Data Exploration, Pattern Detection, and Anomaly Detection

- Identifier for the dataset

14-DataA.csv

- a one-paragraph description of each analysis technique you applied to attempt to identify the relationships in the data

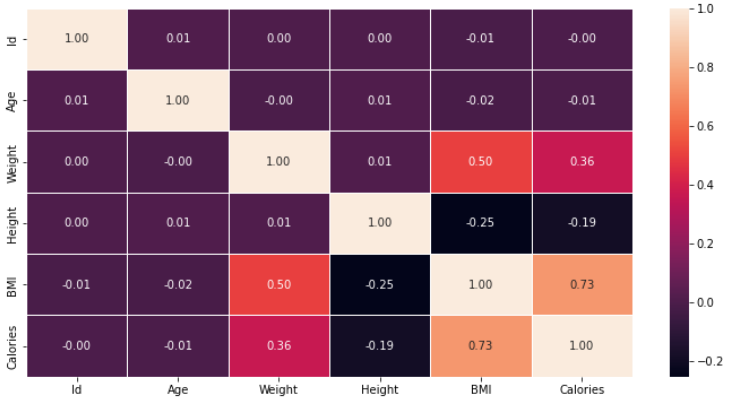
Techniques I used to find relationships in all three datasets are:

- Correlation Matrix:

It helped in showing the coefficient of correlation between different attributes which is the measure of linear relationship between those 2 attributes. The value in each cell of the matrix shows the correlation between two attributes. The values in the correlation matrix can be between -1 and +1. If correlation value is positive, then increase in one attribute increase the other and same for decrease as well. Whereas for a negative correlation value, if one variable increases the other decreases.

-Correlation Heat Map:

Correlation heatmaps are a type of data visualization graph that shows how strongly the attributes are correlated to each other and assigns a value between 0 and 1. Correlation heat maps are useful in showing the correlation between all the variables in and in showing how how strong this relationship is.



-PEARSON Correlation Method:

The Pearson correlation coefficient determines the strength of a linear link between two variables. It has a value between -1 and 1, with -1 indicating total negative linear correlation, 0 indicating no connection, and + 1 indicating total positive correlation.

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

from https://www.sciencedirect.com/topics/computer-science/pearson-correlation#:~:text=The%20Pearson%20correlation%20measures%20the,meaning%20a%20total%20positive%20correlation.

-Kendall Correlation Method:

When the data you're dealing with fails one or more of the test assumptions, Kendall rank correlation (non-parametric) is an alternative to Pearson's correlation (parametric). This is also the best non-parametric alternative to Spearman correlation when your sample size is limited and there are many tied ranks.

$$\text{Kendall's Tau} = (C - D / C + D)$$

From https://towardsdatascience.com/kendall-rank-correlation-explained-dee01d99c535#:~:text=Kendall%20rank%20correlation%20(non%2Dparametric)%20is%20an%20alternative%20to,and%20has%20many%20tied%20ranks.

-Spearman Correlation Method:

The Spearman's rank coefficient of correlation is a nonparametric measure of the statistical dependency of two variables' rankings. It determines the magnitude and direction of the relationship between two ranking variables. However, before we discuss the Spearman correlation coefficient, we must first grasp Pearson's correlation. A Pearson correlation is a statistic that expresses the strength of a linear relationship between two sets of data.

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

From https://www.questionpro.com/blog/spearmans-rank-coefficient-of-correlation/

Relationship:

The correlation methods like Spearman, Kendall and Pearson show that Workhours and Meal_intake are related

- a one-paragraph description of each pattern you found (add a mathematical description

Pattern in Gender Attribute:

Male ≈ Female (approx.) but undisclosed = 0.5% of the entire instances in the Gender Attribute.

Undisclosed ≈ 103 times Male/Female Instances

Undisclosed = 103 * Male

Undisclosed = 103 * Female

Undisclosed instances ≈ 1% of Male/Female Instances

The number of data samples collected for Males is almost the same as data samples collected by females.

Undisclosed could be anomaly

Pattern in Health Status:

Number of instances of Sick people are almost half of that of Healthy people
PATTERN: Okay, Healthy ≈ 2 times Sick
Okay ≈ 1.5 * Sick and Okay ≈ 1.8 * Healthy

Pattern in Screen_time:

Half of the attributes of this data has approximately 20% distribution whereas other half has 80%
Instances of Screen time between 1.5-2.5 is 4 times the Instances of Screen time of 0-1.25.
Mathematical Description
Instances of Screen_time(6-10) = 2.3 * Instances of Screen_time(5-7)
Scatterplot of the Workhours attribute shows some uneven distribution of instances.

Pattern in Workhours:

Almost half of the attributes of this data has approximately 30% distribution, whereas other half has 70% Instances of Workhours between 6-10 is 2.3 times the Instances of Work hours between 5-7.
Mathematical Description
Instances of Workhours (6-10) = 2.3 times Instances of Screen_time (5-7)

Pattern in Meal_Intake:

Half of the attributes of this data has approximately 25% distribution whereas other half has 75%. Meal_intake of 1-2 meals in a day is 3 times the Meal_intake of 3-4 meals in a day
Mathematical Description
Instances of Meal_intake(1-2) = 3 X Instances of Screen_time(3-4)

Pattern for Sleep_Quality:

Data for 45-50, 55-60 are missing, hence making it unevenly distributed attribute. The missi ng data could be anomaly.

- a one-paragraph description of each anomaly detection technique that you applied

Anomaly Detection Techniques that I used for all three datasets:

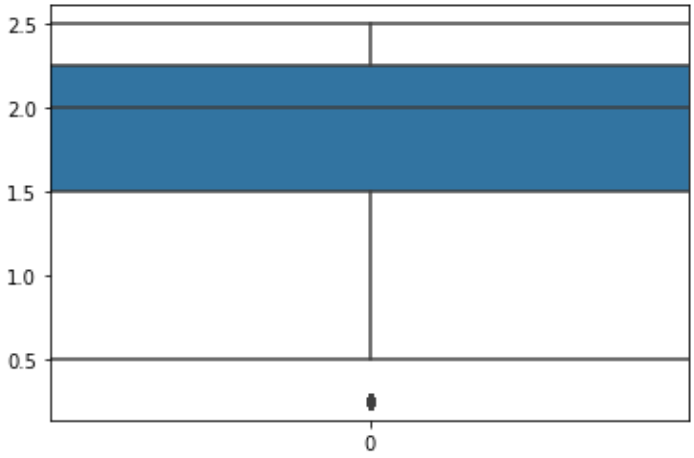
1. Median Absolute Deviation Method:

I first 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)
Then calculated the Median Absolute Deviation where MAD = median (abs (x – median(x)))

2. 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



3. 3 Standard Deviation Method:

First I found the lower bound and the upper bound of the dataset using: (let, data=x)
The lower bound is mean(x) – 3*std(x)
and upper bound = mean(x)+3*std(x)
then Standard Deviation is calculated using

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Anomaly Detection in Screen_time:

Using Quartile deviation Boxplot Method. There was 1 detected outliers in the data 0.25.

Anomaly Detection in Workhours:

Using Mean & Standard Deviation 1 outlier was found 5. while using Boxplot Method, no ou tliers were found.

Anomaly Detection in Sleep_Quality:

No Anomaly Detected

- a brief summary of your analysis that mentions the keys points about the application domain that you see in the data, including the relationships and types of anomalies

- For Dataset1: 14-DataA.csv
- **Application Domain:**
  It is about sleep cycle of human beings and what factors affect their sleep quality and health. It includes screen time on computers, number of workhours, number of meals intake in a day as well as smoking and drinking habits. These factors are responsible for poor health and quality of sleep.
- **Anomalies:**
  - Gender attribute is not evenly distributed among Female, Male and undisclosed.

- Screen time before sleep hours attribute has an anomaly with data [0.25] which I found using Inter Quartile Range method and confirmed by boxplot.
- Workhours attribute has an outlier [5.] which I found by using Mean and 2 Standard Deviation method.
- Sleep quality pattern has some missing data for 45-50 and 55-60 which could be anomaly.

- **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%

- Identifier for the dataset

18-DataB.csv

- a one-paragraph description of each analysis technique you applied to attempt to identify the relationships in the data

Techniques I used, to find relationships in the data are:
- Correlation Matrix
- Correlation Heat Map
- PEARSON Correlation Method
- Kendall Correlation Method
- Spearman Correlation Method

(same as described above)

Relationship:

The correlation methods like Spearman, Kendall and Pearson show that Storage_Capacity (number of containers) is dependent on Ship_length(ft), Ship_beam(ft), Ship_draft(ft).

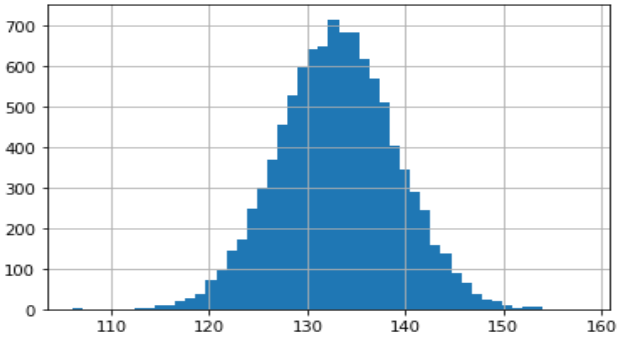
- a one-paragraph description of each pattern you found (add a mathematical description

Pattern in Distance:

PATTERN: Evenly Distributed Scatterplot with
mean 20944.795867
std 3447.750579
min 15000.213313
max 26879.183103

Pattern in Ship_beam:

This attribute shows a bell curve that is a Unimodal Normal Distribution with
mean 133.072053
std 5.976980
min 106.100717
max 158.213400



Mathematical Description

Normal Distribution Formula

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$\mu =$ mean of x
$\sigma =$ standard deviation of x
$\pi \approx 3.14159 \ldots$
$e \approx 2.71828 \ldots$

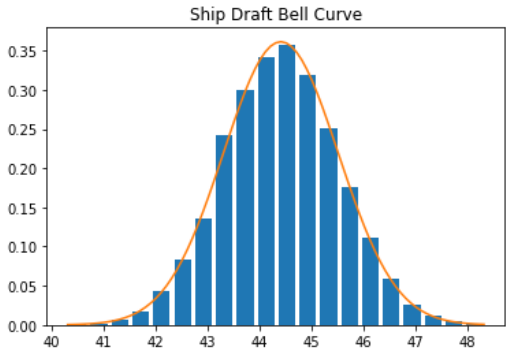
From https://reialcercleartistic.cat/activitats/categoria/festa-de-socis/2018-07-27/?ss=5_6_1_21_37&pp=normal+curve+equation&ii=2700790

Pattern in Ship_draft:

Ship_draft shows a bell curve that is a Unimodal Normal Distribution with

mean 44.406726
std 1.104214
min 40.310158
max 48.333754

Mathematical Description



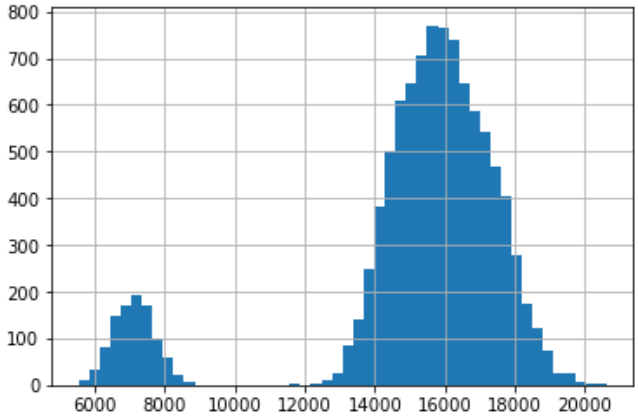
Same Normal Distribution as described above

Pattern in Storage_Capacity

This shows a bell curve that is a Bimodal Normal Distribution. Two bell-curves shows the partition of Normally Distributed data with

mean 15099.460500
std 2944.757835
min 5533.000000
max 20619.000000

Mathematical Description



Same Normal Distribution as described above

- a one-paragraph description of each anomaly detection technique that you applied

Anomaly Detection in Storage_Capacity:

The two bell curves were split into 2 datasets.

Only one Anomaly was found in the first Split of Storage_Capacity and 8 Anomalies were fo und in the second split of Storage_Capacity using Standard Deviation from Mean Method.

Anomaly Detection in Ship_Draft:

Anomalies in Ship_Draft were found using Standard Deviation Method:

[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]

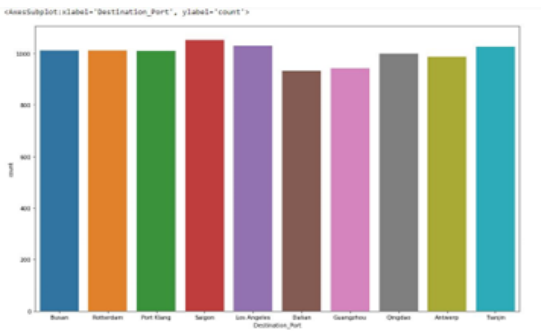
Anomaly Detection in Ship_Beam:

Anomalies in Ship_Beam found using IQR Method & shown with boxplot as well

[111.147938511727,
153.821261505092,
112.990210098791,
106.10071687613,
153.406272364085,
158.213400000499,
106.393449752137,
153.786368776247,
153.119471060909,
153.921585905135,
153.655470780786]

- a brief summary of your analysis that mentions the keys points about the application domain that you see in the data, including the relationships and types of anomalies
- For Dataset2: 18-DataB.csv
- **Application Domain;**
  It is about Container Shipping, which relates to the process of coordinating the transportation of goods by ships smoothly form one place to another and hence there is source and destination ports, holding capacity of a ship with length, beam and draft given. The distance to ship the container also plays a very important factor here.
- **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 outliers 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) using the correlation methods like Spearman, Kendall and Pearson Methods as well as Correlation Matrix and Heatmap.

		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%

- Identifier for the dataset

21-DataA.csv

- a one-paragraph description of each analysis technique you applied to attempt to identify the relationships in the data

Techniques I used, to find relationships in the data are:
- Correlation Matrix
- Correlation Heat Map
- PEARSON Correlation Method
- Kendall Correlation Method
- Spearman Correlation Method

(same as described above)

Relationship:

The correlation methods like Spearman, Kendall and Pearson show

- BMI, Weight, (also height)
- BMI and Calories
- Calories and weight

- a one-paragraph description of each pattern you found (add a mathematical description

Pattern in Membership Type:

Data is Evenly Distributed between "Monthly" and "Annually" column

Number of Monthly Membership Subscriptions are almost the same as number of Annual Membership Subscriptions

Mathematical Description

Monthly Membership ≈ Annual Membership (approx.)

Pattern in Age:

Histogram plot of Age attribute and we can see the number of instance for 60 is very high.

Number of instances of Age 60 is twice as many as the other ages

Mathematical Description

Instances of Age(60) = 2 * Age (10-59)

Pattern in Weight:

Weight attribute is Evenly Distributed with
mean 52.109158
std 27.867158
min 3.424272
max 99.984035

Pattern in Height:

Height attribute is Evenly Distributed with
mean 1.701217
std 0.230416
min 1.300066
max 2.099839

Pattern in Gender:

Number of instances for male is the approximately the same as the number of instances for female

Mathematical Description

Gender(male) ≈ Gender(female)

Pattern in Life_Style:

Number of instances for Active Lifestyle is approximately the same as the number of instances for Low_Active Lifestyle and approximately the same as number of instances of Sedantary Lifestyle

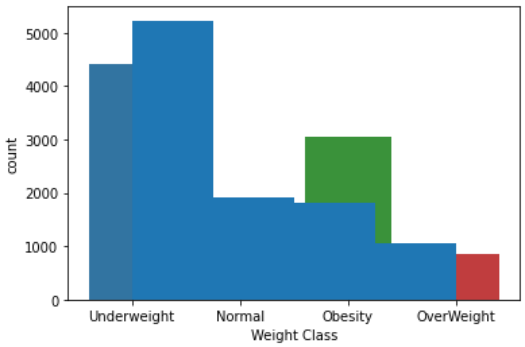
Mathematical Description

Lifestyle(Active) ≈ Lifestyle(Low_Active) ≈ Lifestyle(Sedantary)

- a one-paragraph description of each anomaly detection technique that you applied

Anomaly Detection in BMI:

73 Anomalies in BMI were found using IQR Method



Anomaly Detection in Calories:

687 outliers detected using Median Deviation& Boxplot

Anomaly Detection in Weight:

Anomalies in Weight were found using Standard Deviation Method

[3.42427212 3.42670244 3.4536528 3.45475114 3.45640694 3.46146987 3.46285403 3.4694793 3.47163962 3.48280123 3.48744976 3.50063703 3.51942424 3.52604204 3.52780794 3.57449645 3.57739774 3.57774571 3.57848522 3.58158939 3.59672497 3.60111605 3.60143335 3.60251328 3.62088078]

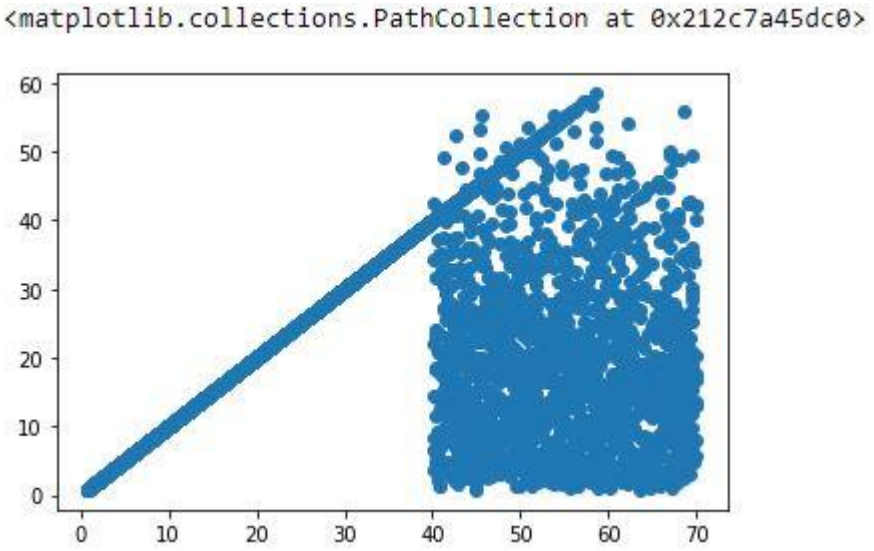
- a brief summary of your analysis that mentions the keys points about the application domain that you see in the data, including the relationships and types of anomalies

- For Dataset3: 21-DataA.csv

- **Application Domain:**
  This dataset is all about employee fitness programs which is decided upon the factors such as Age, Height, Weight, Gender, membership type subscribed by the employee, Body-Mass Index, Calories Intake, Lifestyle of the Employee, whether Sedentary, active, or low active. There is a 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.



  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.
  - 73 Anomalies in BMI were found using IQR Method
  - 687 outliers in Calories detected using Median Deviation Method

- **Relationship:**
  BMI = weight/height^2

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

  Correlation Methods show the relationship between

1. BMI, Weight, (also height)
2. BMI and Calories
3. Calories and weight