Data Exploration, Pattern Detection, and Anomaly Detection

Techniques I used, to find relationships in the data:

-PEARSON Correlation Method:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\{\sum_{i=1}^{n}(x_i - \bar{x})^2\}^{\frac{1}{2}}\{\sum_{i=1}^{n}(y_i - \bar{y})^2\}^{\frac{1}{2}}}$$

-Kendall Correlation Method:

Kendall's Tau = (C – D / C + D)

-Spearman Correlation Method:

$$\rho = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

Pattern Found:

Dataset1:

-Pattern Detected in, Screentime before sleep hrs, Meal intake, and Workhours. The initial values of the columns have less instances in comparison to other half instances. Approximately, 50% data has 20% instances and 50% data has 80% instances.

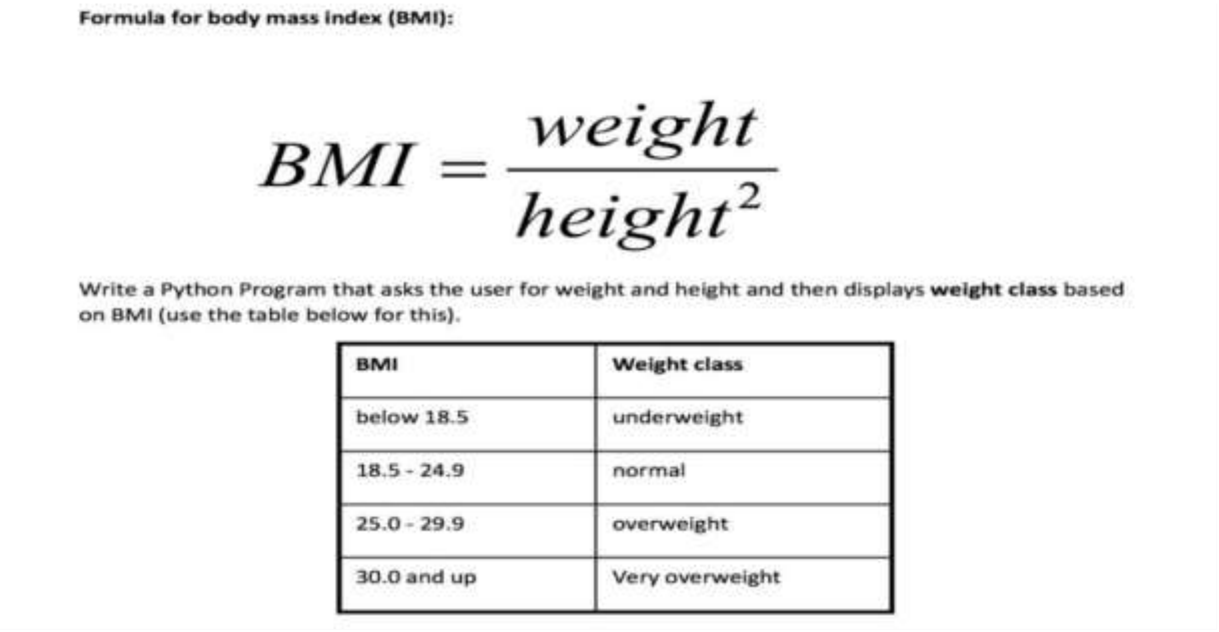
Dataset2:

-Carrier Company Name, Uniquely Identified Carrier, Source Port, Destination Port, have the values that has equal number of instances in each columns.

-Ship beam, Ship draft and Storage Capacity have the bell shaped data, which shows the Normal Distribution.

Dataset3:

- In third dataset, we have a linear dependence on weight and inverse relation with height^2. BMI and Weight Class are also related by this:



Formula for body mass index (BMI):

$$BMI = \frac{weight}{height^2}$$

Write a Python Program that asks the user for weight and height and then displays **weight class** based on BMI (use the table below for this).

BMI	**Weight class**
below 18.5	underweight
18.5 - 24.9	normal
25.0 - 29.9	overweight
30.0 and up	Very overweight

Anomaly Detection Techniques that I used:

-Median Absolute Deviation Method:

-Found the lower bound and the upper bound of the dataset using: (let, data=x)

Lower bound = median(x) –2.5 * MAD(x)

Upper bound =median(x)+2.5*MAD(x)

MAD = median(abs(x – median(x)))

-Inter Quartile Range (IQR method):

-Found the lower bound and the upper bound of the dataset using: (let, data=x)

Lower bound =Q1 – 1.5*IQR

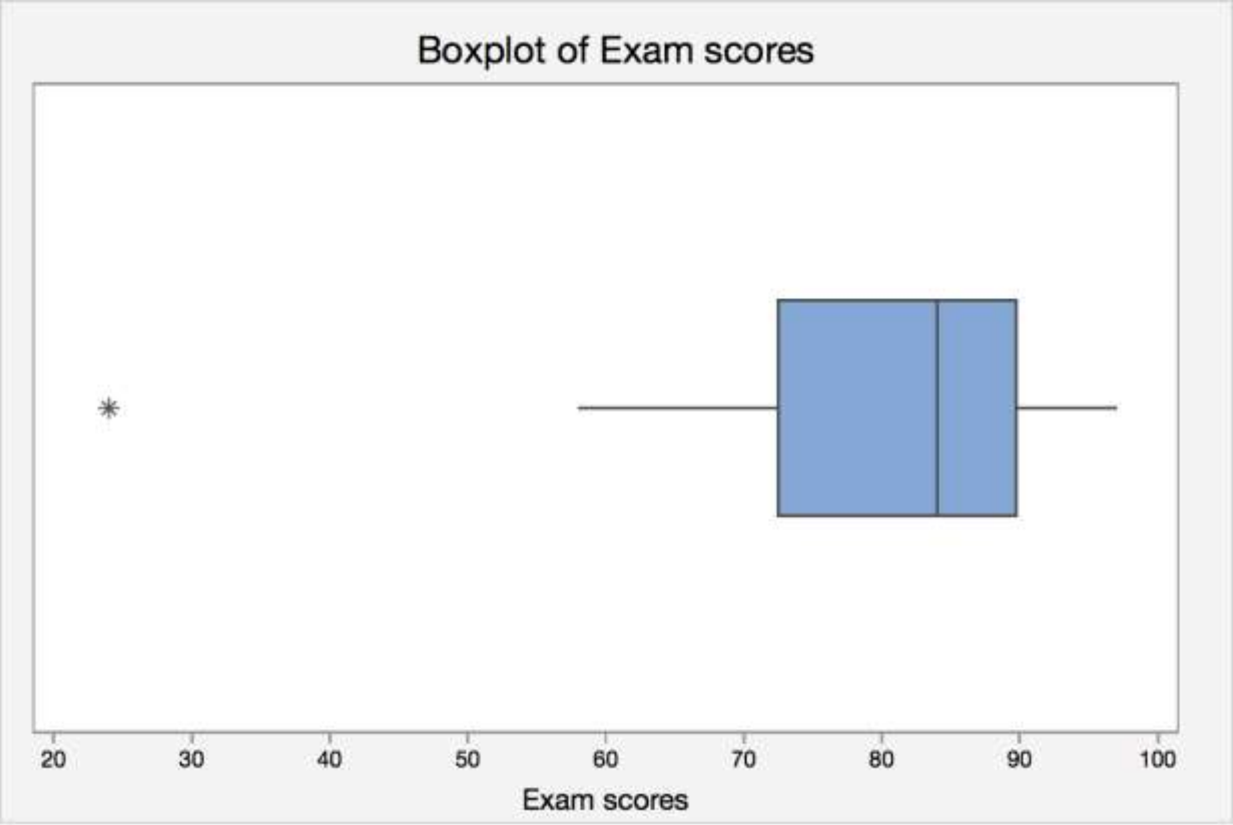
Upper bound =Q3+1.5*IQR

Q1– >first quartile number (Parameter setting: (20-25 percentile))

Q3 -> Second quartile number (Parameter setting: (75-80 percentile))

IQR = Q3-Q1

-Boxplot method:



-3 Standard deviation method:

-Found the lower bound and the upper bound of the dataset using: (let, data=x)

lower bound = mean(x) – 3*std(x)

upper bound = mean(x)+3*std(x)

std -> Standard Deviation