Estimation of the Obesity Levels Based on Eating Habits And Physical Condition.

By Adeyemi Joshua Olasore

1. INTRODUCTION

1.1 Background

The epidemic of obesity is nothing new. It has existed for centuries. The startling increase in frequency during the past few decades is the most recent development. Across the past half century, it has spread all over the globe. Countries that have been known for their low obesity rates are now seeing increases; China's rate of obesity rise has doubled in the last decade.

The CDC claims that this has an impact on a country's health, economy, and military preparedness. One in three adults is overweight, putting them at increased risk for cardiovascular disease, type 2 diabetes, and several cancers.

The problem of obesity, sadly, is not restricted to adults. In the United States, 17% of kids and teens are overweight. Despite this being far greater than in previous years, the rate does appear to be worsening. However, the risk for diabetes in children is increased by a factor of four, and children who are fat are more likely to remain obese into adulthood.

The changes in our environment and way of life have occurred at the same time that obesity rates have risen. Many of our once-manual labours are now performed by machines, our diets and the methods by which they are produced have evolved, and our communities are organized to favour the private automobile over other forms of transportation.

Obesity has become a worldwide epidemic, and if we want to stop it, we need to find ways to treat its underlying causes. Predicting the possibility of obesity based on some accessible and quantitative parameters is intriguing to me as an aspiring data scientist.

1.2 Research Questions

Furthermore, in order to arrive a sizable conclusion, this research would like to answer the following questions;

- What is the relationship between various eating habits and Obesity?
- What is the relationship between various physical conditions and Obesity?
- What are the significant differences in obesity observed among gender, smokers, connection to family history, etc.?
- Can Obesity in terms of Body Mass Index (BMI) be predicted given selected parameters?

1.3 Objectives

1. To determine and visualize the relationship between eating habit, physical condition of people and their corresponding body mass index.
2. To evaluate the effect of the categorical parameters on the body mass index using the different correlation tests.
3. To determine whether the obtained numerical and categorical parameters predicts obesity in terms of body mass index using regression model and classification algorithms.

2.0 METHODOLOGY

2.1 Data Collection and Source

The data gives the estimation of obesity levels in people from the countries of Mexico, Peru and Colombia, with ages between 14 and 61 and diverse eating habits and physical condition. According to UCI, the data was primarily collected using a web platform with a survey where anonymous users answered each question, then the information was processed obtaining 17 attributes and 2111 records. These raw data and related information was obtained from the Center for Machine Learning and Intelligent System, University of California, Irvine (UCI) Machine Learning Repository.

2.2 Data Introduction

The dataset includes eating habits attributes such as: Frequent consumption of high caloric food (FAVC), Frequency of consumption of vegetables (FCVC), Number of main meals (NCP), Consumption of food between meals (CAEC), Consumption of water daily (CH20), and Consumption of alcohol (CALC), as well as physical condition attributes such as: Calories consumption monitoring (SCC), Physical activity frequency (FAF), Time using technology (TUE), Means of Transportation (MTRANS). Furthermore, based on Equation (1) and information from WHO and Mexican Normativity, the class variable NObeyesdad was established with the values: Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II, and Obesity Type III. Because the data comprises both numerical and continuous data, it may be analyzed using classification and prediction algorithms.

2.3 Data Preparation and Preprocessing

The data was imported into the IDE using the ***numpy*** and ***pandas*** library for the needed wrangling and cleaning. Although basic preprocessing has been primarily done on the dataset, there was need to further check for some discrepancies.

Some of the task involved checking for missing data (***is.null()***), confirming of the data types (***.info()***), removal of unessential variables (***.drop()***), renaming of column names of attributes (***.rename()***), confirming the values of the categorical variables (***.unique()***), the determination of the Body Mass Index (BMI) which is the numerical dependent variable needed for the analysis, and finally evaluating the descriptive statistical summary for the measure of central tendency and visible outliers (***.describe()***).

2.4. Data Analysis tools

2.4.1 Exploratory Data Analysis

In order to execute full exploratory analysis of the dataset the ***matplotlib.pyplot*** and ***seaborn*** libraries were imported.

Univariate analysis were carried out using the frequency distribution plots, **histplot, barplot and countplots**. This helps to get summary of selected categorical variables and check for patterns in the data.

Bivariate analysis were as well carried out using the **boxplot, jointplot, regplot, and lineplot**. These helped in understanding the form, strength and dependence of the relationship between two selected variables.

Multivariate Analysis was done by; using the **FacetGrid** to visualize the distribution of multiple variables , and generating the correlation among the numerical and categorical (using the ***pivot table*** function) variables and visualized on the **heatmap and clustermap**.

2.4.2 Normality Test

The normality of the dependent variable was checked whether the sample data was drawn from a normally distributed population. This was actualized by importing the ***statsmodels*** library and performing the function ***qqplot*** as well as the ***Shapiro*** from the ***scipy*** library.

2.4.3 Correlation Test

The linear correlation between the two numerical variables. Basically, the test was done between each independent numerical variable and the dependent variable (BMI). This helped to measure the strength and direction of the relationship between them. The pearsonr and spearmanr function from the scipy.stats library was used to execute this.

2.4.4. T -Test

This statistical test were performed to carry out hypothesis testing and compare the mean of two-grouped categorical variables (Gender, family history, FAVC, SMOKE, SCC). This was executed using the ***researchpy*** library.

2.4.5 Anova Test

For categorical variable with more than two groups, the anova test was performed to check if there is a significant difference in the mean values of the groups . This was done for the variables CAEC, CALC and MTRANS using the ***statsmodels.api*** library, followed by executing the functions ***ols*** and ***.stats.anova_lm.*** The corresponding pvalues and F values were obtained.

2.5 Machine Learning

Based on the data given, there are both categorical and numerical attributes which would permit to perform both regression and classification algorithms to predict the output of the dependent variables.

2.5.1 Regression Algorithm

Regression model was generated by making the machine to learn the relationship between the numerical variable against the output variables. The following steps were taken, namely;

1. Splitting of the data into the X and Y variables.
2. Splitting the variables into the train and test variable using the ***train_test_split*** function from the ***sklearn.model_selection*** library.
3. Instantiating the model using the ***LinearRegression*** function from the ***sklearn.linear_model*** library
4. Fitting the model using the ***LinearRegression.fit*** function.
5. Generating the predictions by using the ***LinearRegression.predict*** function
6. Calculating the residuals by subtracting the predictions from the tested dependent variable.
7. Plotting the distribution of the residual to check the uniform distribution.
8. Evaluating the metrics of the model suing the ***metrics*** function from the ***sklearn*** library
9. Generating the coefficients and the other values of the model.

2.5.2 Classification Algorithm

The output variable is binned into classes based on the value of the Body Mass Index (BMI) obtained and this is being used as the categorical output. This would help to predict the probability of occurrence using a binary form.

The following steps were taken, namely;

1. The dataset was duplicated to avoid complications in computing using the .copy() function
2. The numerical variables in the dataset were removed using the .drop() function
3. The independent categorical variables were turn to binary by using the get_dummies function from the pandas library, whereas the dependent variable, being more than two classes, were turned to integers first using the LabelEncoder function from the sklearn.preprocessing library, followed by turning the integers into binaries using the OneHotEncoder function from the same sklearn.preprocessing library.
4. Splitting of the data into the X and Y variables. The binaries of the independent variable were concatenated in dataframe together. Each of the classes of the dependent variable were indexed and trained and tested.
5. Splitting the variables into the train and test variable using the ***train_test_split*** function from the ***sklearn.model_selection*** library. Each of the classes of the output variables were trained and tested.

6. Instantiating the model using the ***LogisticRegression*** function from the ***sklearn.linear_model*** library
7. Fitting the model using the ***LogisticRegression.fit*** function for each of the classes.
8. Generating the predictions by using the ***LogiscticRegression.predict*** function
9. Evaluating the tests of the model using the ***classification_report*** function from the ***sklearn.meterics*** library
10. Generating the ***confusion_matrix*** for the tested dependent variables and the derived prediction.
11. Visualized into heatmap to analyse the True Positive, True Negative, False Positive and the False Negative.

3.0 ANALYSIS AND RESULTS PRESENTATION

3.1 Descriptive Statistical Summary

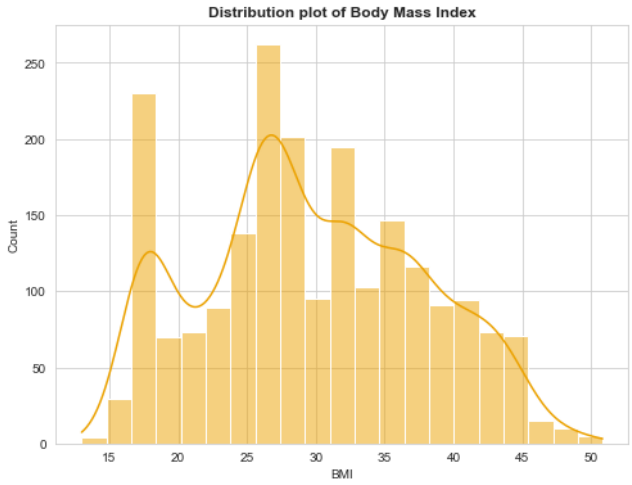
Table 1; The Descriptive summary of all numerical variables

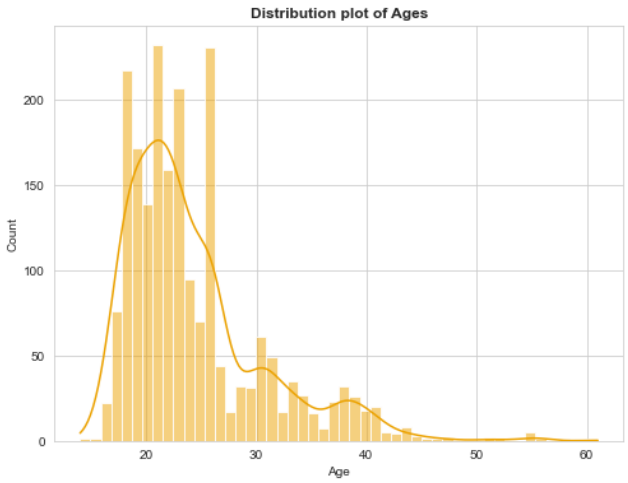
	Age	Height	Weight	FCVC	NCP	CH2O	FAF	TUE	BMI
count	2111	2111	2111	2111	2111	2111	2111	2111	2111
mean	24.31	1.70	86.59	2.42	2.69	2.01	1.01	0.66	29.70
std	6.35	0.09	26.19	0.53	0.78	0.61	0.85	0.61	8.01
min	14.00	1.45	39.00	1.00	1.00	1.00	0.00	0.00	13.00
25%	19.95	1.63	65.47	2.00	2.66	1.58	0.12	0.00	24.33
50%	22.78	1.70	83.00	2.39	3.00	2.00	1.00	0.63	28.72
75%	26.00	1.77	107.43	3.00	3.00	2.48	1.67	1.00	36.02
max	61.00	1.98	173.00	3.00	4.00	3.00	3.00	2.00	50.81

The table gave the measure of central tendency and measure of dispersion using the mean value, interquartile ranges and standard deviation.

3.2 Exploratory Data Analysis

3.2.1 Univariate Data Analysis





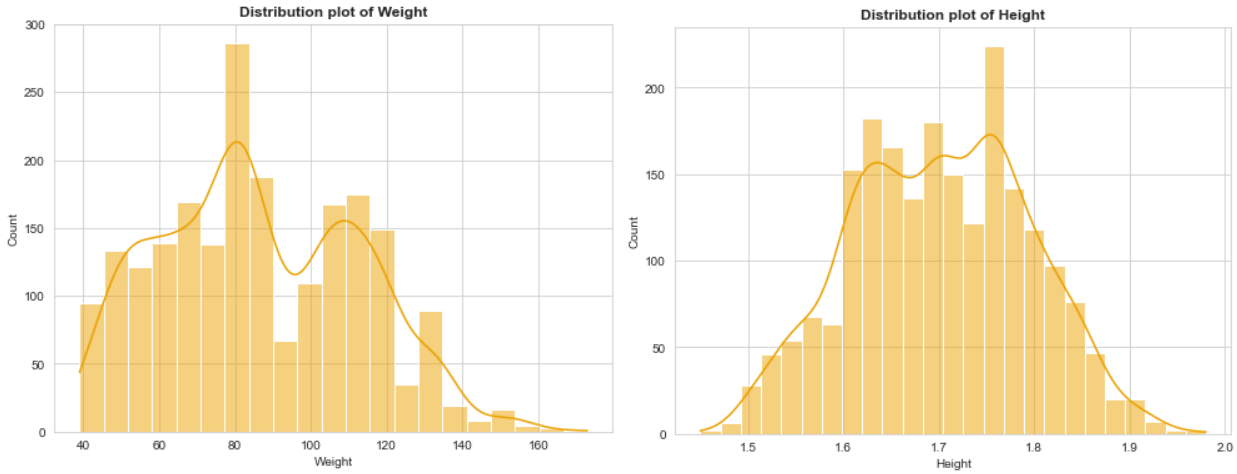


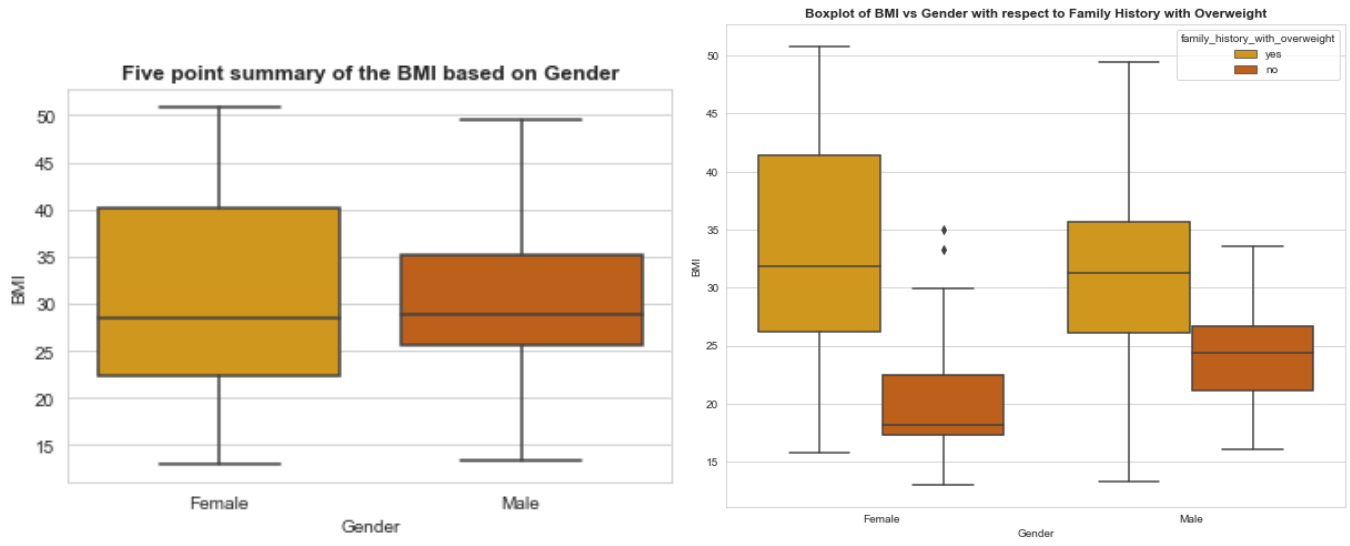
Figure 2; The distribution plot of BMI, Ages, Weight and the Height.

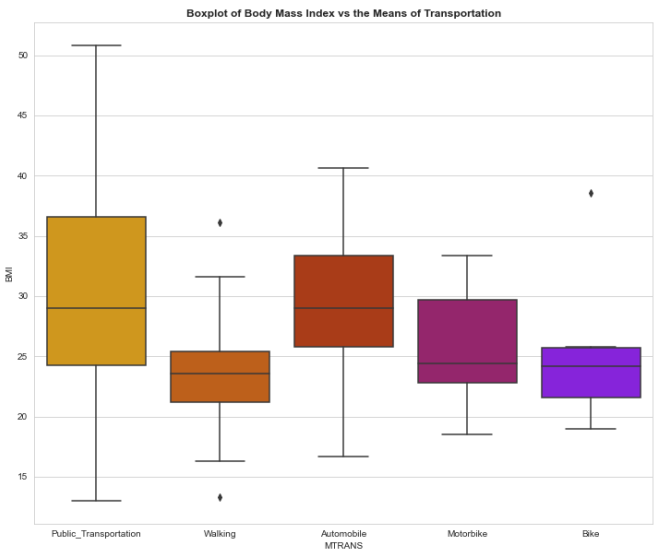
As observed from the BMI plot, although the data is widely spread along the axis, the most index is at 26 and 17, having it's mean value at 29. The majority of the index at about 95% is concentrated at the index ranging from 20 – 45. The plot is slightly skewed to the right and there does not seem to be any outlier.

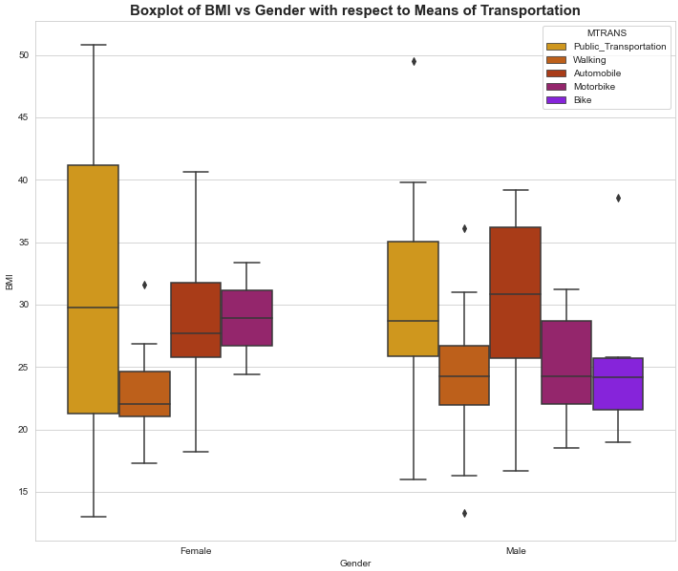
As observed from the Age plot, although the data is not spread along the axis but concentrated at the age range of 14 - 25, having it's mean value at 24. The plot is skewed to the right.

As observed from the Weight plot, the data is widely spread along the axis, the least count weight are observed at 140kg to 173kg. The majority of the weights at about 90% is concentrated at the index ranging from 40kg to 120kg. The plot is slightly skewed to the right and there does not seem to be any outlier. The Height data is evenly distributed with a slight imbalance and most height at 1.6m to 1.8m

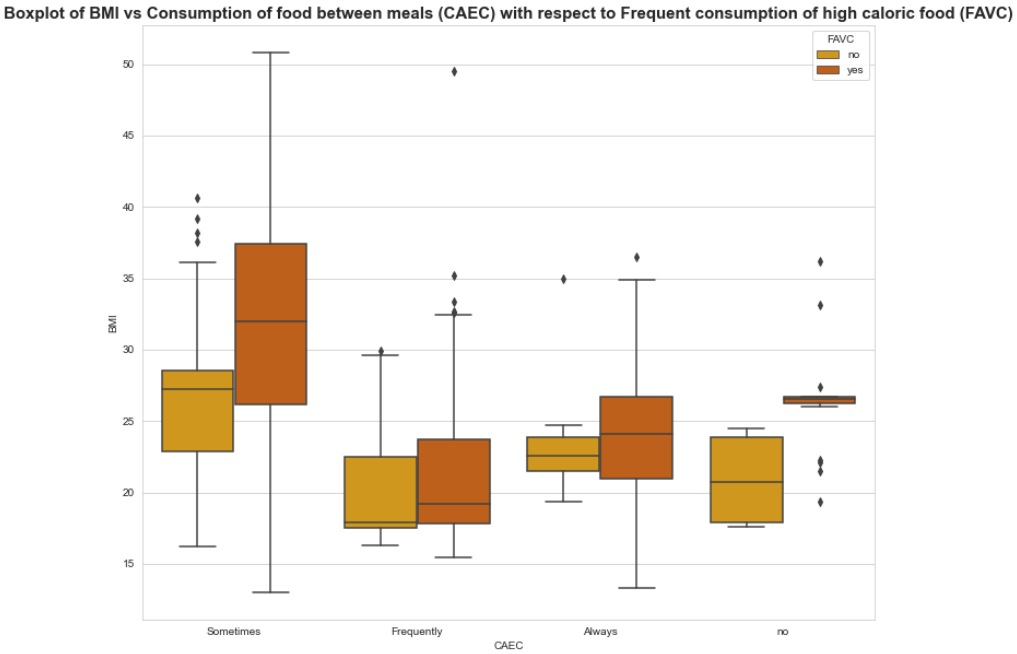
3.2.2 Bivariate Data Analysis



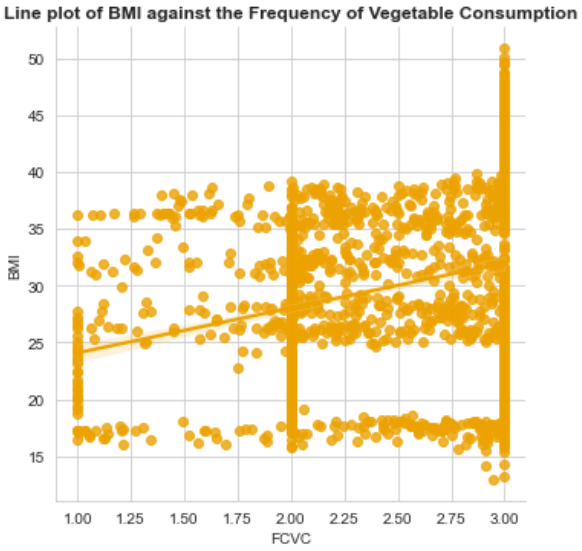


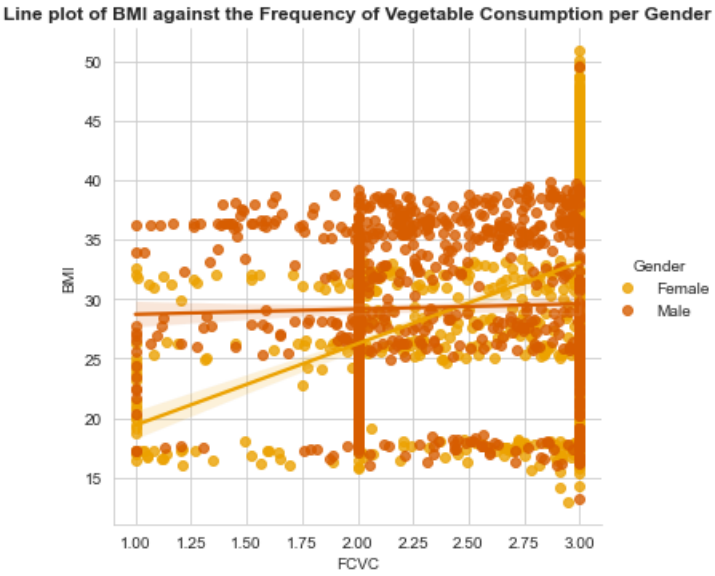


As observed from the plots above, there is higher body mass index in female than in male, but both gender has higher BMI as result of overweight in the family history. There is higher body mass index for those that opted for automobile and public transportation as means of transportation, which can be as a result of less physical activity done relating to exercise.

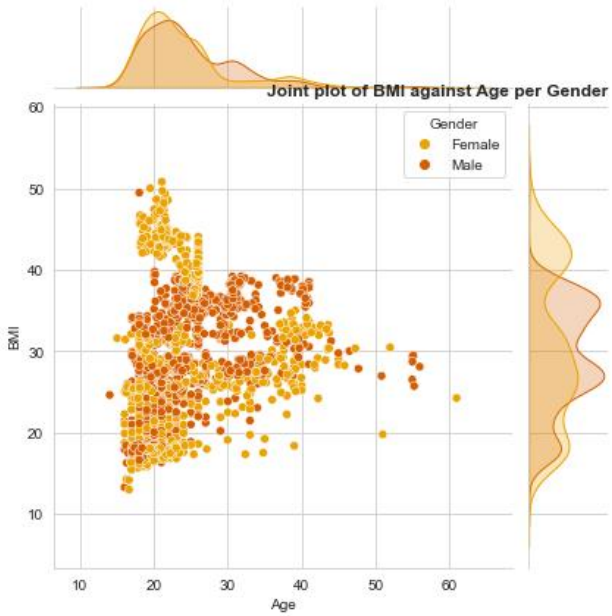


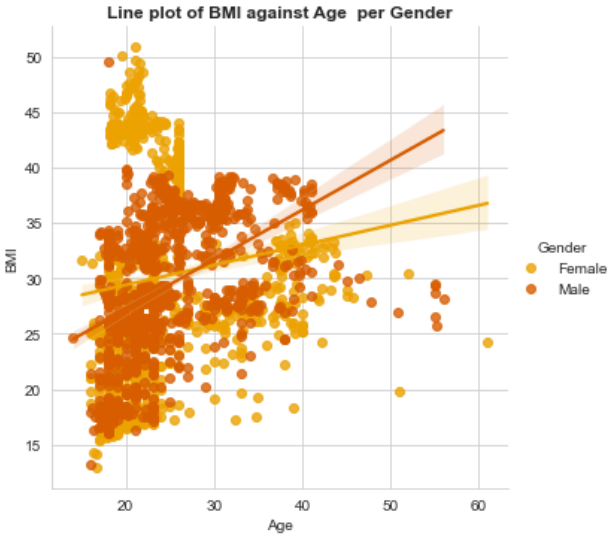
The figure above explains the relationship between the frequency in consuming high caloric food and also eating between meals. It can be seen that those that eat high caloric food have higher BMI. There is higher BMI for those that consume high caloric food seldomly than those that do always. This can be a result of the activities they are involved in and how often the monitor their calories level.



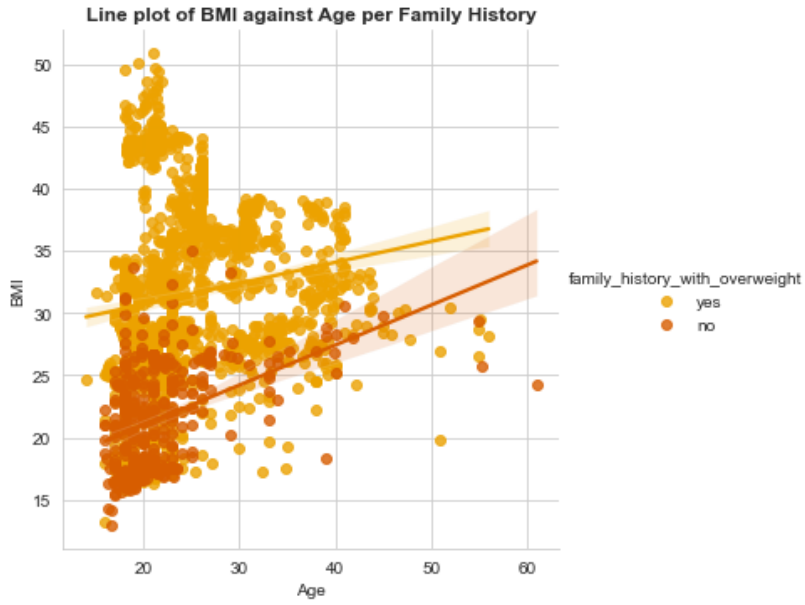


The line plot above explains the relationship between the frequency of vegetable consumption and BMI, this is further compared between genders. It can be observed that an increase in the consumption result in an increase in the BMI and this higher in Females than in Male gender.



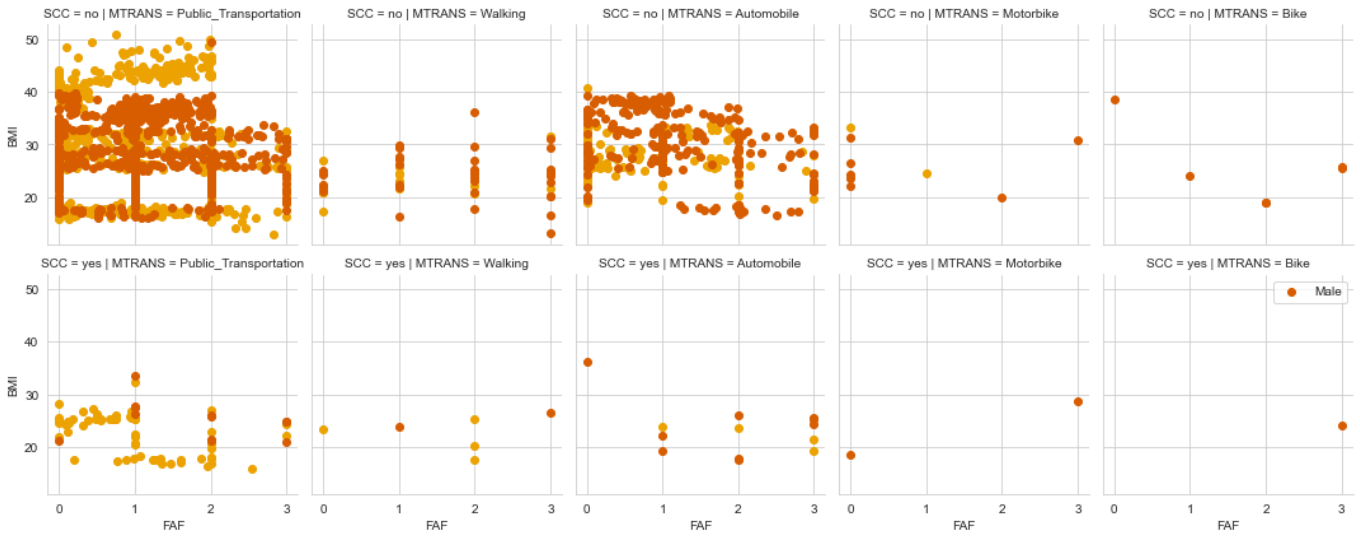


In terms of Age and BMI, there is also an higher BMI in Female gender than in Male, Although both gave positive linear relationship to the BMI.

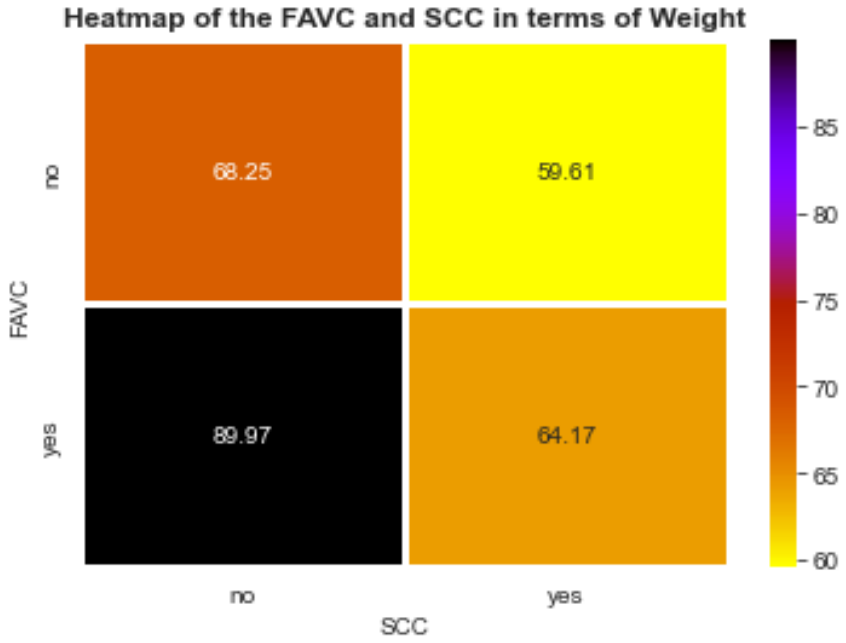


And as seen on the plot, the family with an history of overweight gave an higher value of BMI but there seem to be lower rate in the increase compared to the increase in age for families with history of overweight. This can be explained as a factor of consciousness of their history and hence a close monitoring was done on their consumptions and activity.

3.2.3 Multivariate Data Analysis.



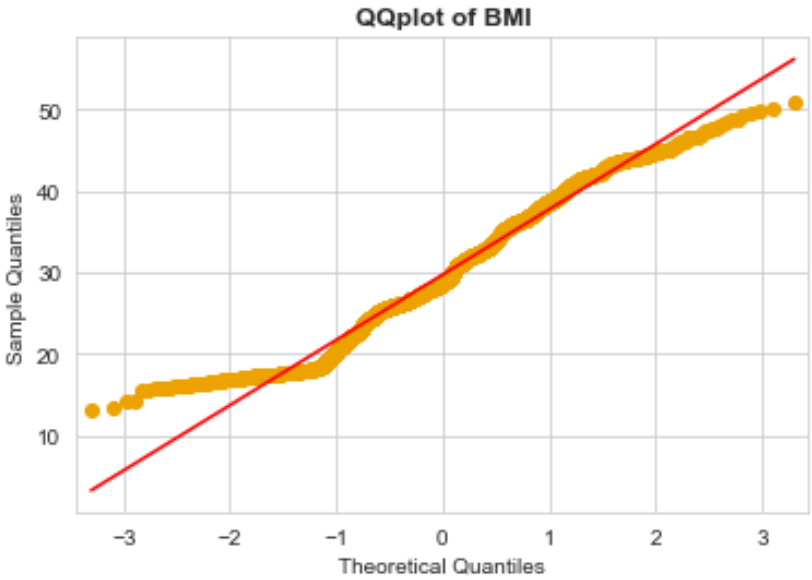
As previously highlighted concerning those opted for public transport and automobile, and how they have higher body mass index. It is interesting to check the extent of the BMI if factors such as regular physical activity and monitoring of caloric consumption is included. It can be observed that with an increase in physical activity and constant monitoring of consumption, there is relatively lower BMI.



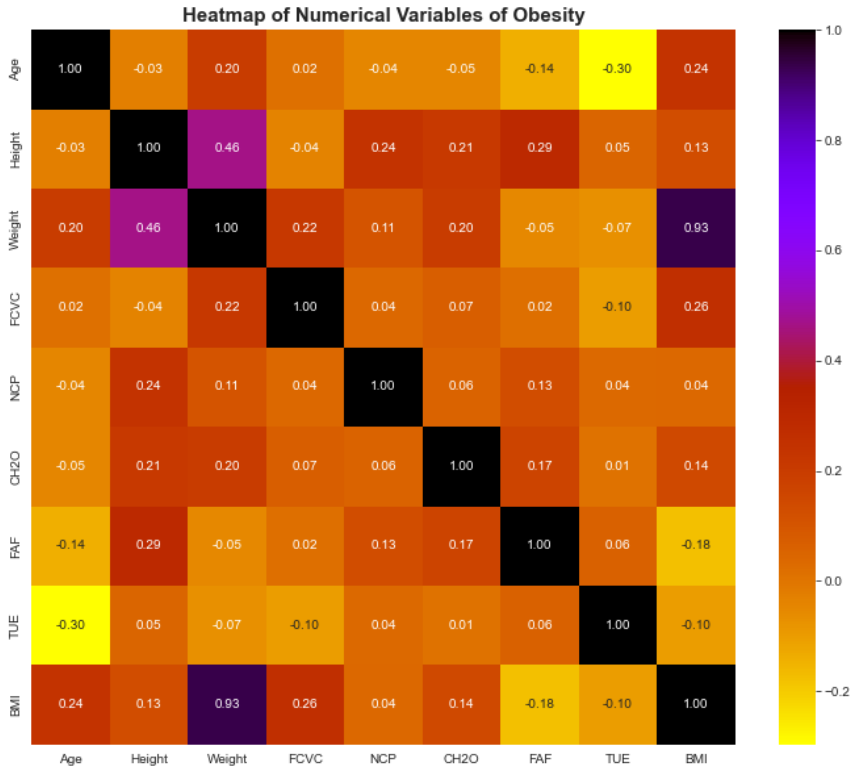
There is an obvious increase in the BMI for those that consumer vegetables frequently and no monitoring of caloric consumption as seen on the heatmap generated.

3.3 Normality Test

The normality test showed that the BMI is not normally distributed as it deviates from the center of linearity using the qqplot.



3.4 Correlation Test



The heatmap of the correlation showed a weak correlation between the all variables except the BMI and Weight which strongly correlates to each other.

3.5 T -Test

a. BMI and Gender T-test

Ho: There is no significant difference between the BMI of the Male vs the Female

H1: There is significant difference between the BMI of the Male vs the Female

Variable	N	Mean	SD	SE	95% Conf.	Interval
Male_BMI	1068	29.28	6.35	0.19	28.90	29.66
Female_BMI	1043	30.13	9.40	0.29	29.56	30.70
combined	2111	29.70	8.01	0.17	29.36	30.04

Independent t-test	results
Difference (Male_BMI - Female_BMI) =	-0.850
Degrees of freedom =	2109.000
t =	-2.439
Two side test p value =	0.015

Difference < 0 p value =	0.007
Difference > 0 p value =	0.993
Cohen's d =	-0.106
Hedge's g =	-0.106
Glass's delta1 =	-0.134
Point-Biserial r =	-0.053

There is a significant difference between the mean as the pvalue is less than 0.05, hence we reject the null hypothesis and accept the alternative hypothesis.

b. BMI and Family History T-test

Ho: There is no significant difference between the BMI of the Family History of Overweight vs the No Family History with Overweight

H1: There is significant difference between the BMI of the Family History of Overweight vs the No Family History with Overweight

Variable	**N**	**Mean**	**SD**	**SE**	**95% Conf.**	**Interval**
FHistory_BMI	1726	31.53	7.50	0.18	31.18	31.88
NFHistory_BMI	385	21.50	4.21	0.21	21.08	21.92
combined	2111	29.70	8.01	0.17	29.36	30.04

Independent t-test	**results**
Difference (FHistory_BMI - NFHistory_BMI) =	10.029
Degrees of freedom =	2109
t =	25.367
Two side test p value =	0.000
Difference < 0 p value =	1.000
Difference > 0 p value =	0.000
Cohen's d =	1.430
Hedge's g =	1.429
Glass's delta1 =	1.338
Point-Biserial r =	0.484

There is a significant difference between the mean as the pvalue is less than 0.05, hence we reject the null hypothesis and accept the alternative hypothesis.

c. BMI and SMOKE T-test

Ho: There is no significant difference between the BMI of the Smokers vs Non-smokers

H1: There is significant difference between the BMI of the Smokers vs Non-smokers

Variable	N	Mean	SD	SE	95% Conf.	Interval
Smokers_BMI	44	29.66	6.60	0.99	27.65	31.66
NonSmokers_BMI	2067	29.70	8.04	0.18	29.35	30.05
combined	2111	29.70	8.01	0.17	29.36	30.04

Independent t-test	results
Difference (Smokers_BMI - NonSmokers_BMI) =	-0.046
Degrees of freedom =	2109
t =	-0.038
Two side test p value =	0.970
Difference < 0 p value =	0.485
Difference > 0 p value =	0.515
Cohen's d =	-0.006
Hedge's g =	-0.006
Glass's delta1 =	-0.007
Point-Biserial r =	-0.001

There is no significant difference between the mean as the p value is greater than 0.05, hence we accept the null hypothesis and reject the alternative hypothesis.

d. BMI and SCC -Calories Monitoring T-test

Ho: There is no significant difference between the BMI of those that do not monitor their calories consumption vs monitor their calories consumption

H1: There is significant difference between the BMI of those that do not monitor their calories consumption vs monitor their calories consumption

Variable	N	Mean	SD	SE	95% Conf.	Interval
noSCC_BMI	2015	30.02	8.01	0.18	29.67	30.37
SCC_BMI	96	22.94	4.04	0.41	22.12	23.76
combined	2111	29.70	8.01	0.17	29.36	30.04

Independent t-test	results
Difference (noSCC_BMI - SCC_BMI) =	7.085
Degrees of freedom =	2109
t =	8.611
Two side test p value =	0.000
Difference < 0 p value =	1.000
Difference > 0 p value =	0.000
Cohen's d =	0.900
Hedge's g =	0.899
Glass's delta1 =	0.884
Point-Biserial r =	0.184

There is a significant difference between the mean as the pvalue is less than 0.05, hence we reject the null hypothesis and accept the alternative hypothesis.

e. BMI and FAVC T-test

Ho: There is no significant difference between the BMI of Frequency of High Calories Consumption and No Frequency of High Calories Cosumption

H1: There is significant difference between the BMI of Frequency of High Calories Consumption and No Frequency of High Calories Cosumption

Variable	N	Mean	SD	SE	95% Conf.	Interval
FAVC_BMI	1866	30.41	8.05	0.19	30.05	30.78
noFAVC_BMI	245	24.26	5.08	0.32	23.62	24.90
combined	2111	29.70	8.01	0.17	29.36	30.04

Independent t-test	results
Difference (FAVC_BMI - noFAVC_BMI) =	6.154
Degrees of freedom =	2109
t =	11.660
Two side test p value =	0.000
Difference < 0 p value =	1.000
Difference > 0 p value =	0.000
Cohen's d =	0.792
Hedge's g =	0.792
Glass's delta1 =	0.764
Point-Biserial r =	0.246

There is a significant difference between the mean as the pvalue is less than 0.05, hence we reject the null hypothesis and accept the alternative hypothesis.

3.6 Anova Test

CAEC

Ho:$\mu1=\mu2=\mu3=\mu4$

$H1$:$\mu1\neq\mu2\neq\mu3\neq\mu4$

	df	sum_sq	mean_sq	F	PR(>F)
C(CAEC)	3	25125.43	8375.14	159.99	2.02E-93
Residual	2107	110297.57	52.35		

There is a significant difference between the mean as the pvalue is less than 0.05, hence we reject the null hypothesis and accept the alternative hypothesis.

CALC

Ho:$\mu1=\mu2=\mu3=\mu4$

$H1$:$\mu1\neq\mu2\neq\mu3\neq\mu4$

	df	sum_sq	mean_sq	F	PR(>F)
C(CALC)	3	7537.67	2512.56	41.40	5.47E-26
Residual	2107	127885.33	60.70		

There is a significant difference between the mean as the pvalue is less than 0.05, hence we reject the null hypothesis and accept the alternative hypothesis.

MTRANS

Ho:$\mu1=\mu2=\mu3=\mu4=\mu5$

$H1$:$\mu1\neq\mu2\neq\mu3\neq\mu4\neq\mu5$

	df	sum_sq	mean_sq	F	PR(>F)
C(MTRANS)	4	2740.645	685.1613	10.87522	9.97E-09
Residual	2106	132682.3	63.00207		

There is a significant difference between the mean as the pvalue is less than 0.05, hence we reject the null hypothesis and accept the alternative hypothesis.

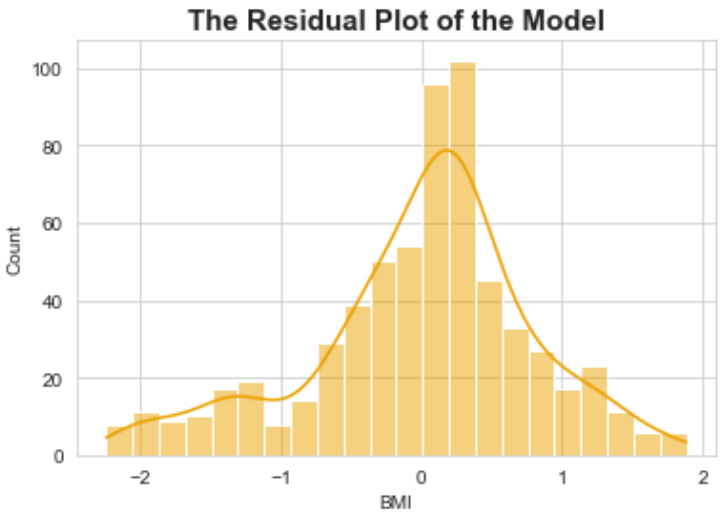
Although there is less significance in MTRANS and CALC compared to the CAEC, judging from the value of the Fvalue.

3.7 Machine Learning

3.7.1 Regression Algorithm

The Regression Model

Y = a + b1X1 + b2X2 + b3X3 + b4X4 + b5X5 + b6X6 + b7X7 ..... + e



Error Terms of the Model

MAE: 0.5973352559387414

MSE: 0.6302154703430548

RMSE: 0.7728746702659761

The RMSE explains the higher accuracy of the model as there is lower value.

Model Coefficients.

	Coefficients
Age	0.014
Height	-32.514
Weight	0.339
FCVC	0.121
NCP	0.139
CH2O	0.058
FAF	-0.103
TUE	0.034

Intercept

54.67599570066156

Model Equation

BMI = 54.67 + 0.014.Age – 32.52.Height + 0.339.Weight + 0.121.FCVC + 0.139.NCP + 0.058.CH2O -0.103.FAF + 0.034.TUE

4.0 CONCLUSION

Body Mass Index (BMI) is a measure of body fat based on height and weight that is used to classify individuals as underweight, normal weight, overweight, or obese. It is calculated by dividing an individual's weight in kilograms by their height in meters squared. While BMI can be a useful tool for identifying trends in population-level data, it has some limitations as a measure of individual body fatness.

One limitation is that BMI does not take into account differences in muscle mass, bone density, and distribution of fat. For example, an athlete with a high level of muscle mass may have a high BMI but may not actually be overweight or obese. Similarly, an older adult with a lower muscle mass may have a lower BMI but may still have a high level of body fat.

Another limitation is that BMI does not account for differences in body shape or the distribution of fat on the body. Some people may have a higher percentage of abdominal fat, which is associated with an increased risk of health problems, even if their BMI is within the normal range.

Overall, BMI can be a useful measure for identifying trends in population-level data and for identifying individuals who may be at risk for health problems due to being overweight or obese. However, it is important to consider other factors, such as muscle mass, bone density, and distribution of fat, when assessing an individual's health and risk of developing related health problems.

According to the analyzed data, there is a high correlation between the family history of obesity and the body mass index, hence an important factor in the determination of the level of obesity. There are limited information from the given data and some of the values given are based on frequency which could a bias in the conclusion derived and the generated model.