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 pollutants 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)
```

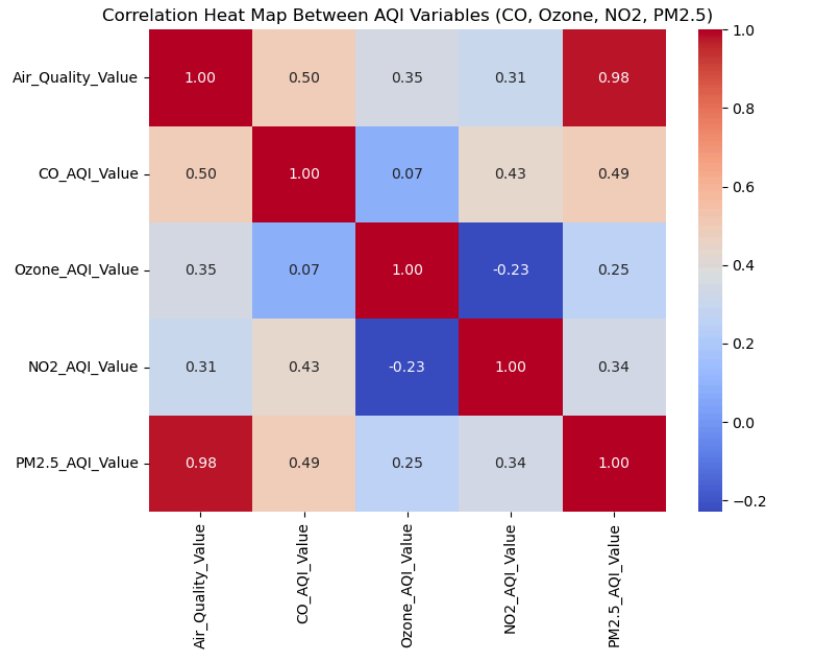
```
                  Air_Quality_Value  CO_AQI_Value  Ozone_AQI_Value  \
Air_Quality_Value           1.000000      0.495144         0.348281
CO_AQI_Value                0.495144      1.000000         0.074985
Ozone_AQI_Value             0.348281      0.074985         1.000000
NO2_AQI_Value               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 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 with a value of 0.98 we find that PM2.5 has the greatest effect on AQI.

```
plt.figure(figsize=(8, 6))  # Adjust figure size
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Heat Map Between AQI Variables (CO, Ozone, NO2, PM2.5)')
plt.show()
```



The result of having a high correlation between AQI Value and PM.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.

```
#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

We then plotted these points on a scatterplot using the Matplotlib visualization tool. From the figure 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 Figure 1. To do so we added a regression model to the scatter plot Figure 2.

```
sns.scatterplot(x='Air_Quality_Value', y='PM2.5_AQI_Value', data=PM_df, color='yellow',edgecolor='black', s=100)  # `s` controls marker size
plt.title('AQI Value vs PM2.5 Value')
plt.xlabel('Air_Quality_Value')
plt.ylabel('PM2.5_AQI_Value')
plt.grid(True)
plt.show()
```

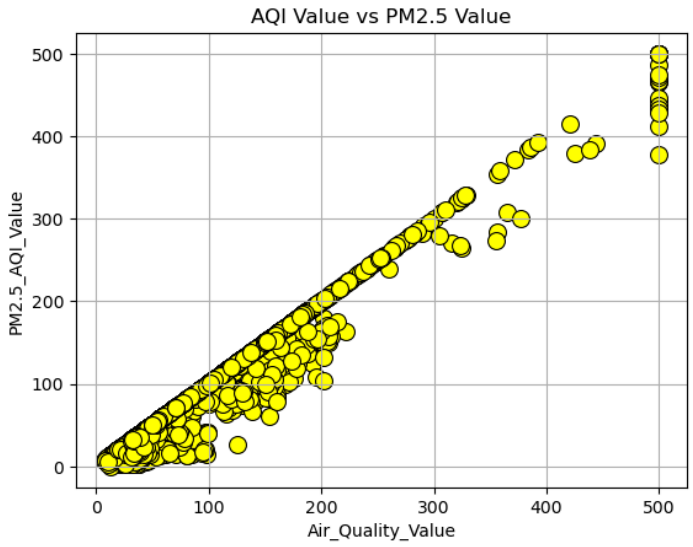


Figure 1: Scatterplot of AQI Value vs PM2.5 Value

```
# 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': 'r

# 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.show()
```

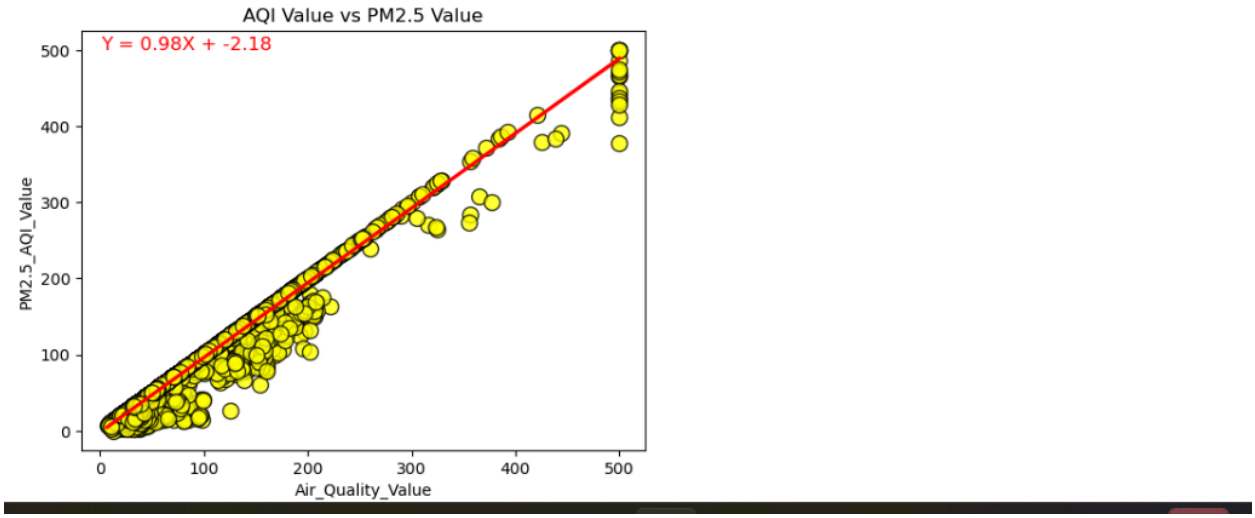


Figure 2: AQI Value Vs PM2.5 Value with a regression model showing line of best fit (Y= 0.98x + -2.18)

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.

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 bigger 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 an key variable to add for measuring AQI Value

Works Cited

“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

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.