Chapter 1: INTRODUCTION

Air pollution is a major problem being faced all nations of the world. Rapid urban and industrial growth has resulted in vast quantities of potentially harmful waste products being released into the atmosphere. As the largest growing industrial nation, India is producing record amount of pollutants $PM_{2.5}$, PM_{10}, 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, CPCB took initiative for developing a national Air Quality Index (AQI) for Indian cities. AQI provides the meaningful information to the people to know what theybreathe and knowing the disadvantages of high AQI.

1.1 Motivation

Being a student of M.Sc. (Statistics) with specialization in Statistics, we were interested in knowing how Air pollutants works to decide Air Quality and Air Quality Index (AQI). As AQI has high then it is quite harmful for us. Also, we got to know more about parameters of Air Quality and analyze the data of various Air pollutants.

1.2 Project Problem

As Mumbai city is one of the most populous city of India and city started facing pollution problems recently, we decided to study different air pollutants responsible at different locations in Mumbai city.

1.3 Objective

The project aims to achieve the following:

- To analyze overall status of air quality through a summation parameter that is easy to understand.
- To classify the levels of air quality for health concern.
- To study Air Quality Index.
- To find hidden patterns in vast quantities of data.
- To forecast values of Air Quality Index (AQI) and Pollutant for different stations.

1.4 About Mumbai City

Mumbai is the most populous city in India, the population of Mumbai city is around 20,961,000. Air pollution is major issue in Mumbai. In 2019, Mumbai recorded 6% of total days with very poor AQI as Compared to 1%-5% in 2017-2018 as per report by SAFAR (system of air quality and weather forecasting)

(Source: Maps-mumbai.com)

The Major contributors to air pollution in Mumbai:

1. Road and construction dust is the major source of pollution in Mumbai. 29% i.e., the highest percentage of PM (Particulate Matter) comes from road and construction dust.
2. Followed by power plants which contribute 20% of the PM levels.
3. As per the reports from an air quality assessment conducted by CPCB and NEERI, the emissions from heavy-duty vehicles running on diesel were also found to be majorly contributing to the pollution
4. Slum rehabilitation projects and private townships and structures.
5. Metro and flyover constructions.

Vehicles Impact:

As per the Maharashtra economic survey, in 2019 Mumbai saw a 9.9% increase in the number of private vehicles registered, around 3.575 million vehicles are in Mumbai and surrounding townships.

In Mumbai city according to CPCB there are 17 different stations. We see the Different stations in the above Map.

Chapter 2: TERMS AND CONCEPTS RELATED WITH AIR POLLUTANTS

2.1 Introduction to (National) Air Quality Index

Air Quality Index (AQI) is a tool for effective communication of air quality status to people in terms, which are easy to understand. AQI is a tool to disseminate information on air quality in qualitative terms (e.g., good, satisfactory, poor) as well as its associated likely health impacts.

An air quality index is defined as an overall scheme that transforms the weighed values of individual air pollution related parameters (for example, pollutant concentrations) into a single number or set of numbers. The result is a set of rules (i.e., set of equations) that translate parameter values into a simpler form by means of numerical manipulation.

1. There are six AQI categories, namely Good, Satisfactory, moderately polluted, Poor, Very Poor, and Severe. Each of these categories is decided based on ambient concentration values of air pollutants and their likely health impacts (known as health breakpoints). Air Quality sub-index and health breakpoints are evolved for eight pollutants (PM10, PM2.5, NO2, SO2, CO, O3, NH3, and Pb) for which short-term (up to 24-hours) National Ambient Air Quality Standards are prescribed.

2. Based on the measured ambient concentrations of a pollutant, sub-index is calculated, which is a linear function of concentration (e.g. the sub-index for PM2.5 will be 51 at concentration 31 μg/m3, 100 at concentration 60 μg/m3 , and 75 at concentration of 45 μg/m3). The worst sub-index determines the overall AQI. AQI categories and health breakpoints for the eight pollutants are as follow:

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

AQI	Color 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 are the foundation that provides a legal framework for air pollution control. An air quality standard is a description of a level of air quality that is adopted by a regulatory authority as enforceable. The basis of development of standards is to provide a rational for protecting public health from adverse effects of air pollutants, to eliminate or reduce exposure to hazardous air pollutants, and to guide national/local authorities for pollution control decisions.

2.2 Parameter related to AQI

The proposed AQI will consider eight pollutants (PM_{10}, $PM_{2.5}$, NO_2, O_3, CO, SO_2, NH_3, 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.

2.2.1 Particulate Matter (PM_{10})

PM_{10} is one of the criteria for Air Quality Index (AQI) calculation. The safe exposure levels for **PM_{10} (24 hour) are 0-100 µg/m³**. India 37 cities have been identified as having the highest pollution levels of PM_{10}. 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 PM_{10} include effects on breathing, respiratory symptoms, decrease in pulmonary function and damage to lung tissue, cancer, and premature death.
- An association between elevated PM_{10} 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/$m^3$of PM_{10} 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 PM_{10} 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/m^3**. Several epidemiological studies (Pope, 1989; Schwartz, 1996) have linked PM_{10}(aerodynamic diameter ≤ 10 μm) and $PM_{2.5}$(aerodynamic diameter ≤2.5μm) 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:

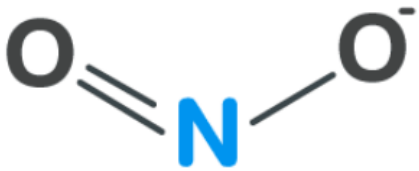
- Premature mortality, chronic respiratory disease, emergency visits and hospital admissions, aggravated asthma, acute respiratory 27 symptoms, and decrease in lung function. $PM_{2.5}$ is of specific concern because it contains a high proportion of

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 (NO_2)

Nitrogen dioxide is a parameter for calculating AQI. As per CPCB 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 (O_3)

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 vapors solvent, chemicals, electronic devices. Nitrogen oxides (NOx) and total Volatile Organic Compounds (tVOCs) 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 CPCB 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 colorless 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 (SO_2)

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 CPCB respectively. Sulphur dioxide is a colorless gas with burnt odor 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 (NH_3) 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 NH_3 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:

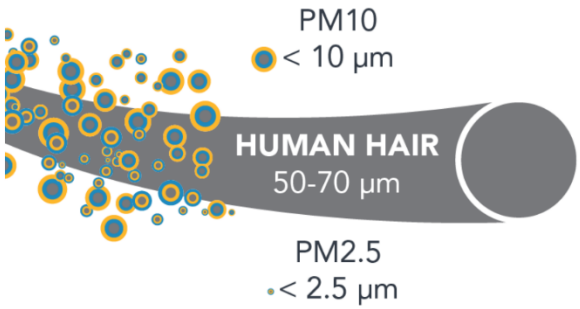
Ammonia and Lead major found in chemical industries, hospitals, and other industrial sectors.

Related Effects:

- Inhalation of high levels of NH3 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.

2.2.8 Unit of Pollutants PM_{10} and $PM_{2.5}$

A mixture of particles with liquid droplets in air forms particulate matter. PM 10 are particles having size of less than or equal to 10 microns whereas PM 2.5 are ultra-fine 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.



2.3 Formulation of AQI

Formulating an AQI: Formation of sub-indices (for each pollutant)

Air quality standards are the foundation that provides a legal framework for air pollution control. An air quality standard is a description of a level of air quality that is adopted by a regulatory authority as enforceable. The basis of development of

standards is to provide a rational for protecting public health from adverse effects of air pollutants, to eliminate or reduce exposure to hazardous air pollutants, and to guide national/local authorities for pollution control decisions.

2.3.1 Structure of an Index

Formation of sub-indices (I1, I2...., In) for n pollutant variables (X1, X2...., Xn) is carried out using subindex functions that are based on air quality standards and health effects. Mathematically,

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

Each sub-index represents a relationship between pollutant concentrations and health effects. The functional relationship between sub-index value (I_i) and pollutant concentrations (X_i) is explained later in the text.

Aggregation of sub-indices, I_i is carried out with some mathematical function (described below) to obtain the overall index (I), referred to as AQI.

$$I = F(I_1, I_2, \dots, I_n \quad (2)$$

The aggregation function usually is a summation or multiplication operation or simply a maximum operator.

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:

$$I = \alpha X + \beta \quad (3)$$

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:

$$I_i = \left[\left\{\frac{(IHI - ILO)}{(BHI - BLO)}\right\} \times (C_P - B_P)\right] + I_{LO} \quad (4)$$

Where,

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

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

I_{HI} =AQI value corresponding to B_{HI}

I_{LO} = AQI value corresponding to B_{LO}

C_p = Pollutant concentration

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

$$AQI = MAX\ (Ii) \quad (5)$$

Where, i= 1,2,…,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: µg/m3 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/m³)	SO_2 24-hr	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	801-1600	1201-1800	3.1-3.5
Severe (401-500)	430+	250+	400+	748+*	34+	1600+	1800+	3.5+

*One hourly monitoring (for mathematical calculation only)

CPCB may consider reviewing the AQI breakpoints every three years after accounting the new research findings on air pollution exposure and health effects.

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

2.3.3 Example

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

$PM_{10} = 202.06\ \mu g/m^3$

$PM_{2.5} = 63.49\ \mu g/m^3$

$NO_2 = 43.61\ \mu g/m^3$

$O_3 = 8.76\ \mu g/m^3$

$CO = 1.81\ mg/m^3$

$SO_2 = 3.14\ \mu g/m^3$

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,

$$I_i = [\{\frac{(I_{HI} - I_{LO})}{(B_{HI} - B_{LO})}\} \times (C_P - B_P)] + I_{LO}$$

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

I) $PM_{10} = 202.06\ \mu g/m^3$

$I_1 = [\{(200 - 101)/(250-101)\} * (202.06-101)] + 101$

$= 168.1472$

II) $PM_{2.5} = 63.49\ \mu g/m^3$

$I_2 = [\{(200 - 101)/(90 -61)\} * (63.49-61)] + 101$

$= 109.500345$

III) $NO_2 = 43.61\ \mu g/m^3$

$I_3 = [\{(100 - 51)/(80 -41)\} * (43.61-41)] + 51$

$= 54.2792$

IV) $O_3 = 8.76\ \mu g/m^3$

$I_4 = [\{(50 - 0)/(50 -0)\} * (8.76-0)] + 0$

$= 8.76$

V) CO = 1.81 mg/m^3

$I_5 = [\{(100 - 51)/(2 - 1.1)\} * (1.81-0)] + 0$

$= 98.54$

VI) $SO_2 = 3.14$ µg/m3

$I_6 = [\{(50 - 0)/(40 - 0)\} * (3.14-0)] + 0$

$= 3.925$

Now we calculate AQI,

$AQI = Max(I_1, I_2, I_3, I_4, I_5, I_6)$

$= 168.14$

Hence, The AQI value is **168.14** (for PM_{10}) of one day (date: 21/11/2020) City – Mumbai (Bandra-Kurla complex).

2.3.4 AQI Calculation Using Spreadsheet XL

AQI for a particular day and at a desired location can be calculated using the MS Excel, wherein a user-friendly evaluation of AQI has been developed. The user needs to input at least three values of pollutant concentration (including at least one of PM10 or PM2.5) in the blue cells and the sub-indices are calculated thus displaying the final AQI along with the colour signifying the AQI category. The health impacts corresponding to the AQI category are detailed in the legend at the bottom of the sheet. This XL program can be obtained from CPCB. Overall AQI is calculated only if data are available for minimum three pollutants out of which one should necessarily be either PM2.5 or PM10. Else, data are considered insufficient for calculating AQI. Similarly, a minimum of 16 hours' data is considered necessary for calculating subindex.

Calculation of AQI

Date			Station	Bandra-Kurla complex
21-Nov-20			City	Mumbai
			State	Maharashtra

Pollutants		concentration in µg/m3 (except for CO)	Sub-Index	check		Air Quality Index
PM10	24-hr avg	202.06	168	1		
PM2.5	24-hr avg	63.49	112	1		
SO2	24-hr avg	3.14	4	1		
NOx	24-hr avg	8.00	10	1	AQI =	168
*CO (mg/m3)	max 8-hr	1.81	91	1		
O3	max 8-hr	8.76	9	1		
NH3	24-hr avg	34.00	9	1		

* 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 1 Spreadsheet for AQI Calculation

2.4 Applications of Air Quality Index

The following six objectives that are served by an AQI:

1.Resource Allocation: To assist administrators in allocating funds and determining priorities. Enable evaluation of trade-offs involved in alternative air pollution control strategies.

2. Ranking of Locations: To assist in comparing air quality conditions at different locations/cities. Thus, pointing out areas and frequencies of potential hazards.

3. Enforcement of Standards: To determine extent to which the legislative standards and existing criteria are being adhered. Also helps in identifying faulty standards and inadequate monitoring programs.

4. Trend Analysis: To determine change in air quality (degradation or improvement) which have occurred over a specified period. This enables forecasting of air quality (i.e., tracking the behavior of pollutants in air) and plan pollution control measures.

5. Public Information: To inform the public about environmental conditions (state of environment). It's useful for people who suffer from illness aggravated or caused by air pollution. Thus, it enables them to modify their daily activities at times when they are informed of high pollution levels.

6. Scientific Research: As a means for reducing a large set of data to a comprehendible form that gives better insight to the researcher while conducting a study of some environmental phenomena. This enables more objective determination of the contribution of individual pollutants and sources to overall air quality. Such tools become more useful when used in conjunction with other sources such as local emission surveys.

Briefly, an AQI is useful for: (i) General public to know air quality in a simplified way. (ii) A decision maker to know the trend of events and to chalk out corrective 3 pollution control strategies. (iii) A government official to study the impact of regulatory actions and (v) a scientist who engages in scientific research using air quality data.

Chapter 3: DATA AND DATA REPRESENTATION

In this chapter we see about dataset of the project. Data is biggest factor in this project. Here we see what kind of data were used and how it looks like? and what are the data sources of the dataset etc. and see the how data is modified or removing of outliers etc.

3.1 Data Source

The **Central Pollution Control Board (CPCB)** www.cpcb.gov.in statutory organization was constituted in September 1974 under the Water (Prevention and Control of Pollution) Act, 1974. Further, CPCB was entrusted with the powers and functions under the Air (Prevention and Control of Pollution) Act, 1981. **Air Quality Monitoring** is an important part of the air quality management. The NATIONAL AIR MONITORING PROGRAMME (NAMP) has been established with objectives to determine the present air quality status and trends and to control and regulate pollution from industries and other source to meet the air quality standards. It also provides background air quality data needed for industrial siting and towns planning.

Central Pollution Control Board (CPCB) provides us a dashboard for the air pollutant concentration data to calculate Air Quality Index (AQI). It provides day wise data of air pollutants.

3.1.1 Data Dashboard:

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

https://app.cpcbccr.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

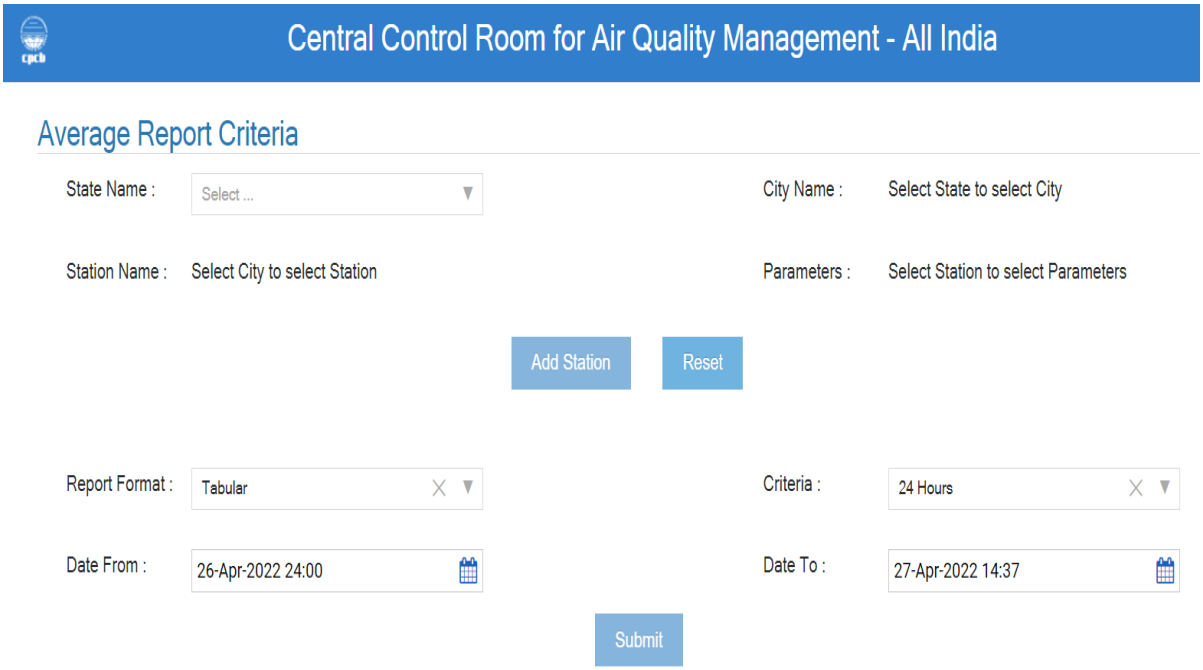


Figure 2: CPCB Dashboard

Here we get the dashboard which includes state name, city name, station name, parameters, from date and to date options. Select the information according to our need and proceed to submit button.

3.2 Data Information

We collect the of air pollutant concentrations of different air pollutants for Mumbai city. In Mumbai there are 17 monitoring stations of air quality measurement which gives the values of air pollutant concentration to calculate Air Quality Index. These are given in following table 3.1.

Table 3.1

Stations	Station Name
S1	Bandra Kurla Complex, Mumbai-IITM
S2	Bandra, Mumbai-MPCB
S3	Borivali East, Mumbai-IITM
S4	Borivali East, Mumbai-MPCB
S5	Chakala – Andheri East, Mumbai-IITM
S6	ChhatrapatiShivajiIntl.Airport(T2),Mumbai-MPCB
S7	Colaba, Mumbai-MPCB
S8	Deonar, Mumbai-IITM
S9	Kandivali East, Mumbai-MPCB
S10	Khindipada-Bhandup West, Mumbai-IITM
S11	Kurla, Mumbai-MPCB
S12	Malad West, Mumbai-IITM
S13	Mazgaon, Mumbai-IITM
S14	Mulund West, Mumbai-MPCB
S15	Powai, Mumbai-MPCB
S16	Siddharth Nagar-Worli, Mumbai-IITM
S17	Vasai West, Mumbai-MPCB

In our data, by removing outliers and missing values there are total 12,737 observations are recorded for 17 stations of Mumbai. These observations are concentration of pollutants and concentration are calculated hourly. To find concentration of pollutant for a day, calculate the average of the concentration for 24 hrs. i.e. the value of concentration of pollutant for a day. CPCB gives us the concentration of pollutant day wise. Using these values of concentrations, we find the Air Quality Index (AQI)

3.3 Data Mining

Our data has collected by CPCB. And it is raw format means there is only concentration values of air pollutants. Which are recorded daily machine or staff of monitoring station. Hence, there are many outliers and missing values present in a data. These outliers may be due to assignable causes. To find these outliers in a data we use some data mining techniques like Exploratory Data Analysis etc. using python software. To detect the outliers from the data we use the graphical tool, Box Plot and descriptive statistics.

In dataset there are many missing values (i.e. None) are present to overcome these problems we use data mining tools. To deal with missing values we replace missing values by median, and if the observed values are too large then we dropped the observation.

3.3.1 Data preprocessing:

It is a technique used in data mining that involves transforming raw data into an understandable format. The data is cleansed through processes such as filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data. As it contains some missing value, the dataset is cleaned, and decimal values are converted into proper float values.

Table 3.2 Air Quality Index Data collected from CPCB

12737	12736	12735	12734	12733	…	5	4	3	2	1	
Mumbai-17	Mumbai-17	Mumbai-17	Mumbai-17	Mumbai-17	…	Mumbai-01	Mumbai-01	Mumbai-01	Mumbai-01	Mumbai-01	**City**
VasaiWest,Mumbai-MPCB	VasaiWest,Mumbai-MPCB	VasaiWest,Mumbai-MPCB	VasaiWest,Mumbai-MPCB	VasaiWest,Mumbai-MPCB	…	BandraKurlaComplex,Mumbai-IITM	BandraKurlaComplex,Mumbai-IITM	BandraKurlaComplex,Mumbai-IITM	BandraKurlaComplex,Mumbai-IITM	BandraKurlaComplex,Mumbai-IITM	**Station**
31-12-202100:00	30-12-202100:00	29-12-202100:00	28-12-202100:00	27-12-202100:00	….	14-11-202000:00	13-11-202000:00	12-11-202000:00	11-11-202000:00	10-11-202000:00	**FromDate**
01-01-202200:00	31-12-202100:00	30-12-202100:00	29-12-202100:00	28-12-202100:00	…	15-11-202000:00	14-11-202000:00	13-11-202000:00	12-11-202000:00	11-11-202000:00	**ToDate**
172.16	183.29	45.57	86.49	76.68	….	27.93	27.24	26.19	90.91	28.85	**PM2.5**
6.24	9.45	127.84	234.8	244.83	…	180.92	243.11	237.63	266.66	289.36	**PM10**
24.06	31.19	5.26	9.14	9.71	…	62.56	86.95	76.37	59.7	70.61	**NO**
9.27	9.06	23.24	30.31	34.67	…	68.73	45.14	92.88	51.01	58.47	**NO2**
20.39	20.34	7.95	9.43	8.92	…	131.72	185.24	157.16	153.5	174.3	**NH3**
1.69	1.9	20.63	20.84	20.31	…	7.24	3.26	8.28	3.24	3.01	**SO2**
		1.56	1.84	1.89	…	1.8	0.76	1.11	0.35	0.2	**CO**
					…	5	5	5	5	5	**Benzene**
					…	19.06	8.11	22.12	8.06	13.72	**Ozone**
					…						**CH4**
					…	323.93	433.62	426.32	417.61	420.08	**CO2**

This is the data for studying air quality index. The data has 15 variables; The following table shows us the variable names, no. of rows and columns etc. In a dataset there are 11 different pollutants in the air namely PM2.5, PM10, NO, NO2, NH3, SO2, CO, Benzene, Ozone, CH4, CO2. And other variables are City, Station Name, FromDate, ToDate. In which city contains the city name along with station number and, FromDate and ToDate contain in which day the values are recorded.

Chapter 4:STATISTICAL TECHNIQUES

In this chapter we see what techniques are used to study the Air Quality Index and Air Pollutants, Information about the statistical techniques below.

4.1 Graphical Representation

Graphical Representation is most important part of the study of the dataset. It shows us meaningful and easily understandable information about the data. The use of Charts and graphs to visualize, analyze and interpret numerical data, functions and other qualitative structures. Using graphs, we conclude many observations about the dataset.

In this project we used different types of graphs like Boxplot, Histogram, Scatter plots, Time series plots, heatmaps, different types of Bar graphs, Line graph and Charts etc. By using these tools helps to give additional information and interpretation about data. By importing some packages like matplotlib, seaborn in python, we visualize our data.

For better graphical visualization we use PowerBI tool. This tool gives us best visualization for the dataset.

4.2 Correlation Analysis

Correlation analysis is used to quantify the degree to which two variables are related. Through the correlation analysis, we evaluate correlation coefficient that tells us how much one variable change when the other one does. Correlation analysis provides us with a linear relationship between two variables.

Correlation Coefficient:A correlation coefficient is a way to put a value to the relationship. Correlation coefficients have a value of between -1 and +1. A “0” means there is no relationship between the variables at all, while -1 or 1 means that there is a perfect negative or positive correlation (negative or positive correlation here refers to the type of graph the relationship will produce).

Pearson Correlation Coefficient: The most common correlation coefficient is the Pearson Correlation Coefficient. It’s used to test for linear relationships between data.

In this dataset we check is there any positive, negative and neutral relationship between the variables in the data.

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

Where,

r = Correlation coefficient

xi = Values of the x-variables in a sample

$\bar{x}$ = mean of the values of the x-variable

yi = Values of the y- variable in a sample

$\bar{y}$ = Mean of the values of the y- variables

- **Multivariate Analysis**

Multivariate analysis (MVA) is based on the principles of multivariate statistics. Multivariate analysis is used to address the situations where multiple measurements are made on each experimental unit and the relations among these measurements and their structures are important. There are many types of multivariate analysis but in this project, we see mainly Principal Component Analysis. Also see the multivariate statistics for the data.

4.3 Principal Component Analysis (PCA)

Principal component analysis was used in order to make possible the visualization of patterns and correlations between the data and hence the identification of possible emission sources. Principal Component Analysis is the type of multivariate analysis. It is the method of analysis which involves finding the linear combination of a set of variables that has maximum variance and removing its effect, repeating this successively.

Now we see how PCA worked on our data, The purpose of principal component analysis is to find the best low-dimensional representation of the variation in a multivariate data set. For example, in the case of the AQI data set, we have 11 different air pollutants in which CO and CH4 has most missing values therefore we take here 9 air pollutants except these two. Now we have 8 air pollutants describing concentration of the pollutant from seventeen different monitoring stations. We can carry out a principal component analysis to investigate whether we can capture most of the variation between the concentration using a smaller number of new variables (principal

components), where each of these new variables is a linear combination of all or some of the 9 air pollutants.

Steps to perform PCA in Python

1. To carry out a principal component analysis (PCA) on a multivariate data set, the first step is often to standardize the variables under study using the scale() function in the Python. This is necessary if the input variables have very different variances. Mathematically, this can be done by subtracting the mean and dividing by the standard deviation for each value of each variable.

$$z = \frac{value - mean}{standard\ deviation}$$

2. Once you have standardized your variables, you can carry out a principal component analysis using the PCA class from 'sklearn.decomposition' package and its fit method, which fits the model with the data X. The default solver is Singular Value Decomposition ("svd"). For more information you can type help(PCA) in the python console. (we see the python code for this in Appendix)
3. In order to decide how many principal components should be retained, it is common to summarize the results of a principal components analysis by making a scree plot, which we can do using the 'screeplot()' function. Here 'screeplot()' is main in PCA which tells us how many component are retained.
4. The loadings for the principal components are stored in a named element 'components_' of the variable returned by 'PCA().fit().' This contains a matrix with the loadings of each principal component, where the first column in the matrix contains the loadings for the first principal component, the second column contains the loadings for the second principal component, and so on.
5. To calculate the values of the first principal component, we can define our own function to calculate a principal component given the loadings and the input variables values.
6. We can then use the function to calculate the values of the first principal component for each sample in our wine data. Similarly, we can calculate second principal component.
7. The values of the principal components can be computed by the 'transform()' (or fit_transform()) method of the PCA class. It returns a matrix

with the principal components, where the first column in the matrix contains the first principal component, the second column the second component, and so on.

The scatterplot shows the first principal component on the x-axis, and the second principal component on the y-axis.

4.4 Multiple linear Regression

MLR is a supervised learning model that attempts to establish the relationship between two or more explanatory variables and a response variable by fitting a linear equation of the data. The MLR block diagram is shown in Figure 4.1. It is an extension to an ordinary least square regression that involves over one explanatory variable. The regression line for*p*explanatory variable x_1, x_2……. x_p is defined as μ. This regression line mentions the mean response of μ_y, which is going to change with explanatory variables. The observed values of y differed about their means μ_y and were assumed to have the standard deviation σ. The formula for MLR is given in equation.

$$y_i = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \cdots + \alpha_p x_{ip} + \varepsilon$$

Where,

i= nth reading
x_i =Explanatory variable
y_i =Dependent variable
α_0 =y-intercept
α_p =Slope value of each explanatory variable
ε= error term

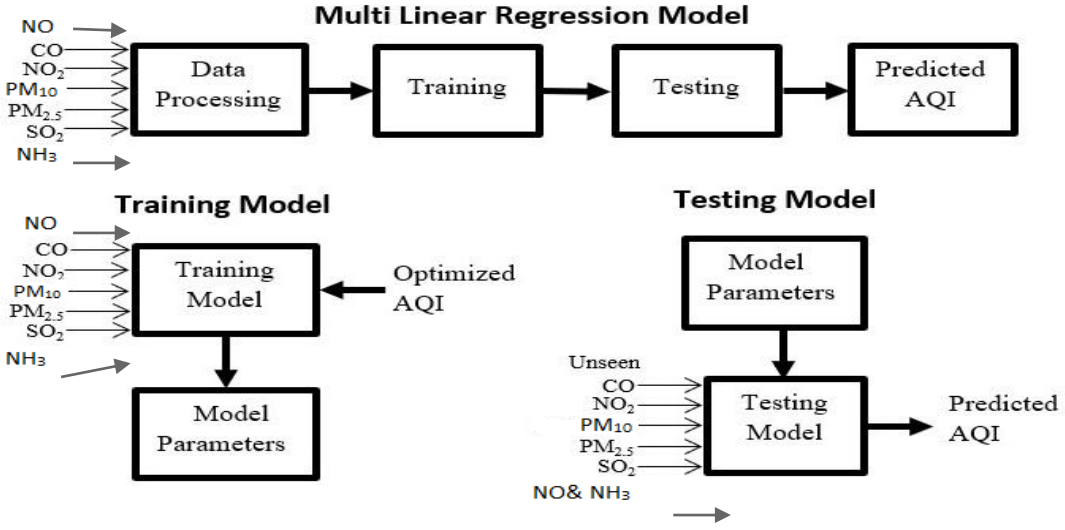


Figure 4.1 Block Diagram of Multi-Linear Regression

4.5 Time Series Analysis and Forecasting

Time series forecasting is the process of analyzing time series data using statistics and modeling to make predictions and inform strategic decision-making. Time series forecasting occurs when you make scientific predictions based on historical time stamped data. While forecasting and "prediction" generally mean the same thing, there is a notable distinction. In some industries, forecasting might refer to data at a specific future point in time, while prediction refers to future data in general. Series forecasting is often used in conjunction with time series analysis.

Time series analysis involves understanding various aspects about the inherent nature of the series so that you are better informed to create meaningful and accurate forecasts.

COMPONENTS OF TIME SERIES:

A time series is a series of data points indexed (or listed or graphed) in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time. Thus, it is a sequence of discrete-time data. Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values.

The various forces of work, affecting the values of a phenomenon in a time series, can be broadly classified into the following four categories, commonly known as the components of a time series, someor all of which are present (in a given time series) in varying degrees.

Stationarity and differencing

A stationary time series is one whose properties do not depend on the time at which the series is observed. Thus, time series with trends, or with seasonality, are not stationary, the trend and seasonality will affect the value of the time series at different times. On the other hand, a white noise series is stationary it does not matter when you observe it, it should look much the same at any point in time.

Some cases can be confusing a time series with cyclic behavior (but with no trend or seasonality) is stationary. This is because the cycles are not of a fixed length, so before we observe the series we cannot be sure where the peaks and troughs of the cycles will be.

In general, a stationary time series will have no predictable patterns in the long-term. Time plots will show the series to be roughly horizontal (although some cyclic behavior is possible), with constant variance.

To check stationarity

To check the stationarity of the data the following are most useful statistical tests.

1. ADF test (Dickey Fuller Test)

Augmented Dickey Fuller test (ADF Test) is a common statistical test used to test whether a given Time series is stationary or not. It is one of the most commonly used statistical test when it comes to analyzing the stationary of a series.

Where the null hypothesis is the time series possesses a unit root and is non-stationary. So, the P-value in ADF test is less than the significance level, you reject the null hypothesis.

2. KPSS test

KPSS test is a statistical test to check for stationarity of a series around a deterministic trend. Like ADF test, the KPSS test is also commonly used to analyze the stationarity of a series. However, it has couple of key differences compared to the ADF test in function and in practical usage.

The KPSS test, short for, Kwiatkowski-Phillips-Schmidt-Shin (KPSS), is a type of Unit root test that tests for the stationarity of a given series around a deterministic trend.In other words, the test is somewhat similar in spirit with the ADF test.

On the other hand, KPSS is used to test for trend stationary. The null hypothesis and the p-value interpretation is just the opposite of ADF test.

ARIMA

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 non-negative integers

Where,

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

d=is the degree of differencing(I) Model

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

Model:

For p, d, q ≥ 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

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

Time-series analysis-ARIMA is used to forecast the AQI. ARIMA model is the combination of three different individual models known as the AutoRegressive (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.

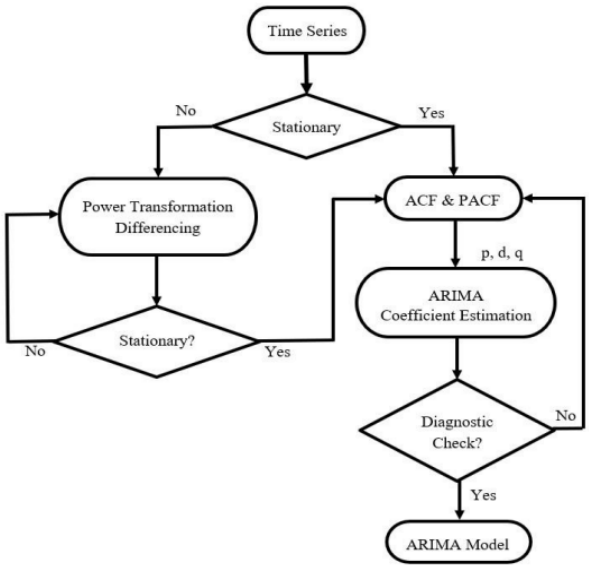


Fig. 4.2 Block Diagram of ARIMA Model

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.

Building ARIMA Model:

1. Plot the time series. Look for trends, seasonal components, step changes, outliers.
2. Nonlinearly transform data, if necessary
3. Identify preliminary values of d, p, andq.
4. Estimate parameters.
5. Use diagnostics to confirm residuals are white/iid/normal.
6. Model selection.

4.5.1Performance Indices: The statistical criteria such as MAPE, RMSE, and MAE are used to evaluate each developed model's performance measure.

Mean Absolute Percentage Error (MAPE)

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

$$\mathbf{MAPE} = \frac{\sum_{i=1}^{n} \left| \frac{(xi - \widehat{xi})}{xi} \right|}{n} \times 100 \, , (xi \neq 0)$$

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

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathbf{xi} - \hat{\mathbf{x}}\mathbf{i})^2}$$

Mean Absolute Error (MAE)

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

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|xi - \hat{\mathbf{x}}\mathbf{i}|$$

Chapter 5: DATA ANALYSIS AND CONCLUSIONS

In this chapter we see the graphical representation of data, data analysis and conclusions using statistical techniques.

5.1 Graphical representation:

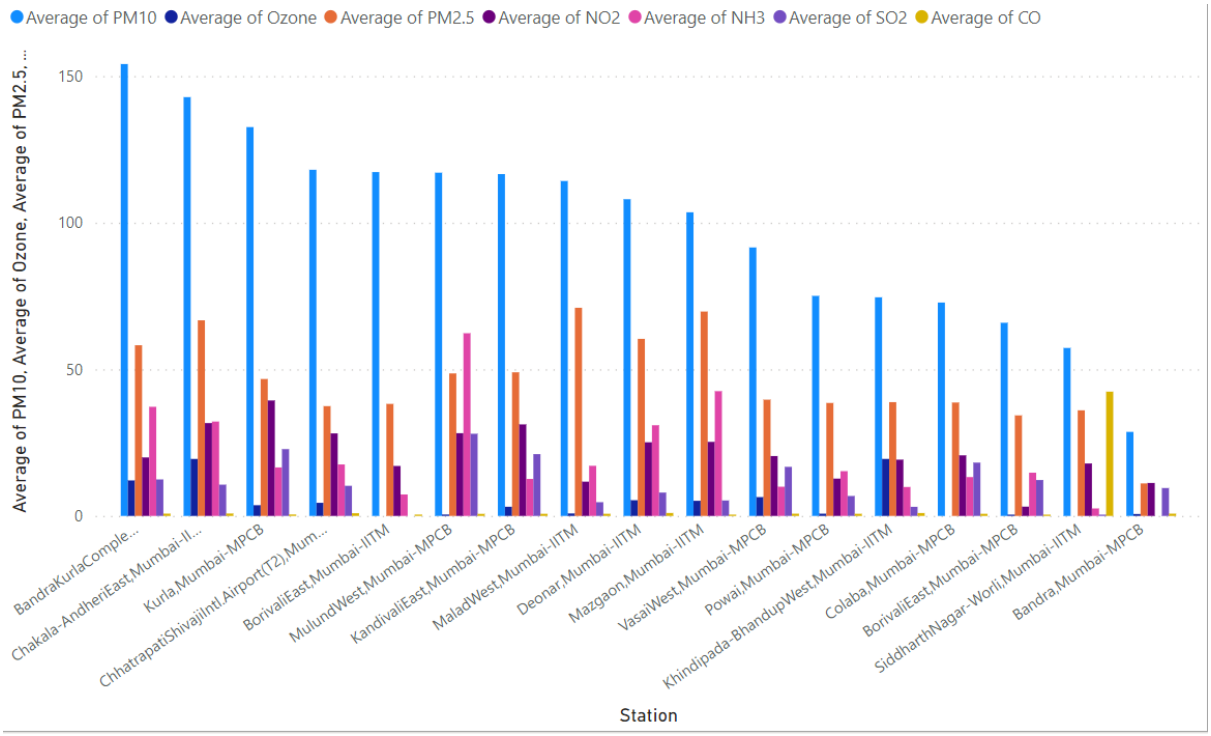


Figure 5.1 Average Air Pollutants Concentration for 17 stations

Conclusion:

1. In above graph we observed that Bandra Kurla Complex is high in pollutant PM_{10} as compared to other stations. We conclude that the PM_{10} is high for all 17 stations.
2. The concentration of $PM_{2.5}$ is high for Chakala-Andheri East Mumbai, Malad West Mumbai-IITM and Mazgaon-Mumbai-IITM as compared to other stations.
3. The concentration of NH_3 is high for Mulund West-Mumbai-MPCB compared to other stations.

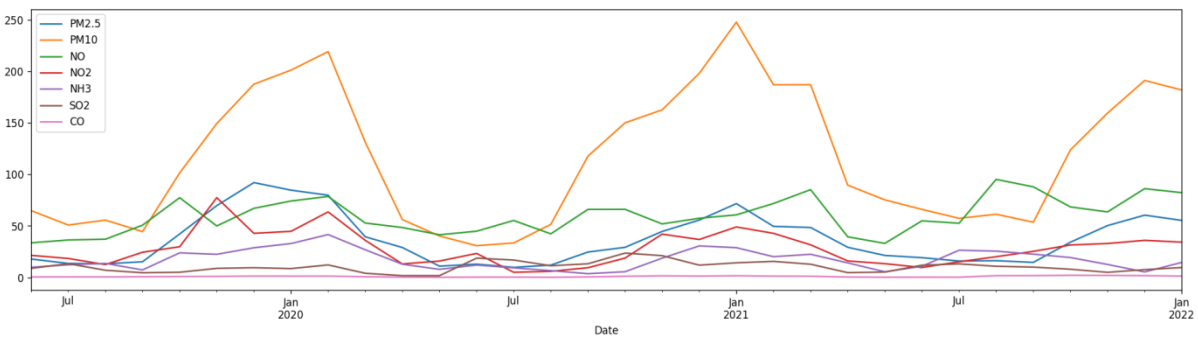


Figure 5.2 Air Pollutants Concentration form (July 2019- Jan 2022

Conclusion:

1. The concentration of PM_{10} high than other pollutant for Chhatrapati Shivaji Int. Airport(T2), Mumbai .
2. NO is continuously increasing

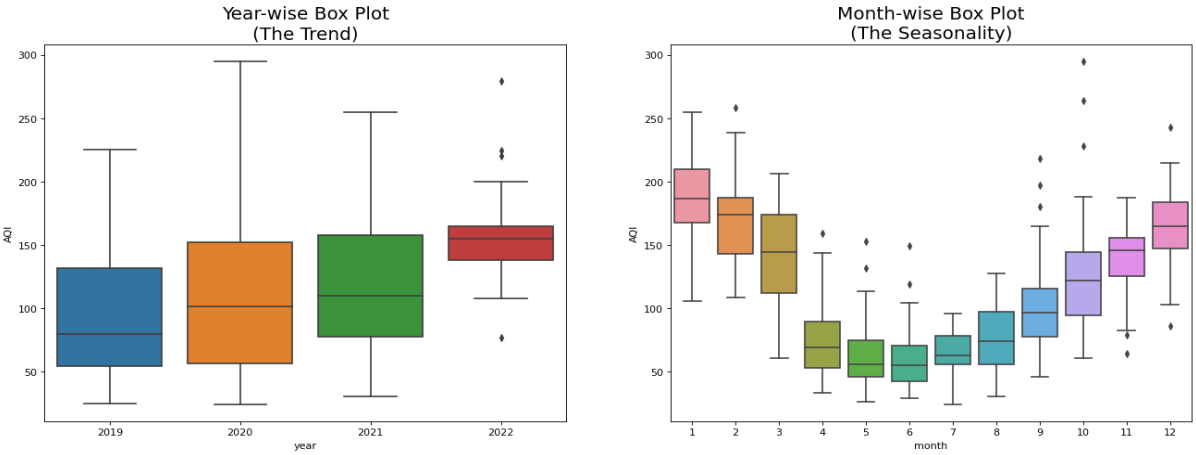


Figure 5.3 Yearly and Monthly box plot for AQI

As we observe that the AQI mean concentration is high in 2022, the 2019 is due to covid-19 low and is increasing as year increases. Monthly box plot is indicated that the Jan to Apr the AQI is low due to summer and after the summer the AQI is as stable due to rainy season and after rainy season AQI is increase.

5.2 Descriptive statistics among the monitoring stations and pollutant:

Table 5.1 Descriptive statistics for $PM_{2.5}$

Station	count	Mean	Std	min	25%	50%	75%	Max
S1	3619	10.99	19.66	0	0	0	16.87	408.07
S2	401	57.94	45.61	1.62	24.36	38.73	76.36	208.81
S3	414	37.98	29.38	2.14	15.22	30.19	52.14	154.19
S4	891	33.89	30.05	1.36	10.64	22.87	53.54	199.68
S5	361	66.71	46.41	5.28	25.44	58.32	98.08	368.03
S6	927	36.64	27.76	4.29	15.38	24.33	53.07	155.28
S7	895	37.84	28.76	2.87	14.63	26.51	54.87	271.56
S8	405	60.20	39.83	2.99	21.52	57.59	92.76	153.74
S9	373	48.98	36.15	7.14	18.76	34.95	73.93	176.34
S10	363	38.80	24.27	1.17	18.71	35.63	56.62	206.24
S11	916	45.62	35.87	0.79	17.83	28.29	74.64	198.7
S12	283	70.81	50.91	1.06	24.35	60.29	107.6	219.2
S13	413	68.79	51.86	4.1	22.87	58.32	106.92	296.93
S14	356	47.08	29.82	5.49	21.99	38.11	70.05	134.64
S15	933	38.142	28.84	0.07	15.3	26.19	59.9	150.15
S16	339	36.01	24.31	3.66	15.68	28.17	55.88	129.29
S17	847	38.63	29.60	0.79	12.95	27.2	63.80	138.38

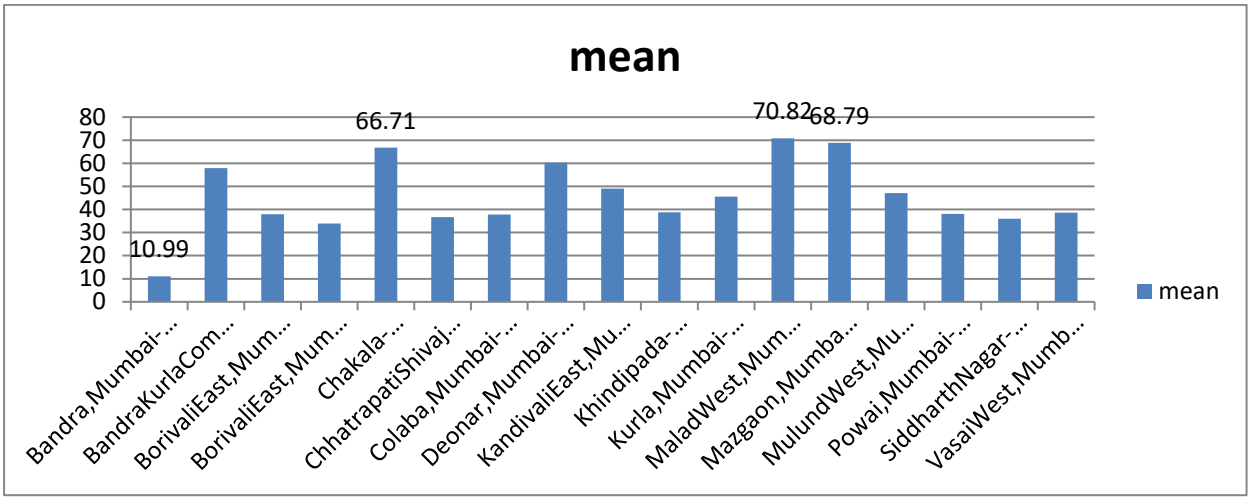


Figure 5.4Mean of pollutant $PM_{2.5}$

Conclusions:

1) Using descriptive statistics we detect the outlier like the S1 (Bandra,Mumbai-MPCB) has 408.07 is outlier because of the 75th percentile of the data is 16.87, mean of data is 10.99 that why we detect it as outlier using the box plot we also detect outlier. Similarly we detect outlier and find the reason behind it.

2) The average concentration of $PM_{2.5}$ is high in stations S2, S5, S12 and S13 (BandraKurlaComplexMumbai-IITM, Chakala Andheri East Mumbai-IITM, Malad West Mumbai-IITM and Mazgaon Mumbai-IITM table 3.1) compare to other station.

3) Low concentration of $PM_{2.5}$is in Bandra,Mumbai-MPCB station 10.99208. That means the bandra is very low concentration of $PM_{2.5}$ we tag it as good station.

Table 5.2 Descriptive statistics for PM_{10}

Station	count	Mean	std	min	25%	50%	75%	Max
S1	3619	30.23	48.41	0	0	0	62.73	446.69
S2	401	153.43	86.77	7.16	73.14	150.7	222.4	371
S3	414	115.8	60.02	24.26	62.73	108.5	162.2	298.75
S4	891	65.70	39.68	4.5	32.52	61.55	92.9	233.44
S5	361	142.8	96.83	11.31	52.14	128.9	211.2	470.84
S6	927	116.3	74.21	10.15	54.96	91.22	175.2	495.21
S7	895	72.48	50.65	11.97	31.99	59.59	100.1	540.44
S8	405	107.7	68.13	5.77	40.76	105.1	166.4	255.74
S9	373	116.0	61.86	24.83	62.6	112.1	168.8	271.86
S10	363	74.55	45.77	1.91	36.01	70.79	110.0	395.01
S11	916	131.6	74.89	10.59	66.90	116.0	183.5	403.49
S12	283	113.4	68.64	11.25	55.40	103.3	157.0	317.18
S13	413	102.7	71.67	9.78	41.77	91.81	156.3	329.6
S14	356	114.1	62.75	21.59	62.73	100.8	165.5	288.31
S15	933	74.71	43.00	7.83	38.22	64.07	106.3	212.23
S16	339	57.28	40.07	5.17	23.7	44.35	90.2	234.86
S17	847	78.59	51.26	4.95	61.83	62.73	85.41	350.52

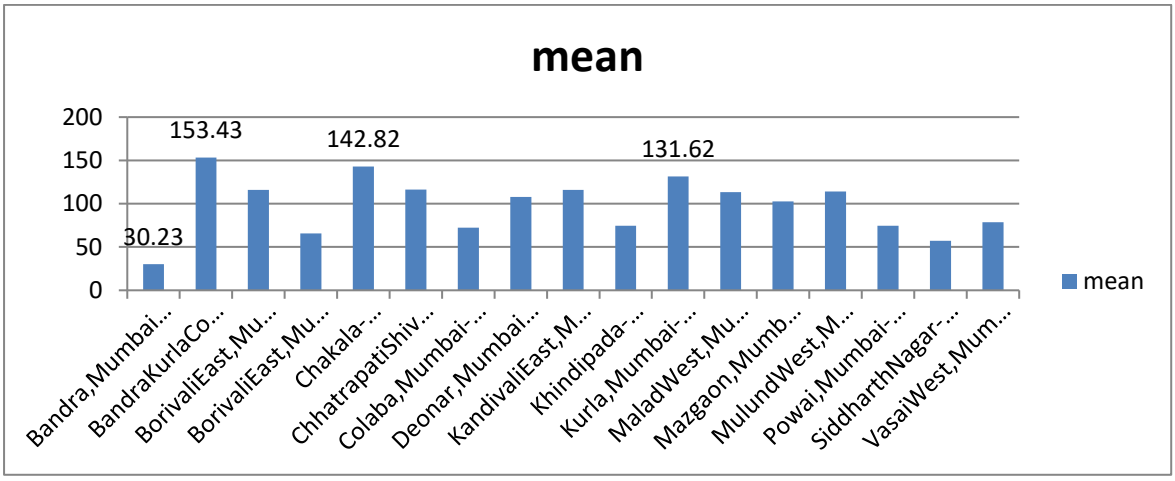


Figure 5.5Mean of pollutant PM_{10}

Conclusions:

1. The average concentration of PM_{10} is high in S2 (BandraKurlaComplex) 153.43 compared to other station.
2. Low concentration of PM_{10} is in S1 (Bandra,Mumbai-MPCB) station 30.23. That means the bandra is very low concentration of PM_{10}.

Table 5.3 Proportion of simple index for pollutant to calculate AQI

Statistics	PM2.5	PM10	NO	NO2	NH3	SO2	co	Total AQI
count	295	8218	934	120	278	269	141	10255
proportion	0.0287	0.8013	0.0910	0.0117	0.0271	0.0262	0.0137	1

In the above table we observe that the proportion of PM10 is high i.e. PM10 is more contributing than other pollutant. Secondly the NO and PM2.5 is more contributing.

Using the above table and section 2.3 we observe the PM10 and PM2.5 pollutant is more contributory and important pollutant to study.

Table 5.4 Descriptive statistics for AQI

Station	Count	mean	Std	min	25%	50%	75%	90%	Max	Color code
S1	1237	82.67	46.80	10.57	50	76	105.32	131.08	471.18	
S2	398	133.31	67.40	11.36	74.37	135.07	181.67	219.09	326.25	
S3	414	103.7	48.03	12.74	58.34	105.7	141.49	162.97	248.75	
S4	867	63.74	36.37	9.67	32.31	56.89	93.85	113.92	188.96	
S5	361	125.24	77.79	17.86	52.97	119.3	175.86	241.44	451.05	
S6	927	108.9	54.44	22.46	61.8	100.9	150.17	182.01	481.51	
S7	891	70.68	44.81	10.06	32.69	60.46	104.07	130.7	538.05	
S8	403	103.0	52.77	12.62	55.92	110.7	147.42	166.83	257.34	
S9	373	104.9	47.81	26.24	62.05	108.08	145.89	167.61	221.86	
S10	363	71.88	41.33	6.89	36.01	70.79	107.25	121.55	356.26	
S11	912	120.7	54.67	19.88	77.5	112.5	156.9	191.77	366.86	
S12	281	102.64	54.58	11.25	54.79	104.8	139.79	173.95	267.18	
S13	405	96.33	56.63	9.78	48.04	99.65	137.95	170.52	279.6	
S14	351	150.38	40.10	70.51	123.8	147.3	175.09	190.57	369.13	
S15	929	72.80	37.41	9.88	41.39	65.73	104.74	123.76	174.82	
S16	339	63.39	33.62	9.9	35.84	55.85	94.32	108.77	189.90	
S17	812	75.35	44.49	15.85	42.75	63.28	98.53	135.35	300.65	

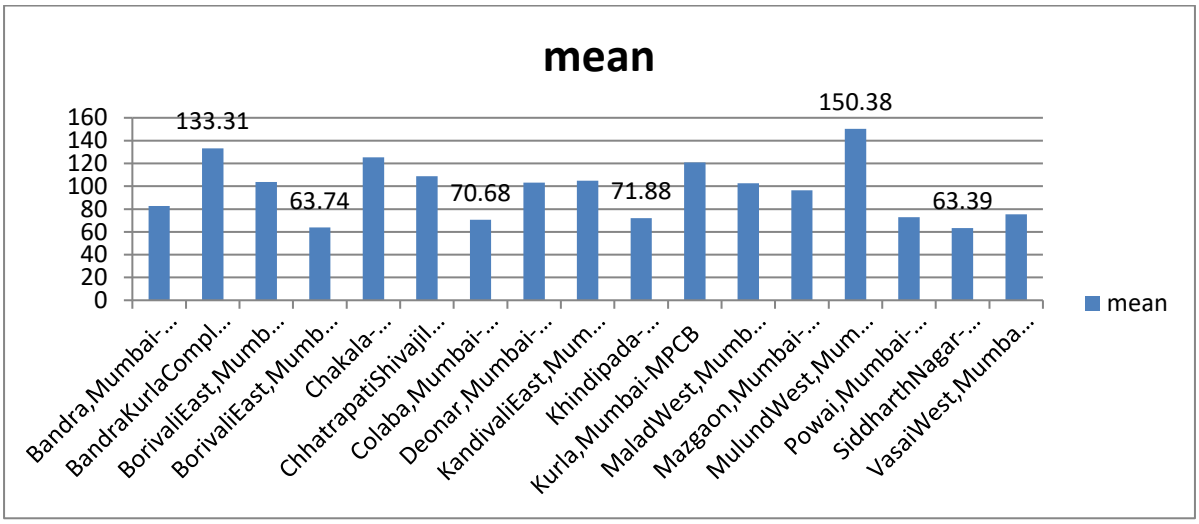


Figure 5.6 Mean of Air Quality Index (AQI)

Conclusion:

1. The average AQI high in S14 (Mulund West,Mumbai-MPCB) compare to other station.
2. Low AQI S16 (Siddhart Nagar Warli-Mumbai) station 63.39.
3. In above table 5.4 show that the AQI is greater than 101, so the Mumbai stations has moderate (see Table 2.1) AQI for Yellow color shaded Station.
4. In above table 5.4 show that the AQI is greater than 50, so the Mumbai stations haveSatisfactory (see Table 2.1) AQI for Light green color shaded Station.

Table 5.5 Station with range of AQI and area and population information

Station no	Station	Count (ni)	$\sum_{i=1}^{ni} Si$	AQI Range	Area KM sq	Population (Wi)	$\sum(Wi \times ni)$
S1	Bandra,Mumbai-MPCB	1237	101826.859	10-472	5.24	176708	218587796
S2	BandraKurlaComplex,Mumbai-IITM	398	53057.56	11-327	3.01	68547	27281706
S3	BorivaliEast,Mumbai-IITM	414	42958.59	12-249	5.22	167265	69247710
S4	BorivaliEast,Mumbai-MPCB	867	55262.82	9-189	5.22	167265	145018755
S5	Chakala-AndheriEast,Mumbai-IITM	361	45204.94	17-452	0.39	17479	6309919
S6	ChhatrapatiShivajiIntl.Airport(T2),Mumbai-MPCB	927	100951.95	22-482	6.61	16737	15515199
S7	Colaba,Mumbai-MPCB	891	62973.28	10-538	4.51	87280	77766480
S8	Deonar,Mumbai-IITM	403	41536.28	12-257	11.65	162967	65675701
S9	KandivaliEast,Mumbai-MPCB	373	39136.22	26-222	10.07	255484	95295532
S10	Khindipada-BhandupWest,Mumbai-IITM	363	26091.06	6-357	0.8	22613	8208519
S11	Kurla,Mumbai-MPCB	912	110103.87	19-367	4.29	149504	136347648
S12	MaladWest,Mumbai-IITM	281	28838.55	11-268	22.19	384137	107942497
S13	Mazgaon,Mumbai-IITM	405	39013.87	9-280	4.53	98364	39837420
S14	MulundWest,Mumbai-MPCB	351	52783.36	70-370	7.18	260209	91333359
S15	Powai,Mumbai-MPCB	929	67636.74	9-175	11.88	216957	201553053
S16	SiddharthNagar-Worli,Mumbai-IITM	339	21488.55	9-190	3.73	99917	33871863
S17	VasaiWest,Mumbai-MPCB	812	61187.76	15-300	39.09	401633	326125996

(Source: geoiq.io)

To find weighted average AQI for Mumbai city

$$Weighted\ Average = \frac{W1 \times \sum_{i=1}^{n1} S1 + W2 \times \sum_{i=1}^{n2} S2 + \cdots W17 \times \sum_{i=1}^{n17} S17}{\sum_{i=1}^{17} ni \times Wi}$$

Using this table we use the population as weight to find weighted average AQI for Mumbai city is 89.98, AQI for Mumbai city is satisfactory in nature.

Table 5.6 Distribution of AQI in Health Standards (In Percentage)

Station	Count	Good (0-50)	Satisfactory (51-100)	Moderate (101-200)	Poor (201-300)	Very Poor (301-400)	Severe (401-500)
S1	1237	25.55	35.81	36.94	1.29	0.32	0.16
S2	398	13.07	24.87	48.74	12.06	1.51	0
S3	414	15.94	31.64	50.72	1.69	0	0
S4	867	45.1	34.83	20.07	0	0	0
S5	361	22.99	17.73	41.83	16.07	1.11	0.28
S6	927	12.41	37.22	45.74	4.53	0	0.11
S7	890	42.13	31.01	26.52	0.34	0	0
S8	403	21.59	23.08	54.59	0.74	0	0
S9	373	15.82	30.56	52.01	1.61	0	0
S10	363	37.47	31.68	30.58	0	0.28	0
S11	912	7.24	32.89	51.75	7.79	0.33	0
S12	281	21	27.05	46.26	5.69	0	0
S13	405	27.41	22.72	45.68	4.2	0	0
S14	351	0	7.12	87.18	4.84	0.85	0
S15	929	35.52	34.23	30.25	0	0	0
S16	339	42.48	38.35	19.17	0	0	0
S17	812	33.25	42.36	22.41	1.85	0.12	0

In above table we classify the AQI (In percentage) according to the health standards in table 2.1 for 17 different stations given in table 3.1

Conclusion:

1. In above table we observe overall the AQI percentage is High for all station in Good, satisfactory and Moderate in condition i.e. the overall station is AQI is less than or equal to 200.
2. The AQI for station 5 (Chakala-AndheriEast,Mumbai-IITM) is nearly 16% and station 2 (BandraKurlaComplex,Mumbai-IITM) is 12% in poor condition compared to other station.
3. The station 14 Muland west Mumbai is high in percentage 87% is lie for Moderate condition.
4. Using the above table we create the action plan for reduce the AQI. And increase the Air quality.

Overall AQI classification in Percentagefor Mumbai City

Health standards	Good (0-50)	Satisfactory (51-100)	Moderate (101-200)	Poor (201-300)	Very Poor (301-400)	Severe (401-500)	total
Count	2660	3267	3992	319	22	4	10264
Percentage	25.92	31.83	38.89	3.11	0.21	0.04	100
Cumsum	25.92	57.75	96.64	99.75	99.96	100	

As above table we conclude that the AQI for Mumbai city is 26% in Good condition, 32% is in satisfactory condition and 39% is in Moderate condition. We see the 96.64% data is below moderate condition.

The AQI is 3% in poor or below condition.

5.3 Correlation of pollutants and AQI among the Monitoring Stations

A correlation Analysis has carried out for different monitoring stations for various pollutants and between air pollutant components, discussed in subsequent sections.

Table 5.7 Correlation analysis for $PM_{2.5}$ (see table 3.1 see station names)

station	1	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	1															
3	0.5	1														
4	-0.46	-0.44	1													
5	0.57	0.61	-0.61	1												
6	-0.54	-0.49	0.52	-0.63	1											
7	-0.41	-0.36	0.55	-0.51	0.52	1										
8	0.58	0.73	-0.5	0.67	-0.59	-0.38	1									
9	0.47	0.39	-0.59	0.58	-0.62	-0.57	0.52	1								
10	0.49	0.45	-0.62	0.61	-0.65	-0.55	0.54	0.68	1							
11	-0.49	-0.46	0.64	-0.65	0.74	0.51	-0.56	-0.6	-0.64	1						
12	0.65	0.78	-0.68	0.71	-0.71	-0.59	0.85	0.68	0.63	-0.67	1					
13	0.49	0.57	-0.46	0.56	-0.53	-0.35	0.61	0.49	0.49	-0.53	0.68	1				
14	0.42	0.27	-0.6	0.47	-0.65	-0.64	0.34	0.73	0.67	-0.61	0.6	0.35	1			
15	-0.51	-0.47	0.63	-0.64	0.73	0.48	-0.55	-0.61	-0.62	0.84	-0.7	-0.49	-0.6	1		
16	0.28	0.21	-0.5	0.33	-0.5	-0.46	0.21	0.54	0.46	-0.49	0.43	0.2	0.63	-0.47	1	
17	-0.13	0.08	0.4	-0.05	0.18	0.49	0.12	-0.37	-0.23	0.18	-0.18	-0.08	-0.55	0.17	-0.39	1

Conclusion:

1. The correlation between station 11 and station 15 (i.e. kurla and powai, table3.1) is 0.84 , it shows strong positive correlation between this two variables.

2. The correlation between station 5 and 3 is 0.61, station 12 and 3 is 0.78, station 8 and 5 is 0.67, station 12 and 1 is 0.65, station 11 and 6 is 0.74, station 10 and 9 is 0.68, and station 12 and 5 is 0.71, here we conclude that the all the above relationship between the station is positive correlation.

3. $PM_{2.5}$ shows positive correlation among most of the monitoring stations. The values ranged between 0.50-0.70. Similarly, some monitoring stations shows the negative correlation values ranged between -0.40 and -0.70.

Table 5.8 Correlation Analysis for PM_{10} (see table 3.1 see station name)

Station	1	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	1															
3	0.66	1														
4	-0.6	-0.57	1													
5	0.61	0.66	-0.72	1												
6	-0.57	-0.51	0.65	-0.68	1											
7	-0.4	-0.29	0.42	-0.54	0.4	1										
8	0.71	0.82	-0.65	0.69	-0.56	-0.38	1									
9	0.6	0.56	-0.7	0.7	-0.66	-0.47	0.63	1								
10	0.54	0.46	-0.69	0.65	-0.66	-0.56	0.55	0.7	1							
11	-0.55	-0.48	0.61	-0.66	0.74	0.44	-0.51	-0.64	-0.57	1						
12	0.71	0.86	-0.68	0.73	-0.72	-0.56	0.87	0.64	0.61	-0.66	1					
13	0.55	0.61	-0.53	0.54	-0.42	-0.3	0.61	0.53	0.46	-0.43	0.6	1				
14	0.63	0.6	-0.74	0.73	-0.71	-0.58	0.67	0.72	0.78	-0.63	0.69	0.53	1			
15	-0.63	-0.6	0.64	-0.71	0.74	0.37	-0.63	-0.68	-0.61	0.76	-0.71	-0.55	-0.69	1		
16	0.21	0.17	-0.4	0.34	-0.47	-0.39	0.18	0.43	0.47	-0.4	0.33	0.15	0.48	-0.37	1	
17	-0.02	0.02	0.29	-0.24	0.13	0.47	-0.1	-0.25	-0.38	0.26	-0.19	-0.26	-0.45	0.09	-0.46	1

Conclusion:

1. The correlation between station 12 and 3 (i.e. malad west and Borivali, table 3.1) is 0.86, station 12 and 8 is 0.87, station 8 and station 3 is 0.82 shows strong positive correlation between variables.
2. PM10 shows high positive correlation among most of the monitoring stations. The values ranged between 0.60-0.85. Similarly, some monitoring stations shows negative correlation values ranged between -0.30 and -0.75.

5.4Principal Component Analysis

Here we perform steps in pythonas below.

We know principal component analysis is type of multivariate analysis, so first we can perform some basic operations on our multivariate data. So we import some packages in python to perform these operations.

Python Code:

```
#importing packages

from datetime import datetime
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import scale
from sklearn.decomposition import PCA
```

```
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from scipy import stats
from IPython.display import display, HTML
# figures inline in notebook
%matplotlib inline
np.set_printoptions(suppress=True)
pd.set_option('display.max_rows', 20)

#reading multivariate data in python using pandas

data=pd.read_csv("D:\Data AQA\mumbai_multi.csv")
X = data.loc[:, "PM2.5":] # independent variables data
y = data.St_no# dependednt variable data
data
```

Output:

	St_no	PM2.5	PM10	NO	NO2	NH3	SO2	CO	Benzene	Ozone	CH4	CO2
0	1	28.85	289.36	70.61	58.47	174.30	3.01	0.20	5.0	13.72	NaN	420.08
1	1	90.91	266.66	59.70	51.01	153.50	3.24	0.35	5.0	8.06	NaN	417.61
2	1	26.19	237.63	76.37	92.88	157.16	8.28	1.11	5.0	22.12	NaN	426.32
3	1	27.24	243.11	86.95	45.14	185.24	3.26	0.76	5.0	8.11	NaN	433.62
4	1	27.93	180.92	62.56	68.73	131.72	7.24	1.80	5.0	19.06	NaN	323.93
...	...	...	...	...	...	...	...	...	...	...	...	...
12728	17	76.68	244.83	9.71	34.67	8.92	20.31	1.89	NaN	NaN	NaN	NaN
12729	17	86.49	234.80	9.14	30.31	9.43	20.84	1.84	NaN	NaN	NaN	NaN
12730	17	45.57	127.84	5.26	23.24	7.95	20.63	1.56	NaN	NaN	NaN	NaN
12731	17	22.87	183.29	9.45	31.19	9.06	20.34	1.90	NaN	NaN	NaN	NaN
12732	17	22.87	172.16	6.24	24.06	9.27	20.39	1.69	NaN	NaN	NaN	NaN

12733 rows × 13 columns

#plot multivariate data

```
#matrix scatterplot
pd.plotting.scatter_matrix(data.loc[:, "PM2.5":"CO2"], diagonal="kde",figsize=(20,15))
plt.show()
```

Output:

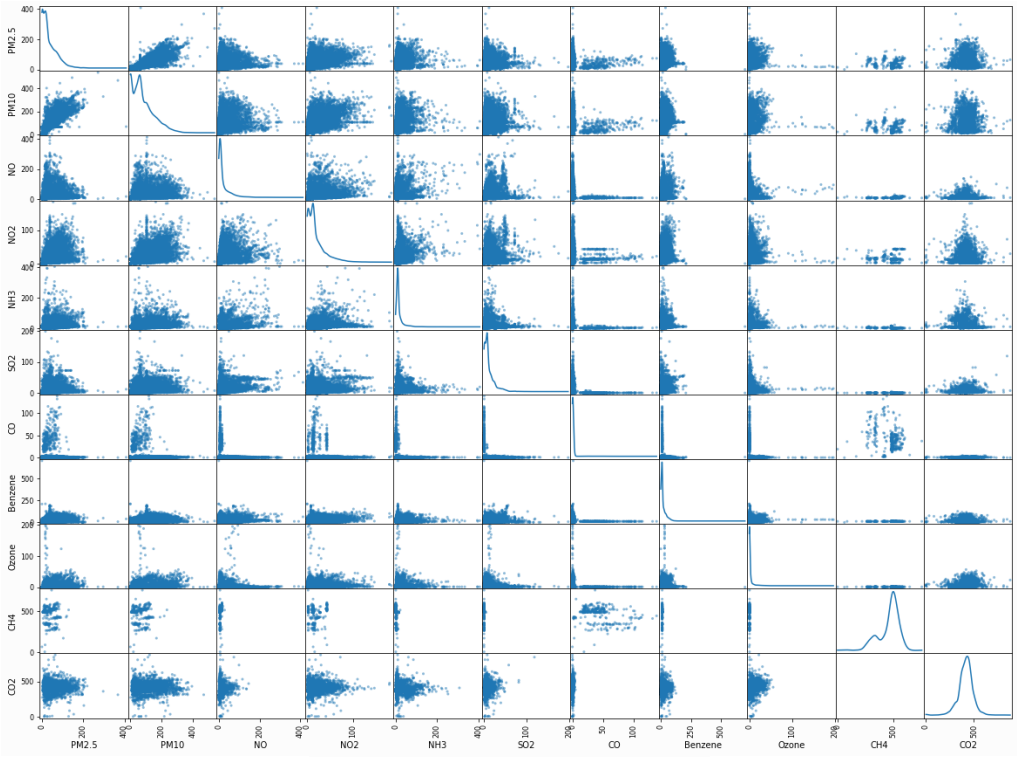


Figure 5.7 Matrix Scatter Plot

Conclusion:

One common way of plotting multivariate data is to make a matrix scatterplot, showing each pair of variables plotted against each other.

1. In this matrix scatterplot, the diagonal cells show histograms of each of the variables, in this case the concentrations of the air pollutants. Each of the off-diagonal cells is a scatter plot of two pollutants.
2. As we show in above graph the distribution of the Air Pollutant is nearly exponential for all pollutant and for CO2 and CH4 is showing different.

Python Code:

```
#A Profile Plot
ax=data[["PM2.5","PM10","NO","NO2","NH3","SO2","CO","Benzene","Ozone","CH
4","CO2"]].plot(figsize=(20,15))
ax.legend(loc='center left', bbox_to_anchor=(1, 0.5));
```

Output:

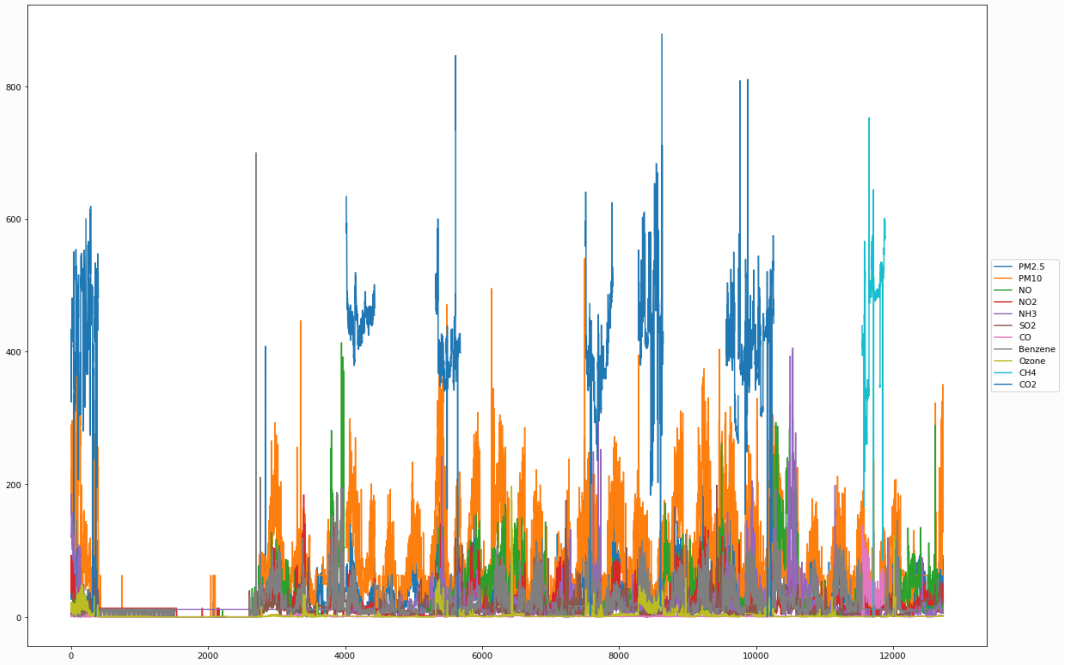


Figure 5.8 Profile plot of pollutants

Conclusion:

Profile plotwhich shows the variation in each of the variables, by plotting the value of each of the variables for each of the samples.

It is clear from the profile plot that the mean and standard deviation for PM2.5 and PM10 is quite a lot higher than that for the other pollutants.

Calculating Summary Statistics for Multivariate Data:

Python Code:

```
X.apply(np.mean)  #Calculate mean
```

Output:

```
PM2.5        34.945293
PM10         80.069920
NO           24.644706
NO2          19.513693
NH3          19.181224
SO2          11.969594
CO            1.925000
Benzene      20.333736
Ozone         3.440804
CH4         459.042358
CO2         414.361779
dtype: float64
```

```
X.apply(np.std)  #Calculate Standard Deviation
```

Output:

```
PM2.5          35.473559
PM10           72.101751
NO             36.091752
NO2            21.107528
NH3            25.848244
SO2            14.021576
CO              8.221628
Benzene        23.980602
Ozone           8.838628
CH4            94.752648
CO2            76.225597
dtype: float64
```

X.apply(np.max) #Maximum of data

Output:

```
PM2.5          408.07
PM10           540.44
NO             413.36
NO2            184.08
NH3            405.31
SO2            198.45
CO             138.05
Benzene        699.84
Ozone          196.88
CH4            752.58
CO2            878.59
dtype: float64
```

X.apply(np.min) # Minimum of data

Output:

```
PM2.5          0.00
PM10           0.00
NO             0.00
NO2            0.00
NH3            0.01
SO2            0.00
CO             0.00
Benzene        0.00
Ozone          0.00
CH4            0.00
CO2            0.00
dtype: float64
```

Conclusion:

The above code gives us mean, standard deviation, maximum and minimum of the multivariate data. We can see here that it would make sense to standardizing order t o compare the variables because the variables have very different standard deviations - the standard deviation of CO2 is 76.225597, while the standard deviation of CO is just 8. 221628.

Thus, in order to compare the variables, we need to standardize each variable s o that it has a sample variance of 1 and sample mean of 0. We will explain below how to standardize the variables.

Similarly, we can calculate Means and Variances Per Group, Between-groups Variance a nd Within-groups Variance for a Variable, Between-groups Covariance and Within-grou ps Covariance for Two Variables etc.

Calculating Correlations for Multivariate Data

Python Code:

```
corr = stats.pearsonr(data1.PM10,data1.NO)
print("p-value:\t", corr[1])
print("cor:\t\t", corr[0])
```

Output:

```
p-value:        6.019957765966942e-256
('cor:\t\t', 0.2960607986462041)
```

```
corrmat=data1.corr()
corrmat
```

Output:

Table 5.9 Correlation analysis of Pollutants

pollutants	PM2.5	PM10	NO	NO2	NH3	SO2	CO	Benzene	Ozone	CH4	CO2
PM2.5	1	0.85	0.14	0.53	0.27	0.16	0.04	0.43	0.22	-0.05	0.01
PM10	0.85	1	0.28	0.6	0.27	0.28	-0.01	0.44	0.23	-0.03	0.06
NO	0.14	0.28	1	0.32	0.34	0.33	-0.06	0.2	0	0.27	-0.03
NO2	0.53	0.6	0.32	1	0.31	0.29	0.02	0.41	0.15	0.14	0.03
NH3	0.27	0.27	0.34	0.31	1	0.03	-0.1	0.11	0.01	-0.18	-0.13
SO2	0.16	0.28	0.33	0.29	0.03	1	-0.11	0.22	-0.01	0.03	0.1
CO	0.04	-0.01	-0.06	0.02	-0.1	-0.11	1	0.22	0.07	-0.25	0.1
Benzene	0.43	0.44	0.2	0.41	0.11	0.22	0.22	1	0.11	0	-0.05
Ozone	0.22	0.23	0	0.15	0.01	-0.01	0.07	0.11	1	0	0.28
CH4	-0.05	-0.03	0.27	0.14	-0.18	0.03	-0.25	0	0	1	0
CO2	0.01	0.06	-0.03	0.03	-0.13	0.1	0.1	-0.05	0.28	0	1

Conclusion:

1. The correlation between PM_{10} and $PM_{2.5}$ is 0.85i.e. There is positive correlation between these pollutants.
2. The correlation between NO_2 and $PM_{2.5}$is 0.52, NO_2 and PM_{10} is 0.59 which is indicating that moderate positive relationship between the two variables.

3. This tells us that the correlation coefficient is about 0.29, which is a very weak correlation. Furthermore, the p-value for the statistical test of whether the correlation coefficient is significantly i.e. is nearly zero. This is much less than0.05, so there is very strong evidence that that the correlation is zero.

A better graphical representation of the correlation matrix is via a correlation matrix pl ot in the form of a heatmap.

Python Code:

```
sns.heatmap(corrmat, vmax=1., square=False).xaxis.tick_top()
```

Output:

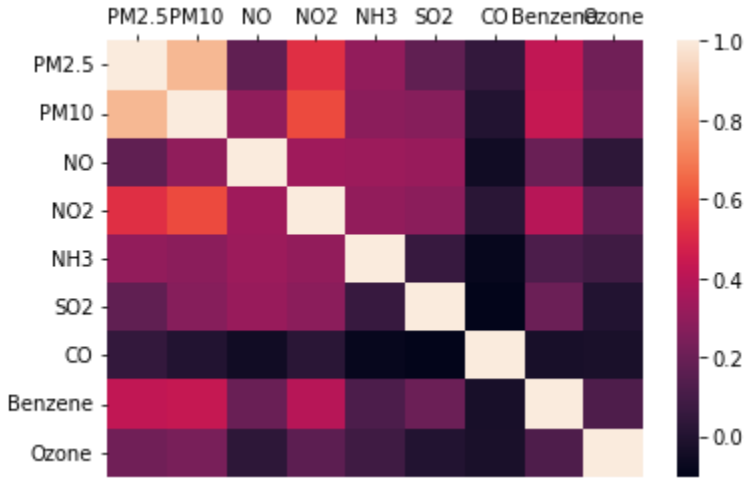


Figure 5.9 Correlation Matrix Heatmap

Standardising Variables:

If you want to compare different variables that have different units, are very different va riances, it is a good idea to first standardise the variables.

Python Code:

```
standardisedX = scale(data1)
standardisedX = pd.DataFrame(standardisedX, index=data1.index, columns=
data1.columns)
standardisedX.apply(np.mean)
```

Output:

```
PM2.5      -1.607133e-16
PM10        7.142816e-17
NO          8.928519e-17
NO2        -8.928519e-17
NH3        -7.142816e-17
SO2         7.589242e-17
CO          0.000000e+00
Benzene     1.071422e-16
Ozone       8.928519e-18
dtype: float64
```

standardisedX.apply(np.std)

Output:

```
PM2.5       1.0
PM10        1.0
NO          1.0
NO2         1.0
NH3         1.0
SO2         1.0
CO          1.0
Benzene     1.0
Ozone       1.0
dtype: float64
```

#PCA

To carry out a principal component analysis (PCA) on a multivariate data set, the first st ep is often to standardise the variables under study using the scale() function

Python Code:

```
data1
pca = PCA().fit(standardisedX)
def pca_summary(pca, standardised_data, out=True):
   names = ["PC"+str(i) for i in range(1, len(pca.explained_variance_ratio_)+1)]
   a = list(np.std(pca.transform(standardised_data), axis=0))
   b = list(pca.explained_variance_ratio_)
   c = [np.sum(pca.explained_variance_ratio_[:i]) for i in range(1, len(pca.explained_vari
ance_ratio_)+1)]
   columns = pd.MultiIndex.from_tuples([("sdev", "Standard deviation"), ("varprop", "Pr
oportion of            Variance"), ("cumprop", "Cumulative Proportion")])
   summary = pd.DataFrame(list(zip(a, b, c)), index=names, columns=columns)
   if out:
print("Importance of components:")
display(summary)
return summary
summary = pca_summary(pca, standardisedX)
```

Output:

Importance of components:

	sdev	varprop	cumprop
	Standard deviation	Proportion of Variance	Cumulative Proportion
PC1	1.773395	0.349437	0.349437
PC2	1.101434	0.134795	0.484232
PC3	1.005087	0.112244	0.596476
PC4	0.982635	0.107286	0.703762
PC5	0.910075	0.092026	0.795788
PC6	0.819837	0.074681	0.870470
PC7	0.733370	0.059759	0.930229
PC8	0.703666	0.055016	0.985245
PC9	0.364411	0.014755	1.000000

Figure 5.10 PCA variability of Pollutant

This gives us the standard deviation of each component, and the proportion of variance explained by each component. The standard deviation of the components is stored in a named row called 'sdev'of the output variable made by the 'pca_summary' function.

Conclusion:

1. As the above output we observe that the Principle component 1 ($PM_{2.5}$) is 35%variability explain form total variance and the PC2 (PM10) and PC3 (NO) is 14% and 12% variability explain out of total variation.
2. The PC1, PC2 and PC3 are 59% variability explain. In above output in cumpropcolumn we conclude that to fit the best model we want to use the at least 6 Principle components (corresponding factors) i.e. the model get the 87% variability explain.
3. In the above scenario we cannot reduce the data dimension because the high variability is obtain when we used at least 6 or 5 PC so in data reducibility is no so important we go with the original data.

```
np.sum(summary.sdev**2)
```

Output:

```
Standard deviation    9.0
dtype: float64
```

Conclusion:

In this case, we see that the total variance is 9, which is equal to the number of standardised variables. This is because for standardised data, the variance of each standardised variable is 1. The total variance is equal to the sum of the variances of the individual variables, and since the variance of each standardised variable is 1, the total variance should be equal to the number of variables.

Deciding How Many Principal Components to Retain:

Python Code:

```
def screeplot(pca, standardised_values):
    y = np.std(pca.transform(standardised_values), axis=0)**2
    x = np.arange(len(y)) + 1
plt.plot(x, y, "o-")
plt.xticks(x, ["Comp."+str(i) for i in x], rotation=60)
plt.ylabel("Variance")
plt.show()
screeplot(pca, standardisedX)
```

Output:

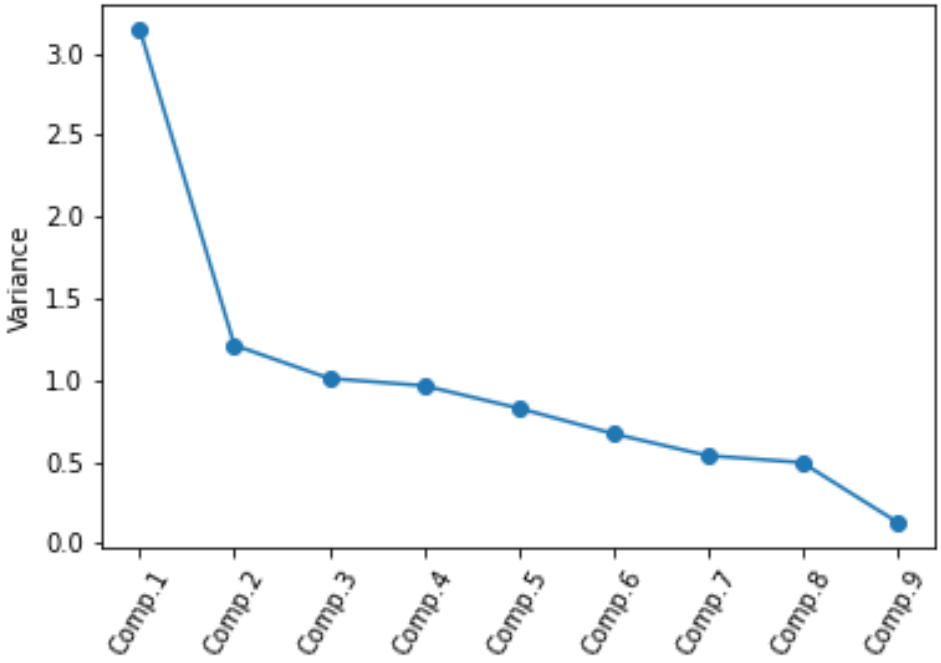


Figure 5.11 Scree Plot

Conclusion:

The most obvious change in slope in the scree plot occurs at component 4, which is the " elbow" of the scree plot. Therefore, it count be argued based on the basis of the scree pl ot that the first three components should be retained.

Scatterplots of the Principal Components

Python Code:

```
def pca_scatter(pca, standardised_values, classifs):
    foo = pca.transform(standardised_values)
    bar = pd.DataFrame(list(zip(foo[:, 0], foo[:, 1], classifs)), columns=["PC1", "PC2",
"Class"])
sns.lmplot("PC1", "PC2", bar, hue="Class", fit_reg=False)

pca_scatter(pca, standardisedX, y)
```

Output:

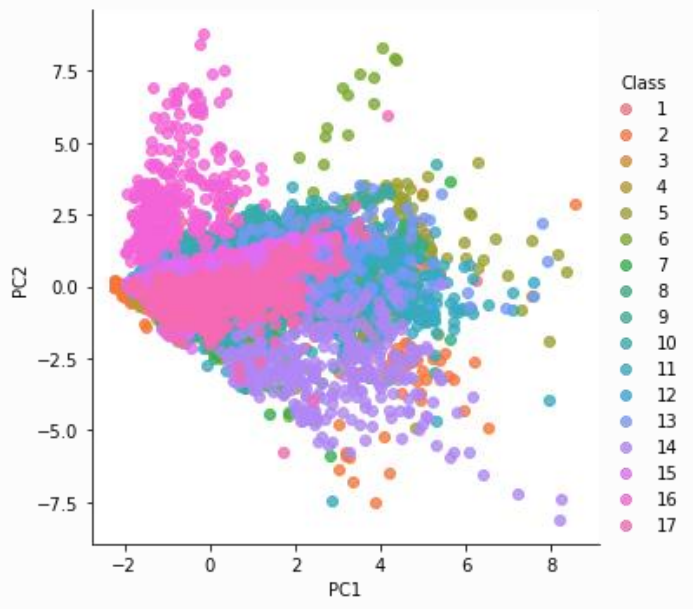


Figure 5.12 Scatter Plot

Conclusion:

The scatterplot shows the first principal component on the x-axis, and the second principal component on the y-axis. We can see from the scatterplot that pollutant concentration samples of station 1 have much lower values of the first principal component than pollutant concentration samples of station 17. Therefore, the first principal component separates pollutant concentration samples of station 1 from those of station 17.

We can also see that pollutant concentration samples of station 7 have much higher values of the second principal component than pollutant concentration samples of station 1 and 3. Therefore, the second principal component separates samples of station 7 from samples of other stations.

5.5 Multiple Linear Regression

As above PCA see in fig 5.10 we see that the first 7 component is explained 93% variability but also using table 5.9 observe the correlation between the PM2.5 and PM_{10} has 0.85 correlation i.e. is very high correlation, to remove multicollinearitywe drop any one of them if we drop we PM_{10} as using information of PCA but the model gives 74% R^2 and very high residual that why we used $PM_{10}$and drop $PM_{2.5}$using this information we fit the Multiple Linear regression.

To fit the multiple linear regression we used PM10, NO, NO_2, NH_3, SO_2, and CO used as independent variable and AQI is used as dependent variable. Correlation between above 6 components given in table 5.10.

The coefficients of the MLR equation are given in Table 5.11

Table 5.10 Correlation coefficients of air pollutants

	PM10	NO	NO2	NH3	SO2	CO
PM10	1	0.2	0.53	0.25	0.14	-0.07
NO	0.2	1	0.28	0.31	0.25	-0.08
NO2	0.53	0.28	1	0.28	0.2	-0.01
NH3	0.25	0.31	0.28	1	0.02	-0.1
SO2	0.14	0.25	0.2	0.02	1	-0.15
CO	-0.07	-0.08	-0.01	-0.1	-0.15	1

Table 5.11 Coefficients of MLR equation.

Model Parameters	Values
y-intercept ($\boldsymbol{\alpha_0}$)	8.9981
$\boldsymbol{\alpha_1}$	0.7151
$\boldsymbol{\alpha_2}$	0.2831
$\boldsymbol{\alpha_3}$	0.0224
$\boldsymbol{\alpha_4}$	0.2054
$\boldsymbol{\alpha_5}$	0.0466
$\boldsymbol{\alpha_6}$	0.0752

The training data consist of 80% (8027 samples) of the aggregate data, and the data sample is taken randomly for training the MLR algorithm. The test data set consists of 20% (2007 data samples) of the whole data set.

The residual between actual and predicted value for test data and accuracy of the MLR algorithm. The calculated quantitative performance indices are given in Table 5.12 and Table 5.13.

Table 5.12 Actual VS predicted value

sr.no	Actual Value	Predicted Value	Residual
4456	31.56	35.38998	-3.82998
4561	33.1	35.98688	-2.88688
4100	196.09	208.3746	-12.2846
1312	50.03	49.47818	0.551821
…	…	…	…
4313	113.45	101.2417	12.20827
2140	66.85	61.07994	5.770065
7586	60.15	55.23889	4.911109
7565	15.16	24.71498	-9.55498
7920	134.59	164.8384	-30.2484

Table 5.13 Performance indices of Multi-Linear Regression model

Performance Indices	Training Data	Testing Data
R^2	90.57	91.31
CORR	0.9517	0.9556
MAE	10.639	9.885
RMSE	16.963	15.50
MAPE	0.1933	0.1750

Conclusion:

1. From table 5.10 correlation coefficients, this inferred that they are no strong correlation between any two air pollutants.
2. Here we can see the difference between Actual values and predicted values which are not very high.
3. We have already seen that the accuracy of the test model is about 91.31% .i.e. the model is good fit.
4. Mean absolute percentage error (MAPE) is 0.1933 for train model and 0.1750 for testing data i.e. both the model is more than 80% accurate.

5.6 Time Series Analysis and Forecasting

Time-series plots were prepared for several stations to compare forecasts with actual observations.

Mumbai city is 17 monitoring stations, so we built the model for onestations Chhatrapati Shivaji International Airport (T2), Mumbai-MPCB.

Station 6 for Chhatrapati Shivaji International Airport (T2), Mumbai-MPCB

1) Time series analysis for PM_{10}

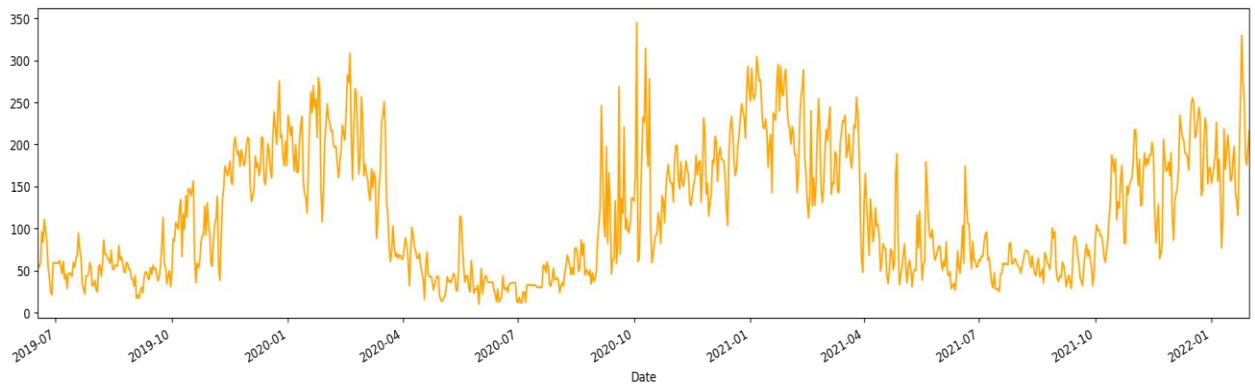


Figure 5.13Time series plot for daily average PM_{10} concentrations(July 2019- Dec 2021)

Conclusion:

In time series plot, we observe that daily concentration PM_{10}data is present of some seasonal component and absent of trend component. To checking a seasonality and stationary we also plot Autocorrelation plot.

Table 5.14 Descriptive statistics (July 2019- Dec 2021)

	PM2.5	PM10	NO	NO2	NH3	SO2	CO	Benzene	Ozone	AQI
Count	928	928	922	919	922	917	913	913	911	928
Mean	36.65	115.09	59.57	28.03	17.61	10.30	0.93	20.69	4.48	108.61
Std	27.61	73.88	26.7	19.62	12.93	7.56	0.64	18.83	20.97	52.89
Min	4.29	10.15	0.03	0.15	0.54	0.03	0	0.07	0	24.43
25%	14.80	52.5	40.96	13.58	6.84	5.48	0.31	6.89	0.4	62.14
50%	25.83	91.1	56.39	24.59	15.87	9.47	0.87	13.47	1.18	100.74
75%	53.11	174.87	74.34	39.24	26.12	13.45	1.48	30.98	2.87	149.91
Max	155.28	344.8	169.79	112.52	86.21	107.97	2.62	107.14	275.7	294.8

Autocorrelation and Partial autocorrelation Plot for PM2.5

Using the autocorrelation, we check the presence of seasonal component and checking of the process is stationary or not. ACF is used to indicate and how similar a value is within a given time series and the previous value.

Python Code:

```
#Autocorrelation plot
from pandas.plotting import autocorrelation_plot
# Draw Plot
plt.rcParams.update({'figure.figsize':(9,5), 'figure.dpi':120})
autocorrelation_plot(df2['PM2.5'].tolist()
```

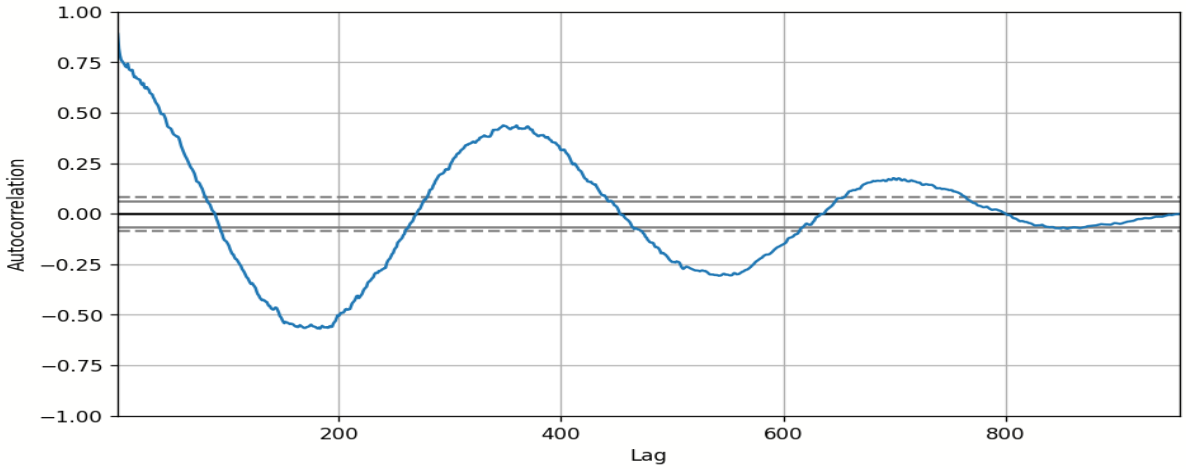


Figure 5.14 Autocorrelation Plot for n-lag

Conclusion:

In the above ACF plot we conclude that there is presence of seasonality, but the graph is not having exponentially decayed so it is not stationary.

We can limit the number of lags on x-axis to 50 to make the plot easier to read

ACF and PACF plot for 50-lags:

PACF is basically instead of finding correlations of present with lags like ACF, it finds correlation of residuals(which remains after removing the effects which are already explained by the earlier lag(s)) with the next lag value hence 'partial' and not 'complete' .

In PACF few examine the spikes at each lag to determine whether they are significance. A significance spikes will extend beyond the significant limits which indicates that the correlation for that lag doesn't equal zero.

Python Code:

```
fig, axes = plt.subplots(1, 2, figsize=(20,7), dpi= 80)
#acf plot of PM2.5
from statsmodels.graphics.tsaplots import plot_acf
plot_acf(df2['PM2.5'], lags = 50,ax=axes[0])
#pacf plot PM2.5
From statsmodels.graphics.tsaplots import plot_pacf
plot_pacf(df2['PM2.5'], lags = 50,ax=axes[1])
plt.show()
```

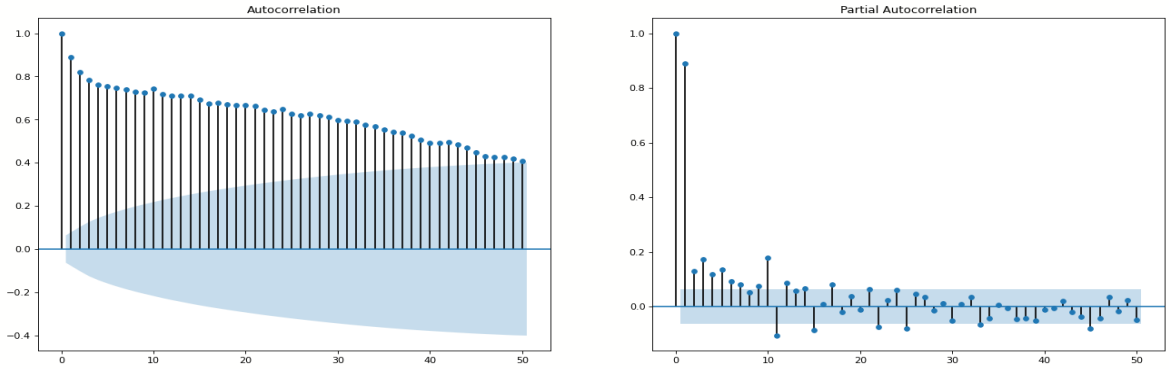


Figure 5.15ACF and PACF

Conclusion:

1. In above ACF plot for 50-lags we clearly see that is slow decay that means the time series is a not stationary time series.
2. In that above pacf plot there is a significant correlation at up to lag 1 and some positive and negative correlation is present that means after lag 1 also other lags are also significant , hence we conclude that a higher order moving average term in the data is present

Test for stationary

To check stationarity of the time series, using the ADF (Augmented Dickey Fuller Test) and KPSS (Kwiatkowski-phillips-schmidt-shin test) tests

The most commonly used is the ADF test, where the null hypothesis is the time series possesses a unit root and is non-stationary i.e. H0: Non-stationary. So, the P-Value in ADF test is less than the significance level (0.05), you reject the null hypothesis.

The KPSS test, on the other hand, is used to test for trend stationary. The null hypothesis and the P-Value interpretation is just the opposite of ADF test i.e. H0: stationary. The below code implements these two using 'statmodels' package in python.

Python Code:

```
from statsmodels.tsa.stattools import adfuller, kpss
# ADF Test
result = adfuller(df2['PM2.5'].values, autolag='AIC')
print(f'ADF Statistic: {result[0]}')
print(f'p-value: {result[1]}')
for key, value in result[4].items():
print('Critial Values:')
print(f'   {key}, {value}')
# KPSS Test
result = kpss(df2['PM2.5'], regression='c')
print('\nKPSS Statistic: %f' % result[0])
print('p-value: %f' % result[1])
for key, value in result[3].items():
print('Critial Values:')
print(f'   {key}, {value}')
```

Output:

```
ADF Statistic: -1.9528820022252606
p-value: 0.3076068669176704
Critial Values:
  1%, -3.4373257950466174
Critial Values:
  5%, -2.864619627202065
Critial Values:
  10%, -2.568409774784971

KPSS Statistic: 0.254518
p-value: 0.100000
Critial Values:
  10%, 0.347
Critial Values:
  5%, 0.463
Critial Values:
  2.5%, 0.574
Critial Values: 1%, 0.739
```

Conclusion:

1) According to ADF test the P-Value=0.3076>0.05 i.e. we fail to reject H0, therefore the Time series is not stationary.

2) As KPSS the P-Value =0.1 > 0.0.5, i.e. fail to reject H0, therefore the Time series is stationary.

3) We most used ADF test, finally we conclude that the process is non stationary we need differencing to get stationary process.

Granger Causality test

The Granger causality test is a statistical hypothesis test of determining whether one time series is useful for forecasting another. If the probability value is less than any α level, then the hypothesis would be rejected at that level.

Python code

```
#Granger Causality test
from statsmodels.tsa.stattools import grangercausalitytests
#.dt.month
data['Date'] = data.Date.dt.month
grangercausalitytests(data[['PM2.5', 'Date']], maxlag=2)
```

Output

```
Granger Causality
number of lags (no zero) 1
ssr based F test:         F=0.0007  , p=0.9791  , df_denom=953, df_num=1
ssr based chi2 test:   chi2=0.0007  , p=0.9791  , df=1
likelihood ratio test: chi2=0.0007  , p=0.9791  , df=1
parameter F test:         F=0.0007  , p=0.9791  , df_denom=953, df_num=1

Granger Causality
number of lags (no zero) 2
ssr based F test:         F=0.4440  , p=0.6416  , df_denom=950, df_num=2
ssr based chi2 test:   chi2=0.8926  , p=0.6400  , df=2
likelihood ratio test: chi2=0.8922  , p=0.6401  , df=2
parameter F test:         F=0.4440  , p=0.6416  , df_denom=950, df_num=2
```

Conclusion: In Granger Causality test at no of lags 1 and 2 both the p-value is greater than 0.05 so we conclude that the above time series is useful for forecasting another.

Method of Differencing

Here, we have a non-stationary Causal time series to get non-stationary to stationary we used method of differencing, in the method of differencing we used iterative method then we plot the ACF , PACF and test for stationary to check the stationarity.

First Order Differencing

Python code:

```
#first order differencing of raw data
df2['PM2.5_with_diff_1'] = df2['PM2.5'] - df2['PM2.5'].shift(1)
```

ACF and PACF

Python code:

```
fig, axes = plt.subplots(1, 2, figsize=(20,7), dpi= 80)
#acf plot of dataframe of diff 1
from statsmodels.graphics.tsaplots import plot_acf
plot_acf(df2['PM2.5_with_diff_1'][1:], lags = 50,ax=axes[0])
#pacf plot of dataframe of diff 1
from statsmodels.graphics.tsaplots import plot_pacf
plot_pacf(df2['PM2.5_with_diff_1'][1:], lags = 50,ax=axes[1])
plt.show()
```

Output:

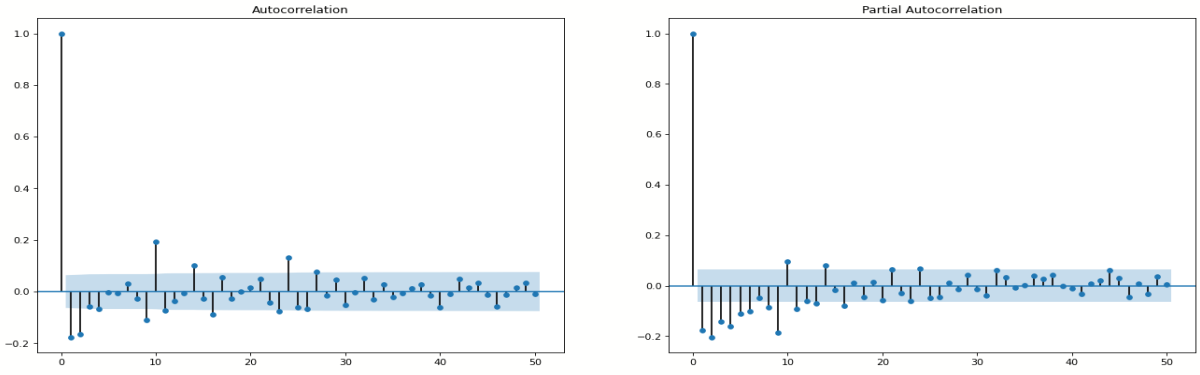


Figure 5.16 ACF and PACF for 1st order differencing

Test for stationarity:

Python code:

```
#ADF test on data with diff 1
 from statsmodels.tsa.stattools import adfuller
result = adfuller(df2['PM2.5_with_diff_1'][1:])
print(result)
print('ADF Statistic: %f' % result[0])
print('p-value ADF: %f' % result[1])
```

```
# KPSS test
from statsmodels.tsa.stattools import kpss
result1 = kpss(df2['PM2.5_with_diff_1'][1:])
print(result1)
print('kpss Statistic: %f' % result1[0])
print('p-value kpss: %f' % result1[1])
```

Output:

(-10.698947102455671, 3.564322887998363e-19, 15, 940, {'1%': -3.4373257950466174, '5%': -2.864619627202065, '10%': -2.568409774784971}, 9099.43073314422)

ADF Statistic: -10.698947

p-value ADF: 0.000000

(0.04384602928008686, 0.1, 22, {'10%': 0.347, '5%': 0.463, '2.5%': 0.574, '1%': 0.739})

kpss Statistic: 0.043846

p-value kpss: 0.100000

Conclusion:

1) Using above ACF plot Figure 5.8 we conclude that there is fast decay of the graph so it is stationary after 1 order differencing similarly we conclude for the pacf there is lag significance as 0 and negative correlation between points significance.

2) According to ADF test the P-Value=0.000<0.05 i.e. we reject H0, therefore the Time series is stationary.

3) As KPSS the P-Value =0.1 > 0.0.5, i.e. fail to reject H0, therefore the Time series is stationary.

4) Finally we conclude that the process is stationary.

ARIMA Model

Python Code:

```
# Create Training and Test
train_pm10 = pm10['PM10']['17-06-2019':'22-12-2021']
test_pm10 = pm10['PM10']['22-12-2021':'31-12-2021']
[train_pm10.count(),test_pm10.count()]
from statsmodels.tsa.arima_model import ARIMA
from pmdarima import auto_arima

model=auto_arima(train,start_p=1,seasonal=True,D=1,trace=True,error_actio
n='ignore',stepwise=True)
```

Output:

The auto.arima is used AIC approach to detect best fit model.

So according to auto.arima

Best model: ARIMA(1,1,2)(0,0,0)[0] with AIC:8926 BIC: 8951

Summary to check SARIMA

Statespace Model Results			
Dep. Variable:	Y	**No. Observations:**	387
Model:	SARIMAX(2, 1, 1)	**Log Likelihood**	-1858.299
Date:	Sun, 29 May 2022	**AIC**	3724.599
Time:	17:36:18	**BIC**	3740.422
Sample:	0	**HQIC**	3730.874
	- 387		
Covariance Type:	Opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.6656	0.051	13.023	0.000	0.565	0.766
ar.L2	-0.1676	0.048	-3.456	0.001	-0.263	-0.073
ma.L1	-0.8635	0.039	-22.158	0.000	-0.940	-0.787
sigma2	888.0476	45.395	19.563	0.000	799.075	977.020

Ljung-Box (Q):	37.18	**Jarque-Bera (JB):**	87.92
Prob(Q):	0.60	**Prob(JB):**	0.00

Conclusion:

1) the best model using the AIC criteria to fit the best model is Best model: ARIMA(2,1,1)(0,0,0)[0]

2) There is no present of SARIMA model i.e. No present of seasonal component after 1 order differencing, here only ARIMA model is used

3) Using above Ljunk-Box test Prob(Q) >0.05 i.e. the data are independently distributed.

During cross validation of auto.arima using arima function we fit the model individually and get the best model.

Python code:

```
## create a ARIMA model
from statsmodels.tsa.arima_model import ARIMA
# 1,1,2 ARIMA Model
model2=ARIMA(train_pm10, order=(3,1,3))
model_fit = model2.fit()
print(model_fit.summary())
```

Output:

ARIMA Model Results

Dep. Variable:	D.PM10	No. Observations:	918
Model:	ARIMA(3, 1, 3)	Log Likelihood	-4448.213
Method:	css-mle	S.D. of innovations	30.756
Date:	Sun, 29 May 2022	AIC	8912.425
Time:	17:45:53	BIC	8951.003
Sample:	1	HQIC	8927.148

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.1675	0.259	0.646	0.519	-0.341	0.676
ar.L1.D.PM10	0.1694	0.044	3.819	0.000	0.082	0.256
ar.L2.D.PM10	-0.7186	0.029	-24.868	0.000	-0.775	-0.662
ar.L3.D.PM10	0.5632	0.037	15.192	0.000	0.491	0.636
ma.L1.D.PM10	-0.5188	0.036	-14.557	0.000	-0.589	-0.449
ma.L2.D.PM10	0.6075	0.038	16.192	0.000	0.534	0.681
ma.L3.D.PM10	-0.8387	0.028	-30.295	0.000	-0.893	-0.784

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-0.2162	-0.9963j	1.0195	-0.2840
AR.2	-0.2162	+0.9963j	1.0195	0.2840
AR.3	1.7083	-0.0000j	1.7083	-0.0000
MA.1	-0.1984	-1.0120j	1.0313	-0.2808
MA.2	-0.1984	+1.0120j	1.0313	0.2808
MA.3	1.1212	-0.0000j	1.1212	-0.0000

We compare the various model apart from the auto.arima function we compare the AIC value of the models so to compare the AIC value of ARIMA(2,1,1) model and ARIMA (3,1,3) the AIC value is less for the ARIMA(3,1,3) and BIC are less.

Table 5.15 Best fitted model

Model Coefficients	**Values**
Auto Regressive Model ($\boldsymbol{p}$)	3
Differencing Model ($\boldsymbol{d}$)	1
Moving Average Model ($\boldsymbol{q}$)	3

Conclusion:

The model summary reveals a lot of information.

1) Notice here the coefficient of the AR3 P-value i.e. P>|Z| column the AR3 is significant in the model, similarly the MA3 term is also significant in the model , therefore the model is ARIMA(3,1,3) Autoregressive term p=3, difference d=1, and Moving average q=3.

2) Root of AR3 polynomial is greater than 1 so the root of AR polynomial is outside the unit circle i.e. the model is causal stationary model.

3) Root of MA3 polynomial is greater than 1 so the root of MA polynomial is outside the unit circle i.e. the model is invertible stationary model.

Residual plot

Lets plot the residuals to ensure there are no patterns (i.e. look for constant mean and variance).

Python Code:

```
# Plot residual errors
residuals = pd.DataFrame(model_fit.resid)
fig, ax = plt.subplots(1,2)
residuals.plot(title="Residuals", ax=ax[0])
residuals.plot(kind='kde', title='Density', ax=ax[1])
plt.show()
```

Output:-

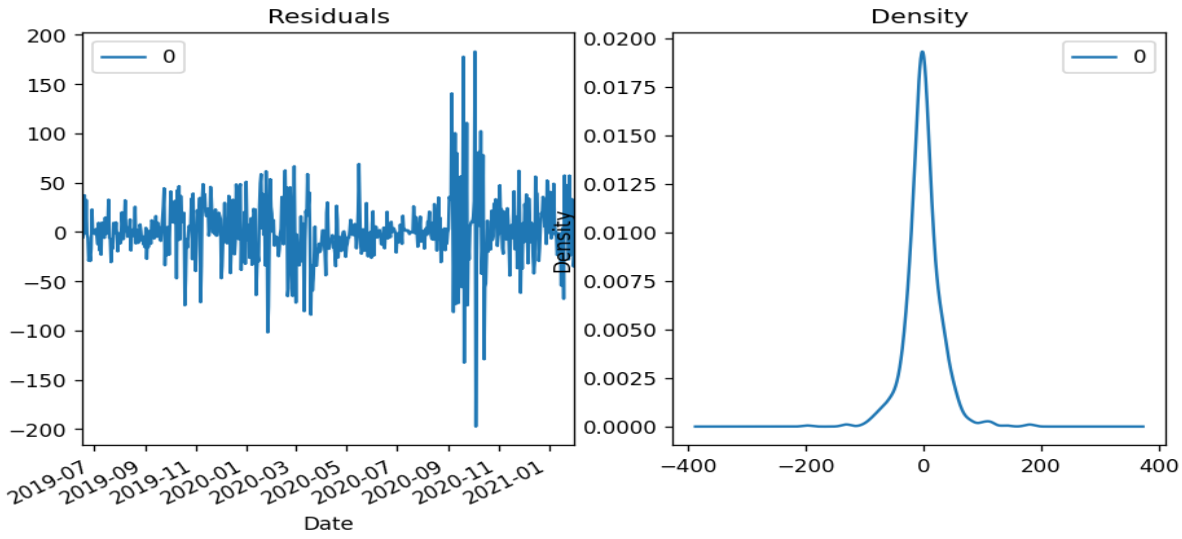


Figure 5.17 Residual plot and density plot

Conclusion:

The residual errors term fine with near zero mean and uniform variance.

ModelDiagnostics

Python Code:

```
model.plot_diagnostics(figsize=(15,8))
plt.show
```

Output:

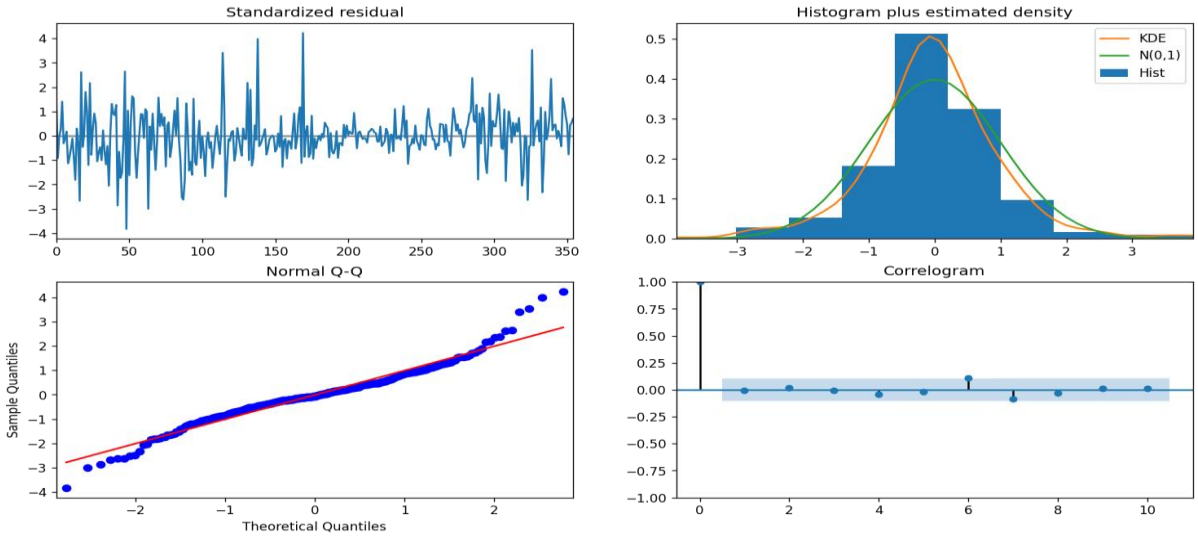


Figure 5.18ModelDiagnostics

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. Any significant deviations would imply the distribution is skewed.

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. So you will need to look for more X's (predictors) to the model.

Actual VS Fitted

Python Code:

```
# Actual vs Fitted
model_fit.plot_predict(dynamic=False)
plt.show()
```

Output:

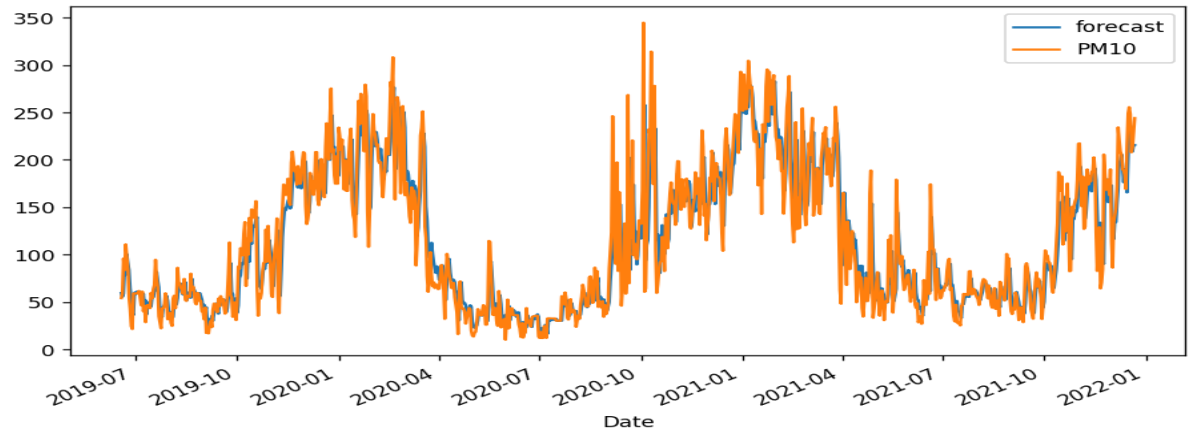


Figure 5.19 Actual vs. forecasted valuesFor train data

Conclusion:

The actual value and forecast value of $PM_{2.5}$ concentration is good fit in 95% CI.

ACF plot of Residual and forecast test set

Python Code:

```
from statsmodels.tsa.stattools import acf
# acf plot of residuals
 plot_acf(residuals, lags = 50)
 plt.show()
 y_pred=pd.Series(model_fit.forecast(30)[0], index=test.index)
 y_true=test
print (np.array(y_pred))
print (np.array(y_true))
 df=pd.DataFrame(y_pred,y_true)
```

Output:

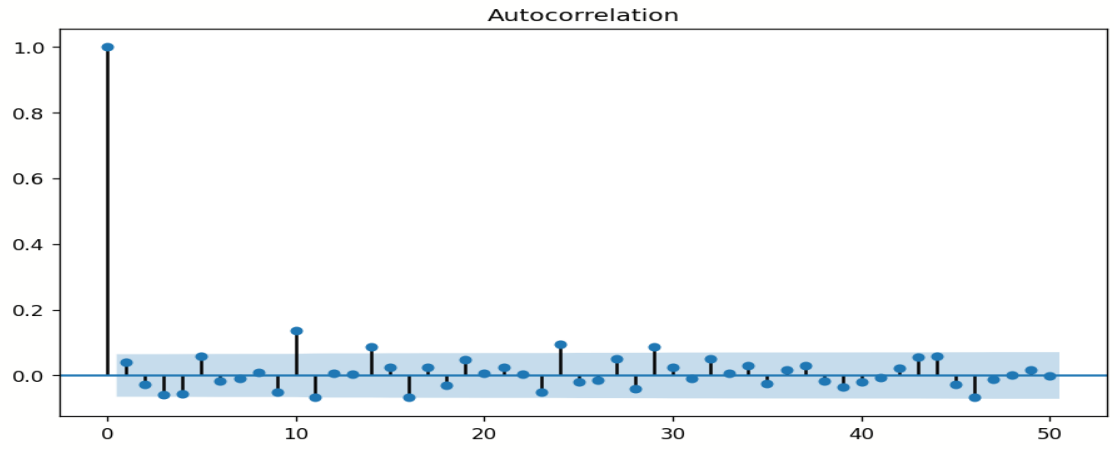


Figure 5.20 ACF for residual

Table 5.16 Actual VS predicted values for PM10

Date	Actual Value	Predicted Value
22/12/2021	244.13	228.4245
23/12/2021	233.29	229.5114
24/12/2021	139.32	226.4552
25/12/2021	144.59	216.4757
26/12/2021	202.54	217.7594
27/12/2021	231.29	223.5921
28/12/2021	220.3	218.2019
29/12/2021	153.13	213.9855
30/12/2021	170.92	220.5953
31/12/2021	172.46	221.874

Conclusion:

1) As observe that the above ACF plot for residual is significant at the 0 lag. We conclude that the residual has follows white noise.

2) As per the table 5.9 the actual values are some values are nearly same and some are deviated from the forecasted values

Plot Actual VS fitted for test data

Python code:

```
df=pd.DataFrame()
df['Actual Value']=y_true
df['Predicted Value']=y_pred
df.plot()
```

Output:

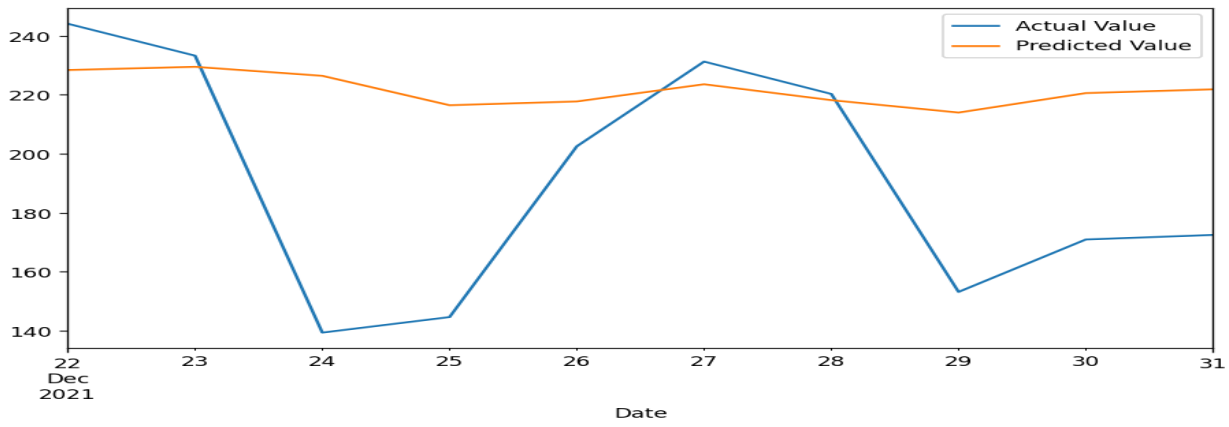


Figure 5.21Validation of forecasting of PM_{10} using ARIMA model with test data

Accuracy Metrics for Time Series Forecast

Typically, if you are comparing forecasts of two different series, the MAPE, Correlation and Min-Max Error can be used.

Because only the above three are percentage errors that vary between 0 and 1. That way, you can judge how good the forecast irrespective of the scale of the series is.

The other error metrics are quantities. That implies, an RMSE of 100 for a series whose mean is in 1000's is better than an RMSE of 5 for series in 10's. So, you can't really use them to compare the forecasts of two different scaled time series.

Python code:

```
mape = np.mean(np.abs(y_pred -
y_true) /np.abs(y_true)) # Mean absolute percentage error -
mae = np.mean(np.abs(y_pred - y_true)) # Mean absolute error
mpe = np.mean((y_pred - y_true)/y_true) # Mean percentage error
rmse = np.mean((y_pred - y_true)**2)**.5 # RMSE
corr = np.corrcoef (y_pred, y_true) [0,1]
# Correlation Coefficient
mins = np.amin(np.hstack([y_pred[:,None], y_true[:,None]]), axis=1)
maxs = np.amax(np.hstack([y_pred[:,None], y_true[:,None]]), axis=1)
minmax = 1 - np.mean(mins/maxs)   # minmax
import pprint
pprint.pprint({'mape':mape,'mae':mae,'mpe':mpe,'rmse':rmse,'corr':corr,'minma
x' :minmax})
```

Output

Table 5.17 Quantitative Performance indices of ARIMA model

Performance Indices	Test Data
CORR	0.4914
MAE	36.3465
RMSE	46.8285
MAPE	0.2295
MPE	0.2048
MINMAX	0.1642
Average Accuracy of Test Data	77.05%

Conclusion:

1) Accuracy and MAE compare to other model this model is high accuracy than other and low MAE.

2) Around 22.95% MAPE implies the model is about 77.05% accurate in predicting the next 10 observations.

3) The performance indices such ask fold cross-validation MAE, and RMSE where found satisfactory.

3) Mean absolute error is high 36.3465 we want MAE approaches to zero, there for forecasted and actual has error i.e. there is strongly deviated. As similarly for RMSE but is less compare to other models

2) Time series analysis for $PM_{2.5}$

Time-series analysis-ARIMA is used to forecast the PM_{10}. As discussed in section 5.4.1 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. Table 5.11 shows the obtained ARIMA model coefficients.

Table 5.18ARIMA model coefficients

Model Coefficients	Values
Auto Regressive Model ($\boldsymbol{p}$)	2
Differencing Model ($\boldsymbol{d}$)	1
Moving Average Model ($\boldsymbol{q}$)	1

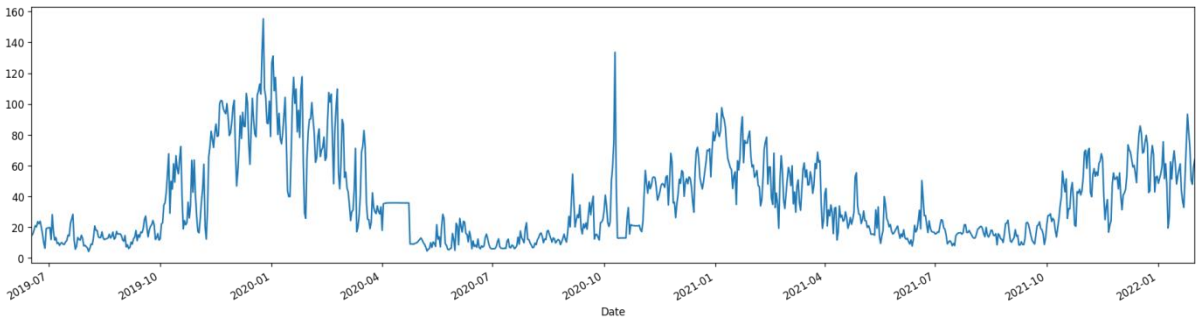


Figure 5.22Validation of Forecasting of PM2.5 using ARIMA model with test data

Figure 5.14presents the comparison of the actual value and the forecasted value obtained by the ARIMA model with unseen test data. The obtained forecasted values present in a 95 % confidence interval zone. The quantitative performance evaluation indices of the ARIMA model for the test data set are calculated and presented in Table 5.12

Table 5.19 Quantitative Performance indices of ARIMA model

Performance Indices	Test Data
CORR	0.8515
MAE	11.039
RMSE	14.009
MAPE	0.2272
MPE	0.2175
Average Accuracy of Test Data	77.28%

Conclusion:

1) Around 22.72% MAPE implies the model is about 77.28% accurate in predicting the next 5 observations. Over the entire model is good fit.

2) The performance indices such ask-fold cross-validation MAE, and RMSE where found satisfactory.

3) Mean absolute error is low 11.039 we want MAE approaches to zero but MAE for this model is less compare to other model we observe, there for forecasted and actual has less error i.e. there is weakly deviated. As similarly for RMSE

3) Time series Analysis for AQI

Similarly as PM_{10} we find the ARIMA model and forecasted AQI for station 6(table 3.1)

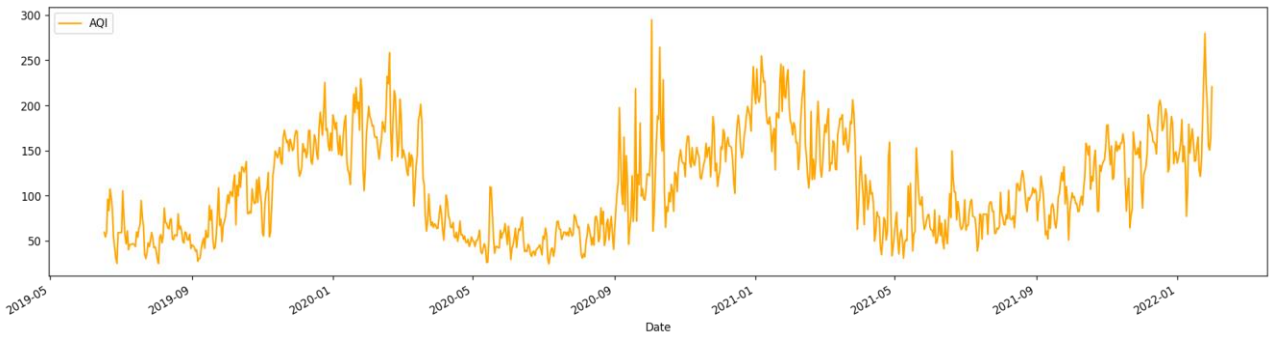


Figure 5.23 Time series plot for AQI

In time series plot, we observe that daily AQI data is present of some seasonal component and absent of trend component. To checking a seasonality and stationary we also plot Autocorrelation plot.

Time-series analysis-ARIMA is used to forecast the AQI. As discussed in section 5.4.1 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. Table 5.4.3.1 shows the obtained ARIMA model coefficients.

Table 5.20ARIMA model coefficients

Model Coefficients	Values
Auto Regressive Model ($\boldsymbol{p}$)	3
Differencing Model ($\boldsymbol{d}$)	1
Moving Average Model ($\boldsymbol{q}$)	3

Summary for ARIMA (3, 1, 3)

ARIMA Model Results

Dep. Variable:	D.AQI	No. Observations:	918
Model:	ARIMA(3, 1, 3)	Log Likelihood	-4195.167
Method:	css-mle	S.D. of innovations	23.346
Date:	Sun, 29 May 2022	AIC	8406.333
Time:	18:43:49	BIC	8444.911
Sample:	1	HQIC	8421.056

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.1229	0.185	0.663	0.507	-0.240	0.486
ar.L1.D.AQI	0.1494	0.050	3.002	0.003	0.052	0.247
ar.L2.D.AQI	-0.7245	0.031	-23.263	0.000	-0.786	-0.663
ar.L3.D.AQI	0.5322	0.037	14.198	0.000	0.459	0.606
ma.L1.D.AQI	-0.5308	0.044	-12.116	0.000	-0.617	-0.445
ma.L2.D.AQI	0.6107	0.045	13.678	0.000	0.523	0.698
ma.L3.D.AQI	-0.8310	0.028	-29.313	0.000	-0.887	-0.775

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	-0.2146	-1.0017j	1.0244	-0.2836
AR.2	-0.2146	+1.0017j	1.0244	0.2836
AR.3	1.7907	-0.0000j	1.7907	-0.0000
MA.1	-0.1934	-1.0175j	1.0357	-0.2799
MA.2	-0.1934	+1.0175j	1.0357	0.2799
MA.3	1.1218	-0.0000j	1.1218	-0.000

Conclusion:

The model summary reveals a lot of information.

1) Notice here the coefficient of the AR3 P-value i.e. P>|Z| column the AR3 is significant in the model, similarly the MA3 term is also significant in the model, therefore the model is ARIMA (3, 1, 3) Autoregressive term p=3, difference d=1, and Moving average q=3.

2) Root of AR3 polynomial is greater than 1 so the root of AR polynomial is outside the unit circle i.e. the model is causal stationary model.

3) Root of MA3 polynomial is greater than 1 so the root of MA polynomial is outside the unit circle i.e. the model is invertible stationary model.

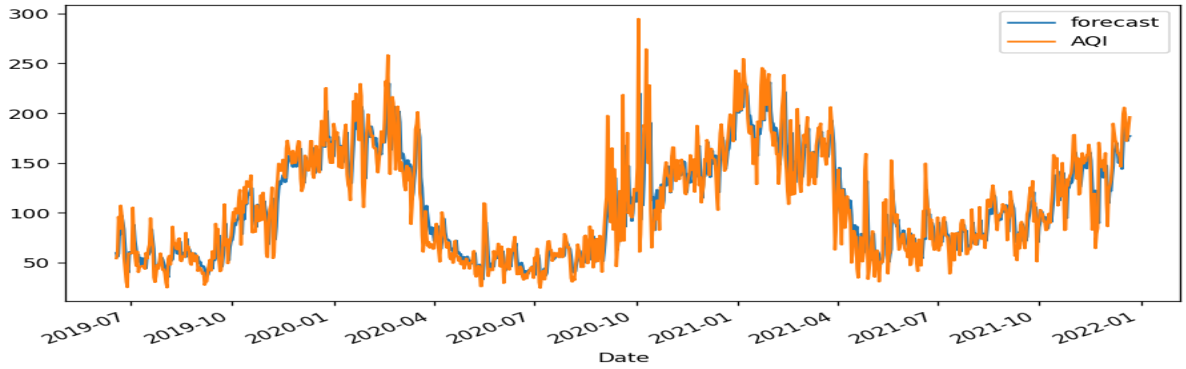


Figure 5.24Validation of Forecasting of AQI using ARIMA model

Table 5.21 Forecasted AQI for station 6

Date	Actual Value	Predicted Value
22/12/2021	196.0867	184.6538
23/12/2021	188.86	185.3563
24/12/2021	126.2133	183.7993
25/12/2021	129.7267	177.1018
26/12/2021	168.36	177.7309
27/12/2021	187.5267	181.977
28/12/2021	180.2	178.7198
29/12/2021	135.42	175.6196
30/12/2021	147.28	179.9039
31/12/2021	148.3067	181.1851

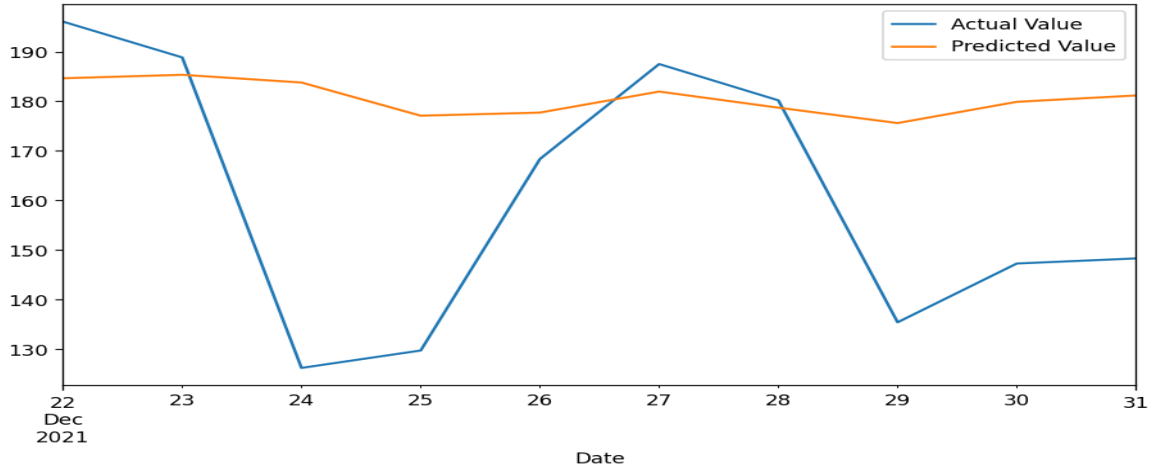


Figure 5.25Validation of Forecasting of AQI using ARIMA model with test data

Table 5.22 Quantitative Performance indices of ARIMA model

Performance Indices	Test Data
MAE	0.4779
RMSE	30.96
MAPE	0.1731
MPE	0.1502
Average Accuracy of Test Data	82.69%

Conclusion:

The prediction and forecasting of AQI comprehend people and the environment from adverse health conditions because of poor AQI. The ARIMA model is formulated and used for the forecasting of AQI. In the first place, the ARIMA model is trained with the help of training data and assessed with test data.

The performance indices for the ARIMA model with test data were found to be acceptable

We observe that the model for accuracy of $PM_{2.5}$, PM_{10} and AQI nearly same see in following table.

Table 5.23 Table Model and Accuracy comparison

Time series	Model	Accuracy (%)
$PM_{2.5}$	ARIMA (2, 1, 1)	77.28
PM_{10}	ARIMA (3, 1, 3)	77.05
AQI	ARIMA (3,1, 3)	82.69

5.7 Overall Conclusion

1. Mumbai city has 17 stations for calculation of Air pollution.
2. In data, several types of pollutants are available, but we usedonly pollutants for calculation of AQI.
3. We observe that the PM_{10} and PM2.5 are most Harmful and increasing pollutant.
4. We see $PM_{2.5}$ is in range 0-30, 31-60 and 61-90 as per Table 2.2 conclude that the $PM_{2.5}$ concentration for Mumbai is Good, Satisfactory and moderate.
5. Similar for PM_{10} is in range 0-50, 51-100 and 101-200 i.e. conclude that the PM_{10} concentration for Mumbai is Good, Satisfactory and moderate.
6. Using Table 5.3 we conclude that the PM10 has maximum contribution formation of AQIi.e. nearly 80%.
7. Overall AQI for Mumbai city is good (25.92%) satisfactory (31.83%) and moderate (38.89%) in nature.
8. Using Table 5.4 we classify the AQI for Mumbai city in Satisfactory and Moderate in nature and using Table 5.5 we calculate weighted average AQI for Mumbai city by treating population as weight to calculate weighted average AQI which is equal to 89.98 i.e. the Mumbai is satisfactory AQI as our data.
9. Overall the ARIMA Model is good fit the time series data, model accuracy is in 70-85 is overall good compare to other model.
10. The accuracy of the forecasted model is good and some values the actual value and forecasted values are deviated from other.
11. The MLR model is used to predict the AQI. Firstly, the correlation between air pollutants was found, and then the MLR model was trained using the training data set and validated with the unseen test data. The performance indices such R2, MAE, CORR, MAPE and RMSE for both training and test data were found to be satisfactory.
12. Secondly The ARIMA model is formulated and used for the forecasting of AQI. In the first place, the ARIMA model is trained with the help of training data and assessed with test data. The performance indices for the ARIMA model with test data were found to be acceptable.

APPENDIX

Appendix 1 calculate AQI using Excel

Sr. no.	City	Station	Date	FromDate	PM2.5	PM10	NO	NO2	NH3	SO2	CO
1	Mumbai-01	BandraKurlaComplex,Mumbai-IITM	10/11/2020	10-11-202000:00	28.85	289.36	70.61	58.47	174.3	3.01	0.2
2	Mumbai-01	BandraKurlaComplex,Mumbai-IITM	11/11/2020	11-11-202000:00	90.91	266.66	59.7	51.01	153.5	3.24	0.35
3	Mumbai-01	BandraKurlaComplex,Mumbai-IITM	12/11/2020	12-11-202000:00	26.19	237.63	76.37	92.88	157.16	8.28	1.11
4	Mumbai-01	BandraKurlaComplex,Mumbai-IITM	13/11/2020	13-11-202000:00	27.24	243.11	86.95	45.14	185.24	3.26	0.76
5	Mumbai-01	BandraKurlaComplex,Mumbai-IITM	14/11/2020	14-11-202000:00	27.93	180.92	62.56	68.73	131.72	7.24	1.8
….	….	….	….	….	….	….	….	….	….	….	….
12733	Mumbai-17	VasaiWest,Mumbai-MPCB	27/12/2021	27-12-202100:00	76.68	244.83	9.71	34.67	8.92	20.31	1.89
12734	Mumbai-17	VasaiWest,Mumbai-MPCB	28/12/2021	28-12-202100:00	86.49	234.8	9.14	30.31	9.43	20.84	1.84
12735	Mumbai-17	VasaiWest,Mumbai-MPCB	29/12/2021	29-12-202100:00	45.57	127.84	5.26	23.24	7.95	20.63	1.56
12736	Mumbai-17	VasaiWest,Mumbai-MPCB	30/12/2021	30-12-202100:00		183.29	9.45	31.19	9.06	20.34	1.9
12737	Mumbai-17	VasaiWest,Mumbai-MPCB	31/12/2021	31-12-202100:00		172.16	6.24	24.06	9.27	20.39	1.69

I1 PM2.5	I2 PM10	I3 NO	I4 NO2	I5 NH3	I6 SO2	I7 CO	I1 PM 2.5	I2 PM10	I3 NO	I4 NO2	I5 NH3	I6 SO2	I7 CO	AQI
28.8 5	239.36	70.6 1	58.47	149.53 33	3.01	0.2	1	1	1	1	1	1	1	239.36
90.91	216.66	59.7	51.01	135.66 67	3.24	0.35	1	1	1	1	1	1	1	216.66
26.19	191.753 3	76.3 7	92.88	138.10 67	8.28	1.11	1	1	1	1	1	1	1	191.753 3
27.24	195.406 7	86.9 5	45.14	156.82 67	3.26	0.76	1	1	1	1	1	1	1	195.406 7
27.93	153.946 7	62.5 6	68.73	121.14 67	7.24	1.8	1	1	1	1	1	1	1	153.946 7
….	….	….	….	….	….	….	….	….	….	….	….	….	….	….
76.68	196.553 3	9.71	34.67	8.92	20.31	1.89	1	1	1	1	1	1	1	196.553 3
86.49	189.866 7	9.14	30.31	9.43	20.84	1.84	1	1	1	1	1	1	1	189.866 7
45.57	118.56	5.26	23.24	7.95	20.63	1.56	1	1	1	1	1	1	1	118.56
0	155.526 7	9.45	31.19	9.06	20.34	1.9	0	1	1	1	1	1	1	155.526 7
0	148.106 7	6.24	24.06	9.27	20.39	1.69	0	1	1	1	1	1	1	148.106 7

Appendix 2 data for Chhatrapati Shivaji International Airport (T2), Mumbai-MPCB

sr.no.	Date	FromDate	ToDate	PM2.5	PM10	NO	NO2	NH3	SO2	CO	Benzene	Ozone	AQI
1	17-06-2019	17-06-2019 00:00	18-06-2019 00:00	14.61	59.44	18.07	16	5.85	7.36	0.29	16.6	1.36	59.44
2	18-06-2019	18-06-2019 00:00	19-06-2019 00:00	15.68	54.09	26.5	21.58	6.43	7.6	0.4	14.12	1.59	54.09
3	19-06-2019	19-06-2019 00:00	20-06-2019 00:00	18.28	58.87	39.33	23.05	12.19	8.33	0.48	9.24	3.43	58.87
4	20-06-2019	20-06-2019 00:00	21-06-2019 00:00	21.07	96.02	45.62	19.11	8.24	9.03	0.44	8.59	3.21	96.02
5	21-06-2019	21-06-201900:00	22-06-2019 00:00	20.27	83.51	37.44	20.19	6.05	8.93	0.41	9.41	2.68	83.51
….	….	….	….	….	….	….	….	….	….	….	….	….	….
956	27-01-2022	27-01-2022 00:00	28-01-2022 00:00	71.64	250.19	76.18	45.46	25.88	13.7	1.28	4.74	38.4	200.19
957	28-01-2022	28-01-2022 00:00	29-01-2022 00:00	50.62	182.83	55.17	40.47	25.89	15.19	0.94	3.55	56.37	155.22
958	29-01-2022	29-01-2022 00:00	30-01-2022 00:00	47.95	175.7	57.13	48.6	26.41	13.13	0.96	2.04	52.69	150.4667
959	30-01-2022	30-01-2022 00:00	31-01-2022 00:00	59.23	196.21	111.2	48.2	24.58	12.54	1.35	3.68	39.18	164.14
960	31-01-2022	31-01-2022 00:00	01-02-2022 00:00	63.91	217.11	270.55	44.02	1.67	17.13	2.86	0.8	3.59	220.55

References

- Practical Time Analysis: master time series data processing, visualization, and modeling using Python (2017) , Avishek Pal, PKS Prakash.
- Data Science Projects with Python (2019), Stephen Klosterman.
- Time Series Analysis Forecasting and Control, 5th Edition, George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, Greta M. Ljung ,Wiley
- Prediction and Forecasting of Air Quality Index in Chennai using Regression and ARIMA time series models, Geetha Mani*, Joshi Kumar Viswanadhapalli* and Albert Alexander Stonier** *School of Electrical Engineering, Vellore Institute of Technology, Vellore *Department of Electrical and Electronics Engineering, Kongu Engineering College, India
- NATIONAL AIR QUALITY INDEX (FINAL REPORT 2014-15),CPCB
- Applied Multivariate Statistical Analysis, 6thEdition, RICHARD A. JOHNSON, DEAN W. WICHERN (for PCA)
- A Machine Learning Approach to Predict Air Quality in California (Research Article) published by Hindawi on 4 August 2020.

WEBSITE:

www.CPCB.nic.in

www.machinelearningplus.com

SOFTWARE:

- MS-EXCEL
- Python
- Power BI