

Course Project: Big Data Concepts

Author: Satya Priyanka Ponduru

Email: sponduru@iu.edu

1. Introduction

Air quality is an important topic in today's world because it directly affects both health and the environment. Analyzing air quality data helps us understand how pollutants like CO, Ozone, NO2, and PM2.5 impact the Air Quality Index (AQI). For this project, I worked on a dataset that records air quality levels across various countries. The dataset provides detailed pollutant information, which makes it useful for identifying patterns and understanding air quality trends globally.

The main goal of this project was to explore the dataset using big data technologies and advanced analytical methods. By focusing on how different pollutants contribute to AQI and identifying trends in the data, I aimed to gain insights into air quality management. This project also helped me improve my technical skills by applying concepts like NoSQL databases, distributed processing, and data modeling to solve real-world environmental challenges.

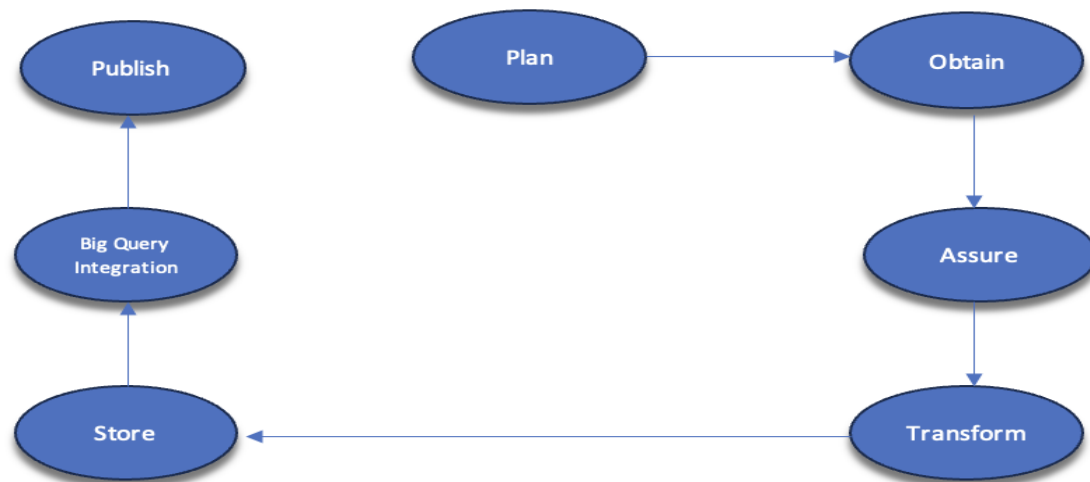
2. Background

The motivation for this project comes from my curiosity to understand how air quality changes across different regions and the factors that influence it. Air quality is a vital environmental parameter that directly impacts human health, ecosystems, and climate. Studying its trends is essential to gaining insights into environmental health and addressing global challenges related to pollution and sustainability. The comprehensive dataset used in this project offers a unique opportunity to explore the relationship between different pollutants, such as CO, Ozone, NO2, and PM2.5, and the Air Quality Index (AQI).

This project allowed me to analyze air quality trends using tools like Google Cloud Platform, Google Colab, BigQuery, and Looker Studio. It was not just about working with data but also about understanding the broader implications of air quality on the environment and public health. By identifying patterns and trends in the data, I aimed to contribute to the growing discussions on global air quality issues and the importance of sustainable practices to mitigate pollution.

3. Methodology

Data Pipeline



1. **Plan:** Set objectives to analyze global air quality trends by aggregating and examining AQI levels across different regions.
2. **Obtain:** Collected air quality data from Kaggle and stored it in a Google Cloud Storage bucket for secure access.
3. **Assure:** Performed data quality checks by handling missing values, correcting errors, and removing outliers to ensure data reliability.
4. **Transform:** Processed the data in Python using Google Colab, normalizing and structuring it for detailed analysis.
5. **Store:** Saved the cleaned dataset back to the GCP bucket for organized and easy accessibility.
6. **BigQuery Integration:** Imported the processed data into BigQuery for efficient querying and advanced analytical tasks.
7. **Publish:** Shared the cleaned dataset, Python scripts, and SQL queries on GitHub to encourage collaboration and further exploration.

3.1 Plan

The main goal of this project was to analyze air quality trends across different countries and understand the impact of pollutants on the Air Quality Index (AQI). To achieve this, I designed a structured plan with the following objectives:

1. **Objective 1:** Collect a detailed dataset containing AQI values, pollutant levels (CO, Ozone, NO₂, PM_{2.5}), and geographic information such as country, city, and coordinates.
2. **Objective 2:** Perform data cleaning by addressing missing values and removing outliers to ensure the reliability of the data for analysis.
3. **Objective 3:** Add new features like pollutant categories and AQI classifications to enhance the depth of analysis and provide meaningful insights.

4. **Objective 4:** Use cloud-based tools like Google Cloud Platform and BigQuery for efficient data storage, querying, and processing.
5. **Objective 5:** Create interactive dashboards using Looker Studio to present the findings in a visually appealing and meaningful way.

Utilized advanced cloud technologies like **Google Cloud Platform (GCP)** for storage, **Google Colab** for preprocessing, **BigQuery** for querying, and **Looker Studio** for visualization.

3.2 Obtain

The first step in this project was to find a dataset that matched the objectives of analyzing global air quality trends and understanding the role of pollutants. After exploring multiple options on Kaggle, I chose the “**AQI and Lat Long of Countries**” dataset. This dataset was selected because it provides a detailed and organized collection of air quality measurements across various countries and cities.

The dataset contains **16,695 records** and includes the following important features:

- **Basic Information:** Details about the country and city where the air quality data was recorded.
- **AQI Values:** Information on the Air Quality Index (AQI), which helps in understanding air quality levels and trends.
- **Geographical Coordinates:** Latitude and longitude fields that allow for a geographic analysis of air quality variations.
- **Pollutant-Specific Data:** Data for individual pollutants such as CO, Ozone, NO2, and PM2.5, categorized into levels like "Good," "Moderate," and "Unhealthy."

This dataset offered a comprehensive view of air quality across different regions, making it suitable for in-depth analysis. It provided me with an excellent opportunity to understand how geographic and environmental factors impact air quality globally. With this rich dataset, I was ready to begin my exploration of air quality trends and pollutant contributions.

3.3 Assure - Infrastructure Setup

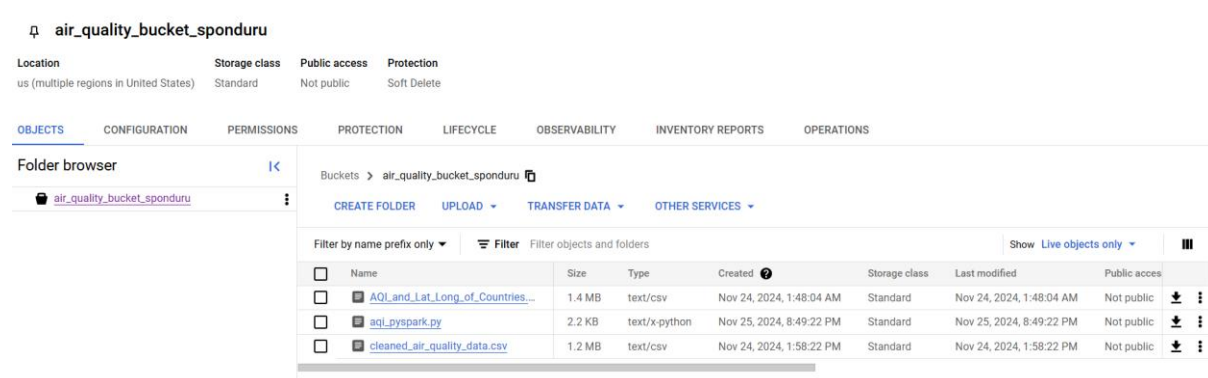
After acquiring the dataset, the next crucial step was to establish a robust infrastructure to ensure secure and organized data storage. I set up a **Google Cloud Platform (GCP)** storage bucket named "**air_quality_bucket_1**". This bucket served as a centralized repository for the dataset, enabling easy access and management throughout the project.

By using GCP, I ensured:

- **Security:** The dataset was stored in a secure and reliable cloud environment, safeguarding it from accidental loss or corruption.

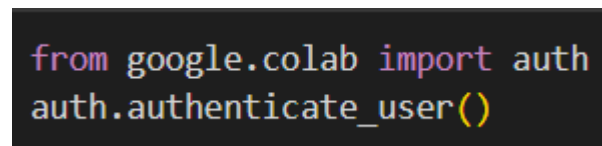
- **Accessibility:** The cloud-based storage allowed seamless access from multiple platforms, including Google Colab and BigQuery.
- **Consistency:** The centralized setup ensured that all preprocessing, analysis, and visualization tasks used the same version of the dataset, maintaining data integrity.

This infrastructure setup laid the foundation for efficient data processing and analysis, leveraging the scalability and reliability of GCP for handling large datasets.



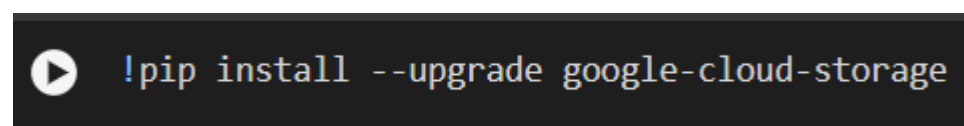
3.4 Transform – Data Preprocessing and Analysis

After successfully uploading the dataset to the **Google Cloud Platform (GCP) bucket**, I accessed it using **Google Colab** to take advantage of the cloud's computational resources for efficient data processing. I established a connection between Google Colab and GCP, downloaded the dataset, and performed several preprocessing steps to clean and prepare the data for analysis.



Accessing the GCP Bucket

1. I first installed the required Google Cloud libraries in the Colab environment to ensure smooth integration.
2. After authenticating my Google account, I connected Colab to the GCP bucket named **'air_quality_bucket_1'**.
3. Using references like **blob objects**, I downloaded the dataset **AQI_and_Lat_Long_of_Countries.csv** directly into my Colab notebook.
4. The dataset was loaded into a **pandas DataFrame**, making it ready for preprocessing.



Preprocessing Steps

The preprocessing phase was an essential step to ensure data quality and reliability. The following steps were carried out:

1. **Handling Missing Values:** The 'Country' column had 302 missing entries.

To address this:

- I initially replaced missing values with 'Unknown' to maintain the dataset's structure.
- Then, using a **city-to-country mapping** derived from existing data, I filled in the missing values wherever possible. The remaining entries were left as 'Unknown.'

```
[ ] print(air_quality_df.isnull().sum())
```

Country	302
City	0
AQI Value	0
AQI Category	0
CO AQI Value	0
CO AQI Category	0
Ozone AQI Value	0
Ozone AQI Category	0
NO2 AQI Value	0
NO2 AQI Category	0
PM2.5 AQI Value	0
PM2.5 AQI Category	0
lat	0
lng	0
dtype: int64	

```
[ ] air_quality_df = air_quality_df.dropna(subset=['Country'])
```

2. **Removing Outliers:** Outliers in numerical columns were identified during the data review. Using **percentile filtering** (retaining values within the 0.05 to 0.995 range), I removed these anomalies to normalize the dataset and improve the reliability of the analysis.

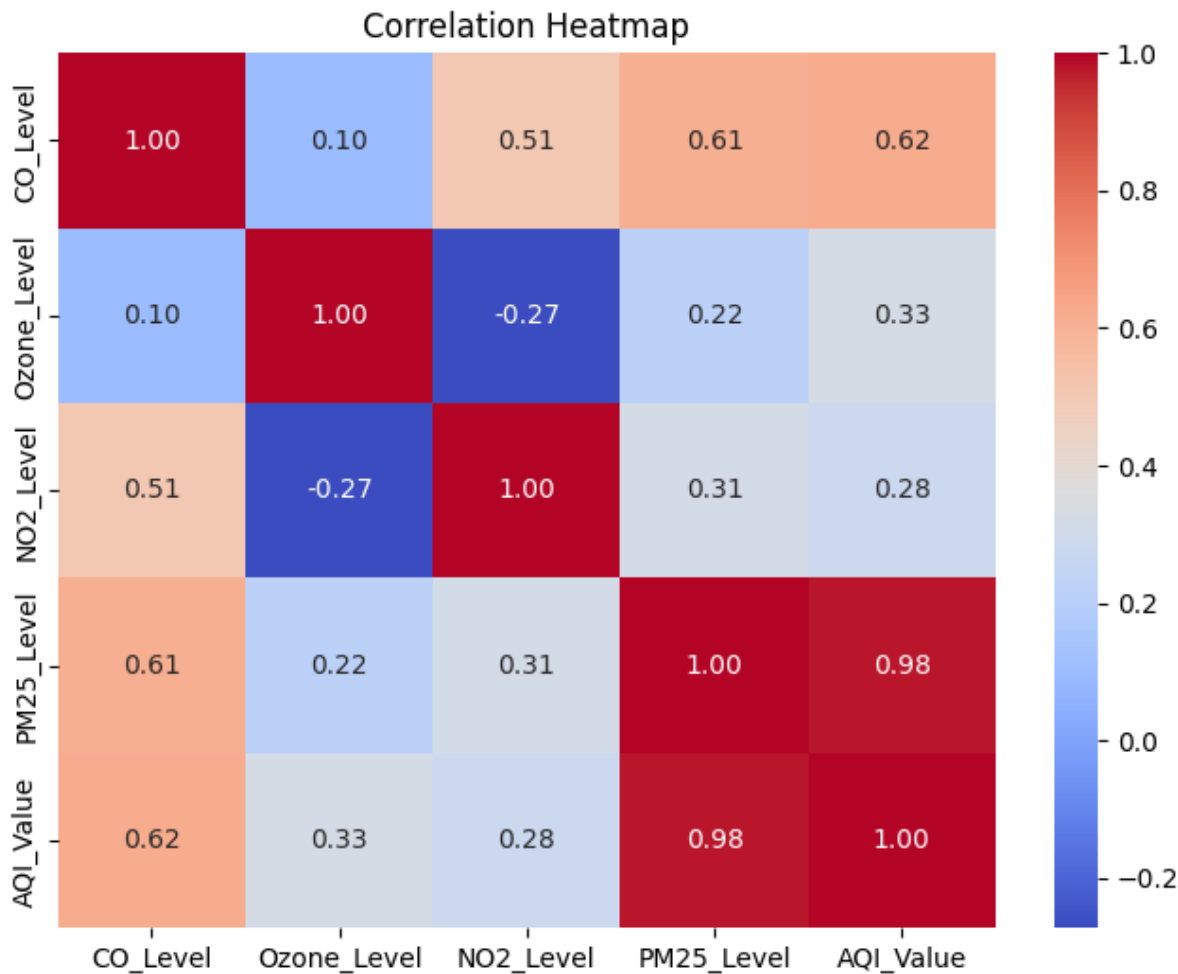
```
# Define pollutant columns
pollutant_columns = ['CO AQI Value', 'Ozone AQI Value', 'NO2 AQI Value', 'PM2.5 AQI Value']

# Remove outliers beyond the 1th and 99th percentiles
for col in pollutant_columns:
    lower_bound = air_quality_df[col].quantile(0.01)
    upper_bound = air_quality_df[col].quantile(0.995)
    air_quality_df = air_quality_df[(air_quality_df[col] >= lower_bound) & (air_quality_df[col] <= upper_bound)]

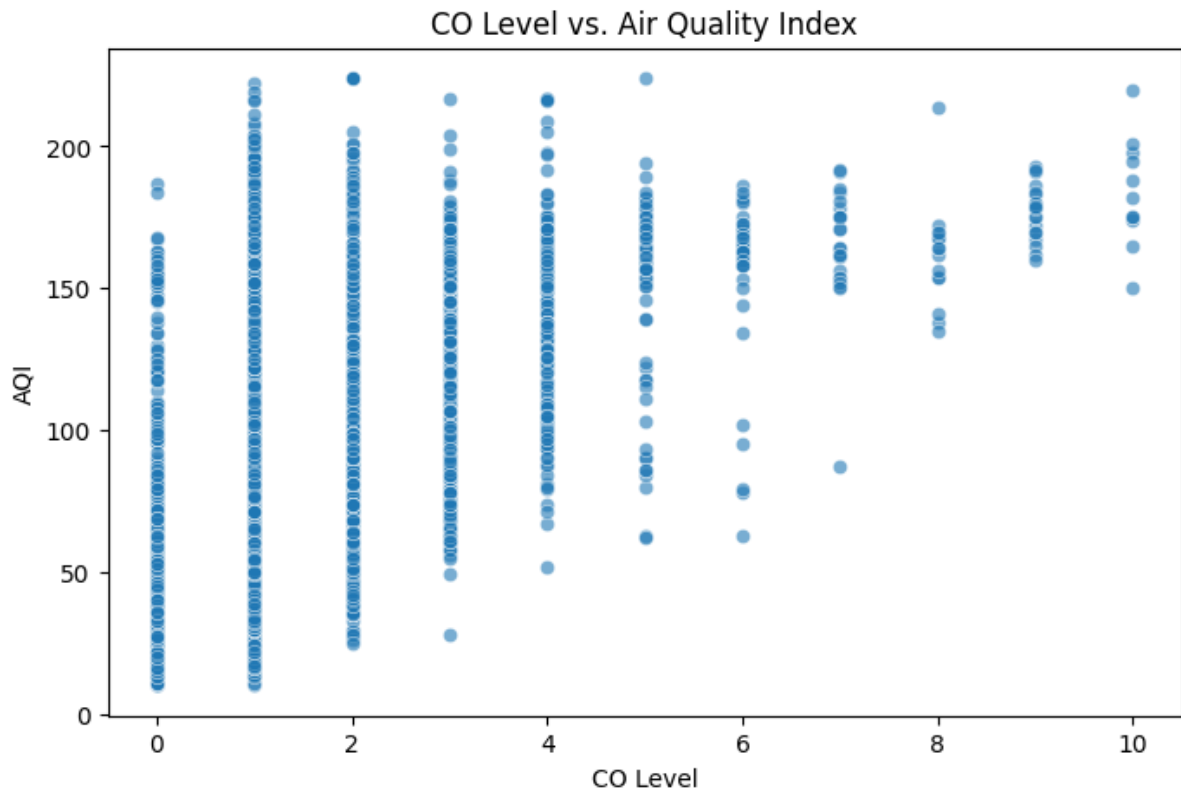
# Verify changes
print(f"Dataset shape after outlier removal: {air_quality_df.shape}")
```

Dataset shape after outlier removal: (15784, 14)

3. **Analyzing Correlations:** I created a **correlation heatmap** to explore relationships between pollutants like CO, NO2, PM2.5, and Ozone. This step provided insights into how these pollutants interact and affect the AQI.



4. **Exploring Data Distributions:**
- Using **scatter plot matrices**, I examined how pollutants such as CO and PM2.5 correlated with one another across different conditions. These visualizations helped identify patterns and potential outliers.
 - **Box plots and histograms** were also generated to better understand the distribution of numerical attributes in the dataset.



5. **Finalizing the Dataset:** After completing the preprocessing steps, I saved the cleaned data as **cleaned_AQI_and_Lat_Long_of_Countries.csv** and uploaded it back to the GCP bucket for further analysis.

Outcome

The preprocessing steps ensured that the data was clean, structured, and ready for advanced analysis. This refined dataset provided a solid foundation for querying in BigQuery and creating meaningful visualizations in Looker Studio, allowing me to uncover patterns and insights about air quality trends across different regions.

3.5 Store

After completing the preprocessing steps, I saved the cleaned dataset, **cleaned_air_quality_data.csv**, back to the **Google Cloud Platform (GCP)** bucket named **'air_quality_bucket_1'**.

This step was important because:

- **Data Security:** Storing the processed data in the GCP bucket ensured it was safe from accidental deletion or corruption.
- **Ease of Access:** The centralized storage made it easy to access the data for further analysis and querying in BigQuery.
- **Organized Workflow:** Having the cleaned data stored in a structured and accessible location helped maintain consistency throughout the project.

This setup ensured that the dataset was readily available for the next stages of the project, like querying and creating visualizations, without any interruptions.

3.6 BigQuery Integration and Visualization in Looker Studio

The next important step in this project was to integrate the cleaned dataset into **BigQuery** to perform detailed analysis and generate insights. The preprocessed data, which was securely stored in the **Google Cloud Platform (GCP) bucket**, was imported into a new dataset and table in BigQuery. This integration was a key step as it allowed me to efficiently explore and analyze the data using SQL queries.

BigQuery Integration Process

1. I first created a new dataset in BigQuery to store the cleaned data.
2. The file **cleaned_AQI_and_Lat_Long_of_Countries.csv** was uploaded directly from the GCP bucket into this dataset as a table.
3. During the upload, I carefully ensured that the schema matched correctly, verifying that all columns and data types were consistent with the original dataset structure to avoid errors.
4. Once the data was successfully imported, I utilized BigQuery's SQL capabilities to perform in-depth analysis and prepare the data for visualization.

This integration with BigQuery not only made it easier to run complex queries but also streamlined the process of connecting the dataset to **Looker Studio** for creating visualizations and dashboards.

3.6 BigQuery Integration and Visualization in Looker Studio

The next important step in this project was to integrate the cleaned dataset into **BigQuery** to perform detailed analysis and generate insights. The preprocessed data, which was securely stored in the **Google Cloud Platform (GCP) bucket**, was imported into a new dataset and table in BigQuery. This integration was a key step as it allowed me to efficiently explore and analyze the data using SQL queries.

BigQuery Integration Process

1. I first created a new dataset in BigQuery to store the cleaned data.
2. The file **cleaned_AQI_and_Lat_Long_of_Countries.csv** was uploaded directly from the GCP bucket into this dataset as a table.
3. During the upload, I carefully ensured that the schema matched correctly, verifying that all columns and data types were consistent with the original dataset structure to avoid errors.

- This integration with BigQuery not only made it easier to run complex queries but also streamlined the process of connecting the dataset to **Looker Studio** for creating visualizations and dashboards.

To extract meaningful insights from the dataset, the following SQL queries were executed in **BigQuery**

1. Average AQI by Country

RUN
 SAVE QUERY ▾
 DOWNLOAD
 SHARE ▾
 SCHEDULE
 OPEN IN ▾
 MORE ▾
 This query

```

1 SELECT
2   Country,
3   AVG(AQI_Value) AS Avg_AQI
4 FROM
5   `aqi-analysis-project.air_quality_analysis.final_air_quality`
6 GROUP BY
7   Country
8 ORDER BY
9   Avg_AQI DESC;
10

```

Press Alt+F1 for Accessibility Options

Query results

SAVE RESULTS ▾
 EXPLORE DATA ▾

JOB INFORMATION	RESULTS	CHART	JSON	EXECUTION DETAILS	EXECUTION GRAPH
	Row Country ▾ Avg_AQI ▾				
1	Mauritania	157.0			
2	Saudi Arabia	147.0			
3	Gambia	146.2			
4	Senegal	137.7272727272...			
5	Pakistan	133.6666666666...			
6	Yemen	131.0			
7	Iraq	127.5			
8	Guinea-Bissau	120.0			
9	Haiti	116.4285714285...			
10	Uzbekistan	114.0000000000...			
11	Tajikistan	111.6153846153...			
12	Dominican Republic	108.0384615384...			
13	Turkmenistan	106.6666666666...			

Query 2: Top Pollutants Contributing to AQI

2. Top Pollutants Contributing to AQI

RUN

SAVE QUERY

DOWNLOAD

SHARE

SCHEDULE

OPEN IN

⋮

This q...

1 SELECT

2 Country,

3 AVG(CO_Level) AS Avg_CO,

4 AVG(NO2_Level) AS Avg_NO2,

5 AVG(Ozone_Level) AS Avg_Ozone,

6 AVG(PM25_Level) AS Avg_PM25,

7 AVG(AQI_Value) AS Avg_AQI

8 FROM

9 'aqi-analysis-project.air_quality_analysis.final_air_quality'

10 GROUP BY

11 Country

12 ORDER BY

13 Avg_AQI DESC;

Press Alt+F1 for Accessibility Options

Query results

SAVE RESULTSEXPLORE DATA

JOB INFORMATIONRESULTSCHARTJSONEXECUTION DETAILSEXECUTION GRAPH

Row	Country	Avg_CO	Avg_NO2	Avg_Ozone	Avg_PM25	Avg_AQI
1	Mauritania	0.0	0.0	36.0	157.0	157.0
2	Saudi Arabia	1.0	0.0	59.0	147.0	147.0
3	Gambia	1.0	1.0	23.4	146.2	146.2
4	Senegal	1.272727272727...	1.727272727272...	21.72727272727...	137.7272727272...	137.7272727272...
5	Pakistan	0.888888888888...	0.611111111111...	44.61111111111...	132.5555555555...	133.6666666666...
6	Yemen	1.0	0.0	44.33333333333...	131.0	131.0
7	Iraq	1.0	1.5	77.0	127.5	127.5
8	Guinea-Bissau	1.0	1.0	27.0	120.0	120.0
9	Haiti	1.857142857142...	4.142857142857...	20.57142857142...	116.4285714285...	116.4285714285...
10	Uzbekistan	1.041666666666...	2.958333333333...	45.37500000000...	114.0000000000...	114.0000000000...
11	Tajikistan	1.0	0.153846153846...	45.30769230769...	111.6153846153...	111.6153846153...

Query 3: Countries with Highest Percent Poor Air Quality

3. Countries with Highest Percent Poor Air Quali...

RUN

SAVE QUERY

DOWNLOAD

SHARE

SCHEDULE

⋮

This query will p

1 SELECT

2 Country,

3 COUNT(CASE WHEN AQI_Value > 100 THEN 1 END) * 100.0 / COUNT(*) AS Percentage_Unhealthy_AQI

4 FROM

5 'aqi-analysis-project.air_quality_analysis.final_air_quality'

6 GROUP BY

7 Country

8 ORDER BY

9 Percentage_Unhealthy_AQI DESC;

10

Press Alt+F1 for Accessibility Options

Query results

SAVE RESULTSEXPLORE DATA

JOB INFORMATIONRESULTSCHARTJSONEXECUTION DETAILSEXECUTION GRAPH

Row	Country	Percentage_Unhealthy
1	Senegal	100.0
2	Saudi Arabia	100.0
3	Mauritania	100.0
4	Gambia	100.0
5	Guinea-Bissau	100.0
6	Iraq	100.0
7	Haiti	85.71428571428...
8	Pakistan	83.33333333333...
9	Turkmenistan	66.66666666666...
10	Uzbekistan	66.66666666666...
11	Yemen	66.66666666666...
12	Jamaica	66.66666666666...
13	Tajikistan	61.53846153846...

Query 4: Regional Trends in Air Quality Index

4. Regional Trends in AQI

RUN

SAVE QUERY

DOWNLOAD

SHARE

SCHEDULE

OPEN IN

```
1 SELECT
2   CASE
3     WHEN Lat BETWEEN -90 AND -30 THEN 'Southern Hemisphere'
4     WHEN Lat BETWEEN -30 AND 30 THEN 'Equatorial Region'
5     WHEN Lat BETWEEN 30 AND 90 THEN 'Northern Hemisphere'
6     ELSE 'Unknown Region'
7   END AS Region,
8   AVG(AQI_Value) AS Avg_AQI
9 FROM
10  `aqi-analysis-project.air_quality_analysis.final_air_quality`
11 GROUP BY
12   Region
13 ORDER BY
14   Avg_AQI DESC;
15
```

Query results

SAVE RESULTS

JOB INFORMATIONRESULTSCHARTJSONEXECUTION DETAILSEXECUTION GRAPH

Row	Region	Avg_AQI
1	Equatorial Region	61.57145950627...
2	Northern Hemisphere	51.93566602803...
3	Southern Hemisphere	44.94117647058...

Query 5: Air Quality Index Distribution by AQI Category

5. AQI Distribution by AQI Category

RUN

SAVE QUERY

DOWNLOAD

SHARE

SCHEDULE

```
1 SELECT
2   CASE
3     WHEN AQI_Value <= 50 THEN 'Good'
4     WHEN AQI_Value <= 100 THEN 'Moderate'
5     WHEN AQI_Value <= 150 THEN 'Unhealthy for Sensitive Groups'
6     WHEN AQI_Value <= 200 THEN 'Unhealthy'
7     WHEN AQI_Value <= 300 THEN 'Very Unhealthy'
8     ELSE 'Hazardous'
9   END AS AQI_Category,
10  COUNT(*) AS Count
11 FROM
12  `aqi-analysis-project.air_quality_analysis.final_air_quality`
13 GROUP BY
14   AQI_Category
15 ORDER BY
16   Count DESC;
17
```

Query results


SAVE RESULTS

JOB INFORMATIONRESULTSCHARTJSONEXECUTION DETAILSEXECUTION GRAPH






Row	AQI_Category	Count
1	Good	7338
2	Moderate	6635
3	Unhealthy for Sensitive Groups	604
4	Unhealthy	213

Query 6: Cities with the Highest Pollutant Levels

	6. Cities with the Highest Pollutant Levels	 RUN	 SAVE QUERY ▾	 DOWNLOAD	 SHARE ▾	 SC
1	SELECT					
2	City,					
3	AVG(CO_Level) AS Avg_CO,					
4	AVG(NO2_Level) AS Avg_NO2,					
5	AVG(Ozone_Level) AS Avg_Ozone,					
6	AVG(PM25_Level) AS Avg_PM25,					
7	AVG(CO_Level)+AVG(NO2_Level)+AVG(Ozone_Level)+AVG(PM25_Level)/4 AS Avg_Pollutants					
8	FROM					
9	`aqi-analysis-project.air_quality_analysis.final_air_quality`					
10	GROUP BY					
11	City					
12	ORDER BY					
13	6 DESC					
14	LIMIT					
15	10;					
16						

Query results								 SAVE RESULTS ▾
JOB INFORMATION		RESULTS	CHART	JSON	EXECUTION DETAILS	EXECUTION GRAPH		
Row	City ▾	Avg_CO ▾	Avg_NO2 ▾	Avg_Ozone ▾	Avg_PM25 ▾	Avg_Pollutants ▾		
1	Chicholi	2.0	0.0	87.0	160.0	129.0		
2	Shenyang	4.0	6.0	85.0	125.0	126.25		
3	Jamshedpur	3.0	1.0	83.0	155.0	125.75		
4	Tekkali	3.0	1.0	82.0	155.0	124.75		
5	Sompeta	2.0	1.0	83.0	154.0	124.5		
6	Xinyu	4.0	3.0	89.0	107.0	122.75		
7	Salar	1.0	3.0	81.0	150.0	122.5		
8	Sibi	1.0	0.0	81.0	159.0	121.75		

Query 7: Correlation Between Pollutants

	7. Correlation Between Pollutants	 RUN	 SAVE QUERY ▾	 DOWNLOAD	 SHAR
1	SELECT				
2	CORR(AQI_Value, CO_Level) AS Correlation_AQI_CO,				
3	CORR(AQI_Value, NO2_Level) AS Correlation_AQI_NO2,				
4	CORR(AQI_Value, Ozone_Level) AS Correlation_AQI_Ozone,				
5	CORR(AQI_Value, PM25_Level) AS Correlation_AQI_PM25				
6	FROM				
7	`aqi-analysis-project.air_quality_analysis.final_air_quality`;				
8					

Query results

JOB INFORMATION		RESULTS	CHART	JSON	EXECUTION DETAILS	EXECUTION GRAPH
Row	Correlation_AQI_CO	Correlation_AQI_NO2	Correlation_AQI_Ozone	Correlation_AQI_PM25		
1	0.5381304806495123	0.264180591390421...	0.0825941923993178...	0.9710125926272074		

Query 8: Top Cities with the Unhealthiest Days

8. Top Cities with the Most Unhealthy Days

RUN

SAVE QUERY

DOWNLOAD

SHARE

SC

```
1 SELECT
2   City,
3   COUNT(CASE WHEN AQI_Value > 100 THEN 1 END) AS Unhealthy_Days
4 FROM
5   `aqi-analysis-project.air_quality_analysis.final_air_quality`
6 GROUP BY
7   City
8 ORDER BY
9   Unhealthy_Days DESC
10 LIMIT
11   21;
12
```

Query results

SAVE RESULTS

JOB INFORMATIONRESULTSCHARTJSONEXECUTION DETAILSEXECUTION GRAPH

Row	City	Unhealthy_Days
1	Santa Ana	15
2	San Miguel	11
3	San Luis	10
4	Valencia	8
5	Santa Fe	7
6	Quezon	6
7	Lima	3
8	San Lorenzo	3
9	Rio Claro	3
10	San Salvador	3
11	Brownsville	3
12	Frontera	3

Query 9: Top Cities with the Best Air Quality

9. Top Cities with the Best Air Quality

RUN

SAVE QUERY

DOWNLOAD

SHARE

SCHEDULE

```
1 SELECT
2   City,
3   COUNT(CASE WHEN AQI_Value <= 50 THEN 1 END) * 100.0 / COUNT(*) AS Percentage_Good_AQI,
4   AVG(CO_Level) AS Avg_CO,
5   AVG(NO2_Level) AS Avg_NO2,
6   AVG(Ozone_Level) AS Avg_Ozone,
7   AVG(PM25_Level) AS Avg_PM25,
8   AVG(CO_Level)+AVG(NO2_Level)+AVG(Ozone_Level)+AVG(PM25_Level)/4 AS Avg_Pollutants
9 FROM
10  `aqi-analysis-project.air_quality_analysis.final_air_quality`
11 GROUP BY
12   City
13 ORDER BY
14   Percentage_Good_AQI DESC,
15   Avg_Pollutants ASC
16 LIMIT 10;
17
```

Query results

SAVE RESULTSEXP

JOB INFORMATIONRESULTSCHARTJSONEXECUTION DETAILSEXECUTION GRAPH

Row	City	Percentage_Good_AQI	Avg_CO	Avg_NO2	Avg_Ozone	Avg_PM25	Avg_Pollutants
1	Huancavelica	100.0	0.0	1.0	2.0	10.0	5.5
2	Leticia	100.0	1.0	0.0	2.0	14.0	6.5
3	Carauari	100.0	1.0	0.0	3.0	16.0	8.0
4	Caranavi	100.0	1.0	0.0	3.0	17.0	8.25
5	Tabatinga	100.0	1.0	0.0	2.0	21.0	8.25
6	Senador Guiomard	100.0	1.0	0.0	4.0	14.0	8.5
7	Nueva Loja	100.0	1.0	0.0	5.0	10.0	8.5
8	Sena Madureira	100.0	1.0	0.0	4.0	14.0	8.5

Query 10: Distribution of AQI and Pollutant Categories by Country

```
10. Distribution of AQI and Pollutant Categories ... RUN SAVE QUERY DOWNLOAD SHARE SCHEDULE Query c

1 SELECT
2   Country,
3   'AQI Category',
4   COUNT(*) AS AQI_Category_Count,
5   'CO AQI Category' AS CO_Category,
6   COUNT(CASE WHEN 'CO AQI Category' IS NOT NULL THEN 1 END) AS CO_Category_Count,
7   'NO2 AQI Category' AS NO2_Category,
8   COUNT(CASE WHEN 'NO2 AQI Category' IS NOT NULL THEN 1 END) AS NO2_Category_Count,
9   'Ozone AQI Category' AS Ozone_Category,
10  COUNT(CASE WHEN 'Ozone AQI Category' IS NOT NULL THEN 1 END) AS Ozone_Category_Count,
11  PM25_AQI_Category AS PM25_Category,
12  COUNT(CASE WHEN PM25_AQI_Category IS NOT NULL THEN 1 END) AS PM25_Category_Count
13 FROM
14   'aqi-analysis-project.air_quality_analysis.final_air_quality'
15 GROUP BY
16   Country, CO_Category, NO2_Category, Ozone_Category, PM25_Category
17 ORDER BY
18   AQI_Category_Count DESC, Country;
```

Query results

with the Best Air Quality in current tab open in new tab, Tap - open in split tab)		RESULTS	CHART	JSON	EXECUTION DETAILS	EXECUTION GRAPH		
		f0_			AQI_Category_Count	CO_Category	CO_Category_Count	NO2_Category
1	United States of America	AQI Category			2031	CO AQI Category	2031	NO2 AQI Category
2	United States of America	AQI Category			1455	CO AQI Category	1455	NO2 AQI Category
3	Russian Federation	AQI Category			692	CO AQI Category	692	NO2 AQI Category
4	Brazil	AQI Category			650	CO AQI Category	650	NO2 AQI Category
5	Germany	AQI Category			594	CO AQI Category	594	NO2 AQI Category
6	Italy	AQI Category			507	CO AQI Category	507	NO2 AQI Category
7	Germany	AQI Category			478	CO AQI Category	478	NO2 AQI Category

4. Results

Bigquery Integration and Looker Studio Visualizations

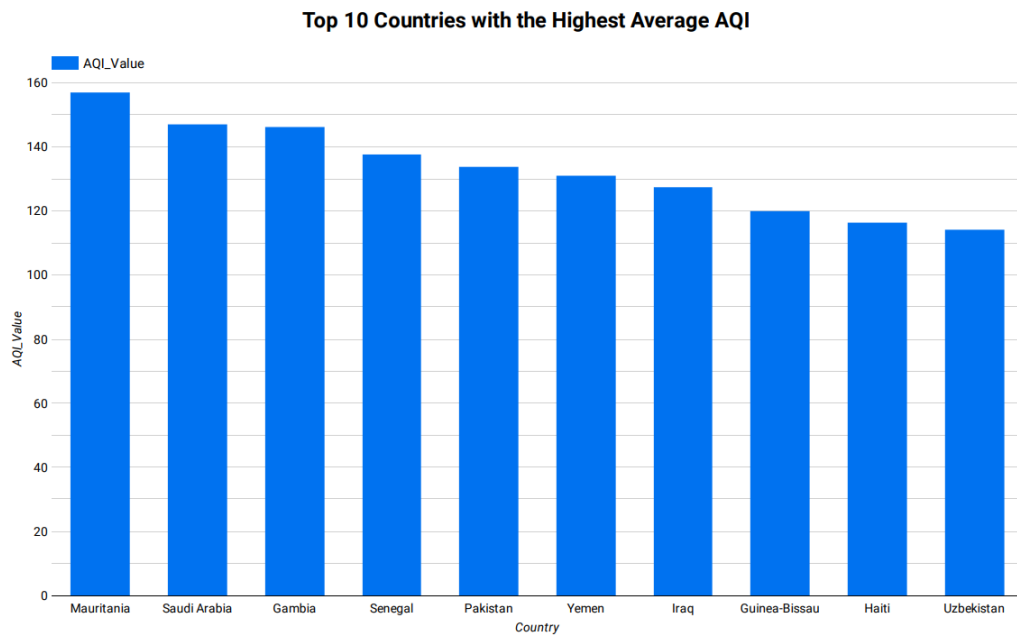


Figure 1

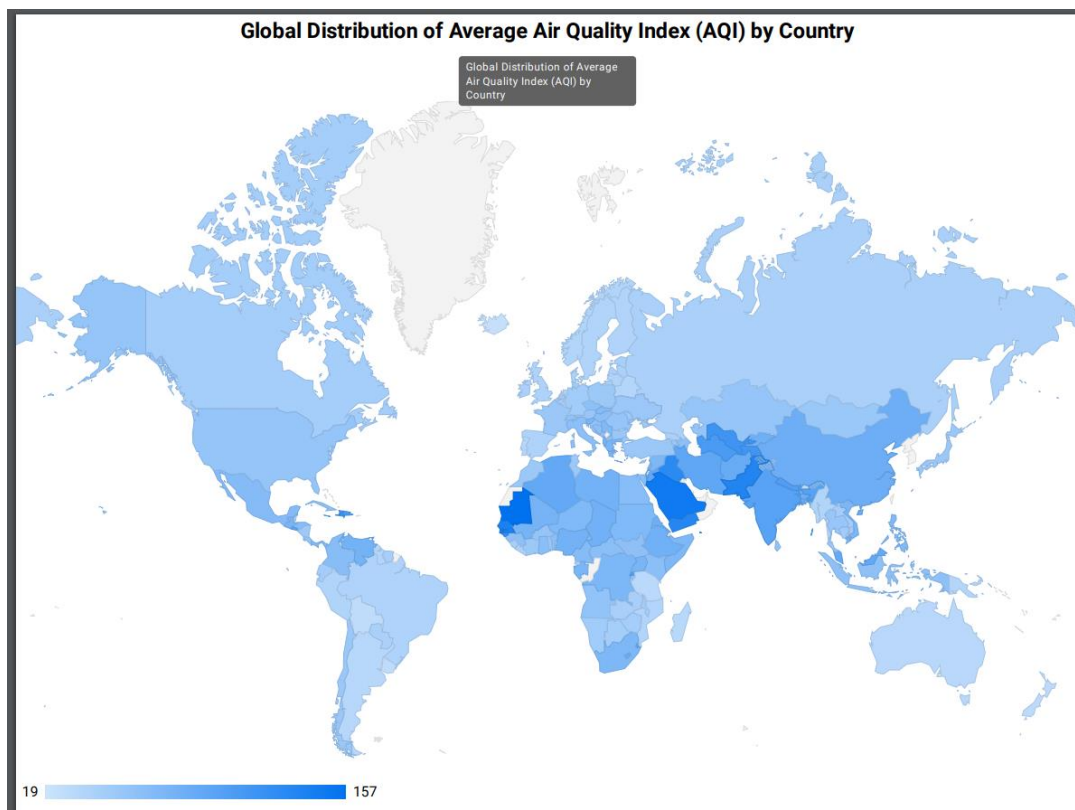


Figure 2

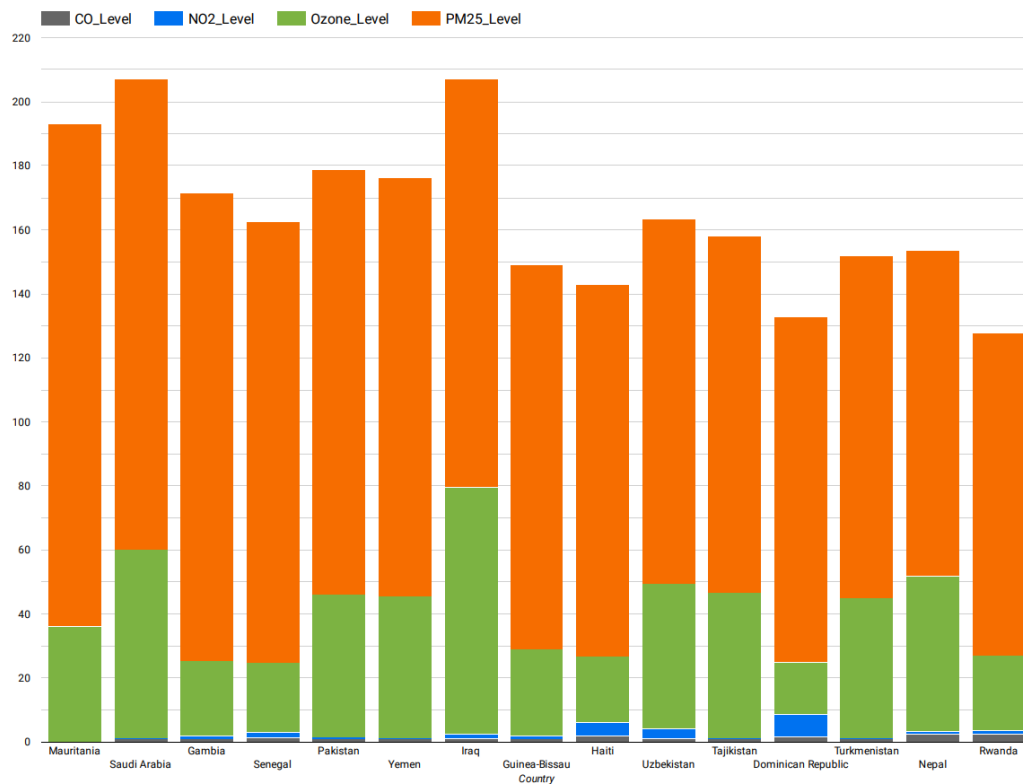


Figure 3

Geographic Distribution of Unhealthy Air Quality Percentages by Country

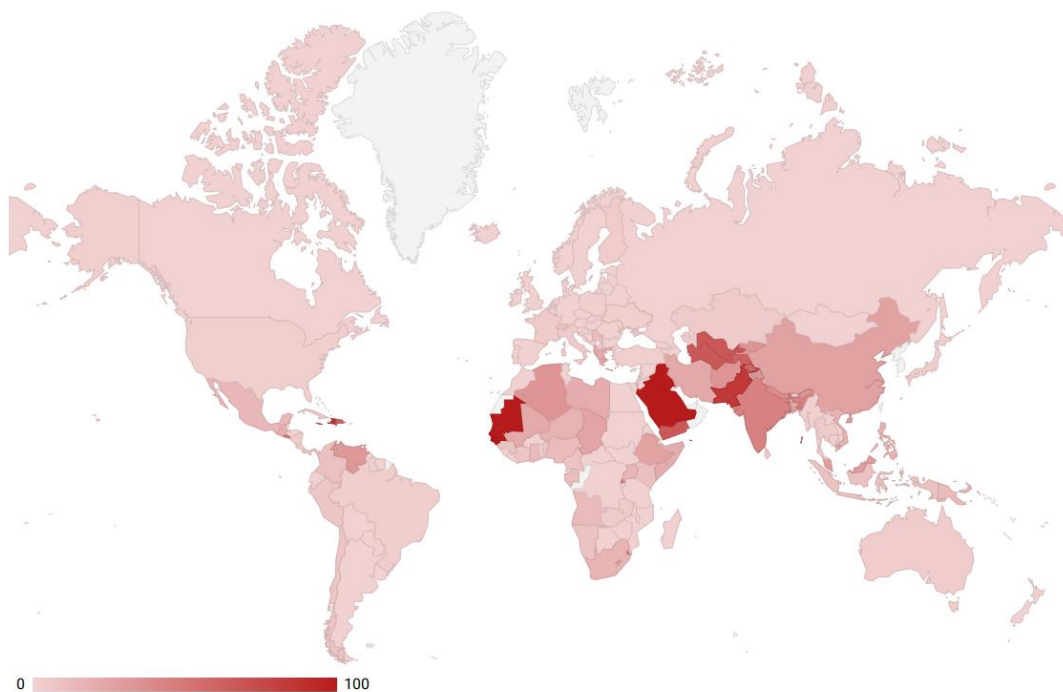


Figure 4

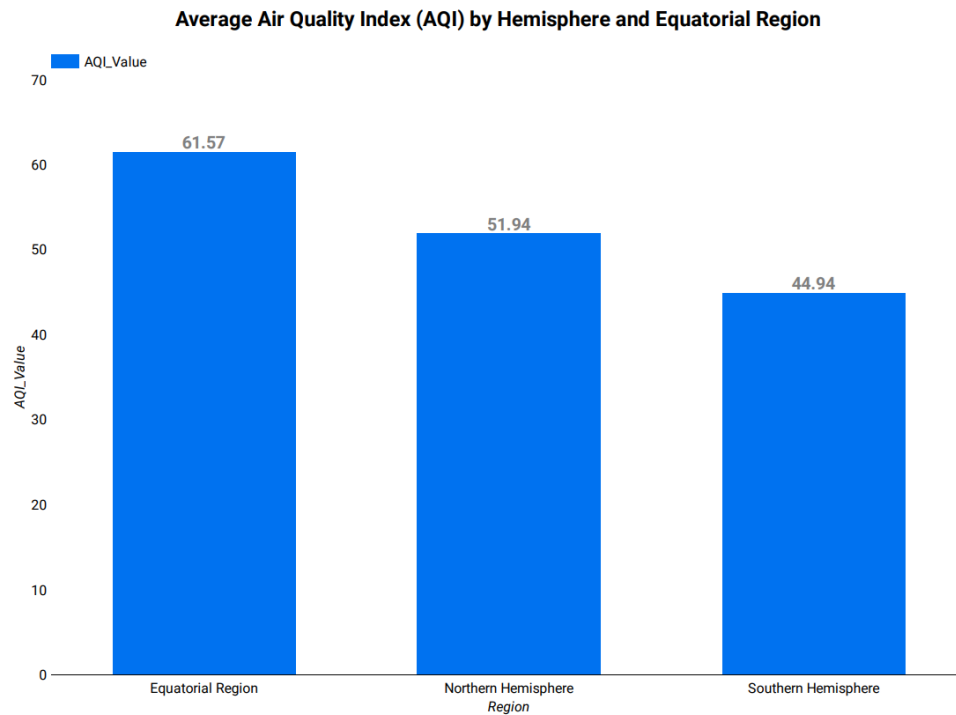


Figure 5

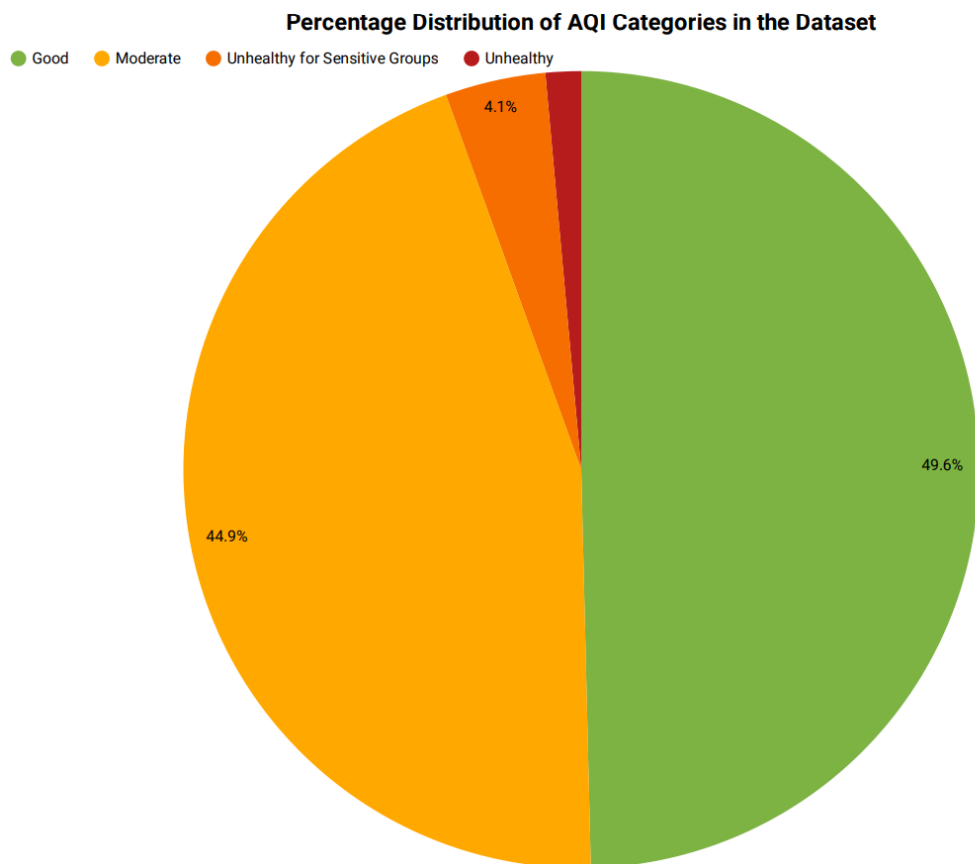


Figure 6

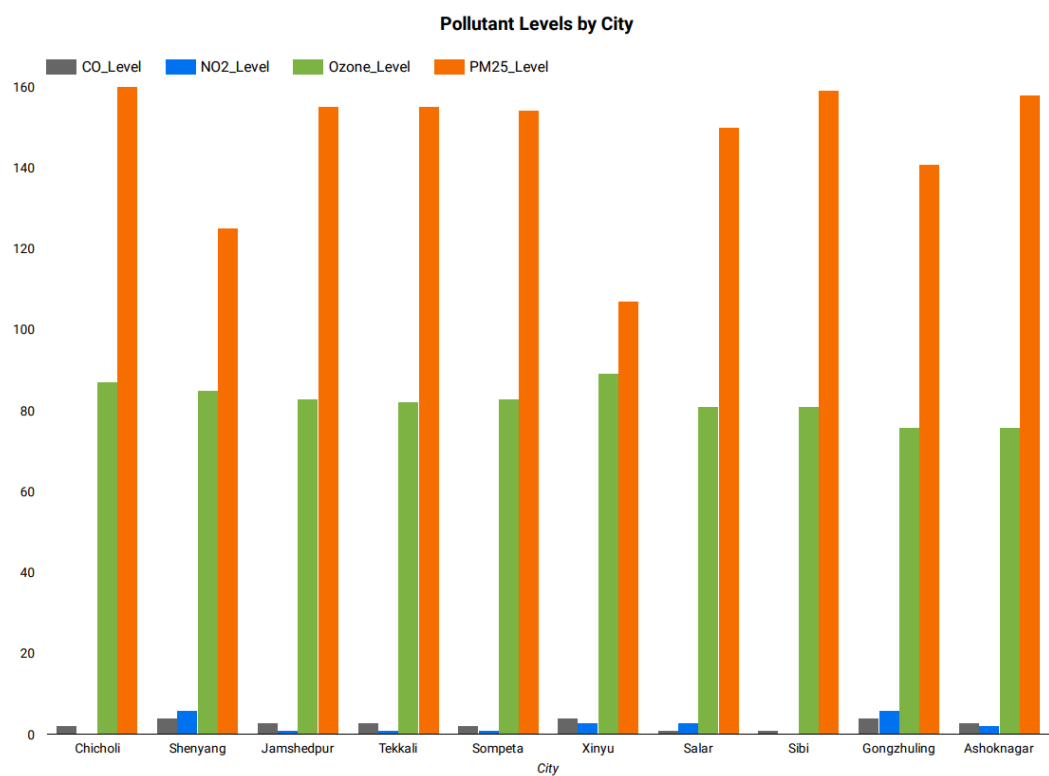


Figure 7

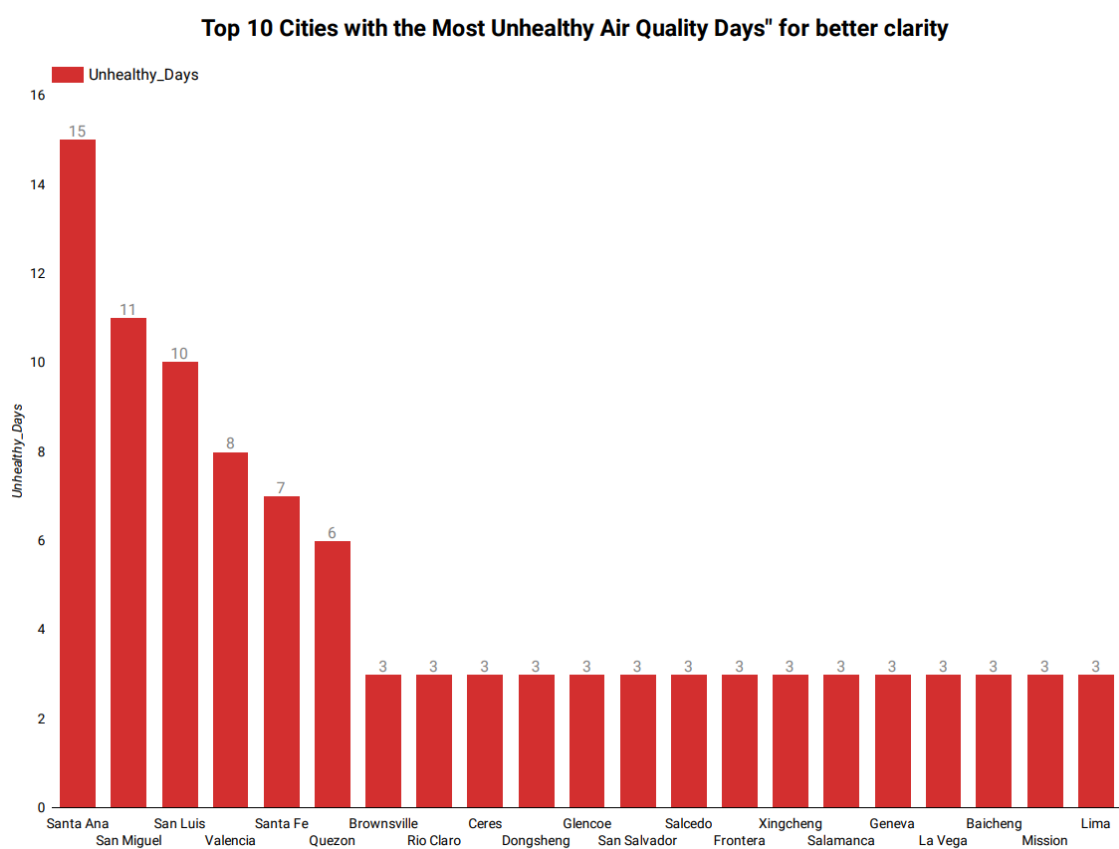


Figure 8

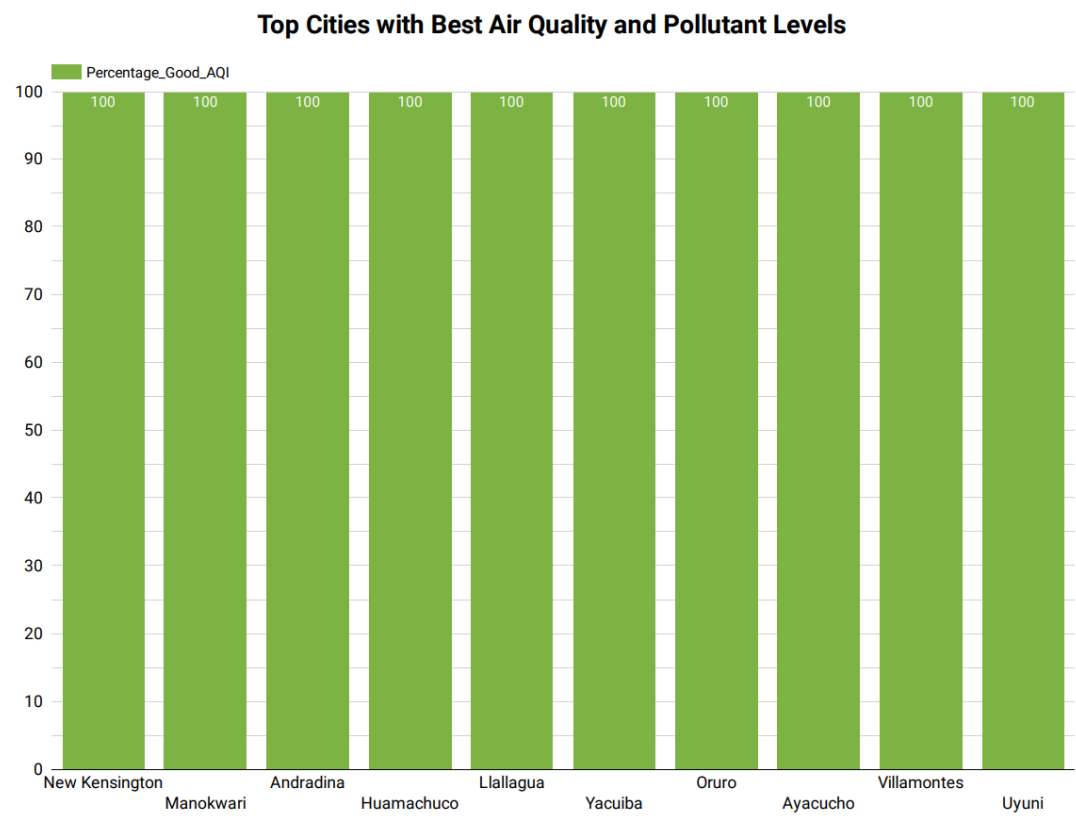


Figure 9

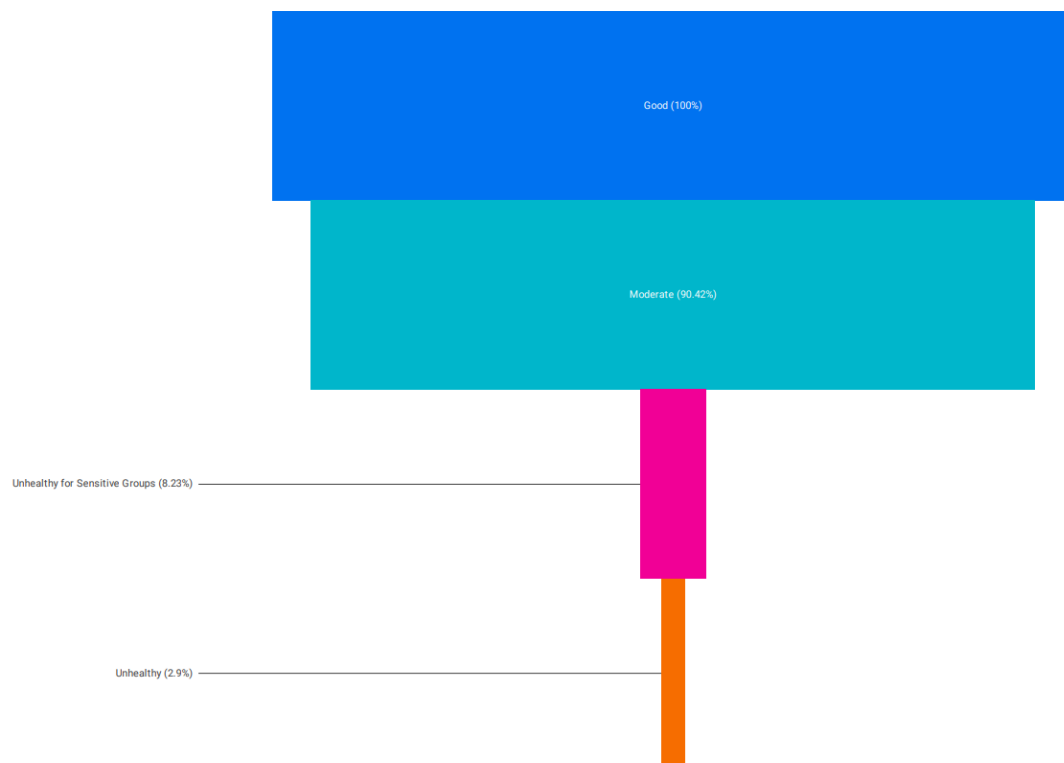


Figure 10

5. Discussion

1. Figure 1 – The bar chart highlights countries like Mauritania, Saudi Arabia, and Gambia with the highest average AQI values. These numbers point towards significant air pollution challenges in these regions, likely driven by industrial activities and environmental factors. It emphasizes the urgent need for air quality management policies in these countries.
2. Figure 2 – This visualization shows how air quality varies across the globe. Countries near urban and industrial centers tend to have higher AQI values, while regions with strict environmental regulations exhibit better air quality. This reflects the critical role of governance in managing air pollution.
3. Figure 3 – This stacked bar chart illustrates the pollutant composition (CO, NO₂, Ozone, and PM_{2.5} levels) in various countries. PM_{2.5} consistently dominates the pollutant levels across all countries, contributing significantly to air quality issues. Other pollutants like Ozone, CO, and NO₂ vary across regions, highlighting the diverse sources and environmental challenges influencing air pollution globally.
4. Figure 4 – The map reveals areas with a high percentage of unhealthy air quality days, particularly in urbanized and industrial regions like Pakistan and Senegal. These findings highlight the importance of targeted pollution control measures in such regions.
5. Figure 5 – The chart illustrates that the Northern Hemisphere, with its dense population and industries, has higher average AQI values than the Southern Hemisphere. This disparity underscores the environmental burden of industrialization and urbanization.
6. Figure 6 – Most data points fall under the "Moderate" and "Unhealthy for Sensitive Groups" categories, making up nearly 95% of the total dataset. This indicates that while some regions maintain acceptable air quality, many still face significant health risks due to pollution.
7. Figure 7 - The chart shows cities like Shenyang and Jamshedpur experiencing high levels of PM_{2.5}, a major contributor to poor air quality. These findings suggest the need for city-specific strategies to reduce pollution and improve living conditions.
8. Figure 8 - Cities like Santa Ana and San Miguel are repeatedly reported to have unhealthy air quality days. This calls for immediate steps to address pollution sources such as industrial emissions and vehicular exhaust.
9. Figure 9 - Cities like New Kensington and Manokwari consistently report "Good" AQI levels. This demonstrates the positive impact of effective policies and environmental advantages in maintaining better air quality.

10. Figure 10 - This visualization represents the distribution of AQI categories globally. A majority of locations report "Good" air quality, followed by "Moderate," while smaller proportions fall under "Unhealthy for Sensitive Groups" and "Unhealthy" categories. These insights highlight the predominance of favorable air conditions worldwide, with room for improvement in areas experiencing poorer air quality.

Key Takeaways

- **Critical Regions:** Countries and cities with consistently high AQI need urgent interventions to control air pollution and improve public health.
- **Pollution Sources:** Pollutants like PM2.5 and CO are key drivers of air quality issues, particularly in urban and industrialized areas.
- **Geographic Patterns:** The Northern Hemisphere's higher AQI values reflect the impact of industrialization and population density, necessitating regional collaboration to tackle pollution.
- **Positive Examples:** Cities with "Good" air quality provide insights into successful strategies for maintaining environmental health and serve as role models for other regions.

These insights highlight the importance of addressing air pollution at both local and global levels to ensure sustainable environmental practices.

5.1 Skills applied from the course

This project was an excellent opportunity to put into practice several concepts and skills learned during the course. I applied knowledge from modules like **Cloud Computing** and **Virtualization** to set up and manage Google Cloud Platform (GCP) for data storage and processing. These modules gave me a strong understanding of how to build a scalable and secure environment for handling large datasets.

The **Data Types and Sources** module was particularly useful when selecting and importing the dataset from Kaggle. I followed the principles from **Data Pipelines and Lifecycle Management** to preprocess the data, ensuring it was clean and ready for analysis. This involved handling missing values, removing outliers, and structuring the dataset for advanced analytics.

The **Modeling and Analytics** module provided me with the foundation for performing statistical analysis and creating visualizations. Using BigQuery and Looker Studio, I was able to generate insights and present findings effectively. Additionally, concepts from **Computing Principles and System Design** helped me structure my work in Google Colab for smooth integration with GCP tools.

Lastly, the lessons from **Big Data Impact** and **Data Governance** taught me the importance of collaboration and open data sharing. I published the cleaned dataset and scripts on GitHub, ensuring transparency and encouraging contributions from others in the data science community.

5.2 Challenges Faced and Learnings

While I successfully completed the project, there were a few challenges that tested my problem-solving skills:

1. **Cluster Setup Issues:** Setting up the Google Cloud Dataproc cluster for PySpark analysis was initially challenging. I faced difficulties in configuring the cluster to match the project requirements, but after researching and adjusting settings, I managed to get it running smoothly.
2. **Integration with BigQuery:** Exporting the output from PySpark to BigQuery was a significant challenge. I encountered issues related to permissions and data format compatibility. Despite trying different configurations, I couldn't resolve this completely and had to save the output locally as a workaround.

Load job details



Error while reading data, error message: CSV processing encountered too many errors, giving up. Rows: 14791; errors: 1; max bad: 0; error percent: 0



Error while reading data, error message: Unable to parse; line_number: 1
byte_offset_to_start_of_line: 0 column_index: 2 column_name: "AQI_Value"
column_type: DOUBLE value: "AQI_Value" File:
gs://air_quality_bucket_sponduru/cleaned_air_quality_data.csv

Load job details

!

Field name 'PM2.5 AQI Category' is not supported by the current character map. Please change your field name or use character map V2 to let the system modify the field names.

Job ID	aqi-analysis-project:US.bquxjob_14601c5a_1935d41382a
User	sponduru@iu.edu
Location	US
Creation time	Nov 24, 2024, 3:19:34 AM UTC-5
Start time	Nov 24, 2024, 3:19:35 AM UTC-5
End time	Nov 24, 2024, 3:19:35 AM UTC-5
Duration	0 sec
Auto-detect schema	true
Ignore unknown values	false
Source format	CSV
Max bad records	0
Destination table	aqi-analysis-project.air_quality_analysis.final_air_quality

REPEAT LOAD JOB

CLOSE

- Visualization Limitations:** While Looker Studio was useful for creating basic visualizations, it had limitations in handling complex visualizations like scatter plot matrices. I had to rely on additional tools to complete some of the required visualizations, which added extra time and effort.
- Time Management:** Balancing the various stages of the project, such as data preprocessing, analysis, and visualization, within the given timeline was tough. Debugging technical issues consumed more time than expected, affecting the overall workflow.

Key Takeaways: These challenges taught me the importance of patience and troubleshooting in real-world projects. I also learned the value of understanding cloud-based permissions and schema settings for seamless integration. Overall, this project improved my technical and problem-solving skills, preparing me for future roles in data science and big data analytics.

6. Conclusion

In conclusion on this project, I found it to be both challenging and rewarding, helping me gain a deeper understanding of global air quality trends. Working on tasks like data collection, preprocessing, analysis, and visualization improved my technical skills significantly. Integrating Google Cloud Platform with Google Colab and using BigQuery for data querying and visualization allowed me to apply the concepts I learned during the course. Although I faced challenges with Google Cloud Dataproc while implementing PySpark, I adapted by exploring alternative approaches, which taught me the importance of persistence and flexibility in solving technical issues.

This project has shown me how the theoretical knowledge gained in class can be applied to solve real-world problems. While I couldn't achieve all my objectives, I am satisfied with the meaningful insights and visualizations I was able to create, which highlighted the trends in air quality across different regions. This experience has motivated me to further explore advanced cloud-based tools and refine my skills, proving that even small steps can contribute to addressing larger environmental challenges.

7. References

- [1] Kaggle Air Quality Dataset: <https://www.kaggle.com/datasets/adityaramachandran27/world-air-quality-index-by-city-and-coordinates>
- [2] Google Cloud VPC networks: <https://cloud.google.com/vpc/docs/vpc>
- [3] Dataproc: Qwik Start – Console :
[https://www.cloudskillsboost.google/focuses/586?catalog_rank=%7B\"rank\"%3A7,\"num_filters\"%3A1,\"has_search\"%3Atrue%7D&parent=catalog&search_id=9028616](https://www.cloudskillsboost.google/focuses/586?catalog_rank=%7B\)
- [4] Introduction to Cloud Dataproc: Hadoop and Spark on Google Cloud:
[https://www.cloudskillsboost.google/focuses/672?catalog_rank=%7B\"rank\"%3A1,\"num_filters\"%3A0,\"has_search\"%3Atrue%7D&parent=catalog&search_id=9026299](https://www.cloudskillsboost.google/focuses/672?catalog_rank=%7B\)
- [5] Connect GCP Bucket to Google Colab : <https://pub.towardsai.net/connect-colab-to-gcs-bucket-using-gcsfuse-29f4f844d074>
- [6] Ingesting New Datasets to BigQuery:
[https://www.cloudskillsboost.google/focuses/3692?catalog_rank=%7B\"rank\"%3A1,\"num_filters\"%3A0,\"has_search\"%3Atrue%7D&parent=catalog&search_id=26980459](https://www.cloudskillsboost.google/focuses/3692?catalog_rank=%7B\)
- [7] Looker Studio Google Cloud Platform Tutorial:
<https://cloud.google.com/bigquery/docs/visualize-looker-studio>