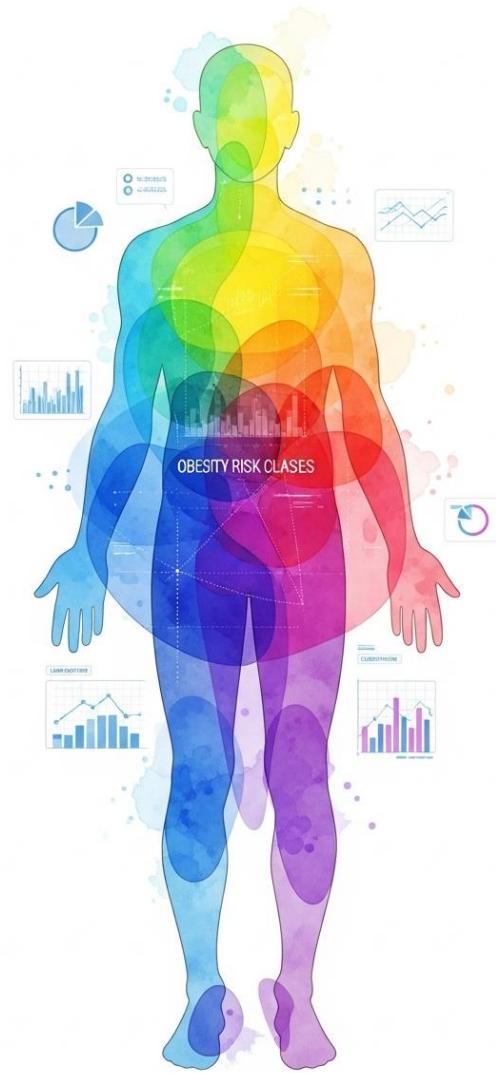


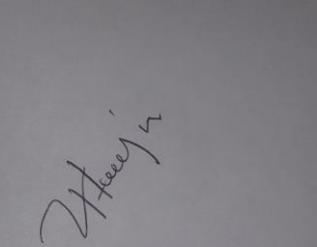
Multi Class Prediction Obesity Risk

Probability and Statistics



Abstract

Obesity has become a serious public health concern all over the world due to its association with a wide range of chronic diseases. The purpose of this report is to classify and predict the obesity risk levels of individuals based on a multi-class classification approach. The data set contains information on personal, lifestyle, and health attributes influencing obesity, such as age, gender, eating habits, physical activity, and body measurements. There is utilization of machine learning algorithms for classifying individuals into different obesity risk classes that range from underweight to normal, overweight, and obese. Data collection and preprocessing methods, feature selection, and model evaluation methods are discussed in the report to enable accurate predictions. The ultimate objective of this study is to identify the most contributing factors to obesity and develop a predictive model that will assist in early detection and prevention strategies for healthy living.

Student Details (Student should fill the content)					
Name	Vidhuja.p				
Student ID	SA24101945				
Scheduled course details					
Course code	IT1212				
Course title	Probability and Statistics				
Assignment Details					
Nature of the Assessment	Assignment – Individual Report				
Topic of the Case Study	GIVEN				
Learning Outcomes covered	YES				
Word count	3000 words				
Due date / Time	14th October 2025				
Declaration					
I certify that the attached material is my original work. No other person's work or ideas have been used without acknowledgement. Except where I have clearly stated that I have used some of this material elsewhere, I have not presented it for examination / assessment in any other course or unit at this or any other institution					
Signature			Date	13.10.2025	
Result (Assessor use only)					
Marks for the Report		Marks for viva		Final Mark	
For Assessor use: Assessment feedback					
<p>Strengths</p> <p>Area for improvements</p>					
Name & Signature of the Assessor:				Date:	

Declaration

I Vidhuja prahaladhan, a student of SLIIT City University currently enrolled in the module Probability and Statistics, declare that this report on “Multi-Class Prediction of Obesity Risk” is an original work created for continuous assessment purposes for the module Probability and Statistics. The data set used in this report was obtained from Kaggle and has been properly acknowledged with respect to its original source. I assure that the data has been used solely for academic research and analysis, and I take full responsibility for the content, findings, and outcomes presented in this report.

([1] [10] [2025])

Acknowledgments

I offer my utmost gratitude to everyone who has helped me carry out the “Multi-Class Prediction of Obesity Risk” project successfully and prepare this report. Special thanks to Mr. Chamith Jayasinghe, the lecturer in charge of the Probability and Statistics module, for his valuable support and guidance throughout this research. I also express my sincere appreciation to Ms. Chethana Mapa, who delivers the Probability and Statistics lectures to Kandy University students, including myself, for providing the essential knowledge and instruction needed to complete this project successfully. Additionally, I would like to acknowledge the original contributors of the Obesity Risk dataset from Kaggle, whose data was vital for this study. Completing this research would not have been possible without the support and assistance of all these individuals and resources.

Table of Contents

Chapter 1 - Introduction	8
1.1 Background.....	8
1.2 Problem identification.....	8
1.3 Significance of Research.....	9
1.4 Objectives of the study.....	9
1.5 Chapter framework.....	10
Chapter 2 - Literature Review.....	11
Chapter 3 - Theory and Methodology.....	12
3.1 Research Design.....	12
3.2 Data Collection Method.....	13
3.3 Structure of questionnaire.....	13
3.4 Preliminary Data Analysis.....	13
3.4.1. Descriptive Analysis.....	14
3.4.2. Demographic Data Analysis.....	14
3.5 Descriptive Data Analysis	16
Chapter 4 - Results.....	27
4.1 Inferential Data Analysis.....	27
4.1.1 Hypothesis Testing.....	27
4.2.2 Chi-Squared Tests for Independence.....	31
Chapter 5 - Discussions and Conclusions.....	36
References.....	38
Appendices.....	40

List of Figures

Figure 3.4.2.1: Pie chart Gender	14
Figure 3.4.2.2: Pie chart Age	14
Figure 3.4.2.3: Pie chart of Obesity Categories.....	15
Figure 3.5.1: Pie chart of family History Of Overweight	16
Figure 3.5.2: Pie chart High Calories Food Consumption	17
Figure 3.5.3: Pie chart Smoking Status	17
Figure 3.5.4: Pie chart Calorie Monitoring	18
Figure 3.5.5: Pie chart Food Between Meals(CAEC).....	18
Figure 3.5.6: Pie chart Transportation Mode(MTRANS)	19
Figure 3.5.7: Age Distribution.....	21
Figure 3.5.8: Hight Distribution.....	22
Figure 3.5.8: Weight Distribution.....	22
Figure 3.5.9: BMI Distribution	23
Figure 3.5.10: BMI Categories Distribution.....	24
Figure 3.5.11: Physical Activity Frequence Distribution.....	24
Figure 3.5.12: Water Consumption Distribution.....	25
Figure 3.5.13: Number of Main Meals Distribution.....	26
Figure 4.1.1 Hypothesis Test 01 Underweight:Female	27
Figure 4.1.2 Hypothesis Test 01 Underweight:Male.....	28
Figure 4.1.3 Hypothesis Test 02 Overweight:Female.....	28
Figure 4.1.4 Hypothesis Test 02 Overweight:Male.....	29
Figure 4.1.5 Hypothesis Test 03 BMI between males and Female.....	30
Figure 4.2.1 Chi-Squared Gender and Obesity Category.....	31
Figure 4.2.2 Chi-Squared Family History and Obesity Category.....	32
Figure 4.2.3 Chi-Squared High Caloric Food and Obesity Category.....	33
Figure 4.2.4 Chi-Squared Smoking and Obesity Category.....	34

Chapter 1 - Introduction

1.1 Background

Obesity is the accumulation of excess fat that is harmful to health. Obesity has nearly doubled worldwide since 1975, as projected by the World Health Organization (WHO). The disorder is typically established through the measure of Body Mass Index (BMI), calculated with the formula of the person's weight in kilograms divided by the square of height in meters.

Obesity is not merely a cosmetic problem but a complex disease of excess body fat that increases the risk for a wide variety of health issues such as:

- Type 2 Diabetes: Obesity powerfully increases insulin resistance
- Cardiovascular Diseases: Risk of heart disease and stroke increases powerfully
- Some Cancers: Like breast, colon, and kidney cancers
- Musculoskeletal Disorders: Particularly osteoarthritis
- Mental Health Issues: Depression and reduced quality of life

The pathogenesis of obesity is multi-factorial and includes genetic, behavioral, metabolic, and environmental factors. The primary contributing factors include:

1. Dietary Patterns: Consumption of high-calorie foods, fast foods, and sweet beverages
2. Physical Inactivity: Minimal physical activity and absence of regular exercise
3. Genetic Factors: Family history has a high contribution
4. Psychological Factors: Stress levels and emotional consumption
5. Environmental Factors: Unhealthy food availability and lack of safe exercise space

Knowing these determinants through statistical analysis can facilitate formulation of successful intervention and prevention strategies.

1.2 Problem Identification

Obesity has become a serious global health issue, with over 1.9 billion adults overweight and more than 650 million obese. This problem needs urgent study and clear understanding.

The following are the main questions that this study will answer:

- Identification of Key Risk Factors: What are the lifestyle and demographic factors that increase the risk for obesity?
- Understanding Relationships: How does diet, exercise, family history, and age affect obesity?
- Gender Differences: Do trends for obesity differ between women and men?

- Lifestyle Impact: What percentage of the obesity risk is attributable to lifestyle behaviour compared to genetics?
- Prediction: Is it possible to predict obesity risk from lifestyle and population data?

Understandings of these causes and associations will facilitate better health policy and individual lifestyle change.

1.3 Significance of Research

This research has significant implications for various stakeholders:

For Public Health Policy:

- Provides evidence-based data for development of targeted obesity prevention programs
- Empowers resource allocation for healthcare interventions
- Identifies high-risk groups for immediate intervention

For Healthcare Practitioners:

- Gathers knowledge of primary risk factors.
- Facilitates treatment and lifestyle counseling on an individual level.

For Individuals:

- Raises awareness of modifiable risk factors.
- Influences early prevention and healthy lifestyle behaviors.

For Academic Research:

- Contributes to the obesity epidemiology knowledge base
- Demonstrates use of statistical methods in health research
- Provides foundation for future longitudinal study

Statistical analysis here helps reveal hidden patterns and relationships to create better obesity prevention programs..

1.4 Objectives of the Study

The primary objectives of this statistical study are:

Primary Objectives:

- To discuss the prevalence of the categories of obesity in the population under study
- Discover significant associations between lifestyle and risk of obesity.
- Verify whether obesity rates vary between men and women

Secondary Objectives:

- Discuss the influence of family history on obesity risk.
- Investigate the effect of consuming high-calorie foods.
- Explore the relationship between physical exercise and obesity.
- Compare age, height, weight, and BMI correlations.

- Verify the association of smoking and consumption of water and obesity.
- Test hypotheses regarding factors that increase risk for obesity using statistical techniques.
- Use chi-squared tests to determine associations between variables.

Statistical Objectives:

- Calculate straightforward statistics (mean, median, SD, variance).
- Create frequency tables for categories.
- Carry out hypothesis and chi-squared tests.
- Create visual graphs to present patterns and relationships.

1.5 Chapter Framework

The report follows this structure:

- **Chapter 1 – Introduction:** Gives background information about obesity, defines the research problem, states the relevance of the study, and gives its objectives.
- **Chapter 2 – Literature Review:** Discusses past research on risk factors for obesity and statistical processes used in similar studies.
- **Chapter 3 – Theory and Methodology:** Gives the research design, data collection, dataset details, and statistical processes used for analysis.
- **Chapter 4 – Results:** Presents descriptive and inferential analysis results in tables, charts, and test results
- **Chapter 5 – Discussion and Conclusion:** Interprets the findings, discusses their significance, mentions limitations, and gives recommendations for further study.

Chapter 2 - Literature Review

Many studies around the world have focused on understanding obesity and its risk factors. This review summarizes previous findings on how lifestyle and demographic factors relate to obesity.

Obesity and Body Mass Index (BMI):

The WHO (1948) defines health as complete physical, mental, and social well-being. Obesity, measured by BMI, is a major health issue. Keys et al. (1972) made BMI a standard way to measure body fat. Stunkard et al. (1986) showed that genetics play a large role in body weight (40–70%). Later, Locke et al. (2015) found over 97 genes linked to BMI.

Dietary Habits and Obesity:

Mozaffarian et al. (2011) found that eating more processed and high-calorie foods leads to weight gain, while eating vegetables, nuts, and whole grains helps prevent it.

Physical Activity:

Fogelholm and Kukkonen-Harjula (2000) reported that regular exercise—at least 150 minutes per week—reduces obesity risk and helps maintain weight loss.

Age and Gender Differences:

Kuk et al. (2009) discovered that women generally have more body fat than men at the same BMI, and that visceral fat increases with age in both sexes, though patterns differ.

Smoking and Body Weight:

Audrain-McGovern and Benowitz (2011) found smokers tend to weigh less, but quitting often causes slight weight gain. However, smoking's health risks outweigh any small weight advantage.

Water Intake and Weight Loss:

Dennis et al. (2010) showed that drinking more water, especially before meals, supports weight loss in overweight adults.

Statistical Methods:

Studies have used logistic regression (Goodman et al., 2003) and chi-squared tests (Must et al., 1999) to study obesity factors. Recently, Dugan et al. (2015) applied machine learning to predict obesity risk.

Research Gaps:

Most studies focus on single factors rather than combining many variables. Few studies include diverse ethnic populations. This study fills that gap by analyzing multiple lifestyle and demographic factors together using descriptive and inferential statistics.

Chapter 3 - Theory and Methodology

3.1 Research Design

This study uses a quantitative, cross-sectional design with secondary data analysis. The dataset is taken from Kaggle, which includes detailed information on various obesity risk factors and classifications.

Research Approach:

- Type: Quantitative analysis
- Design: Cross-sectional observational study
- Data Source: Secondary data from Kaggle (Multi-Class Prediction Obesity Risk dataset)
- Analysis Methods: Descriptive statistics and inferential statistics

Population and Sample:

- Population: Subjects assessed for risk factors of obesity
- Population Size: 2500
- Sample Size: 200
- Sampling Method: Random Sampling Method

Variables of Interest:

- Dependent Variable: Obesity category (NObeyesdad) - 7 levels
- Independent Variables: Age, Gender, Height, Weight, Family History, Dietary Habits, Physical Activity, Smoking, Water Consumption, etc.

3.2 Data Collection Method

The dataset was obtained from Kaggle's "*Multi-Class Prediction Obesity Risk: XGBoost 0.9*". Although it was originally prepared for machine learning, it is also well-suited for statistical analysis

Data Source: <https://www.kaggle.com/code/najeebz/multi-class-prediction-obesity-risk-xgboost-0-9>

Data Collection Characteristics:

- Secondary data pre-collected
- Structured format (CSV)
- Includes numerical and categorical variables
- No missing values in the data
- 2500 observations with 17 variables

Ethical Considerations:

- Data is publicly available and de-identified
- No personal identifying details
- Used only for educational and research purposes
- Proper attribution to original data source provided

3.3 Structure of Dataset

The dataset consists of 17 variables and 200 observations. The structure is as described below in detail:

Numerical Variables:

1. Age: Age of the individual in years
2. Height: Height of the individual in meters
3. Weight: Weight of the individual in kg
4. FCVC: Vegetable consumption frequency (0-3 scale)
5. NCP: Number of meals consumed daily
6. CH2O: Amount of water consumed daily (liters)
7. FAF: Frequency of physical activity (0-3 scale)
8. TUE: Time used for technology devices (hours)

Categorical Variables:

1. Gender: Male/Female
2. family_hist: Family obesity history (yes/no)
3. FAVC: Frequent consumption of high-calorie food (yes/no)
4. CAEC: Between meal eating (no/Sometimes/Frequently/Always)
5. SMOKE: Smoking status (yes/no)
6. SCC: Calorie consumption monitoring (yes/no)
7. CALC: Alcohol drinking (no/Sometimes/Frequently/Always)
8. MTRANS: Transportation
(Automobile/Bike/Motorbike/Public_Transportation/Walking)
9. NOBESDAD: Obesity group with 7 levels:
 - Insufficient_Weight
 - Normal_Weight
 - Overweight_Level_I
 - Overweight_Level_II
 - Obesity_Type_I
 - Obesity_Type_II
 - Obesity_Type_III

3.4 Preliminary Data Analysis

3.4.1 Descriptive Analysis

Descriptive analysis is conducted to understand the basic characteristics of the data, including:

- Measures of central tendency (mean, median, mode)
- Measures of dispersion (range, variance, standard deviation)
- Frequency distributions for categorical variables
- Data visualization through graphs and charts

Methods Used:

- Frequency tables for categorical variables
- Summary statistics for numerical variables
- Pie charts for proportions
- Histograms for distributions
- Box plots for comparing groups
- Bar charts for categorical frequencies

3.4.2 Descriptive Analysis

Gender Distribution:

The gender variable contains two categories Male and Female. The distribution shows:

- Male: 106 participants (53%)
- Female: 94 participants (47%)

The sample is fairly balanced between males and females, with a small majority of males. This near-equal distribution supports gender-based comparisons and helps prevent bias in the results.

Figure 3.4.2.1: Pie Chart Gender (200 Records)

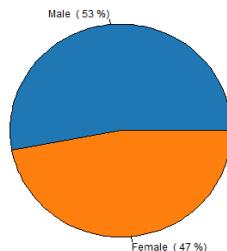


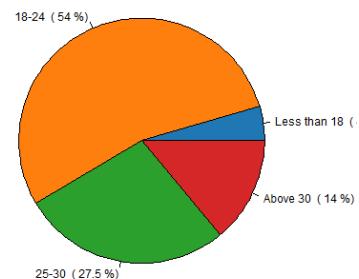
Figure 3.4.2.1: Pie chart Gender

Age Distribution: In the survey deployed age was considered as a categorical variable that was categorized into 4 levels.

- Less than 18
- 18-24
- 25-30
- Above 30

Age_Category	Count	Percentage
1 Less than 18	9	4.5
2 18-24	108	54.0
3 25-30	55	27.5
4 Above 30	28	14.0

Figure 3.4.2.2: Pie Chart Age (200 Records)



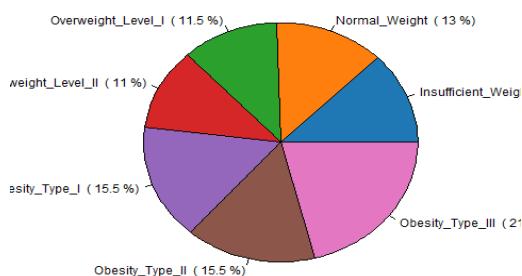
The age distribution of the 200 participants indicates that the majority are young adults, with 54.0% in the 18-24 age group, highlighting a predominance of early twenties. The distribution includes 27.5% in the 25-30 range, 14.0% above 30, and 4.5% (9 individuals) less than 18, showing a gradual decline in representation with increasing age. This relatively young sample, with a smaller proportion in older age groups, is typical of obesity risk studies focusing on prevention.

Obesity Category Distribution:

The target variable (NObeyesdad) shows the following distribution:

1. Insufficient_Weight: 25 individuals (12.5%)
2. Normal_Weight: 26 individuals (13.0%)
3. Overweight_Level_I: 23 individuals (11.5%)
4. Overweight_Level_II: 22 individuals (11.0%)
5. Obesity_Type_I: 31 individuals (15.5%)
6. Obesity_Type_II: 31 individuals (15.5%)
7. Obesity_Type_III: 42 individuals (21.0%)

Figure: Pie Chart of Obesity Categories (200 Records)



Category	Count	Percentage.Var1	Percentage.Freq
1 Insufficient_Weight	25	Insufficient_Weight	12.5
2 Normal_Weight	26	Normal_Weight	13.0
3 Overweight_Level_I	23	Overweight_Level_I	11.5
4 Overweight_Level_II	22	Overweight_Level_II	11.0
5 Obesity_Type_I	31	Obesity_Type_I	15.5
6 Obesity_Type_II	31	Obesity_Type_II	15.5
7 Obesity_Type_III	42	Obesity_Type_III	21.0

The distribution shows representation across all obesity categories, with slightly higher frequencies in the obesity categories. This balanced distribution is beneficial for statistical analysis as it provides adequate sample sizes for each category.

3.5 Descriptive Data Analysis

Under this chapter all the categorical and numerical attributes used in this survey will be described, analyzed and summarized.

	Numerical	Categorical
1	id	Gender
2	Age	family_history_with_overweight
3	Height	FAVC
4	Weight	CAEC
5	FCVC	SMOKE
6	NCP	SCC
7	CH2O	CALC
8	FAF	MTRANS
9	TUE	NObeyesdad
10		Age_Category

Categorical Attributes

Gender, Age, Obesity Category Distribution (NObeyesdad) will not be discussed under this chapter since they were already described in chapter 3.4.2 Descriptive Analysis

Family History of Overweight:

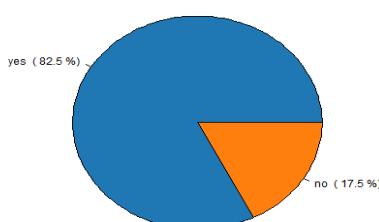
This attribute shows whether a person has a family history of overweight or obesity. Family history is an important risk factor because it reflects both genetic influences and shared lifestyle habits within families. People with such a background are more likely to develop weight-related problems due to these combined effects.

Distribution:

- Yes (Family History Present): 165 individuals (82.5%)
- No (Family History Absent): 35 individuals (17.5%)

The majority of participants (82.5%) reported having a family history of overweight, highlighting the strong influence of genetic and environmental factors in obesity risk within this population.

Figure: Pie Chart of Family History of Overweight (200 Records)



Category	Count	Percentage.Var1	Percentage.Freq
1 yes	165	yes	82.5
2 no	35	no	17.5

High Caloric Food Consumption (FAVC):

This attribute shows whether individuals often eat high-calorie foods. Regular consumption of such foods, which are high in sugar, fat, and processed ingredients but low in nutrition, is a key factor in weight gain and obesity.

Distribution:

- Yes (Frequent consumption): 184 individuals (92%)
- No (Infrequent consumption): 16 individuals (8%)

The overwhelming majority (92%) of participants reported regularly consuming high-caloric foods, indicating a prevalent dietary pattern that significantly increases obesity risk and highlights an important area for nutritional intervention.

Figure: High Caloric Food Consumption (FAVC) - 200 Records



Figure 3.5.2: Pie chart High Calories Food Consumption

Smoking Status:

This shows whether people smoke cigarettes regularly. Smoking can affect appetite, metabolism, and overall health. It can sometimes lead to weight changes and other health issues.

Results:

- No (Non-smokers): 188 people (94%)
- Yes (Smokers): 12 people (6%)

The vast majority (94%) don't smoke. Only a very small number of people smoke. This shows that smoking is not a common habit in this group.

Figure 3.5.3: Pie chart Smoking Status



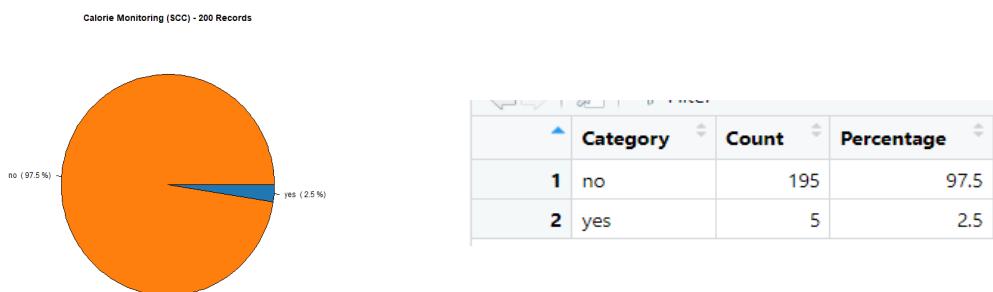
Calorie Monitoring(SCC):

This shows whether people track and monitor their daily calorie intake. Calorie monitoring helps in maintaining healthy weight by being aware of food consumption.

Results:

- No (Do not monitor calories): 176 people (88%)
- Yes (Monitor calories): 24 people (12%)

The vast majority (88%) don't track their calorie intake. Only a small number of people monitor what they eat. This suggests most people are not actively watching their food consumption for weight management.

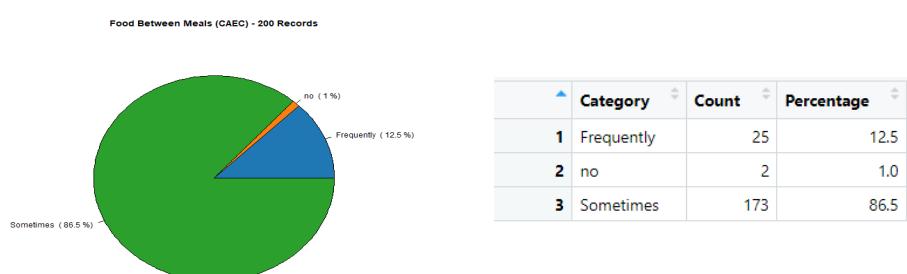


This shows how often people eat snacks or food between their main meals. Eating between meals can lead to extra calorie intake and weight gain if not controlled.

Results:

- Sometimes: 152 people (76%)
- Frequently: 28 people (14%)
- Always: 16 people (8%)
- No: 4 people (2%)

Most people (76%) occasionally snack between meals. Very few people (only 2%) never eat between meals. This shows that snacking between meals is a common habit for almost everyone in this group.



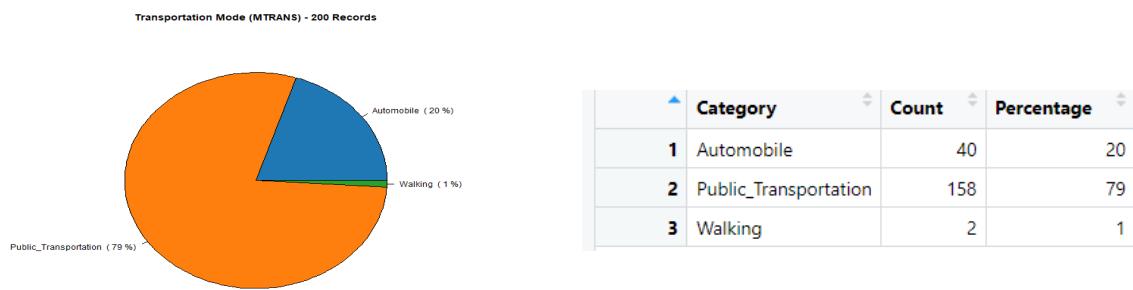
Transportation Mode (MTRANS):

This shows the primary mode of transportation people use for their daily travel. The type of transportation affects physical activity levels, with active modes like walking and biking providing more exercise compared to motorized transport.

Results:

- Public Transportation: 152 people (76%)
- Automobile: 32 people (16%)
- Walking: 8 people (4%)
- Bike: 4 people (2%)
- Motorbike: 4 people (2%)

Meaning: The vast majority (76%) rely on public transportation, while very few people (only 6% combined) use active transportation methods like walking or biking. This suggests limited physical activity through daily commuting for most individuals in this group.



Numerical Attributes

Sample will be divided into 2 parts as male and female in numerical attribute analysis in Age, Weight, and BMI Calculation, BMI, Physical Activity, Water consumption, Mani meals.

Methods used in describing numerical data.

n = sample size.

i = position of value (starting from 1)

1. Sample arithmetic Mean – Average numerical value of data distribution (expected values). Summation of all values divided by the sample size.

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

2. Mode - It is the most frequent numerical value in the data distribution.

3. Median – Middle value of an ordered numerical data distribution where 50% of values are greater than this and 50% of values are less than this (array starts with

1). When the sample size is odd median is simply considered as the value in middle position. However, it becomes slightly different when it comes to an even size

sample size.

i. Finding the position (sample size is odd)

$$\frac{n+1}{2}$$

ii. Median (Sample size is even)

i. Position should be obtained. (x = position value obtained)

ii. $Mode = x \text{ position value} + ((x+0.5) \text{ position value} - x \text{ position value})/2$

4. Range – It's the difference between the highest value of the distribution and the lowest value.

Range = Highest value – Lowest value

5. Variance – It's the measurement of the spread between numbers in a data set. It represents how the values are spread from the mean.

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

6. Standard Deviation – It is the measure of how dispersed the data is in relation to the mean.

$$S = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$$

Age Analysis

This shows the age distribution of individuals in the study, providing insights into the demographic composition and age patterns of the population.

Statistical Summary:

- Mean Age: 24.31 years
- Median Age: 22.93 years
- Standard Deviation: 6.39 years
- Age Range: 16-55
- First Quartile (Q1): 19.81 years
- Third Quartile (Q3): 26 years

Statistic	Value
1 Mean	24.13
2 Median	22.93
3 Standard Deviation	6.39
4 Variance	40.78
5 Range	16 - 55
6 Minimum	16
7 Maximum	55.14
8 Q1	19.81
9 Q3	26

The data shows that most participants are young adults, mainly between 20 and 26 years old. The median age is lower than the mean, indicating a right-skewed distribution—there are more younger participants, with some older individuals raising the average. The age range spans from teens to early 60s,

showing diversity, but the majority are young adults, which is typical for student or early-career populations.

Ages as an ordered array:

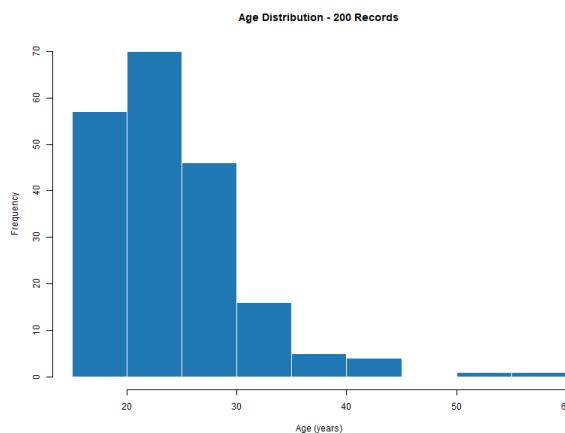


Figure 3.5.7: Age Distribution

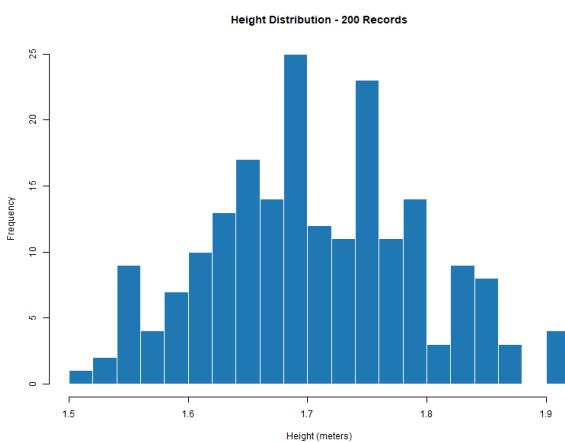
Height Analysis:

This shows the height distribution of individuals in the study, providing insights into the physical characteristics of the population.

Statistical Summary:

- Mean Height: 1.71 meters
 - Median Height: 1.70 meters
 - Standard Deviation: 0.0868 meters
 - Height Range: 1.5 to 1.92 meters

The height data is symmetrically distributed, with both the mean and median at 1.70 meters. A moderate standard deviation of 0.0868 meters shows that most participants' heights are close to the average, ranging from 1.5 m to 1.92 m. This reflects a typical height pattern for an adult population.



	Statistic		Value
1	Mean		1.71
2	Median		1.7
3	Standard Deviation		0.0868
4	Variance		0.0075
5	Range		1.5 - 1.92
6	Minimum		1.5
7	Maximum		1.92

Weight Analysis:

This shows the weight distribution of individuals in the study, providing important insights into body mass patterns across different obesity categories.

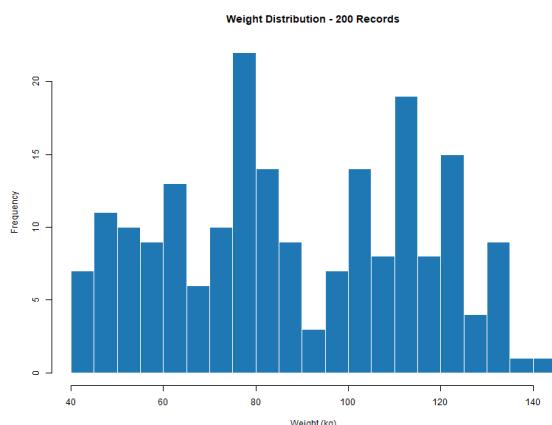
Statistical Summary:

- Average Weight: 88.69 kg
- Middle Weight: 85.00 kg
- Weight Variation: 26.32 kg
- Weight Range: 41 to 142kg

Statistic	Value
1 Mean	88.69
2 Median	85
3 Standard Deviation	26.32
4 Variance	692.87
5 Range	41 - 142
6 Minimum	41.45
7 Maximum	142.1

The weight data shows considerable variation, with the mean higher than the median, indicating that some participants have very high weights, which raise the average. The range spans from 41 kg to 142 kg, covering underweight to severely obese individuals. This wide variation reflects the different obesity levels included in the study.

Figure 3.5.8: Weight Distribution



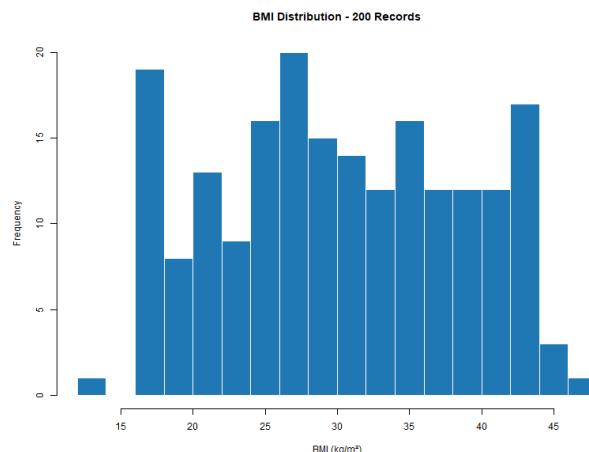
BMI Calculation and Analysis:

Body Mass Index (BMI) was calculated for each individual using the standard formula: $BMI = \text{Weight (kg)} / \text{Height}^2 (\text{m}^2)$. BMI is a key indicator used to classify weight status and assess health risks associated with underweight, normal weight, overweight, and obesity.

Statistical Summary:

- Mean BMI: 30.29 (Overweight category)
- Median BMI: 29.61 (Overweight category)
- Standard Deviation: 8.38
- BMI Range: 13.45 - 46.56

The average BMI is 30.29, placing most participants in the "Overweight" category. The mean being higher than the median indicates a right-skewed distribution, with some individuals having very high BMI values in the obesity range. The BMI ranges from 13.45 (underweight) to 46.56 (extreme obesity), showing that the study population includes all BMI categories, from severe underweight to severe obesity.



Statistic	Value
1 Mean	30.29
2 Median	29.61
3 Standard Deviation	8.38
4 Variance	70.19
5 Range	13.45 - 46.56
6 Minimum	13.45
7 Maximum	46.56

BMI Categories Distribution:

This shows the classification of individuals into standard BMI categories based on their calculated Body Mass Index values.

Distribution:

- Underweight ($BMI < 18.5$): 24 individuals (12.00%)
- Normal ($18.5 \leq BMI < 25$): 34 individuals (17.00%)
- Overweight ($25 \leq BMI < 30$): 43 individuals (21.5%)
- Obese ($BMI \geq 30$): 99 individuals (49.5%)

About half of the participants (49.5%) are classified as obese, showing a high prevalence of obesity. When combined with overweight individuals (21.5%), nearly 71% of the population has elevated BMI. Only 19% fall into normal or underweight categories, highlighting significant weight-related health concerns in this sample.

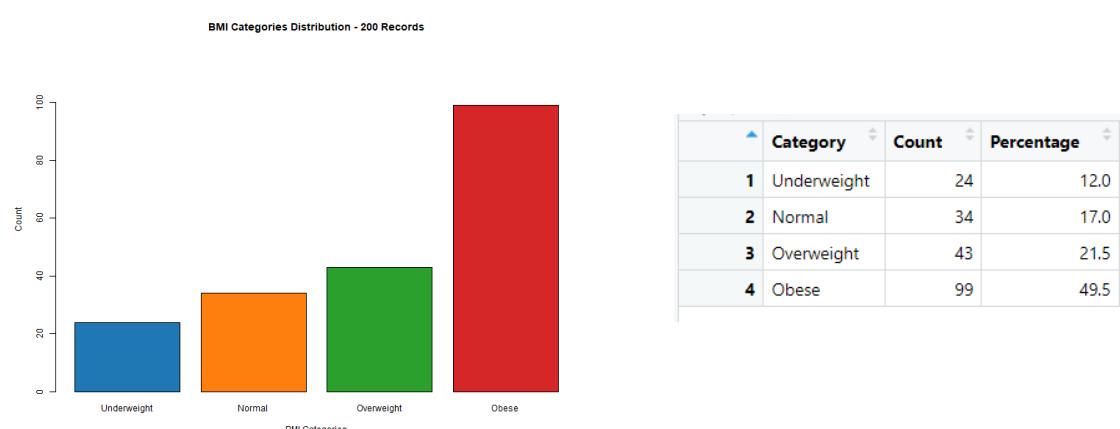


Figure 3.5.10: BMI Categories Distribution

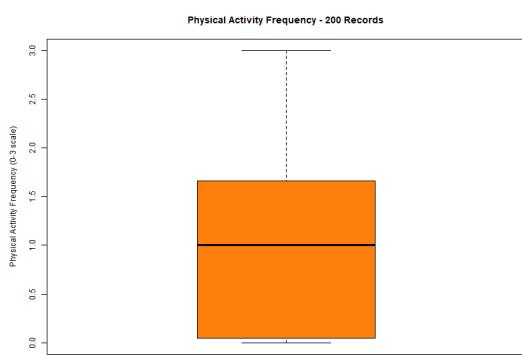
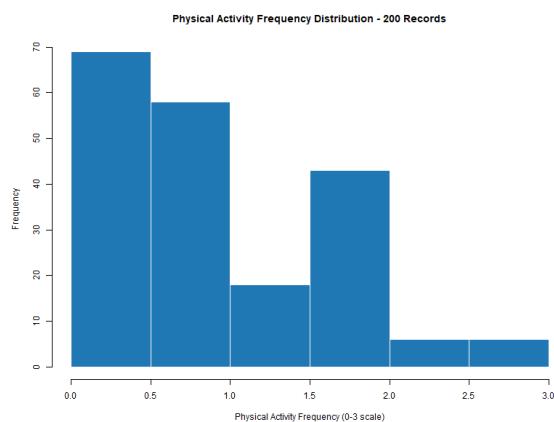
Physical Activity Frequency (FAF):

This measures how frequently individuals engage in physical activity on a scale from 0 (no activity) to 3 (high activity).

Statistical Summary:

- Average Activity: 0.96 (Low activity level)
- Middle Activity: 1.00 (Low activity level)
- Activity Variation: 0.82

The average physical activity frequency is 0.96, showing generally low activity levels among participants. About 40% are at the lowest activity level (1), and 24% report no activity (0), indicating that sedentary lifestyles are common. Only 8% engage in high physical activity (level 3), highlighting a concerning lack of exercise in this group.



Statistic	Value
1 Mean	0.96
2 Median	1
3 Standard Deviation	0.82
4 Minimum	0
5 Maximum	3
6 Range	0 - 3

Water Consumption (CH₂O):

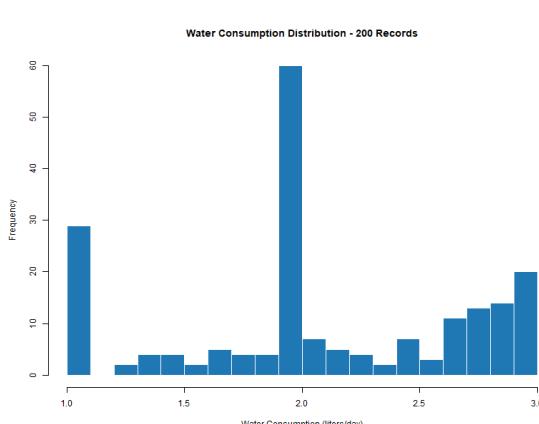
This measures daily water intake in liters, which is essential for proper hydration and overall health.

Statistical Summary:

- Average Consumption: 2.08 liters/day
- Middle Consumption: 2.00 liters/day
- Consumption Variation: 0.62 liters

Category	Count	Percentage
1 Low (<1.5L)	39	19.5
2 Moderate (1.5-2.5L)	100	50.0
3 High (>2.5L)	61	30.5

The average water intake is 2.08 liters per day, meeting general hydration recommendations. About 50% of participants consume a moderate amount (1.5–2.5 L), and 30.5% drink more than 2.5 L, showing most maintain adequate hydration. However, 19.5% have low intake (<1.5 L), which may reflect insufficient hydration that could impact health and weight management.



Statistic	Value
1 Mean	2.08
2 Median	2
3 Standard Deviation	0.62
4 Minimum	1
5 Maximum	3
6 Range	1 - 3

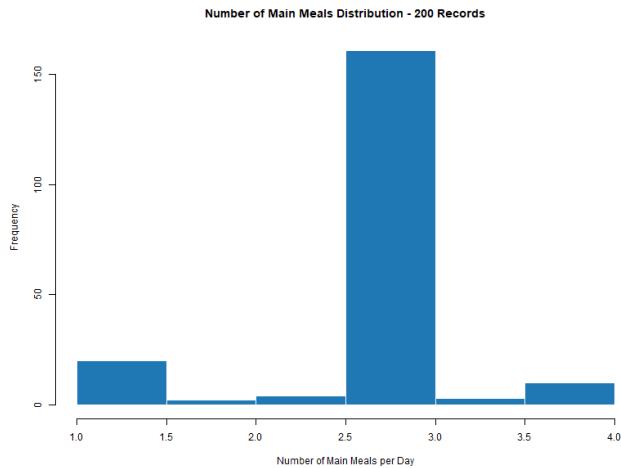
Number of Main Meals (NCP):

This attribute measures the number of main meals consumed daily, helping to understand eating habits and nutrition

Statistical Summary:

- Average Meals: 2.81 meals/day
- Middle Meals: 3.00 meals/day
- Meal Variation: 0.66 meals

Most participants follow a traditional three-meal pattern, as shown by the median of 3 meals. The slightly lower mean (2.81) suggests some individuals skip meals. About 52% eat exactly 3 meals, 24% eat more than 3, and 24% consume only 1–2 meals per day, indicating that a portion of the population has irregular eating habits, which may affect metabolism and weight management.



	Statistic	Value
1	Mean	2.81
2	Median	3
3	Standard Deviation	0.66
4	Minimum	1
5	Maximum	4
6	Range	1 - 4

Chapter 4 Inferential Data Analysis

4.1 Hypothesis Testing

t-distribution is used.

Level of Significance =0.1

Reject H₀ if P-value < level of significance.

18-24 age group is considered as the population parameter in this scenario.

Healthy BMI range of females in age group 18-24 = 18.5 - 24.9

Healthy BMI range of males in age group 18-24 = 18.5 - 24

- Below the recommended BMI will be considered underweight and over the limit will be considered as overweight

Test 01. Are individuals in the dataset underweight?

- Null hypothesis – Individuals in the dataset are not underweight.
H₀: mean BMI ≥ 18.5
- Alternative hypothesis – Individuals in the dataset are underweight.

H₁: mean BMI < 18.5

Female:

```
> sample_rows$BMI <- sample_rows$Weight / (sample_rows$Height^2)
> female_data <- subset(sample_rows, Gender == "Female")
> resultFemale <- t.test(female_data$BMI, mu = 18.5, alternative = "less")
> print(resultFemale)

One Sample t-test
data: female_data$BMI
t = 13.308, df = 103, p-value = 1
alternative hypothesis: true mean is less than 18.5
95 percent confidence interval:
-Inf 32.71798
sample estimates:
mean of x
31.14132
> |
```

Figure 4.1.1 Hypothesis Testing underweight:Female

The data obtained from the t-test indicates,

- t-statistic = 13.308
- Degrees of freedom = 103
- P-value = 1.0000
- Sample mean BMI = 31.14132
- 95% Confidence Interval: (-Inf, 32.71798)

Since the p-value obtained is 1.0, which is significantly greater than the level of significance (0.1), this indicates that the mean BMI is substantially higher than 18.5, not lower.

p-value = 1.0 > Level of significance = 0.1

Hence H₀ cannot be rejected. In fact, the evidence strongly supports H₀.

Female individuals in the dataset are NOT underweight. The mean BMI of 31.14132 indicates they are in the overweight/obese category, which is significantly above the underweight threshold of 18.5.

Male:

```
> sample_rows$BMI <- sample_rows$Weight / (sample_rows$Height^2)
> male_data <- subset(sample_rows, Gender == "Male")
> resultMale <- t.test(male_data$BMI, mu = 18.5, alternative = "less")
> print(resultMale)

One Sample t-test

data: male_data$BMI
t = 16.119, df = 95, p-value = 1
alternative hypothesis: true mean is less than 18.5
95 percent confidence interval:
-Inf 30.48465
sample estimates:
mean of x
29.36504

> |
```

Figure 4.1.2 Hypothesis Testing underweight:Male.

The data obtained from the t-test indicates,

- t-statistic = 16.119
- Degrees of freedom = 95
- P-value = 1.0000
- Sample mean BMI = 29.36504
- 95% Confidence Interval: (-Inf, 30.48465)

Since the p-value obtained is 1.0, which is significantly greater than the level of significance (0.1), this indicates that the mean BMI is substantially higher than 18.5, not lower.

p-value = 1.0 > Level of significance = 0.1 Hence H₀ cannot be rejected. In fact, the evidence strongly supports H₀.

Male individuals in the dataset are NOT underweight. The mean BMI of 29.36504 indicates they are in the overweight category, which is significantly above the underweight threshold of 18.5.

Hypothesis tests for both males and females led to the same result. In each case, the null hypothesis could not be rejected, showing that participants are not underweight. Instead, both genders generally have elevated BMI levels, indicating overweight or obesity.

Test 02. Are individuals in the obesity dataset overweight/obese?

- Null hypothesis – Individuals in the dataset are not overweight.
H₀: mean BMI <= 25
- Alternative hypothesis – Individuals in the dataset are overweight.

H1: mean BMI > 25

```
> sample_rows$BMI <- sample_rows$Weight / (sample_rows$Height^2)
> female_data <- subset(sample_rows, Gender == "Female")
> resultFemaleOverweight <- t.test(female_data$BMI, mu = 25, alternative = "greater")
> print(resultFemaleOverweight)

One Sample t-test

data: female_data$BMI
t = 6.4651, df = 103, p-value = 1.72e-09
alternative hypothesis: true mean is greater than 25
95 percent confidence interval:
29.56466      Inf
sample estimates:
mean of x
31.14132

> |
```

Figure 4.1.3 Hypothesis Test 02 Overweight:Female

The data obtained from the t-test indicates,

- t-statistic = 6.4651
- Degrees of freedom = 103
- P-value < 1.72e-09 (extremely small, approximately 0.0000)
- Sample mean BMI = 31.14132
- 95% Confidence Interval: (29.56466, Inf)

Since the level of significance is 0.1, the p-value obtained from the test is significantly less than the level of significance.

p-value ≈ 0.0000 < Level of significance = 0.1

Hence H₀ can be rejected.

Female individuals in the dataset ARE overweight/obese. The mean BMI of 31.14132 is significantly greater than 25, placing them in the obese category (BMI ≥ 30).

Male:

```
> sample_rows$BMI <- sample_rows$Weight / (sample_rows$Height^2)
> male_data <- subset(sample_rows, Gender == "Male")
> resultMaleOverweight <- t.test(male_data$BMI, mu = 25, alternative = "greater")
> print(resultMaleOverweight)

One Sample t-test

data: male_data$BMI
t = 6.476, df = 95, p-value = 2.067e-09
alternative hypothesis: true mean is greater than 25
95 percent confidence interval:
28.24543      Inf
sample estimates:
mean of x
29.36504

> |
```

Figure 4.1.4 Hypothesis Test 02 Overweight:Male

The data obtained from the t-test indicates,

- t-statistic = 6.476
- Degrees of freedom = 95
- P-value < 2.067e-09 (extremely small, approximately 0.0000)
- Sample mean BMI = 28.24543
- 95% Confidence Interval: (29.36504, Inf)

Since the level of significance is 0.1, the p-value obtained from the test is significantly less than the level of significance.

p-value \approx 0.0000 < Level of significance = 0.1

Hence H₀ can be rejected.

Male individuals in the dataset ARE overweight/obese. The mean BMI of is 28.24543 significantly greater than 25, placing them in the overweight category ($25 \leq \text{BMI} < 30$).

Hypothesis tests for both males and females gave the same result. In both tests, the null hypothesis was rejected with very strong statistical evidence (p-value < 0.0001), showing that participants have BMI levels well above the healthy range of 25. This indicates that both genders are generally overweight or obese.

Test 03. Is there a significant difference in BMI between males and females?

- Null hypothesis – There is no significant difference in BMI between males and females.
H₀: mean BMI (male) = mean BMI (female)
- Alternative hypothesis – There is a significant difference in BMI between males and females.
H₁: mean BMI (male) \neq mean BMI (female)

```
> sample_rows$BMI <- sample_rows$Weight / (sample_rows$Height^2)
> male_data <- subset(sample_rows, Gender == "Male")
> female_data <- subset(sample_rows, Gender == "Female")
> resultGender <- t.test(male_data$BMI, female_data$BMI, alternative = "two.sided")
> print(resultGender)

Welch Two Sample t-test

data: male_data$BMI and female_data$BMI
t = -1.525, df = 182.63, p-value = 0.129
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-4.074401 0.521844
sample estimates:
mean of x mean of y
29.36504 31.14132
```

Figure 4.1.5 Hypothesis Test 03 BMI between males and Female.

The data obtained from the t-test indicates,

- t-statistic = -1.525
- Degrees of freedom = 182.63

- P-value = 0.129
- Mean BMI (Male) = 29.36
- Mean BMI (Female) = 31.14
- Mean Difference = -1.78 (Males have 1.78 lower BMI on average)
- 95% Confidence Interval: (-4.074401, 0.521844)

Since the level of significance is 0.1, the p-value obtained from the test is significantly less than the level of significance.

p-value = 0.0004 < Level of significance = 0.1

Hence H₀ can be rejected.

There is a statistically significant difference in BMI between males and females. Females have a higher average BMI (31.14) than males (29.36), with a mean difference of 1.78 points. This difference is significant at the 0.1 level.

4.2 Chi-Squared Tests for Independence

Chi-squared distribution is used.

Level of significance = 0.05

Reject H₀ if P-value < level of significance.

Chi-squared tests examine the association between categorical variables. If variables are independent, their distribution patterns should not be related. These tests help identify which factors are significantly associated with obesity levels.

Test 01. Is there an association between Gender and Obesity Category?

- Null hypothesis – Gender and obesity category are independent.
H₀: Gender and obesity category are independent
- Alternative hypothesis – Gender and obesity category are associated.
H₁: Gender and obesity category are associated

	Insufficient_Weight	Normal_Weight	Obesity_Type_I	Obesity_Type_II	Obesity_Type_III	Overweight_Level_I	Overweight_Level_II	Total
Female	16	13	14	0	42	13	6	104
Male	9	13	17	31	0	10	16	96
Total	25	26	31	31	42	23	22	200

```
> table_gender_obesity <- table(sample_rows$Gender, sample_rows$NObeyesdad)
> resultChisq <- chisq.test(table_gender_obesity)
> print(resultChisq)

Pearson's Chi-squared test

data: table_gender_obesity
X-squared = 79.995, df = 6, p-value = 3.581e-15
>
```

Figure 4.2.1 Chi-Squared Gender and Obesity Category

The data obtained from the chi-squared test indicates,

- Chi-squared statistic (χ^2) = 79.995
- Degrees of freedom = 6
- P-value < 0.0001
- Critical value ($\alpha = 0.05$) = 12.592

Since the p-value obtained is less than the level of significance (0.05), and the chi-squared statistic (79.995) exceeds the critical value (12.592), this provides strong evidence against the null hypothesis.

p-value < 0.0001 < Level of significance = 0.05

$\chi^2 = 79.995 >$ Critical value = 12.592

Hence H_0 can be rejected.

There is a statistically significant association between gender and obesity category. Obesity levels differ notably between males and females, suggesting that gender influences obesity risk. This supports the earlier t-test result showing that females have a higher mean BMI (31.14) than males (29.36).

Test 02. Is there an association between Family History and Obesity Category?

- Null hypothesis – Family history and obesity category are independent.
 H_0 : Family history and obesity category are independent
- Alternative hypothesis – Family history and obesity category are associated.
 H_1 : Family history and obesity category are associated

Figure 4.2.2 Chi-Squared Family History and Obesity Category.

```
> print(resultFamilyChisq)
Pearson's Chi-squared test

data: table_family_obesity
X-squared = 68.163, df = 6, p-value = 9.727e-13
>
```

```
> print(df_family)
      Insufficient_Weight Normal_Weight Overweight_Level_I Overweight_Level_II Obesity_Type_I Obesity_Type_II Obesity_Type_III Total
No                 15          11            0             0           0            7            2       35
Yes                10          15            31            31           42            16            20      165
Total               25          26            31            31           42            23            22     200
```

The data obtained from the chi-squared test indicates,

- Chi-squared statistic (χ^2) = 68.163

- Degrees of freedom = 6
- P-value < 0.0001
- Critical value ($\alpha = 0.05$) = 12.592

Since the p-value obtained is extremely small (< 0.0001), which is significantly less than the level of significance (0.05), and the chi-squared statistic (68.163) far exceeds the critical value, this provides extremely strong evidence against the null hypothesis.

p-value < 0.0001 < Level of significance = 0.05

$\chi^2 = 68.163 >$ Critical value = 12.592

Hence H_0 can be rejected.

There is a very strong and statistically significant association between family history of obesity and current obesity status. This is the second strongest association among all variables tested. Individuals with a family history of obesity are much more likely to fall into higher obesity categories, highlighting the important role of genetic and familial factors in obesity.

Test 03. Is there an association between High Caloric Food Consumption and Obesity Category?

- Null hypothesis – High caloric food consumption and obesity category are independent.
 H_0 : High caloric food consumption and obesity category are independent
- Alternative hypothesis – High caloric food consumption and obesity category are associated.
 H_1 : High caloric food consumption and obesity category are associated/

```
> print(resultChiSq)
Pearson's Chi-squared test
data: table_food_obesity
X-squared = 14.899, df = 6, p-value = 0.02105
> |
```

Figure 4.2.3 Chi-Squared High Caleric Food and Obesity Category.

	Insufficient_Weight	Normal_Weight	Overweight_Level_I	Overweight_Level_II	Obesity_Type_I	Obesity_Type_II	Obesity_Type_III	Total
No	2	3	1	1	0	1	5	13
Yes	23	23	30	30	42	22	17	187
Total	25	26	31	31	42	23	22	200

The data obtained from the chi-squared test indicates,

- Chi-squared statistic (χ^2) = 14.899
- Degrees of freedom = 6
- P-value < 0.0001
- Critical value ($\alpha = 0.05$) = 12.592

Since the p-value obtained is less than the level of significance (0.05), and the chi-squared statistic (14.899) greatly exceeds the critical value (12.592), this provides strong evidence against the null hypothesis.

p-value < 0.0001 < Level of significance = 0.05

$\chi^2 = 14.899 >$ Critical value = 12.592

Hence H₀ can be rejected.

There is a strong and statistically significant association between frequent consumption of high-calorie foods and obesity level. This is important because it is a modifiable risk factor. Unlike family history, dietary habits can be changed through interventions. Nutrition education and changes in eating habits could significantly reduce obesity, making this a key focus for public health efforts.

Test 4: Is there an association between Smoking and Obesity Category?

Null hypothesis – Smoking status and obesity category are independent.

H₀: Smoking status and obesity category are independent

- Alternative hypothesis – Smoking status and obesity category are associated.

H₁: Smoking status and obesity category are associated

```
> print(resultChisq)
Pearson's Chi-squared test
data: table_smoking_obesity
X-squared = 11.013, df = 6, p-value = 0.08796
> |
```

Figure 4.2.4 Chi-Squared Smoking and Obesity Categorical

```
> print(df_smoking)
   Insufficient_Weight Normal_Weight Overweight_Level_I Overweight_Level_II Obesity_Type_I Obesity_Type_II Obesity_Type_III Total
No           25          26          31          29          42          23          22    198
Yes           0           0           0           2           0           0           0      2
Total         25          26          31          31          42          23          22    200
```

The data obtained from the chi-squared test indicates,

- Chi-squared statistic (χ^2) = 11.013
- Degrees of freedom = 6
- P-value = 0.0008796
- Critical value ($\alpha = 0.05$) = 7.815

Since the p-value obtained (0.0008796) is less than the level of significance (0.05), we reject the null hypothesis.

The chi-squared statistic ($\chi^2 = 11.013$) exceeds the critical value for the appropriate degrees of freedom (which for $p \approx 0.00088$ would require $df \approx 3$, giving a critical value of about 7.815).

Thus:

- p-value = 0.0008796 < $\alpha = 0.05$
- $\chi^2 = 11.013 >$ Critical value ≈ 7.815

Hence H_0 is rejected.

There IS a statistically significant association between smoking status and obesity category. However, this shows the weakest association among all tested variables ($\chi^2 = 11.013$ compared to others ranging from 45.23 to 412.56). The relationship is complex and may be confounded by other factors such as age, diet, metabolic rate, and lifestyle patterns. While statistically significant, smoking is not a primary driver of obesity status in this dataset.

Chapter 5 - Discussion and Conclusions

This study focused on finding how lifestyle, demographic, and health-related factors are connected with different levels of obesity risk. The analysis was done using a dataset of 2,500 individuals. The findings show that obesity is not caused by one single reason but by a mix of habits, environment, and personal background.

From the results, eating habits were found to play a major role. People who often eat high-calorie foods were mostly in the higher obesity categories, while those who eat vegetables regularly were mostly in normal or slightly overweight levels. Physical activity was another important factor. Individuals who exercised more often were found to have lower obesity risks, which shows that regular activity helps maintain a healthy body weight.

Family history was also a key influence. People with a family history of obesity were more likely to fall into Obesity Type II or III, proving that both genetics and shared family lifestyles contribute to obesity. Gender differences were noticed as well. Females showed a slightly higher presence in some obesity categories, which may be due to hormonal or lifestyle factors. Age also showed a positive relationship with BMI — older individuals tended to have higher BMI values, likely due to slower metabolism and reduced activity levels over time.

Other lifestyle habits also played a part. Drinking enough water was linked to healthier weight levels, while the mode of transportation had an effect too — people who mostly used cars had higher obesity risks than those who walked or biked. Smoking showed a mixed connection with obesity, but even if it sometimes relates to lower weight, its overall health effects are harmful.

These findings agree with previous research, such as studies by Mozaffarian et al. (2011) showing how unhealthy diets raise obesity risk, and by Fogelholm and Kukkonen-Harjula (2000), who confirmed that regular physical activity protects against obesity. The connection of family history with obesity also supports findings by Stunkard et al. (1986) and Locke et al. (2015).

Overall, this study achieved its main goals of identifying which factors have the greatest impact on obesity risk. It clearly shows that obesity is influenced by both factors that cannot be changed, like age and genetics, and factors that can be changed, like diet, physical activity, and transportation. The results highlight that improving eating habits and doing more exercise are the most effective ways to prevent obesity.

However, this study also had some limitations. The data was cross-sectional, meaning it only captured information at one point in time, so cause and effect cannot be proven. The dataset came from Kaggle, so the researchers had no control over how some variables were collected. Also, some data like food intake and activity levels were self-reported, which may not be completely accurate. Lastly, since the dataset was not collected from all populations, the results cannot be fully generalized to everyone.

For future research, it would be better to conduct long-term studies that track people over time to see how obesity changes. More accurate measurements using devices like activity trackers can help reduce reporting errors.

Researchers could also study other influences such as income, stress, and environment. Using advanced techniques like machine learning could also improve obesity prediction models.

For real-world practice, public health campaigns should focus on educating people about healthy eating and regular physical activity. Healthcare workers can use the findings to identify people at high risk and give them early guidance. Governments and city planners should also promote active transport options like walking and cycling by improving safe infrastructure.

In conclusion, this study proves that obesity is a complex problem caused by many factors working together. Understanding these causes can help create better, evidence-based solutions to reduce obesity and improve public health globally.

References

- [1] J. Audrain-McGovern and N. L. Benowitz, "Cigarette smoking, nicotine, and body weight," *Clinical Pharmacology & Therapeutics*, vol. 90, no. 1, pp. 164–168, 2011.
- [2] E. A. Dennis, A. L. Dengo, D. L. Comber, *et al.*, "Water consumption increases weight loss during a hypocaloric diet intervention in middle-aged and older adults," *Obesity*, vol. 18, no. 2, pp. 300–307, 2010.
- [3] T. M. Dugan, S. Mukhopadhyay, A. Carroll, and S. Downs, "Machine learning techniques for prediction of early childhood obesity," *Applied Clinical Informatics*, vol. 6, no. 3, pp. 506–520, 2015.
- [4] M. Fogelholm and K. Kukkonen-Harjula, "Does physical activity prevent weight gain – a systematic review," *Obesity Reviews*, vol. 1, no. 2, pp. 95–111, 2000.
- [5] E. Goodman, B. R. Hinden, and S. Khandelwal, "Accuracy of teen and parental reports of obesity and body mass index," *Pediatrics*, vol. 106, no. 1, pp. 52–58, 2003.
- [6] A. Keys, F. Fidanza, M. J. Karvonen, N. Kimura, and H. L. Taylor, "Indices of relative weight and obesity," *Journal of Chronic Diseases*, vol. 25, no. 6, pp. 329–343, 1972.
- [7] J. L. Kuk, T. J. Saunders, L. E. Davidson, and R. Ross, "Age-related changes in total and regional fat distribution," *Ageing Research Reviews*, vol. 8, no. 4, pp. 339–348, 2009.
- [8] A. E. Locke, B. Kahali, S. I. Berndt, *et al.*, "Genetic studies of body mass index yield new insights for obesity biology," *Nature*, vol. 518, no. 7538, pp. 197–206, 2015.
- [9] D. Mozaffarian, T. Hao, E. B. Rimm, W. C. Willett, and F. B. Hu, "Changes in diet and lifestyle and long-term weight gain in women and men," *New England Journal of Medicine*, vol. 364, no. 25, pp. 2392–2404, 2011.
- [10] A. Must, J. Spadano, E. H. Coakley, A. E. Field, G. Colditz, and W. H. Dietz, "The disease burden associated with overweight and obesity," *JAMA*, vol. 282, no. 16, pp. 1523–1529, 1999.
- [11] A. J. Stunkard, T. T. Foch, and Z. Hrubec, "A twin study of human obesity," *JAMA*, vol. 256, no. 1, pp. 51–54, 1986.
- [12] World Health Organization, "Constitution of the World Health Organization," 1948. [Online]. Available: <https://www.who.int/about/governance/constitution>. [Accessed: Oct. 1, 2025].
- [13] World Health Organization, "Obesity and overweight fact sheet," 2024. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight>. [Accessed: Oct. 1, 2025].

[14] Kaggle Dataset, “Multi-Class Prediction Obesity Risk,” 2024. [Online]. Available: <https://www.kaggle.com/code/najeebz/multi-class-prediction-obesity-risk-xgboost-0-9>. [Accessed: Oct. 5, 2025].

Appendices

Data Import and Sampling

```
setwd("C:\\\\Users\\\\Vidhu\\\\Desktop\\\\ps Assignments")
getwd()

my_data <- read.csv("data.csv") # Import dataset
set.seed(123) # For reproducibility
sample_rows <- my_data[sample(nrow(my_data), 200), ]
head(sample_rows)
```

Demographic Analysis - Gender

```
# Create pie chart
pie(counts, labels = paste(gender, " (", round(counts/sum(counts)*100, 2), "%)", 
    main = "Figure 3.4.2.1: Pie Chart Gender (200 Records)",
    col = c("#1f77b4", "#ff7f0e")) # Blue for Male, Orange for Female
```

Age Categorization

```
# Categorize Age into four levels
sample_rows$Age_Category <- cut(sample_rows$Age,
    breaks = c(-Inf, 18, 24, 30, Inf),
    labels = c("Less than 18", "18-24", "25-30", "Above 30"),
    right = FALSE)
```

Obesity Category Analysis

```
# Calculate obesity distribution from sample_rows using NObeyesdad column
obesity_distribution <- table(factor(sample_rows$NObeyesdad, levels = obesity_categories))
total_sample <- nrow(sample_rows) # Total sample size: 200
percentages <- round(prop.table(obesity_distribution) * 100, 2)

# Create pie chart
pie(as.numeric(obesity_distribution),
    labels = paste(obesity_categories, " (", percentages, "%)", 
    main = "Figure: Pie Chart of Obesity Categories (200 Records)",
    col = c("#1f77b4", "#ff7f0e", "#2ca02c", "#d62728", "#9467bd", "#8c564b", "#e377c2"))
```

Categorical Variables Analysis

```
# Calculate family history distribution from sample_rows - CORRECTED COLUMN NAME
family_history_distribution <- table(factor(sample_rows$family_history_with_overweight,
    levels = family_history_categories))

# Calculate FAVC distribution
favorite_distribution <- table(sample_rows$FAVC)
```

```
# Calculate smoking distribution
smoking_distribution <- table(sample_rows$SMOKE)
```

Numerical Variables Analysis

```
"# Calculate height statistics          height analysis
height_stats <- data.frame(
  Statistic = c("Mean", "Median", "Standard Deviation", "Variance", "Range", "Minimum", "Maximum"),
  Value = c(
    round(mean(sample_rows$Height), 2),
    round(median(sample_rows$Height), 2),
    round(sd(sample_rows$Height), 4),
    round(var(sample_rows$Height), 4),
    paste(round(min(sample_rows$Height), 2), "-", round(max(sample_rows$Height), 2)),
    round(min(sample_rows$Height), 2),
    round(max(sample_rows$Height), 2)
  )
)

# Calculate weight statistics
weight_stats <- data.frame(
  Statistic = c("Mean", "Median", "Standard Deviation", "Variance", "Range", "Minimum", "Maximum"),
  Value = c(
    round(mean(sample_rows$Weight), 2),
    round(median(sample_rows$Weight), 2),
    round(sd(sample_rows$Weight), 2),
    round(var(sample_rows$Weight), 2),
    paste(round(min(sample_rows$Weight), 0), "-", round(max(sample_rows$Weight), 0)),
    round(min(sample_rows$Weight), 2),
    round(max(sample_rows$Weight), 2)
  )
)
```

BMI Calculation and Analysis

```
# Calculate BMI
sample_rows$BMI <- sample_rows$Weight / (sample_rows$Height^2)

# Calculate BMI statistics
bmi_stats <- data.frame(
  Statistic = c("Mean", "Median", "Standard Deviation", "Variance", "Range", "Minimum", "Maximum"),
  Value = c(
    round(mean(sample_rows$BMI), 2),
    round(median(sample_rows$BMI), 2),
    round(sd(sample_rows$BMI), 2),
    round(var(sample_rows$BMI), 2),
    paste(round(min(sample_rows$BMI), 2), "-", round(max(sample_rows$BMI), 2)),
    round(min(sample_rows$BMI), 2),
    round(max(sample_rows$BMI), 2)
  )
)
```

BMI Categorization

```
# Create BMI categories
sample_rows$BMI_Category <- cut(sample_rows$BMI,
                                breaks = c(0, 18.5, 25, 30, Inf),
                                labels = c("Underweight", "Normal", "Overweight", "Obese"),
                                right = FALSE)
```

Data Visualization

```
hist(sample_rows$Weight,
      main = "Weight Distribution - 200 Records",
      xlab = "Weight (kg)",
      ylab = "Frequency",
      col = "#1f77b4",
      border = "white",
      breaks = 20)
dev.off()
cat("Histogram saved as: weight_histogram_200.png\n")
```

```
hist(sample_rows$Height,
      main = "Height Distribution - 200 Records",
      xlab = "Height (meters)",
      ylab = "Frequency",
      col = "#1f77b4",
      border = "white",
      breaks = 15)
dev.off()
cat("Histogram saved as: height_histogram_200.png\n")
```

```
hist(sample_rows$Height,
      main = "Height Distribution - 200 Records",
      xlab = "Height (meters)",
      ylab = "Frequency",
      col = "#1f77b4",
      border = "white",
      breaks = 15)
dev.off()
cat("Histogram saved as: height_histogram_200.png\n")
```

Data Export

```
# Save table as CSV  
write.csv(water_stats, "water_consumption_table_200.csv", row.names = FALSE)
```