Nature or Nurture? Estimating Obesity Based on Physical Condition & Eating Habits

Soo-ah Kim (3035661061), Dongjun Yeom (3035666463)

2024-04-28

Introduction

"Nature or Nurture?" This is one of the most important debated questions in the field of human biology. Some say genetic and physical factors play a bigger role, and others say the environments or habits play a bigger role in shaping a condition in a person. In this report, we focused on estimating the obesity level based on physical condition and eating habits. Obesity considered a great threat in the developed economies, and is an increasing threat in the developing countries. According to Public Health England, 63% of adults in England were overweight or obese in 2015. It is an important task to find out what factors influence the obesity levels most and to predict the obesity level based on a person's physical condition and eating habits.

We aim to analyse what factors play a bigger role in the obesity level of a person, using various machine learning models, such as logistic regression, k-nearest neighbours, naive bayes analysis, support vector machine, decision tree and random forest. As our target variable is ordinal, We further employed linear and polynomial regressions after converting the obesity level into numeric scale.

About the Data

```
# Import data

data <- read.csv("obesity.csv")
head(data)
```

```
##   Gender Age Height Weight family_history_with_overweight FAVC FCVC NCP
## 1 Female  21   1.62   64.0                            yes   no    2   3
## 2 Female  21   1.52   56.0                            yes   no    3   3
## 3   Male  23   1.80   77.0                            yes   no    2   3
## 4   Male  27   1.80   87.0                             no   no    3   3
## 5   Male  22   1.78   89.8                             no   no    2   1
## 6   Male  29   1.62   53.0                             no  yes    2   3
##        CAEC SMOKE CH2O SCC FAF TUE       CALC                MTRANS
## 1 Sometimes    no    2  no   0   1         no Public_Transportation
## 2 Sometimes   yes    3 yes   3   0  Sometimes Public_Transportation
## 3 Sometimes    no    2  no   2   1 Frequently Public_Transportation
## 4 Sometimes    no    2  no   2   0 Frequently               Walking
## 5 Sometimes    no    2  no   0   0  Sometimes Public_Transportation
## 6 Sometimes    no    2  no   0   0  Sometimes            Automobile
##            NObeyesdad
## 1       Normal_Weight
## 2       Normal_Weight
## 3       Normal_Weight
```

```
## 4  Overweight_Level_I
## 5 Overweight_Level_II
## 6       Normal_Weight
```

We employed 2,111 records for the estimation of obesity levels in individuals from the countries of Mexico, Peru and Colombia, based on their eating habits and physical condition. The original data was donated to the UC Irvine Machine Learning Repository and can be found from the following link.

https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition

Attributes

```r
# Add BMI into consideration

data$BMI <- with(data, Weight / (Height^2))
head(data)
```

```
##   Gender Age Height Weight family_history_with_overweight FAVC FCVC NCP
## 1 Female  21   1.62   64.0                            yes   no    2   3
## 2 Female  21   1.52   56.0                            yes   no    3   3
## 3   Male  23   1.80   77.0                            yes   no    2   3
## 4   Male  27   1.80   87.0                             no   no    3   3
## 5   Male  22   1.78   89.8                             no   no    2   1
## 6   Male  29   1.62   53.0                             no  yes    2   3
##        CAEC SMOKE CH2O SCC FAF TUE       CALC                MTRANS
## 1 Sometimes    no    2  no   0   1         no Public_Transportation
## 2 Sometimes   yes    3 yes   3   0  Sometimes Public_Transportation
## 3 Sometimes    no    2  no   2   1 Frequently Public_Transportation
## 4 Sometimes    no    2  no   2   0 Frequently               Walking
## 5 Sometimes    no    2  no   0   0  Sometimes Public_Transportation
## 6 Sometimes    no    2  no   0   0  Sometimes            Automobile
##            NObeyesdad      BMI
## 1       Normal_Weight 24.38653
## 2       Normal_Weight 24.23823
## 3       Normal_Weight 23.76543
## 4  Overweight_Level_I 26.85185
## 5 Overweight_Level_II 28.34238
## 6       Normal_Weight 20.19509
```

Note: The original dataset contained 17 attributes, and 1 attribute has been added for the analysis.

The following attributes were considered for the analyses:

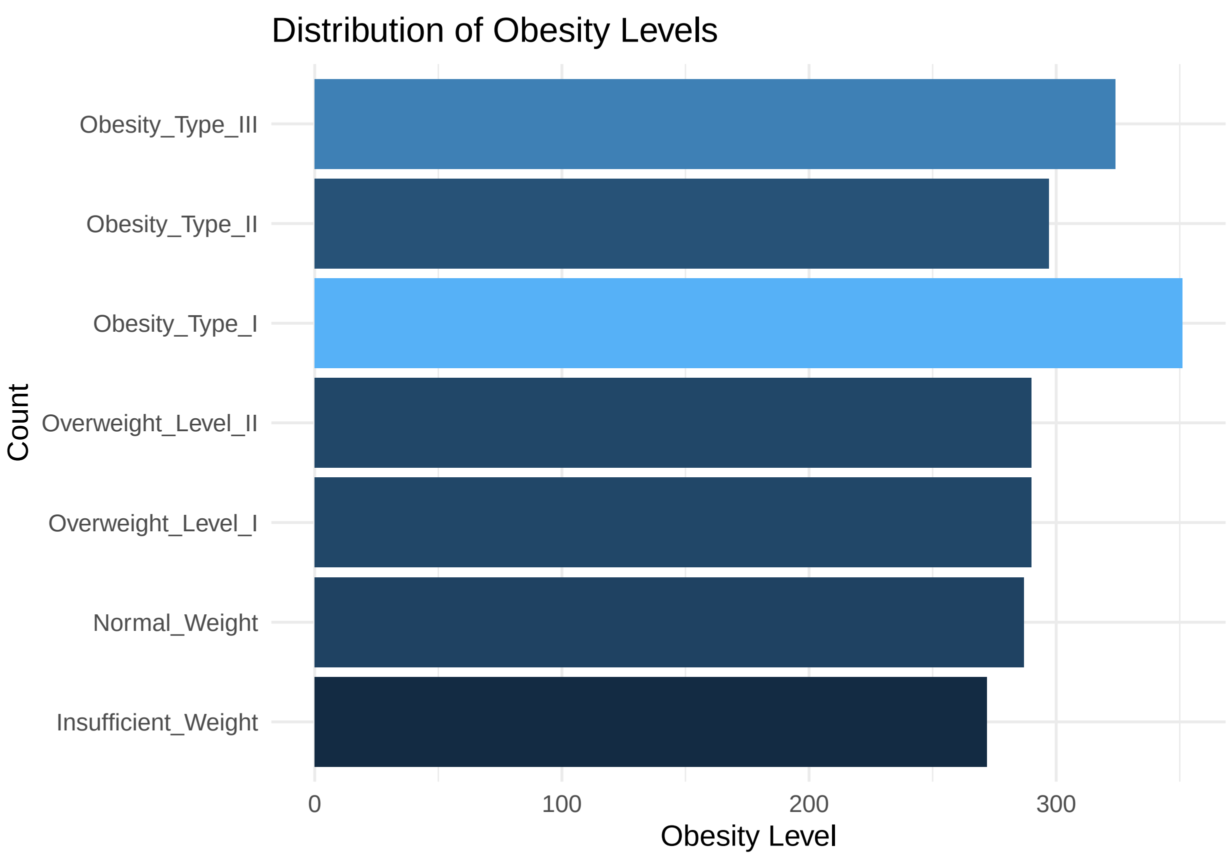
1. *The attributes related with the physical condition:*
   - Gender
   - Age
   - Height
   - Family History with Obesity
   - Body Mass Index (BMI)
2. *The attributes related with eating habits:*
   - Frequent consumption of high caloric food (FAVC)
   - Frequency of consumption of vegetables (FCVC)
   - Number of main meals (NCP)

- Consumption of food between meals (CAEC)
- Whether they smoke cigarettes (SMOKE)
- Consumption of water daily (CH20)
- Consumption of alcohol (CALC)
- Calories consumption monitoring (SCC)
- Physical activity frequency (FAF)
- Time using technology devices (TUE)
- Transportation used (MTRANS)

Exploratory Data Analysis

Bar Plot

```
#bar chart

# Set the factor levels for obesity level in ascending order
data$NObeyesdad <- factor(data$NObeyesdad,
levels = c("Insufficient_Weight", "Normal_Weight", "Overweight_Level_I", "Overweight_Level_II", "Obesity

# Plot
ggplot(data, aes(x = NObeyesdad, fill = ..count..)) +
  geom_bar() +
  coord_flip() +
  labs(x = "Count", y = "Obesity Level", title = "Distribution of Obesity Levels") +
  theme_minimal() +
  theme(legend.position = "none")
```



In figure 1, the bar plot above indicates the distribution of the data. Based on the Body Mass Index calculated, the data was classified into 7 categories:

- Insufficient_Weight: Less than 18.5
- Normal_Weight: 18.5 to 24.9
- Overweight_Level_I: 25.0 to 27.4
- Overweight_Level_II: 27.5 to 29.9
- Obesity_Type_I: 30.0 to 34.9
- Obesity_Type_II: 35.0 to 39.9
- Obesity_Type_III: Higher than 40

The distribution of obesity level shows no class imbalance, with each of the classes having similar numbers of instances all ranging from 250 to 350. This might indicate a sampling bias due to the difference with the real world distribution, which is more or less bell-curved, as shown in Figure 2 below (Al-Malki et al., 2003).

However, having less class imbalance might actually work as an advantage in our analysis. Models trained on balanced data are less likely to overfit to the majority class. They are more likely to generalize better to unseen data since they have had the opportunity to learn representative features from all classes equally.

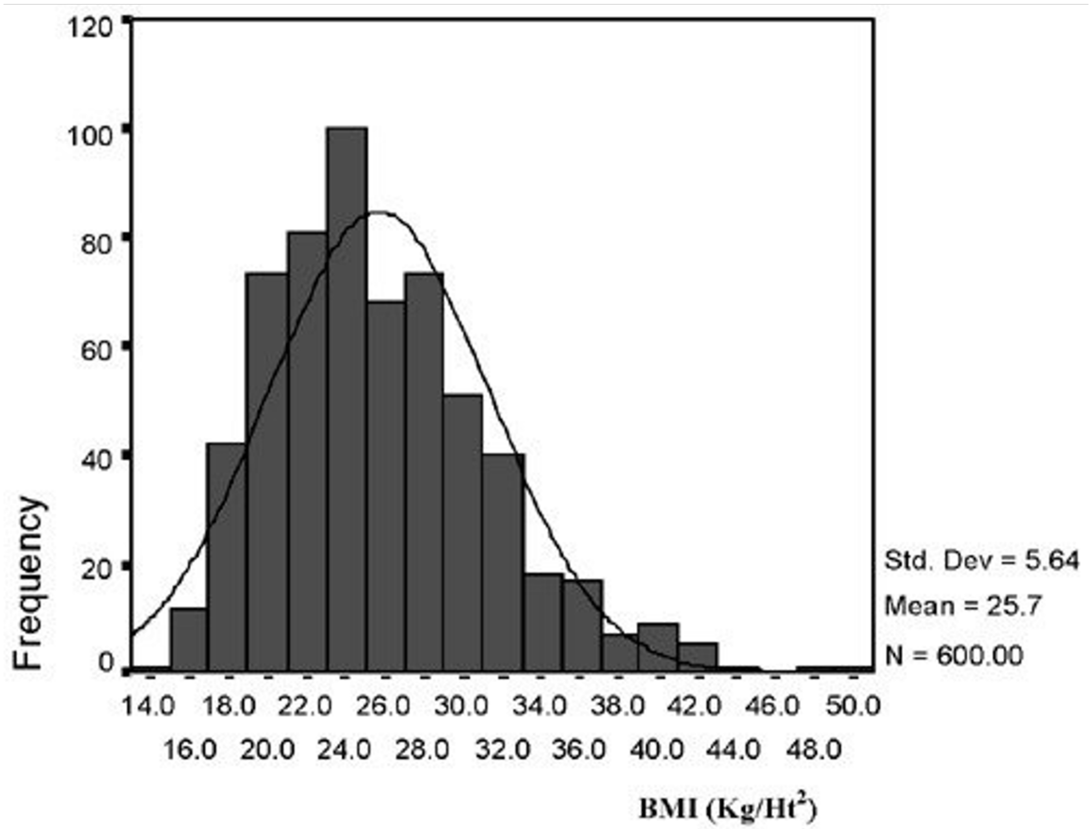
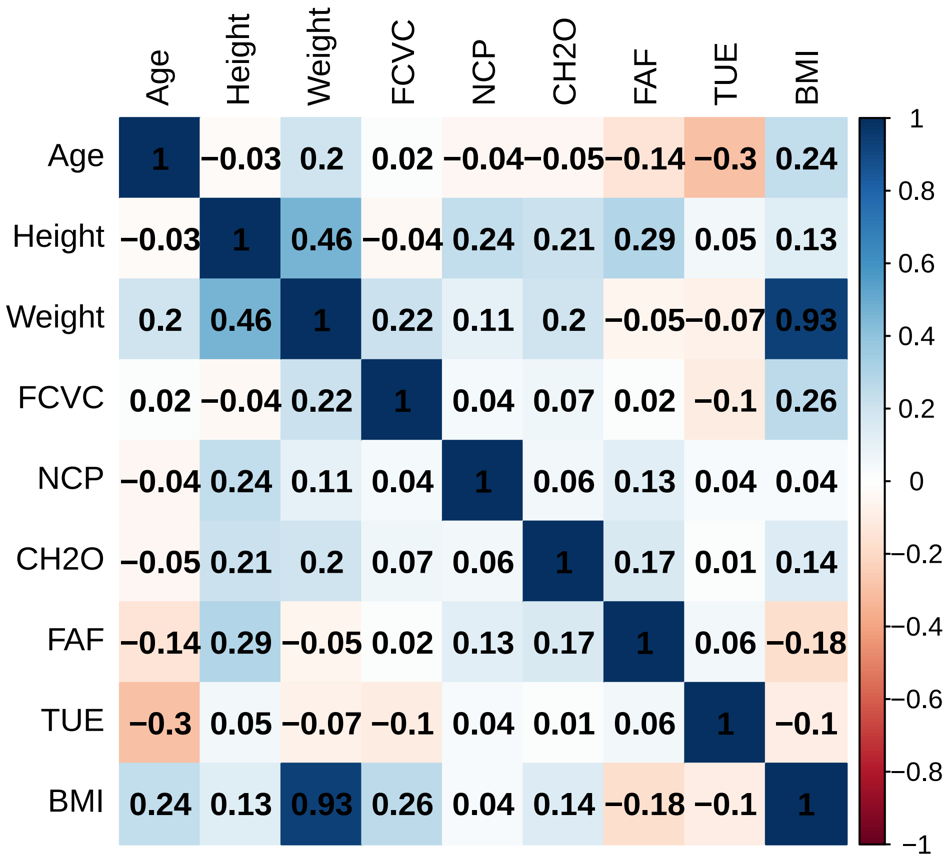


Figure 1: Figure 2. Real-world BMI distribution (Al-Malki et al., 2003)

Correlation Plot

```
# correlation heatmap for numeric columns
```

```r
numeric_columns <- sapply(data, is.numeric) & !names(data) %in% 'NObeyesdad'
numeric_data <- data[numeric_columns]

cor_matrix <- cor(numeric_data)

corrplot(cor_matrix, method = "color",
         tl.col = "black",
         addCoef.col = "black")
```



The above correlation plot presents the inter-relationships between a set of variables. The matrix reveals a positive correlation between *Weight* and *BMI*, where 0.93 signifies a strong direct relationship. Similarly, *Height* and *Weight* shows a moderate strong positive correlation with a coefficient of 0.46, implying that *height* is a contributing factor to *weight*. This is explainable since *BMI* is calculated by the following equation:

$$BMI = \frac{weight(Kg)}{height(m)^2}$$

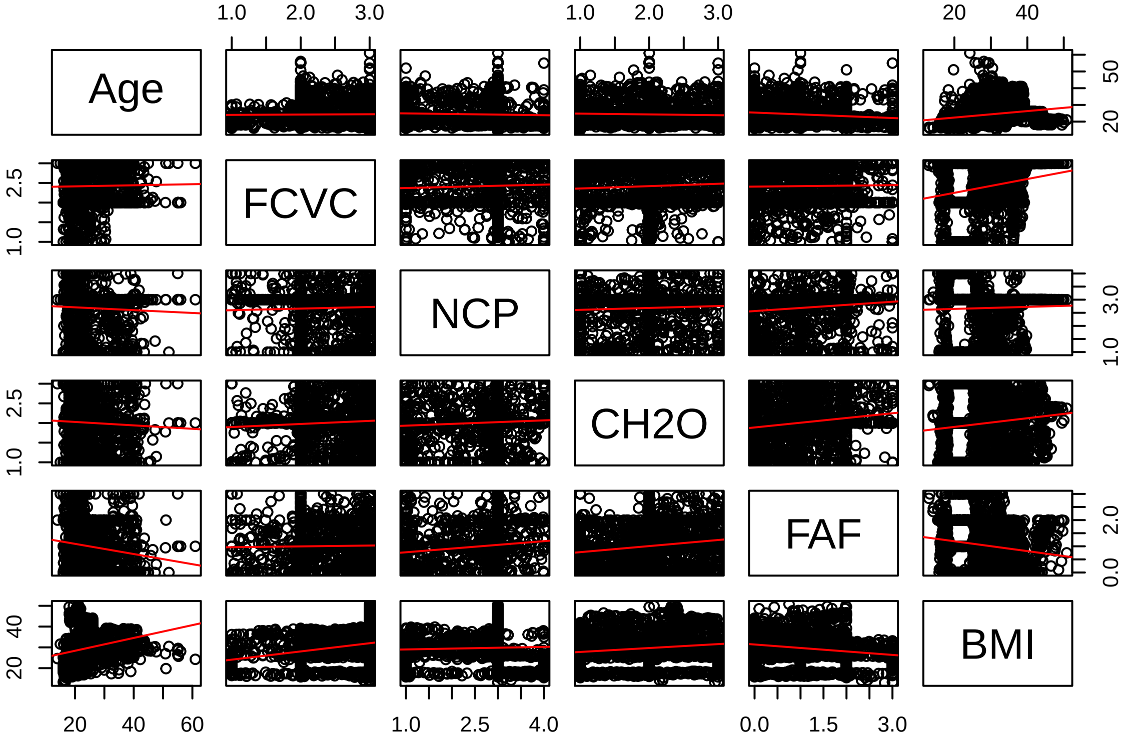
Due to the high correlation, we decided not to include *BMI* as a variable in the following analyses to prevent multicollinearity problems.

On the other hand, a moderate negative correlation with a coefficient of -0.3 is observed between *TUE* and *FAF*, indicating that an increase in one may correspond with a decrease in the other. For other pairs, they exhibited relatively weak correlations, coefficients being lower than 0.3 in absolute manner, suggesting negligible linear associations.

For better visualisation, our correlation plot employs the colour scale: darker shades of blue represent stronger positive correlations and darker shades of red denote stronger negative correlations.

Pairwise Plot

```
# pairwise plot of selected columns

ppcor <- function(x, y, ...) {
  points(x, y, ...)
  abline(lm(y ~ x), col = "red")
}

selected_data <- data[, c("Age", "FCVC", "NCP", "CH2O", "FAF", "BMI")]
pairs(selected_data, panel = ppcor)
```

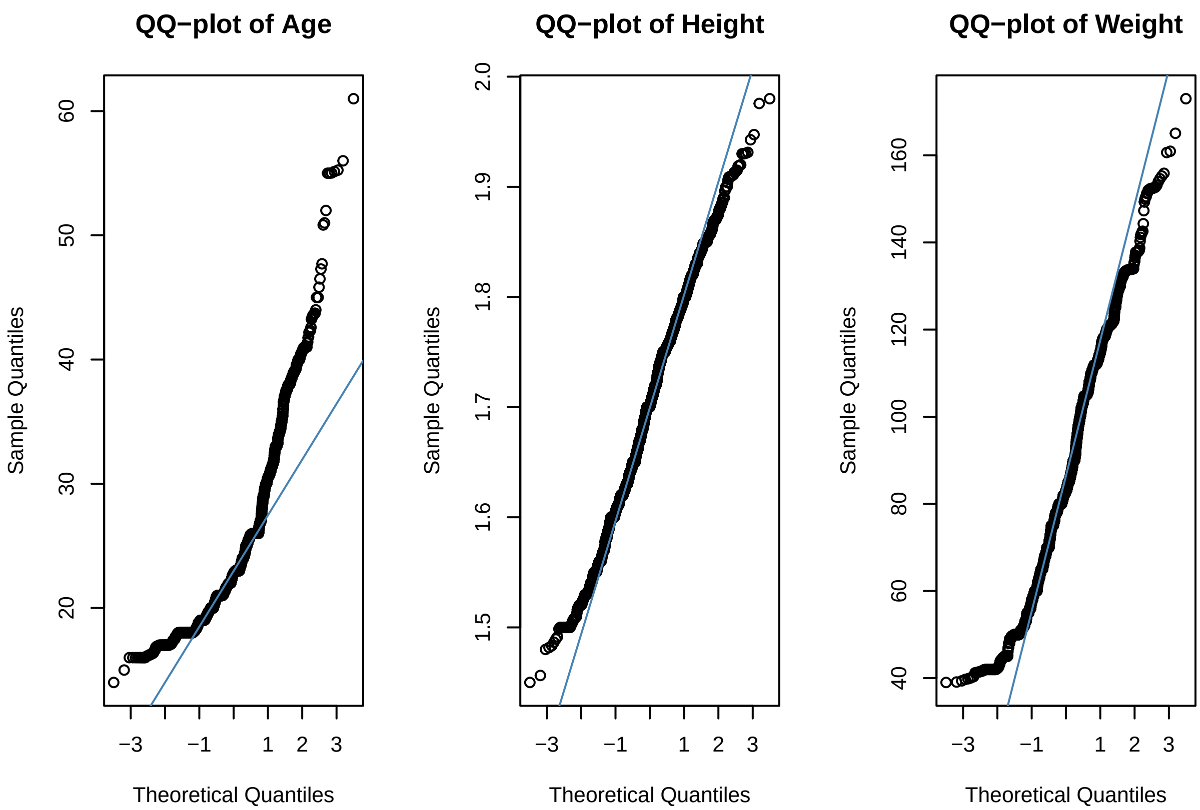


The above pairwise plot examines the potential correlations between the selected variables, including *Age, FCVC, NCP, CH2O, FAF*, and *BMI*. The red linear regression lines within each scatterplot serve as a reference to estimate the linearity between variables. While the regression lines are mostly flat for all the scatterplots, suggesting a weak linear relationship, the *Age* and *BMI* shows a gradual positive linearity. This further supports the decision to eliminate *BMI* to prevent multicollinearity. Overall, while some linear relationship may exist, it is not strong or consistent across all variable pairs. Thus, other than *BMI*, there are no further variables to remove in order to prevent strong multicollinearity between variables.

QQ-Plots

```
# QQ-plots

# Define the columns you want to plot
columns_to_plot <- c("Age", "Height", "Weight")

# Check if all specified columns exist in the data
if (!all(columns_to_plot %in% names(data))) {
  stop("One or more specified columns do not exist in the dataset.")
```

```
}

# Plotting each selected column
par(mfrow = c(1, length(columns_to_plot)))  # Arrange plots in a single row
for (col in columns_to_plot) {
    qqnorm(data[[col]], main = paste("QQ-plot of", col))
    qqline(data[[col]], col = "steelblue")  # Add a reference line
}
```




```
par(mfrow = c(1, 1))  # Reset plot layout
```

```
# Shapiro-Wilk Test
shapiro.test(data$Age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$Age
## W = 0.86606, p-value < 2.2e-16
```

```
# Kurtosis
kurtosis(data$Age)
```

```
## [1] 5.816858
```

From the QQ-plot, we can see that the variables *height* and *weight* generally follows the normal distribution line, but *age* doesn't seem to quite follow a normal distribution. To check the deviance from normal distribution of the variable *age*, we performed the Sapiro-Wilk normality test and the kurtosis test.

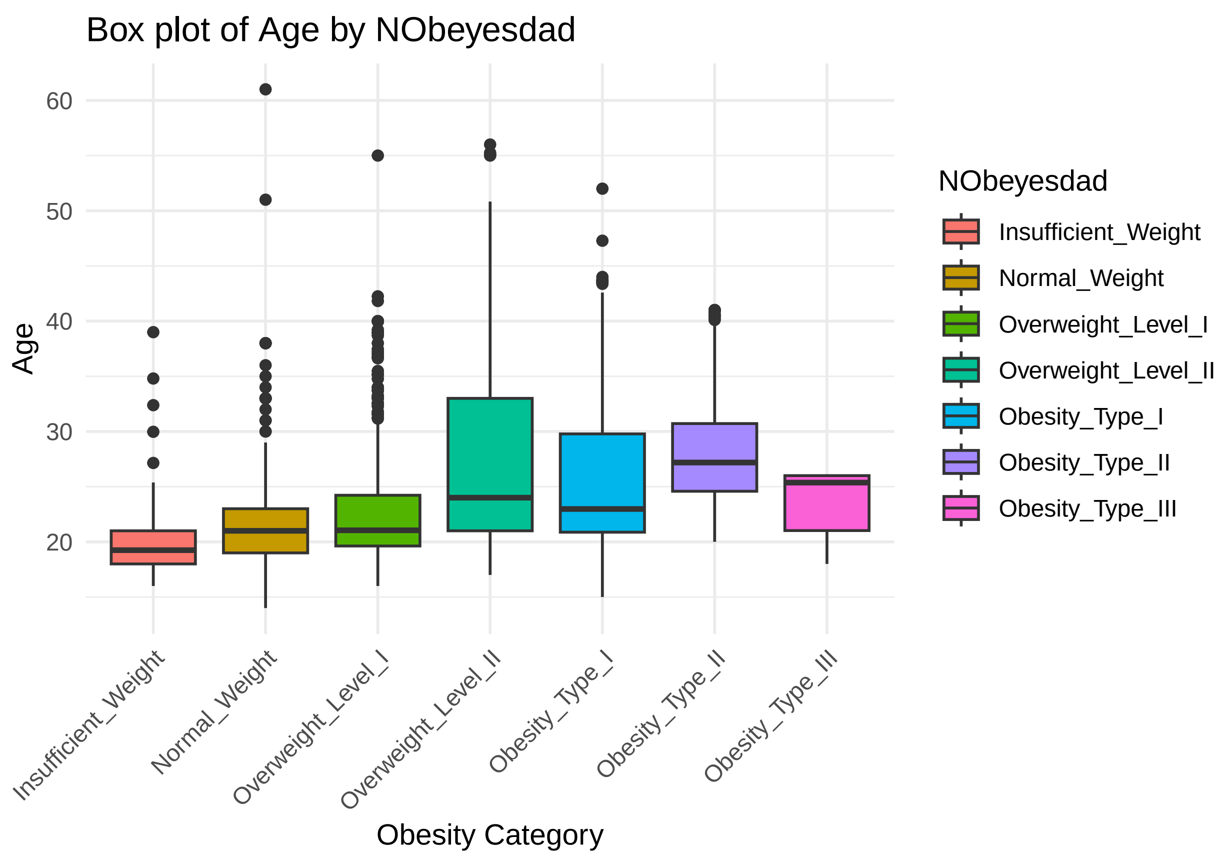
Shapiro-Wilk Normality Test The W-statistic value of W = 0.86606 indicates how well the data conforms to a normal distribution. A value closer to 1 would indicate data that more closely follows a normal distribution. A value of 0.86606 suggests a noticeable deviation from normality. The p-value < 2.2e-16 is highly significant, which strongly rejects the null hypothesis that the data are normally distributed. This result confirms that the distribution of Age significantly deviates from a normal distribution.

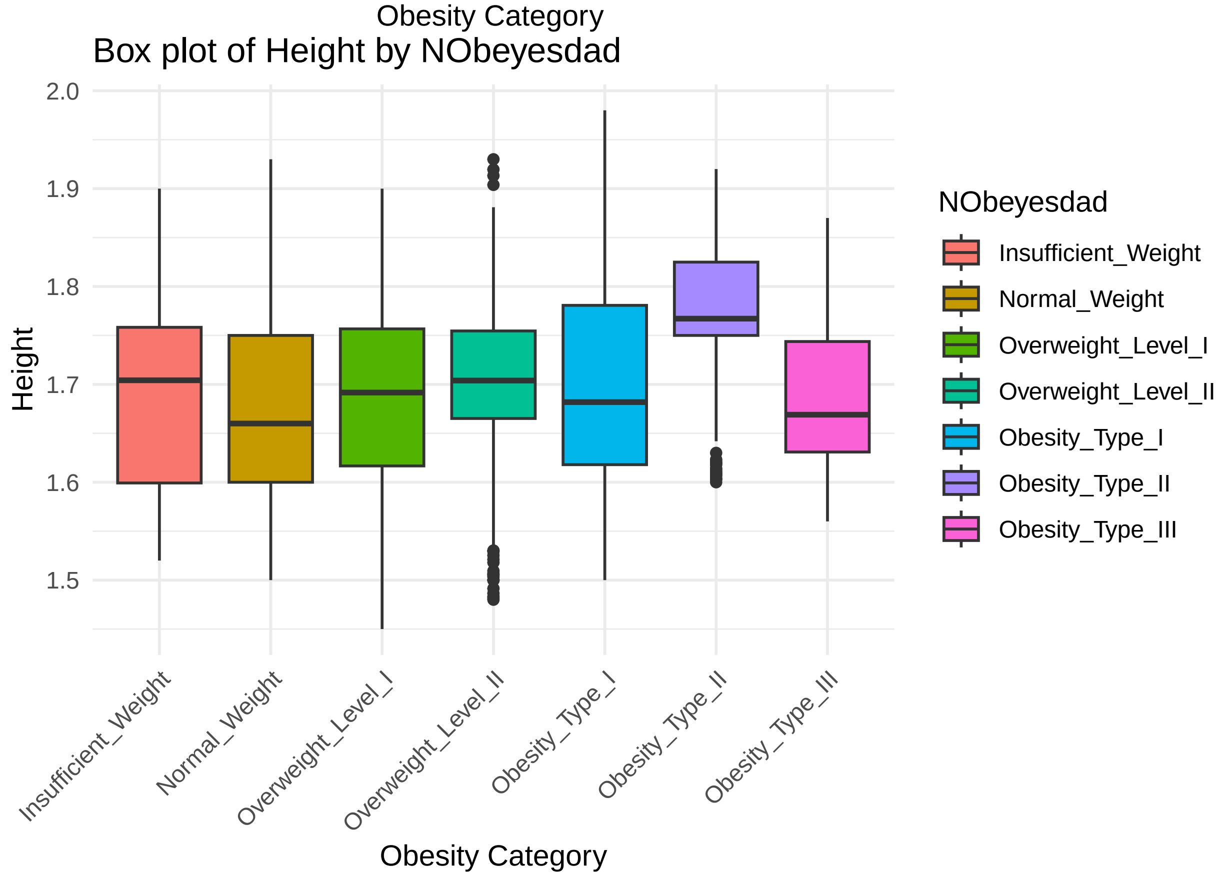
Kurtosis The reported kurtosis value of 5.816858 indicates that the distribution has heavier tails than a normal distribution (which has a kurtosis of 3). This leptokurtic nature is consistent with the QQ-plot observation where both tails were above the normal line, suggesting more extreme values in the tails than expected under normality.

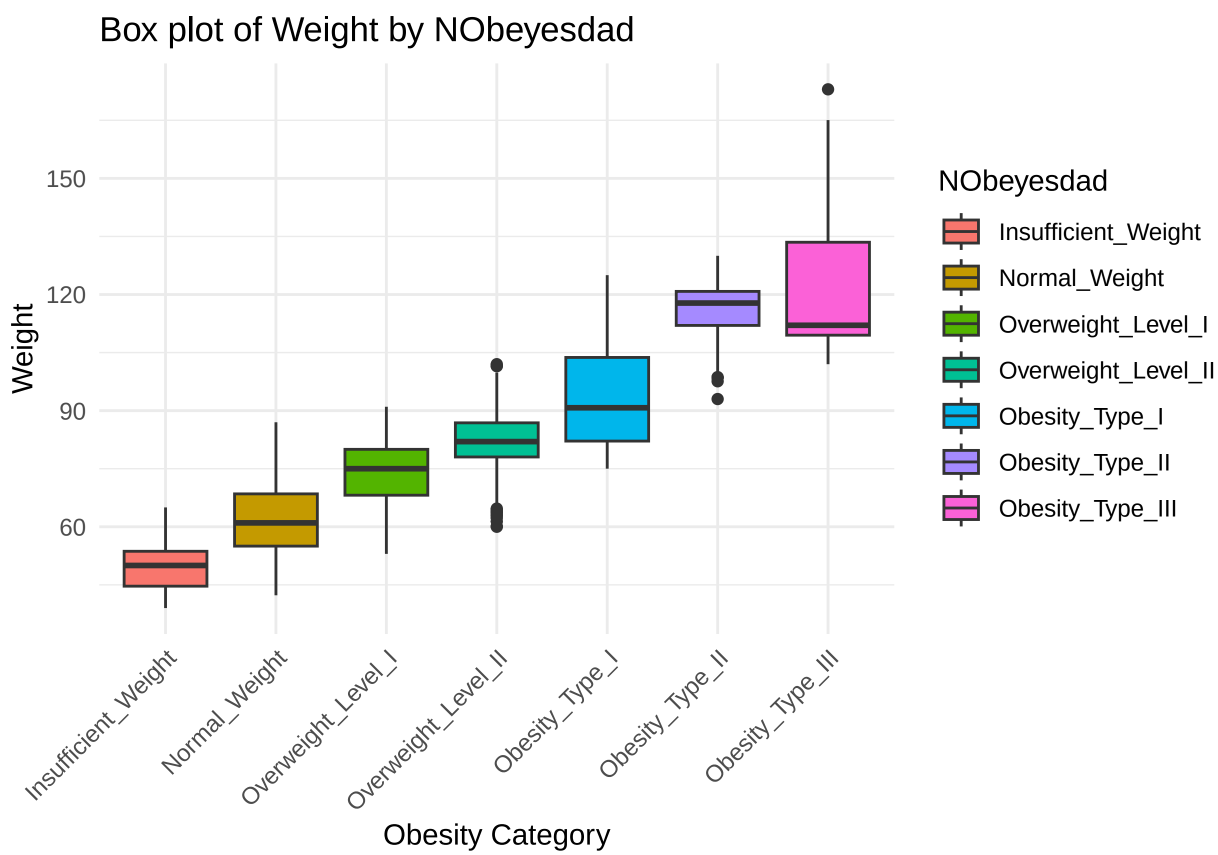
Box Plots for Numeric and Nominal Variables

```r
# Box Plots

# Loop through the numeric columns to create box plots
for (variable_name in names(numeric_data)) {
  p <- ggplot(data, aes_string(x = 'NObeyesdad', y = variable_name, fill = 'NObeyesdad')) +
    geom_boxplot() +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
    labs(title = paste("Box plot of", variable_name, "by NObeyesdad"),
         x = "Obesity Category",
         y = variable_name)

  print(p)
}
```

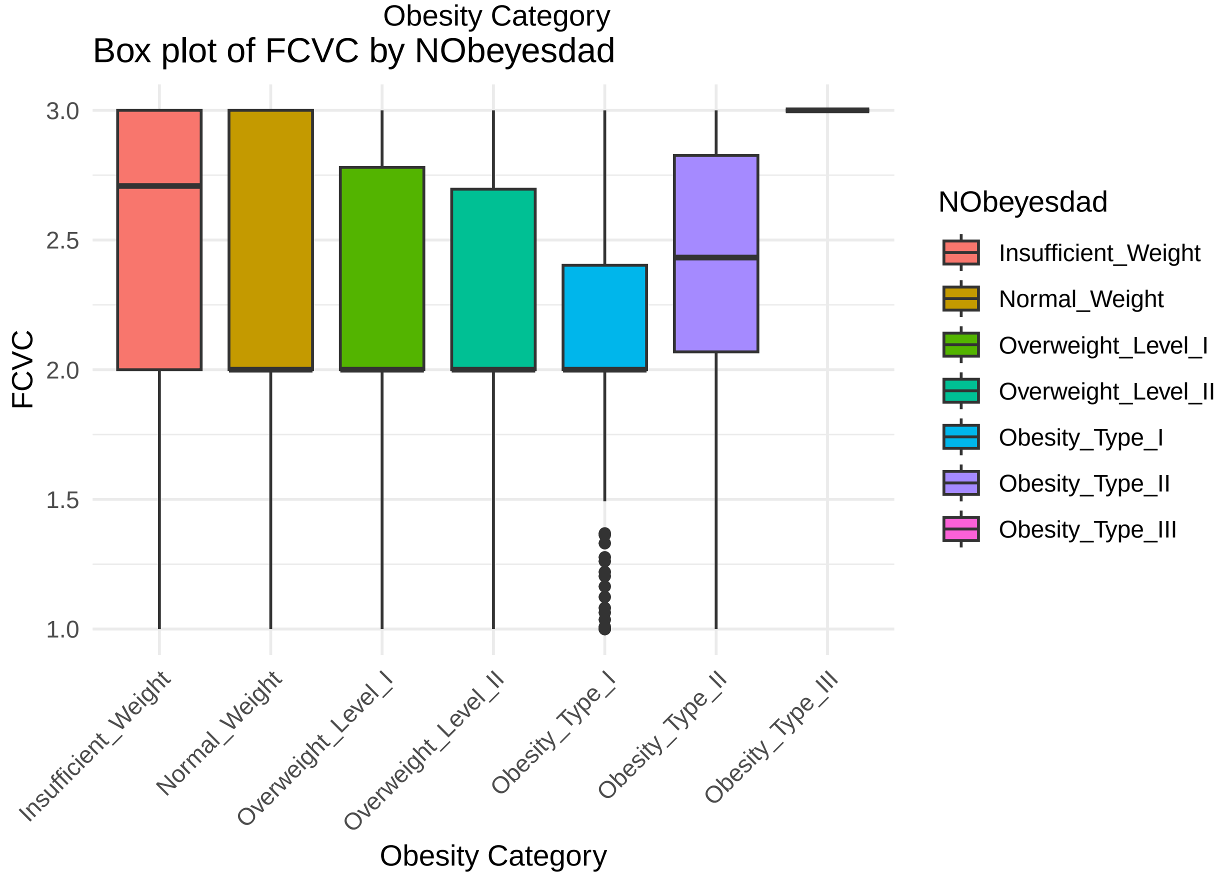
```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()`.
## i See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```


Box plot of Age by NObeyesdad
Age
60
50
40
30
20
Insufficient_Weight
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III
Obesity Category
NObeyesdad
Insufficient_Weight
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III

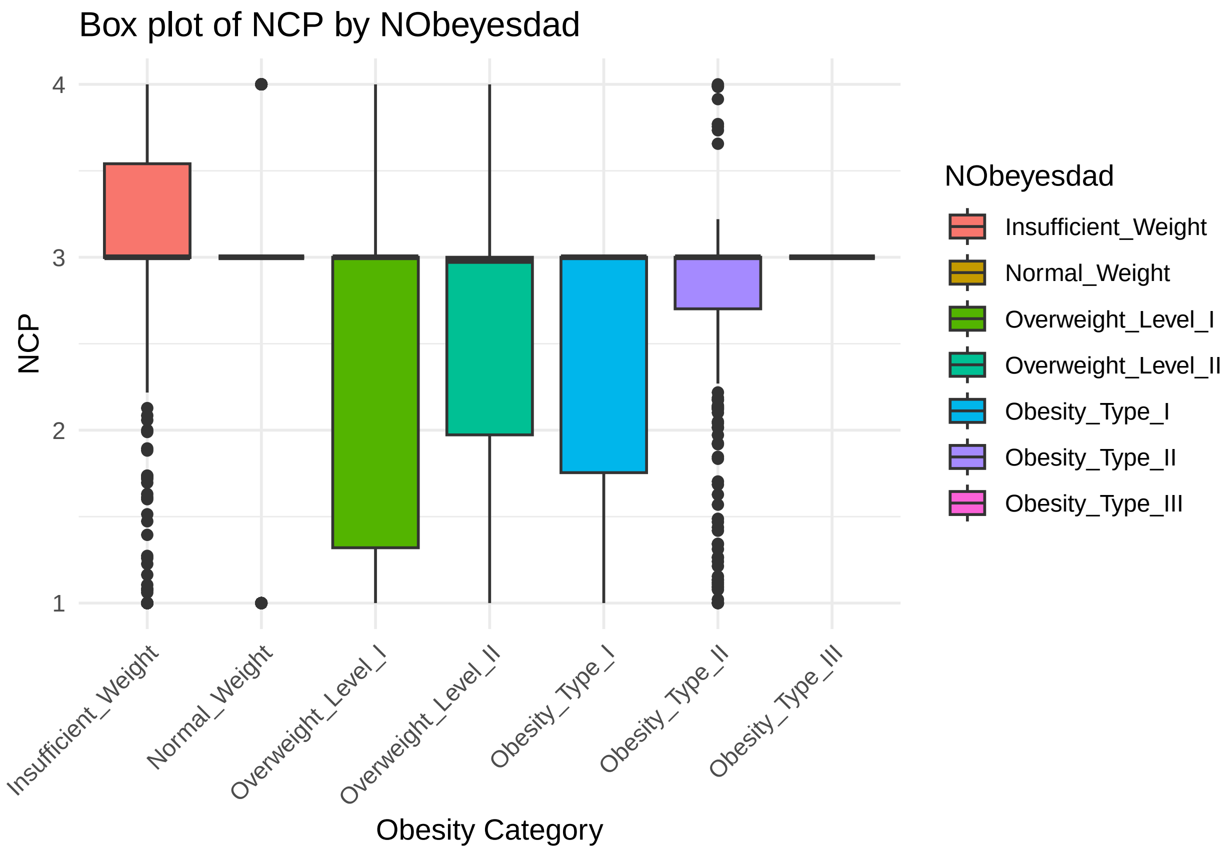

Box plot of Height by NObeyesdad
Height
2.0
1.9
1.8
1.7
1.6
1.5
Insufficient_Weight
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III
Obesity Category
NObeyesdad
Insufficient_Weight
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III

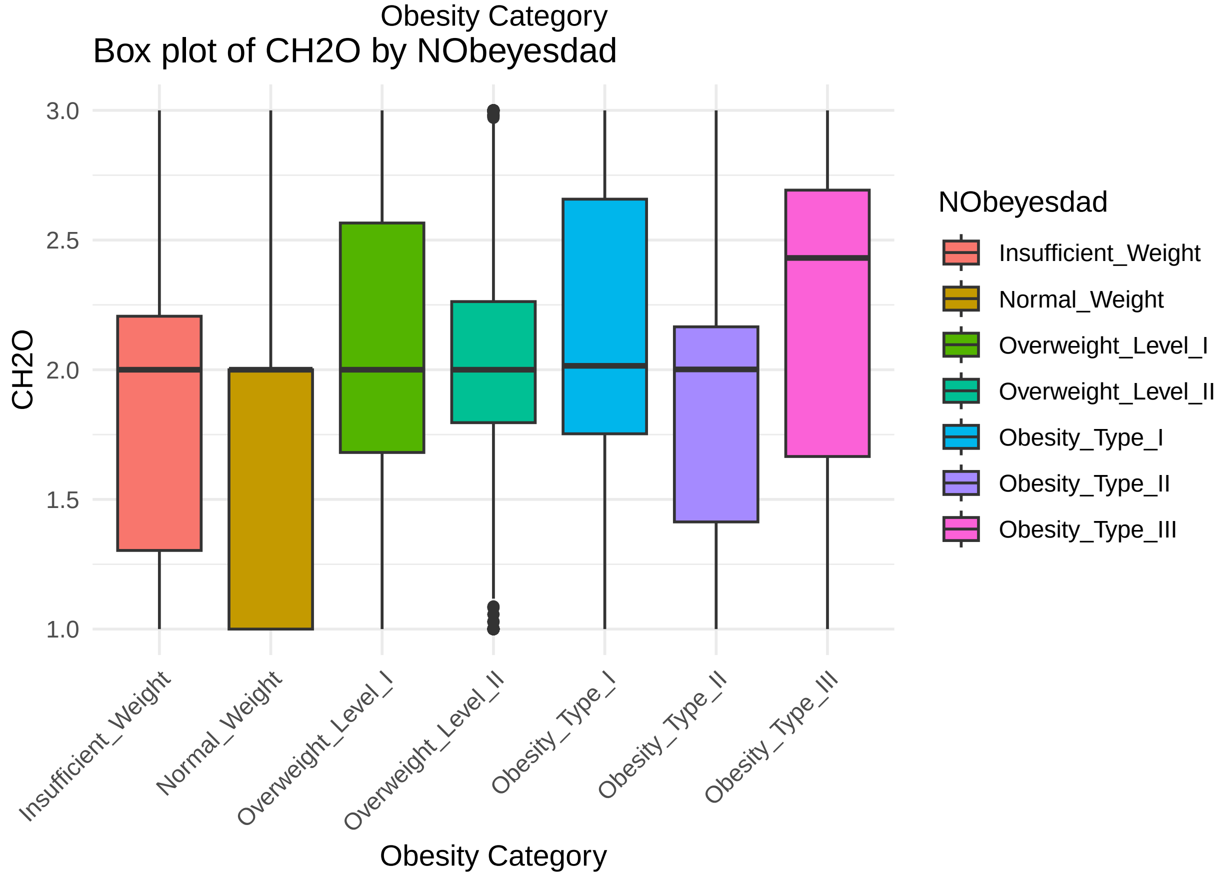


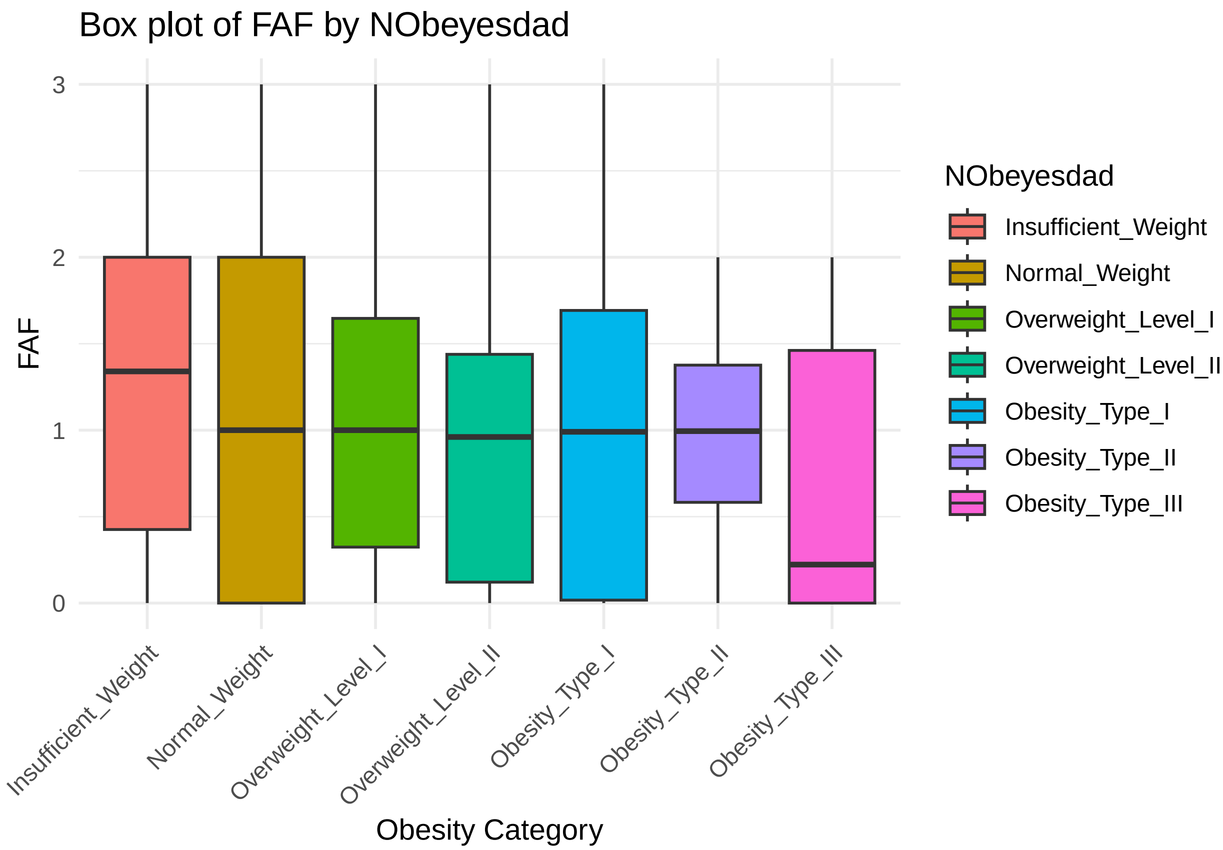
Box plot of Weight by NObeyesdad
Weight
150
120
90
60
Insufficient_Weight
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III
Obesity Category
NObeyesdad
Insufficient_Weight
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III

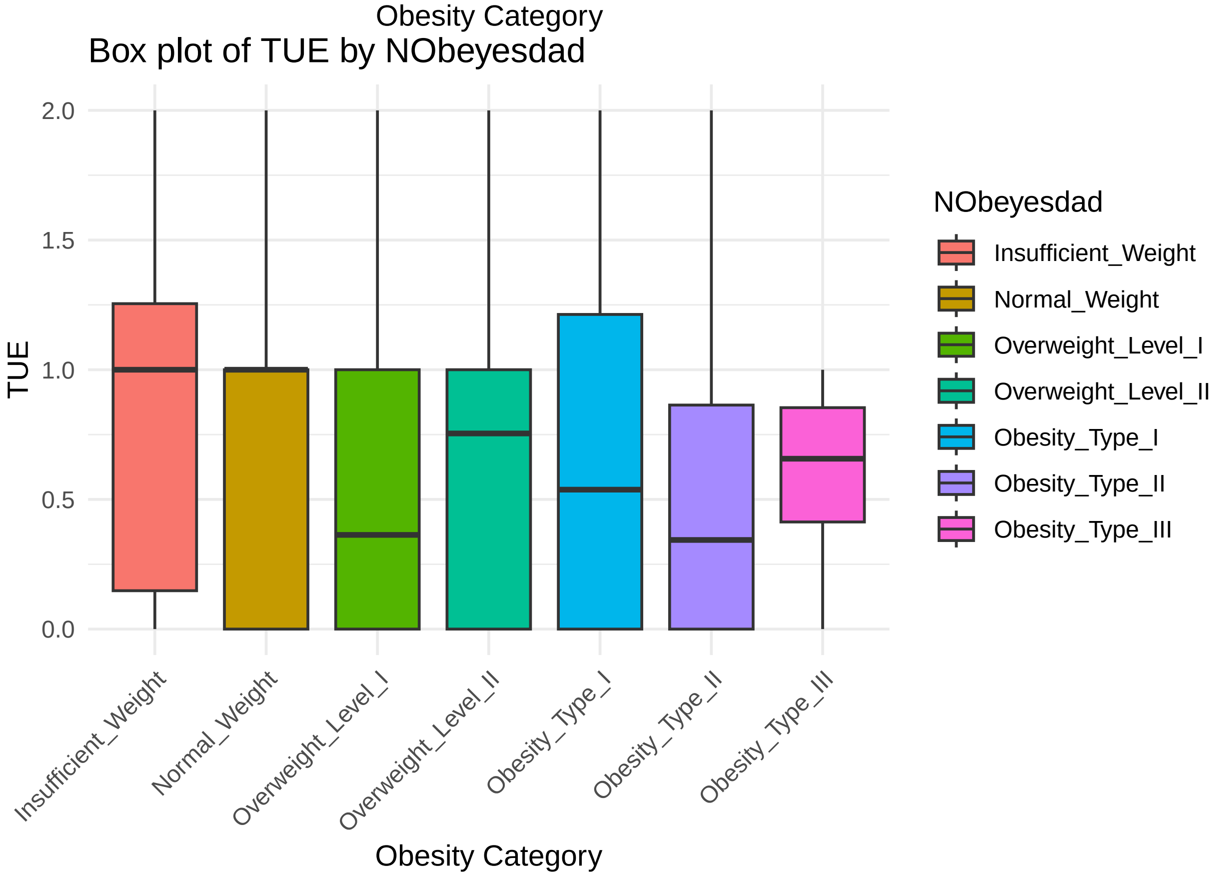


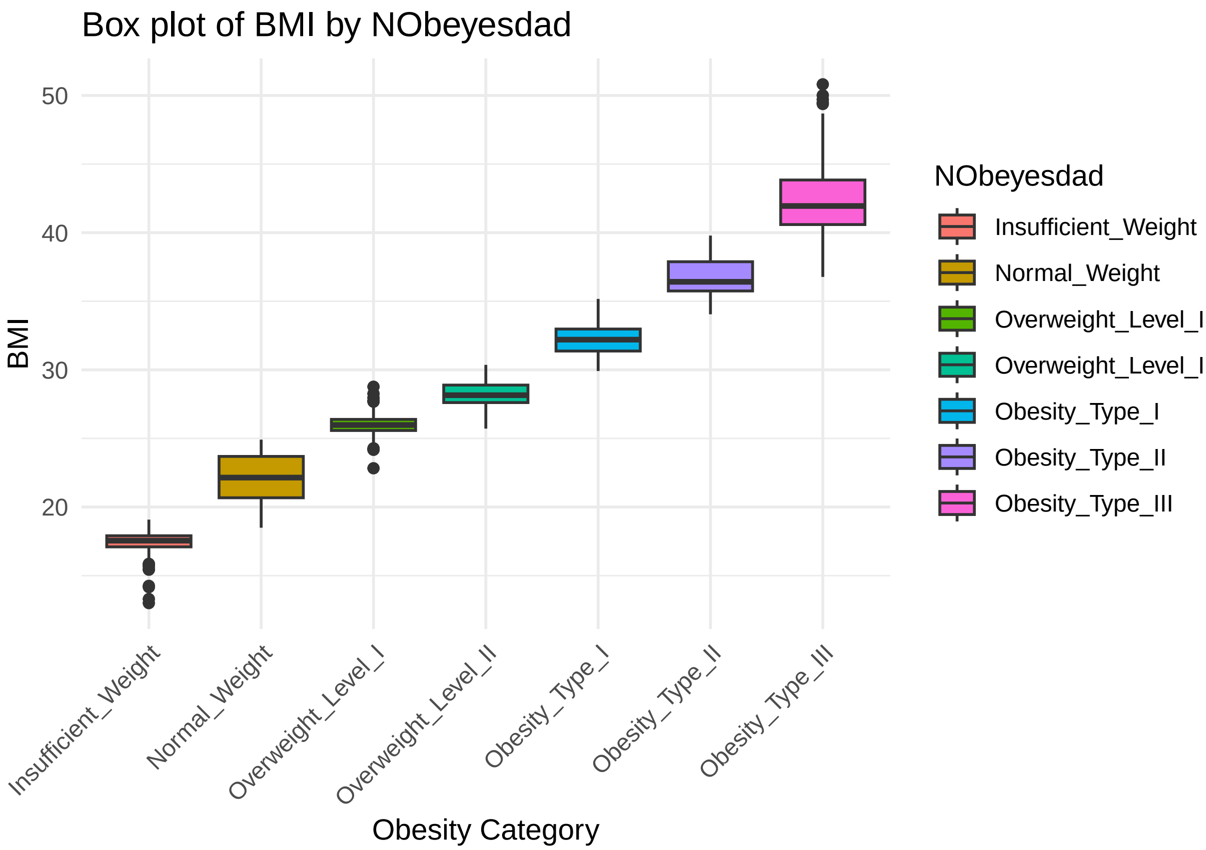
Box plot of FCVC by NObeyesdad
FCVC
3.0
2.5
2.0
1.5
1.0
Insufficient_Weight
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III
Obesity Category
NObeyesdad
Insufficient_Weight
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III


Box plot of NCP by NObeyesdad
NCP
4
3
2
1
Insufficient_Weight
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III
Obesity Category
NObeyesdad
Insufficient_Weight
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III


Box plot of CH2O by NObeyesdad
CH2O
3.0
2.5
2.0
1.5
1.0
Insufficient_Weight
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III
Obesity Category
NObeyesdad
Insufficient_Weight
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III


Box plot of FAF by NObeyesdad
FAF
3
2
1
0
Insufficient_Weight
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III
Obesity Category
NObeyesdad
Insufficient_Weight
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III


Box plot of TUE by NObeyesdad
TUE
2.0
1.5
1.0
0.5
0.0
Insufficient_Weight
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III
Obesity Category
NObeyesdad
Insufficient_Weight
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III



Box Plot of Age by NObeyesdad The median *ages* for all categories are normally in the early twenties, with a slight upward trend as the categories progress from *Insufficient_Weight* towards *Obesity_Type_III*. Additionally, regarding the widening interquartile ranges, the box plot portrays an increasing variability of *age* distribution in the *overweight* and *obese* categories compare to the *Normal* and *Insufficient weights*, suggesting a greater diversity in the ages of individuals as the severity of obesity increases. Moreover, the existence of outliers across all categories indicates that there are ages that deviate significantly from the median, leading to the complexity and variability of *age* within classification.

Box Plot of Height by NObeyesdad The median *height* is consistent across categories, mostly lying between approximately 1.65 and 1.75. The IQR, the middle 50% of the data, are considered stable across categories, proposing a little variation in *height* regardless of *weight*. The absence of notable differences in median *height* across *weight* categories shows that height may not be serve as a strong indicator of obesity status.

Box Plot of Weight by NObeyesdad The box Plot of Weight by NObeyesdad shows an ascending order, median *weights* increasing, indicating a positive correlation between the severity of category and median *weight*. Regarding the IQR, the size of the box are consistent across categories, except for *Obesity_Type_III*. Broader IQR may suggest the greater *weight* variation. For *Overweight_Level_II*, *Obesity_Type_II*, and *Obesity_Type_III* displays a presence of outliers.

Box Plot of FCVC by NObeyesdad The median appears to decrease as the *level of obesity* increases. The IQR is larger in the Insufficient Weight compare to the other categories, implying greater variability. Outliers exclusively exist in the *Obesity_Type_I*, suggesting that individual cases with FCVC measurements for *Obesity_Type_I* are lower than the others. There is an apparent decreasing trend of median, but the relationship is not linear, which can be seen by the increase in median *FCVC* for the *Obesity_Type_II* compared to *Type_I*.

Box Plot of NCP by NObeyesdad The median *NCP* of *Insufficient-Weight* rated 3, with a relatively compact size of IQR, leading to a less variability. The rest of categories has shown similar median *NCP* values, close to 3. The IQR was relatively longer in *Overweight_Level_I, Overweight_Level_II,* and *Obesity_Type_I*. As most of the categories has exhibited outliers, we assume that individual cases within *NCP* significantly differ from the norm.

Box Plot of CH2O by NObeyesdad The median values for each category are mostly consistent, suggesting that *CH2O* may not notably vary among all the individuals. This uniformity further suggests that *CH2O* may not serve as a significant factor influencing the differences in obesity levels. The IQRs are comparable among the categories, further supports a homogeneous *CH2O* pattern across the obesity levels.

Box Plot of FAF by NObeyesdad The plot illustrates a slight decreasing trend in the median value, with *Insufficient_Weight* rating the highest median, while *Obesity_Type_III* having the lowest. The IQRs have shown similarity across the categories indicating a consistency in the spread of *FAF* values within each obesity level. The box plot for *Obesity_Type_II* presents the narrowest IQR with same median as others, implying less variability within this group, further supporting the possibility that the individuals with *Obesity_Type_II* may contain a subgroup with higher physical activity levels. Or else, it may reflect a limitation. Overall, regarding the box plots, the correlation between lower physical activity levels and higher *obesity levels* are likely to be promoted.

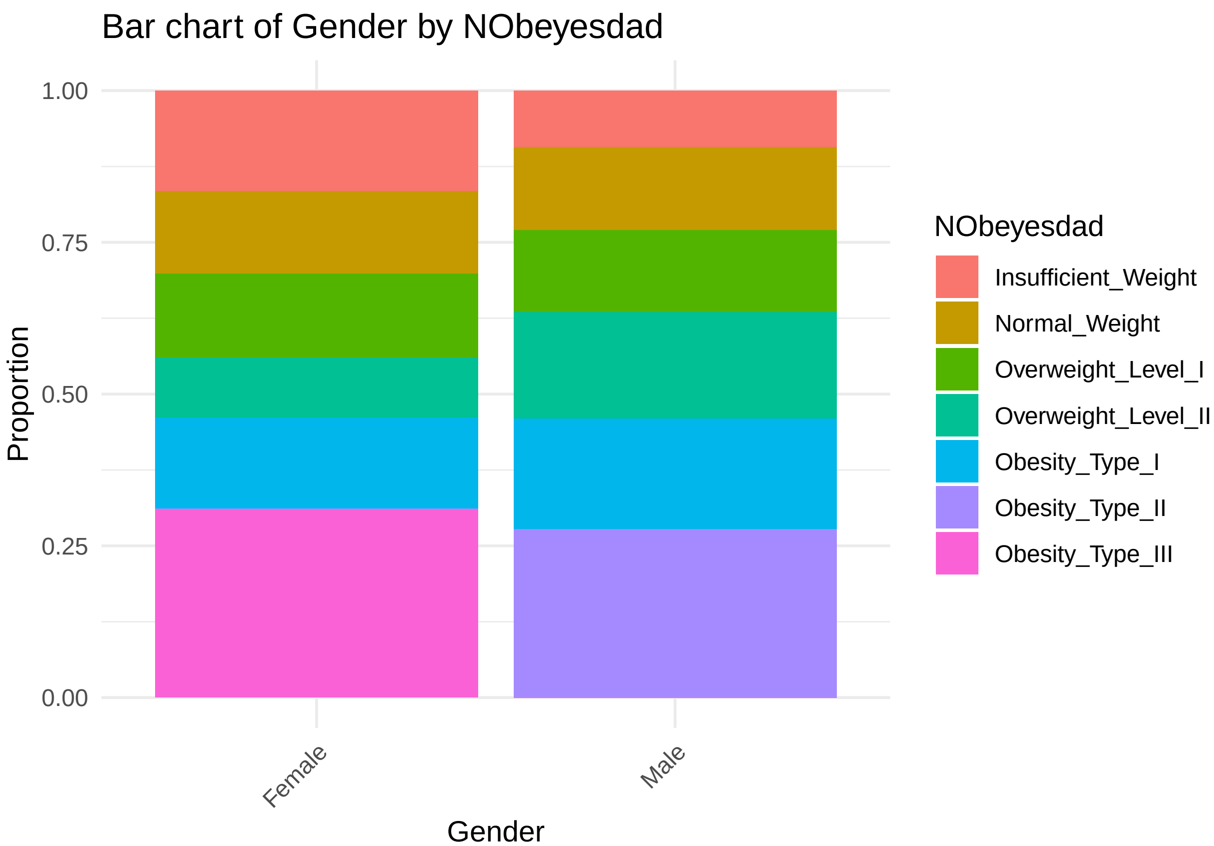
Box Plot of TUE by NObeyesdad The median TUE appears to differ across the obesity levels. The interquartile range for Obesity_Type_III is narrower compared then categories, suggesting less variability in TUE among individuals within the group. The noticeable trends, such as the linearity within median, is not noticeable in this box plot.

Box Plot of BMI by NObeyesdad The median *BMI* increases with each category. The IQR within each category are also significantly narrow. Although some categories presents the potential outliers, the box plot displays clear correlation between *BMI* and *obesity level*, supporting that the individuals with higher *BMI* are likely to fall into higher *obesity*.

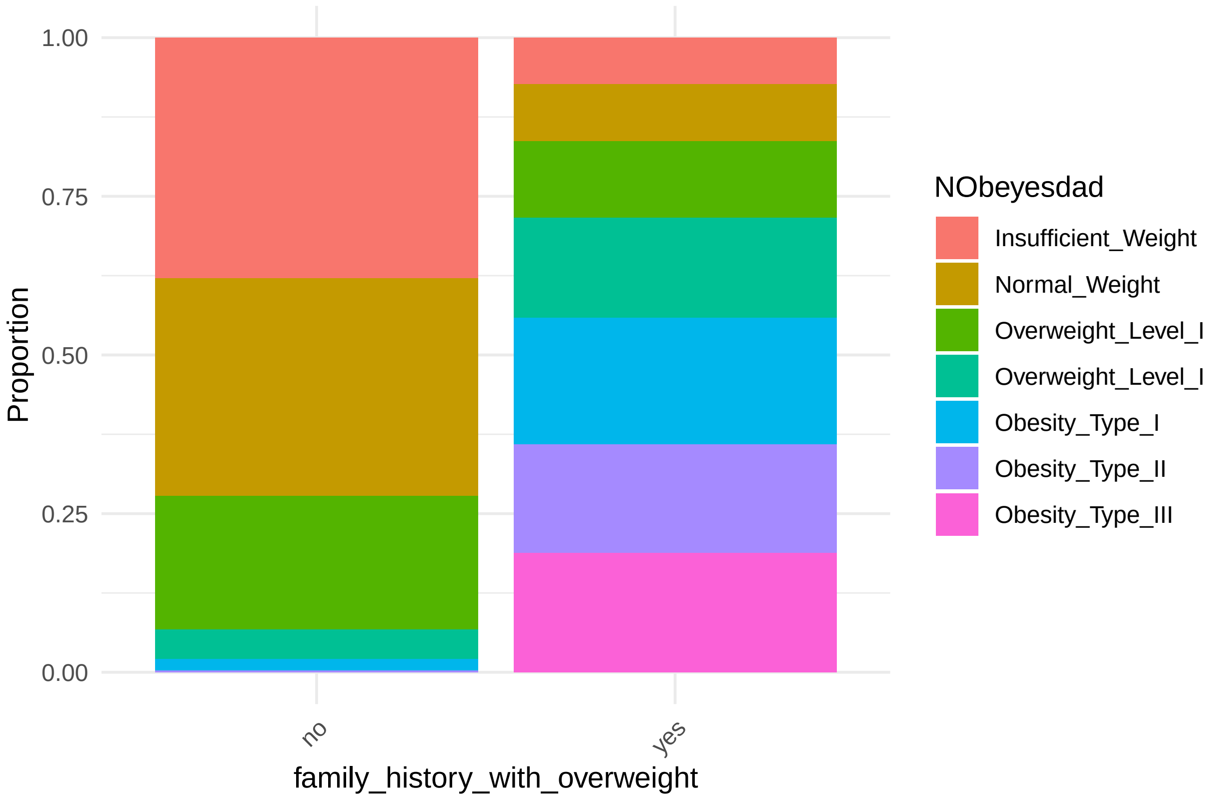
Bar Charts for Binary and Categorical Variables

```r
#plot categorical variables by obesity level

#change order of categorical variables
data$CAEC <- factor(data$CAEC, levels = c("no", "Sometimes", "Frequently", "Always"))
data$CALC <- factor(data$CALC, levels = c("no", "Sometimes", "Frequently", "Always"))
data$MTRANS <- factor(data$MTRANS, levels = c("Walking", "Bike", "Motorbike",
                                              "Public_Transportation", "Automobile"))

# Identify the categorical and binary columns
categorical_columns <- sapply(data, function(x) is.factor(x) || is.character(x) || length(unique(x)) ==
categorical_data <- data[categorical_columns]

# Loop through the categorical columns to create bar charts
for (variable_name in names(categorical_data)) {
  if (variable_name != "NObeyesdad") {
    p <- ggplot(data, aes_string(x = variable_name, fill = 'NObeyesdad')) +
      geom_bar(position = "fill") +
      theme_minimal() +
      theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
      labs(title = paste("Bar chart of", variable_name, "by NObeyesdad"),
           x = variable_name,
```

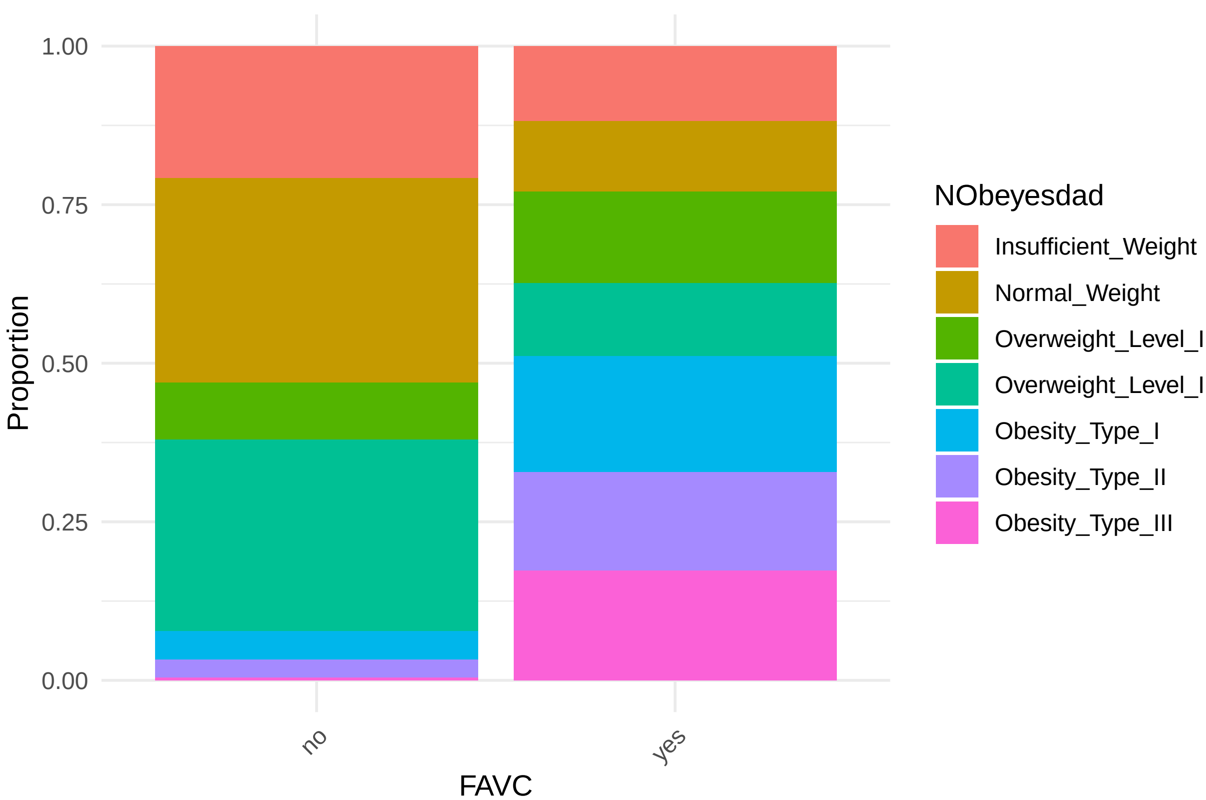
```
          y = "Proportion")

    print(p)
  }
}
```

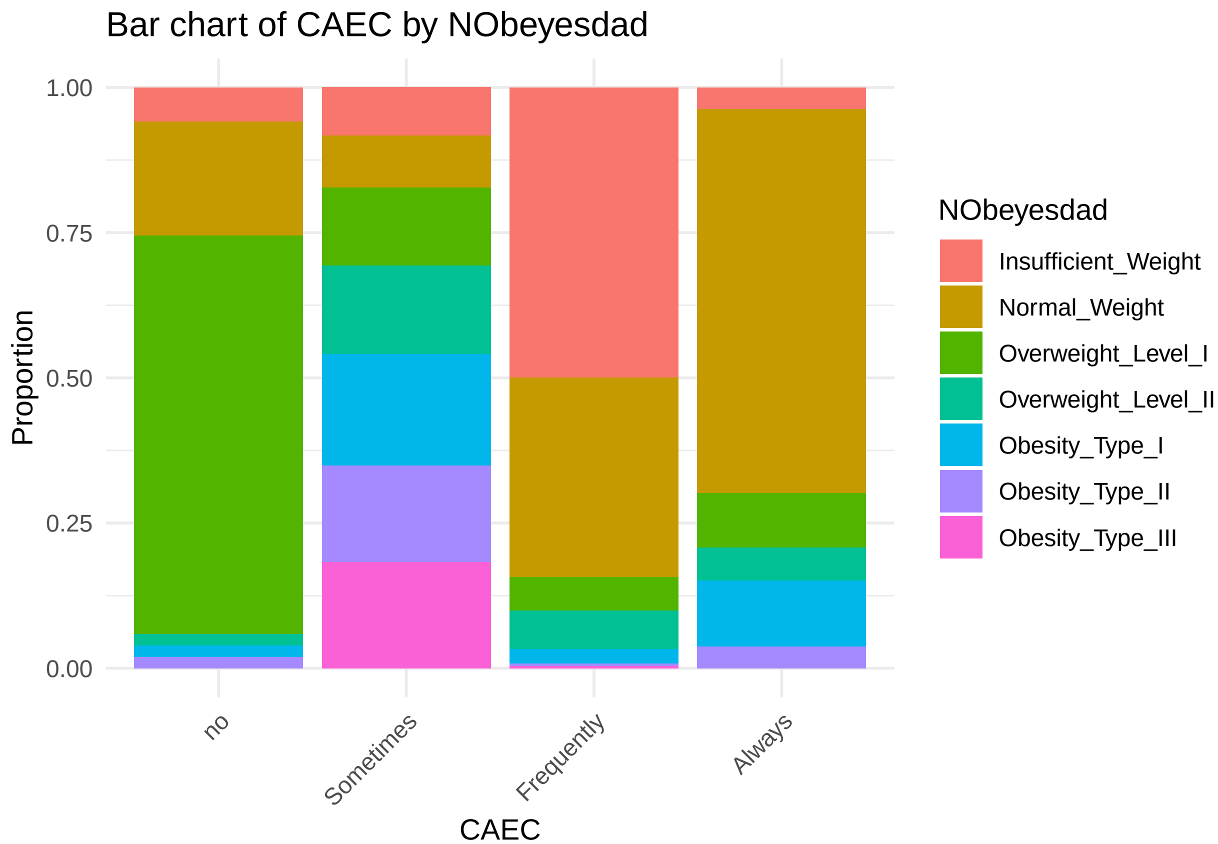


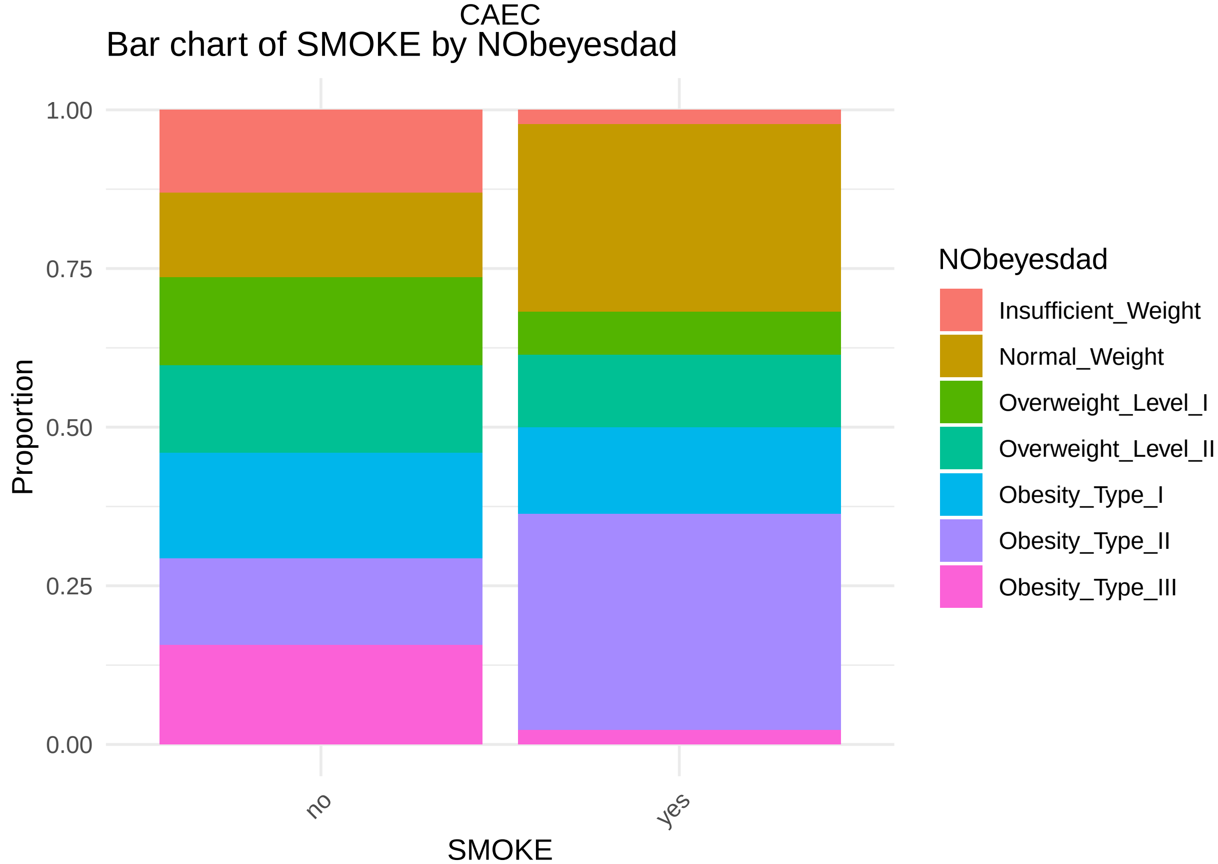
Bar chart of family_history_with_overweight by NObeyesdad

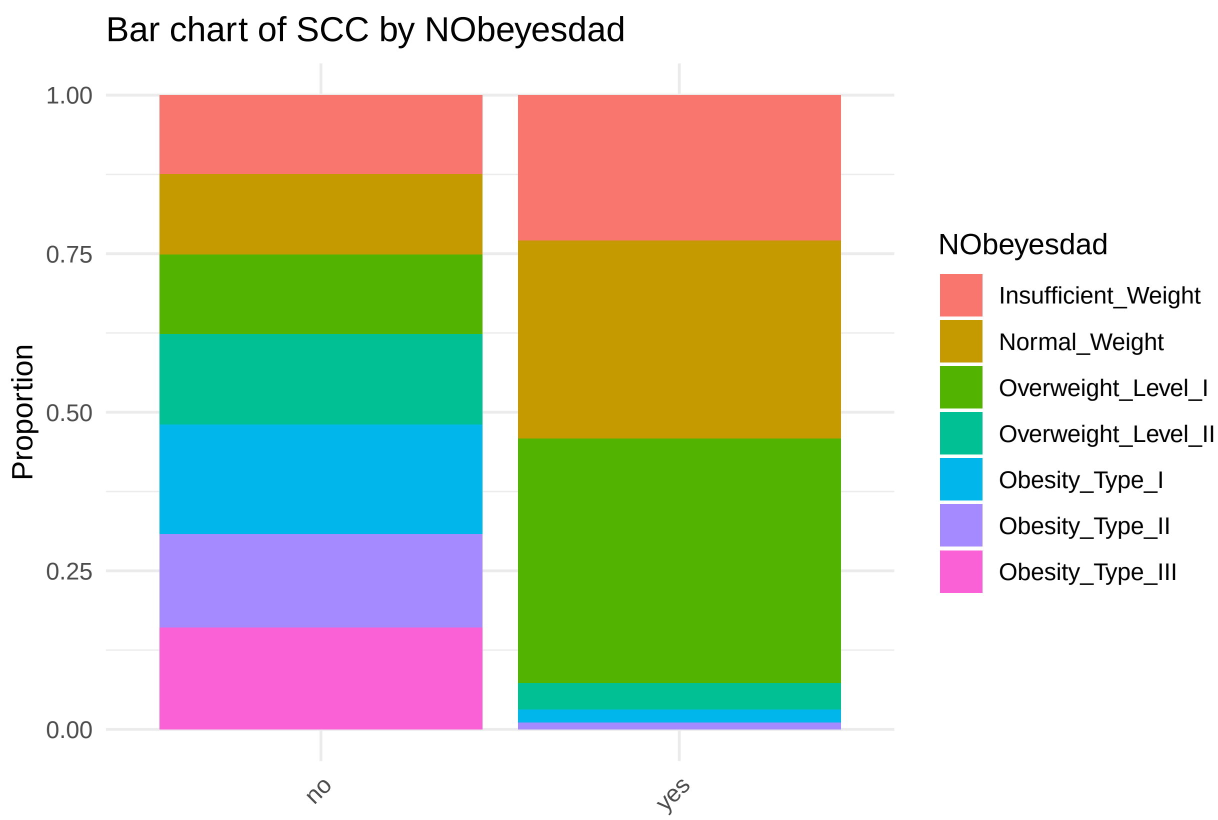


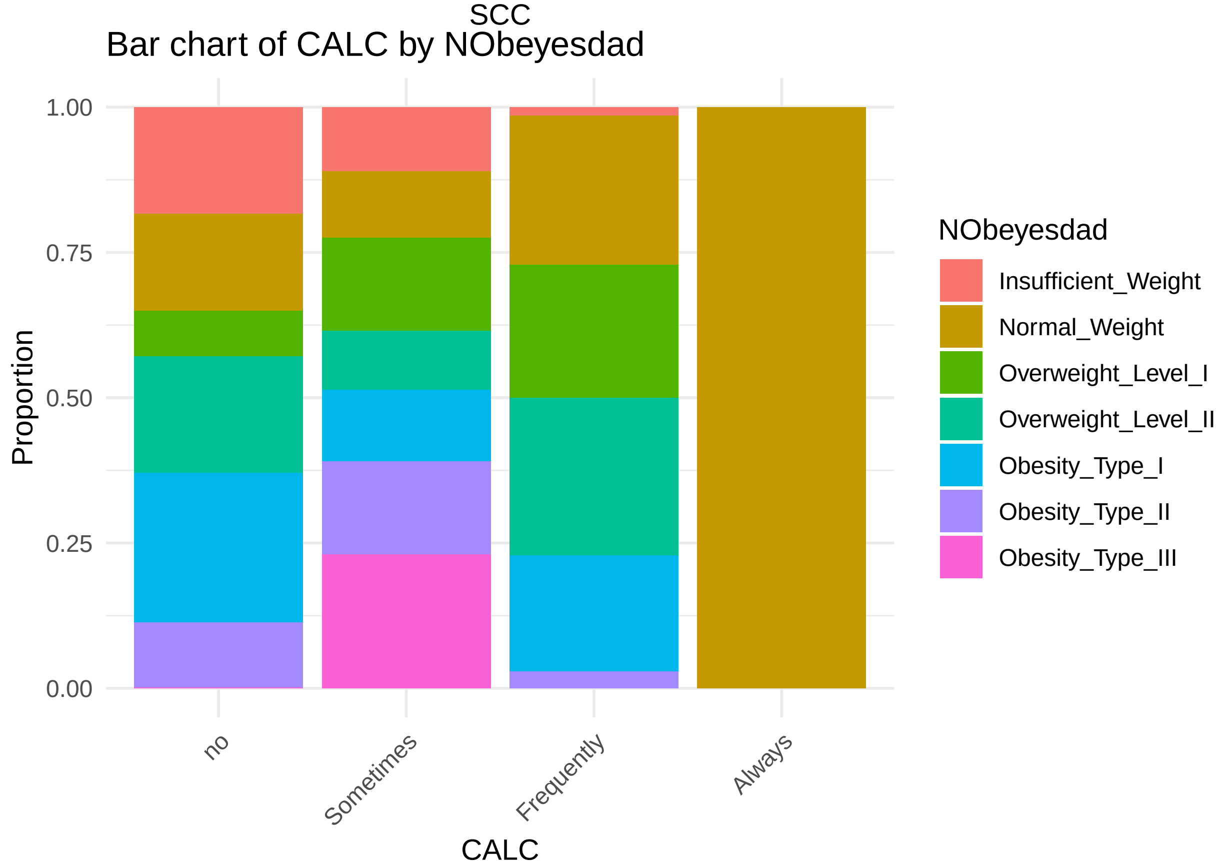
Bar chart of FAVC by NObeyesdad

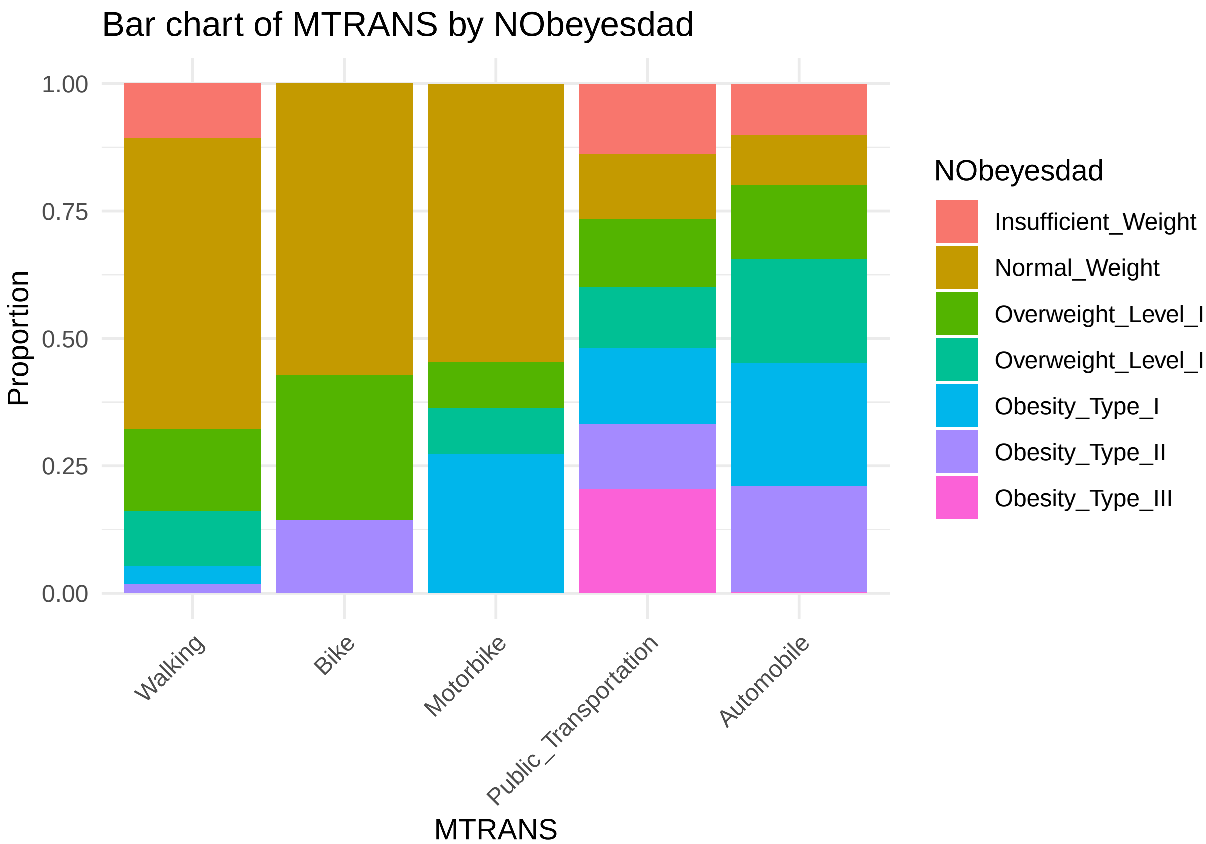



Bar chart of CAEC by NObeyesdad
Proportion
1.00
0.75
0.50
0.25
0.00
no
Sometimes
Frequently
Always
CAEC
NObeyesdad
Insufficient_Weight
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III


Bar chart of SMOKE by NObeyesdad
Proportion
1.00
0.75
0.50
0.25
0.00
no
yes
SMOKE
NObeyesdad
Insufficient_Weight
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III


Bar chart of SCC by NObeyesdad
1.00
0.75
0.50
0.25
0.00
Proportion
no
yes
SCC
NObeyesdad
Insufficient_Weight
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III


Bar chart of CALC by NObeyesdad
1.00
0.75
0.50
0.25
0.00
Proportion
no
Sometimes
Frequently
Always
CALC
NObeyesdad
Insufficient_Weight
Normal_Weight
Overweight_Level_I
Overweight_Level_II
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III



Bar Chart of Gender by NObeyesdad The stacked bar chart depicts the gender-based distribution on 7 *obesity levels*. The vertical axis represents the proportion of obesity classifications within *gender*. For females, *Obesity_Type_III* constitutes the majority, followed by *Insufficient_Weight,* while *Obesity_Type_II* is predominant in males. Regarding the the smallest proportions, females barely had *Overweight_II*, while males were hardly underweight. Overall, the above bar chart clearly illustrates the *gender* disparities within weight distributions, with a higher proportion of females in the lower eight category and a higher proportion of males in the overweight categories.

Bar Chart of Family_history by NObeyesdad The bar chart depicts the *family history of obesity* distribution on 7 *obesity levels*. There is a stark difference between the group with *family history* and the group without. More than 70% of the group with *family history* are overweight or obese, while less than 30% of the group without any *family history* are overweight or obese.

Bar Chart of FAVC by NObeyesdad The bar chart depicts the *FAVC (frequency of consuming high caloric food)* distribution on 7 *obesity levels*. There is a clear difference between the group that does consume high caloric food frequently and the group that does not. More than 75% of the group that responded that they consume high caloric food frequently are overweight or obese, while less than 50% of the group that responded that they don't are overweight or obese.

Bar Chart of CAEC by NObeyesdad The bar chart depicts the *CAEC (consumption of food between meals)* distribution on 7 *obesity levels*. There is a clear difference between the four groups of different *CAEC* levels, but the result is interesting since it is slightly counter-intuitive. *Sometimes* was the most common answer for samples that are overweight or obese, and for the responses *frequently* and *always*, the percentage of overweight and obese samples drop significantly. On the other hand, *frequently* and *always* were the two most common responses for *normal weight* samples.

Bar Chart of SMOKE by NObeyesdad The bar chart depicts the *SMOKE (whether they smoke cigarettes)* distribution on 7 *obesity levels.* The result is interesting. The group that does not smoke have almost a uniform number of instances for every obesity level. However, in the group that does smoke, there are a lot of *normal_weight* and *obesity_type_II* samples.

Bar Chart of SCC by NObeyesdad The bar chart depicts the *SCC (calories consumption monitoring)* distribution on 7 *obesity levels.* For the group that does not monitor calorie intake, the number of instances for every obesity level was near uniform. However, for the group that does track calorie intake, more than 90% were in the range between *insufficient weight* to *overweight_level_I.*

Bar Chart of CALC by NObeyesdad The bar chart depicts the *CALC (consumption of alcohol)* distribution on 7 *obesity levels.* An interesting result is that none of the obese instances were included in the group that always consume alcohol. Most of them were included in the group that does not or only sometimes drink alcohol.

Bar Chart of MTRANS by NObeyesdad The bar chart depicts the *MTRANS (mode of transportation)* distribution on 7 *obesity levels.* For the MTRANS group that usually walks, the overweight to obese instances make up less than 30%. For the groups that usually rides the bicycle or the motorcycle, the overweight to obese instances make up less than 50%. However, for the groups that usually take public transportation or the automobile, the percentage goes up to nearly 75%. We can interpret this as the mode of transportation plays quite an important role in the obesity level, and the more active the mode of transportation is, the less obese the individual is likely to be.

Data Modelling and Analysis

After performing exploratory data analysis, we performed machine learning models to classify the obesity level. We first performed classification by considering the different levels of obesity as categories. Then we performed some regression methods by considering the obesity level as ordinal.

Data Preprocessing

To preprocess the data, we divided the dataset into 80% training set and 20% testing set. To further handle the data, we normalised the numeric variables and converted factors to dummy variables for the categorical variables.

```r
# Preprocessing
set.seed(123)

# Preprocess numerical data: normalise, handle categorical variables, etc.
preprocess_params <- preProcess(data[, -which(names(data) %in% c("BMI"))], method = c("center", "scale")
normalized_data <- predict(preprocess_params, data[, -which(names(data) %in% c("BMI"))])

#train vs test
index <- createDataPartition(normalized_data$NObeyesdad, p = 0.8, list = FALSE)
train_set <- normalized_data[index, ]
test_set <- normalized_data[-index, ]

# X_train & X_test
X_train <- train_set[, -which(names(train_set) == "NObeyesdad")]
X_test <- test_set[, -which(names(test_set) == "NObeyesdad")]
y_train <- train_set[["NObeyesdad"]]
y_test <- test_set[["NObeyesdad"]]
```

```r
# Convert factors to dummy variables
dummies <- dummyVars(" ~ .", data = X_train)
X_train <- predict(dummies, newdata = X_train)
X_test <- predict(dummies, newdata = X_test)
```

Multinomial Logistic Regression

```r
# Multinomial logistic regression
#(instead of linear regression since the target variable NObeyesdad is categorical)

# Convert all categorical variables to factors
categorical_columns <- sapply(data, function(x) is.character(x) || length(unique(x)) < 10)
data[categorical_columns] <- lapply(data[categorical_columns], factor)

# Fit the multinomial logistic regression model
# 'NObeyesdad' is the target, and all other columns are predictors
multinom_model <- multinom(NObeyesdad ~ ., data = data)
```

```
## # weights:  182 (150 variable)
## initial  value 4107.816325
## iter  10 value 3731.010196
## iter  20 value 2595.175304
## iter  30 value 1999.985546
## iter  40 value 1549.084326
## iter  50 value 1259.829931
## iter  60 value 852.850994
## iter  70 value 465.329764
## iter  80 value 227.660863
## iter  90 value 145.213434
## iter 100 value 103.110617
## final  value 103.110617
## stopped after 100 iterations
```

```r
# Summary of the model
summary(multinom_model)
```

```
## Call:
## multinom(formula = NObeyesdad ~ ., data = data)
##
## Coefficients:
##                     (Intercept) GenderMale       Age     Height   Weight
## Normal_Weight          24.12898   9.019703 0.5678375 -119.46866 1.722287
## Overweight_Level_I   -182.83987   6.379296 0.7104995  -93.76631 1.498044
## Overweight_Level_II  -226.05283   7.334804 0.9475116 -193.22331 2.797584
## Obesity_Type_I       -226.86114   6.404880 0.9746652 -325.09458 4.125254
## Obesity_Type_II      -264.96755  61.073223 4.3265853 -587.77838 8.277810
## Obesity_Type_III     -318.30542 -56.600748 2.3246601 -479.52867 7.787516
##                     family_history_with_overweightyes    FAVCyes       FCVC
## Normal_Weight                              -3.9215998   3.448170  -3.409784
## Overweight_Level_I                         -3.5008002   5.318596  -5.780468
## Overweight_Level_II                         0.3689883   1.643515  -7.239650
## Obesity_Type_I                              5.6423957   4.936698  -7.171034
## Obesity_Type_II                           -24.0790751 -27.378347 -13.721485
## Obesity_Type_III                          -25.6704164 -26.874992  19.369521
```

```
##                               NCP CAECSometimes CAECFrequently  CAECAlways
## Normal_Weight        -2.18720866     -9.786192     -13.253382   -3.444783
## Overweight_Level_I   -2.40634380     -7.052549     -13.793132   -5.509582
## Overweight_Level_II  -2.64478872     -9.657530     -16.515767  -10.549070
## Obesity_Type_I       -3.11310138     14.049614       4.208679   11.240685
## Obesity_Type_II     -10.93542357    -80.990719    -104.162274  -60.561584
## Obesity_Type_III      0.02293971    -63.401176      10.431923 -210.697956
##                     SMOKEyes       CH2O     SCCyes         FAF         TUE
## Normal_Weight       8.451060  -4.833812   4.735164  -0.8767041   0.2693296
## Overweight_Level_I  3.979388  -5.344679   9.421321  -1.6452871   0.5217717
## Overweight_Level_II 7.205711  -5.250145   8.844196  -2.5009745   2.0730961
## Obesity_Type_I      7.709162  -6.130897  18.030431  -2.7920804   2.8072726
## Obesity_Type_II     9.959143 -18.288695 -30.573884 -10.5705705   4.2443961
## Obesity_Type_III    8.337026 -25.269798 -29.001114   0.5839267 -19.0485565
##                     CALCSometimes CALCFrequently CALCAlways MTRANSBike
## Normal_Weight           -4.626645     -5.6800022  62.783907   46.34119
## Overweight_Level_I      -3.712233     -5.8957680 -11.267951   47.82871
## Overweight_Level_II     -7.773977     -3.6858005 -41.469290 -142.07219
## Obesity_Type_I          -6.358047      0.7160588 -16.361527  -86.26292
## Obesity_Type_II        -17.026743    -82.1869505   6.880466   57.90373
## Obesity_Type_III        29.796278     41.8000032  10.313250   80.25673
##                     MTRANSMotorbike MTRANSPublic_Transportation
## Normal_Weight              53.55795                   -3.765869
## Overweight_Level_I         49.46278                   -4.728514
## Overweight_Level_II        53.36426                    4.213916
## Obesity_Type_I             58.52646                    4.742811
## Obesity_Type_II           -72.89551                   -2.310561
## Obesity_Type_III          -86.30170                   -7.940910
##                     MTRANSAutomobile       BMI
## Normal_Weight              -2.242739  5.866332
## Overweight_Level_I         -2.844920 13.177482
## Overweight_Level_II         2.607851 17.065439
## Obesity_Type_I              3.754780 19.683792
## Obesity_Type_II           -26.629867 24.126932
## Obesity_Type_III          -24.580097 20.535595
##
## Std. Errors:
##                     (Intercept) GenderMale       Age    Height    Weight
## Normal_Weight         4.9968442   3.169310 0.2345066  8.837240 0.3465287
## Overweight_Level_I    3.8465072   3.480564 0.2524558  6.732803 0.3274132
## Overweight_Level_II   6.0169013   3.655553 0.2655430 10.194053 0.3567939
## Obesity_Type_I        1.6920319   4.265783 0.2947850  2.862268 0.3576754
## Obesity_Type_II       0.9864621   6.189085 0.5003497  1.421703 0.5801805
## Obesity_Type_III      1.2890396   4.449109 0.9591182  1.902815 0.7282484
##                     family_history_with_overweightyes  FAVCyes     FCVC
## Normal_Weight                                1.651003 3.202301 1.667398
## Overweight_Level_I                           1.940000 3.405045 1.978733
## Overweight_Level_II                          2.372489 3.630383 2.233705
## Obesity_Type_I                               3.110138 4.097041 2.616078
## Obesity_Type_II                              8.240967 7.221068 4.513598
## Obesity_Type_III                             5.976591 3.678205 7.878043
##                          NCP CAECSometimes CAECFrequently   CAECAlways
## Normal_Weight       1.130728      3.752709       3.703433 5.286345e+00
## Overweight_Level_I  1.219140      3.512026       3.287176 4.296967e+00
```

```
## Overweight_Level_II 1.313788      3.539225       3.340290 3.464077e+00
## Obesity_Type_I      1.537958      5.153274       5.397400 4.747190e+00
## Obesity_Type_II     4.168705      5.814312       5.142472 4.460149e+00
## Obesity_Type_III    3.002694      3.365435       5.639733 9.653115e-39
##                       SMOKEyes     CH2O       SCCyes      FAF      TUE
## Normal_Weight       4.26044208 1.692256 3.800034e+00 1.096622 1.164496
## Overweight_Level_I  2.93398944 1.864613 4.186554e+00 1.225273 1.423151
## Overweight_Level_II 2.84344649 2.032414 4.487652e+00 1.322640 1.569581
## Obesity_Type_I      4.38313111 2.260161 9.380387e+00 1.613468 1.812899
## Obesity_Type_II     3.94492284 5.601146 3.816024e+00 3.804531 3.651788
## Obesity_Type_III    0.09386118 6.388823 2.532414e-10 6.370490 5.425304
##                     CALCSometimes CALCFrequently    CALCAlways   MTRANSBike
## Normal_Weight            1.814064       3.227314  4.195279e-09 1.729451e+00
## Overweight_Level_I       2.285957       2.959954           NaN 1.729451e+00
## Overweight_Level_II      2.473649       2.829653           NaN          NaN
## Obesity_Type_I           2.738216       3.347825  7.197478e-15 3.194447e-15
## Obesity_Type_II          5.876708       3.952153  5.830840e-27          NaN
## Obesity_Type_III         2.594889       6.824172 5.863410e-158 5.743202e-36
##                     MTRANSMotorbike MTRANSPublic_Transportation
## Normal_Weight          1.587858e+00                    3.561331
## Overweight_Level_I     4.537435e+00                    3.751998
## Overweight_Level_II    3.322124e+00                    4.640777
## Obesity_Type_I         4.880621e+00                    5.087815
## Obesity_Type_II                 NaN                    3.379038
## Obesity_Type_III       9.894522e-38                    6.392352
##                     MTRANSAutomobile       BMI
## Normal_Weight               3.803415 1.0570775
## Overweight_Level_I          4.208075 0.8999161
## Overweight_Level_II         4.709894 0.8409647
## Obesity_Type_I              5.970369 0.9185362
## Obesity_Type_II             3.350241 1.2391354
## Obesity_Type_III            5.781138 2.9954897
##
## Residual Deviance: 206.2212
## AIC: 506.2212
```

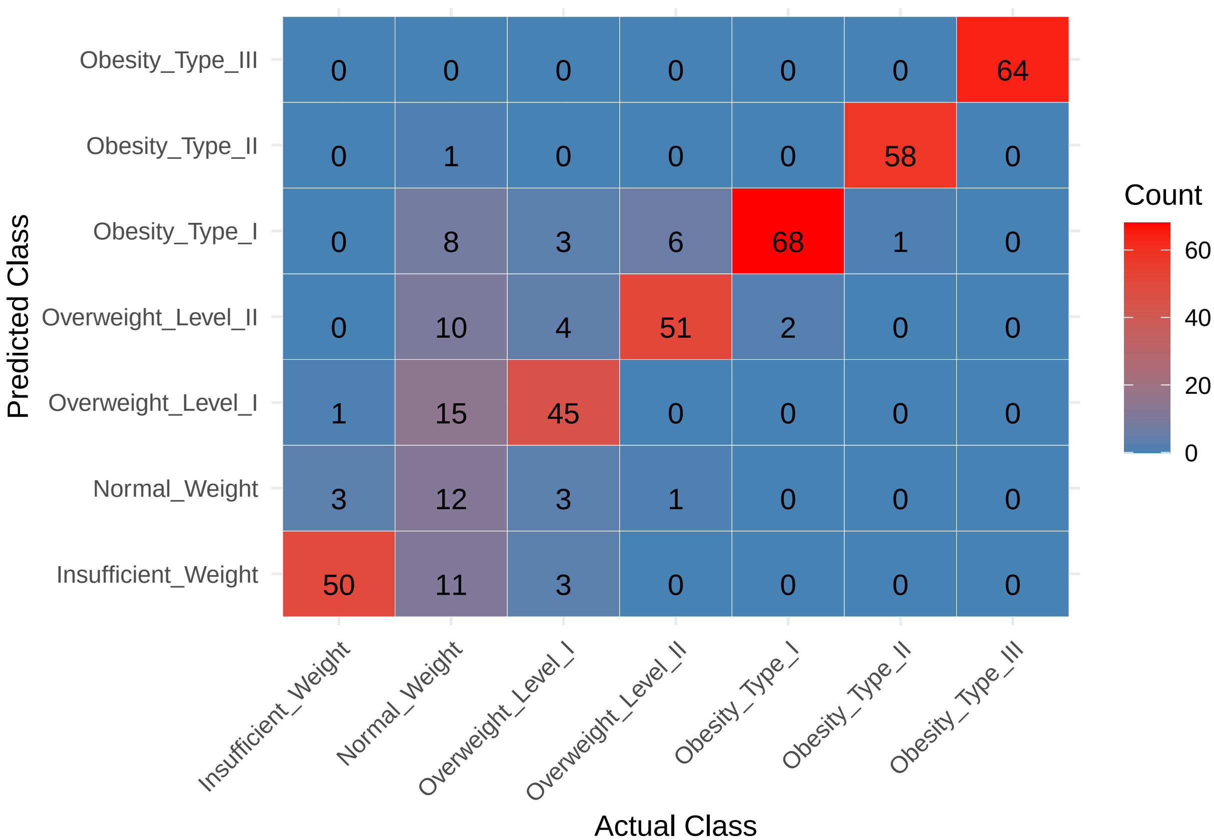
$$DegreesOfFreedom = NumObservations - NumParameters = 2111 - 24 = 2087$$

Since the target variable NObeyesdad is categorical, we performed Multinomial Logistic Regression. The coefficients examine the log-odds impact of predictors such as gender, age, height, weight, and family history of overweight. Positive coefficients represent the increased odds and negative coefficients indicate the decreased odds relative to a reference group. The 'GenderMale' predictor is positively associated with higher weight categories, meaning that males are likely to have increased odds of being in 'Overweight_Level_I' and 'Obesity_Type_I' categories than reference female group. Regarding the residual deviance, a lower value generally indicates a better fit. As residual deviance is significantly smaller than the degree of freedom of 2,087, we assume the model is a very good fit. Likewise, AIC is also preferable once it has lower values. The given value of 506.2212, seems competitive enough, but it has to be further tackled with other models.

K-Nearest Neighbors

```r
# Fit the KNN model using the base 'knn' function since 'knn3' is not a standard function
set.seed(123) # for reproducible results
knn_fit <- knn(train = X_train,
```

```r
                test = X_test,
                cl = y_train, k = 5)


#confusion matrix
confusionMatrix <- table(Predicted = knn_fit, Actual = y_test)
confusion_data <- as.data.frame(as.table(confusionMatrix))

# Create the heatmap
ggplot(confusion_data, aes(x = Actual, y = Predicted, fill = Freq)) +
  geom_tile(color = "white") +
  scale_fill_gradient(low = "steelblue", high = "red") +
  geom_text(aes(label = sprintf("%d", Freq)), vjust = 1) +
  theme_minimal() +
  labs(x = "Actual Class", y = "Predicted Class", fill = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



As obesity levels may have patterns where related features, including eating habits and physical conditions, lead to similar obesity levels, we employed a similarity-based prediction method, KNN. The confusion matrix illustrates the model's predictions across various obesity levels. The diagonal cells represent the correctly classified instances. The model evaluation appears to encounter the issues with some classes not being predicted, which led to zero denominators when generating evaluation metrics. In order to address such issue, micro-averaged metrics were employed as it ensures a more robust evaluation.

```r
# Convert factors to a confusion matrix object
conf_matrix <- confusionMatrix(data = knn_fit, reference = test_set$NObeyesdad)

# Print the overall accuracy
```

```r
cat("Accuracy:", conf_matrix$overall['Accuracy'], "\n")
```

```
## Accuracy: 0.8285714
```

```r
# Calculate micro-average metrics
positive_class <- levels(test_set$NObeyesdad)[1]

micro_precision <- sum(conf_matrix$table[1, 1]) / sum(conf_matrix$table[, 1])
micro_recall <- sum(conf_matrix$table[1, 1]) / sum(conf_matrix$table[1, ])
micro_f1_score <- 2 * (micro_precision * micro_recall) / (micro_precision + micro_recall)

cat("Micro-averaged Precision:", micro_precision, "\n")
```

```
## Micro-averaged Precision: 0.9259259
```

```r
cat("Micro-averaged Recall:", micro_recall, "\n")
```

```
## Micro-averaged Recall: 0.78125
```

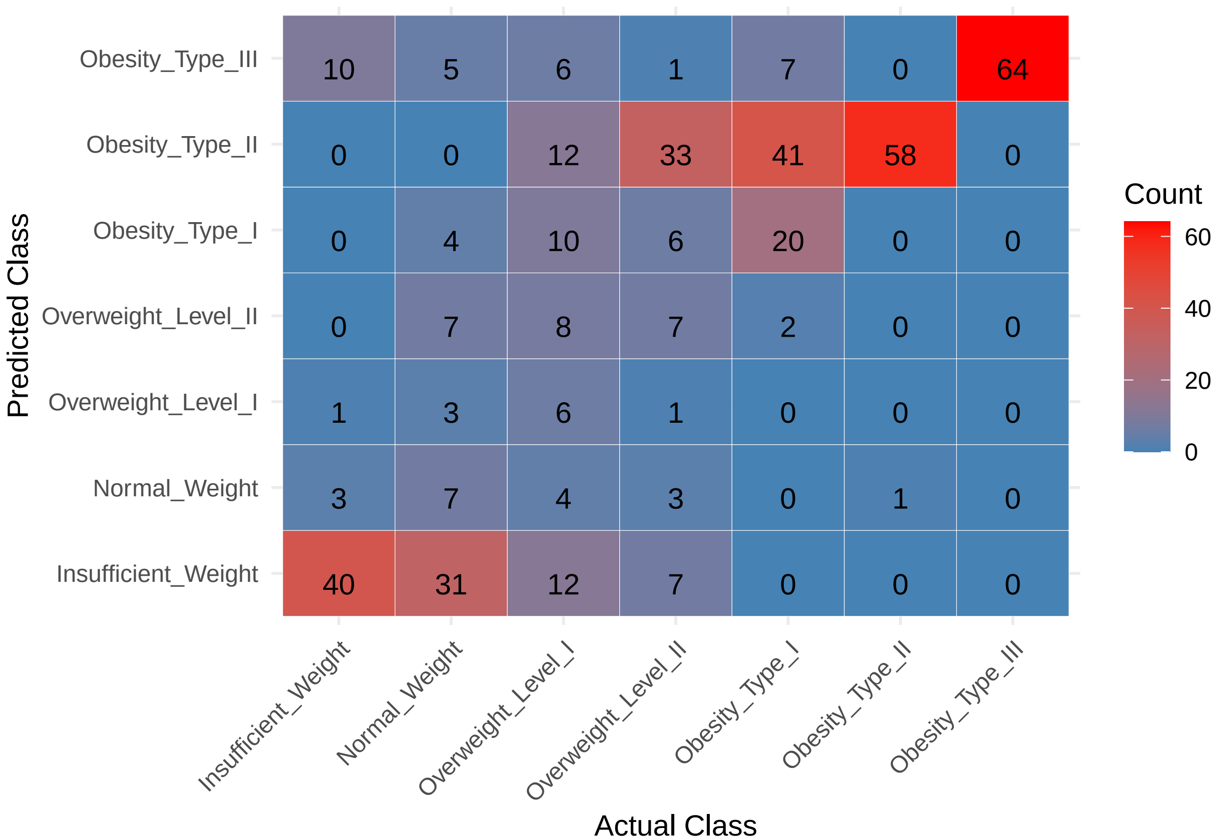
```r
cat("Micro-averaged F1-Score:", micro_f1_score, "\n")
```

```
## Micro-averaged F1-Score: 0.8474576
```

As shown, the metrics used to examine the performance of the model across multiple classes are calculated. The overall accuracy of the model is 0.8285714, with errors occurring between some classes, particularly adjacent classes such as *Normal_Weight* and *Overweight_Level_I*. The micro-average precision is 0.9259259. suggesting that most of the instances predicted as a positive class were likely positive. The micro-average recall is 0.78125. This indicates that the model was able to retrieve 78.12% of actual positive instances. The F1-score. the harmonic mean of Precision and Recall, was 0.8474576. The high recall and F1 score support the model's robustness in identifying the correct classes. The higher values of these metrics confirms the good performance of the model.

Naive Bayes

```r
# Naive Bayes model
nb_model <- naiveBayes(X_train, y_train)
nb_predictions <- predict(nb_model, newdata = X_test)

# Create a confusion matrix
confusionMatrix <- table(Predicted = nb_predictions, Actual = y_test)
confusion_data <- as.data.frame(as.table(confusionMatrix))

# Create the heatmap
ggplot(confusion_data, aes(x = Actual, y = Predicted, fill = Freq)) +
  geom_tile(color = "white") +
  scale_fill_gradient(low = "steelblue", high = "red") +
  geom_text(aes(label = sprintf("%d", Freq)), vjust = 1) +
  theme_minimal() +
  labs(x = "Actual Class", y = "Predicted Class", fill = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Naive Bayes, with its probabilistic approach, can be effective if the features, eating habits and physical conditions, independently contribute to the probability of obesity. The heatmap shows that the model performs well in predicting *Obesity_Type_III*, but has fewer correct predictions for the other classes. This reflects that the model is not differentiating well between similar classes and, therefore, requires further model tuning, or better feature selection, as an improvement.

```
# Evaluate the model performance
accuracy <- sum(diag(confusionMatrix)) / sum(confusionMatrix)
precision <- diag(confusionMatrix) / rowSums(confusionMatrix)
recall <- diag(confusionMatrix) / colSums(confusionMatrix)
f1_score <- 2 * (precision * recall) / (precision + recall)

# Print the metrics
cat("\nAccuracy:", accuracy, "\n")
```

```
##
## Accuracy: 0.4809524
```

```
cat("Precision:", precision, "\n")
```

```
## Precision: 0.4444444 0.3888889 0.5454545 0.2916667 0.5 0.4027778 0.688172
```

```
cat("Recall:", recall, "\n")
```

```
## Recall: 0.7407407 0.122807 0.1034483 0.1206897 0.2857143 0.9830508 1
```

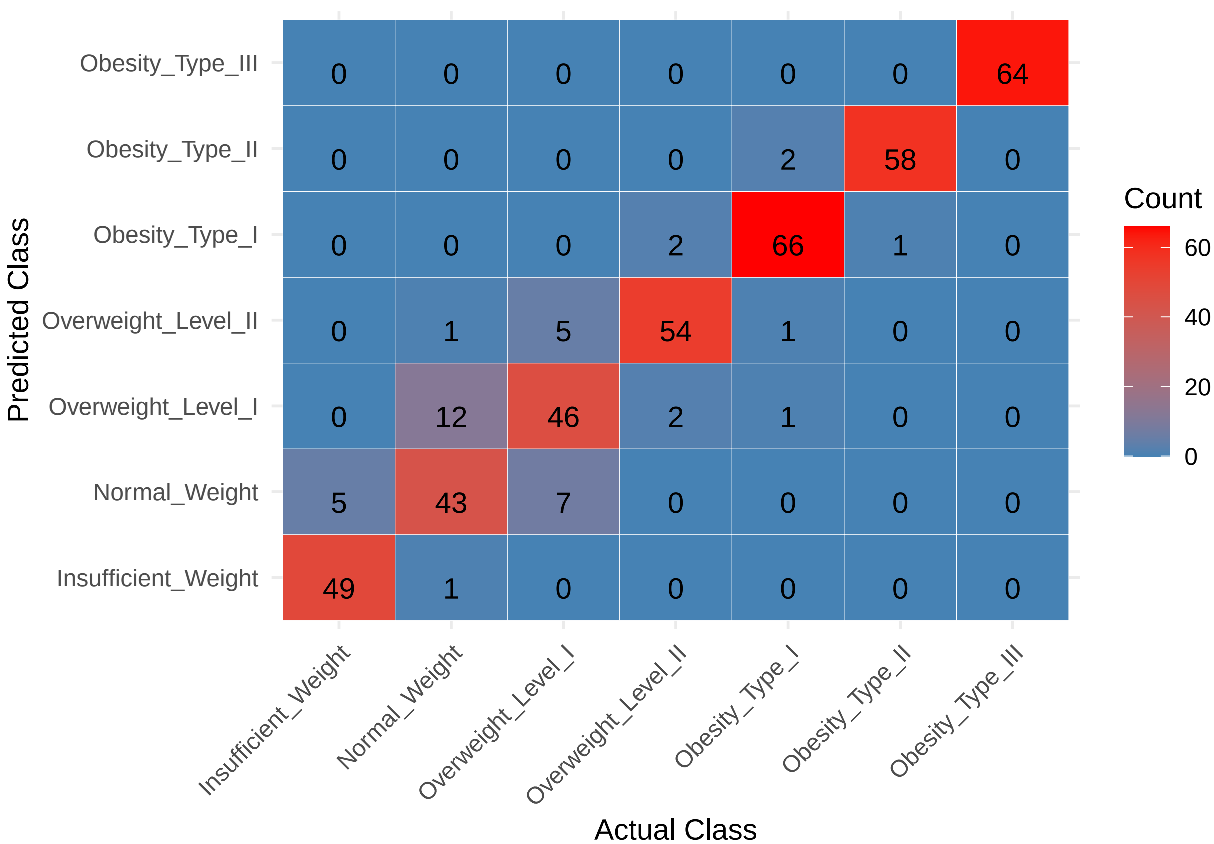
```
cat("F1-Score:", f1_score, "\n")
```

```
## F1-Score: 0.5555556 0.1866667 0.173913 0.1707317 0.3636364 0.5714286 0.8152866
```

The overall accuracy of the model is 0.48059524, a low level of correctness in predictions. Precision values range from 0.291667 to 0.688172, reflecting the ability of the model to identify true positives among all positive predictions. Recall ranges from 0.122807 to 1, examining the success in identifying all actual positives, and the complete recall was shown only for one class. F1-scores are between 0.1707317 and 0.8152866, meaning that the model does not have a balanced performance across the classes.

SVM

```r
# Fit SVM model

svm_model <- svm(x = X_train, y = y_train, type="C-classification", kernel="radial")

# Make predictions on the test set
svm_predictions <- predict(svm_model, newdata = X_test)

# Create a confusion matrix
confusionMatrix <- table(Predicted = svm_predictions, Actual = y_test)
confusion_data <- as.data.frame(as.table(confusionMatrix))


# Create the heatmap
ggplot(confusion_data, aes(x = Actual, y = Predicted, fill = Freq)) +
  geom_tile(color = "white") +
  scale_fill_gradient(low = "steelblue", high = "red") +
  geom_text(aes(label = sprintf("%d", Freq)), vjust = 1) +
  theme_minimal() +
  labs(x = "Actual Class", y = "Predicted Class", fill = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



We also have utilized Support Vector Machine, as it has the complex decision boundaries, which allows to effectively investigate the complex relationships between features and obesity levels, i.e., not being linearly separable. The matrix shows the correct predictions as follows: 49 for *Insufficient_Weight*, 43 for *Normal_Weight*, 46 for *Overweight_Level_I*, 54 for *Overweight_Level_II*, 66 for *Obesity_Type_I*, 58 for *Obesity_Type_II*, and 64 for Obesity_Type_III. The diagonal dominance support our initial assumption that SVM is capable at classifying, validating its application to handle non-linearly separable data like *obesity levels*.

```
# Calculate accuracy
accuracy <- sum(diag(confusionMatrix)) / sum(confusionMatrix)

# Calculate precision and recall for each class
precision <- diag(confusionMatrix) / rowSums(confusionMatrix)
recall <- diag(confusionMatrix) / colSums(confusionMatrix)

# Calculate F1-score for each class
f1_score <- 2 * (precision * recall) / (precision + recall)

# Print the metrics
cat("\nAccuracy:", accuracy, "\n")
```

```
##
## Accuracy: 0.9047619
```

```
cat("Precision:", precision, "\n")
```

```
## Precision: 0.98 0.7818182 0.7540984 0.8852459 0.9565217 0.9666667 1
```

```
cat("Recall:", recall, "\n")
```

```
## Recall: 0.9074074 0.754386 0.7931034 0.9310345 0.9428571 0.9830508 1
```

```
cat("F1-Score:", f1_score, "\n")
```

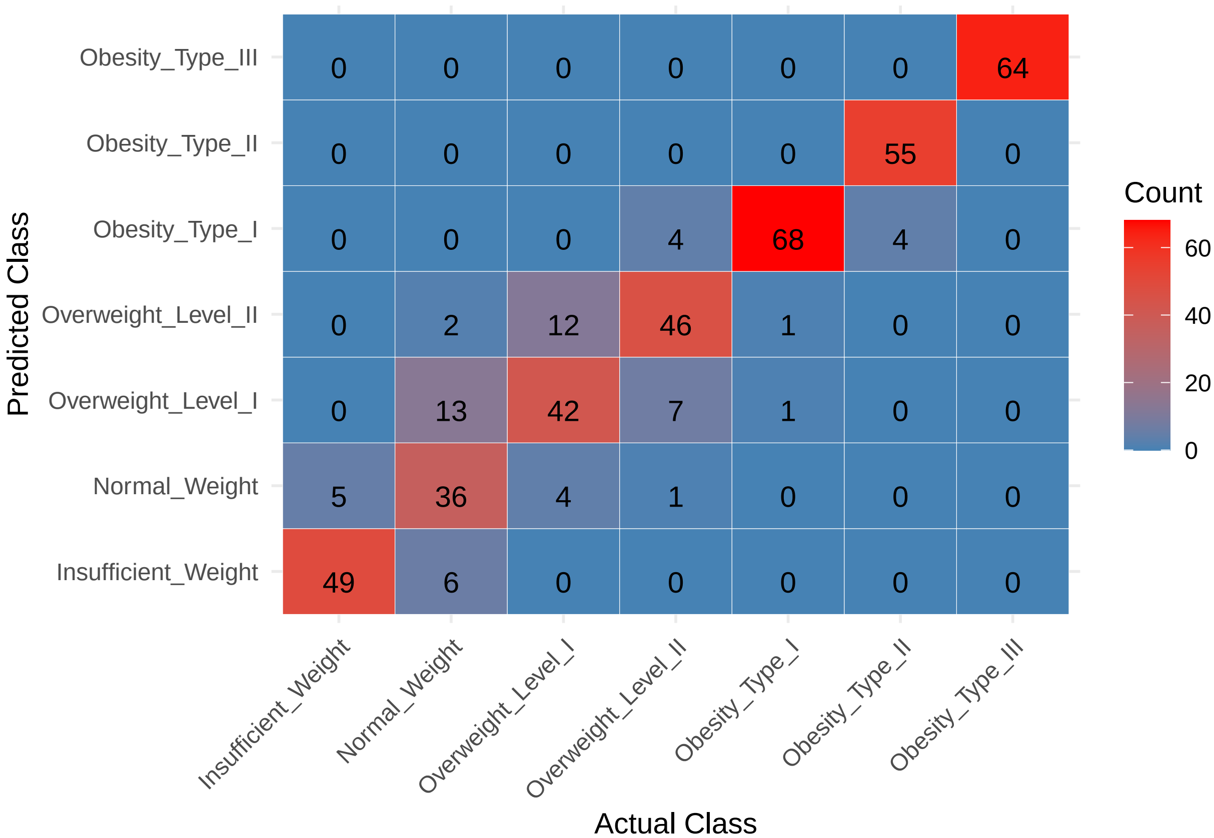
```
## F1-Score: 0.9423077 0.7678571 0.7731092 0.907563 0.9496403 0.9747899 1
```

The accuracy is high with the value of 0.9047619, indicating that the model may correctly predict 90.5% of the outcomes. The precision ranging from 0.7540984 to 1 suggests that the model's prediction will be correct between 75.41% and 100% of the time. The recall values vary from 0.754386. This indicates that the model successfully identifies between 75.44% and 100% of all positive instances for each obesity class. The F1-scores also have shown significance, ranging from 0.7678571 to 1, demonstrating that the model has a robust performance on classifying instances without significant bias or error.

Decision Tree

```
# Fit Decision Tree model
y_train <- as.factor(y_train)
train_data <- data.frame(y_train, X_train)
dt_model <- rpart(y_train ~ ., data = train_data, method="class")
test_data <- data.frame(y_test, X_test)
dt_predictions <- predict(dt_model, newdata = test_data, type = "class")

# Create a confusion matrix
confusionMatrix <- table(Predicted = dt_predictions, Actual = test_data$y_test)
confusion_data <- as.data.frame(as.table(confusionMatrix))

# Create the heatmap
```

```r
ggplot(confusion_data, aes(x = Actual, y = Predicted, fill = Freq)) +
  geom_tile(color = "white") +
  scale_fill_gradient(low = "steelblue", high = "red") +
  geom_text(aes(label = sprintf("%d", Freq)), vjust = 1) +
  theme_minimal() +
  labs(x = "Actual Class", y = "Predicted Class", fill = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



As our dataset includes a lifestyle factors, medical measurements, and demographic information, for ease of interpretation and straightforwardness in visualization, we employed decision tree, as it can handle both numerical and categorical data and are capable of modeling complex relationships. Regarding the confusion matrix the models' performance was high in classifying true positives.

```r
# Calculate accuracy
accuracy <- sum(diag(confusionMatrix)) / sum(confusionMatrix)

# Calculate precision and recall for each class
precision <- diag(confusionMatrix) / rowSums(confusionMatrix)
recall <- diag(confusionMatrix) / colSums(confusionMatrix)

# Calculate F1-score for each class
f1_score <- 2 * (precision * recall) / (precision + recall)

# Print the metrics
cat("\nAccuracy:", accuracy, "\n")
```

```
##
## Accuracy: 0.8571429
```

```
cat("Precision:", precision, "\n")
```

```
## Precision: 0.8909091 0.7826087 0.6666667 0.7540984 0.8947368 1 1
```

```
cat("Recall:", recall, "\n")
```

```
## Recall: 0.9074074 0.6315789 0.7241379 0.7931034 0.9714286 0.9322034 1
```

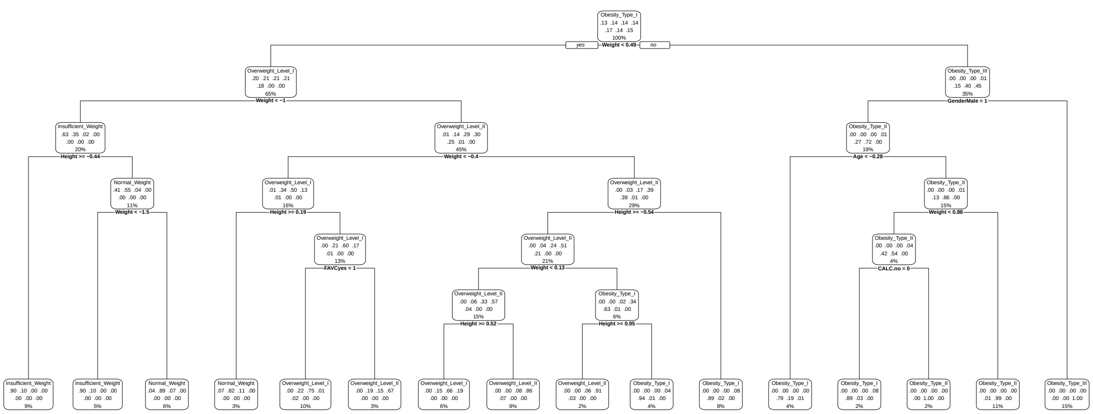
```
cat("F1-Score:", f1_score, "\n")
```

```
## F1-Score: 0.8990826 0.6990291 0.6942149 0.7731092 0.9315068 0.9649123 1
```

The overall accuracy of the model is 0.8571429, so 85.71% of the outcomes are correctly predicted. Precision values for individual classes varies from 0.6666667 and 1, proving a high likelihood that the model's positive predictions are correct. Recall values, which range from 0.6315789 to 1, indicates that the model is capable of detecting true positives. F1-Score's presence in between 0.6942149 and 1 demonstrates that the model is well performing.

Visualization of Decision Tree

```
# Visualize the tree
library(rpart.plot)
rpart.plot(dt_model)
```

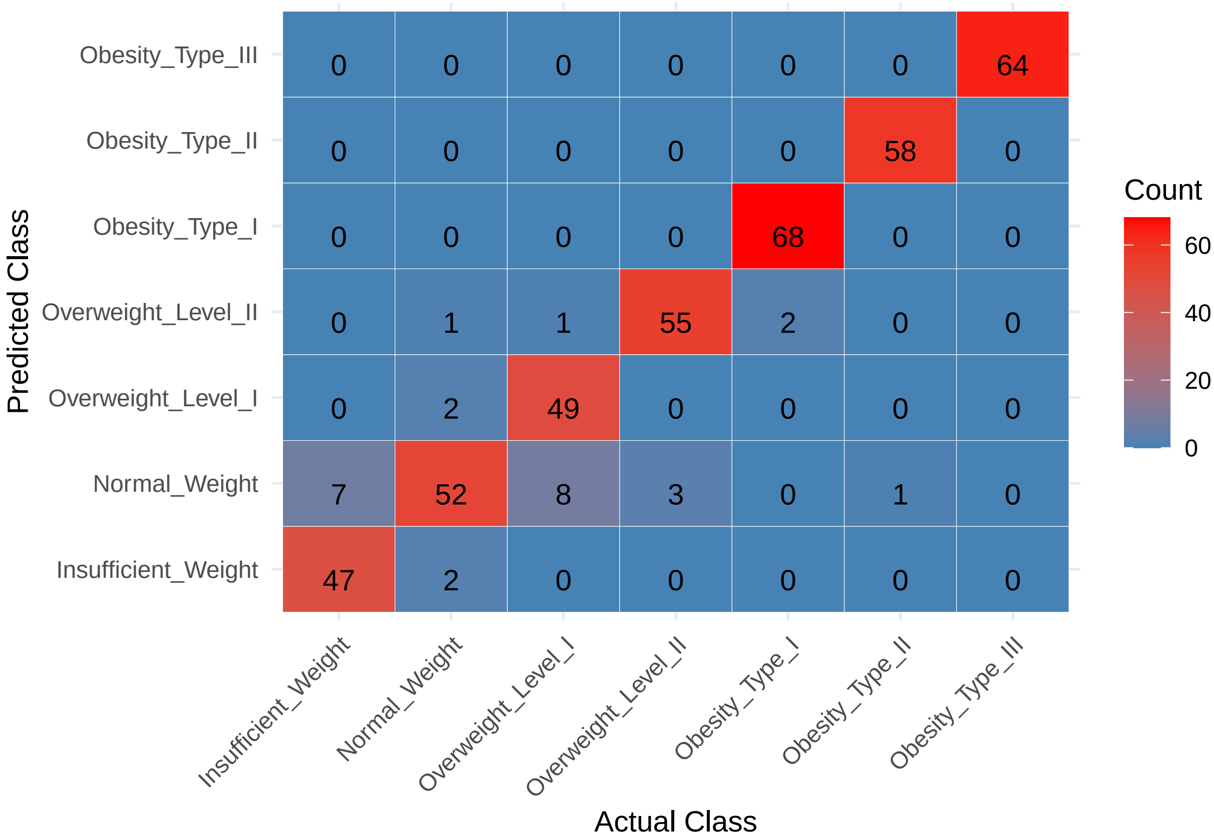


Random Forest

However, decision tree has a possibility of overfitting, which may create the model that is too complex and fail to generalize. Hence, we have further applied random forest.

```
rf_model <- randomForest(y_train ~ ., data = train_data, ntree=500, importance=TRUE)
rf_predictions <- predict(rf_model, newdata = test_data)

# Create a confusion matrix
confusionMatrix <- table(Predicted = rf_predictions, Actual = test_data$y_test)
confusion_data <- as.data.frame(as.table(confusionMatrix))

# Create the heatmap
ggplot(confusion_data, aes(x = Actual, y = Predicted, fill = Freq)) +
  geom_tile(color = "white") +
  scale_fill_gradient(low = "steelblue", high = "red") +
  geom_text(aes(label = sprintf("%d", Freq)), vjust = 1) +
  theme_minimal() +
```

```r
  labs(x = "Actual Class", y = "Predicted Class", fill = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Random forest can exhibit complex relationships without facing the risk of overfitting. The diagonal cells within the confusion matrix show a high number of correct predictions without misclassifications, supported by zero values in most of the off-diagonal cells. This heatmap indicates an exemplary classification performance.

```r
accuracy <- sum(diag(confusionMatrix)) / sum(confusionMatrix)
precision <- diag(confusionMatrix) / rowSums(confusionMatrix)
recall <- diag(confusionMatrix) / colSums(confusionMatrix)
f1_score <- 2 * (precision * recall) / (precision + recall)

# Print the metrics
cat("\nAccuracy:", accuracy, "\n")
```

```
##
## Accuracy: 0.9357143
```

```r
cat("Precision:", precision, "\n")
```

```
## Precision: 0.9591837 0.7323944 0.9607843 0.9322034 1 1 1
```

```r
cat("Recall:", recall, "\n")
```

```
## Recall: 0.8703704 0.9122807 0.8448276 0.9482759 0.9714286 0.9830508 1
```

```r
cat("F1-Score:", f1_score, "\n")
```

```
## F1-Score: 0.9126214 0.8125 0.8990826 0.9401709 0.9855072 0.991453 1
```

The accracy of the model is 0.9380952, followed by precision ranging from 0.7361111 to 1, and recall varying from 0.862089 and 1. The F1-score is also high from 0.8217054 to 1. These statistics reflect the model's high performance, indicating that Random Forest's ensemble nature played a role in increasing the performance from a single Decision Tree model.

Random Forest Variable Importance Table

```r
# Get variable importance
importance(rf_model)
```

```
##                                 Insufficient_Weight Normal_Weight
## GenderFemale                              15.820268   10.45110856
## GenderMale                                15.793933   10.00109971
## Age                                       30.507102    9.27014904
## Height                                    25.153467   19.71832954
## Weight                                    62.499333   48.21815557
## family_history_with_overweightno          18.563805    9.19098224
## family_history_with_overweightyes         20.129553   10.59588497
## FAVCno                                    10.914435    2.68317399
## FAVCyes                                    9.983628    1.69771134
## FCVC                                      24.136569   14.49571601
## NCP                                       23.972012    7.08644356
## CAEC.no                                    4.614357   -2.97842371
## CAEC.Sometimes                            19.156931   14.27840085
## CAEC.Frequently                           23.652244    4.39124430
## CAEC.Always                                5.201334    8.39877066
## SMOKEno                                    2.914411   -1.37932119
## SMOKEyes                                   3.011182   -1.23516252
## CH2O                                      17.512381   15.40220241
## SCCno                                      5.293858   -0.51434946
## SCCyes                                     4.350459    0.03539818
## FAF                                       19.540080    9.14218567
## TUE                                       23.407011   12.41310603
## CALC.no                                   17.155163   -1.36587455
## CALC.Sometimes                            17.595069    0.45210834
## CALC.Frequently                            4.907255   -3.82400654
## CALC.Always                                0.000000    0.00000000
## MTRANS.Walking                             2.252525    1.93270480
## MTRANS.Bike                                0.000000    1.81522692
## MTRANS.Motorbike                           1.001002   -1.77276625
## MTRANS.Public_Transportation              18.638549    0.42538784
## MTRANS.Automobile                         17.855403    0.13815315
##                                 Overweight_Level_I Overweight_Level_II
## GenderFemale                           15.48420466          15.6203099
## GenderMale                             16.26829689          16.2052755
## Age                                    35.04648072          36.1402728
## Height                                 34.73346679          36.7897856
## Weight                                 51.09125925          50.7728536
## family_history_with_overweightno       16.23443546          16.2318208
## family_history_with_overweightyes      16.51157635          18.7403210
## FAVCno                                 12.55628630          19.1184643
## FAVCyes                                14.31636913          19.7383258
## FCVC                                   23.61981882          23.1538981
## NCP                                    25.39763353          23.3358015
```

```
## CAEC.no                                 16.46891615          5.0474508
## CAEC.Sometimes                          15.11719262         16.5338649
## CAEC.Frequently                         16.18503133         11.9076555
## CAEC.Always                              2.92780960          3.6774389
## SMOKEno                                  3.47564600          2.8778407
## SMOKEyes                                 3.40974015          2.1295299
## CH2O                                    22.74126414         18.9926314
## SCCno                                   14.86613159          7.1592132
## SCCyes                                  14.04342654          6.7750907
## FAF                                     19.97864674         13.5117589
## TUE                                     19.03552077         22.5785806
## CALC.no                                 19.00431860         18.5186487
## CALC.Sometimes                          21.72893215         20.2624312
## CALC.Frequently                          5.82776765          5.0936889
## CALC.Always                              0.00000000          0.0000000
## MTRANS.Walking                           2.86060289          0.4856318
## MTRANS.Bike                              0.03322163          1.6045533
## MTRANS.Motorbike                        -0.51569702          1.3223202
## MTRANS.Public_Transportation            16.71999867         17.0441257
## MTRANS.Automobile                       14.81554480         15.7217021
##                                  Obesity_Type_I Obesity_Type_II
## GenderFemale                         19.0527005       17.694384
## GenderMale                           19.2286213       18.308611
## Age                                  33.9620716       31.563511
## Height                               39.9976394       17.183534
## Weight                               61.5003793       74.819428
## family_history_with_overweightno     16.4853250       12.171266
## family_history_with_overweightyes    19.4645333       13.686826
## FAVCno                               14.7119105        6.762124
## FAVCyes                              15.2970654        7.625000
## FCVC                                 30.4221354       23.919085
## NCP                                  25.8297428       21.978274
## CAEC.no                               5.8637090        3.177726
## CAEC.Sometimes                       15.5447733       12.414004
## CAEC.Frequently                      15.1312776        9.807941
## CAEC.Always                           2.3186450        4.197999
## SMOKEno                               1.5626912        2.618901
## SMOKEyes                              0.1681977        1.213314
## CH2O                                 23.6402430       21.701552
## SCCno                                 4.6495385        3.339259
## SCCyes                                5.0273144        2.898204
## FAF                                  21.7351872       19.628102
## TUE                                  23.0263731       14.528877
## CALC.no                              18.4861625       12.928007
## CALC.Sometimes                       17.2933593       13.570181
## CALC.Frequently                       7.4605555        5.778027
## CALC.Always                           0.0000000        0.000000
## MTRANS.Walking                        3.5054380        4.313283
## MTRANS.Bike                           0.0000000       -1.001002
## MTRANS.Motorbike                      1.4170325        1.736723
## MTRANS.Public_Transportation         17.1444631        9.894842
## MTRANS.Automobile                    17.2733407        7.862435
##                                  Obesity_Type_III MeanDecreaseAccuracy
## GenderFemale                            18.891241           21.4248161
```

```
## GenderMale                              21.193571          23.0205872
## Age                                     13.665975          43.8260087
## Height                                   9.961780          45.2164352
## Weight                                  46.756845          83.7143306
## family_history_with_overweightno        10.916706          19.8602921
## family_history_with_overweightyes       12.894637          21.7552280
## FAVCno                                   8.187647          20.6132076
## FAVCyes                                  8.204105          20.9608830
## FCVC                                    24.376175          33.6138624
## NCP                                     14.340426          33.9047344
## CAEC.no                                  2.023319          15.8859968
## CAEC.Sometimes                          11.343580          22.4568525
## CAEC.Frequently                         10.356369          22.6325264
## CAEC.Always                              3.036694          11.4691664
## SMOKEno                                  3.302156           5.5567323
## SMOKEyes                                 2.853187           4.1146952
## CH2O                                     8.252918          37.6901316
## SCCno                                    4.577544          14.7578182
## SCCyes                                   4.965624          14.5435171
## FAF                                      8.870727          31.4685198
## TUE                                     13.691374          31.3590665
## CALC.no                                 10.741395          21.7985604
## CALC.Sometimes                          12.010579          22.9611036
## CALC.Frequently                          2.714985          10.1187951
## CALC.Always                              0.000000           0.0000000
## MTRANS.Walking                           2.261412           6.1706246
## MTRANS.Bike                              0.000000           1.9215979
## MTRANS.Motorbike                         1.001002           0.3650187
## MTRANS.Public_Transportation             9.617012          23.1929467
## MTRANS.Automobile                        7.077731          23.9239795
##                                   MeanDecreaseGini
## GenderFemale                            47.9659555
## GenderMale                              58.9770981
## Age                                    121.7873777
## Height                                 112.3738593
## Weight                                 387.1355557
## family_history_with_overweightno        28.0128242
## family_history_with_overweightyes       31.9114483
## FAVCno                                  16.7345908
## FAVCyes                                 17.9697415
## FCVC                                   109.0785861
## NCP                                     67.0152403
## CAEC.no                                  6.9281146
## CAEC.Sometimes                          28.4569047
## CAEC.Frequently                         27.1743817
## CAEC.Always                              4.9437796
## SMOKEno                                  2.2884452
## SMOKEyes                                 2.3434088
## CH2O                                    59.2478587
## SCCno                                    6.6848151
## SCCyes                                   6.3394498
## FAF                                     55.7182876
## TUE                                     56.8127624
## CALC.no                                 26.3694140
```

```
## CALC.Sometimes                          30.5556411
## CALC.Frequently                          5.8496068
## CALC.Always                              0.0000000
## MTRANS.Walking                           4.3565558
## MTRANS.Bike                              0.9574048
## MTRANS.Motorbike                         0.9550148
## MTRANS.Public_Transportation            21.6083040
## MTRANS.Automobile                       18.2908844
```

Based on the variable importance from the random forest, we have done the analysis based on Mean Decrese Accuracy, or MDA, and Mean Decrease Gini, or MDG. MDA measures the decrease in model accuracy when the values of a particular variable are permuted randomly, and MDG measures the contribution of each feature to the homogeneity of the nodes and leaves in the model. Therefore, MDA and MDG, together examines the significance of the feature. According to the table, *weight* has shown highest importance scores in both MDA and MDG, suggesting that it is likely a direct indicator of an obesity. The *Age*. *Height*, *FCVC* (frequency of consumption of vegetables), *NCP* (number of main meals), *CH2O*, (water intake), *FAF* (physical activity frequency), and *TUE* (time using electronic devices), have shown the moderate importance. The variables, including *SMOKEno*, *SMOKEyes*, *SCCno* (calories consumption monitoring), *SCCyes*, anhd *MTRANS* (different types of transportation), particulary for walking, bike, and motorbike, shows relatively lower importance, suggesting that these factors may have less direct influence on the target variable.

Converting Obesity Levels to Numeric Scale

```r
data$NObeyesdad <- factor(data$NObeyesdad, levels = c("Insufficient_Weight", "Normal_Weight",
                                                      "Overweight_Level_I", "Overweight_Level_II",
                                                      "Obesity_Type_I", "Obesity_Type_II",
                                                      "Obesity_Type_III"), ordered = TRUE)

data$NObeyesdad_numeric <- as.numeric(data$NObeyesdad)
head(data)
```

```
##   Gender Age Height Weight family_history_with_overweight FAVC FCVC NCP
## 1 Female  21   1.62   64.0                            yes   no    2   3
## 2 Female  21   1.52   56.0                            yes   no    3   3
## 3   Male  23   1.80   77.0                            yes   no    2   3
## 4   Male  27   1.80   87.0                             no   no    3   3
## 5   Male  22   1.78   89.8                             no   no    2   1
## 6   Male  29   1.62   53.0                             no  yes    2   3
##        CAEC SMOKE CH2O SCC FAF TUE       CALC                MTRANS
## 1 Sometimes    no    2  no   0   1         no Public_Transportation
## 2 Sometimes   yes    3 yes   3   0  Sometimes Public_Transportation
## 3 Sometimes    no    2  no   2   1 Frequently Public_Transportation
## 4 Sometimes    no    2  no   2   0 Frequently               Walking
## 5 Sometimes    no    2  no   0   0  Sometimes Public_Transportation
## 6 Sometimes    no    2  no   0   0  Sometimes            Automobile
##            NObeyesdad      BMI NObeyesdad_numeric
## 1       Normal_Weight 24.38653                  2
## 2       Normal_Weight 24.23823                  2
## 3       Normal_Weight 23.76543                  2
## 4  Overweight_Level_I 26.85185                  3
## 5 Overweight_Level_II 28.34238                  4
## 6       Normal_Weight 20.19509                  2
```

To further analyze the influence of diverse factors on obesity level, we have converted the obesity levels from categorical to numerical scale to progress the regressions.

Linear Regression

```r
# Explanatory variables (categorical data) to factor type
data <- data %>%
  mutate(Gender = as.factor(Gender),
         family_history_with_overweight = as.factor(family_history_with_overweight),
         FAVC = as.factor(FAVC),
         CAEC = as.factor(CAEC),
         SMOKE = as.factor(SMOKE),
         CALC = as.factor(CALC),
         SCC = as.factor(SCC),
         MTRANS = as.factor(MTRANS))

model <- lm(NObeyesdad_numeric ~ Gender + Age + Height + family_history_with_overweight + FAVC + FCVC +
              SCC + FAF + TUE + CALC  + MTRANS + BMI, data = data)

summary(model)
```

```
##
## Call:
## lm(formula = NObeyesdad_numeric ~ Gender + Age + Height + family_history_with_overweight +
##     FAVC + FCVC + NCP + CAEC + SMOKE + CH2O + SCC + FAF + TUE +
##     CALC + MTRANS + BMI, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.95453 -0.22389 -0.00186  0.24798  1.27790
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -3.6559527  0.2162279 -16.908  < 2e-16 ***
## GenderMale                      0.0506229  0.0231104   2.190 0.028599 *
## Age                             0.0204886  0.0018618  11.004  < 2e-16 ***
## Height                          0.1270187  0.1327395   0.957 0.338728
## family_history_with_overweightyes  0.1362891  0.0265500   5.133 3.11e-07 ***
## FAVCyes                        -0.0180403  0.0284097  -0.635 0.525496
## FCVC                           -0.0685955  0.0173743  -3.948 8.14e-05 ***
## NCP                            -0.0113701  0.0113102  -1.005 0.314871
## CAECSometimes                   0.3144467  0.0584846   5.377 8.44e-08 ***
## CAECFrequently                  0.0597352  0.0636714   0.938 0.348261
## CAECAlways                      0.1274053  0.0778040   1.638 0.101674
## SMOKEyes                        0.0114032  0.0592189   0.193 0.847322
## CH2O                           -0.0051678  0.0145759  -0.355 0.722967
## SCCyes                         -0.0007651  0.0419126  -0.018 0.985437
## FAF                            -0.0560764  0.0109245  -5.133 3.11e-07 ***
## TUE                            -0.0270031  0.0148053  -1.824 0.068313 .
## CALCSometimes                  -0.0627562  0.0196993  -3.186 0.001465 **
## CALCFrequently                 -0.0109437  0.0491381  -0.223 0.823781
## CALCAlways                     -0.1289268  0.3859877  -0.334 0.738400
## MTRANSBike                     -0.0863687  0.1533167  -0.563 0.573267
## MTRANSMotorbike                 0.0663947  0.1273046   0.522 0.602045
## MTRANSPublic_Transportation     0.2013779  0.0539007   3.736 0.000192 ***
## MTRANSAutomobile               -0.0313041  0.0578719  -0.541 0.588621
## BMI                             0.2302690  0.0014660 157.072  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3802 on 2087 degrees of freedom
## Multiple R-squared:  0.9637, Adjusted R-squared:  0.9633
## F-statistic:  2410 on 23 and 2087 DF,  p-value: < 2.2e-16
```

The Multiple R-squared is 0.9637, indicating that approximately 96.37% of the obesity variability can be explained by the model. Likewise, the Adjusted R-squared of 0.9633, suggests that the model has a very good fit. Regarding the coefficients, as the asterisks represent the level of statistical significance, the significant independent variables are as follows: *GenderMale, Age, family_history_with_overweightyes, FCVC, CAECSometimes, FAF, CALCSometimes, MTRANSPublic_Transportation, and BMI.*

Polynomical Regression

```r
# Polynomial Regression
model_poly <- lm(NObeyesdad_numeric ~ Gender + Age + I(Age^2) + Height + BMI + I(BMI^2) + family_histor
                  FAVC + FCVC + NCP + CAEC + SMOKE + CH2O + SCC + FAF + TUE + CALC + MTRANS, data = da

summary(model_poly)
```

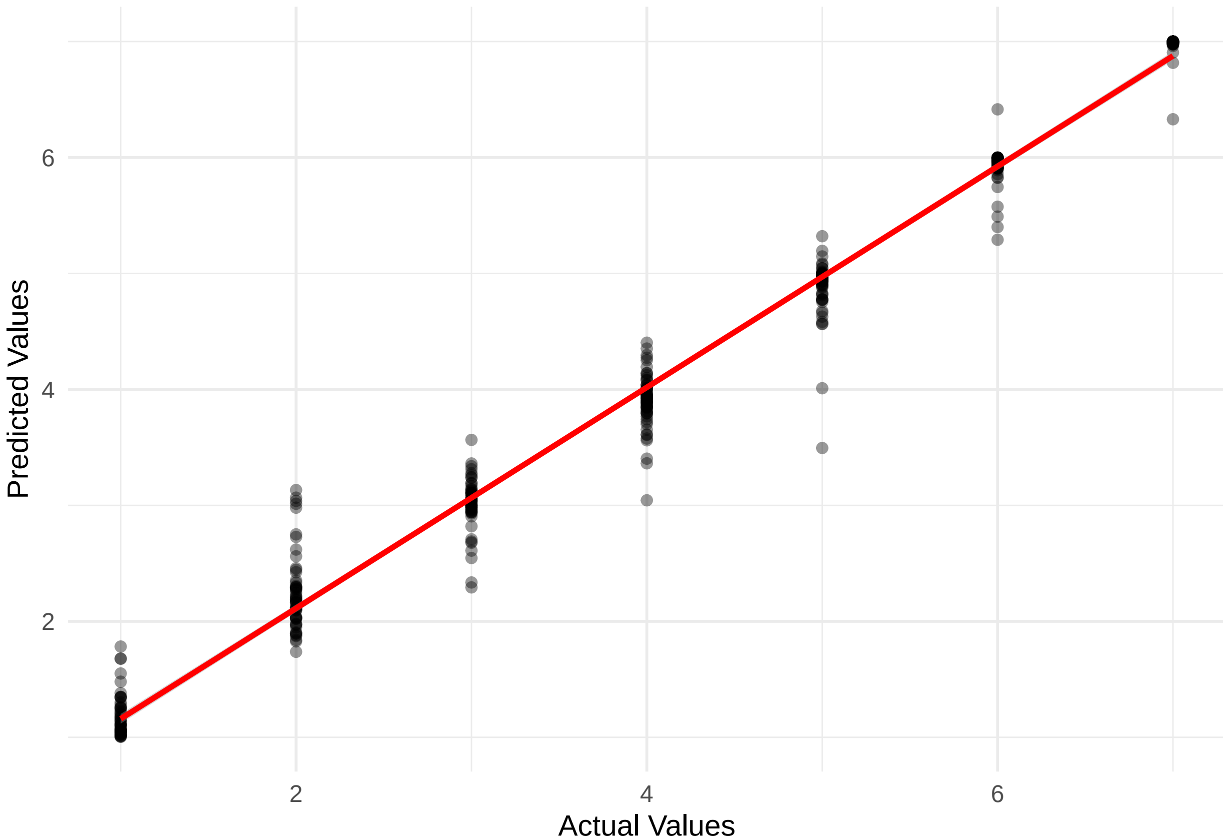
```
##
## Call:
## lm(formula = NObeyesdad_numeric ~ Gender + Age + I(Age^2) + Height +
##     BMI + I(BMI^2) + family_history_with_overweight + FAVC +
##     FCVC + NCP + CAEC + SMOKE + CH2O + SCC + FAF + TUE + CALC +
##     MTRANS, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.24072 -0.20920  0.01958  0.22709  1.79185
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -6.9510163  0.2763559 -25.152  < 2e-16 ***
## GenderMale                    -0.0837832  0.0228873  -3.661 0.000258 ***
## Age                            0.0536840  0.0081972   6.549 7.27e-11 ***
## I(Age^2)                      -0.0006751  0.0001329  -5.079 4.12e-07 ***
## Height                         0.5094604  0.1258220   4.049 5.33e-05 ***
## BMI                            0.3762100  0.0092823  40.530  < 2e-16 ***
## I(BMI^2)                      -0.0023963  0.0001487 -16.110  < 2e-16 ***
## family_history_with_overweightyes  0.0331750  0.0254490   1.304 0.192517
## FAVCyes                        0.0094578  0.0265541   0.356 0.721749
## FCVC                           0.0001354  0.0168418   0.008 0.993584
## NCP                            0.0273545  0.0107831   2.537 0.011260 *
## CAECSometimes                  0.3220862  0.0545838   5.901 4.21e-09 ***
## CAECFrequently                 0.1447566  0.0596726   2.426 0.015357 *
## CAECAlways                     0.1461152  0.0726188   2.012 0.044339 *
## SMOKEyes                       0.0009582  0.0552758   0.017 0.986171
## CH2O                          -0.0103112  0.0136030  -0.758 0.448531
## SCCyes                        -0.0339209  0.0393857  -0.861 0.389199
## FAF                           -0.0440541  0.0102271  -4.308 1.73e-05 ***
## TUE                           -0.0079639  0.0138554  -0.575 0.565497
## CALCSometimes                 -0.0498368  0.0186502  -2.672 0.007595 **
```

```
## CALCFrequently                    -0.0674042  0.0460257  -1.464 0.143211
## CALCAlways                        -0.0445808  0.3601741  -0.124 0.901505
## MTRANSBike                         0.0058487  0.1432081   0.041 0.967427
## MTRANSMotorbike                    0.0605857  0.1188087   0.510 0.610145
## MTRANSPublic_Transportation        0.2158566  0.0503467   4.287 1.89e-05 ***
## MTRANSAutomobile                   0.0018600  0.0541778   0.034 0.972616
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3548 on 2085 degrees of freedom
## Multiple R-squared:  0.9684, Adjusted R-squared:  0.9681
## F-statistic:  2559 on 25 and 2085 DF,  p-value: < 2.2e-16
```

For polynomial regression, to capture the nonlinear relationships and increase the model flexibility, regarding that relationship between *Age* and *BMI* with *Obesity* may illustrate the U-shaped pattern, obesity being increased to a certain age and then decrease, we included the terms *Age^2* and *BMI^2*. The Multiple R-squared is 0.9684, indicating that approximately 96.84% of the obesity variability can be explained by the model. The Adjusted R-squared of 0.9681 portrays that the model has a very good fit. Regarding the coefficients, as the asterisks represent the level of statistical significance, the significant independent variables are as follows: *GenderMale, Age, Age^2, Height, BMI, BMI^2, NCP, CAEC (Sometimes, Frequently, and Always), FAF, CALCSometimes, MTRANSPublic_Transportation.* As the coefficient of *Age^2* is negative, the polynomial regression further captures the nonlinear effect of age.

Random Forest with Numeric Obesity Level

```
# Preprocess numerical data
prep_p_num <- preProcess(data[, -which(names(data) %in% c("NObeyesdad", "BMI"))], method = c("center",
data_normalized_num <- predict(prep_p_num, data[, -which(names(data) %in% c("NObeyesdad", "BMI"))])
data_normalized_num$NObeyesdad_numeric <- data$NObeyesdad_numeric

# Train vs test
set.seed(123)
index_numeric <- createDataPartition(data_normalized_num$NObeyesdad_numeric, p = 0.8, list = FALSE)
trainset_num <- data_normalized_num[index_numeric, ]
testset_num <- data_normalized_num[-index_numeric, ]

X_train_num <- trainset_num[, -which(names(trainset_num) == "NObeyesdad_numeric")]
X_test_num <- testset_num[, -which(names(testset_num) == "NObeyesdad_numeric")]
y_train_num <- trainset_num[["NObeyesdad_numeric"]]
y_test_numeric <- testset_num[["NObeyesdad_numeric"]]
dn <- dummyVars(" ~ .", data = X_train_num)
X_train_num <- predict(dn, newdata = X_train_num)
X_test_num <- predict(dn, newdata = X_test_num)

# Random Forest Model
rf_model_numeric <- randomForest(y_train_num ~ ., data = as.data.frame(cbind(X_train_num, y_train_num))
rf_predictions_numeric <- predict(rf_model_numeric, newdata = X_test_num)

# Scatter plot
ggplot(data = data.frame(y_test_numeric, rf_predictions_numeric), aes(x = y_test_numeric, y = rf_predic
  geom_point(alpha = 0.4) +
  geom_smooth(method = "lm", color = "red") +
  labs(x = "Actual Values", y = "Predicted Values") + theme_minimal()
```




```r
# RMSE
rmse <- sqrt(mean((rf_predictions_numeric - y_test_numeric)^2))
print(paste("RMSE = ", rmse))
```

```
## [1] "RMSE =  0.251817327752757"
```

```r
# R-squared
r_sq <- cor(rf_predictions_numeric, y_test_numeric)^2
print(paste("R-squared = ", r_sq))
```

```
## [1] "R-squared =  0.985111123344303"
```

Root Mean Square Error, or RMSE, measures the average size of the errors between the predicted and actual outcomes, and lower RMSE confirms the high accuracy in prediction. The RMSE value of the model is 0.257060390474687, reflecting that the model's predictions are significantly close to the actual values. Moreover, R-squared value of the model with 0.984430086580748 denotes that the model has extremely good fit.

While the regression and classification model are not directly comparable, as they are applicable for different types of problems, the results present that the both models are well performing.

Limitations

For the limitation, our dataset may have biased towards certain cultural or ethnic backgrounds. The data was collected from Mexico, Peru, and Columbia. However, eating habits and physical conditions may differ in different regions with different cultures, including food and habits, and different ethnicities with different genetic compositions.

Furthermore, there is a difference in the distribution of obesity levels. While obesity follows a normal distribution in the real world, our sample data follows the uniform distribution, which could indicate the sampling bias. However, the uniformity may have been beneficial in the data analysis process, due to the lower likeness of overfitting to the majority class and better generalisation.

Conclusions

In this report, the dataset, including the physical conditions and eating habits, was tackled in detail to examine which factors attribute most to the obesity. Overall, the classification models, excluding the Naive Bayes, and regression models exhibited good performance and, through these methods we employed, it was demonstrated that while physical conditions such as weight inherently have a significant impact on obesity, eating habits and life patterns also substantially affect one's obesity level.