

ANALYSIS OF AIR QUALITY INDEX (PUNE CITY)



AIR QUALITY INDEX

A Guide to Air Quality and Your Health

Department Of Statistics

Fergusson College (Autonomous)

Pune

2023-2024



Deccan Education Society's

Fergusson College (Autonomous), Pune

Department of Statistics

T. Y. B. Sc.

Year 2023-24

STS3609: Statistics Practical III

CERTIFICATE

This is to certify that Mr./Ms. _____
Roll no. _____, has satisfactorily completed the project work entitled
“**ANALYSIS OF AIR QUALITY INDEX (PUNE CITY)**” during the
academic year 2023 – 24 as per the rules and regulations laid down by
FERGUSSON COLLEGE (Autonomous), Pune.

Place: Pune

Date: / / 2024

Project guide: -

Dr. Subhash S. Shende

Dr. Subhash S. Shende

Vice principal & Head Department,

Statistics Department,

Fergusson College, pune

FINAL YEAR PROJECT

Mentor : Dr Subhash Shende

By

VINAY DAMSE (212112)

VINAY HABALE (212268)

MAHEK CHAUDHARI (212279)

NIRAJ MHATRE (212285)

SAEE SHINDE (212286)

PAYAL GARGADE (212287)

Acknowledgement

We extend our sincere appreciation to everyone who contributed to the completion of this project. Special thanks to all those whose expertise and valuable insights helped us.

Firstly, we express our gratitude to our project guide, Dr. Subhash S. Shende, Vice Principal and Head of the Department of Statistics at Fergusson College, Pune, for his valuable guidance, knowledge, constructive criticism, and support throughout the project.

Furthermore, we thank the Pune Smart City office, whose contribution to the project is considerable and of utmost importance.

We also thank the department faculty, our seniors, friends, and everyone involved in the successful completion of the project. Their collective efforts and support were instrumental in achieving our goals.

TABLE OF CONTENTS :

| Sr. No. | Title | Page No. |
|----------------|--------------------------------------|-----------------|
| 1. | Introduction | 6 |
| 2. | Motivation | 7 |
| 3. | Objective | 8 |
| 4. | Data collection | 8 |
| 5. | Data set | 8 |
| 6. | Chemical Theory of Pollutants | 9 |
| 7. | Statistical tests and tools | 12 |
| i. | Data visualization | 12 |
| ii. | Correlation analysis | 15 |
| iii. | Regression analysis | 17 |
| iv. | Non – parametric tests | 29 |
| v. | Estimation of parameters | 31 |
| vi. | Time series analysis | 43 |
| 8. | Conclusion | 55 |
| 9. | References | 56 |
| 10. | Software Used | 57 |

1.INTRODUCTION:

In the face of escalating environmental concerns and the undeniable impact of air pollution on public health, the Air Quality Index (AQI) project emerges as a crucial initiative in the pursuit of a sustainable and healthier future. The pressing need to monitor and assess air quality has become increasingly apparent in recent years, reflecting a growing consciousness about the detrimental effects of pollutants on our surroundings.

As societies worldwide continue to advance technologically, the AQI project takes centre stage by leveraging detailed information from environmental sensors to formulate a comprehensive numerical representation of air quality. This amalgamation of data not only serves as a key indicator but also offers a holistic understanding of potential health risks associated with ambient air. By delving into the patterns and trends inherent in AQI data, the project seeks to empower communities with knowledge that is instrumental in crafting sustainable strategies and protective measures.

Our mission goes beyond mere data collection – it aims to transform raw information into actionable insights. Through a meticulous process of statistical analysis, we intend to unravel the complex tapestry of air quality dynamics, providing communities with the tools they need to make informed decisions regarding their well-being. Real-time updates and user-friendly interfaces will be at the forefront of our efforts, ensuring that individuals can easily access and comprehend the information, thereby fostering heightened awareness.

The AQI project stands as a beacon of environmental responsibility and community well-being. By fostering a deeper understanding of air quality through statistical analysis, we aspire to catalyse a collective effort towards cleaner and healthier air for all. Join us on this journey as we strive to create a positive impact, one breath at a time.

2.MOTIVATION:

Embarking on the final year project focused on analysing the Air Quality Index (AQI) data represents not only a meaningful endeavour but also a timely and impactful one. In the current global climate, concerns surrounding air pollution and its potential ramifications on public health have reached a critical juncture. The relevance of this initiative cannot be overstated, given the increasing urgency to address air quality issues.

The significance of delving into AQI data lies in its potential to yield insightful analyses that can drive public awareness campaigns, inform urban planning strategies, and influence policy decisions. As we navigate a world grappling with environmental challenges, our project stands as a beacon of awareness and action. By deciphering the intricate numbers within the AQI dataset, we equip ourselves with the tools to not only understand the depth of the issue but also to communicate its gravity effectively.

The multifaceted nature of the problems associated with air quality, spanning public health, climate change, and overall environmental impact, necessitates a comprehensive approach. Our project serves as a conduit for exploring these various dimensions, shedding light on the interconnectedness of air quality issues and their far-reaching consequences. By addressing these challenges, we aim to contribute not only to academic knowledge but also to the development of sustainable policies and practices for the betterment of society.

Framing and implementing effective policies are essential components of our project's broader mission. Beyond the academic realm, the practical implications of our work extend to influencing policy decisions that can drive positive change. As we navigate the delicate balance between economic development and environmental preservation, understanding the AQI becomes a crucial factor in ensuring sustainable progress.

In essence, our project is more than a culmination of academic efforts; it is a conscientious endeavour with real-world implications. By immersing ourselves in the intricacies of AQI data, we take on the responsibility of contributing to a healthier, more sustainable future. Through our dedication to understanding, analysing, and communicating the nuances of air quality issues, we aspire to be catalysts for positive change in public awareness, urban planning, and policy development.

3.OBJECTIVE:

Asses relationship between individual pollutants and AQI.

Identify main contributors to AQI fluctuations.

Evaluate the seasonal effects on AQI.

Significance of all the variables affecting AQI.

Regression analysis and model fitting.

Time series analysis and forecasting.

4.DATA COLLECTION :

The data for the project is a secondary data.

The source being Pune Smart City Office, a government agency that records the data for Air Quality.

This data is for the Pune region.

5.DATA SET :

Pollutants : CO, CO₂, SO₂, Ozone, PM_{2.5}, PM₁₀, NO₂

Environmental factors : Temperature Maximum, Temperature Min, Sound, Light, UV Maximum, UV Minimum, Humidity

6.CHEMICAL THEORY OF POLLUTANTS :

The Air Quality Index (AQI) is a measure used to communicate how polluted the air is in a specific area. While the AQI itself is a numerical scale, it reflects the concentration of certain pollutants in the air, many of which have a chemical basis.

Particulate Matter (PM10 and PM2.5) :

These are tiny particles suspended in the air. PM10 refers to particles with a diameter of 10 micrometres or less, while PM2.5 refers to even smaller particles with a diameter of 2.5 micrometres or less. These particles can originate from various sources, including vehicle emissions, industrial processes, and natural sources like wildfires. Chemically, they can consist of a variety of substances, including carbonaceous particles, metals, and organic compounds.

Ozone (O₃) :

Ground-level ozone is formed when pollutants from vehicles, industrial facilities, and other sources react with sunlight. It is a major component of smog. Ozone formation involves complex chemical reactions between nitrogen oxides (NO_x) and volatile organic compounds (VOCs) in the presence of sunlight. High levels of ozone can cause respiratory issues and other health problems.

Nitrogen Dioxide (NO₂) and Nitric Oxide (NO) :

These are nitrogen oxides produced during combustion processes, such as those in vehicles and power plants. NO₂ is a brown gas with a sharp, pungent odour, and it can cause respiratory problems. In the atmosphere, NO₂ can react with other compounds to form ozone and particulate matter.

Sulphur Dioxide (SO₂) :

SO₂ is produced primarily by burning fossil fuels containing sulphur, such as coal and oil. It can also come from industrial processes like metal smelting. SO₂ can react with other compounds in the atmosphere to form sulphate particles, which contribute to particulate matter pollution.

Carbon Monoxide (CO):

CO is a colourless, odourless gas produced by incomplete combustion of carbon-containing fuels. It can come from vehicle exhaust, industrial processes, and wildfires. CO can bind to haemoglobin in the bloodstream, reducing the blood's ability to carry oxygen, which can lead to health problems.

In a chemistry laboratory, there are several methods used to measure the Air Quality Index (AQI) by analysing the concentration of various pollutants.

- i. **Gravimetric Analysis:** This method involves collecting airborne particulate matter on a filter and then weighing the filter before and after sampling to determine the mass of particulate matter. The collected particles can be analysed further to identify specific components. This method is commonly used for measuring PM10 and PM2.5 concentrations.
- ii. **Gas Chromatography (GC):** Gas chromatography is a technique used to separate and analyse volatile organic compounds (VOCs) and other gaseous pollutants in air samples. It involves injecting a sample into a chromatographic column, where the compounds are separated based on their chemical properties and interactions with the stationary phase. Detection methods such as flame ionization detection (FID) or mass spectrometry (MS) can be used to quantify the concentration of individual pollutants.
- iii. **Chemiluminescence:** Chemiluminescence is a method used to measure the concentration of nitrogen dioxide (NO₂) in air samples. It involves reacting NO₂ with ozone (O₃) to produce excited nitrogen dioxide molecules, which then emit light as they return to their ground state. The intensity of the emitted light is proportional to the concentration of NO₂ in the sample.
- iv. **Ultraviolet (UV) Spectroscopy:** UV spectroscopy can be used to measure the concentration of ozone (O₃) in air samples. Ozone absorbs UV light at specific wavelengths, and the extent of absorption is proportional to the concentration of ozone.

in the sample. UV absorption spectroscopy can provide real-time measurements of ozone levels in the atmosphere.

- v. **Electrochemical Sensors:** Electrochemical sensors are commonly used for measuring gases such as carbon monoxide (CO) and sulphur dioxide (SO₂). These sensors work by detecting changes in electrical conductivity or potential that occur when the target gas interacts with an electrode surface. Electrochemical sensors are portable, relatively inexpensive, and can provide real-time measurements in the field.
- vi. **Inductively Coupled Plasma Mass Spectrometry (ICP-MS):** ICP-MS is a sensitive analytical technique used for measuring trace metals in air samples. It involves ionizing the sample in an inductively coupled plasma (ICP) torch and then analysing the ions using mass spectrometry. ICP-MS can detect a wide range of metals, including heavy metals such as lead and mercury, which are harmful air pollutants.

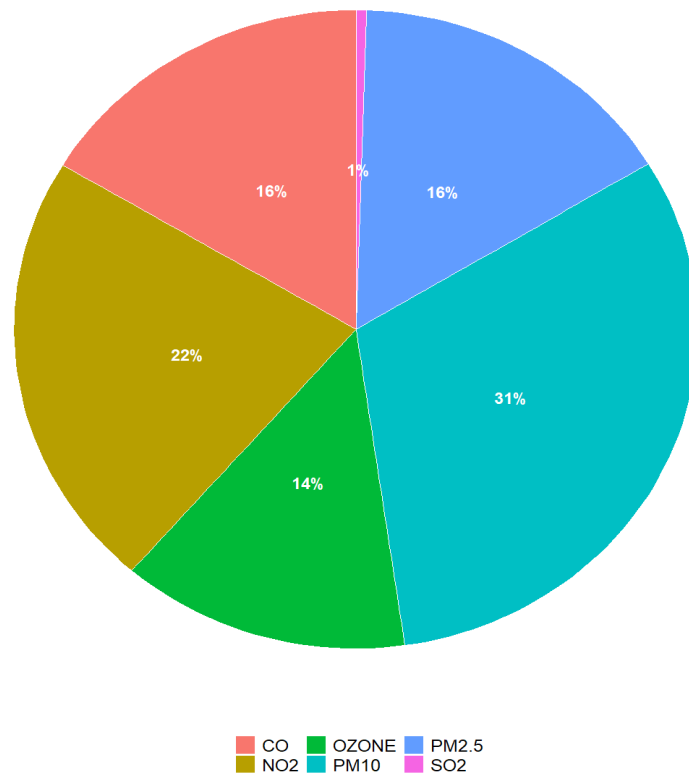
7.STATISTICAL TESTS AND TOOLS

The air we breathe, is it safe? Are we having potential risk of any diseases due to air quality in Pune City,

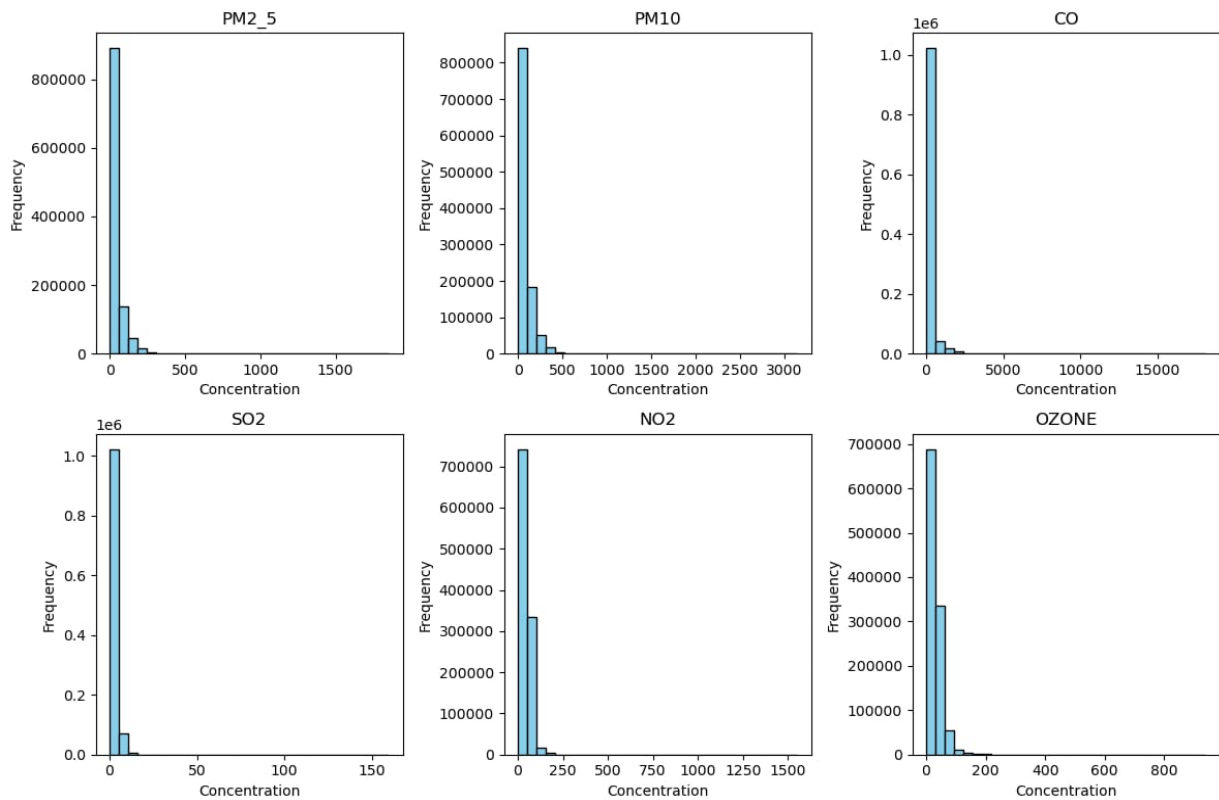
To find out answers to these questions, we will carry out some statistical analysis,

Firstly, we to visualize the data for pollutants (which are our variables) through different forms

I] DATA VISUALIZATION :



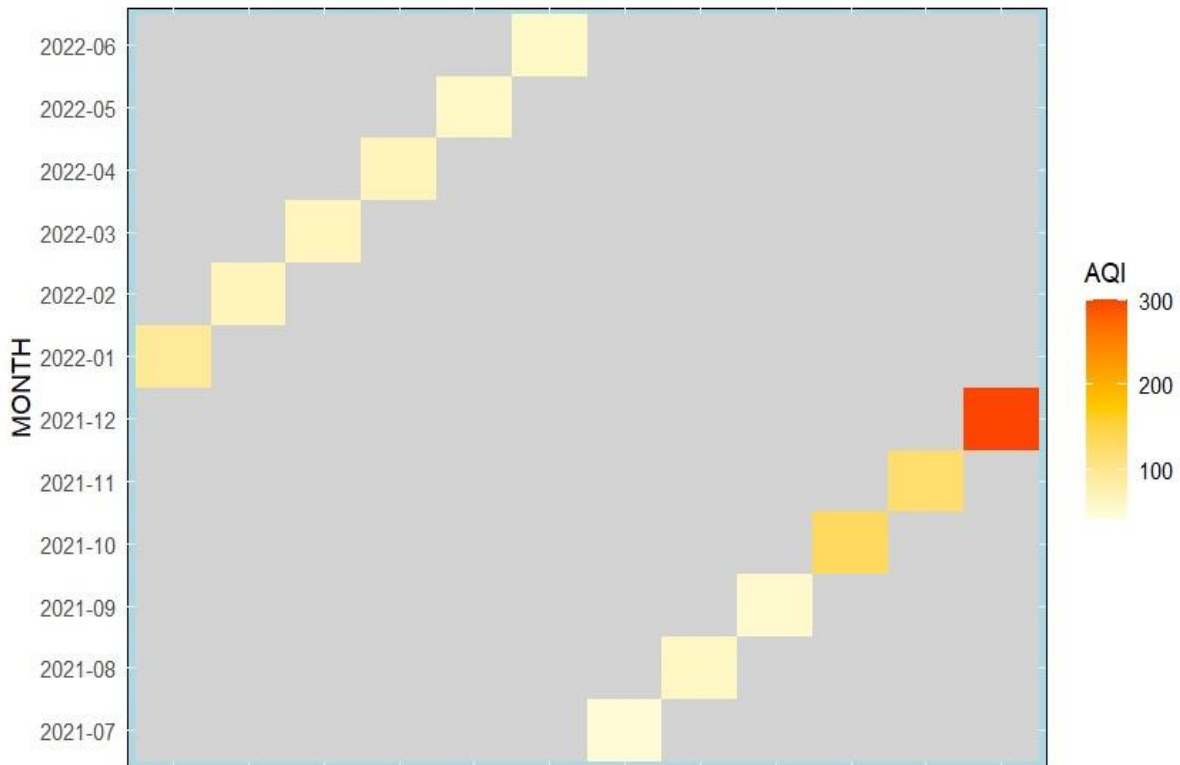
In above pie chart, PM10 has 31% observations out of all observations. i.e. PM10 has more observations as compared to other pollutants. After PM10 , NO2 has more observations and so on.



From above plots we can say almost all pollutants have positively skewed distribution. i.e. All pollutants have more observations in their respective small ranges and as value increases we can see number of observations belong to that range decreases.

HEAT MAP :

A heat map is a two-dimensional representation of data in which various values are represented by colours.



A rising trend can be observed from the month of October, which continues till the end of January and then there's a decline in the trend.

This rise indicates degraded quality of air which can be due to various factors like different festivals , celebrations or sudden change in the climate.

Now we carry out correlation and regression analysis to find out which variables affect the most to AQI hence instead of studying each variable studying these variables and taking measures to reduce/tackle them solely would help in significantly improve the AQI.

III] CORRELATION ANALYSIS :

we carry out a correlation analysis to check how each variable is correlated with AQI

| Column1 | AQI | CO | CO2 | NO2 | OZONE | PM2_5 | PM10 | SO2 | Temp_max | Temp_min | UV_max | UV_min | Humidit | Light | Sound |
|----------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|-------------|-------------|--------|------------|----------|-------|
| AQI | 1 | | | | | | | | | | | | | | |
| CO | -0.120108323 | 1 | | | | | | | | | | | | | |
| CO2 | -0.124714464 | 0.981783952 | 1 | | | | | | | | | | | | |
| NO2 | 0.574939515 | 0.118588128 | 0.080859303 | 1 | | | | | | | | | | | |
| OZONE | 0.06128322 | 0.339273772 | 0.333968144 | 0.063658165 | 1 | | | | | | | | | | |
| PM2_5 | 0.893687569 | -0.153023569 | -0.154722977 | 0.479672349 | -0.064696258 | 1 | | | | | | | | | |
| PM10 | 0.835678375 | -0.162589335 | -0.164279107 | 0.411316815 | -0.043834908 | 0.967301326 | 1 | | | | | | | | |
| SO2 | 0.474687538 | -0.045273095 | -0.087025612 | 0.609873369 | -0.273101499 | 0.532644678 | 0.486629822 | 1 | | | | | | | |
| Temp_max | -0.256502443 | 0.17522799 | 0.175444769 | -0.426074755 | 0.63982695 | -0.298056634 | -0.231527903 | -0.36741665 | 1 | | | | | | |
| Temp_min | -0.487995485 | 0.16734021 | 0.172627212 | -0.651888134 | 0.162827873 | -0.428502785 | -0.351975359 | -0.399893662 | 0.679443811 | 1 | | | | | |
| UV_max | -0.204283073 | 0.709408456 | 0.740136489 | -0.079851952 | 0.135841873 | -0.184357942 | -0.174469414 | 0.007265153 | 0.217905262 | 0.223149213 | 1 | | | | |
| UV_min | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | | | |
| Humidity | -0.135576537 | 0.029578292 | 0.025021498 | 0.042367222 | -0.689263207 | -0.031141858 | -0.052674704 | 0.278833169 | -0.605111132 | -0.01864331 | 0.079614542 | 0 | 1 | | |
| Light | -0.215549027 | 0.794292771 | 0.826739514 | -0.16262517 | 0.366475073 | -0.211661982 | -0.203007011 | -0.232528164 | 0.364904987 | 0.317578166 | 0.812047937 | 0 | -0.114924 | 1 | |
| Sound | 0.001180255 | 0.163687835 | 0.163670835 | -0.207374181 | 0.37729761 | -0.005553303 | 0.010144064 | -0.159790546 | 0.488035845 | 0.205936684 | 0.226556137 | 0 | -0.4402068 | 0.383663 | 1 |

Interpretation:-

Correlation between (CO and AQI),(CO2 and AQI),(Temp min and AQI),(Temp max and AQI),(UV max and AQI),(humidity and AQI),(light and AQI) are negative values which implies there is negative correlation between these pollutants and AQI.

Correlation between UV min and AQI is 0 which implies they are uncorrelated. Correlation between (NO2 and AQI),(Ozone and AQI),(PM2.5 and AQI),(PM10 and AQI),(SO2 and AQI),(Sound and AQI) are all positive values that implies there is positive correlation between these pollutants and AQI. Correlation between AQI and NO2, PM2.5, PM10 are above than 0.5 hence, they are more important to predict AQI

III] REGRESSION ANALYSIS :

Regression analysis is a statistical technique for investigating and modelling the relationship between variables. Regression analysis is one of the most widely used techniques for analysing multi-factor data. Its broad appeal and usefulness result from the conceptually logical process of using an equation to express the relationship between a variable of interest (the response or dependent) and a set of related predictors (Independent) variables. Regression analysis is also interesting theoretically because of elegant underlying mathematics and a well-developed statistical theory. Successful use of regression requires an appreciation of both the theory and the practical problems that typically arise when the technique is employed with real-world data.

Applications of regression are numerous and occur in almost every field, including engineering, the physical and chemical sciences, economics, management, life and biological sciences, and the social sciences.

Regression analysis is primarily used for two conceptually distinct purposes. First, regression analysis is widely used for prediction and forecasting. Second, in some situation's regression analysis can be used to infer the casual relationships between dependent and independent variables.

➤ Underlying Assumptions :-

- i. Errors are normally and independent distributed.
- ii. Variable X is with no error or negligible error.
- iii. The variance of residuals is constant (homoscedasticity).
- iv. The residuals are uncorrelated.

➤ Historical Background :-

- i. Francis Galton started using the term regression in his biology research.
- ii. Karl Pearson and Udny Yule extended Galton's work to the statistical context.
- iii. Legendre and Gauss developed the method of least squares used in regression analysis.
- iv. Ronald Fisher developed the maximum likelihood method used in relates statistical interference (test of significance of regression etc.)

❖ **MODEL BUILDING** (Simple Linear Regression Model)

This analysis considers the Simple Linear Regression Model, that is a model with a single regressor variable 'x' that has a relationship with a response variable 'y' that is a straight line. This is simple linear regression model;

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

Where β_0 is the intercept and β_1 is the slope. β_0 and β_1 are unknown constants and ε is a random error component. The errors are assumed to have mean zero and unknown variance σ^2 . Additionally, we usually assume that the errors are uncorrelated. This means that the value of one error does not depend on the value of any other error. The parameters β_0 and β_1 are usually called regression coefficients.

Dependent Variable: Air Quality index (AQI)

Independent Variable: PM2.5, PM10, OZONE, CO, CO2, SO2, NO2, Temperature, Humidity, Light, Sound, etc.

Where,

PM2.5 : Particulate matter 2.5 , PM10 : Particulate matter 10

CO : Carbon dioxide , CO2 : Carbon dioxide

SO2 : Sulphur dioxide , NO2 : Nitrogen dioxide.

TEST FOR β_j 's

HYPOTHESIS TO BE TESTED :

$H_0: \beta_1 = 0$ (There is no relation between dependent variable and independent variable)

$H_1: \beta_1 \neq 0$ (There is relation between dependent variable and independent variable)

TEST STATISTICS :

$$t = \beta_1 / SE(\beta_1) \sim \text{follow } t_{(n-2)} \text{ DF}$$

TEST CRITERIA :

According to p value, we reject H_0 , if p-value is less than 0.05 or t-statistic is larger than p-value , accept H_0 otherwise.

1. Air quality index (AQI) Vs Particulate matter 10 (PM10)

MODEL: $Y = \beta_0 + \beta_1 * X$

Where, $Y = \text{AQI}$, $X = \text{PM}_{2.5}$

FITTED MODEL : $Y = 49.52492 + 1.18600X$

HYPOTHESIS DEVELOPMENT :

$H_0 : \beta_1 = 0$ (There is no relation between AQI and $\text{PM}_{2.5}$)

$H_1 : \beta_1 \neq 0$ (There is relation between AQI and $\text{PM}_{2.5}$)

DECISION : Here, $p\text{-value} = 2.2e^{-16} < 0.05$, we Reject H_0 .

CONCLUSION : We conclude that the regressor $\text{PM}_{2.5}$ is significant. There is relation between AQI and $\text{PM}_{2.5}$.

2. Air quality index (AQI) Vs Carbon dioxide (CO2)

MODEL : $Y = \beta_0 + \beta_1 * X$ Where, $Y = \text{AQI}$, $X = \text{CO}_2$

FITTED MODEL : $Y = 85.74 - 0.00004316X$

HYPOTHESIS DEVELOPMENT :

$H_0 : \beta_1 = 0$ (There is no relation between AQI and CO_2)

$H_1 : \beta_1 \neq 0$ (There is relation between AQI and CO_2)

DECISION : Here, $p\text{-value} = 0.01874 < 0.05$, we Reject H_0 .

CONCLUSION : We conclude that the regressor CO_2 is significant. There is direct relation between AQI and CO_2 .

3. Air quality index (AQI) Vs OZONE (O3)

MODEL : $Y = \beta_0 + \beta_1 * X$ Where, $Y = \text{AQI}$, $X = \text{OZONE}$

FITTED MODEL : $Y = 78.1989 + 0.2611 X$

HYPOTHESIS DEVELOPMENT:

$H_0 : \beta_1 = 0$ (There is no relation between AQI and O3)

$H_1 : \beta_1 \neq 0$ (There is relation between AQI and O3)

DECISION : Here, $p\text{-value} = 0.249 > 0.05$, we Accept H_0 .

CONCLUSION : We conclude that the regressor OZONE is significant. There is no relationship between AQI and CO2.

4. Air quality index (AQI) Vs Max temperature)

MODEL : $Y = \beta_0 + \beta_1 * X$ Where, $Y = \text{AQI}$, $X = \text{Max temperature}$

FITTED MODEL : $Y = 154.678 - 2.268X$

HYPOTHESIS DEVELOPMENT :

$H_0 : \beta_1 = 0$ (There is no relation between AQI and Max temperature)

$H_1 : \beta_1 \neq 0$ (There is relation between AQI and Max temperature)

DECISION : Here, $p\text{-value} = 9.681e^{-07} > 0.05$, we Reject H_0 .

CONCLUSION : We conclude that the regressor Max temperature is significant. There is relationship between AQI and Max temperature.

❖ MODEL BUILDING (Multiple Linear Regression Model)

Multiple regressions generally explain the relationship between multiple independent or predictor variables and one dependent or criterion variable. The goal of multiple linear regression (MLR) is to model linear relationship between the explanatory (independent) variables and response (dependent) variable.

The MLR model with k regressors is given as

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + e_{ij}$$

where, the parameters β_j are called regression coefficients with $j = 0, 1, 2, \dots, k$.

AQI~(PM2_5+PM10+CO+CO2+NO2+SO2+OZONE+Temp_max+UV_max+Humidity+Light+Sound)

Dependent Variable : AQI

Independent Variable : PM2.5, PM10, CO, CO2, NO2, SO2, OZONE, TEMP_MAX, UV_MAX, HUMIDITY, LIGHT and SOUND

Hypothesis to be tested :

$H_0 : \beta_j's = 0$

H_1 : At least one β_j 's is not equal to zero

Coefficients :

| | Estimate | Std. Error | t value | Pr(> t) | Significance |
|-------------|------------|------------|---------|----------|--------------|
| (Intercept) | 5.420e-01 | 8.140e+01 | 0.007 | 0.9947 | |
| PM10 | -2.625e-01 | 6.273e-02 | -4.186 | 3.62e-05 | *** |
| PM2_5 | 1.568e+00 | 1.219e-01 | 12.858 | < 2e-16 | *** |
| CO2 | 3.578e-05 | 4.376e-05 | 0.818 | 0.4141 | |
| CO | -1.776e-02 | 2.513e-02 | -0.707 | 0.4802 | |
| SO2 | -4.126e+00 | 1.758e+00 | -2.347 | 0.0195 | * |
| NO2 | 6.294e-01 | 1.282e-01 | 4.910 | 1.41e-06 | *** |
| OZONE | 2.188e-01 | 1.683e-01 | 1.301 | 0.1943 | |
| Temp_max | 5.458e-01 | 5.109e-01 | 1.068 | 0.2862 | |

| | | | | | |
|----------|------------|-----------|--------|--------|-----|
| Temp_min | -5.711e-01 | 5.267e-01 | -1.084 | 0.2790 | |
| UV_max | 1.790e+00 | 4.832e+00 | 0.370 | 0.7113 | |
| Humidity | -1.001e-01 | 1.042e-01 | -0.961 | 0.3375 | |
| Light | -7.081e-03 | 4.053e-03 | -1.747 | 0.0815 | . |
| Sound | 3.878e-01 | 1.064e+00 | 0.365 | 0.7157 | |
| UV_min | NA | NA | NA | NA | --- |

Significance. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.56 on 341 degrees of freedom

Multiple R-squared: 0.8547, Adjusted R-squared: 0.8492

F-statistic: 154.3 on 13 and 341 DF, p-value: < 2.2e-16

ANOVA table :

| | Df | Sum Sq | Mean Sq | F Value | Pr(>F) | Significance |
|-----------|-----|--------|---------|----------|----------|--------------|
| PM10 | 1 | 301212 | 301212 | 1639.241 | < 2e-16 | *** |
| PM2_5 | 1 | 48825 | 48825 | 265.715 | < 2e-16 | *** |
| CO2 | 1 | 22 | 22 | 0.119 | 0.7305 | |
| CO | 1 | 125 | 125 | 0.682 | 0.4096 | |
| SO2 | 1 | 205 | 205 | 1.118 | 0.2912 | |
| NO2 | 1 | 12742 | 12742 | 69.344 | 2.03e-15 | *** |
| OZONE | 1 | 3847 | 3847 | 20.937 | 6.66e-06 | *** |
| Temp_max | 1 | 189 | 189 | 1.028 | 0.3112 | |
| Temp_min | 1 | 671 | 671 | 3.651 | 0.0569 | . |
| UV_max | 1 | 117 | 117 | 0.639 | 0.4247 | |
| Humidity | 1 | 135 | 135 | 0.733 | 0.3926 | |
| Light | 1 | 540 | 540 | 2.940 | 0.0873 | . |
| Sound | 1 | 24 | 24 | 0.133 | 0.7157 | |
| Residuals | 341 | 62659 | 184 | | | |

Model will be :

$$Y = 0.5420 - 0.2625(X1) + 1.561(X2) - 3.578e^{-05}(X3) - 0.01776(X4) - 4.126(X5) + 0.694(X6) + 0.2188(X7) + 0.5458(X8) - 0.5711(X9) + 1.790(X10) - 0.1001(X11) - 7.081e^{-03}(X12) + 0.3878(X13)$$

Decision :

We reject H_0 for regressors like PM2.5, PM10, NO2, and OZONE accept other regressors.

Conclusion :

The regressors PM2.5, PM10, SO2 and OZONE are more significant regressors which more contribute in model as well as which is not equal to zero and others regressors may not differ significant, from **R square = 0.8547** the clearly indicate a very good regression model performance with 85.47% variation is explain by regressors on the $Y=AQI$ in above model.

Interpretation :

From the regression model equation we can say, Y represents expected change in the AQI per unit change 1st regressors then other regressor is kept constant. Similarly, Y represents expected change in the AQI per unit change in 2nd regressors when other regressor is kept constant same as Y denote expected change in the AQI per unit change in 3rd regressors when other regressor is kept constant and so on.

RELATIVE IMPORTANCE OF VARIABLES :

Relative importance of variables refers to the extent to which different variables contribute to the variation or outcome of a particular phenomenon. This concept is crucial in various statistical techniques, such as regression analysis. Understanding the relative importance of variables helps researchers and analysts identify the most influential factors and make informed decisions.

| Variable | RIV | Distribution |
|----------|-----|--------------|
|----------|-----|--------------|

| | | |
|----------|---------|-----------|
| PM2.5 | 0.33176 | Lognormal |
| PM10 | 0.25641 | Cauchy |
| NO2 | 0.08879 | Beta |
| SO2 | 0.05544 | Normal |
| TEMP_MIN | 0.05233 | Weibull |
| TEMP_MAX | 0.01827 | Lognormal |
| HUMIDITY | 0.01338 | Beta |
| OZONE | 0.00935 | Beta |
| LIGHT | 0.00891 | Lognormal |
| UV_MAX | 0.00874 | Lognormal |
| CO | 0.00427 | Cauchy |
| CO2 | 0.00402 | Cauchy |
| SOUND | 0.00306 | Beta |

[A] Model building by Variable Selection method

Selection allows us for the construction of an optimal regression equation along with investigation into specific predictor variables. In other words, it helps to determine level of importance of each independent variable.

In regression analysis, we often deal with datasets containing many potential explanatory variables, or features. However, not all these features may be equally relevant for predicting the outcome variable. This is where selection methods come into play. These methods aim to identify the optimal subset of features that contribute most to the model's performance while avoiding overfitting and maintaining interpretability. Including irrelevant or redundant features can lead to a less accurate and more complex model that doesn't generalize well to unseen data.

Method of Backward elimination

In this process all the independent variables are entered into equation, particular variable is deleted at each step if it is not contributing to the regression equation.

For accessing accurate model by using R code

```
#R- CODE To fit Multiple Linear Model by method of BACKWARD ELIMINATION
Data= library(readxl)

model=lm(AQI~PM2_5+PM10+CO+CO2+NO2+SO2+OZONE+Temp_max+UV_max+Humidity+Light+Sound,data="file name")

stepAIC(model,direction = "both")
```

Process :

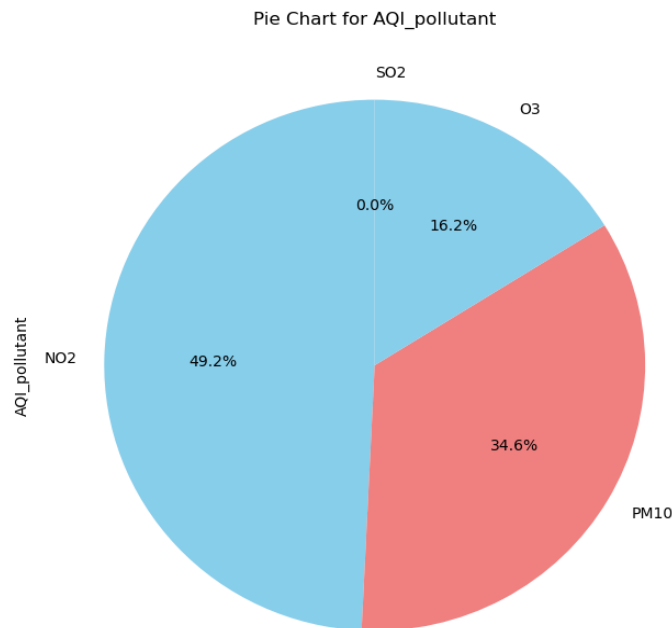
- i. In first step Max temperature is getting rejected.
- ii. In second and third steps a sound and UV max is eliminated respectively.
- iii. In fourth and fifth steps a CO and CO2 is eliminated respectively.

Conclusion :

Reduced Model contain only seven regressors is

$$Y = \beta_0 + \beta_1 PM_{2.5} + \beta_2 PM_{10} + \beta_3 NO_2 + \beta_4 SO_2 + \beta_5 OZONE + \beta_6 Humidity + \beta_7 Light$$

Pie Chart after backward elimination method :



After removing all pollutants which are not present in Backward Elimination. We can still see that NO2 has high number of observations out of all pollutants. After NO2, PM10 has high number of observations and so on.

[B] Diagnostic Checking

Diagnostic checking in regression analysis is a crucial step after fitting a model to your data. It involves a series of techniques to assess the validity of the model and ensure it meets the underlying assumptions of linear regression like multicollinearity, Normality, Independence and Heteroskedasticity etc.

Benefits:

- i. Improved model reliability: Ensures the model accurately represents the data and avoids misleading conclusions.
- ii. Identification of potential issues: Helps in refining the model by addressing violations like non-linearity or outliers.
- iii. Enhanced understanding of the data: Provides insights into the relationship between variables and potential limitations.

B.1) Multicollinearity

Multicollinearity meant the existence of linear relationship among the independent variables. VIF (variance inflation factor) is a method to detect multicollinearity problem. If multicollinearity problem existed in the model, the highly correlated pair of independent variables was to regress again in order to determine the VIF

$$VIF = \frac{1}{1-R^2}$$

```
R CODE to check VIF of all independent variables  
  
library(faraway)  
  
model=lm(AQI~PM2.5+PM10+NO2+SO2+OZONE+Humidity+Light) #Reduced Model  
  
vif(model)
```

Output :

| PM2.5 | PM10 | NO2 | SO2 | OZONE | HUMIDITY | LIGHHT |
|---------|----------|--------|--------|--------|----------|--------|
| 18.2579 | 16.25710 | 1.9824 | 2.1619 | 2.5233 | 2.1445 | 1.2859 |

Conclusion:

If the calculated VIF is larger than 10, the result is serious multicollinearity problem exist in the model. In order to overcome the problem, the less significant variable among the highest correlated pair had to be eliminated. Here the variable with highest VIF in PM2.5.

VIF's after fitting model excluding PM2.5.

| PM10 | NO2 | SO2 | OZONE | HUMIDITY | LIGHT |
|--------|--------|--------|--------|----------|--------|
| 1.4285 | 1.8335 | 2.1460 | 2.4939 | 2.1358 | 1.2848 |

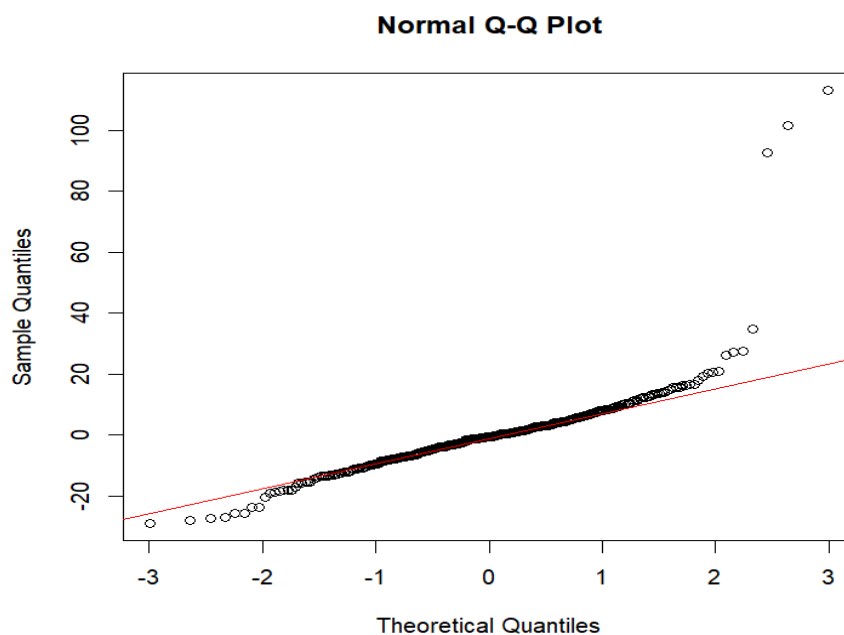
Result :

- VIF of each variable < 10.
- Problem of Multicollinearity is removed.

B.2) Normality :

A very simple method of checking the normality assumption is to construct a normal probability plot of the residuals. This is a graph designed so that the cumulative normal distribution will plot as a straight line.

Normal Probability Plot Small departures from the normality assumption do not affect the model greatly, but gross non normality is potentially more serious as the t or F statistics and confidence and prediction intervals depend on the normality assumption.



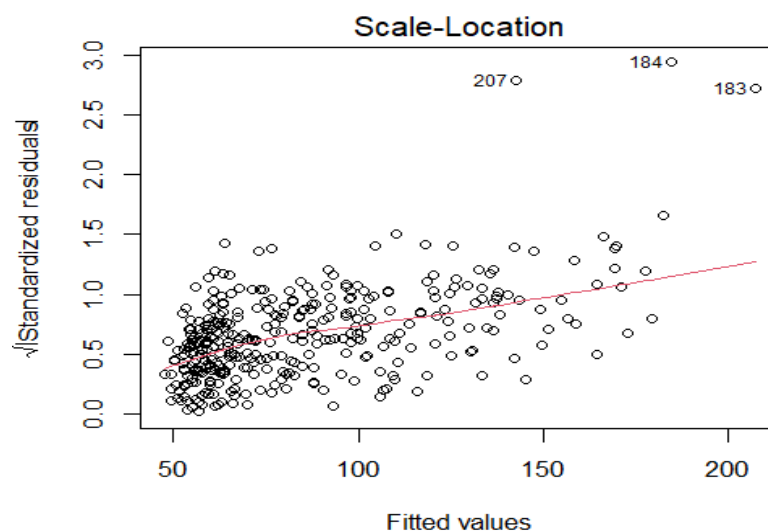
Interpretation :

The resulting points should lie approximately on a straight line. The straight line is usually determined visually, with emphasis on the central value rather than the extremes.

From residual analysis we have observed that the residuals followed normality as most of the points were along the line.

B.3) Heteroskedasticity (constant variance) :

Scale-Location plot shows whether residuals are spread equally along the ranges of input variables (predictor). The assumption of equal variance (homoscedasticity) could also be checked with this plot. If we see a horizontal line with randomly spread points, it means that the model is good. The plot is between fitted values and the square root of standardized residuals.



Interpretation: -

We can observe the above the scale-location plot for this regression model. The red line is roughly horizontal across the plot. then the assumption of homoscedasticity is satisfied for a given regression model. That is, the spread of the residuals is roughly equal at all fitted values.

IV] NON-PARAMETRIC TESTS :

(1) Kruskal Wallis test :

The Kruskal–Walli’s test is a rank-based test that is similar to the Mann–Whitney U test but can be applied to one-way data with more than two groups. It is a non-parametric alternative to the one-way ANOVA test, which extends the two-samples Wilcoxon test. Using the Kruskal-Wallis Test, it can be decided whether the population distributions are similar without assuming them to follow the normal distribution.

Hypothesis :

H0 : There is no significant difference in the median air quality index pollutant levels among the groups.

H1 : There is a significant difference in the median air quality index pollutant levels among the groups.

Output :

Kruskal Wallis chi-squared= 4185.3

Degree of freedom=15

p-value<2.2e⁻¹⁶

Interpretation :

The Kruskal-Walli’s chi-squared value is 4185.3. This is the test statistic that is used to determine if there are significant differences among the groups. The larger the chi-squared value, the more evidence there is against the null hypothesis.

The p-value is a crucial component of hypothesis testing. It represents the probability of observing a test statistic as extreme as the one calculated, assuming the null hypothesis is true. In this case, the p-value is very small (p-value < 2.2e-16), indicating strong evidence against the null hypothesis.

Therefore, there is strong evidence to suggest that there are significant differences in air quality index pollutants among the groups i.e. we reject H0.

(2) Friedman Test :

The Friedman test is a non-parametric statistical test used to detect differences between groups when the dependent variable is ordinal or continuous and the data is paired or matched. It is an extension of the Wilcoxon signed-rank test for more than two related groups. The test is often used when the assumptions for a repeated measures ANOVA cannot be met.

Hypothesis :

H0 : There is no significant difference in the mean ranks of air quality index pollutants.

H1 : There is a significant difference in the mean ranks of air quality index pollutants.

Output :

Friedman chi-squared =4988.2

Degree of freedom=15

P-value= 2.2×10^{-16}

Interpretation :

The Friedman chi-squared value is the test statistic. It is used to evaluate whether there are significant differences among the groups.

The p-value is a crucial indicator in hypothesis testing. It represents the probability of obtaining the observed results (or more extreme) if the null hypothesis is true. A very low p-value, such as 2.2×10^{-16} (scientific notation for $2.2 * 10^{(-16)}$), indicates strong evidence against the null hypothesis.

This suggests that there are statistically significant differences among the mean ranks of the related groups. Hence, we reject H0.

VJ FINDING ESTIMATOR FOR LOCATION PARAMETER OF CAUCHY DISTRIBUTION :

Since we have observed through our correlation and regression analysis that variables like PM10 and PM2_5 which have the most relative importance in AQI, studying these variables and building confidence interval would help us know how much the levels of pollutants can vary throughout the year, hence when compared to standard levels, we can know how safe air we breathe.

We have fitted the distribution for each variable, affecting AQI using easy fit,

Here are the results obtained for the best fit

| Variable | Distribution |
|---------------------|--------------|
| | |
| PM2.5 | Lognormal |
| PM10 | Cauchy |
| NO2 | Beta |
| SO2 | Normal |
| Temperature minimum | Weibull |
| Temperature maximum | Lognormal |
| Humidity | Beta |
| Ozone | Beta |
| Light | Lognormal |
| UV maximum | Lognormal |
| CO | Cauchy |
| CO2 | Cauchy |
| Sound | Beta |

we can see CO2 and PM10 follow Cauchy distribution, knowing the location parameter and finding the corresponding confidence interval would help us in determining to which extent the median values of AQI would vary throughout the year.

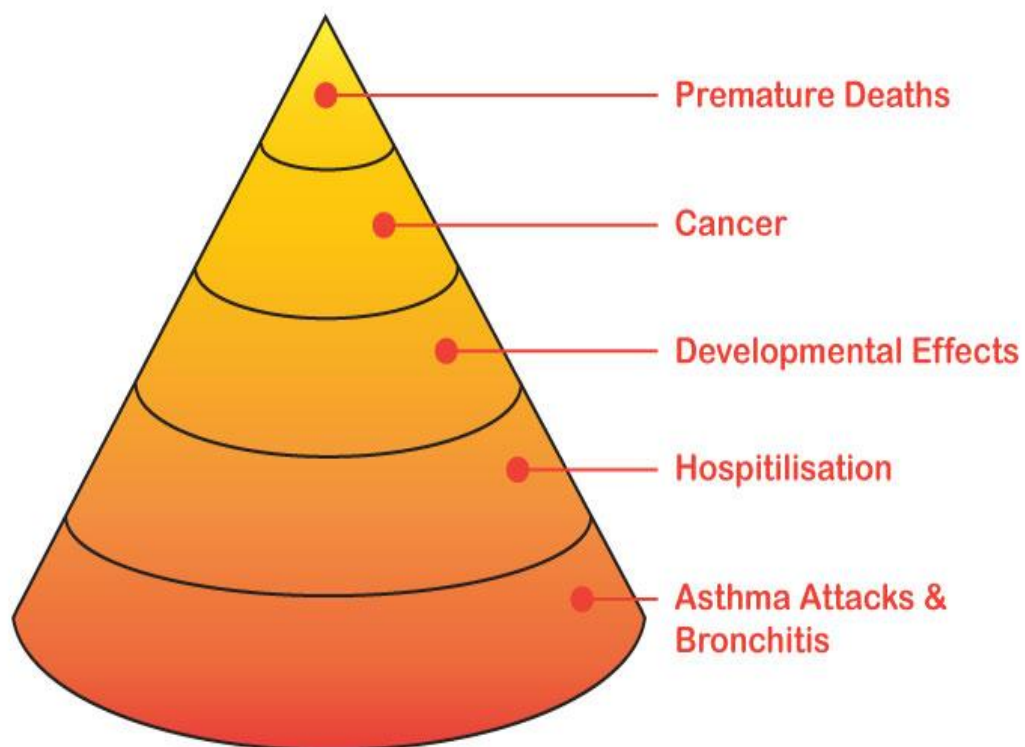
DO WE BREATHE SAFE AIR ?

PM₁₀ are very small particles found in dust and smoke. They have a diameter of 10 micrometres (0.01 mm) or smaller.

PM₁₀ particles are a common air pollutant. We have measure of PM10 in Pune City over a year

Health effects of PM10 particulate

PM10 particles are so small that they effectively act as a gas. When breathed in they penetrate deep into the lungs. Exposure to high concentrations of PM10 can result in a number of health impacts ranging from coughing and wheezing to asthma attacks and bronchitis to high blood pressure, heart attack, strokes and premature death.



The young and old and persons with existing medical conditions are most likely to be adversely affected by exposure to high PM10 concentrations.

Now we find confidence intervals to see if the levels of PM10 are within safe range for humans to breathe in Pune City, for that we would build a confidence interval for PM10.

| Air quality category | PM ₁₀ µg/m ³ averaged over 1 hour |
|----------------------|---|
|----------------------|---|

| | |
|-----------------------|---------------|
| Good | Less than 40 |
| Fair | 40–80 |
| Poor | 80–120 |
| Very poor | 120–300 |
| Extremely poor | More than 300 |

First, we aim to find an estimator for location parameter for the distribution of PM10(Cauchy distribution).

The **Cauchy distribution**, sometimes called the *Lorentz distribution*, is a family of continuous probability distributions which resemble the normal distribution family of curves. While the resemblance is there, it has a taller peak than a normal. And unlike the normal distribution, it's [fat tails](#) decay much more slowly.

The Cauchy distribution is well known for the fact that **it's expected value and other moments do not exist**. The median and mode *do* exist for Cauchy distribution.

The pdf of Cauchy distribution is given by

$$f(x) = \frac{\lambda}{\pi} \cdot \frac{1}{\lambda^2 + (x - u)^2}$$

There are many measures for central tendency

The commonly used are :

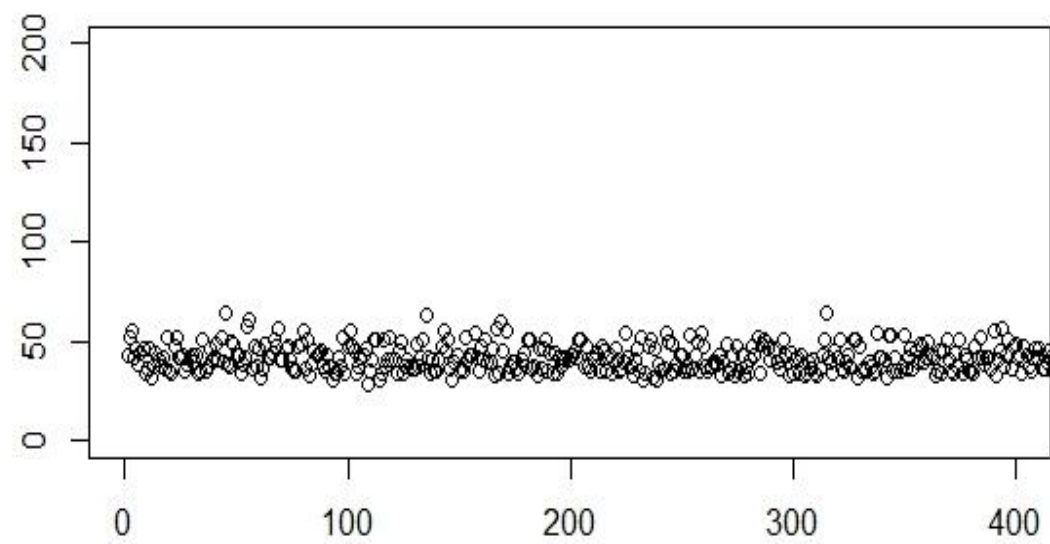
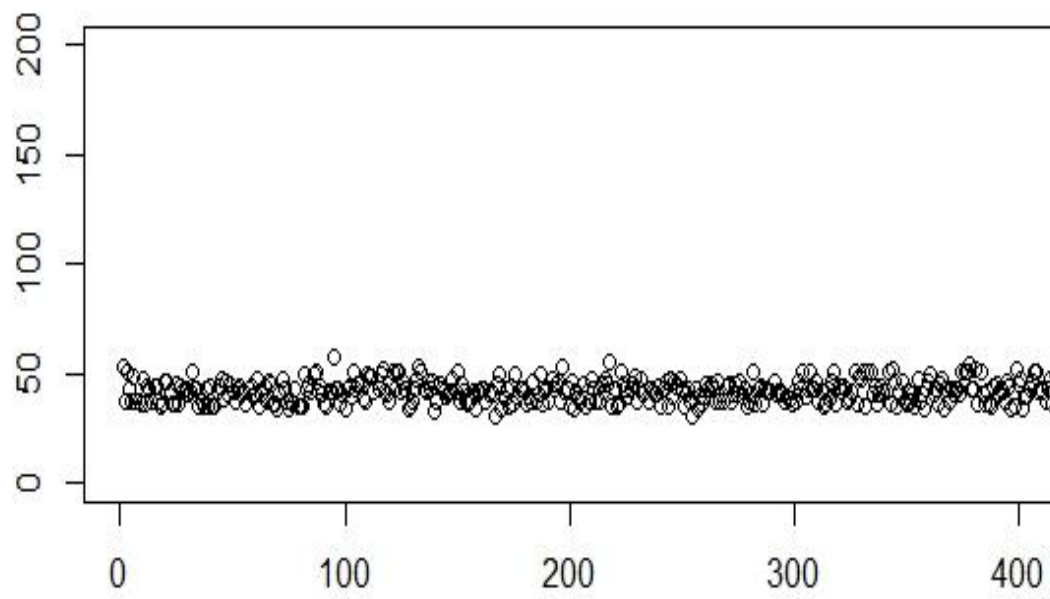
- i. Mean
- ii. Median
- iii. Mode

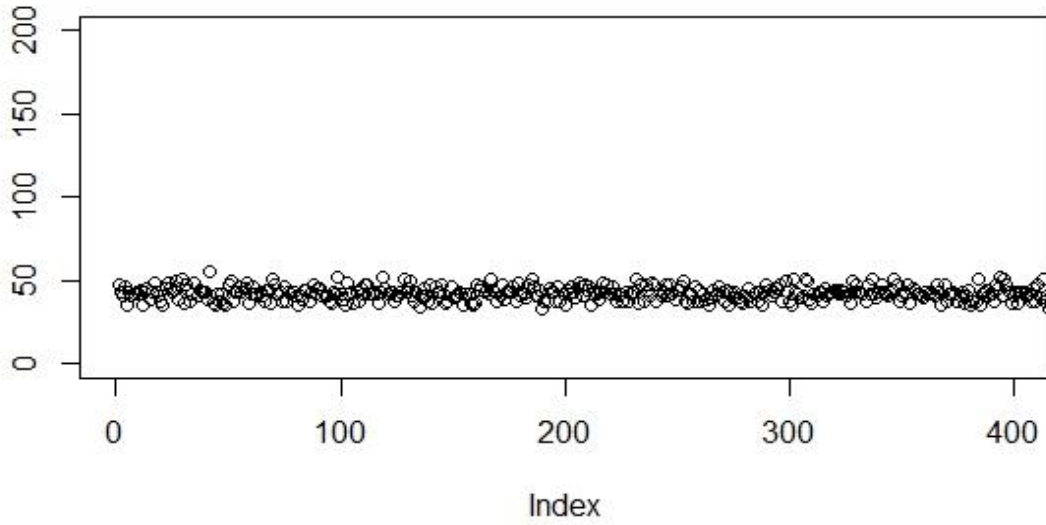
A sequence of estimators is said to be consistent for parameter θ iff for every $\varepsilon > 0$

$$P(|T - \theta| < \varepsilon) \rightarrow 1 \text{ as } n \rightarrow \infty.$$

Since sample mean is not a good estimator for location parameter, we will check if sample median converges in probability to μ , Since we have precalculated value of μ in our case to be 41.6 .

We plot different graphs of sample median over different sample sizes to visualize if it converges to μ





Here, we see, as we increase the sample size from 50 to 120, the plots become more and more concentrated near μ . Which is evident to say that sample mean converges to μ .

Further the distribution for median ordered statistic will be

$$f\left(x_{\frac{m+1}{2}}\right) = \frac{m!}{\left(\frac{m-1}{2}\right)!^2} (F(x))^{\frac{m-1}{2}} (1 - F(x))^{\frac{m-1}{2}} \cdot f(x)$$

Now to find the expected value of this pdf, note the following

Since Cauchy is distribution is symmetric and rest of the expression forms beta with both parameters equal which makes it symmetric, the product of pdf and rest of the expression forms an even function about μ .

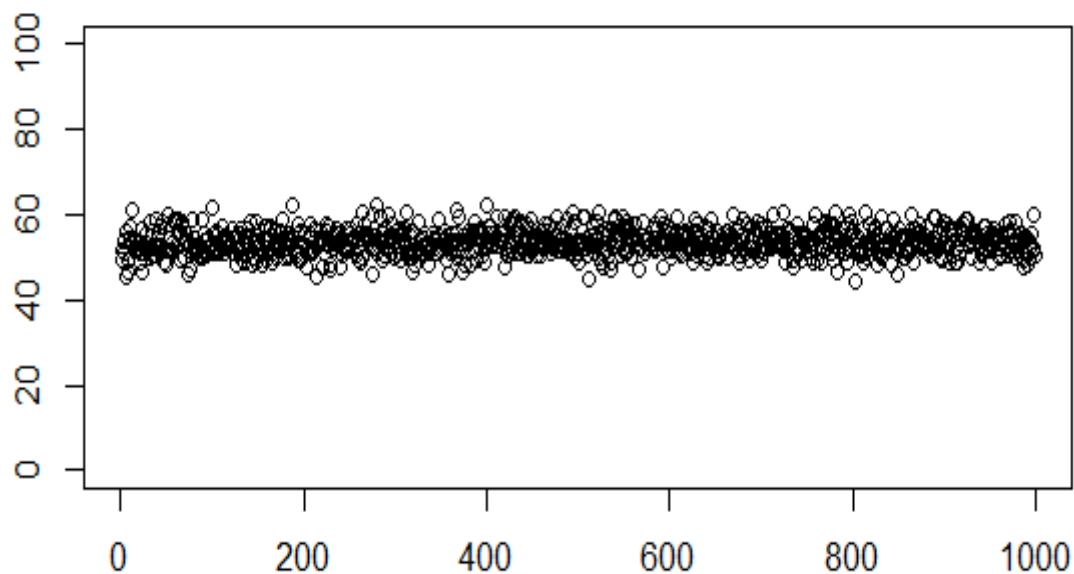
Hence the expectation of median distribution if exists will must be equal to μ .

FINDING BETTER ESTIMATOR :

Now we in effort to find a better estimator than sample median,

We try working with trimmed mean,

To check the consistency we have plotted the following graph



Looking at the graph we can decode that the value of trimmed mean lies around value of location parameter hence, trimmed mean can be used as an estimator.

A trimmed mean is stated as a mean trimmed by $x\%$, where x is the sum of the percentage of observations removed from both the upper and lower bounds. The trimming points are often arbitrary in that they follow rules of thumb rather than some optimized method of setting those thresholds. For example, a trimmed mean of 3% would remove the lowest and highest 3% of values, leaving the mean to be calculated from the 94% of remaining data.

The trimmed mean helps us tame outliers and obtain a robust measure of central tendency. By removing extreme values, this statistic can better represent typical dataset values.

To find which estimator works better sample median or sample trimmed mean,

We compare the variances for both the estimators in the following table.

Here the variances are numerically computed using R.

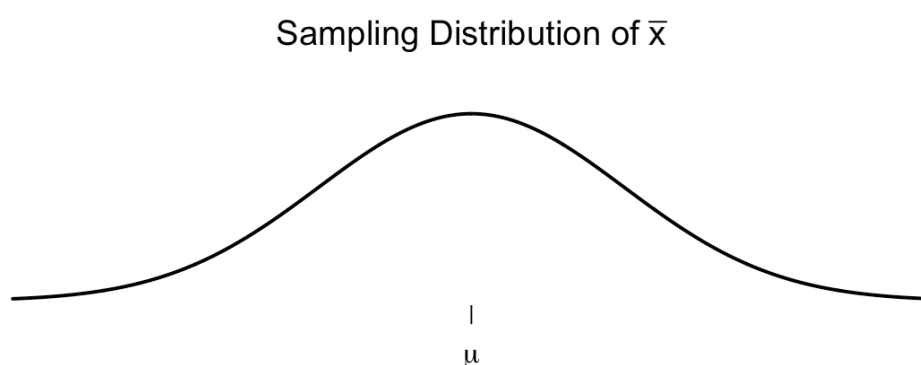
| Sample size (n) | Variance for sample median | Variance for sample trimmed mean |
|-----------------|----------------------------|----------------------------------|
| 50 | 88.39 | 52.00 |
| 60 | 73.49 | 48.90 |
| 70 | 66.41 | 38.59 |
| 80 | 58.07 | 35.24 |
| 90 | 51.67 | 30.10 |
| 100 | 50.43 | 28.24 |
| 200 | 25.39 | 14.11 |
| 300 | 18.30 | 8.78 |

Here it is clearly seen that sample trimmed mean performs better than sample median as there is more variation observed in sample median

Confidence interval :

Here we try to draw out a confidence interval for sample median using bootstrap method

- i. Here we have drawn N samples from the original sample with replacement.
- ii. For each of the samples, find the sample median.
- iii. Arrange these sample medians in order of magnitude.
(we use `boot` function in R for the aforementioned operation)
- iv. There the confidence intervals are calculated by 5 different methods in R.
- v. We will look for percentile method.
- vi. Percentile method and works when the sampling distribution is symmetric and the estimator, we are using is unbiased. For example, we expect that the sample median should be a good estimate of the population median μ and the sampling distribution of $\bar{x}(\text{median})$ should look something like the following.



Since we have already proved that the estimated value for sample median i.e. median ordered statistic is μ , the percentile method works here.

Further,

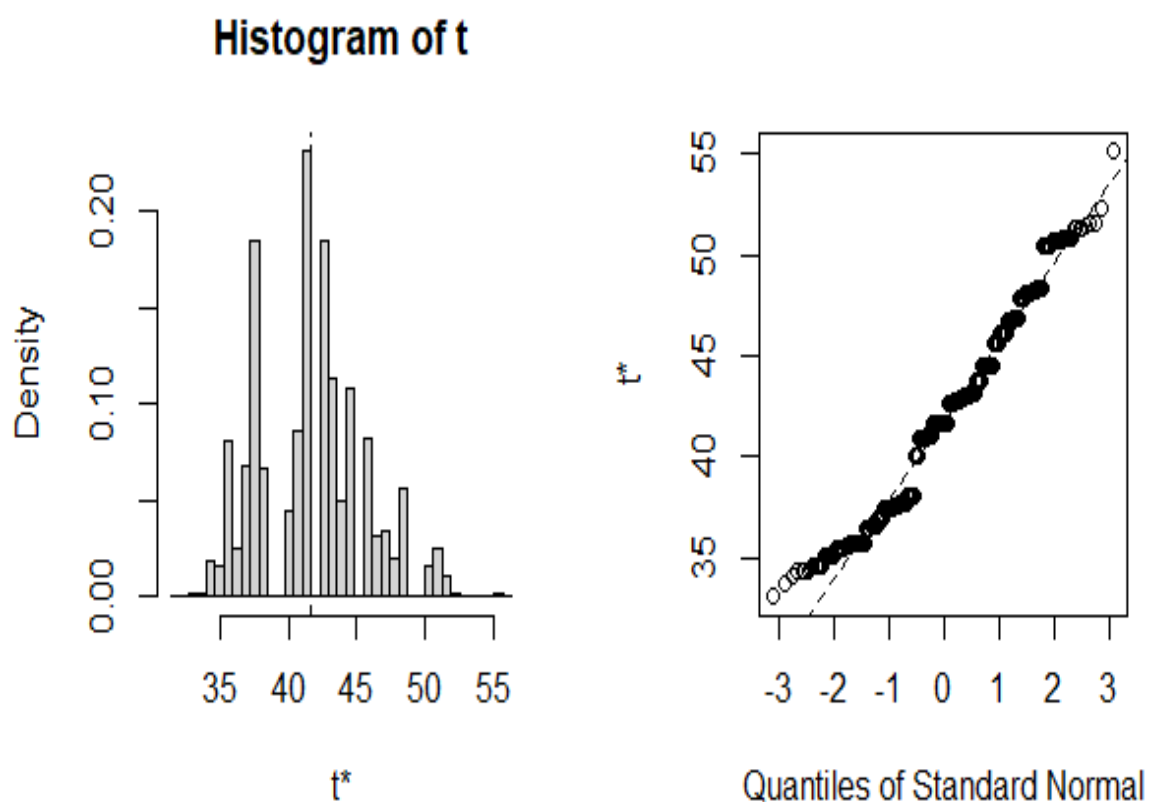
we have a sample of 365 scores with a sample median of \bar{x} . For the percentile method we simply draw a large number of bootstrapped samples (e.g. 1000) *with replacement* from a population made up of the sample data. We determine the median of each sample, call it \bar{X} , and create the sampling distribution of the median. We then take the $\alpha/2$ and $1 - \alpha/2$

percentiles (e.g. the $.025 \cdot 1000$ and $.975 \cdot 1000 = 25\text{th}$ and 975th bootstrapped statistic), and these are the confidence limits.

Percentile CI is : $(\theta_{(1-\alpha)/2}, \theta_{1-(1-\alpha)/2})$

Normal CI, works best when

- i. Statistic is unbiased.
- ii. Bootstrapped distribution is normally distributed.




The plot shows Q-Q plot and histogram for distribution of bootstrapped median for our data

A typical Normal CI would be : $\mu \pm z_{\alpha} \cdot se^*$

Percentile CI is generally not recommended because it performs poorly when it comes to weird-tailed distributions. Basic CI (also called *pivotal* or *empirical* CI) is much more robust to this. The rationale behind it is to compute differences between each bootstrap replication and t_0 and use percentiles of their distribution.

Formula for basic CI is :

$$(2t_0 - \theta_{1 - (1 - \alpha)/2}, 2t_0 - \theta_{(1 - \alpha)/2})$$

```
R 4.1.2 . ~/ 
Call:
boot(data = dt_daily_median_all$PM10, statistic = bootMedian,
      R = 1000)

Bootstrap Statistics :
      original    bias      std. error
t1*    41.675 0.014435    3.927306
> boot.ci(b)
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = b)

Intervals :
Level      Normal              Basic
95%    (33.96, 49.36 )    (32.67, 47.89 )

Level      Percentile          BCa
95%    (35.45, 50.68 )    (35.02, 50.44 )
Calculations and Intervals on Original Scale
Warning message:
In boot.ci(b) : bootstrap variances needed for studentized intervals
```

By confidence intervals calculated by all 5 methods we see that 95% of the time of year the values of PM10 lie between (33,49) $\mu\text{g}/\text{m}^3$ Which lies between good to fair levels of PM10 for year 2021-22. Also comparing to current PM10 levels of Maharashtra which are 62 $\mu\text{g}/\text{m}^3$. Which have worsened in short span of 2 years from good to fair quality.

So how to keep safe from PM10?

Pick an air purifier :

Air purifiers are great ways to directly remove PM10 from your air. Since most of us spend the majority of our lives indoors, taking care to eliminate PM10 from your home is paramount.

Use air quality monitor :

Particle pollution isn't just an outdoor occurrence. In many cases, the air quality in your home can be far worse than ambient air quality, and knowing your local AQI alone won't protect you from indoor PM10, especially if you use certain appliances like wood-burning stoves, fireplaces, and candles that release particulate matter into your home.

Wear a pollution mask :

Sometimes, braving a pollution-heavy day is unavoidable. When this occurs, it's best to wear a pollution mask. While pollution masks won't filter out all of the PM10 dust floating in the air, good-quality pollution masks do keep out the majority when appropriately fitted. If you do go out when PM10 levels are high, wearing a pollution mask will protect you from the worst of it.

Keep your home spick-and-span :

Keeping your house clean is one of the most basic ways to limit exposure to PM10 inside your own home. Pollutants like dander, Mold spores, and dust are all forms of PM10 readily thwarted by regular cleaning.

VI] TIME SERIES ANALYSIS:

Air quality data often exhibit temporal patterns, such as daily, weekly, or seasonal variations. Time series analysis helps in understanding these patterns, which can be essential for identifying trends, seasonal effects, and periodic fluctuations.

Time series models can be used to forecast future air quality levels based on historical data. This forecasting capability is crucial for decision-making and planning, such as implementing pollution control measures or issuing health advisories during periods of poor air quality.

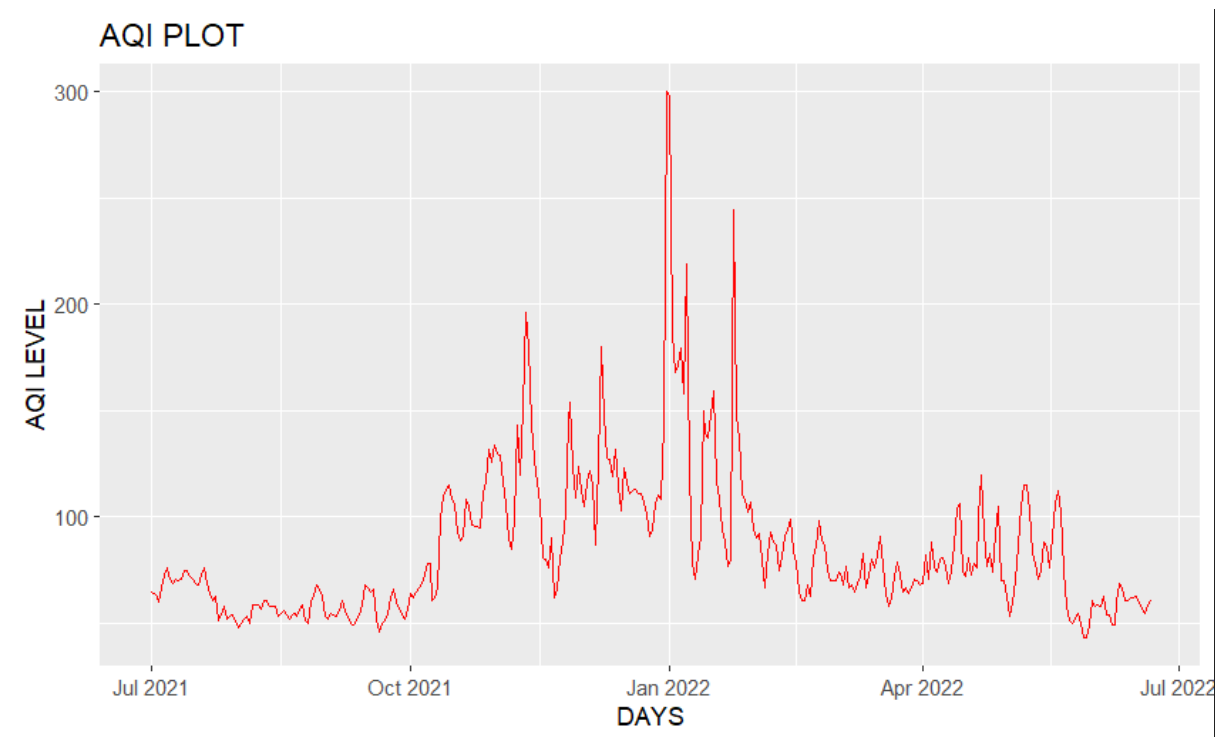
Time series plots and graphs provide a visual representation of air quality data over time, making it easier to identify trends, patterns, and outliers.

Since our data includes 11 months, no seasonality is observed, and regression analysis shows that the data is nonlinear. Due to the lack of a trend and seasonal fluctuation component, we use the Simple Exponential Smoothing technique.

Data :

Air quality index measures the quality of the air and helps us understand the environmental effects due to degraded quality of air. Here we have considered the data for AQI on the daily basis for almost a year. This data is from Pune city region.

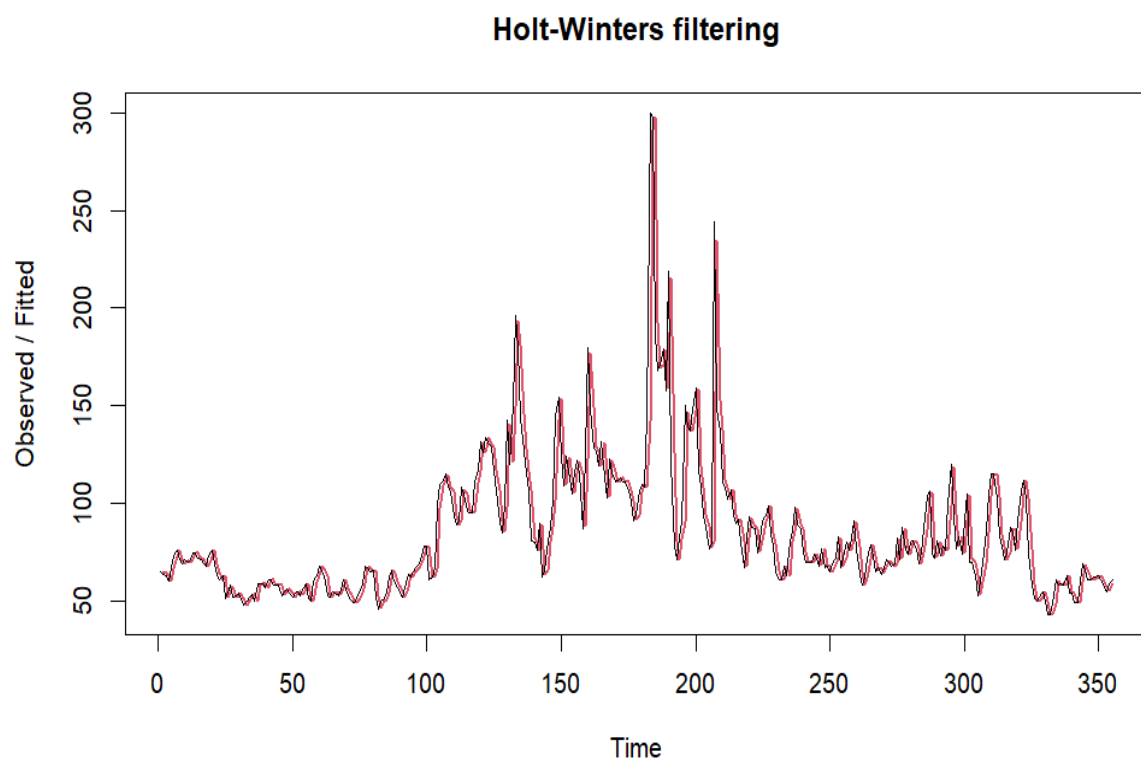
AQI of Pune 2021-22



We can further use these following tools to analyse the given time series :

- i. Smoothing using exponential and double exponential methods.
- ii. Forecasting further values.
- iii. Checking for stationarity in the data.

✚ Single exponential smoothing :



The Equation for simple exponential Smoothing is :

$$L_{t+1} = \alpha \times Y_t + (1 - \alpha) \times L_t$$

F_{t+1} : the forecast for the next time period $t+1$.

Y_t : the actual value observed at time period t .

F_t : the forecast for the current time period t .

α (alpha) : the smoothing parameter.

Output :

Holt-Winters exponential smoothing without trend and without seasonal component

Call :

HoltWinters(x = data1, alpha = NULL, beta = FALSE, gamma = FALSE)

Smoothing parameters :

alpha: 0.9420372

beta : FALSE

gamma: FALSE

Coefficients:

[1]

a 60.87106

Interpretation :

Smoothing Parameters :

Alpha (α) :

The value of alpha (0.9420372 in this case) is the smoothing parameter for the level. It controls the weight given to the most recent observation when updating the estimated level. A higher alpha places more weight on recent observations, making the model more responsive to short-term fluctuations.

Beta (β) :

In your output, beta is set to FALSE, indicating that the model is a simple exponential smoothing model without a trend component. If beta were present, it would represent the smoothing parameter for the trend.

Gamma (γ) :

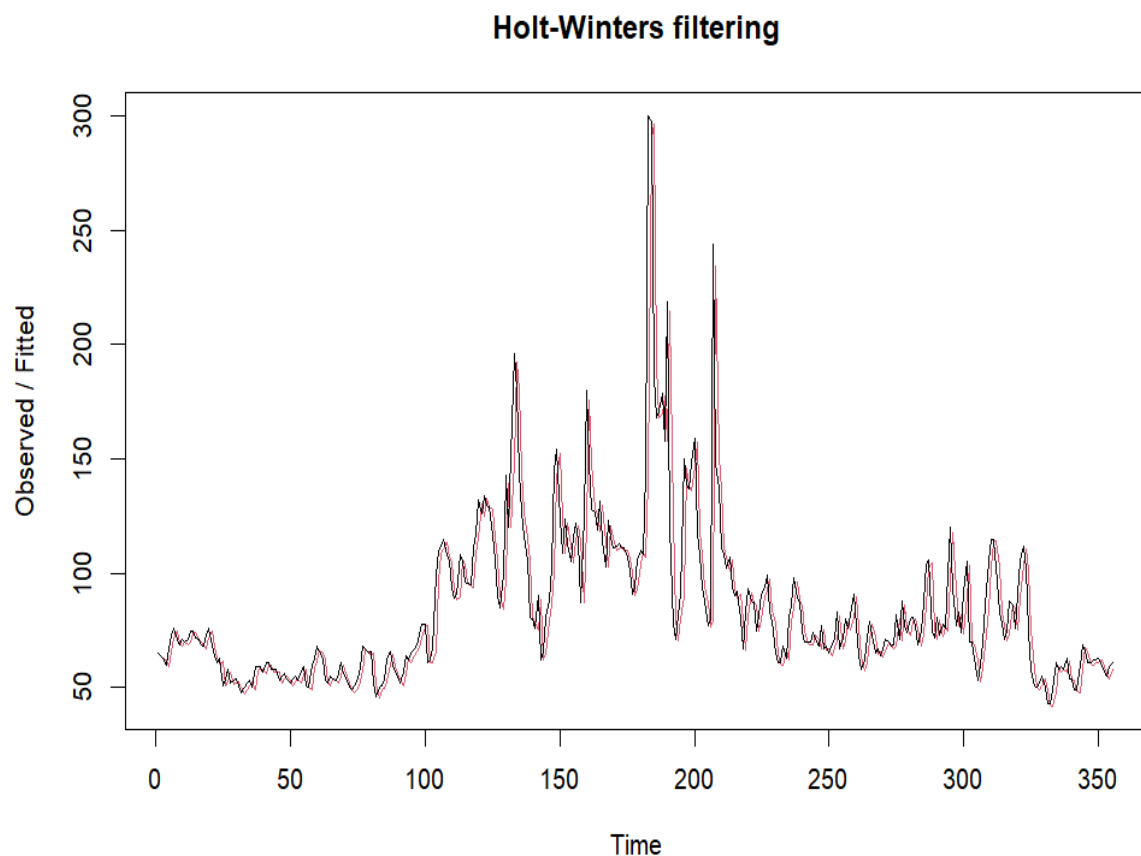
Similarly, gamma is set to FALSE, indicating that there is no seasonality component in your model. If gamma were present, it would represent the smoothing parameter for the seasonality.

The coefficient :

a : Represents the estimated level of time series. In this case, the estimated level is 60.87106., the model is emphasizing recent observations with a high alpha, and it does not include a trend or seasonality component. The estimated level of the time series is approximately 60.87.

🚦 Holt Winter's Double Exponential Smoothing :

Plot :



The equation for double exponential smoothing is :

Level equation : $L_t = \alpha Y_t + (1 - \alpha)(L_{t-1} + T_{t-1})$

Trend equation : $T_t = \beta(L_t - L_{t-1}) + (1 - \beta) T_{t-1}$

Where ,

L_t = smoothed level at time t

T_t = smoothed trend at time t

Y_t = actual observation at time t

α = smoothing parameter for level ($0 < \alpha < 1$)

β = smoothing parameter for trend ($0 < \beta < 1$)

L_{t-1} = smoothed level at time t-1

T_{t-1} = smoothed trend at time t-1

Forecast equation: $Y_{t+1} = L_t + T_t$

$Y_{t+1} = L_t + T_t$

Where, Y_{t+1} is the forecast for the next period

Call :

HoltWinters(x = data1, alpha = NULL, beta = NULL, gamma = FALSE)

Smoothing parameters :

alpha: 0.947276

beta : 0

gamma: FALSE

Coefficients:

[1]

a 60.82809

b -1.00000

Interpretation:

Smoothing Parameters:

Alpha (α) :

The smoothing parameter for the level. It controls the weight given to the most recent observation when updating

the smoothed value. In your case, alpha is approximately 0.947, indicating that a large weight is given to the most recent

observation in forecasting the next value.

Beta (β) :

The smoothing parameter for the trend. In your output, beta is 0, which suggests that your model does not consider any trend.

The forecast is solely based on the level.

Gamma (γ) :

The smoothing parameter for the seasonal component. In your output, gamma is FALSE, indicating that the model does not include seasonality.

Coefficients:

a : The estimated level of the time series. In output, the level (a) is approximately 60.83.

b : The estimated slope or trend. In your case, b is -1. This suggests a constant downward trend.

The simple exponential smoothing model is estimating a level of around 60.83 with no trend (beta = 0) and

no seasonality (gamma = FALSE). The negative trend coefficient suggests a decreasing pattern in the data.

The forecast for the next period is obtained by using the current level as the predicted value.

Stationarity of data :

To check the whether the data is stationary or not , we use augmented Dickey Fuller's test.

Hypothesis:

H0 : Given time series is non-stationary

H1: Given time series is stationary.

Output :

Here the p-value is 0.1007 which is greater than the level of significance

(L.O.S =0.05). This indicates that the given data is not stationary.

Further to make the time series stationary we can take the log of the time series.

Prophet:

Prophet is a forecasting procedure released by Facebook's Core Data Science team in 2017, published in the paper "Forecasting at Scale". It is an open-source software and is available in both Python and R programming language. According to its official website, it is "based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects." The formulation of Prophet can be described as follows:

$$y(t) = g(t) + s(t) + h(t) + \phi(t),$$

where,

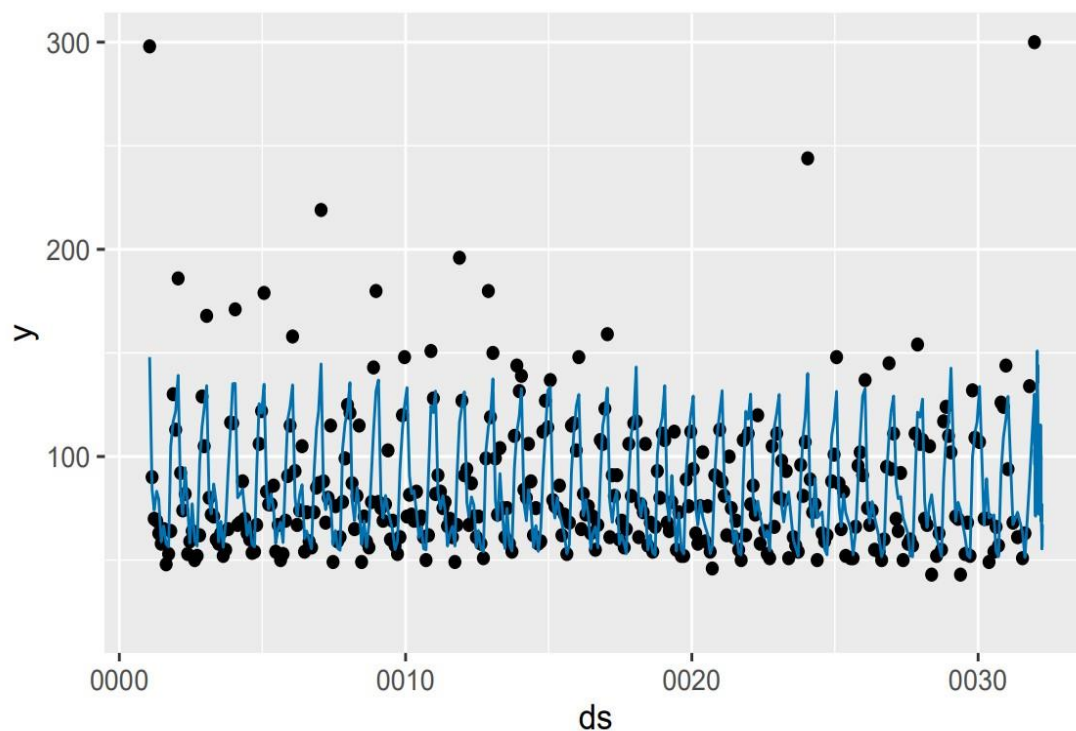
$y(t)$ = the predicted value

$g(t)$ = the trend function

$s(t)$ = seasonal changes

$h(t)$ = the effect of holidays

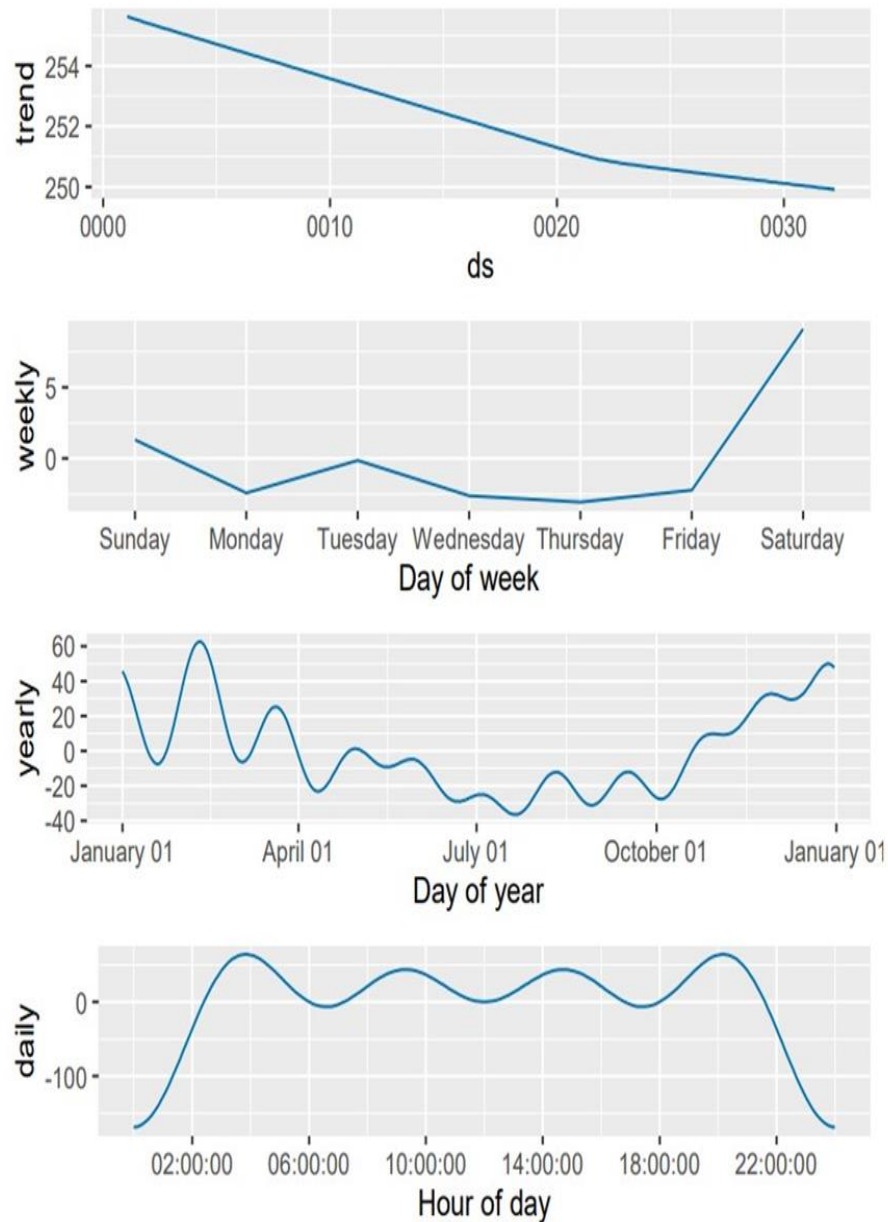
$\phi(t)$ = the normally distributed error.



Interpretation:-

The plot shows us the predicted values of AQI for next 100 days.

The following plot gives us the overall trend, seasonality being daily weekly and yearly about the given AQI data.



✚ Root Mean Square Error (RMSE) :

RMSE stands for Root Mean Square Error. It is a measure of the differences between values predicted by a model or estimator and the actual observed values. In the context of AQI (Air Quality Index) modelling, RMSE can be used as a metric to evaluate the performance of the predictive model used to estimate AQI values.

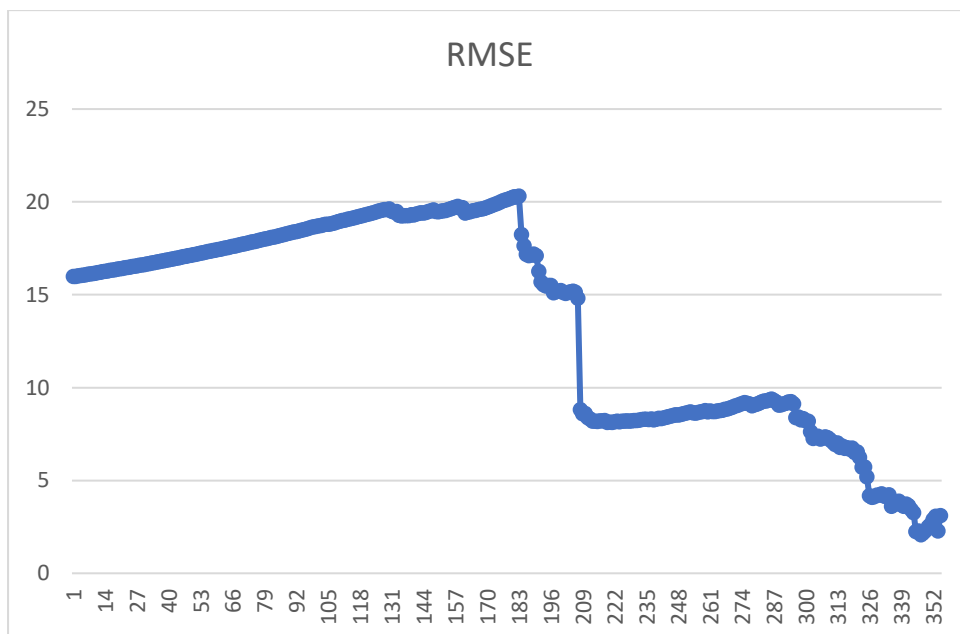
Mathematically, it is expressed as:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2}$$

Where:

- Y_i = Actual AQI value for observation i
- \hat{Y}_i = Predicted AQI value for observation i
- n = Number of observations

Plot :



Interpretation :

At the beginning of the graph, the RMSE is likely relatively low. This indicates that the model's predictions are relatively close to the actual observed values, suggesting good predictive performance during this phase.

The peak in the RMSE indicates a point where the model's predictive accuracy is at its worst. It's a critical point to investigate because it highlights where the model performs the poorest.

Following the peak, the RMSE gradually decreases. This indicates that the model's performance improves over time or with certain adjustments. The decrease may occur as the model learns from the data, adjusts parameters, or incorporates additional features that better capture the underlying patterns in the data.

Confidence Interval for RMSE :

Using R-Software:

95% Confidence Interval: (12.13970,14.09422)

- **Below Confidence Value :**

If the RMSE falls below the confidence value, it suggests that the model's predictions are generally close to the observed AQI values. In other words, the model is performing well within the expected range, and there's confidence that the predictions are reliable. This is an ideal scenario, indicating a good fit between the model and the data. The model can provide reliable information about air quality, enabling better decision-making for public health, environmental policies, and pollution control measures.

- **Within Confidence Value :**

When the RMSE falls within the confidence interval, it indicates that the model's performance is consistent with the expected variability. While the model might not be perfectly accurate, its predictions are still within an acceptable range of error. This suggests moderate confidence in the model's predictions and indicates a reasonable fit between the model and the data.

Environmental policies and interventions should be made with caution, taking into account the uncertainty in the predicted AQI values.

- **Above Confidence Value :**

If the RMSE exceeds the confidence value, it suggests that the model's predictions deviate significantly from the observed AQI values. In this case, the model's performance is worse than expected, and there's less confidence in the reliability of its predictions. This scenario may indicate that the model needs improvement, such as adjusting parameters, incorporating additional data, or using a different modelling approach. In terms of the environment, relying on such predictions for decision-making could lead to incorrect assessments of air quality, potentially resulting in inadequate pollution control measures, health risks for the population, or misallocation of resources.

8.CONCLUSION:

- The project delved into a comprehensive statistical analysis of Air Quality Index (AQI) data, employing various methodologies to gain insights into pollutant dynamics and their implications for human health and the environment.
- Firstly, parameter estimation was conducted, with the determination that the median serves as the most robust estimator for PM10, reflecting its central tendency accurately amidst potential outliers.
- Regression analysis played a crucial role in model building to discern the pollutants that significantly impact AQI. Through the rigorous process of backward elimination, we identified key pollutants whose variations exert substantial influence on AQI readings.
- Moreover, employing non-parametric tests, we investigated whether there existed a significant difference between the mean and median AQI values across different pollutant groups. This analysis provided valuable insights into the distributional characteristics of AQI data, enhancing our understanding of pollutant variability.
- Time series analysis was instrumental in capturing temporal trends in AQI readings, enabling the development of predictive models. By fitting appropriate models and forecasting future AQI values, we gained invaluable insights into the trajectory of air quality dynamics over time.
- The assessment of Root Mean Square Error (RMSE) facilitated the evaluation of forecast accuracy, providing a quantitative measure of model performance. Leveraging estimated parameters and RMSE values, confidence intervals were constructed for PM10 and AQI, offering a probabilistic framework for assessing uncertainty in pollutant concentrations.
- The interpretation of these findings underscores the critical importance of addressing air quality concerns for safeguarding human health and environmental integrity. Elevated levels of pollutants, as indicated by AQI variations, pose significant risks to respiratory health and overall well-being, highlighting the urgent need for proactive measures to mitigate air pollution sources and improve air quality standards. By elucidating the complex interplay between pollutants, AQI dynamics, and their repercussions for human health and the environment, this study contributes valuable insights towards informed decision-making and policy formulation aimed at fostering sustainable air quality management practices.

9. REFERENCES :

Research papers :

- i. Study and Analysis of Air Quality Index and Related Health Impact on Public Health

Department of Computer Science & Engineering, Koneru Lakshamaiah Education foundation, Vaddeshwaram,

- ii. Air Quality Index – A Comparative Study for Assessing the Status of Air Quality

CSIR-National Environmental Engineering Research Institute APC Division, CSIR-National Environmental Engineering Research Institute,

- iii. ANALYSIS OF AIR QUALITY INDEX Coimbatore Institute of Technology

iv. Root mean square error (RMSE) or mean absolute error (MAE) Arguments against avoiding RMSE in the literature. Chai^{1,2} and R. R. Draxler¹ NOAA Air Resources Laboratory (ARL), NOAA Center for Weather and Climate Prediction, 5830 University Research Court, College Park, MD 20740, USA² Cooperative Institute for Climate and Satellites, University of Maryland

- v. Hodges-Lehmann scale estimator for Cauchy distribution

University of Queensland.

- vi. ESTIMATORS FOR THE CAUCHY DISTRIBUTION

Los Alamos National Laboratory, New Mexico USA.

- vii. Estimation of the Location Parameter of Cauchy Distribution Using Some Variations of the Ranked Set Sampling Technique.

1. Department of Statistics, Yarmouk University, Jordan.

2. Department of Mathematics & Statistics, Jordan University of Science and Technology, Jordan.

- viii. Time Series Analysis for Psychological Research

Cancer Prevention Research Center University of Rhode Island.

ix. Statistical Characteristics of Air Quality Index Monitoring Stations: A Case Study of Seoul, Republic of Korea

Department of Urban Climatology, Office for Environmental Protection, Germany

Laboratory of Landscape Architecture, Department of Horticultural Science, Faculty of Bioscience and Industry, College of Applied Life Science, Jeju National University, Republic of Korea

Chair of Environmental Meteorology, Albert-Ludwigs-University of Freiburg, Germany.

x. Environmental and Health Impacts of Air Pollution : A review

University of Glasgow.

xi. Prediction of Air Quality Index Using Supervised Machine Learning

Faculty of Computing, Blekinge Institute of Technology, Karlskrona, Sweden.

References Books :

- i. Statistical Inference - Casella Berger
- ii. A First Course on Parametric Inference - BK Kale
- iii. Theory of Point Estimation - Erich Leo Lehman
- iv. Applied Nonparametric Statistics – Daniel W. W.
- v. Introduction to Statistical Learning - Robert Tibshirani
- vi. Analytics Stories - Wayne Winston
- vii. Forecasting and Time Series Analysis - Montgomery, D.C. and Johnson L.A., McGraw Hill.
- viii. Design and Analysis of Experiments - Angela Dean and Daniel Voss
- ix. Brandon Foltz – YouTube
- x. Statistical learning in R / Introduction to statistical learning – Stanford Online (YouTube series)
- xi. Zedstatistics – YouTube

10.Softwares used :

- i. MS-Excel
- ii. R-studio
- iii. Python

