Report

Module Name: Data Engineering

Module Code: **CM2606**

Module Leader: **Mr. Mohamed Ayoob**

Student Name: **Yenuka Rajapaksha**

IIT ID: **20221359**

RGU ID: **2237044**

Table of Contents

1. Introduction

This report gives a summary of the data preprocessing, findings, visualizations, and the model implementation of the analysis of HCHO gas levels in Sri-Lanka. The relationships between the HCHO levels and external factors were also researched in this assignment. Time series algorithms were used to predict the future HCHO levels, and these data were represented in an interactive way using a power bi dashboard.

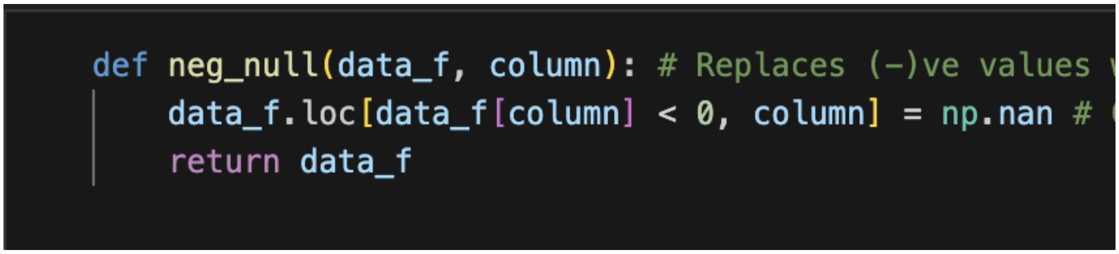
2. Data Preprocessing

2.1 Libraries Imported

- NumPy
- Pandas
- Seaborn
- Matplotlib
- scikit-learn.

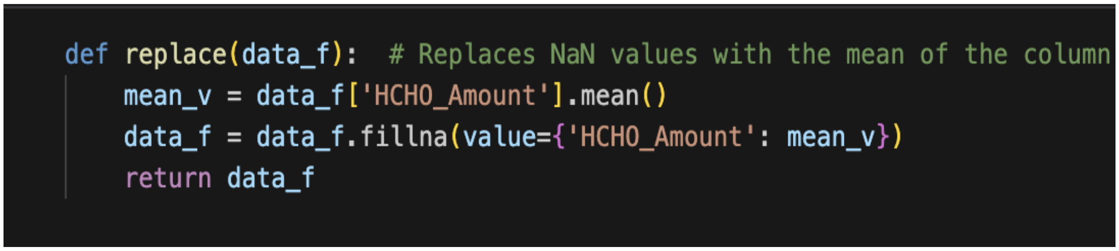
2.2 Functions

2.2.1 Replacing negative values with null values.



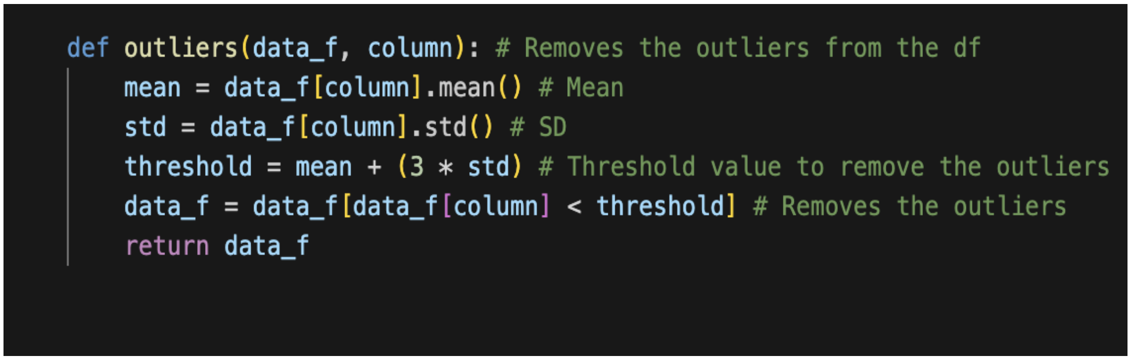
```
def neg_null(data_f, column): # Replaces (-)ve values
    data_f.loc[data_f[column] < 0, column] = np.nan #
    return data_f
```

2.2.2 Replace null values with the mean of the column.



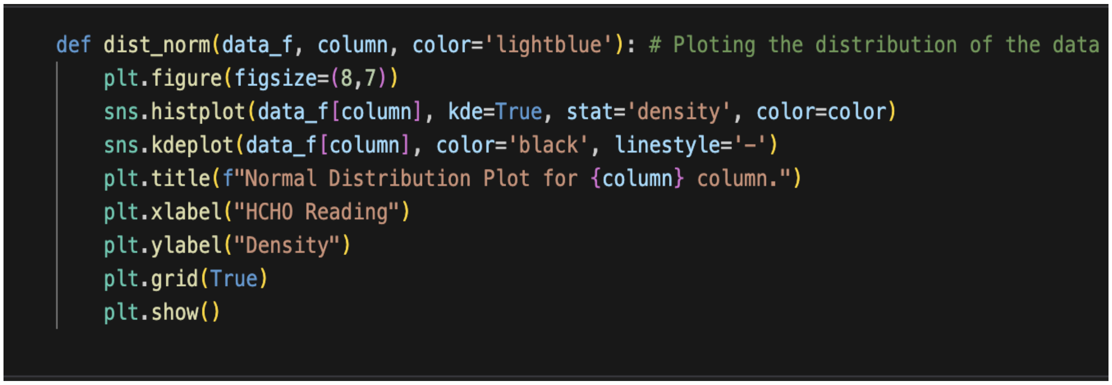
```
def replace(data_f):  # Replaces NaN values with the mean of the column
    mean_v = data_f['HCHO_Amount'].mean()
    data_f = data_f.fillna(value={'HCHO_Amount': mean_v})
    return data_f
```

2.2.3 Removing outliers from the data.



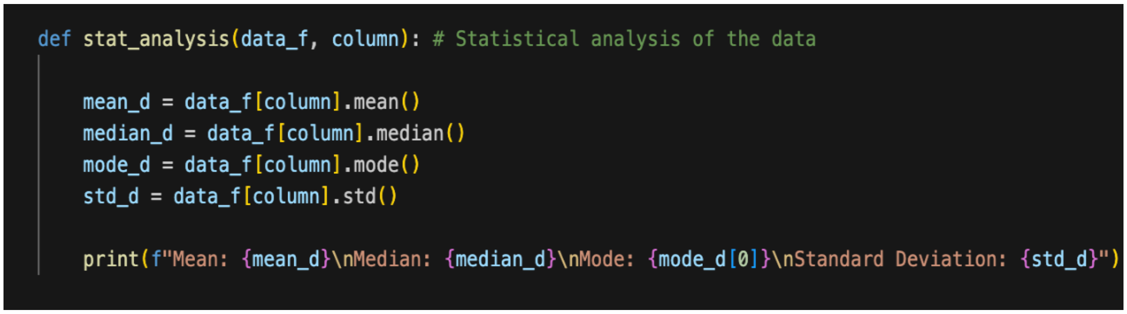
```
def outliers(data_f, column): # Removes the outliers from the df
    mean = data_f[column].mean() # Mean
    std = data_f[column].std() # SD
    threshold = mean + (3 * std) # Threshold value to remove the outliers
    data_f = data_f[data_f[column] < threshold] # Removes the outliers
    return data_f
```

2.2.4 Plotting the normal distribution of the data



```
def dist_norm(data_f, column, color='lightblue'): # Ploting the distribution of the data
    plt.figure(figsize=(8,7))
    sns.histplot(data_f[column], kde=True, stat='density', color=color)
    sns.kdeplot(data_f[column], color='black', linestyle='-')
    plt.title(f"Normal Distribution Plot for {column} column.")
    plt.xlabel("HCHO Reading")
    plt.ylabel("Density")
    plt.grid(True)
    plt.show()
```

2.2.5 Statistical analysis of the data.

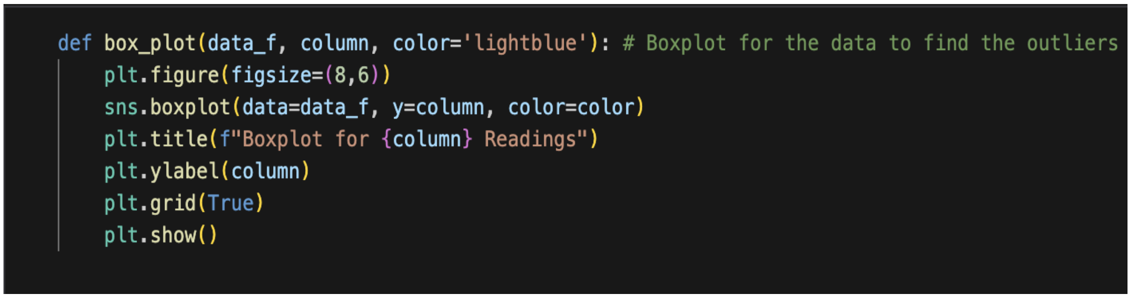


```
def stat_analysis(data_f, column): # Statistical analysis of the data

    mean_d = data_f[column].mean()
    median_d = data_f[column].median()
    mode_d = data_f[column].mode()
    std_d = data_f[column].std()

    print(f"Mean: {mean_d}\nMedian: {median_d}\nMode: {mode_d[0]}\nStandard Deviation: {std_d}")
```

2.2.6 Boxplot to check for outliers.



```
def box_plot(data_f, column, color='lightblue'): # Boxplot for the data to find the outliers
    plt.figure(figsize=(8,6))
    sns.boxplot(data=data_f, y=column, color=color)
    plt.title(f"Boxplot for {column} Readings")
    plt.ylabel(column)
    plt.grid(True)
    plt.show()
```

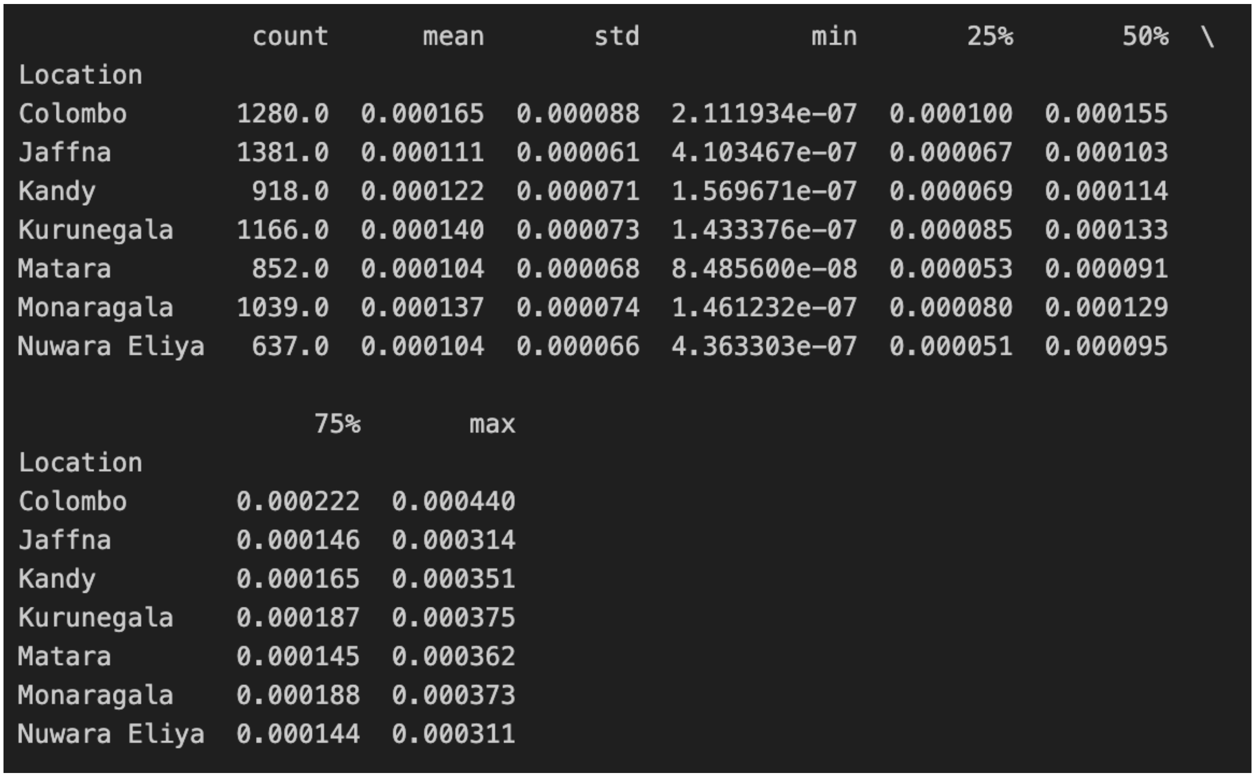
2.3 Removal of outliers

The dataset was broken down to individual datasets for each location and saved into a csv file. For each location the boxplot was drawn to check for outliers. From these boxplots it was evident that there were extreme values and negative values as well. The negative values were converted to null values and then they were filled with the mean of the column. Before replacing these null values, the outliers were dealt with by using the z-score, which was used as a threshold to remove the outliers in the data. This gave a better more related mean for the data.

2.4 Descriptive Statistics for the data

Descriptive statistics for the data gave a better understanding on the data. The data was cleaned, and the outliers were removed before getting these values which gives these values more validity. The mean, median, mode and the standard deviation was calculated here.

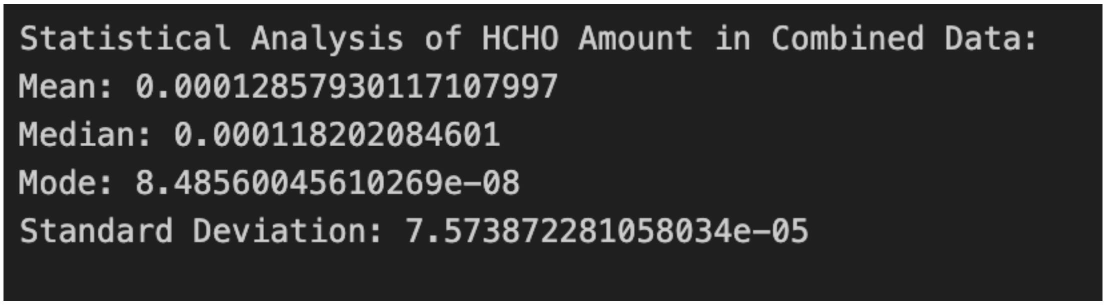
Results –



```
               count      mean       std           min       25%       50%  \
Location
Colombo       1280.0  0.000165  0.000088  2.111934e-07  0.000100  0.000155
Jaffna        1381.0  0.000111  0.000061  4.103467e-07  0.000067  0.000103
Kandy          918.0  0.000122  0.000071  1.569671e-07  0.000069  0.000114
Kurunegala    1166.0  0.000140  0.000073  1.433376e-07  0.000085  0.000133
Matara         852.0  0.000104  0.000068  8.485600e-08  0.000053  0.000091
Monaragala    1039.0  0.000137  0.000074  1.461232e-07  0.000080  0.000129
Nuwara Eliya   637.0  0.000104  0.000066  4.363303e-07  0.000051  0.000095

                   75%       max
Location
Colombo       0.000222  0.000440
Jaffna        0.000146  0.000314
Kandy         0.000165  0.000351
Kurunegala    0.000187  0.000375
Matara        0.000145  0.000362
Monaragala    0.000188  0.000373
Nuwara Eliya  0.000144  0.000311
```

This gives a detailed description of the data in a certain location.

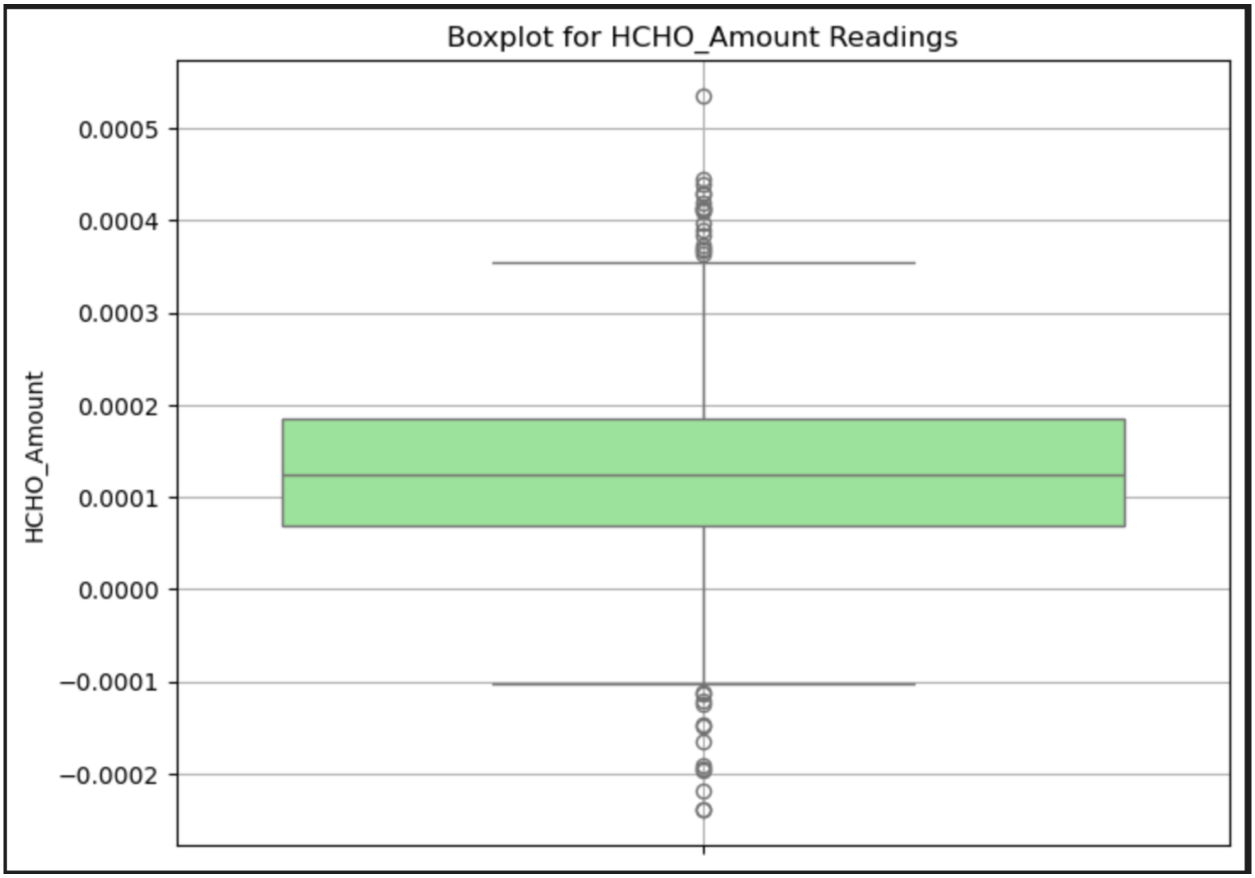


```
Statistical Analysis of HCHO Amount in Combined Data:
Mean: 0.00012857930117107997
Median: 0.000118202084601
Mode: 8.48560045610269e-08
Standard Deviation: 7.573872281058034e-05
```

This gives an overview of all the data in the whole dataset (all seven locations).

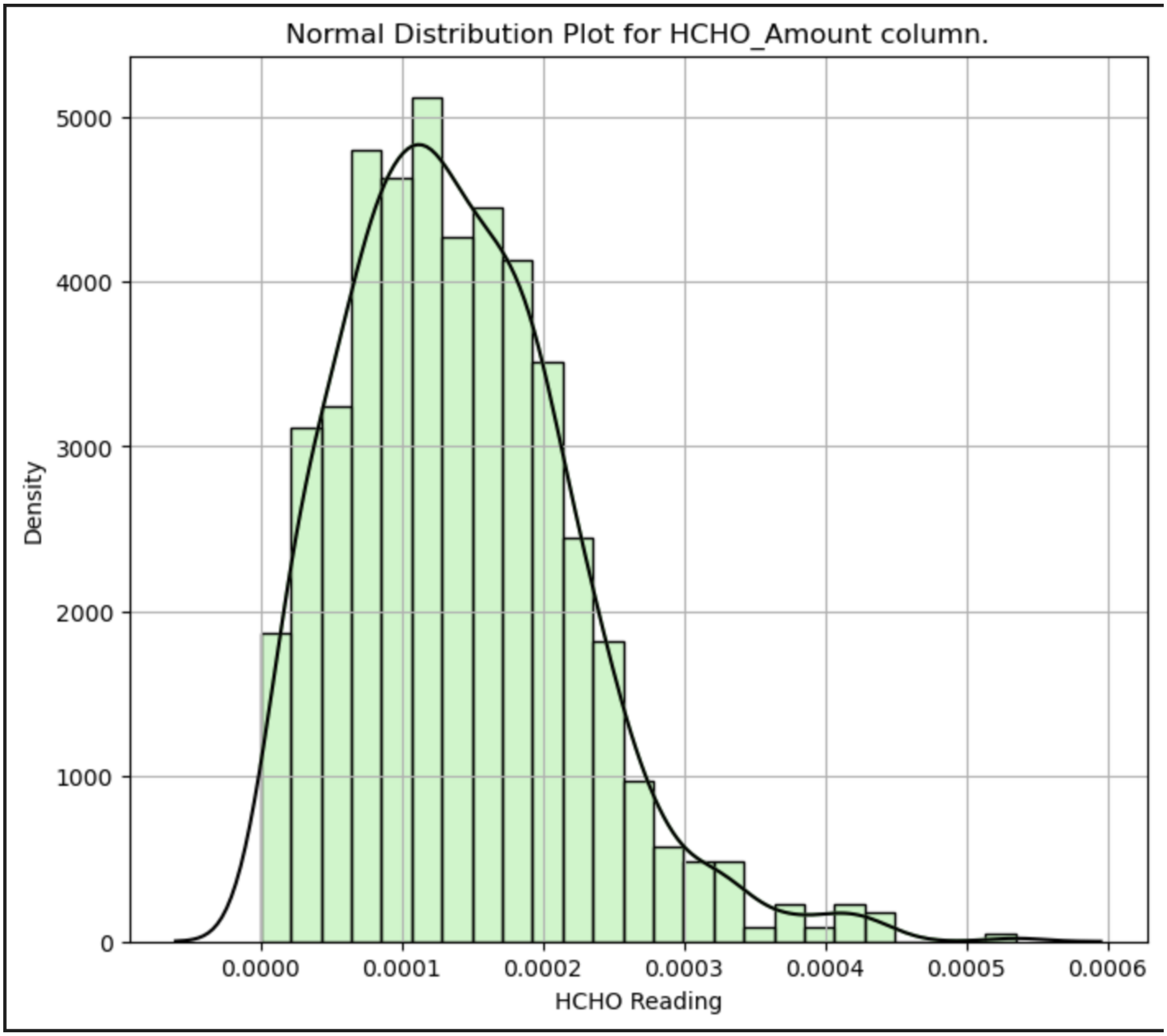
2.5 Visualizations

Example Boxplot -



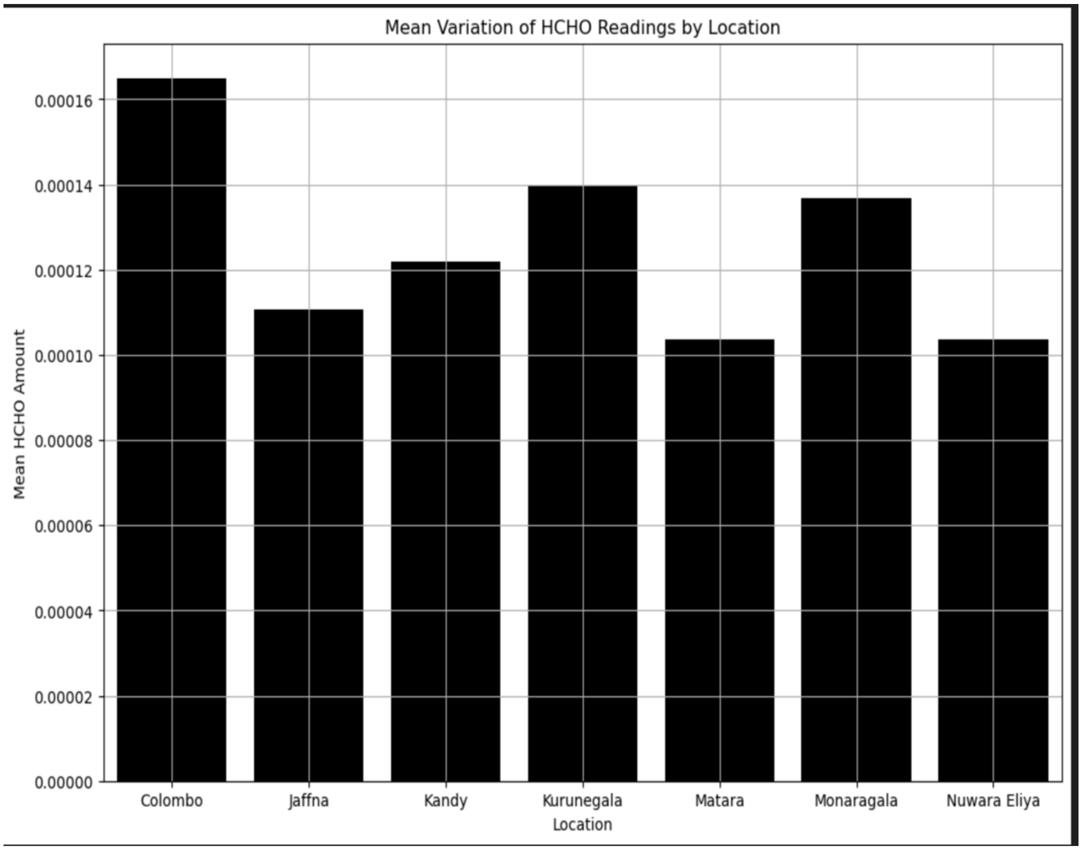
The negative values and the outliers were identified using the boxplot in all the data frames. As boxplots can sometimes be misleading the outliers were handled by calculating the z-score by using the above given function def outliers.

Examples normal distribution graph –



As the boxplot doesn't give a clear picture of the outliers, the normal distribution was used to get a proper understanding of the outliers of the data in the data frame for a certain location.

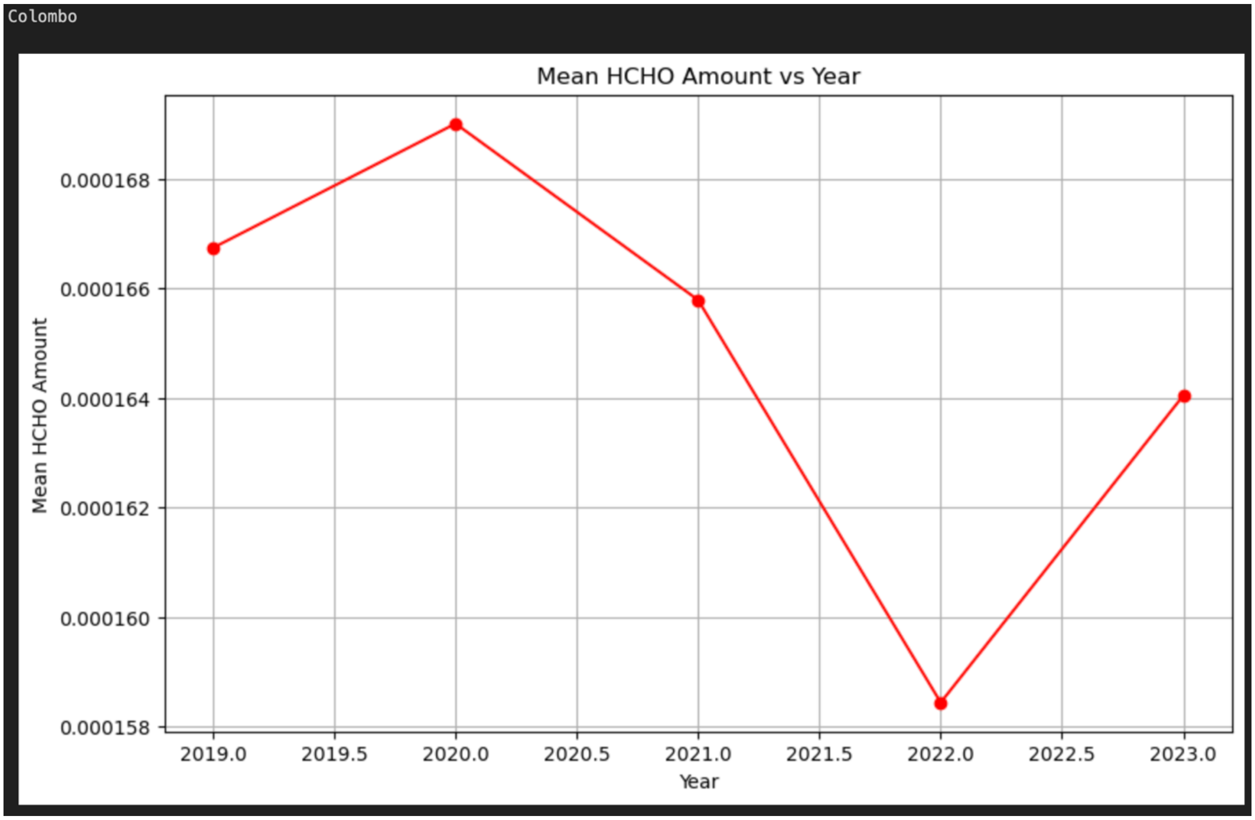
Bar chart –



This bar chart gives the mean HCHO level for four years of the given cities. As the bar chart shows Colombo has the highest mean HCHO level and Matara and Nuwara Eliya has the lowest mean HCHO levels for the four years (2019-2023). There could be several factors contributing to this finding. They are discussed below.

Variation of HCHO levels per year for each location –

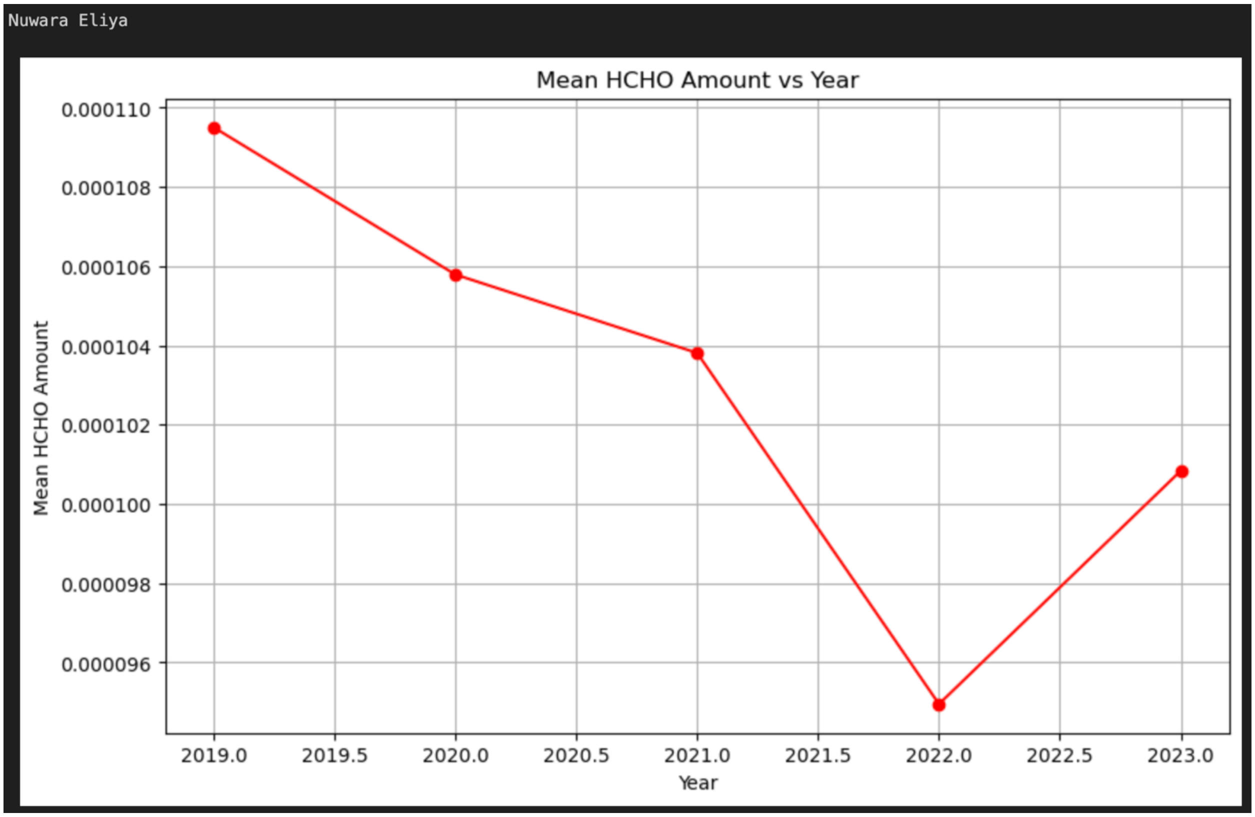
Colombo -



```
Colombo
```

The above line graph gives the levels of HCHO with the year. From 2019 to 2020 the HCHO level increases and then from 2020 to 2022 there's a sharp decline in the levels. The major factor that causes this could be the covid pandemic which resulted in nationwide lockdowns. Since most people were confined to their houses, the city's economic activity came to a halt, which probably resulted in a decrease in formaldehyde (HCHO) emissions into the environment. The decrease in atmospheric formaldehyde (HCHO) levels can also be attributed to reduced vehicle usage and the cessation of industrial plant operations during the lockdowns. From 2022 to 2023 a sharp increase in the HCHO levels can be observed, as covid lockdowns were lifted and people got back to their usual routines and lives.

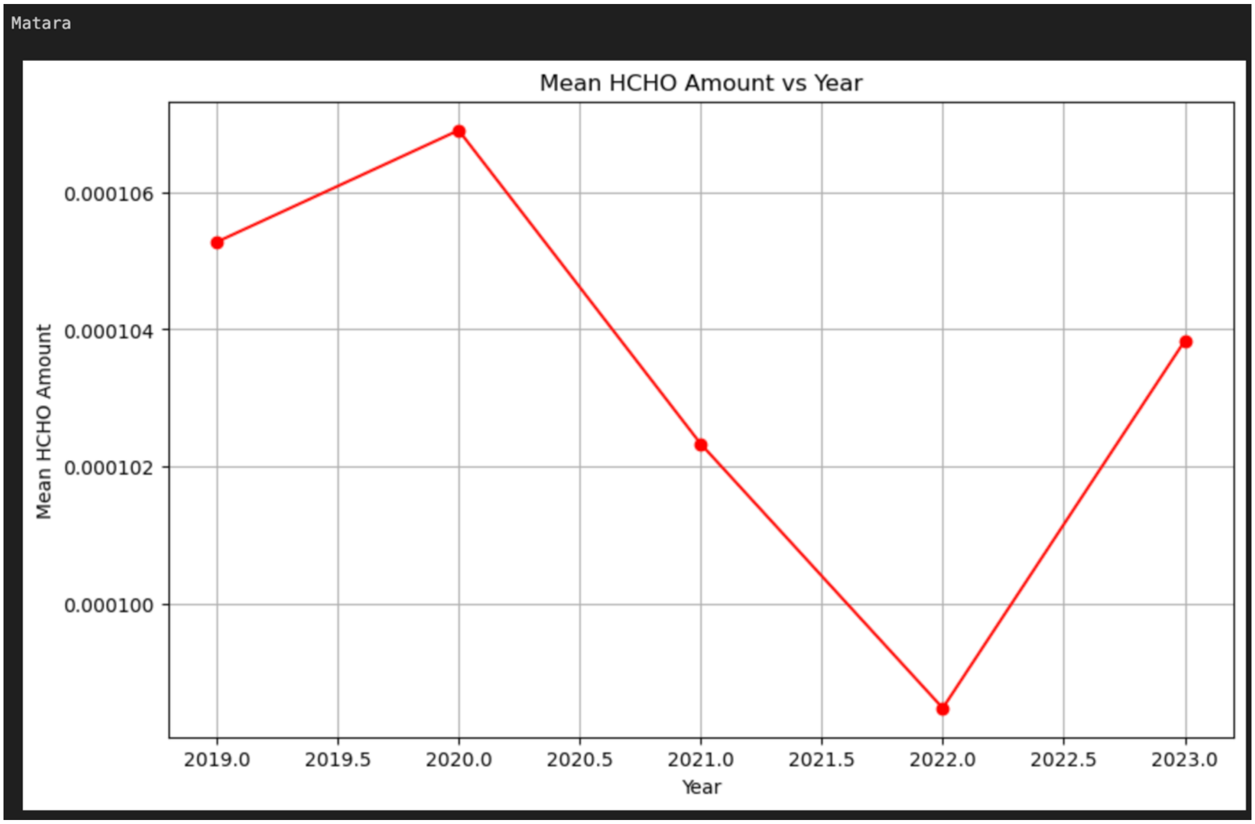
Nuwara Eliya –



```
Nuwara Eliya
```

The HCHO levels in Nuwara Eliya had a decreasing trend from 2019 itself but a sharp decrease is observed from 2021 to 2022. Covid lockdowns could be the main reason here as well. Just like in Colombo there is a sharp increase in HCHO levels from 2022 to 2023 as the covid lockdowns were lifted.

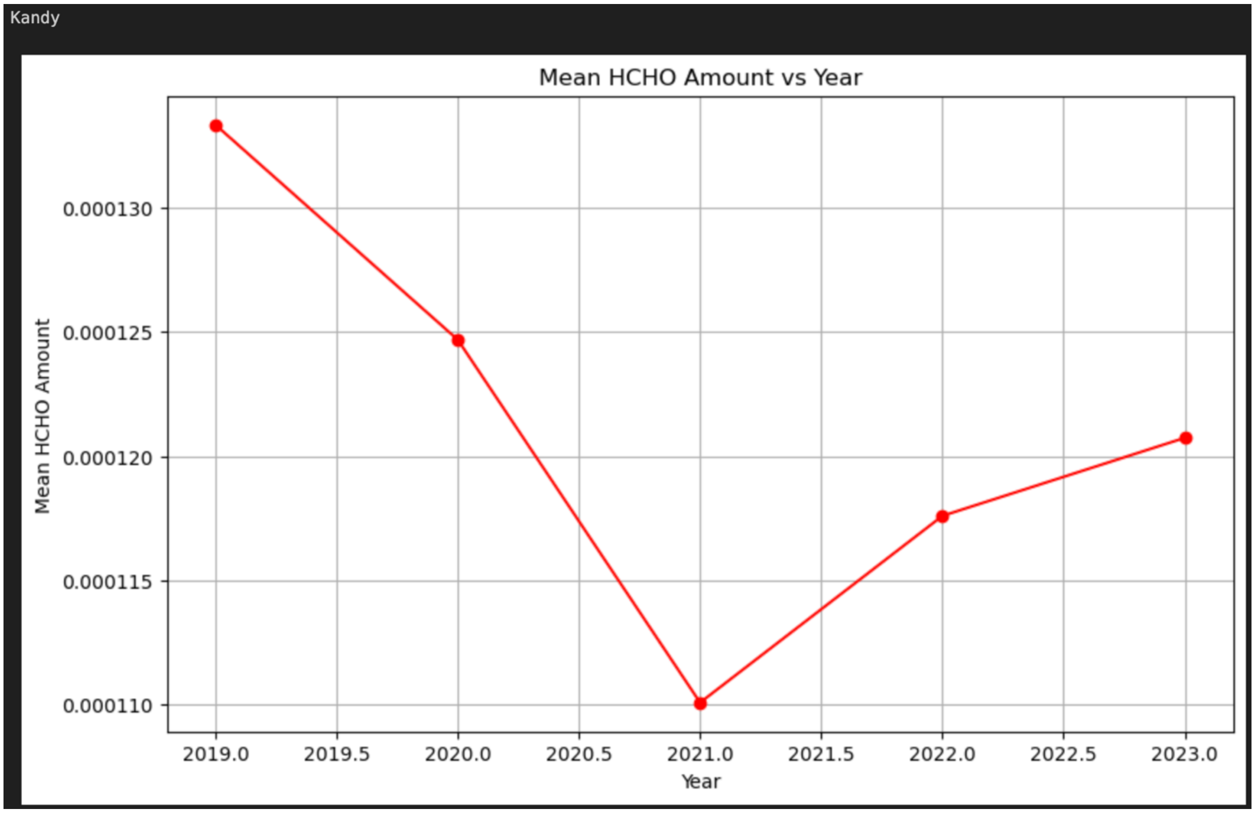
Matara –



```
Matara
```

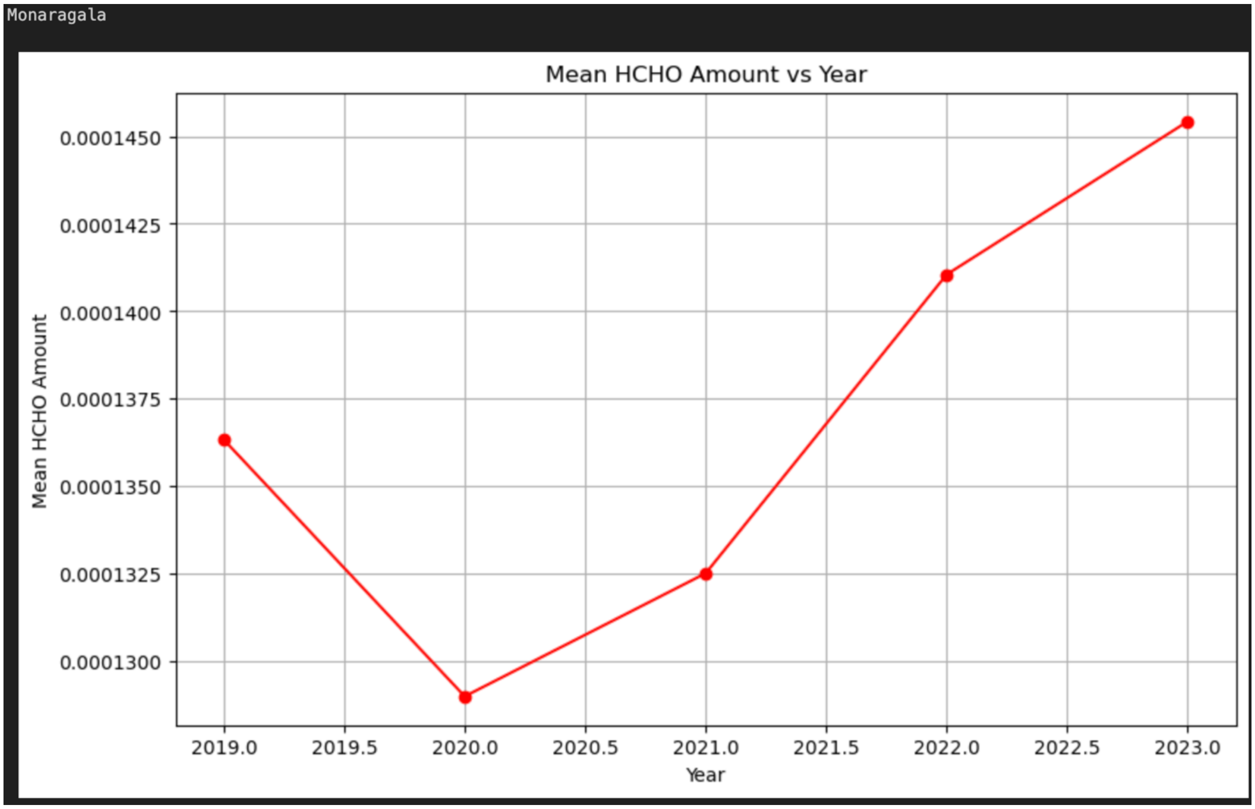
Matara follows the trend of Colombo having a sharp decrease in the HCHO levels during the covid lockdown period and a sharp increase after the lockdowns were lifted.

Kandy –



Kandy follows a similar trend to Nuwara Eliya, but a sharper decrease is seen when the covid lockdowns started and a sharp increase is seen after the covid lockdown was lifted.

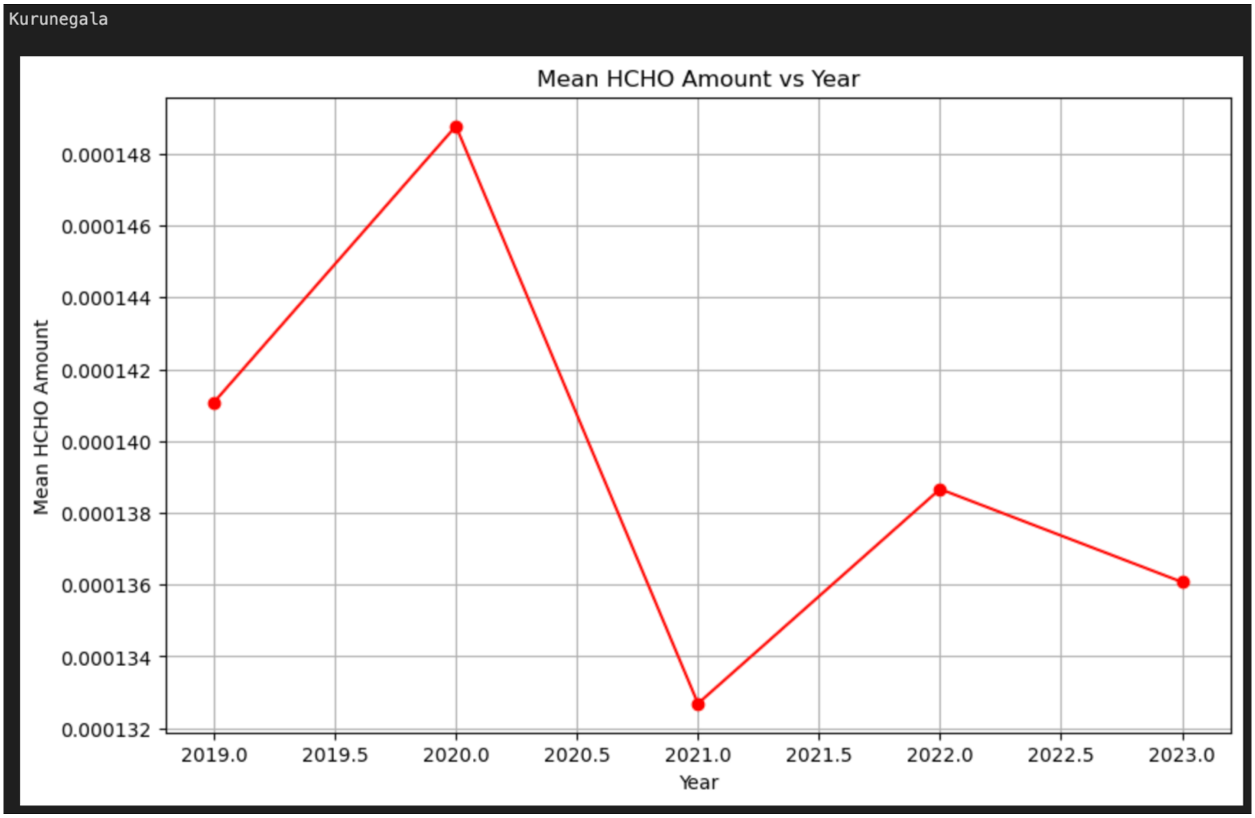
Monaragala –



```
Monaragala
```

Monaragala deviates from the other cities as the HCHO levels start increasing rapidly from 2020 itself, and it decreases from 2019 to 2020.

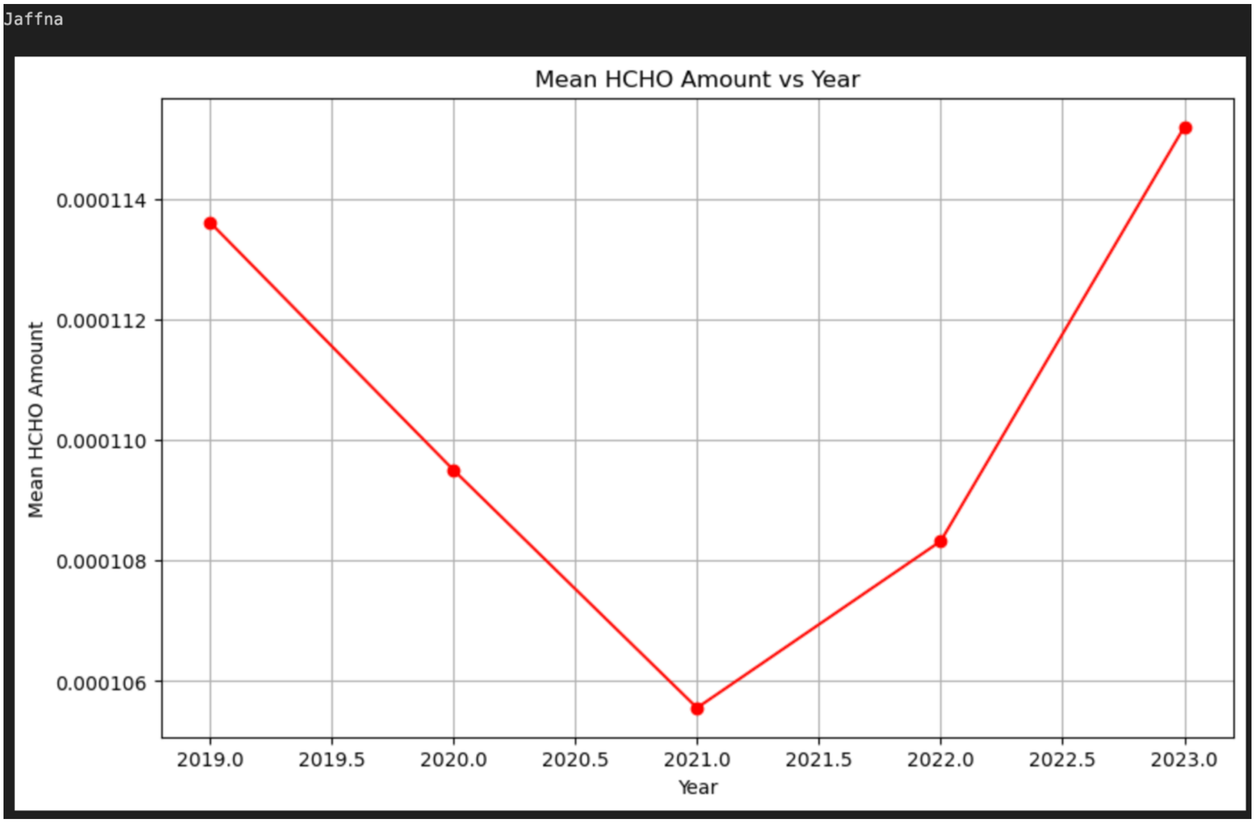
Kurunegala –



```
Kurunegala
```

Kurunegala follows a trend like Colombo. HCHO levels increase from 2019 to 2020 and there’s a sharp decrease from 2020 to 2021. But from 2021 to 2022 there’s an increase which was not seen in Colombo and again it starts to decrease from 2022 to 2023, which was not seen in the other cities.

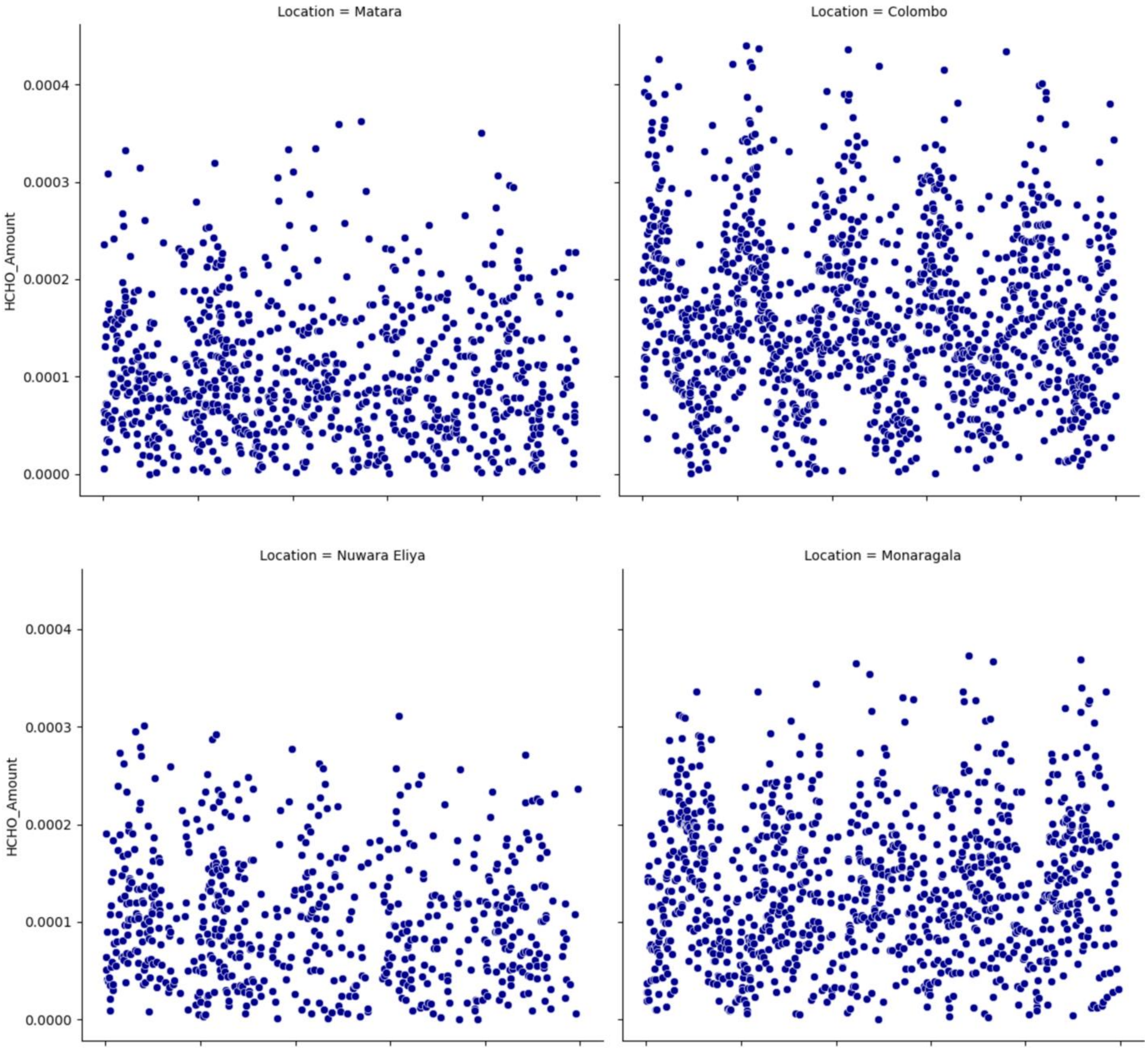
Jaffna –

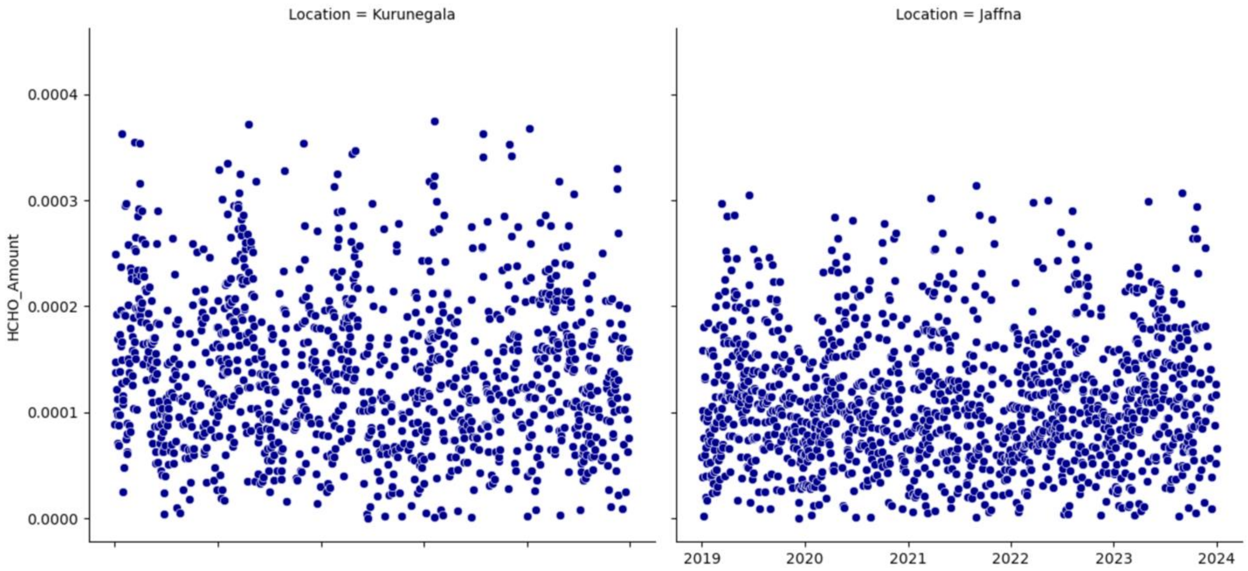


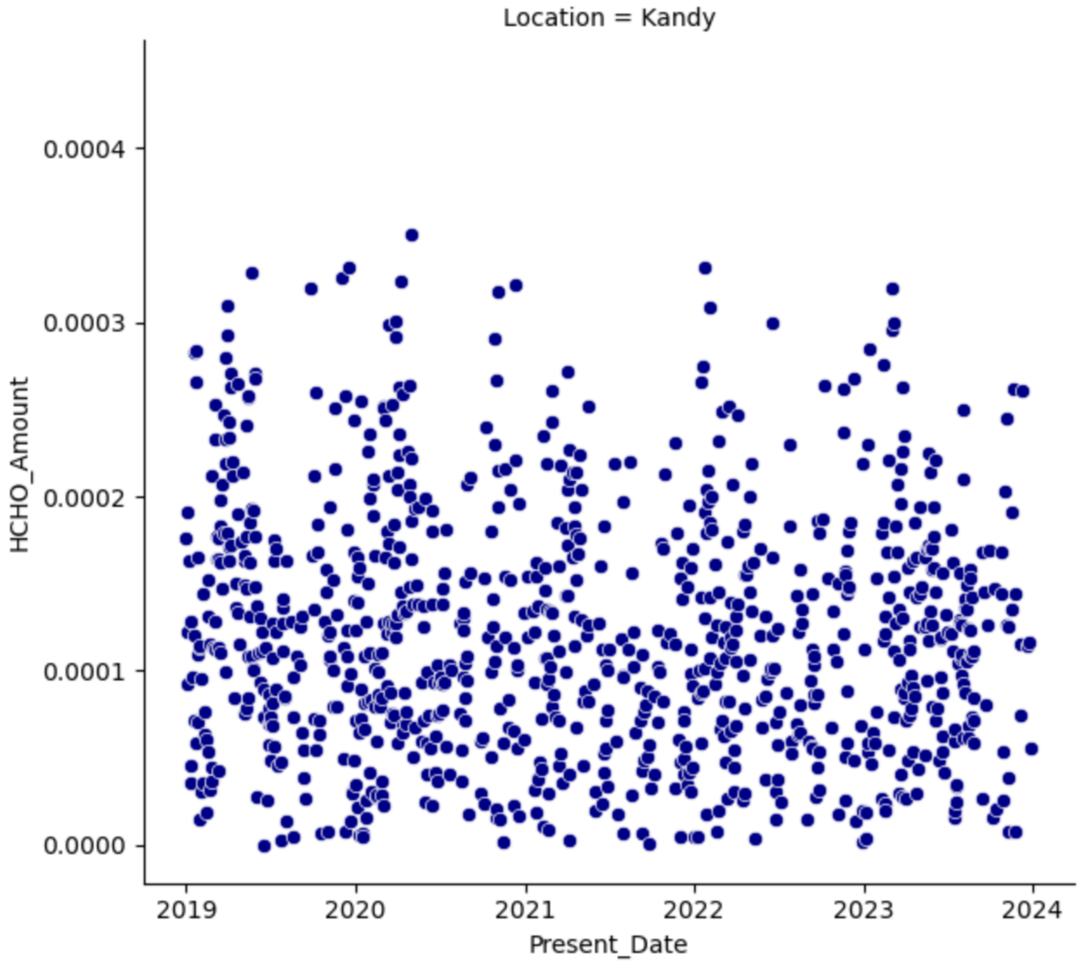
```
Jaffna
```

Jaffna has a different trend when compared to all the other cities. The HCHO levels decrease from 2019 and start increasing from 2021.

Scatter plots –



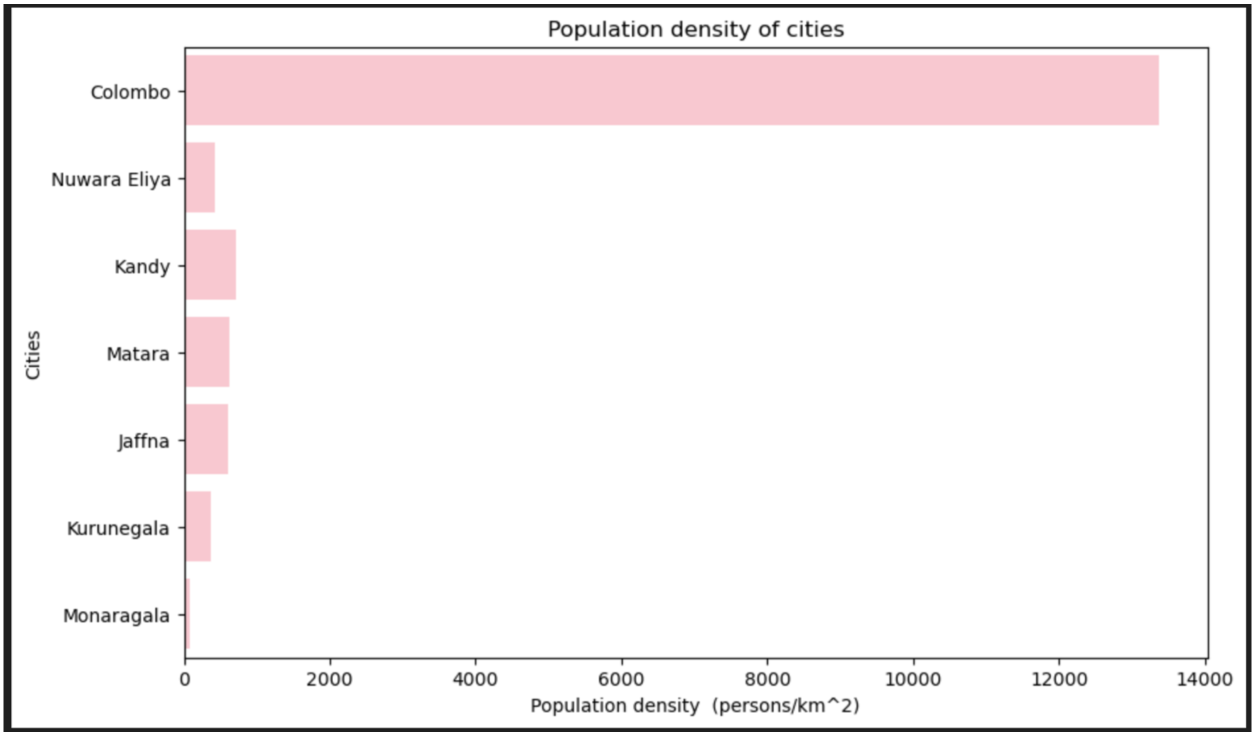

Location = Kurunegala
Location = Jaffna
HCHO_Amount
0.0004
0.0003
0.0002
0.0001
0.0000
2019
2020
2021
2022
2023
2024


Location = Kandy
HCHO_Amount
0.0004
0.0003
0.0002
0.0001
0.0000
2019
2020
2021
2022
2023
2024
Present_Date

From the above given scatter plots Colombo has the highest distribution when compared to the other cities and Jaffna has the most compact distribution which could make it easier to find patters in the Jaffna dataset.

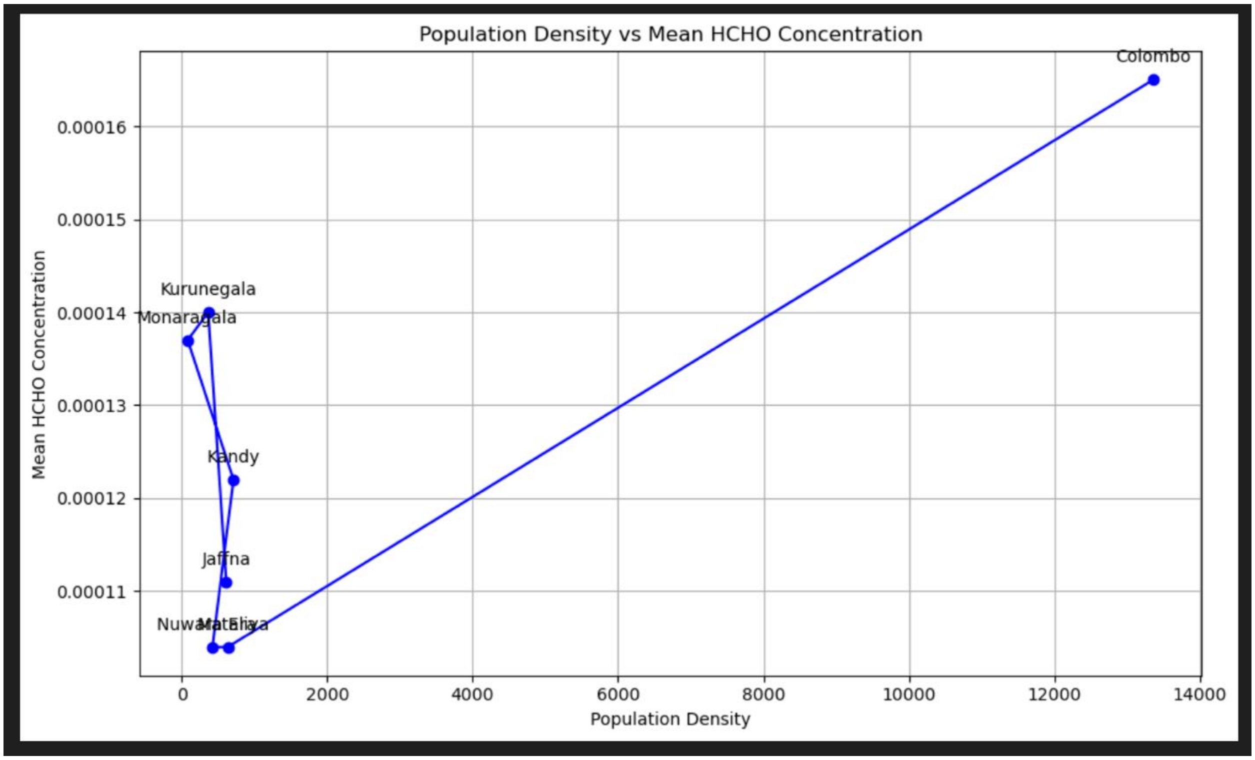
3. Spatial- temporal Analysis

Population Density –



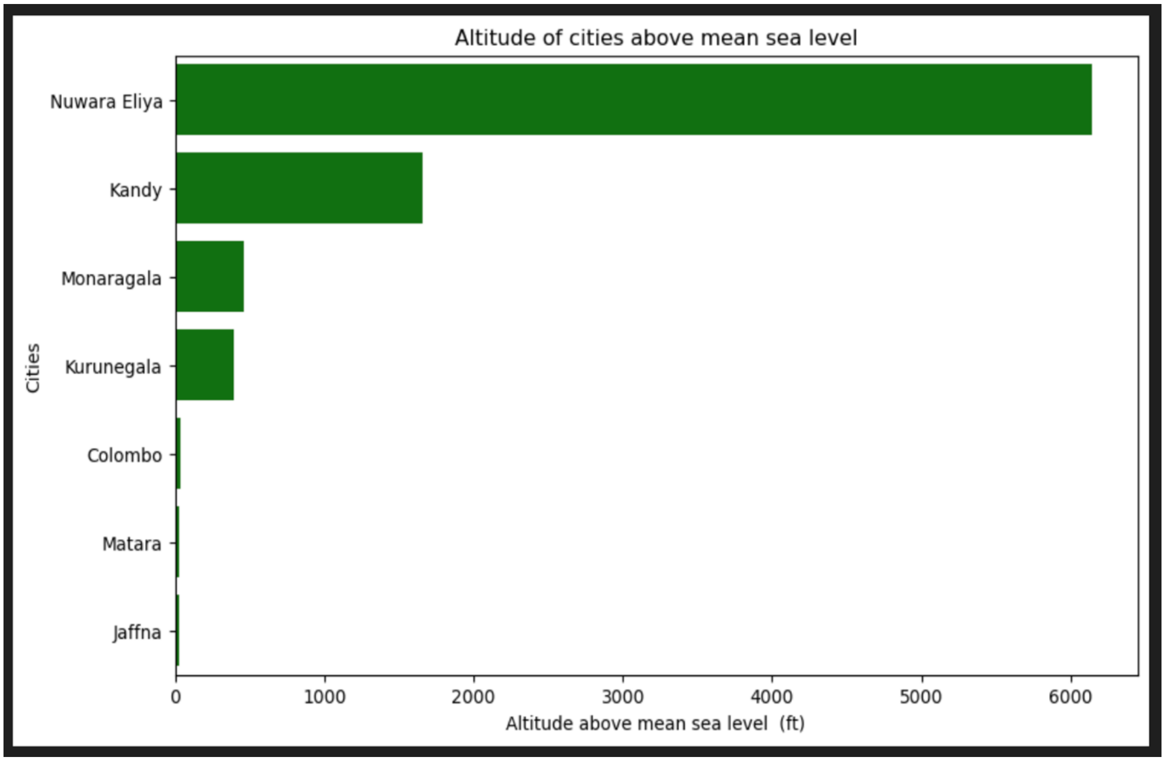
Population density is the measurement of population per unit land area. From the above given bar chart, Colombo has the highest population density and Monaragala has the lowest population density. A line a graph and the correlation coefficient could be used to check if there is a relationship between the HCHO levels in the atmosphere and the population density.

Line Graph (HCHO Levels vs Population Density) –



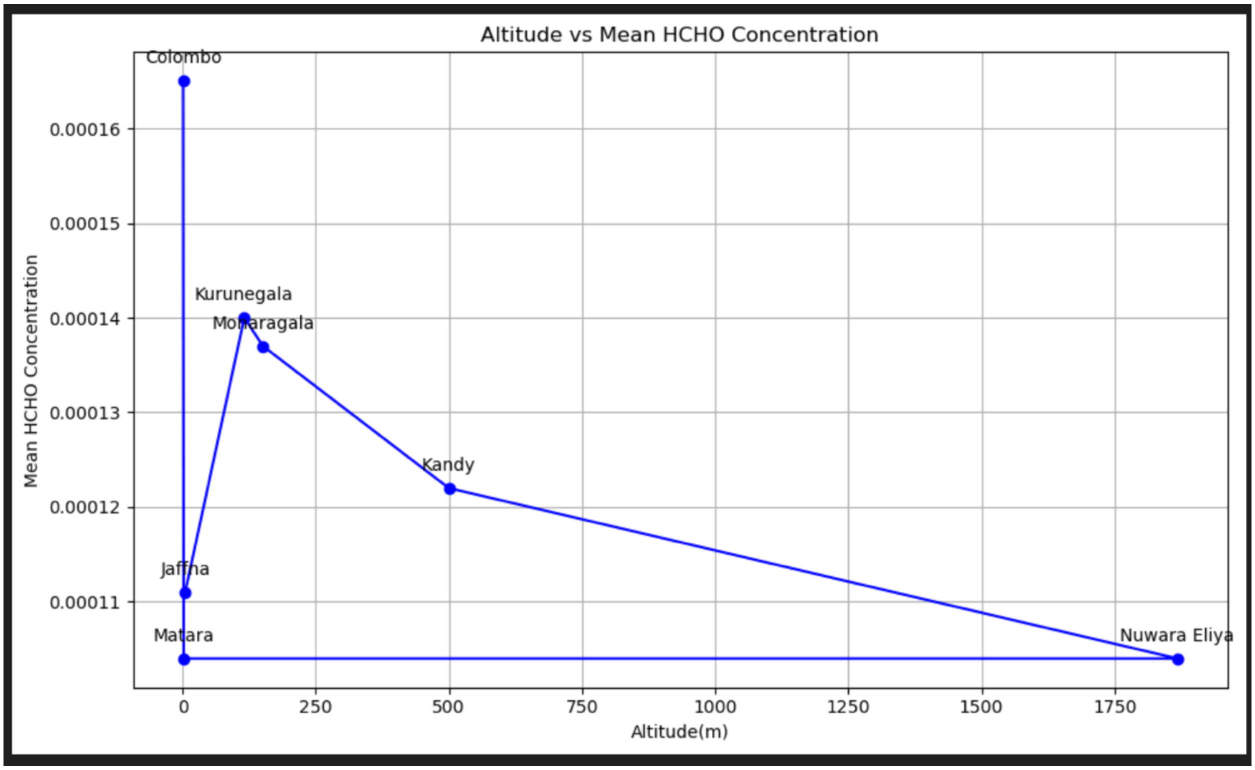
The line graph doesn't give a proper relationship between the mean HCHO levels and the population density. Due to this reason the correlation coefficient was found between the two. The value was positive 0.742. This indicates that there is a positive relationship between the mean HCHO levels and the population density, which means that when the population density increases the mean HCHO levels also increase. This could be due to many reasons. When the population density increases the economic activity of a region tends to increase, which leads to more vehicle usage and the release of more waste gases and pollutants to the environment. These factors may lead to the increase in the mean HCHO levels with the increase in population density. A better understanding can be obtained by observing the bar chart for the mean HCHO levels for four years by each city. Colombo has the highest level of HCHO and Monaragala and Nuwara Eliya has the lowest. This proves our finding as Colombo has the highest population density, Monaragala and Nuwara Eliya has the lowest population densities.

Altitude –



An analysis was done to check if the altitude of a city is a factor which effects the HCHO levels in the atmosphere. From the above that it's observed that Nuwara Eliya has the highest altitude and Colombo, Matara and Jaffna has the lowest altitudes. A line graph was drawn to check if a relationship between the two could be observed.

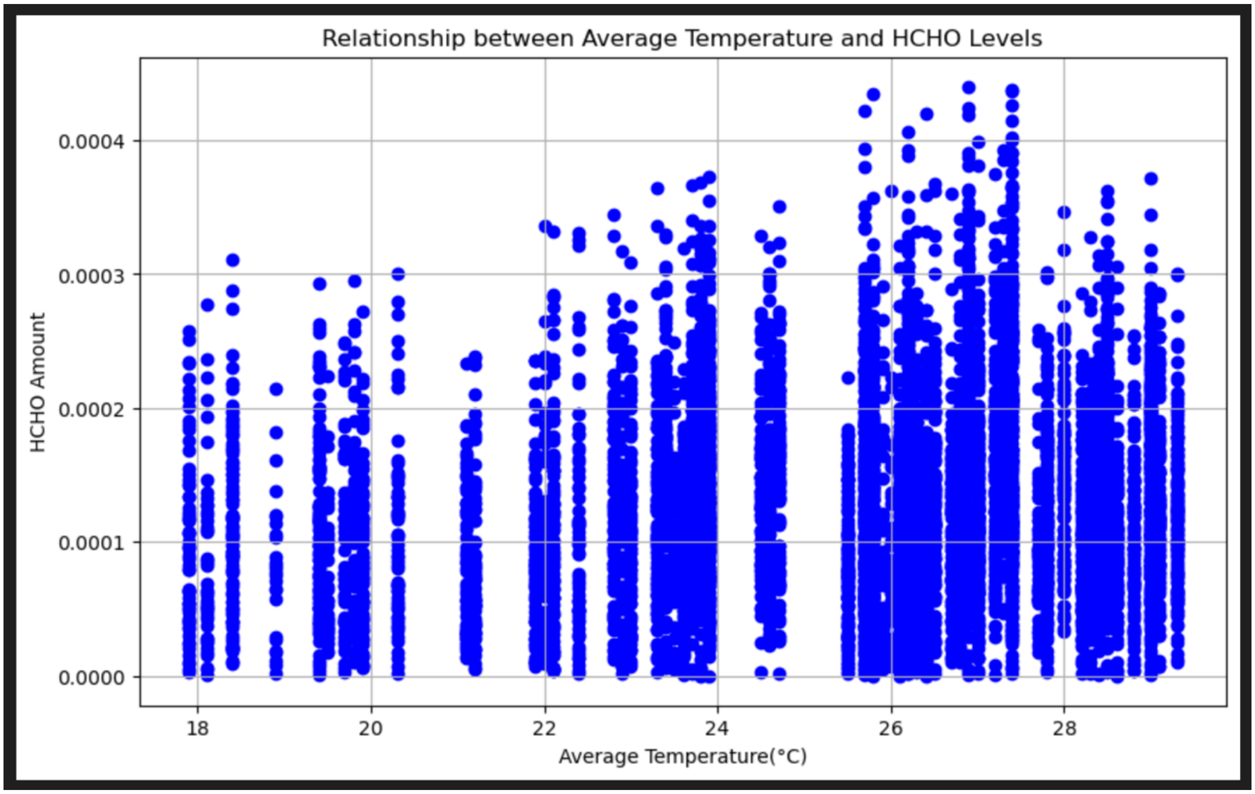
Line Graph (Altitude vs Mean HCHO Levels) –



A proper relationship between the two could not be identified from the line graph, therefore the correlation coefficient between the altitude and the mean HCHO levels were found to check for a relationship. The correlation coefficient between the two was -0.438. This shows that the altitude and the mean HCHO levels have an inverse relationship which means, when the altitude increases the mean HCHO levels should decrease.

Average Temperature –

A scatter plot graph was drawn between the average temperature and HCHO levels.

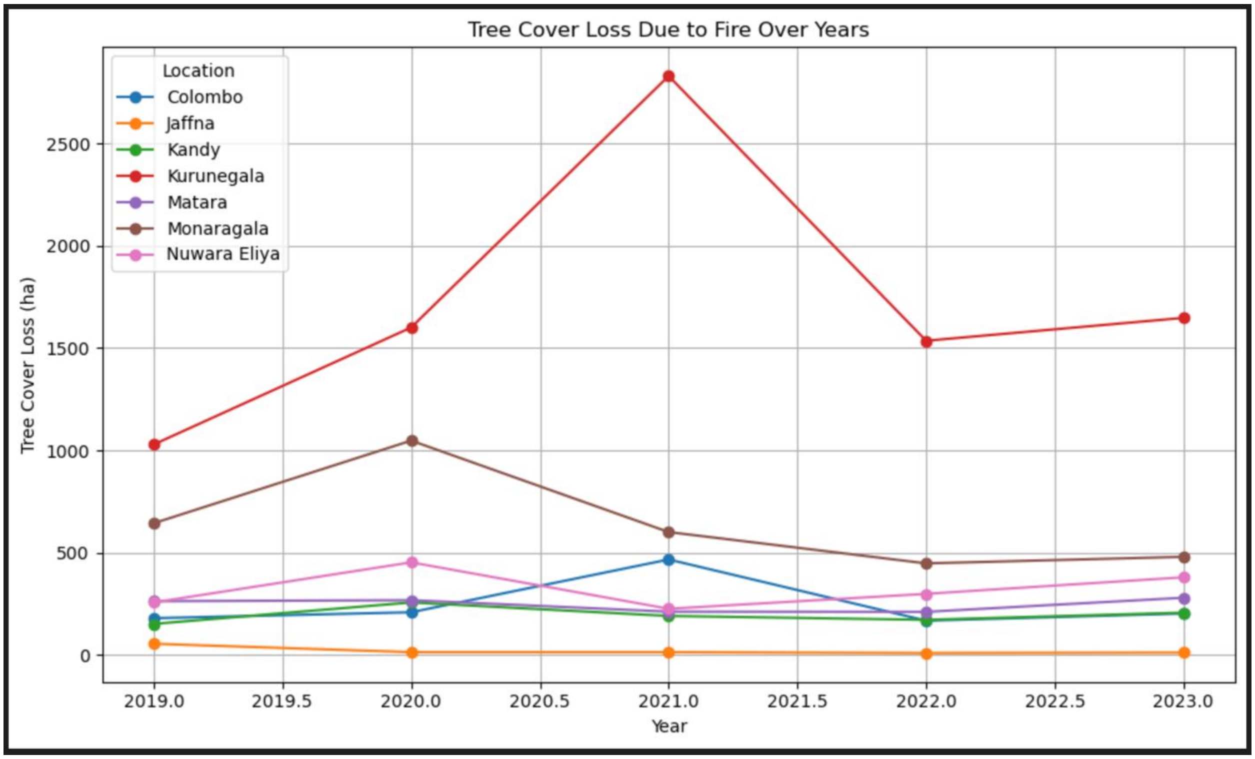


This scatter plot shows that the HCHO levels are high with the higher temperatures. The relationship between the average temperature and the mean HCHO level was also found using the correlation coefficient between the two. The correlation coefficient was 0.117. This indicates that higher temperatures will have higher HCHO levels. Therefore, cities which have a higher temperature will have higher HCHO levels in its atmosphere. The temperature alone won't be the only reason for the increase in HCHO levels, it's a combination of several factors.

Economic Activity (Unemployment Rate) –

Economic activity also has a relationship with the amount of HCHO levels in the atmosphere. The unemployment rate was considered to get a measurement of the economic activity. Lower unemployment rates mean that there's higher economic activity higher unemployment rates mean there is lesser economic activity in a region. There was a negative correlation coefficient of -0.12 between the unemployment rate and the HCHO levels, which means that when the unemployment rate decreases the HCHO levels increase and vice versa, just as discussed above.

Tree Cover Lost Due to Fire –



The loss of trees and fire could be major factors for the increase of HCHO levels in the atmosphere. To check if there is a relationship the correlation coefficient between the two was found. It was 0.069. This indicates that there is a weak positive relationship between the two. There is no strong evidence to suggest that there is a strong positive relationship between the two according to the above value.

4. Model Implementation and Comparison

The models were trained for each location. Three models were used, and the most efficient and robust model was used to get the future predictions of HCHO levels for each city. ARIMA, SARIMAX and Gaussian Process Regressor are the models tested for the above problem.

Model Type	Location	Order	RMSE
SARIMAX	Matara	(0,0,4)	0.0001250
ARIMA	Matara	(0,0,4)	7.11e-05
Gaussian Process	Matara	-	0.00014
SARIMAX	Nuwara Eliya	(5,0,0)	6.44e-05
ARIMA	Nuwara Eliya	(5,0,0)	6.107e-05
Gaussian Process	Nuwara Eliya	-	0.0063
SARIMAX	Kandy	(0,1,3)	0.00035
ARIMA	Kandy	(0,1,3)	0.000125
Gaussian Process	Kandy	-	0.00015
SARIMAX	Colombo	(0,1,4)	0.000257
ARIMA	Colombo	(0,1,4)	8.788e-05
Gaussian Process	Colombo	-	0.00043
SARIMAX	Monaragala	(3,0,3)	8.698e-05
ARIMA	Monaragala	(3,0,3)	9.189e-05
Gaussian Process	Monaragala	-	0.0001996
SARIMAX	Jaffna	(3,0,3)	7.7625e-05
ARIMA	Jaffna	(3,0,3)	6.23e-05
Gaussian Process	Jaffna	-	0.000193
SARIMAX	Kurunegala	(3,0,3)	0.00022
ARIMA	Kurunegala	(3,0,3)	6.633e-05
Gaussian Process	Kurunegala	-	0.00018

The SARIMAX model was used to generate future predictions as the RMSE value was more realistic in most scenarios as the RMSE value of the ARIMA model gave an indication that the model might be overfitting. Colombo being the exception as the prediction from the SARIMAX model for Colombo gave negative values, due to this the predictions from the ARIMA model was taken for Colombo. The generated predictions were stored in a csv file afterwards.

5. Communication and Insights

5.1 Existing Research

There were much researches in many countries, finding the HCHO levels in the atmosphere. Most of these were done by just monitoring the HCHO levels, there were very few researches done using machine learning or deep learning techniques. However, the paper "Using machine learning approach to reproduce the measured feature and understand the model- to- measurement discrepancy of atmospheric formaldehyde" used a machine learning technique for its research purposes. They had used a light grading boosting algorithm for their perditions.

5.2 Summary of findings

From this project several key factors were found, that effects the HCHO levels in the atmosphere.

Unemployment Rate –

There is weak negative relationship between the unemployment rate and the HCHO levels, but this indicates that when the unemployment rate increases the HCHO levels tend to decrease. This indicates that when the economic activity increase there is an increase in the HCHO levels in the atmosphere.

Covid Lockdowns –

During the covid lockdown period almost all the cities showed a sharp reduction in the HCHO levels in the atmosphere. This could be due to the reduction in human activity as less vehicles were used, and many industrial plants were not used and due to this, harmful gases such as HCHO weren't released into the atmosphere.

Average Temperature –

There is a weak positive relationship among the average temperature and the HCHO levels. When the temperatures increase there is a chance that the HCHO levels in the atmosphere will increase.

Tree cover lost by fire –

A very weak relationship between this and the HCHO levels exist. As trees help the reduction of harmful gases from the atmosphere.

Altitude –

There seems to be a mid strong negative relationship with the altitude and the HCHO levels in the atmosphere. When the altitude increases the HCHO levels tend to decrease. Cities such as Nuwara Eliya tend to have lower HCHO levels in the atmosphere, whereas areas with a low altitude have higher HCHO levels (EX – Colombo).

Population Density –

There is a significantly strong positive relationship with the population density and the atmospheric HCHO levels. This could be due to many reasons as more people means more pollution. Humans tend to release toxic gases such as HCHO to the atmosphere through factories etc. This could be the reason why cities like Colombo tend to have higher HCHO levels in the atmosphere.

5.3 Limitations

The dataset had a lot of issues. There were many null values and a significant number of outliers which had to be dealt with. Robust preprocessing techniques had to be used to get the data to a usable standard. The accuracy of the data could have been improved if more data from previous years and more cities were obtained.

5.4 Policy making & research

This type of analysis can help the government come up with policies which can help reduce and maintain the HCHO levels in the atmosphere as this gas is a harmful gas to humans. These policies could set up standards that industries need to follow to prevent such release of harmful gases to the atmosphere.

6. References

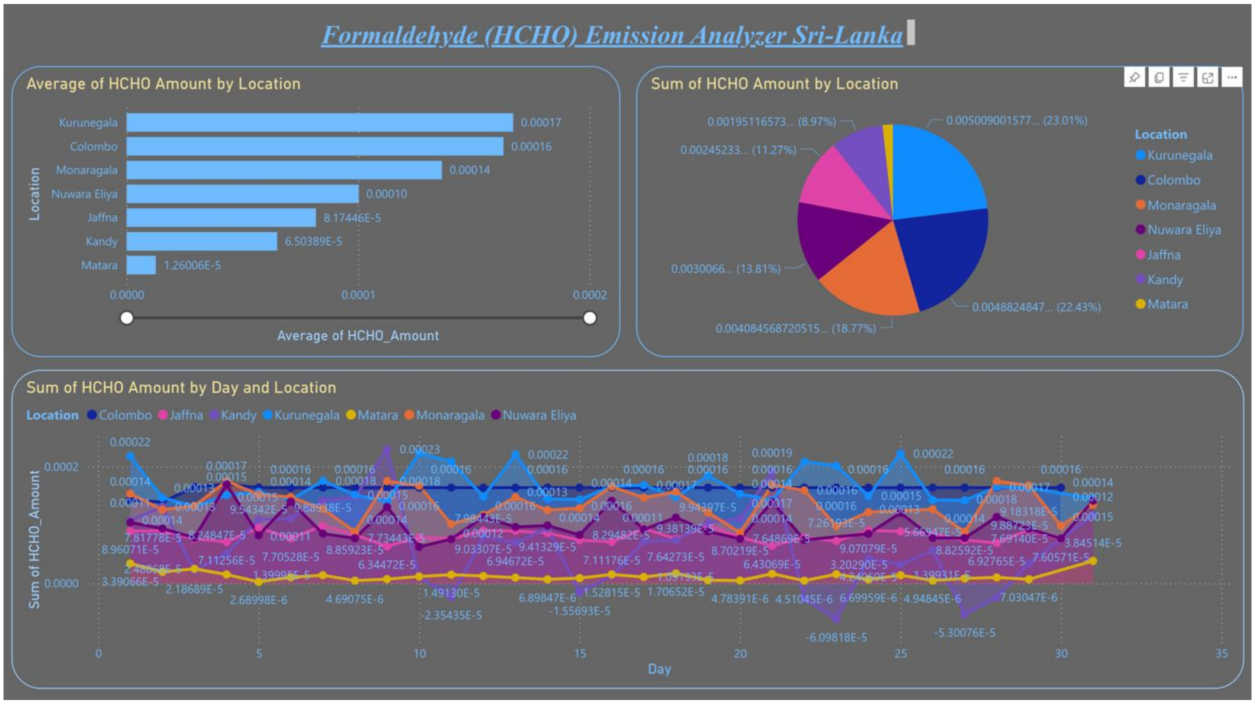
YIN, H. et al., 2022. Using machine learning approach to reproduce the measured feature and understand the model-to-measurement discrepancy of atmospheric formaldehyde. *The Science of the Total Environment*, 851(158271), p. 158271. Available from: http://dx.doi.org/10.1016/j.scitotenv.2022.158271.

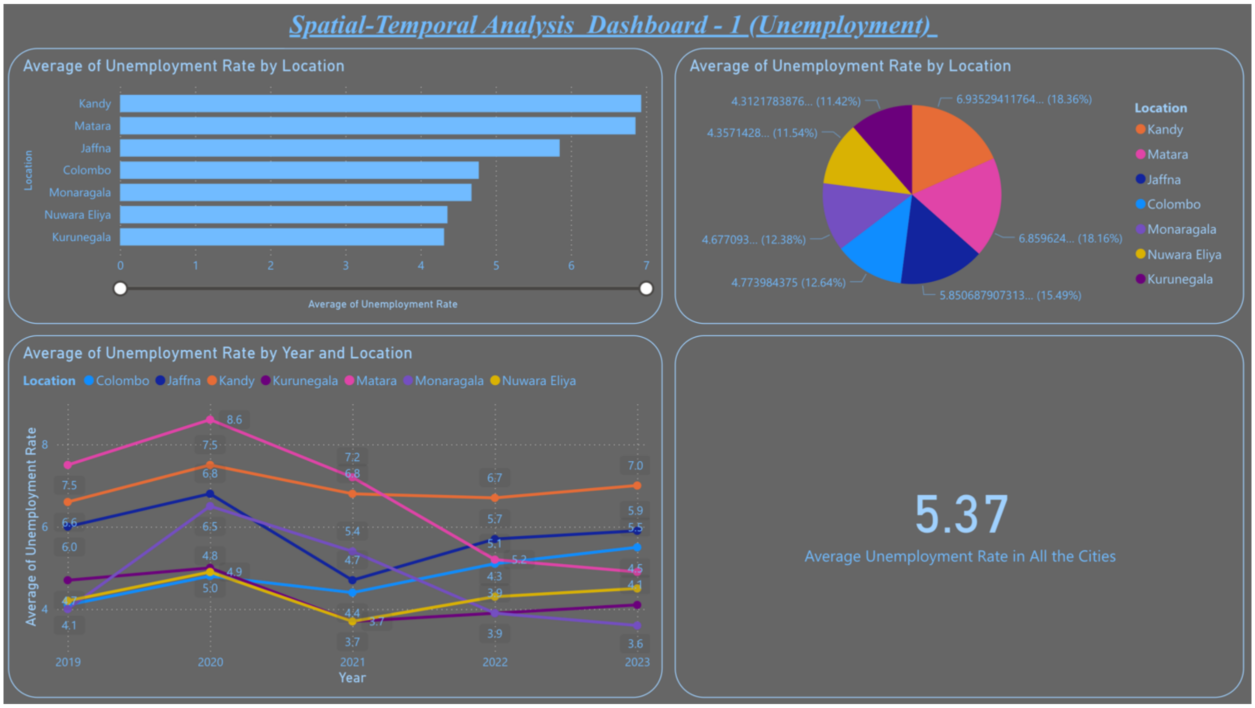
FOLLOW, S., 2020. *Python*. [online]. GeeksforGeeks. Available from: https://www.geeksforgeeks.org/python-arima-model-for-time-series-forecasting/ [Accessed 19 Apr 2024].

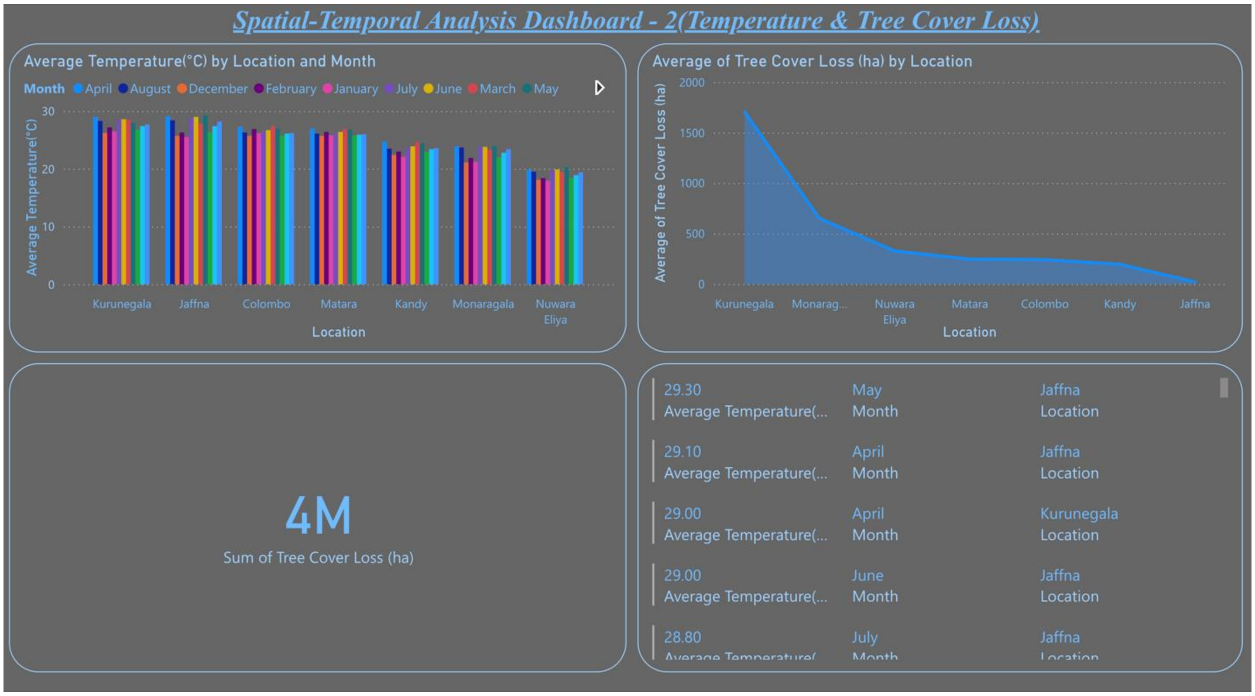
ML, 2017. [online]. GeeksforGeeks. Available from: https://www.geeksforgeeks.org/data-preprocessing-machine-learning-python/ [Accessed 19 Apr 2024].

2024. [online]. Datacamp.com. Available from: https://www.datacamp.com/tutorial/power-bi-dashboard-tutorial [Accessed 19 Apr 2024].

7. Appendix


Formaldehyde (HCHO) Emission Analyzer Sri-Lanka
Average of HCHO Amount by Location
Location
Kurunegala 0.00017
Colombo 0.00016
Monaragala 0.00014
Nuwara Eliya 0.00010
Jaffna 8.17446E-5
Kandy 6.50389E-5
Matara 1.26006E-5
0.0000
0.0001
0.0002
Average of HCHO_Amount
Sum of HCHO Amount by Location
0.00195116573... (8.97%)
0.005009001577... (23.01%)
0.00245233... (11.27%)
0.0030066... (13.81%)
0.0048824847... (22.43%)
0.004084568720515... (18.77%)
Location
Kurunegala
Colombo
Monaragala
Nuwara Eliya
Jaffna
Kandy
Matara
Sum of HCHO Amount by Day and Location
Location Colombo Jaffna Kandy Kurunegala Matara Monaragala Nuwara Eliya
Sum of HCHO_Amount
0.0002
0.0000
Day
0
5
10
15
20
25
30
35


Spatial-Temporal Analysis Dashboard - 1 (Unemployment)
Average of Unemployment Rate by Location
Location
Kandy
Matara
Jaffna
Colombo
Monaragala
Nuwara Eliya
Kurunegala
0
1
2
3
4
5
6
7
Average of Unemployment Rate
Average of Unemployment Rate by Location
4.3121783876... (11.42%)
6.93529411764... (18.36%)
4.3571428... (11.54%)
4.677093... (12.38%)
6.859624... (18.16%)
4.773984375 (12.64%)
5.850687907313... (15.49%)
Location
Kandy
Matara
Jaffna
Colombo
Monaragala
Nuwara Eliya
Kurunegala
Average of Unemployment Rate by Year and Location
Location Colombo Jaffna Kandy Kurunegala Matara Monaragala Nuwara Eliya
Average of Unemployment Rate
8
6
4
2019
2020
2021
2022
2023
Year
5.37
Average Unemployment Rate in All the Cities

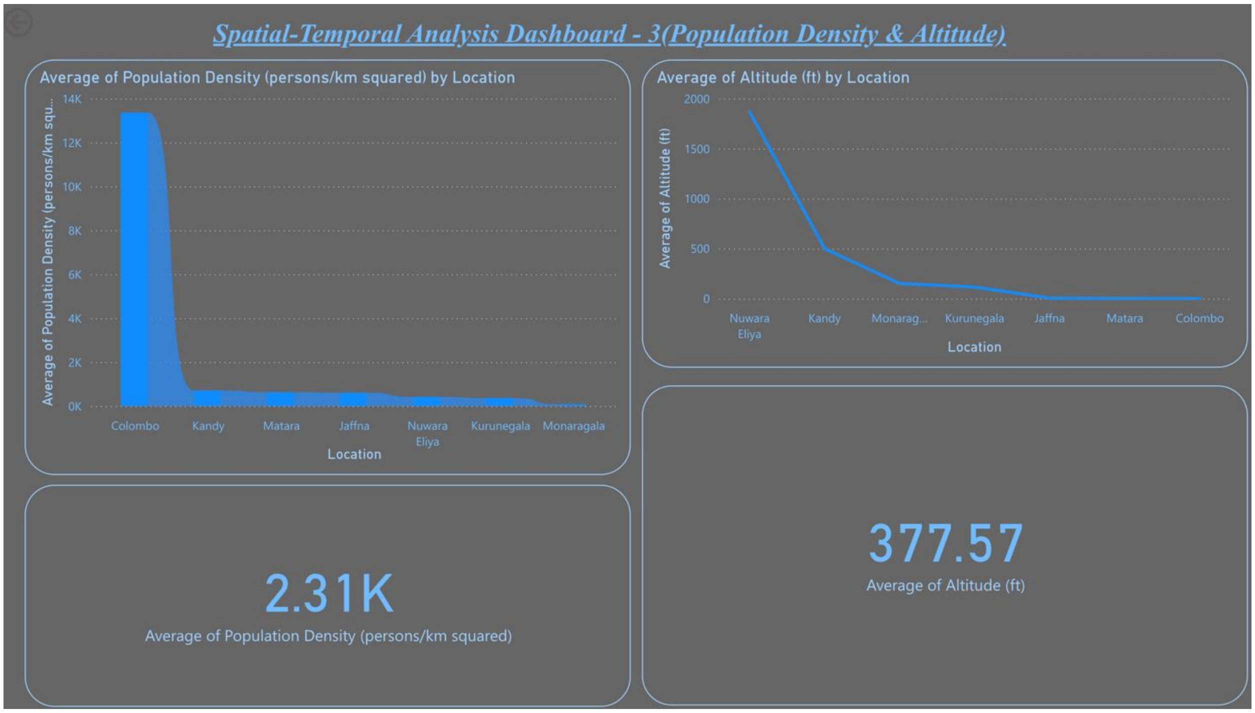


Spatial-Temporal Analysis Dashboard - 2(Temperature & Tree Cover Loss)

Average Temperature(°C) by Location and Month
Month ● April ● August ● December ● February ● January ● July ● June ● March ● May
Average Temperature(°C)
30
20
10
0
Kurunegala
Jaffna
Colombo
Matara
Kandy
Monaragala
Nuwara Eliya
Location
Average of Tree Cover Loss (ha) by Location
Average of Tree Cover Loss (ha)
2000
1500
1000
500
0
Kurunegala
Monarag...
Nuwara Eliya
Matara
Colombo
Kandy
Jaffna
Location

4M
Sum of Tree Cover Loss (ha)

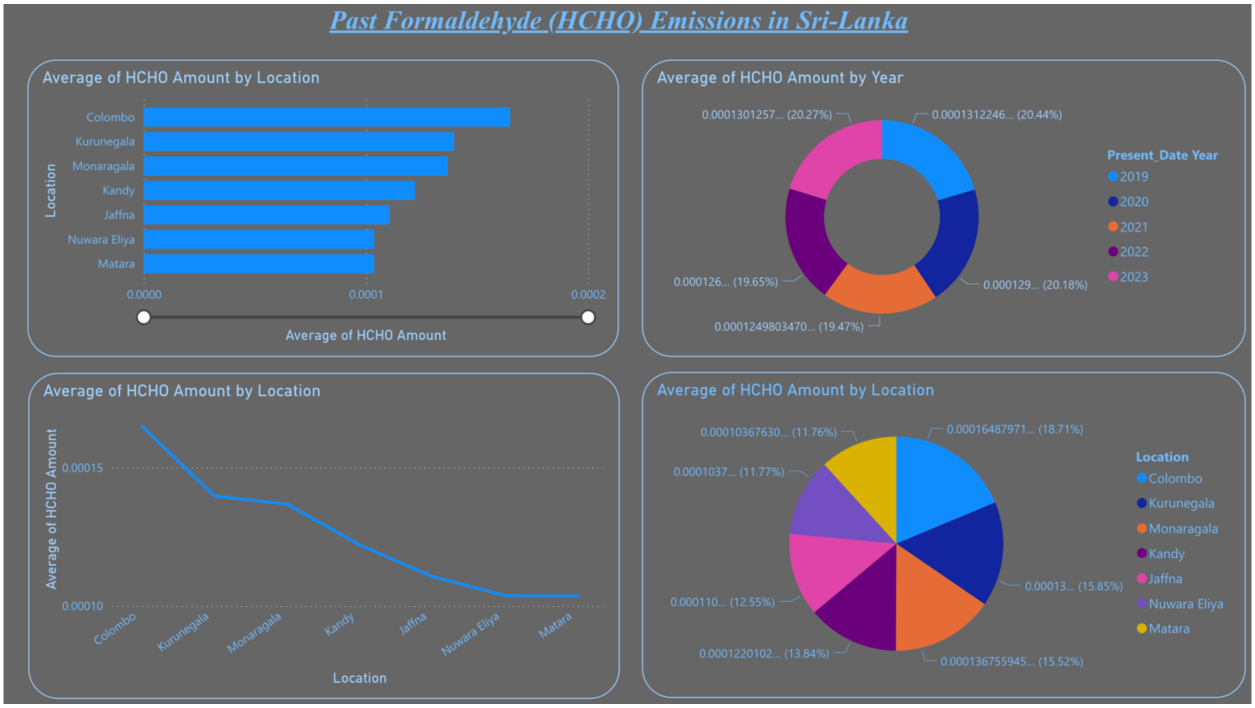
29.30 Average Temperature(...	May Month	Jaffna Location
29.10 Average Temperature(...	April Month	Jaffna Location
29.00 Average Temperature(...	April Month	Kurunegala Location
29.00 Average Temperature(...	June Month	Jaffna Location
28.80 Average Temperature(...	July Month	Jaffna Location



Spatial-Temporal Analysis Dashboard - 3(Population Density & Altitude)

2.31K
Average of Population Density (persons/km squared)

377.57
Average of Altitude (ft)



Past Formaldehyde (HCHO) Emissions in Sri-Lanka