Term Project

Sharma, Biswajit

2023-11-05

Analysis and Prediction of Obesity based on eating habits and physical activity

Introduction

Obesity is a medical condition that is related to the excessive accumulation of body fat. It is not only a cosmetic concern but rather a medical problem that can increase the risk of other health problems and diseases like heart diseases, diabetes, high cholesterol, high blood pressure, liver disease, musculoskeletal disorders, and certain cancers. Since 1997, WHO has considered obesity a global epidemic and a significant health problem. To prevent obesity, various organizations, including government and non-government, are promoting campaigns regarding two main risk factors: eating habits and physical activity (Gozukara et al., 2023).

Although it is known that excessive intake of calories can cause obesity, nutritional factors like low-quality diet, unbalanced diet, processed foods, and alcohol consumption can also increase the risk of obesity. Physical activity has also been very influential in controlling or preventing obesity. The frequency, duration, and intensity of physical activity and exercises play an essential role in the effective prevention and reduction of obesity (Gozukara et al., 2023).

Obesity levels are measured by BMI (Body Mass Index) which is the ratio between body weight and height, calculated as below equation (Body Mass Index, CDC, n.d.).

$BMI = \frac{BodyWeight}{Height^2}$ (Equation. 1)

BMI greater than 30 is considered *Obesity* and between 25 to 30 is considered *Overweight* (Defining Adult Overweight & Obesity, CDC, n.d.).

This analysis will investigate the relationship of Obesity and Body Mass Index (BMI) with eating habits and physical activity. This study can help to identify patterns associated with obesity, such as the quality and quantity of food intake, frequency, duration, and intensity of physical activity. This study can also help to determine the optimal amount and type of physical activity for each obesity level based on the individual's age and gender. Similarly, this study can help to design a balanced and nutritious diet plan for each obesity level. Furthermore, this study will generate a model to estimate obesity and BMI based on eating habits and physical activity.

Research Questions

1. Determine if we can estimate the BMI (Body Mass Index) and obesity based on eating habits and physical activity.
2. Explore how BMI and obesity vary with the weight and height.
3. Investigate the relationship between height and weight.
4. Investigate the relationship between vegetable consumption and weight and BMI.
5. Investigate the relationship between alcohol consumption and weight and BMI
6. Investigate the relationship between taking food between meals and weight and BMI

7. Investigate the relationship between calories consumed and weight and BMI
8. Investigate the relationship of physical activity with weight, height and BMI
9. Investigate the relationship between family history of overweight and weight and BMI
10. Explore if the factors other than eating habits and physical activity like ethnicity and geography also influence obesity.

Approach

As mentioned above (Equation. 1), BMI is calculated from weight and height. Therefore, in order to determine the relationship between obesity and eating habits, and physical activity, we need to identify that there is a relationship between weight, height, eating habits, and physical activity. If we find a relationship between eating habits and physical activity with weight and height. In that case, we can estimate the influence (negative or positive) of these factors on BMI and obesity.

Exploratory data analysis will help to identify relationships, patterns, and gather insights from the dataset. The correlation matrix will help to identify the correlation between the numerical variables and BMI. T-tests will confirm if there is evidence of a relationship between eating habits, physical activity, and BMI. ANOVA will help to analyze the variance and how much variation in the BMI is accounted for by eating habits and physical activity.

Various plots like histograms and QQ plots will identify the distribution of the variables regarding various obesity levels. Box plots will aid in determining five-point summary statistics and locating outliers. Scatter plots will help to identify any linear relationship between BMI, weight, height, eating habits, and physical activity.

Furthermore, we can model the data to predict BMI and obesity based on eating habits and physical activity.

Data

1. *UC Irvine Machine Learning Repository - Obesity levels, Eating Having and Physical activity dataset.*

This dataset include data about eating habits, physical activity, weight, height and obesity levels of individuals from the countries of mexico, Peru and Columbia. The data includes the eating habits and physical activity levels of 498 participants aged between 14 and 61 years (UCI.2019).

The originally collected data was preprocessed, such as the removal of missing values, and normalization was performed. It was also balanced to reduce the skewness of the obesity levels. 23% of the source data is actual responses collected over a 30-day survey, while the remaining 77% was synthetically generated using SMOTE (Palechor & de la Hoz Manotas, 2019).

We will use this dataset to perform analysis and mining of relationship of weight, height and obesity with eating Having and physical activity. We will also use this data for prediction of obesity using eating habits and physical activity as predictors.

Variables:

There are 17 variables in the dataset.

```
- Gender
- Age
- Height
- Weight
- Family History of overweight
- Frequency of consumption of high caloric food (FAVC)
- Frequency of consumption of vegetables (FCVC)
- Number of main meals (NCP)
- Consumption of food between meals (CAEC)
- Daily consumption of water (CH20)
- Consumption of alcohol (CALC)
```

```
    - Calorie consumption monitoring done (SCC)
    - Frequency of Physical activity (FAF)
    - Measuring Physical activity time using devices (TUE)
    - Mode of transportation used (MTRANS)
    - Obesity Level
```

Obesity level is labelled in the source data based on mass body index calculation (Equation. 1) and then compared with data provided by WHO (Palechor & de la Hoz Manotas, 2019).

```
  - Underweight when BMI Less than 18.5
  - Normal when BMI 18.5 to 24.9
  - Overweight when BMI 25.0 to 29.9
  - Obesity I when BMI 30.0 to 34.9
  - Obesity II when BMI 35.0 to 39.9
  - Obesity III when BMI Higher than 40
```

2. *CDC Nutrition, Physical Activity and Obesity.*

This dataset includes data on adults diet, physical activity and weight status from Behavioral Risk Factor Surveillance System (Nutrition, Physical Activity, and Obesity, CDC, 2023).

This dataset has 33 columns, however the main columns which will be used in this study are given below.

```
    - LocationAbbr (Location of the datasoure like particular US State or territories)
    - Topic (Obesity or Physical Activity)
    - Question (Question regarding obesity, overweight and physical activity)
    - Data Value (in percentage of population)
    - Race/Ethnicity
```

This dataset will be used to analyze the distribution and variation in the percentage of obesity in adults based on geography and ethnicity. Analysis of the dataset will help identify whether obesity varies with race/ethnicity and geography.

Required packages

This study will require below R packages for ingestion, analysis, tranaformation, manipulation, visualizations and modeling of data.

- dplyr
- ggplot2
- purrr
- tidyr
- metrics
- car
- mlogit
- magrittr
- reshape2
- tidyr
- usmap

Plots

Various plots like histograms and QQ plots will identify the distribution of the variables regarding various obesity levels. Box plots will aid in determining five-point summary statistics and locating outliers. Scatter plots will help to identify any linear relationship between BMI, weight, height, eating habits, and physical activity. Residual plots and normality plots will be used to identify normality assumption and presence of bias in the model.

Questions for future steps

1. Explore if there are duplicate rows or missing data in the dataset.
2. Explore if multicollinearity exists between the variables.
3. Determine if we need to transform variables or generate new variables for modeling. Check for outliers and identify if outliers need to be removed before modeling. Identify if we need any transformation of categorical to numeric variables.
4. Suppose we create a model to estimate obesity and BMI based on eating habits and physical activity. How well does that model fit the data, whether the model generalizes well or the generated model has any bias? Investigate whether the model follows the assumptions of linearity, homoscedasticity, and normality.
5. Determine which variables from the dataset we need to use as predictors in modeling.

Milestone #2

Data import and Cleaning

Import and inspect structure of the dataset (UCI, 2019).

```
obesity_df <-
↪  read.csv("C:/Users/babub/Documents/Bellevue/DSC520/ObesityDataSet_raw_and_data_sinthetic.csv")

str(obesity_df)
```

```
## '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_
```

View few rows of the dataset

```
##   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
```

Data cleaning and transformation

Check for NAs in the dataset

```r
colSums(is.na(obesity_df))
```

```
##                         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 see that there are no NAs in the dataset. Lets check for duplicate rows.

```r
obesity_df %>% dplyr::group_by_all() %>% filter(n() > 1) %>% ungroup() %>% nrow()
```

```
## [1] 33
```

We see that there are 33 duplicate rows. Lets drop the duplicate rows.

```r
obesity_df <- obesity_df %>% distinct(.keep_all = TRUE)
```

Lets Validate that there are no more duplicate rows.

```r
obesity_df %>% dplyr::group_by_all() %>% filter(n() > 1) %>% ungroup() %>% nrow()
```

```
## [1] 0
```

New feature generation

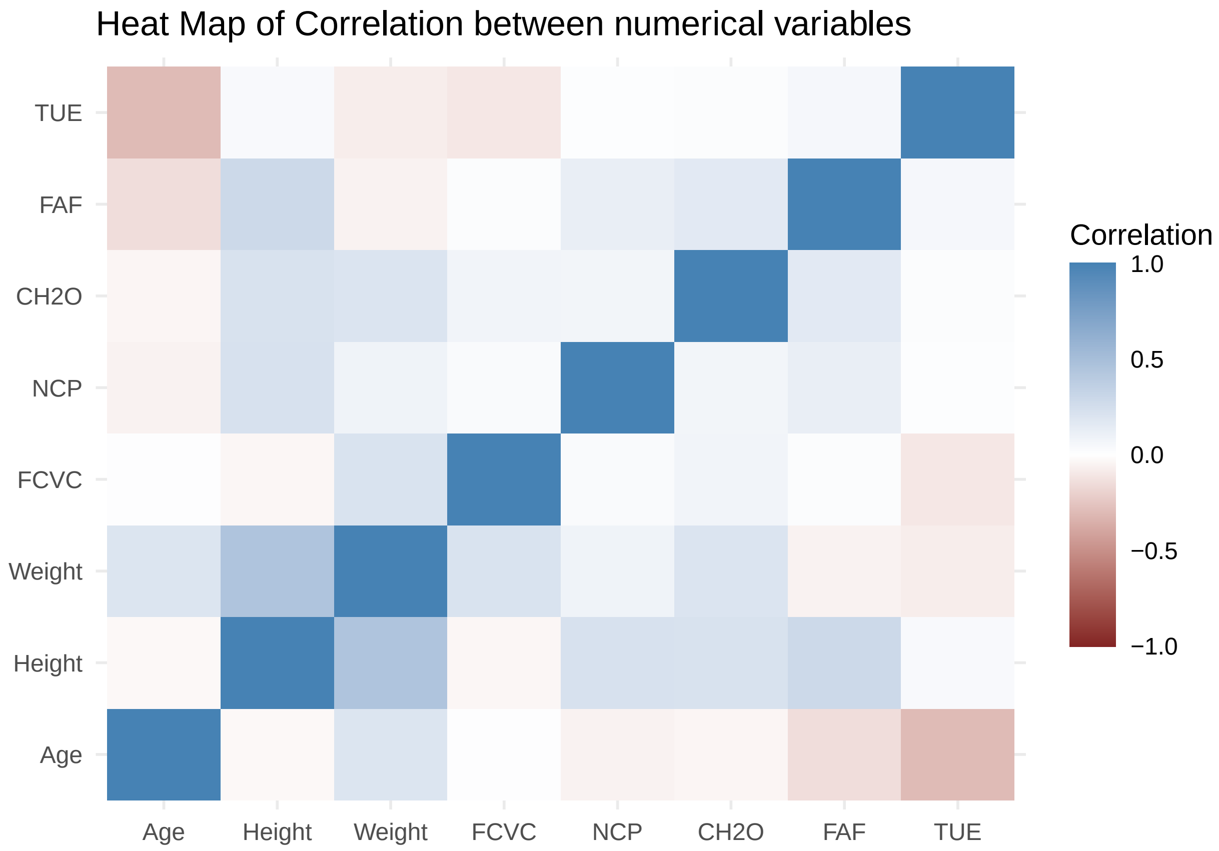
We see that BMI is not available in the dataset. However, we can calculate the BMI using equation 1. Lets create a new column named BMI using equation 1.

```
obesity_df <- obesity_df %>% mutate(BMI=Weight/(Height^2))
```

Correlation between Numeric variables

Lets check the correlation between the numeric variables using ***heat map*** to identify high collinearity.

```
obesity.cor <- cor(obesity_df %>% select(Age, Height, Weight, FCVC, NCP, CH2O, FAF, TUE))
obesity.melt <- melt(obesity.cor, varname=c("x","y"), value.name="Correlation")
ggplot(obesity.melt, aes(x=x,y=y)) + geom_tile(aes(fill=Correlation)) +
  scale_fill_gradient2(low=muted("red"),
                       mid="white", high="steelblue",
                       guide=guide_colorbar(ticks=FALSE, barheight=10),
                       limits=c(-1,1)) +
  theme_minimal() +
  labs(title="Heat Map of Correlation between numerical variables", x=NULL,y=NULL)
```



Based on the above heat map, we observe:

- a moderate positive correlation between *Height* and *Weight*
- a moderate negative correlation between *Age* and *Measuring Physical activity time using devices (TUE)*
- a mild negative correlation between *FAF* and *Age*
- a mild correlation between *Age* and *weight*
- a mild positive correlation between *Physical Activity (FAF)* , *Number of Meals (NCP)* and *Height*
- a mild positive correlation between *Frequency of consumption of vegetables (FCVC)* and *Weight*
- a small positive correlation between *Consumption of water (CH20)* and *Weight*
- a mild positive correlation between *Height* and *Physical Activity (FAF)*
- a mild negative correlation between *Measuring Physical activity time using devices (TUE)* and *Con-*

sumption of vegetables (FCVC)

As we do not see high correlation between the numerical variables, we can say that there is no concern of multicollinearity if these numerical variables are used as predictors in a model.

Investigate the relationship between Height and Weight.

From above *heat map* we see a moderate positive correlation between Height and Weight, lets perform a *t-test* to check if the relationship also exists in population.

```r
cor.test(obesity_df$Height, obesity_df$Weight)
```

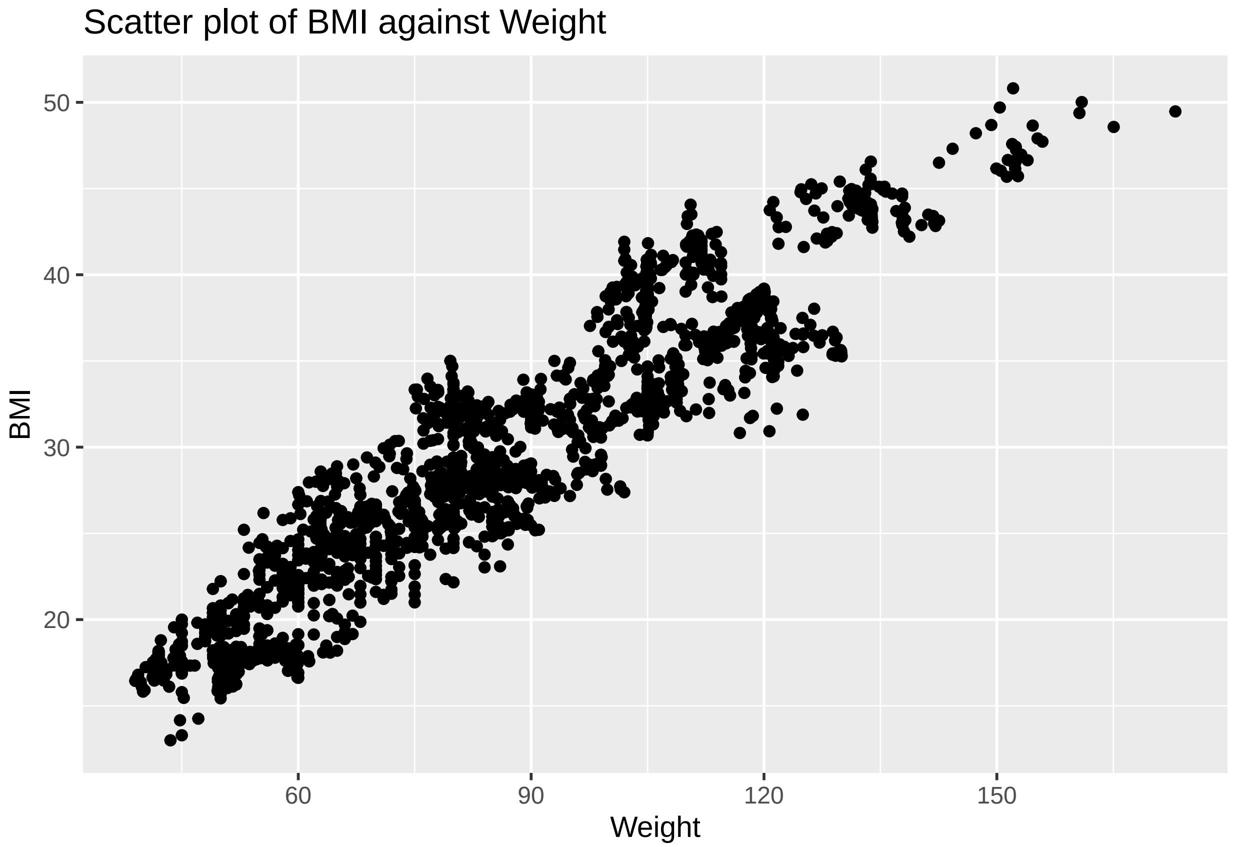
```
##
##  Pearson's product-moment correlation
##
## data:  obesity_df$Height and obesity_df$Weight
## t = 23.491, df = 2085, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4228608 0.4907428
## sample estimates:
##      cor
## 0.457468
```

The t-test show that there is evidence to reject *Null Hypothesis* and we can say that there is relationship between Height and Weight in the population of the sample as p-value < `0.001`. The *correlation coefficient* is `0.46` which shows that the relationship is positive and Weight tends to increase with an increase in Height.

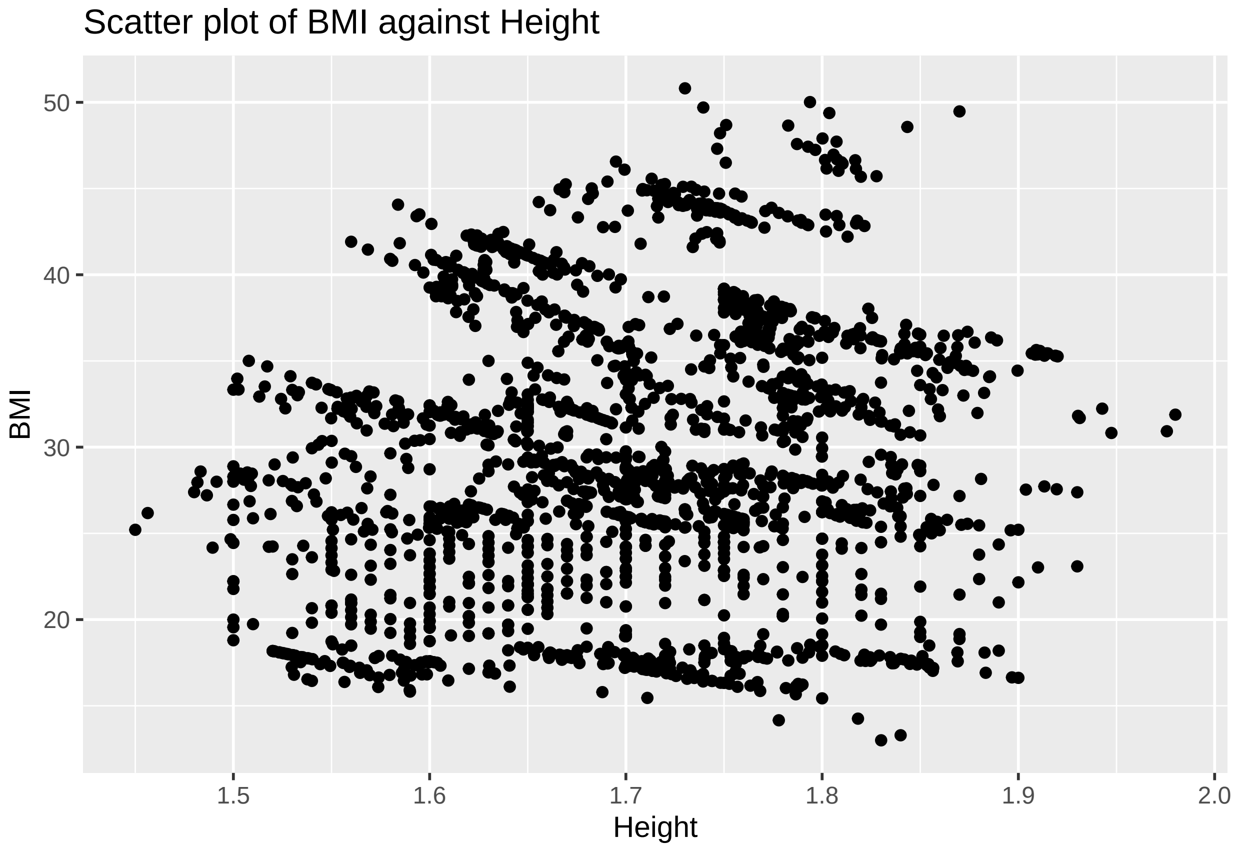
Explore how BMI and obesity vary with the weight and height.

Let plot some scatter plots to check the BMI against weight and height,

```r
ggplot(obesity_df, aes(x=Weight, y=BMI)) + geom_point() + labs(title="Scatter plot of BMI
↪  against Weight",x="Weight", y="BMI")
ggplot(obesity_df, aes(x=Height, y=BMI)) + geom_point() + labs(title="Scatter plot of BMI
↪  against Height", x="Height", y="BMI")
```



Scatter plot of BMI against Weight



The scatter plots show that there is very high *positive correlation* between BMI and Weight. However, the relationship between BMI and Height is inconclusive from the scatter plot as the plot has random points.

Lets perform a `t-test` to investigate the relationship BMI and Weight in the population.

```
cor.test(obesity_df$BMI, obesity_df$Weight)
```

```
##
##  Pearson's product-moment correlation
##
## data:  obesity_df$BMI and obesity_df$Weight
## t = 119.87, df = 2085, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.928830 0.939722
## sample estimates:
##       cor
## 0.9344944
```

The t-test show that there is evidence to reject *Null Hypothesis* and we can say that there is relationship between BMI and Weight in the population of the sample as p < `0.001`. The *correlation coefficient* is `0.93` which shows that there is a strong positive correlation between BMI and weight and BMI tends to increase proportionately with an increase in Weight.

Lets perform a `t-test` to investigate the relationship BMI and Height in the population.

```
cor.test(obesity_df$BMI, obesity_df$Height)
```

```
##
```

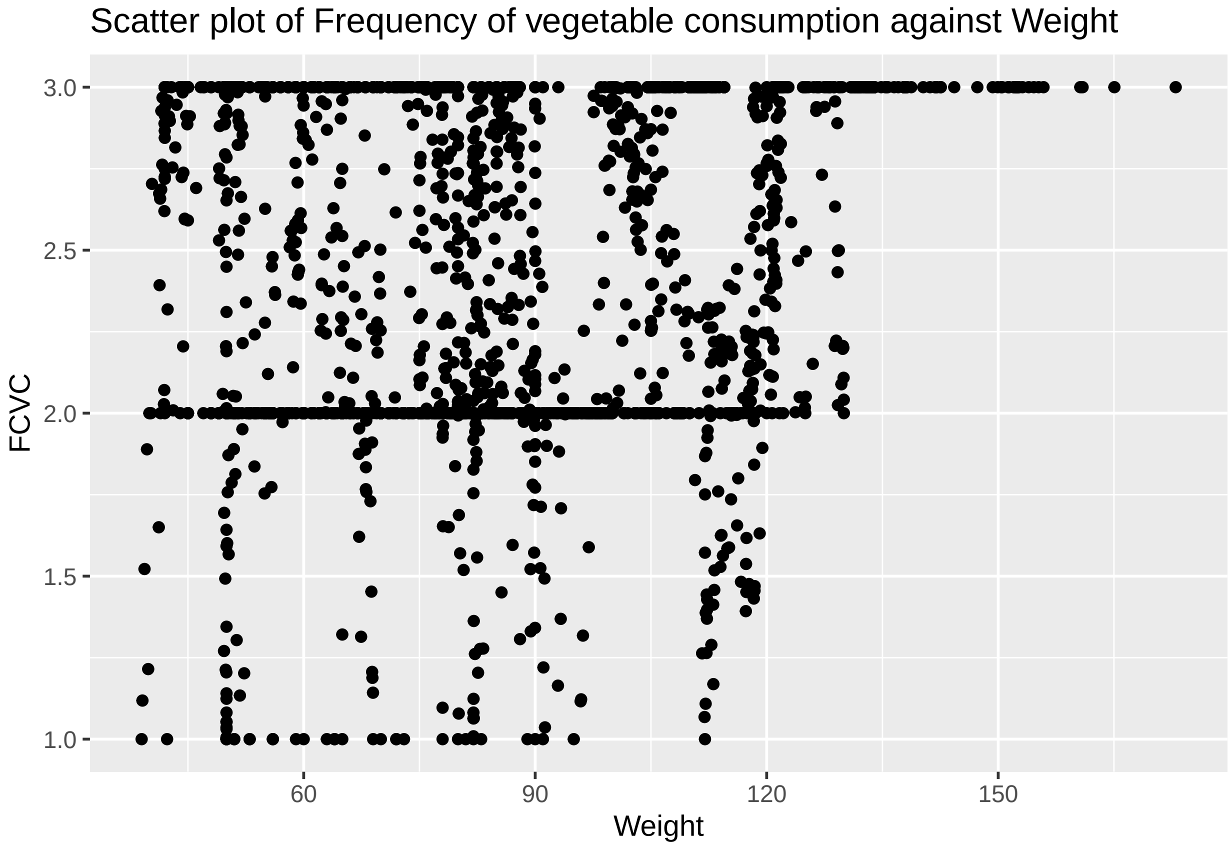
```
##  Pearson's product-moment correlation
##
## data:  obesity_df$BMI and obesity_df$Height
## t = 5.7279, df = 2085, p-value = 1.165e-08
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.08199612 0.16648397
## sample estimates:
##       cor
## 0.1244657
```

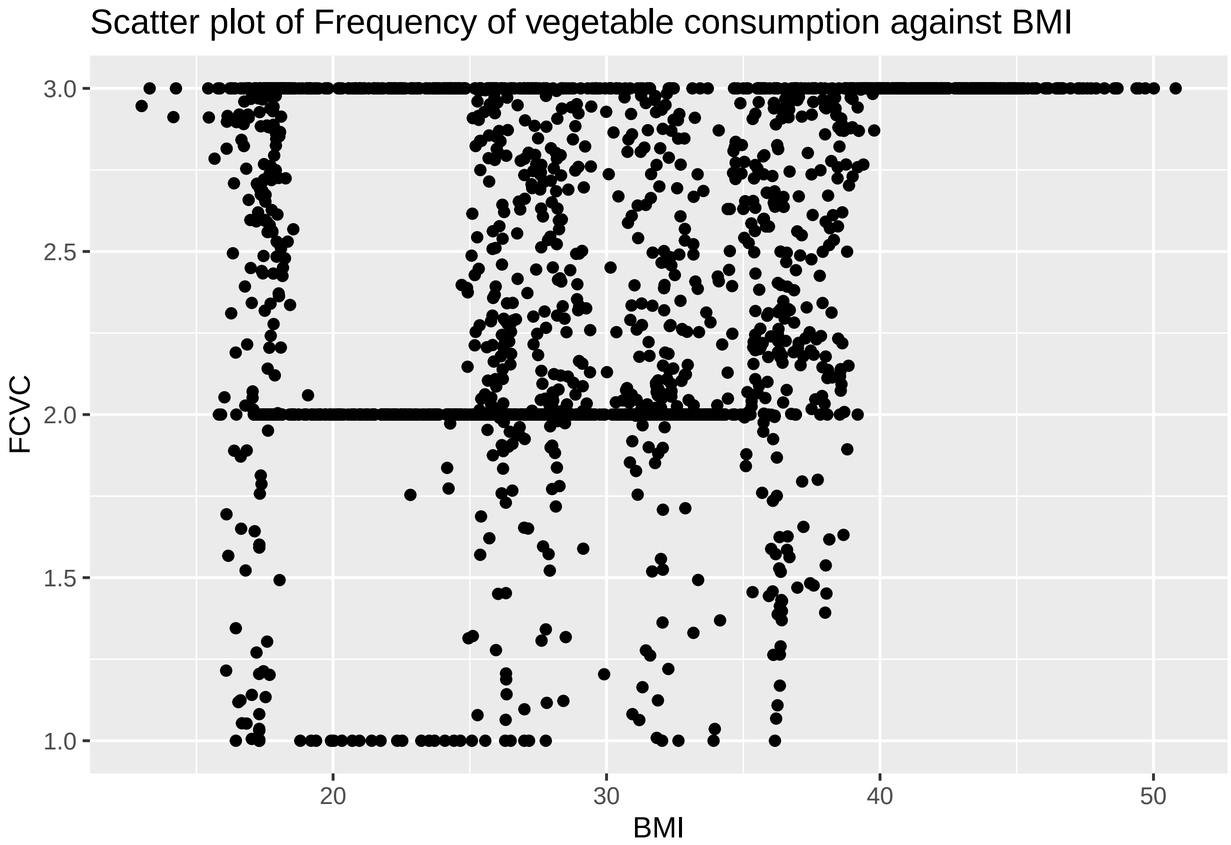
The t-test show that there is evidence to reject *Null Hypothesis* and we can say that there is relationship between BMI and weight in the population of the sample as p < 0.001. The *correlation coefficient* is `0.12` which shows a small positive correlation between BMI and Height.

Investigate the relationship between vegetable consumption and weight and BMI.

Lets plot some scatter plots to check the Frequency of Consumption of vegetables (FCVC) against weight and BMI.

```
ggplot(obesity_df, aes(x=Weight, y=FCVC)) + geom_point() + labs(title="Scatter plot of
↪  Frequency of vegetable consumption against Weight",x="Weight", y="FCVC")
ggplot(obesity_df, aes(x=BMI, y=FCVC)) + geom_point() + labs(title="Scatter plot of
↪  Frequency of vegetable consumption against BMI", x="BMI", y="FCVC")
```





No apparent relationship is conclusive from the scatter plots. Lets perform a `t-test` to check any relationship between these variables in the population.

```
cor.test(obesity_df$Weight, obesity_df$FCVC)
```

```
##
##  Pearson's product-moment correlation
##
## data:  obesity_df$Weight and obesity_df$FCVC
## t = 10.13, df = 2085, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1752959 0.2570928
## sample estimates:
##       cor
## 0.2165744
```

The bove t-test show that there is evidence to reject *Null Hypothesis* and we can say that there is some relationship between Weight and Frequency of vegetable consumption (FCVC) in the population of the sample as $p < 0.001$. The *correlation coefficient* is `0.22` which shows that there is a mild positive correlation between Weight and Frequency of vegetable consumption (FCVC).

```
cor.test(obesity_df$BMI, obesity_df$FCVC)
```

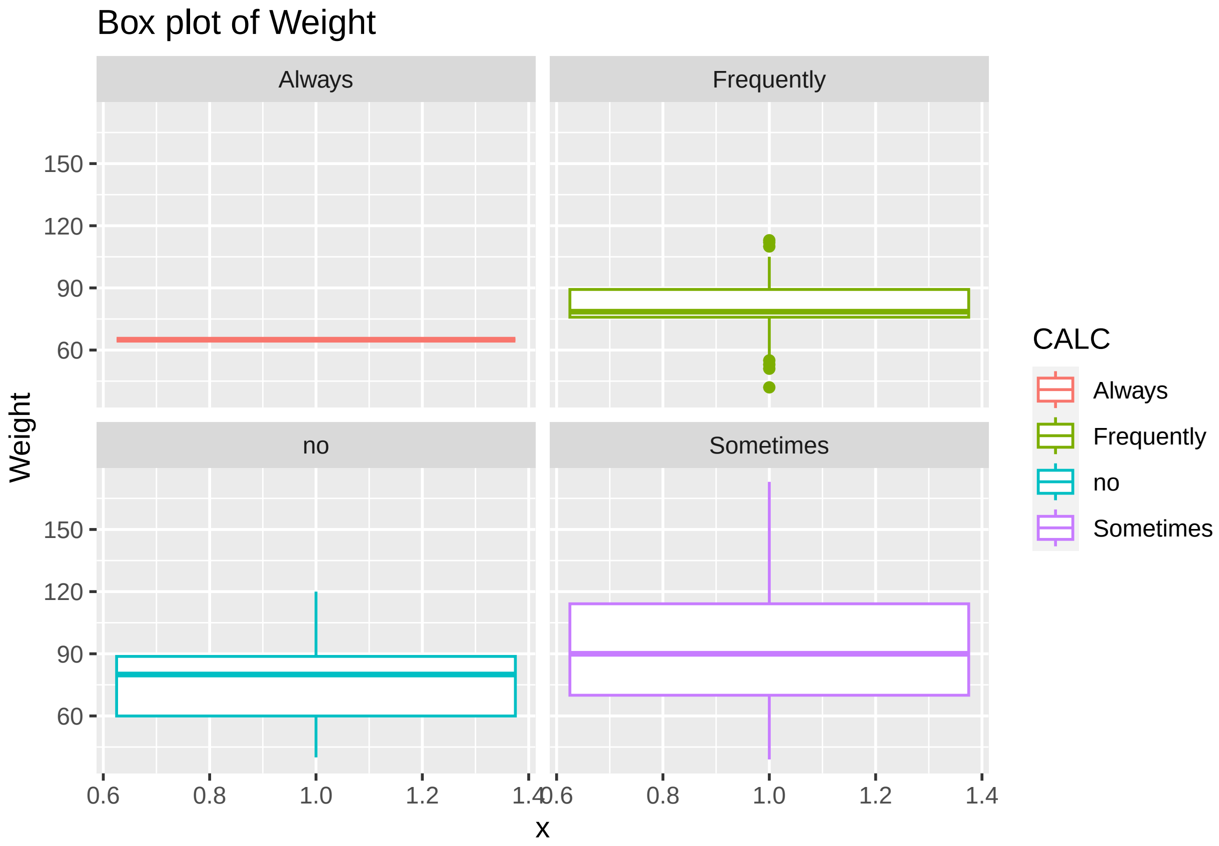
```
##
##  Pearson's product-moment correlation
##
## data:  obesity_df$BMI and obesity_df$FCVC
## t = 12.553, df = 2085, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2247306 0.3045257
## sample estimates:
##      cor
## 0.265082
```

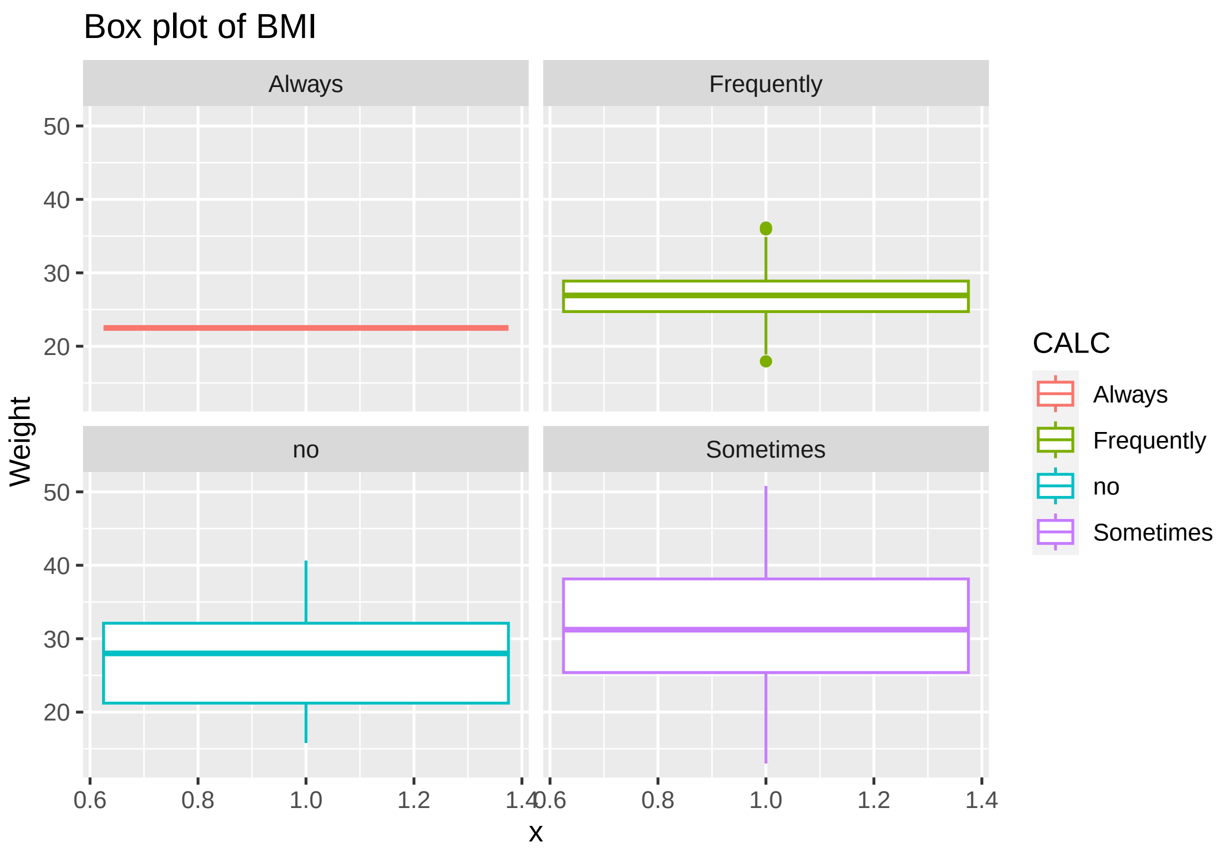
The above t-test show that there is evidence to reject *Null Hypothesis* and we can say that there is some relationship between Weight and FCVC in the population of the sample as p < 0.001. The *correlation coefficient* is 0.26 which shows that there is a mild positive correlation between BMI and Frequency of vegetable consumption (FCVC).

Investigate the relationship between alcohol consumption (CALC) and Weight and BMI

As alcohol consumption (CALC) is a categorical variable we will need to see the distribution of BMI across the alcohol consumption categories.

```
ggplot(obesity_df) +
  geom_boxplot(aes(y=Weight, x=1,color=CALC)) +
  labs(title="Box plot of Weight",y="Weight") + facet_wrap(~CALC)

ggplot(obesity_df) +
  geom_boxplot(aes(y=BMI, x=1,color=CALC)) +
  labs(title="Box plot of BMI",y="Weight") + facet_wrap(~CALC)
```





Based on the above plots, we can see that the consumption of alcohol have a tendency for higher BMI and Weight than those who do not consume alcohol. This might be an indication of relationship between some elevated alcohol consumption and obesity.

Lets create a crosstable between alcohol consumption and obesity levels.

```
##
##              Insufficient_Weight Normal_Weight Obesity_Type_I Obesity_Type_II
##   Always                       0             1              0               0
##   Frequently                   1            18             14               2
##   no                         117           104            165              71
##   Sometimes                  149           159            172             224
##   sum                        267           282            351             297
##
##              Obesity_Type_III Overweight_Level_I Overweight_Level_II  sum
##   Always                    0                  0                   0    1
##   Frequently                0                 16                  19   70
##   no                        1                 50                 128  636
##   Sometimes               323                210                 143 1380
##   sum                     324                276                 290 2087
```

```
((351+297+324+276+290) - (165+71+1+50+128))/(351+297+324+276+290) * 100
```

```
## [1] 73.01691
```

The above table with marginal totals shows that the number of persons having obesity and who consume alcohol are much higher than who doesn't. `73%` of the persons having obesity consume alcohol.

Lets perform a `chi-square` test to check for relationship will exist in the population.

```
chisq.test(obesity_df$CALC, obesity_df$NObeyesdad)
```

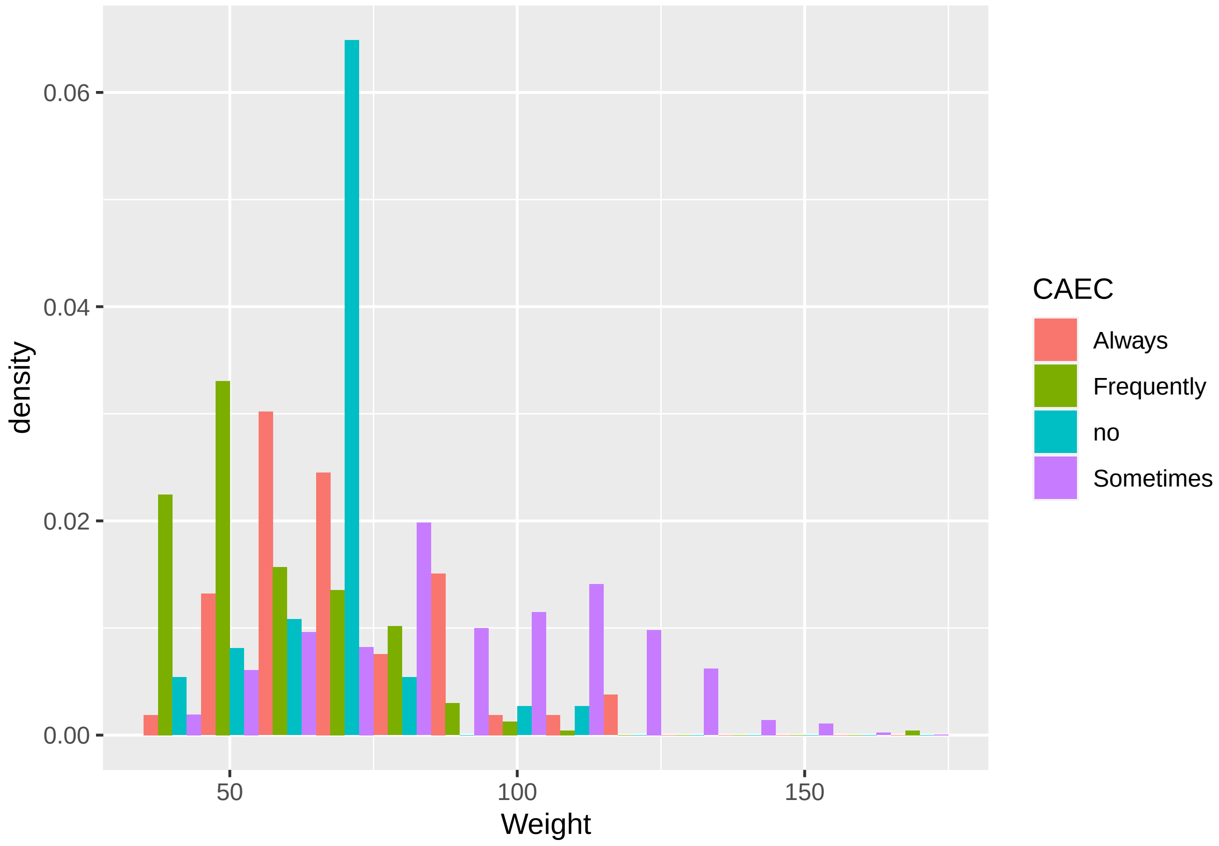
```
##
##  Pearson's Chi-squared test
##
## data:  obesity_df$CALC and obesity_df$NObeyesdad
## X-squared = 335.56, df = 18, p-value < 2.2e-16
```

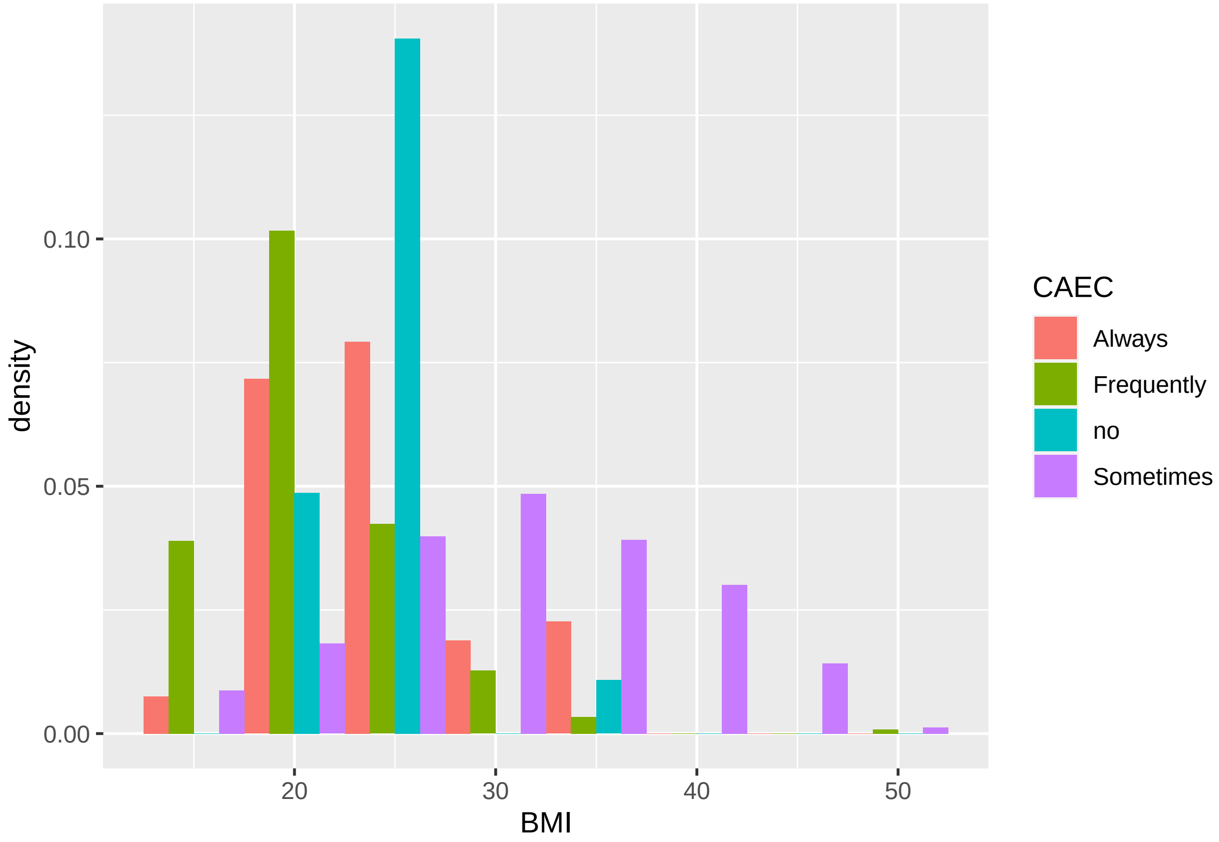
Chi-sqaure test show that there is evidence to reject the *Null Hypothesis* and conclude that there is some *relationship between alcohol consumption and obesity as p-vale* < 0.001.

Investigate the relationship between taking food between meals and weight and BMI

As taking food between meals (CAEC) is a categorical variable, we will need to see the distribution of BMI across the categories.

```
ggplot(obesity_df) + geom_histogram(aes(x=Weight, y=after_stat(density),fill=CAEC),
↪  binwidth=10, position = "dodge")
ggplot(obesity_df) + geom_histogram(aes(x=BMI, y=after_stat(density),fill=CAEC),
↪  binwidth=5, position = "dodge")
```





Based on the above plots, we can see that those who take food between meals do have a tendency of higher Weight and BMI than those who doesn't. This might be an indication of relationship between taking food between meals and and obesity.

Lets create a crosstable between taking food between meals and obesity levels.

```
##
##              Insufficient_Weight Normal_Weight Obesity_Type_I Obesity_Type_II
##   Always                       2            35              6               2
##   Frequently                 117            81              6               1
##   no                           3            10              1               1
##   Sometimes                  145           156            338             293
##   sum                        267           282            351             297
##
##              Obesity_Type_III Overweight_Level_I Overweight_Level_II  sum
##   Always                    0                  5                   3   53
##   Frequently                1                 14                  16  236
##   no                        0                 21                   1   37
##   Sometimes               323                236                 270 1761
##   sum                     324                276                 290 2087
```

```
((351+297+324+276+290) - (1+1+21+1))/(351+297+324+276+290) * 100
```

```
## [1] 98.43953
```

The above table with marginal totals shows that the number of persons having obesity do take food between meals. 98% of the persons having obesity do take food between meals.

Lets perform a `chi-square` test to check for relationship will exist in the population.

```
chisq.test(obesity_df$CAEC, obesity_df$NObeyesdad)
```

```
##
##  Pearson's Chi-squared test
##
## data:  obesity_df$CAEC and obesity_df$NObeyesdad
## X-squared = 723.34, df = 18, p-value < 2.2e-16
```

Chi-sqaure test show that there is evidence to reject the *Null Hypothesis* and conclude that there is *some relationship between taking food between meals and obesity as p-vale < 0.001*.

Investigate the relationship between high calories consumption (FAVC) and weight and BMI

Lets create a crosstable between consumption of high calories and obesity levels.

```
##
##       Insufficient_Weight Normal_Weight Obesity_Type_I Obesity_Type_II
##   no                   50            78             11               7
##   yes                 217           204            340             290
##   sum                 267           282            351             297
##
##       Obesity_Type_III Overweight_Level_I Overweight_Level_II  sum
##   no                 1                 22                  74  243
##   yes              323                254                 216 1844
##   sum              324                276                 290 2087
```

```
(340+290+323+254+216)/(351+297+324+276+290) * 100
```

```
## [1] 92.52276
```

The above table with marginal totals shows that the persons having obesity do take food with high calories. **92.5%** of the persons having obesity take food with high calories.

Lets perform a `chi-square` test to check for relationship will exist in the population.

```
chisq.test(obesity_df$FAVC, obesity_df$NObeyesdad)
```

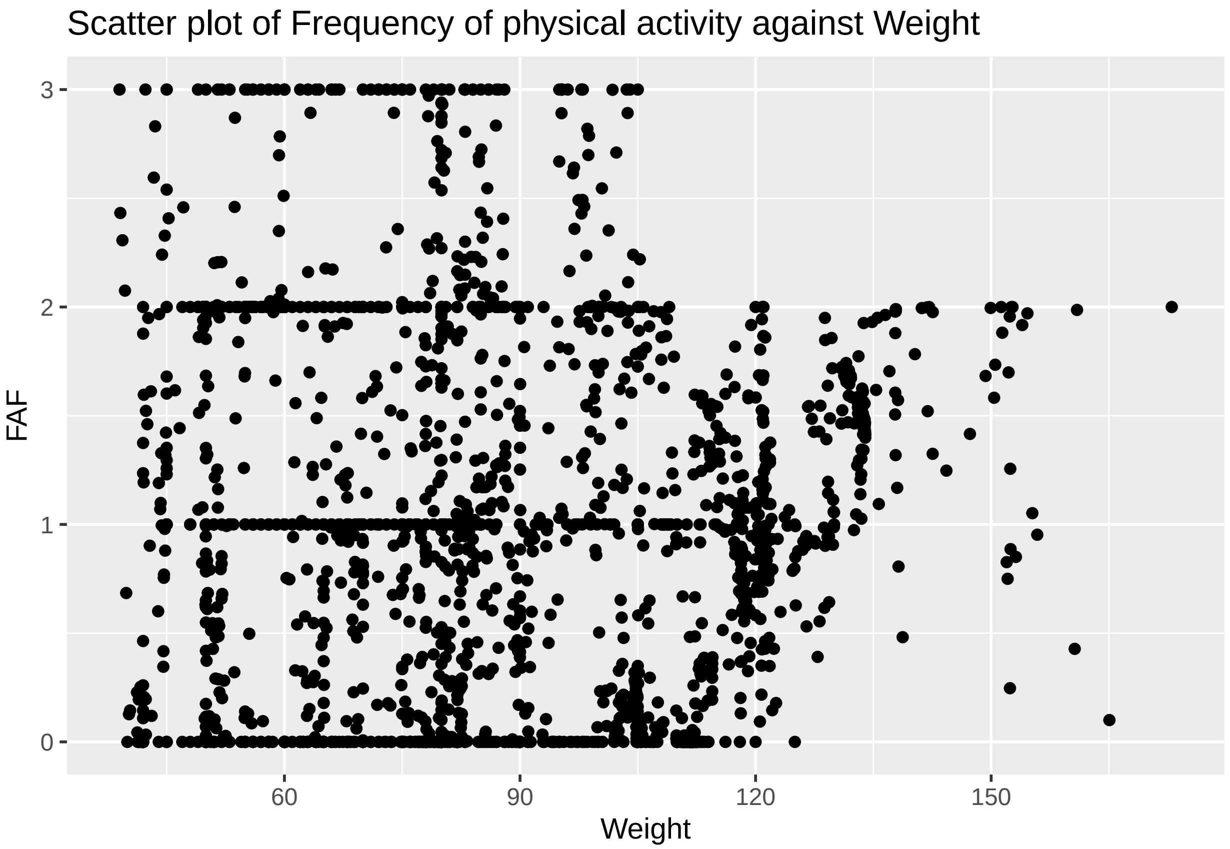
```
##
##  Pearson's Chi-squared test
##
## data:  obesity_df$FAVC and obesity_df$NObeyesdad
## X-squared = 231.28, df = 6, p-value < 2.2e-16
```

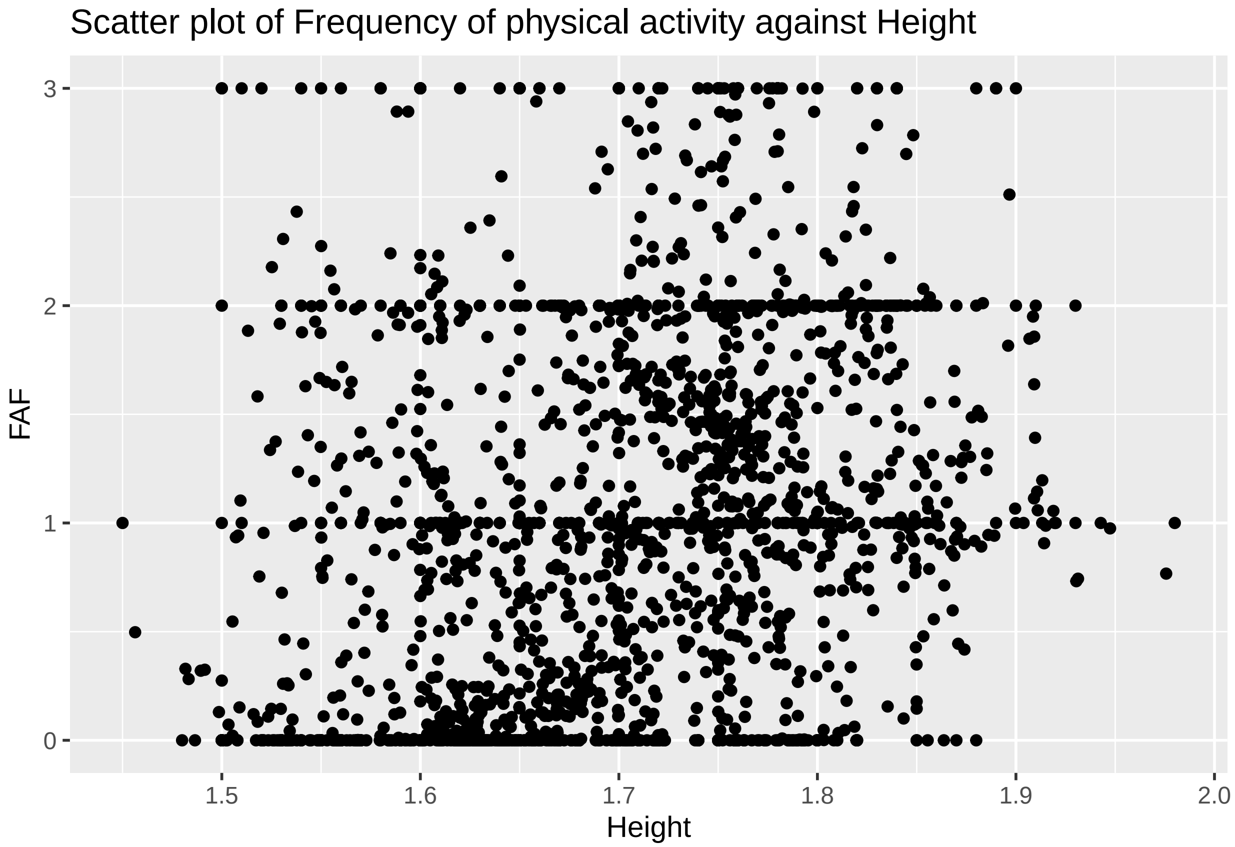
Chi-sqaure test show that there is evidence to reject the *Null Hypothesis* and conclude that there is some *relationship between taking foods with high calories and obesity as p-vale < 0.001.*

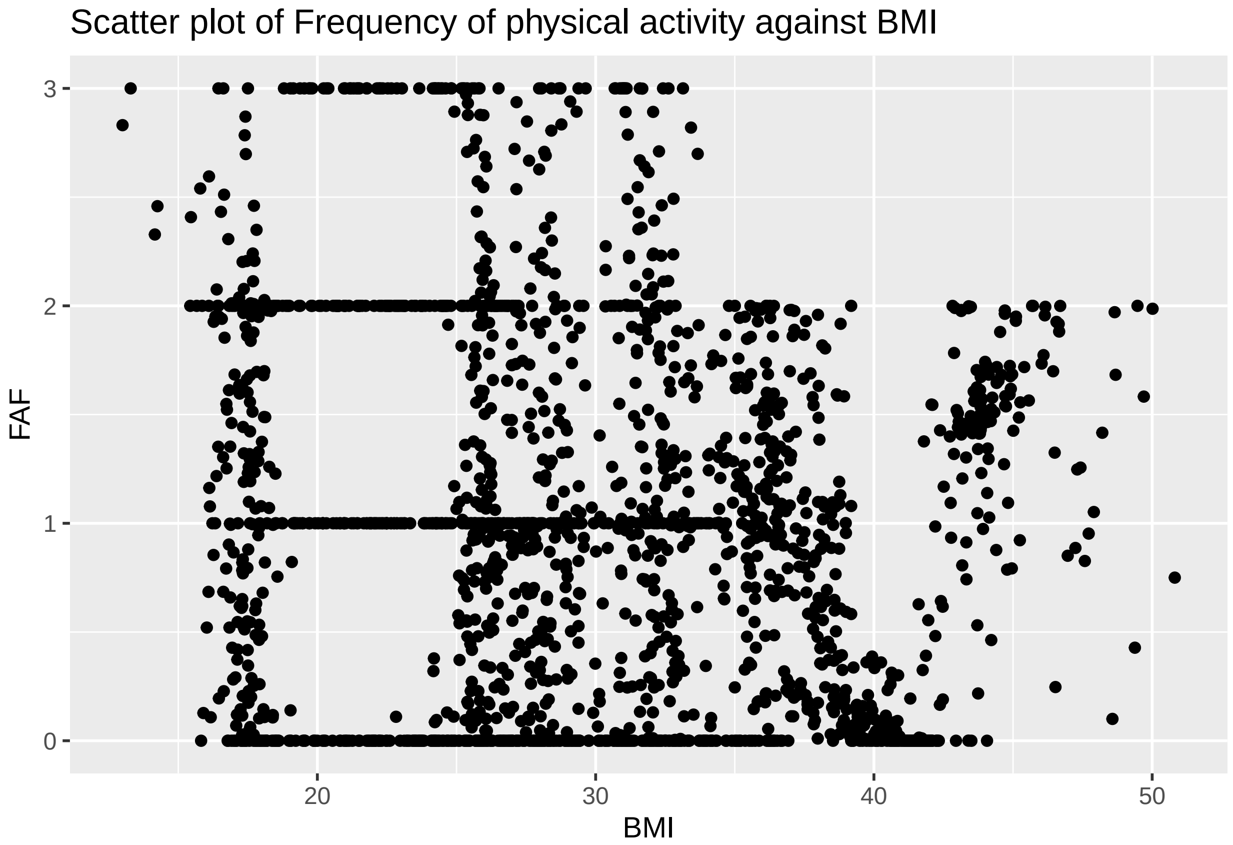
Investigate the relationship of physical activity with weight, height and BMI

Let plot some scatter plots to check the frequency of physical activity (FAF) against Weight, Height and BMI.

```
ggplot(obesity_df, aes(x=Weight, y=FAF)) + geom_point() + labs(title="Scatter plot of
↪  Frequency of physical activity against Weight",x="Weight", y="FAF")
ggplot(obesity_df, aes(x=Height, y=FAF)) + geom_point() + labs(title="Scatter plot of
↪  Frequency of physical activity against Height", x="Height", y="FAF")
ggplot(obesity_df, aes(x=BMI, y=FAF)) + geom_point() + labs(title="Scatter plot of
↪  Frequency of physical activity against BMI", x="BMI", y="FAF")
```


Scatter plot of Frequency of physical activity against Weight
FAF
3
2
1
0
60
90
120
150
Weight


Scatter plot of Frequency of physical activity against Height
FAF
0
1
2
3
1.5
1.6
1.7
1.8
1.9
2.0
Height



No apparent relationship is conclusive from the scatter plots. Lets perform a `t-test` to check any relationship between these variables in the population.

```
cor.test(obesity_df$Weight, obesity_df$FAF)
```

```
##
##  Pearson's product-moment correlation
##
## data:  obesity_df$Weight and obesity_df$FAF
## t = -2.5836, df = 2085, p-value = 0.009846
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.09915716 -0.01361564
## sample estimates:
##         cor
## -0.05649007
```

```
cor.test(obesity_df$Height, obesity_df$FAF)
```

```
##
##  Pearson's product-moment correlation
##
## data:  obesity_df$Height and obesity_df$FAF
## t = 14.024, df = 2085, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2538746 0.3323054
## sample estimates:
##      cor
## 0.293584
```

```r
cor.test(obesity_df$BMI, obesity_df$FAF)
```

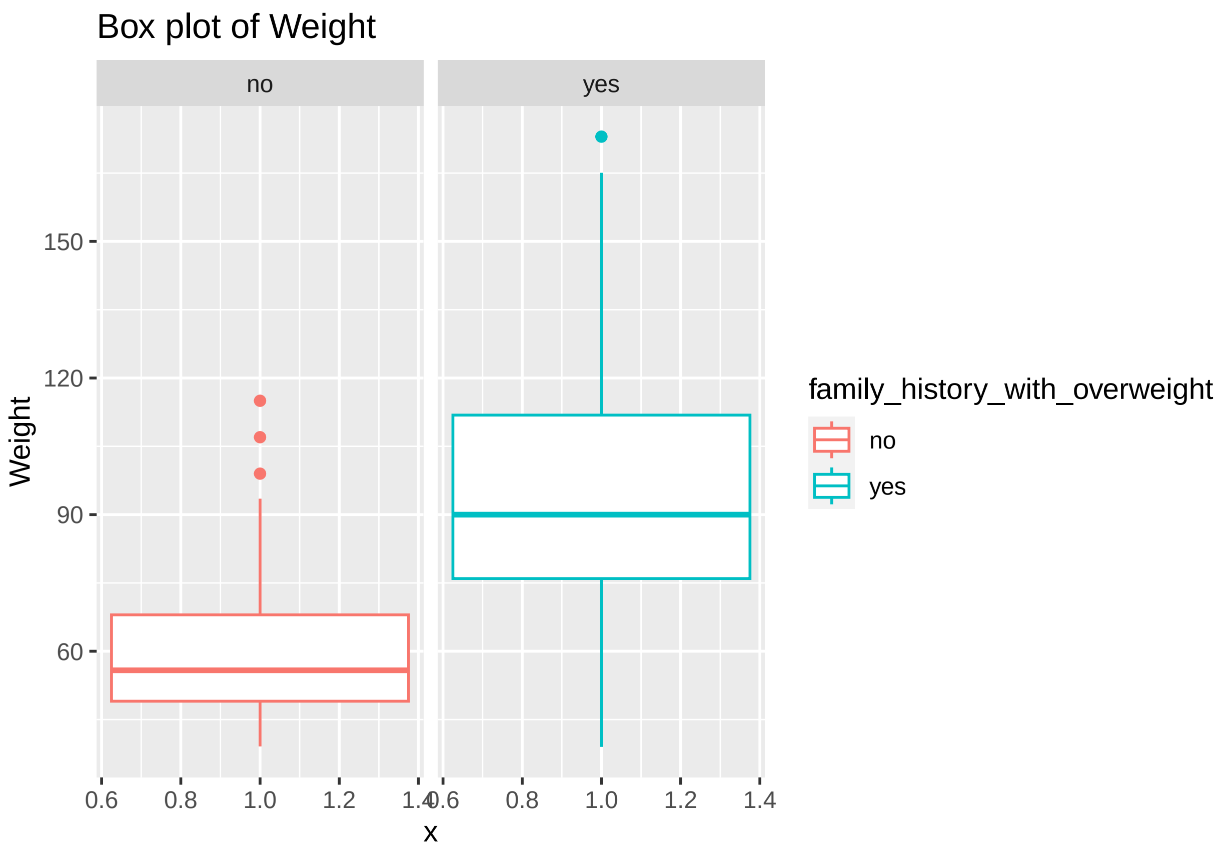
```
##
##  Pearson's product-moment correlation
##
## data:  obesity_df$BMI and obesity_df$FAF
## t = -8.4964, df = 2085, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2240807 -0.1411325
## sample estimates:
##        cor
## -0.1829321
```

From the above t-test, we see that there is a small *negative correlation* between BMI and *physical activity (FAF)*, which makes sense and the data also corroborates the same. Therefore as physical activity increases, BMI tends to decrease, hence *increased physical activity will help in reduction of obesity.*

Investigate the relationship between family history with overweight and Weight and BMI

Lets create a box-plot of Weight across family history with overweight.

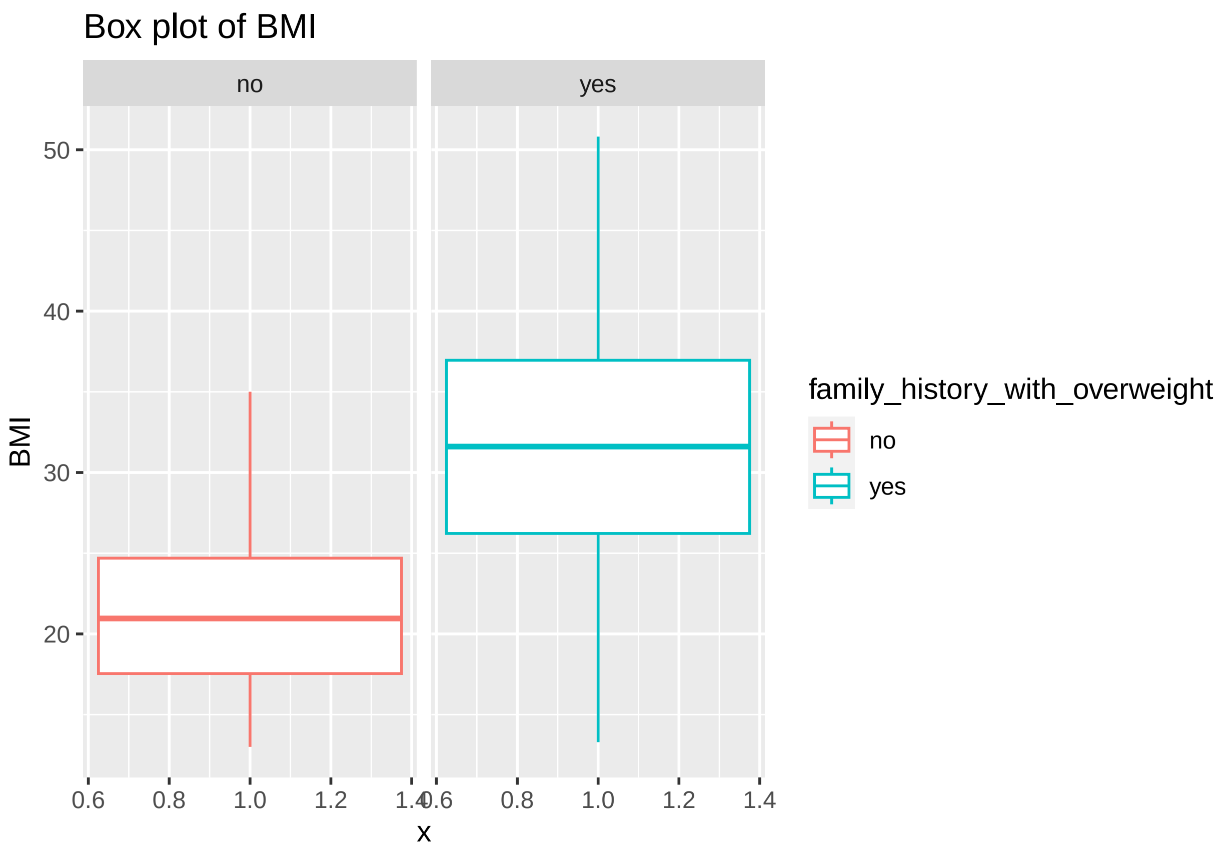
```r
ggplot(obesity_df) +
  geom_boxplot(aes(y=Weight, x=1,color=family_history_with_overweight)) +
  labs(title="Box plot of Weight",y="Weight") +
  ↪  facet_wrap(~family_history_with_overweight)
```



From above box-plots, we clearly see that the Weight of the persons having family history with overweight is much higher than than those *not* having family history with overweight.

Lets create a box-plot of BMI across family history with overweight.

```
ggplot(obesity_df) +
  geom_boxplot(aes(y=BMI, x=1,color=family_history_with_overweight)) +
  labs(title="Box plot of BMI",y="BMI") + facet_wrap(~family_history_with_overweight)
```



From above box-plots we clearly see that the BMI of persons having family history with overweight is much higher than than those *not* having family history with overweight.

Based on the above exploratory data analysis of the obesity dataset, we have observed that there is relationship of physical activity and eating habits with BMI and obesity. Therefore we can fit a model to *predict BMI and Obesity using physical activity and eating habits.*

Linear regression model to predict BMI using eating habits and physical activity.

Lets create a copy of the obesity data and drop the obesity level column as this model will predict BMI.

```
lin_reg_df <- obesity_df %>% select(-NObeyesdad)
```

Lets drop Weight and Height as we will use only eating habits and physical activity variables as predictors.

```
lin_reg_df <- lin_reg_df %>% select(-c(Weight,Height))
```

Fit a linear regression model using all columns as predictors.

```
linear.model.1 <- lm(BMI ~ ., lin_reg_df)
```

Summary of the model

```
summary(linear.model.1)
```

```
##
## Call:
## lm(formula = BMI ~ ., data = lin_reg_df)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -18.1388  -4.0151   0.2586   3.6705  23.3893 
## 
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                     2.80523    5.98235   0.469 0.639178    
## GenderMale                     -0.54837    0.27465  -1.997 0.045999 *  
## Age                             0.31191    0.02695  11.572  < 2e-16 ***
## family_history_with_overweightyes  6.85081    0.36927  18.552  < 2e-16 ***
## FAVCyes                         2.04369    0.42089   4.856 1.29e-06 ***
## FCVC                            3.31506    0.24988  13.267  < 2e-16 ***
## NCP                             0.47239    0.16931   2.790 0.005317 ** 
## CAECFrequently                 -3.47449    0.87747  -3.960 7.76e-05 ***
## CAECno                          1.41717    1.24363   1.140 0.254609    
## CAECSometimes                   3.36572    0.81124   4.149 3.48e-05 ***
## SMOKEyes                       -0.44747    0.88755  -0.504 0.614197    
## CH2O                            0.64430    0.21785   2.957 0.003137 ** 
## SCCyes                         -2.09413    0.62792  -3.335 0.000868 ***
## FAF                            -0.78714    0.15845  -4.968 7.33e-07 ***
## TUE                            -0.47305    0.22261  -2.125 0.033707 *  
## CALCFrequently                 -3.60899    5.82739  -0.619 0.535777    
## CALCno                         -4.78089    5.79242  -0.825 0.409257    
## CALCSometimes                  -2.40072    5.79602  -0.414 0.678769    
## MTRANSBike                      2.05284    2.19627   0.935 0.350054    
## MTRANSMotorbike                 4.27294    1.76496   2.421 0.015564 *  
## MTRANSPublic_Transportation     4.58904    0.39403  11.646  < 2e-16 ***
## MTRANSWalking                   1.60658    0.87333   1.840 0.065968 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 5.706 on 2065 degrees of freedom
## Multiple R-squared:  0.4995, Adjusted R-squared:  0.4944 
## F-statistic: 98.12 on 21 and 2065 DF,  p-value: < 2.2e-16
```

The R^2 value is `0.5` which shows that `50`% variation in BMI is accounted by the predictors. Therefore we can consider that the predictor variables are a good fit to predict BMI.

Next Steps.

- Create a Logistic Regression model to predict obesity.
- Explore for potential improvement areas of the linear regression model using less number of predictors with a lower AIC and better fit. Check for bias and outliers. Investigate if removal of outliers aid in better model fit.
- Check for linearity, homoscedasticity, and normality assumptions of the models.
- Investigate if factors other than eating habits and physical activity like ethnicity and geography also influence obesity using CDC Nutrition, Physical Activity, and Obesity dataset (CDC, 2023).
- Summarize the observations and findings.

Project Step 3

Logistic Regression model to predict obesity

As we are only predicting *risk of obesity* as a binary class (risk: yes or no), so lets create a column named Obesity and assign value `1` if the Obesity level is other than *Insufficient_Weight or Normal_Weight.*

```
obesity_df$Obesity <- ifelse(obesity_df$NObeyesdad == "Insufficient_Weight" |
↪  obesity_df$NObeyesdad == "Normal_Weight", 0, 1 )
```

```
obesity.model1 <- glm(Obesity ~  Gender + Age + family_history_with_overweight + FAVC +
                       FCVC + NCP + CAEC + SMOKE + CH2O + SCC + CALC+
                       FAF + TUE + MTRANS, data=obesity_df, family=binomial())
```

Investigate for most important predictors

```
data.frame(exp(obesity.model1$coefficients))
```

```
##                                   exp.obesity.model1.coefficients.
## (Intercept)                                           3.282257e-09
## GenderMale                                            1.178373e+00
## Age                                                   1.259164e+00
## family_history_with_overweightyes                     1.228893e+01
## FAVCyes                                               1.886560e+00
## FCVC                                                  1.025203e+00
## NCP                                                   7.096630e-01
## CAECFrequently                                        5.168796e-01
## CAECno                                                8.314644e+00
## CAECSometimes                                         1.428178e+01
## SMOKEyes                                              4.284892e-01
## CH2O                                                  1.355385e+00
## SCCyes                                                1.650523e+00
## CALCFrequently                                        6.222719e+04
## CALCno                                                2.285988e+04
## CALCSometimes                                         3.121228e+04
## FAF                                                   7.141416e-01
## TUE                                                   8.087231e-01
## MTRANSBike                                            1.242767e+00
## MTRANSMotorbike                                       1.889142e+00
## MTRANSPublic_Transportation                           4.613922e+00
## MTRANSWalking                                         9.304367e-01
```

From the above odds-ratio table we can clearly see that the below predictors have most influence on obesity.

1. **Family history with overweight**
2. **Consumption of alcohol**
3. **Consumption of food between meals**

Explore for potential improvement areas of the linear regression model using less number of predictors with a lower AIC and better fit. Check for bias and outliers. Investigate the model fit.

Fit another linear regression model using all columns as predictors except `Age`.

```
lin_reg_df <- lin_reg_df %>% select(-c(Age))
linear.model.2 <- lm(BMI ~ ., lin_reg_df)
```

Summary of the model

```
summary(linear.model.2)
```

```
##
## Call:
## lm(formula = BMI ~ ., data = lin_reg_df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.8744  -4.1553   0.6852   4.1352  22.5640
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        12.5840     6.1099   2.060 0.039561 *
## GenderMale                         -0.6345     0.2832  -2.240 0.025188 *
## family_history_with_overweightyes   7.5202     0.3763  19.987  < 2e-16 ***
## FAVCyes                             2.0534     0.4342   4.729 2.41e-06 ***
## FCVC                                3.4254     0.2576  13.297  < 2e-16 ***
## NCP                                 0.3011     0.1740   1.730 0.083726 .
## CAECFrequently                     -3.3527     0.9052  -3.704 0.000218 ***
## CAECno                              1.5003     1.2830   1.169 0.242379
## CAECSometimes                       3.6770     0.8365   4.396 1.16e-05 ***
## SMOKEyes                            0.7432     0.9095   0.817 0.413945
## CH2O                                0.5853     0.2247   2.605 0.009256 **
## SCCyes                             -2.7803     0.6449  -4.311 1.70e-05 ***
## FAF                                -1.0524     0.1617  -6.506 9.63e-11 ***
## TUE                                -1.1289     0.2221  -5.083 4.05e-07 ***
## CALCFrequently                     -2.8975     6.0116  -0.482 0.629865
## CALCno                             -4.7082     5.9758  -0.788 0.430858
## CALCSometimes                      -2.2806     5.9795  -0.381 0.702940
## MTRANSBike                          0.5420     2.2618   0.240 0.810654
## MTRANSMotorbike                     2.6527     1.8151   1.461 0.144038
## MTRANSPublic_Transportation         1.9025     0.3285   5.792 8.01e-09 ***
## MTRANSWalking                      -0.6911     0.8774  -0.788 0.430947
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.887 on 2066 degrees of freedom
## Multiple R-squared:  0.467,  Adjusted R-squared:  0.4618
## F-statistic: 90.51 on 20 and 2066 DF,  p-value: < 2.2e-16
```

Compare AIC between two models

```r
library(AICcmodavg)

#define list of models
models <- list(linear.model.1, linear.model.2)

#specify model names
mod.names <- c('linear.reg.model.1', 'linear.reg.model.1')

#calculate AIC of each model
aictab(cand.set = models, modnames = mod.names)
```

```
##
## Model selection based on AICc:
##
##                      K     AICc Delta_AICc AICcWt Cum.Wt       LL
## linear.reg.model.1 23 13216.39       0.00      1       1 -6584.93
## linear.reg.model.1 22 13345.47     129.08      0       1 -6650.49
```

Based on AIC we see that AIC of Model 2 is higher, so Model 1 is better fit

Check Bias and Outliers

```
#Store Residuals and Cook's distance as columns of a dataframe
df.model.1 <- data.frame(BMI=lin_reg_df$BMI)
df.model.1$residuals <- resid(linear.model.1)
df.model.1$cooks_distance <- cooks.distance(linear.model.1)
df.model.1$standardized_residuals <- rstandard(linear.model.1)
```

Check if any standardized residuals less than -2 or greater than > 2

```
df.model.1$large_residual <- df.model.1$standardized_residuals > 2 |
↪  df.model.1$standardized_residuals < -2
sum(df.model.1$large_residual)
```

```
## [1] NA
```

Check for data points with Cook's distance greater than 1

```
# cook's distance
head(df.model.1[df.model.1$cooks_distance > 1,])
```

```
##    BMI residuals cooks_distance standardized_residuals large_residual
## NA  NA        NA             NA                     NA             NA
```

As there are *no standardized residuals less than -2 or standardized residuals greater than 2*, we can say that there is no substantial outlier problem in the model. Furthermore there are no instances with *Cook's distance greater 1*, so we can say that there is no substantial influential cases for the model. *Therefore, the model seem to have no issues with bias or outliers.*

Based on the above model analysis using AIC, Standardized residuals, and Cook's distance we notice that Model 1 is a good fit and we do not improve the fit by reducing the number of predictors.

Check for Linearity and Normality of both the linear and logistic regression models

Linear Model Lets use VIF to check for multicollinearity

```
vif(linear.model.1)
```

```
##                                    GVIF Df GVIF^(1/(2*Df))
## Gender                         1.208610  1        1.099368
## Age                            1.887849  1        1.373990
## family_history_with_overweight 1.261212  1        1.123037
## FAVC                           1.168101  1        1.080787
## FCVC                           1.143758  1        1.069466
## NCP                            1.073583  1        1.036138
## CAEC                           1.301294  3        1.044871
## SMOKE                          1.042017  1        1.020793
## CH2O                           1.124989  1        1.060655
## SCC                            1.108976  1        1.053079
## FAF                            1.171563  1        1.082387
## TUE                            1.174189  1        1.083600
## CALC                           1.202349  3        1.031189
## MTRANS                         1.887079  4        1.082614
```

```
1/vif(linear.model.1)
```

```
##                                     GVIF        Df GVIF^(1/(2*Df))
## Gender                         0.8273966 1.0000000       0.9096134
## Age                            0.5297034 1.0000000       0.7278072
## family_history_with_overweight 0.7928882 1.0000000       0.8904427
## FAVC                           0.8560901 1.0000000       0.9252514
## FCVC                           0.8743111 1.0000000       0.9350460
## NCP                            0.9314607 1.0000000       0.9651221
## CAEC                           0.7684658 0.3333333       0.9570562
## SMOKE                          0.9596769 1.0000000       0.9796310
## CH2O                           0.8888974 1.0000000       0.9428136
## SCC                            0.9017331 1.0000000       0.9495963
## FAF                            0.8535609 1.0000000       0.9238836
## TUE                            0.8516517 1.0000000       0.9228498
## CALC                           0.8317052 0.3333333       0.9697540
## MTRANS                         0.5299196 0.2500000       0.9236900
```

```
mean(vif(linear.model.1))
```

```
## [1] 1.285656
```

From the above results we see that *VIF is not greater than 10 and 1/VIF is not less than 0.2*. The *mean VIF is not substantially greater than 1*, so we can say there is *no evidence of any problem of multicollinearity in the linear regreesion model*.

Lets plot residuals to check for homoscedasticity, and normality

```
linear_obesity_df <- obesity_df
linear_obesity_df$residuals <- resid(linear.model.1)
linear_obesity_df$fitted <- linear.model.1$fitted.values
linear_obesity_df$standard_residuals <- rstandard(linear.model.1)
linear_obesity_df$cooks_distance <- cooks.distance(linear.model.1)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

We do notice a mild funneling shape, so there is a *slight deviation of assumption of homoscedasticity.*

```
ggplot(linear_obesity_df, aes(sample=residuals)) + stat_qq() + stat_qq_line(color="blue")
↪  + labs(title="QQ plot of residuals",x="Theoritical Values", y="Observed Values")
```

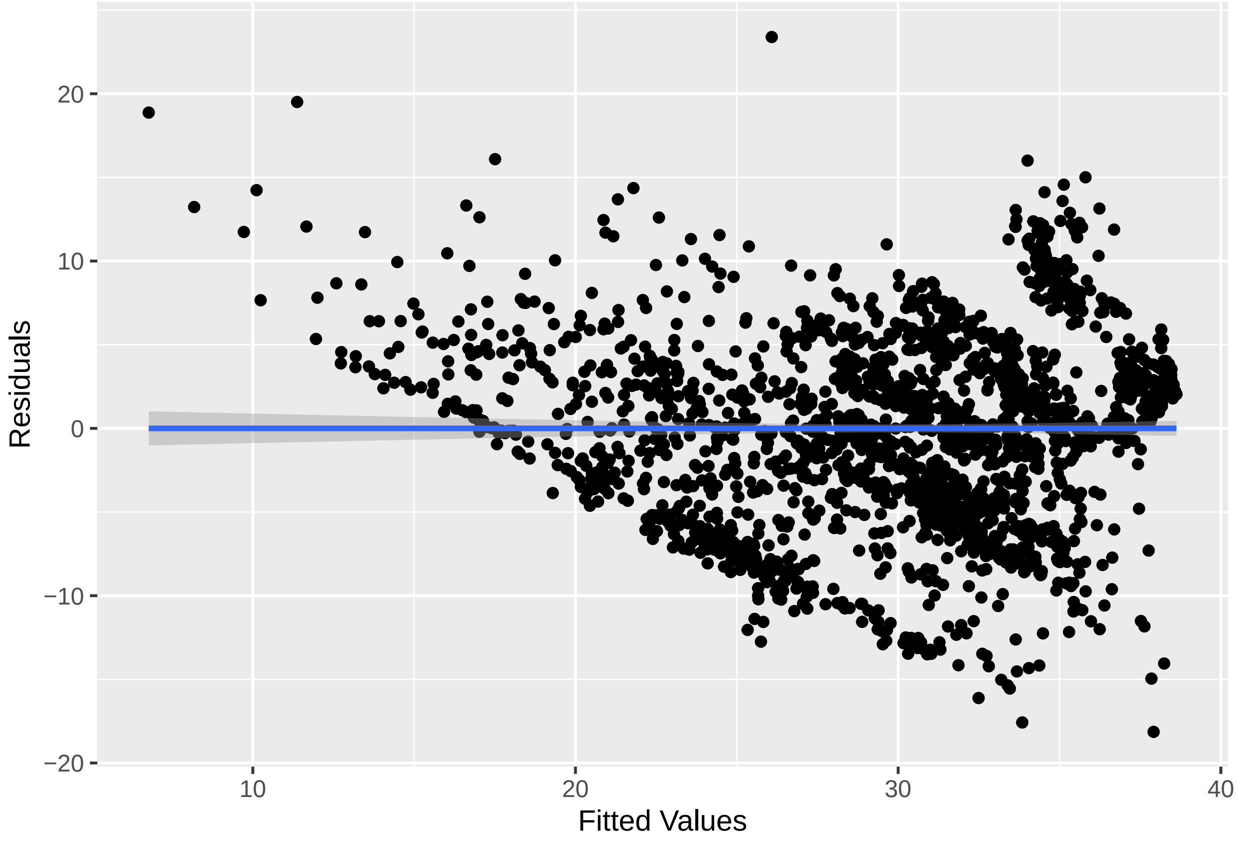
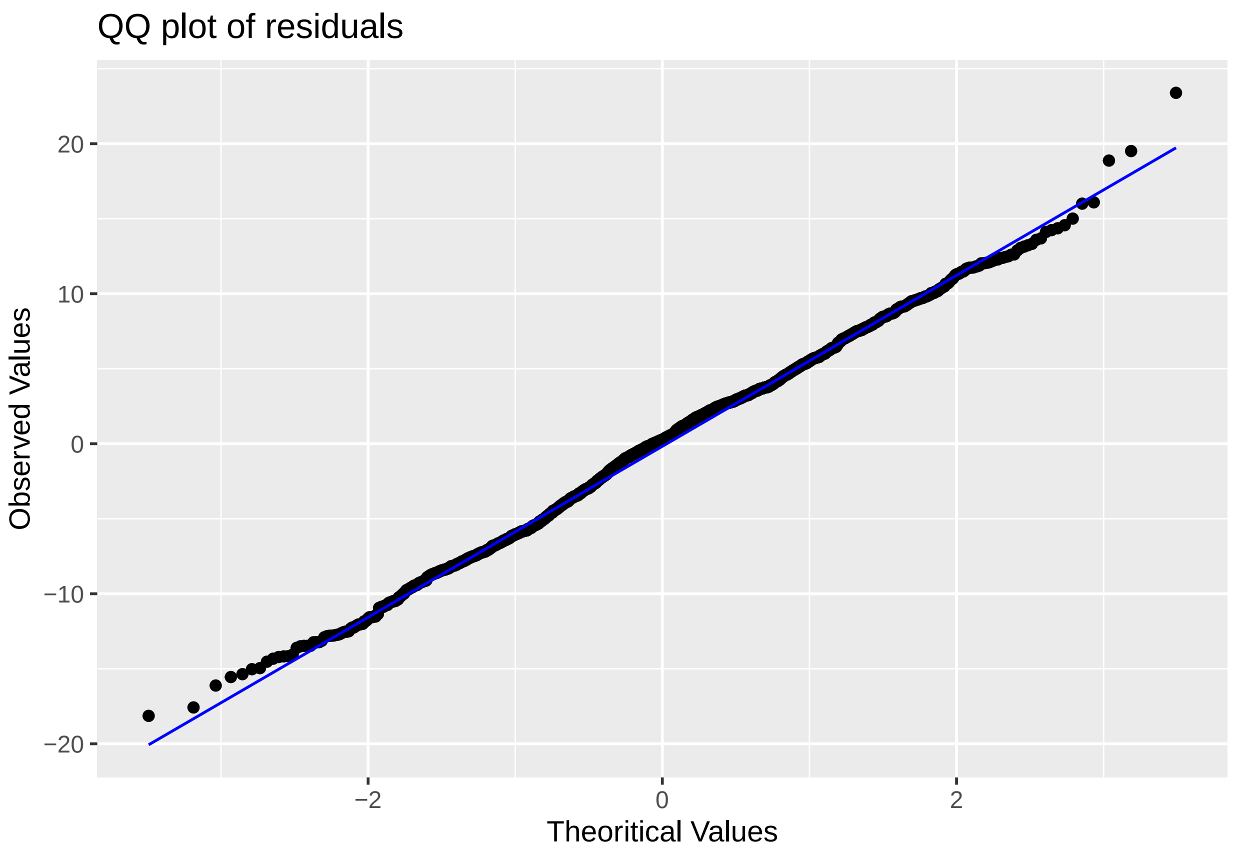


Figure 1: Scatter Plot of Residuals vs Fitted Values



The plot is very well aligned to the normal line (blue), so the assumption of normality is maintained.

Logistic Regression Model Lets use VIF to check for multicollinearity

```
vif(obesity.model1)
```

```
##                                    GVIF Df GVIF^(1/(2*Df))
## Gender                         1.271287  1        1.127514
## Age                            1.414664  1        1.189397
## family_history_with_overweight 1.168547  1        1.080993
## FAVC                           1.089774  1        1.043922
## FCVC                           1.183568  1        1.087919
## NCP                            1.141001  1        1.068176
## CAEC                           1.410102  3        1.058949
## SMOKE                          1.050571  1        1.024974
## CH2O                           1.194657  1        1.093004
## SCC                            1.122612  1        1.059534
## CALC                           1.204533  3        1.031501
## FAF                            1.204309  1        1.097410
## TUE                            1.132611  1        1.064242
## MTRANS                         1.663444  4        1.065678
```

```
1/vif(obesity.model1)
```

```
##                                     GVIF        Df GVIF^(1/(2*Df))
## Gender                         0.7866044 1.0000000       0.8869072
## Age                            0.7068814 1.0000000       0.8407624
## family_history_with_overweight 0.8557638 1.0000000       0.9250750
## FAVC                               0.9176214 1.0000000       0.9579256
## FCVC                               0.8449026 1.0000000       0.9191858
## NCP                                0.8764236 1.0000000       0.9361750
## CAEC                               0.7091685 0.3333333       0.9443324
## SMOKE                              0.9518636 1.0000000       0.9756350
## CH2O                               0.8370604 1.0000000       0.9149101
## SCC                                0.8907798 1.0000000       0.9438113
## CALC                               0.8301971 0.3333333       0.9694607
## FAF                                0.8303514 1.0000000       0.9112362
## TUE                                0.8829160 1.0000000       0.9396361
## MTRANS                             0.6011626 0.2500000       0.9383697
```

```
mean(vif(obesity.model1))
```

```
## [1] 1.270117
```

From the above results we see that *VIF is not greater than 10* and *1/VIF is not less than 0.2*. The *mean VIF is not substantially greater than 1*, so we can say there is no evidence of problem of multicollinearity in the logistic regression model.

Investigate if factors other than eating habits and physical activity like `ethnicity` and `geography` also influence obesity using CDC Nutrition, Physical Activity, and Obesity dataset (CDC, 2023).

Import and inspect first few rows

```
##   YearStart YearEnd LocationAbbr LocationDesc
## 1      2021    2021           AL      Alabama
## 2      2021    2021           AL      Alabama
## 3      2021    2021           AL      Alabama
## 4      2021    2021           AL      Alabama
## 5      2021    2021           AL      Alabama
## 6      2021    2021           AL      Alabama
##                                   Datasource                   Class
## 1 Behavioral Risk Factor Surveillance System Obesity / Weight Status
## 2 Behavioral Risk Factor Surveillance System   Fruits and Vegetables
## 3 Behavioral Risk Factor Surveillance System   Fruits and Vegetables
## 4 Behavioral Risk Factor Surveillance System Obesity / Weight Status
## 5 Behavioral Risk Factor Surveillance System       Physical Activity
## 6 Behavioral Risk Factor Surveillance System       Physical Activity
##                                Topic
## 1          Obesity / Weight Status
## 2 Fruits and Vegetables - Behavior
## 3 Fruits and Vegetables - Behavior
## 4          Obesity / Weight Status
## 5     Physical Activity - Behavior
## 6     Physical Activity - Behavior
##                                                                         Question
## 1 Percent of adults aged 18 years and older who have an overweight classification
## 2           Percent of adults who report consuming fruit less than one time daily
## 3      Percent of adults who report consuming vegetables less than one time daily
## 4                      Percent of adults aged 18 years and older who have obesity
## 5                 Percent of adults who engage in no leisure-time physical activity
## 6                 Percent of adults who engage in no leisure-time physical activity
##   Data_Value_Unit Data_Value_Type Data_Value Data_Value_Alt
## 1              NA           Value       28.5           28.5
## 2              NA           Value       44.8           44.8
## 3              NA           Value       24.7           24.7
## 4              NA           Value       25.4           25.4
## 5              NA           Value       19.3           19.3
## 6              NA           Value       23.3           23.3
##   Data_Value_Footnote_Symbol Data_Value_Footnote Low_Confidence_Limit
## 1                                                                21.6
## 2                                                                37.0
## 3                                                                18.3
## 4                                                                19.2
## 5                                                                14.4
## 6                                                                19.2
##   High_Confidence_Limit Sample_Size Total Age.years. Education Gender Income
## 1                  36.5         232          18 - 24
## 2                  52.9         223          18 - 24
## 3                  32.6         219          18 - 24
## 4                  32.8         232          18 - 24
## 5                  25.4         254          18 - 24
## 6                  28.0         475          25 - 34
##   Race.Ethnicity                             GeoLocation ClassID TopicID
## 1                (32.84057112200048, -86.63186076199969)     OWS    OWS1
## 2                (32.84057112200048, -86.63186076199969)      FV     FV1
## 3                (32.84057112200048, -86.63186076199969)      FV     FV1
## 4                (32.84057112200048, -86.63186076199969)     OWS    OWS1
## 5                (32.84057112200048, -86.63186076199969)      PA     PA1
## 6                (32.84057112200048, -86.63186076199969)      PA     PA1
##   QuestionID DataValueTypeID LocationID StratificationCategory1 Stratification1
## 1       Q037           VALUE          1             Age (years)         18 - 24
## 2       Q018           VALUE          1             Age (years)         18 - 24
## 3       Q019           VALUE          1             Age (years)         18 - 24
## 4       Q036           VALUE          1             Age (years)         18 - 24
## 5       Q047           VALUE          1             Age (years)         18 - 24
## 6       Q047           VALUE          1             Age (years)         25 - 34
##   StratificationCategoryId1 StratificationID1
## 1                     AGEYR         AGEYR1824
## 2                     AGEYR         AGEYR1824
## 3                     AGEYR         AGEYR1824
## 4                     AGEYR         AGEYR1824
## 5                     AGEYR         AGEYR1824
## 6                     AGEYR         AGEYR2534
```

Lets drop columns having NAs in *Data_Value* column

```r
cdc_obesity_df <- cdc_df %>% select(YearStart, LocationAbbr,LocationDesc,Gender,
↪  Data_Value) %>%drop_na()
```

Lets filter only rows having value `Percent of adults aged 18 years and older who have obesity` in *Question* column and *StratificationCategoryId1* column as `GEN`

```r
cdc_obesity_df_gen <- cdc_df %>% filter(Question == "Percent of adults aged 18 years and
↪  older who have obesity" & StratificationCategoryId1 == "GEN")
```

Lets sum of Data_Value column by each State/Location to calculate *total percentage of obesity for each Location by each Year.*

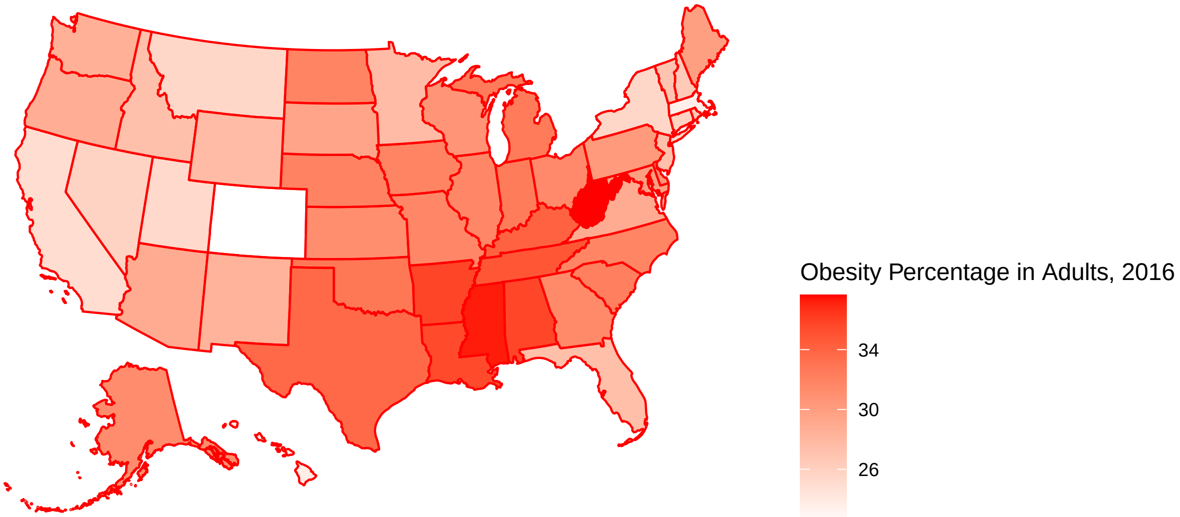
```
cdc_obesity_df_by_year <- cdc_obesity_df_gen %>% group_by(LocationAbbr, YearStart) %>%
↪  summarize(obesity_percentage=sum(Data_Value)/2) # sum is divided by two as we have
↪  summed up two genders and need the average
```

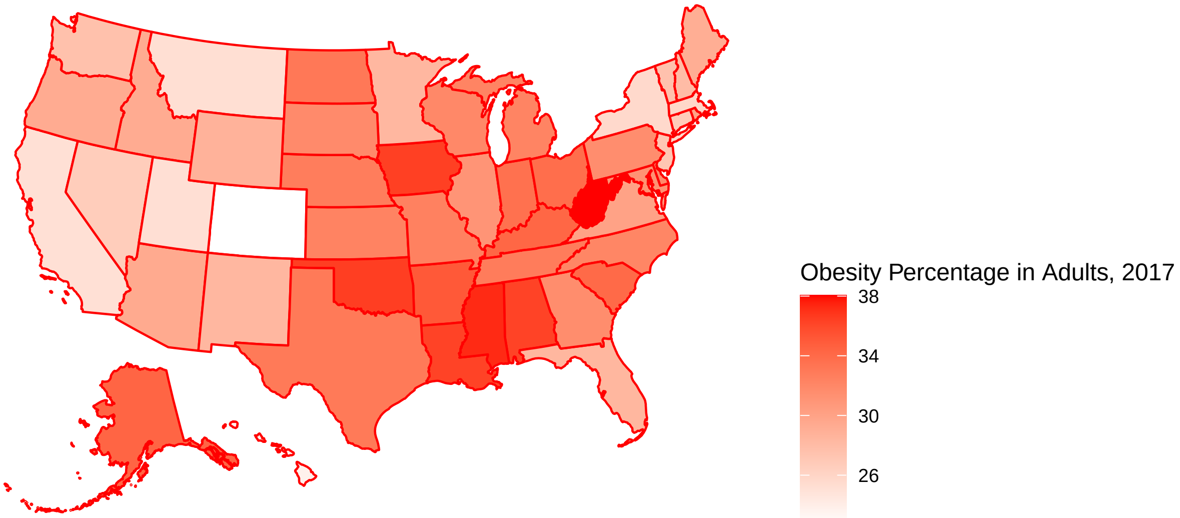
```
## `summarise()` has grouped output by 'LocationAbbr'. You can override using the
## `.groups` argument.
```

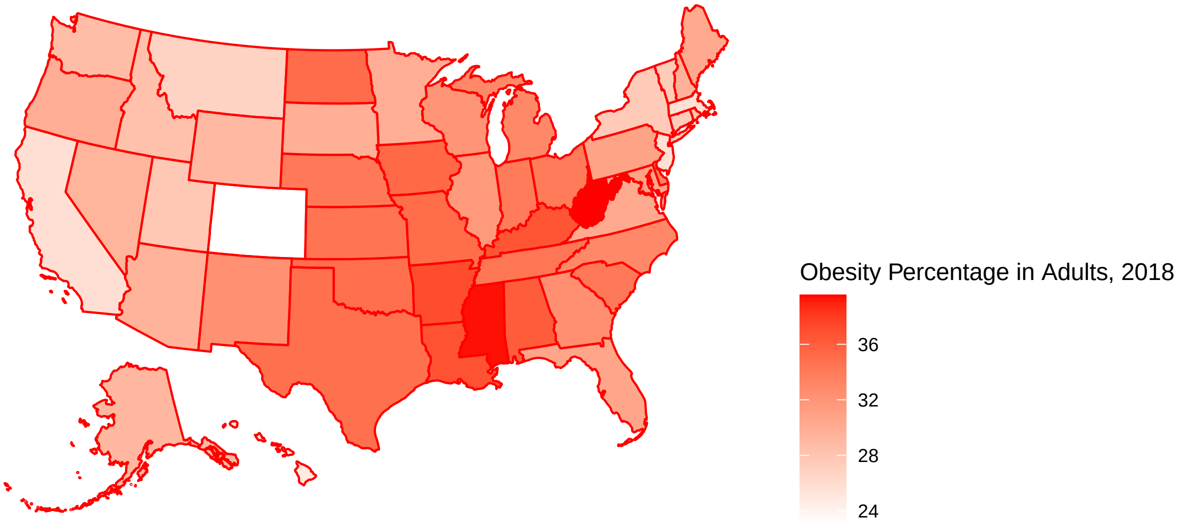
```
cdc_obesity_df_by_year$state <- cdc_obesity_df_by_year$LocationAbbr
```

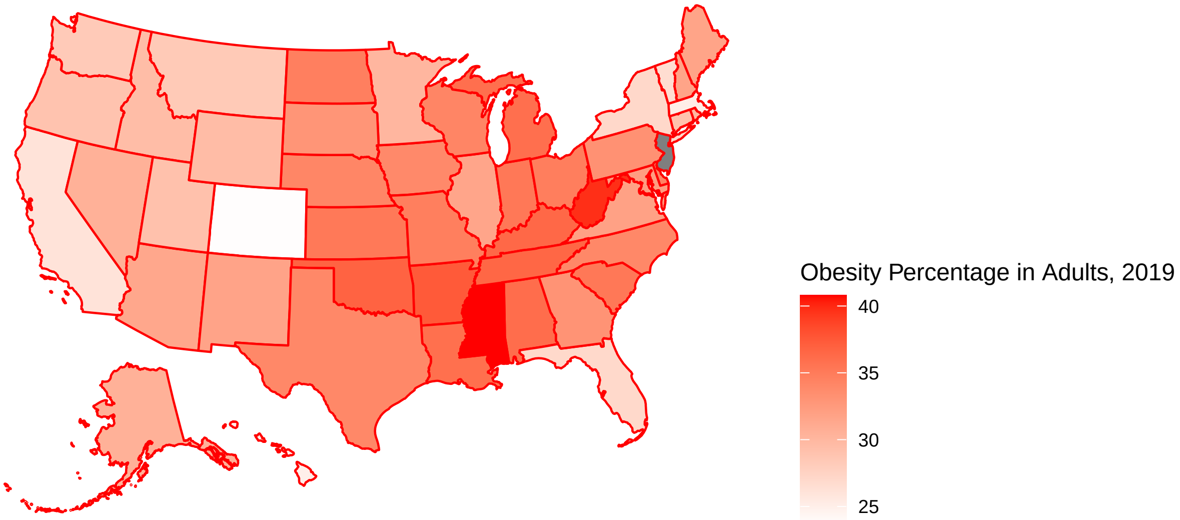
Lets plot 5 years Obesity percentage by states from 2016 to 2021.

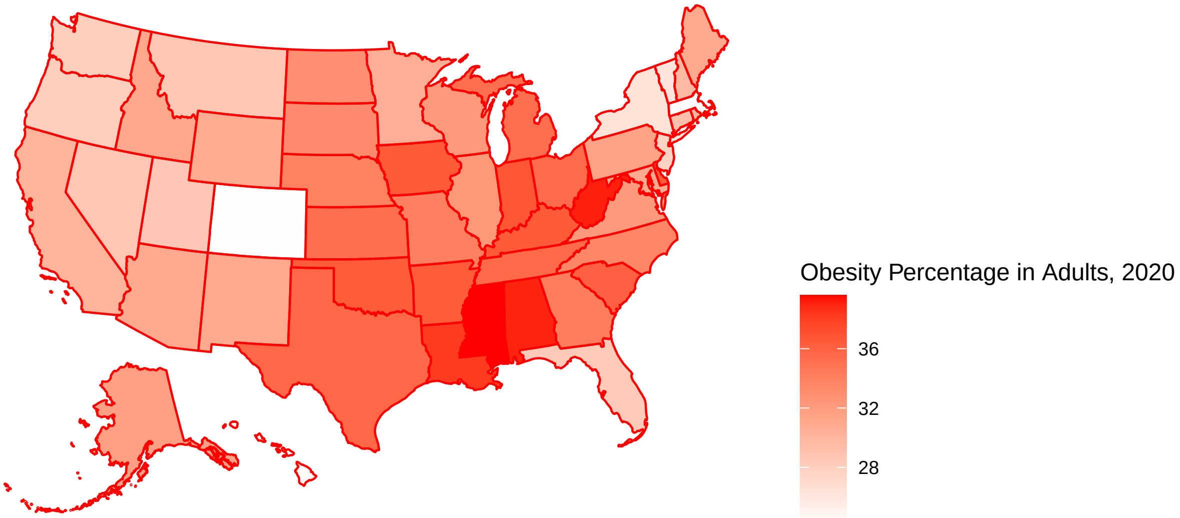
```
for (year in 2016:2021){

  cdc_obesity_by_year_plotting_df <-
↪  cdc_obesity_df_by_year[cdc_obesity_df_by_year$YearStart==year, ]

  us_map <- plot_usmap(data = cdc_obesity_by_year_plotting_df, values =
↪  "obesity_percentage", color = "red",
           labels=FALSE) +
  scale_fill_continuous(
    low = "white", high = "red", name = paste("Obesity Percentage in Adults,",year),
     ↪  label = scales::comma
  ) + theme(legend.position = "right")

  plot(us_map)

}
```

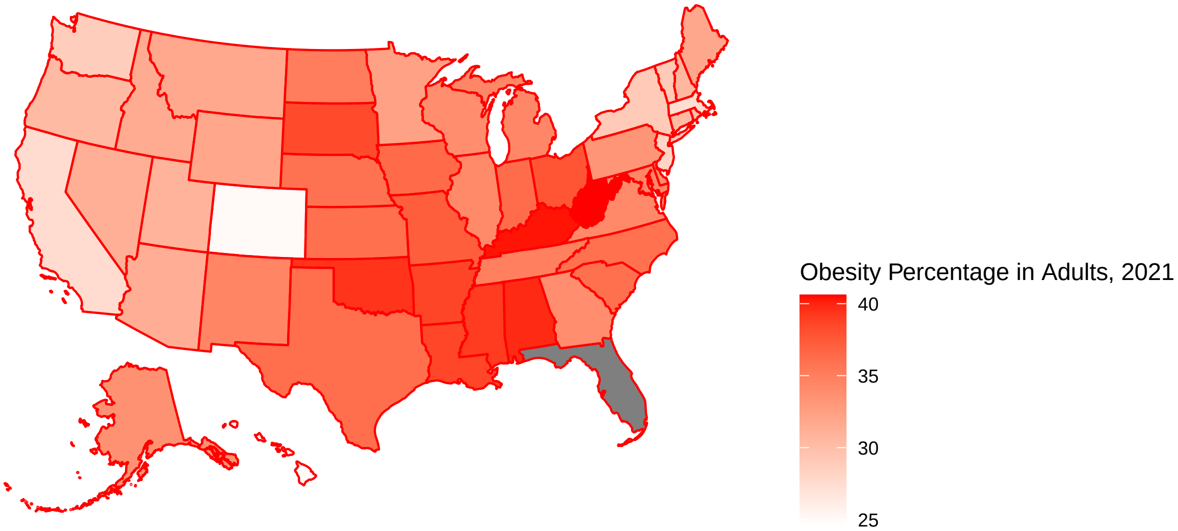

Obesity Percentage in Adults, 2016
34
30
26


Obesity Percentage in Adults, 2017
38
34
30
26


Obesity Percentage in Adults, 2018
36
32
28
24


Obesity Percentage in Adults, 2019
40
35
30
25


Obesity Percentage in Adults, 2020
36
32
28



The above maps clearly shows that the Obesity percentage is much higher in `South Eastern`, `Midwest` and `Southern` states as compared to `Western` states. Hence, we observe that the obesity vary by geography.

Lets filter only rows having value `Percent of adults aged 18 years and older who have obesity` in *Question* column and *StratificationCategoryId1* column as `RACE`

```
cdc_obesity_df_race <- cdc_df %>% filter(Question == "Percent of adults aged 18 years and
↪  older who have obesity" & StratificationCategoryId1 == "RACE") %>% select(YearStart,
↪  LocationAbbr, LocationDesc, Data_Value, Race.Ethnicity) %>% tidyr::drop_na()
```

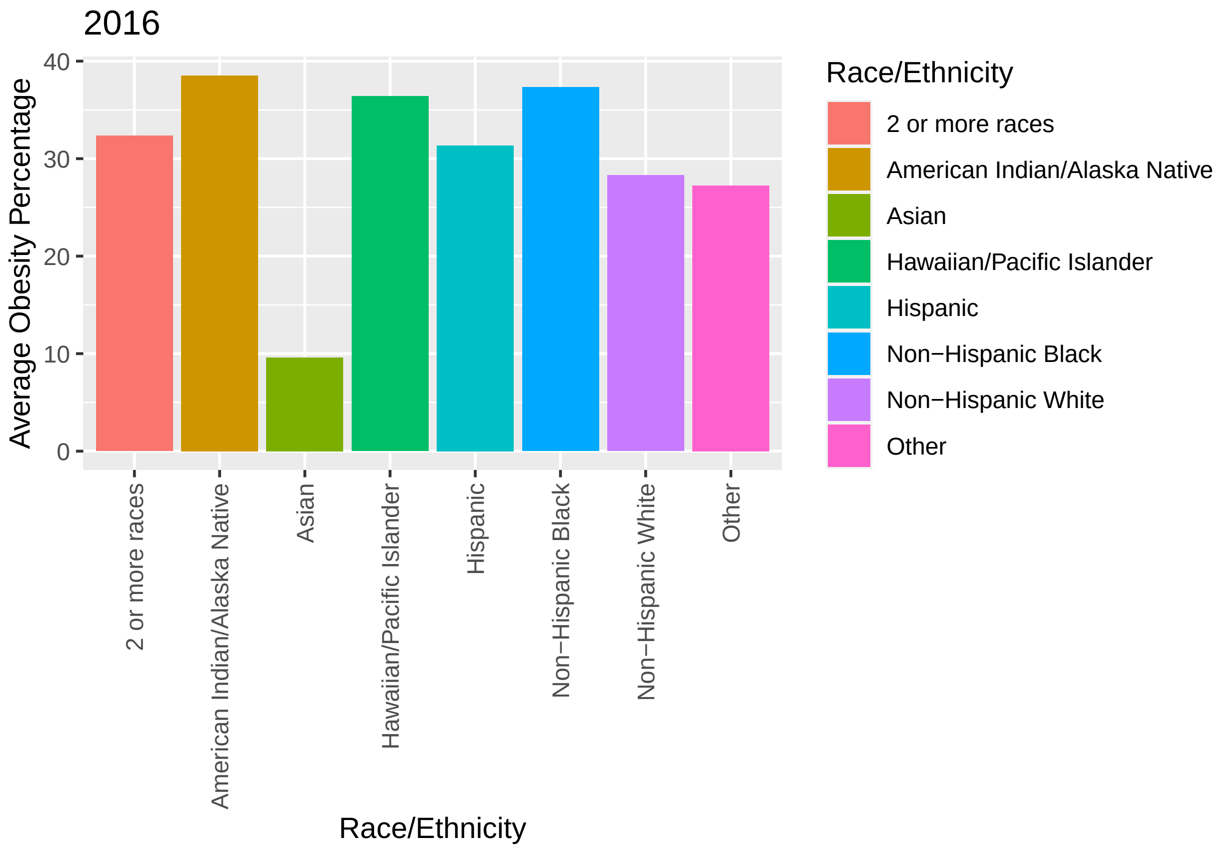
Lets calculate mean of Data_Value column by each Race/Ethnicity to get *average percentage of obesity* for each Race/Ethnicity across entire USA by each Year.

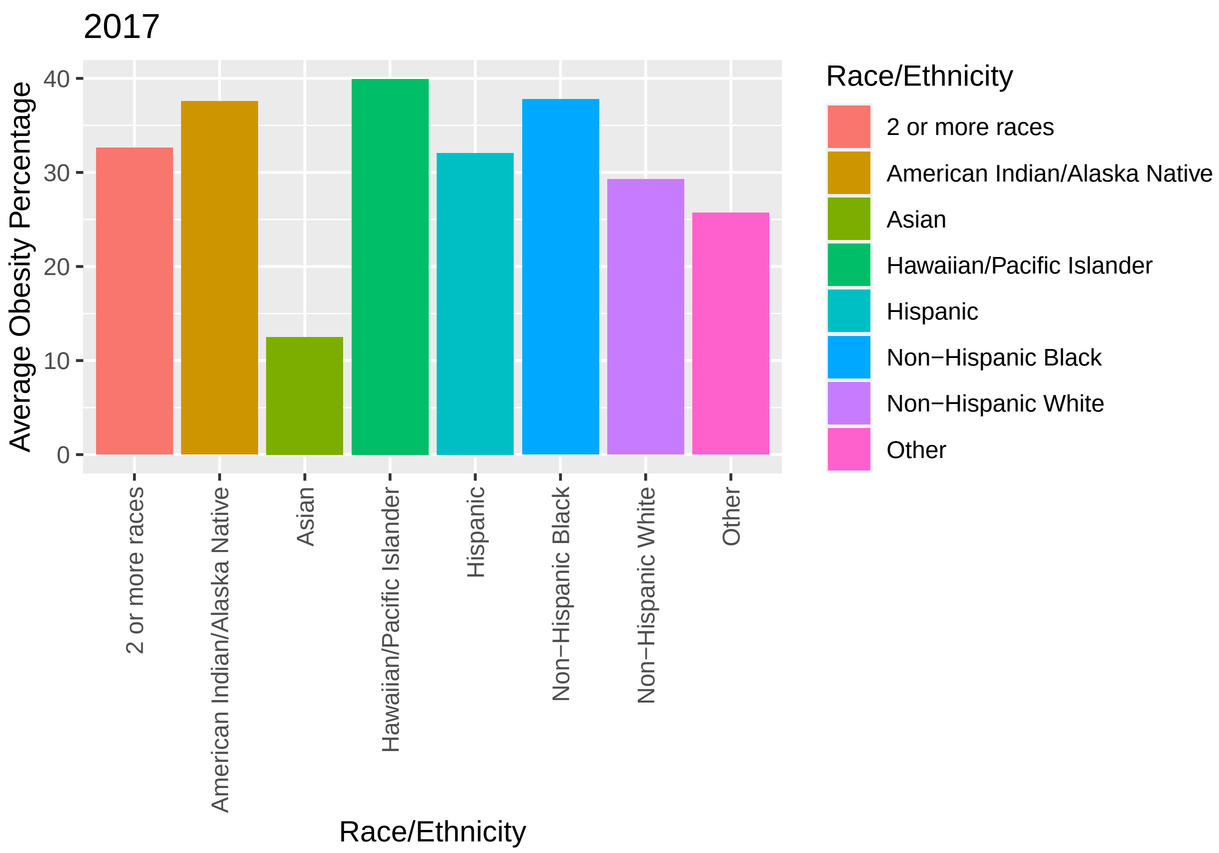
```
cdc_obesity_df_by_year_race <- cdc_obesity_df_race %>% group_by(Race.Ethnicity,
↪  YearStart) %>% summarize(obesity_percentage=mean(Data_Value))
```

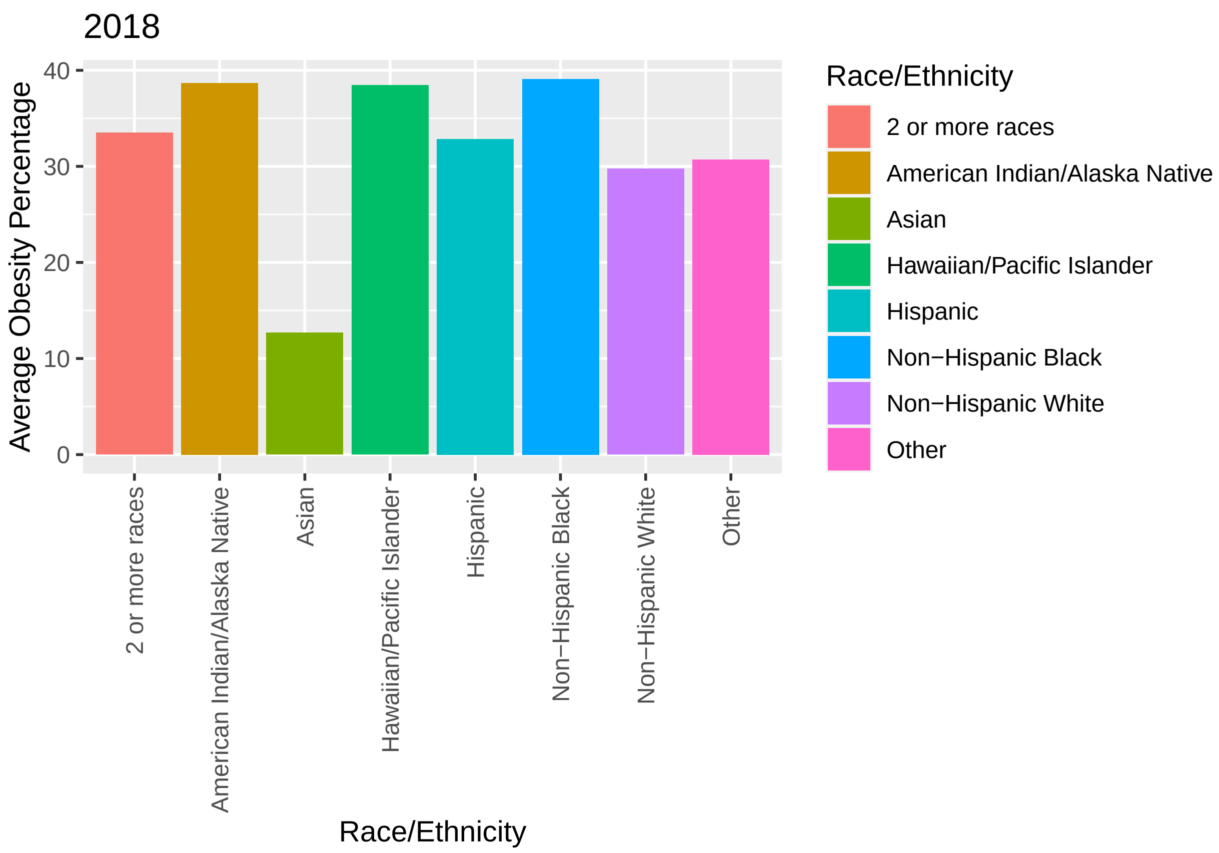
```
## `summarise()` has grouped output by 'Race.Ethnicity'. You can override using
## the `.groups` argument.
```

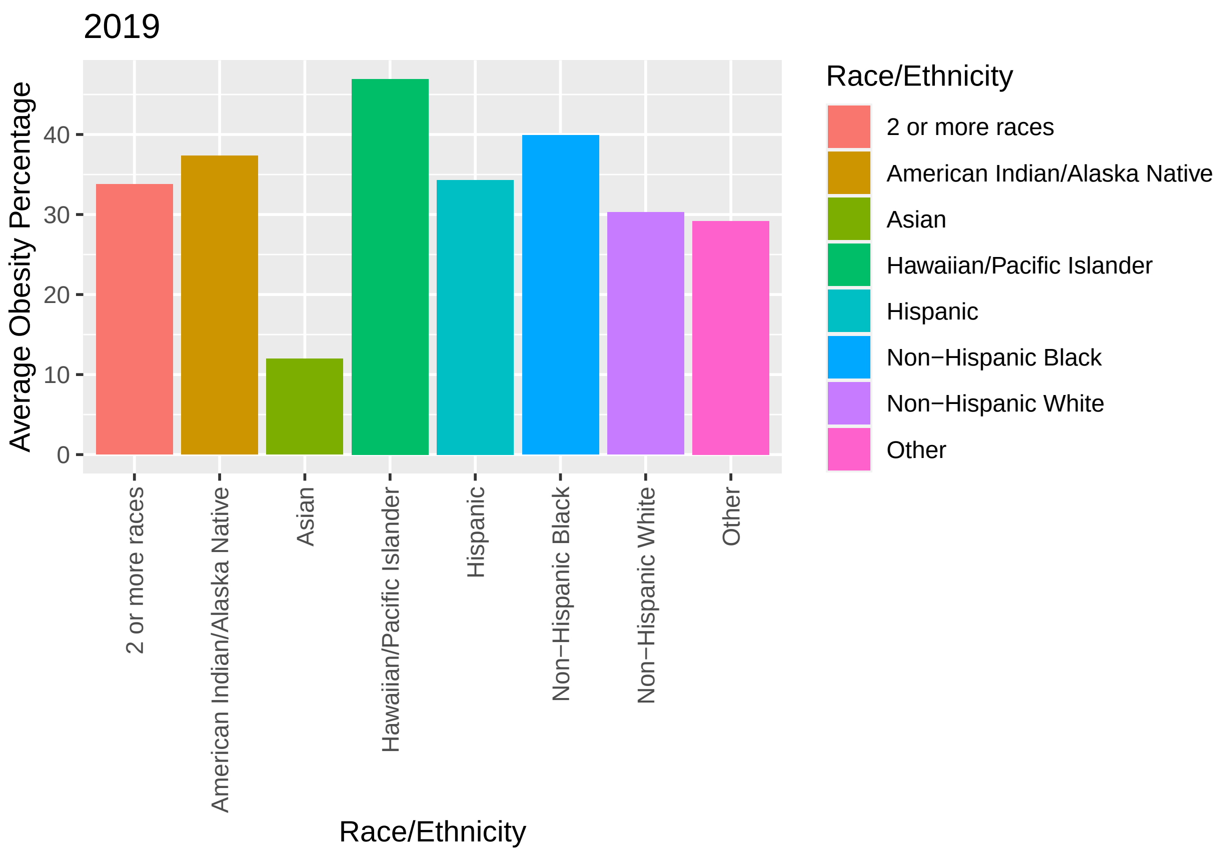
Lets plot 5 years Obesity percentage by Race from 2016 to 2021.

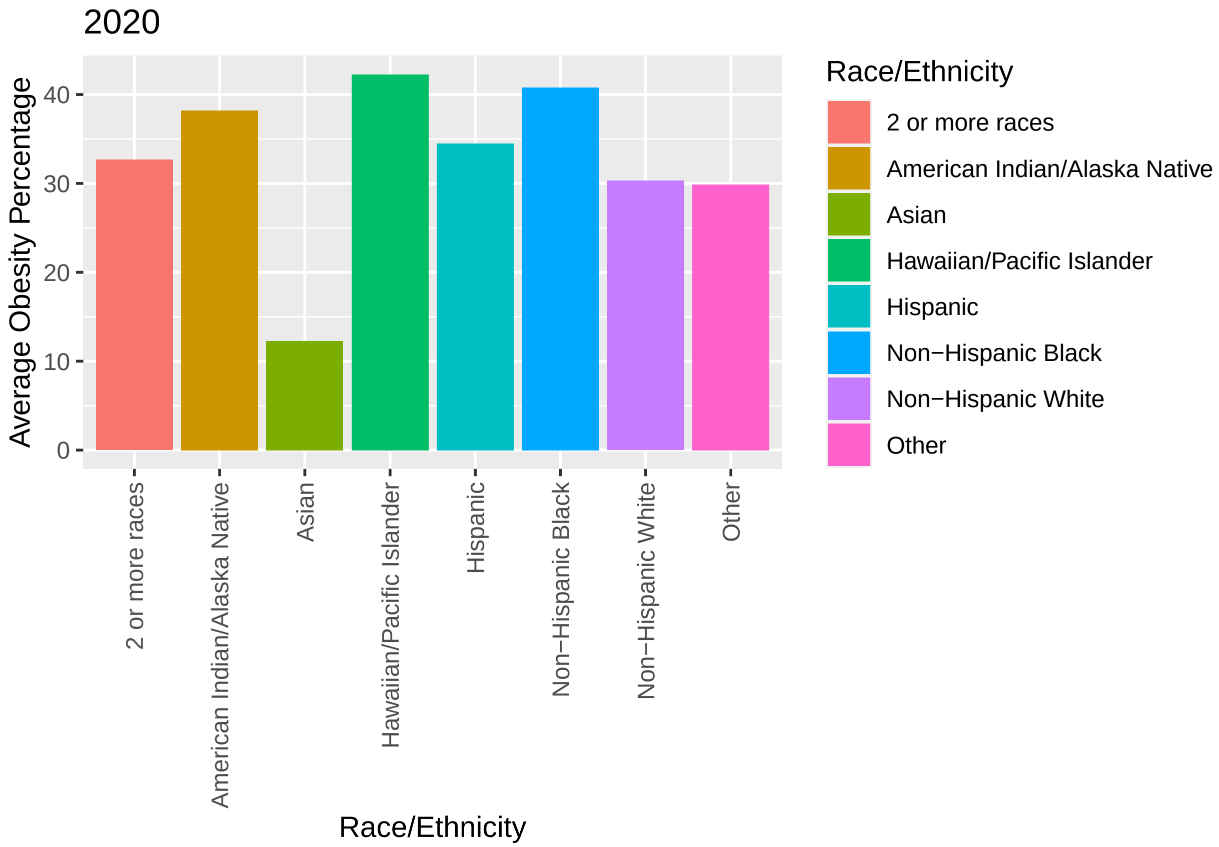
```
for (year in 2016:2021){
  cdc_obesity_df_year_race_plotting_df <-
↪  cdc_obesity_df_by_year_race[cdc_obesity_df_by_year_race$YearStart==year, ]

  bar <- ggplot(cdc_obesity_df_year_race_plotting_df, aes(x=
  Race.Ethnicity, y=obesity_percentage)) +geom_bar(stat="identity", aes(fill=
  Race.Ethnicity)) + theme(axis.text.x = element_text(angle=90, vjust=.5, hjust=1)) +
    labs(title=paste(year),x="Race/Ethnicity", y="Average Obesity Percentage",
     ↪  fill="Race/Ethnicity")
  plot(bar)
}
```

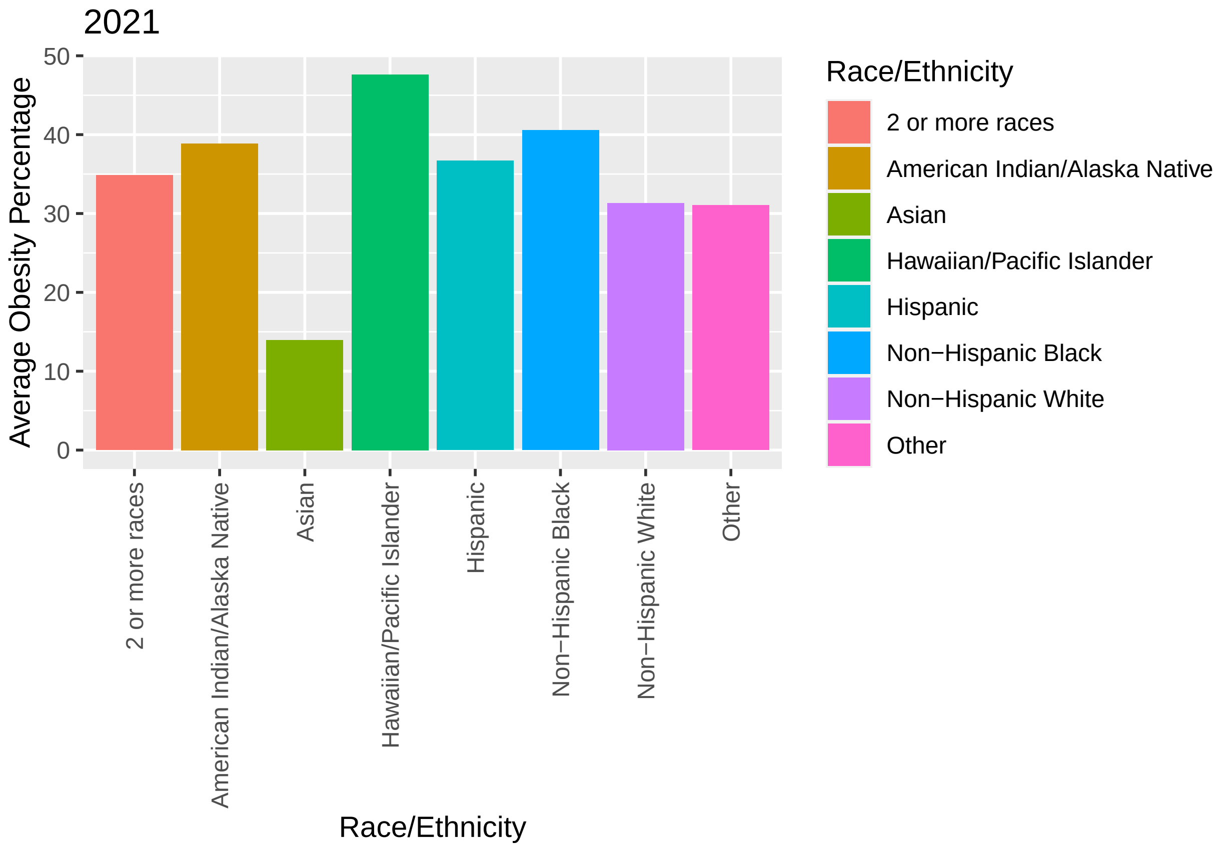



2017
Average Obesity Percentage
40
30
20
10
0
2 or more races
American Indian/Alaska Native
Asian
Hawaiian/Pacific Islander
Hispanic
Non−Hispanic Black
Non−Hispanic White
Other
Race/Ethnicity
Race/Ethnicity
2 or more races
American Indian/Alaska Native
Asian
Hawaiian/Pacific Islander
Hispanic
Non−Hispanic Black
Non−Hispanic White
Other


2018
Average Obesity Percentage
40
30
20
10
0
2 or more races
American Indian/Alaska Native
Asian
Hawaiian/Pacific Islander
Hispanic
Non−Hispanic Black
Non−Hispanic White
Other
Race/Ethnicity
Race/Ethnicity
2 or more races
American Indian/Alaska Native
Asian
Hawaiian/Pacific Islander
Hispanic
Non−Hispanic Black
Non−Hispanic White
Other


2019
Average Obesity Percentage
0
10
20
30
40
2 or more races
American Indian/Alaska Native
Asian
Hawaiian/Pacific Islander
Hispanic
Non−Hispanic Black
Non−Hispanic White
Other
Race/Ethnicity
Race/Ethnicity
2 or more races
American Indian/Alaska Native
Asian
Hawaiian/Pacific Islander
Hispanic
Non−Hispanic Black
Non−Hispanic White
Other


2020
Average Obesity Percentage
0
10
20
30
40
2 or more races
American Indian/Alaska Native
Asian
Hawaiian/Pacific Islander
Hispanic
Non−Hispanic Black
Non−Hispanic White
Other
Race/Ethnicity
Race/Ethnicity
2 or more races
American Indian/Alaska Native
Asian
Hawaiian/Pacific Islander
Hispanic
Non−Hispanic Black
Non−Hispanic White
Other



From the above plots, we notice that the obesity is higher among "Hawaiian/Pacific Islander", "Hispanic", "Non-Hispanic Black" and "American Indian/Alaska Native" as compared to "Asian" ethnicity/race. Hence, we observe that obesity vary by race and ethnicity.

Limitations

Even though we have found evidence to confirm the relationship of BMI and Obesity with eating habits and physical activity, we still cannot imply causation. We cannot say that obesity is caused by bad eating habits or lack of physical activity based on the above analysis. We must perform causal analyses like control testing and Pareto analysis to confirm causation.

In the above analysis, we have already identified that ethnicity and geography might influence obesity. The above study and predictive modeling we have performed do not consider other important factors that might also highly impact obesity. However, we still see that eating habits and physical activity has a significant relationship with BMI and obesity.

Conclusion

From the above research and data analysis, we have identified that BMI and Obesity have a relationship with eating habits and physical activity.

The correlation heatmap, correlation coefficients, and correlation t-tests show evidence of the relationship between BMI and eating habits like number of meals, consumption of vegetables, consumption of water, and frequency of physical activity. Similarly, boxplots, cross tables, and chi-squared tests show that there is evidence of the relationship between obesity and eating habits like high calorific food consumption, taking food between meals, alcohol consumption, and a family history of being overweight.

We also investigated that data model can be fitted to predict BMI and Obesity based on eating habits and physical activity. The above model analysis for bias, outliers, and linearity assumptions shows that the linear regression model to predict BMI and logistic regression to predict Obesity risk are feasible. Therefore, BMI and Obesity can be predicted based on eating habits and physical activity.

From the logistics regression model analysis, we see that a family history of being overweight, consumption of alcohol, and consumption of food between meals are the essential factors in predicting obesity risk, which do line up with general understanding. Hence, we can educate people about these leading factors, which can help to reduce the risk of obesity.

References

Body Mass Index. (n.d.). *Centers for Disease Control and Prevention.* **https://www.cdc.gov/healthyweight/assessing/bmi/index.html#:~:text=Body%20Mass%20Index%20(BMI), %20is,or%20health%20of%20an%20individual.**

Defining Adult Overweight & Obesity. (n.d.). *Centers for Disease Control and Prevention.* **https://www.cdc.gov/obesity/basics/adult-defining.html**

Estimation of obesity levels based on eating habits and physical condition. (2019). *UCI Machine Learning Repository.* **https://doi.org/10.24432/C5H31Z.**

Gozukara Bag, H.G., Yagin, F.H., Gormez, Y., González, P.P., Colak, C., Gülü, M., Badicu, G., Ardigò, L.P. 2023. Estimation of Obesity Levels through the Proposed Predictive Approach Based on Physical Activity and Nutritional Habits. *Diagnostics.* 13(18), 2949. **https://doi.org/10.3390/diagnostics13182949**

Nutrition, Physical Activity, and Obesity - Behavioral Risk Factor Surveillance System. (2023). *Centers for Disease Control and Prevention.* **https://data.cdc.gov/Nutrition-Physical-Activity-and-Obesity/ Nutrition-Physical-Activity-and-Obesity-Behavioral/hn4x-zwk7**

Palechor, F.M., de la Hoz Manotas, A. (2019). Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. *Data Brief.* 25, 104344. **https://doi.org/10.1016/j.dib.2019.104344**