# Estimation of Obesity Levels

Rhea Bhat, Yaashi Khatri, Hiba Ansari, Neha George, Natalie Nguyen, Andrea Flores, Ian Wood

## Introduction

Obesity is a disease characterized by an excessive accumulation of body fat that poses a risk to the individual's health. The tool for measuring whether an individual is overweight or obese is through body mass index (BMI), in which a BMI of 25-29.9 $kg/m^2$ indicates being overweight and a BMI > 30 indicates obesity (Fock & Khoo, 2013). There are a multitude of factors that contribute to the development of obesity. Among these factors include diet, physical conditions, and family history.

*Physical Lifestyle*

Physical conditions refer to exercise, sedentary vs. active lifestyle, and sleep quality. According to a study conducted by Hruby et al. (2016), higher levels of physical activity (defined as activity ≥ 30 min/day) resulted in less weight gained than low physical activity levels (defined as activity < min/day). Conversely, greater times spent engaged in sedentary activities (watching TV, using cellphone, playing video games, etc.) results in increased risk of obesity. The built environment of individuals may influence BMI as well. Higher density areas that are more walkable are likely to promote walking or bicycling as a means of transportation, which would increase average physical activity in these areas (Hruby et al., 2016).

Sleep duration and quality also impacts risk of weight gain. It was found that short duration sleep (time slept ≤ 6 hr per night) and poor sleep quality are linked to increased risk of obesity (Beccuti & Pannain, 2011).

*Diet*

An imbalanced diet is the main contributing factor to the onset of obesity. Individuals that subscribe to diets with higher intakes in saturated and trans fat, refined grains, red meats, sweets, and sweetened sugary-beverages are observed to have increased weight gain and risk of obesity. Conversely, individuals with diets high in whole grains, vegetables, fruit, yogurt, nuts, fish, and poultry were observed to have less weight gain and decreased risk of obesity (Hruby et al., 2016).

*Genetic Factors*

Family history also contributes to the development of obesity and consists of genetic factors and home environment. Genetics may impact dietary palate, food sensitivity, and metabolism of individuals. Variations in these genes could result in adverse eating choices (i.e., selective, limited, or excessive intake of different food groups). Moreover, Families of varying ethnic and cultural backgrounds are exposed to different types of diets that emphasize different food groups. Family environment also influences eating behaviors, which determines diet quality and quantity. The home environment determines if food is provided consistently, prepared at home, and if it is healthy and nutritious (Fruh, 2017).

*Substance use*

The use of substances such as drugs and alcohol have been studied as a potential risk factor to the development of obesity.

These four variables—diet, physical lifestyle, genetic factors, and substance use will be explored using exploratory analysis and machine learning to determine a prediction for our classification model. The hypotheses that will be tested in this project are as follows:

1. Lifestyle: It is claimed that a sedentary lifestyle results in a higher BMI that could put a human at risk for obesity.
2. It is claimed that a diet of high caloric food and poor nutrition results in a higher BMI that could put a human at risk for obesity.
3. It is claimed that having a family history of obesity results in a higher BMI and chance of being overweight.
4. It is claimed that substance use results in higher BMI and may increase risks of developing obesity.
5. It is claimed that weight, gender, and age can have a significant impact on weight classification

## Data

The data will be obtained from the data article by Palechor and Manotas (2019), which consists of text and table data acquired through survey. The set includes data for the estimation of obesity levels in individuals from the countries of Mexico, Peru and Colombia, based on their eating habits and physical condition. The data contains 17 attributes and 2111 records, the records are labeled with the class variable `NObesity` (Obesity Level), which allows classification of the data using the values of Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II and Obesity Type III.

The 17 attributes and their definitions are listed below:

| Variable Name | Variable Description | Survey Answers |
|---|---|---|
| **Gender** | The behavioral, cultural, and physical traits one typically associates with themselves | Female<br>Male |
| **Age** | NA | Integer |
| **Height** | NA | Numeric |
| **Weight** | NA | Numeric |
| **family_history_with_overweight** | Binary variable if there is a history of being overweight in the person's family | Yes<br>No |
| **FAVC** | Frequency consumption of high caloric food | Yes<br>No |
| **FCVC** | Frequency of consumption of vegetables | Sometimes<br>Frequently<br>Always<br>no |
| **NCP** | Number of main meals per day | 1 per day<br>2<br>3 |
| **CAEC** | Consumption of food between meals | Sometimes<br>Frequently<br>Always<br>no |
| **SMOKE** | Consumption of | Yes<br>No |
| **CH20** | Consumption of water daily | Sometimes<br>Frequently<br>Always |

| | | |
|---|---|---|
| | | no |
| **SCC** | Calories consumption monitoring | Yes<br>No |
| **FAF** | Physical activity frequency | Sometimes<br>Frequently<br>Always<br>No |
| **TUE** | Time using technology devices | Sometimes<br>Frequently<br>Always<br>No |
| **CALC** | Consumption of alcohol | Sometimes<br>Frequently<br>Always<br>No |
| **MTRANS** | Transportation used | Public-Transportation<br>Walking<br>Autobombile<br>Motorbike<br>Bike |
| **NObeyesdad** | Weight category | Insufficient_Weight<br>Normal_Weight<br>Overweight_Level_I<br>Overwieght_Level_II<br>Obesity_Type_I<br>Obesity_Type_II<br>Obesity_Type_III |

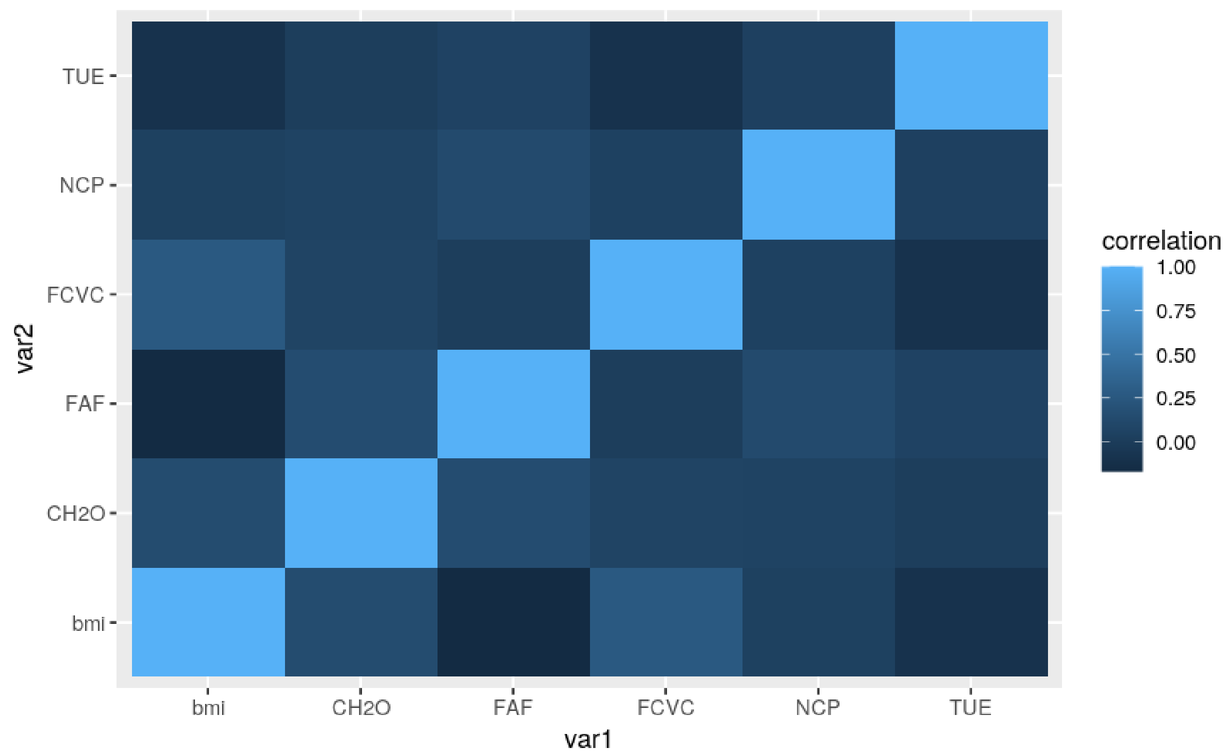
In order to utilize this data, it needs to be tidied and cleaned. In addition, we need to create new variables so that we can continue with our exploratory analysis and modeling. The first step is to save our dataset as a dataframe and then omit the NA values and overwrite our dataset. We introduced a new variable **id** in order to give each person a unique identifier and then we calculated a **bmi** column using the simple equation bmi= (Weight) / $(\text{Height})^2$ . We will be basing our hypotheses on this column. For some of our analysis, binary values will be needed in order to plot and compare. We took variables that had *yes* and *no* responses and converted them into binary numeric variables *0* and *1*. In addition, some variables such as **FAF** had responses that ranged on discrete values such as 0,1,2, and 3 but some responses had decimal answers. We rounded this data so it is easier to visualize in the exploratory visualizations.

## Exploratory Analysis

*Correlation Between Variables*

We constructed a correlation matrix between all the numeric variables and displayed the results as a correlation heatmap. From Figure 1, we can see while the correlation is not as high as we expected, we decided to pull out a few variables of interest for our hypotheses.

**Figure 1. Correlation Matrix**

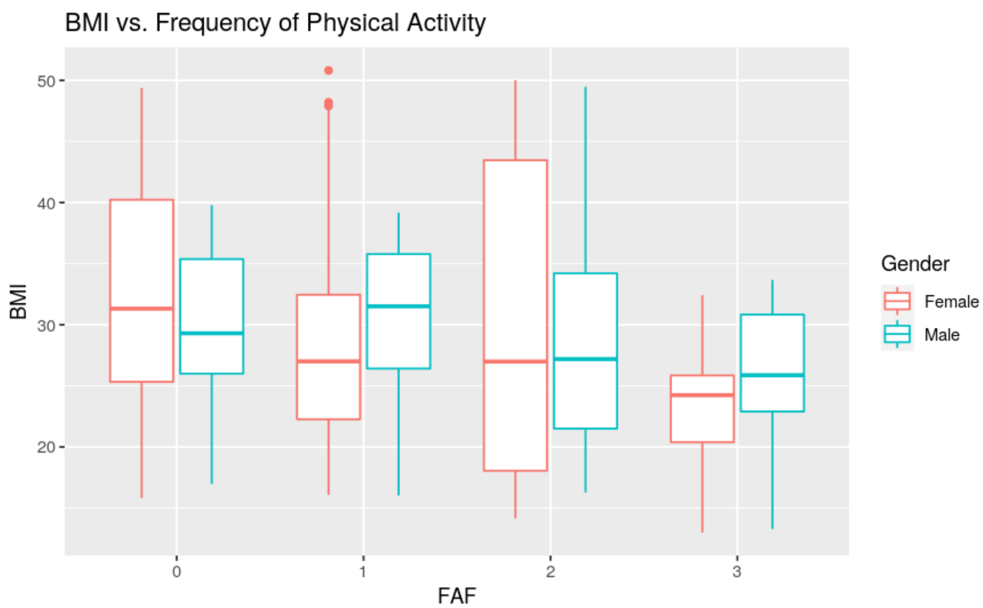



We constructed a correlation heatmap with the numeric variables in the dataset. While the correlations are not high, we used this heatmap to see the relationship between these variables and BMI. The highest correlation was between CH2O (consumption of water daily) and BMI as well as FCVC (frequency consumption of high caloric foods) and BMI. We will be visualizing multiple variables correlation with BMI but these are two noteworthy variables to keep in mind.

*Hypothesis 1: Lifestyle*

It is claimed that a sedentary lifestyle results in a higher BMI that could put a human at risk for obesity. The first step to this process was visualizing the data. We decided to split the data based on gender.

**Figure 2. Box plot: BMI vs. Frequency of Physical Activity**




From Figure 2, we can see that women have a wider range of BMIs than men and seem to be more sedentary. After visualizing the data, we took out the gender filter and decided to use k-means clustering on the data. The silhouette width was maximized at k=2 clusters so we ran the algorithm as follows.

**Figure 3. Silhouette Plot**

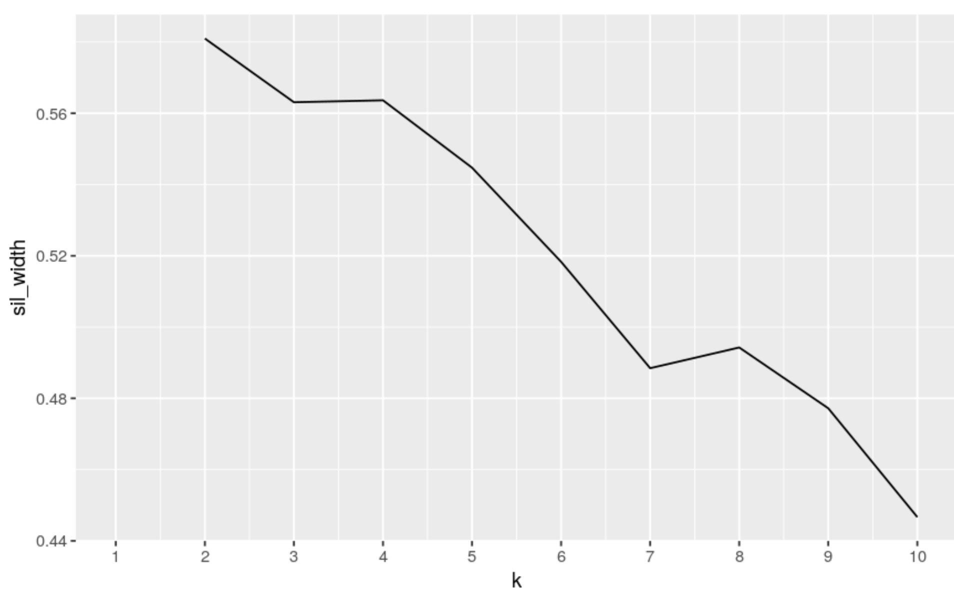
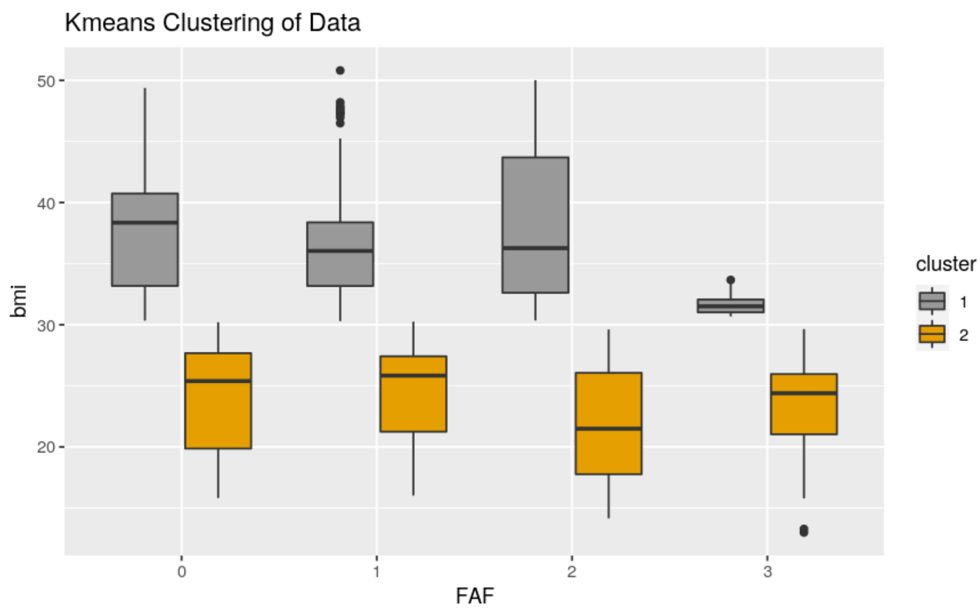


**Figure 4. K-means clusters: BMI vs. FAF**




From the clustering task we can see how the data was split based on the BMI classification. A healthy BMI for an adult is between 25.0 and 29.9. The data is split at a BMI of 30. Within both clusters, it can easily be observed how the higher median BMI tends to be for lower activity levels [FAF= 0 or 1] when compared to higher activity levels [FAF= 2 or 3].

*Hypothesis 2: Diet*

It is claimed that a diet of high caloric food and poor nutrition results in a higher BMI that could put a human at risk for obesity. The data collected by Palechor and Manotas (2019) included survey questions regarding consumption of high caloric food and consumption of water daily – “Do you eat high caloric food frequently?” and “How much water do you drink daily?”. Responses to “Do you eat high caloric food frequently?” were “yes” or “no”, these answers were transformed into binary making a “yes” response equivalent to “1” and a “no” response equivalent to “0”. The answers for “How much water do you drink daily?” were “less than a liter”, “between 1 and 2 liters”, and “more than 2 liters” were rounded into “1”, “2”, and “3” respectively. Palechor and Manotas (2019) also included survey questions regarding consumption of food between meals, consumption of vegetables, and number of main meals – “Do you eat any food between meals?”, “Do you usually eat vegetables”, and “How many main meals do you have daily?”. Responses to “Do you eat any food between meals?” were “No”, “Sometimes”, and “Frequently”, these responses were converted into “0”, “1”, and “2” respectively. The answers for “Do you usually eat vegetables” were indicated with 1,2, or 3. A value of 1 equals rare consumption of vegetables, whereas a value of 3 equals a frequent consumption of vegetables. The answers for “How many main meals do you have daily?” were either one, two, three, or four. After converting the survey responses into numerical values, a visual analysis was done which incorporated histograms, boxplots, scatter plots, and density plots.

**Figure 5. Histogram: BMI vs. Consumption of High Caloric Food Frequency**

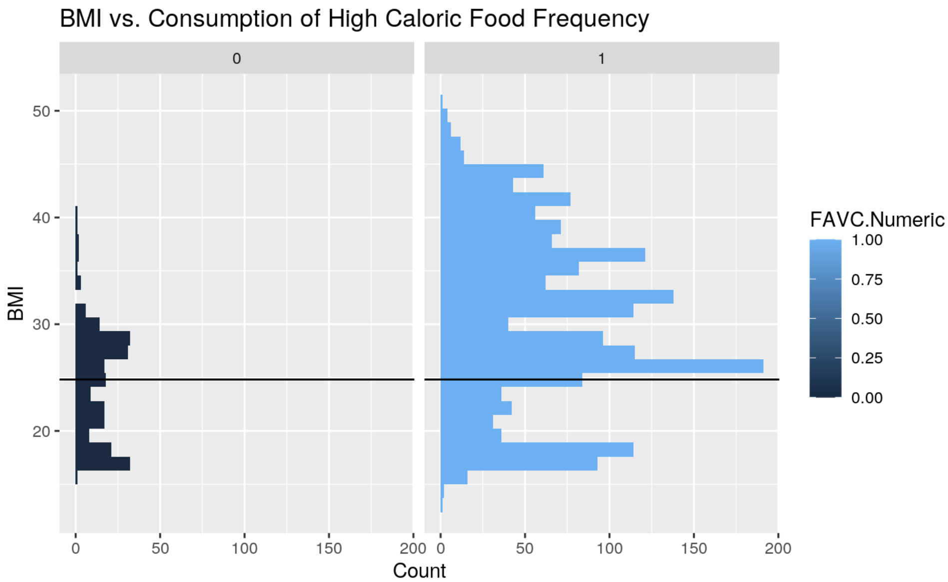
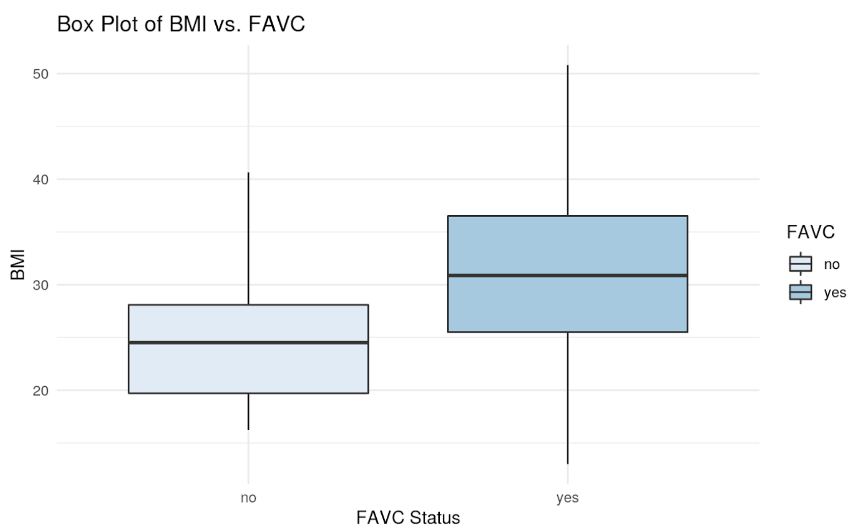



**Figure 6. Boxplot: BMI vs. Consumption of High Caloric Food Frequency (FAVC)**




From figure 5, we are able to observe that individuals who responded “no” to consuming high caloric food had a dramatically lower BMI than individuals who responded “yes” to consuming high caloric food. The overweight BMI, as shown by the geom_line, is 25. The “yes” group had a BMI IQR from 26 to 37 while the “no” group had a BMI IQR from 20 to 27.

From figure 6, we are able to observe that individuals who responded “yes” to consuming high caloric food had a median BMI of 31 and individuals who responded “no” had a median BMI of 24. Individuals from the “yes” group had a higher BMI overall than those in the “no” group.

**Figure 7. Scatterplot: BMI vs. Consumption of Water Daily**

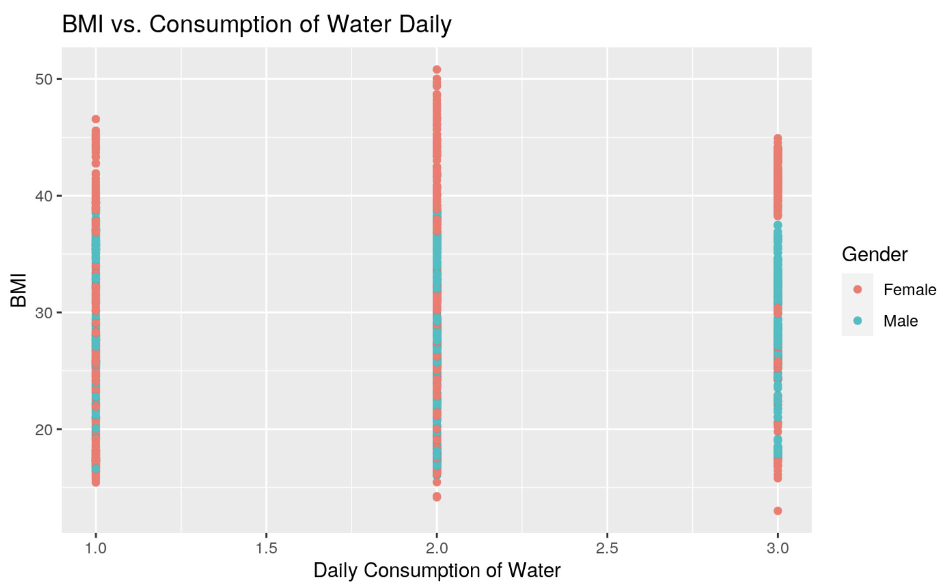
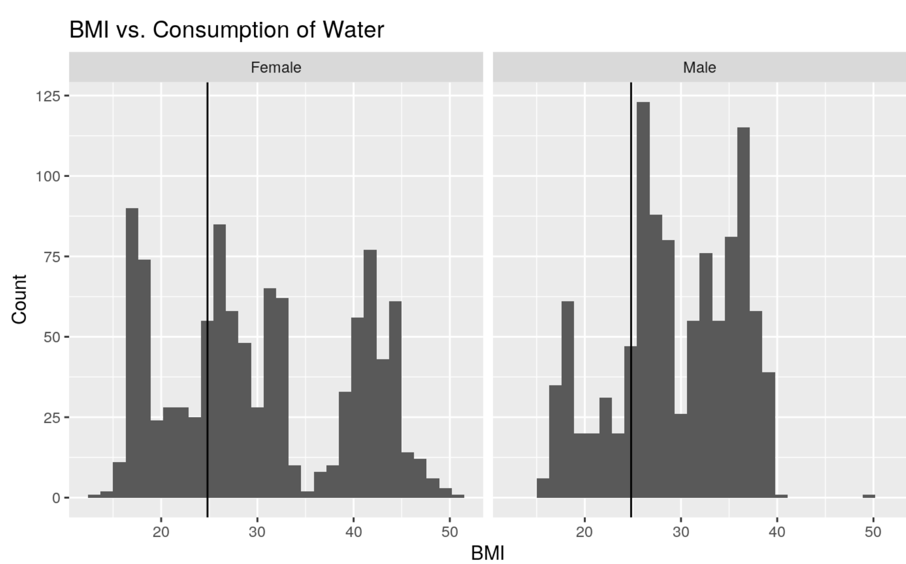



**Figure 8. Histogram: BMI vs. Consumption of Water Daily (CH20)**



The x-axis values on figure 7 correlate with responses to water consumption daily with 1=less than a liter, 2=between 1 and 2 L, and 3=more than 2L. Based on figure 7 there appears to be no clear correlation between BMI and consumption of water daily for females or males. From the histograms in figure 8, females show more variability in BMI than males, with females having a BMI ranging from 5 to 55 while males have a BMI ranging from 10 to 40.

**Figure 9. Boxplot: BMI vs. Consumption of food between meals (CAEC)**

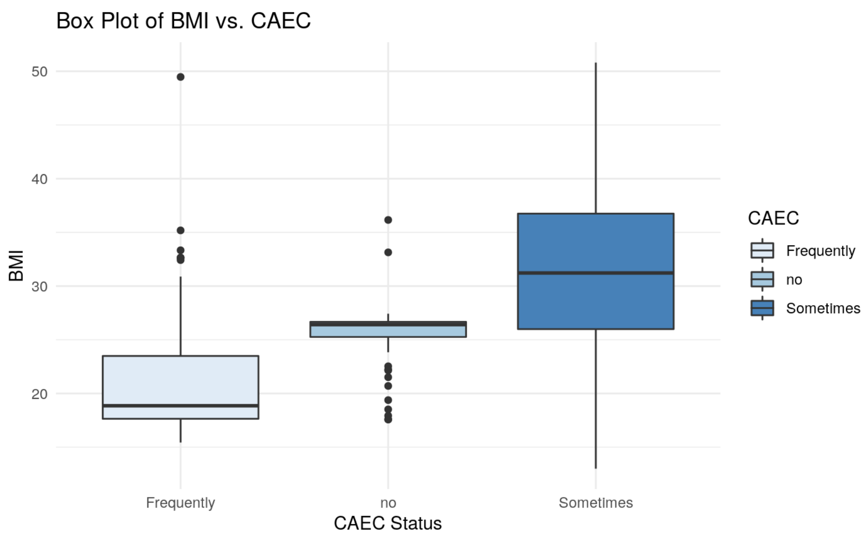
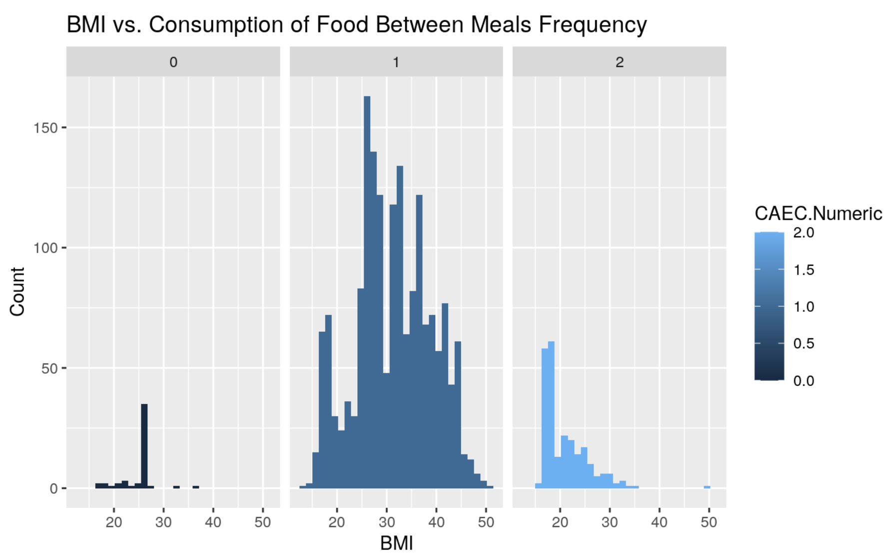



**Figure 10. Histogram: BMI vs. Consumption of food between meals (CAEC)**




Through the visualization in figure 9, we can observe that individuals who responded “sometimes” to having food between meals had a greater average BMI compared to those who answered “no” and those who answered “frequently”. In the “frequently” and “no” groups there are more outliers than in the “sometimes” group.

The histograms in figure 10 represent those who responded with 0=no, 1=sometimes, and 2=frequently. Those in the “no” group had a BMI that peaks at about 25. Those in the “sometimes” group had a BMI that peaks at 28 and those in the “frequently” group had a BMI that peaks at 18. The “sometimes” group had a greater count with high BMI than both the “no” and “frequently” groups.

**Figure 11. Heat map: BMI vs. Frequency of Consumption of Vegetables (FCVC**

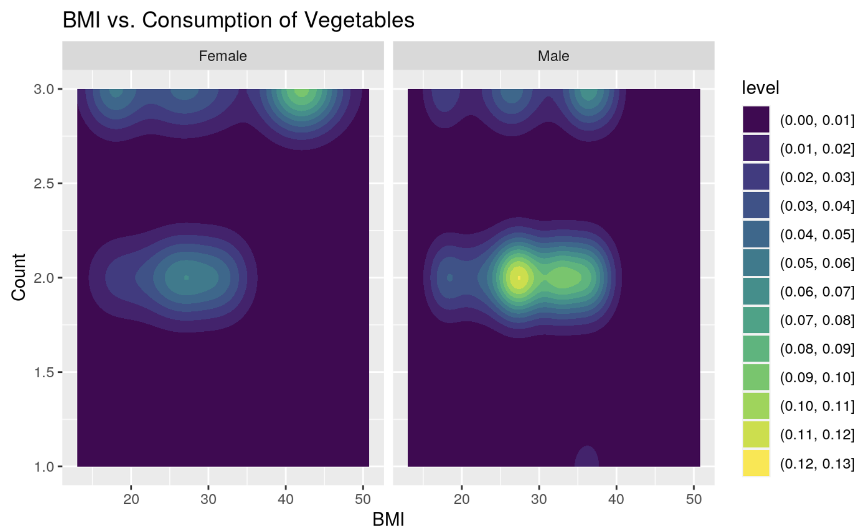
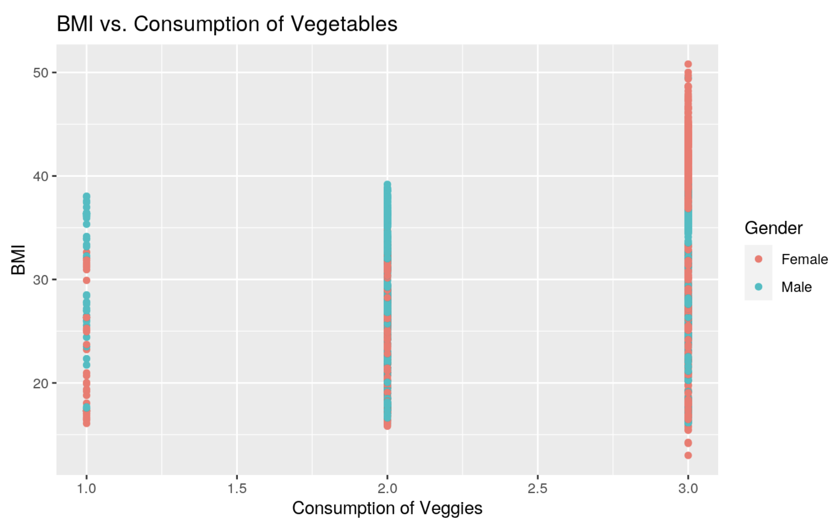



**Figure 12. Scatterplot: BMI vs. Frequency of Consumption of Vegetables (FCVC)**



Through the visualization in figure 11, we can see that males have a greater BMI than females for consumption of vegetables for count at 2.0 whereas for count at 3.0 males appear to have a lower BMI than females. The x-axis values on figure 12 correlate with responses to vegetable consumption daily with 1=never, 2=sometimes, and 3=always. Based on figure 12 there appears to be a greater BMI for females in the always category as compared to both the “sometimes” and “never” categories for females.

**Figure 13. Scatterplot: BMI vs. Number of Main Meals (NCP)**

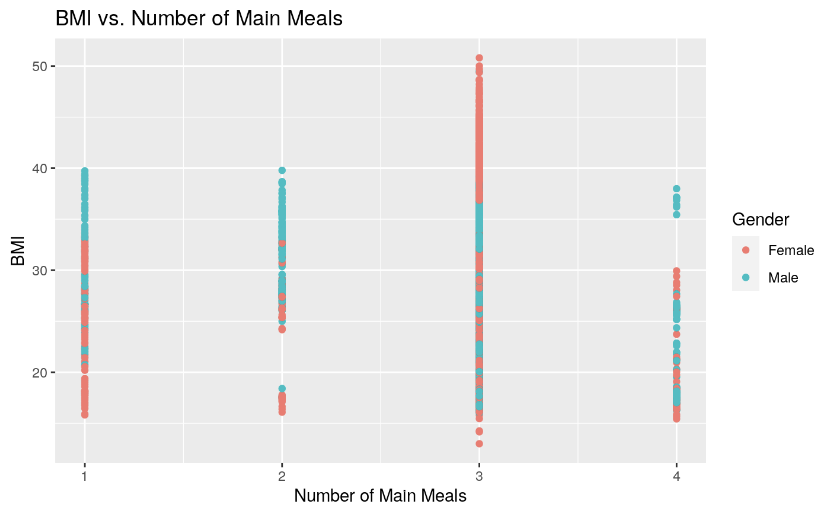
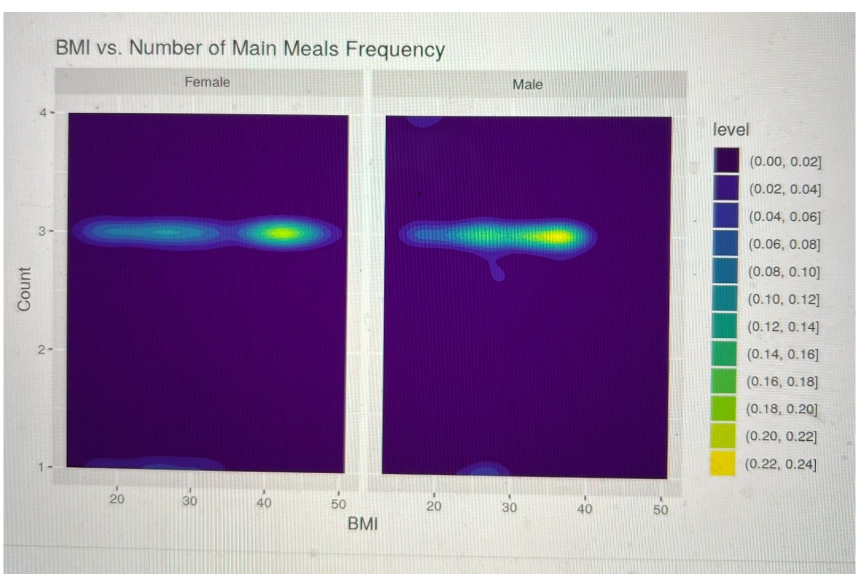



**Figure 14. Heat map: BMI vs. Number of Main Meals (NCP)**

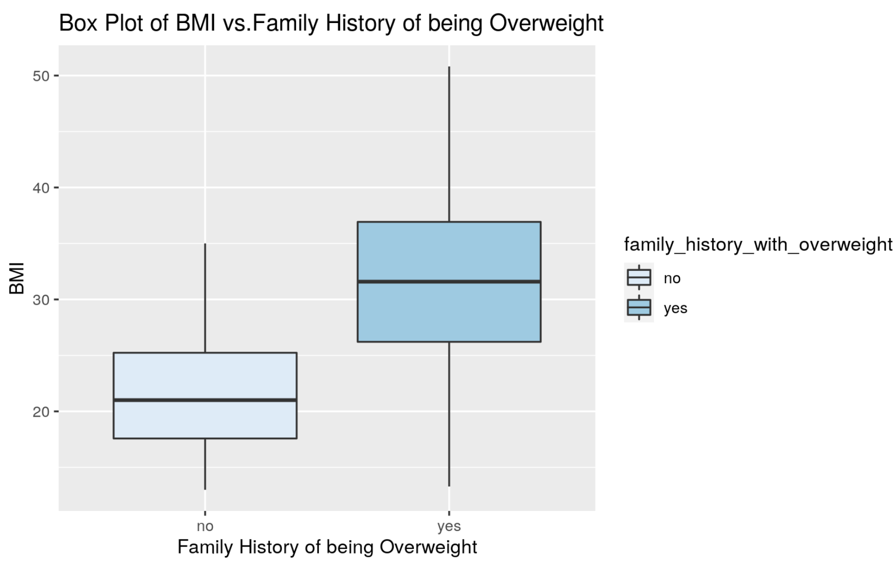


The x-axis values on figure 13 correlate with responses to the number of main meals with 1=never, 2=sometimes, and 3=always. Based on figure 13 there appears to be a greater BMI for females in the always category as compared to both the “sometimes” and “never” categories for females. Through the visualization for figure 14, we can see that males and females appear to have a relatively similar BMI for the number of main meals.

*Hypothesis 3: Genetic Factors*

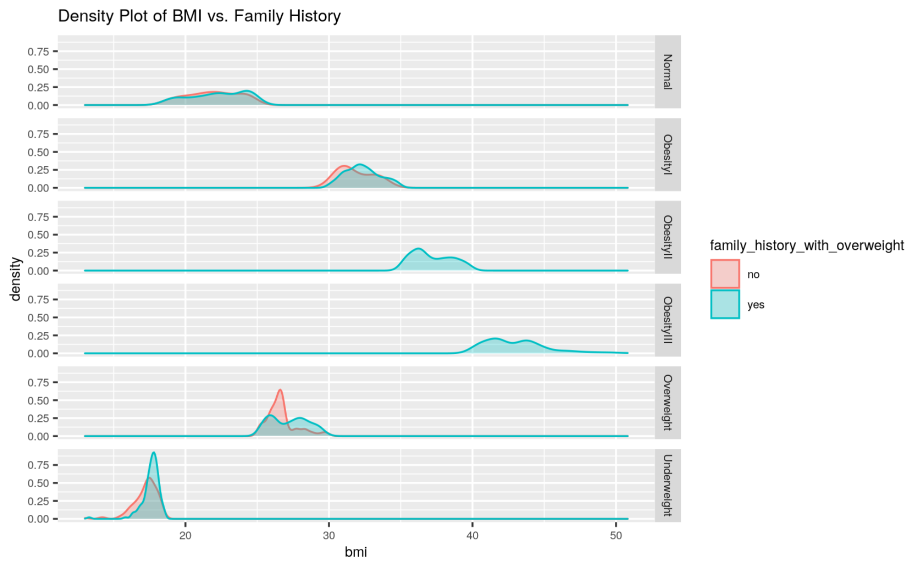
It is predicted that coming from a family with a history of being overweight results in a higher probability of developing a higher BMI. The data collected included survey questions regarding family history – “Has a family member suffered or suffers from overweight?”. The response to this question was a “Yes” or “No” that were converted into binary values for ease of data analysis with Yes corresponding to 1 and No corresponding to 0. After converting the survey answers into numerical values, a visual analysis was done, which incorporated boxplots, density plots, scatter plots and clustering.

**Figure 15. Box plot: BMI vs. Family History of Being Overweight**



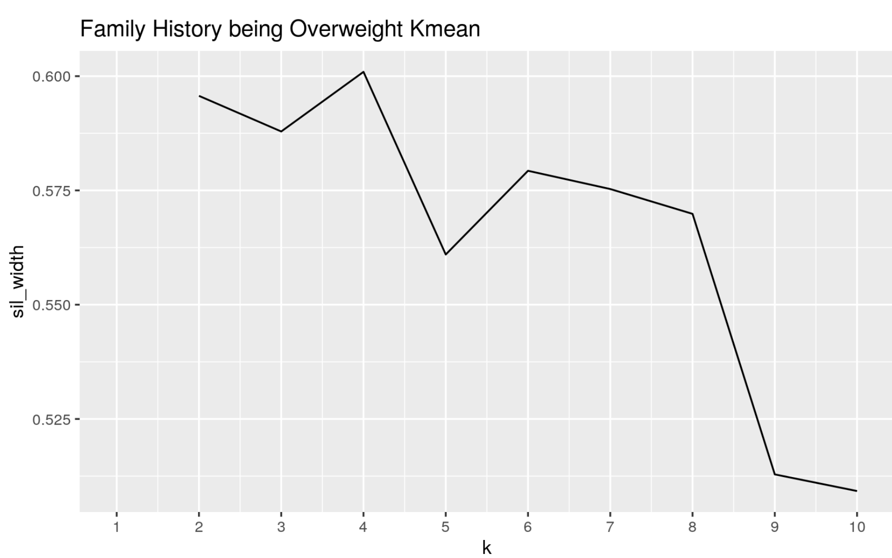

Through this visualization, we can clearly see that the individuals who responded “Yes” for a family member suffering from being overweight resulted in a higher overall BMI that ranged from 27 to 37 based on lower and upper quartiles in the dark blue box plot. On the other hand, individuals who responded “No” for this question had a lower overall BMI that ranged from 18 to 25 based on lower and upper quartiles in the light blue box plot.

**Figure 16.  Density plot: BMI vs. Family History**



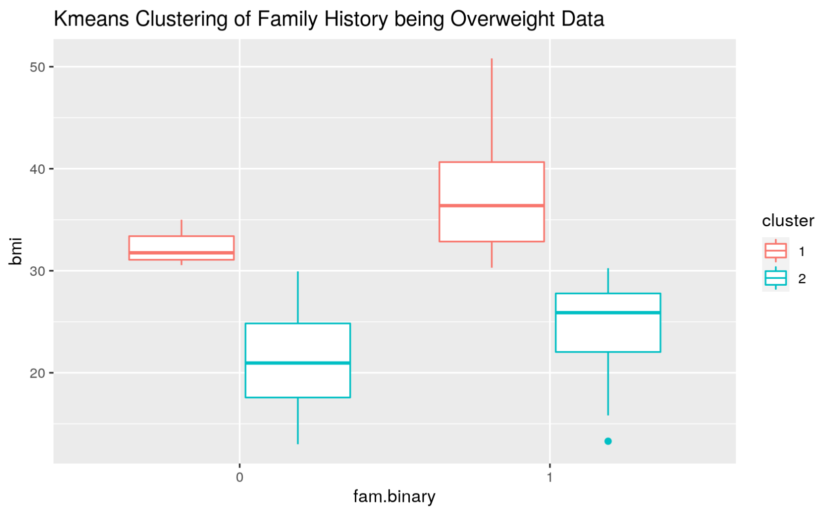

This is another way of visualizing the data and through this density plot we can see that as the levels of weight classifications increase from Underweight to ObesityIII, there is no “red” or “no” portion on the graphs specifically for Obesity II and Obesity III. This depicts that all the individuals who answered “Yes” for family overweight genes had a higher BMI value as they are towards the right side of this plot. There is some overlap with responses with the other weight classes, however, we can attribute this to a multitude of variables such as sedentary lifestyle and diet choices.

**Figure 17. Silhouette Width plot**



After visualizing the data, we decided to use k-means clustering on the data. The silhouette width was maximized at k=4 clusters, however, for running the algorithm, I used k =2 clusters since the sil_width value for both clusters are very close to each other and the smaller number of the clusters is better in order to identify simpler similarities to interpret.

**Figure 18. K-means clusters: BMI vs Family History with Overweight**



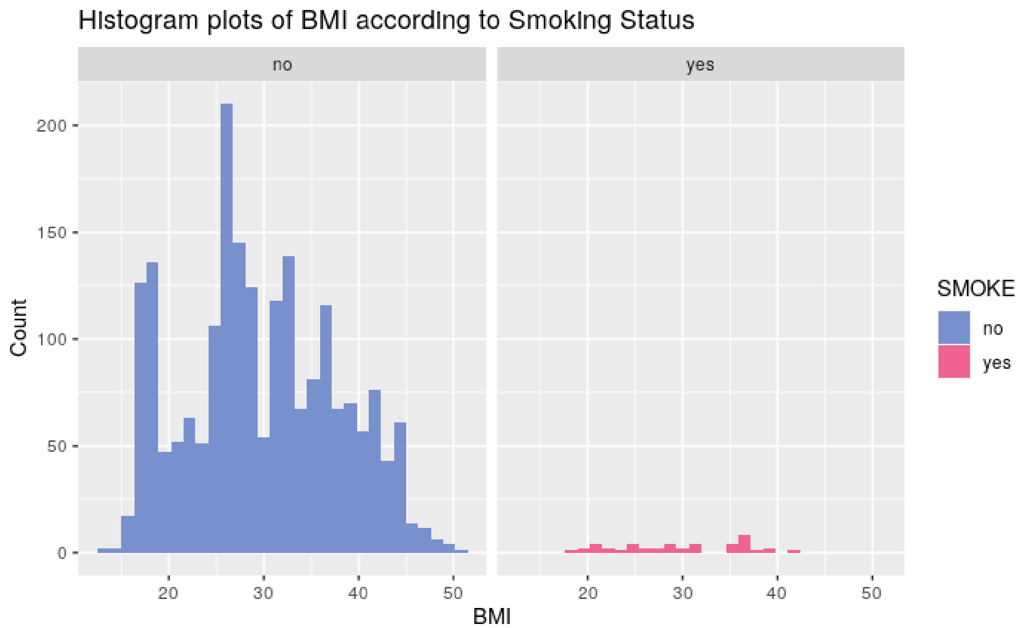

From the clustering task we can see how the data was split based on the BMI classification. A healthy BMI for an adult is between 18.5 and 24.9. The data is split at a BMI of 30. Within both clusters, it can easily be observed how the higher median BMI tends to be correlated with a family history of being overweight. The median BMI values for fam.binary = 1 (Overweight genes) were higher for both clusters compared to the median BMI values for fam.binary = 0.

*Hypothesis 4: Substance use*

It is claimed that substance use such as smoking tobacco products and drinking alcohol results in increased risks of developing a higher BMI. The data collected by Palechor and Manotas (2019) included survey questions regarding smoking status and Alcohol intake—"Do you smoke?" and "How often do you drink alcohol?". The answers for "Do you smoke?" were either "yes" or "no", which were then converted into binary (1 = yes, and 0 = no). The answers for "How often do you drink?" were "I do not drink", "Sometimes", "Frequently", and "Always", which were then given the numerical values "0", "1", "2", and "3" respectively.

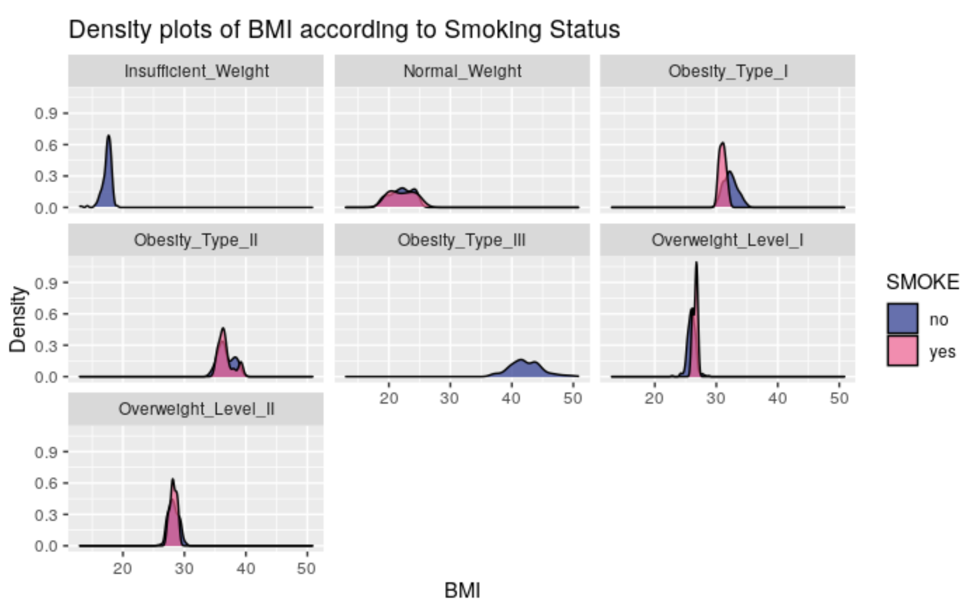
After converting the survey answers into numerical values, a visual analysis was done, which incorporated histograms, density plots, and box plots.

**Figure 19. Histogram plot: BMI according to Smoking status**




From the histograms in figure 19, we are able to observe that the amount of people who reported that they do smoke is dramatically lower than those who do not smoke. The non-smoking distribution (labeled “no”) peaks at a BMI of approximately 28, and the smoking distribution (labeled “yes”) seems to be fairly uniform with no clear median value.

**Figure 20. Density plot: BMI according to smoking status & faceted by obesity levels classification**



The density plots in figure 20 reveal that once the data is faceted by obesity levels classification (Insufficient_Weight, Normal_Weight, Obesity_Type_I, Obesity_Type_II, Obesity_Type_III, Overweight_Level_I, and Overweight_Level_II), there is not a noticeable difference between the BMI of those who smoke and those who do not smoke. Each density plot shows a similar distribution excluding the plots with not enough recorded smokers (Insufficient_Weight & Obesity_Type_III).

**Figure 21. Box plot: BMI vs. Smoking status**

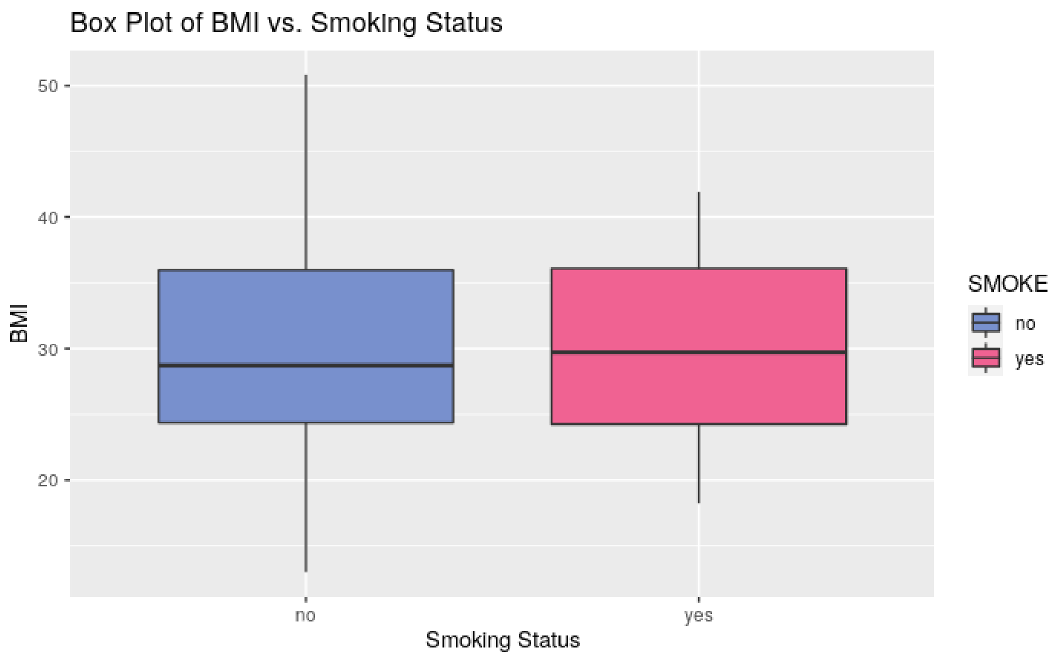



**Table 1. Summary statistics on BMI according to smoking status**

**Summary Statistics on BMI for Smokers and Non-Smokers**

| ***Smoking Status*** | *Does not smoke* | *Smokes* |
|---|---|---|
| ***n (count)*** | 2067 | 44 |
| ***Min BMI*** | 13 | 18.2 |
| ***Max BMI*** | 50.8 | 41.5 |
| ***Median BMI*** | 28.7 | 29.7 |
| ***Mean BMI*** | 29.7 | 29.7 |
| ***Standard Dev.*** | 8.04 | 6.6 |

A box plot and a table of summary statistics on BMI based on smoking status were also produced and are shown in Figure 21 and table 1. From box plot and summary statistics, we observe that the median BMI of non-smokers is 28.7 and 29.7 for smokers, and the means for both groups are the same (29.7). The data for non-smokers has a larger spread with a standard deviation of 8.04 and the spread of the smoker's BMI is smaller with a standard deviation of 6.6.

**Figure 22. Silhouette plot: Smoking status**

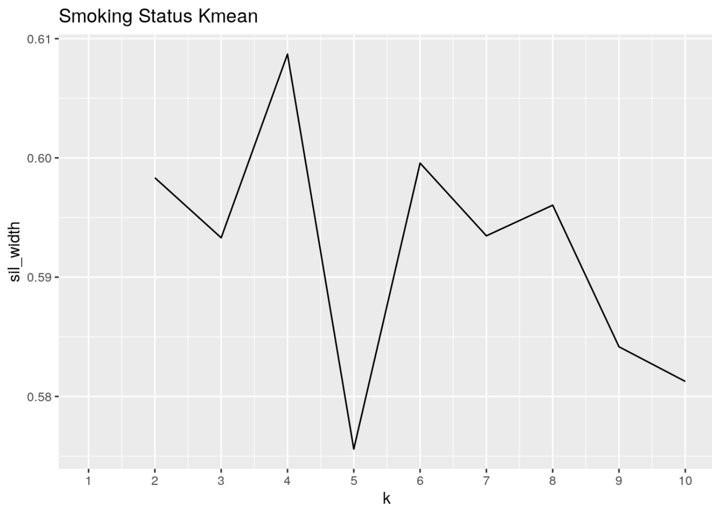
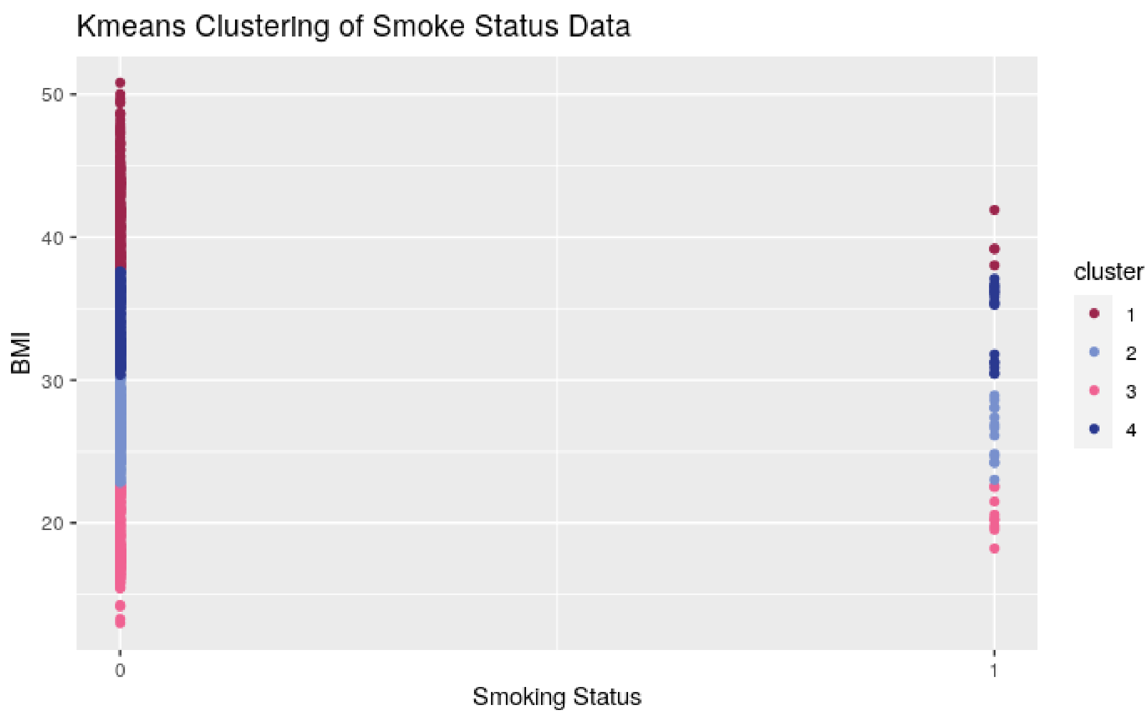
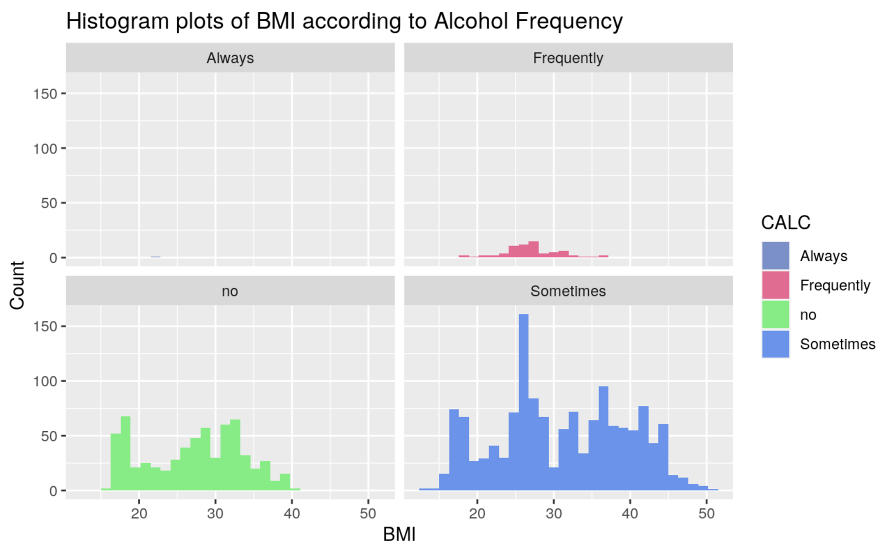



**Figure 23. K-means clusters: Smoking status**



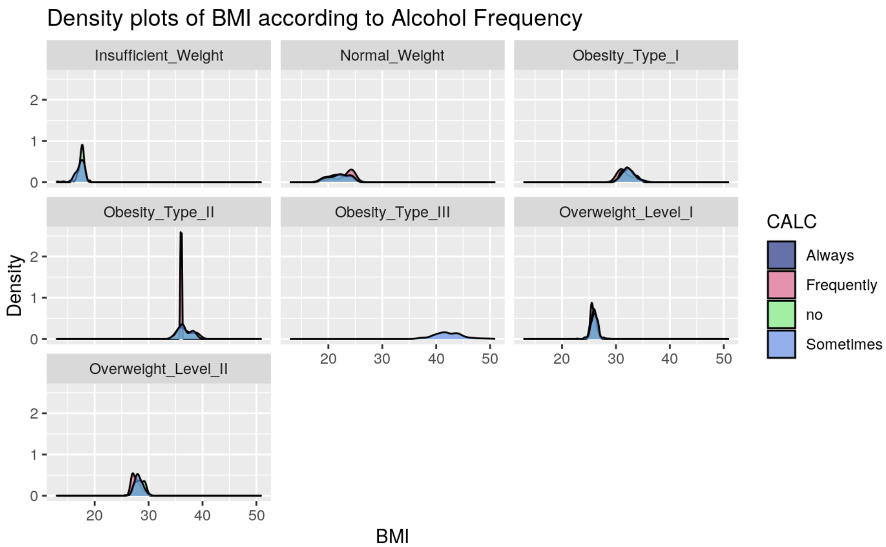

Following the data visualization for smoking status, we performed K-means clustering on the "SMOKE" data. The silhouette width was determined to be maximized at 4 clusters (k = 4) as shown in the silhouette plot in Figure 22. Using k = 4 clusters, the clustering task was performed and visualized in Figure 23. We observe 2 clusters above and below BMI levels that indicate obesity (BMI > 30). There is little to no difference between the clustering when comparing between the data of smokers and non-smokers, which aligns with the previous observations from the visualizations.

**Figure 24. Histogram plot: BMI according to Alcohol Consumption Frequency**




From the histograms in figure 24 , we are able to observe the number of individuals who reported at which frequency they consumed alcohol. The greatest variability in BMI distribution can be seen in those that sometimes consume alcohol with BMIs ranging from around 12 to 50. Those that always drink alcohol there seem to be lower BMIs ranging from 20-40. In addition those that did not consume alcohol also ranged from 20-40 BMIs with a peak at around 27. Finally those that consumed alcohol frequently showed a left right skew and had BMIs that ranged from 20-35. Just by looking at the distribution there does not seem to be a relationship between alcohol consumption and BMI.

**Figure 25. Density plot: BMI according to Alcohol Consumption Frequency & faceted by obesity levels classification**




The density plots in figure 25 reveal that once the data is faceted by obesity levels classification (Insufficient_Weight, Normal_Weight, Obesity_Type_I, Obesity_Type_II, Obesity_Type_III, Overweight_Level_I, and Overweight_Level_II), there is not a noticeable difference between the four categories of those that consume alcohol frequently, sometimes, always, and never. Most of the density plots show a similar distribution, however in the Obesity_Type_II plot we can see that those that frequently consume alcohol have a peak BMI of around 35.

**Figure 26. Box plot: BMI vs.  Alcohol Consumption Frequency**

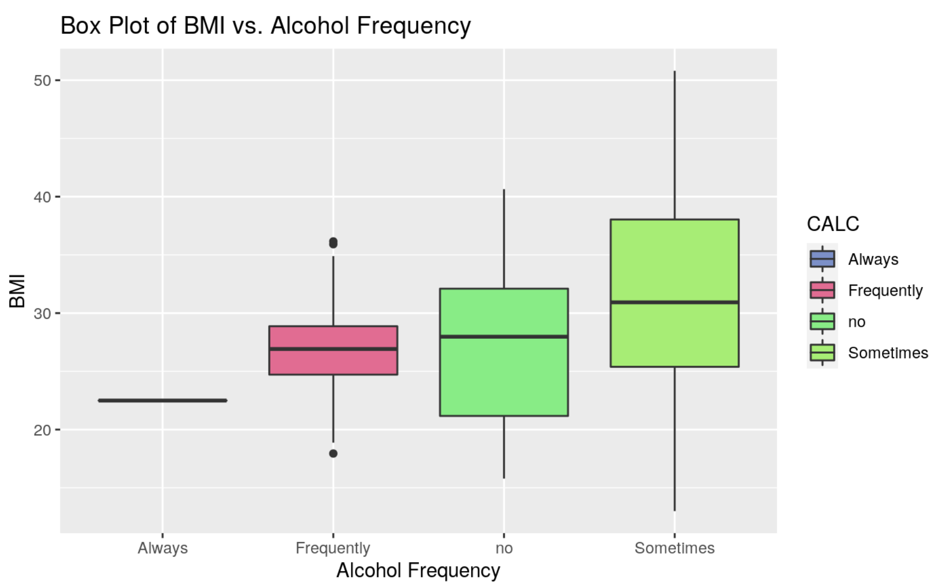


**Table 2. Summary statistics on BMI according alcohol consumption**

**Summary Statistics on BMI and Alcohol Consumption Frequency**

| ***Frequency of Alcohol Consumption*** | *Never* | *Sometimes* | *Frequently* | *Always* |
|---|---|---|---|---|
| ***n (count)*** | 639 | 1401 | 70 | 1 |
| ***Min BMI*** | 15.8 | 13 | 17.9 | 22.5 |
| ***Max BMI*** | 40.6 | 50.8 | 36.2 | 22.5 |
| ***Median BMI*** | 28 | 31 | 26.9 | 22.5 |
| ***Mean BMI*** | 27.1 | 31 | 27 | 22.5 |
| ***Standard Dev.*** | 6.4 | 8.5 | 3.7 | |

A box plot and a table of summary statistics on BMI based on alcohol consumption frequency were also produced and are shown in Figures 26 and table 2. From box plot and summary statistics, we observe that the median BMI of those that frequently consume alcohol is 26.9, those that never consume alcohol 28, those that always consume alcohol 22.5, and those that sometimes consume alcohol is 31. The means for those that Never consume alcohol, those that sometimes do, those that frequently do, and those that always do are 27.1, 31, 27, and 22.5 respectively. The data for those that consume alcohol sometimes has a larger spread with a standard deviation of 8.5, and the spread of those that frequently consume alcohol is smaller with a standard deviation of 3.7.

**Figure 27. Silhouette plot: Alcohol Consumption Frequency**

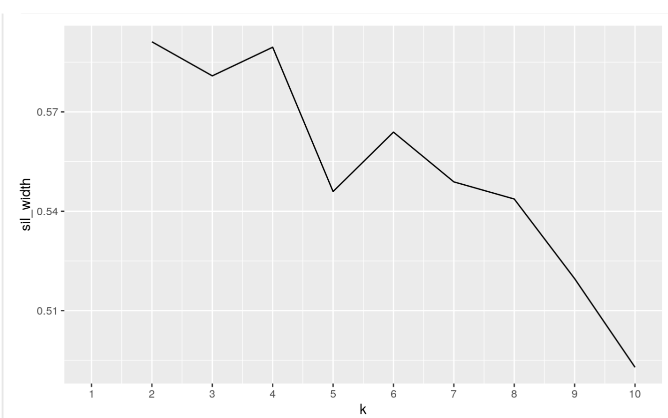
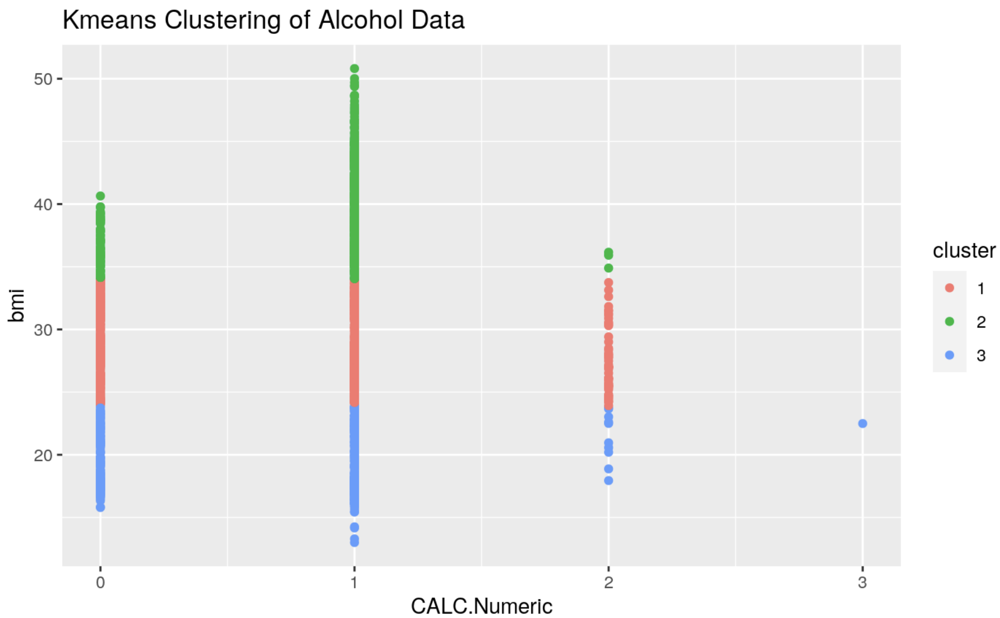


**Figure 28. K-means clusters: Alcohol Consumption Frequency**




Following the data visualization for smoking status, we performed K-means clustering on the “CALC” data. The silhouette width was determined to be maximized at 4 clusters (k = 4) as shown in the silhouette plot in Figure 27. Using k = 4 clusters, the clustering task was performed and visualized in Figure 28. The numbers on the x-axis relate to the four categories of alcohol consumption with 0=never, 1=sometimes, 2=frequently, and 3=always. Based upon the clustering we do not see a clear relationship between alcohol consumption frequency and BMI like we clearly did with the Frequency of Physical Activity cluster analysis.

## Modeling Weight Classification

Setup the Data for Analysis

Before performing any modeling analysis, we needed to convert all qualitative variables into quantitative variables. For example, variables such gender, family history, and weight category were all non-numeric and thus needed to be converted. A table of subset of qualitative variables and their converted quantitative counterpart is shown below:

| Variable Name | Variable Description | Survey Answers |
|---|---|---|
| **Gender** | Female<br>Male | 1<br>0 |
| **Family History With Overweight** | yes<br>no | 1<br>0 |
| **Weight Category** | Insufficient-Weight<br>Normal-Weight<br>Overweight-Level-I<br>Overweight-Level-II<br>Obesity-Type-I<br>Obesity-Type-II<br>Obesity-Type-III | -1<br>0<br>1<br>2<br>3<br>4<br>5 |

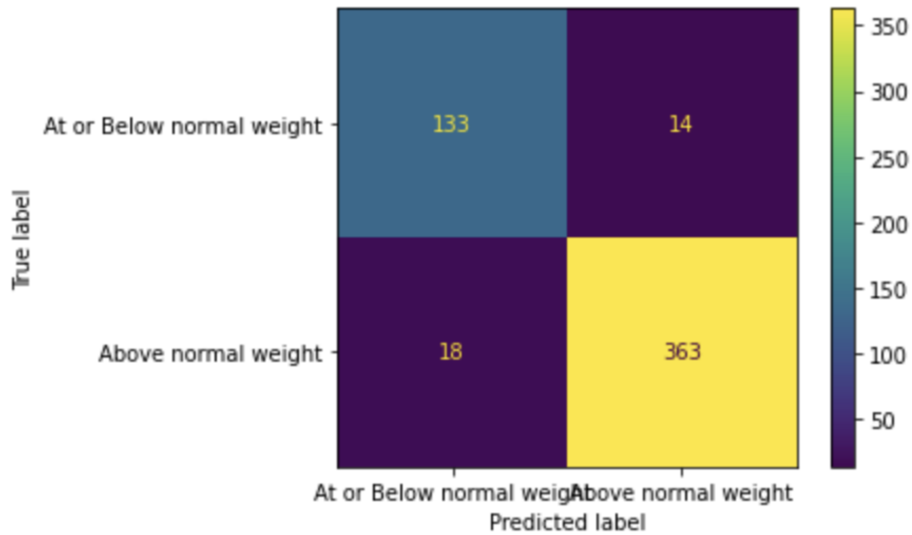
Once the data had fully been converted to numeric values, we selected a hypothesis and models to test this hypothesis.

Select Hypothesis and Models

Our hypothesis was to test if the variables/features of this dataset could accurately predict the weight classification of a new observation. To test this hypothesis, we decided to model using both the SVM and k-Nearest Neighbors models. We decided SVM would be a good first model because it is a linear model and, therefore, less likely to overfit our data. We then compared the results of the simple SVM model to the more complex k-Nearest Neighbors model to obtain an understanding of which model better predicted the weight classification of a new observation.

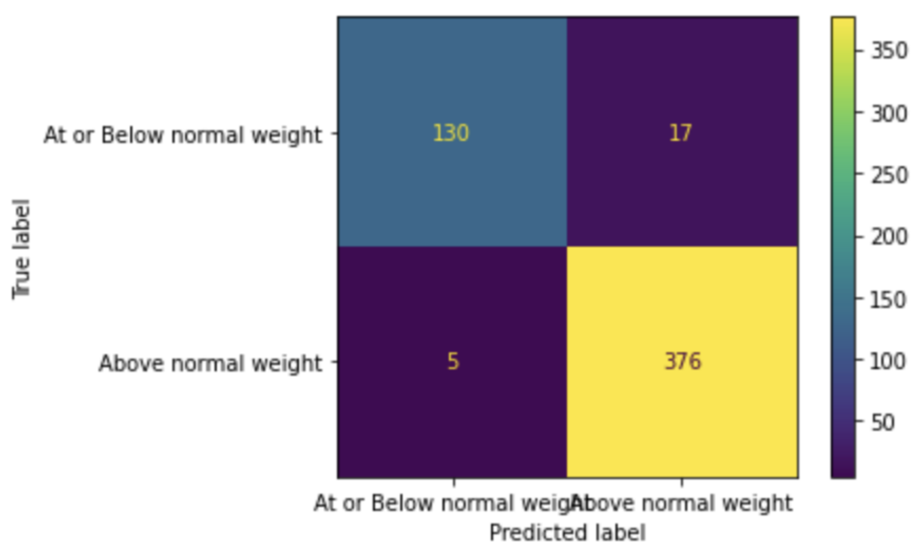
Results from SVM and k-Nearest Neighbors

Before modeling the data, we first split the data into training and test sets. After the dataset had been split, we ran the SVM and k-Nearest fit using the training set and computed an accuracy score and confusion matrix for each using the test set. For SVM, we calculated an accuracy score of 0.9393 meaning that the SVM model was able to accurately predict the weight classification 93% of the time. For k-Nearest, we calculated a recall score of 0.9868 meaning the model predicted new positive samples with a high degree of accuracy. Next, we graphed the confusion matrix for both the SVM and k-Nearest models; these graphs are shown below:




*Confusion matrix for SVM Model*

From the confusion matrix for the SVM model, we can see that the model predicted the correct class more often than it predicted the wrong class. More specifically, we can see that the model correctly classified 363 new observations to be above normal weight whereas the model incorrectly classified only 18 to be at or above normal weight. Similarly, the model correctly classified 133 new observations to be at or below normal weight where it incorrectly classified only 14 to be above normal weight.



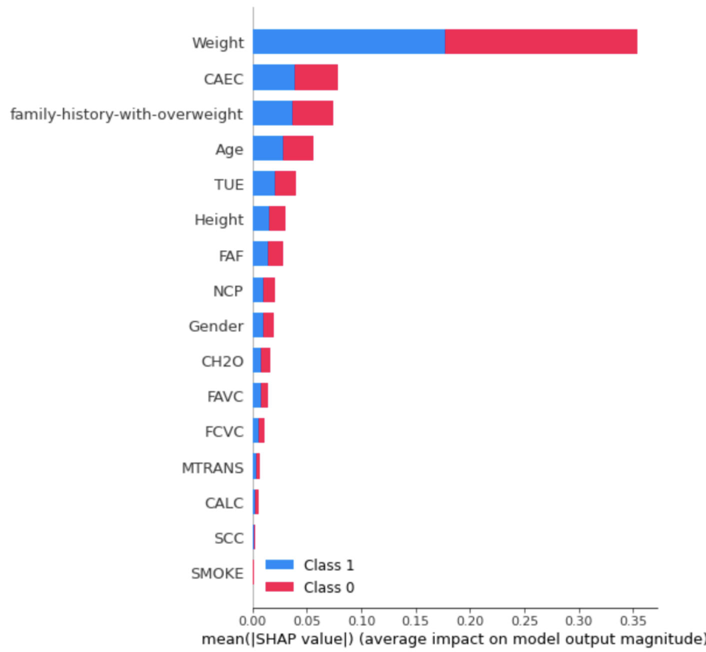

*Confusion matrix for k-Nearest Neighbors Model*

From the confusion matrix for the k-Nearest model, we can immediately see the supremacy of the k-Nearest model over the SVM model in terms of correctly classifying new observations.

Specifically, we can see that k-Nearest was able to correctly classify 376 new observations to be above normal weight (13 more than SVM). Similarly, k-Nearest only incorrectly classified 5 new observations as at or above normal weight (13 fewer than SVM).

<u>Investigating Important Features</u>

After quantifying the efficacy of each model, we investigated the importance of each variable in determining the classification of an observation using SHAP. The figure to the right displays the importance of each variable by its SHAP value (larger SHAP values mean larger importance).




From the figure, we can see that the most significant variable in determining the classification of the observation was weight. The next three most significant variables were CAEC (consumption of food between meals), family history with overweight, and age. This significance of most of these variables comes as unsurprising determinants of weight classification.

## Discussion/Conclusion/Limitations

We found that the SVM model correctly predicted the weight classification for new observations 93% of the time. In addition, the The k-Nearest model had a recall score of 0.9868 meaning the model predicted positive new observations with a high degree of accuracy. We found that the most significant determinants of weight classification were weight, CAEC (consumption of food between meals), family history of overweight, and age Our models suggest that these variables can be used to accurately predict the weight classification of a new observation. For limitations, we can increase the sample size to get a more representative distribution. This can be done by expanding our sample to include other Latin American countries. Another limitation to the study was no individual analysis was done on each country (Mexico, Peru, and Colombia) included in the study. Conducting an individual analysis can provide even further information as to how the explanatory variables relate to the response variable (BMI) depending on which country the individual is from.

**Acknowledgement**


| Team Member Name | Contribution |
|---|---|
| Rhea Bhat | 100% |
| Yaashi Khatri | 100% |
| Hiba Ansari | 100% |
| Natalie Nguyen | 100% |
| Neha George | 100% |
| Andrea Flores | 100% |
| Ian Wood | 100% |


**Bibliography**

Links for Data:

https://archive.ics.uci.edu/ml/machine-learning-databases/00544/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6710633/

References:


Beccuti, G., & Pannain, S. (2011). Sleep and obesity: Current Opinion in Clinical Nutrition and Metabolic Care, 14(4), 402–412. https://doi.org/10.1097/MCO.0b013e3283479109

Fock, K. M., & Khoo, J. (2013). Diet and exercise in management of obesity and overweight: Diet and exercise for weight management. Journal of Gastroenterology and Hepatology, 28, 59–63. https://doi.org/10.1111/jgh.12407

Fruh, S. M. (2017). Obesity: Risk factors, complications, and strategies for sustainable long‐term weight management. Journal of the American Association of Nurse Practitioners, 29(S1), S3–S14. https://doi.org/10.1002/2327-6924.12510

Hruby, A., Manson, J. E., Qi, L., Malik, V. S., Rimm, E. B., Sun, Q., Willett, W. C., & Hu, F. B. (2016). Determinants and consequences of obesity. American Journal of Public Health, 106(9), 1656–1662. https://doi.org/10.2105/AJPH.2016.303326

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

World Health Organization. (n.d.). Obesity and overweight. World Health Organization. Retrieved October 11, 2022, from https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight