اونيۏرسيتي مليسيا ڤهڠ السلطان عبدالله

UNIVERSITI MALAYSIA PAHANG
AL-SULTAN ABDULLAH

BSD2223 DATA SCIENCE PROGRAMMING LANGUAGE II
GROUP PROJECT

PREDICTIVE MODELLING OF OBESITY RISK FACTOR

TOPIC: GOOD HEALTH AND WELLBEING (SDG 3)

PREPARED FOR:
ASSOCIATE PROFESSOR DR. ROSLINAZAIRIMAH ZAKARIA

PREPARED BY: THE WELLNESS WARRIORS

MATRIC ID	NAME	SECTION
SD22003	NUR ATIEKA RAFIEKAH BINTI RAZAK	01G
SD22034	MUHAMMAD AIMAN IRFAN BIN AHMAD MUHYE	
SD22056	ELIANE HO WAN WEN	
SD22017	BATRISYIA BINTI ISMAIL	

ABSTRACT

Obesity is a global health concern that has a big impact on people’s health, well-being, and access to health care. The aim of this project is to use data analytics to predict and reduce diabetes diseases. Also, our project aims to determine the relative contribution of several genetic, behavioral, and environmental risk factors to the overall prediction of obesity by analyzing a wide range of variables, such as age, gender, dietary habits, level of physical activity, and lifestyle factors. The objective is to develop a personalized strategy for obesity prevention and management based on individual risk profiles. This project has the potential to significantly improve public health strategies, identify high-risk individuals, and guide targeted treatments. The insights gained from this project can contribute to reducing obesity-related health issues and advancing global health initiatives. Overcoming these limitations in future projects through expanded datasets and comprehensive temporal analysis will contribute to the development of more reliable and efficient predictive models for obesity risk factors.

TABLE OF CONTENT

1.0 INTRODUCTION

Predictive modeling of obesity risk factors is an important field of study that aims to use data analytics to predict and reduce an individual's chance of becoming obese. Obesity or the collection of excessive body fat, is a global health concern that has a big impact on people's health, well-being, and access to health care. For effective handling of this issue, specific treatments and strategies may be developed with the help of an understanding of the prediction of the factors that lead to obesity.

Obesity is led by a complex interaction of genetic, behavioral, and environmental factors. Factors such as gender, age, height, weight, family history with overweight, frequent consumption of high-caloric food, frequency of consumption of vegetables, number of main meals, consumption of food between meals, smoke, daily water consumption, caloric beverages consumption, physical activity frequency, time spent using technological devices, consumption of alcohol, mode of transportation and target variable representing obesity level play an important role in determining an individual's risk of obesity. Predictive modeling uses these factors to create models that can predict an individual's likelihood of developing obesity.

Recent developments in data collection, storage, and analysis techniques have enabled researchers to obtain huge amounts of data on obesity risk factors. Predictive modelling approaches can evaluate this data to discover patterns and trends that can help forecast obesity risk more accurately than standard statistical methods.

Objective of this study:

- To investigate the impact of different obesity risk factors on the overall prediction of obesity risk and figure out their relative importance.
- To provide knowledge about the development of personalized strategies and actions for obesity prevention and management based on individual risk profiles.
- To visualize the behavior and habits regarding their obesity level.

2.0 PROJECT DESCRIPTION

The growing concern about the worldwide obesity pandemic was the reason for selecting this project on "Predictive Modelling of Obesity Risk Factors." The problem of obesity is increasing globally, thus it is critical to learn about its risk factors and create efficient predictive models in order to identify people at high risk and give ways to prevent in action.

The project is important because it has the potential to improve public health approaches by applying methods for predictive modeling. By an analysis of a wide range of variables, such as age, gender, dietary habits, levels of physical activity, and lifestyle factors, this project aims to identify trends and connections that increase the risk of obesity. To reduce the risk of problems linked to obesity, targeted treatments, and individual healthcare plans can be informed by an understanding of these characteristics.

Moreover, the innovative nature of this project lies in its diverse approach, combining data science, public health, and behavioral research to address the objectives outlined in Sustainable Development Goal 3. By employing advanced predictive modeling techniques alongside domain-specific knowledge, the project has the potential to produce insights that transcend conventional epidemiology and contribute important to the global efforts towards achieving SDG 3. The project addresses one of the most important public health issues of our day by utilizing data and predictive analytics, but it also strongly relates to SDG 3's overall goal of promoting healthy lifestyles and well-being for people of all ages worldwide.

3.0 DATA DESCRIPTION

This dataset contains the risk of obesity which was meticulously curated for research and analysis in the domain of health and lifestyle studies. All the 18 columns in the dataset indicate the comprehensive information on individuals and encompassing key attributes. The dataset contains 20759 rows with 18 columns.

No	Title	Explanation	Type
1	Gender	Gender of the Patients	Qualitative
2	Age	Age of the Patients	Quantitative
3	Height	Height of the Patients	Quantitative
4	Weight	Weight of the Patients	Quantitative
5	family_history_with_overweight	Family History of Patients with Overweight; 'No' and 'Yes'	Qualitative
6	FAVC	Frequent consumption of high-caloric food; 'No' and 'Yes'	Qualitative
7	FCVC	Frequency of consumption of vegetables	Quantitative
8	NCP	Number of main meals	Quantitative
9	CAEC	Consumption of food between meals; 'No', 'Sometimes', 'Frequently' and 'Always'.	Qualitative
10	SMOKE	History of Smoking of Patients; 'No' and 'Yes'	Qualitative
11	CH2O	Daily water consumption	Quantitative

12	SCC	Caloric beverages consumption; ‘No’ and ‘Yes’	Qualitative
13	FAF	Physical activity frequency	Quantitative
14	TUE	Time spent using technological devices	Quantitative
15	CALC	Consumption of alcohol; ‘No’, ‘Sometimes’ and ‘Frequently’	Qualitative
16	MTRANS	Mode of transportation; ‘Automobile’, ‘Bike’, ‘Motorbike’, ‘Public_Transportation’ and ‘Walking’	Qualitative
17	Level	Target variable representing obesity level; ‘Under_Weight’, ‘Normal_Weight’, ‘Obesity_Type_I’, ‘Obesity_Type_II’, ‘Obesity_Type_III’, ‘Overweight_Level_I’ and ‘Overweight_Level_II’	Qualitative
18	BMI	BMI of the patients	Quantitative

4.0 DATA PREPARATION

In this part, we will prepare the obtained data to get a better insight into the future analysis. First, we do the renaming column to make sure the name represents well the data they hold and give the audience an early description of the data in the column. Sometimes, we also rename the column to remove unwanted characters or symbols to make it nice. In this dataset, there is about one column we renamed which is Obe1dad to become column Level.

Next, we change the data types for column Age and NCP from numeric to integer. It is due to age and number of meals are discrete attributes cannot be decimal. The attributes' height, weight, FCVC, FAF, TUE, CH20 are rounded to 2 decimal places. It helps the user to easily observe with 2 decimal places compared to the previous 5 decimal places.

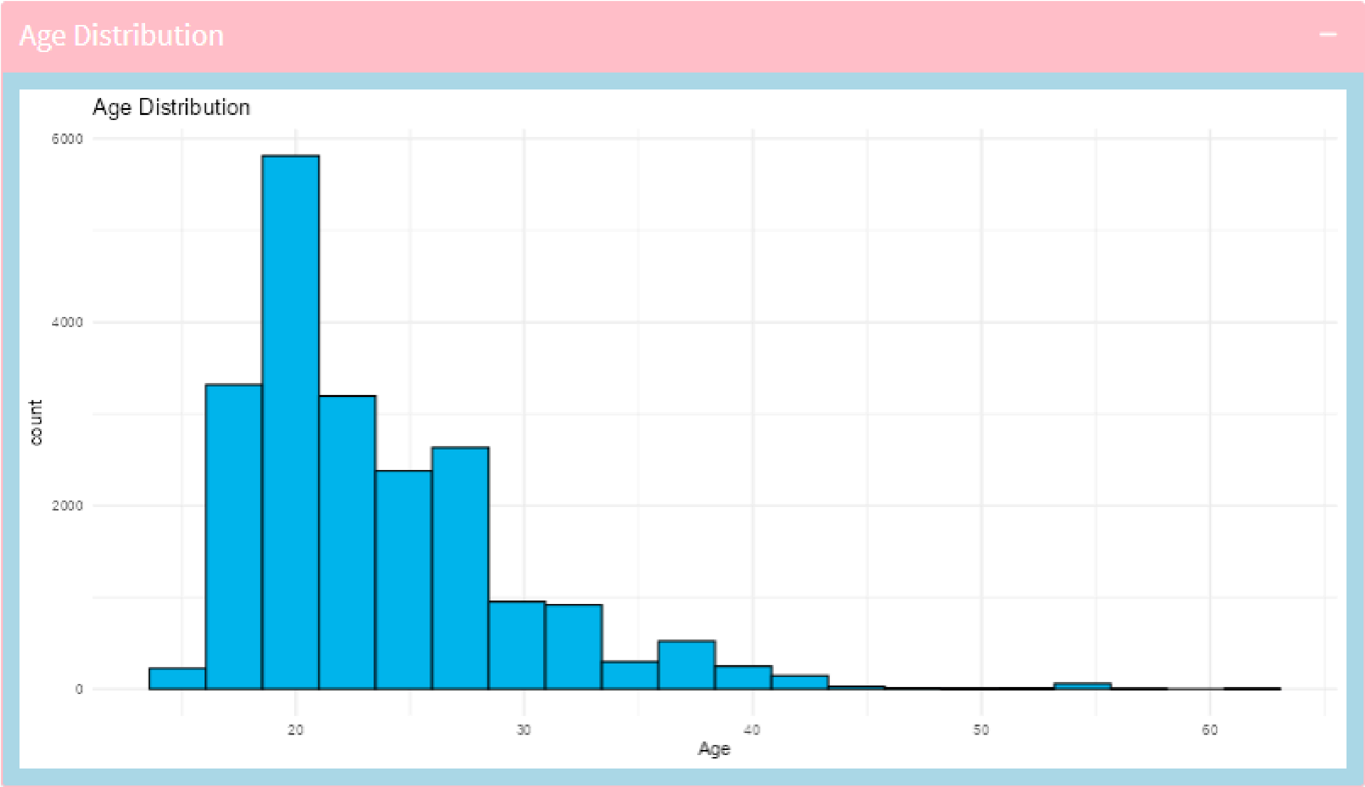
A few columns need to change the data. It is due to prevent any confusion of the data. For example, the value 1 in column family_history_with_overweight represents the family already involved in the overweight issue. Therefore to avoid any confusion we rename from 1 to Yes, and 0 to No. Then, we look at the attributes smoke, CALC, and, FAVC the data we adjust from 1 to Yes and 0 to No. The Level column also need to be changed Ormal_weight to Normal_weight and insufficient_weight to under_weight.

We created a new column names BMI calculated using the formula weight divided by height squared. This provide a better understanding the data related to body mass. Then, check for missing values in the data. Missing value or NA could impact the interpretation of the data when analysis begins and fortunately, our dataset didn’t encounter any missing value.

5.0 DATA ANALYSIS, RESULT AND DISCUSSION

5.1 Dashboard 1: Overview Analysis

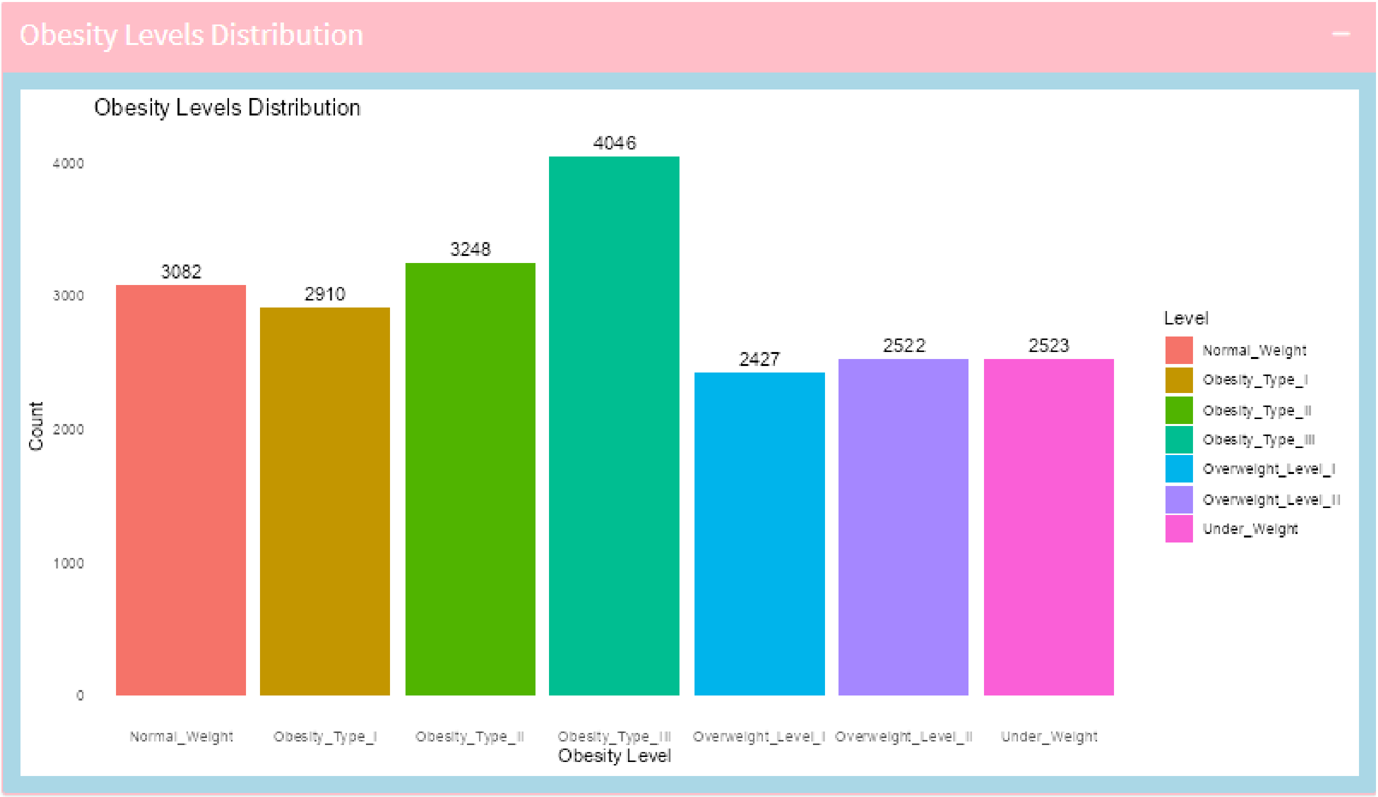
i) Age Distribution



This histogram gives us a clear picture of the age distribution in the obesity risk dataset. It shows that there are more people in their 20s compared to other age groups. The highest number of individuals, around 6000, falls around the age of 20. As we move towards older ages, the number of individuals decreases significantly, especially for those over 50. This age distribution is important because it tells us that the dataset mainly consists of younger people.

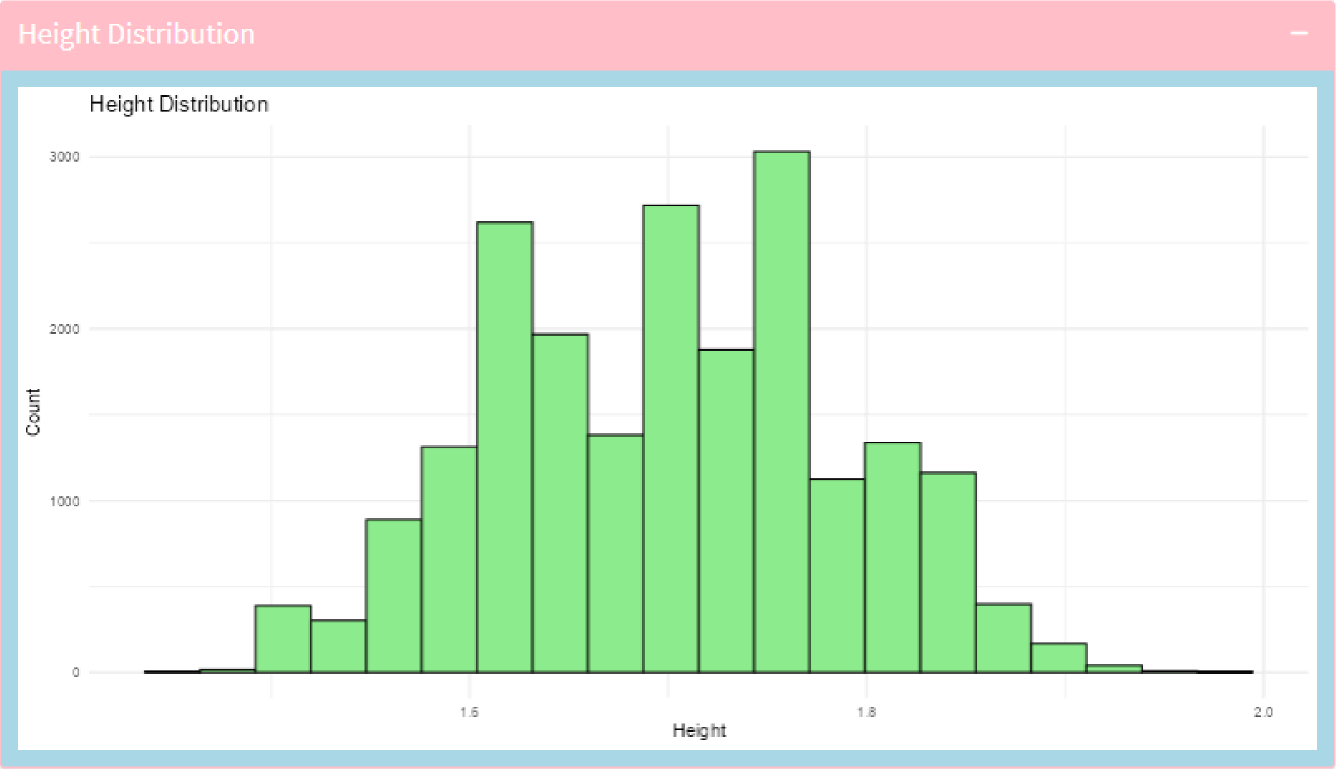
This demographic skew has a big impact on how we predict and interpret obesity risk. Younger individuals may have different factors that contribute to their risk of obesity compared to older people. They may have different eating habits, levels of physical activity, and lifestyle choices. For example, younger people might be more likely to eat high-calorie fast food, spend more time on electronic devices, and have varying levels of physical activity. All of these factors can affect their risk of obesity.

ii) Obesity Levels Distribution



The bar chart shows how many people in the dataset fall into different categories of obesity. The highest count is for Obesity_Type_III, with 4046 individuals, which indicates a significant number of people with severe obesity. Next is the Normal_Weight category, with 3082 individuals. Obesity_Type_II and Obesity_Type_I have 3248 and 2910 individuals respectively. The counts for Overweight_Level_I, Overweight_Level_II, and Under_Weight are all around 2500 individuals, which is relatively close. This distribution highlights the urgent need for targeted interventions, especially for those with severe obesity. It also shows that overweight and obesity are widespread issues within the dataset. These findings are crucial for our project's goal of predicting obesity risk factors. They emphasize the importance of addressing severe obesity and developing public health strategies to reduce obesity rates and promote healthier lifestyles.

iii) Height Distribution

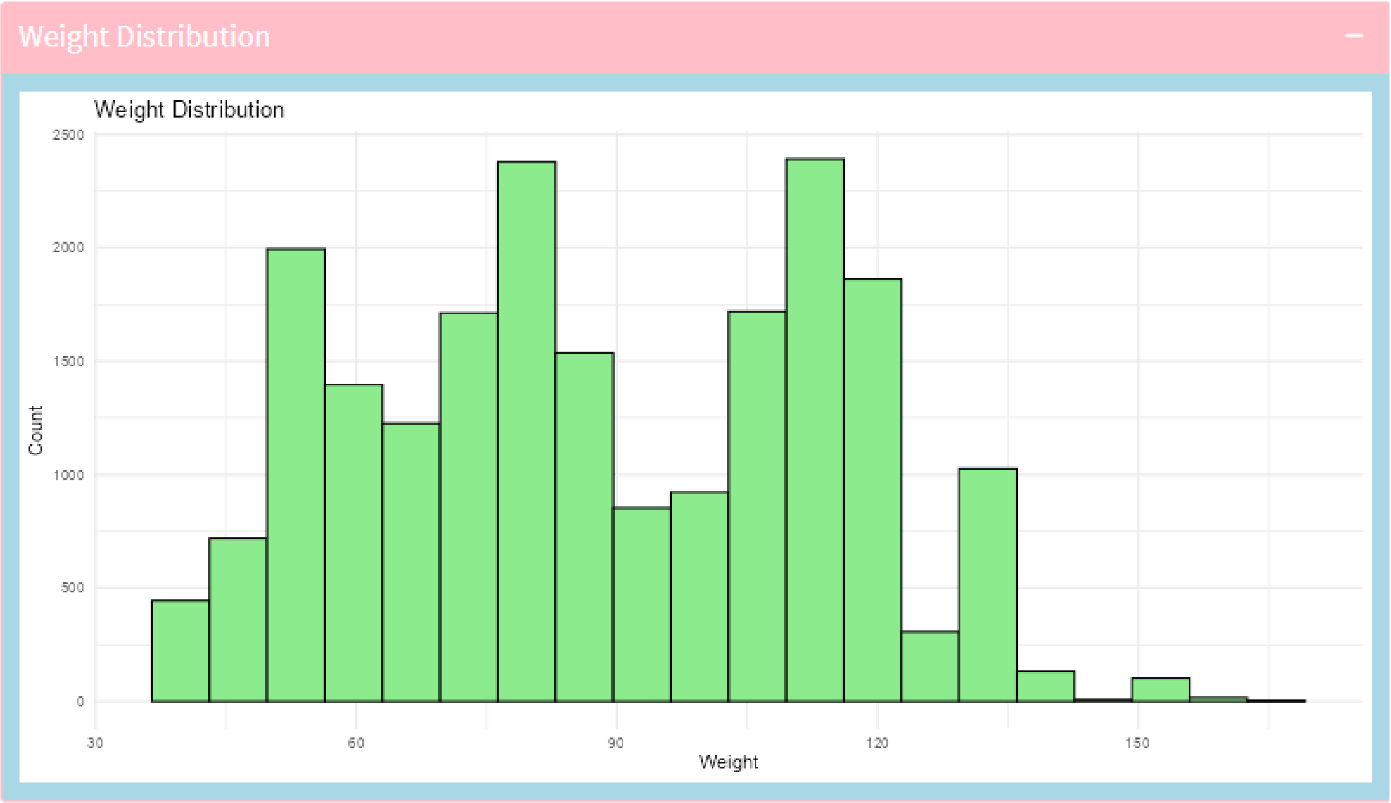


This histogram displays how heights are spread out in obesity risk dataset. The horizontal axis shows height, while the vertical axis shows the number of people with that height. Most people in the data fall between 1.6 and 1.8 meters tall.

This kind of information is helpful for studying the risk of obesity, since height plays a big role in calculating body mass index (BMI). BMI is a way to measure body fat using height and weight, and it's commonly used to assess the risk of obesity.

By looking at how heights are distributed in a dataset, we can pinpoint individuals who might have a higher chance of being obese based on their height. This data can then be used to create specific strategies to prevent and treat obesity.

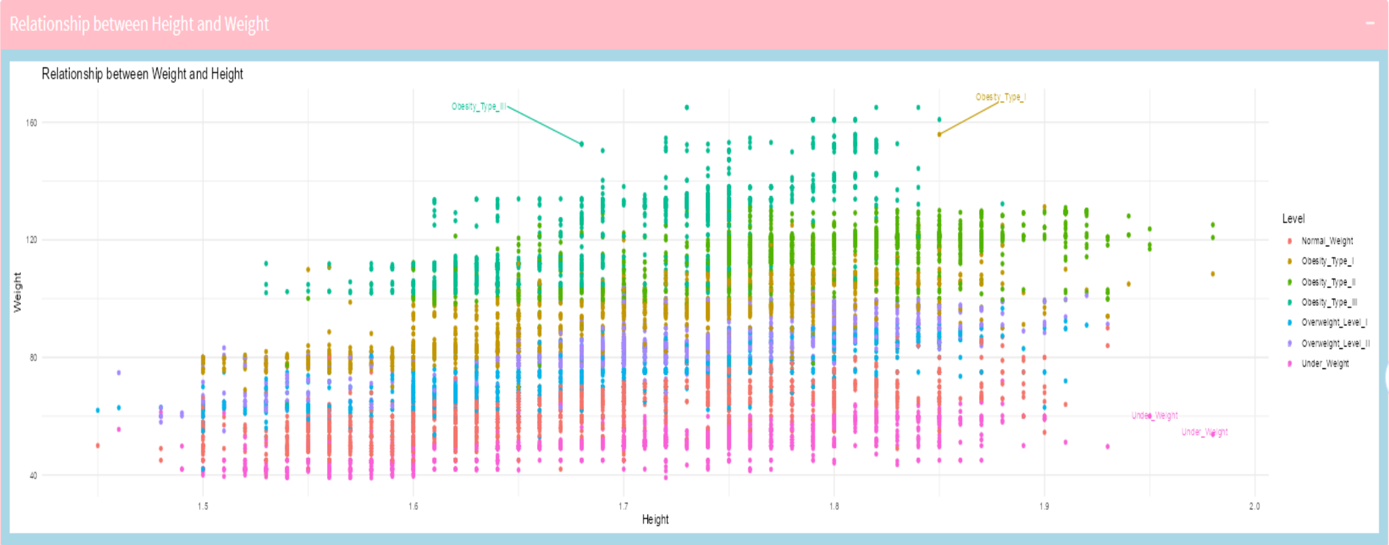
iv)Weight Distribution



This histogram displays how weight is distributed in obesity risk dataset. The horizontal axis shows weight, while the vertical axis shows the number of people in each weight range. The graph indicates that the most common weight range is between 70 and 100, with a peak around 80. As weight increases, there is a noticeable decrease in the number of people.

This chart could help analyze the prevalence of different weight categories in an obesity risk dataset. It could show the percentage of people who are overweight or obese based on their BMI. Additionally, it could highlight areas that need attention for interventions to lower the risk of obesity. For instance, obesity risk dataset includes details about eating habits, this graph could reveal if there is a connection between weight and specific dietary choices.

v)Relationship Between Height and Weight

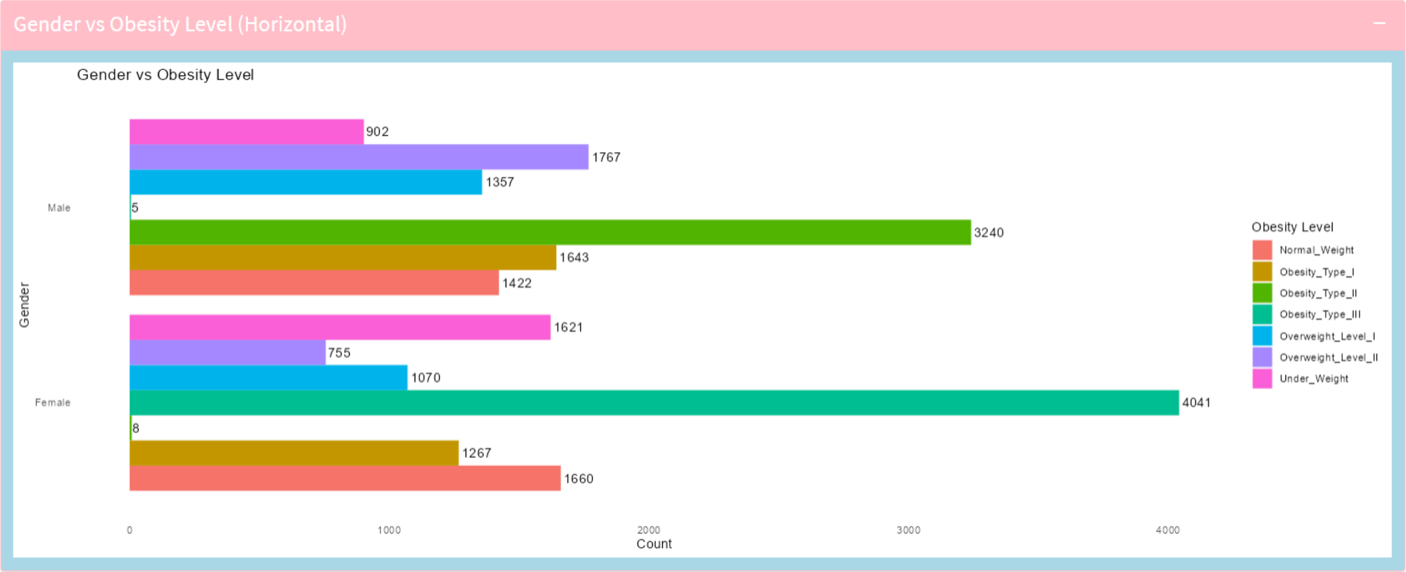


The scatter plot shows how height and weight are related, based on different levels of obesity and weight status. It reveals that taller people generally have higher weights, but there are variations within the same height range. This means that height and weight isn't enough to predict obesity risk. Factors like diet and lifestyle also play important roles. For instance, people of similar heights and weight but different eating habits may end up in different obesity level categories. This shows that preventing and managing obesity requires looking at more than just height. Understanding individual habits and lifestyle is key.

It shows how complicated it is to determine the risk of obesity and pinpoint the importance of looking at more than just height and weight. Knowing a person's eating habits, decisions, and daily routine is crucial for preventing and controlling obesity.

5.2 Dashboard 2: Demographics Analysis

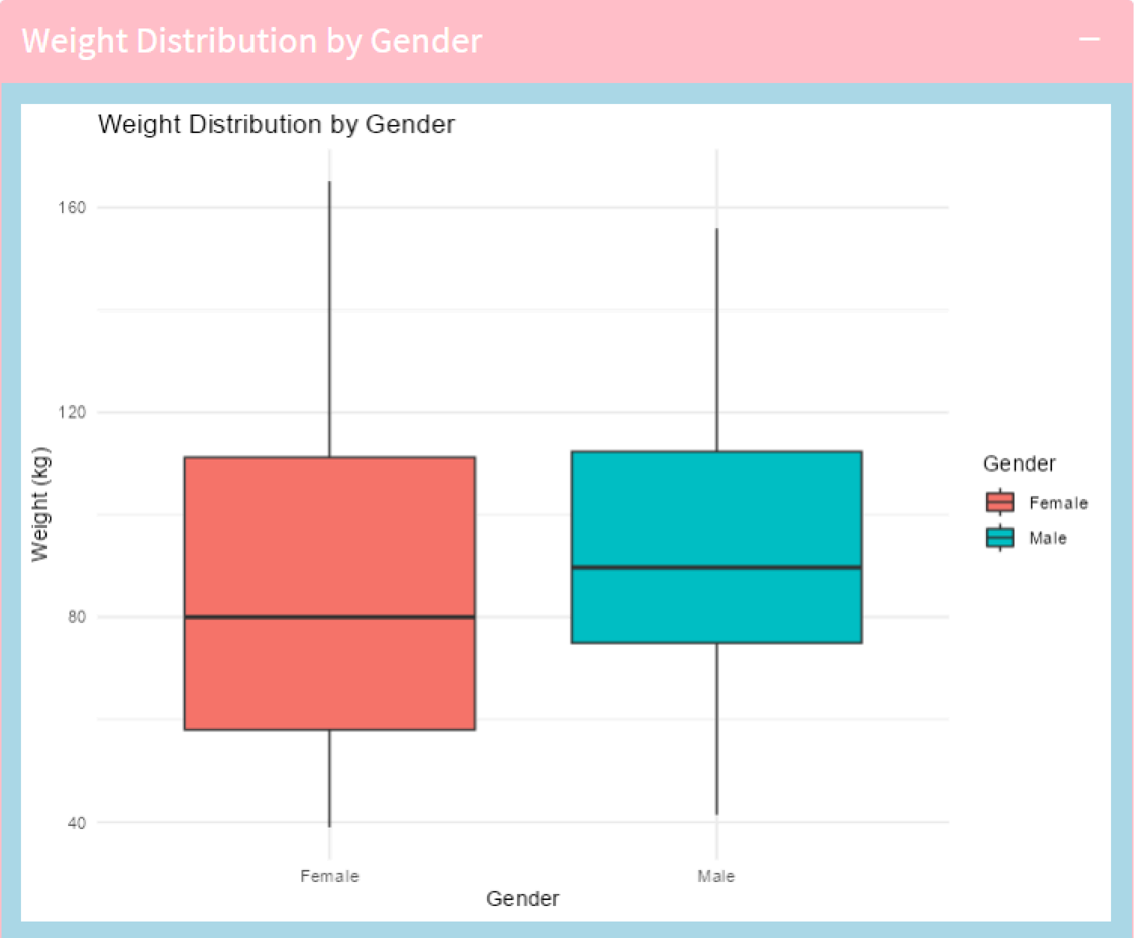
i) Gender vs Obesity Level



The data shows a trend for male and female obesity levels. Females have the highest count which indicates the Obesity type III category with a staggering 4041 patients. It has a huge gap with the second highest category normal weight which has 1660 patients. This significant gap indicates a potential bimodal distribution in which some of the female population demonstrates good dietary and lifestyle habits, while the part may lack proper care and be classified as Obesity Type III. it is crucial for the government to take action on this issue since the obesity type 3 is extremely obesity and may be associated with more severe metabolic dysfunction. In addition, the female patient that suffers from obesity type 2 diabetes has the smallest only 8 people.

Meanwhile, for males, the highest category are Obesity type II with 3240 patients. The overall data for males also needs to be concern. Similar to the female population, a majority of males fall within various obesity categories compared to the Normal Weight group. Then, despite having a lower number of "Obesity Type III" patients (only 5) compared to females, the high prevalence of obesity across other categories warrants attention.

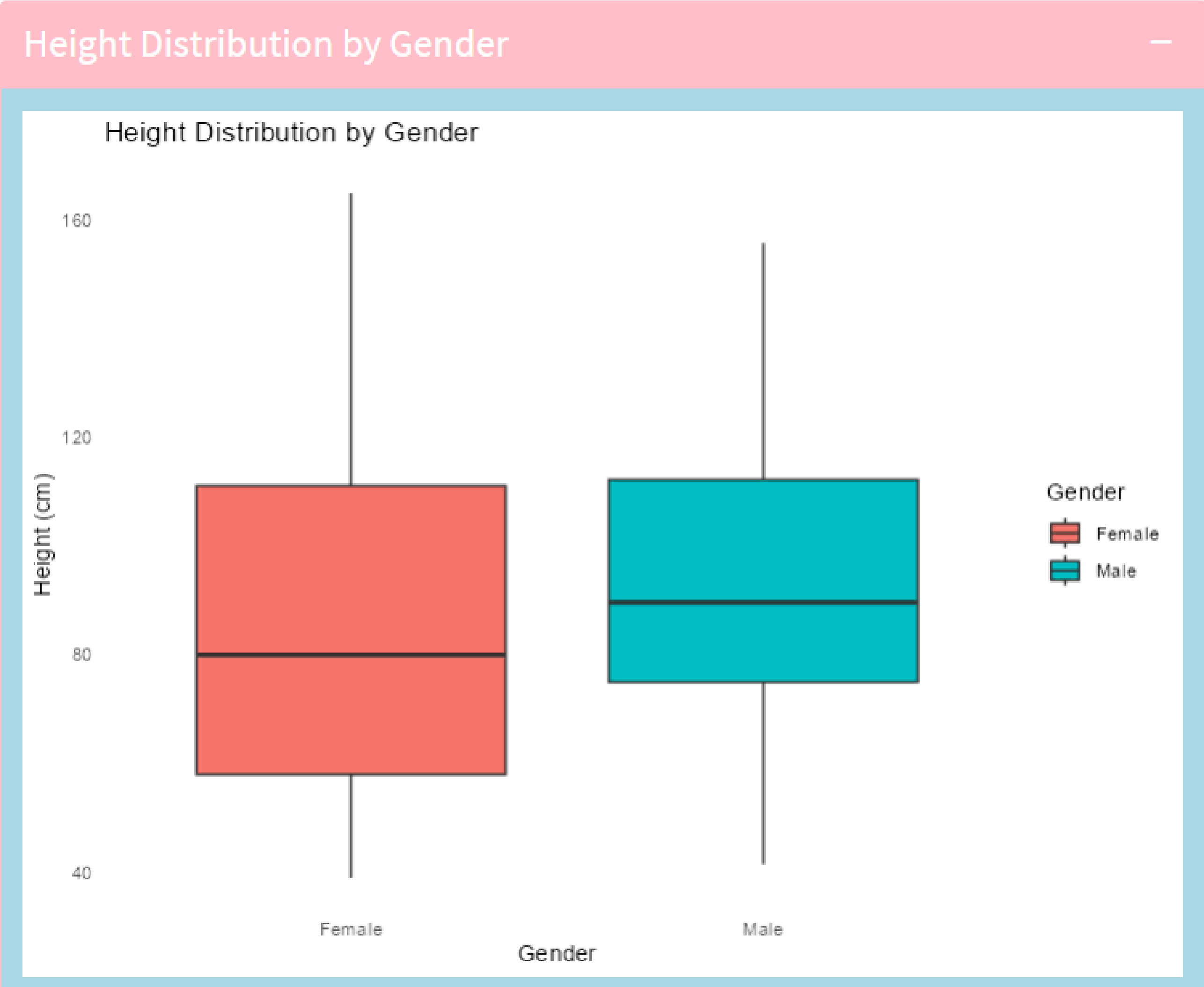
ii) Weight Distribution by Gender



The graph shows the distribution of weight in kg for both males and females. The boxplot graph is used to show the range of weight values and median. For the female gender, the median is around 80kg. This indicates that 50% of the females have a weight below 80 kg and 50% have a weight above 80 kg. Female weight consists of IQR from 60 kg to 100kg with an overall weight range for females of almost 40 kg to 165 kg.

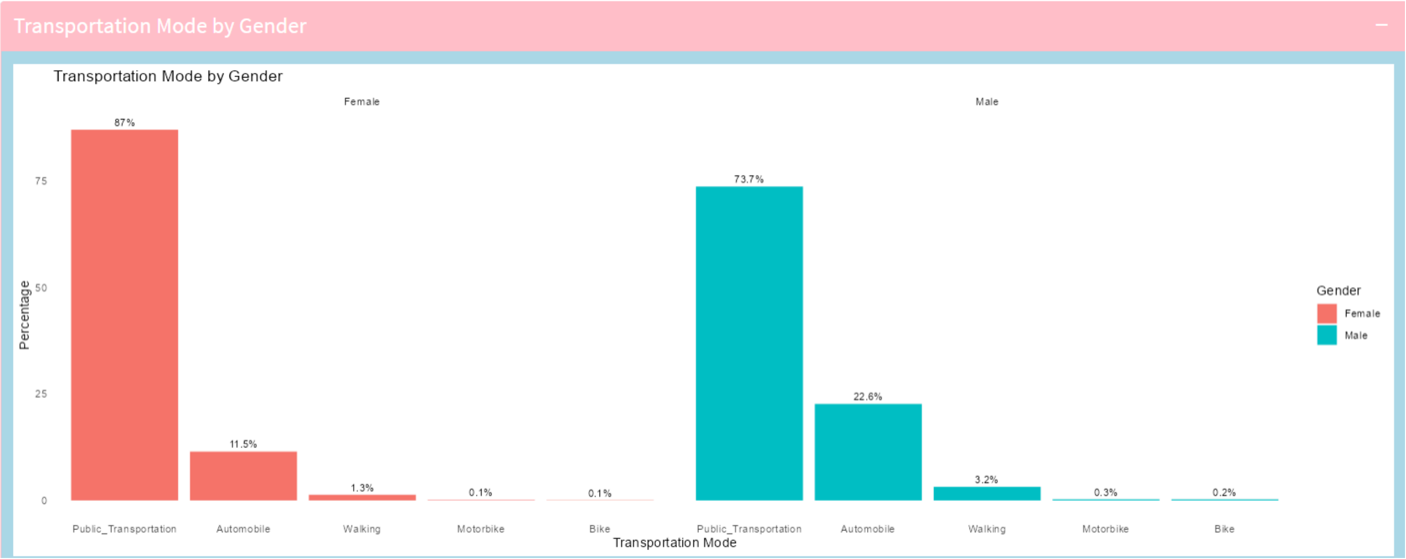
Meanwhile, for the male gender, the median weight is higher compared to females with around 100 kg. This indicates that half of the males weigh more than the females since 50% of males have weight below 100kg. The IQR for males is slightly wider ranging 80kg to 120 kg compared to females. The distribution of the weight for males is slightly higher compared to females. We can reflect that the higher median weight in males parallel to the higher prevalence of obesity observed in the previous chart, particularly in the Obesity Type II category. In addition, there are no outliers shown in this graph. It is good for us to make the analysis future.

iii) Height Distribution by Gender



The graph shows the distribution of the patient's height by graph. It shows the range of the height value for the patient. For the female gender, the median is around 0.8 m with IQR from 0.6m to 1 meter. For the male gender, the median height is higher, around 1.2 m, with IQR from 0.7m to 1.15 m. The distribution of height for males is significantly higher compared to females. It represents the male has a greater height. There are no extreme outliers shown in the boxplots for either gender

iii) Transportation Mode by Gender

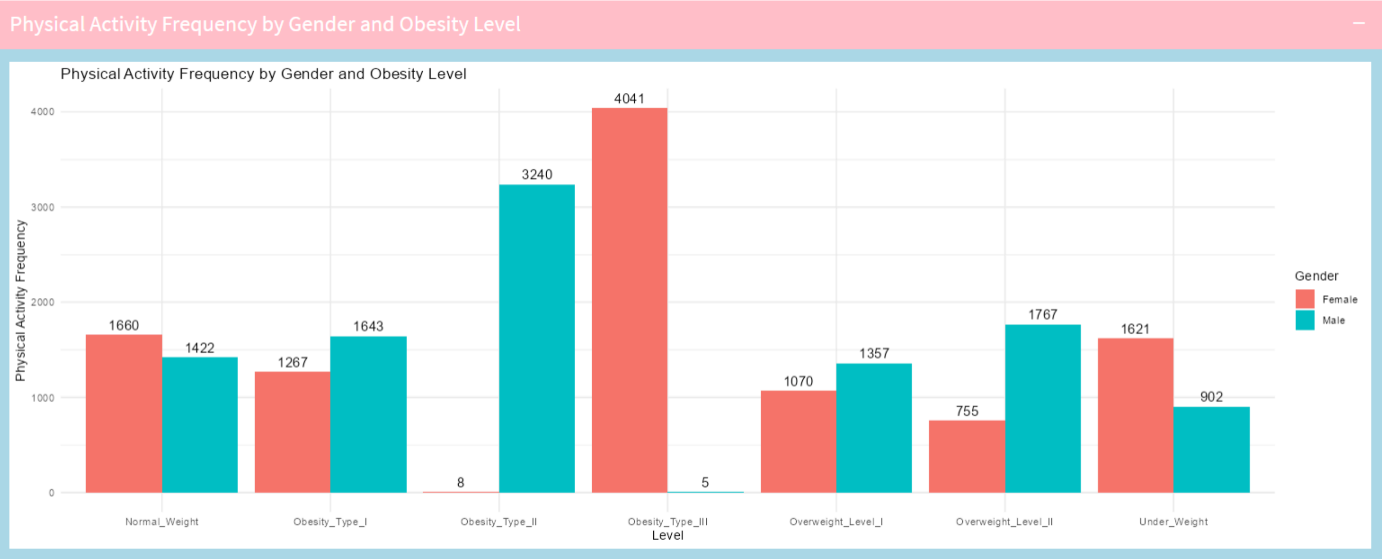


Based on this bar chart, we can see that public transportation has the highest percentage usage among females to use as their daily physical activities. It might be due to the convenient public transportation system for females. The lowest usage for females have two transportation which are Motorbike, and Bike. Same as females public transportation is the most used for transportation formalese. Since both gender categories are the same, we can assume that the data collected in the urban areas that have the most accessible, affordable, or convenient option.

Public transportation for females has a more significant lead which 87%comparede to males (73.7%). Like the females, there are no outliers in the male weight data, suggesting a uniform distribution without extreme deviations

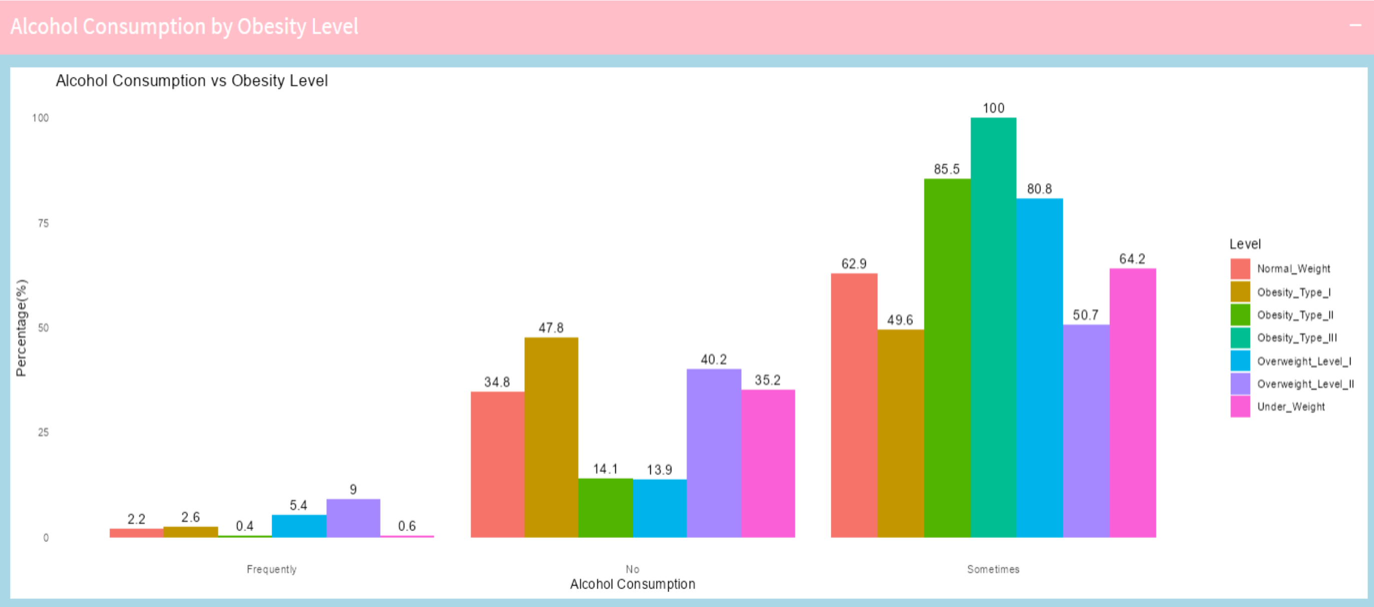
5.3 Dashboard 3: Lifestyle Analysis

i) Physical Activity Frequency by Gender and Obesity Level



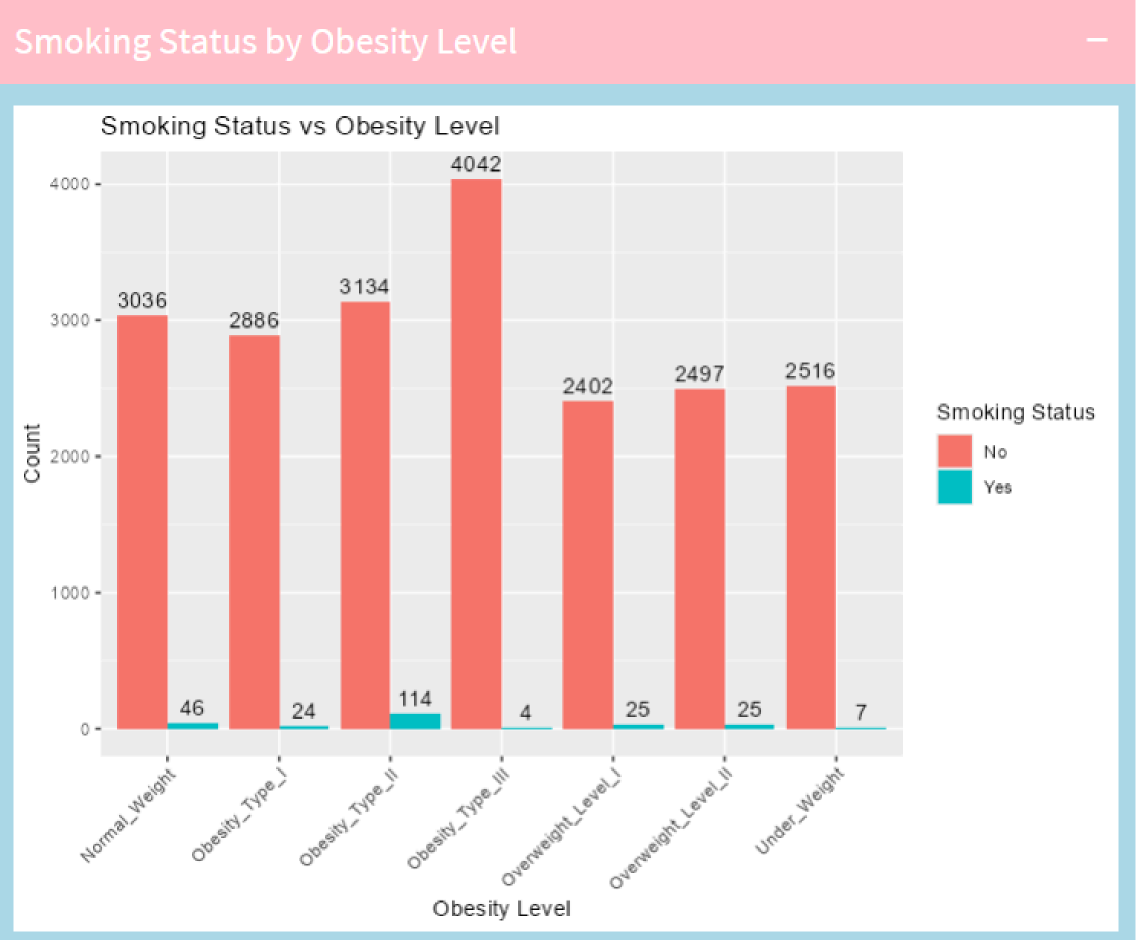
Based on the side-bar-side bar chart view, the highest physical activity frequency is female in Obesity_Type_III which is 4041 while the lowest physical activity frequency is male also in Obesity_Type_III which is only 5. The number of females in the category of Obesity_Type_III is higher than the number of males. This is due to the fact that females will be facing the problem of endocrine dyscrasia whicht can cause the weight of females to increase without any reason or symptoms. The male in Obesity_Type_IIhass the second highest physical activity frequency which is lower than the physical activity frequency of females in Obesity_Type_II but they are still under Obesity_Type_II even though their physical activity frequency is lower than the female in Obesity_Type_III. This is because they is working in a condition that needs a lot of energy, even though their physical activity frequency is lesser than the female in Obesity_Type_III, their body condition is better than the female in Obesity_Type_III even though they are under the condition of obesity.

ii) Alcohol Consumption by Obesity Level



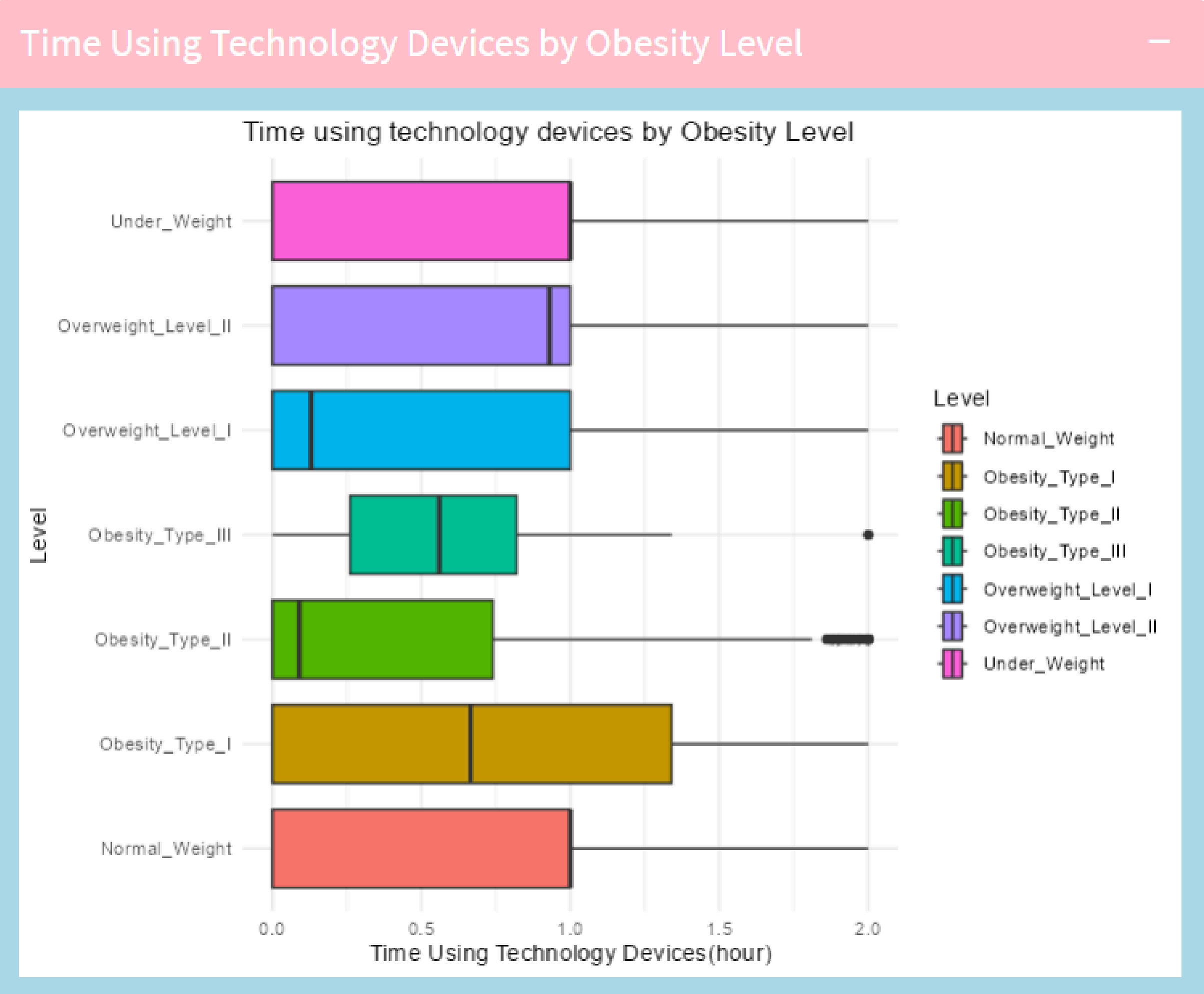
Based on the group bar chart view, the percentage of patients for all stages of level in the category of sometimes alcohol consumption is the highest while the percentage of patients for all stages of level in the category of frequently alcohol consumption is the lowest. It is due to the patient in the category of sometimes of alcohol consumption having a large amount of quantity of alcohol once they start to drink. The patient in the category of frequent alcohol consumption has a small amount of alcohol compared to others. In the category of sometimes alcohol consumption, the percentage of patients in the category of Obesity_Type_III is highest which is 100%. This is because the number of patients in the category of Obesity_Type_III is the highest. So, the category of sometimes alcohol consumption has the highest percentage of patients while the category of frequent alcohol consumption has the lowest percentage of patients.

iii) Smoking Status by Obesity Level



Based on the side-by-side bar chart view, the non-smoking amount is higher than smoker in all category obesity levels. The number of non-smoking in category Obesity_Type_III is highest compared to others which is 4042. This is due to the fact that even though they are not smoking at all, they still perform an unhealthy lifestyle and unhealthy diet to cause them to become overweight and eventually face the disease of obesity. Other than that, the number of smokers is lower in all category obesity levels because even though the risk of disease of obesity for smokers is higher than not smoking but they are performing a healthy diet and exercise regularly that can help them to reduce the risk of obesity. So, smoking is not the main factor that causes obesity but heavy smokers are more likely to be overweight or obese than are light smokers.

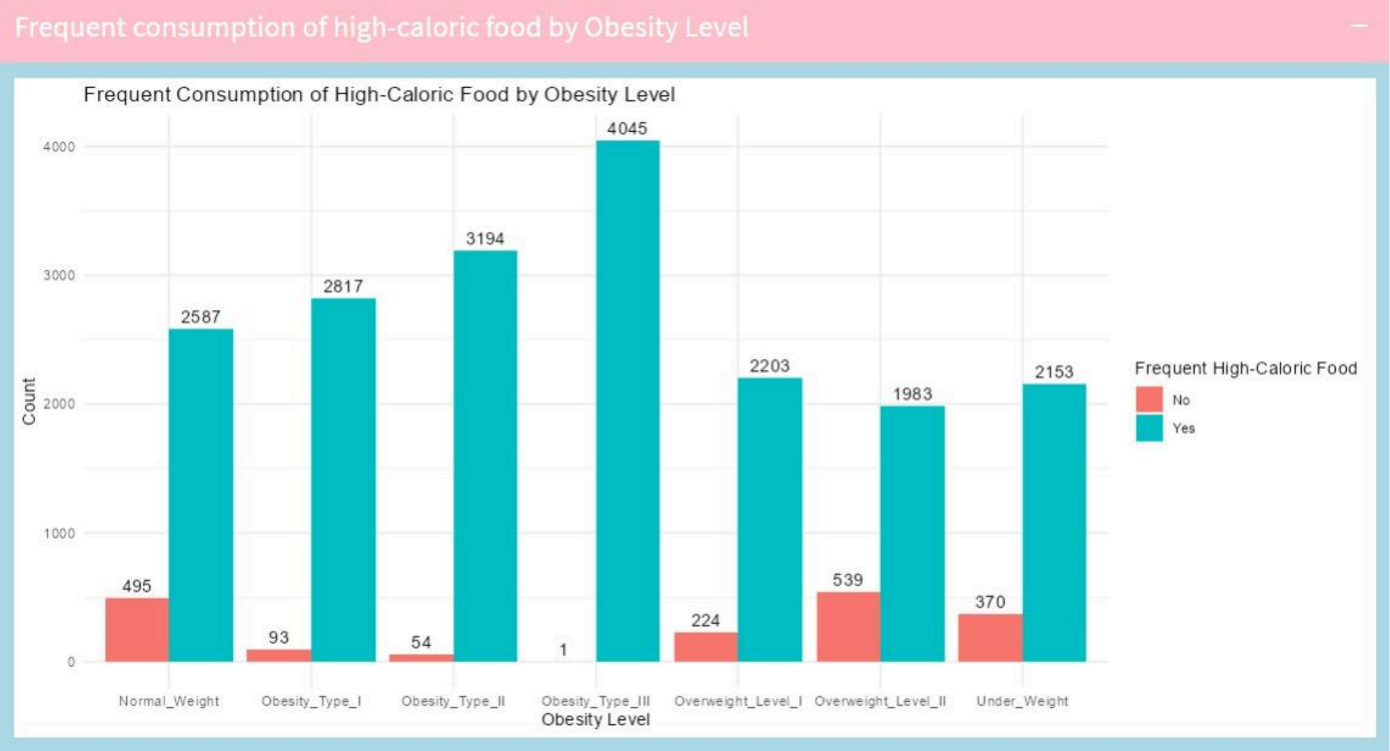
iv) Time using technology devices by Obesity Level



Based on the boxplot view, the two category levels which are Obesity_Type_II and Obesity_Type_III have outliers. The boxplot of Obesity_Type_II is a right-skewed distribution also called a positive skew distribution. With right-skewed distribution, the mean is greater than the median. There are many outliers in the category of Obesity_Type_II compared to Obesity_Type_III. The boxplot of Obesity_Type_III is right-skewed. There is one outlier in the category of Obesity_Type_III. The maximum point of Obesity_Type_III is the lowest compared to the others category levels while the maximum point of Obesity_Type_II is second lowest compared to others. The other category level of maximum points is the same which is 2 hours. Even though the time spent using technology devices in Obesity_Type_III and Obesity_Type_II is lowest and second lowest, it is still the most serious level among all categories. So, the time using technology devices is not the main factor that causes obesity.

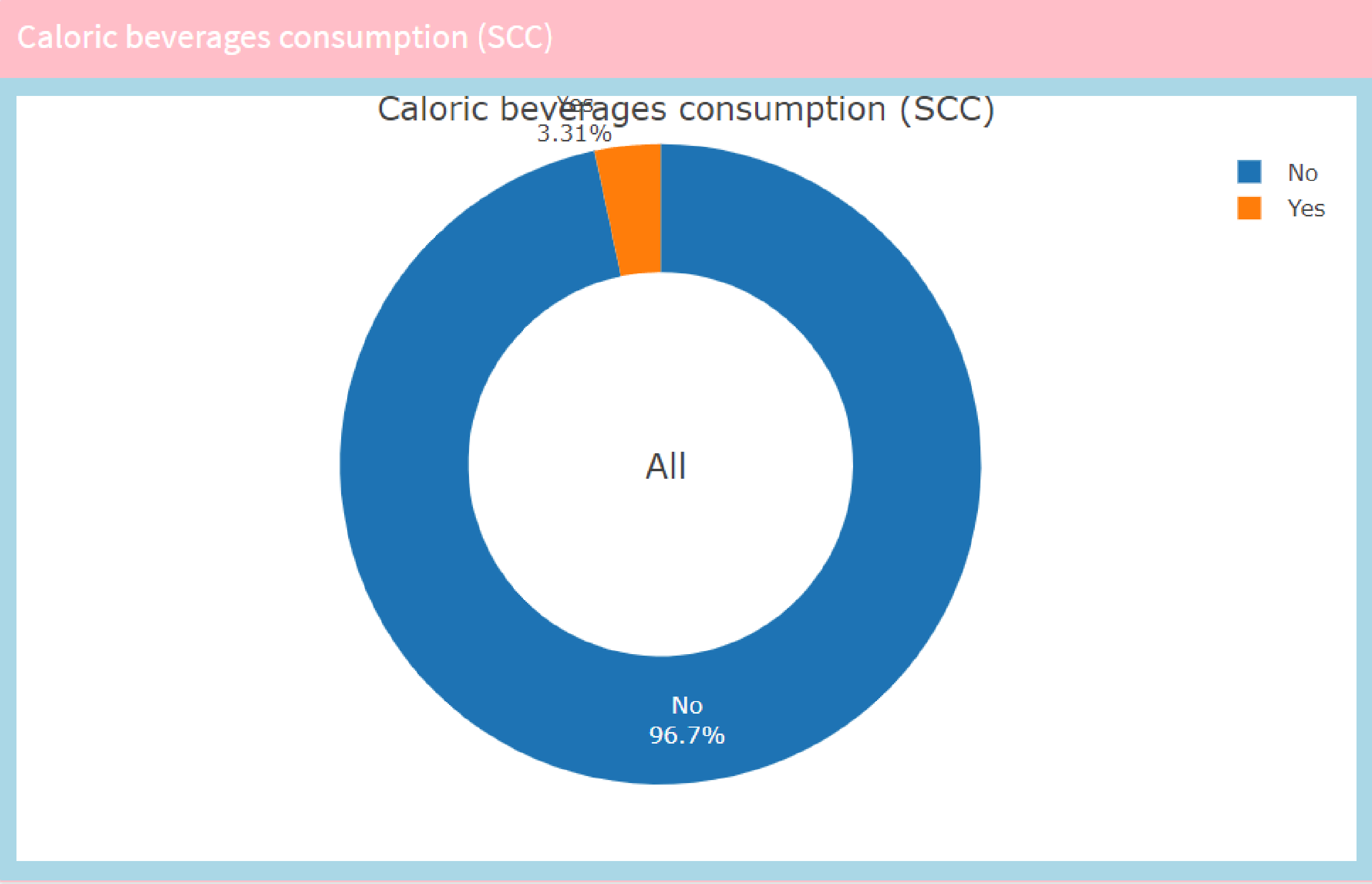
5.4 Dashboard 4: Eating Habit

i) Frequent Consumption of High-Caloric Food by Obesity Level



Based on the side by side view, it shows how different levels of obesity are correlated with frequent consumption of high-caloric foods. It shows that different weight groups consume high-caloric foods in different ways. Among those who are underweight, 370 do not consume high caloric food, which suggests that they may have bad eating habits or other factors influencing their weight status and a significant number of these individuals (2153) do consume. However, 495 people in the normal weight group do not maintain a balanced diet, 2587 individuals often consume high-caloric foods, showing that not all of them do. The majority of overweight level 1 and 2 individuals (2203 and 1983 individuals) consume high-caloric food regularly, which may contribute to their higher weight, whereas a smaller proportion (224 and 539 individuals) do not consume. These tendencies are comparable. For obesity types 1,2 and 3, the vast majority (2817,3194, and 4045 individuals) do consume, showing a strong relationship between this dietary habit and obesity. Just a small number of people in obesity types 1,2 and 3 (93,54,1 individuals) do not consume high-caloric food. These results show a clear correlation between higher obesity levels and increased consumption of high-caloric foods.

ii) Caloric beverages consumption (SCC)



Based on the donut chart, it represents the proportion of individuals who consume caloric beverages (SCC). The majority, 96.7%, indicated that they do not consume caloric beverages, as depicted by the blue section of the donut chart. In contrast, only a small fraction, 3.3%, reported consuming caloric beverages, as shown by the orange section. This implies that caloric beverage consumption is a common dietary practice among individuals. Given the fact that alcoholic beverages appear to be an important component in many diets and may account for a sizable portion of daily calorie intake. This broad use may significantly impact public health, especially in light of initiatives to avoid obesity and manage weight. The limited number of people who do not regularly consume caloric beverages shows that only a small portion of the population avoids these kinds of drinks, which may be due to an intentional decision to limit calorie intake from beverages.

6.0 CONCLUSION

The ‘Predictive Modelling of Obesity Risk Factors’ developed by R Shiny dashboard offers valuable insights that improve user experience and create entrepreneurial opportunities in the health sector.

The purpose of the project is to use data analytics to predict and reduce an individual’s chance of becoming fat, therefore addressing the rising global concern about obesity. It aims to determine the relative contribution of several genetic, behavioural, and environmental risk factors to the overall prediction of obesity by analysing a large dataset. The objective is to develop personalized strategy for obesity prevention and management based on individual risk profiles.

Specifically chosen for health and lifestyle research, the dataset has 20, 759 rows with 18 key attributes, offering a strong basis for analysis. As part of the data preparation process, columns were renamed to improve readability, data types were converted as needed, numerical values were rounded for easier reading, and qualitative data was consistently and meaningfully categorized.

By applying data science, public health, and behavioral research allows for the creation of advanced predictive models that can identify trends and connections beyond conventional epidemiology. This project aligns with Sustainable Development Goal 3, promoting healthy lifestyles and well-being for individuals of all ages globally.

By applying predictive modeling techniques, this project has the potential to significantly improve public health strategies, identify high-risk individuals, and guide targeted treatments. The insights gained from this project can contribute to reducing obesity-related health issues and advancing global health initiatives.

6.1 Limitation

While the predictive modeling of obesity risk factors provides valuable insights, several limitations should be acknowledged:

1. Limited Dataset:

   The results and forecasts of the study are based on a dataset mostly composed of information from a young age. The whole population's obesity risk variables, especially for older persons, might not be included in this age-specific dataset. Due to older age groups being underrepresented in the study, important information about obesity in older age may be missed. Future research initiatives can think about expanding the dataset to incorporate a better representation of different age groups in order to get around this restriction.

2. Limited Time:

   The project does not fully investigate the relationship of age and BMI because of the limitation of time. By not examining how obesity risk factors vary over different age categories, the study may miss important temporal trends and patterns. Future research should examine the relationship of age and BMI in deeper detail, examining how obesity risk factors change as people age to give a more complex picture of the problem.

In summary, while the current study offers valuable insights and findings, it is important to acknowledge the limitations related to the dataset, and temporal limitations. Overcoming these limitations in future projects through expanded datasets, and comprehensive temporal analysis will contribute to the development of more reliable and efficient predictive models for obesity risk factors.

7.0 REFERENCES

Barplot in R shiny. (2022, April 8). Posit Community. https://forum.posit.co/t/barplot-in-r-shiny/133885

Creating a histogram of a selectInput function in R shiny. (n.d.). Stack Overflow. https://stackoverflow.com/questions/74032308/creating-a-histogram-of-a-selectinput-function-in-r-shiny

Efran. (n.d.). Shiny-Boxplot-Application/app.R at master · efran/Shiny-Boxplot-Application. GitHub. https://github.com/efran/Shiny-Boxplot-Application/blob/master/app.R

Finnstats. (2021, September 21). Side-by-Side plots with ggplot2 | R-bloggers. R-bloggers. https://www.r-bloggers.com/2021/09/side-by-side-plots-with-ggplot2/

how to use plotly on donut chart in R Shiny App. (n.d.). Stack Overflow. https://stackoverflow.com/questions/68036881/how-to-use-plotly-on-donut-chart-in-r-shiny-app

Scatterplot in Shiny. (n.d.). Stack Overflow. https://stackoverflow.com/questions/54944804/scatterplot-in-shiny

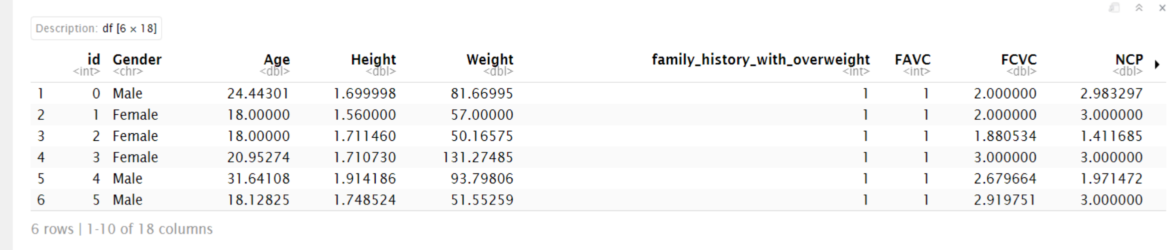
Shiny dashboard. (n.d.). https://rstudio.github.io/shinydashboard/

8.0 APPENDIX

1. Link for dataset, presentation slide, presentation video and r coding: [SD22003 - Google Drive]

2. Cleaning process

```
#read file
```

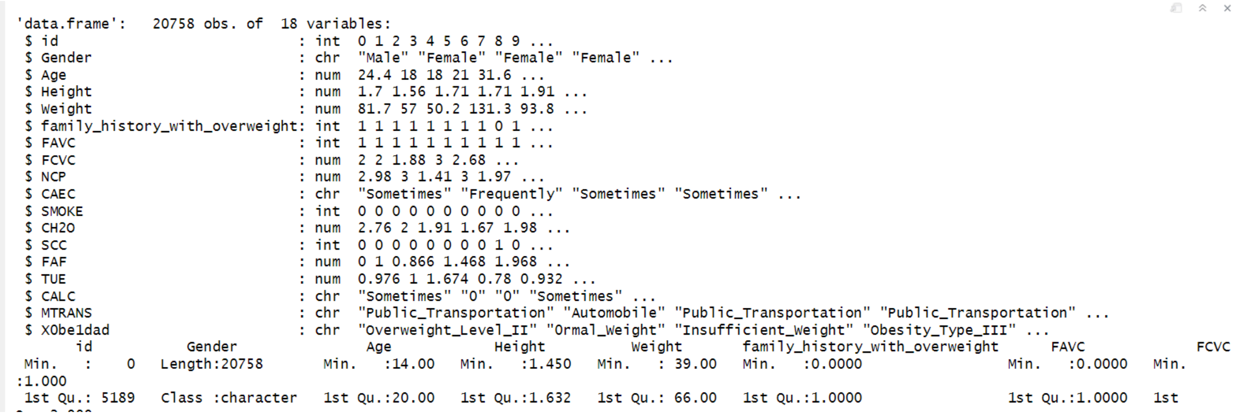


Description: df [6 × 18]

	id <int>	Gender <chr>	Age <dbl>	Height <dbl>	Weight <dbl>	family_history_with_overweight <int>	FAVC <int>	FCVC <dbl>	NCP <dbl>
1	0	Male	24.44301	1.699998	81.66995	1	1	2.000000	2.983297
2	1	Female	18.00000	1.560000	57.00000	1	1	2.000000	3.000000
3	2	Female	18.00000	1.711460	50.16575	1	1	1.880534	1.411685
4	3	Female	20.95274	1.710730	131.27485	1	1	3.000000	3.000000
5	4	Male	31.64108	1.914186	93.79806	1	1	2.679664	1.971472
6	5	Male	18.12825	1.748524	51.55259	1	1	2.919751	3.000000

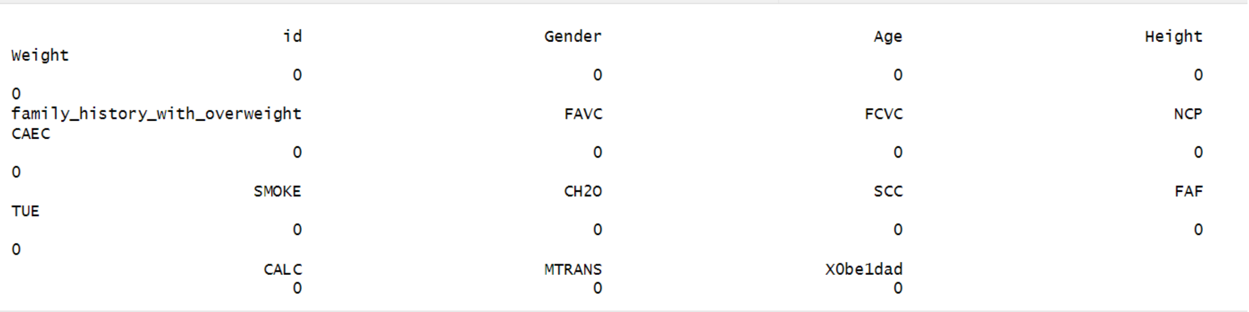
6 rows | 1-10 of 18 columns

```
#structure and summary of the data
```



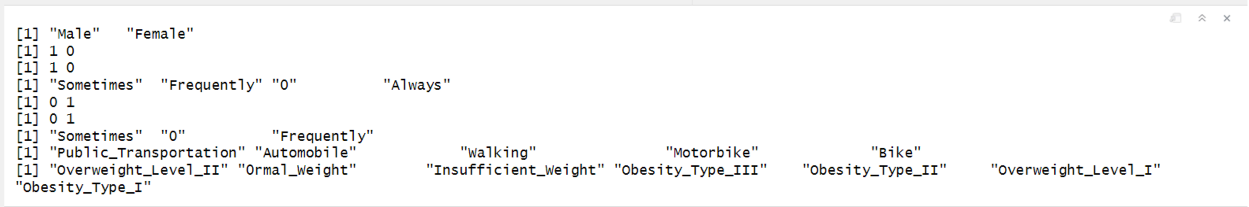
```
'data.frame':	20758 obs. of  18 variables:
 $ id                            : int  0 1 2 3 4 5 6 7 8 9 ...
 $ Gender                        : chr  "Male" "Female" "Female" "Female" ...
 $ Age                           : num  24.4 18 18 21 31.6 ...
 $ Height                        : num  1.7 1.56 1.71 1.71 1.91 ...
 $ Weight                        : num  81.7 57 50.2 131.3 93.8 ...
 $ family_history_with_overweight: int  1 1 1 1 1 1 1 1 0 1 ...
 $ FAVC                          : int  1 1 1 1 1 1 1 1 1 1 ...
 $ FCVC                          : num  2 2 1.88 3 2.68 ...
 $ NCP                           : num  2.98 3 1.41 3 1.97 ...
 $ CAEC                          : chr  "Sometimes" "Frequently" "Sometimes" "Sometimes" ...
 $ SMOKE                         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ CH2O                          : num  2.76 2 1.91 1.67 1.98 ...
 $ SCC                           : int  0 0 0 0 0 0 0 0 1 0 ...
 $ FAF                           : num  0 1 0.866 1.468 1.968 ...
 $ TUE                           : num  0.976 1 1.674 0.78 0.932 ...
 $ CALC                          : chr  "Sometimes" "0" "0" "Sometimes" ...
 $ MTRANS                        : chr  "Public_Transportation" "Automobile" "Public_Transportation" "Public_Transportation" ...
 $ X0be1dad                      : chr  "Overweight_Level_II" "Ormal_Weight" "Insufficient_Weight" "Obesity_Type_III" ...
       id             Gender                Age             Height          Weight       family_history_with_overweight      FAVC               FCVC
 Min.   :    0   Length:20758       Min.   :14.00   Min.   :1.450   Min.   : 39.00   Min.   :0.0000                 Min.   :0.0000   Min.   
:1.000
 1st Qu.: 5189   Class :character   1st Qu.:20.00   1st Qu.:1.632   1st Qu.: 66.00   1st Qu.:1.0000                 1st Qu.:1.0000   1st
```

```
#check null for each column
```



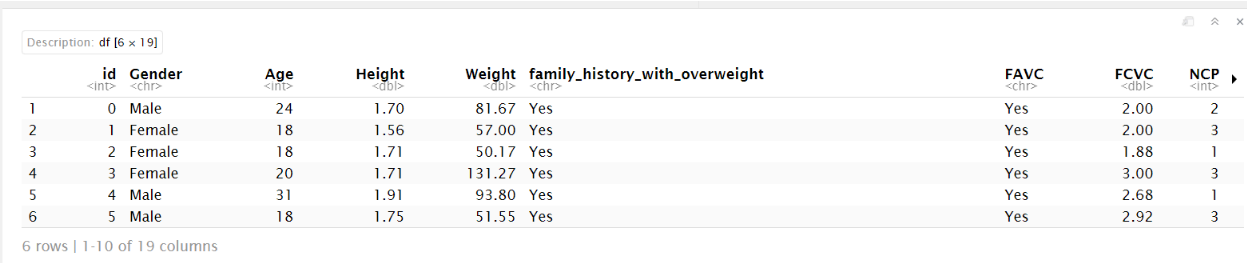
```
                            id                    Gender                       Age                    Height 
Weight 
                             0                         0                         0                         0 
0 
family_history_with_overweight                      FAVC                      FCVC                       NCP 
CAEC 
                             0                         0                         0                         0 
0 
                         SMOKE                      CH2O                       SCC                       FAF 
TUE 
                             0                         0                         0                         0 
0 
                          CALC                    MTRANS                  X0be1dad 
                             0                         0                         0 
```

```
#check distinct value for Gender, family_history_with_overweight, FAVC, CAEC, SMOKE, SCC, CALC, MTRANS, X0be1dad column
```



```
[1] "Male"   "Female"
[1] 1 0
[1] 1 0
[1] "Sometimes"  "Frequently" "0"          "Always"    
[1] 0 1
[1] 0 1
[1] "Sometimes"  "0"          "Frequently"
[1] "Public_Transportation" "Automobile"            "Walking"               "Motorbike"             "Bike"                 
[1] "Overweight_Level_II" "Ormal_Weight"        "Insufficient_Weight" "Obesity_Type_III"    "Obesity_Type_II"     "Overweight_Level_I" 
"Obesity_Type_I"    
```

```
#recheck the data
```



Description: df [6 × 19]

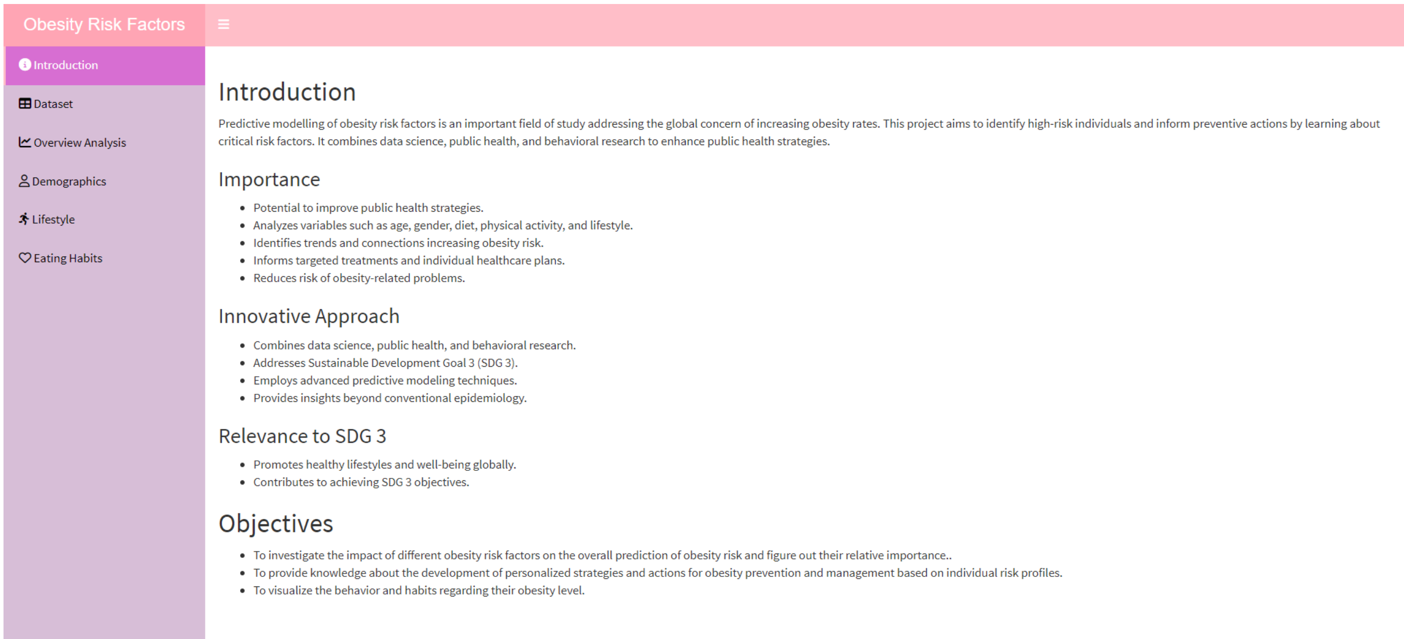
	id <int>	Gender <chr>	Age <int>	Height <dbl>	Weight <dbl>	family_history_with_overweight <chr>	FAVC <chr>	FCVC <dbl>	NCP <int>
1	0	Male	24	1.70	81.67	Yes	Yes	2.00	2
2	1	Female	18	1.56	57.00	Yes	Yes	2.00	3
3	2	Female	18	1.71	50.17	Yes	Yes	1.88	1
4	3	Female	20	1.71	131.27	Yes	Yes	3.00	3
5	4	Male	31	1.91	93.80	Yes	Yes	2.68	1
6	5	Male	18	1.75	51.55	Yes	Yes	2.92	3

6 rows | 1-10 of 19 columns

```
#upload the new dataframe
#get the path new dataframe
```

```
[1] "C:/Users/nurra/OneDrive - ump.edu.my/UMP/SEM 4/DSP 2/PROJ ASSIGMNT/DATA"
```

3. Dashboard



Introduction

Predictive modelling of obesity risk factors is an important field of study addressing the global concern of increasing obesity rates. This project aims to identify high-risk individuals and inform preventive actions by learning about critical risk factors. It combines data science, public health, and behavioral research to enhance public health strategies.

Importance

- Potential to improve public health strategies.
- Analyzes variables such as age, gender, diet, physical activity, and lifestyle.
- Identifies trends and connections increasing obesity risk.
- Informs targeted treatments and individual healthcare plans.
- Reduces risk of obesity-related problems.

Innovative Approach

- Combines data science, public health, and behavioral research.
- Addresses Sustainable Development Goal 3 (SDG 3).
- Employs advanced predictive modeling techniques.
- Provides insights beyond conventional epidemiology.

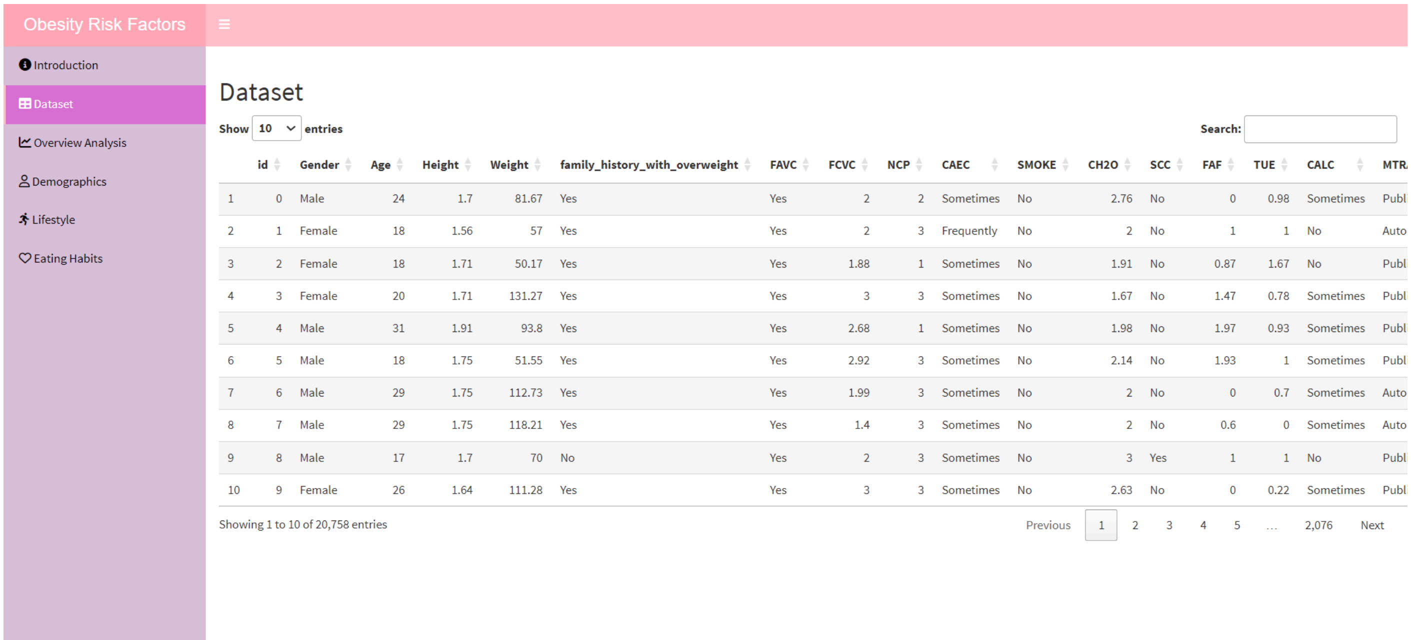
Relevance to SDG 3

- Promotes healthy lifestyles and well-being globally.
- Contributes to achieving SDG 3 objectives.

Objectives

- To investigate the impact of different obesity risk factors on the overall prediction of obesity risk and figure out their relative importance..
- To provide knowledge about the development of personalized strategies and actions for obesity prevention and management based on individual risk profiles.
- To visualize the behavior and habits regarding their obesity level.

Dashboard 1: Introduction



Dataset

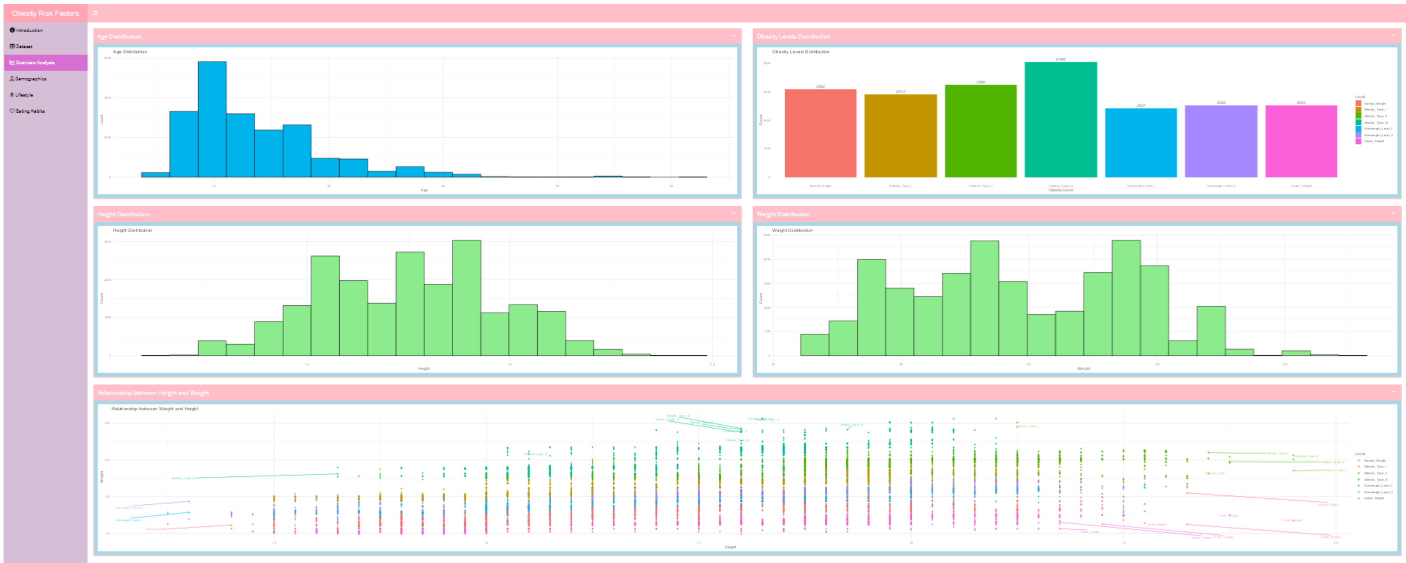
Show 10 entries

Search:

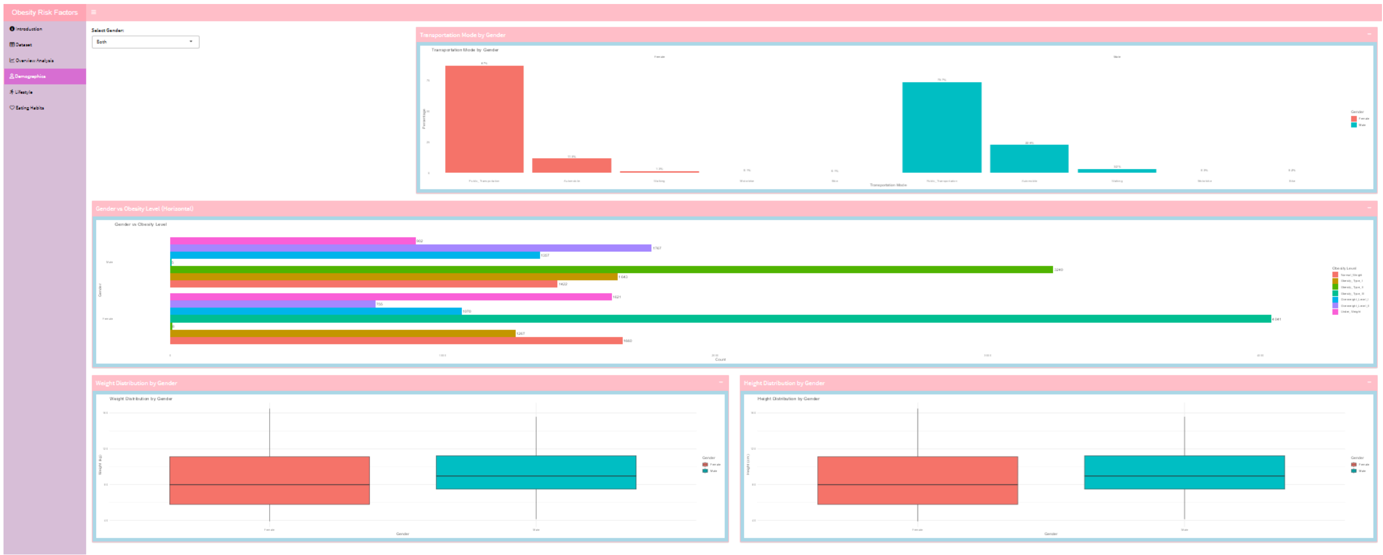
	id	Gender	Age	Height	Weight	family_history_with_overweight	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	FAF	TUE	CALC	MTR/
1	0	Male	24	1.7	81.67	Yes	Yes	2	2	Sometimes	No	2.76	No	0	0.98	Sometimes	Publ
2	1	Female	18	1.56	57	Yes	Yes	2	3	Frequently	No	2	No	1	1	No	Auto
3	2	Female	18	1.71	50.17	Yes	Yes	1.88	1	Sometimes	No	1.91	No	0.87	1.67	No	Publ
4	3	Female	20	1.71	131.27	Yes	Yes	3	3	Sometimes	No	1.67	No	1.47	0.78	Sometimes	Publ
5	4	Male	31	1.91	93.8	Yes	Yes	2.68	1	Sometimes	No	1.98	No	1.97	0.93	Sometimes	Publ
6	5	Male	18	1.75	51.55	Yes	Yes	2.92	3	Sometimes	No	2.14	No	1.93	1	Sometimes	Publ
7	6	Male	29	1.75	112.73	Yes	Yes	1.99	3	Sometimes	No	2	No	0	0.7	Sometimes	Auto
8	7	Male	29	1.75	118.21	Yes	Yes	1.4	3	Sometimes	No	2	No	0.6	0	Sometimes	Auto
9	8	Male	17	1.7	70	No	Yes	2	3	Sometimes	No	3	Yes	1	1	No	Publ
10	9	Female	26	1.64	111.28	Yes	Yes	3	3	Sometimes	No	2.63	No	0	0.22	Sometimes	Publ

Showing 1 to 10 of 20,758 entries

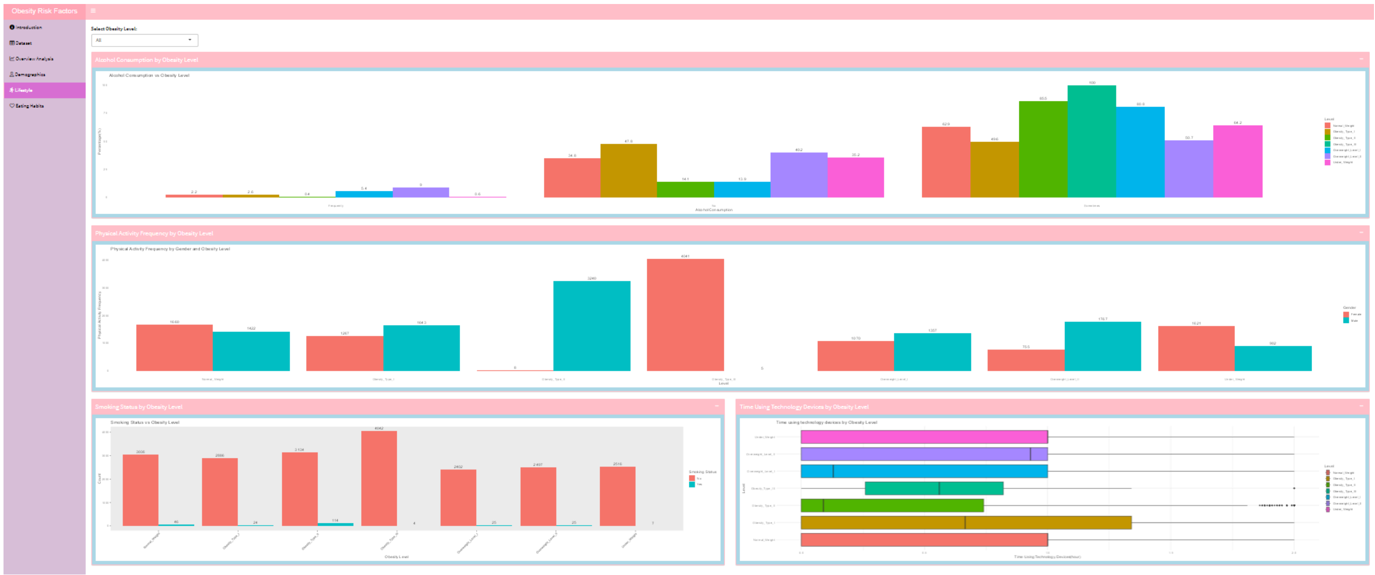
Dashboard 2: Dataset



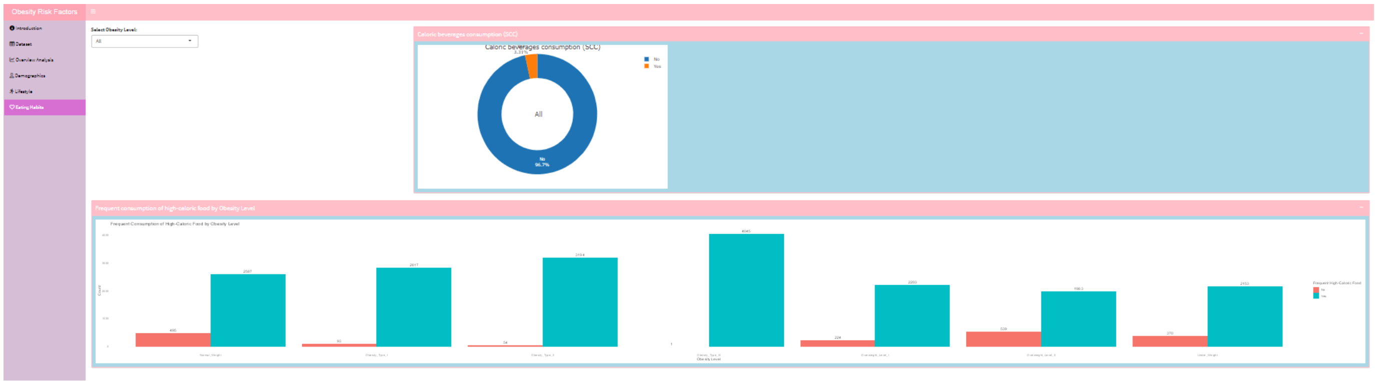
Dashboard 3: Overview



Dashboard 4: Demographic


Obesity Risk Factors
Introduction
Dataset
Overview Analysis
Demographics
Lifestyle
Eating Habits
Select Obesity Level:
All
Alcohol Consumption by Obesity Level
Alcohol Consumption vs Obesity Level
Percentage (%)
2.2
2.6
0.4
5.4
9
0.6
34.8
47.8
14.1
Frequently
Alcohol Consumption
Level
Physical Activity Frequency by Obesity Level
Physical Activity Frequency by Gender and Obesity Level
Physical Activity Frequency
Level
Gender
Smoking Status by Obesity Level
Smoking Status vs Obesity Level
Count
Obesity Level
Smoking Status
Time Using Technology Devices by Obesity Level
Time using technology devices by Obesity Level
Level
Time Using Technology Devices(hours)

Dashboard 5: Lifestyle


Obesity Risk Factors
Introduction
Dataset
Overview Analysis
Demographics
Lifestyle
Eating Habits
Select Obesity Level:
All
Caloric beverages consumption (SCC)
Caloric beverages consumption (SCC)
3.31%
All
No
96.7%
No
Yes
Frequent consumption of high-caloric food by Obesity Level
Frequent Consumption of High-Caloric Food by Obesity Level
Count
Obesity Level
Frequent High-Calorie Food

Dashboard 6: Eating Habit

GROUP PROJECT PLAN & APPROVAL

Note: Print this form and obtain your lecturer’s approval before starting the project’s tasks.

SECTION NO.	01G/ 02G/ 03G/ 04G	
GROUP NAME	THE WELLNESS WARRIORS	
GROUP MEMBERS (Leader’s name is the first in the list)	**Id. No.**	**Name**
	SD22003	NUR ATIEKA RAFIEKAH BINTI RAZAK
	SD22034	MUHAMMAD AIMAN IRFAN BIN AHMAD MUHYE
	SD22056	ELIANE HO WAN WEN
	SD22017	BATRISYIA BINTI ISMAIL
PROJECT TITLE	PREDICTIVE MODELLING OF OBESITY RISK FACTOR	
PROJECT DESCRIPTION	We would like to focus on health and lifestyle study therefore this project aim to analyze which crucial factor contributes to obesity risk. Also our study is to identify who are at high risk obesity of developing obesity based on factors.	
DATA DESCRIPTION	The dataset provide comprehensive information on individual, encompassing key attributes for both qualitative and quantitative variable. the qualitative variables are gender, consumption of food and alcohol, transportation mode and corresponding obesity level. for quantitative variables are age, height, weight, family history, frequent consumption of high caloric food, vegetable, daily water, alcohol, physical activities, and time spend using tech devices	
APPROVED BY (signature)	Associate Professor Dr. Roslinazairimah Zakaria	
DATE	3/4/2024	

Marking Sheet (For Lecturer's use only)

UNIVERSITI MALAYSIA PAHANG PUSAT SAINS MATEMATIK	**DATA SCIENCE PROGRAMMING II (BSD2223)**		**MARKS:** **120 (30%)** /120
	GROUP LEADER: NUR ATIEKA RAFIEKAH BINTI RAZAK	**ID NO :** SD22003	
	GROUP PROJECT	**SECTION NO: 01G/02G/03G/04G** **DUE DATE: 19/6/2023**	

RUBRICS FOR CLO2/PLO2

CLO2: Analyse and summarise data using appropriate programming tools	PLO2: Cognitive Skills and Functional work skills with focus on Numeracy skills. **C4: Analysis**	/40	/10

Criteria	Achievement Level						Weightage	Score
	0	1	2	3	4	5		
	Incompetent	Inadequate	Emerging	Developing	Good	Excellent		
Ability to **obtain appropriate data** for the project	Unable to do the task.	Limited ability to do the task.	Reasonable ability to do the task.	Able to do the task with effort.	Good effort to do the task .	Able to do the task efficiently.	2	10
Ability to **summarise data numerically and graphically using R**.	Unable to do the task.	Limited ability to do the task.	Able to do the task with errors.	Able to do the task with no errors but wrong answer.	Able to do the task with no errors.	Able to do the task efficiently and correctly with no errors.	2	10
Ability to **analyse data using R and obtain data insights**.	Unable to do the task.	Limited ability to do the task.	Able to do the task with errors.	Able to do the task with no errors but wrong answer.	Able to do the task with no errors.	Able to do the task efficiently and correctly with no errors.	2	10
Ability to **provide conclusion** and recommendation from the project.	Unable to do the task.	Limited ability to do the task.	Reasonable ability to do the task.	Able to do the task with effort.	Good effort to do the task .	Able to do the task efficiently.	2	10

RUBRICS FOR CLO3/PLO3

CLO3: Develop programming codes to solve problems	PLO3: Functional work skills with focus on Practical, and Digital skills. **P4: Mechanism**	/40	/10

Criteria	Achievement Level						Weightage	Score
	0	1	2	3	4	5		
	Incompetent	Inadequate	Emerging	Developing	Good	Excellent		
Ability to **construct R codes to summarise data numerically and graphically**.	Unable to do the task.	Limited ability to do the task.	Reasonable ability to do the task.	Able to do the task with effort.	Good effort to do the task .	Able to do the task efficiently.	2	10
Ability to **construct R codes for data analysis**.	Unable to do the task.	Limited ability to do the task.	Able to do the task with errors.	Able to do the task with no errors but wrong answer.	Able to do the task with no errors.	Able to do the task efficiently and correctly with no errors.	2	10
Ability to **develop a dashboard (GUI) to present the data using Rshiny.**	Unable to do the task.	Limited ability to do the task.	Reasonable ability to do the task.	Able to do the task with effort.	Good effort to do the task .	Able to do the task efficiently.	4	20

RUBRICS FOR CLO4/PLO5

CLO4: Demonstrate verbal and written communication skills	PLO5:Functional work skills with focus on communication skills. **A3: Valuing**	/20	/5

	Criteria	Achievement Level						Weightage	Score
		0	1	2	3	4	5		
		Incompetent	Inadequate	Emerging	Developing	Good	Excellent		
Written	Ability to write the report findings coherently.	Unable to do the task.	Limited ability to do the task.	Reasonable ability to do the task.	Able to do the task with effort.	Good effort to do the task .	Able to do the task efficiently.	1	5
	Ability to present the project report in the given format which include data description, data analysis, results and discussion.	Unable to do the task.	Limited ability to do the task.	Reasonable ability to do the task.	Able to do the task with effort.	Good effort to do the task .	Able to do the task efficiently.	1	5
Communication	Ability to present the project proficiently by organizing and communicating the results in a clear, logical, and easy-to-follow manner.	Unable to do the task.	Limited ability to do the task.	Reasonable ability to do the task.	Able to do the task with effort.	Good effort to do the task .	Able to do the task efficiently.	1	5
	Ability to deliver the dashboard to summarise the project findings.	Unable to do the task.	Limited ability to do the task.	Reasonable ability to do the task.	Able to do the task with effort.	Good effort to do the task .	Able to do the task efficiently.	1	5

RUBRICS FOR CLO5/PLO8

CLO5: Relate entrepreneur skills in assigned task	PLO8:Entrepreneural skills **A4: Organising values**	/20	/5

Criteria	Achievement Level						Weightage	Score
	0	1	2	3	4	5		
	Incompetent	Inadequate	Emerging	Developing	Good	Excellent		
Ability to articulate the given/ chosen case study related to entrepreneurship.	Unable to articulate the given/ chosen case study.	Able to articulate the given/ chosen case study fairly weak.	Able to articulate the given/ chosen case study fairly well.	Able to articulate the given/ chosen case study well.	Able to articulate the given/ chosen case study reasonably well.	Able to articulate the given/ chosen case study excellently.	2	10
Ability to deliver entrepreneur ideas.	Unable to deliver any entrepreneur idea.	Delivery of idea is unclear, vague and not systematic.	Delivery of idea is less clear, vague and systematic.	Delivery of idea is moderately clear, vague and systematic.	Delivery of idea is clear and systematic.	Delivery of idea is very clear and systematic.	2	10