ROBERT GORDON UNIVERSITY ABERDEEN

INFORMATICS INSTITUTE OF TECHNOLOGY

Report

Module Name: **Data Engineering**

Module Code: **CM2606**

Module Leader: **Mr. Mohamed Ayoob**

Student Name: **Aadhavan Arkhash Saravanakumar**

IIT ID: **20221213**

RGU ID: **2237045**

Table of Contents

1. Introduction

This project involves the analysis of Tropospheric Formaldehyde (HCHO) gas in the atmospheric regions across different cities in Sri Lanka. Formaldehyde is a primary component in the formation of ozone, which in turn acts as a key constituent of photochemical smog. Formaldehyde emission is a result of certain natural and human influenced activities. Natural causes of this emission include changes of seasons, weather conditions such as atmospheric temperature and health scenarios such as epidemics and pandemics. Human influenced activities include combustion of vegetation and other substances and traffic conditions. This report analyses the effect of each of these factors and their contribution to emission of Tropospheric Formaldehyde.

Data regarding the HCHO readings across six major cities in Sri Lanka have been obtained via the TROPOMI instrument in the Sentinel- 5P satellite. This satellite is used to provide information and collective insights regarding air quality. The cities considered in this project are Colombo, Matara, Nuwara Eliya, Kandy, Jaffna, Kurunegala, Monaragala. For each of these cities, the data is preprocessed, cleaned, and visualized to see any possible trends between the cities. Following this, spatial and temporal analysis is conducted to determine how external factors influence the emission of Formaldehyde. Time series algorithms are implemented to see how the HCHO readings will be in the future, and this is presented as a dashboard using the Microsoft Power BI application.

2. Data Preprocessing

2.1. Importing of Libraries and synthesis of functions

To perform the preprocessing steps, few libraries in Python have been imported and been made use of.

1. Pandas- for handling of data frames
2. Numpy- for numerical calculations
3. Matplotlib- for visualizations on data
4. Seaborn- for advanced statistical visualizations on data

Next, a few functions have been made to perform the same functionalities for the dataframes involving each city.

1. Conversion of negative HCHO readings into null values so that they can be removed.

```python
# function to convert negative values to null values
def negative_val_convertor(df, column):
    df[column] = df[column].apply(lambda x: np.nan if x<0 else x)
    return df
```

2. Creating a function to draw a normal distribution plot to remove outliers.

```python
# defining a function to draw the normal distribution plot
def normal_distribution_plot(df, column, color='skyblue'):
    plt.figure(figsize=(8, 6))
    sns.histplot(df[column], kde=True, stat='density', color=color)
    sns.kdeplot(df[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()
```

3. A function is created to remove outliers by removing all the values which are greater than the threshold which is calculated using the equation below.

$$\text{Threshold} = \text{Mean} + (3 \text{ x Standard Deviation})$$

```python
# defining a function to remove outliers from the data frame
def remove_outliers(df, column):
    # calculating the mean and the standard deviation of the data frame
    mean = df[column].mean()
    std = df[column].std()

    # setting the threshold value for the removal of outliers
    threshold = mean + (3 * std)

    # removing the outliers from the data frame
    df = df[df[column] <= threshold]
    return df
```

4. Imputation has been used to fill in the missing values with the overall mean of the existing values for the respective function. The values could not be dropped as a large amount of data would be lost which would make the predictions of the time series algorithm less efficient. These null values include negative values which were converted into null values and those which already existed in the original dataset.

```python
# method to fill the null values with the mean
def imputation(df):
    mean_val = df['HCHO_Reading'].mean()
    df['HCHO_Reading'].fillna(mean_val, inplace=True)
    return df
```

5. A function has been created to calculate all the statistical measures for each location's HCHO readings which include the mean, median, mode and the standard deviation. These values are returned at the end of the function.

```python
# method to print the statistical analysis of the HCHO readings
def statistical_analysis(df, column):
    df = globals()[df]

    # calculating the statistical measures
    mean_val = df[column].mean()
    median_val = df[column].median()
    mode_val = df[column].mode()[0]
    std_dev_val = df[column].std()

    return mean_val, median_val, mode_val, std_dev_val
```

6. A function is created to remove the next date column from the data frame as it does not prove any use during the engineering of trends and development of the model.

```python
# dropping off the Next_Date column
def drop_next_date(dataframe):
    return dataframe.drop(columns=['Next_Date'], inplace=True)
```

2.2. Cleaning and Removal of Outliers

Following this, the names of the columns are entered into an array based on the column names provided in the specification.

Next, each CSV file is considered, and the following processes are done. First, the name of the location is changed so that it is only a single word. For example, in the original dataset, one location was “Colombo Proper”. This has been changed to “Colombo” to make it easier to deal with. Following this, the dimension of the dataset is obtained by using “.shape”. Next, the data types of each column are printed using “. dtypes”. Upon doing this, it could be noted that the “Current_date” and “Next_date” columns have the object data type. This is then converted to date time data type to make it more efficient for visualizations to be made based on variation of date.

Once this has been done, then the dataset is split into different data frames based on the location. For example, one CSV file contains data for the cities of Colombo, Matara and Nuwara Eliya. This CSV file has been separated based on the location and has been split into three separate data frames.

A box plot is drawn for each location to check for the existence of outliers within the dataframe. Upon drawing the box plot, it was evident that not enough information could be obtained regarding the outliers by using just the box plot. Negative values were also existent in this dataset. There are two options for the removal of these negative values.

1. Drop these rows.
2. Convert into null values and perform imputation.

The second method was performed to prevent data being lost unnecessarily. Following this, a normal distribution plot was drawn to obtain more insights into the outliers, allowing us to remove them using the threshold function defined above. This was performed by calling the respective function.

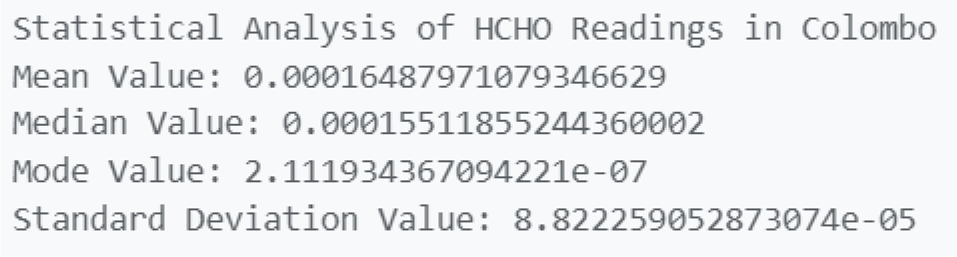
Once this was done, the null values in the data frame were imputed using the existent mean value of the HCHO reading. The number of null values in each column was printed to see if the imputation was performed correctly.

Following this, the statistical measures are printed to the console by calling the function respectively. Finally, the cleaned data frame is saved as a CSV file. This process is performed for the three datasets and at the end of the preprocessing, there are seven new CSV files: one for each location.

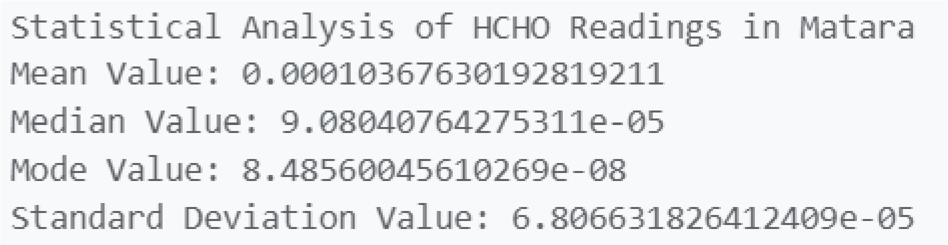
2.3. Exploration of descriptive statistics

By analyzing the statistical measures such as the mean, median, mode and the standard deviation, we could identify the variation of HCHO readings at the said location to a certain extent. The statistical measures are calculated once the data frame has been completely cleaned and imputed to fill in null values. This is done using the function "statistical_analysis" defined above. This function takes the name of the data frame and the column to be considered as arguments.

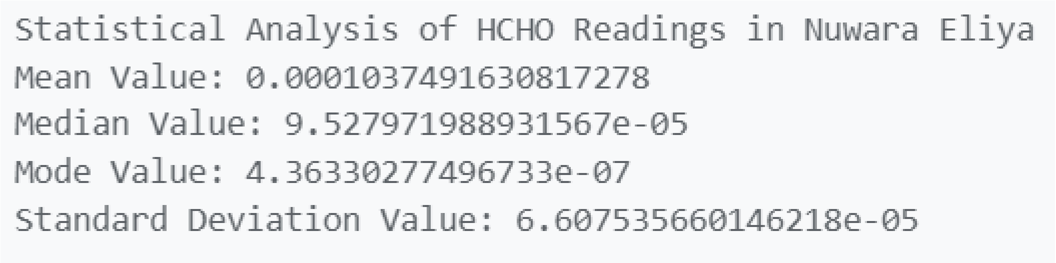
Below are the statistical results obtained.



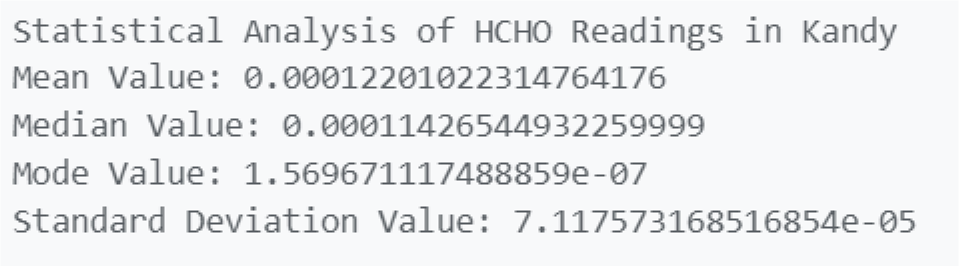
```
Statistical Analysis of HCHO Readings in Colombo
Mean Value: 0.00016487971079346629
Median Value: 0.00015511855244360002
Mode Value: 2.111934367094221e-07
Standard Deviation Value: 8.822259052873074e-05
```



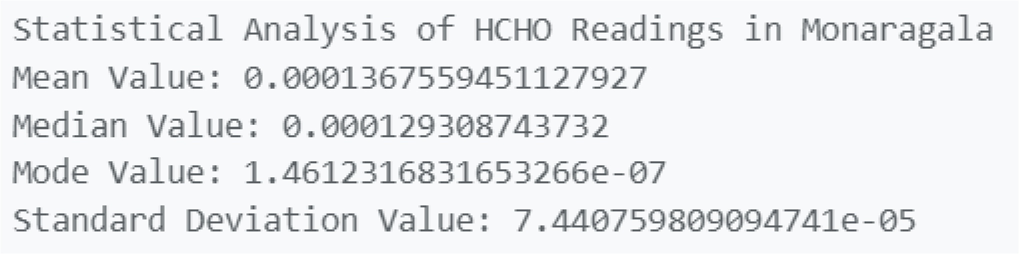
```
Statistical Analysis of HCHO Readings in Matara
Mean Value: 0.00010367630192819211
Median Value: 9.08040764275311e-05
Mode Value: 8.48560045610269e-08
Standard Deviation Value: 6.806631826412409e-05
```



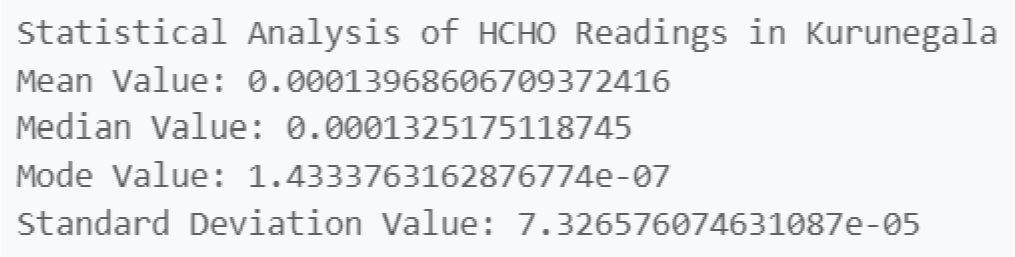
```
Statistical Analysis of HCHO Readings in Nuwara Eliya
Mean Value: 0.0001037491630817278
Median Value: 9.527971988931567e-05
Mode Value: 4.36330277496733e-07
Standard Deviation Value: 6.607535660146218e-05
```



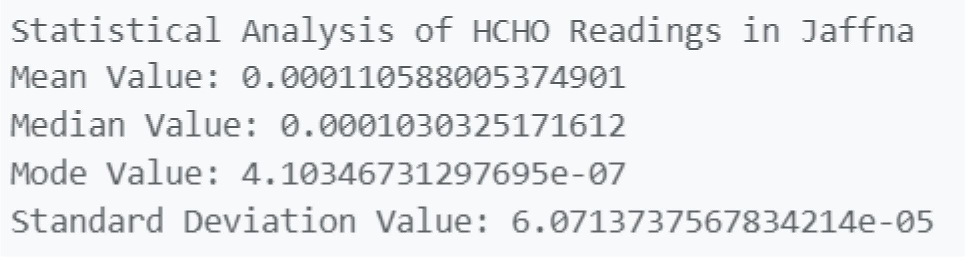
```
Statistical Analysis of HCHO Readings in Kandy
Mean Value: 0.00012201022314764176
Median Value: 0.00011426544932259999
Mode Value: 1.569671117488859e-07
Standard Deviation Value: 7.117573168516854e-05
```



```
Statistical Analysis of HCHO Readings in Monaragala
Mean Value: 0.0001367559451127927
Median Value: 0.000129308743732
Mode Value: 1.4612316831653266e-07
Standard Deviation Value: 7.440759809094741e-05
```

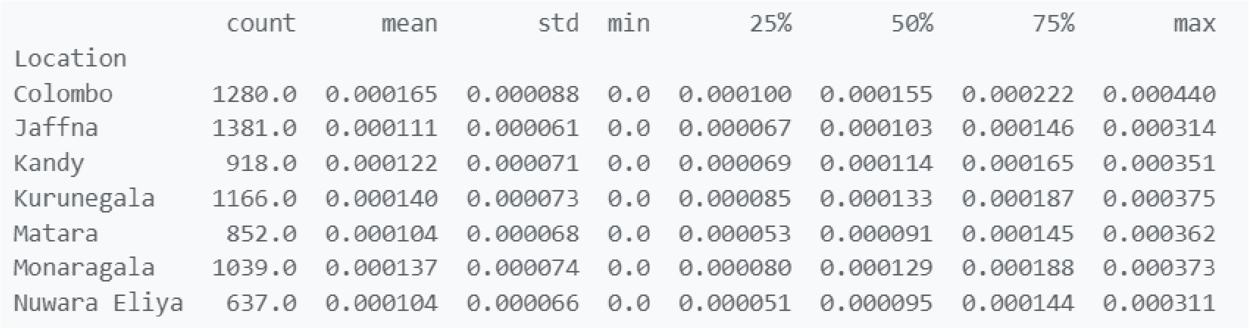


```
Statistical Analysis of HCHO Readings in Kurunegala
Mean Value: 0.00013968606709372416
Median Value: 0.0001325175118745
Mode Value: 1.4333763162876774e-07
Standard Deviation Value: 7.326576074631087e-05
```



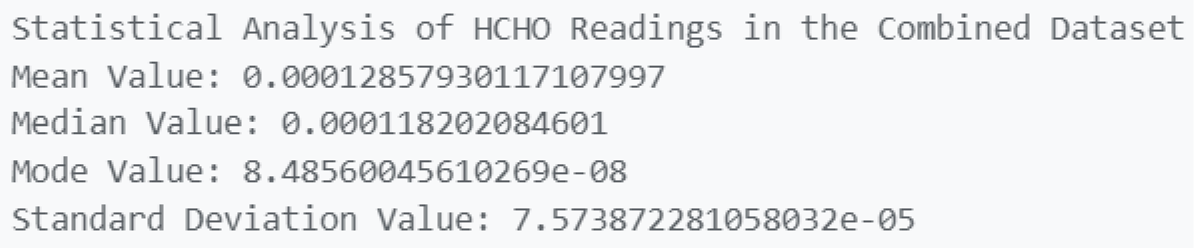
```
Statistical Analysis of HCHO Readings in Jaffna
Mean Value: 0.000110588005374901
Median Value: 0.0001030325171612
Mode Value: 4.10346731297695e-07
Standard Deviation Value: 6.0713737567834214e-05
```

These results could be tabulated to be compared easily as follows.



```
              count      mean       std  min       25%       50%       75%       max
Location
Colombo      1280.0  0.000165  0.000088  0.0  0.000100  0.000155  0.000222  0.000440
Jaffna       1381.0  0.000111  0.000061  0.0  0.000067  0.000103  0.000146  0.000314
Kandy         918.0  0.000122  0.000071  0.0  0.000069  0.000114  0.000165  0.000351
Kurunegala   1166.0  0.000140  0.000073  0.0  0.000085  0.000133  0.000187  0.000375
Matara        852.0  0.000104  0.000068  0.0  0.000053  0.000091  0.000145  0.000362
Monaragala   1039.0  0.000137  0.000074  0.0  0.000080  0.000129  0.000188  0.000373
Nuwara Eliya  637.0  0.000104  0.000066  0.0  0.000051  0.000095  0.000144  0.000311
```

Finally, the combined statistics could be obtained to have an overall idea of how the HCHO emissions are in the whole of Sri Lanka. This is not an accurate analysis as only six cities are being considered, which is a very small proportion.

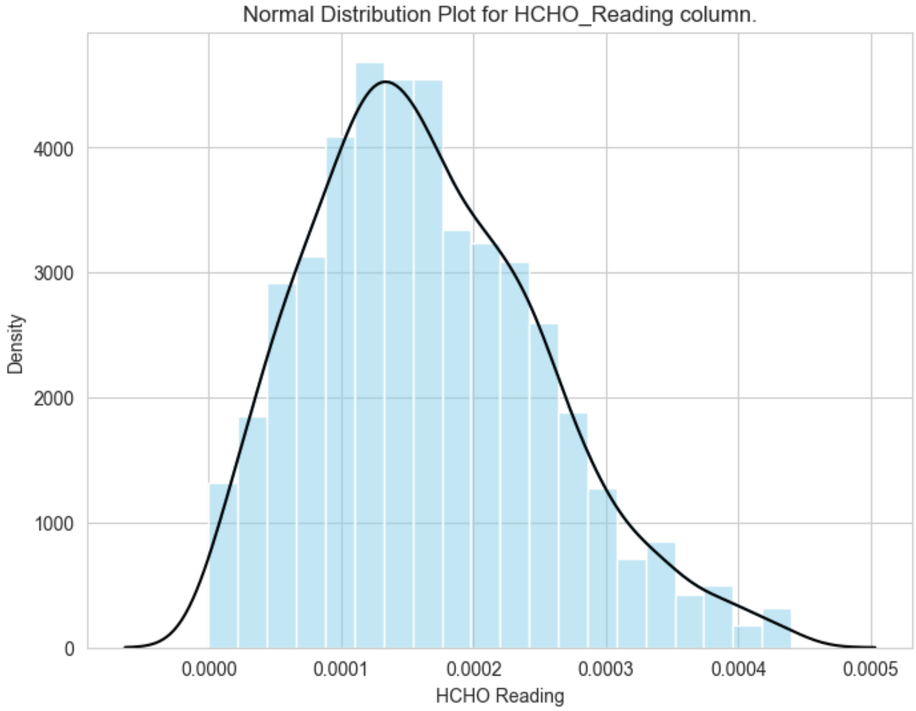


```
Statistical Analysis of HCHO Readings in the Combined Dataset
Mean Value: 0.00012857930117107997
Median Value: 0.000118202084601
Mode Value: 8.48560045610269e-08
Standard Deviation Value: 7.573872281058032e-05
```

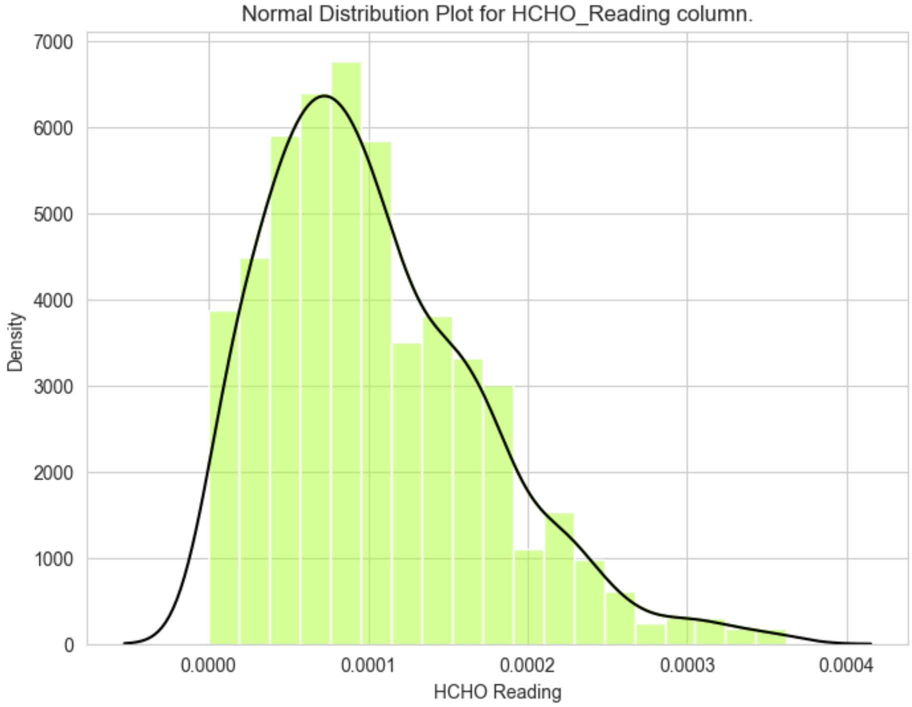
2.4. Visualization of data distribution

The distribution of the data can be visualized using many techniques such as box plots and normal distribution plots. Box plots do not always provide all the details regarding the data so normal distribution proves to be a better option. The following are the plots of the normal distribution for each location after the outliers have been removed using the threshold function.

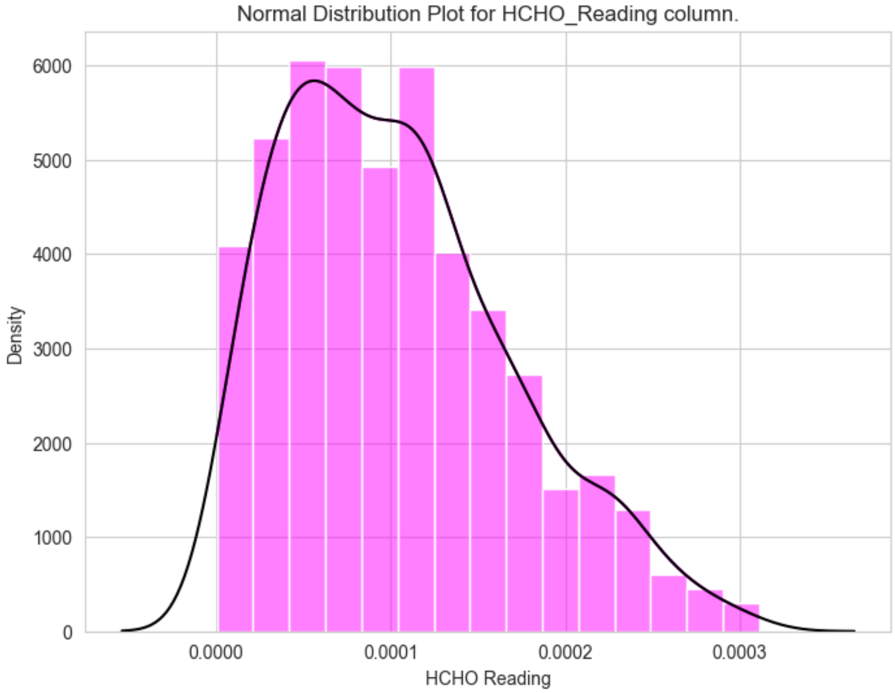
- Colombo



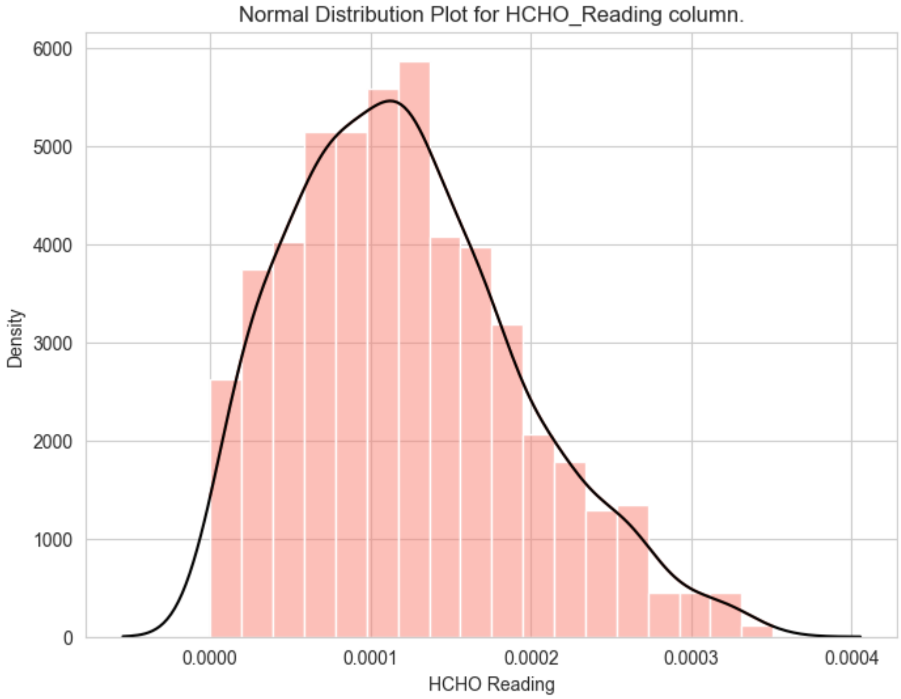
- Matara



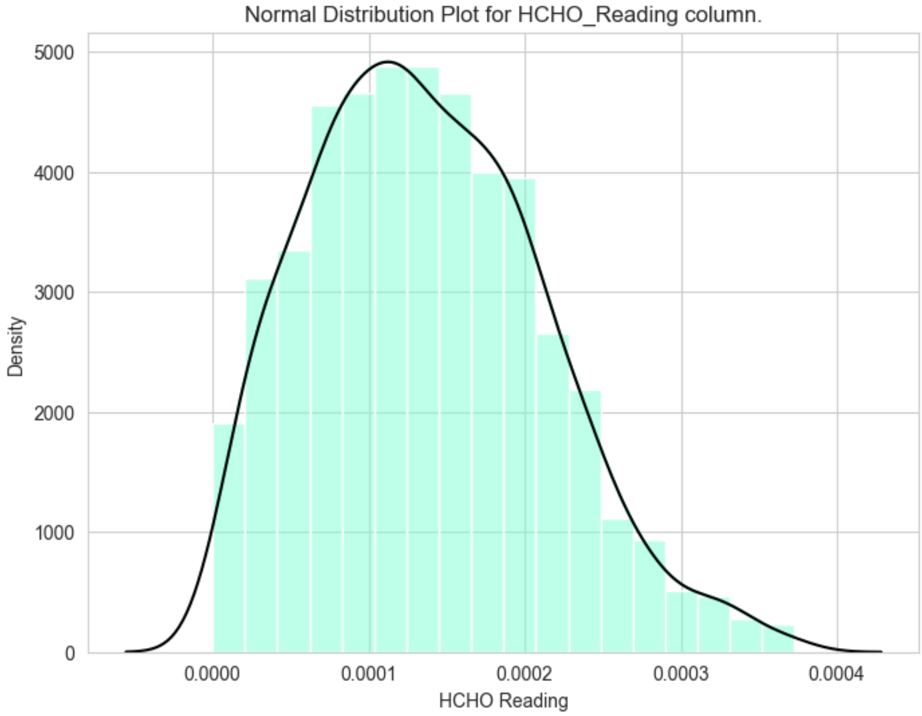
- Nuwara Eliya



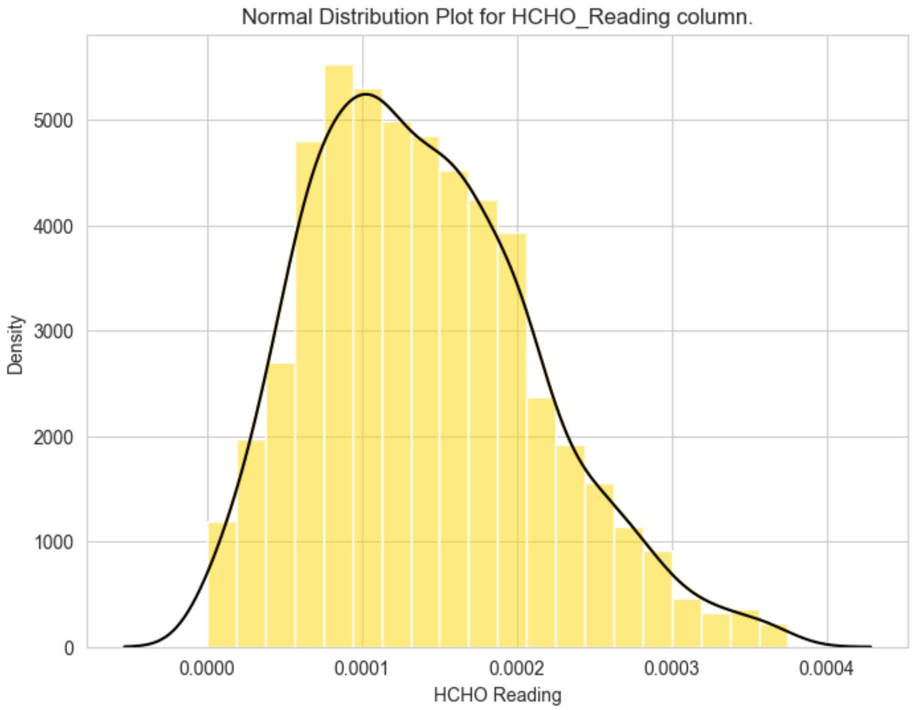
- Kandy



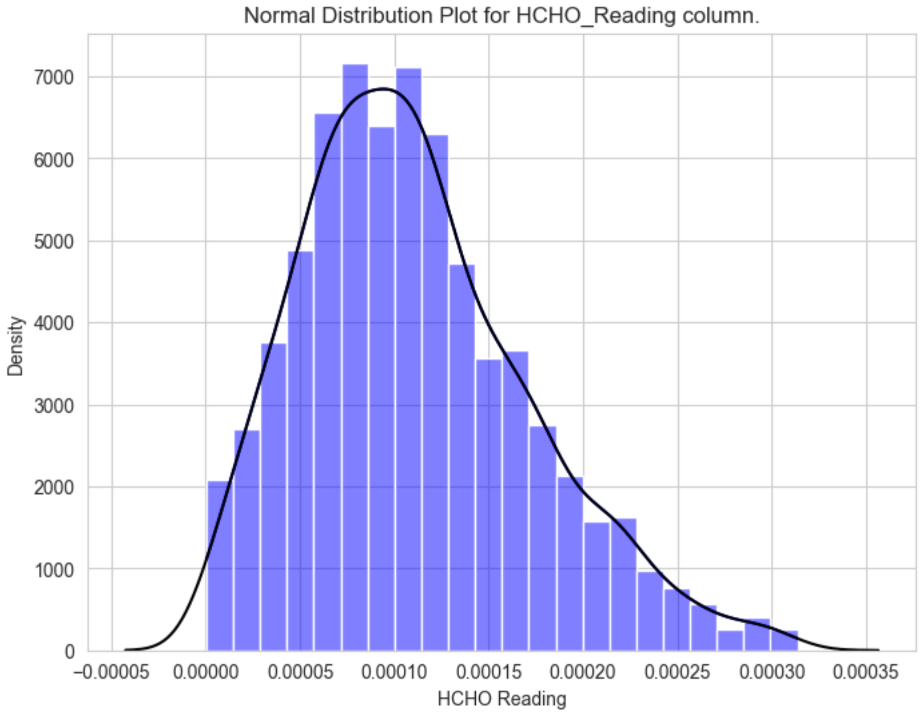
- Monaragala



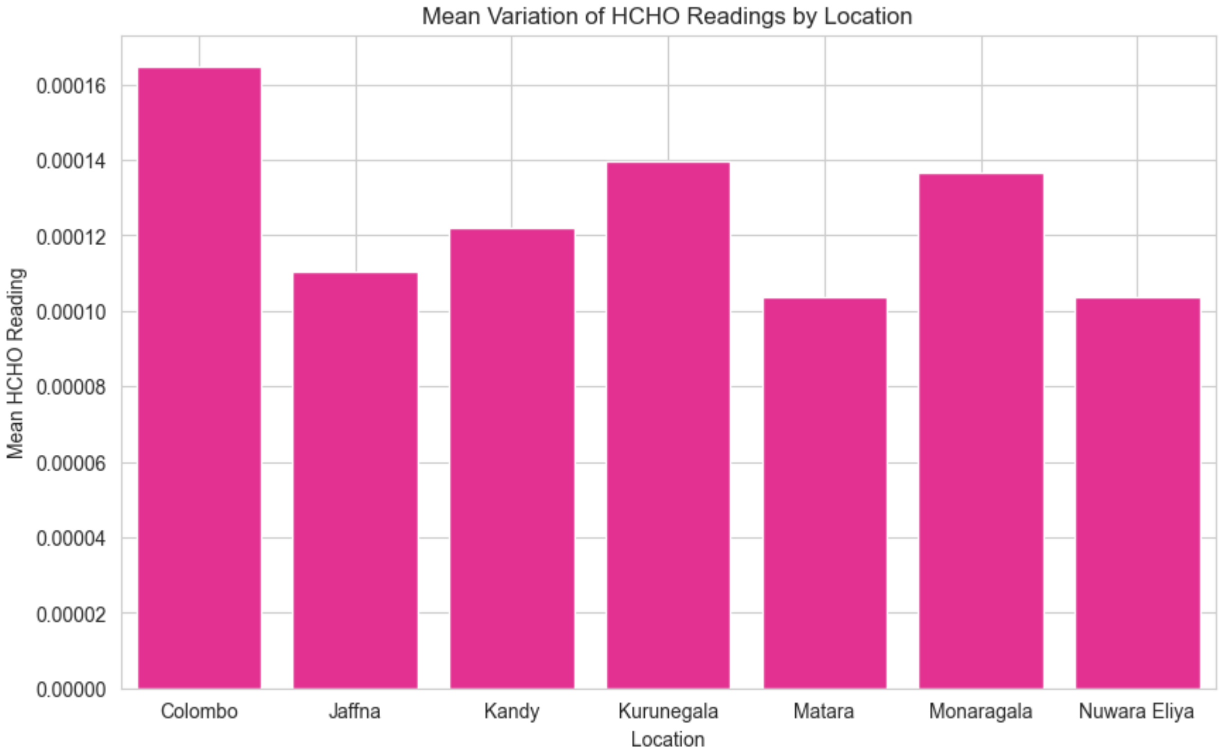
- Kurunegala



- Jaffna

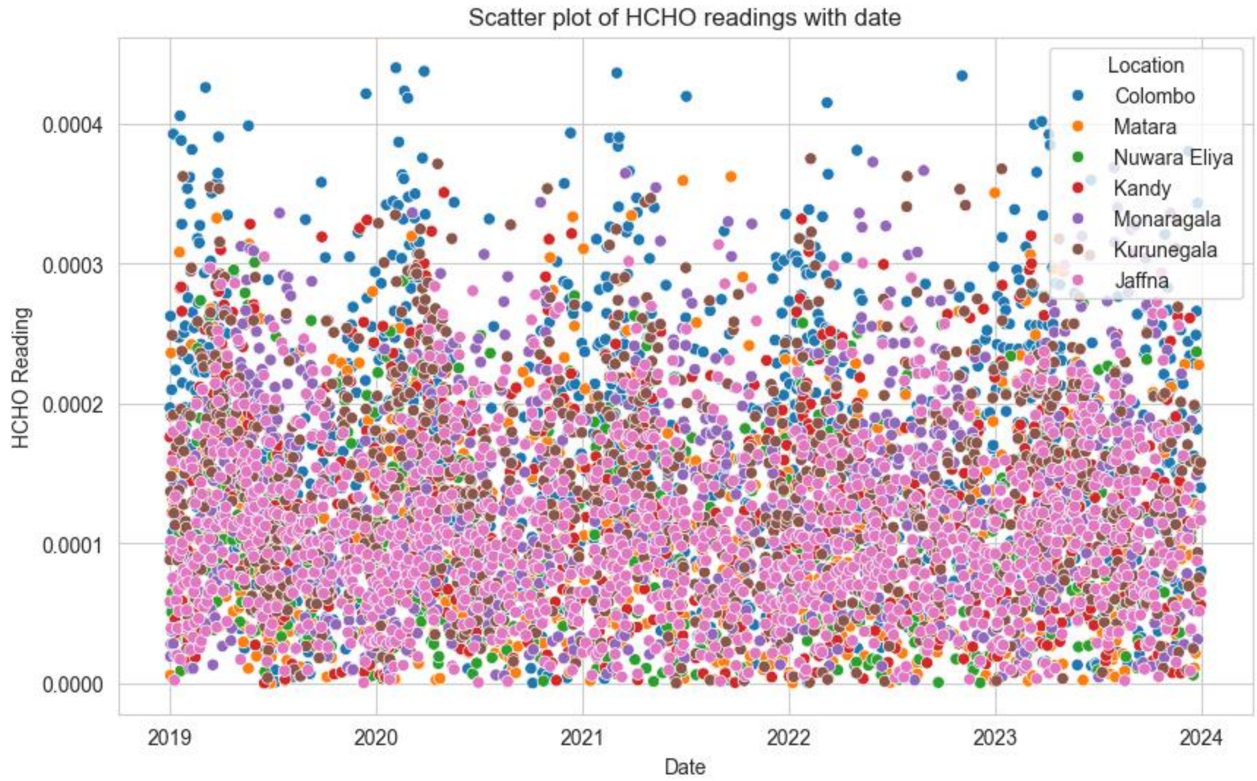


Next, a few generic plots have been made to see how the HCHO readings compare between different cities.



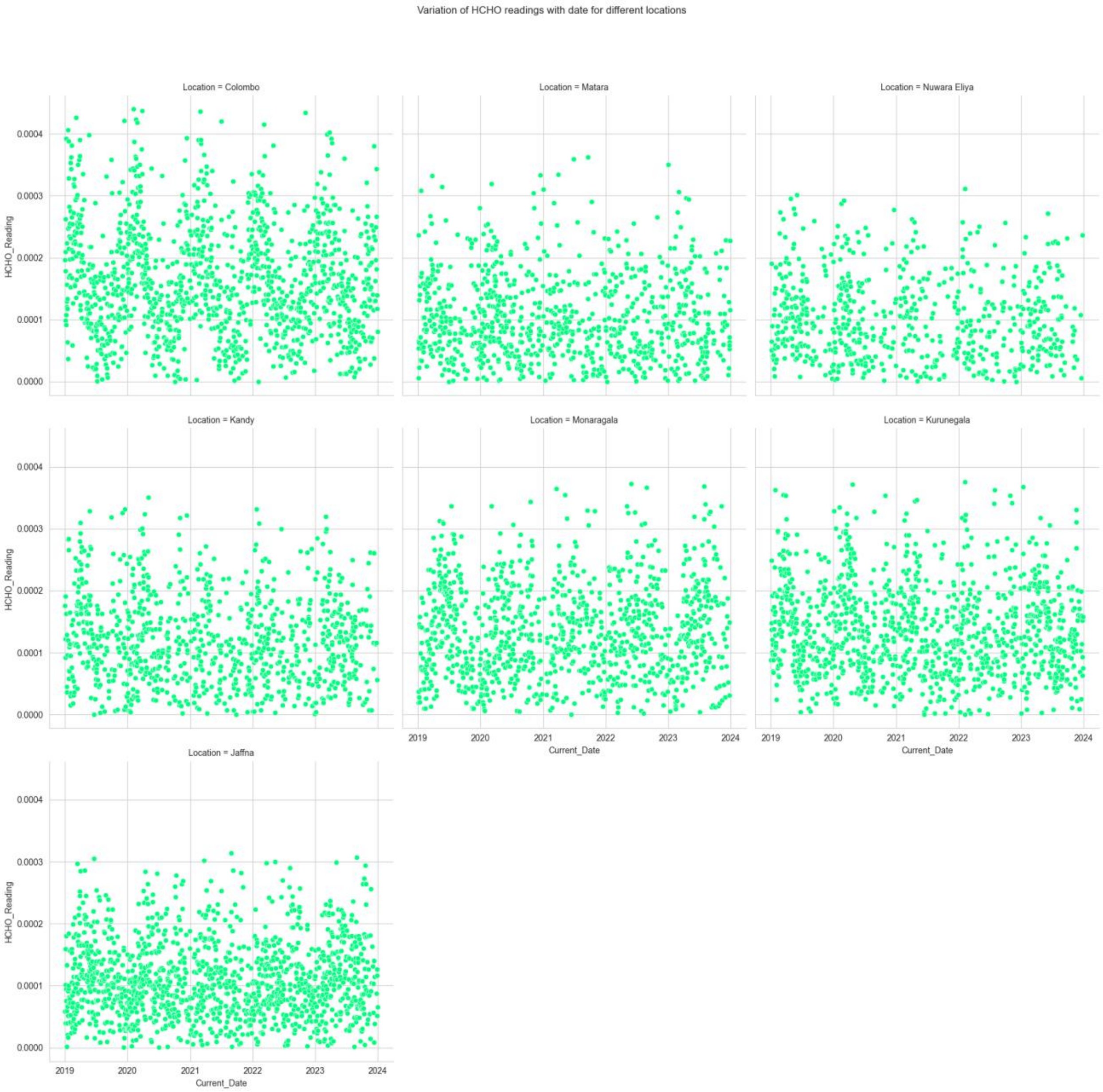
This plot shows how the average HCHO readings vary across the different locations. The plot depicts that Colombo has the highest mean HCHO reading whereas Matara and Nuwara Eliya have the lowest, with Jaffna closely behind. It could be interpreted that these changes occur due to the variation of geographical features across these cities and the levels of industrialization.

To get a better understanding of how these values are spread, a scatter plot could be used.



This scatter plot shows how the readings of the different cities have been spread across the years. However, it is difficult to analyze each city individually using this plot.

Thus, a Facet Grid was used to make subplots and plot each city on a separate set of axes.

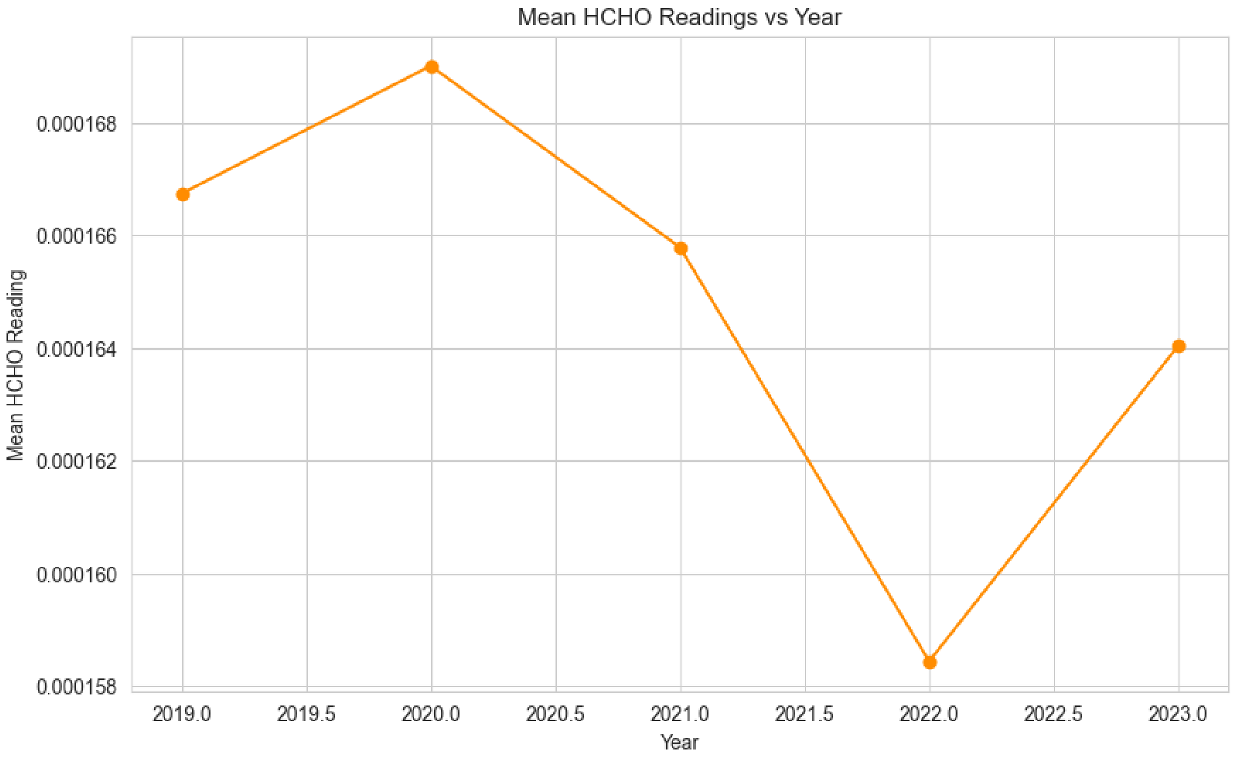


This diagram provides a more detailed analysis of each city. It is noticeable that Colombo has a greater spread over a range of values. However, Jaffna and Nuwara Eliya seem to have a more compact spread within a smaller range of values.

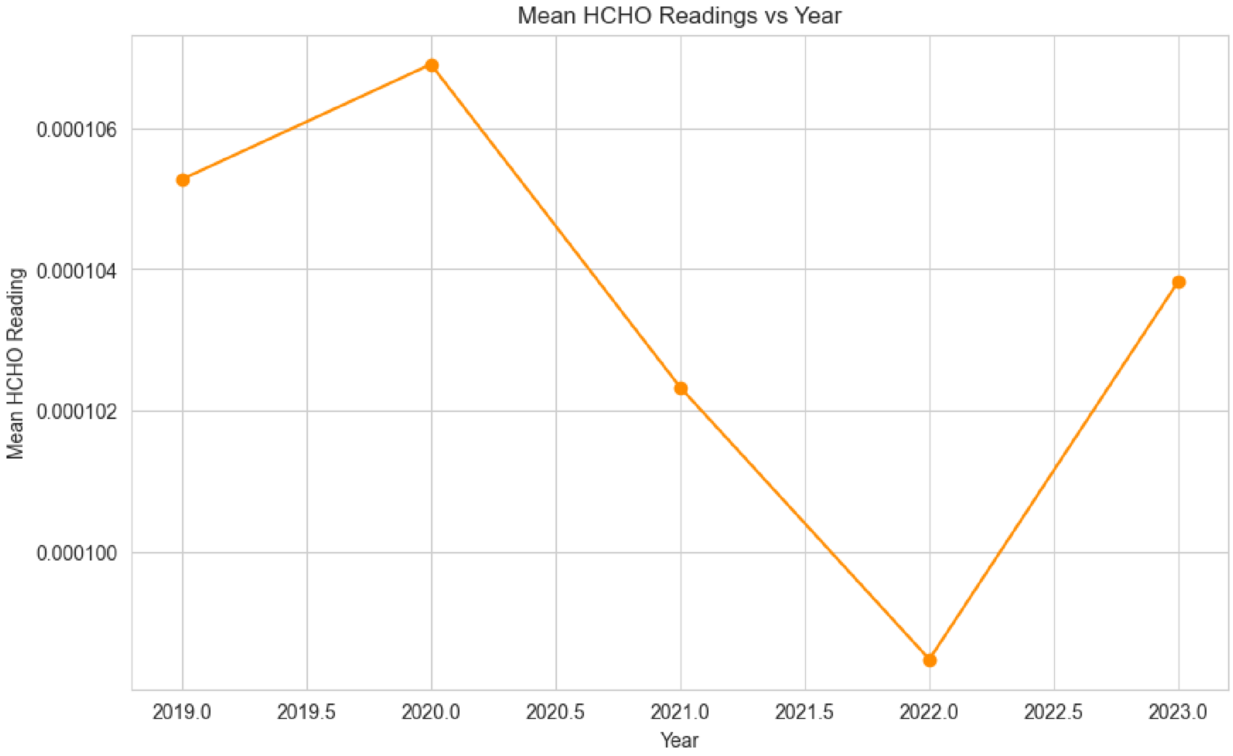
Next, we could plot the mean HCHO reading against the year. This would enable us to see the time periods when emissions were low and when the emissions were high.

Results were as follows.

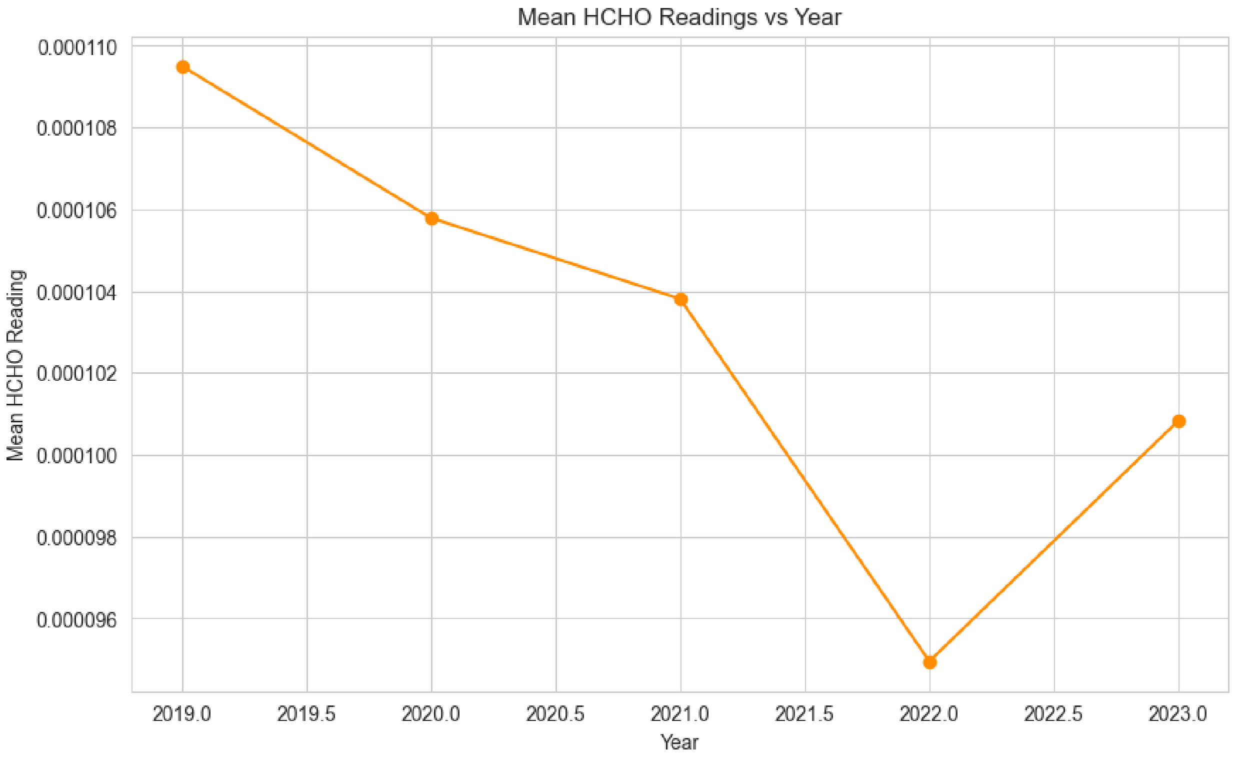
- Colombo



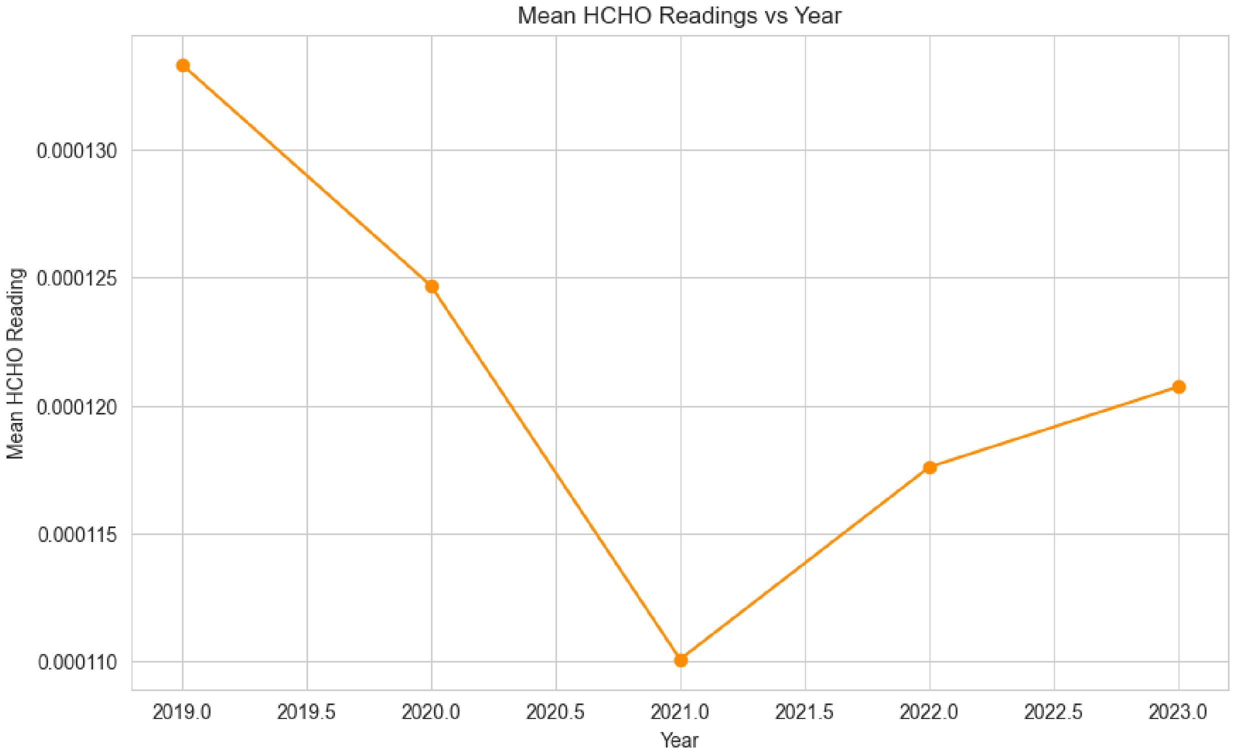
- Matara



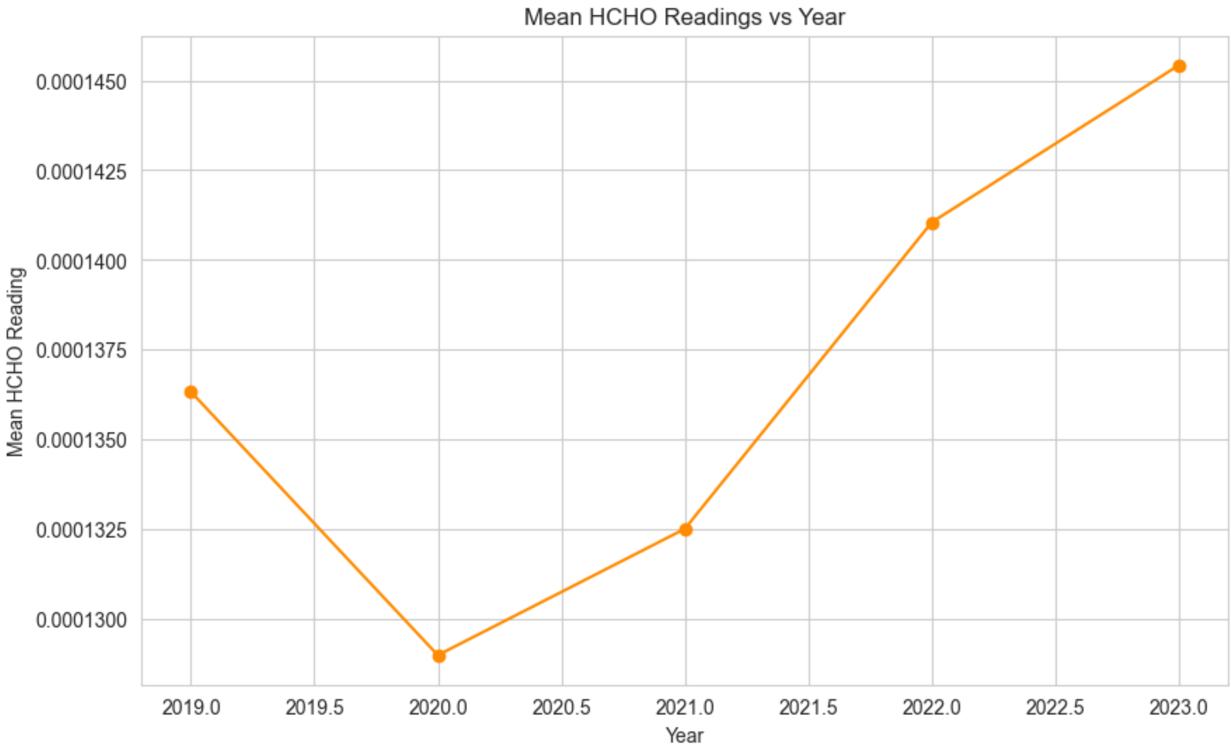
- Nuwara Eliya



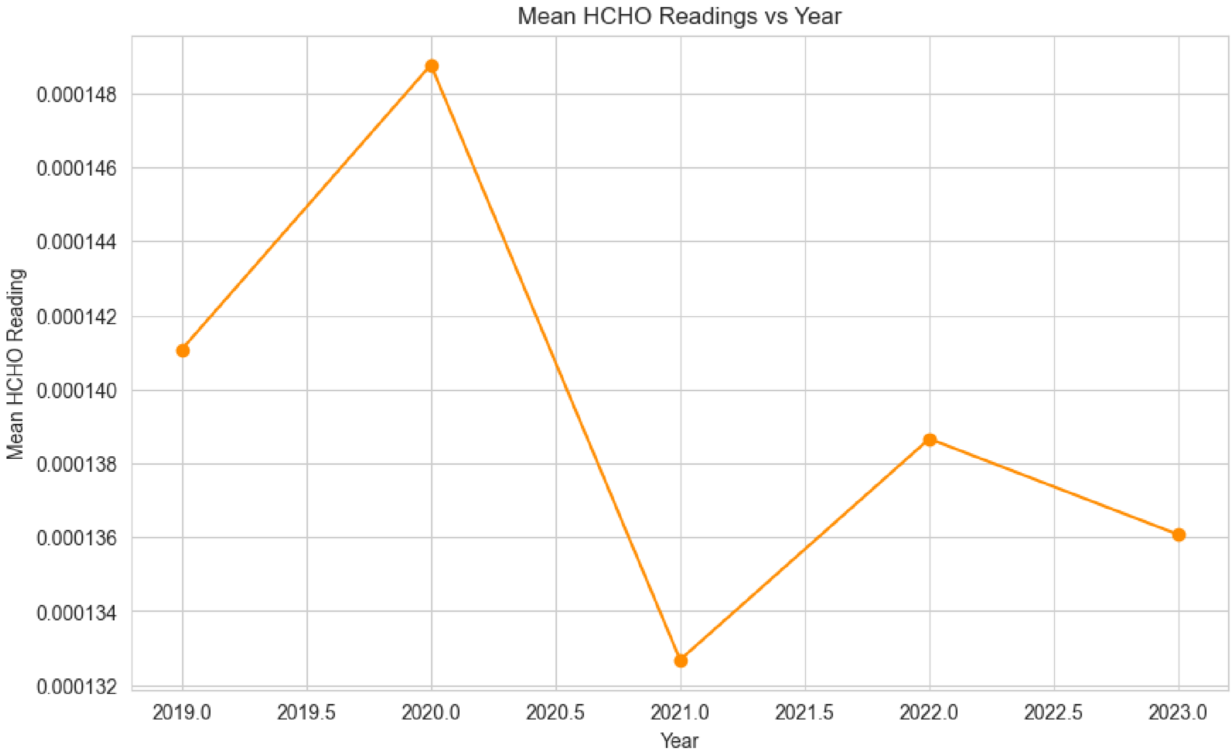
- Kandy



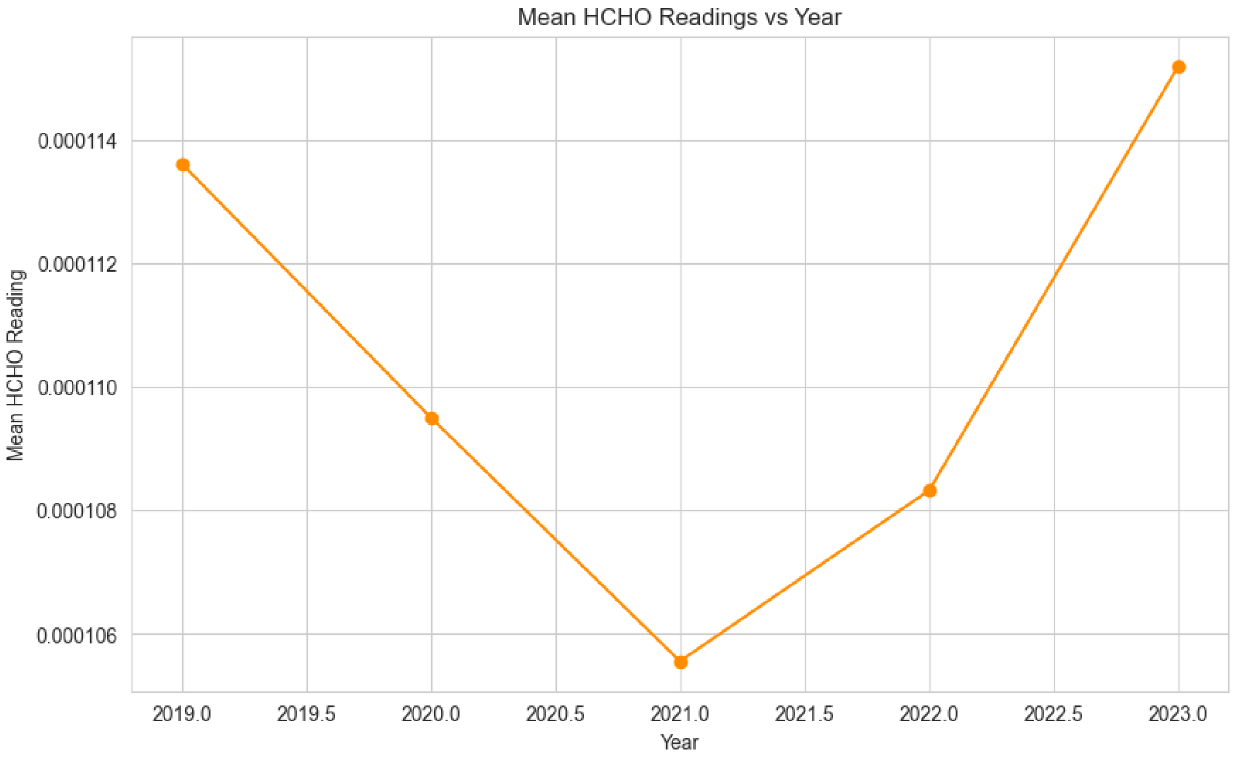
- Monaragala



- Kurunegala



- Jaffna



A common trend which is visible throughout all the cities is that the mean HCHO reading is roughly lower from 2019.5 till 2022, compared to the other time periods. The value significantly drops from 2019 to 2020 and further till 2021 before starting to increase slightly till 2022. However, even in 2022, for most cities the value is comparatively lower than the other periods of time.

Incidentally, it could be said that this reduction in HCHO readings was due to the Covid-19 pandemic which caused a reduction in industrialization, while reducing traffic and combustion in the environment.

This finding proves to a certain extent that the occurrence of the Covid- 19 pandemic caused an impact on the Tropospheric Formaldehyde emissions around the cities of Sri Lanka.

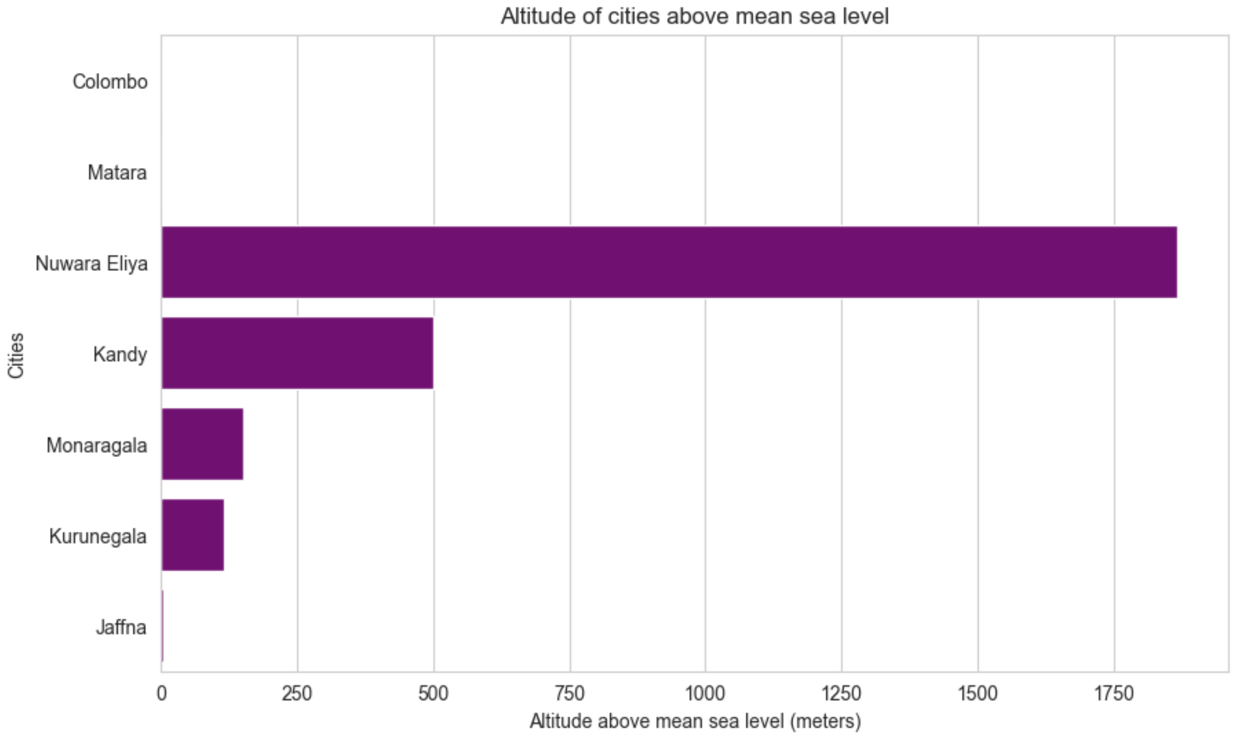
3. Spatial- Temporal Analysis

3.1. General Analysis

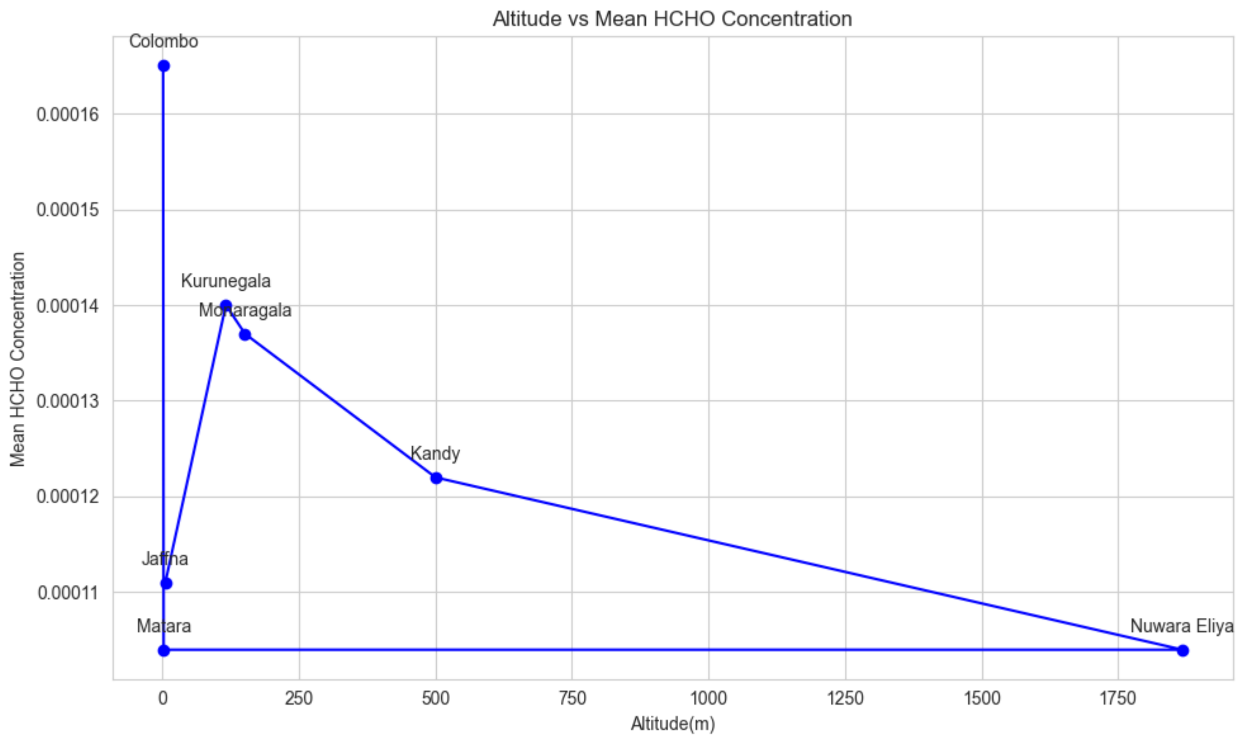
1. Altitude above mean sea level

The following are the altitude of the seven cities above the mean sea level.

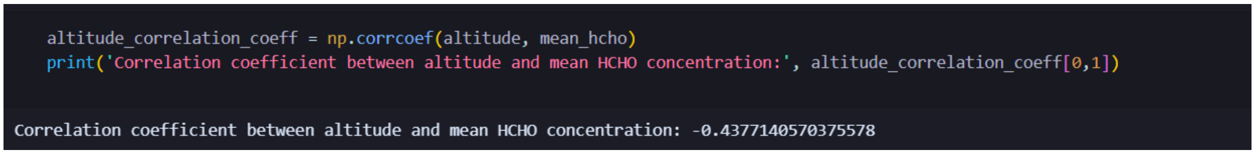
City	Mean height above sea level (m)
Colombo	1
Matara	2
Nuwara Eliya	1,868
Kandy	500
Monaragala	151
Kurunegala	116
Jaffna	5



To identify the relationship between the altitude and the mean HCHO concentration, a line plot was made between the two variables.



This graph does not provide much useful information; thus, the correlation coefficient can be calculated between the two variables.



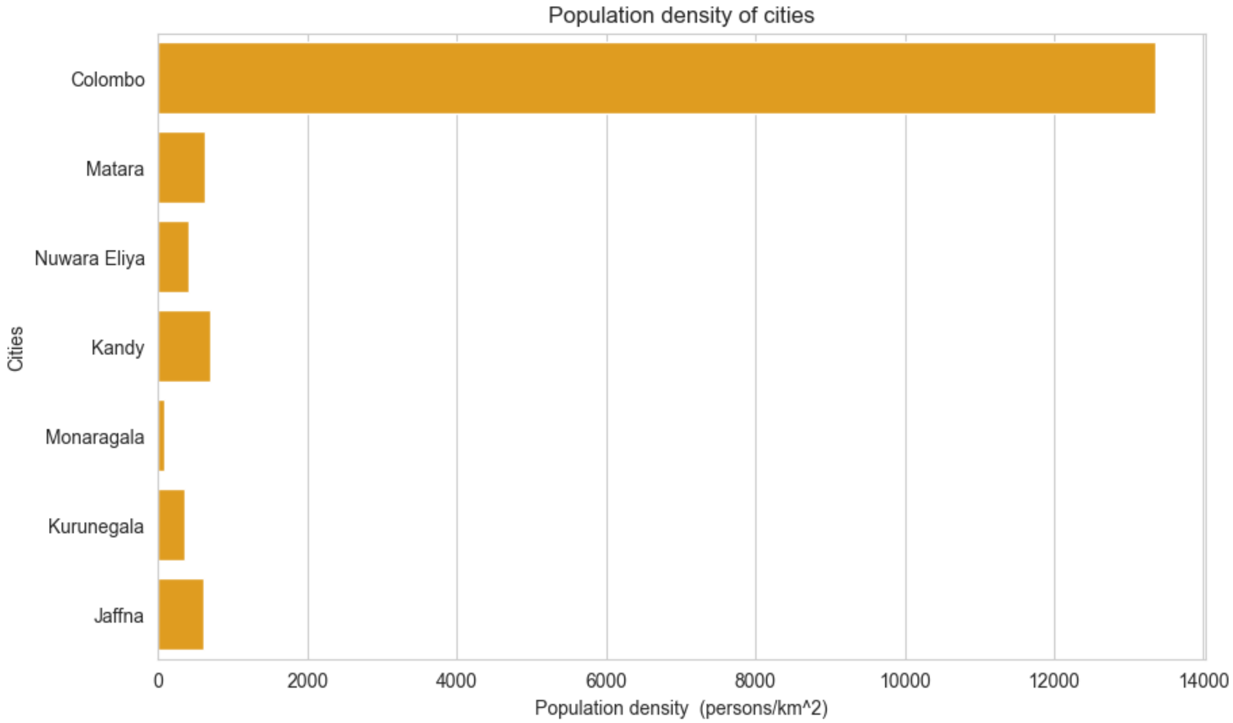
```
    altitude_correlation_coeff = np.corrcoef(altitude, mean_hcho)
    print('Correlation coefficient between altitude and mean HCHO concentration:', altitude_correlation_coeff[0,1])

Correlation coefficient between altitude and mean HCHO concentration: -0.4377140570375578
```

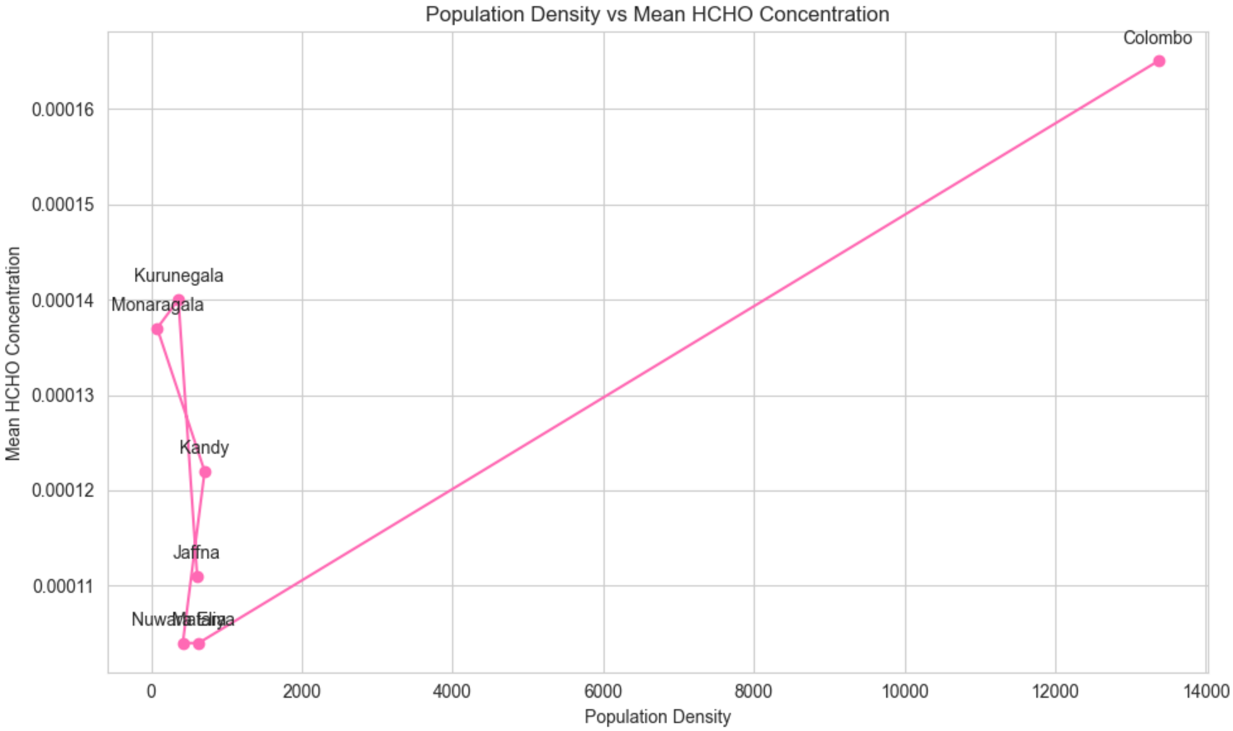
The correlation coefficient returns a value of -0.4377, which suggests an intermediately strong inverse relationship. This means that in cities at higher altitudes, the HCHO concentration is potentially lower than comparatively lower lying cities.

2. Population density of cities

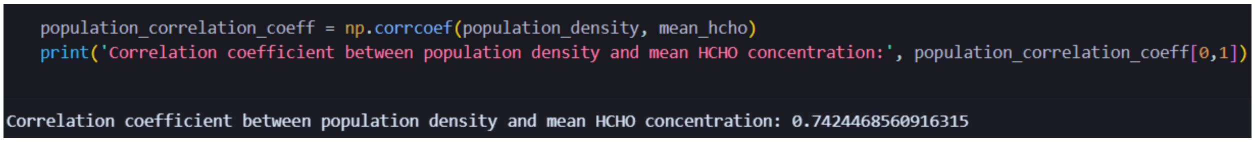
The population density of a city is a measure of human population across a certain area of land. This is a measure to depict the population of individuals within a city or a country. Following are the population densities of the cities measured in persons per square kilometer.

City	Population Density (persons/ km^2)
Colombo	13,364
Matara	630
Nuwara Eliya	420
Kandy	710
Monaragala	80
Kurunegala	362
Jaffna	611





This is a line plot of population density against mean HCHO concentration, which does not suggest a straightforward relationship. Thus, once again the correlation coefficient can be calculated.

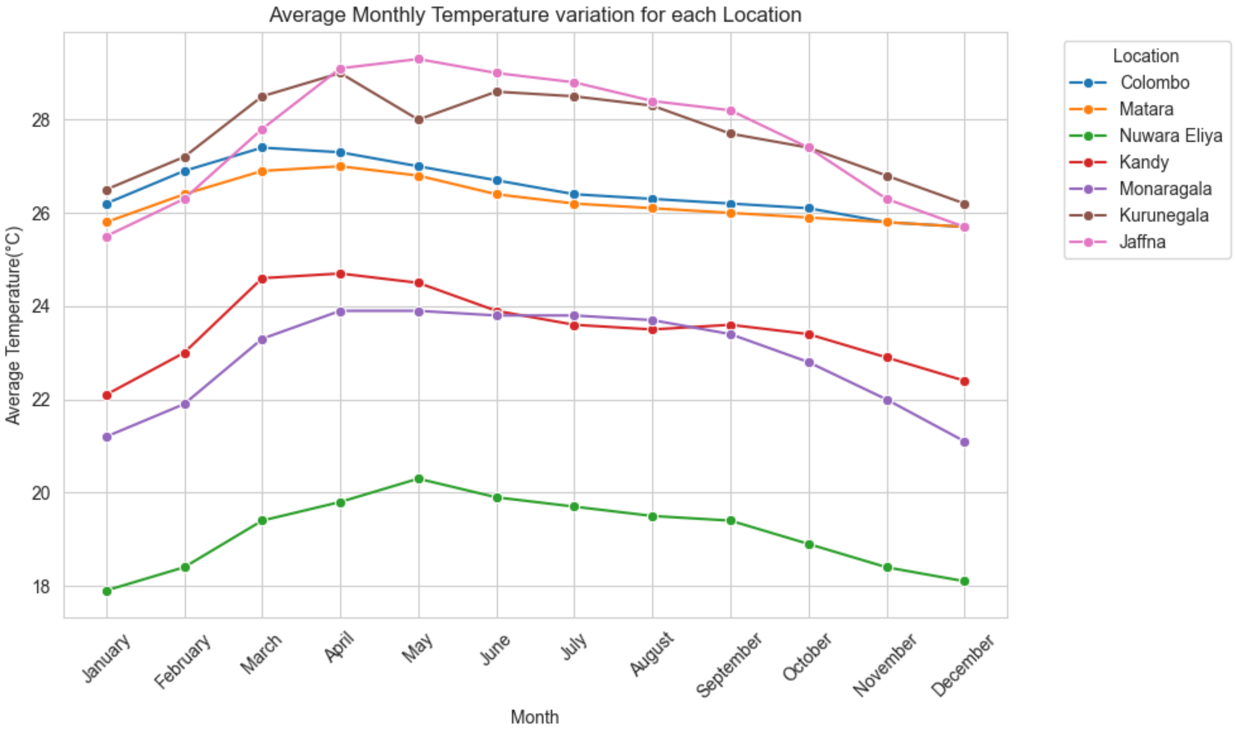


```
    population_correlation_coeff = np.corrcoef(population_density, mean_hcho)
    print('Correlation coefficient between population density and mean HCHO concentration:', population_correlation_coeff[0,1])

Correlation coefficient between population density and mean HCHO concentration: 0.7424468560916315
```

The correlation coefficient returns a value of 0.7424, which suggests a strong direct relationship between the population density and mean HCHO concentration. This means that a higher population density returns a higher mean HCHO concentration.

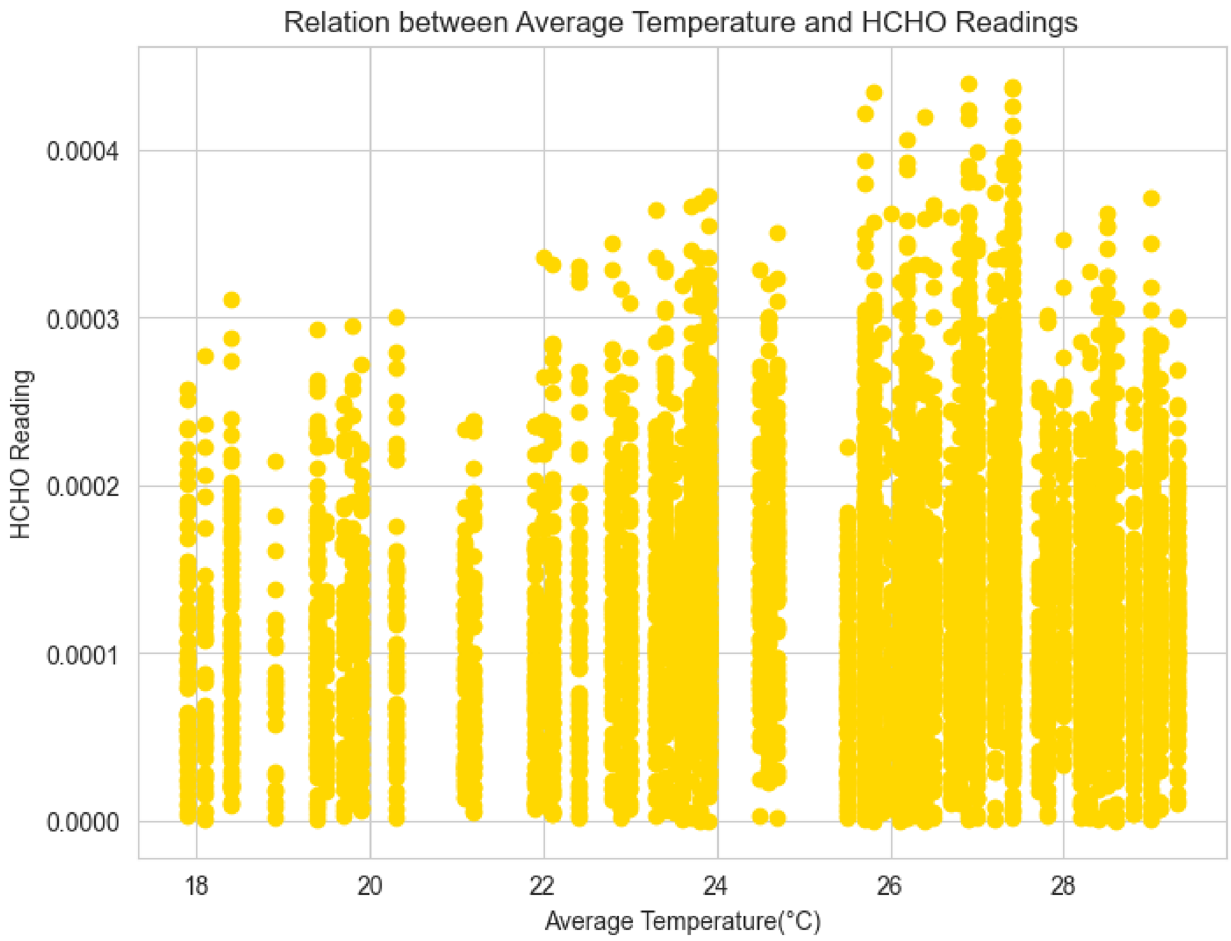
3. Average Temperature

The line plot below suggests the average temperature variation at the locations across the year, for each month.

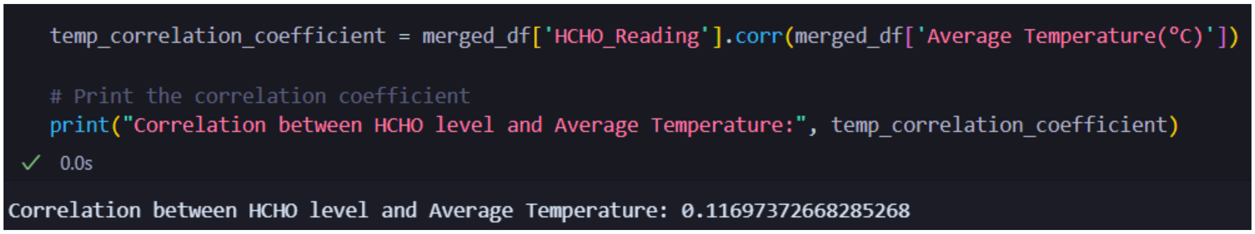


Through this we can identify that Nuwara Eliya is the city with the least average temperature throughout the year whereas Jaffna records the highest temperature in the month of May. Monaragala and Kandy have intermediate temperatures compared to the other locations.

It could also be noted that the temperatures are lower during the months from October to February (possibly due to the major monsoon season in Sri Lanka), and usually experience higher temperatures during the months of March to August. However, it could also be noted that Kandy does not exactly follow this pattern.



The scatter plot shows the relation between the temperature and the HCHO readings. There is a greater spread in the values when the temperatures are closer or in the range of 26°C to 28°C.



```
temp_correlation_coefficient = merged_df['HCHO_Reading'].corr(merged_df['Average Temperature(°C)'])

# Print the correlation coefficient
print("Correlation between HCHO level and Average Temperature:", temp_correlation_coefficient)
```

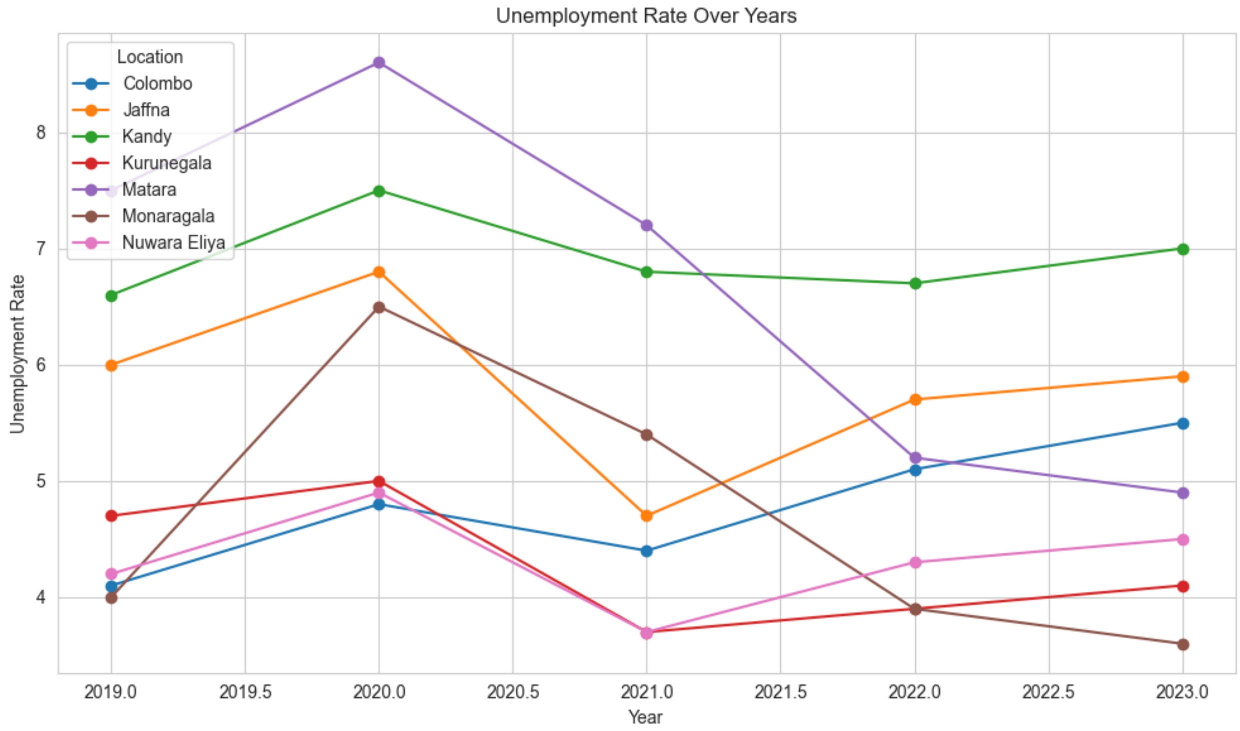
✓ 0.0s

```
Correlation between HCHO level and Average Temperature: 0.11697372668285268
```

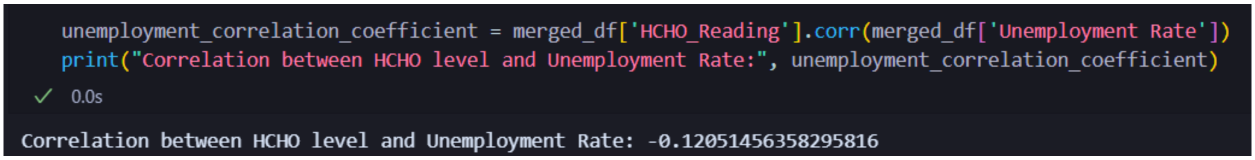
The correlation coefficient suggests that when the average temperature is higher, then the mean HCHO concentration would also be higher. It is a very weak relationship as the value is a positive value close to zero.

4. Unemployment Rate

The line plot below shows the variation of unemployment rate for each year across each location.



There is no significant trend in the unemployment rate across these cities. However, it could be noted that the unemployment rate increases in the year 2020. This could possibly be due to the occurrence of the Covid-19 pandemic.



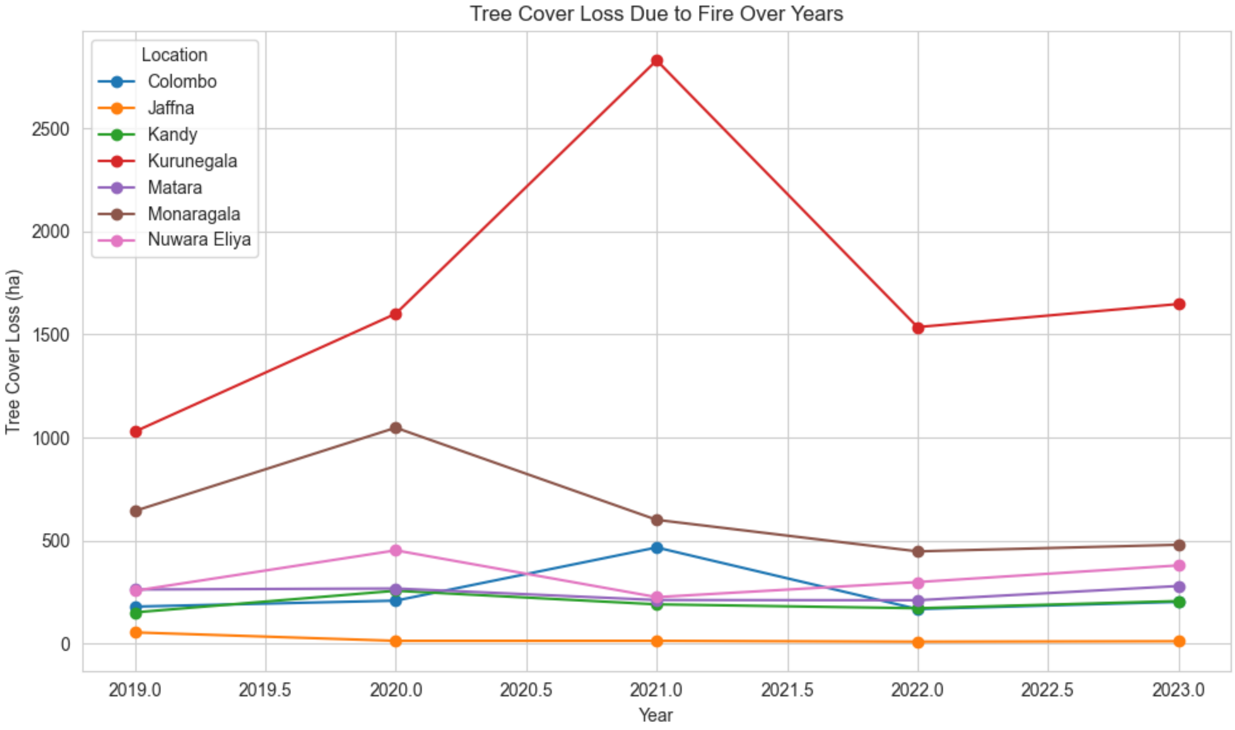
```
unemployment_correlation_coefficient = merged_df['HCHO_Reading'].corr(merged_df['Unemployment Rate'])
print("Correlation between HCHO level and Unemployment Rate:", unemployment_correlation_coefficient)
```

✓ 0.0s

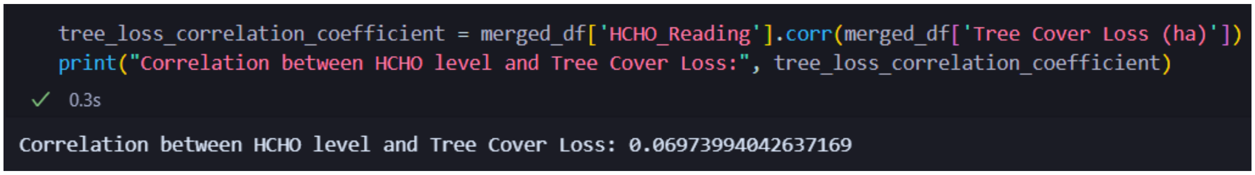
```
Correlation between HCHO level and Unemployment Rate: -0.12051456358295816
```

There is a weak negative relationship between the HCHO level and the unemployment rate. This is suggested by the correlation coefficient of -0.1205.

5. Tree cover loss due to fire related events



The line plot above shows how much of tree cover has been lost over the years of 2019 to 2023 across the seven locations.



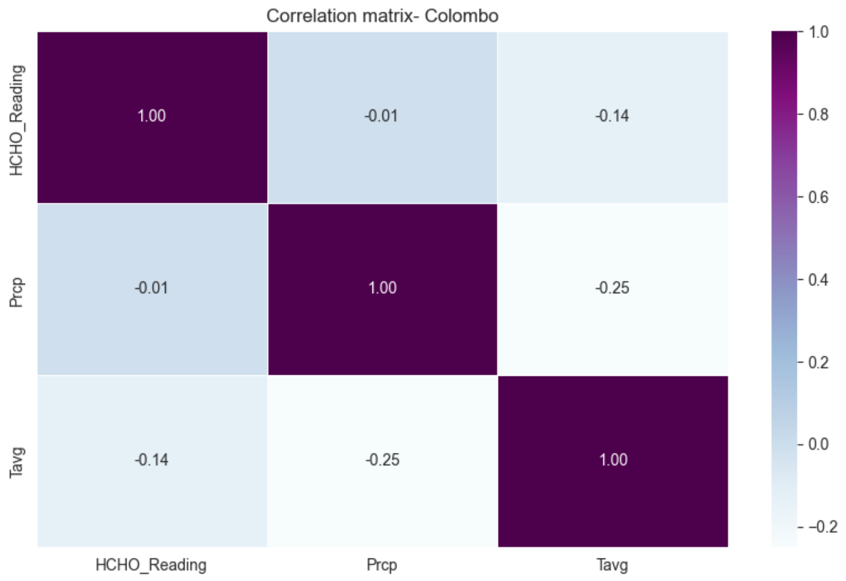
```
    tree_loss_correlation_coefficient = merged_df['HCHO_Reading'].corr(merged_df['Tree Cover Loss (ha)'])
    print("Correlation between HCHO level and Tree Cover Loss:", tree_loss_correlation_coefficient)
 ✓  0.3s

Correlation between HCHO level and Tree Cover Loss: 0.06973994042637169
```

The correlation coefficient between the HCHO level and the tree cover loss is a very weak positive relationship. An increase in fore related accidents directly causes an increased level of formaldehyde emission.

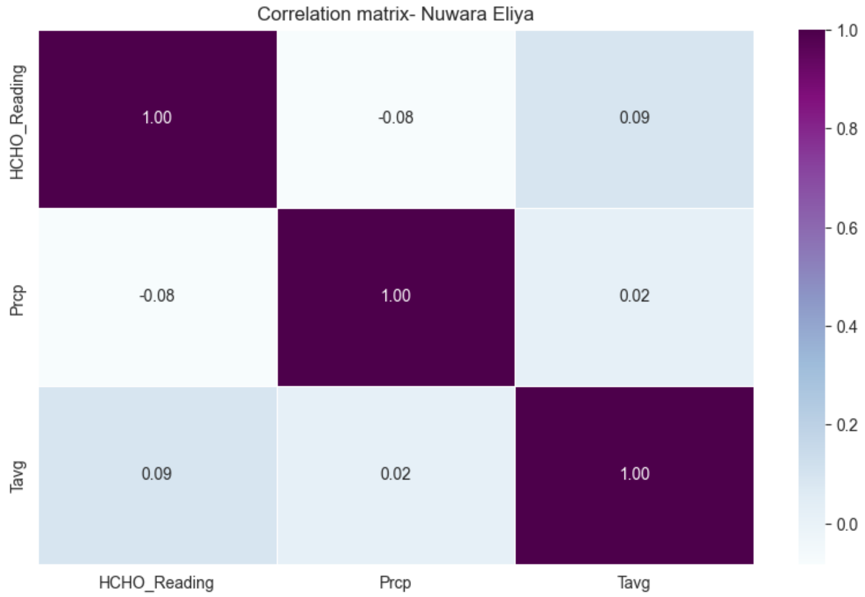
6. Precipitation and Average Temperature (Census Data)

Based on census data found on the internet, a data frame was formed to identify the relationship between the precipitation and the HCHO reading. However, this data was only available for Colombo, Kurunegala and Nuwara Eliya.

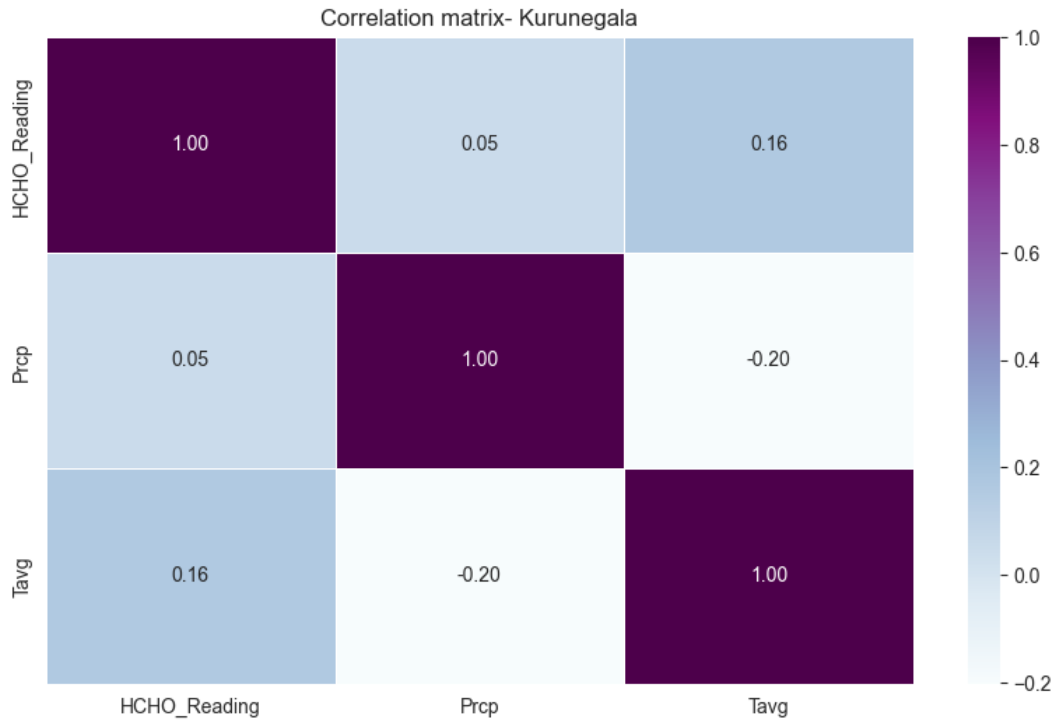
- Colombo



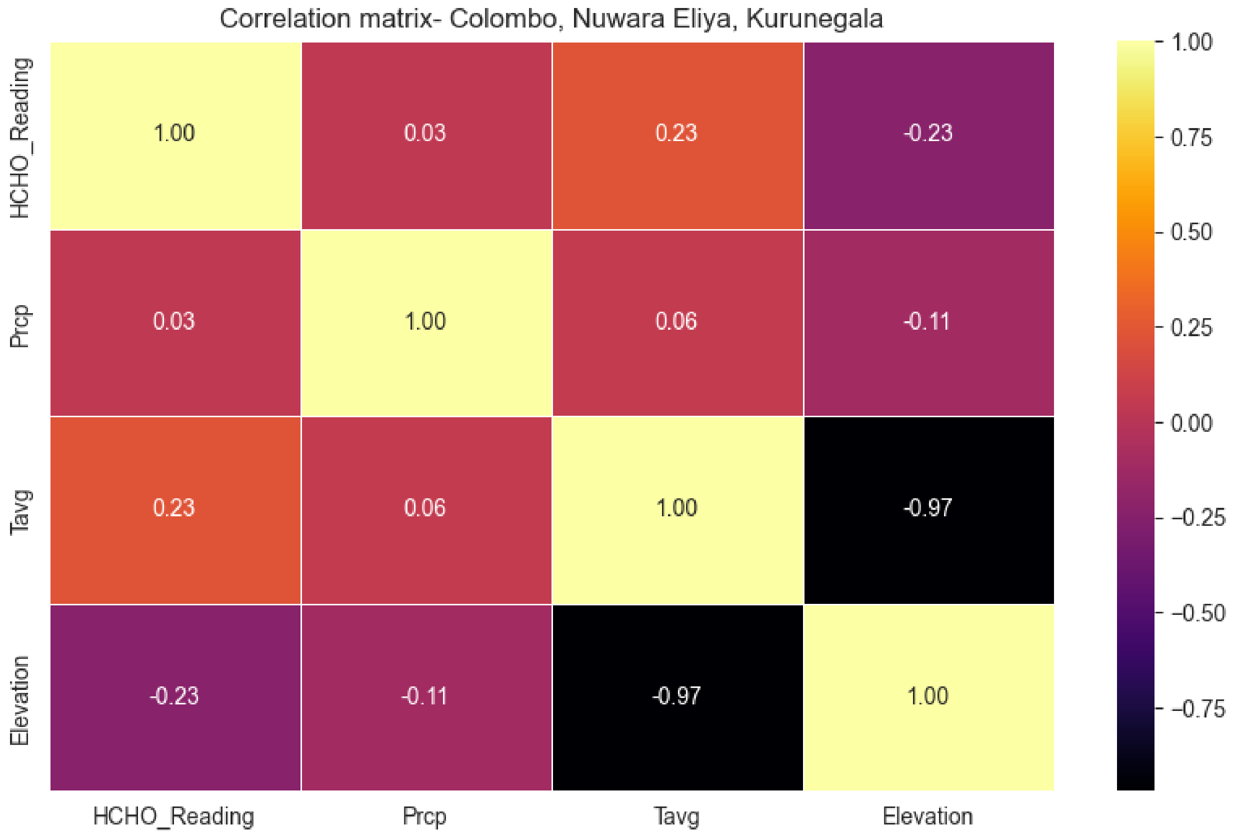
- Nuwara Eliya



- Kurunegala


Correlation matrix- Kurunegala
HCHO_Reading
Prcp
Tavg
1.00
0.05
0.16
0.05
1.00
-0.20
0.16
-0.20
1.00
HCHO_Reading
Prcp
Tavg
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

- Combined- Colombo, Kurunegala and Nuwara Eliya



Based on this correlation matrix, we can confirm the following.

- Precipitation has a very weak positive relationship with the HCHO readings.
- Average temperature of the location has a stronger relationship with the HCHO readings.
- However, the elevation has an inverse relationship with the HCHO readings as discussed above using the data from Google.

4. Machine Learning

4.1.Machine Learning Workflow

Three machine learning models were developed for each location: Autoregressive Integrated Moving Average (ARIMA), Seasonal Autoregressive Integrated Moving Average (SARIMA) and the Gaussian Process Regressor.

Autoregressive Integrated Moving Average (ARIMA)

The ARIMA model makes use of three main forecasting components: autoregression, integration and moving average.

The autoregression component aims to capture a relationship between a stated observation and past observation. An assumption is made that the predicted value is a result of a linear relationship with the past observations. It is denoted by the parameter “p”.

The primary role of the integration component is to make the time series stationary, which means that the statistical measures such as the mean, standard deviation and the correlation between features remain constant throughout time. This is done by extracting the difference between successive observations and removing any possible trends or seasonal patterns. It is denoted by the parameter “q”.

A model is applied to the past observations and the residual errors are calculated. Following this, the moving average component identifies a relationship between the predicted observation and the residual errors. It is denoted by the parameter “q”.

Overall, the ARIMA model is represented as ARIMA (p, d, q)

Seasonal Autoregressive Integrated Moving Average (SARIMA)

The SARIMA model is an extension of the ARIMA model which makes use of seasonality to forecast values in a time series process. It consists of all the components in the ARIMA model along with an additional seasonal component.

The seasonal component makes use of metrics such as seasonal moving averages, seasonal differences and seasonal autoregression to calculate the duration of a specific seasonal pattern. This is denoted by the parameter “m”.

The overall representation of the SARIMA model is SARIMAX (p, d, q) (P, D, Q) m, where P, D, Q are the seasonal parameters.

Gaussian Process Regressor

This is a non- parametric supervised learning method which could be modelled for time series scenarios. A distribution is initially defined based on possible relationships between the input and the output variable before any data is observed. A Gaussian Process consists of a mean function which represents the expected value at each input node, and a kernel function which utilizes covariance to identify the correlations between the features.

Following this, Bayes theorem is used to constantly streamline the relationships identified. This is continued and the predictions are generated for new input nodes by using the distribution and the functions at a said point. The model could be tuned using hyperparameters for optimality.

One of the key pros of the Gaussian process Regressor is that it generates estimates of the uncertainty in predictions along with the mean. This can help in identifying the difference between the actual and the predicted values.

The auto Arima library was used to obtain the optimal values for the order of p, d, q to train the model. An 80% training split was used and 20% split was used for testing.

Root Mean Square Error (RMSE) was used as the evaluation metric to analyze the performance of the machine learning models. Root Mean Square Error is a measure of the difference between the estimated and actual values of the target variable, which is the HCHO Reading in our scenario.

4.2.Model Development Methodology

Following is the process followed for each location in developing the models.

- The csv file is read and converted into a dataframe using pandas.
- The location column is dropped as each location is being trained individually and a separate model is created for each location.
- The “Current_Date” column is converted into datetime format, so that it is understood by the model as the index column.

- The dataset is split into the testing and training components. Eighty percent of the data is used for training and the rest is used for testing.

ARIMA Model

- Auto Arima is used to generate the best parameters for the ARIMA model, and these are applied to an instance of the ARIMA model and trained.
- The training summary and the diagnostics are printed on the console.
- The model is tested, and the predictions are printed along with the Root Mean Squared Error (RMSE).

SARIMA Model

- Auto Arima is used to generate the best parameters for the SARIMA model, and these are applied to an instance of the SARIMA model and trained.
- The training summary and the diagnostics are printed on the console.
- The model is tested, and the predictions are printed along with the Root Mean Squared Error (RMSE).

Gaussian Process Regressor

- A kernel is defined, and the restart optimizer parameter is provided.
- The model is fitted with the training data and the target column is defined.
- The model is made to predict the values and the Root Mean Squared Error (RMSE) is printed to evaluate the performance of the model.

The table below shows the performance of each of the models for each location, along with the optimal parameter order provided by the auto Arima library.

Location	**Order**	**Model Type**	**RMSE**
Colombo	(0,1,4)	ARIMA	8.7882e-05
Colombo	(0,1,4)	SARIMA	0.000257
Colombo	N/A	Gaussian Process	0.000433
Matara	(0,0,4)	ARIMA	7.1118e-05
Matara	(0,0,4)	SARIMA	0.000114

Matara	N/A	Gaussian Process	0.000136
Nuwara Eliya	(5,0,0)	ARIMA	6.1071e-05
Nuwara Eliya	(5,0,0)	SARIMA	9.9250e-05
Nuwara Eliya	N/A	Gaussian Process	0.000630
Kandy	(0,1,0)	ARIMA	0.000125
Kandy	(0,1,0)	SARIMA	8.4498e-05
Kandy	N/A	Gaussian Process	0.000147
Monaragala	(3,0,3)	ARIMA	9.1896e-05
Monaragala	(3,0,3)	SARIMA	0.000101
Monaragala	N/A	Gaussian Process	0.0001200
Kurunegala	(3,0,3)	ARIMA	6.6332e-05
Kurunegala	(3,0,3)	SARIMA	7.7129e-05
Kurunegala	N/A	Gaussian Process	0.000182
Jaffna	(3,0,3)	ARIMA	6.2313e-05
Jaffna	(3,0,3)	SARIMA	9.2893e-05
Jaffna	N/A	Gaussian Process	0.000131

For most regions, the SARIMA model performed better than the ARIMA model. This is because the ARIMA model has an extraordinarily low RMSE value which suggests the model may be overfitting. Thus, predictions were generated for the next 30 days using the SARIMA model. However, for the city of Colombo, ARIMA was used in preference to SARIMA as the latter produced negative values which are not acceptable.

These generated predictions for each location were stored in a separate CSV file under a new folder so that they could be used for the analysis.

5. Communication and Insights

5.1.Existing Research

Plenty of research has been done in bringing up measures to reduce the emission of Formaldehyde to the atmosphere. However, there is very minimal research when it comes to the use of Machine Learning and Deep Learning techniques and frameworks to predict the Formaldehyde levels over the next few days.

The research paper titled "Using machine learning approach to reproduce the measured feature and understand the model- to- measurement discrepancy of atmospheric formaldehyde" makes use of Light Gradient Boosting Algorithm to obtain predictions before using these predictions to compare with actual values to see the difference in readings. Light Gradient Boosting Algorithm is a novel machine learning algorithm developed by using the theory of decision tree gradient boosting. Furthermore, Shapley Additive Explanations (SHAP) has been used to provide a certain level of explainability for the results predicted by the model, thereby contributing to the validity and correctness of the results.

Another journal titled "Trustworthy Modelling of Atmospheric Formaldehyde Powered by Deep Learning" makes use of deep learning techniques to predict the Formaldehyde emissions around the states of India.

5.2.Summary of Findings

Based on this project, it has been proven successfully that the following factors play a role in impacting atmospheric formaldehyde emissions.

- Altitude
    - There is an intermediately strong negative relationship between the altitude and the HCHO concentration.
    - Formaldehyde emission is a result of photochemical reactions, which are reactions caused by sunlight. At higher altitudes, there is reduced solar radiation, which in turn leads to a lower rate of photochemical reactions.
- Population density
    - There is a very strong positive relationship between the population density and the mean HCHO concentration.

  - This could probably be because a higher population density causes higher levels of pollution and traffic flow which in turn releases emissions into the atmosphere.
- Average Temperature
  - There is a very weak positive relationship between the average atmospheric temperature and the HCHO concentration.
  - This could be due to several reasons. One of the main reasons could be the increased volatility of organic compounds. Organic compounds act as the source of formaldehyde. When the volatility of these compounds increases, they undergo photochemical reactions at a higher rate.
  - Moreover, higher temperatures cause an increased rate of chemical reactions. When the temperature increases, the reactants gain more kinetic energy and react at a much faster rate, resulting in higher volumes of HCHO being produced.
  - Higher temperatures could also lead to biogenic emissions from Volatile Organic Compounds (VOCs), which could result in natural disasters such as wildfires and microbial activity in earth soil, which are also causes of formaldehyde emission.
- Unemployment Rate
  - There is a very weak negative relationship between the unemployment rate and the HCHO emission.
  - This may occur as a lower unemployment rate results in more individuals travelling to work, which in turn leads to increased traffic flow volumes.
- Forest cover loss due to fire events
  - There is an extremely weak positive relationship between forest cover loss due to fire events and the mean formaldehyde emission.
  - This may occur because organic compounds may be combusted during fire accidents which could cause the release of Volatile Organic Compounds (VOCs).
- Lockdowns or Reduction in Human Activity (Covid 19)
  - Instances such as the Covid 19 pandemic caused the whole nation to come to a halt. Thus, this can reduce human activity and thereby reduce the levels of pollution.

  - This was shown to have a strong negative relationship with the HCHo readings. When these events occurred, the HCHO readings seemed to drop by a large margin.
- Precipitation
  - There is a weak positive relationship between precipitation and formaldehyde emission.

5.3.Limitations and Uncertainties

- The dataset had a lot of uncertainties which included null values, negative values, and outliers. They had to be preprocessed using preprocessing techniques to ensure the reliability and accuracy of the data. This could have occurred due to instrument or human entry error.
- Data could have been collected for a greater number of years to obtain more valuable visual insights and develop more efficient machine learning models with a more optimal performance.
- Data could have been obtained for a greater number of cities within the country to analyze the trends for the whole country.

5.4.Policy Making and Research

Analysis done could help in providing valuable insights into the standards of air quality in different locations in Sri Lanka. Policymakers and environment protection agencies could use this information to revise and remake regulations aimed at reducing emissions of Formaldehyde. This includes various sources of emission such as complete and incomplete combustion of Volatile Organic Compounds (VOCs), and industrial activities. Implementation of sterns regulations could help in mitigating the adverse health concerns and environmental consequences.

Emission of Formaldehyde into the atmosphere could be catastrophic on a health and well-being point of view. Formaldehyde has been classified as a human carcinogen, which means it could increase the risk of cancer occurrence. It could also cause respiratory concerns and eye irritations as side effects of the emission. Thus, using these findings, stakeholders in the medical industry could plan measures to increase the awareness of the public on these concerns. Use of technologies and methods which emit low levels of Volatile Organic Compounds (VOCs).

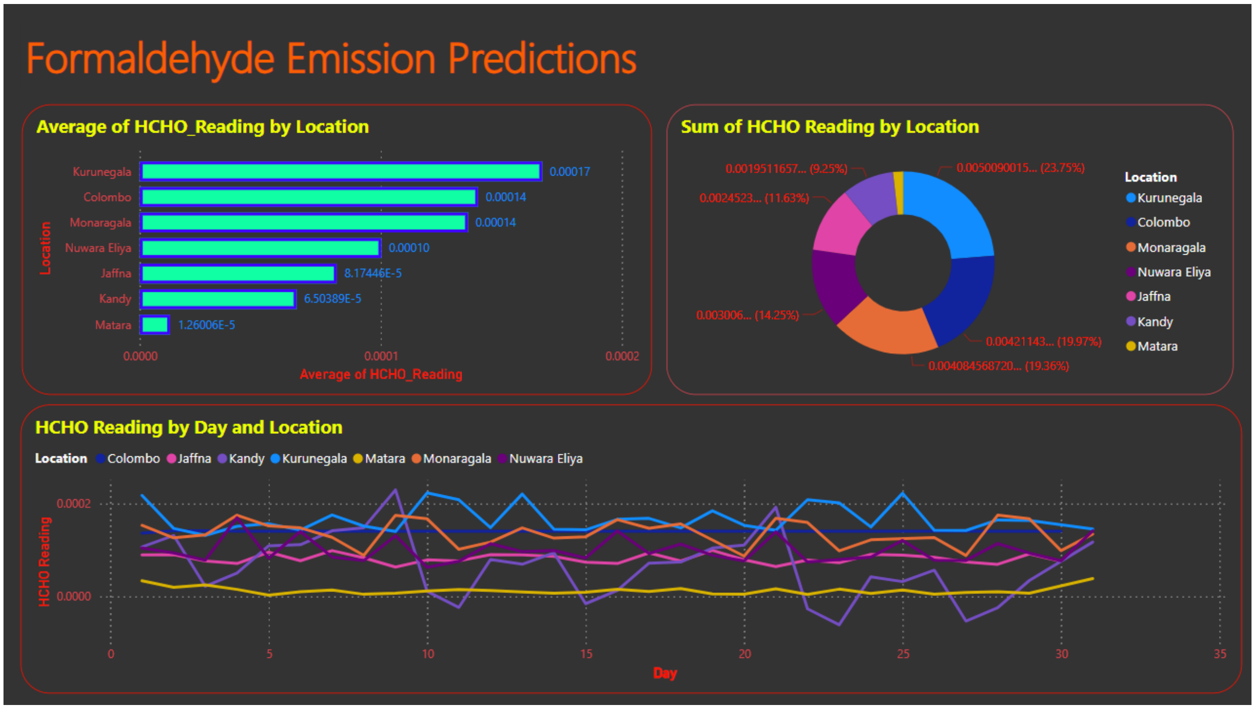
Since a detailed spatial temporal analysis has been conducted on different external factors and how they affect the HCHO emission in different cities, the results of this analysis could be used by individuals involved in urban planning and development. Information could be used in land planning and traffic management while regulating the mining activities. This will help to improve the levels of emissions in the city and in Sri Lanka as a whole.

This report serves academic purposes as well as it could be used by students and individuals involved in academia to research on the topic of Formaldehyde emissions, thereby improving the research gap.

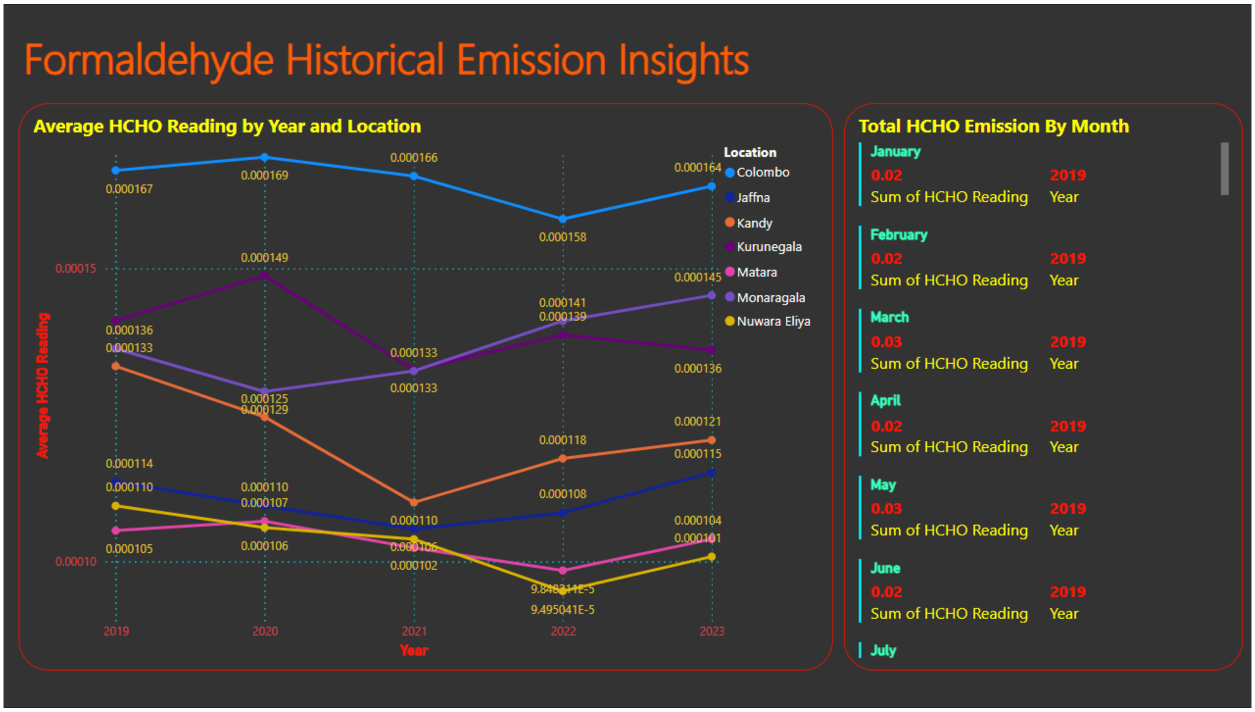
6. References

1. Vizzuality (n.d.). *Sri Lanka Deforestation Rates & Statistics | GFW*. [online] www.globalforestwatch.org. Available at: https://www.globalforestwatch.org/dashboards/country/LKA/?category=fires.
2. ECONOMIC STATISTICS OF SRI LANKA. (2023). Available at: http://www.statistics.gov.lk/Publication/Economic-Statistics-2023.
3. www.ncdc.noaa.gov. (n.d.). *Search | Climate Data Online (CDO) | National Climatic Data Center (NCDC)*. [online] Available at: https://www.ncdc.noaa.gov/cdo-web/search?datasetid=GHCND.
4. GOYAL, C. (2021). *How to Detect and Remove Outliers | Outlier Detection and Removal*. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/05/feature-engineering-how-to-detect-and-remove-outliers-with-python-code/.
5. MERITSHOT, 2023. *Introduction to ARIMA and SARIMA for time series forecasting*. [online]. Medium. Available from: https://medium.com/@meritshot/introduction-to-arima-and-sarima-for-time-series-forecasting-5af5025c8876.
6. THE POWER BI GUY, 2022. *How to build a STUNNING sales dashboard in Power BI - tutorial #2022*. [online]. Youtube. Available from: https://www.youtube.com/watch?v=SGyMYG6C91M.
7. 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.
8. BISWAS, M.S. and SINGH, M., 2024. *Trustworthy modelling of atmospheric formaldehyde powered by deep learning.* [online]. Arxiv.org. Available from: http://arxiv.org/abs/2209.07414.

7. Appendix



Formaldehyde Emission Predictions



Formaldehyde Historical Emission Insights

Total HCHO Emission By Month

January
0.02 2019
Sum of HCHO Reading Year

February
0.02 2019
Sum of HCHO Reading Year

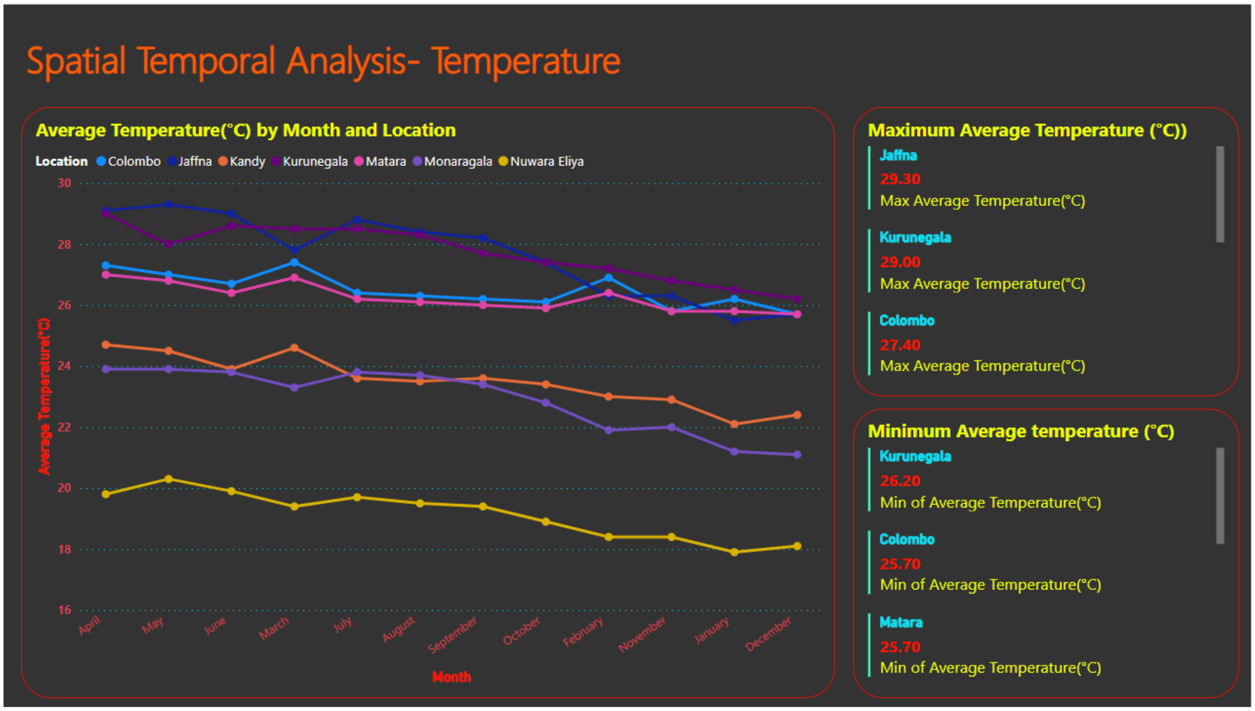
March
0.03 2019
Sum of HCHO Reading Year

April
0.02 2019
Sum of HCHO Reading Year

May
0.03 2019
Sum of HCHO Reading Year

June
0.02 2019
Sum of HCHO Reading Year

July



Spatial Temporal Analysis- Temperature

Average Temperature(°C) by Month and Location
Location ● Colombo ● Jaffna ● Kandy ● Kurunegala ● Matara ● Monaragala ● Nuwara Eliya
30
28
26
24
22
20
18
16
Average Temperature(°C)
April
May
June
March
July
August
September
October
February
November
January
December
Month

Maximum Average Temperature (°C))

Jaffna
29.30
Max Average Temperature(°C)

Kurunegala
29.00
Max Average Temperature(°C)

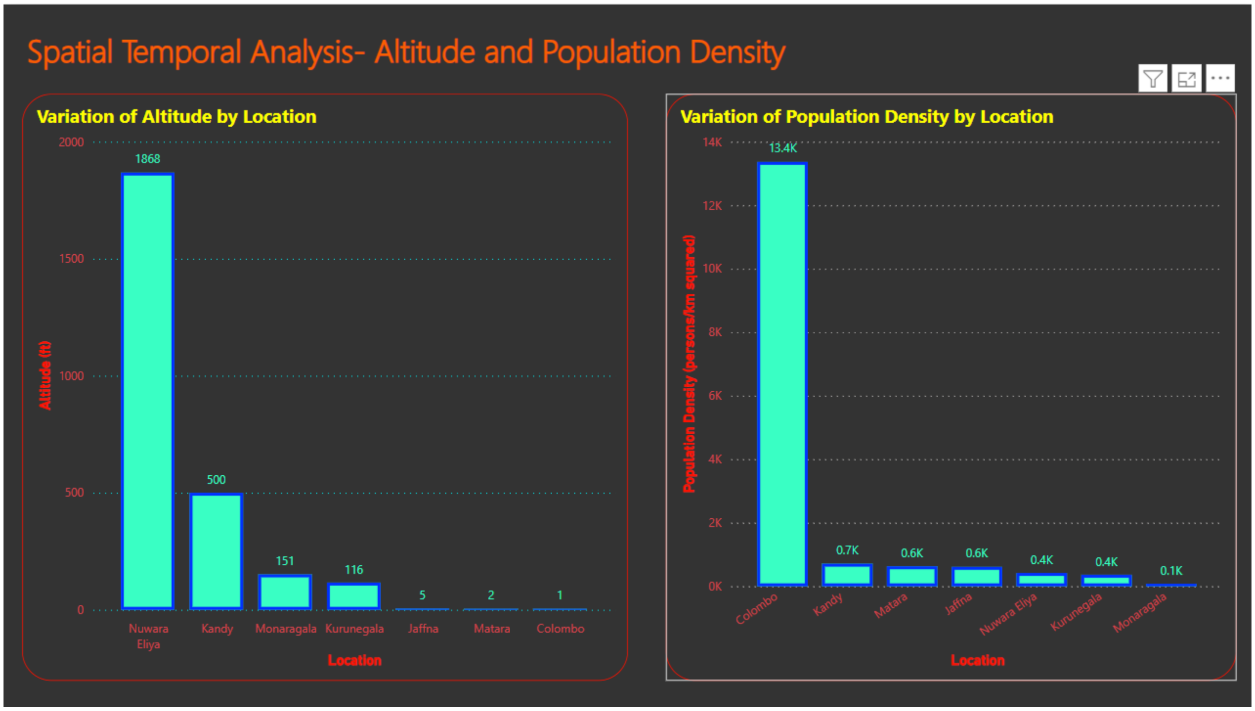
Colombo
27.40
Max Average Temperature(°C)

Minimum Average temperature (°C)

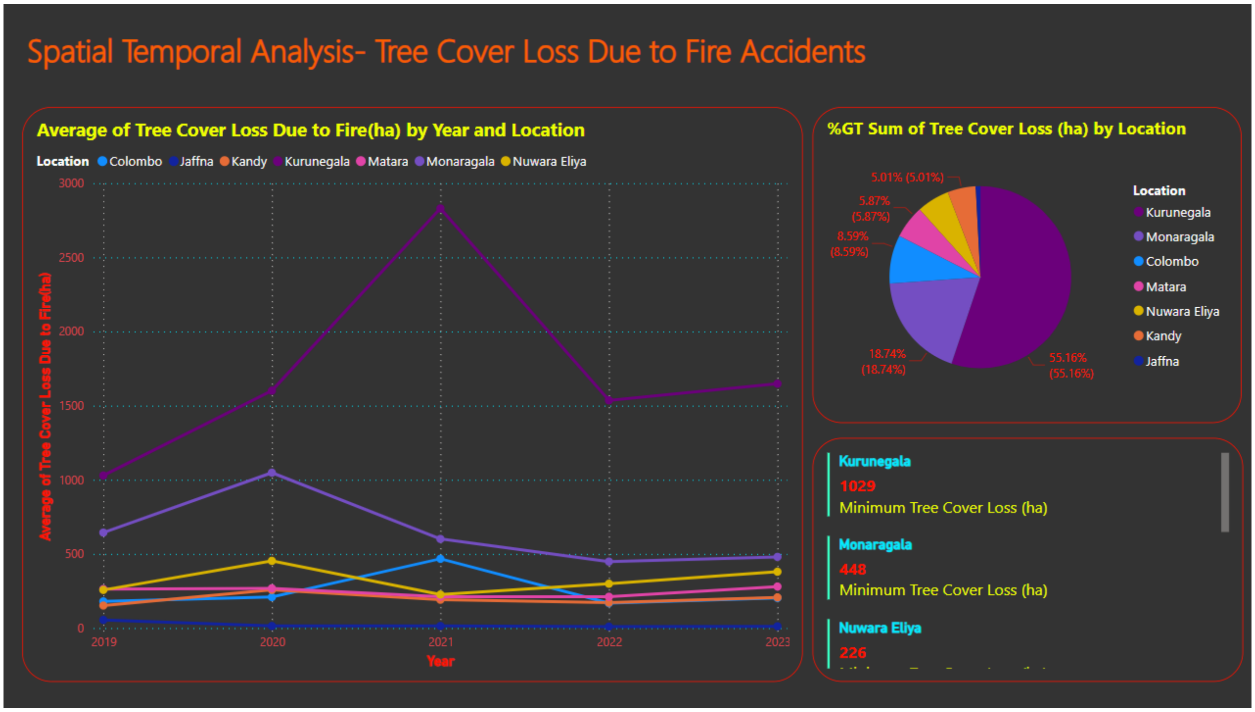
Kurunegala
26.20
Min of Average Temperature(°C)

Colombo
25.70
Min of Average Temperature(°C)

Matara
25.70
Min of Average Temperature(°C)



Spatial Temporal Analysis- Altitude and Population Density



Spatial Temporal Analysis- Tree Cover Loss Due to Fire Accidents

Average of Tree Cover Loss Due to Fire(ha) by Year and Location

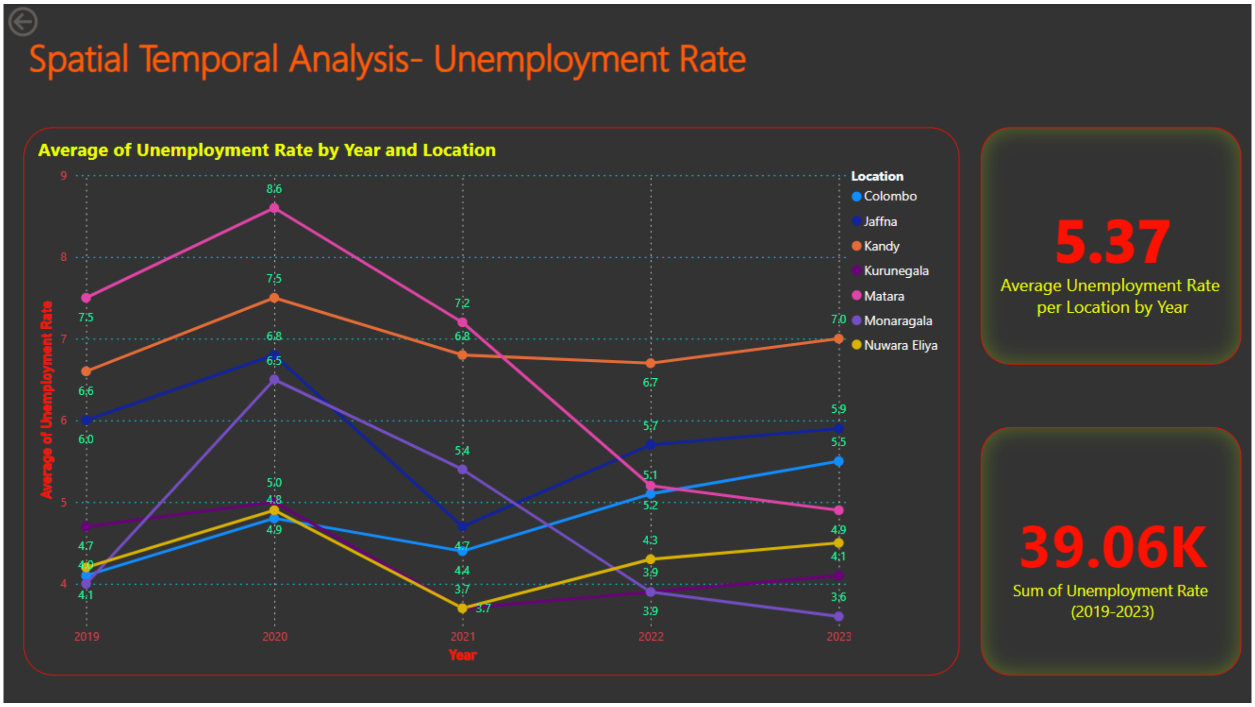
Location ● Colombo ● Jaffna ● Kandy ● Kurunegala ● Matara ● Monaragala ● Nuwara Eliya
3000
2500
2000
1500
1000
500
0
Average of Tree Cover Loss Due to Fire(ha)
2019
2020
2021
2022
2023
Year

%GT Sum of Tree Cover Loss (ha) by Location

Kurunegala
1029
Minimum Tree Cover Loss (ha)

Monaragala
448
Minimum Tree Cover Loss (ha)

Nuwara Eliya
226



Spatial Temporal Analysis- Unemployment Rate

Average of Unemployment Rate by Year and Location

5.37
Average Unemployment Rate per Location by Year

39.06K
Sum of Unemployment Rate (2019-2023)