**INFORMATICS INSTITUTE OF TECHNOLOGY**

**In Collaboration with ROBERT GORDON UNIVERSITY**

**ABERDEEN**

**Course: Artificial Intelligence and Data Science**

**Module Leader: Mohamed Ayoob**

**CM2606: Data Engineering**

**Assignment Type: Individual**

**Student Name: W.A.C.I.V Bandara**

**IIT ID – 20222453**

**RGU ID – 2311689**

Table of content

Introduction ##

In this coursework include analysis and prediction of HCHO gas distribution across the various districts in Sri Lanka using provided data. Three datasets given in the starting point of the coursework that including data of different districts. For further analysis I used a Sri Lankan weather dataset which include different attributes such as wind direction, elevation, temperature to make a deep analysis and finally made a timeseries forecasting for gas distribution in upcoming year and visualized predicted outputs using Power BI.

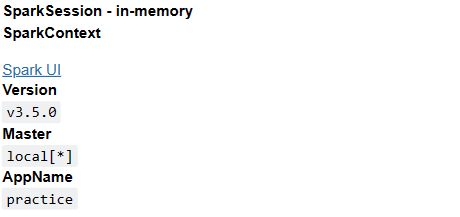
Dataset Selection

In the beginning of the coursework data set was given and it include information about HCHO distribution in Sri Lanka. For further analysis I used a data set from Kaggle.

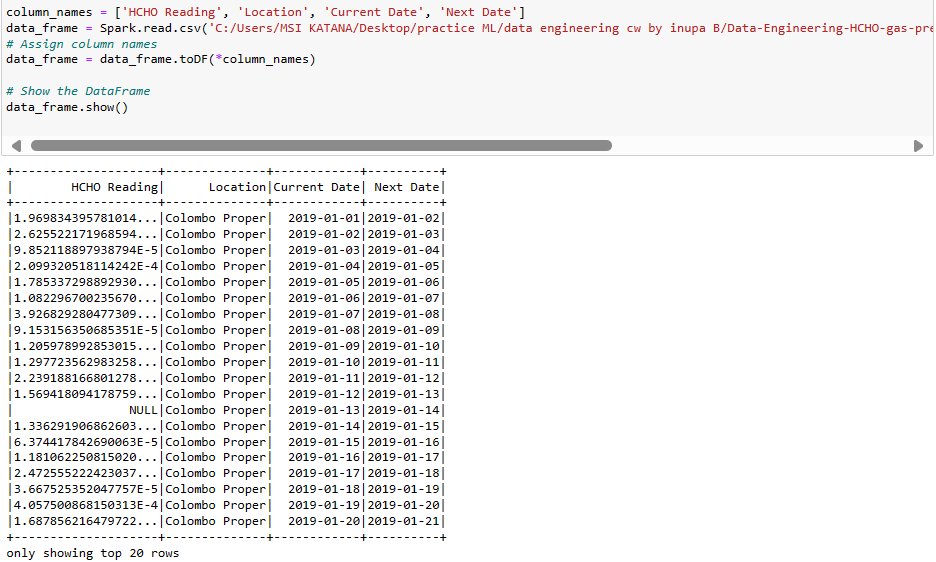
Data Integration and Preprocessing part

For data cleaning and preprocessing part, I used pyspark. Using pyspark first I loaded given data sets separately and then cleaned and handled the missing values and did the preprocessing part separately and finally integrates those datasets and made a single data frame.

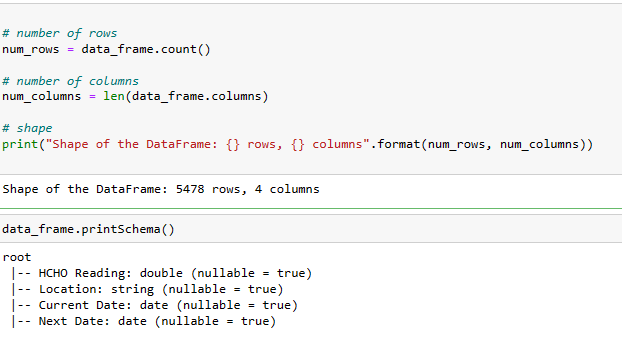
* Making a spark session to start the process



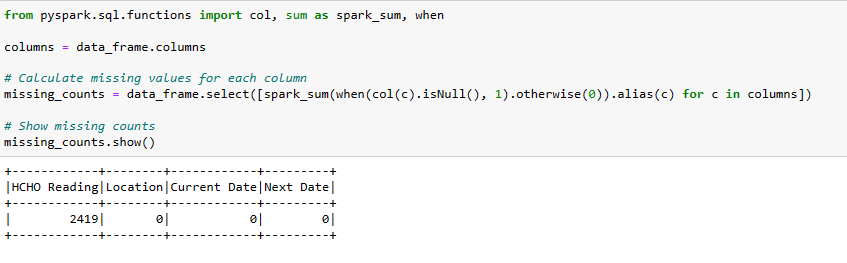
* Loading data frames and assigning column names for the data frame



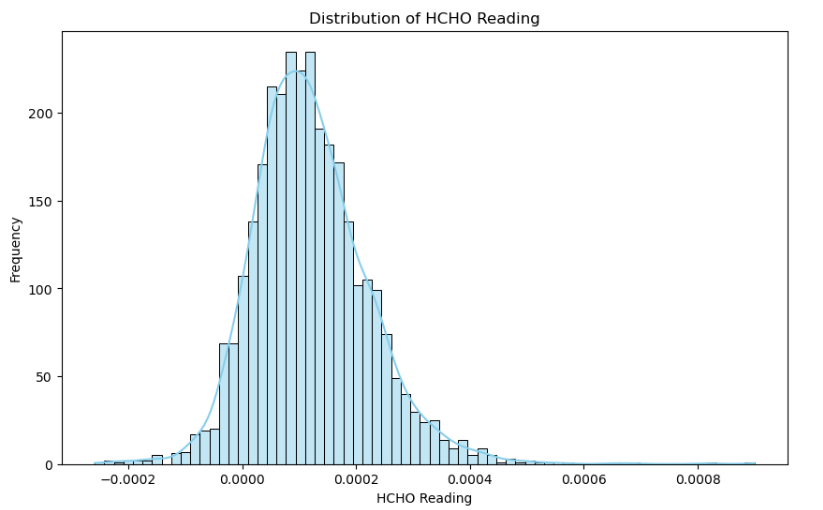
* Checking data types of each attribute



* Calculating the amount of missing values

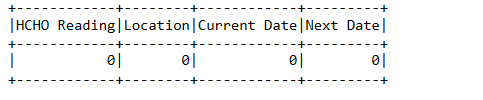


Before handling missing values in HCHO Reading column I extracted that column separately and converted it into a pandas data frame and visualized the distribution of data.

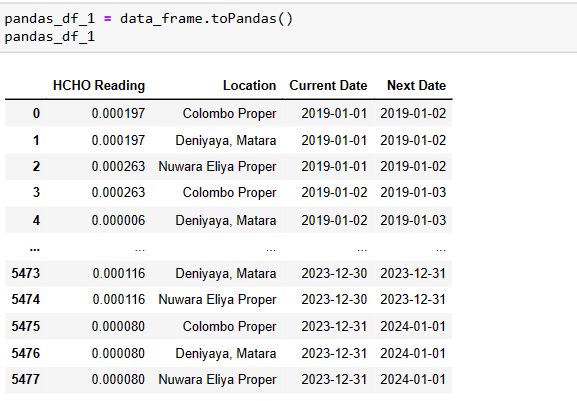


After confirming that the data is distributed like a normal distribution I decided to use forward windowspec and backward windowspec methods to fill the missing values.

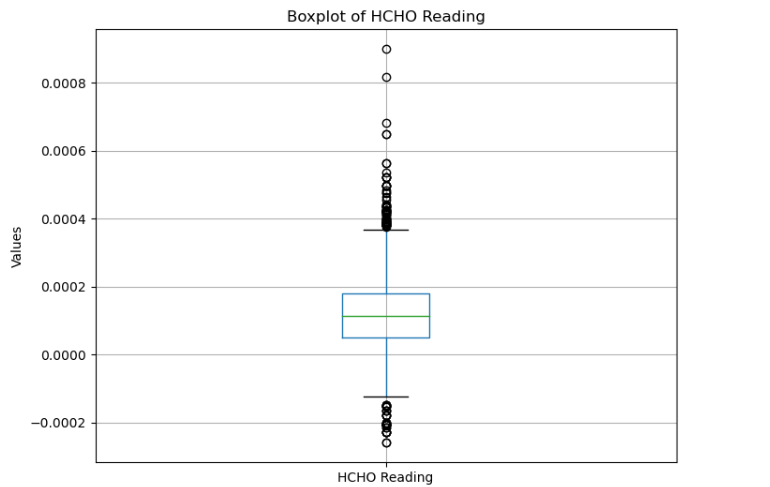
* Confirmed no missing values.

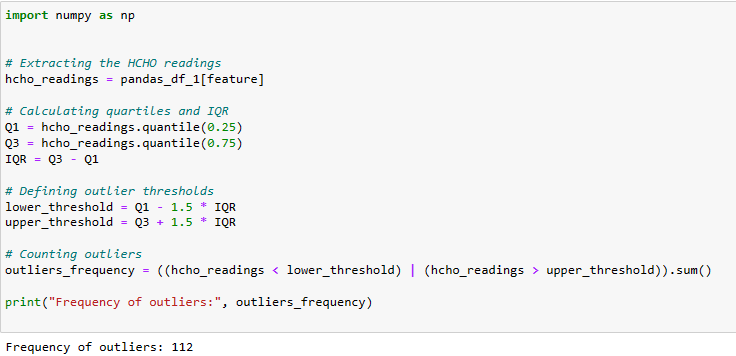


After handling missing values, I converted the data frame into a pandas data frame to continue further inspection.

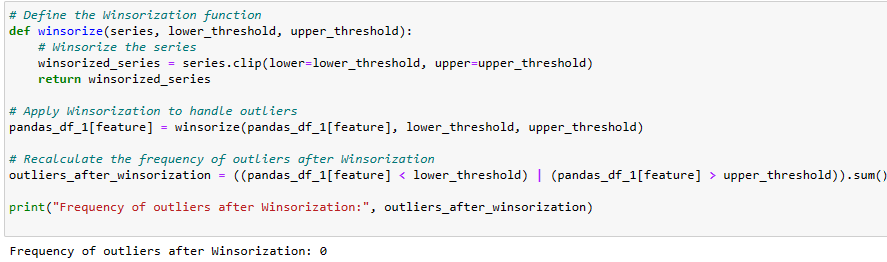


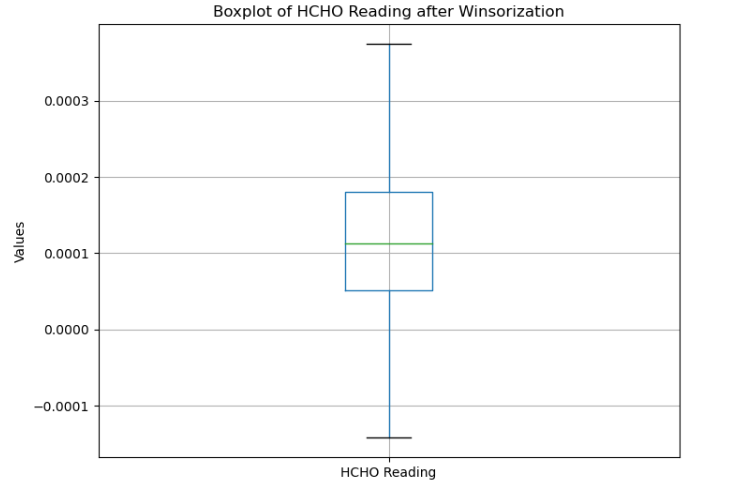
In the pandas data frame first selected the HCHO Reading column and then visualized outliers and calculated the number of outliers.





Since the number of outliers are very smaller when comparing with the original size of the data set, I decided to use winsorization method to winsorize outliers to upper and lower boundaries.



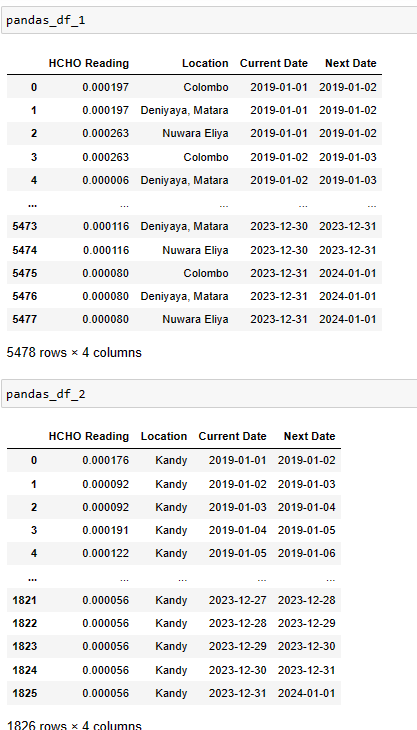


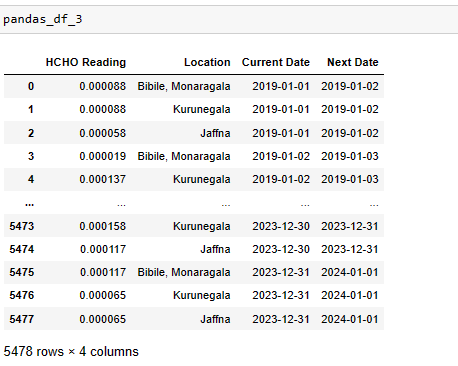
Removed the “Proper” part that include in several districts and converted it into the normal format.



Then saved the data frame for use in future.

Continued the above process for the other data sets in a similar way and saved those datasets as pandas data frames to use in future.





After preprocessing each data set individually, I concatenated all datasets and made a final dataset.

