

Analyzing Ambient PM 2.5 Levels in 5 Indian Metropolitan Cities: A Comprehensive Study from 2016-2022

Project report submitted in partial fulfillment of the requirement for

POST-GRADUATE DIPLOMA IN STATISTICAL METHODS AND ANALYTICS



Submitted by

Atanu Das & Vicky Raj Ray

Roll no: (DST-22/23-004) & (DST-22/23-025)

4th JUNE, 2023

INDIAN STATISTICAL INSTITUTE

North-East Centre, Punioni, Solmara

Tezpur-784501

Assam

Date: 4th June, 2023

Certificate

This is to certify Mr. Atanu Das and Mr. Vicky Raj Ray has done the project under my supervision and guidance (from 27th January to 4th June). This is an original project report based on work carried out by them in partial fulfillment of the requirement for the Post-Graduate Diploma in Statistical Methods and Analytics programmed of the Indian Statistical Institute, North-East Centre, Tezpur, Assam.

(Supervisor's signature)

Dr. Darpa Saurav Jyethi

Acknowledgements

We would like to express our heartfelt gratitude to the individuals who played a vital role in the successful completion of this project. Our deepest appreciation goes to **Dr. Darpa Sourav Jyethi**, our supervisor and Assistant Professor at ISI Tezpur, whose unwavering guidance, valuable resources, insightful ideas, and motivation have been instrumental in shaping this project. We are truly grateful for his constant support throughout the journey.

We also want to extend our sincere thanks to all the professors at ISI Tezpur for accommodating our needs and providing us with valuable suggestions, continuous support, and encouragement. Their expertise and guidance have been invaluable in the development of this project.

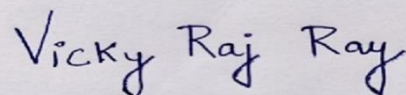
Furthermore, we would like to express our gratitude to our colleagues at ISI Tezpur who have been a part of our journey. Their valuable comments, cooperation, and endorsement have been immensely appreciated. Their insights and collaboration have contributed to the overall success of this endeavor.

Lastly, we would like to express our sincere gratitude to everyone who contributed to the realization of this project. Whether your role was significant or small, your support, guidance, and encouragement made a meaningful impact. We are truly thankful for the opportunity to have worked alongside such remarkable individuals.

Date: 04/June/202



Atanu Das



Vicky Raj Ray

[Signature(s)]

Contents

1.Introduction:	1
2. Methodology	3
2.1 Data Source	3
2.2 Study Area	3
2.2.1 New Delhi	4
2.2.2 Kolkata	4
2.2.3 Mumbai	4
2.2.4 Chennai	4
2.2.5 Hyderabad	4
2.3 Data Preprocessing:	5
2.3.1 Effective Strategies for Handling Missing, Invalid, and Suspect Data in the PM2.5 Dataset for Indian Metropolitan Cities:	6
2.4 Data visualization:	8
2.4.1 Trend:	8
2.4.2 Box plot:	10
2.4.3 Heat map:	11
2.5 Average PM-2.5 Concentration by Day and 6- Hour interval	13
2.5.1 Delhi:	13
2.5.2 Kolkata	15
2.5.3 Mumbai:	16
2.5.4 Chennai	18
2.5.5 Hyderabad	20
2.6 Average PM-2.5 Concentration by Day and Hour	22
2.6.1 New Delhi	22
2.6.2 Kolkata	22
2.6.3 Mumbai	23
2.6.4 Chennai	23
2.6.5 Hyderabad	24
3 Results	24
3.1 Statistical tools	24
3.1.1 Normality tests	24

3.2 Wilcoxon signed-rank test.....	28
I).Delhi:	29
II)Kolkata	30
III) Mumbai	30
IV) Chennai:	31
V) Hyderabad.....	32
3.3 Seasonal Variation	33
3.3.1 New Delhi	33
3.3.2 Kolkata	35
3.3.3 Mumbai	38
3.3.4 Chennai	40
3.3.5 Hyderabad	42
4. Conclusions	44
5.References :	45
I)Reference from Journal	45
II) Reference from Book	45
III) References from data collected sites	45

1.Introduction:

Air pollution is a critical global environmental issue that poses significant threats to the health and well-being of communities worldwide. India, in particular, has experienced a troubling surge in air pollution levels within its major metropolitan cities. This project aims to address this urgent issue by conducting an analysis of the ambient concentration of PM 2.5, which refers to fine particulate matter with a diameter of 2.5 micrometers or smaller, in five prominent cities across India from 2016 to 2022. Inhaling PM 2.5 can lead to severe health problems such as respiratory illnesses, cardiovascular diseases, strokes, and even lung cancer. These harmful particles originate from various sources, including vehicle emissions, industrial activities, construction sites, and the burning of fossil fuels.

In India, the presence of fine particulate matter (PM_{2.5}) has emerged as a major environmental concern, with approximately 50% of the country's population exposed to PM_{2.5} levels exceeding the Indian air quality standard of 40 $\mu\text{g m}^{-3}$. This exposure has resulted in over 1 million premature deaths in India. Satellite data indicates a worsening trend of air pollution in India, with a significant increase in PM_{2.5} levels observed from 2010 to 2019 compared to the previous decade. It is crucial to swiftly understand the underlying mechanisms driving particulate pollution in India to effectively manage air quality.

The root cause of air pollution in India lies in emissions from various sources, including residential biomass combustion, power plants, and industrial coal combustion. The Indo-Gangetic Plain (IGP) in Northern India stands out as one of the most affected regions, with extremely high PM_{2.5} levels. Emissions from the IGP contribute to approximately 46% of the total premature mortality across India. Household emissions have been identified as the largest contributor to PM_{2.5} in India. Taking measures to mitigate household emissions, such as addressing biomass burning for cooking, domestic heating, and the use of kerosene for lighting, could lead to a significant 30.7% improvement in annual PM_{2.5} concentrations.

While emissions play a significant role in air pollution, meteorology is also considered an external factor that influences day-to-day variations in PM_{2.5} levels. Several studies have explored the correlation between local meteorological parameters and daily PM_{2.5} concentrations. For example, during winter in the IGP, pollution buildup is favored by meteorological conditions such as low wind speeds, shallow boundary layer heights, and high relative humidity. However, the associated circulation characteristics and underlying processes have not been extensively investigated. The impact of weather conditions on PM_{2.5} pollution

remains poorly understood, despite the use of pollution-favorable meteorological indices that incorporate correlated meteorological parameters to project future conditions under a warming climate. The question of whether India will experience more or fewer polluted days in the future remains a topic of debate, and the unique topography of northern India adds complexity to this issue.

This study aims to provide valuable insights into air quality conditions in major metropolitan areas of India by examining trends and patterns in PM_{2.5} levels over time and comparing them among different cities. Rapid urbanization, industrialization, rural-urban migration, and increased vehicular usage have led to heightened air pollution, with PM_{2.5} emerging as a critical environmental challenge.

Delhi, Kolkata, Chennai, Hyderabad, and Mumbai, the major cities of India, are grappling with severe air pollution problems. Delhi, in particular, has gained international attention as the fourth most polluted city worldwide, consistently exceeding national air quality standards for respiratory suspended particulate matter (RSPM). The primary contributor to pollution in Delhi is road traffic. Kolkata has also gained notoriety as a highly polluted city, especially during the winter season, earning it the moniker of a "dusty city."

Over the years, Delhi has consistently remained the most polluted state, with a declining percentage of the population complying with the annual ambient standard of 40 mg m⁻³. Fuel combustion is identified as a significant factor contributing to PM_{2.5} pollution in India, accounting for approximately 81% of the pollution. This includes various activities such as personal and freight transportation, electricity generation, industrial manufacturing, cooking, heating, construction, road dust resuspension, and waste burning. By analyzing the pollution trend alongside fuel consumption and activity patterns, a clearer understanding of this evolution can be obtained.

The two-month COVID-19 lockdown in 2020 served as evidence that achieving "clean air" necessitates emission reductions across all sources and regions. However, substantial and sustained efforts are required to not only meet the national ambient standard but also the World Health Organization's guideline of 5 mg m⁻³. Achieving this goal may involve making difficult decisions to enable and sustain larger reductions across sectors. Addressing air pollution in major Indian cities is a complex challenge that requires comprehensive and coordinated measures. By leveraging the findings of this study and adopting targeted interventions, we can work towards mitigating the adverse effects of PM_{2.5} pollution and improving the overall air quality in these metropolitan areas.

2. Methodology

2.1 Data Source

The data set was meticulously obtained from the official website of the esteemed United States Embassy and Consulates in India, which can be accessed at the highly credible link:

<https://www.airnow.gov/international/us-embassies-and-consulates/#India>.

This website serves as a reliable platform for reporting real-time air quality information and daily forecasts using the EPA Air Quality Index (AQI). The AQI, a numerical scale, provides insights into the levels of air pollution and associated health risks by measuring various pollutants. By utilizing the historical data available on this website, it becomes possible to conduct an analysis of the ambient PM 2.5 levels across five major metropolitan cities in India, covering the period from 2016 to 2022.

2.2 Study Area

The study area can be easily identified by referring to the accompanying map.



Fig 1: Map of India highlighting the selected cities for the study, indicated by red pointers. (The map source can be found at: <https://in.usembassy.gov/u-s-citizen-services/consular-posts-india/graphical-india-map-emb-cons/>.)

2.2.1 New Delhi

New Delhi, the capital of India, is located at a latitude of 28.61 and a longitude of 77.21. The city experiences extreme temperatures, with May being the hottest month at an average of 33°C, while January is the coldest at 13°C. July is the wettest month, receiving an average of 180mm of rain. The majority of the annual rainfall, about 87%, occurs during the monsoon months from June to September.

2.2.2 Kolkata

Kolkata, the capital of the Indian state of West Bengal, is located at a latitude of 22.57 and a longitude of 88.36. The city experiences a Tropical wet and dry climate, with an annual mean temperature of 24.8 °C. The presence of the Bay of Bengal greatly influences Kolkata's climate, resulting in an average annual temperature of around 27°C.

2.2.3 Mumbai

Mumbai, the capital city of Maharashtra, is located at a latitude of 19.07 and a longitude of 72.88. The hottest month in Mumbai is May, with an average temperature of 30°C, while January is the coldest month at 24°C. Summers in Mumbai receive significantly more rainfall compared to winter, with an average annual rainfall of 2386 mm.

2.2.4 Chennai

Chennai, the capital of Tamil Nadu, is the sixth-largest city and fourth-most populous urban agglomeration in India, according to the 2011 census. Situated on the eastern coast of South India, its coordinates are 13.08 latitude and 80.27 longitude. Chennai experiences a tropical wet and dry climate, with May being the hottest month at an average temperature of 91°F (33°C) and January being the coldest at 79°F (26°C). The city receives an average annual rainfall of about 140 cm (55 in), with the majority coming from the north-east monsoon winds between mid-October and mid-December.

2.2.5 Hyderabad

Hyderabad, the capital of Telangana, India, is located at a latitude of 17.38 and a longitude of 78.49. It experiences a hot climate with a mean annual temperature of 26.7 degrees Celsius. The hottest month is May, with an average temperature of 33 degrees Celsius, while the coldest month, December, is relatively warm with an average temperature of 21.6 degrees Celsius. The city receives heavy rainfall during the south-west summer monsoon, which occurs between June and September and contributes to the majority of its mean annual rainfall.

2.3 Data Preprocessing:

Objective:

The objective of this section is to analyze and preprocess the air quality data for the five major Indian cities: Delhi, Kolkata, Hyderabad, Chennai, and Mumbai. The focus will be on examining the total data, missing data, invalid data, suspected data, and the corresponding percentages for each city.

City	Total Data	Valid Data	Missing data	Invalid data	Suspect data	Percentage of Missing data	Percentage of Invalid data	Percentage of valid data
Delhi	59661	58112	733	816	0	1.228	1.367	97.403
Kolkata	57874	56450	1003	421	7	1.732	0.727	97.527
Mumbai	58068	55450	2002	616	4	3.450	1.061	95.481
Hyderabad	58922	57494	1179	249	1	2.000	0.422	97.574
Chennai	59554	56418	2245	891	1	3.785	1.495	94.716

Results and Insights:

1. Chennai has the highest percentage of missing data (3.785%) and invalid data (1.495%) among the five cities. In contrast, Hyderabad has the lowest percentage of missing data (2.000%), and Kolkata has the lowest percentage of invalid data (0.727%).
2. The data quality can affect the reliability of air quality readings, which can be used to guide public health interventions.
3. Data collection and storage systems in each city need improvement to ensure accurate and complete data.
4. The level of trust and credibility of the data can be affected by its quality, which can have consequences for decision-making and policy development.
5. It is important to identify and address issues with data collection and storage systems to improve data quality.
6. Effective communication of improvements in data quality is crucial. Transparently addressing data quality issues and sharing updates with stakeholders helps maintain the credibility of the data.

2.3.1 Effective Strategies for Handling Missing, Invalid, and Suspect Data in the PM2.5 Dataset for Indian Metropolitan Cities:

Handling missing, invalid, and suspect data is essential in data cleaning and analysis. The PM2.5 dataset for Indian metropolitan cities is a prime example of a dataset that may contain such data due to errors in measurement or recording. Therefore, it is important to understand the dataset, identify any missing, invalid, or suspect data, and choose an appropriate method for handling the data.

Methods:

Various common approaches exist for handling missing data, such as imputation, deletion, or the use of specialized models. In cases of invalid data, complete removal or recoding as missing values can be employed. For suspect data, options include removal or recoding, or performing sensitivity analyses to assess their impact on the overall dataset.

Linear interpolation is a valuable method for replacing missing values, especially in cases involving invalid or suspect data. Multiple imputation, on the other hand, involves creating multiple estimates of missing values using statistical models that incorporate information from the observed data. Other methods for replacing missing values include regression analysis, mean substitution, and hot deck imputation.

Steps:

To effectively handle missing, invalid, suspect, and outlier data in the PM2.5 dataset for Indian metropolitan cities, the following steps are undertaken:

1. Understanding the dataset: Begin by acquiring a comprehensive understanding of the PM2.5 dataset. PM2.5 refers to fine particulate matter with a diameter of less than 2.5 micrometers, known for its respiratory health risks due to its ability to penetrate deep into the lungs.

2. Identifying missing, invalid, suspect, and outlier data: Conduct a meticulous examination of the dataset to identify any missing, invalid, suspect, or outlier data points. This involves scrutinizing values that are clearly incorrect or fall outside the expected range.
3. Selecting the appropriate methods: Once missing, invalid, suspect, and outlier data are identified, choose appropriate methods to handle the data. Linear interpolation is a suitable method for estimating missing values based on the trend or pattern observed in neighboring data points. Additionally, the Boxplot method can be utilized to detect and address outliers.
4. Applying data handling techniques: Implement the selected methods, such as linear interpolation for missing data and the Boxplot method for outliers. Linear interpolation fills in gaps in the dataset by estimating missing values, while the Boxplot method identifies and removes outlier values based on their deviation from the majority of the dataset.
5. Evaluating the results: Evaluate the outcomes after applying the data handling techniques. Compare the estimated values to the actual values and assess the overall quality of the dataset after handling missing, invalid, suspect, and outlier data. This evaluation ensures the accuracy and reliability of the processed dataset.

Combining effective strategies for handling missing, invalid, suspect, and outlier data in the PM2.5 dataset for Indian metropolitan cities leads to accurate and reliable results. Linear interpolation serves as a valuable method for estimating missing values based on the trend observed in the surrounding data points. Additionally, the Boxplot method effectively identifies and removes outliers, minimizing their impact on the analysis. However, it is crucial to consider the limitations and assumptions associated with these methods and select the most appropriate approach based on the dataset's characteristics and research question. By employing these strategies thoughtfully, the PM2.5 dataset can be processed and cleaned to provide trustworthy information for further analysis. Furthermore, in the context of outliers, it is relevant to consider the specific range of PM2.5 concentration values that are considered outliers. For example, in this case, values above 600 and below 10 are considered outliers and are subsequently removed from the dataset. By removing these extreme values, the impact of erroneous readings on the overall analysis can be mitigated.

2.4 Data visualization:

Data visualization is an essential tool for analyzing and interpreting large and complex datasets. It allows us to identify patterns and relationships within the data, enabling us to draw meaningful insights and communicate our findings effectively.

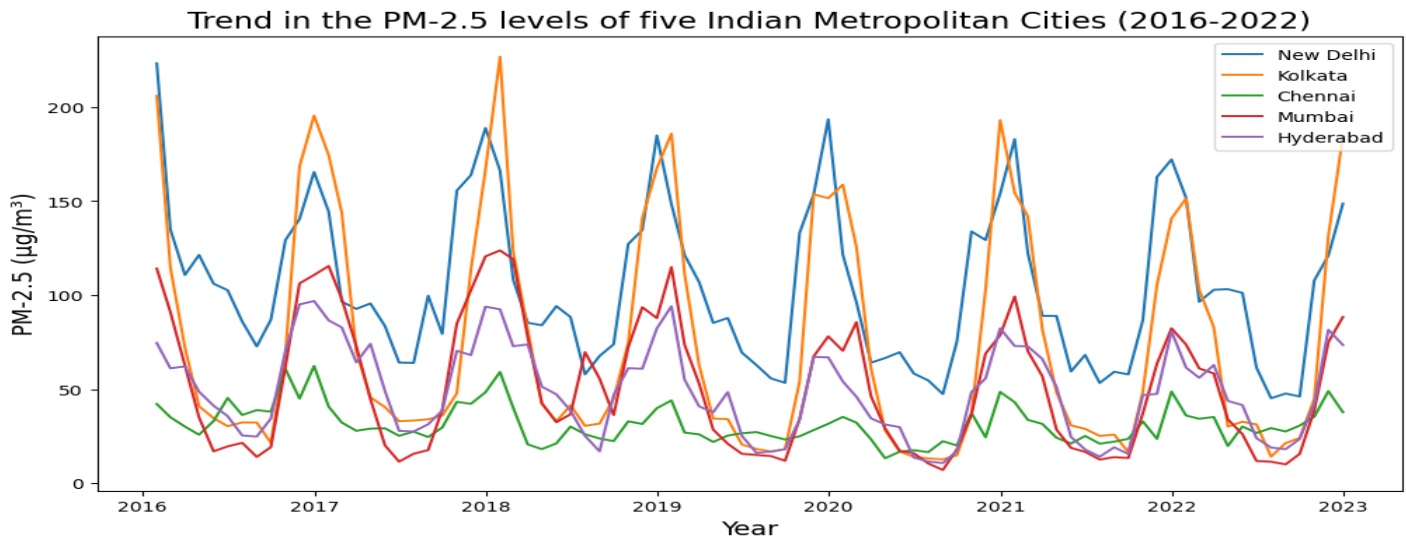
There are various visualization techniques available, each suitable for different types of data and analysis goals. Among these techniques, trend analysis, box plots, and heat maps are commonly used to present data in a clear and informative way.

2.4.1 Trend:

In general, a trend is a pattern or direction of change in a particular area over time. It is a term that is commonly used in economics, finance, and statistics, but can also be applied to other fields such as fashion, technology, or social behavior.

In statistics, a trend can refer to the long-term pattern or direction of change in a time series or data set. This can be analyzed using various techniques, such as linear regression, moving averages, or exponential smoothing. The trend component of a time series is often used to make predictions about future values or to identify potential anomalies in the data. A trend can be a useful tool for understanding and predicting changes in various areas and can help individuals and organizations make informed decisions based on the direction of the trend.

The plot shows the monthly mean PM-2.5 data for five metropolitan cities in India (New Delhi, Kolkata, Chennai, Mumbai, Hyderabad) from 2016 to 2022.

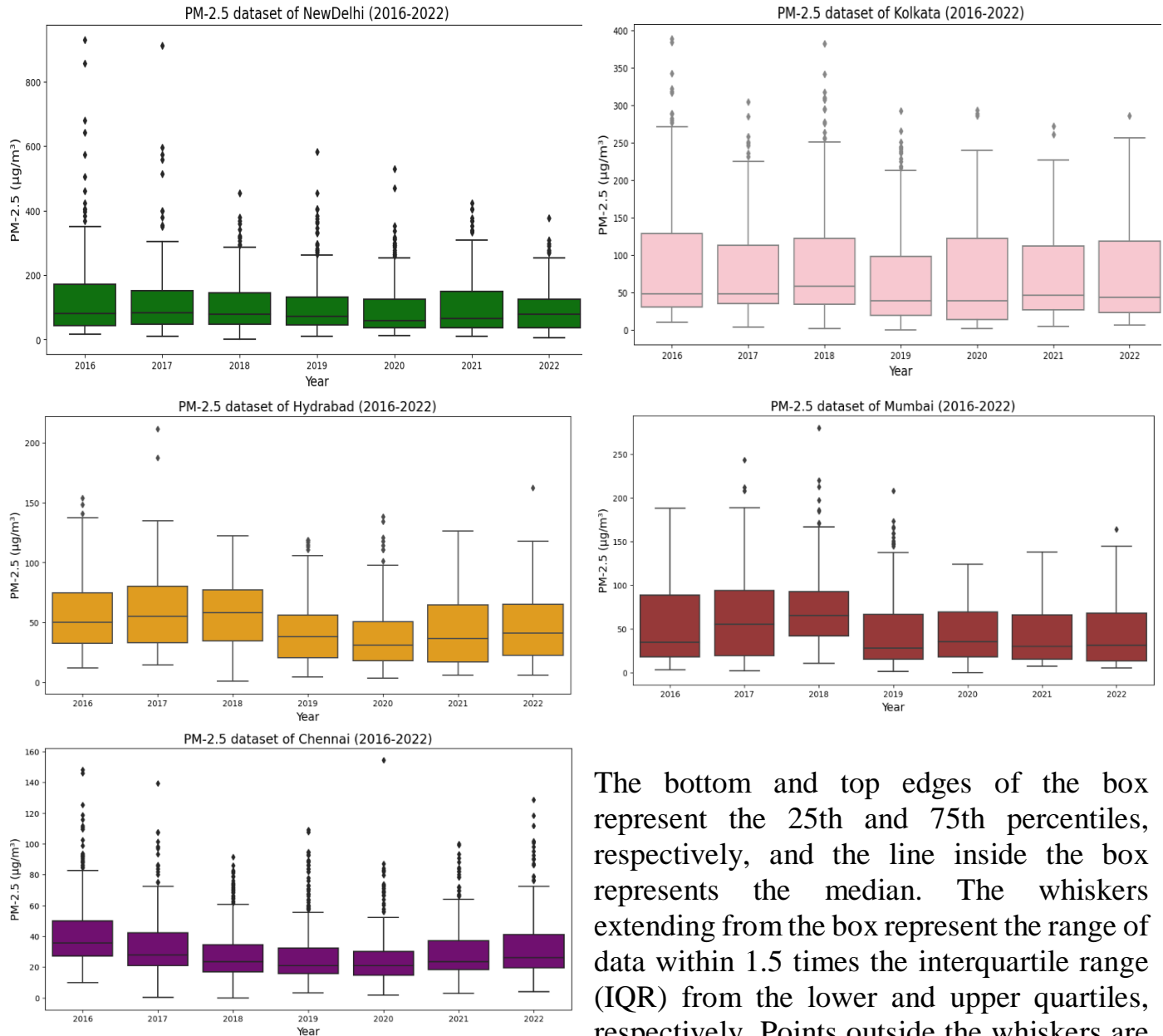


- ❖ The trend in the PM-2.5 levels can provide important insights into the overall behavior of the data and can be used to make predictions about future values. From the plot, we can see that the PM-2.5 levels vary considerably across the five cities, with New Delhi having the highest levels overall. However, there are some general trends that are visible across all the cities.
- ❖ Overall, we can see a decreasing trend in PM-2.5 levels in all the cities from 2017 to 2019, followed by a slight increase in 2020 and 2021. The COVID-19 pandemic had a significant impact on the PM-2.5 levels in the cities, as we can see a sharp decrease in levels in 2020 during the lockdown period, followed by a gradual increase in 2021 as restrictions were lifted.
- ❖ The PM-2.5 concentration levels in Chennai and Hyderabad are generally lower compared to the other three cities. This could be due to factors such as lower population density, fewer industrial and vehicular sources of pollution, and proximity to the coast.

The analysis of ambient PM 2.5 levels across the five major Indian metropolitan cities from 2016 to 2022 is significant as air pollution is a serious issue affecting public health and the environment. The project aims to identify trends and patterns in the PM 2.5 levels over time and compare them between different cities.

2.4.2 Box plot:

The boxplot provides a graphical representation of the distribution of PM-2.5 levels across the five Indian cities (New Delhi, Kolkata, Chennai, Mumbai, and Hyderabad) over the period from 2016 to 2022.



The bottom and top edges of the box represent the 25th and 75th percentiles, respectively, and the line inside the box represents the median. The whiskers extending from the box represent the range of data within 1.5 times the interquartile range (IQR) from the lower and upper quartiles, respectively. Points outside the whiskers are considered outliers and are shown as individual data points.

Key Insights:

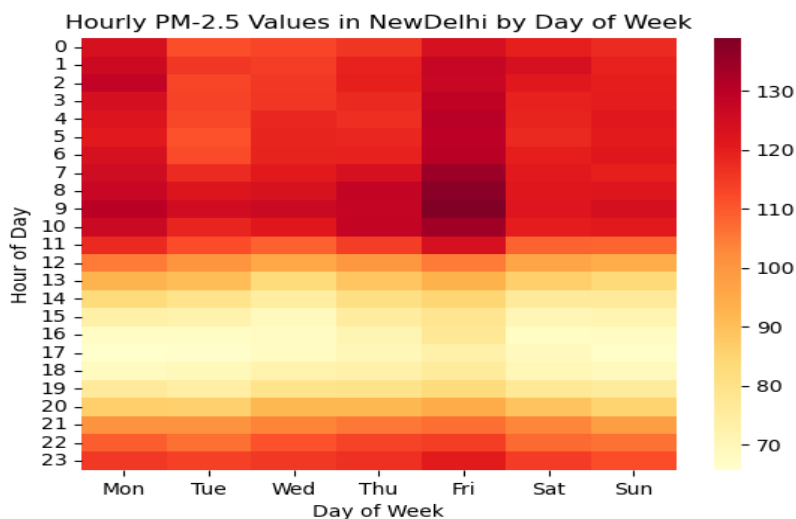
The box plot provides several insights into the distribution of PM-2.5 levels in the five Indian cities.

- Firstly, it shows that the median PM-2.5 levels in Chennai and Hyderabad are generally lower than in the other three cities, with Delhi having the highest median levels.
- Secondly, the plot highlights the variability of the PM-2.5 levels across the cities, with Mumbai having the widest range of PM-2.5 levels and Chennai having the narrowest range.
- Finally, the plot shows that all five cities have some extreme values of PM-2.5 levels, which are indicated as outliers.

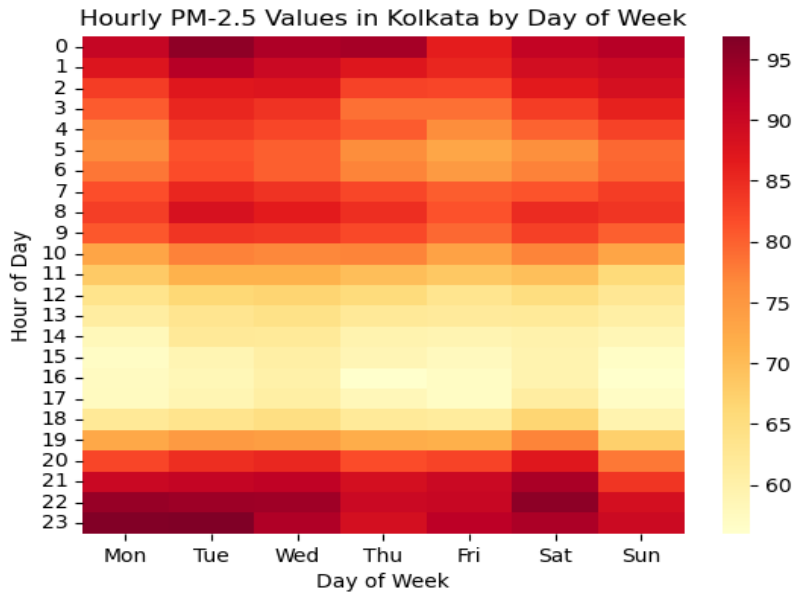
Overall, the box plot provides a clear and concise summary of the distribution of PM_{2.5} levels in the five Indian cities, which can help researchers to understand the differences between the cities and identify areas where interventions to reduce PM_{2.5} levels are needed.

2.4.3 Heat map:

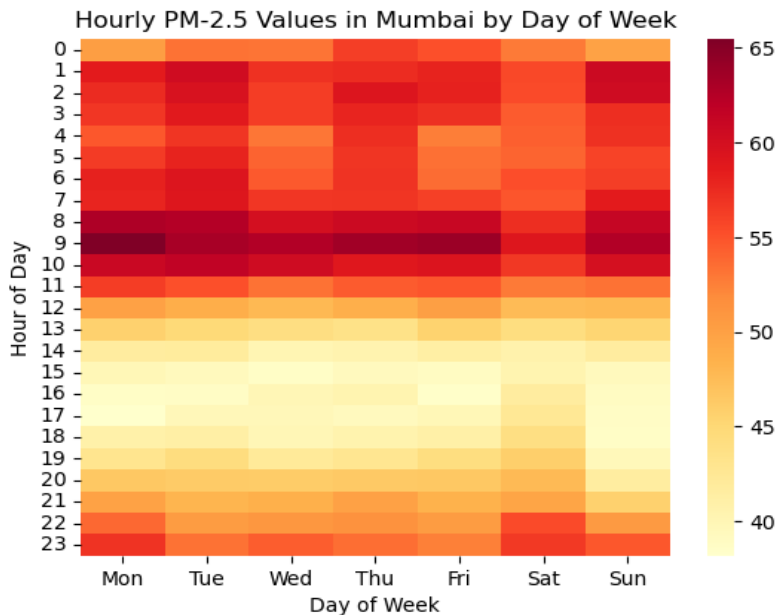
A heat map is a graphical representation of a matrix or table of data where values are represented as colors, and each row and column are a separate entity. The right-side scale on the heatmap represents the magnitude of the PM-2.5 values in micrograms per cubic meter ($\mu\text{g}/\text{m}^3$). This scale allows us to easily see the relative concentration of PM-2.5 at different times of the day and days of the week. The colors on the heatmap correspond to the values on the scale, with darker colors indicating higher concentrations of PM-2.5. This information can be useful for identifying patterns in PM-2.5 pollution and for developing strategies to reduce pollution levels. Here are some insights from the heatmaps:



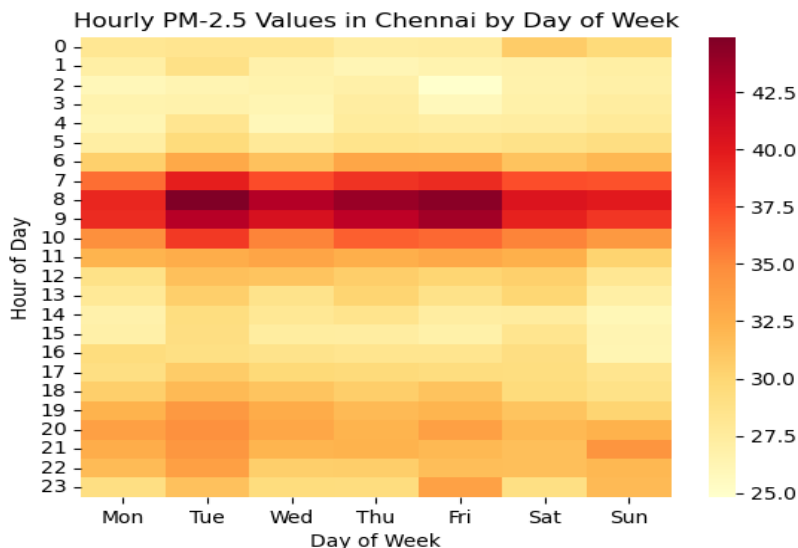
i) New Delhi: The heatmap shows high PM-2.5 values throughout the week, with the highest levels on weekdays during the morning and evening rush hours (7-9 AM and 5-7 PM). The lowest PM-2.5 values are observed on Sundays during the daytime.



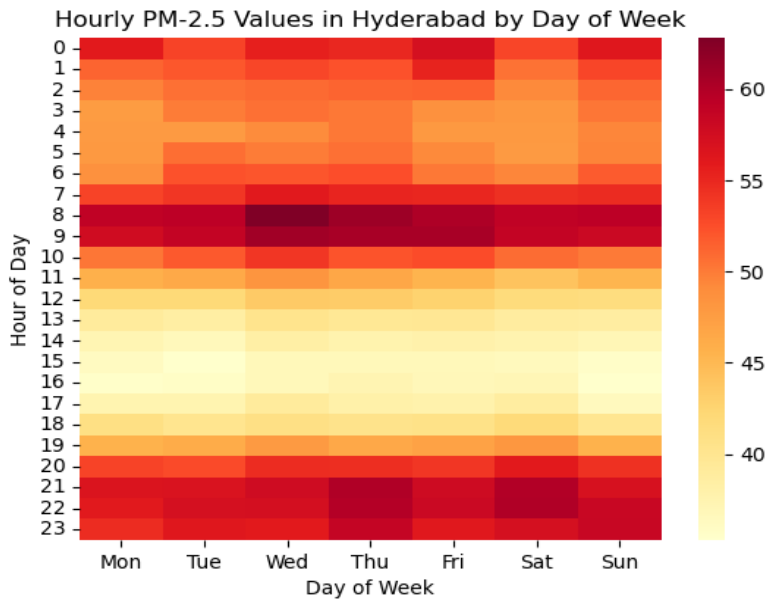
ii)Kolkata:The heatmap shows moderate to high PM-2.5 values throughout the week, with the highest levels during the late evening hours (8-11 PM) and the lowest levels during the early morning hours (5-7 AM). There is a slight increase in PM-2.5 values on weekdays during the morning and evening rush hours (7-9 AM and 5-7 PM).



iii)Mumbai:The heatmap shows high PM-2.5 values on weekdays during the morning and evening rush hours (7-9 AM and 5-7 PM), with lower values during the rest of the day. On weekends, the PM-2.5 values are lower overall than on weekdays, but there is still a slight increase during the morning rush hour.



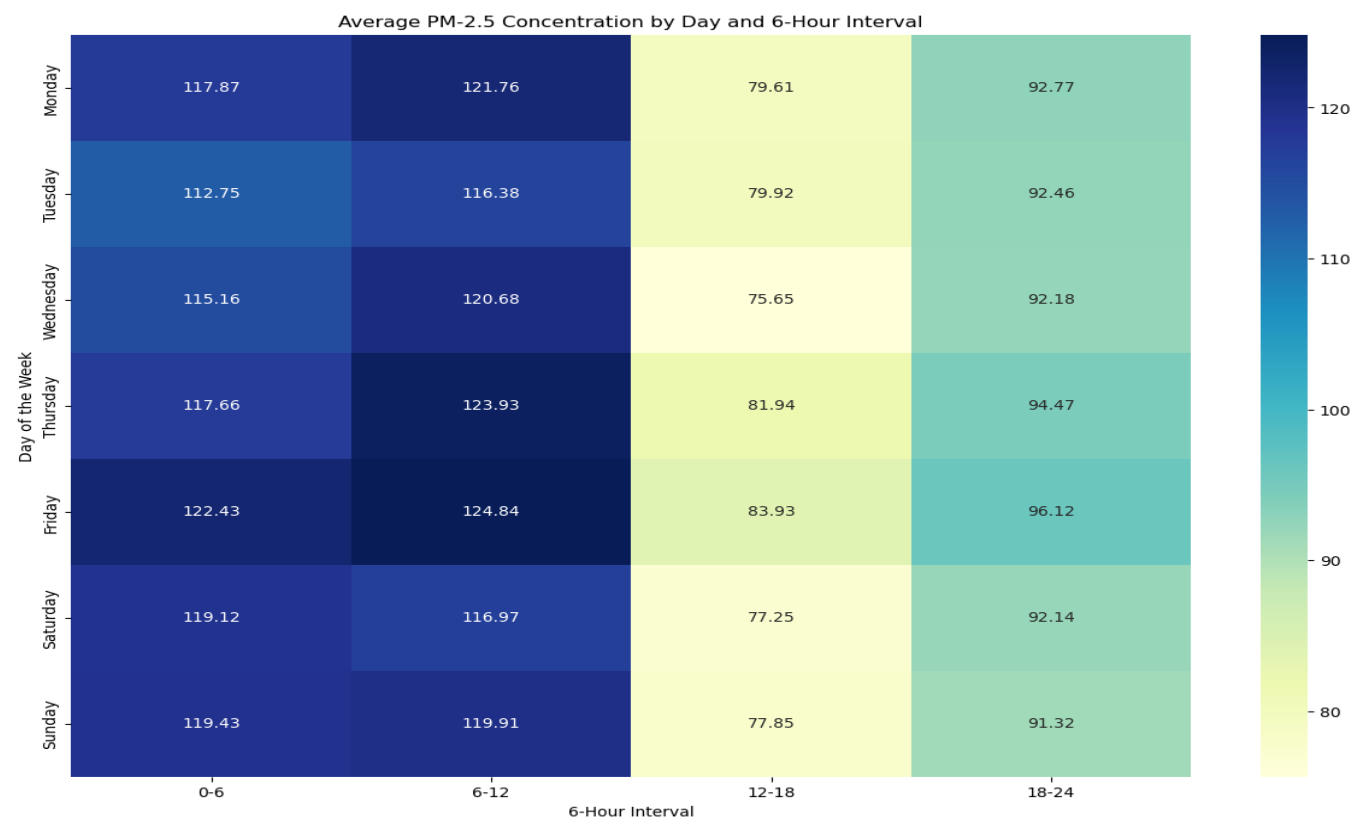
iv)Chennai:The heatmap shows moderate to high PM-2.5 values throughout the week, with the highest levels during the late evening hours (8-11 PM) and the lowest levels during the early morning hours (5-7 AM). There is a slight increase in PM-2.5 values on weekdays during the morning and evening rush hours (7-9 AM and 5-7 PM).



v)Hyderabad:The heatmap shows moderate to high PM-2.5 values throughout the week, with the highest levels during the early morning hours (5-7 AM) and late evening hours (8-11 PM). The lowest PM-2.5 values are observed on Sundays during the daytime.

2.5 Average PM-2.5 Concentration by Day and 6- Hour interval

2.5.1 Delhi:



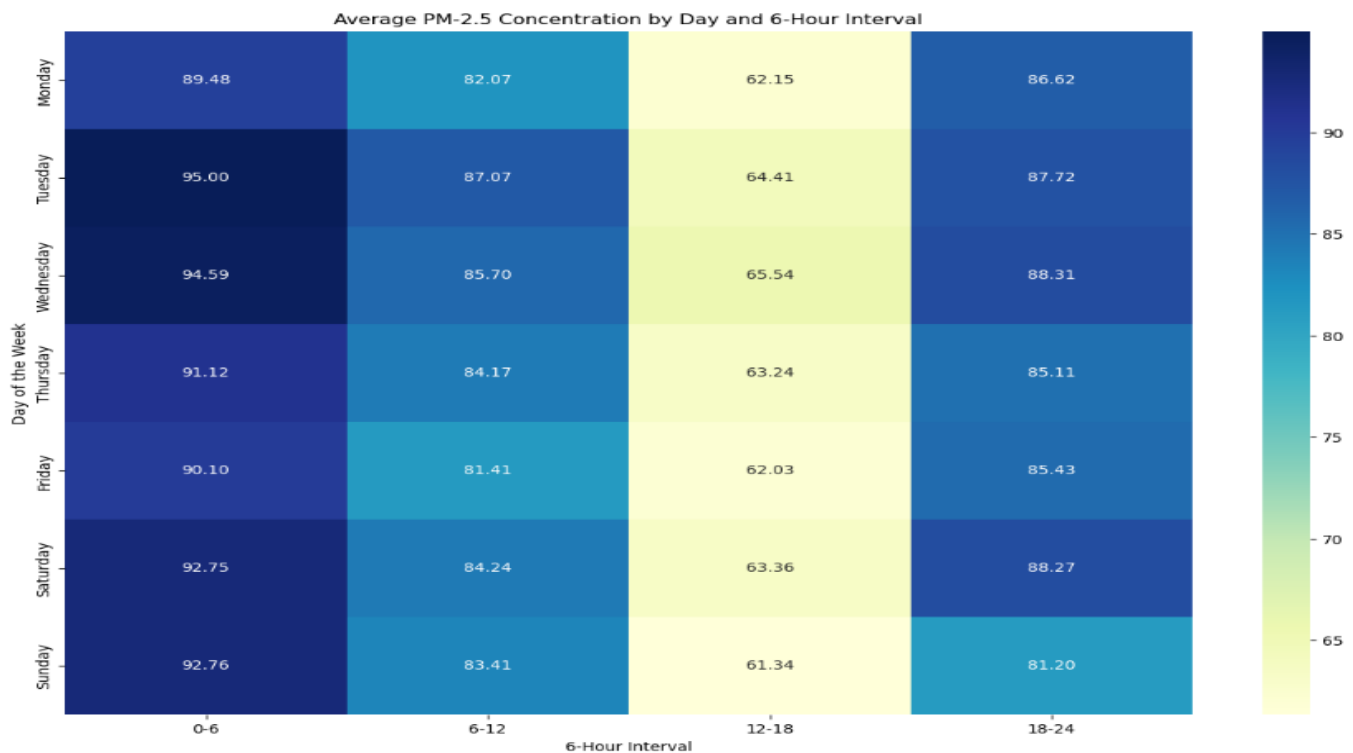
	0-6	6-12	12-18	18-24
Monday	117.8701	121.7648	79.6067	92.7658
Tuesday	112.7469	116.3793	79.91765	92.4576
Wednesday	115.1576	120.68	75.64893	92.17971
Thursday	117.6588	123.9263	81.94354	94.46647
Friday	122.4287	124.8376	83.93115	96.11792
Saturday	119.1212	116.9654	77.24978	92.14176
Sunday	119.4318	119.915	77.84802	91.31862

The descriptive table provides insights into the average PM-2.5 concentrations in New Delhi across different time intervals and days of the week:

- Monday shows relatively high PM-2.5 concentrations during the early morning '0-6' interval (117.87) and the mid-morning '6-12' interval (121.76). The concentrations decrease during the afternoon '12-18' interval (79.61) and slightly rise again during the evening '18-24' interval (92.77).
- Tuesday exhibits similar patterns to Monday, with higher concentrations in the early morning and mid-morning intervals (112.75 and 116.38) and lower concentrations in the afternoon (79.92) and evening (92.46) intervals.
- Wednesday follows a similar trend as Tuesday, but with a slightly lower concentration during the afternoon '12-18' interval (75.65).
- Thursday shows a relatively high PM-2.5 concentration during the mid-morning '6-12' interval (123.93) and a slightly higher concentration during the evening '18-24' interval (94.47).
- Friday experiences a significant increase in the average PM-2.5 concentration during the mid-morning '6-12' interval (124.84) and maintains higher concentrations throughout the day, peaking during the evening '18-24' interval (96.12).
- Saturday displays a relatively consistent and moderate concentration throughout the day, with slightly higher levels during the early morning and evening intervals (119.12 and 92.14).
- Sunday shows a similar pattern to Saturday, with slightly higher concentrations during the early morning and mid-morning intervals (119.43 and 119.91) and slightly lower concentrations during the afternoon and evening intervals (77.85 and 91.32).

These insights provide a clear picture of the variations in PM-2.5 concentrations across different intervals and days of the week, allowing for a better understanding of pollution levels in New Delhi throughout the week.

2.5.2 Kolkata



	0-6	6-12	12-18	18-24
Monday	89.47811	82.07077	62.14739	86.6159
Tuesday	94.99501	87.07423	64.40785	87.72025
Wednesday	94.58846	85.69713	65.53647	88.30765
Thursday	91.12359	84.16811	63.23539	85.10622
Friday	90.09806	81.41432	62.03105	85.43175
Saturday	92.75467	84.23955	63.35758	88.26819
Sunday	92.75676	83.41033	61.33777	81.19904

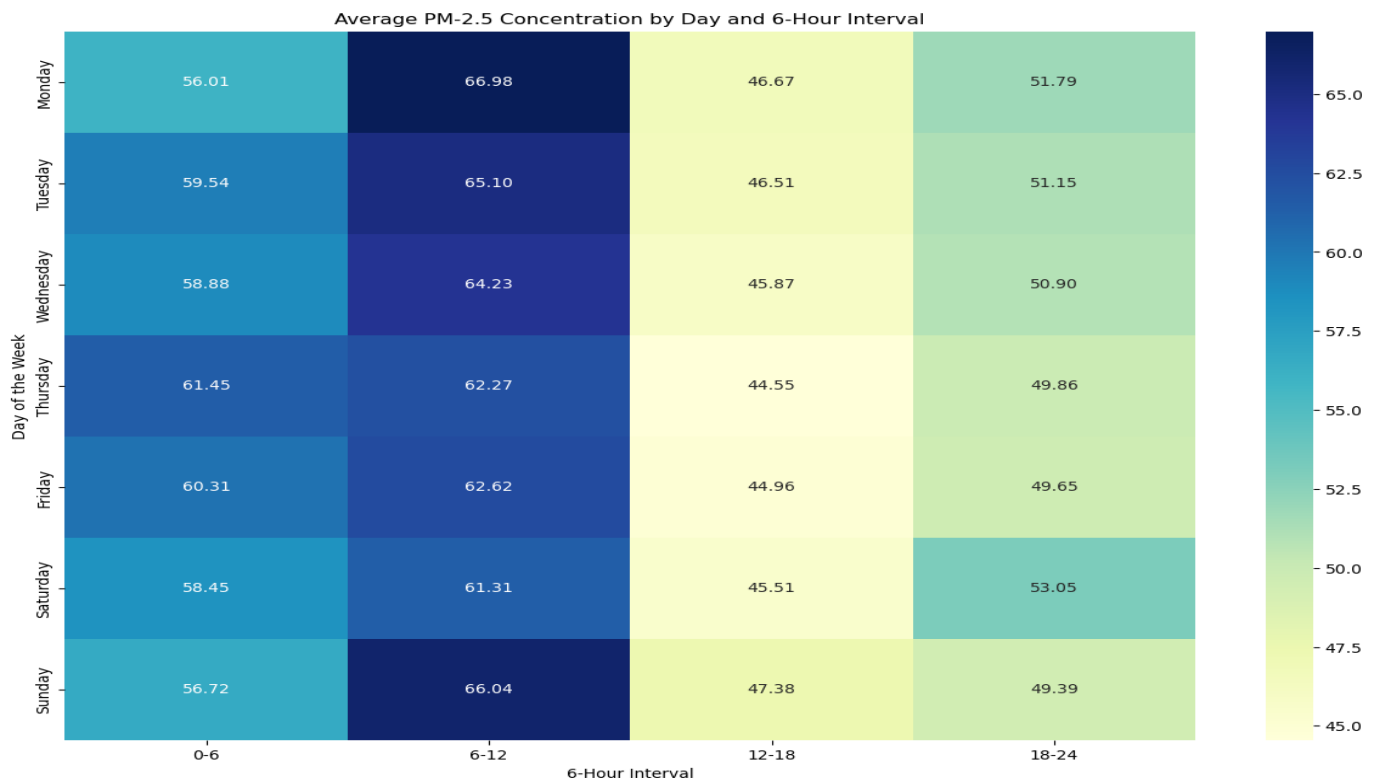
The descriptive table provides insights into the average PM_{2.5} concentrations in Kolkata across different time intervals and days of the week:

- Monday shows relatively lower PM_{2.5} concentrations during the early morning '0-6' interval (89.48) and mid-morning '6-12' interval (82.07). The concentrations increase during the afternoon '12-18' interval (62.15) and slightly rise again during the evening '18-24' interval (86.62).
- Tuesday exhibits a similar pattern to Monday, with slightly higher concentrations during the early morning and mid-morning intervals (94.99 and 87.07), and a further increase during the afternoon and evening intervals (64.41 and 87.72).

- Wednesday shows a consistent trend similar to Tuesday, with slightly lower concentrations during the early morning and mid-morning intervals (94.59 and 85.70), and a slight increase during the afternoon and evening intervals (65.54 and 88.31).
- Thursday displays relatively lower PM2.5 concentrations throughout the day, with the lowest concentration during the afternoon '12-18' interval (63.24).
- Friday shows a similar pattern to Thursday, with slightly higher concentrations during the early morning and mid-morning intervals (90.10 and 81.41), and a slight increase during the evening interval (85.43).
- Saturday exhibits a consistent trend similar to Friday, with slightly higher concentrations during the early morning and evening intervals (92.75 and 88.27).
- Sunday displays relatively lower PM2.5 concentrations throughout the day, with the lowest concentration during the afternoon '12-18' interval (61.34).

These insights provide a clear understanding of the variations in PM2.5 concentrations in Kolkata across different intervals and days of the week, enabling a better comprehension of pollution levels throughout the week.

2.5.3 Mumbai:

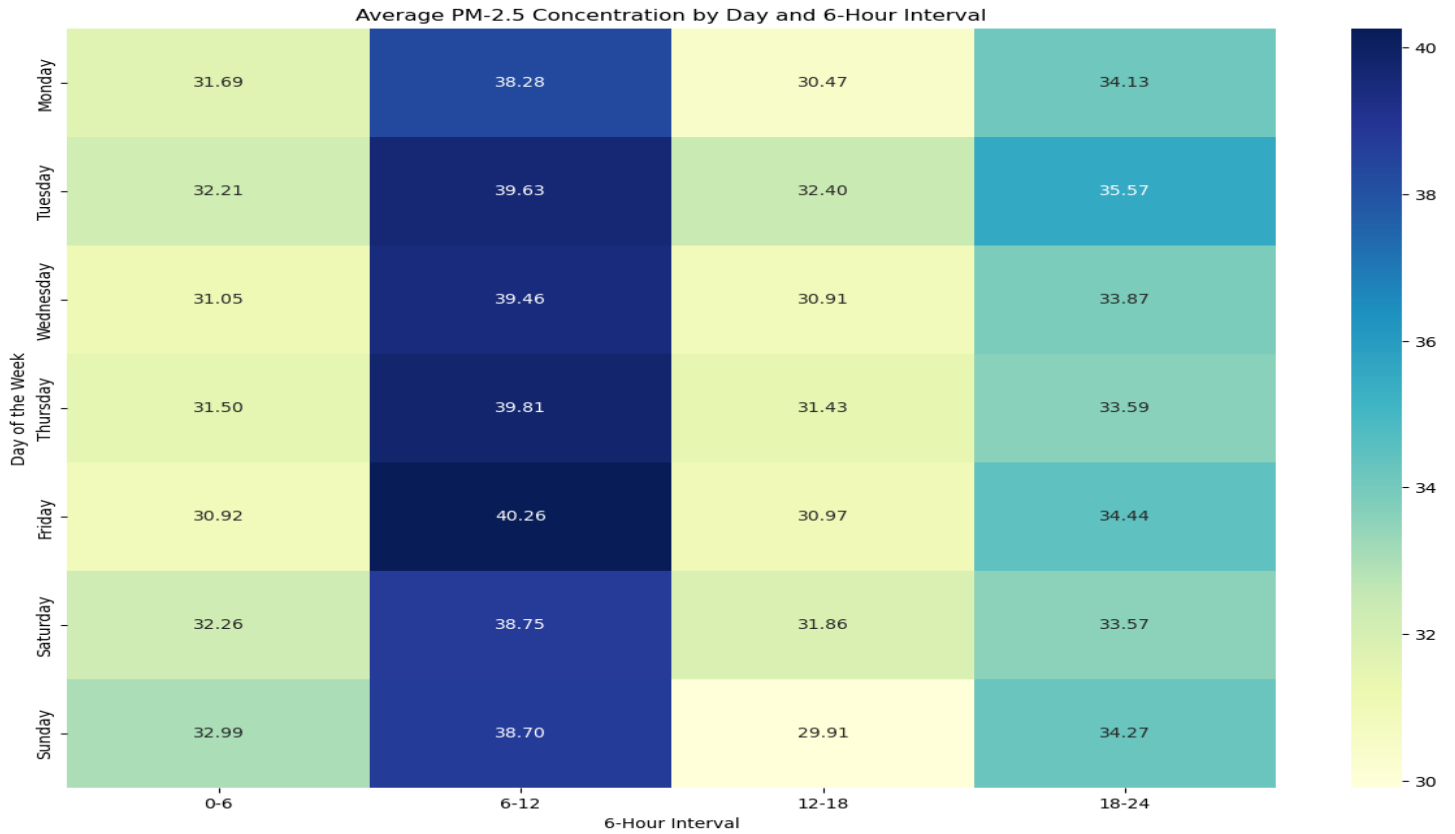


	0-6	6-12	12-18	18-24
Monday	56.00727	66.98395	46.67317	51.79103
Tuesday	59.54427	65.09842	46.51437	51.14973
Wednesday	58.87642	64.22608	45.87266	50.89649
Thursday	61.45499	62.27116	44.55182	49.85512
Friday	60.3088	62.62465	44.96235	49.64727
Saturday	58.45018	61.30748	45.50864	53.05425
Sunday	56.71595	66.0442	47.38477	49.39225

The descriptive table provides insights into the average PM2.5 concentrations in Mumbai across different time intervals and days of the week:

- Monday shows relatively lower PM2.5 concentrations during the early morning '0-6' interval (56.01) and evening '18-24' interval (51.79). The concentrations increase during the mid-morning '6-12' interval (66.98) and slightly decrease during the afternoon '12-18' interval (46.67).
- Tuesday exhibits a similar pattern to Monday, with slightly higher concentrations during the early morning and mid-morning intervals (59.54 and 65.10), and a slight decrease during the afternoon and evening intervals (46.51 and 51.15).
- Wednesday displays a consistent trend similar to Tuesday, with slightly lower concentrations during the early morning and mid-morning intervals (58.88 and 64.23), and a slight decrease during the afternoon and evening intervals (45.87 and 50.90).
- Thursday shows a gradual decrease in PM2.5 concentrations throughout the day, with the highest concentration during the early morning '0-6' interval (61.45) and the lowest concentration during the evening '18-24' interval (49.86).
- Friday exhibits a similar pattern to Thursday, with slightly lower concentrations during the early morning and mid-morning intervals (60.31 and 62.62), and a slight increase during the afternoon and evening intervals (44.96 and 49.65).
- Saturday displays a consistent trend similar to Friday, with slightly lower concentrations during the early morning and mid-morning intervals (58.45 and 61.31), and a slight increase during the afternoon and evening intervals (45.51 and 53.05).
- Sunday shows relatively lower PM2.5 concentrations during the early morning '0-6' interval (56.72) and evening '18-24' interval (49.39). The concentrations increase during the mid-morning '6-12' interval (66.04) and slightly increase during the afternoon '12-18' interval (47.38).

2.5.4 Chennai



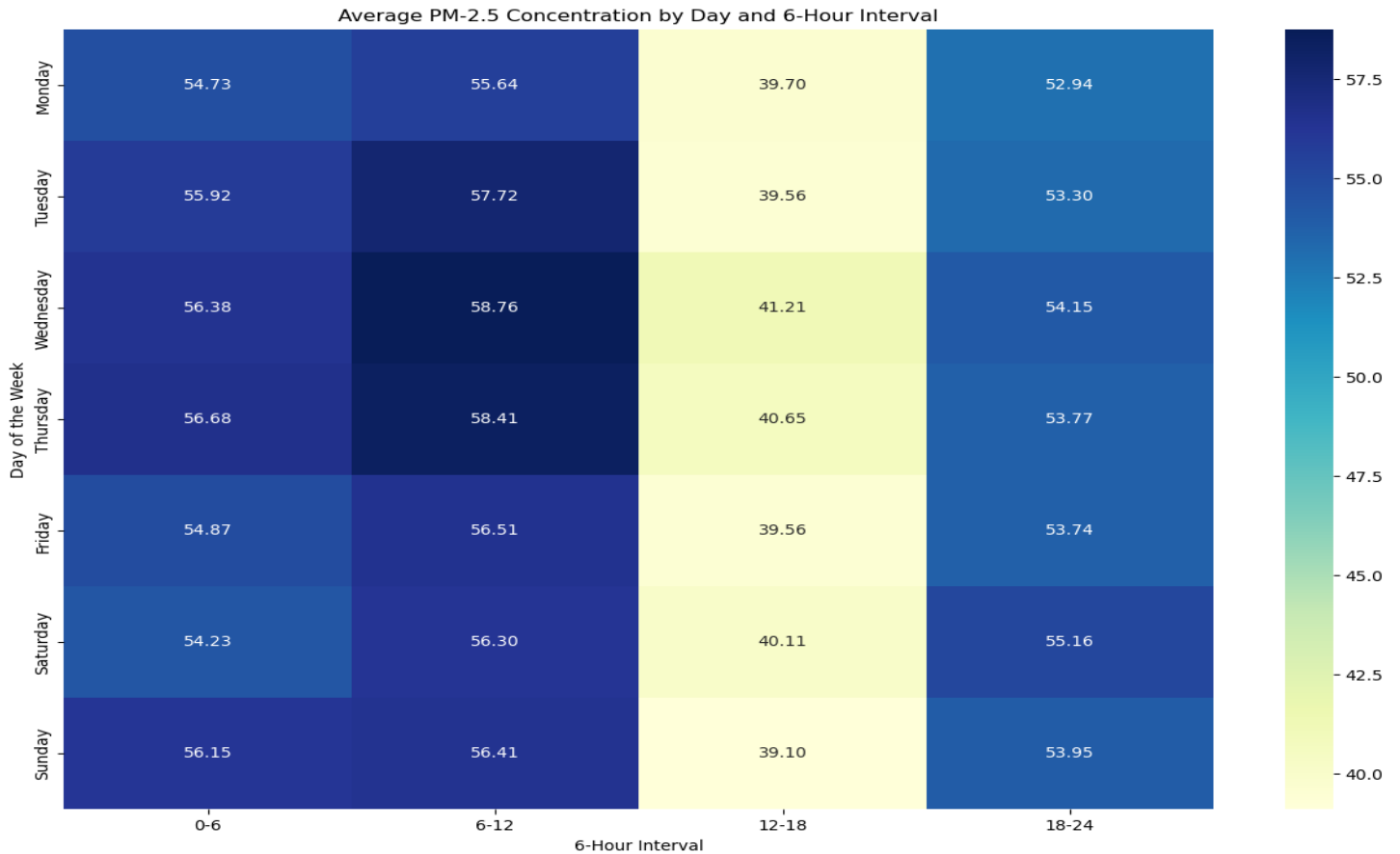
	0-6	6-12	12-18	18-24
Monday	31.68654	38.27907	30.47197	34.13246
Tuesday	32.21018	39.6252	32.39878	35.56781
Wednesday	31.05187	39.46035	30.91288	33.86803
Thursday	31.50247	39.80695	31.43492	33.58946
Friday	30.91725	40.26051	30.96635	34.43804
Saturday	32.25877	38.75405	31.85625	33.56851
Sunday	32.98601	38.69732	29.90736	34.27248

The descriptive table provides insights into the average PM_{2.5} concentrations in Chennai across different time intervals and days of the week:

- Monday shows relatively lower PM_{2.5} concentrations during the early morning '0-6' interval (31.69) and evening '18-24' interval (34.13). The concentrations increase during the mid-morning '6-12' interval (38.28) and slightly decrease during the afternoon '12-18' interval (30.47).

- Tuesday exhibits a similar pattern to Monday, with slightly higher concentrations during the early morning and mid-morning intervals (32.21 and 39.63), and a slight increase during the afternoon and evening intervals (32.40 and 35.57).
- Wednesday displays a consistent trend similar to Monday and Tuesday, with slightly lower concentrations during the early morning and mid-morning intervals (31.05 and 39.46), and a slight decrease during the afternoon and evening intervals (30.91 and 33.87).
- Thursday shows a relatively stable pattern with consistent PM2.5 concentrations throughout the day, with similar values during the early morning and mid-morning intervals (31.50 and 39.81), and a slight increase during the afternoon and evening intervals (31.43 and 33.59).
- Friday exhibits a similar pattern to Thursday, with slightly lower concentrations during the early morning and mid-morning intervals (30.92 and 40.26), and a slight increase during the afternoon and evening intervals (30.97 and 34.44).
- Saturday displays a consistent trend similar to Monday and Wednesday, with slightly higher concentrations during the early morning and mid-morning intervals (32.26 and 38.75), and a slight decrease during the afternoon and evening intervals (31.86 and 33.57).
- Sunday shows relatively lower PM2.5 concentrations during the early morning '0-6' interval (32.99) and afternoon '12-18' interval (29.91). The concentrations increase during the mid-morning '6-12' interval (38.70) and slightly increase during the evening '18-24' interval (34.27).

2.5.5 Hyderabad



	0-6	6-12	12-18	18-24
Monday	54.73191	55.63709	39.69679	52.94223
Tuesday	55.91786	57.72369	39.56371	53.29665
Wednesday	56.38399	58.76275	41.20963	54.14998
Thursday	56.67546	58.41485	40.65166	53.7681
Friday	54.87175	56.50683	39.56193	53.74182
Saturday	54.23296	56.29513	40.11192	55.1623
Sunday	56.15459	56.4118	39.10395	53.94707

The descriptive table provides insights into the average PM_{2.5} concentrations in New Delhi across different time intervals and days of the week:

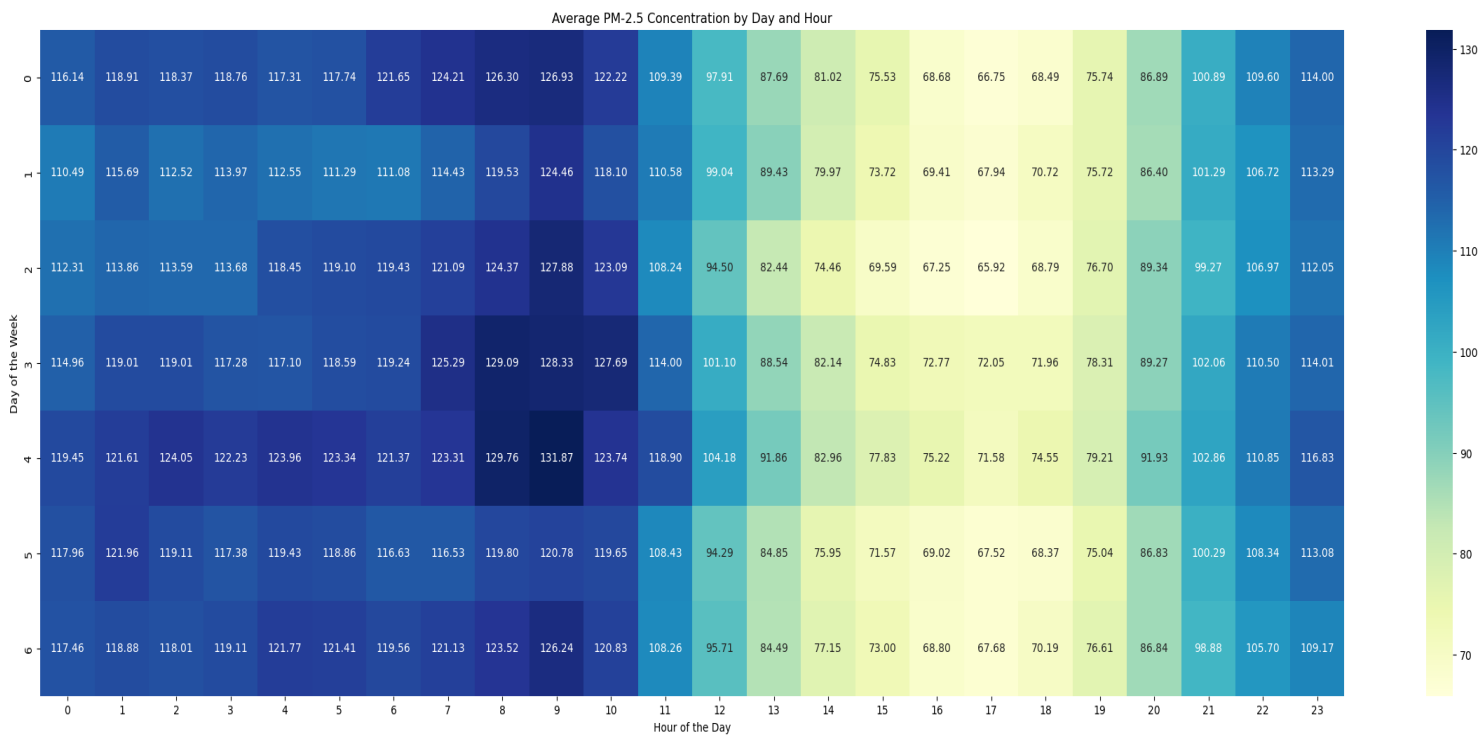
- Monday shows relatively lower PM_{2.5} concentrations during the early morning '0-6' interval (54.73) and evening '18-24' interval (52.94). The concentrations increase slightly during the mid-morning '6-12' interval (55.64) and significantly during the afternoon '12-18' interval (39.70).

- Tuesday exhibits a similar pattern to Monday, with slightly higher concentrations during the early morning and mid-morning intervals (55.92 and 57.72), and a slight decrease during the afternoon and evening intervals (39.56 and 53.30).
- Wednesday displays a consistent trend similar to Tuesday, with slightly higher concentrations during the early morning and mid-morning intervals (56.38 and 58.76), and a further increase during the afternoon and evening intervals (41.21 and 54.15).
- Thursday shows relatively stable PM2.5 concentrations throughout the day, with similar values during the early morning, mid-morning, and evening intervals (56.68, 58.41, and 53.77). There is a slight decrease during the afternoon '12-18' interval (40.65).
- Friday exhibits a similar pattern to Thursday, with slightly lower concentrations during the early morning and mid-morning intervals (54.87 and 56.51), and a slight increase during the afternoon and evening intervals (39.56 and 53.74).
- Saturday displays relatively stable PM2.5 concentrations throughout the day, with similar values during the early morning, mid-morning, and evening intervals (54.23, 56.30, and 55.16). There is a slight decrease during the afternoon '12-18' interval (40.11).
- Sunday shows relatively stable PM2.5 concentrations throughout the day, with similar values during the early morning, mid-morning, and evening intervals (56.15, 56.41, and 53.95). There is a slight decrease during the afternoon '12-18' interval (39.10).

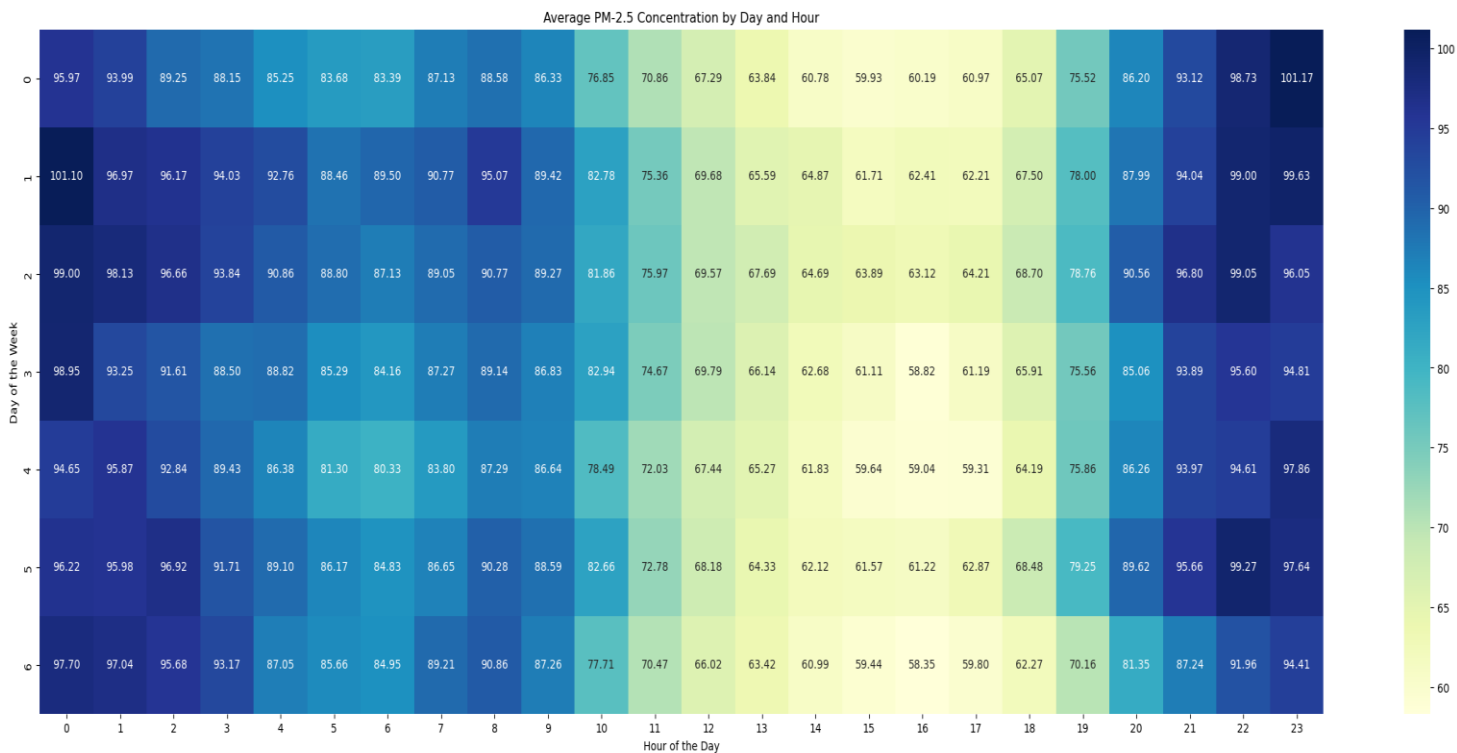
These insights provide a clear understanding of the variations in PM2.5 concentrations in New Delhi across different intervals and days of the week, helping to identify periods of higher pollution and potential patterns.

2.6 Average PM-2.5 Concentration by Day and Hour

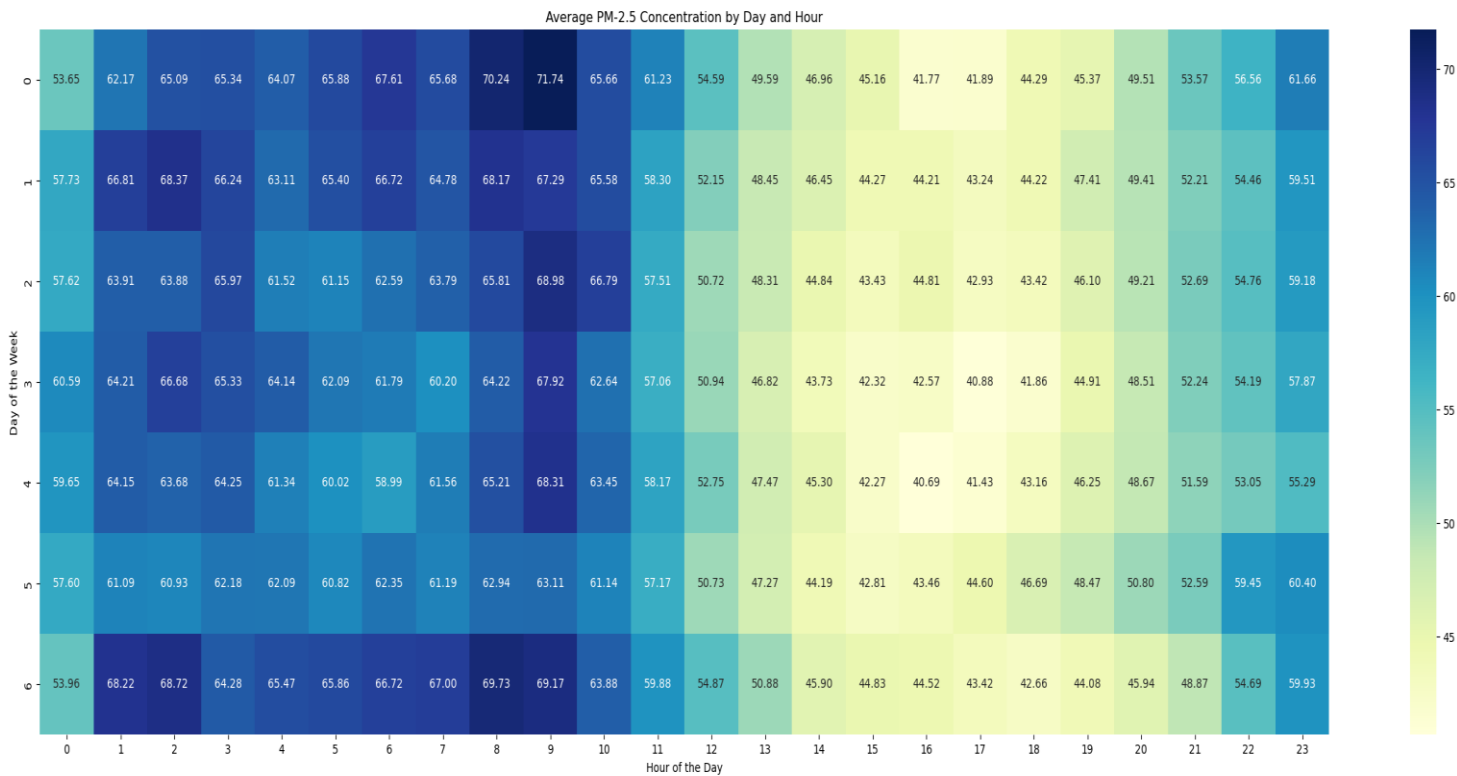
2.6.1 New Delhi



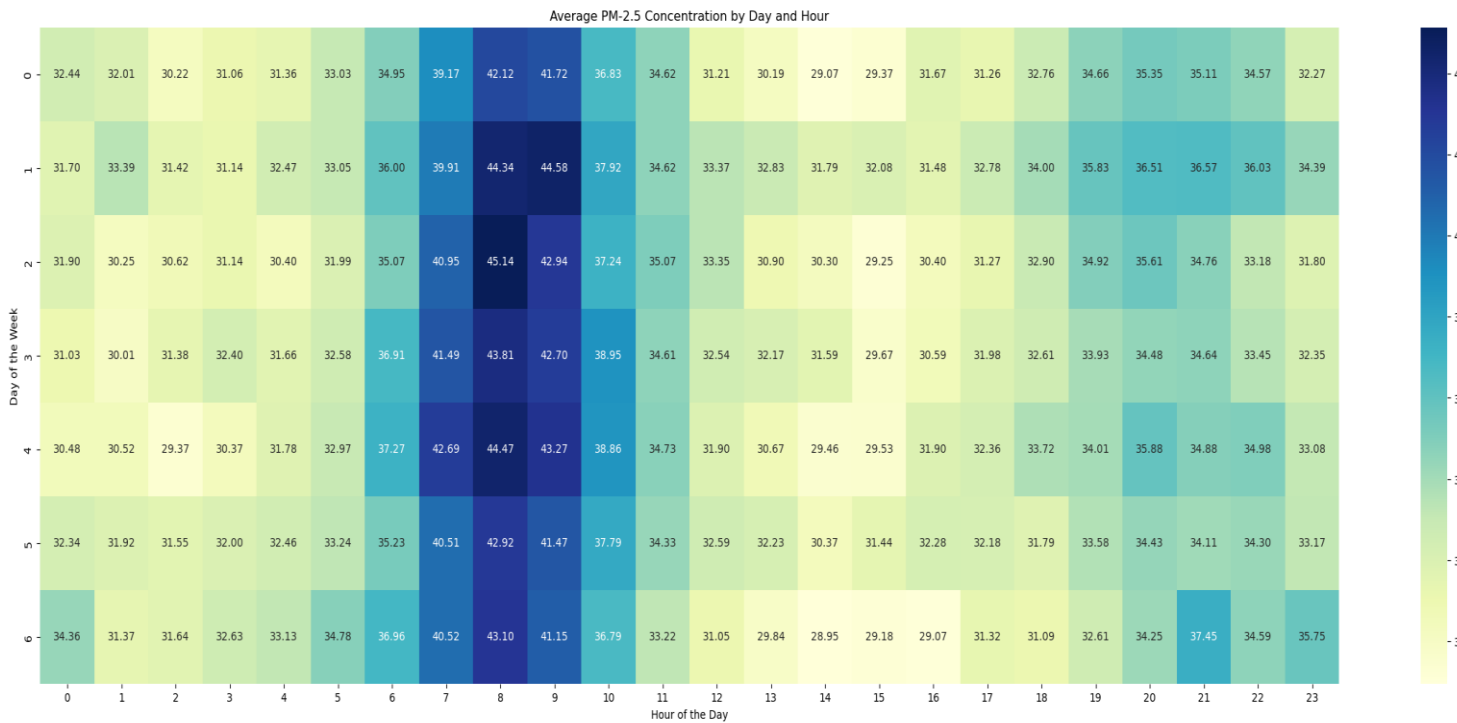
2.6.2 Kolkata



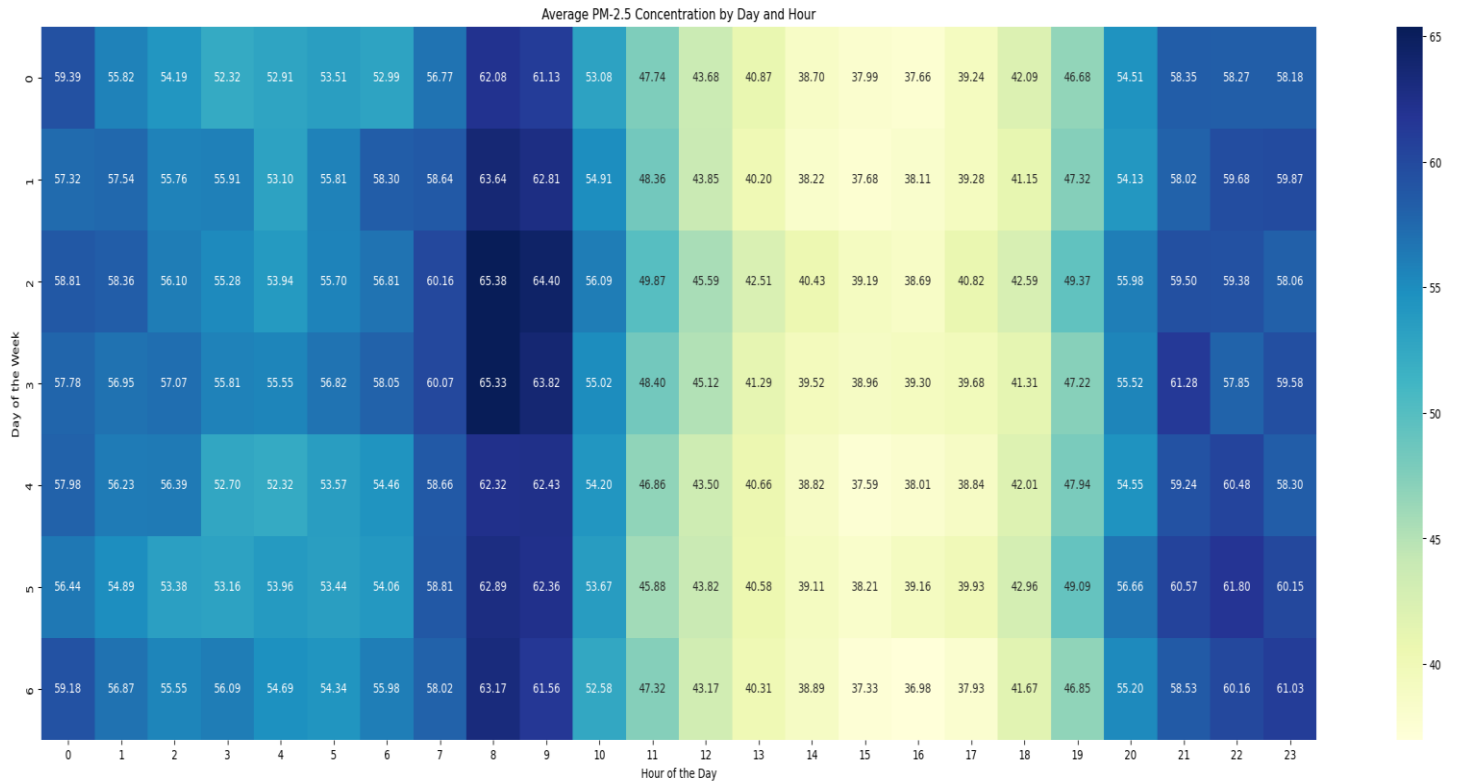
2.6.3 Mumbai



2.6.4 Chennai



2.6.5 Hyderabad



3 Results

3.1 Statistical tools

The Statistical tools to be used are as follows:

3.1.1 Normality tests

Normality tests are essential statistical tools used to assess the adequacy of representing a dataset with a normal distribution. These tests help determine the likelihood of a random variable underlying the dataset being normally distributed. There are two main types of normality tests: graphical methods and statistical methods.

3.1.1.1. Graphical Methods:

Probability Plots (Q-Q Plots): Probability plots, such as Q-Q plots, compare the dataset's quantiles to those of a theoretical normal distribution, providing a visual assessment of normality.

3.1.1.2. Statistical Methods:

- **Shapiro-Wilk Test:** The Shapiro-Wilk test calculates a test statistic and compares it to a critical value. If the test statistic exceeds the critical value, the null hypothesis of normality is rejected, indicating non-normality.
- Anderson-Darling Test, Kolmogorov-Smirnov Test, Lilliefors Test: These are other statistical tests used to assess normality, each with its own assumptions and limitations. The choice of test depends on factors like sample size and analysis requirements.

I) New Delhi

Test	Statistic	Df	Significance	Normality
Shapiro-wilk	0.851944	2464	0.0	Not Normal

Insights:

- The Shapiro-Wilk test statistic value of 0.851944 suggests that the dataset deviates significantly from a normal distribution.
- The p-value obtained from the Shapiro-Wilk test is 0.0, indicating strong evidence against the null hypothesis of normality.
- With a significance level of 0.05, we can confidently conclude that the dataset is not normally distributed.

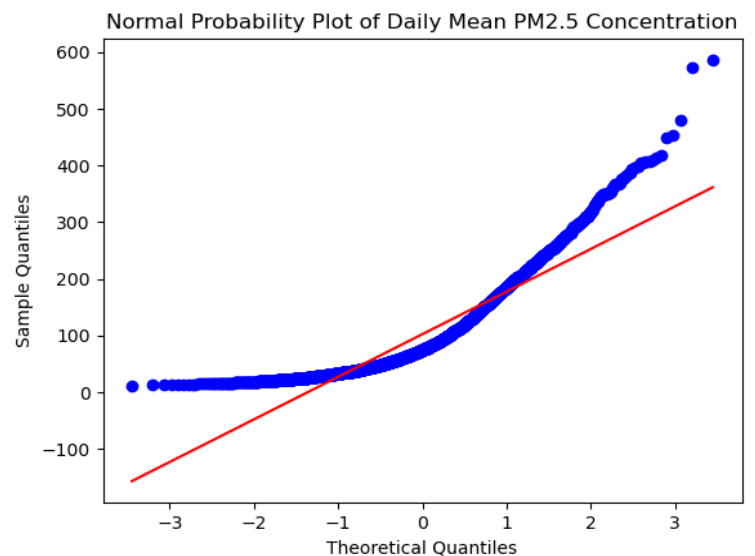


Fig 2: Normal Q-Q plot of New Delhi PM2.5 concentration level

II) Kolkata:

Test	Statistic	Df	Significance	Normality
Shapiro-wilk	0.842291	2423	0.0	Not Normal

Insights:

The Shapiro-Wilk test was conducted to assess the normality of the dataset, which consists of the daily mean PM2.5 concentration in Kolkata. The test yielded a test statistic of 0.842291, degrees of freedom (df) of 2423, and a significance value of 0.0. Based on these results, it can be concluded that the dataset is not normally distributed.

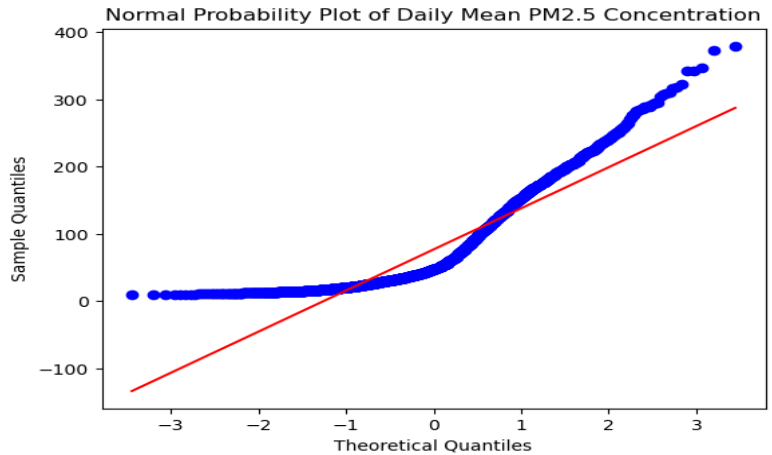


Fig:Normal Q-Q plot of Kolkata PM2.5 concentration level

III) Mumbai

Test	Statistic	Df	Significance	Normality
Shapiro-wilk	0.884064	2399	0.0	Not Normal

Insights:

- The Shapiro-Wilk test statistic value of 0.884064 suggests that the dataset deviates significantly from a normal distribution.
- The p-value obtained from the Shapiro-Wilk test is 0.0, indicating strong evidence against the null hypothesis of normality.
- With a significance level of 0.05, we can confidently conclude that the dataset is not normally distributed.

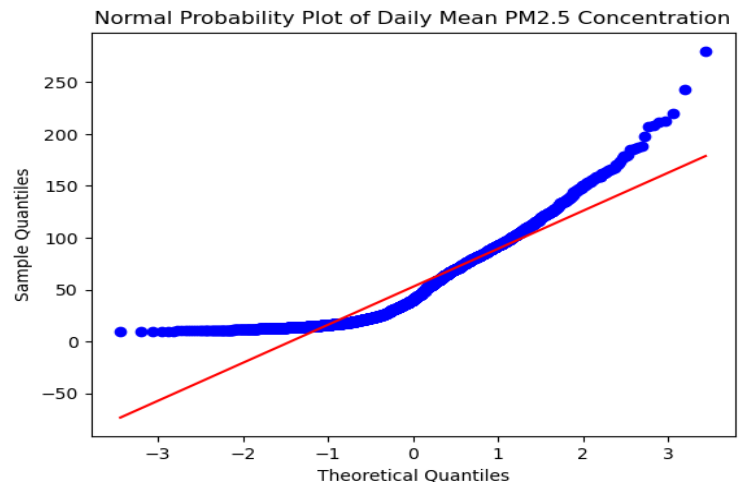


Fig:Normal Q-Q plot of Mumbai PM2.5 concentration level

IV) Chennai:

TEST	STATISTIC	DF	SIGNIFICANCE	NORMALITY
Shapiro-wilk	0.827992	2438	0.0	Not Normal

Insights:

The Shapiro-Wilk test was conducted to assess the normality of the dataset, which consists of the daily mean PM2.5 concentration in Kolkata. The test yielded a test statistic of 0.82799, degrees of freedom (df) of 2438, and a significance value of 0.0. Based on these results, it can be concluded that the dataset is not normally distributed.

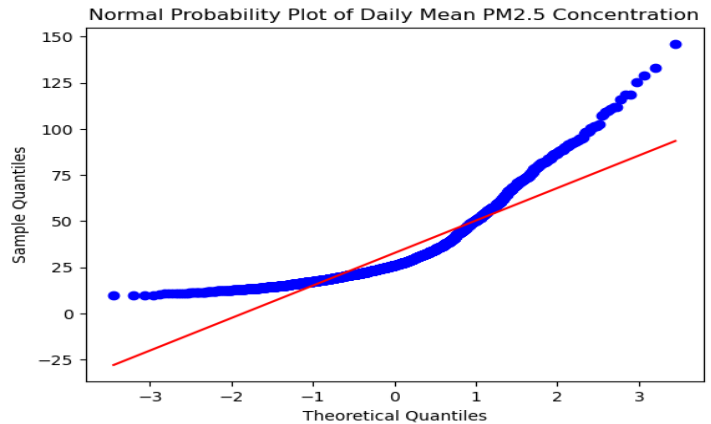


Fig:Normal Q-Q plot of Chennai PM2.5 concentration level

V)Hyderabad:

Test	Statistic	Df	Significance	Normality
Shapiro-wilk	0.942431	2471	0.0	Not Normal

Insights:

- The Shapiro-Wilk test statistic value of 0.942431 suggests that the dataset deviates significantly from a normal distribution.
- The p-value obtained from the Shapiro-Wilk test is 0.0, indicating strong evidence against the null hypothesis of normality.
- With a significance level of 0.05, we can confidently conclude that the dataset is not normally distributed.

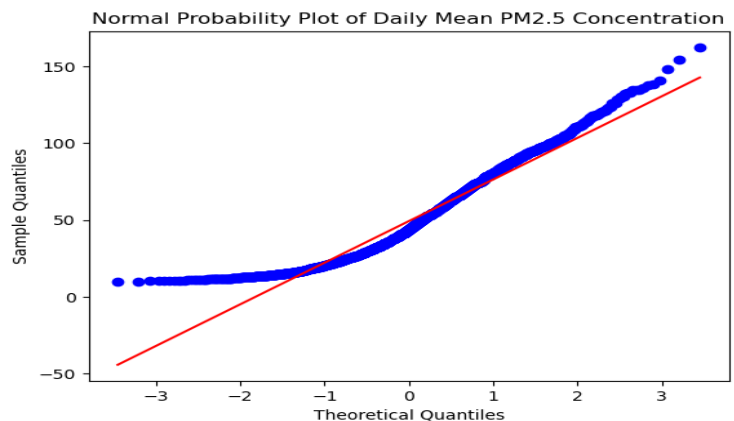


Fig:Normal Q-Q plot of Hyderabad PM2.5 concentration level

3.2 Wilcoxon signed-rank test:

The Wilcoxon signed-rank test is a non-parametric statistical test used to determine if there is a significant difference between paired observations. It is often used as an alternative to the paired Student's t-test when the data does not meet the assumptions of normality or when the sample size is small.

In this project, we employed the Wilcoxon signed-rank test to evaluate the differences in mean PM2.5 values between weekdays and weekends in a specific dataset. This test is appropriate for this type of analysis because it does not require the data to be normally distributed.

Dataset Description:

The dataset involved calculating mean PM2.5 values for weekdays and weekends across different years. We observed variations in mean PM2.5 values between weekdays and weekends, suggesting potential differences in pollution levels based on the day of the week. It should be noted that this dataset does not follow normal distribution.

Wilcoxon Signed-Rank Test Procedure:

1. Null Hypothesis (H_0): There is no significant difference between the mean PM2.5 values on weekdays and weekends.
2. Alternative Hypothesis (H_a): There is a significant difference between the mean PM2.5 values on weekdays and weekends.
3. Calculate the differences between the paired observations (weekday and weekend mean PM2.5 values).
4. Rank the absolute values of the differences obtained in Step 3, from smallest to largest.
5. Assign positive ranks to the observations with positive differences and negative ranks to the observations with negative differences.
6. Calculate the sum of the ranks for the positive differences (test statistic).
7. Determine the critical value or p-value associated with the test statistic. This can be obtained from a Wilcoxon signed-rank table or through statistical software.
8. Compare the obtained p-value to the chosen significance level (e.g., 0.05) to make a decision.

Interpretation of Results:

1. If the obtained p-value is less than the chosen significance level, reject the null hypothesis. This indicates a significant difference between weekday and weekend mean PM2.5 values.
2. If the obtained p-value is greater than the chosen significance level, fail to reject the null hypothesis. This suggests no significant difference between weekday and weekend mean PM2.5 values.

I).Delhi:

Mean PM2.5 values were calculated for weekdays and weekends in each year. The results show that there are variations in the mean PM2.5 values between weekdays and weekends across different years. For example, in 2016, the mean PM2.5 values were slightly higher on weekends compared to weekdays, while in 2019, the mean PM2.5 values were lower on weekends. To further investigate the significance of the difference between weekday and weekend mean PM2.5 values, a Wilcoxon signed-rank test was performed.

Date (LT)	Day	Raw Conc.
2016	Weekday	115.0594
2016	Weekend	115.6215
2017	Weekday	107.6861
2017	Weekend	105.6439
2018	Weekday	108.5641
2018	Weekend	103.3116
2019	Weekday	107.1221
2019	Weekend	94.35794
2020	Weekday	86.79371
2020	Weekend	95.92522
2021	Weekday	100.4709
2021	Weekend	101.8984
2022	Weekday	96.81894
2022	Weekend	94.65612

The test resulted in a test statistic of 9.0 and a p-value of 0.46875.

- With a significance level of 0.05, the p-value obtained (0.46875) is greater than the significance level. Therefore, we fail to reject the null hypothesis. This suggests that there is no significant difference between the mean PM2.5 values on weekdays and weekends.
- Overall, the analysis indicates that while there are variations in the mean PM2.5 values between weekdays and weekends across different years, the observed differences are not statistically significant according to the Wilcoxon signed-rank test.

II) Kolkata

Mean PM2.5 values were calculated for weekdays and weekends in each year. The results show that there are variations in the mean PM2.5 values between weekdays and weekends across different years.

Wilcoxon signed-rank test result:

Statistic: 12.0

P-value: 0.8125

Date (LT)	Day	Raw Conc.
2016	Weekday	85.13371
2016	Weekend	87.33149
2017	Weekday	78.92373
2017	Weekend	74.3686
2018	Weekday	91.13546
2018	Weekend	98.54471
2019	Weekday	74.67426
2019	Weekend	76.35222
2020	Weekday	84.91307
2020	Weekend	76.16354
2021	Weekday	76.94238
2021	Weekend	74.90459
2022	Weekday	77.9749
2022	Weekend	77.88352

With a significance level of 0.05, the obtained p-value (0.8125) is greater than the significance level.

- Consequently, the null hypothesis cannot be rejected, indicating that there is no significant difference between the mean PM2.5 values on weekdays and weekends.
- In summary, while variations in mean PM2.5 values between weekdays and weekends were observed across different years, the Wilcoxon signed-rank test did not indicate a statistically significant difference. These findings suggest that weekdays and weekends exhibit similar levels of PM2.5 pollution in Kolkata.

III) Mumbai

Mean PM2.5 values were calculated for weekdays and weekends in each year. The results show that there are variations in the mean PM2.5 values between weekdays and weekends across different years.

Wilcoxon signed-rank test result:

Statistic: 9.0

P-value: 0.46875

Date (LT)	Day	Raw Conc.
2016	Weekday	58.8882
2016	Weekend	59.22395
2017	Weekday	71.71581
2017	Weekend	69.80159
2018	Weekday	77.61576
2018	Weekend	72.61388
2019	Weekday	47.78239
2019	Weekend	50.57694
2020	Weekday	48.35589
2020	Weekend	49.21941
2021	Weekday	46.89939
2021	Weekend	45.18892
2022	Weekday	48.80366
2022	Weekend	46.35012

- The test result shows that the p-value is 0.46875, which is greater than the significance level of 0.05. Therefore, we cannot reject the null hypothesis that there is no significant difference between the mean PM-2.5 values for weekdays and weekends.
- This means that there is no evidence to suggest that there is a significant difference in air pollution levels between weekdays and weekends in Mumbai based on the PM-2.5 data we have analyzed.

IV) Chennai:

Mean PM2.5 values were calculated for weekdays and weekends in each year. The results show that there are variations in the mean PM2.5 values between weekdays and weekends across different years. To assess the significance of the difference between weekday and weekend mean PM2.5 values, a Wilcoxon signed-rank test was performed.

Wilcoxon signed-rank test result:

Statistic: 11.0

P-value: 0.6875

Date (LT)	Day	Raw Conc.
2016	Weekday	42.08722
2016	Weekend	42.01018
2017	Weekday	35.79795
2017	Weekend	34.68302
2018	Weekday	33.48348
2018	Weekend	35.08693
2019	Weekday	31.30081
2019	Weekend	32.09493
2020	Weekday	28.98758
2020	Weekend	30.29854
2021	Weekday	33.55544
2021	Weekend	28.90217
2022	Weekday	34.54135
2022	Weekend	36.3476

- The test result shows that the p-value is 0.6875, which is greater than the significance level of 0.05. Therefore, we cannot reject the null hypothesis that there is no significant difference between the mean PM-2.5 values for weekdays and weekends.
- This means that there is no evidence to suggest that there is a significant difference in air pollution levels between weekdays and weekends in Mumbai based on the PM-2.5 data.

V) Hyderabad

Mean PM2.5 values were calculated for weekdays and weekends in each year. The results show that there are variations in the mean PM2.5 values between weekdays and weekends across different years. To assess the significance of the difference between weekday and weekend mean PM2.5 values, a Wilcoxon signed-rank test was performed.

Wilcoxon signed-rank test result:

Statistic: 10.0

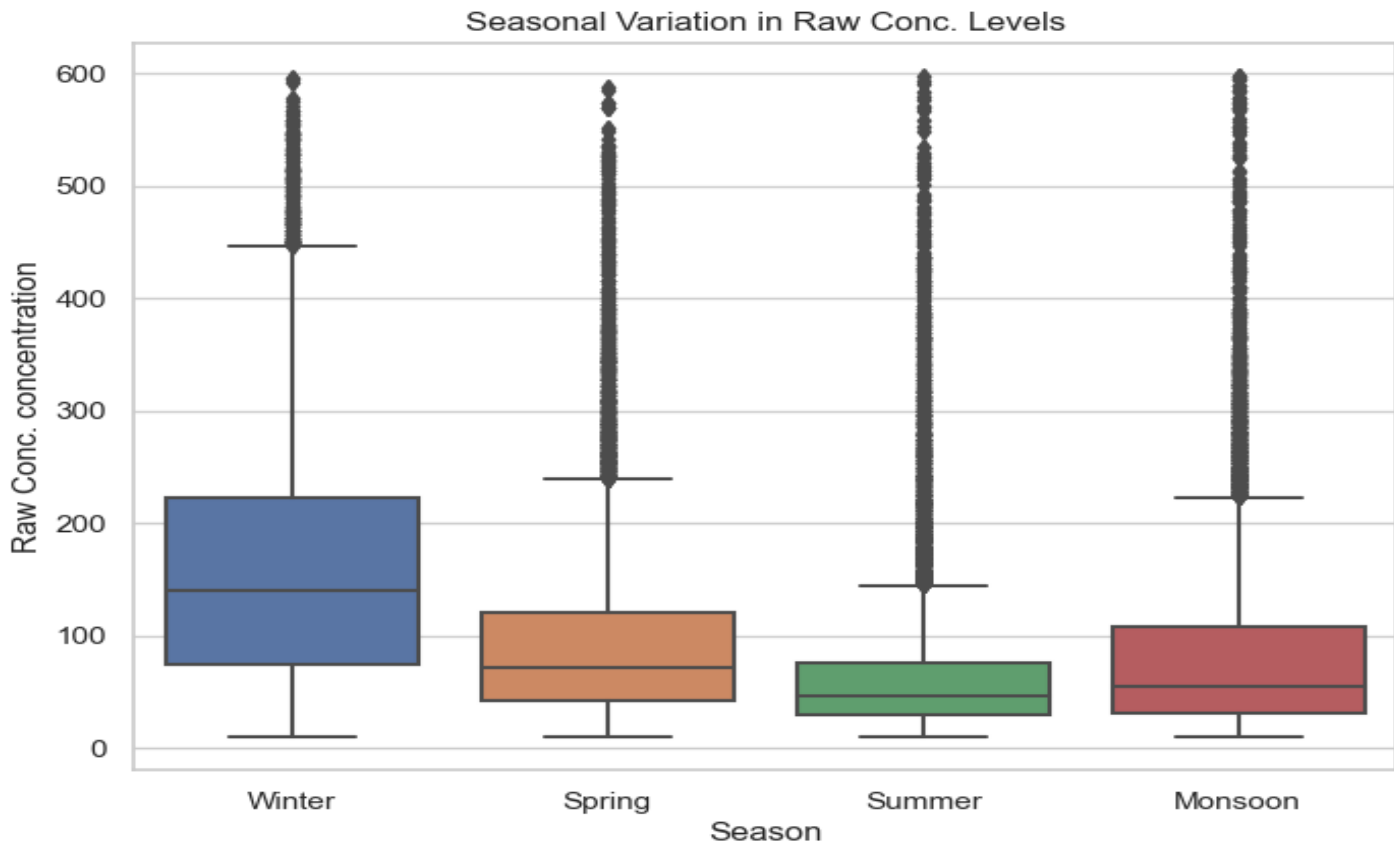
P-value: 0.578125

Date (LT)	Day	Raw Conc.
2016	Weekday	56.35347
2016	Weekend	57.10763
2017	Weekday	58.87237
2017	Weekend	59.98908
2018	Weekday	60.72611
2018	Weekend	59.29832
2019	Weekday	46.73917
2019	Weekend	45.59604
2020	Weekday	41.25376
2020	Weekend	43.66568
2021	Weekday	47.89101
2021	Weekend	47.1337
2022	Weekday	49.40634
2022	Weekend	45.87196

- With a significance level of 0.05, the p-value obtained (0.578125) is greater than the significance level. Therefore, we fail to reject the null hypothesis. This suggests that there is no significant difference between the mean PM2.5 values on weekdays and weekends.
- This means that there is no evidence to suggest that there is a significant difference in air pollution levels between weekdays and weekends in Mumbai based on the PM-2.5 data we have analyzed.

3.3 Seasonal Variation

3.3.1 New Delhi



Season	count	mean	std	min	25%	50%	75%	max
Monsoon	14183	83.54481	78.64267	10	31	55	108	597
Spring	13609	95.15304	77.69686	10	42	72	121	587
Summer	14262	69.35984	69.70258	10	30	47	76	597
Winter	14891	159.7758	106.6418	10	74	141	223	596

Interpretation of Statistics:

Monsoon Season: The mean concentration of PM-2.5 during the Monsoon season is 83.54, with a standard deviation of 78.64. The values range from a minimum of 10 to a maximum of 597. This indicates that PM-2.5 levels during the Monsoon season vary widely, with an average concentration of 83.54.

Spring Season: During the Spring season, the mean PM-2.5 concentration is 95.15, with a standard deviation of 77.70. The values range from a minimum of 10 to a maximum of 587. This suggests that the PM-2.5 levels in Spring are slightly higher on average compared to the Monsoon season.

Summer Season: The average PM-2.5 concentration during the Summer season is 69.36, with a standard deviation of 69.70. The values range from a minimum of 10 to a maximum of 597. This indicates relatively lower levels of PM-2.5 during the Summer season compared to the other seasons.

Winter Season: The Winter season has the highest average PM-2.5 concentration among the seasons, with a mean of 159.78 and a standard deviation of 106.64. The values range from a minimum of 10 to a maximum of 596. This suggests that the Winter season experiences the highest levels of PM-2.5 pollution on average.

These insights provide an overview of the PM-2.5 concentrations in different seasons in New Delhi based on the available data, indicating the varying levels of pollution throughout the year.

Based on the analysis of the PM-2.5 data in New Delhi, the following insights can be derived from the statistical tests:

The Kruskal-Wallis test:

The Kruskal-Wallis test is a non-parametric test that is used to compare two or more independent groups on a continuous or ordinal dependent variable. It is considered the nonparametric alternative to the one-way ANOVA, and an extension of the Mann-Whitney U test to allow the comparison of more than two independent groups.

The Dunn's test:

The Dunn's test was used to conduct pairwise comparisons between the different seasons. The results of the Dunn's test showed that all of the comparisons between seasons resulted in extremely low p-values, indicating significant differences in PM-2.5 pollution levels.

In general, the Dunn's test is a good choice for comparing two groups after a Kruskal-Wallis test has found a statistically significant difference between the groups. The Duan test is a good choice for comparing two groups when the data is not normally distributed.

Kruskal-Wallis Test:

p-value: 6.671668876616056e-54

The Kruskal-Wallis test yielded a significant test statistic and an extremely low p-value. This indicates that there are significant differences in the PM-2.5 pollution levels among the seasons in New Delhi.

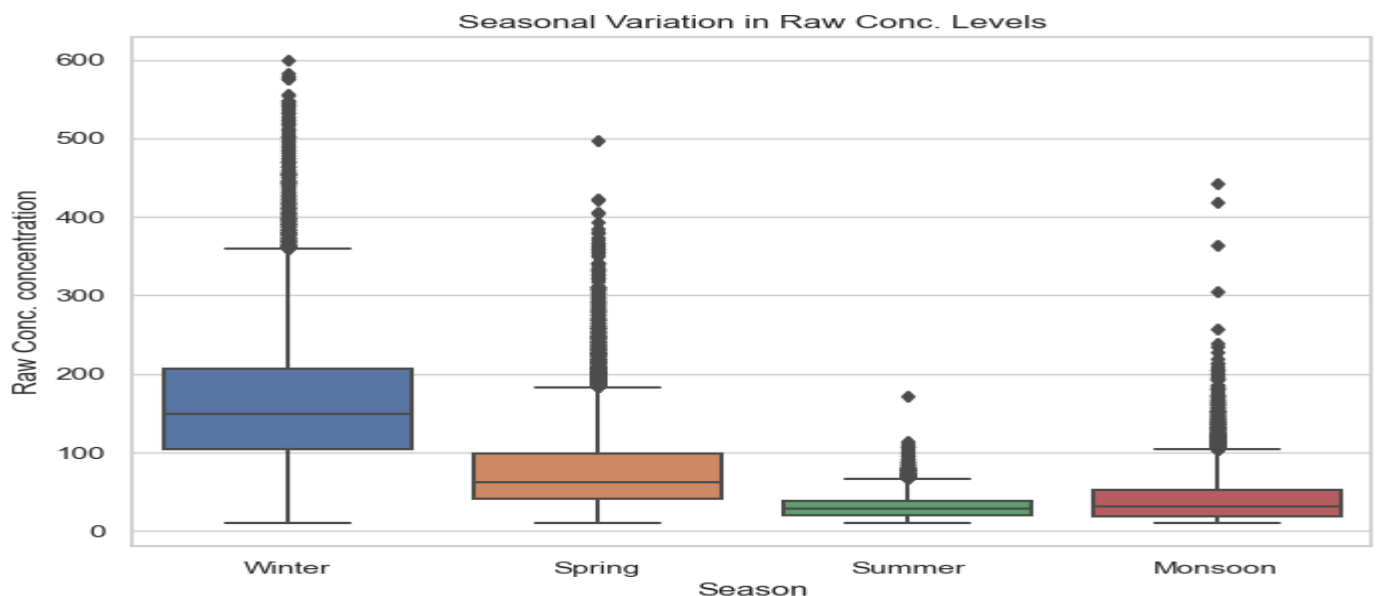
Dunn's Test Results:

- Winter vs Spring: p-value = 3.3171413071420265e-194
- Winter vs Summer: p-value = 0.0
- Winter vs Monsoon: p-value = 0.0
- Spring vs Summer: p-value = 0.0
- Spring vs Monsoon: p-value = 0.0
- Summer vs Monsoon: p-value = 6.671668876616056e-54

The pairwise post hoc tests using the Dunn's Test revealed that all the comparisons between seasons resulted in extremely low p-values, indicating significant differences in PM-2.5 pollution levels.

Overall, the statistical analysis confirms that there are significant differences in PM-2.5 pollution levels among the seasons in New Delhi. The results suggest that the pollution levels vary significantly depending on the time of the year.

3.3.2 Kolkata



Season	count	mean	std	min	25%	50%	75%	max
Monsoon	11551	40.70297	30.73253	10	19	31	53	442
Spring	14135	77.56944	53.97793	10	41	62	98	497
Summer	12566	30.4427	13.98403	10	20	28	39	172
Winter	14228	162.6832	82.91936	10	104	149	206	599

1. Monsoon Season: During the Monsoon season, the mean PM-2.5 concentration is 40.70, with a standard deviation of 30.73. The minimum PM-2.5 concentration recorded is 10, while the maximum is 442. The 25th percentile of the data is 19, the median (50th percentile) is 31, and the 75th percentile is 53. These statistics indicate that PM-2.5 levels during the Monsoon season generally range from 10 to 442, with an average concentration of around 40.70.

2. Spring Season: In the Spring season, the average PM-2.5 concentration is 77.57, with a standard deviation of 53.98. The minimum concentration observed is 10, while the maximum is 497. The 25th percentile is 41, the median is 62, and the 75th percentile is 98. These statistics suggest that PM-2.5 levels during Spring exhibit a wider range, with an average concentration of approximately 77.57.

3. Summer Season: During the Summer season, the mean PM-2.5 concentration is 30.44, with a standard deviation of 13.98. The minimum recorded value is 10, and the maximum is 172. The 25th percentile is 20, the median is 28, and the 75th percentile is 39. These statistics indicate relatively lower levels of PM-2.5 during the Summer season, with an average concentration of around 30.44.

4. Winter Season: The Winter season has the highest average PM-2.5 concentration among the seasons, with a mean of 162.68 and a standard deviation of 82.92. The minimum PM-2.5 concentration observed is 10, while the maximum is 599. The 25th percentile is 104, the median is 149, and the 75th percentile is 206. These statistics suggest that the Winter season experiences the highest levels of PM-2.5 pollution, with an average concentration of approximately 162.68.

These insights provide an overview of the PM-2.5 concentrations in different seasons in an undisclosed location based on the available data.

Based on the analysis of the PM-2.5 data in Kolkata, the following insights can be derived from the statistical tests:

i) Kruskal-Wallis Test:

p-value: 9.388092166819055e-63

The Kruskal-Wallis test yielded a significant test statistic and an extremely low p-value. This indicates that there are significant differences in the PM-2.5 pollution levels among the seasons in Kolkata.

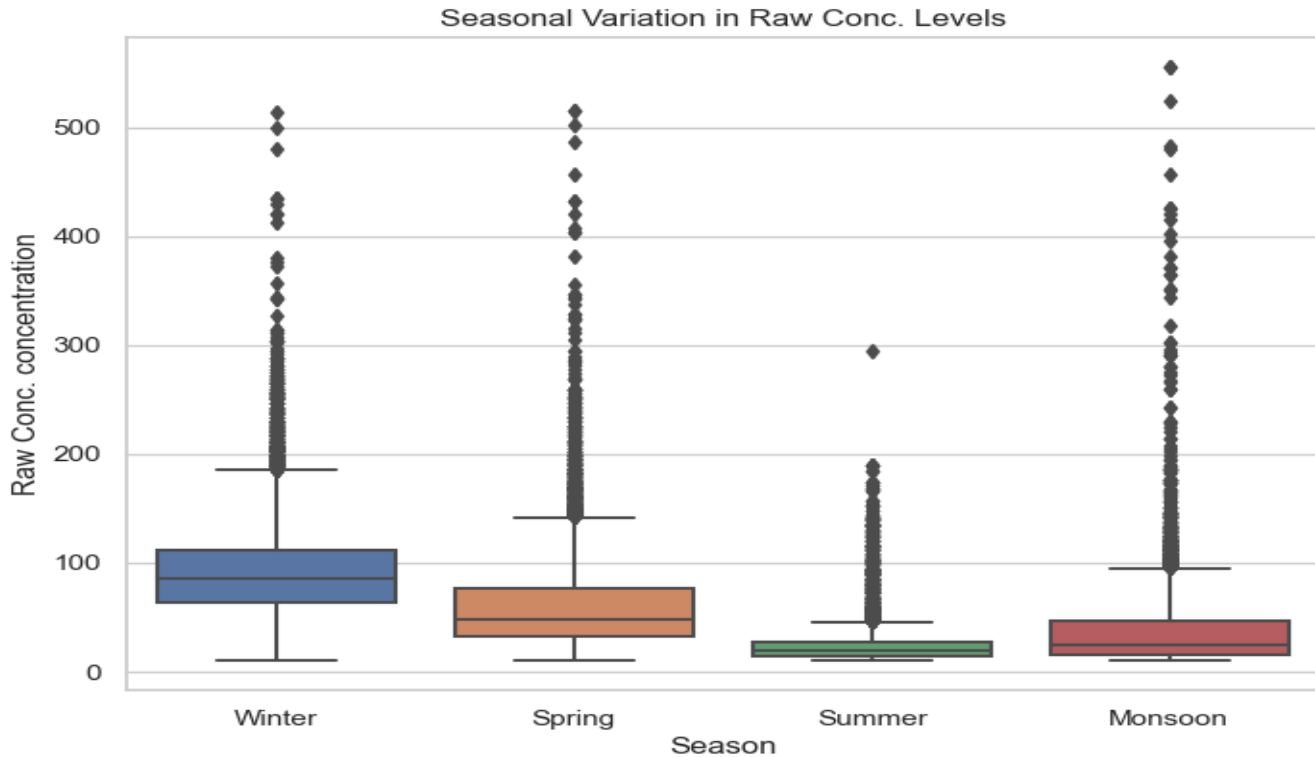
ii) Dunn's Test Results:

- Winter vs Spring: p-value = 0.0
- Winter vs Summer: p-value = 0.0
- Winter vs Monsoon: p-value = 0.0
- Spring vs Summer: p-value = 0.0
- Spring vs Monsoon: p-value = 0.0
- Summer vs Monsoon: p-value = 9.388092166819055e-63

The pairwise post hoc tests using the Dunn's Test revealed that all the comparisons between seasons resulted in extremely low p-values, indicating significant differences in PM-2.5 pollution levels.

Overall, the statistical analysis confirms that there are significant differences in PM-2.5 pollution levels among the seasons in Kolkata. The results suggest that the pollution levels vary significantly depending on the time of the year. All pairwise comparisons resulted in extremely low p-values, indicating significant differences in pollution levels between seasons.

3.3.3 Mumbai



Here are the insights from the table:

- Monsoon Season:** The PM-2.5 concentration during the Monsoon season has a mean of 36.41, with a standard deviation of 34.57. The values range from a minimum of 10 to a maximum of 555.
- Spring Season:** During the Spring season, the average PM-2.5 concentration is 60.47, with a standard deviation of 41.24. The values range from a minimum of 10 to a maximum of 515.

3. **Summer Season:** The Summer season exhibits a lower average PM-2.5 concentration of 24.16, with a standard deviation of 19.04. The values range from a minimum of 10 to a maximum of 294.

4. **Winter Season:** The Winter season has the highest average PM-2.5 concentration with a mean of 92.28, and a standard deviation of 44.32. The values range from a minimum of 10 to a maximum of 514.

These insights provide a summary of the PM-2.5 concentrations in different seasons, indicating the varying pollution levels throughout the year based on the available data.

i)Kruskal-Wallis Test:

p-value: 5.419714226577545e-190

The Kruskal-Wallis test yielded a significant test statistic and an extremely low p-value. This result indicates that there are significant differences in the PM-2.5 pollution levels among the seasons in Mumbai.

ii) Dunn's Test Results:

Winter vs Spring: p-value = 0.0

Winter vs Summer: p-value = 0.0

Winter vs Monsoon: p-value = 0.0

Spring vs Summer: p-value = 0.0

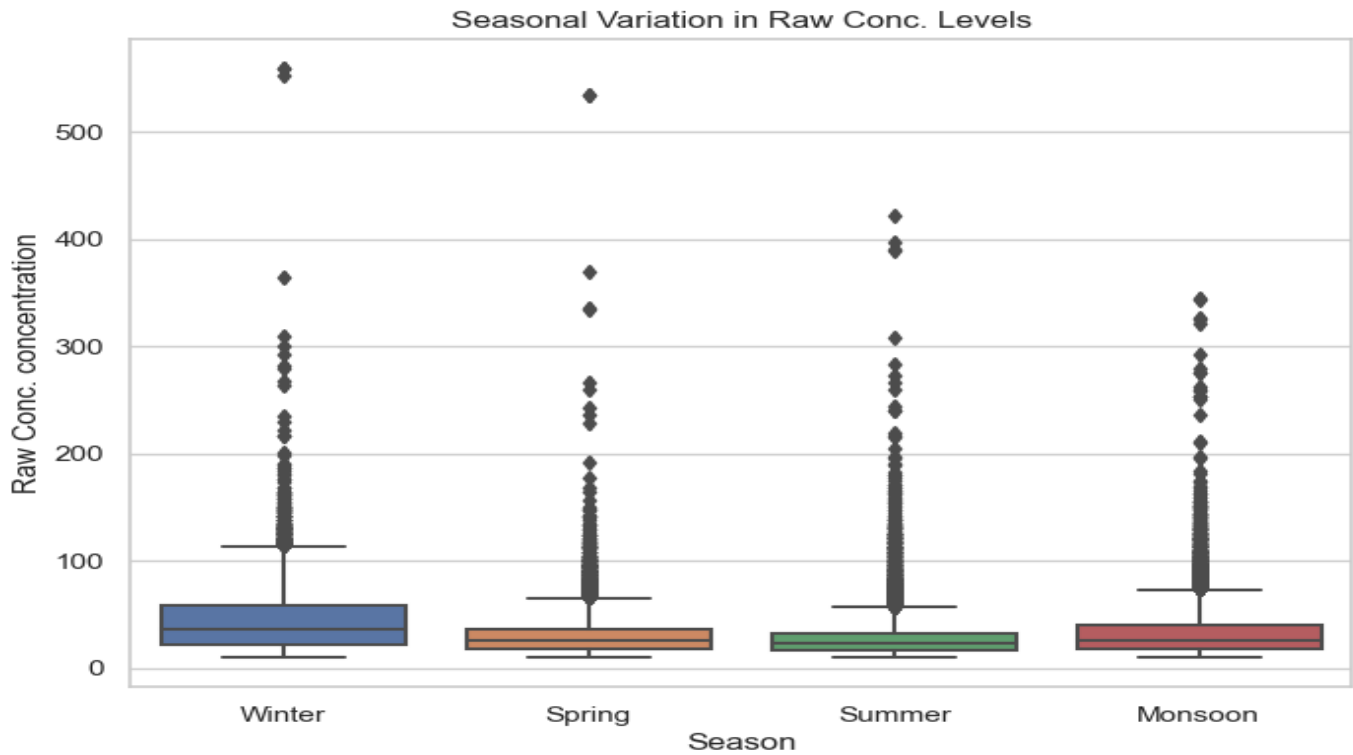
Spring vs Monsoon: p-value = 0.0

Summer vs Monsoon: p-value = 5.419714226577545e-190

The pairwise post hoc tests using the Dunn's Test revealed that all the comparisons between seasons resulted in extremely low p-values. This implies that there are significant differences in the PM-2.5 pollution levels between each pair of seasons.

Overall, the statistical analysis confirms that there are significant differences in PM-2.5 pollution levels among the seasons in Mumbai. The results suggest that the pollution levels vary significantly depending on the time of the year, with all pairwise comparisons showing significant differences in pollution levels.

3.3.4 Chennai



Season	count	mean	std	min	25%	50%	75%	max
Monsoon	12501	33.5435	24.8565	10	18	26	40	345
Spring	12130	30.6277	20.5310	10	18	26	37	534
Summer	12858	28.7176	21.1298	10	17	24	33	422
Winter	12574	43.8558	29.5171	10	22	37	59	559

Here are the insights from the table:

- The highest average PM-2.5 concentration is in winter, followed by spring, summer, and monsoon.
- The standard deviation is highest in winter, followed by spring, summer, and monsoon. This indicates that there is more variation in PM-2.5 concentrations in winter than in the other seasons.
- The minimum PM-2.5 concentration is the same in all four seasons.

- The 25th percentile PM-2.5 concentration is lowest in winter and highest in summer.
- The 50th percentile PM-2.5 concentration is lowest in winter and highest in summer.
- The 75th percentile PM-2.5 concentration is lowest in winter and highest in summer.
- The maximum PM-2.5 concentration is highest in winter and lowest in monsoon.

These insights suggest that PM-2.5 concentrations are highest in winter and lowest in monsoon in Chennai. The reasons for this are not clear, but it may be due to factors such as weather patterns, agricultural practices, and industrial emissions.

Based on the analysis of the PM-2.5 data in Chennai, the following insights can be derived from the statistical tests:

Kruskal-Wallis Test:

p-value: 1.1925504427547453e-51

The Kruskal-Wallis test yielded a significant test statistic and an extremely low p-value. This indicates that there are significant differences in the PM-2.5 pollution levels among the seasons in Chennai.

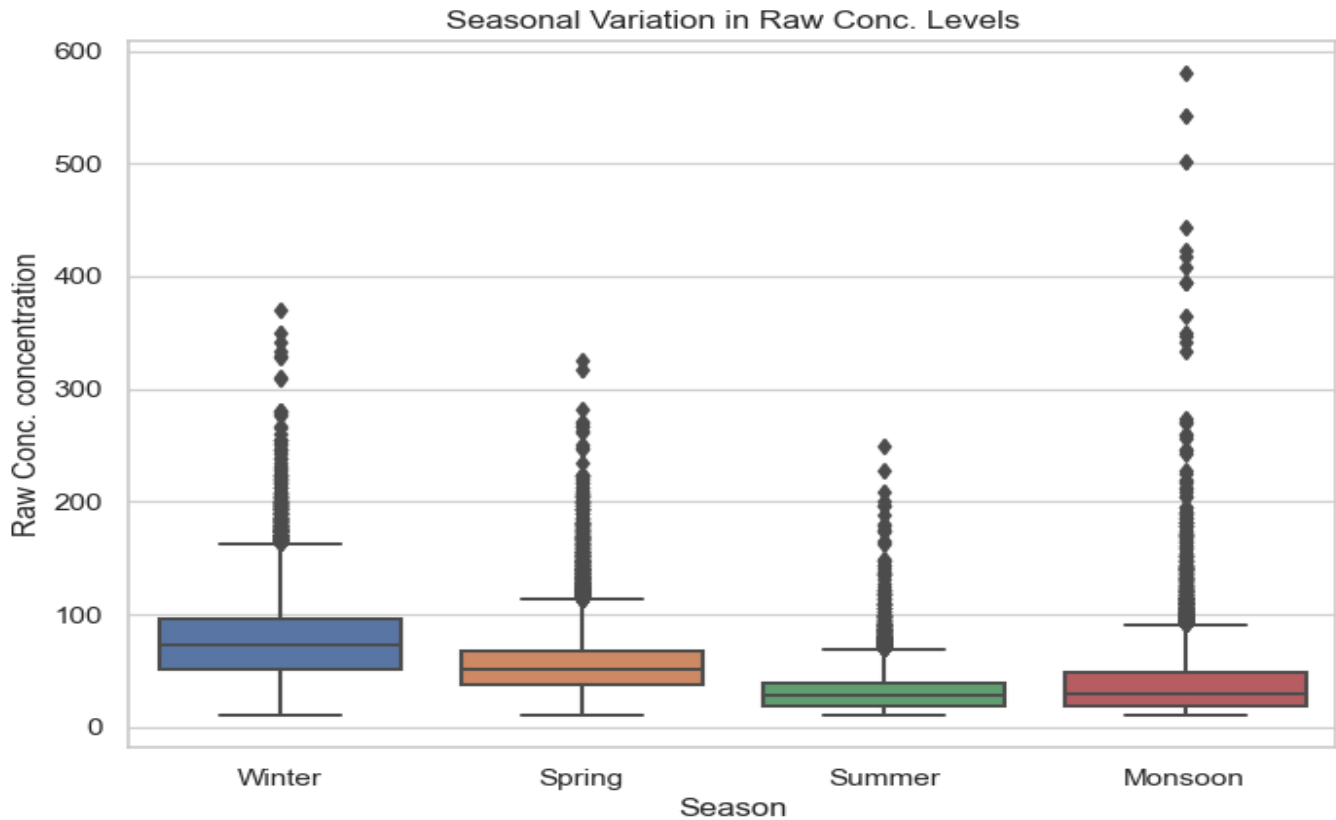
Dunn's Test Results:

- Winter vs Spring: p-value = 0.0
- Winter vs Summer: p-value = 0.0
- Winter vs Monsoon: p-value = 3.091633560214505e-246
- Spring vs Summer: p-value = 9.64956944688156e-31
- Spring vs Monsoon: p-value = 1.0664187217811742e-05
- Summer vs Monsoon: p-value = 1.1925504427547453e-51

The pairwise post hoc tests using the Dunn's Test revealed that all the comparisons between seasons resulted in extremely low p-values, indicating significant differences in PM-2.5 pollution levels.

Overall, the statistical analysis confirms that there are significant differences in PM-2.5 pollution levels among the seasons in Chennai. The results suggest that the pollution levels vary significantly depending on the time of the year. All pairwise comparisons resulted in extremely low p-values, indicating significant differences in pollution levels between seasons.

3.3.5 Hyderabad



Season	count	mean	std	min	25%	50%	75%	max
Monsoon	12024	37.69212	29.56512	10	18.75	29	48	580
Spring	14228	56.05096	28.23254	10	38	51	68	325
Summer	12797	31.21693	17.1786	10	19	28	39	249
Winter	14837	76.16405	35.18687	10	51	73	96	370

Based on the table, we can see that:

- The highest average PM-2.5 concentration is in winter, followed by summer, spring, and monsoon.
- The standard deviation is highest in winter, followed by summer, spring, and monsoon. This indicates that there is more variation in PM-2.5 concentrations in winter than in the other seasons.
- The minimum PM-2.5 concentration is the same in all four seasons.

- The 25th percentile PM-2.5 concentration is lowest in winter and highest in summer.
- The 50th percentile PM-2.5 concentration is lowest in winter and highest in summer.
- The 75th percentile PM-2.5 concentration is lowest in winter and highest in summer.
- The maximum PM-2.5 concentration is highest in winter and lowest in monsoon.

These insights suggest that PM-2.5 concentrations are highest in winter and lowest in monsoon in Hyderabad. The reasons for this are not clear, but it may be due to factors such as weather patterns, agricultural practices, and industrial emissions.

Based on the analysis of the PM-2.5 data in Hyderabad, the following insights can be derived from the statistical tests:

i) Kruskal-Wallis Test:

p-value: 1.1588509605919164e-28

The Kruskal-Wallis test yielded a significant test statistic and an extremely low p-value. This indicates that there are significant differences in the PM-2.5 pollution levels among the seasons in Hyderabad.

ii) Dunn's Test Results:

- Winter vs Spring: p-value = 0.0
- Winter vs Summer: p-value = 0.0
- Winter vs Monsoon: p-value = 0.0
- Spring vs Summer: p-value = 0.0
- Spring vs Monsoon: p-value = 0.0
- Summer vs Monsoon: p-value = 1.1588509605919164e-28

The pairwise post hoc tests using the Dunn's Test revealed that all the comparisons between seasons resulted in extremely low p-values, indicating significant differences in PM-2.5 pollution levels.

Overall, the statistical analysis confirms that there are significant differences in PM-2.5 pollution levels among the seasons in Hyderabad. The results suggest that the pollution levels vary significantly depending on the time of the year. All pairwise comparisons resulted in extremely low p-values, indicating significant differences in pollution levels between seasons.

4. Conclusions

This comprehensive analysis of ambient PM_{2.5} levels in Indian metropolitan cities from 2016 to 2022 has provided valuable insights into the air quality situation and emphasized the urgent need for targeted actions and interventions to combat air pollution and improve overall air quality. The study highlights the importance of addressing data quality issues, improving data collection systems, and utilizing effective visualization techniques to inform public health interventions.

The findings demonstrate the non-normal distribution of PM_{2.5} concentration data across the cities, emphasizing the necessity for tailored measures and interventions to address air pollution and enhance air quality. Moreover, the lack of significant differences in mean PM_{2.5} values between weekdays and weekends underscores the importance of considering non-normality and employing appropriate statistical tests when analyzing air pollution data.

The analysis also demonstrates distinct patterns in PM_{2.5} concentrations across different days of the week and time intervals in Delhi, Kolkata, Mumbai, Chennai, and Hyderabad. Each city exhibits unique characteristics, highlighting the need for tailored pollution control strategies. Additionally, significant seasonal variations in pollution levels are observed, with winter consistently showing the highest pollution levels and summer the lowest.

While variations in mean PM_{2.5} values are observed between weekdays and weekends in multiple cities, statistical tests indicate no statistically significant difference in air pollution levels between these two categories. This suggests that pollution control efforts should not focus solely on weekdays but should consider comprehensive measures for reducing pollution throughout the week.

Overall, this analysis provides valuable insights into the air quality situation in Indian metropolitan cities, emphasizing the urgency for targeted actions, considering seasonal variations, and adopting a holistic approach to combat air pollution and improve public health.

5.References :

I)Reference from Journal

1. Wilmar Hernandez; Alfredo Mendez; Angela Maria Diaz-Marquez; Rasa Zalakeviciute, "PM_{2.5} Concentration Measurement Analysis by Using Non-Parametric Statistical Inference".
2. Alexis Dinno, Nonparametric pairwise multiple comparisons in independent groups using Dunn's test, School of Community Health Portland State University Portland
3. Doreswamy, Harishkumar K S1, Yogesh KM, Ibrahim Gad , " Forecasting Air Pollution Particulate Matter (PM_{2.5}) Using Machine Learning Regression Models ", Third International Conference on Computing and Network Communications (CoCoNet'19).
4. SarathGuttikunda , and Nishadh KA, " Evolution of India's PM_{2.5} pollution between 1998 and 2020 using global reanalysis fields coupled with satellite observations and fuel consumption patterns".
5. S. Chowdhury, S. Dey, L. Di Girolamo, K.R. Smith, A. Pillarisetti, A. Lyapustin, Tracking ambient PM_{2.5} build-up in Delhi national capital region during the dry season over 15 years using a high-resolution (1 km) satellite aerosol dataset, Atmos. Environ. 204 (2019) 142–150

II) Reference from Book

1. A.B. Tsybakov , "Introduction to Nonparametric Estimation ".
2. Andrew Blann, "Data Handling and Analysis".
3. Al-Malawi Dhaif-Allah R , Shaltout Abdallah A (2014), "Assessment of Environmental Pollution of Pm_{2.5} at Taif, Saudi Arabia".

III) References from data collected sites

1. <https://www.airnow.gov/international/us-embassies-and-consulates/#India>
2. <https://in.usembassy.gov/u-s-citizen-services/consular-posts-india/graphical-india-map-emb-cons/>