**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.

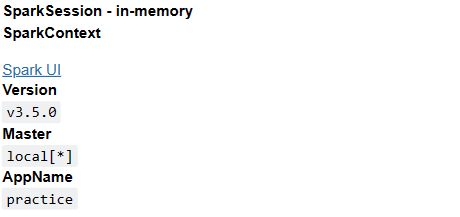
Dataset Selection

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.

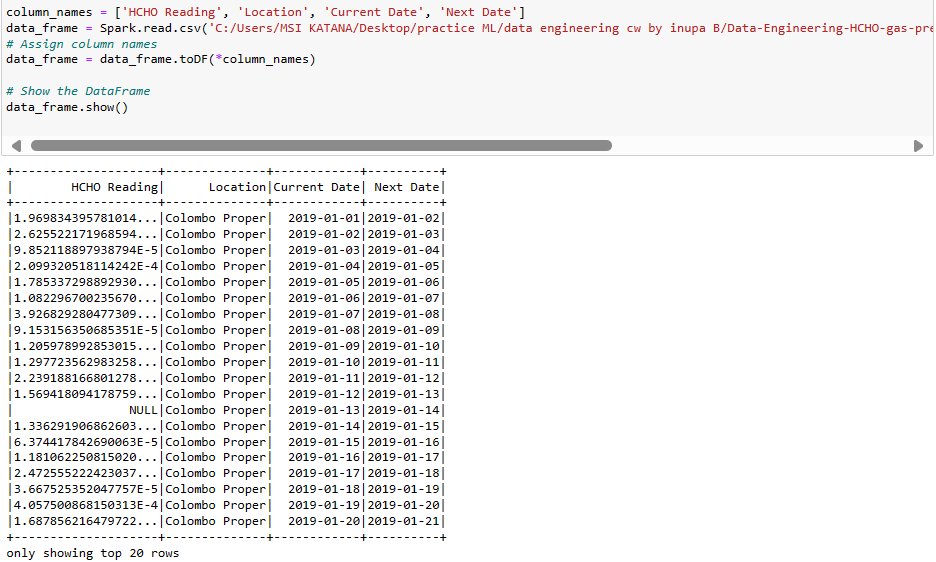
Data Integration and Preprocessing part

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.

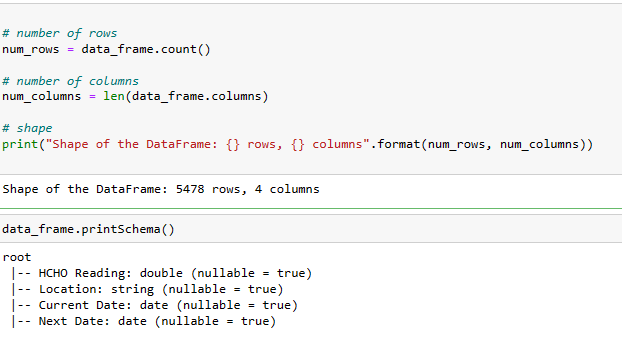
* Making a spark session to start the process



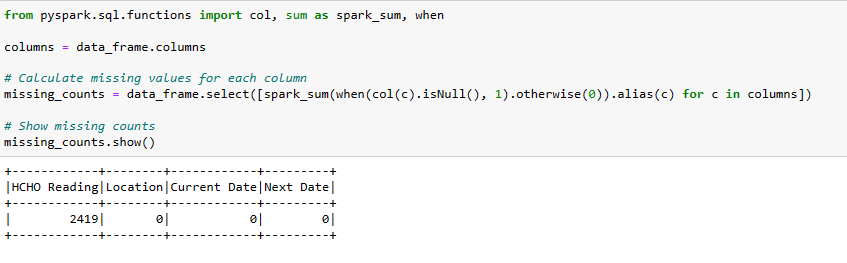
* Loading data frames and assigning column names for the data frame



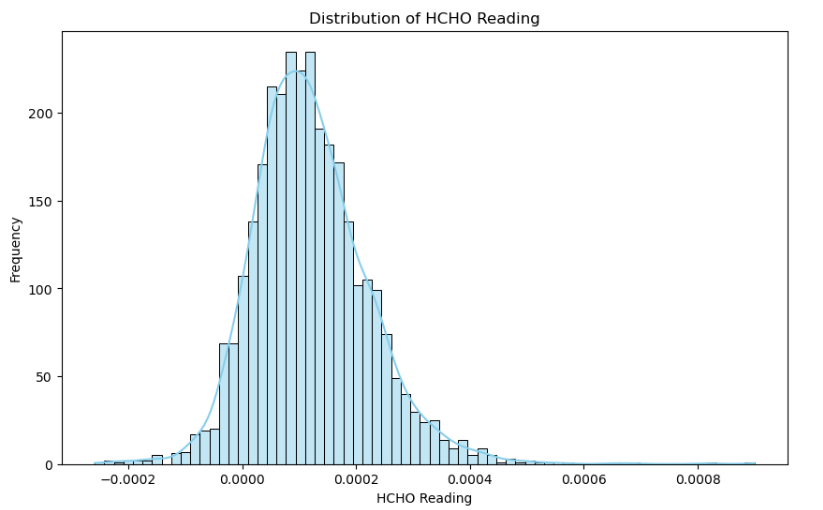
* Checking data types of each attribute



* Calculating the amount of missing values

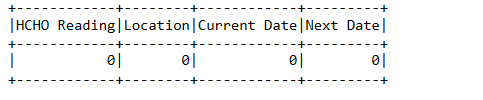


Before handling missing values in HCHO Reading column I extracted that column separately and converted it into a pandas data frame and visualized the distribution of data.

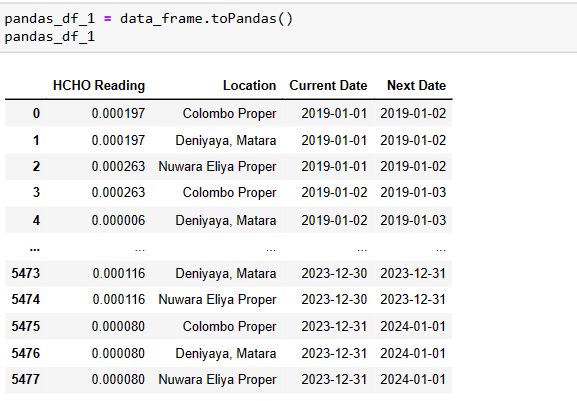


After confirming that the data is distributed like a normal distribution I decided to use forward windowspec and backward windowspec methods to fill the missing values.

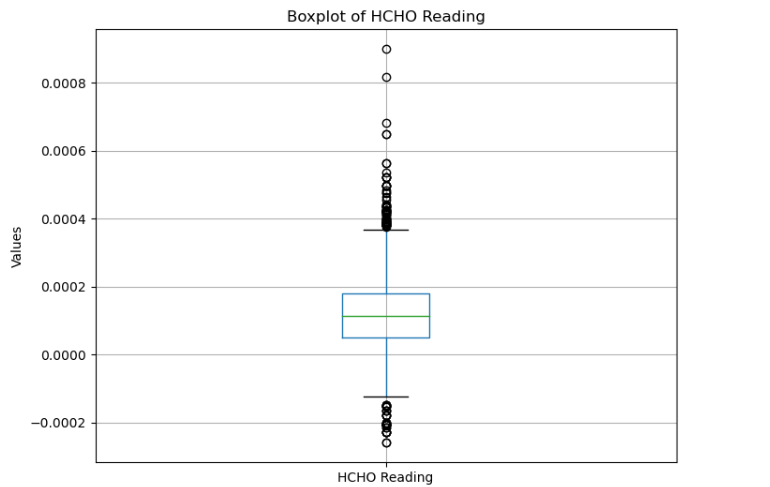
* Confirmed no missing values.

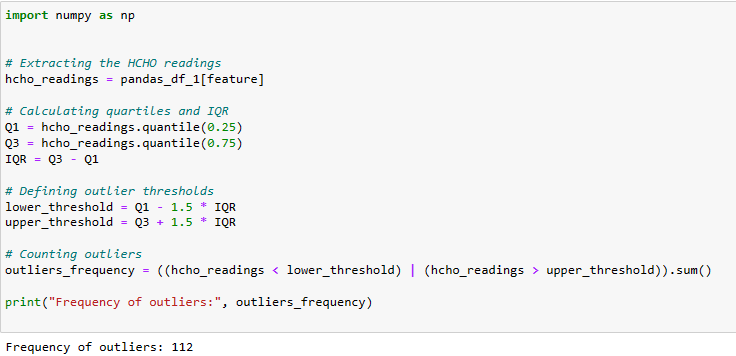


After handling missing values, I converted the data frame into a pandas data frame to continue further inspection.

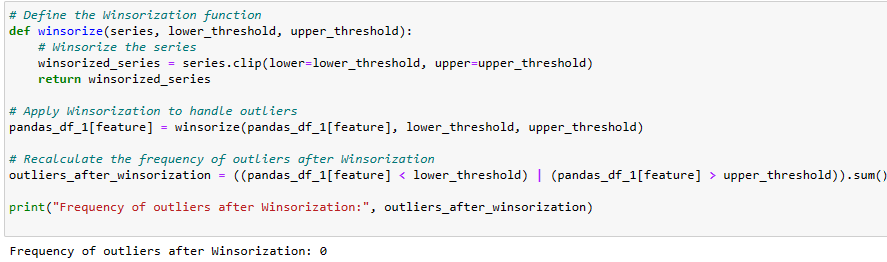


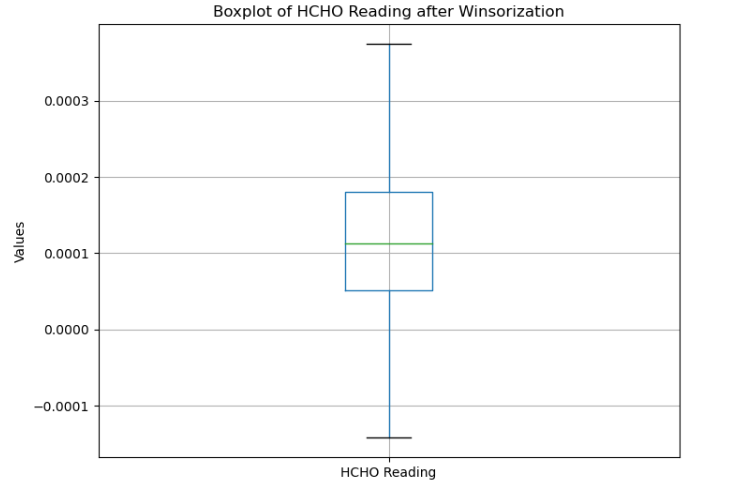
In the pandas data frame first selected the HCHO Reading column and then visualized outliers and calculated the number of outliers.





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.



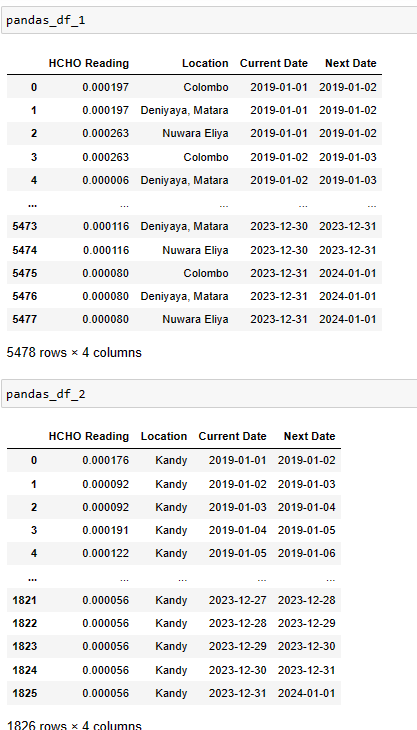


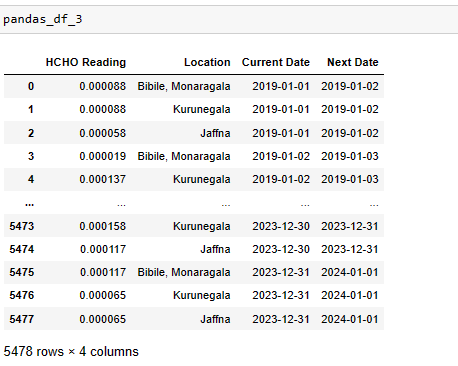
Removed the “Proper” part that include in several districts and converted it into the normal format.



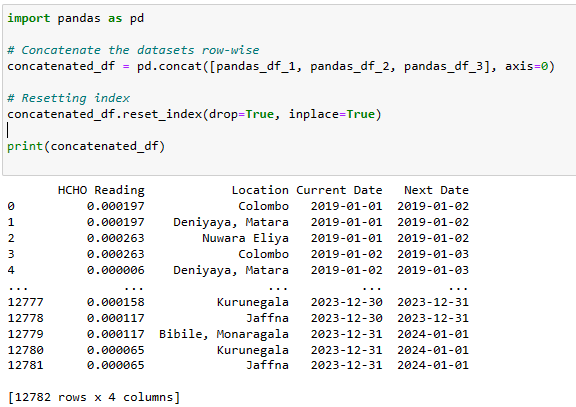
Then saved the data frame for use in future.

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.

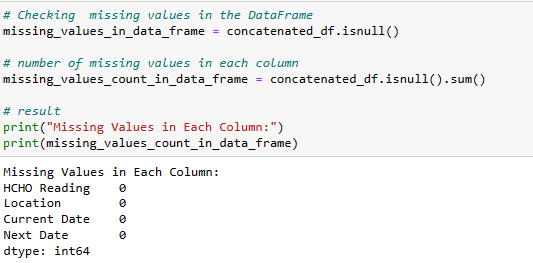




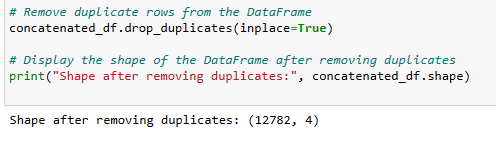
After preprocessing each data set individually, I concatenated all datasets and made a final dataset.



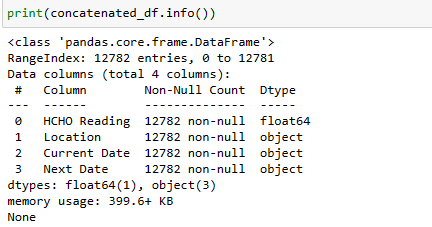
Checking missing value in concatenated data frame



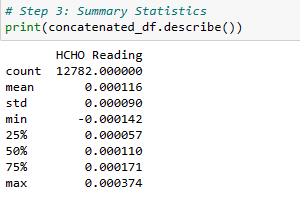
After that checked and dropped duplicated values in concatenated data frame.



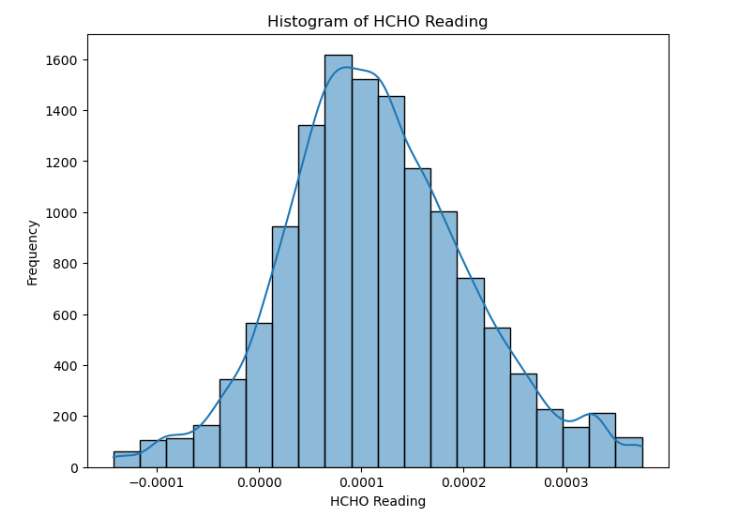
Checking data types and value count of concatenated data frame



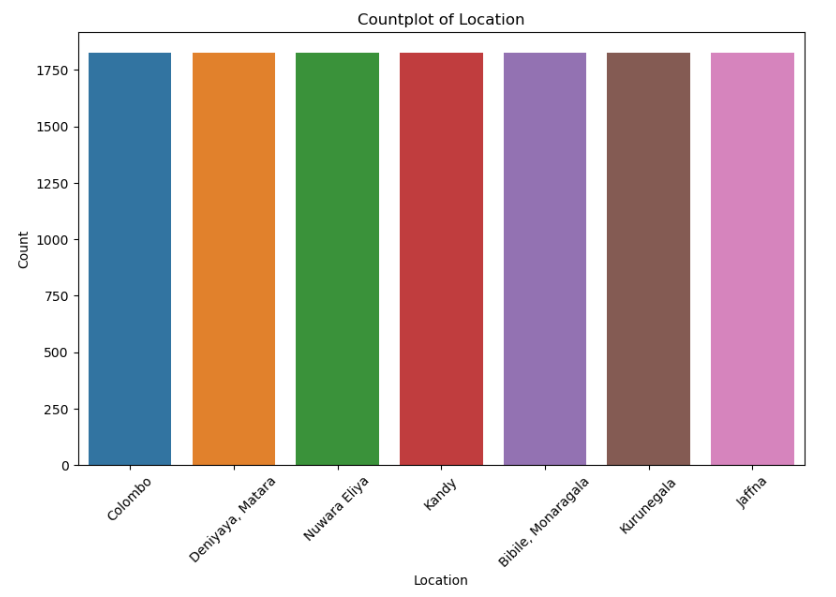
Taking the statistical summery of concatenated data frame



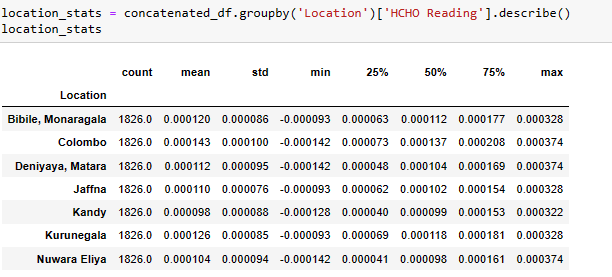
In the concatenated data frame most important attribute is “HCHO Reading” and in that attribute first I visualized the distribution of HCHO reading for an inspection and it follows a normal distribution and it is a good sign of data is distributed in a proper manner.



Then made a visualization with the count vs city.



Then taken the HCHO Reading location wise and then taken a detailed statistical distribution



Data Analysis ####

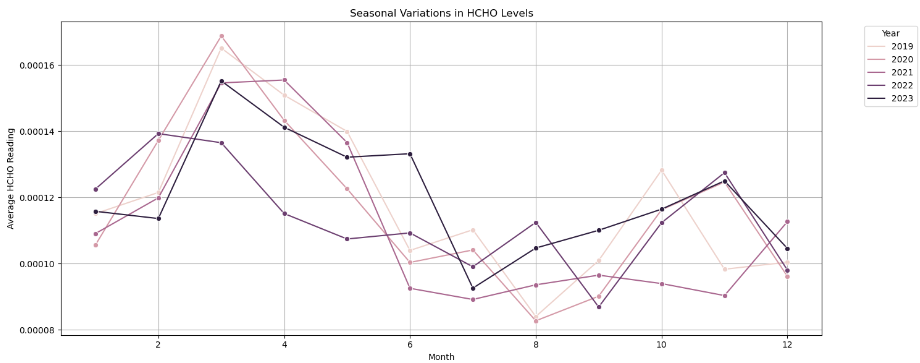
For the analysis of the data I started with analysis of seasonal data.

* Seasonal analysis

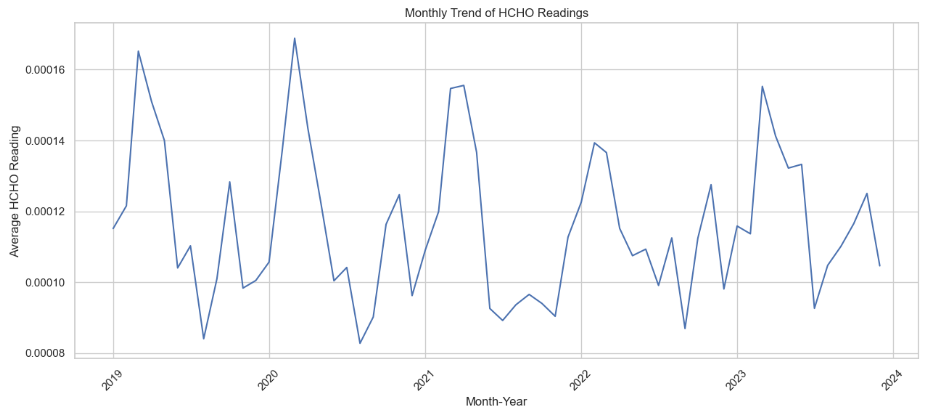
For the analysis purpose I used “groupby” function for grouping data by year and month.



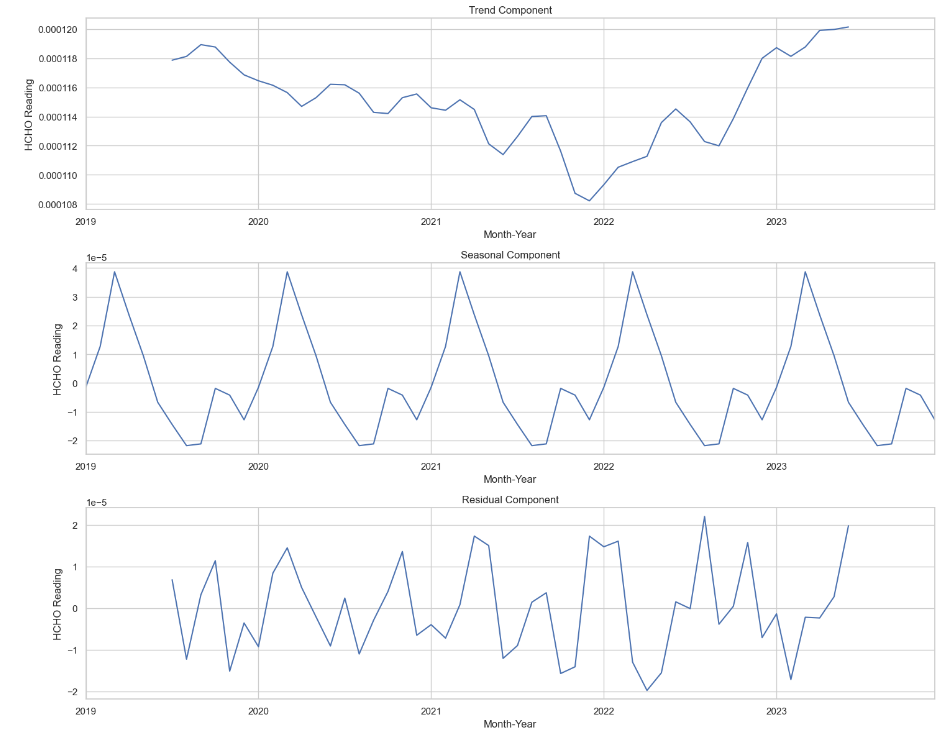
* Then I analyzed seasonal variations in HCHO levels with the months in each year.



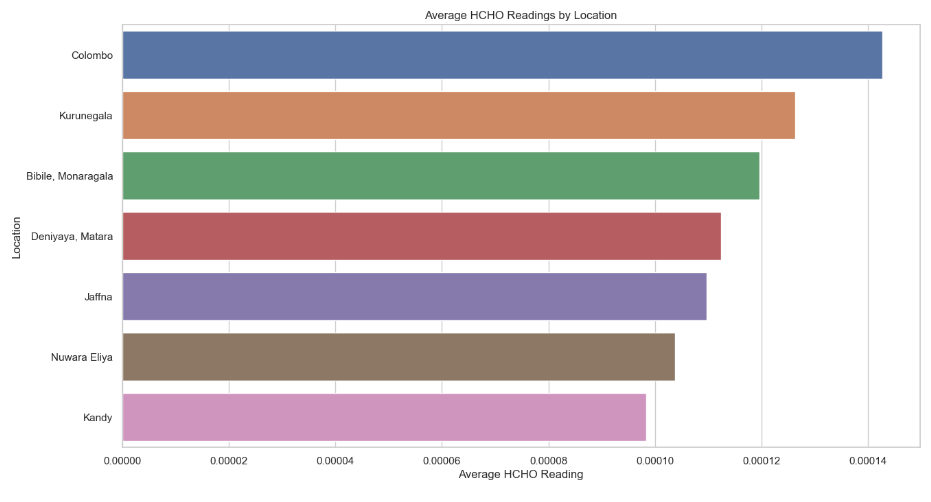
* Then analyzed average HCHO Reading with the time



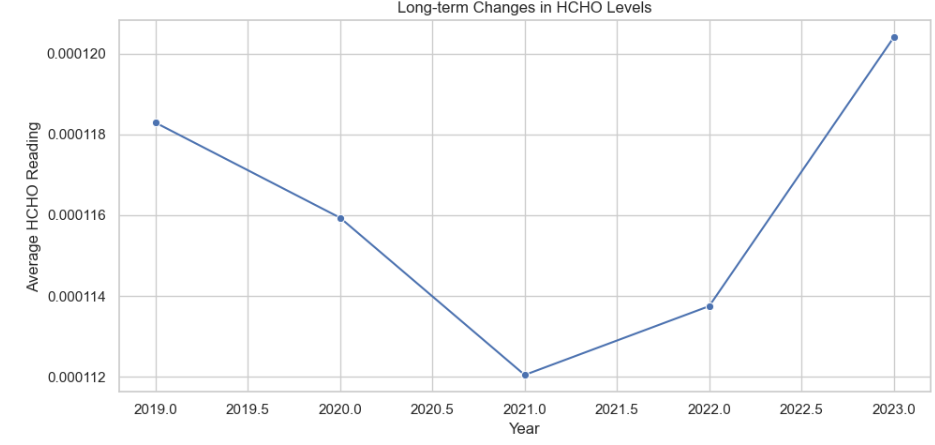
* Then started analysis of trend, seasonality and residual component.



* Grouped data with the location and visualized the average HCHO distribution by location

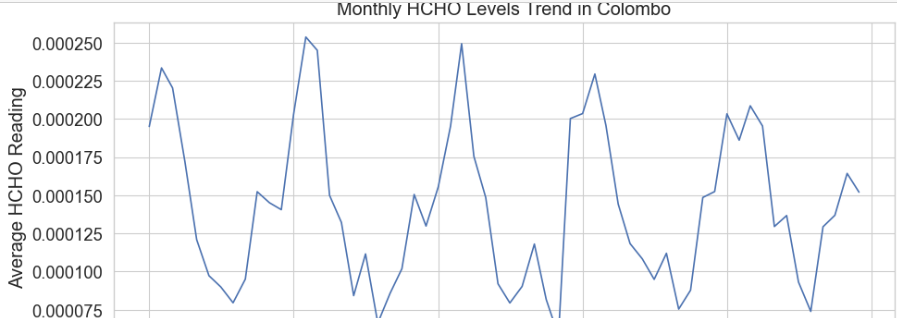


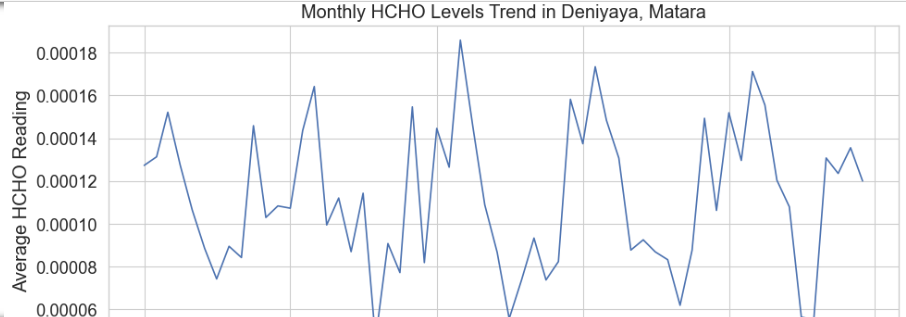
* After that plotted long term average HCHO levels with the time and this clearly showing the deviation of gas levels with the lockdown period.

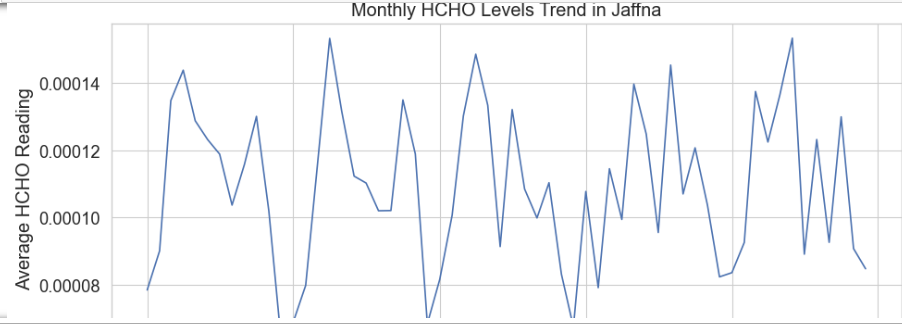


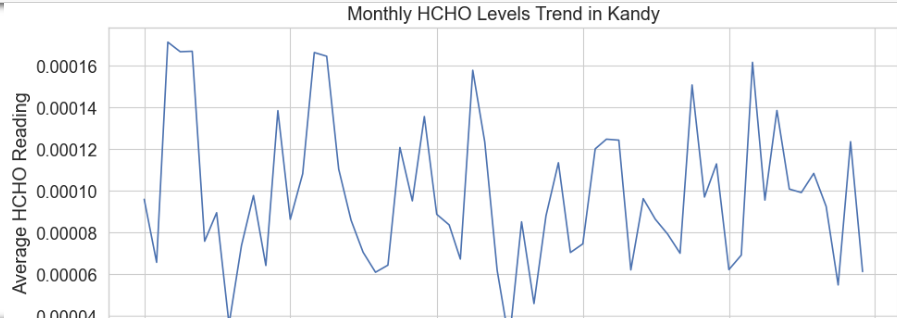
* Then I plotted monthly changes of average HCHO levels. For this I used two different ways.

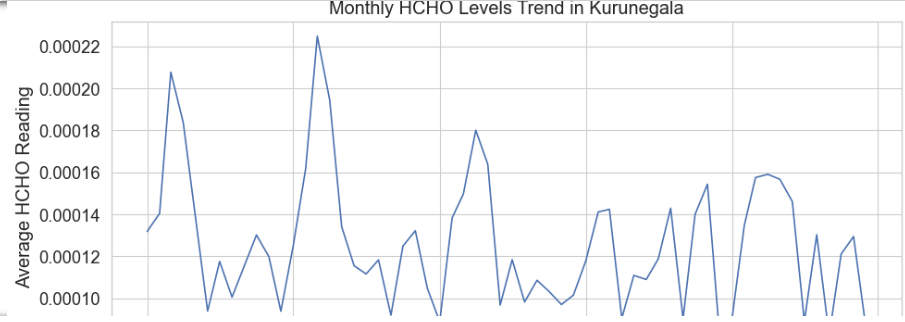
1. Plotting data in separate diagrams to analysis with respect to each city.

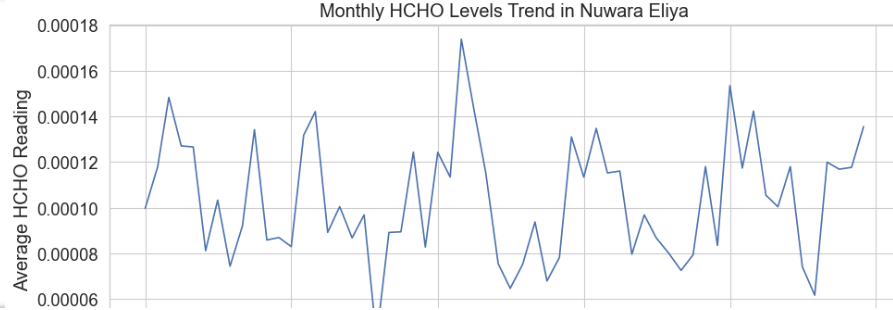




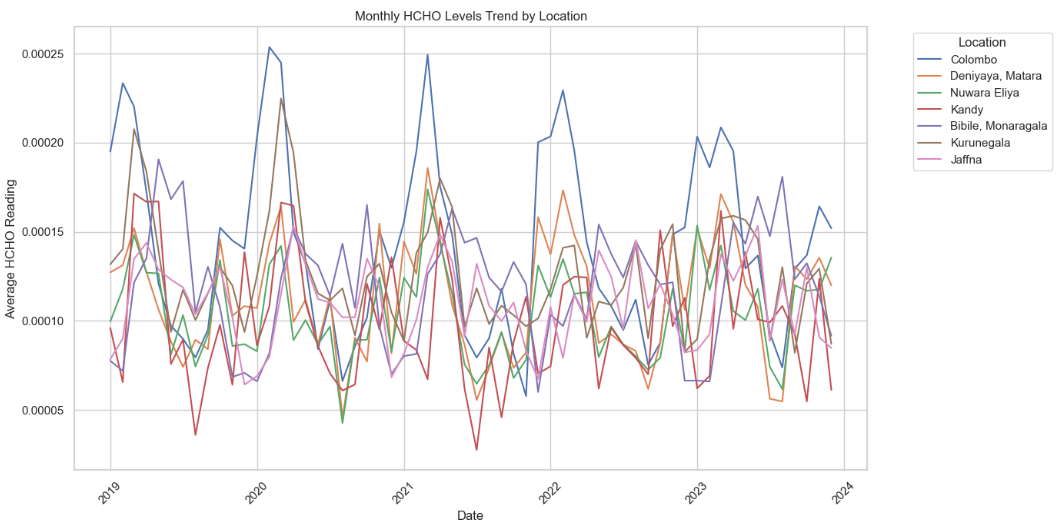




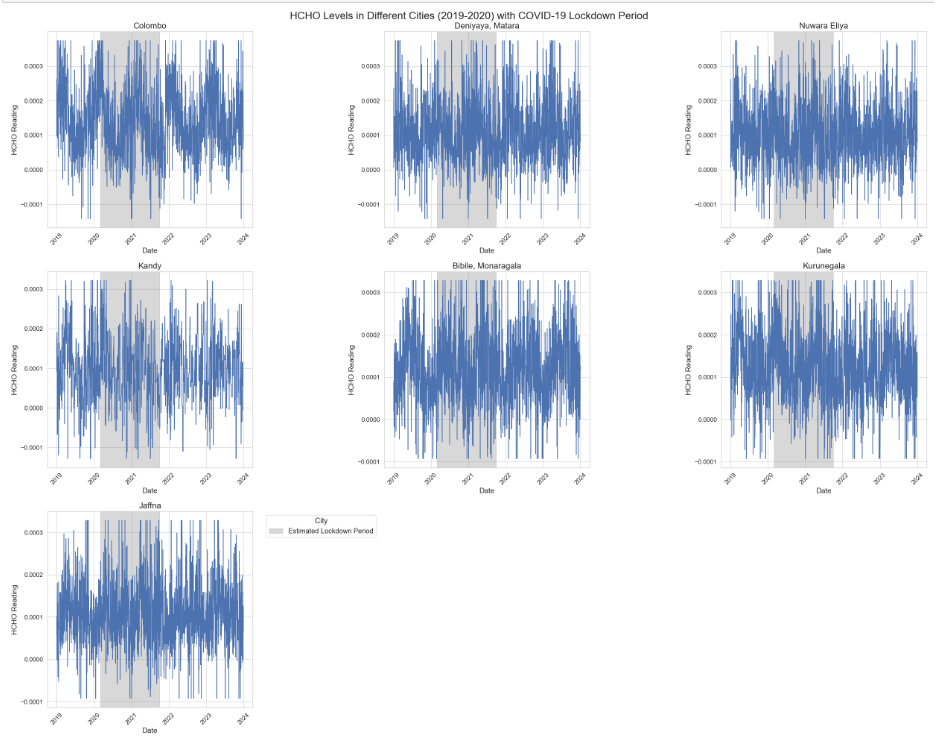




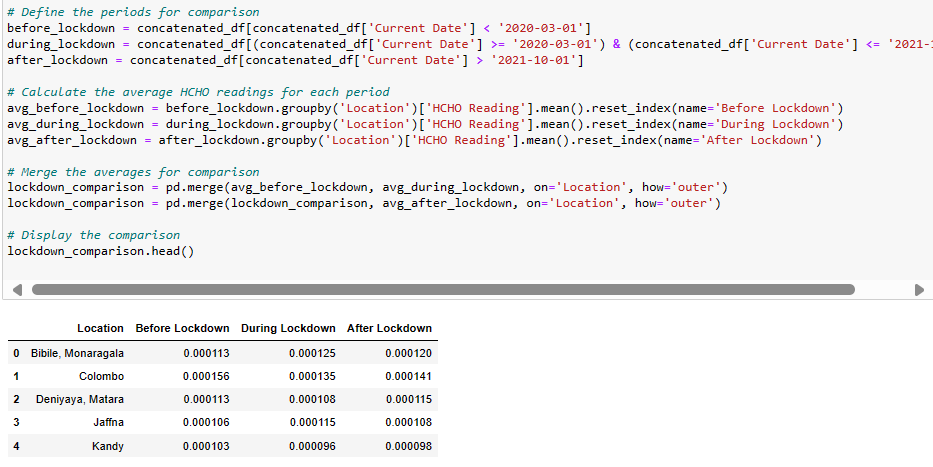
1. Plotted all in one diagram to analyze it city wise.



* Then analyzed the distribution of HCHO gas variation in each city in lockdown period. In this situation I assumed that lockdown period is from 2020-03-01 to 2021-10-01 due to various situations in the country.



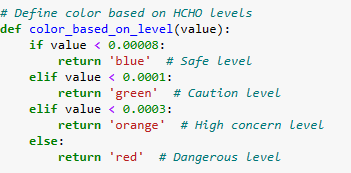
* Then taken the average HCHO level of each city before, after and during lockdown period.

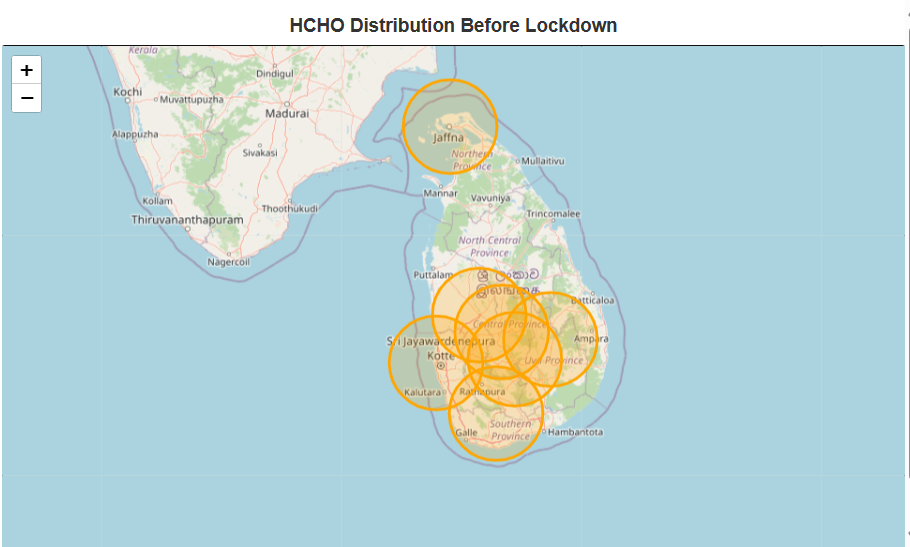


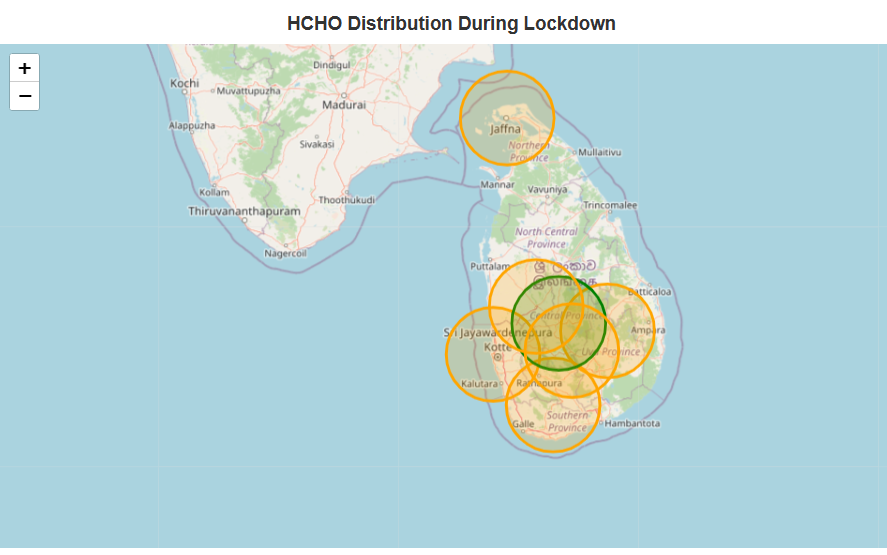
* Taken the HCHO levels and how they effect for human. # reference
* safe level to prevent irritation for most people: 0.00008 mol/m²
* caution level where discomfort and irritation can begin: 0.0001 mol/m²
* high concern level indicating significant risk of respiratory issues: 0.0003 mol/m²
* dangerous level, potentially lethal or causing severe health effects: 0.01 mol/m²

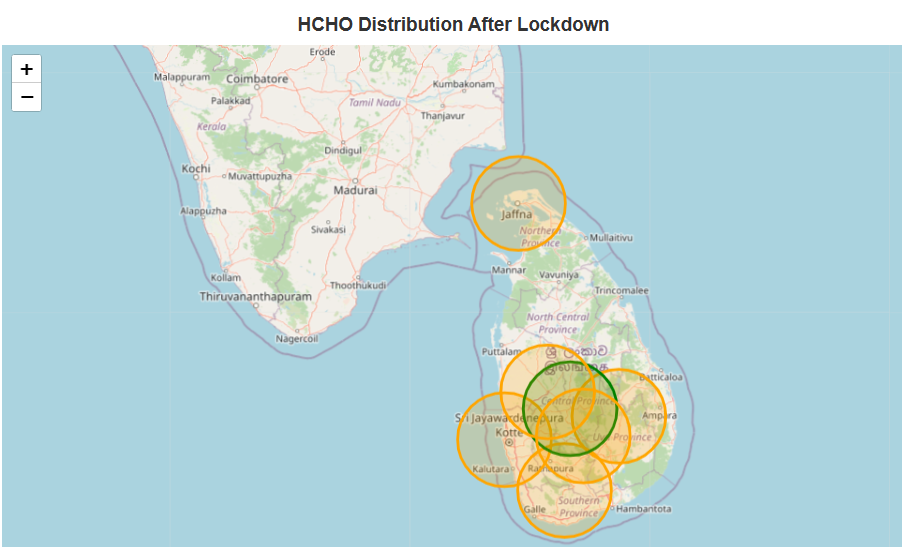
using this I visualized how gas levels varies in different cities using periods such as before, during and after lockdown using folium heatmaps.

Logic used for making the heatmap

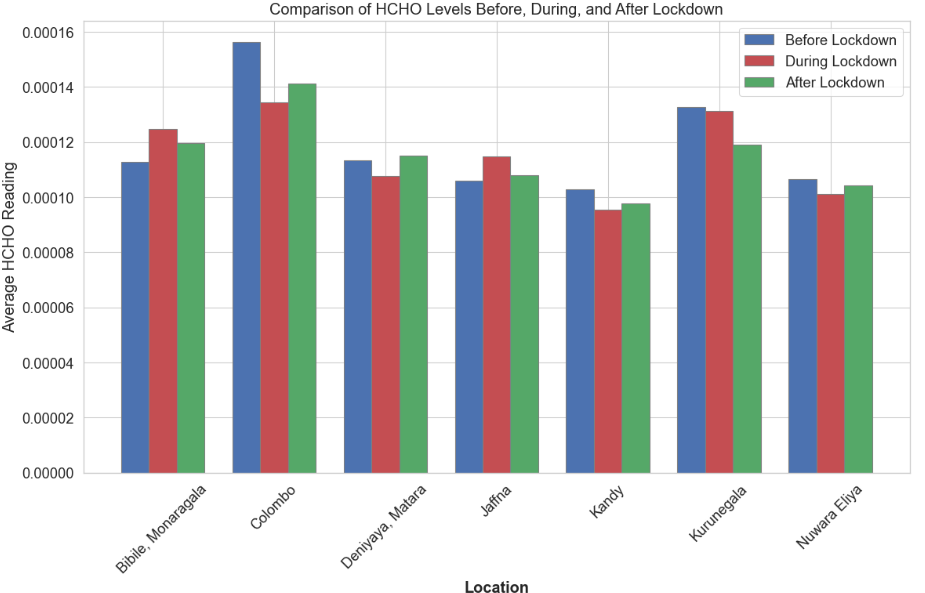








* Taken the average HCHO gas distribution with in above 3 seasons.

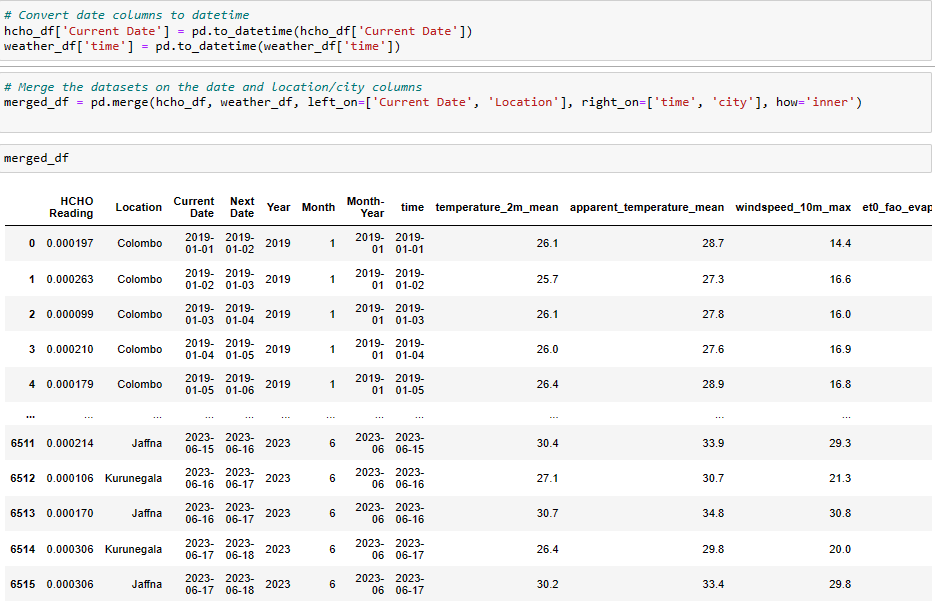


For further analysis I taken a Sri Lankan weather dataset separately and preprocessed that dataset separately and loaded in this notebook for continue advanced analysis using main factors that contributing for distribution of a gas.

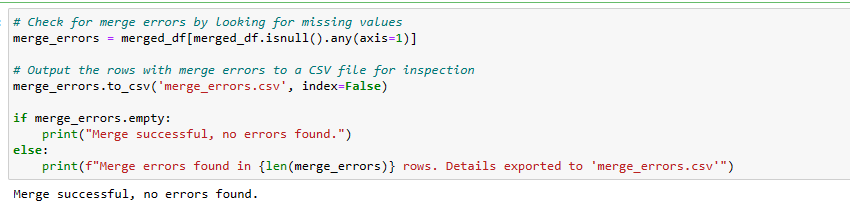
That dataset is a cleaned dataset but it has details belongs to following cities.

'Colombo’, ‘Kandy’, ‘Jaffna’, ‘Matara’, ‘Kurunegala'.

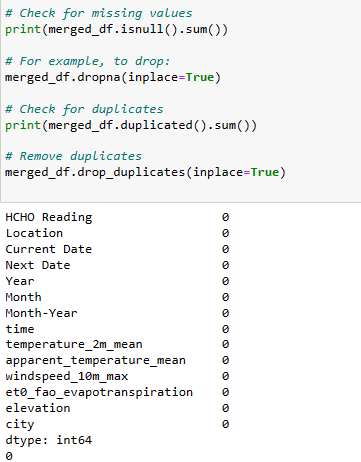
* Then I merged my concatenated dataset with my new weather dataset using date and location.



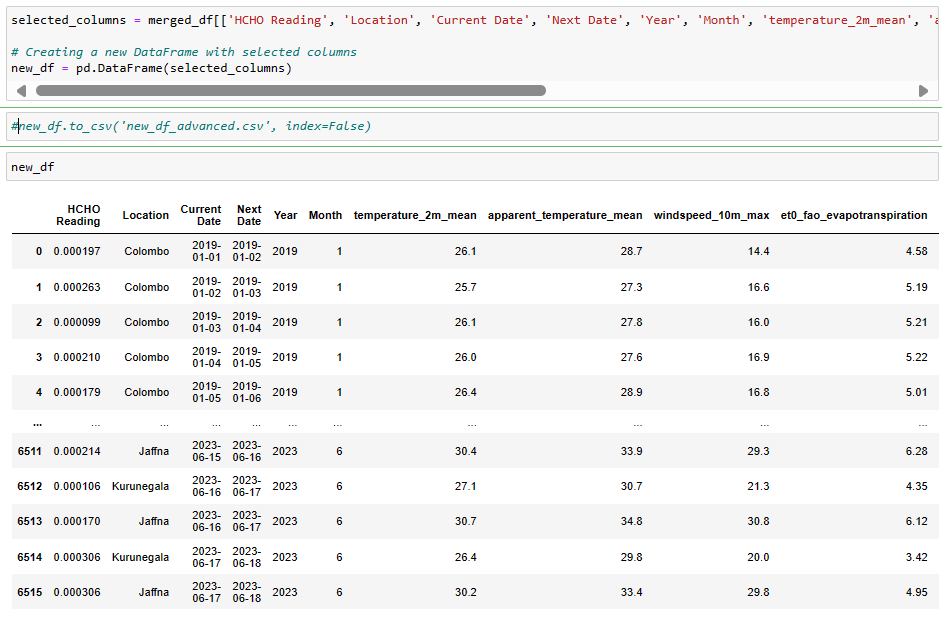
* Checking for merging errors



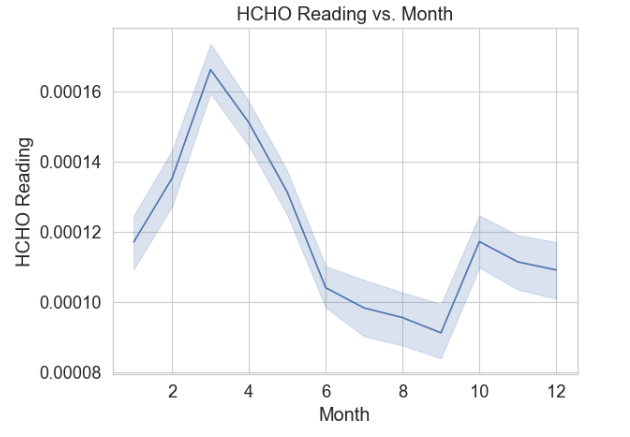
* Checking duplicates and missing values and dropping those values from the new dataset.



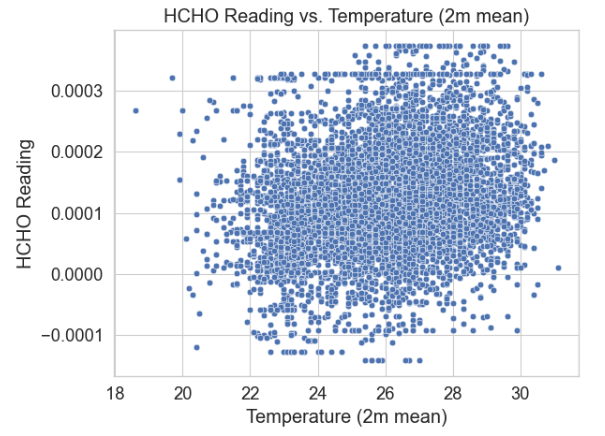
* Saving the final dataset for further analysis.

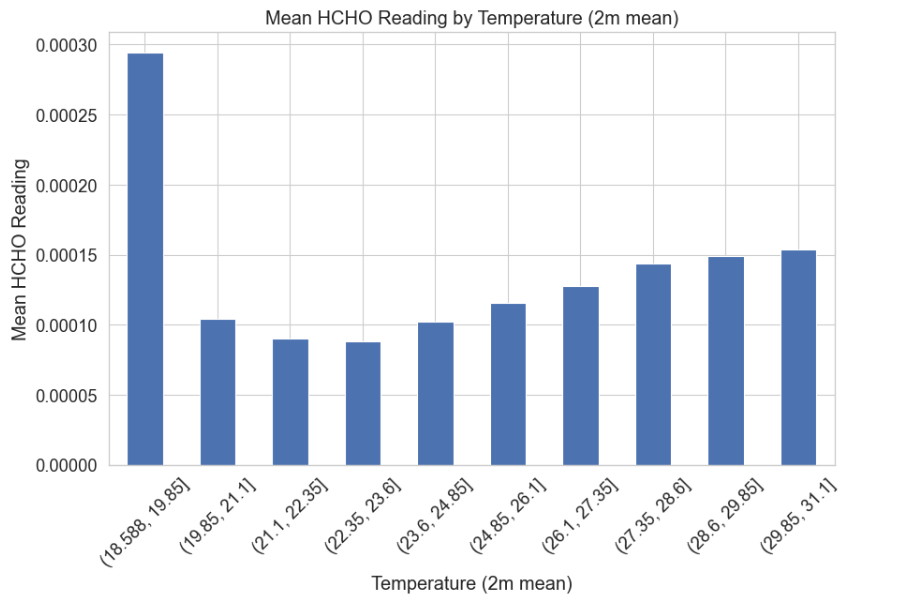


First, I analyzed HCHO distribution over months.

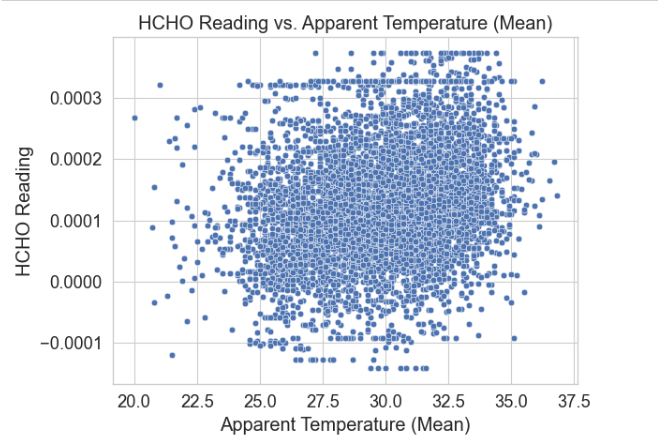


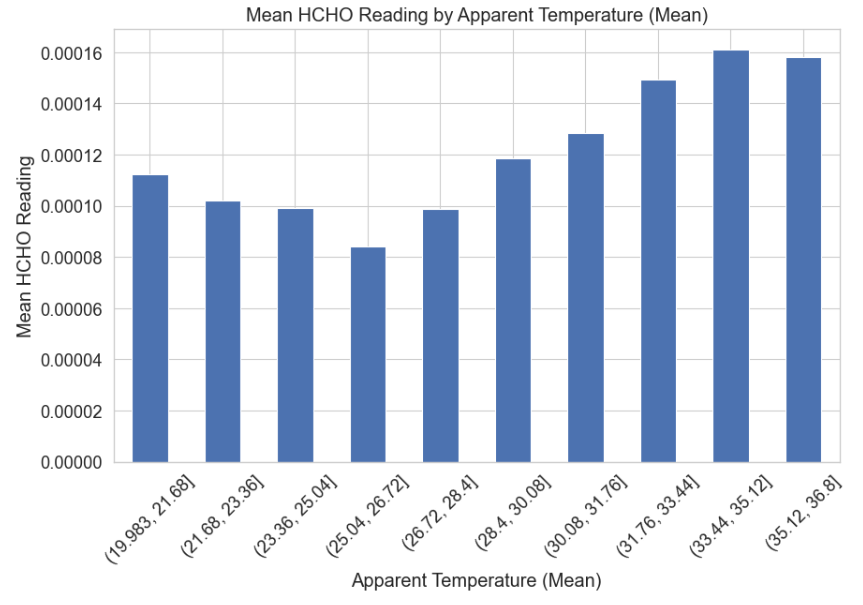
* HCHO distribution vs mean temperature and temperature ranges.



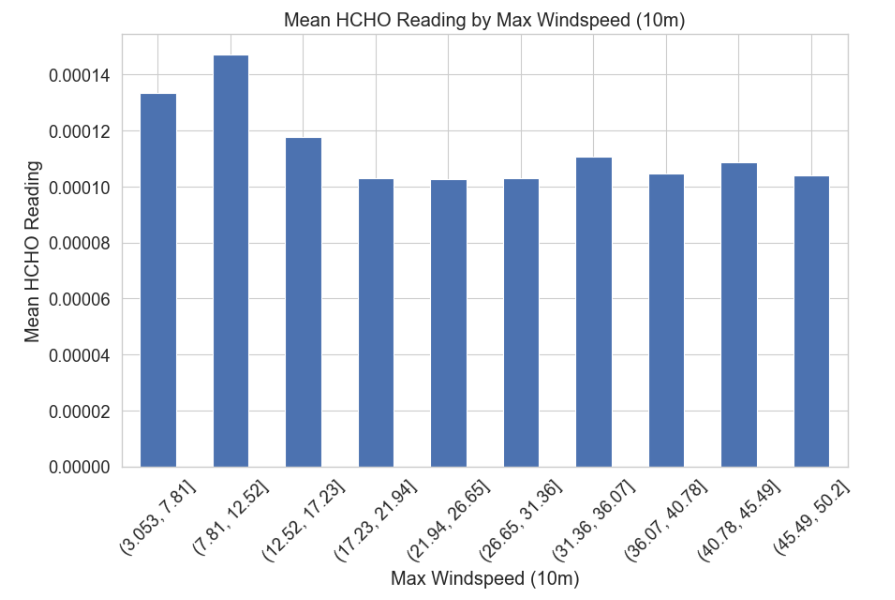


* HCHO distribution vs apparent temperature and temperature ranges.

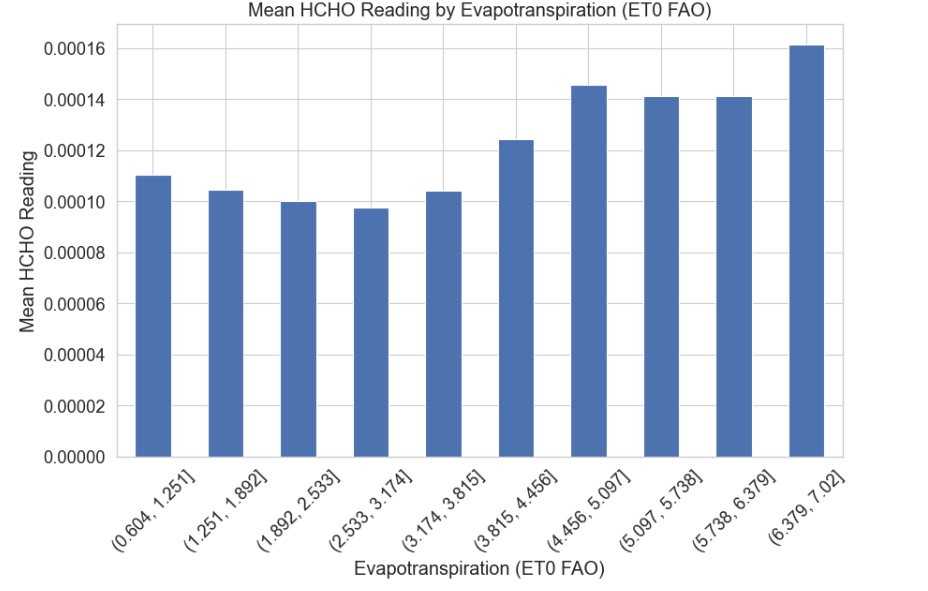




* HCHO distribution vs Max windspeed.



* HCHO distribution vs Evaporation



* HCHO distribution vs Elevation

