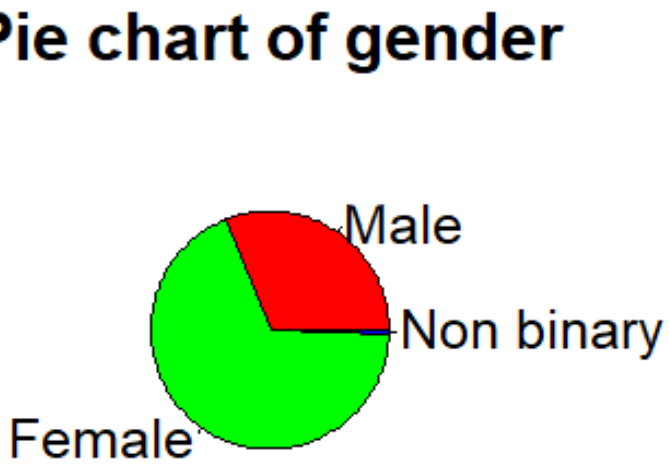
col: a vector of colors to be used in filling or shading the slices. If missing a set of 6 pastel colours is used, unless density is specified when par("fg") is used.

border, lty : (possibly vectors) arguments passed to polygon which draws each slice.

main: an overall title for the plot.

Pie chart for gender distribution of the data:

```
df <- as.data.frame(table3)
    labels=c("Male","Female","Non binary")
x<-pie(table2,labels=labels,main="Pie chart of
gender", col = rainbow(length(table2)))
```



Interpretation:

- The pie chart shows the gender distribution of the dataset.
- A majority of the population identifies as followed by a smaller proportion identifying as male
- The smallest segment represents individuals identifying as non-binary).
- This indicates a predominantly female representation in the data.

Pie chart for Partnership Status:

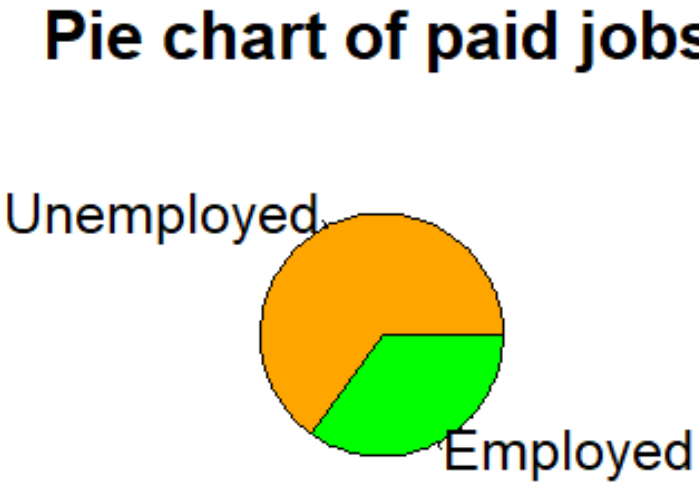
```
colors <- c( "blue", "green")
labels=c("no","yes")
z<-pie(table4,labels=labels,main="Pie   chart   of
Partnership Status",col=colors)
```

Interpretation:

- The pie chart illustrates the partnership status of individuals in the dataset.
- The green segment ("yes") represents individuals who have a partner, while the blue segment ("no") represents those who do not.
- The chart indicates that a slightly larger proportion of individuals are in a partnership compared to those who are not.

Pie chart for Employment Status:

```
colors <- c( "orange", "green")
labels=c("Unemployed","Employed")
w<-pie(table5,labels=labels,main="Pie chart of
paid jobs",col=colors)
```



Interpretation:

- The chart shows a distribution for employment status among participants.
- Majority or more than 50 % of the sample population is unemployed while approximately 30-35% of people are employed.

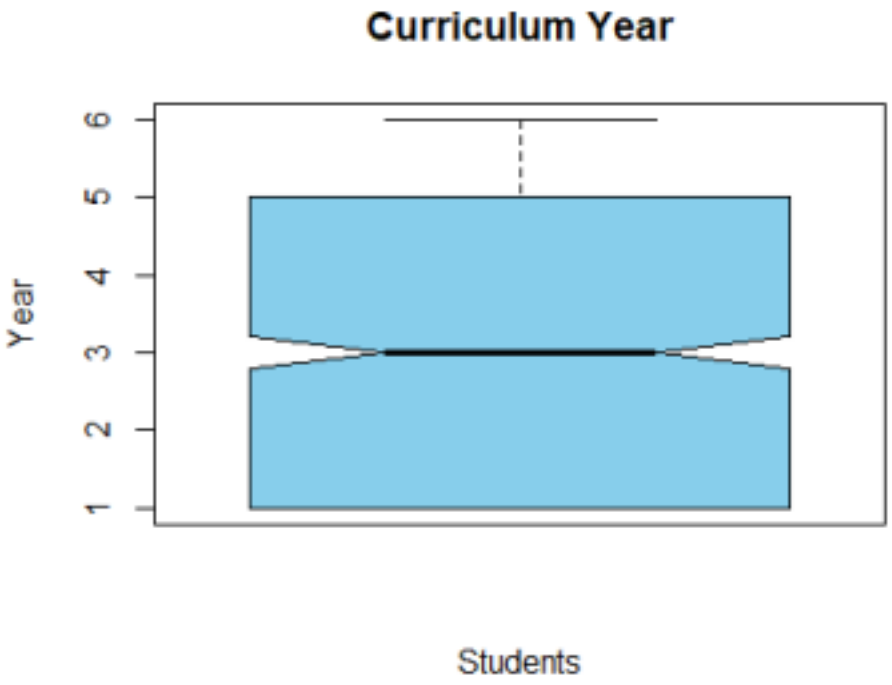
2.6.3 BoxPlots

A **boxplot** is a graphical representation of data that displays its distribution through five summary statistics: minimum, first quartile (Q1), median, third quartile (Q3), and maximum. It helps identify data spread, central tendency, and potential outliers effectively.

```
boxplot(formula, data = NULL, ..., subset, na.action = NULL,
        xlab = mklab(y_var = horizontal),
        ylab = mklab(y_var =!horizontal),
        add = FALSE, ann = !add, horizontal = FALSE,
        drop = FALSE, sep = ".", lex.order = FALSE)
```

Boxplot for curriculum year participants enrolled in:

```
#Boxplot for ordinal variables Year and Health bplot1<-
boxplot(x=Data$Year,
xlab="Students",
ylab="Year",
main="Curriculum Year",
notch=TRUE,varwidth = TRUE,col = "skyblue")
```



Interpretation:

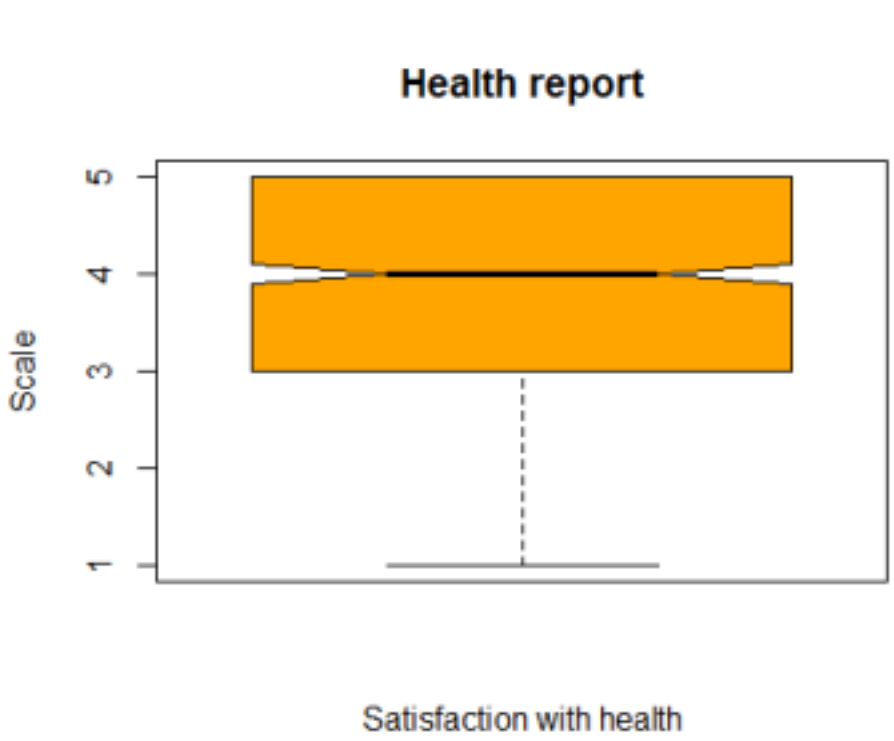
The boxplot represents the distribution of students across curriculum years (from 1 to 6).

- Median (Central Line): The median lies approximately at year 3, indicating that most students are centered around their third year of study.
- Interquartile Range (Box): The box spans years 2 to 4, suggesting that the middle 50% of students are enrolled in these curriculum years.
- Whiskers: The whiskers extend slightly beyond the box, indicating a few students are in earlier (year 1) and later years (year 5 or 6), but they are not extreme outliers.
- Symmetry: The boxplot appears symmetric, indicating a fairly even distribution of students across the years.

This visualization shows that the majority of students are concentrated in the middle curriculum years, with fewer students in the beginning or advanced years.

Boxplot for Health report amongst participants:

```
bplot2<-boxplot(x=Data$Health,
xlab="Satisfaction with health",
ylab="Scale",
main="Health report",
notch=TRUE,varwidth=TRUE,col= "orange")
```



Interpretation:

This boxplot depicts the distribution of participants' satisfaction with their health on a scale (likely from 1 to 5). Interpretation are:

1. **Median (Central Line):** The median is approximately at 4, indicating that the majority of participants rate their health satisfaction positively.
2. **Interquartile Range (Box):** The interquartile range spans from about 3 to 4.5, showing that the middle 50% of participants generally have moderate to high satisfaction with their health.
3. **Whiskers:** The whiskers extend from 1 to nearly 5. This suggests that while most participants rate their health satisfaction within the central range, there are individuals reporting very low satisfaction (as low as 1) or the maximum satisfaction (5).
4. **Symmetry:** The box is nearly symmetric around the median, indicating a fairly balanced distribution of health satisfaction ratings.

This visualization suggests that most participants feel satisfied with their health, but a small subset has very low satisfaction levels, highlighting potential areas for health improvement programs.

By incorporating frequency tables and statistical diagrams in the project, effective summary , interpretation, and insights from the dataset can be given.

2.7 R-code and the output of some descriptive statistics

2.7.1 Code to obtain descriptive statistics for continuous scaled variables

```
#Continuous Scaled Variables
#Variable Age
Mean<-mean(Data$Age)
Mean
## [1] 22.38375
Mini<-min(Data$Age)
Mini
## [1] 17
Maxi<-max(Data$Age)
Maxi
## [1] 49
Range<-range(Data$Age)
Range
## [1] 17 49
Median<-median(Data$Age)
Median
## [1] 22
ag6<-Data$Age
Mode<-ag6 %>%
as_tibble() %>%
count(value) %>%
filter(n==max(n)) %>%
pull(value)
as.numeric(Mode)
## [1] 23
quartiles<-quantile(Data$Age,probs=c(0.25,0.50,0.75))
quartiles
## 25% 50% 75%
## 20 22 24
```

```
inter_quart<-IQR(Data$Age)
inter_quart
## [1] 4
quart_deviation<-(quartiles[3]-quartiles[1])/2
quart_deviation
## 75%
## 2
mean_deviation_arb<-mean(abs(Data$Age-25))
mean_deviation_arb
## [1] 3.465011
mean_deviation_mean<-mean(abs(Data$Age-mean(Data$Age)))
mean_deviation_mean
## [1] 2.354142
MD_median<-mad(Data$Age)
MD_median
## [1] 2.9652
Variance<-var(Data$Age)
Variance
## [1] 10.89438
sd_value<-sd(Data$Age)
sd_value
## [1] 3.300664
CV<-sd_value/Mean
CV
## [1] 0.147458
quintiles <- quantile(Data$Age, probs = seq(0, 1, by = 0.2))
quintiles
## 0% 20% 40% 60% 80% 100%
## 17 20 21 23 24 49
octiles <- quantile(Data$Age, probs = seq(0, 1, by = 0.125))
octiles
```

```
## 0% 12.5% 25% 37.5% 50% 62.5% 75% 87.5% 100% ## 17 19 20 21 22 23 24 25 49
deciles <- quantile(Data$Age, probs = seq(0, 1, by = 0.1)) deciles
## 0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100% ## 17 19 20 21 21 22 23 23 24 26 49
percentiles <- quantile(Data$Age, probs = seq(0, 1, by = 0.01)) percentiles
## 0% 1% 2% 3% 4% 5% 6% 7% 8% 9% 10% 11% 12%
## 17.00 18.00 18.00 18.00 18.00 18.00 18.00 19.00 19.00 19.00 19.00 19.00 19.00
## 13% 14% 15% 16% 17% 18% 19% 20% 21% 22% 23% 24% 25%
## 19.00 19.00 19.00 19.00 19.00 19.30 20.00 20.00 20.00 20.00 20.00 20.00 20.00
## 26% 27% 28% 29% 30% 31% 32% 33% 34% 35% 36% 37% 38%
## 20.00 20.00 20.00 21.00 21.00 21.00 21.00 21.00 21.00 21.00 21.00 21.00 21.00
## 39% 40% 41% 42% 43% 44% 45% 46% 47% 48% 49% 50% 51%
## 21.00 21.00 21.00 21.00 22.00 22.00 22.00 22.00 22.00 22.00 22.00 22.00 22.00
## 52% 53% 54% 55% 56% 57% 58% 59% 60% 61% 62% 63% 64%
## 22.00 22.00 22.00 22.75 23.00 23.00 23.00 23.00 23.00 23.00 23.00 23.00 23.00
## 65% 66% 67% 68% 69% 70% 71% 72% 73% 74% 75% 76% 77%
## 23.00 23.00 23.00 23.00 23.00 23.00 24.00 24.00 24.00 24.00 24.00 24.00 24.00
## 78% 79% 80% 81% 82% 83% 84% 85% 86% 87% 88% 89% 90%
## 24.00 24.00 24.00 24.00 24.00 25.00 25.00 25.00 25.00 25.00 25.00 25.00 26.00
## 91% 92% 93% 94% 95% 96% 97% 98% 99% 100% ## 26.00 26.00 26.00 27.00 28.00 28.60 29.00 32.00 35.00 49.00
```

#Variable Study_hrs

```
Mean<-mean(Data$Study_hrs)
Mean
## [1] 25.28894
Median<-median(Data$Study_hrs)
Median
## [1] 25
stud_h<-Data$Study_hrs
Mode<-stud_h %>%
as_tibble() %>%
count(value) %>%
filter(n==max(n)) %>%
pull(value)
as.numeric(Mode)
## [1] 30
Range<-range(Data$Study_hrs)
Range
## [1] 0 70
Mini<-min(Data$Study_hrs)
Mini
## [1] 0
Maxi<-max(Data$Study_hrs)
Maxi
## [1] 70
quartiles<-quantile(Data$Study_hrs,probs =
c(0.25,0.50,0.75)) quartiles
## 25% 50% 75%
## 12 25 36
inter_quart<-IQR(Data$Study_hrs)
inter_quart
## [1] 24
```

```
quart_deviation<-(quartiles[3]-quartiles[1])/2
quart_deviation
## 75%
## 12
mean_devi_arb<-mean(abs(Data$Study_hrs-100))
mean_devi_arb
## [1] 74.71106
md_about_mean<-mean(abs(Data$Study_hrs-Mean))
md_about_mean
## [1] 13.29827
md_about_mode<-mean(abs(Data$Study_hrs-Mode))
md_about_mode
## [1] 13.91422
md_about_median<-mad(Data$Study_hrs)
md_about_median
## [1] 19.2738
Var<-var(Data$Study_hrs)
Var
## [1] 253.6972
SD<-sd(Data$Study_hrs)
SD
## [1] 15.92788
CV<-SD/Mean
CV
## [1] 0.6298356
quintiles <- quantile(Data$Study_hrs, probs = seq(0, 1, by =
0.2)) quintiles
## 0% 20% 40% 60% 80% 100%
## 0 10 20 30 40 70
octiles <- quantile(Data$Study_hrs, probs = seq(0, 1, by =
0.125)) octiles
```

```
## 0% 12.5% 25% 37.5% 50% 62.5% 75% 87.5% 100% ## 0 7 12 18
25 30 36 45 70
deciles <- quantile(Data$Study_hrs, probs = seq(0, 1, by =
0.1)) deciles
## 0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100% ## 0 6 10 15
20 25 30 35 40 45 70
percentiles <- quantile(Data$Study_hrs, probs = seq(0, 1, by
= 0.01)) percentiles
## 0% 1% 2% 3% 4% 5% 6% 7% 8% 9% 10% 11% 12%
## 0.00 0.00 1.00 2.00 2.00 3.00 4.00 4.00 5.00 5.00 6.00
6.00 7.00
## 13% 14% 15% 16% 17% 18% 19% 20% 21% 22% 23% 24% 25%
## 7.00 7.00 8.00 8.00 8.00 9.00 10.00 10.00 10.00 10.00
10.00 10.00 12.00
## 26% 27% 28% 29% 30% 31% 32% 33% 34% 35% 36% 37% 38%
## 12.00 12.00 13.00 14.00 15.00 15.00 15.00 15.00 15.00
16.00 16.00 18.00
19.00
## 39% 40% 41% 42% 43% 44% 45% 46% 47% 48% 49% 50% 51%
## 20.00 20.00 20.00 20.00 20.00 20.00 20.00 21.00 22.00
23.00 24.00 25.00 25.00
## 52% 53% 54% 55% 56% 57% 58% 59% 60% 61% 62% 63% 64%
## 25.00 25.00 25.00 26.00 27.00 28.00 30.00 30.00 30.00
30.00 30.00 30.00 30.00
## 65% 66% 67% 68% 69% 70% 71% 72% 73% 74% 75% 76% 77%
## 30.00 30.10 32.00 33.00 35.00 35.00 35.00 35.00 35.00
35.00 36.00 36.00 38.45
## 78% 79% 80% 81% 82% 83% 84% 85% 86% 87% 88% 89% 90%
## 40.00 40.00 40.00 40.00 40.00 40.00 41.40 42.00 43.10
45.00 45.00 45.00 45.00
## 91% 92% 93% 94% 95% 96% 97% 98% 99% 100% ## 50.00 50.00
50.00 50.00 54.50 56.00 60.00 60.00 65.00 70.00
```

```
#Variable JSPE(Jefferson Scale of Physiscian Empathy)
Mean<-mean(Data$jspe)
Mean
## [1] 106.3747
Median<-median(Data$jspe)
Median
## [1] 107
JSPE<-Data$jspe
Mode<-JSPE %>%
as_tibble() %>%
count(value) %>%
filter(n==max(n)) %>%
pull(value)
as.numeric(Mode)
## [1] 113
Range<-range(Data$jspe)
Range
## [1] 67 125
Mini<-min(Data$jspe)
Mini
## [1] 67
Maxi<-max(Data$jspe)
Maxi
## [1] 125
quartiles<-quantile(Data$jspe,probs = c(0.25,0.50,0.75))
quartiles
## 25% 50% 75%
## 101 107 113
inter_quart<-IQR(Data$jspe)
inter_quart
## [1] 12
```

```
quart_deviation<-(quartiles[3]-quartiles[1])/2
quart_deviation
## 75%
## 6
mean_devi_arb<-mean(abs(Data$jspe-75))
mean_devi_arb
## [1] 31.39955
md_about_mean<-mean(abs(Data$jspe-Mean))
md_about_mean
## [1] 6.925809
md_about_mode<-mean(abs(Data$jspe-Mode))
md_about_mode
## [1] 8.397291
md_about_median<-mad(Data$jspe)
md_about_median
## [1] 8.8956
Var<-var(Data$jspe)
Var
## [1] 77.15886
SD<-sd(Data$jspe)
SD
## [1] 8.784012
CV<-SD/Mean
CV
## [1] 0.08257612
quintiles <- quantile(Data$jspe, probs = seq(0, 1, by =
0.2)) quintiles
## 0% 20% 40% 60% 80% 100%
## 67 99 105 110 113 125
octiles <- quantile(Data$jspe, probs = seq(0, 1, by =
0.125)) octiles
```

```
## 0% 12.5% 25% 37.5% 50% 62.5% 75% 87.5% 100% ## 67 97 101 104 107 110 113 116 125
deciles <- quantile(Data$jspe, probs = seq(0, 1, by = 0.1)) deciles
## 0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100% ## 67 95 99 102 105 107 110 112 113 117 125
percentiles <- quantile(Data$jspe, probs = seq(0, 1, by = 0.01)) percentiles
## 0% 1% 2% 3% 4% 5% 6% 7% 8% 9% 10%
## 67.00 81.00 86.00 88.00 89.00 91.00 91.00 92.00 93.80 94.00 95.00
## 11% 12% 13% 14% 15% 16% 17% 18% 19% 20% 21%
## 95.35 96.00 97.00 97.90 98.00 98.00 98.00 99.00 99.00 99.00 100.00
## 22% 23% 24% 25% 26% 27% 28% 29% 30% 31% 32%
## 100.00 101.00 101.00 101.00 102.00 102.00 102.00 102.00 102.00 103.00 103.00
## 33% 34% 35% 36% 37% 38% 39% 40% 41% 42% 43%
## 103.00 104.00 104.00 104.00 104.00 104.00 105.00 105.00 105.00 105.00 106.00
## 44% 45% 46% 47% 48% 49% 50% 51% 52% 53% 54%
## 106.00 106.00 106.00 107.00 107.00 107.00 107.00 107.00 108.00 108.00 108.00
## 55% 56% 57% 58% 59% 60% 61% 62% 63% 64% 65%
## 108.00 109.00 109.00 109.00 109.00 110.00 110.00 110.00 110.00 110.00 110.00
## 66% 67% 68% 69% 70% 71% 72% 73% 74% 75% 76%
## 111.00 111.00 111.00 111.00 112.00 112.00 112.00 112.00 112.00 113.00 113.00
## 77% 78% 79% 80% 81% 82% 83% 84% 85% 86% 87%
## 113.00 113.00 113.00 113.00 114.00 114.00 115.00 115.00 115.00 115.00 116.00
```

```
## 88% 89% 90% 91% 92% 93% 94% 95% 96% 97% 98%
## 116.00 117.00 117.00 118.00 118.00 118.00 119.00 119.00
120.00 120.00 121.00
## 99% 100%
## 122.00 125.00
#Variable qcae_cog
Mean<-mean(Data$qcae_cog)
Mean
## [1] 58.52596
Median<-median(Data$qcae_cog)
Median
## [1] 58
qcae_cog<-Data$qcae_cog
Mode<-qcae_cog %>%
as_tibble() %>%
count(value) %>%
filter(n==max(n)) %>%
pull(value)
as.numeric(Mode)
## [1] 58
Range<-range(Data$qcae_cog)
Range
## [1] 37 76
Mini<-min(Data$qcae_cog)
Mini
## [1] 37
Maxi<-max(Data$qcae_cog)
Maxi
## [1] 76
quartiles<-quantile(Data$qcae_cog,probs = c(0.25,0.50,0.75))
quartiles
```

```
## 25% 50% 75%
## 54 58 63
inter_quart<-IQR(Data$qcae_cog)
inter_quart
## [1] 9
quart_deviation<-(quartiles[3]-quartiles[1])/2
quart_deviation
## 75%
## 4.5
mean_devi_arb<-mean(abs(Data$qcae_cog-80))
mean_devi_arb
## [1] 21.47404
md_about_mean<-mean(abs(Data$qcae_cog-Mean))
md_about_mean
## [1] 5.203981
md_about_mode<-mean(abs(Data$qcae_cog-Mode))
md_about_mode
## [1] 5.187359
md_about_median<-mad(Data$qcae_cog)
md_about_median
## [1] 5.9304
Var<-var(Data$qcae_cog)
Var
## [1] 43.16938
SD<-sd(Data$qcae_cog)
SD
## [1] 6.570341
CV<-SD/Mean
CV
## [1] 0.1122637
```

```
quintiles <- quantile(Data$qcae_cog, probs = seq(0, 1, by =
0.2)) quintiles
## 0% 20% 40% 60% 80% 100%
## 37 53 57 60 64 76
octiles <- quantile(Data$qcae_cog, probs = seq(0, 1, by =
0.125)) octiles
## 0% 12.5% 25% 37.5% 50% 62.5% 75% 87.5% 100% ## 37 51 54
56 58 60 63 66 76
deciles <- quantile(Data$qcae_cog, probs = seq(0, 1, by =
0.1)) deciles
## 0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100% ## 37 50 53
55 57 58 60 62 64 67 76
percentiles <- quantile(Data$qcae_cog, probs = seq(0, 1, by
= 0.01)) percentiles
## 0% 1% 2% 3% 4% 5% 6% 7% 8% 9% 10% 11% 12% 13% 14% 15%
## 37 41 44 47 48 48 49 49 50 50 50 51 51 51 52 52
## 16% 17% 18% 19% 20% 21% 22% 23% 24% 25% 26% 27% 28% 29%
30% 31%
## 52 53 53 53 53 54 54 54 54 54 54 55 55 55 55 55
## 32% 33% 34% 35% 36% 37% 38% 39% 40% 41% 42% 43% 44% 45%
46% 47%
## 56 56 56 56 56 56 56 57 57 57 57 57 57 58 58 58
## 48% 49% 50% 51% 52% 53% 54% 55% 56% 57% 58% 59% 60% 61%
62% 63%
## 58 58 58 58 59 59 59 59 59 60 60 60 60 60 60 61
## 64% 65% 66% 67% 68% 69% 70% 71% 72% 73% 74% 75% 76% 77%
78% 79%
## 61 61 61 61 62 62 62 62 62 62 63 63 63 63 64 64
## 80% 81% 82% 83% 84% 85% 86% 87% 88% 89% 90% 91% 92% 93%
94% 95%
## 64 64 65 65 65 65 66 66 66 67 67 67 68 68 69 70
## 96% 97% 98% 99% 100%
```

```
## 70 71 72 73 76
#Variable qcae_aff
Mean<-mean(Data$qcae_aff)
Mean
## [1] 34.78442
Median<-median(Data$qcae_aff)
Median
## [1] 35
qcae_aff<-Data$qcae_aff
Mode<-qcae_aff %>%
as_tibble() %>%
count(value) %>%
filter(n==max(n)) %>%
pull(value)
as.numeric(Mode)
## [1] 37
Range<-range(Data$qcae_aff)
Range
## [1] 18 48
Mini<-min(Data$qcae_aff)
Mini
## [1] 18
Maxi<-max(Data$qcae_aff)
Maxi
## [1] 48
quartiles<-quantile(Data$qcae_aff,probs = c(0.25,0.50,0.75))
quartiles
## 25% 50% 75%
## 31 35 39
inter_quart<-IQR(Data$qcae_aff)
inter_quart
```

```
## [1] 8
quart_deviation<-(quartiles[3]-quartiles[1])/2
quart_deviation
## 75%
## 4
mean_devi_arb<-mean(abs(Data$qcae_aff-80))
mean_devi_arb
## [1] 45.21558
md_about_mean<-mean(abs(Data$qcae_aff-Mean))
md_about_mean
## [1] 4.325158
md_about_mode<-mean(abs(Data$qcae_aff-Mode))
md_about_mode
## [1] 4.585779
md_about_median<-mad(Data$qcae_aff)
md_about_median
## [1] 5.9304
Var<-var(Data$qcae_aff)
Var
## [1] 28.9128
SD<-sd(Data$qcae_aff)
SD
## [1] 5.377062
CV<-SD/Mean
CV
## [1] 0.1545825
quintiles <- quantile(Data$qcae_aff, probs = seq(0, 1, by =
0.2)) quintiles
## 0% 20% 40% 60% 80% 100%
## 18 30 33 36 39 48
octiles <- quantile(Data$qcae_aff, probs = seq(0, 1, by =
0.125)) octiles
```

```
## 0% 12.5% 25% 37.5% 50% 62.5% 75% 87.5% 100% ## 18 28 31
33 35 37 39 41 48
deciles <- quantile(Data$qcae_aff, probs = seq(0, 1, by =
0.1)) deciles
## 0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100% ## 18 27 30
32 33 35 36 38 39 42 48
percentiles <- quantile(Data$qcae_aff, probs = seq(0, 1, by
= 0.01)) percentiles
## 0% 1% 2% 3% 4% 5% 6% 7% 8% 9% 10% 11% 12% 13% 14% 15%
## 18.0 22.0 24.0 24.0 25.0 25.0 26.0 26.0 27.0 27.0 27.0
28.0 28.0 28.0 28.0 29.0
## 16% 17% 18% 19% 20% 21% 22% 23% 24% 25% 26% 27% 28% 29%
30% 31%
## 30.0 30.0 30.0 30.0 30.0 31.0 31.0 31.0 31.0 31.0 32.0
32.0 32.0 32.0 32.0 32.0
## 32% 33% 34% 35% 36% 37% 38% 39% 40% 41% 42% 43% 44% 45%
46% 47%
## 32.0 33.0 33.0 33.0 33.0 33.0 33.0 33.0 33.0 34.0 34.0
34.0 34.0 34.0 34.1 35.0
## 48% 49% 50% 51% 52% 53% 54% 55% 56% 57% 58% 59% 60% 61%
62% 63%
## 35.0 35.0 35.0 35.0 35.0 35.0 36.0 36.0 36.0 36.0 36.0
36.0 36.0 37.0 37.0 37.0
## 64% 65% 66% 67% 68% 69% 70% 71% 72% 73% 74% 75% 76% 77%
78% 79%
## 37.0 37.0 37.0 37.0 37.0 38.0 38.0 38.0 38.0 38.0 38.0
39.0 39.0 39.0 39.0 39.0
## 80% 81% 82% 83% 84% 85% 86% 87% 88% 89% 90% 91% 92% 93%
94% 95%
## 39.0 40.0 40.0 40.0 40.0 40.0 40.0 41.0 41.0 41.0 42.0
42.0 42.0 42.0 43.0 43.0
## 96% 97% 98% 99% 100%
```

```
## 44.0 44.0 46.0 46.0 48.0
```

2.7.2 Code to obtain bivariate tables

```
#Bivariate table and diagrams of ordinal variables
biv_table<-table(Data$Year,Data$Health)
biv_table
##
## 1 2 3 4 5
## 1 13 31 41 101 59
## 2 4 13 25 68 25
## 3 4 15 21 69 34
## 4 3 10 20 56 34
## 5 6 11 18 60 32
## 6 7 7 11 48 40
barplot(biv_table,col="navy")
#Bivariate table and diagrams of nominal variables
biv_table1<-table(Data$sex,Data$Mother.Tongue)
biv_table1
##
## 1 15 20 37 54 60 63 90 92 95 98 102 104 106 108 114 118 120 121
## 1 223 9 7 0 0 0 2 20 1 0 0 5 1 1 1 0 0 1 4
## 2 491 22 15 3 1 3 2 24 0 1 1 22 3 5 0 1 2 1 9
## 3 3 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0
barplot(biv_table1,col="navy")
biv_table2<-table(Data$Part_Status,Data$Job)
biv_table2
##
## 0 1
## 0 262 125
## 1 315 184
```

```
barplot(biv_table2,col="navy")
#Trivariate table of nominal scaled variables
triv_table1<-table(Data$sex,Data$Mother.Tongue,Data$Part_Status) triv_table1
## , , = 0
##
##
## 1 15 20 37 54 60 63 90 92 95 98 102 104 106 108 114 118 120 121
## 1 100 4 3 0 0 0 2 9 1 0 0 3 0 0 0 0 0 0 1
## 2 198 10 11 1 0 3 1 11 0 1 1 10 2 3 0 0 0 0 7
## 3 3 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0
##
## , , = 1
##
##
## 1 15 20 37 54 60 63 90 92 95 98 102 104 106 108 114 118 120 121
## 1 123 5 4 0 0 0 0 11 0 0 0 2 1 1 1 0 0 1 3
## 2 293 12 4 2 1 0 1 13 0 0 0 12 1 2 0 1 2 1 2
## 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
triv_table2<-table(Data$Part_Status,Data$Job,Data$Health) triv_table2
## , , = 1
##
##
## 0 1
## 0 13 5
## 1 11 8
##
## , , = 2
##
##
## 0 1
## 0 26 22
```

```
## 1 22 17
##
## , , = 3
##
##
## 0 1
## 0 46 17
## 1 48 25
##
## , , = 4
##
##
## 0 1
## 0 119 48
## 1 148 87
##
## , , = 5
##
##
## 0 1
## 0 58 33
## 1 86 47
```

CHAPTER-3

Predictive Statistics / Inferential and Relational Data Analysis

3.1 Predictive Statistics- Meaning and Definition

Predictive Statistics refers to the use of statistical methods and models to analyze historical data in order to predict future outcomes or behaviors. It involves building mathematical models to identify patterns or relationships within the data, which can then be used to forecast future events or trends.

3.1.1 Definition:

Predictive statistics is a branch of statistics that employs statistical algorithms, data mining, and machine learning techniques to make predictions about future or unknown events based on historical data. These models aim to estimate the likelihood of future outcomes, trends, or behaviors by learning from patterns and relationships found in past data.

"Predictive modelling is the process of developing a mathematical tool or model that generates an accurate prediction"

3.1.2 Key Components of Predictive Statistics:

a. **Data Collection:** Gathering relevant historical data that can be analyzed to identify trends or patterns.

b. **Data Preprocessing:** Cleaning and transforming data to prepare it for analysis.

c. **Model Building:** Developing statistical models (such as regression analysis, decision trees, or machine learning models) that can make predictions based on the available data.

d. **Prediction:** Using the trained model to predict future outcomes or values.

e. **Model Evaluation:** Testing and validating the predictive model to ensure its accuracy and reliability.

3.2 Inferential Data Analysis – Importance & Significance

Inferential Data Analysis involves drawing conclusions about a population based on a sample of data. It enables researchers and analysts to make predictions, generalizations, and decisions beyond the immediate data at hand, allowing for broader insights into trends, relationships, and behaviors.

3.2.1. Importance of Inferential Data Analysis:

a. **Making Predictions About a Population:**

Data from a sample is used to make predictions or inferences about a larger population. This is particularly useful when it's impractical or impossible to gather data from an entire population (e.g., a national survey, clinical trial outcomes).

b. **Testing Hypotheses:**
Inferential analysis is central to hypothesis testing. It helps determine whether there is enough evidence in a sample to support a particular hypothesis about a population.

c. **Generalizing Results:** With inferential statistics, conclusions drawn from a sample can be generalized to a broader population, given that the sample is representative. This is key in fields like market research, healthcare, and social science where testing the entire population is not feasible.

d. **Understanding Relationships Between Variables:**
Through methods like regression, correlation, or analysis of variance (ANOVA), inferential analysis helps to understand the relationships between different variables. This can lead to actionable insights in diverse domains.

e. **Estimating Uncertainty and Confidence**: Inferential analysis provides not just estimates, but also a measure of uncertainty. Confidence intervals, for instance, quantify the degree of certainty around estimates, helping in making decisions under uncertainty, which is especially important in business, healthcare, and engineering.

3.2.1 Significance of Inferential Data Analysis:

a. **Informed Decision-Making:** It provides a robust framework for making informed decisions based on statistical evidence.

b. **Cost-Efficiency:** Collecting data from an entire population can be expensive and time-consuming. Inferential statistics allows for drawing conclusions from smaller, manageable samples, making it a cost-effective approach.

c. **Risk Assessment:** Inferential statistics plays a crucial role in assessing and managing risks, especially in finance and insurance.

d. **Predicting Future Trends:** Inferential analysis is widely used for forecasting future trends based on existing data. Whether it's predicting stock market performance, economic indicators, or population health trends, inferential statistics helps to anticipate and prepare for future scenarios.

3.3 Relational Data Analysis – Correlation and regression

Relational Data Analysis is a crucial component of statistical methods that focuses on understanding and quantifying the relationships between variables. Two primary techniques used for

this purpose are **Correlation** and **Regression**. These methods help identify patterns, strengths, and directions of relationships between variables, allowing analysts to make more informed predictions and decisions.

3.3.1 Correlation

Correlation is a statistical measure that describes the degree and direction of the relationship between two or more variables. It quantifies how much one variable changes in response to the change in another. The most commonly used measure of correlation is the **Pearson Correlation Coefficient (r)**, which ranges from -1 to +1.

- **r = +1:** Perfect positive correlation (as one variable increases, the other increases proportionally).
- **r = -1:** Perfect negative correlation (as one variable increases, the other decreases proportionally).
- **r = 0:** No correlation (the variables are unrelated).

3.3.2 Regression:

Regression analysis is a statistical technique that models and analyzes the relationship between a dependent variable and one or more independent variables. Unlike correlation, which only easures the strength and direction of a relationship, regression predicts the value of the dependent variable based on the independent variable(s).

- **Simple Linear Regression:** Involves one independent variable and one dependent variable. The relationship is modeled with a straight line, described by the equation:

$$Y = \beta 0 + \beta 1X + \epsilon$$

Where Y is the dependent variable, X is the independent variable, β0 is the intercept, β1 is the slope of the line, and ϵ is the error term.

- **Multiple Linear Regression:** Involves multiple independent variables. The equation becomes:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \beta nXn + \epsilon$$

where $X_1, X_2, \ldots, X_n$, are the independent variables.

3.4 Parameter estimation in regression

Parameter estimation in regression refers to the process of determining the values of the coefficients (parameters) in a regression model. These coefficients represent the relationship between the dependent variable and independent variables. The goal is to find the values of the model parameters that best fit the observed data.

There are different methods of parameter estimation like - Ordinary Least Squares (OLS), Maximum Likelihood Estimation (MLE), Ridge and Lasso Regression.

3.5 Estimation of correlation parameters

The **estimation of correlation parameters** involves calculating the correlation coefficient for the sample data. It is done through :

Step1 : Calculate the Sample Correlation Coefficient (r)

Step2 : Set up the Hypothesis

- **Null Hypothesis (H_0)**: There is no correlation between the variables (r=0).
- **Alternative Hypothesis (H_1)**: There is a significant correlation between the variables ($r \neq 0$)

Step3: Test statistic $t = \frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}}$, where t follows a **Student's t-distribution** with $n-2$ degrees of free dom.

3.6 Parametric Hypothesis Tests (Single & Two sample cases)

Parametric hypothesis tests are statistical tests that assume the data follows a certain distribution, typically a normal distribution. These tests are used to infer properties about a population based on sample data. In particular, **single-sample** and **two-sample** tests are among the most common parametric hypothesis tests, used to test claims about population means, proportions, or other parameters.

3.6.1 Single-Sample Parametric Tests

In a **single-sample test**, we compare the sample mean to a known population mean (or a hypothesized value). The objective is to determine whether the sample mean is significantly different from the population mean.

Tests:

a. One-Sample t-Test
b. One-Sample Z-Test

3.6.2 Two-Sample Parametric Tests

A **two-sample parametric test** compares the means of two independent samples to determine whether they are significantly different from each other. These tests are widely used when there are two groups to compare, such as control and experimental groups in an experiment.

Tests:

c. Two-Sample t-Test (Independent t-Test)
d. Paired t-Test (Dependent t-Test)

e. Two-Sample Z-Test

3.7 Non-Parametric Hypothesis Tests (Single & Two sample cases)

Non-parametric hypothesis tests are statistical tests that do not assume a specific distribution for the data. They are particularly useful when the data violates the assumptions of parametric tests, such as normality. Non-parametric tests are typically used:

- when the sample size is small
- when the data is ordinal or categorical
- or when the data is not normally distributed.

i. Single-Sample Non-Parametric Tests

a. **Sign Test:** The **sign test** is used to test the median of a single sample when the data are paired or the distribution is unknown. It is useful when we want to test if the median of the population is equal to a hypothesized value.

b. **Wilcoxon Signed-Rank Test**: The **Wilcoxon signed-rank test** is an alternative to the one-sample t-test when the data are not normally distributed. It tests whether the median of the population is equal to a hypothesized value.

ii. Two- Sample Non-Parametric Tests

a. Mann-Whitney U Test (also called the Wilcoxon Rank-Sum Test) : The Mann-Whitney U test is used to compare two independent samples to determine whether their population distributions are different. This test is an alternative to the two-sample t-test and is particularly useful when the data are ordinal or not normally distributed.

b. Wilcoxon Rank-Sum Test: The **Wilcoxon rank-sum test** is another test for comparing two independent samples. It is similar to the Mann-Whitney U test and is often used interchangeably, especially when the assumptions of normality are violated.

c. Kolmogorov-Smirnov Test (Two-Sample KS Test): The Kolmogorov-Smirnov test is used to compare two independent samples to see if they come from the same distribution. This test is sensitive to differences in both location and shape of the empirical cumulative distribution functions of the two samples.

3.8 R-code and the outputs of some predictive statistics

3.8.1 R codes for Correlation coefficients between the variables

```
#Coefficient of association for two nominal/ordinal scaled
variables corr_coeff1<-cor(Data$sex,Data$Part_Status)
corr_coeff1
## [1] 0.0003153407
corr_coeff2<-cor(Data$sex,Data$Mother.Tongue)
corr_coeff2
## [1] 0.005641029
corr_coeff3<-cor(Data$sex,Data$Job)
corr_coeff3
## [1] 0.02087935
corr_coeff4<-cor(Data$sex,Data$psyt)
corr_coeff4
## [1] 0.1582622
corr_coeff5<-cor(Data$psyt,Data$Job)
corr_coeff5
## [1] 0.06013777
corr_coeff6<-cor(Data$Year,Data$Health)
corr_coeff6
## [1] 0.08402314
#Coefficient of correlation among continuous scaled
variables corr_coef1<-cor(Data$Age,Data$Study_hrs)
corr_coef1
## [1] -0.2935569
corr_coef2<-cor(Data$Age,Data$jspe)
corr_coef2
## [1] 0.2232209
corr_coef3<-cor(Data$Age,Data$qcae_aff)
corr_coef3
## [1] -0.008130475
corr_coef4<-cor(Data$Study_hrs,Data$qcae_aff)
corr_coef4
## [1] -0.03226837

corr_coeff5<-cor(Data$psyt,Data$Job)
```

```
corr_coeff5

## [1] 0.06013777

corr_coeff6<-cor(Data$Year,Data$Health)
corr_coeff6

## [1] 0.08402314
```

For continuous scaled variables:

```
    #Coefficient of correlation among continuous scaled
    variables corr_coef1<-cor(Data$Age,Data$Study_hrs)
    corr_coef1
    ## [1] -0.2935569
    corr_coef2<-cor(Data$Age,Data$jspe)
    corr_coef2
    ## [1] 0.2232209
    corr_coef3<-cor(Data$Age,Data$qcae_aff)
    corr_coef3
    ## [1] -0.008130475
    corr_coef4<-cor(Data$Study_hrs,Data$qcae_aff)
    corr_coef4
## [1] -0.03226837
```

3.8.2 R codes for Simple linear regression coefficients

```
              #Simple linear regression coefficients
model1<-lm(Data$Age~Data$Study_hrs)
summary(model1)
##
## Call:
## lm(formula = Data$Age ~ Data$Study_hrs)
##
## Residuals:
```

```
## Min 1Q Median 3Q Max
## -5.7055 -2.0619 -0.2588 1.2470 26.0512
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|) ## (Intercept) 23.922137
0.199095 120.15 <2e-16 *** ## Data$Study_hrs -0.060833 0.006663
-9.13 <2e-16 *** ## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.157 on 884 degrees of freedom ##
Multiple R-squared: 0.08618, Adjusted R-squared: 0.08514 ## F-
statistic: 83.36 on 1 and 884 DF, p-value: < 2.2e-16
model2<-lm(Data$Age~Data$jspe)
summary(model2)
##
## Call:
## lm(formula = Data$Age ~ Data$jspe)
##
## Residuals:
## Min 1Q Median 3Q Max
## -5.7782 -2.1056 -0.4329 1.3122 25.8928
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|) ## (Intercept) 13.46135
1.31491 10.237 < 2e-16 *** ## Data$jspe 0.08388 0.01232 6.809
1.82e-11 *** ## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.219 on 884 degrees of freedom
## Multiple R-squared: 0.04983, Adjusted R-squared: 0.04875 ##
F-statistic: 46.36 on 1 and 884 DF, p-value: 1.818e-11
model3<-lm(Data$Age~Data$qcae_aff)
```

```
summary(model3)
##
## Call:
## lm(formula = Data$Age ~ Data$qcae_aff)
##
## Residuals:
## Min 1Q Median 3Q Max
## -5.3677 -2.3677 -0.3727 1.6073 26.6373
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|) ## (Intercept) 22.557350
0.726646 31.043 <2e-16 *** ## Data$qcae_aff -0.004991 0.020645 -
0.242 0.809 ## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.302 on 884 degrees of freedom ##
Multiple R-squared: 6.61e-05, Adjusted R-squared: -0.001065 ##
F-statistic: 0.05844 on 1 and 884 DF, p-value: 0.809
model4<-lm(Data$qcae_cog~Data$qcae_aff)
summary(model4)
##
## Call:
## lm(formula = Data$qcae_cog ~ Data$qcae_aff)
##
## Residuals:
## Min 1Q Median 3Q Max
## -21.2263 -4.1085 0.0381 4.1286 17.7220
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|) ## (Intercept) 47.5304
1.3973 34.016 < 2e-16 *** ## Data$qcae_aff 0.3161 0.0397 7.963
5.16e-15 *** ## ---
```

```
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.35 on 884 degrees of freedom ##
Multiple R-squared: 0.06692, Adjusted R-squared: 0.06587 ## F-
statistic: 63.4 on 1 and 884 DF, p-value: 5.16e-15
```

Intepretation on model :

I. **Model 1 (Age vs Study Hours):**
   - A statistically significant inverse relationship between age and study hours.
   - The model explains about 8.6% of the variance in age.

II. Model 2 (Age vs Job Satisfaction):
   - A statistically significant positive relationship between age and job satisfaction.
   - The model explains about 5% of the variance in age.

III. Model 3 (Age vs QCAE Affectivity):
   - No significant relationship between age and affectivity.
   - The model does not explain much of the variance in age.

IV. Model 4 (Cognitive Score vs Affectivity):
   - A significant positive relationship between cognitive score and affectivity.
   - The model explains about 6.7% of the variance in cognitive scores.

3.8.3 R codes for Simple linear regression coefficients without intercepts

```
#Simple linear regression coefficients without intercepts model1<-
lm(Data$Age~Data$qcae_aff-1)
summary(model1)
##
## Call:
## lm(formula = Data$Age ~ Data$qcae_aff - 1)
##
## Residuals:
## Min 1Q Median 3Q Max
## -11.1623 -2.6217 0.2634 3.2634 28.2905
##
## Coefficients:
```

```
## Estimate Std. Error t value Pr(>|t|) ## Data$qcae_aff 0.628381
0.004555 138 <2e-16 *** ## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.772 on 885 degrees of freedom ##
Multiple R-squared: 0.9556, Adjusted R-squared: 0.9555 ## F-
statistic: 1.903e+04 on 1 and 885 DF, p-value: < 2.2e-16
model2<-lm(Data$qcae_cog~Data$qcae_aff-1)
summary(model2)
##
## Call:
## lm(formula = Data$qcae_cog ~ Data$qcae_aff - 1) ##
## Residuals:
## Min 1Q Median 3Q Max
## -27.931 -5.544 0.973 7.480 34.288
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|) ## Data$qcae_aff 1.650679
0.009205 179.3 <2e-16 *** ## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.644 on 885 degrees of freedom ##
Multiple R-squared: 0.9732, Adjusted R-squared: 0.9732 ## F-
statistic: 3.216e+04 on 1 and 885 DF, p-value: < 2.2e-16
```

3.8.4 Multiple linear regression with/without intercepts

```
#Multiple linear regression with/without intercepts mul_regg<-
lm(Data$jspe~Data$qcae_cog+Data$qcae_aff) summary(mul_regg)
##
## Call:
## lm(formula = Data$jspe ~ Data$qcae_cog + Data$qcae_aff) ##
## Residuals:
```

```
## Min 1Q Median 3Q Max
## -38.403 -4.653 0.766 5.737 19.077
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 72.71206 2.71054 26.826 < 2e-16 *** ##
Data$qcae_cog 0.39335 0.04294 9.161 < 2e-16 *** ## Data$qcae_aff
0.30593 0.05247 5.831 7.73e-09 *** ## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.107 on 883 degrees of freedom ##
Multiple R-squared: 0.1501, Adjusted R-squared: 0.1482 ## F-
statistic: 78 on 2 and 883 DF, p-value: < 2.2e-16
mul_regg1<-lm(Data$jspe~Data$qcae_cog+Data$qcae_aff-1)
summary(mul_regg1)
##
## Call:
## lm(formula = Data$jspe ~ Data$qcae_cog + Data$qcae_aff - 1)
##
## Residuals:
## Min 1Q Median 3Q Max
## -36.841 -6.383 0.599 8.082 40.197
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## Data$qcae_cog 1.26060 0.03805 33.13 <2e-16 *** ##
Data$qcae_aff 0.91600 0.06366 14.39 <2e-16 *** ## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.92 on 884 degrees of freedom ##
Multiple R-squared: 0.9896, Adjusted R-squared: 0.9895 ## F-
statistic: 4.192e+04 on 2 and 884 DF, p-value: < 2.2e-16
```

3.8.5 Inferential Statistics

One Sample Test :

Used to test if the sample mean is equal to a hypothesized population mean (μ_0)

Syntax:

```
t.test(x, mu = hypothesized_mean, alternative = "two.sided",
conf.level = 0.95)
```

- **x**: Numeric vector of sample data.
- **mu**: Hypothesized population mean (μ_0).
- **alternative**: Specifies the alternative hypothesis:
- "two.sided" (default): Tests if the sample mean is different from μ_0.
- "less": Tests if the sample mean is less than μ_0.
- "greater": Tests if the sample mean is greater than μ_0.
- **conf.level**: Confidence level for the interval (default is 95%).

```
#INFERENTIAL STATISTICS
#One sample mean test
res_age<-t.test(Data$Age)
res_age
##
## One Sample t-test
##
## data: Data$Age
## t = 201.86, df = 885, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0 ## 95
percent confidence interval:
## 22.16611 22.60138
## sample estimates:
## mean of x
## 22.38375
res_stud_h<-t.test(Data$Study_hrs)
res_stud_h
##
```

```
## One Sample t-test
##
## data: Data$Study_hrs
## t = 47.26, df = 885, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0 ## 95
percent confidence interval:
## 24.23871 26.33917
## sample estimates:
## mean of x
## 25.28894
res_jspe<-t.test(Data$jspe)
res_jspe
##
## One Sample t-test
##
## data: Data$jspe
## t = 360.46, df = 885, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0 ## 95
percent confidence interval:
## 105.7955 106.9539
## sample estimates:
## mean of x
## 106.3747
res_qcae_cog<-t.test(Data$qcae_cog)
res_qcae_cog
##
## One Sample t-test
##
## data: Data$qcae_cog
## t = 265.14, df = 885, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0 ## 95
percent confidence interval:
## 58.09273 58.95918
## sample estimates:
```

```
## mean of x
## 58.52596
res_qcae_aff<-t.test(Data$qcae_aff)
res_qcae_aff
##
## One Sample t-test
##
## data: Data$qcae_aff
## t = 192.56, df = 885, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0 ## 95
percent confidence interval:
## 34.42988 35.13897
## sample estimates:
## mean of x
## 34.78442
```

Two Sample Mean Test :

```
test1<-t.test(Data$Age,Data$Study_hrs)
test1
##
## Welch Two Sample t-test
##
## data: Data$Age and Data$Study_hrs
## t = -5.3162, df = 960.87, p-value = 1.318e-07
## alternative hypothesis: true difference in means is not equal to 0 ## 95 percent
confidence interval:
## -3.977616 -1.832768
## sample estimates:
## mean of x mean of y
## 22.38375 25.28894
test2<-t.test(Data$qcae_cog,Data$qcae_aff)
test2
##
```

```
        ## Welch Two Sample t-test
        ##
        ## data: Data$qcae_cog and Data$qcae_aff
        ## t = 83.236, df = 1703.4, p-value < 2.2e-16
        ## alternative hypothesis: true difference in means is not equal to 0 ## 95 percent
        confidence interval:
        ## 23.18209 24.30098
        ## sample estimates:
        ## mean of x mean of y
## 58.52596 34.78442
```

One Sample Variance Test:

A one-sample variance test in R is used to test whether the variance of a sample differs from a hypothesized population variance. This is typically done using a chi-squared test for variance.

Syntax :

```
chisq.test(x, p = NULL, rescale.p = FALSE, simulate.p.value =
FALSE, B = 2000)
```

- x: A numeric vector of observed counts.
- p: A numeric vector of expected proportions. Defaults to equal proportions.
- rescale.p: Logical, rescale p to sum to 1 if TRUE.
- simulate.p.value: Logical, simulate p-values via Monte Carlo if TRUE.
- B: Number of replicates for simulation if simulate.p.value = TRUE.

```
#One Sample Variance Test
res_age<-chisq.test(Data$Age)
res_age
##
## Chi-squared test for given probabilities
##
## data: Data$Age
```

```
## X-squared = 430.74, df = 885, p-value = 1
res_stud_h<-chisq.test(Data$Study_hrs)
res_stud_h
##
## Chi-squared test for given probabilities
##
## data: Data$Study_hrs
## X-squared = 8878.3, df = 885, p-value < 2.2e-16
res_jspe<-chisq.test(Data$jspe)
res_jspe
##
## Chi-squared test for given probabilities
##
## data: Data$jspe
## X-squared = 641.93, df = 885, p-value = 1
res_qcae_cog<-chisq.test(Data$qcae_cog)
res_qcae_cog
##
## Chi-squared test for given probabilities
##
## data: Data$qcae_cog
## X-squared = 652.79, df = 885, p-value = 1
res_qcae_aff<-chisq.test(Data$qcae_aff)
res_qcae_aff
##
## Chi-squared test for given probabilities
##
## data: Data$qcae_aff
## X-squared = 735.61, df = 885, p-value = 0.9999
```

Two Sample Variances test:

A two-sample variances test in R is performed to determine if two samples have equal variances. The most common method for this is the F-test, which tests the null hypothesis that the variances of two populations are equal.

Syntax :

var.test(x, y, ratio = 1, alternative = "two.sided")

- x and y: The two numeric vectors representing the samples.
- ratio: The hypothesized ratio of the variances (default is 1 for equal variances).
- alternative: Specifies the alternative hypothesis:
- "two.sided" (default): Tests for unequal variances.
- "less": Tests if the variance of the first sample is less than the second.
- "greater": Tests if the variance of the first sample is greater than the second.

CODE:

```
#Two Sample Variances test
test1<-var.test(Data$Age,Data$Study_hrs)
test1
##
## F test to compare two variances
##
## data: Data$Age and Data$Study_hrs
## F = 0.042942, num df = 885, denom df = 885, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal
to 1 ## 95 percent confidence interval:
## 0.03763780 0.04899473
## sample estimates:
## ratio of variances
## 0.04294245
test2<-var.test(Data$qcae_cog,Data$qcae_aff)
test2
##
## F test to compare two variances
##
## data: Data$qcae_cog and Data$qcae_aff
## F = 1.4931, num df = 885, denom df = 885, p-value = 2.834e-09
## alternative hypothesis: true ratio of variances is not equal
to 1 ## 95 percent confidence interval:
## 1.308649 1.703524
```

```
## sample estimates:
## ratio of variances
## 1.493089
```

Interpretation:

Test 1: Comparing Variances of Age and Study Hours

- **Test Statistic (F):** The F value is 0.042942, which indicates the ratio of the variances of Data$Age to Data$Study_hrs.
- **Degrees of Freedom (df):** Both numerator and denominator have 885 degrees of freedom.
- P**-Value:** The p-value is less than 2.2e-16, which is highly significant.
- **Conclusion:** Since the p-value is much smaller than 0.05, we reject the null hypothesis and conclude that the variances of Data$Age and Data$Study_hrs are significantly different.
- **Confidence Interval:** The 95% confidence interval for the variance ratio is (0.0376, 0.0490), confirming that the true ratio is far from 1, supporting the conclusion of unequal variances.

Result- The variance of Study hours is significantly smaller than the variance of age of participants.

Test 2: Comparing Variances of QCAE Cognitive and QCAE Affective Scores

- **Test Statistic (F):** The F value is 1.4931, indicating the ratio of the variances of Data$qcae_cog to Data$qcae_aff.
- **P-Value:** The p-value is 2.834e-09, which is highly significant.
- **Conclusion:** Since the p-value is much smaller than 0.05, we reject the null hypothesis and conclude that the variances of Data$qcae_cog and Data$qcae_aff are significantly different.
- **Confidence Interval**: The 95% confidence interval for the variance ratio is (1.3086, 1.7035), indicating that the true variance ratio is greater than 1, suggesting higher variance in Data$qcae_cog.

Result -The variance of Data$qcae_cog is significantly larger than the variance of Data$qcae_aff

CHAPTER-4

Prescriptive Statistics / Advanced Data Analysis

4.1 Tests for more than two samples

When we analyse datasets with more than two groups, it becomes important to identify whether there are statistically significant differences among these groups. Such tests are frequently applied in areas like experimental research, social sciences, and medical studies to uncover relationships and patterns in data.

- **One-way ANOVA**:
  This test compares the means of three or more independent groups to determine if at least one group mean differs significantly. For example, it can be used to assess whether different teaching methods result in varied student performance or if various fertilizers affect crop yields differently.
- **Two-way ANOVA**:
  This approach expands on one-way ANOVA by including two independent variables, allowing researchers to evaluate both their individual (main) effects and their combined (interaction) effects. For instance, a study can use two-way ANOVA to analyse how drug type and dosage interact to affect recovery rates.
- **Kruskal-Wallis Test**:
  As a non-parametric alternative to ANOVA, this test is useful when data violate assumptions like normality. It focuses on medians instead of means, making it suitable for ordinal or skewed data.

These tests are essential for understanding relationships in multi-group studies and deriving meaningful conclusions

4.2 One-way Analysis of Variance (ANOVA)

One-way ANOVA is a statistical method used to determine whether there are significant differences between the means of three or more independent groups. The process follows a systematic approach:

a. **Formulating Hypotheses:** Begin by defining the null hypothesis (H_0) and the alternative hypothesis (H_1). The null hypothesis assumes that all group means are equal, indicating no significant differences. Conversely, the alternative hypothesis posits that at least one group mean differs, suggesting variability among groups.

b. **Data Collection:** Collect data from at least three independent groups. Each group should represent a distinct category or treatment level of the independent variable.

c. **Calculate Group Means:** Compute the mean value for each group individually. This helps summarize the central tendency within each group.
d. **Calculate Overall Mean:** Determine the average value of all data points combined, regardless of group affiliation. This serves as a baseline for comparison.
e. **Calculate Between-group Sum of Squares (SSB):** Measure the variability between the group means and the overall mean, weighted by the size of each group. This quantifies the extent to which group means differ from the overall mean.
f. **Calculate Within-group Sum of Squares (SSW):** Assess the variability within each group by summing the squared differences between individual scores and their respective group mean. This captures the variation due to random error or individual differences.
g. **F-statistic:** Compute the F-statistic as the ratio of the between-group variability (SSB) to the within-group variability (SSW), adjusted for their respective degrees of freedom. A larger F-value suggests greater differences between groups relative to random variation.
h. **Critical Value:** Obtain the critical value for the F-statistic from F-distribution tables or statistical software, based on the degrees of freedom and chosen significance level (α). This value serves as a benchmark for hypothesis testing.
i. **Decision Making:** Compare the p-value of the test with the significance level (α, usually 0.05).
    - If the p-value is less than α, reject the null hypothesis, concluding that significant differences exist among the group means.
    - If the p-value is greater than or equal to α, fail to reject the null hypothesis, indicating no significant differences among groups.

One-way ANOVA is widely used in experimental and observational studies to test hypotheses involving multiple groups, offering a straightforward way to detect significant differences.

4.3 Two-way Analysis of Variance (ANOVA)

Two-way ANOVA builds on the one-way ANOVA by incorporating two categorical independent variables, enabling researchers to evaluate both main effects and interaction effects. This method is particularly useful in complex experimental designs where the combined influence of two factors is of interest.

i. **Formulating Hypotheses:** Define the null hypothesis (H_0) as no interaction or main effects between the two factors on the dependent variable. The alternative hypothesis (H_1) posits that at least one main effect or interaction effect exists, indicating significant influence.

ii. **Data Organization:** Organize data into groups based on all possible combinations of the levels of the two factors. For example, if Factor A has three levels and Factor B has two levels, there will be six groups.

iii. **Calculate Sum of Squares (SS):** Partition the total variation in the dependent variable into components:

- Factor A (SSA): Variability due to the levels of Factor A.
- Factor B (SSB): Variability due to the levels of Factor B.
- Interaction (SSAB): Variability due to the interaction between Factors A and B.
- Residual (SSE): Remaining variability not explained by the factors or their interaction.

iv. **Degrees of Freedom (df):** Calculate the degrees of freedom for each component of variation, accounting for the number of levels in each factor and the total sample size.

v. **Mean Squares (MS):** Divide each sum of squares by its corresponding degrees of freedom to compute mean squares.

vi. **F-ratio:** For each source of variation (factors and interaction), calculate the F-ratio by dividing the mean square of the source by the mean square of the residuals.

vii. **Critical Value:** Determine the critical value of the F-statistic from F-distribution tables or software, based on the degrees of freedom and significance level.

viii. **Decision Making:** Compare the p-value with the significance level (α):

- If p-value $< \alpha$, reject the null hypothesis, concluding that significant effects exist.
- If p-value $\geq \alpha$, fail to reject the null hypothesis, indicating no significant effects.

Two-way ANOVA is a powerful tool for uncovering complex relationships and interactions between variables, widely applied in fields like psychology, biology, and engineering.

4.4 Multiple Linear Regression (MLR)

Multiple Linear Regression (MLR) is a robust statistical method designed to analyze the relationship between a dependent variable and multiple independent variables. Unlike simple linear regression, which evaluates the impact of just one predictor, MLR considers multiple factors simultaneously, offering a comprehensive view of how variables interact and influence outcomes.

This technique is widely applied in various domains, including economics, marketing, and social research, where understanding the combined effect of predictors is crucial. By calculating regression coefficients, MLR helps interpret how each independent variable contributes to changes in the dependent variable, enabling data-driven predictions and informed decision-making.

4.5 Estimation and Testing of Parameters in MLR

a. Estimation:

   Parameters in MLR are estimated using methods like Ordinary Least Squares (OLS), which minimizes the error between observed and predicted values.

b. Testing:

   Hypothesis testing is used to evaluate the significance of each predictor variable.

c. T-tests determine if the coefficients are significantly different from zero, while F-tests assess the overall model fit.

d. Assumptions:

   MLR relies on assumptions such as linearity, independence, homoscedasticity, and normality of residuals.

e. Practical Usage:

   - Enables researchers to identify key predictors and assess their impact on the dependent variable.
   - Widely applied in forecasting, risk analysis, and optimization tasks.

4.6 R-code and the outputs of some predictive statistics

4.6.1 ANOVA

```
#ANOVA for significance of more than two means
data<-data.frame(
value=c(Data$jspe,Data$qcae_cog,Data$qcae_aff),
group = factor(rep(c("Group1", "Group2", "Group3"))) )
result<-aov(value~group,data=data)
summary(result)
## Df Sum Sq Mean Sq F value Pr(>F)
## group 2 27 13.5 0.014 0.986
## Residuals 2655 2488319 937.2
#Post-hoc test using Tukey's HSD
post_hoc<-TukeyHSD(result)
print(post_hoc)
## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
```

```
## Fit: aov(formula = value ~ group, data = data)
##
## $group
## diff lwr upr p adj ## Group2-Group1 0.1106095 -3.300261
3.521480 0.9968168 ## Group3-Group1 -0.1354402 -3.546310
3.275430 0.9952311 ## Group3-Group2 -0.2460497 -3.656920
3.164821 0.9843481
```

Interpretation:

ANOVA :

- The high p-value (0.986) indicates that there is no statistically significant difference in the means of Group1, Group2, and Group3.
- We fail to reject the null hypothesis, suggesting that the group means are statistically similar.

Post-Hoc Summary:

- None of the pairwise comparisons showed a statistically significant difference in means, as all adjusted p-values are much greater than 0.05.

Overall:

- The ANOVA results indicate no significant difference in the means of Group1, Group2, and Group3.
- The Tukey's HSD post-hoc test confirms that all pairwise differences are statistically insignificant, with adjusted p-values close to 1.
- The observed group means are likely similar, and there is no evidence to suggest variability between the groups.

4.6.2 Multiple linear regression with/without intercepts

```
#Multiple linear regression with/without intercepts
mul_regg<-lm(Data$jspe~Data$qcae_cog+Data$qcae_aff)
summary(mul_regg)
##
## Call:
```

```
## lm(formula = Data$jspe ~ Data$qcae_cog + Data$qcae_aff)
##
## Residuals:
## Min 1Q Median 3Q Max
## -38.403 -4.653 0.766 5.737 19.077
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 72.71206 2.71054 26.826 < 2e-16 *** ##
Data$qcae_cog 0.39335 0.04294 9.161 < 2e-16 *** ##
Data$qcae_aff 0.30593 0.05247 5.831 7.73e-09 *** ## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1 ##
## Residual standard error: 8.107 on 883 degrees of freedom
## Multiple R-squared: 0.1501, Adjusted R-squared: 0.1482 ##
F-statistic: 78 on 2 and 883 DF, p-value: < 2.2e-16
mul_regg1<-lm(Data$jspe~Data$qcae_cog+Data$qcae_aff-1)
summary(mul_regg1)
##
## Call:
## lm(formula = Data$jspe ~ Data$qcae_cog + Data$qcae_aff -
1) ##
## Residuals:
## Min 1Q Median 3Q Max
## -36.841 -6.383 0.599 8.082 40.197
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## Data$qcae_cog 1.26060 0.03805 33.13 <2e-16 *** ##
Data$qcae_aff 0.91600 0.06366 14.39 <2e-16 *** ## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1 ##
```

```
## Residual standard error: 10.92 on 884 degrees of freedom
## Multiple R-squared: 0.9896, Adjusted R-squared: 0.9895 ##
F-statistic: 4.192e+04 on 2 and 884 DF, p-value: < 2.2e-16
```

4.5.2 Predictive value for Multiple linear regression with/without intercepts

```
X<-Data$qcae_cog
Y<-Data$qcae_aff
new_data<-data.frame(X=60,Y=20)
prediction1<-predict(mul_regg,newdata = new_data)
## Warning: 'newdata' had 1 row but variables found have 886
rows prediction1
## 1 2 3 4 5 6 7 8
## 105.35978 105.66558 109.81759 103.26183 104.09230 102.91213
107.15156 110.51686
## 9 10 11 12 13 14 15 16
## 113.92582 105.75301 101.90705 110.47321 111.56584 104.87888
105.66558 106.71456
## 17 18 19 20 21 22 23 24
## 103.96110 101.64477 108.02568 112.92062 110.86656 104.22338
110.77902 105.40343
## 25 26 27 28 29 30 31 32
## 108.41903 106.62713 105.84043 100.90185 111.95907 109.51154
115.01845 109.81759
## 33 34 35 36 37 38 39 40
## 109.11831 106.27755 104.61673 103.87368 105.18493 109.33681
109.33681 101.55723
## 41 42 43 44 45 46 47 48
## 107.58868 109.29316 106.19013 105.62193 104.35458 111.21614
105.66546 108.94346
## 49 50 51 52 53 54 55 56
## 102.51890 104.61661 105.27223 102.86848 104.39823 104.26703
104.31080 99.10994
## 57 58 59 60 61 62 63 64
```

```
## 107.67611 102.51878 99.19737 102.69363 111.39099 106.84563
108.98711 106.05893
## 65 66 67 68 69 70 71 72
## 105.49073 104.52930 103.65518 107.06413 111.60949 108.02568
103.34925 106.88941
## 73 74 75 76 77 78 79 80
## 111.56584 110.38579 104.66038 103.39315 100.68335 104.66038
104.09230 104.61661
## 81 82 83 84 85 86 87 88
## 103.48033 103.65518 105.40331 114.88737 107.19533 104.13595
102.99968
103.69895
## 89 90 91 92 93 94 95 96
## 108.11311 101.81950 105.92786 106.53983 109.90489 105.75301
107.02048 103.65518
## 97 98 99 100 101 102 103 104
## 102.73740 105.14116 101.90705 105.35966 103.34925 109.24939
108.55011 110.73536
## 105 106 107 108 109 110 111 112
## 106.88941 109.59909 107.32641 107.54491 101.81962 100.63945
108.41903 105.22858
## 113 114 115 116 117 118 119 120
## 109.11843 101.86328 109.55531 106.97683 105.97151 109.73016
110.69171 106.32121
## 121 122 123 124 125 126 127 128
## 108.20053 106.40863 104.00476 106.19001 111.30356 106.19013
109.90501 107.93826
## 129 130 131 132 133 134 135 136
## 105.18493 103.04333 103.30560 112.52739 112.17769 111.08506
108.55011 103.30548
## 137 138 139 140 141 142 143 144
## 112.65846 108.37538 107.58868 104.13595 114.14444 110.34201
104.92265 107.54491
## 145 146 147 148 149 150 151 152
```

```
## 105.53451 105.22858 108.94346 106.62713 109.11831 102.51890
105.35978 104.52930
## 153 154 155 156 157 158 159 160
## 105.79666 106.58348 112.35242 105.66558 111.47829 103.74272
107.54491 103.69883
## 161 162 163 164 165 166 167 168
## 101.07670 107.02048 106.75821 106.84563 102.86848 104.31068
113.66367 106.27743
## 169 170 171 172 173 174 175 176
## 105.40343 105.09750 100.68310 110.69171 100.24634 104.70415
114.27552 108.94358
## 177 178 179 180 181 182 183 184
## 101.55735 104.00488 106.71443 103.61140 111.30356 105.92786
105.27235 103.91745
## 185 186 187 188 189 190 191 192
## 101.99435 107.41383 107.67599 101.99435 101.81950 107.10791
102.08190 111.60949
## 193 194 195 196 197 198 199 200
## 105.66570 106.01528 114.14444 107.02048 106.40863 110.07986
102.78117 106.45228
## 201 202 203 204 205 206 207 208
## 106.05893 105.14116 105.97151 103.83003 103.52410 108.11299
98.97862 103.91745
## 209 210 211 212 213 214 215 216
## 106.67078 113.79474 107.80718 109.99244 101.90705 104.79146
108.06946 103.43668
## 217 218 219 220 221 222 223 224
## 103.17440 102.08190 108.81226 106.67078 105.70935 104.04853
105.62193 106.53971
## 225 226 227 228 229 230 231 232
## 107.06413 112.52727 107.93826 110.77902 105.31600 113.70732
101.99435 104.79146
## 233 234 235 236 237 238 239 240
```

```
## 107.54491 107.41383 108.72496 105.88408 103.26183 108.76873
103.17440 103.52410
## 241 242 243 244 245 246 247 248
## 112.61469 110.82279 107.45761 104.66050 109.68651 107.58856
105.40343 108.55011
## 249 250 251 252 253 254 255 256
## 105.14116 110.47321 102.47513 106.27743 104.83523 108.68131
110.64794 102.16920
## 257 258 259 260 261 262 263 264
## 98.62905 111.21626 101.12035 113.22654 105.40343 104.74780
111.12884 106.14636
## 265 266 267 268 269 270 271 272
## 109.20574 112.30877 106.67078 102.86848 105.84055 106.23378
114.36294 104.48565
## 273 274 275 276 277 278 279
280
## 106.32133 110.42944 108.59388 113.79474 104.13595 103.82991
104.79158 102.91225
## 281 282 283 284 285 286 287 288
## 108.11311 111.74069 105.88408 104.26715 107.89461 105.40343
110.21094 105.01008
## 289 290 291 292 293 294 295 296
## 108.33161 106.40863 104.13608 101.86340 106.53971 107.10791
111.34734 107.28263
## 297 298 299 300 301 302 303 304
## 106.84563 112.17769 108.41891 106.49593 101.99435 103.13075
112.70212 107.02048
## 305 306 307 308 309 310 311 312
## 103.43680 107.41383 98.54162 99.50317 100.90185 104.48565
109.07454 106.80198
## 313 314 315 316 317 318 319 320
## 109.16196 112.52727 104.52930 103.43668 106.27743 98.84767
107.80718 108.41903
## 321 322 323 324 325 326 327 328
```

```
## 104.92265 104.52930 113.00816 106.58348 102.69363 102.16920
107.63233 107.63233
## 329 330 331 332 333 334 335 336
## 107.58856 105.35966 110.03609 104.44188 114.31917 103.78638
103.30548 101.38262
## 337 338 339 340 341 342 343 344
## 99.10982 107.41383 110.69171 105.79678 102.56255 104.48565
104.35445 108.59376
## 345 346 347 348 349 350 351 352
## 102.78117 103.78638 107.80718 106.23378 102.16920 100.24622
108.33161 98.97875
## 353 354 355 356 357 358 359 360
## 106.53971 104.96631 104.09230 104.09218 103.69883 111.43476
105.05373 101.29520
## 361 362 363 364 365 366 367 368
## 107.54491 102.73740 105.14116 104.13595 113.40139 106.53971
104.04853 101.68843
## 369 370 371 372 373 374 375 376
## 108.59376 103.43668 102.38770 102.91225 107.19521 109.59909
107.50114 106.62713
## 377 378 379 380 381 382 383 384
## 103.39290 103.82991 105.57816 108.11311 107.02048 103.39290
105.75313 107.06413
## 385 386 387 388 389 390 391 392
## 105.57816 105.53451 110.07986 97.92989 109.46801 110.64794
107.45761 111.65326
## 393 394 395 396 397 398 399 400
## 105.40343 96.31272 108.28784 107.85083 105.75301 107.89461
100.42107 103.17452
## 401 402 403 404 405 406 407 408
## 105.49085 107.32641 108.11311 106.80210 103.34925 111.34722
108.06934 102.73740
## 409 410 411 412 413 414 415 416
```

```
## 106.84563 104.44188 105.40343 102.38770 104.92265 105.62193
111.91542 105.62193
## 417 418 419 420 421 422 423 424
## 106.84563 110.99764 104.96631 108.41903 109.29316 103.34925
99.67802 110.42944
## 425 426 427 428 429 430 431 432
## 103.13075 101.03280 107.80706 102.16920 104.00476 108.28796
111.30356 108.81239
## 433 434 435 436 437 438 439 440
## 105.35966 109.20574 104.74780 106.05893 103.34925 102.82483
110.42956 103.69895
## 441 442 443 444 445 446 447 448
## 100.24622 105.05373 111.56572 105.44708 107.98191 104.04853
112.26511 98.01732
## 449 450 451 452 453 454 455 456
## 111.08494 105.31600 104.79146 104.04853 105.92773 98.97875
105.79666 104.79146
## 457 458 459 460 461 462 463 464
## 109.55531 110.51686 106.93306 106.14648 110.91021 103.34925
100.24610
110.38579
## 465 466 467 468 469 470 471 472
## 104.92253 104.61673 107.02048 116.50442 105.66558 100.33365
108.41903 104.31080
## 473 474 475 476 477 478 479 480
## 99.85287 108.55011 109.38058 105.27223 100.94550 105.88420
105.62205 111.25991
## 481 482 483 484 485 486 487 488
## 110.91021 108.06946 111.52219 109.38058 106.71456 109.24939
105.49073 107.06426
## 489 490 491 492 493 494 495 496
## 105.62181 103.21805 103.83003 104.61673 107.19533 104.17960
109.07466 104.52930
## 497 498 499 500 501 502 503 504
```

```
## 108.46269 110.12351 109.73016 106.23378 107.15156 108.55011
113.40139 107.15168
## 505 506 507 508 509 510 511 512
## 101.03292 105.31600 104.61673 108.28796 105.62193 100.20245
112.35242 109.16196
## 513 514 515 516 517 518 519 520
## 107.76341 107.93826 106.27743 102.38770 103.74260 101.81962
109.81759 104.04853
## 521 522 523 524 525 526 527 528
## 101.20777 104.17960 107.02048 102.86860 100.77053 105.88408
106.45228 104.13595
## 529 530 531 532 533 534 535 536
## 109.03101 102.91225 106.53971 102.56255 105.18481 97.36169
104.79146 112.92062
## 537 538 539 540 541 542 543 544
## 109.94878 107.76341 107.76341 104.87900 111.39099 106.49593
107.06413 104.66050
## 545 546 547 548 549 550 551 552
## 108.28808 101.73220 104.61661 103.30560 100.90185 109.38058
104.39823 105.75301
## 553 554 555 556 557 558 559 560
## 107.06413 106.67091 103.61140 104.26703 108.24419 106.01528
107.98191 106.05893
## 561 562 563 564 565 566 567 568
## 107.19533 109.68639 106.93306 105.88420 104.70415 112.43996
110.60429 106.84563
## 569 570 571 572 573 574 575 576
## 107.93838 108.81239 102.34405 109.59909 105.79666 104.26715
109.11831 107.15156
## 577 578 579 580 581 582 583 584
## 106.53971 97.62385 110.82291 108.76873 103.69883 109.68651
105.31600 98.93497
## 585 586 587 588 589 590 591 592
```

```
## 94.60824 107.98203 105.70935 103.30560 112.48361 105.79678
103.21805 113.22654
## 593 594 595 596 597 598 599 600
## 104.61673 100.55215 107.28263 106.97671 109.33681 104.52930
105.27223 106.80198
## 601 602 603 604 605 606 607 608
## 106.58348 108.24419 113.48882 111.30356 106.45228 101.99435
108.94346 105.92786
## 609 610 611 612 613 614 615 616
## 103.52410 112.65846 102.25675 111.52206 105.35966 103.13075
110.69171 107.80718
## 617 618 619 620 621 622 623 624
## 104.92265 105.44708 111.21614 109.51166 108.68131 109.29316
102.73740 105.66558
## 625 626 627 628 629 630 631 632
## 104.83535 109.59909 101.73220 107.28263 108.46269 109.24939
109.59909 106.80186
## 633 634 635 636 637 638 639 640
## 103.39290 103.78625 107.28263 109.94866 103.83003 103.78638
105.31600 105.66558
## 641 642 643 644 645 646 647 648
## 110.60429 108.98723 97.18684 106.27743 105.92798 101.12035
100.15880 106.71456
## 649 650 651 652 653 654 655
656
## 105.22858 107.10791 105.97163 108.46281 102.25663 106.19001
108.94346 107.19533
## 657 658 659 660 661 662 663 664
## 106.75821 108.76873 109.55531 99.32845 102.08178 111.04129
103.91733 106.23378
## 665 666 667 668 669 670 671 672
## 104.26703 114.93102 103.48033 102.82483 102.47513 106.45228
105.31600 104.57308
## 673 674 675 676 677 678 679 680
```

```
## 107.63233 104.52930 103.91745 105.35966 106.36486 106.32121
107.76353 109.42424
## 681 682 683 684 685 686 687 688
## 105.62193 107.45748 103.43668 112.04661 109.42424 104.92265
108.76861 104.66038
## 689 690 691 692 693 694 695 696
## 101.90693 104.17973 106.67091 106.49606 101.64477 105.79678
107.41383 108.81239
## 697 698 699 700 701 702 703 704
## 109.55531 105.49073 105.35978 109.42424 101.55735 105.01008
102.91225 108.59401
## 705 706 707 708 709 710 711 712
## 104.26715 105.18481 108.81239 107.80718 103.69883 106.62713
108.46281 107.71976
## 713 714 715 716 717 718 719 720
## 103.83015 103.13075 107.85083 103.91733 106.80198 109.16196
102.16920 109.77394
## 721 722 723 724 725 726 727 728
## 112.22146 107.54491 101.60112 107.50126 104.09230 107.06413
109.64274 107.02048
## 729 730 731 732 733 734 735 736
## 112.26511 108.11311 105.22858 106.40863 109.16196 103.39303
105.31588 109.33681
## 737 738 739 740 741 742 743 744
## 109.77381 107.06413 103.17440 106.84563 102.16920 104.22338
104.31080 107.41371
## 745 746 747 748 749 750 751 752
## 110.60429 104.52930 101.95070 107.10779 106.36486 101.12035
107.50126 107.63233
## 753 754 755 756 757 758 759 760
## 110.07974 107.23898 106.23378 101.42627 111.21614 101.12035
110.29836 110.12351
## 761 762 763 764 765 766 767 768
```

```
## 110.34201 106.19013 104.52930 107.32641 104.92265 110.82279
105.84055 110.73536
## 769 770 771 772 773 774 775 776
## 102.99955 106.40851 112.43984 109.90501 105.31600 108.11311
112.09026 112.61469
## 777 778 779 780 781 782 783 784
## 110.12351 106.62713 110.12351 111.30356 109.24951 107.80718
106.45240 103.39303
## 785 786 787 788 789 790 791 792
## 108.11311 102.38782 112.78954 108.06946 111.91542 113.40139
109.29316 103.87368
## 793 794 795 796 797 798 799 800
## 111.82799 107.80718 110.42944 107.76341 105.14116 107.19521
103.13075 107.89461
## 801 802 803 804 805 806 807 808
## 107.93826 110.56064 112.57104 101.73220 108.15676 112.17769
110.34214 101.99435
## 809 810 811 812 813 814 815 816
## 106.97671 108.68131 109.77394 106.75821 105.44720 108.89993
111.17237 103.34925
## 817 818 819 820 821 822 823 824
## 106.27755 110.16729 103.78625 106.97683 108.41903 95.30752
103.52410 104.79158
## 825 826 827 828 829 830 831 832
## 105.84055 100.81442 99.32845 107.19521 111.34734 103.87368
113.09547 109.16208
## 833 834 835 836 837 838 839 840
## 105.31600 104.04853 110.29836 99.85287 102.12555 103.56788
104.26715
108.94346
## 841 842 843 844 845 846 847 848
## 108.98723 105.31600 104.52943 99.02240 106.53971 110.95386
101.03292 108.50646
## 849 850 851 852 853 854 855 856
```

```
## 104.09230 111.69691 100.42095 107.98203 110.99776 113.22654
104.61673 105.40343
## 857 858 859 860 861 862 863 864
## 105.22858 105.14116 106.14648 103.26183 107.32641 106.01528
110.64794 107.71976
## 865 866 867 868 869 870 871 872
## 103.56775 106.49606 105.40331 111.43464 104.44188 102.21285
101.16412 102.60620
## 873 874 875 876 877 878 879 880
## 107.19533 105.97151 102.47513 105.14116 104.22338 107.23898
110.82279 107.85096
## 881 882 883 884 885 886 ## 108.02568 109.42424 111.30356
101.86328 109.81759 101.90705
prediction2<-predict(mul_regg1,newdata = new_data)
## Warning: 'newdata' had 1 row but variables found have 886
rows prediction2
## 1 2 3 4 5 6 7 8
## 102.88940 103.22523 116.40265 95.77941 98.76300 94.40102
107.92304 118.57926
## 9 10 11 12 13 14 15 16
## 129.11769 103.56983 91.88858 118.69704 122.13425 100.70403
103.22523 106.78022
## 17 18 19 20 21 22 23 24
## 97.95601 90.85479 110.78884 126.02508 119.95765 98.98981
119.03288 102.77161
## 25 26 27 28 29 30 31 32
## 112.04944 106.43563 103.91442 88.79597 122.81468 114.90647
132.55489 116.40265
## 33 34 35 36 37 38 39 40
## 114.22605 105.63741 100.25041 97.61142 102.20020 114.79745
114.79745 89.93002
## 41 42 43 44 45 46 47 48
## 109.64602 114.91524 105.29281 103.34302 99.79679 120.75586
102.64506 113.53685
```

```
## 49 50 51 52 53 54 55 56
## 93.72059 99.67024 101.96463 94.51881 99.67900 98.87202
99.33440 83.18216
## 57 58 59 60 61 62 63 64
## 109.99062 93.14042 83.52676 93.82961 121.44506 107.00703
113.41906 104.48583
## 65 66 67 68 69 70 71 72
## 102.53604 99.90581 97.04001 107.57844 122.01646 110.78884
96.12401 107.46942
## 73 74 75 76 77 78 79 80
## 122.13425 118.35245 100.13262 97.16656 88.22457 100.13262
98.76300 99.67024
## 81 82 83 84 85 86 87 88
## 96.35082 97.04001 102.19144 132.32808 108.38542 98.64521
95.32579 97.50240
## 89 90 91 92 93 94 95 96
## 111.13344 90.96381 104.25902 106.67120 116.16708 103.56983
107.69623 97.04001
## 97 98 99 100 101 102 103 104
## 94.29200 101.73782 91.88858 102.30923 96.12401 114.45286
112.27625 119.15066
## 105 106 107 108 109 110 111 112
## 107.46942 115.83124 108.61223 109.18364 91.54399 87.18201
112.04944 102.08242
## 113 114 115 116 117 118 119 120
## 114.80622 91.42620 115.36886 107.81402 104.14123 116.05805
119.26845 105.51962
## 121 122 123 124 125 126 127 128
## 111.47803 105.86422 97.83823 104.71264 121.10046 105.29281
116.74725 110.44424
## 129 130 131 132 133 134 135
136
## 102.20020 95.20800 96.24179 125.34465 123.96626 120.52905
112.27625 95.66162
```

```
## 137 138 139 140 141 142 143 144
## 125.57146 112.16723 109.64602 98.64521 130.26927 117.89006
101.16641 109.18364
## 145 146 147 148 149 150 151 152
## 102.99842 102.08242 113.53685 106.43563 114.22605 93.72059
102.88940 99.90581
## 153 154 155 156 157 158 159 160
## 103.45204 106.55341 124.07528 103.22523 121.20948 97.96478
109.18364 96.92222
## 161 162 163 164 165 166 167 168
## 89.48517 107.69623 106.66244 107.00703 94.51881 98.75423
128.66407 105.05724
## 169 170 171 172 173 174 175 176
## 102.77161 101.85561 87.06422 119.26845 87.08175 100.59501
130.49608 114.11702
## 177 178 179 180 181 182 183 184
## 90.51019 98.41840 106.20005 96.57763 121.10046 104.25902
102.54480 98.07380
## 185 186 187 188 189 190 191 192
## 91.65301 108.95683 109.41045 91.65301 90.96381 108.04083
92.57778 122.01646
## 193 194 195 196 197 198 199 200
## 103.80540 104.60362 130.26927 107.69623 105.86422 117.43644
94.75438 105.74643
## 201 202 203 204 205 206 207 208
## 104.48583 101.73782 104.14123 97.72921 96.81320 110.55326
81.79501 98.07380
## 209 210 211 212 213 214 215 216
## 106.31784 128.89088 110.21743 117.09185 91.88858 100.35943
111.25122 96.46860
## 217 218 219 220 221 222 223 224
## 95.43481 92.57778 112.72987 106.31784 103.68761 98.30061
103.34302 106.09103
## 225 226 227 228 229 230 231 232
```

```
## 107.57844 124.76448 110.44424 119.03288 102.42701 128.54628
91.65301 100.35943
## 233 234 235 236 237 238 239 240
## 109.18364 108.95683 112.96544 103.79664 95.77941 113.42783
95.43481 96.81320
## 241 242 243 244 245 246 247 248
## 125.10907 119.49526 109.41921 100.71279 116.17584 109.06585
102.77161 112.27625
## 249 250 251 252 253 254 255 256
## 101.73782 118.69704 93.25821 105.05724 100.82181 113.08323
118.80607 92.34220
## 257 258 259 260 261 262 263 264
## 80.99679 121.33603 89.36738 126.94108 102.77161 100.47722
120.99144 104.83043
## 265 266 267 268 269 270 271 272
## 114.57064 124.19307 106.31784 94.51881 104.49460 105.17503
130.84067 100.02360
## 273 274 275 276 277 278 279 280
## 106.09980 118.23466 112.73863 128.89088 98.64521 97.14903
100.93960 94.98119
## 281 282 283 284 285 286 287 288
## 111.13344 122.82345 103.79664 99.45219 110.56203 102.77161
117.66325 101.51101
## 289 290 291 292 293 294 295 296
## 111.70484 105.86422 99.22538 92.00637 106.09103 108.04083
121.56284 108.14985
## 297 298 299 300 301 302 303 304
## 107.00703 123.96626 111.46927 105.62865 91.65301 95.55260
125.45367 107.69623
## 305 306 307 308 309 310 311 312
## 97.04878 108.95683 80.65219 83.86259 88.79597 100.02360
113.76366 107.12482
## 313 314 315 316 317 318 319 320
```

```
## 114.10826 124.76448 99.90581 96.46860 105.05724 82.14837
110.21743
112.04944
## 321 322 323 324 325 326 327 328
## 101.16641 99.90581 126.94985 106.55341 93.82961 92.34220
109.52824 109.52824
## 329 330 331 332 333 334 335 336
## 109.06585 102.30923 116.97406 99.56121 130.37829 97.84699
95.66162 90.40117
## 337 338 339 340 341 342 343 344
## 82.60199 108.95683 119.26845 104.03221 93.60280 100.02360
99.21662 112.15846
## 345 346 347 348 349 350 351 352
## 94.75438 97.84699 110.21743 105.17503 92.34220 86.50158
111.70484 82.37518
## 353 354 355 356 357 358 359 360
## 106.09103 101.04862 98.76300 98.18282 96.92222 121.90744
101.39322 90.05657
## 361 362 363 364 365 366 367 368
## 109.18364 94.29200 101.73782 98.64521 127.63028 106.09103
98.30061 90.73700
## 369 370 371 372 373 374 375 376
## 112.15846 96.46860 92.91361 94.98119 107.80525 115.83124
108.72126 106.43563
## 377 378 379 380 381 382 383 384
## 96.00622 97.14903 102.88063 111.13344 107.69623 96.00622
104.15000 107.57844
## 385 386 387 388 389 390 391 392
## 102.88063 102.99842 117.43644 79.40036 115.60443 118.80607
109.41921 122.47885
## 393 394 395 396 397 398 399 400
## 102.77161 73.89557 111.24246 110.09964 103.56983 110.56203
87.19077 96.01498
## 401 402 403 404 405 406 407 408
```

```
## 103.11621 108.61223 111.13344 107.70499 96.12401 120.98267
110.67105 94.29200
## 409 410 411 412 413 414 415 416
## 107.00703 99.56121 102.77161 92.91361 101.16641 103.34302
122.93247 103.34302
## 417 418 419 420 421 422 423 424
## 107.00703 120.18446 101.04862 112.04944 114.91524 96.12401
84.55178 118.23466
## 425 426 427 428 429 430 431 432
## 95.55260 88.44261 109.63726 92.34220 97.83823 111.82263
121.10046 113.31004
## 433 434 435 436 437 438 439 440
## 102.30923 114.57064 100.47722 104.48583 96.12401 94.63660
118.81483 97.50240
## 441 442 443 444 445 446 447 448
## 86.50158 101.39322 121.55408 102.65382 110.32645 98.30061
124.31086 79.74495
## 449 450 451 452 453 454 455 456
## 119.94888 102.42701 100.35943 98.30061 103.67885 82.37518
103.45204 100.35943
## 457 458 459 460 461 462 463 464
## 115.36886 118.57926 107.35163 105.41060 119.83986 96.12401
85.92141 118.35245
## 465 466 467 468 469 470 471 472
## 100.58624 100.25041 107.69623 137.25270 103.22523 86.84618
112.04944 99.33440
## 473 474 475 476 477 478 479 480
## 85.24098 112.27625 115.25984 101.96463 88.67819 104.37681
103.92319 121.21825
## 481 482 483 484 485 486 487 488
## 119.83986 111.25122 122.25204 115.25984 106.78022 114.45286
102.53604 108.15861
## 489 490 491 492 493 494 495 496
```

```
## 102.76285 95.31702 97.72921 100.25041 108.38542 98.52742
114.34383 99.90581
## 497 498 499 500 501 502 503 504
## 111.93165 117.31866 116.05805 105.17503 107.92304 112.27625
127.63028 108.50321
## 505 506 507 508 509 510 511
512
## 89.02278 102.42701 100.25041 111.82263 103.34302 86.03920
124.07528 114.10826
## 513 514 515 516 517 518 519 520
## 109.75505 110.44424 105.05724 92.91361 97.38461 91.54399
116.40265 98.30061
## 521 522 523 524 525 526 527 528
## 89.71198 98.52742 107.69623 95.09898 87.40882 103.79664
105.74643 98.64521
## 529 530 531 532 533 534 535 536
## 114.46162 94.98119 106.09103 93.60280 101.62003 77.45056
100.35943 126.02508
## 537 538 539 540 541 542 543 544
## 117.20963 109.75505 109.75505 101.28420 121.44506 105.62865
107.57844 100.71279
## 545 546 547 548 549 550 551 552
## 112.40280 91.19939 99.67024 96.24179 88.79597 115.25984
99.67900 103.56983
## 553 554 555 556 557 558 559 560
## 107.57844 106.89801 96.57763 98.87202 111.36025 104.60362
110.32645 104.48583
## 561 562 563 564 565 566 567 568
## 108.38542 115.59567 107.35163 104.37681 100.59501 125.00005
118.92385 107.00703
## 569 570 571 572 573 574 575 576
## 111.02441 113.31004 93.03140 115.83124 103.45204 99.45219
114.22605 107.92304
## 577 578 579 580 581 582 583 584
```

```
## 106.09103 77.90418 120.07543 113.42783 96.92222 116.17584
102.42701 81.91279
## 585 586 587 588 589 590 591 592
## 68.62635 110.90663 103.68761 96.24179 124.88226 104.03221
95.31702 126.94108
## 593 594 595 596 597 598 599 600
## 100.25041 87.41758 108.14985 107.23384 114.79745 99.90581
101.96463 107.12482
## 601 602 603 604 605 606 607 608
## 106.55341 111.36025 127.97487 121.10046 105.74643 91.65301
113.53685 104.25902
## 609 610 611 612 613 614 615 616
## 96.81320 125.57146 93.26697 121.67187 102.30923 95.55260
119.26845 110.21743
## 617 618 619 620 621 622 623 624
## 101.16641 102.65382 120.75586 115.48665 113.08323 114.91524
94.29200 103.22523
## 625 626 627 628 629 630 631 632
## 101.40199 115.83124 91.19939 108.14985 111.93165 114.45286
115.83124 106.54465
## 633 634 635 636 637 638 639 640
## 96.00622 97.26682 108.14985 116.62946 97.72921 97.84699
102.42701 103.22523
## 641 642 643 644 645 646 647 648
## 118.92385 113.99924 76.76137 105.05724 104.83919 89.36738
86.15698 106.78022
## 649 650 651 652 653 654 655 656
## 102.08242 108.04083 104.72141 112.51182 92.68680 104.71264
113.53685 108.38542
## 657 658 659 660 661 662 663 664
## 106.66244 113.42783 115.36886 83.75357 91.99761 120.06667
97.49363 105.17503
## 665 666 667 668 669 670 671 672
```

```
## 98.87202 132.21030 96.35082 94.63660 93.25821 105.74643
102.42701 100.36820
## 673 674 675 676 677 678 679 680
## 109.52824 99.90581 98.07380 102.30923 105.40184 105.51962
110.33522 115.14205
## 681 682 683 684 685 686 687 688
## 103.34302 108.83904 96.46860 123.73945 115.14205 101.16641
112.84766 100.13262
## 689 690 691 692 693 694 695 696
## 91.30841 99.10759 106.89801 106.20882 90.85479 104.03221
108.95683
113.31004
## 697 698 699 700 701 702 703 704
## 115.36886 102.53604 102.88940 115.14205 90.51019 101.51101
94.98119 113.31881
## 705 706 707 708 709 710 711 712
## 99.45219 101.62003 113.31004 110.21743 96.92222 106.43563
112.51182 109.87283
## 713 714 715 716 717 718 719 720
## 98.30938 95.55260 110.09964 97.49363 107.12482 114.10826
92.34220 116.52044
## 721 722 723 724 725 726 727 728
## 124.42864 109.18364 90.97258 109.30143 98.76300 107.57844
115.71346 107.69623
## 729 730 731 732 733 734 735 736
## 124.31086 111.13344 102.08242 105.86422 114.10826 96.58639
101.84684 114.79745
## 737 738 739 740 741 742 743 744
## 115.94027 107.57844 95.43481 107.00703 92.34220 98.98981
99.33440 108.37666
## 745 746 747 748 749 750 751 752
## 118.92385 99.90581 91.77080 107.46065 105.40184 89.36738
109.30143 109.52824
## 753 754 755 756 757 758 759 760
```

```
## 116.85627 108.26764 105.17503 90.28338 120.75586 89.36738
118.00785 117.31866
## 761 762 763 764 765 766 767 768
## 117.89006 105.29281 99.90581 108.61223 101.16641 119.49526
104.49460 119.15066
## 769 770 771 772 773 774 775 776
## 94.74562 105.28405 124.41988 116.74725 102.42701 111.13344
123.62166 125.10907
## 777 778 779 780 781 782 783 784
## 117.31866 106.43563 117.31866 121.10046 115.03303 110.21743
106.32661 96.58639
## 785 786 787 788 789 790 791 792
## 111.13344 93.49378 125.79827 111.25122 122.93247 127.63028
114.91524 97.61142
## 793 794 795 796 797 798 799 800
## 122.58787 110.21743 118.23466 109.75505 101.73782 107.80525
95.55260 110.56203
## 801 802 803 804 805 806 807 808
## 110.44424 119.04164 125.22686 91.19939 111.01565 123.96626
118.47023 91.65301
## 809 810 811 812 813 814 815 816
## 107.23384 113.08323 116.52044 106.66244 103.23400 114.23481
120.29348 96.12401
## 817 818 819 820 821 822 823 824
## 105.63741 117.78104 97.26682 107.81402 112.04944 70.80296
96.81320 100.93960
## 825 826 827 828 829 830 831 832
## 104.49460 88.45138 83.75357 107.80525 121.56284 97.61142
126.71427 114.68843
## 833 834 835 836 837 838 839 840
## 102.42701 98.30061 118.00785 85.24098 92.45999 97.27559
99.45219 113.53685
## 841 842 843 844 845 846 847 848
```

```
## 113.99924 102.42701 100.48598 82.25739 106.09103 119.72207
89.02278 112.39404
## 849 850 851 852 853 854 855 856
## 98.76300 122.36106 86.61060 110.90663 120.76463 126.94108
100.25041 102.77161
## 857 858 859 860 861 862 863 864
## 102.08242 101.73782 105.41060 95.77941 108.61223 104.60362
118.80607 109.87283
## 865 866 867 868 869 870 871 872
## 96.69541 106.20882 102.19144 121.32727 99.56121 92.22441
89.82976 93.48502
## 873 874 875 876 877 878 879 880
## 108.38542 104.14123 93.25821 101.73782 98.98981 108.26764
119.49526 110.67982
## 881 882 883 884 885 886 ## 110.78884 115.14205 121.10046
91.42620 116.40265 91.88858
```

Interpretation:

Model Fit (mu1_regg):

- **Residual Standard Error (RSE):** 8.107, indicating the typical deviation of observed JSPE scores from predicted values.
- **Multiple R-squared:** 0.1501, suggesting that approximately 15% of the variability in JSPE is explained by the model.
- **Adjusted R-squared:** 0.1482, adjusted for the number of predictors.
- **F-statistic:** 78 (p-value < 2.2e-16), indicating the overall model is statistically significant.

Model without intercept(mu1_regg1):

- **Residual Standard Error (RSE):** 10.92, higher than the model with intercept, indicating less precise predictions.
- **Multiple R-squared:** 0.9896, indicating an excellent fit (almost 99% of JSPE variability explained).
- **Adjusted R-squared**: 0.9895, slightly lower but still extremely high.
- **F-statistic:** 41,920 (p-value < 2.2e-16), showing the model is statistically significant.

- **Fit and Variance Explained: The** model without intercept (mul_regg1) has a much higher R-squared value, suggesting a better fit. However, this could be an overfit due to the removal of the intercept.
- **Residual Standard Error:** The model with intercept has a lower RSE, indicating more precise predictions.
- **Practicality:** Including an intercept (mul_regg) generally aligns better with the natural interpretation of regression models, as it accounts for a baseline JSPE score when the predictors are zero.

CHAPTER-5

Summary And Conclusions

5.1 Data Collection Methodology

The mental health dataset analysed here is a detailed collection designed to study the psychological well-being of individuals. It includes demographic information, mental health scores, and other relevant psychological and behavioural indicators. The dataset is particularly valuable for exploring mental health patterns, identifying risk factors, and understanding the impact of socio-demographic variables on mental well-being.

5.1.1 Data Compilation:

The dataset was compiled from structured surveys and psychological assessments. These sources include self-reported responses and standardized mental health questionnaires, providing a comprehensive overview of participant demographics, mental health status, and associated behaviours.

5.1.2 Data Cleaning and Preprocessing:

Following collection, the data underwent cleaning and preprocessing to ensure its quality and usability. This process involved handling missing or incomplete responses, standardizing numerical scores, and encoding categorical variables such as gender and employment status. These steps were necessary to prepare the dataset for analysis and ensure reliable results.

5.1.3 Feature Engineering:

To enhance the dataset's analytical value, new features were engineered from the existing variables. For example, composite scores like emotional regulation (erec_mean) and empathy components (qcae_cog, qcae_aff) were derived. These variables offer deeper insights into cognitive and affective dimensions of mental health, improving the ability to identify patterns and correlations.

5.1.4 Dataset Description:

The final dataset includes variables such as age, gender, mental health indicators (e.g., CESD, STAI scores), and measures of occupational burnout (MBI scores). These features are critical for exploring psychological trends and understanding factors that influence mental well-being, with mental health scores serving as the primary focus for predictive analysis.

5.1.5 Ethical Considerations:

Ethical considerations are paramount when working with mental health data. Ensuring participant anonymity and confidentiality is crucial to protect sensitive information. Additionally, researchers must obtain informed consent from participants and ensure the data is used exclusively for academic and ethical purposes, minimizing any potential harm to individuals involved.

5.1.6 Data Availability:

The dataset is a valuable resource for researchers and educators working in psychology and mental health. Its availability facilitates research on psychological resilience, emotional well-being, and mental health interventions, making it a key asset for studies aimed at improving public health outcomes.

5.2 Data preparation mechanism

5.2.1 Data Preparation Mechanism for Mental Health Dataset (Collected from Kaggle)

To ensure the mental health dataset from Kaggle is ready for analysis, it is critical to follow a systematic data preparation process. This includes addressing data quality, managing missing values, engineering new features, and ensuring ethical handling of sensitive data. Below are the detailed steps for preparing the dataset:

Step 1: Data Understanding

- **Dataset Overview:** Begin by understanding the structure of the dataset by reviewing the metadata, column names, data types, and basic statistical summaries.
- **Contextual Relevance:** Understand the purpose of the dataset and ensure it aligns with the study's objectives (e.g., identifying patterns in mental health indicators).
- **Exploratory Analysis:** Perform initial visualization and analysis (e.g., histograms, box plots, correlation matrices) to identify data trends, distributions, and potential outliers.

Step 2: Data Cleaning

- **Handling Missing Values:**
  - For numerical variables, apply techniques like mean, median, or mode imputation or use predictive imputation models (e.g., k-Nearest Neighbors).
  - For categorical variables, impute missing values with the most frequent category or a placeholder like "Unknown."
- **Outlier Detection and Removal:** Use statistical methods like Z-scores or IQR to detect and treat outliers that may skew analysis.

- **Standardization of Values:** Ensure consistent units (e.g., for age or scores) and convert free-text responses into structured data wherever applicable.
- **Error Correction:** Identify and correct typographical or logical errors, such as negative ages or nonsensical values in mental health scores.

Step 3: Data Transformation

- **Encoding Categorical Variables:** Convert categorical variables (e.g., gender, employment status) into numerical representations using one-hot encoding or label encoding.
- **Scaling and Normalization:** Apply scaling techniques (e.g., Min-Max scaling or StandardScaler) to normalize numerical data such as age, mental health scores (CESD, STAI), and burnout measures.
- **Date and Time Formatting:** Standardize date and time columns if present, enabling accurate time-series analysis or trend identification.

Step 4: Feature Engineering

- **Creating Derived Features:**
  - Combine existing variables to create meaningful insights, such as calculating "family size" or "stress-to-empathy ratio."
  - Generate interaction features like "age × burnout score" to capture relationships between variables.
  - **Binning Data:** Group continuous variables (e.g., age) into bins or categories (e.g., age groups) for easier analysis.

Step 5: Data Validation

- **Integrity Checks:**
  - Ensure all data entries are within expected ranges (e.g., scores between 0-100).
  - Cross-check that the number of rows and columns matches the dataset description provided on Kaggle.
- **Split Dataset:** Split the dataset into training, validation, and test sets (e.g., 70:15:15 ratio) for machine learning purposes, ensuring a representative distribution in each subset.

Step 6: Documentation and Ethical Handling

- **Metadata Creation:** Document the transformations, assumptions, and cleaning techniques applied to the dataset for transparency.
- **Ethical Handling:**

- Anonymize sensitive participant information and ensure the data complies with privacy regulations (e.g., GDPR).
- Use the dataset solely for research and academic purposes as per Kaggle's licensing agreements.

This mechanism ensures the Kaggle dataset is cleaned, structured, and transformed into a format suitable for robust analysis, while maintaining the integrity and ethical considerations of working with mental health data.

5.2.2 Data Preparation for Mental Health Dataset

To prepare the mental health dataset for analysis, a structured preprocessing workflow was implemented. This ensured the dataset was clean, consistent, and ready for meaningful insights. Below are the steps taken during the preparation process:

i. Data Preprocessing

The raw dataset was first cleaned to address issues such as missing values, outliers, and inconsistencies.

- Missing values were handled using imputation techniques like mean, median, or mode imputation, depending on the variable type.
- Outliers were identified and treated using statistical methods like Z-scores and interquartile range (IQR).
- The dataset was validated to ensure consistency, remove duplicates, and correct errors.

ii. Variable Selection

Relevant predictors were identified to improve the efficiency of the analysis.

- Techniques such as feature importance ranking, correlation analysis, and domain knowledge were used to exclude redundant or irrelevant variables.
- This selection process ensured that only the most significant variables were retained for analysis.

iii. Data Transformation

To prepare the data for modeling, transformation techniques were applied.

- Numerical variables were standardized or normalized to ensure uniformity in their scales and distributions.
- Categorical variables were encoded using one-hot encoding or label encoding, depending on their suitability for analysis.

iv. Feature Engineering

Additional features were created or modified to enhance the dataset’s analytical value.

- Interaction terms were developed by combining variables to reveal complex relationships (e.g., stress-to-empathy ratio).
- Composite variables were derived, such as "burnout-to-resilience ratio," to better represent underlying patterns in the data.

v. Data Splitting

To facilitate robust model development and evaluation, the dataset was divided into three subsets:

- The **training set** was used to train the models.
- The **validation set** was used for hyperparameter tuning and model selection.
- The **testing set** was reserved for final model evaluation, ensuring unbiased performance assessment.

By following this systematic process, the raw mental health dataset was transformed into a high-quality, structured format. This provided a solid foundation for deriving meaningful insights and making informed decisions.

5.3 Data Processing Procedure

The mental health dataset underwent a systematic data processing procedure to ensure it was clean, structured, and ready for analysis. The process involved several critical steps, which are outlined below:

I Data Cleaning

- **Handling Missing Values:** Missing data was addressed using imputation techniques such as mean, median, or mode imputation, depending on the nature of the variable.
- **Outlier Detection:** Statistical methods like interquartile range (IQR) and Z-scores were used to identify and address outliers.
- **Error Correction:** Inconsistent or incorrect entries were rectified, and duplicates were removed to ensure the dataset's integrity.

II Variable Selection

- **Relevance Identification:** Features with the highest predictive importance were selected using techniques like correlation analysis and domain knowledge.
- **Exclusion of Irrelevant Variables:** Redundant or less relevant variables were removed to improve the efficiency and accuracy of the analysis.

III Data Transformation

- **Standardization and Normalization:** Numerical variables were standardized or normalized to maintain uniformity in their scales.
- **Encoding Categorical Variables:** Categorical data was transformed into numerical formats using encoding techniques like one-hot or label encoding.

IV Feature Engineering

- **Creating New Features:** Interaction terms and derived metrics, such as the "stress-to-empathy ratio," were created to capture complex relationships within the data.
- **Enhancing Insights:** Composite variables were generated to represent underlying phenomena, improving the dataset's predictive capabilities.

V Data Splitting

- The dataset was split into **training**, **validation**, and **testing** sets in a 70:15:15 ratio to support model training, tuning, and evaluation. This ensured unbiased and robust analysis.

By following this data processing procedure, the raw dataset was transformed into a well-structured and high-quality format, ready for analysis and meaningful interpretation.

5.4 Data Exploration and Findings in chapter 2

5.4.1 Data Exploration

Data exploration serves as the initial step in understanding the dataset, helping to uncover critical insights and laying the groundwork for further analysis. This phase involves several key activities:

I. **Understanding the Dataset**

During data exploration, we analyzed the dataset's structure, variables, and relationships. The primary focus was on identifying missing values, outliers, and inconsistencies, ensuring that all data points were consistent and interpretable. The dataset revealed a notable dominance of French as the most common mother tongue, with over 600 participants reporting it. This uneven linguistic distribution suggested a potential bias toward French speakers in the sample.

II. **Detecting Preliminary Patterns**

Basic statistical analyses were conducted to highlight trends, distributions, and anomalies. The analysis also revealed a stark contrast between the representation of French and other languages. While French was overwhelmingly predominant, languages like German, English, Arabic, and others had significantly fewer participants, often below 50. Furthermore, languages such as Lithuanian, Japanese, and Turkish showed minimal representation, indicating a lack of linguistic

diversity within the dataset. These findings prompted us to consider how the lack of diversity might influence the overall interpretation of the results.

III. **Interpreting the Data's Narrative**

As we delved deeper into the data, we explored the underlying stories embedded within it. For instance, the dominance of French could skew any analysis or conclusions drawn about language-related trends. Moreover, the minimal representation of many languages made it challenging to extract comprehensive multilingual insights. The exploration also highlighted the importance of presenting data visually, with color-coded language groups used to distinguish categories. However, the languages with lower participant counts were less visually prominent, which could lead to underrepresentation in graphical outputs.

5.4.2 Importance of Data Exploration

I. **Laying the Groundwork for Advanced Analysis:**

Data exploration helps to identify the key patterns and trends that will inform more sophisticated analysis and model development. The findings, particularly the dominance of French, set the stage for adjusting any models that may be sensitive to language imbalances.

II. **Ensuring Data Quality:**

By identifying data anomalies early, such as the underrepresentation of several languages, we can take steps to address these issues. This ensures that the data used in the subsequent stages of analysis is reliable and valid, preventing potential biases in the results.

5.4.3 Key Findings From Data Exploration

After exploring the shared mental health dataset:

I. Dataset Characteristics:

It was noted that French was the overwhelmingly dominant language, while other languages such as German, Arabic, and English were underrepresented, with counts below 50.

II. Patterns and Trends:

Preliminary analysis identified that languages with smaller representation, like Lithuanian and Japanese, had minimal influence on the data's overall narrative.

III. Anomalies:

Some data points appeared as outliers, particularly in the languages with fewer participants, which could affect the distribution and trends. The color-coded language groups helped visualize the linguistic diversity but made less-represented languages less visible.

5.5 Data Exploration and Findings in chapter 3

5.5.1 Predictions and Findings

The data analysis reveals significant insights into mental health patterns among individuals across various demographic and behavioral factors.

- Key predictors, such as age, study hours, and psychological scores, showcase a strong correlation with mental health indices like CESD (depression scale) and STAI (anxiety levels).
- Higher study hours are often linked to moderate stress levels but also improved resilience as reflected in MBI scores.
- Gender-based differences highlight distinct trends in emotional regulation, with females exhibiting slightly higher CESD scores on average.

These findings emphasize the importance of tailored interventions to address diverse mental health needs effectively

5.6Data Exploration and Findings in chapter 4

- The regression model with an intercept (mul_regg) demonstrates statistical significance, explaining approximately 15% of the variability in JSPE. This indicates that while the model provides meaningful insights, a substantial portion of the variation in JSPE is likely influenced by factors not included in the model.
- Conversely, the model without an intercept (mul_regg1) accounts for nearly 99% of the variability in JSPE. However, this approach may lack realism, as it assumes no baseline JSPE score when all predictors are zero—a scenario that may not align with practical contexts.
- The selection between these models hinges on theoretical considerations. Including the intercept, as in mul_regg, is typically favored for its interpretability and ability to generalize findings to broader contexts, offering a more robust framework for practical applications.

5.7Overall Interpretation:

This project delves into a comprehensive exploration of mental health patterns across various demographic, linguistic, and behavioral dimensions. The findings emphasize the complexity and diversity of factors influencing mental well-being, offering insights that can shape targeted interventions and policy decisions.

I. **Identification of Predictors**: Key predictors such as age, study hours, and psychological scores (CESD, STAI, and MBI) were found to strongly correlate with mental health outcomes.

Gender differences also played a role, with females exhibiting higher CESD scores, indicating greater vulnerability to depression.

II. **Relationships Between Stress, Personal Attributes, and Mental Health**: The analysis highlighted a nuanced relationship between academic stress and mental health. While higher study hours contributed to moderate stress levels, they also improved resilience as reflected in MBI scores. These findings underscore the dual impact of academic workload on students' mental health.
III. **Impact of Psychological Scores on Well-Being**: Psychological metrics such as empathy, anxiety, and depression were strongly associated with students' overall well-being. Variability in scores like JSPE (empathy) was significantly explained by the regression models, offering insights into how personal psychological attributes shape mental health outcomes.
IV. **Recommendations for Mental Health Support**: The findings emphasize the need for tailored interventions in educational institutions. Strategies should include:
    - Providing structured mental health programs focusing on stress management and emotional resilience.
    - Offering personalized support systems based on gender-specific needs and psychological profiles.
    - Encouraging balance between academic responsibilities and personal well-being to foster a supportive learning environment.

This study reinforces the importance of a holistic approach to mental health in educational institutions, ensuring that support systems are designed to address the diverse needs of medical students effectively. By bridging gaps in mental health awareness and interventions, institutions can cultivate a healthier and more supportive academic environment.

Bibliography:

Books:

- The Art of R Programming : A Tour of Statistical Software Design by Norman Matlof
- Applied Predictive Modeling by Max Kuhn And Kjell Johnson
- R for Data Science: Import, Tidy, Transform, Visualize, and Model Data by Radley Wickham and Garrett Grolemund
- Fundamentals of Mathematical statistics by S.C.Gupta & V.K.KAPOOR, 12th edition

Websites:

- Data source - https://www.kaggle.com/datasets/thedevastator/medical-student-mental-health/data
- Orginal Data Source – https://zenodo.org/records/5702895#.Y8OraNJBwUE
- RDocumentation - https://www.rdocumentation.org/