Atal Bihari Vajpayee-Indian Institute of Information Technology and Management Gwalior

Data Analytics Project

Analysis of India’s Air Quality (2015-2020)

Submitted to:
Dr. Santosh Singh Rathore

Submitted by:
Narra Abhigna 2021BCS-048
Ravi Jwalana 2021BCS-056

Table of Contents

Abstract: The amount of contaminants in ambient air has been rising between 2015 and 2020, which has resulted in an alarming decline in India's air quality. Based on data made publicly available by the Central Pollution Control Board, the official portal of the Indian government, the current study looked at variations in the concentration of air pollutants and patterns of different air pollutants, such as sulphur dioxide (SO2), nitrogen dioxide (NO2), suspended particulate matter (PM), ozone (O3), carbon monoxide (CO), and benzene, have been studied at daily levels of numerous stations throughout multiple cities in India using descriptive analysis.We also examined the impact of the COVID-19 lockdown on India's pollution levels.

1. Introduction

The invisible element known as air surrounds the Earth, giving us all access to breathable oxygen and playing a critical part in maintaining life as we know it. However, as time goes on, pure and fresh air becomes progressively tainted as a result of rising air pollution. When one or more substances are present in the air at concentrations higher than they naturally occur and have the potential to cause harm, this is known as air pollution.

Rapid urbanisation and industrialization have had a negative impact on the environment recently. The severity of the air pollution problem is rising. Elevated air pollution levels are leading to severe health problems. Millions of people are directly impacted, including those who have severe asthma attacks, pneumonia, eye irritation, shortness of breath, and other chronic respiratory conditions . According to a 2018 study by the Health Effects Institute on air pollution in India, air pollution caused 1.1 million fatalities in that country in 2015.

1.1 Causes of Air Pollution

The following discusses a few of the main causes of air pollution.

- Industrial emissions

The main sources of air pollution are thermal power plants and other industrial areas, which release toxic chemicals like SO2 and NOx into the atmosphere .

- Automobile emissions

Vehicle emissions and traffic congestion are the main causes of the declining air quality.

- Burning agricultural stubble in Haryana and Punjab. In order to expeditiously prepare their fields for the Rabi crop of wheat, farmers in Punjab and Haryana burn the stubble of their rice crop.

- Construction and destruction Ongoing construction and demolition are harmful because they raise the amount of dust-borne particulate matter in the air.

- Added elements

Overcrowding, road dust, Diwali cracker smoke, and other issues are some that could contribute indirectly to declining air quality.

1.2 Type of Pollutants in ambient air

Particulate matter (PM2.5 and PM10): Particulate matter is a mixture of liquids and solids that are suspended in the air. Examples of these substances include carbon, water, sulphates, nitrates, complex organic compounds, and mineral dust. PM's size varies. Certain particles, such smoke, soot, dust, or dirt, are big enough or dark enough to be visible to the unaided eye. However, PM2.5 and PM10, two of the smallest particles, are the most harmful.PM10 is particulate matter 10 micrometres or less in diameter, PM2.5 is particulate matter 2.5 micrometres or less in diameter. PM2.5 is generally described as fine particles.Particles are defined by their diameter for air quality regulatory purposes. Those with a diameter of 10 microns or less (PM10) are inhalable into the lungs and can induce adverse health effects.

Oxides of nitrogen (NO, NO2, NOx): Often referred to as NOx gases, nitrogen oxides are a class of seven gases and compounds made up of nitrogen and oxygen. Nitric oxide (NO) and nitrogen dioxide (NO2) are the two most prevalent and dangerous forms of nitrogen oxides.

Sulphur Dioxide (SO2): SO2 is a colourless gas with a potent smell.

Ammonia (NH3): Ammonia pollution is the result of industrial and agricultural processes that produce ammonia (NH3), a chemical compound consisting of nitrogen and hydrogen.

Oxygen (O3): Colourless and extremely unpleasant, ground-level ozone develops just above the surface of the earth. Because it is created when two principal pollutants combine in the presence of sunshine and stagnant air, it is referred to as a "secondary" pollutant. Nitrogen oxides (NOx) and volatile organic compounds (VOCs) are these two main pollutants.

1.3 Air Quality Index

The government uses the Air Quality Index (AQI) to inform the public about the state of the air. Pollutant concentration rises lead to a decline in air quality. For the average person, the level of pollution is represented by the Air Quality Index.

The range of the Indian AQI, according to the Indian Government (CPCB), is 0-500, with 0 denoting good and 500 denoting severe. Particulate matter (PM 10 and PM 2.5), carbon monoxide (CO), ozone (O3), nitrogen dioxide (NO2), sulphur dioxide (SO2), ammonia (NH3), and lead (Pb) are the eight main pollutants that must be considered when calculating the AQI. A minimum of three pollutants' data must be available for the AQI to be calculated, one of which must be PM2.5 or PM10. The AQI, which ranges from 0-500, has varying concentrations of each pollutant and corresponding health impacts.

1.4 Indian AQI range & probable impacts

0-50: This range defines air quality as good as it shows minimal or no impact on health.

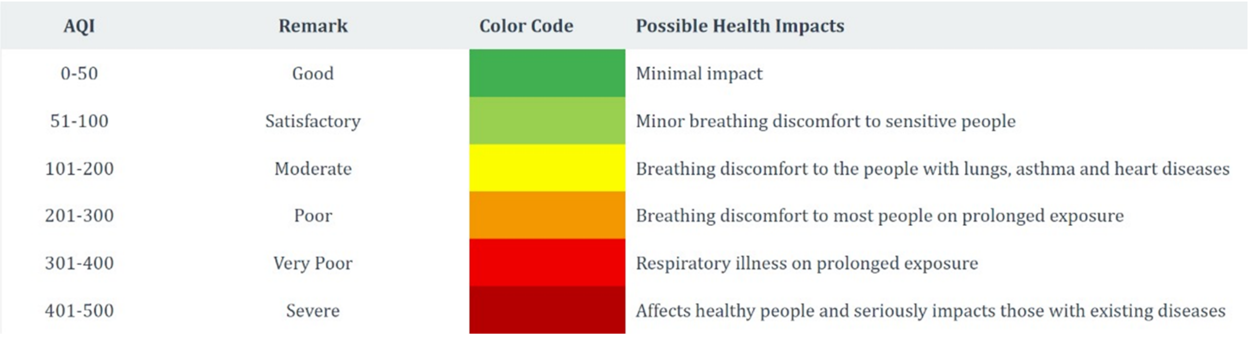
51-100: This is a satisfactory air quality range and it can show effects such as breathing difficulty in sensitive groups.

101-200: The range shows moderate air quality with impacts such as breathing discomfort for children and elderly people, and people already suffering from lung disorders and heart disease.

201-300: AQI falling in this range communicates that the air quality is poor and shows health effects on people when exposed for the long term. People already suffering from heart diseases can experience discomfort from short exposure.

301-400: This range shows very poor air quality and causes respiratory illness for a longer duration of exposure.

401-500: This is the severe range of AQI causing health impacts to normal and diseased people. It also causes severe health impacts on sensitive groups.



AQI	Remark	Color Code	Possible Health Impacts
0-50	Good	🟩	Minimal impact
51-100	Satisfactory	🟩	Minor breathing discomfort to sensitive people
101-200	Moderate	🟨	Breathing discomfort to the people with lungs, asthma and heart diseases
201-300	Poor	🟧	Breathing discomfort to most people on prolonged exposure
301-400	Very Poor	🟥	Respiratory illness on prolonged exposure
401-500	Severe	🟥	Affects healthy people and seriously impacts those with existing diseases

National Air Quality Index given on CPCB (Central Pollution Control Board) India

2. Key objectives

Data Collection:

Collect extensive data on air quality parameters from multiple cities, including PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene, Toluene, Xylene, AQI (Air Quality Index), and AQI Bucket classifications. Deploy air quality monitoring stations strategically within each city to record pollutant concentrations at regular intervals.

Data Preprocessing:

Clean and prepare the air quality data for analysis. Address missing values,duplicates and outliers to ensure the dataset's integrity and accuracy in capturing air quality variations across different locations and time periods.

Exploratory Data Analysis (EDA):

Conduct a thorough exploration of the air quality dataset. Gain insights into the distribution of pollutant concentrations, examine trends over time, and identify potential patterns or anomalies that may influence air quality.

Feature Engineering:

Derive additional features or metrics that enhance the dataset's richness. Explore relationships between different pollutants and identify composite indicators that may provide deeper insights into air quality dynamics.

Statistical Analysis:

Utilise statistical methods to investigate correlations between individual pollutant concentrations and the overall Air Quality Index. Understand how specific pollutants contribute to the overall air quality assessment.

Visualisation:

Create compelling visualizations, such as line charts, bar graphs, and geographical maps, to represent air quality trends and variations. Visualize the impact of different pollutants on the Air Quality Index.

Interpretation:

Analyze the results to identify which pollutants have the most significant impact on the Air Quality Index. Provide insights into the factors influencing air quality and potential sources of pollution in different urban environments.

Validation:

Validate the analysis by comparing the predicted air quality index values to actual measurements. Consider historical air quality data to assess the accuracy of the model in predicting air quality variations.

Conclusion and Recommendations:

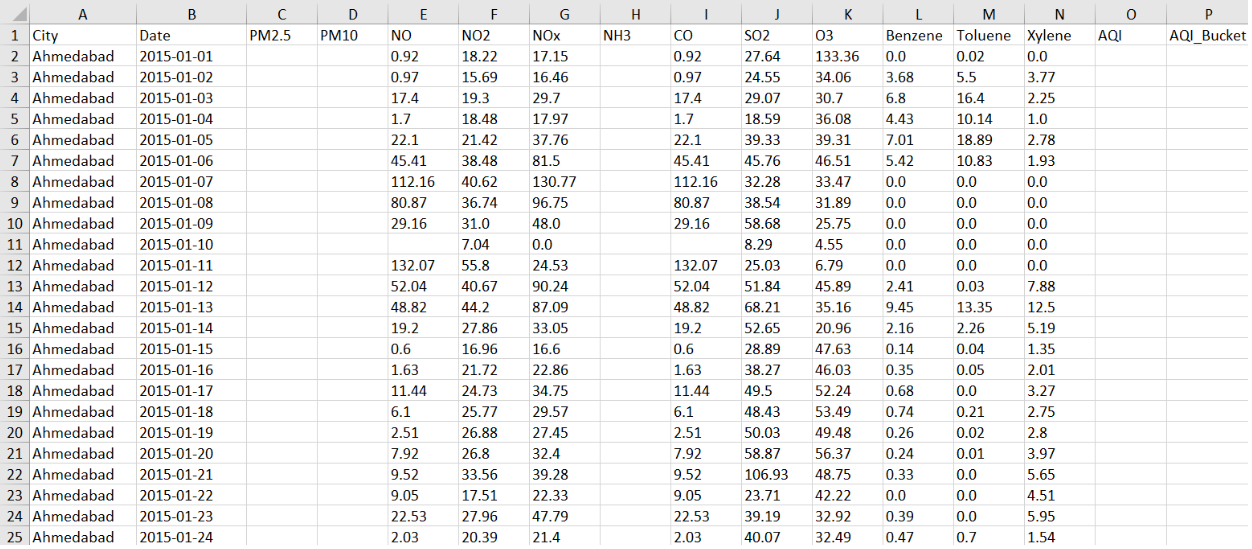
Summarize the findings and offer recommendations for policymakers, environmental agencies, and the public on effective measures to improve air quality. Provide insights into areas that require targeted interventions to address specific pollution sources.

Documentation and Reporting:

Present the results and analysis in a clear and accessible format, potentially through reports or presentations. Ensure that the information is understandable to a broad audience, facilitating informed decision-making regarding air quality management.

3. Dataset used

The Dataset that we’ve used is collected from Kaggle. Following figure, a dataset snapshot is displayed. City, Date, PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene, Toluene, Xylene, AQI, and AQI Index are all included in the dataset in csv file. The sampling date is described by the 'Date' attribute, whereas the other attributes provide the specific air concentration for each. The air quality index is described by AQI and AQI Index. For analysis, data from 2015 to 2020 has been gathered.



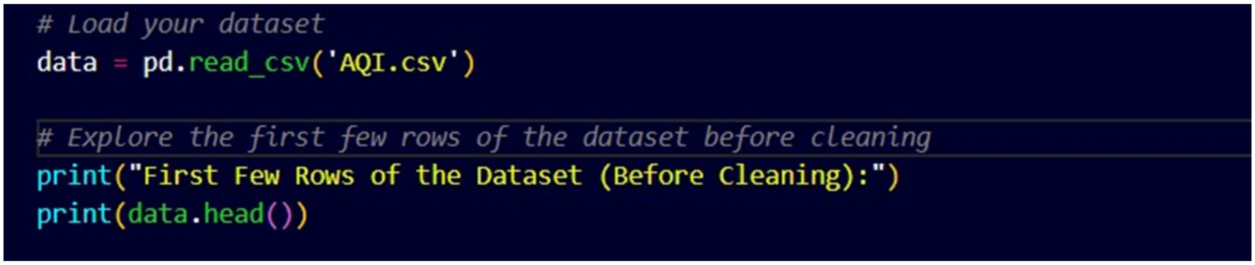
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	City	Date	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene	AQI	AQI_Bucket
2	Ahmedabad	2015-01-01			0.92	18.22	17.15		0.92	27.64	133.36	0.0	0.02	0.0		
3	Ahmedabad	2015-01-02			0.97	15.69	16.46		0.97	24.55	34.06	3.68	5.5	3.77		
4	Ahmedabad	2015-01-03			17.4	19.3	29.7		17.4	29.07	30.7	6.8	16.4	2.25		
5	Ahmedabad	2015-01-04			1.7	18.48	17.97		1.7	18.59	36.08	4.43	10.14	1.0		
6	Ahmedabad	2015-01-05			22.1	21.42	37.76		22.1	39.33	39.31	7.01	18.89	2.78		
7	Ahmedabad	2015-01-06			45.41	38.48	81.5		45.41	45.76	46.51	5.42	10.83	1.93		
8	Ahmedabad	2015-01-07			112.16	40.62	130.77		112.16	32.28	33.47	0.0	0.0	0.0		
9	Ahmedabad	2015-01-08			80.87	36.74	96.75		80.87	38.54	31.89	0.0	0.0	0.0		
10	Ahmedabad	2015-01-09			29.16	31.0	48.0		29.16	58.68	25.75	0.0	0.0	0.0		
11	Ahmedabad	2015-01-10				7.04	0.0			8.29	4.55	0.0	0.0	0.0		
12	Ahmedabad	2015-01-11			132.07	55.8	24.53		132.07	25.03	6.79	0.0	0.0	0.0		
13	Ahmedabad	2015-01-12			52.04	40.67	90.24		52.04	51.84	45.89	2.41	0.03	7.88		
14	Ahmedabad	2015-01-13			48.82	44.2	87.09		48.82	68.21	35.16	9.45	13.35	12.5		
15	Ahmedabad	2015-01-14			19.2	27.86	33.05		19.2	52.65	20.96	2.16	2.26	5.19		
16	Ahmedabad	2015-01-15			0.6	16.96	16.6		0.6	28.89	47.63	0.14	0.04	1.35		
17	Ahmedabad	2015-01-16			1.63	21.72	22.86		1.63	38.27	46.03	0.35	0.05	2.01		
18	Ahmedabad	2015-01-17			11.44	24.73	34.75		11.44	49.5	52.24	0.68	0.0	3.27		
19	Ahmedabad	2015-01-18			6.1	25.77	29.57		6.1	48.43	53.49	0.74	0.21	2.75		
20	Ahmedabad	2015-01-19			2.51	26.88	27.45		2.51	50.03	49.48	0.26	0.02	2.8		
21	Ahmedabad	2015-01-20			7.92	26.8	32.4		7.92	58.87	56.37	0.24	0.01	3.97		
22	Ahmedabad	2015-01-21			9.52	33.56	39.28		9.52	106.93	48.75	0.33	0.0	5.65		
23	Ahmedabad	2015-01-22			9.05	17.51	22.33		9.05	23.71	42.22	0.0	0.0	4.51		
24	Ahmedabad	2015-01-23			22.53	27.96	47.79		22.53	39.19	32.92	0.39	0.0	5.95		
25	Ahmedabad	2015-01-24			2.03	20.39	21.4		2.03	40.07	32.49	0.47	0.7	1.54		

Dataset before cleaning

4. Descriptive Analysis

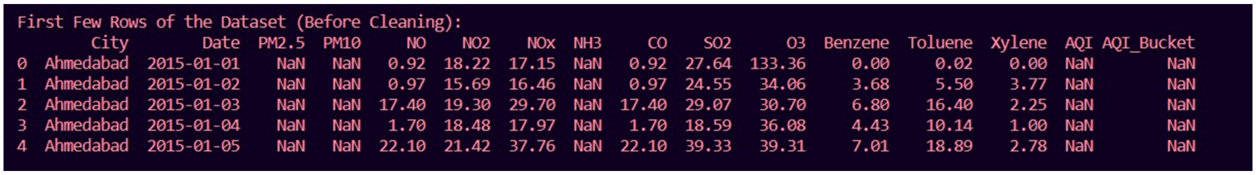
4.1 Data Collection

This dataset compiles air quality measurements from various urban areas, covering pollutants such as PM2.5, PM10, NO, NO2, NOx, NH3, CO, SO2, O3, Benzene, Toluene, and Xylene. Collected systematically by strategically placed monitoring stations across multiple cities, the dataset captures temporal patterns and seasonal variations in air quality. Each observation includes a specific date, enabling detailed analysis, and incorporates the Air Quality Index (AQI) for an overall assessment. The dataset's geographic diversity ensures representation across regions, supporting comparative analyses and insights into urban air quality dynamics.



```python
# Load your dataset
data = pd.read_csv('AQI.csv')

# Explore the first few rows of the dataset before cleaning
print("First Few Rows of the Dataset (Before Cleaning):")
print(data.head())
```

Code to print first few rows of dataset



```
First Few Rows of the Dataset (Before Cleaning):
        City        Date  PM2.5  PM10     NO    NO2    NOx  NH3     CO    SO2      O3  Benzene  Toluene  Xylene  AQI AQI_Bucket
0  Ahmedabad  2015-01-01    NaN   NaN   0.92  18.22  17.15  NaN   0.92  27.64  133.36     0.00     0.02    0.00  NaN        NaN
1  Ahmedabad  2015-01-02    NaN   NaN   0.97  15.69  16.46  NaN   0.97  24.55   34.06     3.68     5.50    3.77  NaN        NaN
2  Ahmedabad  2015-01-03    NaN   NaN  17.40  19.30  29.70  NaN  17.40  29.07   30.70     6.80    16.40    2.25  NaN        NaN
3  Ahmedabad  2015-01-04    NaN   NaN   1.70  18.48  17.97  NaN   1.70  18.59   36.08     4.43    10.14    1.00  NaN        NaN
4  Ahmedabad  2015-01-05    NaN   NaN  22.10  21.42  37.76  NaN  22.10  39.33   39.31     7.01    18.89    2.78  NaN        NaN
```

First few rows of dataset

Data Types of each attribute

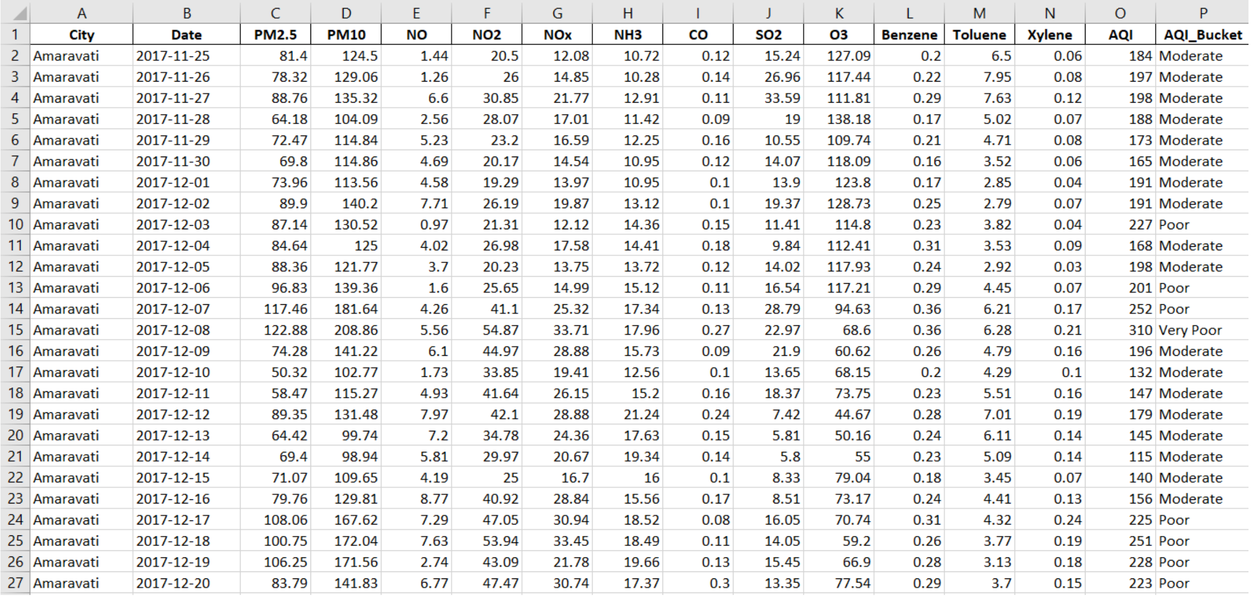
- **City :** Nominal
- **Date :** Interval
- **PM2.5 :** Ratio

- **PM10 :** Ratio
- **NO :** Ratio
- **NO2 :** Ratio
- **NOx :** Ratio
- **NH3 :** Ratio
- **CO :** Ratio
- **SO2 :** Ratio
- **O3 :** Ratio
- **Benzene :** Ratio
- **Toluene :** Ratio
- **Xylene :** Ratio
- **AQI :** Ratio
- **AQI_Bucket :** Ordinal

4.2 Data Preprocessing

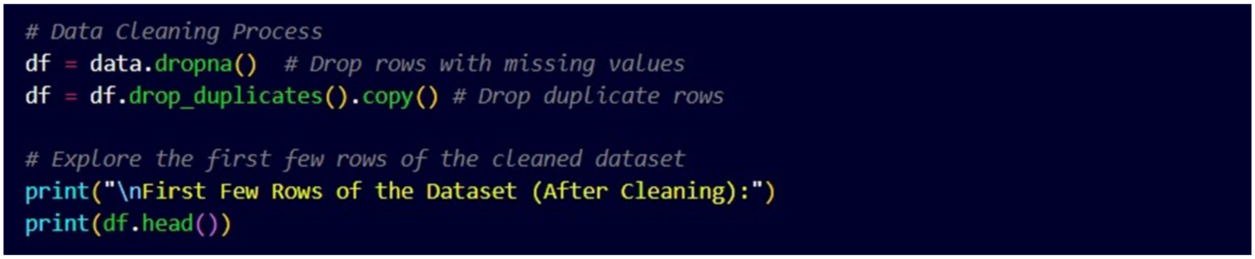
During the preprocessing phase, the dataset undergoes rigorous **cleaning** to ensure data integrity. Steps include **handling missing values,duplicates** and outliers, promoting a reliable foundation for subsequent analyses. This meticulous process aims to enhance the accuracy and consistency of the air quality data, preparing it for insightful exploration and interpretation.

Data set after cleaning:



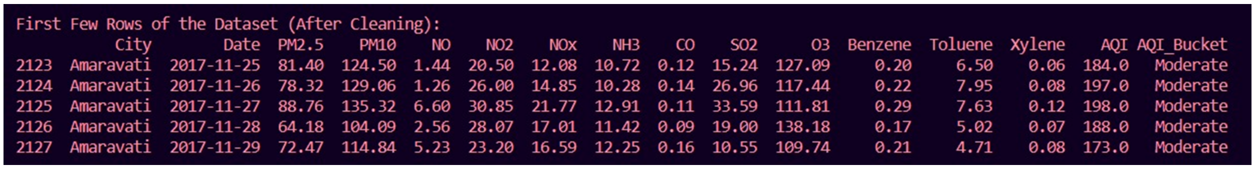
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	**City**	**Date**	**PM2.5**	**PM10**	**NO**	**NO2**	**NOx**	**NH3**	**CO**	**SO2**	**O3**	**Benzene**	**Toluene**	**Xylene**	**AQI**	**AQI_Bucket**
2	Amaravati	2017-11-25	81.4	124.5	1.44	20.5	12.08	10.72	0.12	15.24	127.09	0.2	6.5	0.06	184	Moderate
3	Amaravati	2017-11-26	78.32	129.06	1.26	26	14.85	10.28	0.14	26.96	117.44	0.22	7.95	0.08	197	Moderate
4	Amaravati	2017-11-27	88.76	135.32	6.6	30.85	21.77	12.91	0.11	33.59	111.81	0.29	7.63	0.12	198	Moderate
5	Amaravati	2017-11-28	64.18	104.09	2.56	28.07	17.01	11.42	0.09	19	138.18	0.17	5.02	0.07	188	Moderate
6	Amaravati	2017-11-29	72.47	114.84	5.23	23.2	16.59	12.25	0.16	10.55	109.74	0.21	4.71	0.08	173	Moderate
7	Amaravati	2017-11-30	69.8	114.86	4.69	20.17	14.54	10.95	0.12	14.07	118.09	0.16	3.52	0.06	165	Moderate
8	Amaravati	2017-12-01	73.96	113.56	4.58	19.29	13.97	10.95	0.1	13.9	123.8	0.17	2.85	0.04	191	Moderate
9	Amaravati	2017-12-02	89.9	140.2	7.71	26.19	19.87	13.12	0.1	19.37	128.73	0.25	2.79	0.07	191	Moderate
10	Amaravati	2017-12-03	87.14	130.52	0.97	21.31	12.12	14.36	0.15	11.41	114.8	0.23	3.82	0.04	227	Poor
11	Amaravati	2017-12-04	84.64	125	4.02	26.98	17.58	14.41	0.18	9.84	112.41	0.31	3.53	0.09	168	Moderate
12	Amaravati	2017-12-05	88.36	121.77	3.7	20.23	13.75	13.72	0.12	14.02	117.93	0.24	2.92	0.03	198	Moderate
13	Amaravati	2017-12-06	96.83	139.36	1.6	25.65	14.99	15.12	0.11	16.54	117.21	0.29	4.45	0.07	201	Poor
14	Amaravati	2017-12-07	117.46	181.64	4.26	41.1	25.32	17.34	0.13	28.79	94.63	0.36	6.21	0.17	252	Poor
15	Amaravati	2017-12-08	122.88	208.86	5.56	54.87	33.71	17.96	0.27	22.97	68.6	0.36	6.28	0.21	310	Very Poor
16	Amaravati	2017-12-09	74.28	141.22	6.1	44.97	28.88	15.73	0.09	21.9	60.62	0.26	4.79	0.16	196	Moderate
17	Amaravati	2017-12-10	50.32	102.77	1.73	33.85	19.41	12.56	0.1	13.65	68.15	0.2	4.29	0.1	132	Moderate
18	Amaravati	2017-12-11	58.47	115.27	4.93	41.64	26.15	15.2	0.16	18.37	73.75	0.23	5.51	0.16	147	Moderate
19	Amaravati	2017-12-12	89.35	131.48	7.97	42.1	28.88	21.24	0.24	7.42	44.67	0.28	7.01	0.19	179	Moderate
20	Amaravati	2017-12-13	64.42	99.74	7.2	34.78	24.36	17.63	0.15	5.81	50.16	0.24	6.11	0.14	145	Moderate
21	Amaravati	2017-12-14	69.4	98.94	5.81	29.97	20.67	19.34	0.14	5.8	55	0.23	5.09	0.14	115	Moderate
22	Amaravati	2017-12-15	71.07	109.65	4.19	25	16.7	16	0.1	8.33	79.04	0.18	3.45	0.07	140	Moderate
23	Amaravati	2017-12-16	79.76	129.81	8.77	40.92	28.84	15.56	0.17	8.51	73.17	0.24	4.41	0.13	156	Moderate
24	Amaravati	2017-12-17	108.06	167.62	7.29	47.05	30.94	18.52	0.08	16.05	70.74	0.31	4.32	0.24	225	Poor
25	Amaravati	2017-12-18	100.75	172.04	7.63	53.94	33.45	18.49	0.11	14.05	59.2	0.26	3.77	0.19	251	Poor
26	Amaravati	2017-12-19	106.25	171.56	2.74	43.09	21.78	19.66	0.13	15.45	66.9	0.28	3.13	0.18	228	Poor
27	Amaravati	2017-12-20	83.79	141.83	6.77	47.47	30.74	17.37	0.3	13.35	77.54	0.29	3.7	0.15	223	Poor

Code to remove missing values & duplicates & print first few lines of dataset:



```
# Data Cleaning Process
df = data.dropna()  # Drop rows with missing values
df = df.drop_duplicates().copy() # Drop duplicate rows

# Explore the first few rows of the cleaned dataset
print("\nFirst Few Rows of the Dataset (After Cleaning):")
print(df.head())
```

First few lines of dataset after cleaning:



```
First Few Rows of the Dataset (After Cleaning):
           City        Date  PM2.5    PM10    NO    NO2    NOx    NH3    CO    SO2      O3  Benzene  Toluene  Xylene    AQI AQI_Bucket
2123  Amaravati  2017-11-25  81.40  124.50  1.44  20.50  12.08  10.72  0.12  15.24  127.09     0.20     6.50    0.06  184.0   Moderate
2124  Amaravati  2017-11-26  78.32  129.06  1.26  26.00  14.85  10.28  0.14  26.96  117.44     0.22     7.95    0.08  197.0   Moderate
2125  Amaravati  2017-11-27  88.76  135.32  6.60  30.85  21.77  12.91  0.11  33.59  111.81     0.29     7.63    0.12  198.0   Moderate
2126  Amaravati  2017-11-28  64.18  104.09  2.56  28.07  17.01  11.42  0.09  19.00  138.18     0.17     5.02    0.07  188.0   Moderate
2127  Amaravati  2017-11-29  72.47  114.84  5.23  23.20  16.59  12.25  0.16  10.55  109.74     0.21     4.71    0.08  173.0   Moderate
```

4.3 Exploratory Data Analysis

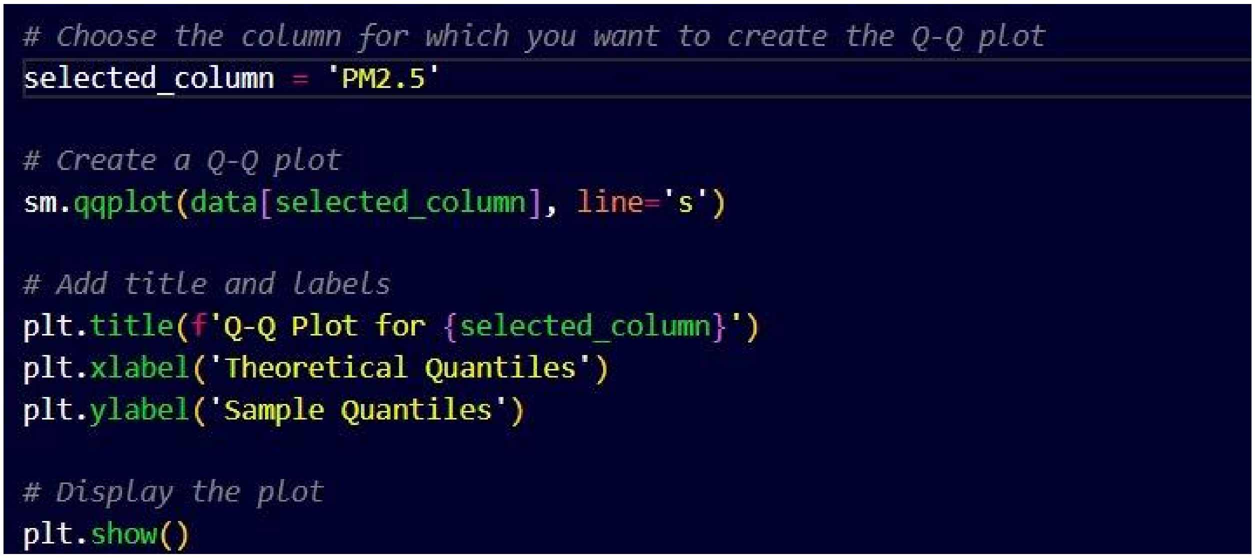
i) Finding correlation between attributes

Checking for normal distribution is an important step before applying certain statistical tests, especially those that assume normality.

There are many methods like histogram,Q-Q Plot (Quantile-Quantile Plot),Shapiro-Wilk Test,Anderson-Darling Test,Kolmogorov-Smirnov Test to check normality.no test can definitively prove that data follows a normal distribution. They are tools to provide evidence for or against the assumption of normality.

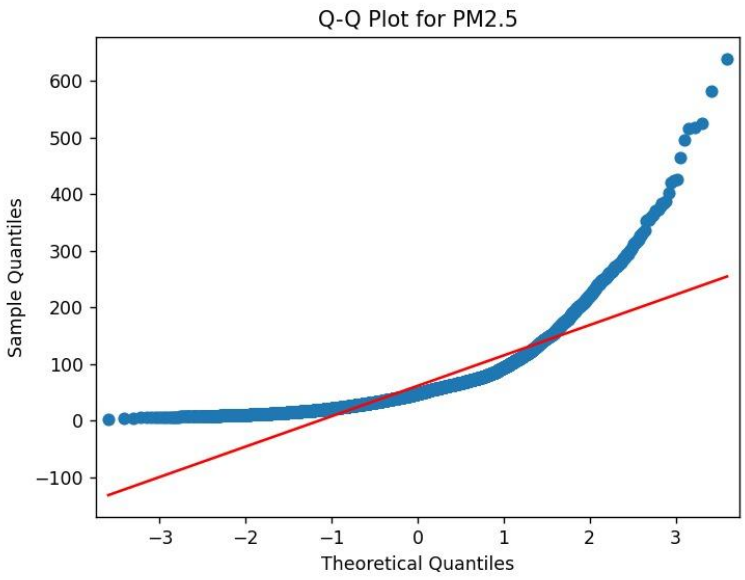
Q-Q Plot:

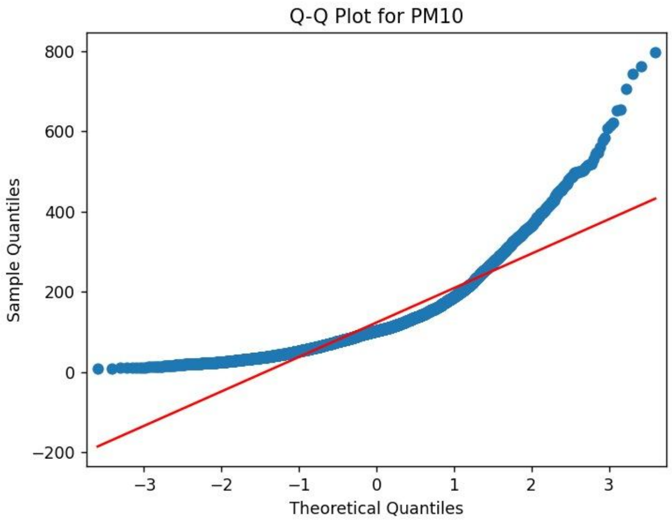
- If the points lie approximately on the straight line, it suggests that the data follows a normal distribution.
- It is a graphical tool to assess if a dataset follows a particular theoretical distribution (such as the normal distribution).
- Deviations from the straight line indicate departures from normality.

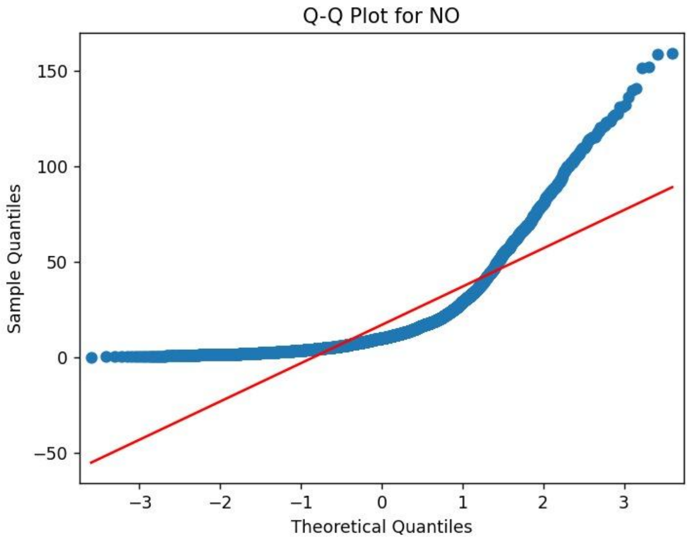


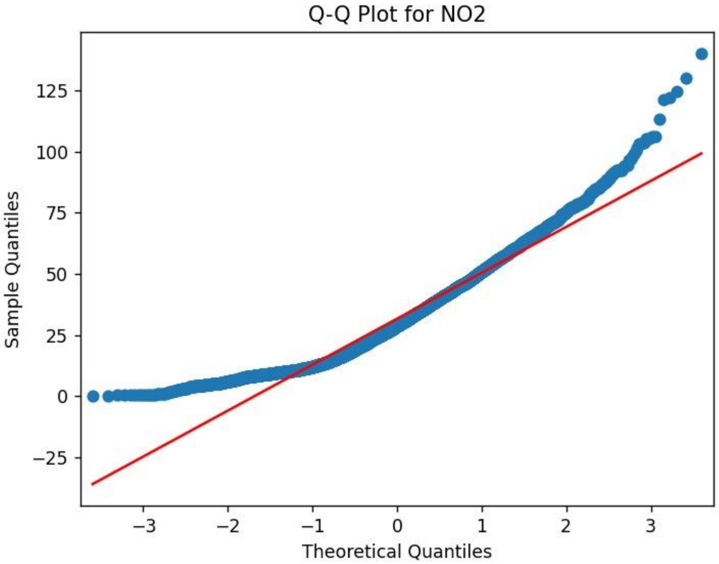
```
# Choose the column for which you want to create the Q-Q plot
selected_column = 'PM2.5'

# Create a Q-Q plot
sm.qqplot(data[selected_column], line='s')

# Add title and labels
plt.title(f'Q-Q Plot for {selected_column}')
plt.xlabel('Theoretical Quantiles')
plt.ylabel('Sample Quantiles')

# Display the plot
plt.show()
```

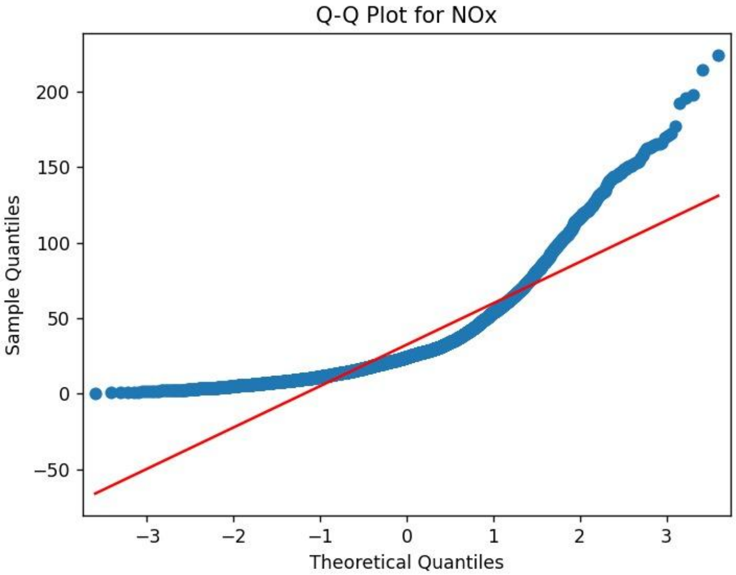
Code for plotting Q-Q plot

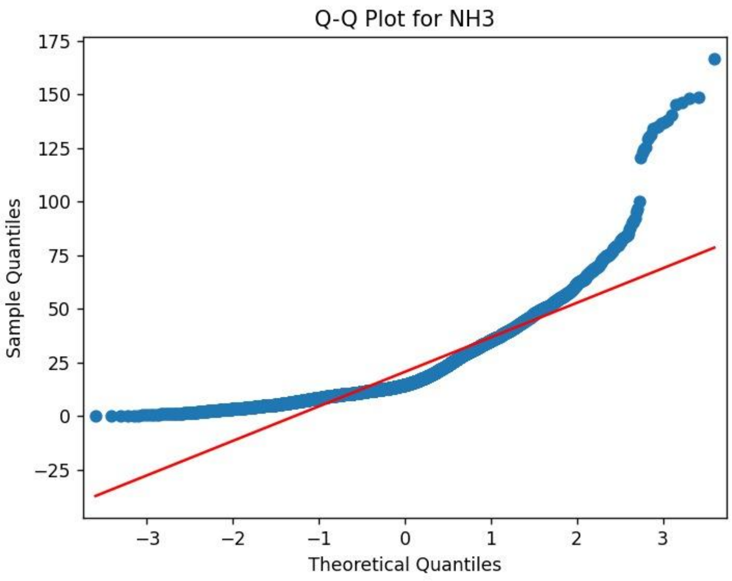


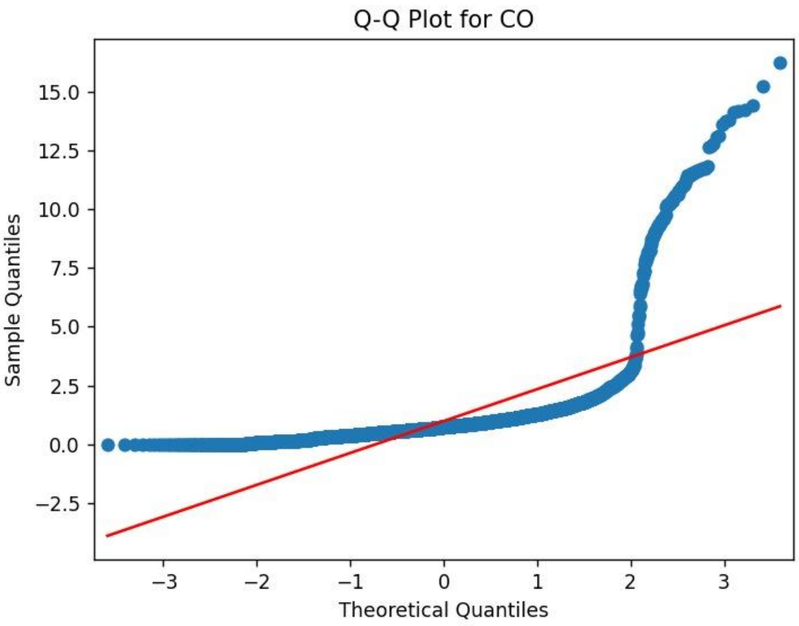

Q-Q Plot for PM10
Sample Quantiles
Theoretical Quantiles
800
600
400
200
0
−200
−3
−2
−1
0
1
2
3

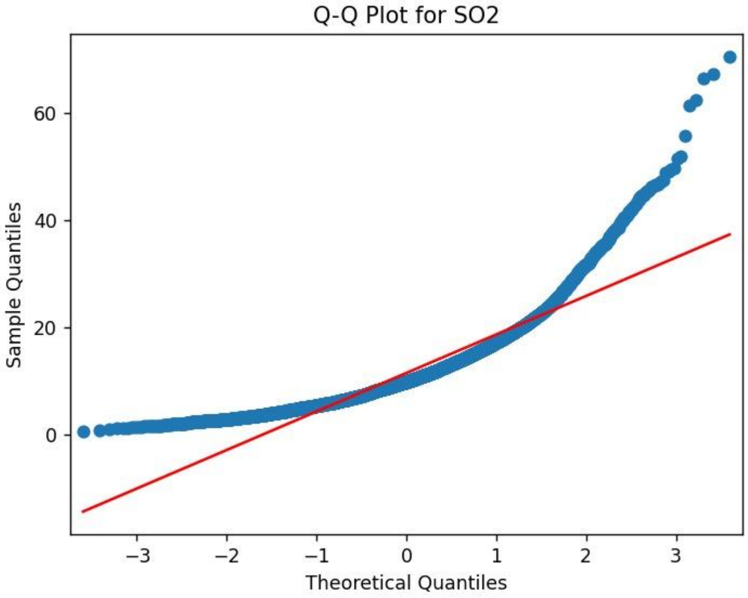

Q-Q Plot for NO
Sample Quantiles
Theoretical Quantiles
150
100
50
0
−50
−3
−2
−1
0
1
2
3

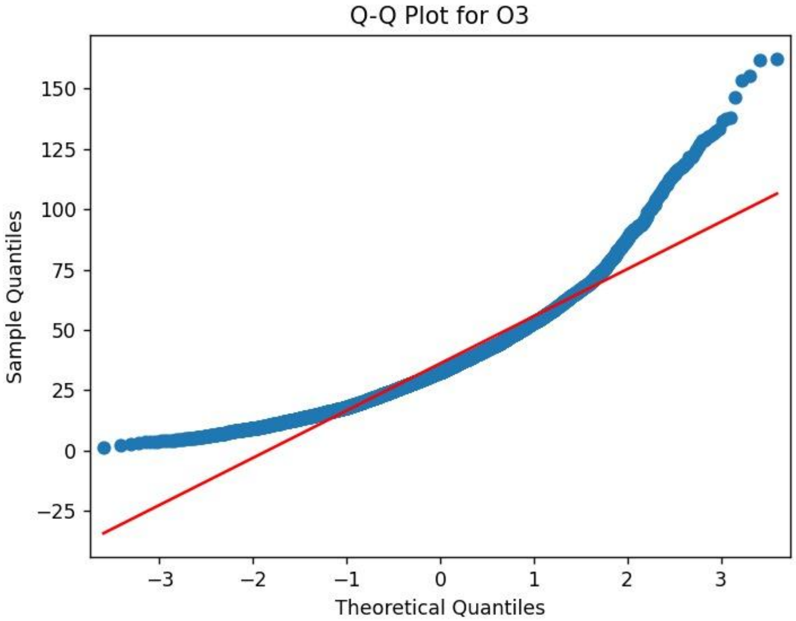

Q-Q Plot for NO2
Sample Quantiles
125
100
75
50
25
0
−25
−3
−2
−1
0
1
2
3
Theoretical Quantiles

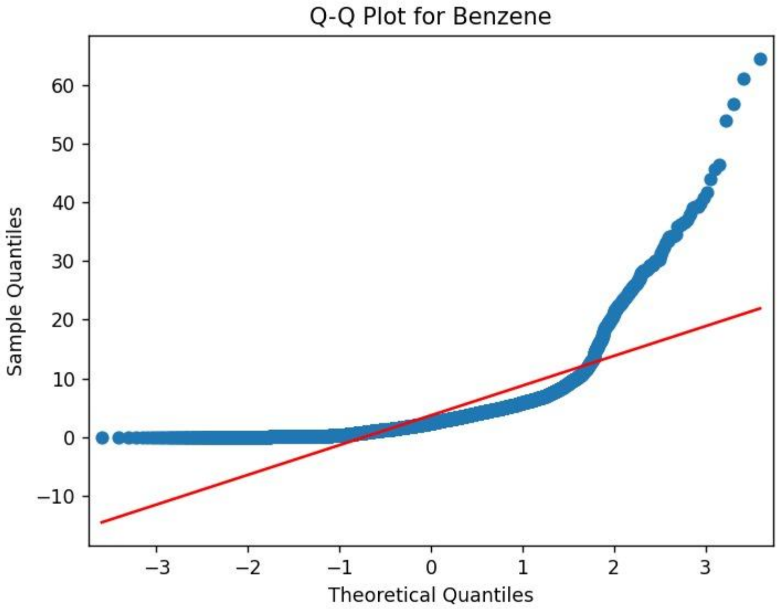

Q-Q Plot for NOx
Sample Quantiles
200
150
100
50
0
−50
−3
−2
−1
0
1
2
3
Theoretical Quantiles

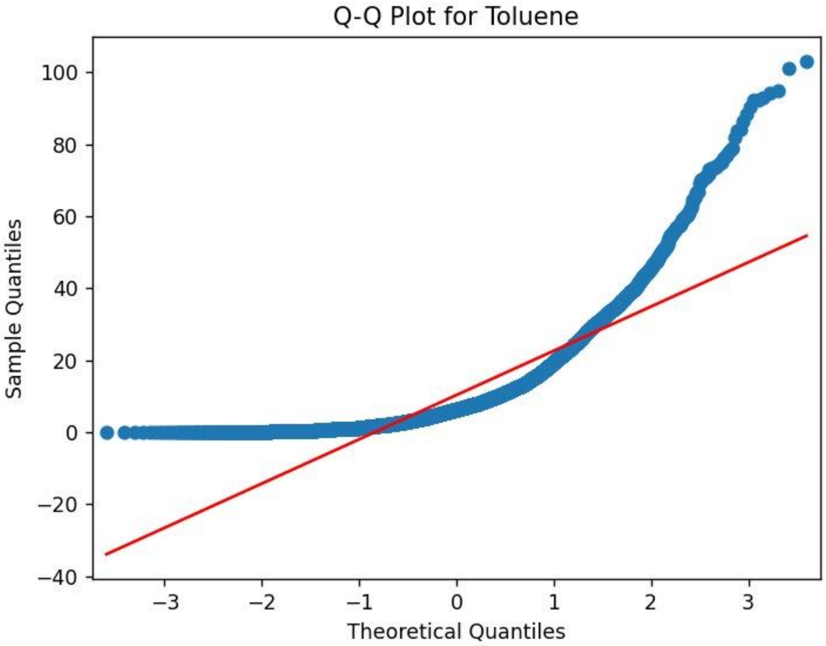

Q-Q Plot for NH3
Sample Quantiles
Theoretical Quantiles
175
150
125
100
75
50
25
0
−25
−3
−2
−1
0
1
2
3

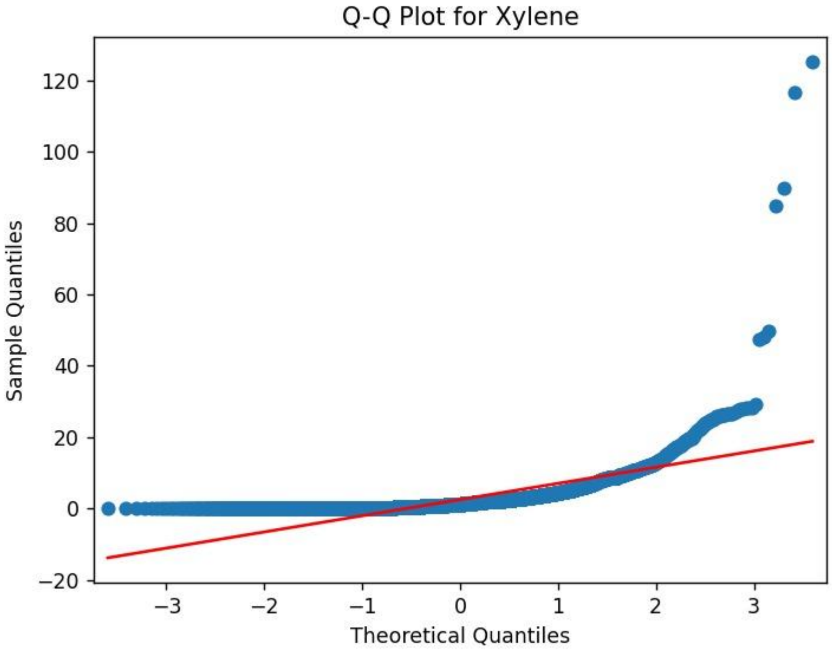

Q-Q Plot for CO
Sample Quantiles
Theoretical Quantiles
15.0
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
−3
−2
−1
0
1
2
3

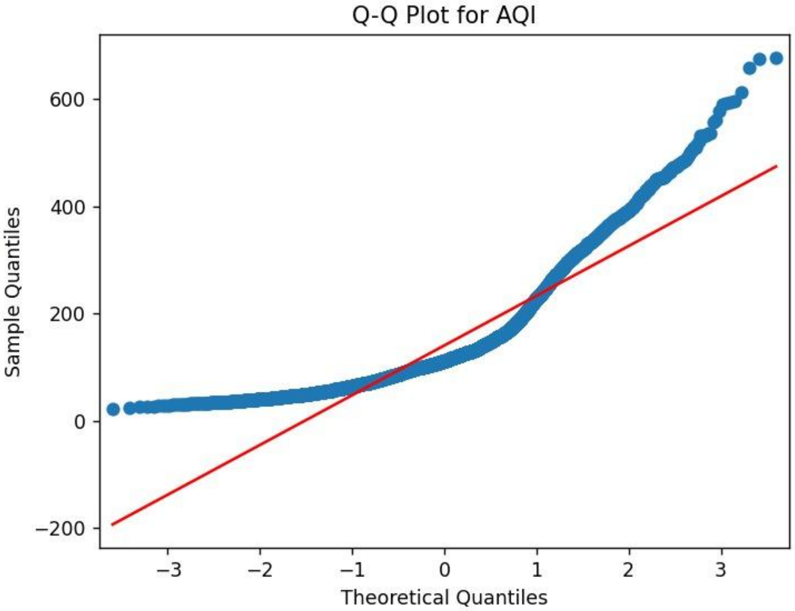

Q-Q Plot for SO2
Sample Quantiles
Theoretical Quantiles
60
40
20
0
−3
−2
−1
0
1
2
3


Q-Q Plot for O3
Sample Quantiles
Theoretical Quantiles
150
125
100
75
50
25
0
−25
−3
−2
−1
0
1
2
3


Q-Q Plot for Benzene
Sample Quantiles
60
50
40
30
20
10
0
−10
−3
−2
−1
0
1
2
3
Theoretical Quantiles


Q-Q Plot for Toluene
Sample Quantiles
100
80
60
40
20
0
−20
−40
−3
−2
−1
0
1
2
3
Theoretical Quantiles





Interpretation:

Upon visual inspection of the Q-Q plot, it is evident that the observed points deviate from the expected straight line, indicating **not a normal distribution**. This departure from normality may have implications for statistical analyses that assume normal distribution. Given these observations, alternative non-parametric or robust statistical methods may be considered for analyses where normality assumptions cannot be met.

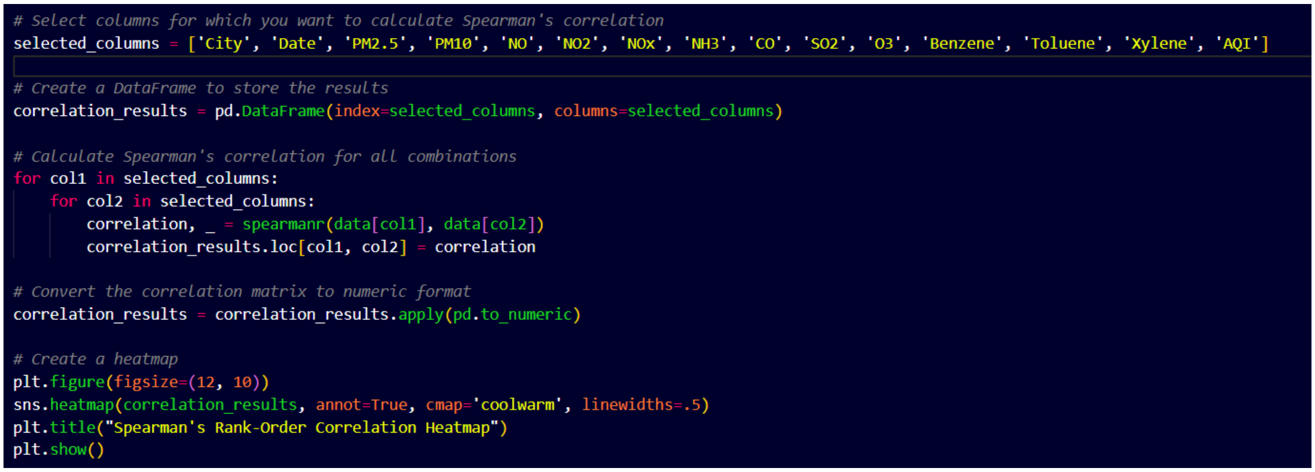
Correlation test

When dealing with data that is not normally distributed, or when the assumption of normality is violated, non-parametric correlation tests are often recommended. Non-parametric tests are robust to deviations from normality and do not rely on specific distributional assumptions.

There are various correlation non-parametric tests such as Spearman's Rank Correlation Coefficient,Kendall's Tau.

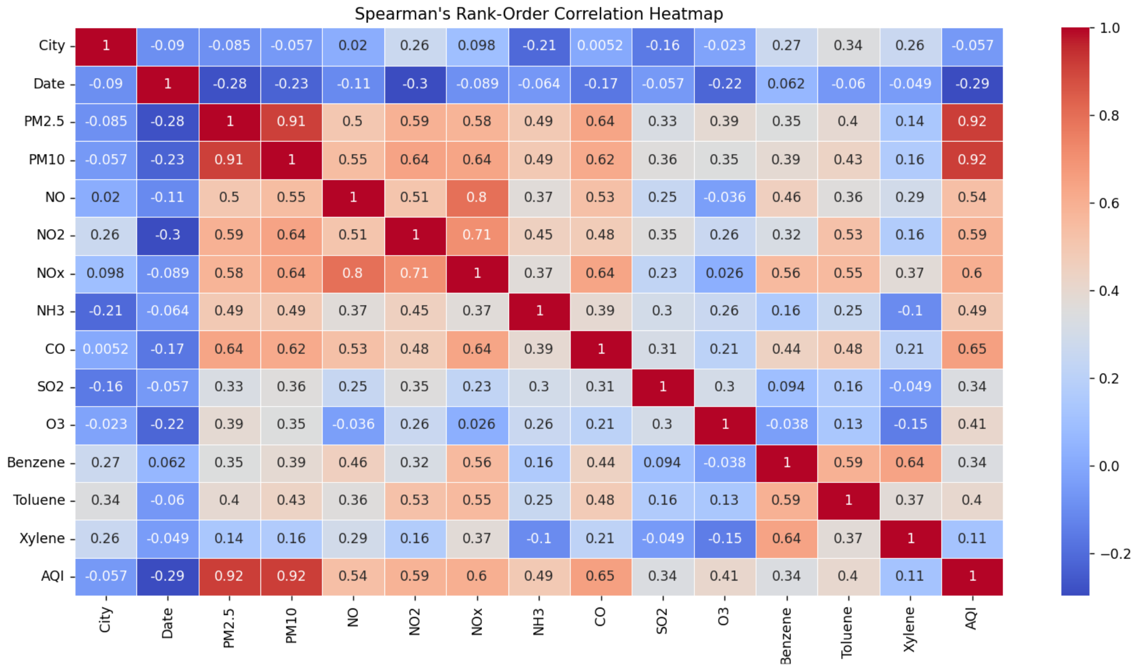
Spearman Correlation:

- Measures the strength and direction of the monotonic relationship between pair of attributes.
- Robust to outliers and non-normality.
- This non-parametric nature is especially pertinent in our analysis, where we are dealing with real-world data that can often deviate from idealised parametric conditions.
- By opting for the Spearman correlation, we can gain a deeper understanding of the monotonic relationships between these attribute values, allowing us to account for the complexities and variations present in our dataset, thus enhancing the quality and reliability of our analysis.
- Alternative to this is Kendall's Tau Rank Correlation; this is better for a smaller dataset.



```
# Select columns for which you want to calculate Spearman's correlation
selected_columns = ['City', 'Date', 'PM2.5', 'PM10', 'NO', 'NO2', 'NOx', 'NH3', 'CO', 'SO2', 'O3', 'Benzene', 'Toluene', 'Xylene', 'AQI']

# Create a DataFrame to store the results
correlation_results = pd.DataFrame(index=selected_columns, columns=selected_columns)

# Calculate Spearman's correlation for all combinations
for col1 in selected_columns:
    for col2 in selected_columns:
        correlation, _ = spearmanr(data[col1], data[col2])
        correlation_results.loc[col1, col2] = correlation

# Convert the correlation matrix to numeric format
correlation_results = correlation_results.apply(pd.to_numeric)

# Create a heatmap
plt.figure(figsize=(12, 10))
sns.heatmap(correlation_results, annot=True, cmap='coolwarm', linewidths=.5)
plt.title("Spearman's Rank-Order Correlation Heatmap")
plt.show()
```

Code for Spearman's Rank-Order Correlation



	City	Date	PM2.5	PM10	NO	NO2	NOx	NH3	CO	SO2	O3	Benzene	Toluene	Xylene	AQI
City	1	-0.09	-0.085	-0.057	0.02	0.26	0.098	-0.21	0.0052	-0.16	-0.023	0.27	0.34	0.26	-0.057
Date	-0.09	1	-0.28	-0.23	-0.11	-0.3	-0.089	-0.064	-0.17	-0.057	-0.22	0.062	-0.06	-0.049	-0.29
PM2.5	-0.085	-0.28	1	0.91	0.5	0.59	0.58	0.49	0.64	0.33	0.39	0.35	0.4	0.14	0.92
PM10	-0.057	-0.23	0.91	1	0.55	0.64	0.64	0.49	0.62	0.36	0.35	0.39	0.43	0.16	0.92
NO	0.02	-0.11	0.5	0.55	1	0.51	0.8	0.37	0.53	0.25	-0.036	0.46	0.36	0.29	0.54
NO2	0.26	-0.3	0.59	0.64	0.51	1	0.71	0.45	0.48	0.35	0.26	0.32	0.53	0.16	0.59
NOx	0.098	-0.089	0.58	0.64	0.8	0.71	1	0.37	0.64	0.23	0.026	0.56	0.55	0.37	0.6
NH3	-0.21	-0.064	0.49	0.49	0.37	0.45	0.37	1	0.39	0.3	0.26	0.16	0.25	-0.1	0.49
CO	0.0052	-0.17	0.64	0.62	0.53	0.48	0.64	0.39	1	0.31	0.21	0.44	0.48	0.21	0.65
SO2	-0.16	-0.057	0.33	0.36	0.25	0.35	0.23	0.3	0.31	1	0.3	0.094	0.16	-0.049	0.34
O3	-0.023	-0.22	0.39	0.35	-0.036	0.26	0.026	0.26	0.21	0.3	1	-0.038	0.13	-0.15	0.41
Benzene	0.27	0.062	0.35	0.39	0.46	0.32	0.56	0.16	0.44	0.094	-0.038	1	0.59	0.64	0.34
Toluene	0.34	-0.06	0.4	0.43	0.36	0.53	0.55	0.25	0.48	0.16	0.13	0.59	1	0.37	0.4
Xylene	0.26	-0.049	0.14	0.16	0.29	0.16	0.37	-0.1	0.21	-0.049	-0.15	0.64	0.37	1	0.11
AQI	-0.057	-0.29	0.92	0.92	0.54	0.59	0.6	0.49	0.65	0.34	0.41	0.34	0.4	0.11	1

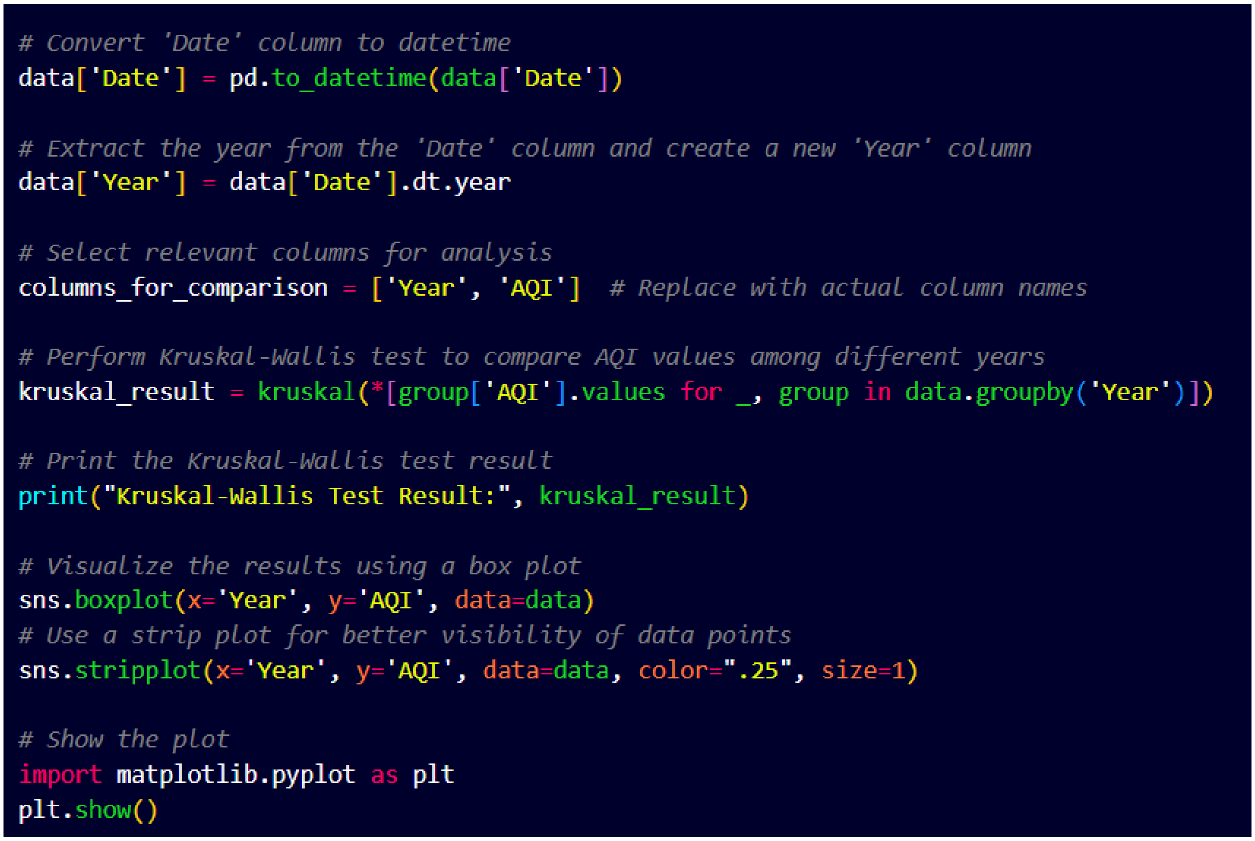
ii)compare AQI year wise

We use the Kruskal-Wallis test as we have **more than two independent samples** and the assumption of normality is violated, making it a suitable **non-parametric** alternative to the one-way analysis of variance (ANOVA).

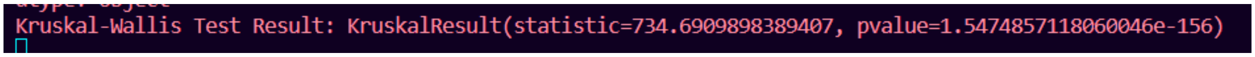
Interpreting the result of the Kruskal-Wallis test:

Null Hypothesis (H0): There is no significant difference in the distribution of AQI values among the different years.

Alternative Hypothesis (H1): There is a significant difference in the distribution of AQI values among the different years.

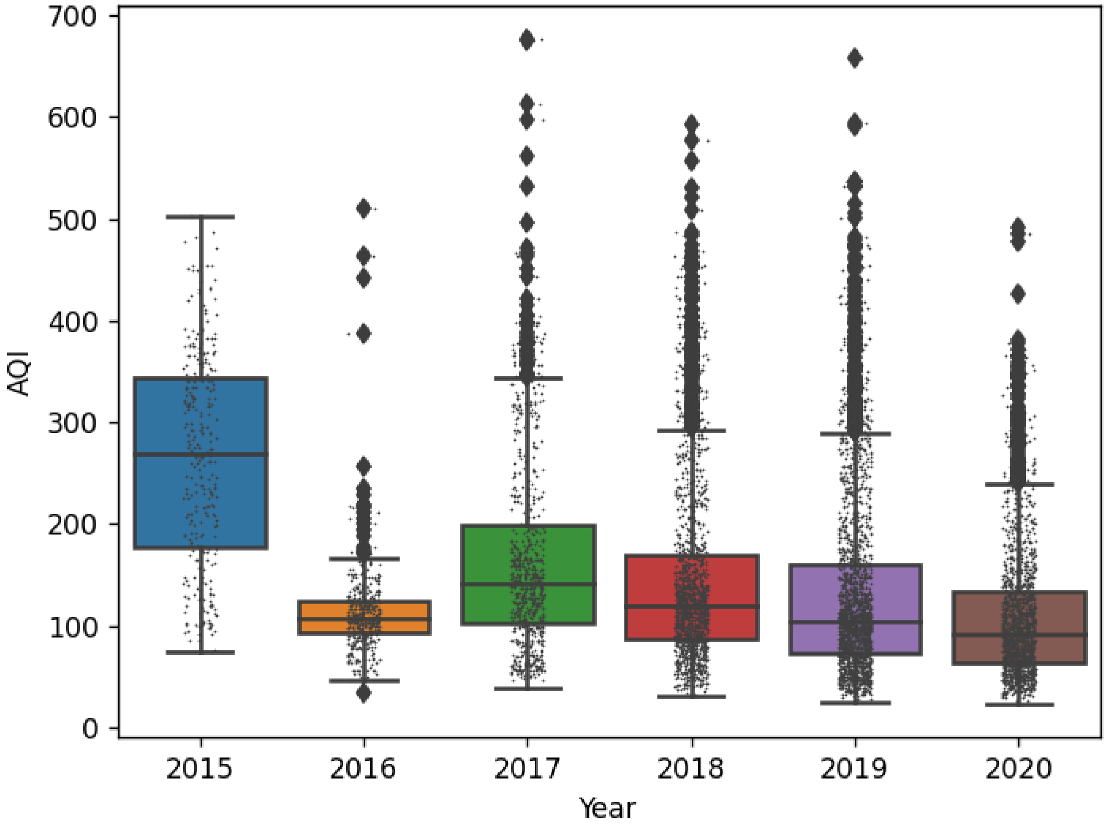


```
# Convert 'Date' column to datetime
data['Date'] = pd.to_datetime(data['Date'])

# Extract the year from the 'Date' column and create a new 'Year' column
data['Year'] = data['Date'].dt.year

# Select relevant columns for analysis
columns_for_comparison = ['Year', 'AQI']  # Replace with actual column names

# Perform Kruskal-Wallis test to compare AQI values among different years
kruskal_result = kruskal(*[group['AQI'].values for _, group in data.groupby('Year')])

# Print the Kruskal-Wallis test result
print("Kruskal-Wallis Test Result:", kruskal_result)

# Visualize the results using a box plot
sns.boxplot(x='Year', y='AQI', data=data)
# Use a strip plot for better visibility of data points
sns.stripplot(x='Year', y='AQI', data=data, color=".25", size=1)

# Show the plot
import matplotlib.pyplot as plt
plt.show()
```



```
dtype: object
Kruskal-Wallis Test Result: KruskalResult(statistic=734.6909898389407, pvalue=1.5474857118060046e-156)
```

As the p-value is less than 0.05 we reject the null hypothesis. There are significant differences in AQI values among the different years.



iii)Are there any notable trends in air quality improvements or deteriorations during the COVID-19 pandemic across the cities in the dataset?

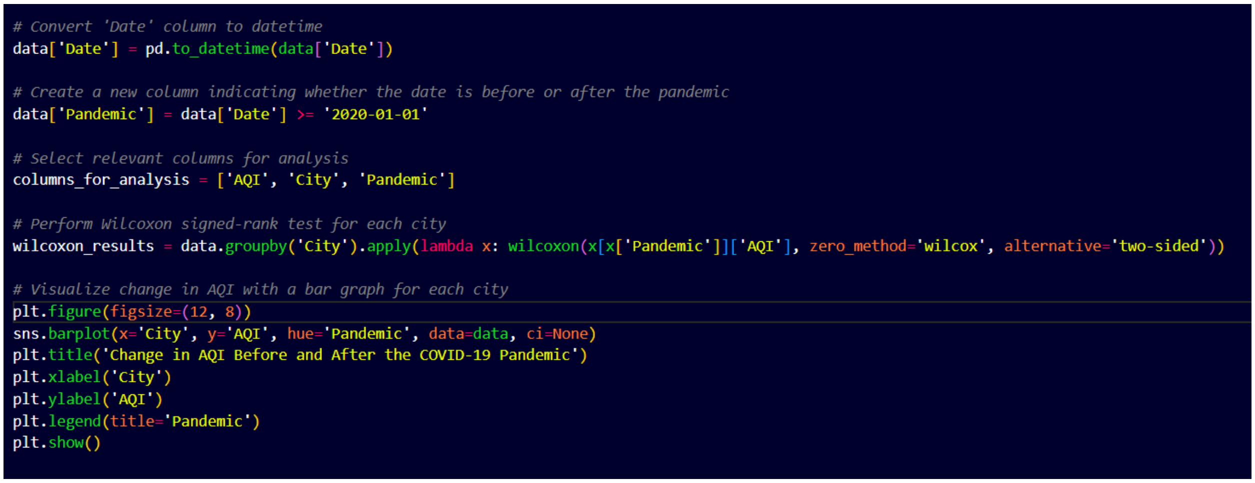
The Wilcoxon signed-rank test is a non-parametric test used to assess whether the distribution of paired differences between two groups is symmetric around zero. It is commonly used when the assumptions of a parametric test (such as the paired t-test) are not met, particularly when the data is not normally distributed.

Here we have **two dependent samples** and data doesn't follow normal distribution(**non-parametric**) hence we are using Wilcoxon signed-rank test.

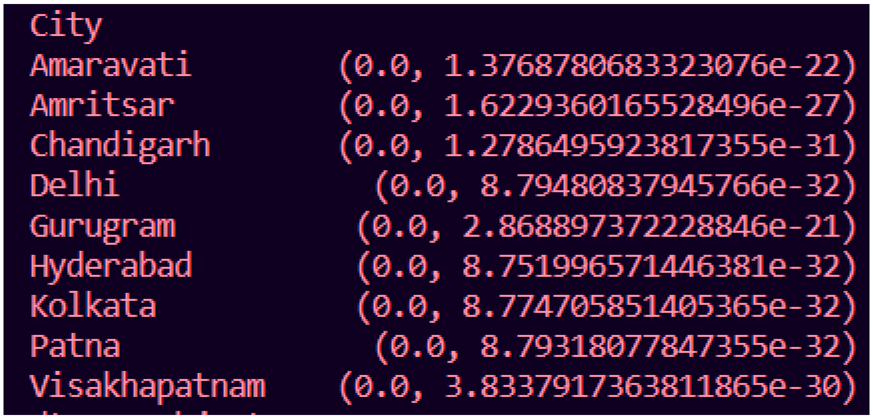
Interpreting the result of the Wilcoxon signed-rank test:

Null Hypothesis (H0): The null hypothesis assumes that there is no difference between the AQI values before and after the pandemic.

Alternative Hypothesis (H1): The alternative hypothesis suggests that there is a significant difference between the AQI values before and after the pandemic.

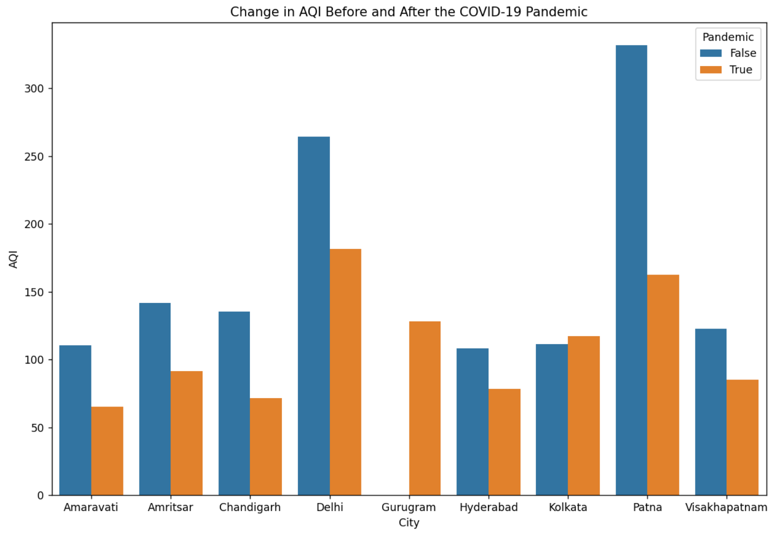


```
# Convert 'Date' column to datetime
data['Date'] = pd.to_datetime(data['Date'])

# Create a new column indicating whether the date is before or after the pandemic
data['Pandemic'] = data['Date'] >= '2020-01-01'

# Select relevant columns for analysis
columns_for_analysis = ['AQI', 'City', 'Pandemic']

# Perform Wilcoxon signed-rank test for each city
wilcoxon_results = data.groupby('City').apply(lambda x: wilcoxon(x[x['Pandemic']]['AQI'], zero_method='wilcox', alternative='two-sided'))

# Visualize change in AQI with a bar graph for each city
plt.figure(figsize=(12, 8))
sns.barplot(x='City', y='AQI', hue='Pandemic', data=data, ci=None)
plt.title('Change in AQI Before and After the COVID-19 Pandemic')
plt.xlabel('City')
plt.ylabel('AQI')
plt.legend(title='Pandemic')
plt.show()
```



```
City
Amaravati        (0.0, 1.3768780683323076e-22)
Amritsar         (0.0, 1.6229360165528496e-27)
Chandigarh       (0.0, 1.2786495923817355e-31)
Delhi              (0.0, 8.79480837945766e-32)
Gurugram          (0.0, 2.868897372228846e-21)
Hyderabad         (0.0, 8.751996571446381e-32)
Kolkata           (0.0, 8.774705851405365e-32)
Patna              (0.0, 8.79318077847355e-32)
Visakhapatnam    (0.0, 3.8337917363811865e-30)
```

As p-value ≤ 0.05 we reject the null hypothesis. This indicates that there is a significant difference in AQI values before and after the pandemic for all cities.



5. Conclusion and Future Scope:

Conclusion:

i) Correlation Between Attributes:

The analysis revealed significant correlations between various air quality attributes, providing insights into potential relationships among pollutants and overall air quality.

ii) Year-Wise Comparison of AQI:

Year-wise comparisons of Air Quality Index (AQI) across cities highlighted variations and trends in air quality over time. This information is crucial for understanding the temporal dynamics of air pollution.

iii) Impact of COVID-19 Pandemic on Air Quality:

The investigation into air quality changes during the COVID-19 pandemic demonstrated notable variations across cities. These findings contribute to understanding the influence of lockdowns and reduced human activities on air pollution levels.

Future Scope:

1) Longitudinal Analysis:

Extend the analysis over a more extended period to identify long-term trends and patterns in air quality, considering seasonal variations and potential contributing factors.

2) Detailed Pollutant Analysis:

Conduct in-depth analyses focusing on specific pollutants to identify their sources, seasonal variations, and the impact on overall air quality.

3) Machine Learning Predictions:

Implement machine learning models to predict future air quality based on historical data, considering various influencing factors.

4) Geospatial Analysis:

Integrate geospatial data to explore spatial patterns of air quality, identifying hotspots and areas with consistent improvements or deteriorations.

5) Public Health Implications:

Collaborate with public health authorities to correlate air quality data with health outcomes, providing valuable insights into the potential health risks associated with varying pollution levels

6. References

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

[2] https://cpcb.nic.in/National-Air-Quality-Index/

[3]https://blogs.worldbank.org/endpovertyinsouthasia/india-air-quality-has-been-improving-despite-covid-19-lockdown