**TASK ONE – NARRATIVE**

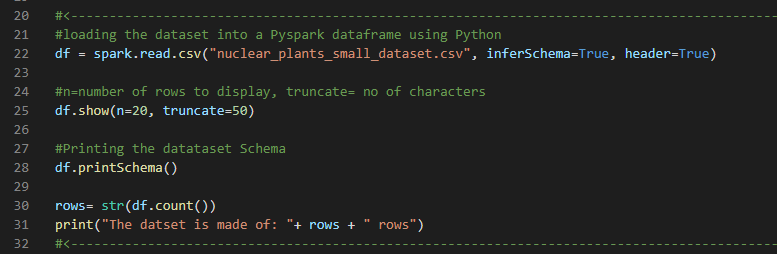
Start here

Talk about ETL- extraction, transformation, L

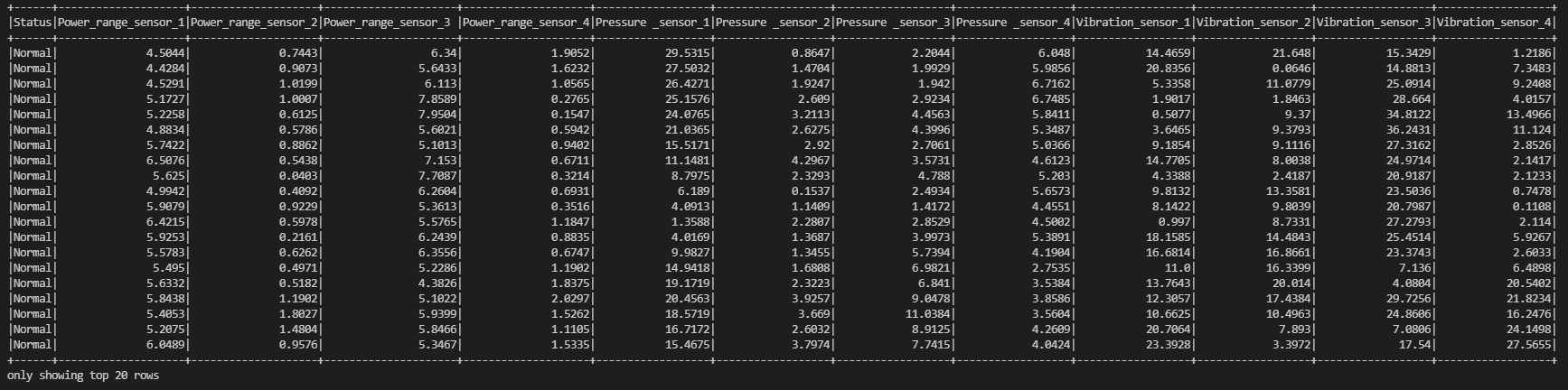
**TASK TWO – ANALYSIS**

**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.

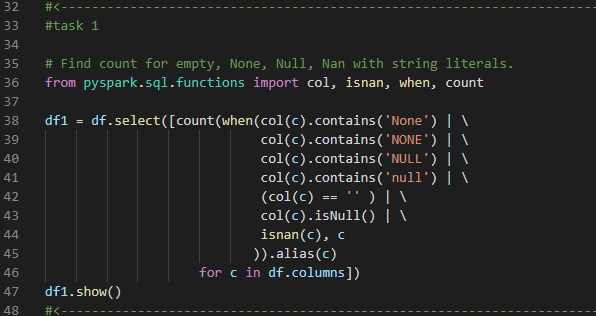


The below images show what first 20 rows of the DataFrame and it is printed using df.show() command.

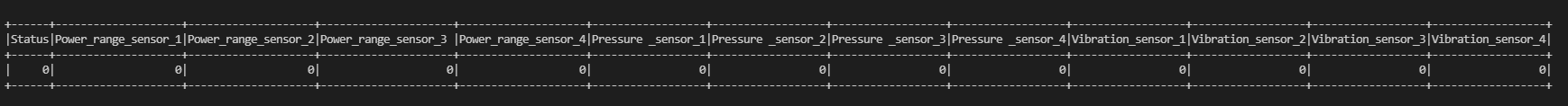


**Task 1**

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.



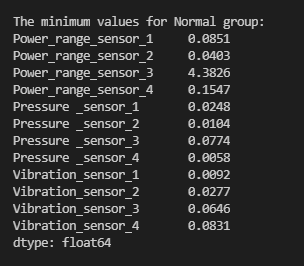
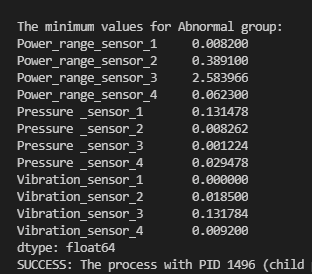
One way to deal with missing values is to remove the rows or columns that have any or more than half of the values missing (Kumar, 2022). This will improve the model accuracy but this could lead to loss of information if there are excessive missing values in comparison to the complete dataset. Generally,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

Minimum values of features for Normal group

Minimum values of features for Abormal group

Next to calculate the maximum values can be accomplised by using dataframe.max()

**Task 3**

**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].