**INFORMATICS INSTITUTE OF TECHNOLOGY**

**In Collaboration with**

**ROBERT GORDON UNIVERSITY ABERDEEN**

**Artificial Intelligence and Data Science**

**Module Leader: Mohamed Ayoob**

**CM2604 - Data Engineering**

**Assignment Type: Individual Coursework**

**Student Name: Vanuja Thihansith Sooriyaarachchi**

**IIT ID - 20222408**

**RGU ID - 2311130**

© The copyright for this project and all its associated products resides with Informatics Institute of Technology

# Table of Contents

[Table of Contents 2](#_Toc164376321)

[Introduction 3](#_Toc164376322)

[1. Data Preprocessing 4](#_Toc164376323)

[1.1Clean and Prepare the Data 4](#_Toc164376324)

[1.1.1 Load the Data 4](#_Toc164376325)

[1.1.2 Explore Descriptive Statistics 4](#_Toc164376326)

[1.1.3 Handling Missing Values 5](#_Toc164376327)

[1.1.4 Handling Duplicates 6](#_Toc164376328)

[1.1.5 Handling Outliers 7](#_Toc164376329)

[2.Spatio-Temporal Analysis 13](#_Toc164376330)

[2.1 Analyze Trends Over Time 13](#_Toc164376331)

[2.1.1 Seasonal Variations 13](#_Toc164376332)

[2.1.2 Long-term Changes 14](#_Toc164376333)

[2.1.3 Trends Across Cities 15](#_Toc164376334)

[2.2 Changes in Gas Emissions due to the COVID-19 Lockdowns 22](#_Toc164376335)

# Introduction

# 1. Data Preprocessing

## Clean and Prepare the Data

### Load the Data

Load the col\_mat\_nuw\_output.csv file.

data = spark.read.csv("D:/IIT/2 nd Year/2nd Sem/Data Engineering/Course Work/HCHO\_Prediction/dataset/col\_mat\_nuw\_output.csv", header=True, inferSchema=True)

Load the kan\_output.csv file.

data\_2 = spark.read.csv("D:/IIT/2 nd Year/2nd Sem/Data Engineering/Course Work/HCHO\_Prediction/dataset/kan\_output.csv", header=True, inferSchema=True)

Load the mon\_kur\_jaf\_output.csv file.

data\_3 = spark.read.csv("D:/IIT/2 nd Year/2nd Sem/Data Engineering/Course Work/HCHO\_Prediction/dataset/mon\_kur\_jaf\_output.csv", header=True, inferSchema=True)

### Explore Descriptive Statistics

Summarize (mean, median, standard deviation) HCHO levels for each city on and across the entire dataset.

# Describe the 'HCHO reading' column

data.select('HCHO reading').describe().show()

+-------+--------------------+

|summary| HCHO reading|

+-------+--------------------+

| count| 3058|

| mean|1.200178195763001...|

| stddev|1.009287188756533...|

| min|-2.59296176552668...|

| max|8.997101837438971E-4|

+-------+--------------------+

+-------+--------------------+

|summary| HCHO reading|

+-------+--------------------+

| count| 1825|

| mean|9.890951713730535E-5|

| stddev|9.651844491820422E-5|

| min|-2.99702863135199...|

| max|7.051621763962024E-4|

+-------+--------------------+

+-------+--------------------+

|summary| HCHO reading|

+-------+--------------------+

| count| 5477|

| mean|1.192770268341103...|

| stddev|8.860002918894764E-5|

| min|-3.52473024357239...|

| max|5.837611392919413E-4|

+-------+--------------------+

### Handling Missing Values

# Check for null values in the DataFrame

This is obtained through the code below.

data.select([count(when(col(c).isNull(), c)).alias(c) for c in data.columns]).show()

Dataset 1

+------------+--------+------------+---------+

|HCHO reading|Location|Current Date|Next Date|

+------------+--------+------------+---------+

| 2419| 0| 0| 0|

+------------+--------+------------+---------+

Dataset 2

+------------+--------+------------+---------+

|HCHO reading|Location|Current Date|Next Date|

+------------+--------+------------+---------+

| 793| 0| 0| 0|

+------------+--------+------------+---------+

Dataset 3

+------------+--------+------------+---------+

|HCHO reading|Location|Current Date|Next Date|

+------------+--------+------------+---------+

| 1651| 0| 0| 0|

+------------+--------+------------+---------+

After removing null values

+------------+--------+------------+---------+

|HCHO Reading|Location|Current Date|Next Date|

+------------+--------+------------+---------+

| 0| 0| 0| 0|

+------------+--------+------------+---------+

### Handling Duplicates

**Calculate the length of DataFrame.**

# Count the number of rows in the DataFrame

data\_count = data.count()

# Show the length of the DataFrame

print("Length of DataFrame:", data\_count)



**Drop duplicate values.**

# Drop duplicates from the DataFrame

data\_no\_duplicates = data.dropDuplicates()

**Calculate the length of DataFrame after dropping duplicate values.**

# Count the number of rows in the DataFrame

data\_count = data.count()

# Show the length of the DataFrame

print("Length of DataFrame:", data\_count)



Length of DataFrame(Before drop duplicates) = Length of DataFrame(After drop duplicates)

So, there are no duplicate values in each of the data sets.

### Handling Outliers

A graph of a distribution of hcho reading

Description automatically generatedDistribution of HCHO reading of first dataset (col\_mat\_nuw\_output.csv)

Figure 1: Distribution of HCHO reading of first dataset.

Distribution of HCHO reading of first dataset after handling outliers

A graph of a distribution of hcho reading

Description automatically generated

Figure 2: Distribution of HCHO reading of first dataset after handling outliers.

Distribution of HCHO reading of second dataset (kan\_output.csv)

A graph of a distribution of hcho reading

Description automatically generated

Figure 3: Distribution of HCHO reading of second dataset.

Distribution of HCHO reading of second dataset after handling outliers

A graph of distribution of hcho reading

Description automatically generated

Figure 4: Distribution of HCHO reading of second dataset after handling outliers.

Distribution of HCHO reading of third dataset (mon\_kur\_jaf\_output.csv)

A diagram of a distribution of hcho reading

Description automatically generated

Figure 5: Distribution of HCHO reading of third dataset.

Distribution of HCHO reading of third dataset after handling outliers

A diagram of a distribution of hcho reading

Description automatically generated

Figure 6: Distribution of HCHO reading of third dataset after handling outliers.

Boxplot of HCHO reading of first dataset (col\_mat\_nuw\_output.csv)

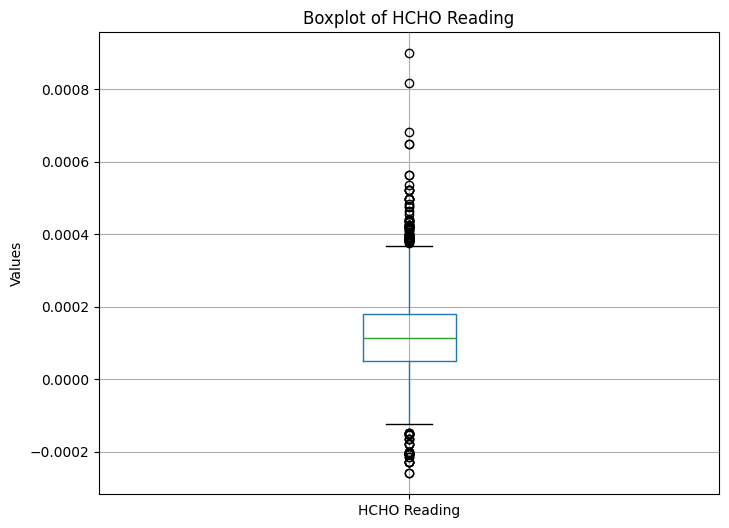


Figure 7: Boxplot of HCHO reading of first dataset.

Boxplot of HCHO reading of first dataset after handling outliers

A graph with a blue and green box

Description automatically generated

Figure 8: Boxplot of HCHO reading of first dataset after handling outliers.

Boxplot of HCHO reading of second dataset (kan\_output.csv)

A graph with a box and a line

Description automatically generated with medium confidence

Figure 9: Boxplot of HCHO reading of second dataset.

Boxplot of HCHO reading of second dataset after handling outliers

A diagram of a box with a green and blue line

Description automatically generated

Figure 10: Boxplot of HCHO reading of second dataset after handling outliers.

Boxplot of HCHO reading of third dataset (mon\_kur\_jaf\_output.csv)

A diagram of a box plot

Description automatically generated

Figure 11: Boxplot of HCHO reading of third dataset.

Boxplot of HCHO reading of third dataset after handling outliers

A graph with a blue and green line

Description automatically generated

Figure 12: Boxplot of HCHO reading of third dataset after handling outliers.

# Spatio-Temporal Analysis

## Analyze Trends Over Time

### Seasonal Variations

A graph with lines and points

Description automatically generated

Figure 13:Seasonal Variations in HCHO Level

### Long-term Changes

A graph of different colored lines

Description automatically generated

Figure 14: Average HCHO Reading by Year for Each City

A graph showing different colored bars

Description automatically generated

Figure 15: Average HCHO Readings by Location

### Trends Across Cities

**Bibile, Monaragala**

A graph of different colored lines

Description automatically generated with medium confidence

Figure 16:Bibile, Monaragala - Trend, Seasonal, Residual Component

**Colombo**

**A graph of different colored lines

Description automatically generated with medium confidence**

Figure 17: Colombo - Trend, Seasonal, Residual Component

**Deniyaya, Matara**

**A graph of different colored lines

Description automatically generated**

Figure 18: Deniyaya, Matara - Trend, Seasonal, Residual Component

**Jaffna**

**A group of graphs showing different colored lines

Description automatically generated**

Figure 19: Jaffna - Trend, Seasonal, Residual Component

**Kandy**

**A graph of different colored lines

Description automatically generated**

Figure 20: Kandy - Trend, Seasonal, Residual Component

**Kurunegala**

**A graph of different colored lines

Description automatically generated with medium confidence**

Figure 21: Kurunegala - Trend, Seasonal, Residual Component

**Nuwara Eliya**

**A graph of different colored lines

Description automatically generated with medium confidence**

Figure 22: Nuwara Eliya - Trend, Seasonal, Residual Component

## Changes in Gas Emissions due to the COVID-19 Lockdowns

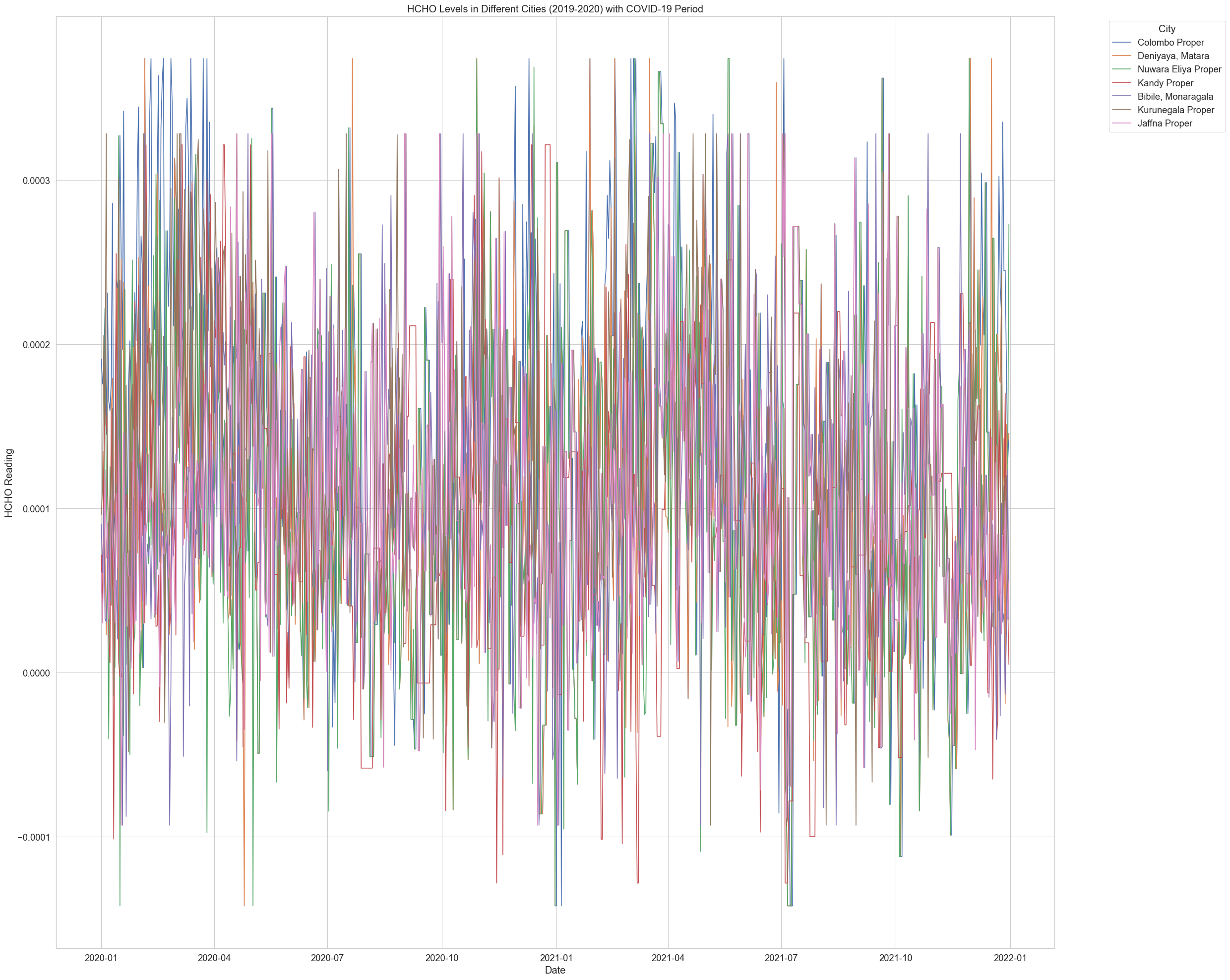


Figure 23: HCHO Levels in Different Cites with Covid-19 Period

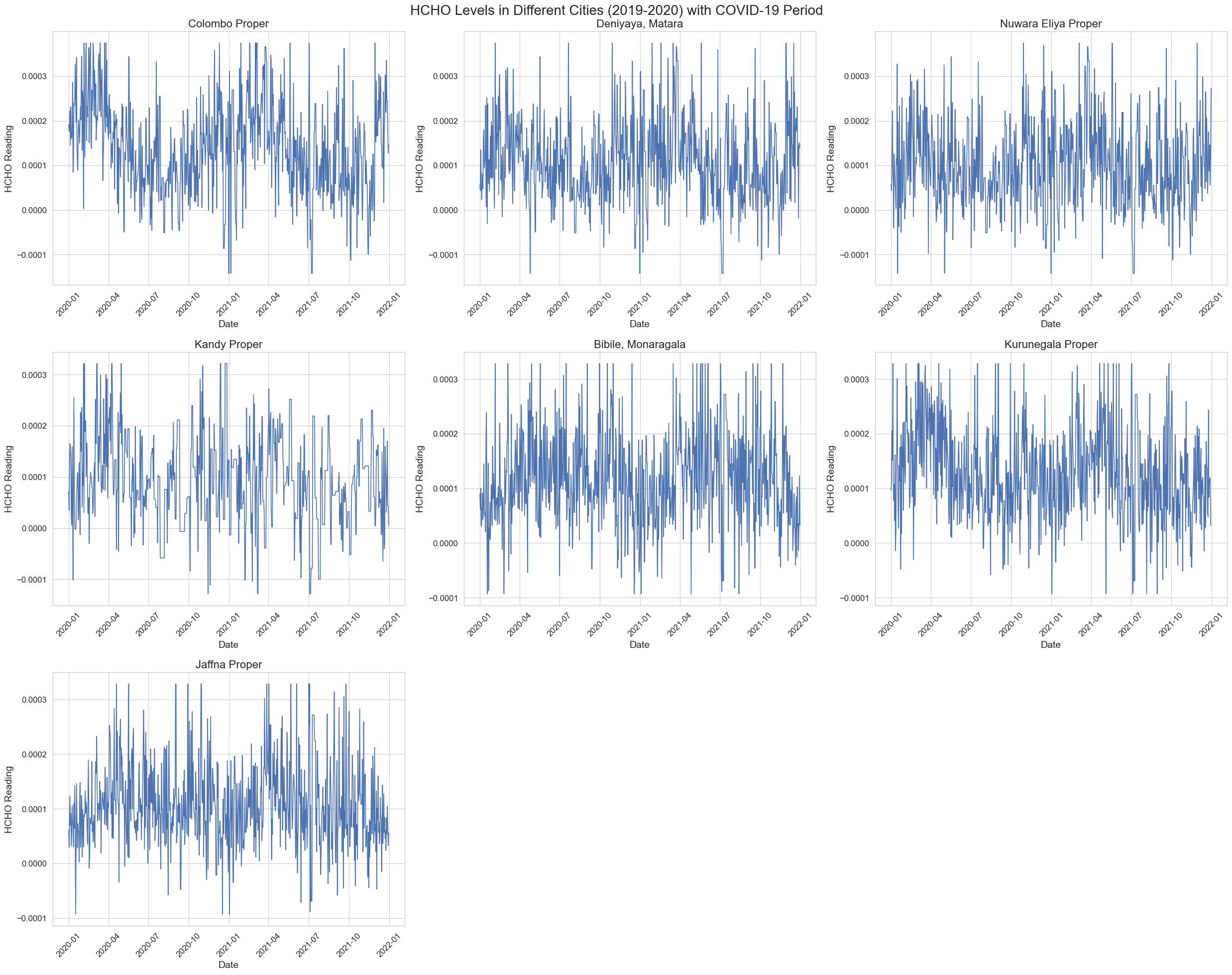


Figure 24: HCHO Levels in Different Cites with Covid-19 Period (One graph for each city)

This graph presents HCHO (formaldehyde) levels in various cities from 2019-2020, including a highlighted section that likely indicates the COVID-19 lockdown period.

**Colombo Proper:** Shows significant variability with some periods of elevated HCHO levels. During the lockdown, there is a visible reduction in the upper range of the fluctuations, suggesting a decrease in HCHO emissions during that period.

**Deniyaya, Matara:** Also exhibits variability, but with fewer extreme spikes compared to Colombo. The lockdown period does not show a substantial change in pattern.

**Nuwara Eliya Proper:** Displays a relatively consistent range of variability. Like Deniyaya, the lockdown doesn’t appear to have a pronounced effect on the HCHO levels.

**Kandy Proper:** The HCHO levels vary widely over time. The lockdown period doesn’t demonstrate a significant deviation from the overall variability pattern.

**Bibile, Monaragala:** The data show a high degree of fluctuation, and it seems the lockdown period corresponds with a section of reduced variability, although not as marked as in Colombo.

**Kurunegala Proper:** The fluctuations are less extreme than in Colombo but still show a wide range of variability. The lockdown period here also does not indicate a clear shift in the pattern of HCHO levels.

**Jaffna Proper:** The variability is consistent with no evident change during the lockdown period.

**Comparative Analysis:**

**Volatility:** All cities show significant fluctuations in HCHO levels. Colombo stands out with higher peaks, suggesting periods of elevated HCHO concentration, possibly due to more industrial activities or traffic.

**Lockdown Impact:** The impact of the lockdown on HCHO levels is inconsistent across the cities. Colombo shows a reduction in variability, which might indicate an environmental response to reduced human and industrial activity.

**Baseline Levels:** Cities like Jaffna, Kandy, and Kurunegala exhibit a stable pattern throughout, with no apparent impact from the lockdown. This could imply that sources of HCHO in these areas were less affected by lockdown measures, or other factors may be influencing the levels.

**Data Consistency:** Bibile, Monaragala has a wide range but less pronounced spikes, suggesting different sources or dynamics of HCHO emissions.

In general, HCHO levels are influenced by various factors, including industrial activities, traffic emissions, and possibly natural sources. The lockdown may have impacted these activities differently in each city. However, environmental factors such as temperature, humidity, and wind patterns can also affect HCHO levels and should be considered when analyzing these patterns. Moreover, the consistency of data collection and any local events (such as fires or industrial incidents) can also impact the data.