Exploring the Global Air Quality Index: A Data-Driven Analysis of Air Pollution

Presented by:
Adriana E Kuhl
Misha Mambully Muralidharan
Val Sanhueza
Neelam Prasad

November 25th, 2024

Introduction

What is AQI?

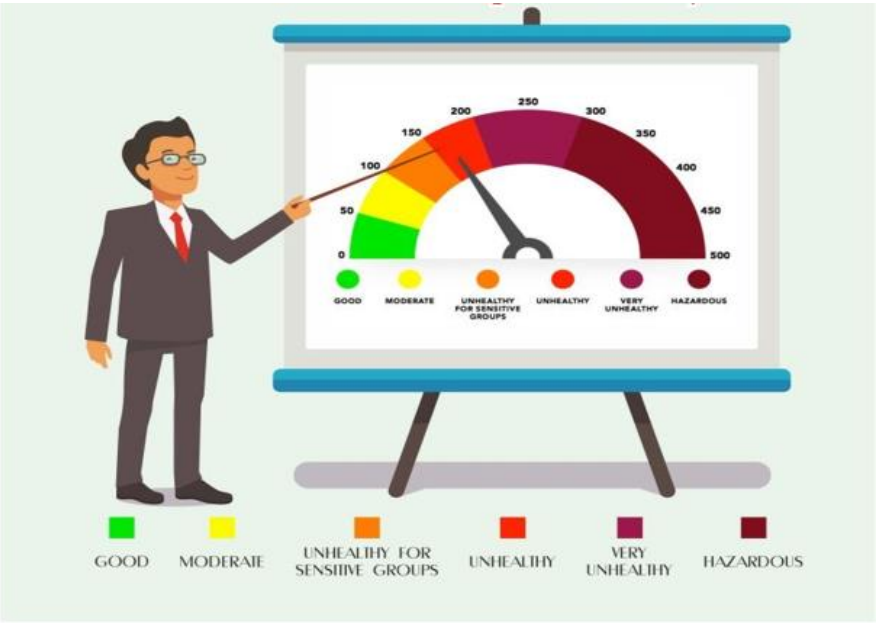
The Air Quality Index (AQI) is a tool used to measure and communicate air quality levels.

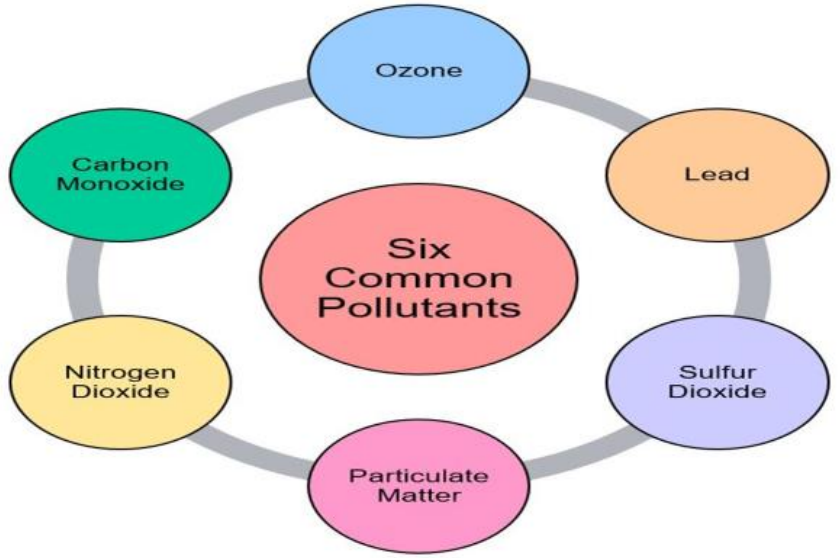
The US Environmental Protection Agency (EPA) developed an Air Quality Index designed to display the air quality level that is mapped to the concern for human health.

The AQI measures five different types of air pollutants; Ozone & Particulate Matter (PM2.5, PM10) also carbon monoxide, sulfur dioxide and nitrogen dioxide.

The worse the air quality is on a given day, the stronger the warning displayed on the Air Quality Index.

All five pollutants are measured, typically every 8 hours but on severe days measurement can be hourly.





Objective and Dataset Overview

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.

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.

Data Engineering

Exploratory Data Analysis (EDA) on the World Air Quality Index dataset.

The dataset contains information about air quality measurements in different cities around the world.

Data Cleaning

Dropped Country rows with null values.
Dropped duplicate city rows
Renamed column headers

Feature Engineering

Merged two datasets to form a new data frame to consolidate data on continent basis.

```
Data columns (total 15 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Country               13956 non-null  object
 1   City                  13956 non-null  object
 2   Air_Quality_Value     13956 non-null  int64
 3   Air_Quality_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
dtypes: float64(2), int64(5), object(8)
```

High Level Questions

Question 1

What are the effects of pollutants on the Air Quality Index Value?

Question 2

What is the distribution of all the four pollutants across the globe ?

Question 3

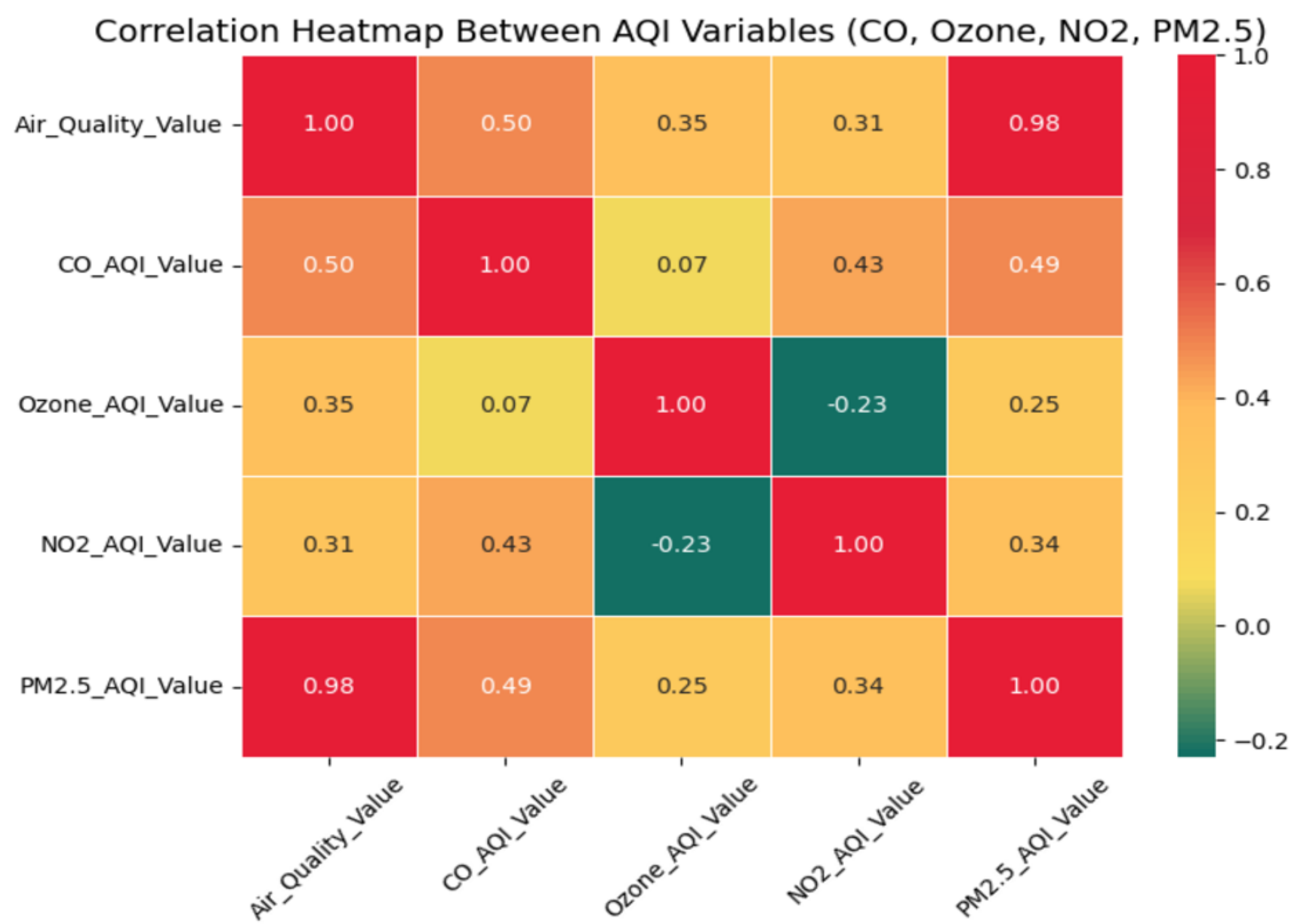
Is there a relationship between geographical location and AQI values across different regions?

Question 1 :What are the effects of pollutants on the Air Quality Index Value?

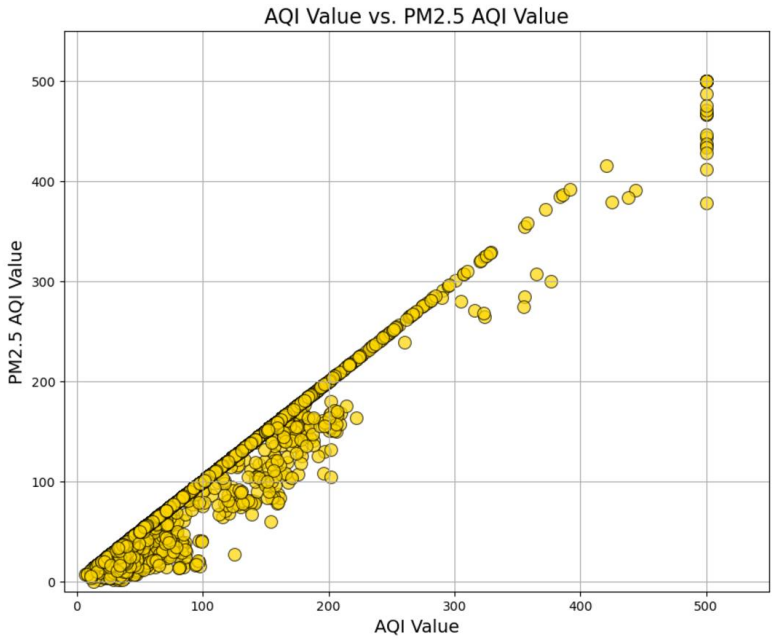
Four major pollutants in our dataset is PM2.5, O3, NO2 and CO. Each of these pollutants influences AQI values in different ways, with **PM2.5** being the most significant contributor. In this analysis, we will explore the relationship between various pollutants and AQI through various visualizations to better understand their impact on air quality.

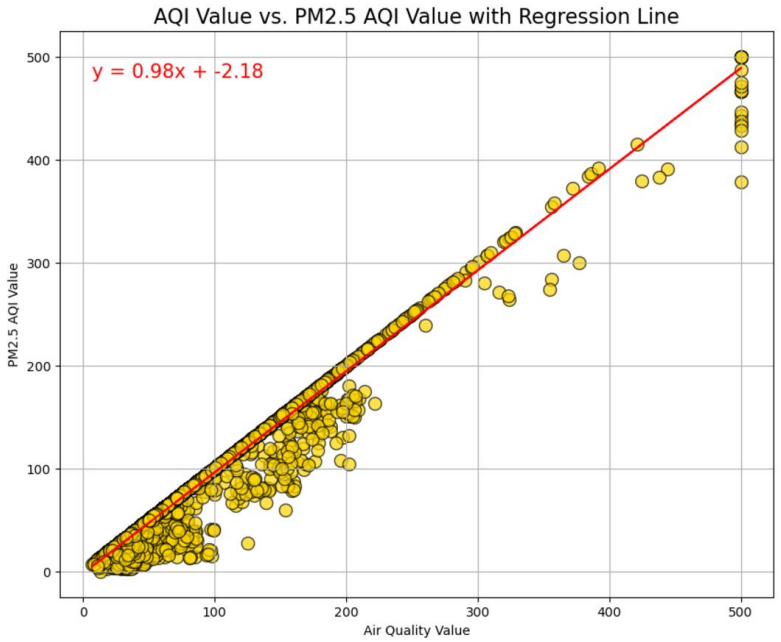
AQI Values for Key Pollutants by Country

Country	Air_Quality_Value	PM2.5_AQI_Value	CO_AQI_Value	Ozone_AQI_Value	NO2_AQI_Value
Afghanistan	259	259	1	126	0
Albania	694	689	9	383	5
Algeria	425	425	16	140	103
Andorra	32	24	1	32	0
Angola	1785	1733	71	487	47


Correlation Heatmap Between AQI Variables (CO, Ozone, NO2, PM2.5)
Air_Quality_Value
CO_AQI_Value
Ozone_AQI_Value
NO2_AQI_Value
PM2.5_AQI_Value
1.00
0.50
0.35
0.31
0.98
0.50
1.00
0.07
0.43
0.49
0.35
0.07
1.00
-0.23
0.25
0.31
0.43
-0.23
1.00
0.34
0.98
0.49
0.25
0.34
1.00
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Scatter plot and linear regression between AQI value and PM2.5 AQI value



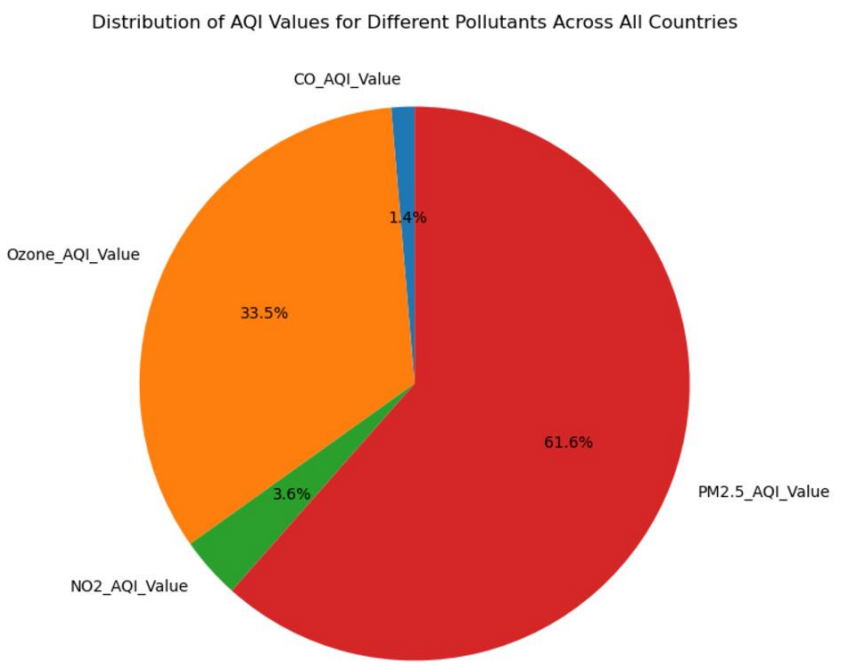


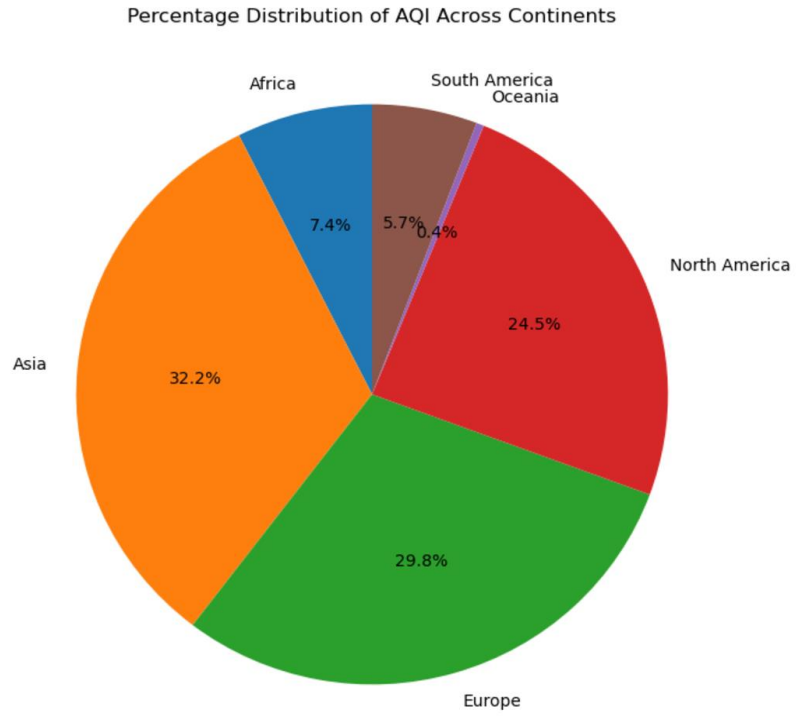
From the scatter plot between PM2.5 AQI and Air Quality Value (AQI), we observe a clear positive trend. Tracking PM2.5 levels closely can help predict AQI fluctuations, enabling timely interventions to mitigate adverse health effects. The correlation between Air Quality Value and PM2.5 AQI Value is 0.98.This value is very high which indicates a very strong positive relationship between these two variables

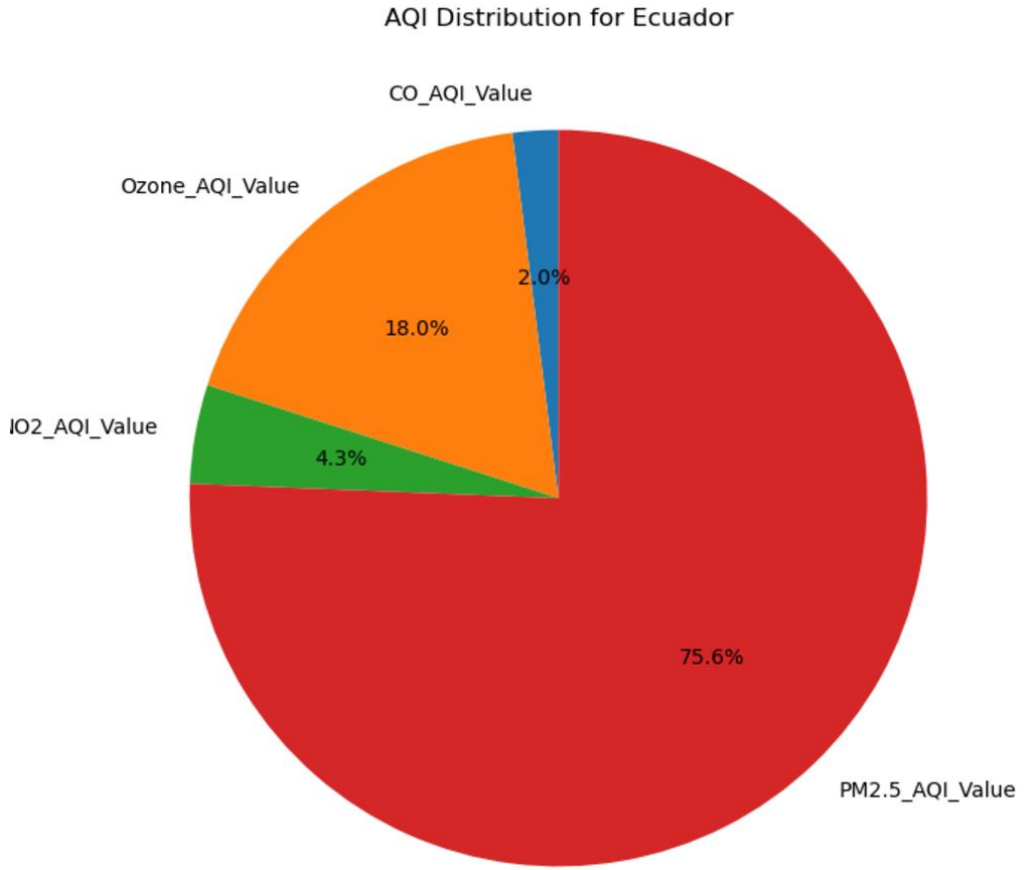
Question 2:

What is the distribution of all the four pollutants across the globe?

- The global distribution of the four major pollutants (AQI, Ozone, NO_2, and CO) shows varying proportions.
- In the top graph we can see pollutants across all countries and which may impact our world the most
- The bottom graph visualizes the distribution of all pollutants across countries specifically

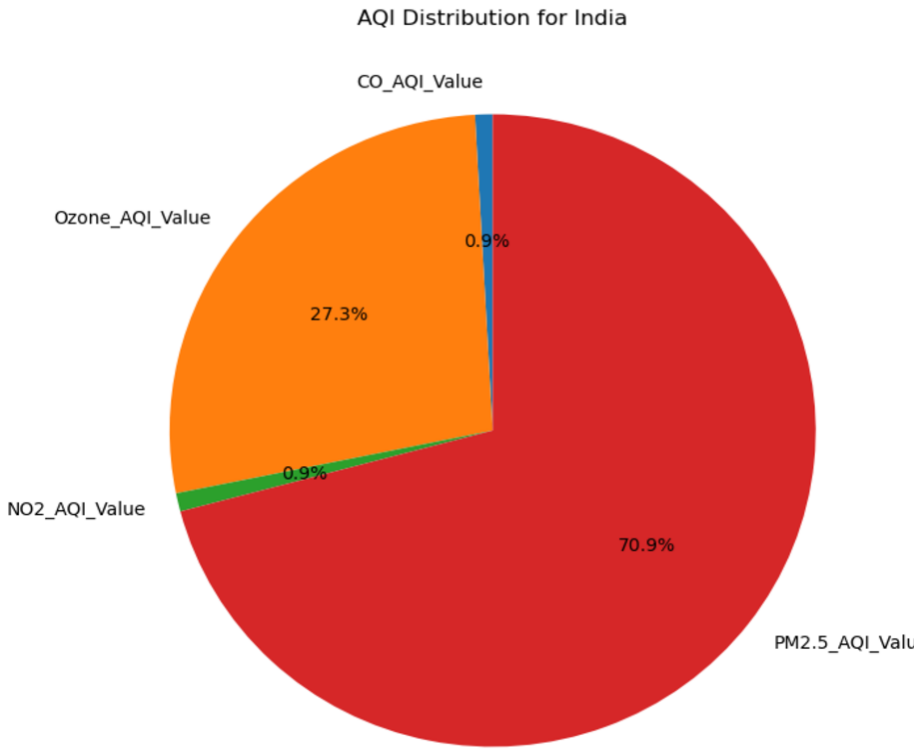




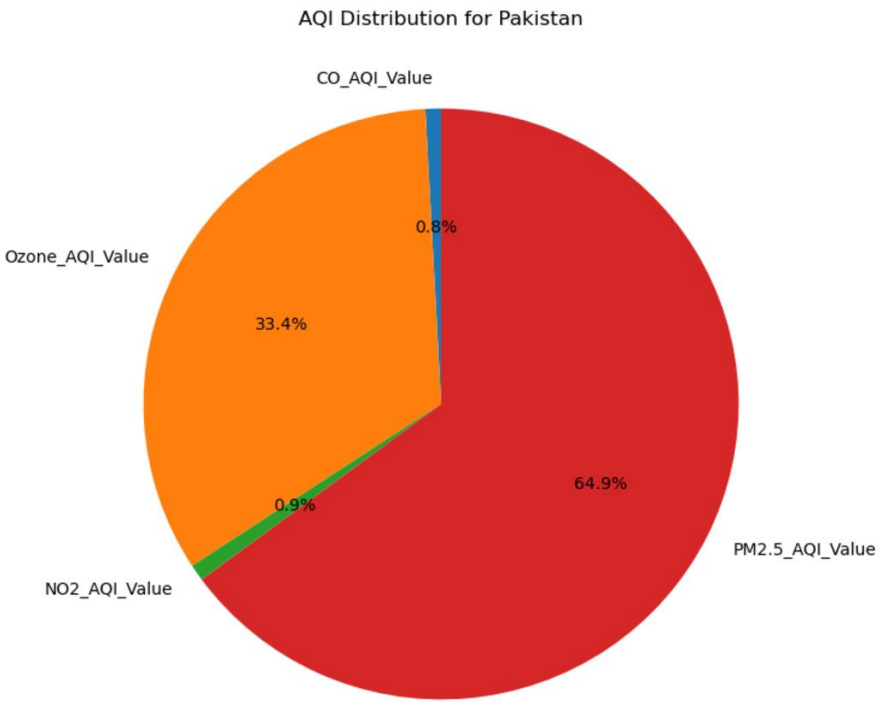


The Least Polluted Country based on AQI Value – Pie Chart

INDIA – 16 Cities

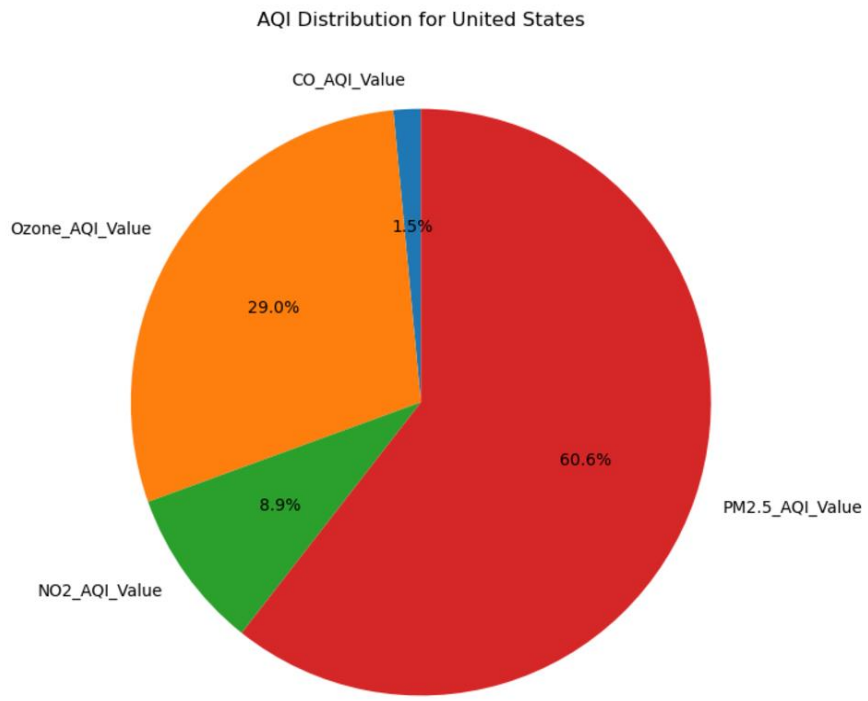


PAKISTAN – Two Cities

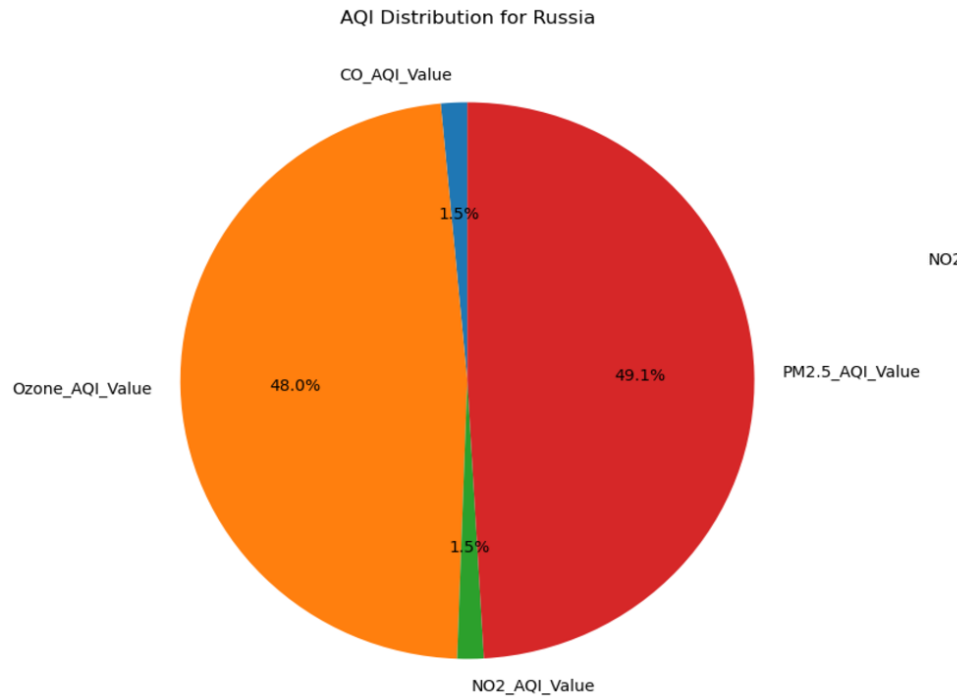


The Most Polluted Countries Based on AQI Value – Pie Chart

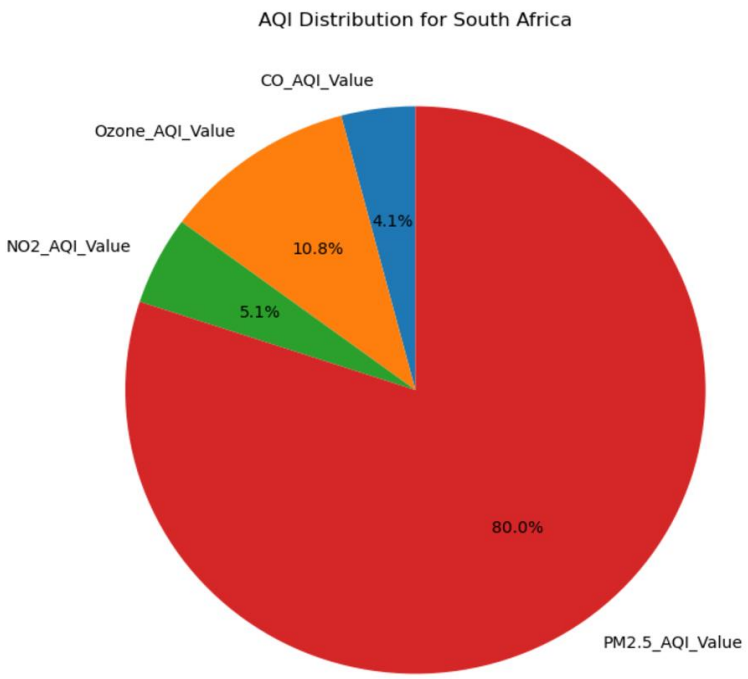
UNITED STATES – One City



RUSSIA – One City



SOUTH AFRICA – One City



The Most Polluted Countries Based on AQI Value – Pie Chart (More Possibly Outliers)

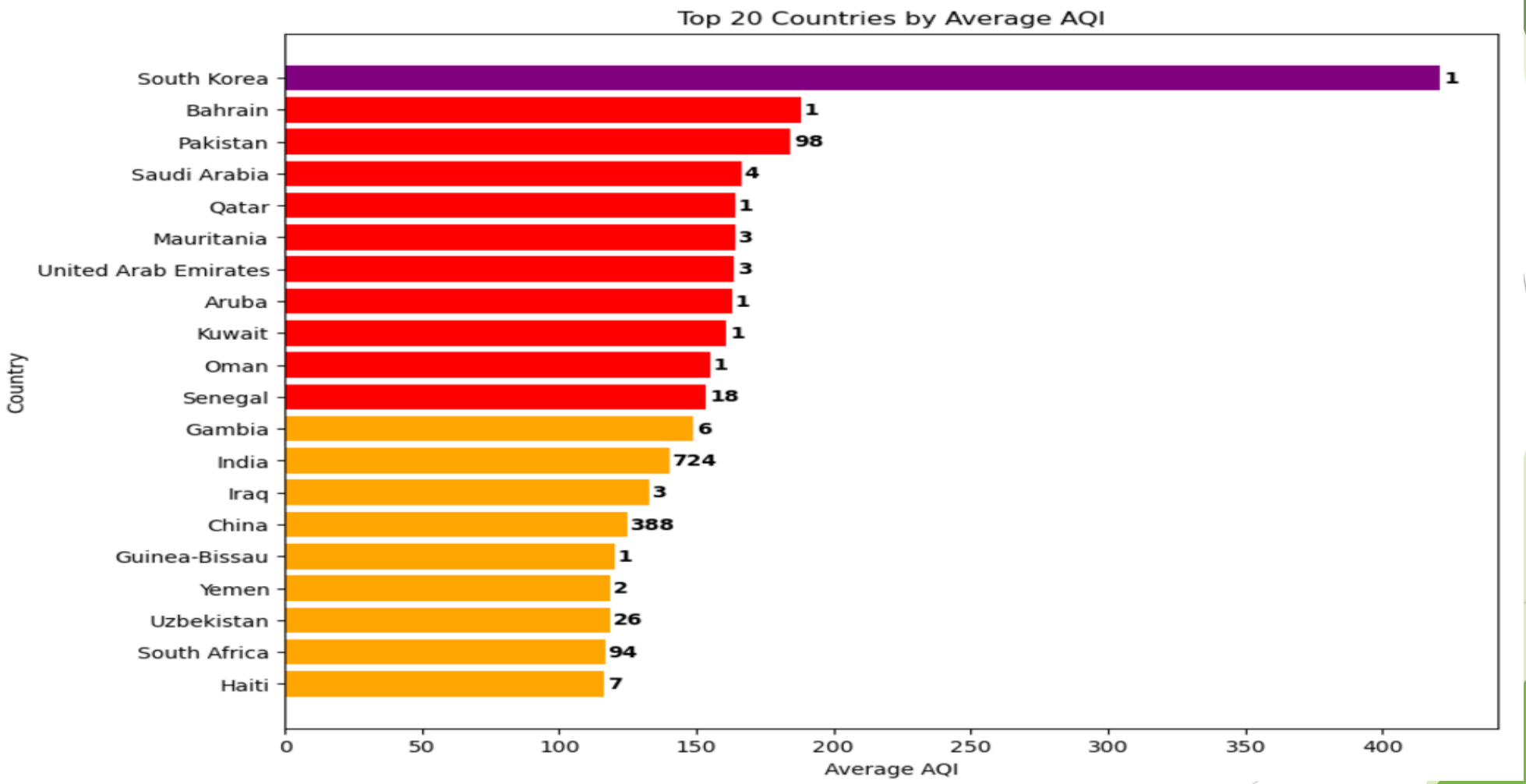
DataFrame of the Information Found in Our Pie Charts

		City	Country	Air_Quality_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

		City	Country	Air_Quality_Value
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

Question3:Is there a relationship between geographical location and AQI values across different regions?

Bar Graph representing top 20 countries with number of cities contributing to the average AQI



TOP 5

The leaderboards suggest that AQI is affected more by urbanization, with more populated cities having higher values of AQI.

	Country	Air_Quality_Value	City
143	South Korea	421.000000	1
11	Bahrain	188.000000	1
116	Pakistan	184.336735	98
132	Saudi Arabia	166.250000	4
126	Qatar	164.000000	1

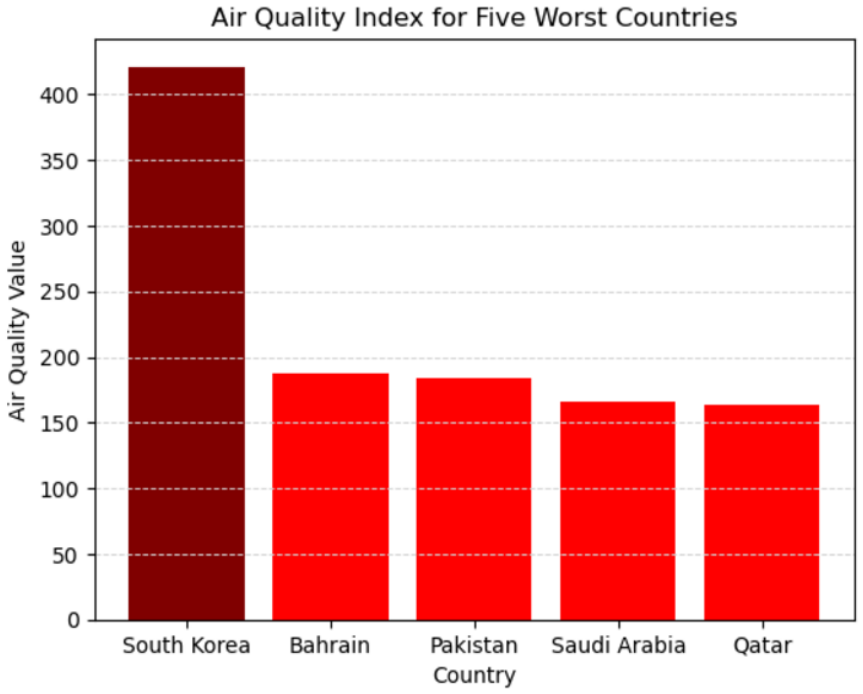
Worst AQI Value per country

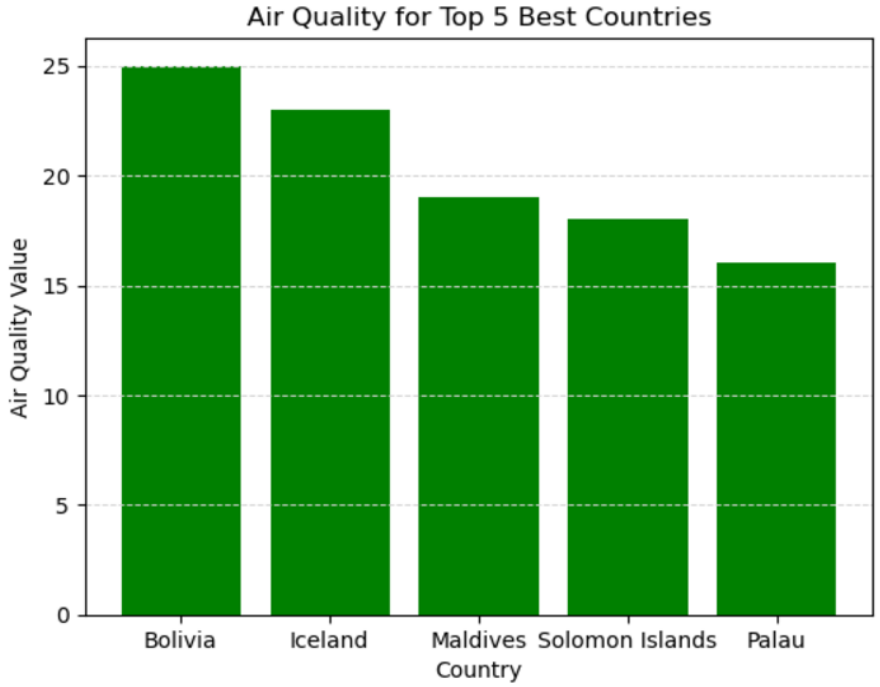
	Country	Air_Quality_Value	City
19	Bolivia (Plurinational State of)	24.875	24
67	Iceland	23.000	1
95	Maldives	19.000	1
140	Solomon Islands	18.000	1
117	Palau	16.000	1

Best AQI Value per Country

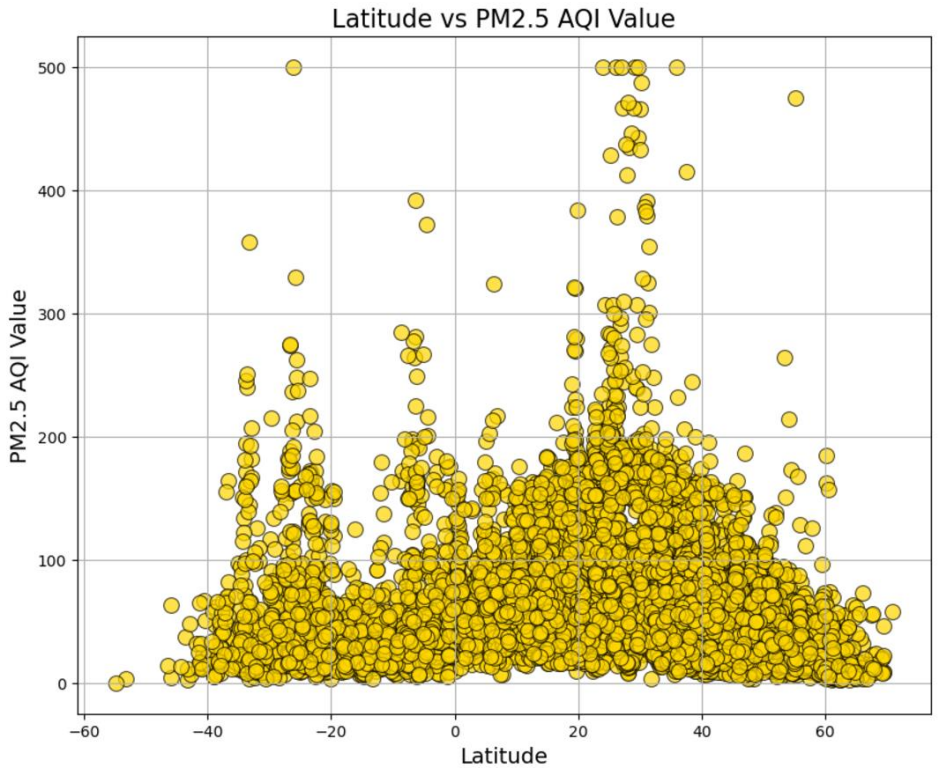
TOP 5 Best and Worst Countries based on AQI Value

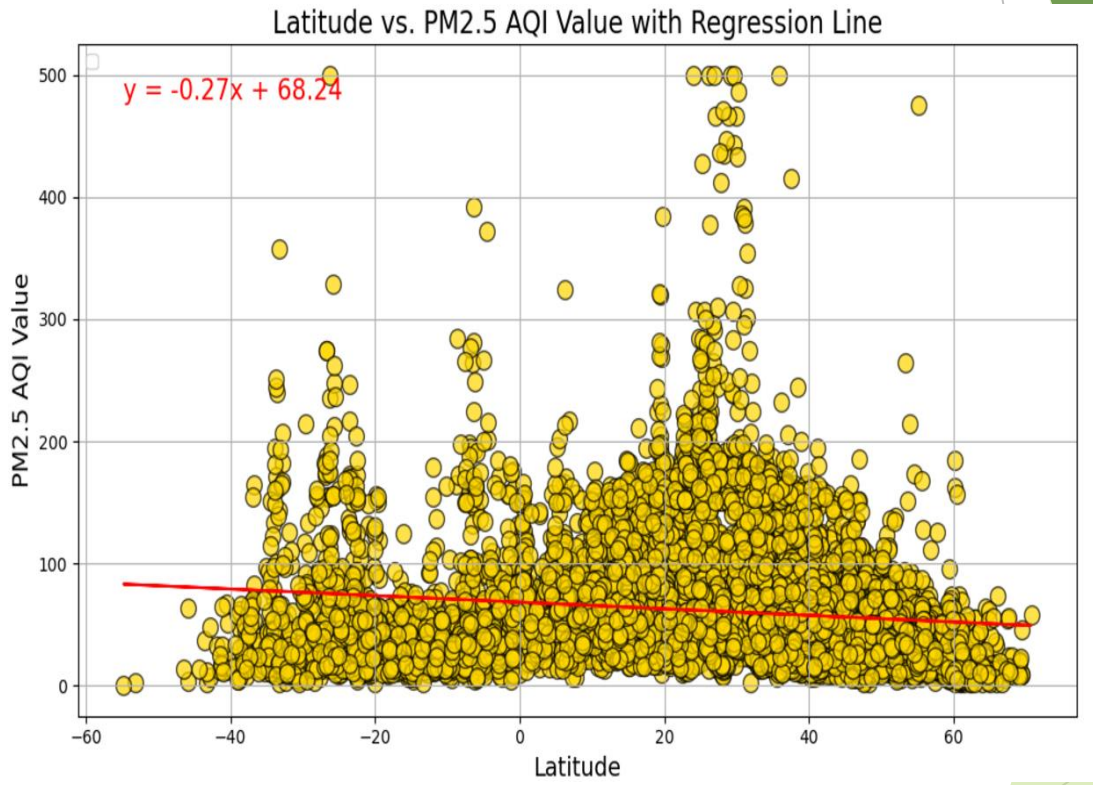
Reference: 0-50 AQI in good standing, 301-500 AQI is considered hazardous





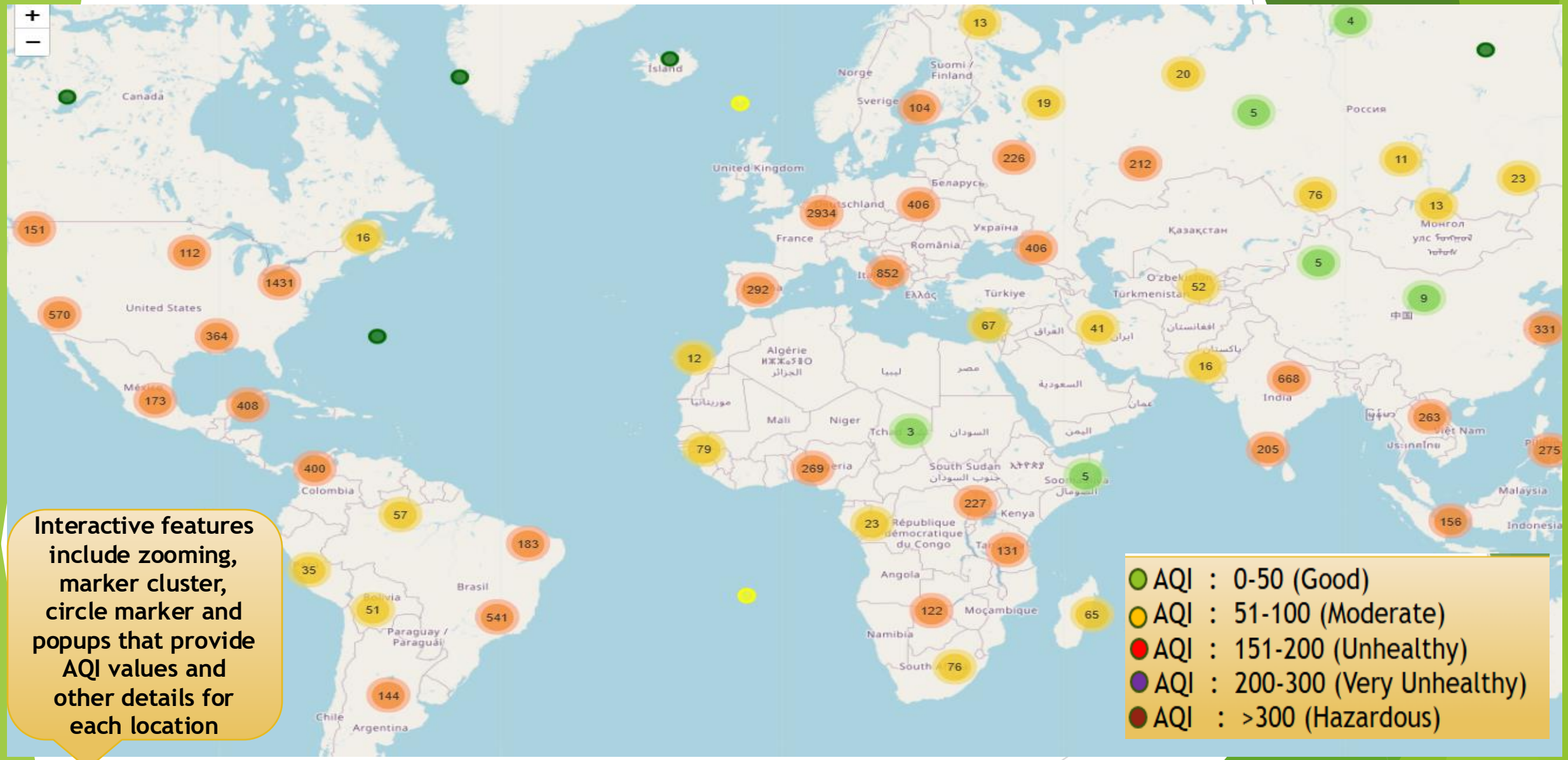
Scatter plot and linear regression for Latitude and PM2.5 AQI value



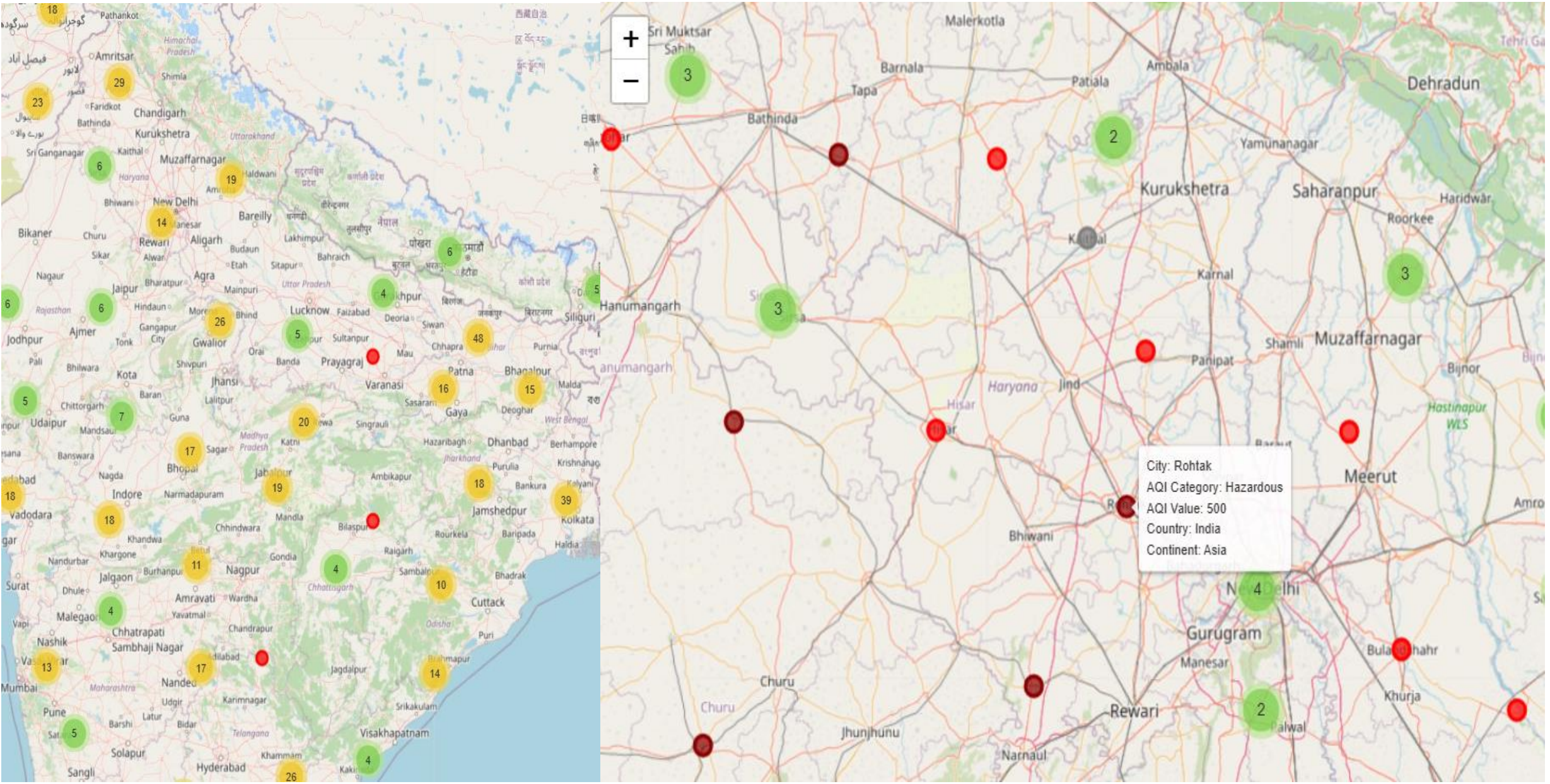


The scatter plot and linear regression showed no clear pattern between the AQI values and Latitude/Longitude. This suggests that the AQI values are influenced by urbanization, industrialization and other environmental factors rather than geographical location.

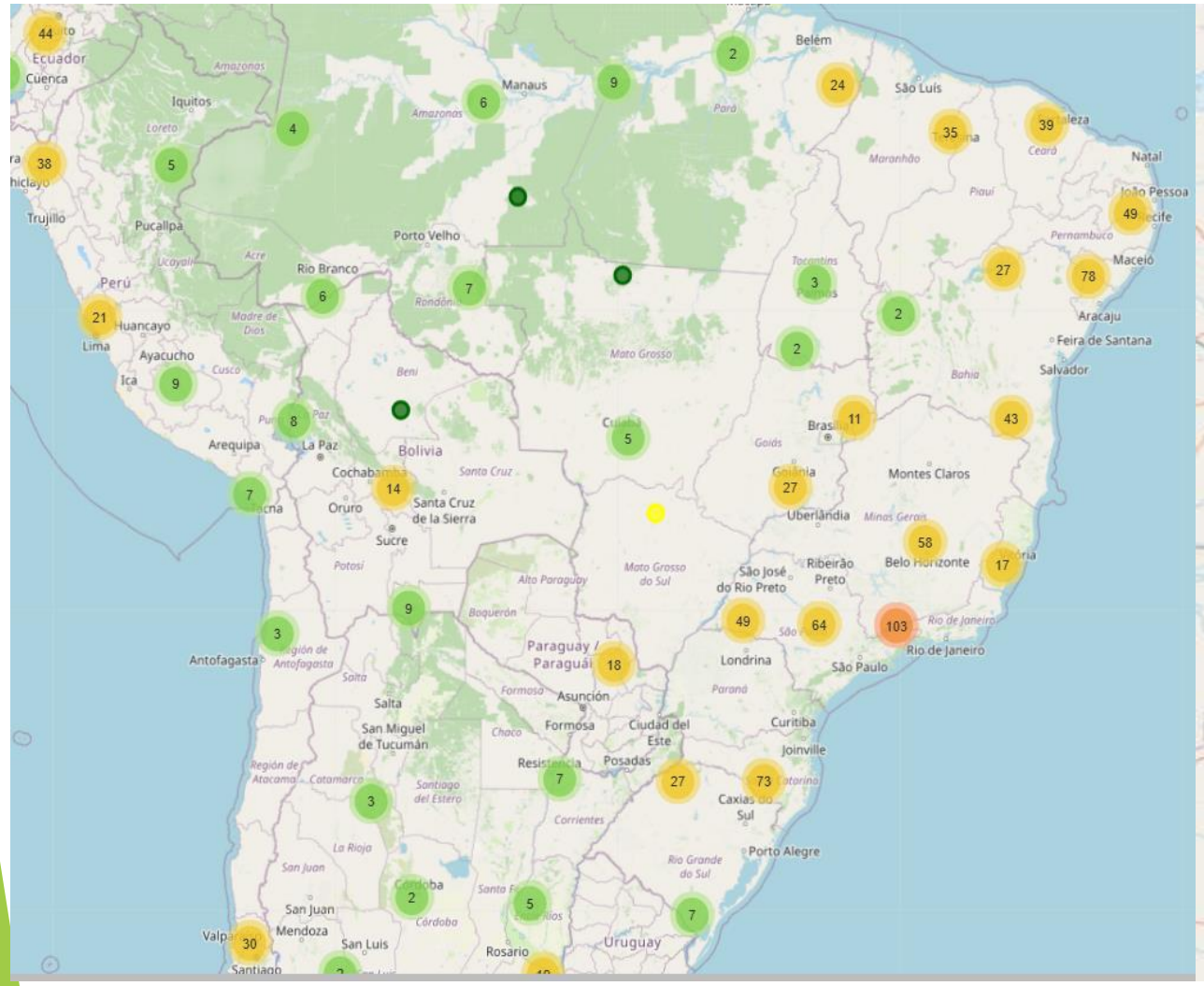
Air Quality Overview

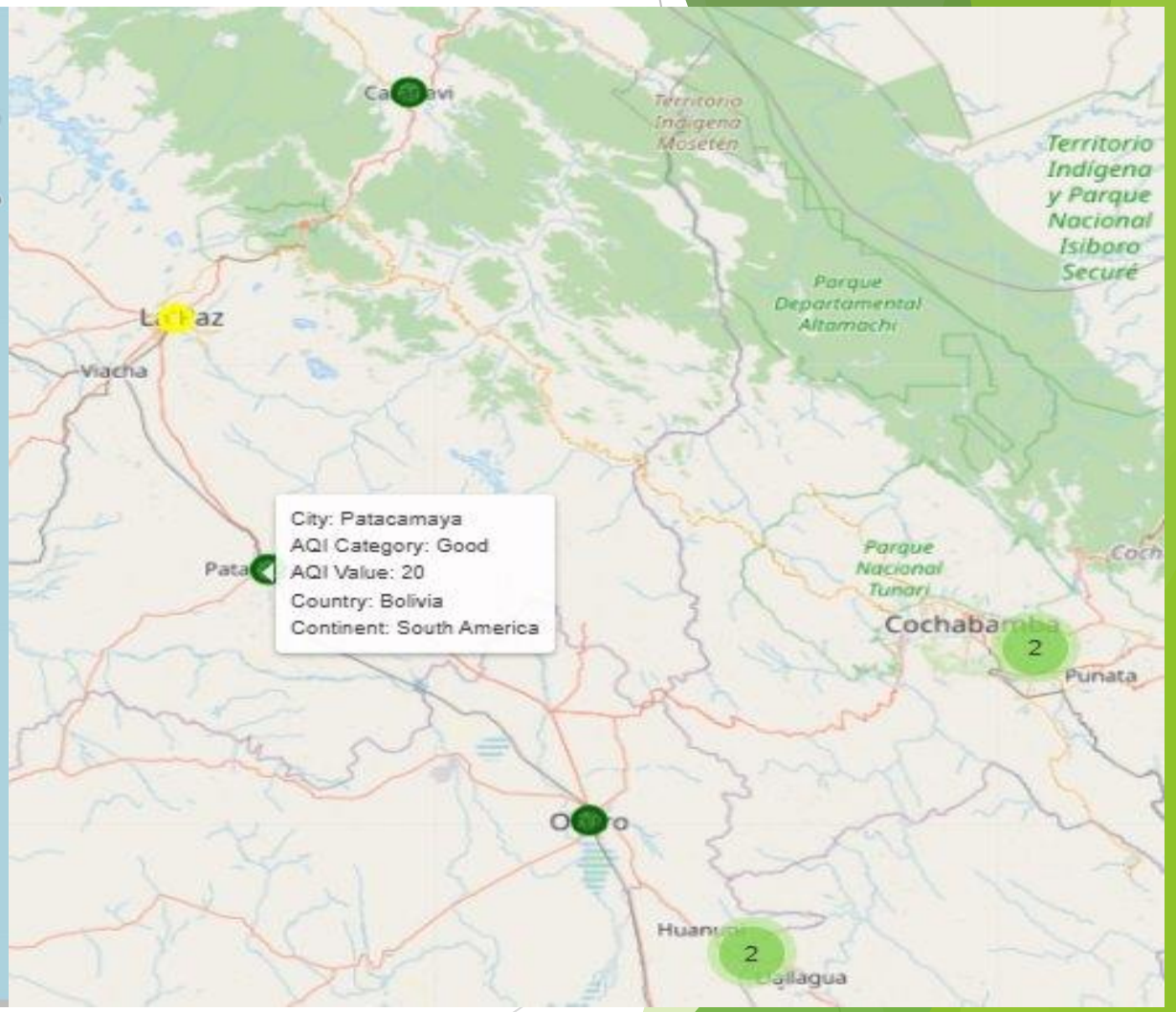


Map showing AQI levels in India



Map showing AQI levels in Bolivia, South America

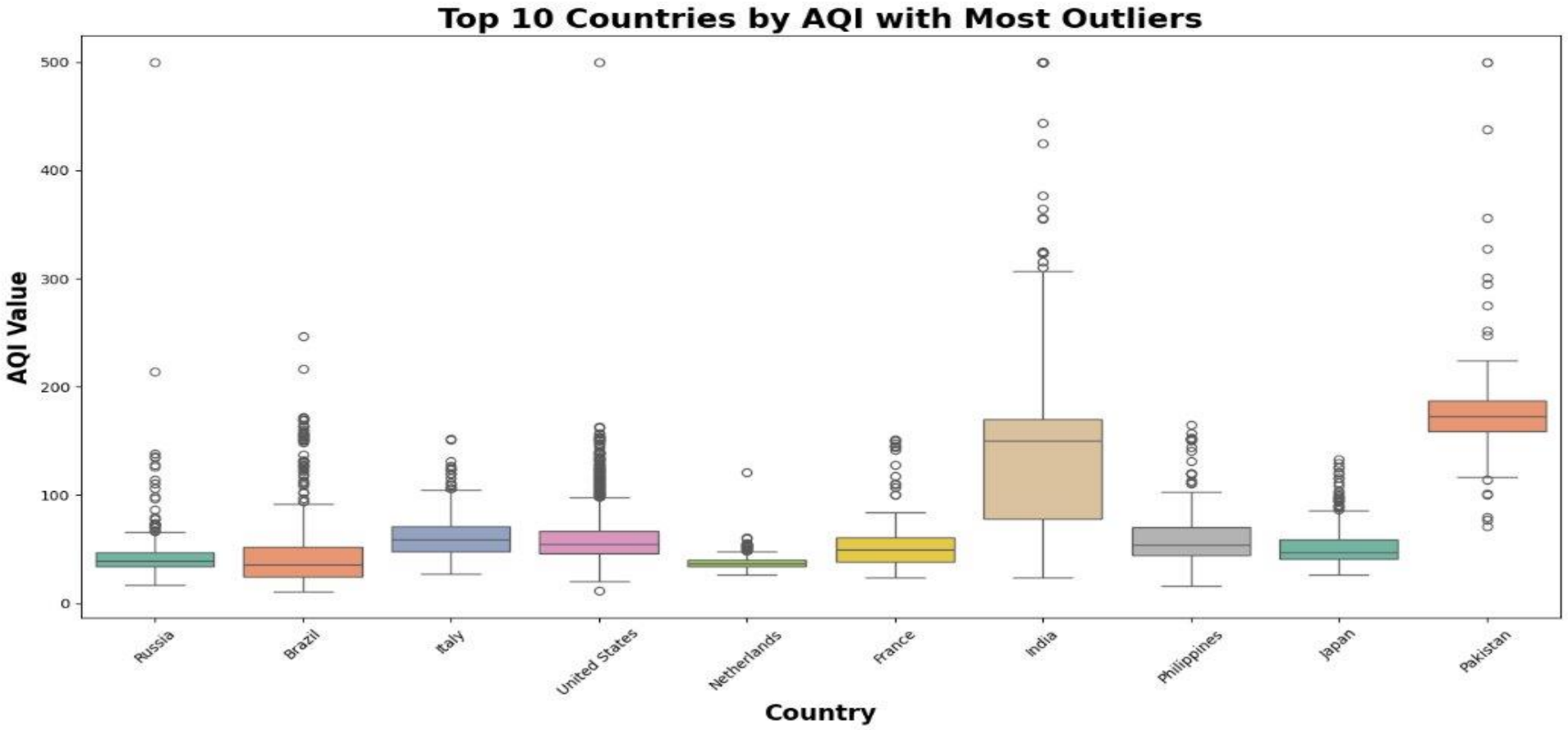




Leaderboard Showing the Quartiles and Outliers for top 10 countries

	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

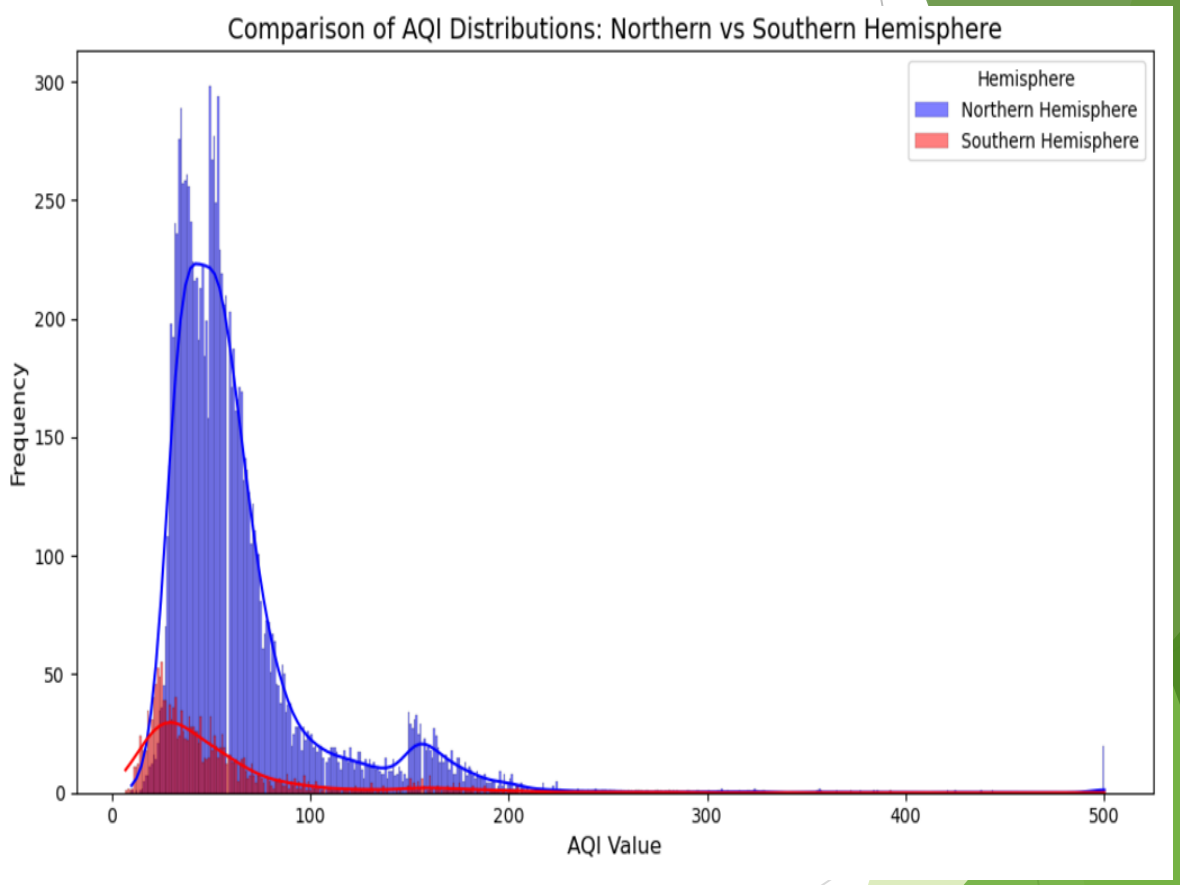
Box Plot for top 10 countries with most outliers for AQI value



Histogram for comparison of AQI distribution across northern and southern hemisphere

	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

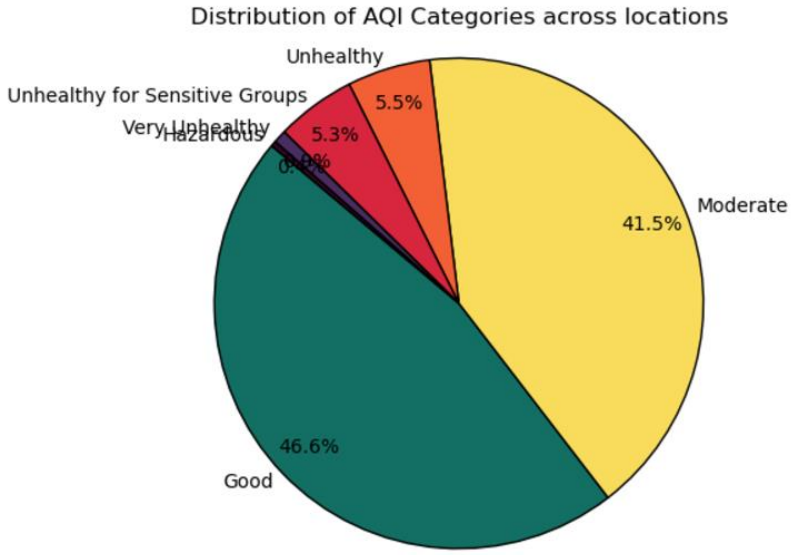
	Air_Quality_Category	Count	Hemisphere	Percentage
0	Good	1141	Southern Hemisphere	64.499717
1	Moderate	432	Southern Hemisphere	24.420577
2	Unhealthy	91	Southern Hemisphere	5.144149
3	Unhealthy for Sensitive Groups	68	Southern Hemisphere	3.843980
4	Very Unhealthy	32	Southern Hemisphere	1.808932
5	Hazardous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



Skewness of the data in northern hemisphere: 0.672
Skewness of the data in southern hemisphere: 1.376

Percentage Distribution Of AQI categories across the world

	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



T-test to compare AQI values

T-test: A T-test was performed to compare AQI values between Asia and Europe.

TtestResult(statistic=31.05031627218465, pvalue=4.150145543241 63e-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 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.

Anova Test to compare AQI across continents

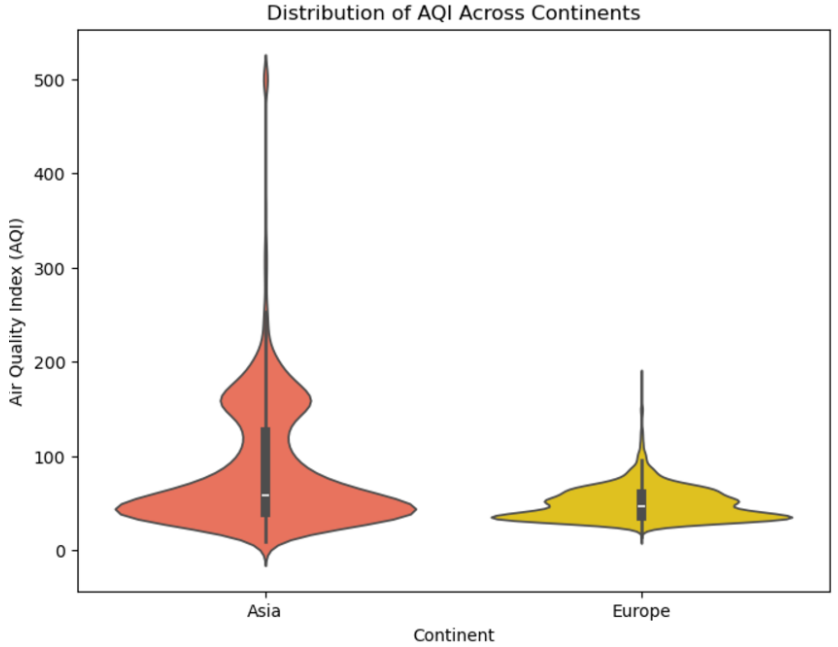
ANOVA: An ANOVA test was conducted to compare AQI values across all continents.

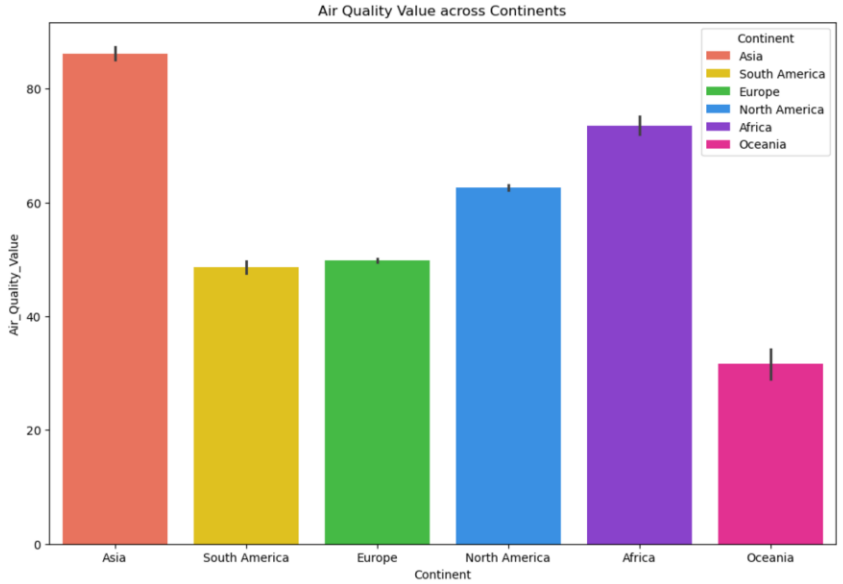
F_onewayResult(statistic=369.3766024923413, pvalue=0.0)

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.

Comparison of AQI values across continents with data visualization

We compared the AQI values across different continents to find out which regions are more polluted. 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 level of pollution.





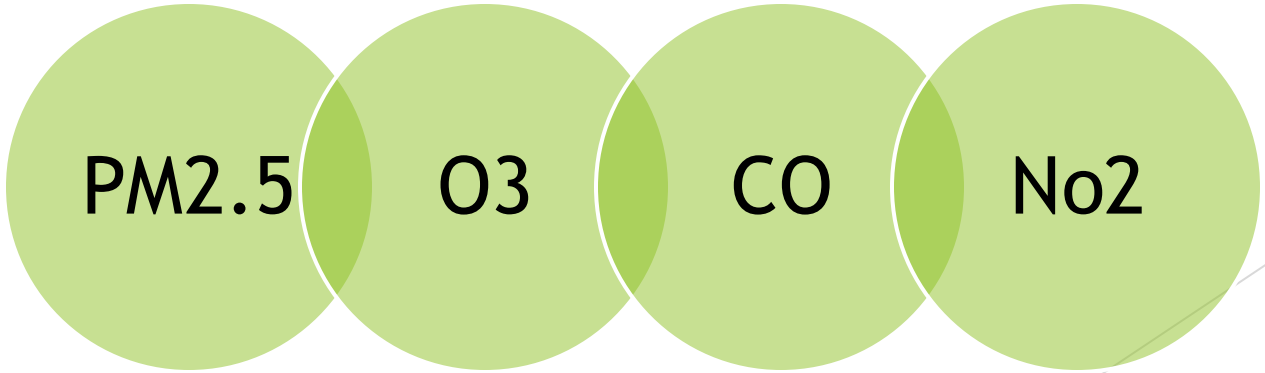
Bar plots and violin plots of AQI distribution across continents

Call to Action

- All four pollutants are caused by the burning of fuels and natural gas. Vehicle emissions are a main source of contamination, and industrial emissions are being linked to both PM2.5 and O3

- Industries should make an environmental movement to reduce their emissions.
  - Good news!
    - We see many companies taking action, becoming more sustainable.

- As a world we should look for alternative sources or cut our everyday emissions for travel in transportation.
  - Public Transportation, electrical vehicles, biking



Bias and Limitations

- Absence of source of pollution(industrial, vehicular) in the dataset.
- It does not include a date column, which means we cannot track or predict changes in air quality over time.
- Dropping of some country rows due to missing values reduced the overall sample size.

Future Work

- We would like to look into the renewable sources of energy and if there is a statistically significant difference with pollutant emissions.
- We would like to take closer look at industrial processing and how we can cut back pollutant sources.
- We would like to look at more datasets with the pollution source added as a column.
- We would like to have a dataset to include a population count, to verify our results from today.

Works Cited

- https://www.kaggle.com/datasets/adityaramachandran27/world-air-quality-index-by-city-and-coordinates/data
- https://www.epa.gov/ground-level-ozone-pollution/ground-level-ozone-basics
- https://www.iqair.com/us/world-air-quality
- Chatgpt for helping with coding

Questions?