

CHAPTER 1: INTRODUCTION

Air pollution refers to the presence of harmful substances or pollutants in the air that we breathe. These pollutants can come from natural sources like volcanic eruptions and wildfires, but human activities are the primary cause of air pollution. Some of the main sources of human-caused air pollution include Transportation, Industrial Activities, Agricultural practices, Household Activities. As the largest growing industrial nation, India is producing record amount of pollutants Particulate Matter 2.5, Particulate Matter 10, Carbon Monoxide(CO) etc. and other harmful aerial contaminants. Air quality of a particular state or a country is a measure on the effect of pollutants on the respected regions, as per the Indian air quality standard pollutants are indexed in terms of their scale, and these air quality indexes indicate the levels of major pollutants on the atmosphere. In view of this, Central Pollution Control Board(CPCB) took initiative for developing a national Air Quality Index (AQI) for Indian cities. Air pollution can have serious impacts on human health, including respiratory problems like asthma, lung cancer, and heart disease. It can also harm the environment by damaging crops and forests, polluting water bodies, and contributing to climate change.

1.1 Motivation

As a student studying Statistics, we were interested in understanding how air pollutants contribute to determining the Air Quality Index (AQI). The AQI serves important purposes like protecting public health, preserving the environment, following regulations, advancing scientific research, and engaging with the community. By collecting and analysing data, we can identify sources of pollution, work towards reducing emissions, and raise awareness about air quality issues. Sharing information about air quality with the public helps people make informed decisions to keep themselves and the environment healthy.

1.2 Project Problem

With the increasing concern about the rising levels of air pollution, our curiosity led us to explore the major harmful air pollutants in different locations. Conducting a study on the Air Quality Index (AQI) becomes crucial, especially in cities like Delhi. Such studies play a vital role in identifying the sources of pollution, mitigating the adverse economic and health effects of air pollution, safeguarding the environment, and formulating informed policies to enhance air quality. By understanding the key pollutants and their

impact, we can work towards implementing effective measures to combat air pollution and ensure a healthier and sustainable living environment.

1.3 Objective

➤ **The project aims to achieve the following:**

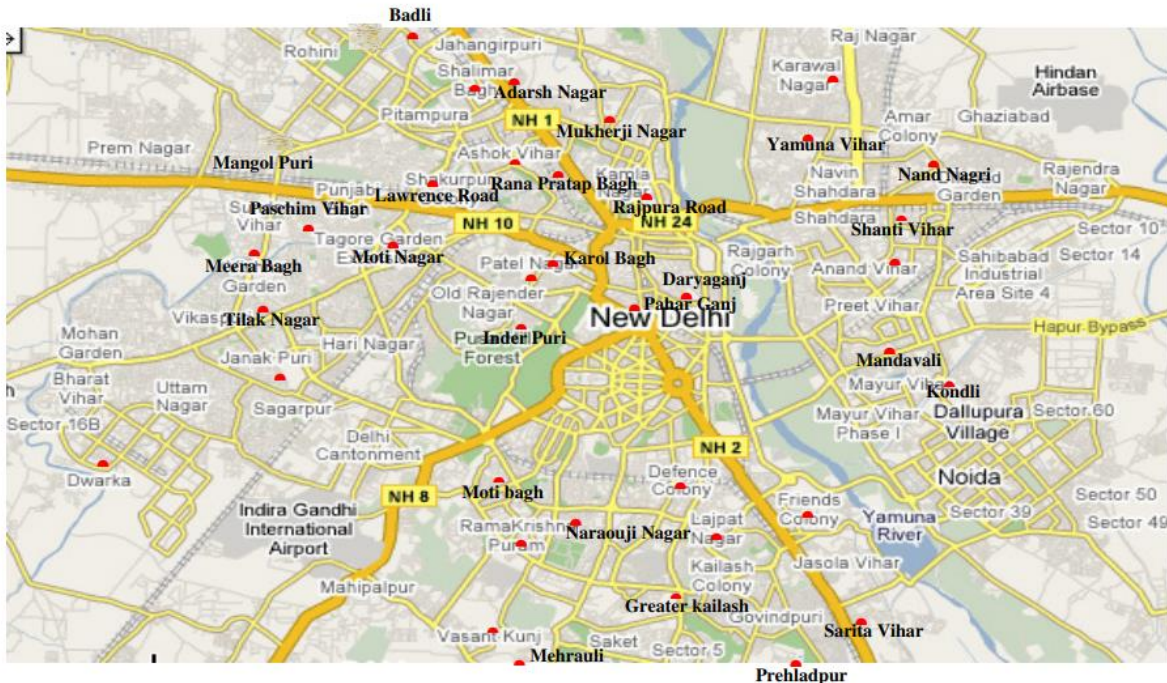
- Develop an easy-to-understand parameter to provide an overall status of air quality in Delhi.
- Identify the most critical pollutants responsible for high AQI levels at various locations in the city.
- Classify the levels of air quality at different locations to highlight the health concerns.
- To classify locations based on air quality levels using clustering techniques.
- To determine and compare the levels of air pollutants between summer and winter.
- Uncover hidden patterns and trends in vast quantities of air quality data through advanced data analysis techniques.
- Forecast the Air Quality Index (AQI) and pollutant levels at various stations to provide insights into future air quality scenarios.

1.4 Expected Outcomes and Significance of the Study:

- Developing a uniform AQI that considers health impacts, air quality standards, and monitoring scenarios to facilitate informed decision-making by policymakers, businesses, and individuals in reducing air pollution.
- Identifying the major air pollutants contributing to poor air quality in specific areas of Delhi, allowing for targeted interventions such as promoting public transport, electric vehicles, and improved traffic management.
- Classifying monitoring station locations based on air quality levels to assist individuals in making informed decisions about their activities and potential exposure to air pollution.
- Analysing vast quantities of air quality data to uncover hidden patterns and gain insights into pollution sources and patterns, leading to more effective policies and interventions.
- Classifying monitoring station locations based on their air quality to provide a comprehensive understanding of pollution levels across different areas of Delhi.

- Forecasting AQI and pollutant levels for monitoring stations to enable proactive measures by policymakers to mitigate pollution levels before they become hazardous.

1.5 Overview of Delhi City



(Map Source: DPCC)

Figure 1.1 Map of Delhi

Delhi, the capital city of India, is known for its rich historical heritage, bustling streets, and diverse culture. However, it also faces a significant challenge when it comes to air pollution, which has emerged as a major concern affecting the city's residents and overall environmental health. The current metro area population of Delhi in 2023 is 32,941,000.

Delhi, one of the most heavily polluted cities in the world, faces a significant air pollution problem. The sources of air pollution in Delhi are varied and complex, ranging from vehicular emissions and industrial activities to construction dust, open burning of waste and biomass, and natural factors such as dust storms and crop burning. According to Delhi Pollution Control Board reports, vehicular emissions account for around 40% of the total pollution in Delhi, with industrial emissions being responsible for 30%. Construction activities also contribute significantly to the particulate matter (PM) 2.5 levels in Delhi, with biomass burning accounting for around 17% of the particulate matter in the city's air.

In addition, the neighbouring states of Punjab and Haryana burn crop residue during the months of October and November, which contributes to a significant increase in air pollution in Delhi. The city's rapid urbanization has also led to increased construction activity, which produces large amounts of dust and particulate matter. Furthermore, Delhi's location in a region with high levels of natural dust and sand, as well as its geography, which traps pollutants due to low wind speed and high humidity, exacerbate the city's air pollution problems.

The government of Delhi and other stakeholders have implemented a range of measures to address air pollution, including the introduction of the Odd-Even rule for private vehicles, shutting down of polluting industries, and the implementation of the Graded Response Action Plan (GRAP). Despite these efforts, the issue of air pollution in Delhi remains a significant health and environmental concern. Vehicular emissions: According to a report by the Central Pollution Control Board (CPCB), vehicular emissions account for around 40% of the total pollution in Delhi.

➤ **Sources and Contributors of Air Pollution in Delhi:**

1. Vehicular emissions: According to a report by the Central Pollution Control Board (CPCB), vehicular emissions account for around 40% of the total pollution in Delhi. The high number of private vehicles on the roads, including diesel-powered ones, is a major contributor to air pollution in the city.
2. Power plants: Thermal power plants located in and around Delhi are significant contributors to air pollution. These plants emit large quantities of sulphur dioxide, nitrogen oxide, and particulate matter, which can cause respiratory problems and other health issues.
3. Agricultural Burning: Crop residue burning in neighbouring states, particularly during the post-harvest season, leads to the transport of smoke and pollutants to Delhi, exacerbating air pollution levels.
4. Residential and commercial activities: Activities such as cooking with biomass, burning garbage, and using diesel generators for electricity contribute significantly to air pollution in Delhi.
5. Geographical location: Delhi's geographical location, combined with the lack of wind movement during winter months, worsens the situation by trapping pollutants and leading to a build-up of smog.

6. Dust and construction materials: The construction industry in Delhi also contributes to air pollution, with the use of construction materials such as cement, sand, and bricks releasing particulate matter into the air. The city's dusty roads also add to the problem, with the wind easily picking up and spreading dust particles.
7. Inadequate waste management: Improper waste management practices, such as open burning of waste and uncontrolled dumping of garbage, also contribute to air pollution in Delhi.

➤ **Preventive measures and schemes implemented in Delhi to address air pollution**

- Bharat Stage VI (BS-VI) Fuel: Delhi transitioned to BS-VI fuel standards from April 2018. This move aimed to reduce the sulphur content in fuels and curb vehicular emissions. BS-VI compliant vehicles have also been introduced to ensure cleaner transportation.
- Graded Response Action Plan (GRAP): The GRAP is a comprehensive plan implemented by the Central Pollution Control Board (CPCB) and the Delhi Pollution Control Committee (DPCC). It outlines specific actions to be taken based on the severity of air pollution levels. These actions include measures like banning certain activities, increasing public transportation, and enforcing stricter emission standards.
- Construction of the Eastern Peripheral Expressway (EPE): The EPE, completed in 2018, is a 135-kilometer peripheral road around Delhi. It diverts a significant amount of non-Delhi bound traffic away from the city, reducing congestion and pollution.
- Greening Initiatives: Delhi has undertaken various initiatives to increase green cover, such as the planting of trees and the development of green spaces. For instance, the "Paudhe Lagao, Paryavaran Bachao" (Plant Trees, Save Environment) campaign aims to increase the city's green cover by planting millions of trees.
- Electric Vehicles (EV) Promotion: The Delhi government has introduced incentives and subsidies to promote the adoption of electric vehicles. These measures include financial incentives, exemptions from road tax and registration fees, and the development of charging infrastructure.





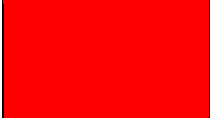

CHAPTER 2: TERMS AND CONCEPTS RELATED WITH AIR POLLUTANTS

In this chapter, we will explore several important terms and concepts related to air pollutants. Understanding these concepts is crucial for assessing and managing air quality. We will discuss the Air Quality Index (AQI), Pollutants measurement, the unit of measurement used, subindex, and the contribution of the Central Pollution Control Board (CPCB) in monitoring and controlling air pollution.

2.1 Air Quality Index (AQI):

The Air Quality Index (AQI) is a numerical scale used to communicate the quality of air in a specific location at a given time. It provides a simplified representation of air pollution levels and their potential health impacts. The AQI typically ranges from 0 to 500, with higher values indicating poorer air quality. The index is often categorized into different color-coded ranges, such as good, moderate, unhealthy, and hazardous, to help people easily understand the air quality conditions.

Table 2.1: Health Statements for AQI Categories and AQI standards by CPCB

AQI	Colour Code	Associated Health Impacts
Good (0-50)		Minimal Impact
Satisfactory (51-100)		May cause minor breathing discomfort to sensitive people
Moderate (101-200)		May cause breathing discomfort to the people with lung disease such as asthma and discomfort to people with heart disease, children, and older adults
Poor (201-300)		May cause breathing discomfort to the people on prolonged exposure and discomfort to people with heart disease with short exposure
Very Poor (301-400)		May cause respiratory illness to the people on prolonged exposure. Effect may be more pronounced in people with lung and heart diseases
Severe (401-500)		May cause respiratory effects even on healthy people and serious health impacts on people with lung/heart diseases. The health impacts may be experienced even during light physical activity

Air quality standards serve as a crucial framework for controlling air pollution and ensuring public health. These standards, adopted by regulatory authorities, establish enforceable levels of air quality. The primary objective behind developing these standards is to protect public health from the adverse effects of air pollutants, minimize or eliminate exposure to hazardous pollutants, and guide national and local authorities in making effective pollution control decisions. They provide a rational basis for implementing measures that reduce air pollution and promote a healthier environment. By adhering to air quality standards, we can safeguard the well-being of communities and ensure sustainable development for future generations.

➤ **Pollutants Measurement:**

Measuring air pollutants involves the collection and analysis of air samples to determine the concentration of various pollutants present in the atmosphere. This process helps in understanding the extent of pollution and evaluating compliance with air quality standards. Different methods are used for measuring different pollutants, including particulate matter (PM), nitrogen dioxide (NO₂), sulphur dioxide (SO₂), carbon monoxide (CO), ozone (O₃), and others. These measurements are conducted using specialized instruments and techniques, such as air quality monitoring stations.

➤ **Subindex:**

The AQI is composed of different subindices, each corresponding to a specific pollutant or group of pollutants. These subindices represent the relative contribution of individual pollutants to the overall AQI value. For example, a subindex for PM_{2.5} and another subindex for NO₂ might be used to calculate the overall AQI. By examining the subindices, it is possible to identify the primary pollutants responsible for the observed air quality conditions.

➤ **Contribution of CPCB:**

The Central Pollution Control Board (CPCB) is a regulatory body in India responsible for monitoring and controlling air pollution. The CPCB plays a crucial role in developing and implementing policies, standards, and guidelines related to air quality. It establishes air quality monitoring stations across the country to measure pollutant concentrations and assess compliance with national air quality standards. The CPCB also collaborates with state pollution control boards and other agencies to conduct research, raise public awareness, and enforce pollution control measures.

2.2 Parameter related to AQI

The proposed AQI will consider eight pollutants (PM₁₀, PM_{2.5}, NO₂, O₃, CO, SO₂, NH₃, and Pb) for which short-term (up to 24-hourly averaging period) National Ambient Air Quality Standards are prescribed.

Using this calculator, we calculate AQI in excel for all over data. In our data the Pb is not present. It is not collected by CPCB.

- **Meaning of (up to 24-hourly/8-hourly averaging period)**

The terms "24-hour average" and "8-hour average" are commonly used in air quality monitoring to describe the length of time over which air quality data is averaged.

The 24-hour average refers to the average concentration of a pollutant over a 24-hour period. This is calculated by adding up the concentration of the pollutant at various times over the course of 24 hours and then dividing by the number of measurements taken. This is a commonly used metric for PM₁₀, PM_{2.5}, NO₂, and SO₂.

For example, if PM_{2.5} levels are monitored every hour for 8 hours and the concentrations are 50 µg/m³, 60 µg/m³, 40 µg/m³, 70 µg/m³, 80 µg/m³, 90 µg/m³, 100 µg/m³, and 70 µg/m³ respectively, then the 8-hour average would be calculated as follows:

$$(50 + 60 + 40 + 70 + 80 + 90 + 100 + 70) \div 8 = 70 \text{ µg/m}^3$$

The 8-hour average, on the other hand, is used for measuring ozone (O₃) levels. It refers to the average concentration of ozone over an 8-hour period, which is typically the period of highest ozone concentrations during the day. This is calculated in a similar way to the 24-hour average, by adding up the concentration of ozone at various times over the course of 8 hours and then dividing by the number of measurements taken.

Overall, both the 24-hour average and the 8-hour average are useful metrics for assessing air quality, as they provide an indication of the average concentration of pollutants over a given period of time.

2.2.1 Particulate Matter (PM₁₀)

PM₁₀ is one of the criteria for Air Quality Index (AQI) calculation. The safe exposure levels for PM₁₀ (24 hour) are 0-100 µg/m³. India 37 cities have been identified as having the highest pollution levels of PM₁₀. Rapid industrialization and urbanization in India have

resulted in highly polluted cities and a large proportion of the Indian population is exposed to high levels of particulate pollutants.

Sources:

Particulate Matter is released from constructions, smoking, cleanings, renovations, demolitions, constructions, natural hazards such as earthquakes, volcanic eruptions and emissions from industries such as brick kilns, paper & pulp etc.

Related Effects:

- Major concerns for human health from exposure to PM₁₀ include effects on breathing, respiratory symptoms, decrease in pulmonary function and damage to lung tissue, cancer, and premature death.
- An association between elevated PM₁₀ levels and hospital admissions for pneumonia, bronchitis, and asthma was observed. Long-term particulate exposure was associated with an increase in risk of respiratory illness in children.
- An increase of 10µg/m³ of PM₁₀ levels resulted in a 3-6 % increase in visits for asthma people and a 1-3 % increase in visits for upper respiratory diseases not with asthma to hospitals.
- The findings are consistent with the result of previous studies of particulate pollution in other urban areas and provide evidence that the coarse fraction of PM₁₀ may affect the health of working people.

2.2.2 Particulate Matter (PM_{2.5})

PM_{2.5} is one of the criteria for Air Quality Index (AQI) calculation. The safe exposure levels for PM_{2.5} (24 hour) is 0-60 µg/m³. Several epidemiological studies (Pope,1989,Schwartz, 1996) have linked PM₁₀ (aerodynamic diameter ≤ 10 µm) and PM_{2.5} (aerodynamic diameter ≤ 2.5 µm) with significant health problems.

Sources:

PM_{2.5} comes primarily from combustion. Fireplaces, car engines, and coal or natural gas fired power plants are all major pm 2.5 sources. Beside this, Particulate Matter is released from constructions, smoking, cleanings, renovations, demolitions,

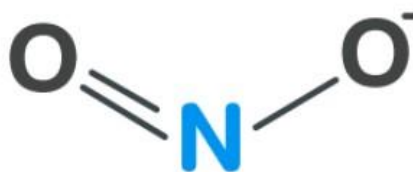
constructions, natural hazards such as earthquakes, volcanic eruptions and emissions from industries such as brick kilns, paper & pulp etc.

Related Effects:

- Premature mortality, chronic respiratory disease, emergency visits and hospital admissions, aggravated asthma, acute respiratory 27 symptoms, and decrease in lung function. PM_{2.5} is of specific concern because it contains a high proportion of various toxic metals and acids, and aerodynamically it can penetrate deeper into the respiratory tract.
- Long-term (months to years) exposure to PM_{2.5} has been linked to premature death, particularly in people who have chronic heart or lung diseases, and reduced lung function growth in children.
- Particulate matter has been shown in many scientific studies to reduce visibility, affects by altering the way light is absorbed and scattered in the atmosphere.

2.2.3 Nitrogen Dioxide (NO₂)

Nitrogen dioxide is a parameter for calculating AQI. As per Indian Government safe exposure is 0-80 ug/m³ (24 hour). Nitrogen dioxide is a known highly reactive gas present in the atmosphere. It is one of the major atmospheric pollutants that absorb UV light and stops to reach it to the earth's surface.



Sources:

It is released into the environment from automobile emissions, generation of electricity, burning of fuel, combustion of fossil fuel and different industrial processes.

Related Effects:

- Nitrogen dioxide poisoning is as much as hazardous as carbon monoxide poisoning.
- It is when inhaled can cause serious damage to the heart, absorbed by lungs, inflammation and irritation of airways. Smog formation and foliage damage are some environmental impacts of nitrogen dioxide.

2.2.4 Ozone (O₃)

Ozone is a parameter for calculating AQI. The safe exposure is 0-80 ug/m³ (24 hour). Ozone is composed of three oxygen atoms. It forms the protective layer which prevents entry of harmful ultraviolet radiation into the earth. The ground ozone is very harmful to human beings and the environment.



Sources:

It is released from industries, automobile emissions, gasoline vapours solvents, chemicals, electronic devices. Nitrogen oxides (NO_x) and total Volatile Organic Compounds (toss) also contribute to ground ozone formation.

Related Effects:

- Ground ozone interferes with the plant's respiration process and enhances environmental stressor susceptibility. When ozone is inhaled by humans, reduced lung function, inflammation of airways and irritation in eyes, nose & throat are seen.

2.2.5 Carbon Mono-oxide (CO)

Carbon Mono-oxide is a parameter for calculating AQI. Safe level of exposure according to the Indian government is 0-04 mg/m³ (1-hour).

Carbon monoxide (CO) is an important criteria pollutant which is ubiquitous in urban environment. CO production mostly occurs from sources having incomplete combustion. Due to its toxicity and appreciable mass in atmosphere, it should be considered as an important pollutant in AQI scheme.



Sources:

It is a colourless gas, releasing from automobile emissions, fires, industrial processes, gas stoves, kitchen chimneys, generators, wood burning smoking etc. into the atmosphere.

Related Effects:

- The initial symptoms of CO poisoning may include headache, dizziness, drowsiness, and nausea.
- These initial symptoms may advance to vomiting, loss of consciousness, and collapse if prolonged or high exposures are encountered and may lead to Coma or death if high exposures continue.

2.2.6 Sulphur Dioxide (SO₂)

Sulphur Dioxide is used as a parameter for Air Quality Index (AQI) calculation. The safe exposure level is 0-80 ug/m³ (24 hour) according to the Indian government respectively. Sulphur dioxide is a colourless gas with burnt odour and chemical formula SO₂. The gas is acidic & corrosive in nature and can react in the atmosphere with other compounds to form sulfuric acid and other oxides of Sulphur.



Sources:

Emissions from automobiles, industries, combustion of fossil fuel, generation of electricity etc. are reasons for the entry of Sulphur dioxide into the atmosphere.

Related Effects:

- Sulphur dioxide is a major cause of haze production, acid rain, damage to foliage, monuments & buildings, reacts and forms particulate matter.
- In humans, it causes breathing discomfort, asthma, eyes, nose and throat irritation, inflammation of airways and heart diseases.

2.2.7 Ammonia (NH₃) and Lead (Pb)

It is to be noted that most of the countries have taken only six pollutants (described above) for formulation of AQI. An attempt has been made to propose breakpoints for NH₃

and Pb as these two pollutants also have short-term standards of 24-hr. While NH_3 can be measured on continuous basis and can be included in the list of real time parameters for AQI, such measurements are not possible for Pb. However, Pb levels can be utilized in calculation of AQI of past days to assess impact of lead pollution.

Sources:

Ammonia and Lead major found in chemical industries, hospitals and other industrial sectors.

Related Effects:

- Inhalation of high levels of NH_3 causes irritation to the nose, throat and respiratory tract. Increased inhalation may result in cough and an increased respiratory rate as well as respiratory distress.
- An association has been reported between exposure to ammonia and cough, phlegm, wheezing, and asthma at high concentration.
- Pb is a toxic metal and its exposure through all routes results in increased blood lead level.

• Measurement Methods and Units Used by CPCB for Air Pollutants in India

The Central Pollution Control Board (CPCB) measures various air pollutants using different methods, depending on the type of pollutant. Some of the methods used by CPCB are:

1. Particulate Matter (PM₁₀ and PM_{2.5}): CPCB uses a High-Volume Sampler (HVS) to collect samples of particulate matter from the air. The collected samples are then weighed and the mass of PM₁₀ or PM_{2.5} is calculated. The units used for measuring PM₁₀ and PM_{2.5} are micrograms per cubic meter ($\mu\text{g}/\text{m}^3$).

Unit of Pollutants PM₁₀ and PM_{2.5}

A mixture of particles with liquid droplets in air forms particulate matter. PM₁₀ are particles having size of less than or equal to 10 microns whereas PM_{2.5} are ultrafine particles having size less than or equal to 2.5 microns. For the sake of comparison, most bacteria are at least five microns across. The diameter of a red blood cell is six microns. A strand of hair is around 70 microns wide. You could fit several thousand PM_{2.5} particles on a period.

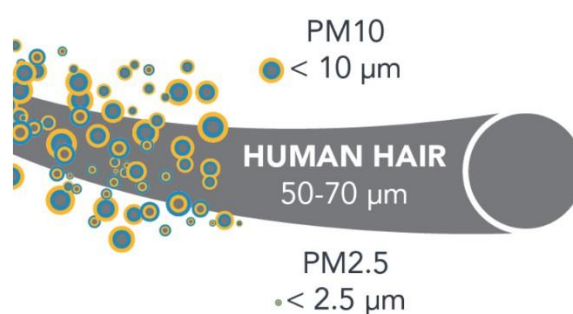


Figure 2.1 PM2.5 and PM10 size measurement

2. Sulphur Dioxide (SO₂): CPCB uses a UV Fluorescence Analyzer to measure the concentration of SO₂ in the air. The analyser works by measuring the amount of UV light absorbed by SO₂. The unit used for measuring SO₂ is parts per billion (ppb).

3. Nitrogen Dioxide (NO₂): CPCB uses a chemiluminescence analyser to measure the concentration of NO₂ in the air. The analyser works by measuring the amount of light produced when NO₂ reacts with ozone. The unit used for measuring NO₂ is parts per billion (ppb).

4. Ammonia (NH₃): The measurement method used by CPCB for NH₃ in India is based on the principle of gas diffusion through a tube and subsequent reaction with boric acid in a buffered solution. The resulting ammonium borate is then titrated with standardized hydrochloric acid to determine the amount of NH₃ present in the air sample. The unit used for measuring CO is parts per million (ppb)

5. Carbon Monoxide (CO): CPCB uses a Non-Dispersive Infrared (NDIR) analyser to measure the concentration of CO in the air. The analyser works by measuring the amount of infrared light absorbed by CO. The unit used for measuring CO is parts per million (ppm).

6. Ozone (O₃): CPCB uses an Ozone Analyzer to measure the concentration of O₃ in the air. The analyser works by measuring the amount of UV light absorbed by O₃. The unit used for measuring O₃ is parts per billion (ppb).

To convert the units of measurement from ppb to $\mu\text{g}/\text{m}^3$, CPCB uses conversion factors specific to each pollutant. These factors are based on the molecular weight of the pollutant and the temperature and pressure of the air.

2.3 Formulation of AQI

Formation of sub-indices (for each pollutant)

Air quality standards play a crucial role in controlling air pollution and protecting public health from the harmful effects of air pollutants. The development of these standards is based on scientific research that determines the level of air quality required to ensure public health and safety.

To formulate an AQI, sub-indices are created for each pollutant to provide an overall picture of the air quality. These sub-indices are calculated based on the concentration of pollutants in the air, measured by sophisticated instruments and equipment. The use of sub-indices allows for a more comprehensive assessment of the air quality, considering the different pollutants present in the air. This information can then be used to make informed decisions regarding air pollution control measures. By adopting air quality standards as enforceable regulations, regulatory authorities can ensure that the general public is protected from the adverse effects of air pollutants. This not only improves public health and safety but also helps in guiding national and local authorities in making informed pollution control decisions.

The formulation of an AQI is crucial in maintaining a healthy and sustainable environment for present and future generations.

2.3.1 Structure of an Index

Air Quality Index (AQI) is a structured system that is used to measure the air quality of a particular area. To calculate the AQI, sub-indices are formed for each pollutant variable based on air quality standards and health effects. These sub-indices, represented by I_1, I_2, \dots, I_n , are derived from pollutant concentrations X_1, X_2, \dots, X_n using sub-index functions.

The relationship between sub-index value (I_i) and pollutant concentrations (X_i) is crucial in understanding the health effects of air pollutants. These sub-indices are then aggregated using a mathematical function (F) to obtain the overall index, known as the AQI. The aggregation function could be summation, multiplication or the maximum operator. The AQI provides a comprehensive and easily understandable representation of the air quality in a particular area.

Here are the formulas mentioned in the text:

The sub-index formula:

$$I_i = f(X_i), i = 1, 2, \dots, n$$

The overall index formula:

$$I = F(I_1, I_2, \dots, I_n)$$

Where:

I_i = sub-index for pollutant i

X_i = concentration of pollutant i

n = number of pollutants being measured

I = overall index (AQI)

F = function for aggregating sub-indices (e.g. summation, multiplication, maximum)

2.3.2 Sub-indices

Sub-index function represents the relationship between pollutant concentration X_i and corresponding sub-index I_i . It is an attempt to reflect environmental consequences as the concentration of specific pollutant changes. It may take a variety of forms such as linear, non-linear and segmented linear. Typically, the I - X relationship is represented as follows:

$$I = \alpha X + \beta$$

Where,

α = slope of the line,

β = intercept at $X=0$.

The general equation for the sub-index (I_i) for a given pollutant concentration (C_p); as based on 'linear segmented principle' is calculated as:

$$I_i = \left[\left\{ \frac{I_{HI} - I_{LO}}{B_{HI} - B_{LO}} \right\} \times (C_p - B_{(LO)}) \right] + I_{LO}$$

Where,

B_{HI} = Breakpoint concentration greater or equal to given concentration (C_p).

B_{LO} = Breakpoint concentration smaller or equal to given concentration (C_p).

I_{HI} = AQI value corresponding to the breakpoint concentration (B_{HI})

I_{LO} = AQI value corresponding to the breakpoint concentration (B_{LO})

C_p = Pollutant concentration

Now, for to calculate AQI using the sub-indices of air pollutants,

$$AQI = MAX (I_i)$$

Where, $i = 1, 2, \dots, n$; denotes n pollutants

There are two reasons for adopting a maximum operator function:

- Free from eclipsing and ambiguity
- Health effects of combination of pollutants (synergistic effects) are not known and thus a health-based index cannot be combined or weighted

Table 2.2: Breakpoints for AQI Scale 0-500 (units: $\mu\text{g}/\text{m}^3$ unless mentioned otherwise)

AQI Category (Range)	PM ₁₀ 24- hr	PM _{2.5} 24- hr	NO ₂ 24- hr	O ₃ 8-hr	CO 8-hr (mg/m^3)	SO ₂ 24hr	NH ₃ 24- hr	Pb 24-hr
Good (0-50)	0-50	0-30	0-40	0-50	0-1.0	0-40	0-200	0-0.5
Satisfactory (51-100)	51- 100	31- 60	41- 80	51-100	1.1-2.0	41- 80	201- 400	0.6 – 1.0
Moderate (101-200)	101- 250	61- 90	81- 180	101- 168	2.1-10	81- 380	401- 800	1.1- 2.0
Poor (201-300)	251- 350	91- 120	181- 280	169- 208	10.1-17	381- 800	801- 1200	2.1- 3.0
Very poor (301-400)	351- 430	121- 250	281- 400	209- 748*	17.1-34	801- 160 0	1201- 1800	3.1- 3.5
Severe (401-500)	430+	250+	400+	748+*	34+	160 0+	1800 +	3.5+

*hourly monitoring (for mathematical calculation only)

Breakpoints are provided by CPCB for each pollutant to segment the pollutant concentration range into sub-ranges. The breakpoints are set based on the National Ambient Air Quality Standards (NAAQS) for India, which define the permissible limits of pollutant concentrations in ambient air. The sub-ranges are then assigned specific AQI values based on the sub-index formula. By using breakpoints, the sub-index values can be computed for a range of pollutant concentrations and then combined to calculate the overall AQI for a particular location.

Using above Breakpoints (table 2.2) and formula 4 and 5 we calculate the AQI. Some example is given below:

2.3.3 Example

1. Consider the values as concentration of pollutants from the dataset as day wise (One day) to find Air Quality Index (AQI)?

$$PM_{10} = 45.88 \mu\text{g}/\text{m}^3$$

$$PM_{2.5} = 39.9 \mu\text{g}/\text{m}^3$$

$$NO_2 = 50.92 \mu\text{g}/\text{m}^3$$

$$O_3 = 28.89 \mu\text{g}/\text{m}^3$$

$$CO = 0.46 \text{mg}/\text{m}^3$$

$$SO_2 = 20.75 \mu\text{g}/\text{m}^3$$

Solution:

Here we are given the values of Pollutant Concentrations for one day,

First, Calculate the sub-index (I_i) (for Breakpoint concentration for each pollutant see the table 2.2 above) we know the formula for sub-index,

$$I_i = \left[\left\{ \frac{I_{HI} - I_{LO}}{B_{HI} - B_{LO}} \right\} \times (C_p - B_{(LO)}) \right] + I_{LO}$$

Now, putting the values in above formula, Calculating sub-index for each pollutant,

I) $PM_{10} = 45.88 \mu\text{g}/\text{m}^3$

$$\begin{aligned} I_1 &= \left[\left\{ \frac{(50 - 0)}{(50 - 0)} \right\} \times (45.88 - 0) \right] + 0 \\ &= 45.88 \end{aligned}$$

II) $PM_{2.5} = 39.9 \mu\text{g}/\text{m}^3$

$$\begin{aligned} I_2 &= \left[\left\{ \frac{(100 - 51)}{(60 - 31)} \right\} \times (39.9 - 31) \right] + 51 \\ &= 66.5 \end{aligned}$$

$$\text{III) } \text{NO}_2 = 50.92\mu\text{g}/\text{m}^3$$

$$\begin{aligned} I_3 &= [\{(100 - 51) / (80 - 41)\} * (50.92 - 41)] + 51 \\ &= 63.65 \end{aligned}$$

$$\text{IV) } \text{O}_3 = 28.89\mu\text{g}/\text{m}^3$$

$$\begin{aligned} I_4 &= [\{(50 - 0) / (50 - 0)\} * (28.89 - 0)] + 0 \\ &= 28.89 \end{aligned}$$

$$\text{V) } \text{CO} = 0.46\text{mg}/\text{m}^3$$

$$\begin{aligned} I_5 &= [\{(50 - 0) / (1 - 0)\} * (0.46 - 0)] + 0 \\ &= 23 \end{aligned}$$

$$\text{VI) } \text{SO}_2 = 20.75\mu\text{g}/\text{m}^3$$

$$\begin{aligned} I_6 &= [\{(50 - 0) / (40 - 0)\} * (20.75 - 0)] + 0 \\ &= 25.9375 \end{aligned}$$

Now we calculate AQI,

$$\begin{aligned} \text{AQI} &= \text{Max} (I_1, I_2, I_3, I_4, I_5, I_6) \\ &= 66.5 \end{aligned}$$

Hence, The AQI value is 66.5 (for PM₁₀) of one day (date: 22/09/2019) City – Delhi (Alipur).

2.3.4 AQI Calculation Using Spreadsheet XL

To calculate the Air Quality Index (AQI) for a specific location on a given day, CPCB has developed a user-friendly MS Excel spreadsheet. The user needs to input the concentrations of at least three pollutants, including one of either PM_{2.5} or PM₁₀, in the designated blue cells. Once these values are entered, the sub-indices are automatically calculated, and the final AQI is displayed along with a colour code that signifies the AQI

category. The legend at the bottom of the sheet provides detailed information on the health impacts associated with each AQI category.

It is important to note that the overall AQI can only be calculated if data is available for at least three pollutants, out of which one must be either PM2.5 or PM10. If data is insufficient or missing for any of these pollutants, the AQI cannot be calculated. Similarly, a minimum of 16 hours of data is required to calculate the sub-index.

This Excel program can be obtained from CPCB and is a useful tool for individuals, organizations, and government agencies to monitor air quality and take necessary actions to mitigate pollution levels.

Calculation of AQI						
Date 28-Jan-15		Station City State		IIT Kanpur Kanpur Uttar Pradesh		
Pollutants		concentration in $\mu\text{g}/\text{m}^3$ (except for CO)	Sub-Index			Air Quality Index
PM10	24-hr avg	64.00	64	check 1		
PM2.5	24-hr avg	58.00	97	1		
SO ₂	24-hr avg	5.00	6	1		
NO _x	24-hr avg	22.00	28	1	AQI =	97
*CO (mg/m ³)	max 8-hr		0	0		
O ₃	max 8-hr	35.00	35	1		
NH ₃	24-hr avg	0.00	0	0		
* Concentrations of minimum three pollutants are required; one of them should be PM10 or PM2.5						
* The check displays "1" when a non-zero value is entered						
Good (0–50)	Minimal Impact		Poor (201–300)	Breathing discomfort to people on prolonged exposure		
Satisfactory (51–100)	Minor breathing discomfort to sensitive people		Very Poor (301–400)	Respiratory illness to the people on prolonged exposure		
Moderate (101–200)	Breathing discomfort to the people with lung, heart disease, children and older adults		Severe (>401)	Respiratory effects even on healthy people		

Figure 2.2 Spreadsheet for AQI Calculation

2.4 Applications of Air Quality Index

Air Quality Index (AQI) has several applications that serve various objectives. Here are six objectives that are served by an AQI:

1. Resource Allocation: AQI helps administrators in allocating funds and determining priorities for air pollution control strategies. It enables evaluation of trade-offs involved in alternative air pollution control strategies.
2. Ranking of Locations: AQI helps in comparing air quality conditions at different locations/cities, pointing out areas and frequencies of potential hazards.

3. **Enforcement of Standards:** AQI determines the extent to which the legislative standards and existing criteria are being adhered to. It also helps in identifying faulty standards and inadequate monitoring programs.
4. **Trend Analysis:** AQI helps to determine the change in air quality (degradation or improvement) that has occurred over a specified period. This enables forecasting of air quality and planning pollution control measures.
5. **Public Information:** AQI informs the public about the state of the environment, especially those who suffer from illness aggravated or caused by air pollution. It enables them to modify their daily activities when they are informed of high pollution levels.
6. **Scientific Research:** AQI is useful for reducing a large set of data to a comprehensible form that gives better insight to researchers while conducting a study of some environmental phenomena. It enables more objective determination of the contribution of individual pollutants and sources to overall air quality.

AQI is useful for the general public to know air quality in a simplified way, decision-makers to know the trend of events and to chalk out corrective pollution control strategies, government officials to study the impact of regulatory actions, and scientists who engage in scientific research using air quality data.

Briefly, an AQI is useful for:

- (i) General public to know air quality in a simplified way.
- (ii) Decision maker to know the trend of events and to check out corrective pollution control strategies.
- (iii) Government official to study the impact of regulatory actions and
- (v) Scientist who engages in scientific research using air quality data.

CHAPTER 3: DATA AND DATA REPRESENTATION

In this chapter, we take a closer look at the dataset used in our project. Data plays a crucial role in our analysis, so let's explore what kind of information was included and how it looked. We also examine where the data came from and how we made sure it was accurate and reliable. We discuss the steps taken to modify the data and remove any unusual values that could affect our findings. Understanding our dataset is key to unravelling the mysteries of Delhi's air pollution issue.

3.1 About Data Source

The Central Pollution Control Board (CPCB) is a statutory organization under the Ministry of Environment, Forest and Climate Change, Government of India. Established in 1974, CPCB aims to promote cleanliness and pollution control in India by monitoring environmental quality and implementing pollution control measures.

CPCB's website, www.cpcb.gov.in, launched in 2006, is a centralized location for accessing information on environmental quality and pollution control measures in India. The website provides real-time data on air, water, and noise pollution, hazardous waste management, and various environmental regulations and standards in India. The CPCB has played a crucial role in promoting transparency and accountability in environmental governance in India. It has enabled citizens and civil society organizations to monitor and evaluate environmental conditions, bringing about significant improvements in environmental quality in India. The CPCB provides access to real-time data from over 700 monitoring stations across India, including 39 in Delhi alone. The CPCB and its website have become a vital part of environmental governance in India. They provide critical information and data to support informed decision-making, policy development, and public awareness on environmental issues. The CPCB website has enabled citizens and policymakers to work towards achieving cleaner and healthier environments in India.

3.1.1 Data Dashboard:

The National AQI data dashboard, which can be accessed through the CPCB website, provides a user-friendly interface to explore air quality information for different cities. Users can select a specific city and view the current AQI value, the dominant pollutant contributing to the index, and an associated health advisory. Additionally, historical data is available, allowing users to analyse trends and patterns in air quality over time.

The following link of CPCB provides dashboard, from which we extract data of various stations.

<https://app.cpcbccc.com/ccr/#/caaqm-dashboard-all/caaqm-landing>

OR

Go through www.cpcb.gov.in -> select Air Quality Data ->Live Air Quality Data of Monitoring stations -> Captcha Verification -> Comparison Data

Central Control Room for Air Quality Management - All India

Average Report Criteria

State Name :

City Name :

Station Name :

Parameters :

Report Format :

Criteria :

Date From :

Date To :

Figure 3.1 CPCB Dashboard

The CPCB website offers a convenient dashboard where we can access a wealth of historical data. It provides options to select specific information such as state name, city name, station name, parameters, and date range. We can customize our selections according to our project requirements. Once we have chosen the desired information, we simply click on the submit button to retrieve the valuable historical data.

3.2 Data Information

To monitor the air quality, CPCB has set up several air quality monitoring stations in various cities across the country. Some of the major stations include DPCC, IMD, IITM, and CPCB itself.

1. DPCC (Delhi Pollution Control Committee) is a state-level pollution control committee that monitors air quality in the National Capital Region (NCR). It operates several air quality monitoring stations across Delhi, including residential, industrial, and commercial areas.
2. IMD (India Meteorological Department) is a national-level organization that provides weather forecasting services in India. IMD also operates air quality monitoring stations in several cities across India.
3. IITM (Indian Institute of Tropical Meteorology) is a research institute under the Ministry of Earth Sciences in India. It conducts research on meteorology, climate, and air quality. IITM operates air quality monitoring stations in Pune, Mumbai, and Hyderabad.
4. CPCB (Central Pollution Control Board) also operates several air quality monitoring stations across India. These stations provide real-time data on air quality parameters such as PM_{2.5}, PM₁₀, ozone, nitrogen dioxide, and sulphur dioxide. The data collected by these monitoring stations is used to create air quality indexes that help citizens and policymakers make informed decisions about their health and the environment.

Air pollution is a serious concern in Delhi, India. With 40 monitoring stations across the city, we have collected data on the concentrations of various air pollutants. By analysing this data, we can better understand the current state of air quality in Delhi and take steps towards improving it for the health and well-being of the community. These are given in following table 3.1.

Table 3.1 Air pollutants Monitoring Stations(Delhi)

Stations	Station Names
Delhi1	Alipur, DPCC
Delhi2	Anand Vihar, DPCC
Delhi3	Ashok Vihar, DPCC
Delhi4	Aya nagar, IMD
Delhi5	Bawana, DPCC
Delhi6	Burani Crossing, IMD
Delhi7	CRRRI Mathurd road, IMD
Delhi8	Chandni Chawk, IITM
Delhi9	DTU, CPCB
Delhi10	Dr. Karni Singh Shootin Range, DPCC
Delhi11	Dwarka Sector-8, DPCC
Delhi12	East Arjun nagar, CPCB
Delhi13	IGI Airport, IMD
Delhi14	IHBAS Dilshad Garden , CPCB
Delhi15	ITO, CPCB
Delhi16	Jahangirpuri, DPCC
Delhi17	Jawaharlal Nehru Stadiaum, DPCC
Delhi18	Lodhi road, IITM
Delhi19	Lodhi road, IDM
Delhi20	Major Dhyanchand National stadium, DPCC
Delhi21	Mandir Marg, DPCC
Delhi22	Mundka, DPCC
Delhi23	NSIT Dwarka, CPCB
Delhi24	Najafgarh, DPCC
Delhi25	Narela, DPCC
Delhi26	Nehru nagar, DPCC
Delhi27	North Campus-DU, IMD
Delhi28	Okhala Phase-2, DPCC
Delhi29	Patparganj, DPCC
Delhi30	Panjabi Bagh, DPCC
Delhi31	Pusa DPCC
Delhi32	Pusa IMD
Delhi33	R.K. Puram, DPCC
Delhi34	Rohini, DPCC
Delhi35	Shadipur, CPCB
Delhi36	Sirifort, CPCB
Delhi37	Sonia Vihar, DPCC
Delhi38	Sri Aurabindo Marg, DPCC
Delhi39	Vivek Vihar, DPCC
Delhi40	Wazipur, DPCC

We have collected hourly concentration values of air pollutants from 40 monitoring stations across the city, resulting in a total of 65,062 observations. By removing outliers

and missing values, we have ensured the accuracy and reliability of our data. To determine the daily concentration of pollutants. These values are then used to calculate the Air Quality Index (AQI), as determined by the Central Pollution Control Board (CPCB). We Explore our dataset to gain insights into the air quality of Delhi and discover trends in air pollutant concentrations across the city. Let our data guide you in making informed decisions to protect your health and well-being.

3.3 Data Mining

Our data, obtained from the Central Pollution Control Board (CPCB), comprises daily concentration values of air pollutants recorded at 39 monitoring stations in Delhi. However, the raw nature of the data introduces outliers and missing values, which can impact the accuracy of our analysis. To tackle this challenge, we harnessed the power of Python software and employed various data mining techniques.

Exploratory Data Analysis (EDA) played a vital role in our data processing. Through the utilization of Box Plots and descriptive statistics, we effectively identified and managed outliers within our dataset. Moreover, we encountered instances of missing values, which we addressed by substituting them with median values and excluding observations with exceptionally high pollutant concentrations.

In the data pre-processing phase, Excel proved to be an invaluable tool. Its user-friendly interface and powerful functionalities allowed us to efficiently clean and organize our dataset, facilitating smoother data analysis. We leveraged Excel's capabilities to handle missing values, filter out outliers, and perform basic statistical calculations. We processed an impressive 65,062 observations, enabling us to derive meaningful insights into the concentration of air pollutants in Delhi. These insights shed light on the Air Quality Index (AQI) and provide valuable guidance for implementing measures to improve the city's air quality. Excel's contribution to our data processing journey cannot be overstated, as it played a pivotal role in ensuring the accuracy and reliability of our analysis.

3.3.1 Data pre-processing:

Data pre-processing is an essential step in data mining, which involves transforming raw data into a more understandable format. In our dataset, we encountered missing values, which were filled with the median value. Additionally, the decimal values were converted into proper float values, which makes it easier to analyse the data. By cleaning and transforming the dataset, we can ensure that our analysis is accurate and reliable.

- **Data Cleaning:**

- Irrelevant columns were removed, retaining only the ones necessary for analysis.
- Missing values were handled by imputing them using appropriate techniques or removing rows/columns with missing values.
- Inconsistent or incorrect data formats were identified and resolved to ensure data integrity.
- Duplicate entries were detected and removed from the dataset.

- **Data Integration:**

- Multiple files or sources within the CPCB historical dataset were merged to create a comprehensive dataset for analysis.
- Column names and data types were standardized across different datasets to ensure consistency.

- **Feature Selection:**

- Relevant features/columns were identified based on the analysis goals.

- **Handling Outliers:**

- Outliers in the dataset were identified using statistical methods and visualization techniques.
- An appropriate approach (e.g., removal or replacement) was employed to address outliers and maintain data integrity.

- **Handling Time-Series Data:**

- Timestamps in the dataset were standardized to a common format.

- **Normalization:**

- Data normalization was performed to bring all variables to a common scale and this is done by calculating sub index of air pollutants as they are unitless.

The pre-processing steps described above successfully cleaned and prepared the air pollutants dataset from the CPCB historical dataset. The resulting pre-processed dataset is now ready for subsequent analysis, modelling, and gaining insights into air pollution patterns and trends.

M.Sc. (Statistics) Project Report

Table 3.2 Air Quality Index related Air Pollutants Dataset collected from CPCB

	Location	Data Station	Collection	From Date	To Date	PM2.5	PM10	NO	NO2	NOx	NH3	SO2	CO	Ozone	Benzene	Toluene	Eth-Benzene
0	delhi1	Alipur, DPCC	Delhi -	01-01-2019 00:00	02-01-2019 00:00	309.68	455.39	33.18	109.18	84.52	49.8	24.83	1.69	23.91	5.18	53.94	
1	delhi1	Alipur, DPCC	Delhi -	02-01-2019 00:00	03-01-2019 00:00	339.24	470.53	26.17	114.99	82.5	50.43	24.89	1.87	22	6.5	30.93	
2	delhi1	Alipur, DPCC	Delhi -	03-01-2019 00:00	04-01-2019 00:00	367.25	508.77	50.08	110.56	99.33	51.87	25.19	1.94	19.59	5.53	85.54	
3	delhi1	Alipur, DPCC	Delhi -	04-01-2019 00:00	05-01-2019 00:00	222.16	346	12.63	100.05	63.5	46.38	19.16	1.36	12.58	3.65	34.52	
4	delhi1	Alipur, DPCC	Delhi -	05-01-2019 00:00	06-01-2019 00:00	226.4	326.12	8.18	95.78	57.27	46.2	22.14	1.39	19.91	4.12	16.22	
...
66399	NaN	Wazirpur, DPCC	Delhi -	27-12-2022 00:00	28-12-2022 00:00	232.43	337.28	27	31.48	58.51	85.25	4.44	2.55	15.89	None	None	
66400	NaN	Wazirpur, DPCC	Delhi -	28-12-2022 00:00	29-12-2022 00:00	182.6	306.27	36.24	46.21	82.4	89.07	4.51	2.92	22.38	0	0	
66401	NaN	Wazirpur, DPCC	Delhi -	29-12-2022 00:00	30-12-2022 00:00	258.46	436.62	45.86	54.09	100.03	95.28	5.85	3.62	24.6	0	0	
66402	NaN	Wazirpur, DPCC	Delhi -	30-12-2022 00:00	31-12-2022 00:00	325.15	508.34	58.72	46.15	104.89	109.18	5.12	3.36	21.01	0	0	
66403	NaN	Wazirpur, DPCC	Delhi -	31-12-2022 00:00	01-01-2023 00:00	196.01	295.49	16.41	22.85	39.28	100.95	3.49	0.75	18.69	0	0	

Data description:

1. Location – Id of physical location
2. Data Collection Station - the name of the station where the data was collected
3. From Date - the start date of the data collection period
4. To Date - the end date of the data collection period
5. PM2.5 - the concentration of fine particulate matter with a diameter of 2.5 micrometres or less, measured in micrograms per cubic meter ($\mu\text{g}/\text{m}^3$)
6. PM10 - the concentration of particulate matter with a diameter of 10 micrometres or less, measured in micrograms per cubic meter ($\mu\text{g}/\text{m}^3$)
7. NO - the concentration of nitrogen monoxide, measured in micrograms per cubic meter ($\mu\text{g}/\text{m}^3$)
8. NO2 - the concentration of nitrogen dioxide, measured in micrograms per cubic meter ($\mu\text{g}/\text{m}^3$)
9. NOx - the concentration of nitrogen oxides, measured in micrograms per cubic meter ($\mu\text{g}/\text{m}^3$)
10. NH3 - the concentration of ammonia, measured in micrograms per cubic meter ($\mu\text{g}/\text{m}^3$)
11. SO2 - the concentration of sulphur dioxide, measured in micrograms per cubic meter ($\mu\text{g}/\text{m}^3$)
12. CO - the concentration of carbon monoxide, measured milligrams per cubic meter (mg/m^3)
13. Ozone - the concentration of ozone, measured in micrograms per cubic meter ($\mu\text{g}/\text{m}^3$)
14. Benzene - the concentration of benzene, measured in micrograms per cubic meter ($\mu\text{g}/\text{m}^3$)
15. Toluene - the concentration of toluene, measured in micrograms per cubic meter ($\mu\text{g}/\text{m}^3$)
16. Eth-Benzene - the concentration of ethylbenzene, measured in micrograms per cubic meter ($\mu\text{g}/\text{m}^3$)

These variables provide information on the levels of various pollutants in the air at different locations and time periods.

CHAPTER 4: STATISTICAL TECHNIQUES

In this chapter, we will explore various techniques used to study Air Quality Index and Air Pollutants. We will cover statistical techniques that are commonly used for data analysis and visualization. These techniques help us to identify trends, patterns, and outliers in the data, which can provide valuable insights into the quality of the air we breathe. So, let's dive into the world of statistical techniques and learn how they can be applied to study Air Quality Index and Air Pollutants.

4.1 Analysis using Data Visualization:

To better understand the data, we have used various graphical representation techniques, including box plots, histograms, scatter plots, time series plots, heatmaps, bar graphs, line graphs, and charts. These visualizations have helped us to identify trends and patterns in the data, such as the distribution of pollutant levels across different locations and the relationship between AQI and individual pollutant levels.

We have used packages like Matplotlib and Seaborn in Python to create these graphs, which offer a wide range of customization options and allow us to tailor the visualizations to our specific needs. Additionally, we have used the PowerBI tool to create more interactive and dynamic visualizations for even better graphical representation of the data.

Using graphical representation has helped us to gain a deeper understanding of the data and communicate our findings more effectively.

4.2 Correlation Analysis

Correlation analysis is a statistical method used to determine the relationship between two variables. By quantifying the degree of correlation, we can evaluate the strength of the relationship between two variables. This analysis helps us understand how much one variable change when the other one does.

The correlation coefficient is used to assign a value to the relationship between two variables. It ranges from -1 to +1, where 0 indicates no relationship, -1 indicates a perfect negative correlation, and +1 indicates a perfect positive correlation. The most commonly used correlation coefficient is the Pearson Correlation Coefficient, which tests for linear relationships between data.

To determine if there is a positive, negative, or neutral relationship between the variables in the dataset, we calculate the Pearson Correlation Coefficient using the following formula:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}}$$

Where,

r is the correlation coefficient,

x_i are the values of the x-variables in a sample,

\bar{x} is the mean of the values of the x-variable,

y_i are the values of the y-variables in a sample, and

\bar{y} is the mean of the values of the y-variable.

By conducting correlation analysis, we can determine the strength and direction of the relationship between variables in the dataset. This information can be used to make informed decisions and draw meaningful insights from the data.

4.3 Testing of Hypothesis:

The algorithm for testing the hypothesis of a significant difference in air pollutant levels between winter and summer using the Shapiro-Wilk test and the two-sample independent t-test can be described as follows:

- Shapiro-Wilk test for normality:
 - a. Separate air pollutant data and AQI measurements for both winter and summer.
 - b. Conduct the Shapiro-Wilk test separately for the air pollutant data in winter and summer.

H_0 : Data is normally distributed.

H_1 : Data not normally distributed.

- c. Calculate the test statistic and corresponding p-value for each season's data.
- d. Evaluate the p-values to determine if the data is normally distributed in both winter and summer. If the p-values are greater than the significance level (e.g., $\alpha = 0.05$), we fail to reject the null hypothesis, indicating that the data can be assumed to be normally distributed.

- Two-sample independent t-test:
 - a. Assuming the normality assumption is satisfied for both winter and summer data, proceed with the two-sample independent t-test.
 - b. Define the null hypothesis
$$H_0 : \text{There is no significant difference in the mean pollutant levels between winter and summer.}$$
$$H_1 : \text{There is significant difference in the mean pollutant levels between winter and summer.}$$
 - c. Collect the mean pollutant levels for winter and summer, respectively.
 - d. Calculate the test statistic and the associated p-value using the two-sample independent t-test.
 - e. Evaluate the p-value against the significance level (e.g., $\alpha = 0.05$) to determine the statistical significance of the difference in means.
 - f. If the p-value is less than the significance level, we reject the null hypothesis and conclude that there is a significant difference in the mean pollutant levels between winter and summer.

4.4 Cluster Analysis:

Cluster analysis is a powerful data analysis technique used to group similar objects or data points into distinct clusters. In the context of our objective to classify locations based on air pollutant levels, cluster analysis provides a systematic approach to identify and categorize locations that exhibit similar patterns of air pollution.

The primary goal of cluster analysis is to maximize the homogeneity within clusters while maximizing the heterogeneity between clusters. By grouping locations with similar air pollutant levels together, cluster analysis helps to uncover underlying patterns and relationships in the data, facilitating effective decision-making and targeted interventions in managing air pollution.

- **Algorithm (K-Means Clustering):**
 1. Initialize the algorithm: Begin by specifying the number of clusters, k . This determines the number of centroids that will be used to represent the clusters.
 2. Select initial centroids: Choose k initial centroid points randomly or based on some predefined criteria. These centroids serve as the starting points for the clusters.
 3. Assign observations to clusters:

- a. Calculate the Euclidean distance between each observation and the centroids.
- b. Assign each observation to the cluster whose centroid is closest to it. This is done by minimizing the Euclidean distance.
- c. Repeat steps a and b for all observations until each observation is assigned to a cluster.
4. Update centroid positions:
 - a. Recalculate the centroids of each cluster by taking the mean of all observations assigned to that cluster.
 - b. The centroid represents the centre of the cluster and is recalculated as the average of all observations within that cluster.
 - c. After updating the centroids, reassign observations to clusters based on the new centroid positions.
5. Repeat steps 3 and 4:
 - a. Iteratively perform steps 3 and 4 until the centroids stabilize and the observations' assignments to clusters remain unchanged or show minimal change.
 - b. This means that the algorithm has converged and no further adjustments are required.
6. Cluster formation: At convergence, you will have formed 4 clusters based on the observations of air pollutants and AQI. Each cluster will contain a group of observations that are similar to each other in terms of their pollutant and AQI values.
7. Classification of locations:
 - a. Assign each location's observation to the cluster it belongs to. This is determined by the proximity of the observation to the centroid of the respective cluster.
 - b. By classifying the locations' observations based on their cluster assignments, you can identify which cluster each location falls into, indicating the similarity of air pollutant levels and AQI at those locations.

4.5 Time Series Analysis and Forecasting

Time series analysis is a statistical technique used to analyse and interpret data that is collected and recorded over regular time intervals. It involves studying the patterns, trends, and dependencies within the data to make predictions or understand the underlying processes driving the observed behaviour.

In various fields such as economics, finance, meteorology, engineering, and social sciences, time series analysis is employed to gain insights, make forecasts, and support decision-making. By considering the temporal ordering of data points, time series

analysis accounts for the fact that observations at different time points may be interrelated, allowing us to capture and model the dynamic nature of the data.

- **Performing time series analysis typically involves the following steps:**

1. **Visualize Time Series Plot:** Time series plot is of great importance as it allows us to visually examine the patterns, trends, and seasonality present in the data. It provides an intuitive understanding of the temporal behaviour, helps identify outliers or irregularities, and guides the selection of appropriate time series analysis techniques.
2. **Decomposition:** Decompose the time series plot into its constituent components: trend, seasonality, and residuals. This step helps in understanding the individual contributions of these components and their impact on the overall behaviour of the series. Various decomposition methods, such as additive or multiplicative decomposition, can be applied.

The Classical Decomposition Model of Time Series Analysis is

$$X_t = m_t + s_t + z_t$$

Where,

$X_t = \text{Time Series}$

$m_t = \text{Trend Component}$

$s_t = \text{Seasonal Component}$

$z_t = \text{Noise Component}$

3. **Stationarity Analysis:** Assess the stationarity of the time series. Stationarity is desirable for many time series analysis techniques. Conduct statistical tests (e.g., ADF test) or inspect plots (e.g., rolling mean and standard deviation) to determine if the series is stationary. If non-stationarity is observed, consider differencing to achieve stationarity.

The time series $\{X_t, t \in Z\}$ is Stationary (weakly stationary) if,

$$1) E|X_t| < \infty \quad \forall t \in Z$$

$$2) E(X_t) = m \quad \forall t \in Z$$

$$3) \gamma_x(r, s) = \gamma_x(r + t, s + t) \quad \forall r, s, t \in Z$$

I. ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function)

These plots are commonly used in time series analysis to determine the order of autoregressive (AR) and moving average (MA) terms in a time series model. Additionally, stationarity tests are employed to assess the stationarity of the time series. Here's a breakdown of the plots and tests in terms of AR and MA parameters, along with their hypotheses and interpretations:

- **ACF Plot:**

AR(p) Parameter Interpretation: The ACF plot helps determine the order of the autoregressive parameter, p .

Interpretation:

If the ACF values decay gradually and become insignificant after a certain lag, it suggests an AR(p) process, where p is the last significant lag.

If the ACF values cut off abruptly after a certain lag, it indicates an AR(p) process, where p is the lag at which the ACF cuts off.

- **PACF Plot:**

MA(q) Parameter Interpretation: The PACF plot helps determine the order of the moving average parameter, q .

Interpretation:

If the PACF values cut off abruptly after a certain lag, it suggests an MA(q) process, where q is the last significant lag.

If the PACF values gradually decrease and become insignificant after a certain lag, it indicates an MA(q) process, where q is the lag at which the PACF cuts off.

If the time series is non-stationary, differencing can be applied to make it stationary. The order of differencing required can be determined by examining the ACF and PACF plots to see when the autocorrelation becomes insignificant or when the partial autocorrelation cuts off.

II. Stationarity Tests:

- **Augmented Dickey-Fuller (ADF) Test:**

Hypotheses:

Null Hypothesis (H_0): The time series is non-stationary.

Alternative Hypothesis (H_1): The time series is stationary.

Interpretation:

If the p-value from the ADF test is less than a chosen significance level (e.g., 0.05), we reject the null hypothesis and conclude that the time series is stationary.

If the p-value is greater than the significance level, we fail to reject the null hypothesis, indicating non-stationarity.

4. **Model Selection:** Choose an appropriate time series model based on the characteristics of the data and the analysis objectives. Common models include ARIMA (AutoRegressive Integrated Moving Average), SARIMA (Seasonal ARIMA), exponential smoothing, or

state-space models. The choice of model depends on the presence of trends, seasonality, and other patterns in the data.

- **ARIMA Model Fitting :**

ARIMA is defined as Auto Regressive Integrating Moving Average. ARIMA model combines three different models: The Auto-Regressive model, integrated model, and moving average model. ARIMA model can be applied to data, which is of non-stationary type. Non-stationary information is the data that does not have continuous successive intervals in the series. ARIMA model are generally denoted with p, d, q which are nonnegative integers

Where,

p is the number of time lags in the Auto-Regressive (AR) Model

d is the degree of differencing of Model

q is the order of the Moving Average (MA) Model

For $p, d, q \geq 0$, we say that a time series $\{X_t\}$ is an ARIMA (p, d, q) process if

$$Y_t = \nabla^d X_t = (1 - B)^d X_t$$

is ARMA (p, q). We can write

$$\varphi(B)(1 - B)^d X_t = \theta(B)Z_t$$

Time-series analysis-ARIMA is used to forecast the AQI. ARIMA model is the combination of three different individual models known as the Auto Regressive (AR) model denoted by p , Differencing (I) model indicated by d , Moving Average (MA) model denoted by q . The coefficients AR model and MR model are calculated with the help of Partial Auto-Correlation Function (PACF) and Auto-Correlation Function (ACF). The coefficient of the Differencing model depends on the number of times the data is differentiated. Differentiation relies on the stationarity of the data. The dickey-fuller test is performed to find whether the given data is stationary or not. The results of the dickey fuller test confirmed that the dataset is non-stationary. Hence, the data is differentiated by two times to make it stationary, and the coefficient of the differencing model (d) is calculated as 2. The p and q coefficients were obtained from PACF and ACF graphs. The flowchart for to check given model is ARIMA or not is given below.

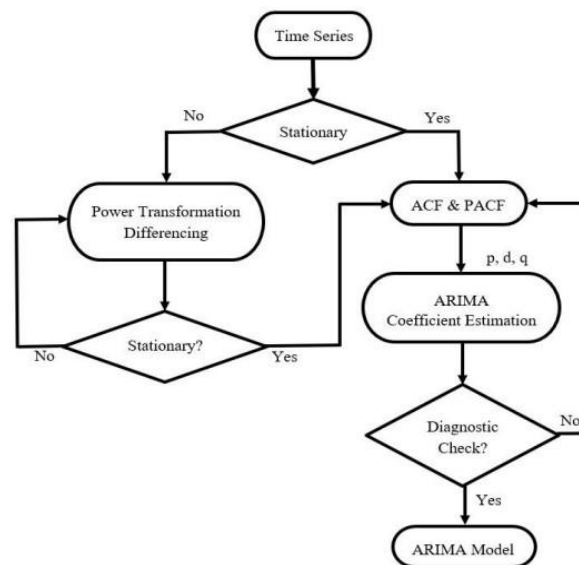


Figure 4.1 Block Diagram of ARIMA Model

Data transformation has been performed in the ARIMA model during data identification to make the non-stationary data to stationary data. A Stationary is a necessary condition for ARIMA Model. The stationary of the data is characterized by mean, standard -deviation, and auto-correlation structure. If the data present any trend, then applying the differencing and power transformation trend will be removed. Once the ARIMA model is identified, model parameters are estimated, and the final selected model is used for prediction purposes.

5. **Model Estimation:** Estimate the parameters of the selected time series model using the available data. This can be done through various techniques, such as maximum likelihood estimation or optimization algorithms, depending on the chosen model.
6. **Model Evaluation:** Evaluate the performance of the estimated model. Compare the predicted values against the actual values using evaluation metrics like MSE, RMSE, MAPE, or AIC/BIC. Assess the residuals for any remaining patterns or autocorrelation, which may indicate model inadequacy.

➤ **Performance Indices:**

The statistical criteria such as MAPE, RMSE, and MAE are used to evaluate each developed model's performance measure.

a) **Mean Absolute Percentage Error (MAPE)**

MAPE measures the accuracy of fitted time series values. It expresses accuracy as a percentage

$$MAPE = \frac{\sum_{i=1}^n \left| \frac{x_i - \hat{x}_i}{x_i} \right|}{n} \times 100, (x_i \neq 0)$$

b) Root Mean Squared Error (RMSE)

RMSE is the square root of the mean of the squared errors. RMSE indicates how close the predicted values are to the actual values. Hence, the lower RMSE value signifies that the model performance is good. It is calculated as

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \hat{x}_i)^2} \quad , (x_i \neq 0)$$

c) Mean Absolute Error (MAE)

MAE is the mean or average of the absolute value of the errors, the Predicted – Actual. It is calculated as

$$MAE = \frac{1}{n} \sum_{i=1}^n |x_i - \hat{x}_i| \quad , (x_i \neq 0)$$

d) Mean Absolute Deviation (MAD):

The average absolute difference between each data point and the mean of the dataset, indicating the average variability or dispersion of the data.

$$MAD = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}| \quad , (x_i \neq 0)$$

e) Mean Squared Deviation (MSD):

The average squared difference between each data point and the mean of the dataset, providing a measure of the overall variability or spread of the data.

$$MSD = \frac{\sum_{i=1}^n (x_i - \bar{x}_i)^2}{n} \quad . (x_i \neq 0)$$

7. Forecasting: Use the estimated model to make future predictions or forecasts. Extrapolate the model into the future based on the available historical data. Consider the uncertainty associated with the forecasts by calculating prediction intervals or using simulation techniques.

8. Model Monitoring and Updating: Continuously monitor the performance of the time series model as new data becomes available. Update the model and forecasts accordingly, ensuring that the analysis stays up-to-date and adapts to any changes in the underlying patterns.

CHAPTER 5: DATA ANALYSIS AND CONCLUSIONS

5.1 Graphical Analysis of Air Pollutants in Delhi:

Graphical representation can help to provide a quick and easy way to understand and compare air quality data, identify trends and patterns, and communicate key findings to stakeholders. It is an effective way to present complex data in a simple, visually appealing manner, making it easier for the reader to understand and engage with the information being presented.

5.1.1 Bar plot

The graph below shows the concentration of different air pollutants at various monitoring stations in Delhi. Each line represents a different pollutant, and the colours represent the AQI levels. The AQI is a measure of air quality that ranges from 0 to 500+, with higher values indicating poorer air quality.

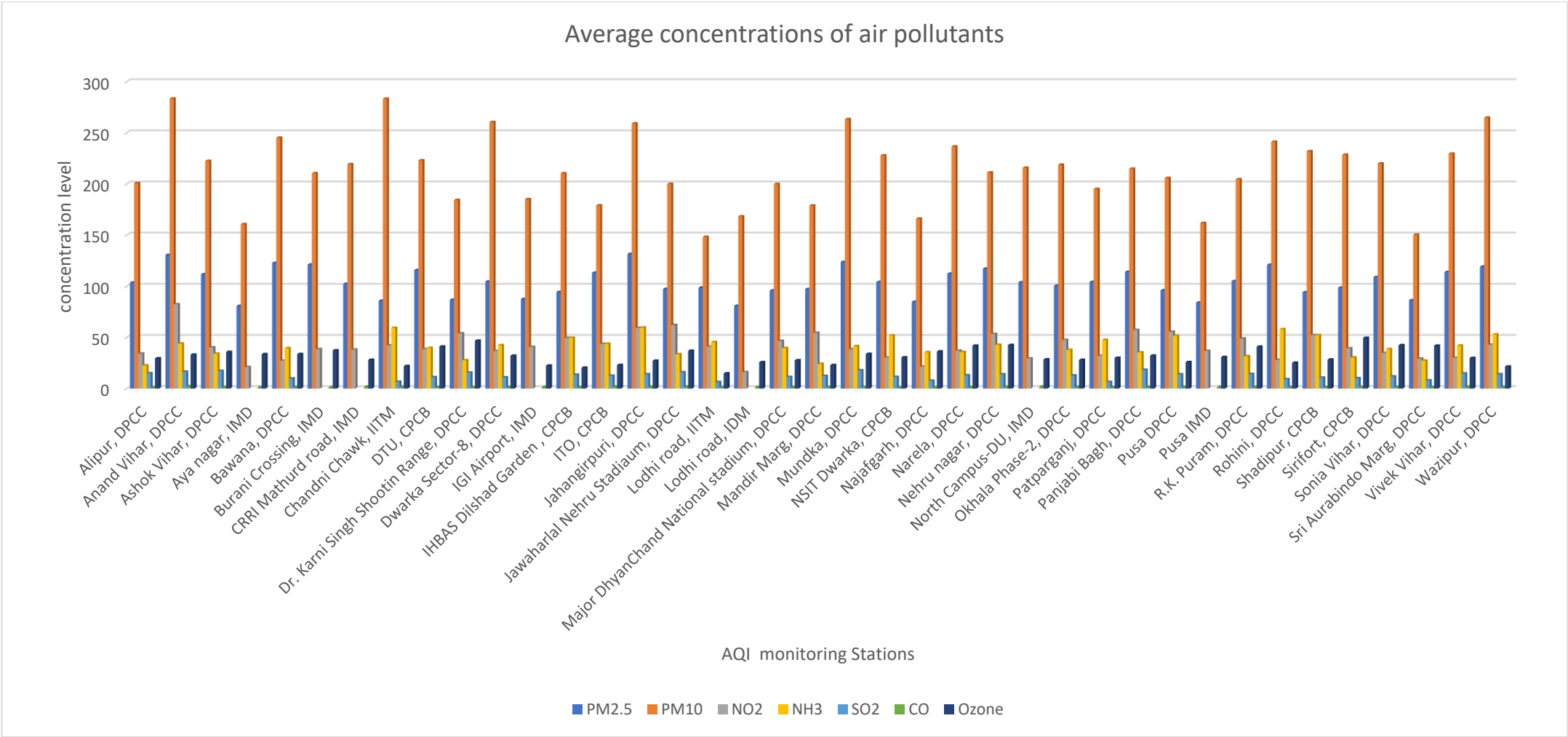


Figure 5.1 Average Air Pollutants Concentration for 39 Air monitoring station

Conclusion:

Based on the above plot, here are the main overall conclusions regarding the average concentration of air pollutants with respect to different locations:

1. Anand Vihar has the highest average concentrations of PM_{2.5}, PM₁₀, NO₂, and NH₃ among all locations, indicating severe pollution levels in that area.
2. Lodhi Road has relatively lower concentrations of all pollutants compared to other locations, suggesting relatively better air quality.
3. IHBAS Dilshad Garden and Mundka have higher concentrations of NH₃ and SO₂ compared to other pollutants, which might be attributed to specific local factors or sources.
4. Alipur, Dr.Karni singh shooting range, and Dwarka sector-8 exhibit moderate concentrations of various pollutants, indicating a moderate level of pollution in those areas.
5. Several locations have missing values for NH₃ and SO₂, indicating a lack of data or measurement for these pollutants in those areas.

Overall, PM_{2.5}, PM₁₀, NO₂, and CO are the prominent pollutants across most locations, highlighting the need for targeted measures to address these pollutants and improve air quality.

5.1.2 Line Chart

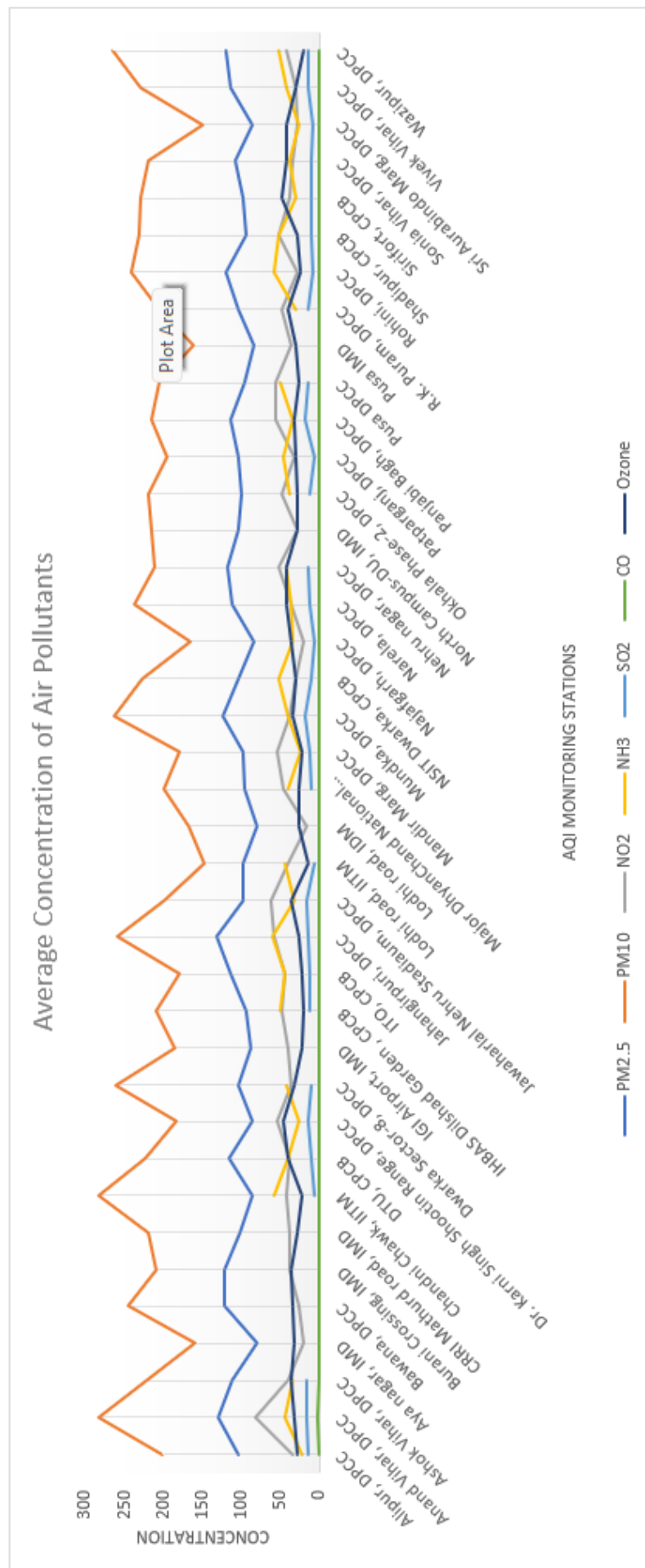


Figure 5.2 Air pollutants concentration (from Jan 2018-Dec 2022)

Conclusion:

1. Variation in Air Pollutant Concentrations: from above line plot we can observe that there is variation in the concentrations of different pollutants, indicating spatial differences in pollution levels within the city.
2. High Concentrations of PM_{2.5} and PM₁₀: PM_{2.5} and PM₁₀ are particulate matter pollutants, and their concentrations tend to be relatively high across most locations. This suggests a significant presence of fine and coarse particles in the air, which can have adverse health effects.
3. Variation in Gaseous Pollutants: Gaseous pollutants like NO₂, SO₂, CO, and Ozone show varying concentrations across different locations. This indicates variations in emission sources, local meteorological conditions, and the effectiveness of pollution control measures in different areas.

5.1.3 Yearly and Monthly box plot for AQI

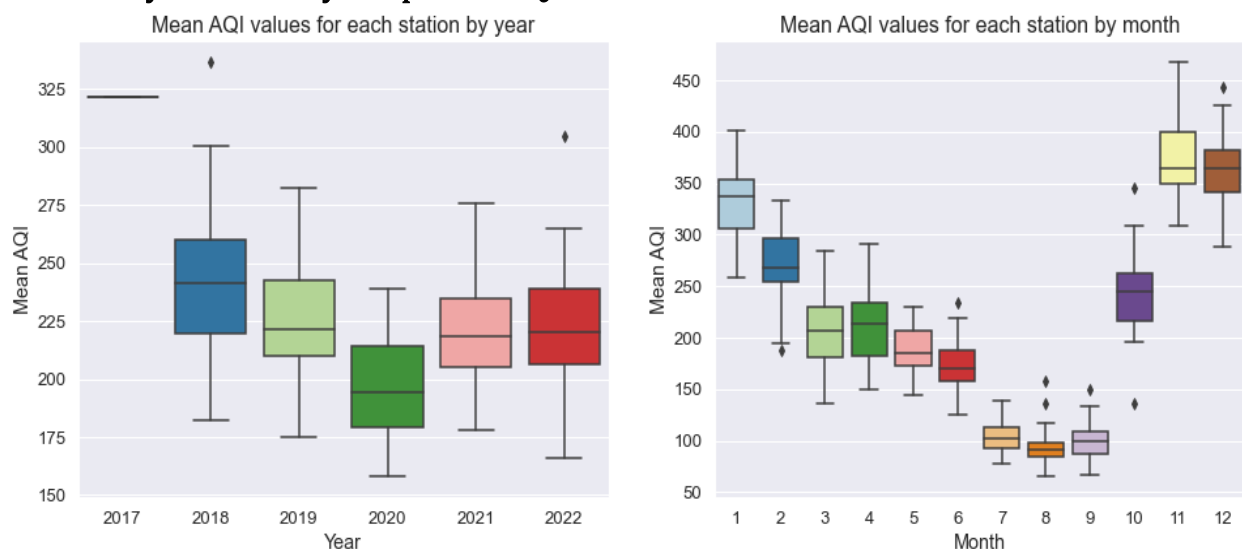


Figure 5.3 Yearly and Monthly box plot for AQI

- The analysis of the AQI concentration in Delhi from 2018 to 2022 reveals important findings, as follows:

The year 2018 had the highest AQI concentration among all the years observed, while 2020 had the lowest AQI concentration. This decrease in AQI concentration can be attributed to the COVID-19 pandemic-induced lockdown, which restricted human activities and reduced air pollution.

The monthly box plot indicates that the AQI level decreases from January to September, with a slight increase in October, followed by a significant increase from November to December and again a slight decrease in January. This pattern may be

attributed to increased crop residue burning and a decrease in temperature during the winter months.

The year-wise AQI box plot suggests that there is a decrease in the AQI concentration average for each monitoring station from 2018 to 2020. However, after 2020, the AQI concentration has started to increase again. This indicates that the measures implemented during the COVID-19 lockdowns were effective in reducing air pollution, but further action is required to maintain these levels.

The month-wise average AQI box plot indicates that the AQI level is highest during November and December and lowest during July, August, and September. This is because of increased crop residue burning during the winter months, and the onset of monsoons in the summer months, which helps to reduce air pollution.

The analysis of the AQI concentration in Delhi from 2018 to 2022 suggests that air pollution remains a significant concern for the city's residents. While the COVID-19 lockdowns helped to reduce air pollution levels, further measures are required to sustain these levels. The findings of this report can guide policymakers and researchers to develop effective strategies to improve air quality in Delhi and ensure a healthier environment for the citizens of the city.

5.2 Descriptive statistics among the monitoring stations and pollutant:

Table 5.1 Descriptive statistics for PM2.5

Location	count	mean	std	min	25%	50%	75%	max
delhi1	1445	103.5	83.0	4.8	42.9	76.4	139.7	712.1
delhi2	1739	130.6	99.4	9.5	57.1	97.7	175.1	592.3
delhi3	1773	111.7	94.1	6.0	43.8	79.9	150.4	601.5
delhi4	1782	80.7	62.6	7.1	37.5	60.8	105.3	560.9
delhi5	1633	122.8	92.3	7.8	52.9	95.2	167.2	696.8
delhi6	793	121.1	79.9	8.0	68.6	100.7	146.8	572.1
delhi7	1812	102.3	79.0	7.6	43.8	78.4	136.6	521.8
delhi8	577	85.9	50.8	17.1	42.5	78.4	115.2	391.3
delhi9	2139	115.7	96.6	0.0	45.5	87.2	158.2	741.0
delhi10	1750	86.8	77.2	0.5	34.3	56.4	119.0	571.7
delhi11	1793	104.6	83.6	8.1	42.2	78.7	143.7	600.4
delhi13	1658	87.5	67.7	1.3	37.3	68.2	115.8	506.2
delhi14	1817	94.3	68.7	6.7	47.2	78.2	120.1	684.9
delhi15	1785	113.2	84.5	0.0	54.5	87.4	147.6	659.3
delhi16	1791	131.6	105.3	8.8	51.4	95.4	184.6	658.3
delhi17	1778	97.4	83.4	3.8	37.5	68.2	131.1	793.6
delhi18	545	98.6	75.9	1.3	49.5	83.6	124.7	750.5
delhi19	1781	80.8	59.7	4.6	39.0	62.4	103.2	485.5
delhi20	1792	95.8	74.9	6.7	40.7	71.7	128.5	525.6
delhi21	1818	97.2	75.2	4.7	42.0	76.4	129.3	563.9
delhi22	1634	123.7	102.6	7.4	44.5	92.6	176.3	699.0
delhi23	1822	104.0	65.3	8.3	55.4	91.1	135.5	429.9
delhi24	1775	84.7	64.7	0.3	35.4	69.2	114.3	574.4
delhi25	1774	112.3	84.4	6.6	48.7	87.9	153.8	689.1
delhi26	1792	117.2	102.7	6.1	42.2	76.5	162.3	636.0
delhi27	1005	103.8	82.8	4.4	44.3	77.1	136.8	571.4
delhi28	1793	100.4	85.6	0.0	39.2	71.2	135.2	547.6
delhi29	1787	104.1	86.5	3.1	42.6	75.3	138.0	634.0
delhi30	1813	114.0	89.1	6.8	49.1	84.7	151.4	609.2
delhi31	1615	96.1	80.9	3.7	36.8	68.1	137.2	570.6
delhi32	1682	84.0	65.8	6.9	38.3	62.9	108.0	561.8
delhi33	1810	105.0	82.9	6.4	44.4	76.8	142.0	558.9
delhi34	1778	120.9	99.0	6.1	49.2	87.5	162.0	762.0
delhi35	1821	94.1	67.9	9.4	41.9	77.0	125.8	448.5
delhi36	1774	98.6	78.2	0.0	40.8	74.8	132.9	573.1
delhi37	1785	109.0	86.9	6.0	46.8	80.1	146.8	598.8
delhi38	1641	86.3	72.8	5.3	34.2	61.0	117.5	534.9
delhi39	1780	113.9	95.1	2.6	46.3	80.0	152.7	650.9
delhi40	1788	119.2	94.8	6.9	51.8	86.4	155.9	630.1

Bar plot of average concentration of PM2.5

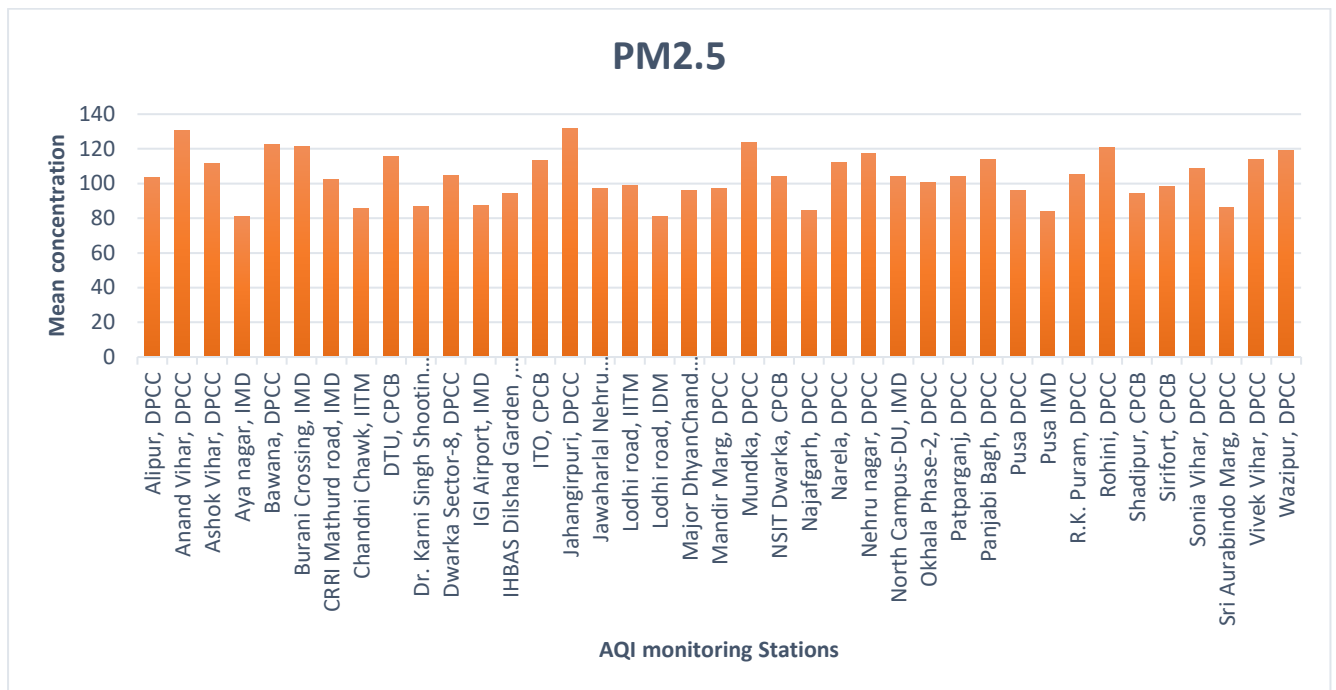


Figure 5.4 Mean of pollutant PM2.5 (Jan-2018 to Dec-2022)

Conclusions:

- The table shows the mean, standard deviation, minimum, maximum, and percentiles of PM2.5 concentration in 39 locations in Delhi. The data are based on the readings from the years prior to 2018-22. We can observe that here the overall average level of PM2.5 is $103.93 \mu\text{g}/\text{m}^3$ for Delhi.
- The mean PM2.5 concentration ranges from $80.74 \mu\text{g}/\text{m}^3$ to $131.57 \mu\text{g}/\text{m}^3$ across different locations, with an overall mean of $102.86 \mu\text{g}/\text{m}^3$. The highest mean PM2.5 concentration is observed in Jahangirpuri, while the lowest is observed in Lodhi road, IITM.
- The standard deviation of PM2.5 concentration ranges from $50.82 \mu\text{g}/\text{m}^3$ to $105.32 \mu\text{g}/\text{m}^3$, with an overall standard deviation of $74.59 \mu\text{g}/\text{m}^3$. The highest standard deviation is observed in Jahangirpuri, while the lowest is observed in Chandni Chawk.
- The minimum and maximum PM2.5 concentrations observed in the locations are $0.00 \mu\text{g}/\text{m}^3$ and $793.55 \mu\text{g}/\text{m}^3$, respectively. DTU and Jawaharlal Nehru Stadium have the highest and the second-highest maximum PM2.5 concentrations, respectively.
- In the barplot we can see that the average minimum concentration of PM2.5 is 8.7 (Aya nagar) and average maximum concentration of PM2.5 is 131.6(Jahangirpuri).

Table 5.2 Descriptive statistics for PM₁₀

Location	Count	mean	std	min	25%	50%	75%	max
delhi1	1446	200.7	125.5	10.0	97.8	182.6	281.4	758.4
delhi2	1727	283.4	155.6	16.3	154.9	268.7	388.3	729.9
delhi3	1791	222.4	134.4	11.8	112.4	198.6	307.4	935.6
delhi4	1773	160.7	97.5	9.8	86.0	146.3	214.4	705.6
delhi5	1603	245.1	144.5	12.0	122.8	222.8	344.3	810.9
delhi6	798	210.4	121.6	19.4	121.8	183.1	275.1	781.6
delhi7	1756	219.3	131.6	9.5	108.0	205.8	303.8	872.1
delhi8	580	283.2	138.5	42.3	180.4	271.1	364.5	790.2
delhi9	1711	222.9	133.1	1.0	115.0	206.3	306.7	966.5
delhi10	1790	184.2	110.7	8.0	93.5	169.6	250.9	756.5
delhi11	1788	260.5	140.2	14.6	142.0	253.7	358.7	866.3
delhi13	1651	185.1	106.9	12.5	100.6	165.4	248.0	746.4
delhi14	893	210.4	119.1	33.8	112.5	198.8	275.4	710.4
delhi15	1633	178.9	100.4	16.0	101.6	166.4	234.7	691.8
delhi16	1787	259.1	149.7	15.5	134.0	241.6	354.3	928.9
delhi17	1780	199.9	121.7	1.0	102.9	184.2	272.3	934.2
delhi18	543	148.2	90.2	2.4	81.7	133.6	196.1	704.6
delhi19	1780	168.3	96.8	8.2	91.1	155.4	228.3	860.3
delhi20	1785	199.9	118.0	12.4	104.3	182.2	274.5	837.9
delhi21	1818	178.9	99.7	17.1	97.5	167.6	239.2	755.2
delhi22	1629	263.2	150.2	10.7	131.9	246.1	369.4	791.0
delhi23	908	227.9	128.0	31.7	118.0	221.3	311.8	826.7
delhi24	1762	166.0	99.0	4.3	89.0	156.2	222.9	904.6
delhi25	1776	236.6	131.4	20.5	129.4	215.2	327.1	906.3
delhi26	1785	211.2	132.4	10.5	102.3	190.4	288.6	900.5
delhi27	984	215.8	136.8	21.4	109.0	190.6	289.4	859.3
delhi28	1782	218.8	127.3	9.7	115.8	202.2	295.9	869.8
delhi29	1760	195.1	120.1	8.4	98.1	175.6	269.3	873.6
delhi30	1808	214.9	124.2	20.6	115.0	192.6	289.4	768.3
delhi31	1615	205.8	119.3	10.0	107.5	194.2	283.9	726.9
delhi32	1677	161.8	102.0	12.1	81.2	143.0	218.1	824.6
delhi33	1801	204.5	116.6	11.5	105.4	193.5	278.2	727.2
delhi34	1786	241.2	145.6	11.1	121.3	215.2	336.9	942.0
delhi35	907	231.9	134.8	24.3	120.7	213.1	314.2	740.9
delhi36	1685	228.5	127.0	11.0	120.7	220.8	310.2	902.0
delhi37	1771	219.9	129.7	13.0	112.6	200.0	302.6	815.5
delhi38	1639	150.5	93.8	8.9	72.7	136.7	207.1	596.1
delhi39	1785	229.6	134.1	12.5	116.8	205.0	313.5	729.3
delhi40	1699	264.8	146.0	19.3	151.5	239.1	349.7	952.0

Bar plot of average concentration of PM10

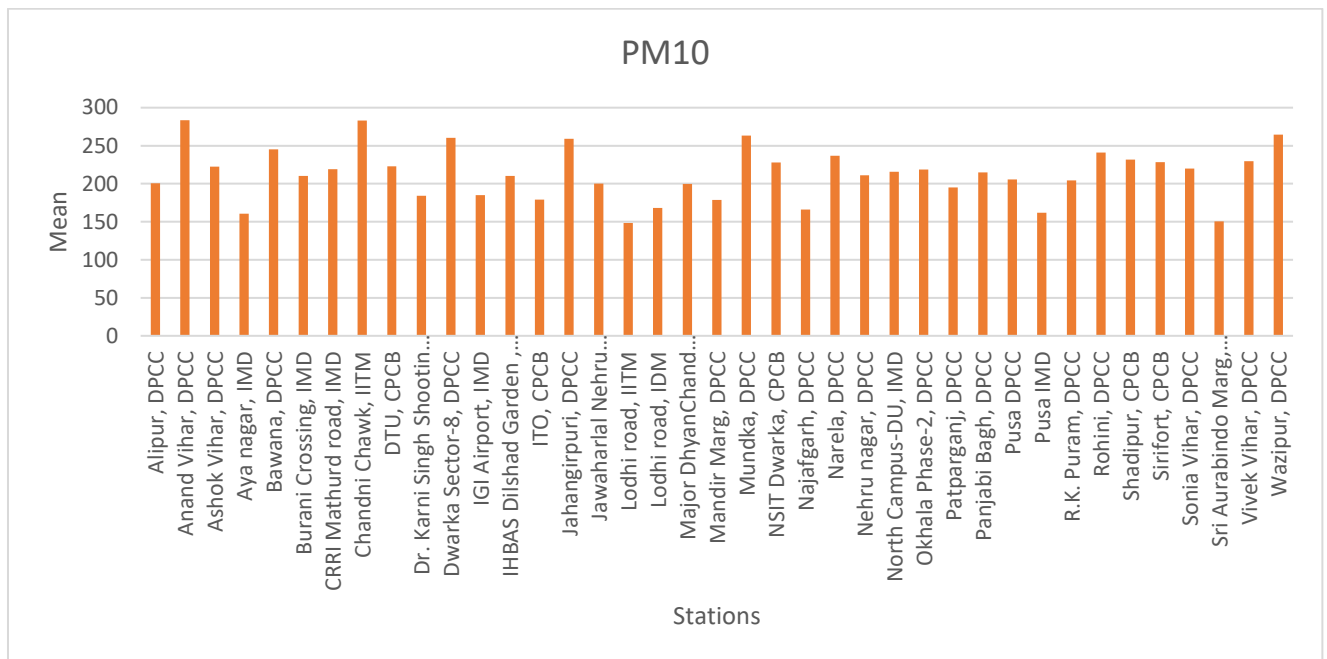


Figure 5.5 Mean of pollutant PM10 (Jan-2018 to Dec-2022)

Conclusions:

- We can observe that here the overall average level of PM10 is $213.06 \mu\text{g}/\text{m}^3$ for Delhi.
- The mean PM10 values for each location vary from 148.23 to 283.37, with a standard deviation ranging from 90.22 to 155.57. This indicates that there is considerable variation in PM10 levels across the different locations in Delhi.
- The highest maximum PM10 values are observed in DTU, Dwarka sector-8, and jahangirpuri, with values of 966.46, 866.33, and 928.87, respectively. These locations may warrant closer attention in terms of air quality management.
- The median PM10 values for each location range from 133.62 to 268.74. Locations with median PM10 values above 200 include Anand vihar, Bawana, CRRI mathurd road, Chandni Chowk, Dwarka sector-8, jahangirpuri, Mundka, Narela, Nehru nagar, Okhala Phase-2, Panjabi bagh, and Pusa DPCC. These locations may also require further monitoring and air quality improvement measures.

Table 5.3 Descriptive statistics for AQI

Location	count	mean	std	min	25%	50%	75%	max	Colour code
delhi1	1445	213.0	131.9	22.2	100.8	183.0	319.2	810.5	
delhi2	1741	280.2	154.4	42.5	144.4	270.8	374.5	774.9	
delhi3	1793	229.2	140.4	25.0	109.9	199.5	330.2	1032.0	
delhi4	1800	177.8	107.7	28.3	96.3	139.7	257.8	744.5	
delhi5	1640	251.0	144.2	19.0	121.2	239.7	350.0	876.1	
delhi6	829	233.2	125.8	34.6	129.4	213.3	317.9	839.5	
delhi7	1813	224.0	128.6	30.2	112.8	198.1	316.0	952.6	
delhi8	580	263.9	139.1	43.1	158.8	234.8	331.0	850.2	
delhi9	1881	241.5	143.9	7.1	115.4	226.1	339.5	1070.6	
delhi10	1793	193.4	120.6	19.7	98.5	156.1	301.5	808.1	
delhi11	1794	247.9	138.8	24.5	129.6	231.6	337.6	945.4	
delhi13	1659	195.5	113.0	29.7	103.0	160.8	292.5	795.5	
delhi14	1817	200.8	118.2	21.0	98.1	178.1	300.4	750.5	
delhi15	1774	220.3	124.2	29.0	107.9	199.5	320.9	727.3	
delhi16	1792	258.4	153.3	24.0	123.5	244.1	356.8	1023.6	
delhi17	1782	209.4	128.1	25.0	105.0	172.4	312.5	1030.2	
delhi18	545	196.9	120.7	25.8	91.4	178.6	303.6	785.0	
delhi19	1783	181.6	105.9	25.7	99.2	151.9	255.7	937.9	
delhi20	1792	205.7	124.0	26.5	103.5	172.2	310.8	909.9	
delhi21	1820	202.1	120.4	24.0	99.3	171.2	308.9	806.5	
delhi22	1635	263.7	151.5	23.0	123.9	260.1	362.5	851.2	
delhi23	1822	223.5	119.4	24.0	110.6	215.8	316.1	895.9	
delhi24	1776	189.2	114.8	17.0	93.2	170.6	290.3	993.2	
delhi25	1776	238.3	133.3	29.2	120.0	215.6	333.6	995.3	
delhi26	1795	226.2	144.3	35.0	103.3	186.0	333.6	988.1	
delhi27	1004	224.4	142.8	27.3	105.2	195.1	321.3	936.6	
delhi28	1794	216.7	130.6	26.7	110.5	182.8	314.6	949.7	
delhi29	1789	208.6	132.5	21.3	97.3	174.3	316.7	954.5	
delhi30	1814	228.9	131.6	37.5	114.0	200.8	326.0	822.9	
delhi31	1616	207.6	124.3	24.1	105.9	177.3	315.9	771.1	
delhi32	1683	182.4	112.2	14.0	99.9	142.7	268.0	893.3	
delhi33	1817	216.7	126.0	28.3	105.8	187.4	319.3	771.5	
delhi34	1787	245.8	148.6	15.5	116.5	227.8	343.4	1040.0	
delhi35	1821	210.7	127.3	33.5	97.4	181.7	309.5	788.6	
delhi36	1774	228.9	125.3	20.5	123.8	208.5	318.4	990.0	
delhi37	1785	225.0	133.2	19.5	110.2	194.3	325.6	881.9	
delhi38	1641	180.9	119.5	21.5	78.2	143.1	299.2	619.1	
delhi39	1789	234.0	137.7	19.5	113.1	200.8	330.7	774.1	
delhi40	1788	255.3	149.4	27.6	133.9	223.0	343.5	1052.5	

Bar plot of average concentration of AQI

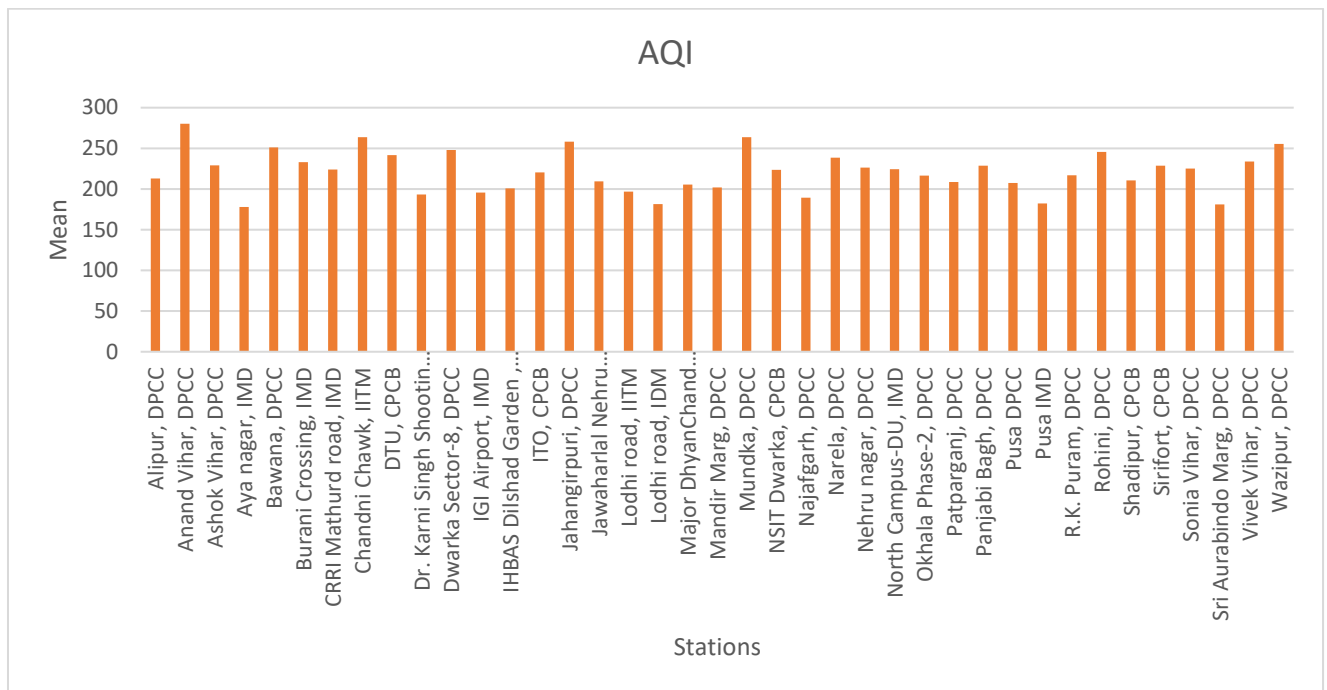


Figure 5.6 Mean of Air Quality Index (AQI)

Conclusion:

- We can observe that here the overall average level of AQI is 221.34 for Delhi.
- There is a significant variation in AQI levels across different monitoring stations in Delhi. For example, Anand vihar has the highest mean AQI of 280.2, while Aya nagar has the lowest mean AQI of 177.8.
- Certain locations are identified as pollution hotspots due to consistently higher AQI values. For instance, Mundka has the highest maximum AQI of 1052.5, indicating potentially severe air pollution levels. Similarly, Wazipur has a relatively high mean AQI of 255.3.
- Some areas demonstrate relatively better air quality conditions. Lodhi road has a lower mean AQI of 196.9 and delhi38 has the lowest maximum AQI of 619.1, suggesting comparatively cleaner air in these locations.
- The standard deviation values indicate the variability in AQI levels within each location. For instance, Anand vihar has a higher standard deviation of 154.4, implying greater fluctuations in air quality over time compared to other locations.
- The variations in AQI levels among different monitoring stations highlight the need for targeted interventions and pollution control measures in specific areas. For example,

Mundka, Wazipur, and Anand vihar may require focused efforts to address the factors contributing to higher air pollution levels.

➤ **Proportion of simple index for pollutant to calculate AQI**

The table below presents the final AQI values for each pollutant in Delhi, calculated by taking the maximum of the simple indices. The counts represent the number of observations where the final AQI value corresponds to the maximum simple index.

Table 5.4 Proportion of days simple index of pollutant was responsible to calculate AQI

Statistics	PM2.5	PM10	NO2	NH3	CO	SO2	Ozone	Total AQI
Count	29610	31122	771	45	2754	28	732	65062
Proportion	45.51	47.83	1.19	0.07	4.23	0.04	1.13	100

The analysis of the AQI data for Delhi from 2018 to 2022 reveals important findings, as follows:

1. The final AQI values are determined by taking the maximum of the simple indices calculated for each pollutant. The maximum simple index is indicative of the severity of the air pollution, and hence, the final AQI value.
2. The highest number of observations for the final AQI values are for PM10 and PM2.5, with 31,122 and 29,610 observations, respectively. This indicates that these two pollutants are major contributors to air pollution in Delhi.
3. The proportions of the final AQI values for each pollutant indicate that PM10 and PM2.5 are the most dominant pollutants, accounting for 47.8% and 45.5% of the total AQI values, respectively. This is followed by CO, with 4.2% of the total AQI values, and Ozone, with 1.1% of the total AQI values.
4. The total number of observations for AQI values is 65,062, which is the sum of all final AQI values calculated for the 39 monitoring locations in Delhi. This indicates that air pollution is a significant concern for the city's residents, and measures are required to improve the air quality.

In the above table, it can be observed that the proportion of NH3 and SO2 is relatively low compared to the main air pollutants, PM2.5 and PM10. Among all the AQI monitoring stations, the highest concentrations of NH3 and SO2 are recorded at Bawana,

DPCC, IHBAS Dilshad Garden, CPCB, Okhla Phase-2, DPCC, and Patparganj, DPCC. The reason behind this is as follows:

- 1) Agricultural activities: Bawana, DPCC and Patparganj, DPCC are located in areas with significant agricultural activities. Ammonia emissions can occur from the use of fertilizers and the management of livestock waste in these agricultural areas.
- 2) Industrial emissions: IHBAS Dilshad Garden, CPCB and Okhla Phase-2, DPCC are situated near industrial zones where various manufacturing and industrial processes take place. Industries such as chemical manufacturing, pharmaceuticals, and waste treatment facilities can release Ammonia as a by-product, contributing to its presence in the air.

- **Distribution of AQI in Health Standards (In Percentage)**

This table provides information on the distribution of Air Quality Index (AQI) in health standards for different locations in Delhi. The AQI is categorized into seven categories - Good, Satisfactory, Moderate, Poor, Very Poor, Severe, and Most Severe. In addition to this, the table provides the percentage of AQI values falling below 300 and above 300.

Table 5.5 Distribution of AQI in Health Standards (In Percentage)

Location	Good	Satisfactory	Moderate	Poor	Very Poor	Severe	Most Severe	Below 300	Above 300
	0-50	51-100	101-200	201-300	301-400	401-500	>500		
delhi1	6.5	18.0	28.9	14.0	24.5	6.3	1.8	53.4	46.6
delhi2	0.2	10.2	28.2	16.2	24.0	11.3	10.0	38.5	61.5
delhi3	3.4	17.4	29.3	15.4	23.2	7.9	3.4	50.1	49.9
delhi4	3.1	24.1	39.4	13.3	17.7	1.9	0.6	66.5	33.5
delhi5	1.6	14.3	28.0	13.6	28.9	8.8	4.9	43.8	56.2
delhi6	1.6	12.1	33.7	18.2	24.5	7.6	2.4	47.3	52.7
delhi7	1.3	17.7	31.9	17.6	23.7	4.7	3.0	50.9	49.1
delhi8	0.2	6.0	31.7	26.9	21.6	6.7	6.9	37.9	62.1
delhi9	3.1	16.1	26.3	15.2	26.6	9.3	3.6	45.4	54.6
delhi10	5.7	20.1	36.3	12.2	20.4	4.1	1.3	62.1	37.9
delhi11	1.6	11.8	30.4	18.3	25.1	7.9	4.9	43.8	56.2
delhi13	1.9	20.8	36.6	16.6	20.6	2.5	1.1	59.3	40.7
delhi14	4.0	21.5	30.4	18.8	19.6	4.0	1.6	55.9	44.1
delhi15	2.4	18.3	29.5	15.2	27.4	5.6	1.7	50.1	49.9
delhi16	1.5	14.3	27.7	13.8	26.0	10.2	6.4	43.5	56.5
delhi17	3.4	19.2	33.3	14.6	22.6	5.0	1.9	55.9	44.1
delhi18	5.9	25.0	22.8	19.8	22.9	2.6	1.1	53.6	46.4
delhi19	2.7	22.8	40.2	14.5	17.3	2.0	0.6	65.7	34.3
delhi20	4.1	19.6	32.5	14.1	23.5	4.7	1.6	56.1	43.9
delhi21	3.5	22.1	31.6	14.5	22.6	4.6	1.2	57.1	42.9
delhi22	1.3	13.5	27.8	13.5	26.5	10.5	7.0	42.5	57.5
delhi23	0.9	20.5	24.9	19.6	28.1	4.5	1.5	46.3	53.7
delhi24	7.8	19.6	32.0	17.1	20.5	2.6	0.4	59.4	40.6
delhi25	1.0	14.4	31.8	14.8	27.9	7.2	3.0	47.1	52.9
delhi26	3.5	20.3	28.7	11.6	23.8	8.7	3.3	52.5	47.5
delhi27	3.5	18.3	29.6	15.9	23.0	5.5	4.2	51.4	48.6
delhi28	2.3	18.3	33.3	15.1	23.2	4.8	3.0	54.0	46.0
delhi29	5.9	20.2	29.5	13.8	21.9	6.9	1.8	55.6	44.4
delhi30	1.3	17.3	31.4	13.6	26.7	7.0	2.8	49.9	50.1
delhi31	5.6	16.5	33.4	14.2	24.3	4.5	1.5	55.6	44.4
delhi32	3.9	21.3	39.9	13.0	18.7	2.4	0.8	65.1	34.9
delhi33	3.1	19.0	30.5	14.7	25.6	5.3	1.8	52.6	47.4
delhi34	1.7	16.5	28.4	12.9	26.0	9.1	5.3	46.7	53.3
delhi35	2.4	23.4	28.2	17.0	21.7	4.9	2.3	54.0	46.0
delhi36	1.4	13.4	33.7	18.4	25.5	5.3	2.3	48.5	51.5
delhi37	1.7	18.2	30.8	16.5	23.1	6.8	3.0	50.7	49.3
delhi38	8.8	26.0	28.0	12.2	21.6	2.7	0.6	62.9	37.1
delhi39	0.4	17.4	32.0	14.1	24.1	8.0	4.0	49.8	50.2
delhi40	0.8	9.9	35.0	15.8	23.6	7.5	7.4	45.7	54.3

Based on the distribution of AQI in different locations in Delhi, the following conclusions can be drawn:

1. Delhi experiences air pollution levels that exceed the National Ambient Air Quality Standards (NAAQS) for most of the year. This is indicated by the fact that in all the locations analysed, the percentage of days with AQI in the "Poor" or worse category (i.e., above 100) is higher than the percentage of days with AQI in the "Good" or "Satisfactory" category (i.e., below 100).
2. The severity of air pollution in Delhi varies across different locations, with some areas experiencing higher levels of pollution than others. For example, the percentage of days with AQI in the "Severe" or worse category (i.e., above 400) ranges from 1.1%(IGI Airport, Lodhi road) to 11.3%(Anand Vihar) across the locations analysed.
3. The most common level of air pollution in Delhi is "Poor," with the percentage of days with AQI in this category ranging from 11.6%(Nehru nagar) to 26.9%(Chandni Chawk) across the locations analysed.
4. The least common level of air pollution in Delhi is "Good," with the percentage of days with AQI in this category ranging from 0.2%(Anand Vihar) to 7.8%(Najafgarh) across the locations analysed.
5. The analysis shows that a significant proportion of days in Delhi (ranging from 37.9% to 66.5% across the locations analysed) have AQI levels above 300, which is classified as "Severe" or "Most Severe" and poses a serious health risk to the population.

These findings highlight the urgent need for effective measures to address air pollution in Delhi, such as reducing vehicular emissions, increasing green cover, promoting the use of clean energy sources, and implementing stricter industrial regulations.

➤ **Overall AQI classification in Percentage**

Table 5.6 Overall AQI classification in Percentage

	Good (0-50)	Satisfactory (51-100)	Moderate (101- 200)	Poor (201-300)	Very Poor (301- 400)	Severe (401- 500)	Most Severe >500	total
Count	1923	11702	20380	9886	15373	3902	1896	65062
Percentage	2.96	17.99	31.32	15.19	23.63	6.00	2.91	100

Conclusions:

- 1) **Overall AQI Distribution:** The majority of the observations fall within the Moderate to Very Poor AQI categories, with a significant proportion in the Poor category. This indicates that the air quality in Delhi is generally not meeting the desired standards.
- 2) **Poor and Very Poor Air Quality:** The combined percentage of observations in the Poor and Very Poor categories is approximately 46.51% (31.32% + 15.19%). This indicates that almost half of the recorded AQI observations in Delhi indicate poor air quality, posing potential health risks to the population.
- 3) **Most Severe Air Quality:** While the percentage of observations in the Most Severe category is relatively low (2.91%), it is still a cause for concern. These high AQI levels suggest extremely poor air quality conditions in certain instances, which can have severe health impacts on the population.
- 4) **Satisfactory and Good Air Quality:** The percentage of observations in the Satisfactory and Good categories is relatively low (2.96% and 17.99%, respectively). This indicates that periods of satisfactory or good air quality are relatively rare in the dataset, emphasizing the need for measures to improve the overall air quality in Delhi.

5.3 Weighted Average of AQI

Table 5.7 Station with range of AQI and area and information

Station Name	AQI Count n_i	AQI Sum $\sum_{i=1}^{40} s_i$	Area (Kmsq) A_i	$A_i \times \sum_{i=1}^{40} s_i$	$n_i * A_i$
Alipur	1445	307785	56.2	17297517	81209
Anand Vihar	1741	487828.2	5.28	2575732.896	9192.48
Ashok Vihar	1793	410955.6	7.3	2999975.88	13088.9
Aya nagar	1800	320040	NA	NA	NA
Bawana	1640	411640	19.08	7854091.2	31291.2
Burani Crossing	829	193322.8	7.68	1484719.104	6366.72
CRRRI Mathurd road	1813	406112	1.2	487334.4	2175.6
Chandni Chawk	580	153062	0.6	91837.2	348
DTU	1881	454261.5	119	54057118.5	223839
Dr. K. S. S. R.	1793	346766.2	0.1	34676.62	179.3
Dwarka Sector-8	1794	444732.6	56.3	25038445.38	101002.2
IGI Airport	1659	324334.5	5.36	1738432.92	8892.24
IHBAS Dilshad Garden	1817	364853.6	5	1824268	9085
ITO	1774	390812.2	3.85	1504626.97	6829.9
Jahangirpuri	1792	463052.8	1.16	537141.248	2078.72
Jawaharlal Nehru Stadium	1782	373150.8	0.12	44778.096	213.84
Lodhi road, IITM	545	107310.5	10.6	1137491.3	5777
Lodhi road, IDM	1783	323792.8	10.6	3432203.68	18899.8
Major D. N. stadium	1792	368614.4	0.334	123117.2096	598.528
Mandir Marg	1820	367822	0.92	338396.24	1674.4
Mundka	1635	431149.5	8.77	3781181.115	14338.95
NSIT Dwarka	1822	407217	0.607	247180.719	1105.954
Najafgarh	1776	336019.2	41.64	13991839.49	73952.64
Narela	1776	423220.8	57.8	24462162.24	102652.8
Nehru nagar	1795	406029	3.5	1421101.5	6282.5
North Campus-DU	1004	225297.6	7	1577083.2	7028
Okhala Phase-2	1794	388759.8	NA	NA	NA
Patparganj	1789	373185.4	8.53	3183271.462	15260.17
Panjabi Bagh	1814	415224.6	7.12	2956399.152	12915.68
Pusa DPCC	1616	335481.6	0.4	134192.64	646.4
Pusa IMD	1683	306979.2	0.4	122791.68	673.2
R.K. Puram	1817	393743.9	9.87	3886252.293	17933.79
Rohini	1787	439244.6	13.14	5771674.044	23481.18
Shadipur	1821	383684.7	3.3	1266159.51	6009.3
Sirifort	1774	406068.6	4.3	1746094.98	7628.2
Sonia Vihar	1785	401625	4.22	1694857.5	7532.7
Sri Aurabindo Marg	1641	296856.9	NA	NA	NA
Vivek Vihar	1789	418626	9.69	4056485.94	17335.41
Wazipur	1788	456476.4	6	2738858.4	10728

The weighted AQI is.

$$\text{Weighted AQI} = \frac{A_i \times \sum_{i=1}^{40} S_i}{n_i \times A_i} = \frac{195639489.7}{848246.7} = 230.6$$

Where, n_i = AQI count

A_i = Area of Station

$\sum S_i$ = Sum of AQI for i^{th} Station

Using this formula, we calculated the weighted AQI for Delhi city to be 230.6. This value is slightly higher than the mean AQI of 39 monitoring stations 221.3, both indicates that the air quality in the city is poor.

Our analysis shows that the weighted AQI provides a more accurate measure of the overall air quality in Delhi city than the individual AQI values for each station. By considering the size of each station's coverage area, the weighted AQI gives a more representative picture of the air quality in the city as a whole. Our analysis of the weighted AQI for Delhi city provides valuable insights into the overall air quality in the city. By considering the size of each station's coverage area, the weighted AQI provides a more accurate measure of the air quality than the individual AQI values for each station. Our findings highlight the need for urgent action to address the air pollution problem in Delhi and protect the health and wellbeing of its residents.

5.4 Correlation Analysis:

The correlation Heatmap shows the correlation coefficients between different air pollutants and air quality index (AQI) in Delhi. The correlation coefficient ranges from -1 to 1, where -1 represents a perfect negative correlation, 0 represents no correlation, and 1 represents a perfect positive correlation. In our project, we employed a powerful visualization technique called a heatmap to represent the correlation between the Air Quality Index (AQI) and various air pollutants including PM2.5, PM10, NO2, NH3, SO2, CO, and Ozone.

The intensity of the colours in the heatmap indicates the degree of correlation between the variables. Strong positive correlations are depicted by vibrant colours, while weak or negative correlations are represented by lighter shades. This visual representation enables us to identify patterns and relationships between the AQI and the different pollutants.

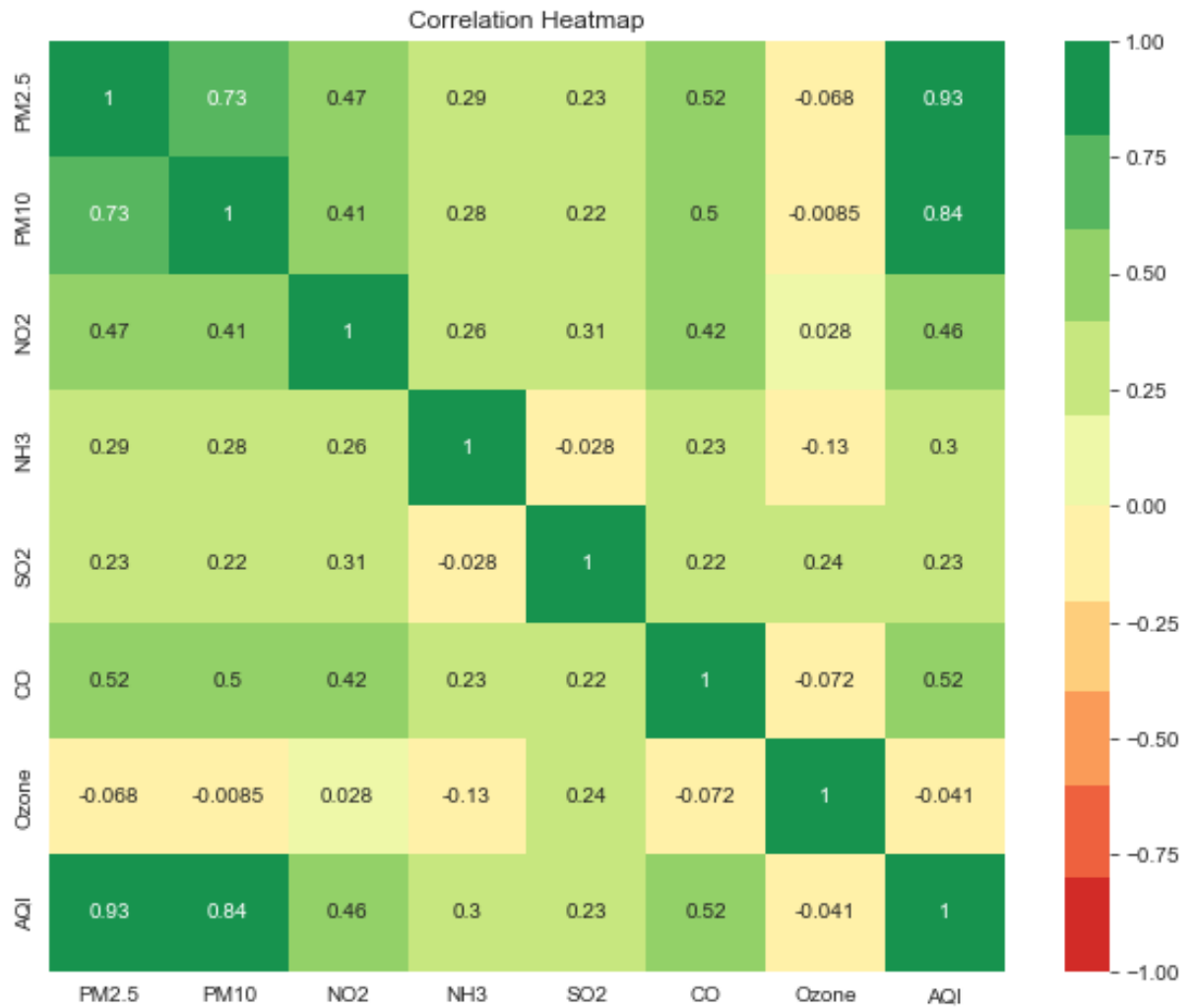


Figure 5.7 Correlation Heatmap

From the correlation heatmap, we can conclude the following:

1. There is a strong positive correlation between PM2.5 and PM10 (0.73), which means that both pollutants have a similar source and are highly related.
2. PM2.5 and AQI are highly correlated (0.93), which indicates that PM2.5 has a significant impact on overall air quality in Delhi.
3. There is a moderate positive correlation between NO2 and PM2.5 (0.47), suggesting that both pollutants have common sources such as vehicular emissions.
4. NH3 shows a weak positive correlation with PM2.5 (0.29), which implies that agricultural activities may contribute to PM2.5 levels to some extent.
5. SO2 and PM2.5 have a weak positive correlation (0.23), indicating that industrial emissions may contribute to PM2.5 levels to some extent.
6. CO has a moderate positive correlation with PM2.5 (0.52), which indicates that both pollutants share similar sources such as vehicular and industrial emissions.
7. Ozone has a weak negative correlation with PM2.5 (-0.07), which indicates that ozone concentrations decrease with an increase in PM2.5 levels.

8. There is a moderate positive correlation between PM10 and AQI (0.84), indicating that PM10 has a significant impact on overall air quality in Delhi.
9. In summary, PM2.5 is the most important pollutant in Delhi, and its sources are vehicular and industrial emissions, along with some contributions from agricultural activities. Therefore, controlling vehicular and industrial emissions is crucial to improve air quality in Delhi.

5.5 Testing of Hypothesis:

In this analysis, we wanted to compare the levels of pollutants between winter and summer seasons. To do this, we first divided the dataset into two groups: one for winter and one for summer. This allowed me to focus specifically on the data from these periods.

5.5.1 Test Method: Shapiro-Wilk test

We performed the Shapiro-Wilk test, which is a commonly used test for checking the normality assumption.

Null Hypothesis

H_0 : The data in the winter/summer group is normally distributed.

Alternative Hypothesis

H_1 : The data in the winter/summer group is not normally distributed.

Summer AQI:

Shapiro-Wilk test statistic: 0.8381

p-value: 0.2090

The summer AQI follows a normal distribution.

Winter AQI:

Shapiro-Wilk test statistic: 0.9818

p-value: 0.7417

The winter AQI follows a normal distribution.

The results of the test indicated that the data in both groups were normally distributed, which was an important prerequisite for the subsequent analysis.

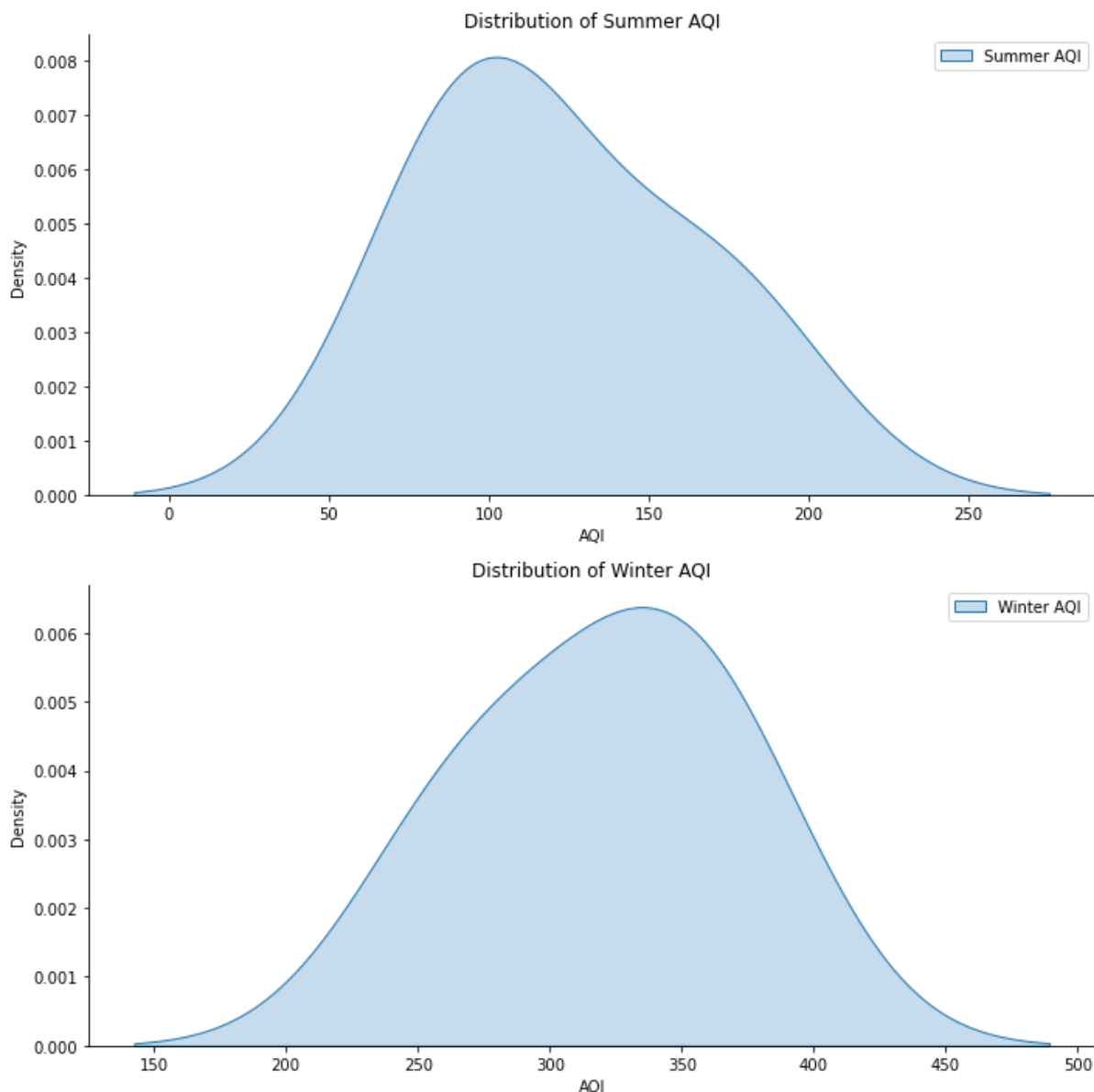


Figure 5.8 Distribution of Summer and Winter AQI

5.5.2 Test Method: Two-sample t-test

With the normality assumption satisfied, then proceeded to perform the two-sample independent t-test. This test is appropriate for comparing the means of two independent groups, in this case, the mean pollutant levels in winter and summer. The t-test allowed us to assess whether there was a significant difference between the average pollutant levels in the two seasons.

Null Hypothesis

H_0 : There is no significant difference between the mean levels of pollutants in winter and summer.

Alternative Hypothesis

H_1 : There is a significant difference between the mean levels of pollutants in winter and summer.

Table 5.8 Outcomes of Two sample t-test

Pollutant	Winter Mean	Summer Mean	t-statistic	Degrees of Freedom	Critical Value (2-tailed)	p-value	Results
PM2.5	311.11	78.46	249.28	32353.00	1.96	0.00E+00	Reject the null hypothesis. There is a significant difference between winter and summer levels. (For all pollutants)
PM10	232.42	108.16	100.11	32353.00	1.96	0.00E+00	
NO2	64.30	33.85	87.06	31891.00	1.96	0.00E+00	
NH3	12.43	8.98	42.02	25463.00	1.96	0.00E+00	
SO2	15.99	12.04	35.69	26427.00	1.96	3.31E-272	
CO	74.59	49.13	82.18	31751.00	1.96	0.00E+00	
Ozone	25.18	30.88	-27.31	31245.00	1.96	4.01E-162	
AQI	321.58	121.20	183.11	32072.00	1.96	0.00E+00	

Based on the results of the t-test, we found that there was a significant difference between the winter and summer levels of pollutants. This means that the pollutant levels varied significantly depending on the season. The rejection of the null hypothesis suggests that the observed differences in pollutant levels are not due to chance alone, but rather indicate a genuine disparity between the seasons.

Conclusions:

1. The results of the tests show that there are significant differences in the levels of pollutants and air quality index (AQI) between winter and summer in Delhi.
2. During winter, the levels of PM2.5, PM10, NO2, NH3, SO2, and CO are higher compared to summer. This means that the air pollution is generally worse in winter than in summer. On the other hand, ozone levels are lower in winter compared to summer.
3. The AQI, which represents the overall air quality, is also significantly higher in winter than in summer. This indicates that the air quality is generally poorer during the winter season.

Reasons behind this result:

1. **Temperature Inversion:** During winter, temperature inversion occurs more frequently, where a layer of warm air traps pollutants close to the ground. This prevents the dispersion of pollutants and leads to higher pollution levels.
2. **Calm Weather Conditions:** Winter is characterized by calmer weather conditions, with lower wind speeds. This lack of wind reduces the dispersion of pollutants, allowing them to accumulate in the air and leading to higher pollution levels.
3. **Increased Biomass Burning:** Winter in Delhi coincides with the crop residue burning season in nearby agricultural regions. The burning of crop residues releases significant amounts of particulate matter (PM) and other pollutants into the air, contributing to higher pollution levels in Delhi.
4. **Increased Industrial and Domestic Activities:** In winter, there is an increased demand for heating in residential and commercial buildings, leading to higher emissions from fossil fuel combustion. Additionally, industrial activities may also contribute to higher pollutant emissions during this season.
5. **Traffic Congestion:** Winter months in Delhi coincide with increased traffic congestion due to weather conditions and festivals. Higher traffic volumes result in increased vehicular emissions, contributing to higher pollution levels.

Regarding the lower ozone levels in winter compared to summer, it can be attributed to the following reasons:

1. **Reduced Photochemical Activity:** Ozone formation is a photochemical process that depends on sunlight and the presence of nitrogen oxides (NO_x) and volatile organic compounds (VOCs). During winter, the reduced sunlight and lower levels of NO_x and VOCs result in lower ozone formation.
2. **Ozone Reaction with Pollutants:** Ozone is reactive and can react with certain pollutants, such as nitrogen dioxide (NO₂) and volatile organic compounds. These reactions can lead to a decrease in ozone levels, especially when pollutant concentrations are high during winter.

5.6 Cluster Analysis: (K-Means Clustering)

By using k-means cluster analysis, we will find patterns in the AQI data and group similar data points together. This will allow us to identify which clusters represent different levels of health impact. Additionally, we will analyze the distribution of these

clusters across various locations to determine which areas experience certain health impacts more frequently.

5.6.1 Scatter plot of clusters:

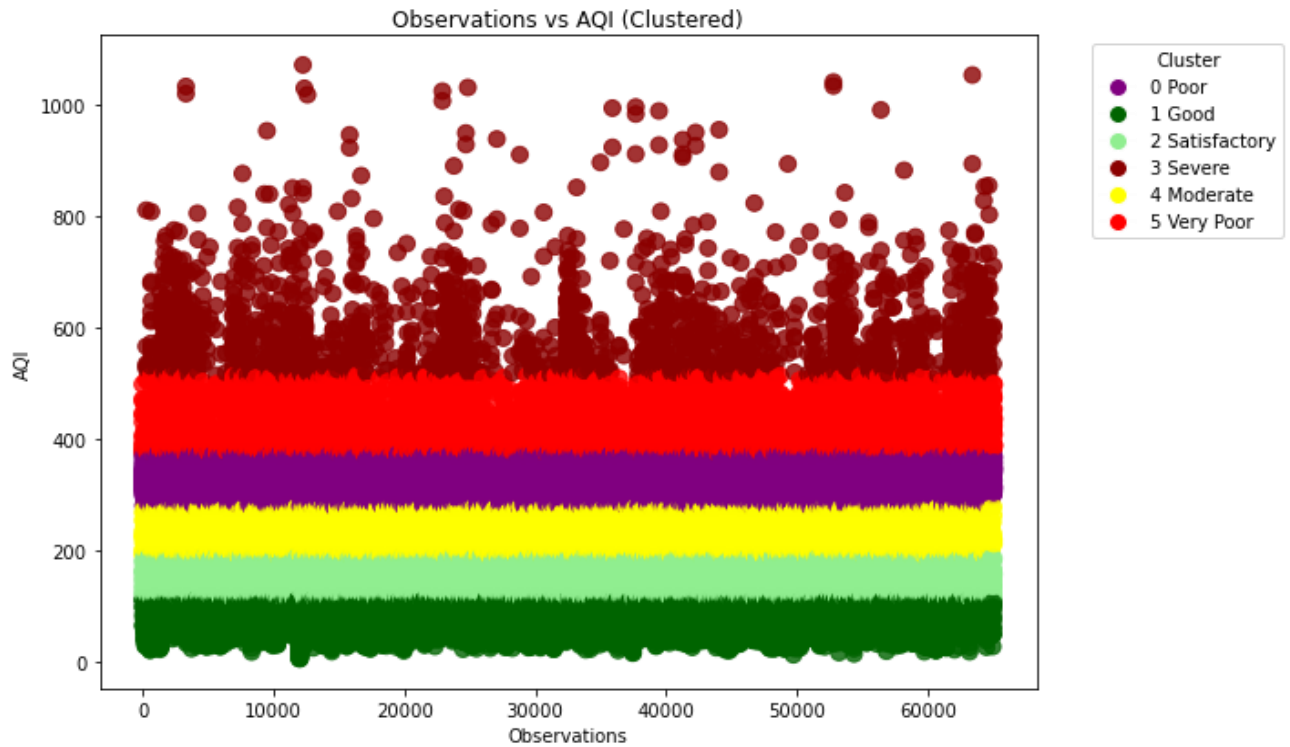


Figure 5.9 Cluster plot of Observations count Vs AQI

By applying K-means clustering to data, we obtained clusters that represent different levels of air quality based on the AQI values. The cluster plot visually demonstrates the grouping of data points into these clusters, allowing us to compare and interpret the air quality characteristics associated with each cluster.

5.6.2 Summary of Cluster Analysis:

Table 5.9 Summary of Cluster analysis

Cluster	Cluster size	Cluster mean	std	min	25%	50%	75%	max
0	14979	325.830	24.09	278.85	307.00	323.49	344.72	374.42
1	18069	80.177	22.12	7.08	63.59	81.53	99.84	114.68
2	14567	148.856	21.98	114.69	129.35	146.80	167.63	190.49
3	1609	607.354	89.98	515.09	542.69	580.69	644.53	1070.58
4	9489	231.677	25.72	190.50	208.67	230.53	253.35	278.80
5	6348	423.487	37.62	374.42	391.80	414.43	448.92	514.95

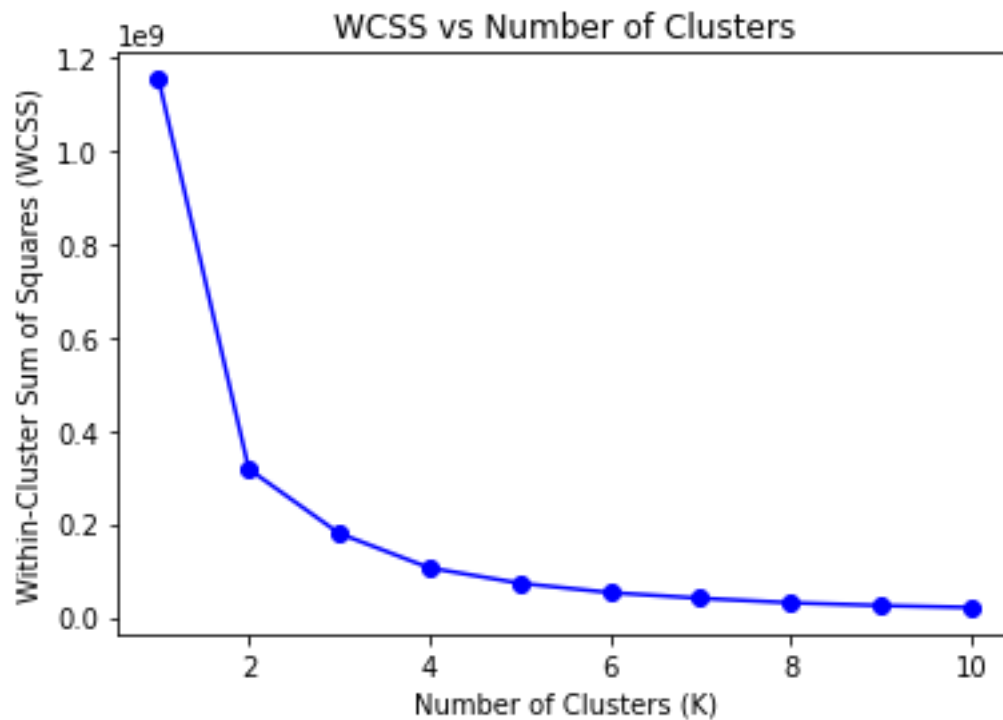


Figure 5.10 Elbow Plot

Within Cluster Sum of Square(WCSS) Values:

```
K=1: WCSS=1153124695.97
K=2: WCSS=317566695.54
K=3: WCSS=179936305.17
K=4: WCSS=105793103.90
K=5: WCSS=72902124.13
K=6: WCSS=52845743.03
K=7: WCSS=40621533.50
K=8: WCSS=31002437.04
K=9: WCSS=25036933.65
K=10: WCSS=20839349.94
```

we can see that the WCSS values decrease significantly as K increases from 1 to 4. After that, the decrease becomes less pronounced. The "elbow" point in the plot could be considered around K=4 or K=5, where the rate of decrease in WCSS starts to level off. Selecting K=4 or K=5 could be a reasonable choice based on the WCSS values.

Interpretation:

- A. Cluster 0:** This cluster has a size of 14,979 data points. The mean AQI value for this cluster is 325.830, indicating air quality falling within the "Very Poor" category (301-400). The standard deviation is 24.09, suggesting moderate variability within this cluster. The minimum AQI is 278.85, and the maximum AQI is 374.42. The majority of AQI values in this cluster fall between 307.00 and 344.72, corresponding to the "Very Poor" AQI category.

- B. Cluster 1:** This cluster has the largest size of 18,069 data points. The mean AQI value is 80.177, representing air quality in the "Satisfactory" category (51-100). The standard deviation is 22.12, indicating relatively low variability within this cluster. The minimum AQI is 7.08, and the maximum AQI is 114.68. The majority of AQI values in this cluster fall between 63.59 and 99.84, which corresponds to the "Satisfactory" AQI category.
- C. Cluster 2:** This cluster has a size of 14,567 data points. The mean AQI value is 148.856, indicating air quality in the "Moderate" category (101-200). The standard deviation is 21.98, suggesting relatively low variability within this cluster. The minimum AQI is 114.69, and the maximum AQI is 190.49. The majority of AQI values in this cluster fall between 129.35 and 167.63, corresponding to the "Moderate" AQI category.
- D. Cluster 3:** This cluster has a size of 1,609 data points. The mean AQI value is 607.354, representing air quality in the "Severe" category (>500). The standard deviation is 89.98, indicating high variability within this cluster. The minimum AQI is 515.09, and the maximum AQI is 1070.58. The majority of AQI values in this cluster fall between 542.69 and 644.53, corresponding to the "Severe" AQI category.
- E. Cluster 4:** This cluster has a size of 9,489 data points. The mean AQI value is 231.677, indicating air quality in the "Poor" category (201-300). The standard deviation is 25.72, suggesting moderate variability within this cluster. The minimum AQI is 190.50, and the maximum AQI is 278.80. The majority of AQI values in this cluster fall between 208.67 and 253.35, corresponding to the "Poor" AQI category.
- F. Cluster 5:** This cluster has a size of 6,348 data points. The mean AQI value is 423.487, representing air quality in the "Very Poor" category (401-500). The standard deviation is 37.62, indicating moderate variability within this cluster. The minimum AQI is 374.42, and the maximum AQI is 514.95. The majority of AQI values in this cluster fall between 391.80 and 448.92, corresponding to the "Very Poor" AQI category.

5.6.3 Visualization:

The clusters are visualized using a stacked bar plot. Each bar represents a location, and the height of the bar corresponds to the count of that location in each cluster. This visualization helps in understanding the distribution of locations across different clusters.

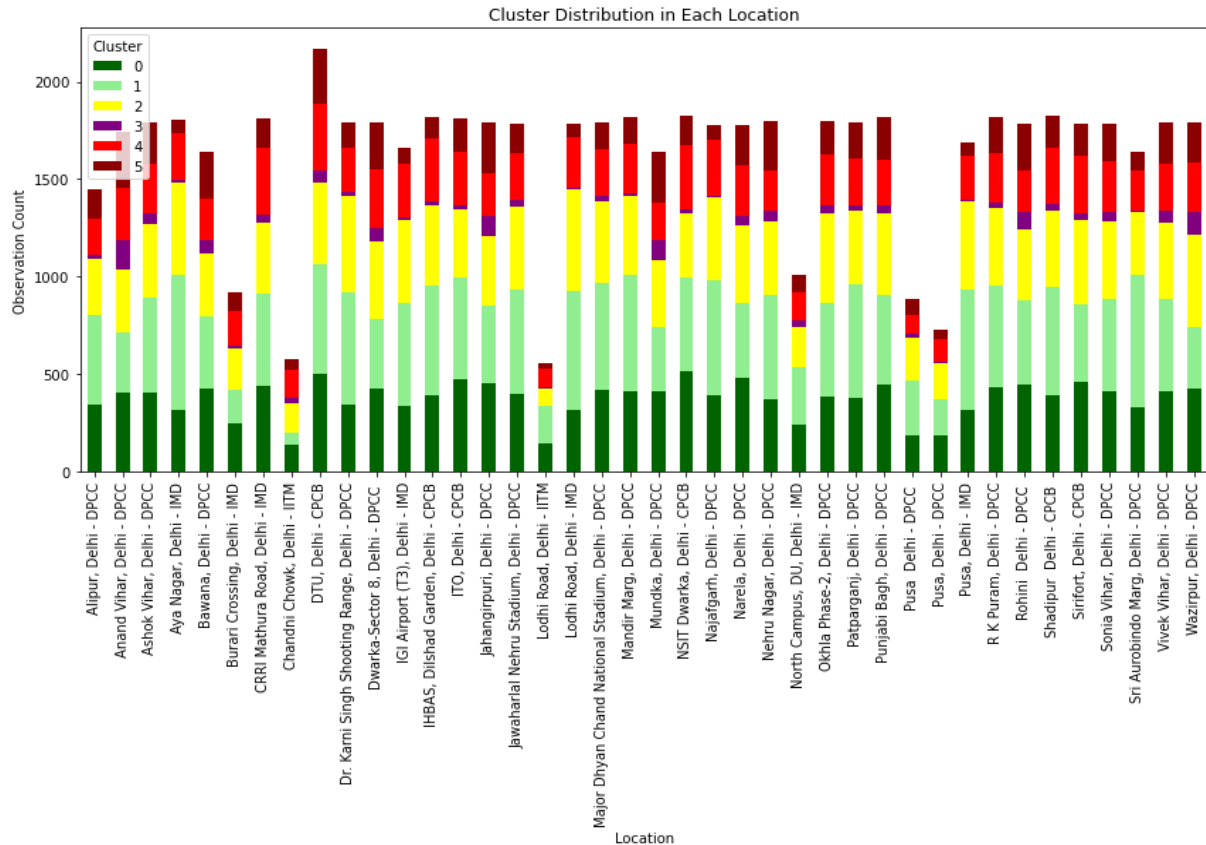


Figure 5.11 Distribution of Clusters according to Locations

5.6.4 Analysis of Proportion of Cluster observations with respect to locations:

In this Analysis, the data points are the different locations, and the similarity is determined by the percentage values in the six clusters (Cluster 0, Cluster 1, Cluster 2, Cluster 3, Cluster 4 and Cluster 5).

The analysis involves calculating the percentage distribution of each location in the four clusters. Based on these percentages, we grouped the Clusters into two distinct Groups: Group 1 and Group 2. Group 1 consists of locations that have a high percentage in Cluster 5 (Severe AQI), while Group 2 consists of locations with a high percentage in Cluster 0 (Good AQI).

Table 5.10 Proportion of Cluster observations with respect to locations

Location	Clusters					
	0	1	2	3	4	5
Alipur	23.81	31.76	20.14	1.38	12.46	10.45
Anand Vihar	23.38	17.59	18.40	8.54	15.64	16.45
Ashok Vihar	22.59	27.16	20.97	3.07	14.28	11.94
Aya nagar	17.69	38.16	26.51	0.61	13.26	3.77
Bawana	26.10	22.32	19.88	4.02	12.99	14.70
Burani Crossing	26.57	18.76	23.21	1.74	19.31	10.41
CRRRI Mathurd road	24.38	25.87	20.08	2.54	18.59	8.55
Chandni Chawk	24.14	10.17	25.52	5.86	24.31	10.00
DTU	23.12	25.93	19.43	2.68	16.01	12.83
Dr. K. S. S. R.	19.07	32.12	27.72	1.06	12.49	7.53
Dwarka Sector-8	23.86	19.84	21.91	3.96	16.83	13.60
IGI Airport	20.43	31.83	25.38	0.90	16.46	5.00
IHBAS Dilshad Garden	21.67	30.86	22.55	1.16	17.66	6.11
ITO	26.28	28.49	19.33	1.16	15.18	9.55
Jahangirpuri	25.38	22.25	19.91	5.69	12.05	14.72
Jawaharlal Nehru Stadium	22.39	29.91	24.13	1.68	13.69	8.19
Lodhi road, IITM	25.63	35.13	15.23	1.43	17.74	4.84
Lodhi road, IDM	17.84	34.10	29.22	0.45	14.64	3.76
Major D. N. stadium	23.44	30.75	23.27	1.45	13.34	7.76
Mandir Marg	22.79	32.73	22.19	0.71	13.78	7.80
Mundka	25.03	20.39	20.76	6.23	11.78	15.81
NSIT Dwarka	28.29	26.15	18.26	1.21	17.87	8.22
Najafgarh	22.10	33.18	23.73	0.34	16.37	4.27
Narela	26.90	21.78	22.45	2.53	14.74	11.59
Nehru nagar	20.67	29.75	21.17	2.90	11.48	14.04
North Campus-DU	24.06	29.32	20.28	3.68	14.31	8.35
Okhala Phase-2	21.34	27.02	25.29	2.28	14.76	9.30
Patparganj	20.95	32.57	21.06	1.73	13.30	10.39
Panjabi Bagh	24.60	25.43	22.73	2.26	13.04	11.94
Pusa DPCC	13.15	32.78	28.32	1.32	15.10	9.32
Pusa IMD	18.70	36.57	26.69	0.65	13.14	4.26
R.K. Puram	23.61	28.94	21.53	1.48	14.11	10.32
Rohini	24.85	24.34	20.48	4.70	12.03	13.60
Shadipur	21.41	30.57	21.51	1.81	15.92	8.78
Sirifort	25.74	22.38	23.95	2.07	16.51	9.35
Sonia Vihar	22.96	26.76	22.12	2.52	14.67	10.97
Sri Aurabindo Marg	20.22	41.23	19.67	0.43	12.61	5.85
Vivek Vihar	23.20	26.10	22.02	3.52	13.30	11.85
Wazipur	23.92	17.44	26.44	6.54	14.25	11.40

Group 1: Locations with high percentage in Cluster 1 (Good air quality as compare to other locations)

- 1) Sri Aurabindo Marg
- 2) Aya Nagar
- 3) Pusa IMD
- 4) Lodhi road
- 5) Najafgarh

Group 2: Locations with high percentage in Cluster 2 (Poor air quality as compare to other locations)

- 1) Anand Vihar
- 2) Wazipur
- 3) Mundka
- 4) Chandni Chawk
- 5) Jahangirpur

5.7 Time Series Analysis:

The analysis of the time series of AQI in Pusa, Delhi provides insights into the air quality trends and pollution patterns in a centrally located area with high population density, contributing to a comprehensive understanding of the city's air pollution scenario.

Analysing the time series data for the Pusa location in Delhi is important due to its central location and high population density (4718 people per km²) , making it representative of air quality conditions in the city. It provides insights into the general air pollution levels and health risks faced by a significant portion of the population. Pusa's urban environment, with its proximity to agricultural activities, adds a unique aspect to the analysis, allowing for the examination of pollution sources such as vehicular emissions, industrial activities, crop burning, and their impact on air quality. The findings from this analysis have policy implications, helping evaluate the effectiveness of pollution control measures and guiding potential interventions or policy changes. Additionally, comparative analysis with other locations provides a broader understanding of localized pollution patterns and the overall air pollution scenario in Delhi.

➤ **Purpose of Time Series Analysis:**

The purpose of time series analysis (TSA) for the AQI data in Pusa, Delhi is to gain insights into the temporal patterns, trends, and fluctuations in air quality over the given period. TSA allows for the examination of how AQI values change over time, identifying any seasonality, trends, or irregularities in the data. The analysis helps in understanding the overall air pollution scenario, evaluating the effectiveness of pollution control measures, and guiding future interventions or policy decisions. It also aids in forecasting future AQI values, assisting in proactive measures for mitigating air pollution and improving public health.

5.7.1 Time Series Plot for Pusa, Delhi over the period Jul-2018 to Dec-2022

The time series plot for Pusa, Delhi from Jul-2018 to Dec-2022 shows the variation of AQI values over time. The plot reveals the fluctuating pattern of air quality, with periods of high and low pollution levels. The plot helps visualize the overall trend and identify any seasonal or long-term patterns in the AQI data.

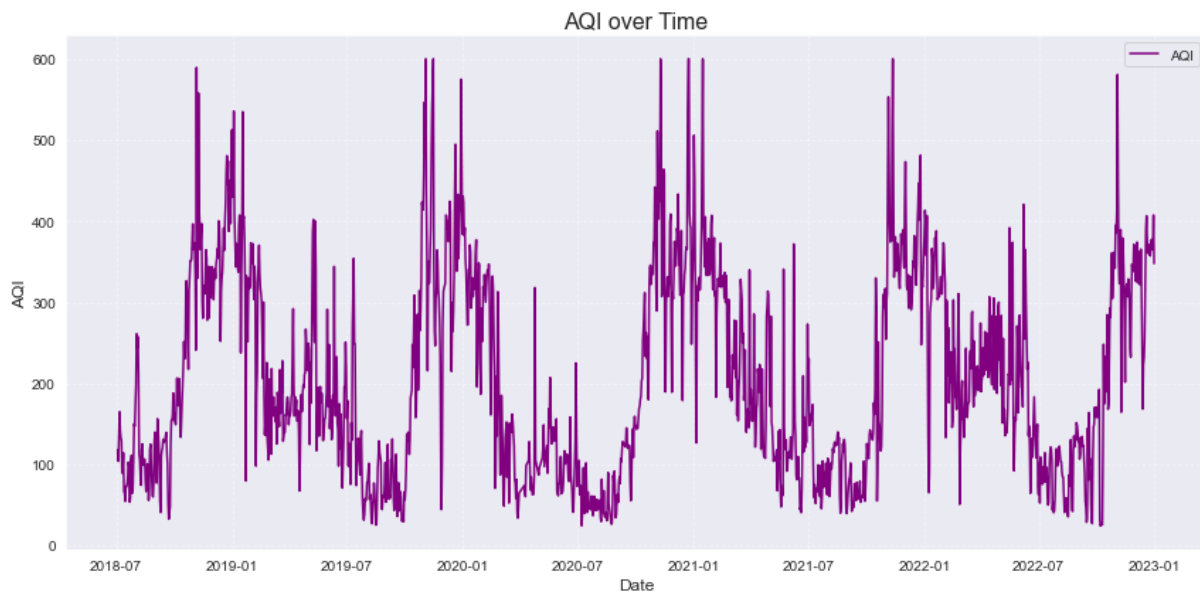


Figure 5.12 Time series plot for AQI(Jul-2018 to Dec-2022)

In addition to analysing the overall time series of AQI values in Pusa, Delhi, it is beneficial to incorporate additional components such as monthly, weekly, or daily averages. By calculating these averages, we can observe underlying patterns and variations that might be obscured in the raw daily data.

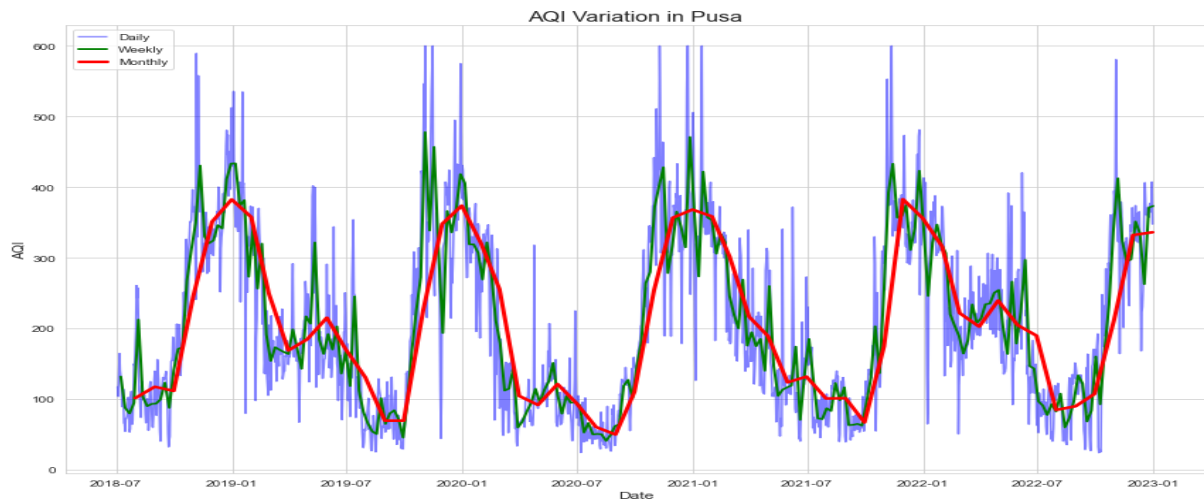


Figure 5.13 Time series plot for AQI(Jul-2018 to Dec-2022)

5.7.2 Classical Decomposition of Time Series:

Classical decomposition plot provides insights into the various components of the time series data. It indicates that there is negligible trend, significant seasonality, and the residuals are relatively close to zero with small variations. This information helps in understanding the underlying patterns and characteristics of the AQI data, which is valuable for developing accurate forecasting models and making informed decisions based on the forecasted values.

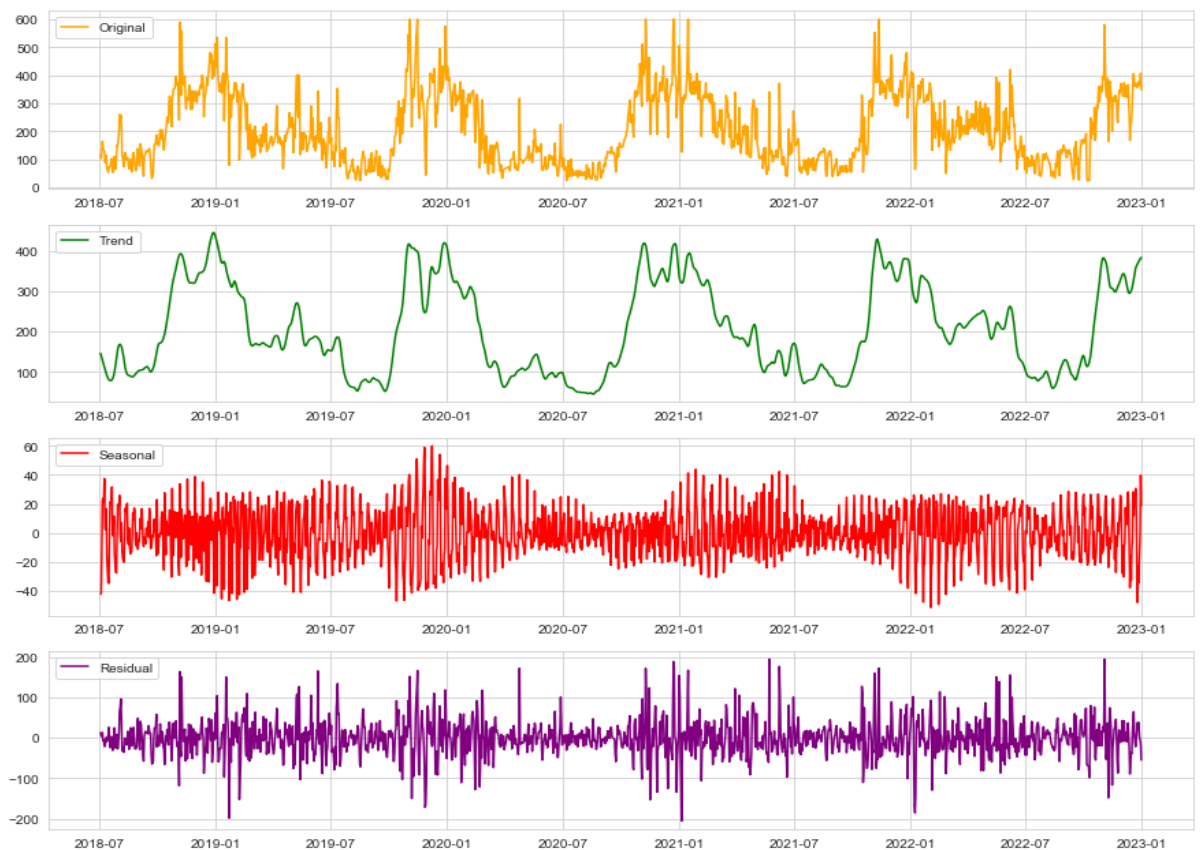


Figure 5.14 Classical Decomposition of Time Series (Jul-2018 to Dec-2022)

➤ From the classical decomposition plot, several observations can be made:

1. The original time series shows the actual values of the AQI over time.
2. The trend plot indicates a negligible trend, suggesting no significant upward or downward movement in AQI values.
3. The seasonal plot reveals predictable patterns in the data, capturing regular variations at fixed intervals.
4. The residuals plot shows small deviations from the predicted values, indicating that the trend and seasonal components explain most of the data's variation.

5.7.3 Autocorrelation and Partial autocorrelation Plot for AQI

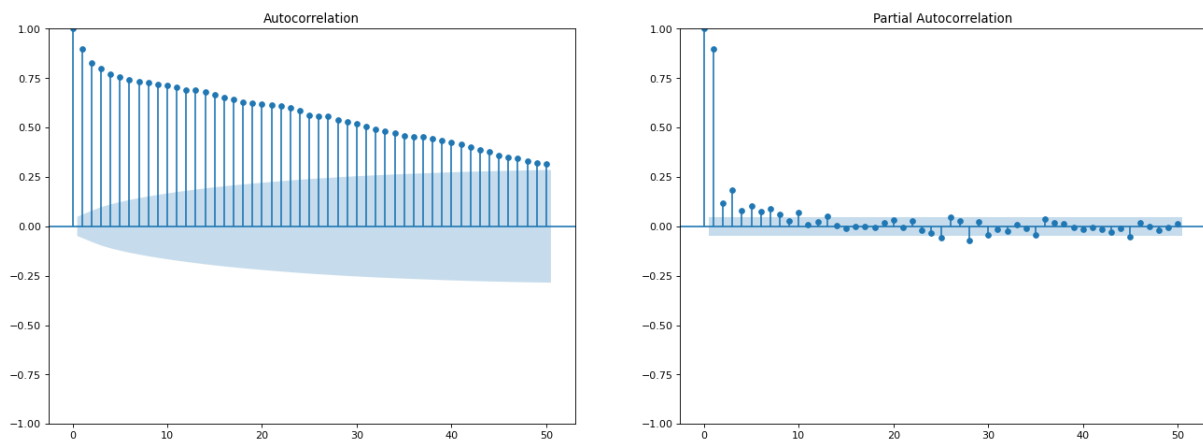


Figure 5.15 ACF and PACF

- For the AR component: The model will be AR(2) (based on the significant spike in the PACF at the second lag)
- For the MA component: The model will be MA(1) (based on the significant spike in the ACF at the first lag)
- Therefore, the specified order for the model would be ARIMA(2,0,1).

5.7.4 Stationarity:

To test the stationarity of a time series, two commonly used tests are the Augmented Dickey-Fuller (ADF) test.

By conducting these tests on a time series, we can determine whether the series is stationary or not, providing insights into the statistical properties and suitability for further analysis and modelling.

Output:

```
ADF Statistic: -3.240
p-value: 0.0178
Critical Values:
  1% : -3.4343
  5% : -2.8633
 10% : -2.568
```

Conclusions:

The ADF statistic is lower than the critical values at the 1%, 5%, and 10% levels, and the p-value is less than the significance level of 0.05. This suggests that the null hypothesis of a unit root (non-stationarity) is rejected, indicating that the data is likely stationary.

5.7.5 Parameter Selection and Model Fitting:

Parameter selection for ARIMA models is an automated process to determine the optimal values of the order parameters (p, d, q) based on statistical techniques. The grid search algorithm evaluates different combinations of parameter values using performance metrics like AIC or BIC. These metrics balance accuracy and model complexity. The Python program output provides the selected parameter values for the ARIMA model.

Output: Performing stepwise search to minimize BIC:

```
ARIMA(0,0,0)(0,0,0)[0] : BIC=22674.440, Time=0.03 sec
ARIMA(0,0,1)(0,0,0)[0] : BIC=20917.242, Time=0.17 sec
ARIMA(0,0,2)(0,0,0)[0] : BIC=19975.380, Time=0.33 sec
ARIMA(0,0,3)(0,0,0)[0] : BIC=19401.759, Time=0.58 sec
ARIMA(0,0,4)(0,0,0)[0] : BIC=19054.029, Time=0.94 sec
ARIMA(0,0,5)(0,0,0)[0] : BIC=18819.781, Time=1.00 sec
ARIMA(1,0,0)(0,0,0)[0] : BIC=17861.419, Time=0.05 sec
ARIMA(1,0,1)(0,0,0)[0] : BIC=17798.720, Time=0.28 sec
ARIMA(1,0,2)(0,0,0)[0] : BIC=17697.092, Time=0.39 sec
ARIMA(1,0,3)(0,0,0)[0] : BIC=17691.003, Time=0.64 sec
ARIMA(1,0,4)(0,0,0)[0] : BIC=17687.577, Time=0.54 sec
ARIMA(2,0,0)(0,0,0)[0] : BIC=17829.689, Time=0.22 sec
ARIMA(2,0,1)(0,0,0)[0] : BIC=17681.616, Time=0.78 sec
ARIMA(2,0,2)(0,0,0)[0] : BIC=17681.939, Time=1.00 sec
ARIMA(2,0,3)(0,0,0)[0] : BIC=17687.659, Time=1.51 sec
ARIMA(3,0,0)(0,0,0)[0] : BIC=17763.856, Time=0.34 sec
ARIMA(3,0,1)(0,0,0)[0] : BIC=17683.041, Time=1.09 sec
ARIMA(3,0,2)(0,0,0)[0] : BIC=17686.336, Time=1.55 sec
ARIMA(4,0,0)(0,0,0)[0] : BIC=inf, Time=0.19 sec
ARIMA(4,0,1)(0,0,0)[0] : BIC=17688.317, Time=1.62 sec
ARIMA(5,0,0)(0,0,0)[0] : BIC=inf, Time=0.25 sec
```

```
Best model: ARIMA(2,0,1)(0,0,0)[0]
Total fit time: 13.501 seconds
```

The best model selected for the ARIMA analysis is ARIMA(2,0,1)(0,0,0)[0]. This model specifies an autoregressive order of 2, a differencing order of 0, and a moving

average order of 1. The seasonal order parameters are set to (0,0,0) with a seasonal period of 0. This model configuration was determined to be the best based on the evaluation of statistical techniques and performance metrics.

The model fitting process involves using the specified parameters (ARIMA(2,0,1)(0,0,0)[0]) to estimate the coefficients and fit the ARIMA model to the data. This includes identifying the appropriate lagged values, differencing the data if required, and determining the moving average terms. The fitted model is then used to generate forecasts and analyse the time series data based on the selected parameters.

SARIMAX Results

```
=====
=====
Dep. Variable:          y      No. Observations:          1643
Model:                SARIMAX(2, 0, 1)      Log Likelihood          -8825.999
Date:                 Sat, 03 Jun 2023      AIC                      17659.999
Time:                  19:50:27             BIC                      17681.616
Sample:               07-03-2018           HQIC                     17668.015
                  - 12-31-2022
Covariance Type:      opg
=====
=====
              coef      std err          z      P>|z|      [0.025      0.975]
-----
ar.L1         1.5516         0.027     58.306      0.000         1.499         1.604
ar.L2        -0.5526         0.026    -20.999      0.000        -0.604        -0.501
ma.L1        -0.8640         0.017   -49.615      0.000        -0.898        -0.830
sigm    2707.5102        52.043     52.025      0.000    2605.508    2809.513
=====
Ljung-Box (L1) (Q):          1.30      Jarque-Bera (JB):          1528.94
Prob(Q):                     0.25      Prob(JB):                   0.00
Heteroskedasticity (H):      0.87      Skew:                       0.16
Prob(H) (two-sided):         0.10      Kurtosis:                   7.72
=====
```

After selecting the optimal parameters (2,0,1) for the time series model, the next step is to fit the model to the data. Model fitting involves estimating the coefficients or parameters of the chosen model using the available time series data.

The stepwise search for minimizing the Bayesian Information Criterion (BIC) resulted in the selection of the best model as ARIMA (2,0,1) (0,0,0) [0] with an intercept term. This model was found to have the lowest BIC value of **17681.616** indicating its superior fit compared to other models evaluated.

Table 5.11 Best fitted model

Model	Coefficients
Autoregressive Model	2
Differencing Model	0
Moving Average Model	1

5.7.6 Create diagnostic plots for residuals

```
results.plot_diagnostics(figsize=(10, 8))
```

```
plt.show()
```

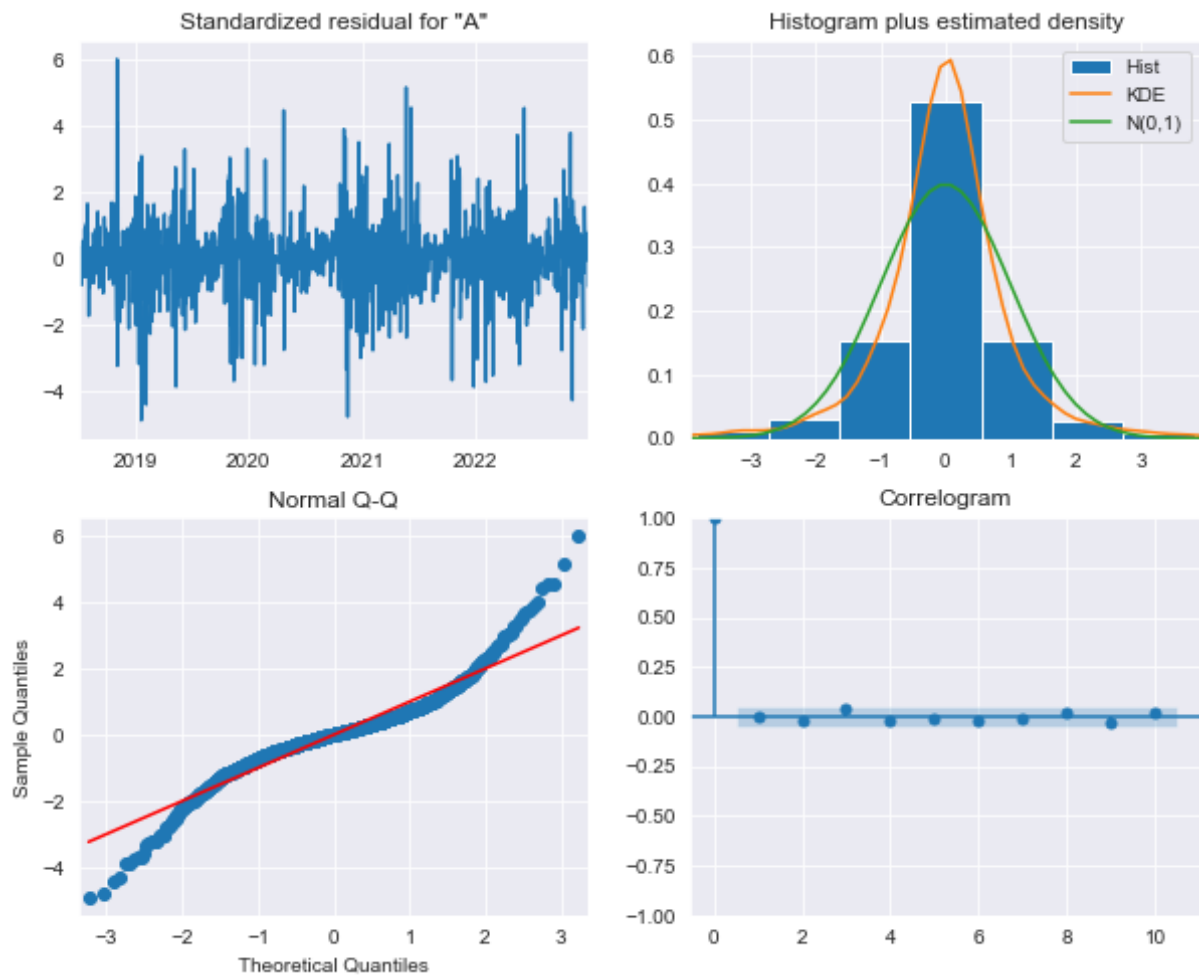


Figure 5.16 Diagnostic plot of Residuals

Conclusion:

- 1) The standardized residual plot show that the residual errors seem to fluctuate around a mean of zero and hence a uniform variance.
- 2) The Histogram plus estimated density plot suggest normal distribution with zero mean.
- 3) Using Normal Q-Q plot, nearly all the dots should fall perfectly in line with the red line.
- 4) The Correlogram, aka, ACF plot shows the residual errors are not autocorrelated. Any autocorrelation would imply that there is some pattern in the residual errors which are not explained in the model.

5.7.7 Forecasting of Time Series:

The visual plot of the actual versus fitted AQI data shows that the forecasted values closely match the actual values. This indicates a good fit of the ARIMA(2,0,1) model to the data.

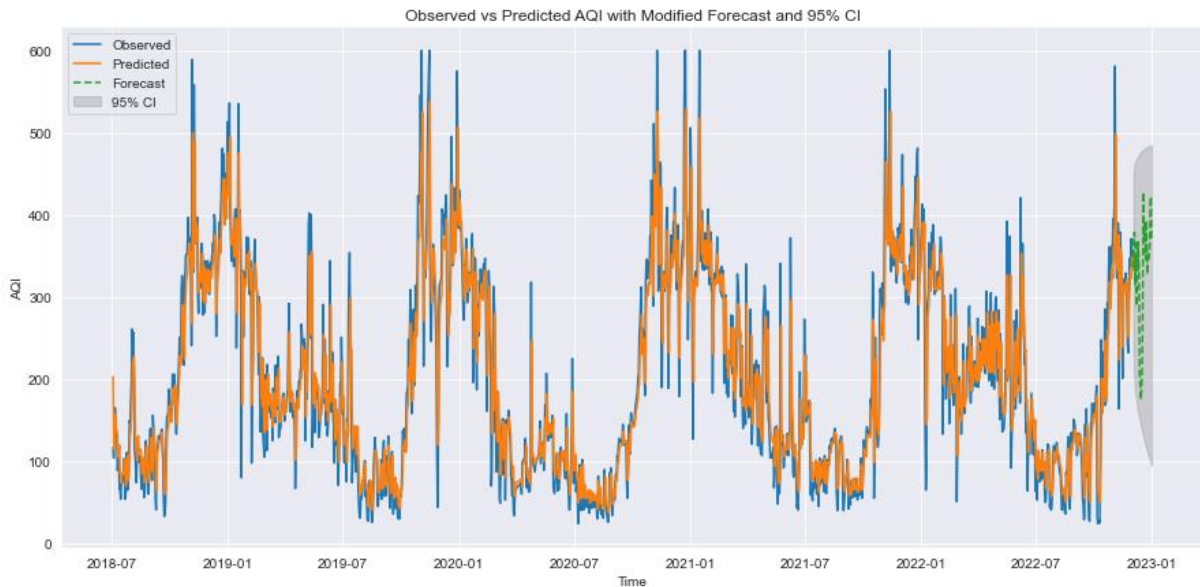


Figure 5.17 Actual AQI vs Predicted AQI(Jul-2018 to Dec-2022)

In the plot, the actual AQI values are represented by data points, typically shown as dots or markers along the y-axis. The fitted AQI values, which are the predictions generated by the ARIMA model using the estimated parameters, are represented by a line or curve that closely follows the actual values.

- **Dataframe of Date, Actual AQI, Forecasted AQI, CI Lower, CI Upper:**

Table 5.12 Actual and Forecasted AQI

NO.	Date	Actual	Forecast	CI Lower	CI Upper
0	02-12-22	332.30	311.24	239.35	443.73
1	03-12-22	331.30	305.96	209.50	456.57
2	04-12-22	369.14	366.68	193.93	461.60
3	05-12-22	338.18	382.10	184.11	464.40
4	06-12-22	356.16	363.43	176.99	466.42
...
25	27-12-22	384.15	358.65	104.35	483.71
26	28-12-22	359.42	357.64	101.87	484.00
27	29-12-22	366.98	401.21	99.45	484.26
28	30-12-22	414.07	456.52	97.08	484.49
29	31-12-22	354.35	402.26	94.76	484.69

5.7.8 Accuracy Metrics for Time Series Forecast:

accuracy measures are important for evaluating the performance of forecasting models like the ARIMA(2,0,1) model for AQI data. They provide a quantitative assessment, enable model comparison, support decision-making, facilitate model improvement, and aid in effective communication of results. Including accuracy measures in the project report ensures a comprehensive and reliable evaluation of the model's forecasting accuracy.

Performance Indices:

Mean Absolute Deviation (MAD) :	25.095
Mean Squared Deviation (MSD) :	841.398
Mean Absolute Percentage Error (MAPE) :	8.078
Average Accuracy:	91.922
Mean Absolute Error (MAE) :	25.0949
Root Mean Squared Error (RMSE) :	29.006

- Based on the calculated accuracy measures for the AQI time series forecast:
 - a. Mean Absolute Deviation (MAD): The average absolute difference between the actual AQI values and the forecasted values is 25.094. This indicates that, on average, the forecasted values deviate from the actual values by approximately 25.09.
 - b. Mean Squared Deviation (MSD): The average squared difference between the actual AQI values and the forecasted values is 841.398. It provides a measure of the overall dispersion of the forecasted values from the actual values.
 - c. Mean Absolute Percentage Error (MAPE): The average percentage difference between the actual AQI values and the forecasted values is 8.078. MAPE is expressed in percentage and represents the magnitude of the forecast errors relative to the actual values. In this case, on average, the forecasted values deviate from the actual values by approximately 8.08%.
 - d. Average Accuracy: The average accuracy of the forecasted AQI values is 91.92%. It represents the percentage of accuracy in predicting the AQI values. Higher accuracy values indicate better performance.
(Average Accuracy of Test Data = $(1 - (\text{MAPE} / 100)) * 100$)
 - e. Mean Absolute Error (MAE): The average absolute difference between the actual AQI values and the forecasted values is 25.095. MAE provides a measure of the average forecast error magnitude.
 - f. Root Mean Squared Error (RMSE): The square root of the average squared difference between the actual AQI values and the forecasted values is 29.007. RMSE provides a measure of the standard deviation of the forecast errors, indicating the overall goodness of fit of the model.

Conclusions

- 1) Jahangirpuri exhibits the highest pollution levels, while Lodhi Road has relatively better air quality. Targeted measures should be implemented to reduce concentrations of PM_{2.5}, PM₁₀, NO₂, and CO in these areas.
- 2) Variations in PM_{2.5} concentration range from 80.74 µg/m³ to 131.57 µg/m³, with Jahangirpuri having the highest and Lodhi Road having the lowest mean concentrations. Standard deviations also vary, with Jahangirpuri having the highest and Lodhi Road having the lowest.
- 3) PM₁₀ concentrations vary across locations in Delhi, with highest maximum values in Delhi Technological University(DTU), Dwarka Sector-8, and Jahangirpuri. Median values above 200 indicate the need for effective air quality management.
- 4) Delhi's monitoring stations show significant variation in AQI levels, with Anand Vihar having the highest mean AQI and Aya Nagar having the lowest. Some locations, such as Mundka, are identified as pollution hotspots, while Lodhi Road and Sri Aurabindo Marg demonstrate comparatively better air quality conditions.
- 5) PM₁₀ and PM_{2.5} dominate Delhi's air pollution, accounting for a significant proportion of the total AQI values. Targeted measures should focus on mitigating the impact of these pollutants on air quality.
- 6) Delhi experiences persistent air pollution exceeding national standards, with a higher percentage of days categorized as "Poor" or worse. Variations in pollution severity exist across locations, and "Good" air quality days are relatively rare. A significant proportion of days in Delhi fall into the "Severe" category or worse, posing a serious health risk.
- 7) The majority of AQI observations indicate moderate to very poor air quality, with a significant proportion falling into the poor category. Instances of extremely poor air quality are a cause for concern.
- 8) The weighted AQI for Delhi city (230.6) is slightly higher than the observed AQI (221.3), indicating the need for improvement. Urgent action is required to address air pollution in Delhi.
- 9) Strong positive correlations exist between PM_{2.5} and PM₁₀, PM_{2.5} and AQI, and moderate positive correlations between NO₂ and PM_{2.5}, CO and PM_{2.5}, and PM₁₀ and AQI, highlighting common sources and the impact of these pollutants on air quality.
- 10) Winter generally exhibits higher mean levels of PM_{2.5}, PM₁₀, NO₂, NH₃, SO₂, CO, and AQI compared to summer, while ozone levels are higher in summer.

- 11) Important locations with relatively good air quality include Sri Aurabindo Marg, Aya Nagar, Pusa, Lodhi Road, Najafgarh. Locations with poor air quality include Anand Vihar, Wazipur, Mundka, Chandni Chawk, Jahangirpur.
- 12) The time series analysis provides valuable insights into forecasted AQI values, with average deviations of 25.09 and an accuracy of approximately 91.92%, indicating reasonably accurate predictions of air quality trends.

APPENDIX

Statistical Analysis codes using Python Software

1) Boxplot

```
import pandas as pd
df=pd.read.csv("C:/python directory/Boxplot.csv")
import matplotlib.pyplot as plt
import seaborn as sns
df['Date'] = pd.to_datetime(df['Date'])
df['Year'] = df['Date'].dt.year
df['Month'] = df['Date'].dt.month
grouped_year = df.groupby(['Year', 'Data Collection
Station'])['AQI'].mean()
df_mean_year = pd.DataFrame(grouped_year).reset_index()
# Group the data by month and station, and calculate the mean AQI for each
group
grouped_month = df.groupby(['Month', 'Data Collection
Station'])['AQI'].mean()
# Convert the resulting group to a dataframe
df_mean_month = pd.DataFrame(grouped_month).reset_index()
sns.set_style('whitegrid')
sns.set(font_scale=1.2)
sns.set_palette('Paired')
# Create a figure with two subplots
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(16, 6))
# Create the year-wise boxplot
sns.boxplot(data=df_mean_year, x='Year', y='AQI', ax=ax1)
ax1.set(xlabel='Year', ylabel='Mean AQI')
ax1.set_title('Mean AQI values for each station by year', fontsize=16)
# Create the month-wise boxplot
sns.boxplot(data=df_mean_month, x='Month', y='AQI', ax=ax2)
ax2.set(xlabel='Month', ylabel='Mean AQI')
ax2.set_title('Mean AQI values for each station by month', fontsize=16)
# Add data labels to the boxplots
for i, box in enumerate(ax1.artists):
    mean = round(df_mean_year[df_mean_year['Year'] ==
int(box.get_x())]['AQI'].values[0], 2)
    ax1.text(box.get_x() + box.get_width() / 2, mean + 1, mean,
ha='center', fontsize=10)
for i, box in enumerate(ax2.artists):
    mean = round(df_mean_month[df_mean_month['Month'] ==
int(box.get_x())]['AQI'].values[0], 2)
    ax2.text(box.get_x() + box.get_width() / 2, mean + 1, mean,
ha='center', fontsize=10)
plt.show()
```

2) Correlation Heatmap

```
import seaborn as sns
import matplotlib.pyplot as plt
my_palette = sns.color_palette("RdYlGn", 10)
sns.set_style("whitegrid")
# Calculate correlation matrix
corr = df.corr()
# Plot heatmap
```

```
plt.figure(figsize=(10,8))
sns.heatmap(corr, annot=True, cmap=my_palette, center=0, vmin=-1, vmax=1)
plt.title('Correlation Heatmap')
plt.show()
```

3) Testing of Hypothesis

```
import pandas as pd
df=pd.read.csv("C:/python directory/AQIDataset.csv")

from scipy.stats import shapiro
# Convert the 'Date' column to datetime format
df['Date'] = pd.to_datetime(df['Date'])
# Extract the month from the 'Date' column
df['Month'] = df['Date'].dt.month
# Calculate the mean AQI for each month
monthly_aqi = df.groupby('Month')['AQI'].mean()
# Perform Shapiro-Wilk test on summer AQI
statistic, p_value = shapiro(summer_aqi)
print("Summer AQI:")
print("Shapiro-Wilk test statistic:", statistic)
print("p-value:", p_value)
if p_value > 0.05:
    print("The summer AQI follows a normal distribution.")
else:
    print("The summer AQI does not follow a normal distribution.")
# Perform Shapiro-Wilk test on winter AQI
statistic, p_value = shapiro(winter_aqi)
print("\nWinter AQI:")
print("Shapiro-Wilk test statistic:", statistic)
print("p-value:", p_value)
if p_value > 0.05:
    print("The winter AQI follows a normal distribution.")
else:
    print("The winter AQI does not follow a normal distribution.")
import seaborn as sns
import matplotlib.pyplot as plt
# Create density plots
fig, axes = plt.subplots(nrows=2, figsize=(10, 10))
# Plot Summer AQI in the first panel
sns.kdeplot(summer_aqi, label='Summer AQI', fill=True, ax=axes[0])
axes[0].set_title('Distribution of Summer AQI')
axes[0].set_xlabel('AQI')
axes[0].set_ylabel('Density')
sns.despine(ax=axes[0])
axes[0].legend()
# Plot Winter AQI in the second panel
sns.kdeplot(winter_aqi, label='Winter AQI', fill=True, ax=axes[1])
axes[1].set_title('Distribution of Winter AQI')
axes[1].set_xlabel('AQI')
axes[1].set_ylabel('Density')
sns.despine(ax=axes[1])
axes[1].legend()
plt.tight_layout()
plt.show()

# t-test
summary_results = {
    'Pollutant': [],
    'Winter Mean': [],
```



```
'Summer Mean': [],
't-statistic': [],
'p-value': [],
'Result': []
}
# Perform hypothesis tests for each pollutant and AQI
pollutants = ['PM2.5', 'PM10', 'NO2', 'NH3', 'SO2', 'CO', 'Ozone', 'AQI']
for pollutant in pollutants:
    winter_values = winter_data[pollutant].dropna()
    summer_values = summer_data[pollutant].dropna()
    # Perform a two-sample independent t-test if data is available for
    both seasons
    if len(winter_values) > 0 and len(summer_values) > 0:
        t_statistic, p_value = ttest_ind(winter_values, summer_values,
equal_var=False)
        # Interpret the results based on the p-value
        alpha = 0.05
        if p_value < alpha:
            result = "Reject the null hypothesis. There is a significant
difference between winter and summer levels."
        else:
            result = "Fail to reject the null hypothesis. There is no
significant difference between winter and summer levels."
    else:
        t_statistic = np.nan
        p_value = np.nan
        result = "Data not available for both seasons."
    # Add the results to the summary dictionary
    summary_results['Pollutant'].append(pollutant)
    summary_results['Winter Mean'].append(np.mean(winter_values))
    summary_results['Summer Mean'].append(np.mean(summer_values))
    summary_results['t-statistic'].append(t_statistic)
    summary_results['p-value'].append(p_value)
    summary_results['Result'].append(result)
summary_df = pd.DataFrame(summary_results)
summary_df
```

4) Cluster analysis (k-means Clustering):

```
import pandas as pd
import pandas as pd
df=pd.read_csv("C:/python directory/AQIDataset.csv")
df
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
df['AQI'] = df['AQI'].interpolate()
# Perform K-means Clustering
X = np.array(df['AQI']).reshape(-1, 1)
kmeans = KMeans(n_clusters=6, random_state=42)
kmeans.fit(X)
df['Cluster'] = kmeans.labels_
cluster_colors = {
    0: 'purple',
    1: 'darkgreen',
    2: 'lightgreen',
    3: 'darkred',
    4: 'yellow',
    5: 'red'
}
```

```
plt.figure(figsize=(10, 6))
scatter = plt.scatter(df.index, df['AQI'],
c=df['Cluster'].map(cluster_colors), s=80, alpha=0.8)
plt.xlabel('Observations')
plt.ylabel('AQI')
plt.title('Observations vs AQI (Clustered)')
cluster_labels = {
    0: '0 Poor',
    1: '1 Good',
    2: '2 Satisfactory',
    3: '3 Severe',
    4: '4 Moderate',
    5: '5 Very Poor'
}
legend_elements = [
    plt.Line2D([0], [0], marker='o', color='w', markerfacecolor=color,
markersize=10)
    for color in cluster_colors.values()
]
plt.legend(legend_elements, cluster_labels.values(), title='Cluster',
bbox_to_anchor=(1.05, 1), loc='upper left')
plt.tight_layout()
plt.show()
cluster_summary = df.groupby('Cluster')['AQI'].describe()
cluster_summary
Stackplot
observation_counts = df.groupby(['Location',
'Cluster']).size().unstack(fill_value=0).reset_index()
observation_counts.columns.name = 'Cluster'
cluster_colors = {
    0: 'darkgreen',
    1: 'lightgreen',
    2: 'yellow',
    3: 'purple',
    4: 'red',
    5: 'darkred'
}
observation_counts.plot(x='Location', kind='bar', stacked=True,
figsize=(12, 9), color=[cluster_colors.get(i) for i in
observation_counts.columns[1:]])
plt.xlabel('Location')
plt.ylabel('Observation Count')
plt.title('Cluster Distribution in Each Location')
plt.legend(title='Cluster')
plt.tight_layout()
plt.show()
```

5) Time Series:

```
import pandas as pd
df25=pd.read_csv("C:/python directory/Pusa.csv")
df25
#TS plot
import matplotlib.pyplot as plt
import seaborn as sns
sns.set_style('darkgrid')
plt.figure(figsize=(12, 6))
plt.plot(df25.index, df25['AQI'], color='purple')
plt.title('AQI over Time', fontsize=16)
plt.xlabel('Date', fontsize=12)
```

```
plt.ylabel('AQI', fontsize=12)
plt.grid(True, linestyle='--', alpha=0.5)
plt.legend(['AQI'], loc='best')
plt.tight_layout()
plt.show()
# Decomposition
from statsmodels.tsa.seasonal import STL
import matplotlib.pyplot as plt
stl = STL(df25['AQI'], seasonal=13, period=12)
res = stl.fit()
fig, axes = plt.subplots(nrows=4, ncols=1, figsize=(15, 12))
axes[0].plot(df25['AQI'], label='Original', color='orange')
axes[0].legend(loc='upper left')
axes[1].plot(res.trend, label='Trend', color='green')
axes[1].legend(loc='upper left')
axes[2].plot(res.seasonal, label='Seasonal', color='red')
axes[2].legend(loc='upper left')
axes[3].plot(res.resid, label='Residual', color='purple')
axes[3].legend(loc='upper left')
plt.show()
# ACF and PACF
fig, axes = plt.subplots(1, 2, figsize=(20,7), dpi= 80)
from statsmodels.graphics.tsaplots import plot_acf
plot_acf(df25['AQI'], lags = 50, ax=axes[0])
# pacf plot AQI
from statsmodels.graphics.tsaplots import plot_pacf
plot_pacf(df25['AQI'], lags = 50, ax=axes[1])
plt.show()
# Stationarity:
from statsmodels.tsa.stattools import adfuller
adf_result = adfuller(df25['AQI'])
adf_statistic = adf_result[0]
adf_p_value = adf_result[1]
adf_critical_values = adf_result[4]
# Model parameter selection:
from pmdarima.arima import auto_arima
max_p = 5 # Maximum value for p
max_d = 2 # Maximum value for d
max_q = 5 # Maximum value for q
# automatic parameter selection and Model Fitting:
model = auto_arima(df['AQI'], start_p=1, start_d=1, start_q=1,
                  max_p=max_p, max_d=max_d, max_q=max_q,
                  seasonal=False, trace=True, suppress_warnings=True,
                  information_criterion='bic', stepwise=False)
print(model.summary())
# Residual plot:
results.plot_diagnostics(figsize=(10, 8))
plt.show()
# Forecasting:
import pandas as pd
import matplotlib.pyplot as plt
from statsmodels.tsa.arima.model import ARIMA
# Set the frequency to daily
df = df.asfreq('D')
# Exclude the last month from the observed data
observed_data = df[:-30]['AQI']
# Fit the ARIMA model
model = ARIMA(observed_data, order=(2, 0, 1))
model_fit = model.fit()
# Generate predictions for the observed period
```

```
predictions = model_fit.predict(start=0, end=len(observed_data)-1)
# Generate forecast for the last month with 95% CI
forecast = model_fit.get_forecast(steps=30)
forecast_values = forecast.predicted_mean
ci = forecast.conf_int()
# Plot observed vs predicted values with forecast and 95% CI
plt.figure(figsize=(12, 6))
plt.plot(df.index[:-30], df['AQI'][:-30], label='Observed')
plt.plot(df.index[:-30], predictions, label='Predicted')
plt.plot(df.index[-30:], forecast_values, label='Forecast')
plt.fill_between(df.index[-30:], ci.iloc[:, 0], ci.iloc[:, 1],
color='gray', alpha=0.3, label='95% CI')
plt.title('Observed vs Predicted AQI with Forecast and 95% CI')
plt.xlabel('Date')
plt.ylabel('AQI')
plt.legend()
plt.show()
```

References:

- Bloomberg Philanthropies. (2021). Assessment of air quality during lockdowns in Delhi.
- National Ambient Air Quality Series. (2012-13). NAAQMS/36/2012-13: Guidelines for the Measurement of Ambient Air Pollutants.
- Kurinji, L. S., Khan, A., & Ganguly, T. (2021). Bending Delhi's Air Pollution Curve: Learnings from 2020 to Improve 2021. Issue brief, June 2021.
- Sahu, J., Sharma, K., & Khanna, A. (2018). Delhi Air Pollution. Reference Note, No. 6/RN/Ref/March/2018. Members Reference Service, LARRDIS, Lok Sabha Secretariat, New Delhi.
- Automotive Research Association of India & The Energy and Resources Institute. (August 2018). Source Apportionment of PM_{2.5} & PM₁₀ of Delhi NCR for Identification of Major Sources. Prepared for the Department of Heavy Industry, Ministry of Heavy Industries and Public Enterprises, New Delhi. Pune: The Automotive Research Association of India. New Delhi: The Energy and Resources Institute.
- Guttikunda, S.K., Dammalapati, S.K., Pradhan, G., Krishna, B., Jethwa, H.T., & Jawahar, P. (2023). What Is Polluting Delhi's Air? A Review from 1990 to 2022. Sustainability, 15, 4209. <https://doi.org/10.3390/su15054209>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning with Applications in R (2nd ed.).
- Brockwell, P. J., & Davis, R. A. (2002). Introduction to Time Series and Forecasting (2nd ed.).
- System of Air Quality and Weather Forecasting and Research (SAFAR) - <https://sifar.tropmet.res.in>
- Delhi Pollution Control Committee. Retrieved from <http://dpcc.delhigovt.nic.in/>
- Central Pollution Control Board. (n.d.). Retrieved from <http://www.cpcb.nic.in/>
- World Health Organization (WHO). Ambient air pollution: A global assessment of exposure and burden of disease.
- Kaggle. Air Quality Index (AQI) Data Analysis
<https://www.kaggle.com/datasets?search=air+quality+index>