**DATA ANALYSIS OF AIR QUALITY IN INDIA**

*A project report submitted to ICT Academy of Kerala*

*in partial fulfillment of the requirements*

*for the certification of*

**CERTIFIED SPECIALIST**

**IN**

**DATA SCIENCE & ANALYTICS**

submitted by

**Team 5**

**Members**

**ALIYA P SUBEER**

**HARITHA C M**

**IRFANA ISMATH**

**PRINCY P**

**SETHULEKSHMI S**

**SREEJAYAN S**



**ICT ACADEMY OF KERALA**

**THIRUVANANTHAPURAM, KERALA, INDIA**

**July 2023**

**LIST OF FIGURES**

|  |  |
| --- | --- |
| Figure 1 | AQI health impacts |
| Figure 2 | AQI dataset info |
| Figure 3 | Statistical values of all the numerical attributes using describe() method |
| Figure 4 | Histogram for each numerical attribute |
| Figure 5 | Correlation between every pair of attributes |
| Figure 6 | Box plot for each attribute |
| Figure 7 | City wise analysis of PM2.5&PM10 |
| Figure 8 | Year wise analysis of PM2.5&PM10 |
| Figure 9 | Year wise analysis of NO&NO2 |
| Figure 10 | Yearly trend of SO2 |
| Figure 11 | Yearly trend of NH3 & NOx |
| Figure 12 | Yearly trend of BTX |
| Figure 13 | Year wise analysis of CO & O3 |
| Figure 14 | Count of AQI bucket |
| Figure 15 | City wise AQI range |
| Figure 16 | Code for calculating NOx feature |
| Figure 17 | Code for AQI calculation |
| Figure 18 | Code for AQI bucket filling |
| Figure 19 | Missing values of the features and their percentages |
| Figure 20 | Describe function after outlier handling using flooring/capping method |
| Figure 21 | Visualization of IQR method |
| Figure 22 | Describe function after outlier handling using IQR method |
| Figure 23 | Describe function after outlier handling using log method |
| Figure 24 | Describe function after outlier handling using Z-Score method |
| Figure 25 | Describe function after outlier handling using Isolation Forest |
| Figure 26 | Describe function after outlier handling using DBSCAN |
| Figure 27 | Describe function after outlier handling using Robust Scaler Method. |

**LIST OF ABBREVIATIONS**

|  |  |
| --- | --- |
| AQI | Air Quality Index |
| EDA | Exploratory Data Analysis |
| IQR | Inter Quartile Range |
| ML | Machine Learning |
| CPCB | Central Pollution Control Board |
| WHO | World Health Organization |
| PCA | Principal Component Analysis |
|  |  |

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **CHAPTER** | **TITLE** | **PAGE NUMBER** |
|  | LIST OF FIGURES | 2 |
|  | LIST OF ABBREVIATIONS | 4 |
|  | ABSTRACT | 7 |
| 1 | PROBLEM DEFINITION  1.1 Overview  1.2 Problem Statement | 8 |
| 2 | INTRODUCTION | 9 |
| 3 | LITERATURE SURVEY | 10 |
| 4 | EXPLORATORY DATA ANALYSIS    4.1 Data Set  4.2 Overview of the dataset  4.3 Descriptive Statistical Summary  4.4 Distribution of each attribute  4.5 Correlation matrix  4.6 Outliers 4.7 Analysis of features | 12  15  15  16  17  17  18 |
| 5 | DATA PREPROCESSING  5.1 Dataset edited  5.2 Feature selection  5.3 Handling missing values  5.4 Outliers  5.4.1 Detection & Handling outliers  5.5 Encoding & Feature scaling | 21  22  23  24  27 |
| 6 | DATA MODELLING | 28 |
| 7 | RESULT | 0 |
| 8 | WEB APPLICATION | 0 |
| 9 | CONCLUSION | 33 |
| 10 | REFERENCES | 34 |
|  |  |  |

**ABSTRACT**

The survival of mankind cannot be imagined without air. Consistent developments in almost all realms of modern human society have adversely affected the health of the air. Daily industrial, transport, and domestic activities are stirring hazardous pollutants in our environment. Monitoring and predicting air quality have become essentially important in this era, especially in developing countries like India. In contrast to the traditional methods, the prediction technologies based on machine learning techniques are proved to be the most efficient tools to study such modern hazards. The present work investigates **Air Quality Data in India (2015 - 2020) and (2020-2023) Dataset** taken from Kaggle consisting of AQI data published by CPCB (Central Pollution Control Board).

The main focus of this project is learning about EXPLORATORY DATA ANALYSIS, DATA PRE-PROCESSING, FEATURE ENGINEERING, MODELING AND PREDICTION of data by supervised algorithms i.e. (Linear Regression (regression), Logistic Regression, Random Forest Classifier, Decision Tree Classifier (classification)). The main focus of this particular project is AQI (Air Quality Index) prediction, and factors that affects AQI i.e. (SO2, NO2, SPM, RSPM). In this project, a basic data analysis is done on India's Air Quality data, and the value of the air quality index is predicted based on the given features of concentration of sulphur dioxide, nitrogen dioxide, respirable suspended particulate matter, and suspended particulate matter. The air quality is classified as good, satisfactory, moderately polluted, poor, very poor, and severe.

1. **PROBLEM DEFINITION**
   1. **Overview**

Machine learning is used to devise complex models and algorithms that lend themselves to prediction; in commercial use, this is known as predictive analytics. These analytical models allow researchers, data scientists, engineers, and analysts to “produce reliable, repeatable decisions and results” and uncover “hidden insights” through learning from historical relationships and trends in the dataset (input).

We forecast the air quality of India by using machine learning to predict the air quality index of a given area. In contrast to the traditional methods, the prediction technologies based on machine learning techniques are proved to be the most efficient tools to study such modern hazards. The present work investigates nine years of air pollution data from 26 Indian cities for air quality analysis and prediction. This study assessed the performance of the best data mining models (-- , -- and --) for predicting the accurate AQI data in some of India’s most populous and polluted cities.

**1.2 Problem Statement**

The prediction of air quality index is a complex problem that requires the consideration of multiple factors. Industrial, transport, and domestic activities release hazardous pollutants into the environment, affecting air quality. Additionally, atmospheric patterns such as rain, air pressure, and temperature can influence the volume of each pollutant in the air. To accurately predict air quality index, it is necessary to develop a model that takes into account these factors and their interactions. Machine learning techniques have shown promise in this regard, allowing for the development of predictive models that can accurately forecast air quality.

1. **INTRODUCTION**

Air pollution occurs when harmful or excessive quantities of substances including gases, particles, and biological molecules are introduced into the Earth's atmosphere. Air pollution in India is a serious issue, ranking higher than smoking, high blood pressure, child and maternal malnutrition, and risk factors for diabetes. At least 140 million people breathe air 10 times or more over the WHO safe limit and 13 of the world's 20 cities with the highest annual levels of air pollution are in India. Air pollution contributes to the premature deaths of 2 million Indians every year. Poor air conditions lead to other contemporary environmental issues like global warming, acid rain, reduced visibility, smog, aerosol formation, climate change, and premature deaths. Scientists have realized that air pollution bears the potential to affect historical monuments adversely. Vehicle emissions, atmospheric releases of power plants and factories, agriculture exhausts, etc. are responsible for increased greenhouse gases. The greenhouse gases adversely affect climate conditions and consequently, the growth of plants. In urban areas, most emissions come from vehicles and industry, whereas in rural areas, much of the pollution stems from biomass burning for cooking and keeping warm. In autumn and winter months, large scale crop residue burning in agriculture fields – a low cost alternative to mechanical tilling - is a major source of smoke, smog and particulate pollution.

The ***Air Quality Index (AQI)***, an assessment parameter is related to public health directly. A higher level of AQI indicates more dangerous exposure for the human population. Therefore, the urge to predict the AQI in advance motivated the scientists to monitor and model air quality. Monitoring and predicting AQI, especially in urban areas has become a vital and challenging task with increasing motor and industrial developments.

Various models have been exercised in the literature to predict AQI, like statistical, deterministic, physical, and Machine Learning (ML) models. The traditional techniques based on probability, and statistics are very complex and less efficient. The ML-based AQI prediction models have been proved to be more reliable and consistent. Advanced technologies and sensors made data collection easy and precise. The accurate and reliable predictions through such huge environmental data require rigorous analysis which only ML algorithms can deal with efficiently. The present work investigates nine years of air pollution data of the Indian cities and analyzes twelve air pollutants and AQI. The dataset is preprocessed and cleaned first, then methods of data visualization are applied to develop better insights and to investigate hidden patterns and trends. This work exploits the essence of correlation coefficient with ML models which has been exercised by very few scholars in the literature. The data imbalance is identified and addressed with a resampling technique. Popular ML models are exercised in context with this resampling technique. Their performances are then compared through standard metrics.

**3. LITERATURE SURVEY**

Gopalakrishnan ([2021](https://link.springer.com/article/10.1007/s13762-022-04241-5#ref-CR15)) combined Google’s Street view data and ML to predict air quality at different places in Oakland city, California. He targeted the places where the data were unavailable. The author developed a web application to predict air quality for any location in the city neighbourhoods. Sanjeev ([2021](https://link.springer.com/article/10.1007/s13762-022-04241-5#ref-CR28)) studied a dataset that included the concentration of pollutants and meteorological factors. The author analysed and predicted the air quality and claimed that the *Random Forest (RF)* classifier performed the best as it is less prone to over-fitting.

Castelli et al. ([2020](https://link.springer.com/article/10.1007/s13762-022-04241-5#ref-CR9)) endeavoured to forecast air quality in California in terms of pollutants and particulate levels through the *Support Vector Regression (SVR)* ML algorithm. The authors claimed to develop a novel method to model hourly atmospheric pollution. Doreswamy et al. ([2020](https://link.springer.com/article/10.1007/s13762-022-04241-5#ref-CR12)) investigated ML predictive models for forecasting PM concentration in the air. The authors studied six years of air quality monitoring data in Taiwan and applied existing models. They claimed that predicted values and actual values were very close to each other. Liang et al. ([2020](https://link.springer.com/article/10.1007/s13762-022-04241-5#ref-CR18)) studied the performances of six ML classifiers to predict the AQI of Taiwan based on 11 years of data. The authors reported that *Adaptive Boosting (AdaBoost)* and *Stacking Ensemble* are most suitable for air quality prediction but the forecasting performance varies over different geographical regions. Madan et al. ([2020](https://link.springer.com/article/10.1007/s13762-022-04241-5#ref-CR19)) compared twenty different literary works over pollutants studied, ML algorithms applied, and their respective performances. The authors found that many works incorporated meteorological data such as humidity, wind speed, and temperature to predict pollution levels more accurately. They found that the *Neural Network (NN)* and boosting models outperformed the other eminent ML algorithms. Madhuri et al. ([2020](https://link.springer.com/article/10.1007/s13762-022-04241-5#ref-CR20)) mentioned that wind speed, wind direction, humidity, and temperature played a significant role in the concentration of air pollutants. The authors employed supervised ML techniques to predict the AQI and found that the *RF* algorithm exhibited the least classification errors. Monisri et al. ([2020](https://link.springer.com/article/10.1007/s13762-022-04241-5#ref-CR22)) collected air pollution data from various sources and endeavoured to develop a mixed model for predicting air quality. The authors claimed that the proposed model aims to help people in small towns to analyse and predict air quality. Nahar et al. ([2020](https://link.springer.com/article/10.1007/s13762-022-04241-5#ref-CR23)) developed a model to predict AQI based on ML classifiers. Their authors studied the data collected over the tenure of 28 months by the ministry of environment, Jordan, and identified the concentrations of pollutants. Their proposed model detected the most contaminated areas with satisfying accuracy. Patil et al. ([2020](https://link.springer.com/article/10.1007/s13762-022-04241-5#ref-CR24)) presented some literary works on various ML techniques for AQI modelling and forecasting. The authors found that *Artificial Neural Network (ANN)*, *Linear Regression (LR),* and *Logistic Regression (LogR)* models were exploited by most of the scholars for AQI prediction.

Bhalgat et al. ([2019](https://link.springer.com/article/10.1007/s13762-022-04241-5#ref-CR8)) applied the ML technique to predict the concentration of SO2 in the environment of Maharashtra, India. The authors concluded that being highly polluted, some cities of this Indian province require grave attention. The authors mentioned that their model was not capable of exhibiting expected outputs. Mahalingam et al. ([2019](https://link.springer.com/article/10.1007/s13762-022-04241-5#ref-CR21)) developed a model to predict the AQI of smart cities and tested it in Delhi, India. The authors reported that the medium Gaussian *Support Vector Machine (SVM)* exhibited maximum accuracy. The authors claim that their model can be used in other smart cities too. Soundari et al. ([2019](https://link.springer.com/article/10.1007/s13762-022-04241-5#ref-CR30)) developed a model based on *NNs* to predict the AQI of India. The authors claimed that their proposed model could predict the AQI of the whole county, of any province, or of any geographical region when the past data on concentration of pollutants were available.

Sweileh et al. ([2018](https://link.springer.com/article/10.1007/s13762-022-04241-5#ref-CR31)) came up with a very interesting study about the analysis of global peer-reviewed literature about air pollution and respiratory health. The authors extracted 3635 documents from the Scopus database published between 1990 and 2017. They observed that there was a substantial increase in publications from 2007 to 2017. The authors reported active countries, institutions, journals, authors, international collaborations in the realm and concluded that research works on air pollution and respiratory health had been receiving a lot of attention. They suggested securing public opinions about mitigation of outdoor air pollution and investment in green technologies. Zhu et al. ([2018](https://link.springer.com/article/10.1007/s13762-022-04241-5#ref-CR32)) refined the problem of AQI prediction as a multi-task learning problem. The authors utilized large-scale optimization techniques and endeavoured to reduce the number of parameters. Based on their empirical results, they claimed that the proposed model exhibited better results than existing regression models.

Bellinger et al. ([2017](https://link.springer.com/article/10.1007/s13762-022-04241-5#ref-CR7)) carried out a detailed literature analysis on the application of ML and data mining methods toward air pollution epidemiology. The authors found that the researchers from Europe, China, and the USA were very active in this realm and the following classifiers had been widely applied: *Decision Tree (DT)*, *SVMs*, *K-means clustering,* and the *APRIORI* algorithm. Rybarczyk and Zalakeviciute ([2017](https://link.springer.com/article/10.1007/s13762-022-04241-5#ref-CR26)) endeavoured to develop a model that correlated traffic density with air pollution. The author mentioned that such traffic data collection was economical, and integrating it with meteorological features boosted accuracy. The authors found that the hybrid model performed the best and accuracy based on morning time data was the highest.

It has been observed that research works in air quality analysis and prediction for Indian cities acquired lesser attention from scholars. In spite of the fact that out of the ten most polluted cities in the world, nine cities are Indian (Deshpande [2021](https://link.springer.com/article/10.1007/s13762-022-04241-5#ref-CR11)), very few researchers investigated AQI prediction from the Indian perspective. The present work endeavours to fill this gap by studying 5 years of substantial air pollution data from twenty-three Indian cities. The current study is an earnest attempt to contribute to the literature with novel ideas of data visualizations, exploiting correlation coefficient-based statistical outliers for analytics, and comparison of five key ML models over standard performance metrics.

**4. EXPLORATORY DATA ANALYSIS**

This section of the present study deals with data exploration and analysis for finding various hidden patterns present in the dataset. Exploratory data analysis is the first step in data analytics which is performed before applying any ML model. Under this, the following important things are being analyzed exploring statuses and trends of air pollutants over the past nine years i.e. from 2015 to 2023, exploring the distribution of pollutants in the air along with polluted cities with their average AQI values, estimating top pollutants which are directly involved in increasing the AQI values.

**4.1** **Data Set**

* The dataset contains air quality data and AQI (Air Quality Index) at hourly and daily level of various stations across multiple cities in India.
* The analysis combines different datasets for the year (2015-2020) and (2020-2023) to provide a comprehensive understanding of the trends and patterns in various locations over time. The dataset can be accessed through the following link, which hosts a collection of diverse datasets for machine learning research.

<https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india>

* This dataset possesses observations from 2015 to 2023 and it is comprised of 16 features with 53035 instances from 26 different Indian cities.
* Data set contains 3 categorical columns and 13 numerical columns.
* We have 15 independent variables and 2 target variables, i.e. AQI and AQI\_Bucket.
* Measurement units
  + ug/m3: micro gram/cubic meter
  + ppm: Parts Per Million
* **Features description**

1. **City**: Ahmedabad, Aizawl, Amaravati, Amritsar, Bengaluru, Bhopal, Brajrajnagar, Chandigarh, Chennai, Coimbatore, Delhi, Ernakulam, Gurugram, Guwahati, Hyderabad, Jaipur, Jorapokhar, Kochi, Kolkata, Lucknow, Mumbai, Patna, Shillong, Talcher, Thiruvananthapuram, Visakhapatnam.
2. **Date:** Date in which the samples are taken.
3. **PM2.5:** It refers to particles that have diameter less than 2.5 micrometers (more than 100 times thinner than human air) and remain suspended for longer. These particles are formed as a result of burning fuel and chemical reactions that take place in the atmosphere.
4. **PM10:** PM10 are the particles with a diameter of 10 micrometers and they are also called fine particles. An environmental expert says that PM10 is also known as respirable particulate matter. Particulate matter is a complex mixture of smoke, metals, nitrates, sulphates, dust water and rubber etc.
5. **NO:** Nitric oxide is not considered to be hazardous to health at typical ambient concentrations, but nitrogen dioxide can be. NOx gases react to form smong and acid rain as well as being central to the formation of fine particles (PM) and ground level ozone.
6. **NO2:** Nitrogen dioxide is part of a group of gaseous air pollutants produced as a result of road traffic and other fossil fuel combustion processes. Its presence in air pollutants.
7. **NOx:** NOx pollution refers to the release of nitrogen oxides (NOx) into the atmosphere. NOx is a group of toxic gases formed during high-temperature combustion processes, especially in vehicles, power plants, and industrial facilities. It is a major contributor to air pollution, which can have harmful effects on human health, the environment, and wildlife. The adverse effects of NOx pollution include respiratory problems, reduced visibility, acid rain, and eutrophication (excess nutrients leading to algal blooms and loss of aquatic life). Governments and industries have implemented various measures to reduce NOx emissions, such as stricter emissions standards, cleaner technologies, and improved transportation systems.
8. **NH3**: Gaseous ammonia (NH3) is the most abundant alkaline gas in the atmosphere. NH3 plays a significant role in the formation of atmospheric particulate matter, visibility degradation and atmospheric deposition of nitrogen to sensitive ecosystem.
9. **CO:** It is produced in the incomplete combustion of carbon-containing fuels such as gasoline, natural gas, oil, coal and wood.
10. **SO2:** It is formed when fuel containing sulphur, such as coal and oil, is burned, creating air pollution. It affects the environment when it reacts with substances in the atmosphere to form acid rain.
11. **O3:** It is harmful to air quality outside of the ozone layer. Ground level ozone is a colourless and highly irritating gas that forms just above the earth's surface.
12. **Benzene:** The benzene in indoor comes from product that contain benzene such as glues, paints, furniture wax, and detergents. The air around hazardous waste sites or gas stations can contain higher level of benzene than in other areas.
13. **Toluene:** Motor vehicle and industrial emissions are the major sources of pollutants.
14. **Xylene:** Motor vehicle emissions are the predominant source of xylene in the urban air environment. Evaporation from petroleum fuel storage facilities and service stations.
15. **AQI:** The air quality index is an index for reporting air quality on a daily basis. It is a measure of how air pollution affects one's health with in short time period. The purpose of the AQI is to help know how the local air quality impacts their health.
16. **AQI\_Bucket:** It is variable of AQI, values such as good, satisfactory, moderate, poor, very poor, and severe.

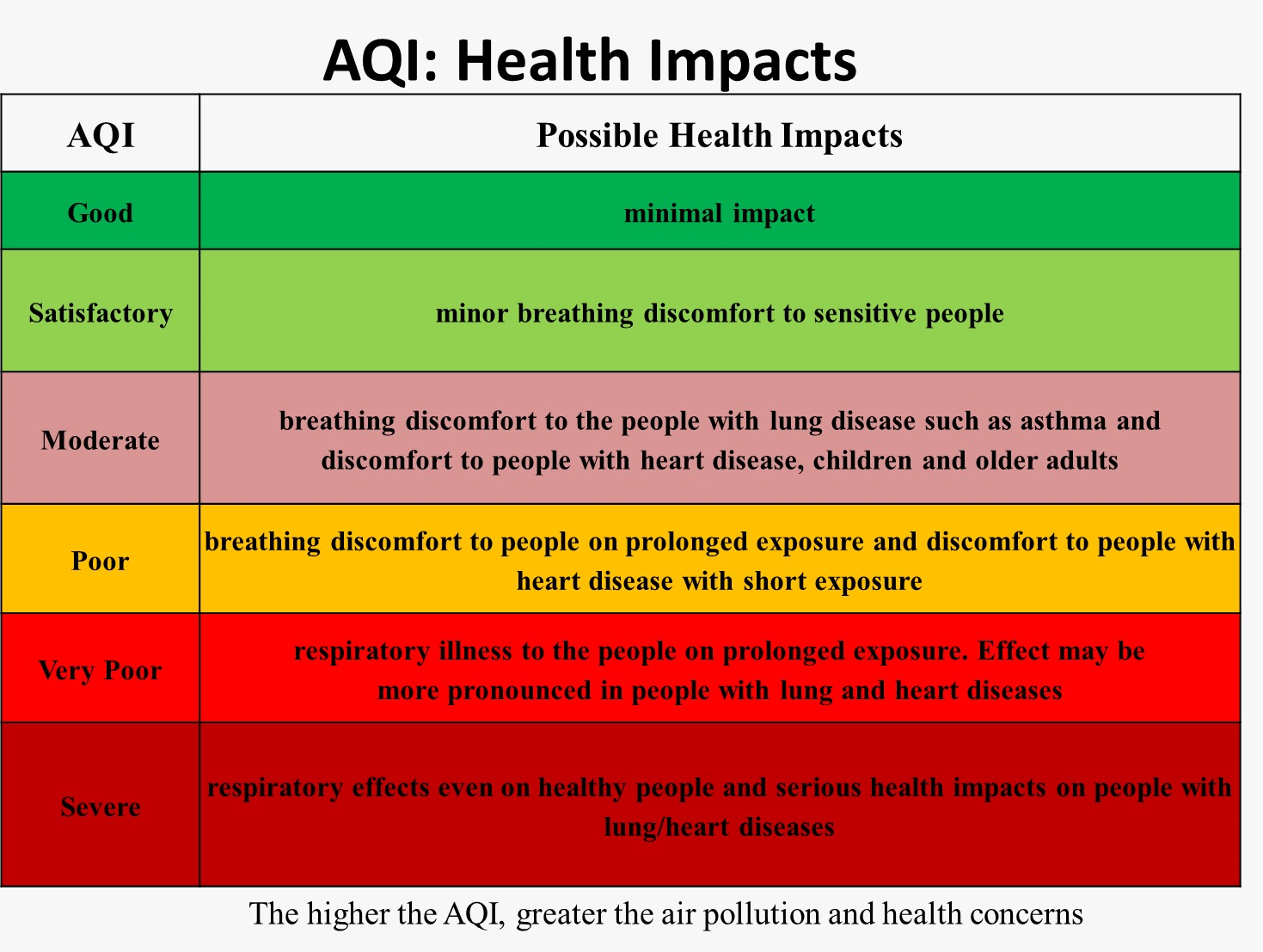


Figure 1 - AQI health impacts

* The overall percentage of missing values is 24.28 and no duplicate values. Observe that among all other features, Xylene has the most missing values and CO has the least missing values.

**4.2 overview of the dataset**

The ‘info()’ function of the Data Frame provides a summary of the dataset, including the total number of rows, the types of attributes present, and the count of non-null values for each attribute. By calling ‘df.info()’ a concise description of the dataset is obtained as below (Fig 4.2).

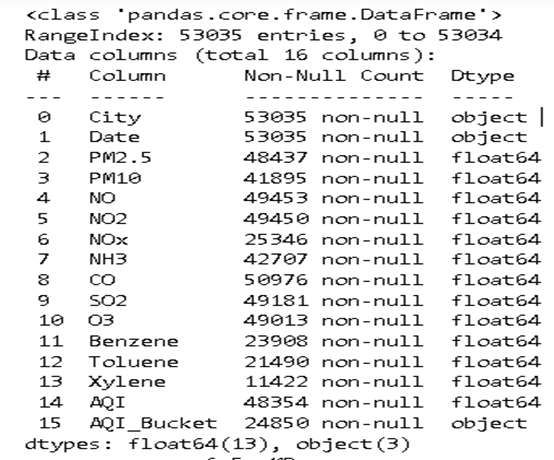
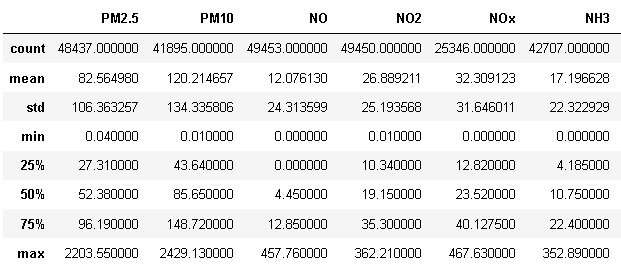


Figure 2 - AQI dataset info

**4.3 Descriptive Statistical Summary**

Descriptive statistical summary provides a concise overview of the key statistical measures for each attribute in the dataset. This summary includes measures such as mean, standard deviation, minimum, maximum, quartiles, and count. It helps to understand the central tendency, dispersion, and distribution of the data. To obtain the descriptive statistical summary, the describe() function in pandas is used.



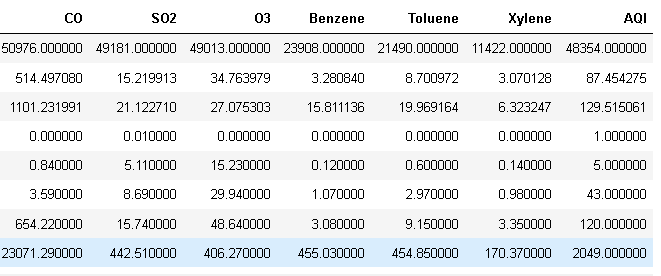


Figure 3 - Statistical values of all the numerical attributes using describe() method

**4.4 Distribution of each attribute**

A histogram was created for each numerical attribute. It allow us to identify the central tendency (mean, median) and the presence of outliers. It’s a skewed histogram suggesting a non-normal distribution. This information is valuable for understanding the data characteristics and potential insights it may provide. By calling the `hist()` method on the entire dataset, a histogram will be plotted for each numerical attribute, as show in Figure 3.

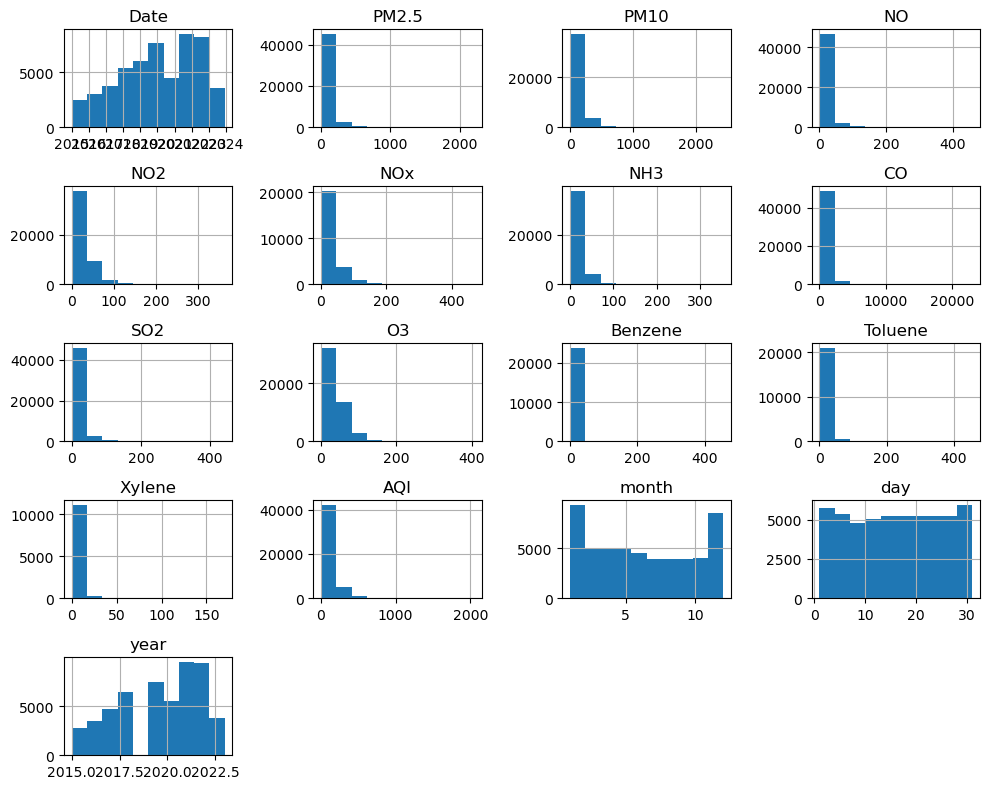


Figure 4 - Histogram for each numerical attribute

**4.5 Correlation matrix**

A correlation-based feature selection method has been exploited in the present work to determine the optimal number of input variables (pollutants) when developing a predictive model. Statistical correlation-based feature selection algorithms compute correlations between every pair of the input variable and the target variable. The variables possessing the strongest correlation with the target variable are then filtered for further study. To compute and visualize a correlation matrix, the corr() function from the pandas library is used along with a heatmap plot from the seaborn library.

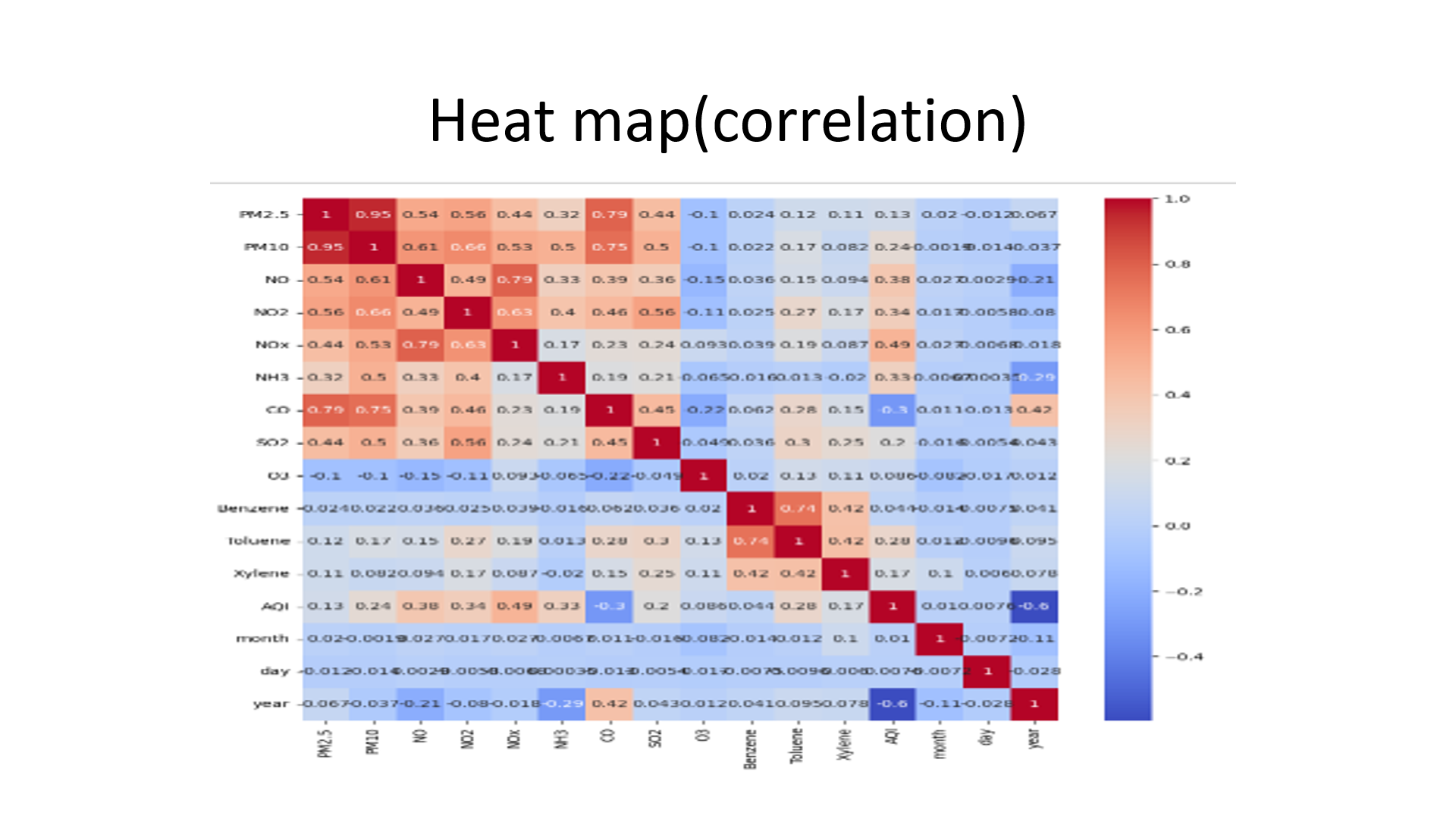


Figure 5 - Correlation between every pair of attributes

The correlation ranges between -1 to 1. When it is close to 1(Vmax), it means that there is strong positive correlation. When the coefficient is close to -1(Vmin), it means that there is strong negative correlation. While coefficients close to 0 mean that there is no linear correlation. The heatmap plotted indicates that there are many features that are strongly correlated and we can eliminate some of them during Data preprocessing. From the figure 4, we can understand that AQI is highly correlated with NOx and less correlated with CO.

**4.6 Outliers**

Fig-1

In addition, a box plot was generated for each attribute, categorized by class. A boxplot is a standardized visualization that presents the distribution of data based on a five-number summary, including the minimum, first quartile (Q1), median, third quartile (Q3), and maximum values. This plot helps identify outliers and provides insights into their values. Refer to Figure-5 for the visual representation of these boxplots. From the figure we can conclude that CO has more outliers.

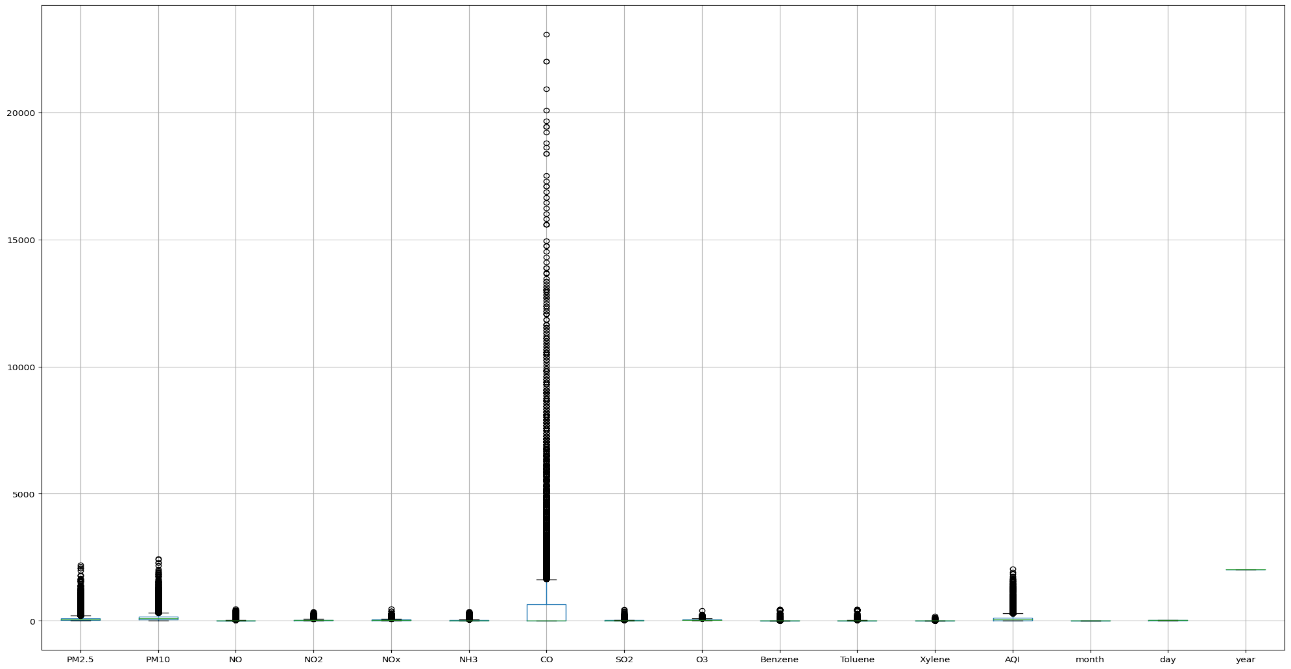


Figure 6 - Box plot for each attribute

### 4.7 Analysis of features

India has become one of the few countries having the most severe air pollution resulting from rapid industrialization and booming urbanization over the last several years. Based on the emissions of PM2.5 and other pollutants, the World Health Organization (WHO) ranked India as the fifth most polluted country. The trends of various pollutants from 2015 to 2023 are observed and shown below.

All the pollutants exhibited a significant fall in the year 2020. The year 2020 witnessed the most strict lockdown in the history of mankind and ceased industrial, automobile, and aviation activities in India and the world served as some ambrosia for the ailing environment and air.

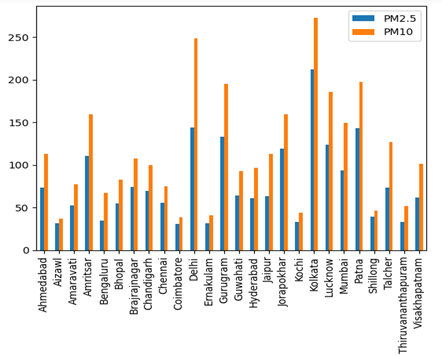
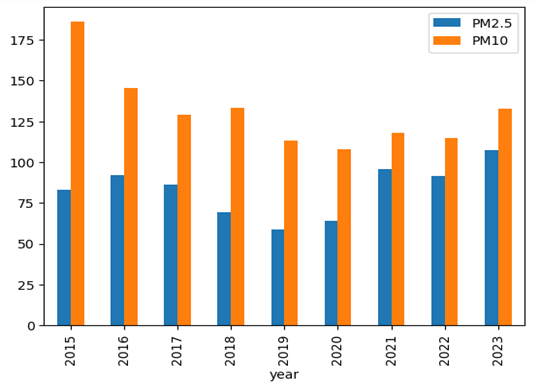
 

Figure 7 - City wise analysis of PM2.5 & PM10 Figure 8 - Year wise analysis of PM2.5&PM10

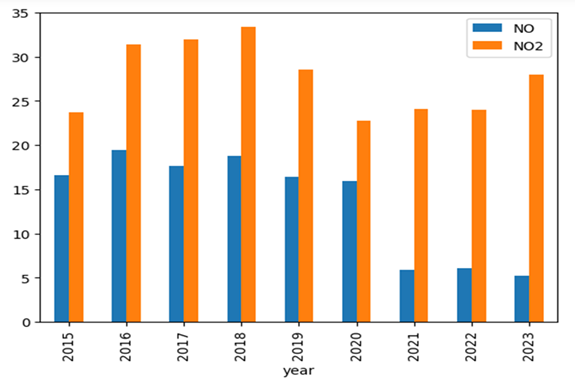
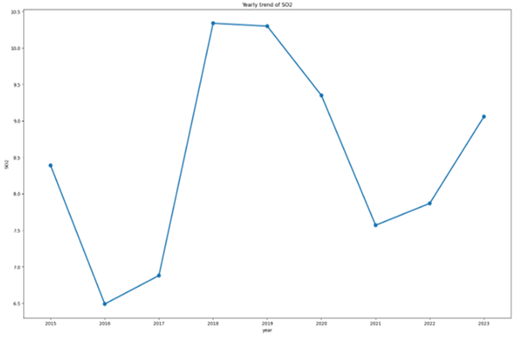
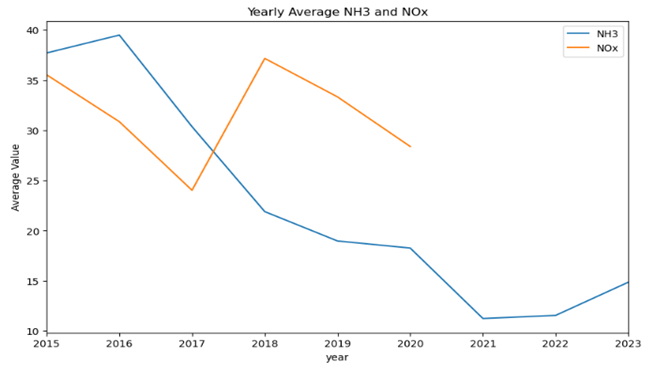
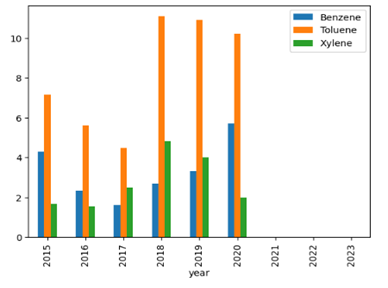
 

Figure 9 - Year wise analysis of NO&NO2 Figure 10 - Yearly trend of SO2

**Figure 11 -** **Yearly trend of NH3 & NOx Figure 12 - Yearly trend of BTX**

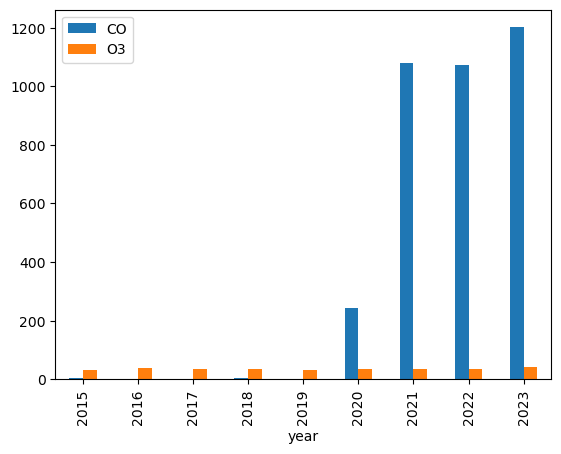
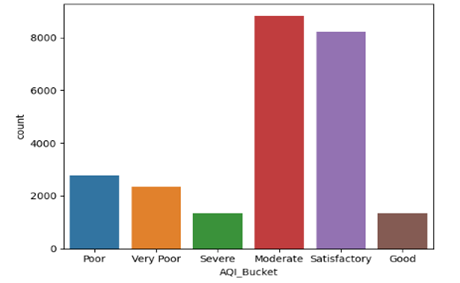
 

Figure 13 - Year wise analysis of CO & O3 Figure 14 - Count of AQI bucket

* From the analysis we can understand Kolkata is having highest range and Aizawl having lowest range for PM2.5, PM10 and NO pollutant. Air pollution levels can vary depending on many factors such as the presence of industries, traffic congestion, population density, and local geography.
* As an analysis of the month, we can understand that the first and last three months pollution is high, and the middle month’s pollution is comparably low. That may be due to the summer and winter climates. Pollutants are lower in the middle of the rainy season. Air pollution levels can vary depending on the season due to changes in weather patterns and human behaviour. In winter, cold air can trap pollutants close to the ground, a process called an inversion." This can cause pollution levels to increase. Additionally, people may leave their cars idling to defrost or warm up, which can increase emissions.
* In summer, air pollution can also be high due to increased emissions from sources such as power plants and vehicles. High temperatures and sunlight can also cause chemical reactions between pollutants, forming ground-level ozone, which is a harmful air pollutant.
* In this, we analyzed the SO2 levels in different cities and observed that Ahmedabad had the highest SO2 concentration, while Ernakulam had the lowest. The comparison of NH3 and NOx concentrations reveals that NOx has higher levels of air pollution compared to NH3. The correlation value between NOx and NH3 indicates a weak relationship between the two pollutants.
* The data analysis reveals that Kolkata has elevated levels of Carbon Monoxide (CO), suggesting potential issues with vehicular emissions or industrial activities that contribute to air pollution. On the other hand, Ernakulam exhibits high Ozone (O3) levels, which could be due to the presence of ozone-forming pollutants and weather conditions that favor ozone formation.
* AQI bucket analysis states that most of the cities have a moderate AQI range, which means breathing discomfort for people with lung disease such as asthma and discomfort for people with heart disease, children, and other adults.
* Figure [15](https://link.springer.com/article/10.1007/s13762-022-04241-5#Fig6) shown below shows the city of Ahmedabad has a severe range of AQI values, which represent the worst air quality, while Aizwal shows a lower AQI range, which represents the purest air to breathe.

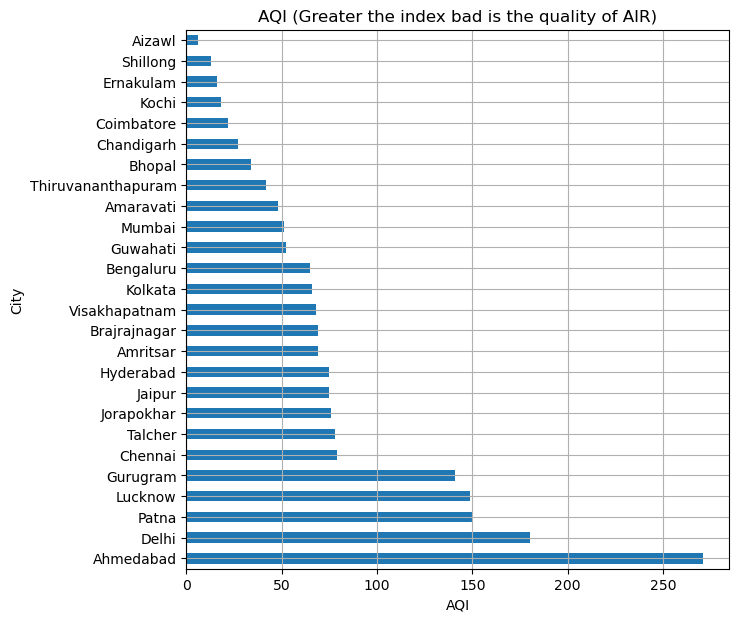
****

Figure 15 - City wise AQI range

**5. DATA PREPROCESSING**

Data pre-processing refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for building and training Machine Learning models. It is an integral step in Machine Learning as the quality of data and the useful information that can be derived from it directly affects the ability of our model to learn. Therefore, it is extremely important to pre-process our data before feeding it into our model.

Data pre-processing mainly includes the steps of Data Cleaning, Data Transformation and Data Reduction. Data Cleaning involves identifying and correcting errors or inconsistencies in the data, such as missing values, outliers, and duplicates. Data Transformation is the step of converting the data into a suitable format for analysis. Data Reduction means reducing the size of the dataset while preserving the important information. The steps of pre-processing performed with the AQI dataset are detailed in this chapter.

**5.1 Dataset edited**

Before starting the preprocessing steps, we edited the 2020-2023 AQI dataset that we used for the combining. Since the NOx feature is not in the 2020-2023 AQI dataset, we calculated the entire NOx attribute using NO and NO2 observations. The code used for the calculation is displayed below.

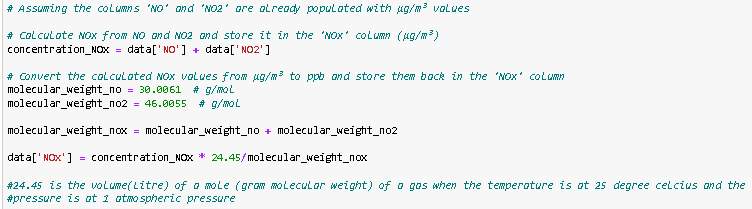


Figure 16 – Code for calculating NOx feature

Since the Air quality Index values are not displayed correctly in the 2020–2023 dataset, we calculated the AQI after the NOx calculation and, using the calculated AQI, filled in the AQI\_Bucket attribute. The code used for AQI calculation is displayed in Figure 17. The edited 2020–2023 AQI dataset is combined with the 2015–2020 dataset for preprocessing.

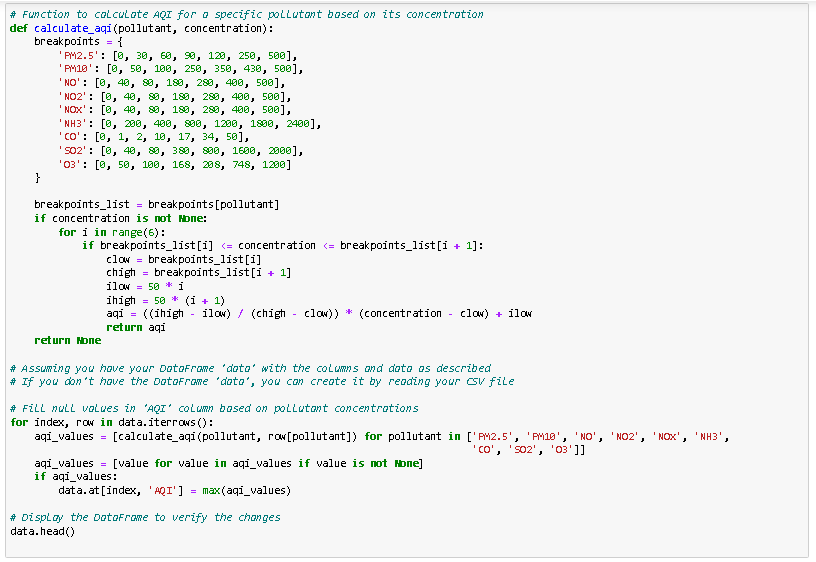
z

Figure 17 – Code for AQI calculation

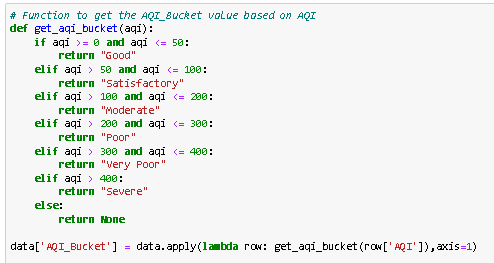


Figure 18 – Code for AQI bucket filling

**5.2 Feature selection**

Feature selection involves selecting a subset of relevant features from the dataset, and this is a technique used for data reduction. It is often performed to remove irrelevant or redundant features from the dataset. We have selected features using correlation analysis. There were 16 features in the dataset, and many of them were strongly correlated. Upon evaluating the heatmap we could understand that the features Benzene, Toluene and Xylene are less correlated; hence, we dropped these features.

**5.3 Handling missing values**

The concept of missing values is important to comprehend in order to efficiently manage data. If the researcher, programmer, or academician does not properly handle the missing figures, he or she may get to the wrong conclusion about the data, which will have a significant impact on the modeling phase. It is a significant problem in data analysis since it has an impact on the outcomes. It is difficult to have total faith in the insights when you know that several items are missing data. It may reduce the statistical power of research and lead to erroneous results owing to skewed estimates. Problems due to missing values are, Statistical power, or the chance that the test would reject the null hypothesis when it is erroneous, is lowered in the absence of evidence, The loss of data might cause parameter estimations to be skewed, It has the ability to reduce the representativeness of the sample, It might also make the analysis of the study more challenging.

Figure 19 shown below presents a view of the missing values in each feature of the dataset. Observe that among all other features, PM10 has the most missing values and CO has the least missing values. A large number of missing values may exist due to a variety of factors, such as a station that can sense data.

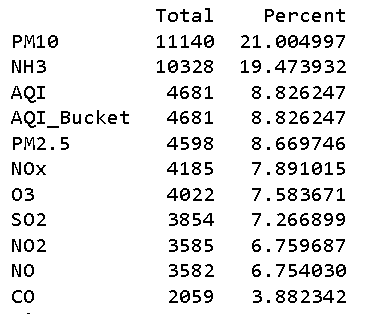


Figure 19 - Missing values of the features and their percentages

Here we treated all the missing values by the method of forward and backward filling by grouping the city. bfill() is used to backfill the missing values in the dataset. i.e., it will fill the missing values with the next data point. The ffill() function is used for forward filling using previous data.

**5.4 Outliers**

An outlier is a piece of data that is an abnormal distance from other points. In other words, it’s data that lies outside the other values in the set. They are the extremely high or extremely low values in the data set. A simple way to find an outlier is to examine the numbers in the data set. We will see that most numbers are clustered around a range and some numbers are way too low or too high compared to the rest of the numbers. Such numbers are known as outliers.

### 5.4.1 Detection & Handling Outliers

The simplest way to detect an outlier is by graphing the features or the data points. Visualisation is one of the best and easiest ways to have an inference about the overall data and the outliers. Scatter plots and box plots are the most preferred visualisation tools to detect outliers. Interquartile range (IQR) technique method can be used to find the maximum and minimum values of data points that are outliers by calculating the boundaries. It is crucial to carefully consider the presence and impact of outliers on the analysis of AQI dataset. Outliers should be detected, evaluated, and handled appropriately for a cleaned dataset.

In this project, we have visualised the outliers using boxplots for each attribute. It is already visualised in Figure 6 of the section **‘Exploratory Data Analysis’**. From the boxplots, it is clear that outliers are present for each attribute.

We treated the outliers using different methods, which are described below.

1. Quantile based flooring and capping: In this technique, the outlier is capped at a certain value above the 90th percentile value or floored at a factor below the 10th percentile value.

The describe function after handling using this method:

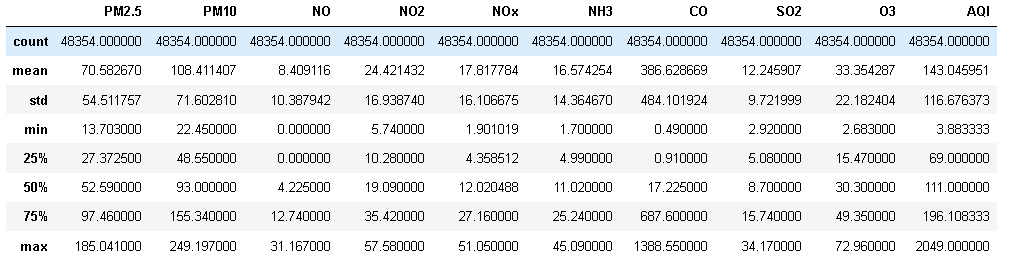
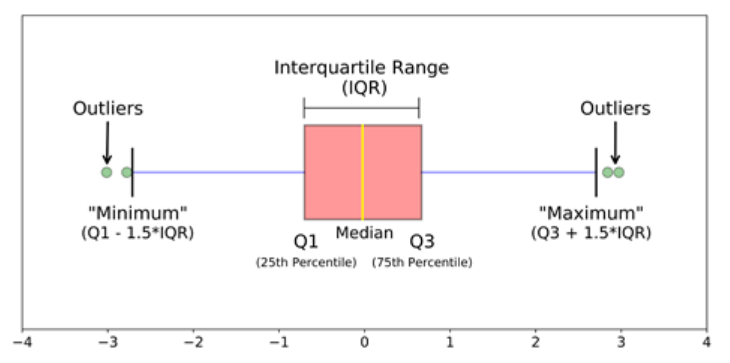
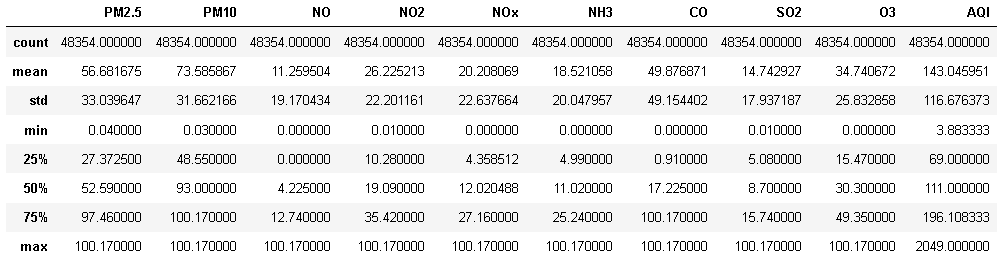


Figure 20 – Describe function after outlier handling using flooring/capping method

1. Flooring & capping with minimum (upper limit) and maximum (lower limit) value by the method of inter quartile range.

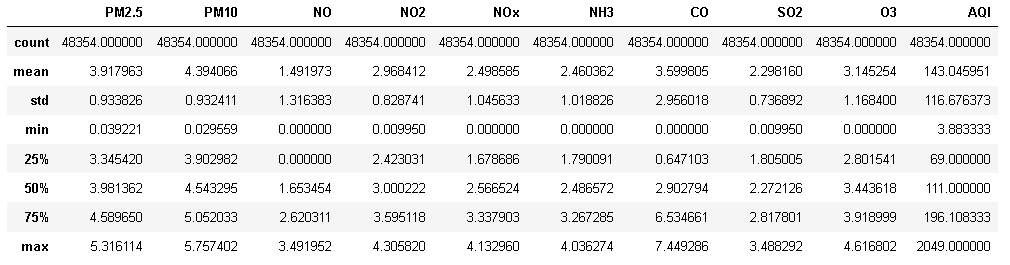


**Figure 21 – Visualization of IQR method**



**Figure 22 – Describe function after outlier handling using IQR method**

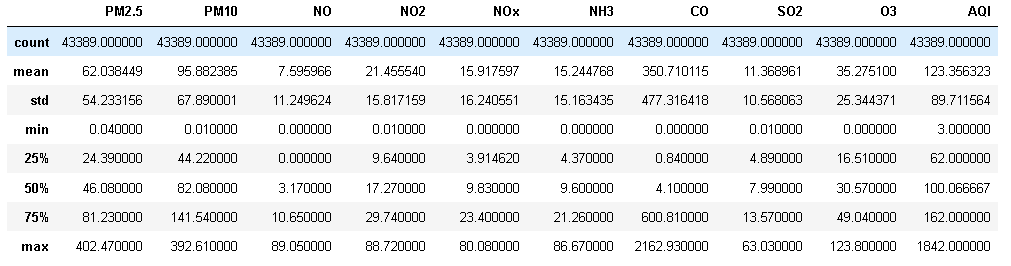
1. Log transformation: Transforming the data using mathematical functions can sometimes reduce the impact of outliers. Common transformations include taking the logarithm of the data. These transformations can help make the data more normally distributed and stabilize the variance.



**Figure 23 - Describe function after outlier handling using log method**

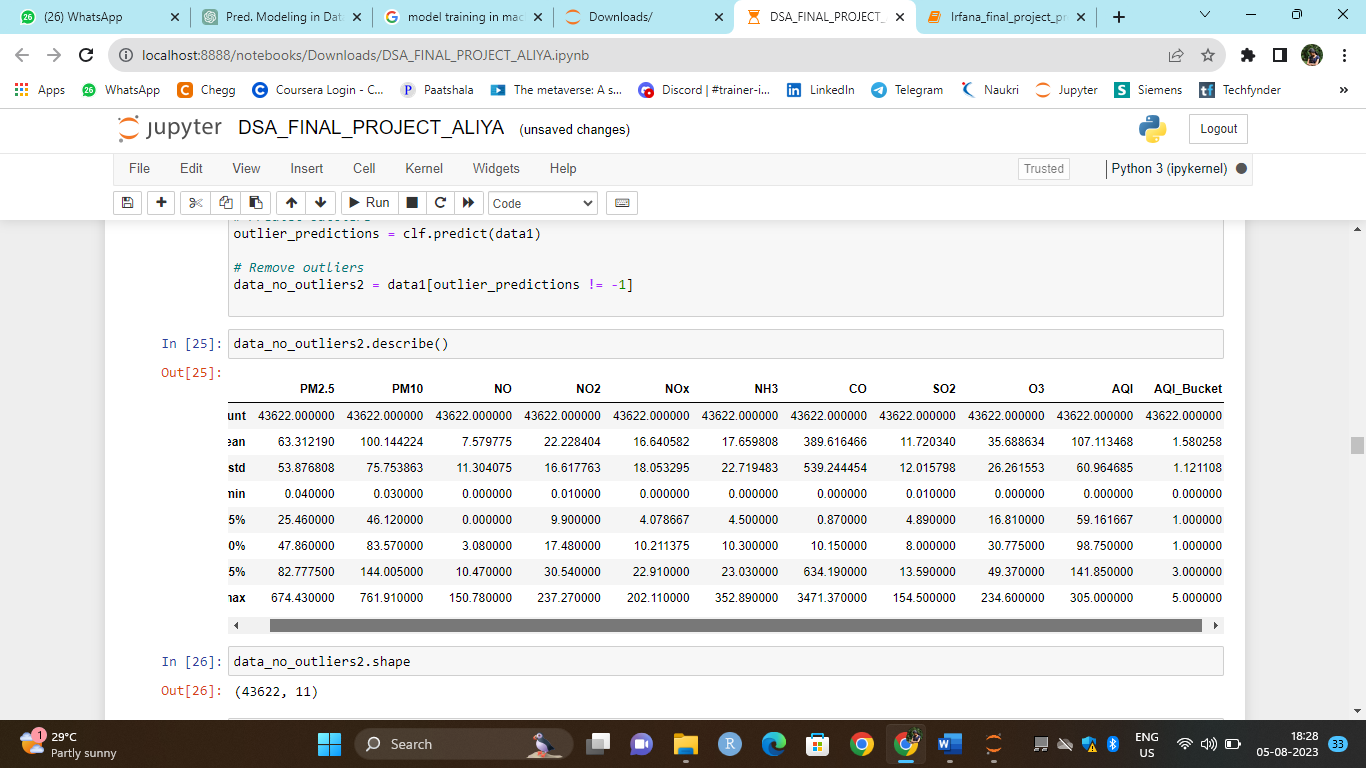
1. Z-score method: [Z- Score](https://www.geeksforgeeks.org/z-score-in-statistics/) is also called a standard score. This value/score helps to understand that how far is the data point from the mean. And after setting up a threshold value one can utilize z score values of data points to define the outliers.

*Z-score = (data point -mean) / std. deviation*



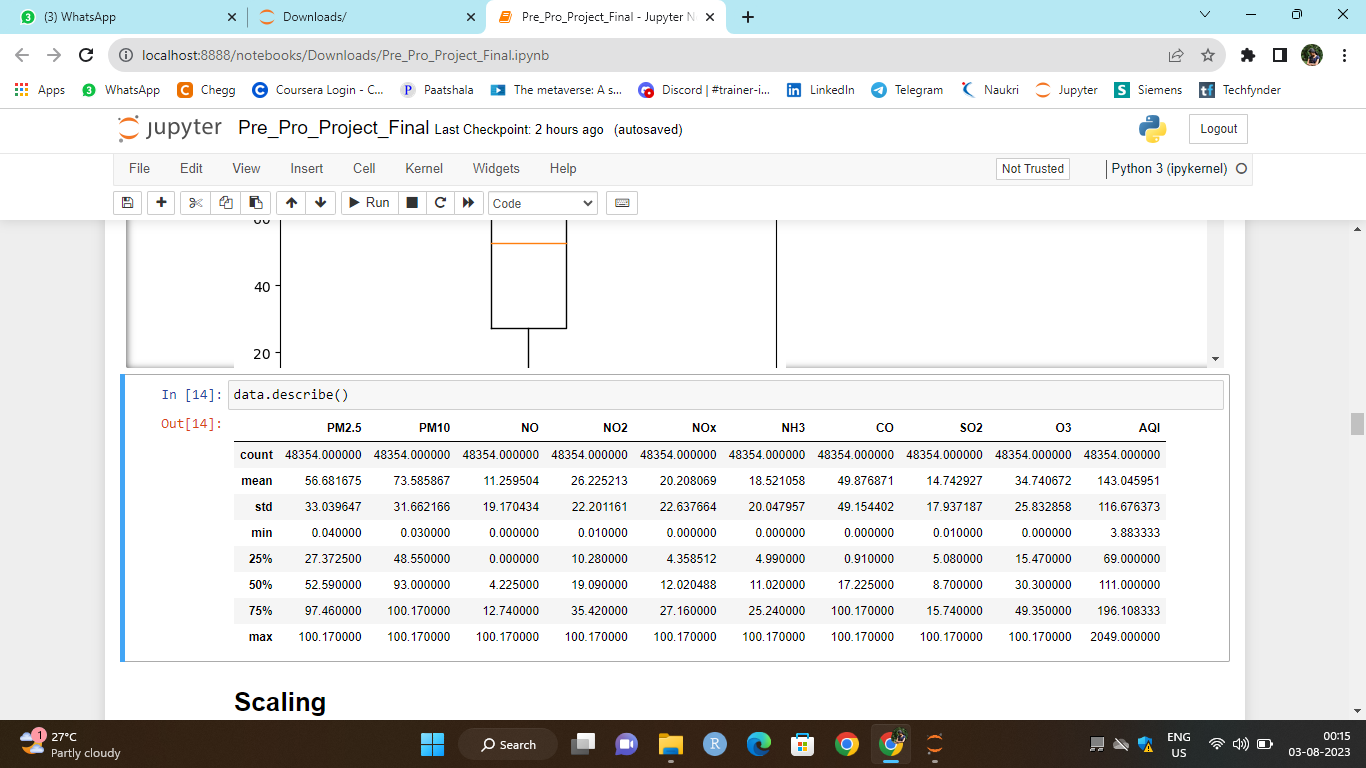
**Figure 24 - Describe function after outlier handling using Z-Score method**

1. Isolation Forest: Isolation Forest is a model-based outlier detection method that attempts to isolate anomalies from the rest of the data using an ensemble of decision trees.



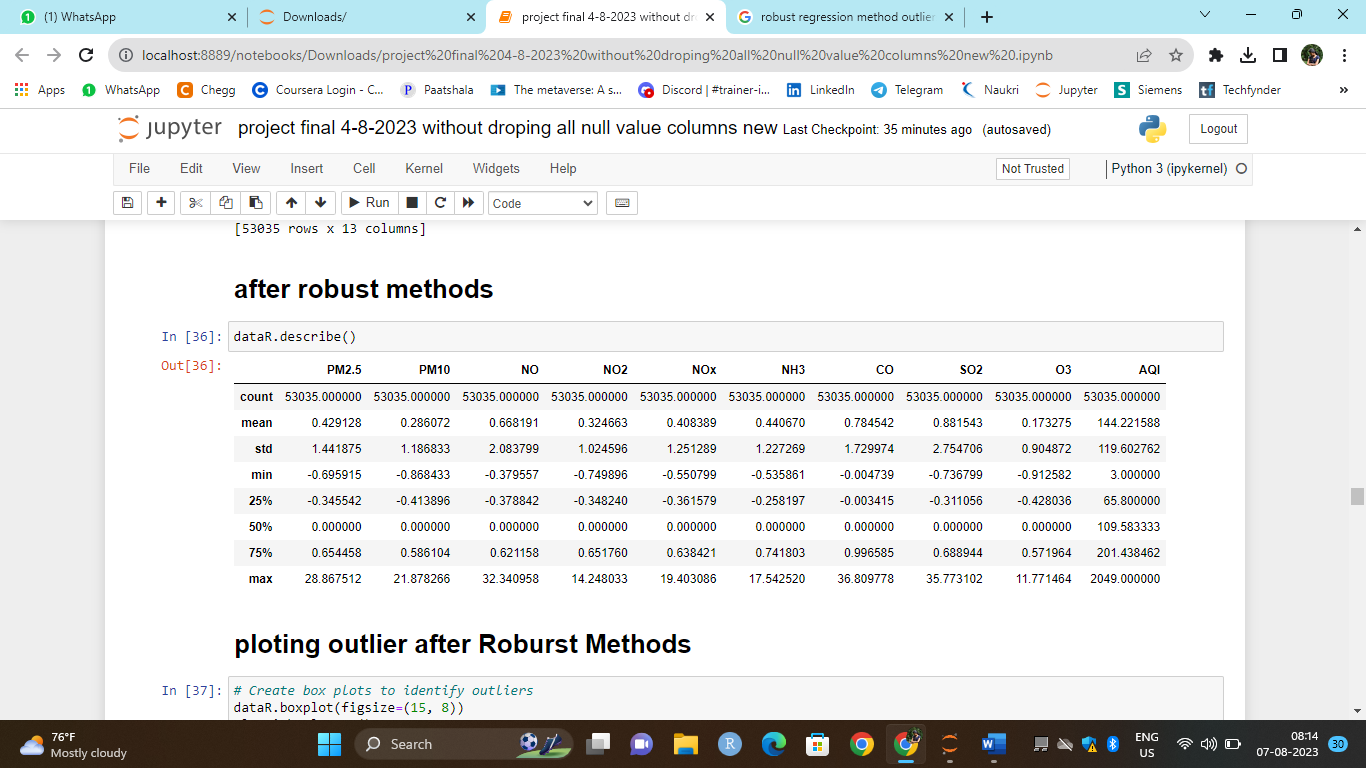
**Figure 25 - Describe function after outlier handling using Isolation Forest**

1. DBSCAN: DBSCAN is effective in discovering arbitrary-shaped clusters in data. DBSCAN considers two main parameters to form a cluster with the nearest data point and based on the high or low-density region, it detects Inliers or outliers.



**Figure 26 - Describe function after outlier handling using DBSCAN**

1. Robust Scaler Method: Robust Scaler is a preprocessing technique used in data normalization or standardization to make numerical features more robust to outliers. The Robust Scaler works by subtracting the median from each data point and then dividing by the interquartile range (IQR).

****

**Figure 27 - Describe function after outlier handling using Robust Scaler Method.**

**5.5 Encoding & Feature Scaling**

Both encoding and scaling are crucial parts of data transformation, aiming to enhance the quality, compatibility, and fairness of the data. Categorical variables need to be encoded into numerical values to effectively process them. This ensures that categorical data can be used as input for algorithms that only accept numerical inputs. The only categorical feature in the dataset is our dependent feature “AQI Bucket”. We are using the technique of Label Encoding to convert it into numerical values.

Feature scaling is the final step of data preprocessing. It involves transforming the independent variables of a dataset to a consistent range. By scaling the features, we ensure that no single variable dominates others, promoting fair comparisons. In this dataset we have applied min max scaling which rescale the features with distribution value between 0 and 1.

After preprocessing, the dataset now consists of 48,354 rows and 13 attributes, which includes the target feature.

**6. DATA MODELLING**

Predictive modeling in data science is a process of using data and statistical algorithms to make predictions about future outcomes or events. It's a subset of machine learning and is commonly used to forecast trends, behaviors, or events based on patterns and relationships discovered in data. Modeling techniques are based around the use of algorithms. We have to choose an appropriate algorithm or model to use for the prediction.

The choice of the model depends on the nature of the data and the problem you're trying to solve. Classification and Regression are two major prediction problems that are usually dealt with in data science.  The key distinction between Classification vs Regression algorithms is Regression algorithms are used to determine continuous values such as price, income, age, etc. and Classification algorithms are used to forecast or classify the distinct values such as True or False, Male or Female, Spam or Not Spam, etc. Common algorithms include linear regression, logistic regression, K-NN, decision trees, random forests and support vector machines.

In our dataset, we have two target variables, AQI and AQI\_Bucket. So we established both regression and classification algorithms. For the regression problem, the feature ‘AQI’ is the target variable, and for the classification, ‘AQI\_Bucket’ is the target variable. As the initial step, the data has been split into training data and testing data, with a test size of 0.2. The subsequent task is to determine the most suitable algorithm, which can be challenging. In order to identify the most appropriate algorithm, the accuracy scores for different algorithms have been evaluated.

We treated outliers using different methods. So the model fitted by each one is different and checked for accuracy.

1. **Flooring & capping with minimum (upper limit) and maximum (lower limit) value by the method of inter quartile range.**

Here mode of evaluation is Accuracy score.

|  |  |
| --- | --- |
| Model | Accuracy Value |
| Linear Regression | 0.596450499734192 |
| Lasso Regression | 0.597180004636446 |
| Logistic Regression | 0.6441939820080653 |
| K-NN | 0.8012615034639644 |
| Decision Tree | 0.7904043015200083 |
| Random Forest | 0.8500672112501293 |
| SVM | 0.792358243 |

Same method of outlier handling with different evaluation methods:

Regression Algorithms

|  |  |  |  |
| --- | --- | --- | --- |
| Model | RMSE training Data | RMSE testing Data | R squared Value |
| Linear Regression | 73.14676959341448 | 75.08443600502682 | Train: 0.6064014953277181  Test: 0.5881995549494626 |
| Decision Tree Regressor | 3.160900524789261 | 49.26542637803148 | Train: 0.9992650043055749  Test: 0.8227152180811381 |
| Random Forest Regressor | 13.541018153939932 | 34.282092329616 | Train: 0.9865114340861656  Test: 0.9141536951391646 |

Classification Algorithms

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy on train | Accuracy on test | Kappa Score |
| Logistic Regression | 0.5778930147853196 | 0.5754841135551796 | 0.43845555419157267 |
| Decision Tree Classifier | 0.9987035836651542 | 0.8439556307576612 | 0.7946288620462526 |
| Random Forest Classifier | 0.9986727166095626 | 0.8842514257065864 | 0.8473142325595101 |
| K-NN | 0.8878908540914282 | 0.8335526728081719 | 0.780419098490543 |
| SVM Model | 0.7838071426366638 | 0.7799085041047816 | 0.7073744953519233 |
| Naïve Bayes | 0.623144118282557 | 0.6273108980384784 | 0.528501694623405 |

1. **Log transformation**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Mean Squared Error | Mean absolute error | Root mean squared error | R-Squared |
| Linear Regression | 5595.834151150718 | 46.62030142194522 | 79.92 | 0.55 |
| Random Forest Regressor | 1354.3909087803306 | 36.80205033391931 | 11.234850025159906 | 0.8920703714052964 |
| SVM | 6387.603245379221 | 79.92248272782335 | 41.02744060939148 | 0.4909803060440371 |

1. **Z-score method.**

Regression Algorithms

|  |  |  |
| --- | --- | --- |
| Model | Mode of Evaluation | Value |
| Logistic Regression | M.S.E. | 2308.3336131539954 |
| Logistic Regression | R-Squared | 0.7157152826887571 |

Classification Algorithms

Here mode of evaluation is Accuracy score.

|  |  |
| --- | --- |
| Model | Accuracy Value |
| Z-Score | 0.9996542982253975 |
| Logistic Regression | 0.5792809403088269 |
| Decision Tree Classifier | 1.0 |

1. **Quantile based flooring and capping:**

Regression Algorithms

|  |  |  |
| --- | --- | --- |
| Model | Mean Squared Error | R2 Score |
| Linear Regression | 4660.460992995359 | 0.6286139985190387 |
| Decision Tree | 2828.8659228253 | 0.7745713985412851 |
| Random Forest Regressor | 1593.5432319295987 | 0.8730126375946752 |
| Support Vector Regressor | 5128.510132238651 | 0.591315778754635 |

Regression models after Min-Max Scaling

Here mode of evaluations are Mean Squared Error and R^2 score.

|  |  |  |
| --- | --- | --- |
| Model | Mean Squared Error | R^2 score |
| Linear Regression | 4660.4609929953585 | 0.6286139985190388 |
| Decision Tree | 2763.617674834474 | 0.7797709455306004 |
| Random Forest Regressor | 1544.4848132934487 | 0.8769220383950773 |
| Support Vector Regressor | 3573.5925204451178 | 0.7152251163382488 |

Regression models after Standard Scaling

|  |  |  |
| --- | --- | --- |
| Model | Mean Squared Error | R2 Score |
| Linear Regression | 4660.460992995362 | 0.6286139985190385 |
| Decision Tree | 2812.436791122264 | 0.7758806144193205 |
| Random Forest Regressor | 8985.231628853837 | 0.2839787196940896 |
| Support Vector Regressor | 3573.5925204451178 | 0.7171508335086666 |

Classification Algorithms

Here mode of evaluation is accuracy score.

|  |  |
| --- | --- |
| Model | Accuracy Value |
| Logistic Regression | 0.5680901664770964 |
| SVM | 0.7816151380415676 |
| Random Forest Classifier | 0.8788129459207942 |
| Decision Tree Classifier | 0.830110640057905 |

After Min Max Scaling

|  |  |
| --- | --- |
| Model | Accuracy Value |
| Logistic Regression | 0.7643470168545135 |
| SVM | 0.8279391996691139 |
| Random Forest Classifier | 0.880674180539758 |
| Decision Tree Classifier | 0.8821218074656189 |

After Standard Scaling

|  |  |
| --- | --- |
| Model | Accuracy Value |
| Logistic Regression | 0.7647606245476166 |
| SVM | 0.8280426015923896 |
| Random Forest Classifier | 0.8791231516906215 |
| Decision Tree Classifier | 0.8798469651535519 |

1. **DBSCAN:**

|  |  |  |
| --- | --- | --- |
| Model | Mode of evaluation | Evaluation Score |
| DBSCAN | Silhouette Coefficient | 0.76 |
| Decision Tree Classifier | Accuracy | 0.98 |
| Decision Tree Regressor | M.S.E. | 334.3 |
| Linear Regression | R-Squared | 0.87 |
| Linear Regression | M.S.E. | 227.6 |

1. **Isolation Forest**

|  |  |  |
| --- | --- | --- |
| Model | Mode of evaluation | Evaluation Score |
| Decision Tree Classifier | Accuracy | 0.99 |
| Decision Tree Regressor | M.S.E. | 633.15 |
| K-NN Regressor | R^2 | 0.98 |
| SVM Regressor | R^2 | 0.89 |
| Random Forest | M.S.E. | 28.20 |
| K-NN Classifier | Accuracy | 0.93 |
| SVM Classifier | Accuracy | 0.85 |

1. **Robust Method:**

Regression Algorithms

|  |  |  |
| --- | --- | --- |
| Model | Mode of Evaluation | Value |
| Logistic Regression | M.S.E. | 5970.115247092257 |
| Logistic Regression | R-Squared | 0.5668908474183252 |

Classification Algorithms

|  |  |
| --- | --- |
| Model | Accuracy Value |
| Random Forest | 0.9999057226359951 |
| Logistic Regression | 0.7926840765532196 |

The table illustrates a comparison of accuracy scores for different models. The selection of the model is based on the evaluation metrics and cross-validation results, aiming to identify the model with the highest accuracy or the desired trade-off among various metrics.

**CONCLUSION**

Air Quality Index is a measure of how polluted the air is in a specific area. It takes into account various pollutants such as carbon monoxide, sulfur dioxide, nitrogen dioxide, particulate matter, and ozone. The higher the AQI value, the greater the level of pollution and the greater the risk to public health and the environment.

Air Quality Index is an important topic that affects us all. Whether you live in a big city or a rural area, the air you breathe can have a significant impact on your health. Poor air quality can cause respiratory problems, heart disease, and even cancer. It also affects the environment by causing acid rain, smog, and climate change. Therefore, it is crucial that we understand what Air Quality Index is and how it affects us.

**REFERENCES**

* [https://www.analyticsvidhya.com](https://www.analyticsvidhya.com/blog/2021/05/introduction-to-data-science-project-lifecycle/)
* [https://towardsdatascience.com](https://towardsdatascience.com/stoend-to-end-data-science-life-cycle-6387523b5afc)
* [https://www.geeksforgeeks.org](https://www.geeksforgeeks.org/data-science-lifecycle/)
* https://www.kaggle.com