**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**

Contents

[Introduction 3](#_Toc164541361)

[Dataset Selection 3](#_Toc164541362)

[Data Integration and Preprocessing 3](#_Toc164541363)

[Spatial Temporal analysis 13](#_Toc164541364)

[Further analysis using external data 25](#_Toc164541365)

[Machine Learning 35](#_Toc164541366)

[Model training and evaluation 36](#_Toc164541367)

[Model training using external factors 47](#_Toc164541368)

[Limitations and future work. 49](#_Toc164541369)

[References 49](#_Toc164541370)

# 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.

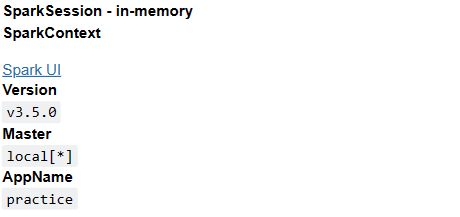
# Dataset Selection

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.

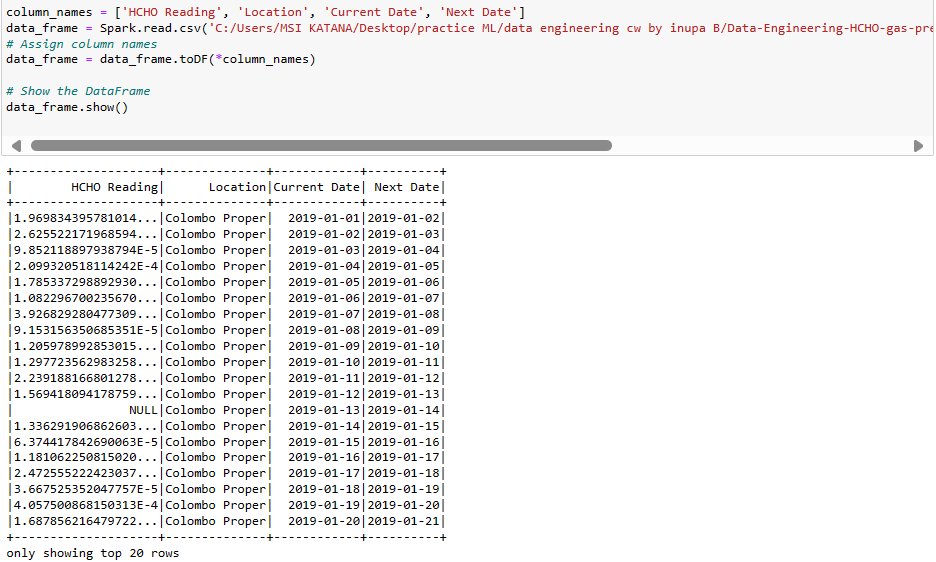
# Data Integration and Preprocessing

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

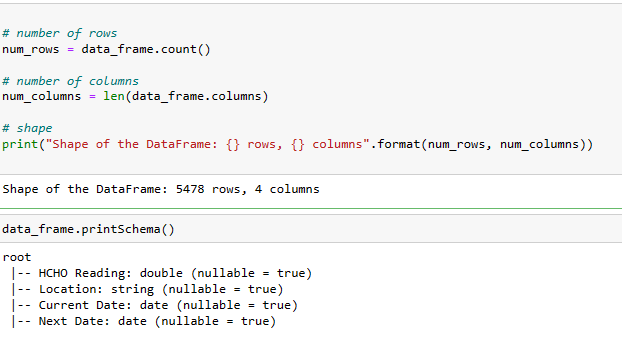


* Loading data frames and assigning column names for the data frame



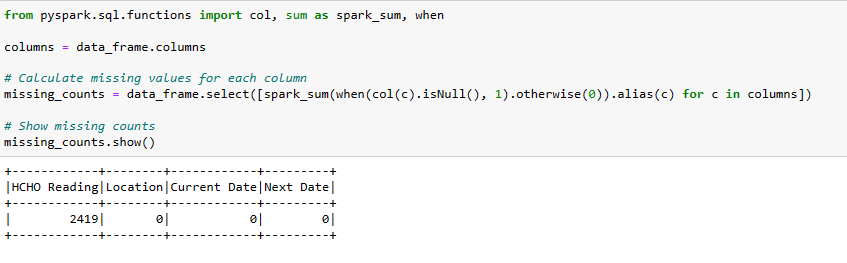
Assigned suitable column names and then assigned them to the data frame. Data is loaded from a csv file.

* Checking data types of each attribute



Checked the shape and schema of each attribute.

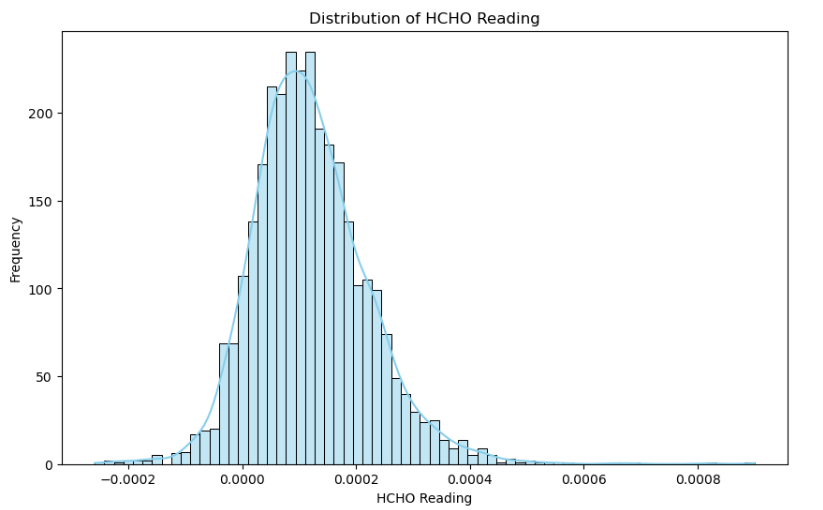
* Calculating the amount of missing values



Calculated missing value count of each column using spark and taken it into a table.

* Handling 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 because data distribution is an important factor when taking decisions in handling missing values.

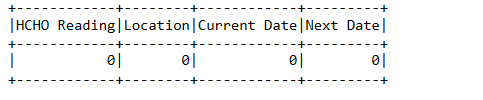


After confirming that the data is distribution following a normal distribution, I decided to use forward windowspec and backward windowspec methods to fill the missing values.

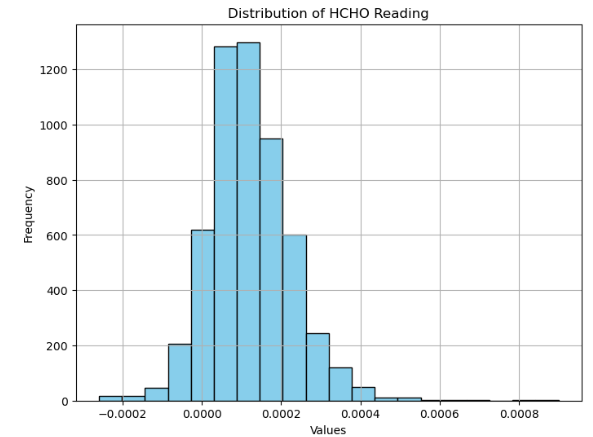
The forward windowspec method defines a moving window that extends forward from a specified point in the time series. This approach makes it possible to analyse data points that happen after the initial point. Usually, the width of the window and the increment by which it advances along the time series are determined by specifying parameters like window size and step size.

On the other hand, the reverse windowspec approach entails creating a moving window that begins at a specific location and moves backward along the time series. By using this strategy, data points that come before the starting point can be examined, providing insights into past trends and patterns.

* Confirmed no missing values.

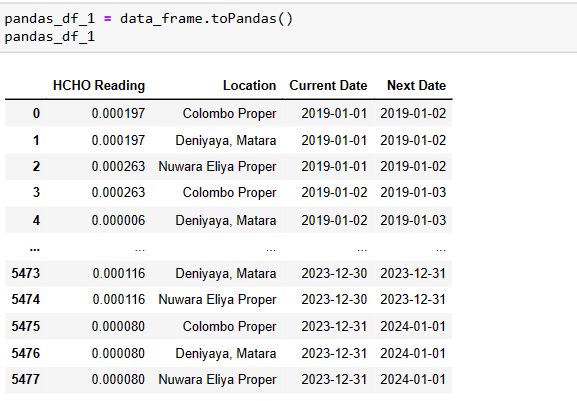


* Data distribution of the data frame after handling missing values.

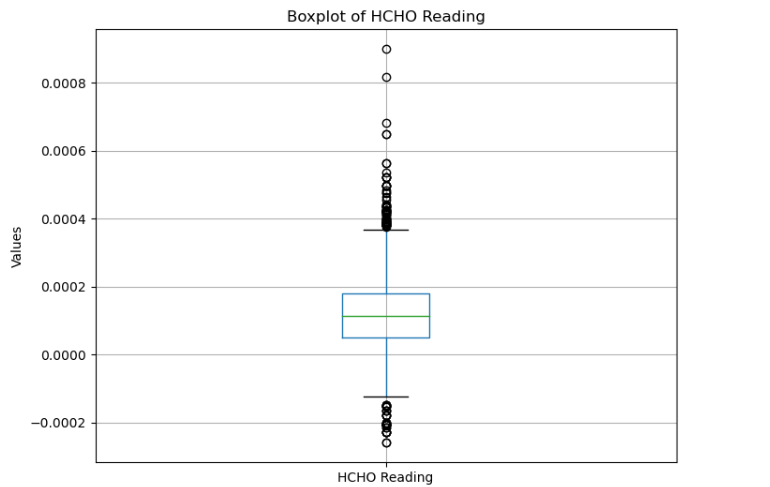


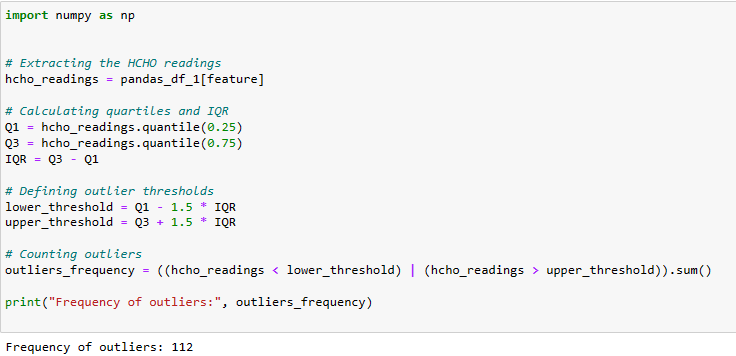
No considerable deviation in the distribution after handling missing values.

After handling missing values, I converted the full data frame into a pandas data frame to continue further inspection because we cannot do visualizations using pyspark.

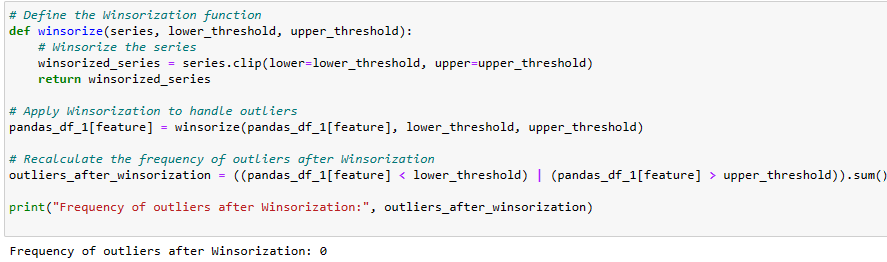


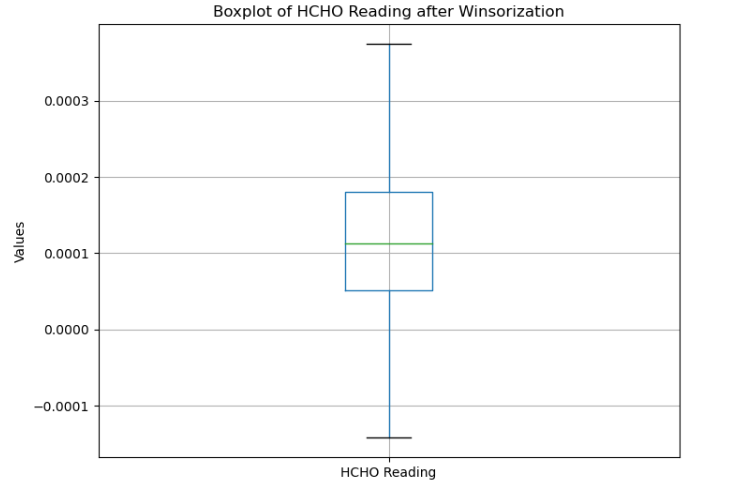
* Removing outliers

In the pandas data frame first selected the “HCHO Reading” column and then visualized outliers and calculated the number of outliers. For calculating outliers, I used interquartile range method that used in boxplots and the logic used for calculating non outlier range is given below.



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.



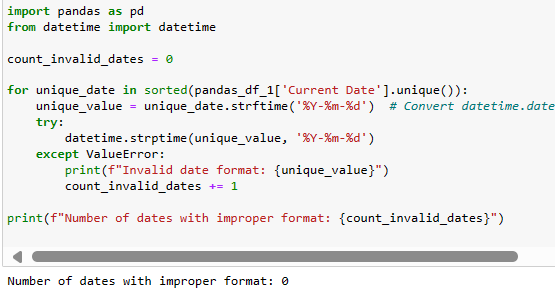


* Removing inconsistencies in district names

Removed the “Proper” part that include in several districts and converted it into the normal format. In this scenario I decided that we can identify each district. For an example we cannot identify proper difference between “Colombo” and “Colombo Proper”.

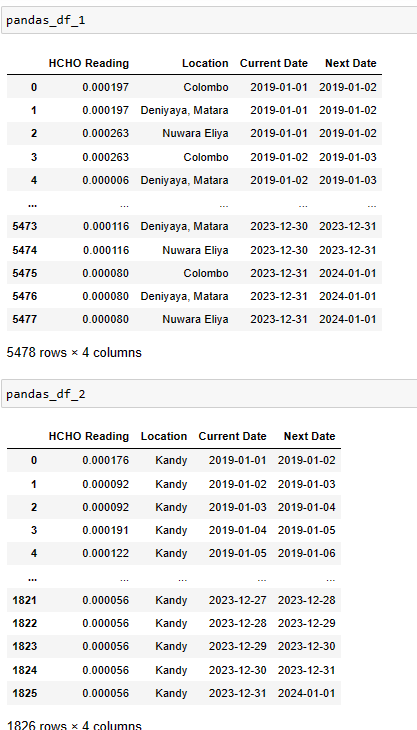


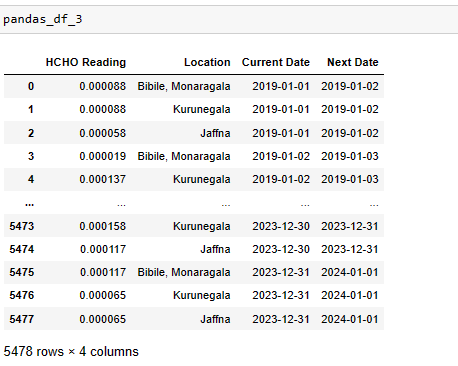
* After handling missing values, made a manual inspection and checked other columns are in a proper format



Then saved the data frame for use in future.

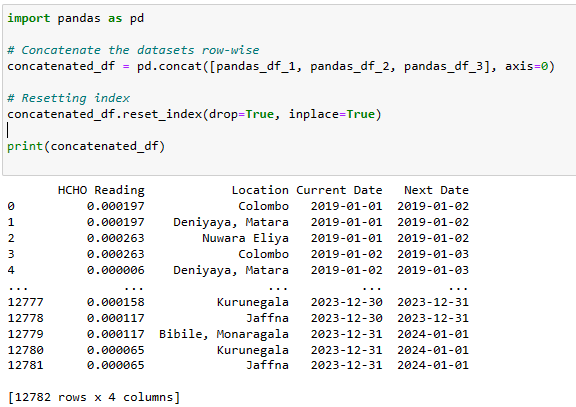
* Continued the above process of preprocessing in a same manner to other datasets and saved the datasets to use in future.



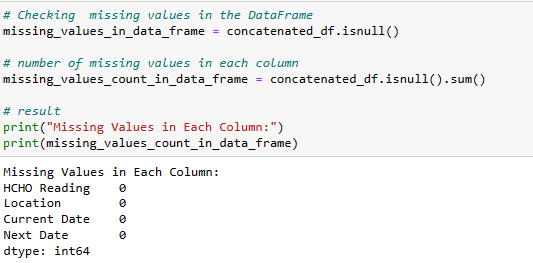


* Concatenating datasets

After preprocessing each data set individually, I concatenated all datasets in row wise and then made the final dataset for my analysis.



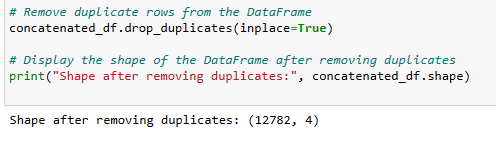
* Checking missing value in concatenated data frame



After concatenating the final dataset, again checked for missing values and confirmed that the preprocessing part is done without any error and no missing values in the final data frame.

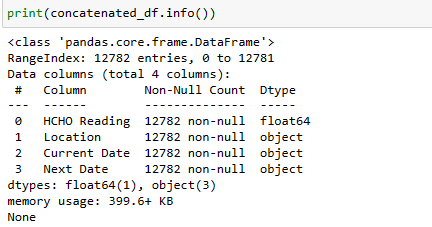
* Checking for duplicate values in concatenated data frame.

Checked and dropped duplicated values in concatenated data frame.



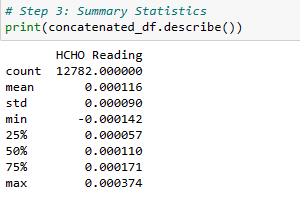
* Checking data types in concatenated data frame.

Then checked about the data types in the final data frame after concatenation.



* Statistical summery of concatenated data frame

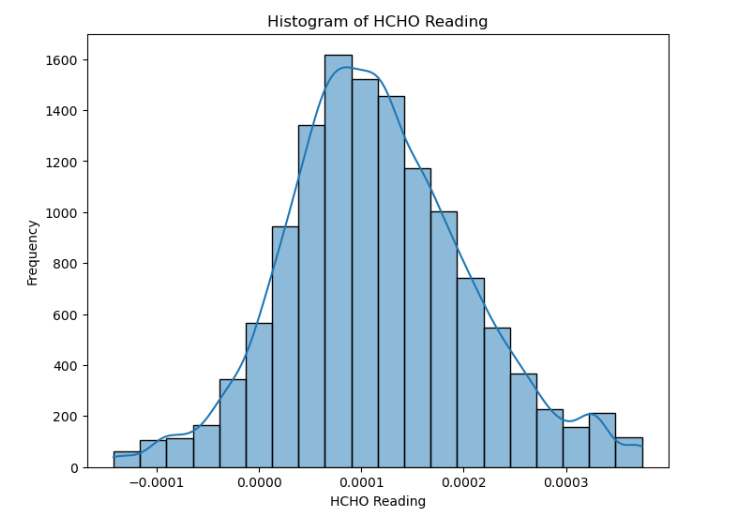
Taking the statistical summary of concatenated data frame



# Spatial Temporal analysis

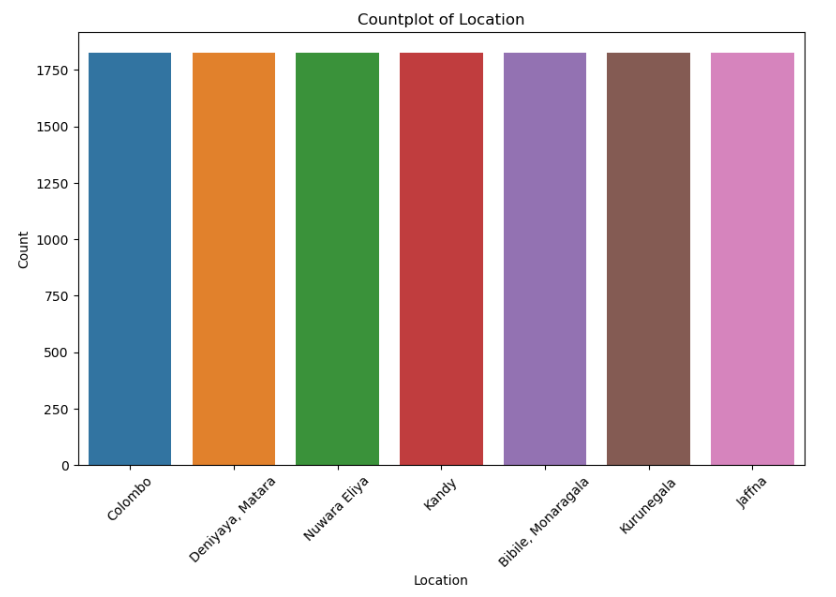
* Analyzing “HCHO Reading” variable

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.



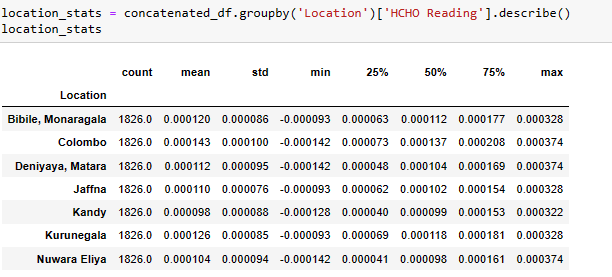
In the final dataset the “HCHO Reading” attribute follows a normal distribution. This distribution confirmed that the data related to this column distributed in a proper way. This is a good sign when taking decisions related to data and handling data

* Analyzing city wise count



This bar graph displays the count of records related to each city in the data set. From this we can identify similar amount of records are allocated to each city. This is a good sign when training the model.

* “HCHO Reading” in location wise and taken a detailed statistical distribution



This table describe in detail statistical description of HCHO Reading in each city.

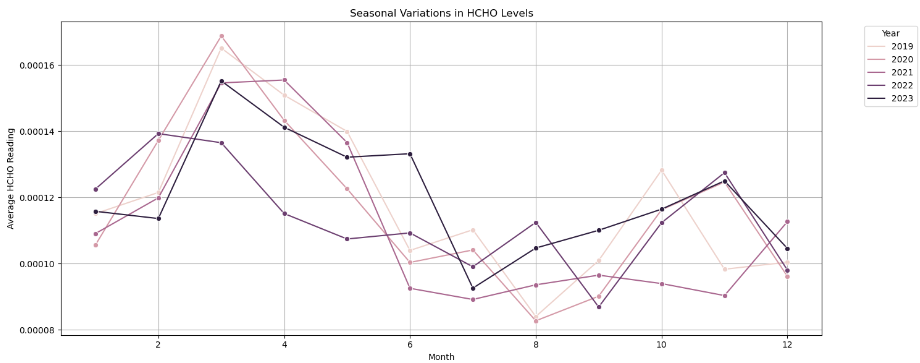
For the analysis of the data started with analysis of seasonal data.

* Seasonal analysis

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

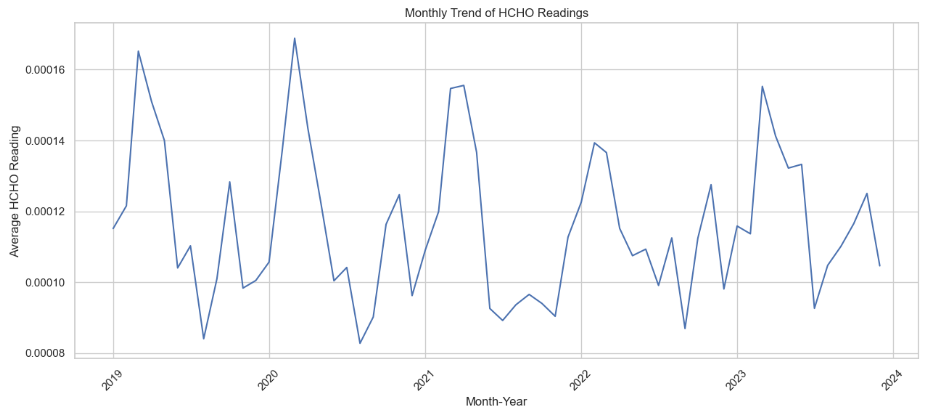


* Then analyzed seasonal variations in HCHO levels with the months in each year.



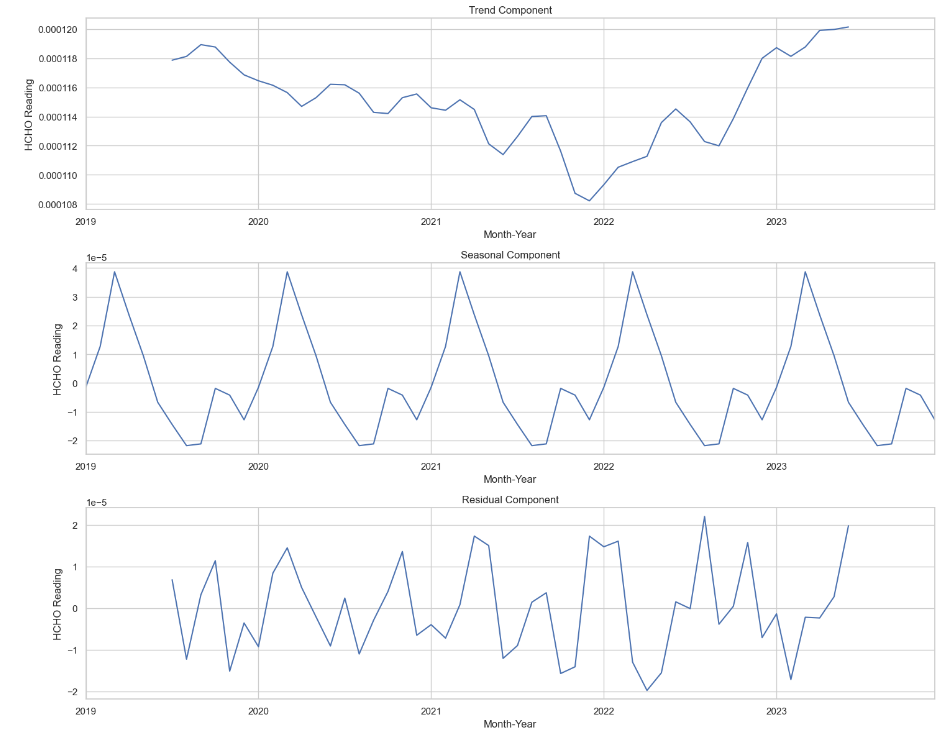
As a country that closer to the equator, Sri Lanka did not show noticeable seasonal changes such as winter, summer, spring, autumn but there are some seasonal changes that we can notice in each year during some seasons. In Sri Lanka first few 3 – 4 months are dry season for most of the part of the country and in that season, temperature is also higher than other seasons. In our analysis HCHO reading is little high in that season. After April we can notice two monsoon seasons in the country and in that period rainfall is changing and other factors such as wind speed also may affect for gas distribution.

* Analyzed average HCHO Reading and trend with the time



For these first aggregated HCHO readings in a monthly basis and taken the mean of the distribution and after that monthly data is used to plot the trend from 2019 to 2023. In the above graph we can identify a trend in first few months of the year.

* analysis of trend, seasonality and residual component.

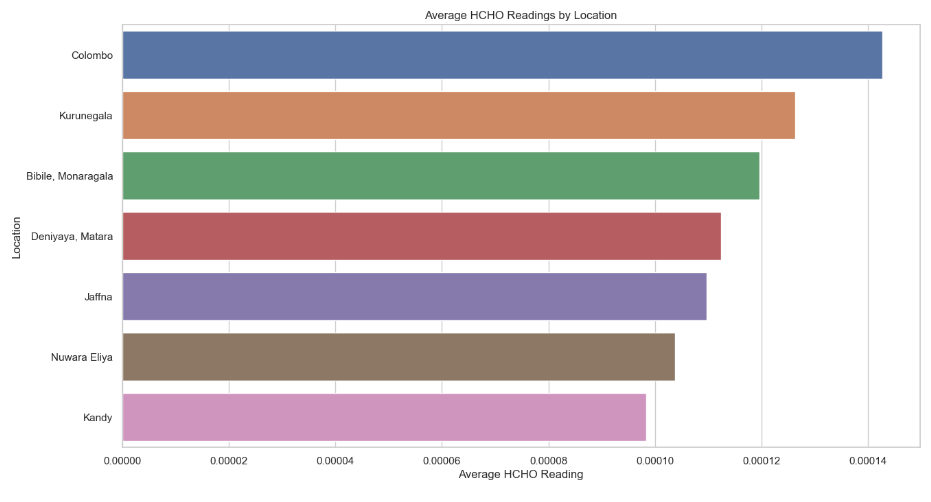


From the above analysis trend component is changing in a different way. Form 2019 to end of 2021 it decreases gradually and after 2022 middle season it stated increasing step by step. Covid lockdown period, economic situation of the country and electricity problem that made a direct impact on industries and that may be the reason for this.

In the seasonality component it shows a clear seasonality from 2019 to 2024 across the dataset.

For the residual component, there is a changing movement in that also.

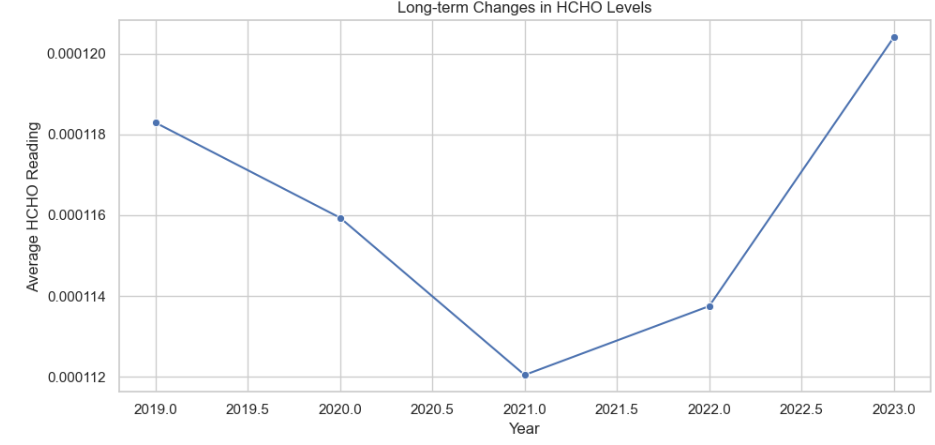
* Grouped data with the location and visualized the average HCHO distribution by location



This visualization describes about change of average HCHO reading in each city. Colombo is the region that has maximum average distribution and it is the commercial capital of the country and Colombo has many industries and that also may be a reason for this distribution. Kurunegala takes the second highest distribution. Kurunegala belongs to north western province of Sri Lanka which main contributor for the paddy harvest and coconut cultivation in large scale. Kurunegala there are lot of farm lands and when using chemicals for harvesting activities and other reasons paddy fields emit different gases to the atmosphere such as methane, HCHO and other gasses. Other than that, Bibile, Monaragala, Matara, Deniyaya are famous for rubber, tea, and many other cultivations and some industries also in those regions agriculture may be the main reason for emission of toxic gasses. In Jaffna also there are some cultivations and Jaffna is closer to south India which produce large scale toxic gases and that may be a reason for increasing level of HCHO gas in Jaffna. Other than that Jaffna maintains high temperature and that also indirectly impactful for distribution of the gasses.

Nuwara Eliya and Kandy are popular for tea cultivation and that may be the reason for spreading HCHO in those hill country regions and also there are some industries like tea factories as well they also make an impact on this.

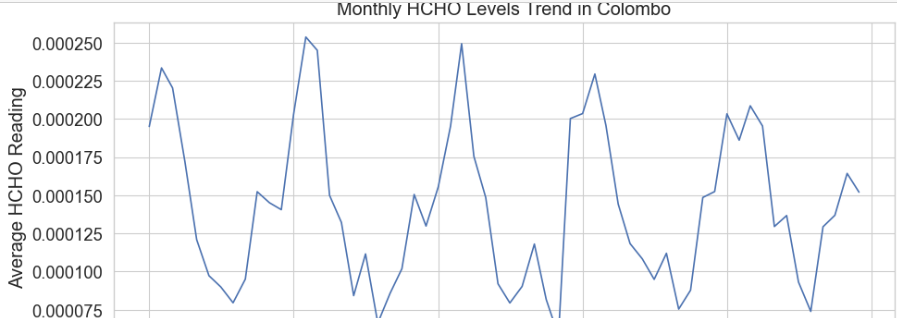
* plotted long term average HCHO levels with the time.

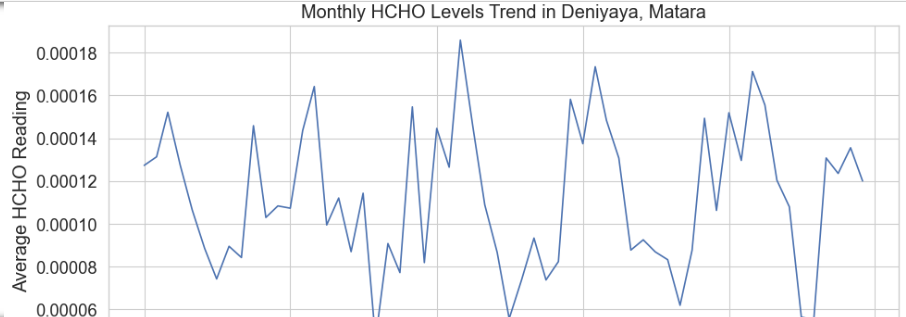


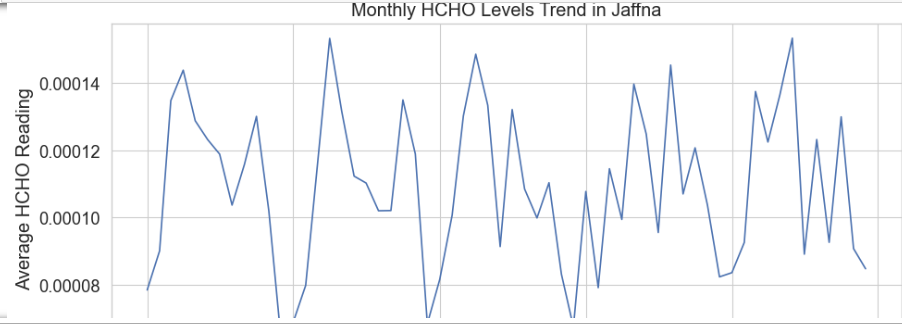
This indicate an average long term HCHO distribution in all cities form 2019 to end of 2023. In this analysis we can clearly identify a decreasing movement from 2019 to 2021 and increasing movement from 2021 to 2023. In 2021 period gas distribution decreases in an unprecedented proportion and Covid pandemic and lockdown period and financial situation of the country may be the major reasons for this. After lockdown period we can identify a clear deviation of gas distribution and In beginning of the 2023 gas distribution is in the highest point and the after pandemic new normal situation of the country may be the reason for this.

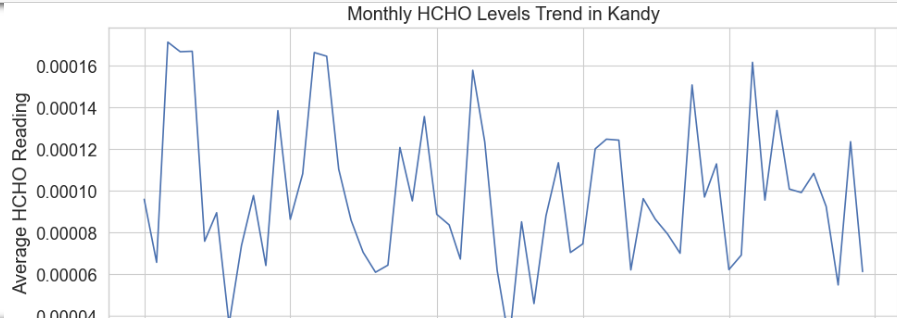
* plotted monthly changes of average HCHO levels. For this I used two different ways.

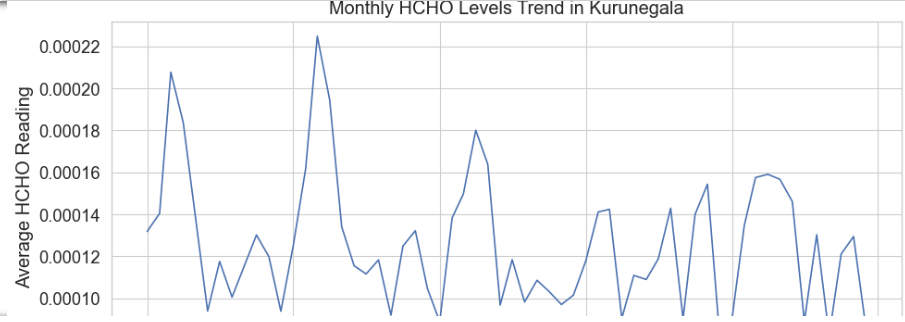
1. Plotting data in separate diagrams to analysis with respect to each city.

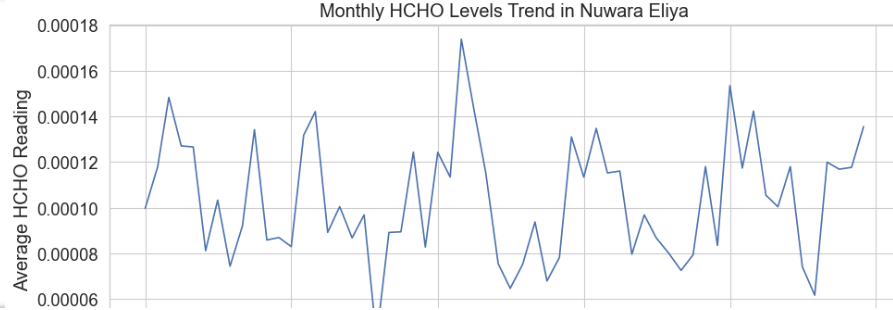






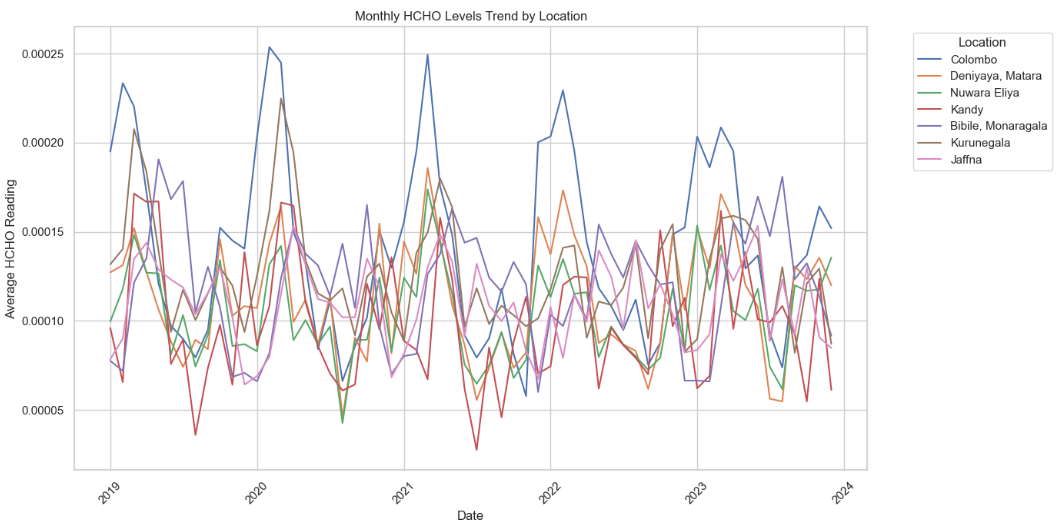




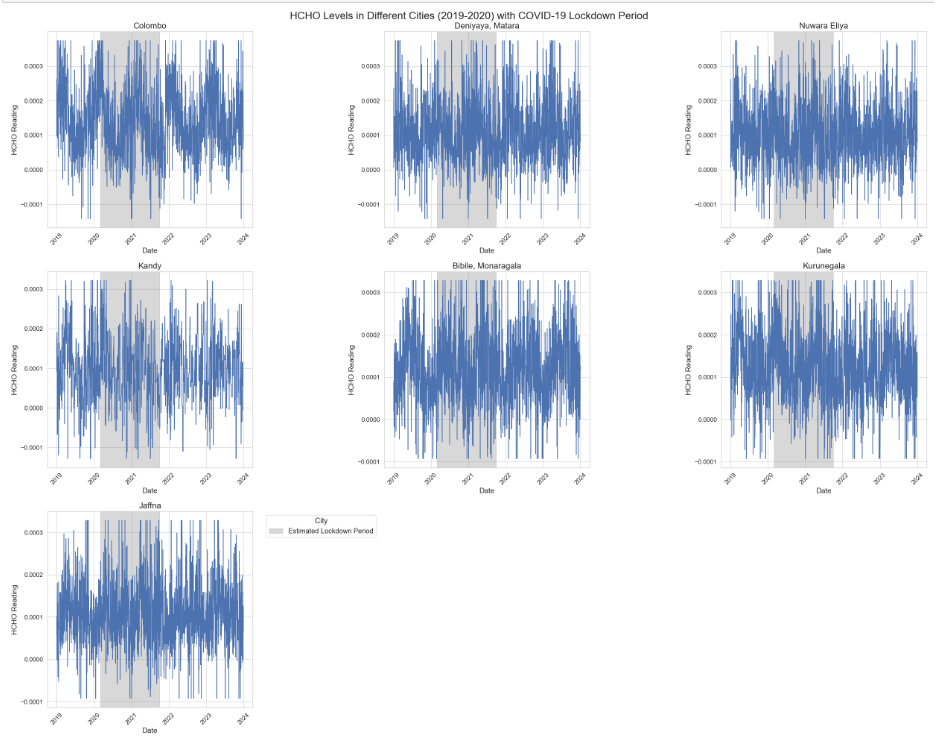


In above visualizations we can see the average distribution of HCHO gas in each city in a different angle and combination of many factors affected for this distribution such as geographical location of the city, economic situation in the country, covid lockdown and many more reasons.

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



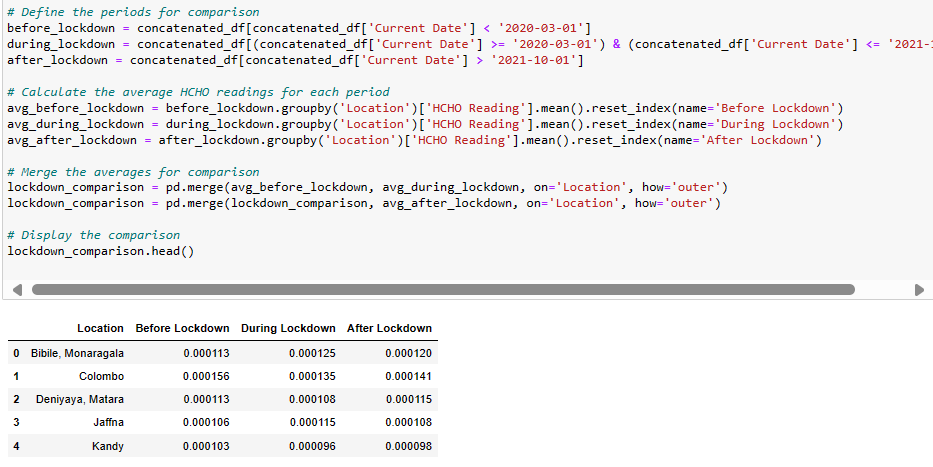
* Analyzed the distribution of HCHO gas variation in each city in lockdown period. In this situation I assumed that lockdown period is from 2020-03-01 to 2021-10-01 due to various situations in the country.



In most of cities there is a clear downfall of HCHO emission in this lockdown period. Situation of the country and stopping industrial activities may be the main reason for this.

After lockdown season we can identify clear deviation of gas levels.

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

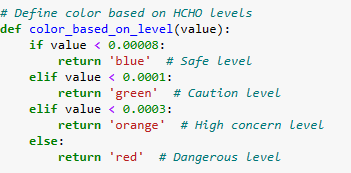


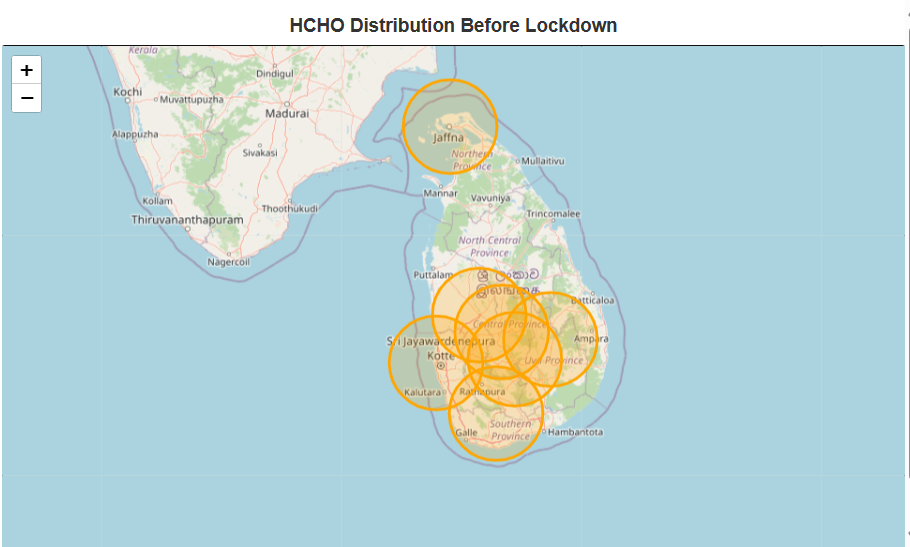
According to data in Sri Lanka I assumed that covid lockdown season extends from 2020 / 03 / 01 to 2021 / 10 / 01 and in this period, Sri Lanka faced for 3 main lockdowns and many other restricted seasons.

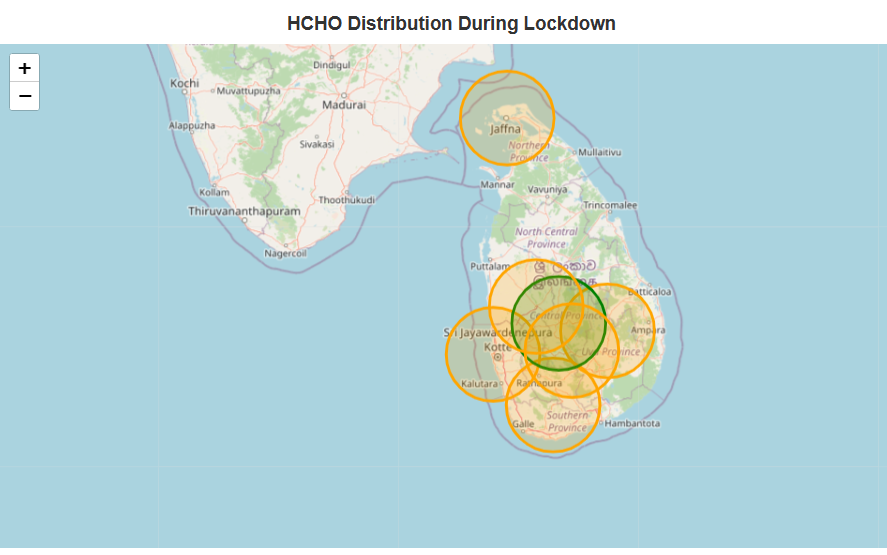
* Taken the HCHO levels and how they effect for human (organization, n.d.).
* safe level to prevent irritation for most people: 0.00008 mol/m²
* caution level where discomfort and irritation can begin: 0.0001 mol/m²
* high concern level indicating significant risk of respiratory issues: 0.0003 mol/m²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

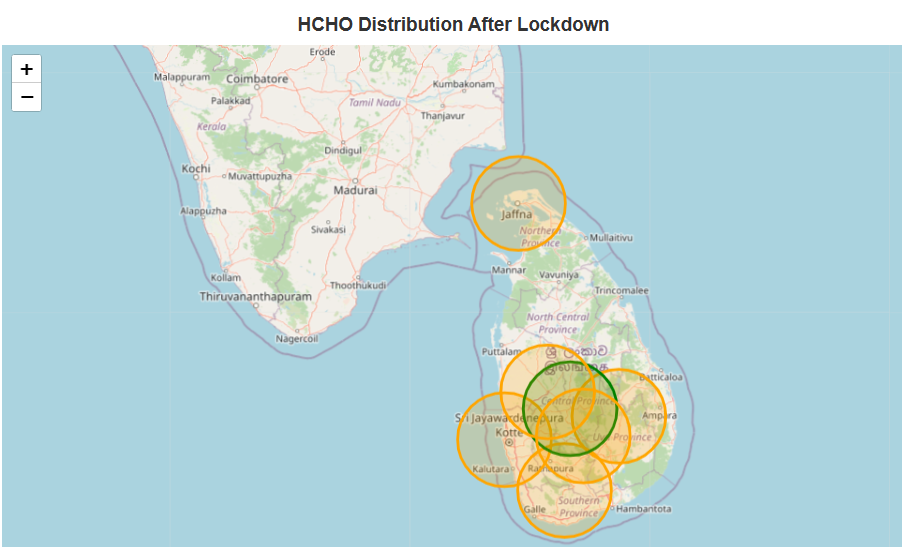
using this visualized how gas levels varies in different cities using periods such as before, during and after lockdown using folium heatmaps.

Logic used for making the heatmap

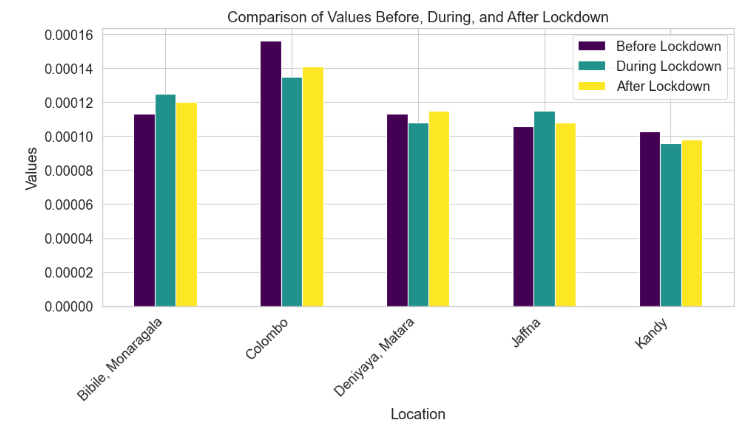








* Taken the average HCHO gas distribution with in above 3 seasons.



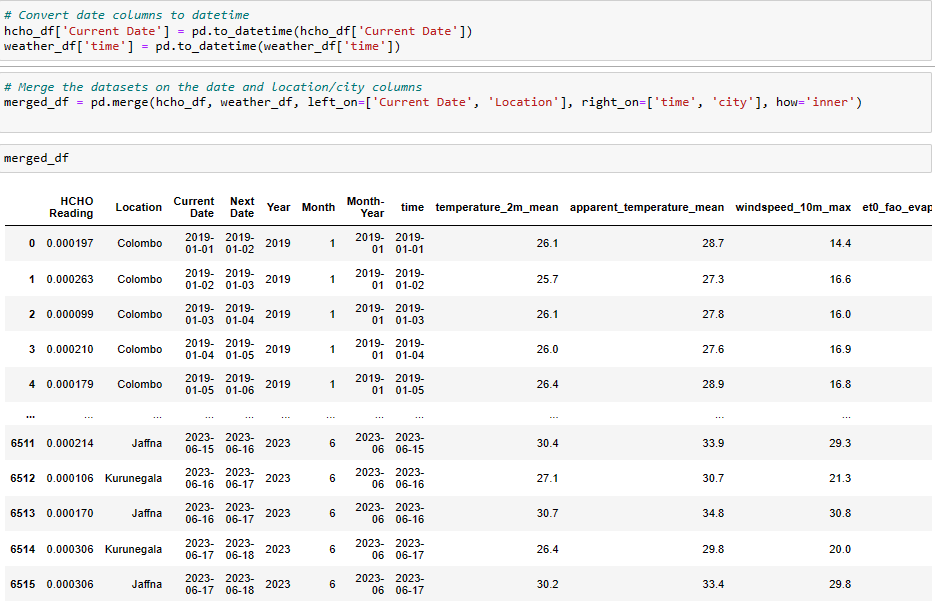
## Further analysis using external data

For further analysis used 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.

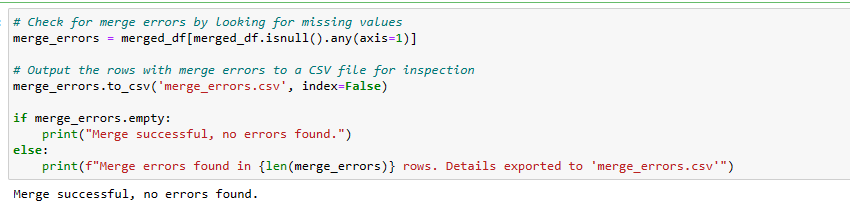
'Colombo’, ‘Kandy’, ‘Jaffna’, ‘Matara’, ‘Kurunegala'.

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



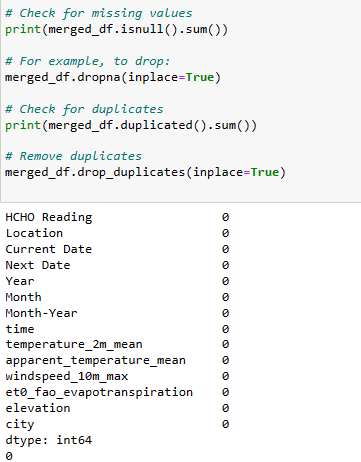
For merging these datasets I used inner join were used and from left table, current date and Location is used and from right table, time and city is used.

* Checking for merging errors



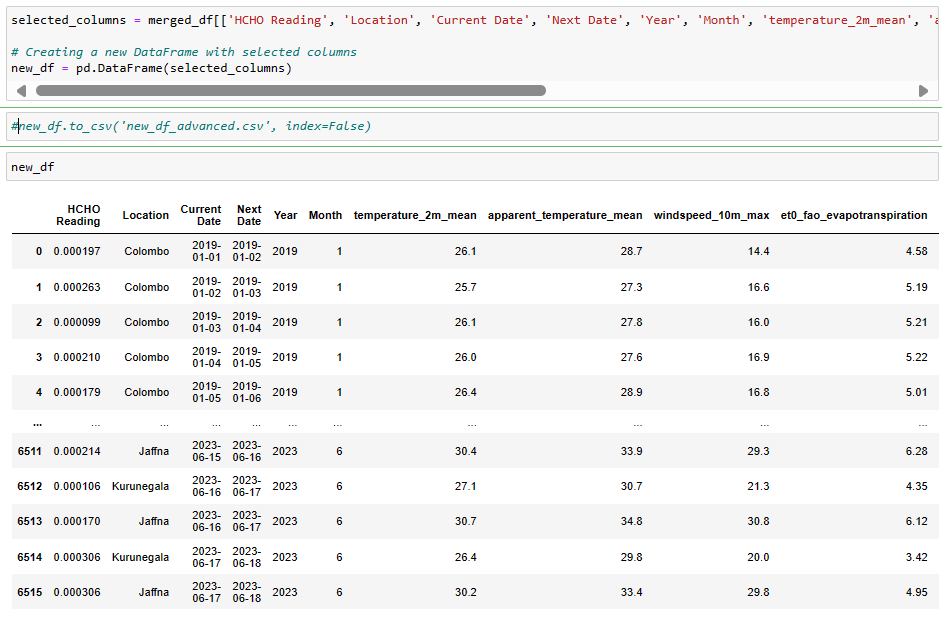
No any errors in merging and data merged successfully.

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

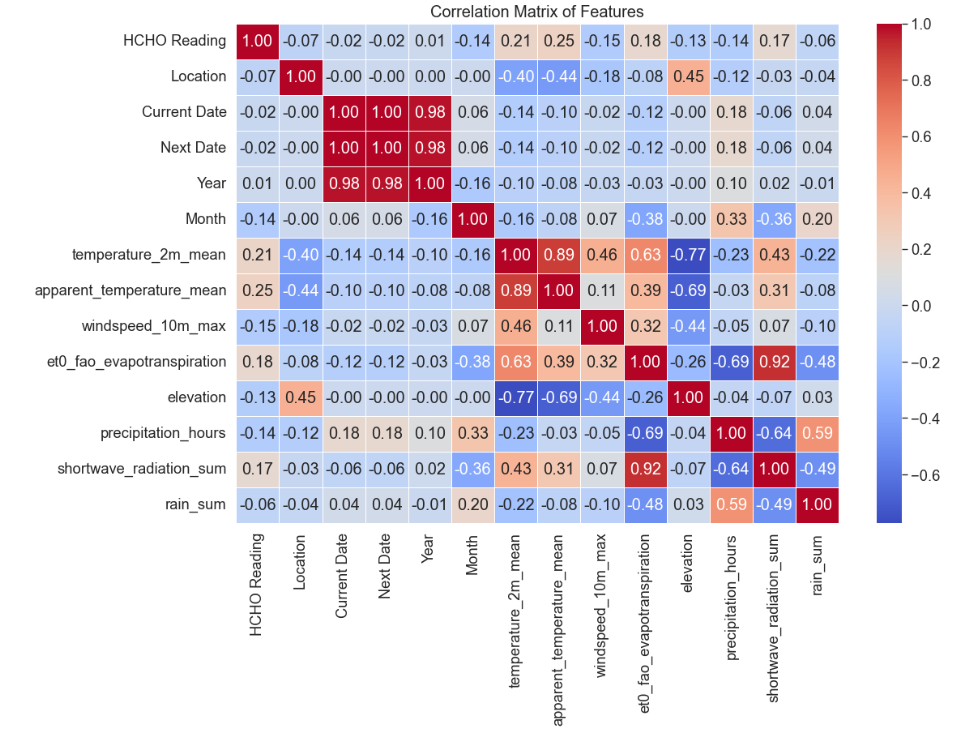


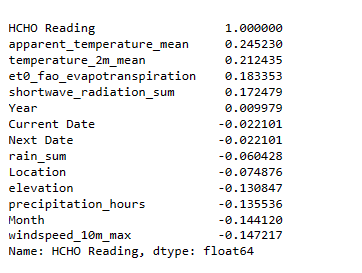
No any duplicated values in new data frame.

* Saving the final dataset for further analysis.

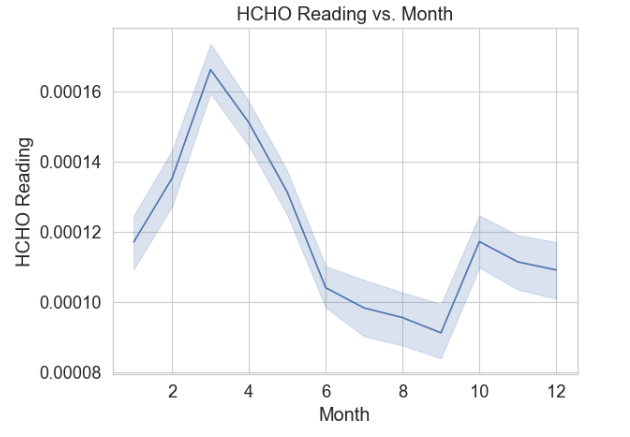


* Correlation matrix



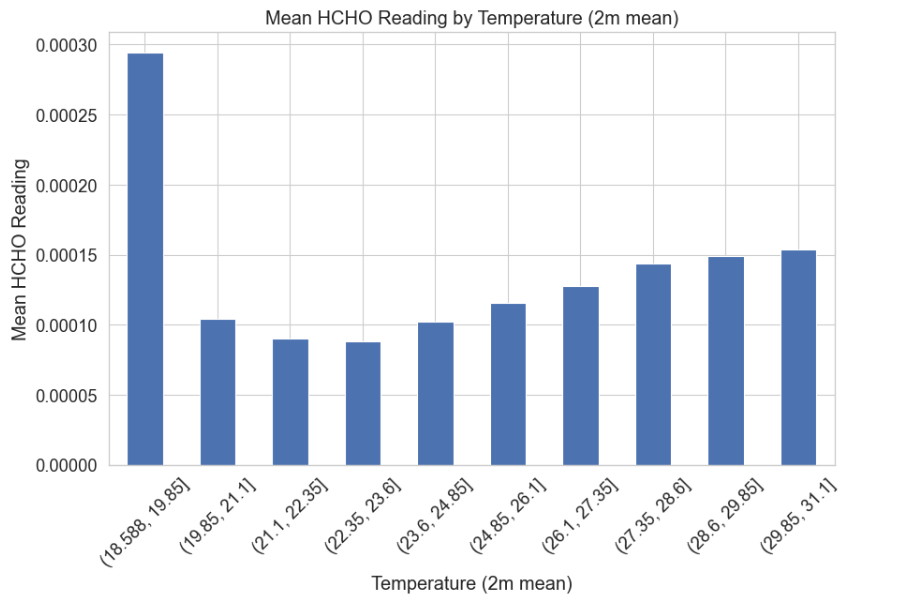


* Above matrix shows the correlation of features and this is an important feature when training the model. Some features such as temperature, apparent temperature, evaporation and shortwave radiation sum are highly positively correlated and there are slightly correlated features and some other features are negatively correlating.
* First, analyzed HCHO distribution over months.



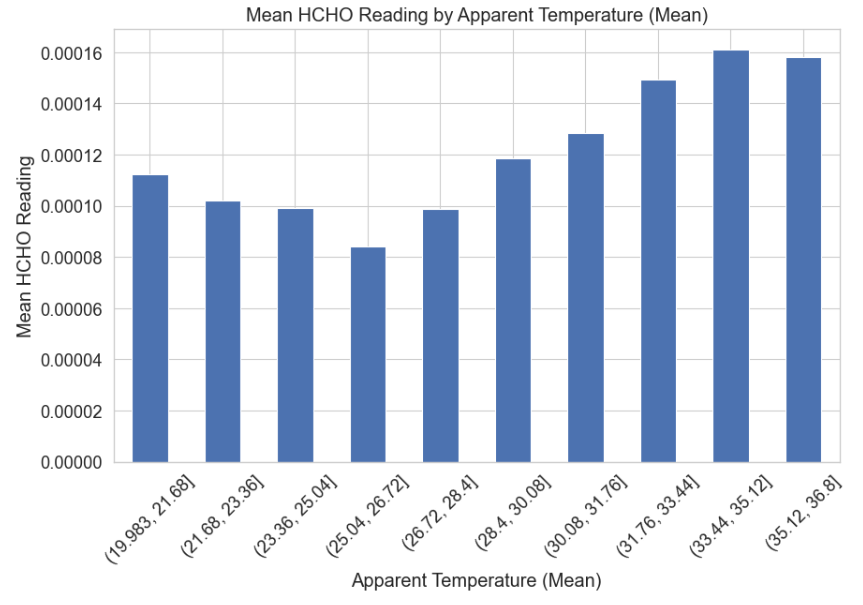
This analysis is done using the previous original data frame with all cities and this also following a close distribution. In first few months reading was high and in middle season distribution is quite low and in last few months we can see some deviation.

* HCHO distribution vs mean temperature and temperature ranges.



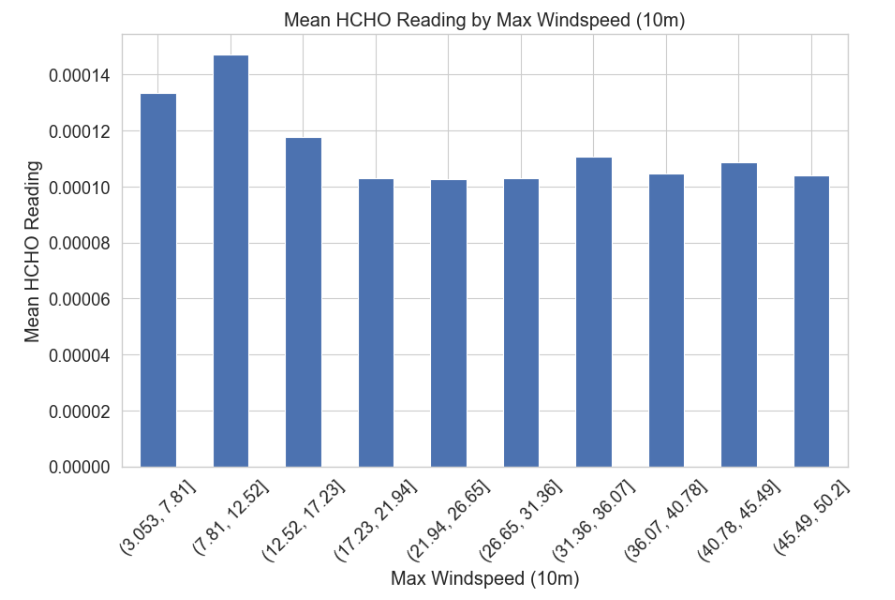
Temperature is a main factor that affect for a distribution of a gas. In this chart less ranges affect mostly for higher distribution of the gas and average temperature levels affect for less distribution and again high temperature levels affect for more distribution. In this graph we can say that in less temperatures and high temperatures spread of gas is high.

* HCHO distribution vs apparent temperature and temperature ranges.



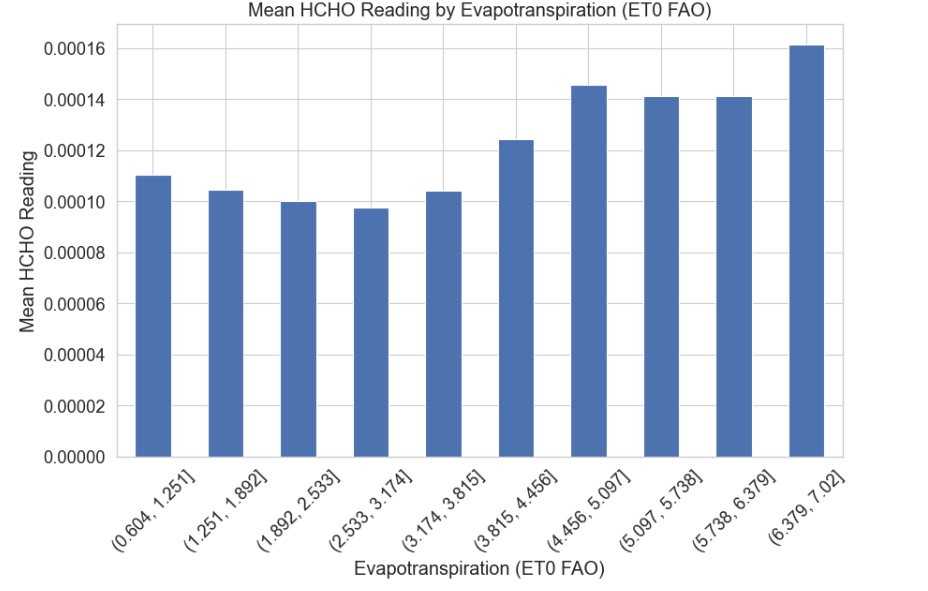
When considering about the apparent mean of the temperatures vs mean HCHO reading we can identify some opposite insights because in this, when temperature is increasing the gas distribution is more increasing. However, this graph follows a main feature of last distribution which is high distribution in low and high temperatures and low distribution in mid temperatures.

* HCHO distribution vs Max windspeed.



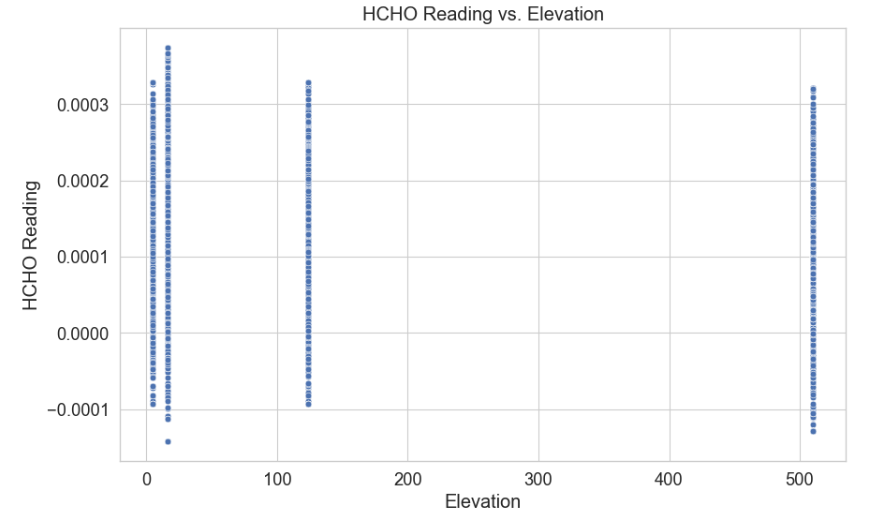
Wind speed plays an important role in distribution of gasses. When we analyze the above bar chart we can identify that low speeds affects for high distribution of the gas and when the speed of the wind gradually increasing the distribution of the gas did not show more deviations and it remains same.

* HCHO distribution vs Evaporation



Above bar chart describes HCHO gas distribution with evaporation. When analyzing this gas distribution is quite low in low and moderate evaporation levels and when evaporation level is high gas distribution is also increasing.

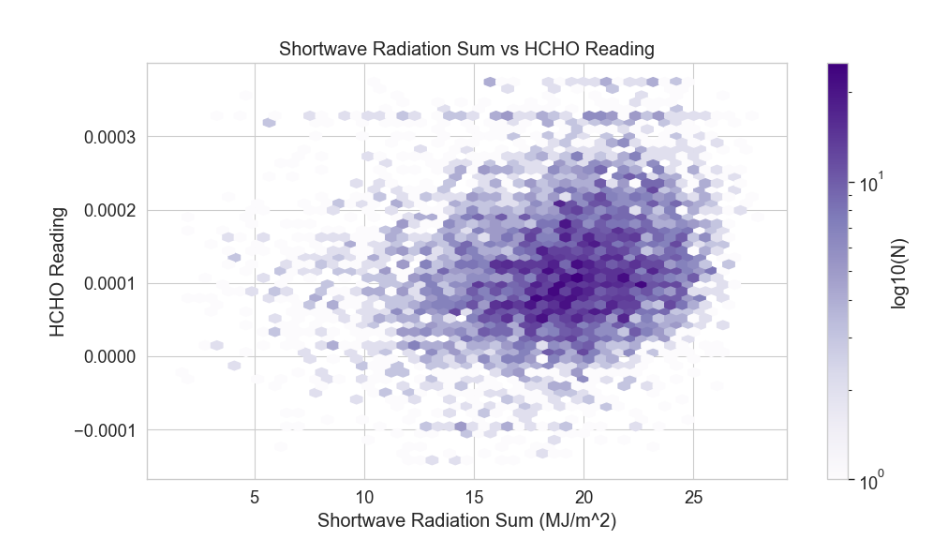
* HCHO distribution vs Elevation



In this analysis we can say that number of readings are high in low elevations that high elevations. When elevation is changing there may be differences in gas distributions because sometimes external features such as pressure and some other factors affect in many ways.

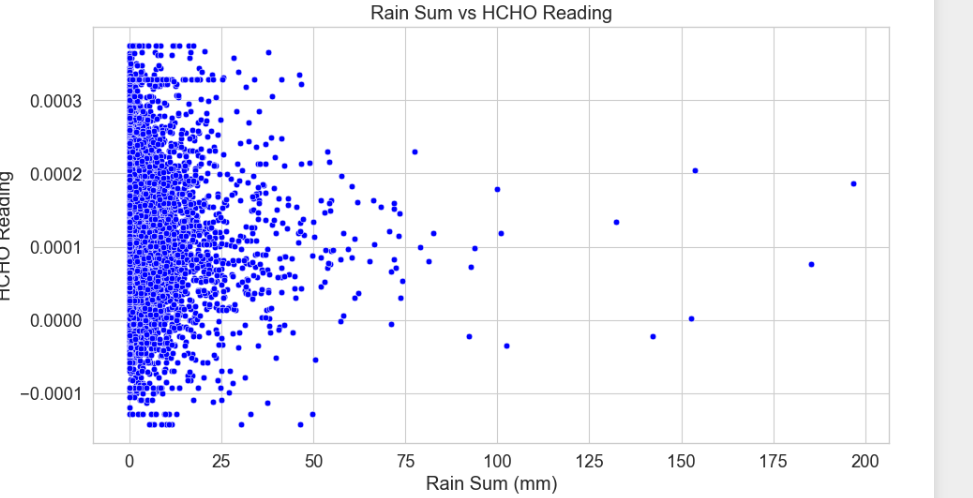
* Shortwave radiation sum vs HCHO Reading





Above plots describes the changes of HCHO readings with the changes of shortwave radiation in two different ways. HCHO distribution is quite low in low radiation levels such as 5, 10. When radiation level increasing, we can have more HCHO readings. Most occurrences are in 15, 20 radiation range and 0.0001 – 0.0002 range. However, HCHO readings are high when radiation is high.

* Rain sum vs HCHO Reading

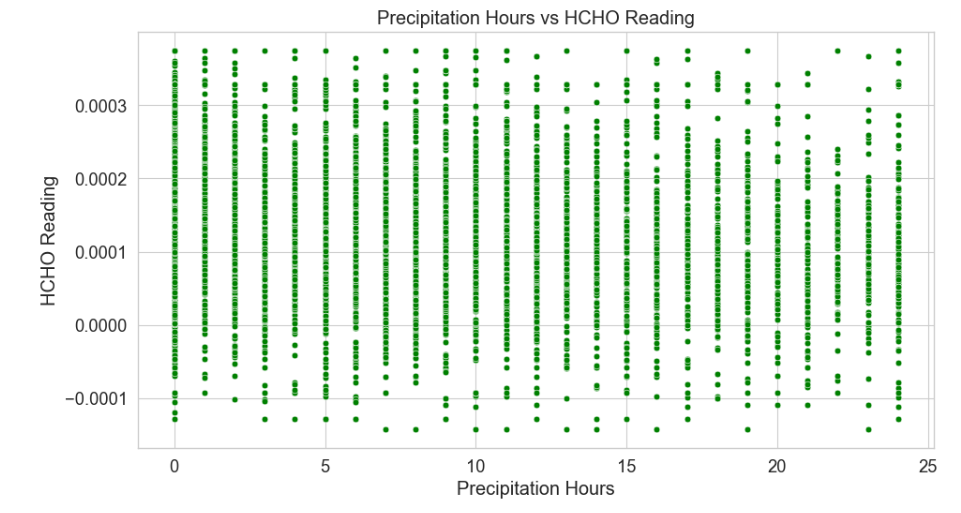




In this analysis we can analyze distribution of sum of the rain fall with HCHO reading. In this visualization we can understand that most of readings are good in less rainfalls. When there is no rainfall, we can say that HCHO distribution is high or moderate.

* Precipitation hours vs HCHO Reading





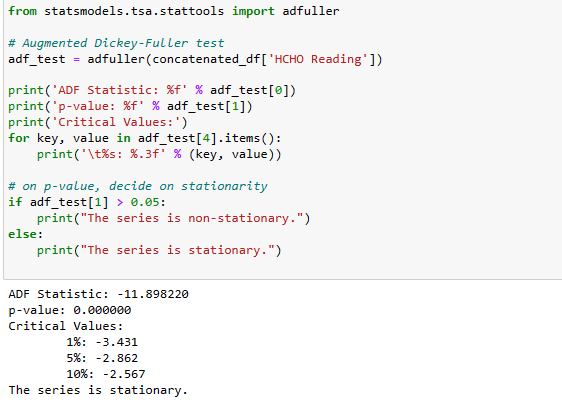
Above analysis explains about the HCHO reading with precipitation hours. In this we cannot make a clearer decision but we can say that when there are less hours of precipitation readings are quite high and moderate. However, rainfall makes less impact for the gas distribution.

## Conclusion of spatial analysis

In conclusion this analysis mainly discussing how external factors affecting for changes in HCHO gas emission, changes in gas levels and changes in gas distributions. In Sri Lanka, mainly few features such as industries, agriculture affect for the emission of HCHO gas (Lanka, n.d.). When considering about the distribution of gasses, temperature, windspeed, precipitation and many other factors are contributing. Covid lockdown season is also playing a major role in HCHO distribution according to this dataset. Likewise, there is a detailed analysis about HCHO distribution in Sri Lanka over the time.

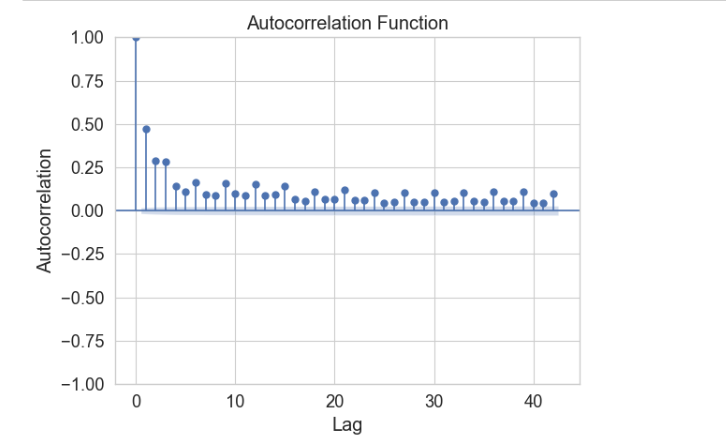
# Machine Learning

* Before training the model, tested weather the data set is stationary. For that I used ADF statistics, P value and critical values.



With the output values, can make a confirmation that the model is stationary and this dataset is suitable for making time series models.

* Autocorrelation function



Plotted the autocorrelation function of the “HCHO Reading” feature of the concatenated dataset.

## Model training and evaluation

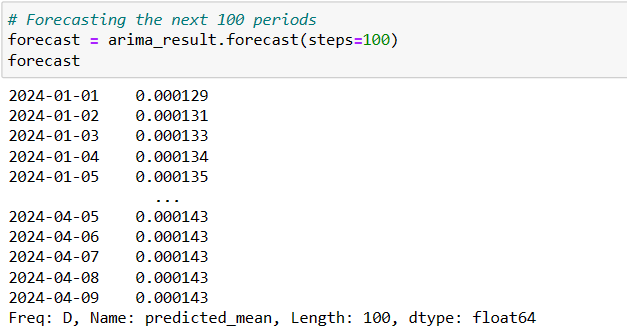
When making the time series model I started with a base model and then tried more complex models and evaluated them.

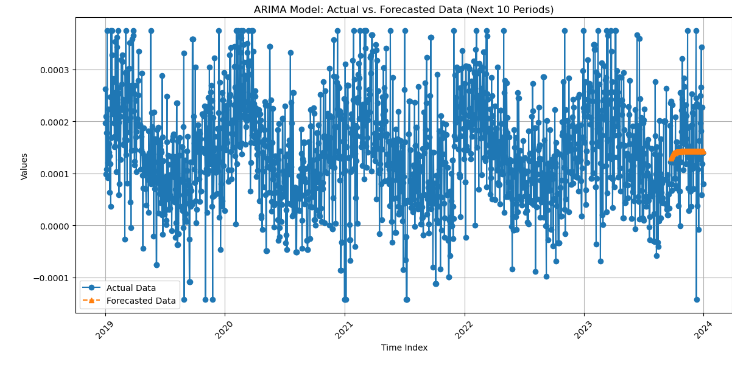
1. ARIMA (**Autoregressive Integrated Moving Average)**

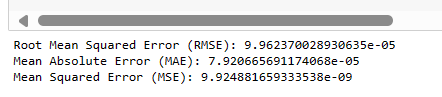
ARIMA is a good base model for time series forecasting started forecasting using this model. Applied this model for only data in Colombo district.



Results



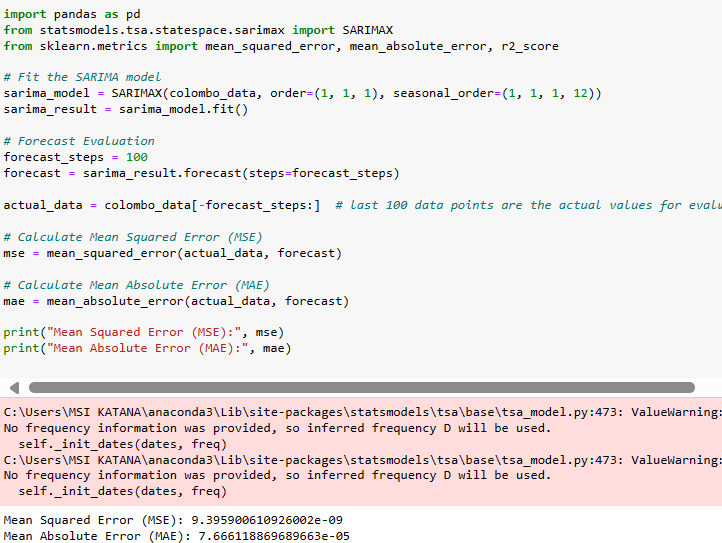


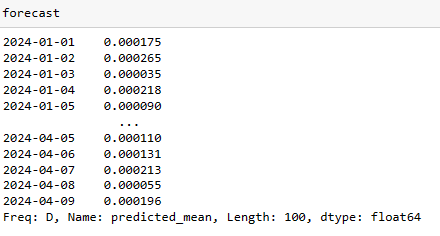


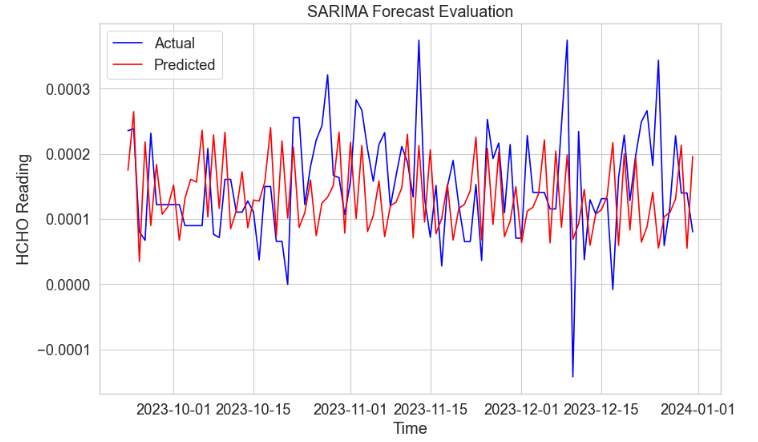
These results indicated that the ARIMA model not giving expected results. Then I used SARMAX model which predicting better results with seasonal data

2. SARIMAX (**Seasonal Autoregressive Integrated Moving Average with eXogenous regressors**.)

* Applying SARIMAX model for Colombo district

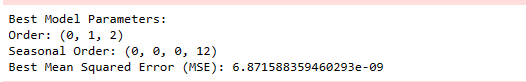


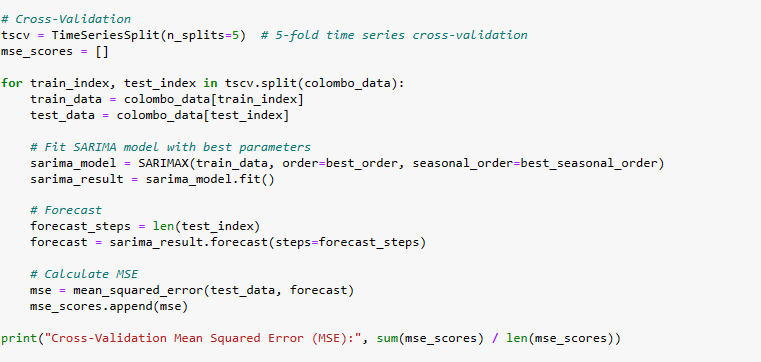


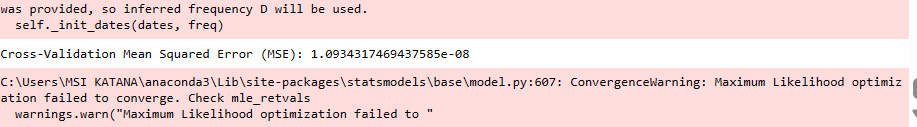


* Used hyperparameter tunning and cross validation techniques to tune the model and increasing accuracy. I did this tunning for Colombo data as a demonstration because model tunning is computationally expensive.

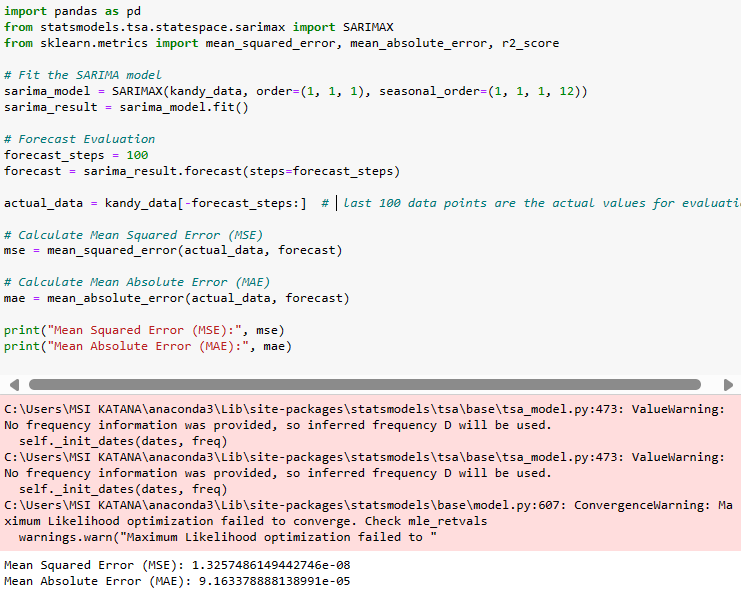


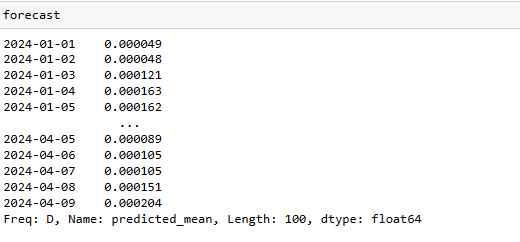


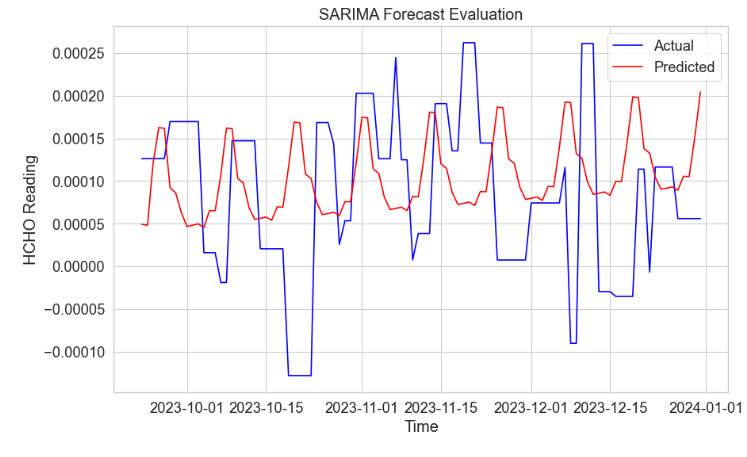




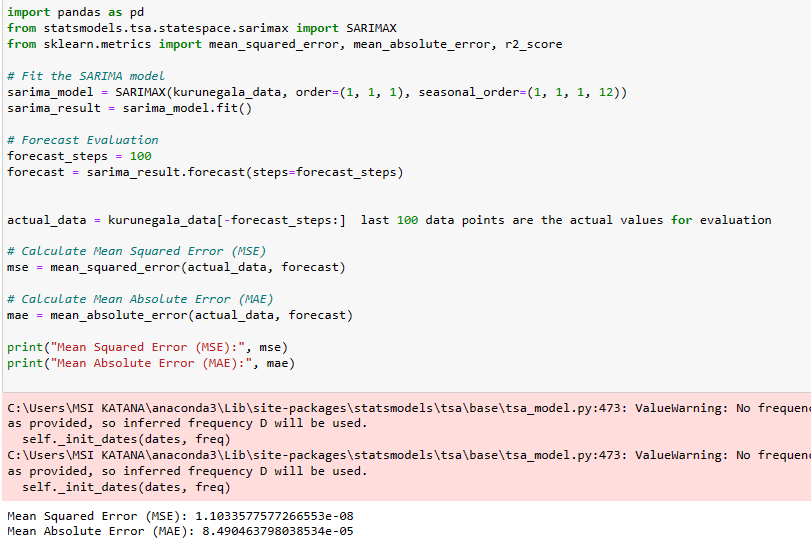
* Applying SARIMAX model for Kandy

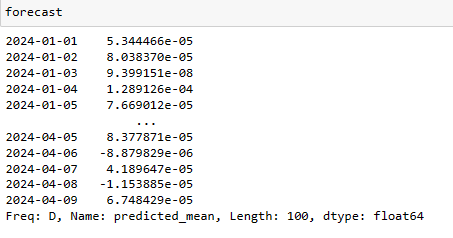


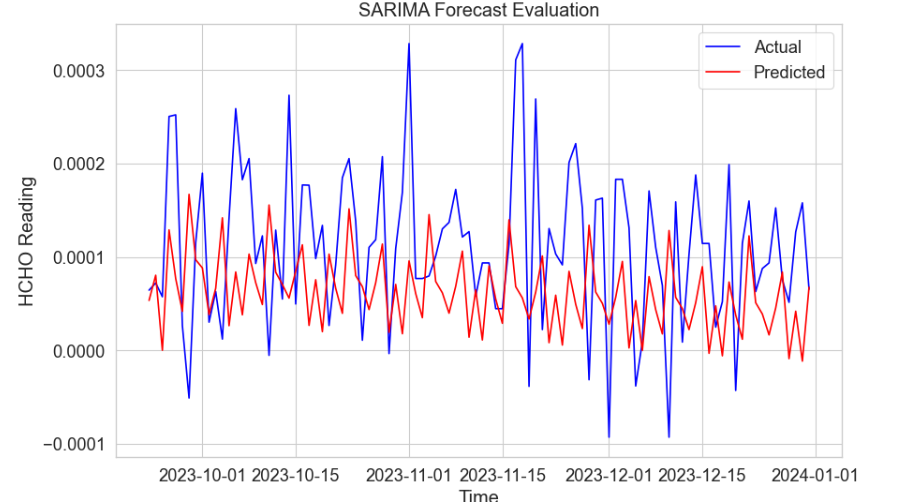




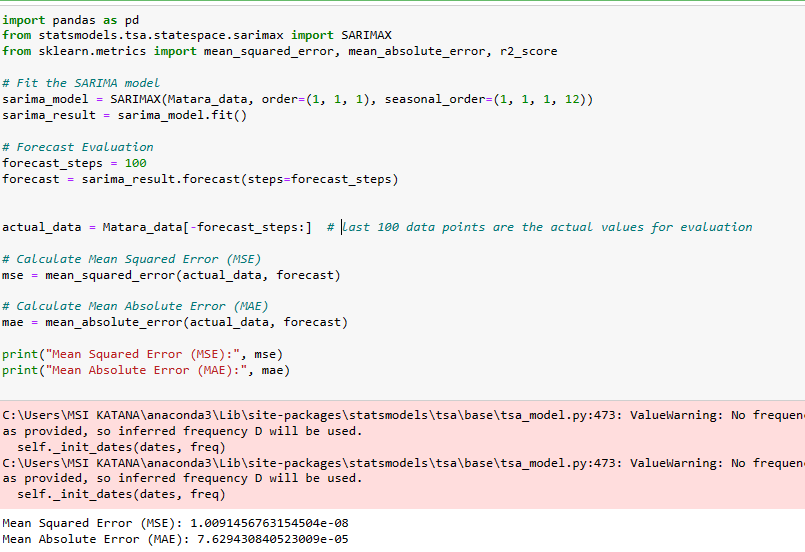
* Applying SARIMAX model for Kurunegala

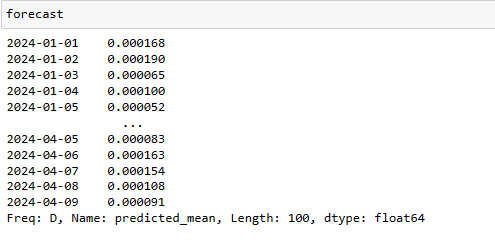






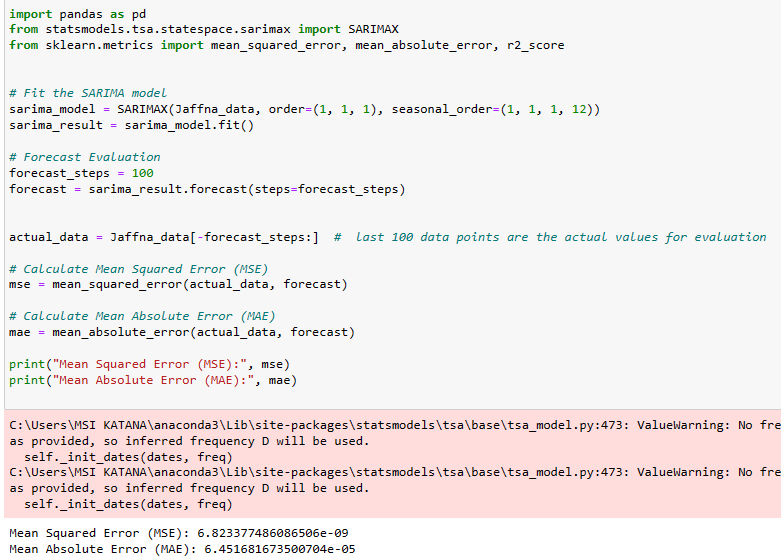
* Applying SARIMAX model for Matara / Deniyaya

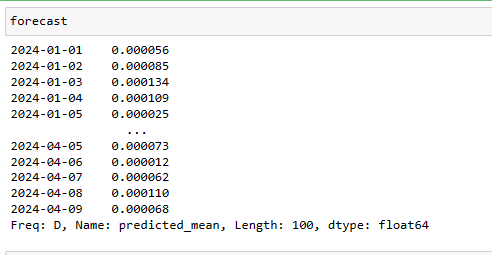


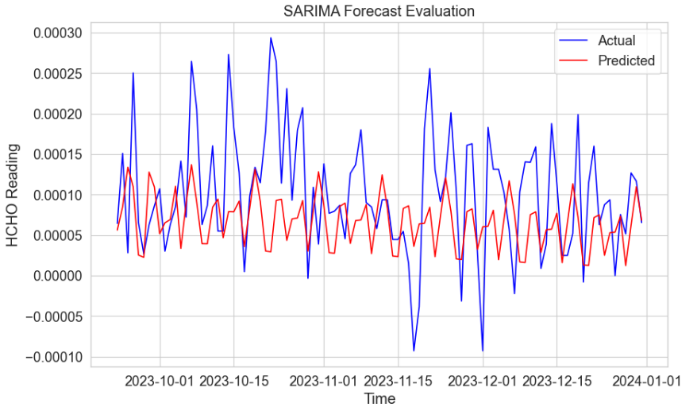




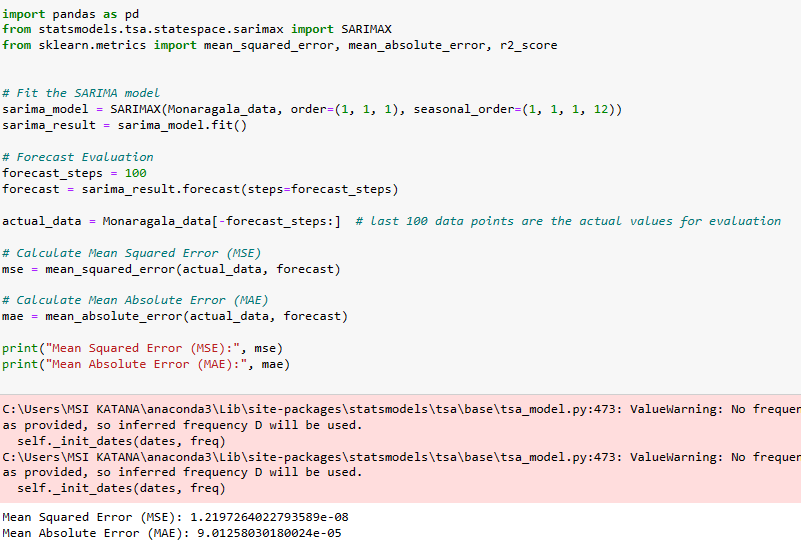
* Applying SARIMAX model for Jaffna

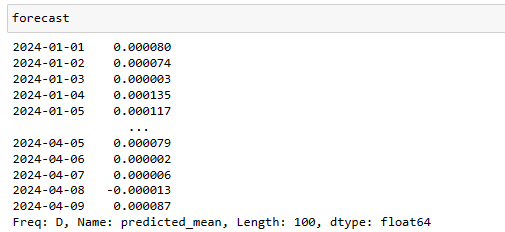


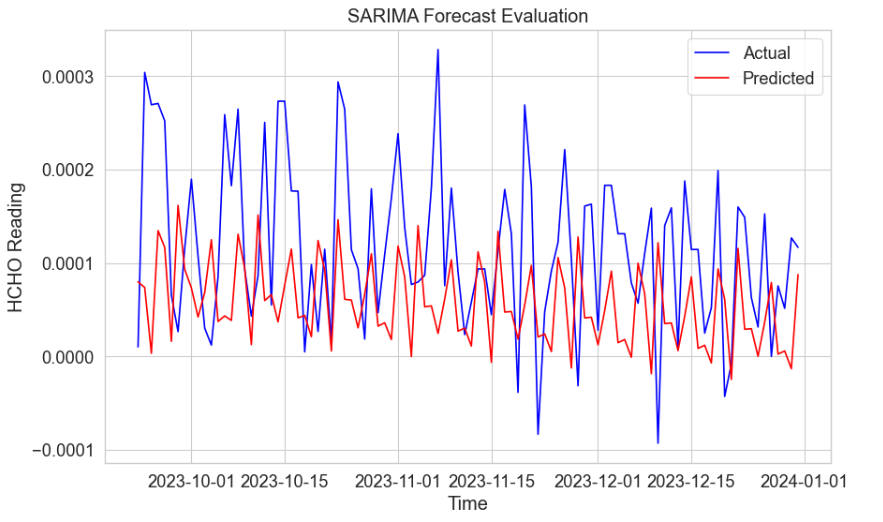




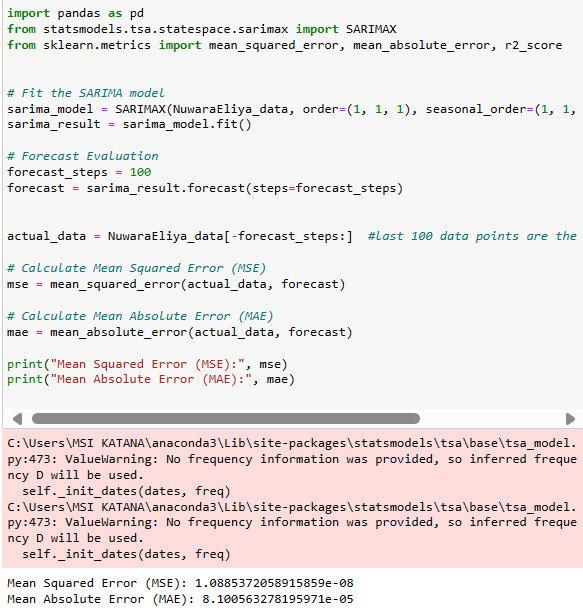
* Applying SARIMAX model for Monaragala / Bibile

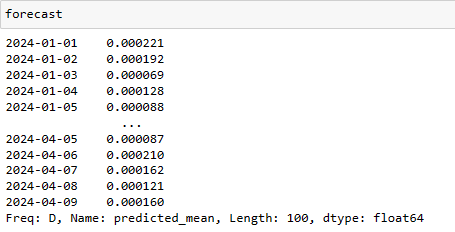


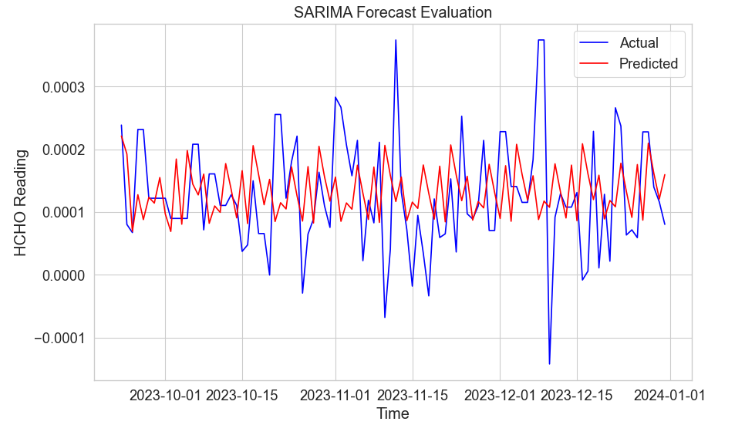




* Applying SARIMAX model for Nuwara Eliya





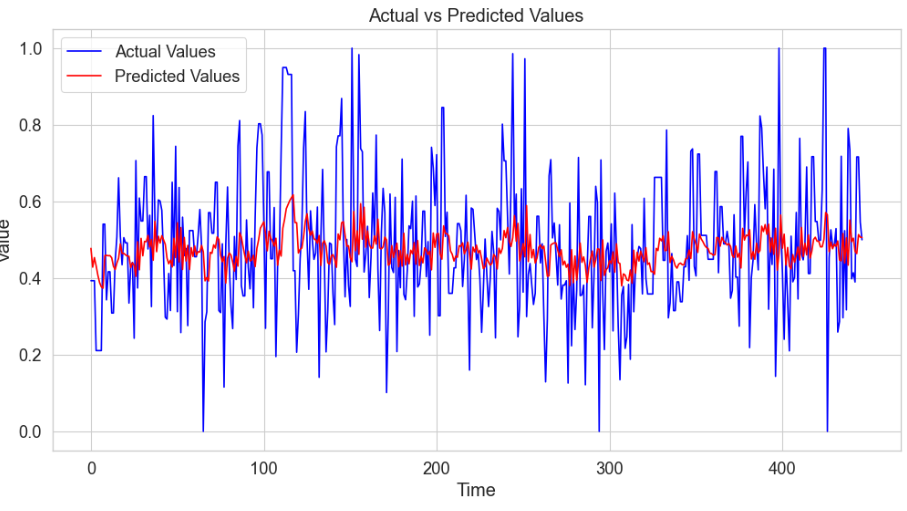


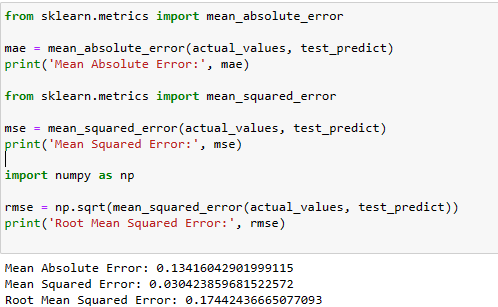
When applying SARIMAX model the evaluations are ok but these model predictions fit for some districts but not fit for some other districts. For solving this problem, I applied LSTM model.

3. LSTM (Long Short Term – Memory)

* LSTM for Nuwara Eliya



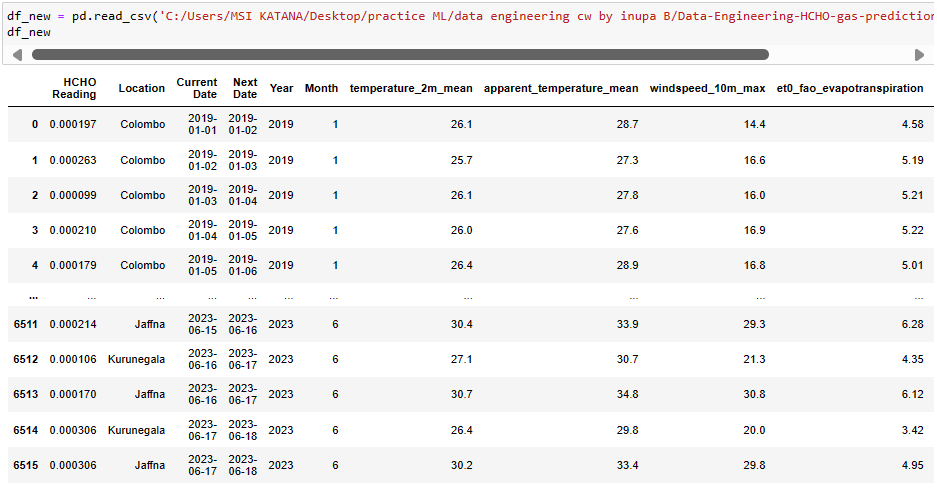




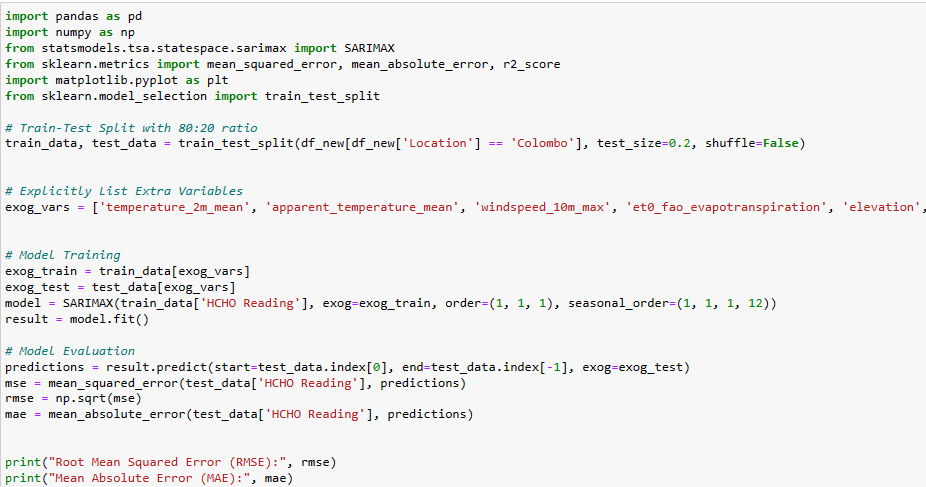
## Model training using external factors

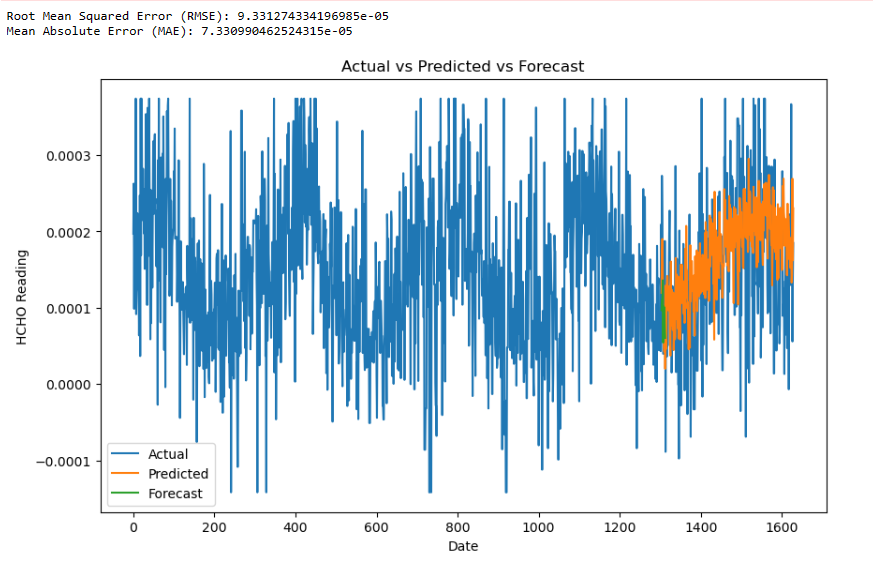
3. When considering about external factors we can make timeseries models for predicting future values with the impact of external factors. Used data only in Colombo and made timeseries and for that applied SARIMAX model.

For this I used my dataset that I used for advanced analysis with many external factors.



For external factors I used temperature\_2m\_mean, apparent\_temperature\_mean, windspeed\_10m\_max, et0\_fao\_evapotranspiration, elevation, precipitation\_hours, shortwave\_radiation\_sum, rain\_sum.





Using SARIMAX with external factor make accurate predictions but in this case absence of data related to all districts is the main problem. Therefore, I made a demonstration for data using Colombo district.

# Limitations and future work.

* When discussing about limitations, main limitation is the data limitation. In the provided dataset number of records for each city is enough but if there any more records that will be a reason for making more accurate predictions.
* In the selected dataset with external features we cannot find data related to some cities like Monaragala and Deniyaya etc.
* Hyperparameter tunning and cross validation is a best way to improve model performance but it is computationally expensive and taking lot of time.
* When using LSTM models, we can use hyperparameter tunning for improve the performance and that is the best way to capture the seasonality of actual data but it also computationally expensive.

# References

* <https://www.atsdr.cdc.gov/formaldehyde/> describes the HCHO gas levels that affect to human.
* External dataset Link [https://www.kaggle.com/datasets/rasulmah/sri-lanka-weather-dataset](#_References).
* SL agriculture data <http://www.statistics.gov.lk/Agriculture/StaticalInformation/new>.