Exploring and Visualising Real-World Data

1 Contents

1. Data Selection [Task 1]

Choose a real-world dataset:
The dataset we have chosen is Global Air Pollution Dataset
https://www.kaggle.com/datasets/hasibalmuzdadid/global-air-pollution-dataset

Reason for selecting this data set is it covers air quality data of different cities of different countries. Also because of societal relevance it is an important topic on environmental health and pollution impact. One of the major factors of selecting this dataset as it contains pollutants air quality index based on their cities all the variable are mentioned below.

This dataset in the most upvoted on Kaggle (one of the most reliable datasets collections) it has the most view and most downloads when you search global air pollution.

However, there are some secondary issues in this dataset one of them is that is it has been 2 years since its last update and the other that there are no timestamps.

But even with those little issue this dataset is much cleaner and contain pollutants values which helps in generating insights regarding countries and their policies in tackling air pollution than some of the other datasets available on Kaggle or other platforms.

It also has several factors for analysis:

- **Country:** Name of the country
- **City:** Name of the city
- **AQI Value:** Overall AQI value of the city
- **AQI Category:** Overall AQI category of the city
- **CO AQI Value:** AQI value of Carbon Monoxide of the city
- **CO AQI Category:** AQI category of Carbon Monoxide of the city
- **Ozone AQI Value:** AQI value of Ozone of the city
- **Ozone AQI Category:** AQI category of Ozone of the city
- **NO2 AQI Value:** AQI value of Nitrogen Dioxide of the city
- **NO2 AQI Category:** AQI category of Nitrogen Dioxide of the city
- **PM2.5 AQI Value:** AQI value of Particulate Matter with a diameter of 2.5 micrometers or less of the city
- **PM2.5 AQI Category:** AQI category of Particulate Matter with a diameter of 2.5 micrometers or less of the city

2. Data Exploration

Exploratory Data analysis:

- Most of the Exploration has been done through visualizations in the tableau workbook. However, some of the key insights are shown below

Key insights generated from data:

- The United States of America has the greatest number of values in the dataset however, it is still not the top country with the highest sum of AQI value as shown below
  The data is ordered in descending order with respect to AQI Values.

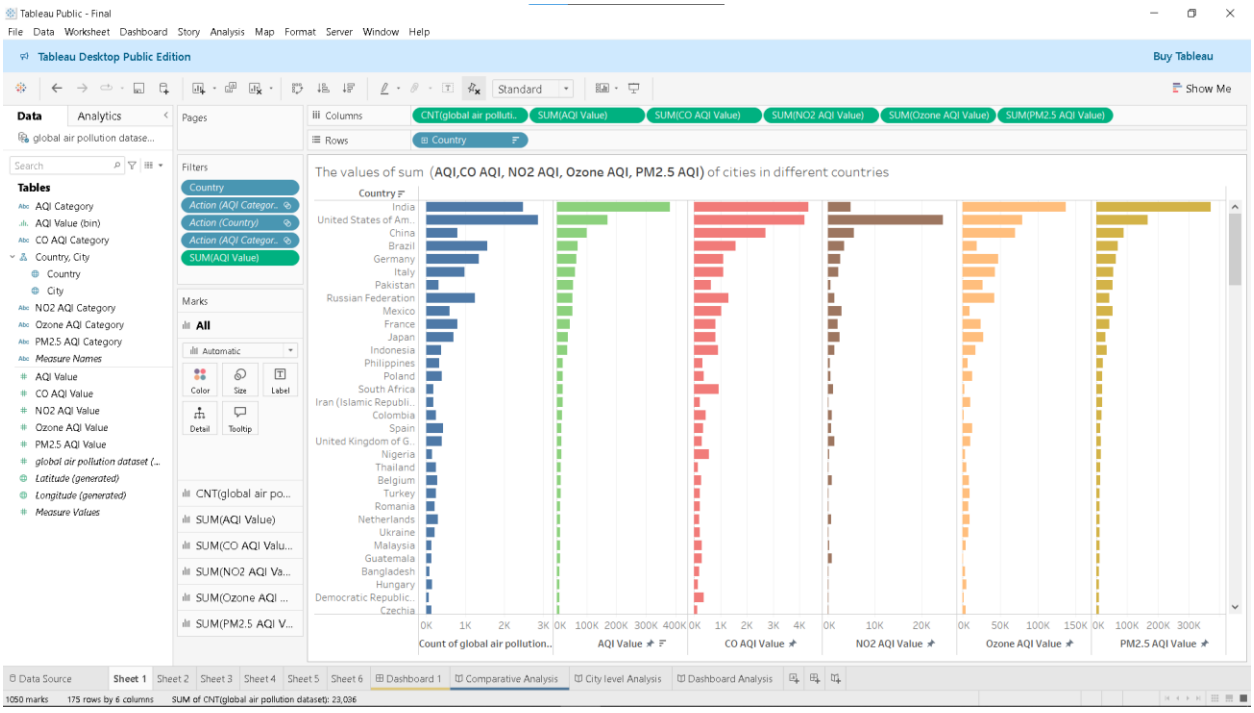


Figure 2. 1

- When countries are seen, whose count is above 100 values in the dataset the country with the highest avg of AQI level comes out to be Pakistan

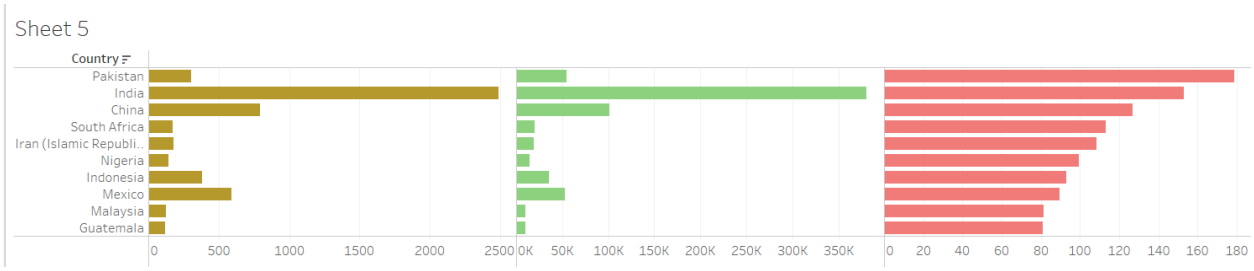


Figure 2. 2

- It is also important to highlight that there are 6 categories in which AQI levels are divided those are:

Figure 2. 3

The majority of cities fall under unhealthy for sensitive levels emphasizing the widespread of pollution issue

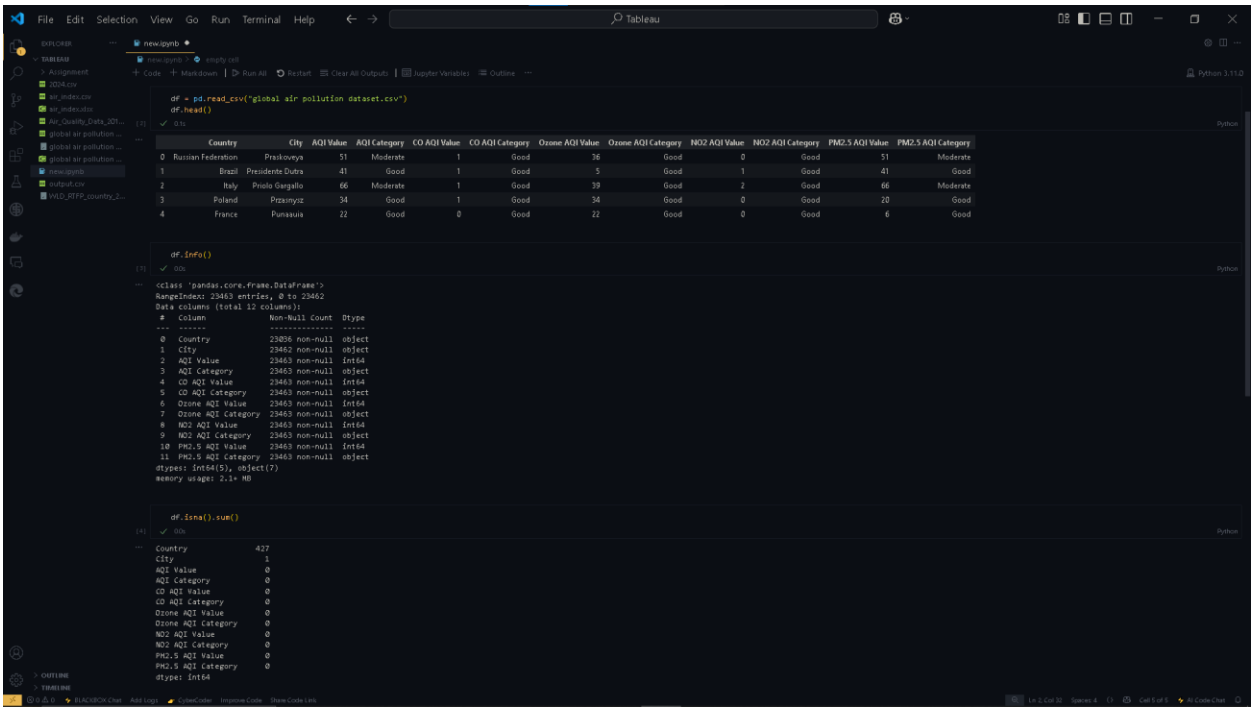
Air Quality varies drastically across regions, with Asia and South America experiencing high air pollution.

India, China and Pakistan face major AQI challenges particularly with NO2 and PM2.5 pollutants with most of the cities of these countries not even in the good category of Air Quality index

Countries with better environmental policies (Countries from Europe and North America) maintain healthier air Quality.

3. Data Cleaning and Transformation:

Before performing task 2 it is important to check for null values before performing visualizations, A python script was performed on the dataset to check for null values in columns there were two columns which contained null values as you can see here



```
df = pd.read_csv("global air pollution dataset.csv")
df.head()
```

	Country	City	AQI Value	AQI Category	CO AQI Value	CO AQI Category	Ozone AQI Value	Ozone AQI Category	NO2 AQI Value	NO2 AQI Category	PM2.5 AQI Value	PM2.5 AQI Category
0	Russian Federation	Praskoveya	51	Moderate	1	Good	36	Good	0	Good	51	Moderate
1	Brazil	Presidente Dutra	41	Good	1	Good	5	Good	1	Good	41	Good
2	Italy	Priolo Gargallo	66	Moderate	1	Good	39	Good	2	Good	66	Moderate
3	Poland	Przasnysz	34	Good	1	Good	34	Good	0	Good	20	Good
4	France	Punaauia	22	Good	0	Good	22	Good	0	Good	6	Good

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23463 entries, 0 to 23462
Data columns (total 12 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Country             23036 non-null  object
 1   City                23462 non-null  object
 2   AQI Value           23463 non-null  int64
 3   AQI Category        23463 non-null  object
 4   CO AQI Value        23463 non-null  int64
 5   CO AQI Category     23463 non-null  object
 6   Ozone AQI Value     23463 non-null  int64
 7   Ozone AQI Category  23463 non-null  object
 8   NO2 AQI Value       23463 non-null  int64
 9   NO2 AQI Category    23463 non-null  object
 10  PM2.5 AQI Value     23463 non-null  int64
 11  PM2.5 AQI Category  23463 non-null  object
dtypes: int64(5), object(7)
memory usage: 2.1+ MB
```

```
df.isna().sum()
```

```
Country             427
City                  1
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
dtype: int64
```

Figure 3. 1

This can also be seen in tableau after loading the dataset

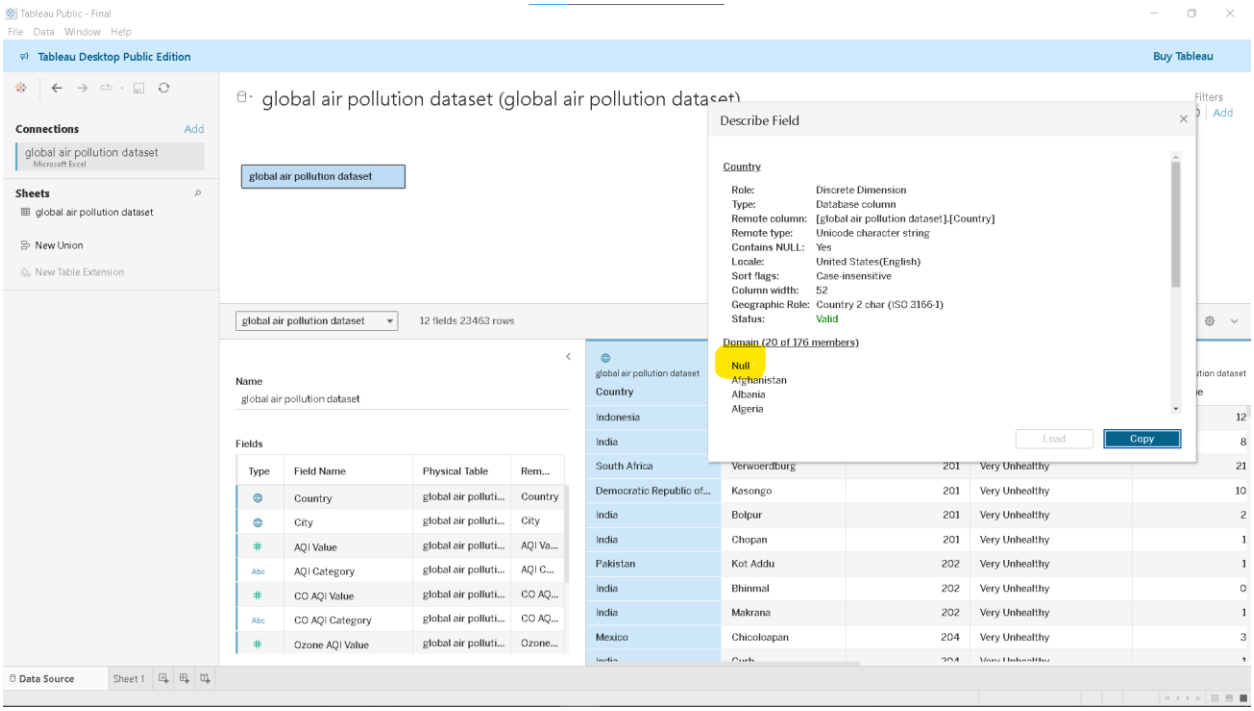


Figure 3. 2

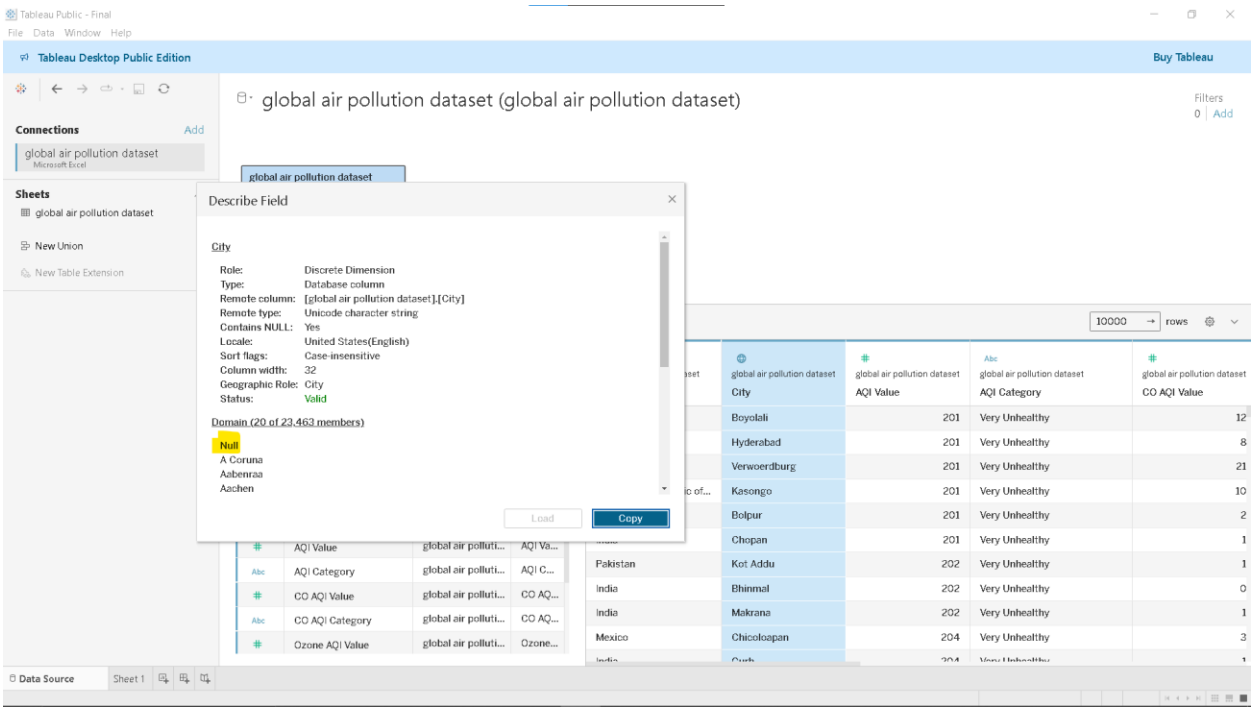


Figure 3. 3

Finding names of these countries/cities is one of the most challenging tasks in this dataset but to tackle this challenge joining technique in tableau is employed to find and fill missing data.

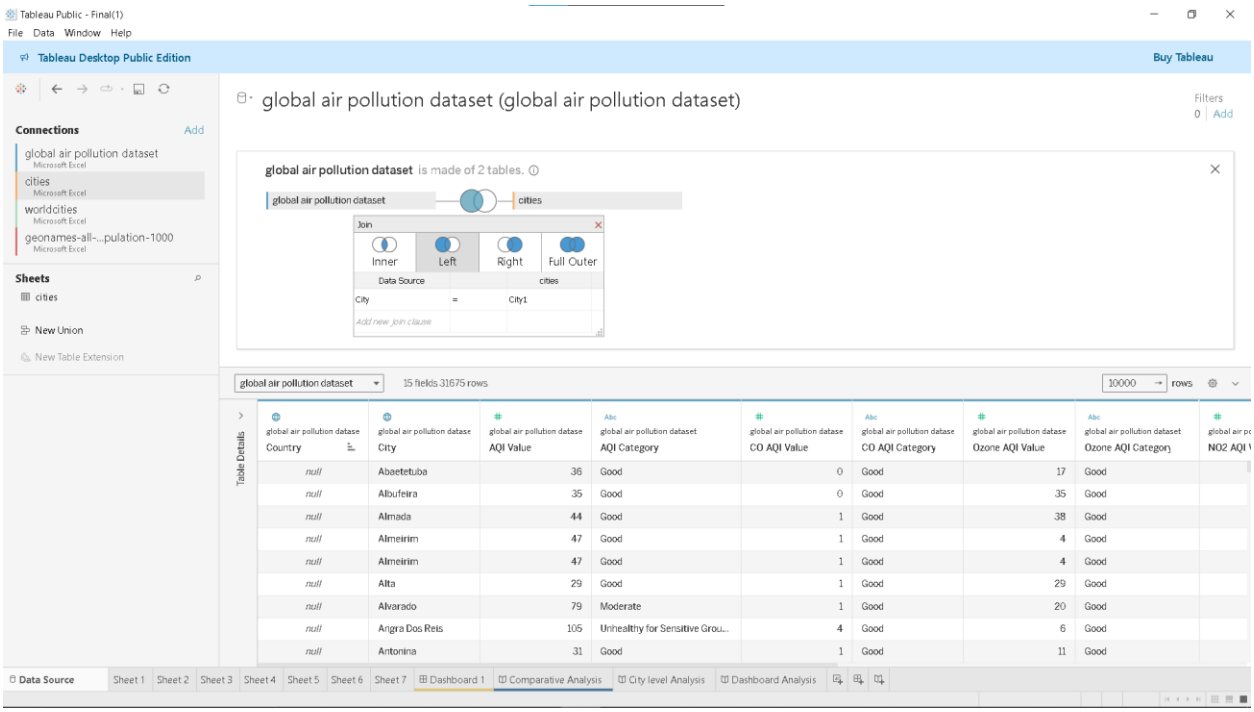


Figure 3. 4

The cities dataset contains values of 150455 cities based on their countries compared to the dataset which is around 23000 by applying left join even applying left join there were still 111 values left.

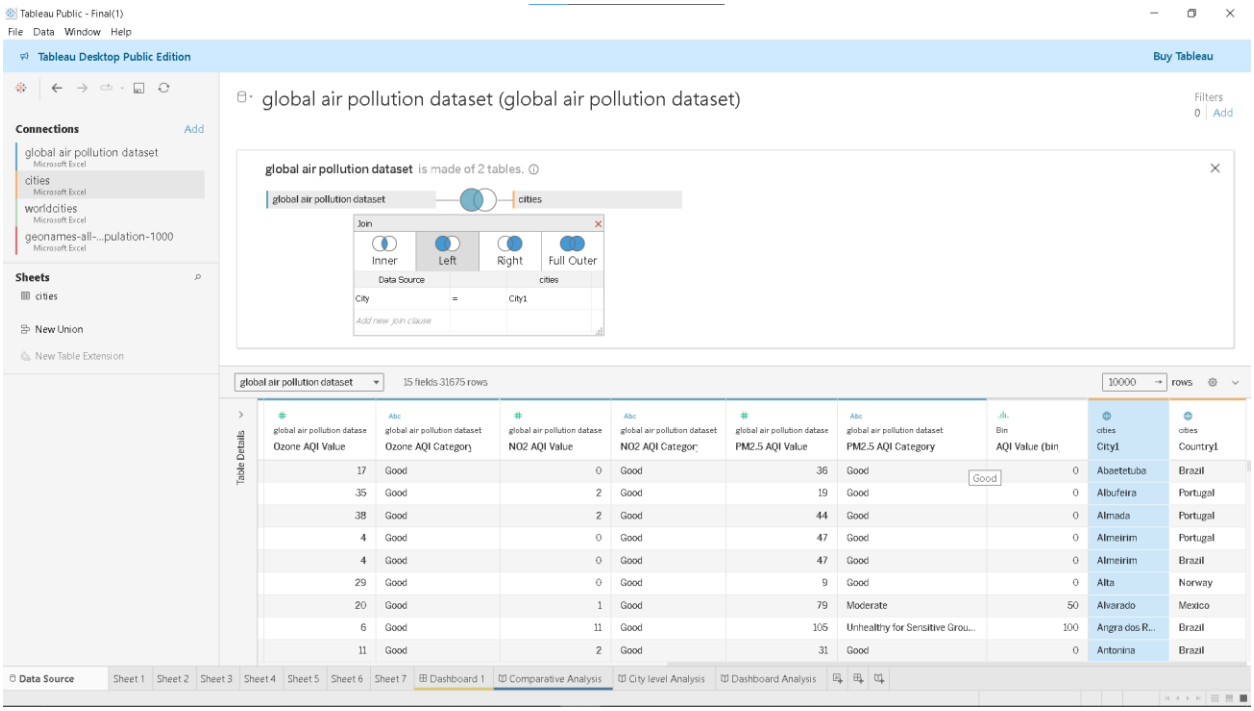


Figure 3. 5

hence it was better to just omit all the rows which contained empty values of cities and countries more importantly the values identified by performing left join were not identified on map that's why it is better to omit all those values.

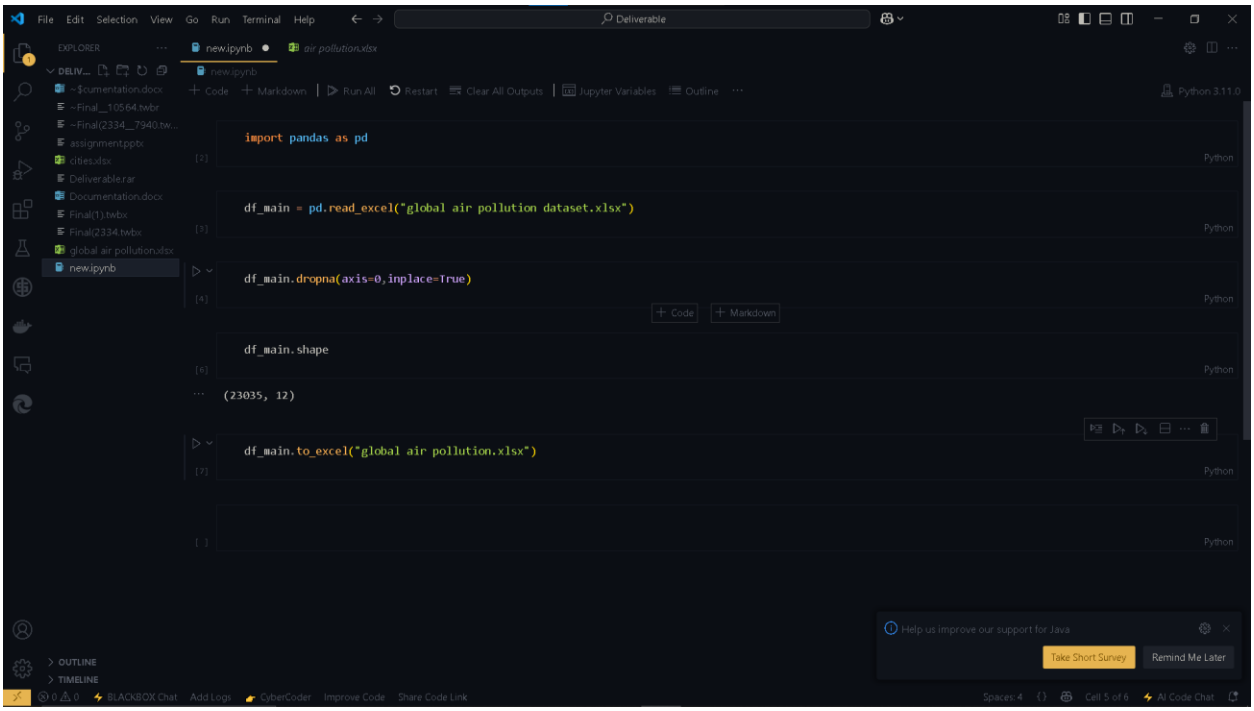


Figure 3. 6

The omitting step is performed in python where we have imported our original dataset and used dropna() function of pandas DataFrame to drop all null values along axis = 0(rows) hence leaving us with 23035 rows along with 12 original columns.

4. Visualization Design

There are 6 worksheets that are generated using tableau to visualize the dataset and generate key insights.

Purpose of Each Visualization:

- **Sheet 1: Bar Plot.**

Purpose: A bar plot is deployed to find the average of Air Quality index along with the average of Air Quality index of different pollutants against countries.

Filters and conditions:

- Countries whose count of cities is above 10 are considered for this plot and **Pakistan** turned to be the country with highest average of Air quality index with the average of 178.8 for the count of 307. India came in second.
- Every Column is assigned different color.
- The sheet is formatted into descending order based on average air quality index values.

- **Sheet 2: Map**

Purpose: Geographic representation of different cities based on Air quality index

- Every city is divided into Air quality index category based on their AQI value
- Tooltip has been used to show the AQI values of the city along with its country.
- This Visualization allows the user to view from a geographic point of view

- Sheet 3: **Bar Plot**

Purpose: This bar plot has been deployed to categorize AQI categories.

- There are six AQI categories according to international standard with each category having its own color identification.

• AQI Value	• AQI Category	• Color
• 0-50	• Good	• Green
• 50-100	• Moderate	• Yellow
• 100-150	• Unhealthy for sensitive people	• Orange
• 150-200	• Unhealthy	• Red
• 200-300	• Very Unhealthy	• Violet
• 300-500	• Hazardous	• Maroon

- **Sheet 4: Bar Plot**

Purpose: City Level analysis of different countries the major purpose of this visualization is to be used in storytelling

- **Sheet 5: Bar Plot**

Purpose: Find the top 10 countries which are most polluted

Filter:

- The countries are ordered in descending form based on their average the top 10 are selected for visualization
- The city count was limited to at least 100.
- There were no Suprises that Pakistan came on top but most important insight was that US was nowhere near the top highlighting its strong policies in tackling Air pollution and managing the rise of industrialization.

- **Sheet 6: Text Table**

Purpose: To check statistics of different countries

- Count, Average, Minimum, Maximum standard deviation are analyzed in this visualization
- A Filter on count of cities has been added which allows the user to find the specific count range.

5. Interactive Dashboards

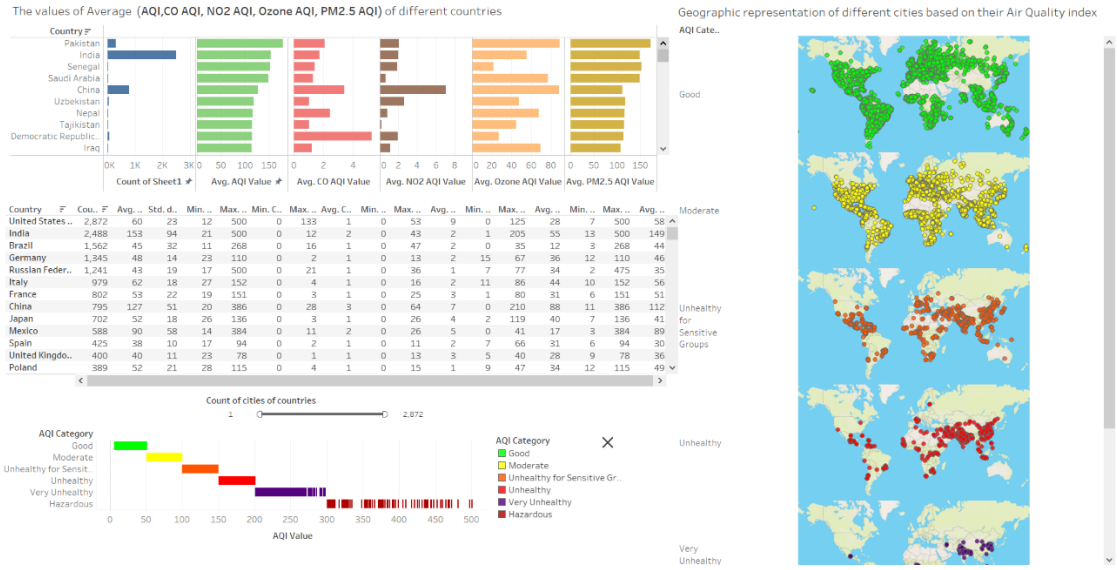


Figure 5. 1

This interactive dashboard allows users find AQI values of different pollutants based on cities/countries. This dashboard uses a geographic visualization to clearly communicate the region of different countries it also helps in analyzing the cities based on AQI categories. It can be found that in this dataset there are no cities in Pakistan that is in good category of the AQI.

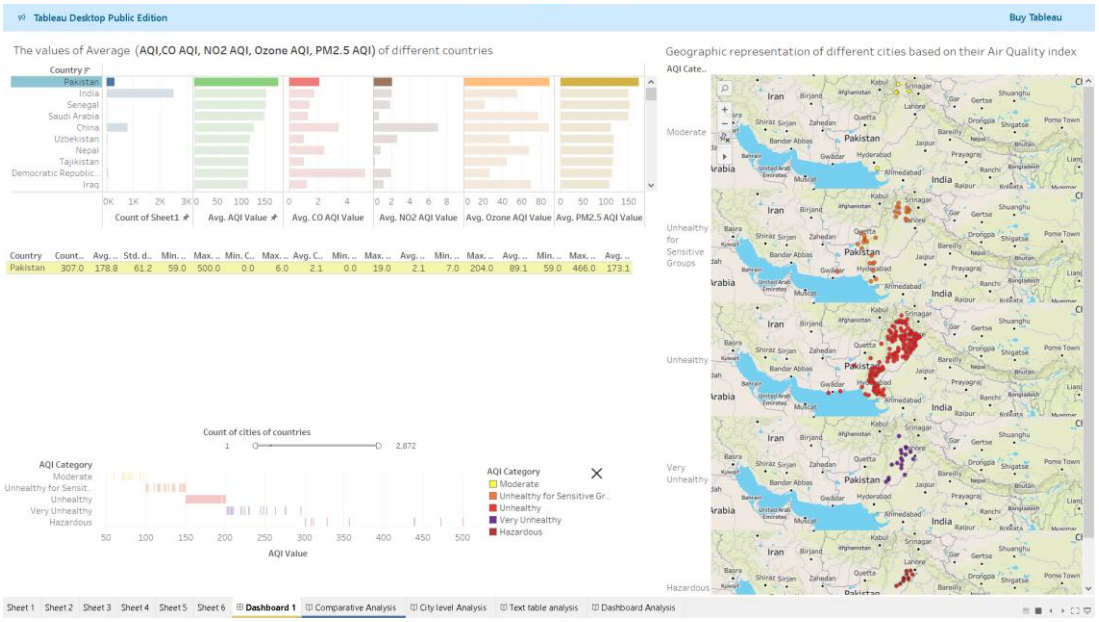


Figure 5. 2

As you can see when Pakistan is seen good category is not found on map also on the bar part below categorizing it.

6. Data Storytelling

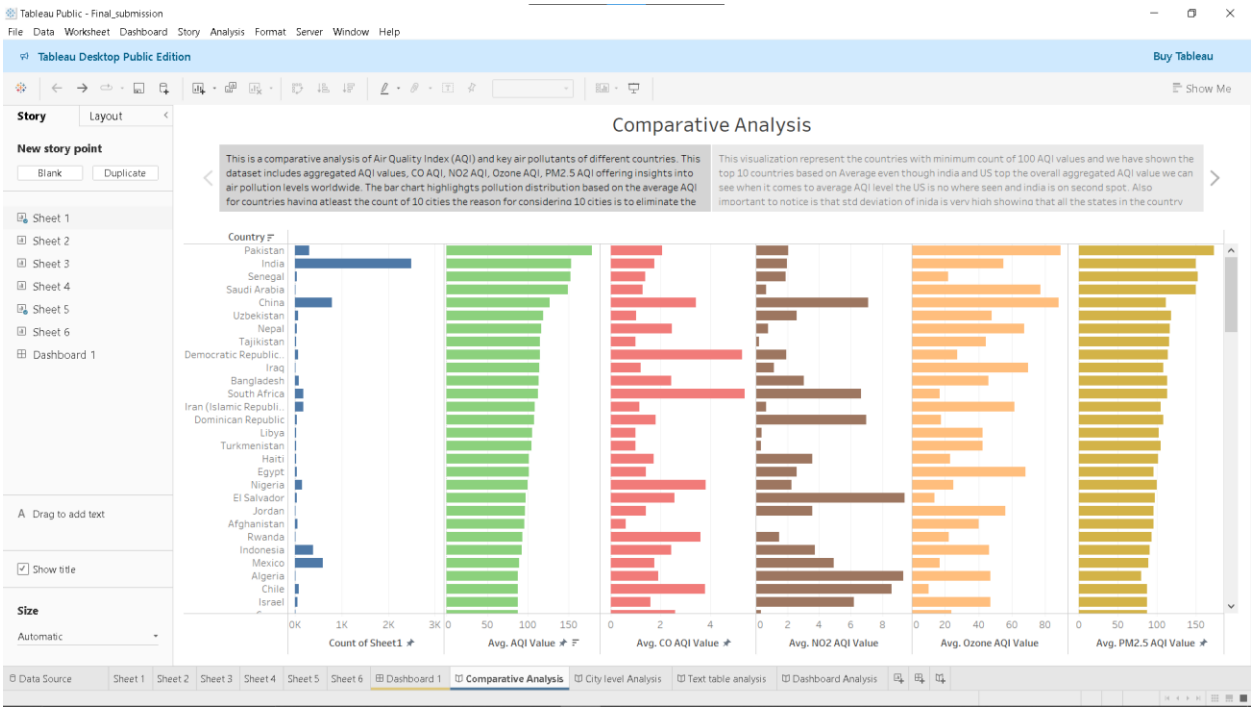


Figure 6. 1

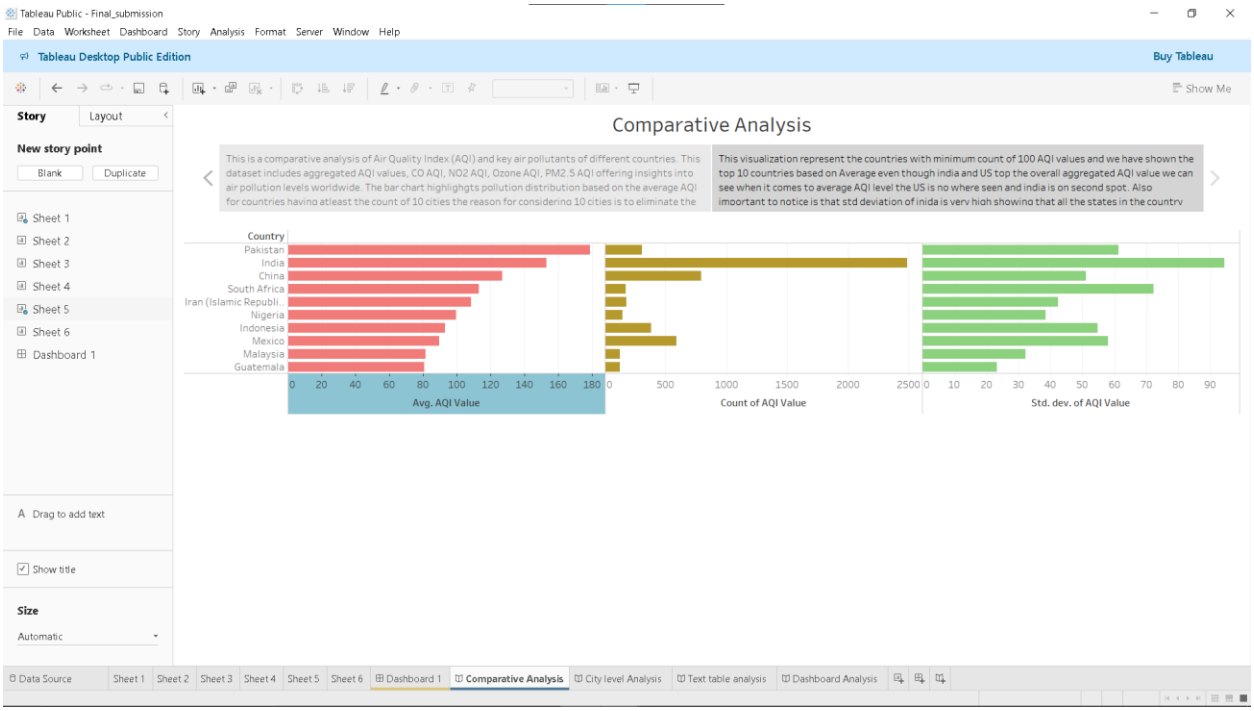


Figure 6. 2

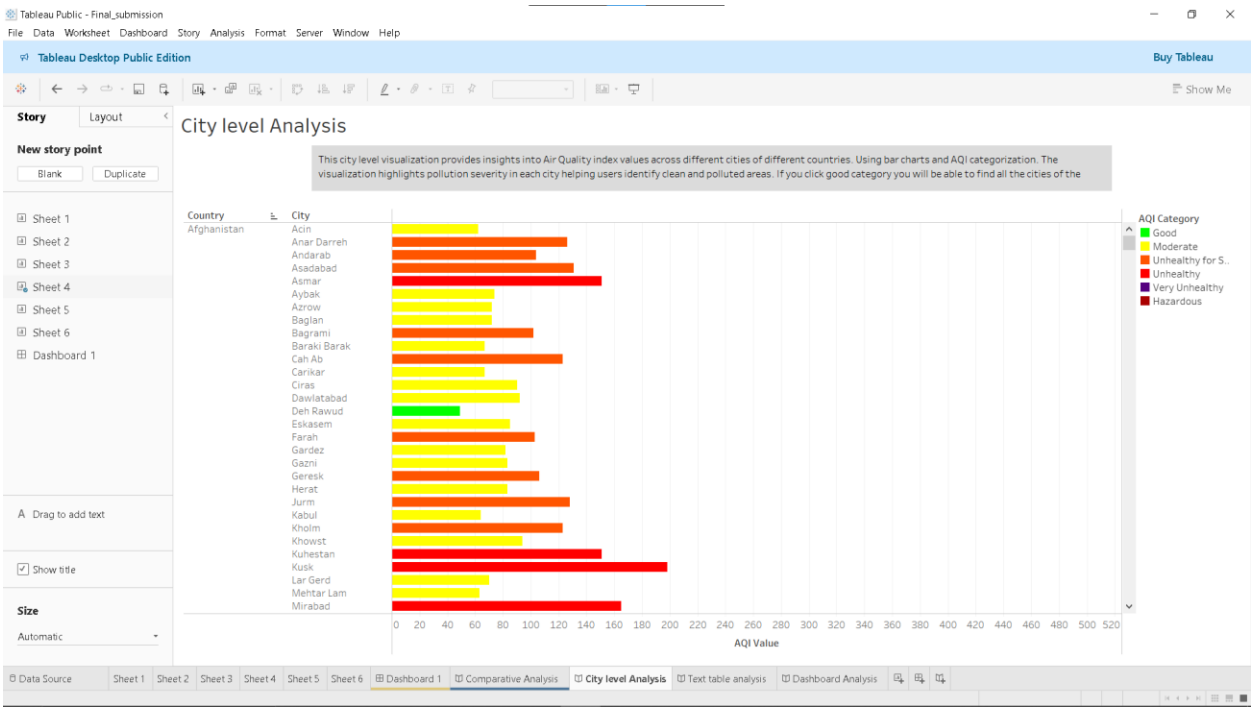


Figure 6. 3

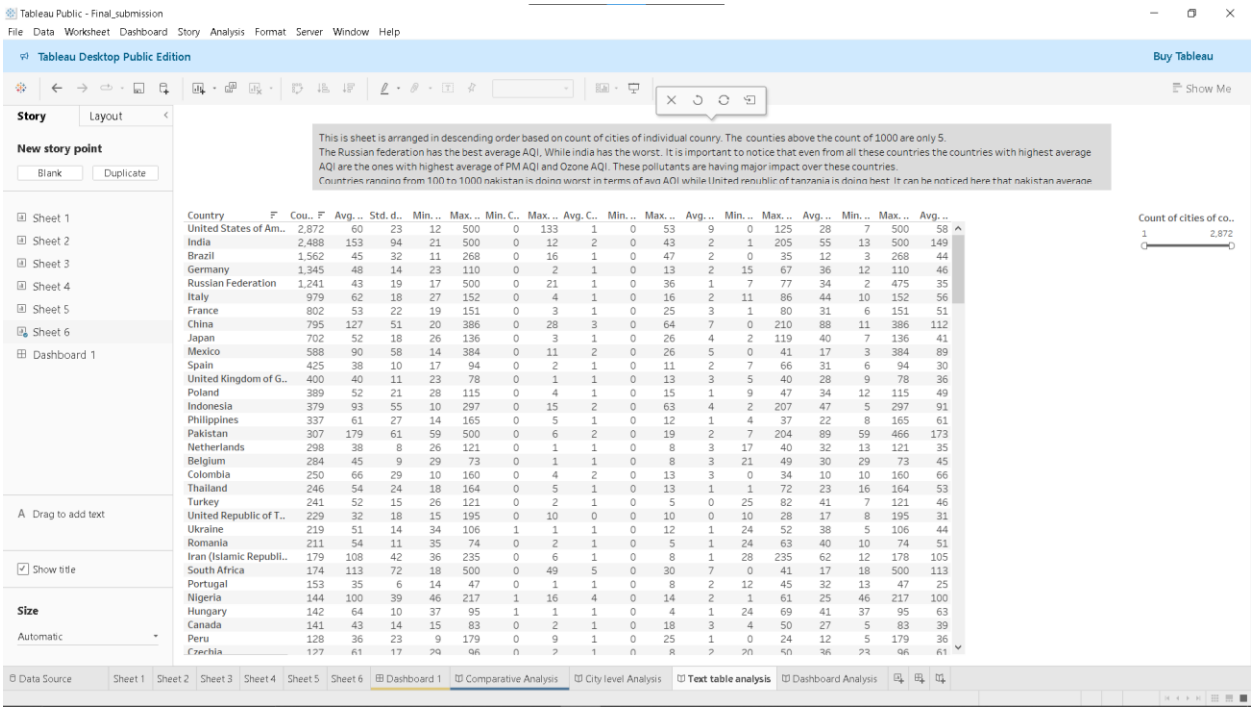


Figure 6. 4

7. Presentation

Presentation is provided assignment.ppt