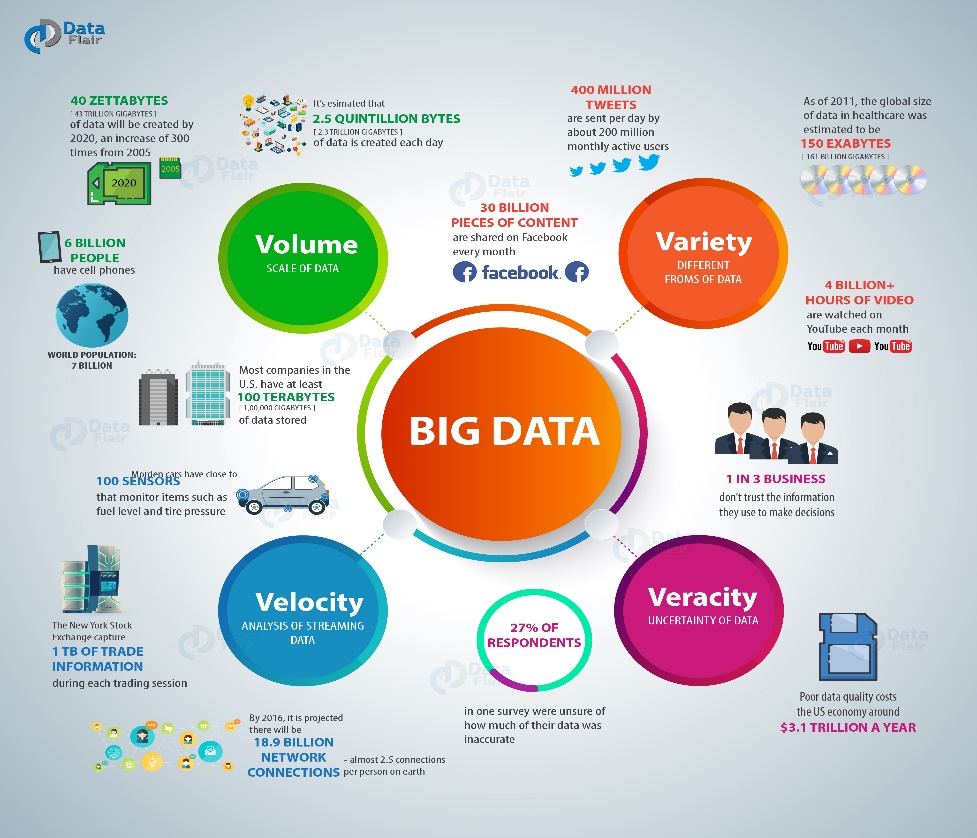
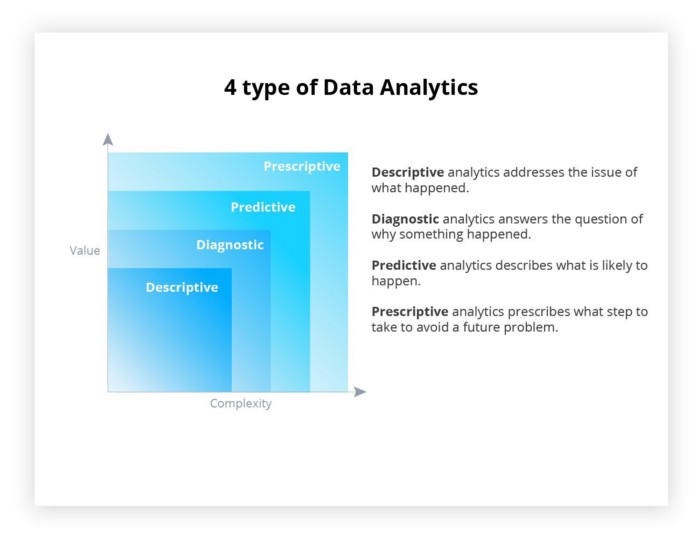
**TASK ONE – NARRATIVE**



5 Vs of bad data (Patidar, 2022)

The best approaches in big data according to me are **Descriptive Analytics, Diagnostic Analytics, Predictive analysis (Machine learning using Support Vector Machine), Prescriptive Analysis**

**Data Science problem:** One of the biggest food brands PepsiCo wanted to improve their supply chain moving the ready products to supermarket shelves quickly and efficiently. Their main motivation was to boost sales numbers in different regions and avoid losses caused due to understocking, overstocking, and product loss.



4 types of Big Data Analysis approaches (onix-systems, 2022)

**Descriptive Analysis:** is used to address the question of ‘what happened’ using data (onix-systems, 2022). This is accomplished using data (current/past). PepsiCo can use the descriptive approach to analyse their past shipping activity, sales activity from retail partners, helping them to understand their market position and KPI’s. This will give PepsiCo insight into how their business is doing. It cannot predict future trends/events, as it only looks at old data on the surface. It can bring attention to factors that might be leading to the company’s success or failure. The analysis can help the sales departments easily visualize the company’s performance, sales number, growth, losses.

Advantages: Analysts can use it without having a complete technical understanding of all the analytical or statistical concepts. Prebuilt resources can be used to perform the analysis, but the data has to be collected by PepsiCo which can be passed into the software. It can run through huge datasets and find the data features the company is interested in without it being hard to understand for the average person (e.g., shipping and sales data).

Disadvantages: Descriptive analysis method works well only on data from the past. The analysed data only answers the “what’s happening” part of the data but does not find the root cause of the issue.

Limitation: This approach cannot predict future events/trends since it only uses data from the past

**Diagnostic Analysis:** Used to determine the cause of an event based on trends and correlations between given variables, using historical data. Problems such as stock being lost on a particular shipping route, then the reason for this could be narrowed down using root cause analysis on the shipping dataset. Essentially this approach can be used to answer “why” something happened. PepsiCo could use this approach to find where most stock is lost and improve it. Or they could use the sales data from its partners to check the region and the flavour of the drink that is not selling well. And based on the data, either choose to advertise it more or improve it in the given region. Analyst must understand some concepts like correlation and causation and diagnostic regression analysis.

Advantage: Companies can with the right data get to the root cause of success or losses in the business. Since machine learning models are used for detecting patterns, they are very efficient and accurate along with eliminating bias and not being misled by causation instead of correlation.

Disadvantage: Is more complex than descriptive analytics, also not all types of business data can be analysed using diagnostic analytics.

Limitations: Machine learning cannot be fully relied on to produce the best results, human intervention is also important to put the output into context.

**Predictive analysis:** Predictive analysis can use big data Pipelines on the dataset using techniques like statistical modelling, data mining and machine learning to predict future trends or outcomes. These insights can give the management the ability to identify future threats to the company and act on them, or make business decisions that will lead to more profits in the future. PepsiCo can use Predictive analysis to learn more about product consumption in different areas of the world around different times of the year. Using this, they can then decide when to manufacture more drinks, which type of drink, and regions to ship them to. This will increase the profitability of the company. It can also be used to predict problems that the company might run into and mitigate them. Predictive analytics uses historical data to perform forecasting the likeliness of an event happening in the future, the magnitude of the events on the business and a time frame of when it might occur. Machine learning models such as Decision trees, Regression, and neural networks can be used as training models and then used to predict the future results

Advantages: Insight gained can be used to increase the company’s growth and revenue. It can help the company management prepare for future changes in shipping lines or manufacturing upgrades to improve productivity. The predictions will give the company a good idea of how much product their shipping line is going to need

Disadvantages: The quality of the data used to train the ML model needs to be accurate and relevant to get the best results.

Limitations: The data that goes into the model needs to be of good quality to get the best predictions.

**Prescriptive analysis:** This is the evolution of Predictive analysis and Descriptive analysis as it can recommend a course of action for the company. This is a complex process so it is PepsiCo could not use it for short term day-to-day operations. To make this work PepsiCo would need data from internal sources and external partners to be used with machine learning models. PepsiCo can use it to get stock numbers they would require for shipping during different times of the year, they ask for recommendations on which region is going to have a higher demand for their products. Also based on current routes data, build more efficient new routs that would be even better.

Advantages: It can be used to get recommendations on a wide variety of business decisions. Since the processing is done by computers the risk of human error is low.

Disadvantages: They cannot replace humans in making the final decision, it is just a tool used to help in the process.

Limitations: Without huge amounts of data to use for training the results can be inaccurate. Sometimes the system cannot account for all variables related to the data.

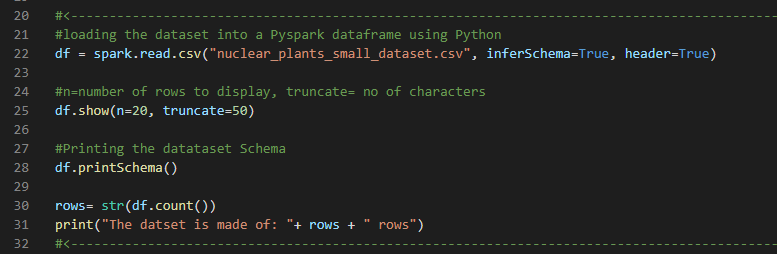
|  |  |  |  |
| --- | --- | --- | --- |
| Big Data Approach | Advantages | Disadvantages | Limitations |
| Descriptive Analysis | * Analysts can perform descriptive analysis without having the complete technical knowhow of all the analytical or statistical concepts * It can answer most queries about the business KPI and performance * It is very really well with huge datasets | * This method works well on the data only from the past * Cannot find the root cause of the cause of an event or trend | * This method cannot provide establish relationships for data pertaining to future events * Only works with data from the past |
| Diagnostic analysis | * Can find correlations and relationships between given variables * Uses a range of techniques to work such as data mining and discovery, statistical analysis, algorithms sensitivity and error analysis * Transforms complex huge datasets into manageable and visualisable data for the average person | * Cannot be used to predict future trends, since it mainly focuses on the data from the past and its relationship * Not all business use cases can rely on diagnostic analysis to base their decisions on | * Diagnostic analysis focuses on past events; hence it is not able to deliver any useful future insights * Some human intervention is needed to give context to the output data |
| Predictive analysis | * Information can be used to form a strong business strategy leading to the company’s growth * Hep the company’s management be better prepared for future events and be better prepared for future challenges * Can predict how much product the company is going to need to supply in the future * Can be used to perform risk and loss reduction * Help to forecast inventory needs | * The quality of the data used is very crucial to the quality of the insight received from the analysed data * The data collection process is very unreliable as some sources of data might not have consistent standards, also people might lie when asked certain types of questions | * If the data used for training the model is bad then the Predicted analysis will be very inaccurate and cannot be relied on * It is not 100% accurate since it is based on probabilities |
| Prescriptive Analysis | * Low risk of human error * It can provide advice on possibilities in many areas of the business * Its output can give management insight into making good decisions for the uncertain future | * Uses complex machine learning algorithms which increases its complexity * Still requires humans to make the final decision * Poor quality input data can result in forecasts that are faulty | * Needs large amounts of data to be fed into the system to get useful results * The data need to be of high quality |

**TASK TWO – ANALYSIS**

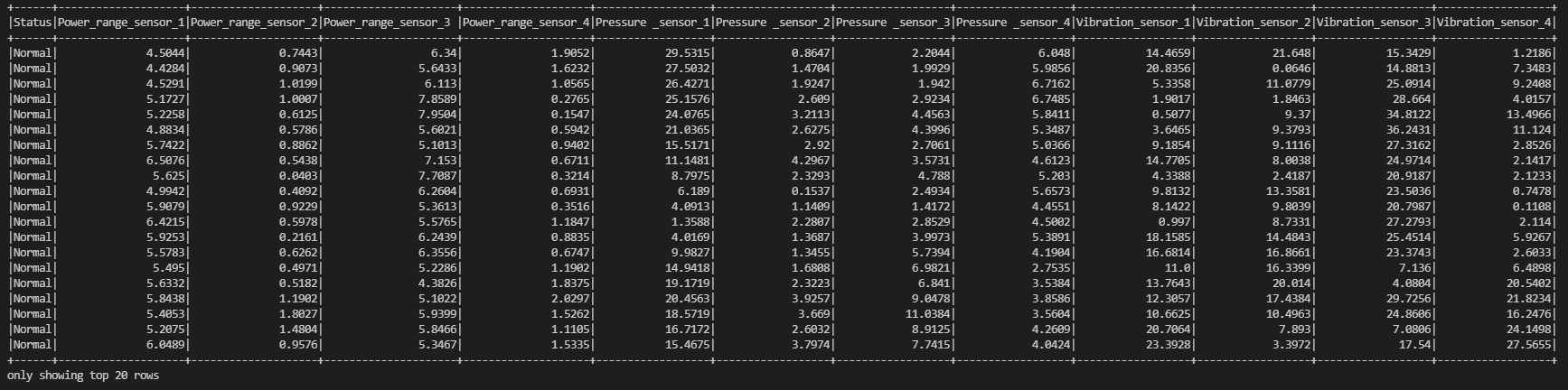
**Provide a critique of each step taken**

**Section 1: Data summary, Understanding and Visualisation**

The dataset is held in a CSV file. As can be seen in the code below the data has been read into the Pyspark DataFrame named **df** using the spark.read.csv() command to read the file. The rows variable has been used to convert the output of df.count(), which is int value into a string value. Lastly, the print statement displays the number of rows in df to help visualise the data.

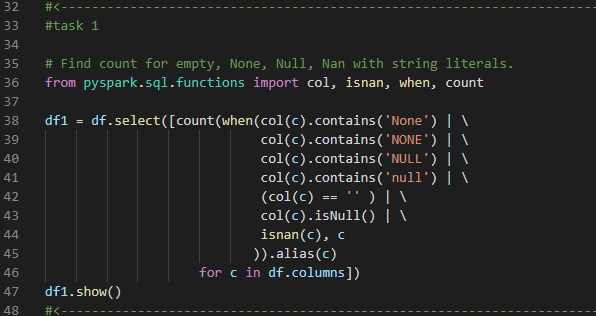


The below images show what first 20 rows of the DataFrame and it is printed using df.show() command.

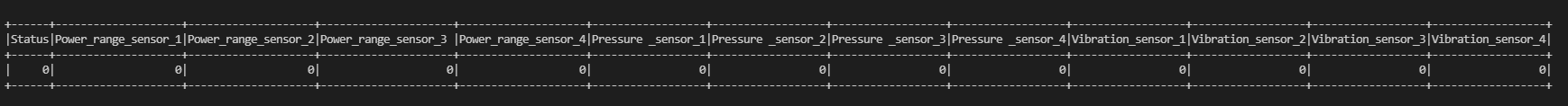


**Task 1**

For the provided reactor dataset, **there were no missing values** in it. Using the code shown in the image below the data set was checked for “None”, “NONE”, “NULL”, “null”, and blank values(“ ”). The missing data vales were counted for each column using the count(when(col(c).contains(“values”)))method and saved the count value into a new DataFrame called **df1**



Using the code above we get df1. Since the DataFrame does not have any missing values, all columns will display the values 0, shown by the table below.



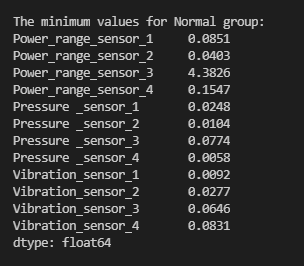
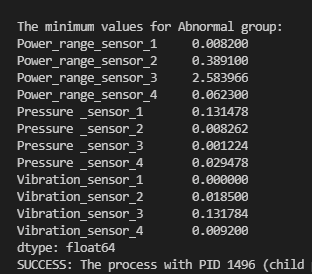
One way to deal with missing values is to remove the rows or columns that have any or more than half of the values missing (Kumar, 2022). This will improve the model accuracy but this could lead to loss of information if there are excessive missing values in comparison to the complete dataset. Generally, it is better to avoid this method but when there are few missing values it can be used.

Second method is to replace the missing value with either the mean or median value of the column. The advantage here is that it prevents loss of data. But, it can cause data leakage,also it only works with numerical values and not with string type values.

Another method to fill missing values is by predicting them using the correlation of the variable containing the missing vlaue, using values of other features around it that are not null or missing. “A regression or classification model” (Kumar, 2022) can be used to predict missing values based on if the missing values are continuous or categorical.

**Task 2**

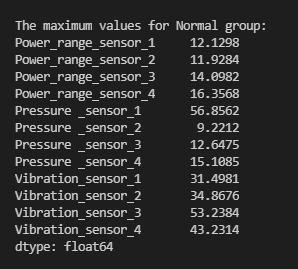
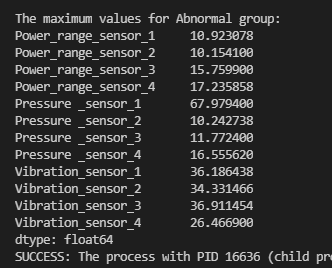
To calculate the minimum values the DataFrame only need to consist of numeric values. So, the ‘Status column is dropped’ for both group’s DataFrame’s and they are saved to **normalDropDF** and **abnormalDropDF** df variables. To calculate the minimum for features of both dataframes, the dataframe.min() Pandas function. The values minimum value’s series for both groups are shown below:

1Minimum values of features for Normal group

2Minimum values of features for Abnormal group

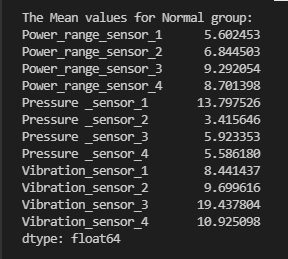
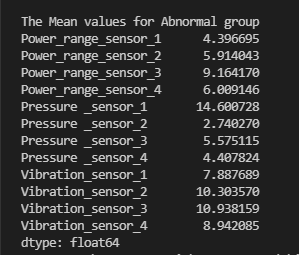
Next to calculate the maximum values can be accomplished by using dataframe.max() Pandas df function that return the maximum of the values in each feature:

3-Maximum values of features for Normal group

4-Maximum values of features for Abnormal group

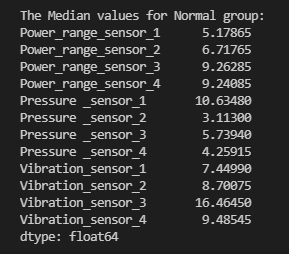
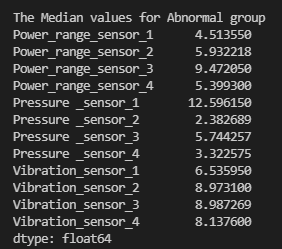
The mean of the features were calculated by used the dataframe.mean() function using pandas and series returned are shown below:

5-Mean values of features for Normal group

6-Mean values of features for Abnormal group

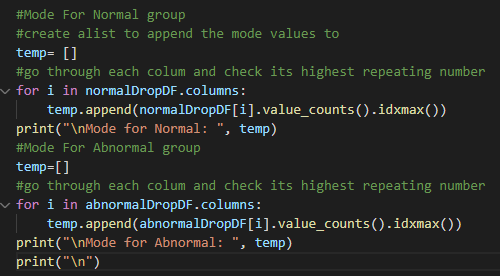
The median was calculated using the dataframe.median() function, using pandas and the median values returned are shown below:

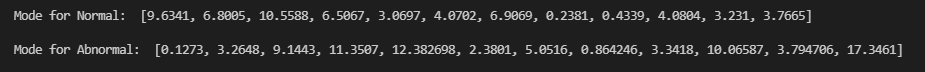
7-Median values of features for Normal group

8-Median values of features for Abnormal group

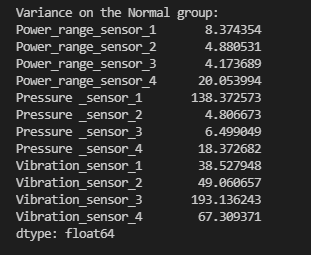
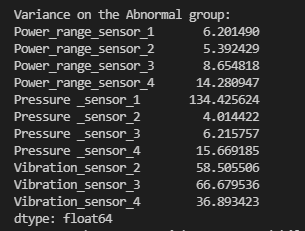
To calculate the mode (most repeated value) of the features a list temp has been created to store the values. Then, using a for loop we iterate through the columns of the dataframe’s and use dataframe[i].value\_counts().idxmax() where value\_counts() get the most repeated values in the column and a randomly selected value among these is appended to temp.



The results for mode in the dataframe’s are shown below:



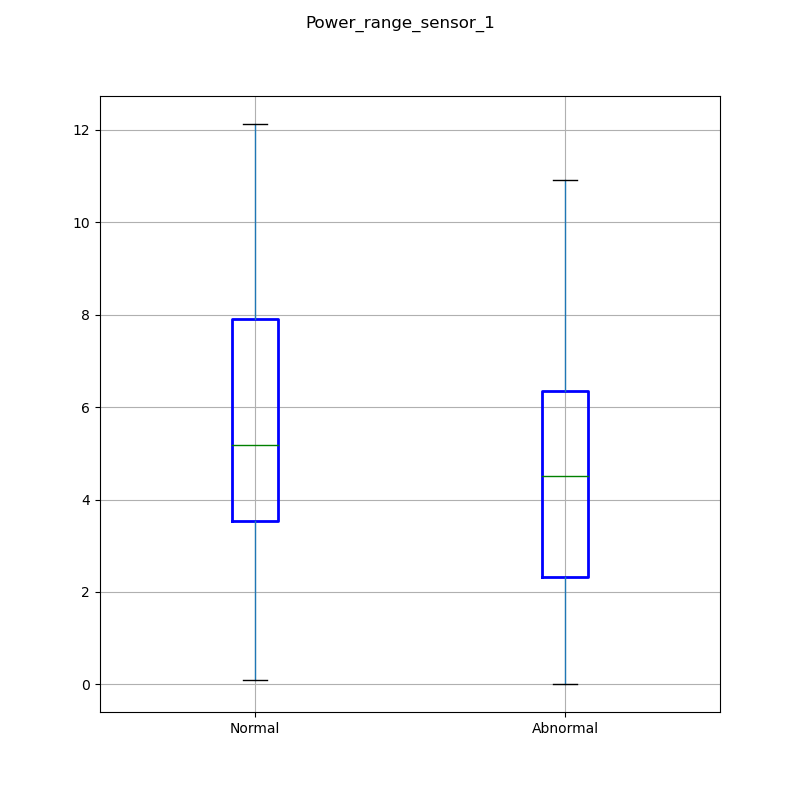
The variance of the DataFrame is calculated using dataframe.var() function using pandas. It is used to calculate the degree of spread in the dataset, the larger the variance the bigger the bigger the data spread is going to be.

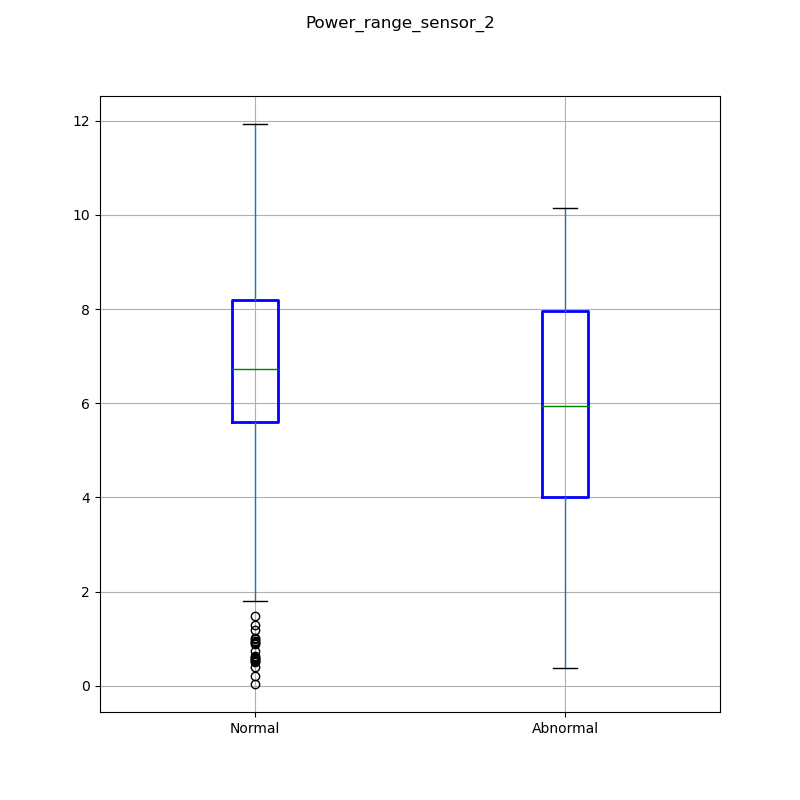
 

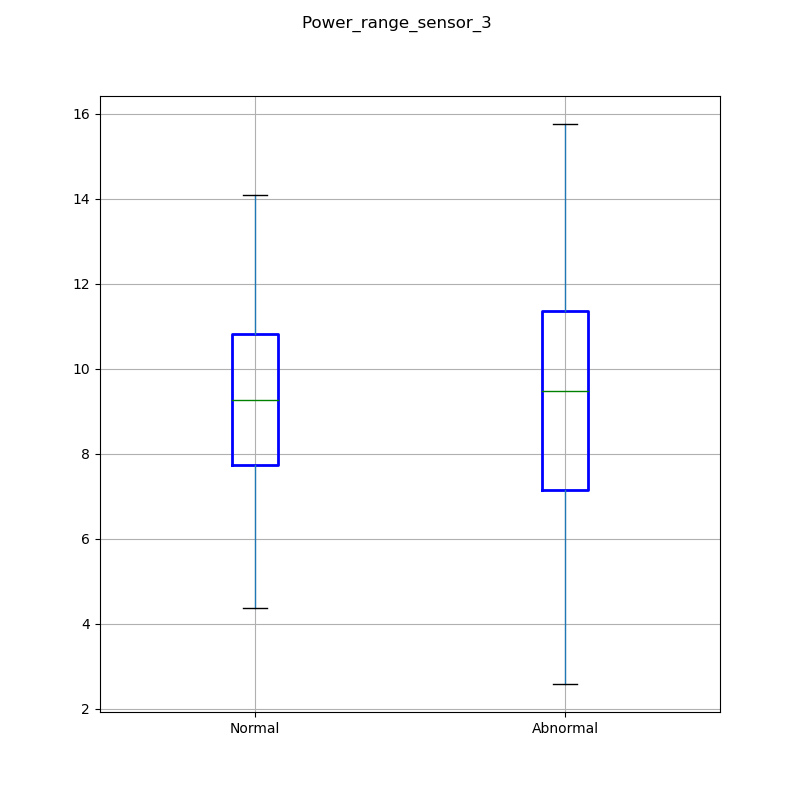
9-Variance values of features for Normal group

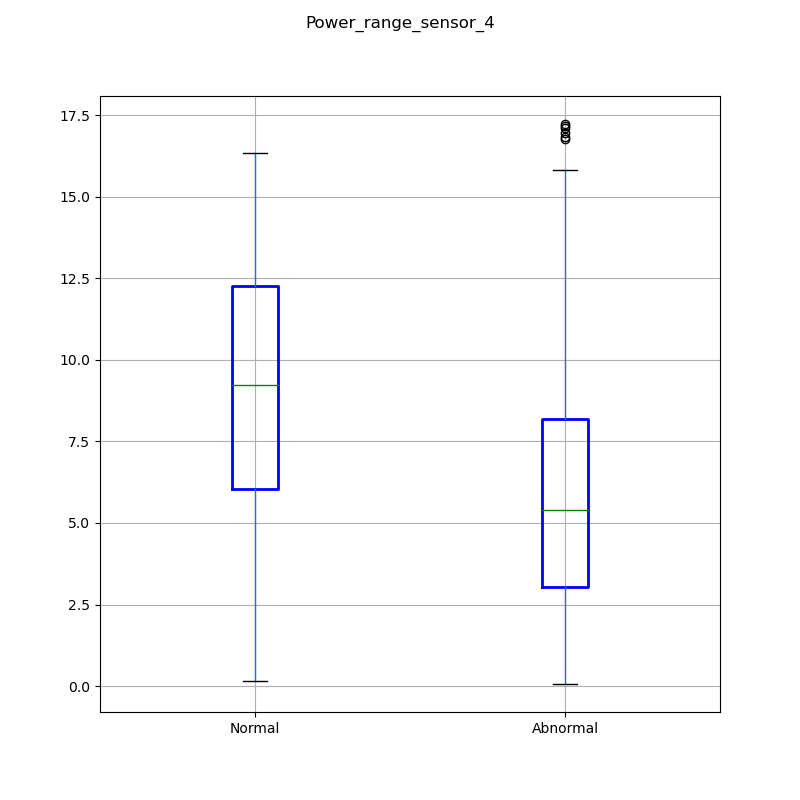
10-Variance values of features for Abnormal group

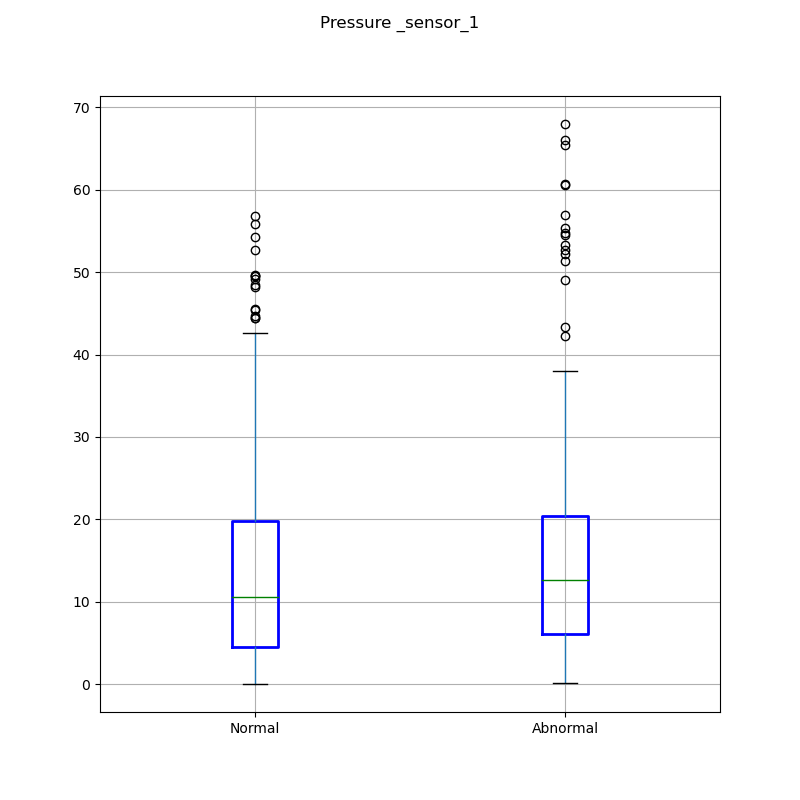
Box plots for features of the dataset:

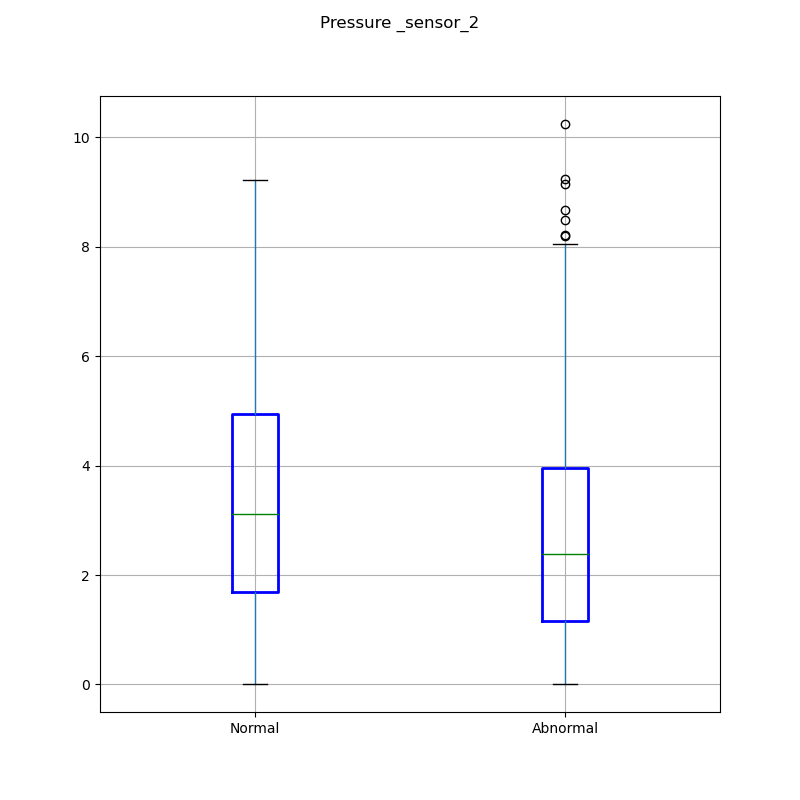


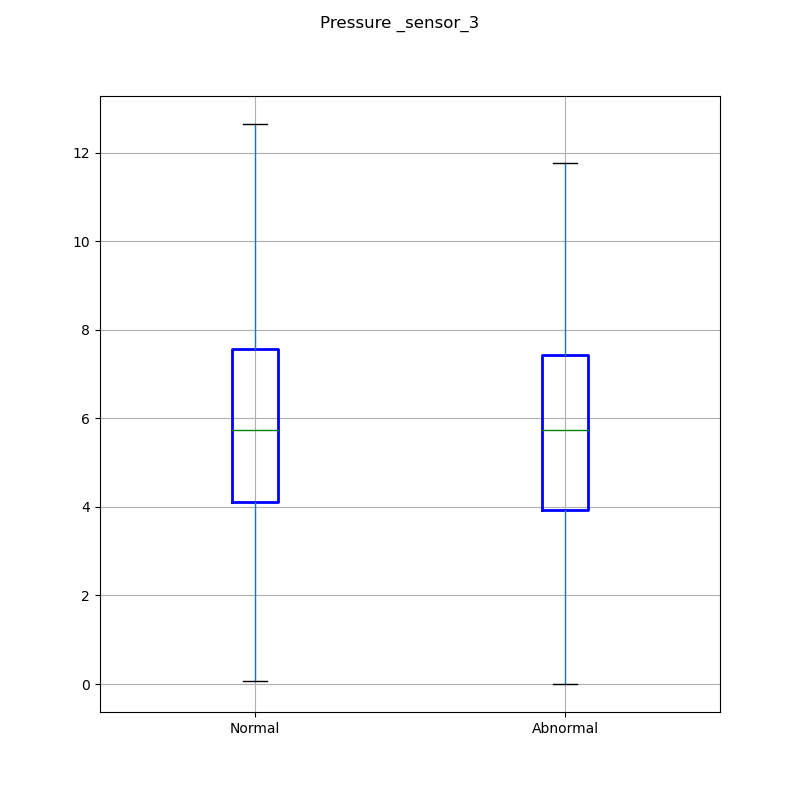


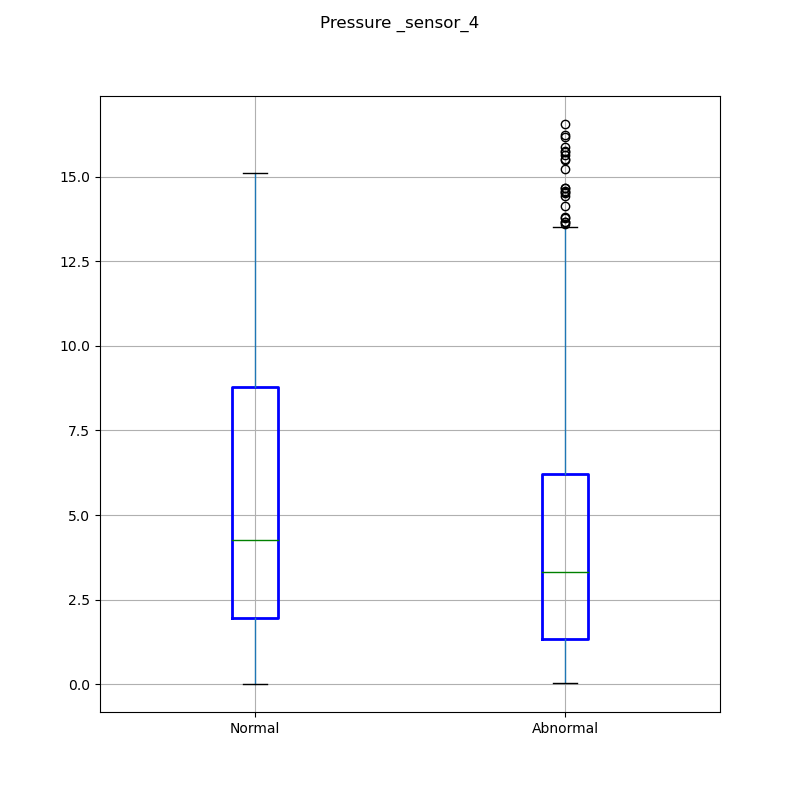


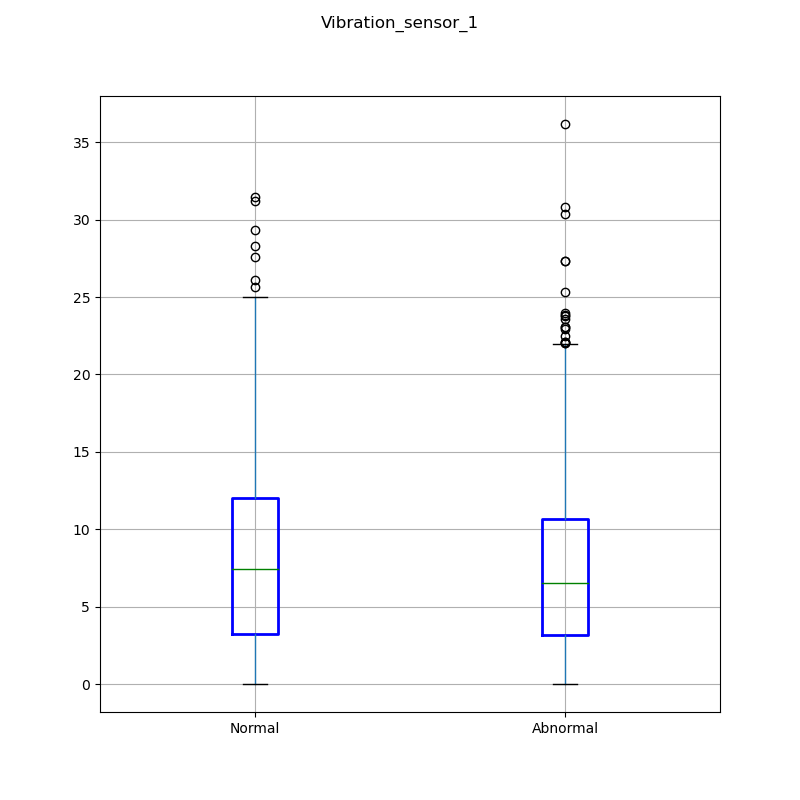


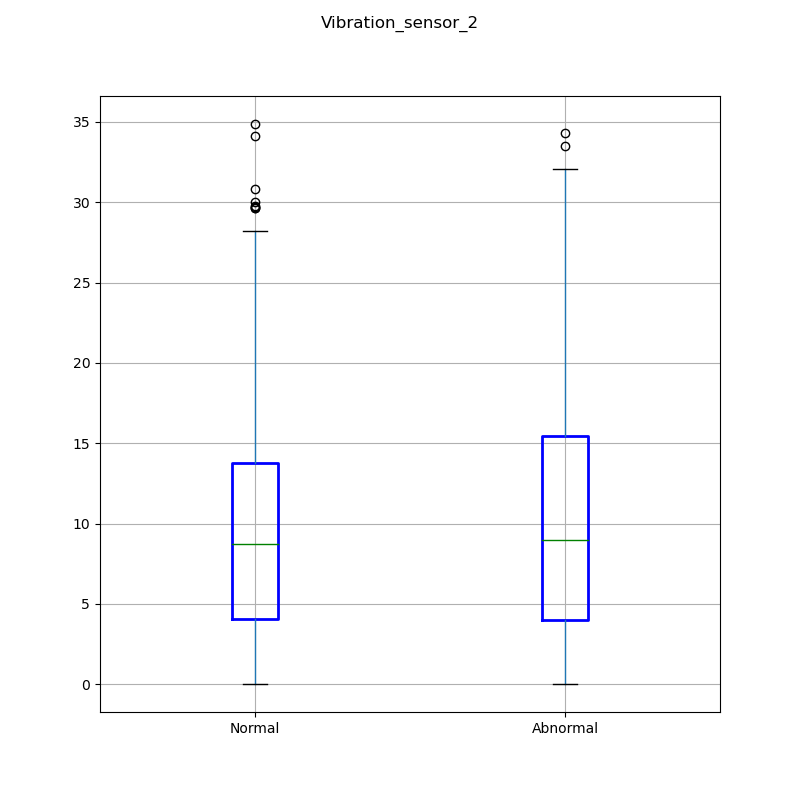


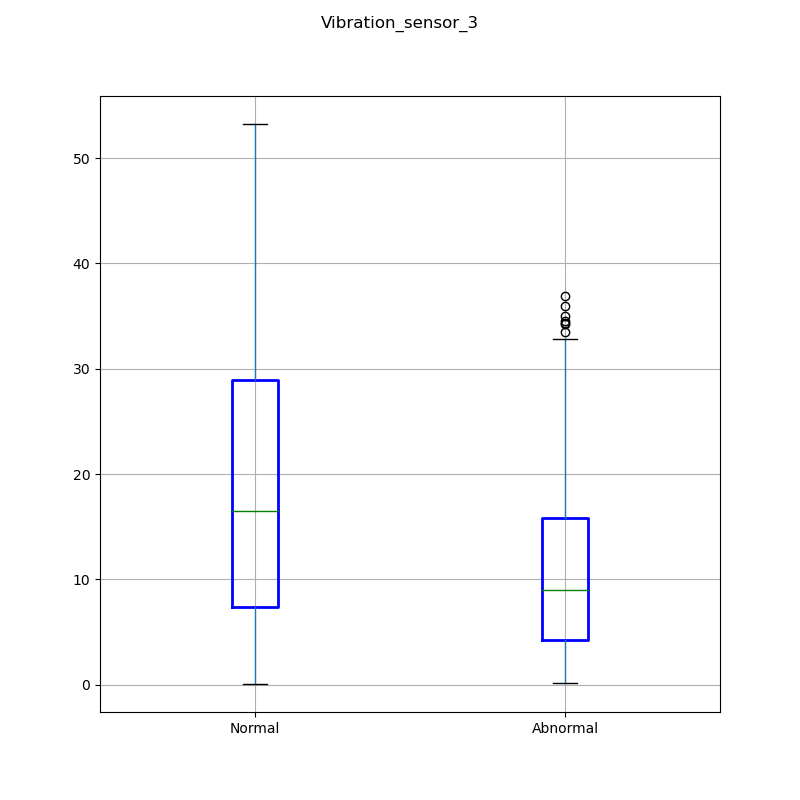


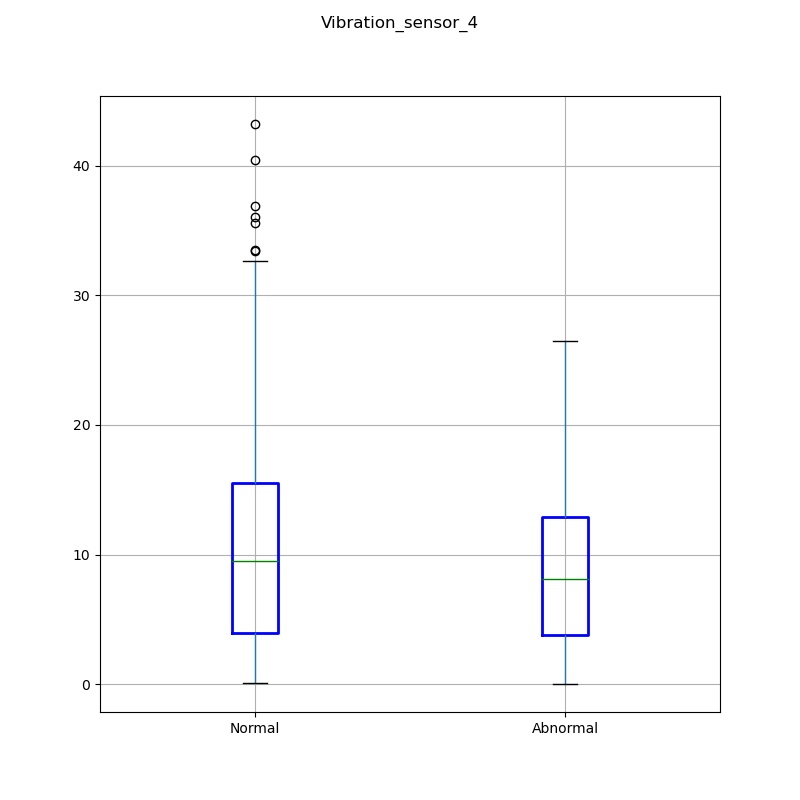












**Task 3**

Correlation shows the normalised measure of the covariance between the given two random feature columns. It makes it easier to understand their statistical dependence by providing quantitative measurement. The correlation measure goes from 0 to 1 where:

0= there is no correlation at all

1= they are perfectly correlated

[**https://www.geeksforgeeks.org/python-pandas-dataframe-corr/**](https://www.geeksforgeeks.org/python-pandas-dataframe-corr/)

**Section 2: Classification & Big data analysis**

**Task 4**

**References**

sparkbyexamples.com (2022) *PySpark – Find Count of null, None, NaN Values*. [online] Available from https://sparkbyexamples.com/pyspark/pyspark-find-count-of-null-none-nan-values/ [Accessed 24 January 2022].

Kumar, S. (2022) *7 Ways to Handle Missing Values in Machine Learning*. [online] Medium. Available from https://towardsdatascience.com/7-ways-to-handle-missing-values-in-machine-learning-1a6326adf79e [Accessed 24 January 2022].

Burris, P., Finos, R., Kobielus, J. and Gilbert, G. (2022). *Wikibon's 2018 Big Data and Analytics Market Share Report - Wikibon Research*. [online] Wikibon Research. Available from https://wikibon.com/wikibons-2018-big-data-analytics-market-share-report/ [Accessed 29 January 2022].

GetSmarter Blog. 2022. *Big Data Analysis Techniques - GetSmarter Blog*. [online] Available from https://www.getsmarter.com/blog/career-advice/big-data-analysis-techniques/ [Accessed 29 January 2022].

Patidar, S. (2022). [image] Available from https://dz2cdn1.dzone.com/storage/temp/9159302-big-data-infographic.jpg [Accessed 29 January 2022].

Durcevic, S. (2022). *What is Ad Hoc Reporting and Analysis? Meaning & Examples*. [online] BI Blog | Data Visualization & Analytics Blog | datapine. Available from https://www.datapine.com/blog/ad-hoc-reporting-analysis-meaning-benefits-examples/ [Accessed 29 January 2022].

Kopanakis, J. (2022). *5 Real-World Examples of How Brands are Using Big Data Analytics | Mentionlytics Blog*. [online] Mentionlytics. Available from https://www.mentionlytics.com/blog/5-real-world-examples-of-how-brands-are-using-big-data-analytics/ [Accessed 29 January 2022].

onix-systems. (2022). *4 types of data analytics Approaches*. [image] Available from https://onix-systems.medium.com/things-you-should-know-about-types-of-data-analysis-40aacf06f13d [Accessed 30 January 2022].