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

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

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

-I calculated calories using weight, height and age. #source : https://www.thejakartapost.com/life/2016/09/27/how-to-calculate-your-ideal-calorie-intake.html

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

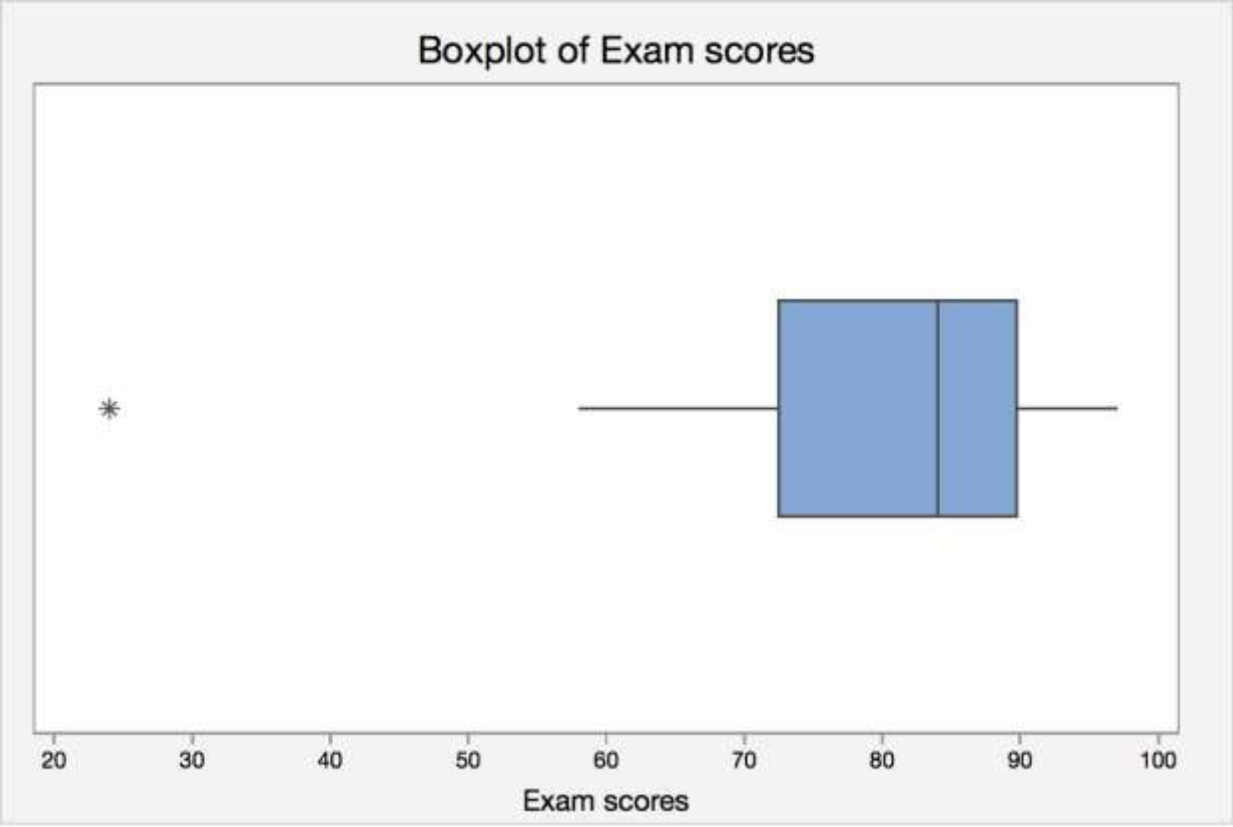
Upper bound =Q3+1.5*IQR

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

Summary:

For Dataset1:

It is about the sleep cycle of the human being, and what factors affect Sleep quality and Health, it includes Screen time on Computers, Workhours, Meals, Smoking, Drinking. These are responsible for poor health and the quality of sleep.

Anomalies:

-Gender column is not evenly distributed among, (Female, Male, Undisclosed).

-Screen time before sleep hours column has an anomaly with data [0.25] which I found using Inter Quartile Range method. And confirmed it with Boxplot method.

-Workhours column have an outlier of [5] which I found by using Mean and 2 Standard Deviation method.

Relationship:

According to Pearson and Spearman correlation method we found the relation between workhours and meal intake.

Pearson correlation for Workhours and Meal intake = 70.8638%

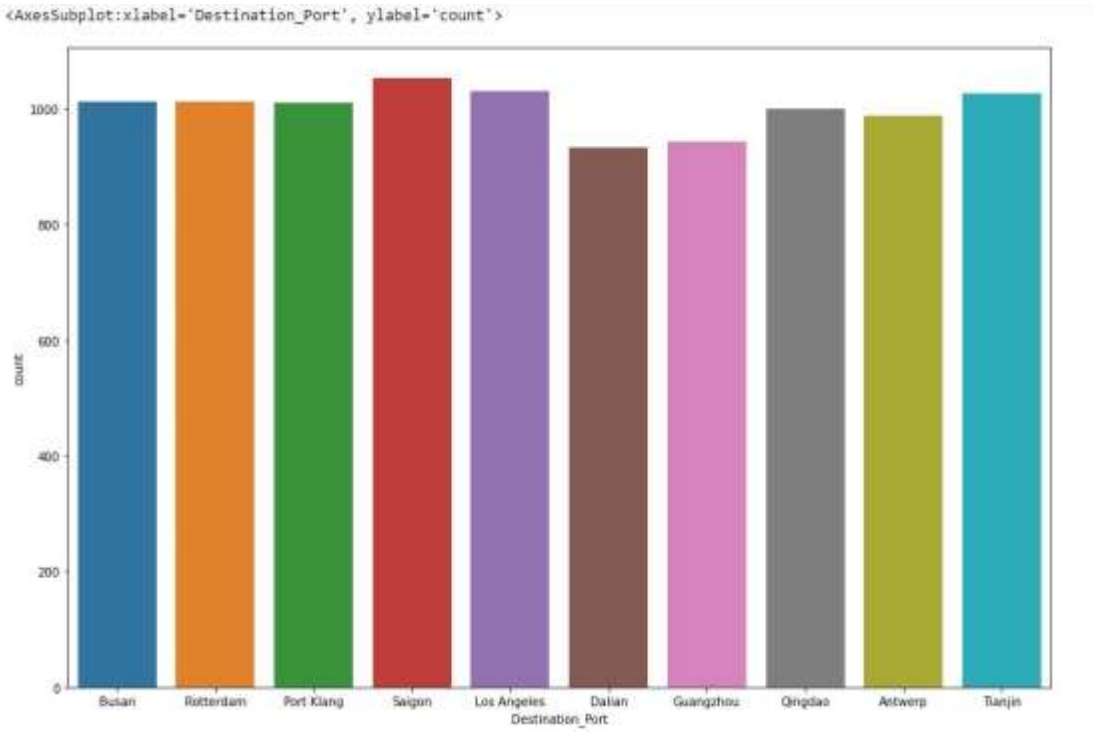
Spearman correlation between Workhours and Meal intake = 58.1%

For Dataset2:

It is about the building of ship and holding capacity of a ship with length, beam and draft given.

Anomalies:

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



-Ship Beam column have outliers that I calculated using Inter Quartile Range and boxplot. Outliers:

[111.147938511727, 153.119471060909, 153.921585905135,153.655470780786
153.821261505092, 112.990210098791,106.10071687613,153.40627236408,
158.213400000499, 106.393449752137, 153.786368776247,]

-Ship draft column has outlier that has been calculated by 3 Standard Deviation method. Detected outliers:

[48.1551360923857, 40.8977325058241, 40.6162748467276, 47.8801960468583,
47.792715741625, 40.6634415840681, 47.8101926167058, 40.3624707853361,
40.6695635348504, 40.8117469832801, 40.7416881650151, 48.3157005167571,
47.7791569845079, 48.3185339309961, 48.0116312737464, 47.8180883656121,
47.8365135464337, 41.0447229075354, 48.3337543026321, 48.1318635205562,
47.8611337234894, 47.9852063769665, 40.7286567440906, 40.7298254655064,
40.9327776367044, 40.3101583955569, 40.8340614579516, 40.6873036811692,
40.9457529433716, 40.4426407497061]

-Found 9 outliers in Storage capacity column using 3 standard deviations from mean method. Detected outliers:

[20309, 20068, 19956, 11699, 11676, 20039, 20619, 20352, 8937]

Relationship:

We found the relationship between Storage container with (Ship length, ship draft, ship beam).

		Ship length	Ship beam	Ship draft
Pearson	Storage	32.8%	23.7%	11.3%
Kendall	Storage	47.9%	30.3%	15.62%
Spearman	Storage	64.82%	43.42%	23.11%

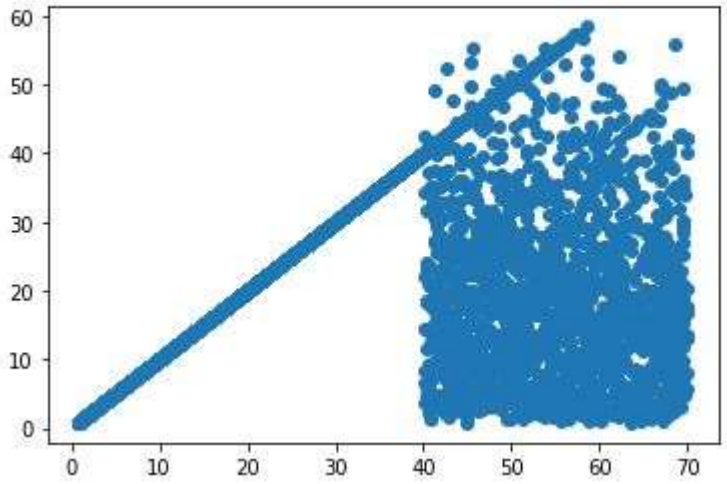
For dataset3:

-This dataset is all about the relation between Weight and height of the body. How we can use these parameters to Calculate the BMI and then Weight class of a person.

Anomalies:

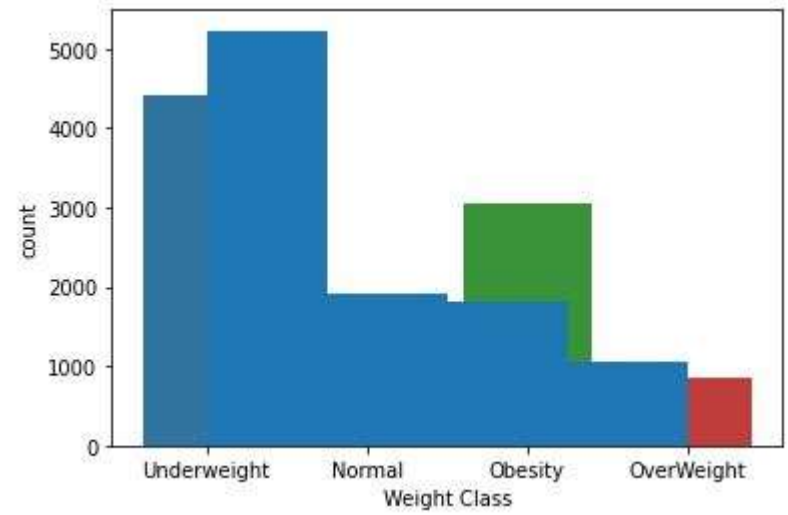
-BMI column have approximately 50% data error. We know, BMI = weight(kg)/height^2 (metre), but 50% Instances in this column are not following the Formula.

<matplotlib.collections.PathCollection at 0x212c7a45dc0>



This is the scatter plot of Calculated BMI and BMI column in the dataset. Approx. 50% instances are linear and rest are not.

-due to data errors in BMI, we have an error in Classification of Weight. Here's the plot for expected graph and actual Graph:



Relationship:

BMI = weight/height^2

Calories = (66.5 + 13.8*weight + 5*height*100)/(6.8*age)

#source :
https://www.thejakartapost.com/life/2016/09/27/how-to-calculate-your-ideal-calorie-intake.html