BSc (Hons) Artificial Intelligence and DataScience

CM2606 Data Engineering

Individual Coursework

Level 05

Module	Data Engineering
Module Code	CM2606
Stage	Year 02 (2nd Semester)
Name	B.G.C. GOMES
IIT ID	20220578
RGU ID	2313082

Table of Contents

List of Figures

Introduction

As part of this coursework in Data Engineering, I explore the complex field of formaldehyde (HCHO) data processing using observations made by the Sentinel-5P satellite in seven major Sri Lankan cities. The dataset provides a wealth of information for deciphering the complex temporal and spatial patterns of HCHO levels, spanning an extensive time period from 2019 to 2023. To ensure the integrity and dependability of ensuing studies, I want to solve issues like missing values, outliers, and format inconsistencies by carefully investigating data pretreatment strategies. In addition, the research broadens its scope to include spatiotemporal analysis, which makes it easier to identify seasonal fluctuations, long-term patterns, and the possible effects of noteworthy occurrences like the COVID-19 pandemic. I hope to clarify the complex interactions between many environmental variables and their combined impact on air quality dynamics by establishing a correlation between HCHO levels and external factors including weather, fire incidents, and human activity.

All things considered, this course on data engineering is an attempt to use data-driven insights to solve urgent environmental problems. Through exploring the complex subtleties of HCHO data analysis, I hope to make a significant contribution to the body of knowledge about the dynamics of air quality in Sri Lanka. My goal is to offer stakeholders useful information that may guide policy decisions, propel advocacy campaigns, and eventually promote a more sustainable and healthful future for everybody. I will accomplish this by thoroughly exploring, analyzing, and interpreting facts collected from satellites.

Dataset Clarification

Data Description

This dataset includes formaldehyde (HCHO) levels measured every day between January 1, 2019, and December 31, 2023, based on Sentinel-5P satellite observations. "Colombo Proper," "Deniyaya, Matara," "Nuwara Eliya Proper," "Bibile, Monaragala," "Kurunegala Proper," "Jaffna Proper," and "Kandy Proper" are the seven cities in Sri Lanka that form its scope. The properties of each instance are as follows: 'Location' (city), 'Current Date', 'Next Date', and 'HCHO reading' (tropospheric HCHO column number density). Through the data, regional and temporal patterns of HCHO will be easier to analyze, supporting the development of environmental and public health policies as well as the knowledge of the dynamics of air quality.

There are three datasets for represent the HCHO Distribution values. All three datasets have 12,782 records and 4 attributes for data distribution (‘HCHO reading’ , ‘Location’ , ‘Current Date’ , ‘Next Date’)

01. col_mat_nuw_output

This data set contains HCHO reading values of

- Colombo Proper
- Deniyaya, Matara
- Nuwara Eliya Proper

02. mon_kur_jaf_output

This data set contains HCHO reading values of

- Bibile, Monaragala
- Kurunagala Proper
- Jaffna Proper

03. kan_output

This data set contains HCHO reading values of

- Kandy Proper

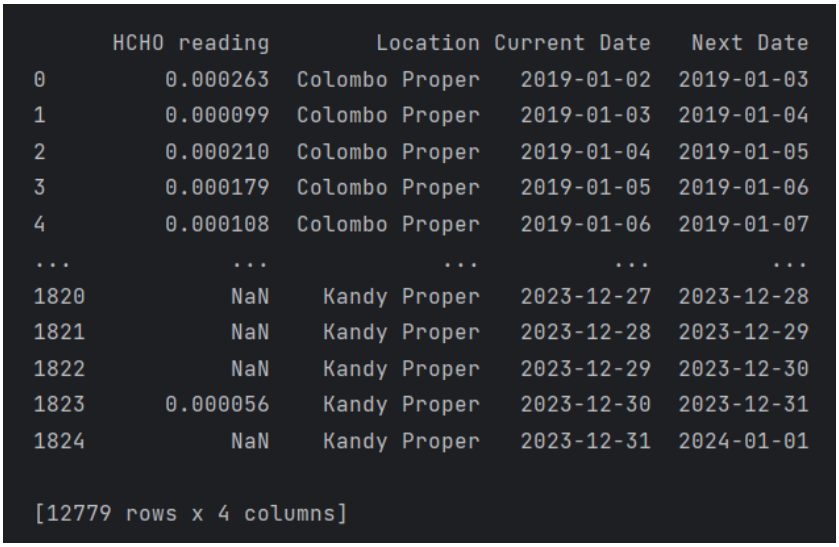
Source of the Dataset	**Sentinel-5P satellite observations**
Number of instances	12779
Number of attributes	04
Missing values	Yes (4863)
Number of classes	01 (HCHO reading value)
Outliers	Yes (2725)
Duplicate Values	No
Attribute Types	Float, Object, Datetime

Data Preprocessing Steps

Preprocessing Step	Description
Handling Missing Values	- Missing values in numeric columns were imputed using forward fill (ffill) and backward fill (bfill) methods. - For categorical columns, missing values were handled based on domain knowledge or imputed using mode.
Removing Duplicates	- There are no any duplicate values in the data frame
Handle Outliers	- Outliers in numeric columns were detected using the interquartile range (IQR) method and winsorized. - Winsorization involved replacing outlier values with the upper or lower bound of the IQR range.
Dataset Size Before Cleaning	- 12779 rows
Final Dataset Size After Cleaning	- 12779 rows

Handling Missing Values

Data frame before handling missing values



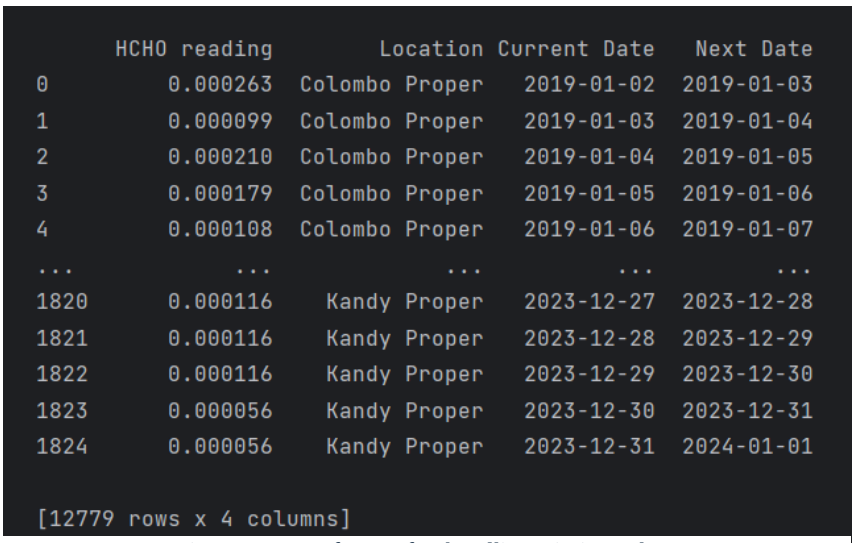
	HCHO reading	Location	Current Date	Next Date
0	0.000263	Colombo Proper	2019-01-02	2019-01-03
1	0.000099	Colombo Proper	2019-01-03	2019-01-04
2	0.000210	Colombo Proper	2019-01-04	2019-01-05
3	0.000179	Colombo Proper	2019-01-05	2019-01-06
4	0.000108	Colombo Proper	2019-01-06	2019-01-07
...	...	...	...	...
1820	NaN	Kandy Proper	2023-12-27	2023-12-28
1821	NaN	Kandy Proper	2023-12-28	2023-12-29
1822	NaN	Kandy Proper	2023-12-29	2023-12-30
1823	0.000056	Kandy Proper	2023-12-30	2023-12-31
1824	NaN	Kandy Proper	2023-12-31	2024-01-01

[12779 rows x 4 columns]

Figure 1: Data frame before handling missing values.

Data frame after handling missing values.

Missing values in numeric columns were imputed using forward fill (ffill) and backward fill (bfill) methods. For categorical columns, missing values were handled based on domain knowledge or imputed using model.



	HCHO reading	Location	Current Date	Next Date
0	0.000263	Colombo Proper	2019-01-02	2019-01-03
1	0.000099	Colombo Proper	2019-01-03	2019-01-04
2	0.000210	Colombo Proper	2019-01-04	2019-01-05
3	0.000179	Colombo Proper	2019-01-05	2019-01-06
4	0.000108	Colombo Proper	2019-01-06	2019-01-07
...	...	...	...	...
1820	0.000116	Kandy Proper	2023-12-27	2023-12-28
1821	0.000116	Kandy Proper	2023-12-28	2023-12-29
1822	0.000116	Kandy Proper	2023-12-29	2023-12-30
1823	0.000056	Kandy Proper	2023-12-30	2023-12-31
1824	0.000056	Kandy Proper	2023-12-31	2024-01-01

[12779 rows x 4 columns]

Figure 2 : Data frame after handling missing values.

Removing Duplicates

There are no any duplicate values in the data frame

```
# Display the count of rows before removing duplicates
print("Count of duplicate rows in the data frame: ",concatenated_data.duplicated().sum())
Executed at 2024.04.20 04:47:57 in 129ms

 Count of duplicate rows in the data frame:  0
```

Figure 3 :Removing Duplicates

Handle Outliers

Outliers in numeric columns were detected using the interquartile range (IQR) method and winsorized. (before removing outliers)

```
Rows with outlier values:
      HCHO reading        Location Current Date   Next Date
5         0.000393  Colombo Proper   2019-01-07  2019-01-08
17        0.000406  Colombo Proper   2019-01-19  2019-01-20
19        0.000388  Colombo Proper   2019-01-21  2019-01-22
30        0.000354  Colombo Proper   2019-02-01  2019-02-02
34        0.000362  Colombo Proper   2019-02-05  2019-02-06
...            ...             ...          ...         ...
1753     -0.000300    Kandy Proper   2023-10-21  2023-10-22
1754     -0.000300    Kandy Proper   2023-10-22  2023-10-23
1755     -0.000300    Kandy Proper   2023-10-23  2023-10-24
1756     -0.000300    Kandy Proper   2023-10-24  2023-10-25
1757     -0.000300    Kandy Proper   2023-10-25  2023-10-26

[331 rows x 4 columns]
```

Figure 4 : Before removing outliers

Winsorization involved replacing outlier values with the upper or lower bound of the IQR range. (After removing outliers)

```
Concatenated dataset after handling outliers (Winsorized):
      HCHO reading        Location Current Date   Next Date
0         0.000263  Colombo Proper   2019-01-02  2019-01-03
1         0.000099  Colombo Proper   2019-01-03  2019-01-04
2         0.000210  Colombo Proper   2019-01-04  2019-01-05
3         0.000179  Colombo Proper   2019-01-05  2019-01-06
4         0.000108  Colombo Proper   2019-01-06  2019-01-07
...            ...             ...          ...         ...
1820      0.000116    Kandy Proper   2023-12-27  2023-12-28
1821      0.000116    Kandy Proper   2023-12-28  2023-12-29
1822      0.000116    Kandy Proper   2023-12-29  2023-12-30
1823      0.000056    Kandy Proper   2023-12-30  2023-12-31
1824      0.000056    Kandy Proper   2023-12-31  2024-01-01

[12779 rows x 4 columns]
```

Figure 5: After removing outliers.

Plot the outliers.

Figure 6 : Plot the outliers.

Exploratory Data Analysis

Summary Statistics

The summary statistics section provides a comprehensive overview of key descriptive measures such as mean, median, and standard deviation of formaldehyde (HCHO) levels. These statistics are presented across various locations and time periods, offering insights into the distribution and variability of HCHO concentrations in the dataset.

- Descriptive statistics for each city

```
Descriptive statistics for each city:
                          mean    median       std
Location
Bibile, Monaragala    0.000120  0.000120  0.000082
Colombo Proper        0.000148  0.000145  0.000082
Deniyaya, Matara      0.000093  0.000082  0.000079
Jaffna Proper         0.000106  0.000099  0.000068
Kandy Proper          0.000101  0.000100  0.000079
Kurunegala Proper     0.000124  0.000118  0.000078
Nuwara Eliya Proper   0.000093  0.000083  0.000078
```

Figure 7: Descriptive statistics for each city

- Descriptive statistics for the entire data frame

```
Descriptive statistics for the entire dataset:
mean      0.000112
median    0.000107
std       0.000080
Name: HCHO reading, dtype: float64
```

Figure 8 : Descriptive statistics for the entire data frame

Visualization of HCHO Distributions

HCHO distribution of each city

01. Distribution of HCHO levels in Colombo Proper (Histogram and Boxplot)

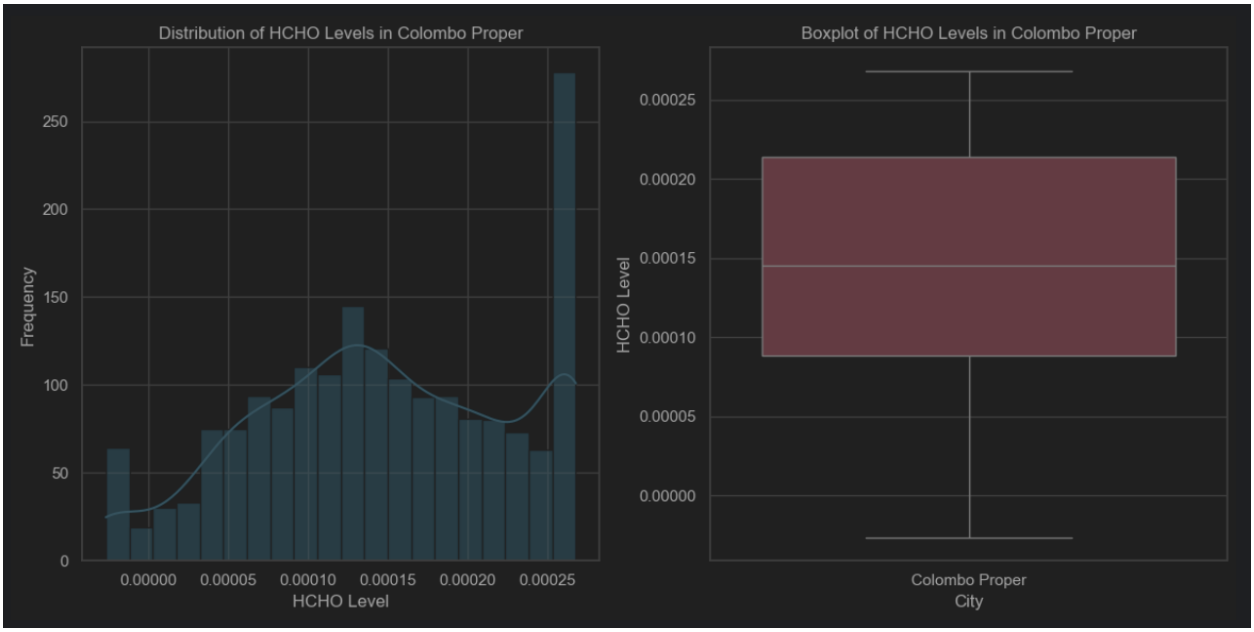


Figure 9: Distribution of HCHO levels in Colombo Proper

02. Distribution of HCHO levels in Deniyaya, Matara (Histogram and Boxplot)

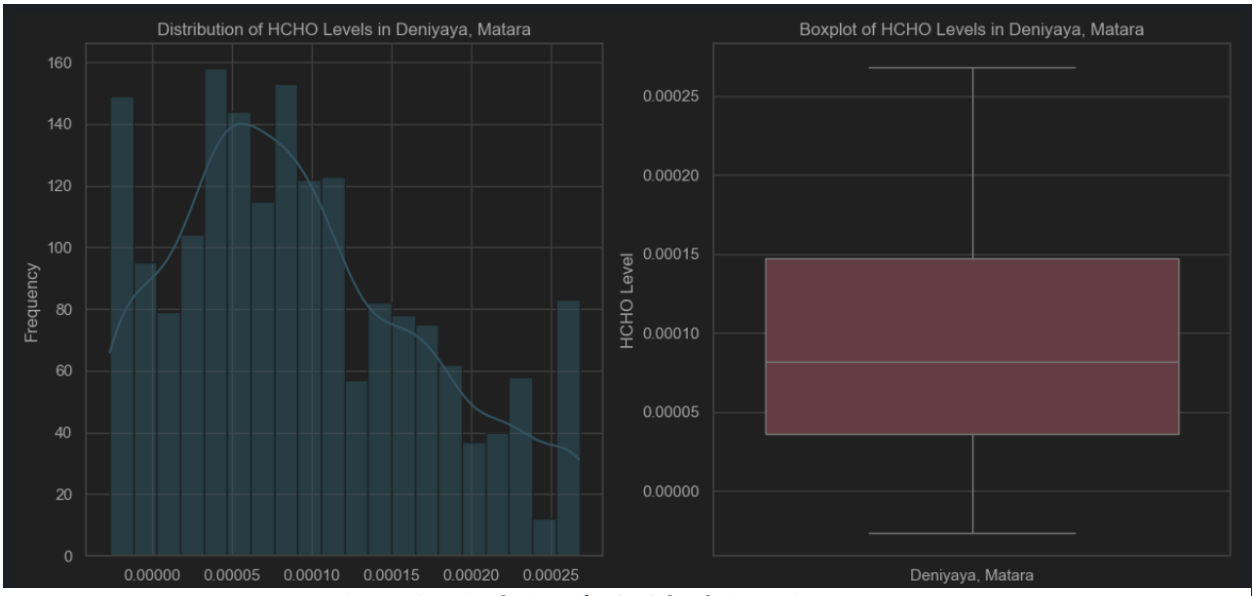


Figure 10 : Distribution of HCHO levels in Deniyaya, Matara

03. Distribution of HCHO levels in Nuwara Eliya Proper (Histogram and Boxplot)

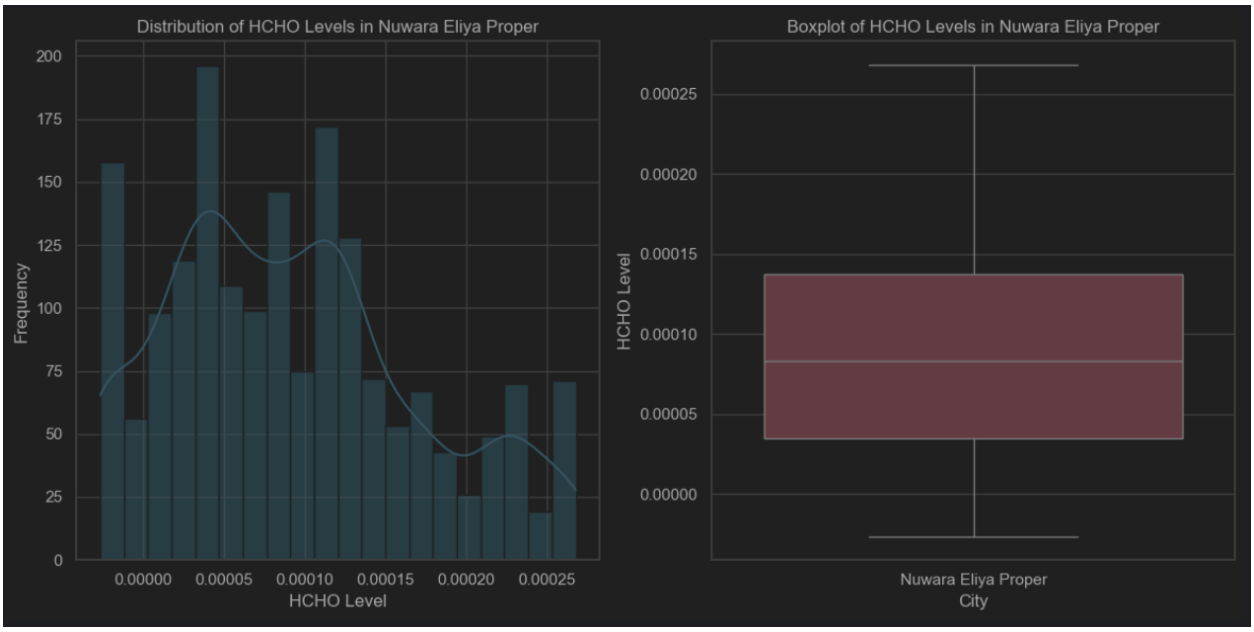


Figure 11 : Distribution of HCHO levels in Nuwara Eliya Proper

04. Distribution of HCHO levels in Bibile, Monaragala (Histogram and Boxplot)

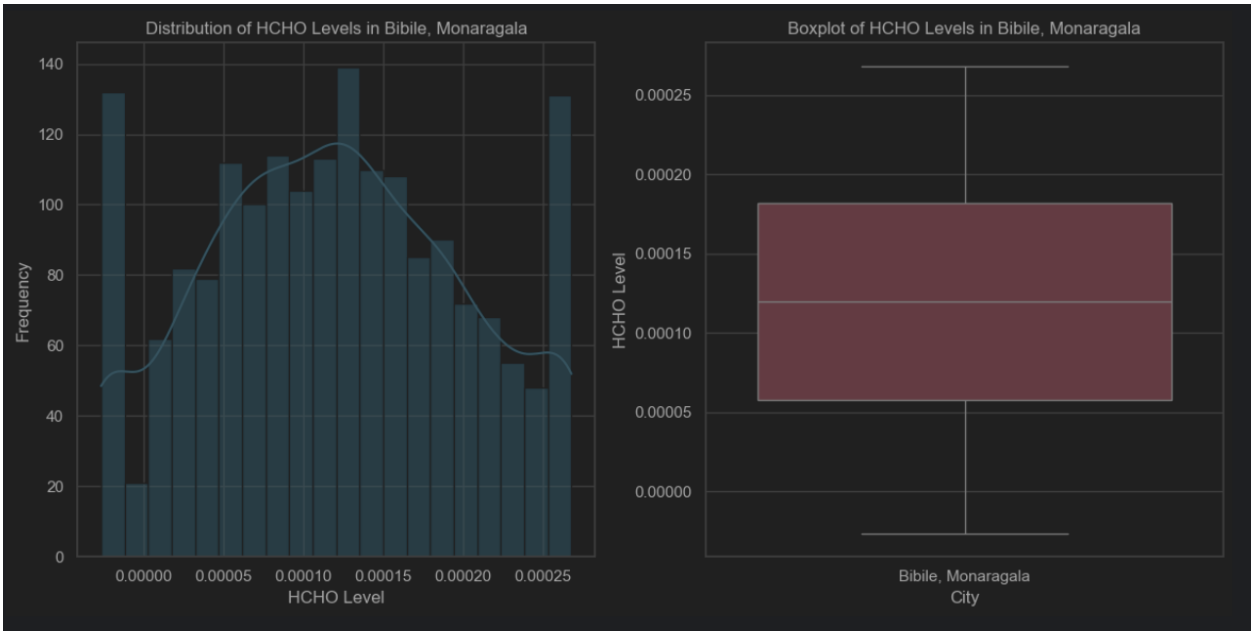


Figure 12 : Distribution of HCHO levels in Bibile, Monaragala

05. Distribution of HCHO levels in Kurunagala Proper (Histogram and Boxplot)

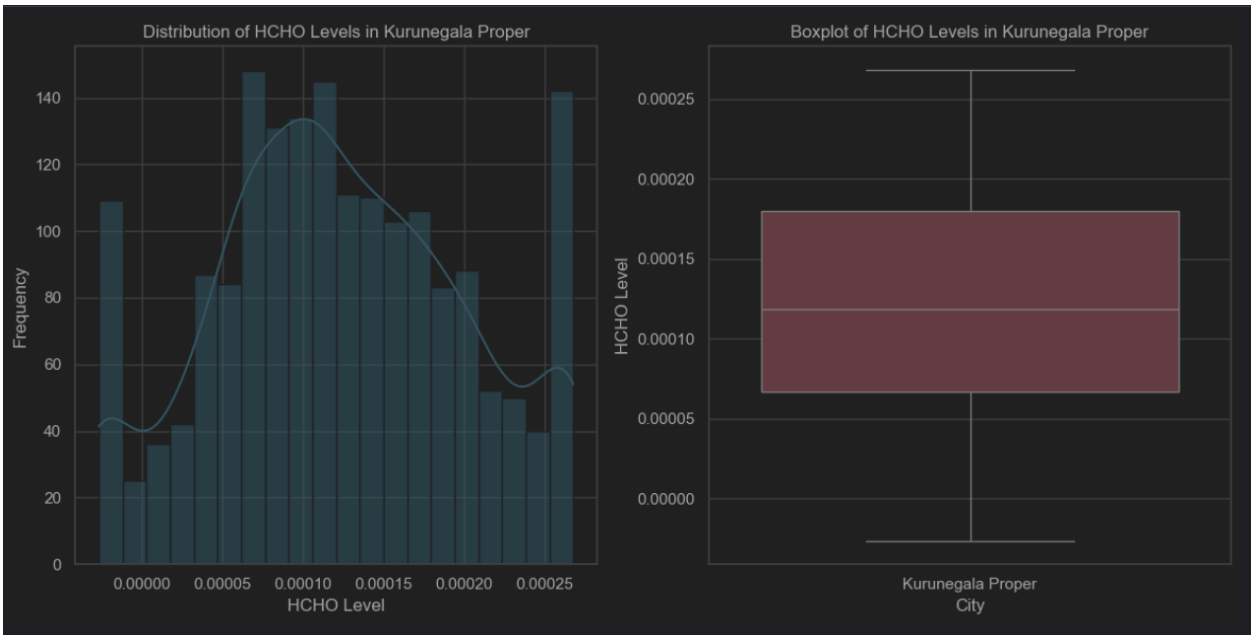


Figure 13 : Distribution of HCHO levels in Kurunagala Proper

06. Distribution of HCHO levels in Jaffna Proper (Histogram and Boxplot)

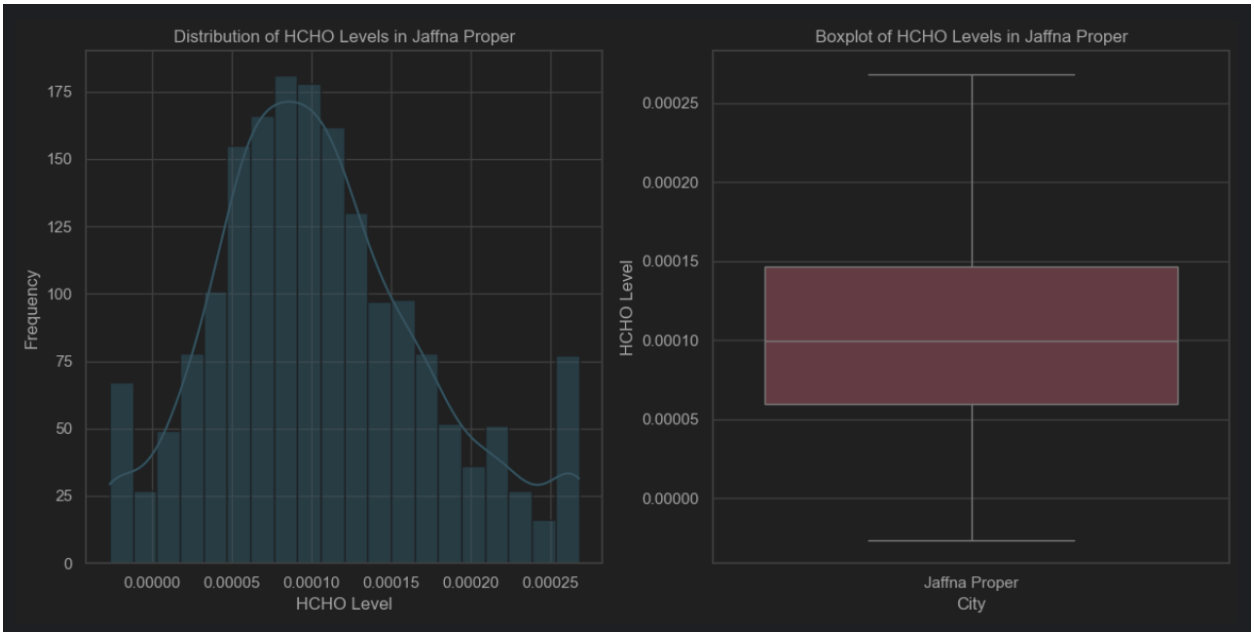


Figure 14 : Distribution of HCHO levels in Jaffna Proper

07. Distribution of HCHO levels in Kandy Proper (Histogram and Boxplot)

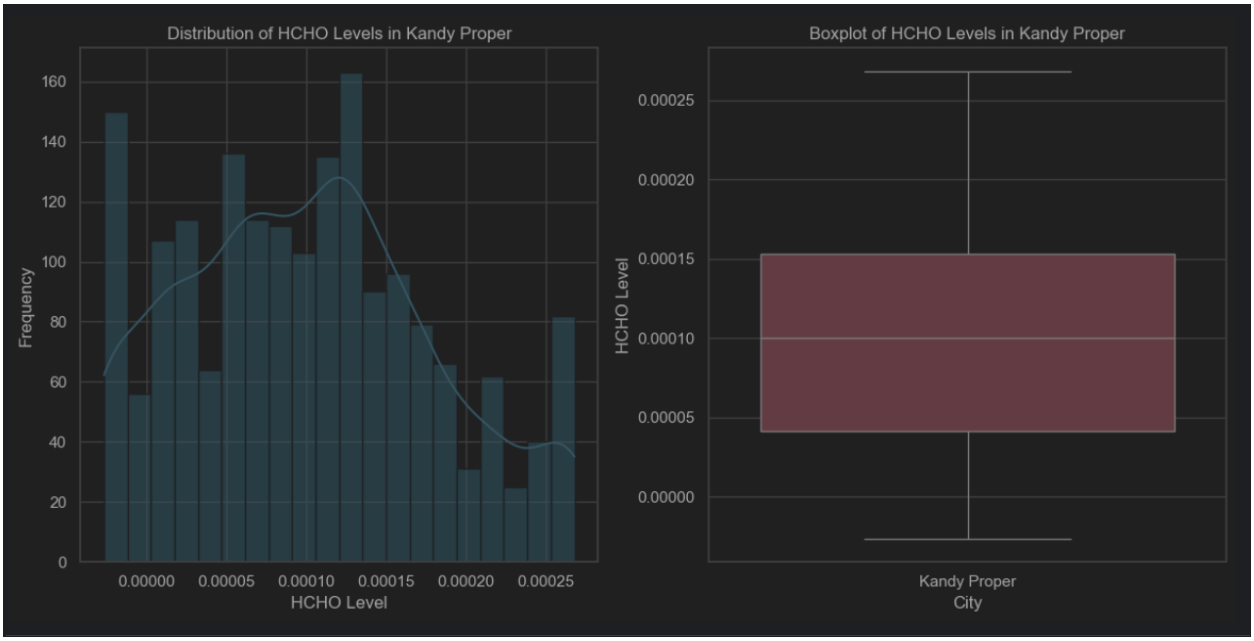


Figure 15 : Distribution of HCHO levels in Kandy Proper

08. Overall distribution of HCHO levels for all cities (Histogram and Boxplot)

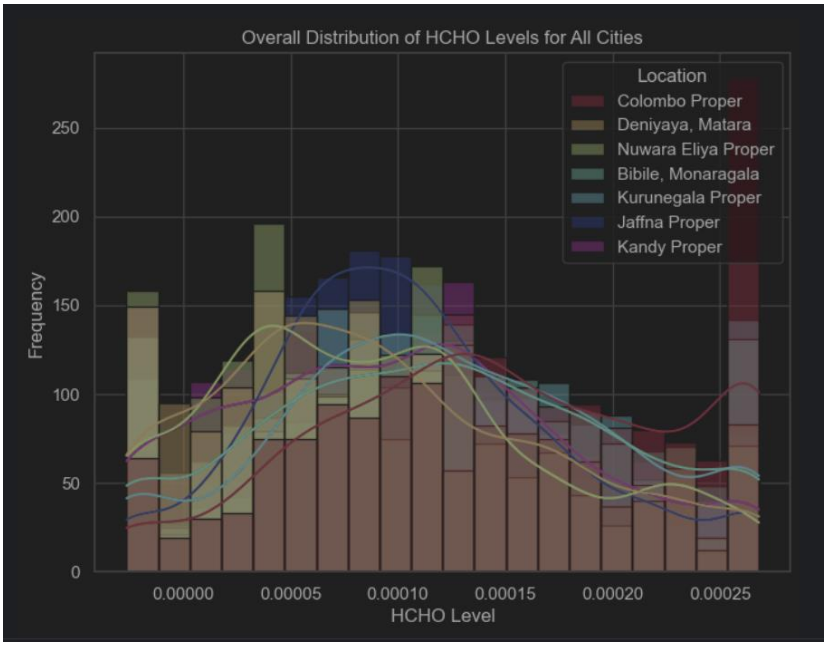


Figure 16 : Overall distribution of HCHO levels for all cities (Histogram)

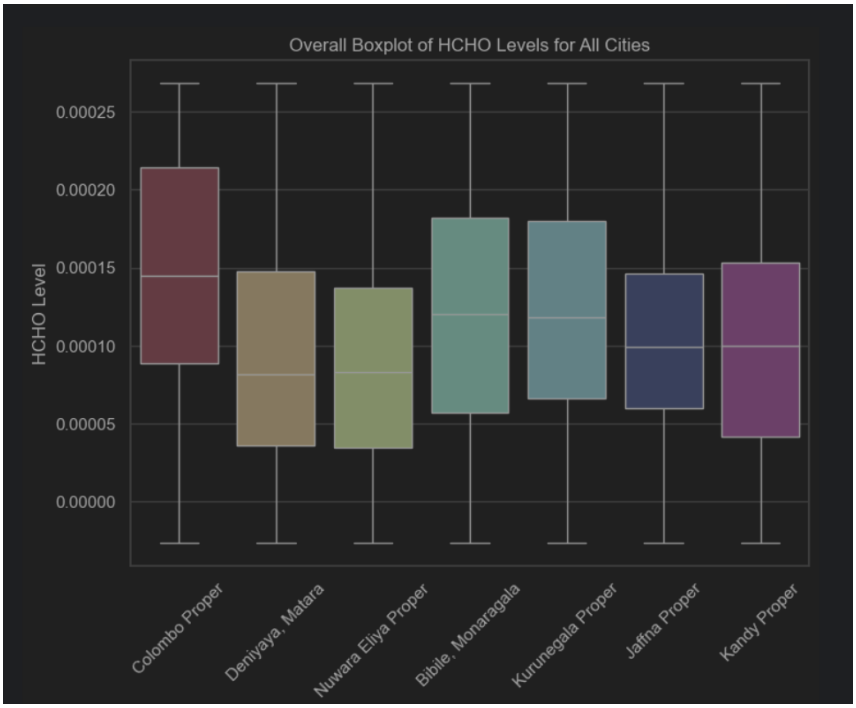


Figure 17 : Overall distribution of HCHO levels for all cities (Boxplot)

Seasonal Variations

This section delves into the seasonal patterns of formaldehyde (HCHO) levels, examining fluctuations across different seasons such as spring, summer, monsoon, and autumn. By comparing HCHO concentrations between seasons, we aim to identify any significant trends or variations influenced by seasonal factors.

*Distribution of HCHO levels in **Spring***

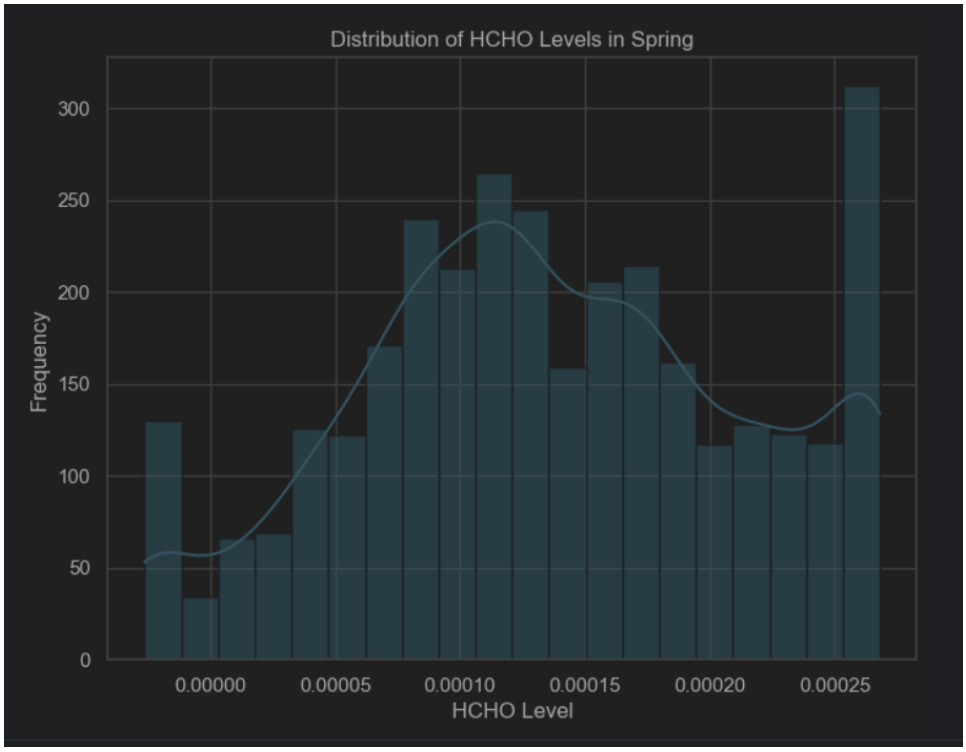


Figure 18 : Distribution of HCHO levels in Spring

*Distribution of HCHO levels in **Summer***

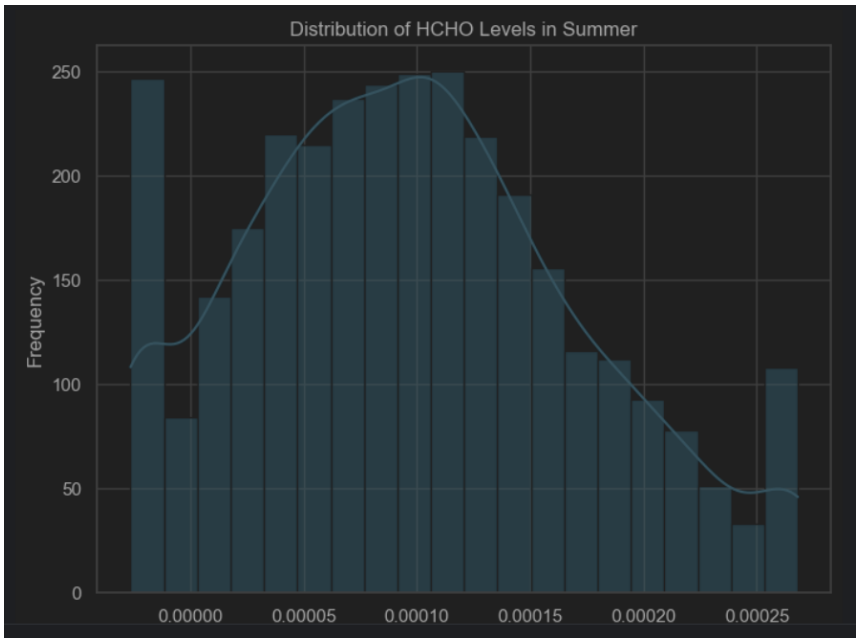


Figure 19 : Distribution of HCHO levels in Summer

*Distribution of HCHO levels in **Monsoon***

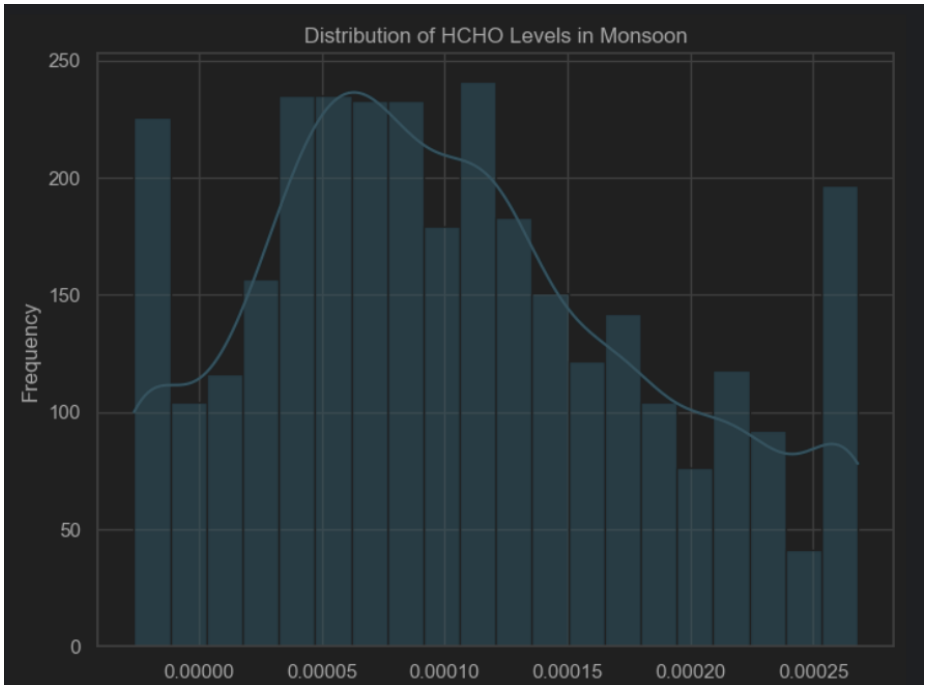


Figure 20 : Distribution of HCHO levels in Monsoon

*Distribution of HCHO levels in **Monsoon***

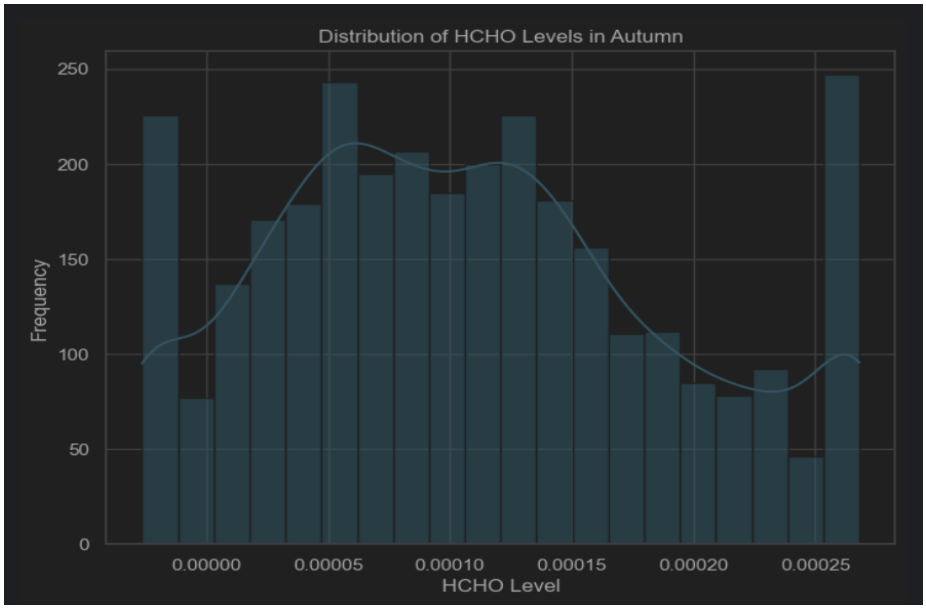
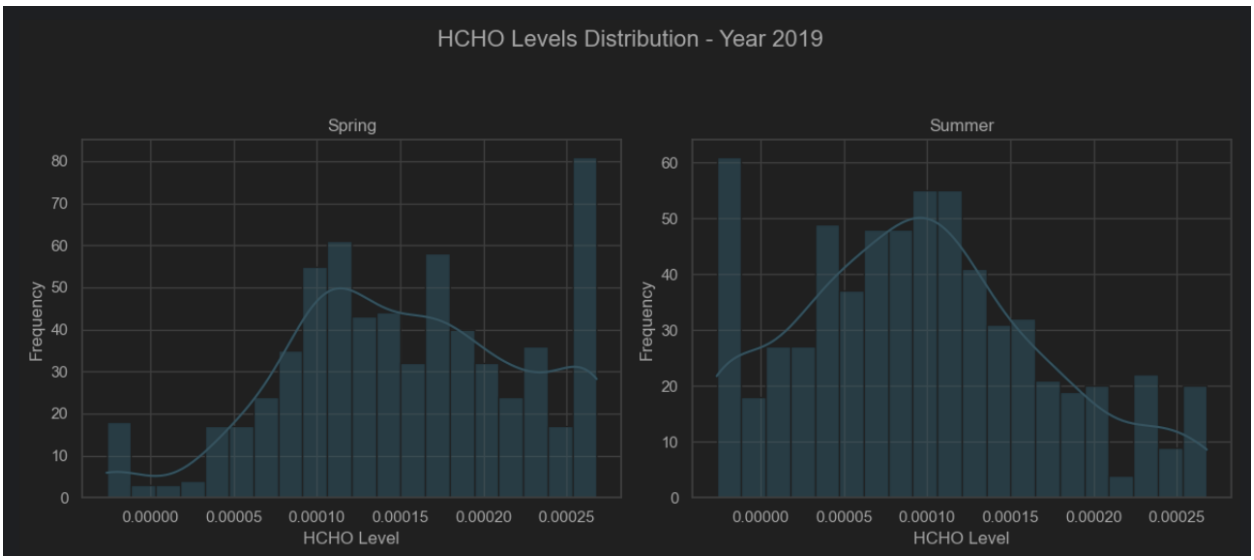


Figure 21: Distribution of HCHO levels in Monsoon

Distribution of HCHO Levels Across Years

2019



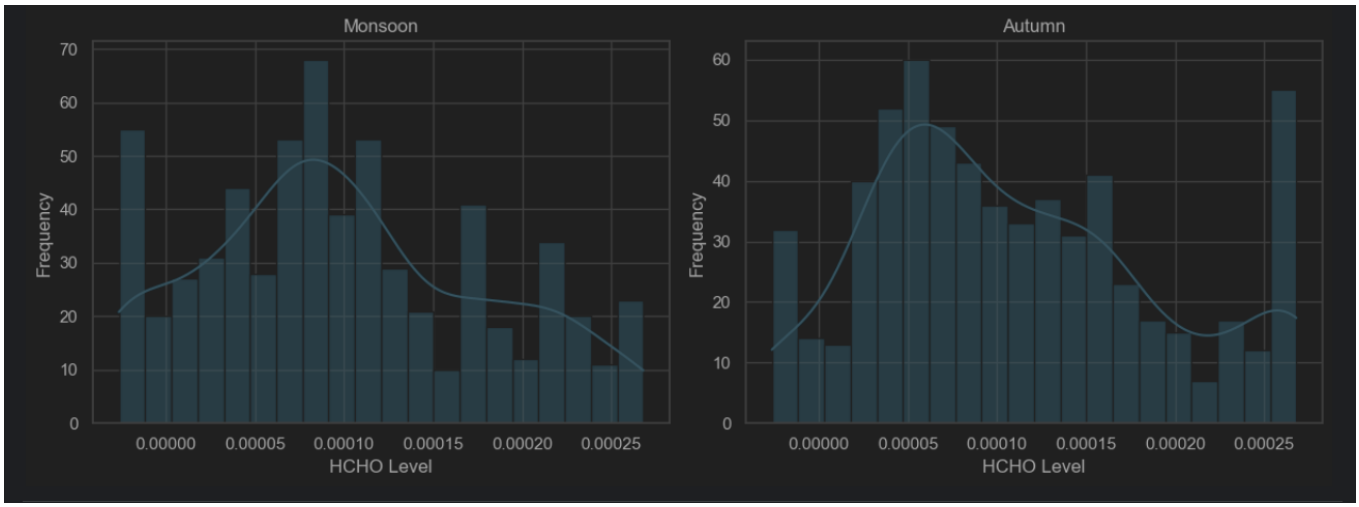


Figure 22 : Distribution of HCHO Levels of each city in 2019

2020

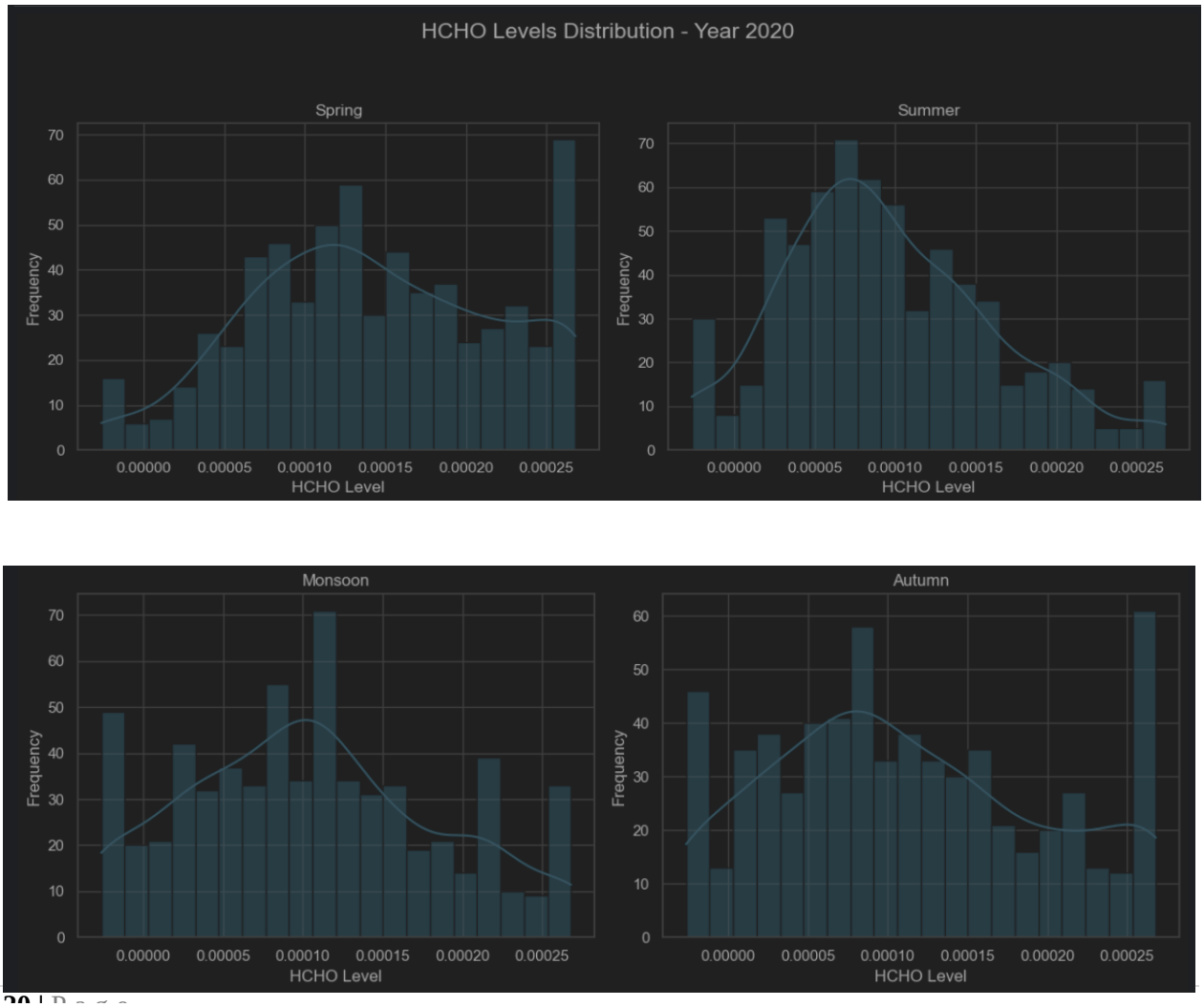


Figure 23: Distribution of HCHO Levels of each city in 2020

2021

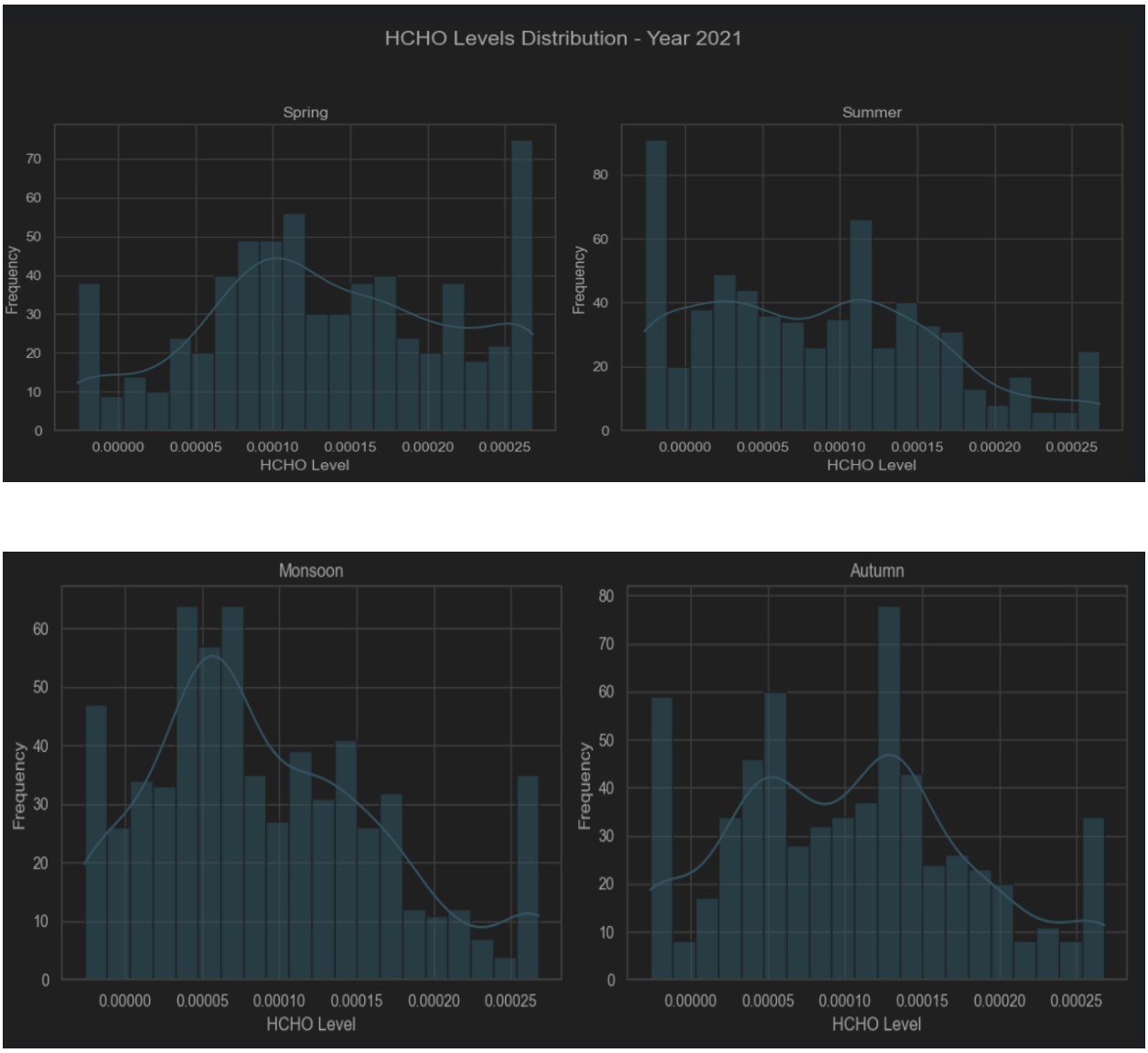


Figure 24: Distribution of HCHO Levels of each city in 2021

2022

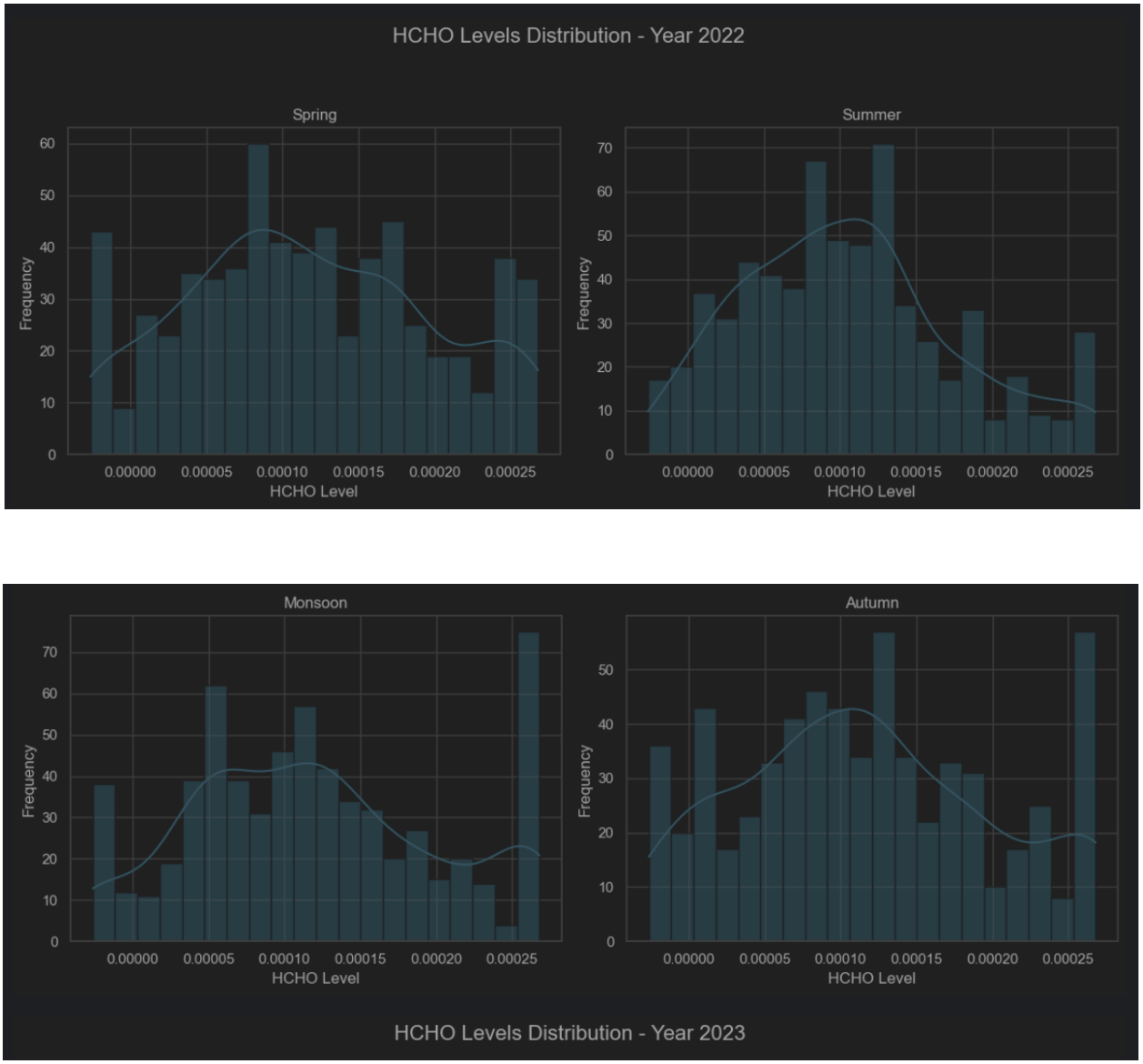


Figure 25: Distribution of HCHO Levels of each city in 2022

2023

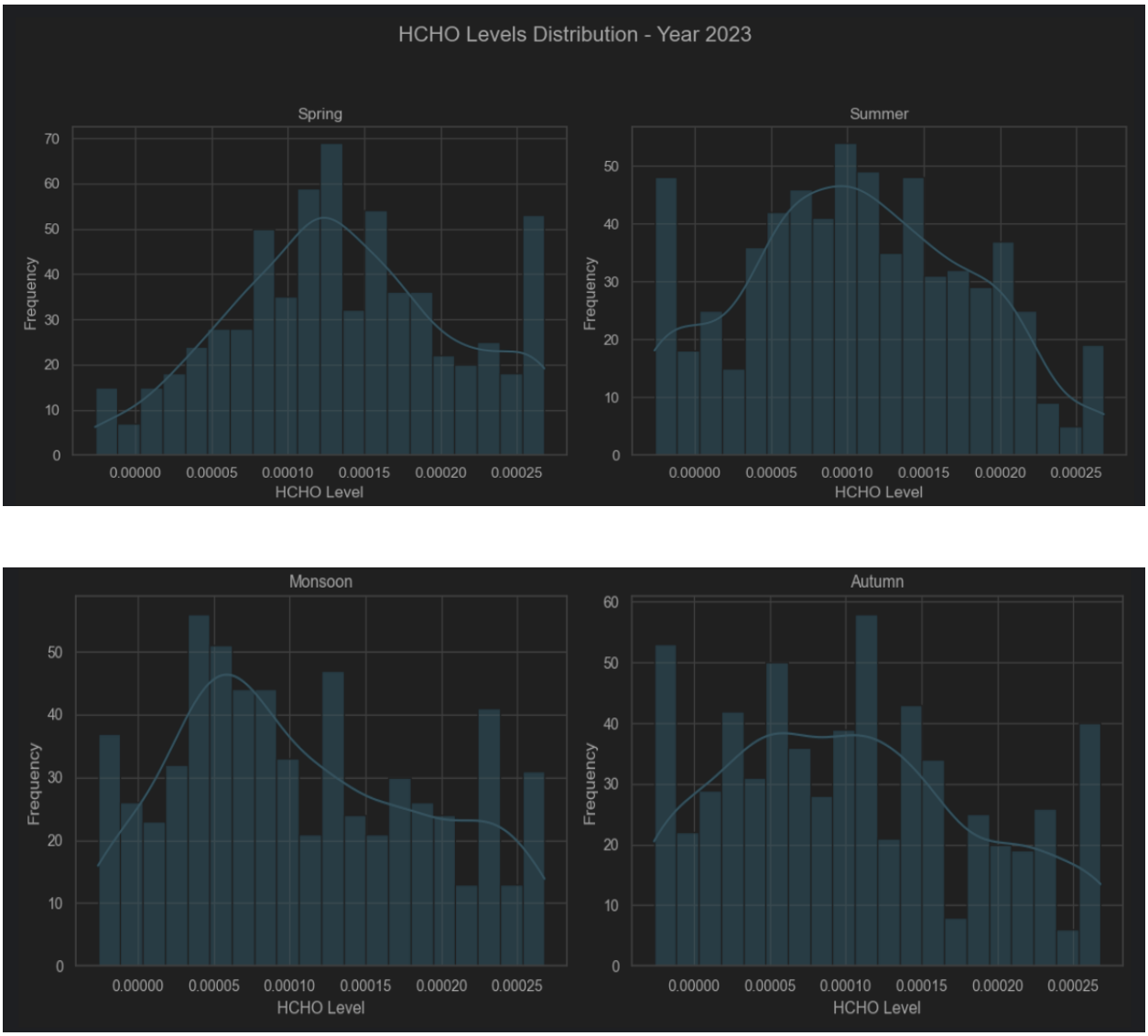
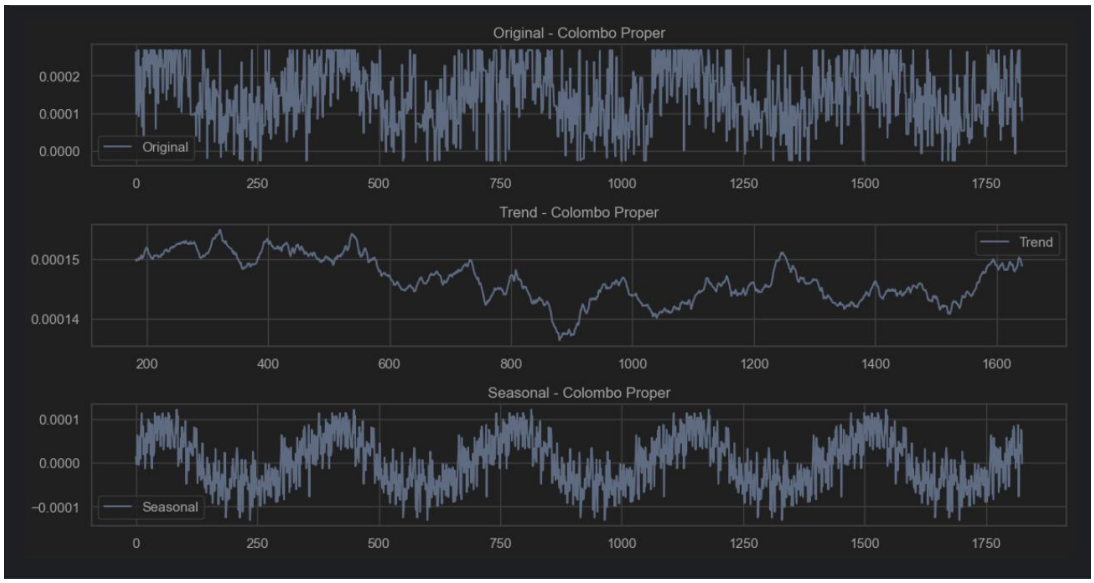


Figure 26: Distribution of HCHO Levels of each city in 2023

Trends in HCHO levels in each city by Seasons

Colombo Proper



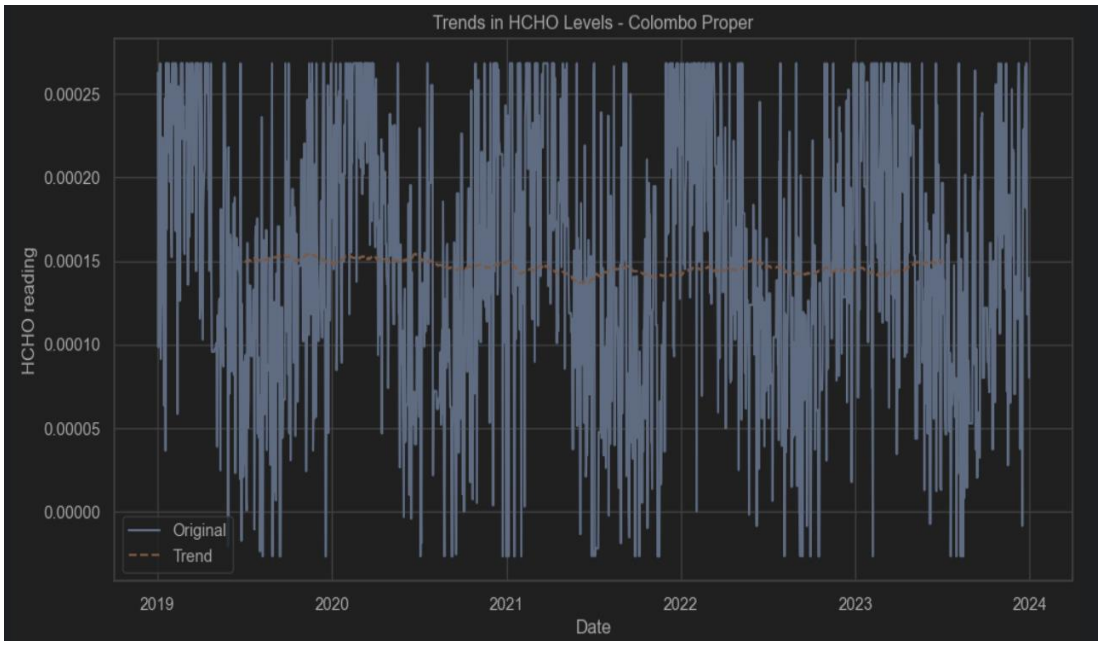
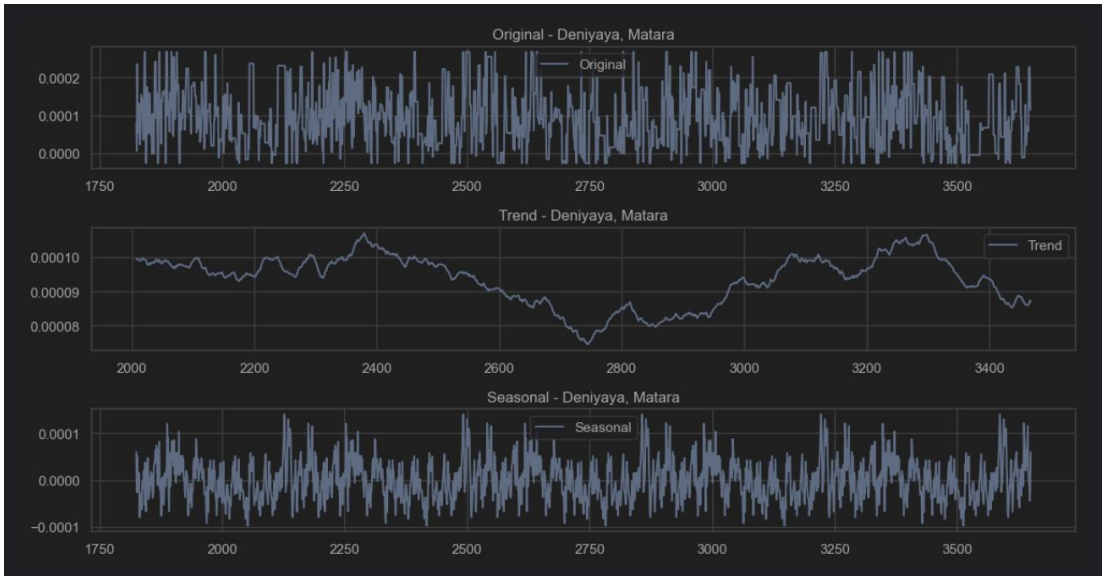


Figure 27: Trends in HCHO levels in Colombo Proper

Deniyaya, Matara



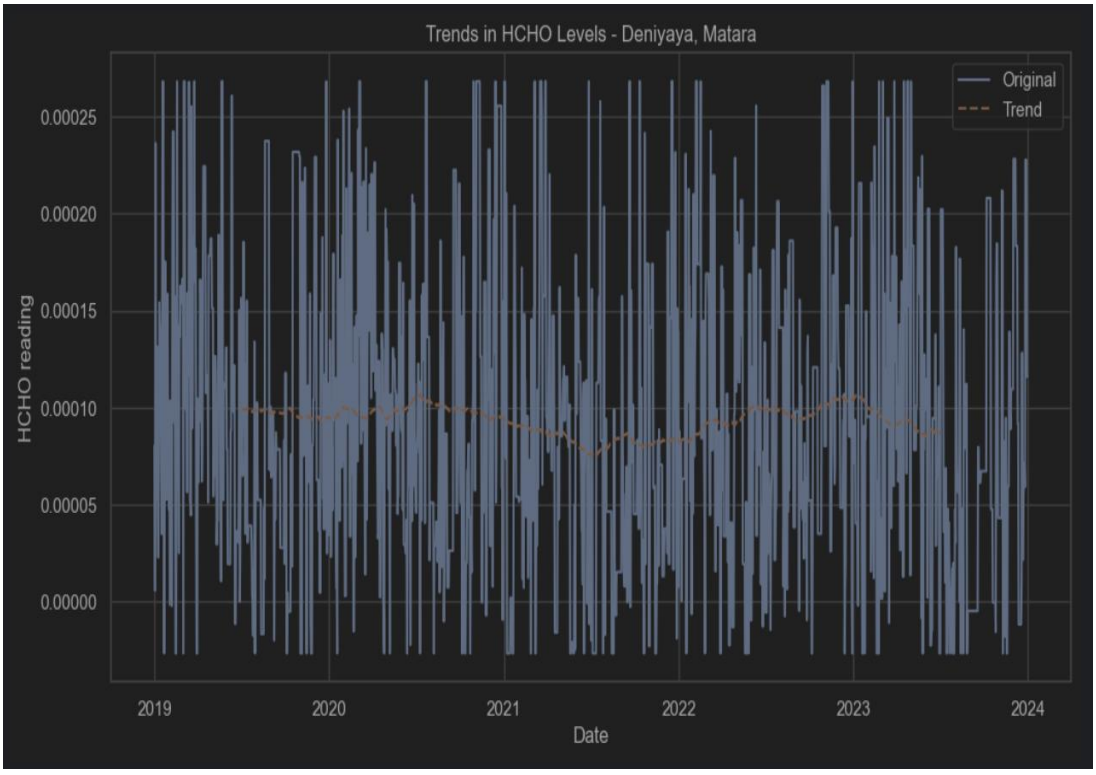
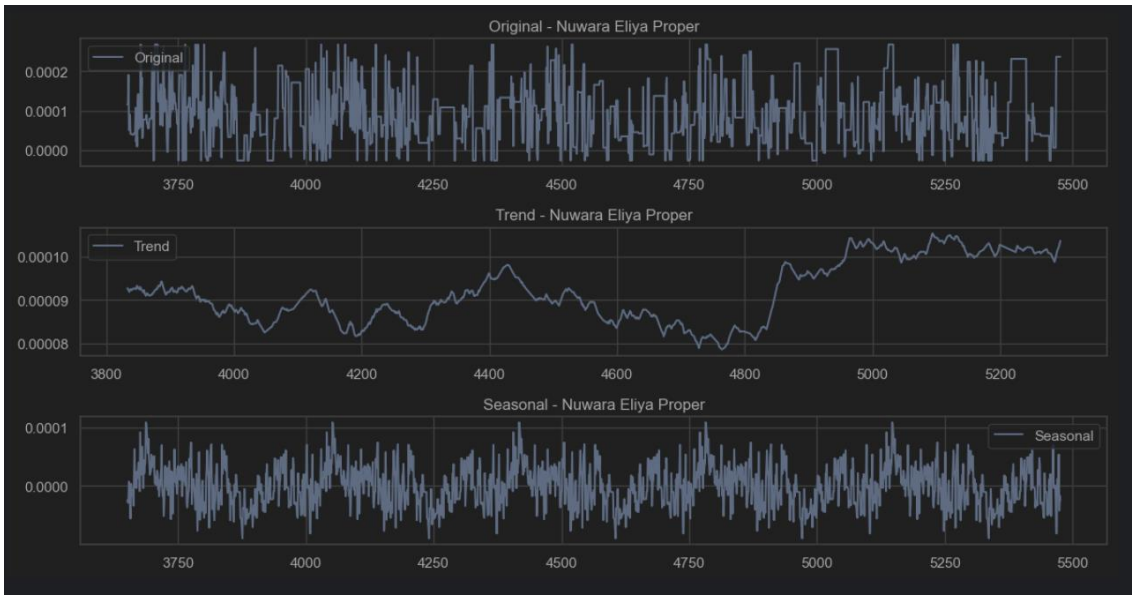


Figure 28 : Trends in HCHO levels in Deniyaya, Matara

Nuwara Eliya Proper



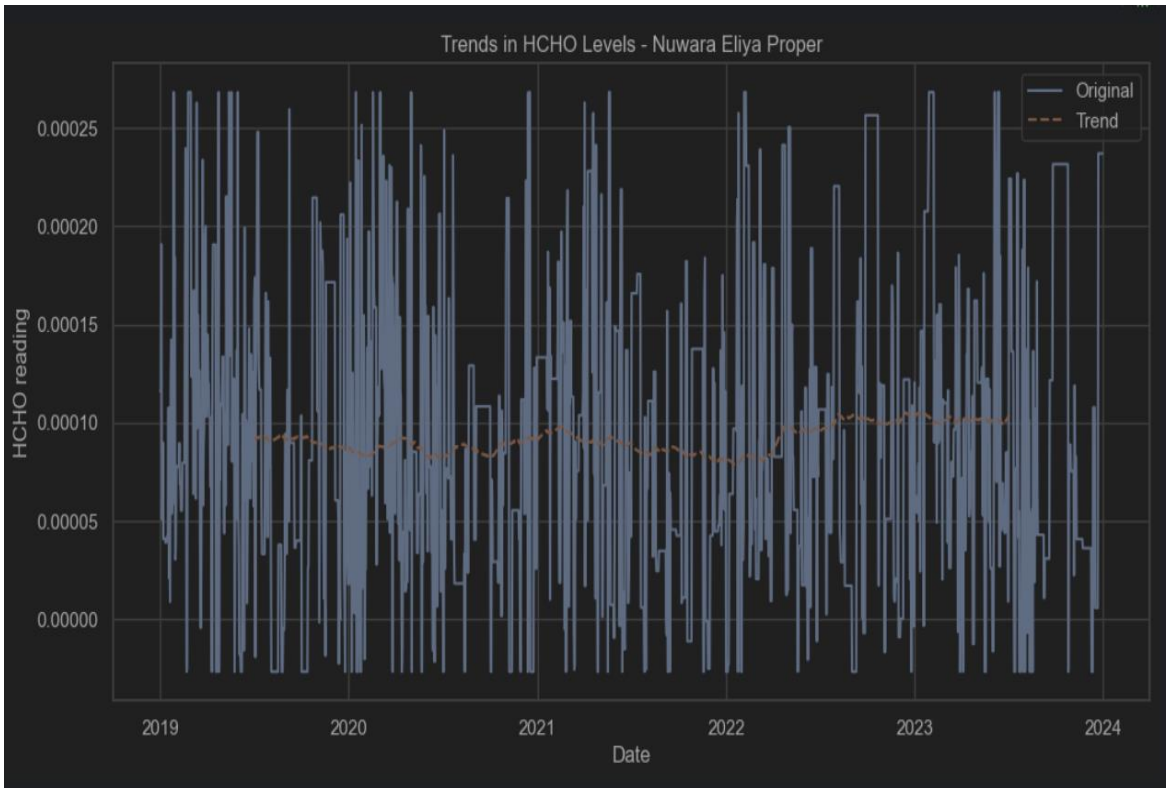
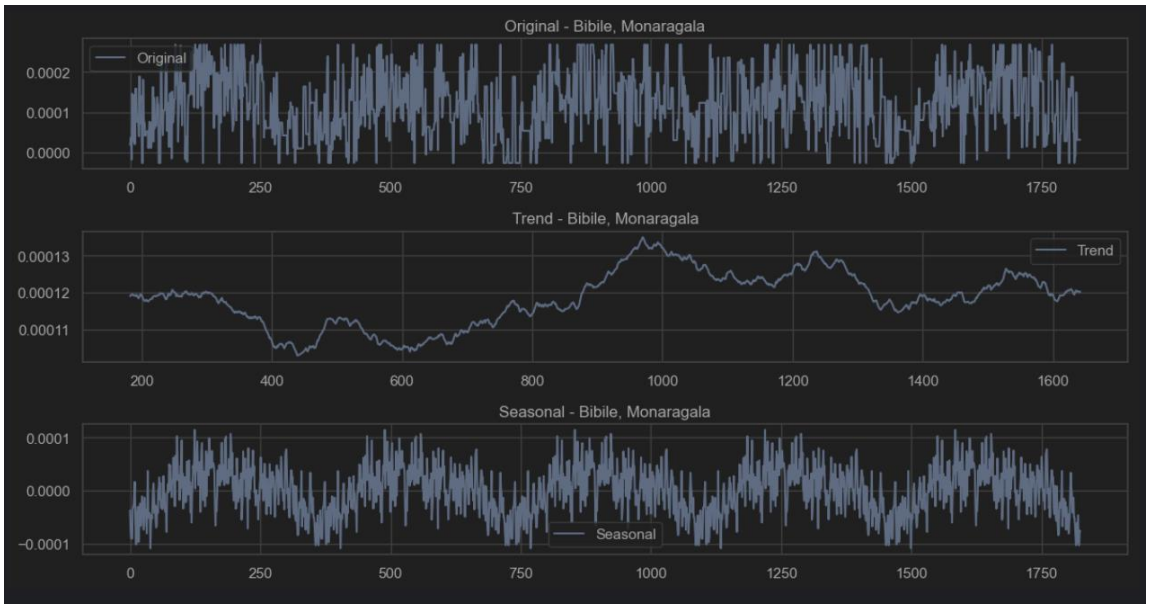


Figure 29 : Trends in HCHO levels in Nuwara Eliya Proper

Bibile, Monaragala



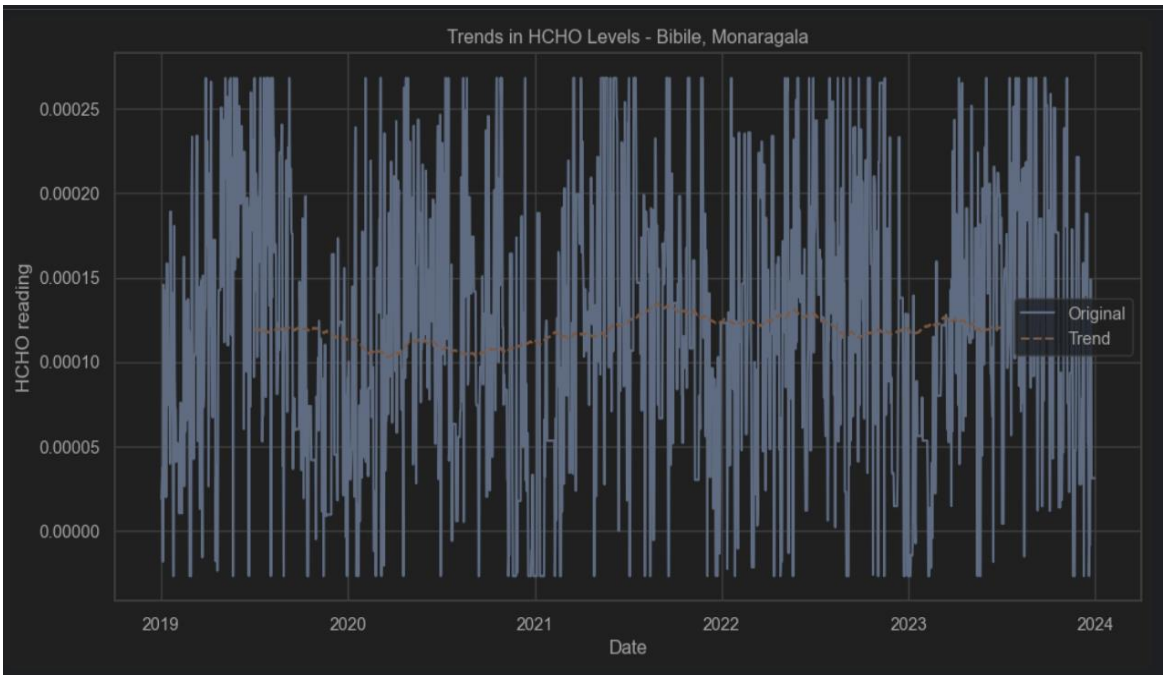
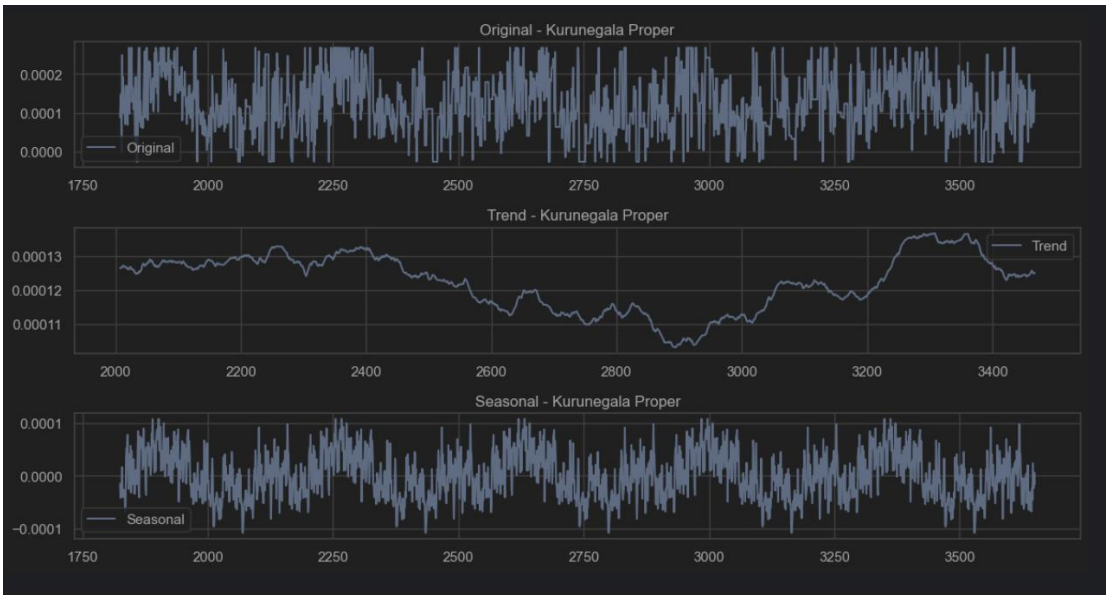


Figure 30 : Trends in HCHO levels in Bibile, Monaragala

Kurunagala Proper



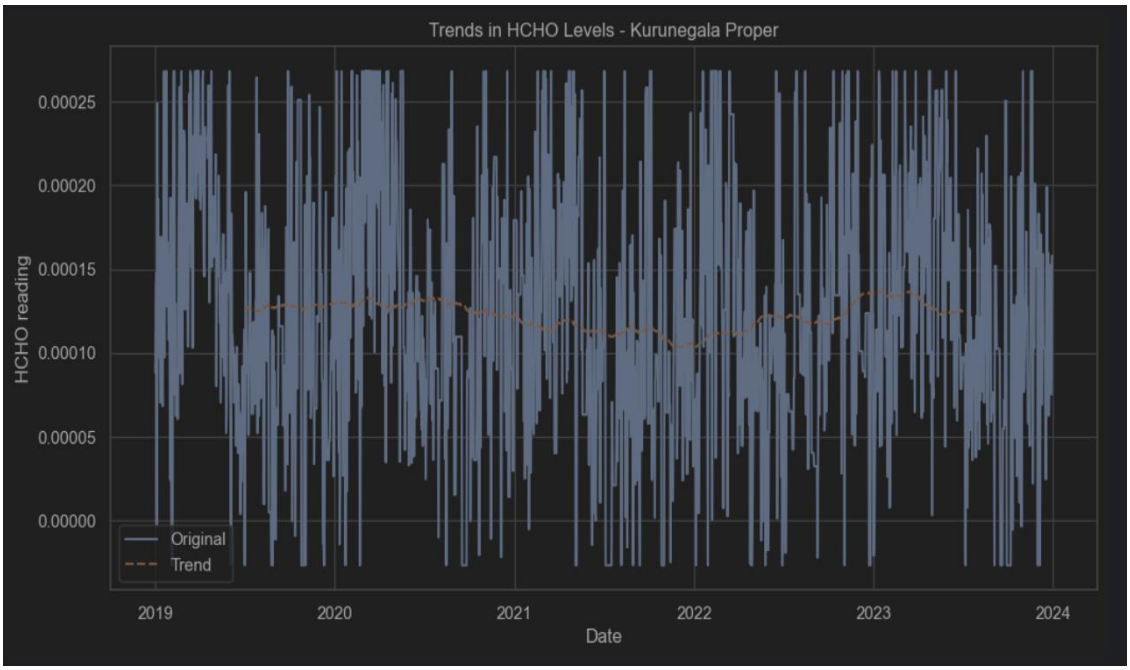
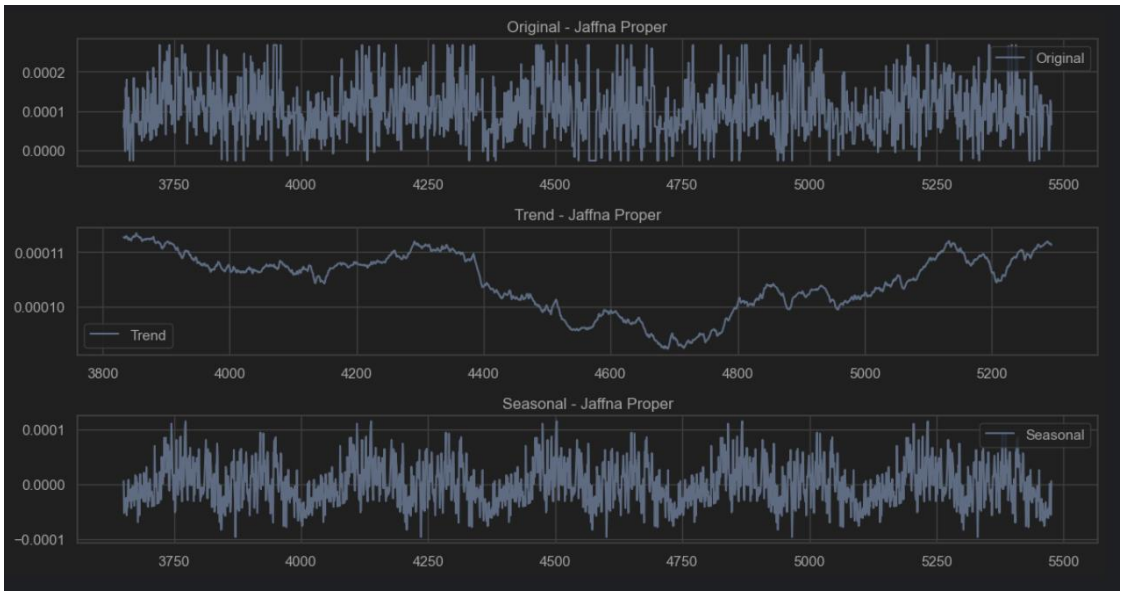


Figure 31: Trends in HCHO levels in Kurunagala Proper

Jaffna Proper



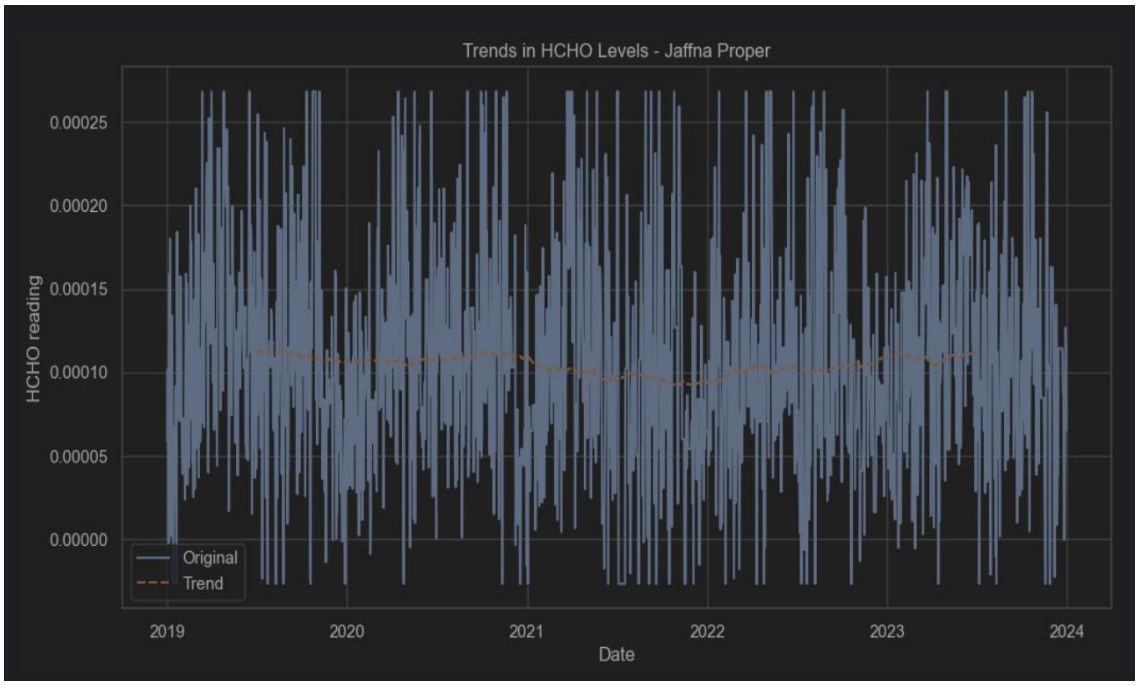
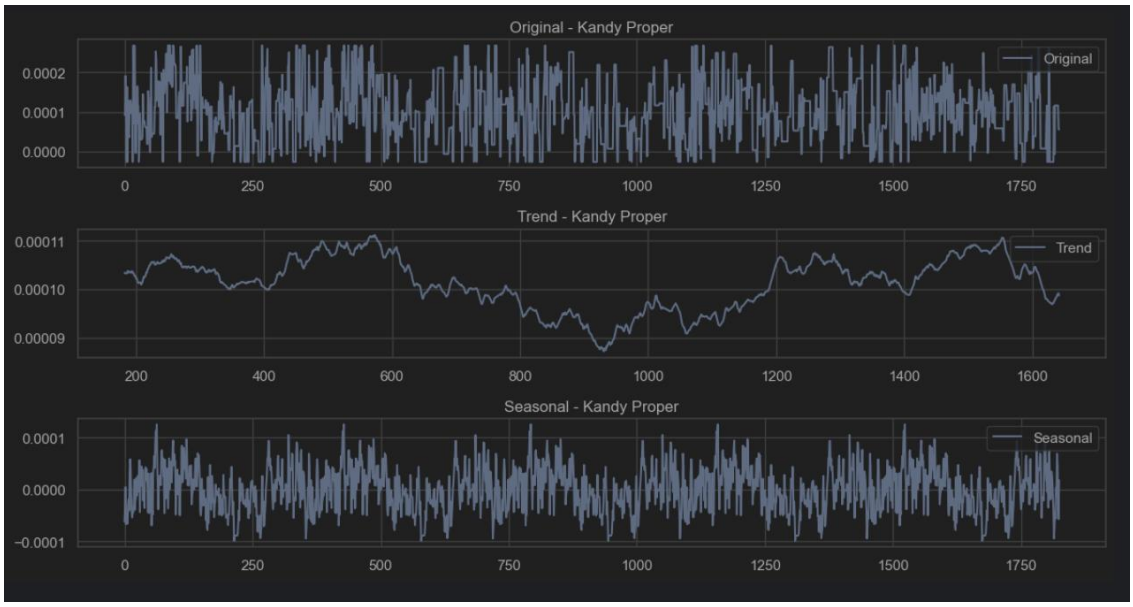


Figure 32 : Trends in HCHO levels in Jaffna Proper

Kandy Proper



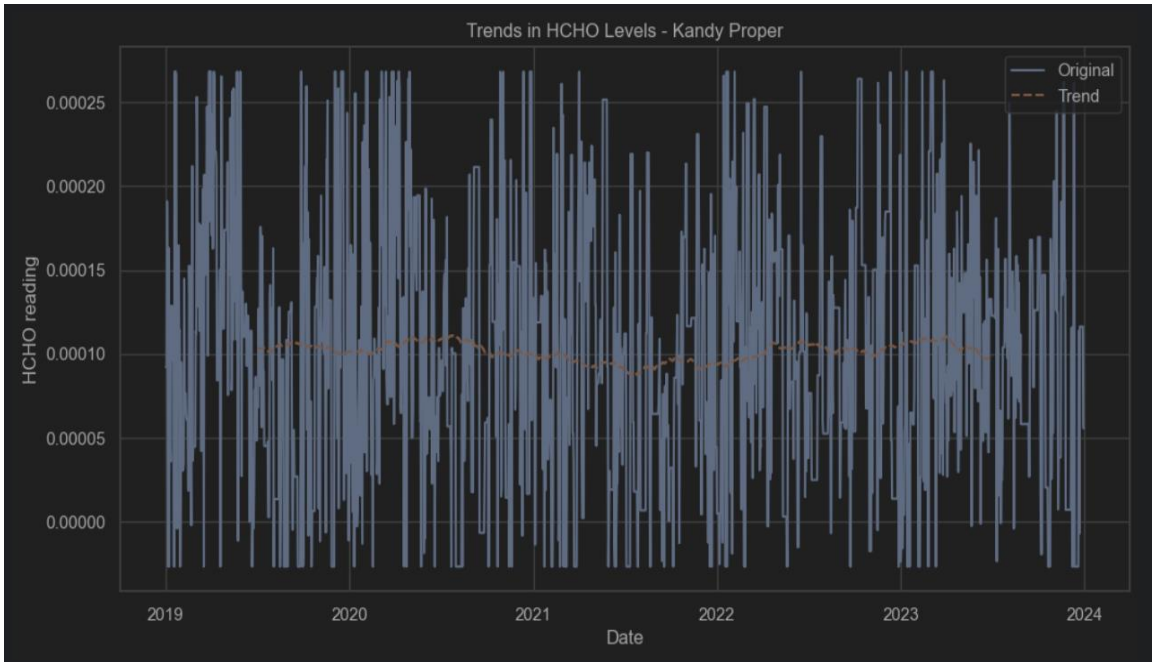


Figure 33: Trends in HCHO levels in Kandy Proper

Changes in gas emissions due to the Covid – 19 lockdowns

Pre – Pandemic

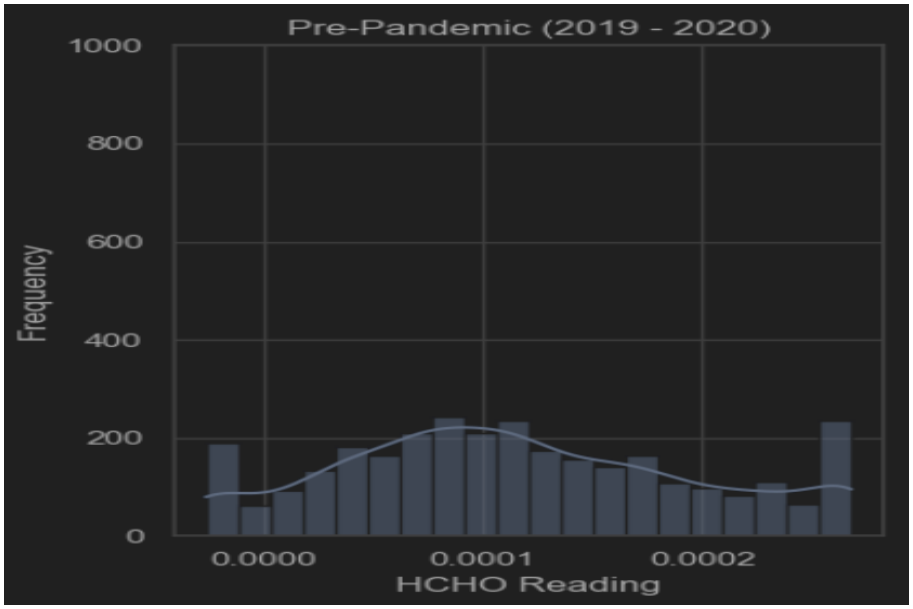


Figure 34: Changes in gas emissions due to the Covid – 19 Pre – Pandemic Period

Pandemic

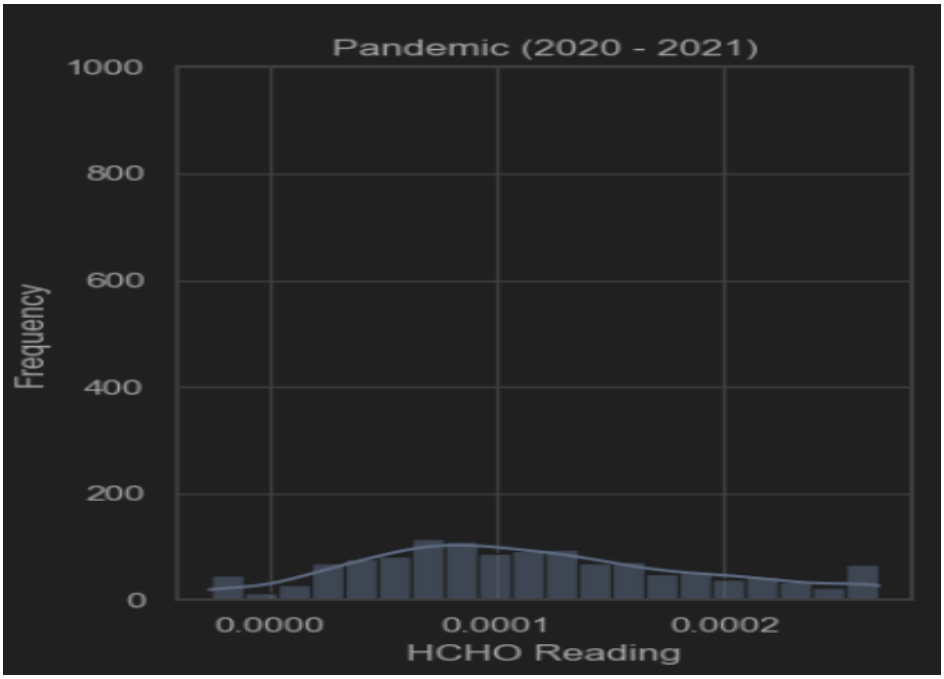


Figure 35 : Changes in gas emissions due to the Covid – 19 Pandemic Period

Post Pandemic

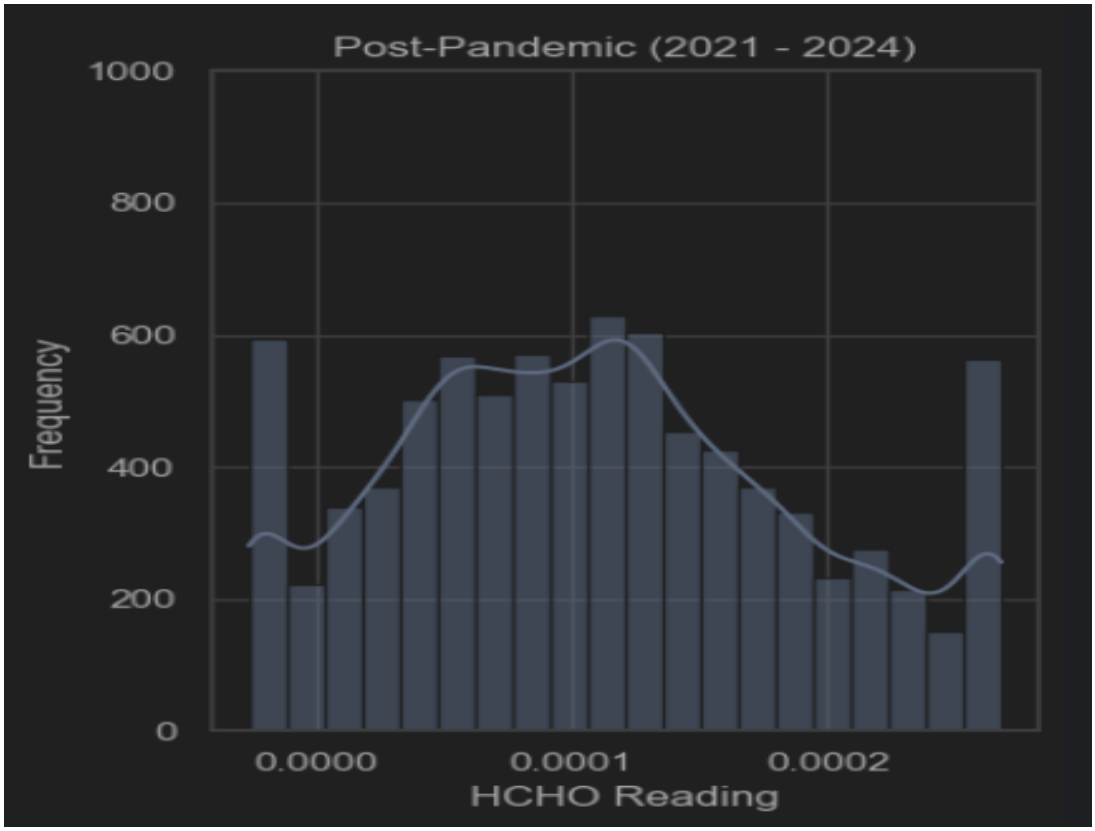


Figure 36: Changes in gas emissions due to the Covid – 19 Post - Pandemic Period

ARIMA model Implementation

In our project, we applied ARIMA models to forecast formaldehyde (HCHO) levels, which is a common approach in time series analysis. ARIMA stands for Auto Regressive Integrated Moving Average, and it's a technique used to model time-dependent data by capturing patterns in the series' past values and incorporating information about how the data changes over time. Essentially, it combines autoregression (AR), differencing (I), and moving average (MA) components to make predictions. We chose ARIMA because it's well-suited for capturing the complex dynamics of air quality data, which often exhibit seasonality and other temporal patterns.

To evaluate the effectiveness of our ARIMA models, we used performance metrics like Mean Absolute Error (MAE) and Mean Squared Error (MSE). These metrics help us understand how well our forecasts align with the actual observed values. MAE measures the average magnitude of errors between predicted and actual values, while MSE provides a measure of the average squared differences between them. By assessing these metrics, we can gauge the accuracy of our predictions and make informed decisions about the reliability of our forecasting models. Overall, employing ARIMA models and evaluating their performance with MAE and MSE allowed us to gain valuable insights into the future trends of HCHO levels, enhancing our understanding of air quality dynamics and contributing to more effective environmental management strategies.

```
from statsmodels.tsa.arima.model import ARIMA
```

Figure 37: Import ARIMA

Forecasts for Colombo Proper

```
Forecasts for Colombo Proper:

2023-12-31    0.000116
2024-01-01    0.000122
2024-01-02    0.000131
2024-01-03    0.000135
2024-01-04    0.000139
2024-01-05    0.000142
2024-01-06    0.000144
2024-01-07    0.000145
2024-01-08    0.000146
2024-01-09    0.000146
2024-01-10    0.000147
2024-01-11    0.000147
2024-01-12    0.000147
2024-01-13    0.000147
2024-01-14    0.000147
2024-01-15    0.000147
2024-01-16    0.000147
2024-01-17    0.000148
2024-01-18    0.000148
2024-01-19    0.000148
2024-01-20    0.000148
```

Figure 38 : Forecasts for Colombo Proper

Forecasts for Deniyaya, Matara

```
Forecasts for Deniyaya, Matara:

2023-12-31    0.000107
2024-01-01    0.000102
2024-01-02    0.000099
2024-01-03    0.000097
2024-01-04    0.000096
2024-01-05    0.000095
2024-01-06    0.000094
2024-01-07    0.000094
2024-01-08    0.000094
2024-01-09    0.000094
2024-01-10    0.000093
2024-01-11    0.000093
2024-01-12    0.000093
2024-01-13    0.000093
2024-01-14    0.000093
2024-01-15    0.000093
2024-01-16    0.000093
2024-01-17    0.000093
2024-01-18    0.000093
2024-01-19    0.000093
2024-01-20    0.000093
```

```
2024-01-21    0.000093
2024-01-22    0.000093
2024-01-23    0.000093
2024-01-24    0.000093
2024-01-25    0.000093
2024-01-26    0.000093
2024-01-27    0.000093
2024-01-28    0.000093
2024-01-29    0.000093
2024-01-30    0.000093
Freq: D, Name: Arima Predictions, dtype: float64
```

Figure 39: Forecasts for Deniyaya, Matara

Forecasts for Nuwara Eliya Proper

a

```
Forecasts for Nuwara Eliya Proper:

2023-12-31    0.000191
2024-01-01    0.000166
2024-01-02    0.000146
2024-01-03    0.000131
2024-01-04    0.000121
2024-01-05    0.000113
2024-01-06    0.000108
2024-01-07    0.000104
2024-01-08    0.000101
2024-01-09    0.000099
2024-01-10    0.000097
2024-01-11    0.000096
2024-01-12    0.000095
2024-01-13    0.000094
2024-01-14    0.000094
2024-01-15    0.000094
2024-01-16    0.000093
2024-01-17    0.000093
2024-01-18    0.000093
```

Figure 40: Forecasts for Nuwara Eliya Proper

```
2024-01-19    0.000093
2024-01-20    0.000093
2024-01-21    0.000093
2024-01-22    0.000093
2024-01-23    0.000093
2024-01-24    0.000093
2024-01-25    0.000093
2024-01-26    0.000093
2024-01-27    0.000093
2024-01-28    0.000093
2024-01-29    0.000093
2024-01-30    0.000093
Freq: D, Name: Arima Predictions, dtype: float64
```

Forecasts for Bibile, Monaragala

```
Forecasts for Bibile, Monaragala:

2023-12-31    0.000066
2024-01-01    0.000082
2024-01-02    0.000094
2024-01-03    0.000103
2024-01-04    0.000108
2024-01-05    0.000112
2024-01-06    0.000115
2024-01-07    0.000116
2024-01-08    0.000117
2024-01-09    0.000118
2024-01-10    0.000119
2024-01-11    0.000119
2024-01-12    0.000119
2024-01-13    0.000120
2024-01-14    0.000120
2024-01-15    0.000120
2024-01-16    0.000120
2024-01-17    0.000120
2024-01-18    0.000120
```

Figure 41: Forecasts for Bibile, Monaragala

```
2024-01-19    0.000120
2024-01-20    0.000120
2024-01-21    0.000120
2024-01-22    0.000120
2024-01-23    0.000120
2024-01-24    0.000120
2024-01-25    0.000120
2024-01-26    0.000120
2024-01-27    0.000120
2024-01-28    0.000120
2024-01-29    0.000120
2024-01-30    0.000120
Freq: D, Name: Arima Predictions, dtype: float64
```

Forecasts for Kurunagala Proper

```
Forecasts for Kurunegala Proper:

2023-12-31    0.000144
2024-01-01    0.000137
2024-01-02    0.000133
2024-01-03    0.000130
2024-01-04    0.000127
2024-01-05    0.000126
2024-01-06    0.000125
2024-01-07    0.000125
2024-01-08    0.000124
2024-01-09    0.000124
2024-01-10    0.000124
2024-01-11    0.000124
2024-01-12    0.000124
2024-01-13    0.000124
2024-01-14    0.000124
2024-01-15    0.000124
2024-01-16    0.000124
2024-01-17    0.000124
2024-01-18    0.000124
```

Figure 42: Forecasts for Kurunagala Proper

```
2024-01-19    0.000124
2024-01-20    0.000124
2024-01-21    0.000124
2024-01-22    0.000124
2024-01-23    0.000124
2024-01-24    0.000124
2024-01-25    0.000124
2024-01-26    0.000124
2024-01-27    0.000124
2024-01-28    0.000124
2024-01-29    0.000124
2024-01-30    0.000124
Freq: D, Name: Arima Predictions, dtype: float64
```

Forecasts for Jaffna Proper

```
Forecasts for Jaffna Proper:

2023-12-31     0.000093
2024-01-01     0.000098
2024-01-02     0.000102
2024-01-03     0.000104
2024-01-04     0.000105
2024-01-05     0.000106
2024-01-06     0.000106
2024-01-07     0.000106
2024-01-08     0.000106
2024-01-09     0.000106
2024-01-10     0.000106
2024-01-11     0.000106
2024-01-12     0.000106
2024-01-13     0.000106
2024-01-14     0.000106
2024-01-15     0.000106
2024-01-16     0.000106
2024-01-17     0.000106
2024-01-18     0.000106
```

Figure 43: Forecasts for Jaffna Proper

```
2024-01-19    0.000106
2024-01-20    0.000106
2024-01-21    0.000106
2024-01-22    0.000106
2024-01-23    0.000106
2024-01-24    0.000106
2024-01-25    0.000106
2024-01-26    0.000106
2024-01-27    0.000106
2024-01-28    0.000106
2024-01-29    0.000106
2024-01-30    0.000106
Freq: D, Name: Arima Predictions, dtype: float64
```

Forecasts for Kandy Proper

```
Forecasts for Kandy Proper:

2023-12-31    0.000074
2024-01-01    0.000083
2024-01-02    0.000089
2024-01-03    0.000093
2024-01-04    0.000096
2024-01-05    0.000098
2024-01-06    0.000099
2024-01-07    0.000100
2024-01-08    0.000100
2024-01-09    0.000101
2024-01-10    0.000101
2024-01-11    0.000101
2024-01-12    0.000101
2024-01-13    0.000101
2024-01-14    0.000101
2024-01-15    0.000101
2024-01-16    0.000101
2024-01-17    0.000101
2024-01-18    0.000101
```

Figure 44: Forecasts for Kandy Proper

```
2024-01-19    0.000101
2024-01-20    0.000101
2024-01-21    0.000101
2024-01-22    0.000101
2024-01-23    0.000101
2024-01-24    0.000101
2024-01-25    0.000101
2024-01-26    0.000101
2024-01-27    0.000101
2024-01-28    0.000101
2024-01-29    0.000101
2024-01-30    0.000101
Freq: D, Name: Arima Predictions, dtype: float64
```

Evaluate the model's performance using appropriate metrics.

In this part of the assignment, we implemented an ARIMA model to forecast HCHO levels for various locations in the dataset. The model's performance was evaluated using key metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). These metrics provide insights into the accuracy and effectiveness of the ARIMA model in predicting HCHO levels. For each location, the MAE, MSE, and RMSE were calculated and printed to assess the model's performance. These metrics help us understand the extent of errors between the predicted and actual HCHO levels, providing valuable information on the model's accuracy and reliability.

Overall, evaluating the ARIMA model's performance using these metrics enables us to gauge its effectiveness in forecasting HCHO levels across different locations. This analysis aids in identifying areas where the model performs well and areas where improvements may be necessary, contributing to more accurate and reliable predictions in the future.

Metrics for Colombo Proper

```
Metrics for Colombo Proper:
Mean Absolute Error (MAE): 5.7860768595170 82e-05
Mean Squared Error (MSE): 5.418757248689996e-09
Root Mean Squared Error (RMSE): 7.361220855734458e-05
```

Figure 45: Metrics for Colombo Proper

Metrics Fore Deniyaya, Matara

```
Metrics for Deniyaya, Matara:
Mean Absolute Error (MAE): 8.002013827752353e-05
Mean Squared Error (MSE): 8.412778236712646e-09
Root Mean Squared Error (RMSE): 9.172119840425465e-05
```

Figure 46 : Metrics Fore Deniyaya, Matara

Metrics for Nuwara Eliya, Proper

```
Metrics for Nuwara Eliya Proper:
Mean Absolute Error (MAE): 9.999935991578739e-05
Mean Squared Error (MSE): 1.18715704909861e-08
Root Mean Squared Error (RMSE): 0.00010895673678568985
```

Figure 47:Metrics for Nuwara Eliya, Proper

Metrics for Bibile, Monaragala

```
Metrics for Bibile, Monaragala:
Mean Absolute Error (MAE): 6.852536809034371e-05
Mean Squared Error (MSE): 5.9865675843014375e-09
Root Mean Squared Error (RMSE): 7.737291247136453e-05
```

Figure 48 : Metrics for Bibile, Monaragala

Metric for Kurunagala Proper

```
Metrics for Kurunegala Proper:
Mean Absolute Error (MAE): 5.582259990018723e-05
Mean Squared Error (MSE): 4.476126589048634e-09
Root Mean Squared Error (RMSE): 6.690386079329528e-05
```

Figure 49 : Metric for Kurunagala Proper

Metrics for Jaffna Proper

```
Metrics for Jaffna Proper:
Mean Absolute Error (MAE): 3.833979528493925e-05
Mean Squared Error (MSE): 2.6956977343943516e-09
Root Mean Squared Error (RMSE): 5.192010915237324e-05
```

Figure 50 : Metrics for Jaffna Proper

Metrics for Kandy Proper

```
Metrics for Kandy Proper:
Mean Absolute Error (MAE): 7.995444149691989e-05
Mean Squared Error (MSE): 9.057766974245994e-09
Root Mean Squared Error (RMSE): 9.517230150756046e-05
```

Figure 51: Metrics for Kandy Proper

Conclusion

Summary of Findings

Formaldehyde (HCHO) level analysis in Sri Lanka provides interesting new information about temporal and spatial trends. There are noticeable seasonal fluctuations, with HCHO concentrations showing unique patterns in each season. Elevated levels of HCHO have been recorded at specific periods, such as the monsoon season. This phenomenon may be attributed to increased humidity and meteorological conditions that facilitate the development of pollutants. Moreover, there have been noteworthy variations in HCHO levels before, during, and after the Covid-19 pandemic, indicating the pandemic's impact on these levels. These results emphasize the dynamic character of air quality dynamics and show how outside events impact Sri Lanka's pollution levels. It is essential to comprehend these patterns in order to create policies for managing air quality that are both successful and safe for the environment and public health.

Recommendations for Further Research

Further investigation is necessary to improve our knowledge of HCHO dynamics in Sri Lanka. Future research could examine the complex interactions that exist between HCHO levels and certain environmental factors like temperature, humidity, wind patterns, and land use characteristics. Researchers can clarify the causes of HCHO emissions and pinpoint viable mitigation techniques by looking more closely at these variables. Furthermore, examining how well urban planning projects and emission control strategies work to lower HCHO concentrations may be able to provide policymakers with important information. Future studies can help make better decisions and move Sri Lanka closer to implementing sustainable environmental management strategies by filling in these knowledge gaps.

Acknowledgement

We thank the European Space Agency in this part for granting us access to the dataset that we used for our investigation. The Sentinel-5P satellite data has been invaluable in our research endeavors to comprehend formaldehyde (HCHO) concentrations and their consequences for environmental and public health.

Data Source

The dataset from the Sentinel-5P satellite of the European Space Agency is a major source of data for our investigation. As it circles the planet, this satellite gathers a plethora of information for remote sensing, which includes atmospheric composition measurements. The sophisticated equipment of the Sentinel-5P allows for accurate and thorough monitoring of formaldehyde (HCHO) concentrations over a wide range of geographic areas. By utilizing this abundant data set, we can examine temporal and regional patterns in HCHO levels in unprecedented detail. We can obtain important insights into the dynamics of HCHO distribution and its consequences for human and environmental health by utilizing the capabilities of the Sentinel-5P satellite.

Limitations and Uncertainties

It's important to recognize the Sentinel-5P dataset's inherent limitations and uncertainties, despite the abundance of information it provides. Satellite measurement mistakes are a major concern. These inaccuracies can be caused by a variety of variables, including atmospheric interference, differences in sensor calibration, and algorithmic constraints in data processing. Furthermore, atmospheric circumstances, equipment deterioration, and shifts in satellite orbit will inevitably cause differences in data quality over time. Due diligence in interpreting the data is vital, and taking potential biases into account in our analysis is made even more crucial by these constraints and uncertainties.

References

- Hans, A. (2021) *Data preprocessing in Python: All important steps explained, Medium*. Available at: https://medium.com/nerd-for-tech/data-preprocessing-and-cleaning-in-python-all-important-steps-explained-6093b8cb0864 (Accessed: 21 April 2024).

- jth359jth359 80911 gold badge99 silver badges1010 bronze badges, user1827356user1827356 6 and bunjibunji 5 (1962) *Fill forwards or backwards depending on another row, Stack Overflow*. Available at: https://stackoverflow.com/questions/42004778/fill-forwards-or-backwards-depending-on-another-row (Accessed: 21 April 2024).

- YAĞCI, H.E. (2021) *Detecting and handling outliers with pandas, Medium*. Available at: https://hersanyagci.medium.com/detecting-and-handling-outliers-with-pandas-7adbfcd5cad8 (Accessed: 21 April 2024).

- *Master power bi essentials in just 15 minutes* (2023a) *YouTube*. Available at: https://www.youtube.com/watch?v=nkmHqs1I_z0 (Accessed: 21 April 2024).

- Pathak, P. (2024) *Building an Arima model for time series forecasting in python, Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2020/10/how-to-create-an-arima-model-for-time-series-forecasting-in-python/ (Accessed: 21 April 2024).

- *Arima model in Python| Time Series forecasting #6|* (2020) *YouTube*. Available at: https://www.youtube.com/watch?v=8FCDpFhd1zk (Accessed: 21 April 2024).