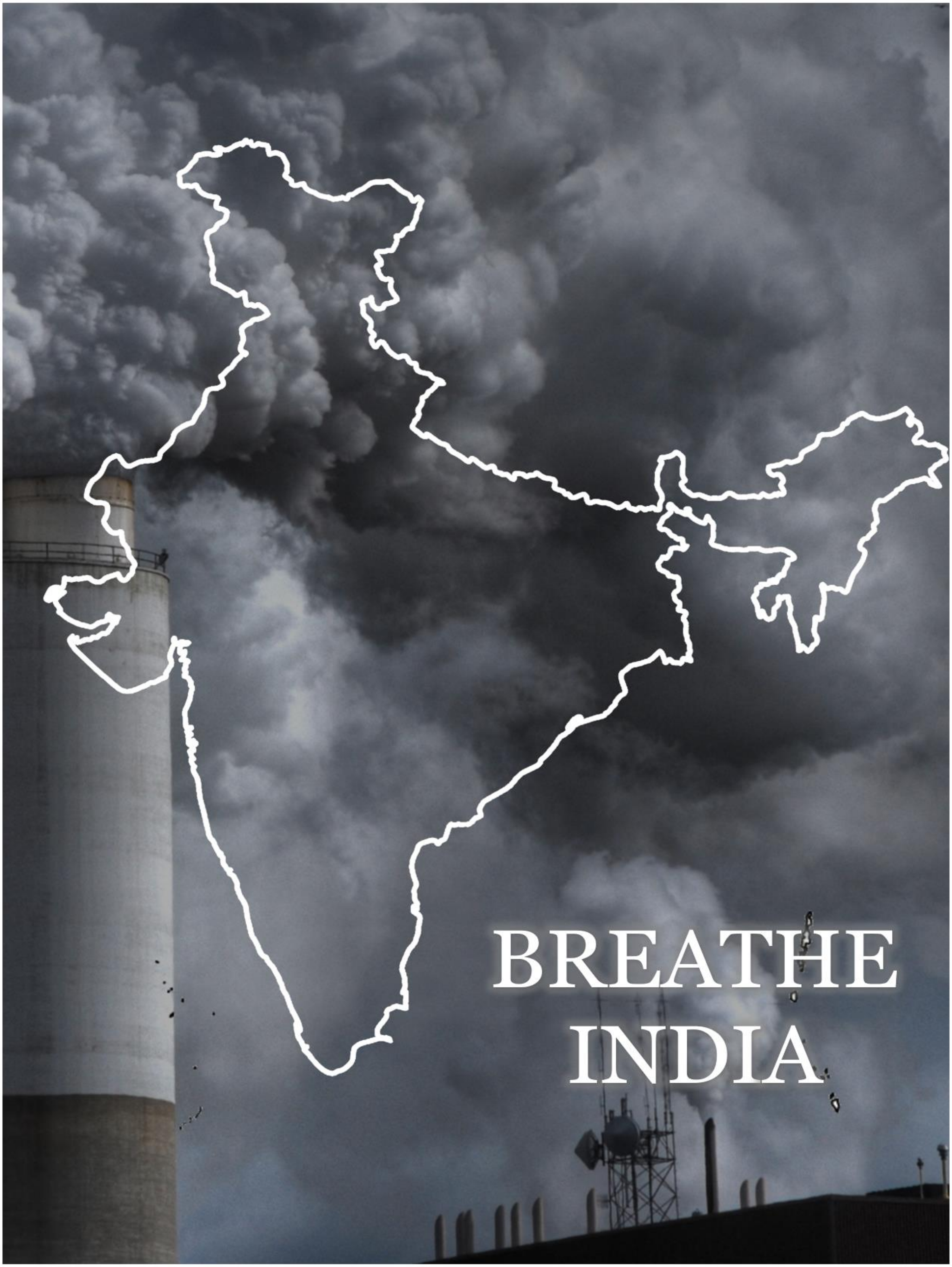

BREATHE
INDIA

**UNIVERSITY OF MUMBAI**

**DEPARTMENT OF STATISTICS**

**Vidyanagari, Mumbai-400098.**

**CERTIFICATE**

This is to certify that the following students of M.Sc. Part-II have successfully completed the project **“BREATHE INDIA”** during the academic year 2020-2021. This work has been done independently to the best of our knowledge and Awareness.

Students involved in this academic project group are:

- Abhishek Bandivadekar. (Roll no. 04)
- Nikita Bansode. (Roll no. 05)
- Chetan Borghare. (Roll no. 09)
- Sunita Choudhary. (Roll no. 10)
- Vinayak Gaikar. (Roll no. 13)
- Ram Jadhav. (Roll no. 20)
- Rutija Parab. (Roll no. 37)
- Rajesh Patil. (Roll no. 38)

**Dr. Alok D. Dabade.** — **Dr. Mrs. Vaijayanti U. Dixit.**

***(Guide and Mentor)*** — ***(Head of the Department, Statistics)***

# ACKNOWLEDGEMENT


We are extremely grateful to our Head of Statistics Department **Dr. Mrs. Vaijayanti Dixit** Madam for providing us with this opportunity of widening our horizons of knowledge and diving deep into the ocean of Statistics through this project.

We are immensely indebted to our guide and mentor **Dr. Alok Dabade** Sir who gave us his extremely valuable suggestions and ideas throughout the duration of our project. We are thankful to him for his time and effort.

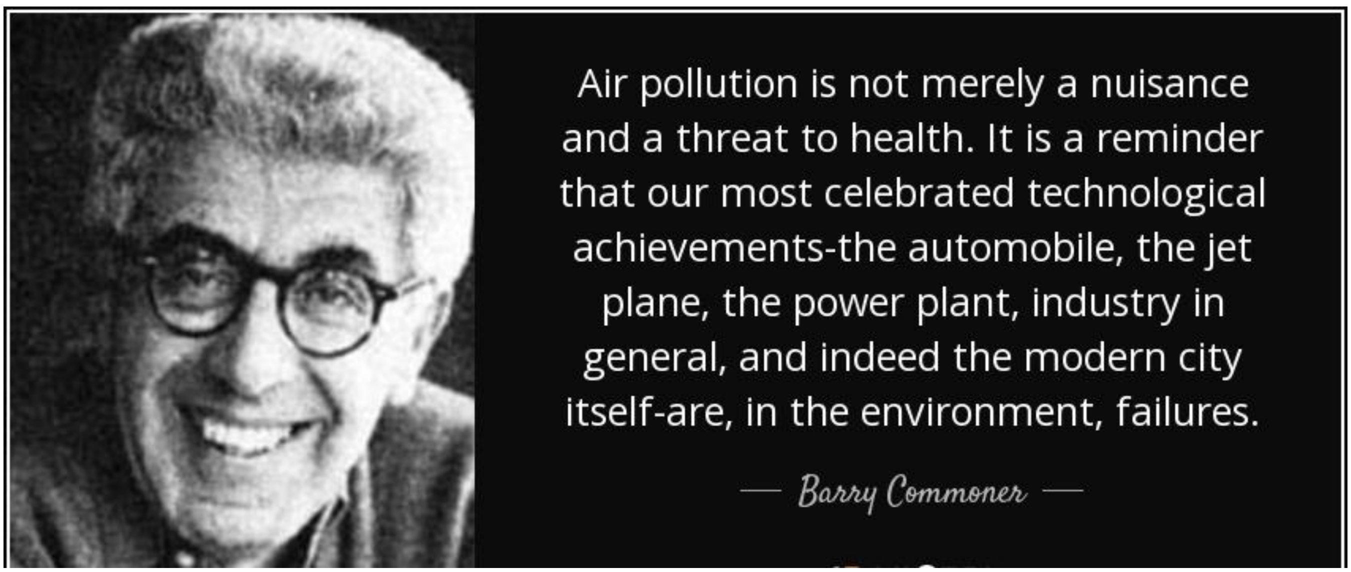
We express our gratitude towards our Professor **Priyesh Tiwari** Sir & Dr. **Leela Subramaniyam** Madam for their support and guidance which helped us complete our project duly. We also thank other Professors of the Department of Statistics for providing us necessary knowledge without which this project could have not been complete.

Last but not the least, we would like to give our special thanks to all involved in this project including the team members and all the external supporters as without their inspiration and valuable suggestions, it would have not been possible to develop the project with ease and within prescribed time.

# INDEX

*"Air pollution may not always be visible, but it can be deadly. It is an invisible killer that lurks all around us, preying on young and old. It is causing deaths from heart attack, strokes, lung disease and cancer."*

*-WHO*



# Introduction

Clean air is the foremost requirement to sustain healthy lives of humankind and those of the supporting ecosystems which in return affect the human wellbeing. Due to the rapid growth of economy and fossil fuel consumption and lack of emission controls, we have experienced substantially elevated concentrations of air pollutants, which not only degrade regional air quality, but also exert significant impacts on public health and global climate.

**Air pollution** is a mixture of solid particles and gases in the air. Car emissions, chemicals from factories, dust, pollen and mold spores may be suspended as particles. Ozone, a gas, is a major part of air pollution in cities. Air pollution is a type of environmental pollution that affects the air and is usually caused by smoke or other harmful gases, Nitrogen oxide ($NO_x$), Particulate matter (PM), Carbon dioxide ($CO_2$), Sulphur oxide ($SO_x$), dioxins and furans, etc. mainly oxides of carbon, Sulphur and nitrogen. In other words, air pollution is the contamination of air due to the presence or introduction of a substance which has a poisonous effect.

**COVID-19** is considered as one of the major disasters, which has impacted the whole world. Wuhan city, capital of Hubei province of China, faced the first outbreak of this COVID-19 during December 2019 and all nations of the world are affected by COVID-19 in a gradual manner. After China, most South Asian Countries like Japan, South Korea, and others are affected by the cross-border travels. The return of Chinese workers spread COVID-19 in Italy. The Government of India issued an advisory for travelers from China during early January and also started screening the travelers from China. In response to the global COVID-19 pandemic, the Indian Prime Minister announced Janata (people's) curfew on 22 March 2020 from 7 am until 9 pm. Soon after, the Government of India announced a complete nationwide lockdown, from 24 March 2020 for 21 days (14 April 2020) then second phase from 14 April 2020 to 3 May 2020 and further it extended in other cities having high number of covid cases. All the domestic and international flights, trains, and vehicular transport except for non-essential purposes were stopped and banned. Such lockdown was unique in India; total lockdown was not seen in any other countries. The Northern parts of India are subjected to poor air quality and atmospheric pollution, mainly due to emissions from vehicles, industry, brick kilns, coal-based power plants, and crop residue burning. For instance, New Delhi, capital of India, suffers with sustained poor air quality where pollution levels are higher compared with Beijing. In recent past, the Delhi Government conducted experiments of permitting odd or even licensed vehicles on the road to curb the pollution level (like Beijing). However, such experiments have generally not helped or improved the air quality of Delhi. Recently, someone carried out an analysis of PM2.5 data in a number of Chinese cities, Beijing, Shanghai, Guangzhou, and Wuhan during COVID-19, and found a pronounced reduction in air pollution attributed to the reduction of emissions in transportation and industrial sectors. They found 20–30% reduction in emission of NO2 in China, Spain, France, Italy, and the USA due to lockdown. During complete lockdown in India, roads were deserted without any vehicle except the emergency vehicles.

# WHAT IS AQI?

The **Air Quality Index** (**AQI**) is an index for reporting air quality on a daily basis. Air quality index is usually the standardized formula to indicate how polluted the air currently is and is also used for simplified public information and data interpretation. AQI has the scale of about 0-500. Higher the AQI, higher is the pollution rate.

## AQI CALCULATION

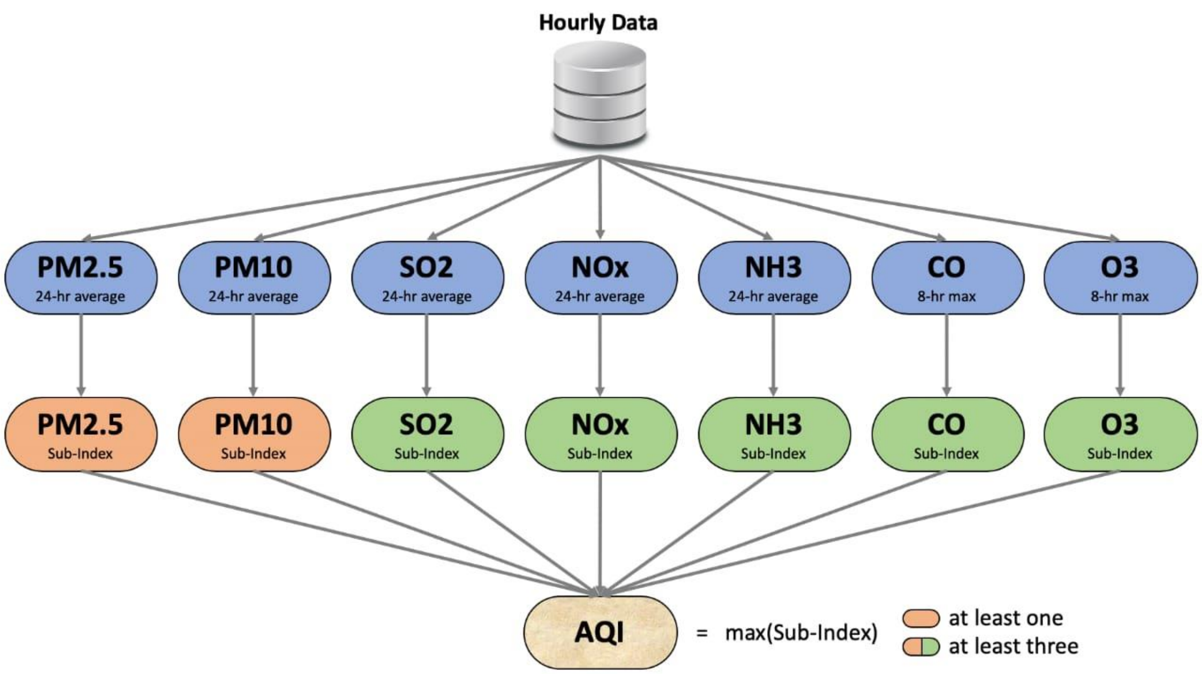
The Air Quality Index is based on measurement of particulate matter (PM2.5 and PM10), Ozone (O3), Nitrogen Dioxide (NO2), Sulfur Dioxide (SO2) and Carbon Monoxide (CO) emissions. Most of the stations on the map are monitoring both PM2.5 and PM10 data, but there are few exceptions where only PM10 is available. There is no theoretical upper value of AQI but it is rare to find values above 1000.

These classifications are done for the calculated Air Quality Index values. This is done as, if the Air Quality Index value is between 0-50 then it is classified as good, 51-100 as satisfactory, 101-200 as moderate, 201-300 as poor, 301-400 is classified as very poor and above 401 considered as severe. There are also color barriers for this classification. Each type has separate color which shows the quality easily. Green represents good, sap green as satisfactory, yellow as moderate, orange as poor, red as very poor and maroon as the severe. This is for the easier understanding.

The pre-defined buckets of AQI are as follows:

| | | | |
|---|---|---|---|
| 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 |

## FORMULA FOR CALCULATING AQI




The AQI calculation uses 7 measures: PM2.5, PM10, SO2, NOx, NH3, CO and O3.For PM2.5, PM10, SO2, NOx and NH3 the average value in last 24-hrs is used with the condition of having at least 16 values. For CO and O3 the maximum value in last 8-hrs is used. Each measure is converted into a Sub-Index based on pre-defined groups.

Sometimes measures are not available due to lack of measuring or lack of required data points.

Final AQI is the maximum Sub-Index with the condition that at least one of PM2.5 and PM10 should be available and at least three out of the seven should be available. To get AQI at day level, the AQI values are averaged over the hours of the day.

# Why this topic?

Clean air is the basic amenity when it comes to healthy living for mankind. Today poor air quality is the main reason for several acute health diseases. Poor air quality brings many health problems like cardiovascular disease and respiratory problems like asthma, allergies, pneumonia and bronchitis, etc. It is essential to know the air quality of our locality, city and nation to assess its impact on our health.

Covid-19 has affected the world on a huge scale and it continues to spread its claws. Almost every country had imposed lockdown in the year 2020 to mitigate the spread of this virus. India imposed lockdown on 22$^{nd}$ March, 2020. The lockdown has led to colossal economic loss to India; however, it has come as a respite to the environment. Utilizing the Air Quality Index data recorded before and during this adverse time, our project is aimed to determine the impact of lockdown on Air Quality Index (AQI) in various cities of India, also the factors that affect AQI and predict the future values of AQI for some of those cities.

# Objectives:

- ❖ To study Air Quality Index before lockdown and during lockdown.
- ❖ To forecast future values for Air Quality Index for four cities.
- ❖ To find association between Air Quality Index and other factors such as temperature, humidity, dew point, wind speed and pressure.

## SOFTWARES USED:

- Minitab
- R Software
- Python
- MS-Excel
- SPSS (Statistical Package for the Social Sciences)

# DATA PREPARATION

| VARIABLES | DESCRIPTION |
|---|---|
| **AQI** | The air quality index (AQI) is an index for reporting air quality on a daily basis. It is a measure of how air pollution affects one's health within a short time period. The purpose of the AQI is to help people know how the local air quality impacts their health.<br>It is a continuous variable. |
| **AQI Bucket** | AQI Category or Bucket is used to group the AQI values into six categories based on the value namely:<br>Good (0–50), Satisfactory (51–100), Moderate (101–200), Poor (201–300), Very Poor (301–400), Severe (401–500).<br>It is a categorical variable. |
| **Average Temperature** | Temperature is continuous variable as it does have fractional value too. The average temperature of the air as indicated by a properly exposed thermometer during a given time period, usually a day, a month, or a year.<br>For climatological tables, the mean temperature is generally calculated for each month and for the year. |
| **Average Humidity** | Humidity is the concentration of water vapor present in the air. Humidity indicates the likelihood for precipitation, dew, or fog to be present.<br>It is a continuous variable. |
| **Average Windspeed** | The "wind speed" reported in each observation is an average speed for the most recent two-minute period prior to the observation time.<br>This is also considered the & "sustained wind" for routine surface observations.<br>It is a continuous variable. |
| **Average Dew Point** | The dew point is the temperature to which air must be cooled to become saturated with water vapor.<br>When cooled further, the airborne water vapor will condense to form liquid water (dew). When, air cools to its dew point through contact with a surface that is colder than the air, water will condense on the surface.<br>It is a continuous variable. |
| **Average Pressure** | It is the force exerted on a surface by the air above it as gravity pulls it to Earth. It is commonly measured with a barometer.<br>It is also a continuous variable. |

# DATA EXTRACTION

The data is taken from open source websites viz kaggle.com and wunderground.com. From Kaggle we get data of 11 cities namely Ahmedabad, Jaipur, Delhi, Mumbai, Chennai, Bengaluru, Gurgaon, Hyderabad, Brajrajnagar, Amritsar and Thiruvananthapuram. The data has daily observations for various Air pollutants, AQI, AQI bucket, Temperature, Humidity, Dew points, Wind speed and Pressure from 1$^{st}$ January, 2018 to 30$^{th}$ June, 2020 for each above mentioned cities. Because of time factor, we have chosen only 4 cities for our analysis, namely, Jaipur, Delhi, Chennai, Hyderabad. The data is basically time series data.

# DATA VISUALIZATION

The overview of all data consisting all four cities:

## OVERVIEW

### Dataset Statistics

| | |
|---|---|
| **Number of Variables** | 20 |
| **Number of Rows** | 3647 |
| **Missing Cells** | 0 |
| **Missing Cells (%)** | 0.0% |
| **Duplicate Rows** | 0 |
| **Duplicate Rows (%)** | 0.0% |
| **Total Size in Memory** | 1.2 MB |
| **Average Row Size in Memory** | 330.8 B |

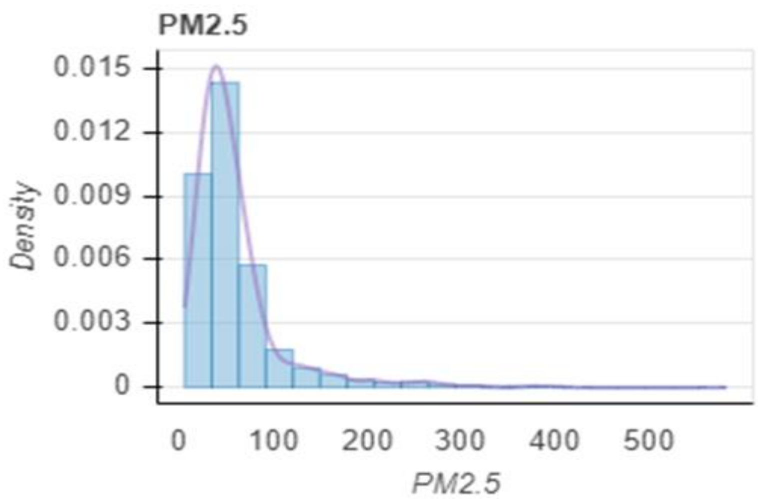
### Variable Types

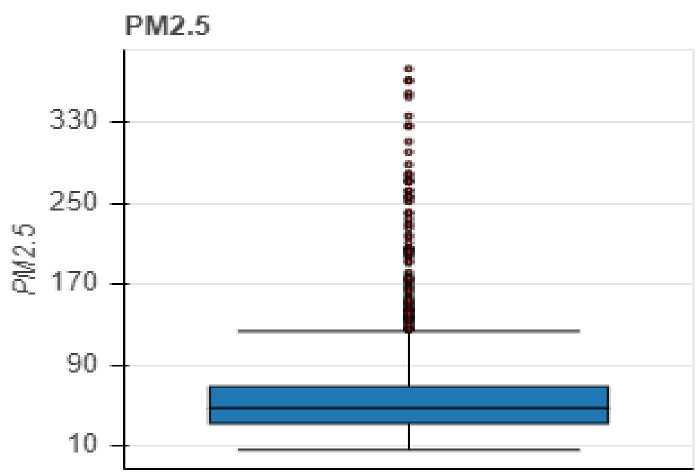
| | |
|---|---|
| **Categorical** | 3 |
| **Numerical** | 17 |

**PM2.5**

**Numerical**

| | |
|---|---|
| Distinct Count | 3028 |
| **Unique (%)** | 83.00% |
| **Missing** | 0 |
| **Minimum** | 6.24 |
| **5-th Percentile** | 17.193 |
| **Q1** | 32.735 |
| **Median** | 47.88 |
| **Q3** | 69.165 |
| **95-th Percentile** | 158.282 |
| **Maximum** | 582.28 |
| **Range** | 576.04 |
| **IQR** | 36.43 |
| **Mean** | 60.7374 |
| **Standard Deviation** | 50.5423 |
| **Variance** | 2554.52 |
| **Sum** | 221509.3 |
| **Skewness** | 3.1512 |
| **Kurtosis** | 13.9122 |
| **Coefficient of Variation** | 0.8321 |

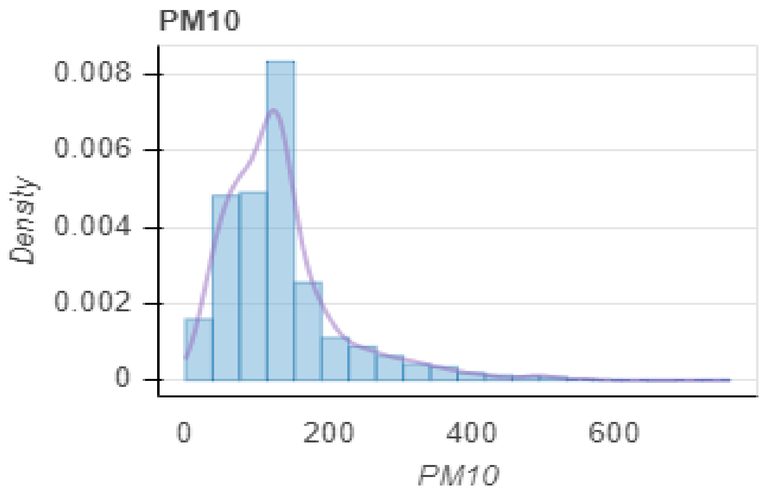


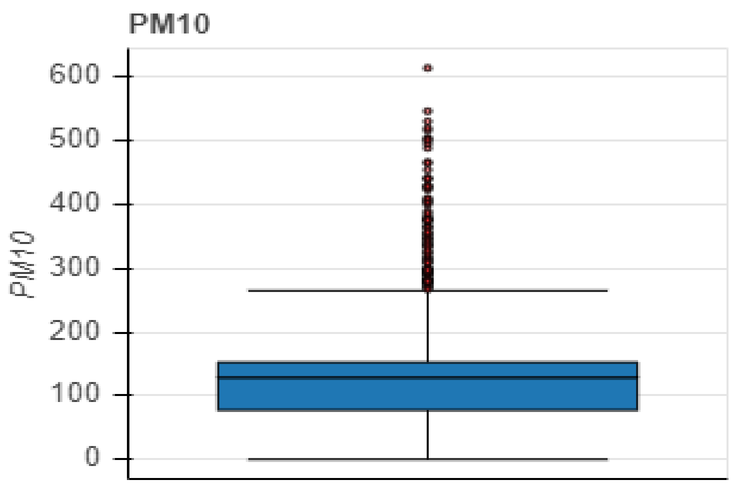



## PM10

**Numerical**

| | |
|---|---|
| Distinct Count | 2853 |
| **Unique (%)** | 78.20% |
| **Missing** | 0 |
| **Minimum** | 0.21 |
| **5-th Percentile** | 35.175 |
| **Q1** | 77.555 |
| **Median** | 128.66 |
| **Q3** | 152.525 |
| **95-th Percentile** | 311.86 |
| **Maximum** | 761.91 |
| **Range** | 761.7 |
| **IQR** | 74.97 |
| **Mean** | 133.1024 |
| **Standard Deviation** | 85.8704 |
| **Variance** | 7373.723 |
| **Sum** | 485424.6 |
| **Skewness** | 1.9782 |
| **Kurtosis** | 5.9381 |
| **Coefficient of Variation** | 0.6451 |

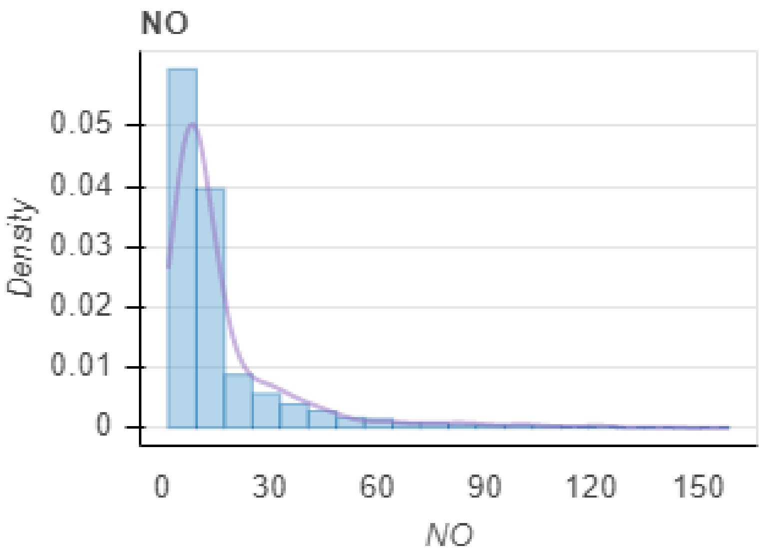


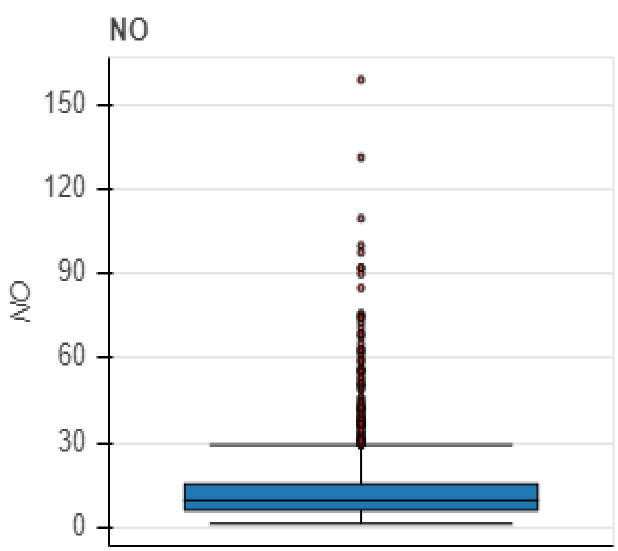



## NO

**Numerical**

| | |
|---|---|
| Distinct Count | 1948 |
| **Unique (%)** | 53.40% |
| **Missing** | 0 |
| Minimum | 1.46 |
| **5-th Percentile** | 3.223 |
| **Q1** | 6.2 |
| **Median** | 9.71 |
| **Q3** | 15.46 |
| **95-th Percentile** | 52.392 |
| **Maximum** | 158.63 |
| **Range** | 157.17 |
| **IQR** | 9.26 |
| **Mean** | 15.5739 |
| **Standard Deviation** | 18.0088 |
| **Variance** | 324.3173 |
| **Sum** | 56797.96 |
| **Skewness** | 3.2231 |
| **Kurtosis** | 12.7263 |
| **Coefficient of Variation** | 1.1563 |

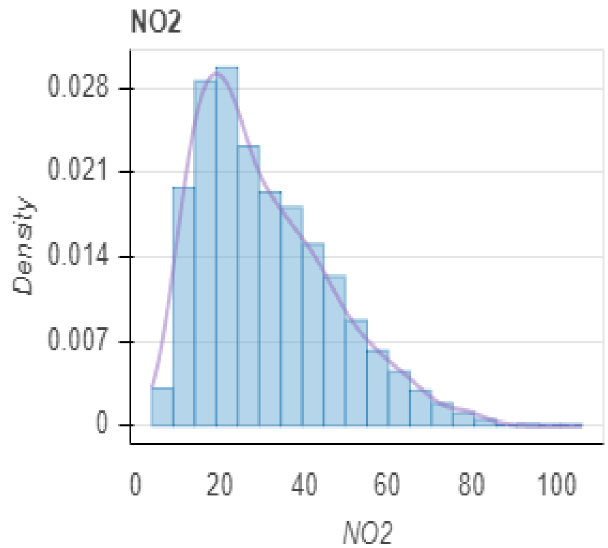


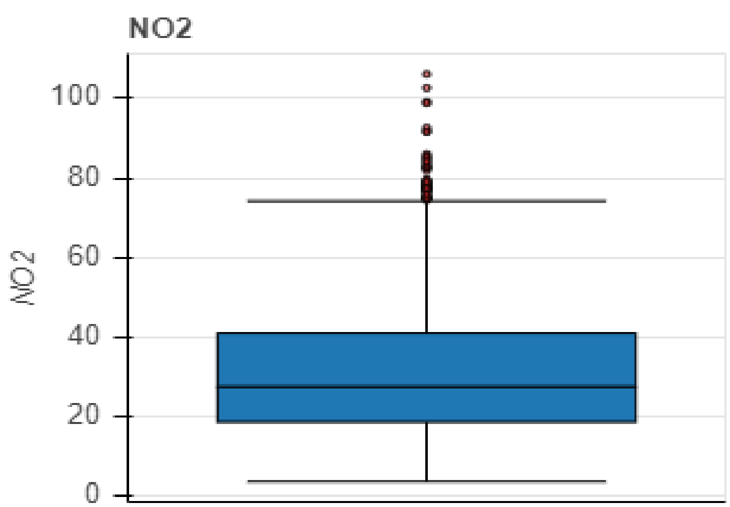



## NO2

**Numerical**

| | |
|---|---|
| Distinct Count | 2611 |
| **Unique (%)** | 71.60% |
| **Missing** | 0 |
| Minimum | 3.73 |
| **5-th Percentile** | 11.32 |
| **Q1** | 18.725 |
| **Median** | 27.59 |
| **Q3** | 40.985 |
| **95-th Percentile** | 61.354 |
| **Maximum** | 106.04 |
| **Range** | 102.31 |
| **IQR** | 22.26 |
| **Mean** | 31.0953 |
| **Standard Deviation** | 15.8966 |
| **Variance** | 252.7004 |
| **Sum** | 113404.4 |
| **Skewness** | 0.877 |
| **Kurtosis** | 0.5005 |
| **Coefficient of Variation** | 0.5112 |

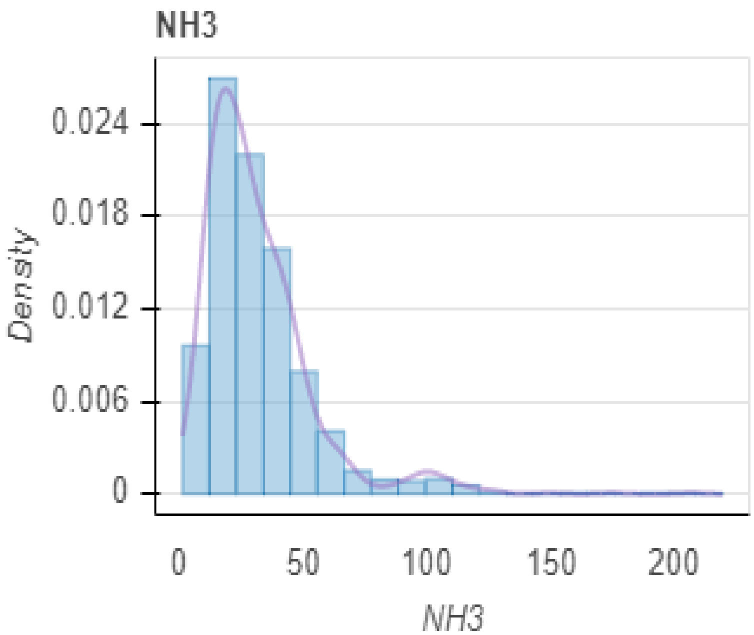


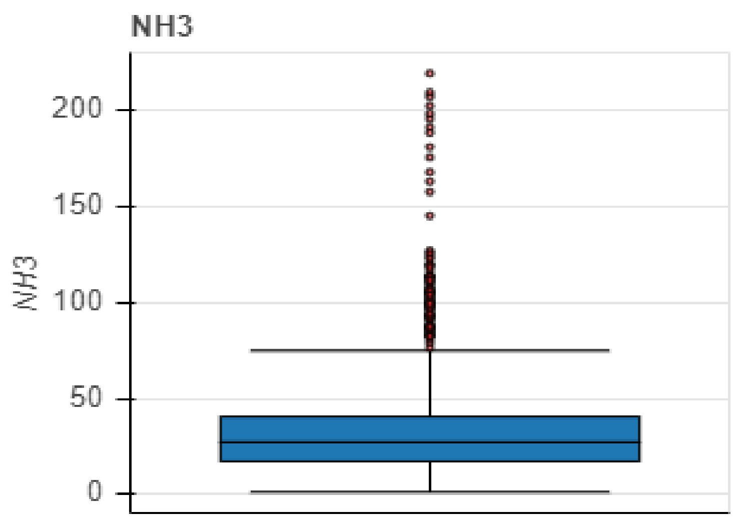



## NH3

**Numerical**

| | |
|---|---|
| Distinct Count | 2636 |
| **Unique (%)** | 72.30% |
| **Missing** | 0 |
| **Minimum** | 1.33 |
| **5-th Percentile** | 9.963 |
| **Q1** | 17.29 |
| **Median** | 27.4 |
| **Q3** | 40.5 |
| **95-th Percentile** | 70.528 |
| **Maximum** | 219.26 |
| **Range** | 217.93 |
| **IQR** | 23.21 |
| **Mean** | 32.3789 |
| **Standard Deviation** | 22.5604 |
| **Variance** | 508.9718 |
| **Sum** | 118085.8 |
| **Skewness** | 2.5081 |
| **Kurtosis** | 10.9783 |
| **Coefficient of Variation** | 0.6968 |

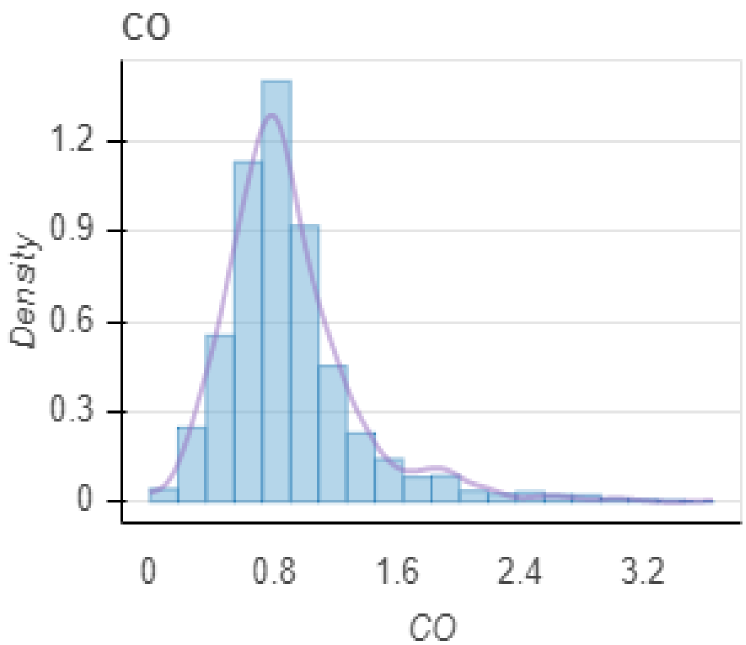


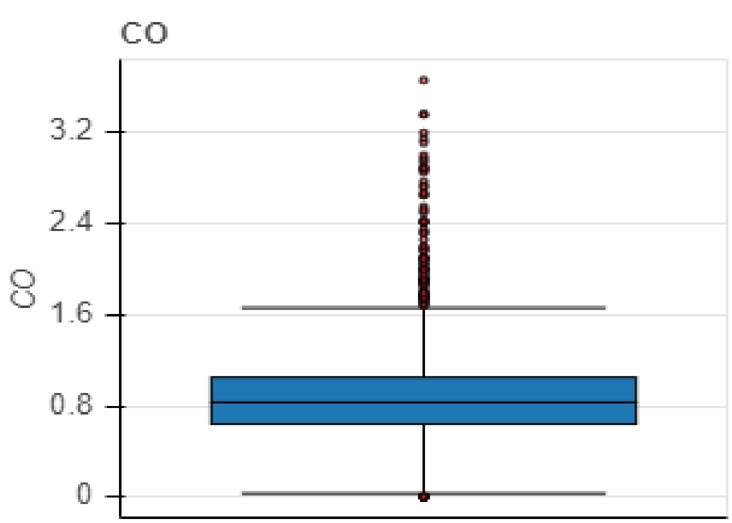



## CO

**Numerical**

| | |
|---|---|
| Distinct Count | 254 |
| **Unique (%)** | 7.00% |
| **Missing** | 0 |
| **Minimum** | 0 |
| **5-th Percentile** | 0.36 |
| **Q1** | 0.64 |
| **Median** | 0.83 |
| **Q3** | 1.05 |
| **95-th Percentile** | 1.787 |
| **Maximum** | 3.66 |
| **Range** | 3.66 |
| **IQR** | 0.41 |
| **Mean** | 0.9061 |
| **Standard Deviation** | 0.4508 |
| **Variance** | 0.2033 |
| **Sum** | 3304.55 |
| **Skewness** | 1.8664 |
| **Kurtosis** | 5.5914 |
| **Coefficient of Variation** | 0.4976 |

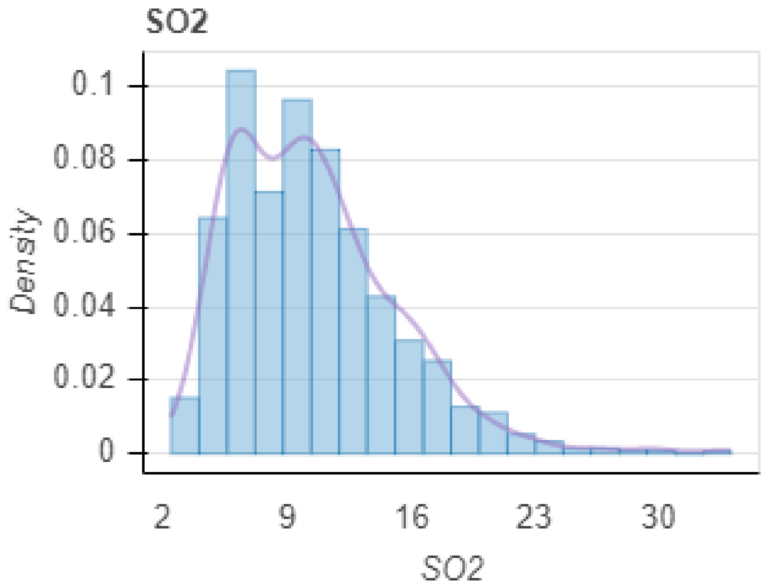


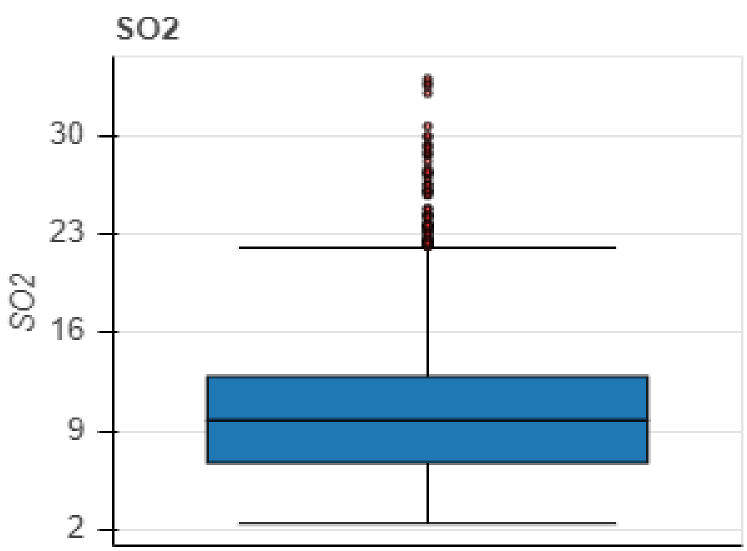



**SO2**

numerical

| | |
|---|---|
| Distinct Count | 1463 |
| **Unique (%)** | 40.10% |
| **Missing** | 0 |
| **Minimum** | 2.47 |
| **5-th Percentile** | 4.59 |
| **Q1** | 6.78 |
| **Median** | 9.83 |
| **Q3** | 12.91 |
| **95-th Percentile** | 19.06 |
| **Maximum** | 34.03 |
| **Range** | 31.56 |
| **IQR** | 6.13 |
| **Mean** | 10.4431 |
| **Standard Deviation** | 4.5842 |
| **Variance** | 21.0146 |
| **Sum** | 38085.83 |
| **Skewness** | 1.0082 |
| **Kurtosis** | 1.4352 |
| **Coefficient of Variation** | 0.439 |

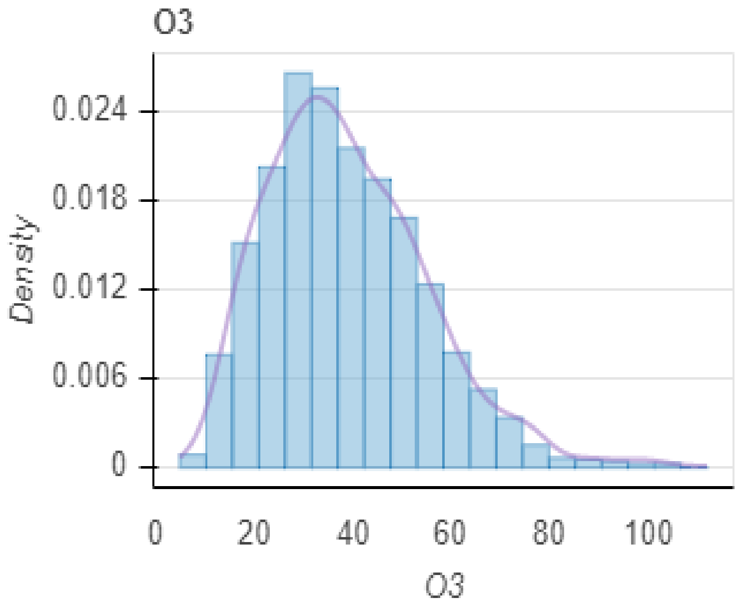


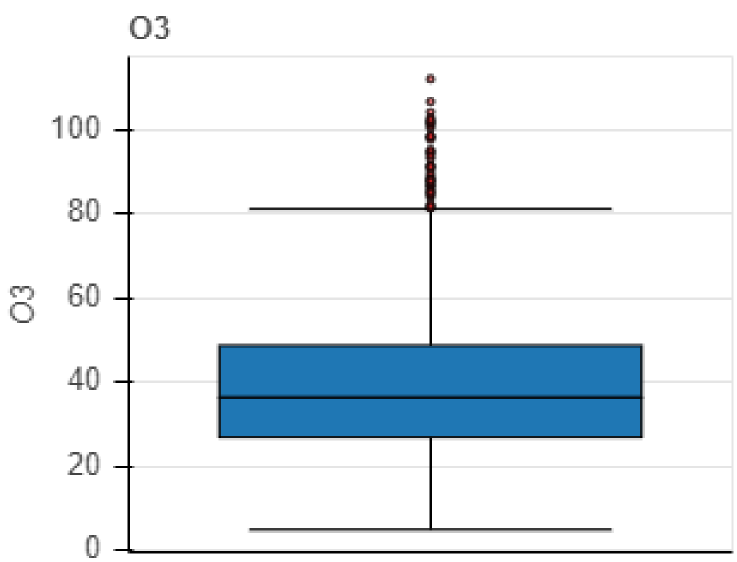



**O3**

`numerical`

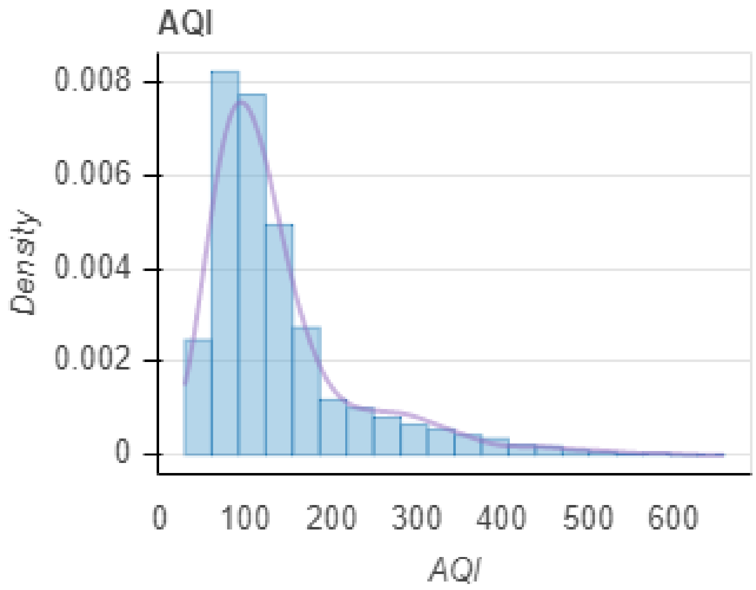
| | |
|---|---|
| Distinct Count | 2700 |
| **Unique (%)** | 74.00% |
| **Missing** | 0 |
| **Minimum** | 4.86 |
| **5-th Percentile** | 16.05 |
| **Q1** | 26.825 |
| **Median** | 36.37 |
| **Q3** | 48.6 |
| **95-th Percentile** | 66.401 |
| **Maximum** | 111.96 |
| **Range** | 107.1 |
| **IQR** | 21.775 |
| **Mean** | 38.459 |
| **Standard Deviation** | 15.7923 |
| **Variance** | 249.3982 |
| **Sum** | 140260 |
| **Skewness** | 0.7071 |
| **Kurtosis** | 0.6372 |
| **Coefficient of Variation** | 0.4106 |

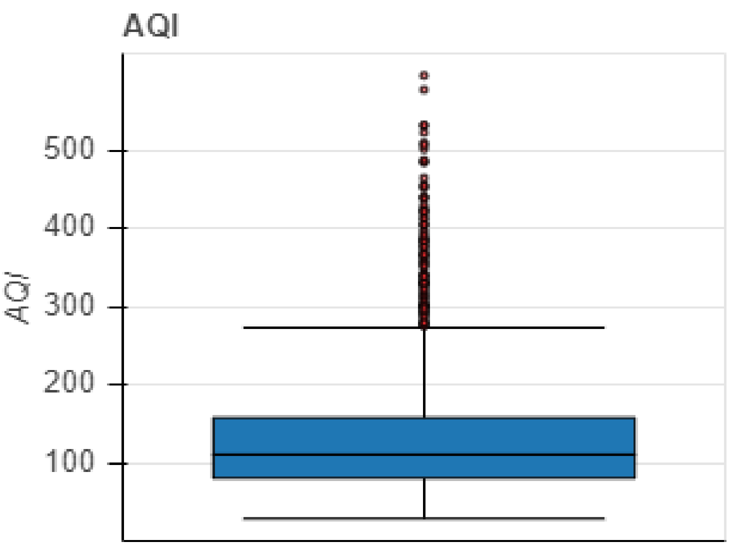


## AQI
**Numerical**

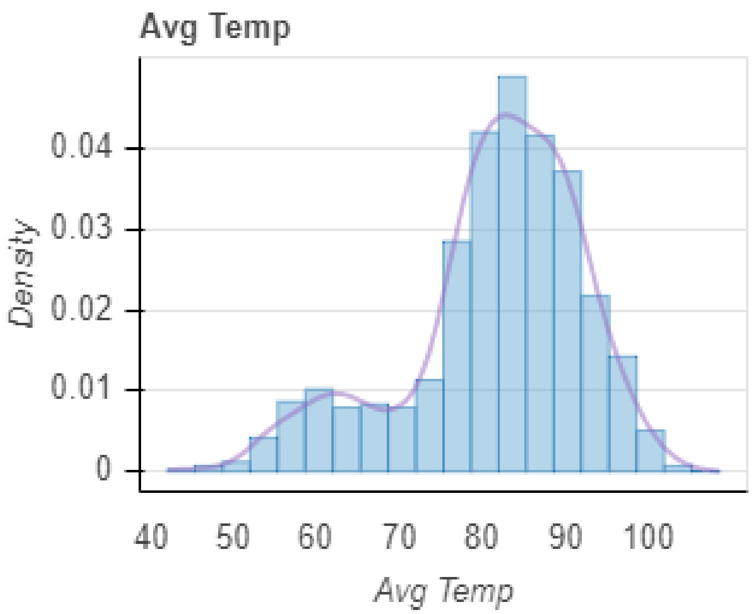
| | |
|---|---|
| Distinct Count | 409 |
| **Unique (%)** | 11.20% |
| **Missing** | 0 |
| **Minimum** | 29 |
| **5-th Percentile** | 54.3 |
| **Q1** | 81 |
| **Median** | 111 |
| **Q3** | 158 |
| **95-th Percentile** | 332 |
| **Maximum** | 659 |
| **Range** | 630 |
| **IQR** | 77 |
| **Mean** | 137.5053 |
| **Standard Deviation** | 86.7255 |
| **Variance** | 7521.314 |
| **Sum** | 501482 |
| **Skewness** | 1.953 |
| **Kurtosis** | 4.2881 |
| **Coefficient of Variation** | 0.6307 |

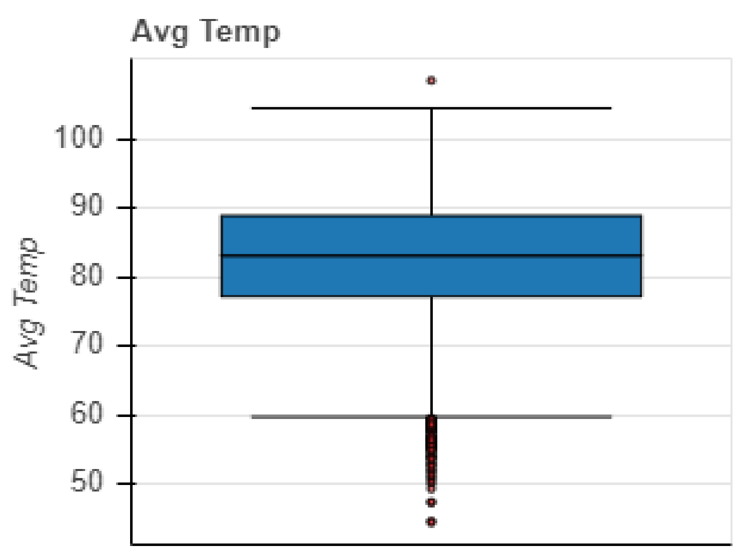


**Avg Temp**
`numerical`

| | |
|---|---|
| Distinct Count | 506 |
| **Unique (%)** | 13.90% |
| **Missing** | 0 |
| **Minimum** | 42 |
| **5-th Percentile** | 58.63 |
| **Q1** | 77.2 |
| **Median** | 83.1 |
| **Q3** | 88.9 |
| **95-th Percentile** | 96.1 |
| **Maximum** | 108.5 |
| **Range** | 66.5 |
| **IQR** | 11.7 |
| **Mean** | 81.4781 |
| **Standard Deviation** | 10.7843 |
| **Variance** | 116.3011 |
| **Sum** | 297150.5 |
| **Skewness** | -0.8555 |
| **Kurtosis** | 0.4597 |
| **Coefficient of Variation** | 0.1324 |

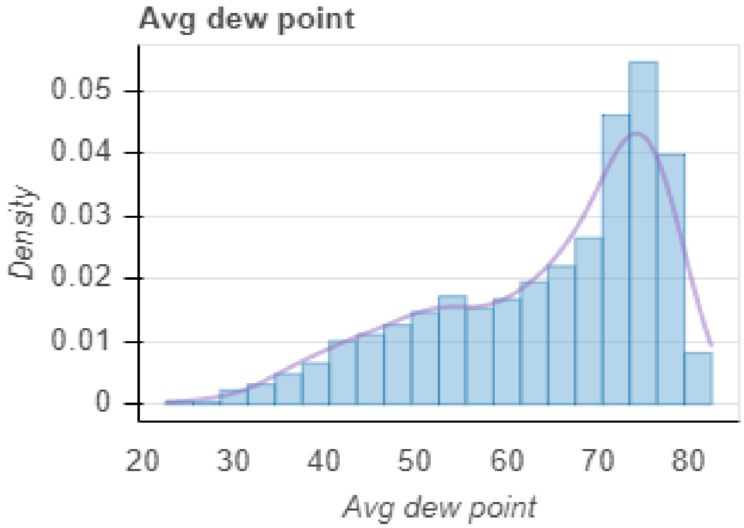


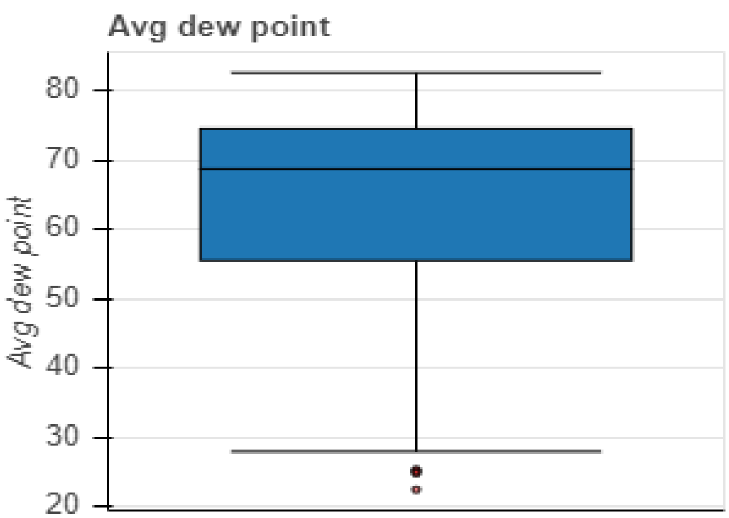



## Avg dew point

numerical

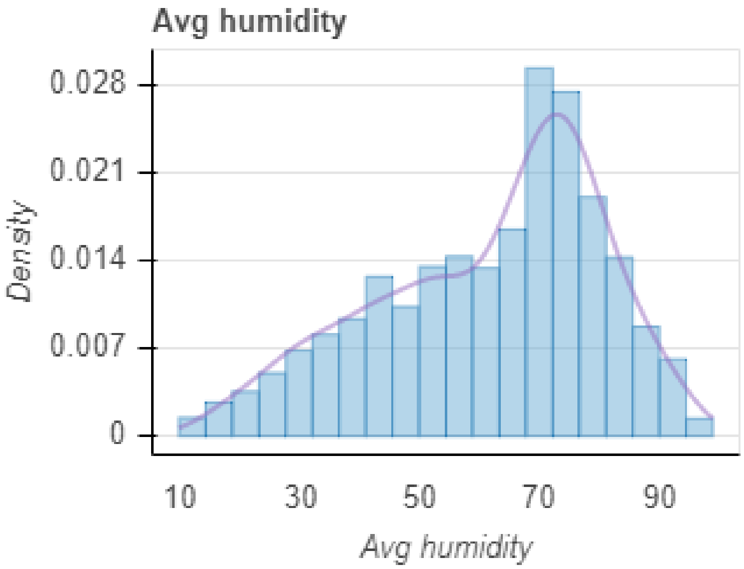
| | |
|---|---|
| Distinct Count | 514 |
| **Unique (%)** | 14.10% |
| **Missing** | 0 |
| **Minimum** | 22.4 |
| **5-th Percentile** | 40 |
| **Q1** | 55.5 |
| **Median** | 68.7 |
| **Q3** | 74.55 |
| **95-th Percentile** | 78.6 |
| **Maximum** | 82.6 |
| **Range** | 60.2 |
| **IQR** | 19.05 |
| **Mean** | 64.4658 |
| **Standard Deviation** | 12.5922 |
| **Variance** | 158.5625 |
| **Sum** | 235106.7 |
| **Skewness** | -0.8393 |
| **Kurtosis** | -0.2674 |

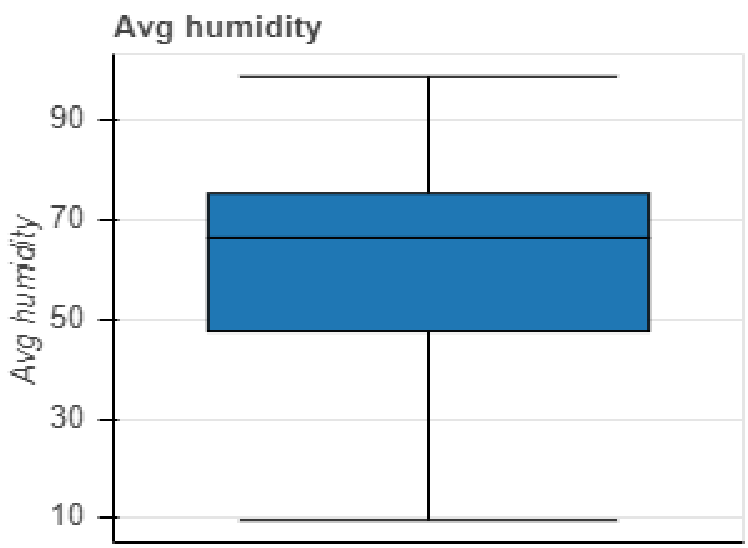


**Avg humidity**
`numerical`

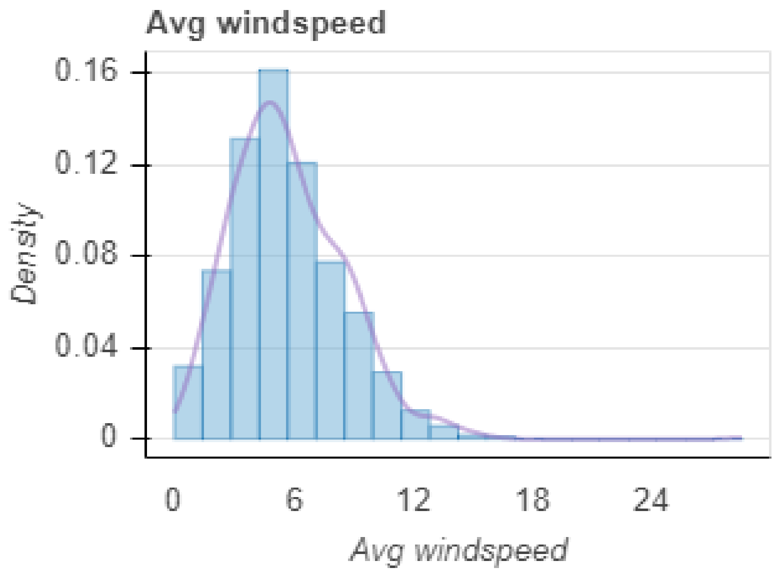
| | |
|---|---|
| Distinct Count | 788 |
| **Unique (%)** | 21.60% |
| **Missing** | 0 |
| **Minimum** | 9.6 |
| **5-th Percentile** | 26.1 |
| **Q1** | 47.6 |
| **Median** | 66.4 |
| **Q3** | 75.4 |
| **95-th Percentile** | 87.7 |
| **Maximum** | 98.9 |
| **Range** | 89.3 |
| **IQR** | 27.8 |
| **Mean** | 61.5547 |
| **Standard Deviation** | 18.8899 |
| **Variance** | 356.829 |
| **Sum** | 224490.1 |
| **Skewness** | -0.5454 |
| **Kurtosis** | -0.4846 |
| **Coefficient of Variation** | 0.3069 |

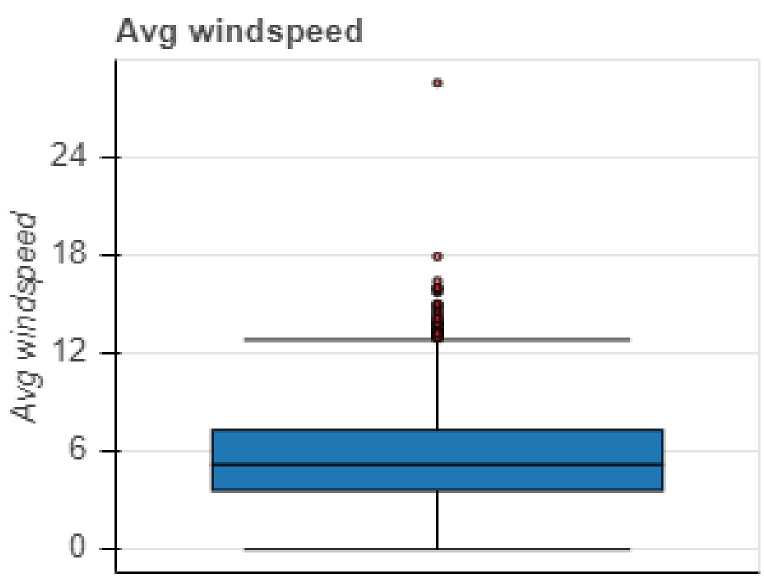


## Avg windspeed
numerical

| | |
|---|---|
| Distinct Count | 156 |
| **Unique (%)** | 4.30% |
| **Missing** | 0 |
| **Minimum** | 0 |
| **5-th Percentile** | 1.5 |
| **Q1** | 3.6 |
| **Median** | 5.2 |
| **Q3** | 7.3 |
| **95-th Percentile** | 10.5 |
| **Maximum** | 28.5 |
| **Range** | 28.5 |
| **IQR** | 3.7 |
| **Mean** | 5.5848 |
| **Standard Deviation** | 2.7723 |
| **Variance** | 7.6858 |
| **Sum** | 20367.6 |
| **Skewness** | 0.7534 |
| **Kurtosis** | 1.4803 |
| **Coefficient of Variation** | 0.4964 |



## From above EDA we conclude that:

PM2.5:-

- It is positively skewed.
- It is not normally distributed.
- It has outliers.
- It does not contain any negative value and zeros.

PM10:-

- It is right/positively skewed.
- It is not normally distributed.
- It has outliers.
- It does not contain any negatives and zeros.

NO:-

- It is positively skewed.
- It is not normally distributed.
- It has outliers.
- It does not contain any negative value and no zeros.

NO2:-

- It is right skewed.
- It is not normally distributed.
- It has outliers.
- It does not contain any negative value and no zeros.

NH3:-

- It is positively skewed.
- It is not normally distributed.
- It has outliers.
- It does not contain any negative value and no zeros.

CO:-

- It is positively skewed.
- It is not normally distributed.
- It has outliers.
- It does not contain any negative value but has 0.7% zeros.

SO2:-

- It is slightly right skewed.
- It is approximately normally distributed.
- It has outliers.
- It does not contain any negative value and no zeros.

O3:-

- It is right skewed.
- It is approximately normally distributed.
- It has outliers.
- It does not contain any negative value and no zeros.

AQI:-

- It is right skewed.
- It is not normally distributed.
- It has outliers.
- It does not contain any negative value and no zeros.

Temperature:-

- It is negatively skewed.
- It is not normally distributed.
- It has outliers.
- It does not contain any negative value and no zeros.

Dew Point:-

- It is left or negatively skewed.
- It is not normally distributed.

- It has outliers.
- It does not contain any negative value and no zeros.

Humidity:-

- It is slightly negatively skewed.
- It is not normally distributed.
- It has no outliers.
- It does not contain any negative value and no zeros.

Wind Speed:-

- It is positively skewed.
- It is not normally distributed.
- It has outliers.
- It does not contain any negative value and no zeros.

# Objective 1

To study Air Quality Index before lockdown and during lockdown.

# TIME SERIES ANALYSIS

A Time series is a collection of random variables at time t, represented as {X(t), t ∈ T} where T is an infinite set of time periods and t is an indexing parameter. The set T is called as time parameter space and the set of collection of all possible values taken by X(t) is called state space. **Time Series Analysis** comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data.

## Assumptions in Time Series Analysis:

### Stationarity:

The first assumption is that the series are stationary. Essentially, this means that the series are normally distributed and the mean and variance are constant and independent of time over a long time period.

### Uncorrelated random error:

We assume that the error term is randomly distributed and the mean and variance are constant over a time period. The Durbin-Watson test is the standard test for correlated errors.

### No outliers:

We assume that there is no outlier in the series. Outliers may affect conclusions strongly and can be misleading. Random shocks (a random error component): If shocks are present, they are assumed to be randomly distributed with a mean of 0 and a constant variance.

### Random shocks (a random error component):

If shocks are present, they are assumed to be randomly distributed with a mean of 0 and a constant variance.

## Some important concepts and terms:

### Dependence:

Dependence refers to the association of two observations with the same variable, at prior time points. Stationarity: Shows the mean value of the series that remains constant over a time period; if past effects accumulate and the values increase towards infinity, then stationarity is not met.

## Differencing:

Used to make the series stationary, to De-trend, and to control the auto-correlations; however, some time series analyses do not require differencing and over-differenced series can produce inaccurate estimates.

## Exponential smoothing in time series analysis:

This method predicts the one next period value based on the past and current value. It involves averaging of data such that the non-systematic components of each individual case or observation cancel out each other. The exponential smoothing method is used to predict the short-term predication. Alpha, Gamma, Phi, and Delta are the parameters that estimate the effect of the time series data. Alpha is used when seasonality is not present in data. Gamma is used when a series has a trend in data. Delta is used when seasonality cycles are present in data. A model is applied according to the pattern of the data.

## Curve fitting in time series analysis:

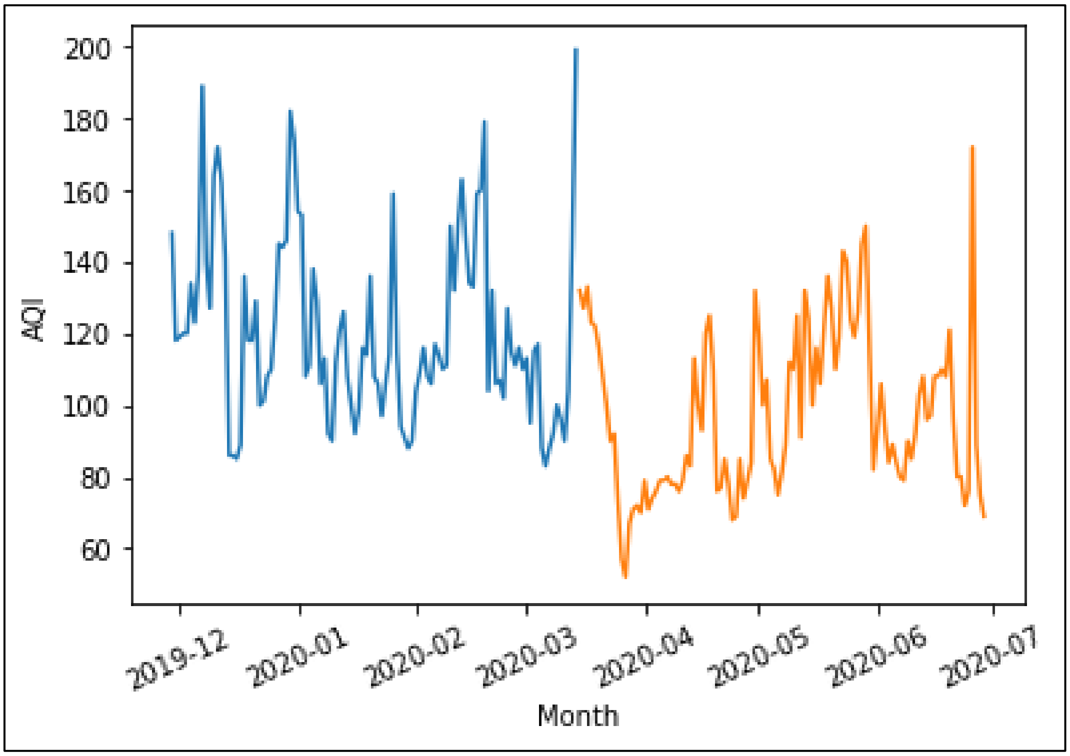
Curve fitting regression is used when data is in a non-linear relationship.

## Time Plot

A time series plot (sometimes called a time series graph) displays **values against time**. They are similar to Cartesian plane x-y graphs, but while an x-y graph can plot a variety of "x" variables (for example, height, weight, age), time plots can only display time on the x-axis. Unlike pie charts and bar charts, these plots do not have categories. Time plots are good for showing how data changes over time. For example, this type of chart would work well if you were sampling data at random times.

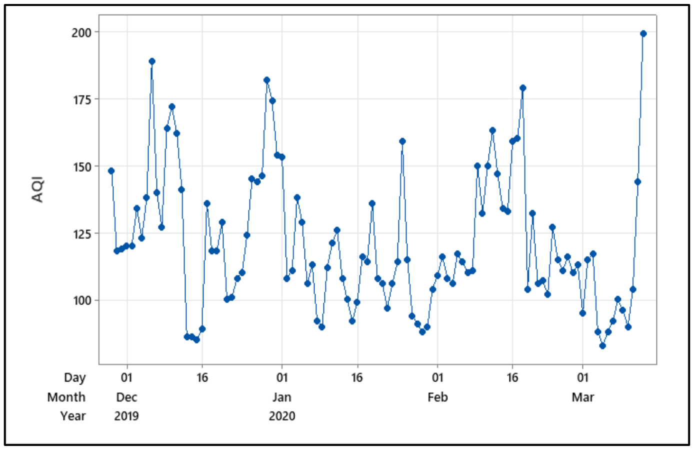
## City-Jaipur

Time plot for city Jaipur from December 2019 to June 2020.

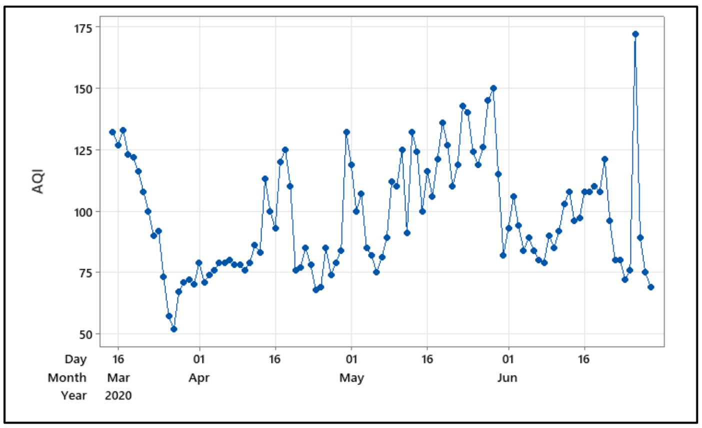



In above time plot for city Jaipur, blue line shows AQI before announcing lockdown and orange line shows AQI during the lockdown. Here, we clearly see that during lockdown, AQI decreases. This might be because there were less vehicular emissions, less human activities which contribute to Air Pollution. When government announced some relaxations, AQI showed an increase.

Also, to show it statistically first we make the two separate time plots of AQI before lockdown and during lockdown. Then we plot time series for both the observations.




Time plot before lockdown




Time plot during lockdown

From these plots it is visible that both the time series plots are not stationary. So, we make them stationary by differencing and then compare these time series by their respective values of means.

Mean of Time series data before lockdown= -0.0274

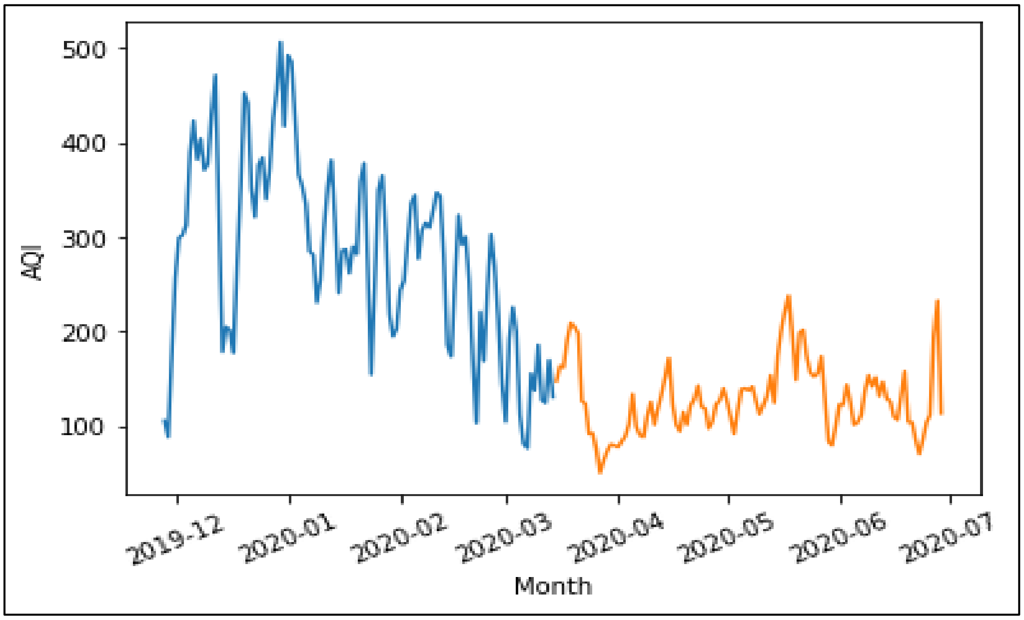
Mean of Time series data during lockdown= -0.59434

Mean of Time series data before lockdown ≥ Mean of Time series data during lockdown

So, we conclude that AQI level decreases during lockdown.

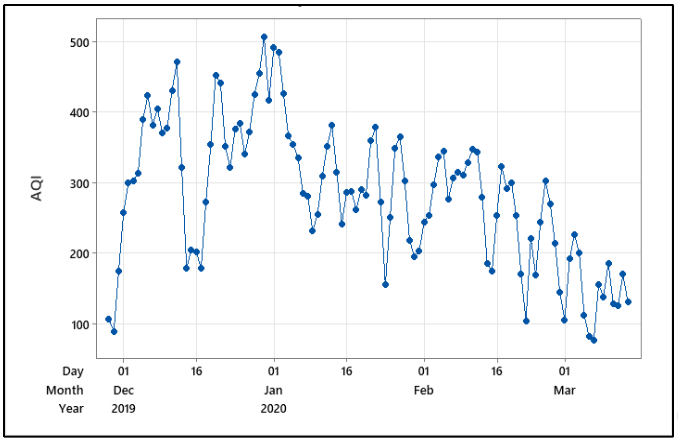
# City-Delhi

Time plot for city Delhi from December 2019 to June 2020.

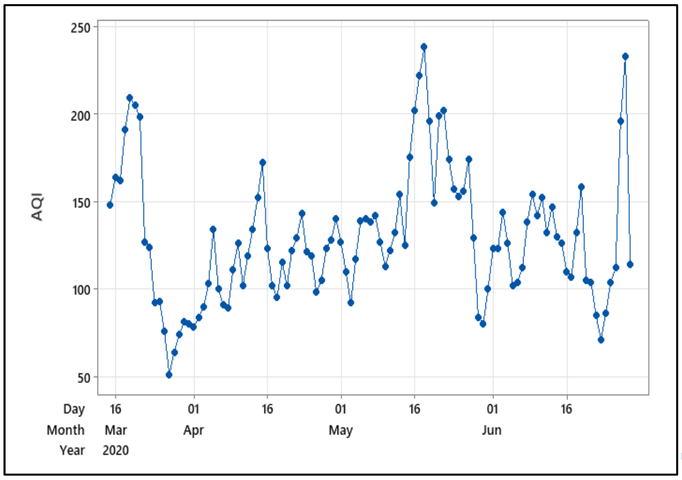



In above time plot for city Delhi, blue line shows Air Quality Index before announcement of lockdown and orange line shows Air Quality Index during the lockdown. Here, we clearly see that before the lockdown, AQI for Delhi is too high affecting human health severely. But after the announcement of lockdown AQI shows immense decrease. Delhi is believed to have highly contaminated air because of excessive vehicular activities, industrial outputs, human contribution to poor air quality.

Also, to show it statistically first we make the two separate time plots of AQI before lockdown and during lockdown. Then we plot time series for both the observations.




Time plot before lockdown




Time plot during lockdown

From these plots it is visible that both the time series plots are not stationary. So, we make them stationary by differencing and then compare these time series by their respective values of means.

Mean of Time series data before lockdown= -0.32075

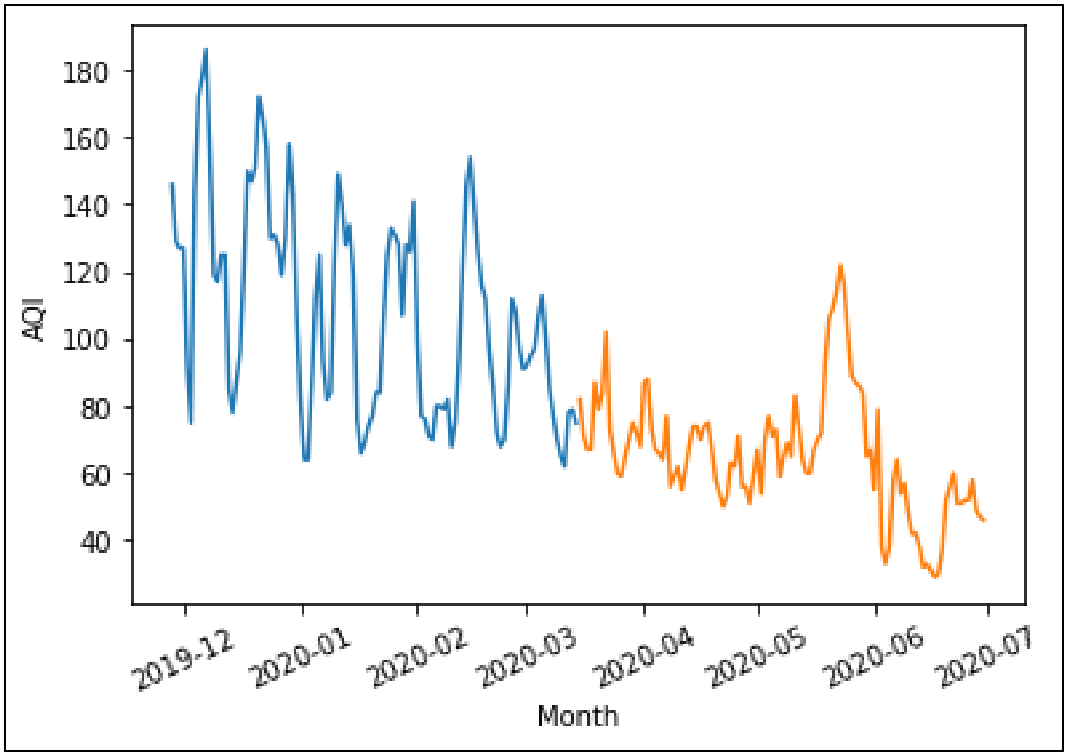
Mean of Time series data during lockdown= -0.4122

Mean of Time series data before lockdown ≥ Mean of Time series data during lockdown

So, we conclude that AQI level decreases during lockdown.

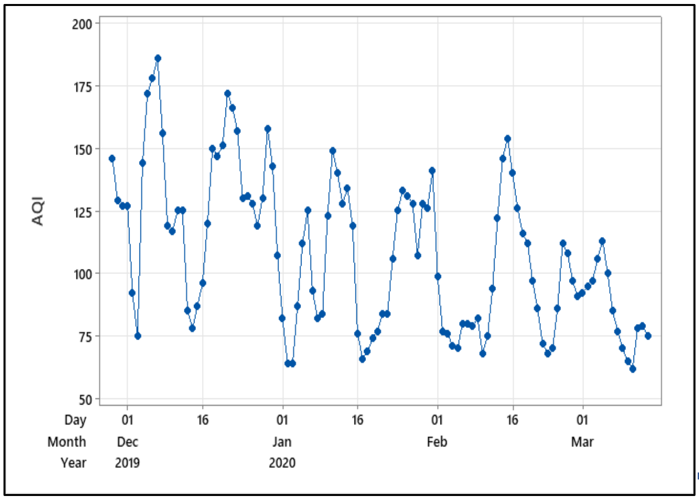
## City-Hyderabad

Time plot for city Hyderabad from December 2019 to June 2020.

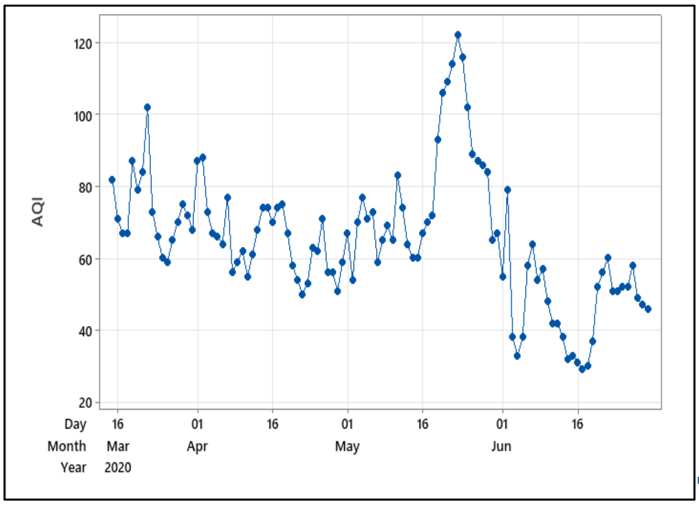



In above time plot for city Hyderabad, blue line shows AQI before the announcement of lockdown and orange line shows AQI during the lockdown. Here, we clearly see that during lockdown AQI decreases due to fewer industrial and human emissions. Even after providing relaxations from lockdown, AQI still shows a decreasing behavior.

Also, to show it statistically first we make the two separate time plots of AQI before lockdown and during lockdown. Then we plot time series for both the observations.



Time plot before lockdown



Time plot during lockdown

From these plots it is visible that both the time series plots are not stationary. So, we make them stationary by differencing and then compare these time series by their respective values of means.

Mean of Time series data before lockdown= -0.058553

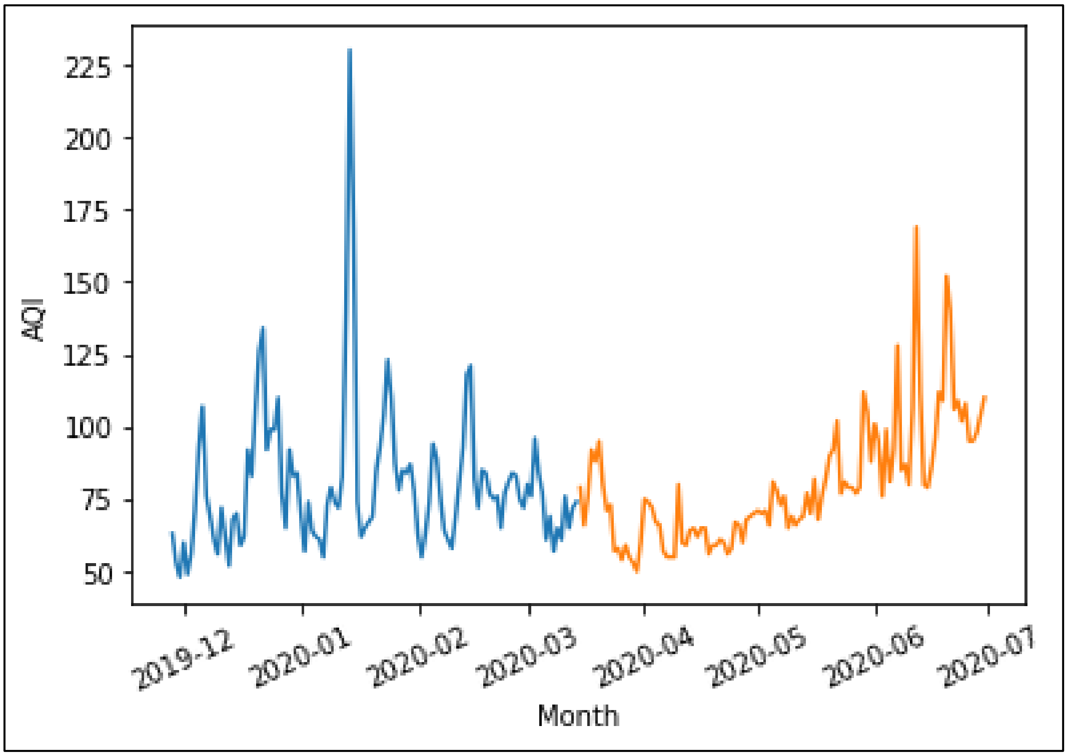
Mean of Time series data during lockdown= -0.33645

Mean of Time series data before lockdown ≥ Mean of Time series data during lockdown

So, we conclude that AQI level decreases during lockdown.

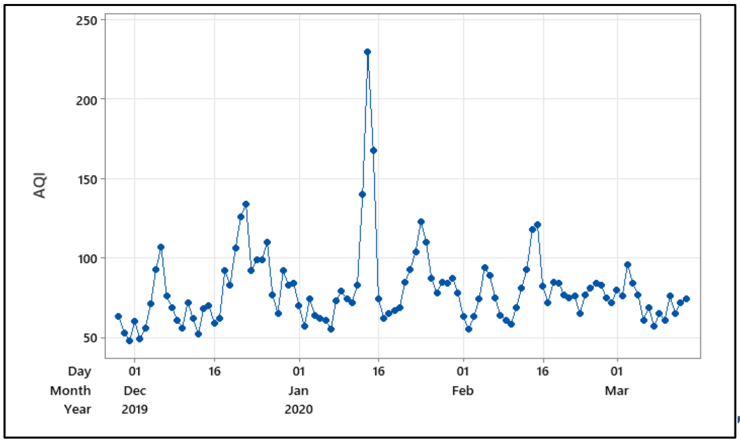
# City-Chennai

Time plot for city Chennai from December 2019 to June 2020.

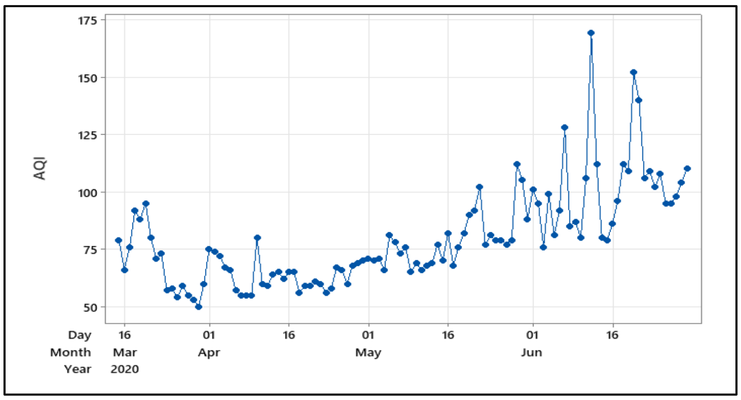



In above time plot for city Chennai, blue line shows AQI before announcing lockdown and orange line shows AQI during the lockdown. Visibly during lockdown AQI decreases. After announcement of relaxations from lockdown, AQI again started increasing to higher values.

Also, to show it statistically first we make the two separate time plots of AQI before lockdown and during lockdown. Then we plot time series for both the observations.



Time plot before lockdown



Time plot during lockdown

From these plots it is visible that both the time series plots are not stationary. So, we make them stationary by differencing and then compare these time series by their respective values of means.

Mean of Time series data before lockdown= -0.01868

Mean of Time series data during lockdown= -0.28972

Mean of Time series data before lockdown ≥ Mean of Time series data during lockdown

So, we conclude that AQI level decreases during lockdown.

# Conclusion

Using time plot we conclude that Air Quality Index decreases during lockdown.

# Objective 2

To forecast future values for Air Quality Index for four cities.

# TIME SERIES ANALYSIS

## Autoregressive model:

An Autoregressive (AR) model predicts future behavior based on past behavior.

An autoregressive model of order p, denoted as $AR(p)$, is of the form,

$$x_t = \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + \ldots + \varphi_p x_{t-p} + w_t$$

Where, $x_t$ is stationary, and $\varphi_1$, $\varphi_2$ and $\varphi_p$ are constants ($\varphi_p \neq 0$).

Although it is not necessary yet, we assume that $w_t$ is a Gaussian white noise series with mean zero and variance $\sigma_w^2$, unless otherwise stated. If the mean, μ, of $x_t$ is not zero,

replace $x_t$ by $x_t - \mu$ and the model becomes,

$$x_t - \mu = \varphi_1(x_{t-1} - \mu) + \varphi_2(x_{t-2} - \mu) + \ldots + \varphi_p(x_{t-p} - \mu) + w_t$$

Which can be simplified to,

$$x_t = \alpha + \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + \ldots + \varphi_p x_{t-p} + w_t$$

Where, $\alpha = \mu(1 - \varphi_1 - \ldots - \varphi_p)$.

The model can be expressed using backshift operator B, which is very useful in determining the various properties of the time series model, as

$$(1 - \varphi_1 B_1 - \varphi_2 B_2 \ldots - \varphi p B p)x_t = w_t$$

or even more concisely as,

$$\varphi(B)x_t = w_t$$

Where, $\varphi(B) = 1 - \varphi_1 B_1 - \varphi_2 B_2 \ldots - \varphi p B p$ and called as autoregressive operator.

## Moving Average model:

Moving Average (MA) model uses past forecast errors to predict future values.

The moving average model of order q, or MA(q) model, is defined to be

$$x_t = w_t + \theta_1 w_{t-1} + \theta 2 w_{t-2} + \cdots + \theta_q w_{t-q}$$

Where, there are q lags in the moving average and $\theta_1, \theta_2, \dots, \theta_q$ ($\theta_q \neq 0$) are parameters. Although it is not necessary yet, we assume that $w_t$ is a Gaussian white noise series with mean zero and variance $\sigma_w^2$, unless otherwise stated. We may also write the MA(q) process in the equivalent form

$$x_t = \theta(B)w_t$$

Where, $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 \cdot \cdot \cdot - \theta_q B^q$, called as moving average operator. Unlike the autoregressive process, the moving average process is stationary for any values of the parameters $\theta_1, \theta_2, \dots, \theta_q$.

## ARMA Model:

A time series $\{x_t, t = 0, \pm 1, \pm 2, \dots\}$ is ARMA(p, q) if it is stationary and

$$x_t = \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + \dots + \varphi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \cdot \cdot \cdot + \theta_q w_{t-q}$$

with $\varphi_p \neq 0$, $\theta_q \neq 0$. We assume that $w_t$ is a Gaussian white noise series with mean zero and variance $\sigma_w^2 > 0$.

When q = 0, the model is called an autoregressive model of order p, AR(p), and when p = 0, the model is called a moving average model of order q, MA(q). To aid in the investigation of

ARMA models, it will be useful to write them using the AR operator and the MA operator. In particular, The ARMA(p, q) model can then be written in concise form as

$$\varphi(B)\, x_t = \theta(B)w_t$$

## ARIMA:

Autoregressive Integrated Moving Average model (ARIMA) is a generalization of an Autoregressive Moving Average (ARMA) model.

ARIMA models are applied in some cases where data show evidence of non-stationarity in the sense of mean (but not variance/autocovariance), where an initial differencing step (corresponding to the "**Integrated**" part of the model) can be applied one or more times to eliminate the non-stationarity of the mean function (i.e., the trend). When the seasonality shows in a time series, the seasonal-differencing could be applied to eliminate the seasonal component.

## Parameters of ARIMA model :

**Autoregressive component**: In ARIMA model, AR stands for autoregressive. Autoregressive parameter is denoted by p. When p=0, it means that there is no auto-correlation in the series. When p=1, it means that the series auto-correlation is till one lag.

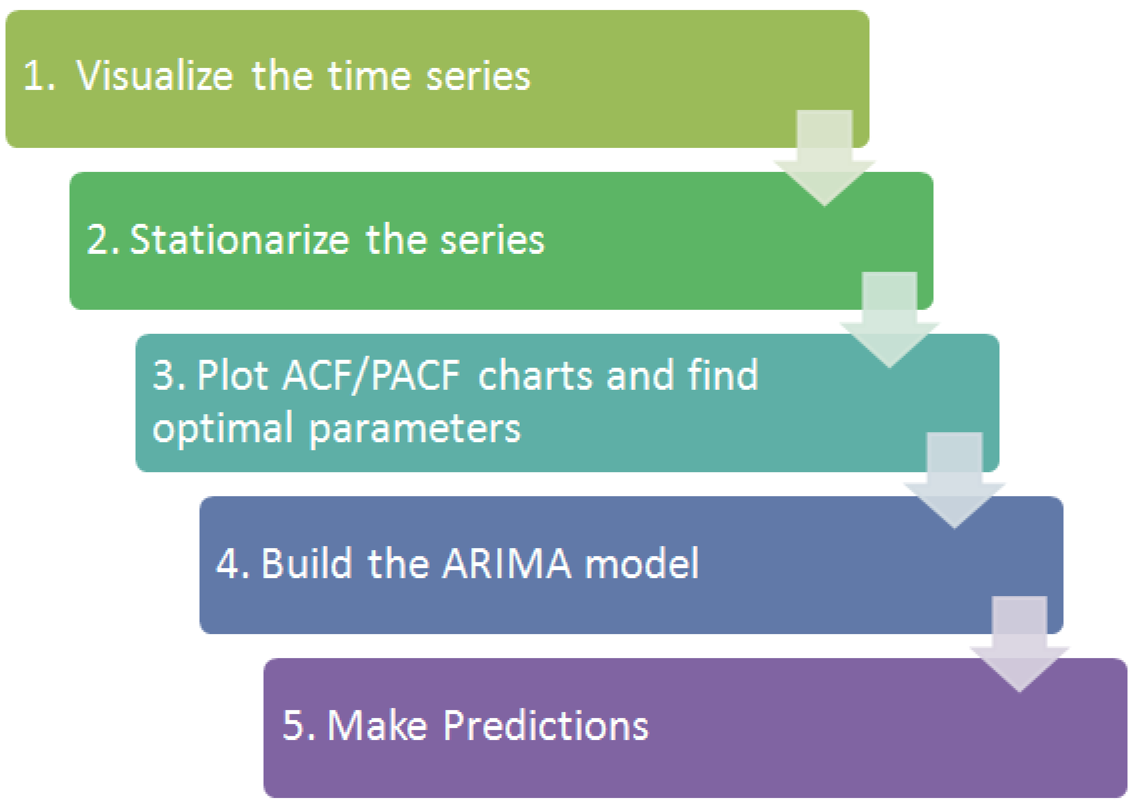
**Integrated**: In ARIMA time series analysis, integrated is denoted by d. Integration is the inverse of differencing. When d=0, it means the series is stationary and we do not need to take the difference of it. When d=1, it means that the series is not stationary and to make it stationary, we need to take the first difference. When d=2, it means that the series has been differenced twice. Usually, more than two-time difference is not reliable.

**Moving average component**: In ARIMA model, MA stands for moving the average. MA parameter is denoted by q. In ARIMA, moving average q=1 means that it is an error term and there is auto-correlation with one lag. In order to test whether or not the series and their error term is auto correlated, we usually use W-D test, ACF, and PACF.

## Decomposition:

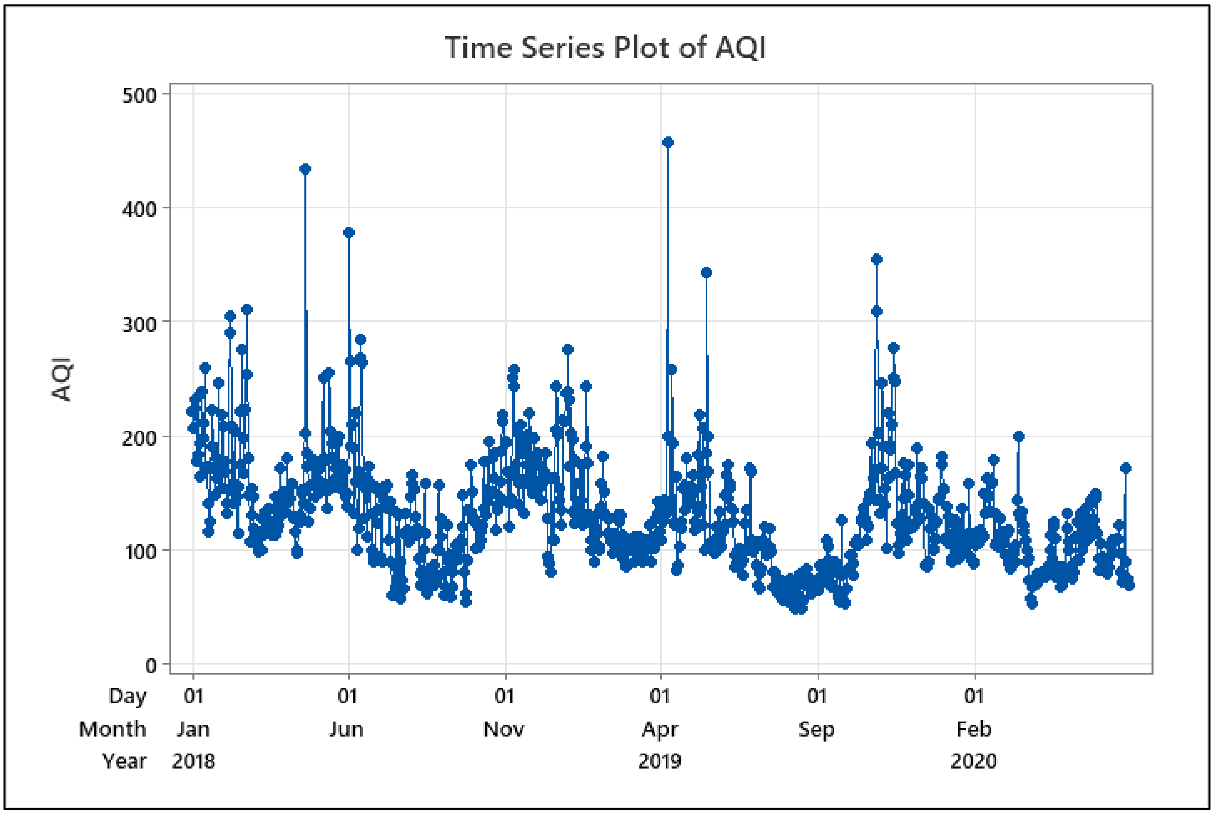
Refers to separating a time series into trend, seasonal effects, and remaining variability.

# How to do a Time Series Analysis:

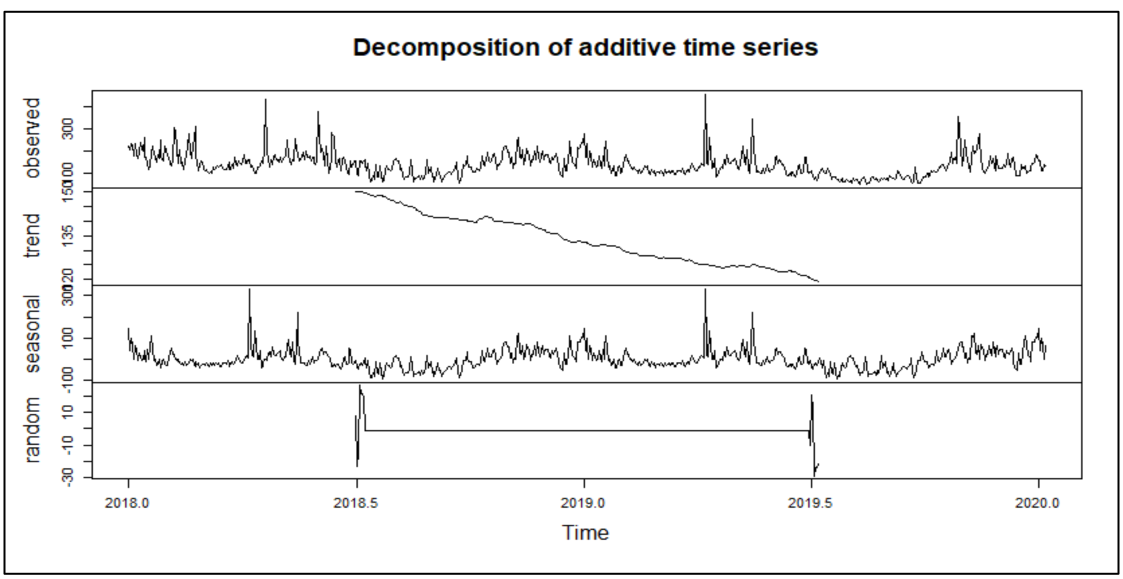


## For city JAIPUR

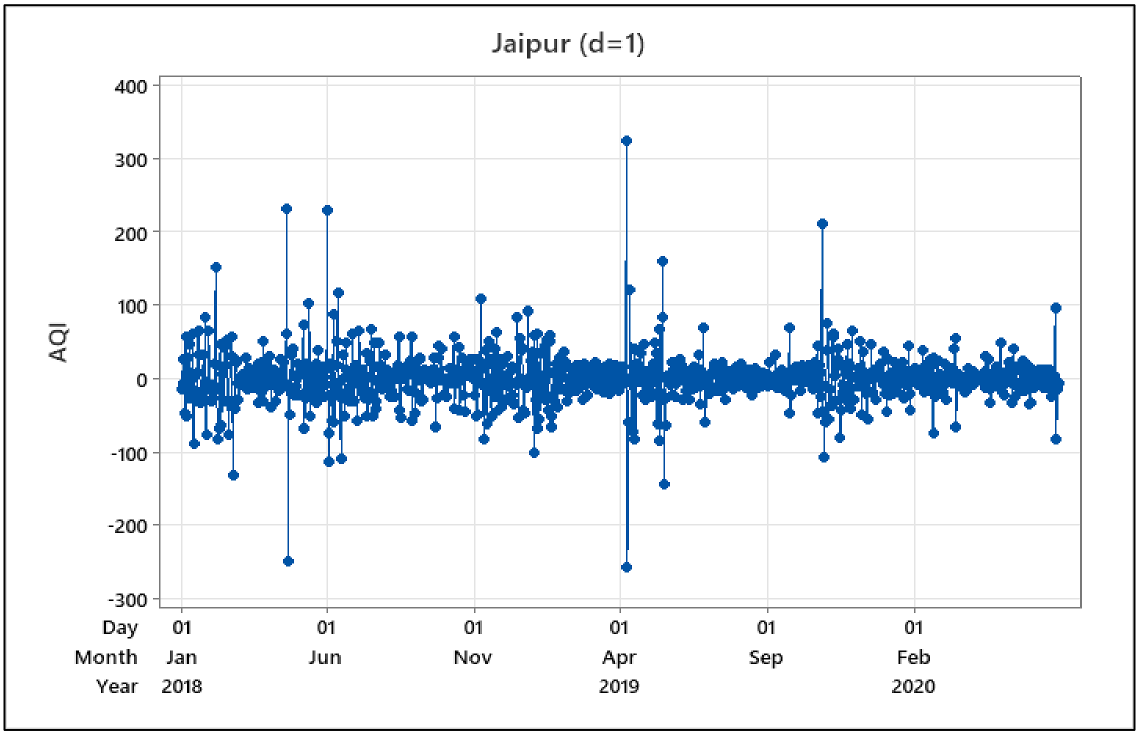
Time Series plot for AQI of city Jaipur from 01/01/2018 to 30/06/2020.




It is a Time Series plot for city Jaipur with days along the x-axis and AQI figures on the y-axis. As we can see our time series is not stationary. Also, by using decomposition.



We can see there is decreasing trend. This suggests that the time series is not stationary and will require differencing to make it stationary, at least a difference order of 1.




It is a Time Series plot of AQI data differenced for lag 1.

## Estimation of Parameters (p, d, q)

Since we difference data with order one to make time series stationary results d=1.

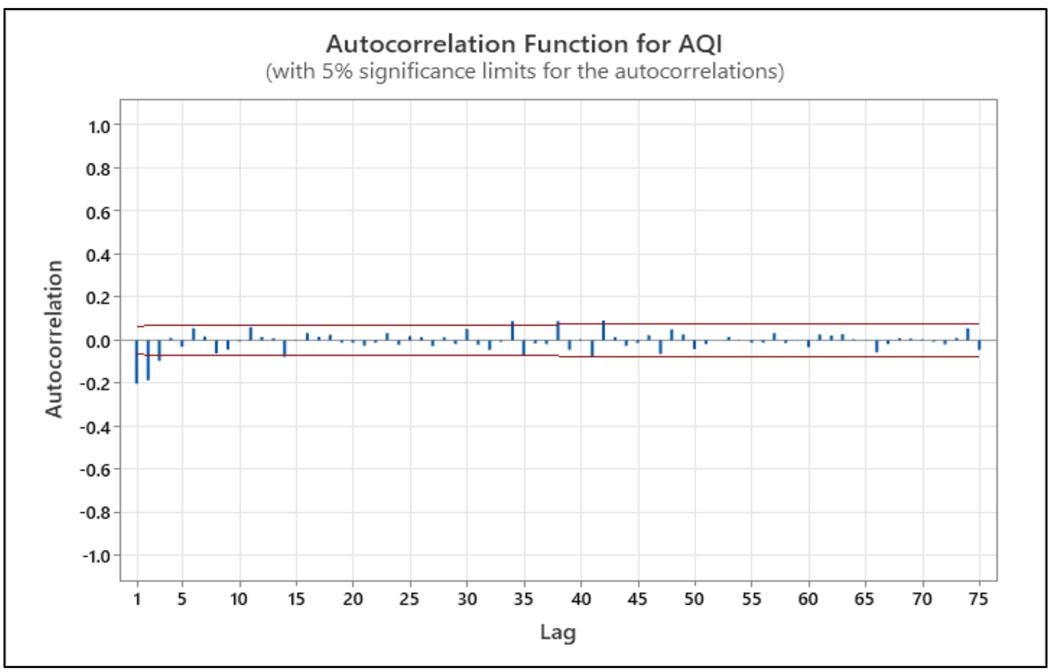
Now we have to plot the ACF and PACF plots for the estimation of parameters (p, q).

## Plot of ACF and PACF:

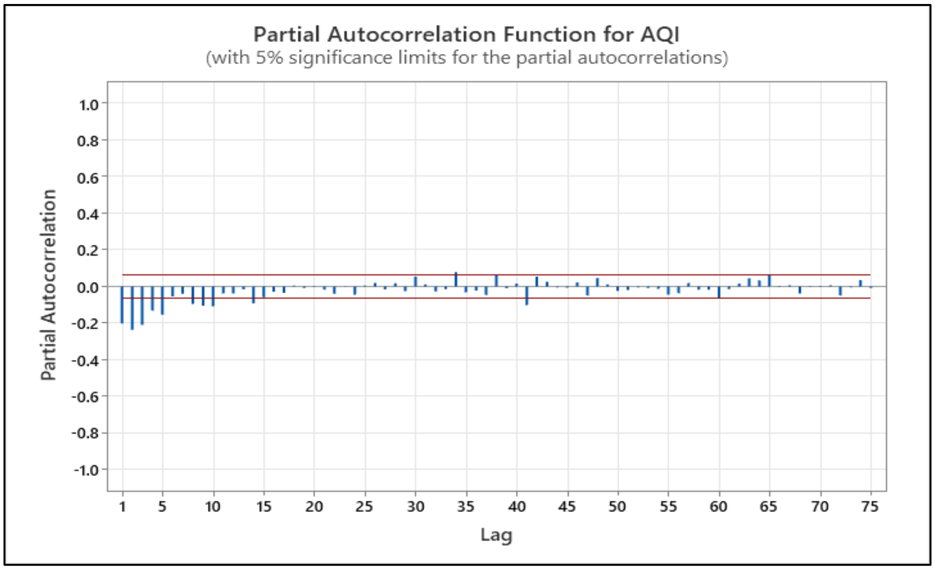
To determine the order of autoregressive (AR) and moving average (MA) series we use ACF and PACF plots. ACF and PACF plots are used to obtain the values of q and p to feed into the ARIMA model

**ACF** is an (complete) auto-correlation function which gives us values of auto-correlation of any series with its lagged values. In simple terms, it describes how well the present value of the series is related with its past values. Order q of the MA process is obtained from the ACF plot, this is the lag after which ACF crosses the upper confidence interval for the first time.

**PACF** is a partial auto-correlation function. Basically, instead of finding correlations of present with lags like ACF, it finds correlation of the 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' as we remove already found variations before we find the next correlation. Order p of the AR process is obtained from PACF plot.




From the above ACF plot we conclude that ACF is significant up to more than 3 lags. So, we consider the value of q=3.




From PACF plot we see that it increases at second lag again it decreases and significant up to 5 lag So, we consider the value of p=1.

Hence, we fit several different models for AQI. Finally, we consider p=1 and q=1. The best fit model we found is ARIMA(1,1,1) with no constant. The final estimates of the parameters are given below.

## Final Estimates of Parameters:

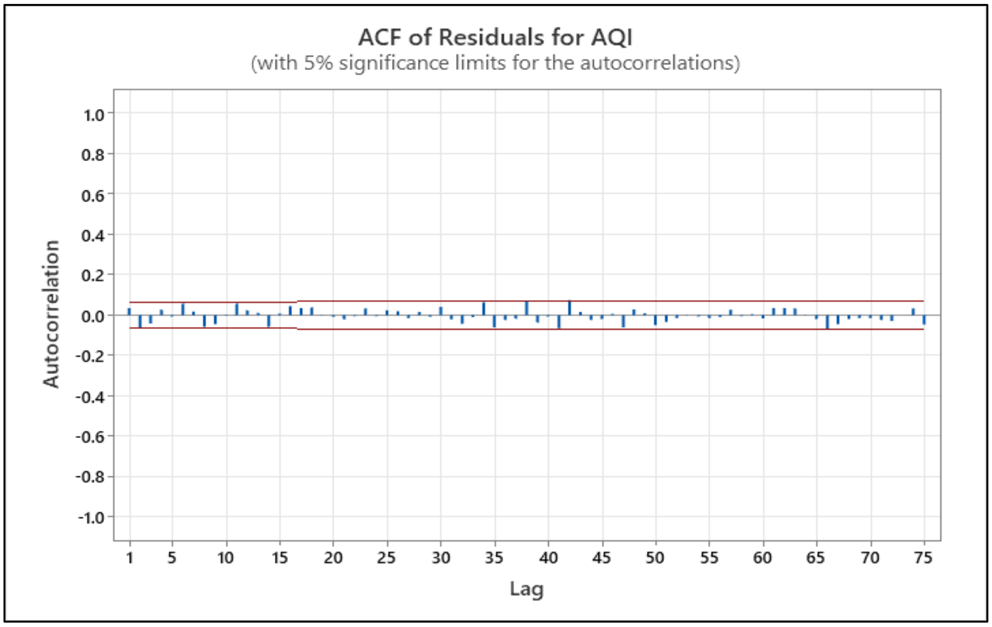
| Type | Coeff | SE Coeff | T-Value | P-Value |
|---|---|---|---|---|
| AR 1 | 0.4795 | 0.0363 | 13.23 | 0.00 |
| MA1 | 0.9114 | 0.0158 | 57.54 | 0.00 |

The significance of the parameters is tested using t-test with p-value very small for all. This indicates that all the parameters are significant.

## Residual Sums of Squares

| DF | SS | MS |
|---|---|---|
| 907 | 981591 | 1082.24 |

The residuals must not be auto correlated and this is apparent below with autocorrelation being insignificant for all lags.



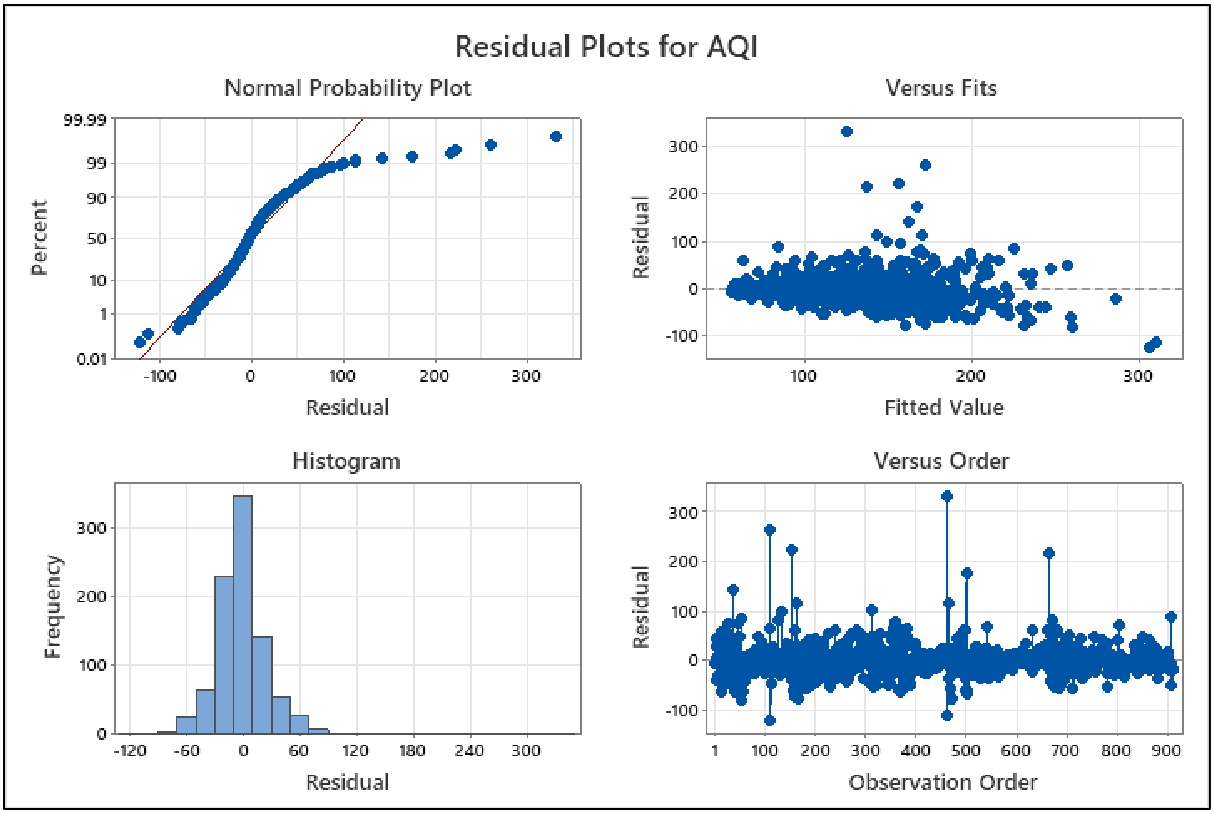

The significance of the autocorrelation among the residuals at four lags is tested using Ljung-Box Chi-Square Statistics and output is shown below.

## Modified Box-Pierce (Ljung-Box) Chi-Square Statistic

| **Lag** | **12** | **24** | **36** | **48** |
|---|---|---|---|---|
| Chi-Square | 18.24 | 27.17 | 40.24 | 60.79 |
| DF | 9 | 21 | 33 | 45 |
| P-Value | 0.032 | 0.165 | 0.18 | 0.058 |

The p-values for the test at all the four lags are large pointing towards the insignificance of autocorrelation among residuals.

Also, the residuals are normally distributed which can be observed in the NPP and histogram. The error has constant variance which is evident from the graph of residual against the fitted values.

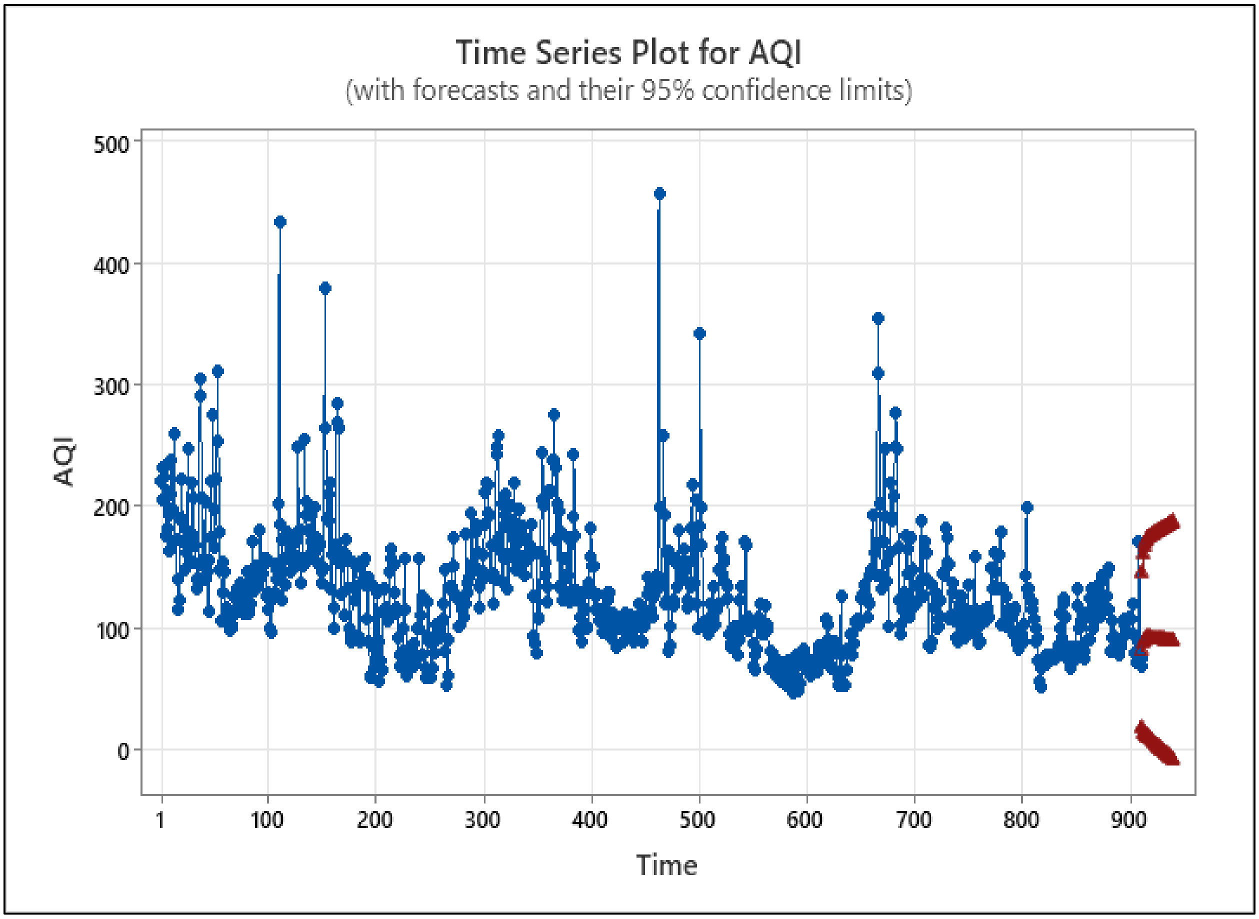



Now that we have fit the best possible model with all the assumptions being satisfied, we move to our main purpose of forecasting.

## Forecast value from observation 912 (for next 30 days)

| Observations | Forecast | 95% Limits | |
|---|---|---|---|
| | | Lower | Upper |
| 912 | 80.9882 | 16.4963 | 145.48 |
| 913 | 86.6738 | 12.5004 | 160.847 |
| 914 | 89.3374 | 11.5949 | 167.08 |
| 915 | 90.5519 | 10.9976 | 170.106 |
| 916 | 91.0715 | 10.3273 | 171.816 |
| 917 | 91.258 | 9.5765 | 172.94 |
| 918 | 91.2848 | 8.7791 | 173.791 |

| | | | |
|---|---|---|---|
| 919 | 91.2349 | 7.9597 | 174.51 |
| 920 | 91.1484 | 7.1324 | 175.164 |
| 921 | 91.0442 | 6.3041 | 175.784 |
| 922 | 90.9316 | 5.4784 | 176.385 |
| 923 | 90.8149 | 4.6567 | 176.973 |
| 924 | 90.6962 | 3.8399 | 177.553 |
| 925 | 90.5767 | 3.0283 | 178.125 |
| 926 | 90.4567 | 2.2219 | 178.691 |
| 927 | 90.3365 | 1.4208 | 179.252 |
| 928 | 90.2162 | 0.6247 | 179.808 |
| 929 | 90.0958 | -0.1663 | 180.358 |
| 930 | 89.9754 | -0.9524 | 180.903 |
| 931 | 89.855 | -1.7337 | 181.444 |
| 932 | 89.7346 | -2.5102 | 181.979 |
| 933 | 89.6142 | -3.2821 | 182.51 |
| 934 | 89.4938 | -4.0494 | 183.037 |
| 935 | 89.3734 | -4.8124 | 183.559 |
| 936 | 89.253 | -5.5709 | 184.077 |
| 937 | 89.1325 | -6.3252 | 184.59 |
| 938 | 89.0121 | -7.0753 | 185.1 |
| 939 | 88.8917 | -7.8213 | 185.605 |
| 940 | 88.7713 | -8.5633 | 186.106 |
| 941 | 88.6509 | -9.3014 | 186.603 |




Time series plot after fitting ARIMA model. It shows AQI of Jaipur city since 2018 (blue line) and forecasts and their 95% confidence limits (red line) for next month. On close observation we can see that AQI values slightly decrease in the forecasted values.

## Model Summary for checking model efficiency :

**Model Summary**

**Model Fit**

| Fit Statistic | Mean | SE | Minimum | Maximum | Percentile 5 | 10 | 25 | 50 | 75 | 90 | 95 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Stationary R-squared | .194 | . | .194 | .194 | .194 | .194 | .194 | .194 | .194 | .194 | .194 |
| R-squared | .558 | . | .558 | .558 | .558 | .558 | .558 | .558 | .558 | .558 | .558 |
| RMSE | 32.914 | . | 32.914 | 32.914 | 32.914 | 32.914 | 32.914 | 32.914 | 32.914 | 32.914 | 32.914 |
| MAPE | 15.791 | . | 15.791 | 15.791 | 15.791 | 15.791 | 15.791 | 15.791 | 15.791 | 15.791 | 15.791 |
| MaxAPE | 95.510 | . | 95.510 | 95.510 | 95.510 | 95.510 | 95.510 | 95.510 | 95.510 | 95.510 | 95.510 |
| MAE | 20.978 | . | 20.978 | 20.978 | 20.978 | 20.978 | 20.978 | 20.978 | 20.978 | 20.978 | 20.978 |
| MaxAE | 331.437 | . | 331.437 | 331.437 | 331.437 | 331.437 | 331.437 | 331.437 | 331.437 | 331.437 | 331.437 |
| Normalized BIC | 7.010 | . | 7.010 | 7.010 | 7.010 | 7.010 | 7.010 | 7.010 | 7.010 | 7.010 | 7.010 |

**Model Statistics**

| Model | Number of Predictors | Model Fit statistics: Stationary R-squared | R-squared | RMSE | MAPE | MaxAPE | Ljung-Box Q(18): Statistics | DF | Sig. | Number of Outliers |
|---|---|---|---|---|---|---|---|---|---|---|
| AQI-Model_1 | 0 | .194 | .558 | 32.914 | 15.791 | 95.510 | 25.574 | 16 | .060 | 0 |

**Residual ACF**

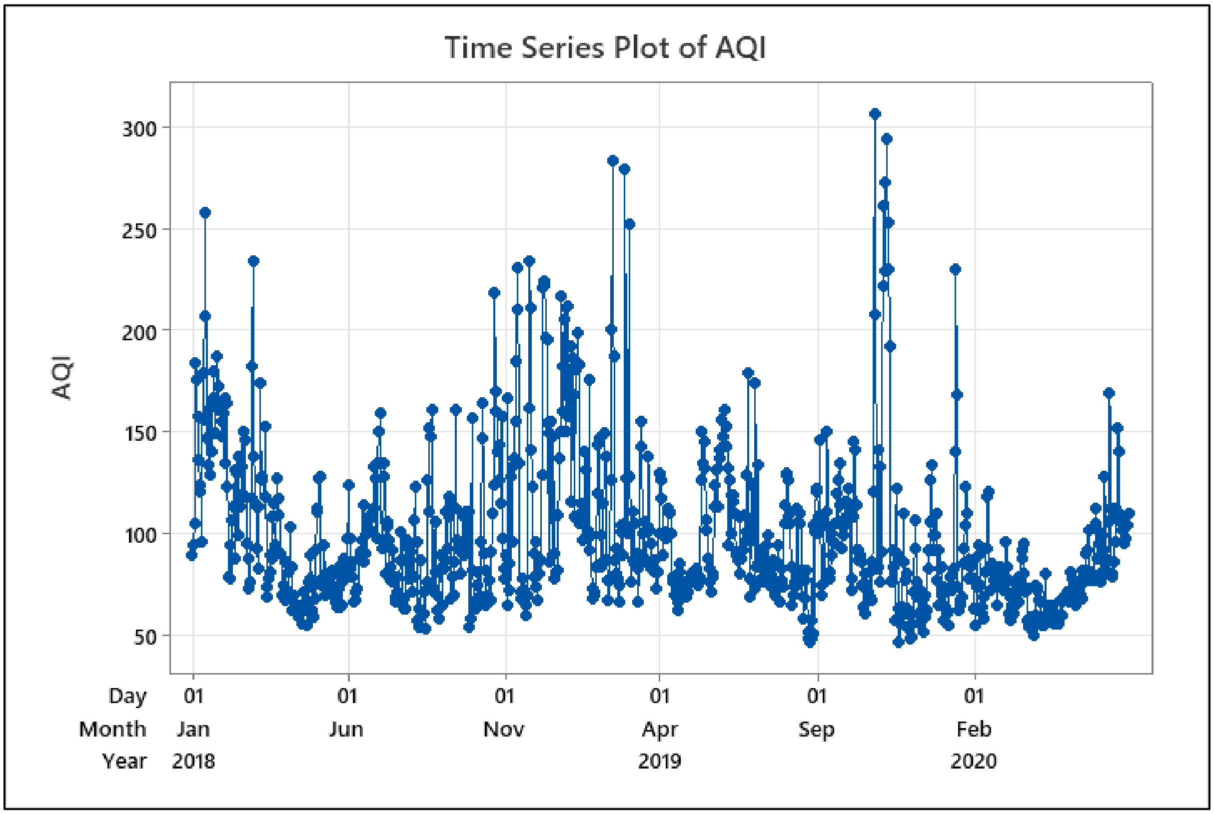
| Model | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AQI-Model_1 | ACF | .034 | -.062 | -.041 | .027 | -.006 | .056 | .017 | -.057 | -.044 | -.003 | .057 | .024 | .011 |
| | SE | .033 | .033 | .033 | .033 | .033 | .033 | .033 | .034 | .034 | .034 | .034 | .034 | .034 |

**Residual PACF**

Since the MAPE value is 15.791, on an average, the forecast is off by 15.79%

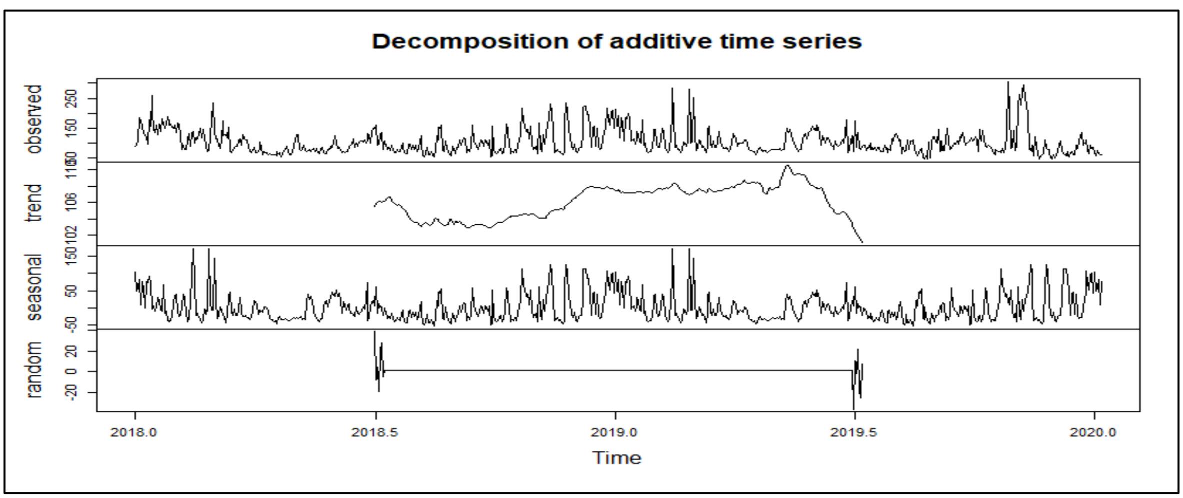
# For city Chennai

Time Series plot for AQI of city Chennai from 01/01/2018 to 30/06/2020.




It is a Time Series plot for city Chennai with days along the x-axis and AQI figures on the y-axis.

This time plot suggests that the time series is stationary. Also, by using decomposition.



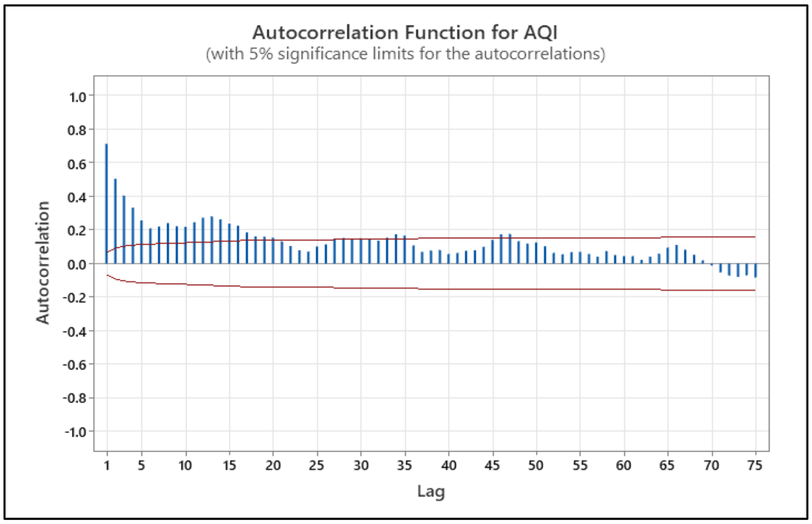
Also, by using decomposition there is no trend or seasonality. This suggests that the time series is stationary. Now we find parameters.

## Estimation of Parameters (p, d, q)

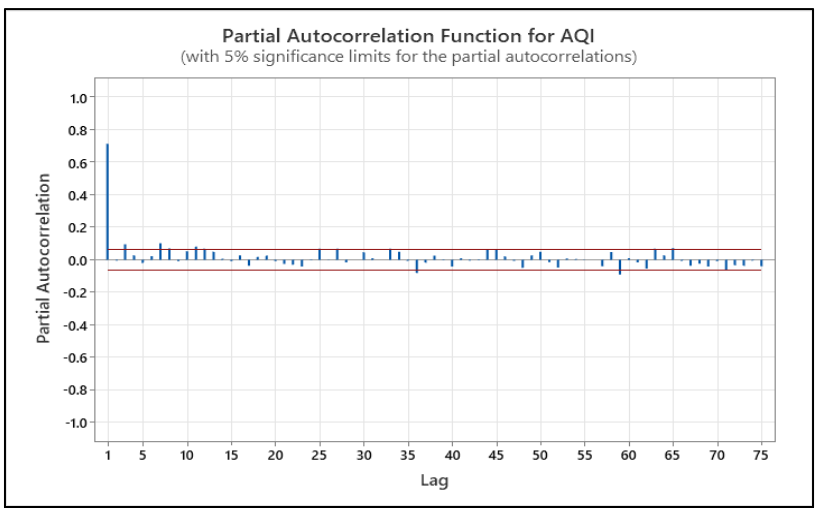
Since our time series is stationary results d=0.

Now we have to plot the ACF and PACF plots for the estimation of parameters (p, q).

# Plot of ACF and PACF:




The ACF is significant up to more than 20 lags. Here to fit model we consider the value of q=5.




From PACF, we consider the value of p=4.

Hence, we fit several different models for AQI. Finally, we consider p=4 and q=3 the best fit model we found is ARIMA(4,0,3) with no constant. The final estimates of the parameters are given below.

## Final Estimates of Parameters:

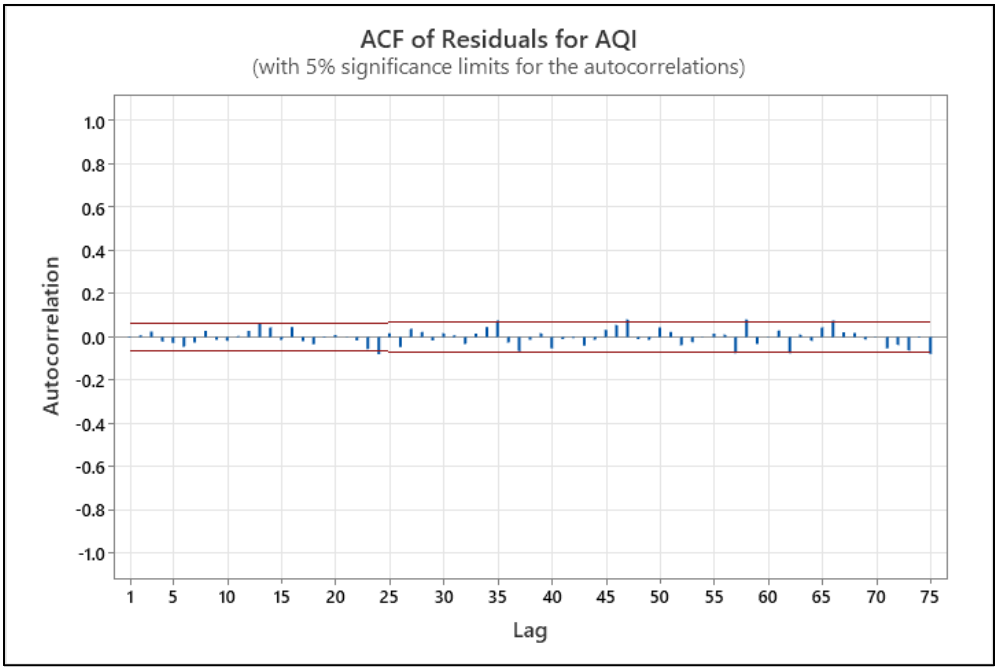
| Type | Coeff | SE Coeff | T-Value | P-Value |
|---|---|---|---|---|
| AR1 | 1.393 | 0.237 | 5.88 | 0.000 |
| AR2 | -1.014 | 0.357 | -2.84 | 0.005 |
| AR3 | 1.052 | 0.271 | 3.89 | 0.000 |
| AR4 | -0.444 | 0.106 | -4.17 | 0.000 |
| MA1 | 0.7 | 0.24 | 2.92 | 0.004 |
| MA2 | -0.454 | 0.233 | -1.95 | 0.052 |
| MA3 | 0.644 | 0.168 | 3.83 | 0.000 |

The significance of the parameters is tested using t-test with p-value very small for all. This indicates that all the parameters are significant.

## Residual Sums of Squares

| DF | SS | MS |
|---|---|---|
| 904 | 672312 | 743.708 |

The residuals must not be auto correlated and this is apparent below with autocorrelation being insignificant for all lags.



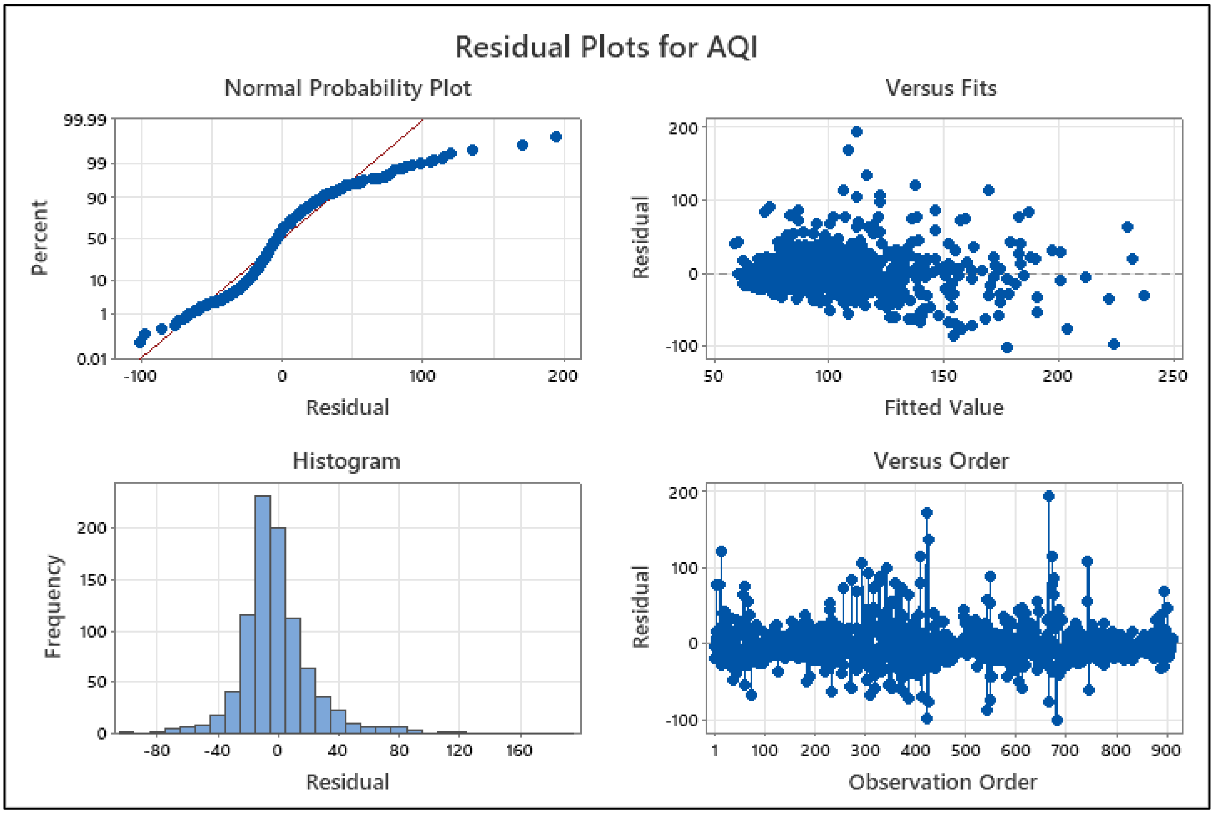

The significance of the autocorrelation among the residuals at four lags is tested using Ljung-Box Chi-Square Statistics and output is shown below.

## Modified Box-Pierce (Ljung-Box) Chi-Square Statistic

| **Lag** | **12** | **24** | **36** | **48** |
|---|---|---|---|---|
| Chi-Square | 6.14 | 24.73 | 38.97 | 58.08 |
| DF | 4 | 16 | 28 | 40 |
| P-Value | 0.189 | 0.075 | 0.081 | 0.032 |

The p-values for the test at all the four lags are large pointing towards the insignificance of autocorrelation among residuals.

Also, the residuals are normally distributed which can be observed in the NPP and histogram. The error has constant variance which is evident from the graph of residual against the fitted values.

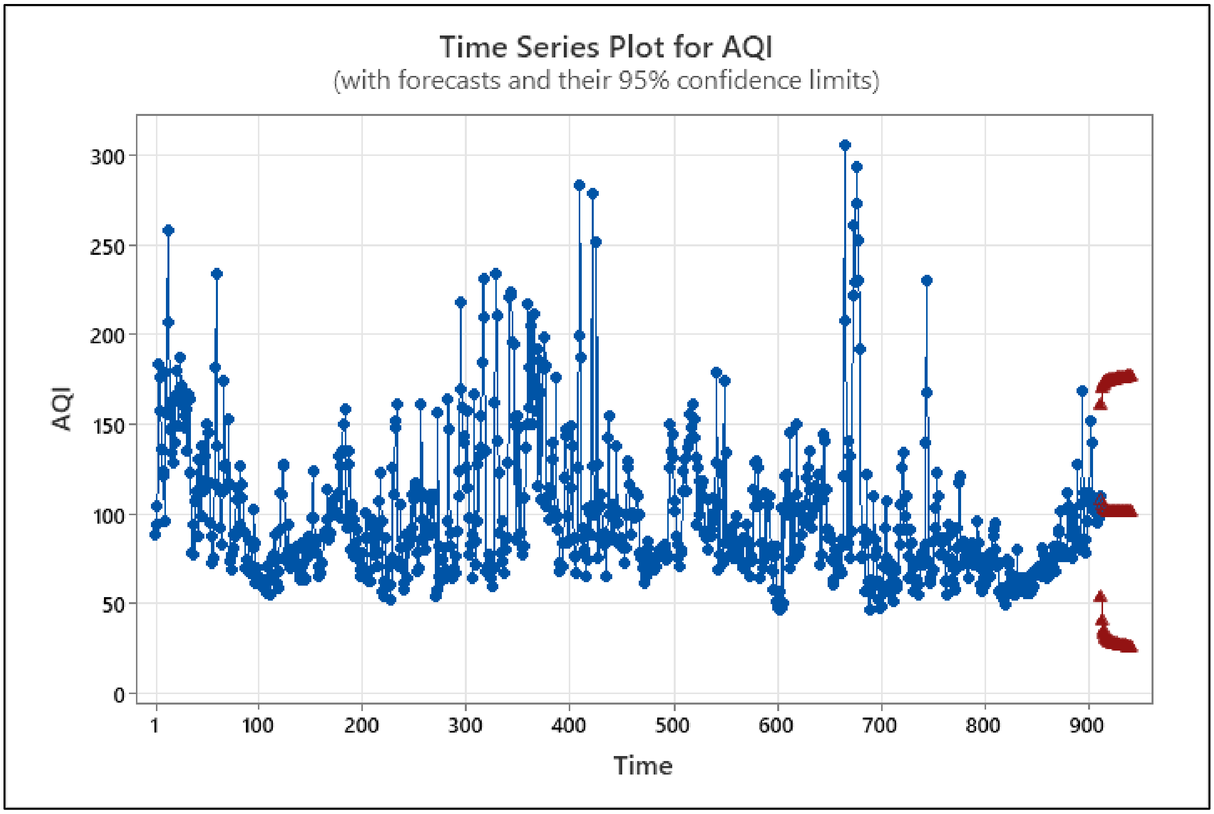



Now that we have fit the best possible model with all the assumptions being satisfied, we move to our main purpose of forecasting.

## Forecast value from observation 912 (for next 30 days)

| Observations | Forecast | 95% Limits | |
|---|---|---|---|
| | | Lower | Upper |
| 913 | 106.927 | 53.4648 | 160.389 |
| 914 | 104.743 | 39.6988 | 169.787 |
| 915 | 103.019 | 34.4631 | 171.575 |
| 916 | 102.246 | 32.1861 | 172.306 |
| 917 | 101.984 | 30.6547 | 173.313 |
| 918 | 101.558 | 29.4801 | 173.636 |
| 919 | 101.183 | 28.8185 | 173.548 |

| | | | |
|---|---|---|---|
| 920 | 101.159 | 28.5351 | 173.783 |
| 921 | 101.175 | 28.2163 | 174.134 |
| 922 | 101.016 | 27.8368 | 174.195 |
| 923 | 100.919 | 27.6064 | 174.231 |
| 924 | 100.973 | 27.4881 | 174.458 |
| 925 | 100.972 | 27.2894 | 174.654 |
| 926 | 100.884 | 27.0617 | 174.707 |
| 927 | 100.863 | 26.9215 | 174.805 |
| 928 | 100.897 | 26.8082 | 174.986 |
| 929 | 100.875 | 26.6401 | 175.11 |
| 930 | 100.826 | 26.4759 | 175.175 |
| 931 | 100.825 | 26.3625 | 175.287 |
| 932 | 100.835 | 26.2468 | 175.424 |
| 933 | 100.808 | 26.1037 | 175.513 |
| 934 | 100.782 | 25.9759 | 175.588 |
| 935 | 100.783 | 25.874 | 175.692 |
| 936 | 100.779 | 25.7634 | 175.794 |
| 937 | 100.756 | 25.6432 | 175.868 |
| 938 | 100.741 | 25.5382 | 175.944 |
| 939 | 100.739 | 25.4445 | 176.033 |
| 940 | 100.728 | 25.344 | 176.113 |
| 941 | 100.711 | 25.2429 | 176.178 |
| 942 | 100.701 | 25.1532 | 176.248 |




Time series plot after fitting ARIMA model. It shows AQI of Chennai city since 2018 (blue line) and forecasts and their 95% confidence limits (red line) for next month. We can observe that AQI values remains approximately constant in the forecasted values.

## Model Summary for checking model efficiency

**Model Description**

| | | | Model Type |
|---|---|---|---|
| Model ID | AQI | Model_1 | ARIMA(4,0,3) |

### Model Summary

**Model Fit**

| Fit Statistic | Mean | SE | Minimum | Maximum | Percentile 5 | 10 | 25 | 50 | 75 | 90 | 95 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Stationary R-squared | .524 | . | .524 | .524 | .524 | .524 | .524 | .524 | .524 | .524 | .524 |
| R-squared | .524 | . | .524 | .524 | .524 | .524 | .524 | .524 | .524 | .524 | .524 |
| RMSE | 27.230 | . | 27.230 | 27.230 | 27.230 | 27.230 | 27.230 | 27.230 | 27.230 | 27.230 | 27.230 |
| MAPE | 17.869 | . | 17.869 | 17.869 | 17.869 | 17.869 | 17.869 | 17.869 | 17.869 | 17.869 | 17.869 |
| MaxAPE | 138.373 | . | 138.373 | 138.373 | 138.373 | 138.373 | 138.373 | 138.373 | 138.373 | 138.373 | 138.373 |
| MAE | 17.977 | . | 17.977 | 17.977 | 17.977 | 17.977 | 17.977 | 17.977 | 17.977 | 17.977 | 17.977 |
| MaxAE | 193.463 | . | 193.463 | 193.463 | 193.463 | 193.463 | 193.463 | 193.463 | 193.463 | 193.463 | 193.463 |
| Normalized BIC | 6.668 | . | 6.668 | 6.668 | 6.668 | 6.668 | 6.668 | 6.668 | 6.668 | 6.668 | 6.668 |

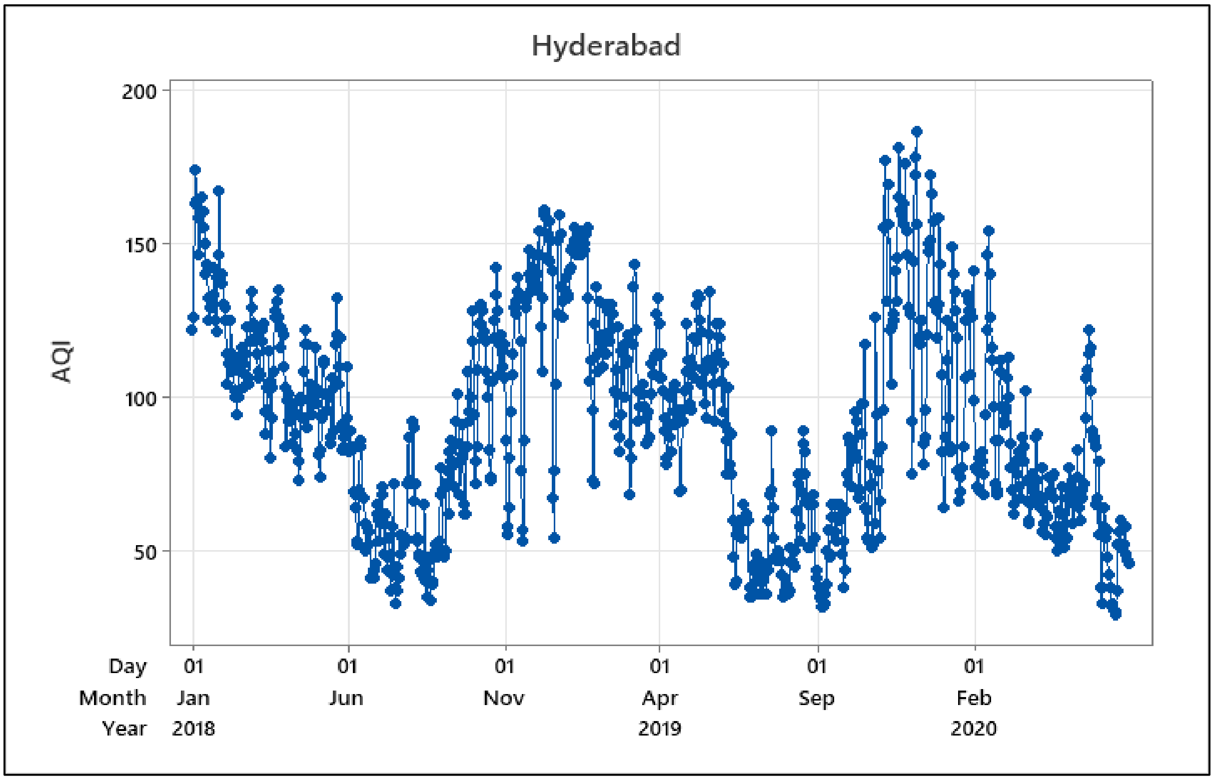
**Model Statistics**

| Model | Number of Predictors | Model Fit statistics: Stationary R-squared | R-squared | RMSE | MAPE | Ljung-Box Q(18): Statistics | DF | Sig. | Number of Outliers |
|---|---|---|---|---|---|---|---|---|---|
| AQI-Model_1 | 0 | .524 | .524 | 27.230 | 17.869 | 15.372 | 11 | .166 | 0 |

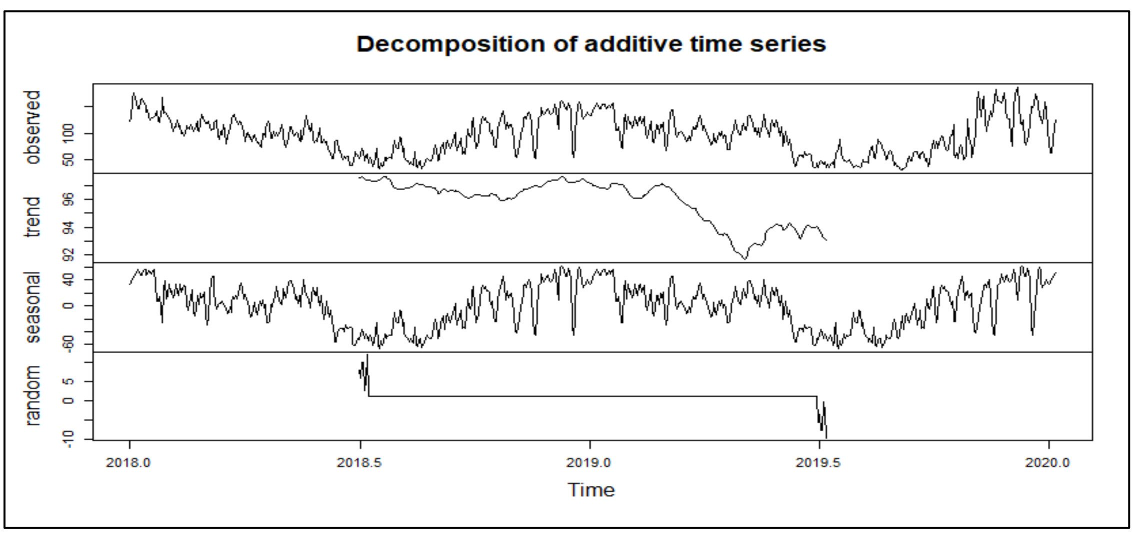
Since the MAPE value is 17.869, on an average, the forecast is off by 17.87%

# For city Hyderabad

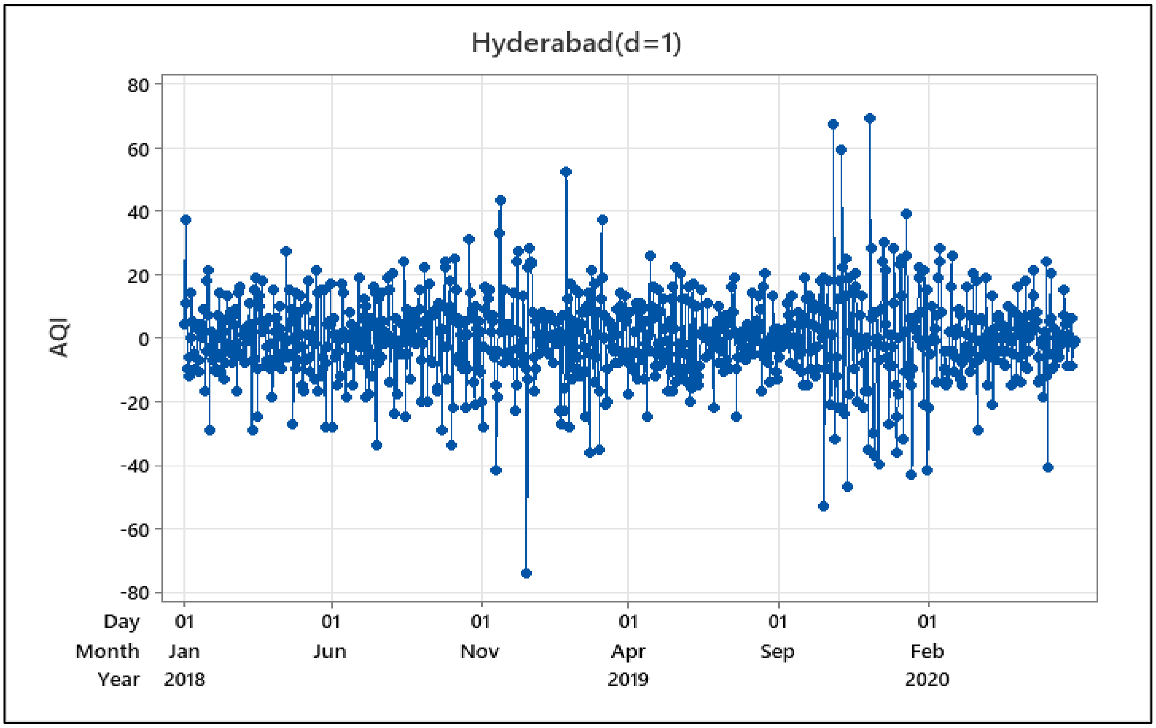
Time Series plot for AQI of city Hyderabad from 01/01/2018 to 30/06/2020.




It is a Time Series plot for city Hyderabad with days along the x-axis and AQI figures on the y-axis. This time plot suggests that the time series is not stationary. Also, by using decomposition.



We can see there is decreasing trend and again it increases. This suggests that the time series is not stationary and will require differencing to make it stationary, at least a difference order of 1.



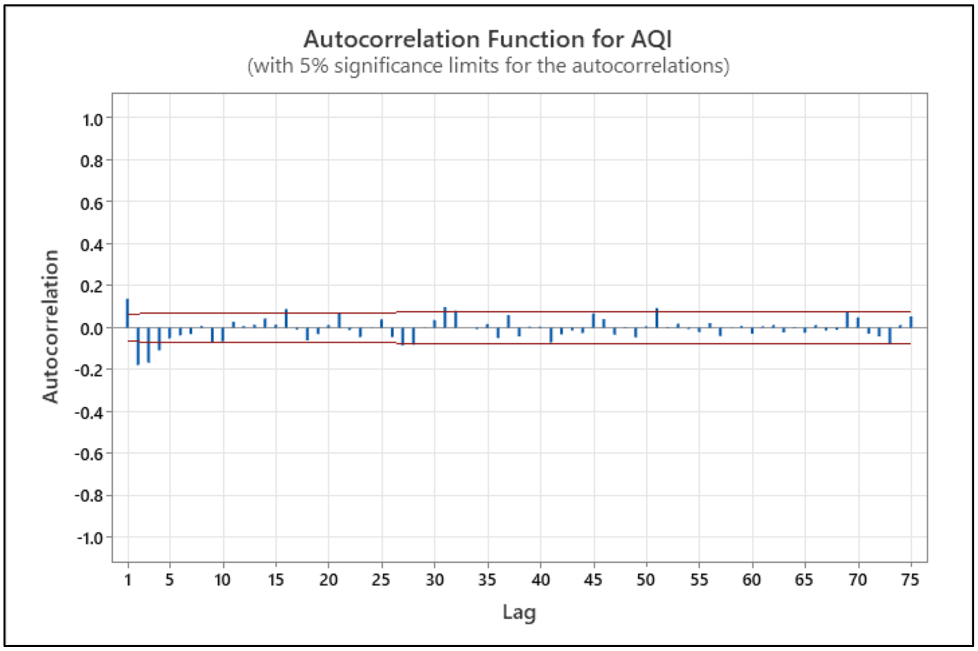

It is a Time Series plot of AQI data differenced for lag 1. Here we can see our time series is now stationary.

## Estimation of Parameters (p, d, q)

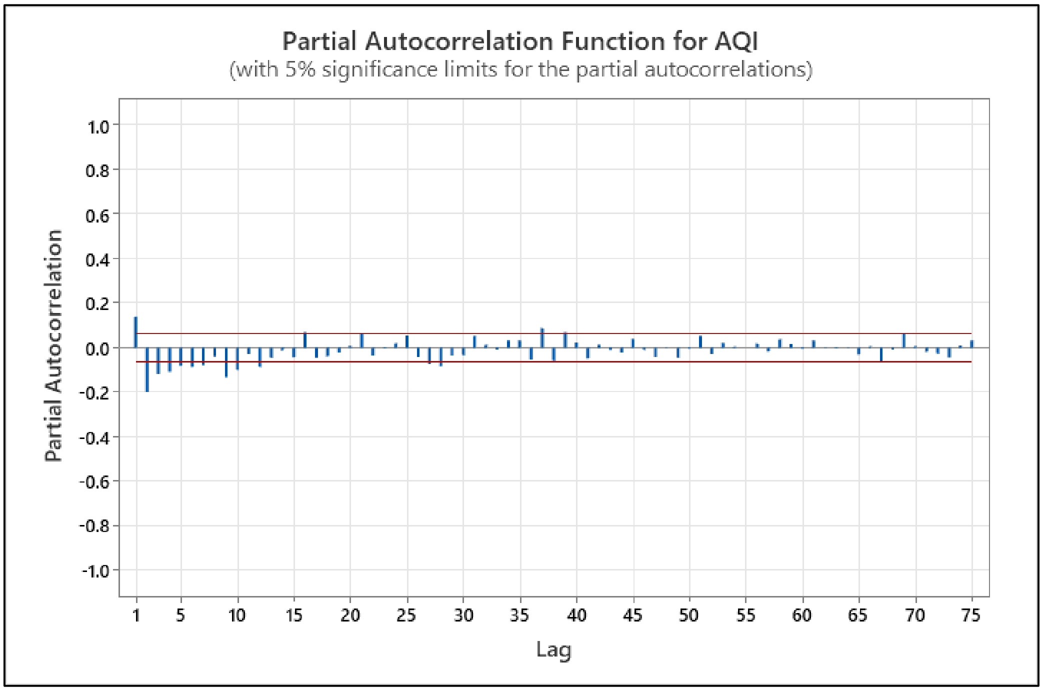
Since our time series is non stationary, we take first difference. After taking first difference our time series being stationary. Hence, here we take d=1.

Now we have to plot the ACF and PACF plots for the estimation of parameters (p, q).

## Plot of ACF and PACF:




From ACF plot after taking first difference we see that it is significant to lag3. So, we consider q as 3.



From PACF plot after taking first difference we can see that it is significant up to lag 1. So, we consider p as 1.

Hence, we fit various models for AQI. Finally we consider p=1 and q=2, the best fit model we found is ARIMA(1,1,2) with constant.

The final estimates of the parameters are given below.

## Final Estimates of Parameters :

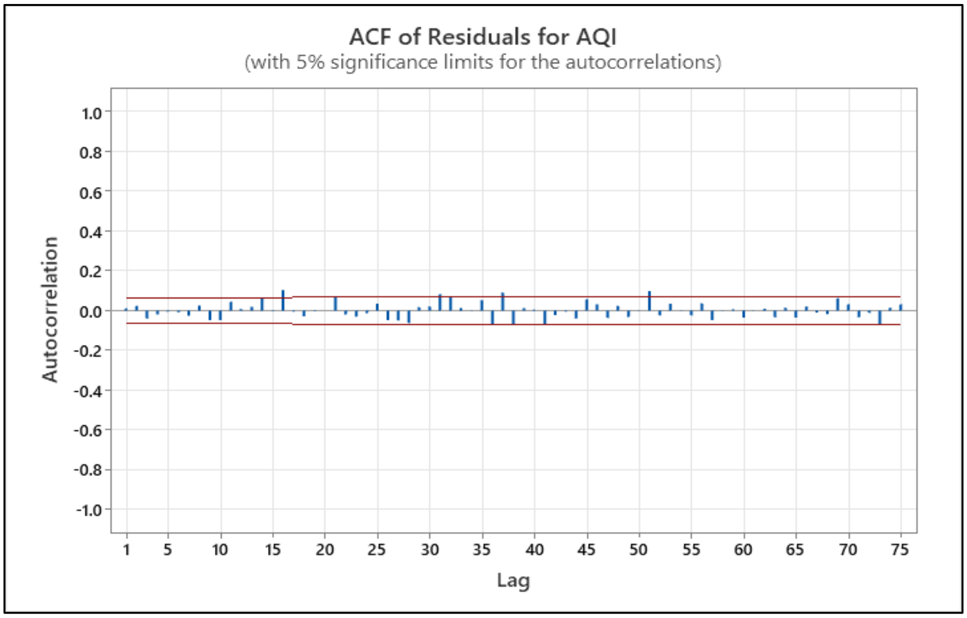
| Type | Coeff | SE Coeff | T-Value | P-Value |
|---|---|---|---|---|
| AR 1 | 0.6373 | 0.046 | 13.85 | 0.00 |
| MA 1 | 0.5728 | 0.0486 | 11.78 | 0.00 |
| MA 2 | 0.3138 | 0.0355 | 8.83 | 0.00 |

The significance of the parameters is tested using t-test with p-value very small for all. This indicates that all the parameters are significant.

## Residual Sums of Squares

| DF | SS | MS |
|---|---|---|
| 907 | 142021 | 156.584 |

The residuals must not be auto correlated and this is apparent below with autocorrelation being insignificant for all lags.



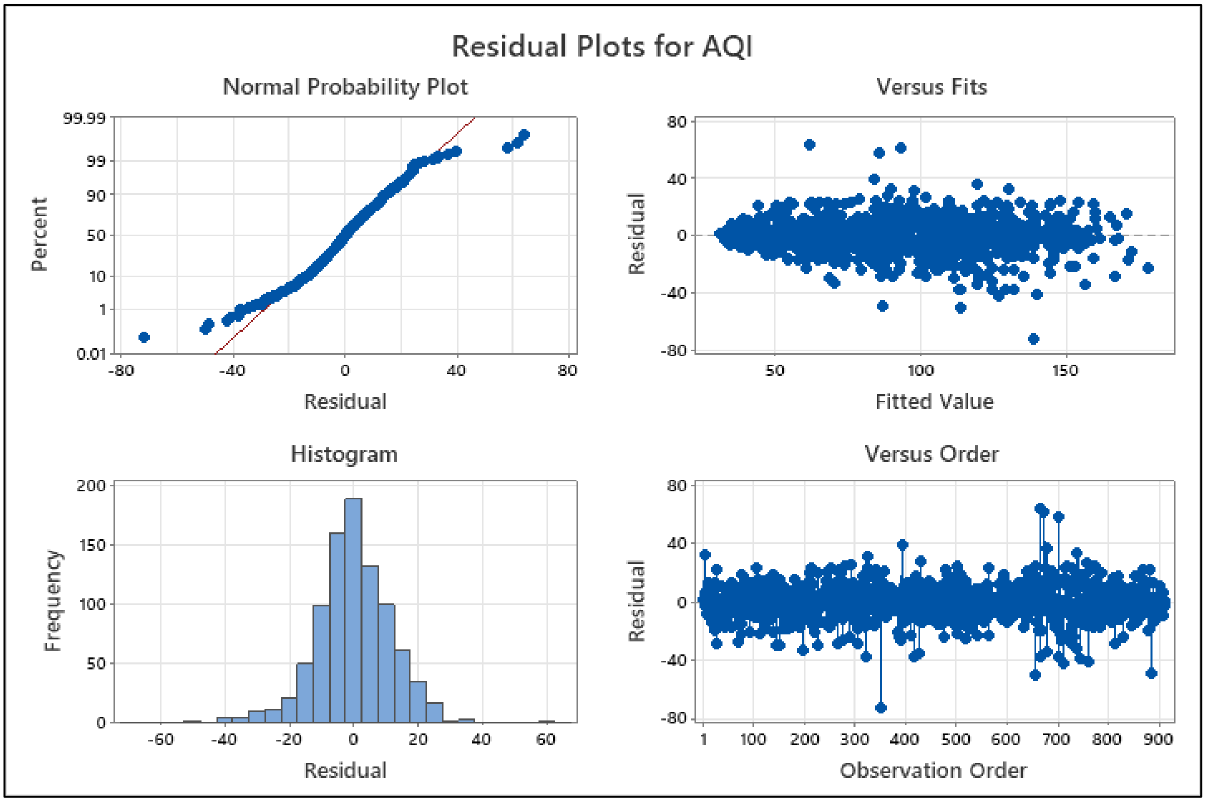

The significance of the autocorrelation among the residuals at four lags is tested using Ljung-Box Chi-Square Statistics and output is shown below.

## Modified Box-Pierce (Ljung-Box) Chi-Square Statistic

| **Lag** | **12** | **24** | **36** | **48** |
|---|---|---|---|---|
| Chi-Square | 9.73 | 30.31 | 58.88 | 83.69 |
| DF | 8 | 20 | 32 | 44 |
| P-Value | 0.285 | 0.065 | 0.053 | 0.042 |

The p-values for the test at all the four lags are large pointing towards the insignificance of autocorrelation among residuals.

Also, the residuals are normally distributed which can be observed in the NPP and histogram. The error has constant variance which is evident from the graph of residual against the fitted values.

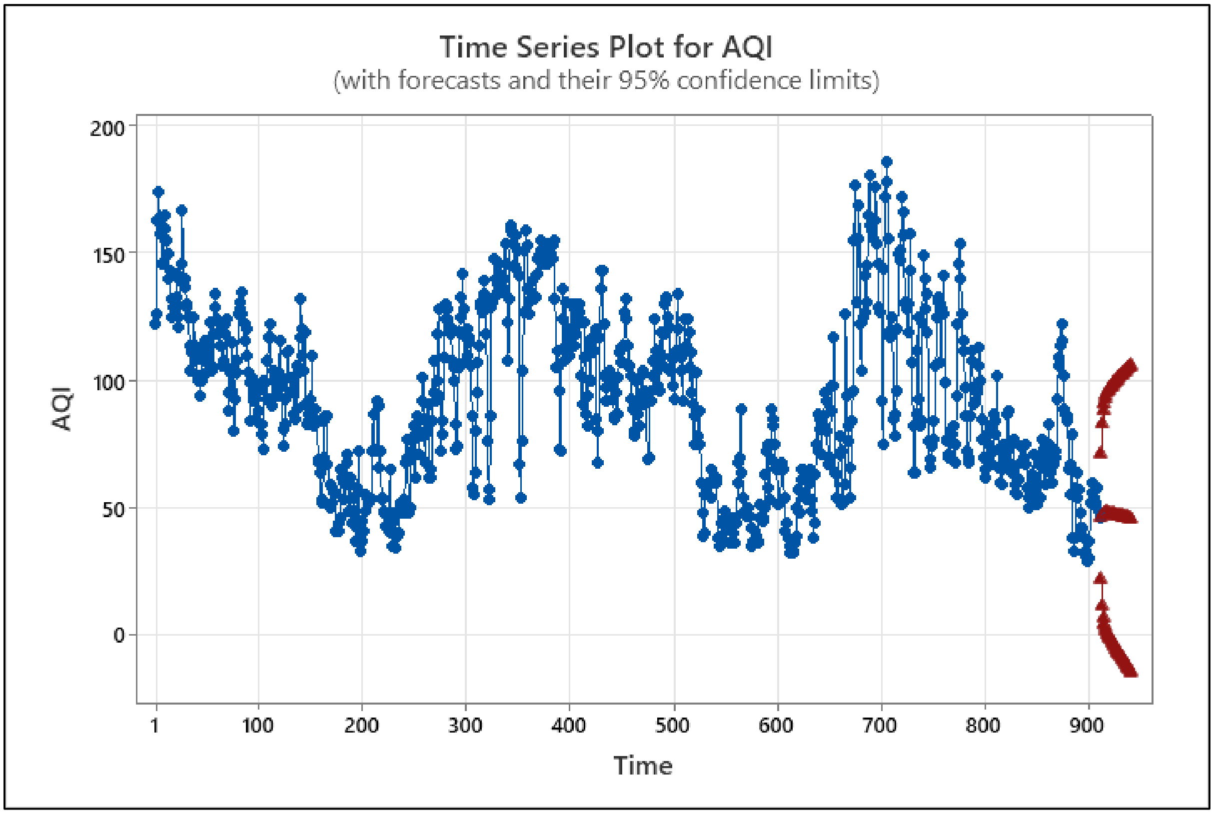



Now that we have fit the best possible model with all the assumptions being satisfied, we move to our main purpose of forecasting.

## Forecast value from observation 912 (for next 30 days)

| Observations | Forecast | 95% Limits | |
|---|---|---|---|
| | | **Lower** | **Upper** |
| 913 | 46.1767 | 21.6456 | 70.708 |
| 914 | 46.8638 | 11.0358 | 82.692 |
| 915 | 47.2645 | 6.5105 | 88.019 |
| 916 | 47.4827 | 4.0007 | 90.965 |
| 917 | 47.5846 | 2.3575 | 92.812 |
| 918 | 47.6123 | 1.1344 | 94.090 |
| 919 | 47.5928 | 0.1324 | 95.053 |
| 920 | 47.5432 | -0.7451 | 95.831 |

| | | | |
|---|---|---|---|
| 921 | 47.4743 | -1.5486 | 96.497 |
| 922 | 47.3933 | -2.3057 | 97.092 |
| 923 | 47.3045 | -3.0323 | 97.641 |
| 924 | 47.2107 | -3.7378 | 98.159 |
| 925 | 47.1137 | -4.4280 | 98.656 |
| 926 | 47.0148 | -5.1065 | 99.136 |
| 927 | 46.9145 | -5.7755 | 99.604 |
| 928 | 46.8134 | -6.4364 | 100.063 |
| 929 | 46.7118 | -7.0903 | 100.514 |
| 930 | 46.6099 | -7.7378 | 100.958 |
| 931 | 46.5077 | -8.3794 | 101.395 |
| 932 | 46.4054 | -9.0155 | 101.826 |
| 933 | 46.3031 | -9.6462 | 102.252 |
| 934 | 46.2007 | -10.2720 | 102.673 |
| 935 | 46.0982 | -10.8928 | 103.089 |
| 936 | 45.9957 | -11.5089 | 103.500 |
| 937 | 45.8932 | -12.1204 | 103.907 |
| 938 | 45.7907 | -12.7275 | 104.309 |
| 939 | 45.6882 | -13.3303 | 104.707 |
| 940 | 45.5857 | -13.9288 | 105.100 |
| 941 | 45.4832 | -14.5233 | 105.490 |
| 942 | 45.3807 | -15.1137 | 105.875 |




Time series plot after fitting ARIMA model. It shows AQI of Hyderabad city since 2018 (blue line) and forecasts and their 95% confidence limits (red line) for next month. On close observation we can see that AQI value slightly decrease in the forecasted values.

## Model Summary for checking model efficiency

**Model Description**

| | | | Model Type |
|---|---|---|---|
| Model ID | AQI | Model_1 | ARIMA(1,1,2) |

# Model Summary

**Model Fit**

| Fit Statistic | Mean | SE | Minimum | Maximum | Percentile 5 | 10 | 25 | 50 | 75 | 90 | 95 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Stationary R-squared | .122 | . | .122 | .122 | .122 | .122 | .122 | .122 | .122 | .122 | .122 |
| R-squared | .868 | . | .868 | .868 | .868 | .868 | .868 | .868 | .868 | .868 | .868 |
| RMSE | 12.525 | . | 12.525 | 12.525 | 12.525 | 12.525 | 12.525 | 12.525 | 12.525 | 12.525 | 12.525 |
| MAPE | 10.944 | . | 10.944 | 10.944 | 10.944 | 10.944 | 10.944 | 10.944 | 10.944 | 10.944 | 10.944 |
| MaxAPE | 128.396 | . | 128.396 | 128.396 | 128.396 | 128.396 | 128.396 | 128.396 | 128.396 | 128.396 | 128.396 |
| MAE | 9.091 | . | 9.091 | 9.091 | 9.091 | 9.091 | 9.091 | 9.091 | 9.091 | 9.091 | 9.091 |
| MaxAE | 71.892 | . | 71.892 | 71.892 | 71.892 | 71.892 | 71.892 | 71.892 | 71.892 | 71.892 | 71.892 |
| Normalized BIC | 5.085 | . | 5.085 | 5.085 | 5.085 | 5.085 | 5.085 | 5.085 | 5.085 | 5.085 | 5.085 |

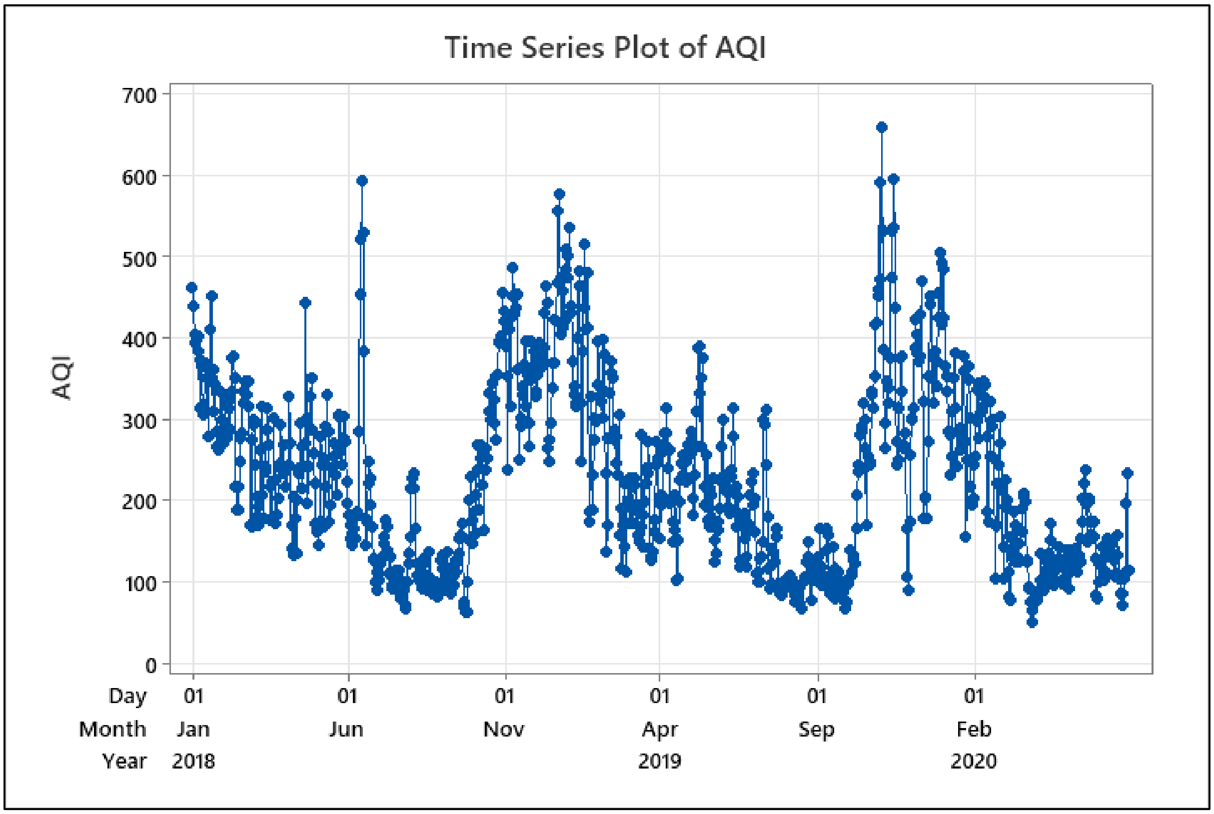
**Model Statistics**

| Model | Number of Predictors | Model Fit statistics: Stationary R-squared | R-squared | RMSE | MAPE | Ljung-Box Q(18): Statistics | DF | Sig. | Number of Outliers |
|---|---|---|---|---|---|---|---|---|---|
| AQI-Model_1 | 0 | .122 | .868 | 12.525 | 10.944 | 24.326 | 15 | .060 | 0 |

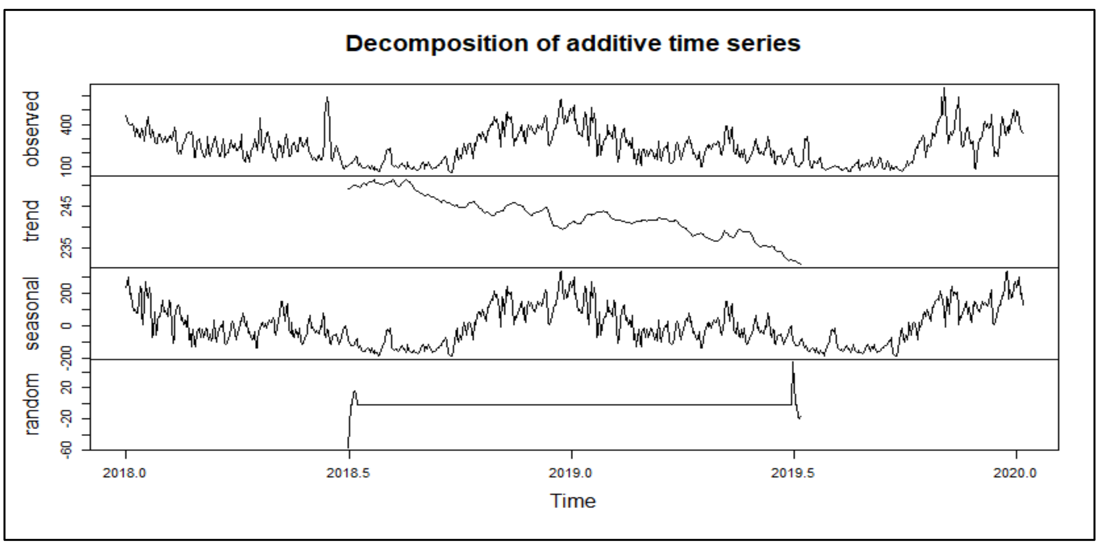
Since the MAPE value is 10.944, on an average, the forecast is off by 10.94%

## For city Delhi

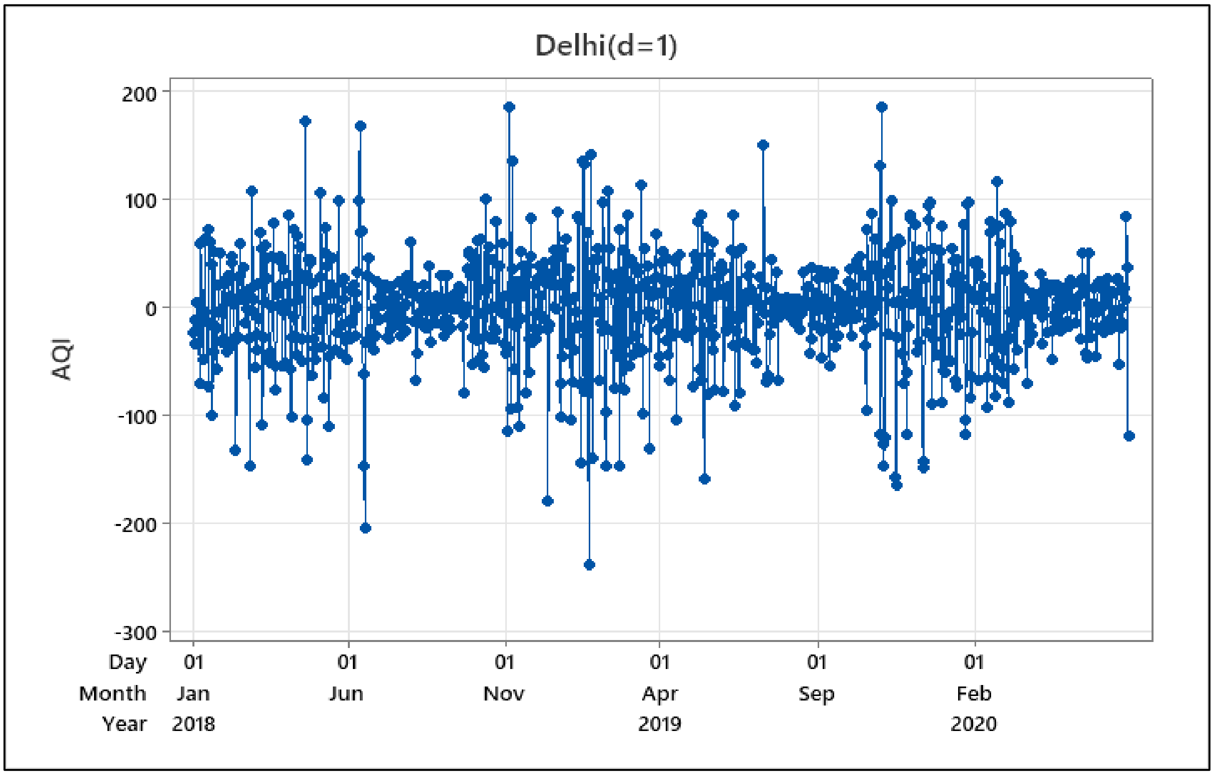
Time Series plot for AQI of city Delhi from 01/01/2018 to 30/06/2020.




It is a Time Series plot for city Delhi with days along the x-axis and AQI figures on the y-axis. This time plot suggests that the time series is not stationary. Also, by using decomposition.



We can see there is decreasing trend. This suggests that the time series is not stationary and will require differencing to make it stationary, at least a difference order of 1.



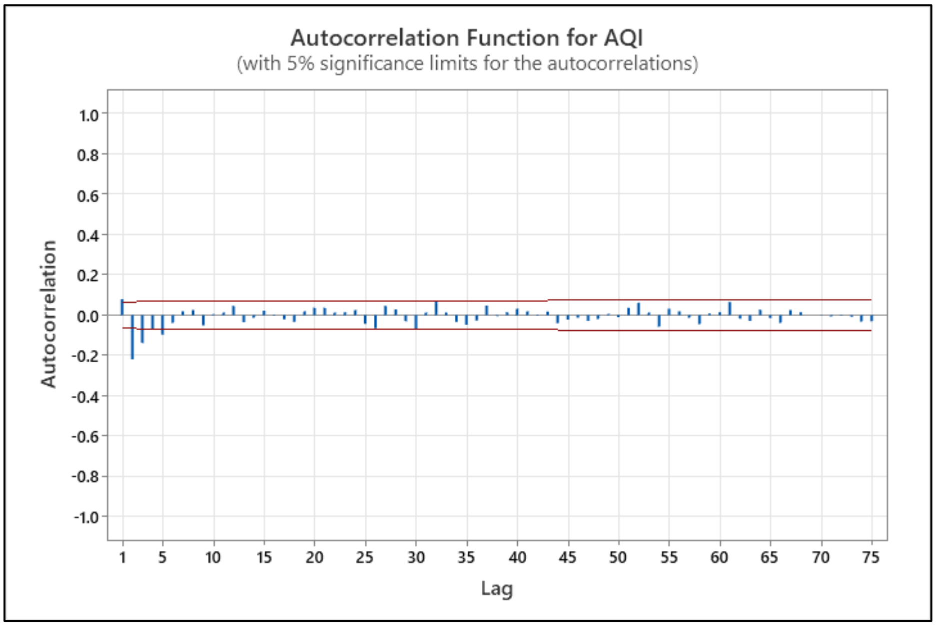

It is a Time Series plot of AQI data differenced for lag 1. Here we can see our time series is now stationary.

## Estimation of Parameters (p, d, q)

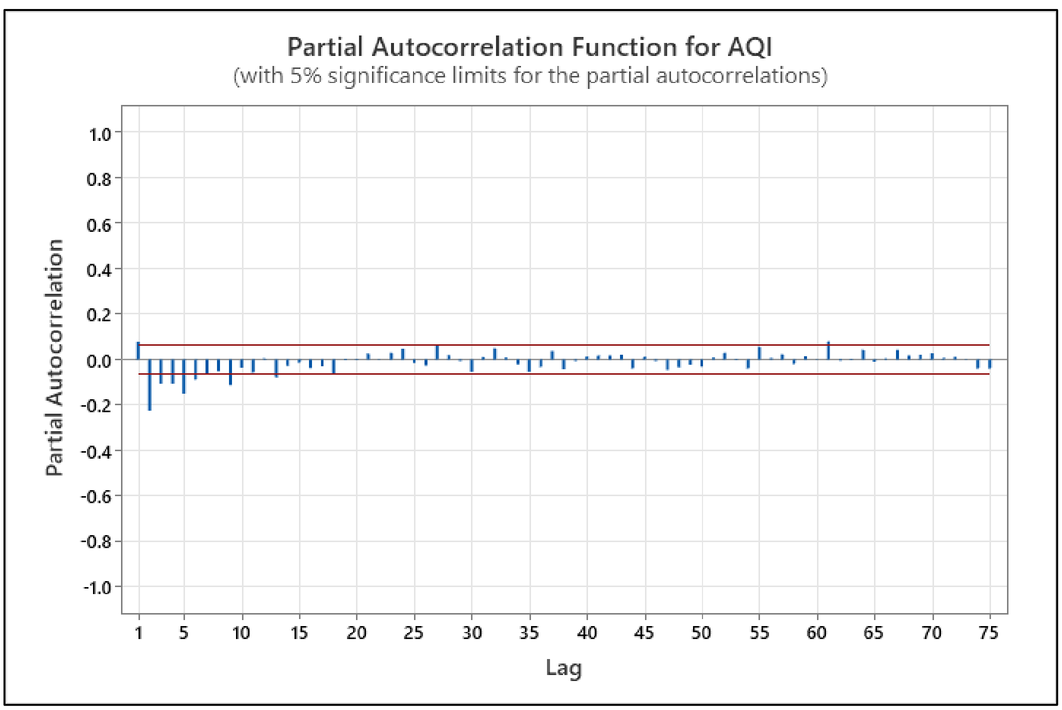
Since our time series is non stationary, we took first difference. After taking first difference our time series being stationary. Hence, here d=1.

Now we have to plot the ACF and PACF plots for the estimation of parameters (p, q).

# Plot of ACF and PACF:




From ACF plot after taking first difference we can see that it is significant till lag 3. So, we consider q as 2.



From PACF plot after taking first difference, we can see it is significant up to some lags. To fit the model we consider p as 1.

Hence, we fit various models for AQI. Finally, we consider p=1 and q=2, the best fit model we found is ARIMA(1,1,2) with constant. The final estimates of the parameters are given below.

## Final Estimates of Parameters :

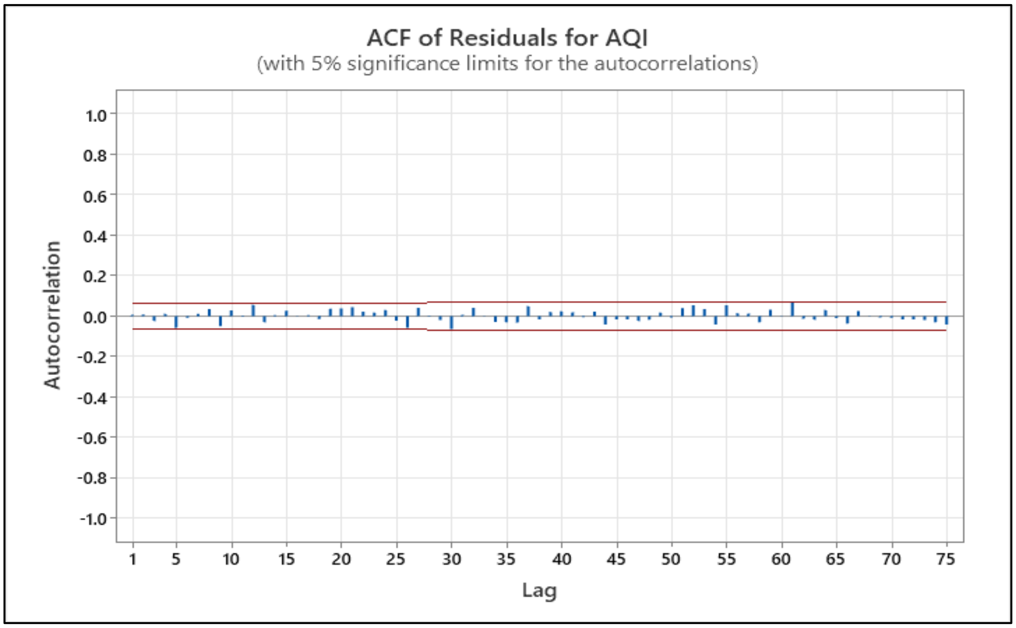
| Type | Coeff | SE Coeff | T-Value | P-Value |
|---|---|---|---|---|
| AR1 | 0.5567 | 0.0527 | 10.56 | 0.00 |
| MA1 | 0.5564 | 0.0541 | 10.29 | 0.00 |
| MA2 | 0.3052 | 0.0374 | 8.16 | 0.00 |

The significance of the parameters is tested using t-test with p-value very small for all. This indicates that all the parameters are significant.

## Residual Sums of Squares

| DF | SS | MS |
|---|---|---|
| 904 | 1911842 | 2107.87 |

The residuals must not be auto correlated and this is apparent below with autocorrelation being insignificant for all lags.



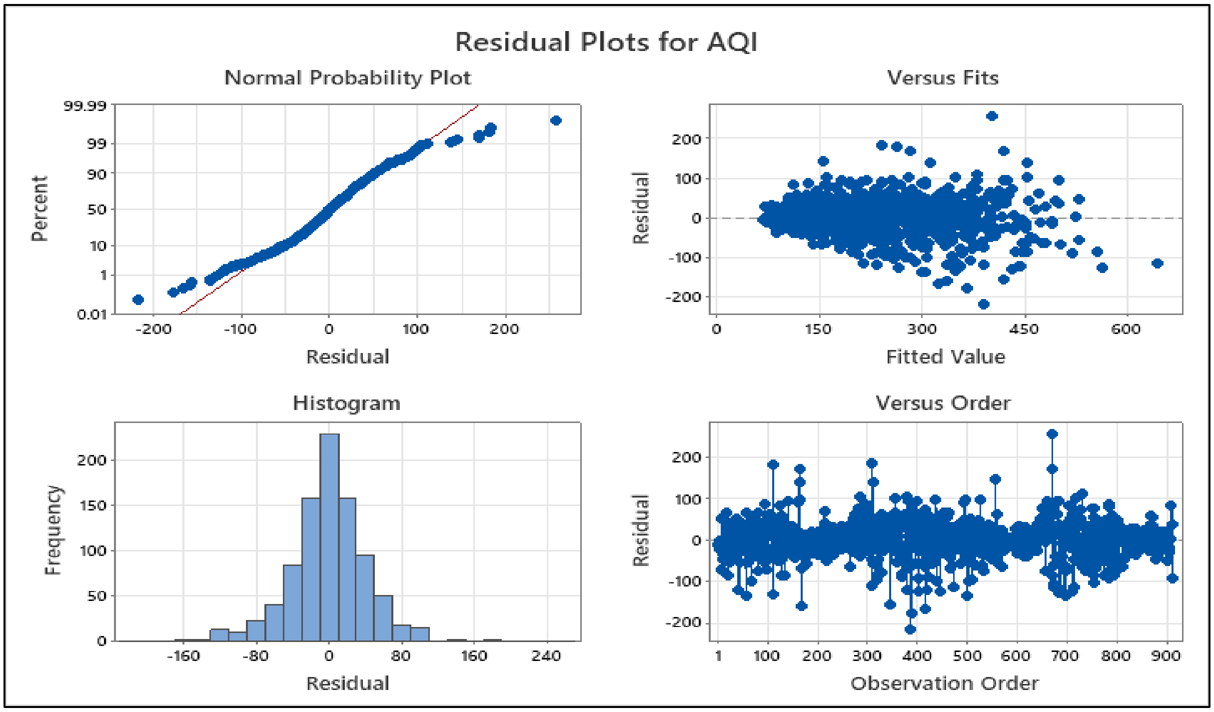

The significance of the autocorrelation among the residuals at four lags is tested using Ljung-Box Chi-Square Statistics and output is shown below.

## Modified Box-Pierce (Ljung-Box) Chi-Square Statistic

| **Lag** | **12** | **24** | **36** | **48** |
|---|---|---|---|---|
| Chi square | 10.76 | 18.60 | 32.11 | 39.42 |
| DF | 8 | 20 | 32 | 44 |
| P-Value | 0.215 | 0.548 | 0.461 | 0.668 |

The p-values for the test at all the four lags are large pointing towards the insignificance of autocorrelation among residuals.

Also, the residuals are normally distributed which can be observed in the NPP and histogram. The error has constant variance which is evident from the graph of residual against the fitted values.

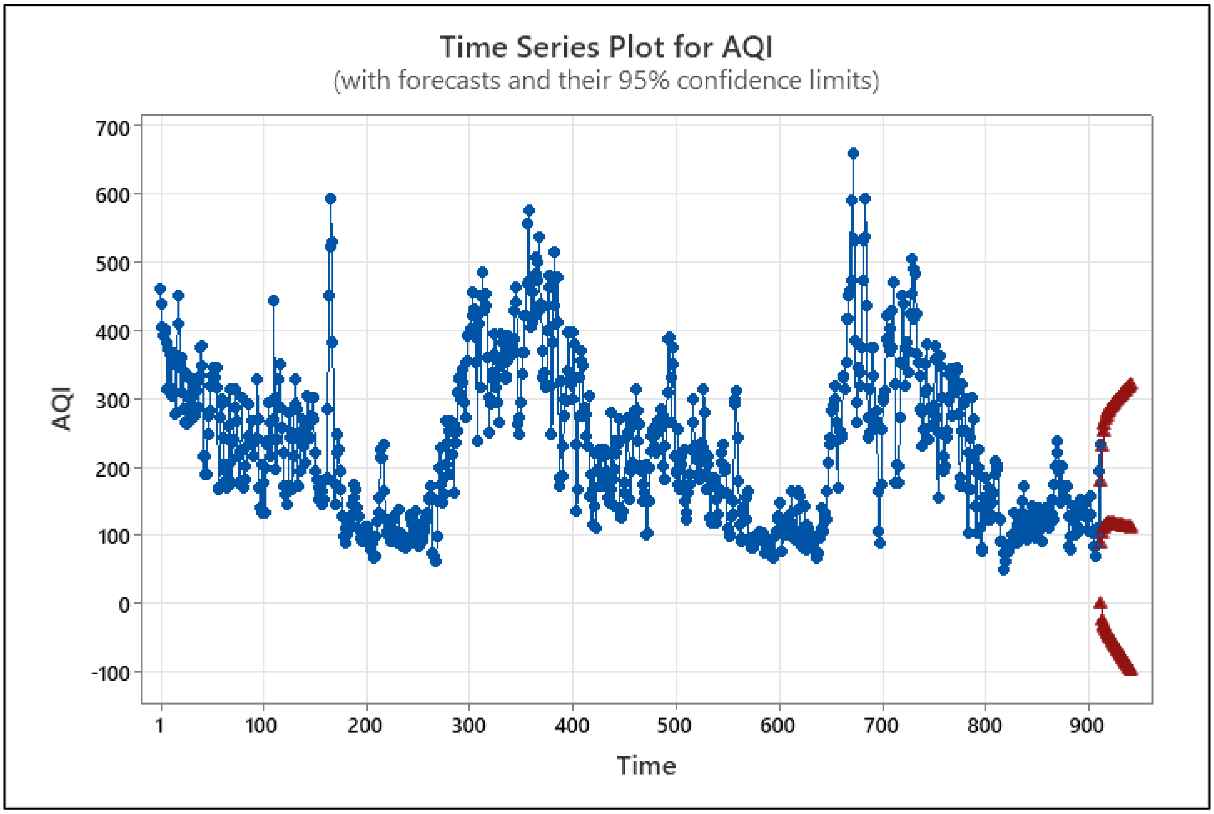



Now that we have fit the best possible model with all the assumptions being satisfied, we move to our main purpose of forecasting.

## Forecast value from observation 912 (for next 30 days)

| Observations | Forecast | 95% Limits | |
|---|---|---|---|
| | | Lower | Upper |
| 913 | 87.603 | -2.402 | 177.608 |
| 914 | 101.079 | -26.229 | 228.387 |
| 915 | 108.446 | -33.411 | 250.303 |
| 916 | 112.412 | -37.121 | 261.946 |
| 917 | 114.485 | -39.996 | 268.966 |
| 918 | 115.504 | -42.685 | 273.692 |
| 919 | 115.936 | -45.341 | 277.212 |
| 920 | 116.041 | -47.991 | 280.073 |

| | | | |
|---|---|---|---|
| 921 | 115.965 | -50.633 | 282.562 |
| 922 | 115.787 | -53.26 | 284.833 |
| 923 | 115.552 | -55.865 | 286.97 |
| 924 | 115.287 | -58.447 | 289.021 |
| 925 | 115.004 | -61.003 | 291.011 |
| 926 | 114.711 | -63.533 | 292.955 |
| 927 | 114.413 | -66.037 | 294.862 |
| 928 | 114.111 | -68.515 | 296.738 |
| 929 | 113.809 | -70.968 | 298.585 |
| 930 | 113.505 | -73.397 | 300.406 |
| 931 | 113.2 | -75.802 | 302.202 |
| 932 | 112.896 | -78.184 | 303.975 |
| 933 | 112.591 | -80.543 | 305.725 |
| 934 | 112.286 | -82.881 | 307.453 |
| 935 | 111.981 | -85.198 | 309.16 |
| 936 | 111.676 | -87.495 | 310.847 |
| 937 | 111.371 | -89.772 | 312.514 |
| 938 | 111.066 | -92.03 | 314.162 |
| 939 | 110.761 | -94.269 | 315.791 |
| 940 | 110.456 | -96.491 | 317.402 |
| 941 | 110.151 | -98.694 | 318.996 |
| 942 | 109.846 | -100.881 | 320.572 |




Time series plot after fitting ARIMA model. It shows AQI of Delhi city since 2018 (blue line) and forecasts and their 95% confidence limits (red line) for next month. On close observation we can see that AQI values slightly increases and again slightly decreases in the forecasted values.

## Model Summary for checking model efficiency

**Model Description**

| | | | Model Type |
|---|---|---|---|
| Model ID | AQI | Model_1 | ARIMA(1,1,2) |

**Model Summary**

**Model Fit**

| Fit Statistic | Mean | SE | Minimum | Maximum | Percentile 5 | 10 | 25 | 50 | 75 | 90 | 95 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Stationary R-squared | .123 | . | .123 | .123 | .123 | .123 | .123 | .123 | .123 | .123 | .123 |
| R-squared | .837 | . | .837 | .837 | .837 | .837 | .837 | .837 | .837 | .837 | .837 |
| RMSE | 45.935 | . | 45.935 | 45.935 | 45.935 | 45.935 | 45.935 | 45.935 | 45.935 | 45.935 | 45.935 |
| MAPE | 15.741 | . | 15.741 | 15.741 | 15.741 | 15.741 | 15.741 | 15.741 | 15.741 | 15.741 | 15.741 |
| MaxAPE | 124.270 | . | 124.270 | 124.270 | 124.270 | 124.270 | 124.270 | 124.270 | 124.270 | 124.270 | 124.270 |
| MAE | 32.974 | . | 32.974 | 32.974 | 32.974 | 32.974 | 32.974 | 32.974 | 32.974 | 32.974 | 32.974 |
| MaxAE | 256.694 | . | 256.694 | 256.694 | 256.694 | 256.694 | 256.694 | 256.694 | 256.694 | 256.694 | 256.694 |
| Normalized BIC | 7.684 | . | 7.684 | 7.684 | 7.684 | 7.684 | 7.684 | 7.684 | 7.684 | 7.684 | 7.684 |

**Model Statistics**

| Model | Number of Predictors | Model Fit statistics: Stationary R-squared | R-squared | RMSE | MAPE | Ljung-Box Q(18): Statistics | DF | Sig. | Number of Outliers |
|---|---|---|---|---|---|---|---|---|---|
| AQI-Model_1 | 0 | .123 | .837 | 45.935 | 15.741 | 12.422 | 15 | .647 | 0 |

Since the MAPE value is 15.741, on an average, the forecast is off by 15.74%

# Objective 3

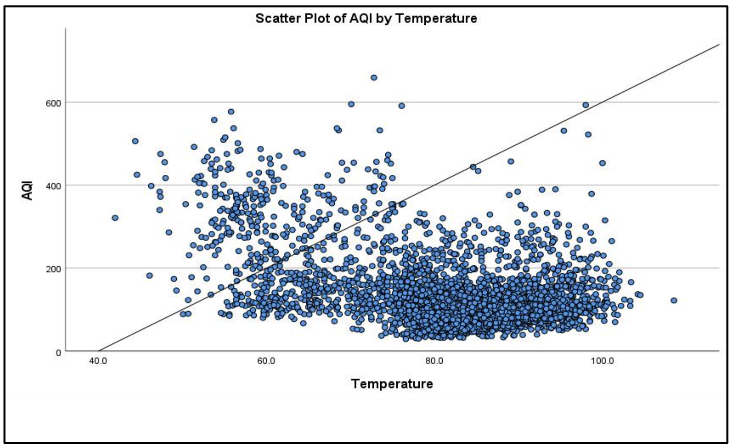
To find association between Air Quality Index and other factors such as temperature, humidity, dew point, wind speed and pressure.

# To check Linearity

To check linearity we use scatter plot.

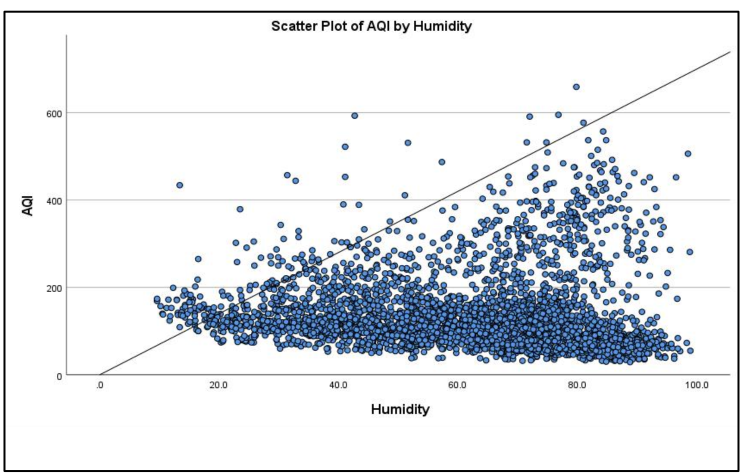
A **scatter plot** (also called a scatterplot, scatter graph, scatter chart, scattergram, or scatter diagram) is a type of plot or mathematical diagram using Cartesian Coordinates to display values for typically two variables for a set of data.

## For AQI and Average Temperature



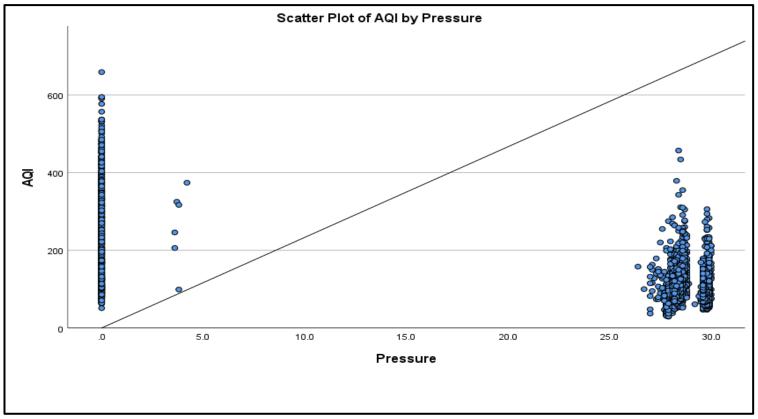

There is no linear relationship between AQI and Average Temperature.

## For AQI and Average Humidity



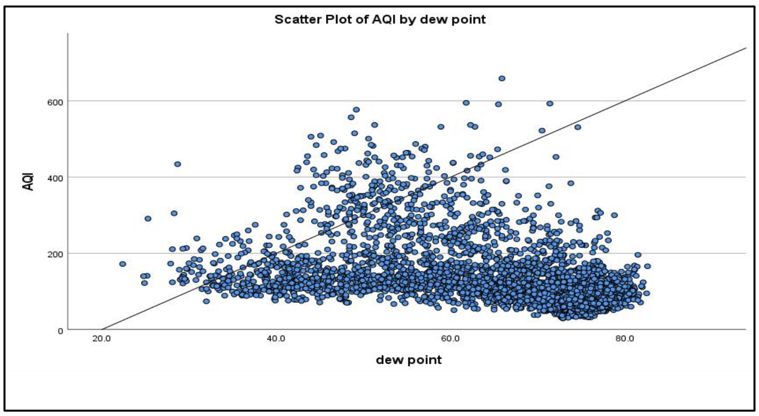

There is no linear relationship between AQI and Average Humidity.

## For AQI and Average Pressure



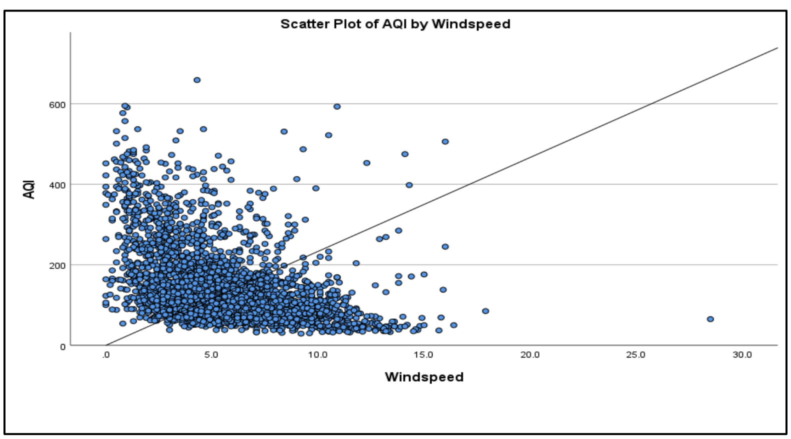

There is no linear relationship between AQI and Average Pressure.

## For AQI and Average Dew Point




There is no linear relationship between AQI and Average Dew point.

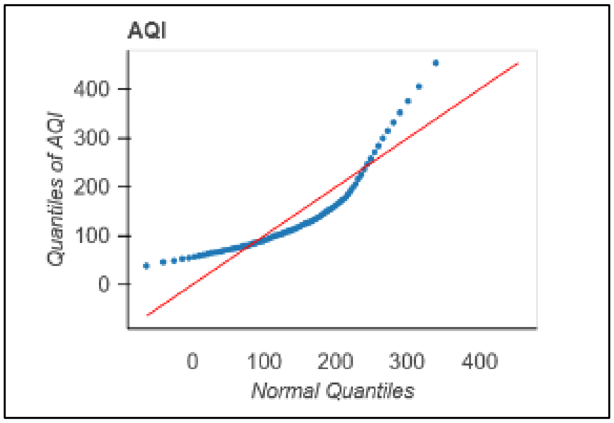
## For AQI and Average Windspeed

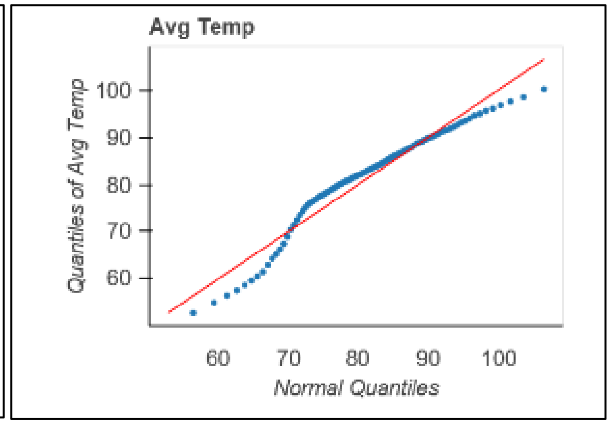


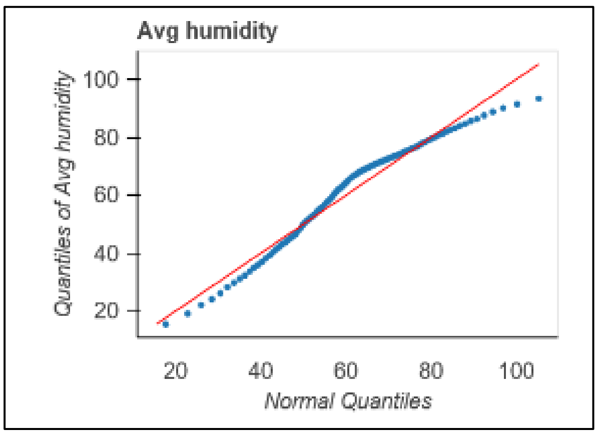

There is no linear relationship between AQI and Average Windspeed.

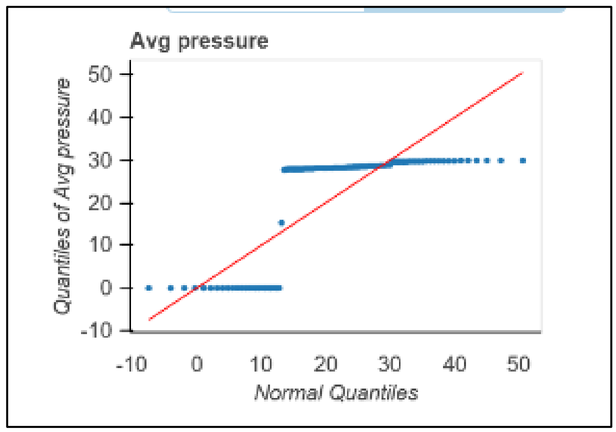
## To check Normality

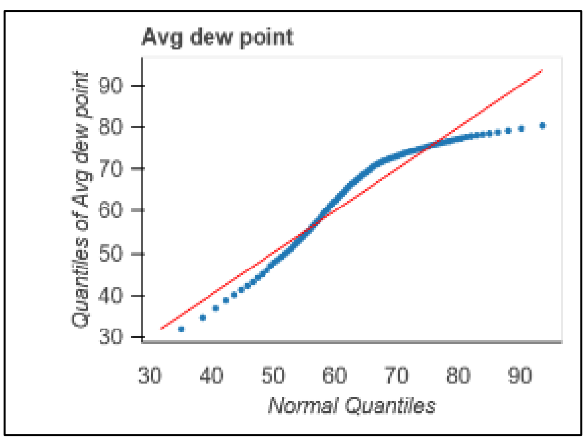
To check normality we use Q-Q plot.

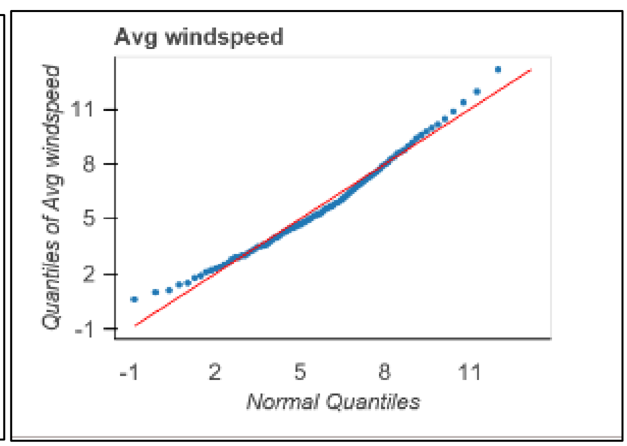



Also, by using Shapiro-Wilcoxon test and Kolmogorov-Smirnov test,

**Hypothesis:**

$H_0$: Data follows Normal Distribution.

$H_1$: Data does not follow Normal Distribution.

**Using SPSS,**

**Tests of Normality**

| | Kolmogorov-Smirnov[a] | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|
| | Statistic | df | Sig. | Statistic | df | Sig. |
| AQI | .171 | 3647 | <.001 | .802 | 3647 | <.001 |
| Avg Temp | .101 | 3647 | <.001 | .944 | 3647 | <.001 |
| Avg dew point | .135 | 3647 | <.001 | .908 | 3647 | <.001 |
| Avg humidity | .104 | 3647 | <.001 | .962 | 3647 | <.001 |
| Avg windspeed | .068 | 3647 | <.001 | .970 | 3647 | <.001 |
| Avg pressure | .431 | 3647 | .000 | .585 | 3647 | <.001 |

Since, all p-values are ≤ 0.05. We reject $H_0$.

Hence, all variables are not normally distributed.

Since, linearity and normality are violated by all variables, we carry out Spearman's rank correlation and Kendall's tau coefficient to find association.

# <u>SPEARMAN'S RANK CORRELATION:</u>

Spearman's correlation coefficient, ($\rho$, also signified by $r_s$) measures the strength and direction of association between two ranked variables. Spearman's correlation determines the strength and direction of the **monotonic relationship** between your two variables. A monotonic relationship is a relationship that does one of the following:

(1) as the value of one variable increases, so does the value of the other variable;

(2) as the value of one variable increases, the other variable value decreases.

The Spearman correlation coefficient, $\rho$ can take values from -1 to +1. A $\rho$ of +1 indicates a perfect association of ranks, a $\rho$ of zero indicates no association between ranks and a ρ of -1 indicates a perfect negative association of ranks. The closer ρ is to zero, the weaker the association between the ranks. There is no requirement of <u>normality</u> and hence it is a nonparametric statistic.

We can verbally describe the strength of the correlation using the following guide for the absolute value of:

- .00-.19 "very weak"
- .20-.39 "weak"
- .40-.59 "moderate"
- .60-.79 "strong"
- .80-1.0 "very strong"

There are two methods to calculate Spearman's correlation depending on whether:

(1) your data does not have tied ranks

(2) your data has tied ranks.

The formula for when there are <u>**no tied ranks**</u> is:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

where $d_i$ = difference in paired ranks

$n$ = number of cases

$i$ = paired score.

The formula to use when there are **tied ranks** is:

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

## KENDALL'S TAU COEFFICIENT:

We have another approach like spearman rank correlation i.e. **Kendall's Tau (τ)** correlation coefficient. It assesses statistical associations based on the ranks of the data.
τ takes the values between -1 and 1. Spearman's ρ usually have larger values than Kendall's Tau.
The distribution of Kendall's tau has better statistical properties.
The interpretation of Kendall's tau in terms of the probabilities of observing the agreeable (concordant) and non-agreeable (discordant) pairs is very direct.

**Concordant pairs:** if both elements of one pair are either greater than, equal to, or less than the corresponding elements of the other pair.
**Discordant pairs:** if the two numbers in one observation differ in opposite directions.

The Kendall τ coefficient is defined as:

$$\tau = \frac{(number\ of\ concordant\ pairs) - (number\ of\ discordant\ pairs)}{c\binom{n}{2}}$$

where $\binom{n}{2} = \frac{n(n-1)}{2}$ is the binomial coefficient for the number of ways to choose two items from n items.

## 1.Checking Association between AQI and Average Temperature:

**To test:**

$H_0$: There is no association between AQI and Average temperature i.e. $\rho=0$
against
$H_1$: There is association between AQI and Average temperature i.e. $\rho\neq0$

By Spearman's Rank correlation Coefficient-

| p-value | ρ |
|---|---|
| 2.2e-16≈0 | -0.2184905 |

By Kendall's Tau-

| p-value | τ |
|---|---|
| 2.2e-16≈0 | -0.1880942 |

**Interpretation:** Here, by both methods we get p-value≤0.05, hence we reject $H_0$.

**Conclusion:** There is weak negative association between AQI and Average temperature.

## 2.Checking Association between AQI and Average Humidity:

**To test:**

$H_0$: There is no association between AQI and Average Humidity i.e. $\rho=0$
against
$H_1$: There is association between AQI and Average Humidity i.e. $\rho\neq0$

By Spearman's Rank correlation Coefficient-

| p-value | ρ |
|---|---|
| 0.00 | -0.2184905 |

By Kendall's Tau-

| p-value | τ |
|---|---|
| 0.00 | -0.1517189 |

**Interpretation**: Here, by both methods we get p-value≤0.05, hence we reject $H_0$.

**Conclusion:** There is weak negative association between AQI and Average Humidity.

## 3.Checking Association between AQI and Average Pressure:

**To test:**

$H_0$: There is no association between AQI and Average Pressure i.e. $\rho=0$
against
$H_1$: There is association between AQI and Average Pressure i.e. $\rho\neq0$

By Spearman's Rank correlation Coefficient-

| p-value | ρ |
|---|---|
| 0.00 | -0.3352159 |

By Kendall's Tau-

| p-value | τ |
|---|---|
| 0.00 | -0.2274167 |

**Interpretation:** Here, by both methods we get p-value≤0.05, hence we reject $H_0$.

**Conclusion:** There is weak negative association between AQI and Average Pressure.

## 4.Checking Association between AQI and Average Dew point:

**To test:**

$H_0$: There is no association between AQI and Average Dew point i.e. $\rho=0$

against

$H_1$: There is association between AQI and Average Dew point i.e. $\rho\neq0$

By Spearman's Rank correlation Coefficient-

| p-value | $\rho$ |
|---|---|
| 0.00 | -0.5437234 |

By Kendall's Tau-

| p-value | $\tau$ |
|---|---|
| 0.00 | -0.3618658 |

**Interpretation:** Here, by both methods we get p-value$\leq$0.05, hence we reject $H_0$.

**Conclusion:** There is moderate negative association between AQI and Average Dew point.

## 5.Checking Association between AQI and Average Wind Speed:

**To test:**

$H_0$: There is no association between AQI and Average Wind Speed i.e. $\rho=0$
against
$H_1$: There is association between AQI and Average Wind Speed i.e. $\rho\neq0$

By Spearman's Rank correlation Coefficient-

| p-value | ρ |
|---|---|
| 0.00 | -0.4748167 |

By Kendall's Tau-

| p-value | τ |
|---|---|
| 0.00 | -0.3362965 |

**Interpretation:** Here, by both methods we get p-value≤0.05, hence we reject $H_0$.

**Conclusion:** There is moderate negative association between AQI and Average Wind Speed.

Also, we check association using cross tabulation method too.

# CROSS TABULATION

Cross Tabulation table is the basic technique for examining between two categorical (nominal or ordinal) variables, possibly controlling for additional of variables. Cross tabulation procedure offers several measures of and tests association. Additionally, you can obtain estimates of the 0relative risk of an event given the presence or absence of a characteristic. A number of tests are available to determine if the relationship between 2x2 tabulated variables is significant.

**Pearson chi square tests**: Pearson chi-square used to test the independence of two attributes. A test of independence assesses whether paired observations on two attributes, expressed in a contingency table, independent of each other i.e. unassociated with each other. For the test of independence, a chi-square probability of less than or equal to 0.05 (or the chi square statistic being larger than the 0.05 critical point) is commonly interpreted by applied workers as justification for rejecting the null hypothesis.

**Hypothesis to be tested:**

$H_0$: The two attributes are independent of each other.

against

$H_1$: The two attributes are dependent of each other.

The first step in the chi-square test is to calculate the chi-square statistic. The chi-square statistic is calculated by finding the difference between each observed and theoretical frequency for each possible outcome, squaring them, dividing each by the theoretical frequency, and taking the sum of the results.

The test statistic is defined as:

$$\chi^2 = \sum \frac{(o_i - E_i)^2}{E_i}$$

$\chi^2$ = Pearson's cumulative test statistic, which asymptotically approaches a $\chi^2$ Distribution.

$O_i$ = the number of observations of type i.

n= total number of observations.

$E_i$ = the expected (theoretical) frequency of type i, asserted by the null hypothesis.

The chi square statistic can then be used to calculate p-value by comparing the value of the statistic to a chi square distribution. The number of degrees of freedom is equal to (k-1) $\times$ (r-1) where, k and r are the levels of two attributes.

<u>In our data we have one variable AQI bucket as categorical variable but other variables such as temperature, pressure, etc. are continuous variable. So, to carry out Chi square test we have converted them into categorical variable and proceeded ahead.</u>

# 1.Checking Independence AQI Bucket to Avg Temperature:

Attribute 1=AQI bucket

Attribute 2=Average Temperature

**To test:**

$H_0$: There is no association between AQI Bucket and Average Temperature.

against

$H_1$: There is association between AQI Bucket and Average Temperature.

| AQI Bucket | Average Temperature | | Total |
|---|---|---|---|
| | 40-80 | 81-120 | |
| Good | 45 | 79 | 124 |
| Satisfactory | 234 | 1132 | 1366 |
| Moderate | 613 | 968 | 1581 |
| Poor | 181 | 146 | 327 |
| Very Poor | 148 | 22 | 170 |
| Severe | 72 | 7 | 79 |
| **Total** | 1293 | 2354 | 3647 |

**Chi Square Test:**

| Pearson Chi-Square Value | Degree of Freedom | p-value |
|---|---|---|
| 569.52 | 5 | 2.2e-16≈0 |

**Interpretation**: We get p-value≤0.05. Hence, we reject $H_0$.

**Conclusion:** There is an association between AQI and Average Temperature.

## 2.Checking Independence AQI Bucket to Average Humidity:

Attribute 1=AQI bucket

Attribute 2= Average Humidity

**To test:**

$H_0$: There is no association between AQI Bucket and Average Humidity.

against

$H_1$: There is association between AQI Bucket and Average Humidity.

| AQI Bucket | Average Humidity | | Total |
|---|---|---|---|
| | 0-50 | Above 50 | |
| Good and Satisfactory | 204 | 1286 | 1490 |
| Moderate | 635 | 946 | 1581 |
| Poor | 123 | 204 | 327 |
| Very Poor | 18 | 152 | 170 |
| Severe | 6 | 73 | 79 |
| **Total** | 986 | 2661 | 3647 |

**Chi Square Test:**

| Pearson Chi-Square Value | Degree of Freedom | p-value |
|---|---|---|
| 341.03 | 4 | 2.2e-16≈0 |

**Interpretation:** We get p-value≤0.05. Hence, we reject $H_0$.

**Conclusion:** There is an association between AQI Bucket and Average Humidity.

## 3.Checking Independence AQI Bucket to Average Pressure:

Attribute 1=AQI bucket

Attribute 2= Average Pressure

**To test**:

$H_0$: There is no association between AQI Bucket and Average Pressure.

against

$H_1$: There is association between AQI Bucket and Average Pressure.

| AQI Bucket | Average Pressure | | Total |
|---|---|---|---|
| | Below 28 | Above 28 | |
| Good | 113 | 11 | 124 |
| Satisfactory | 370 | 996 | 1366 |
| Moderate | 468 | 1113 | 1581 |
| Poor | 246 | 81 | 327 |
| Very Poor and Severe | 240 | 9 | 249 |
| **Total** | 1437 | 2210 | 3647 |

**Chi Square Test:**

| Pearson Chi-Square Value | Degree of Freedom | p-value |
|---|---|---|
| 803.8 | 4 | 2.2e-16≈0 |

**Interpretation:** We get p-value≤0.05. Hence, we reject $H_0$.

**Conclusion:** There is an association between AQI and Average pressure.

## 4.Checking Independence AQI Bucket to Average Dew point:

Attribute 1=AQI bucket

Attribute 2= Average Dew Point

**To test:**

$H_0$: There is no association between AQI Bucket and Average Dew point.
against
$H_1$: There is association between AQI Bucket and Average Dew point.

| AQI Bucket | Average Dew Point | | Total |
|---|---|---|---|
| | 0-60 | 61-120 | |
| Good and Satisfactory | 119 | 1371 | 1490 |
| Moderate | 693 | 888 | 1581 |
| Poor | 178 | 149 | 327 |
| Very Poor | 130 | 40 | 170 |
| Severe | 59 | 20 | 79 |
| **Total** | 1179 | 2468 | 3647 |

**Chi. Square Test:**

| Pearson Chi-Square Value | Degree of Freedom | p-value |
|---|---|---|
| 792.39 | 4 | 2.2e-16≈0 |

**Interpretation:** We get p-value≤0.05. Hence, we reject $H_0$.

**Conclusion:** There is an association between AQI Bucket and Average Dew point.

## 5.Checking Independence AQI Bucket to Average Wind Speed:

Attribute 1=AQI bucket

Attribute 2= Average Wind Speed

**To test:**

$H_0$: There is no association between AQI Bucket and Average Wind Speed.

against

$H_1$: There is association between AQI Bucket and Average Wind Speed.

| AQI Bucket | Average Wind Speed | | Total |
|---|---|---|---|
| | 0-10 | Above 10 | |
| Good | 53 | 71 | 124 |
| Satisfactory | 1242 | 124 | 1366 |
| Moderate | 1539 | 42 | 1581 |
| Poor | 319 | 8 | 327 |
| Very Poor and Severe | 243 | 6 | 249 |
| **Total** | 3396 | 251 | 3647 |

**Chi Square Test:**

| Pearson Chi-Square Value | Degree of Freedom | p-value |
|---|---|---|
| 565.93 | 4 | 2.2e-16≈0 |

**Interpretation:** We get p-value≤0.05. Hence we reject $H_0$.

**Conclusion:** There is an association between AQI Bucket and Average Wind Speed.

# Conclusion:

By Pearson chi square test, Spearman's Rank correlation coefficient and Kendall's Tau we conclude that,

- There is weak negative association between AQI and Average temperature, that means as temperature increases, AQI slightly decreases and vice a versa.
- There is weak negative association between AQI and Average Humidity, that means as humidity increases, AQI slightly decreases and vice a versa.
- There is weak negative association between AQI and Average Pressure, that means as Pressure increases, AQI slightly decreases and vice a versa.
- There is moderate negative association between AQI and Average Dew point, that means as Dew point increases, AQI decreases and vice a versa.
- There is moderate negative association between AQI and Average Wind Speed, that means as Wind speed increases, AQI decreases and vice a versa.

# Conclusion

**Objective 1**- We conclude that Air Quality Index decreases during lockdown.

**Objective 2-** We conclude that forecasted future AQI value for next month for city Hyderabad and Jaipur shows slightly decreasing trend, for city Chennai it is slightly constant and for Delhi it increases in the beginning and then shows a decrease.

**Objective 3-** We arrived at a conclusion that there is weak association in AQI and temperature, humidity, pressure, wind speed and dew point. Also, because of negative association as AQI increases, all other factors such as temperature, humidity, pressure, wind speed and dew point decrease.

# Coding

# R codes

## for Cross Tabulation

```
>data<-read.csv("data of 4 cities.csv")
>data
>temp<-table(data$AQI_bucket, data$Temperature)
>temp
>pre<-table(data$AQI_bucket, data$Pressure)
>pre
>hum<-table(data$AQI_bucket, data$humidity)
>hum
>dew<-table(data$AQI_bucket, data$Dewpoint)
>dew
>wind<-table(data$AQI_bucket, data$Windspeed)
>wind
```

For Chi square test

```
>chisq.test (temp, correct=F)
>chisq.test (pre, correct=F)
>chisq.test (hum, correct=F)
>chisq.test (dew, correct=F)
>chisq.test (wind, correct=F)
```

For Spearman's Rank correlation coefficient:

```
>cor.test(data$AQI, Avg Temp, method="spearman",exact=F,data=data)
>cor.test(data$AQI, Avg pressure, method="spearman",exact=F,data=data)
>cor.test(data$AQI, Avg humidity, method="spearman",exact=F,data=data)
>cor.test(data$AQI, Avg dewpoint, method="spearman",exact=F,data=data)
```

>cor.test(data$AQI, Avg windspeed, method=”spearman”,exact=F,data=data)

For Kendall’s Tau coefficient:

>cor.test(data$AQI, Avg Temp, method=”kendall”, data=data)

>cor.test(data$AQI, Avg pressure, method=”kendall”, data=data)

>cor.test(data$AQI, Avg humidity, method=”kendall”, data=data)

>cor.test(data$AQI, Avg dewpoint, method=”kendall”, data=data)

>cor.test(data$AQI, Avg windspeed, method=”kendall”, data=data)

For time plot Using python

```
import pandas as pd
import seaborn as sns

dv=pd.read_excel('Chennai during lockdown.xlsx')
dv
dk=pd.read_excel('Chennai before lockdown.xlsx')
dk

import matplotlib.pyplot as plt
sns.lineplot(x = "Month", y = "AQI",
             data = dk)
sns.lineplot(x = "Month", y = "AQI",
             data = dv)

plt.xticks(rotation = 25)


dv=pd.read_excel('Delhi during lockdown.xlsx')
dv
dk=pd.read_excel('Delhi before lockdown.xlsx')
dk

import matplotlib.pyplot as plt
sns.lineplot(x = "Month", y = "AQI",
             data = dk)
sns.lineplot(x = "Month", y = "AQI",
             data = dv)

plt.xticks(rotation = 25)


dv=pd.read_excel('Jaipur during lockdown.xlsx')
dv
dk=pd.read_excel('Jaipur before lockdown.xlsx')
dk

import matplotlib.pyplot as plt
```

```
sns.lineplot(x = "Month", y = "AQI",
             data = dk)
sns.lineplot(x = "Month", y = "AQI",
             data = dv)

plt.xticks(rotation = 25)


dv=pd.read_excel('Hyderabad during lockdown.xlsx')
dv
dk=pd.read_excel('Hyderabad before lockdown.xlsx')
dk

import matplotlib.pyplot as plt
sns.lineplot(x = "Month", y = "AQI",
             data = dk)
sns.lineplot(x = "Month", y = "AQI",
             data = dv)

plt.xticks(rotation = 25)
```

# Bibliography


1. https://www.kaggle.com/rohanrao/air-quality-data-in-india

2. https://www.wunderground.com/health/mood.asp

3. https://www.analyticsvidhya.com/blog/2015/12/complete-tutorial-time-series-modeling/