



REAL TIME AIR POLLUTION, COMPARING CONTINENTS AND EVALUATING PERFORMANCE

ABSTRACT

By reviewing live data of pollution indicators it will be shown which are the best and worst performers by continent. The data obtained can be expanded to include historical data to analyse long term performance. The snapshot provided can give an insight into where improvements could be targeted.

John Truong
Wei Ke (William)
Callum Linnegan
Karissa Malseed
James Rydlewski

Data Analytics Bootcamp: Project 1

Contents

| | |
|---|----|
| Introduction | 2 |
| Data Presentation | 2 |
| Background | 2 |
| Methodology..... | 4 |
| 1. World Cities by Continent | 4 |
| 2. API Calls..... | 5 |
| 3. Storing pollution data | 5 |
| 4. Data frame creation | 5 |
| 5. Cursory Graphs..... | 6 |
| 6. Saving csv files..... | 7 |
| 7. Map Visualisation..... | 8 |
| Research Questions | 8 |
| Data Analyses..... | 9 |
| • Bar Graphs showing AQI for cities in each continent | 9 |
| 2. Pie Charts Showing contents of Air pollution in each continent | 12 |
| 3. Scatter plots of Polutants to check correlation | 14 |
| Challenges and Limitations | 14 |
| Conclusion..... | 15 |

Introduction

Air Pollution is produced from a host of different sources and is closely linked to population size and heavy industries in a locality. This report is focussing on considering a cross section of cities from different continents to make a comparison of the pollution data. From this it will be possible to glean which region is the greatest contributor to global air pollutants.

The data obtained is real time snap shot of air pollution in different cities this will be taken as a typical output for each of the cities and regions considered. For future iterations of this project historical data can be used which would show how air pollution has increased or decreased with respect to time.

Data Presentation

The joy of data analytics is the story that can be presented illustrating the findings of the research performed. The data gathered for our analysis is real time air pollution data. The graphics elected to show this data are:

- Pie chart showing the concentration of pollutants in each regions worst performing city
- Bar chart showing the air quality index for each city within a region.
- Top 10 worst performing cities by region for each pollutant category
- Scatter plot of pollutants to see what corellations exist
- Google map to visualise the location and intensity of Air pollution.

Background

This project is based on API calls for air quality data from the World Air Quality Index¹.

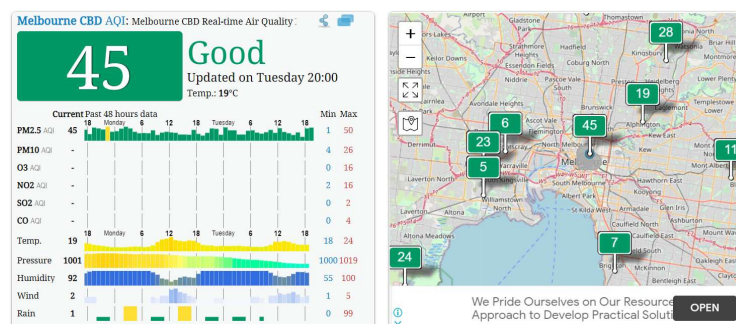


Figure 1 Sample of Data that can be obtained from WAQI

¹ Website data obtained from is waqi.info. Wonderful resource with interactive map featuring air quality data from around the world.

Figure 1 illustrates the sample data that will be explored in this analysis. Our objective is to obtain the following parameters for a number of cities across each continent.

| <u>Parameter</u> | <u>Definition</u> |
|-------------------------|---|
| PM _{2.5} | PM _{2.5} is particulate matter that is less than 2.5µm in diameter. These particles are small enough to be inhaled deeply which can have significant health implications. Often in regions where there is a high PM _{2.5} count people will experience difficulty breathing. These particles are also problematic as they can be absorbed into your blood stream which can lead to further health complications of particular concern is cancer. |
| PM ₁₀ | PM ₁₀ is particulate matter that is less than 10µm in diameter. Although the particle size is greater than PM _{2.5} it is still problematic. Typically found in smoke and smog in terms of air pollution but can also occur as suspended solids during the manufacturing process. An example would be rock dust whilst quarrying. In the context of air pollution. PM pollution causes significant health concerns. A particularly vicious episode occurred in London during the 1950's which led to the death of over 4000 people. |
| O ₃ (Ozone) | Whilst Ozone occurring in the upper atmosphere has a beneficial effect for humanity as it shields the earth from harmful solar rays, at ground level it has the propensity to damage health. It is caused by the reaction of Nitrates and Volatile Organic Compounds (VOC's) in the presence of sunlight. It is the main pollutant in "smog" and causes breathing difficulties for many. In addition to harming human health it can damage the wellbeing of plant life. Photosynthesis can be reduced slowing the plants growth and potentially killing affected species. As this happens there is less biodiversity which has a knock on effect to other entities dependent on the affected plant. |
| NO ₂ | Motor Vehicle exhaust gasses and heavy industry are responsible for the emission of NO. In the presence of air this reacts to form NO ₂ . The effects are breathing difficulties especially for the elderly and children. Significant quantities of aerosol NO ₂ can lead to acid rain which damages other elements of the eco system. Acid rain can further damage plant and marine life. |
| SO ₂ | SO ₂ pollution is heavily caused by the burning of fossil fuels. Similar to NO ₂ It leads to acid rain and harms the respiratory tract and eyes. |
| CO | CO is caused by the incomplete combustion of fuels. This occurs when not enough oxygen is present during the combustion process. The major issue of CO pollution is the tendency for it to form a blanket at surface level. This again causes breathing difficulties, and potentially death. It also affects photosynthesis in plants further damaging the environment |
| Air Quality Index (AQI) | AQI takes the level of pollutants present and combines this data with the prevailing weather conditions to index the overall air quality for a location. The data obtained from this can be used to see which pollutants contribute to AQI performance, and by how much. |

Methodology

The process used to collate and analyse global pollution data is as follows:

- 📁 Generating lists of world cities divided by continent
- 📁 API calls from World Air Quality Index Site for each city.
- 📁 Storing pollution data for each city within each continent.
- 📁 Building data frames for each continents data removing null values.
- 📁 Generating rudimentary graphs of the AQI and Pollutant data.
- 📁 Storing all data frames as .csv for further analysis.
- 📁 Visualising the data on a map.

An .ipynb and .config file to store api keys were created in addition to the above. Upon creating the file dependencies were imported. The dependencies used are there to help obtain and manipulate the data for world pollution.

```
In [1]: 1 #Import Dependencies
        2 import gmaps
        3 import requests
        4 import statistics
        5
        6 import pandas as pd
        7 import matplotlib.pyplot as plt
        8 import scipy.stats as stats
        9
       10 from scipy import stats
       11 from config6 import gkey
       12 from pprint import pprint
       13 from config6 import api_key
```

Figure 2 Dependencies Imported for further analysis

1. World Cities by Continent

To perform this research the most populated cities by continent. The data for this has been taken from an open source website, which has then been copied into the.ipynb environment

```
1 #Save config information
2 #List of most populated cities in North America (Len = 50)
3 namerican_cities = ["Mexico City", "New York City", "Los Angeles", "Toronto", "Chicago", "Houston", "Havana", "Montreal",
4                     "Ecatepec de Morelos", "Philadelphia", "Phoenix", "San Antonio", "Guadalajara", "Puebla", "San Diego",
5                     "Juárez", "León", "Dallas", "Tijuana", "Calgary", "Tegucigalpa", "Zapopan", "Monterrey", "Managua",
6                     "Nezahualcóyotl", "San Jose", "Santo Domingo", "Guatemala City", "Port-au-Prince", "Naucalpan", "Ottawa",
7                     "Austin", "Edmonton", "Mérida", "Querétaro", "Toluca", "Jacksonville", "Chihuahua", "San Francisco",
8                     "Indianapolis", "Columbus", "Fort Worth", "Charlotte", "Hermosillo", "Saltillo", "Aguascalientes",
9                     "Mississauga", "San Luis Potosí", "Veracruz", "San Pedro Sula"]
10 #shorten for ease
11 na_cities = namerican_cities
```

Figure 3 North American Cities list, This process has been repeated for each continent granting 6 lists featuring the most populated cities in each continent

2. API Calls

The city list data was used in combination with API keys for each parameter that is needed to be observed. To discover where the pertinent information was being stored in the API call a trial run for a single city was run to check how the data was being called.

```

1 #Save config information (use a city as an example)
2 city = "Melbourne"
3
4 url = 'http://api.waqi.info/feed/' + city + '/?token='
5
6 main_url = url + api_key

1 #Print JSON response
2 response = requests.get(main_url)
3 data = response.json()
4 pprint(data)

{'data': {'aqi': 44,
          'attributions': [{'logo': 'Australia-Victoria.png',
                             'name': 'Environment Protection Authority | EPA ',
                             'url': 'http://epa.vic.gov.au/'},
                           {'name': 'World Air Quality Index Project',
                             'url': 'https://waqi.info/'}],
          'city': {'geo': [-37.8073959, 144.97],
                    'name': 'Melbourne CBD',
                    'url': 'https://aqicn.org/city/australia/melbourne/melbourne-cbd'},
          'debug': {'sync': '2021-03-25T16:19:27+09:00'},
          'dominantpol': 'pm25',
          'forecast': {'daily': {'o3': [{'avg': 3,
                                         'day': '2021-03-23',
                                         'max': 9,
                                         'min': 1},
                                       {'avg': 10,
                                         'day': '2021-03-24',
                                         'max': 20,
                                         'min': 1}]}}}}

```

Figure 4 Process showing how data was found in the API

3. Storing pollution data

The Pollution Data that was obtained was found using this method, figure 5 illustrates the code used for this. For each pollutant an exception clause was added to ensure that the code did not stall when null value were found. The excerpt featured shows one such call which had the data appended to a list.

```

for city in na_cities:
    response = requests.get('http://api.waqi.info/feed/' + city + '/?token=' + api_key).json()
    try:
        pm2_5.append(response['data']['iaqi']['pm25']['v'])
        print(f"{city}'s PM2.5 found! Appending stats")
    except:
        pm2_5.append("")
        print("Data not found")
        pass

```

Figure 5 Calling and Appending pollution data

4. Data frame creation

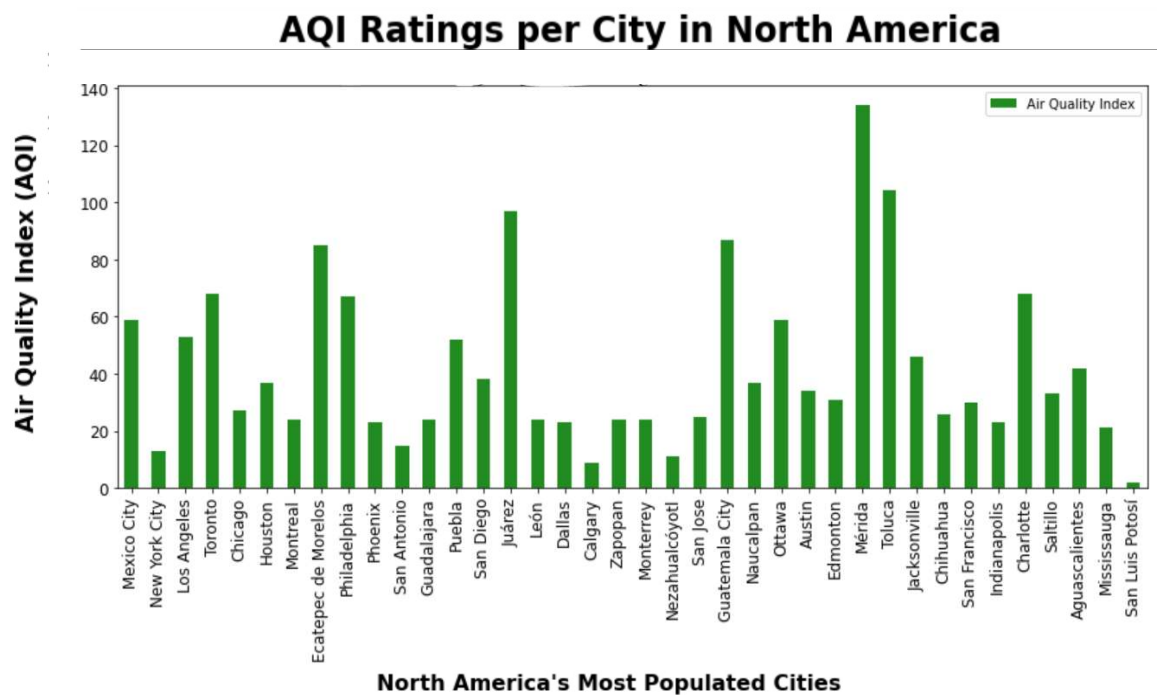
The list data for each pollutant was combined into a single data frame for the first step of visualisation. This single data frame is the backbone for the visualisations that will be displayed of the pollution data.

| Rank | North America's Most Populated Cities | Geolocation | Air Quality Index | Carbon Monoxide (µg/m3) | Hydrogen (µg/m3) | Ozone (µg/m3) | Sulphur Dioxide (µg/m3) | Fine Particle Matter (µg/m3) | Coarse Particulate Matter (µg/m3) | Timestamp |
|------|---------------------------------------|---------------------------|-------------------|-------------------------|------------------|---------------|-------------------------|------------------------------|-----------------------------------|---------------------|
| 1 | Mexico City | [19.42461, -99.119594] | 59 | 13.4 | 23 | 1.6 | 2.9 | 59 | 35 | 2021-03-25 01:00:00 |
| 2 | New York City | [40.7127837, -74.0059413] | 13 | 0 | 99 | 0 | 0 | 13 | 0 | 2021-03-25 01:00:00 |
| 3 | Los Angeles | [34.06653, -118.22676] | 53 | 3.7 | 51.7 | 13.6 | 1.5 | 53 | 39 | 2021-03-24 23:00:00 |
| 4 | Toronto | [43.653226, -79.3831843] | 68 | 2.3 | 87 | 16.8 | 0.2 | 68 | 0 | 2021-03-25 03:00:00 |

Figure 6 North American Pollution Dataframe

5. Cursory Graphs

For the purpose of quickly visualising the data for each continent Bar and Pie charts have been used to show how each city is tracking with air quality and what proportions of pollutants are present.



North America's Most Populated Cities

Figure 7 AQI of N. Americas most populated cities

```

20
21 #Create a bar graph to represents the concentration AQI for corresponding City
22 na_bar_pollution_df.plot('North America's Most Populated Cities',
23                             'Air Quality Index',
24                             kind = 'bar',
25                             figsize = (12, 7),
26                             rot = 90,
27                             color = 'forestgreen',
28                             title = 'AQI Ratings per City in North America',
29                             fontsize = 12
30                             )
31 plt.xlabel('North America\'s Most Populated Cities', fontsize = 16, fontweight = 'bold')
32 plt.ylabel('Air Quality Index (AQI)', fontsize = 16, fontweight = 'bold')
33 plt.title('AQI Ratings per City in North America', fontsize = 24, fontweight = 'bold')
34 plt.tight_layout()
35 plt.show()

```

Figure 8 Bar Plot Code

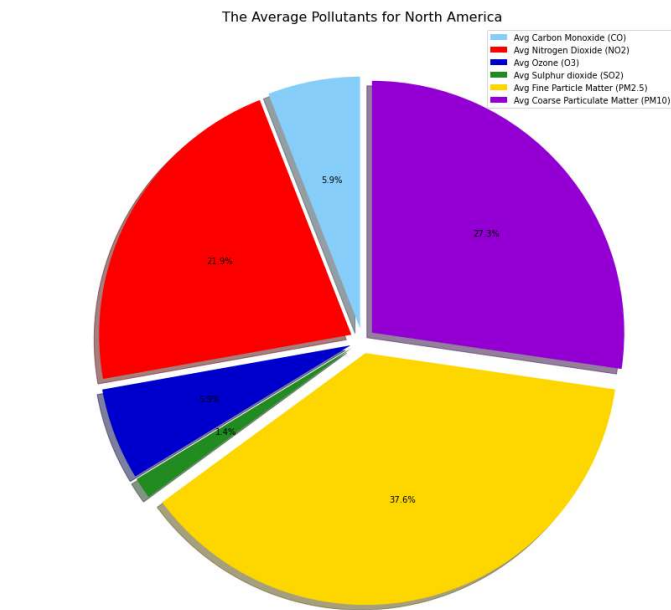


Figure 9 Pie chart of average N.American Pollutants

```

labels = na_pollutants
colors = ['lightskyblue', 'red', 'mediumblue', 'forestgreen', 'gold', 'darkviolet',]
explode = (0.05, 0.05, 0.05, 0.05, 0.05, 0.05)
fig, ax1 = plt.subplots(figsize = (11, 11))
plt.pie(na_values,
        explode = explode,
        colors = colors,
        startangle = 90,
        autopct = '%.1f%%',
        shadow = True)

plt.title("The Average Pollutants for North America", fontsize = 16)
ax1.legend(labels = ['Avg Carbon Monoxide (CO)',
                    'Avg Nitrogen Dioxide (NO2)',
                    'Avg Ozone (O3)',
                    'Avg Sulphur dioxide (SO2)',
                    'Avg Fine Particle Matter (PM2.5)',
                    'Avg Coarse Particulate Matter (PM10)'],
          loc = 'upper right')
plt.tight_layout()
plt.show()
print("All values are converted from µg/m3 to AQI levels using the EPA standard")

```

Figure 10 Pie Chart Code

6. Saving csv files

By saving each data frame as a CSV further analysis can be performed by importing the data into a new .ipynb file.

```

In [203]: 1 #Save dataframe as csv file
          2 na_pollution_df.to_csv('NorthAmericanPollution.csv')

In [204]: 1 #Save dataframe as csv file
          2 sa_pollution_df.to_csv('SouthAmericanPollution.csv')

In [205]: 1 #Save dataframe as csv file
          2 oceania_pollution_df.to_csv('OceaniaPollution.csv')

In [206]: 1 #Save dataframe as csv file
          2 eu_pollution_df.to_csv('EuropePollution.csv')

In [207]: 1 #Save dataframe as csv file
          2 af_pollution_df.to_csv('AfricanPollution.csv')

In [208]: 1 #Save dataframe as csv file
          2 asian_pollution_df.to_csv('AsianPollution.csv')

```

Figure 11 Saving Data Frames as .csv files

7. Map Visualisation

TO complete the general visualisation to show how polluted each city is a map showing the location of the city and its AQI is used. This can be performed on an individual city basis and also for all cities that have been considered. Shown below is the visualisation and code for a single city.

```
1 # git bash 'conda install cartopy'
2 import cartopy.crs as ccrs
3 from cartopy.mpl.ticker import LongitudeFormatter, LatitudeFormatter
4 geo = data['city']['geo']
5 fig = plt.figure(figsize=(12,8))
6 ax = plt.axes(projection = ccrs.PlateCarree())
7 ax.stock_img()
8 tick_proj = ccrs.PlateCarree()
9 ax.set_xticks(np.arange(-180, 180+60, 60), crs=tick_proj)
10 ax.set_xticks(np.arange(-180, 180+30, 30), minor=True, crs=tick_proj)
11 ax.set_yticks(np.arange(-90, 90+30, 30), crs=tick_proj)
12 ax.set_yticks(np.arange(-90, 90+15, 15), minor=True, crs=tick_proj)
13 plt.scatter(geo[1],geo[0],color='blue')
14 ax.xaxis.set_major_formatter(LongitudeFormatter())
15 ax.yaxis.set_major_formatter(LatitudeFormatter())
16 plt.text(geo[1]+3,geo[0]-2,f'{city} AQI \n {aqi}', color='red')
17 plt.show()
```

Figure 12 Map printing code

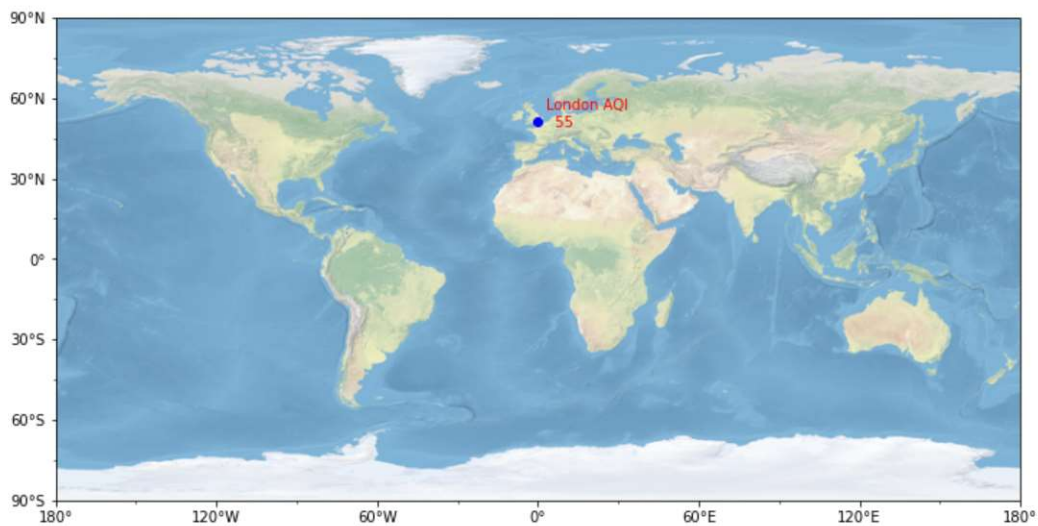


Figure 13 Single city and AQI visualised on map

Research Questions

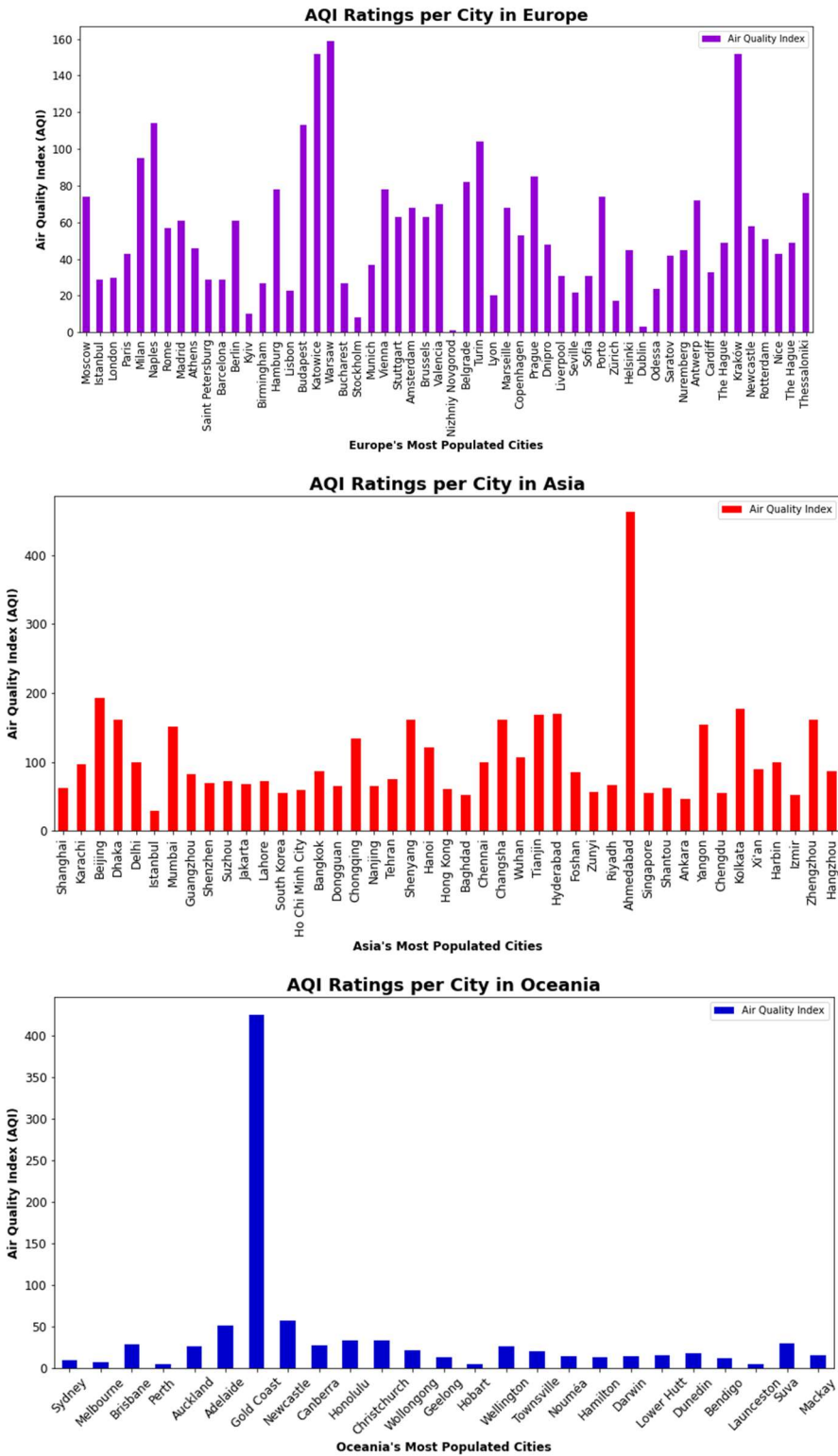
The purpose of this research is to gain an understanding of how pollution varies by continent. The questions that are posed are a small section of what can be asked of the data obtained. Further questions will be suggested upon conclusion of this report which would be worthy of further analysis

- 📊 Does AQI performance correlate to continent?
- 📊 How do the extremities of pollution differ based on location?
- 📊 If pollution levels correlate then which pollutants correlate closely and which do not
- 📊 If PM 25 and PM 10 have a strong correlation then the r value will be over 0.7
- 📊 If SO₂ and PM 25 have a strong correlation then the r value will be over 0.7
- 📊 If O₃ and PM 25 have a strong correlation then the r value will be over 0.7

Data Analyses

1. Bar Graphs showing AQI for cities in each continent

The following bar graphs show the AQI for each city considered in each continent. It shows that some regions are more consistent in their pollutant output with fewer peaks but others are erratic in their AQI. This is demonstrated in the Oceania graph which has a peak reading for AQI for the gold coast, where a more consistent continent in terms of AQI is North America. Asia has the most consistently poor AQI. This could be due to several factors and would be worth further research. A possible suggestion for this could be population density in each city which may be higher than western cities. Typical fuel sources could also affect this too.





2. Pie Charts Showing contents of Air pollution in each continent

The following pie charts illustrate the components of pollution for a typical city in each region. Whilst viewing them it is shown that the content of air pollution do vary depending on the location.

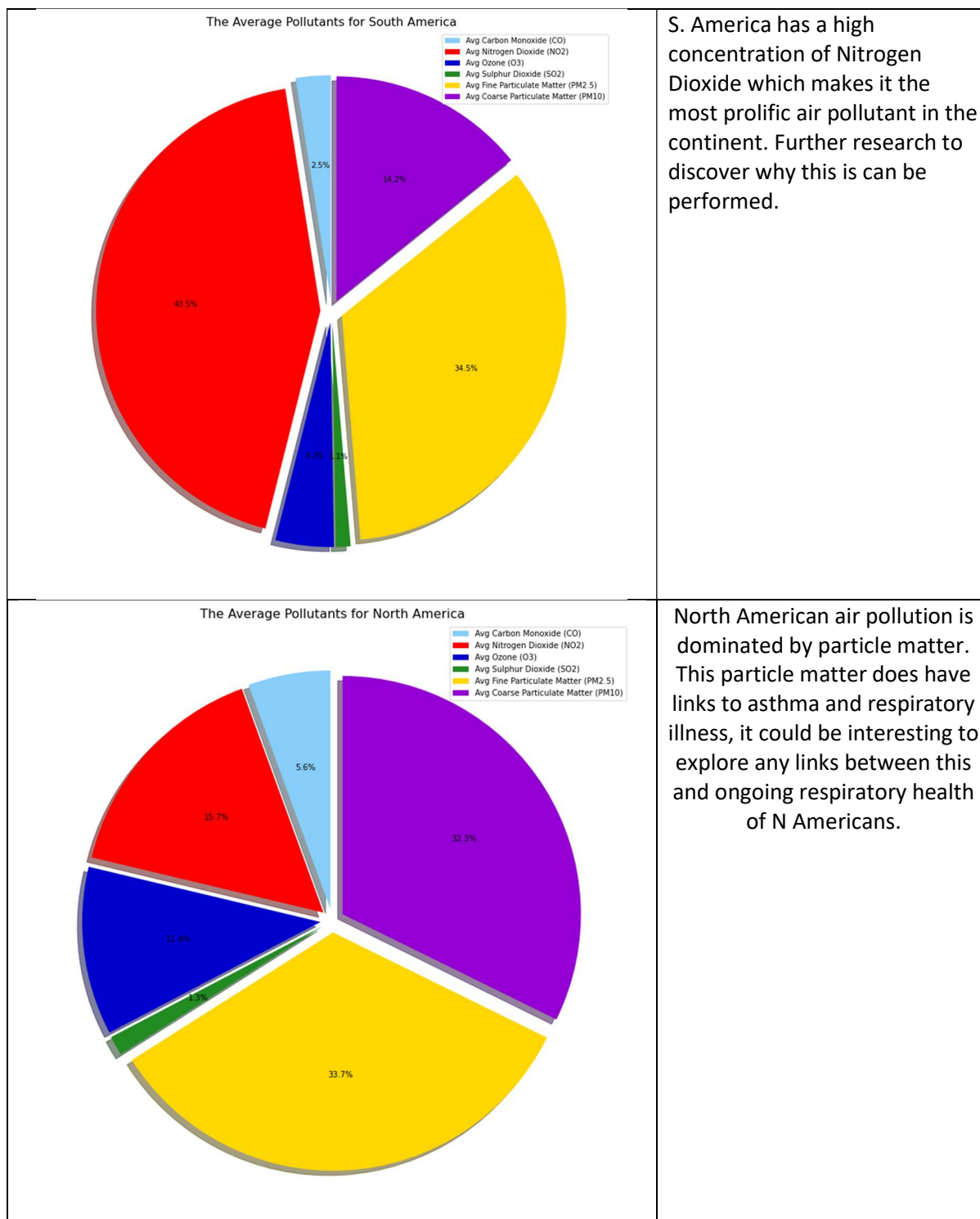


Figure 14 Air pollution proportion N.America & S. America

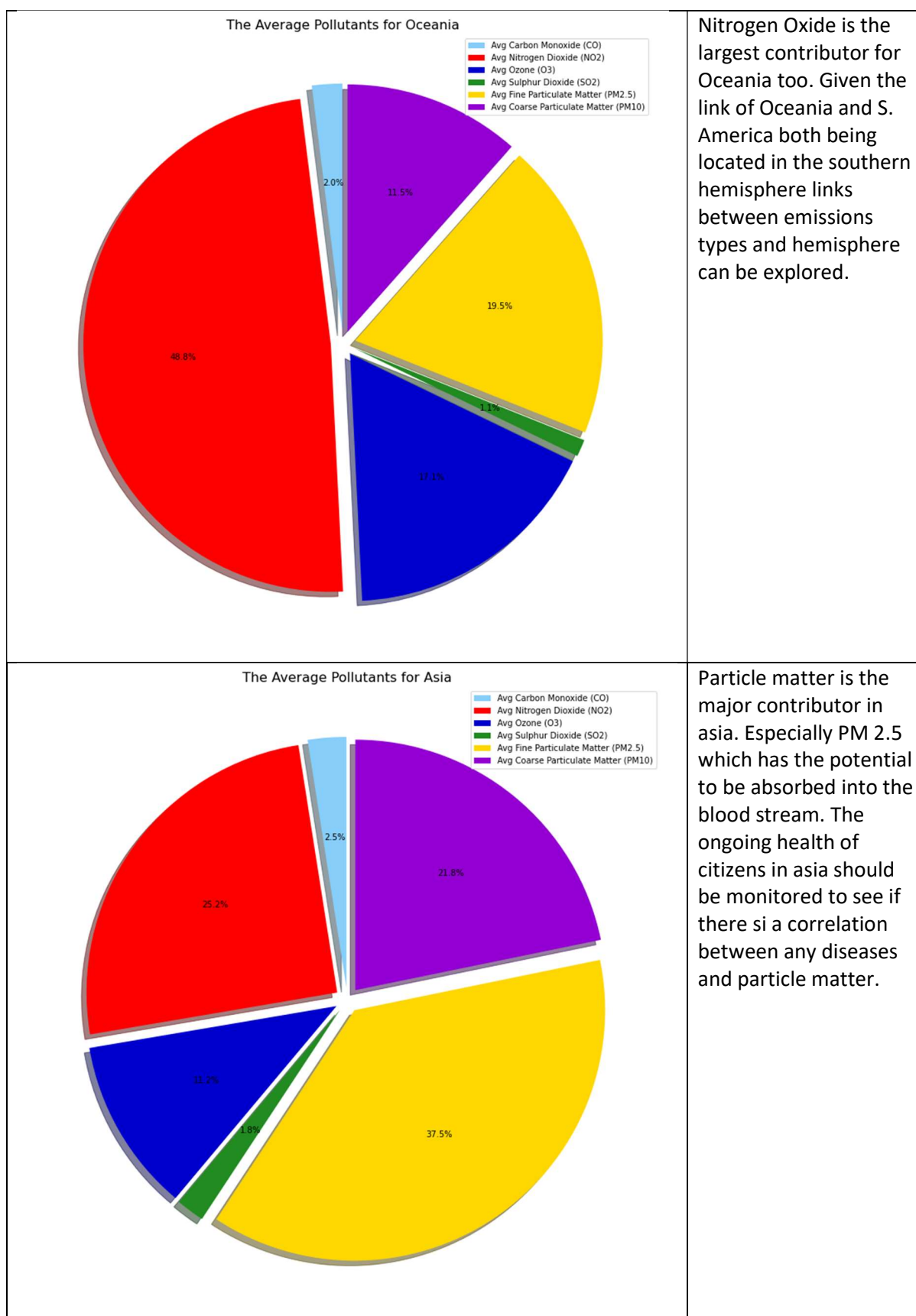
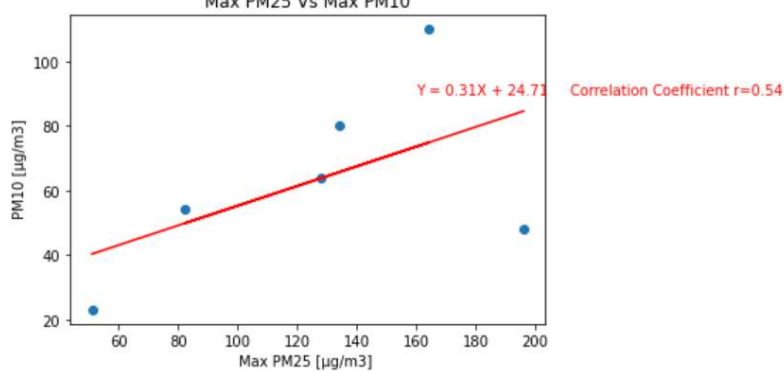
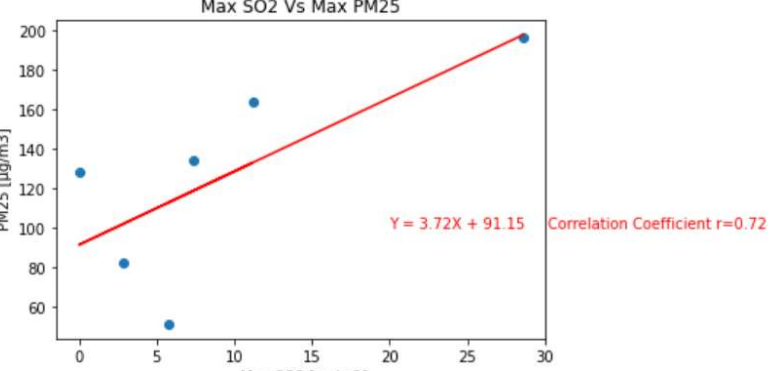
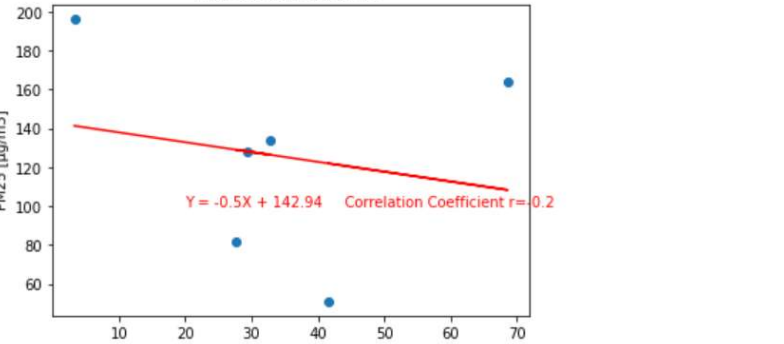


Figure 15 Air pollution proportion Oceania & Asia

3. Scatter plots of Pollutants to check correlation

| | |
|---|--|
|  <p>Max PM25 Vs Max PM10</p> <p>PM10 [$\mu\text{g}/\text{m}^3$]</p> <p>Max PM25 [$\mu\text{g}/\text{m}^3$]</p> <p>$Y = 0.31X + 24.73$</p> <p>Correlation Coefficient $r=0.54$</p> | <p>The correlation between Particle matter pollutants for each continent is relatively weak as it is only 0.54. It is interesting to see that there is only a moderate correlation between particulate matter air pollutants. It was expected that there would be a stronger correlation</p> |
|  <p>Max SO2 Vs Max PM25</p> <p>PM25 [$\mu\text{g}/\text{m}^3$]</p> <p>Max SO2 [$\mu\text{g}/\text{m}^3$]</p> <p>$Y = 3.72X + 91.15$</p> <p>Correlation Coefficient $r=0.72$</p> | <p>The correlation between PM 25 and Sulphur dioxide is over 0.72. The hypotheses of there being a correlation greater than 0.7 is proved valid</p> |
|  <p>Max O3 Vs Max PM25</p> <p>PM25 [$\mu\text{g}/\text{m}^3$]</p> <p>Max O3 [$\mu\text{g}/\text{m}^3$]</p> <p>$Y = -0.5X + 142.94$</p> <p>Correlation Coefficient $r=-0.2$</p> | <p>The correlation between Ozone and PM 25 is very weak, only 0.2. This suggests no correlation, Our hypotheses cannot be verified from the results present further testing will be required to discount it.</p> |

Challenges and Limitations

The data displayed in this report is a cross section of the results generated that were noteworthy. Ultimately the data obtained is limited to live api call and not historical data. This is due to the cost involved of buying historic data for this project. The results presented do show that each continent does have a varying level of AQI and contributing pollutants. It does not show the clear link between any one type of pollutant and the overall AQI. It is proposed in future iterations of this to gather historic data and further check the links between pollutant and AQI.

Conclusion

Whilst this report is able to offer some insight into Air quality globally it only peels back the curtain of the data that can be explored. The visualisations do illustrate that there are links between pollutants and regions that warrant further exploration. For this Historical data for each city would be considered to see how the pollutant levels vary. It has also been postulated that by using historical data it could be possible to find out what effect lockdowns during covid had on AQI. Historical data could also be used to monitor the impact on respiratory health over the past 100 years by checking medical statistics against air quality indicators. If there is a correlation between the pollutants and certain health problems it could be quantified and mitigated potentially.

Pollution data is an interesting resource to analyse as it is readily available for a broad cross section of global cities. It has been sufficient to provide a high level overview of air quality and the parameters impacting it. It has also opened up further elements for discussion which would be elaborated on in future incarnations of AQI research.