URL: [Global Air Quality (2024) - 6 Cities]

License: MIT License

Download Date: June 1st, 2025

Raw shape before cleaning: 11 columns / 52705 rows

Dataset Report: Global Air Quality

In this report, I will be analyzing the global air quality in six major cities; Brasilia, Cairo, Dubai, London, New York, and Sydney. This dataset records CO, CO_2, NO_2, SO_2, O_3, PM2.5, and PM10 levels from December 31st, 2023 to December 31st, 2024. This is a dataset downloaded from Kaggle.

My analysis will be based on the Air Quality Index (AQI), which determines the overall air quality throughout the year. I will analyze when AQI rises (which months hold the highest record), what causes the index to rise, and how each city is driving this AQI.

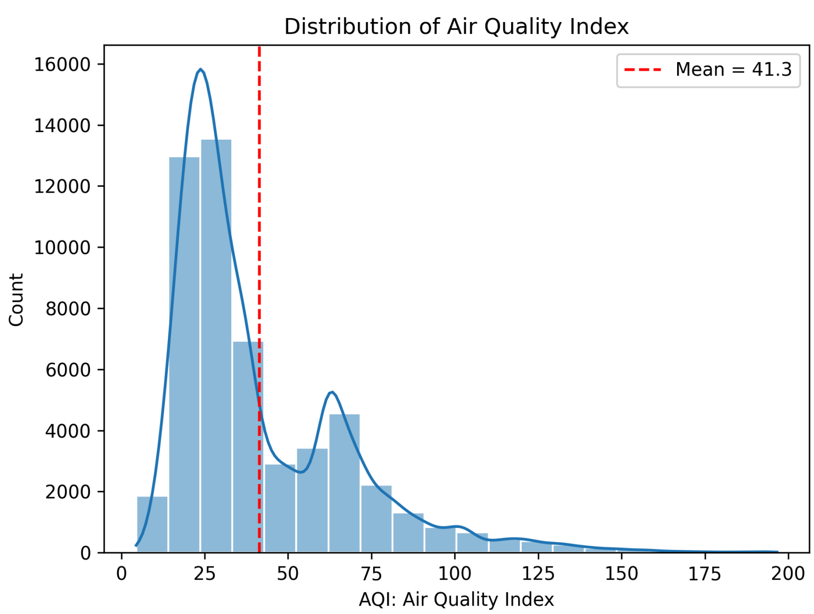

Fig.1: Histogram based on Distribution of Air Quality Index

```
Mean (average):      41.35
Median:              31.27
Standard Deviation:  26.63
Variance:            708.95
Skewness:            1.62
Kurtosis:            3.09
```

Fig.2: Descriptive statistics on AQI, including skewness and kurtosis

```
Mean AQI: 41.35
95% Confidence Interval for AQI: (41.12, 41.58)
```

Fig.3: AQI mean and 95% Confidence Interval

The histogram represents the overall distribution of the AQI in 2024 (Fig. 1). This is based on the count of AQI, which is trending between 25.0 and 45.0. After plotting this histogram, I created a descriptive statistics and AQI 95% confidence interval (CI) formula (Fig. 2 & 3). The descriptive statistics and histogram determine that AQI is right-skewed, meaning the median and standard deviation is less than the mean. However, the kurtosis on AQI is showing a value greater than 3, meaning that the data has a heavier tail with many outliers. The AQI mean is true to the dataset since the CI is narrow.

Analyzing this data, it's interesting to see how the AQI is right-skewed, but kurtosis is high. Based on these results, AQI is consistently 25.0-45.0 overall which creates a good impression (since the lower the AQI, the safer the air quality), but outliers break these patterns.

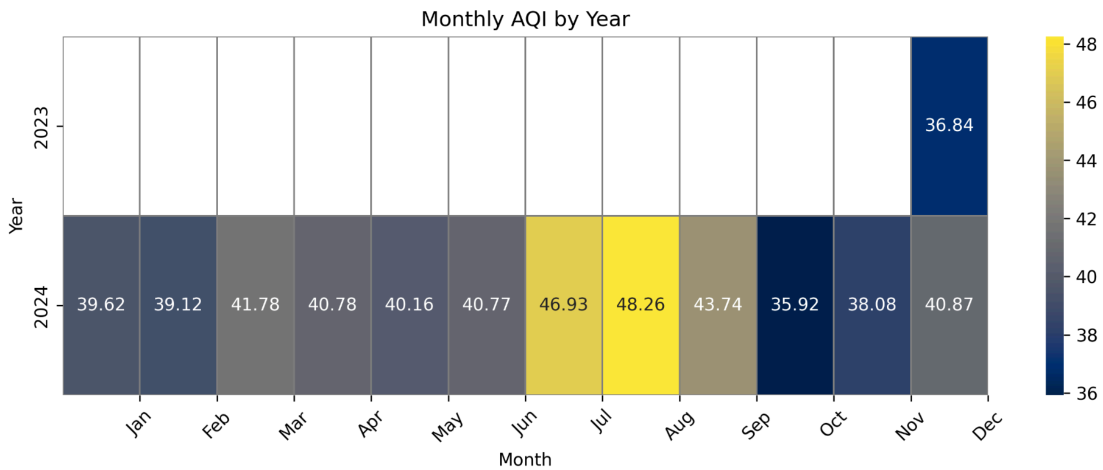

Fig.4: Heat map of average AQI per month in 2023-24

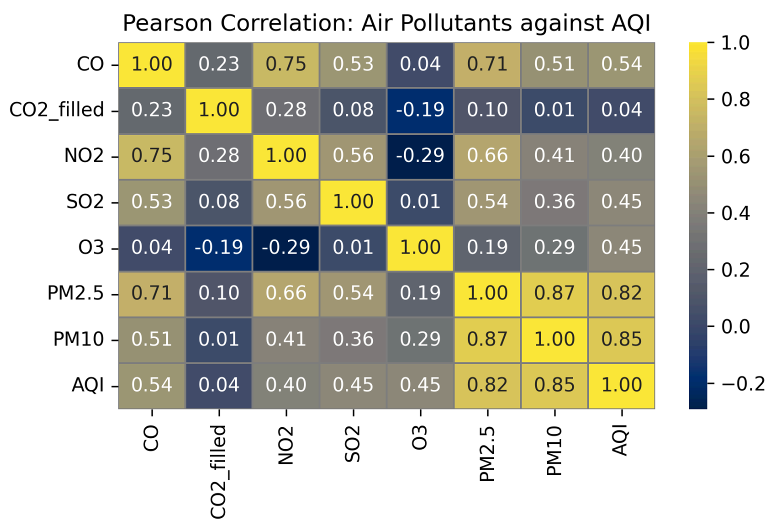

Fig. 5: Heat map of Pearson correlation between AQI and air pollutants

Here, I have created two heat maps to understand the AQI average per month and the correlation between each air pollutant in our dataset (which is CO, CO2, NO2, SO2, O3, PM2.5, and PM10) (Fig. 4 & 5). Based on Fig. 4, AQI tends to rise during the warmer months of the year, especially in July and August. In Fig. 5, I created a pearson correlation heat map to focus my analysis on the air pollutants correlation to AQI. It appears that PM10 tends to have the biggest relationship to AQI.

```
T-statistic: 44.3242
P-value: 0.0000
Result: Reject the null hypothesis. PM10 has a significant effect on AQI.
```

Fig. 6: T-test and p-value analysis based on AQI and PM10

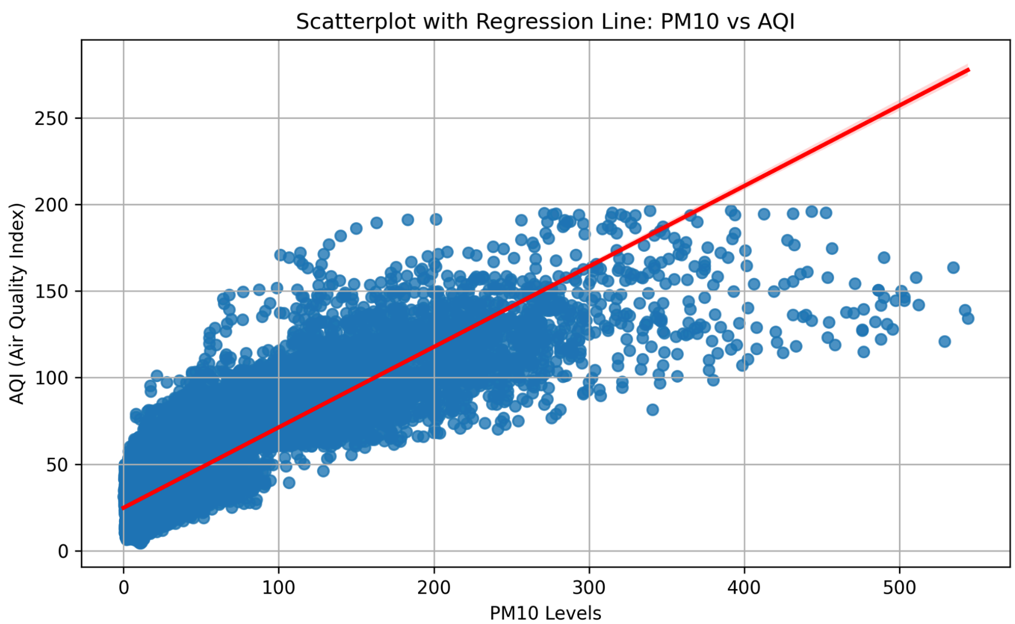

Fig. 7: Scatterplot and Regression Line based on PM10 and AQI

Using the t-test and p-value analysis based on AQI and PM10, I want to see if PM10 makes a difference on the AQI. Since this data is paired, and with support from GPT, I was recommended to use ttest_rel for comparing. After creating this formula, I have determined that the p value is less than 0.05. In this case, PM10 does have a significant effect on the AQI. To visualize this key insight, I scatterplot the data with a regression line to forecast the overall relationship (Fig. 7). Here we can see a linear relationship between the two fields, as PM10 tends to rise, AQI rises with it. However, thinking back to our kurtosis, this plot shows scattered outliers in our dataset, which could potentially be manipulating the t-test.

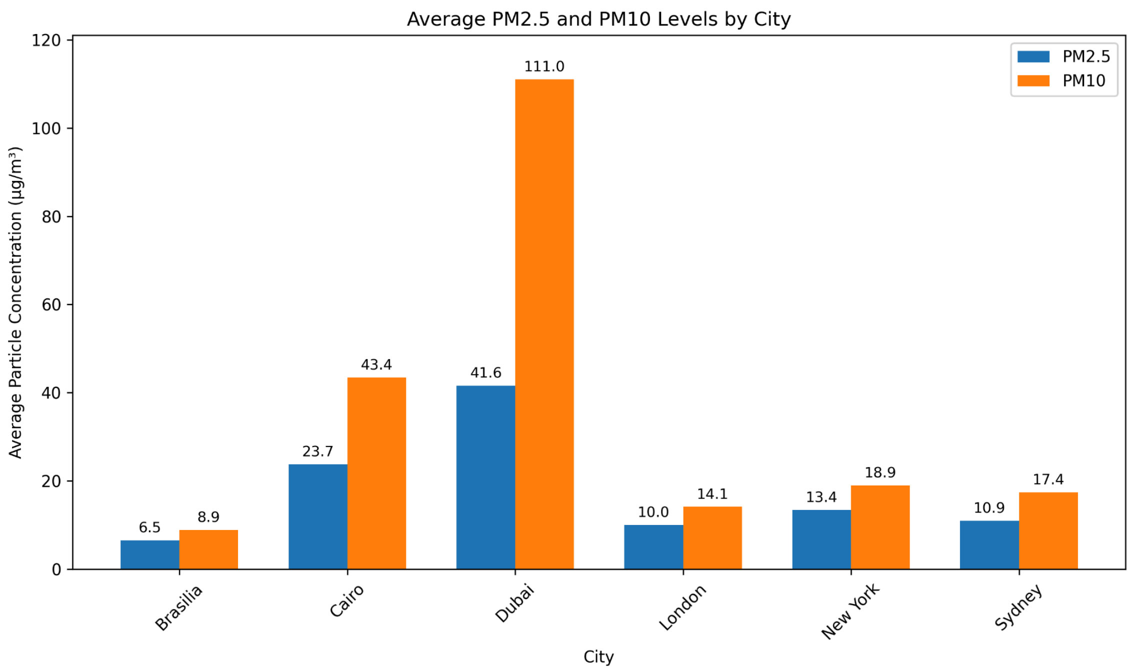

Fig. 8: PM2.5 & PM10 distribution between six major cities

Understanding PM10 level's relationship to the AQI, I also created a grouped bars chart to distribute PM2.5 and PM10 by city in our dataset. Although all these cities contribute to PM10 levels, Dubai holds the most based on this visualization. In fact, Dubai might be the main outlier in our dataset. This brings into question why Dubai has the most PM10 levels out of the six major cities.

My key take-away on this analysis is that, although AQI tends to average between 25.0 and 45.0, the levels tend to rise during the summer. Although each air pollutant is a key contributor to rising AQI, the biggest factor to the rising levels have been from PM10. Although the dataset is only based on six cities in the world, Dubai tends to hold the most PM10 levels, which could be a main contributor in the AQI outliers.

My recommendation to domain holders is to investigate PM10 levels in Dubai and consider factors other than greenhouse gases and emissions (such as biome, temperature, etc). I would recommend domain holders to consider how to lower AQI levels in the summer, since they record higher during this season. This dataset is only based on six cities, so I would recommend branching out this dataset to more cities like Tokyo, Shanghai, and Singapore to create a more accurate analysis on AQI and determine if PM10 is still a significant factor to rising levels.