Exploratory Data Analysis on

Effect of Covid-19 Lockdown on Air Quality in India

by

Group 3

Gunesha Shahu
ID: 202201022
Course: BTech(ICT)

Lakshya Jain
ID: 202312106
Course: MSc(IT)

Megha Maheshwari
ID: 202312118
Course: MSc(IT)

Course Code: IT 462
Semester: Autumn 2024

Under the guidance of

Dr. Gopinath Panda

Dhirubhai Ambani Institute of Information and Communication Technology

December 2, 2024

Acknowledgment

I am writing this letter to express my heartfelt gratitude for your guidance and support throughout the duration of my project titled "Effect of Covid-19 Lockdown on Air Quality in India". Your invaluable assistance has played a pivotal role in shaping the successful completion of this endeavor.

I am extremely fortunate to have had the opportunity to work under your mentorship. Your expertise, encouragement, and willingness to share your knowledge have been instrumental in elevating the quality and scope of my project. Your constructive feedback and insightful suggestions have helped me overcome challenges and develop a deeper understanding of the subject matter.

I would also like to express my gratitude to my peers and colleagues who have been supportive throughout this journey. Their valuable input and camaraderie have been a constant source of motivation.

Completing this project has been a tremendous learning experience, and I am confident that the knowledge and skills acquired during this endeavor will serve as a solid foundation for my future endeavors.

Once again, thank you for your unwavering guidance and belief in my abilities. Your mentorship has been invaluable, and I am truly grateful for the opportunity to work with you.

Sincerely,
Gunesha Shahu, 202201022
Lakshya Jain, 202312106
Megha Maheshwari, 202312118.

DECLARATION

We, Gunesha, Lakshya and Megha hereby declare that the EDA project work presented in this report is our original work and has not been submitted for any other academic degree. All the sources cited in this report have been appropriately referenced.

We acknowledge that the data used in this project is obtained from the CPCB(Central Pollution Control Board) site and data.gov.in. We also declare that we have adhered to the terms and conditions mentioned in the website for using the dataset. We confirm that the dataset used in this project is true and accurate to the best of our knowledge.

We acknowledge that we have received no external help or assistance in conducting this project, except for the guidance provided by our mentor Prof. Gopinath Panda. We declare that there is no conflict of interest in conducting this EDA project.

We hereby sign the declaration statement and confirm the submission of this report on 2nd December, 2024.

Gunesha Shahu
ID: 202201022
Course: BTech(ICT)

Lakshya Jain
ID: 202312106
Course: MSc(IT)

Megha Maheshwari
ID: 202312118
Course: MSc(IT)

CERTIFICATE

This is to certify that Group 3 comprising Gunesha, Lakshya, and Megha has successfully completed an exploratory data analysis (EDA) project on Effect of Covid-19 Lockdown on Air Quality in India, which was obtained from cpcb.nic.in and data.gov.in.

The EDA project presented by Group 3 is their original work and has been completed under the guidance of the course instructor, Prof. Gopinath Panda, who has provided support and guidance throughout the project. The project is based on a thorough analysis of the air quality dataset, and the results presented in the report are based on the data obtained from the dataset.

This certificate is issued to recognize the successful completion of the EDA project on the Air Quality, which demonstrates the analytical skills and knowledge of the students of Group 3 in the field of data analysis.

Signed,
Dr. Gopinath Panda,
IT 462 Course Instructor
Dhirubhai Ambani Institute of Information and Communication Technology
Gandhinagar, Gujarat, INDIA.

December 2, 2024

Contents

List of Figures 5

1 Introduction 1
- 1.1 Our Project idea . . . 1
- 1.2 Data Collection . . . 1
- 1.3 Dataset Description . . . 1
  - 1.3.1 States chosen . . . 2
  - 1.3.2 Pollutants . . . 2
  - 1.3.3 Metropolitan city . . . 2
  - 1.3.4 State Population . . . 3
- 1.4 Packages required . . . 3

2 Data Cleaning 4
- 2.1 2015 . . . 4
  - 2.1.1 Missing data analysis . . . 4
  - 2.1.2 Imputation . . . 5
  - 2.1.3 Changing Data Types . . . 6
- 2.2 2016 . . . 7
  - 2.2.1 Missing data analysis . . . 7
  - 2.2.2 Imputation . . . 8
  - 2.2.3 Changing Data Types . . . 8
- 2.3 2017 . . . 8
  - 2.3.1 Missing data analysis . . . 8
  - 2.3.2 Imputation . . . 8
  - 2.3.3 Changing Data Types . . . 8
- 2.4 2018 . . . 9
- 2.5 2019 . . . 10

3 Visualization 11
- 3.1 Univariate analysis . . . 11
  - 3.1.1 SO_2 emissions . . . 11
  - 3.1.2 NO_2 emissions . . . 19
  - 3.1.3 PM_{10} emissions . . . 27
- 3.2 Multivariate analysis . . . 34

4 Feature Engineering **38**
4.1 Feature extraction 38
4.2 Feature selection 41
4.3 Outlier Detection 41

5 Model fitting **42**
5.1 Regression 42
5.1.1 Ahmedabad 42
5.1.2 More results 44

6 Conclusion & future scope **47**
6.1 Conclusion 47
6.2 Findings/observations 47
6.3 Challenges 47
6.4 Future plan 48

List of Figures

2.1.1 Basic column information for year 2015. 4
2.1.2 Missing value count. 5
2.1.3 Basic column information after changing data type. 6
2.2.4 Basic column information for year 2016. 7
2.2.5 Missing value analysis. 7
2.3.6 Basic information for the year 2017. 8
2.4.7 Final table for 2018 after cleaning. 9
2.5.8 Final table for 2019 after cleaning. 10

3.1.1 Bar Plot of Ahmedabad's SO_2 Emissions 11
3.1.2 Box Plot of Ahmedabad's SO_2 Emissions 12
3.1.3 Bar Plot of Badlapur's SO_2 Emissions 12
3.1.4 Box Plot of Badlapur's SO_2 Emissions 12
3.1.5 Bar Plot of Aizawal's SO_2 Emissions 13
3.1.6 Box Plot of Aizawal's SO_2 Emissions 13
3.1.7 Bar Plot of Bongaigaon's SO_2 Emissions 14
3.1.8 Box Plot of Bongaigaon's SO_2 Emissions 14
3.1.9 Bar Plot of Bangalore's SO_2 Emissions 15
3.1.10 Box Plot of Bangalore's SO_2 Emissions 15
3.1.11 Bar Plot of Belgaum's SO_2 Emissions 16
3.1.12 Box Plot of Belgaum's SO_2 Emissions 16
3.1.13 Bar Plot of Chandigarh's SO_2 Emissions 17
3.1.14 Box Plot of Chandigarh's SO_2 Emissions 17
3.1.15 Bar Plot of Allahabad's SO_2 Emissions 18
3.1.16 Box Plot of Allahabad's SO_2 Emissions 18
3.1.17 Bar Plot of Ahmedabad's NO_2 Emissions 19
3.1.18 Box Plot of Ahmedabad's NO_2 Emissions 19
3.1.19 Bar Plot of Badlapur's NO_2 Emissions 20
3.1.20 Box Plot of Badlapur's NO_2 Emissions 20
3.1.21 Bar Plot of Aizawal's NO_2 Emissions 21
3.1.22 Box Plot of Aizawal's NO_2 Emissions 21
3.1.23 Bar Plot of Bongaigaon's NO_2 Emissions 22
3.1.24 Box Plot of Bongaigaon's NO_2 Emissions 22
3.1.25 Bar Plot of Bangalore's NO_2 Emissions 23
3.1.26 Box Plot of Bangalore's NO_2 Emissions 23
3.1.27 Bar Plot of Belgaum's NO_2 Emissions 24

3.1.28 Box Plot of Belgaum's NO_2 Emissions . . . 24
3.1.29 Bar Plot of Chandigarh's NO_2 Emissions . . . 25
3.1.30 Box Plot of Chandigarh's NO_2 Emissions . . . 25
3.1.31 Bar Plot of Allahabad's NO_2 Emissions . . . 26
3.1.32 Box Plot of Allahabad's NO_2 Emissions . . . 26
3.1.33 Bar Plot of Ahmedabad's PM_{10} Emissions . . . 27
3.1.34 Box Plot of Ahmedabad's PM_{10} Emissions . . . 27
3.1.35 Bar Plot of Badlapur's PM_{10} Emissions . . . 28
3.1.36 Box Plot of Badlapur's PM_{10} Emissions . . . 28
3.1.37 Bar Plot of Aizawal's PM_{10} Emissions . . . 29
3.1.38 Box Plot of Aizawal's PM_{10} Emissions . . . 29
3.1.39 Bar Plot of Bongaigaon's PM_{10} Emissions . . . 30
3.1.40 Box Plot of Bongaigaon's PM_{10} Emissions . . . 30
3.1.41 Bar Plot of Bangalore's PM_{10} Emissions . . . 31
3.1.42 Box Plot of Bangalore's PM_{10} Emissions . . . 31
3.1.43 Bar Plot of Belgaum's PM_{10} Emissions . . . 32
3.1.44 Box Plot of Belgaum's PM_{10} Emissions . . . 32
3.1.45 Bar Plot of Chandigarh's PM_{10} Emissions . . . 33
3.1.46 Box Plot of Chandigarh's PM_{10} Emissions . . . 33
3.1.47 Bar Plot of Allahabad's PM_{10} Emissions . . . 34
3.1.48 Box Plot of Allahabad's PM_{10} Emissions . . . 34
3.2.50 Bar Plot of Allahabad's SO_2, PM_10, NO_2 Emissions . . . 35
3.2.51 Bar Plot of Aizwal's SO_2, PM_10, NO_2 Emissions . . . 35
3.2.49 Bar Plot of Ahmedabad's SO_2, PM_{10}, NO_2 Emissions . . . 35
3.2.52 Bar Plot of Badlapur's SO_2, PM_10, NO_2 Emissions . . . 36
3.2.53 Bar Plot of Belgaum's SO_2, PM_10, NO_2 Emissions . . . 36
3.2.54 Bar Plot of Bongaigaon's SO_2, PM_10, NO_2 Emissions . . . 36
3.2.55 Bar Plot of Chandigarh's SO_2, PM_10, NO_2 Emissions . . . 37

4.1.1 2015 . . . 38
4.1.2 2016 . . . 39
4.1.3 2017 . . . 39
4.1.4 2018 . . . 40
4.1.5 2019 . . . 40

5.1.1 Dropping rows for model training . . . 42
5.1.2 Train - test split . . . 42
5.1.3 Results . . . 43
5.1.4 Comparative Results . . . 43
5.1.5 NO2 prediction for Ahmedabad . . . 44
5.1.6 SO2 prediction for Allahabad . . . 44
5.1.7 PM10 prediction for Allahabad . . . 45
5.1.8 SO2 prediction for Aizawal . . . 45
5.1.9 NO2 prediction for Aizawal . . . 46
5.1.10 PM10 prediction for Aizawal . . . 46

List of Tables

Abstract

Air pollution is a significant environmental challenge in India, with severe implications for public health, ecosystems, and economic development. This study examines air quality trends across major Indian cities, focusing on a comparative analysis of pollution levels during the COVID-19 lockdowns and subsequent periods. The research utilizes Air Quality Index (AQI) data and concentrations of key pollutants, including PM2.5, PM10, NO2, and SO2, to evaluate the impact of reduced human activity and industrial operations during the pandemic. By analyzing these variations, the project explores the relationship between human activities, regulatory policies, and environmental conditions. The findings provide critical insights into sustainable air quality management strategies, offering valuable recommendations for policymakers, researchers, and environmental advocates striving to mitigate pollution and promote healthier living conditions in India.

Chapter 1. Introduction

1.1 Our Project idea

The Covid - 19 pandemic brought an unexpected halt in various activities including emissions from industries, transportation and from other sources. It has been seen that the quality of air quality is decreasing day by day with increasing population and thus increasing use of resources. But the pandemic brought an interesting change in the trends observed in air quality.

The objective of this project is to observe the air quality data from 2015 to 2019 on the basis of PM_{10}, NO_2 and SO_2 and the effect of population and whether the city is a metropolitan or not and then make predictions about the air quality from 2020 to 2023. Thereafter, observe the actual air quality and the predicted air quality and deduce what have been the effects of the pandemic on air quality in India.

By using statistical tools, the analysis aims to quantify the changes in air quality and understand how reduced human activity influenced environmental conditions. This research can inform policymakers on the benefits of sustainable practices for long-term air quality improvement.

1.2 Data Collection

Data collection from authoritative sources like the Cental Pollution Control Board and data.gov.in ensures reliability and credibility. These institutions employ rigorous methodologies to gather and analyze data, providing valuable insights for policymaking, research, and decision-making processes. By accessing data from such reputable sources, one can enhance the accuracy and robustness of their analyses and contribute to evidence- based decision-making on a global scale.

Furthermore, we got the year-wise data in the pdf and converted it to csv format on our own and also did the manual concatenation of multiple datasets. We have mainly merged 3 datasets that included information about the pollutants, population information and whether a city is metropolitan or not.

1.3 Dataset Description

We have formulated multiple year wise datasets which include the data from years 2015-2023.

1.3.1 States chosen

The states included in our dataset are:

- Andhra Pradesh
- Assam
- Bihar
- Chattisgarh
- Chandigarh
- Dadra and Nagar Haveli
- Daman and Diu
- Delhi
- Goa
- Gujarat
- Himachal Pradesh
- Jammu and Kashmir
- Jharkhand
- Karnataka
- Kerala
- Madhya Pradesh
- Maharashtra
- Meghalaya
- Mizoram
- Nagaland
- Orissa
- Punjab
- Puducherry
- Rajasthan
- Sikkim
- Tamil Nadu
- Telangana
- Uttar Pradesh
- Uttarakhand
- West Bengal

1.3.2 Pollutants

Our data set contains the following pollutants: PM_{10}, $PM_{2.5}$, NO_2 and SO_2. They are the main contributors of air pollution and are therefore crucial for our analysis.

Our dataset contains the minimum observed values, maximum observed values and the annual average values of these pollutants. The year 2015 is missing the $PM_{2.5}$ field which couldn't be found by us but the rest of the data is complete.

1.3.3 Metropolitan city

Metropolitan cities generally have more air pollution compared to smaller towns or rural areas. This is primarily due to the concentration of human activities and infrastructure in urban centers.Here are some key effects metropolitan areas have on air quality:

- Vehicular Emissions
- Industrial Pollution

- Construction and Urbanization
- Household Emissions
- Waste Management
- Population Density

All these factors have a big impact on the air quality and so we have factored in if a city is a metropolitan city or not.

1.3.4 State Population

Population also has a major role in the total emissions of a state and so we factored in the population of the states to make predictions based on them.

1.4 Packages required

We have used various packages during our project to help us analyse and visualise the data efficiently. The package are:

1. NumPy (np) NumPy is a fundamental package for scientific computing with Python.We have used it for numerical computations, array manipulation, and linear algebra operations.

2. Pandas (pd) Pandas is a powerful data manipulation and analysis library. We have used it for data cleaning, transformation, exploration, and analysis tasks.

3. Seaborn (sns) Seaborn is a statistical data visualization library based on Matplotlib. We have used it for creating visually appealing plots for exploring relationships in data.

4. Matplotlib.pyplot (plt) Matplotlib is a comprehensive plotting library for Python. We have used for creating various types of plots and visualizations, including line plots, scatter plots, histograms, etc.

5. Missingno (msno) Missingno is a Python library for visualizing missing data in datasets. It provides informative visualizations such as bar charts, heatmaps, and dendograms to identify missing values in datasets. It helped us to know the completeness of data.

6. Scikit-learn (sklearn) Scikit-learn is a machine-learning library for Python. We have used for tasks such as regression, and model selection.

Chapter 2. Data Cleaning

We are going to make use of datasets from years 2015 to 2019.

2.1 2015

Here is the basic information about the columns of the dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 539 entries, 0 to 538
Data columns (total 15 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Unnamed: 0                539 non-null    int64
 1   State                     536 non-null    object
 2   City                      536 non-null    object
 3   Location                  536 non-null    object
 4   2015 SO2 Minimum          536 non-null    float64
 5   2015 SO2 Maximum          536 non-null    float64
 6   2015 SO2 Annual Average   536 non-null    float64
 7   2015 NO2 Minimum          538 non-null    float64
 8   2015 NO2 Maximum          538 non-null    float64
 9   2015 NO2 Annual Average   538 non-null    float64
 10  2015 PM10 Minimum         539 non-null    float64
 11  2015 PM10 Maximum         539 non-null    int64
 12  2015 PM10 Annual Average  539 non-null    int64
 13  Is metropolitan city?     539 non-null    object
 14  State Population 2011     536 non-null    float64
dtypes: float64(8), int64(3), object(4)
memory usage: 63.3+ KB
```

Figure 2.1.1: Basic column information for year 2015.

2.1.1 Missing data analysis

We did not find any missing data for this year.

```
Unnamed: 0                    0
State                         0
City                          0
Location                      0
2015 SO2 Minimum              0
2015 SO2 Maximum              0
2015 SO2 Annual Average       0
2015 NO2 Minimum              0
2015 NO2 Maximum              0
2015 NO2 Annual Average       0
2015 PM10 Minimum             0
2015 PM10 Maximum             0
2015 PM10 Annual Average      0
Is metropolitan city?         0
State Population 2011         0
dtype: int64
```

Figure 2.1.2: Missing value count.

2.1.2 Imputation

Since there are no missing values, no imputation is required.

2.1.3 Changing Data Types

We have four columns with object data type, they are State, City, Location and Is it metropolitan city. While we can't change the 1st three, we can map the values of metropolitan city to 0 and 1.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 536 entries, 0 to 535
Data columns (total 15 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Unnamed: 0                536 non-null    int64
 1   State                     536 non-null    object
 2   City                      536 non-null    object
 3   Location                  536 non-null    object
 4   2015 SO2 Minimum          536 non-null    int64
 5   2015 SO2 Maximum          536 non-null    int64
 6   2015 SO2 Annual Average   536 non-null    int64
 7   2015 NO2 Minimum          536 non-null    float64
 8   2015 NO2 Maximum          536 non-null    int64
 9   2015 NO2 Annual Average   536 non-null    float64
 10  2015 PM10 Minimum         536 non-null    float64
 11  2015 PM10 Maximum         536 non-null    int64
 12  2015 PM10 Annual Average  536 non-null    int64
 13  Is metropolitan city?     536 non-null    int64
 14  State Population 2011     536 non-null    int64
dtypes: float64(3), int64(9), object(3)
memory usage: 62.9+ KB
```

Figure 2.1.3: Basic column information after changing data type.

2.2 2016

Here is the basic information about the columns of the dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 585 entries, 0 to 584
Data columns (total 18 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Unnamed: 0                 584 non-null    float64
 1   State                      584 non-null    object
 2   City                       584 non-null    object
 3   Location                   584 non-null    object
 4   2016 SO2 Minimum           584 non-null    float64
 5   2016 SO2 Maximum           584 non-null    float64
 6   2016 SO2 Annual Average    584 non-null    float64
 7   2016 NO2 Minimum           584 non-null    float64
 8   2016 NO2 Maximum           584 non-null    float64
 9   2016 NO2 Annual Average    584 non-null    float64
 10  2016 PM10 Minimum          584 non-null    float64
 11  2016 PM10 Maximum          584 non-null    float64
 12  2016 PM10 Annual Average   584 non-null    float64
 13  2016 PM2.5 Minimum         164 non-null    float64
 14  2016 PM2.5 Maximum         164 non-null    float64
 15  2016 PM2.5 Annual Average  164 non-null    float64
 16  Is metropolitan city?      585 non-null    object
 17  State Population 2011      585 non-null    int64
dtypes: float64(13), int64(1), object(4)
memory usage: 82.4+ KB
```

Figure 2.2.4: Basic column information for year 2016.

2.2.1 Missing data analysis

There are a lot of missing values for $PM_{2.5}$.

2016 PM10 Maximum	0
2016 PM10 Annual Average	0
2016 PM2.5 Minimum	418
2016 PM2.5 Maximum	418
2016 PM2.5 Annual Average	418

Figure 2.2.5: Missing value analysis.

2.2.2 Imputation

Since a lot of values are missing for $PM_{2.5}$ we can do the analysis with PM_{10} and drop the $PM_{2.5}$ columns.

2.2.3 Changing Data Types

Similar to 2015, we map the yes and no values of the metropolitan column to 1 and 0 respectively.

2.3 2017

Here is the basic information about the columns of the dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 669 entries, 0 to 668
Data columns (total 18 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Unnamed: 0                 669 non-null    int64
 1   State                      669 non-null    object
 2   City                       669 non-null    object
 3   Location                   669 non-null    object
 4   2017 SO2 Minimum           669 non-null    int64
 5   2017 SO2 Maximum           669 non-null    int64
 6   2017 SO2 Annual Average    669 non-null    int64
 7   2017 NO2 Minimum           669 non-null    int64
 8   2017 NO2 Maximum           669 non-null    int64
 9   2017 NO2 Annual Average    669 non-null    int64
 10  2017 PM10 Minimum          669 non-null    int64
 11  2017 PM10 Maximum          669 non-null    int64
 12  2017 PM10 Annual Average   669 non-null    int64
 13  2017 PM2.5 Minimum         205 non-null    float64
 14  2017 PM2.5 Maximum         205 non-null    float64
 15  2017 PM2.5 Annual Average  205 non-null    float64
 16  Is metropolitan city?      669 non-null    object
 17  State Population 2011      669 non-null    int64
dtypes: float64(3), int64(11), object(4)
memory usage: 94.2+ KB
```

Figure 2.3.6: Basic information for the year 2017.

2.3.1 Missing data analysis

We performed the exact analysis as done for the year 2016.

2.3.2 Imputation

Here as well $PM_{2.5}$ had to be dropped.

2.3.3 Changing Data Types

Same as all the years above.

2.4 2018

The exact same cleaning is performed for 2018 as for 2017.

	0
Unnamed: 0	0
State	0
City	0
Location	0
No. of\nmonitoring\ndays	0
2018 SO2 Minimum	0
2018 SO2 Maximum	0
2018 SO2 Annual Average	0
2018 NO2 Minimum	0
2018 NO2 Maximum	0
2018 NO2 Annual Average	0
2018 PM10 Minimum	0
2018 PM10 Maximum	0
2018 PM10 Annual Average	0
Is metropolitan city?	0
State Population 2011	0

dtype: int64

Figure 2.4.7: Final table for 2018 after cleaning.

2.5 2019

The exact same cleaning is performed for 2019 as for 2017.

	0
Unnamed: 0	0
State	0
City	0
Location	0
2019 SO2 Minimum	0
2019 SO2 Maximum	0
2019 SO2 Annual Average	0
No. of monitoring days	0
2019 NO2 Minimum	0
2019 NO2 Maximum	0
2019 NO2 Annual Average	0
2019 PM10 Minimum	0
2019 PM10 Maximum	0
2019 PM10 Annual Average	0
Is metropolitan city?	0
State Population 2011	0

dtype: int64

Figure 2.5.8: Final table for 2019 after cleaning.

placeins

Chapter 3. Visualization

3.1 Univariate analysis

We will perform univariate analysis of the year wise emissions of a few cities per zone i.e. North, West, East, South.

3.1.1 SO_2 emissions

3.1.1.1 Ahmedabad - West Zone

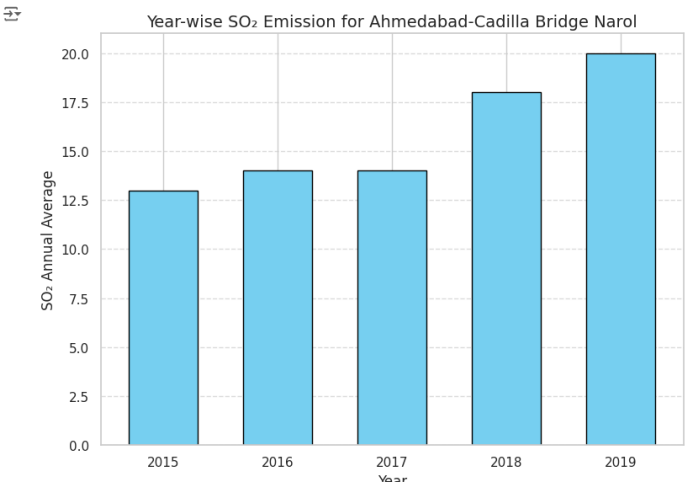

Figure 3.1.1: Bar Plot of Ahmedabad's SO_2 Emissions

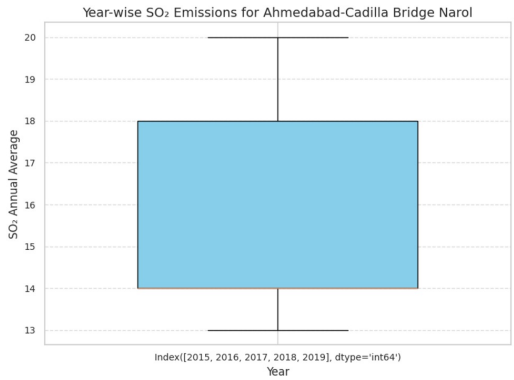

Figure 3.1.2: Box Plot of Ahmedabad's SO_2 Emissions

3.1.1.2 Badlapur - West Zone

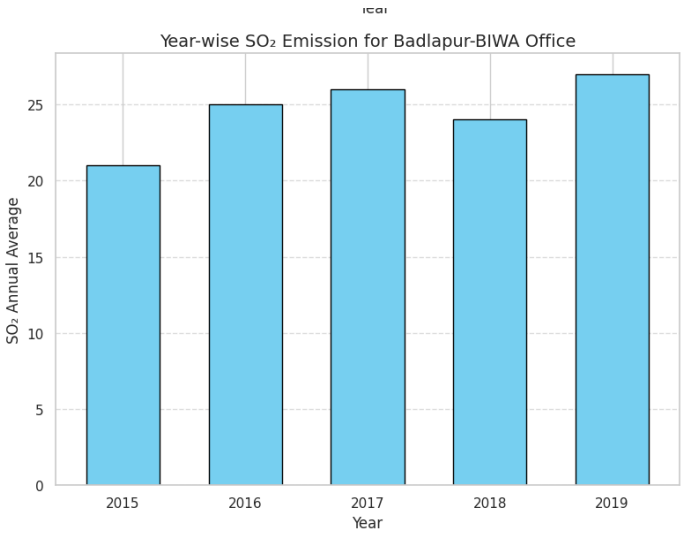

Figure 3.1.3: Bar Plot of Badlapur's SO_2 Emissions

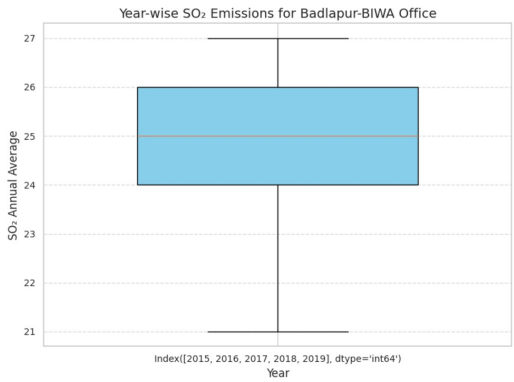

Figure 3.1.4: Box Plot of Badlapur's SO_2 Emissions

3.1.1.3 Aizawal - East Zone

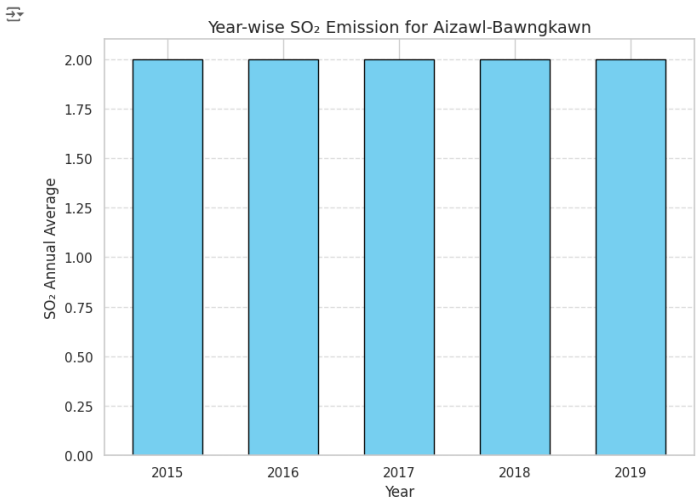

Figure 3.1.5: Bar Plot of Aizawal's SO_2 Emissions

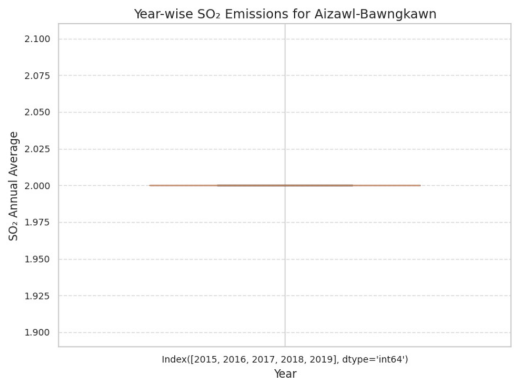

Figure 3.1.6: Box Plot of Aizawal's SO_2 Emissions

3.1.1.4 Bongaigaon - East Zone

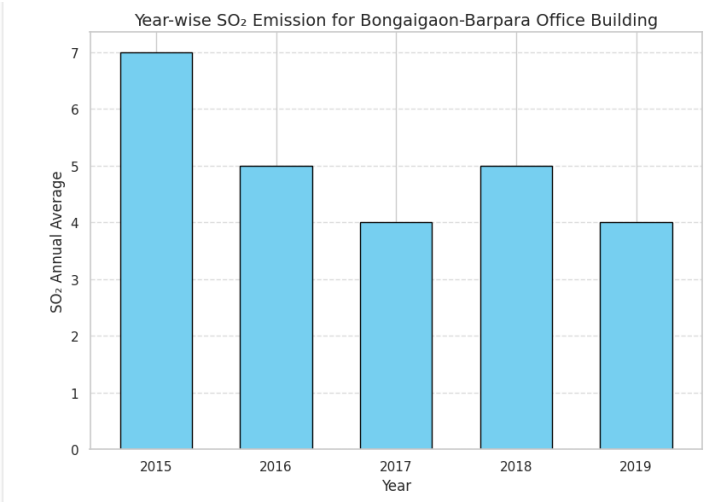

Figure 3.1.7: Bar Plot of Bongaigaon's SO_2 Emissions

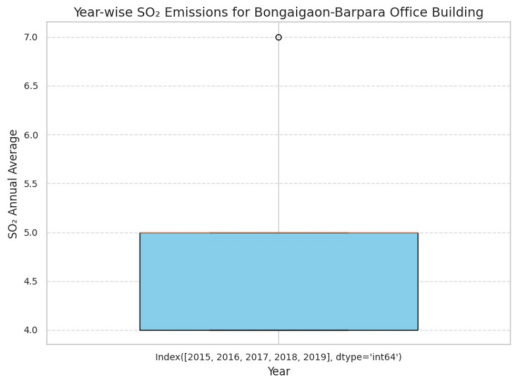

Figure 3.1.8: Box Plot of Bongaigaon's SO_2 Emissions

3.1.1.5 Bangalore - South Zone

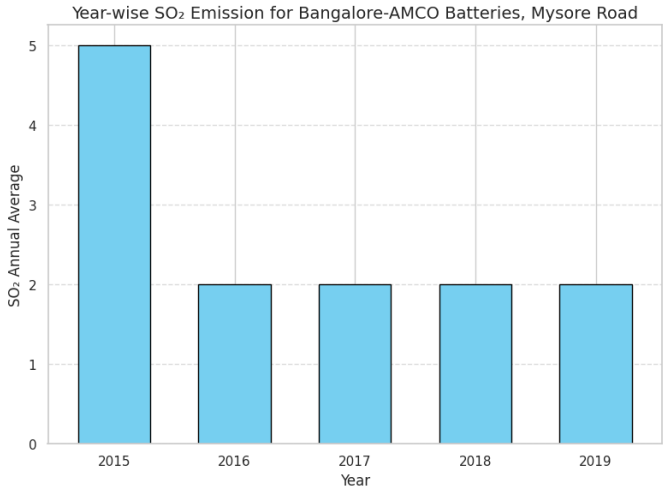

Figure 3.1.9: Bar Plot of Bangalore's SO_2 Emissions

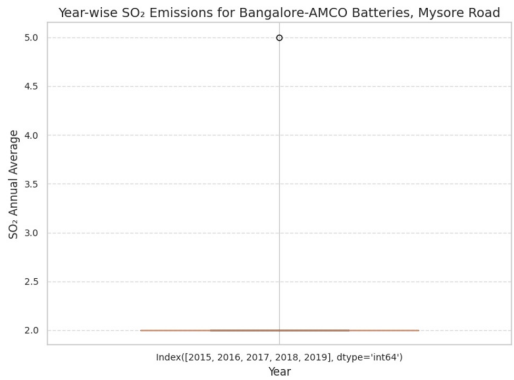

Figure 3.1.10: Box Plot of Bangalore's SO_2 Emissions

3.1.1.6 Belgaum - South Zone

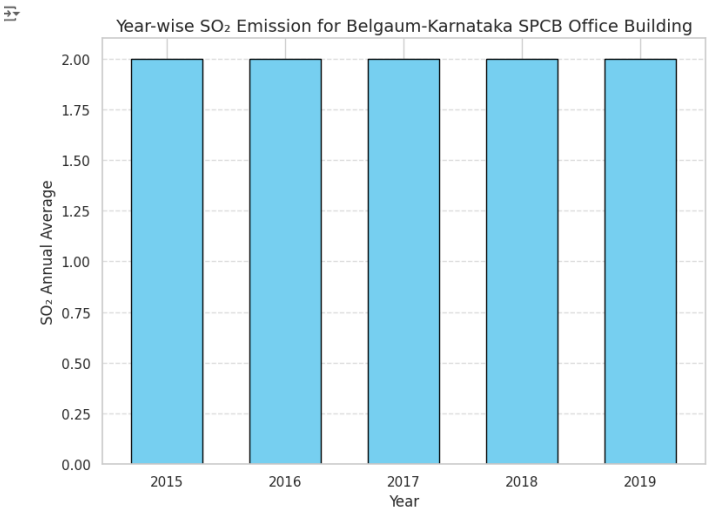

Figure 3.1.11: Bar Plot of Belgaum's SO_2 Emissions

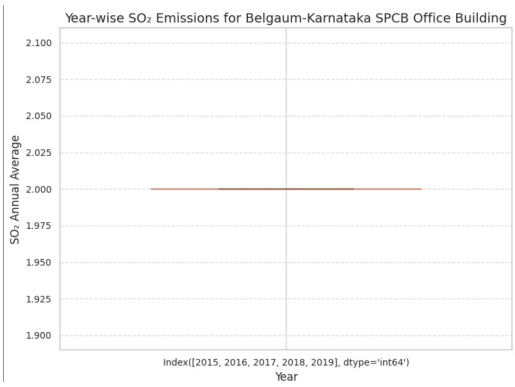

Figure 3.1.12: Box Plot of Belgaum's SO_2 Emissions

3.1.1.7 Chandigarh - North Zone

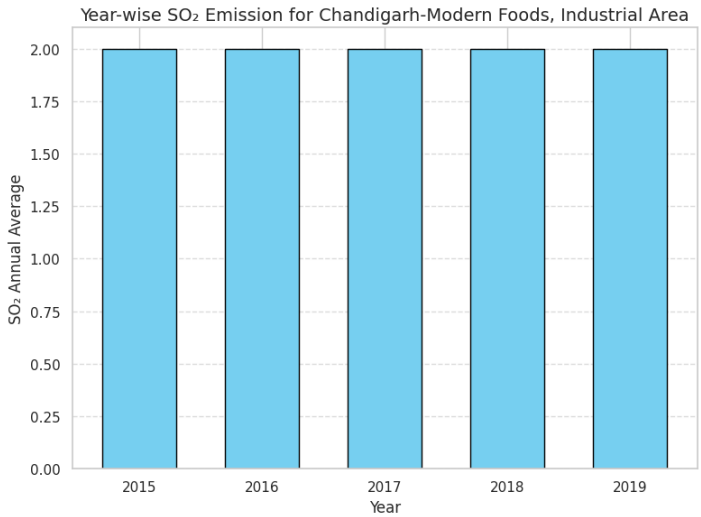

Figure 3.1.13: Bar Plot of Chandigarh's SO_2 Emissions

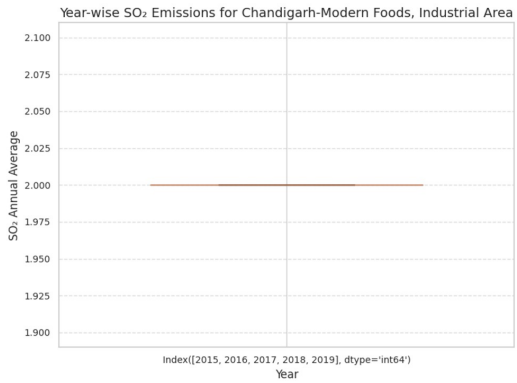

Figure 3.1.14: Box Plot of Chandigarh's SO_2 Emissions

3.1.1.8 Allahabad - North Zone

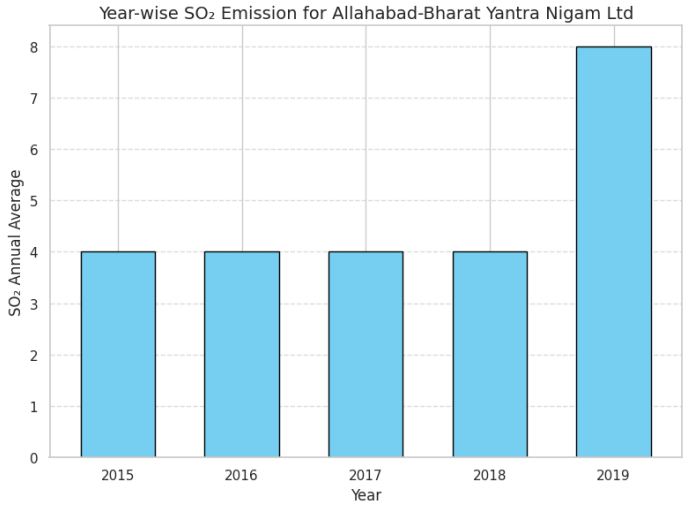

Figure 3.1.15: Bar Plot of Allahabad's SO_2 Emissions

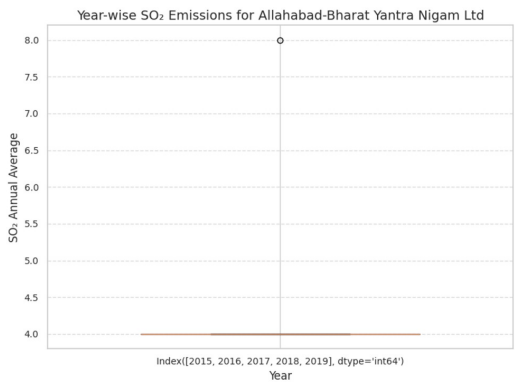

Figure 3.1.16: Box Plot of Allahabad's SO_2 Emissions

3.1.2 NO_2 emissions

3.1.2.1 Ahmedabad - West Zone

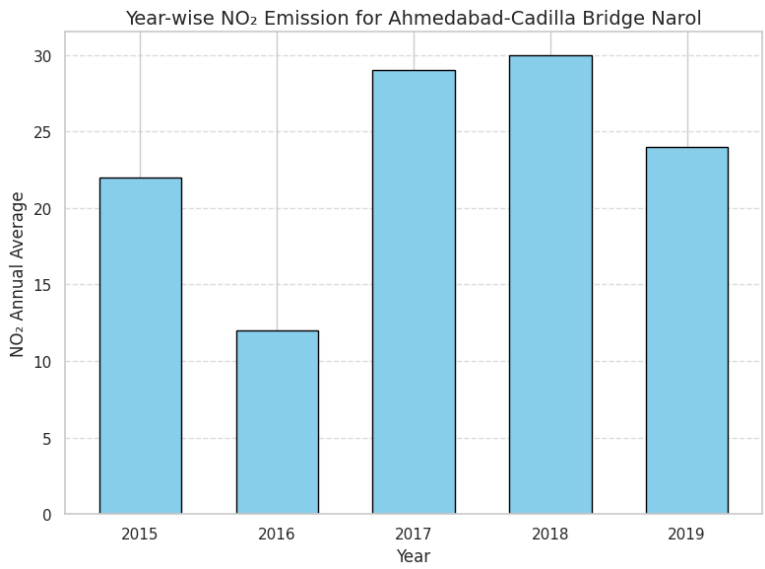

Figure 3.1.17: Bar Plot of Ahmedabad's NO_2 Emissions

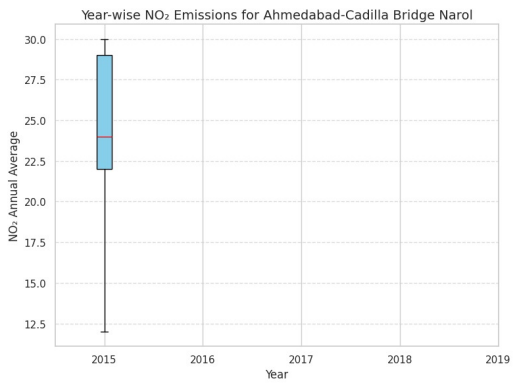

Figure 3.1.18: Box Plot of Ahmedabad's NO_2 Emissions

3.1.2.2 Badlapur - West Zone

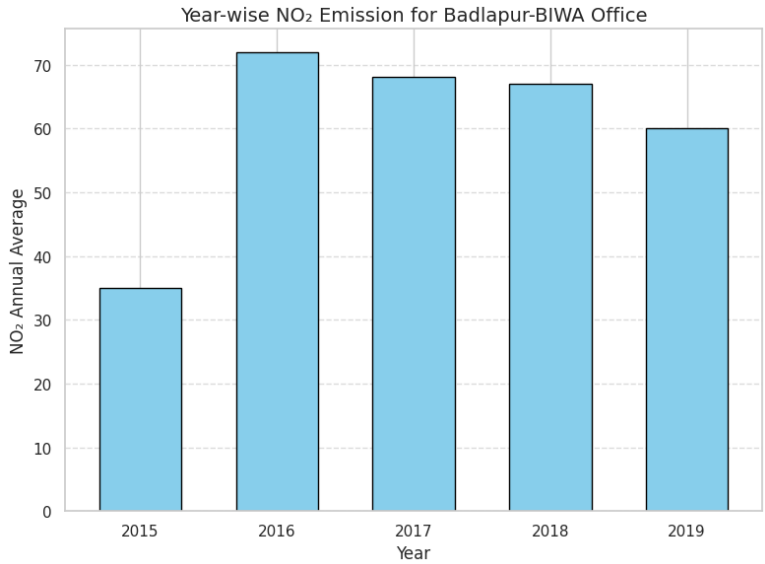

Figure 3.1.19: Bar Plot of Badlapur's NO_2 Emissions

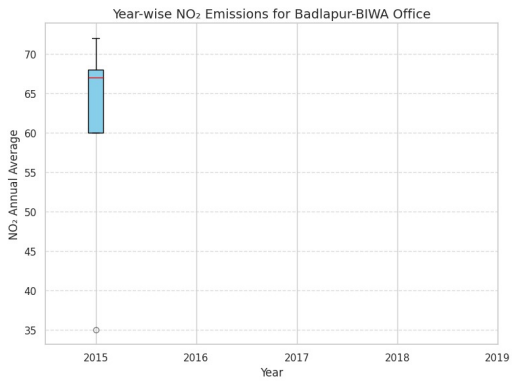

Figure 3.1.20: Box Plot of Badlapur's NO_2 Emissions

3.1.2.3 Aizawal - East Zone

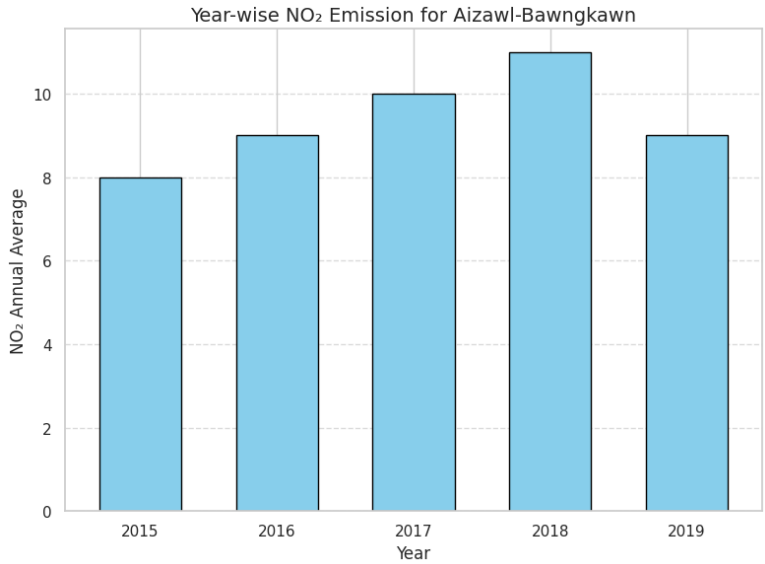

Figure 3.1.21: Bar Plot of Aizawal's NO_2 Emissions

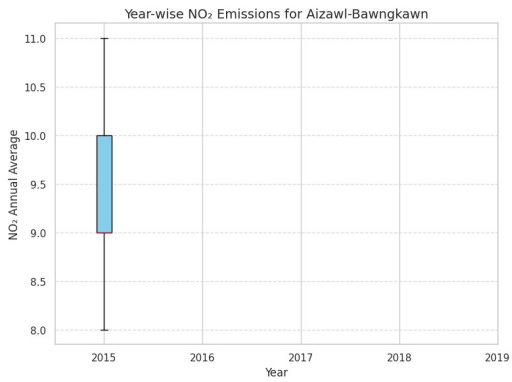

Figure 3.1.22: Box Plot of Aizawal's NO_2 Emissions

3.1.2.4 Bongaigaon - East Zone

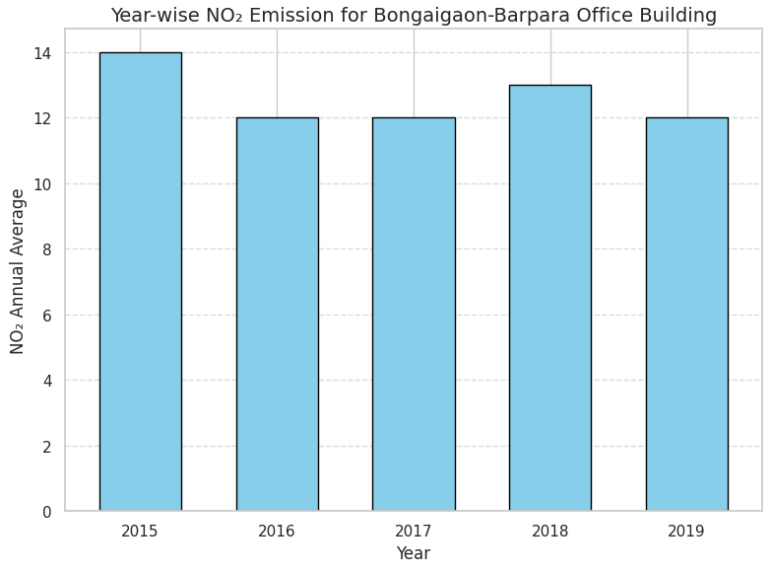

Figure 3.1.23: Bar Plot of Bongaigaon's NO_2 Emissions

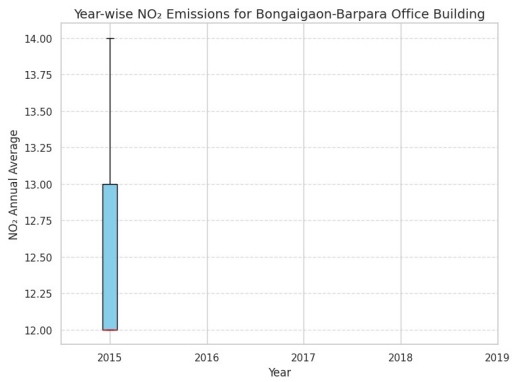

Figure 3.1.24: Box Plot of Bongaigaon's NO_2 Emissions

3.1.2.5 Bangalore - South Zone

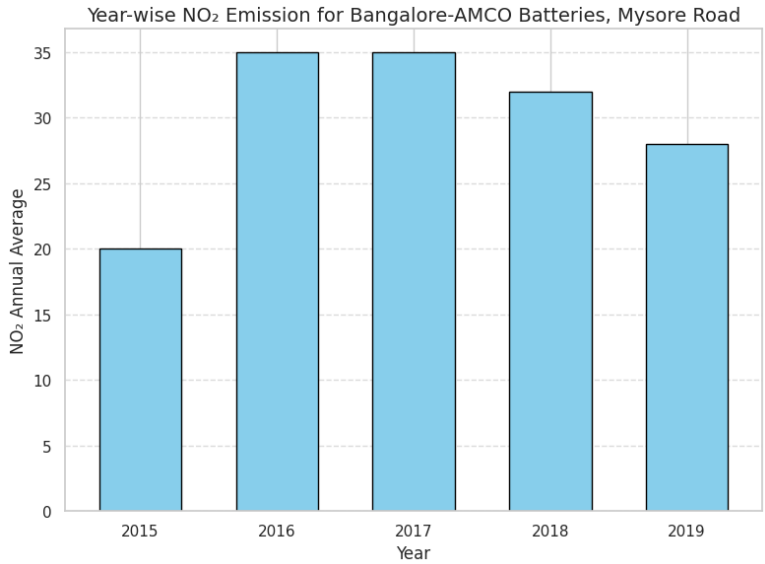

Figure 3.1.25: Bar Plot of Bangalore's NO_2 Emissions

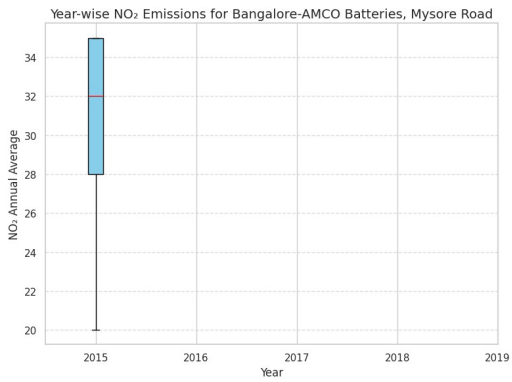

Figure 3.1.26: Box Plot of Bangalore's NO_2 Emissions

3.1.2.6 Belgaum - South Zone

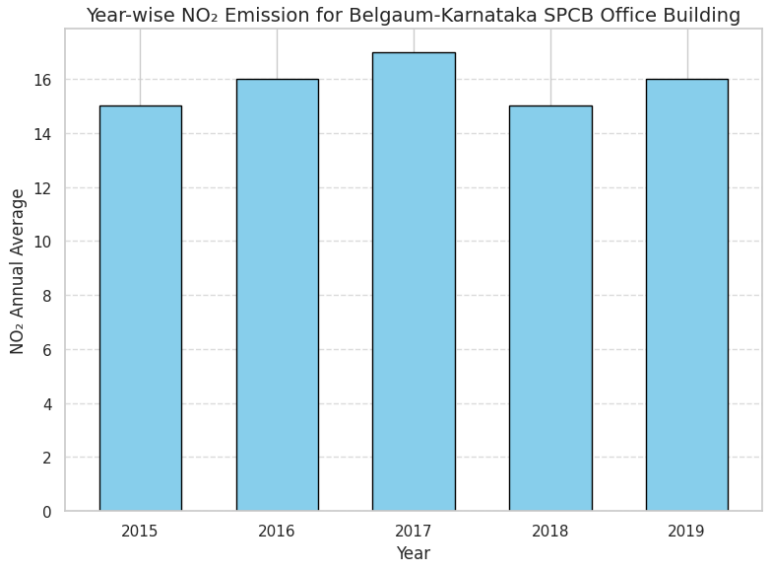

Figure 3.1.27: Bar Plot of Belgaum's NO_2 Emissions

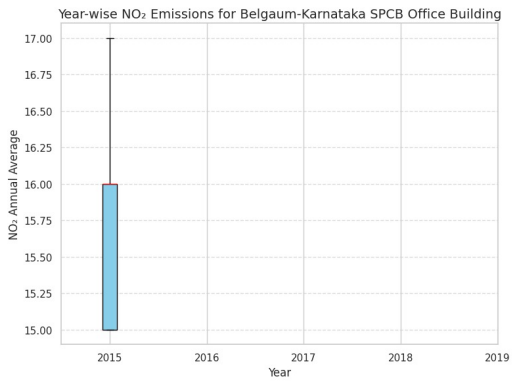

Figure 3.1.28: Box Plot of Belgaum's NO_2 Emissions

3.1.2.7 Chandigarh - North Zone

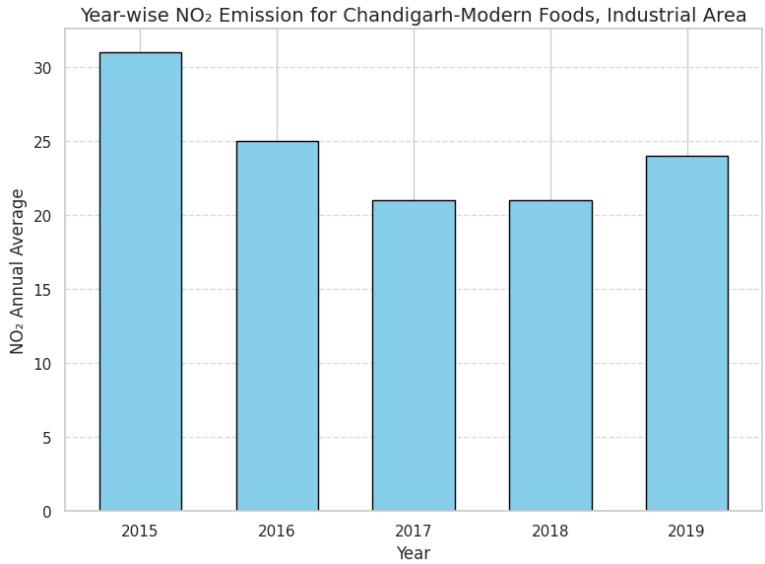

Figure 3.1.29: Bar Plot of Chandigarh's NO_2 Emissions

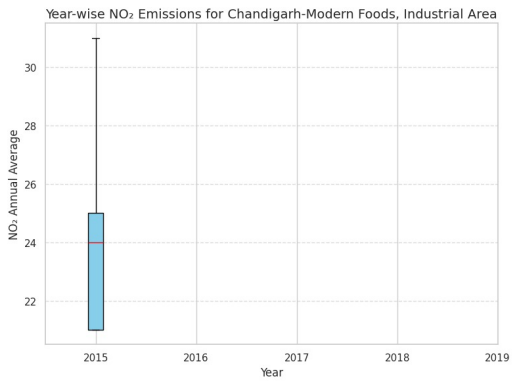

Figure 3.1.30: Box Plot of Chandigarh's NO_2 Emissions

3.1.2.8 Allahabad - North Zone

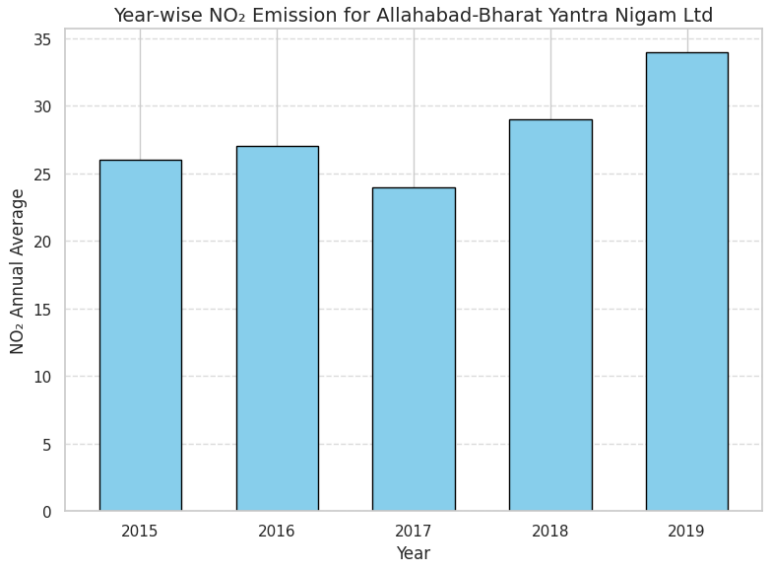

Figure 3.1.31: Bar Plot of Allahabad's NO_2 Emissions

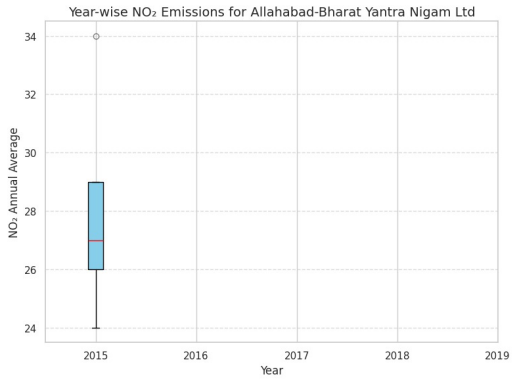

Figure 3.1.32: Box Plot of Allahabad's NO_2 Emissions

3.1.3 PM_{10} emissions

3.1.3.1 Ahmedabad - West Zone

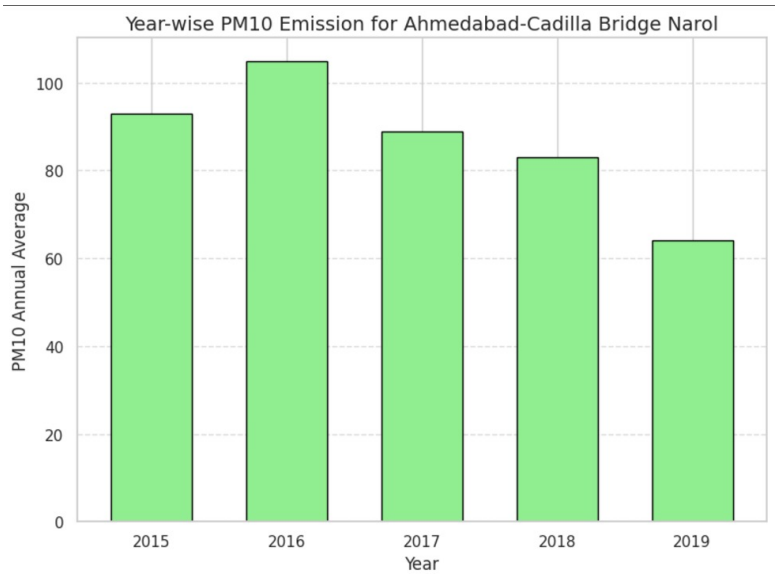

Figure 3.1.33: Bar Plot of Ahmedabad's PM_{10} Emissions

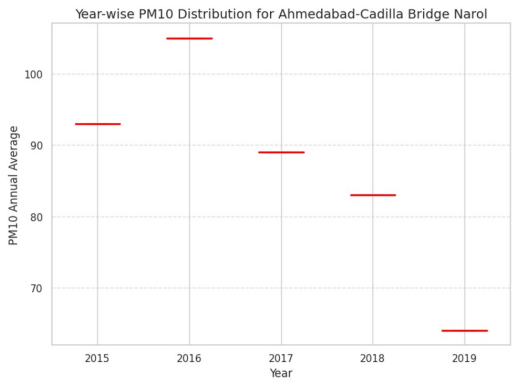

Figure 3.1.34: Box Plot of Ahmedabad's PM_{10} Emissions

3.1.3.2 Badlapur - West Zone

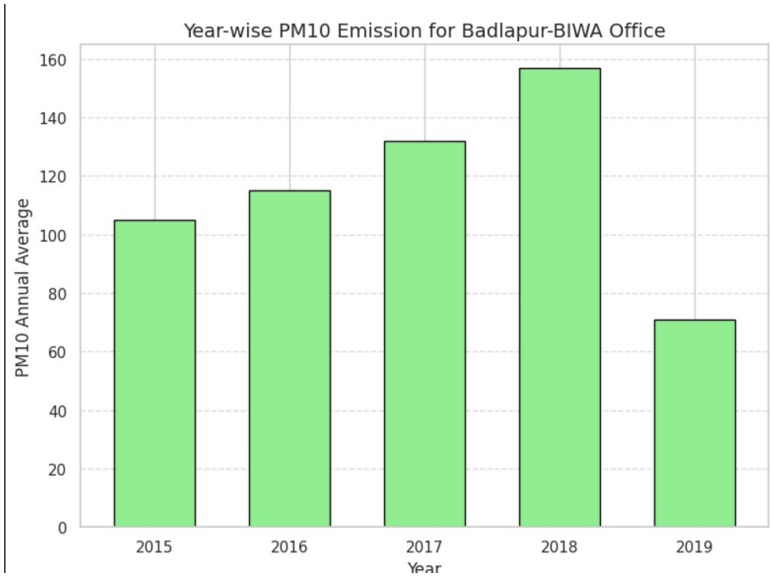

Figure 3.1.35: Bar Plot of Badlapur's PM_{10} Emissions

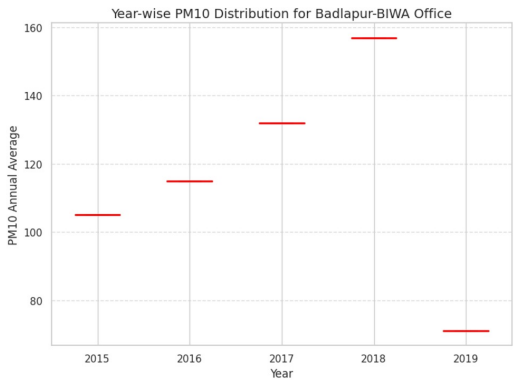

Figure 3.1.36: Box Plot of Badlapur's PM_{10} Emissions

3.1.3.3 Aizawal - East Zone

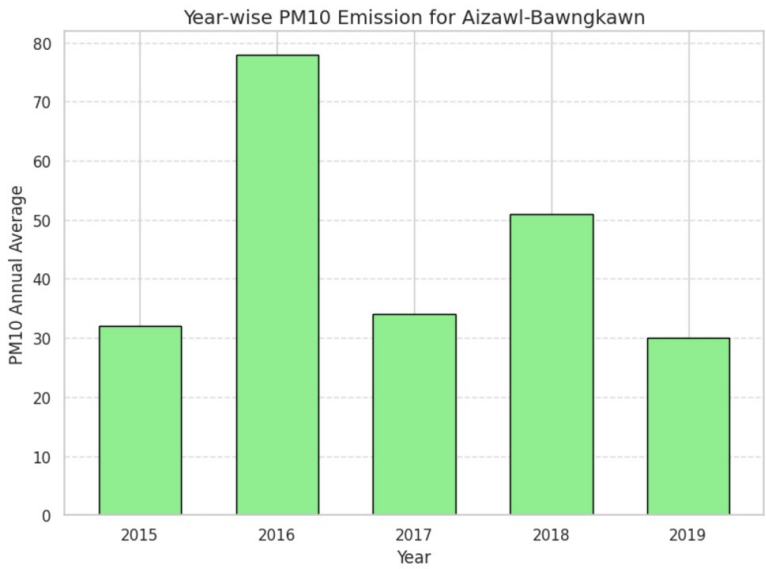

Figure 3.1.37: Bar Plot of Aizawal's PM_{10} Emissions

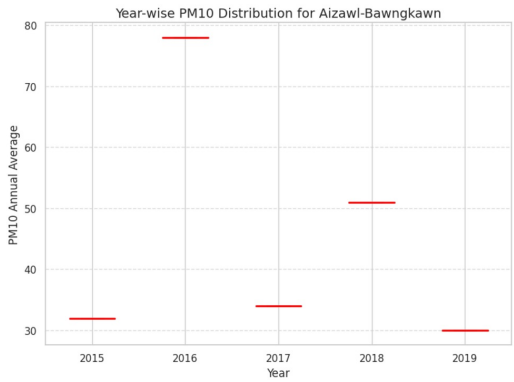

Figure 3.1.38: Box Plot of Aizawal's PM_{10} Emissions

3.1.3.4 Bongaigaon - East Zone

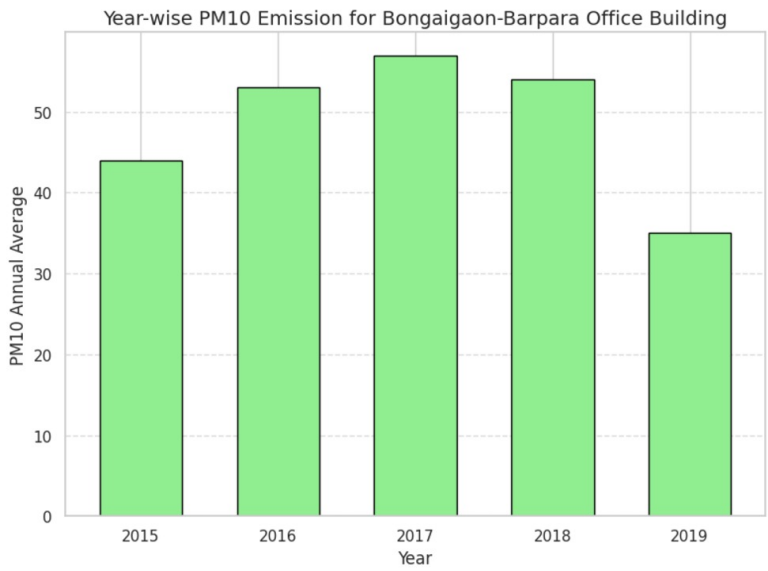

Figure 3.1.39: Bar Plot of Bongaigaon's PM_{10} Emissions

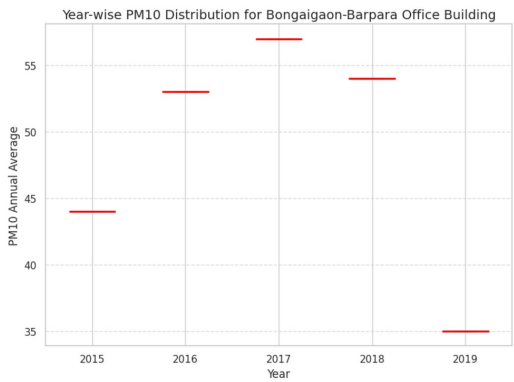

Figure 3.1.40: Box Plot of Bongaigaon's PM_{10} Emissions

3.1.3.5 Bangalore - South Zone

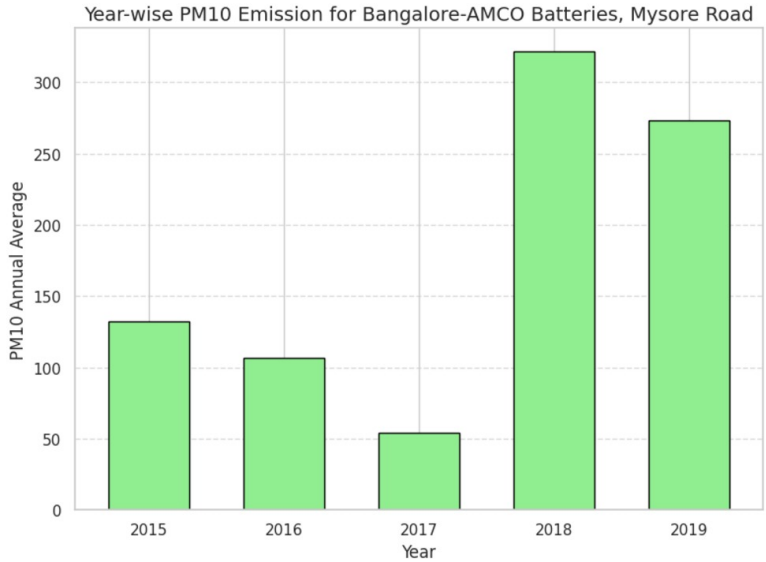

Figure 3.1.41: Bar Plot of Bangalore's PM_{10} Emissions

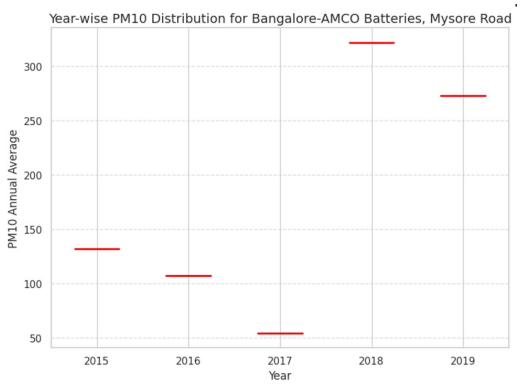

Figure 3.1.42: Box Plot of Bangalore's PM_{10} Emissions

3.1.3.6 Belgaum - South Zone

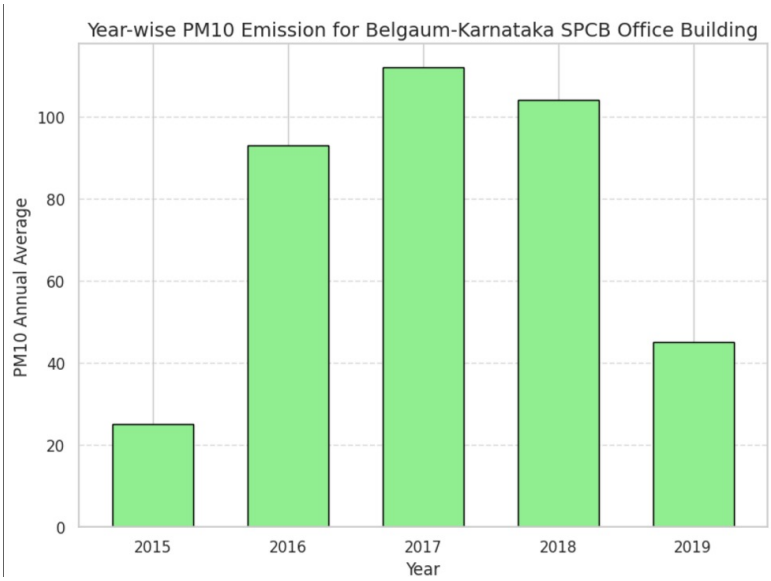

Figure 3.1.43: Bar Plot of Belgaum's PM_{10} Emissions

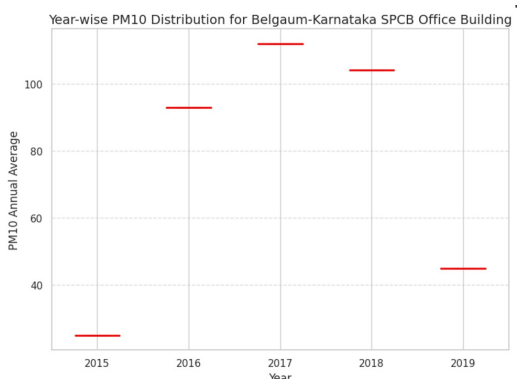

Figure 3.1.44: Box Plot of Belgaum's PM_{10} Emissions

3.1.3.7 Chandigarh - North Zone

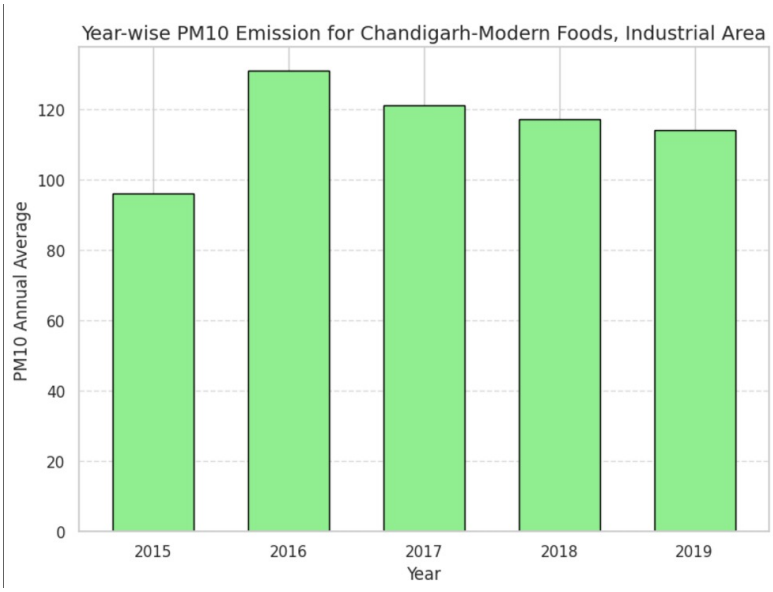

Figure 3.1.45: Bar Plot of Chandigarh's PM_{10} Emissions

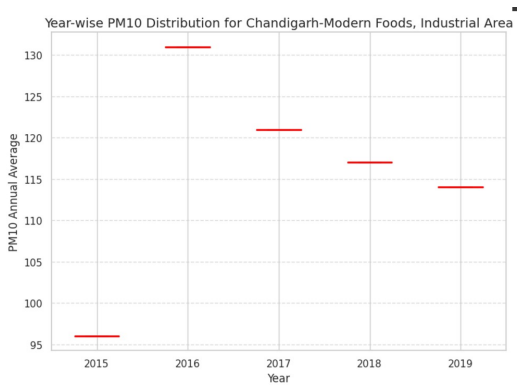

Figure 3.1.46: Box Plot of Chandigarh's PM_{10} Emissions

3.1.3.8 Allahabad - North Zone

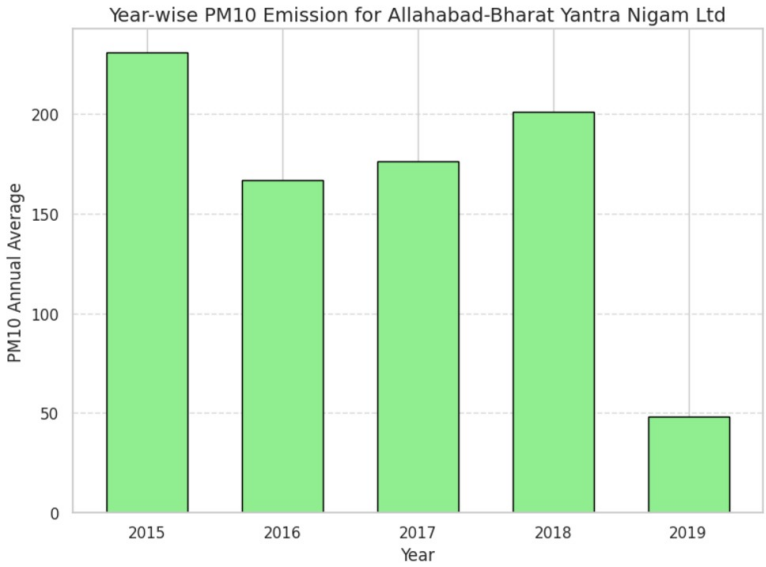

Figure 3.1.47: Bar Plot of Allahabad's PM_{10} Emissions

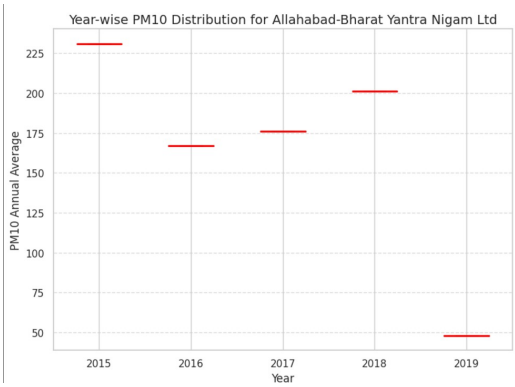

Figure 3.1.48: Box Plot of Allahabad's PM_{10} Emissions

3.2 Multivariate analysis

We will perform multivariate analysis of the year wise emissions of a few cities per zone i.e. North, West, East, South.

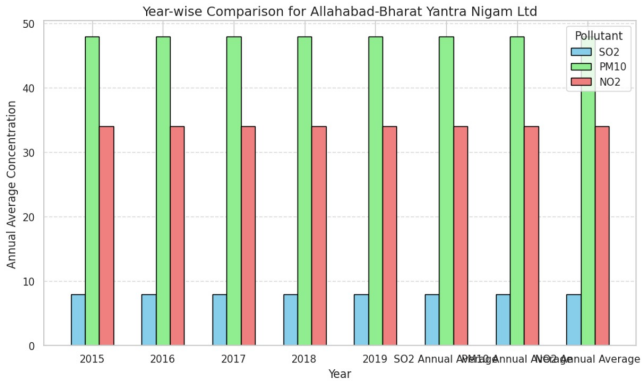

Figure 3.2.50: Bar Plot of Allahabad's SO_2, PM_10, NO_2 Emissions

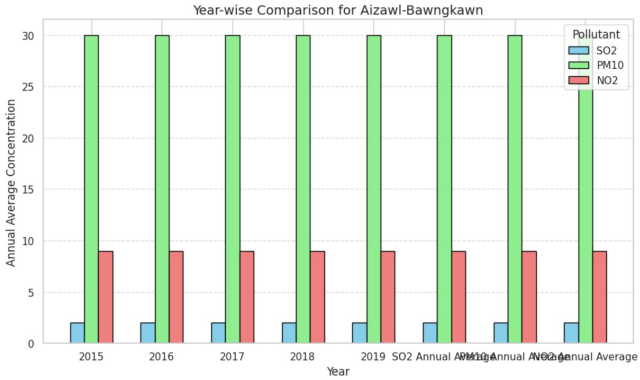

Figure 3.2.51: Bar Plot of Aizwal's SO_2, PM_10, NO_2 Emissions

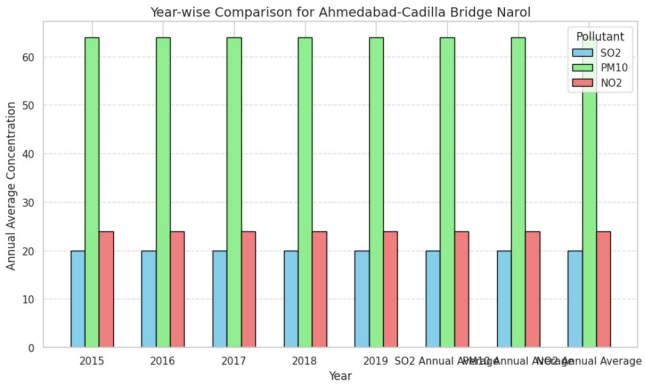

Figure 3.2.49: Bar Plot of Ahmedabad's SO_2, PM_{10}, NO_2 Emissions

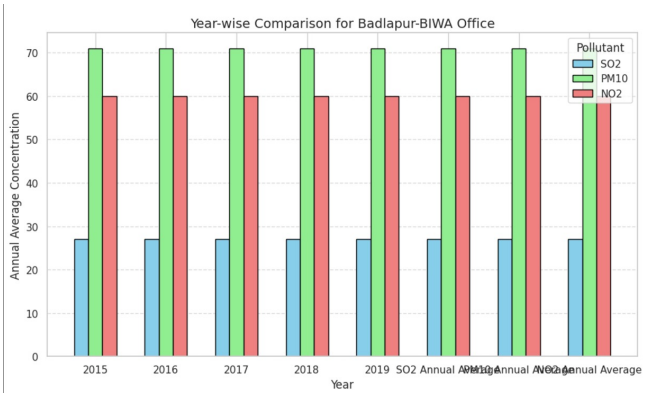

Figure 3.2.52: Bar Plot of Badlapur's SO_2, PM_10, NO_2 Emissions

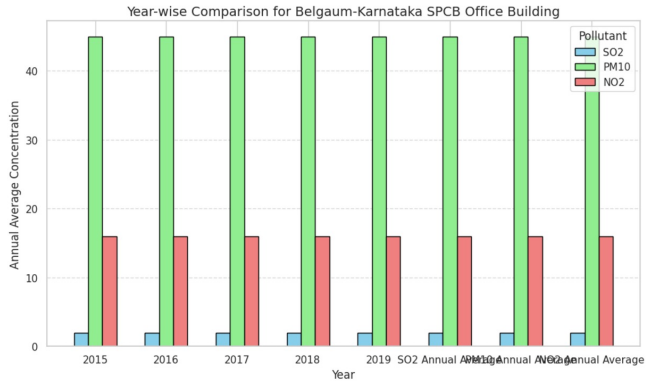

Figure 3.2.53: Bar Plot of Belgaum's SO_2, PM_10, NO_2 Emissions

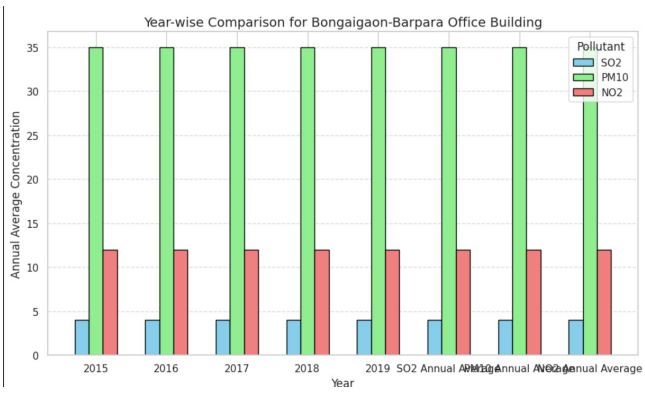

Figure 3.2.54: Bar Plot of Bongaigaon's SO_2, PM_10, NO_2 Emissions

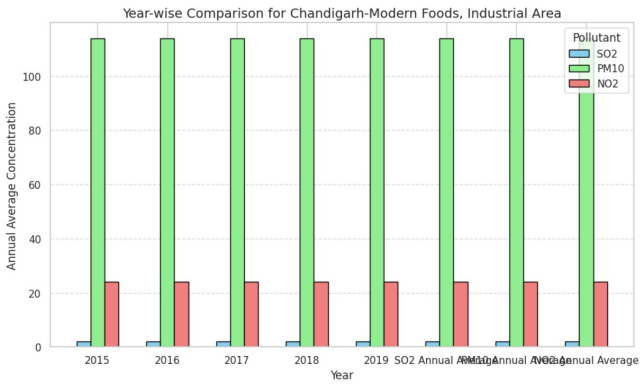

Figure 3.2.55: Bar Plot of Chandigarh's SO_2, PM_10, NO_2 Emissions

Chapter 4. Feature Engineering

4.1 Feature extraction

For feature engineering we first look at the heatmap for correaltion. Following the heatmaps for all 5 years:

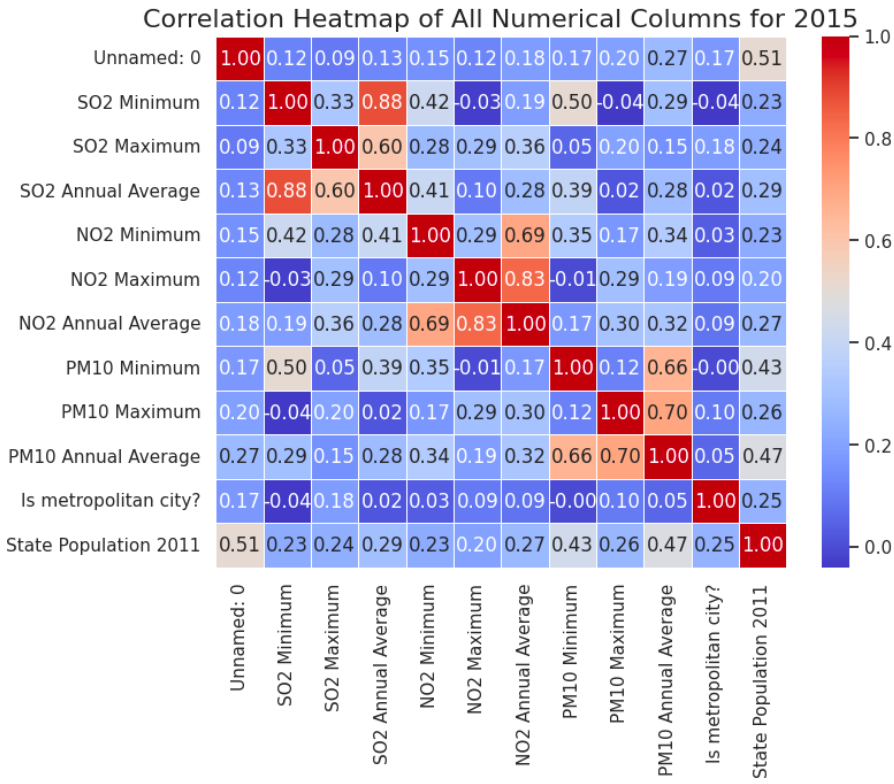

Figure 4.1.1: 2015

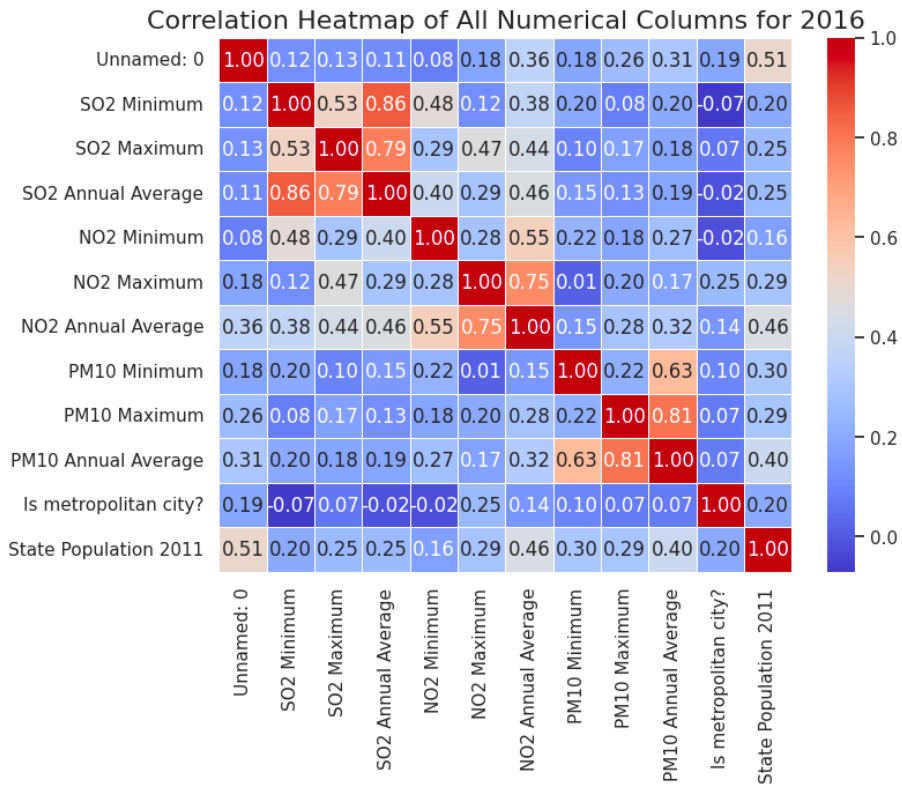

Figure 4.1.2: 2016

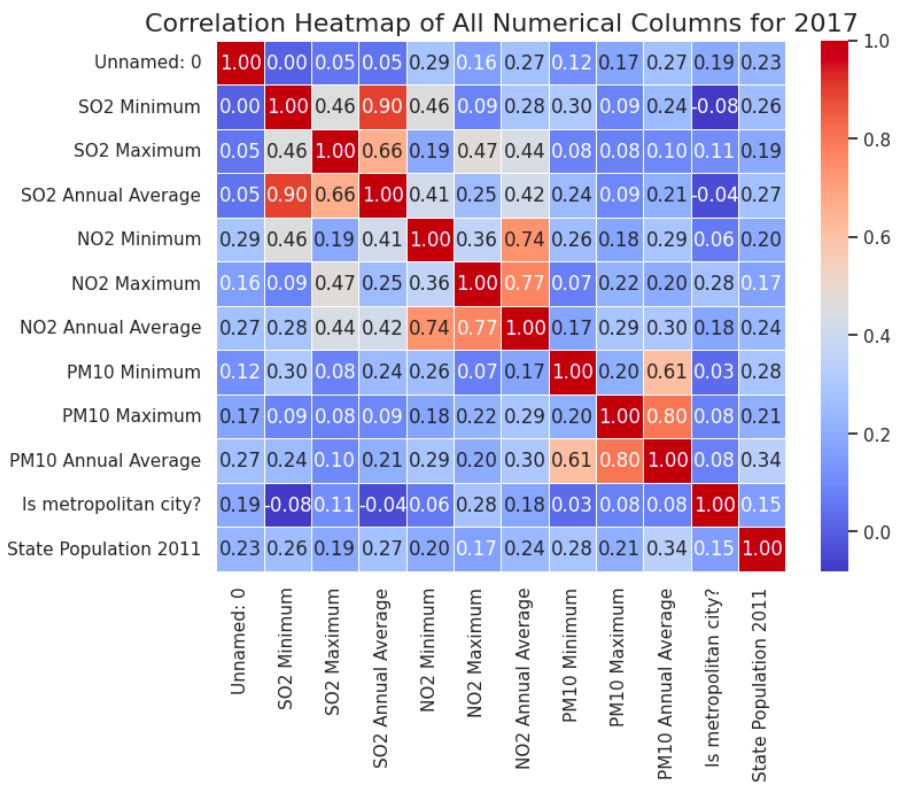

Figure 4.1.3: 2017

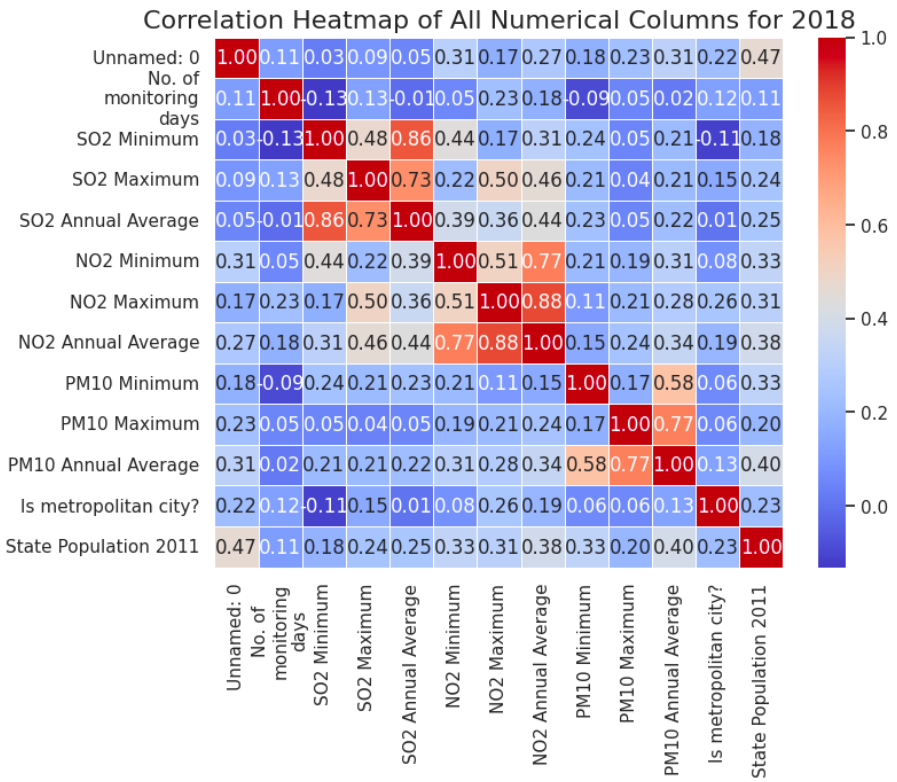

Figure 4.1.4: 2018

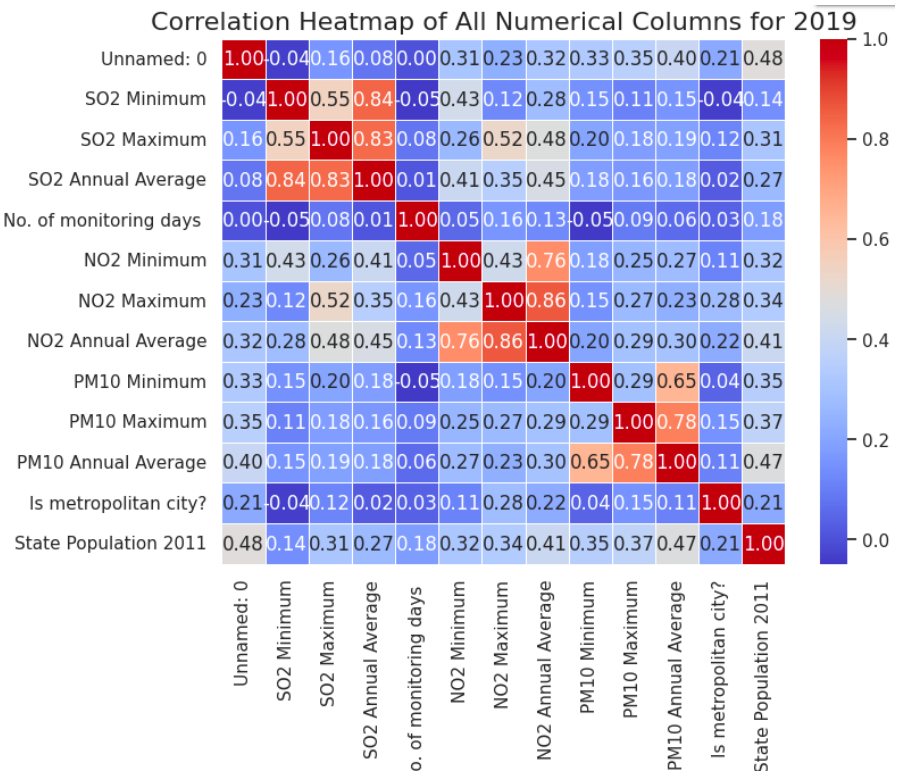

Figure 4.1.5: 2019

4.2 Feature selection

1. State Population 2011: Correlation with SO2 Annual Average: Moderate positive correlation (0.29). This suggests that states with higher populations tend to have slightly higher SO2 levels, likely due to increased industrial and vehicular activities.
Correlation with NO2 Annual Average: Moderate correlation (0.43). This indicates that NO2 levels are influenced by population density, which aligns with urbanization trends and emissions from transportation. Correlation with PM10 Annual Average: Moderate positive correlation (0.47). This implies that particulate pollution is impacted by population, likely from dust, industrial activity, and residential emissions.

2. SO2 Annual Average: Correlation with SO2-related metrics: Strongly correlated with SO2 Maximum (0.88) and SO2 Minimum (0.83). These strong relationships indicate that SO2 Annual Average is a comprehensive summary of sulfur dioxide levels and captures seasonal trends effectively.
Correlation with NO2 Annual Average: Moderate positive correlation (0.39). This reflects a shared emission source (e.g., combustion processes in vehicles and industries).
Correlation with PM10 Annual Average: Moderate positive correlation (0.29). This suggests a link between SO2 levels and particulate matter, possibly due to secondary particle formation from SO2.

3. NO2 Annual Average: Correlation with NO2-related metrics: Strongly correlated with NO2 Maximum (0.83) and NO2 Minimum (0.69). This makes NO2 Annual Average a reliable indicator of nitrogen dioxide trends over time.
Correlation with PM10 Annual Average: Strong positive correlation (0.66). This indicates a strong interplay between NO2 and PM10, likely arising from vehicular and industrial emissions.
Correlation with SO2 Annual Average: Moderate correlation (0.39), highlighting interdependencies between gaseous pollutants.

4. PM10 Annual Average: Correlation with PM10-related metrics: Strong correlation with PM10 Maximum (0.70) and PM10 Minimum (0.66), showing that PM10 Annual Average captures particulate pollution comprehensively.
Correlation with NO2 Annual Average: Strong positive correlation (0.66), indicating that PM10 levels often rise alongside NO2, both being markers of urban pollution.
Correlation with SO2 Annual Average: Moderate positive correlation (0.29), suggesting some link between particulate and gaseous pollutants.

Therefore we chose annual SO2, annual NO2, annual PM10 for our model prediction.

4.3 Outlier Detection

There were no outliers in our dataset and so outlier removal was not required. We move on to model fitting next to predict the emissions and air quality for 2020 and 2021.

Chapter 5. Model fitting

Since we are predicting the air quality of the year 2020 based on historical data, we use Linear Regression model for the same.

5.1 Regression

5.1.1 Ahmedabad

In training and testing the model we applied the following approach.

5.1.1.0.1 Dropping Unwanted Rows We dropped the rows with NaN values.

```
[84] columns_to_drop = df_new.select_dtypes(include=['float64']).columns
     df_new = df_new.drop(columns=columns_to_drop)

     # Display the resulting DataFrame
     print("Columns dropped:", columns_to_drop)
     print("Updated DataFrame:")
     print(df_new)
```

Figure 5.1.1: Dropping rows for model training

5.1.1.0.2 Defining X and y and splitting the data We defined the features i.e. the independent variables and the dependent variables. We then split the data to train the data from 2015-19 and test on actual data from 2020.

```
[85] # Add a feature for modeling
     X = df_new[['Year']]  # Independent variable (Year)
     y = df_new['Annual SO2 emission']  # Dependent variable (SO2 Emission)

     # Train on 2015-2019
     X_train = X[:-1]  # All rows except the last (2020)
     y_train = y[:-1]  # All rows except the last (2020)

     # Test on 2020
     X_test = X[-1:]  # The last row (2020)
     y_test = y[-1:]  # The last row (2020)
```

Figure 5.1.2: Train - test split

```
[87] from sklearn.linear_model import LinearRegression
     from sklearn.metrics import mean_squared_error, r2_score

     # Train the model
     model = LinearRegression()
     model.fit(X_train, y_train)

     # Predict for 2020
     y_pred = model.predict(X_test)

     print(f"Predicted SO2 Emission for 2020: {y_pred[0]:.2f}")
     print(f"Actual SO2 Emission for 2020: {y_test.values[0]:.2f}")

     Predicted SO2 Emission for 2020: 21.20
     Actual SO2 Emission for 2020: 14.00
```

Figure 5.1.3: Results

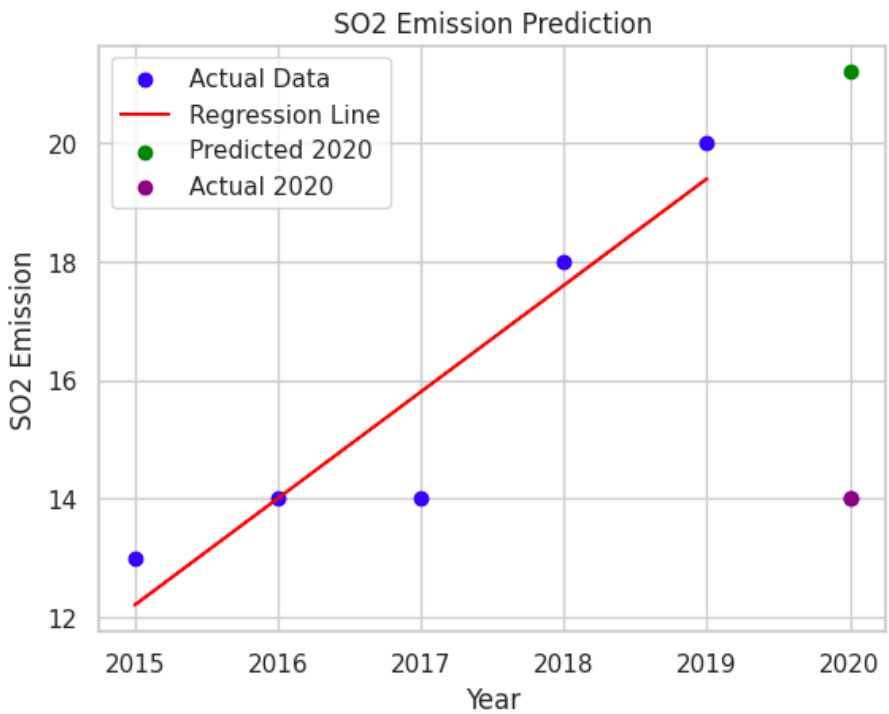

Figure 5.1.4: Comparative Results

5.1.1.0.3 Model output and comparison It can be seen from our observations that the predicted value is way higher than the actual value of SO_2 emissions in 2020 and this is because of the covid - 19 Pandemic. Due to the lockdown all the sources of pollutants had lessened and so this unexpected result is seen.

5.1.2 More results

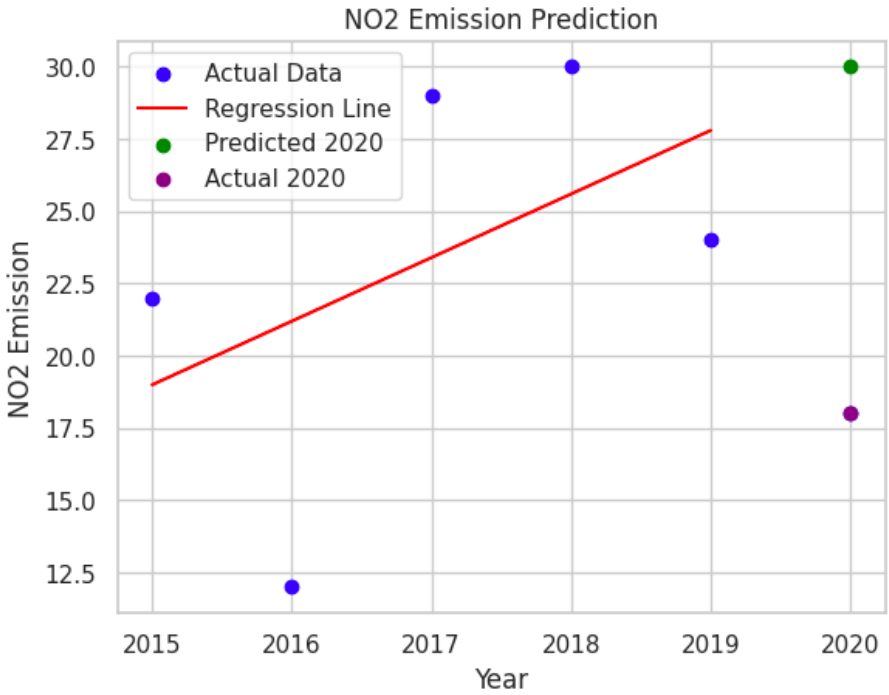

Figure 5.1.5: NO2 prediction for Ahmedabad

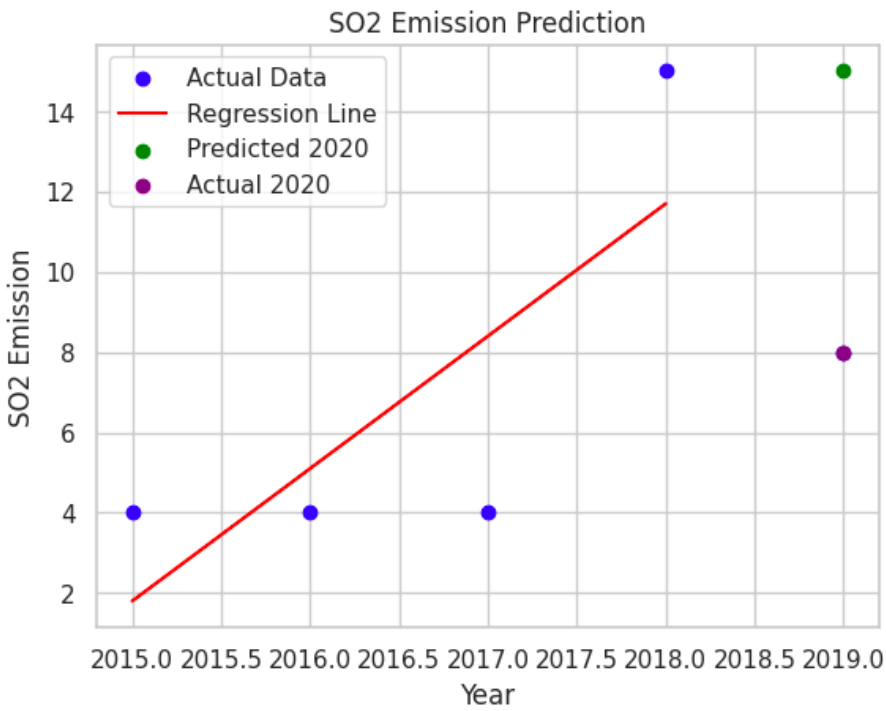

Figure 5.1.6: SO2 prediction for Allahabad

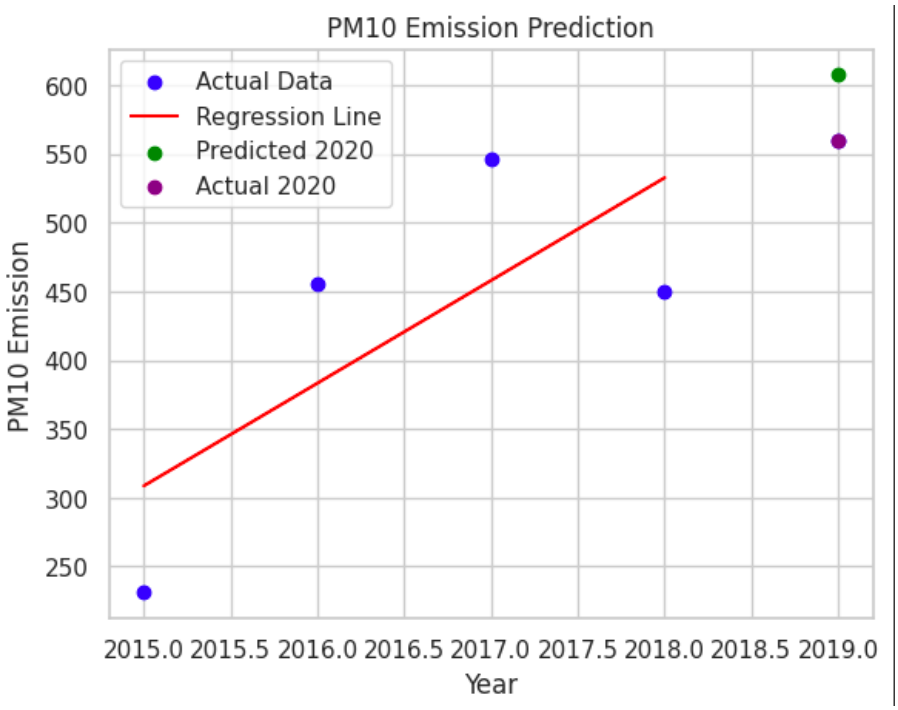

Figure 5.1.7: PM10 prediction for Allahabad

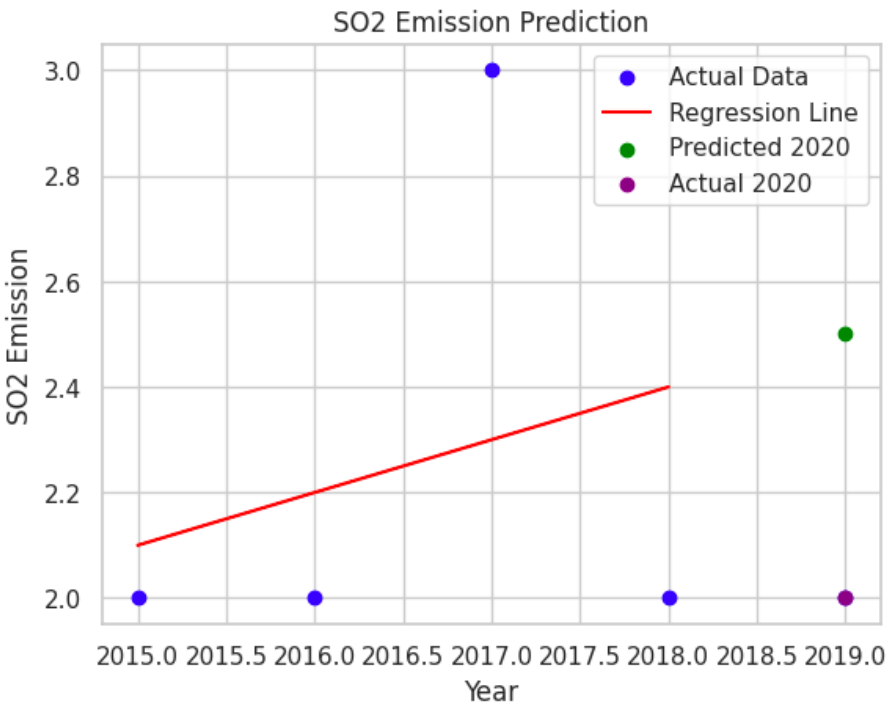

Figure 5.1.8: SO2 prediction for Aizawal

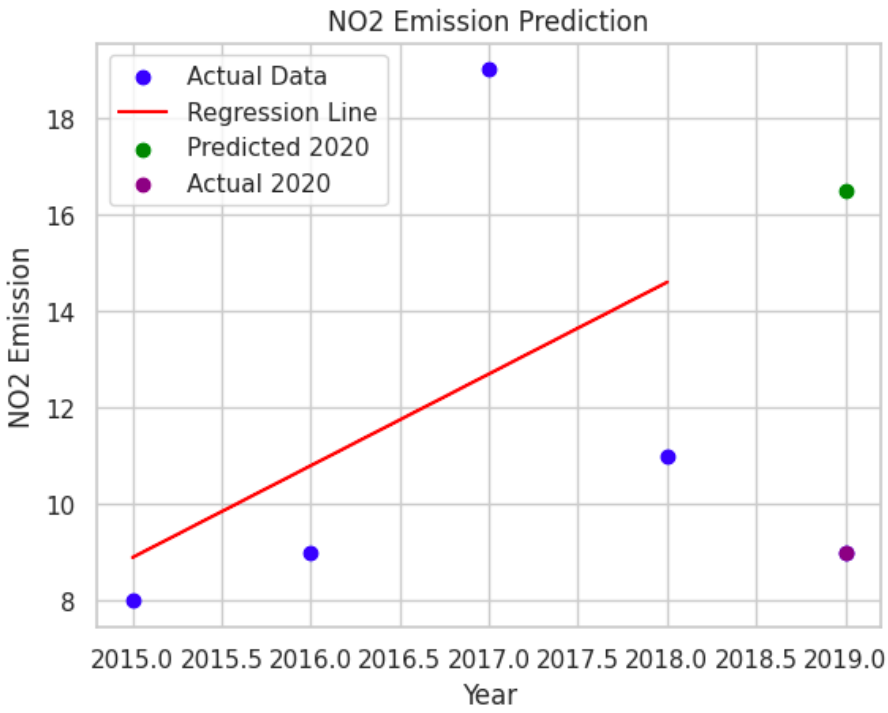

Figure 5.1.9: NO2 prediction for Aizawal

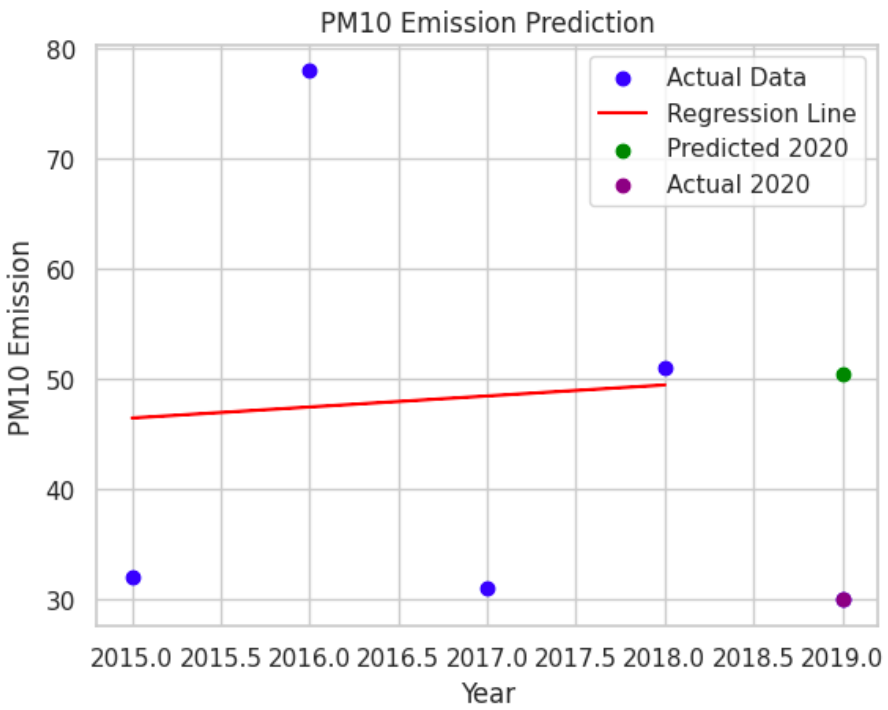

Figure 5.1.10: PM10 prediction for Aizawal

Chapter 6. Conclusion & future scope

6.1 Conclusion

In this study, we analyzed year-wise emissions of SO_2, NO_2, PM_{10} in Indian cities from 2015 to 2019 and used linear regression models to predict pollutant levels for 2020. While the models demonstrated a strong ability to identify trends and forecast emissions, the predictions for 2020 were higher than the actual observed values. This discrepancy can be attributed to the COVID-19 pandemic, which caused unprecedented reductions in industrial activity, vehicular emissions, and other pollution sources during the lockdown periods.

The results highlight the impact of unforeseen global events on air quality and emphasize the importance of integrating flexibility into predictive models to account for such anomalies. Despite these challenges, the models provided valuable insights into long-term trends and pollution patterns, underscoring their utility for environmental planning and management.

6.2 Findings/observations

The cities that are metropolitan have higher pollution.

The cities with higher population have higher pollution.

Air pollution has a trend of increasing every year but the pandemic lowered it a little.

All predictions made before the pandemic need to be adjusted later on.

6.3 Challenges

This study faced several challenges in predicting air quality and analyzing year-wise emissions of SO_2, NO_2, PM_{10}:

Impact of Anomalies (COVID-19): The COVID-19 pandemic introduced an unprecedented anomaly in pollution trends. The reduced industrial and vehicular activities during lockdowns caused actual pollutant levels to be significantly lower than predicted. Models trained on historical data could not account for such unexpected global events, leading to overpredictions for 2020.

Data Gaps and Inconsistencies: Some cities and locations were not consistently represented in all datasets from 2015 to 2019. This limited the analysis to cities with complete data, reducing the scope of the study.

Feature Limitations: The datasets included only emissions data for specific pollutants and lacked additional contextual variables such as meteorological data, industrial output, traffic density, and population growth, which are known to influence air quality.

Simplistic Modeling Approach: While linear regression provided insights into trends, it assumes a linear relationship between time and pollution levels, which may not capture the complexities of real-world emissions influenced by non-linear and interdependent factors.

Data Quality and Measurement Errors: Emission data are subject to inconsistencies due to differences in measurement techniques, monitoring station placements, and reporting standards, which may introduce noise into the analysis.

Limited Temporal Range: The dataset spanned only five years (2015–2019), restricting the model's ability to capture long-term trends or cyclic patterns that might better inform predictions.

Inter-Pollutant Relationships: The analysis treated each pollutant independently, overlooking potential interactions and correlations between pollutants that could influence predictions.

6.4 Future plan

1. Integration of Anomalies in Modeling:
Incorporate anomaly detection methods to account for events like pandemics, natural disasters, or policy shifts that can significantly alter pollution levels. Use hybrid models that combine historical trends with real-time data for more accurate predictions.
2. Temporal and Event-Specific Analysis:
Conduct detailed studies of the COVID-19 lockdown's impact on emissions across different cities and pollutants. Compare pollution levels during lockdowns with pre- and post-lockdown phases to quantify the effect of reduced human activities.
3. Advanced Time-Series Techniques:
Utilize advanced time-series models (e.g., ARIMA, Prophet, or LSTM) that can adapt to disruptions in historical trends. Incorporate external variables, such as lockdown durations and mobility data, into the prediction models.
4. Scenario-Based Predictions:
Develop simulation models to predict emissions under various scenarios, such as prolonged industrial shutdowns or the introduction of greener technologies. Evaluate the potential long-term impact of sustained behavioral and policy changes initiated during COVID-19.
5. Regional and Pollutant-Specific Studies:
Expand the analysis to focus on regional variations in the pandemic's impact, identifying areas with the most significant improvements in air quality. Study interactions between pollutants to understand how reductions in one may affect others.
6. Policy Evaluation and Development:
Use the models to assess the effectiveness of pollution control measures during COVID-19, such as reduced traffic and industrial slowdowns. Advocate for sustainable urban planning and policies that replicate these reductions without requiring disruptions.
7. Broader Data Integration:
Integrate mobility, industrial production, and energy consumption data with pollution data to build more robust models. Use satellite imagery and remote sensing data to supplement ground-level monitoring.

8. Health and Climate Correlations:

Explore the health benefits of reduced pollution during the COVID-19 period to highlight the importance of maintaining low emission levels. Study how temporary improvements in air quality affected climate-related factors like temperature and precipitation.

Group Contribution

Gunesha Shahu, 202201022

Collecting data, cleaning data, creating visualization.
Model fitting, Model training.
Writing Latex Document.

Lakshya Jain, 202312106

Data Collection.
Data visualization.
Writing Latex Document.

Megha Maheshwari, 202312118

Data Collection, data cleaning, visualization.
Model fitting.
Writing Latex Document.

Short Bio

1. **Gunesha Shahu** is a third-year B.Tech student with a keen interest in data science and web development. She enjoys exploring new opportunities and expanding her knowledge in the tech field. Outside of academics, Gunesha has a passion for traveling, building connections, and immersing herself in music. Her dynamic personality reflects a balance between technical expertise and a love for engaging with the world around her.
2. **Megha Maheshwari** is a second-year MSc IT student, passionate about software development with expertise in web development and data structures and algorithms (DSA). Offers strong interpersonal and task prioritization skills.She combines her technical skills with a creative side, enjoying reading, traveling, music, and painting during her leisure time. Megha is dedicated to expanding her knowledge and making meaningful contributions to the tech world.
3. **Lakshya Jain** is a driven professional with a strong focus on personal and professional development. Currently pursuing an MSc in IT, Lakshya thrives on solving challenges and working on impactful projects. With a deep appreciation for teamwork, innovation, and perseverance, Lakshya is dedicated to making meaningful contributions while continuously growing through every opportunity encountered.

References

[1] CPCB *URL:* https://cpcb.nic.in/manual-monitoring/

[2] Government data website. *URL:* data.gov.in