UNIVERSITY OF
SOUTH DAKOTA

Semester Project Report

Riti Dahal

Course Code : DSCI 505

Course Name : Business Analytics Fundamentals

Instructor : Chad Birger

Business Problem

Identifying an obesity level can be difficult to measure as it requires various factors such as age, weight, family history, calorie intake, physical activities, etc. For instance, if an individual follows a proper diet plan and is obese; obesity data can help dietitians identify problems based on their calorie intake and physical activity. This will help the dietician to update an individual's diet plan. Likewise, if an individual has high blood pressure and is obese, dietitians can identify the family history and provide a personalized diet plan accordingly. Be it complex or simple data, Python can be used by businesses for data cleaning, data transformation, and data visualization. Thus, python is one of the easiest tools for data analysis that can be used to enhance business operations.

Dataset Description

Data set:
https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition

a. How many variables are in the data set?

There are 17 variables in the data set which are listed below:

Gender: Number of Male and Female

Age: Provides individual's age

Height: Provides individual's height (in meters)

Weight: Provides individual's weight (in kilograms)

Family_history_with_overweight: Provides an individual's family history suffered from being overweight.

FCVC: If an individual eats caloric food frequently

FAVC: If an individual includes vegetables in their daily meals

NCP: Number of meals

CAEC: Shows a food intake of an individuals between meals

Smoke: Shows if an individual smokes or not

CH2O: Individuals water intake in a daily basis

SCC: Shows if an individual monitor their calorie intake

FAF: Shows an individual's physical activity

TUE: Shows the use of technological devices by an individual

CALC: Shows if an individual drink alcohol or not

MTRANS: Shows the mode of transportation

NObeysedad: Shows the obesity level, which is categorized as Insufficient Weight, Normal Weight, Obesity Type I, Obesity Type II, Obesity Type III, Overweight Level I, Overweight Level II.

b. Data types and levels of measurement for each of the variables

The data set has five categorical, six continuous, three binary, and two Integers.

The table below shows the data types and level of measurement for each of the variables:

Variables	Data Type	Level of Measurement
Gender	Object	Nominal
Age	Float	Ratio
Height	Float	Ratio
Weight	Float	Ratio
Family History With Overweight	Object	Nominal
FAVC	Object	Nominal
FCVC	Float	Ordinal
NCP	Float	Ratio
CAEC	Object	Nominal
SMOKE	Object	Nominal
CH2O	Float	Ratio
SCC	Object	Nominal
FAF	Float	Ratio
TUE	Float	Ordinal
CALC	Object	Nominal
MTRANS	Object	Nominal
Nobeyesdad	Int	Ordinal

Table: Showing the data types and levels of measurement for each variables

c. How many observations are in the data set?

There are 2111 number of observation for each variable.

d. Are there any missing values or outliers in the data set?

There are no missing values in the data set. However, outliers were found for Age, Weight and Height using an Interquartile Range.

e. How did you choose to handle the missing values and outliers?

The outliers were handled by an Interquartile Range was used to check if any variable has outliers or not. The Interquartile range was calculated by calculating the upper and lower quartile value of a variable. In order to remove the outliers, the capping method has been used.
The capping method limits the maximum and minimum value for the range of numbers. The box plot for each numeric variable is shown below:

Age

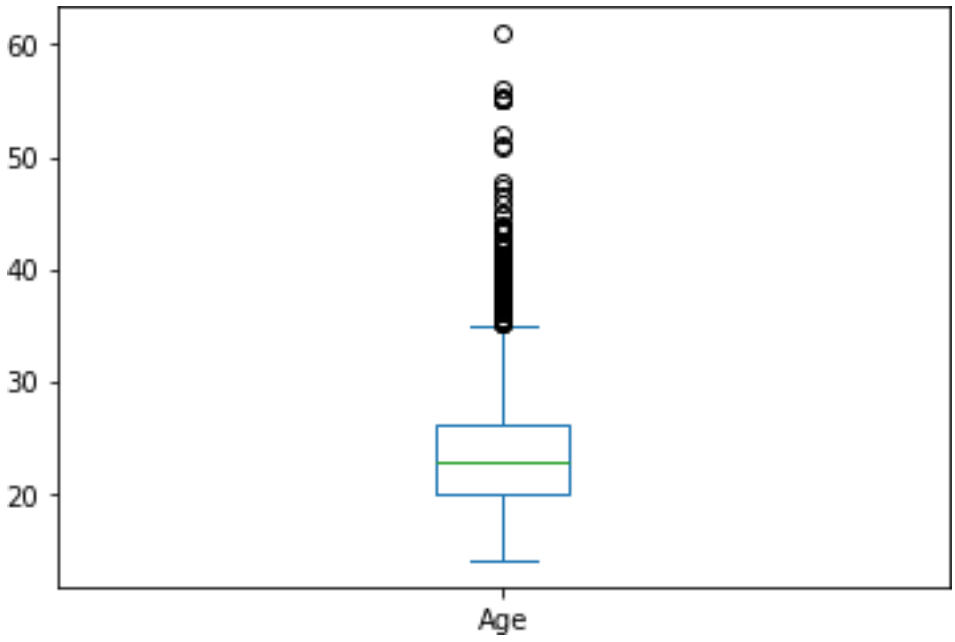


Fig 1. Box Plot for Age with Outliers

The above box plot for Age shows the mid value above 20 and the distribution of age within the Interquartile Range. The smallest number of data set can be seen at the lower end. However, the outliers can be seen in the upper whisker that indicates an individual above 35.

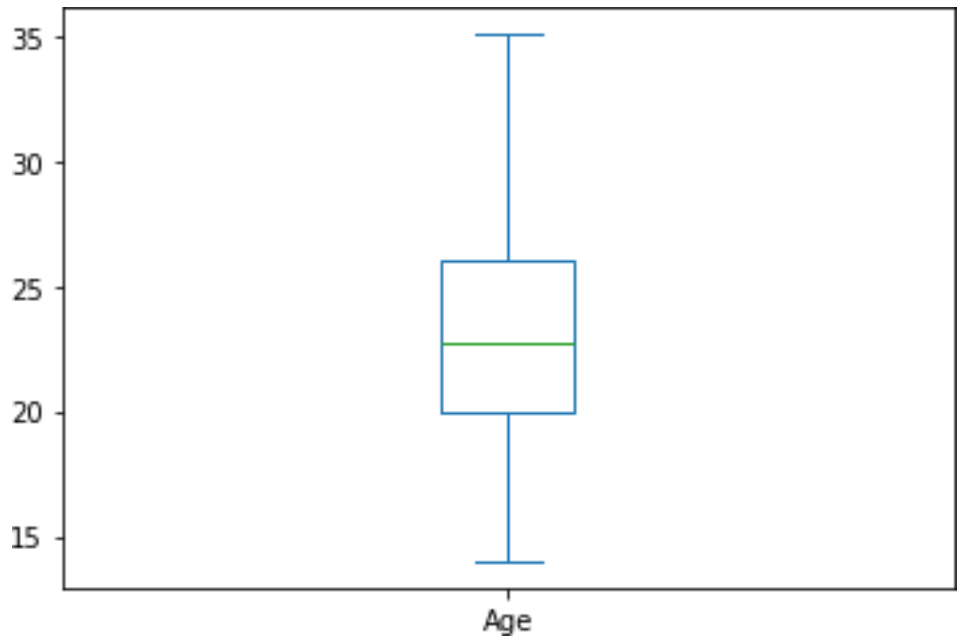


Fig 2. Box plot for Age after removing outliers

The above box plot for age after removing outliers shows that the upper limit is capped at 35. Here, the first quartile (Q1) is 20, the middle value is 22.5 and the third quartile (Q3) is around 26.

Weight

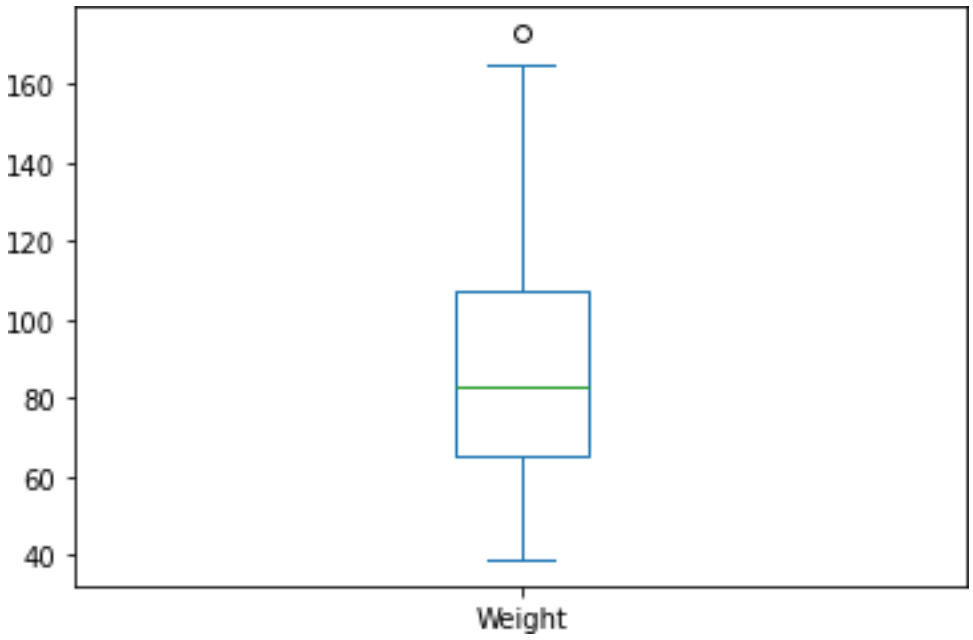


Fig 3. Box Plot for Weight with Outliers

The above box plot for Weight shows the mid value around 82. The smallest number of data set can be seen at the lower end. However, the outliers can be seen in the upper whisker that indicates an individual's weight above 160.

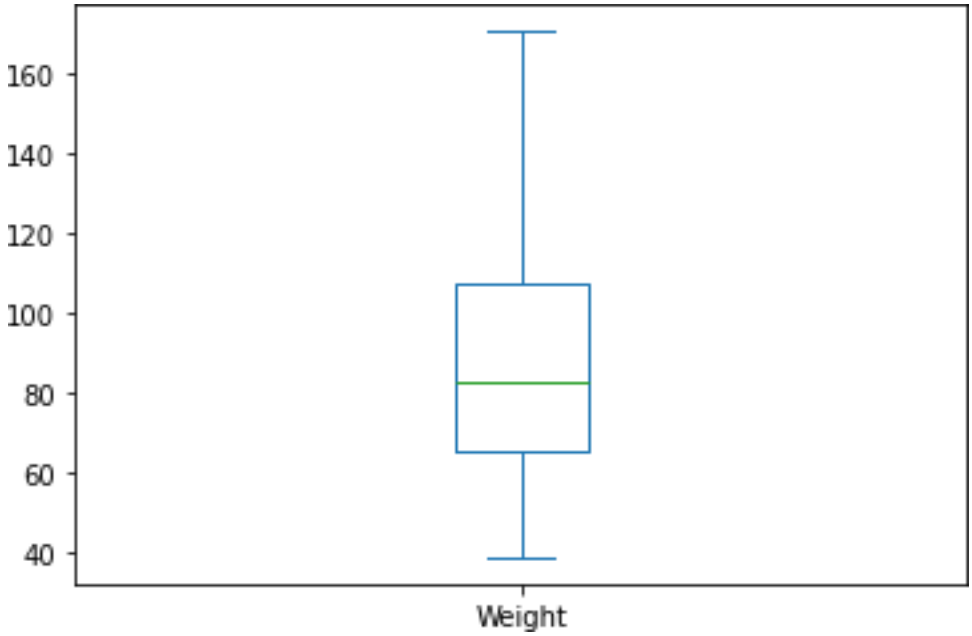


Fig.4 Box plot for weight after removing outliers

The above box plot for weight after removing outliers shows that the upper limit is capped at 160. Here, the first quartile (Q1) is slightly above 60, the middle value is 80 and the third quartile (Q3) is around 110. The plot indicates the weight distribution is positively skewed as the upper whisker is longer than the lower whisker.

Height

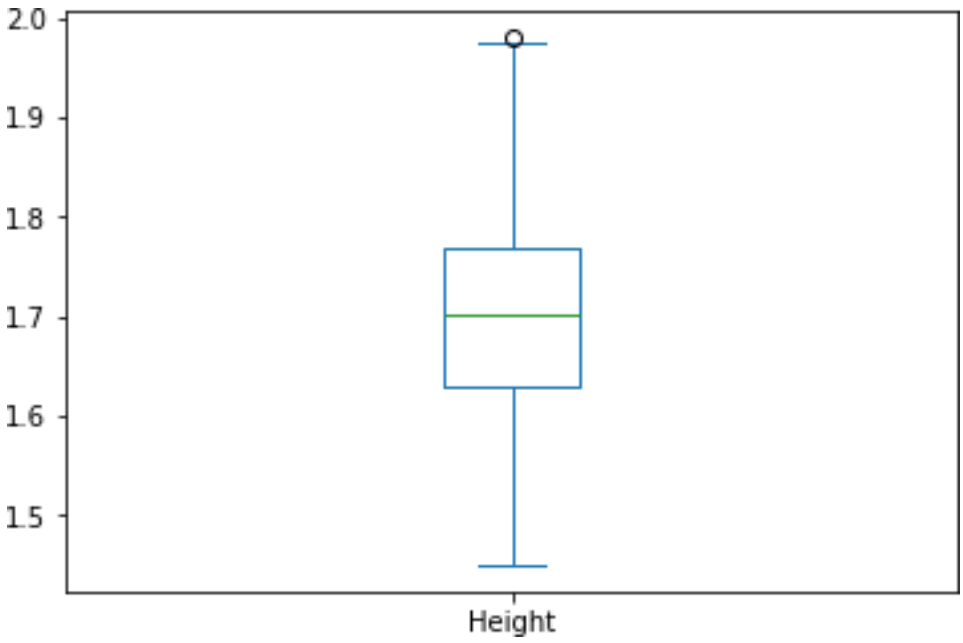


Fig 5. Box Plot for Height with Outliers

The above box plot for Height shows the mid value at 1.7 meters. The outliers can be seen in the upper whisker that indicates an individual's height slightly below 2.0.

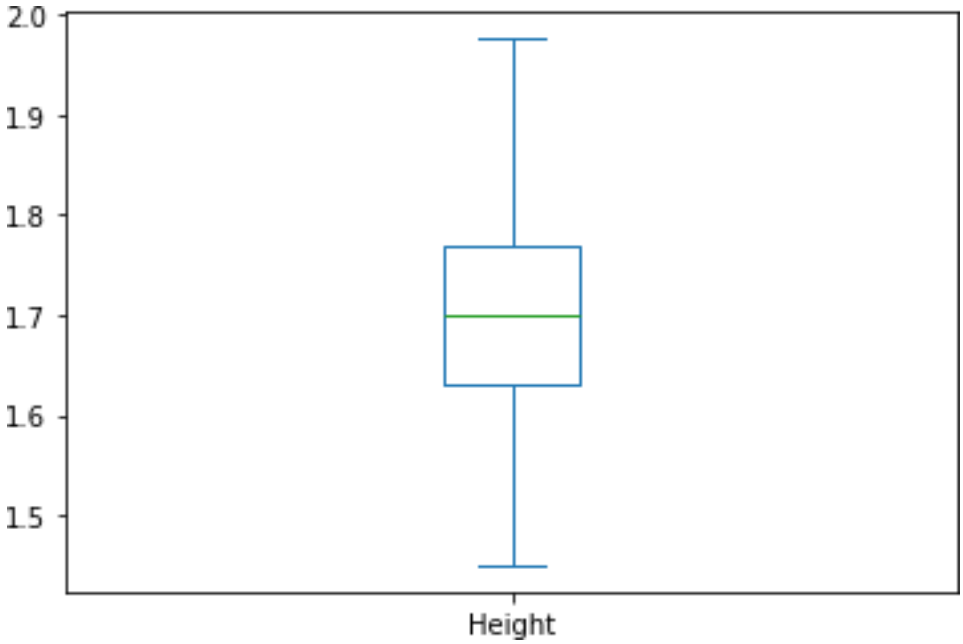


Fig.6 Box plot for Height after removing outliers

The above plot for height after removing outliers shows the evenly distribution of the upper and lower whiskers. This indicates the symmetrical distribution of height (in meters). The first quartile (Q1) is around 1.65 meters, the middle value is 1.7 meters and the third quartile (Q3) is slightly below 1.8 meters.

CH2O (Water Intake)

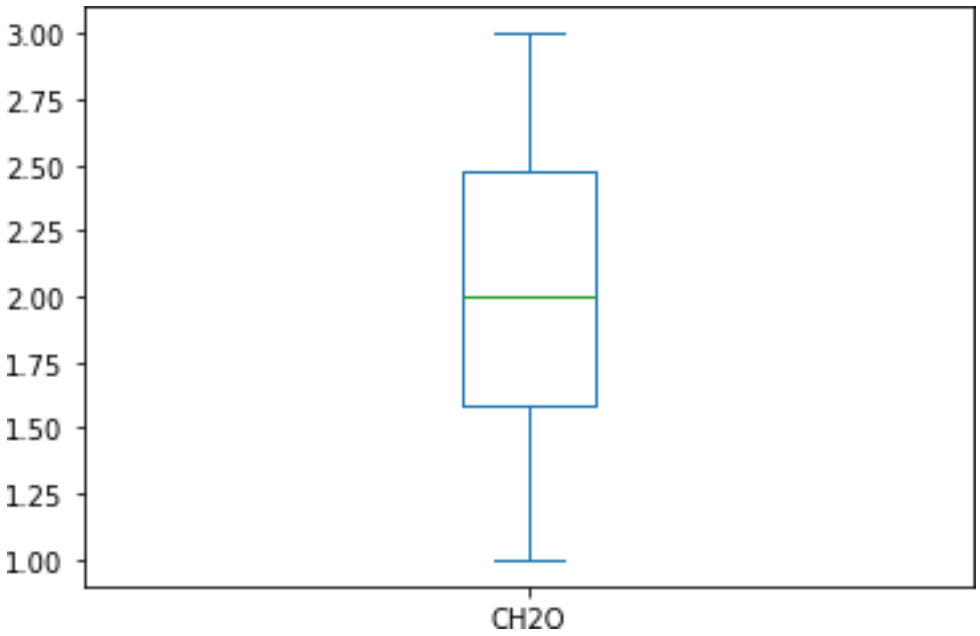


Fig.7 Box Plot for CH2O

The above box plot shows the water intake of an individual on a daily basis. The first quartile (Q1) is 1.55 liters, median is 2 liters and third quartile (Q3) is 2.5 liters. There are no outliers in CH2O and the upper and lower whiskers are evenly distributed indicating symmetrical distribution of data.

FAF (Physical Activity)

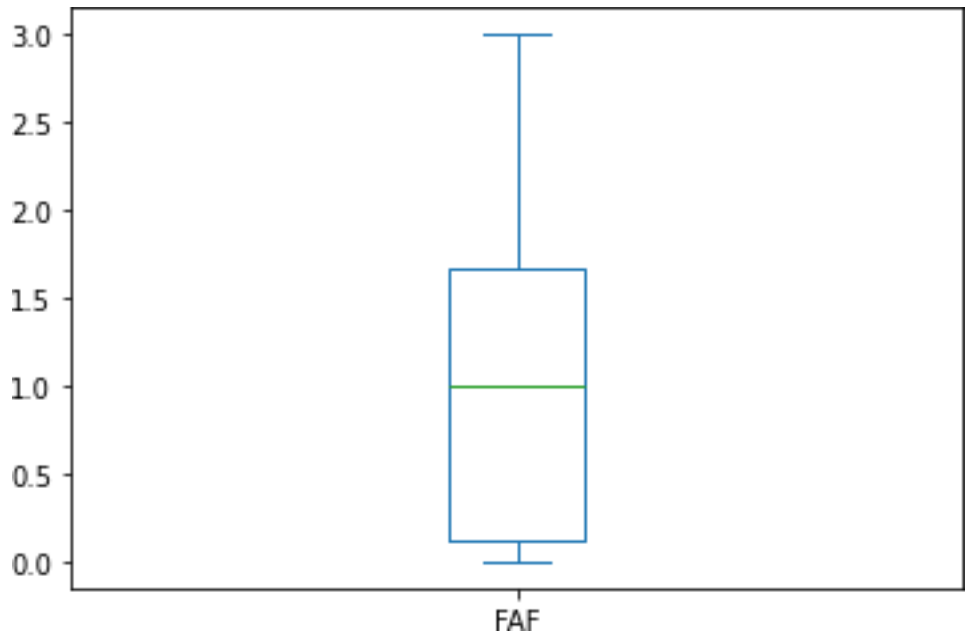


Fig 8. Box Plot for FAF(Physical Activity)

The plot for FAF variable indicates the distribution of data is positively skewed as the upper whisker is longer than the lower whisker. Outliers were not found for FAF.

TUE (Use of Technological Devices)

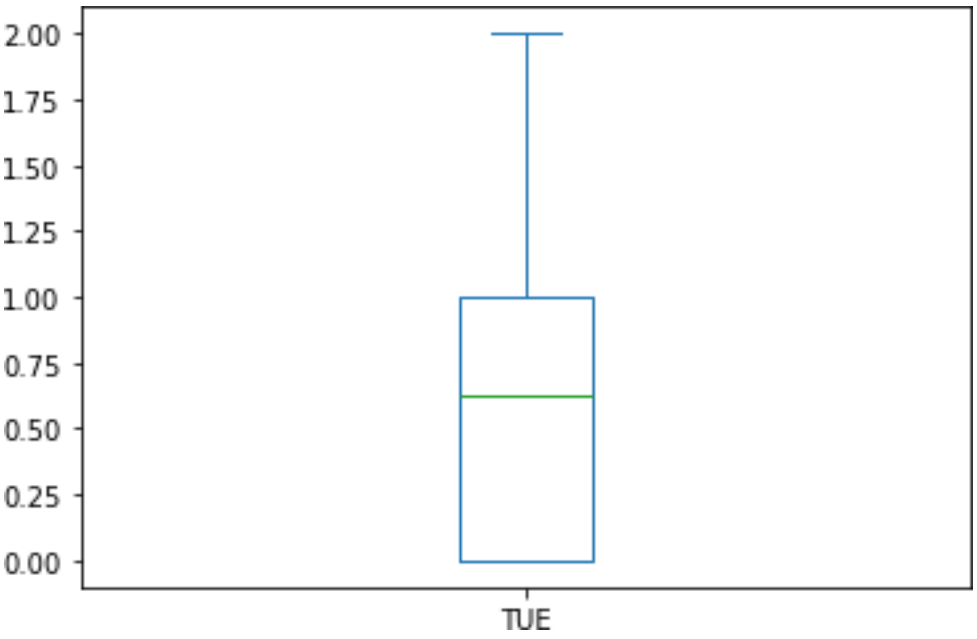


Fig. 9 Box plot for TUE

The above box plot for TUE illustrates the hours spent using technological devices. There were no outliers in this case. Also, the distribution of data is positively skewed as the upper whisker is longer.

Nobeyesdad (Obesity Level)

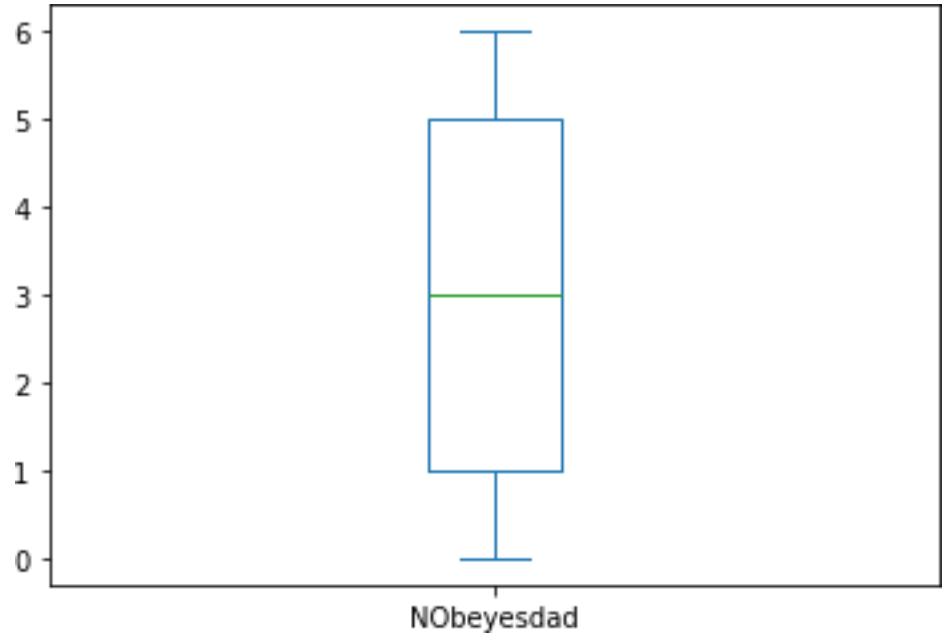


Fig 10. Box plot for Obesity Level

The box plot for Nobeyesdad indicates the obesity level from 0 to 6, where

0 = Insufficient Weight

1 = Normal Weight

2 = Obesity Type I

3 = Obesity Type II

4 = Obesity Type III

5 = Overweight Level I

6 = Overweight Level II

The first quartile (Q1) is 1, median is 3 and third quartile (Q3) is 5. There are no outliers in obesity level and the upper and lower whiskers are evenly distributed indicating symmetrical distribution of data.

f. What are some of the descriptive statistics for key variables?

Some of the descriptive statistics for key variables are listed as below:

Qualitative Variables

- Mean
- Median
- Mode
- Variance
- Standard Deviation
- Range
- Interquartile Range
- Outliers
- Skewness

Categorical Variables

- Number of Counts
- Visualization that includes bar chart , count plot and pie chart

Summary Statistics

	Age	Height	Weight	FCVC	NCP	CH2O	FAF	TUE	Age_clean
Count	2111	2111	2111	2111	2111	2111	2111	2111	1998
Mean	24.3126	1.70168	86.5861	2.41904	2.68563	2.00801	1.0103	0.65787	24.731231
Std	6.34597	0.0933	26.1912	0.53393	0.77804	0.61295	0.85059	0.60893	6.2808048
Min	14	1.45	39	1	1	1	0	0	18
25%	19.9472	1.63	65.4733	2	2.65874	1.58481	0.12451	0	20
50%	22.7779	1.7005	83	2.3855	3	2	1	0.62535	23
75%	26	1.76846	107.431	3	3	2.47742	1.66668	1	26
Max	61	1.98	173	3	4	3	3	2	61

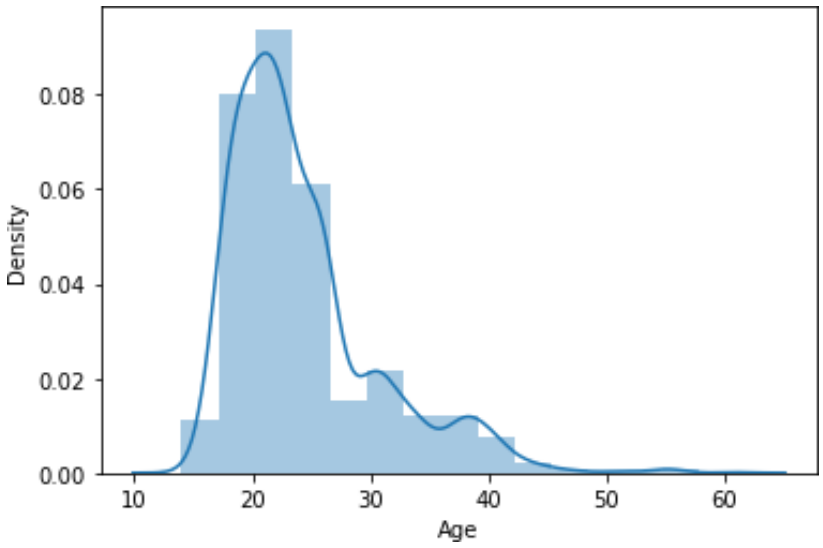
Table 1. Descriptive Statistics for Numeric Variables

From the descriptive statistics we can observe the following:

- The total number of count for each numeric variable is 2111.

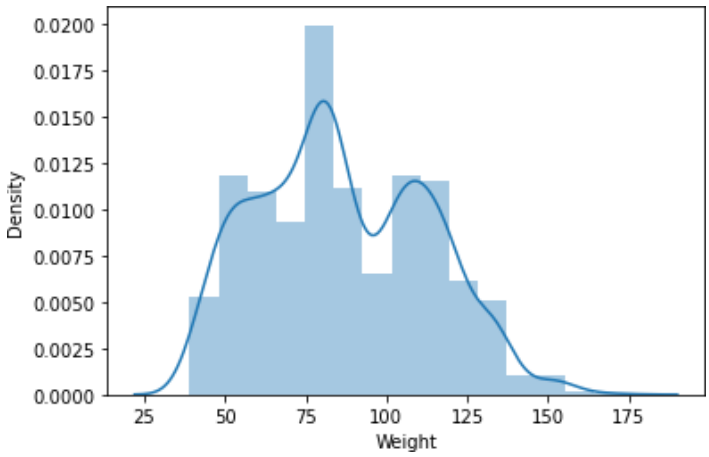
- The mean shows an average age of an individual which is 24.3126. The standard deviation of age is 6.34597. The percentile shows that 25% of an individual are of age 20 and 75% of an individuals are of age 26. Likewise, the maximum age of an individual is 61.
- The mean shows an average height of an individual to be 1.70 meters. The standard deviation is 0.093 which indicates the moderate variance in height. Likewise, 25% of an individual fall under first quartile with a height of 1.63 meters and 75% of them fall under third quartile with a height of 1.768 meters. The maximum height of an individual is 1.98 and minimum height is 1.45.
- The average weight of an individual is 86.586 and the standard deviation is 26.192. The weight of an individual ranges from minimum 39 (in kilograms) to maximum 173 (in kilograms). Likewise, 25% of an individual fall under first quartile with a weight of 65.47 kg and 75% of them fall under third quartile with a weight of 107.431 kg.
- The physical activity of an individual's seems to be low. The minimum time spent during physical activities is 1 (in hours) and maximum time spent is 3 (in hours).
- The CH2O which indicates the average water intake of an individual to be 2 (in liters) per day.

Distribution Plot



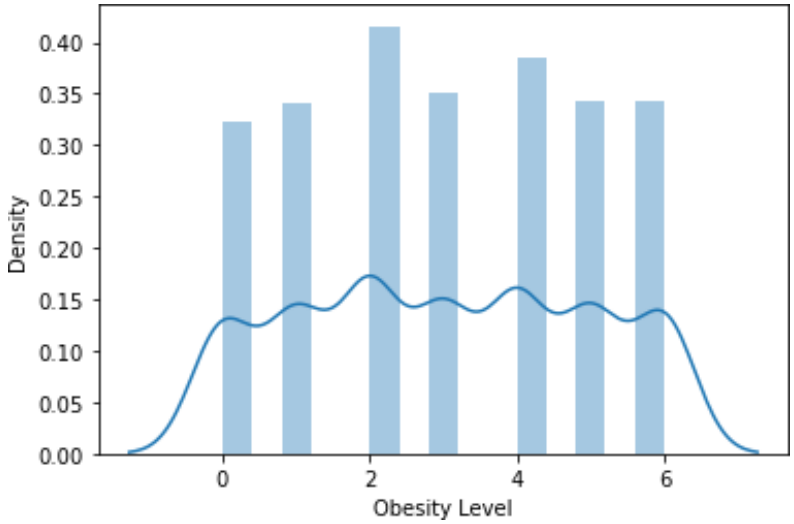
Distribution Plot for Age

The above plot illustrates the highest distribution of age around 20. There is a significant drop that shows an individual's age between 40 -60 is less as compared to the young individuals. Likewise we can see the distribution is positively skewed as the tail seems to be pulled towards the older group of an individual's.



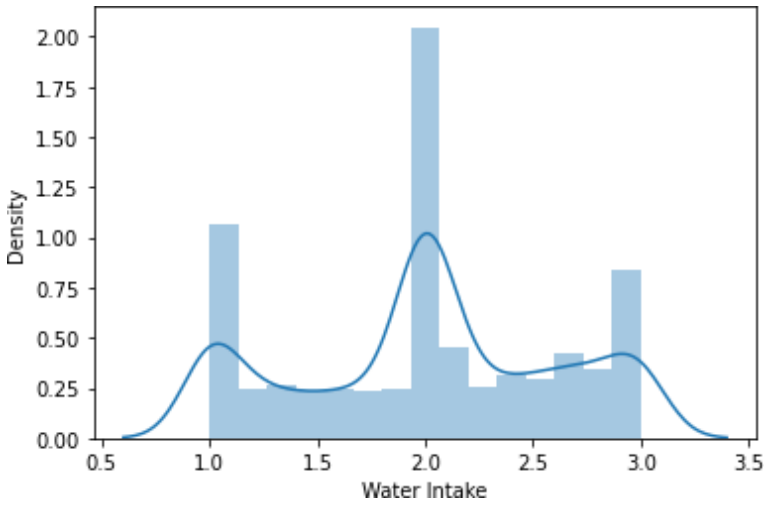
Distribution Plot for Weight

The above plot illustrates the multimodal distribution of the data. The highest distribution of weight can be seen at 75. There is a significant drop at 100 and again peaks around 110.



Distribution Plot for Obesity Level

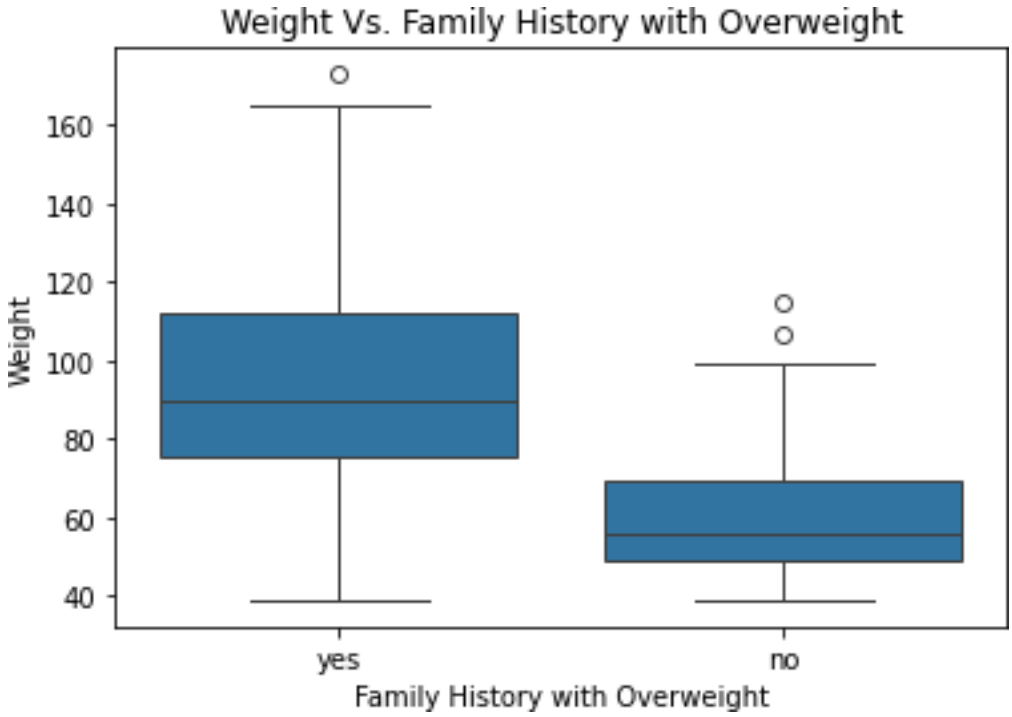
The plot illustrates the uniform distribution of obesity level of an individual's. However, the slightly high distribution of data can be seen in an individual with Obesity Type I and Obesity Type II.



The plot for water intake (in liters) per day shows a multimodal distribution of a data. The highest intake of water is 2 liters per day. This indicates that most of an individual prefer to drink only 2 liters of water per day.

Visualization of Data

1. **Relationship between weight and family history with overweight**



This shows the weight of an individual based on family history. The median value with family history is above 80, whereas the median value without family history is below 60. This show an individual with a family history is overweight.

Relationship between Obesity Level and Weight

The scatter plot illustrates the relationship between Obesity Level and Weight. The obesity level in the data set is from 0 to 6, is assigned as:

0 = Insufficient Weight

1 = Normal Weight

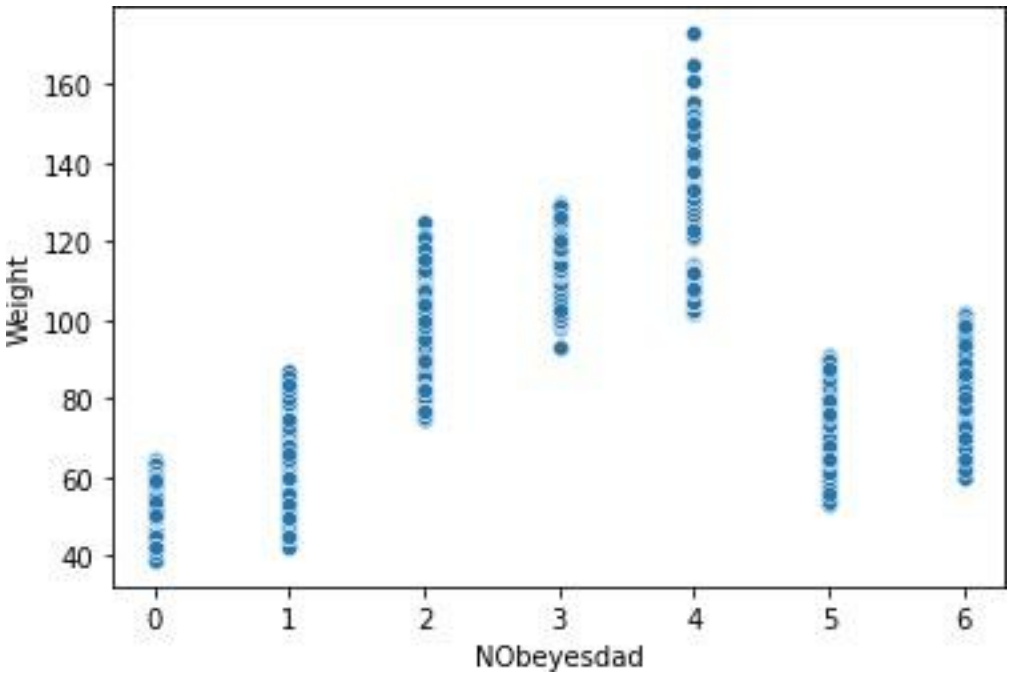
2 = Obesity Type I

3 = Obesity Type II

4 = Obesity Type III

5 = Overweight Level I

6 = Overweight Level II

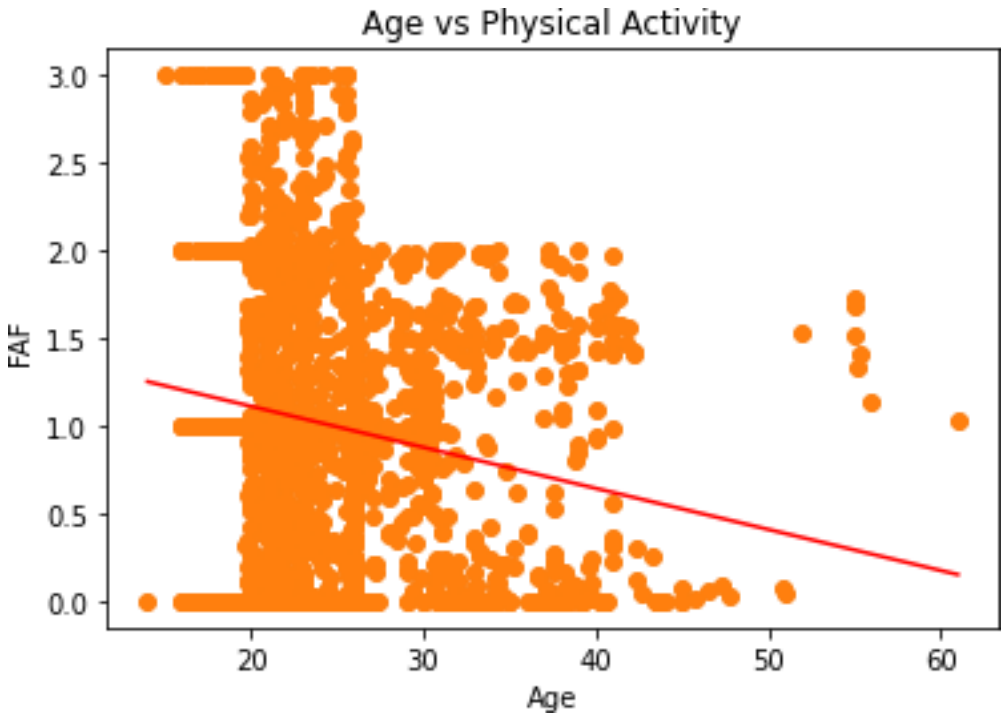


Scatter plot for weight and obesity level

Here, the x-axis represents obesity level and the y-axis represents weight. Obesity levels 0 and 1 represent insufficient weight and normal weight respectively. This shows an individual's weight between 40 – 80 is considered to be low. Individuals with an obesity level 2 and 3 show a weight between 80 -130. Likewise, obesity level 4 i.e. Obesity Type III is seen to be higher, that ranges from 100 to 160. Thus, we can see a positive relationship between obesity level and the weight of an individual.

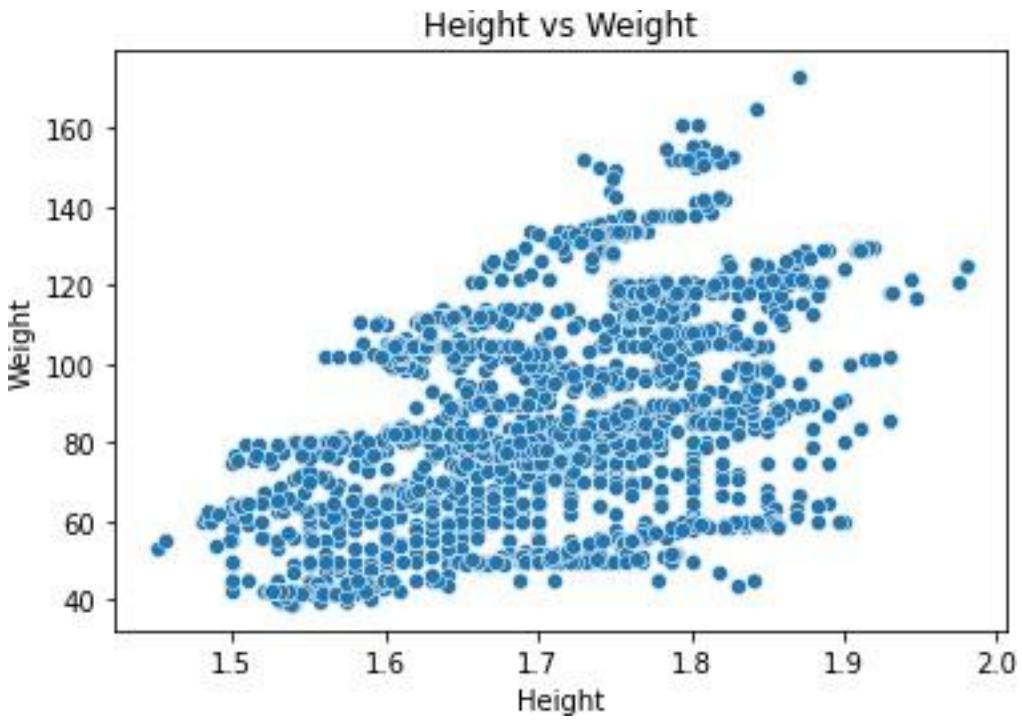
2. **Relationship Between Age and Physical Activity (FAF)**

The scatter plot with a regression line shows a relationship between age and physical activity. An individual between the ages of 20-30 is highly active, whereas an individual between the ages of 40-60 is less active. This shows that with an increasing age, the physical activity significantly decreases.



Scatter plot with regression line for Age and Physical Activity

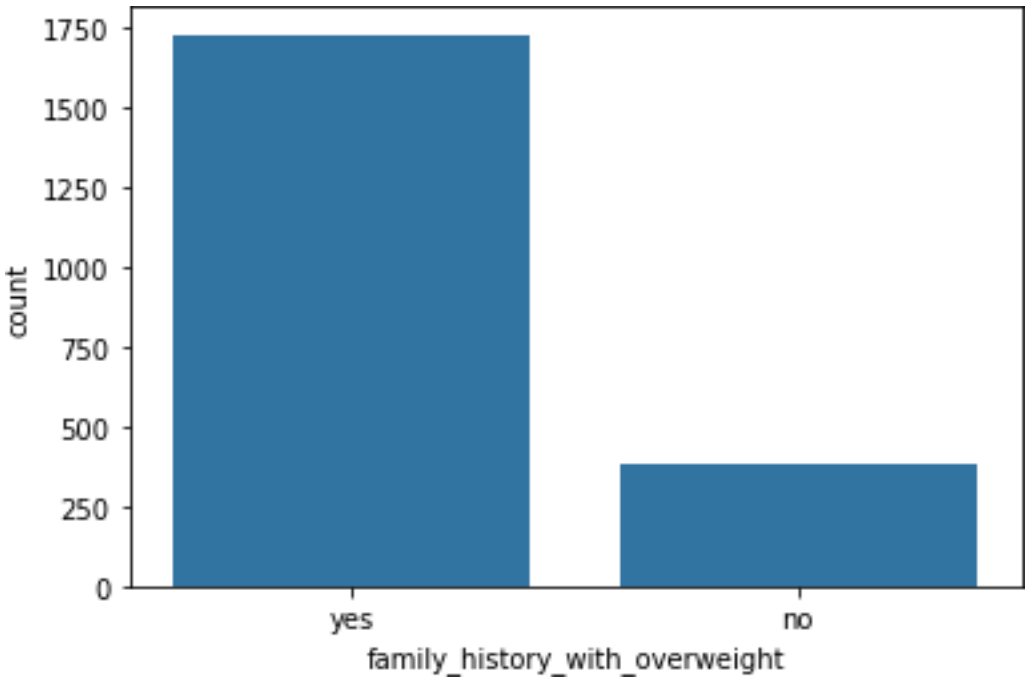
3. **Relationship between height and weight**



Scatterplot showing Height vs. Weight

The scatter plot shows the positive relationship between height and weight. As we can see, an individual weight tends to increase when a height increases. For instance, when the height increase from 1.7 meters to 1.8 meters weight of an individual also increases.

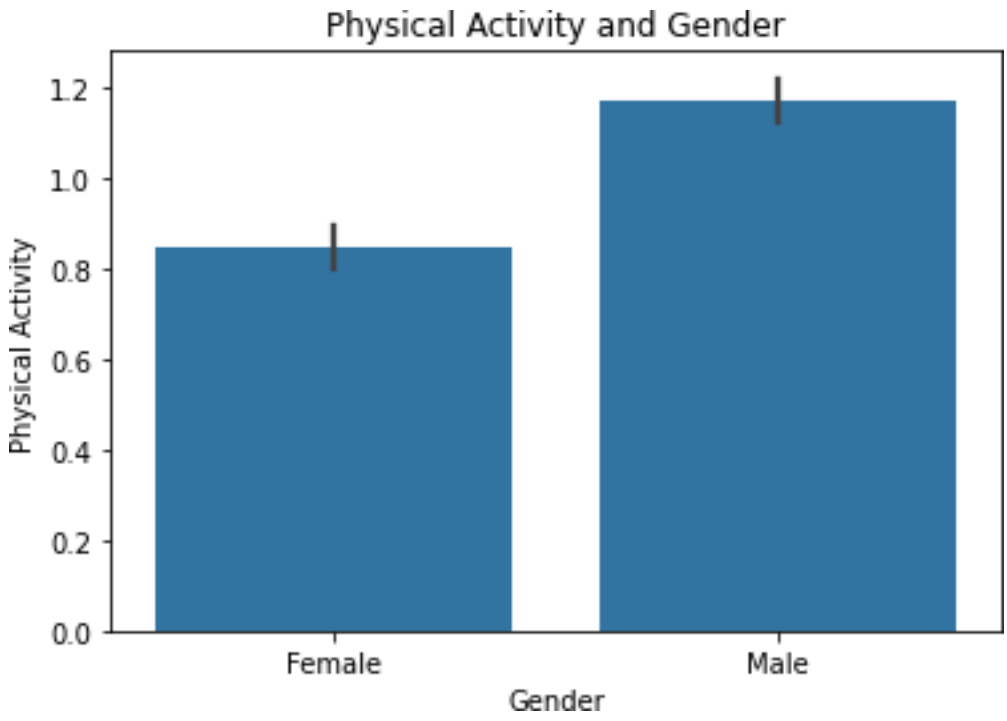
4. **Bar Chart for Family History with Overweight**



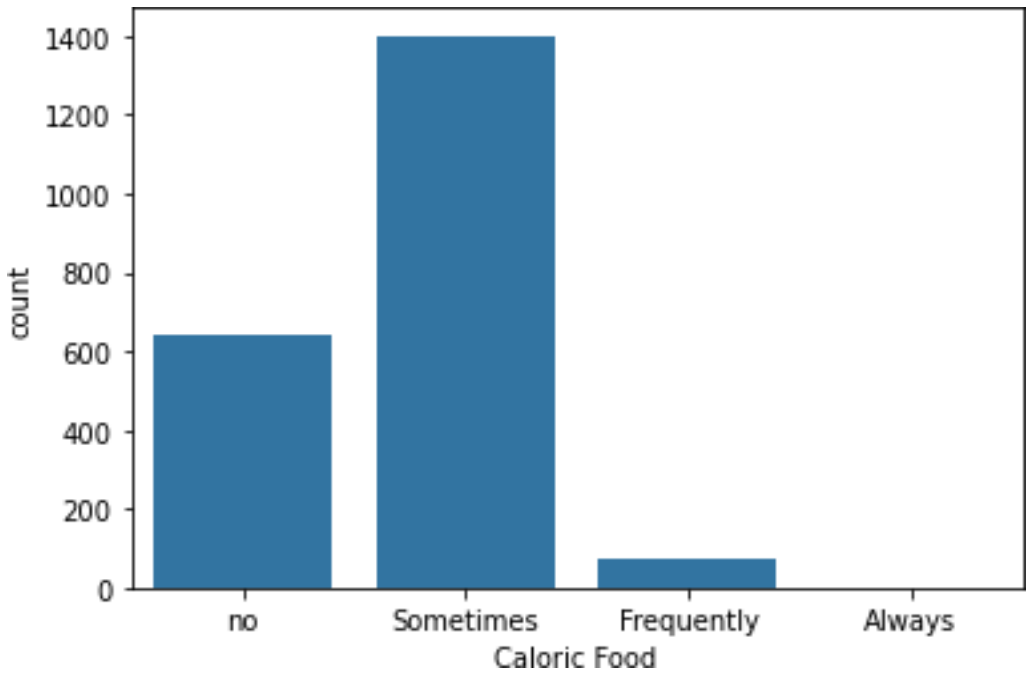
The bar chart illustrates the number of individuals with a family history with overweight. This shows that more than 1700 an individual have a family history with overweight. Whereas, around 400 individuals do not have a family history with overweight.

5. **Relationship between Gender and Physical Activity**

The bar chart shows the relationship between gender and physical activity, It illustrates that male are physically active as compared to female. Since, the physical activity is important to reduce obesity level; we can say that the different obesity level can be seen in females.



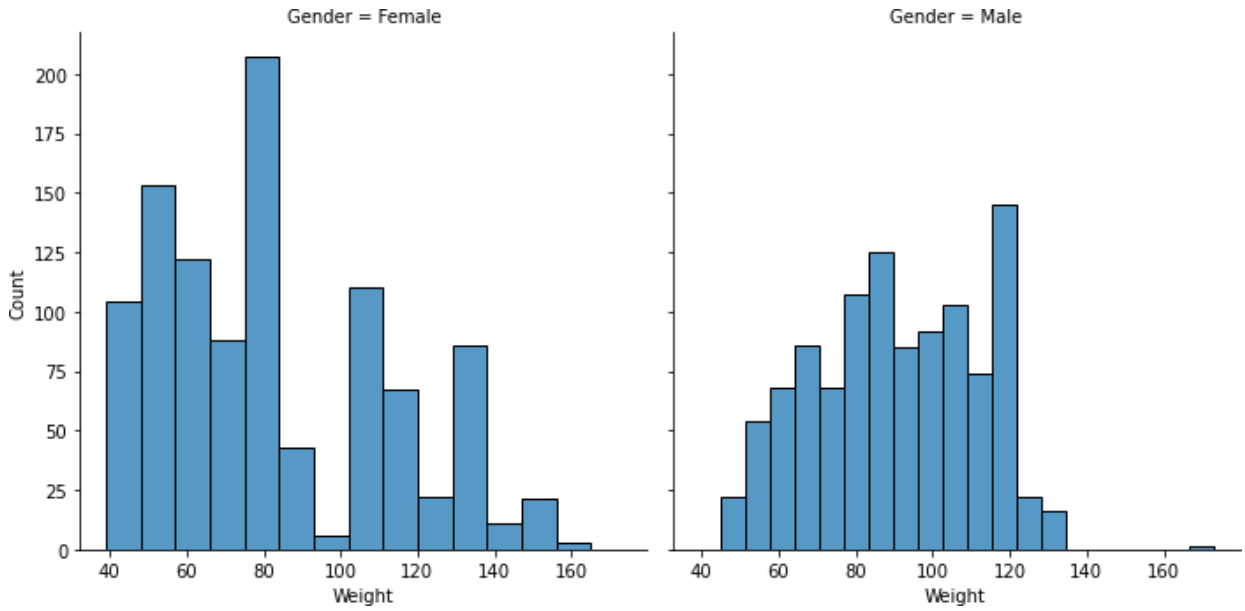
6. Bar Chart for Consumption of Caloric Food



The bar chart represents the consumption of caloric food of an individual. It shows the highest number of individuals prefers caloric food sometimes and lowest number of an individual fall

under frequently. Likewise, significant number of an individual does not prefer caloric food which focuses on the unhealthy eating habits of individuals.

7. Relationship between Female and Male Weight



The above chart illustrates weight of male and female individuals. We can see that males have higher weight i.e. 120 kg, as compared to female. The highest number of females fall under 80.

Findings and Conclusion

The data set shows the age distribution of an individual's 20-26. The average weight and height of an individual is 86.58 kg and 1.70 meters respectively. The data also shows that the individual's physical activity is low which has affected the weight of individuals. Further, the analysis of a data set shows that family history with overweight, physical activity, consumption of caloric food, are some of the factors that affects the obesity level of an individual. The data shows a positive relationship between obesity level and weight, where an individual weight between 40-80 kg fall under normal weight and, 100-160 kg fall under obesity type III. Likewise, there is a positive relationship between weight and height of an individual. Likewise, the consumption of caloric food is very low that reflects their unhealthy eating patterns. The water intake of an individual's seems to be balanced as the water intake in 2 liters per day.

The finding and visualization of data can help dietitians to deal with an individual from different age group and plan diet based on portion size and nutrition. This can help dietician to find the obesity level that might include an individual's family history with overweight. Identifying this in the early stage can help dietitians to analyze the health risks such as heart diseases, high blood

pressure, cancer etc. This can help dietitians to focus on these groups and encourage healthy eating habits and physical activities.

Results

The data set has no missing values however we can see the outliers in age, height and weight variable. Capping method has been used to deal with outliers. The distribution plot helped me to understand the skewness of the data. Likewise, the visualization part has helped me to explore different types of plot that can be used for numeric and categorical variables. The descriptive statistics provides a better understanding on interpreting data which has enhanced my analysis skill. In addition, the project has provided me roadmap on how we can analyze data from scratch. Since, the data set has 17 variables; I would focus on showcasing the relationship with each variable differently in future. The challenges were to use the variables to plot stacked bar chart and line graph. Lastly, the Python was a good tool to analysis the data set as it helped me to analyze the outliers, handle outliers, provided results using a line of code and made visualization of data easy.