BSc (Hons) Artificial Intelligence and Data Science

Module: CM2606 Data Engineering

Coursework Report

Module Leader: Mr. Mohamed Ayoob

RGU Student ID : 2313081

IIT Student ID : 20221470

Student Name : Withanage Lakshan Anjana Cooray

Table of Contents

Introduction

This report discusses how formaldehyde gas distribution has been impacted in several regions of Sri Lanka during 2019 to 2023. This formaldehyde (HCHO) distribution is analysed from data gathered by the TROPOMI instrument in the Sentinel-5P satellite. Colombo, Matara Deniyaya, Nuwara Eliya, Bibile Monaragala, Kurunegala, Jaffna, and Kandy are the main regions analysed from the above-mentioned dataset. This report mainly discusses how data preprocessing is done, including details of the data engineering framework used, how missing values are handled, how outliers are detected, and what the format inconstancies of the datasets. In addition, it also discusses the statistical summaries of each region's HCHO distribution using visualizations. Furthermore, it discusses how seasonal variation has been impacted for HCHO gas distribution by considering the COVID-19 lockdown periods in Sri Lanka. It also examines how other weather conditions, including precipitation and temperature, have been impacted and how anthropogenic activities have impacted HCHO distribution based on other gas emissions, including carbon monoxide, nitrogen dioxide, and ozone, which are the main causes of HCHO emissions. However, the limitations of finding external datasets and the uncertainty of the data are also addressed in this report. It also discusses how time series-related models can be used to forecast the HCHO distribution of each region and how these models have been evaluated using regression-based matrices. Additionally, it discusses how these key findings can be applied to future research and how they can be applied to public health and other environmental-related activities.

Data Source and Code Repository

The datasets downloaded from the sources and the analysis code are available in a public Git repository. The git link is mentioned below.

Git link: https://github.com/Lakshan2023/Fomaldihyde-Analysis-Sri-Lanka

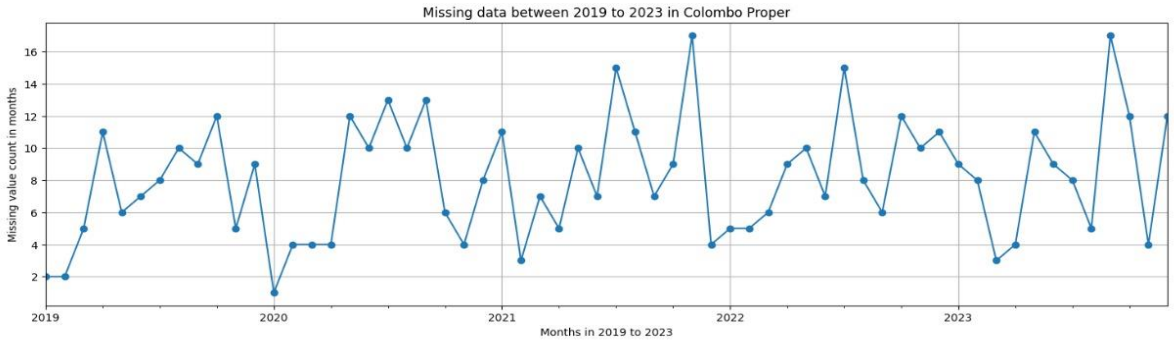
Data Preprocessing and Region wise Formaldehyde Emission Rate Analysis

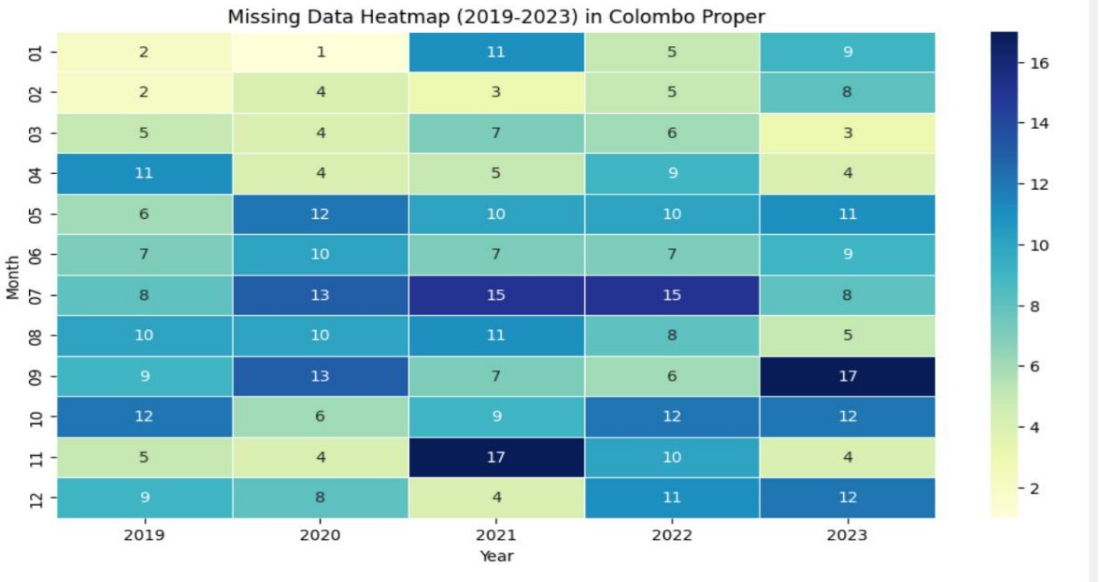
Due to the unavailability of satellite coverage in some days, formaldehyde emission data has been missed in many regions. Therefore, time-series-related null value handling methods had to be used to get the HCHO distribution without making any changes to the original distribution seasonality. This is done using pandas framework. The HCHO distribution is given on the scale molcm^{-2}, but the provided dataset consists of negative values that cannot be physically present. However, the NASA Earth Data Forum has mentioned that negative values should be considered when doing the analysis. Therefore, the negative values did not drop. (Kusterer, 2019)

Colombo Proper Region Formaldehyde Distribution Analysis

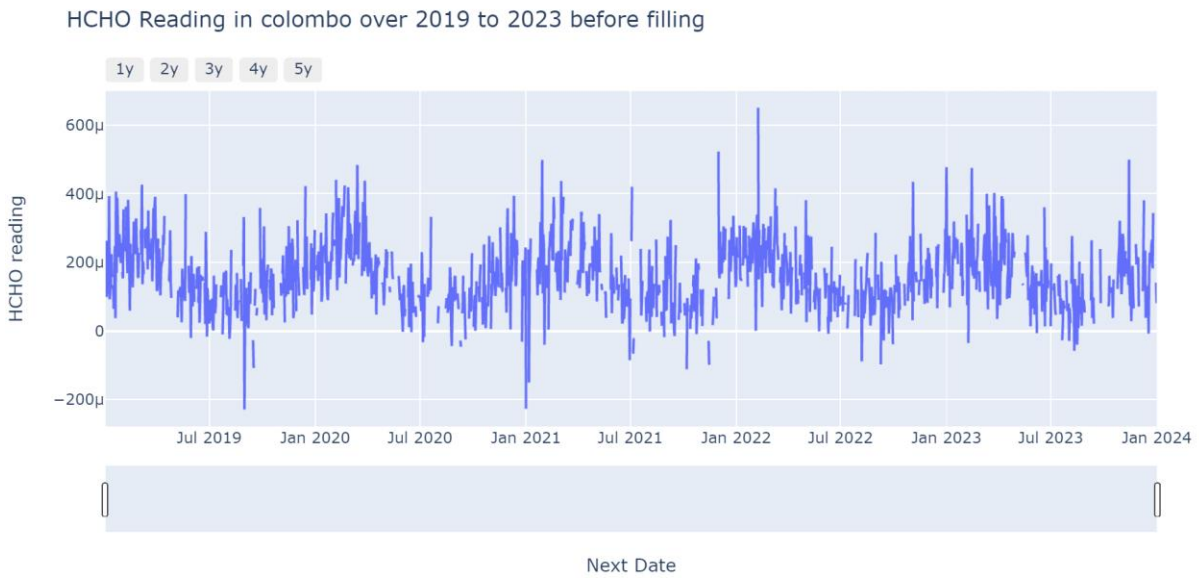
Data Preprocessing Colombo

When missing values checked ,there were 487 missing values out of 1826 values. The below line chart and heatmap shows how null values are distributed in each year in Colombo Region.

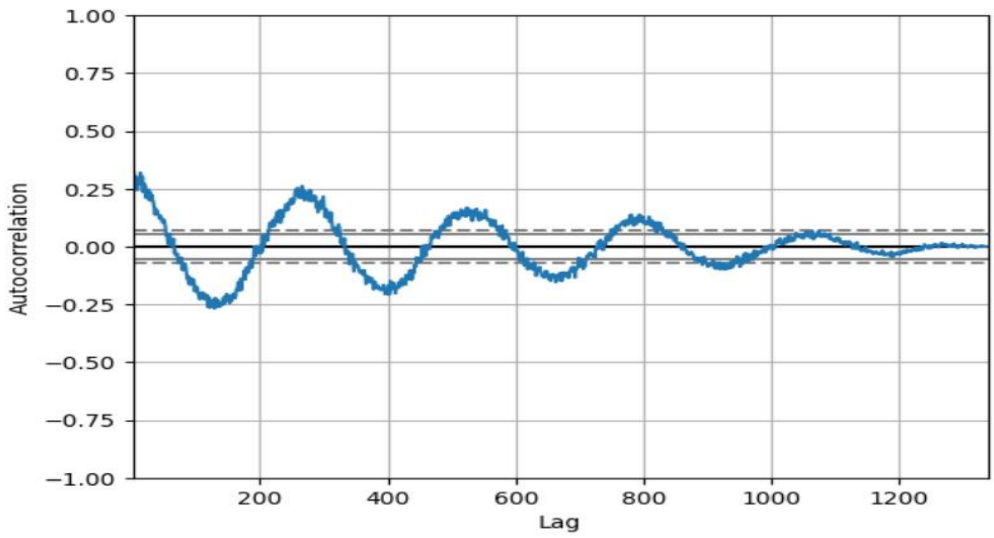




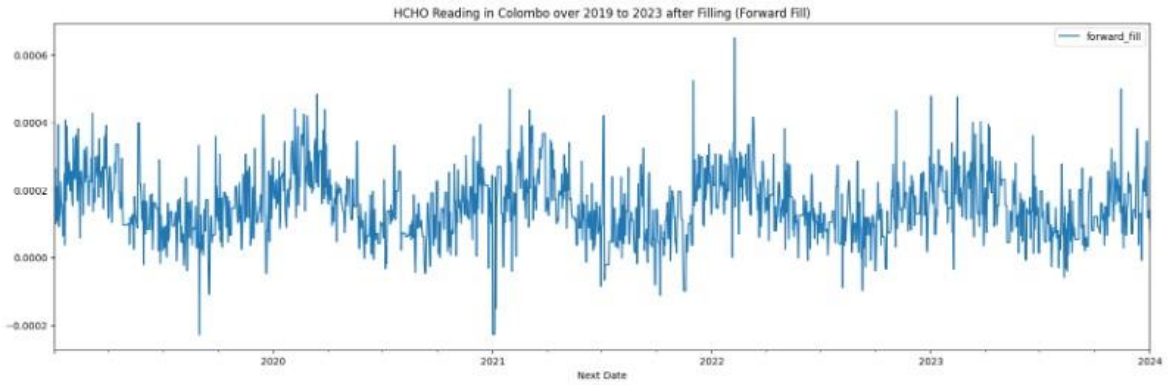
The above visualizations show that there are some months that consist of missing HCHO reading for 17 days in the month. This depicts these null values impacts a lot to the seasonality of the dataset. The below chart shows how the Colombo distribution was existed before handling missing values.

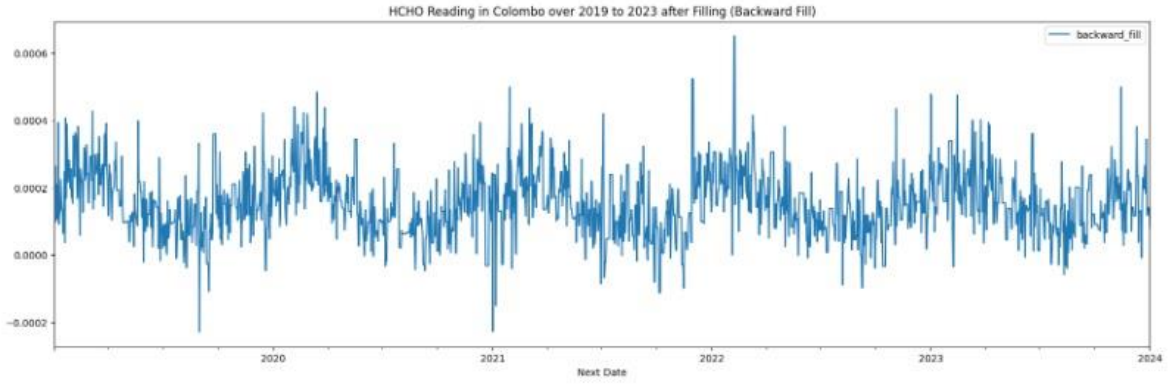


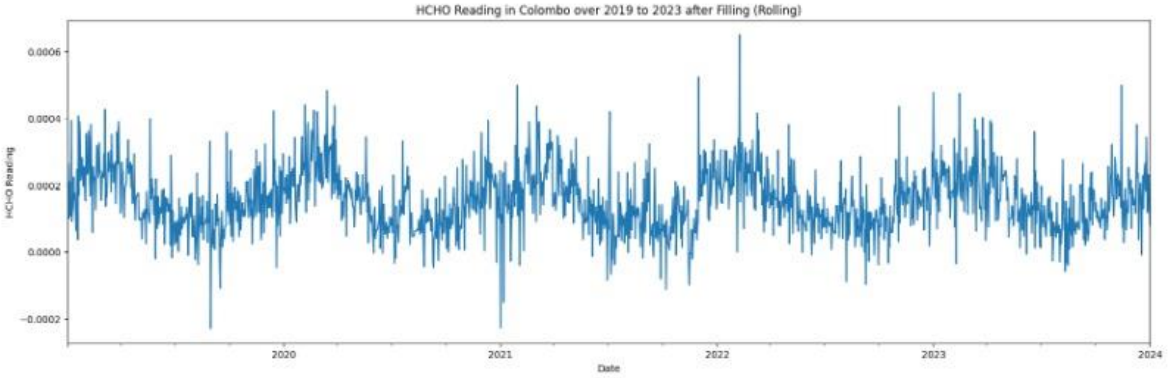
The below auto-correlation plot shows there is a seasonality in Colombo distribution. Therefore, time series based rolling method is used to fill data based on past dates. Since there are large number of missing values are there in the dataset, rolling window 8 was used to handle null values. This rolling window will replace the null values considering last 8 days mean value. Other than that checked with filling the null values using backward and forward filling methods. However, it did not maintain the fluctuation seasonality pattern.

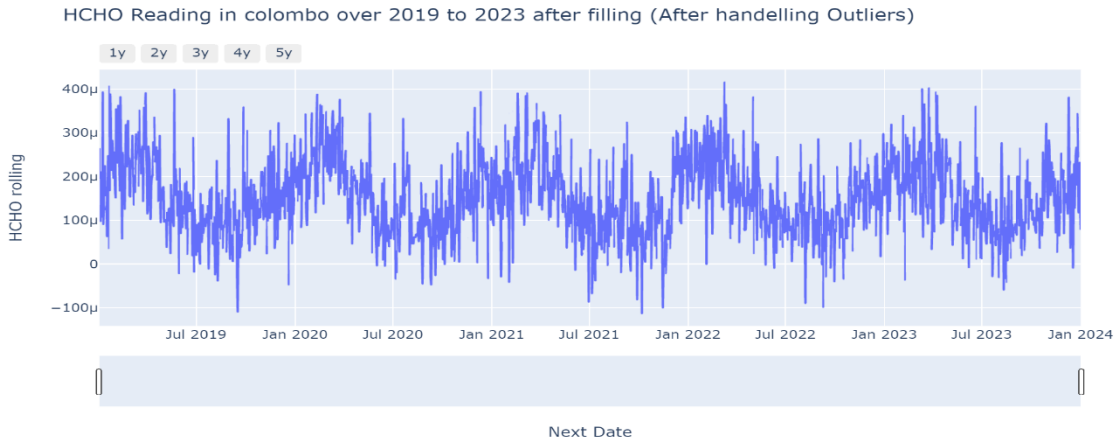


The below line chart shows the backward, forward and rolling filled HCHO distribution.

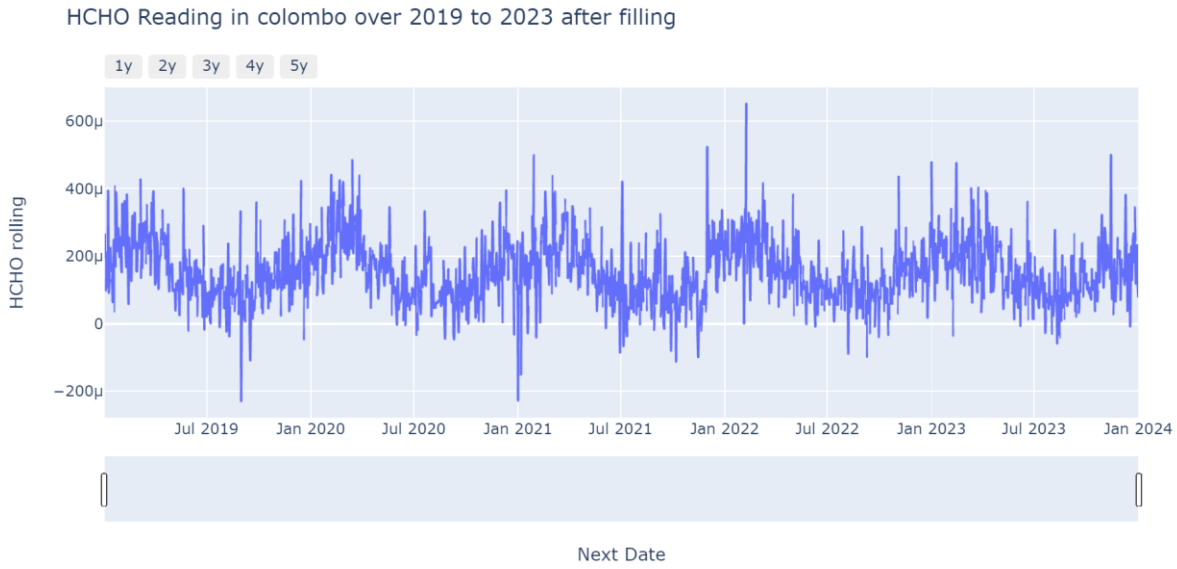






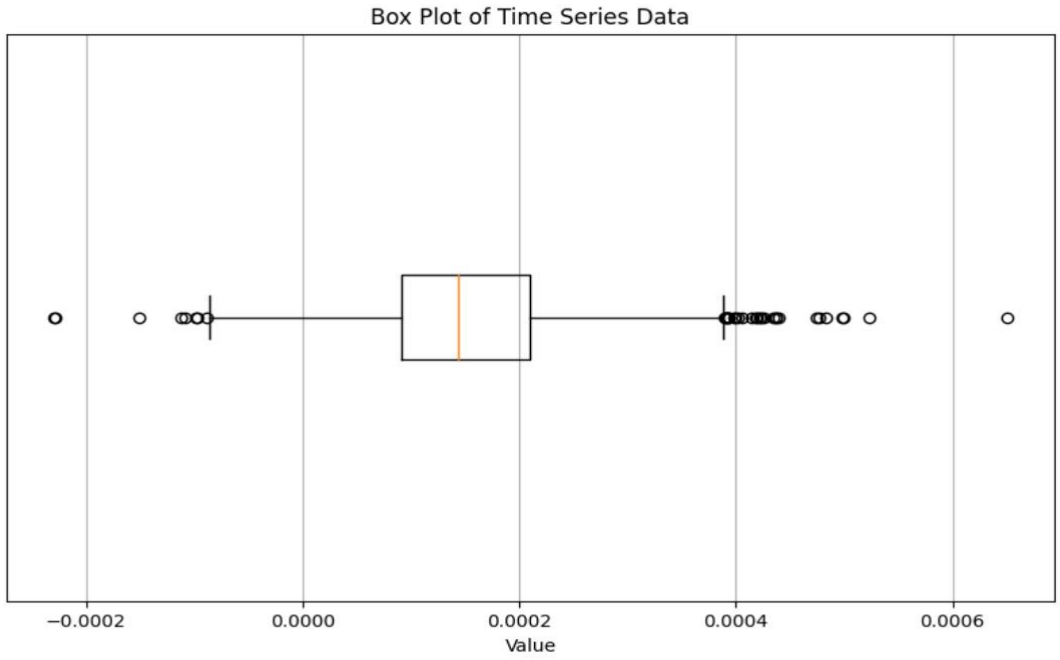


The below plot shows how Colombo HCHO distribution looks like after handling null values with rolling window 8.

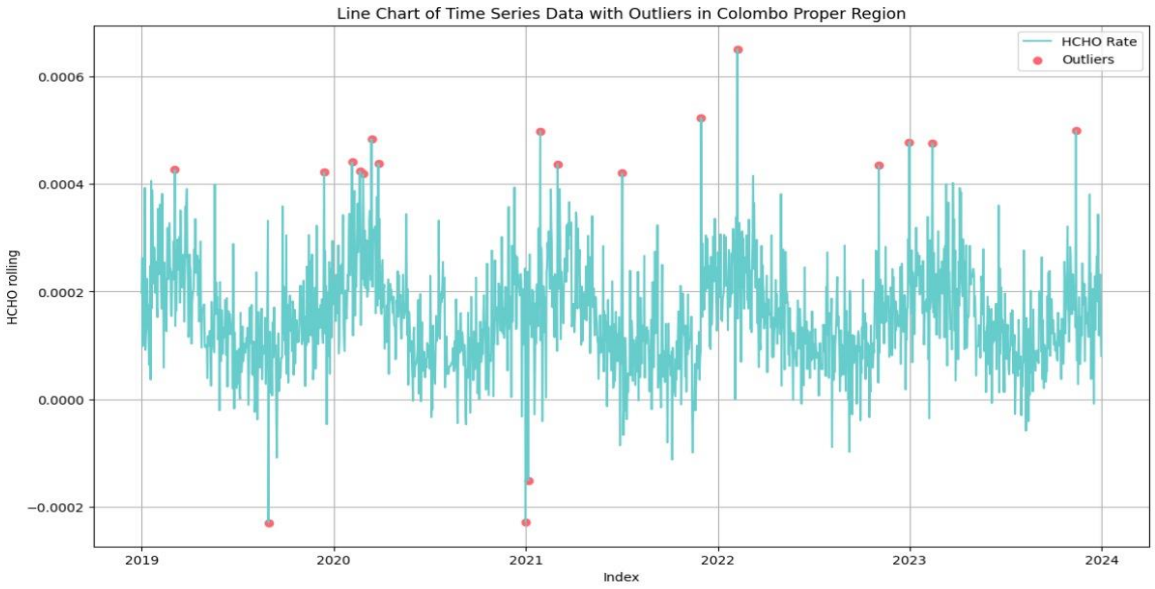


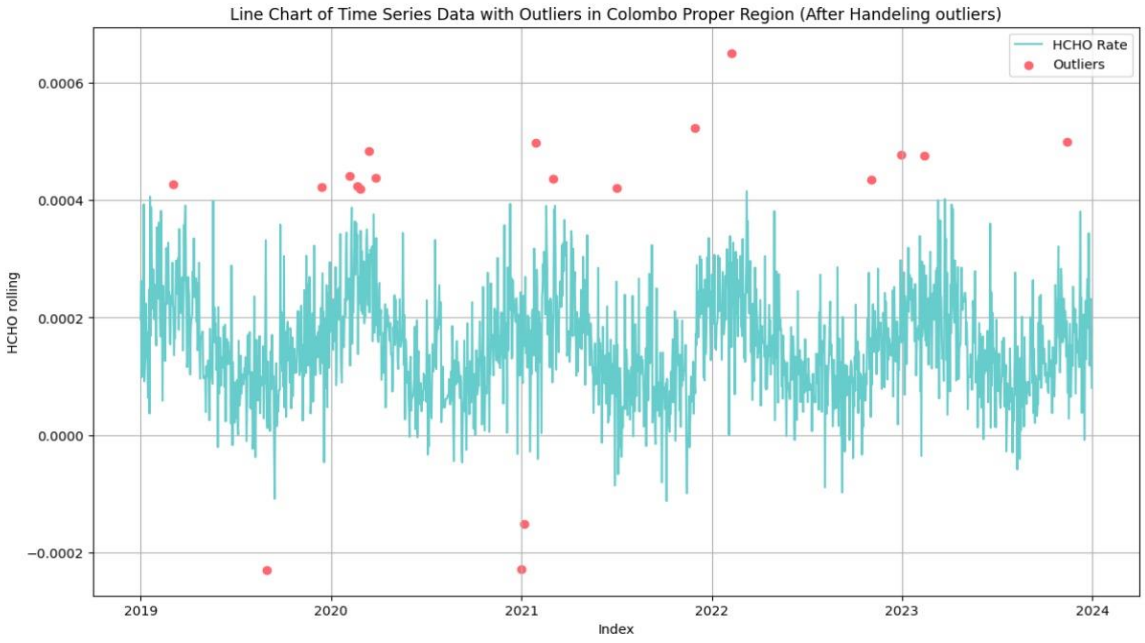
Then a box plot is used to check the outliers, and dropped them using a 1.75 threshold value based on inter quartile range. This is done due to it will drop only limited points by keeping the original pattern.

```
# Identify outliers
q1 = colomboProperData['HCHO rolling'].quantile(0.25)
q3 = colomboProperData['HCHO rolling'].quantile(0.75)
iqr = q3 - q1

lower_bound = q1 - 1.75 * iqr
upper_bound = q3 + 1.75 * iqr
```



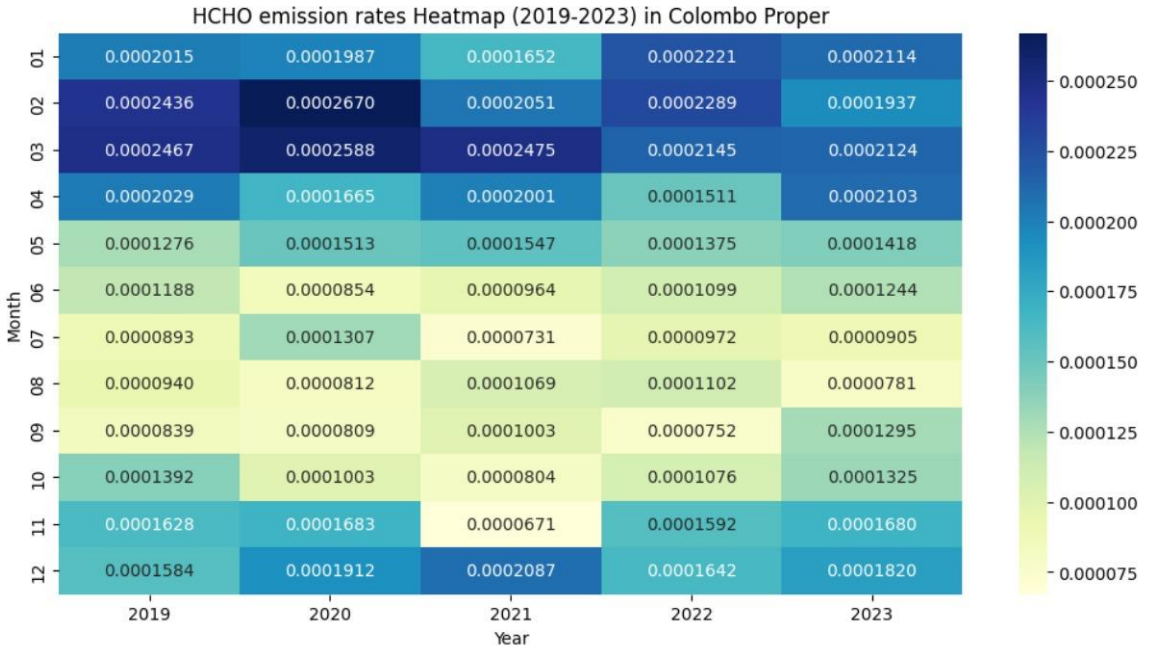
The below plots show the identified outliers data and how the distribution look like after handling the outliers.



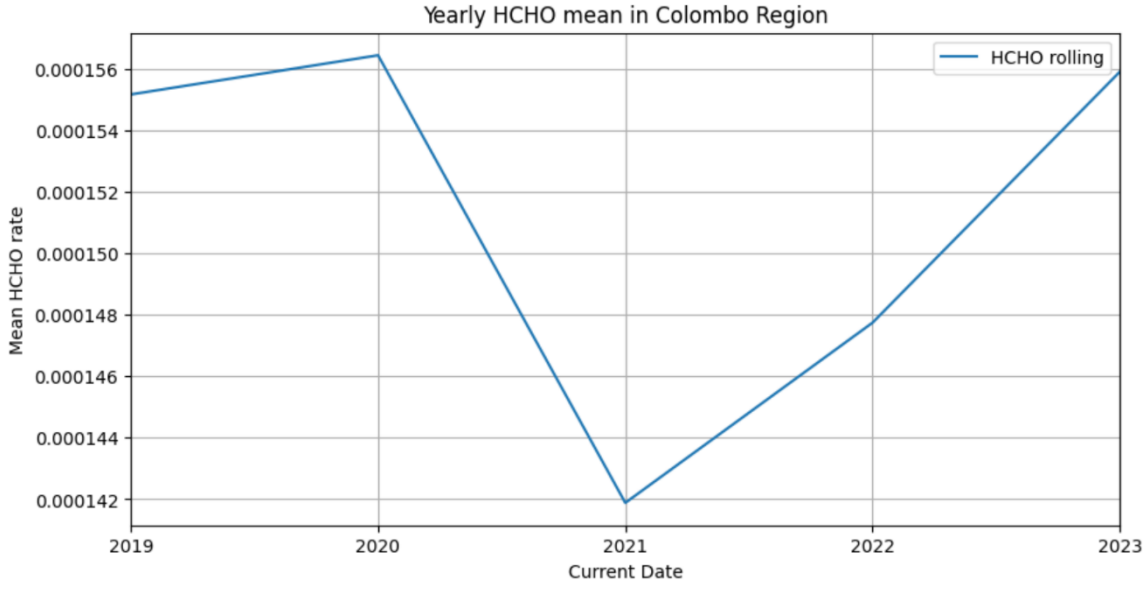


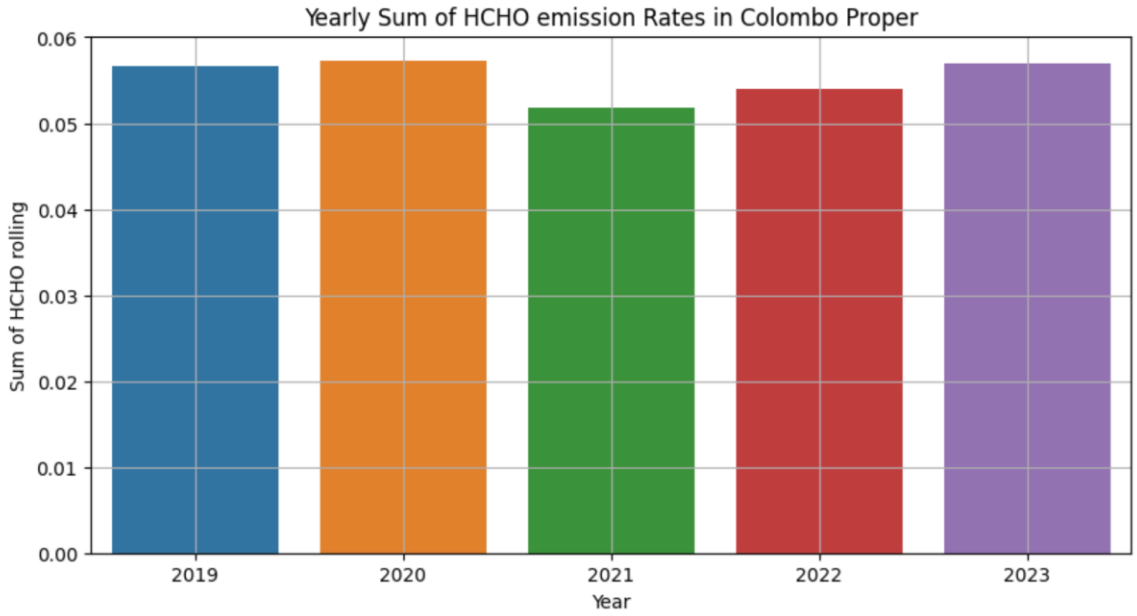
HCHO distribution insights Colombo

Then analysed how the HCHO distribution exists in the Colombo region by doing statistical calculations and using visualizations. The below heatmap shows mean HCHO emission rates in each month of the year. It depicts that when it comes to the middle months of the year, HCHO rates gradually decrease, and there is a high HCHO emission in the end and starting months of the year.

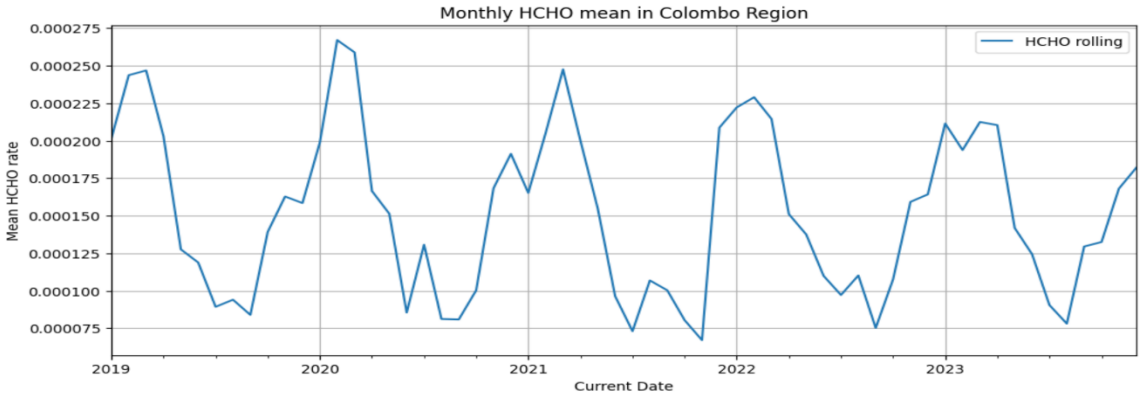


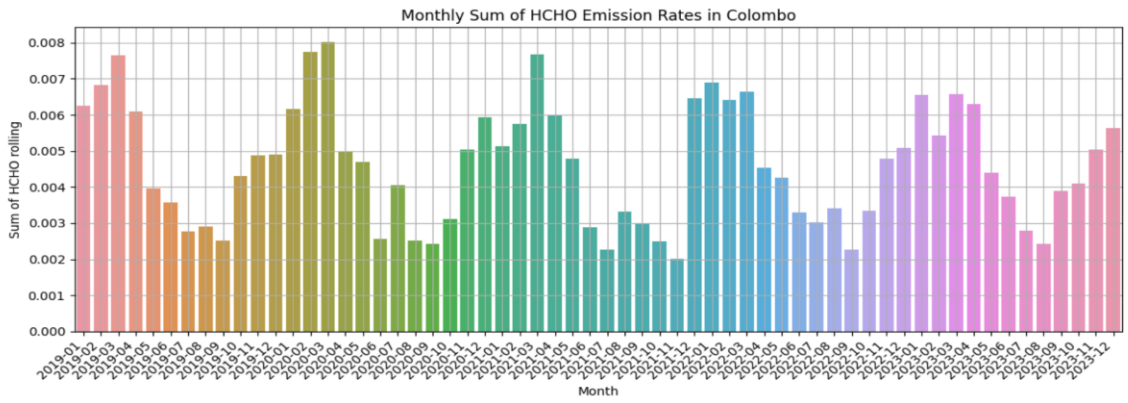
The below line chart and bar chart shows that the HCHO rate has decreased in 2021, and it has skyrocketed after year 2021.This will be discussed with the covid impact data.

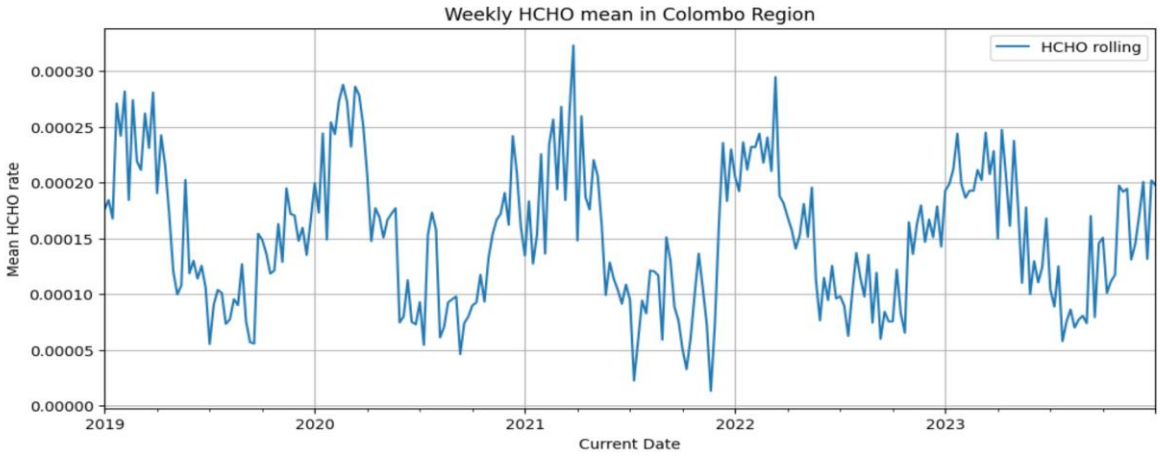


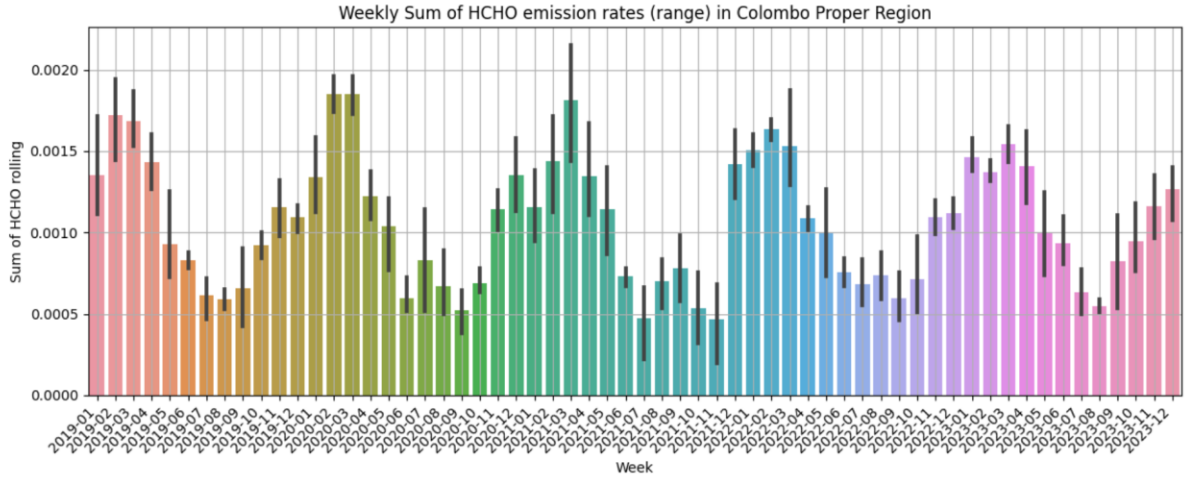


As mentioned above the below weekly and monthly plotted plots depict that there is a low HCHO emission in the middle months and high HCHO emission rate in starting and ending months of the year.





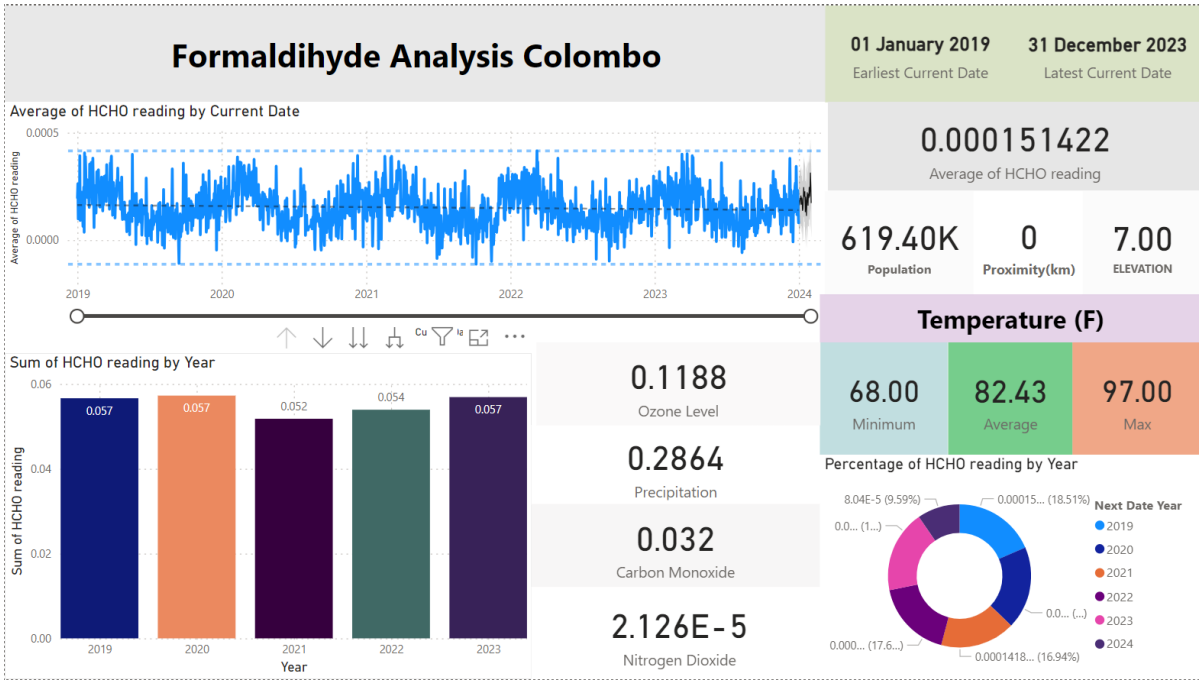


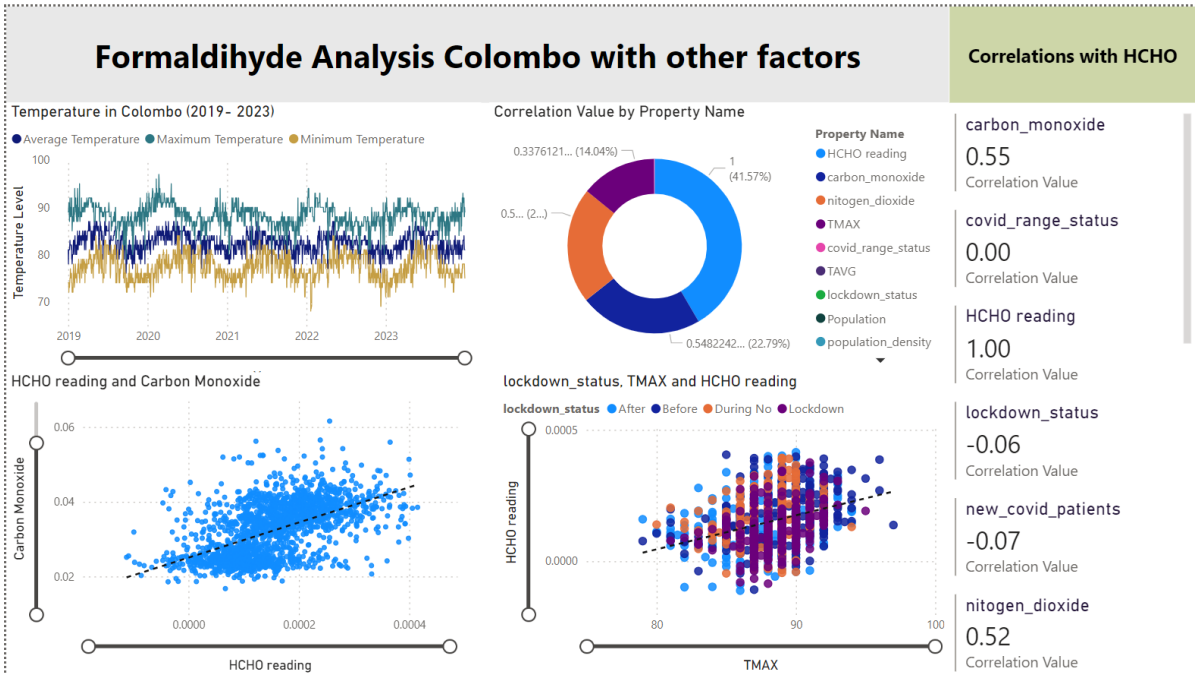


Statistical Measurements taken for Colombo Region

The below table shows the statistical calculations done for Colombo Region.

	HCHO reading
count	1826.000000
mean	0.000151
std	0.000086
min	-0.000112
25%	0.000092
50%	0.000144
75%	0.000209
max	0.000415

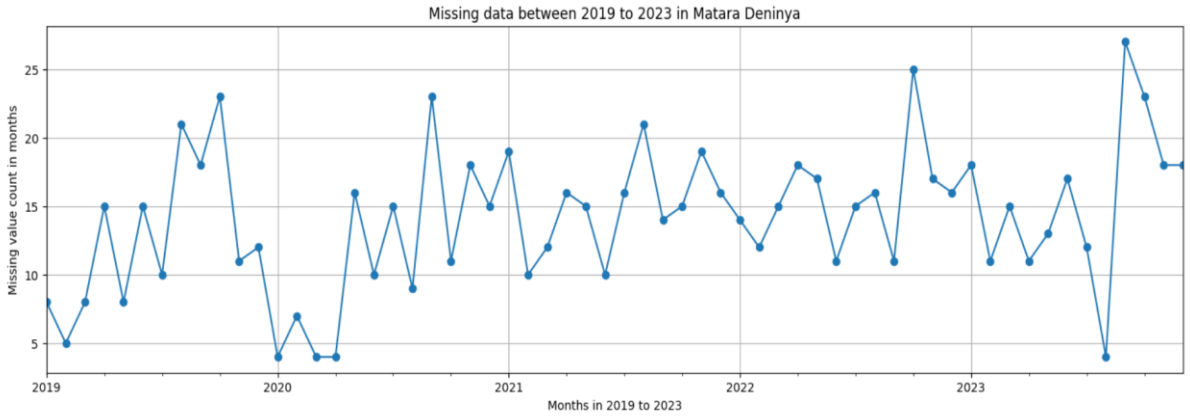

Formaldihyde Analysis Colombo
01 January 2019
Earliest Current Date
31 December 2023
Latest Current Date
Average of HCHO reading by Current Date
0.0005
0.0000
Average of HCHO reading
2019
2020
2021
2022
2023
2024
0.000151422
Average of HCHO reading
619.40K
Population
0
Proximity(km)
7.00
ELEVATION
Temperature (F)
68.00
Minimum
82.43
Average
97.00
Max
Sum of HCHO reading by Year
0.06
0.04
0.02
0.00
Sum of HCHO reading
0.057
0.057
0.052
0.054
0.057
2019
2020
2021
2022
2023
Year
0.1188
Ozone Level
0.2864
Precipitation
0.032
Carbon Monoxide
2.126E-5
Nitrogen Dioxide
Percentage of HCHO reading by Year
8.04E-5 (9.59%)
0.00015... (18.51%)
0.0... (1...)
0.0... (...)
0.000... (17.6...)
0.0001418... (16.94%)
Next Date Year
2019
2020
2021
2022
2023
2024

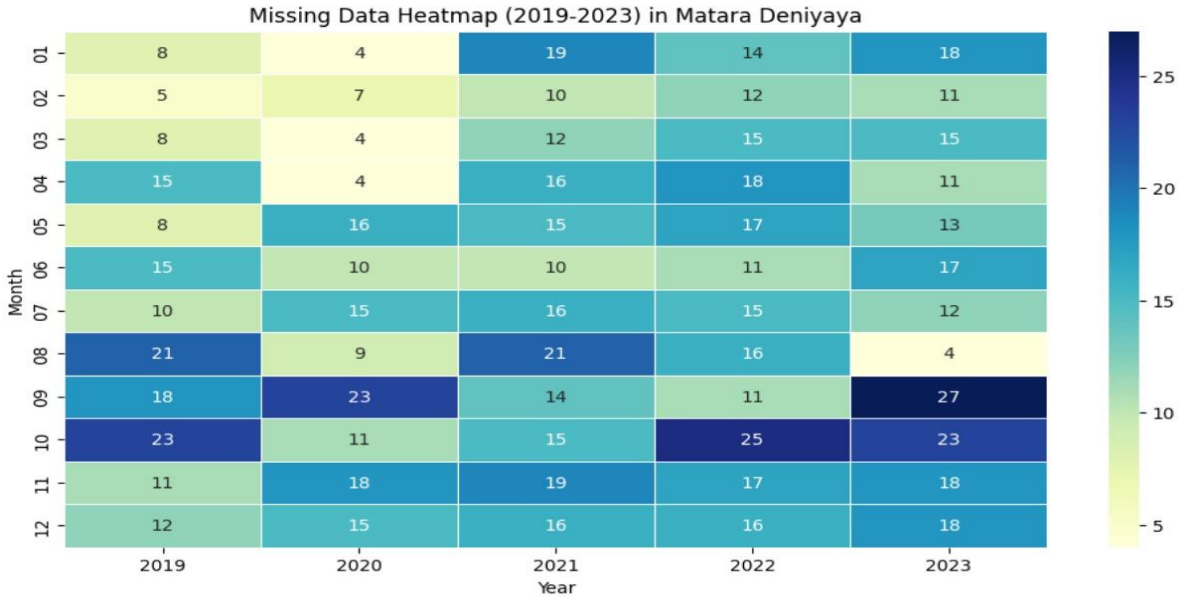

Formaldihyde Analysis Colombo with other factors
Correlations with HCHO
Temperature in Colombo (2019- 2023)
Average Temperature
Maximum Temperature
Minimum Temperature
100
90
80
70
Temperature Level
2019
2020
2021
2022
2023
Correlation Value by Property Name
0.3376121... (14.04%)
1 (41.57%)
0.5... (2...)
0.5482242... (22.79%)
Property Name
HCHO reading
carbon_monoxide
nitogen_dioxide
TMAX
covid_range_status
TAVG
lockdown_status
Population
population_density
HCHO reading and Carbon Monoxide
0.06
0.04
0.02
Carbon Monoxide
0.0000
0.0002
0.0004
HCHO reading
lockdown_status, TMAX and HCHO reading
lockdown_status
After
Before
During No
Lockdown
0.0005
0.0000
HCHO reading
80
90
100
TMAX
carbon_monoxide
0.55
Correlation Value
covid_range_status
0.00
Correlation Value
HCHO reading
1.00
Correlation Value
lockdown_status
-0.06
Correlation Value
new_covid_patients
-0.07
Correlation Value
nitogen_dioxide
0.52
Correlation Value

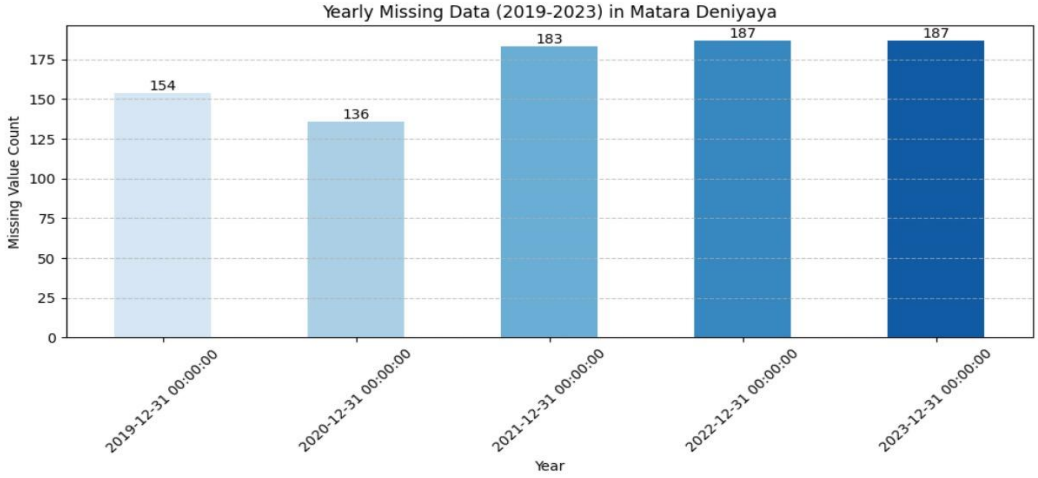
Matara Region Formaldehyde Distribution Analysis

Data Preprocessing Deniyaya, Matara

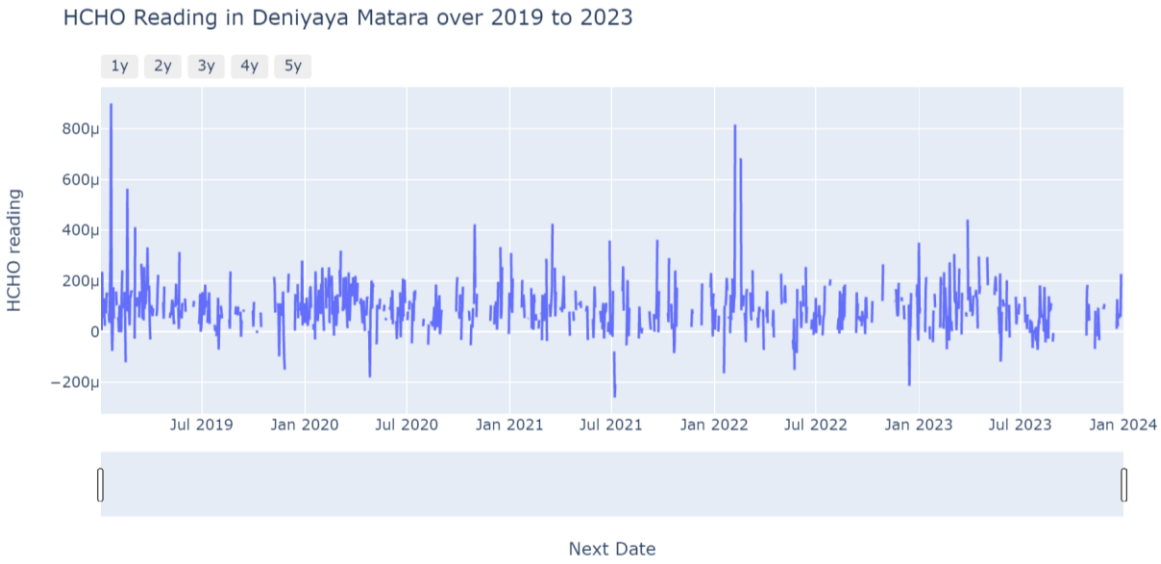
When missing values checked in Matara Deniyaya Region, there were 847 null values in the dataset out of 1826 values. It is approximately 46 percent of the dataset. The below line chart, heatmap and bar chart shows how missing values are distributed in each year.



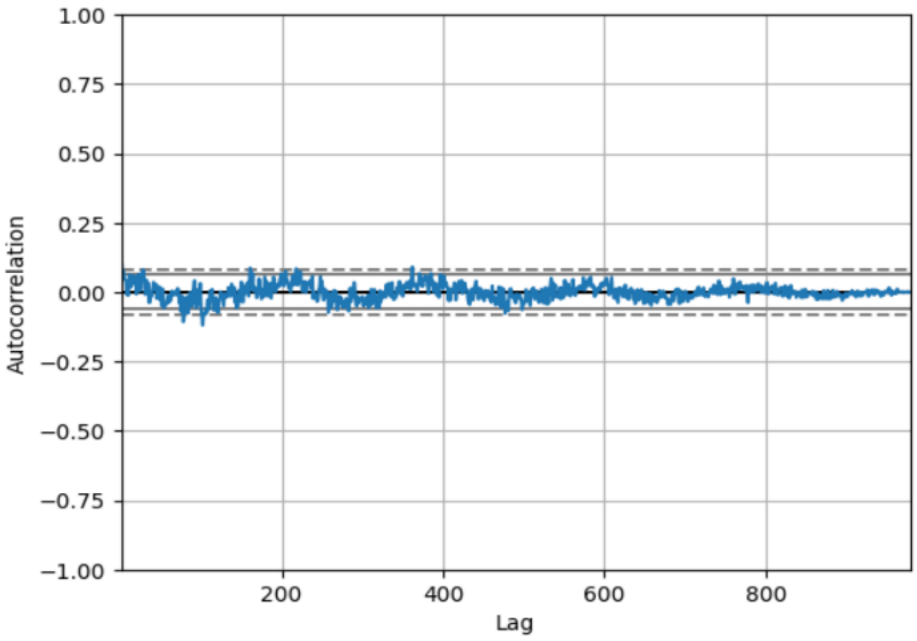




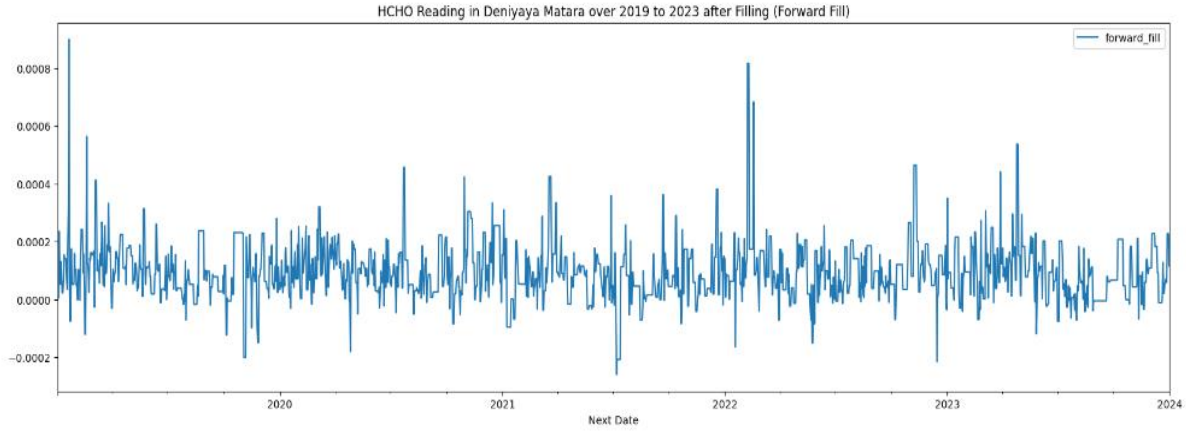
The above visualizations show that there are months that consist of 27 missing values. It can be identified as a limitation of the dataset since it is very difficult to identify the original seasonality pattern in the dataset to handle missing values. The below plot shows Matara HCHO distribution before handling missing values.

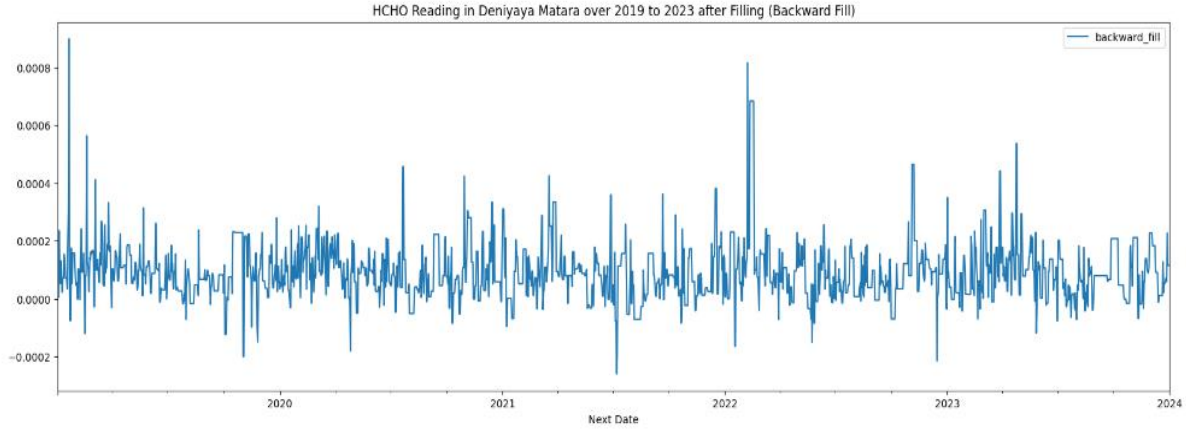


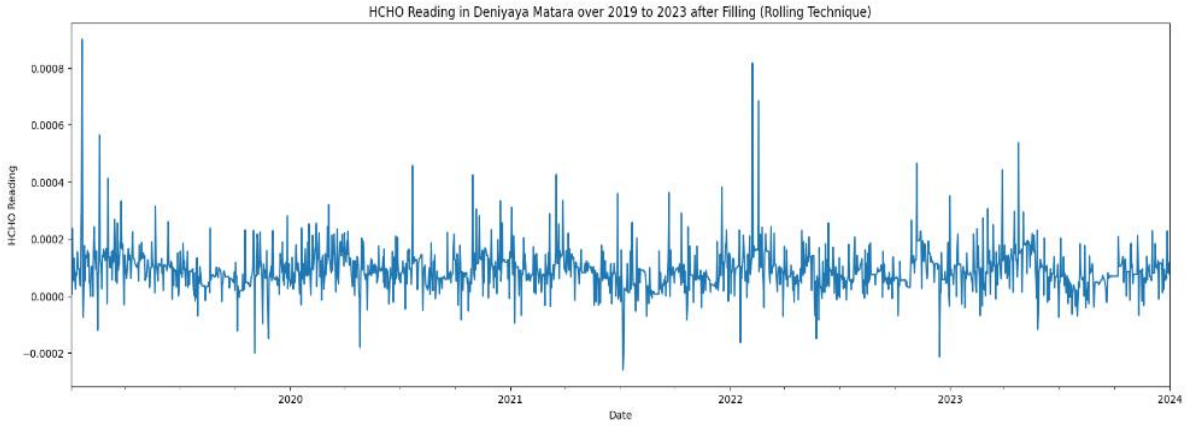
Matara Deniyaya Auto-correlation plot shows that the dataset is more likely to be stationary and there is a small seasonal pattern as well. Since there are large number of null values consist in the above diagram, the finalized null value handled data set is created using a window size of 13 (large window size). Forward filling and backward filling also checked for null value handling.



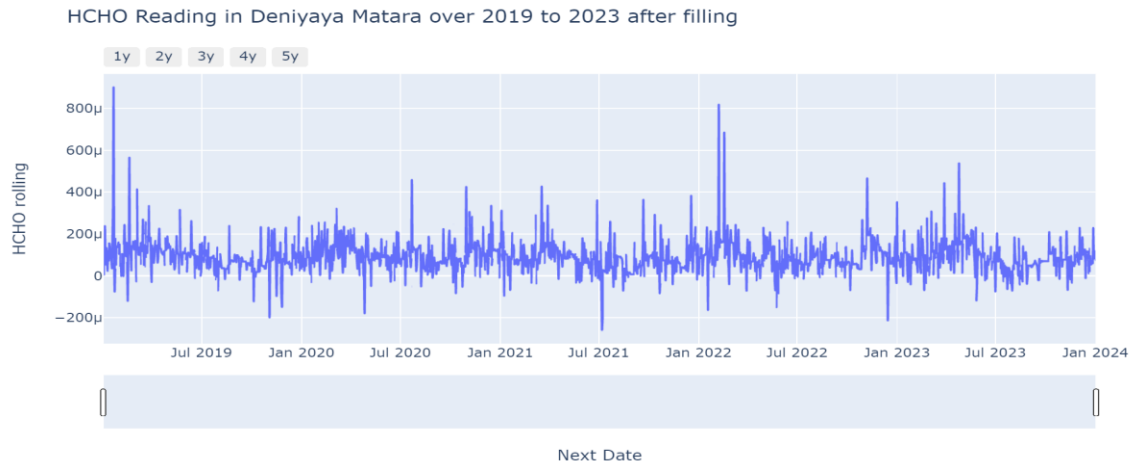
The below plots show that rolling technique is able to keep up the original pattern of the dataset.





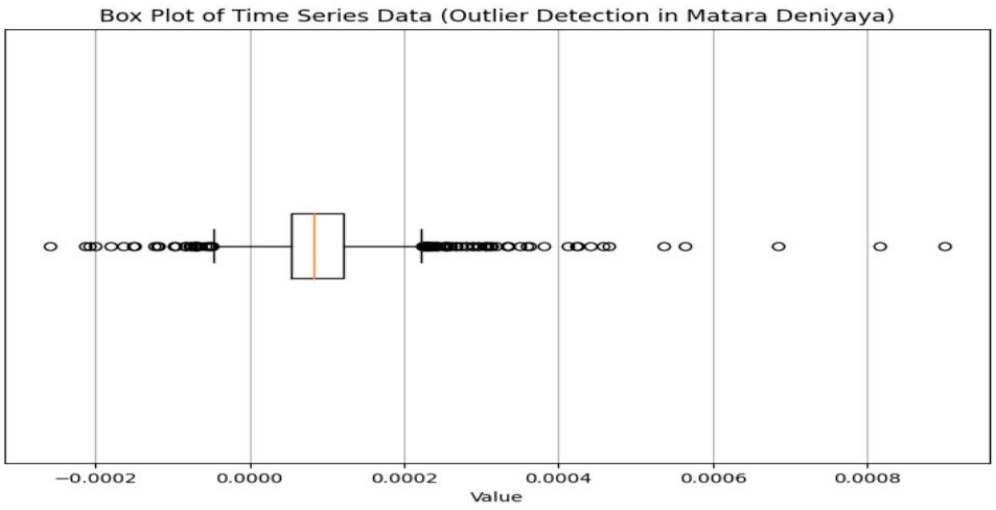


The below plot shows how Matara Deniyaya HCHO distribution looks like after handling null values with rolling window 13.

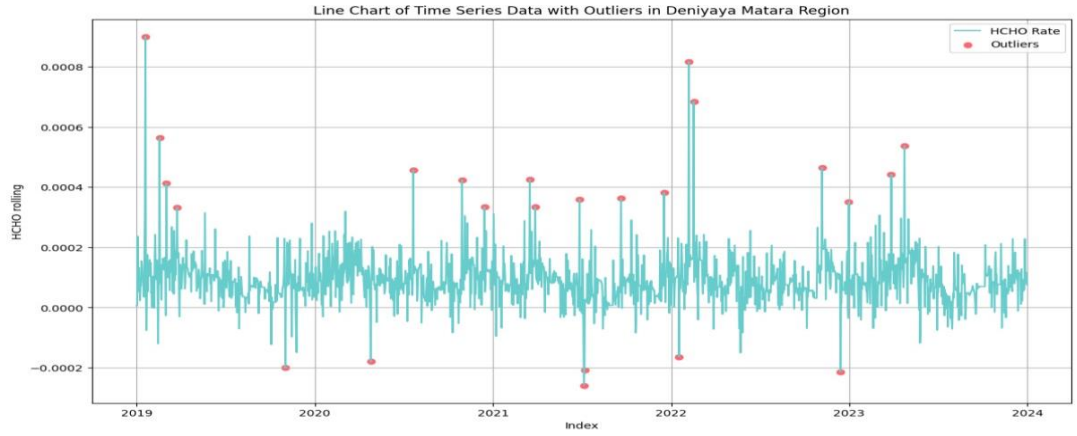


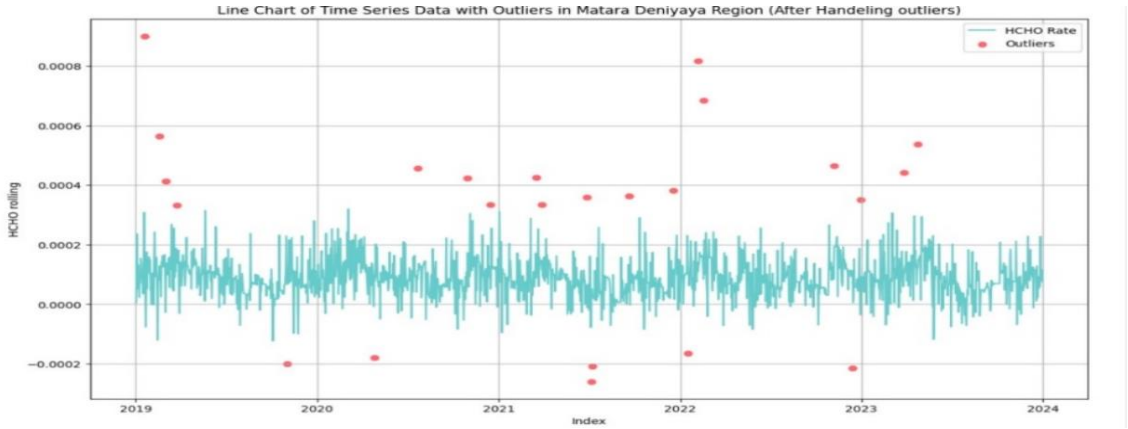
The below box plot shows the outliers given based on the quartiles, but the outliers are removed using a threshold value of 3 from the quartile range.

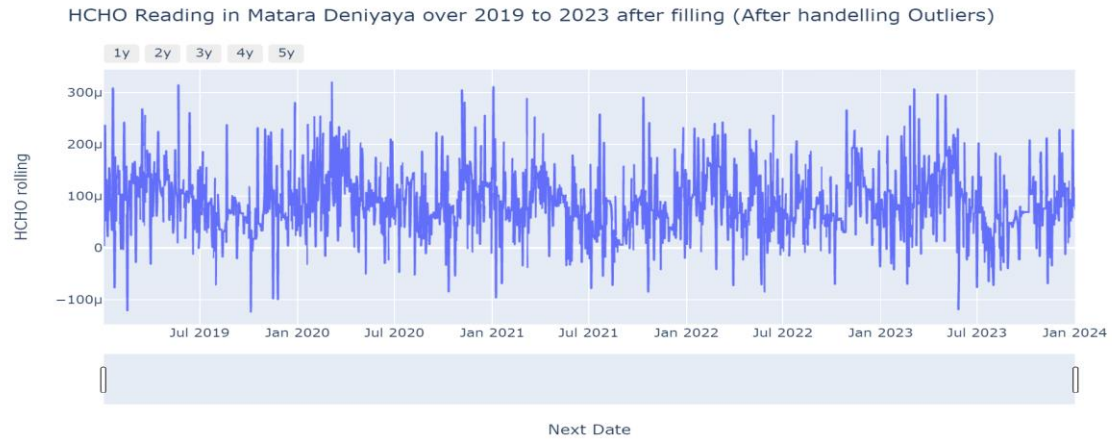
```
# Identify outliers
q1 = MataraDeniyayaData['HCHO rolling'].quantile(0.25)
q3 = MataraDeniyayaData['HCHO rolling'].quantile(0.75)
iqr = q3 - q1

lower_bound = q1 - 3 * iqr
upper_bound = q3 + 3 * iqr
```



The below plots show the identified outliers in Matara Region and how the distribution look like after handling the outliers.



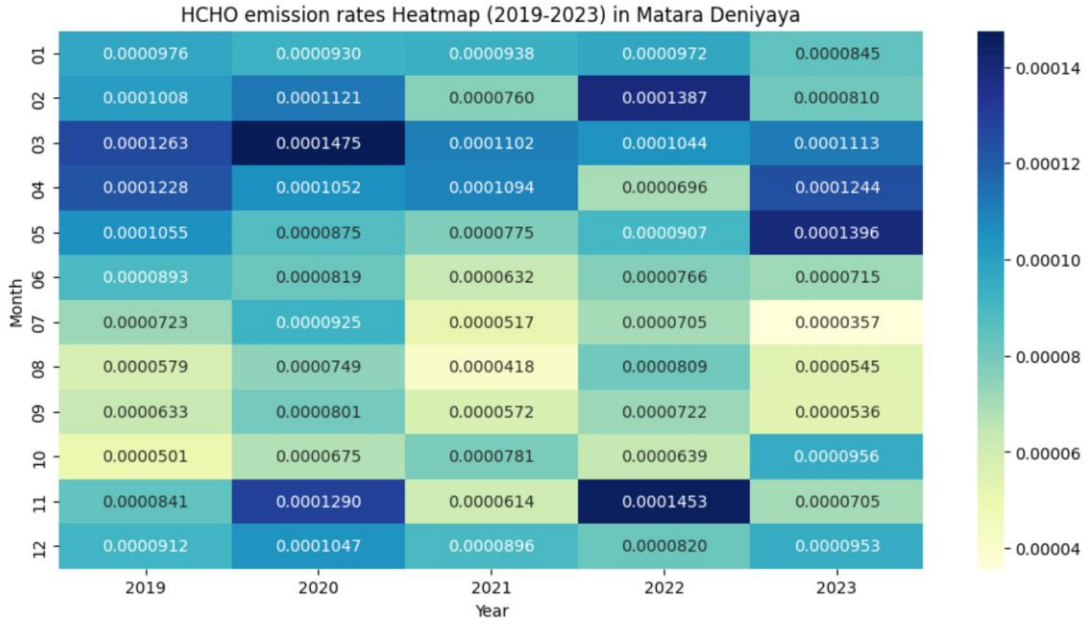


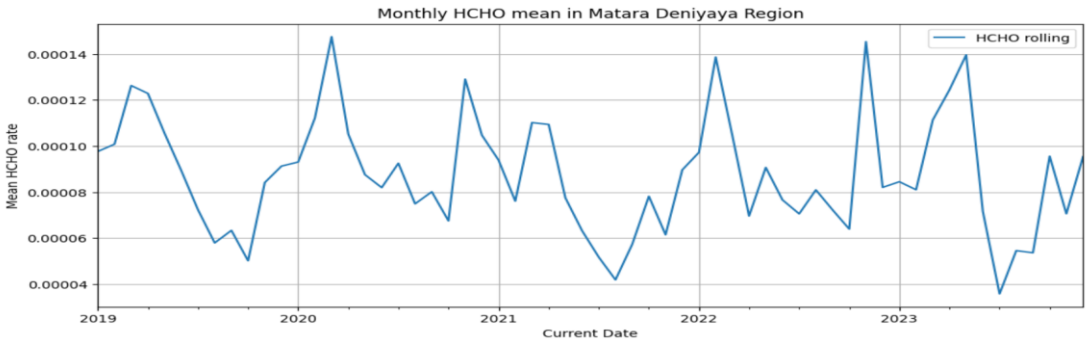


HCHO distribution Insights Matara

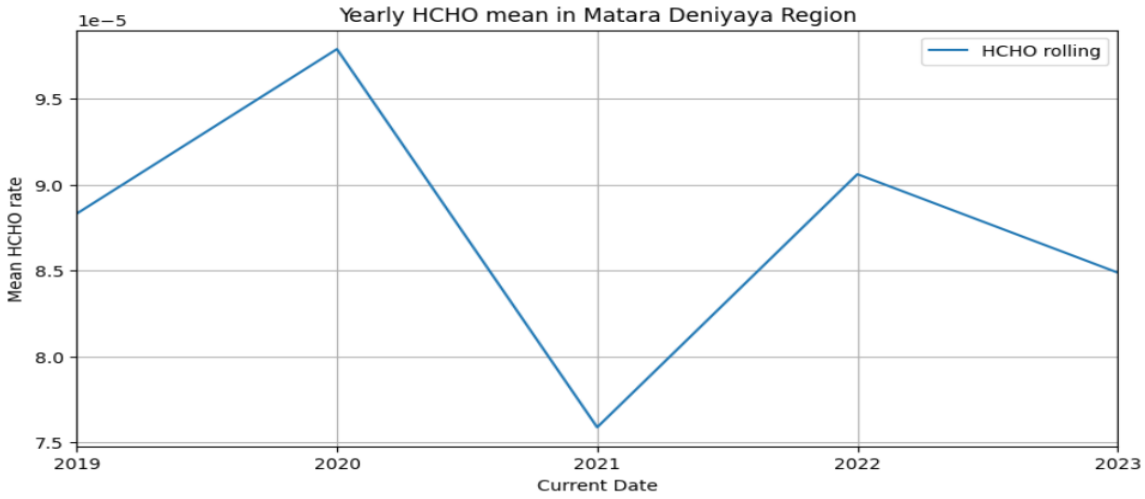
The HCHO distribution in the Matara Deniyaya region was then analysed using statistical calculations and visualizations, and the below plots show how HCHO rates vary monthly, yearly, and weekly. As in Colombo, it shows that the HCHO rate decreases in the mid-months of the year, and it reports its maximum value at the beginning and end of the year.

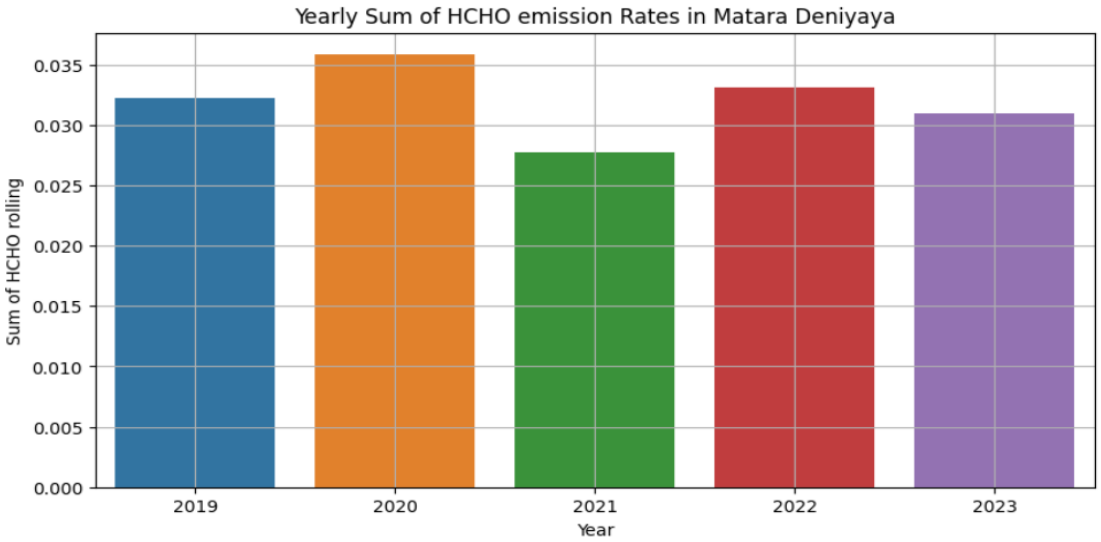
The below heatmap shows how mean HCHO rates in each month is distributed.



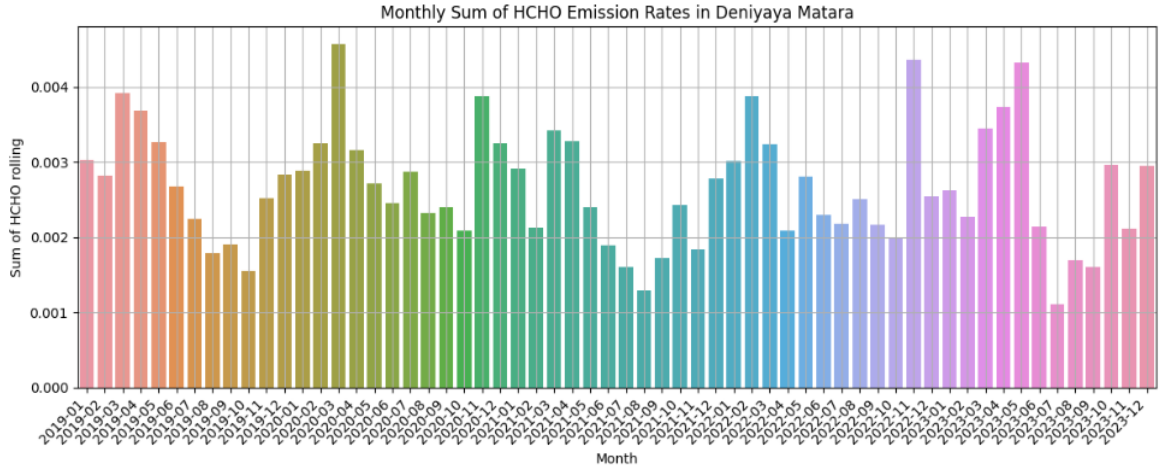


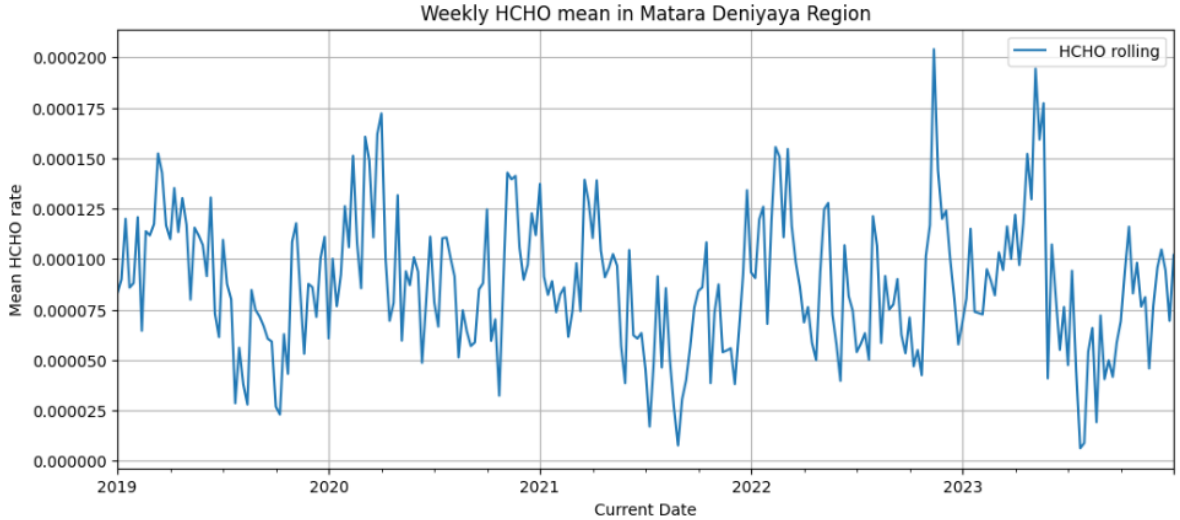
There is a fluctuation of the mean HCHO rate in each year in Matara region. It has reported the lowest HCHO rate in 2021 and highest in 2020.

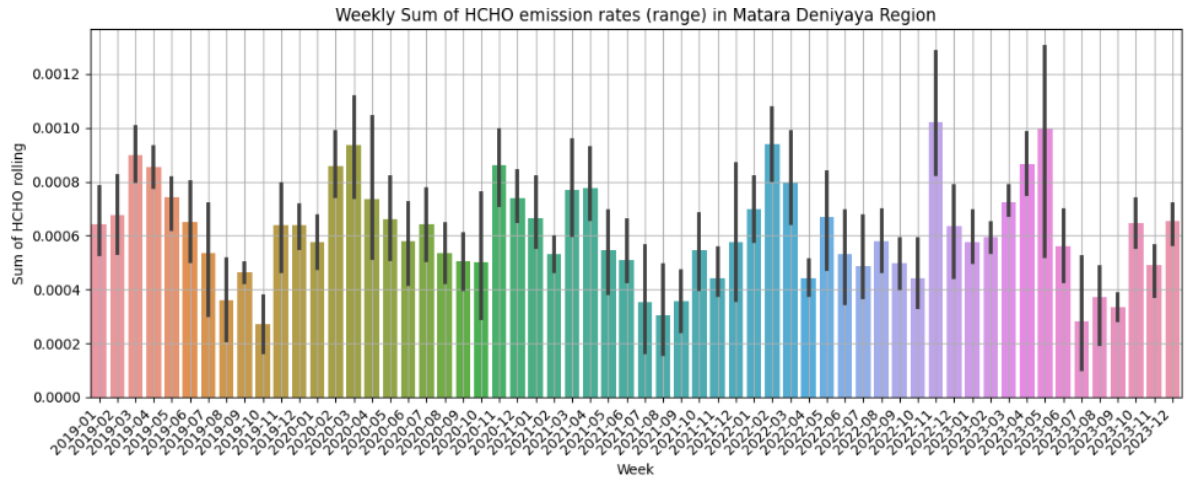




The below visualizations show how weekly and monthly HCHO rates are distributed in Matara Deniyaya Region.

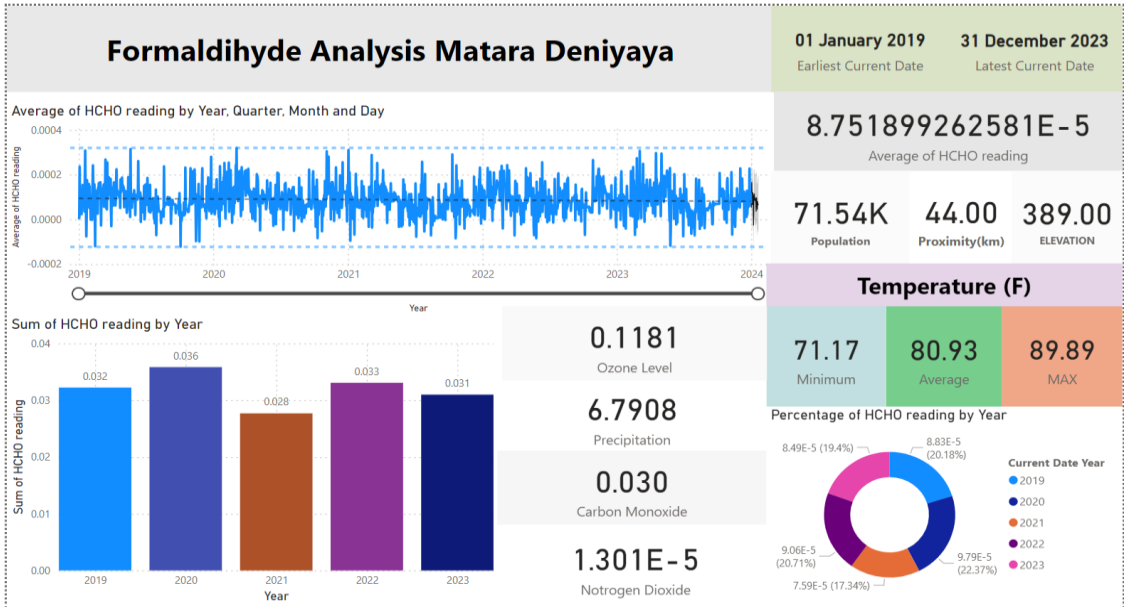


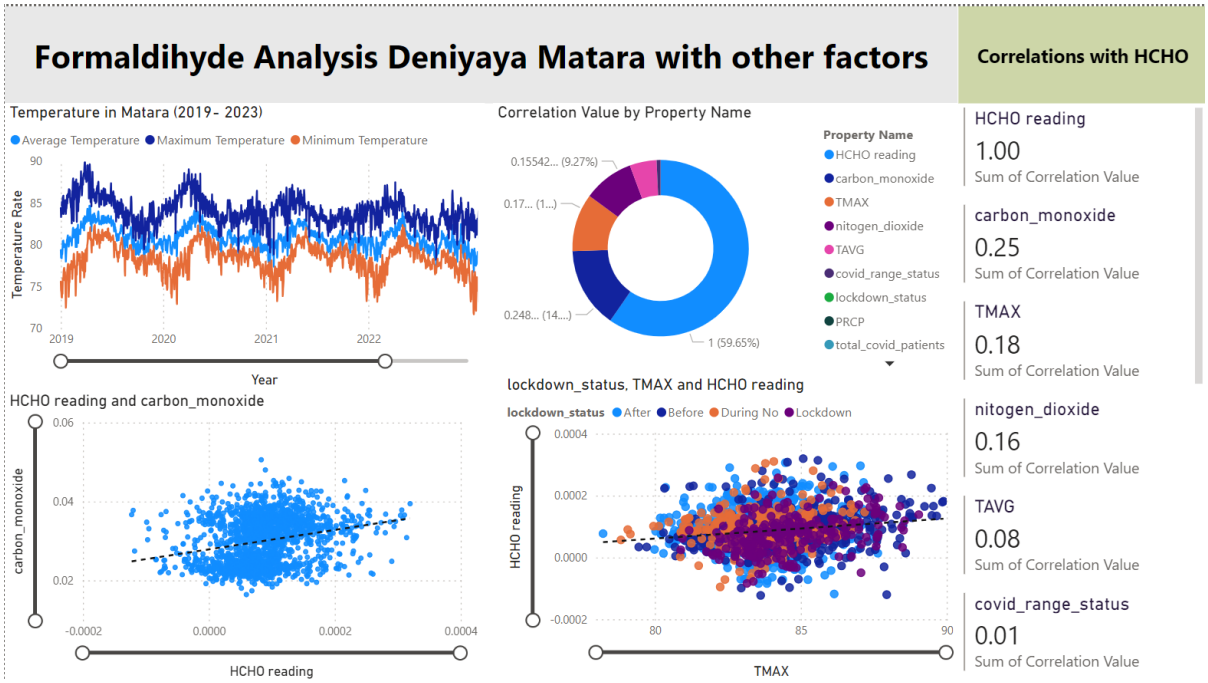




Statistical Measurements done for Matara Region

	HCHO reading
count	1826.000000
mean	0.000088
std	0.000060
min	-0.000123
25%	0.000054
50%	0.000083
75%	0.000120
max	0.000320

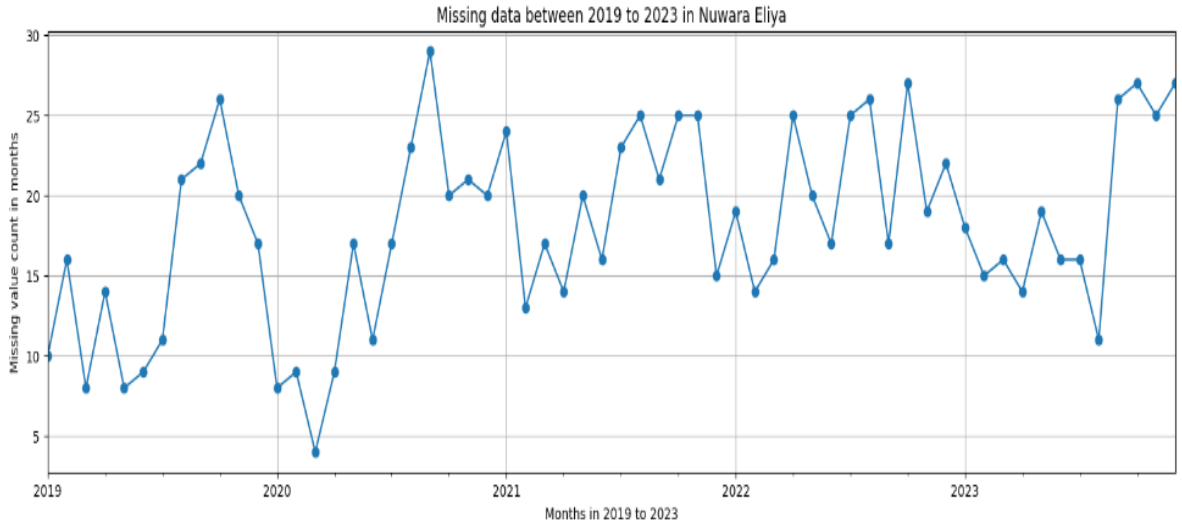


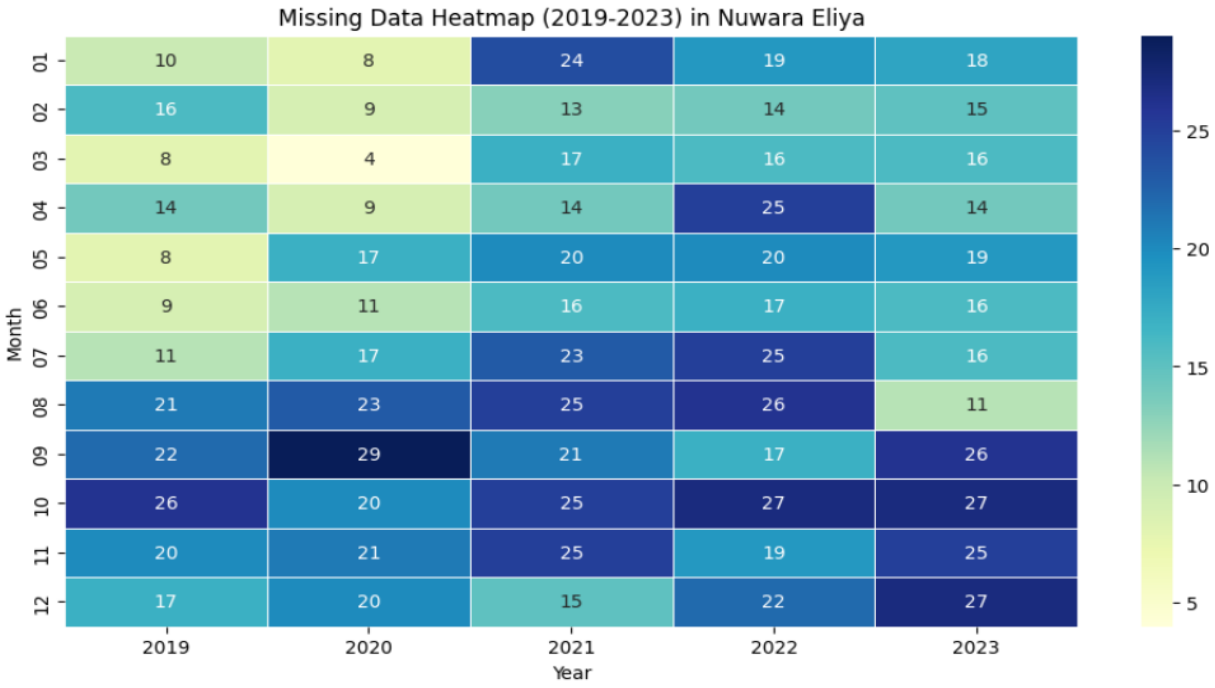


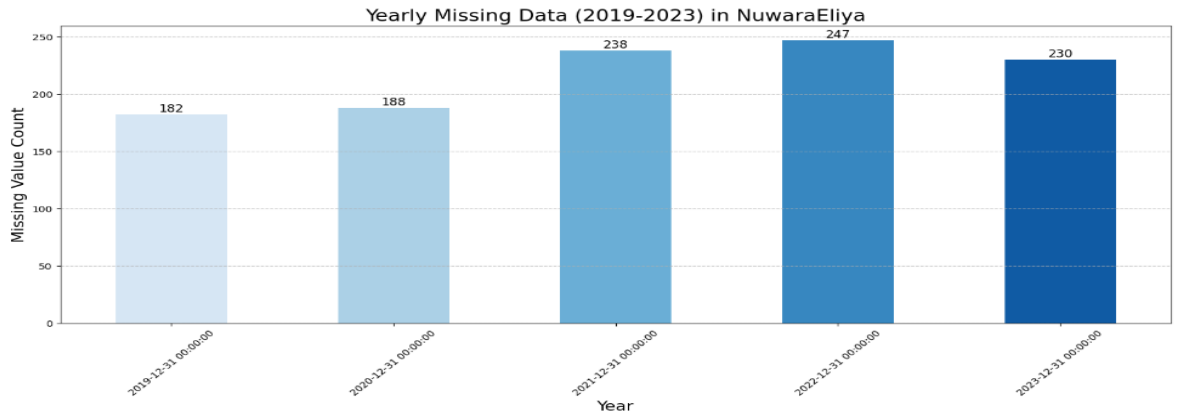
Nuwara Eliya Region Formaldehyde Distribution Analysis

Data Preprocessing Nuwara Eliya

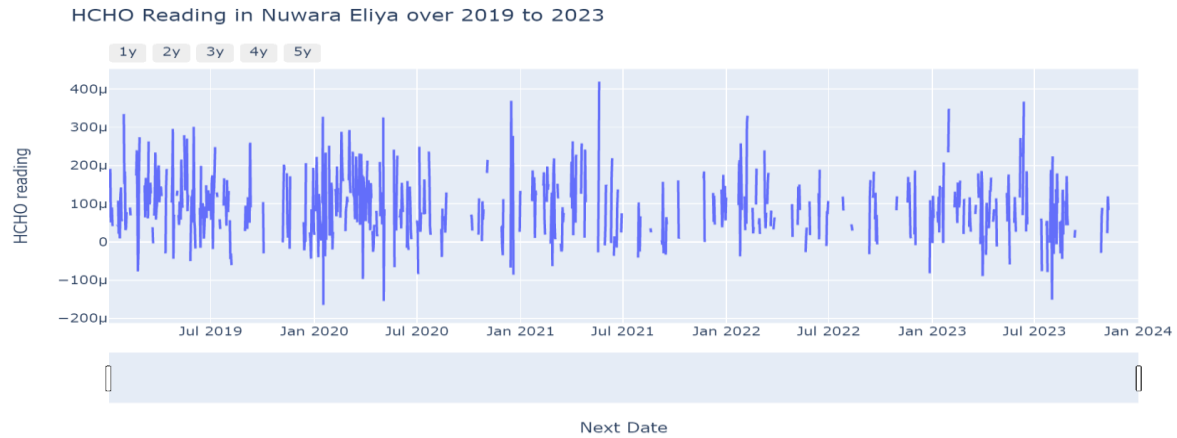
Nuwara Eliya dataset consisted of 1085 null values out of 1826 values, which can be considered as a limitation to identify its original seasonality due to it missing more than 50% of its data. The following plots show the null value distribution.



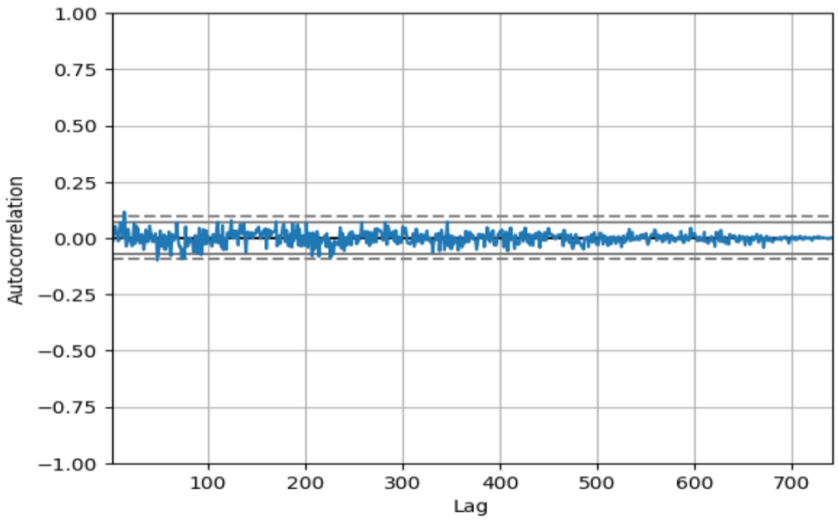




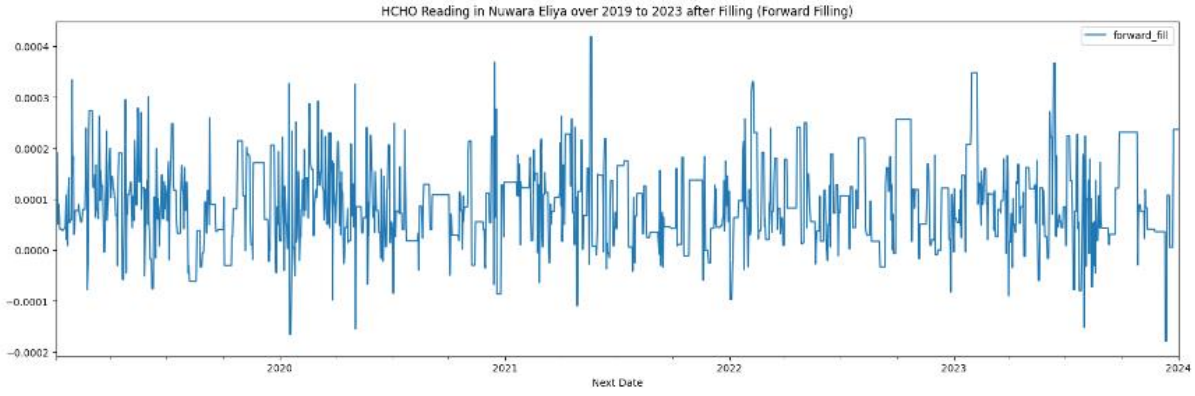
There are some months that contain 29 null values as well. The below plot shows the Nuwara Eliya HCHO distribution before handling null values.

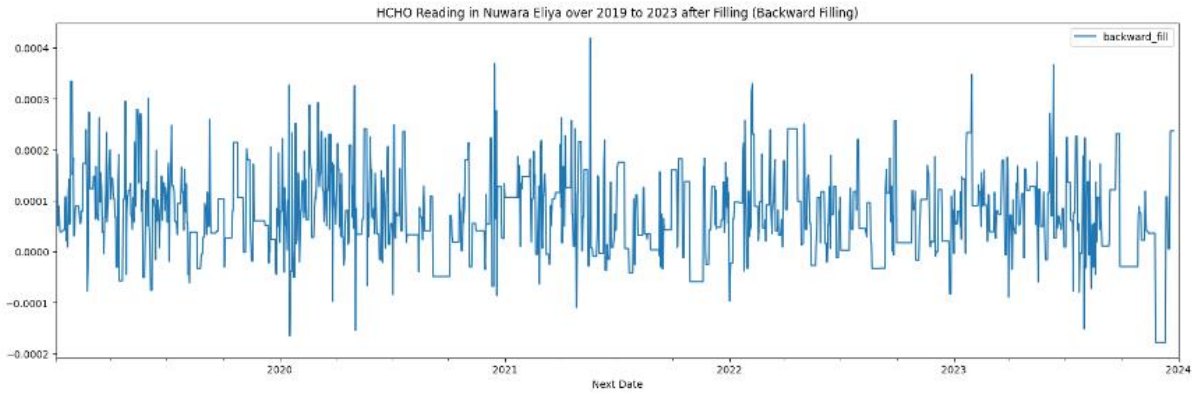


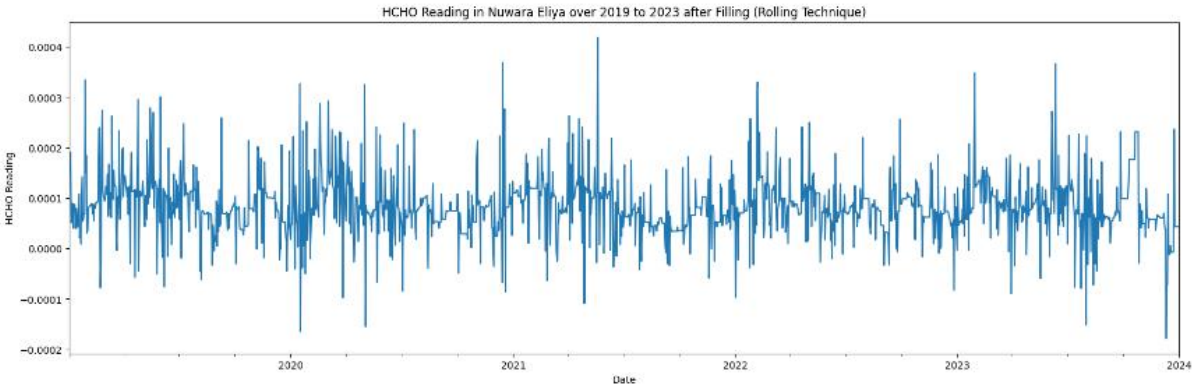
Since most values contain null values, it is very difficult to find seasonality before handling null values. This auto-correlation plot shows it is more likely to be stationary. However, the window size of 30 is used to fill in the null values in the dataset using the rolling technique because some months contain 29 missing values.



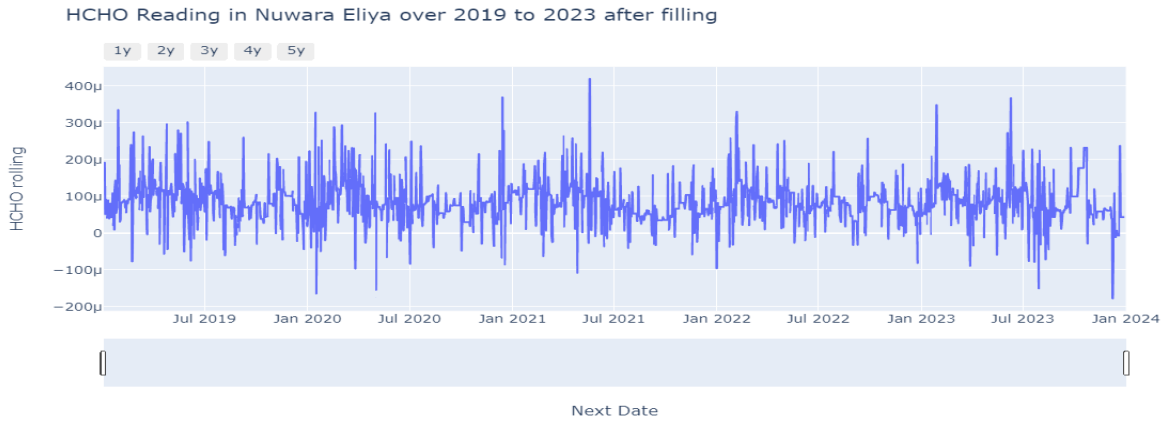
The below line charts show how the HCHO distribution of Nuwara Eliya looks when handling missing values with rolling, forward filling, and backward filling techniques.





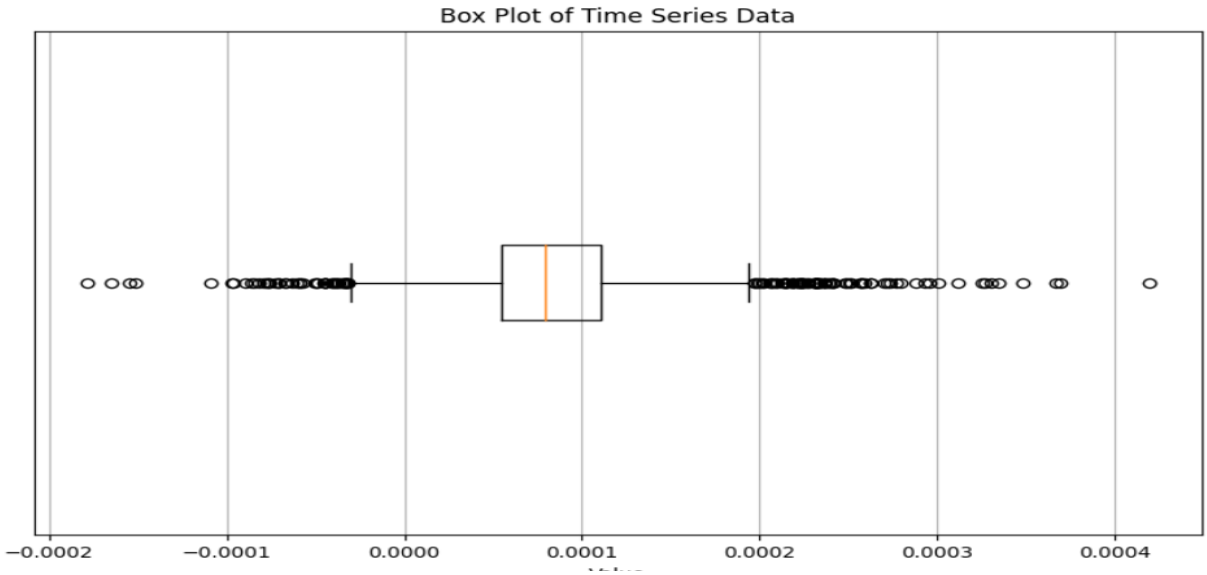


Since Rolling window applied chart maintains a seasonal pattern, it has been considered as the final HCHO distribution for Nuwara Eliya. The finalized null value handled chart shows below.

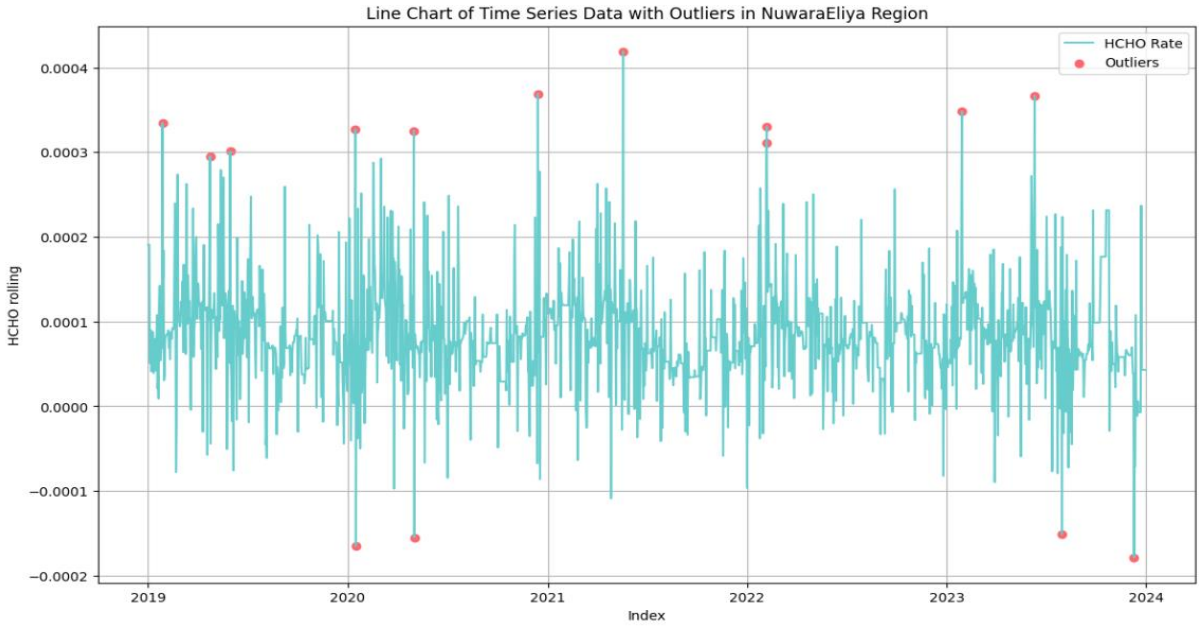


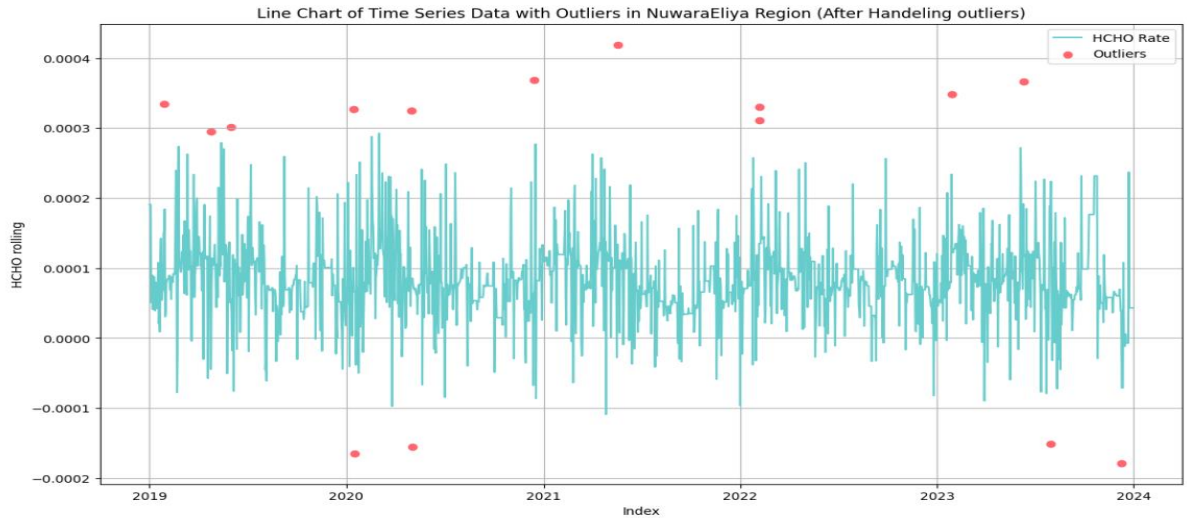
The below box plot shows the outliers given based on the quartiles, but the outliers are removed using a threshold value of 3.25 from the quartile range.

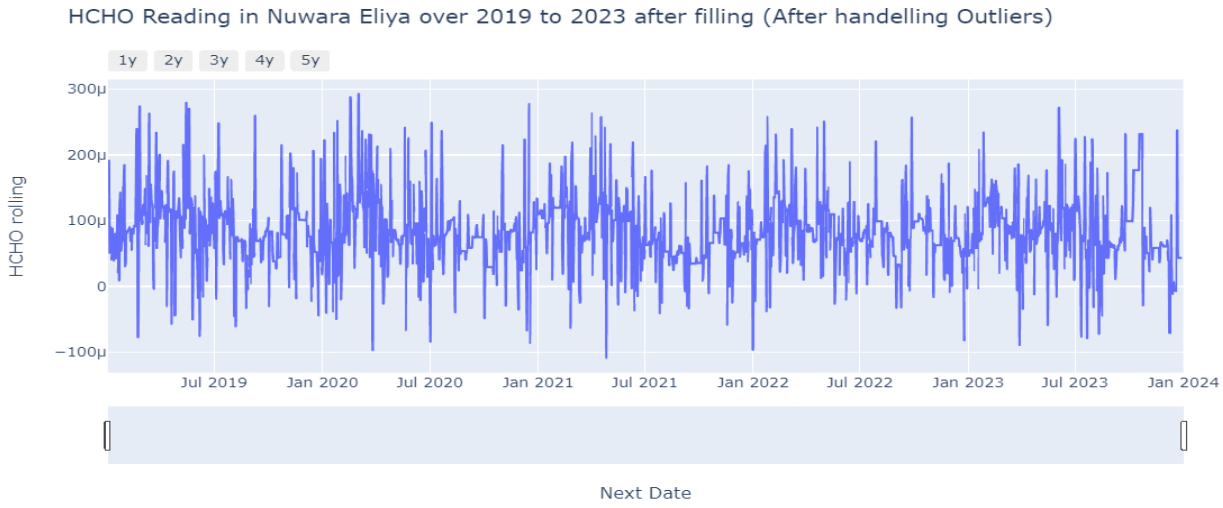
```
# Identify outliers
q1 = NuwaraEliyaData['HCHO rolling'].quantile(0.25)
q3 = NuwaraEliyaData['HCHO rolling'].quantile(0.75)
iqr = q3 - q1

lower_bound = q1 - 3.25 * iqr
upper_bound = q3 + 3.25 * iqr
```



The below plots show the identified outliers in Nuwara Eliya and how the distribution look like after handling the outliers.

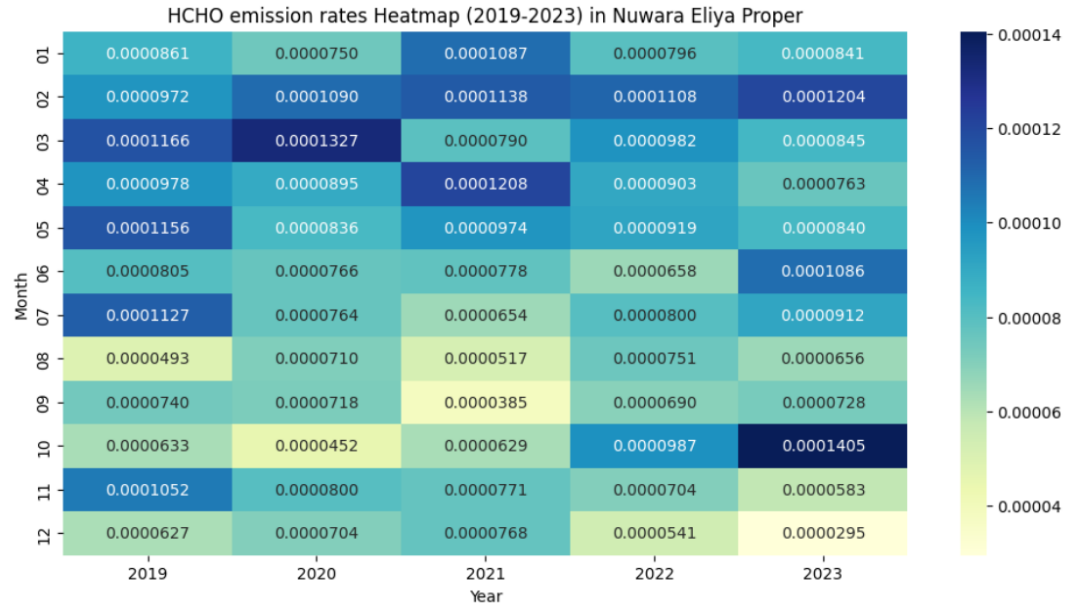


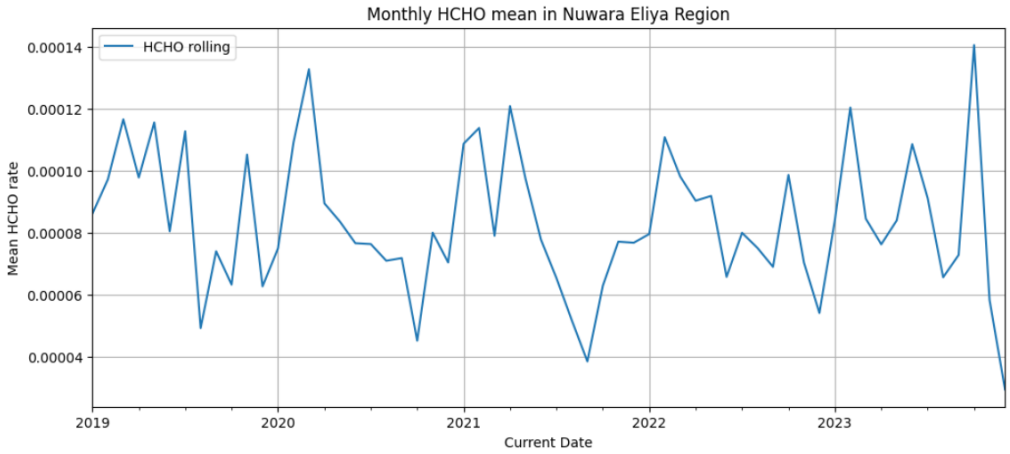




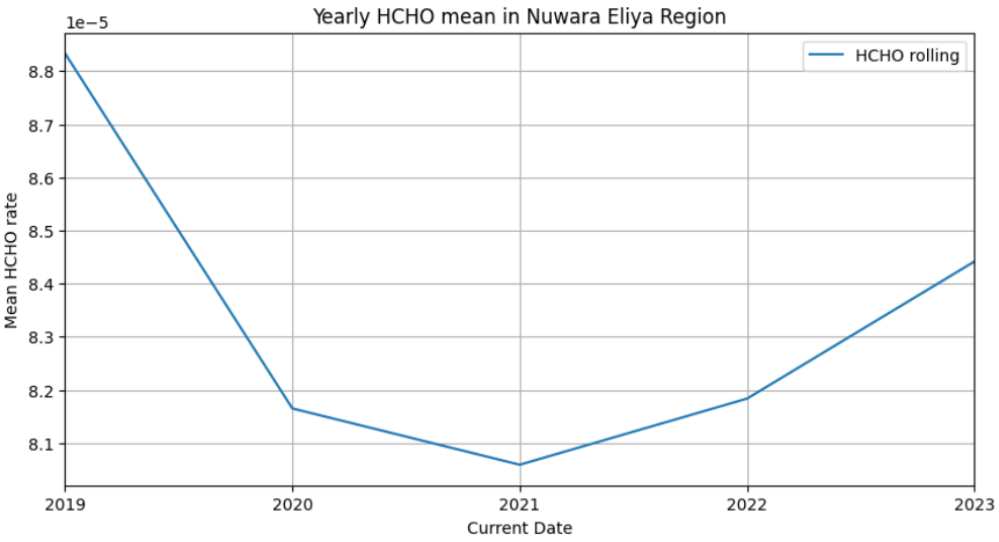
HCHO distribution Insights Nuwara Eliya

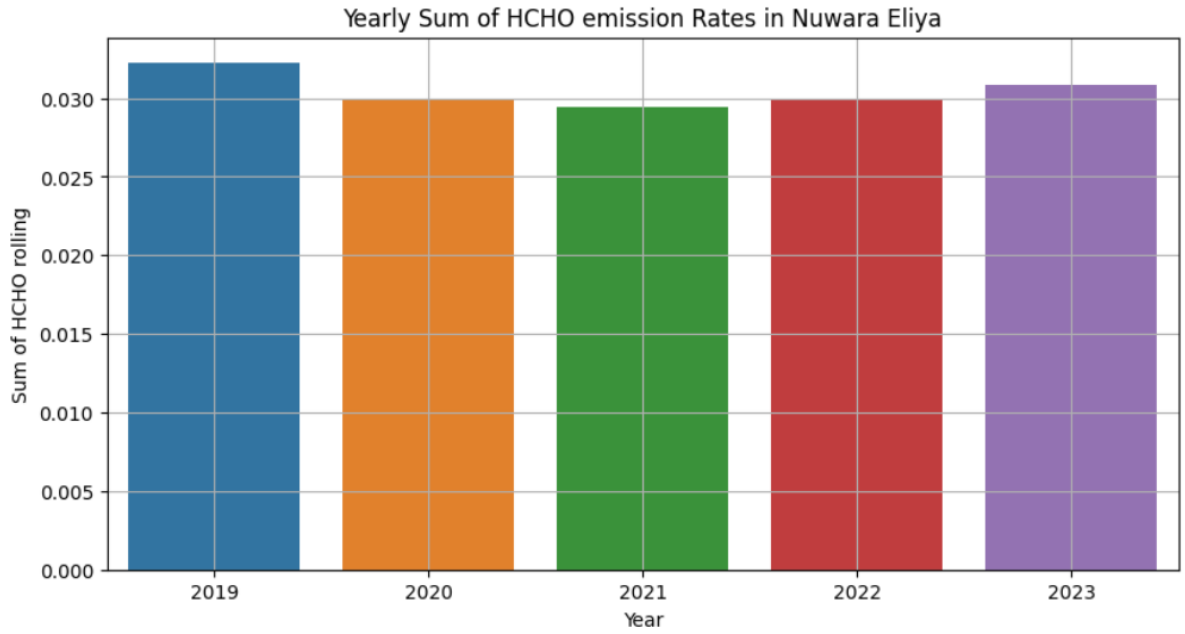
The below visualizations and statistical calculations show how Nuwara Eliya HCHO distribution varies its values monthly, weekly, and yearly. In Nuwara Eliya region it is difficult to find a complete seasonality, but it shows that mean HCHO value decreases in the mid months of the year. The below heatmap and line chart shows the mean HCHO distribution of each month.



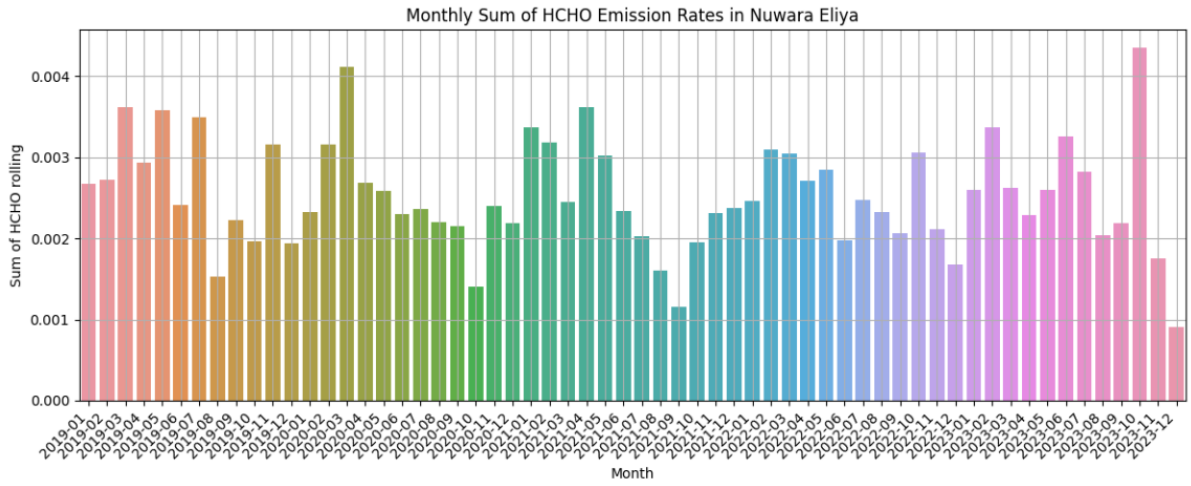


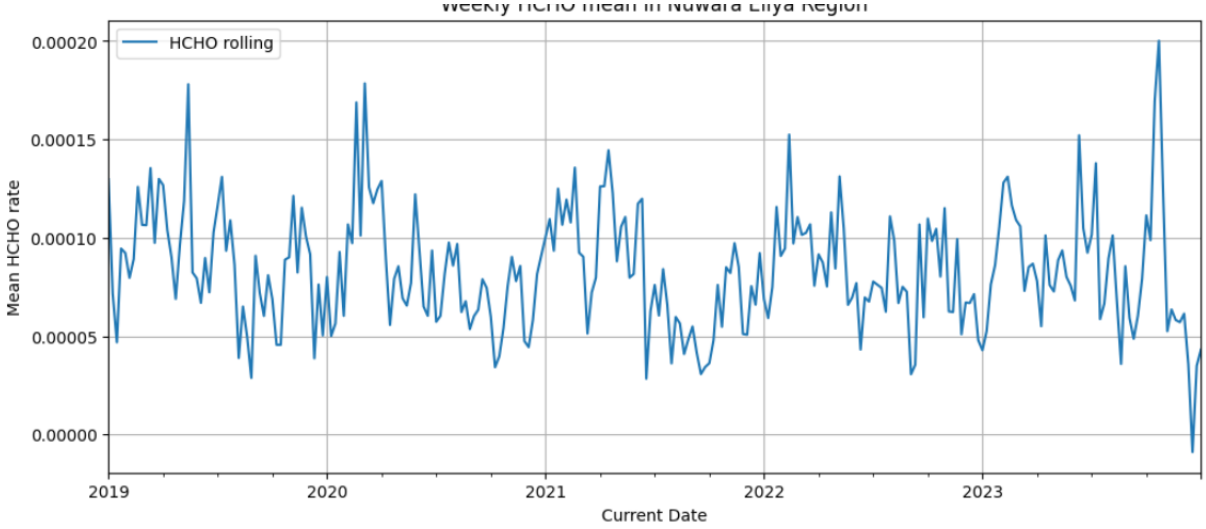
The yearly HCHO distribution has decreased up to year 2021 and it has started skyrocketing after 2021.

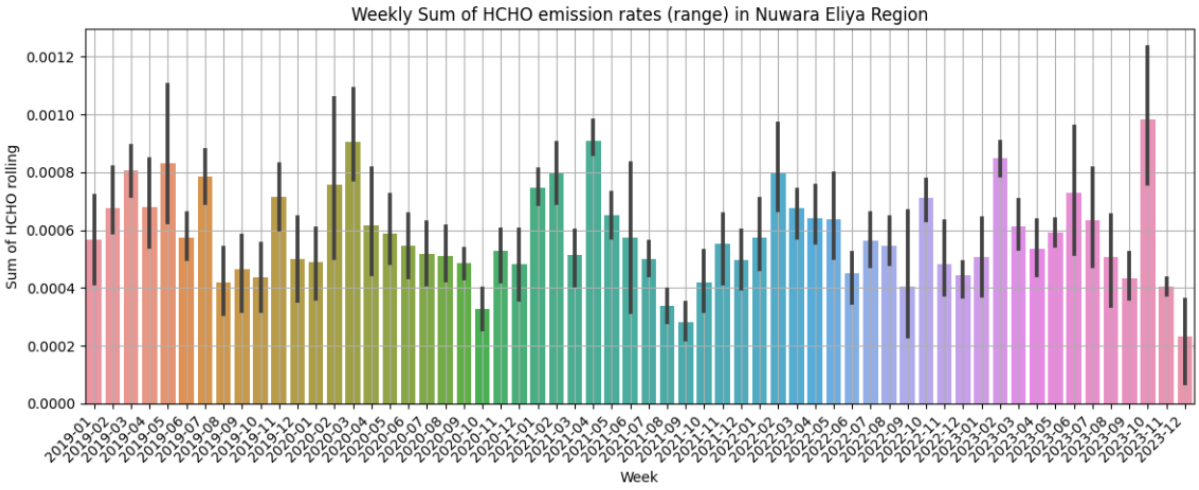




The below visualizations show how weekly and monthly HCHO rates are distributed in Nuwara Eliya Region.

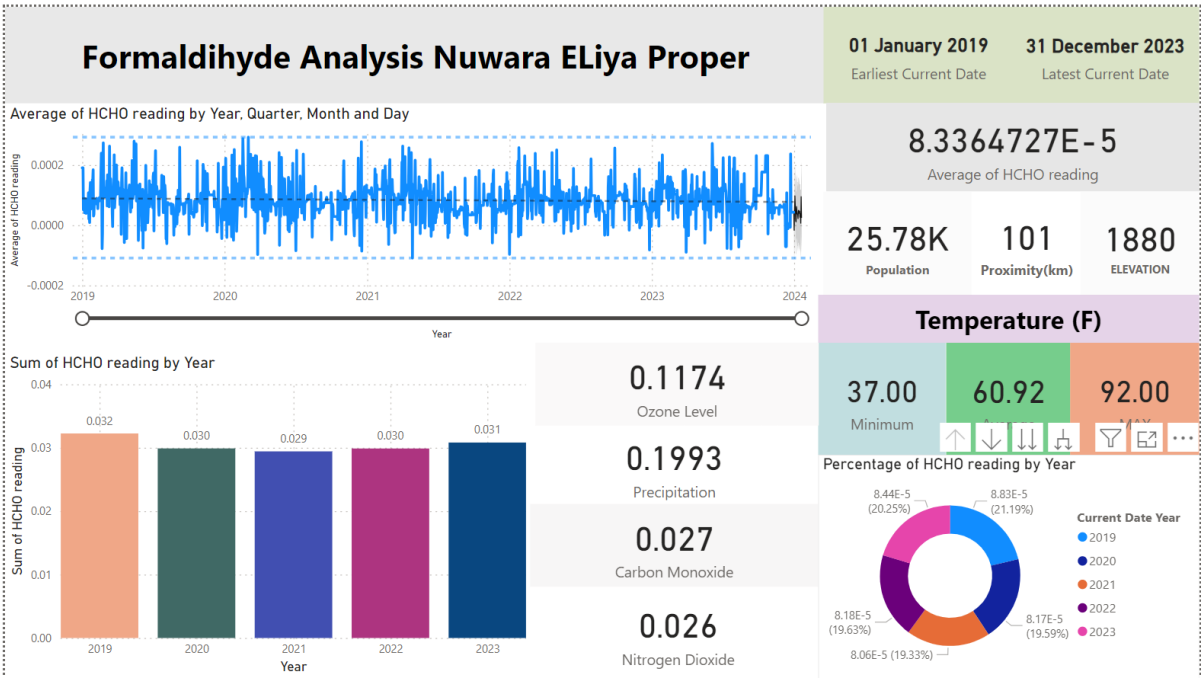


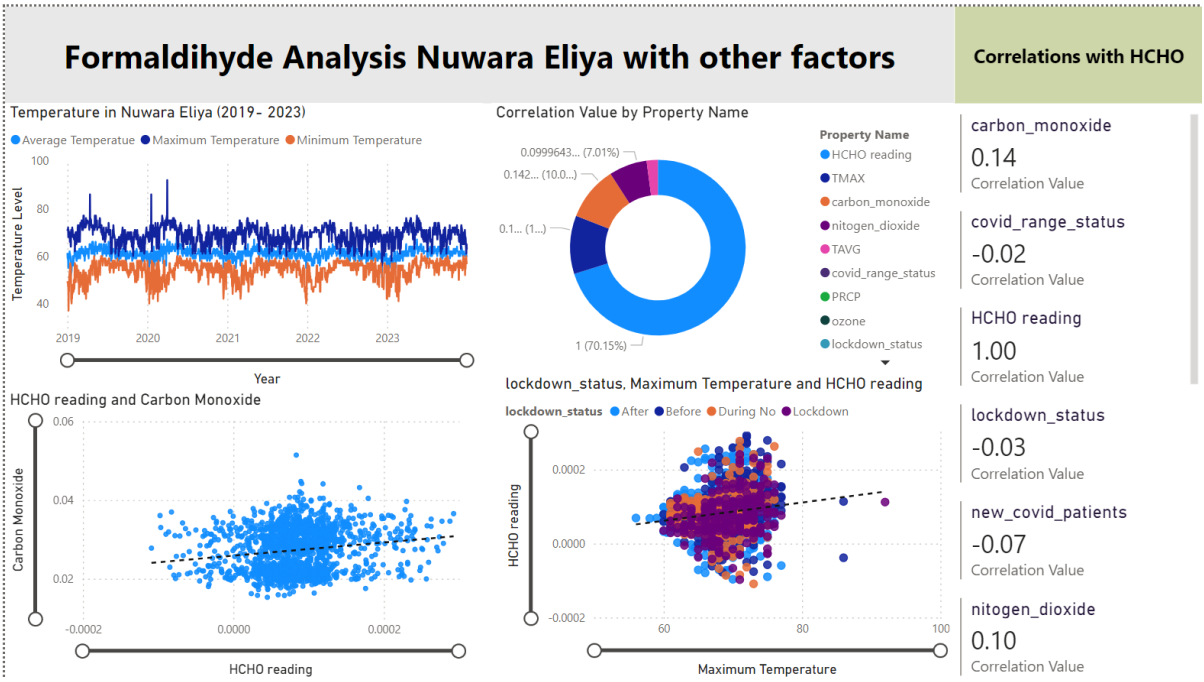




Statistical Measurements done for Nuwara Eliya Region

	HCHO reading
count	1826.000000
mean	0.000083
std	0.000054
min	-0.000109
25%	0.000054
50%	0.000079
75%	0.000110
max	0.000293

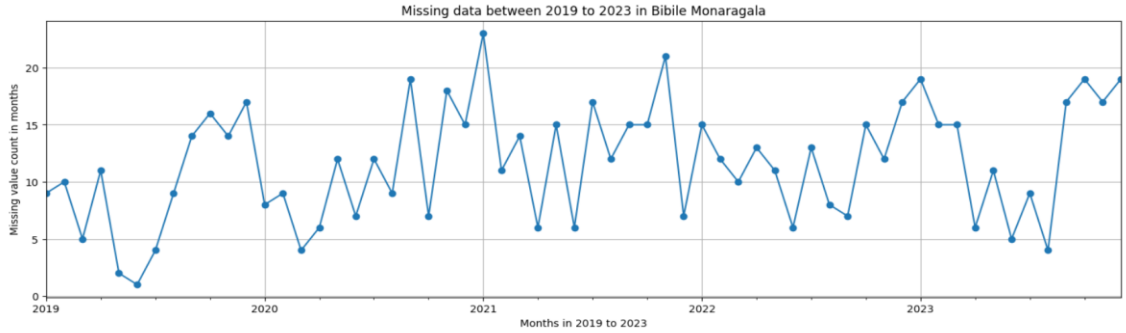

Formaldihyde Analysis Nuwara ELiya Proper
01 January 2019
Earliest Current Date
31 December 2023
Latest Current Date
Average of HCHO reading by Year, Quarter, Month and Day
Average of HCHO reading
0.0002
0.0000
-0.0002
2019
2020
2021
2022
2023
2024
Year
8.3364727E-5
Average of HCHO reading
25.78K
Population
101
Proximity(km)
1880
ELEVATION
Temperature (F)
37.00
Minimum
60.92
92.00
Sum of HCHO reading by Year
Sum of HCHO reading
0.04
0.03
0.02
0.01
0.00
0.032
0.030
0.029
0.030
0.031
2019
2020
2021
2022
2023
Year
0.1174
Ozone Level
0.1993
Precipitation
0.027
Carbon Monoxide
0.026
Nitrogen Dioxide
Percentage of HCHO reading by Year
8.44E-5 (20.25%)
8.83E-5 (21.19%)
8.18E-5 (19.63%)
8.17E-5 (19.59%)
8.06E-5 (19.33%)
Current Date Year
2019
2020
2021
2022
2023

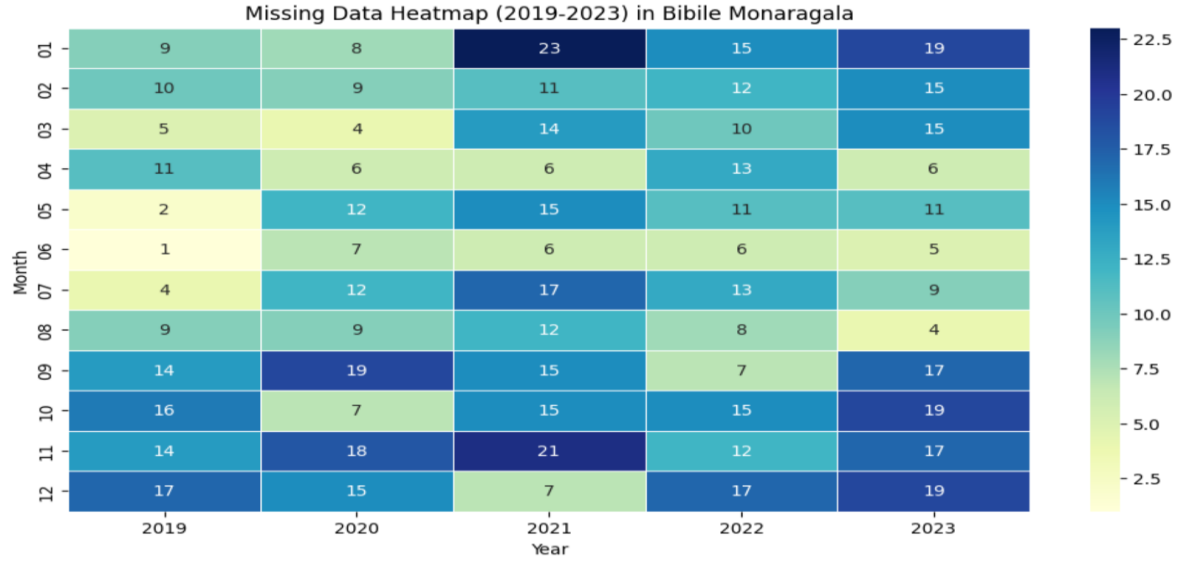

Formaldihyde Analysis Nuwara Eliya with other factors
Correlations with HCHO
Temperature in Nuwara Eliya (2019- 2023)
Average Temperatue
Maximum Temperature
Minimum Temperature
Temperature Level
100
80
60
40
2019
2020
2021
2022
2023
Year
Correlation Value by Property Name
0.0999643... (7.01%)
0.142... (10.0...)
0.1... (1...)
1 (70.15%)
Property Name
HCHO reading
TMAX
carbon_monoxide
nitogen_dioxide
TAVG
covid_range_status
PRCP
ozone
lockdown_status
HCHO reading and Carbon Monoxide
Carbon Monoxide
0.06
0.04
0.02
-0.0002
0.0000
0.0002
HCHO reading
lockdown_status, Maximum Temperature and HCHO reading
lockdown_status
After
Before
During No
Lockdown
HCHO reading
0.0002
0.0000
-0.0002
60
80
100
Maximum Temperature
carbon_monoxide
0.14
Correlation Value
covid_range_status
-0.02
Correlation Value
HCHO reading
1.00
Correlation Value
lockdown_status
-0.03
Correlation Value
new_covid_patients
-0.07
Correlation Value
nitogen_dioxide
0.10
Correlation Value

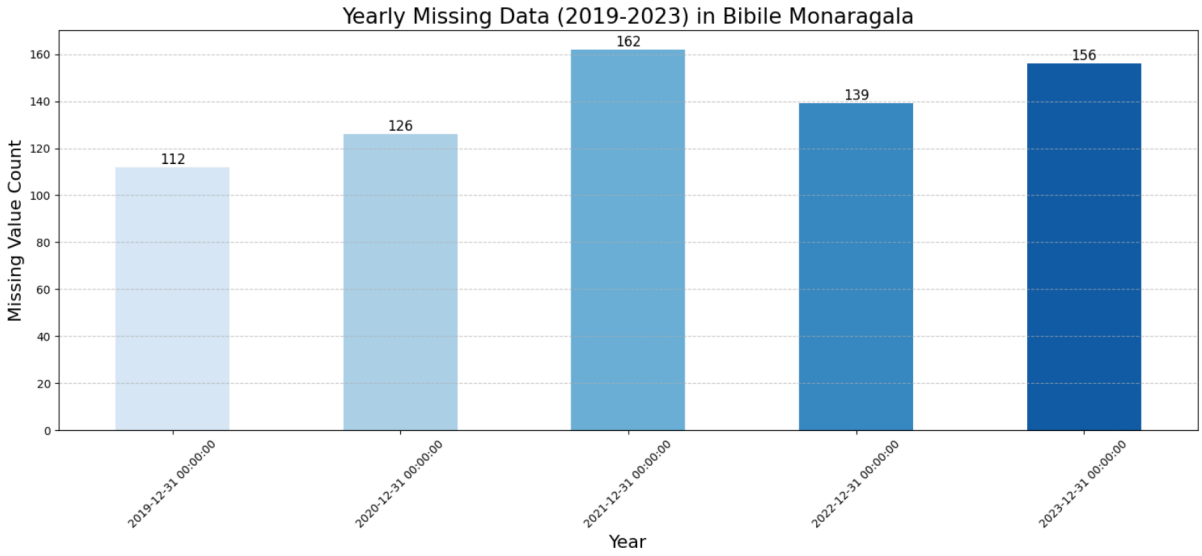
Bibile Monaragala Region Formaldehyde Distribution Analysis

Data Preprocessing Bibile Monaragala

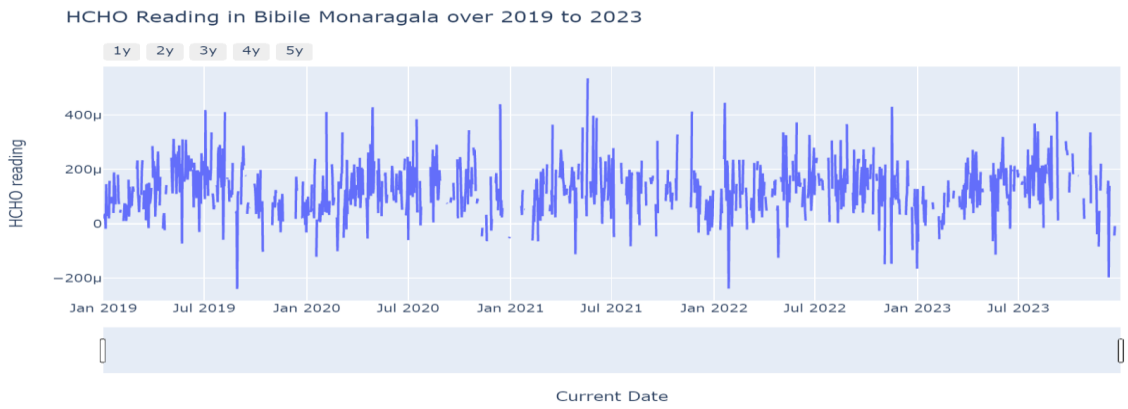
The initial Bibile Monaragala dataset consisted with 695 null values out of 1826 records. Compared to other cities it consisted with a smaller number of missing values. The below heatmap, line plot and bar chart show how null values distributed, yearly and monthly.



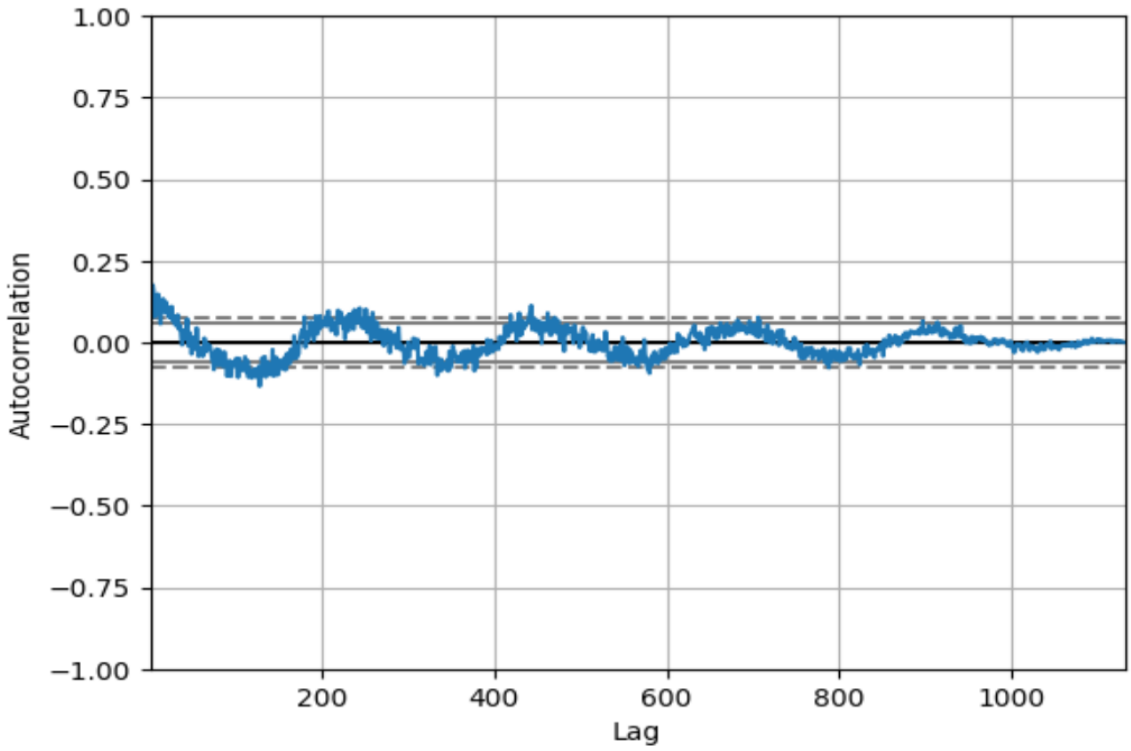




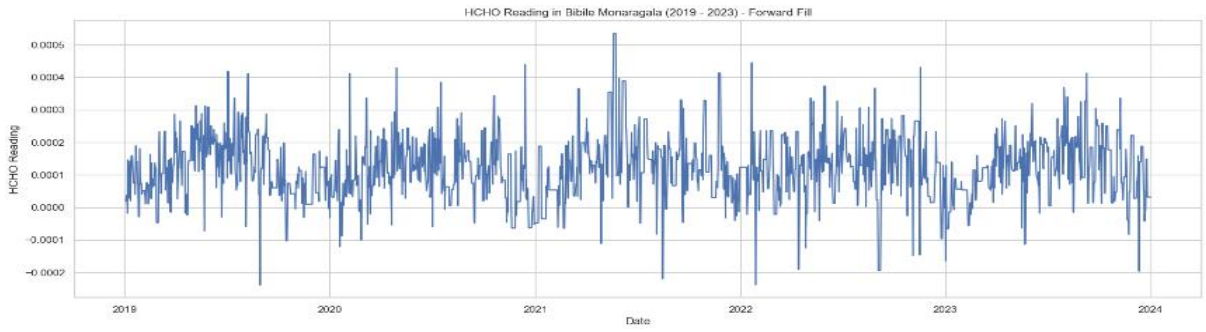
The above heatmap shows there are some months that consist of maximum 23 null values. The below plot shows the Monaragala HCHO distribution before handling missing values.

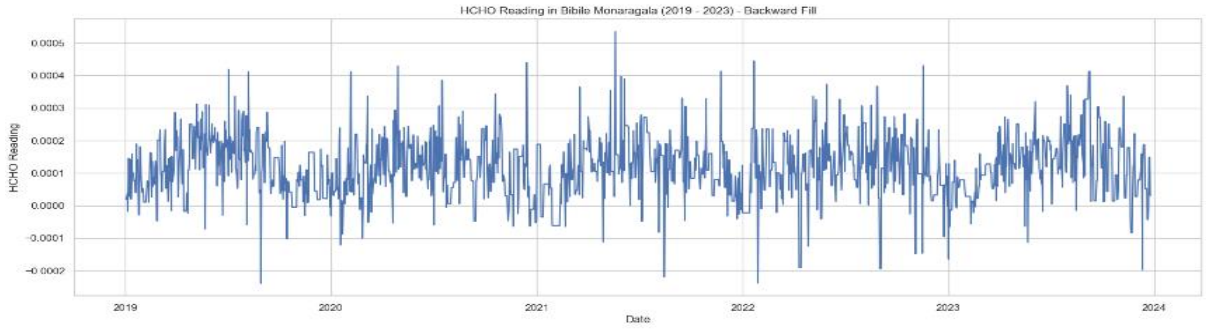


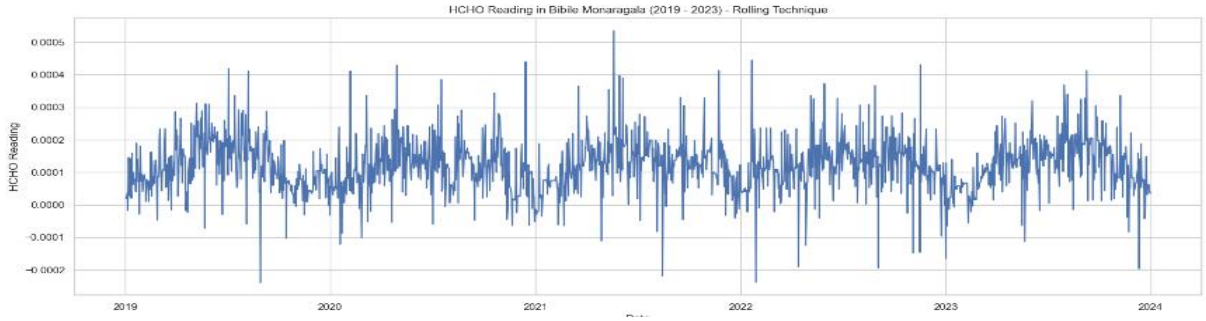
The below Auto-Correlation plot shows there is a clear fluctuation and a seasonality in Monaragala Distribution. However, it shows that fluctuation has not spread in a wide range. However, the window size 15 is used to handle all the null values.



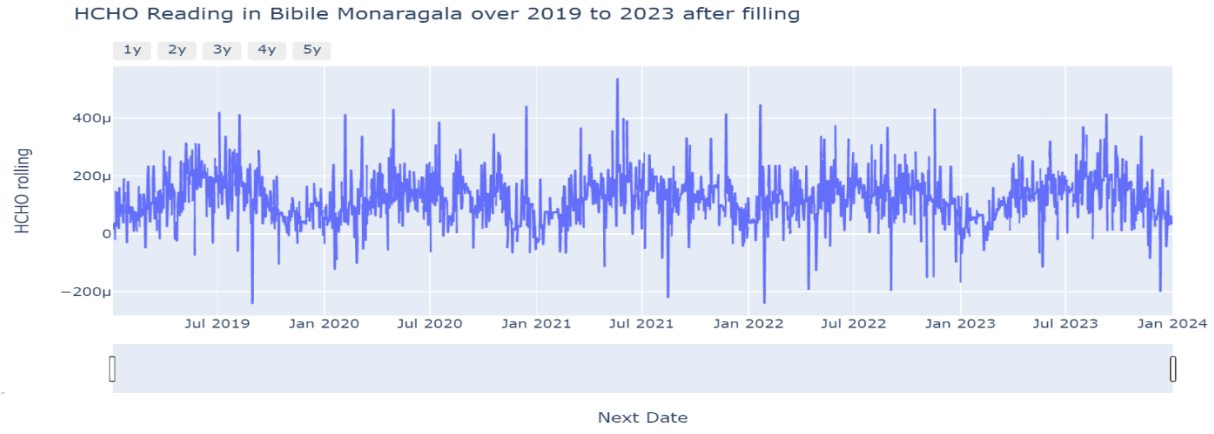
The below line charts show how the HCHO distribution of Monaragala looks when handling missing values with rolling, forward, and backward filling techniques.



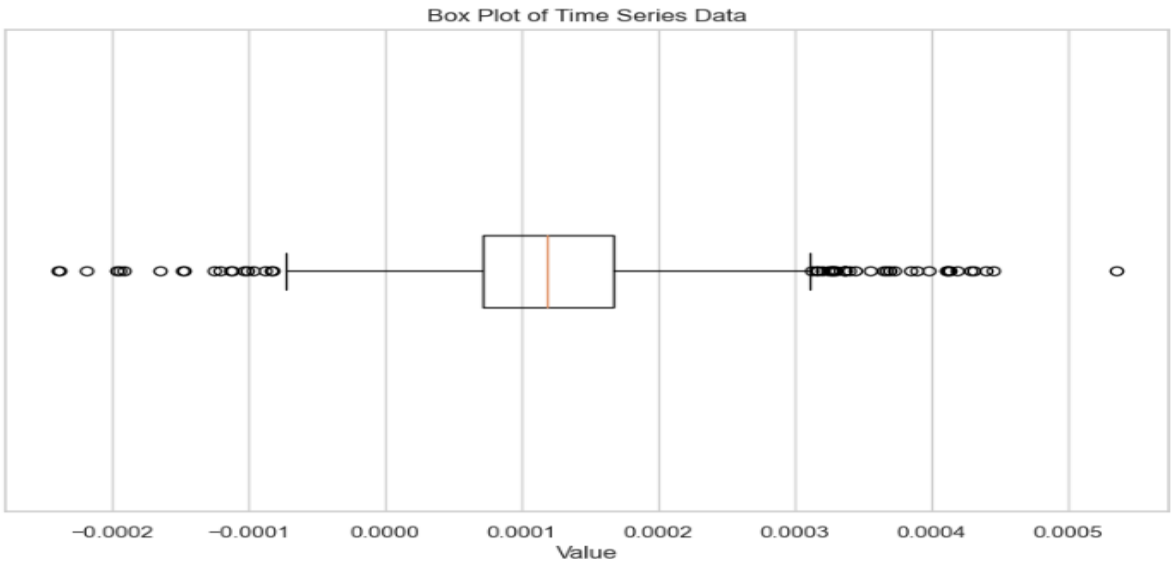




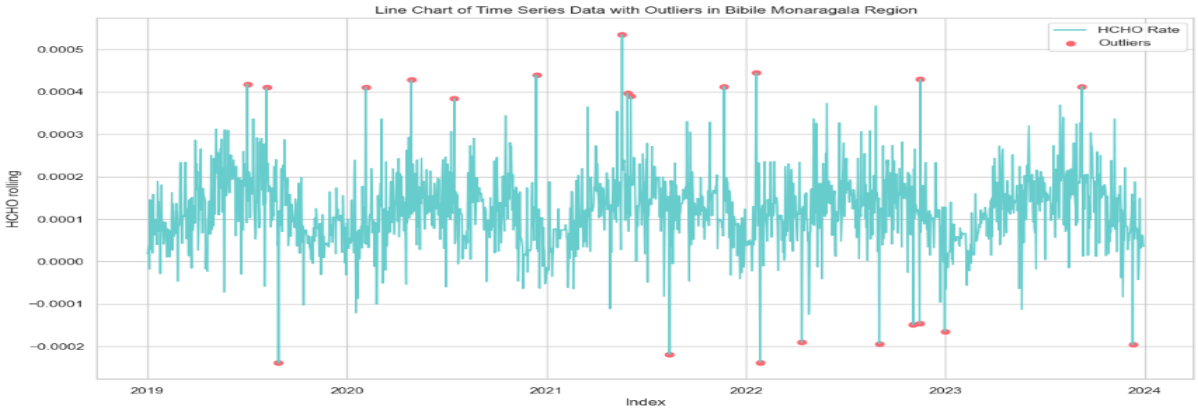
The below line chart shows how HCHO values distributed after handling null values in Monaragala Region. The rolling window 15 filled distribution is considered as the final distribution.

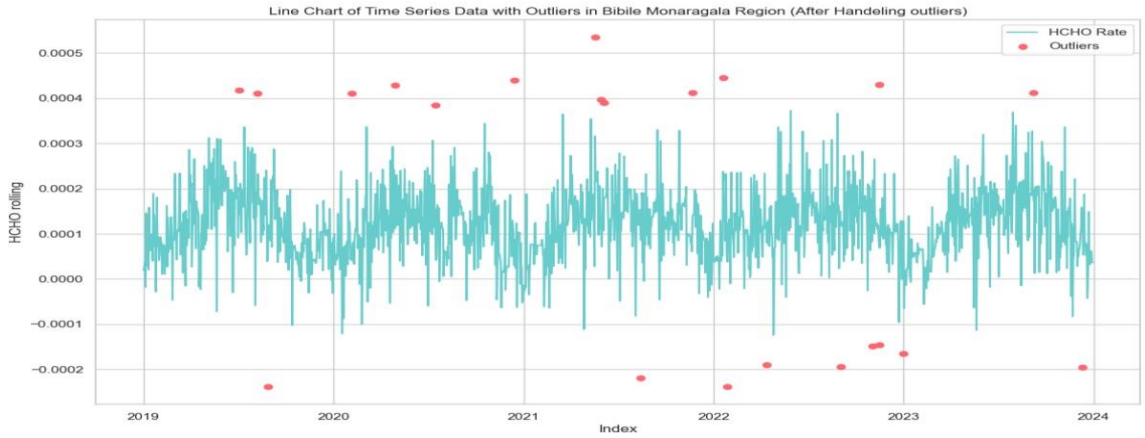


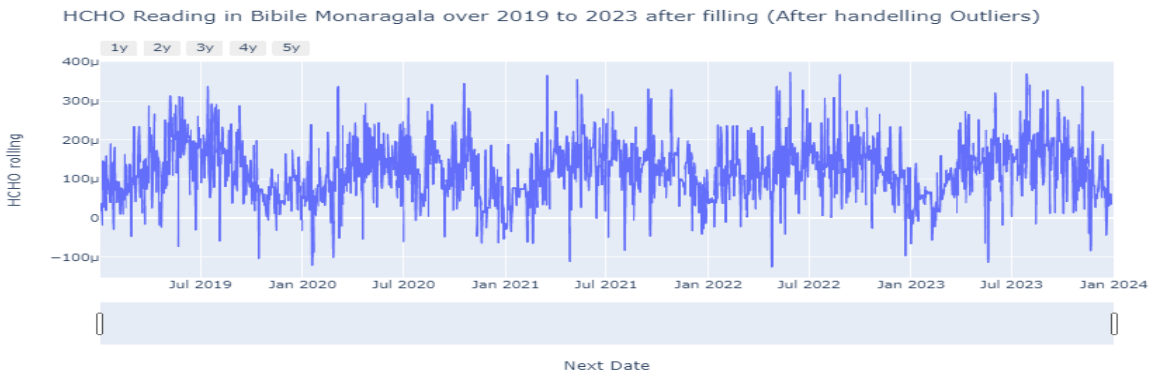
The given boxplot shows the outliers detected by considering the quartiles, but the outliers are removed using a threshold value of 2.25 with inter quartile range.



The below plots show the identified outliers in Monaragala and how the distribution look like after handling the outliers.

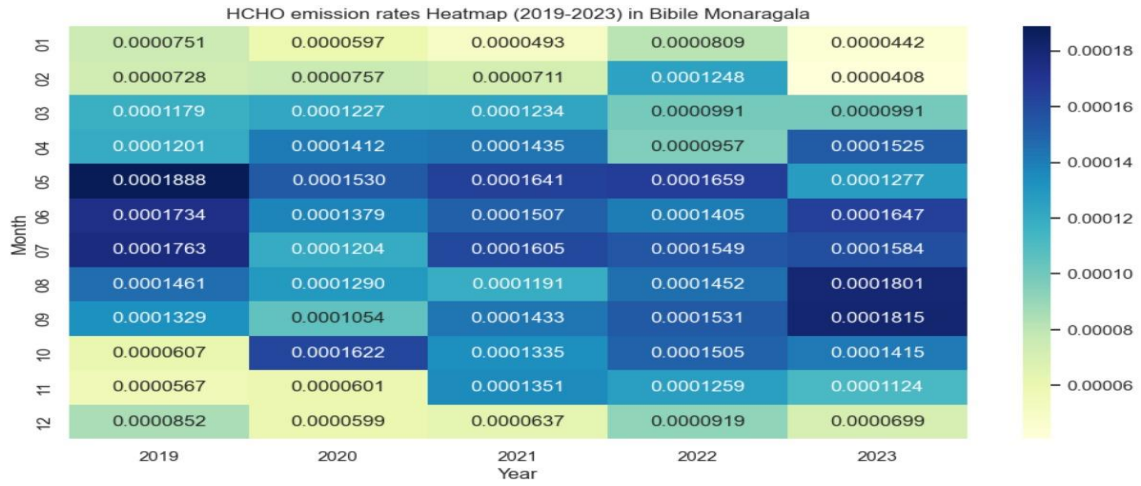




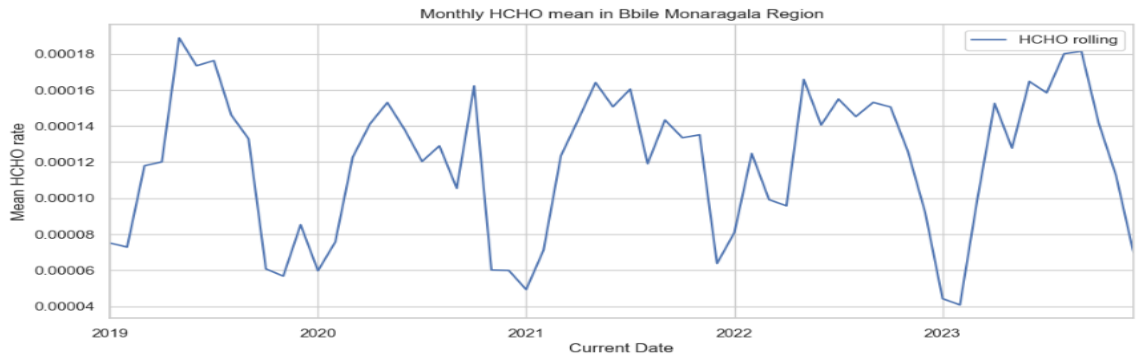


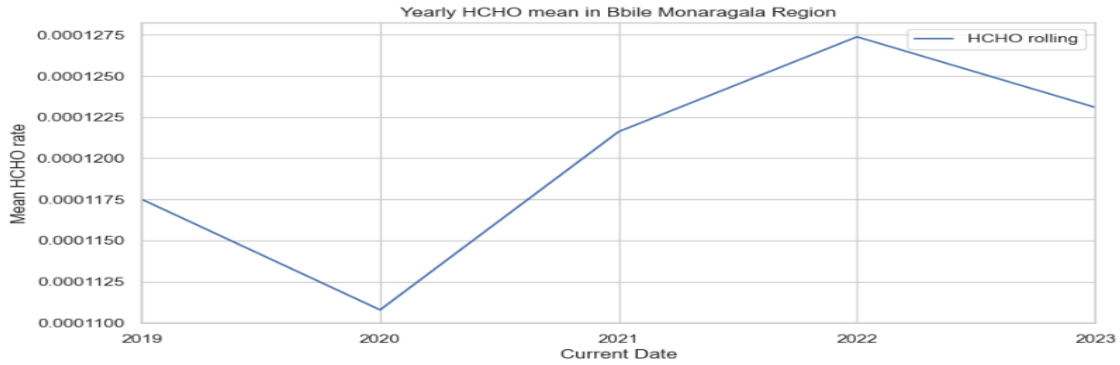
HCHO distribution Insights Bibile Monaragala

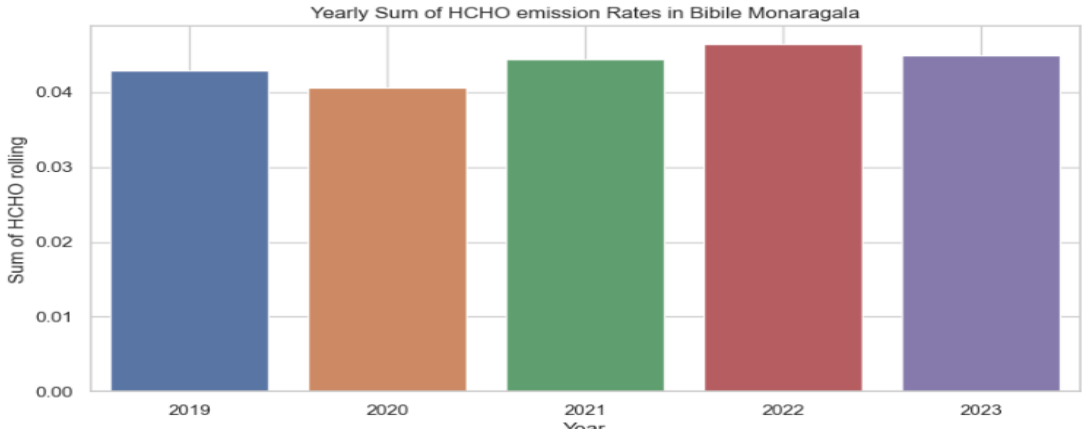
The below visualizations show that not like in other regions there is an increment of HCHO rates in the year 2021. There is a fluctuation of mean rates in all the years by reporting the minimum rate in year 2020 and maximum rate in year 2022. The below visualizations depict how Monaragala HCHO distribution varies its values monthly, weekly, and yearly.

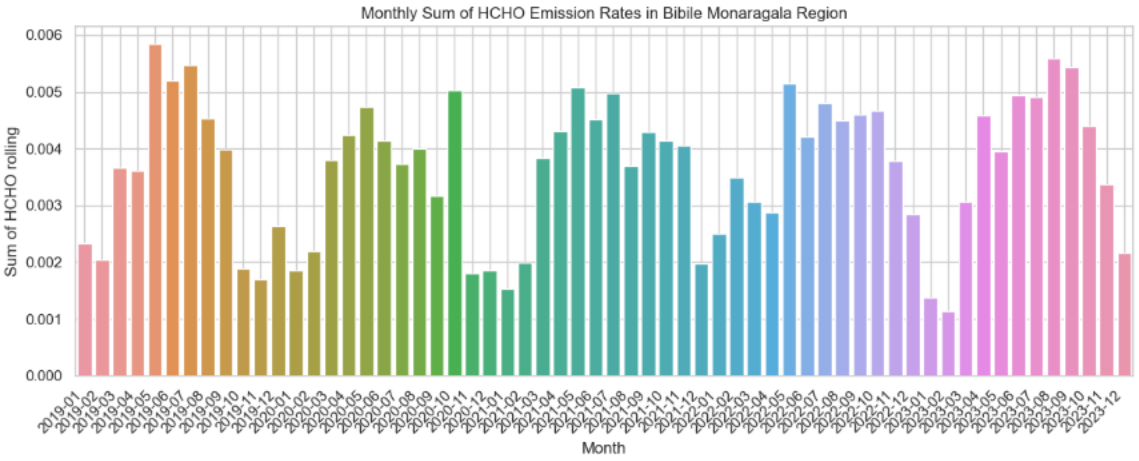


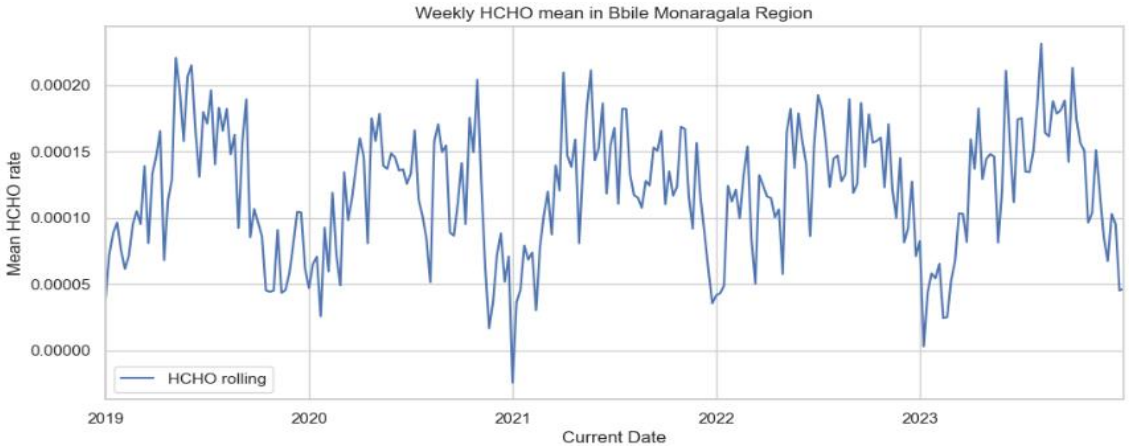
It is not like in other cities, it has reported the maximum HCHO emission in the mid months of the year.

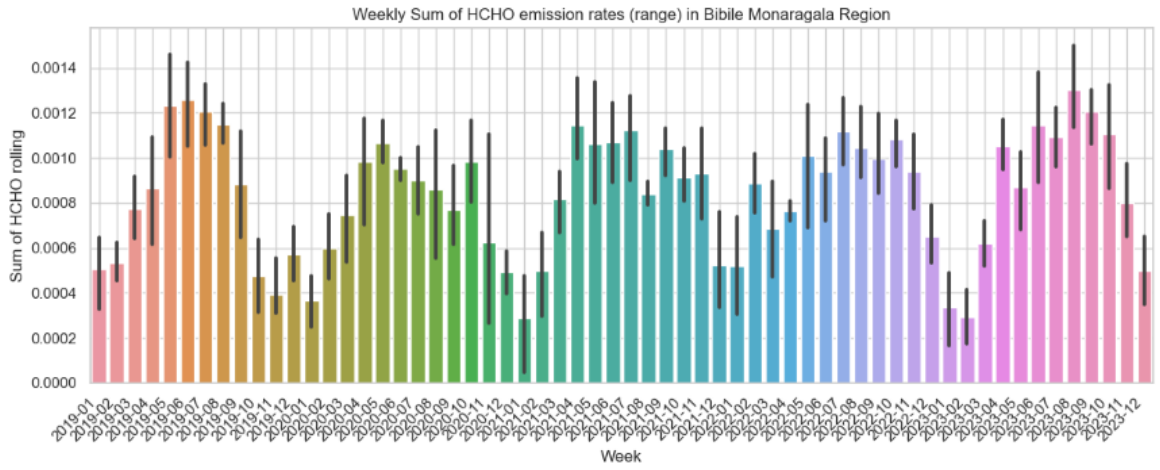



Yearly HCHO mean in Bbile Monaragala Region
HCHO rolling
Mean HCHO rate
0.0001275
0.0001250
0.0001225
0.0001200
0.0001175
0.0001150
0.0001125
0.0001100
2019
2020
2021
2022
2023
Current Date


Yearly Sum of HCHO emission Rates in Bibile Monaragala
Sum of HCHO rolling
0.04
0.03
0.02
0.01
0.00
2019
2020
2021
2022
2023


Monthly Sum of HCHO Emission Rates in Bibile Monaragala Region
Sum of HCHO rolling
0.006
0.005
0.004
0.003
0.002
0.001
0.000
Month

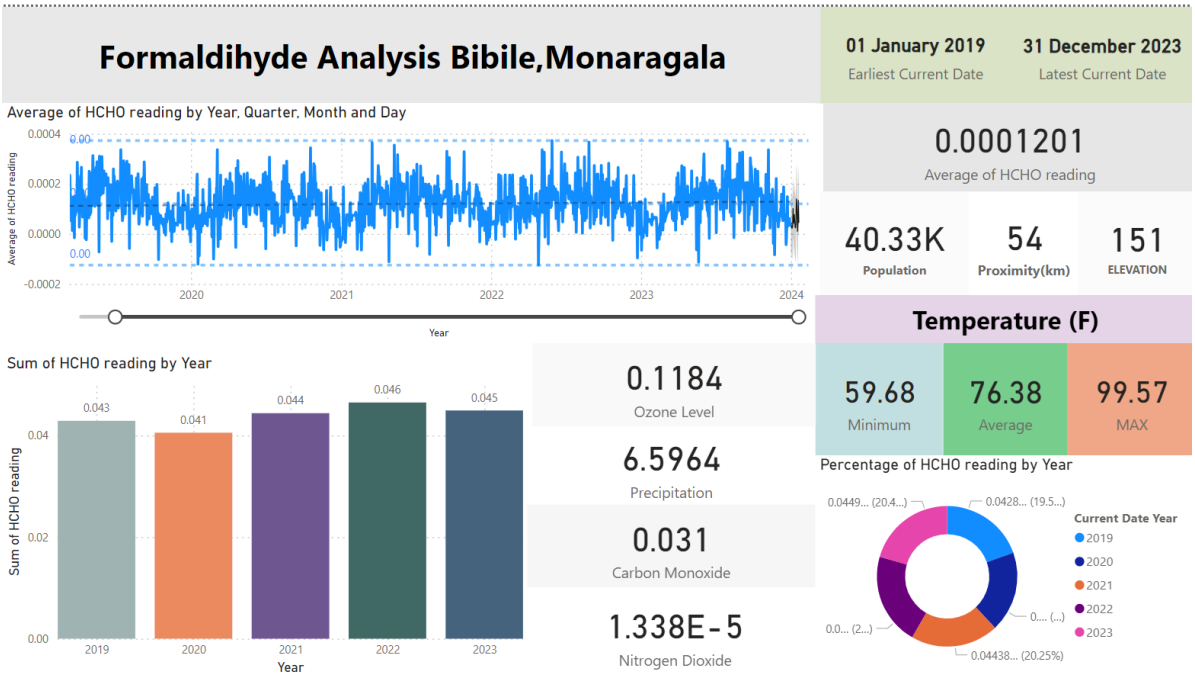

Weekly HCHO mean in Bbile Monaragala Region
Mean HCHO rate
0.00020
0.00015
0.00010
0.00005
0.00000
HCHO rolling
2019
2020
2021
2022
2023
Current Date

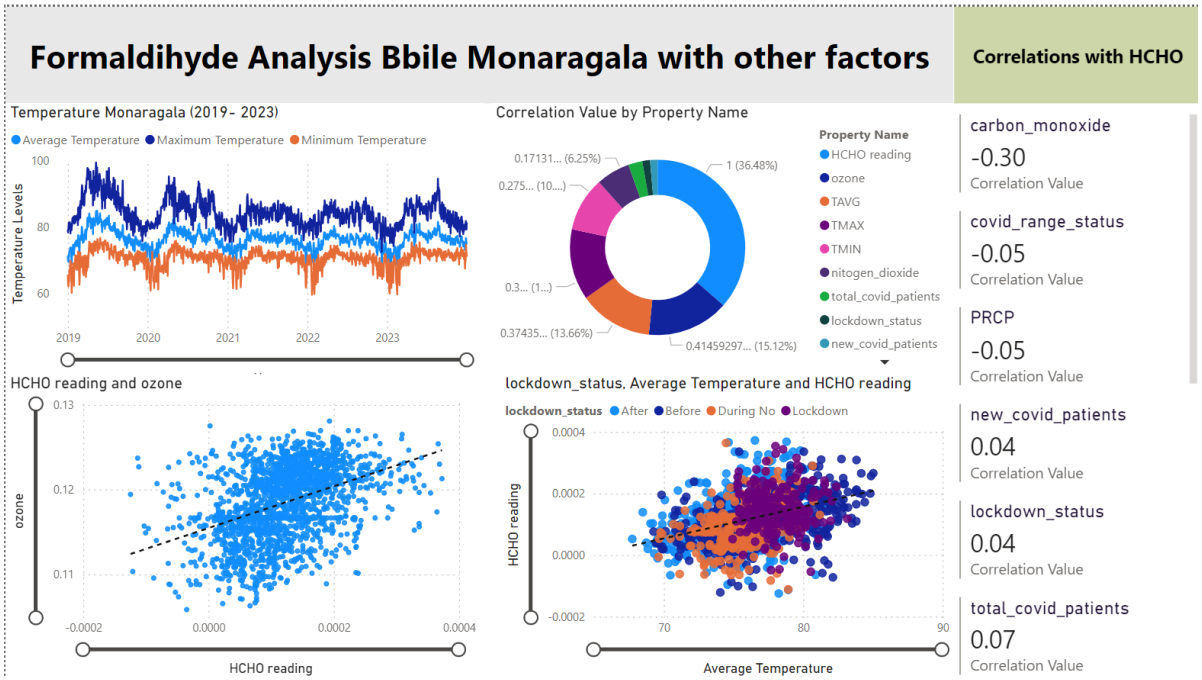


Statistical Measurements done for Monaragala Region

The below table shows the statistical calculations done for Monaragala Region.

	HCHO reading
count	1826.000000
mean	0.000120
std	0.000073
min	-0.000125
25%	0.000072
50%	0.000119
75%	0.000166
max	0.000373

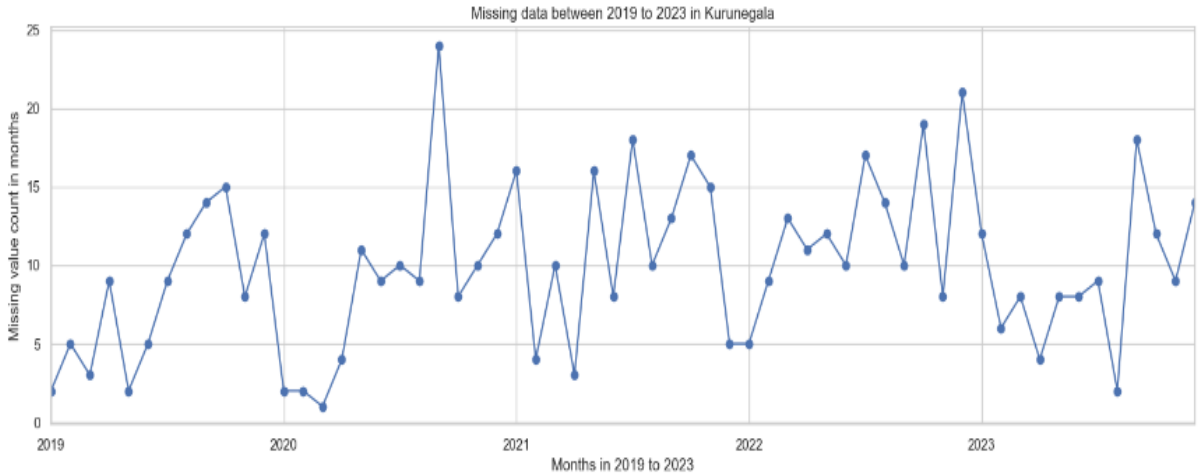


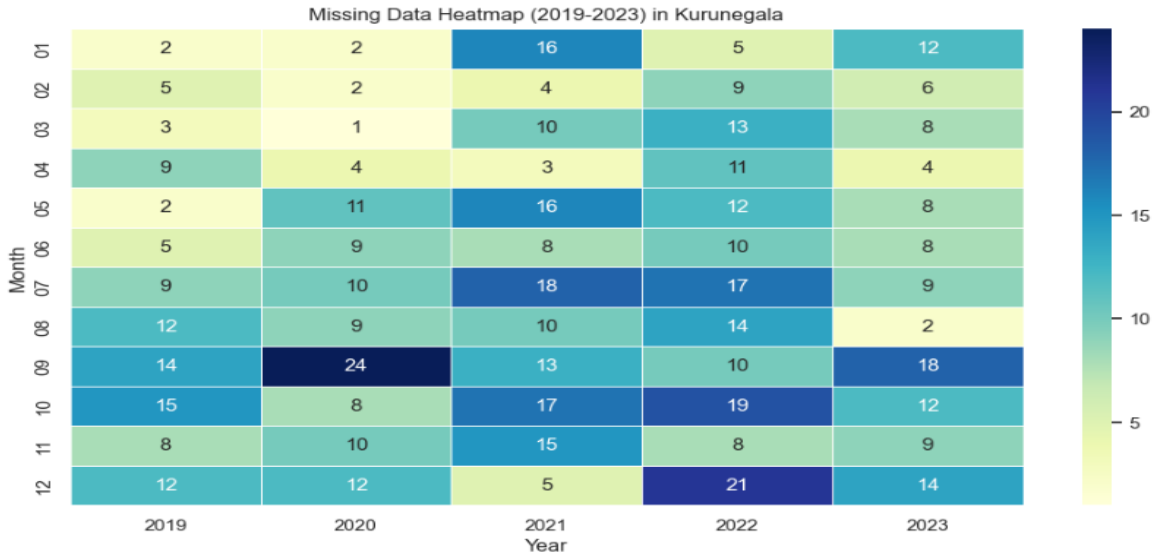


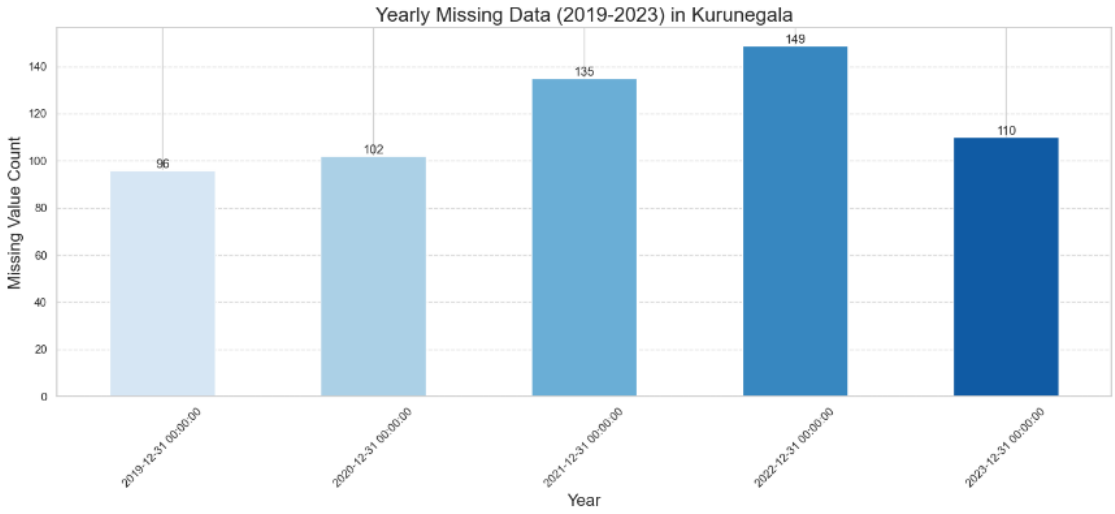
Kurunegala Proper Region Formaldehyde Distribution Analysis

Data Preprocessing Kurunegala

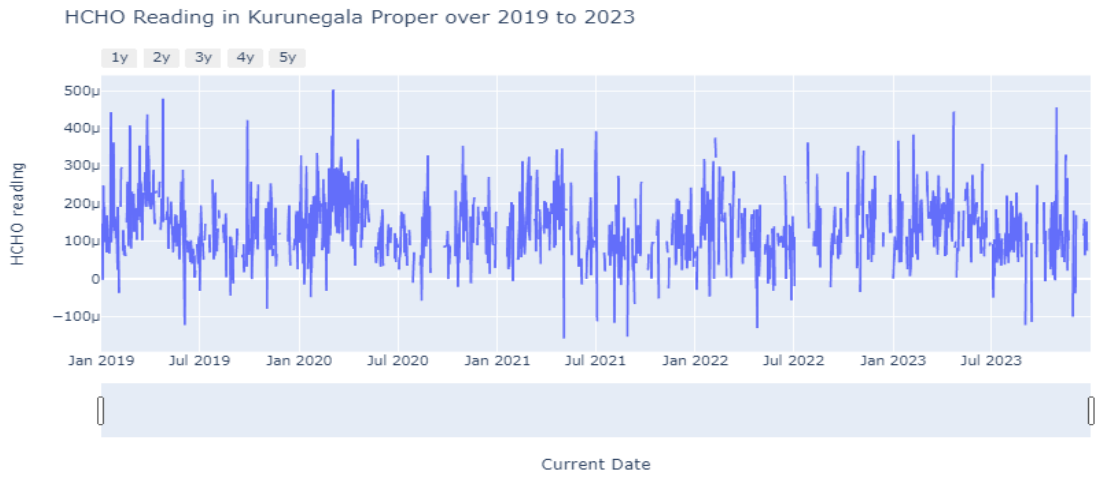
The Kurunegala HCHO distribution consisted of 592 missing values out of 1826 values. It consisted of a smaller number of null values in the Kurunegala dataset. Therefore, it is easy to find a seasonal pattern in the dataset. The below plots show the null value distribution in Kurunegala region data.



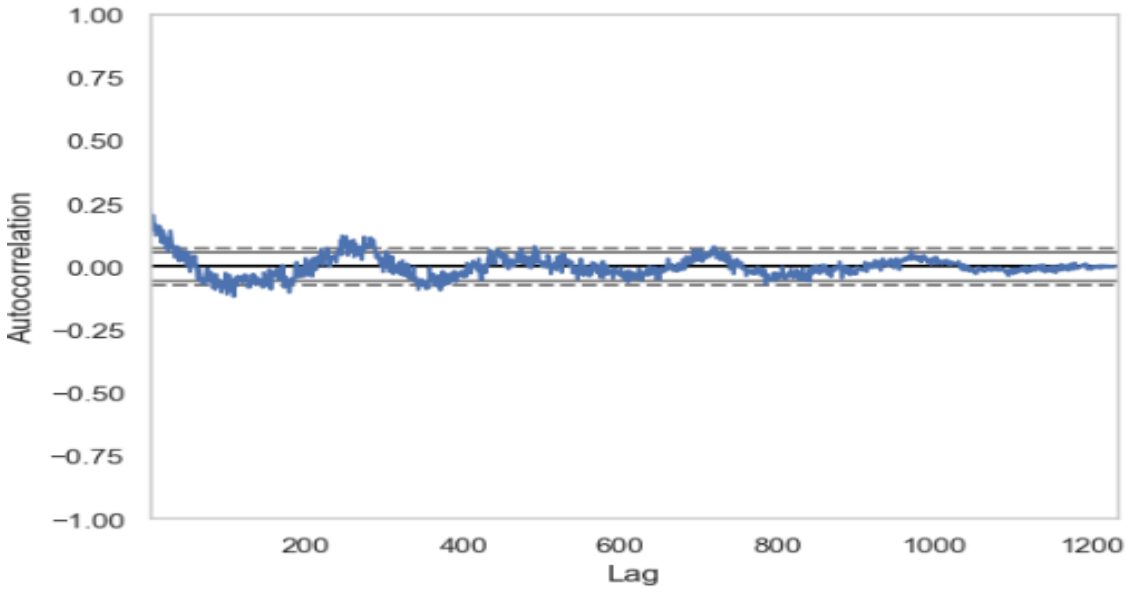




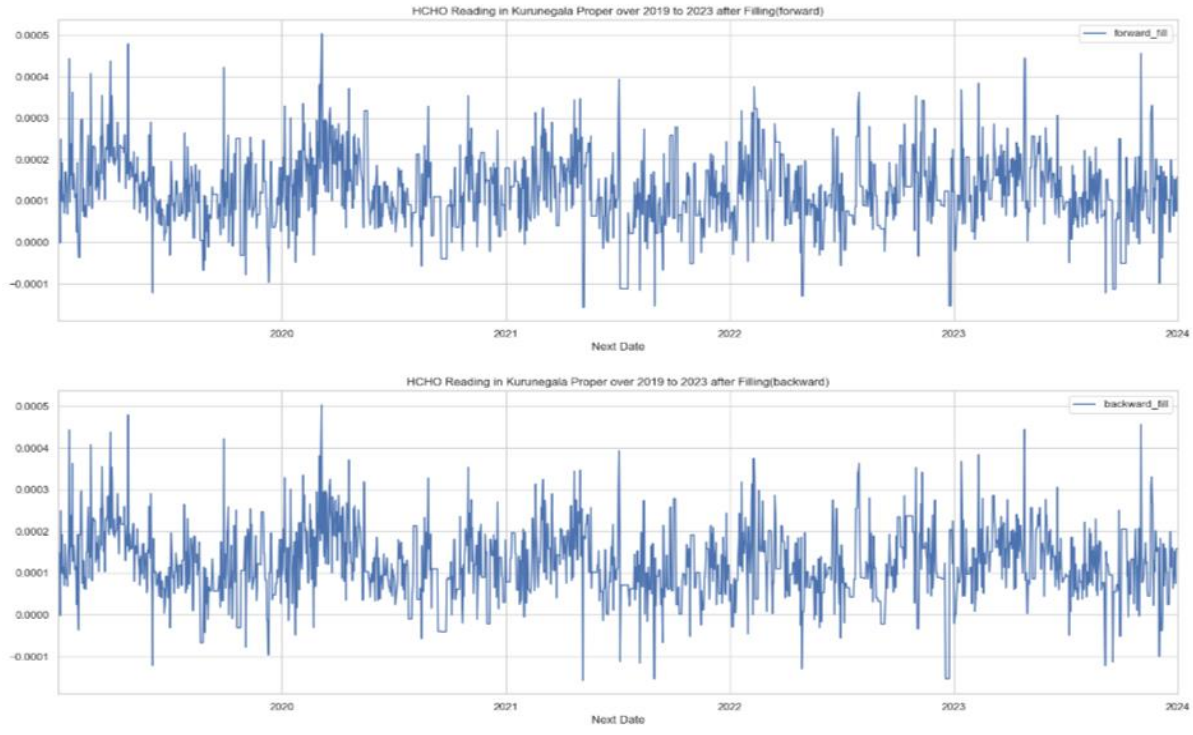
The above heatmap shows there are some months that consist of 24 null values as well. The below plot shows the Kurunegala HCHO distribution before handling null values.

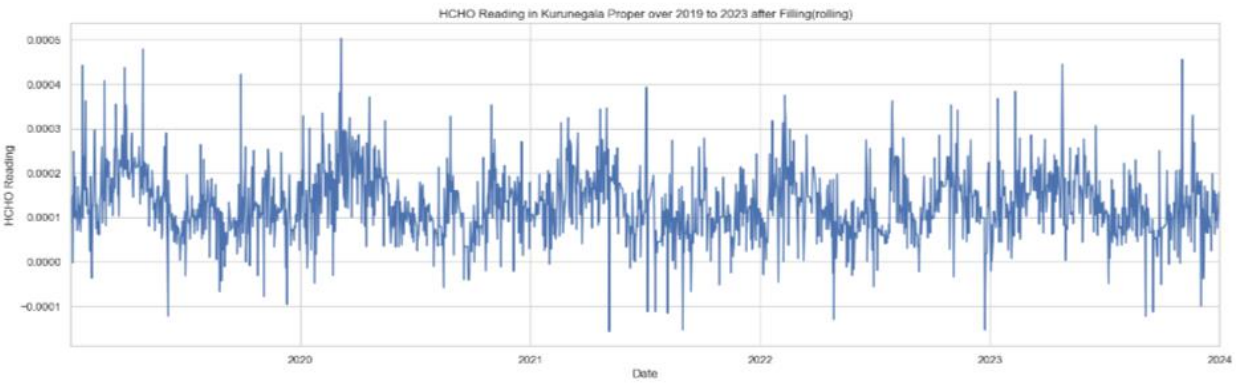


The below Auto-Correlation plot shows there is a clear fluctuation and a seasonality in Kurunegala data. However, it shows that fluctuation has not spread in a wide range as in Colombo Region. The window size 13 was used to handle the null values by keeping the original distribution pattern.

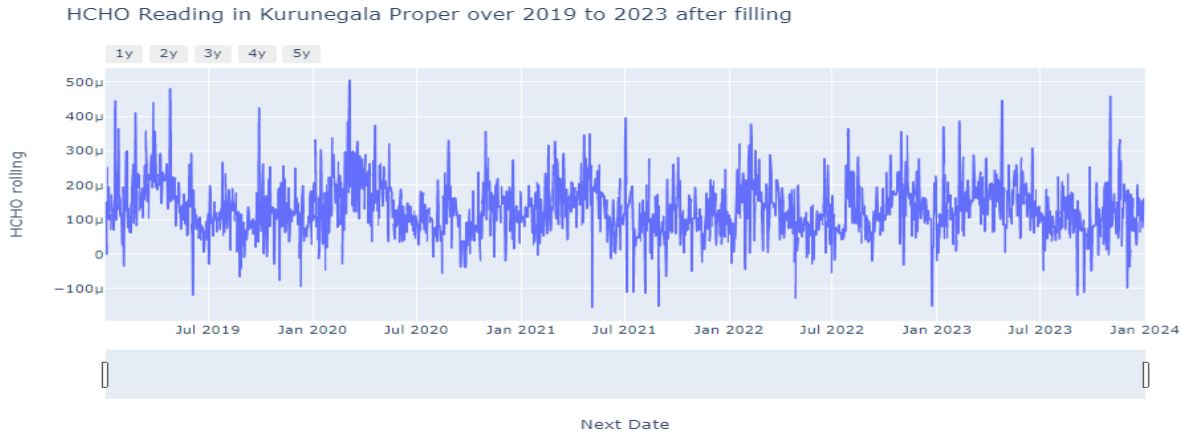


The below line charts show how the HCHO distribution looks when handling missing values with rolling, forward, and backward filling techniques.

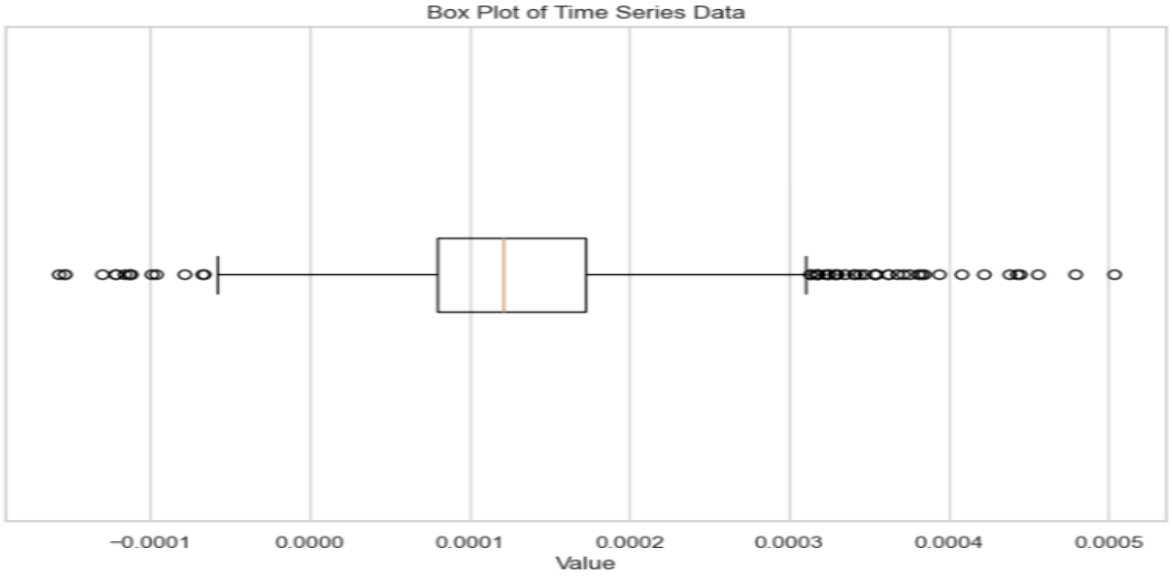




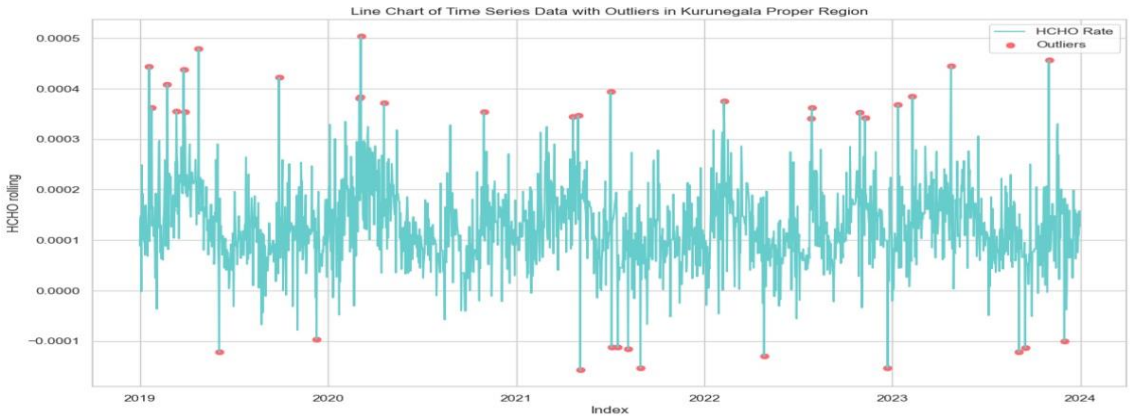
Since the rolling technique with window 13 method has maintained the original pattern with fluctuations, it has been considered as the final null value handled distribution. The below plot shows the final null value handled distribution.

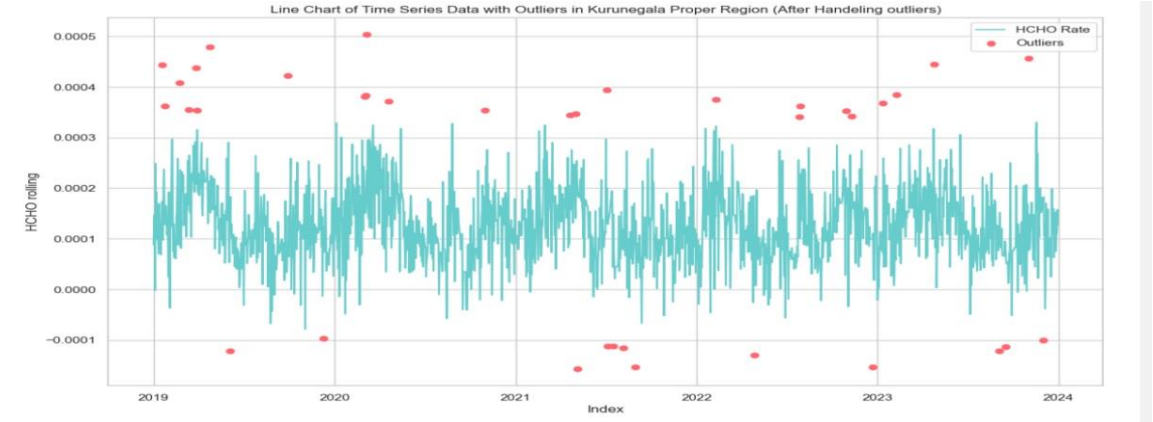


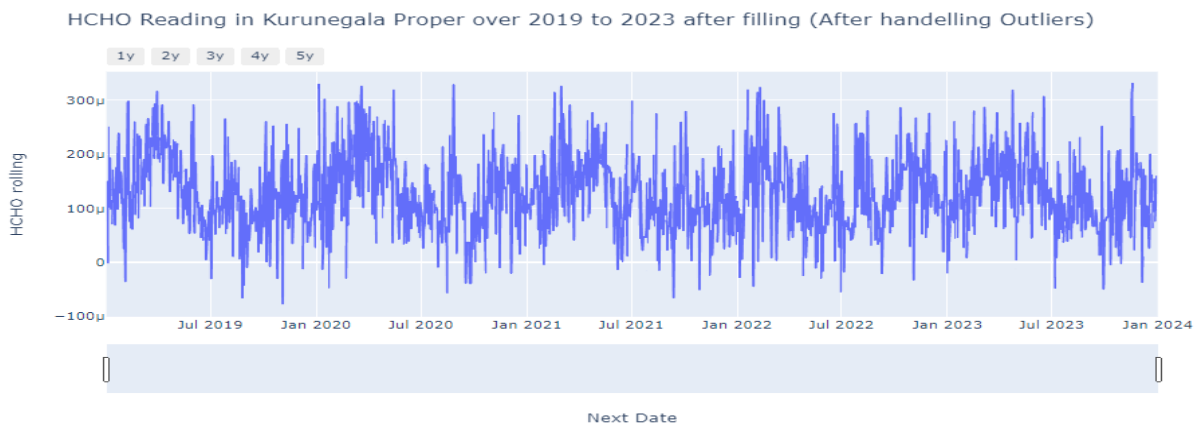
The given boxplot shows the outliers detected by considering the quartiles, but the outliers are removed using a threshold value of 1.75 with inter quartile range.



The below plots show the identified outliers in Kurunegala and how the distribution look like after handling the outliers.

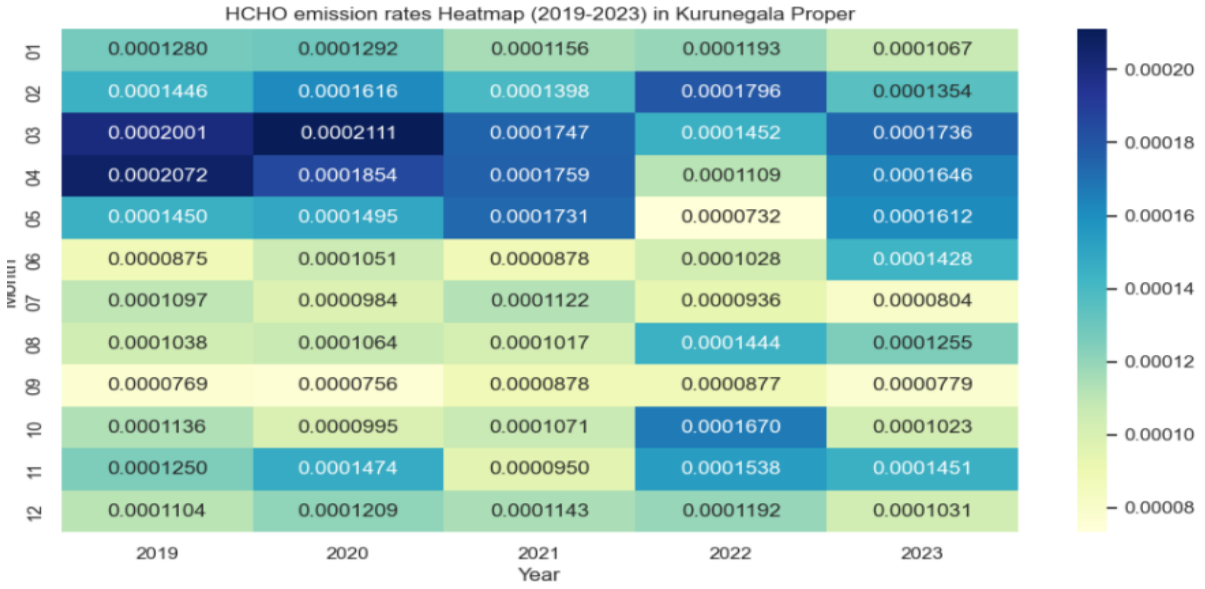




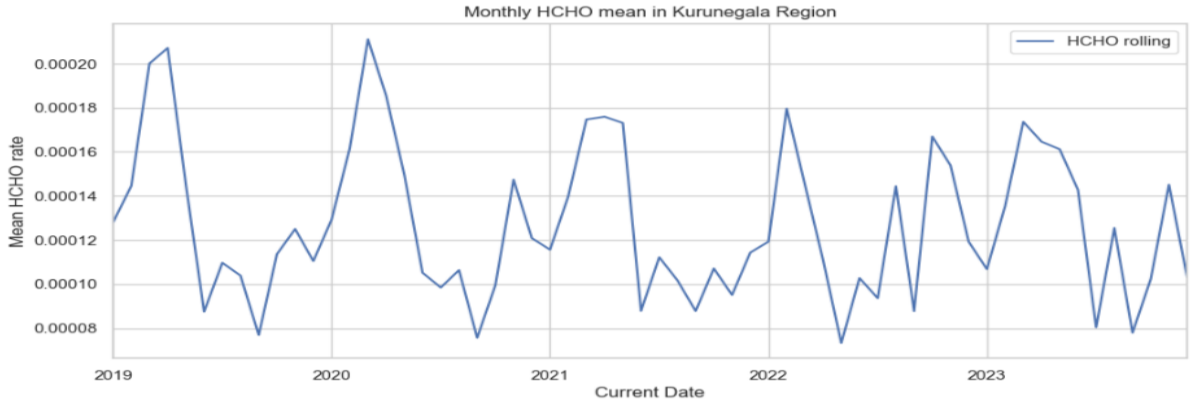


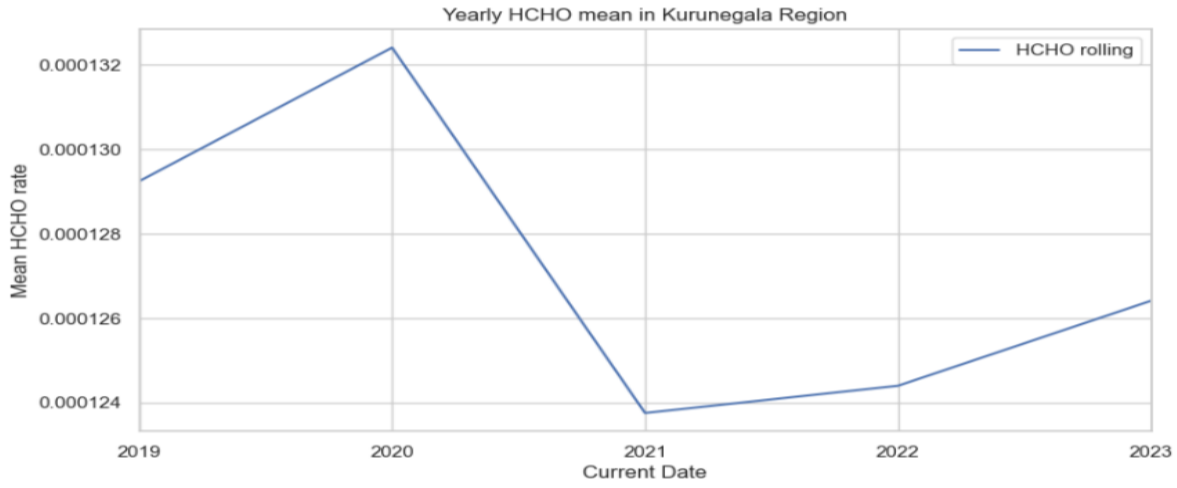
HCHO distribution Insights Kurunegala Proper

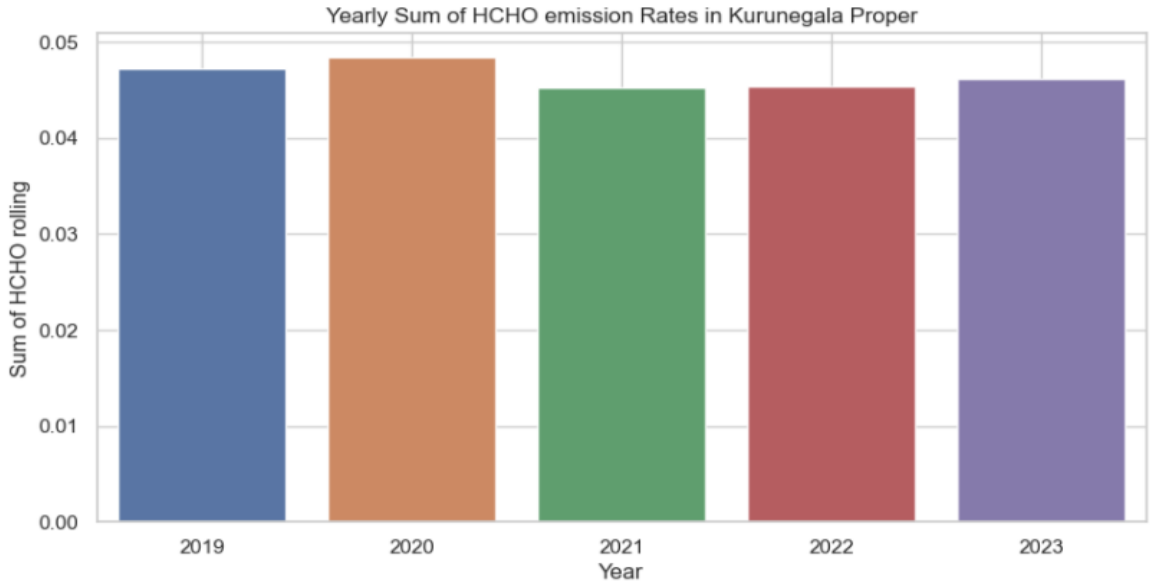
The below visualizations show that in the Kurunegala region, there were high HCHO emissions in the first four months of the year and low HCHO emission rates in other months of the year. As in many regions, it has reported the lowest HCHO emission rate in 2021. In addition, it recorded its maximum HCHO emission in 2020. The below visualizations describe how HCHO distributions vary their values monthly, weekly, and yearly.

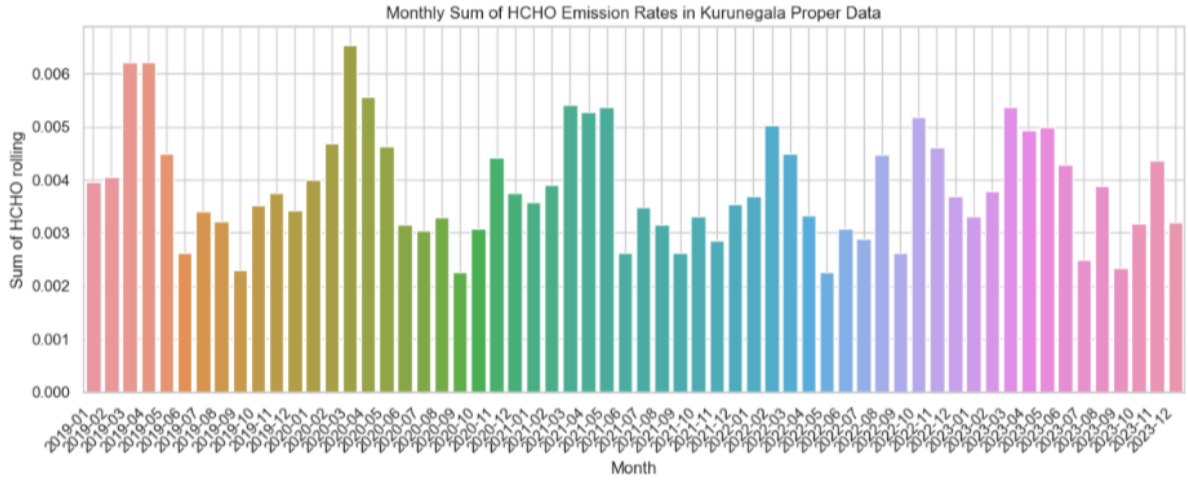


It is not like in other cities, Kurunegala has reported the maximum HCHO rate in the first four months of the year.

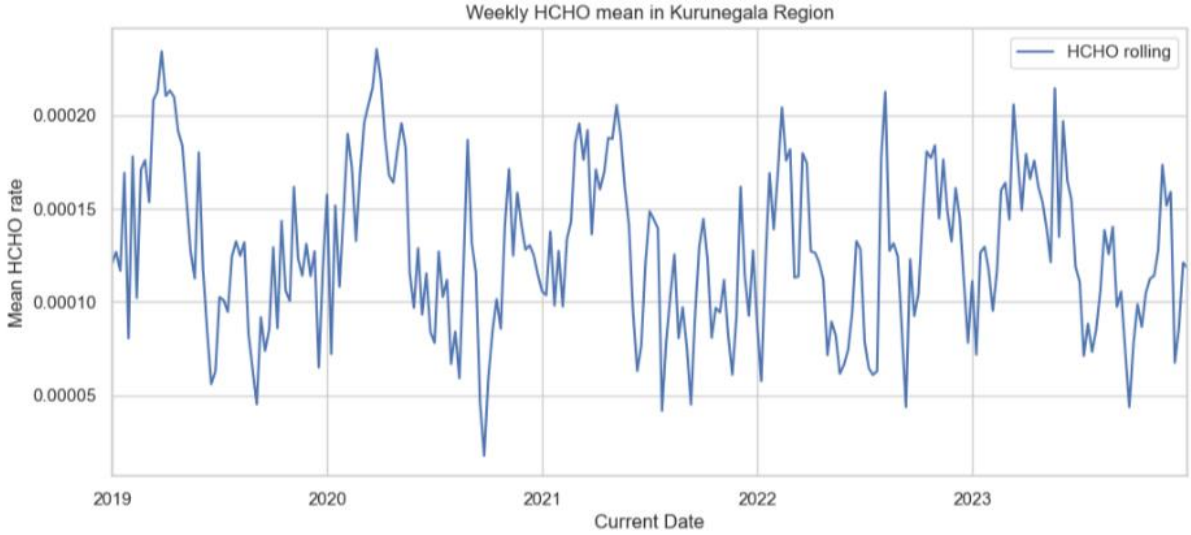


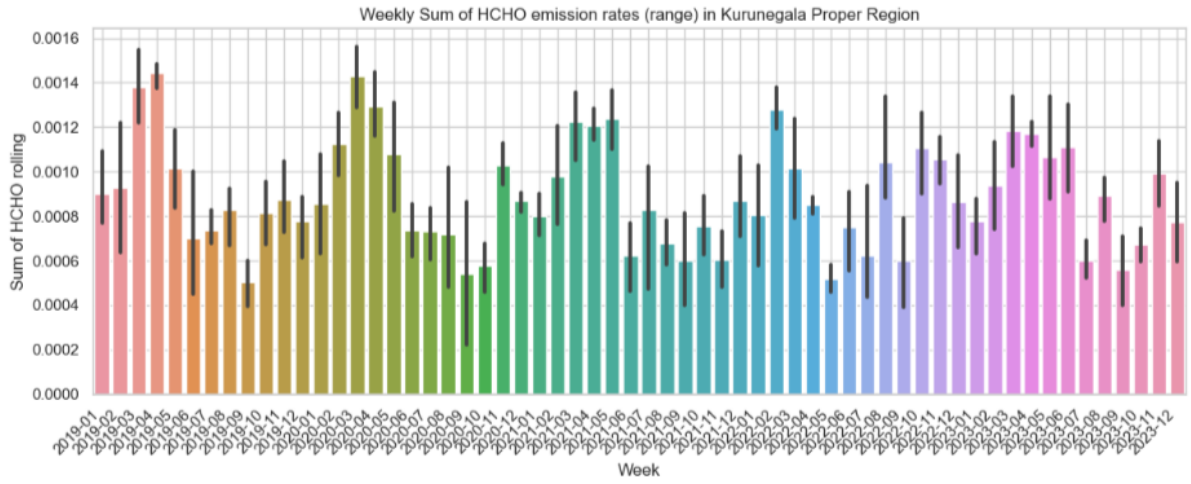






The below plots show the weekly distribution of HCHO rates in Kurunegala region.

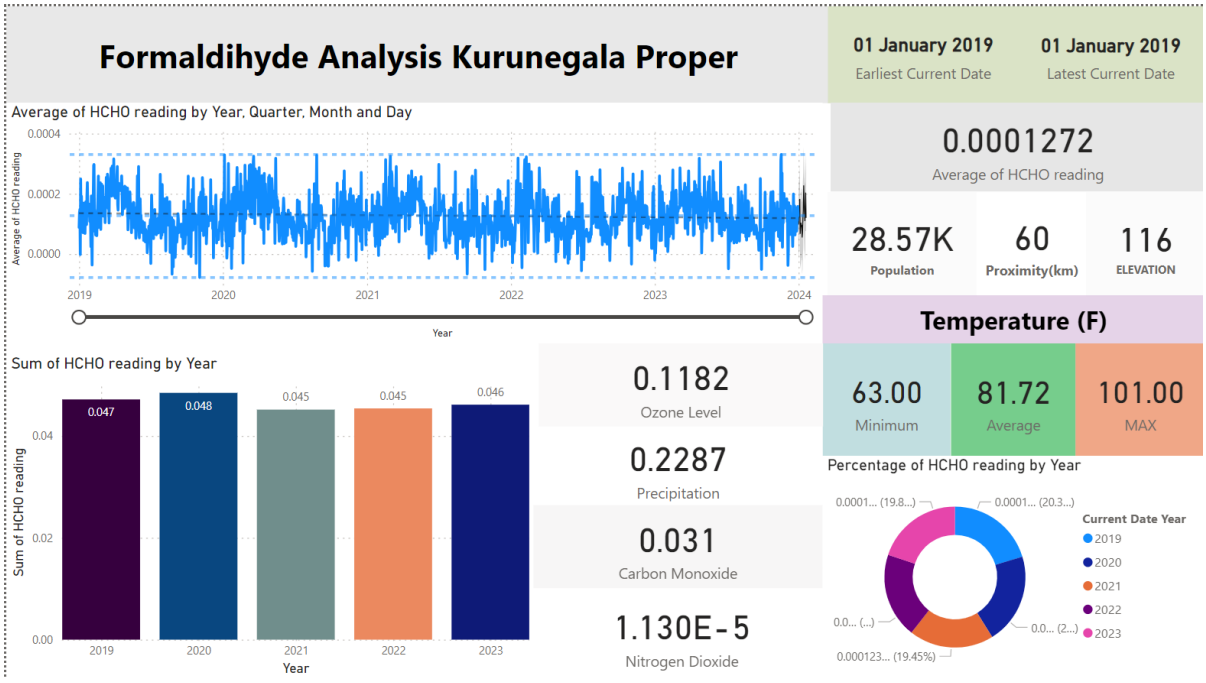


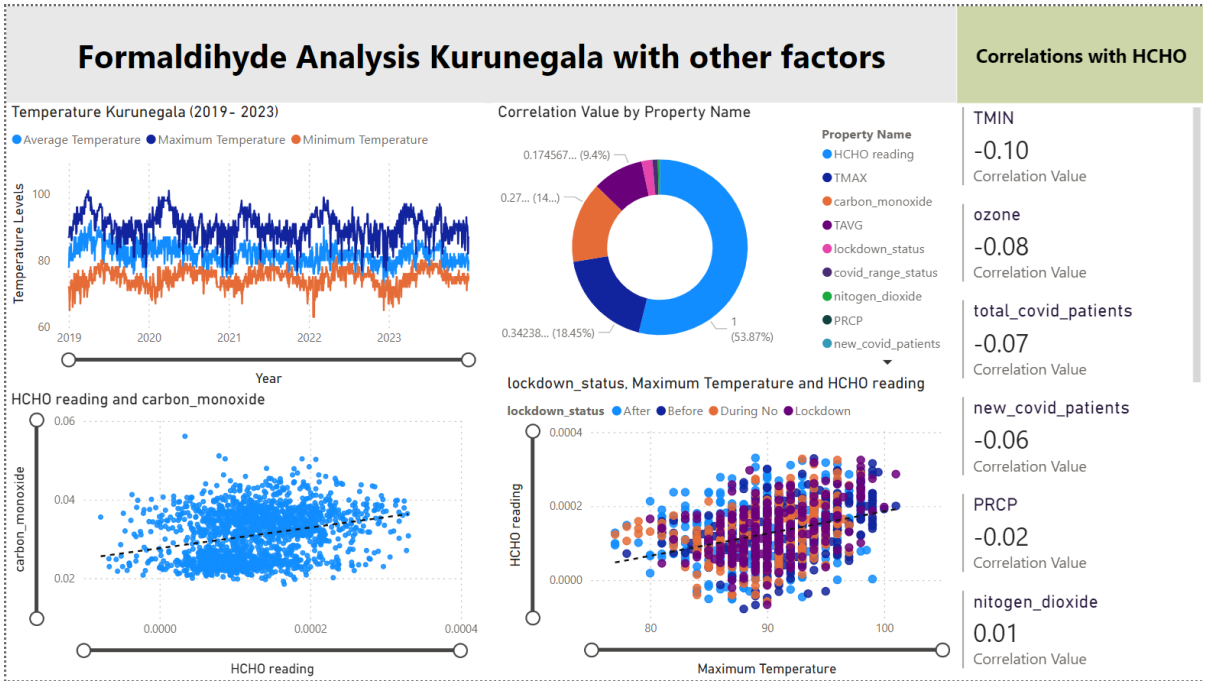


Statistical Measurements done for Kurunegala Region

The below table shows the statistical calculations done for Kurunegala Region.

	HCHO reading
count	1826.000000
mean	0.000127
std	0.000067
min	-0.000078
25%	0.000081
50%	0.000121
75%	0.000171
max	0.000330

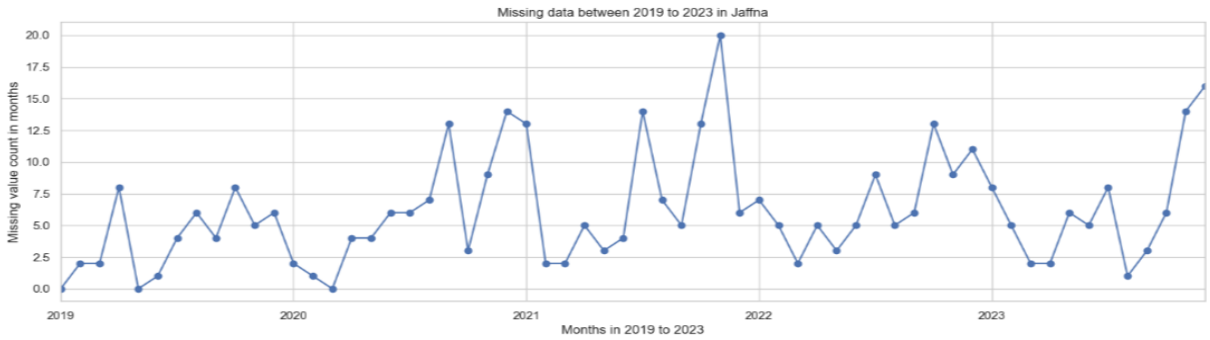

Formaldihyde Analysis Kurunegala Proper
01 January 2019
Earliest Current Date
01 January 2019
Latest Current Date
Average of HCHO reading by Year, Quarter, Month and Day
Average of HCHO reading
0.0001272
Average of HCHO reading
28.57K
Population
60
Proximity(km)
116
ELEVATION
Temperature (F)
63.00
Minimum
81.72
Average
101.00
MAX
Sum of HCHO reading by Year
0.047
0.048
0.045
0.045
0.046
Year
0.1182
Ozone Level
0.2287
Precipitation
0.031
Carbon Monoxide
1.130E-5
Nitrogen Dioxide
Percentage of HCHO reading by Year
Current Date Year
2019
2020
2021
2022
2023

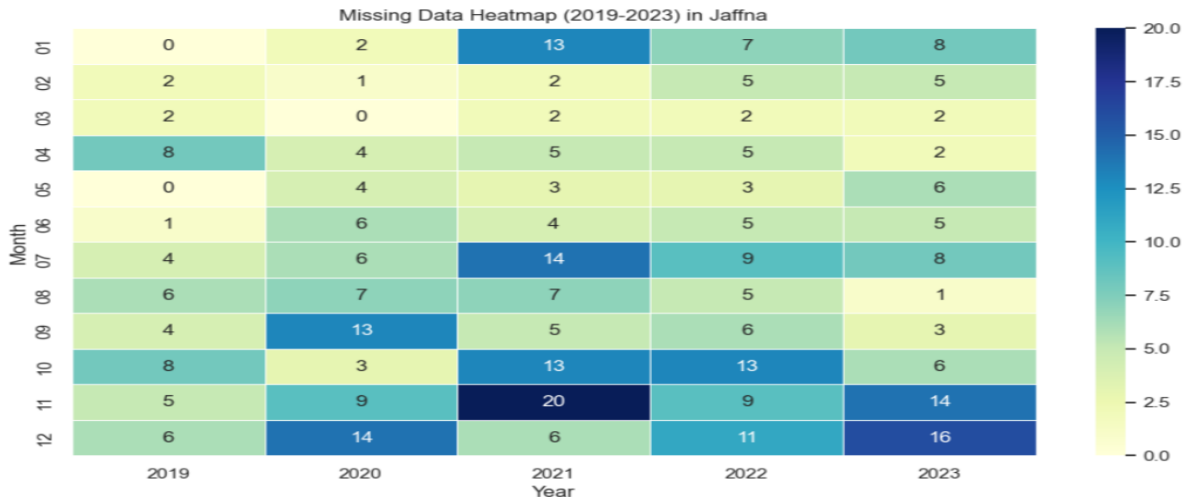

Formaldihyde Analysis Kurunegala with other factors
Correlations with HCHO
Temperature Kurunegala (2019- 2023)
Average Temperature
Maximum Temperature
Minimum Temperature
Temperature Levels
Year
Correlation Value by Property Name
Property Name
HCHO reading
TMAX
carbon_monoxide
TAVG
lockdown_status
covid_range_status
nitogen_dioxide
PRCP
new_covid_patients
HCHO reading and carbon_monoxide
carbon_monoxide
HCHO reading
lockdown_status, Maximum Temperature and HCHO reading
lockdown_status
After
Before
During No
Lockdown
HCHO reading
Maximum Temperature
TMIN
-0.10
Correlation Value
ozone
-0.08
Correlation Value
total_covid_patients
-0.07
Correlation Value
new_covid_patients
-0.06
Correlation Value
PRCP
-0.02
Correlation Value
nitogen_dioxide
0.01
Correlation Value

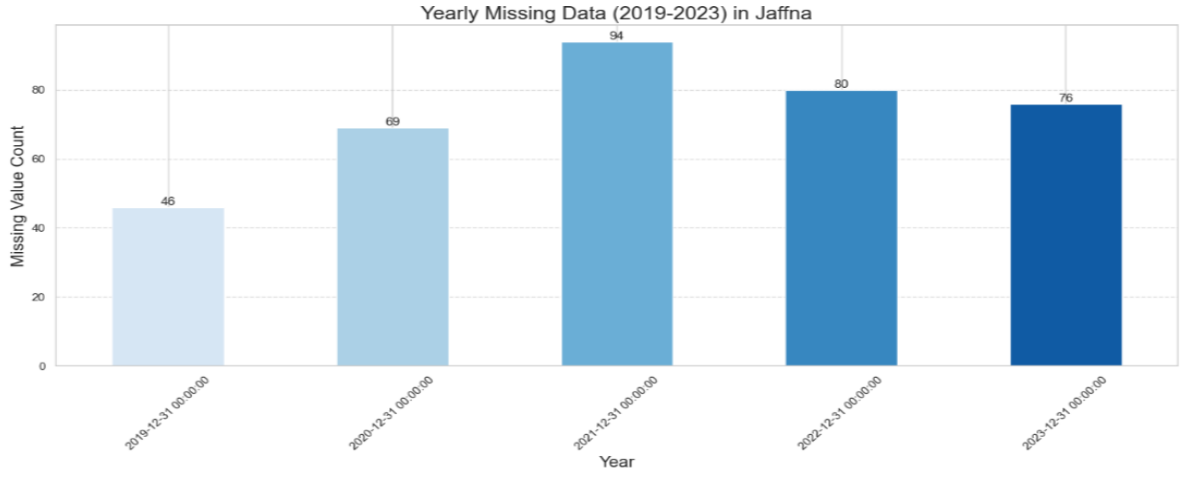
Jaffna Region Formaldehyde Distribution Analysis

Data Preprocessing Jaffna Proper

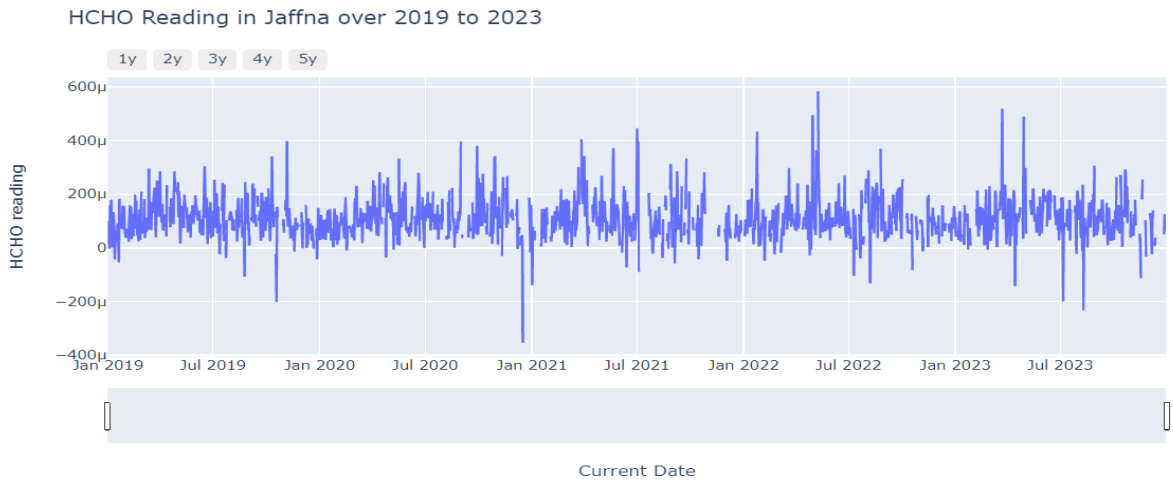
The initial Jaffna Proper dataset consisted of 365 missing values out of 1826 records. It had the fewest null values from the given datasets. The below heatmap, line plot, and bar chart show the distribution of null values.



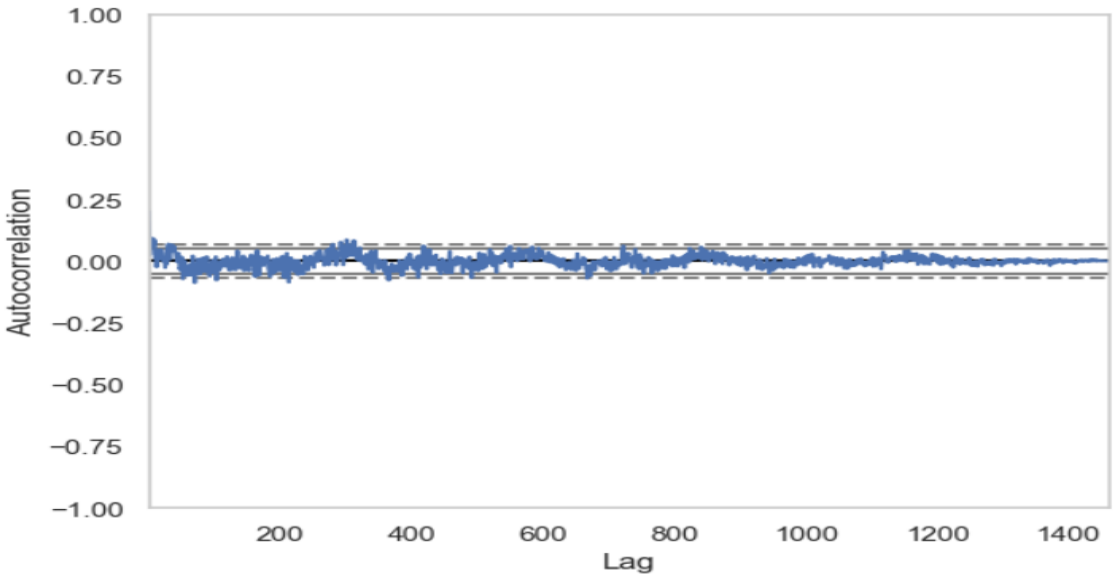




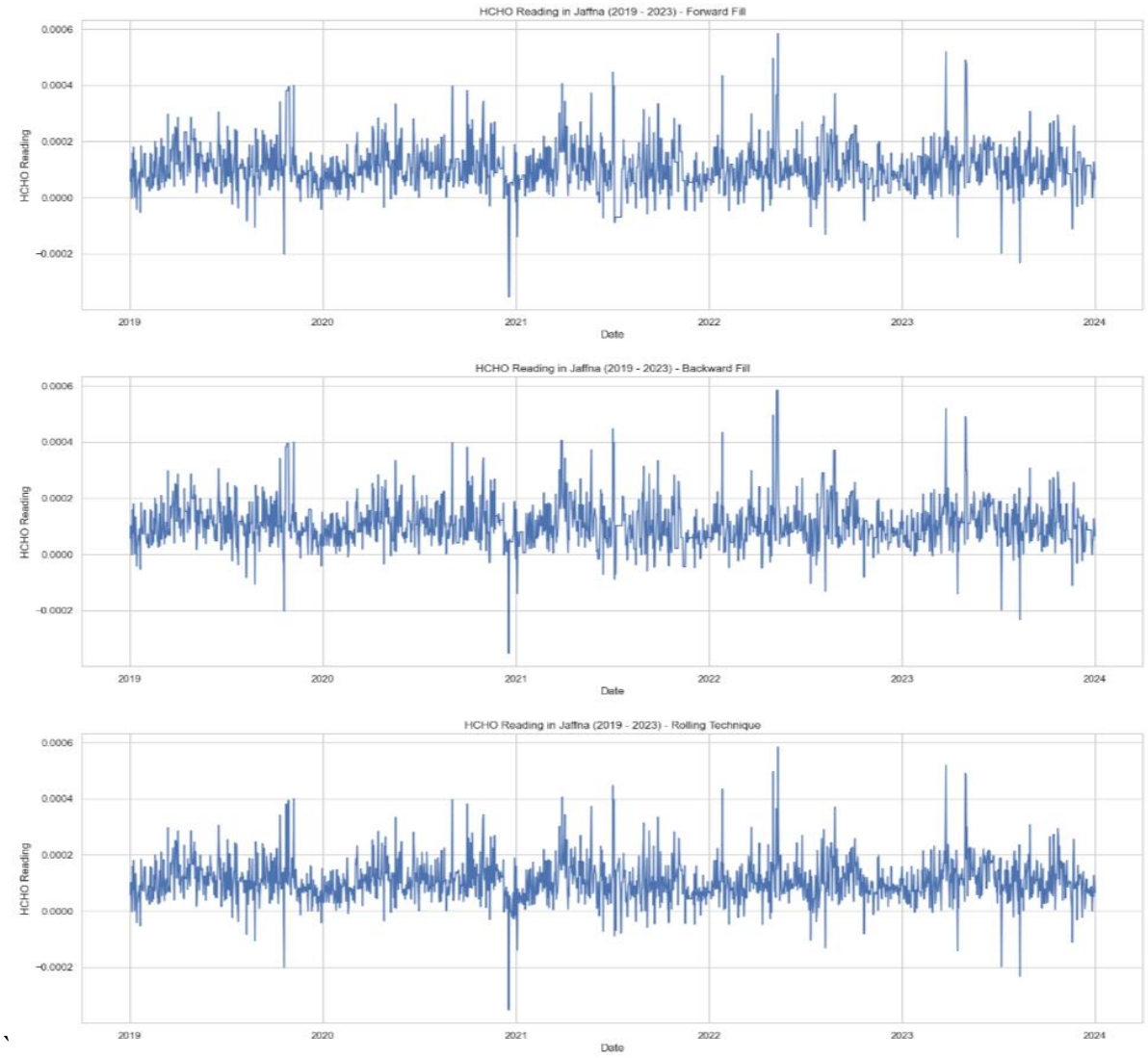
As mentioned in the above heatmap there are months that had 20 null values. The below plot shows the null value distribution of Jaffna Region.



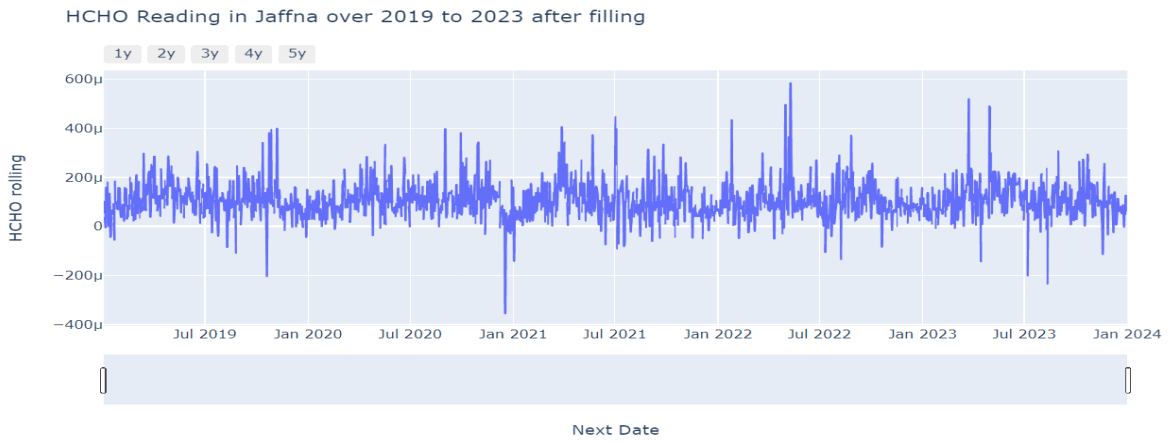
The below Auto-Correlation plot shows there is a very light seasonality in the Jaffna Region dataset. However, it is more likely to be stationary. The window size 13 is used to handle the null values of the Jaffna distribution to maintain its original pattern.



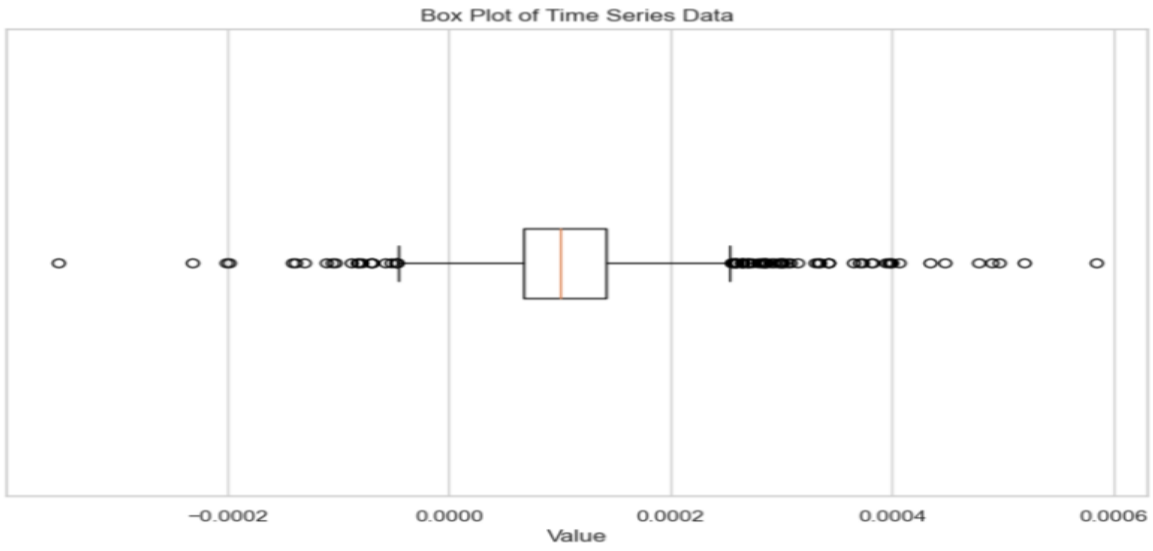
The below line charts show how the HCHO distribution looks when handling missing values with rolling, forward, and backward filling techniques.



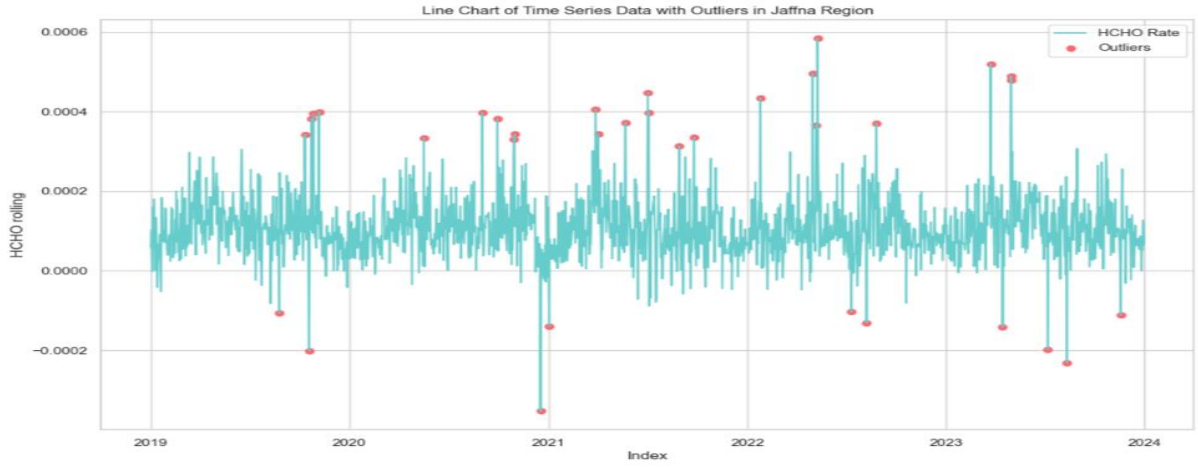
Since the rolling technique with window 13 method has maintained the original pattern with fluctuations, it has been considered as the final null value handled distribution. The below plot shows the final null value handled distribution.

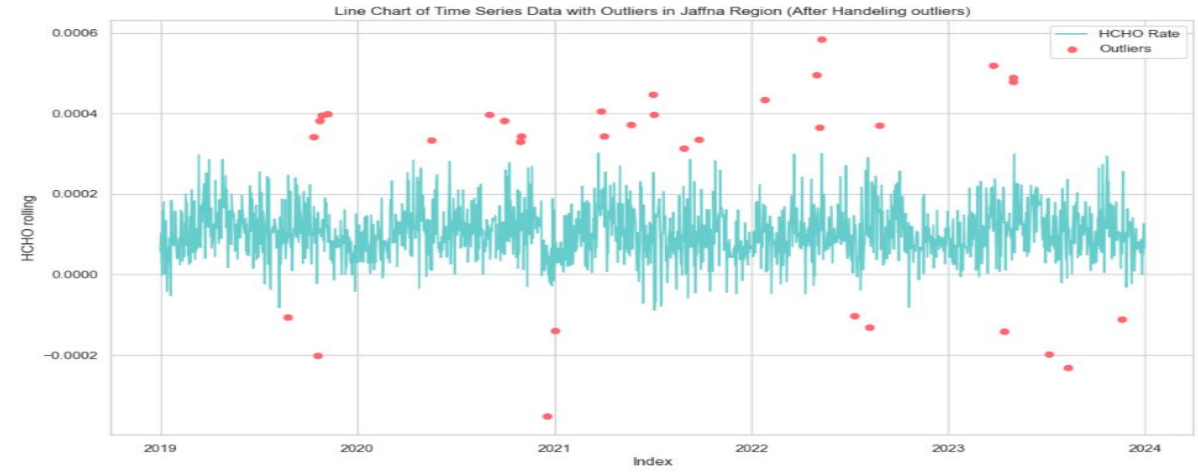


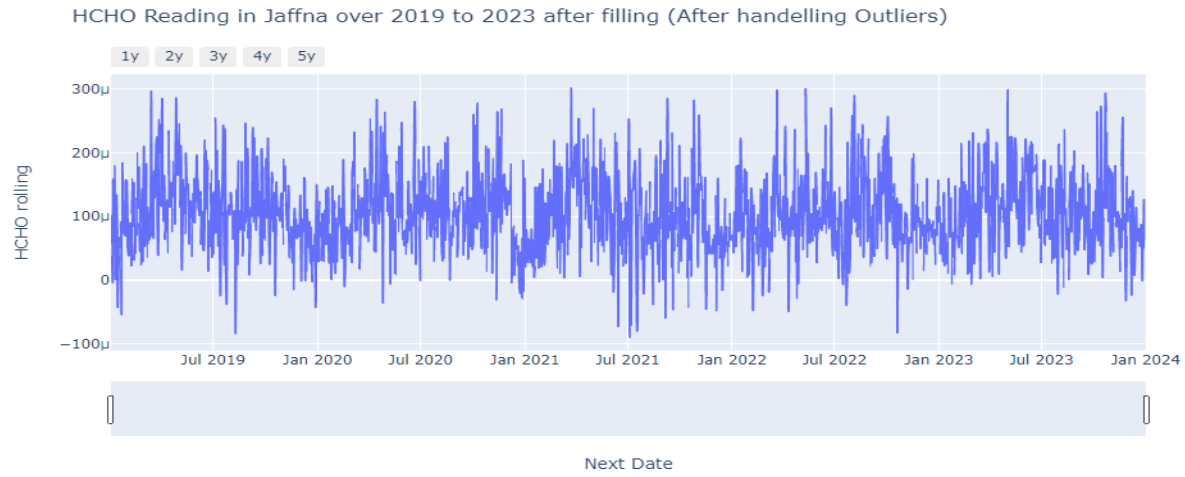
The given boxplot shows the outliers detected by considering the quartiles, but the outliers are removed using a threshold value of 2.25 with inter quartile range.



The below plots show the identified outliers in Jaffna and how the distribution look like after handling the outliers.



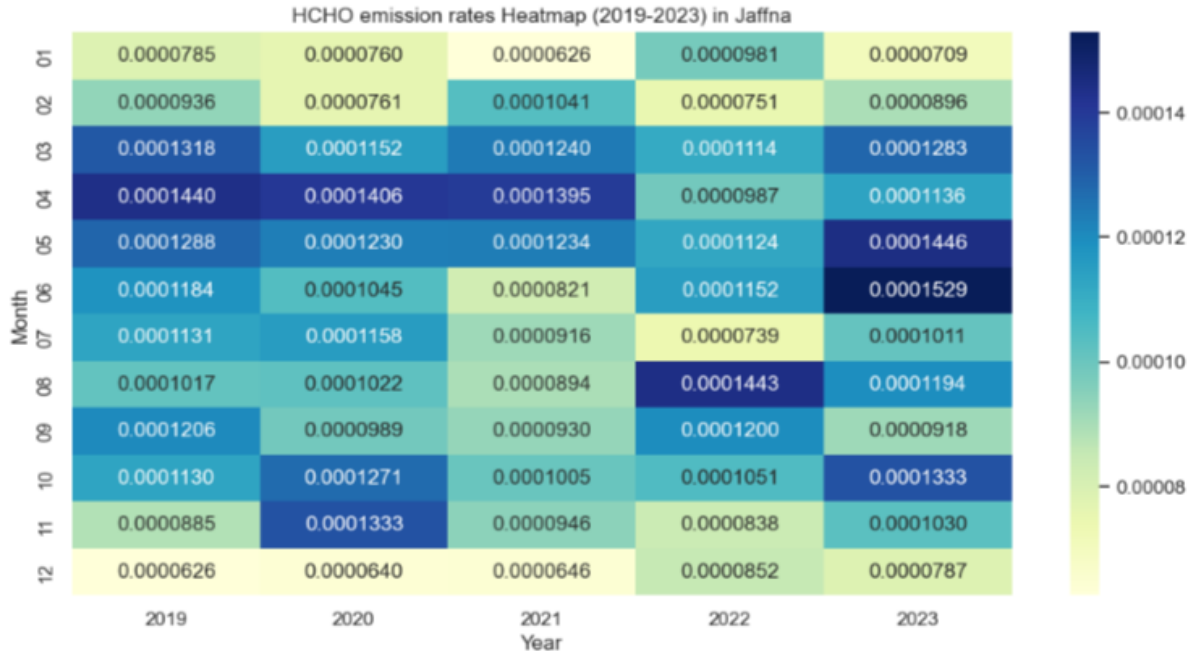


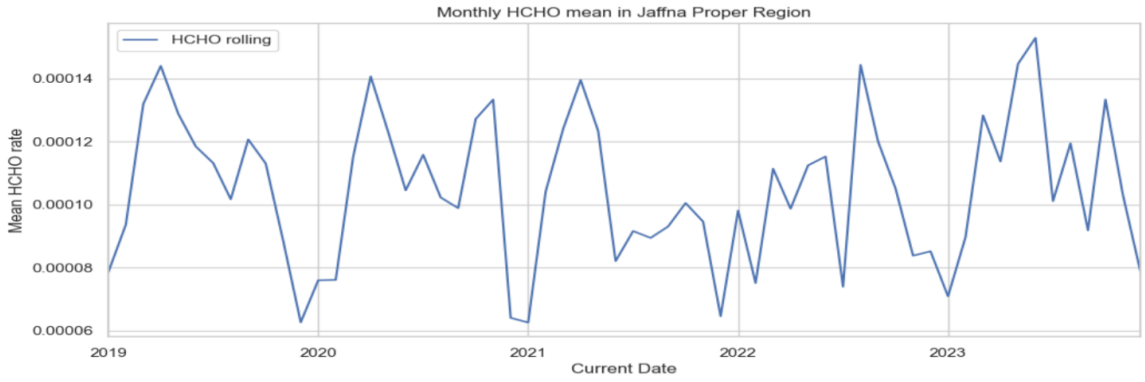


HCHO distribution Insights Jaffna Proper

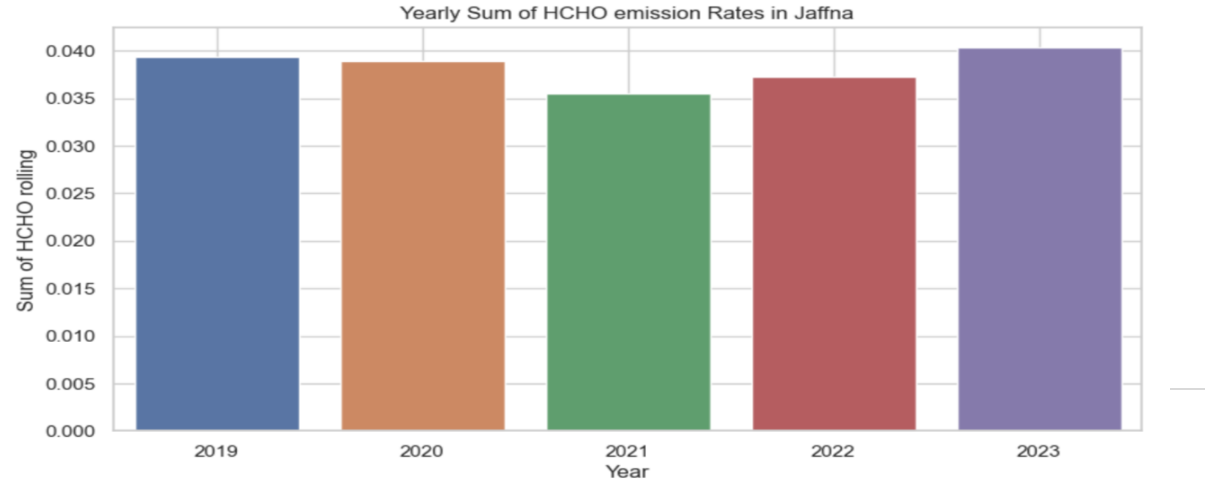
The below visualizations show that in the Jaffna Proper region, there is a high HCHO emission in the middle months of the year and it has reported the least emission rate in December and

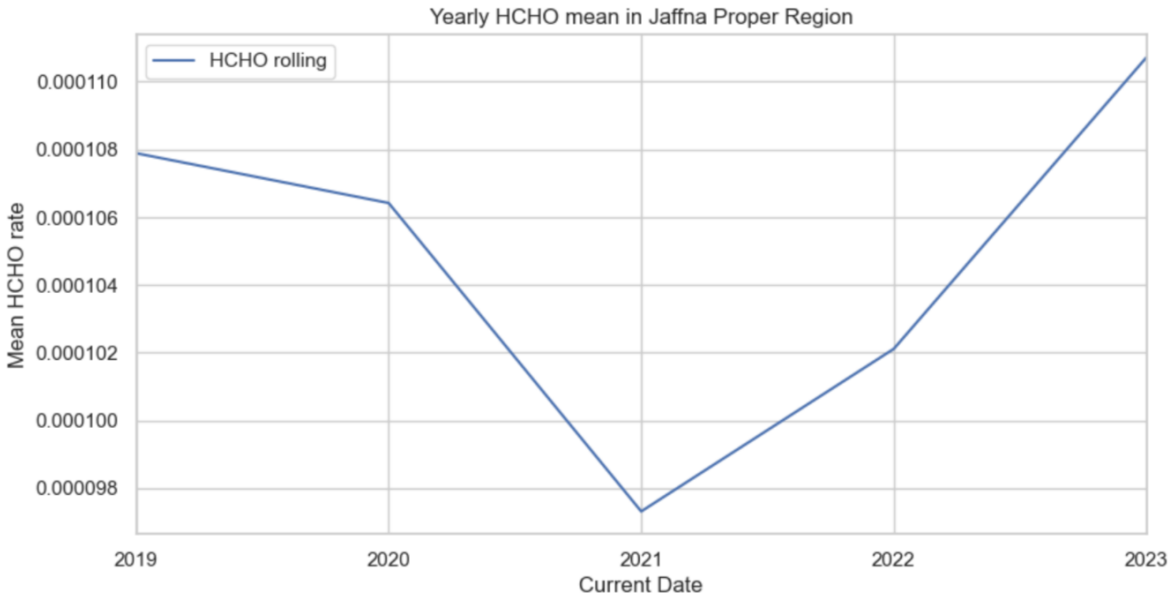
January. It is not like in other regions, it has reported a high HCHO emission constantly in other months of the year. The below visualizations describe how HCHO distributions vary their values monthly, weekly, and yearly.

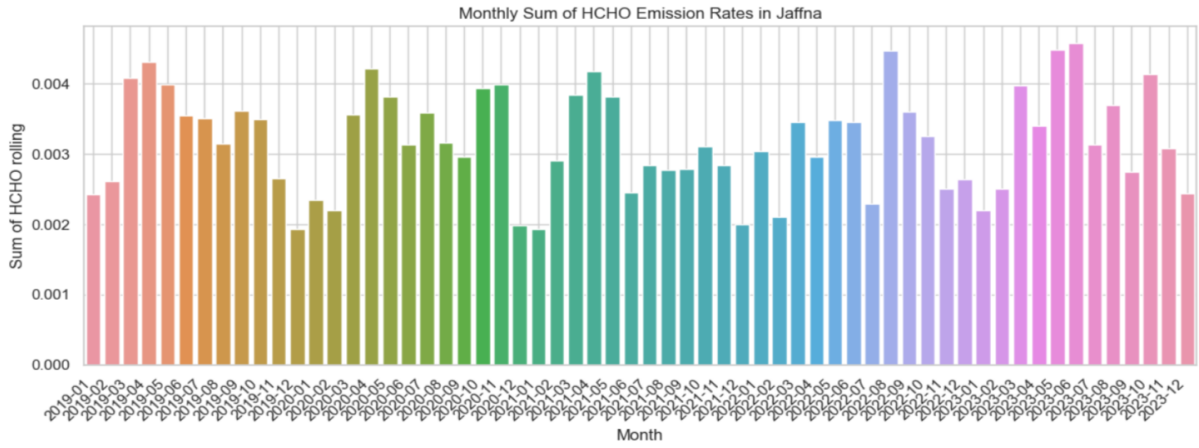


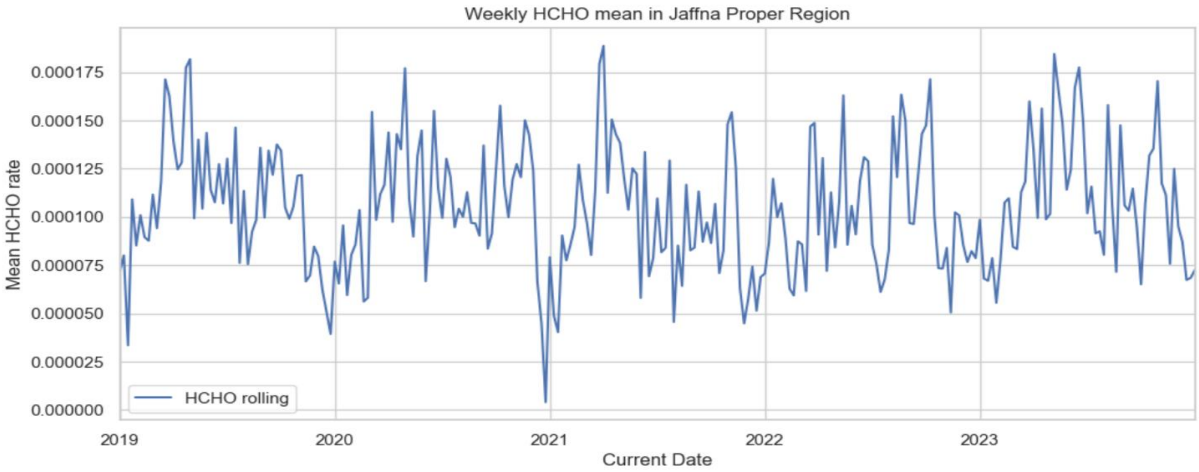


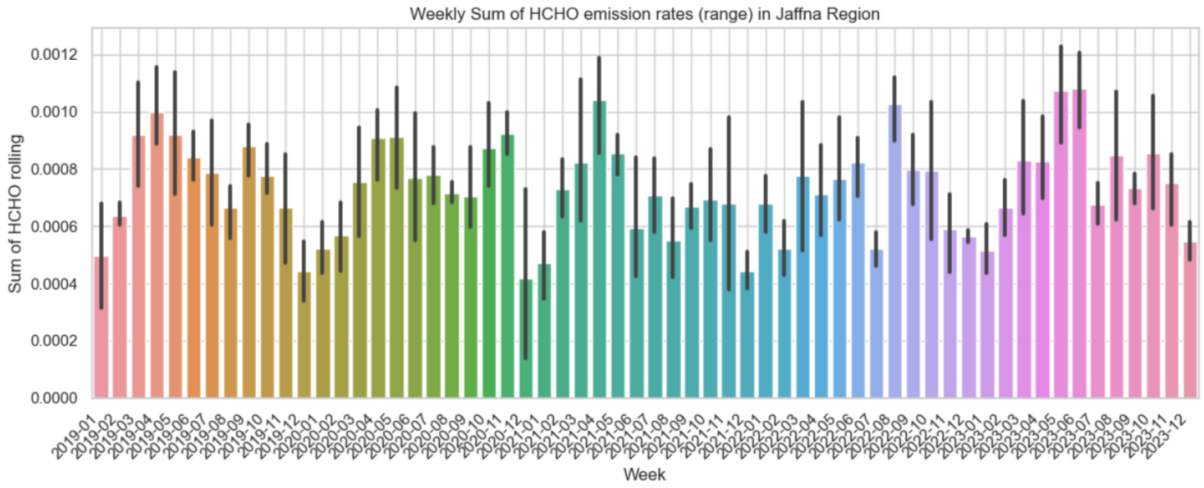
Jaffna region has reported the least HCHO emission in 2021, as in the majority of regions. HCHO emissions have decreased until 2021, and it has started to increase after that year.




Yearly HCHO mean in Jaffna Proper Region
HCHO rolling
Mean HCHO rate
0.000110
0.000108
0.000106
0.000104
0.000102
0.000100
0.000098
2019
2020
2021
2022
2023
Current Date


Monthly Sum of HCHO Emission Rates in Jaffna
Sum of HCHO rolling
0.004
0.003
0.002
0.001
0.000
Month

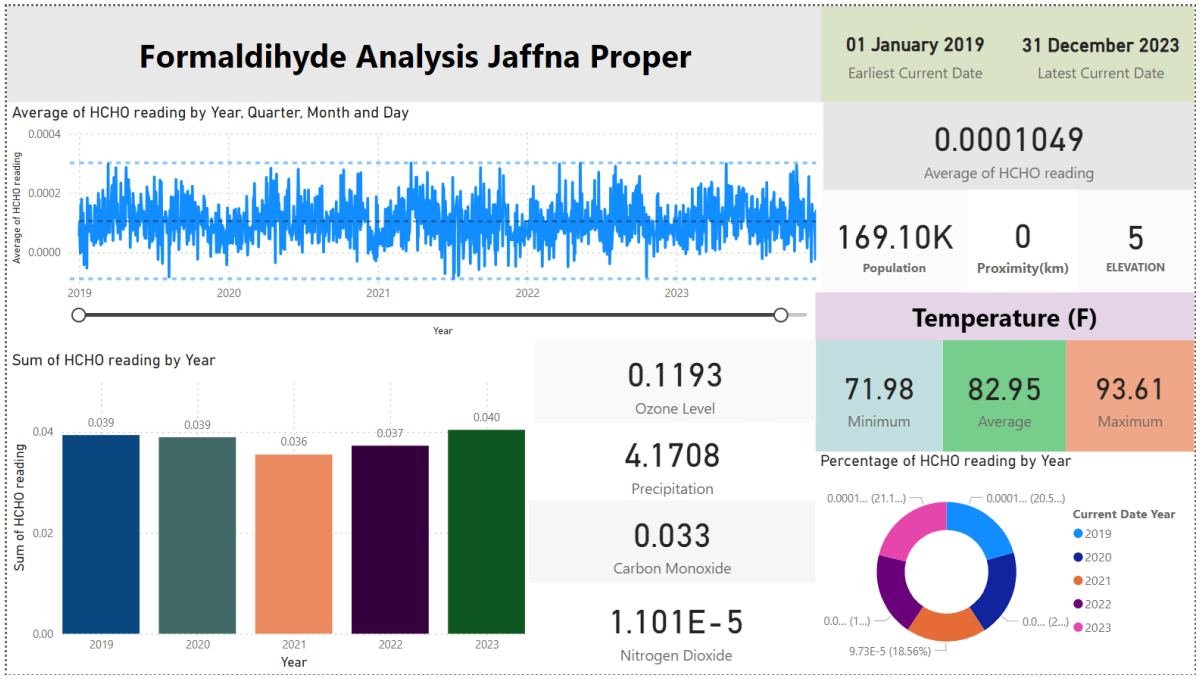

Weekly HCHO mean in Jaffna Proper Region
Mean HCHO rate
0.000175
0.000150
0.000125
0.000100
0.000075
0.000050
0.000025
0.000000
HCHO rolling
2019
2020
2021
2022
2023
Current Date

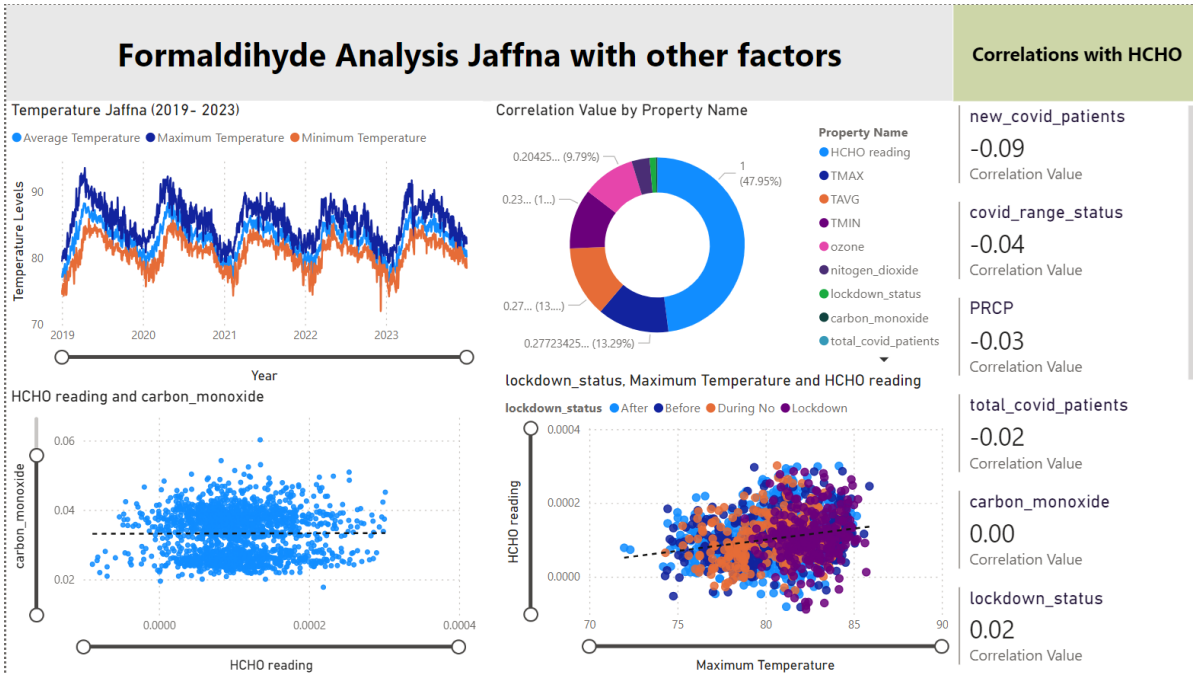


Statistical Measurements done for Jaffna Region

The below table shows the statistical calculations done for Jaffna Region.

	HCHO reading
count	1826.000000
mean	0.000105
std	0.000060
min	-0.000089
25%	0.000067
50%	0.000101
75%	0.000139
max	0.000302

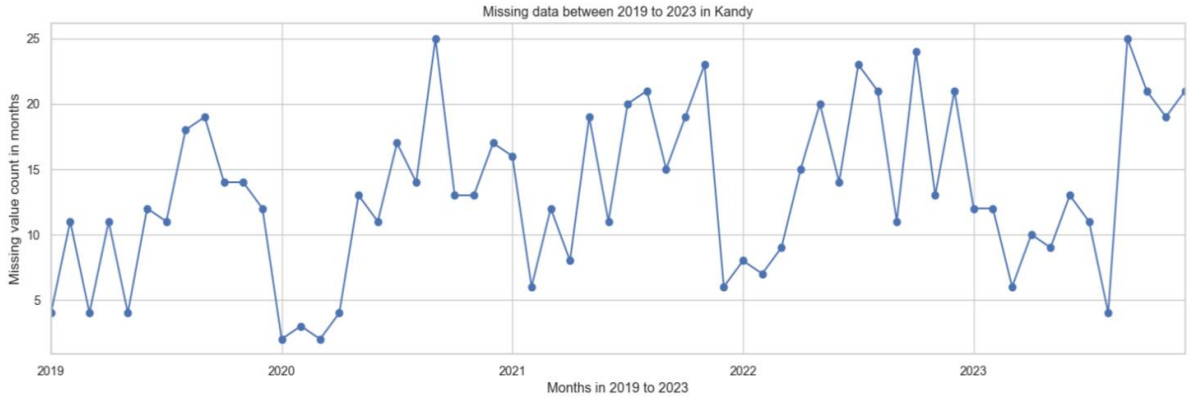

Formaldihyde Analysis Jaffna Proper
01 January 2019
Earliest Current Date
31 December 2023
Latest Current Date
Average of HCHO reading by Year, Quarter, Month and Day
Average of HCHO reading
0.0001049
Average of HCHO reading
169.10K
Population
0
Proximity(km)
5
ELEVATION
Temperature (F)
71.98
Minimum
82.95
Average
93.61
Maximum
Sum of HCHO reading by Year
Sum of HCHO reading
Year
0.1193
Ozone Level
4.1708
Precipitation
0.033
Carbon Monoxide
1.101E-5
Nitrogen Dioxide
Percentage of HCHO reading by Year
Current Date Year
2019
2020
2021
2022
2023

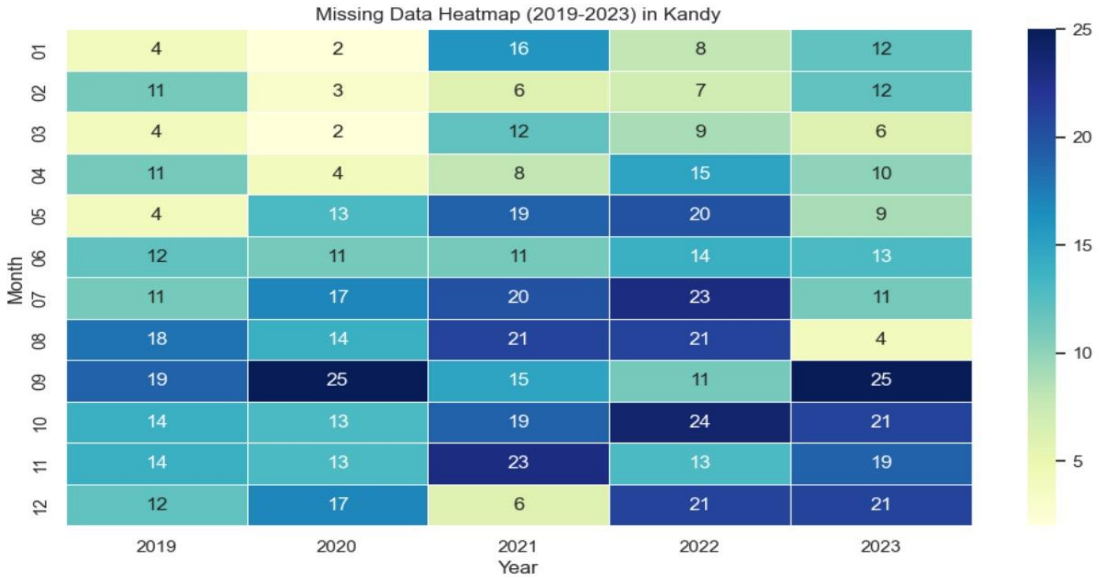

Formaldihyde Analysis Jaffna with other factors
Correlations with HCHO
Temperature Jaffna (2019- 2023)
Average Temperature
Maximum Temperature
Minimum Temperature
Temperature Levels
Year
Correlation Value by Property Name
Property Name
HCHO reading
TMAX
TAVG
TMIN
ozone
nitogen_dioxide
lockdown_status
carbon_monoxide
total_covid_patients
HCHO reading and carbon_monoxide
carbon_monoxide
HCHO reading
lockdown_status, Maximum Temperature and HCHO reading
lockdown_status
After
Before
During No
Lockdown
HCHO reading
Maximum Temperature
new_covid_patients
-0.09
Correlation Value
covid_range_status
-0.04
Correlation Value
PRCP
-0.03
Correlation Value
total_covid_patients
-0.02
Correlation Value
carbon_monoxide
0.00
Correlation Value
lockdown_status
0.02
Correlation Value

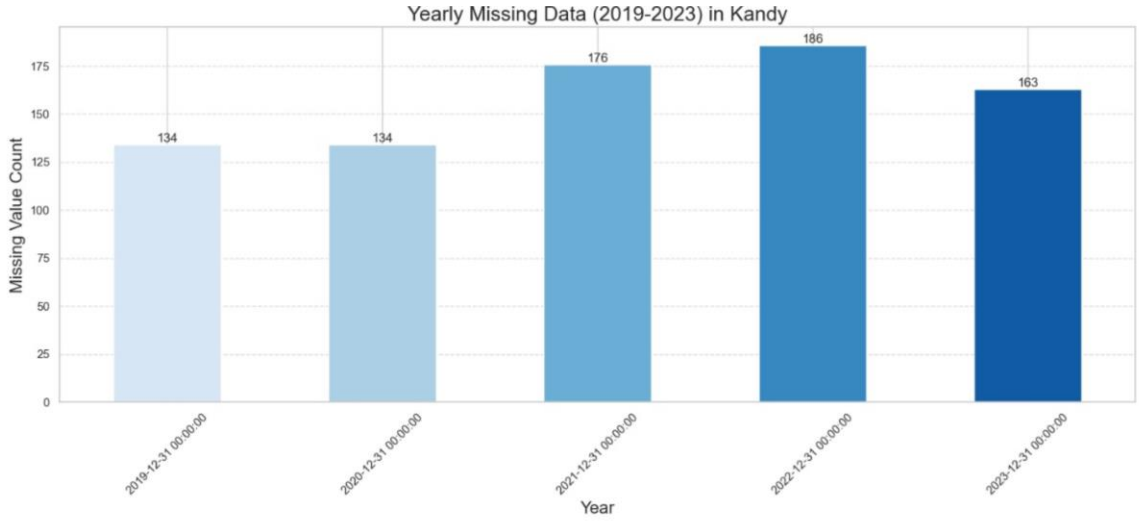
Kandy Region Formaldehyde Distribution Analysis

Data Preprocessing Kandy Proper

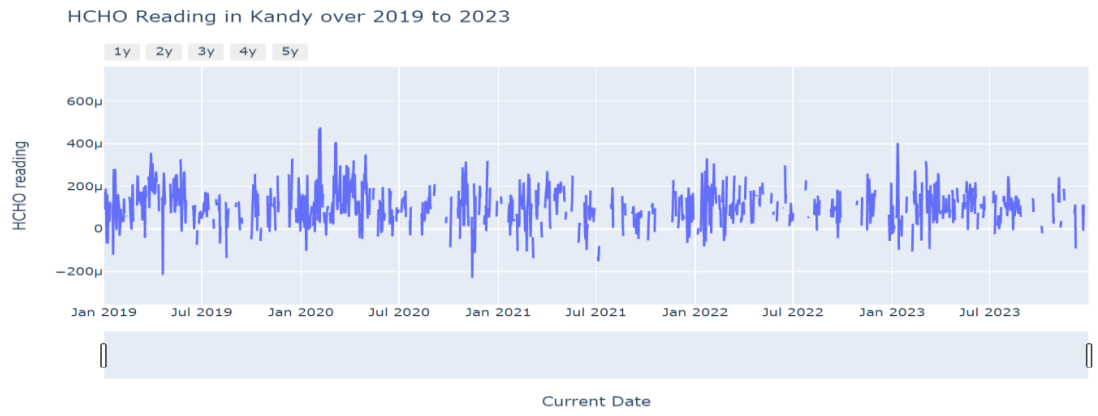
In the initial dataset there were 793 null values out 1826 records. It is around 43 percent of the whole dataset. The below plots show the null value distribution in Kandy Proper region.



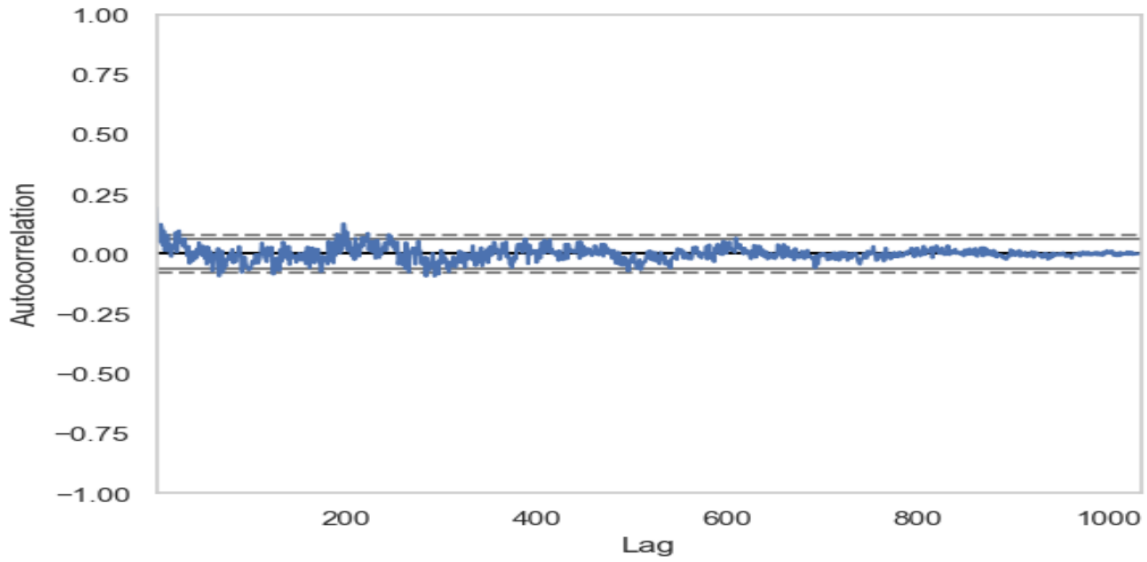




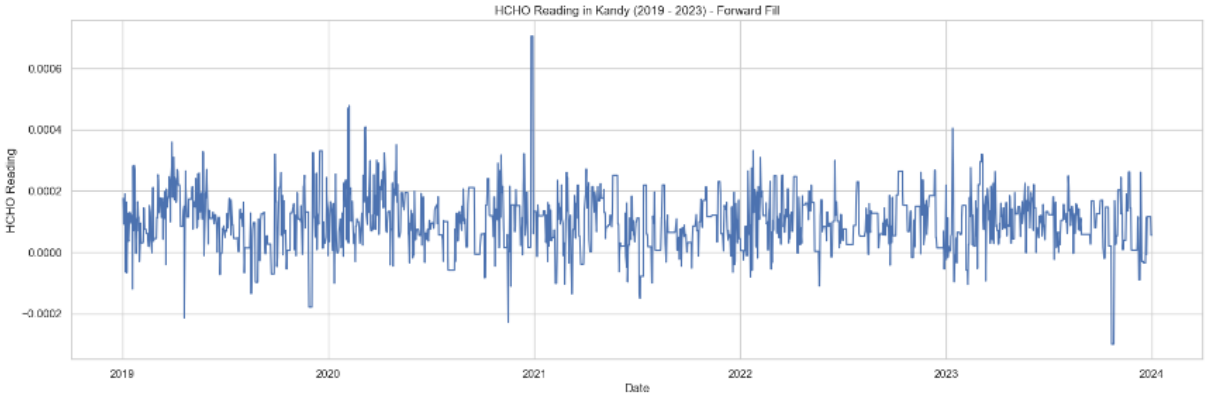
The above heatmap shows there are some months that consisted with 25 null values. The below line chart shows the distribution of HCHO emission rates in Kandy Proper region before handling the null values.

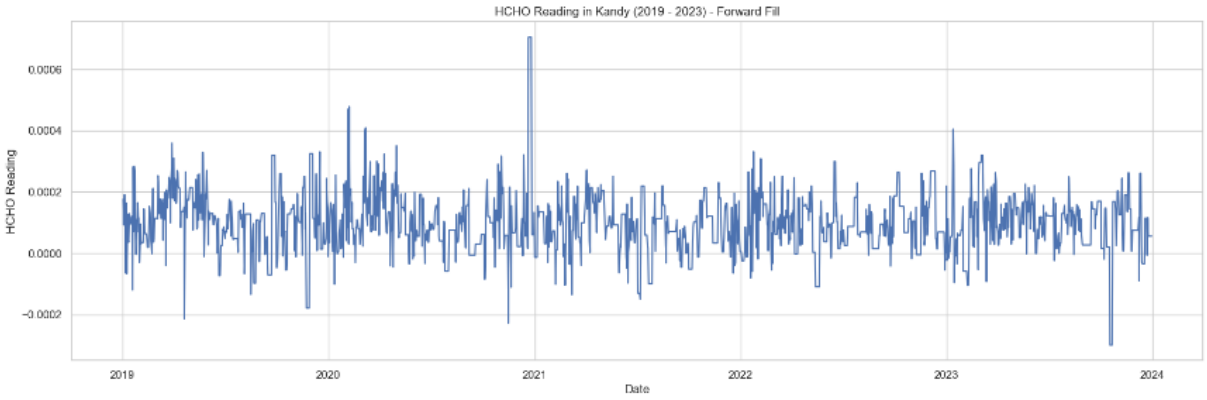


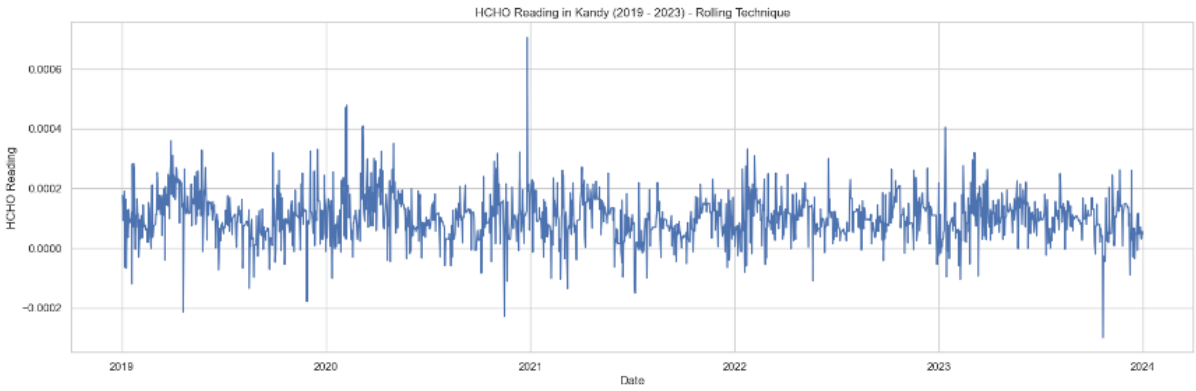
The below auto-correlation plot shows there is a very low seasonality in the Kandy dataset with slight fluctuations. It is more likely a stationary dataset. The window size 17 is used to fill the null values to maintain its seasonality.



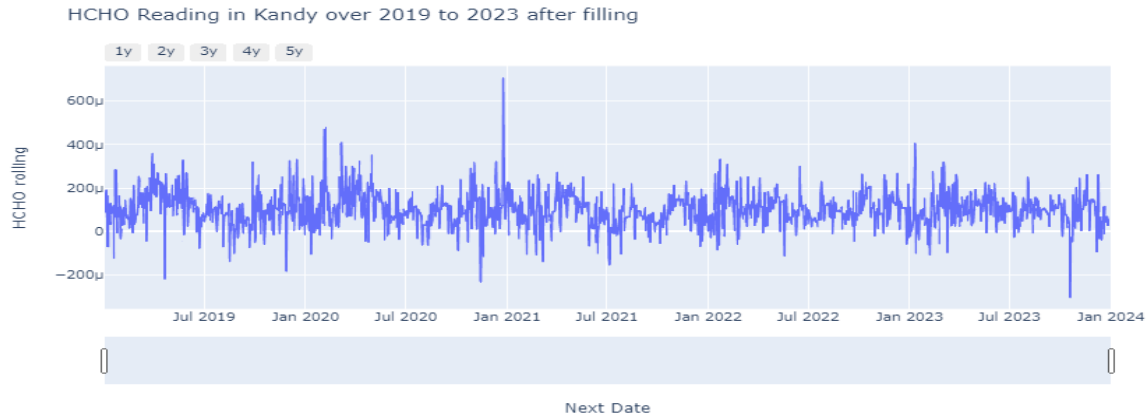
The below line charts show how the HCHO distribution of Kandy looks when handling missing values with rolling, forward, and backward filling techniques.



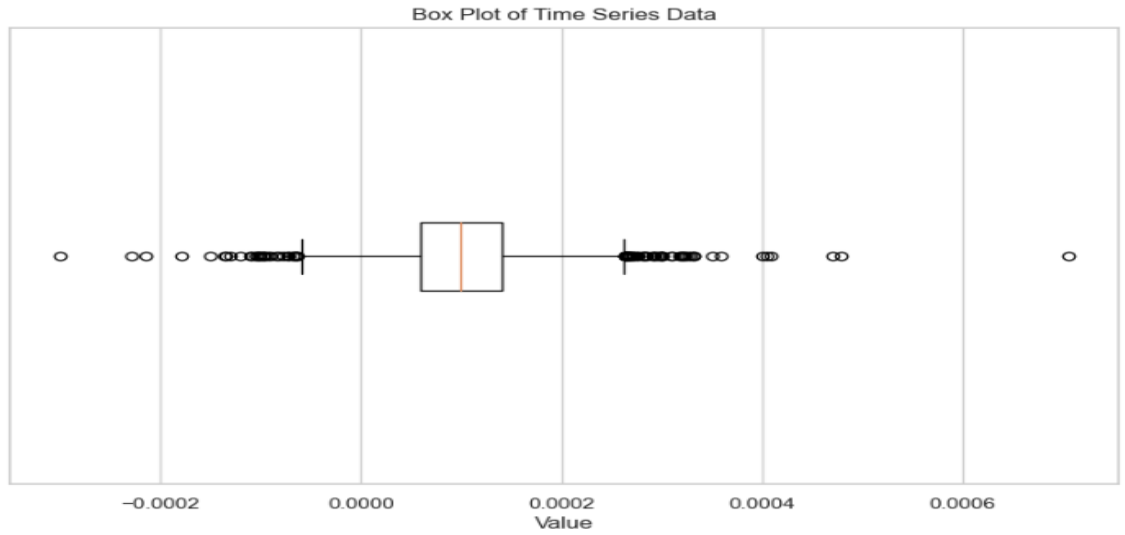




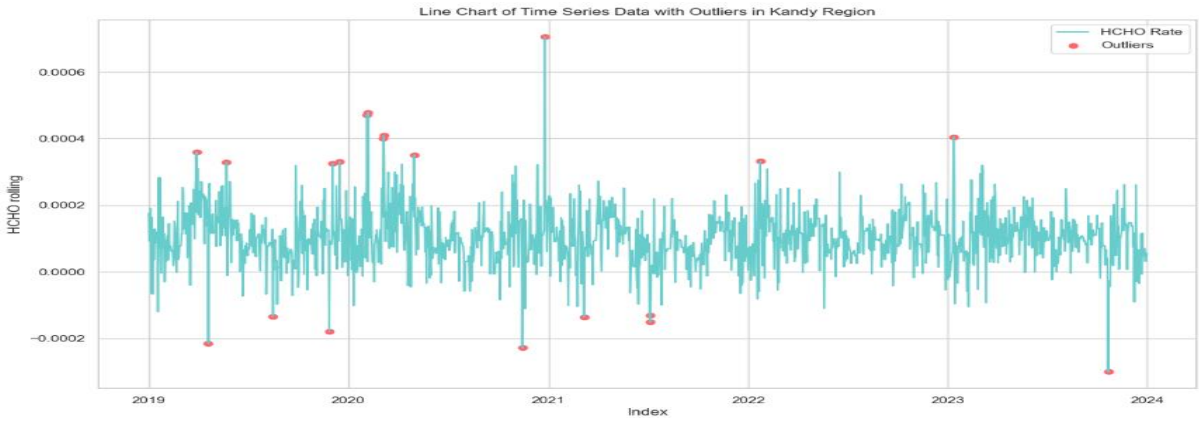
When compared to the above three plots, it shows that the HCHO distribution filled with the window size 17 by using the rolling technique has maintained a better fluctuation compared to forward and backward filling techniques. The below plot shows the final null value handled distribution of Kandy.

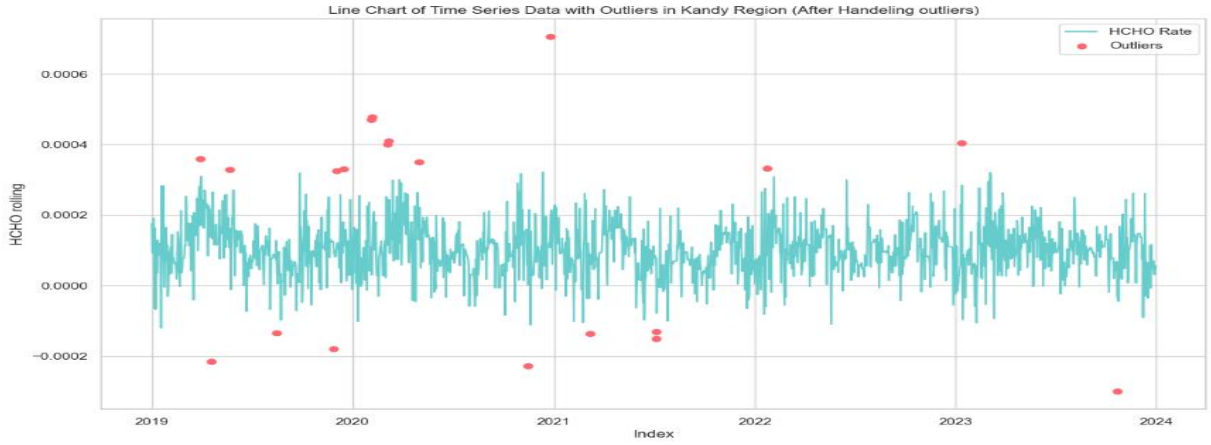


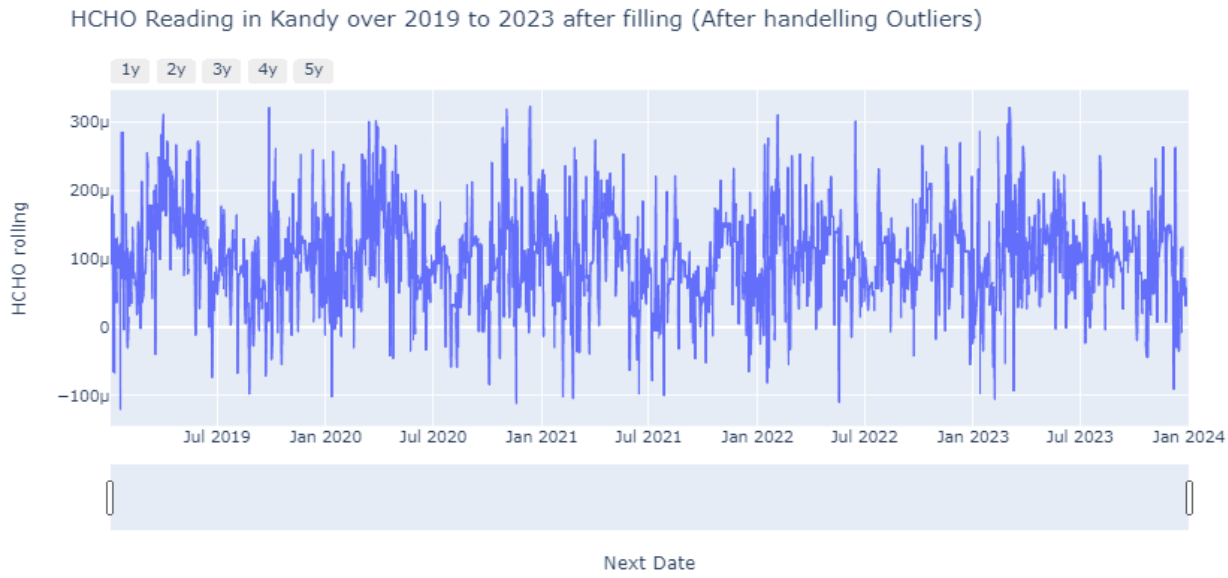
The given boxplot shows the outliers detected by considering the quartiles, but the outliers are removed using a threshold value of 2.25 by considering the inter quartile range.



The below plots show the identified outliers in Kandy and how the distribution look like after handling the outliers.

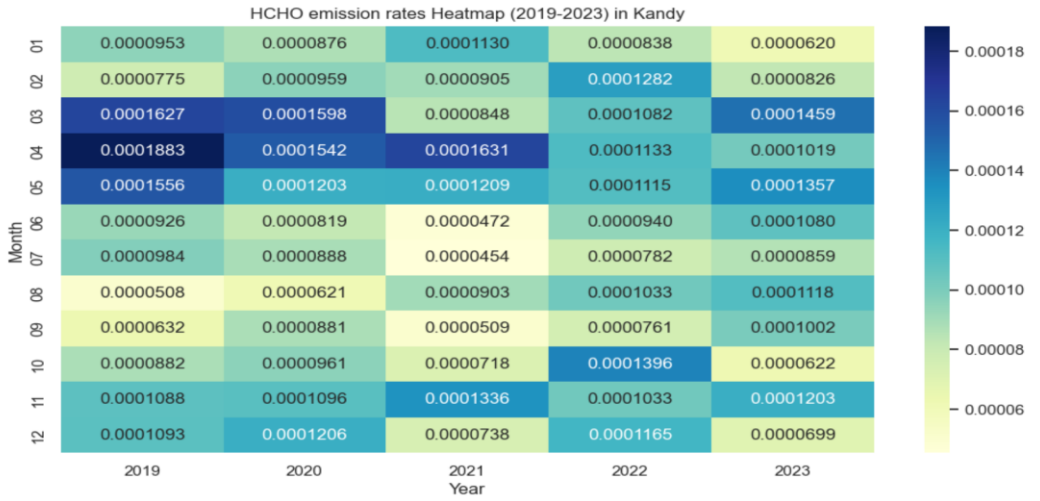


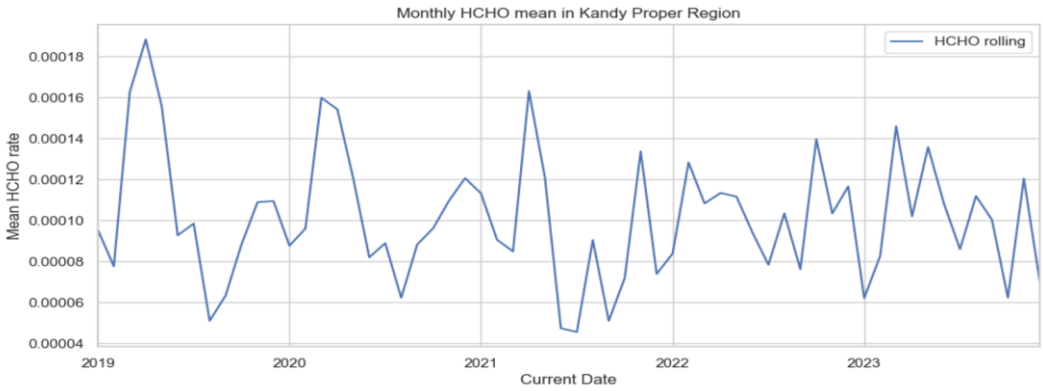




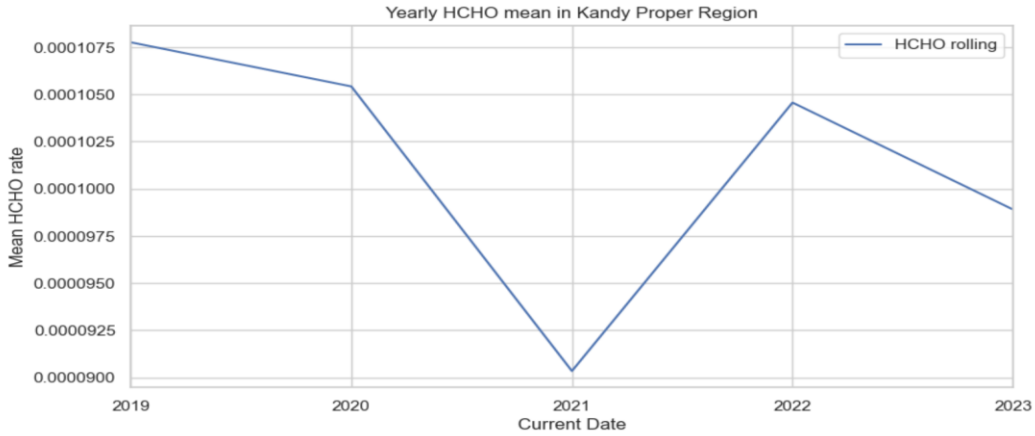
HCHO distribution Insights Kandy Proper Region

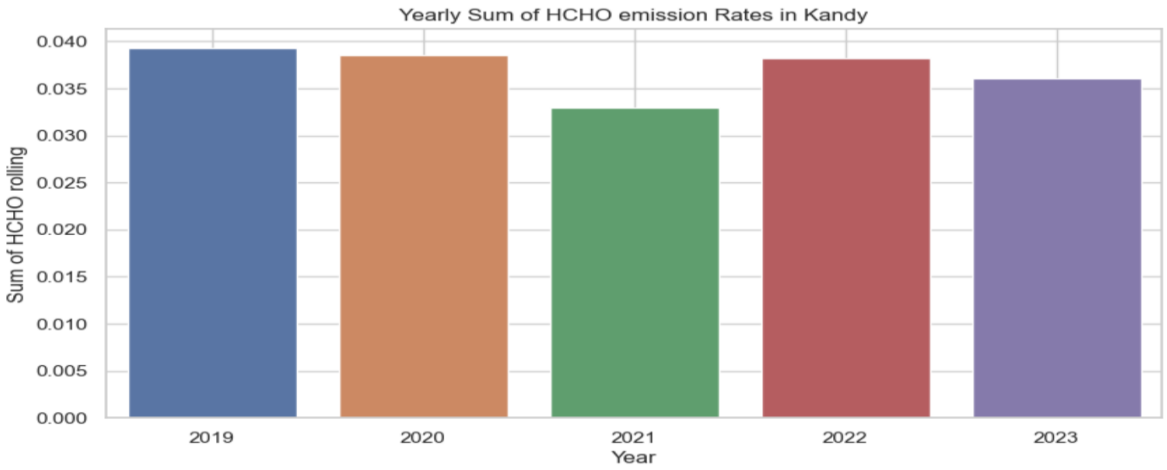
As not in other regions, Kandy region has recorded its highest emission rates in March, April, and May. It has a low emission rate in other months. In addition, it has reported the lowest HCHO mean emission in the year 2021 and it has reported it highest HCHO emission rate in year 2019. The below visualizations describe how HCHO distributions vary their values monthly, weekly, and yearly.

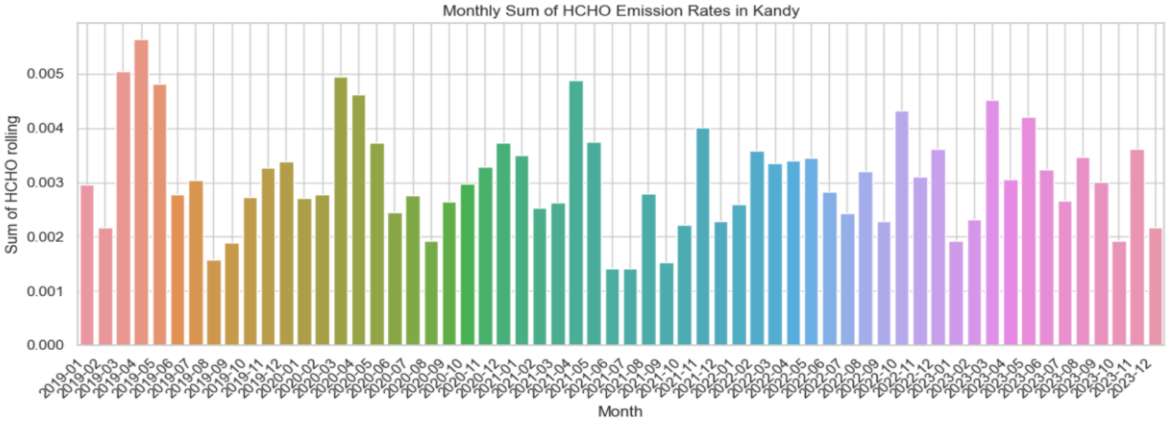


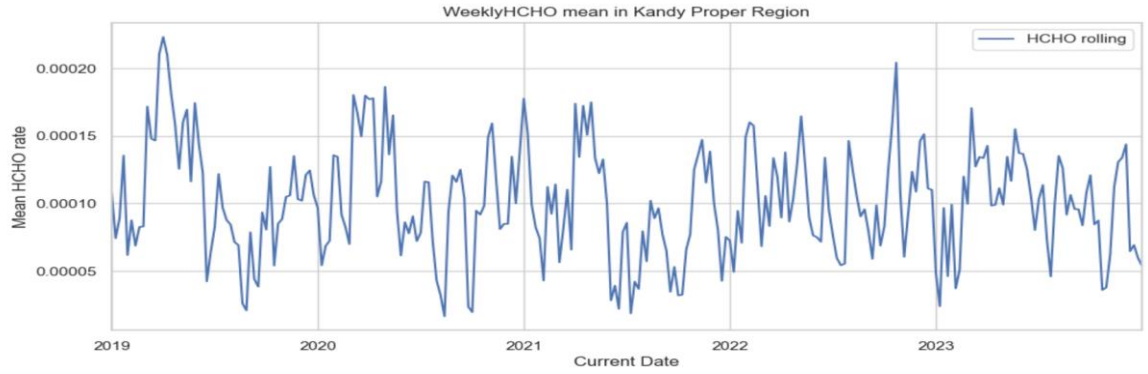


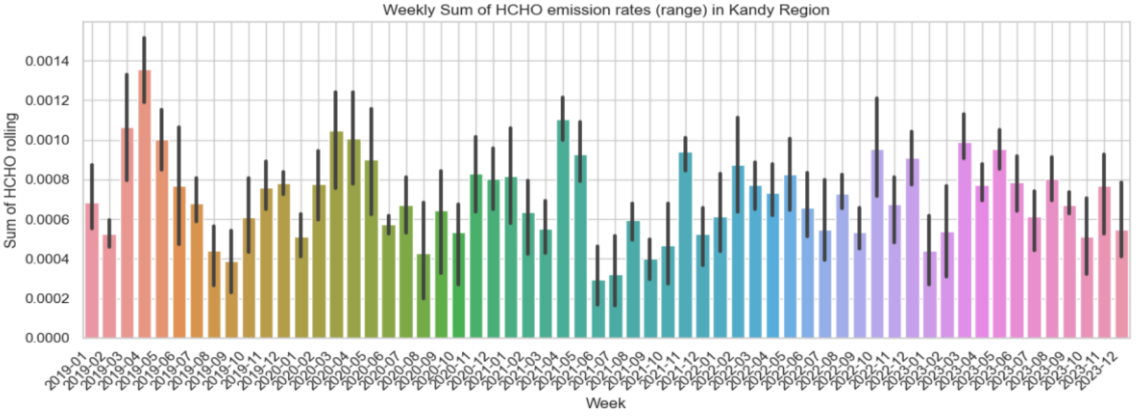
It has recorded the lowest mean HCHO emission rate in year 2021.




Yearly Sum of HCHO emission Rates in Kandy
Sum of HCHO rolling
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
2019
2020
2021
2022
2023
Year


Monthly Sum of HCHO Emission Rates in Kandy
Sum of HCHO rolling
0.005
0.004
0.003
0.002
0.001
0.000
Month

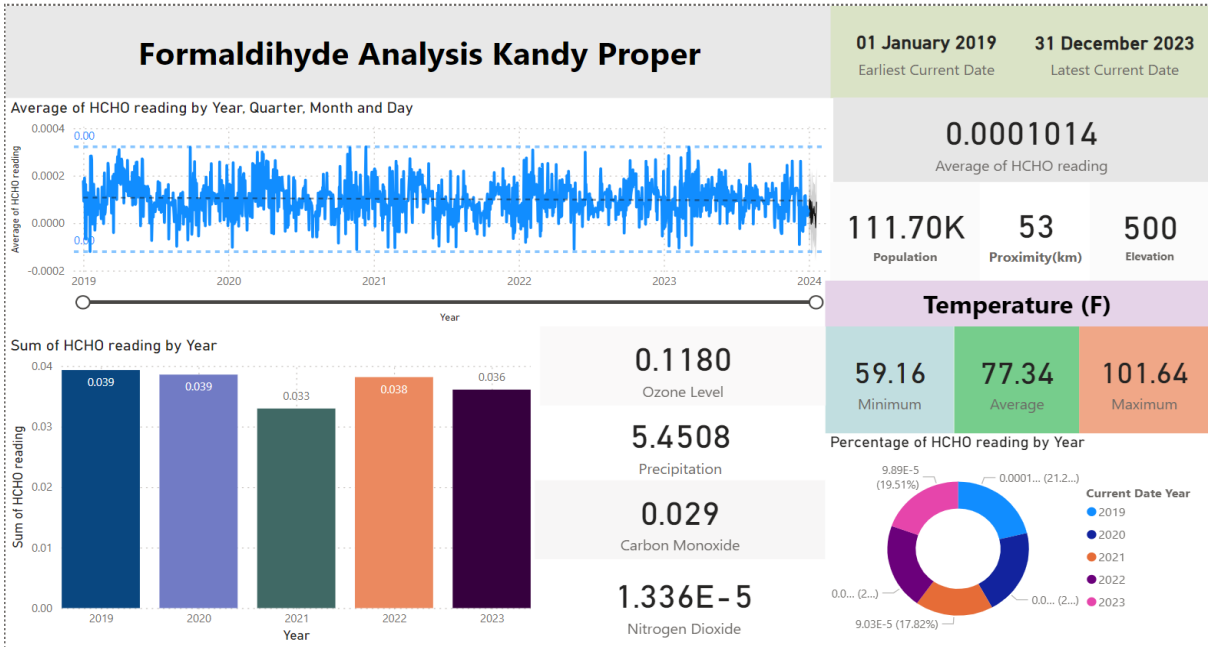

WeeklyHCHO mean in Kandy Proper Region
HCHO rolling
Mean HCHO rate
0.00020
0.00015
0.00010
0.00005
2019
2020
2021
2022
2023
Current Date


Weekly Sum of HCHO emission rates (range) in Kandy Region
Sum of HCHO rolling
0.0014
0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
Week

Statistical Measurements done for Kandy Region

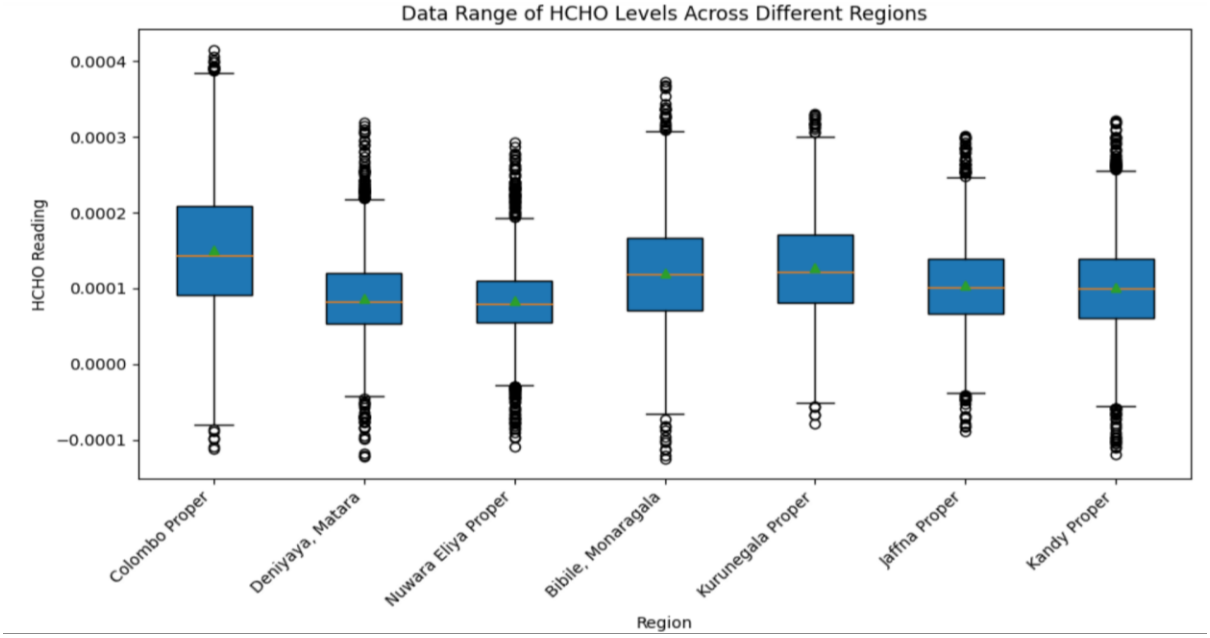
The below table shows the statistical calculations done for the Kandy Region.

	HCHO reading
count	1826.000000
mean	0.000101
std	0.000068
min	-0.000120
25%	0.000061
50%	0.000099
75%	0.000139
max	0.000322

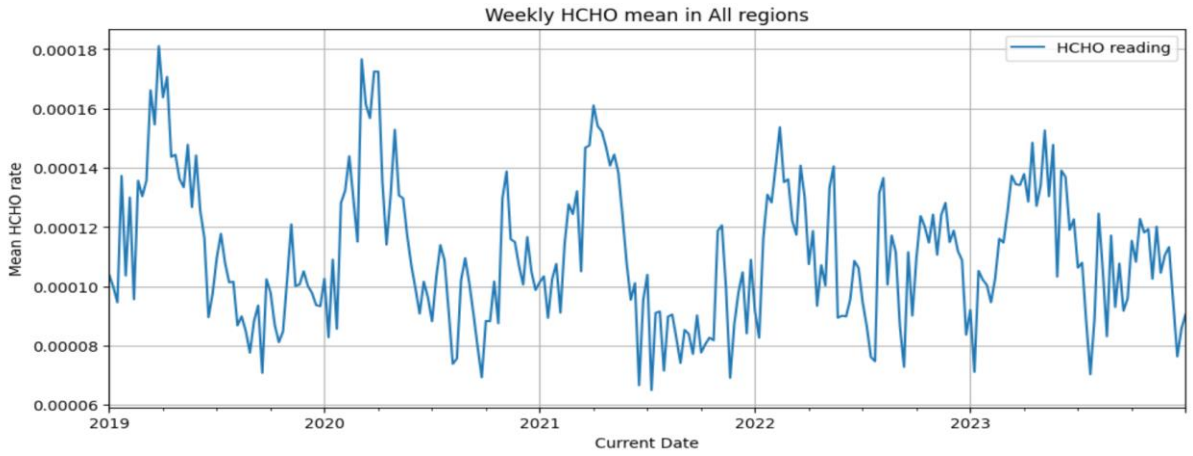


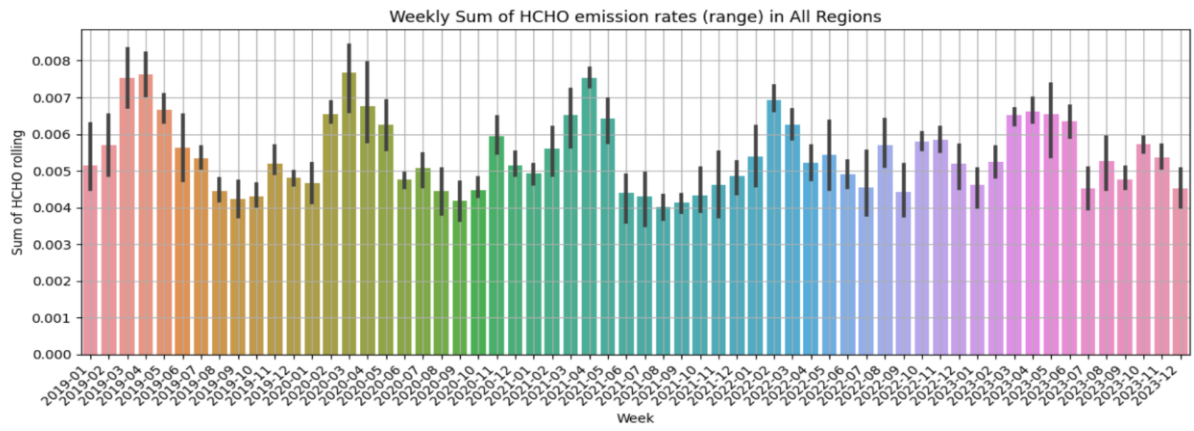
All Regions HCHO distribution Statistics

The below plot shows how the HCHO emission rate spread in each region.

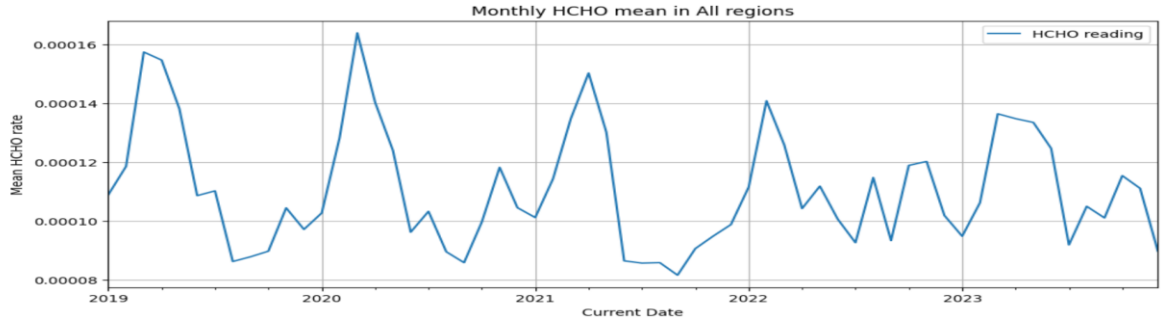


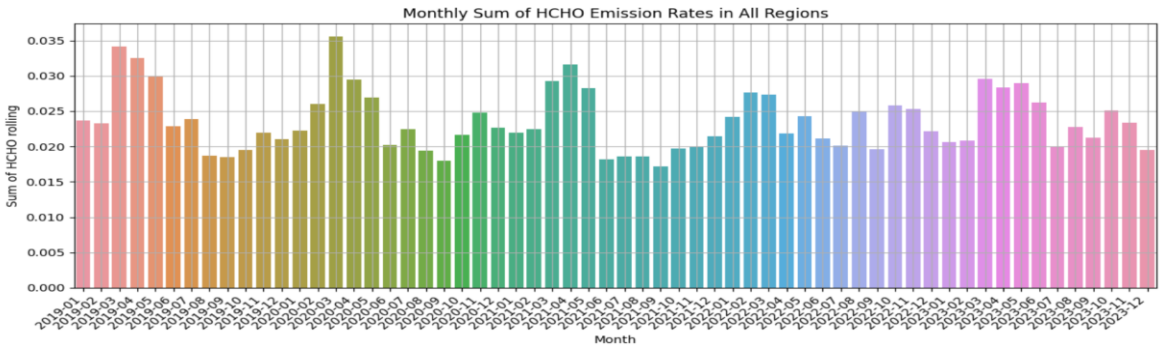
The below plots show the mean and summation of HCHO emission in all regions yearly, weekly, and monthly. It shows there is a high emission rate in March, April, and May.

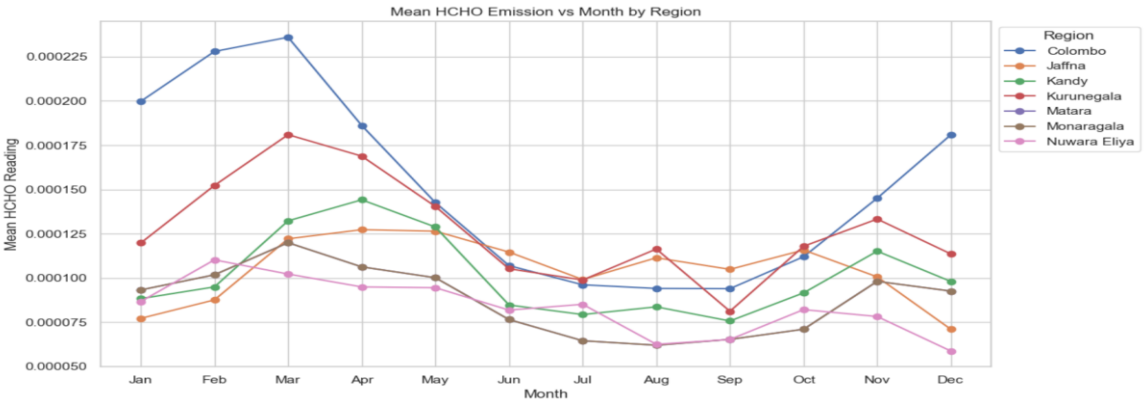




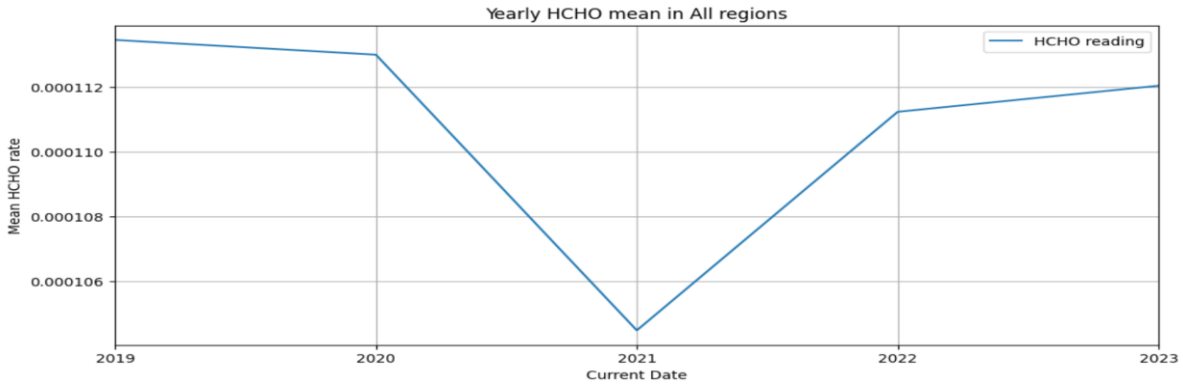
The below plot shows the monthly mean and sum of HCHO distribution in all regions.

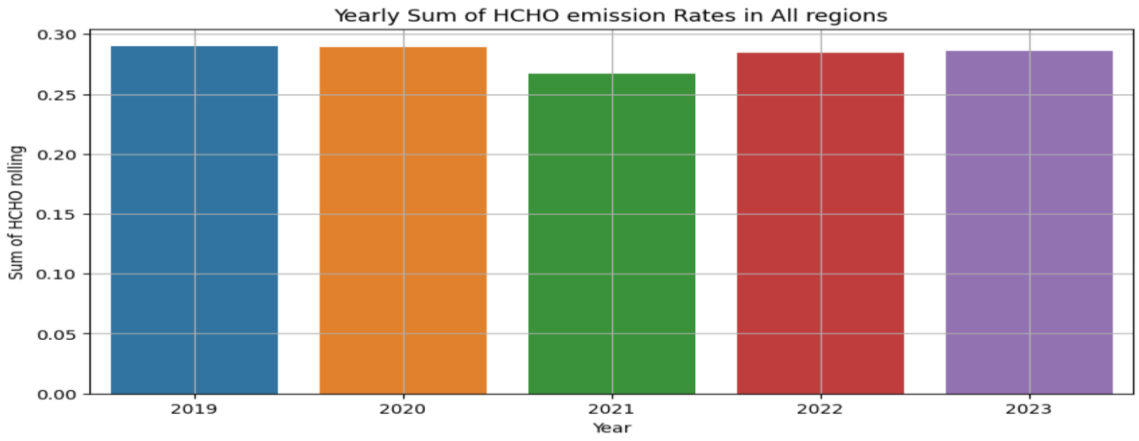






The below plot shows the yearly mean and sum of HCHO distribution in all regions. It also shows that the mean and sum of HCHO distribution has decreased in the year 2021 in all region combined data.





The below table shows the mean, minimum, maximum, and standard deviation values of all regions HCHO distribution.

	HCHO reading
count	12782.000000
mean	0.000111
std	0.000071
min	-0.000125
25%	0.000066
50%	0.000103
75%	0.000152
max	0.000415

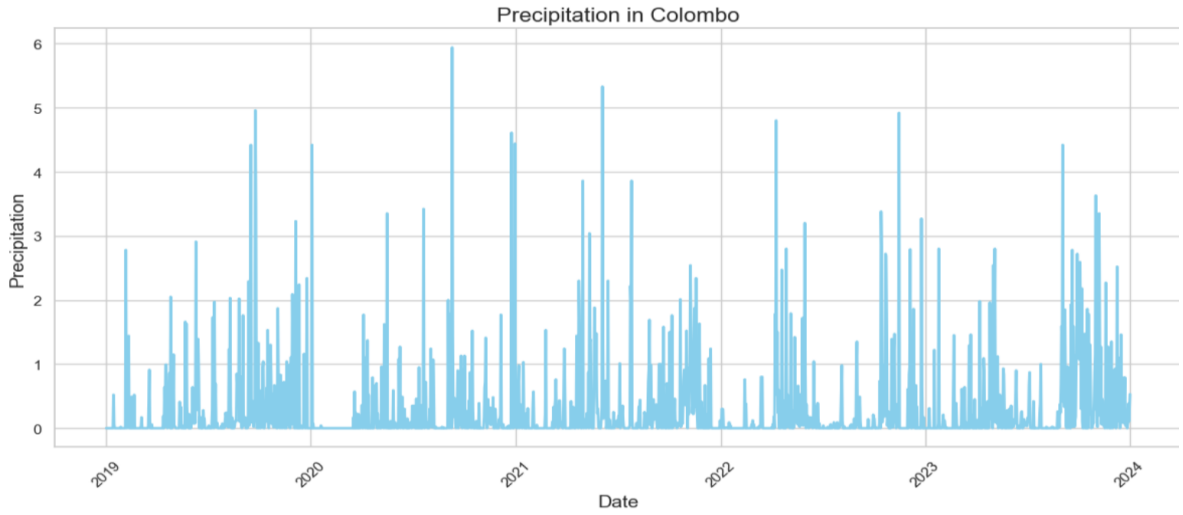
Spatio -Temparal Analysis

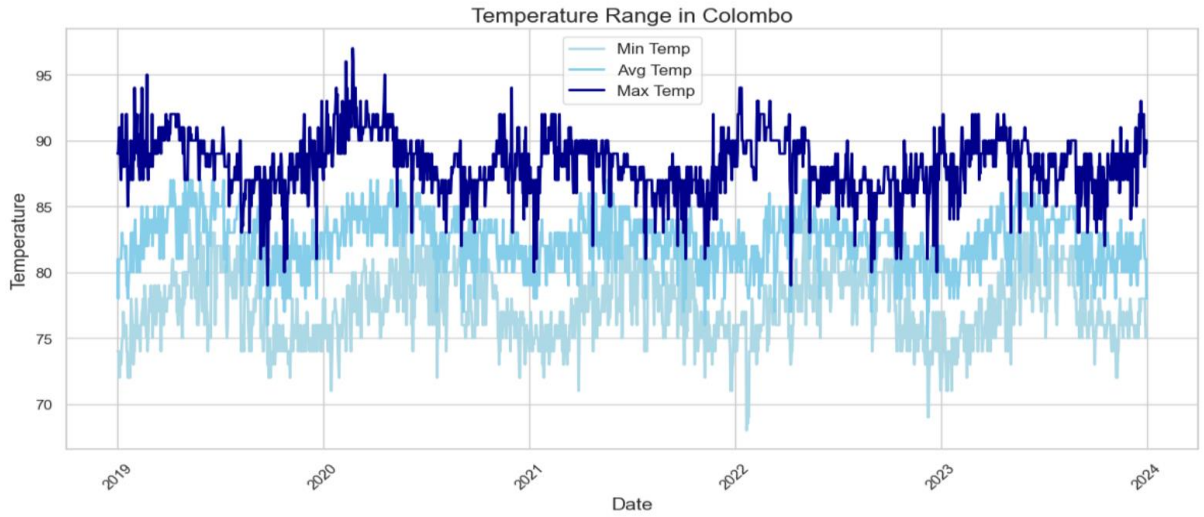
Weather Data Collection and Limitations

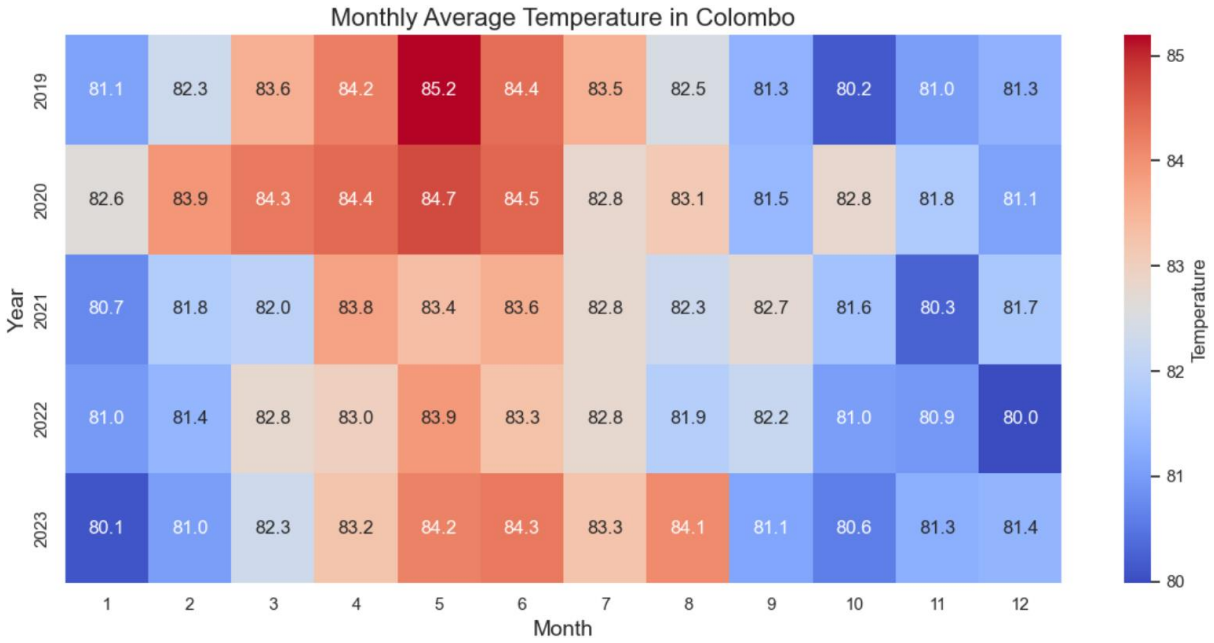
The weather analysis with HCHO levels is done by considering precipitation, average, minimum, and maximum temperatures. For weather data collection in each region, two datasets were mainly used. Colombo Proper, Kurunegala, and Nuwara Eliya regions data were collected from the National Climatic Data Centre's online database (Government, n.d.). However, there were many dates that did not have any records of weather data in the above dataset. Therefore, the missing dates have been added, and the missing record values were handled using rolling, forward filling, and backward filling techniques in time series analysis. For other regions, precipitation and temperature-based data were collected from the NASA Data Access Viewer API, which is an online database that facilitates getting past weather data for any region by providing latitude and longitude data. (NASA, n.d.). The data collected from this dataset did not contain any missing values. However, the temperature values were given on a centigrade scale. The dataset that was used for Colombo, Kurunegala, and Nuwara Eliya consisted of Fahrenheit scale temperature values. Therefore, the temperature data collected from the Nasa dataset was converted to a Fahrenheit scale. These weather data were analysed with HCHO data by joining the regional tables created. The precipitation data was given in two scales, but it was not clear what are the provided scales in both datasets to get them into one scale. Before Analyzing these data by comparing the distribution of HCHO readings, they were analyzed individually. The below plots show the analysed weather data for each region.

Colombo Weather Data

The below plots show how weather data is distributed.



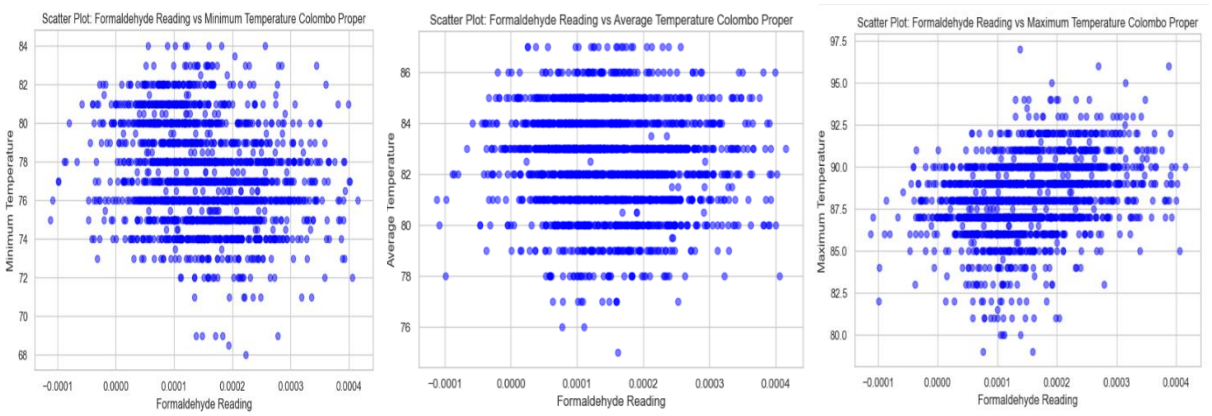




Colombo Weather data Correlation Analysis

When correlations checked with Colombo weather data, there is a slight correlation with maximum and Minimum Temperature in the dataset. However, Precipitation had a very low correlation.

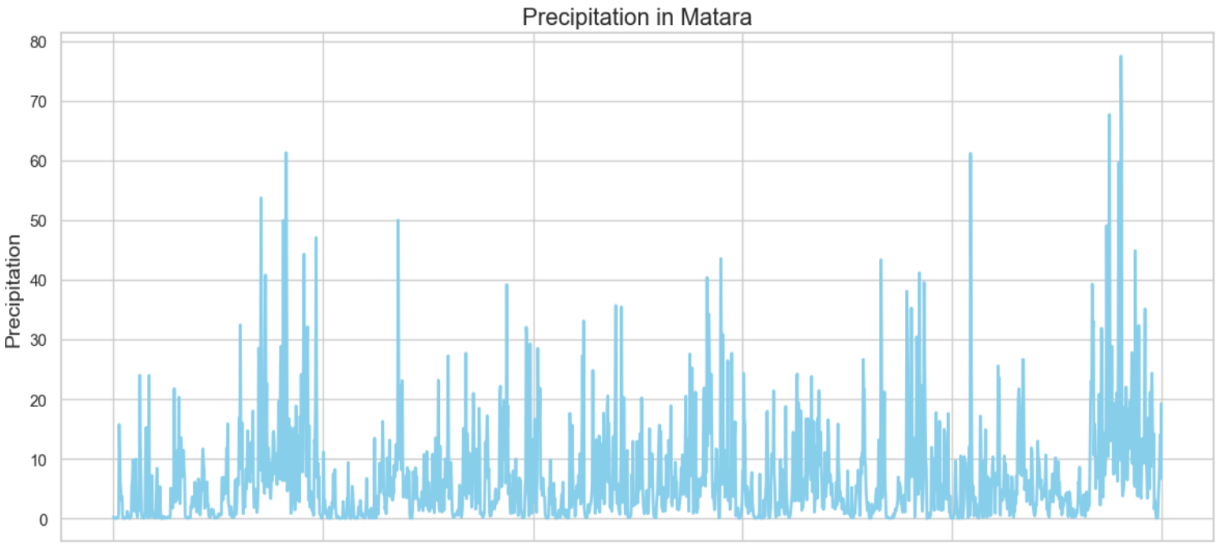
	HCHO reading	PRCP	TAVG	TMAX	TMIN
HCHO reading	1.000000	-0.093883	-0.056770	0.337612	-0.235119
PRCP	-0.093883	1.000000	-0.359612	-0.321459	-0.277047
TAVG	-0.056770	-0.359612	1.000000	0.511461	0.659455
TMAX	0.337612	-0.321459	0.511461	1.000000	0.096884
TMIN	-0.235119	-0.277047	0.659455	0.096884	1.000000

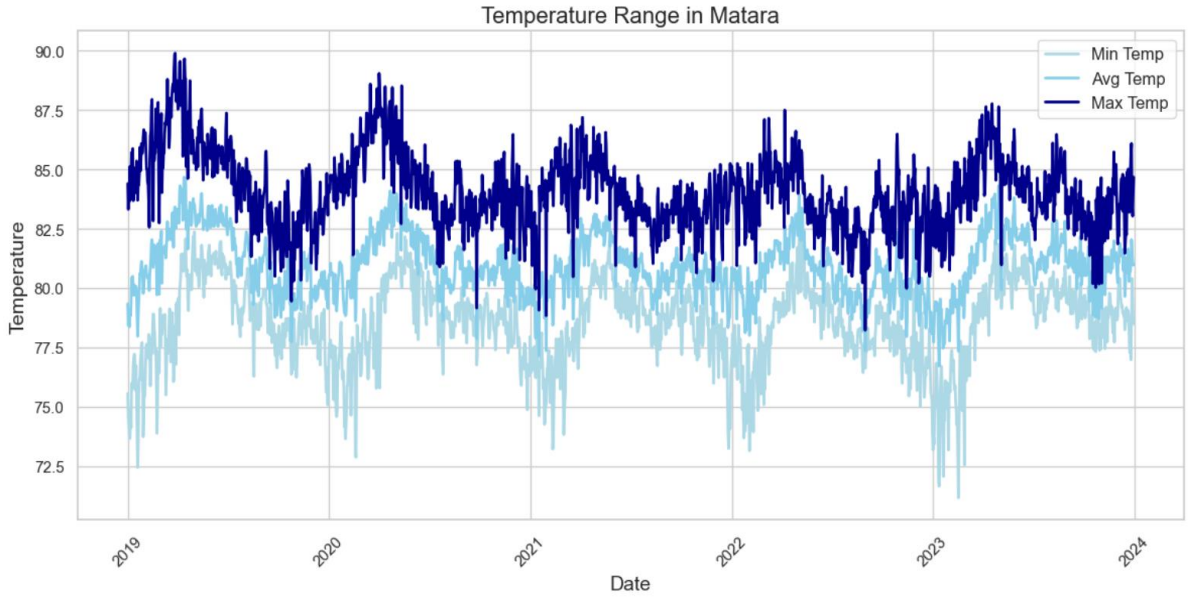


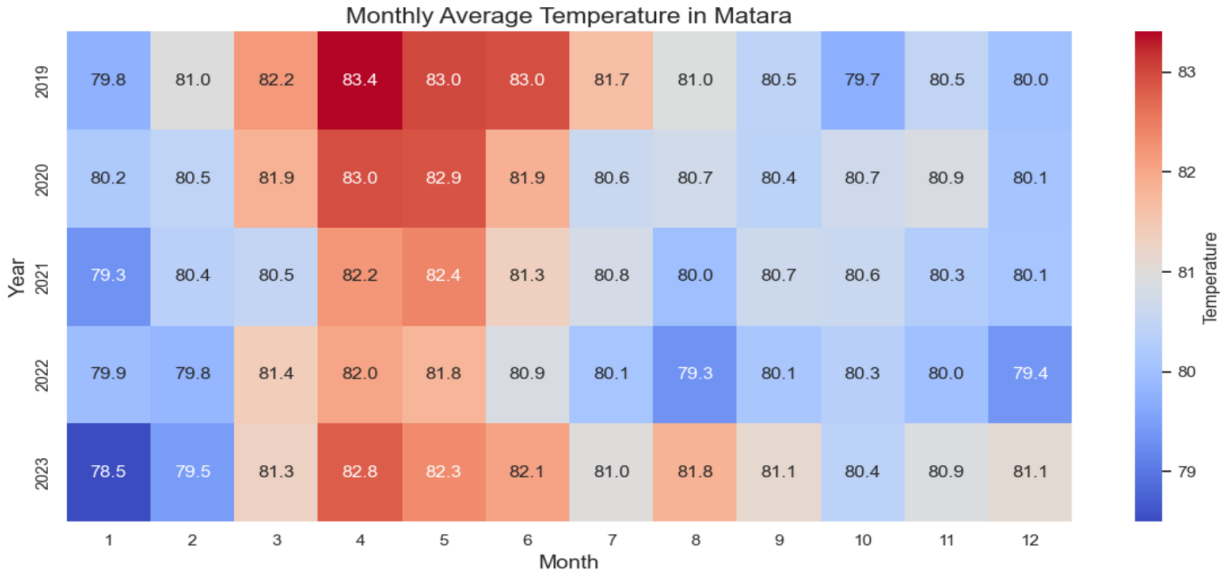
The above plots show how temperature rates are distributed with HCHO readings in Colombo.

Matara DeniyayaWeather Data

The below plots show how Deniyaya Matara weather data is distributed.





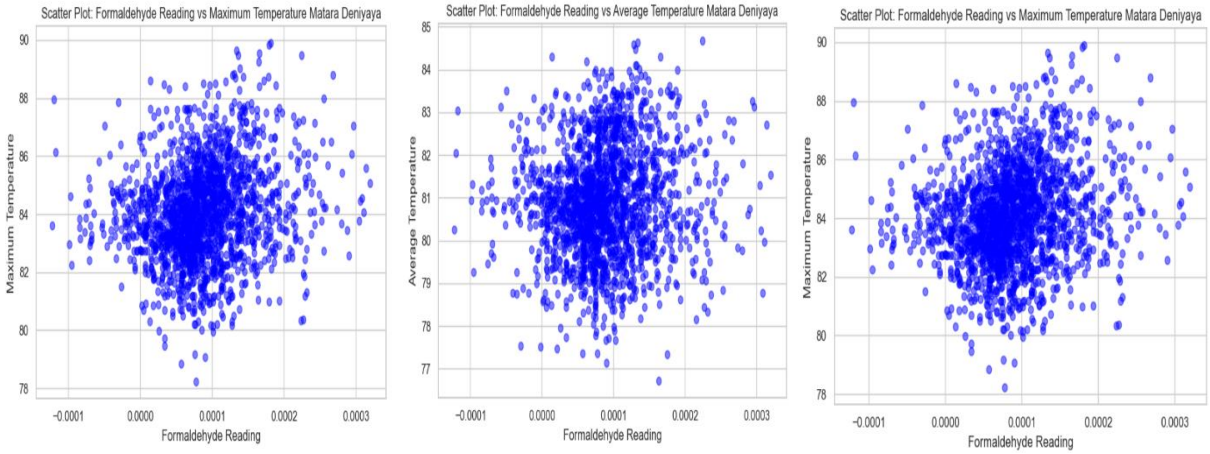


Matara Weather Data Correlation Analysis

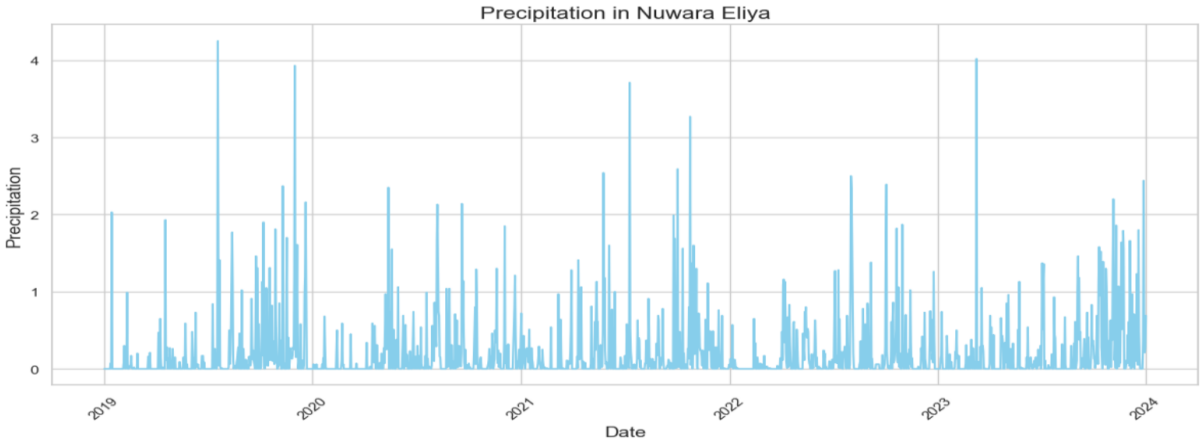
The below table shows there is a very low correlation with both precipitation and temperature levels. Maximum Temperature had the highest correlation with HCHO reading out of them.

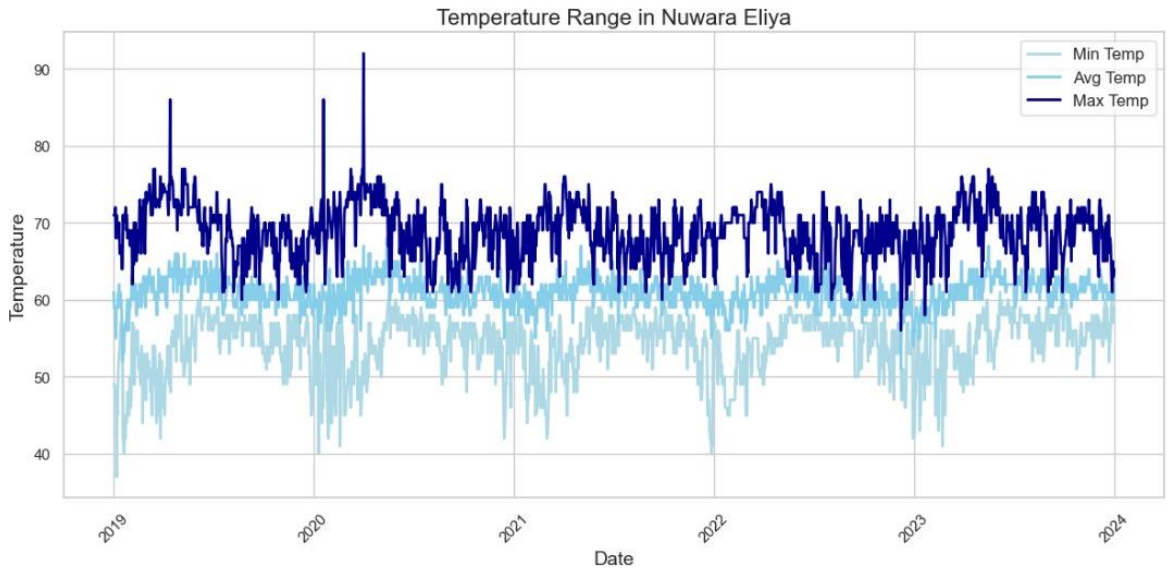
	HCHO reading	PRCP	TAVG	TMAX	TMIN
HCHO reading	1.000000	-0.033231	0.082761	0.177749	-0.063090
PRCP	-0.033231	1.000000	-0.135281	-0.365819	0.109113
TAVG	0.082761	-0.135281	1.000000	0.767972	0.795179
TMAX	0.177749	-0.365819	0.767972	1.000000	0.250005
TMIN	-0.063090	0.109113	0.795179	0.250005	1.000000

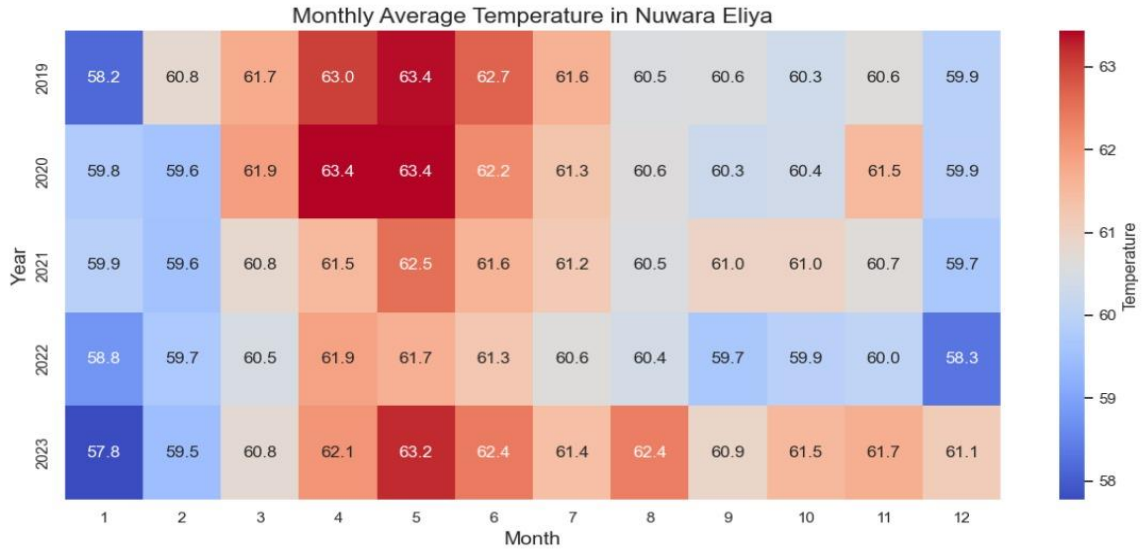
These plots show how temperature-based measurements are correlated with Matara Deniyaya HCHO readings.



Nuwara Eliya Weather Data



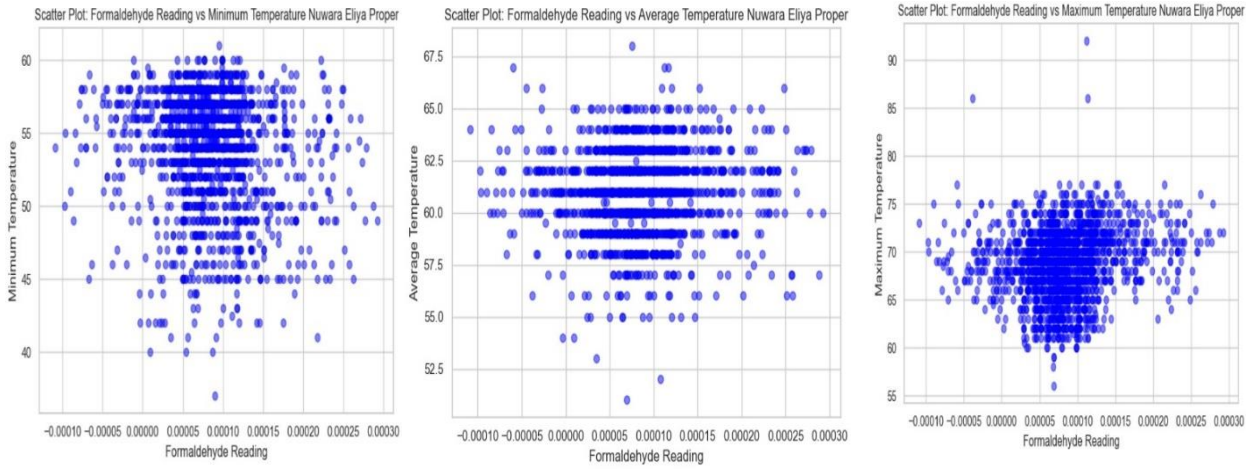




Nuwara Eliya Weather data Correlation Analysis

When the Nuwara Eliya weather data correlations checked, there were no much correlation between HCHO readings and weather data.However, the Maximum had the best correlation out of them with a pearson correlation of 0.15.

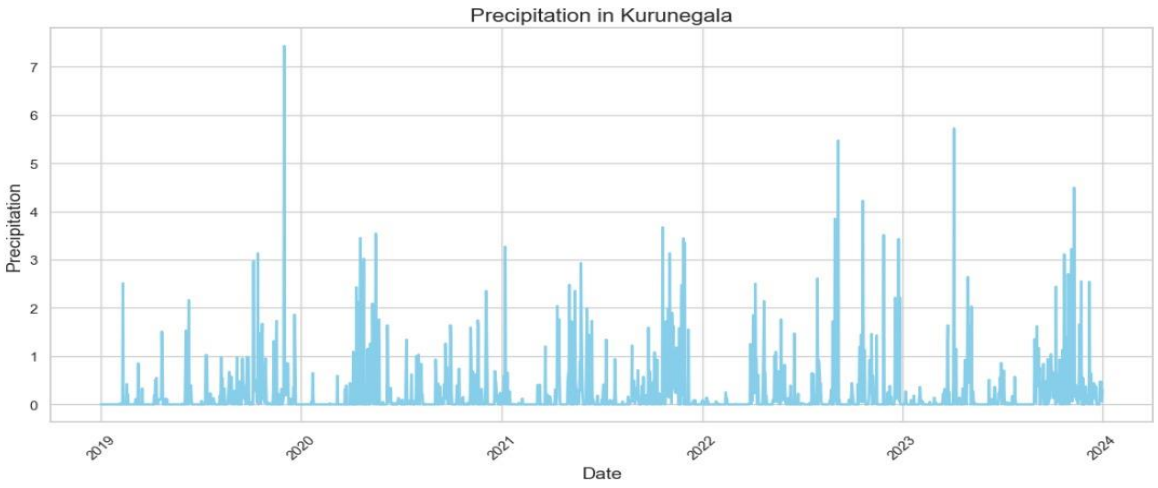
	HCHO reading	PRCP	TAVG	TMAX	TMIN
HCHO reading	1.000000	-0.025234	0.029537	0.153014	-0.139301
PRCP	-0.025234	1.000000	-0.086181	-0.273322	0.201576
TAVG	0.029537	-0.086181	1.000000	0.522359	0.326624
TMAX	0.153014	-0.273322	0.522359	1.000000	-0.363431
TMIN	-0.139301	0.201576	0.326624	-0.363431	1.000000

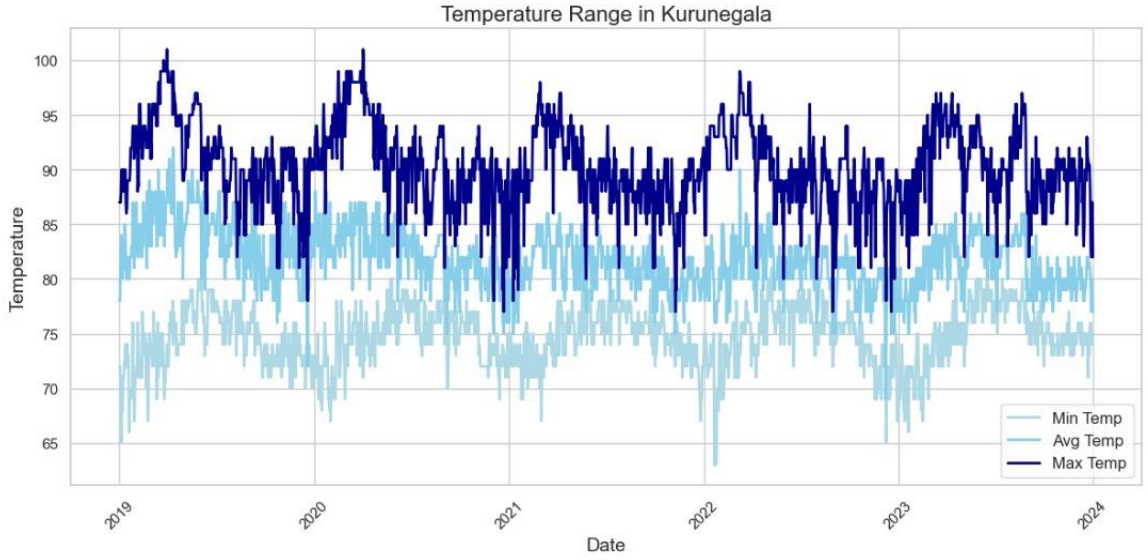


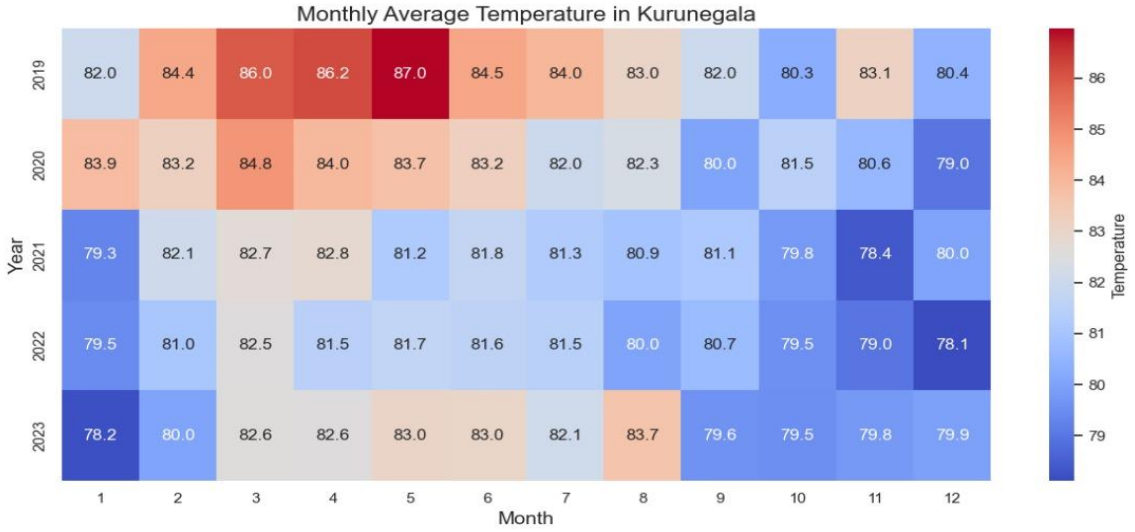
The above plots show how Temperature data is correlated with HCHO readings in Nuwara Eliya region.

Kurunegala Weather Data Analysis

The below plots show how Kurunegala weather data is distributed.





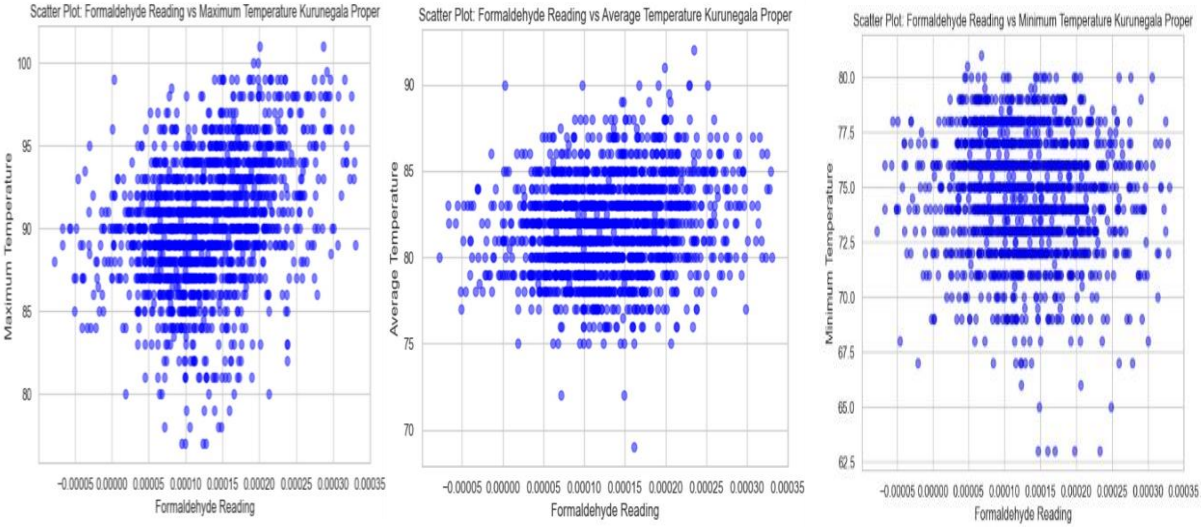


Kurunegala Weather Data Correlation Analysis

The below table shows there is a slight correlation between the HCHO readings and the maximum temperature in the Kurunegala Region. However, other precipitation and temperature-based calculations did not have much correlation with Kurunegala data.

	HCHO reading	PRCP	TAVG	TMAX	TMIN
HCHO reading	1.000000	-0.020175	0.174568	0.342388	-0.095334
PRCP	-0.020175	1.000000	-0.271417	-0.242136	-0.111818
TAVG	0.174568	-0.271417	1.000000	0.726114	0.330163
TMAX	0.342388	-0.242136	0.726114	1.000000	0.039618
TMIN	-0.095334	-0.111818	0.330163	0.039618	1.000000

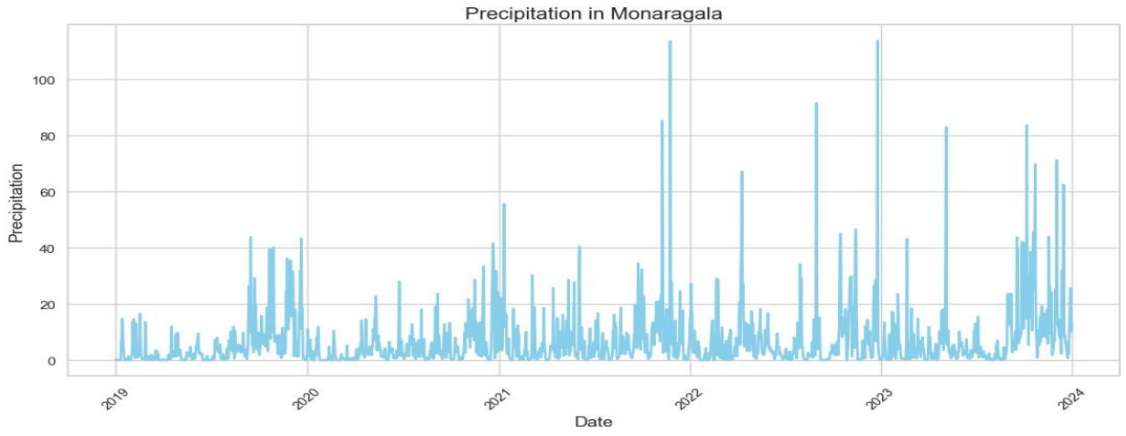
These plots show how Kurunegala weather data is distributed with HCHO readings.

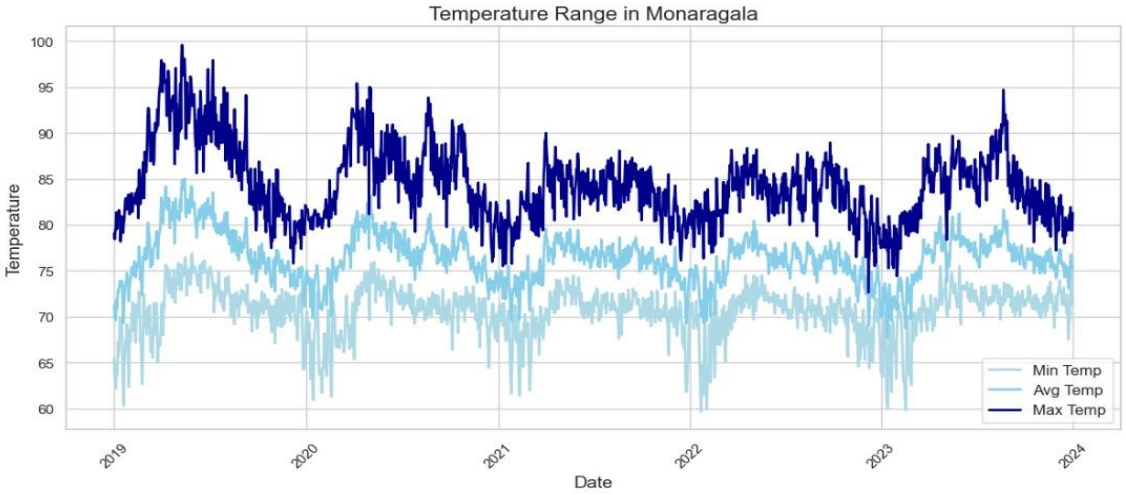


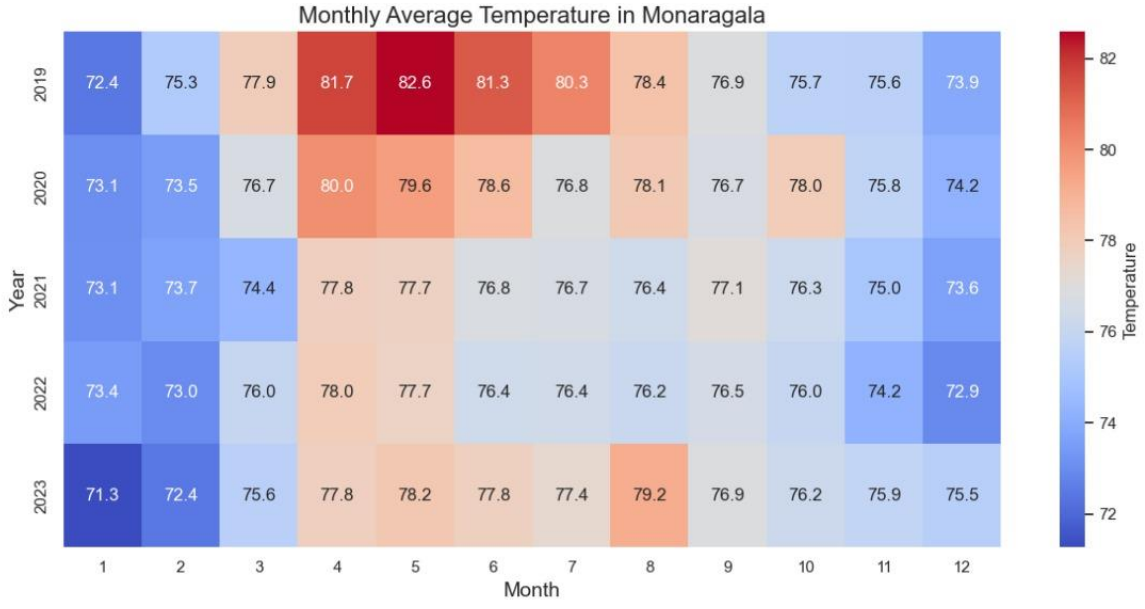
-

Bibile Monaragala Weather Data Analysis

The below plots show how Monaragala weather data is distributed.





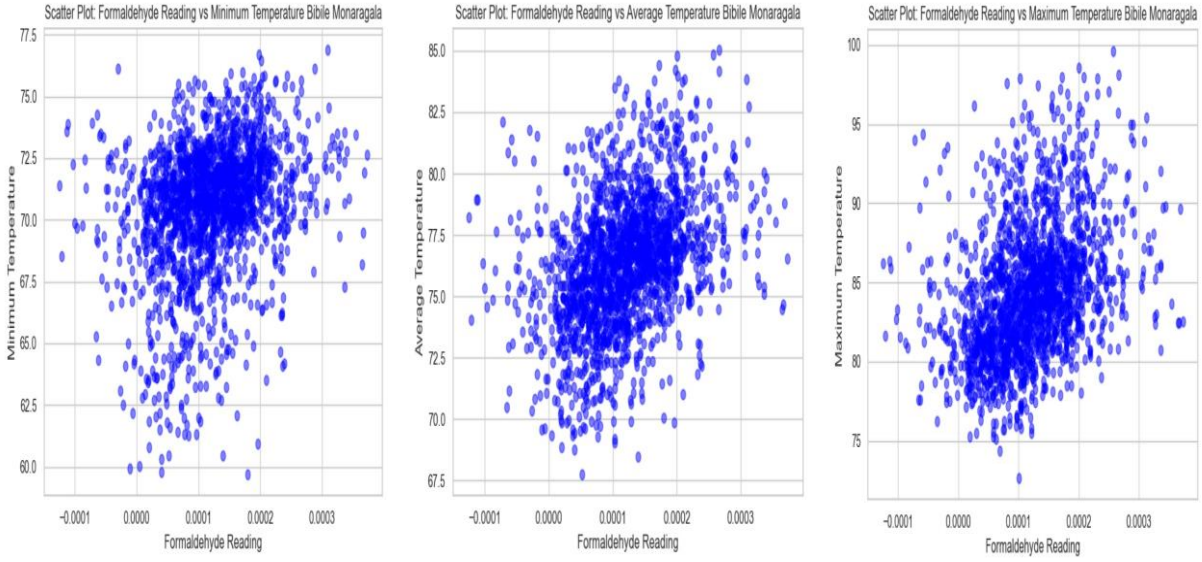


Monaragala Weather Data Correlation Analysis

Compared to other regions Monaragala had better correlations for All temperature related calculations with Formaldehyde distribution. However, precipitation does not correlate with the HCHO emissions.

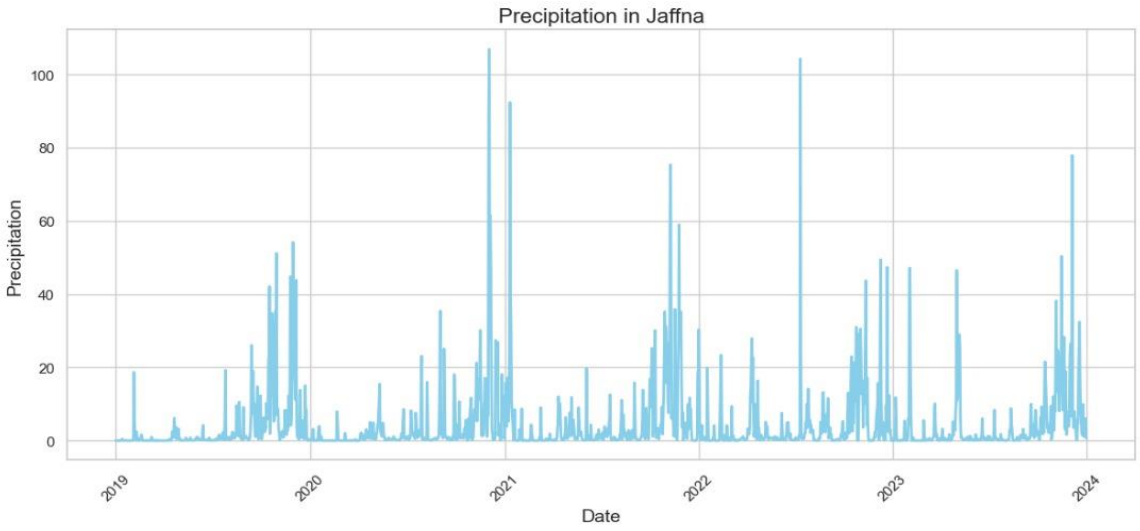
	HCHO reading	PRCP	TAVG	TMAX	TMIN
HCHO reading	1.000000	-0.045412	0.374357	0.357828	0.275180
PRCP	-0.045412	1.000000	-0.107879	-0.306182	0.176686
TAVG	0.374357	-0.107879	1.000000	0.890991	0.786938
TMAX	0.357828	-0.306182	0.890991	1.000000	0.451086
TMIN	0.275180	0.176686	0.786938	0.451086	1.000000

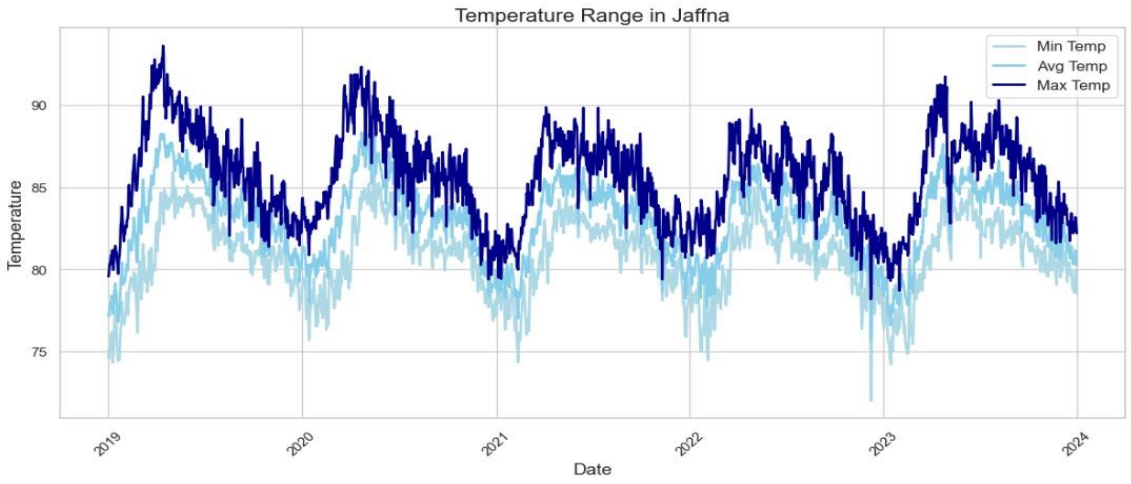
The below plots show how weather temperature-based data is correlated with HCHO emissions.

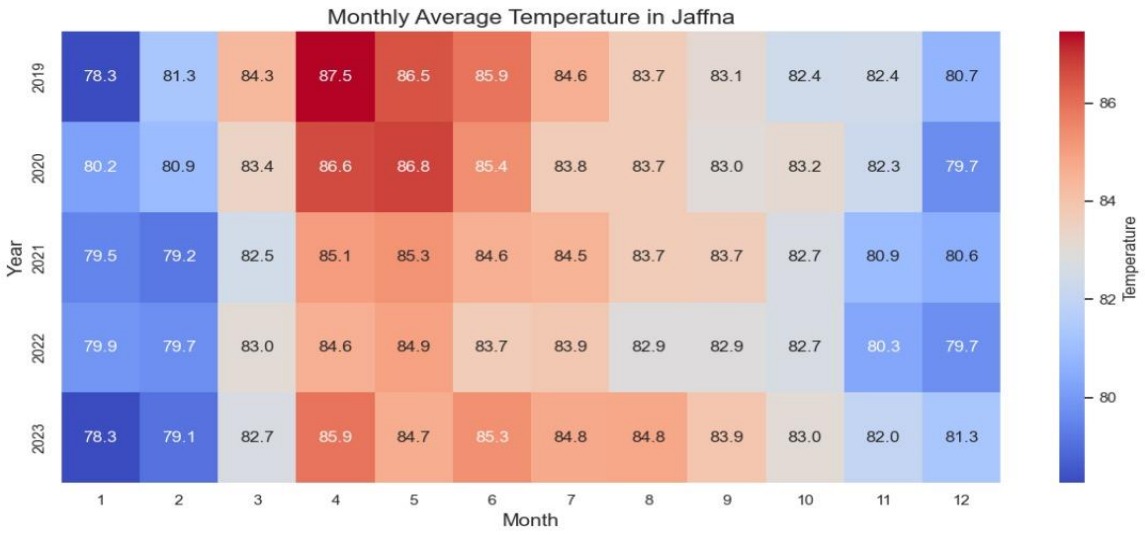


Jaffna Weather Data Analysis

The below plots show how Monaragala weather data is distributed.





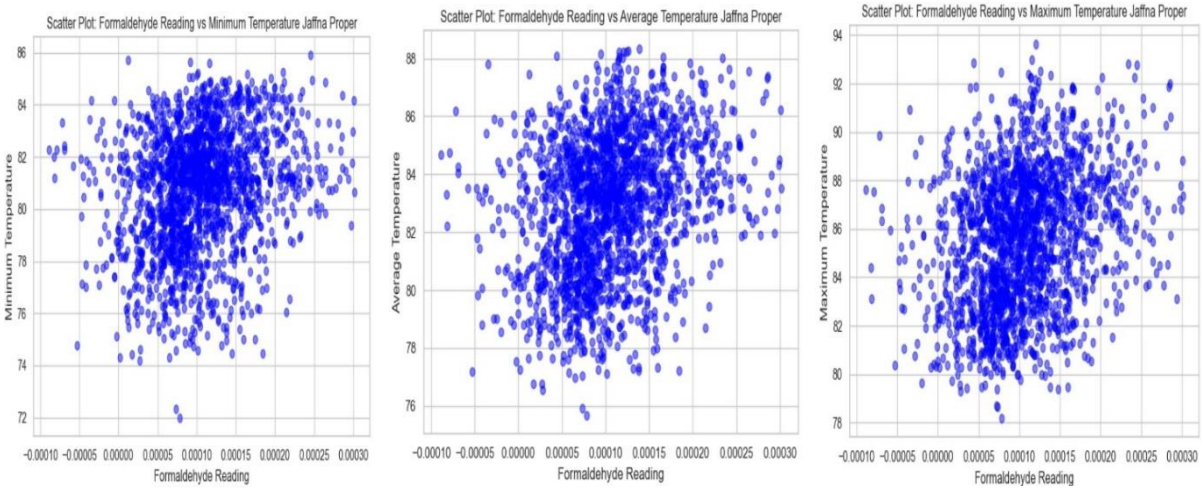


Jaffna Weather Data Correlation Analysis

Jaffna Weather data shows a slight correlation between all temperature-based measurements with the HCHO reading. However, as in other cities there is a very low correlation for precipitation.

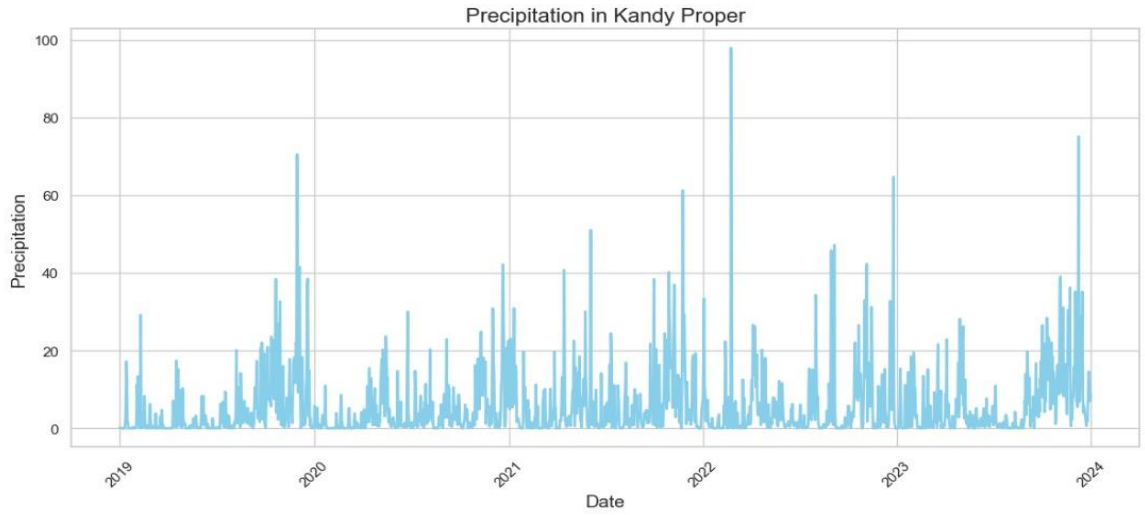
	HCHO reading	PRCP	TAVG	TMAX	TMIN
HCHO reading	1.000000	-0.030000	0.273649	0.277234	0.231368
PRCP	-0.030000	1.000000	-0.213309	-0.318674	-0.095004
TAVG	0.273649	-0.213309	1.000000	0.947952	0.934553
TMAX	0.277234	-0.318674	0.947952	1.000000	0.785812
TMIN	0.231368	-0.095004	0.934553	0.785812	1.000000

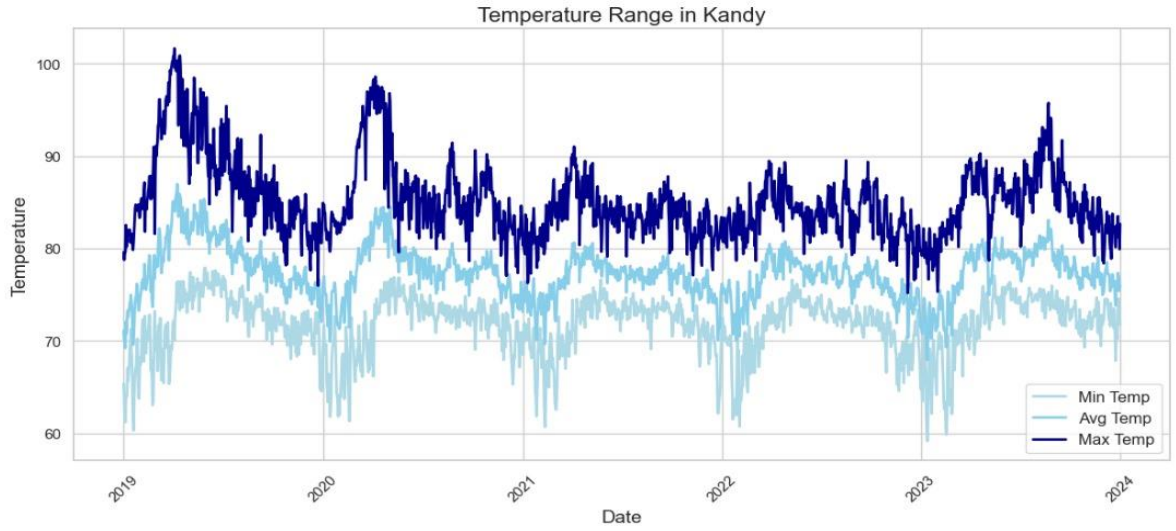
The below plot shows how maximum, average, and minimum temperatures correlated with HCHO readings in Jaffna.

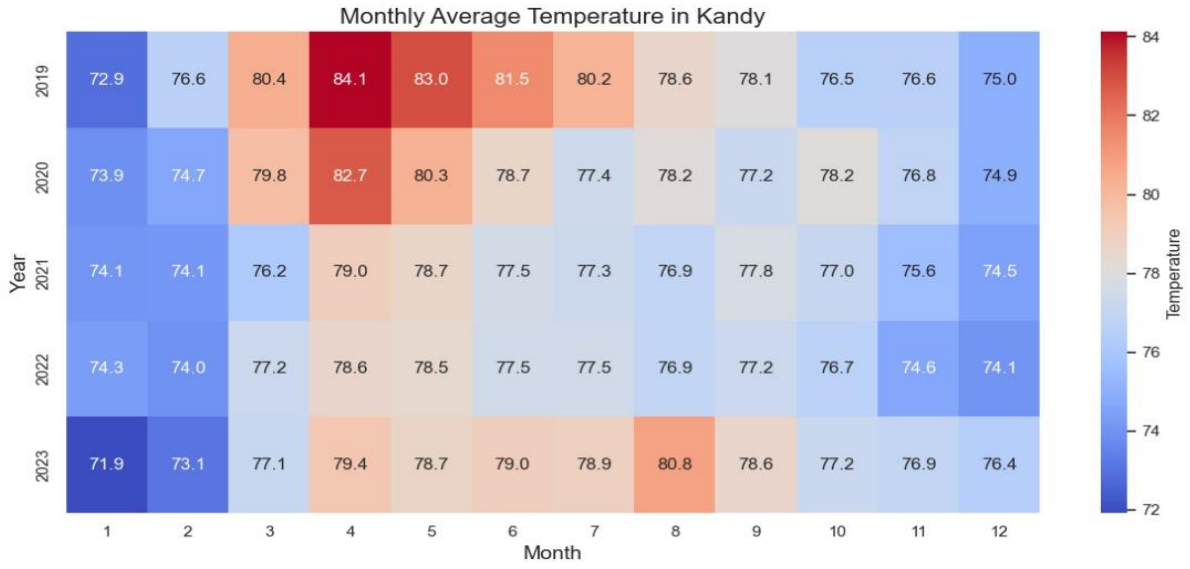


Kandy Weather Data Analysis

The below plots show how Kandy weather data is distributed.





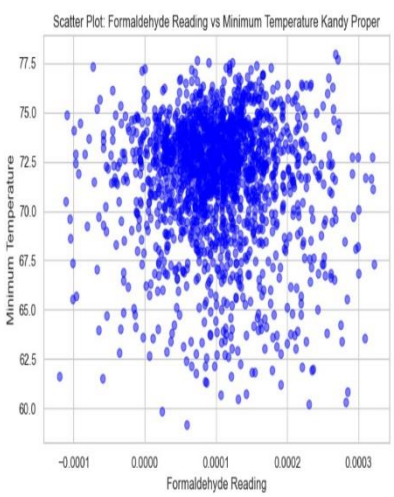


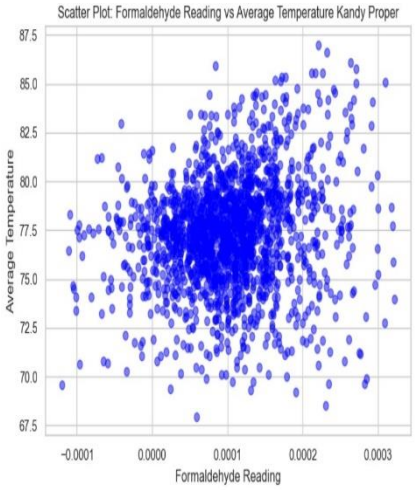
Kandy Weather Data Correlation Analysis

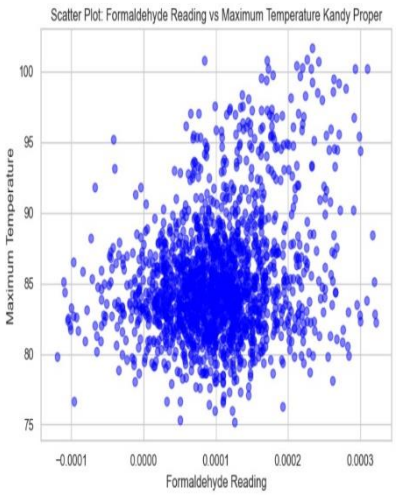
There is an average correlation between HCHO reading and maximum temperature in Kandy. However, other weather-related calculations do not correlate much with the HCHO reading.

	HCHO reading	PRCP	TAVG	TMAX	TMIN
HCHO reading	1.000000	-0.023929	0.162298	0.248444	-0.044744
PRCP	-0.023929	1.000000	-0.117315	-0.296111	0.123072
TAVG	0.162298	-0.117315	1.000000	0.848373	0.745154
TMAX	0.248444	-0.296111	0.848373	1.000000	0.304645
TMIN	-0.044744	0.123072	0.745154	0.304645	1.000000

The below scatter plots show how temperature-based values correlated with HCHO readings.







Conclusion on overall weather analysis with HCHO emissions

As mentioned in the above tables, it shows that there is a very low correlation between HCHO emissions and precipitation in all the regions. However, there were some regions that had a slight correlation up to 0.3 with temperature-related data. It shows that temperature can be taken as an estimator that helps to predict HCHO rates.

Spatial Data Collection and Limitations

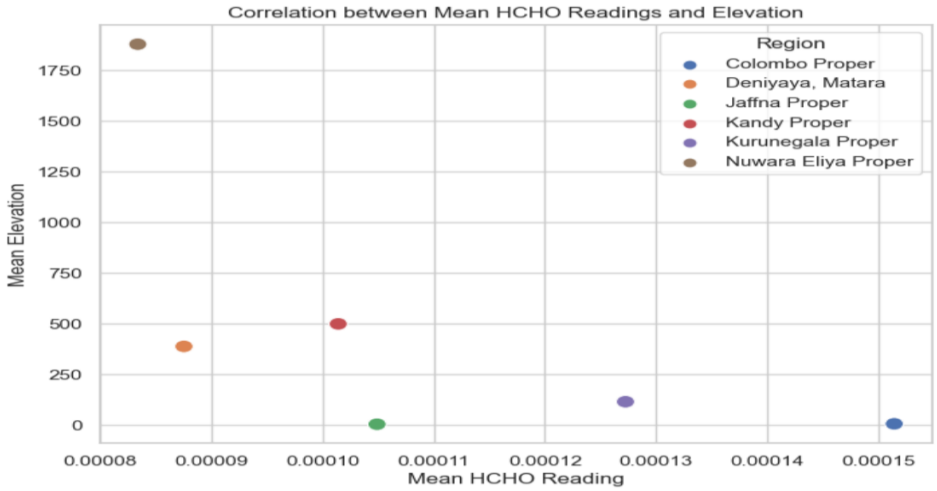
The latitude and longitude data were gathered from the same source for several cities as weather data: the National Climatic Data Centre's online database. In addition, other regions latitude data were collected separately from the Mind Data Organization (Anon., 2024)and using Google Search. Since satellite cites are different compared to the actual cities that are considered, there might be some slight changes in the collected spatial-related data. The city population and area-related data are gathered from several sources for each city, namely, worldpopulationreview.com (Anon., n.d.), some divisional secretariat websites (Kotapola, n.d.)and Google searches. However, as mentioned on the websites, there is no proper update of population-based data from 2012 in Sri Lankan cities. Therefore, there might be slight changes in population density from 2019 to 2023. The city-wise area-related data was gathered from Wikipedia and Google search. Since Wikipedia is an open-source website, there might be some issues with the reliability of the gathered data. For elevation-based correlation analysis, data were mainly gathered from Google searches. However, there was a lack of sources to get data on the proximity levels of each city from the sea. The FreeMapTools online application was used to get the proximity levels of each city from the sea (FreeMapTools, n.d.). These measurements were done using miles; therefore, they were converted to kilometers using google converters. Before doing the analysis, all the spatial pattern related data were combined with each district wise HCHO data. The below table shows an example of a data combined table.

	Current Date	Next Date	HCHO reading	Region	LATITUDE	LONGITUDE	ELEVATION	PRCP	TAVG	TMAX	TMIN	Population	Area_sq_km	population_density
0	2019-01-01	2019-01-02	0.000176	Kandy Proper	7.29	80.63	500	0.09	71.096	79.592	65.354	111701	28.53	3915.21
1	2019-01-02	2019-01-03	0.000092	Kandy Proper	7.29	80.63	500	0.01	70.124	78.764	63.716	111701	28.53	3915.21
2	2019-01-03	2019-01-04	0.000134	Kandy Proper	7.29	80.63	500	0.02	69.728	79.448	62.240	111701	28.53	3915.21
3	2019-01-04	2019-01-05	0.000191	Kandy Proper	7.29	80.63	500	0.00	69.224	79.610	61.214	111701	28.53	3915.21
4	2019-01-05	2019-01-06	0.000122	Kandy Proper	7.29	80.63	500	0.04	72.068	82.526	64.040	111701	28.53	3915.21

Spatial Data Analysis with HCHO Emission Levels

Correlation between Mean HCHO readings and elevation

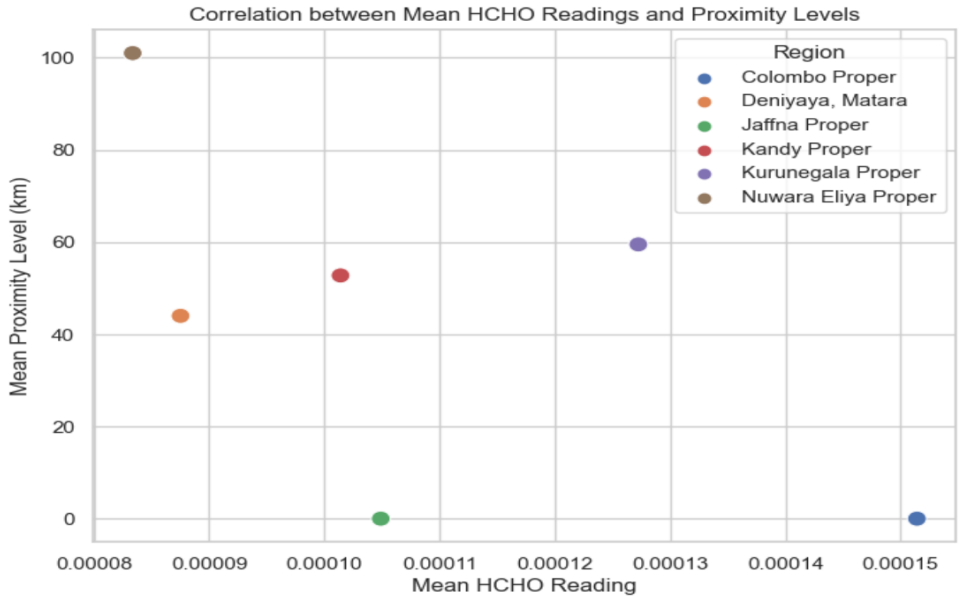
This shows there is a strong correlation between mean HCHO readings and elevation with a Pearson Correlation of -0.63. It shows there is a possibility of decreasing mean HCHO levels when height is increased.




```
Correlation between mean HCHO readings and mean Elevation: -0.6390074936539603
```

Correlation between Mean HCHO readings and Proximity

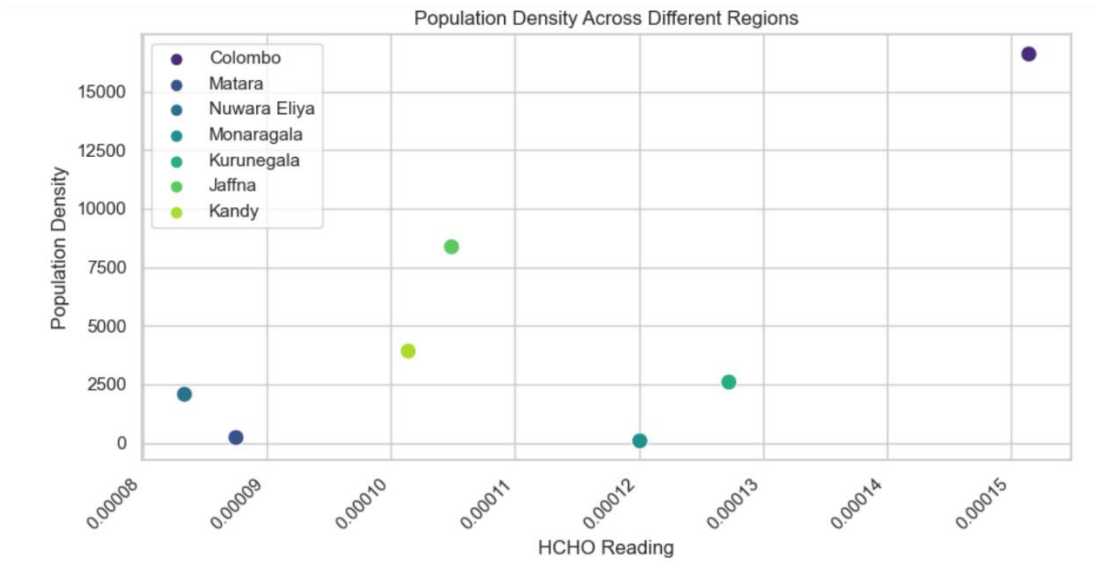
AS in the elevation there is a strong correlation of Proximity and Mean HCHO emissions in each city with a Pearson correlation of -0.58. It shows that when Proximity increases, there is a possibility of decreasing HCHO emission rates.




```
Correlation between mean HCHO readings and mean proximity levels: -0.5891985166833652
```

Correlation between Mean HCHO readings and Population Density

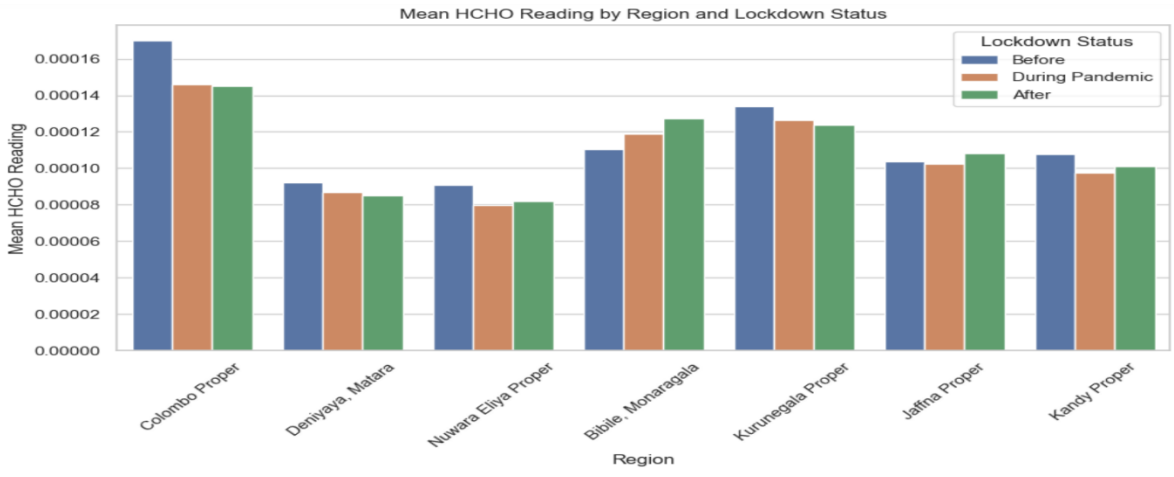
There is a clear correlation between the mean population and the mean HCHO readings of each region with a Pearson correlation of 0.6725. It shows that there is a possibility of increasing the mean HCHO reading when Population density increases.



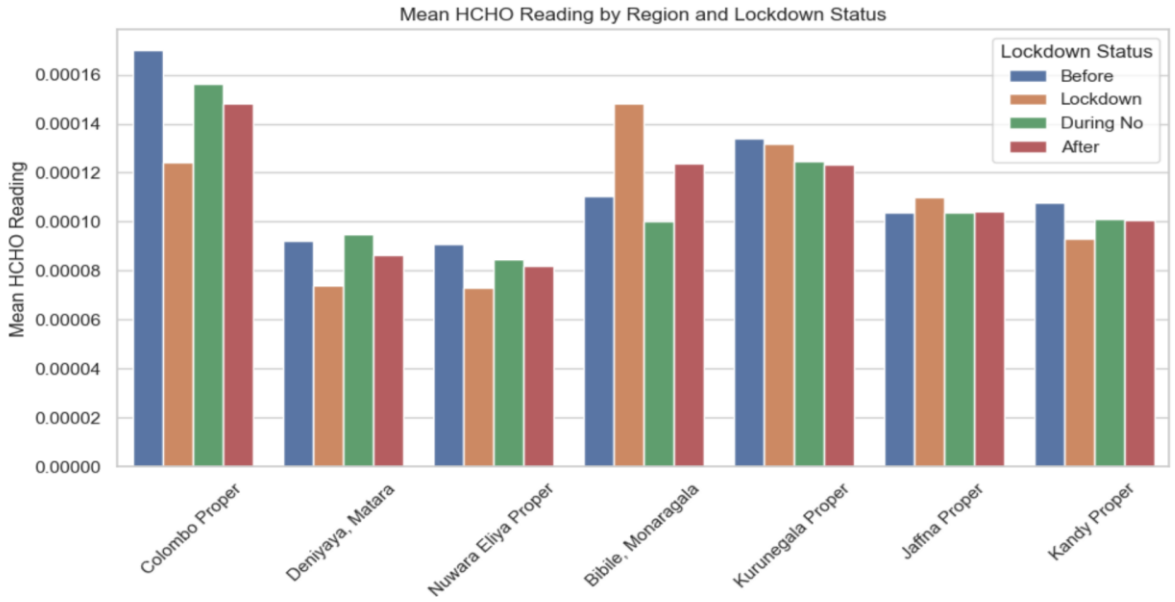

```
Correlation between mean HCHO readings and mean proximity levels: 0.672551940206931
```

Covid Data Analysis with HCHO Emissions

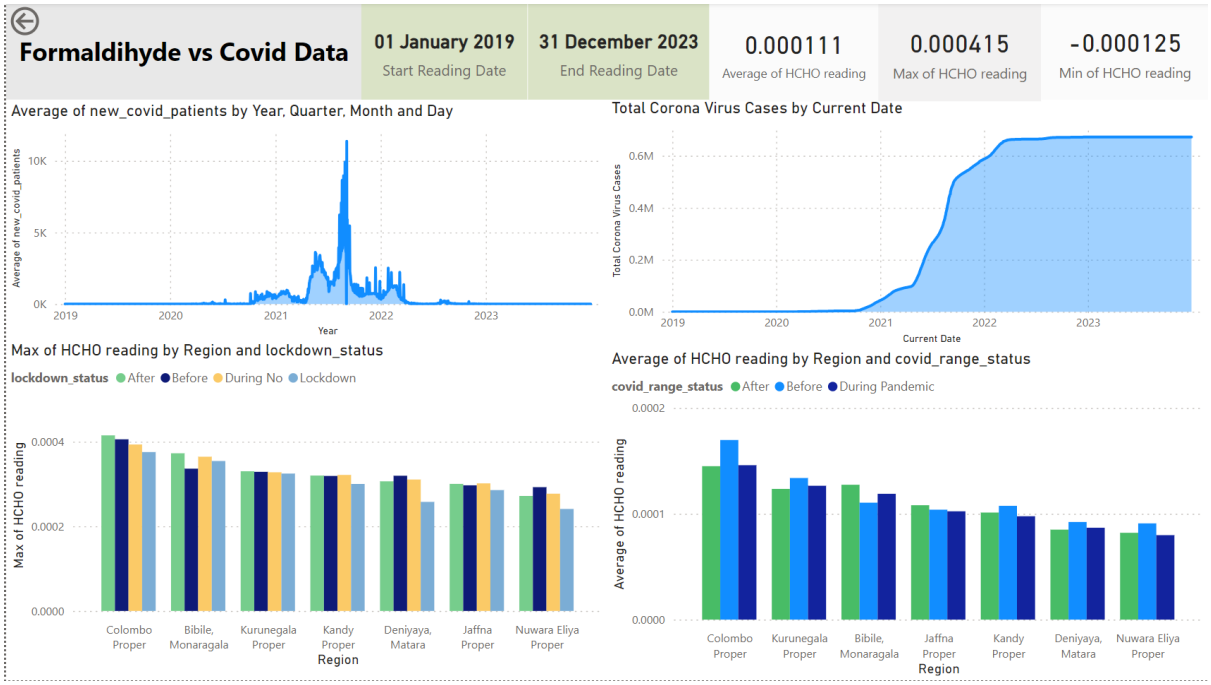
The COVID data analysis was done based on the HCHO emission rates in lockdown periods and the number of COVID patients found. The first two COVID lockdown periods data were collected from Wikipedia, and other lockdown periods data were collected from the A3M global monitoring website (Anon., n.d.). There were some periods when lockdown was imposed only in several areas. Therefore, it was difficult to find the exact days that lockdown was imposed for all the analysed regions. The newly reported COVID patient numbers and related data were gathered from covid 19 Ninjas API for powerbi analysis. Finally, the covid data were combined with the HCHO emission tables.



For the above plot, the time period between 2020-03-18 and 2022-03-01 was considered as pandemic periods. It clearly depicts in majority of cities instead of Bibile Monaragala, has a reduction of Mean HCHO emission rates in pandemic period. In addition, in majority of cities, it has started increasing the mean HCHO emission after the pandemic.



For the above plot, the exact covid lockdown dates were used. It depicts there is a decrement of HCHO readings during lockdown periods except Bibile and Jaffna Cities.

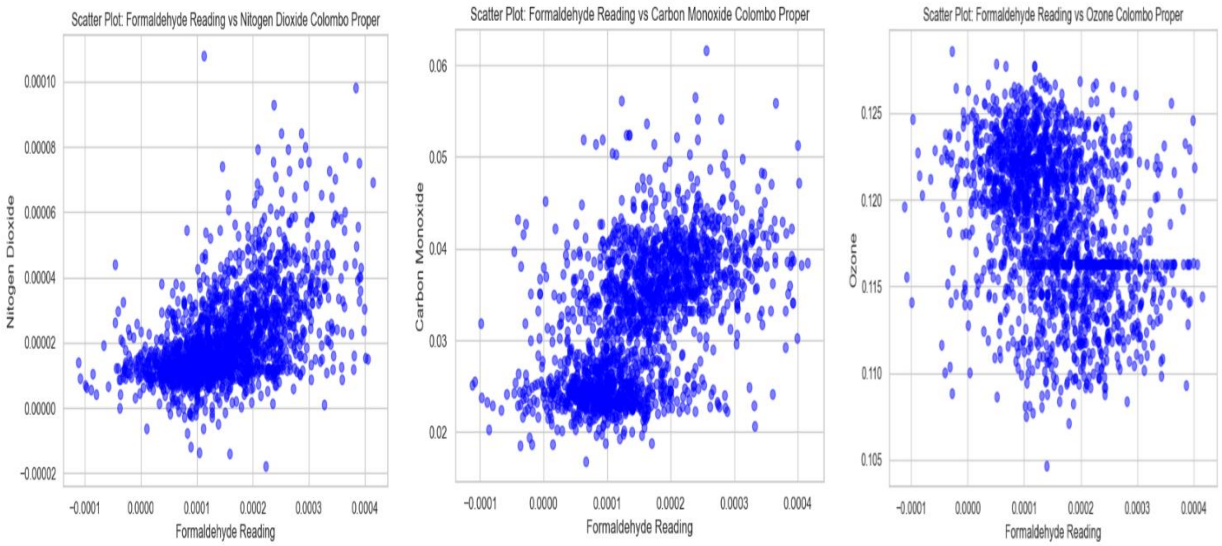


Anthropogenic and Industrial Activity Impact on HCHO Emissions

The anthropogenic impact on HCHO emissions was analysed using the emission rates of carbon monoxide, nitrogen dioxide, and ozone emissions. Carbon monoxide is selected to analyse the industrial impact because it is primarily emitted from the incomplete combustion of fossil fuels, biomass burning, and industrial processes. Nitrogen dioxide is selected because it is mainly emitted from combustion processes such as vehicle engines, power plants, and industrial facilities, and it can participate in atmospheric reactions that lead to the formation of HCHO, particularly under sunlight. However, the ozone gas emission selected due to it is considered as a primary measurement on Air quality. These gas emission data were collected from Swagger Emissions API by providing latitude and longitude data (Developers, n.d.).However, there were missing values in the collected datasets. They were handled using rolling techniques in time series analysis. These collected data were analysed after combining them with HCHO initial datasets.

Colombo HCHO and Other Gas Emission Analysis

The below plots show how HCHO readings are correlated with CO2, NO2 and Ozone emission rates.

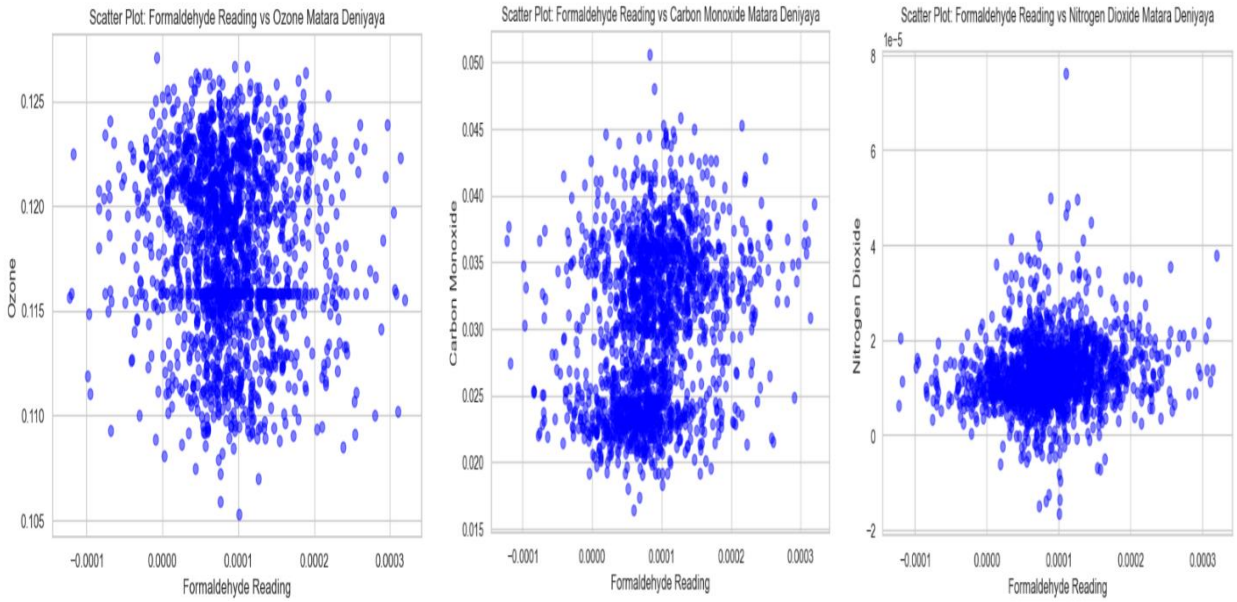


The table below indicates a high association between HCHO emission rates and NO2, CO2 emission rates, with correlation coefficients of 0.54 and 0.51, respectively. In addition, there is a 0.35 correlation between Ozone emissions.

	carbon_monoxide	nitogen_dioxide	ozone	HCHO reading
carbon_monoxide	1.000000	0.590862	-0.601463	0.548224
nitogen_dioxide	0.590862	1.000000	-0.459146	0.516773
ozone	-0.601463	-0.459146	1.000000	-0.353562
HCHO reading	0.548224	0.516773	-0.353562	1.000000

Deniyaya Matara HCHO and Other Gas Emission Analysis

The plots below show how HCHO values relate to other emission rates in Deniyaya.

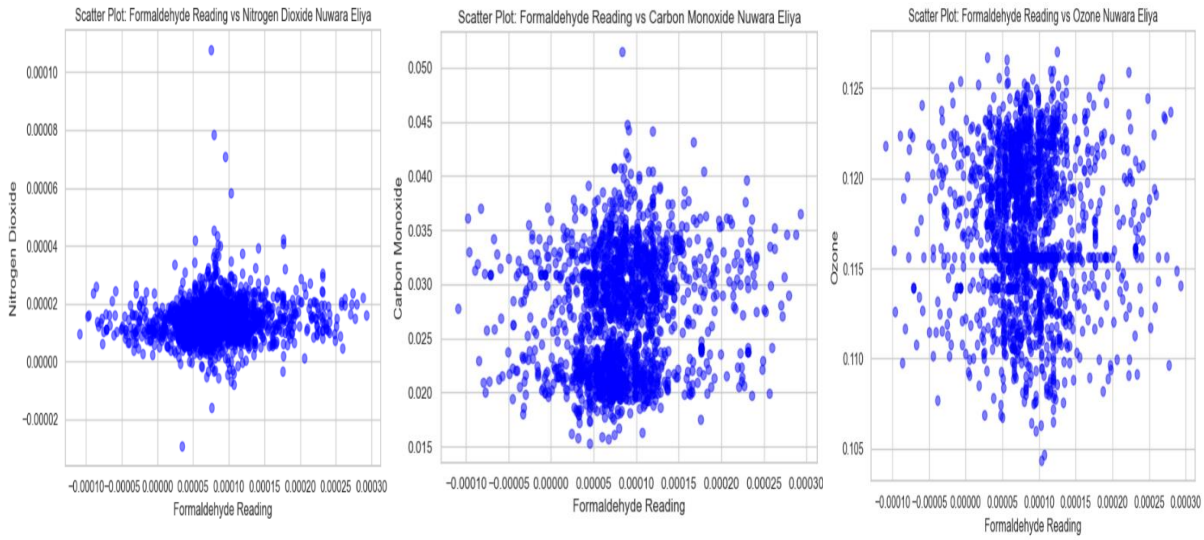


As not in Colombo, the below table shows there is not much correlation between HCHO reading and other gas emissions. Carbon monoxide emission had the highest correlation with 0.248.

	carbon_monoxide	nitogen_dioxide	ozone	HCHO reading
carbon_monoxide	1.000000	0.298125	-0.601841	0.248222
nitogen_dioxide	0.298125	1.000000	-0.242843	0.155423
ozone	-0.601841	-0.242843	1.000000	-0.089257
HCHO reading	0.248222	0.155423	-0.089257	1.000000

Nuwara Eliya HCHO and Other Gas Emission Analysis

These visualizations show how HCHO readings are correlated with CO2, NO2 and Ozone emission rates in Nuwara Eliya.

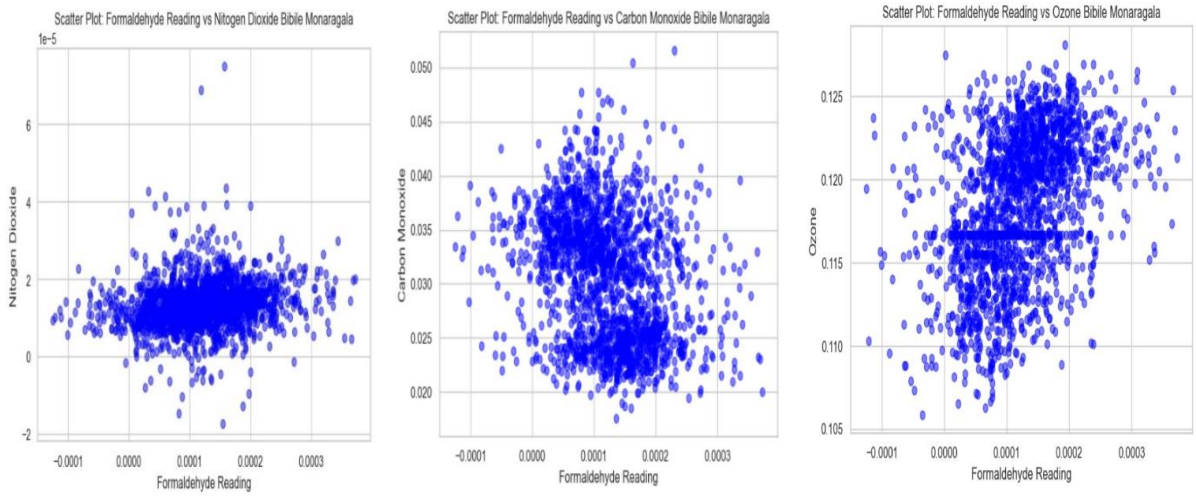


The below table shows that there is very low correlation between HCHO rates and other gas emissions. The highest correlation is recorded for Carbon Monoxide with a Pearson Correlation of 0.1429.

	carbon_monoxide	nitogen_dioxide	ozone	HCHO reading
carbon_monoxide	1.000000	0.300874	-0.582331	0.142956
nitogen_dioxide	0.300874	1.000000	-0.188048	0.099964
ozone	-0.582331	-0.188048	1.000000	-0.030131
HCHO reading	0.142956	0.099964	-0.030131	1.000000

Bibile Monaragala HCHO and Other Gas Emission Analysis

The below visualizations show how Bibile Monaragala HCHO readings are correlated with other gas emissions.

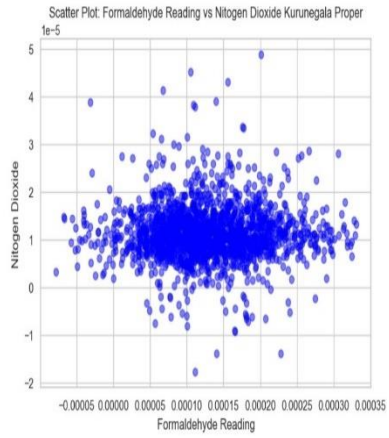


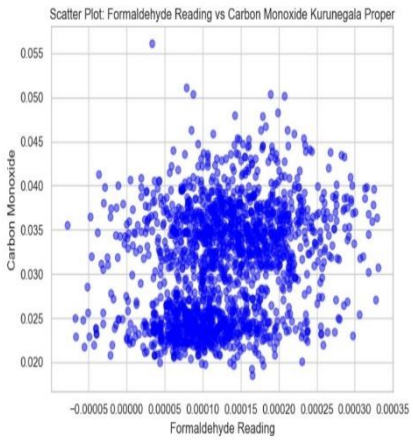
The below table shows comparatively better correlations for carbon monoxide and ozone with HCHO readings, with a Pearson correlation of -0.3 and 0.41, respectively.

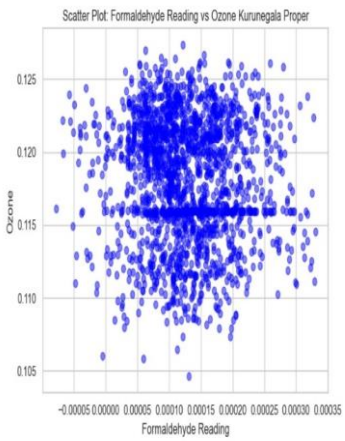
	carbon_monoxide	nitogen_dioxide	ozone	HCHO reading
carbon_monoxide	1.000000	-0.089995	-0.581060	-0.301561
nitogen_dioxide	-0.089995	1.000000	0.171249	0.171317
ozone	-0.581060	0.171249	1.000000	0.414593
HCHO reading	-0.301561	0.171317	0.414593	1.000000

Kurunegala HCHO and Other Gas Emission Analysis

The graphs below illustrate how Bibile Monaragala HCHO values are related to CO2, NO2 and Ozone gas emissions.





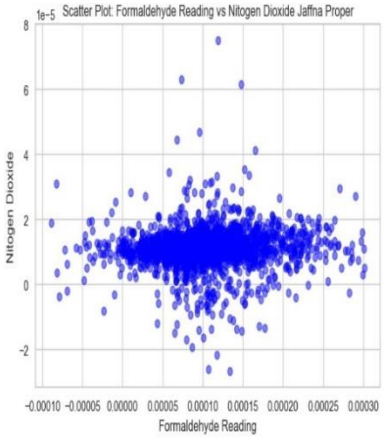


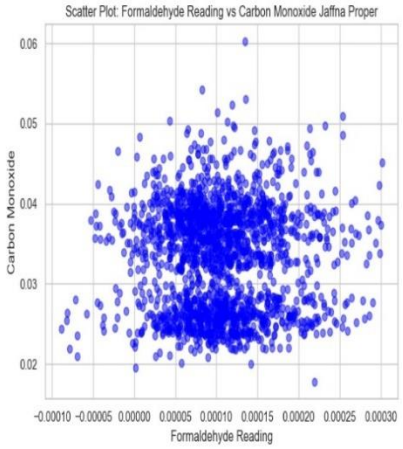
The below table shows there is a slight correlation for Carbon Monoxide rate and HCHO emissions in Kurunegala district with a correlation of 0.27. Other two gases are having a very low correlation.

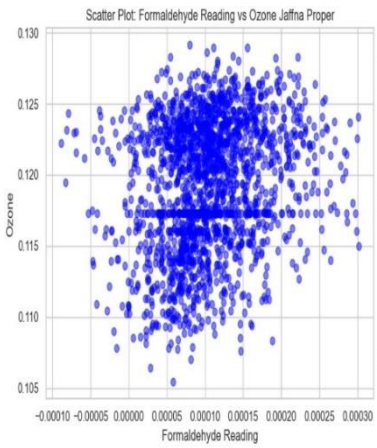
	carbon_monoxide	nitogen_dioxide	ozone	HCHO reading
carbon_monoxide	1.000000	-0.099692	-0.605354	0.276864
nitogen_dioxide	-0.099692	1.000000	0.154155	0.007006
ozone	-0.605354	0.154155	1.000000	-0.075194
HCHO reading	0.276864	0.007006	-0.075194	1.000000

Jaffna HCHO and Other Gas Emission Analysis

The below visualizations show how Jaffna HCHO readings are correlated with CO2, NO2 and Ozone gas emissions.





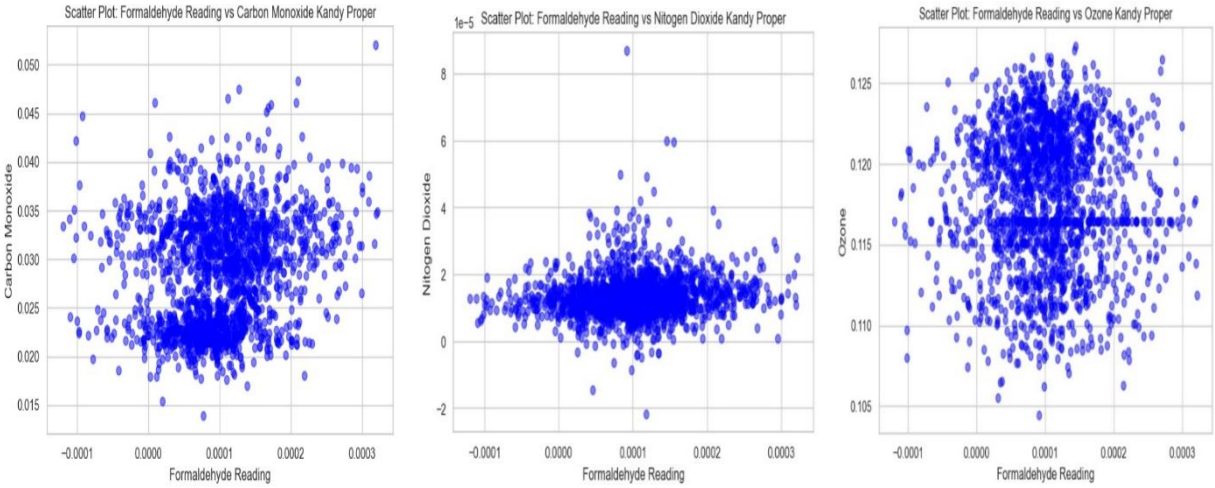


Compared to other cities, Jaffna is having a very low correlation with HCHO emissions rates for the above-mentioned gases. Ozone has the maximum correlation with HCHO emissions with a Pearson Correlation of 0.204.

	carbon_monoxide	nitogen_dioxide	ozone	HCHO reading
carbon_monoxide	1.000000	0.154001	-0.560436	0.004840
nitogen_dioxide	0.154001	1.000000	0.025412	0.069345
ozone	-0.560436	0.025412	1.000000	0.204257
HCHO reading	0.004840	0.069345	0.204257	1.000000

Kandy HCHO and Other Gas Emission Analysis

The plots below show how HCHO values relate to CO2, NO2, and Ozone emission rates in Kandy.



As in majority of cities, Kandy Proper had a very low correlation for CO2, NO2 and ozone gases. Carbon Monoxide had the highest correlation with a Pearson correlation of 0.1806.

	carbon_monoxide	nitogen_dioxide	ozone	HCHO reading
carbon_monoxide	1.000000	0.106242	-0.583233	0.186016
nitogen_dioxide	0.106242	1.000000	-0.013274	0.114099
ozone	-0.583233	-0.013274	1.000000	-0.014814
HCHO reading	0.186016	0.114099	-0.014814	1.000000

Conclusion Industrial Activity impact on HCHO emissions

Out of the analysed areas Colombo is considered as the industrial hub of Sri Lanka. It had a high correlation for all three gases analysed with HCHO gas emissions compared to all other cities. Therefore, it depicts that there is a huge impact from industrial activities for HCHO distribution.

Finalized Data Tables used to do the Powerbi Analysis

For powerbi Analysis, several data tables were used. The main table was used that consisted of the details of all regions including details such HCHO reading, weather data, other gas emission data and covid lockdown related data. The below table shows an example row of the main table used for analysis.

	Current Date	Next Date	HCHO reading	Region	LATITUDE	LONGITUDE	ELEVATION	PRCP	TAVG	TMAX	...	Area_sq_km	population_density
0	2019-01-01	2019-01-02	0.000176	Kandy Proper	7.29	80.63	500	0.09	71.096	79.592	...	28.53	3915.21
1	2019-01-02	2019-01-03	0.000092	Kandy Proper	7.29	80.63	500	0.01	70.124	78.764	...	28.53	3915.21
2	2019-01-03	2019-01-04	0.000134	Kandy Proper	7.29	80.63	500	0.02	69.728	79.448	...	28.53	3915.21
3	2019-01-04	2019-01-05	0.000191	Kandy Proper	7.29	80.63	500	0.00	69.224	79.610	...	28.53	3915.21
4	2019-01-05	2019-01-06	0.000122	Kandy Proper	7.29	80.63	500	0.04	72.068	82.526	...	28.53	3915.21

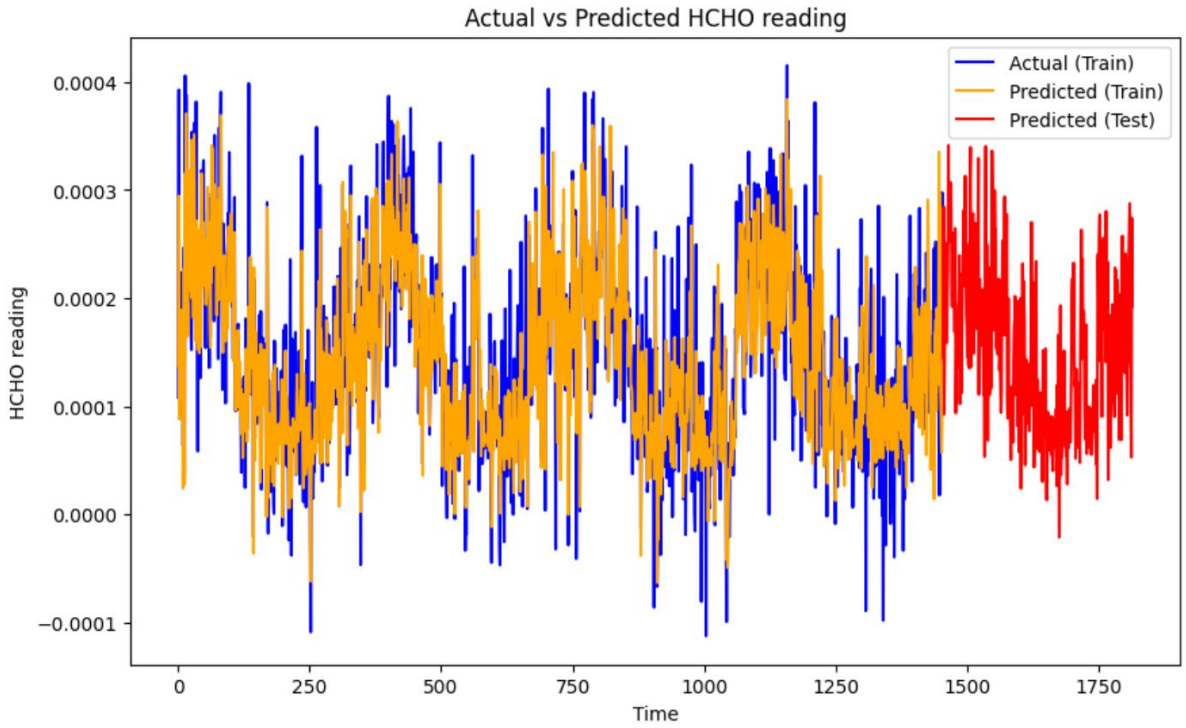
5 rows × 22 columns

Area_sq_km	population_density	Proximity(km)	carbon_monoxide	nitogen_dioxide	ozone	new_covid_patients	total_covid_patients	lockdown_status	covid_range_status
28.53	3915.21	52.76	0.033072	0.000012	0.116477	0	0	Before	Before
28.53	3915.21	52.76	0.032599	0.000013	0.116477	0	0	Before	Before
28.53	3915.21	52.76	0.031031	0.000016	0.116477	0	0	Before	Before
28.53	3915.21	52.76	0.030439	0.000010	0.116477	0	0	Before	Before
28.53	3915.21	52.76	0.031700	0.000014	0.116477	0	0	Before	Before

In addition, for each region, a separate table was used to analyse the data and display the correlations with external factors. Tables that consisted of correlation data were used. Furthermore, for COVID data analysis, another table was used.

Use of Time Series and LSTM Models for HCHO Forecasting

For each region, models were created separately, and both single-variate and multi-variate models were used to do the time series forecasting for all regions. Single-variate models were trained based on the HCHO emission rate each day. For all the regions, single-variate LSTM models performed well. For multi-variate time series forecasting, external factors were also considered based on the correlations with HCHO emissions in each region. For multi-variate time series models, the SARIMAX models performed well compared to other models in the majority of regions. However, cross-validation is done using ARIMA, SARIMA, auto-regression, and bayesian ridge models for both multi-variate and single-variate models. The auto-Arima function was used to find the best-fit parameters for seasonal variables for time series models. The models are evaluated based on pattern recognition and using regression-based evaluation metrics, namely mean absolute error, mean squared error, R2 score, root mean squared error, mean absolute error, and explained variance scores. An example of a prediction done from Colombo regions single variate LSTM model is given below.



Research Questions Arising from the Work

This chapter discusses the research-based questions arising from this work related to HCHO distribution, industrial, and spatial pattern impacts on HCHO distribution. These research questions can be applied to both Sri Lanka and other countries.

1. Are there any discernible seasonal trends in HCHO concentration, and if so, what factors might contribute to these patterns?
2. Can changes in HCHO concentration during lockdowns provide insights into the impact of reduced human activities on air quality?
3. Can variations in industrial emissions be linked to differences in HCHO distribution among various cities or regions, and if so, what factors contribute to these disparities?
4. Are there differences in HCHO distribution related to geographical factors such as elevation or proximity to the coastline?
5. How do meteorological variables such as temperature and precipitation influence formaldehyde emission rates in different cities and regions?

Comparison of study with other work

There was a similar study that was done to identify the impact of HCHO and NO2 distributions on spatial patterns in Southeast Asia. This study was conducted by the University of Panjab (RANA, 2019).This study has revealed that forest fires have had a lot of impact on HCHO emission rates. However, this study has not checked the HCHO emission with other gas emissions to find the industrial impact. In addition, it has found high HCHO emission regions in Southeast Asia. There was another study done to identify sources of formaldehyde (HCHO) to reduce ground-level pollution of HCHO and ozone (O3) by Taiyuan University, China (Cui, et al., 2020). This study shows that HCHO emissions help identify ozone emissions to manage air quality-based improvements. However, this study has not discussed HCHO distribution levels. There was not any research conducted to analyse HCHO distributions in Sri Lanka.

Future Recommendations for the Research

This research has encountered many obstacles due to the unavailability of data on HCHO emissions on certain days. Furthermore, some external factors related to data, such as population statistics, are more likely to be outdated since Sri Lanka has not conducted a proper study on population since 2012. In addition, for proximity-based calculations, online tools had to be used. These difficulties limited the reliability of some of the analyses done in this research. However, it provides information on how HCHO distribution has been impacted due to covid 19 lockdowns, what are the seasonal variations, and the spatial and industry-based impact on HCHO distributions.

To improve the investigations, it is important to address these challenges. Improving access to and utilization of available data resources and conducting the above research for other cities in Sri Lanka will be helpful to come up with meaningful conclusions of HCHO distribution and its impact. If Institutions like NASA can be used as collaborators, it will be useful to come up with more reliable research due to other external factors related data can be gathered by satellite data. These enhancements are critical to advance the understanding of HCHO distribution and its impact on external factors.

How can the findings be used for air quality, public health, and environmental management in Sri Lanka and how to use ethical implications?

Since this study provides data-driven evidence on the sources and spatial-temporal patterns of formaldehyde (HCHO) pollution, it can be used to develop targeted policies and regularizations for Sri Lanka based on HCHO seasonality patterns to reduce emissions from industrial processes, vehicular traffic, and other sources that contribute to HCHO pollution. In addition, by identifying the areas with high HCHO emissions, environmental management strategies can be implemented through conservation efforts. If these study-based results can be communicated to local communities and stakeholders, it will be helpful for fostering community engagement and participation in air quality improvement efforts. However, by raising awareness about the health risks associated with HCHO pollution, it will help individuals take proactive measures to reduce their exposure to harmful pollutants. Ethical considerations in formaldehyde analysis include accurate data interpretation and transparency. Unintended consequences could include economic costs or insufficient mitigating attempts.

Conclusion

In conclusion, this report examines the HCHO distribution in Sri Lanka from 2019 to 2023 using data from the TROPOMI instrument on the Sentinel-5P satellite. Through rigorous data preprocessing and analysis, it covers spatial and temporal variations of the HCHO concentrations, including seasonal trends, covid lockdown impacts and correlations with industrial emission and metrological factors. Despite the challenges had with the collected data, the findings provide actionable insights for air quality management and public health. If future research can be conducted by addressing the data limitations by collaborating with institutions like NASA and extend analysis to more Sri Lankan cities to inform policy and community driven efforts for environmental protection.

References

Anon., 2024. *mLocality Maps - Deniyaya, Matara District, Southern Province, Sri Lanka*. [Online]
Available at: https://www.mindat.org/maps.php?id=214407
[Accessed 03 April 2024].

Anon., n.d. *COVID-19 pandemic - Sri Lanka*. [Online]
Available at: https://global-monitoring.com/gm/page/events/epidemic-0002015.fTDtGCxti2qN.html?lang=en
[Accessed 03 April 2024].

Anon., n.d. *Population of Cities in Sri Lanka 2024*. [Online]
Available at: https://worldpopulationreview.com/countries/cities/sri-lanka
[Accessed 03 April 2024].

Cui, Y. et al., 2020. Comparison of three source apportionment methods based on observed and initial HCHO in Taiyuan, China. *Science of The Total Environment,* 926(171828).

Developers, S. A. E., n.d. *Emissions API.* [Online]
Available at: https://api.v2.emissions-api.org/ui/
[Accessed 02 April 2024].

FreeMapTools, n.d. *FreeMapTools.* [Online]
Available at: https://www.freemaptools.com/#google_vignette
[Accessed 04 April 2024].

Government, U., n.d. *National Centers For Environmental Information*. [Online]
Available at: https://www.ncdc.noaa.gov/cdo-web/search?datasetid=GHCND
[Accessed 2 April 2024].

Kotapola, D. S. -., n.d. *Divisional Secetriat - Kotapala*. [Online]
Available at: http://kotapola.ds.gov.lk/index.php/en/statistical-information.html
[Accessed 05 April 2024].

Kusterer, N. O. J. M., 2019. *Sentinel 5P NO2 Tropospheric vertical column data mapping*. [Online]
Available at: https://forum.earthdata.nasa.gov/viewtopic.php?t=4057
[Accessed 2024 20 April].

NASA, n.d. *Power | Data Access Viewer*. [Online]
Available at: https://power.larc.nasa.gov/data-access-viewer/
[Accessed 03 April 2024].

RANA, A. D., 2019. *ANTHROPOGENIC, BIOGENIC AND PYROGENIC EMISSION SOURCES AND ATMOSPHERIC FORMALDEHYDE (HCHO) AND NITROGEN DIOXIDE (NO2) COLUMNS OVER DIFFERENT LANDUSE/LANDCOVERS OF SOUTH ASIA,* s.l.: Institute of Geology, University of the Punjab, Lahore, Pakistan.