Estimation of Obesity Levels

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

Problem Introduction

- According to WHO, being overweight or obese is defined as abnormal or excessive fat accumulation that presents a risk to health. A body mass index (BMI) over 25 is considered overweight, and over 30 is obese. The issue has grown to epidemic proportions, with over 4 million people dying each year as a result of being overweight or obese in 2017 according to the global burden of disease.(WHO, 2021).
- As a result, the collaborative efforts to understand overweight and obesity and to promote healthy weight is an area to be researched on. In line with this, this project aims at predicting the obesity level of individuals.

Motivation

- Obesity is one of the biggest health problems we have globally and it has a pervasive impact on health, affecting every organ system in the body.
- Obesity-related chronic conditions cause a lot of physical suffering and emotional suffering from social stigma at work and in relationships with other people.
- Our detailed analysis of this dataset can validate the impact of several factors that propitiate the apparition of obesity problems such as diet,physical activity, lifestyle choices and different family upbringings.
- From our work, we could identify what nutritional or lifestyle changes are needed to be made to ensure a healthier individual.

Data Description

- The dataset under study includes data for the estimation of obesity levels in individuals based on their eating habits and physical condition in Mexico, Peru ,and Columbia.
- The data contains 17 attributes and 2111 individuals aged 14 to 61. The 17 attributes are related to individual habits that are likely to determine obesity levels, such as number of main meals, time using technology devices, genetics,gender, and transportation used.
- The 2111 individuals are labeled with the class variable NObeyesdad (Obesity Level), that allows classification of the data using the values of Insufficient Weight (Underweight), Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II and Obesity Type III.

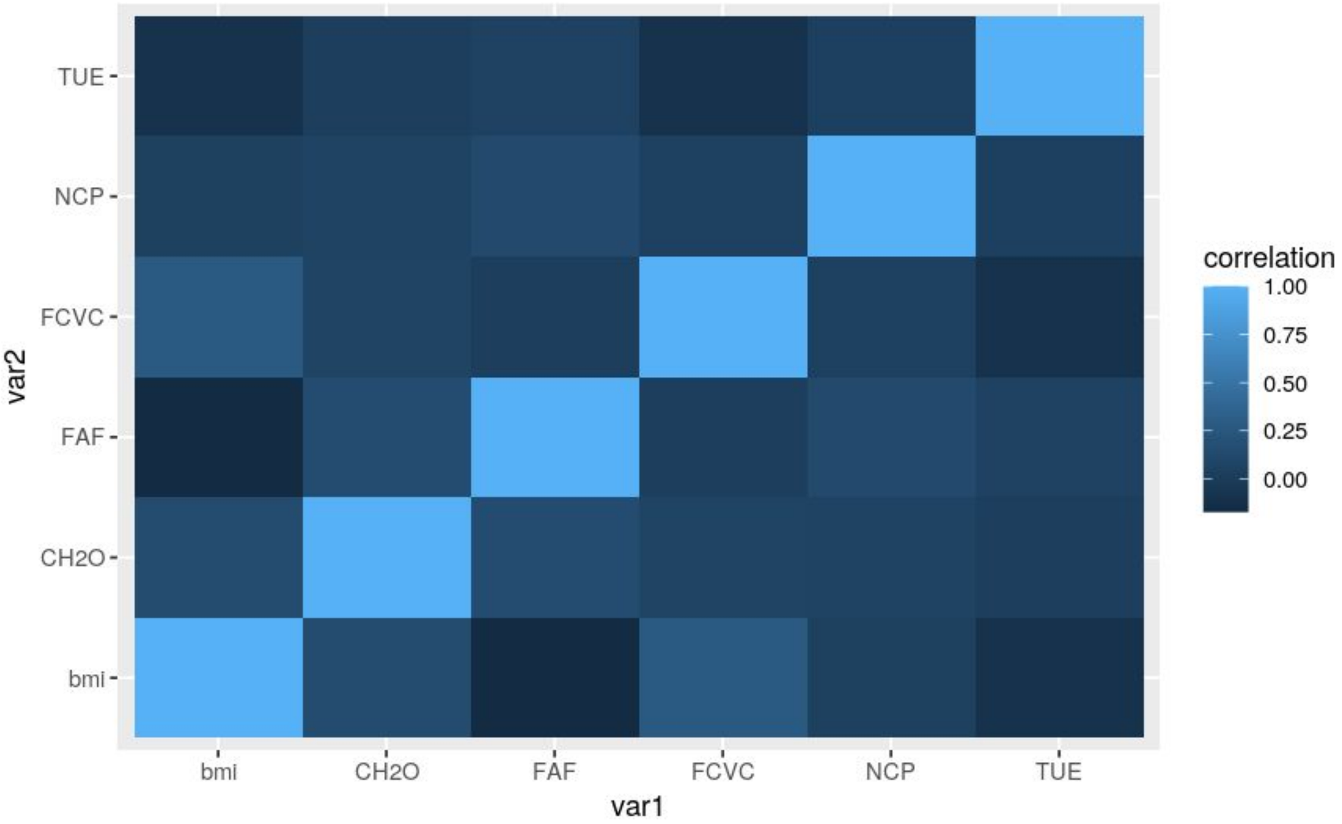
Cleaning Procedure

The data was already set in a format that was usable for analysis. We just need to create new variables for our exploratory analysis and modeling.

1. Save as a dataframe
2. Omit the NA values
3. New variable **id**
4. New variable **bmi**: bmi= (Weight) / $(\text{Height})^2$.
5. Binary variable conversions & rounding
    a. `odata <- odata %>% mutate(SMOKE.binary= case_when(SMOKE=="yes" ~ 1, SMOKE=="no" ~ 0 ))`

Exploratory Analysis

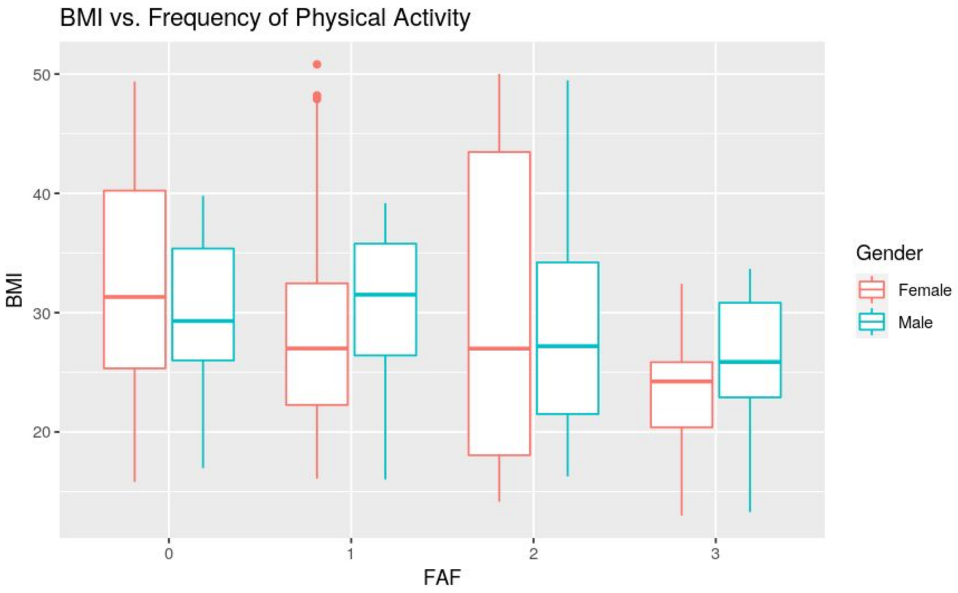
Correlation Matrix of Numeric Variables



Variable	Definition
TUE	Time using technology devices
NCP	Number of main meals
FCVC	Frequency consumption of high caloric food
FAF	Frequency of Physical Activity
CH2O	Consumption of water daily

Hypothesis 1: Lifestyle

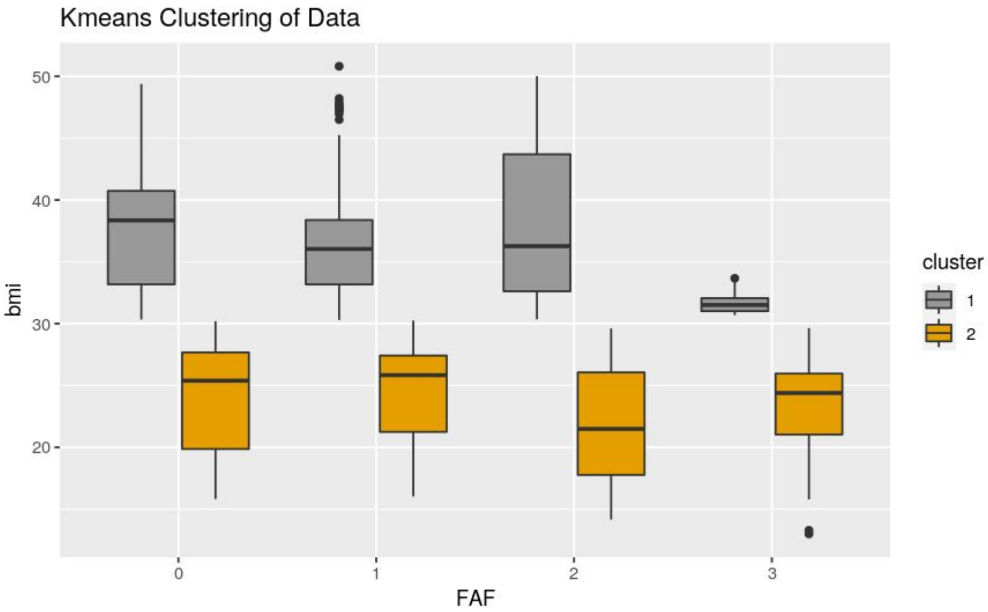
- It is claimed that a sedentary lifestyle results in a higher BMI that could put a human at risk for obesity.
- Women have a wider range of BMIs and are more sedentary
- Variable: FAF



Hypothesis 1: Lifestyle

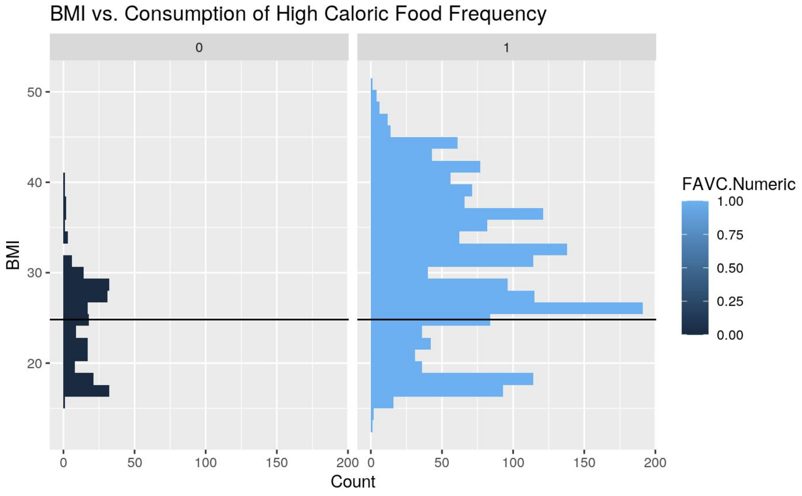
- Clusters were based off of BMI range
- Clear results of the inverse relationship between FAF and BMI
- Variable: FAF

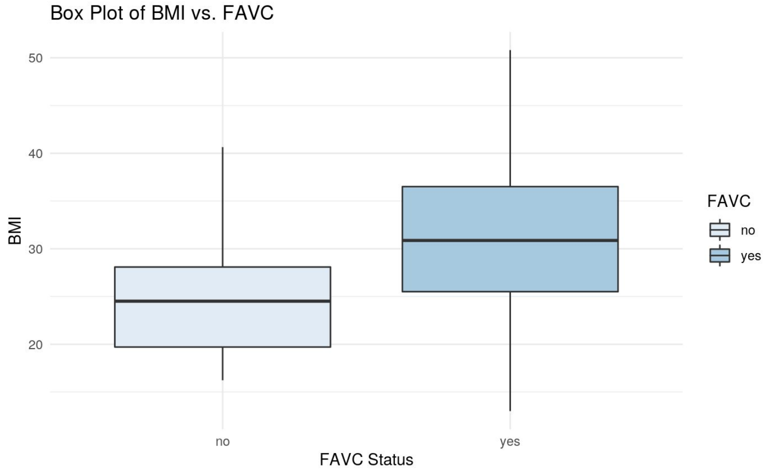
Healthy BMI: 25.0 - 29.9



Hypothesis 2: Diet

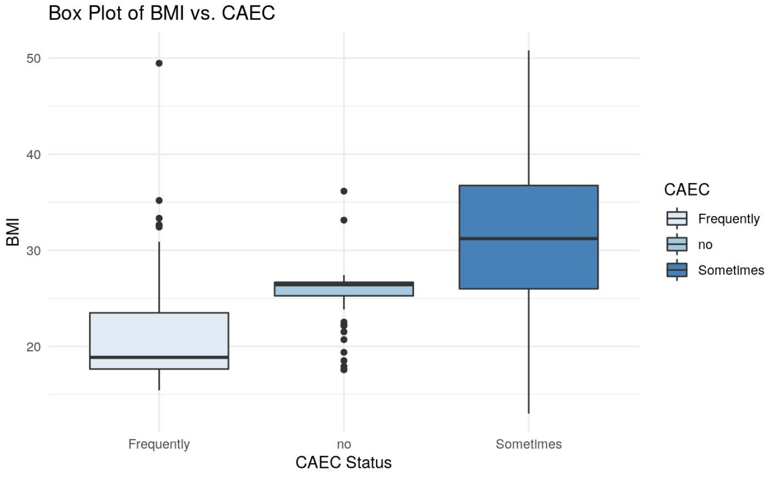
- It is claimed that a diet of high caloric food and poor nutrition results in a higher BMI that could put a human at risk for obesity.
- Variables: FAVC

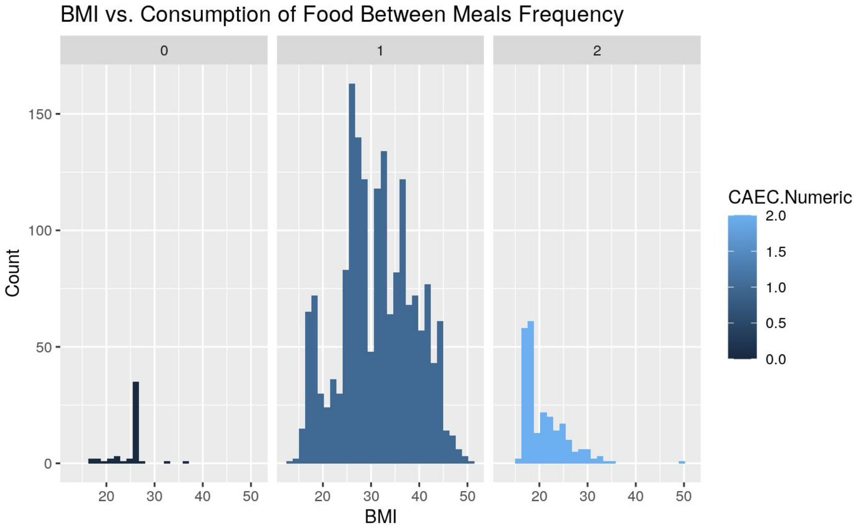




Hypothesis 2: Diet

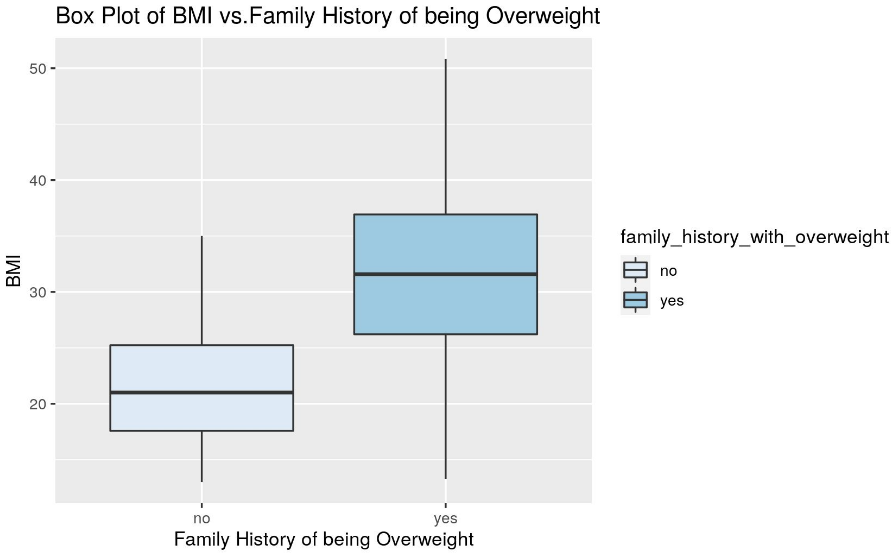
- It is claimed that a diet of high caloric food and poor nutrition results in a higher BMI that could put a human at risk for obesity.
- Variable: CAEC

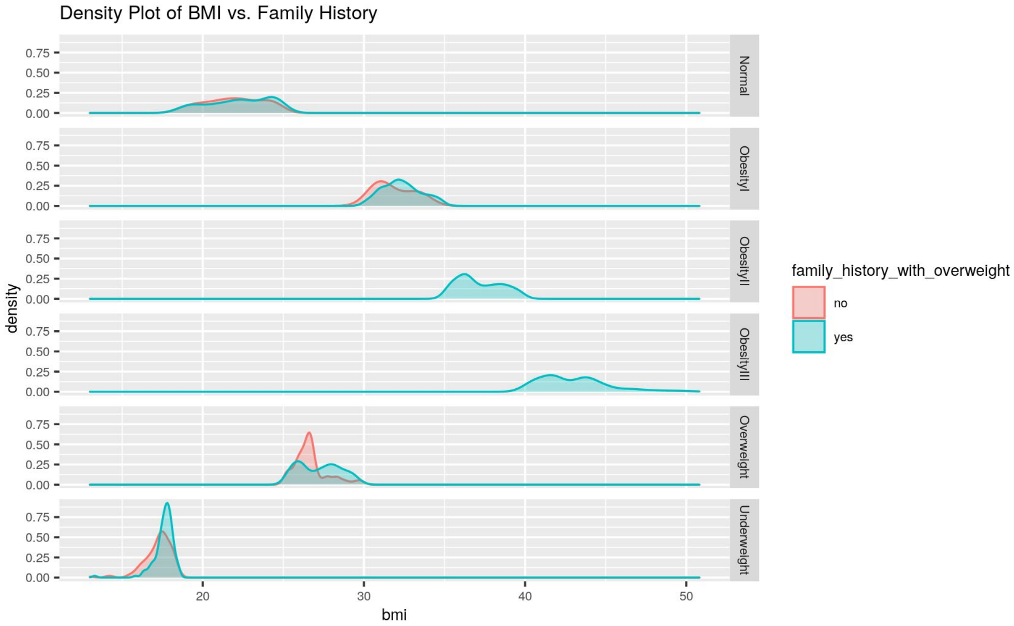




Hypothesis 3: Genetic Factors

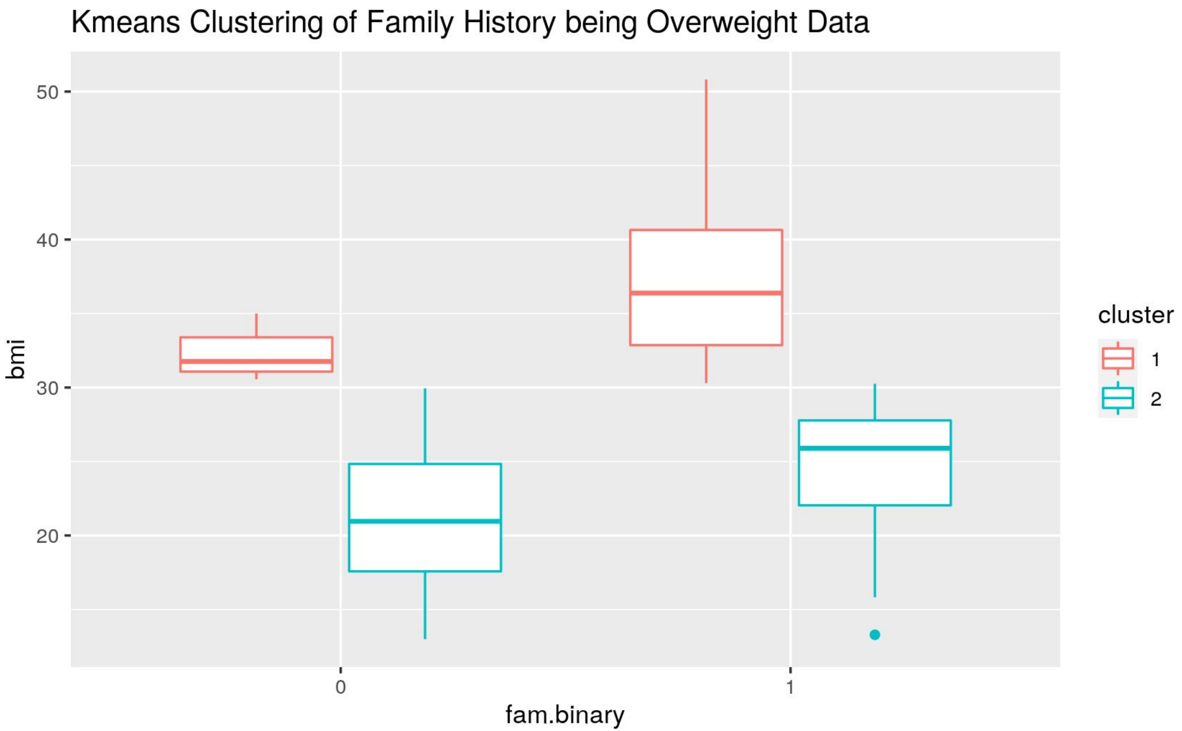
- It is claimed that having a family history of obesity results in a higher BMI and chance of being overweight.
- Variable: family_history_with_overweight





Hypothesis 3: Genetic Factors

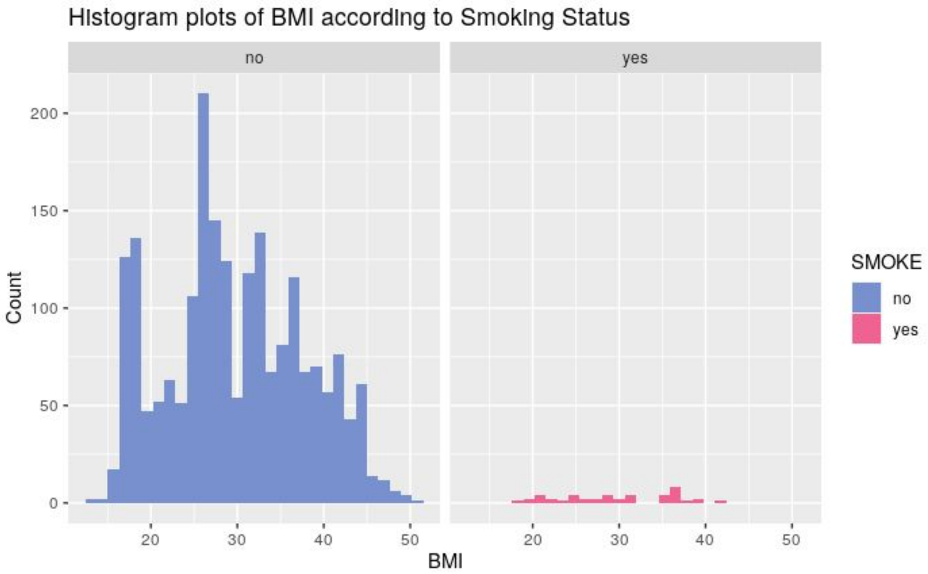
Below is the K-Means Clustering visualization that clearly show that family genes of being overweight does translate to higher BMI values to some extent.

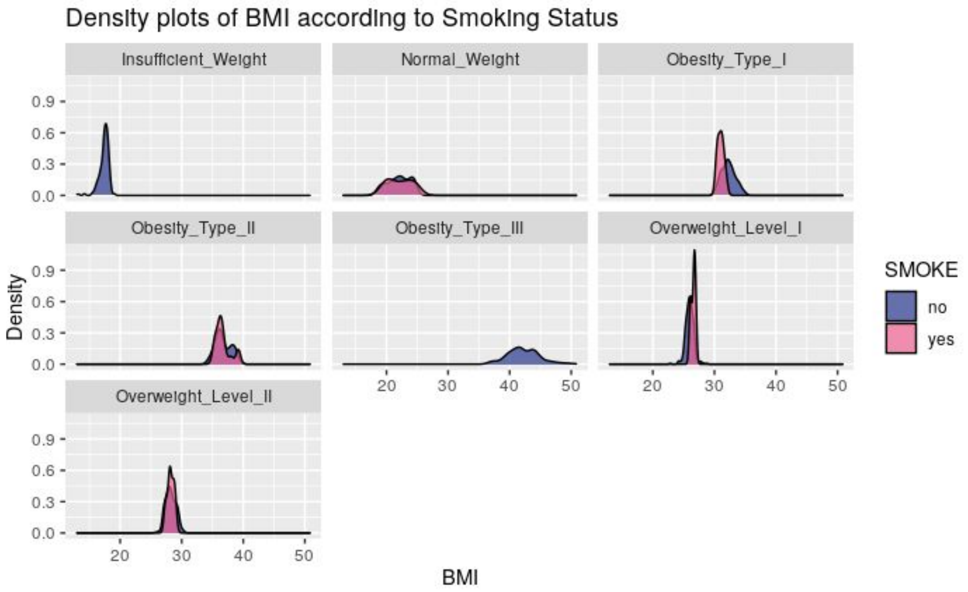


Hypothesis 4: Substance Use

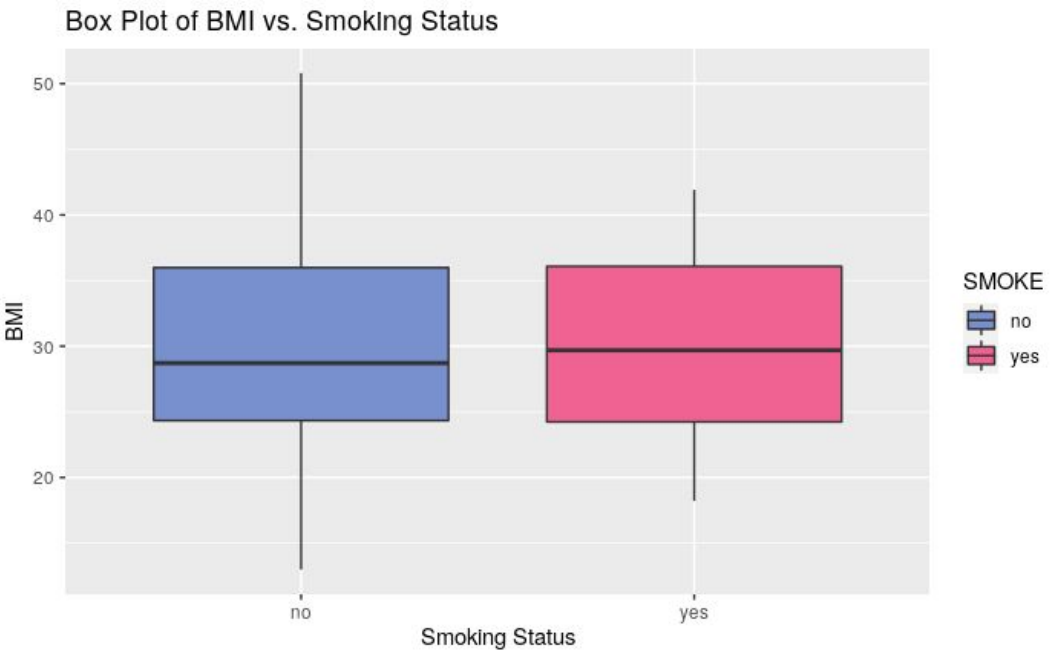
- Substance abuse is claimed to result in a higher BMI.
- Variables explored:
  - Smoking status (“SMOKE”)
  - Alcohol intake (“CALC”)
- Visualizations include histograms, density plots, and box plots
- K Means Clustering for further analysis

Hypothesis 4: Substance Use, Smoking





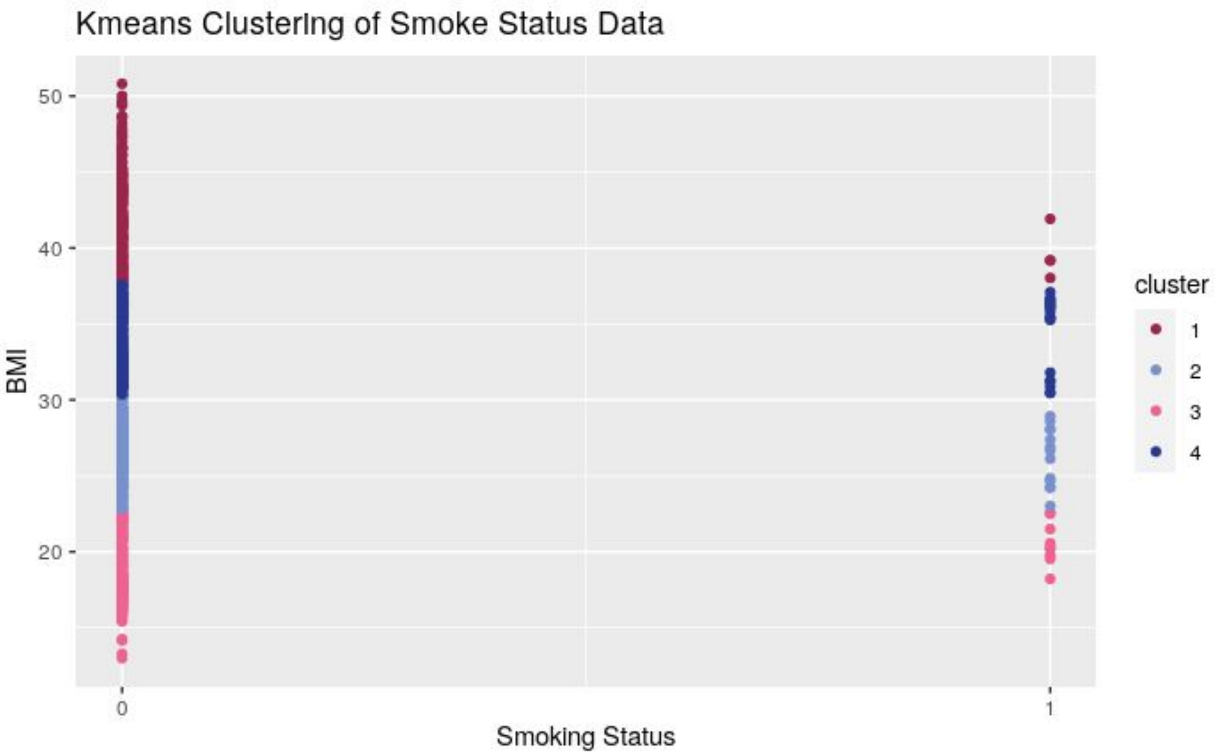
Hypothesis 4: Substance Use, Smoking

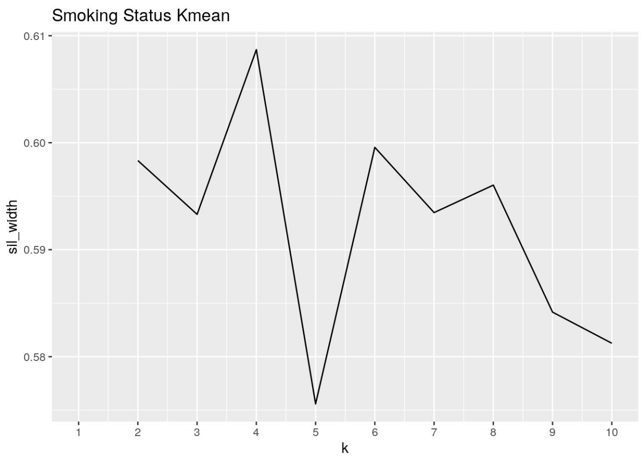


Summary Statistics on BMI for Smokers and Non-Smokers

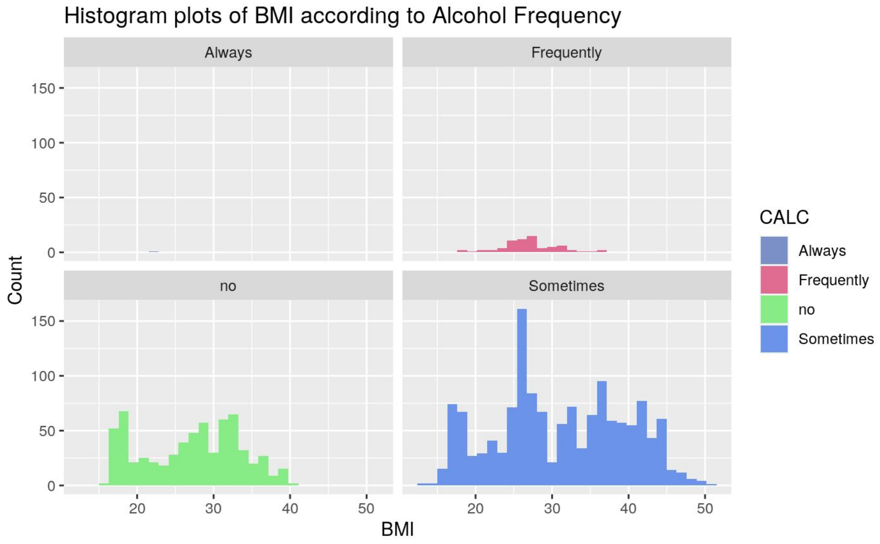
Smoking Status	*Does not smoke*	*Smokes*
n (count)	2067	44
Min BMI	13	18.2
Max BMI	50.8	41.5
Median BMI	28.7	29.7
Mean BMI	29.7	29.7
Standard Dev.	8.04	6.6

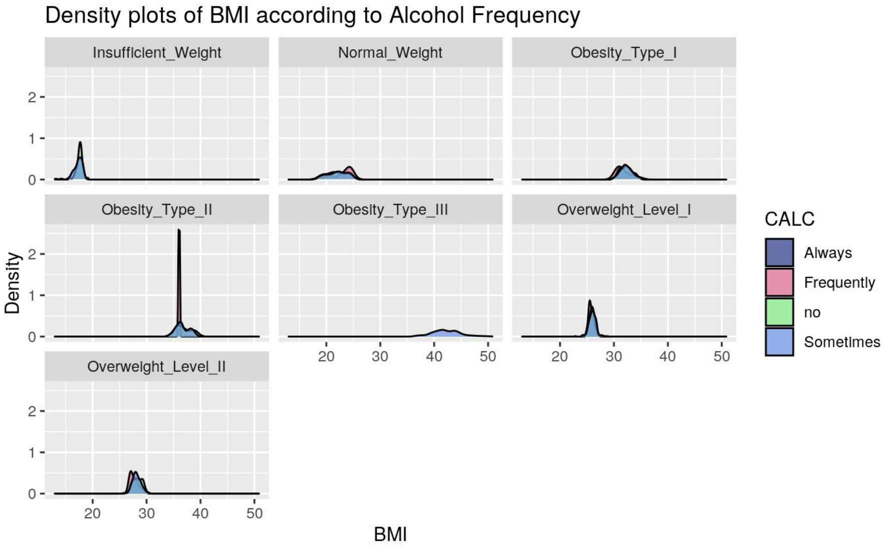
Hypothesis 4: Substance Use, Smoking



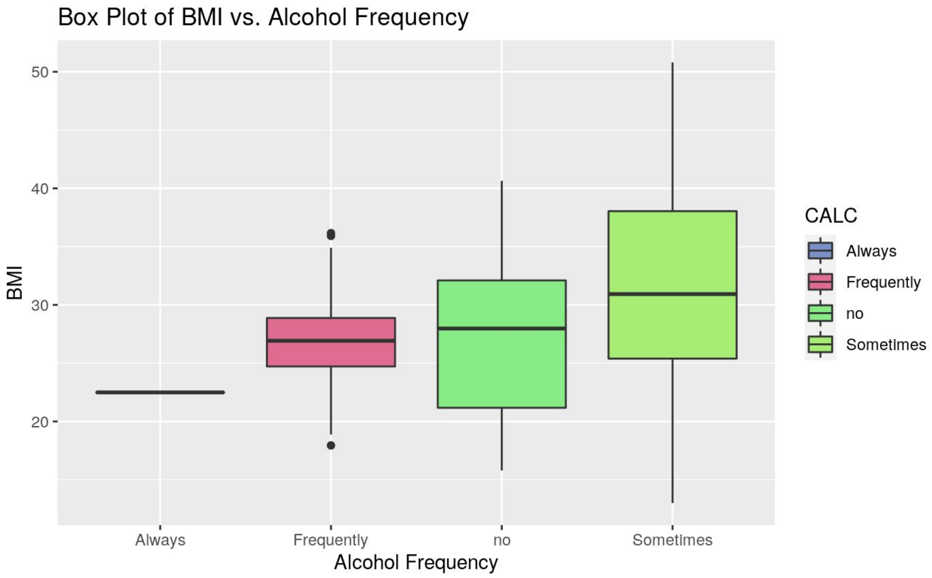


Hypothesis 4: Substance Use, Alcohol





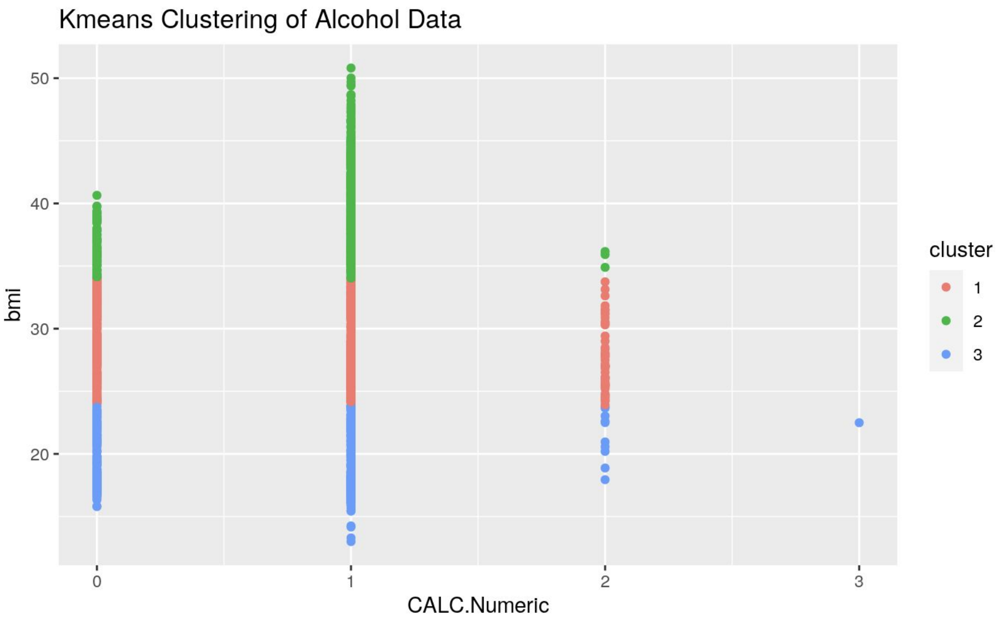
Hypothesis 4: Substance Use, Alcohol

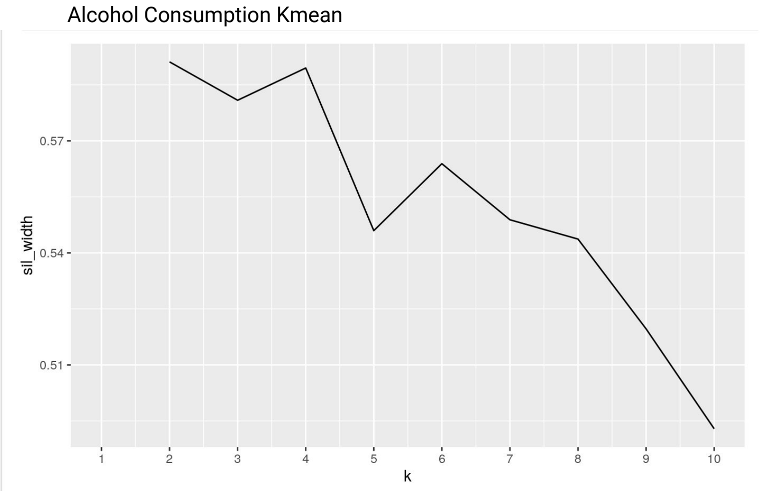


Summary Statistics on BMI and Alcohol Consumption Frequency

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	

Hypothesis 4: Substance Use, Alcohol





Set Up Prediction Task

1. Convert categorical variables to quantitative variables
   a. Such as gender, family history, weight category, etc.
2. Choose two different models
   a. SVM
   b. k-Nearest Neighbors
      i. Number of clusters = 7
3. Choose a binary variable to predict
   a. Above normal weight or not
   b. Above normal weight (BMI = 25 or above) was taken to be > 0

Variable	Possible Values	Converted Value
Gender	Female Male	1 0
Family History With Overweight	yes no	1 0
Weight Category	Insufficient-Weight Normal-Weight Overweight-Level-I Overweight-Level-II Obesity-Type-I Obesity-Type-II Obesity-Type-III	-1 0 1 2 3 4 5

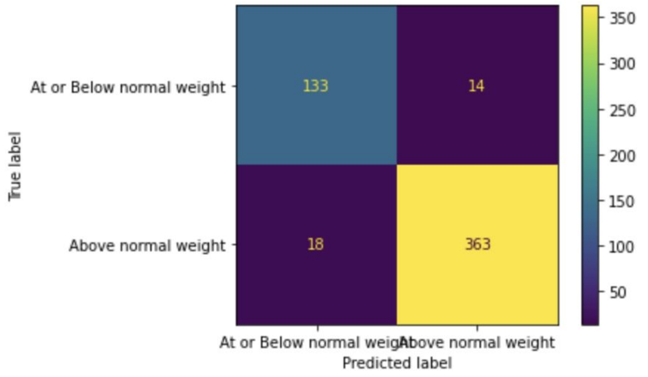
Evaluation of Models

SVM

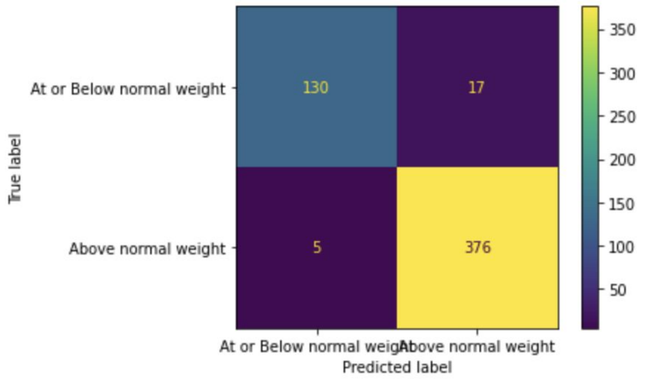
- Evaluated the accuracy using *accuracy_score*
- **Accuracy score for SVM: 0.9393**
    - Model predicted the correct label about 93% of the time
- Visualize accuracy with confusion matrix

k-Nearest Neighbors

- Evaluated the accuracy using *recall_score*
- **Recall score for k-Nearest: 0.9868**
    - The ability of the model to predict positive samples is very close to 1



Confusion matrix for SVM model



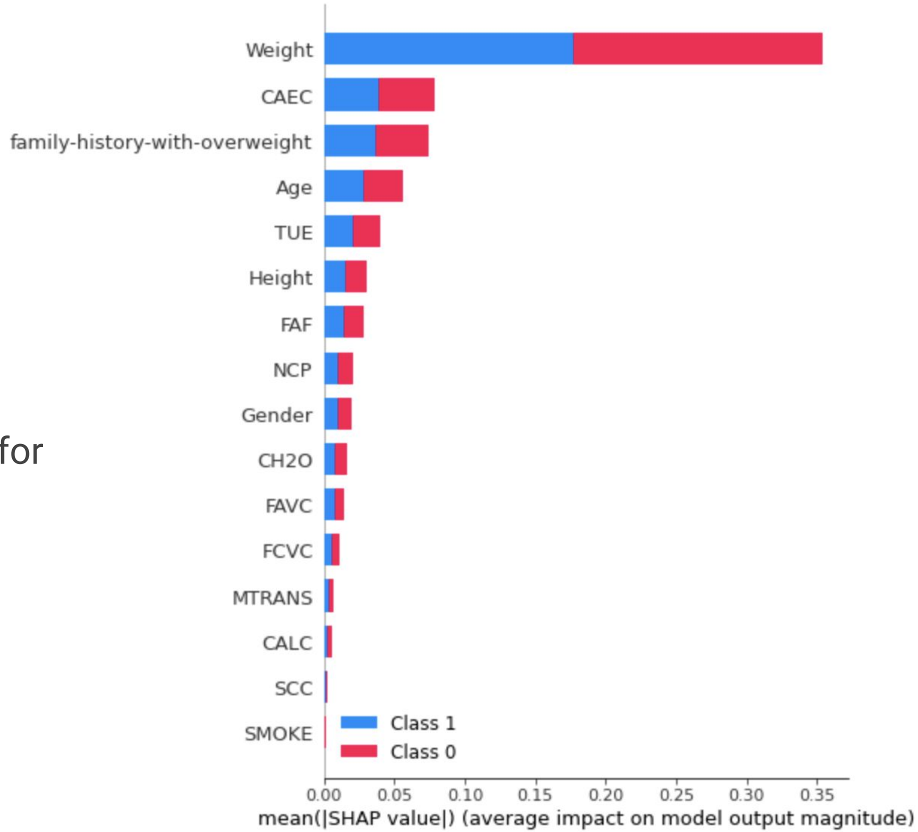
Confusion matrix for k-Nearest model

Important Features

1. Weight
2. CAEC
    a. Consumption of food between meals
3. Family history
4. Age

Observations

- Weight is the most significant feature for both classes
- Features seem to impact both classes equally
- Consuming food between meals was more significant than expected



Limitations

- 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.
- Non-US dataset so results can’t be applied here.
- We were limited in our tools, there are better analysis techniques to fit categorical data that we did not learn in class.

Conclusion and Next Steps

- The SVM model correctly predicted the weight classification for new observations 93% of the time
- The k-Nearest model had a recall score of 0.9868 meaning the model predicted positive new observations with a high degree of accuracy
- The most significant determinants of weight classification were weight, CAEC (consumption of food between meals), family history of overweight, and age
- Conclude: Our models suggest that these variables can be used to accurately predict the weight classification of a new observation