INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

ROBERT GORDON UNIVERSITY ABERDEEN

CM 2606 - Data Engineering

Coursework Report by

Binuka Rajapaksha - 20221332 / 2237043

Module Leader - Mr. Mohamed Ayoob

Submitted in partial fulfillment of the requirements for the BSc (Hons) in Artificial Intelligence and Data Science degree at the Robert Gordon University.

April 2024

Table Of Contents

Introduction

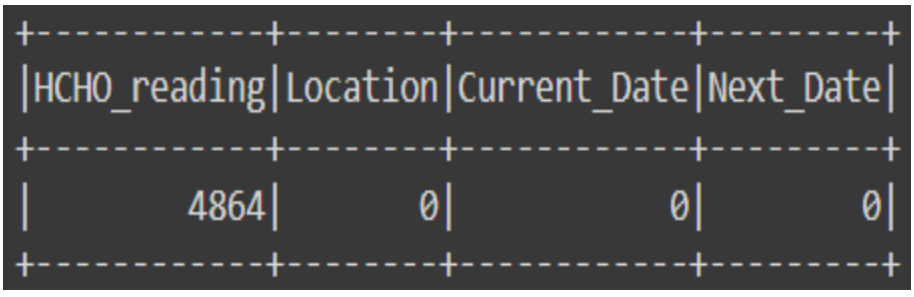
Formaldehyde (HCHO) emissions in urban environments present a critical concern for public health and environmental sustainability. As a hazardous air pollutant, formaldehyde is associated with various health risks, including respiratory irritation and potential carcinogenic effects. Consequently, comprehending the spatial and temporal patterns of HCHO levels is imperative for effective air quality management and public health protection in Sri Lankan cities.

Data Preprocessing

Before initiating any preprocessing, each dataset was loaded into Spark DataFrames, with each representing formaldehyde (HCHO) readings along with location and corresponding date information. To ensure consistency and clarity in data representation, the columns were renamed. Subsequently, the individual DataFrames were merged into a single DataFrame. The dataset consists of 12,782 rows, covering data from seven cities of Sri Lanka spanning from January 1, 2019 to December 31, 2023.

Handling null values

Overall, 4864 null values were identified, all within HCHO readings. Due to the time series nature of the data, I opted for the forward and backward fill method to address these null values. This approach ensured the preservation of temporal integrity while effectively handling missing data.



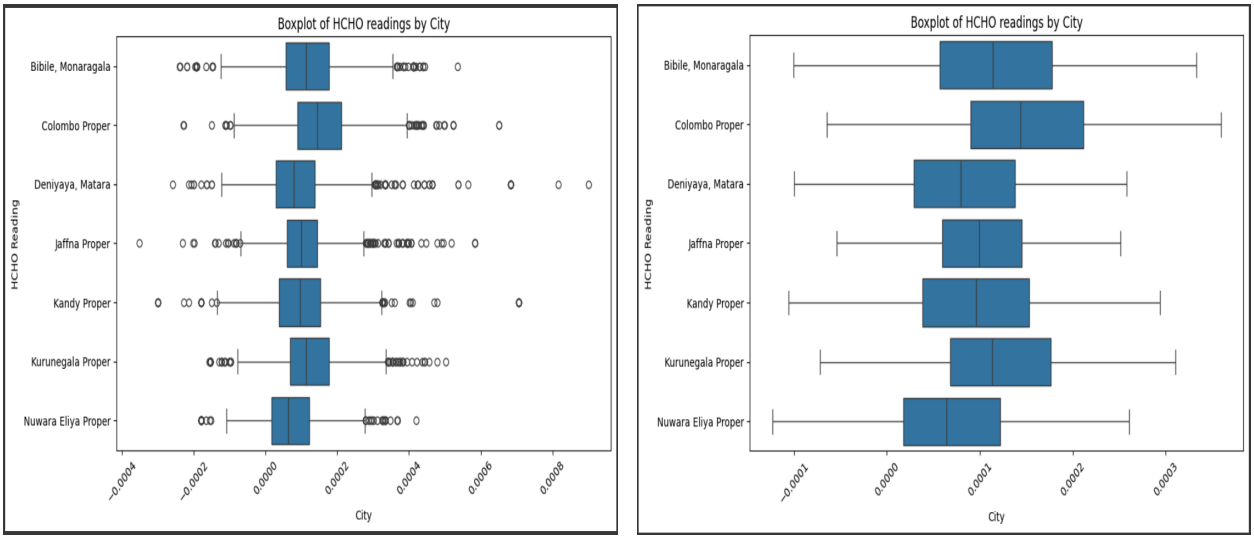
HCHO_reading	Location	Current_Date	Next_Date
4864	0	0	0

Handling outliers

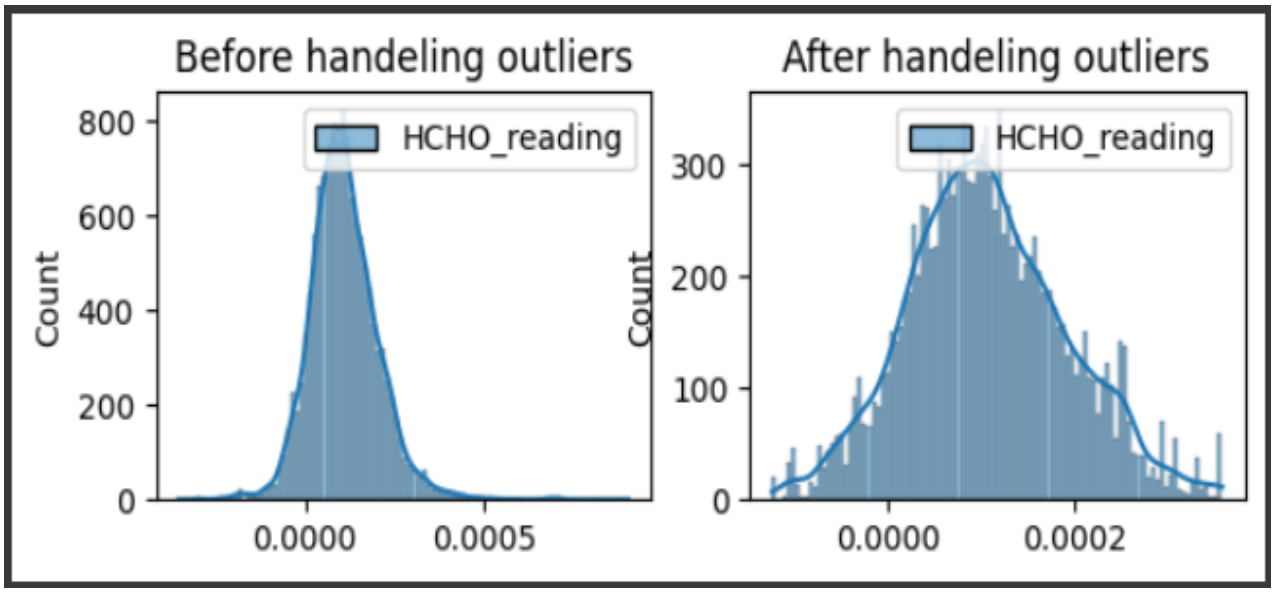
To address outliers in the dataset, the IQR method was employed, which is well-suited for handling extreme values in time series data. Through iteration, quartiles were calculated for each location in the dataset. Then , lower and upper bounds were defined based on the quartiles, and

outlier values were capped without removal. This approach ensures that extreme values are handled while retaining the overall distribution of the data.

Boxplots of each city in the dataset before and after handling outliers,



Distribution of the overall dataset before and after handling outliers,



Descriptive statistics

For entire dataset,

```
Descriptive Statistics for the entire Dataset:
+--------------------+--------------------+--------------------+
|          mean_HCHO|        median_HCHO|         stddev_HCHO|
+--------------------+--------------------+--------------------+
|1.075021429912154...|1.022405582927609E-4|8.648995502405537E-5|
+--------------------+--------------------+--------------------+
```

```
+-------+--------------------+
|summary|         HCHO_reading|
+-------+--------------------+
|  count|               12782|
|   mean|1.075021429912154...|
| stddev|8.648995502405537E-5|
|    min|-1.23359561118189...|
|    max| 3.59186346017419E-4|
+-------+--------------------+
```

For each city,

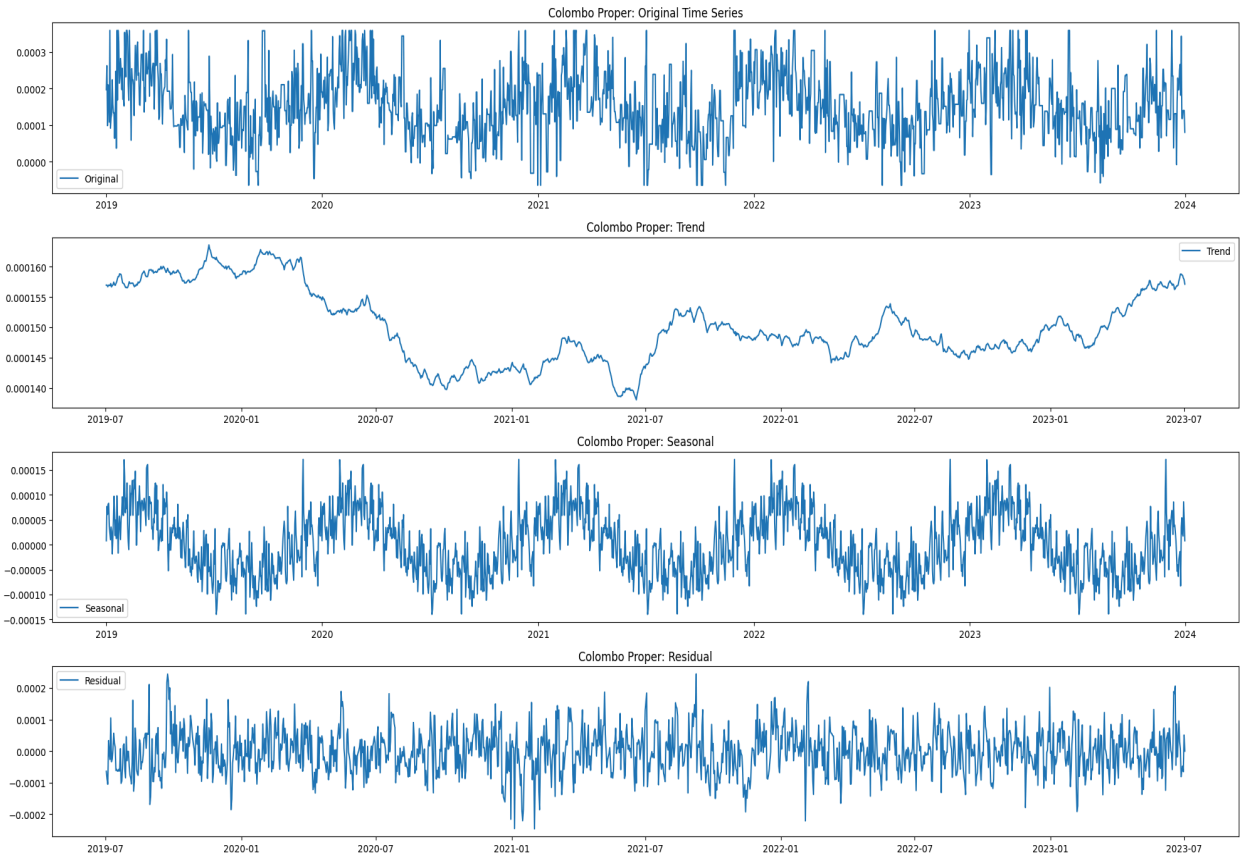
```
Descriptive Statistics for each city:
+------------------+--------------------+--------------------+--------------------+
|          Location|          mean_HCHO|        median_HCHO|         stddev_HCHO|
+------------------+--------------------+--------------------+--------------------+
| Bibile, Monaragala|1.180012245749719...|1.138090338532037...|8.914282712384695E-5|
|     Colombo Proper|1.517972267589955...|1.437350140173650...|9.306577078775164E-5|
|   Deniyaya, Matara|8.571720741113839E-5|7.966551476995192E-5|8.077190678726046E-5|
|      Jaffna Proper|1.046341388532473...|9.957104696875889E-5|6.666719492920703E-5|
|       Kandy Proper|9.720393395054452E-5|9.608622488355643E-5|8.691349003278773E-5|
|  Kurunegala Proper|1.222907015534841...|1.134143316690149...|8.111995471041423E-5|
|Nuwara Eliya Proper|7.287056783612621E-5|6.379205298112547E-5|8.121607602731918E-5|
+------------------+--------------------+--------------------+--------------------+
```

Spatio-Temporal Analysis

Seasonal decomposition analysis

The goal of conducting seasonal decomposition analysis on the HCHO readings is to reveal seasonal patterns within the data. This analysis aids in understanding the fluctuations of HCHO levels over time and in identifying recurring trends or anomalies. By decomposing the data into its seasonal components, we gain insights into the influence of seasonal factors on HCHO emissions. This information is valuable for environmental monitoring and management efforts, as it helps in understanding and mitigating the impact of HCHO emissions on environmental health.

For Colombo,

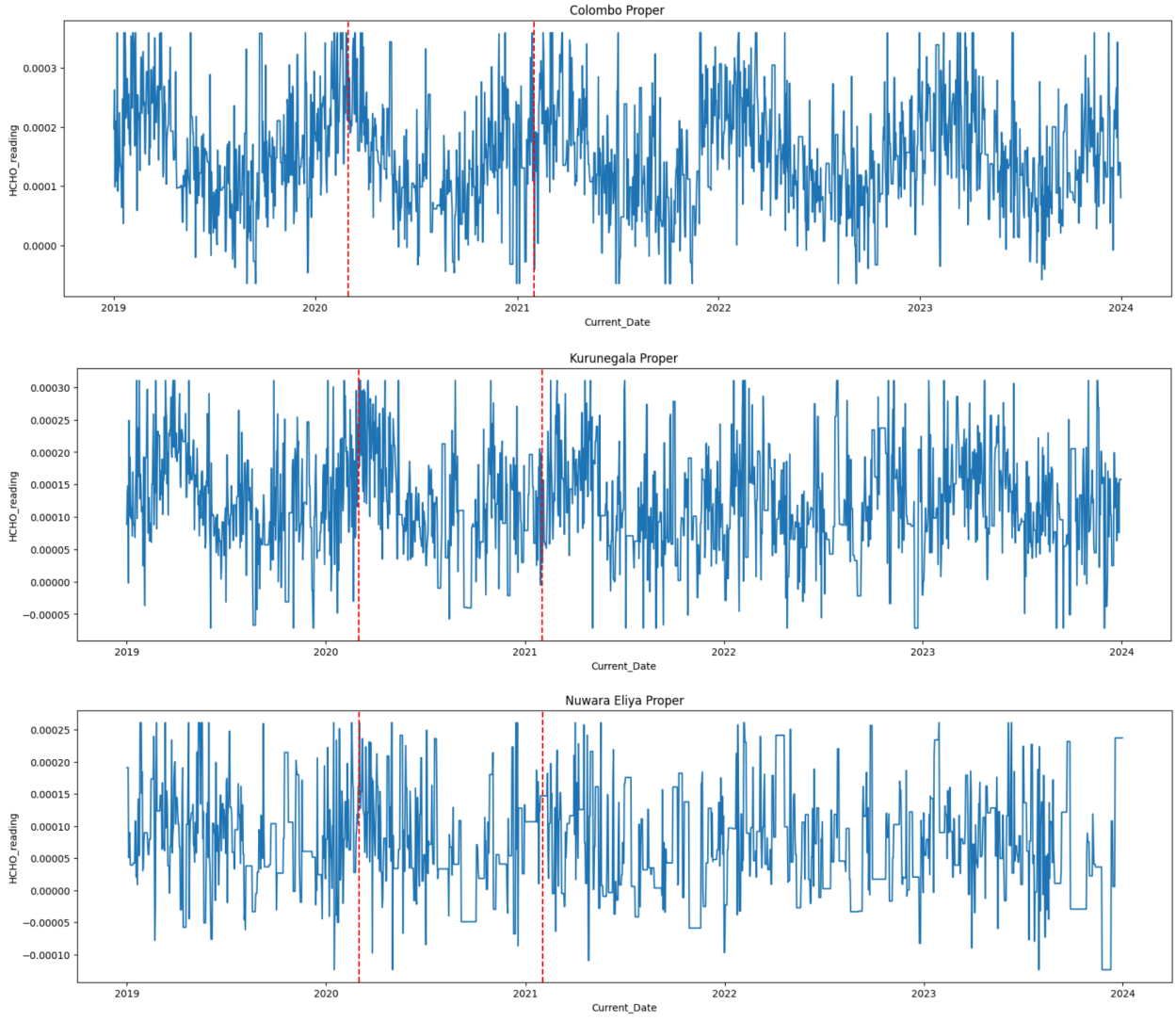


The first graph shows the original HCHO distribution over the time and the next three graphs display the trends, seasonal variations, and residual variations respectively. These graphs help to understand how HCHO levels fluctuate over time, the repeating patterns linked with seasonal

changes and what's left unexplained by trends or seasons. This breakdown gives a better grasp of how HCHO emissions evolve over time. This has been applied to each location separately.

COVID-19 lockdowns

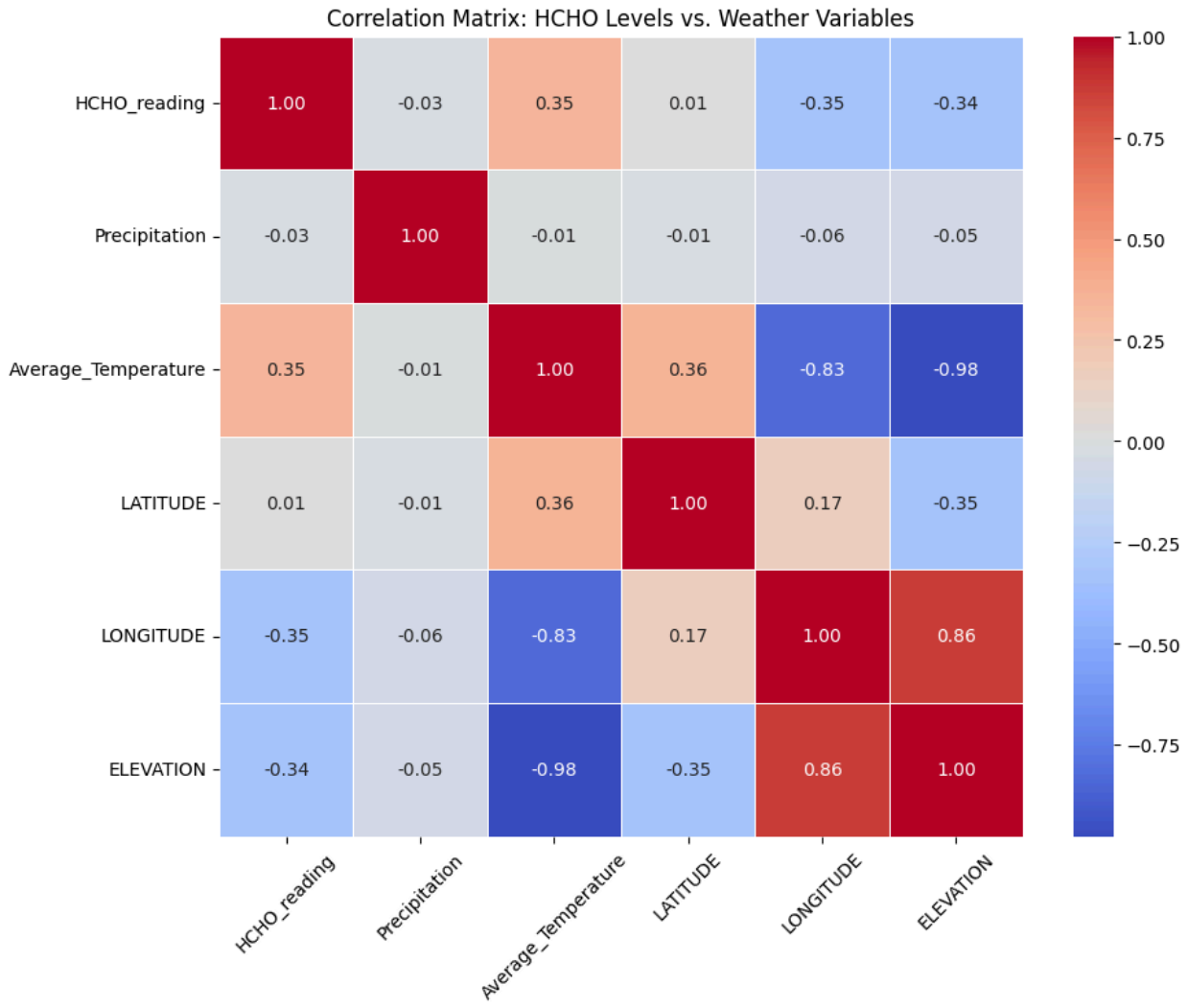
Investigating the influence of external factors such as COVID-19 lockdowns on HCHO emissions is crucial for understanding environmental dynamics.



External Factors

Analyzing facts that could potentially impact HCHO emissions is also crucial. So, consideration was given to climatic and geographical factors that may affect HCHO emission. The data utilized for this analysis was sourced from the National Climatic Data Center (NCDC). It contains climatic factors like precipitation, average temperature and geographical factors like elevation, latitude and longitude. However, it should be noted that the source has some limitations. It only contains data for 3 locations out of the 7 locations we have.

Correlation of HCHO levels with external factors,

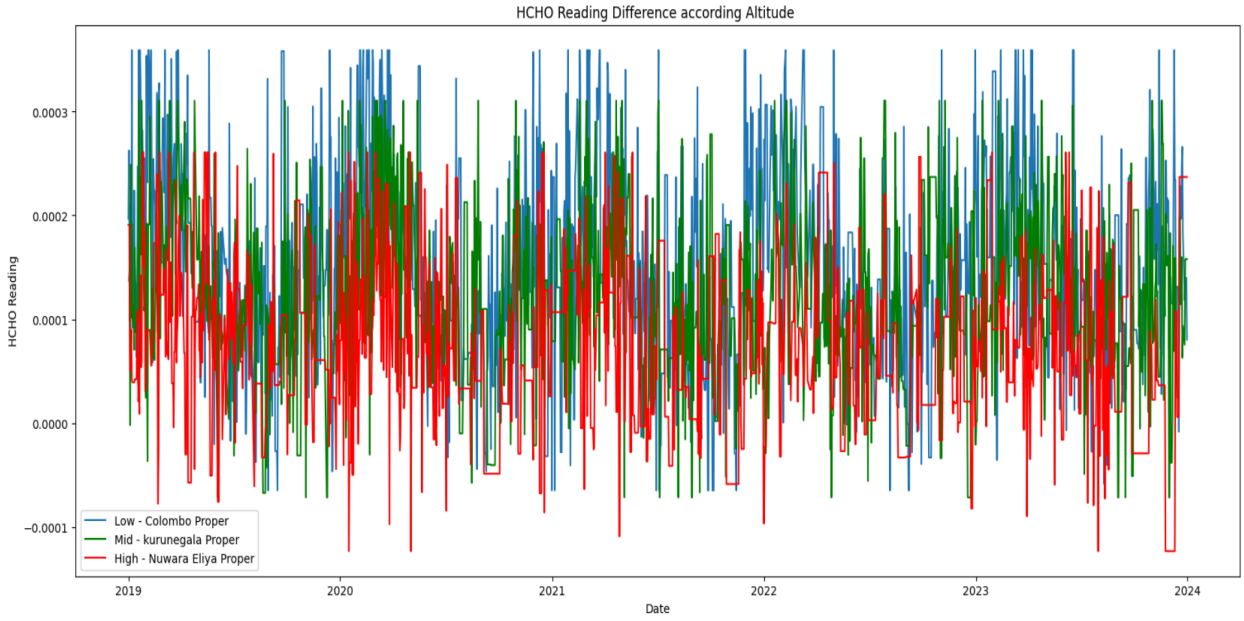


Here we can observe that climatic factors like temperature and geographical facts like elevation and longitude show moderate correlation with the HCHO readings.

HCHO Difference between low lying cities and high lying cities

Location	ELEVATION
Colombo Proper	7.0
Kurunegala Proper	116.0
Nuwara Eliya Proper	1880.0

Locations such as Colombo can be considered as low-elevation cities, while Nuwara Eliya can be considered as a high-elevation city. By comparing these two cities, insights can be gained regarding how elevation affects HCHO readings.



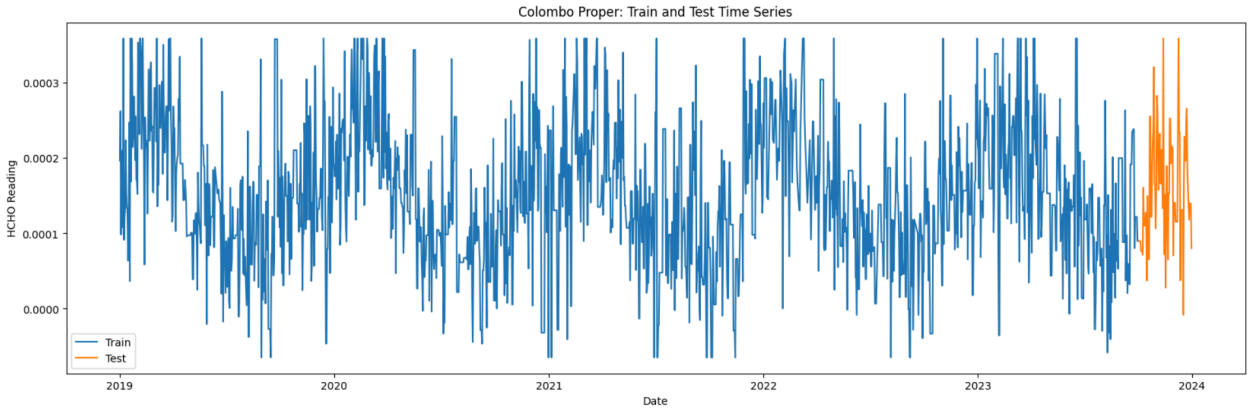
Low-elevation cities like Colombo showing higher HCHO levels compared to other cities. On the other hand, the High-elevation city, Nuwara Eliya, showed less HCHO levels compared to others. The high population density and intense industrial activities might also be crucial facts here.

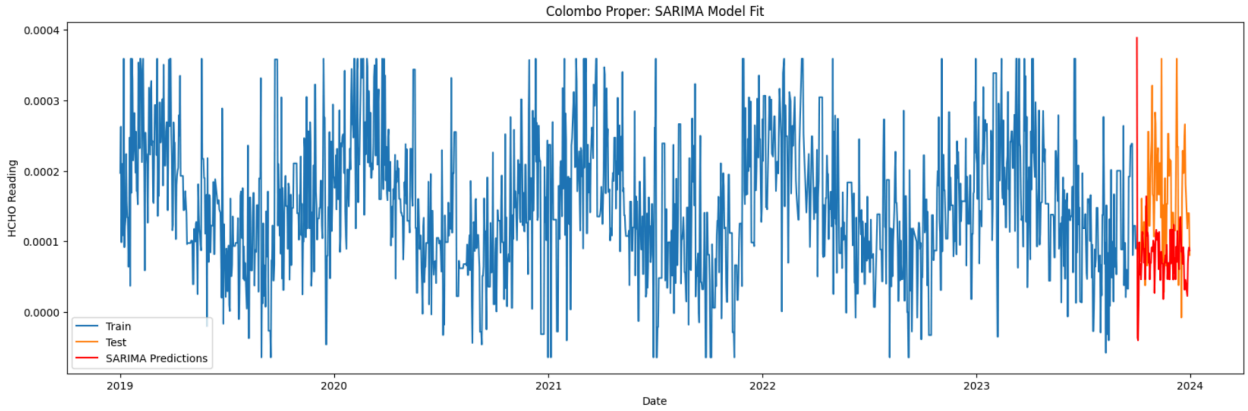
SARIMA Model

Model Development

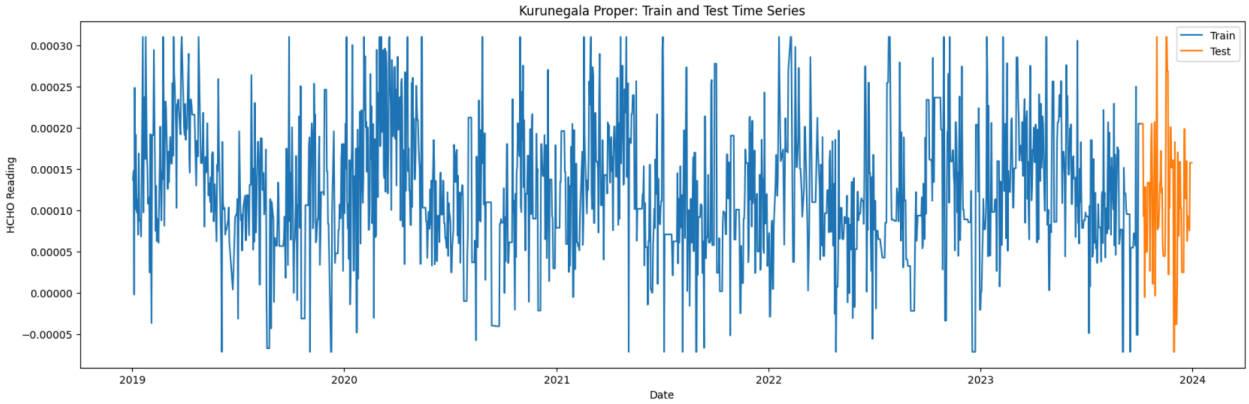
The SARIMA (Seasonal AutoRegressive Integrated Moving Average) model was employed to forecast HCHO levels for each city based on historical data. The dataset was divided into training and testing sets, with 95% of the data utilized for training and the remaining 5% for testing. The Auto ARIMA algorithm was utilized to automatically determine the optimal parameters for the SARIMA model, including the order and seasonal order. Exogenous variables such as average temperature, elevation, and precipitation were considered as additional features in the model. Time series plots were generated to visualize the HCHO readings, and model predictions were visualized alongside the actual HCHO readings for the testing period, facilitating a comparison between observed and forecasted values.

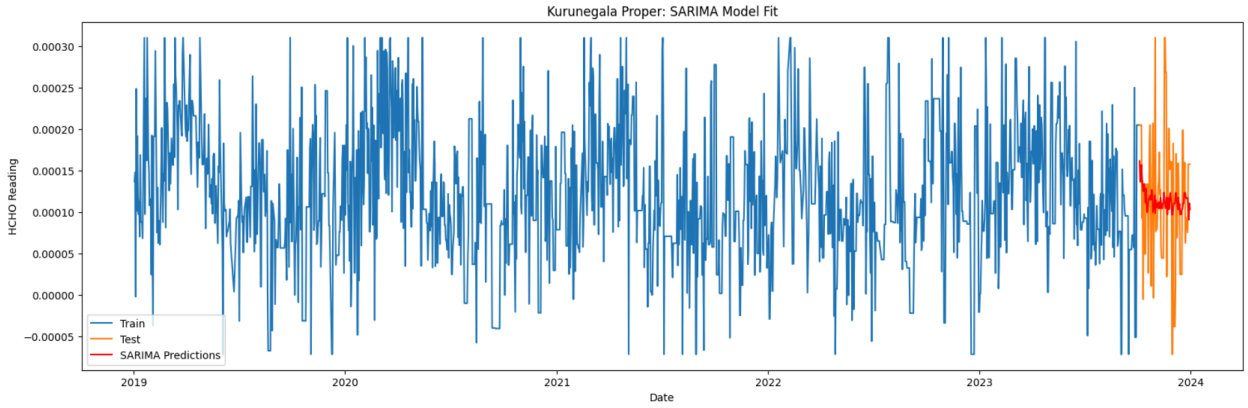
For Colombo,



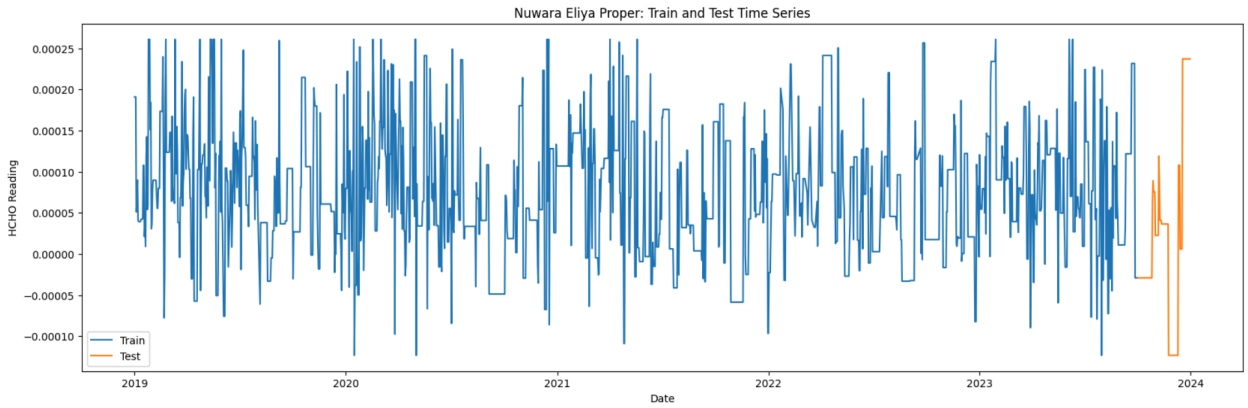


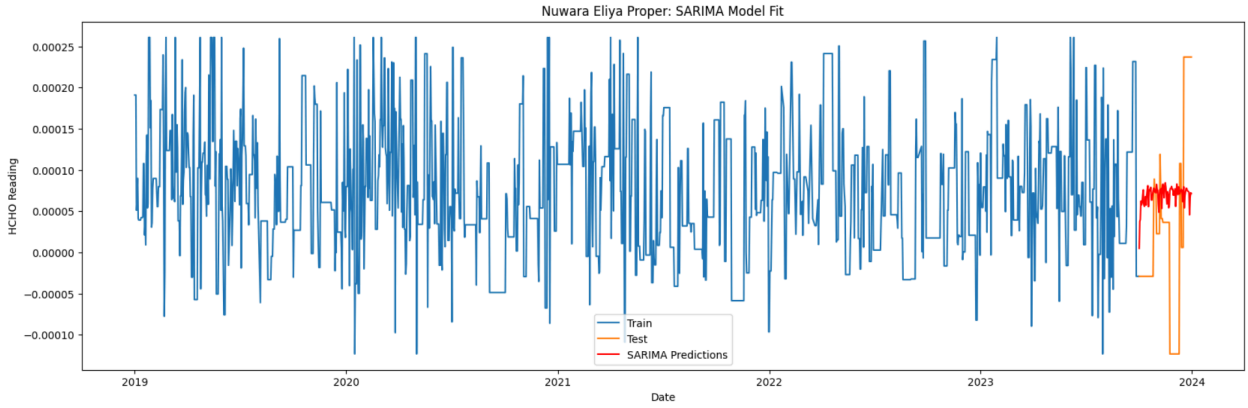
For Kurunegala,





For Nuwara Eliya,





Model evaluation

Evaluation metrics including Mean Squared Error (MSE), R-squared, and Mean Absolute Error (MAE) are calculated to assess the performance of the model. These metrics provide insights into the accuracy and reliability of the forecasted HCHO levels.

Below are the evaluation metrics of each model,

Colombo

```
Metrics for Colombo Proper:
MSE: 1.3433438104734688e-08
R-squared: -1.5740450128831238
MAE: 9.30267118181487e-05
```

Kurunegala

```
Metrics for Kurunegala Proper:
MSE: 6.503348072667198e-09
R-squared: 0.022874098700192413
MAE: 6.368880406380037e-05
```

Nuwara Eliya

```
Metrics for Nuwara Eliya Proper:
MSE: 1.4313404695975252e-08
R-squared: -0.2333921227784177
MAE: 9.981307507556583e-05
```

Limitation and Future Improvements

Limitations:-

- The negative R-squared values for some cities indicate that the model performs worse than a horizontal line. This suggests that the model fails to capture the underlying patterns in the data.
- Limited sources for external factors.

Improvements:-

- Conduct more extensive hyperparameter tuning to find the best parameters for the model.
- The HCHO readings contain negative values, this can be outliers since a concentration can not be a negative value. So handling them in earlier stages might also improve the overall model performance.
- Using Ensemble learning methods.
- Incorporate additional external data sources that may impact on HCHO levels.

Appendix

Power BI Dashboard