Obesity Estimation

2024-03-20

Reading the csv file

```
# Read CSV file with a specific encoding
obesity_data <- read.csv("ObesityDataSet_raw_and_data_sinthetic.csv", fileEncoding = "UTF-8")
```

Loading the required libraries

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

Data Exploration

```
# To display the first few rows of the dataset
head(obesity_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
```

```
# Check structure of the dataset
str(obesity_data)
```

```
## 'data.frame':    2111 obs. of  17 variables:
##  $ Gender                        : chr  "Female" "Female" "Male" "Male" ...
##  $ Age                           : num  21 21 23 27 22 29 23 22 24 22 ...
##  $ Height                        : num  1.62 1.52 1.8 1.8 1.78 1.62 1.5 1.64 1.78 1.72 ...
##  $ Weight                        : num  64 56 77 87 89.8 53 55 53 64 68 ...
##  $ family_history_with_overweight: chr  "yes" "yes" "yes" "no" ...
##  $ FAVC                          : chr  "no" "no" "no" "no" ...
##  $ FCVC                          : num  2 3 2 3 2 2 3 2 3 2 ...
##  $ NCP                           : num  3 3 3 3 1 3 3 3 3 3 ...
##  $ CAEC                          : chr  "Sometimes" "Sometimes" "Sometimes" "Sometimes" ...
##  $ SMOKE                         : chr  "no" "yes" "no" "no" ...
##  $ CH2O                          : num  2 3 2 2 2 2 2 2 2 2 ...
##  $ SCC                           : chr  "no" "yes" "no" "no" ...
##  $ FAF                           : num  0 3 2 2 0 0 1 3 1 1 ...
##  $ TUE                           : num  1 0 1 0 0 0 0 0 1 1 ...
##  $ CALC                          : chr  "no" "Sometimes" "Frequently" "Frequently" ...
##  $ MTRANS                        : chr  "Public_Transportation" "Public_Transportation" "Public_Trans
##  $ NObeyesdad                    : chr  "Normal_Weight" "Normal_Weight" "Normal_Weight" "Overweight_
```

Based on the output of `str(obesity_data)`, we have a dataset with 2111 obesityervations and 17 variables. Here's a brief description of each variable:

1. `Gender`: Character variable indicating the gender of the individual (e.g., "Female", "Male").
2. `Age`: Numeric variable representing the age of the individual.
3. `Height`: Numeric variable representing the height of the individual.
4. `Weight`: Numeric variable representing the weight of the individual.
5. `family_history_with_overweight`: Character variable indicating whether the individual has a family history of overweight ("yes" or "no").
6. `FAVC`: Character variable indicating whether the individual consumes high caloric food frequently ("yes" or "no").
7. `FCVC`: Numeric variable representing the frequency of consumption of high caloric food.
8. `NCP`: Numeric variable representing the number of main meals consumed daily.
9. `CAEC`: Character variable indicating the frequency of consumption of food between meals (e.g., "Sometimes", "Frequently").
10. `SMOKE`: Character variable indicating whether the individual smokes ("yes" or "no").
11. `CH2O`: Numeric variable representing daily water consumption.
12. `SCC`: Character variable indicating whether the individual monitors the calories consumed ("yes" or "no").

13. `FAF`: Numeric variable representing physical activity frequency.
14. `TUE`: Numeric variable representing time spent using technology devices.
15. `CALC`: Character variable indicating whether the individual monitors the calorie intake ("yes" or "no").
16. `MTRANS`: Character variable indicating the mode of transportation (e.g., "Public_Transportation", "Walking").
17. `NObeyesdad`: Character variable indicating the obesity level of the individual (e.g., "Normal_Weight", "Overweight_Level_I").

```
# Check for missing values
na_count <- colSums(is.na(obesity_data))
print(na_count)
```

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

We notice that the count of missing values for each variable is 0, which denotes that there are no missing values in any of the variables within the dataset.

With no missing values to address, we can go ahead with data exploration and analysis. So here is the summary of the dataset.

```
# Summary statistics
summary(obesity_data)
```

```
##     Gender                Age              Height          Weight
##  Length:2111        Min.   :14.00   Min.   :1.450   Min.   : 39.00
##  Class :character   1st Qu.:19.95   1st Qu.:1.630   1st Qu.: 65.47
##  Mode  :character   Median :22.78   Median :1.700   Median : 83.00
##                     Mean   :24.31   Mean   :1.702   Mean   : 86.59
##                     3rd Qu.:26.00   3rd Qu.:1.768   3rd Qu.:107.43
##                     Max.   :61.00   Max.   :1.980   Max.   :173.00
##  family_history_with_overweight     FAVC               FCVC
##  Length:2111                    Length:2111        Min.   :1.000
##  Class :character               Class :character   1st Qu.:2.000
##  Mode  :character               Mode  :character   Median :2.386
##                                                    Mean   :2.419
##                                                    3rd Qu.:3.000
```

```
##                                                 Max.   :3.000
##       NCP             CAEC              SMOKE                CH2O
##  Min.   :1.000   Length:2111        Length:2111        Min.   :1.000
##  1st Qu.:2.659   Class :character   Class :character   1st Qu.:1.585
##  Median :3.000   Mode  :character   Mode  :character   Median :2.000
##  Mean   :2.686                                         Mean   :2.008
##  3rd Qu.:3.000                                         3rd Qu.:2.477
##  Max.   :4.000                                         Max.   :3.000
##      SCC                 FAF               TUE             CALC
##  Length:2111        Min.   :0.0000   Min.   :0.0000   Length:2111
##  Class :character   1st Qu.:0.1245   1st Qu.:0.0000   Class :character
##  Mode  :character   Median :1.0000   Median :0.6253   Mode  :character
##                     Mean   :1.0103   Mean   :0.6579
##                     3rd Qu.:1.6667   3rd Qu.:1.0000
##                     Max.   :3.0000   Max.   :2.0000
##     MTRANS           NObeyesdad
##  Length:2111        Length:2111
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

```r
# Load required libraries
library(caret)
```

```
## Loading required package: lattice
```

```r
# Encode categorical variables using one-hot encoding
obesity_data_encoded <- dummyVars(~., data = obesity_data)
obesity_data_encoded <- data.frame(predict(obesity_data_encoded, newdata = obesity_data))

# Scale numerical features
preproc_train_data <- preProcess(obesity_data[, !(names(obesity_data) %in% c("Gender", "family_history_
                                 method = c("center", "scale"))
obesity_data_scaled <- predict(preproc_train_data, obesity_data[, !(names(obesity_data) %in% c("Gender"
```

Descriptive Statistics

```r
# Compute mean, median, and standard deviation
mean_values <- sapply(obesity_data, mean, na.rm = TRUE)
median_values <- sapply(obesity_data, median, na.rm = TRUE)
sd_values <- sapply(obesity_data, sd, na.rm = TRUE)

# Create a summary dataframe
summary_stats <- data.frame(
  Mean = mean_values,
  Median = median_values,
  SD = sd_values
)
```

```r
# Display summary statistics
print(summary_stats)
```

```
##                                    Mean                Median          SD
## Gender                               NA                  Male          NA
## Age                          24.3125999              22.77789  6.34596827
## Height                        1.7016774              1.700499  0.09330482
## Weight                       86.5860581                    83 26.19117175
## family_history_with_overweight       NA                   yes          NA
## FAVC                                 NA                   yes          NA
## FCVC                          2.4190431              2.385502  0.53392658
## NCP                           2.6856280                     3  0.77803865
## CAEC                                 NA              Sometimes          NA
## SMOKE                                NA                    no          NA
## CH2O                          2.0080114                     2  0.61295345
## SCC                                  NA                    no          NA
## FAF                           1.0102977                     1  0.85059243
## TUE                           0.6578659               0.62535  0.60892726
## CALC                                 NA              Sometimes          NA
## MTRANS                               NA Public_Transportation          NA
## NObeyesdad                           NA       Obesity_Type_II          NA
```

```r
# Compute correlation coefficients
numeric_data <- obesity_data[, sapply(obesity_data, is.numeric)]
correlation_matrix <- cor(numeric_data)

# Display correlation matrix
print(correlation_matrix)
```

```
##                Age      Height      Weight        FCVC         NCP        CH2O
## Age     1.00000000 -0.02595813  0.20256010  0.01629089 -0.04394373 -0.04530386
## Height -0.02595813  1.00000000  0.46313612 -0.03812106  0.24367173  0.21337592
## Weight  0.20256010  0.46313612  1.00000000  0.21612471  0.10746899  0.20057539
## FCVC    0.01629089 -0.03812106  0.21612471  1.00000000  0.04221630  0.06846147
## NCP    -0.04394373  0.24367173  0.10746899  0.04221630  1.00000000  0.05708800
## CH2O   -0.04530386  0.21337592  0.20057539  0.06846147  0.05708800  1.00000000
## FAF    -0.14493833  0.29470900 -0.05143627  0.01993940  0.12950431  0.16723649
## TUE    -0.29693059  0.05191167 -0.07156136 -0.10113485  0.03632557  0.01196534
##                FAF         TUE
## Age    -0.14493833 -0.29693059
## Height  0.29470900  0.05191167
## Weight -0.05143627 -0.07156136
## FCVC    0.01993940 -0.10113485
## NCP     0.12950431  0.03632557
## CH2O    0.16723649  0.01196534
## FAF     1.00000000  0.05856207
## TUE     0.05856207  1.00000000
```

The correlation matrix shows the correlation coefficients between pairs of numerical variables in your dataset. From the output, we understand that:

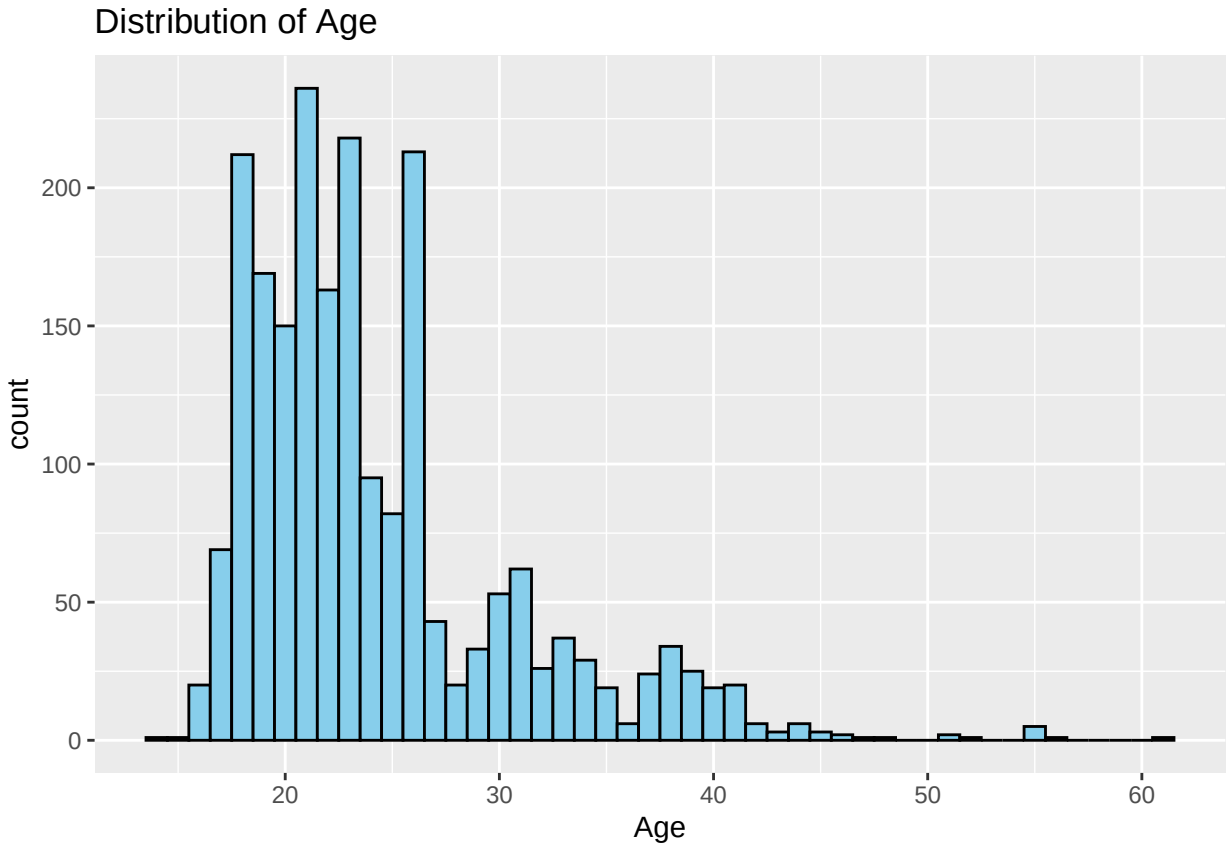
1. **Interpretation of Correlation Coefficients:**

- Age, Height, and Weight:
  - Age and Weight have a positive correlation coefficient of approximately 0.20, indicating a weak positive correlation. This suggests that older individuals tend to have slightly higher weights.
  - Height and Weight have a stronger positive correlation coefficient of approximately 0.46, indicating a moderate positive correlation. This suggests that taller individuals tend to have higher weights.
  - Age and Height have a weak negative correlation coefficient, indicating a very slight negative relationship.
- FCVC (Frequency of Consumption of Vegetables) and NCP (Number of Main Meals Consumed Daily):
  - FCVC and NCP have a very weak positive correlation coefficient of approximately 0.04, suggesting a very slight positive relationship.
- Other variables:
  - The remaining variables (CH2O, FAF, TUE) also have correlation coefficients with each other, indicating their respective relationships.

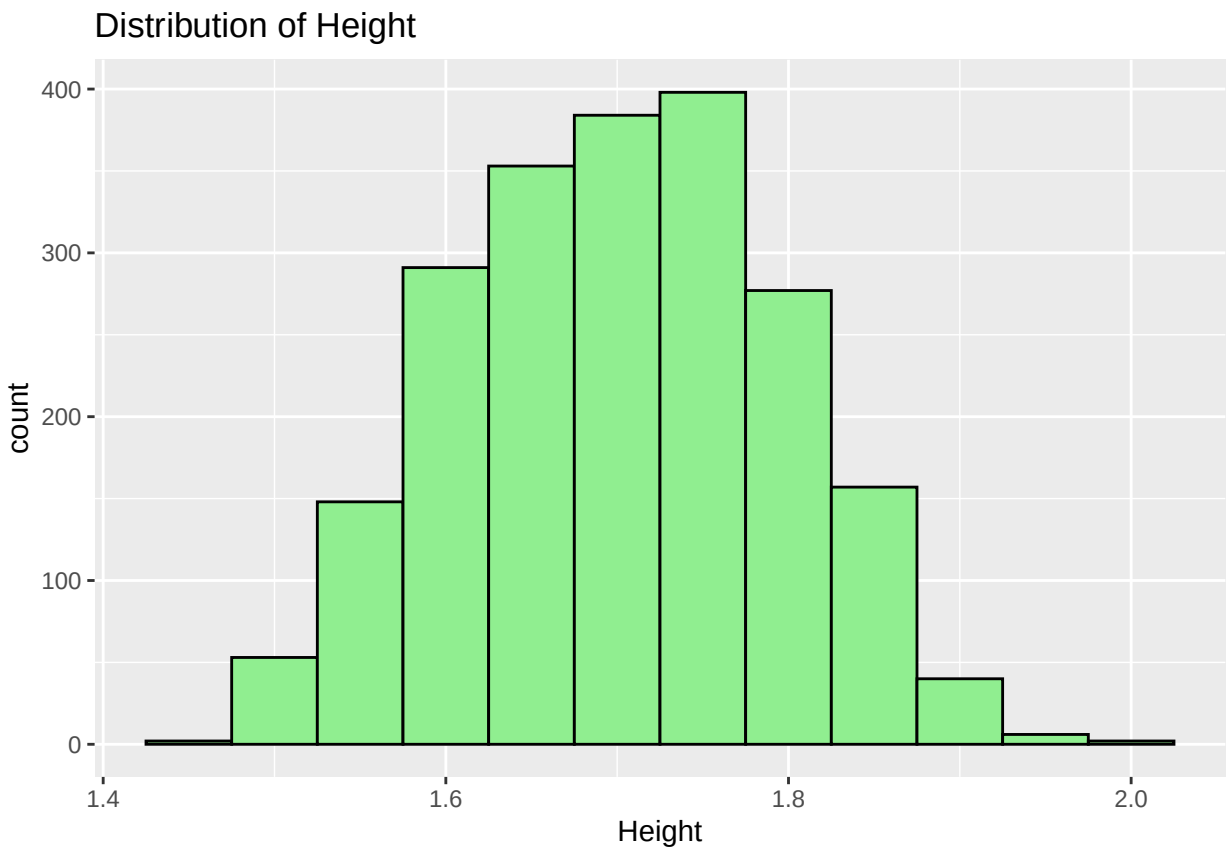
2. **Interpretation of Negative Correlation:**
   - Age and TUE (Time Spent Using Technology Devices) have a negative correlation coefficient of approximately -0.30. This suggests that as age increases, the time spent using technology devices tends to decrease.
3. **Interpretation of Weak Correlation:**
   - Most of the correlations in the matrix are weak, indicating that the variables have little linear relationship with each other.

Data Visualization

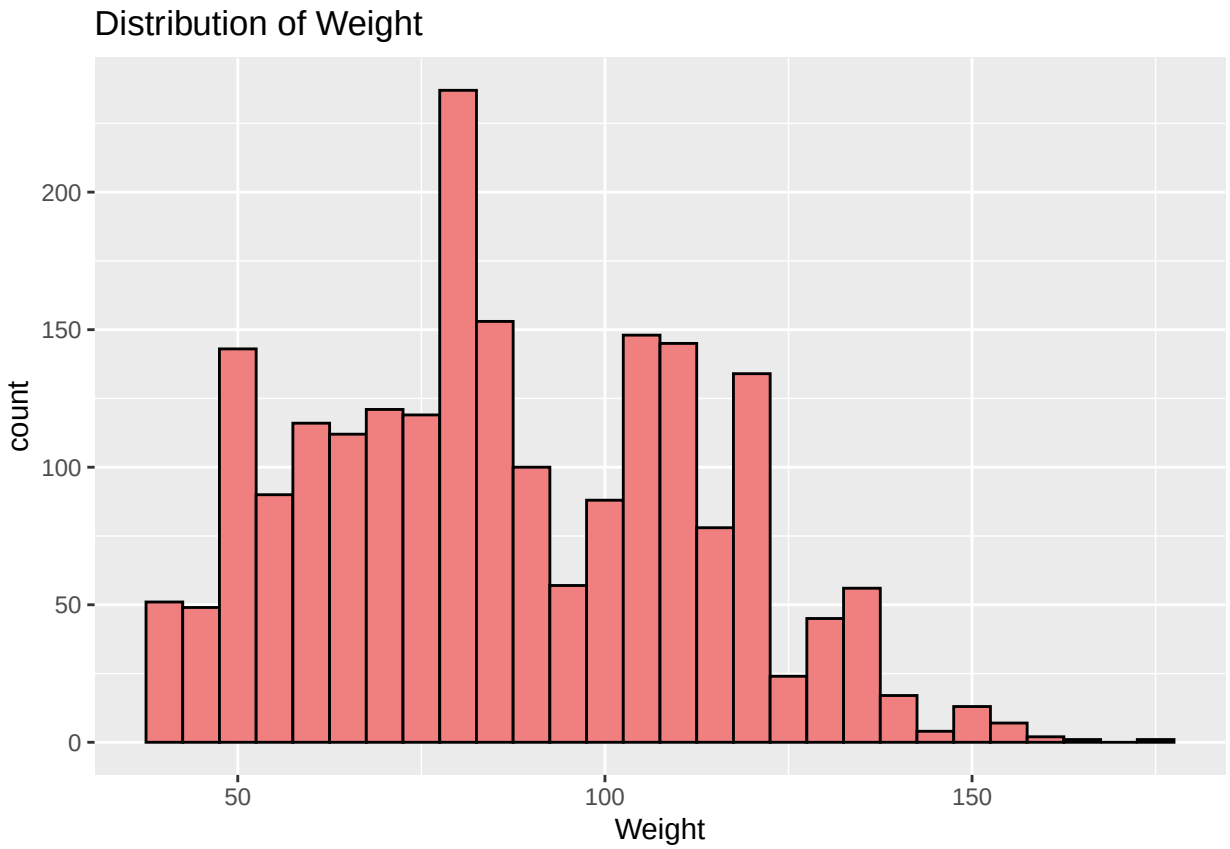
```
# Histogram of Age
ggplot(obesity_data, aes(x = Age)) +
  geom_histogram(binwidth = 1, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Age")
```



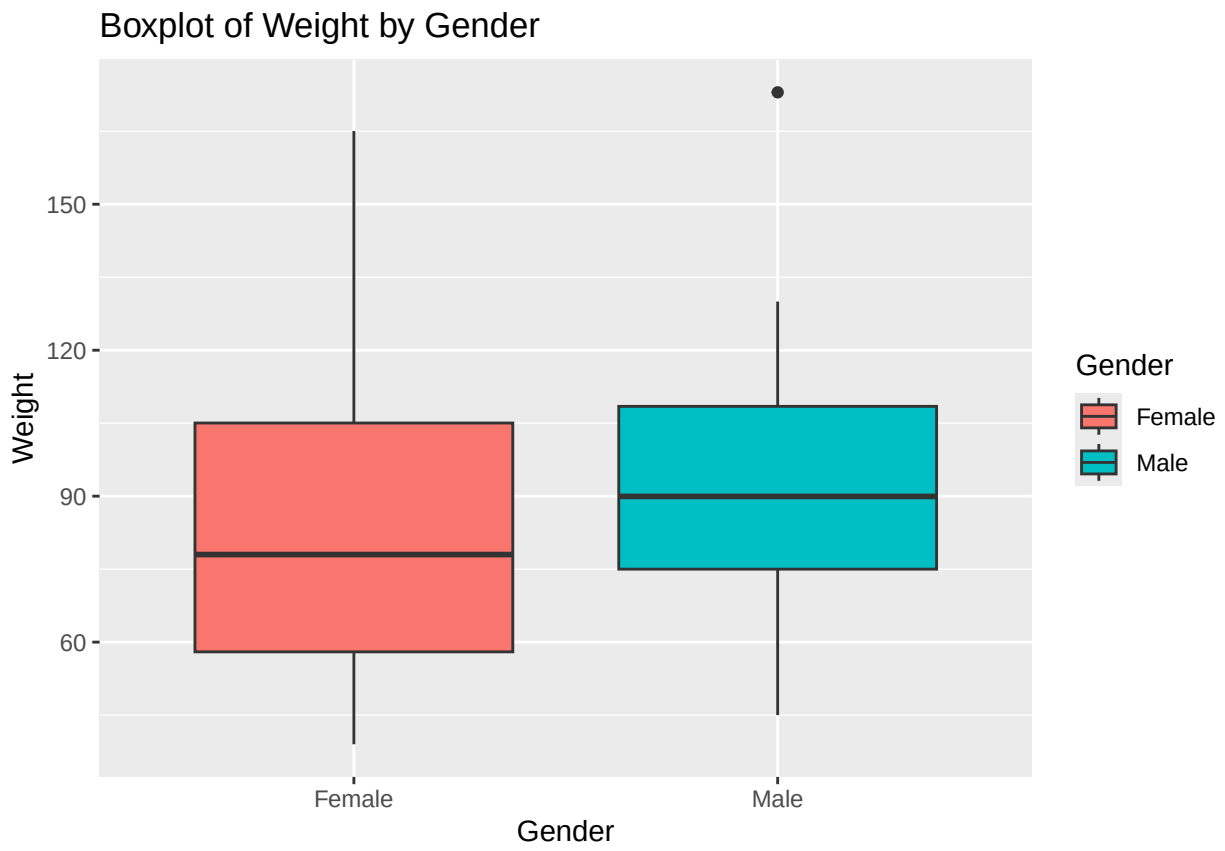

```r
# Histogram of Height
ggplot(obesity_data, aes(x = Height)) +
  geom_histogram(binwidth = 0.05, fill = "lightgreen", color = "black") +
  labs(title = "Distribution of Height")
```




```r
# Histogram of Weight
ggplot(obesity_data, aes(x = Weight)) +
  geom_histogram(binwidth = 5, fill = "lightcoral", color = "black") +
  labs(title = "Distribution of Weight")
```

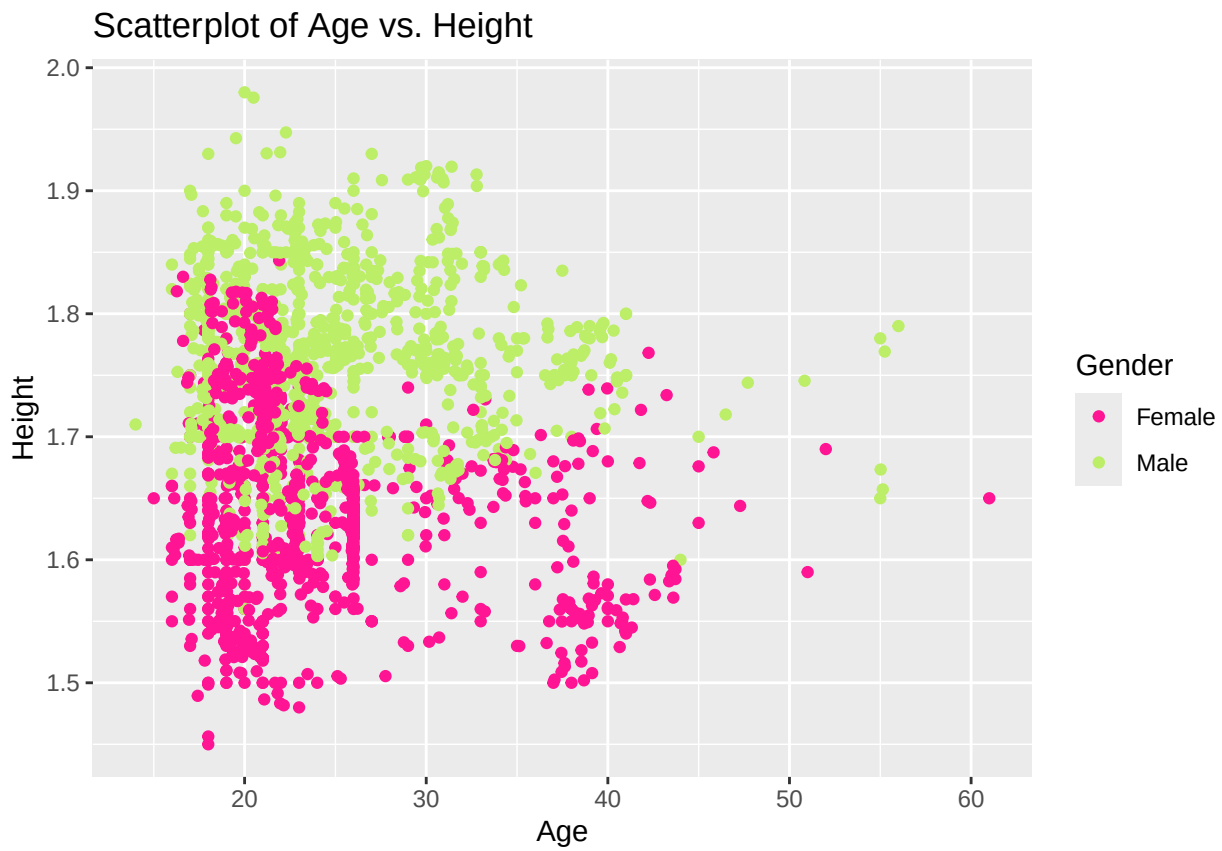



```r
# Boxplot of Weight by Gender
ggplot(obesity_data, aes(x = Gender, y = Weight, fill = Gender)) +
  geom_boxplot() +
  labs(title = "Boxplot of Weight by Gender")
```

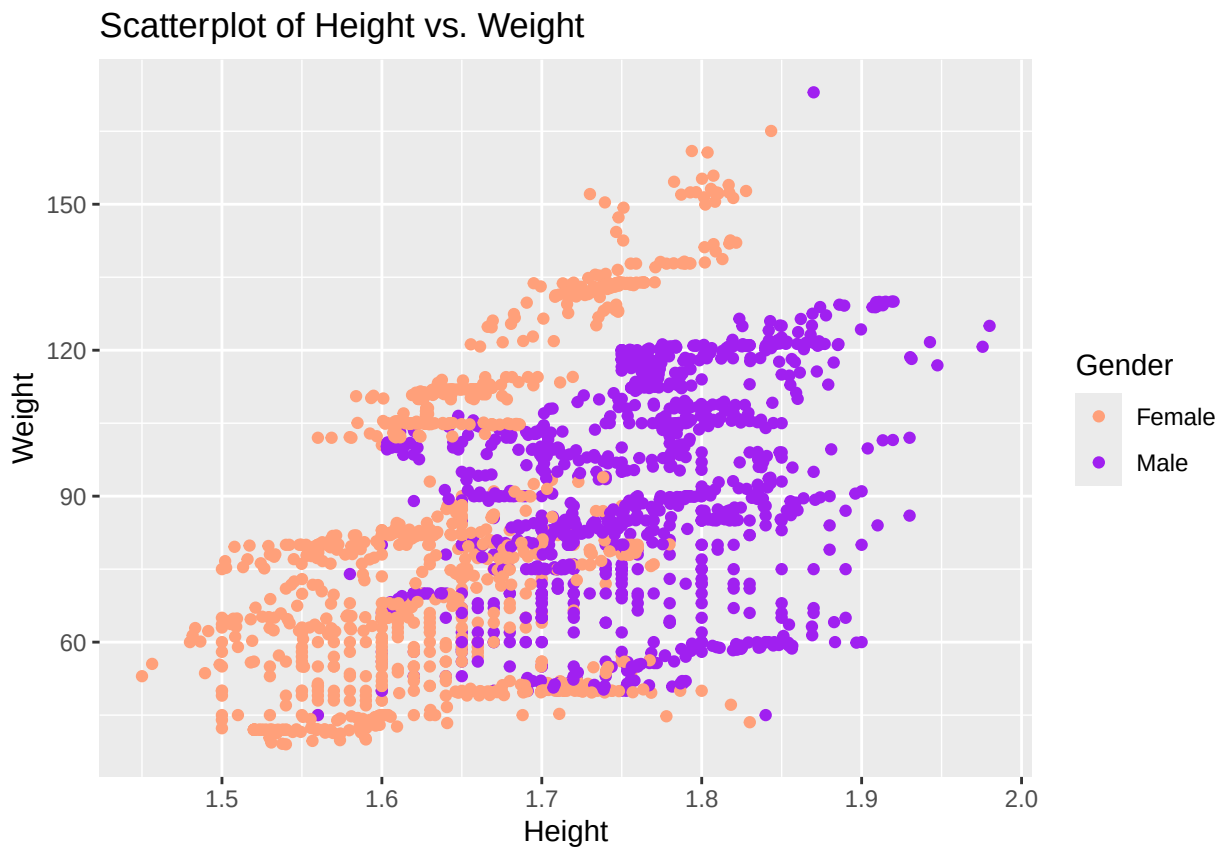



```r
# Boxplot of Age by Obesity Level
ggplot(obesity_data, aes(x = NObeyesdad, y = Age, fill = NObeyesdad)) +
  geom_boxplot() +
  labs(title = "Boxplot of Age by Obesity Level")
```

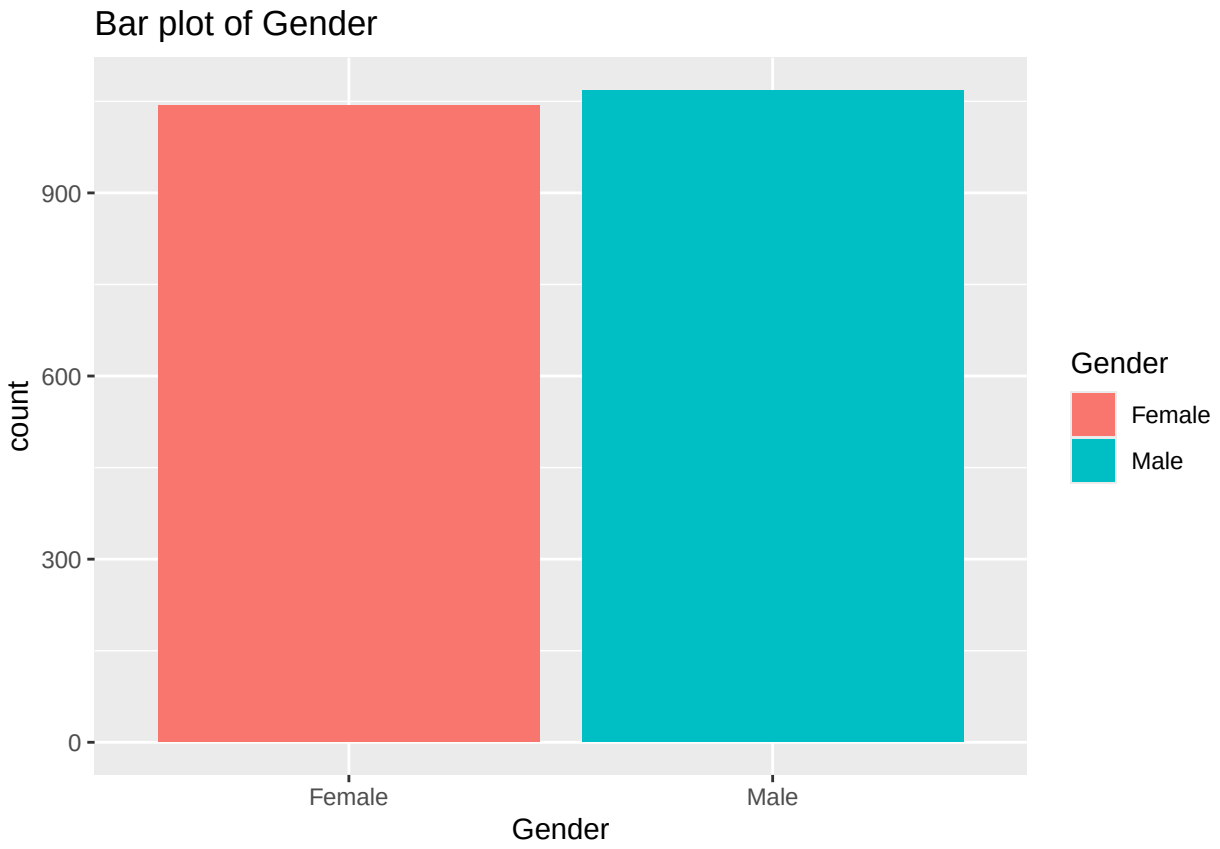

```r
# Scatterplot of Age vs. Height with color
ggplot(obesity_data, aes(x = Age, y = Height, color = Gender)) +
  geom_point() +
  labs(title = "Scatterplot of Age vs. Height") +
  scale_color_manual(values = c("Female" = "deeppink", "Male" = "darkolivegreen2"))  # Custom color sca
```



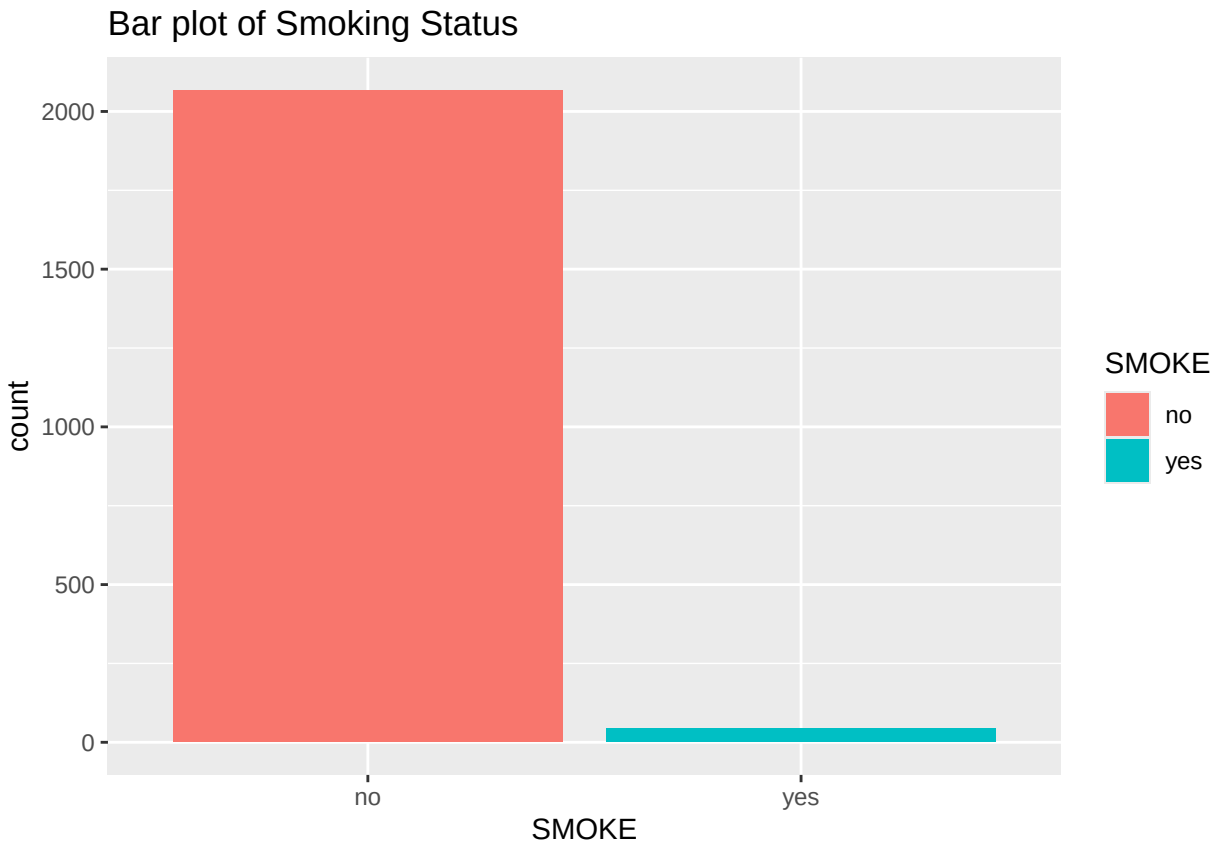

```r
# Scatterplot of Height vs. Weight with color
ggplot(obesity_data, aes(x = Height, y = Weight, color = Gender)) +
  geom_point() +
  labs(title = "Scatterplot of Height vs. Weight") +
  scale_color_manual(values = c("Female" = "lightsalmon", "Male" = "purple"))  # Custom color scale
```



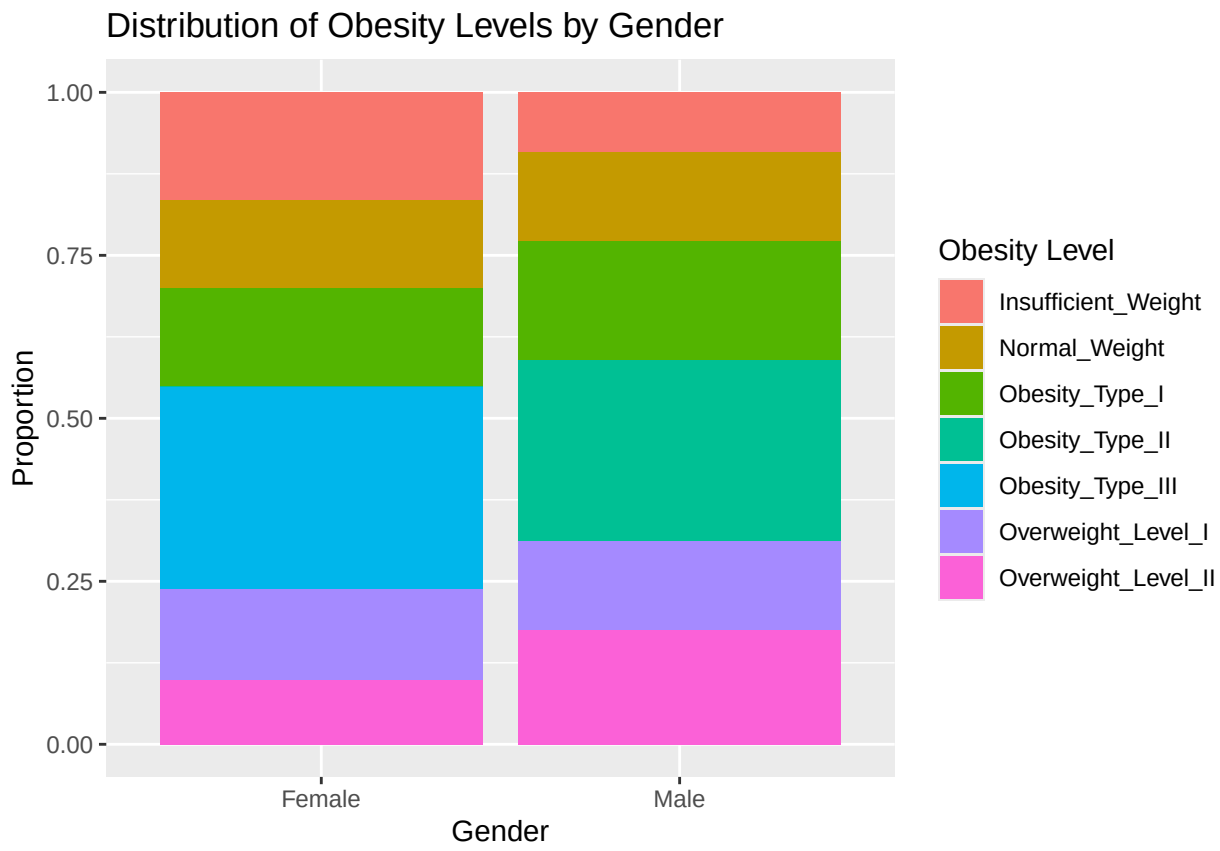

```r
# Bar plot of Gender
ggplot(obesity_data, aes(x = Gender, fill = Gender)) +
  geom_bar() +
  labs(title = "Bar plot of Gender")
```



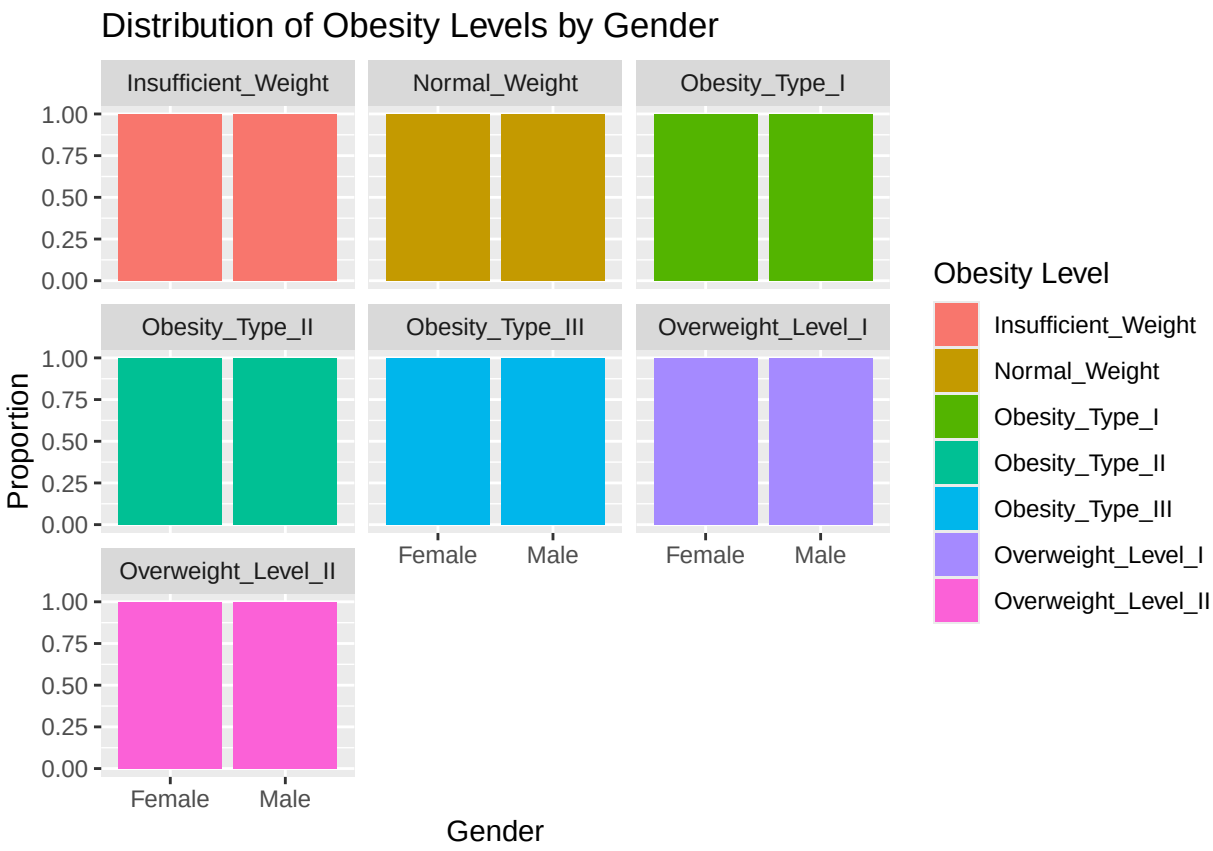

```
# Bar plot of Smoking Status
ggplot(obesity_data, aes(x = SMOKE, fill = SMOKE)) +
  geom_bar() +
  labs(title = "Bar plot of Smoking Status")
```



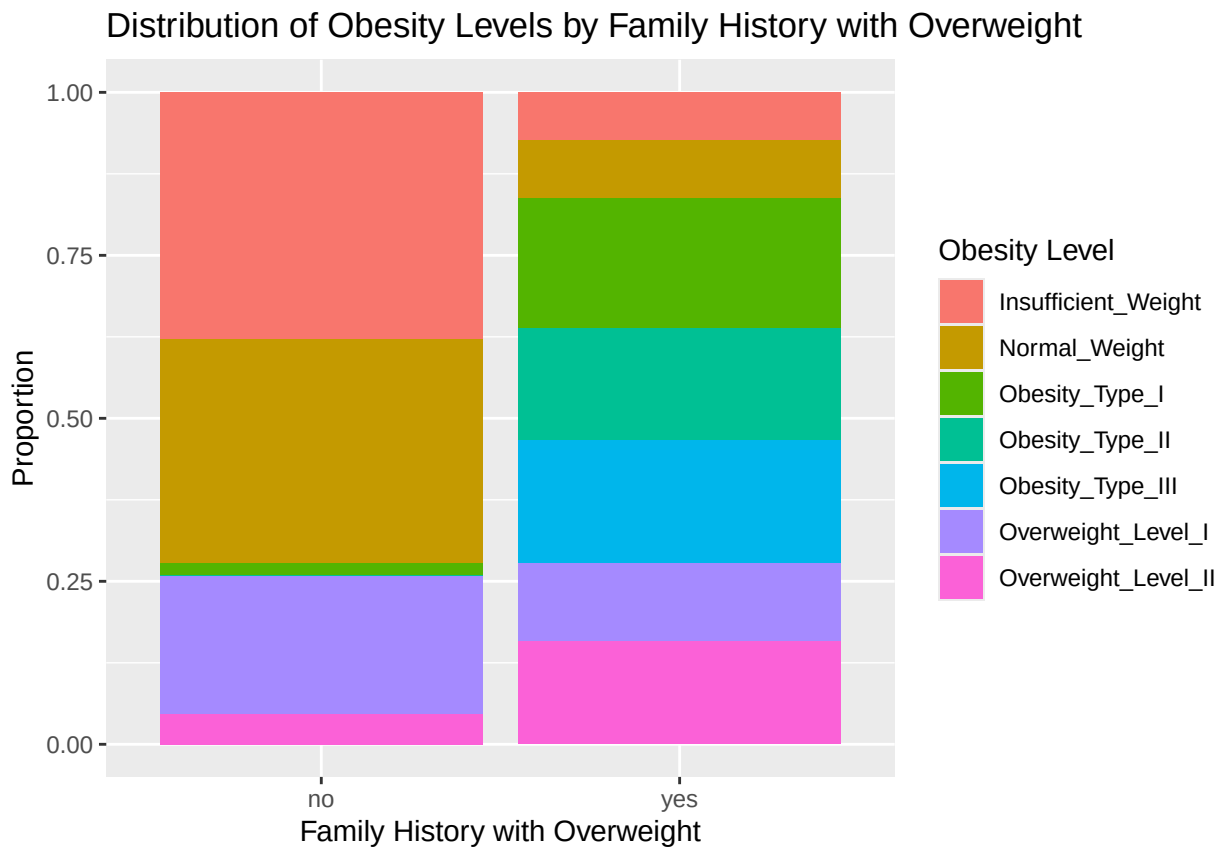

```
# Visualize the distribution of Obesity levels by Gender
ggplot(obesity_data, aes(x = Gender, fill = NObeyesdad)) +
  geom_bar(position = "fill") +
  labs(title = "Distribution of Obesity Levels by Gender", x = "Gender", y = "Proportion", fill = "Obesi
```



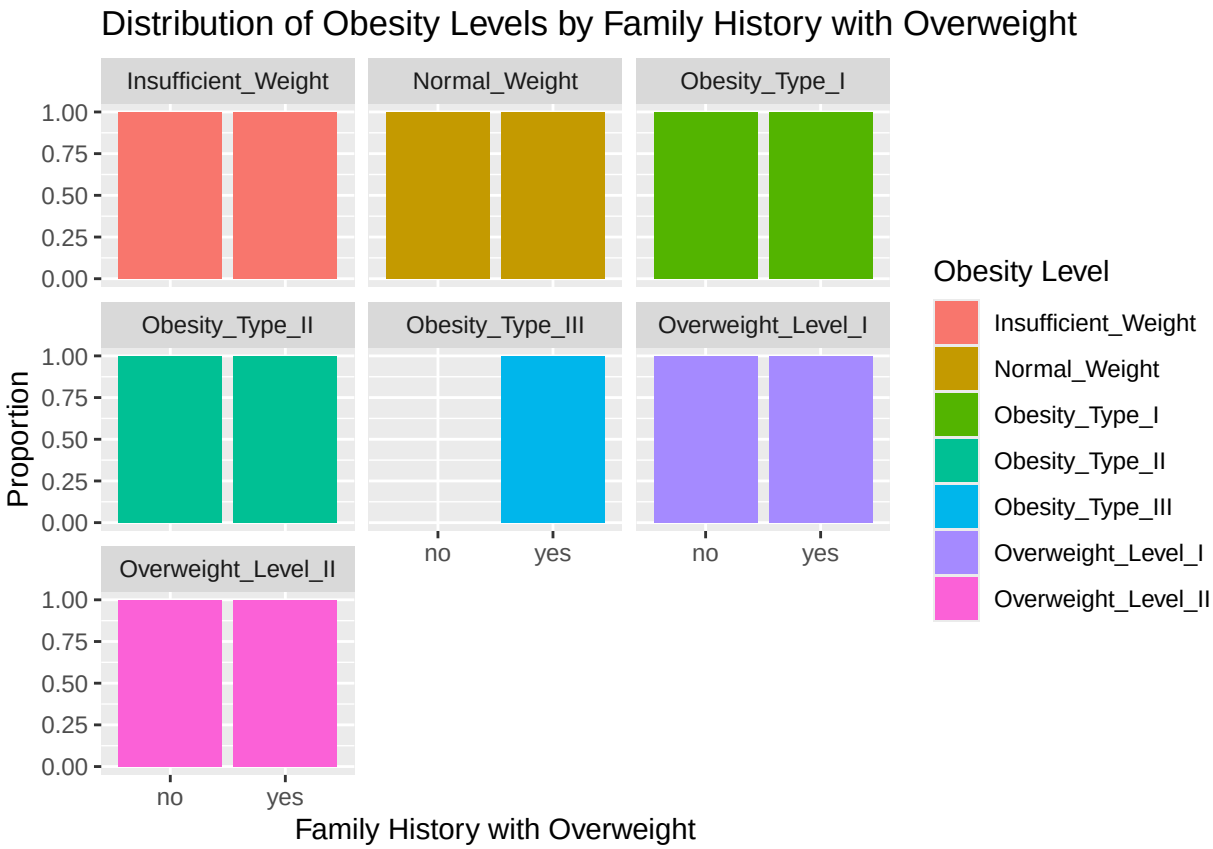

```r
# Visualize the distribution of Obesity levels by Gender for each Obesity Level
ggplot(obesity_data, aes(x = Gender, fill = NObeyesdad)) +
  geom_bar(position = "fill") +
  facet_wrap(~ NObeyesdad) +
  labs(title = "Distribution of Obesity Levels by Gender", x = "Gender", y = "Proportion", fill = "Obesi
```



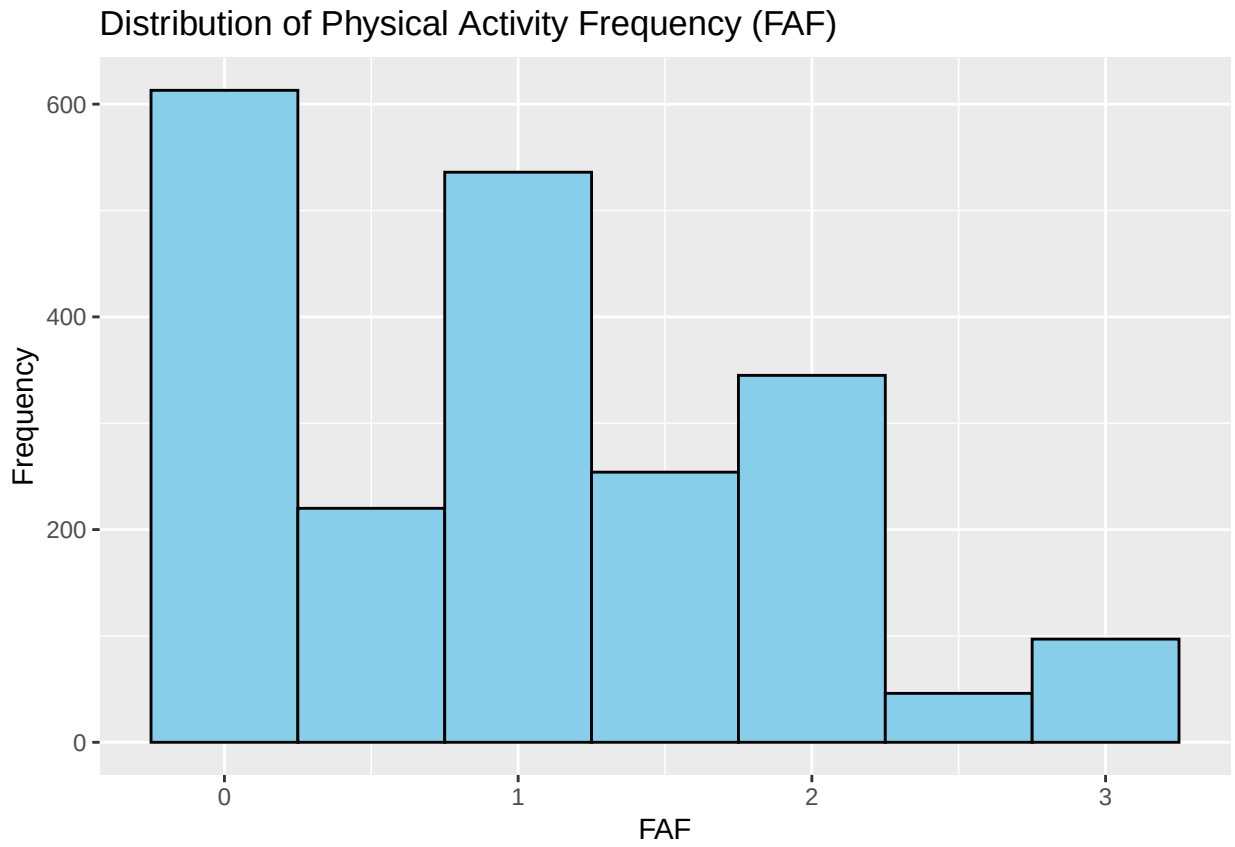

```r
# Visualize the distribution of family history with overweight
ggplot(obesity_data, aes(x = family_history_with_overweight, fill = NObeyesdad)) +
  geom_bar(position = "fill") +
  labs(title = "Distribution of Obesity Levels by Family History with Overweight",
       x = "Family History with Overweight", y = "Proportion", fill = "Obesity Level")
```



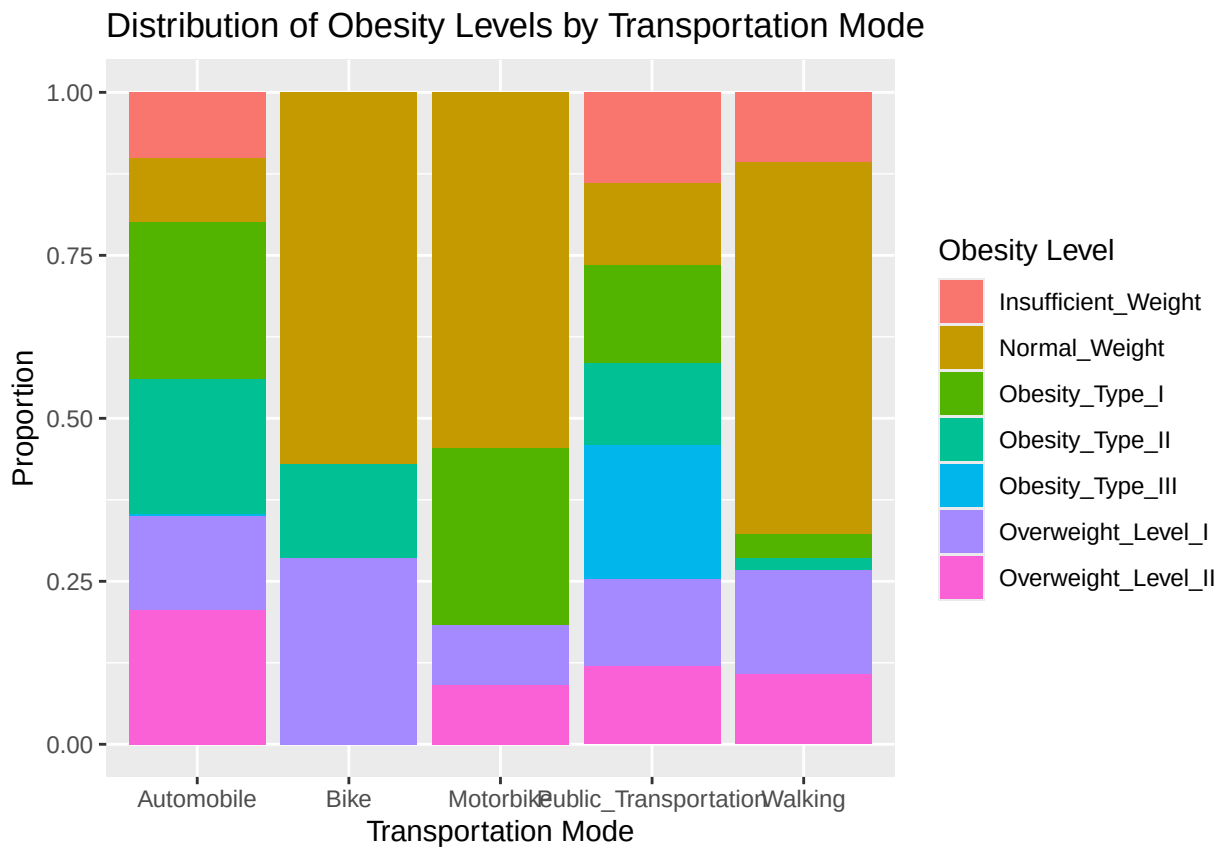

```r
# Visualize the distribution of Obesity levels by Family History with Overweight for each Obesity Level
ggplot(obesity_data, aes(x = family_history_with_overweight, fill = NObeyesdad)) +
  geom_bar(position = "fill") +
  facet_wrap(~ NObeyesdad) +
  labs(title = "Distribution of Obesity Levels by Family History with Overweight",
       x = "Family History with Overweight", y = "Proportion", fill = "Obesity Level")
```



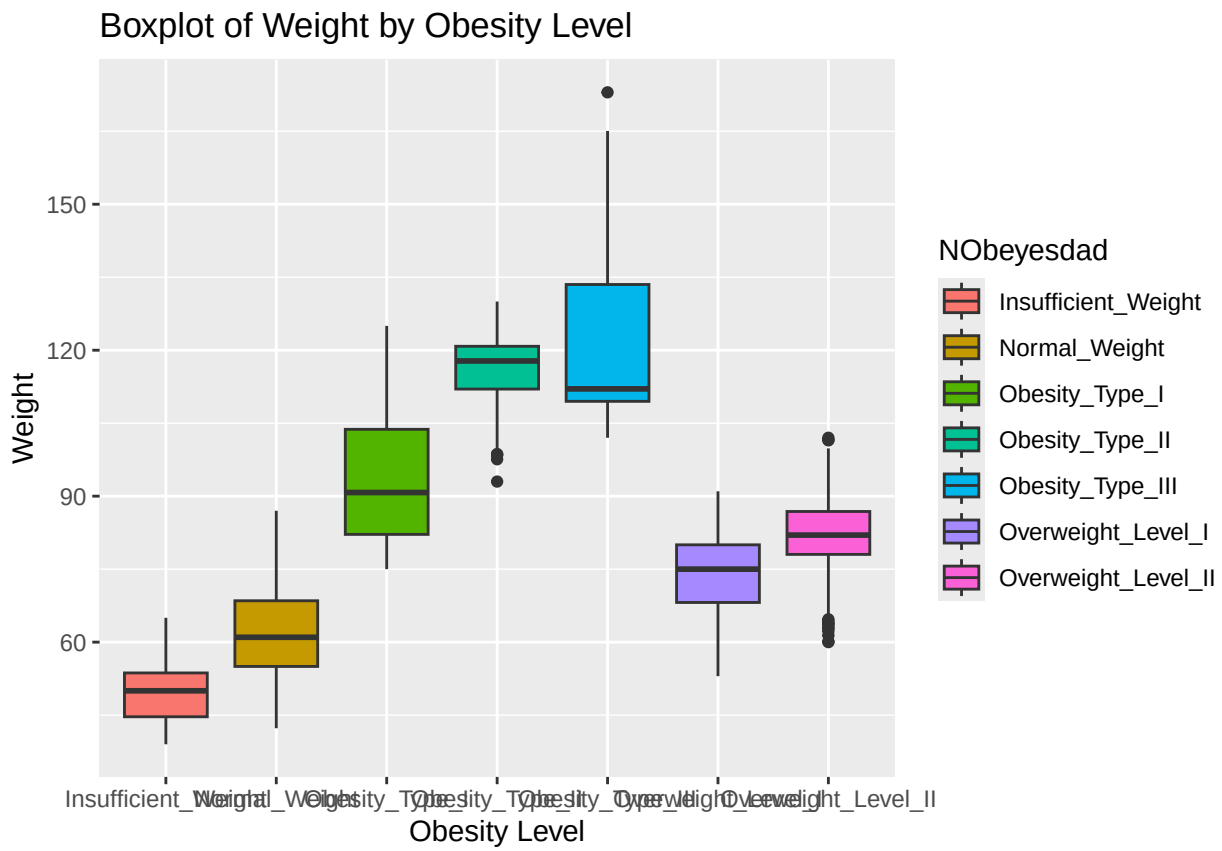

```r
# Visualize the distribution of physical activity frequency (FAF)
ggplot(obesity_data, aes(x = FAF)) +
  geom_histogram(binwidth = 0.5, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Physical Activity Frequency (FAF)", x = "FAF", y = "Frequency")
```



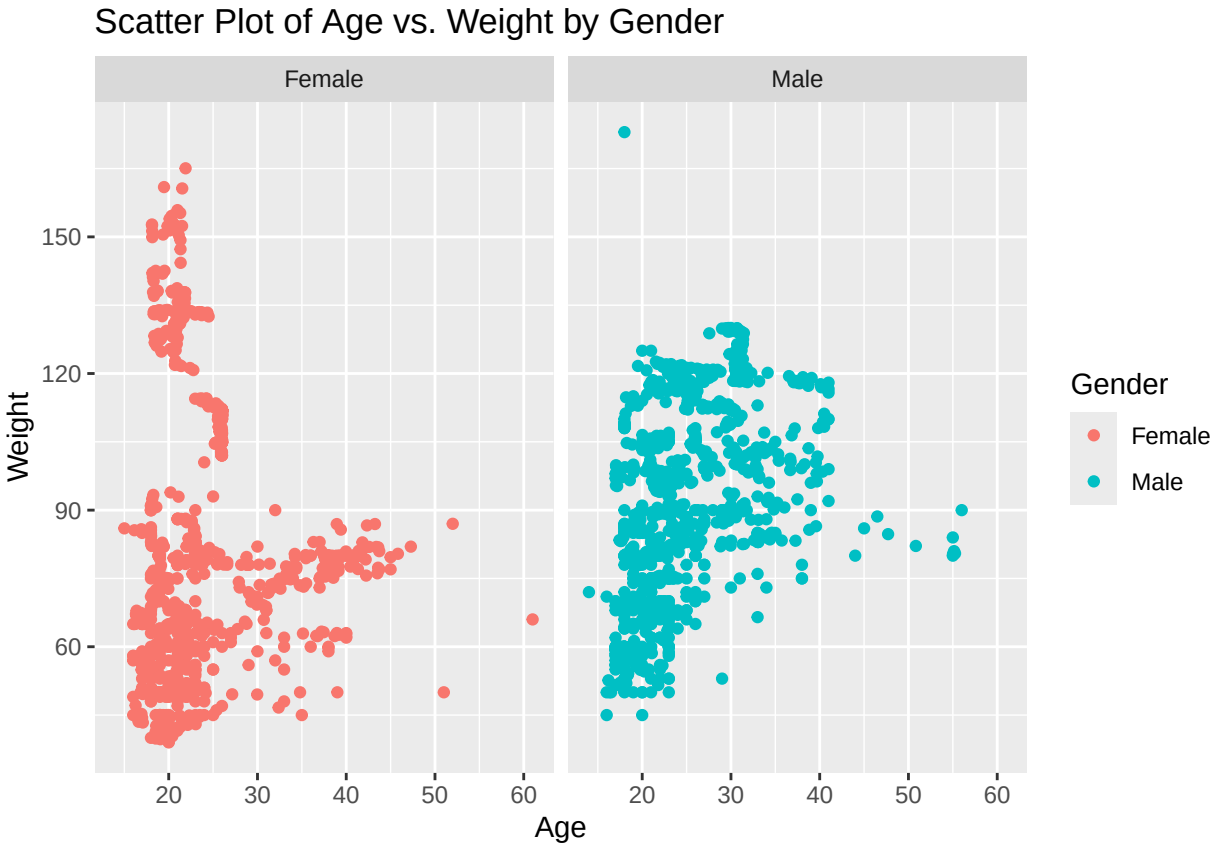

```
# Visualize the distribution of transportation mode (MTRANS)
ggplot(obesity_data, aes(x = MTRANS, fill = NObeyesdad)) +
  geom_bar(position = "fill") +
  labs(title = "Distribution of Obesity Levels by Transportation Mode",
       x = "Transportation Mode", y = "Proportion", fill = "Obesity Level")
```



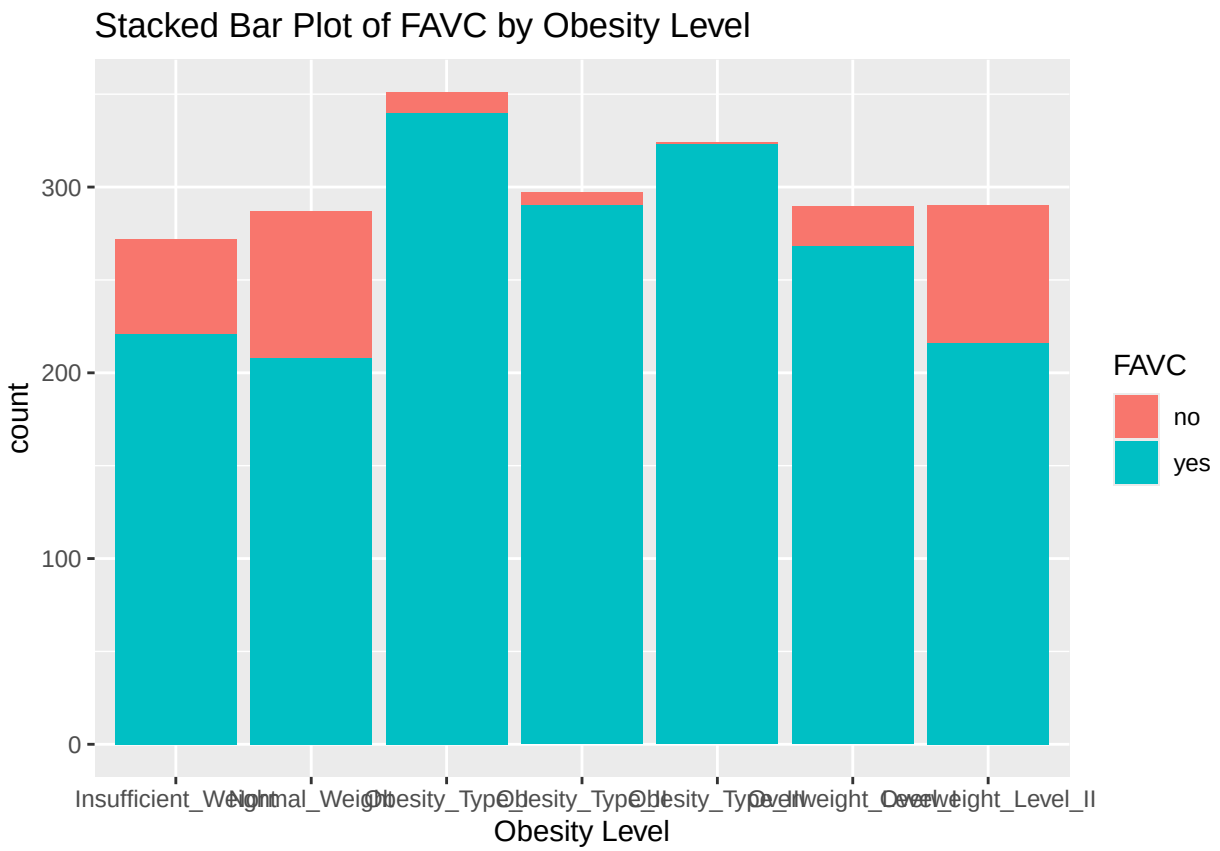

```r
# Boxplot of Weight by Obesity Level
ggplot(obesity_data, aes(x = NObeyesdad, y = Weight, fill = NObeyesdad)) +
  geom_boxplot() +
  labs(title = "Boxplot of Weight by Obesity Level", x = "Obesity Level", y = "Weight")
```



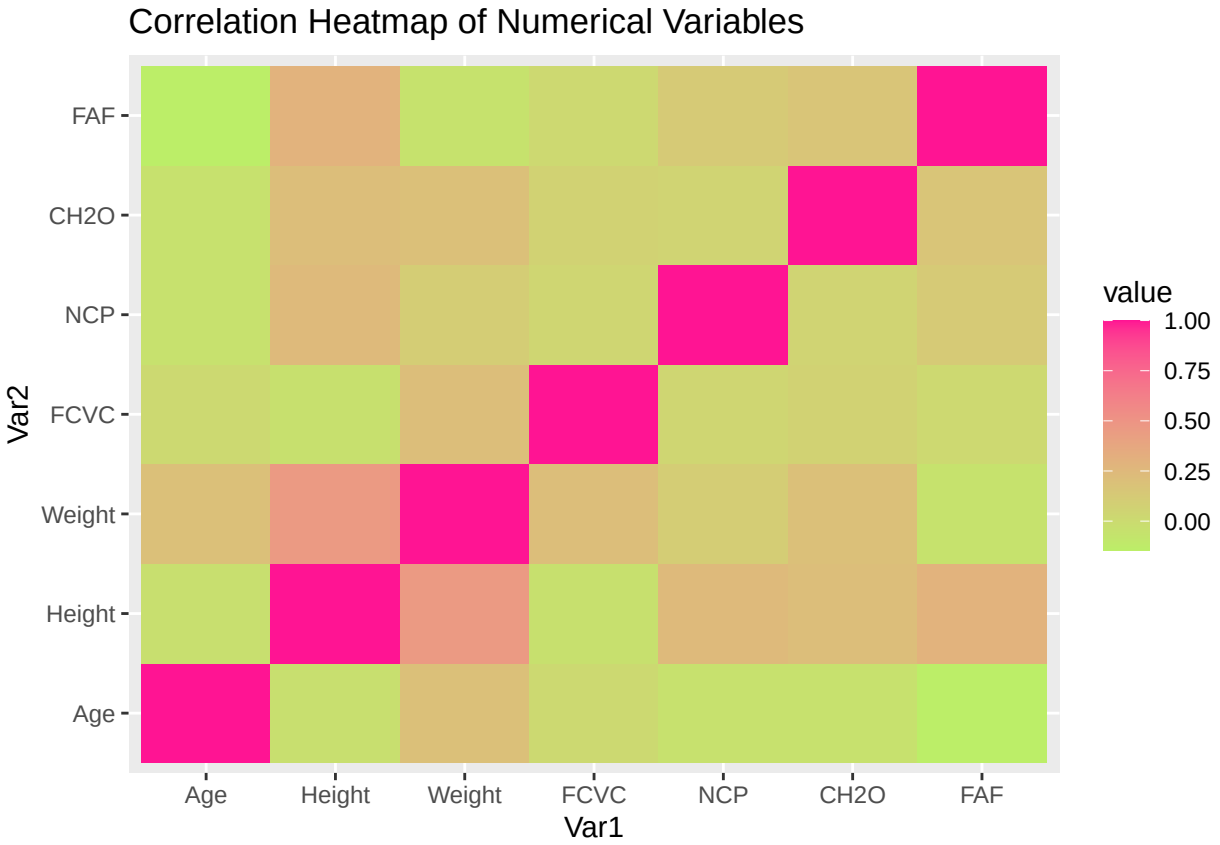

```r
# Faceted scatter plot of Age vs. Weight by Gender
ggplot(obesity_data, aes(x = Age, y = Weight, color = Gender)) +
  geom_point() +
  facet_wrap(~ Gender) +
  labs(title = "Scatter Plot of Age vs. Weight by Gender", x = "Age", y = "Weight")
```




```r
# Stacked bar plot of FAVC (Frequency of consumption of high caloric food) by Obesity Level
ggplot(obesity_data, aes(x = NObeyesdad, fill = FAVC)) +
  geom_bar(position = "stack") +
  labs(title = "Stacked Bar Plot of FAVC by Obesity Level", x = "Obesity Level", fill = "FAVC")
```




```r
# Calculate correlation matrix
correlation_matrix <- cor(select(obesity_data, c("Age", "Height", "Weight", "FCVC", "NCP", "CH2O", "FAF

# Visualize correlation heatmap
library(reshape2)
correlation_melted <- melt(correlation_matrix)
ggplot(correlation_melted, aes(Var1, Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient(low = "darkolivegreen2", high = "deeppink") +
  labs(title = "Correlation Heatmap of Numerical Variables")
```



Inferential Statistics

```
# Load required libraries
library(dplyr)
library(tidyr)
```

```
##
## Attaching package: 'tidyr'

## The following object is masked from 'package:reshape2':
##
##     smiths
```

```
library(ggplot2)
library(ggpubr)

# Convert NObeyesdad to factor for analysis
obesity_data$NObeyesdad <- as.factor(obesity_data$NObeyesdad)

# Function to conduct chi-square test
conduct_chi_square_test <- function(data, x_var, y_var) {
  contingency_table <- table(data[[x_var]], data[[y_var]])
  chi_square_test <- chisq.test(contingency_table)
```

```r
  return(chi_square_test)
}

# Function to conduct ANOVA test
conduct_anova_test <- function(data, x_var, y_var) {
  anova_result <- aov(data[[y_var]] ~ data[[x_var]], data = data)
  return(anova_result)
}

# Perform hypothesis tests for each categorical variable
categorical_vars <- c("Gender", "family_history_with_overweight", "FAVC", "CAEC", "SMOKE", "SCC", "MTRA

for (var in categorical_vars) {
  # Chi-square test
  chi_square_test <- conduct_chi_square_test(obesity_data, var, "NObeyesdad")
  print(paste("Chi-square test for", var))
  print(chi_square_test)

  # ANOVA test
  if (length(levels(obesity_data[[var]])) > 2) {
    anova_result <- conduct_anova_test(obesity_data, var, "NObeyesdad")
    print(paste("ANOVA test for", var))
    print(summary(anova_result))
  }
}
```

```
## [1] "Chi-square test for Gender"
##
##  Pearson's Chi-squared test
##
## data:  contingency_table
## X-squared = 657.75, df = 6, p-value < 2.2e-16
##
## [1] "Chi-square test for family_history_with_overweight"
##
##  Pearson's Chi-squared test
##
## data:  contingency_table
## X-squared = 621.98, df = 6, p-value < 2.2e-16
##
## [1] "Chi-square test for FAVC"
##
##  Pearson's Chi-squared test
##
## data:  contingency_table
## X-squared = 233.34, df = 6, p-value < 2.2e-16
##
## [1] "Chi-square test for CAEC"
##
##  Pearson's Chi-squared test
##
## data:  contingency_table
## X-squared = 802.98, df = 18, p-value < 2.2e-16
```

```
##
## [1] "Chi-square test for SMOKE"
##
##  Pearson's Chi-squared test
##
## data:  contingency_table
## X-squared = 32.138, df = 6, p-value = 1.535e-05
##
## [1] "Chi-square test for SCC"
##
##  Pearson's Chi-squared test
##
## data:  contingency_table
## X-squared = 123.02, df = 6, p-value < 2.2e-16

## Warning in chisq.test(contingency_table): Chi-squared approximation may be
## incorrect

## [1] "Chi-square test for MTRANS"
##
##  Pearson's Chi-squared test
##
## data:  contingency_table
## X-squared = 292.59, df = 24, p-value < 2.2e-16
```

The results of the chi-square tests provide insights into the relationship between each categorical variable and the obesity levels (`NObeyesdad`).

1. **Gender:**
    - The chi-square test for Gender yields a p-value of less than 2.2e-16, indicating that there is a significant relationship between Gender and obesity levels. Therefore, we reject the null hypothesis and accept the alternative hypothesis that there is a significant relationship between Gender and obesity levels.
2. **Family History with Overweight:**
    - The chi-square test for Family History with Overweight also yields a p-value of less than 2.2e-16, indicating a significant relationship between Family History with Overweight and obesity levels. Therefore, we reject the null hypothesis and accept the alternative hypothesis that there is a significant relationship between Family History with Overweight and obesity levels.
3. **FAVC (Frequency of Consumption of Vegetables):**
    - The chi-square test for FAVC yields a p-value of less than 2.2e-16, indicating a significant relationship between FAVC and obesity levels. Therefore, we reject the null hypothesis and accept the alternative hypothesis that there is a significant relationship between FAVC and obesity levels.
4. **CAEC (Consumption of Food Between Meals):**
    - The chi-square test for CAEC yields a p-value of less than 2.2e-16, indicating a significant relationship between CAEC and obesity levels. Therefore, we reject the null hypothesis and accept the alternative hypothesis that there is a significant relationship between CAEC and obesity levels.
5. **SMOKE (Smoking Status):**
    - The chi-square test for SMOKE yields a p-value of 1.535e-05, which is less than the significance level of 0.05. Therefore, we reject the null hypothesis and accept the alternative hypothesis that there is a significant relationship between Smoking Status and obesity levels.

6. **SCC (Calories Consumption Monitoring):**
    - The chi-square test for SCC yields a p-value of less than 2.2e-16, indicating a significant relationship between SCC and obesity levels. Therefore, we reject the null hypothesis and accept the alternative hypothesis that there is a significant relationship between SCC and obesity levels.
7. **MTRANS (Mode of Transportation):**
    - The chi-square test for MTRANS yields a p-value of less than 2.2e-16, indicating a significant relationship between MTRANS and obesity levels. Therefore, we reject the null hypothesis and accept the alternative hypothesis that there is a significant relationship between MTRANS and obesity levels.

Overall, these results suggest that all the examined categorical variables have a significant relationship with obesity levels, indicating that they may be important factors influencing obesity.

While other statistical tests like ANOVA (Analysis of Variance) could also be applied to certain categorical variables, ANOVA is typically used for continuous variables with categorical factors. Since the variables under consideration are all categorical, the chi-square test is the appropriate choice for assessing their relationship with obesity levels.

Regression analysis

```
obesity_data <- obesity_data %>% rename( eats_high_calor_food = FAVC, eats_veggies = FCVC,
                              num_meals = NCP, eats_snacks = CAEC, drinks_water = CH2O,
                              counts_calories = SCC, exercises_often = FAF,
                              time_using_tech = TUE, drinks_alcohol = CALC,
                              method_trans = MTRANS, weight_category = NObeyesdad ) %>%
  mutate( bmi = Weight / Height^2 ) %>%
  mutate( weight_cat_num = case_when( ( weight_category == "Insufficient_Weight" ) ~ -1,
          ( weight_category == "Normal_Weight" ) ~ 0,
          ( weight_category == "Overweight_Level_I" ) ~ 1,
          ( weight_category == "Overweight_Level_II" ) ~ 2,
          ( weight_category == "Obesity_Type_I" ) ~ 3,
          ( weight_category == "Obesity_Type_II" ) ~ 4,
          ( weight_category == "Obesity_Type_III" ) ~ 5 ) )
```

```
glimpse(obesity_data)
```

```
## Rows: 2,111
## Columns: 19
## $ Gender                          <chr> "Female", "Female", "Male", "Male", "Ma~
## $ Age                             <dbl> 21, 21, 23, 27, 22, 29, 23, 22, 24, 22,~
## $ Height                          <dbl> 1.62, 1.52, 1.80, 1.80, 1.78, 1.62, 1.5~
## $ Weight                          <dbl> 64.0, 56.0, 77.0, 87.0, 89.8, 53.0, 55.~
## $ family_history_with_overweight <chr> "yes", "yes", "yes", "no", "no", "no", ~
## $ eats_high_calor_food            <chr> "no", "no", "no", "no", "no", "yes", "y~
## $ eats_veggies                    <dbl> 2, 3, 2, 3, 2, 2, 3, 2, 3, 2, 3, 2, 3, ~
## $ num_meals                       <dbl> 3, 3, 3, 3, 1, 3, 3, 3, 3, 3, 3, 3, 3, ~
## $ eats_snacks                     <chr> "Sometimes", "Sometimes", "Sometimes", ~
## $ SMOKE                           <chr> "no", "yes", "no", "no", "no", "no", "n~
## $ drinks_water                    <dbl> 2, 3, 2, 2, 2, 2, 2, 2, 2, 2, 3, 2, 3, ~
## $ counts_calories                 <chr> "no", "yes", "no", "no", "no", "no", "n~
```

```
## $ exercises_often                <dbl> 0, 3, 2, 2, 0, 0, 1, 3, 1, 1, 2, 2, 2, ~
## $ time_using_tech                <dbl> 1, 0, 1, 0, 0, 0, 0, 0, 1, 1, 2, 1, 0, ~
## $ drinks_alcohol                 <chr> "no", "Sometimes", "Frequently", "Frequ~
## $ method_trans                   <chr> "Public_Transportation", "Public_Transp~
## $ weight_category                <fct> Normal_Weight, Normal_Weight, Normal_We~
## $ bmi                            <dbl> 24.38653, 24.23823, 23.76543, 26.85185,~
## $ weight_cat_num                 <dbl> 0, 0, 0, 1, 2, 0, 0, 0, 0, 0, 3, 2, 0, ~
```

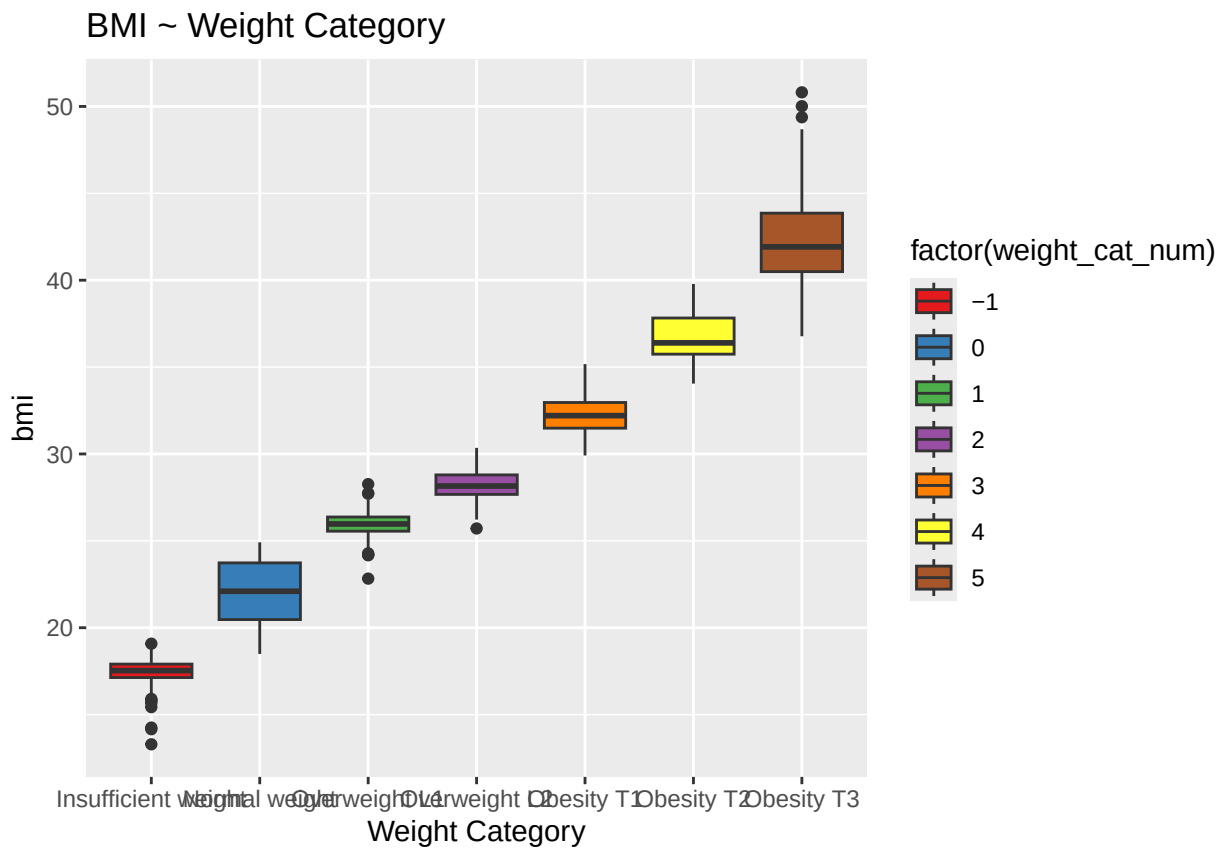
```
#train-test split
library(tidymodels)
```

```
## -- Attaching packages --------------------------------------- tidymodels 1.2.0 --

## v broom        1.0.5      v rsample      1.2.1
## v dials        1.2.1      v tibble       3.2.1
## v infer        1.0.7      v tune         1.2.1
## v modeldata    1.3.0      v workflows    1.1.4
## v parsnip      1.2.1      v workflowsets 1.1.0
## v purrr        1.0.2      v yardstick    1.3.1
## v recipes      1.0.10

## -- Conflicts ------------------------------------------ tidymodels_conflicts() --
## x purrr::discard()         masks scales::discard()
## x dplyr::filter()          masks stats::filter()
## x dplyr::lag()             masks stats::lag()
## x purrr::lift()            masks caret::lift()
## x yardstick::precision()   masks caret::precision()
## x yardstick::recall()      masks caret::recall()
## x yardstick::sensitivity() masks caret::sensitivity()
## x yardstick::specificity() masks caret::specificity()
## x recipes::step()          masks stats::step()
## * Dig deeper into tidy modeling with R at https://www.tmwr.org
```

```
set.seed( 138 )
obesity_split <- initial_split( obesity_data)
train_obesity <- training( obesity_split )
test_obesity <- testing( obesity_split )
```

Visualizing bmi with weight_cat_num

```
ggplot(train_obesity, aes(y = bmi, x = factor(weight_cat_num), fill = factor(weight_cat_num))) +
  geom_boxplot() +
  ggtitle('BMI ~ Weight Category') +
  xlab('Weight Category') +
  scale_x_discrete(breaks = c('-1', '0', '1', '2', '3', '4', '5'),
                   labels = c("Insufficient weight", "Normal weight", "Overweight L1", "Overweight L2",
  scale_fill_manual(values = c('#E41A1C', '#377EB8', '#4DAF4A', '#984EA3', '#FF7F00', '#FFFF33', '#A656
```



Fitting a linear regression model

```
model_lm <- lm( bmi ~ weight_cat_num, data = train_obesity )
summary( model_lm )
```

```
##
## Call:
## lm(formula = bmi ~ weight_cat_num, data = train_obesity)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.297 -1.145 -0.081  0.841  9.740
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    21.34955    0.06240   342.1   <2e-16 ***
## weight_cat_num  3.94438    0.02151   183.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.683 on 1581 degrees of freedom
## Multiple R-squared:  0.9551, Adjusted R-squared:  0.955
## F-statistic: 3.361e+04 on 1 and 1581 DF,  p-value: < 2.2e-16
```

```r
# Hypothesis testing
# Null Hypothesis: There is no significant relationship between weight category and BMI
# Alternative Hypothesis: There is a significant relationship between weight category and BMI

# Extract coefficients and p-values
coef_estimate <- coef(summary(model_lm))[2, 1]  # Coefficient estimate for weight_cat_num
p_value <- coef(summary(model_lm))[2, 4]        # p-value for weight_cat_num

# Print coefficient estimate and p-value
print(paste("Coefficient Estimate:", coef_estimate))
```

```
## [1] "Coefficient Estimate: 3.94437856682777"
```

```r
print(paste("p-value:", p_value))
```

```
## [1] "p-value: 0"
```

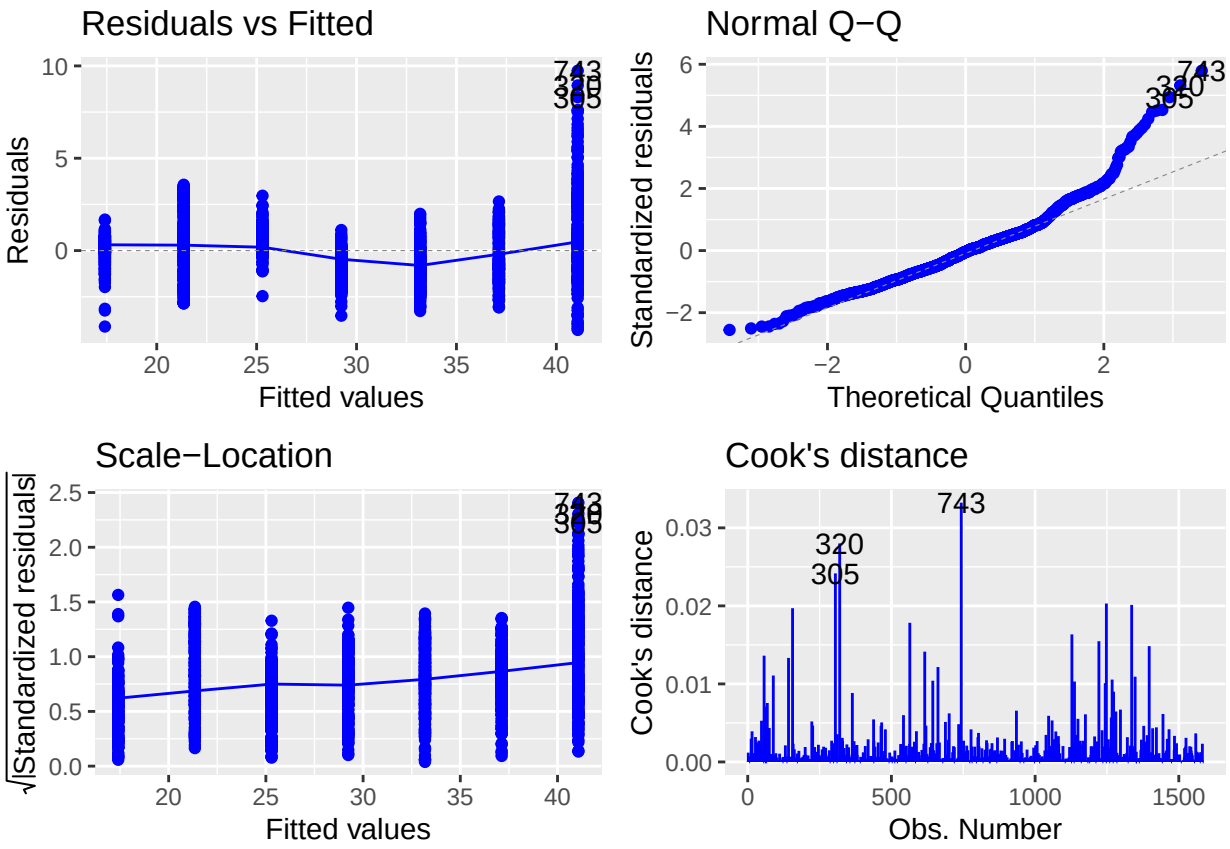
```r
# Conclusion based on p-value
if (p_value < 0.05) {
  print("Reject the null hypothesis: There is a significant relationship between weight category and BM
} else {
  print("Fail to reject the null hypothesis: There is no significant relationship between weight catego
}
```

```
## [1] "Reject the null hypothesis: There is a significant relationship between weight category and BMI
```

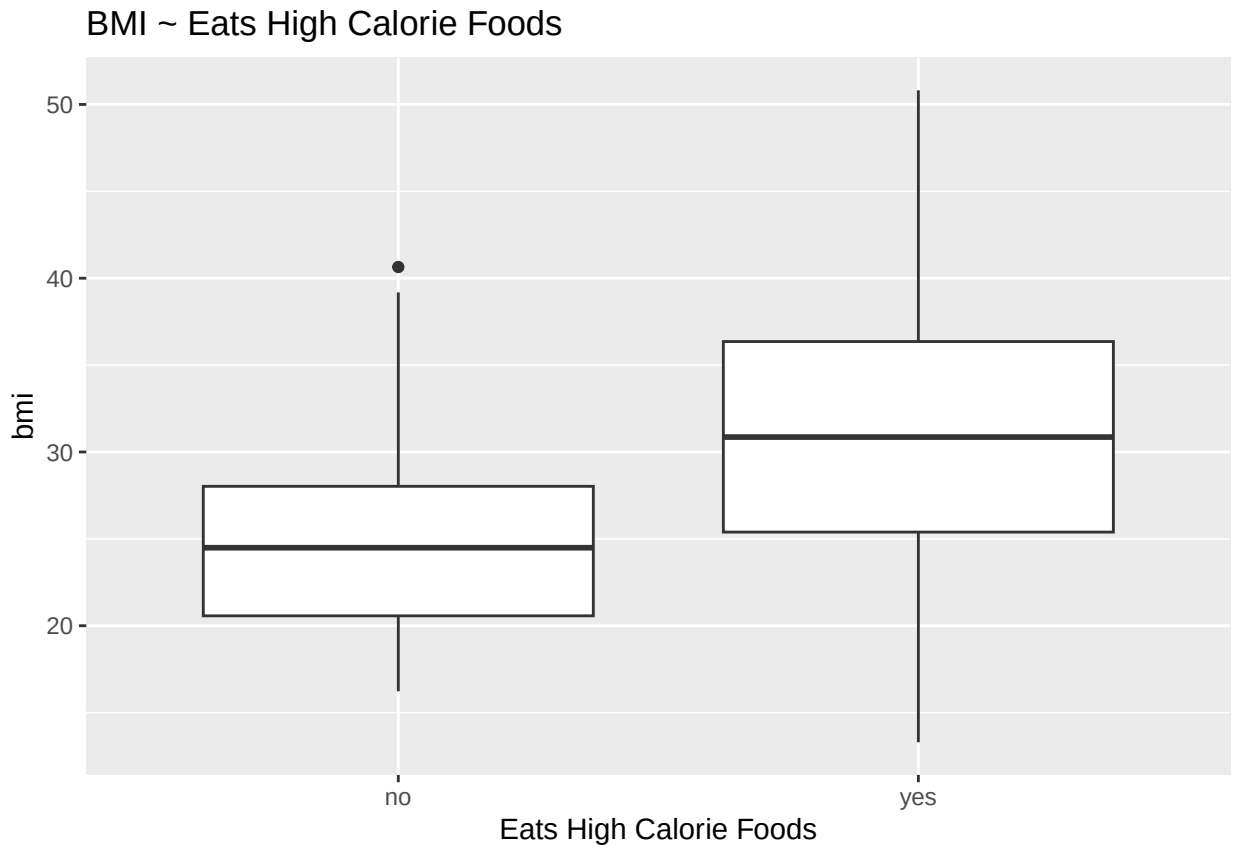
```r
library(ggfortify)
```

```
## Registered S3 method overwritten by 'ggfortify':
##   method          from
##   autoplot.glmnet parsnip
```

```r
autoplot(model_lm, which=1:4, colour= "blue")
```




```
ggplot( train_obesity, aes( y = bmi, x = factor( eats_high_calor_food ) ) ) +
  geom_boxplot() +
  ggtitle( 'BMI ~ Eats High Calorie Foods' ) +
  xlab( 'Eats High Calorie Foods' )
```




```
multi_lm <- lm( bmi ~ eats_high_calor_food + family_history_with_overweight +
                      SMOKE + counts_calories + Gender, data = train_obesity )
summary( multi_lm )
```

```
##
## Call:
## lm(formula = bmi ~ eats_high_calor_food + family_history_with_overweight +
##     SMOKE + counts_calories + Gender, data = train_obesity)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.7283  -4.7842  -0.1968   5.1778  18.8219
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       19.9755     0.6267  31.875  < 2e-16 ***
## eats_high_calor_foodyes            3.8229     0.5599   6.828 1.23e-11 ***
## family_history_with_overweightyes  9.1228     0.4637  19.674  < 2e-16 ***
## SMOKEyes                          -0.1663     1.4022  -0.119    0.906
## counts_caloriesyes                -3.2950     0.8317  -3.962 7.77e-05 ***
## GenderMale                        -1.9013     0.3448  -5.515 4.07e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.798 on 1577 degrees of freedom
```

```
## Multiple R-squared:  0.2688, Adjusted R-squared:  0.2664
## F-statistic: 115.9 on 5 and 1577 DF,  p-value: < 2.2e-16
```

```
# Hypothesis testing
# Null Hypothesis: There is no significant relationship between the predictor variables and BMI
# Alternative Hypothesis: At least one predictor variable has a significant relationship with BMI

# Extract coefficients and p-values
coef_summary <- summary(multi_lm)
coef_estimates <- coef_summary$coefficients[, 1]  # Coefficient estimates
p_values <- coef_summary$coefficients[, 4]        # p-values

# Print coefficient estimates and p-values
print("Coefficient Estimates:")
```

```
## [1] "Coefficient Estimates:"
```

```
print(coef_estimates)
```

```
##                       (Intercept)          eats_high_calor_foodyes
##                        19.9754914                        3.8228735
## family_history_with_overweightyes                         SMOKEyes
##                         9.1228015                       -0.1663324
##                counts_caloriesyes                       GenderMale
##                        -3.2949754                       -1.9012721
```

```
print("p-values:")
```

```
## [1] "p-values:"
```

```
print(p_values)
```

```
##                       (Intercept)          eats_high_calor_foodyes
##                     1.633072e-172                     1.226269e-11
## family_history_with_overweightyes                         SMOKEyes
##                      3.076426e-77                     9.055917e-01
##                counts_caloriesyes                       GenderMale
##                      7.765633e-05                     4.071527e-08
```

```
# Conclusion based on p-values
if (any(p_values < 0.05)) {
  print("Reject the null hypothesis: At least one predictor variable has a significant relationship wit
} else {
  print("Fail to reject the null hypothesis: There is no significant relationship between the predictor
}
```

```
## [1] "Reject the null hypothesis: At least one predictor variable has a significant relationship with
```

```r
multi_lm <- lm( bmi ~ eats_high_calor_food + family_history_with_overweight +
                      counts_calories + Gender, data = train_obesity )
summary( multi_lm )
```

```
##
## Call:
## lm(formula = bmi ~ eats_high_calor_food + family_history_with_overweight +
##     counts_calories + Gender, data = train_obesity)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.7259  -4.7829  -0.1944   5.1755  18.8293
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        19.9699     0.6247  31.965  < 2e-16 ***
## eats_high_calor_foodyes             3.8276     0.5583   6.856 1.02e-11 ***
## family_history_with_overweightyes   9.1218     0.4635  19.682  < 2e-16 ***
## counts_caloriesyes                 -3.2962     0.8313  -3.965 7.67e-05 ***
## GenderMale                         -1.9018     0.3446  -5.519 3.99e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.796 on 1578 degrees of freedom
## Multiple R-squared:  0.2687, Adjusted R-squared:  0.2669
## F-statistic:   145 on 4 and 1578 DF,  p-value: < 2.2e-16
```

```r
# Hypothesis testing
# Null Hypothesis: There is no significant relationship between the predictor variables and BMI
# Alternative Hypothesis: At least one predictor variable has a significant relationship with BMI

# Extract coefficients and p-values
coef_summary <- summary(multi_lm)
coef_estimates <- coef_summary$coefficients[, 1]  # Coefficient estimates
p_values <- coef_summary$coefficients[, 4]        # p-values

# Print coefficient estimates and p-values
print("Coefficient Estimates:")
```

```
## [1] "Coefficient Estimates:"
```

```r
print(coef_estimates)
```

```
##                       (Intercept)           eats_high_calor_foodyes
##                         19.969939                          3.827577
## family_history_with_overweightyes                counts_caloriesyes
##                          9.121823                         -3.296238
##                        GenderMale
##                         -1.901825
```

```r
print("p-values:")
```

```
## [1] "p-values:"
```

```r
print(p_values)
```

```
##                         (Intercept)           eats_high_calor_foodyes
##                        2.678122e-173                      1.015724e-11
## family_history_with_overweightyes                 counts_caloriesyes
##                        2.716119e-77                      7.666742e-05
##                          GenderMale
##                        3.985029e-08
```

```r
# Conclusion based on p-values
if (any(p_values < 0.05)) {
  print("Reject the null hypothesis: At least one predictor variable has a significant relationship wit
} else {
  print("Fail to reject the null hypothesis: There is no significant relationship between the predictor
}
```

```
## [1] "Reject the null hypothesis: At least one predictor variable has a significant relationship with
```

```r
bmi_prediction <- multi_lm %>% predict(test_obesity)
rsquare_test <- rsq_vec(test_obesity$bmi, bmi_prediction)
rsquare_test
```

```
## [1] 0.2910367
```