Analysis of the air pollution levels in Delhi using Python

Abstract

This report aims to analyze pollution levels in Delhi using Python programming language. The analysis aims to investigate the levels of pollution for different pollutants, including CO, NO, NO2, O3, SO2, PM2.5, PM10 and NH3. The Delhi air analysis process includes cleaning the dataset, plotting histograms, boxplots and correlation analysis, analyzing the distributions and calculating key statistical values for each pollutant, building the QQ plots and finding the confidence intervals.

Delhi data analysis using Python

Dataset

The Delhi dataset consists of 18,776 records on air pollution levels, including particulate matter (PM2.5 and PM10) levels, nitrogen dioxide (NO2), sulfur dioxide (SO2), carbon dioxide (CO2), ozone (O3), and Ammonia (NH3). The data was collected hourly from monitoring stations located in various areas of Delhi between January 1, 2020, and June 30, 2023. The information is stored in a CSV format, featuring a timestamp for each record.

Missing Values

In the Delhi dataset, there is no missing values. Occasionally the dataset has 0 values for some of the elements at random intervals, sometimes spanning for a week or more, which could be considered as empty values. When the elements are dropped, the missing data can impact the analysis if the amount of missing data is big. However, in this case the missing data shouldn't impact the dataset substantially. Each pollutant column has the same number of elements

Analysis

The analysis started with the general inspection of the dataset. The dataset consists of 18776 rows × 9 columns, which indicate date stamps and the values for pollutant at these date stamps.

The basic metrics such as mean, standard deviation and quartile ranges were computed in a table using function df.describe(). This table shows brief overview of the general data trend for each pollutant and how its distributed.

	co	no	no2	o3	so2	pm2_5	pm10	nh3
Mean	2929.228628	33.660702	66.221299	60.346239	66.693633	238.130309	300.092966	25.109815
Median	1842.5	5.25	54.15	27.18	52.93	157.445	209.705	17.48
Mode	1895.9	0.0	45.93	0.0	62.94	98.81	117.93 / 176.21	8.61
Standard Deviation	2854.523506	62.127118	48.527492	80.464932	49.439191	226.533625	267.165827	26.402108
1st Quartile (25%)	1068.12	0.68	33.93	0.34	34.81	84.44	118.7975	9.63
2nd Quartile (50%)	1842.5	5.25	54.15	27.18	52.93	157.445	209.705	17.48
3rd Quartile (75%)	3685.0	35.76	83.63	92.98	82.02	313.0	387.965	30.4
Interquantile Range (IQR)	2616.88	35.08	49.7	92.64	47.21	228.56	269.1675	20.77

Figure 1: The table shows the general information about the distribution of each pollutant.

The carbon monoxide (CO) is the pollution with the highest distribution of concentration, which was followed by particulate matter of 10 millimetres (PM10), and 2.5 millimetres (PM2.5).

The histograms for each pollutant were plotted with pollution measured in µg/m³ on x axis and counts on y axis. Matplotlib library was used to generate histograms. It creates a subplot grid for each air quality metric, assigning different colour to each. Within a loop, histograms are generated for each metric, with 20 bins for each histogram to visualize the general distribution and avoid showing random noise. Excess subplots are removed, and layout adjustments are made to prevent overlap.

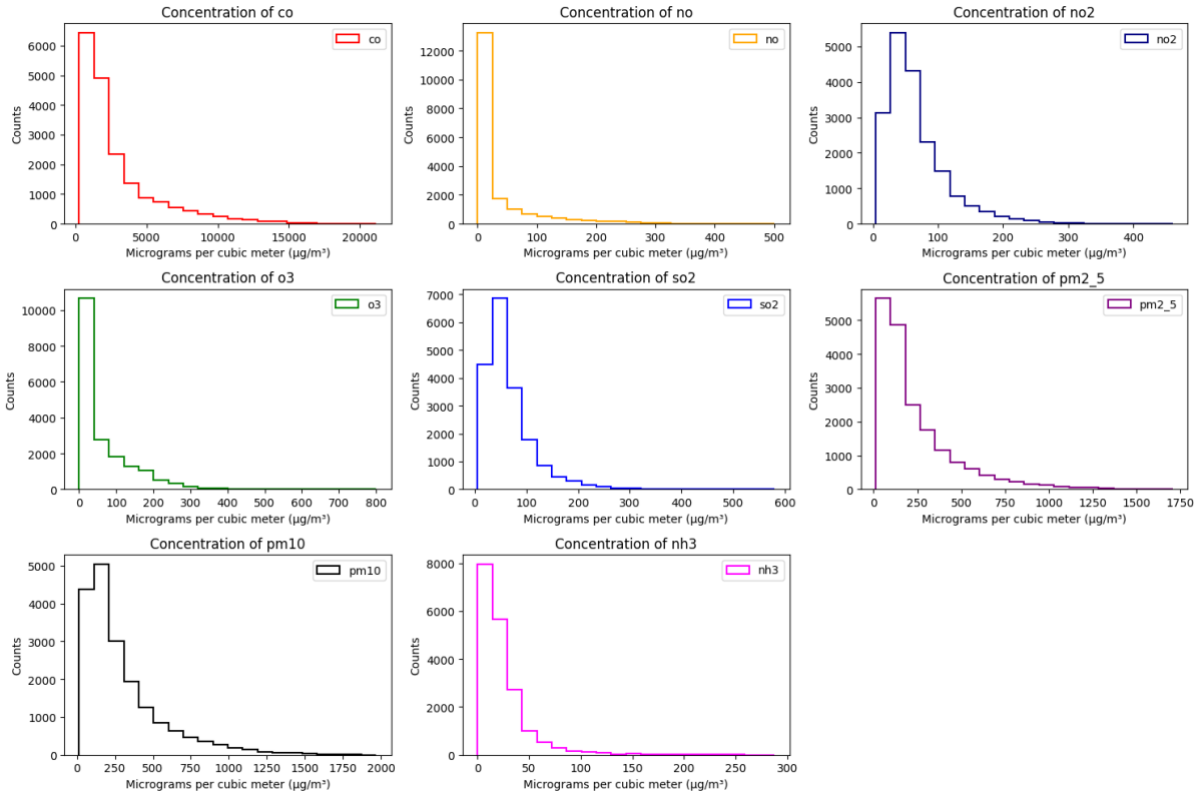

Figure 2: Histograms for each pollutant in the New Delhi, with counts on x axis and pollution in µg/m³.

The histograms show that the pollution for each distribution has the positive skew, where the tail is explicitly pronounced more on the right side than the left. It means that moderate pollution levels are common and most polluation values are located to the left of the mean. However, there are occasional peaks in concentrations where the level of pollution is high, which is represented by a long tail.

For the pollutants with distributions in similar range, the graphs were plotted to compare their distributions.

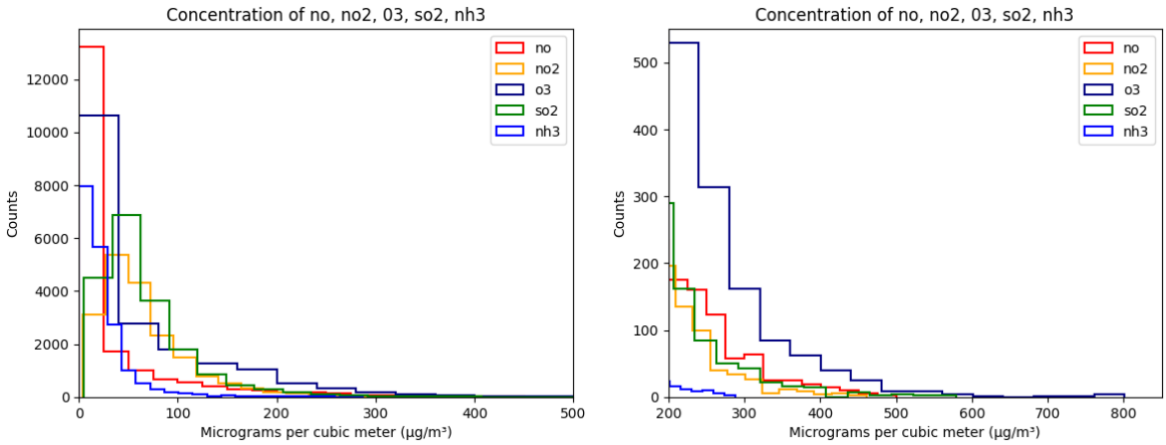

Figure 3: Histogram showing the pollution distribution of NO, NO2, O3, SO2 and NH3 showing the general trend on the left and outliers on the right.

The boxplots were created for all the pollutants in the dataset. It can be seen that CO has a much higher value range than the other pollutants. Only information about the CO can be properly read from this figure due to inappropriate axis scaling. Each pollutant from this dataset contains big number of outliers due to dataset size and a skewed distribution. Outliers were removed from the graph for the visualization purposes.

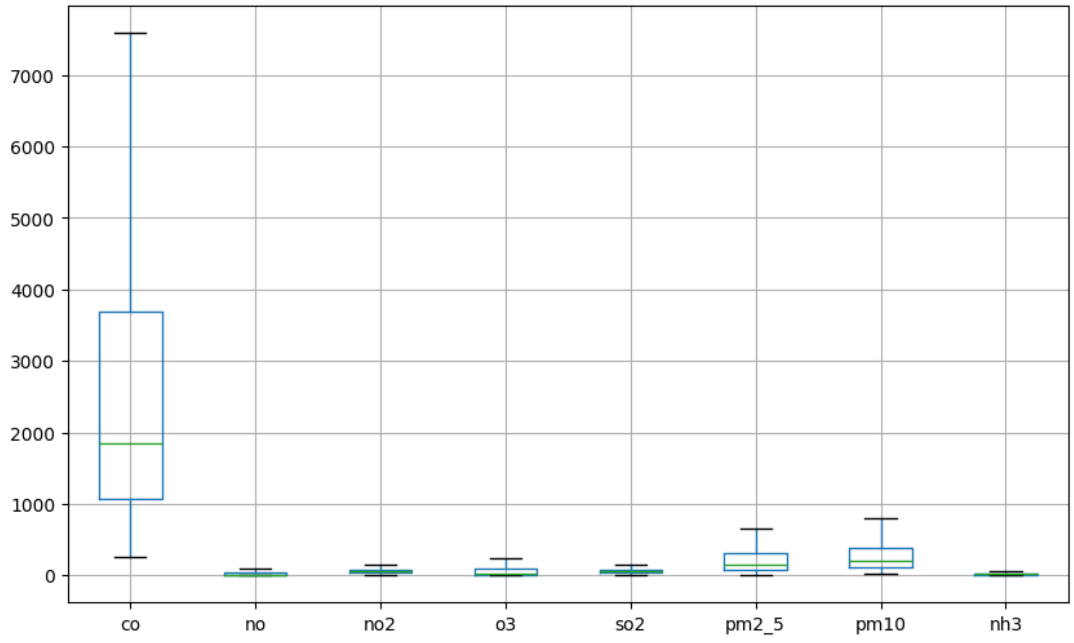

Figure 4: Boxplot graphs for all of the pollutants. The axis range is inappropriate, so another boxplot has to be created.

The boxplots were created for the NO, NO2, O3, SO2 and NH3, as the values for CO, PM2.5 and PM10 are far out of the range for the visual representation.

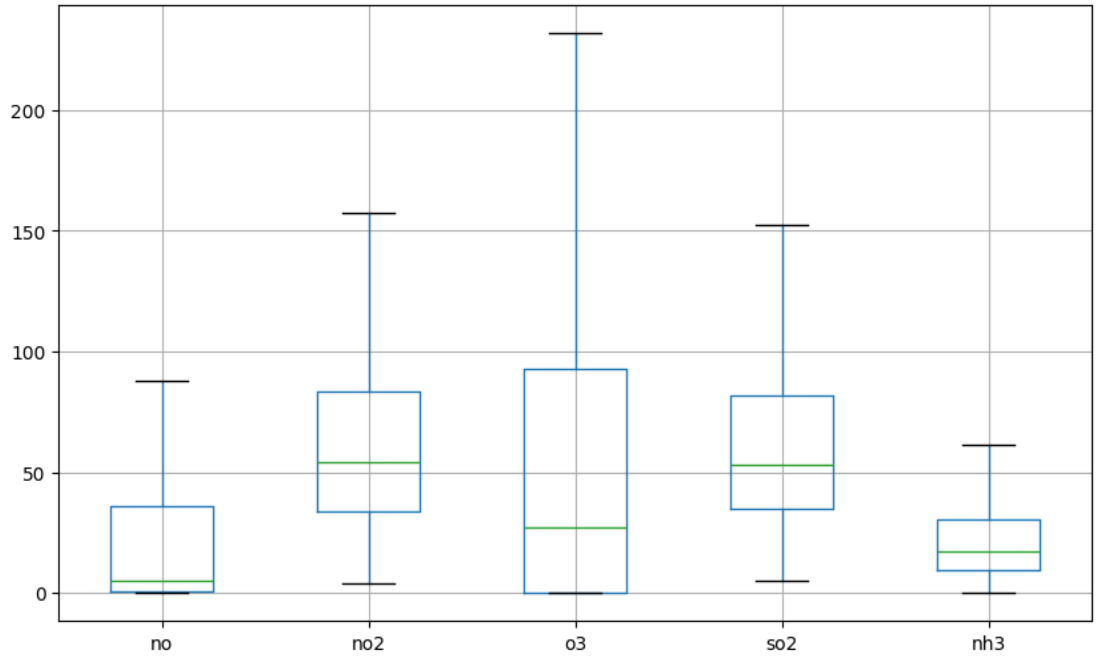

Figure 5: The Image shows the boxplots for NO, NO2, O3, SO2 and NH3.

Boxplots are useful for statistical analysis because they give an overview of the value distribution for each pollutant. They include the measurements of median and interquartile range (IQR), which represents variability. They enable easy comparison of distributional between the pollutants.

The green line shows the mean value of the distribution, blue box represents the 2nd and 3rd quartiles from bottom to the top accordingly. The bottom and upper horizontal lines represent the minimum and maximum values, which are calculated by $Q1 - 1.5IQR$ and $Q3 + 1.5IQR$ respectively.

The general pattern of the prolonged region between the end of the box and the maximum value can be seen, whereas the region between the 2nd quartile and the minimum value is much shorter, or non-existent as in cases of O3 and NO.

The boxplot can also be compared to the pollutant histograms, where most of the values are located in low pollution/high count regions. The shorter region below the median represents high concentration of measurements in the first two quartiles, and the region above represent the long tails in the histograms, where the concentration of measured values is much lower.

The Pearson correlation coefficients were calculated to show how the pollutant values correlate to each other. The information was visualized in the form of heatmap, using the correlation values for each of the pollutants to plot it.

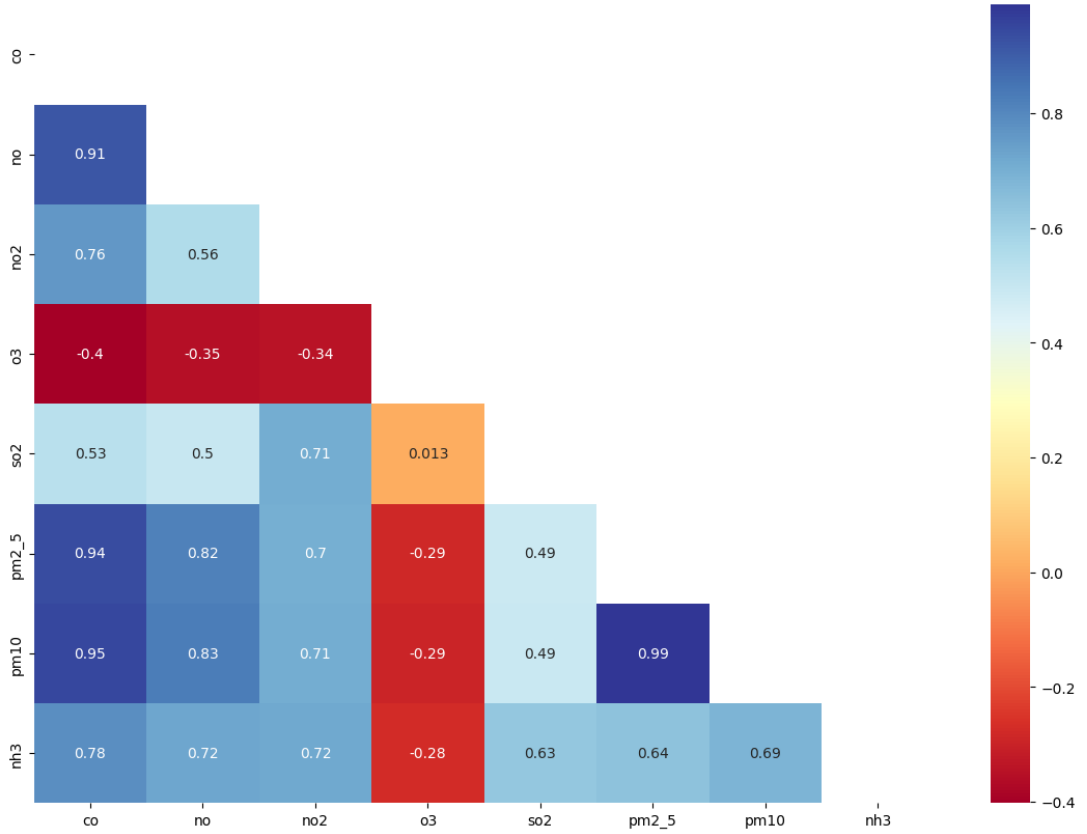

Figure 6: The correlation heatmap of pollutant values.

This heatmap includes only the lower triangle of the correlation values to exclude the unnecessary information, as the correlation matrix is symmetric. Each box has the specific correlation value and a specific colour assigned to it, which represent the correlation between two variables. The correlation value is ranging from -1 to +1, where -1 represents strong negative correlation and +1 indicates strong positive correlation.

Most of the pollutants have the positive correlation, with majority of them being above 0.5. This fact stems from several reasons, described below.

- **Common sources**: Pollutants originating from similar are likely to exhibit positive correlations. For example, pollutants emitted from vehicular exhaust, such as nitrogen oxides (NOx) and particulate matter (PM), tend to be positively correlated due to their common source.
- **Meteorological Conditions**: Certain weather conditions, such as temperature inversions or stagnant air masses, can trap pollutants near the ground. This can result in positive correlations among pollutants during periods of poor air quality.
- **Chemical Reactions**: Some pollutants can undergo chemical reactions in the atmosphere, leading to the formation of secondary pollutants. For instance, nitrogen oxides (NOx) can react with volatile organic compounds (VOCs) in the presence of sunlight to produce ozone (O3), contributing to positive correlations between NOx and O3.
- **Seasonal Variations**: Seasonal factors, such as changes in temperature, humidity, and atmospheric stability, can influence pollutant concentrations.

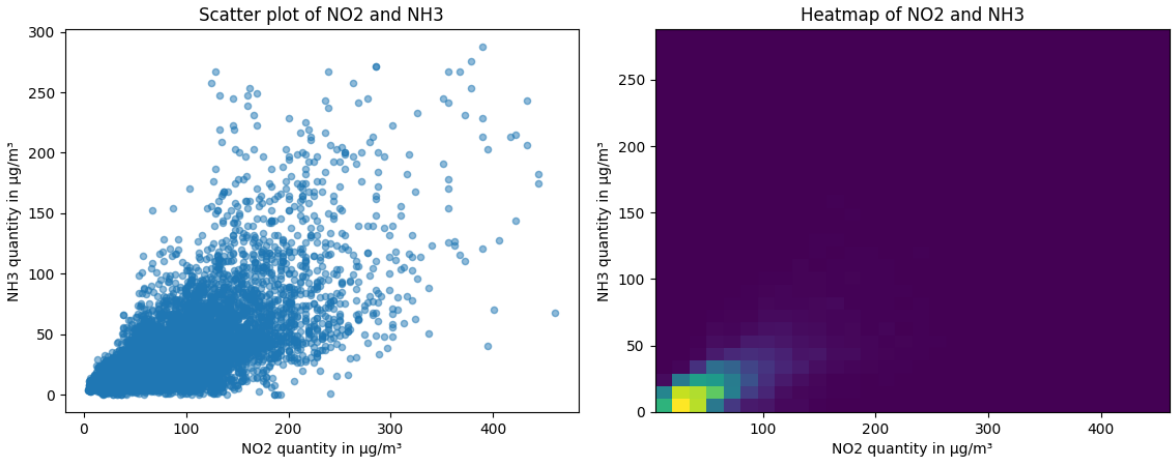

Figure 7: Scatter plot and heatmap showing the correlation between NO2 and NH3. Their positive correlation is linked due to the shared chemical processes in the atmosphere.

However, the only pollutant that has negative correlation value with all other pollutants except SO2, is ozone (O3). This happens due to the chemical reactions that ozone undergoes with other main pollutants in the atmosphere. Ozone formation primarily relies on photochemical reactions involving NOx and VOCs in sunlight; however, ozone is consumed in reactions with NOx, leading to its depletion.

Ozone concentrations can also be influenced by volatile organic compounds VOCs, which contribute to ozone formation but subsequently react with ozone, depleting its levels. Additionally, CO indirectly affects ozone concentrations by altering NOx and VOC levels and directly reacting with ozone, reducing its quantity in the atmosphere.

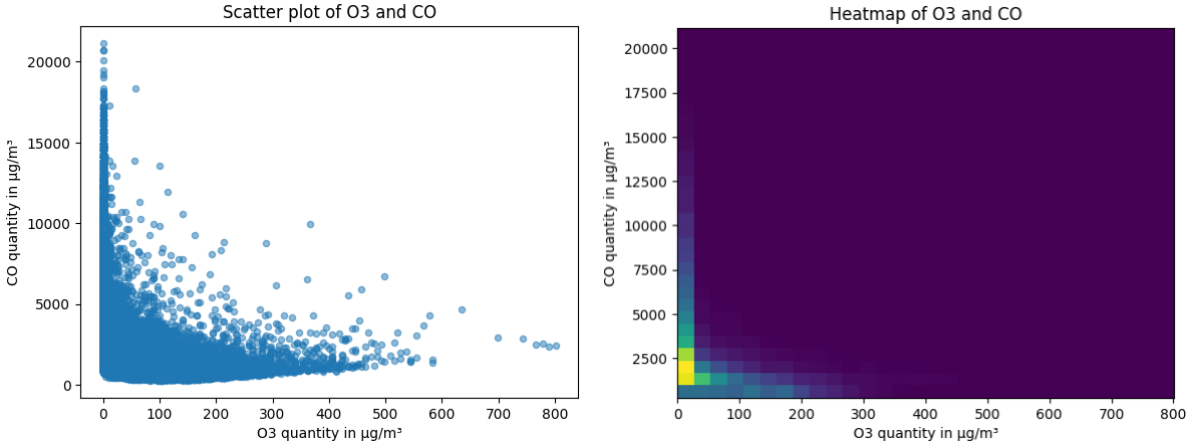

Figure 8: Scatter plot and heatmap showing the correlation between O3 and CO. From the heatmap it can be observed that the concentration is scattered along the x and y axis, meaning absence of one of the pollutants in the atmosphere when other is present.

Both heatmaps and scatterplots were chosen to visualize the correlation of individual correlation patterns, as they complement each other. The scatter plots are good for visualizing the general distribution of the correlation and the heatmaps are good for showing the regions with high concentration of measurements.

Bootstrap

Bootstrap is a resampling technique used in statistics for estimating the sampling distribution of a statistic or for assessing the uncertainty associated with a sample estimate. It involves repeatedly sampling observations from the original dataset with replacement to create samples. These samples are then used to compute estimates of parameters, construct confidence intervals, or perform hypothesis testing.

Sampling distributions of median and standard deviation were made using bootstrap for particulate matter of less than 10 micrometers in diameter (PM10). For both metrics Student's t-distribution and QQ plots were plotted. The confidence intervals were then constructed for both metrics.

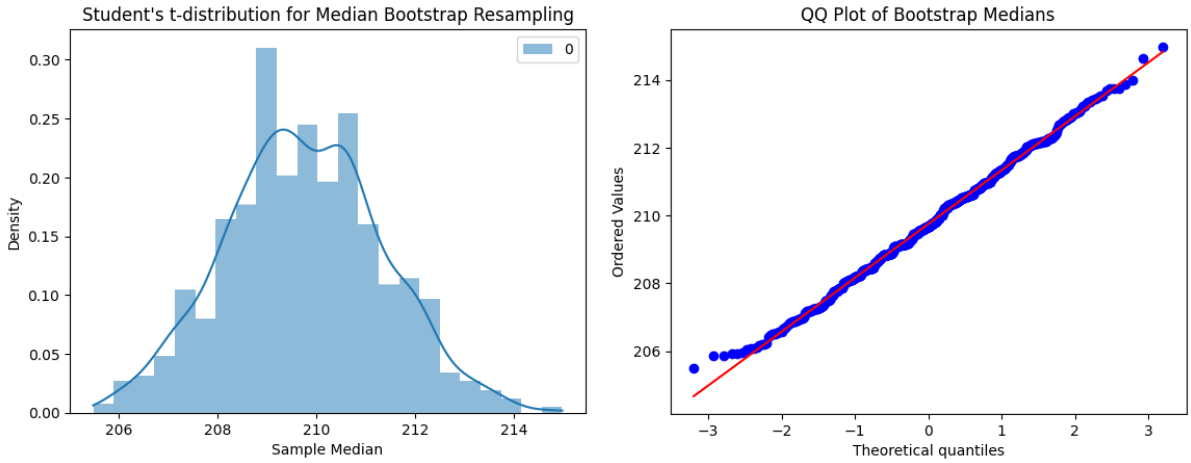

Figure 9: Normalised histogram of Median Bootstraps overlayed by Student's t-distribution on the left and QQ plot of Bootstrap medians on the right.

A QQ plot is a type of scatterplot generated by plotting the quantiles of one dataset against the quantiles of another dataset. When both datasets come from the same distribution, the points in the plot are expected to align approximately along a straight line.

By looking at the QQ plot on the right, it can be seen that the blue dots fit the red line, which indicates that the median values follow the normal distribution. Slight deviation at the end of the diagonal line indicate the outlier values which are often hard to predict by distribution.

On the normalised histogram, several peaks can be observed. Although the data points match well with the expected distribution in terms of quantiles, there may be underlying subpopulations within the data that result in several peaks. One of the potential reasons for this may be multiple pollution sources, seasonal effects and detector sensitivity, which may have an activation value threshold at certain values.

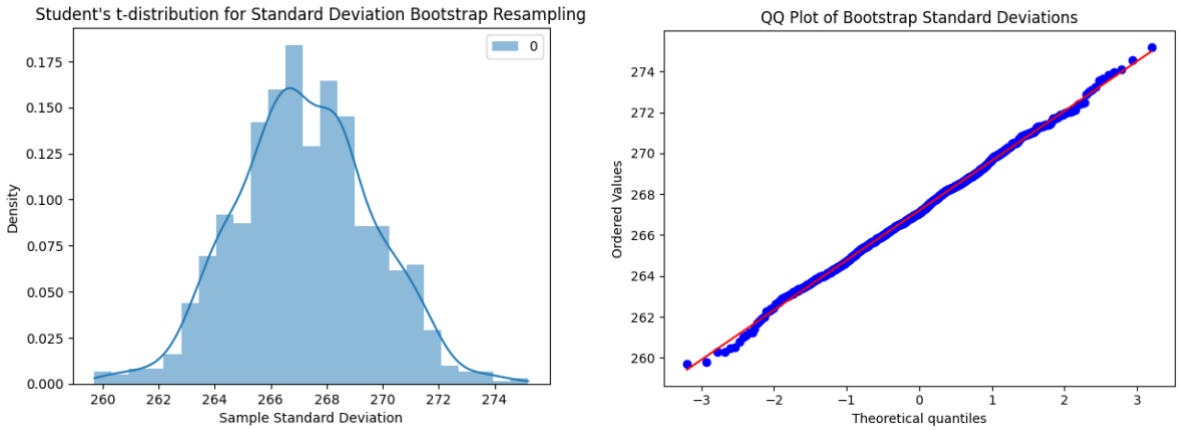

Figure 10: Normalised histogram of Median Bootstraps overlayed by Student's t-distribution on the left and QQ plot of Bootstrap medians on the right.

For the standard deviation, the blue dots also fit the red line, which indicates that the values follow normal distribution, with the outlier values slightly deviating from the norm at the tails. The histogram also has several peaks, which indicates variability of the data inside the distribution.

Conclusion

The examination of air pollution levels in Delhi focused on finding causes and patterns in pollution data. Initial data pre-processing ensured data integrity, followed by statistical computations and visualization techniques.

In New Delhi, the dataset spanned from January 2020 to June 2023. In Delhi, analysis of data from January 2020 to June 2023 revealed periodic zero entries but no missing values. Statistical parameters quantified pollutant distribution characteristics, with carbon monoxide (CO) prevailing prominently. Visualization methods like histograms and boxplots were used to assess the pollutant distribution patterns, while correlation heatmaps showed the correlation patterns. Positive correlations among nitrogen oxides (NOx) were linked to common emission sources, whereas negative correlations involving ozone (O3) were attributed to atmospheric chemical reactions. Bootstrap resampling affirmed normal distributions for particulate matter (PM), bolstering statistical validity.

Data set link:

Delhi: https://www.kaggle.com/datasets/deepaksirohiwal/delhi-air-quality