Albukhary International University

SCHOOL OF COMPUTER SCIENCE

Bachelor of Computer Science (Hons)

ASSIGNMENT

Statistical Programming CCS2233

Name: Thinley Yeshey Choden **Matrix No :AIU22102188**

FOR EXAMINER'S USE ONLY

Total Marks (40 marks)

INSTRUCTIONS

Write a report on the following tasks.

1. Import a dataset from https://www.kaggle.com/datasets with the following characteristics:
   a. The dataset contains within 300 – 500 number of observations
   b. The number of independent variables is five (5) with a mixture of three numeric and two categorical variables.
   c. The dependent variable is in the form of categorical variable with at least three groups.

2. Perform data validation and report your findings.

3. Check your dataset whether there exist:
   a. duplication in the observation. If exists, remove the duplication.
   b. missing values. For numerical variable, replace the missing values using the average value of the variable. While for the categorical variable, replace with the suspected category based on the dataset.
   c. encode the categorical variables accordingly.
   d. outliers. Generate boxplot and check if there exist outliers, report the outliers. Then delete the outliers from your dataset. Generate your new boxplot

4. Perform statistical data analysis for the numerical variables using measure of central tendency (mean, mode, median) and measure of dispersion (variance and standard deviation) according to the subgroups in the dependent variable. Explain your findings.

5. Create suitable visualizations for all the numerical and categorical variables of your dataset considering the dependent variables.

6. Prepare your report using Microsoft words and please include all the R programming codes, outputs and interpretation for each of your results.

Table of Contents

Introduction

This report focuses on analysing obesity levels using a dataset that includes attributes like age, height, weight, gender, and family history of overweight. The goal is to explore patterns and relationships between these variables and obesity levels, using data preprocessing, outlier detection, and statistical analysis to ensure accurate results. The cleaned and encoded dataset will form the basis for further machine learning models aimed at predicting obesity levels and identifying key influencing factors.

Question 1

R Script:

```
> #Question 1
> setwd("C:/Users/USER/OneDrive/Desktop/SP/Assignment")
> dataset <- read.csv("filtered_final_dataset.csv")
```

#Import dataset to working directory.

Question 2 - Data Validation (Summary Statistics and Data Profiling)

R Script:

```
> #Question 2
> head(dataset)
  Age Height Weight Gender Family_history      Obesity_level
1  21   1.62     64 Female            yes      Normal_Weight
2  21   1.52     56 Female            yes      Normal_Weight
3  23   1.80     77   Male            yes      Normal_Weight
4  27   1.80     87   Male             no Overweight_Level_I
5  29   1.62     53   Male             no      Normal_Weight
6  23   1.50     55 Female            yes      Normal_Weight
```

#View the first 6 rows of the dataset.

R Script:

```
> summary(dataset)
      Age            Height          Weight          Gender
 Min.   :14.00   Min.   :1.450   Min.   : 42.30   Length:928
 1st Qu.:19.22   1st Qu.:1.612   1st Qu.: 65.05   Class :character
 Median :21.84   Median :1.676   Median : 76.65   Mode  :character
 Mean   :23.83   Mean   :1.687   Mean   : 77.56
 3rd Qu.:25.00   3rd Qu.:1.760   3rd Qu.: 87.28
 Max.   :61.00   Max.   :1.980   Max.   :125.00
 Family_history     Obesity_level
 Length:928         Length:928
 Class :character   Class :character
 Mode  :character   Mode  :character
```

Interpretation:

The dataset contains 928 records with the following characteristics:

- Age: Ranges from 14 to 61 years, with a median of 21.84 and a mean of 23.83.
- Height: Ranges from 1.45 to 1.98 meters, with a median of 1.676 and a mean of 1.687.
- Weight: Ranges from 42.30 to 125.00 kg, with a median of 76.65 and a mean of 77.56.
- Gender: Recorded as a categorical variable with no further detail provided.
- Family_history: Categorical variable indicating whether there is a family history of overweight.
- Obesity_level: Categorical variable classifying obesity into various levels.

R Script:

```
> library(DataExplorer)
> create_report(dataset)
processing file: report.rmd
output file: C:/Users/USER/OneDrive/Desktop/SP/Assignment/report.knit.md
Output created: report.html
```

Output:

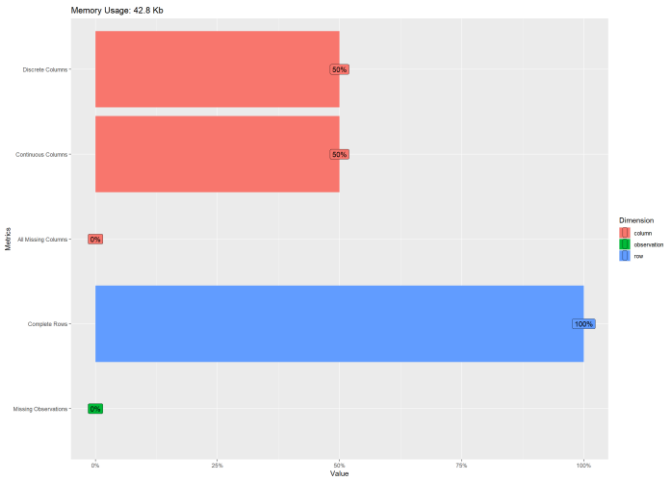


Figure 1. Basic Statistic (Percentage)

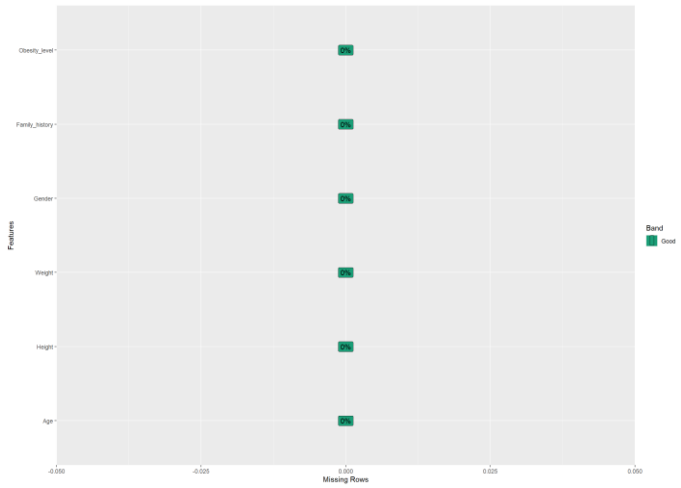


Figure 2. Missing Data Profile

Interpretation:

Figure 1 show the dataset has an equal distribution of discrete and continuous variables; Discrete columns (Gender, Family History, Obesity Level) and Continuous columns (Age, Height, Weight) Figure 2 shows there are no missing data and all 928 rows are complete.

Output:

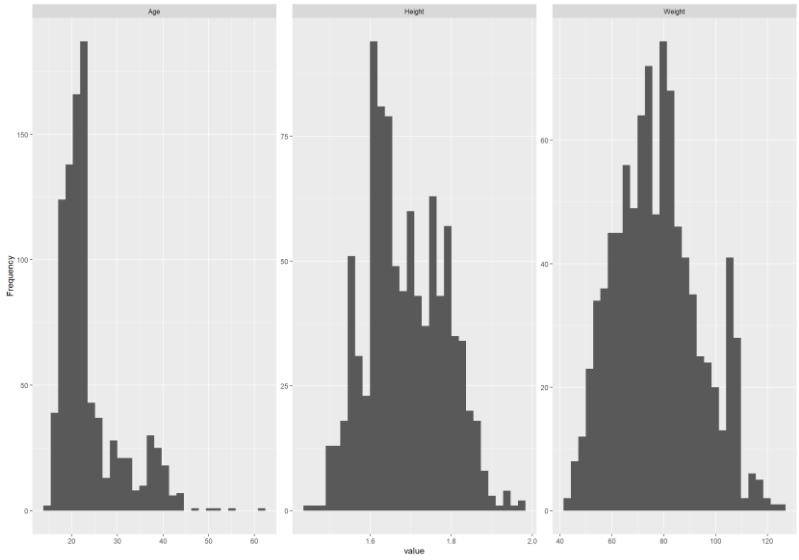


Figure 3. Histogram

Interpretation:

Age:

- The distribution of age is right-skewed, with the majority of individuals aged between 19 and 25. A smaller number of individuals are older, with a maximum age of 61.

Height:

- Heights are normally distributed, centered around the mean of 1.687m, with most individuals falling between 1.6m and 1.8m. The distribution is fairly symmetrical, indicating a typical spread.

Weight:

- The weight distribution is slightly right-skewed, with a higher concentration of individuals weighing between 60kg and 80kg. A few individuals have significantly higher weights, up to 125kg.

Output:

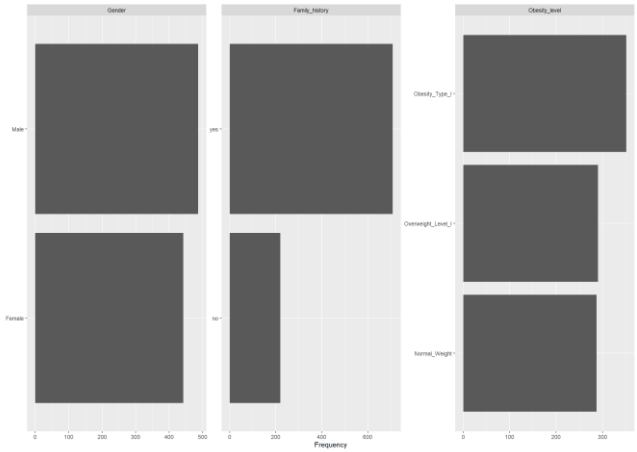


Figure 4. Bar Chart

Interpretation:

Gender:

- The dataset contains a relatively balanced distribution between males and females, with a slightly higher count of Male individuals.

Family History with Overweight:

- A significant proportion of individuals have a family history of overweight, with a noticeable difference between the "Yes" and "No" categories.

Obesity Level:

- The Obesity Type I category has the highest count, with approximately 350 individuals.
- Overweight Level I follows closely, with around 280 individuals.
- Normal Weight is slightly less than Overweight Level I, showing a notable, though smaller, proportion of individuals in the dataset.

Output:

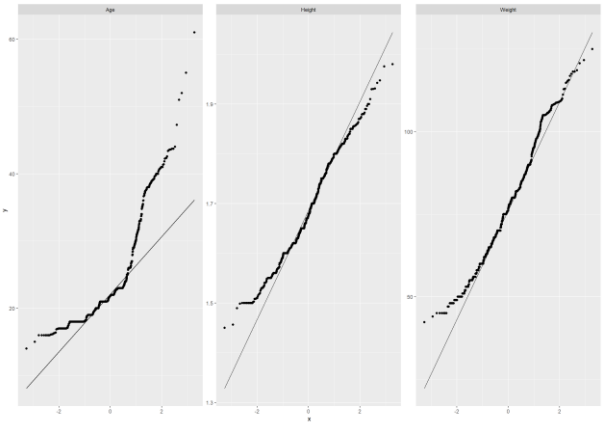


Figure 5. QQ plot

Interpretation:

- Age:
    - The QQ plot for Age shows a deviation from the diagonal line, especially at the tails, indicating that Age is not normally distributed. This aligns with the earlier observation of a right-skewed distribution.
- Height:
    - The QQ plot for Height aligns well with the diagonal line, suggesting that Height follows a nearly normal distribution, which matches the symmetrical shape seen in the histogram.
- Weight:
    - The QQ plot for Weight also aligns well with the diagonal line, again suggesting a nearly normal distribution.

Question 3.a

R Script:

```
> #Question 3
> #a
> #Check for duplicate rows
> duplicate_rows <- dataset[duplicated(dataset), ]
> num_duplicates <- nrow(duplicate_rows)

> #Report the number of duplicate rows
> cat("Total number of duplicate rows:", num_duplicates, "\n")
Total number of duplicate rows: 22

> #View the duplicate rows              ******* 22 duplictaes
>  print(duplicate_rows)
     Age Height Weight Gender Family_history      Obesity_level
81    25   1.57     55 Female             no      Normal_Weight
134   21   1.62     70   Male             no Overweight_Level_I
139   21   1.62     70   Male             no Overweight_Level_I
142   21   1.62     70   Male             no Overweight_Level_I
155   23   1.66     60 Female            yes      Normal_Weight
164   22   1.69     65 Female            yes      Normal_Weight
226   21   1.70     65   Male            yes      Normal_Weight
241   16   1.66     58 Female             no      Normal_Weight
252   38   1.75     75   Male            yes      Normal_Weight
363   18   1.62     55 Female            yes      Normal_Weight
369   22   1.74     75   Male            yes      Normal_Weight
411   21   1.62     70   Male             no Overweight_Level_I
412   21   1.62     70   Male             no Overweight_Level_I
472   21   1.62     70   Male             no Overweight_Level_I
478   21   1.62     70   Male             no Overweight_Level_I
479   21   1.62     70   Male             no Overweight_Level_I
480   21   1.62     70   Male             no Overweight_Level_I
481   21   1.62     70   Male             no Overweight_Level_I
482   21   1.62     70   Male             no Overweight_Level_I
569   21   1.62     70   Male             no Overweight_Level_I
570   21   1.62     70   Male             no Overweight_Level_I
571   21   1.62     70   Male             no Overweight_Level_I

> #Identify and remove actual duplicates for the specific case of the 21-year-old male
> dataset_clean <- dataset[!(duplicated(dataset) &
+                     dataset$Age == 21 &
+                     dataset$Height == 1.62 &
+                     dataset$Weight == 70 &
+                     dataset$Gender == "Male" &
+                     dataset$Family_history == "no" &
+                     dataset$Obesity_level == "Overweight_Level_I"), ]

> #Verify that specific duplicates have been removed
> cat("Number of rows before removing duplicates:", nrow(dataset), "\n")
Number of rows before removing duplicates: 928

> cat("Number of rows after removing duplicates:", nrow(dataset_clean), "\n")
Number of rows after removing duplicates: 914

#Removed all duplicates
```

Question 3.b

R Script:

```
> #Check for missing values in the dataset
> missing_values <- colSums(is.na(dataset_clean))

> #Display the number of missing values for each column
> cat("Missing values in each column:\n")
Missing values in each column:
> print(missing_values)              #******* no missing values
          Age        Height        Weight        Gender Family_history  Obesity_level
            0             0             0             0             0                 0
```

There are no missing values in the dataset

Question 3.c - Encoding Categorical Values using Label encoding.

R Script:

```
> dataset_clean$Gender <- as.factor(dataset_clean$Gender)
> dataset_clean$Family_history <- as.factor(dataset_clean$Family_history)
> dataset_clean$Obesity_level <- as.factor(dataset_clean$Obesity_level)

> # Apply label encoding
> dataset_clean$Gender <- as.numeric(dataset_clean$Gender)
> dataset_clean$Family_history <- as.numeric(dataset_clean$Family_history)
> dataset_clean$Obesity_level <- as.numeric(dataset_clean$Obesity_level)

> # View the first few rows of the label-encoded dataset
> head(dataset_clean)
  Age Height Weight Gender Family_history Obesity_level
1  21   1.62     64      1              2             1
2  21   1.52     56      1              2             1
3  23   1.80     77      2              2             1
4  27   1.80     87      2              1             3
5  29   1.62     53      2              1             1
6  23   1.50     55      1              2             1
> dim(dataset_clean)
[1] 914   6
```

Interpretation:

The original categorical values ("Female", "Male") were converted into numeric values, with 1 representing "Female" and 2 representing "Male".

Family History with Overweight: This variable originally had values "yes" and "no", which were encoded as 2 ("yes") and 1 ("no").

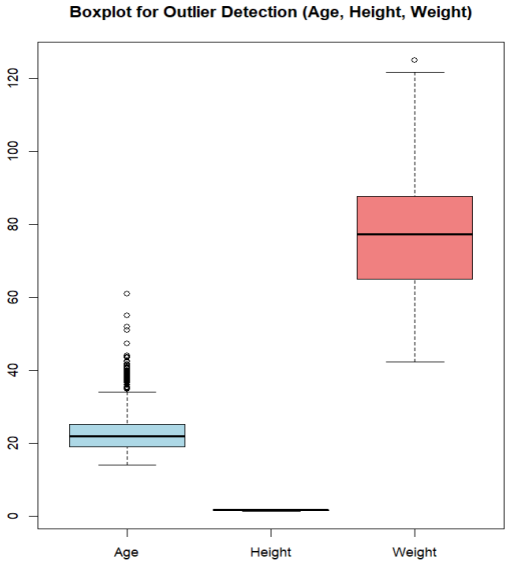
Obesity Level: This categorical variable had multiple levels (“Normal_Weight”, “Overweight_Level_I” and “Obesity_Type_I”), and was also label-encoded into numeric values. Each level of obesity was assigned a unique integer value, with 1 representing “Normal_Weight”, 2 representing “Overweight_Level_I”, and 3 representing “Obesity_Type_I”.

Question 3.d - Outlier detection and Removal

R Script:

```
> #d
> #Generate boxplot for detecting outliers for Age, Height, and Weight together
> boxplot(dataset_clean[, c("Age", "Height", "Weight")],
+         main="Boxplot for Outlier Detection (Age, Height, Weight)",
+         col=c("lightblue", "lightgreen", "lightcoral"),
+         names=c("Age", "Height", "Weight"))
```

Output:



Interpretation:

There are a lot of outliers in Age attribute and 1 outlier in the weight attribute.

Need to be removed.

Figure 6. Box plot for numerical Variables.

R Script:

```
> # Identify outliers for Age
> outliers_age <- boxplot.stats(dataset_clean$Age)$out
> num_outliers_age <- length(outliers_age)
> print(paste("Number of outliers in Age:", num_outliers_age))
[1] "Number of outliers in Age: 102"

> # Identify outliers for Weight
> outliers_weight <- boxplot.stats(dataset_clean$Weight)$out
> num_outliers_weight <- length(outliers_weight)
> print(paste("Number of outliers in Weight:", num_outliers_weight))
[1] "Number of outliers in Weight: 1"

# There are 102 outliers in the age attribute and 1 outlier in the weight attribute.
```

R Script:

```
> # Define the function to identify valid (non-outliers) using the IQR method
> remove_outliers <- function(x) {
+   Q1 <- quantile(x, 0.25)
```

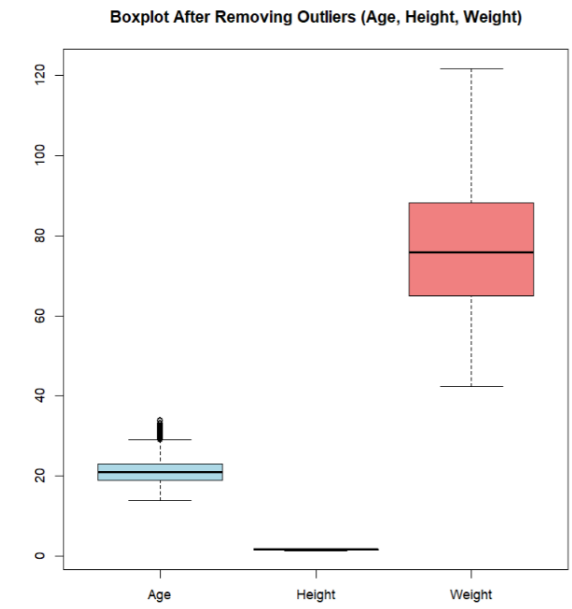
```
+   Q3 <- quantile(x, 0.75)
+   IQR <- Q3 - Q1
+   lower_bound <- Q1 - 1.5 * IQR
+   upper_bound <- Q3 + 1.5 * IQR
+   return(x >= lower_bound & x <= upper_bound)
+ }

> # Apply the function to the Age and Weight columns
> valid_age <- remove_outliers(dataset_clean$Age)
> valid_weight <- remove_outliers(dataset_clean$Weight)

> # Combine valid indices to ensure consistency (both columns should have valid values)
> valid_data <- valid_age & valid_weight
> # Filter the dataset to remove outliers from Age and Weight
> dataset_no_outliers <- dataset_clean[valid_data, ]

> # Check the dimensions of the cleaned data frame (no outliers in Age and Weight)
> print(dim(dataset_no_outliers))
[1] 811   6
```

#Outliers removed.

R Script:

```
> # Generate a boxplot again after removing outliers
> boxplot(dataset_no_outliers[, c("Age", "Height", "Weight")],
+         main="Boxplot After Removing Outliers (Age, Height, Weight)",
+         col=c("lightblue", "lightgreen", "lightcoral"),
+         names=c("Age", "Height", "Weight"))
```

Output:



Interpretation: The outlier from the weight attribute has been removed but there are still some outliers left in the age attribute.

Justification:
I opted not to aggressively remove all outliers to preserve the integrity of the dataset, as these values could represent genuine observations.

Figure 7. Boxplot after removing outliers.

R Script:

```
> # Density for Age
> p5 <- ggplot(dataset_clean, aes(x=Age)) +
+   geom_density(fill="lightblue") +
+   ggtitle("Age Density (Before Outlier Removal)") + theme_minimal()
>
> p6 <- ggplot(dataset_no_outliers, aes(x=Age)) +
+   geom_density(fill="lightblue") +
+   ggtitle("Age Density (After Outlier Removal)") + theme_minimal()
>
> # Density for Weight
> p7 <- ggplot(dataset_clean, aes(x=Weight)) +
+   geom_density(fill="lightcoral") +
+   ggtitle("Weight Density (Before Outlier Removal)") + theme_minimal()
>
> p8 <- ggplot(dataset_no_outliers, aes(x=Weight)) +
+   geom_density(fill="lightcoral") +
+   ggtitle("Weight Density (After Outlier Removal)") + theme_minimal()
>
> # Side-by-side display
> (p5 + p6) / (p7 + p8)
```

Output:

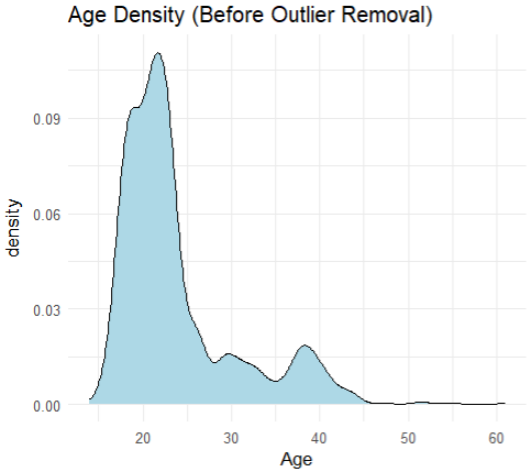


Figure 8. Density Plot for age before outlier removal.

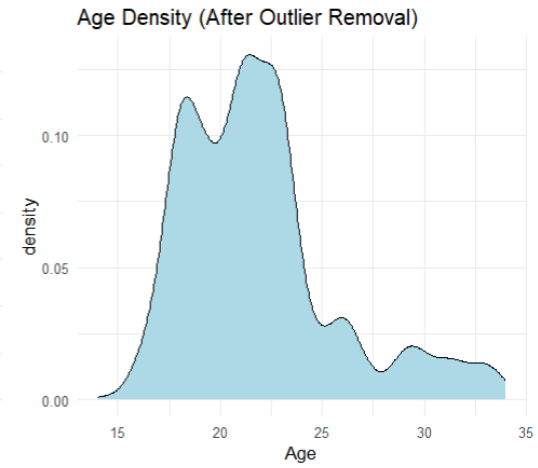


Figure 9. Density Plot for age after outlier removal.

Interpretation:

Figure 9 is less skewed as compared to figure 8. This is due to the removal of outliers from the age attribute. However it still remains slightly skewed due to the few outliers left.

Figure 10. Density Plot for weight before outlier removal.

Figure 11. Density Plot for weight after outlier removal.

Interpretation:

Figure 11 is much more normally distributed as compared to figure 10 due to the removal of the 1 outlier.

Question4- Descriptive statistics

R Script:

```
> age_stats <- dataset_no_outliers %>%
+   group_by(Obesity_level) %>%
+   summarise(
+     Age_Mean = mean(Age, na.rm = TRUE),
+     Age_Median = median(Age, na.rm = TRUE),
+     Age_Mode = as.numeric(names(sort(table(Age), decreasing=TRUE))[1]),
+     Age_SD = sd(Age, na.rm = TRUE),
+     Age_Variance = var(Age, na.rm = TRUE)
+   )
> print(age_stats)  # Descriptive statistics for Height by Obesity Level
# A tibble: 3 × 6
  Obesity_level Age_Mean Age_Median Age_Mode Age_SD Age_Variance
          <dbl>    <dbl>      <dbl>    <dbl>  <dbl>        <dbl>
1             1     21.2         21       21   3.55         12.6
2             2     22.4       22.4       18   3.73         13.9
3             3     22.1       21.0       21   4.17         17.4
```

Interpretation:

- Mean Age: The average age increases slightly with the obesity level. Obesity Level 2 has the highest mean age (22.4 years), while Level 1 has the lowest (21.2 years).

- Median Age: The median age is close to the mean age for Levels 1 and 3, but for Level 2, the median is slightly higher at 22.4 years. This suggests that in Level 2, ages are more symmetrically distributed around the median.
- Mode: Age 21 is the most frequent for Levels 1 and 3, while 18 is the mode for Level 2. This indicates that 21 is a common age for Levels 1 and 3, and 18 is particularly common for Level 2.
- Standard Deviation (SD): The variability in age is higher in Level 3 (SD = 4.17), indicating more diversity in ages. Levels 1 and 2 have similar but slightly lower SDs.
- Variance: The variance is consistent with the SD, with Level 3 showing the highest variance, reflecting greater spread in age.

```
> height_stats <- dataset_no_outliers %>%
+   group_by(Obesity_level) %>%
+   summarise(
+     Height_Mean = mean(Height, na.rm = TRUE),
+     Height_Median = median(Height, na.rm = TRUE),
+     Height_Mode = as.numeric(names(sort(table(Height), decreasing=TRUE))[1]),
+     Height_SD = sd(Height, na.rm = TRUE),
+     Height_Variance = var(Height, na.rm = TRUE)
+   )
> print(height_stats)      # Descriptive statistics for Weight by Obesity Lev
el
# A tibble: 3 × 6
  Obesity_level Height_Mean Height_Median Height_Mode Height_SD Height_Variance
          <dbl>       <dbl>         <dbl>       <dbl>     <dbl>           <dbl>
1             1        1.68          1.66         1.6    0.0946         0.00895
2             2        1.71          1.70         1.65   0.0870         0.00756
3             3        1.70          1.70         1.6    0.0951         0.00904
```

Interpretation:

height_stats

- Mean Height: The average height is slightly higher for Level 2 (1.71 meters) compared to Levels 1 and 3 (1.68 and 1.70 meters, respectively).
- Median Height: The median height for Levels 2 and 3 is the same (1.70 meters), indicating central tendency is similar. Level 1 has a slightly lower median.
- Mode: Height mode is 1.6 meters for Levels 1 and 3, and 1.65 meters for Level 2. This reflects common heights for each level.
- Standard Deviation (SD): The SD is quite similar across levels, with Level 3 having the highest SD, indicating slightly more variation in height.
- Variance: Variance aligns with SD, with Level 3 showing more spread in height data.

```
> weight_stats <- dataset_no_outliers %>%
+   group_by(Obesity_level) %>%
+   summarise(
+     Weight_Mean = mean(Weight, na.rm = TRUE),
+     Weight_Median = median(Weight, na.rm = TRUE),
+     Weight_Mode = as.numeric(names(sort(table(Weight), decreasing=TRUE))[1]),
+     Weight_SD = sd(Weight, na.rm = TRUE),
+     Weight_Variance = var(Weight, na.rm = TRUE)
```

```
+ )
> print(weight_stats)
# A tibble: 3 × 6
  Obesity_level Weight_Mean Weight_Median Weight_Mode Weight_SD Weight_Variance
          <db1>       <db1>         <db1>       <db1>     <db1>           <db1>
1             1        62.2          61            60      9.27            86.0
2             2        95.0          94.5          90     10.8            117.
3             3        75.0          75            75      8.46            71.5
```

Interpretation:

Weight Statistics
Mean Weight: The average weight increases with obesity level, from 62.2 kg in Level 1 to 95.0 kg in Level 2, and 75.0 kg in Level 3.

- Median Weight: The median weight mirrors the mean trends, being lowest in Level 1 and highest in Level 2.
- Mode: The mode is the most frequent weight value. For Level 1, it's 60 kg, for Level 2, it's 90 kg, and for Level 3, it's 75 kg. This shows typical weight values associated with each obesity level.
- Standard Deviation (SD): Weight variability is highest in Level 2 (SD = 10.8), indicating a broader range of weights. Level 1 has the lowest SD.
- Variance: Variance values follow SD patterns, showing Level 2 has the most spread in weight data.

Question 5-Visualization

i. Age

R Script:

```
> #Q5
> library(ggplot2)
> # Ensure Obesity_level is a factor with proper labels
> dataset$Obesity_level <- factor(dataset$Obesity_level,
+                                 levels = c("Normal_Weight", "Overweight_Level_I", "Obesity_Type_I"),
+                                 labels = c("Normal Weight", "Overweight Level I", "Obesity Type I"))
> # Boxplot for Age by Obesity Level
> ggplot(dataset, aes(x=Obesity_level, y=Age, fill=Obesity_level)) +
+   geom_boxplot() +
+   labs(title="Boxplot of Age by Obesity Level", x="Obesity Level", y="Age") +
+   theme_minimal()
```

Output:

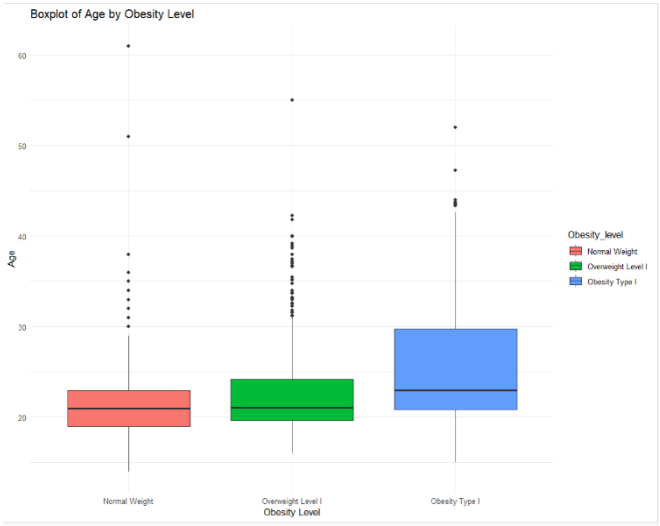


Figure 12. Boxplot for Age against Obesity Levels.

Interpretation:

The median age increases with obesity level, with Normal Weight having a median of around 23-24 years, Overweight Level I around 22-23 years, and Obesity Type I around 28-29 years. The variability (IQR) also increases with obesity, indicating greater age spread as obesity level rises. Outliers are most frequent in the Obesity Type I group, reflecting more age diversity, with a pattern of older individuals being more likely to have higher obesity levels.

ii. Height

R Script:

```
> #height
> ggplot(dataset, aes(x=Height, fill=Obesity_level)) +
+   geom_histogram(binwidth=0.05, position="dodge") +
+   facet_wrap(~Obesity_level, scales="free_y") +
+   labs(title="Histogram of Height by Obesity Level", x="Height", y="Count") +
+   theme_minimal()
```

Output:

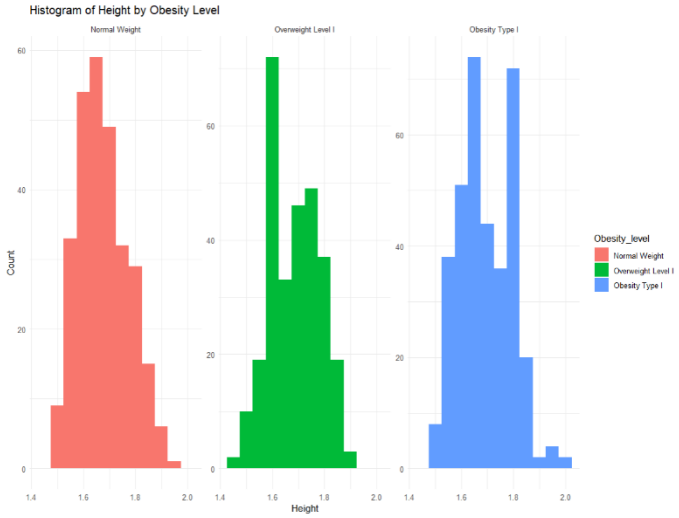


Figure 13. Histogram of Height against Obesity Levels.

Interpretation:
Figure13. illustrates the distribution of height across three obesity categories: Normal Weight, Overweight Level I, and Obesity Type I. For both Normal Weight and Overweight Level I groups, the height distribution is centered around 1.7 to 1.8 meters, with a relatively symmetrical and concentrated pattern, indicating a consistent central tendency. In contrast, the Obesity Type I group exhibits a more varied distribution, with a notable bimodal pattern that includes peaks around 1.7 meters and 1.9 meters, suggesting a wider range of heights within this group. Overall, while height remains fairly consistent between the Normal and Overweight groups, there is greater height variation in individuals with Obesity Type I.

iii. Weight

R Script:

```
> #weight
> ggplot(dataset, aes(x=Obesity_level, y=Weight, fill=Obesity_level)) +
+   geom_boxplot() +
+   labs(title="Boxplot of Weight by Obesity Level", x="Obesity Level", y="Weight") +
+   theme_minimal()
```

Output:

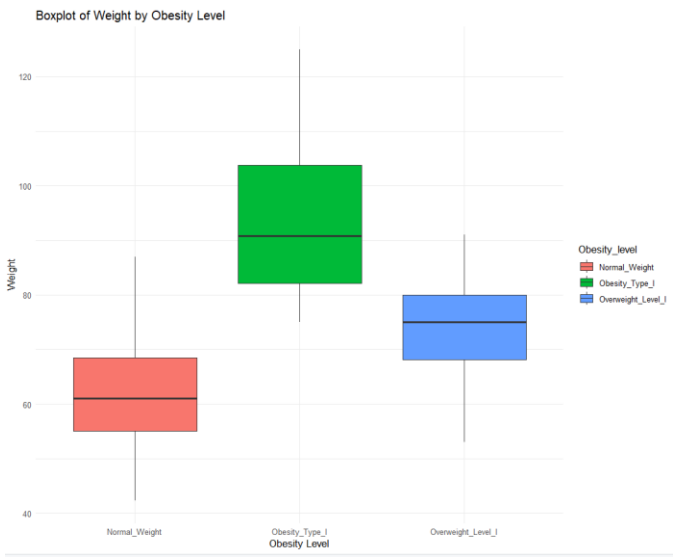


Figure 14. Boxplot for Weight against Obesity Levels.

Interpretation:
The median weight increases with obesity level, with Normal Weight having a median of around 65 kg, Overweight Level I around 75 kg, and Obesity Type I around 95 kg. The variability (IQR) also increases with obesity, indicating greater weight spread as obesity level rises. Obesity Type I shows the widest range, reflecting more weight diversity, while Normal Weight and Overweight Level I have more consistent weight distributions.

iv. Gender

R Script:

```
> #gender
> library(ggplot2)
> # Plot for Gender with faceting
> ggplot(dataset, aes(x = Gender, fill = Gender)) +
+   geom_bar(position = "dodge") +
+   facet_wrap(~Obesity_level, scales = "free_y") +
+   labs(title = "Bar Plot of Gender by Obesity Level", x = "Gender", y = "Count") +
+   scale_fill_manual(values = c("Female" = "lightblue", "Male" = "lightcoral")) +
+   theme_minimal()
```

Output:

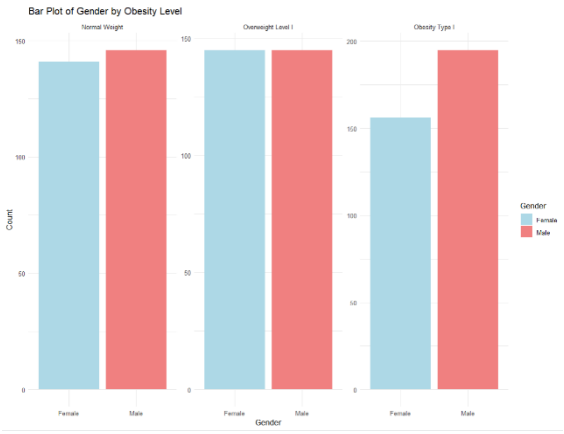


Figure 15. Bar graph for Gender against Obesity Levels.

Interpretation:

For the Normal Weight group, the number of males slightly exceeds females, but the difference is minimal, indicating near parity between the genders in this category.
In the Overweight Level I group, the number of males and females is equal, showing no gender disparity at this level of obesity.
However, in the Obesity Type I group, there is a more noticeable difference, with males having a higher count (nearly 200) compared to females (just over 150). This suggests that males are more represented in the highest obesity category, while the earlier categories show more balance between the genders.

v. Family_history

R Script:

```
> #family history
> library(ggplot2)
> # Grouped bar plot for Family History
> ggplot(dataset, aes(x = Obesity_level, fill = Family_history)) +
+   geom_bar(position = "dodge") +
+   labs(title = "Grouped Bar Plot of Family History by Obesity Level", x = "Obesity Level", y =
"Count") +
+   scale_fill_manual(values = c("yes" = "lightblue", "no" = "lightcoral")) +
+   theme_minimal()
```

Output:

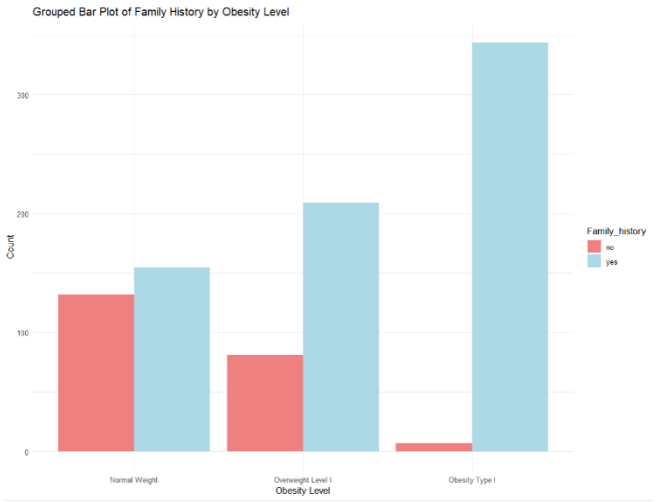


Figure 16. Bar graph for Family history against Obesity Levels.

Interpretation:

Normal Weight Group: The counts for individuals with and without a family history of obesity are fairly balanced, though slightly more individuals in this group report no family history.

Overweight Level I Group: There is a noticeable shift, with more individuals having a family history of obesity compared to those without. The number of people with a family history exceeds those without by a small margin.

Obesity Type I Group: The difference becomes much more pronounced, with a significant majority of individuals reporting a family history of obesity. Very few people in this group do not have a family history, suggesting a strong relationship between family history and Obesity Type I.

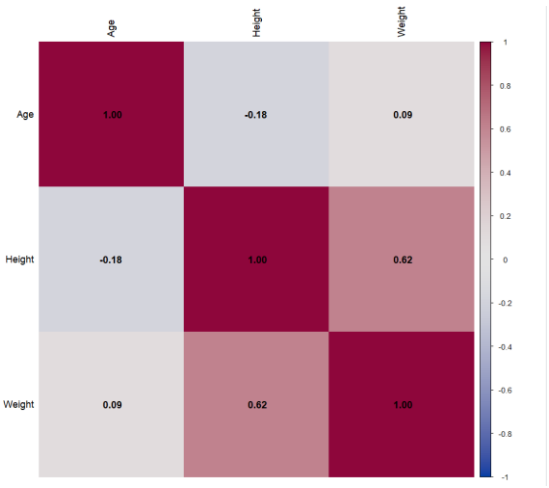
In summary, the plot suggests that as obesity level increases, so does the likelihood of having a family history of obesity, with the most dramatic difference observed in the Obesity Type I group. This may indicate a hereditary factor contributing to higher obesity levels.

vi. Correlation matrix

R Script:

```
> #Correlation
> library(corrplot)
> # Plot correlation matrix
> corrplot(cor_matrix, method="color",
+          col=colorspace::diverge_hcl(180),
+          addCoef.col = "black",
+          tl.col = "black",
+          title = "Correlation Matrix of Numerical Variables")
> # Plot correlation matrix
> corrplot(cor_matrix, method="color",
+          col=colorspace::diverge_hcl(200),
+          addCoef.col = "black",
+          tl.col = "black",
+          title = "Correlation Matrix of Numerical Variables")
```

Output:



Interpretation:

Figure 17. Correlation matrix for numerical attributes.

Interpretation:

Age and Height have a slight negative correlation (-0.18), suggesting that as age increases, there is a small tendency for height to decrease. However, this correlation is weak.
Age and Weight have a weak positive correlation (0.09), indicating little to no relationship between the two variables. In this case, age does not seem to have a significant impact on weight.
Height and Weight have a moderately strong positive correlation (0.62), meaning that as height increases, weight tends to increase as well. This is the strongest relationship shown in the matrix and suggests that height is more predictive of weight compared to age.

Overall, the most notable finding is the moderately strong correlation between height and weight, while age has a weak or minimal relationship with the other two variables.

Conclusion

In this report, I explored and analyzed the obesity levels dataset Colombia using various data preprocessing techniques and machine learning models. By examining the relationships between independent variables such as age, height, weight, gender, and family history with overweight, I aimed to identify key predictors of obesity levels. The results highlight the significance of these variables in predicting different obesity classifications and their correlations.