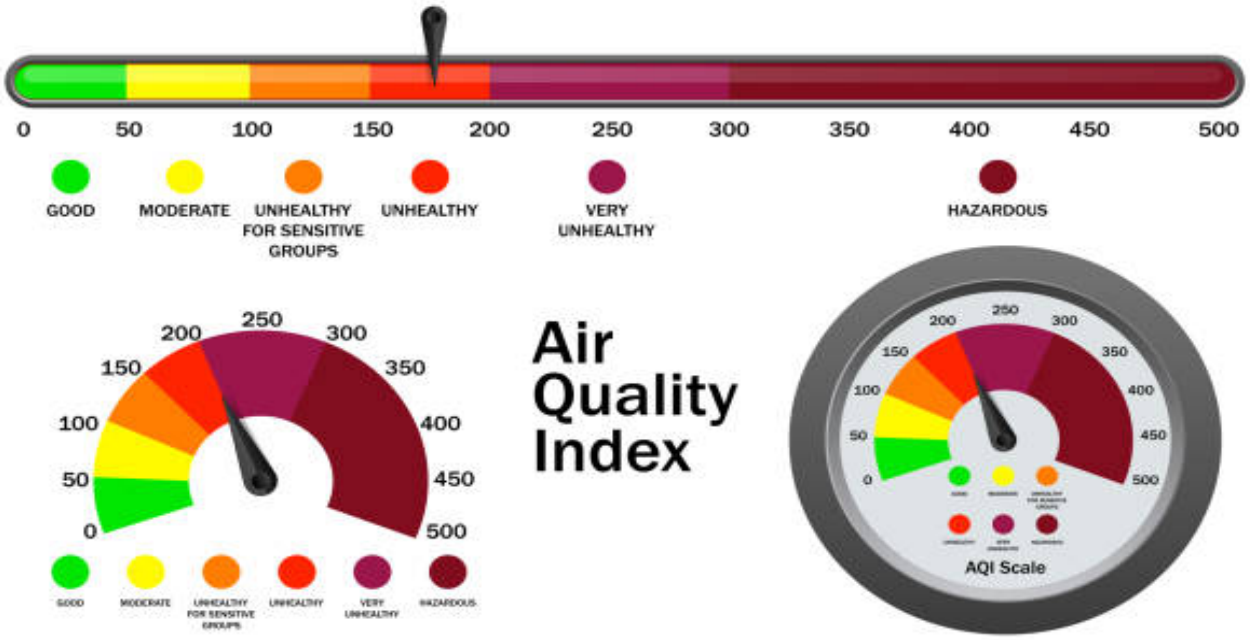
Global Air Quality Index - A Data Driven Analysis on Air Pollution

What is Exploratory Data Analysis?

Exploratory Data Analysis (EDA) is a method of analyzing datasets to understand their main characteristics. It involves summarizing data features, detecting patterns, and uncovering relationships through visual and statistical techniques. EDA helps in gaining insights and formulating hypotheses for further analysis.

Our Motivation

Air pollution is a growing issue worldwide, affecting millions of people each year. From smog-filled cities to rural areas near industrial plants, air quality is a pressing concern. Air quality is crucial for human health, environmental well-being, and economic growth.



Introduction

The project aims to explore global air quality data by analyzing pollutants like **(PM2.5, NO2, CO, and O3)** to identify key patterns, trends and correlations. This project gives the opportunity to analyze real-world data, uncover trends in pollutant levels, and raise awareness about how air quality impacts our daily lives.

PM2.5 : PM2.5 refers to tiny particles or droplets in the air that are 2.5 micrometers or less in width. They can be harmful to human health when inhaled, especially in high concentrations.

Ozone : Ozone is a gas that can form in the atmosphere through a chemical reaction between sunlight and other pollutants. High levels of ozone can be harmful to human health, particularly for those with respiratory issues.

Carbon Monoxide (CO) : CO is a colorless, odorless gas that is produced by the incomplete burning of fossil fuels. High levels of CO can be toxic to humans and can cause headaches, dizziness, and nausea

Nitrogen Dioxide (NO_2) :NO2 is a harmful air pollutant that is primarily produced by the combustion of fossil fuels, such as in vehicle engines, power plants, and industrial activities.It is a significant component of smog and can cause respiratory problems

- # Key Research Questions

1. What is the impact of various pollutants on the Air Quality Index (AQI) value?

This question aims to explore how the concentrations of different pollutants (such as CO, Ozone, NO_2, and PM2.5) affect the overall AQI value. The objective is to understand the relative contribution of each pollutant to the overall air quality index and identify which pollutants have the greatest influence on AQI values across various regions.

2. How are the four key pollutants (CO, Ozone, NO_2, and PM2.5) distributed across the globe?

This research question focuses on the spatial distribution of the four primary pollutants across different regions. It will examine which areas of the world experience the highest levels of pollution for each of these pollutants and provide insight into global air quality patterns and potential hotspots of pollution.

3. Is there a relationship between geographical location (continent, country, city) and AQI values across different regions?

This question seeks to investigate if there are patterns in air quality that are influenced by geographical factors such as the continent, country, or city. The analysis would explore how geographic location might correlate with higher or lower AQI values.

• Dataset Overview

Source of the dataset - Kaggle

[World Air Quality Index by City and Coordinates]

[Countries by Continent]

• Data Preprocessing and Feature Engineering

Step 1: Import Python Libraries

Import all libraries which are required for our analysis, such as Data Loading, Statistical analysis, Visualizations, Data Transformations, Merge and Joins, etc.

```python
# Import Data Science libraries
import pandas as pd
import numpy as np
from scipy.stats import linregress

# Import Data Visualization
import matplotlib.pyplot as plt
import seaborn as sns
import folium
from folium.plugins import MarkerCluster
```

Step 2: Dataset Overview

```
(16695, 14)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16695 entries, 0 to 16694
Data columns (total 14 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Country             16393 non-null  object
 1   City                16695 non-null  object
 2   AQI Value           16695 non-null  int64
 3   AQI Category        16695 non-null  object
 4   CO AQI Value        16695 non-null  int64
 5   CO AQI Category     16695 non-null  object
 6   Ozone AQI Value     16695 non-null  int64
 7   Ozone AQI Category  16695 non-null  object
 8   NO2 AQI Value       16695 non-null  int64
 9   NO2 AQI Category    16695 non-null  object
 10  PM2.5 AQI Value     16695 non-null  int64
 11  PM2.5 AQI Category  16695 non-null  object
 12  lat                 16695 non-null  float64
 13  lng                 16695 non-null  float64
dtypes: float64(2), int64(5), object(7)
memory usage: 1.8+ MB
```

Step 2a: Null Value in Country Column

In this dataset, the Country column is a crucial feature for understanding the geographic location of the air quality measurements. Since the country is a key variable for grouping and analyzing the data based on regional patterns, missing values in this column could lead to incomplete or skewed analysis.

Minimal Impact on the Dataset: With 13,956 total records, removing 300 rows represents a small fraction (around 2%) of the dataset. This means that the loss of these rows has a minimal impact on the overall size and integrity of the data, ensuring that the remaining data is high quality and can be effectively used for Analysis.

```
Countries_AQI_data=Countries_AQI.dropna(subset=["Country"])
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 16393 entries, 0 to 16694
Data columns (total 14 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   Country             16393 non-null  object 
 1   City                16393 non-null  object 
 2   AQI Value           16393 non-null  int64  
 3   AQI Category        16393 non-null  object 
 4   CO AQI Value        16393 non-null  int64  
 5   CO AQI Category     16393 non-null  object 
 6   Ozone AQI Value     16393 non-null  int64  
 7   Ozone AQI Category  16393 non-null  object 
 8   NO2 AQI Value       16393 non-null  int64  
 9   NO2 AQI Category    16393 non-null  object 
 10  PM2.5 AQI Value     16393 non-null  int64  
 11  PM2.5 AQI Category  16393 non-null  object 
 12  lat                 16393 non-null  float64
 13  lng                 16393 non-null  float64
dtypes: float64(2), int64(5), object(7)
memory usage: 1.9+ MB
```

Step 2b: Duplicate Data Based On Country and City

Ensuring Unique Data Entries: The **Country** and **City** columns together represent a unique geographical location where air quality data is collected. Removing these duplicates ensures that each unique city-country combination is represented by a single entry.

```
# Drop duplicate cities
duplicate_AQI_data = Countries_AQI_data[Countries_AQI_data.duplicated(subset=['City','Country'])]
duplicate_AQI_data
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 13956 entries, 0 to 16694
Data columns (total 14 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Country             13956 non-null  object
 1   City                13956 non-null  object
 2   AQI Value           13956 non-null  int64
 3   AQI Category        13956 non-null  object
 4   CO AQI Value        13956 non-null  int64
 5   CO AQI Category     13956 non-null  object
 6   Ozone AQI Value     13956 non-null  int64
 7   Ozone AQI Category  13956 non-null  object
 8   NO2 AQI Value       13956 non-null  int64
 9   NO2 AQI Category    13956 non-null  object
 10  PM2.5 AQI Value     13956 non-null  int64
 11  PM2.5 AQI Category  13956 non-null  object
 12  lat                 13956 non-null  float64
 13  lng                 13956 non-null  float64
dtypes: float64(2), int64(5), object(7)
memory usage: 1.6+ MB
```

Step 2c: Renaming Columns

```
(13956, 14)
<class 'pandas.core.frame.DataFrame'>
Index: 13956 entries, 0 to 16694
Data columns (total 14 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Country               13956 non-null  object
 1   City                  13956 non-null  object
 2   Overall_AQI_Value     13956 non-null  int64
 3   Overall_AQI_Category  13956 non-null  object
 4   CO_AQI_Value          13956 non-null  int64
 5   CO_AQI_Category       13956 non-null  object
 6   Ozone_AQI_Value       13956 non-null  int64
 7   Ozone_AQI_Category    13956 non-null  object
 8   NO2_AQI_Value         13956 non-null  int64
 9   NO2_AQI_Category      13956 non-null  object
 10  PM2.5_AQI_Value       13956 non-null  int64
 11  PM2.5_AQI_Category    13956 non-null  object
 12  Latitude              13956 non-null  float64
 13  Longitude             13956 non-null  float64
dtypes: float64(2), int64(5), object(7)
memory usage: 1.6+ MB
```

Step 2d: Feature Engineering

The dataset is merged with a country-to-continent csv file which will bring the continent column and with important continent information, the scope and quality of our analysis is improved and ensured that the data is well-organized for both global and regional studies.

```
cities_continents_merged = AQI_data_cleaned.merge(continent_mapping_data, on='Country', how='left')
cities_continents_merged.info()
```

```
#   Column                Non-Null Count  Dtype
---  ------                --------------  -----
0   Country               13956 non-null  object
1   City                  13956 non-null  object
2   Overall_AQI_Value     13956 non-null  int64
3   Overall_AQI_Category  13956 non-null  object
4   CO_AQI_Value          13956 non-null  int64
5   CO_AQI_Category       13956 non-null  object
6   Ozone_AQI_Value       13956 non-null  int64
7   Ozone_AQI_Category    13956 non-null  object
8   NO2_AQI_Value         13956 non-null  int64
9   NO2_AQI_Category      13956 non-null  object
10  PM2.5_AQI_Value       13956 non-null  int64
11  PM2.5_AQI_Category    13956 non-null  object
12  Latitude              13956 non-null  float64
13  Longitude             13956 non-null  float64
14  Continent             13954 non-null  object
```

- # Data Visualization

Question 1: What are the effects on the Air Quality Index (AQI) value by the pollutants?

This question aims to explore how the concentrations of different pollutants (such as CO, Ozone, NO_2, and PM2.5) affect the overall AQI value. The objective is to understand the relative contribution of each pollutant to the overall air quality index and identify which pollutants have the greatest influence on AQI values across various regions.

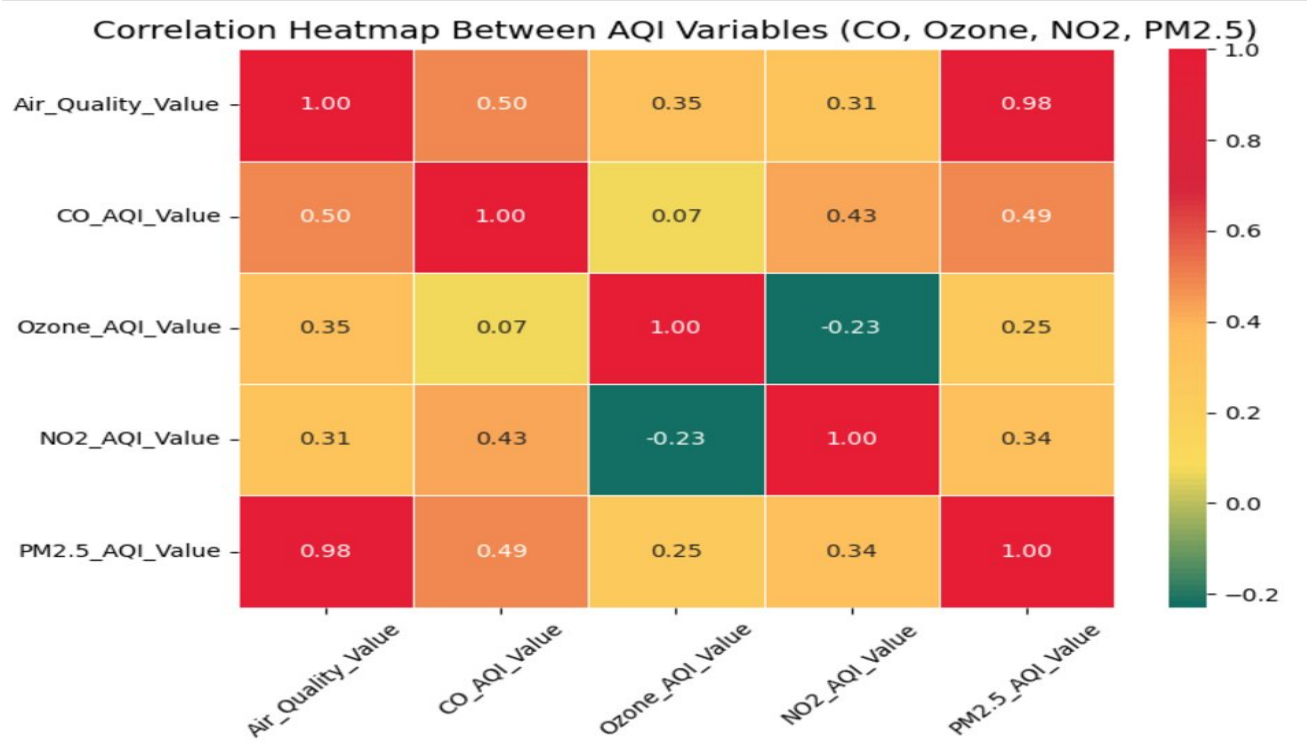
In order to answer this question we first took a look at the dataframe of our cleaned dataset. A new table was then formed to have only the AQI values and the values of the pollutant columns. From this new dataframe we could find the correlation of the pollutants to the AQI Value using the correlation matrix equation in Pandas.

```
#Create Correlation Matrix
corr_matrix = HM_df.corr()
print(corr_matrix)
```

```
                  Air_Quality_Value  CO_AQI_Value  Ozone_AQI_Value  \
Air_Quality_Value           1.000000      0.495144         0.348281
CO_AQI_Value                0.495144      1.000000         0.074985
Ozone_AQI_Value             0.348281      0.074985         1.000000
NO2_AQI_Value               0.308366      0.433509        -0.229369
PM2.5_AQI_Value             0.979874      0.493940         0.252082

                  NO2_AQI_Value  PM2.5_AQI_Value
Air_Quality_Value      0.308366         0.979874
CO_AQI_Value           0.433509         0.493940
Ozone_AQI_Value       -0.229369         0.252082
NO2_AQI_Value          1.000000         0.339855
PM2.5_AQI_Value        0.339855         1.000000
```

Table 1: Correlation Matrix for Pollutants (CO, O3, NO2, PM2.5) on AQI Value



From the correlation matrix results, a heat map (Figure 1) was able to be plotted using Seaborn to give a visualization of the findings. The results of the heat mapped identified that the pollutant of PM2.5 or Particulate matter to have the greatest correlation to Air Quality Index Value. With a value of 1 being the highest correlation, a value of 0.98, we find that PM2.5 has the greatest effect on AQI.

The result of having a high correlation between AQI Value and PM2.5 we wanted to further our findings by using a scatter plot to display similar results. Taking only the columns AQI Value and PM2.5 Value, a new data frame was formed(Table 2).

```
#Scatter plot
PM_df = cities_continents_merged[['Air_Quality_Value', 'PM2.5_AQI_Value']]
print(PM_df)
```

```
       Air_Quality_Value  PM2.5_AQI_Value
0                     51               51
1                     41               41
2                     66               66
3                     34               20
4                     54               54
...                  ...              ...
13951                160               79
13952                 54               54
13953                 71               71
13954                 50               50
13955                 71               71

[13956 rows x 2 columns]
```

Table 2: Data Frame (PM_df) for AQI Value and PM2.5 Value for all cities in CSV

These points were then plotted on a scatterplot using the Matplotlib visualization tool. From the figure(2) below, we find a strong positive correlation between AQI Value and PM2.5. However, as a group we wanted to show a line of best fit for the scatterplot. To do so we added a regression model to the scatter plot to form Figure 3.

```
sns.scatterplot(x='Air_Quality_Value', y='PM2.5_AQI_Value', data=PM_df, color='yellow',edgecolor='black', s=100)
plt.title('AQI Value vs PM2.5 Value')
plt.xlabel('Air Quality Value')
plt.ylabel('PM2.5 AQI Value')
plt.grid(True, linestyle='--', color='gray', alpha=0.5)
plt.show()
```

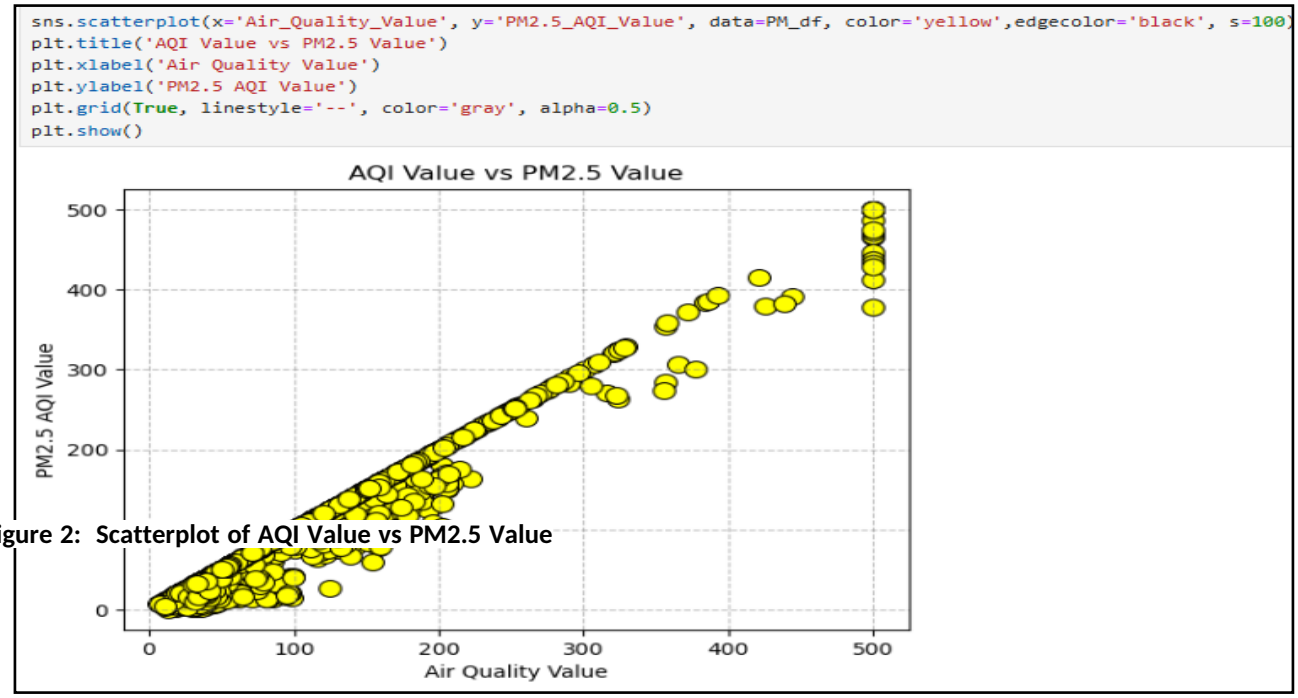


Figure 2: Scatterplot of AQI Value vs PM2.5 Value

```
# Fit the regression model
coefficients = np.polyfit(PM_df['Air_Quality_Value'], PM_df['PM2.5_AQI_Value'], 1)  # Linear regression (degree=1)
slope, intercept = coefficients

# Scatterplot with regression line
sns.regplot(x='Air_Quality_Value', y='PM2.5_AQI_Value', data=PM_df, scatter_kws={'s': 100, 'color':'yellow','edgecolor': 'black'}, line_kws={'color':

# Add the regression equation as text
plt.text(1, 500, f"Y = {slope:.2f}X + {intercept:.2f}", fontsize=12, color='red')


plt.title('AQI Value Vs PM2.5 Value')
plt.xlabel('Air Quality Value')
plt.ylabel('PM2.5 AQI Value')
plt.grid(True, linestyle='--', color='gray', alpha=0.5)
plt.show()
```

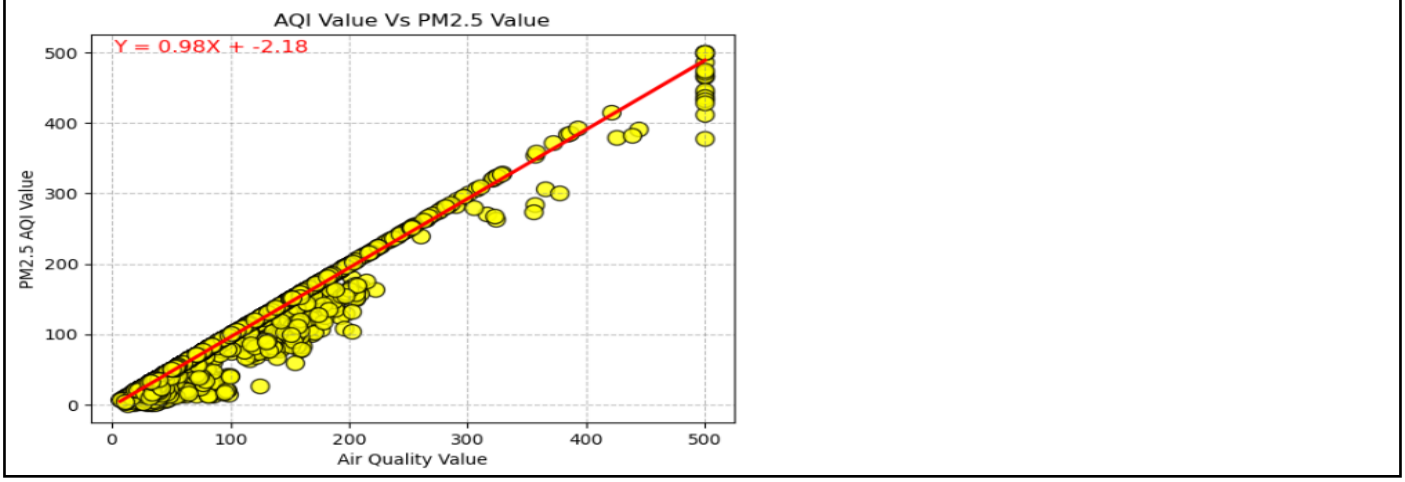


Figure 3: AQI Value Vs PM2.5 Value with a regression model showing line of best fit (Y= 0.98x + -2.18)

In Summary, our question is: What are the effects of AQI value by pollutants? We found that the pollutant PM2.5, better known as particulate matter, has the greatest impact on the AQI Value. With a correlation coefficient of 0.98 it relates more to the AQI Value than the other pollutants (CO, NO2, O3). We found again through the scatterplot that as Air quality value goes up the particulate matter rises making a positive correlation.

Question 2: How are the four key pollutants (CO, Ozone, NO_2, and PM2.5) distributed across the globe?

1. Distribution of Pollutants

To better understand the global distribution of air pollutants and identify trends or anomalies, we utilized **pie charts** to visualize the percentage distribution of each pollutant across different regions. The main objective was to explore potential correlations, understand the relative contribution of each pollutant to global air quality, and investigate how pollutants vary across regions.

The following leaderboard illustrates the distribution of different AQI categories across four major pollutants: **CO (Carbon Monoxide)**, **Ozone**, **NO2 (Nitrogen Dioxide)**, and **PM2.5 (Particulate Matter)**. Each pollutant is classified into six primary AQI categories, which reflect the potential impact on health and the environment.

	CO_AQI_Category	Ozone_AQI_Category	NO2_AQI_Category	PM2.5_AQI_Category
Good	13954	12899	13947	6689
Moderate	1	730	9	5648
Unhealthy	0	146	0	728
Unhealthy for Sensitive Groups	1	157	0	746
Very Unhealthy	0	24	0	102
Hazardous	0	0	0	43

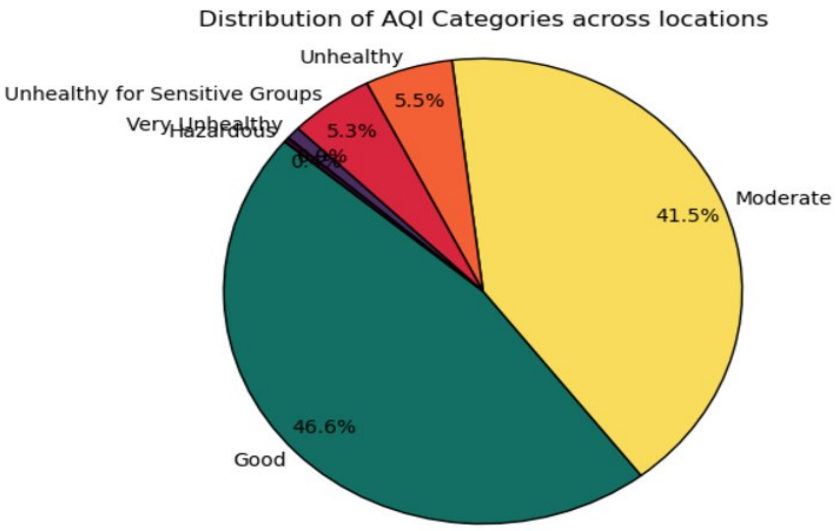
CO and **NO2** levels are generally well within safe limits for the majority of the data, with very few instances exceeding the **Moderate** threshold.

Ozone levels show some variability, with a small number of instances in the **Unhealthy for Sensitive Groups** and **Unhealthy** categories.

PM2.5, however, shows a more concerning pattern with a significant number of instances in the **Moderate**, **Unhealthy**, and **Unhealthy for Sensitive Groups** categories. The presence of instances

in the **Very Unhealthy** and **Hazardous** categories emphasizes the importance of monitoring and controlling particulate pollution to protect public health.

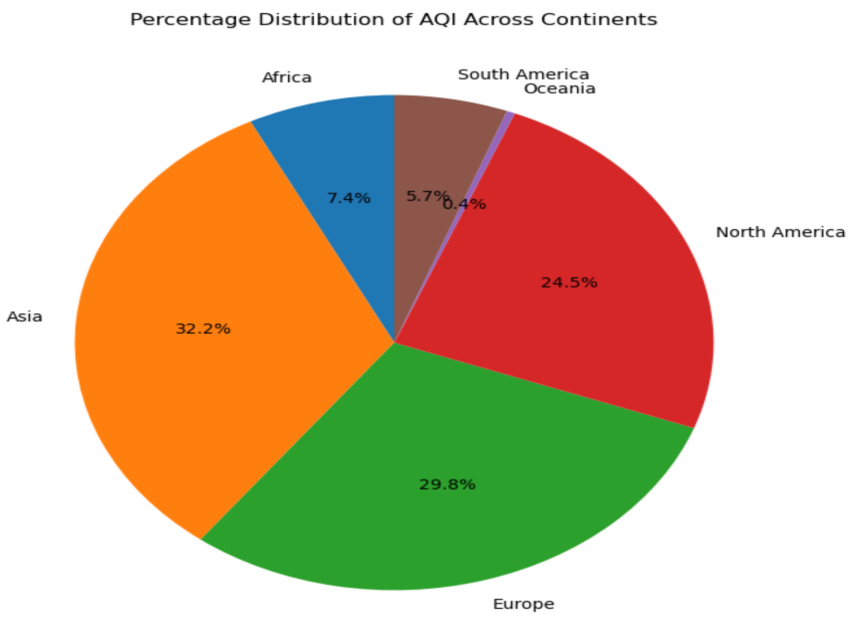
2. Pie Chart showing the Distribution of AQI Categories across geographical locations



88.02% of the cities or countries in this dataset indicates that the majority of locations have relatively acceptable air quality(good and moderate), whereas only 5.5% of the locations have Unhealthy air quality, and 5.27% fall under Unhealthy for Sensitive Groups.

Very Unhealthy and Hazardous categories are extremely rare, contributing only 1.21% of the total. This data can also be useful for studies related to public health, as areas in the Unhealthy" or "Hazardous" categories may need more focus due to the potential health risks associated with poor air quality.

3. Pollutant Distribution Across Continents: This chart helped us examine how pollutants are distributed geographically, giving us insights into which continents are most and least affected by specific pollutants.



The top 20 least and most polluted countries based on AQI value.

Most Polluted	493	Mahendragarh	India	500
	1395	Phalodi	India	500
	11730	Ratangarh	India	500
	7645	Jalalabad	India	500
	6236	Maur	India	500
	10846	Boksburg	South Africa	500
	13268	Tynda	Russia	500
	4686	Etah	India	500
	3111	Delhi	India	500
	3122	Durango	United States	500
	12294	Pokaran	India	500
	3190	Nohar	India	500
	381	Harunabad	Pakistan	500
	180	Bahawalnagar	Pakistan	500
	4928	Sardulgarh	India	500
	8852	Hasanpur	India	500
	4207	Nawalgarh	India	500
	6469	Dhanaura	India	500
	1523	Jodhpur	India	500
	4304	Rohtak	India	500

Least Polluted	5507	Macas	Ecuador	7
	13839	Tari	Papua New Guinea	8
	8710	Azogues	Ecuador	8
	4520	Huaraz	Peru	9
	4707	Huancavelica	Peru	10
	5740	Manokwari	Indonesia	10
	2163	Nueva Loja	Ecuador	10
	8294	Andradina	Brazil	11
	5692	Mendi	Papua New Guinea	11
	12558	Nazca	Peru	11
	3345	Comodoro Rivadavia	Argentina	11
	13014	Mount Hagen	Papua New Guinea	11
	13885	La Rioja	Argentina	11
	6118	Puerto Madryn	Argentina	11
	10163	Huamachuco	Peru	11
	13223	Correntina	Brazil	11
	11780	Young	Uruguay	11
	2978	Puquio	Peru	11
	10051	Uyuni	Bolivia (Plurinational State of)	12
	12597	General Roca	Argentina	12

4. Key Findings

- **Continent-based Pollution Levels**: The charts also highlighted the continents with the highest and lowest pollutant concentrations. For instance, **Asia** was found to have the highest levels of **PM2.5**, while **Europe** and **North America** had relatively lower levels, particularly in countries with stringent environmental regulations.

- **PM2.5 Dominance**: The analysis revealed that **PM2.5** is the most prevalent pollutant worldwide, indicating that fine particulate matter significantly impacts global air quality. This insight aligns with the increasing focus on the health risks associated with PM2.5 exposure.

5. Outliers and Unexplained Trends

During this analysis, we encountered certain **outliers** that contradicted the overall trends observed in the data. These outliers, which appeared in some regions, suggested extreme pollution levels that were not reflected in the general averages. For example, some cities in highly industrialized or rapidly urbanizing countries exhibited pollution levels far higher than those in neighboring regions. These discrepancies are particularly interesting but remain unexplored due to limited time and data constraints.

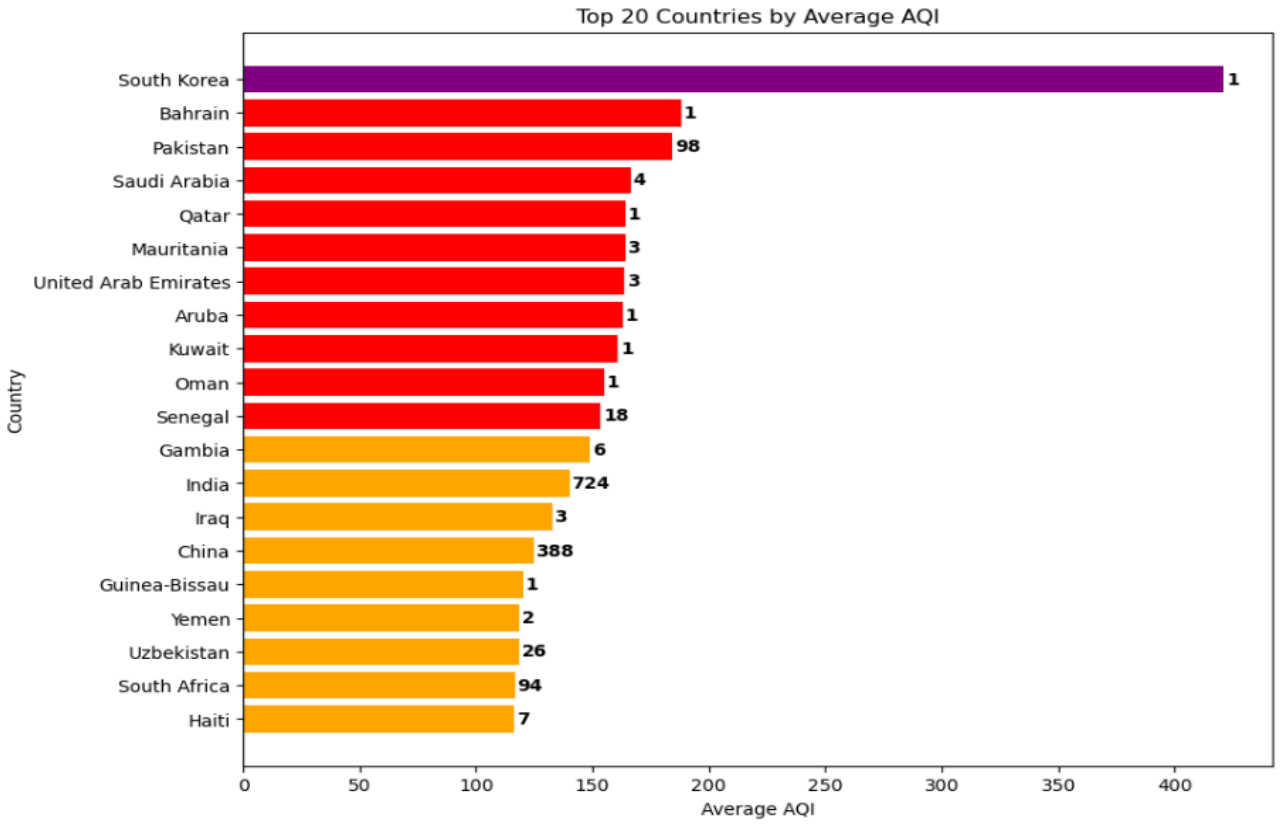
The discovery of these **outliers** points to the complexity of global pollution dynamics. Factors such as local industrial activities, transportation patterns, and geographical features (e.g., valleys trapping pollutants) could be contributing to the anomalies observed.

Question 3 : Is there a relationship between geographical location (continent, country, city) and AQI values across different regions?

This analysis aims to explore whether geographical location—such as continent, country, or city—has a significant relationship with AQI values across different regions. By understanding this relationship, we can gain insights into global air quality patterns and identify regions that may require stricter air quality management policies.

1. Bar Plot to show top 20 countries with Number of Cities Contributing to the AQI Value:

In this bar graph, we are trying to find if there are patterns in air quality that are influenced by geographical factors such as the continent, country, or city.



Each bar is annotated with the number of cities in each country contributing to that country's average AQI value. The number of cities provides context about how representative the average AQI value is for that country. For example, a country with a high AQI but few cities may not have nationwide

pollution problems like South Korea with only one city with poor air quality while Countries like China and India indicate significant air quality challenges for a large proportion of their population. Larger countries with higher population densities may naturally have higher pollution levels.

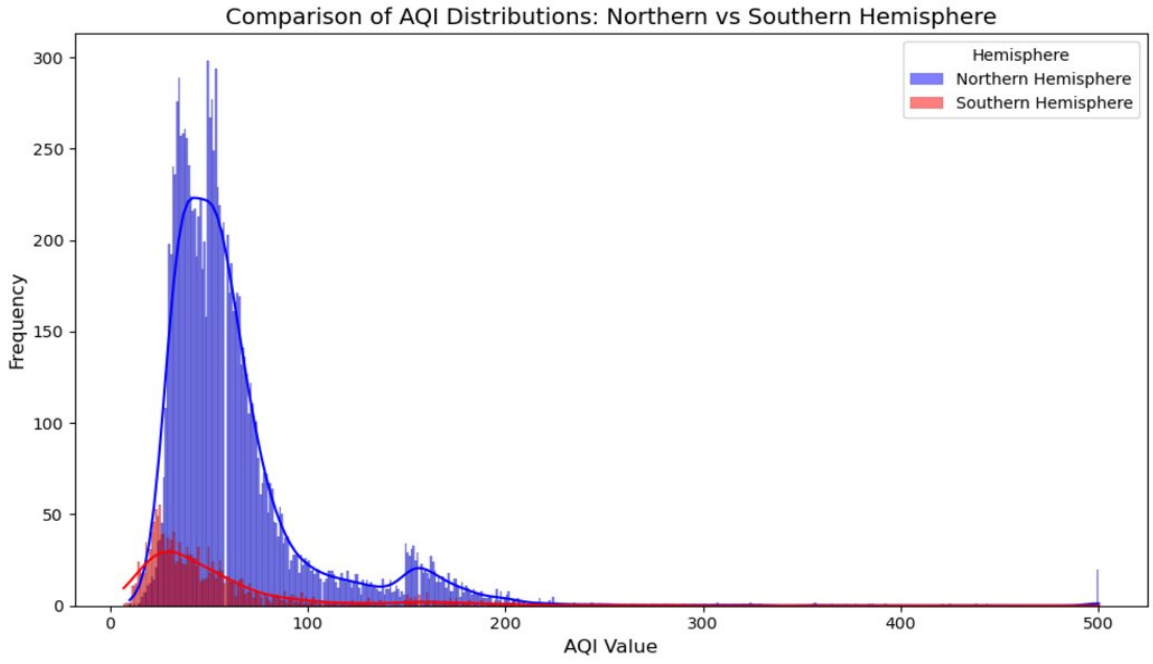
Limitations:

The bar chart only shows average AQI values, but it does not provide information on the time period over which these averages were calculated. Air quality can vary significantly over time due to seasonal changes, weather conditions, or specific events (e.g., wildfires, industrial shutdowns). Without time-based data, it's difficult to assess whether the air quality is improving or deteriorating.

2. Air quality variation in Northern and Southern Hemispheres

While the air quality in the Northern Hemisphere and Southern hemisphere is generally categorized as Good or Moderate, there is still a considerable portion of the population affected by Unhealthy air quality, particularly in certain regions or during specific times of the year. The presence of extreme air quality conditions, though limited, indicates areas of concern that may require attention for public health and environmental policy.

Given the heavy concentration of the data in the Good and Moderate categories, this shows data is right skewed.



Northern Hemisphere: The positive skew suggests that while most of the data is favorable, there are still enough poor air quality regions that they need attention.

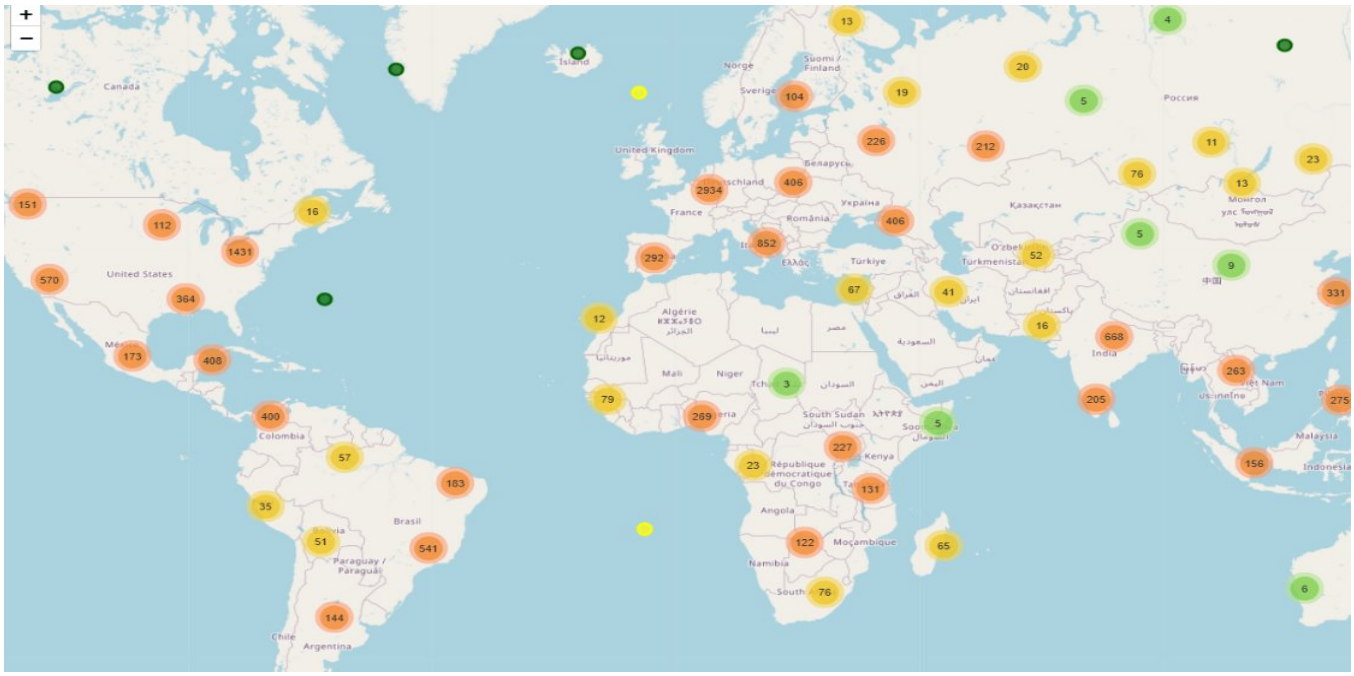
Southern Hemisphere: The positive skew suggests that air quality in the Southern

	Air_Quality_Category	Count	Hemisphere	Percentage
0	Moderate	5356	Northern Hemisphere	43.948470
1	Good	5356	Northern Hemisphere	43.948470
2	Unhealthy	676	Northern Hemisphere	5.546894
3	Unhealthy for Sensitive Groups	667	Northern Hemisphere	5.473045
4	Very Unhealthy	88	Northern Hemisphere	0.722081
5	Hazardous	44	Northern Hemisphere	0.361040

	Air_Quality_Category	Count	Hemisphere	Percentage
0	Good	1141	Southern Hemisphere	64.499717
1	Moderate	432	Southern Hemisphere	24.420577
2	Unhealthy	91	Southern Hemisphere	5.144149
3	Unhealthy for Sensitive Groups	68	Southern Hemisphere	3.843980
4	Very Unhealthy	32	Southern Hemisphere	1.808932
5	Hazardous	5	Southern Hemisphere	0.282646

Hemisphere is predominantly good, with very few regions experiencing poor air quality. The outliers are minimal and should not overly influence summary statistics, but they should still be considered when evaluating air quality challenges in specific locations.

3. Exploring AQI Patterns Based on Geographic Locations



The goal of this interactive map is to allow users to visualize the AQI across different regions , giving a spatial representation of air quality. This makes it easy to pinpoint areas with poor air quality and compare them geographically.

On this interactive map, **Folium** is used to plot cities from across the world. Each city is represented by a **circle marker,** whose color corresponds to its air quality category.

Folium has some cool features like **Markercluster** which groups nearby points into clusters.

In our dataset, there are too many cities across the world so plotting it would make it all cluttered.

Popups are also another cool feature which gives detailed description when the user clicks on a circle marker.

4. Leaderboard showing quartiles and outliers for top 10 countries

	Country	Q1	Q2	Q3	IQR	Lower Bound	Upper Bound	Outliers
165	United States	46.00	55.0	67.00	21.00	14.500	98.500	157.0
22	Brazil	25.00	36.0	52.00	27.00	-15.500	92.500	48.0
128	Russia	34.00	39.0	47.00	13.00	14.500	66.500	29.0
68	India	77.75	150.0	170.00	92.25	-60.625	308.375	28.0
77	Japan	41.00	47.0	59.00	18.00	14.000	86.000	26.0
109	Netherlands	34.00	37.0	39.75	5.75	25.375	48.375	25.0
123	Philippines	44.50	54.0	70.50	26.00	5.500	109.500	17.0
74	Italy	48.00	59.0	71.00	23.00	13.500	105.500	16.0
53	France	38.00	50.0	61.00	23.00	3.500	95.500	16.0
116	Pakistan	159.00	173.0	187.75	28.75	115.875	230.875	16.0

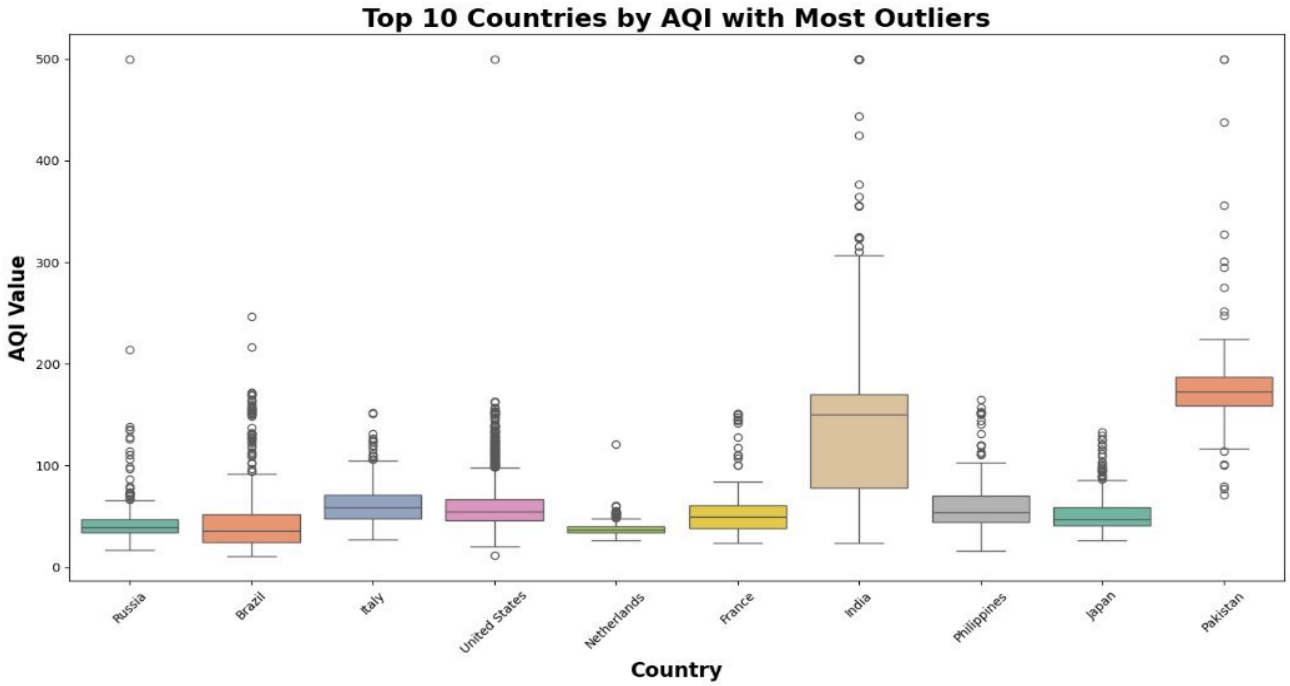
The **Leaderboard of AQI quartiles and outliers** helps us understand how **geographical location (continent, country, city)** relates to AQI values. Countries in **Asia**, especially **India** and **Pakistan**,

show higher AQI values and greater variability, highlighting the impact of **urbanization**, **industrialization**, and **population density** on air quality.

On the other hand, **European countries** like **Netherlands** and **France** display better air quality with fewer extreme values, suggesting that **environmental regulations**, **lower population density**, and **industrial activity control** contribute to lower pollution levels.

This analysis provides a clearer picture of how **geographical factors** play a role in shaping the air quality, suggesting that **location, urbanization**, and **industrialization** are key factors influencing AQI values across different regions.

5. Box Plot for the top 10 countries with most outliers for AQI value



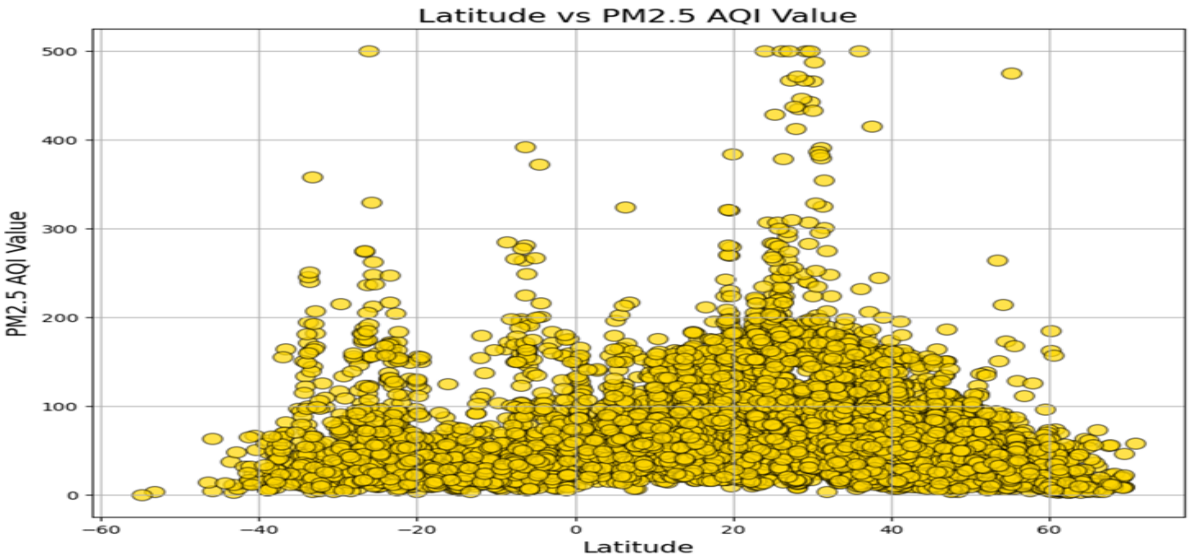
Some countries have a significant number of outliers (e.g, the United States with 157 outliers), indicating there may be some extreme air quality data points that deviate far from the general distribution. This suggests unusual air quality events or reporting anomalies in certain regions.

Countries like **India** and **Brazil** show wider IQRs, indicating greater variability in air quality levels. In contrast, countries like **Netherlands** and **France** have narrower IQRs, indicating more consistent air quality.

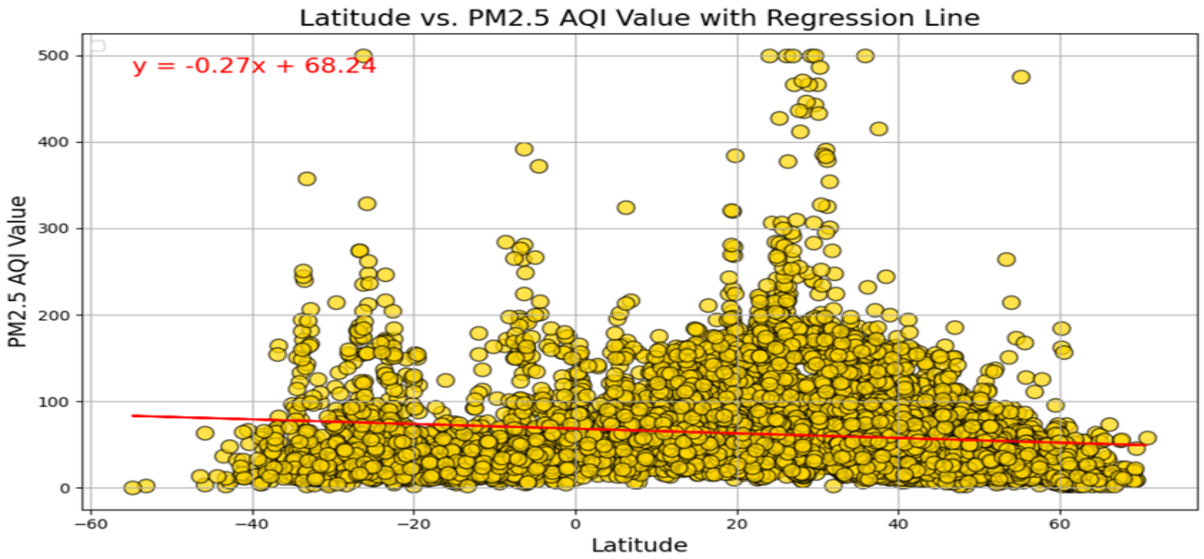
Countries like **India** and **Pakistan** show a wide range of air quality values, with lower and upper bounds that span a much broader range compared to countries like **Russia** or **Netherlands**. This could reflect more extreme pollution events or fluctuations in air quality.

6. PM2.5 Air Quality Index value vs Latitude

A scatter plot was plotted between Latitude and PM2.5 to examine the relationship between these two variables. There was no noticeable pattern between them. This suggests that latitude alone may not have any significant influence on the PM2.5 AQI values.



A linear regression analysis was done to further investigate the relationship between latitude and PM2.5 values. The correlation between Latitude and PM2.5 AQI Value is -0.14.This indicates a very weak negative correlation, so a meaningful conclusion cannot be drawn.



The equation of the regression line is y = -0.27x +68.24.Here y represents the PM2.5 AQI value and x represents latitude.So for an increase in unit for the latitude we find that PM2.5 AQI values decrease by 0.27 points.We could thus conclude that industrialization, population density, and the environmental regulations of a given region could be the deciding factors that contribute to the PM2.5 AQI values,rather than the latitude for that particular region.

7. TTest and Anova

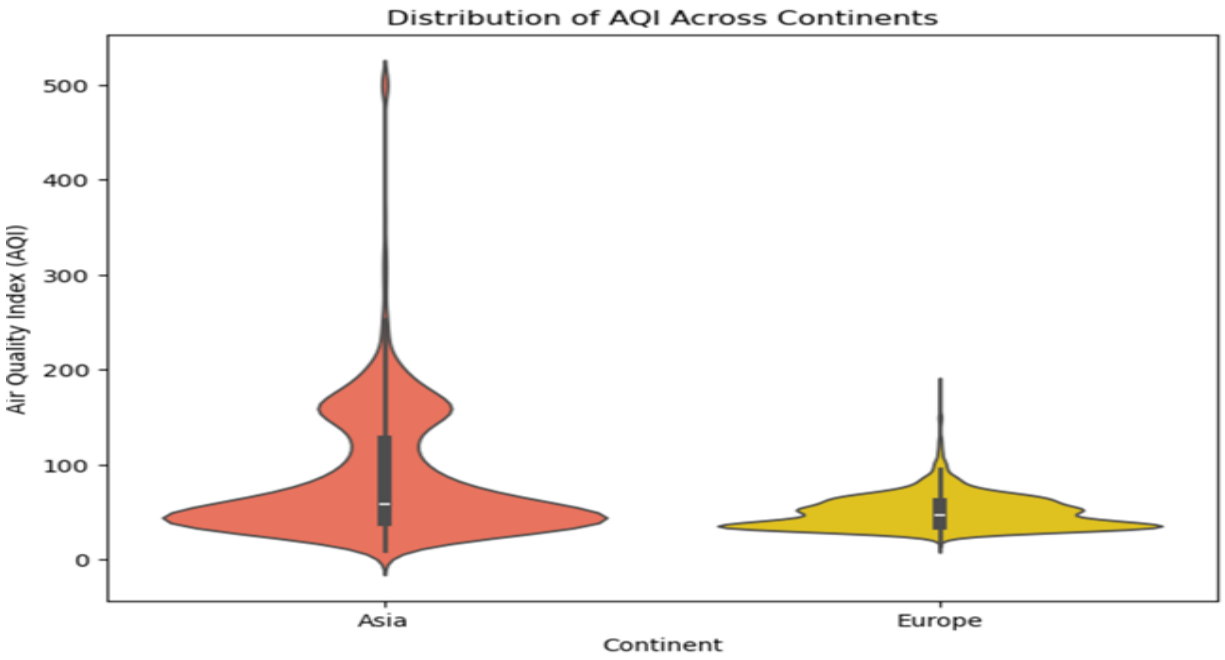
Understanding how AQI values differ across regions is crucial, as it provides insights into the air quality conditions in different parts of the world. Higher AQI values typically correspond to poorer air quality, which can have significant health, environmental, and impacts. By examining AQI values across continents, we can identify regions with more severe pollution problems and target efforts to improve air quality in those areas. A TTest was done to compare the AQI values of two continents, Europe and Asia.A T-test is a statistical test used to compare the means of two groups and determine if the difference between them is statistically significant. The Basic Hypothesis for T-test: Null Hypothesis (H_0): There is no significant difference between the AQI values in Asia and Europe are not significantly different).Alternative Hypothesis (H_1): There is a significant difference between the AQI values in Asia and Europe are significantly different

TTest Result(statistic=31.05031627218465, p value=4.15014554324163e-188, df=3661.611569440565)

The high T-statistic which is 31.05 suggests a large difference in AQI between the two continents.The low p-value of 4.15014554324163e-188 which is way less than the accepted threshold of 0.05, indicates a statistically significant difference in AQI values between the two continents. This result allows us to reject the null hypothesis which suggests there is no difference in AQI values between Asia and Europe. T-test specifically compares AQI values between Asia and Europe, providing a focused analysis of these two continents. Meanwhile, the ANOVA test gives a broader perspective by comparing AQI values across all continents, allowing us to assess the global distribution of air quality and understand how regions around the world compare to each other.
F_onewayResult(statistic=369.3766024923413, pvalue=0.0)

The F-statistic 369.38 is very high, indicating that the variability between continents is much greater than the variability within each continent. This result suggests that AQI values differ significantly across continents. The p-value of 0 further supports this conclusion, confirming that the observed differences in AQI across the continents are statistically significant.

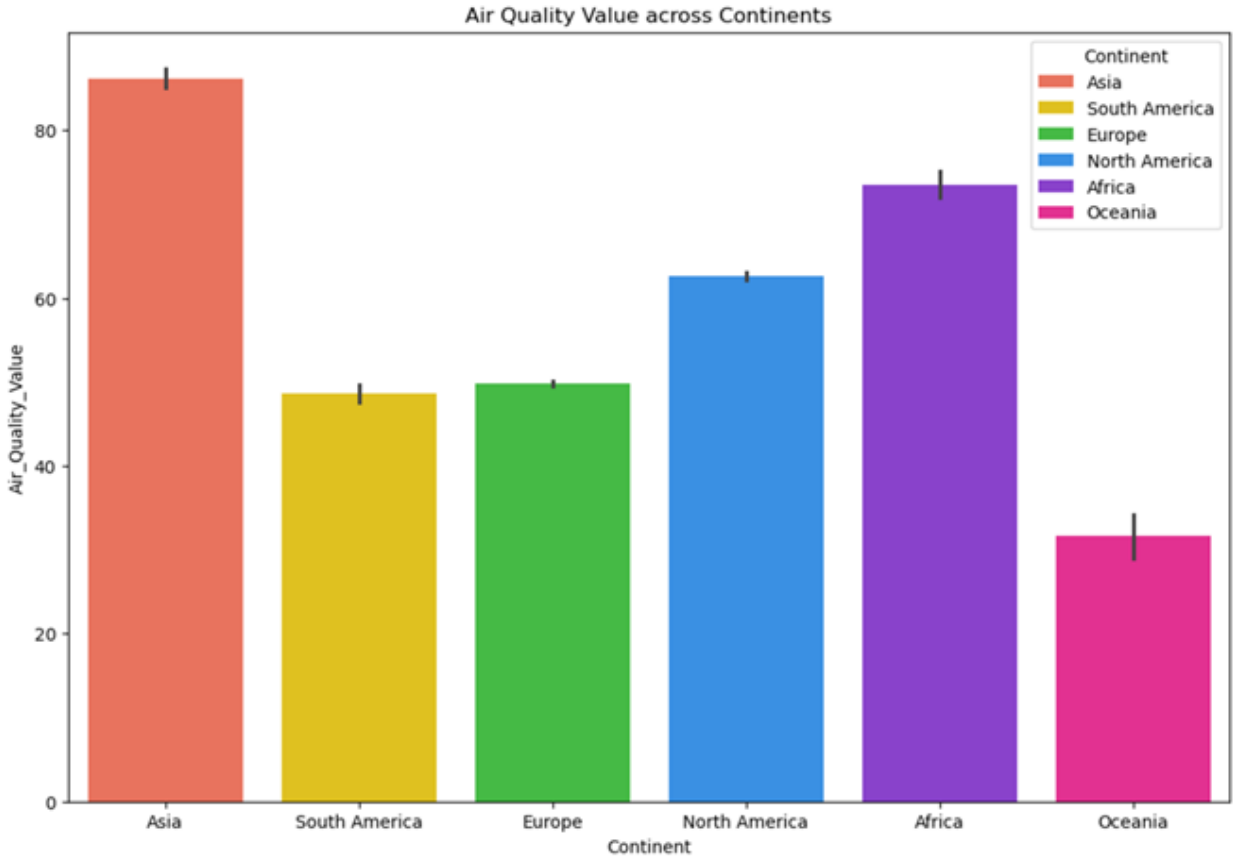
We used data visualization such as bar plots and violin plots to understand the air quality pattern. The plots highlighted which continents have got higher levels of pollution.



This is a violin plot comparing the AQI values for the continent Asia and Europe.We can see that Asia is more polluted than Europe.Also violin plots indicate a higher outlier for the continent of

Asia.Looking at the IQR of Europe which is much smaller than Asia ,we can suggest that air quality is more consistent across regions in Europe,whereas for Asia there is a greater variability for air quality index values.This could be because of environmental regulations, air quality management, or industrial activity .

Looking at barplot for all the continents of the world we can clearly see that Asia has got the worst air quality whereas Oceania has got the best air quality among the continents.



8. Key Findings

The findings support the hypothesis that geographical location has a relationship with AQI values, but this relationship is more complex than a direct connection to latitude, longitude, or even hemisphere. Several key factors influence AQI values across different regions:

- **Urbanization and Industrialization**: Countries with large, densely populated urban areas, such as China and India, tend to have poorer air quality. High industrial activity, transportation emissions, and limited green spaces contribute significantly to elevated AQI values in these regions.
- **Environmental Policies and Regulations**: Countries with stricter environmental regulations and lower industrial activity, such as those in Europe and Oceania, tend to have better air quality. This indicates the role of governmental policies in managing air quality.
- **Climate and Geography**: Climate conditions and geographical features also play a role. For example, cities in valleys or areas prone to temperature inversions often experience worse air quality due to trapped pollutants. On the other hand, cities with favorable climatic conditions for dispersing pollutants (e.g., coastal cities) tend to have better air quality.

Call to Action

The four major pollutants (PM2.5, O3, NO2, and CO) are primarily caused by the burning of fossil fuels and natural gas, with vehicle and industrial emissions being the biggest contributors.Industries must take proactive steps toward reducing emissions. Many companies are already adopting more sustainable practices to help combat this issue. As individuals, we can all play a part in reducing air pollution.We can embrace alternative transportation like using public transportation, biking, and walking.We could also think about transitioning to electric vehicle reduce personal carbon

Bias and Limitations

The dataset does not include information about air quality over time (e.g., hourly, daily, or seasonal trends). Air quality can vary throughout the day and absence of time information made it difficult to draw accurate conclusions from this dataset.

Dropping some country rows due to missing values reduced the overall sample size which meant that we had to lose some data points that could have contributed to our data visualization.

The absence of sources of pollution (industrial,vehicular) in the dataset was also a major limitation since it did not give us a complete picture.

Future Work

After finishing the project we had even more questions about air quality and how the subject is affected by our daily lives. This leads us to want to look into future projects, if given the time. Our data set did not include the pollution source; and therefore, some future work would include looking at the pollutant sources more closely i.e car and industrial emissions. We can study the emission factors and see which source plays a larger role in total air quality. Additionally, we would like to look at a dataset that includes population count. We gathered from our data that population size and urbanization plays a large role in the AQI of a region. Future work would include looking at a dataset that includes population along with the cities. This would be a key variable to add for measuring AQI Value.

Works Cited

- Ramchandran, Aditya. “World Air Quality Index by City and Coordinates- Comprehensive Dataset on Cities, Latitude, Longitude, and Pollution Levels.” Kaggle, 28 November 2024, http://www.kaggle.com/datasets/adityaaramachandran27/world-air-quality-index -by-city-and-coordinates/data.
- “Air Quality in the World Air Quality Index (AQI+) and PM2.5 air pollution in the world.” IQAir, Accessed: 6 December 2024, https://www.iqair.com/us/world-air-quality.
- Booth, Alexander. “Data Visualization.” Data Analytics Bootcamp,18 November 2024, MSU
- "Ground-Level Ozone Basics." USA.gov, 14 May 2024,
- https://www.epa.gov/ground-level-ozone-pollution/ground-level-ozone-basics.
- “How do you get rid of duplicated data?” Xpert Learning Assistant, 26 Nov. version, Open AI,
- 26 Nov. 2024, https://bootcampspot.instructure.com/courses/6994/external_tools/313