

Exploring and Visualising Real-World Data

1 Contents

1. Data Selection	2
2. Data Exploration.....	3
3. Data Cleaning and Transformation:	4
4. Visualization Design.	9
5. Interactive Dashboards	11
6. Data Storytelling.....	12
7. Presentation.....	13
Figure 2. 1.....	3
Figure 2. 2.....	3
Figure 2. 3.....	4
Figure 3. 1.....	5
Figure 3. 2.....	5
Figure 3. 3.....	6
Figure 3. 4.....	6
Figure 3. 5.....	7
Figure 3. 6.....	8
Figure 5. 1.....	11
Figure 5. 2.....	11
Figure 6. 1.....	12
Figure 6. 2.....	12
Figure 6. 3.....	13
Figure 6. 4.....	13

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 is 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.

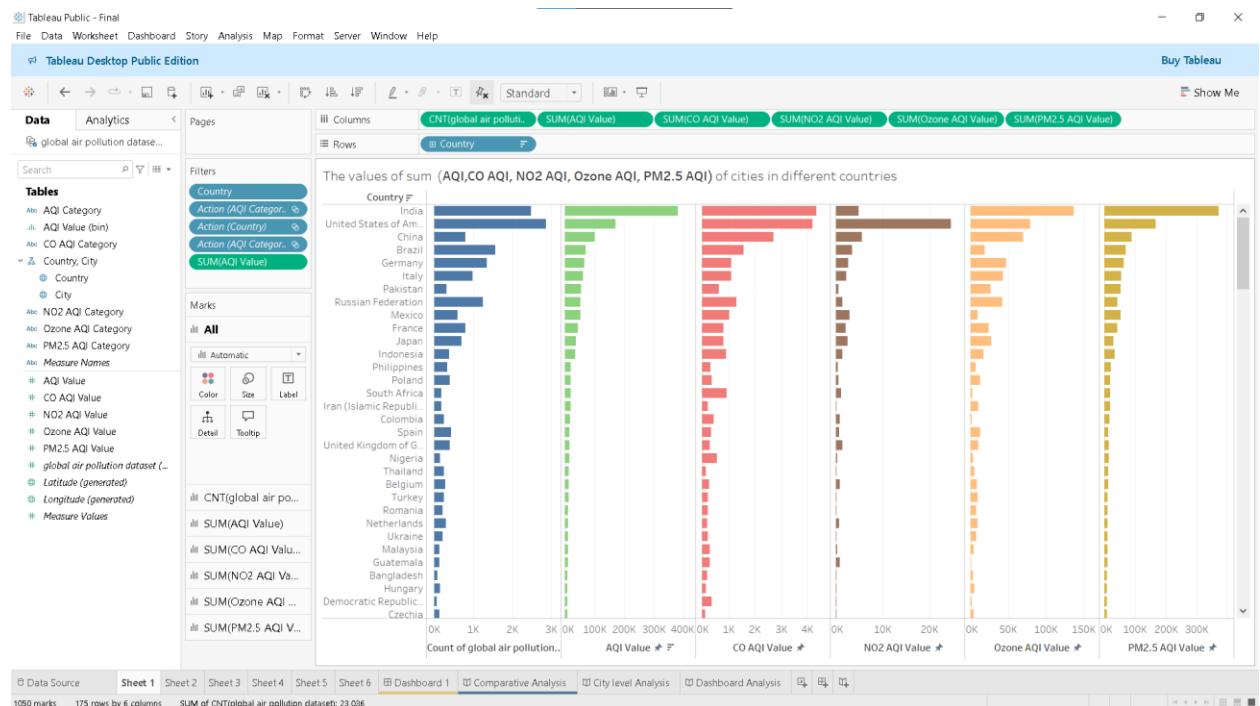


Figure 2. 1

- 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

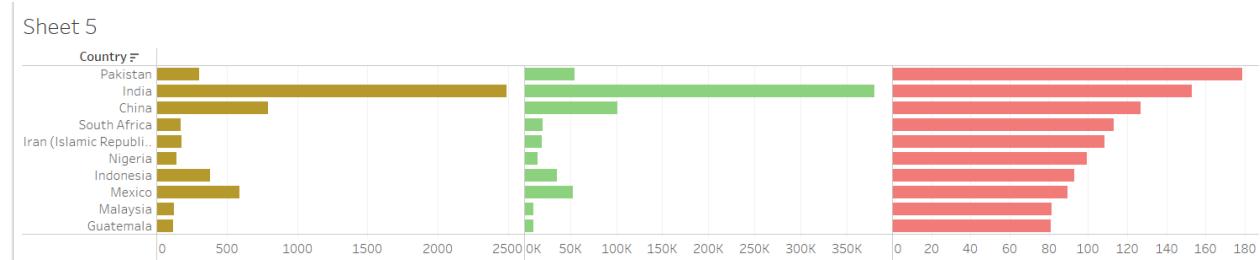


Figure 2. 2

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

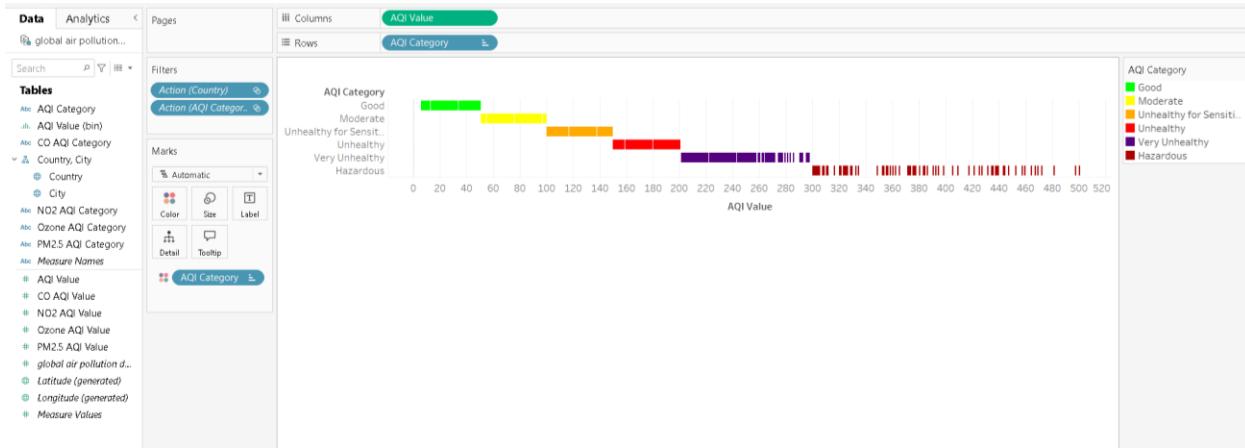


Figure 2. 3

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

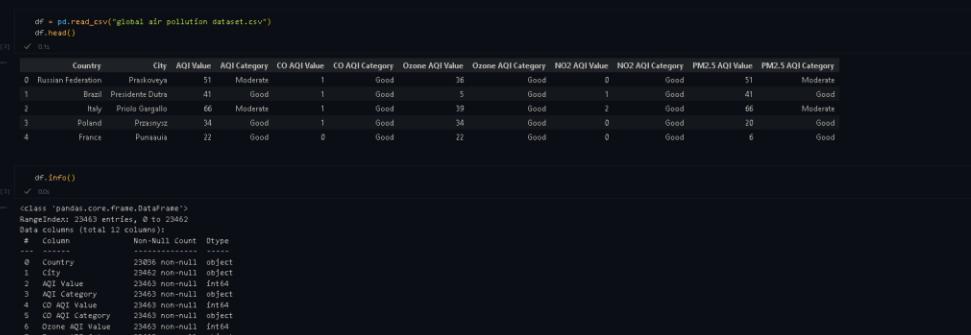
Air Quality varies drastically across regions, with Asia and South America experiencing high air pollution.

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



The screenshot shows a Jupyter Notebook interface with the following content:

- Code Cell 1:** Displays the code `df = pd.read_csv("global air pollution dataset.csv")` and `df.head()`. The output shows the first 5 rows of a DataFrame with columns: 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, and PM2.5 AQI Category.
- Code Cell 2:** Displays the code `df.info()`. The output provides detailed information about the DataFrame, including the number of rows (23463), column names, data types, and non-null counts.
- Code Cell 3:** Displays the code `df.isna().sum()`. The output shows the count of missing values for each column.

Figure 3. 1

This can also be seen in tableau after loading the dataset

Tableau Public - Final

File Data Window Help

Tableau Desktop Public Edition

Buy Tableau

Connections Add

global air pollution dataset Microsoft Excel

Sheets

global air pollution dataset

New Union

New Table Extension

global air pollution dataset 12 fields 23463 rows

Name

global air pollution dataset

Fields

Type	Field Name	Physical Table	Rem...
Country	global air pollut...	Country	
City	global air pollut...	City	
AQI Value	global air pollut...	AQI Va...	
AQI Category	global air pollut...	AQI C...	
CO AQI Value	global air pollut...	CO AQ...	
CO AQI Category	global air pollut...	CO AQ...	
Ozone AQI Value	global air pollut...	Ozone...	

global air pollution dataset (global air pollution dataset)

Describe Field

Country

Role: Discrete Dimension
Type: Database column
Remote column: [global air pollution dataset].[Country]
Remote type: Unicode character string
Contains NULL: Yes
Locale: United States(English)
Sort flags: Case-insensitive
Column width: 52
Geographic Role: Country 2 char (ISO 3166-1)
Status: Valid

Domain (20 of 176 members)

Null
Afghanistan
Albania
Algeria

Load Copy

Verwoerdburg 201 Very Unhealthy

Kasongo 201 Very Unhealthy

Bolpur 201 Very Unhealthy

Chopan 201 Very Unhealthy

Kot Addu 202 Very Unhealthy

Bhimgarh 202 Very Unhealthy

Makrana 202 Very Unhealthy

Chicoloapan 204 Very Unhealthy

Cook 204 Very Unhealthy

12 8 10 2 1 0 1 3 1

Data Source Sheet 1

Figure 3.2

global air pollution dataset (global air pollution dataset)

Describe Field

City

Role: Discrete Dimension
Type: Database column
Remote column: [global air pollution dataset].[City]
Remote type: Unicode character string
Contains NULL: Yes
Locale: United States(English)
Sort flags: Case-insensitive
Column width: 32
Geographic Role: City
Status: Valid

Domain (20 of 23,463 members)

Null
A Coruna
Asenraa
Aschen

Load Copy

10000 rows

City	AQI Value	global air pollution dataset	AQI Category	global air pollution dataset	AQI Value	global air pollution dataset	AQI Category	global air pollution dataset	CO AQI Value		
Boyolali	201	Very Unhealthy	12	Hyderabad	201	Very Unhealthy	8	Verwoerdburg	201	Very Unhealthy	21
Kasongo	201	Very Unhealthy	10	Bolpur	201	Very Unhealthy	2	Chopan	201	Very Unhealthy	1
Kot Addu	202	Very Unhealthy	1	Bhimtal	202	Very Unhealthy	0	Makrana	202	Very Unhealthy	1
India	India	India	Mexico	Chicloapan	204	Very Unhealthy	3	Oru	Oru	Very Unhealthy	1
Pakistan	Pakistan	Pakistan									

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.

global air pollution dataset (global air pollution dataset)

global air pollution dataset is made of 2 tables.

Join

Left

City = City1

Table Details

Country, City, AQI Value, AQI Category, CO AQI Value, CO AQI Category, Ozone AQI Value, Ozone AQI Category, NO2 AQI 1

15 fields 31675 rows

Country	City	AQI Value	AQI Category	CO AQI Value	CO AQI Category	Ozone AQI Value	Ozone AQI Category	NO2 AQI 1
null	Abaeteuba	36	Good	0	Good	17	Good	
null	Albufeira	35	Good	0	Good	35	Good	
null	Almada	44	Good	1	Good	38	Good	
null	Almeirim	47	Good	1	Good	4	Good	
null	Almeirim	47	Good	1	Good	4	Good	
null	Alta	29	Good	1	Good	29	Good	
null	Alvarado	79	Moderate	1	Good	20	Good	
null	Angra Dos Reis	105	Unhealthy for Sensitive Groups	4	Good	6	Good	
null	Antonina	31	Good	1	Good	11	Good	

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.

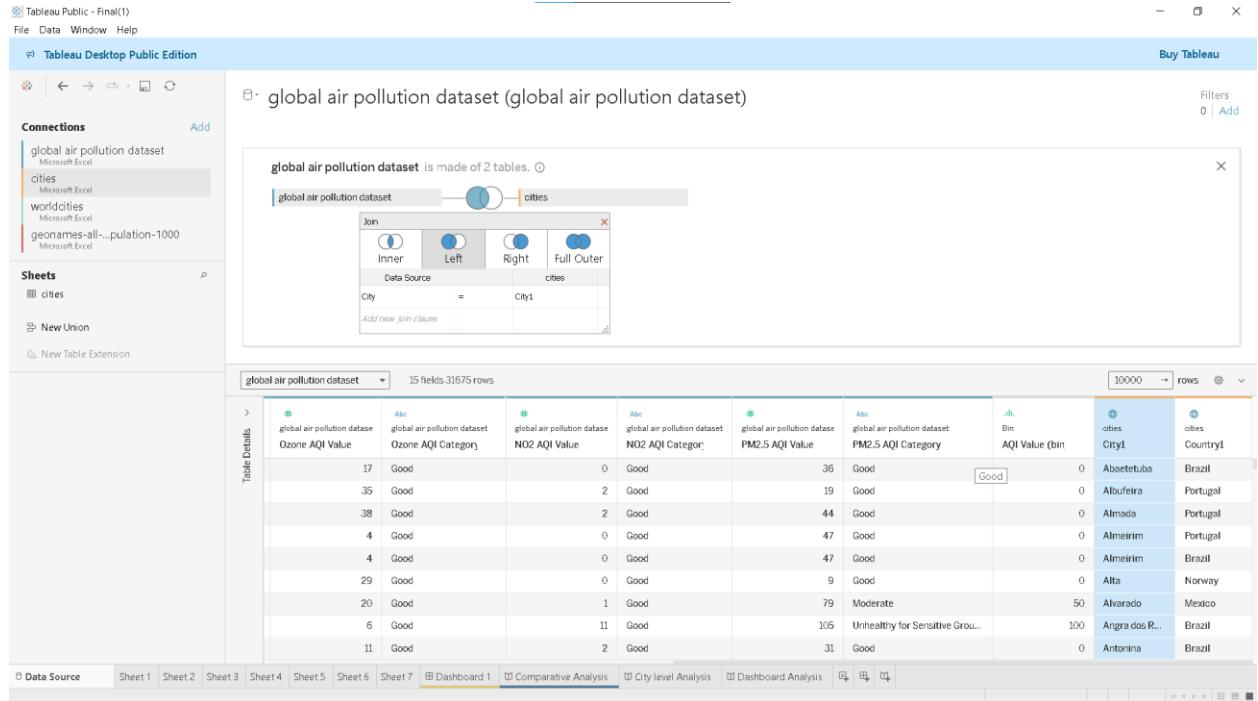
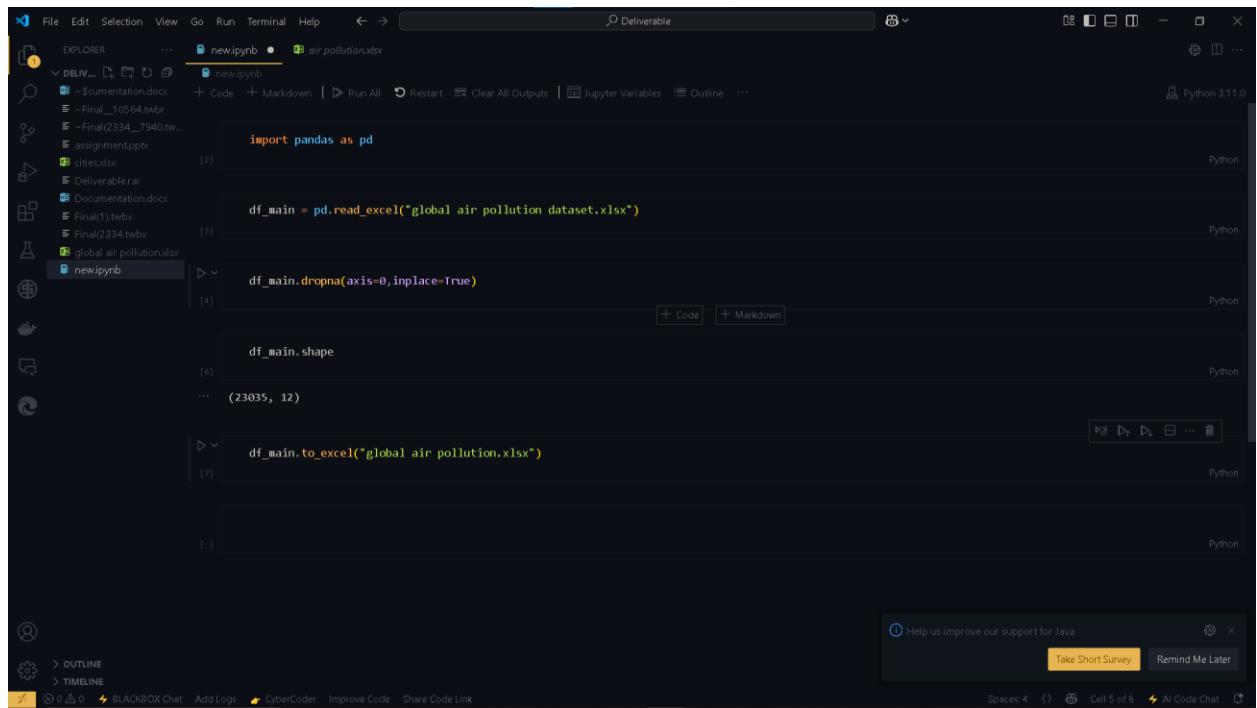


Figure 3. 5

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.



```

File Edit Selection View Go Run Terminal Help ⏪ ⏪ Deliverable
EXPLORER new.ipynb air pollution.xlsx
Documentation.docx
~Final_10564.twb
~Final(2334_7940.tw...
assignment.pptx
dites.xlsx
Deliverable.rar
Documentation.docx
Final(1).twbx
Final(2334).twbx
global air pollution.xlsx
new.ipynb
import pandas as pd
df_main = pd.read_excel("global air pollution dataset.xlsx")
df_main.dropna(axis=0,inplace=True)
df_main.shape
... (23035, 12)
df_main.to_excel("global air pollution.xlsx")

```

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 surprises 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

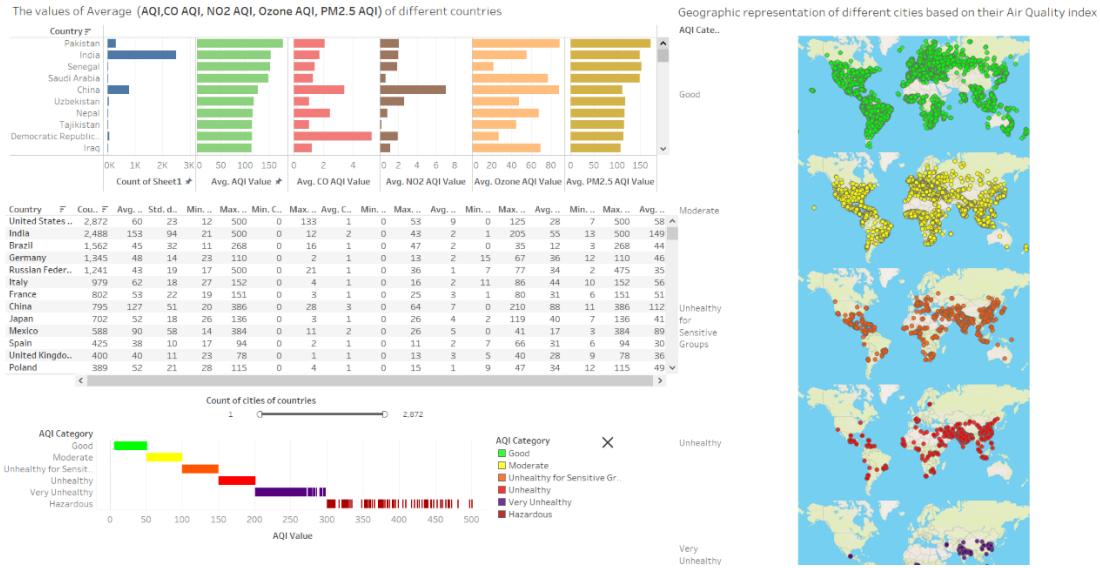


Figure 5. 1

This interactive dashboard allows users to 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.

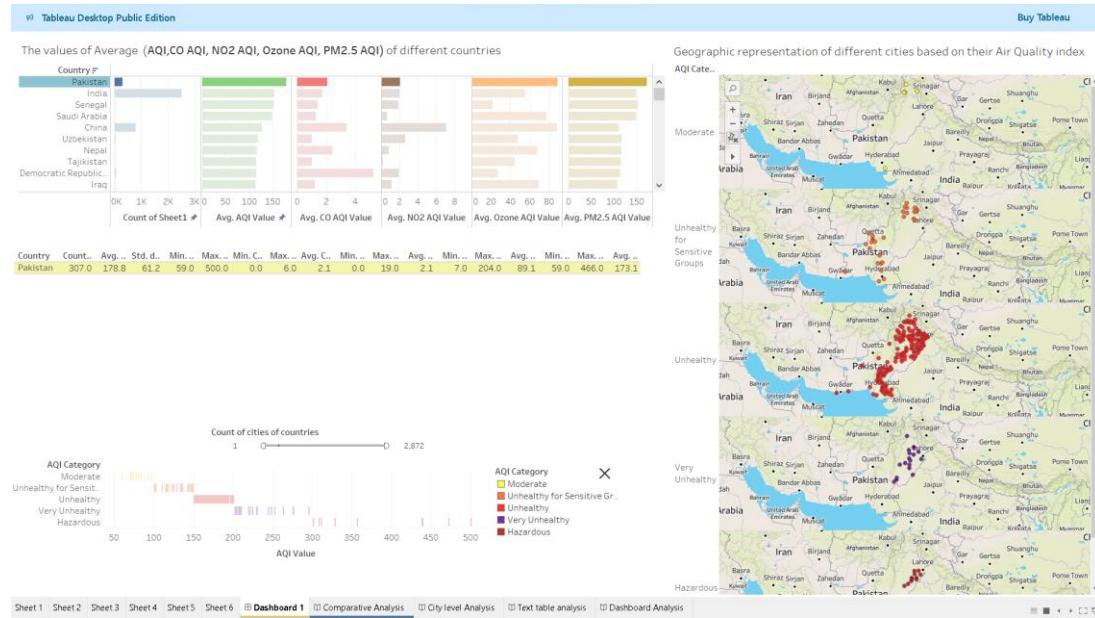


Figure 5. 2

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

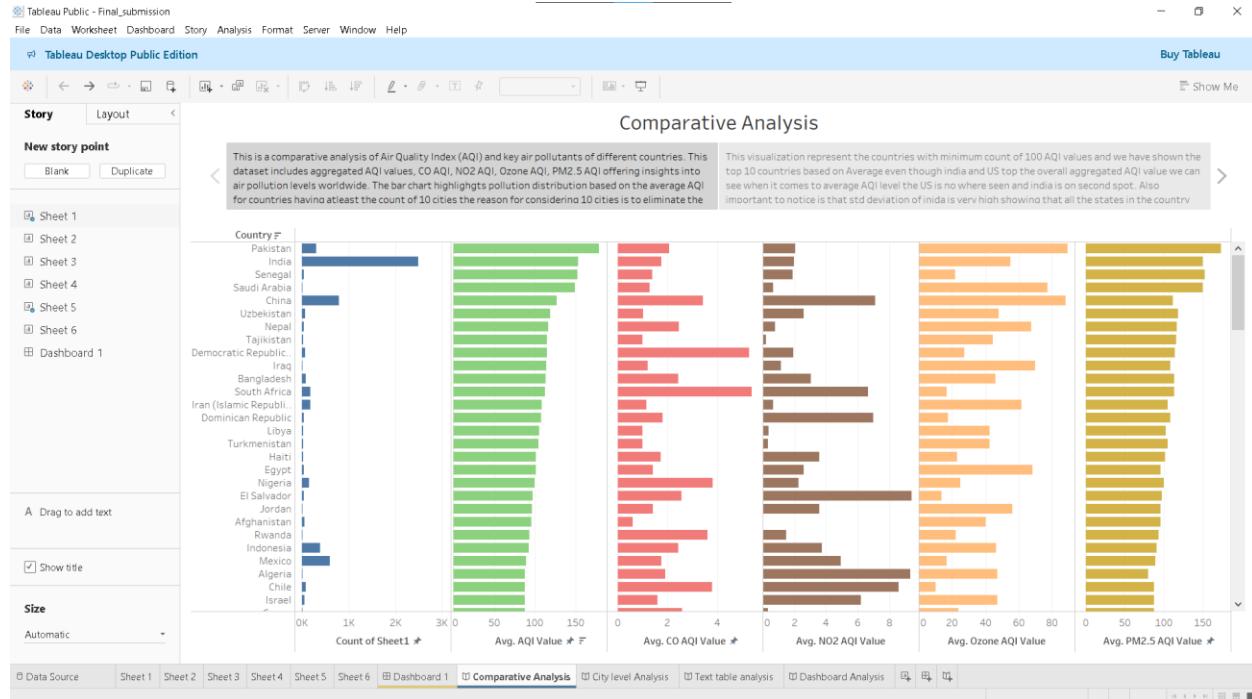


Figure 6.1

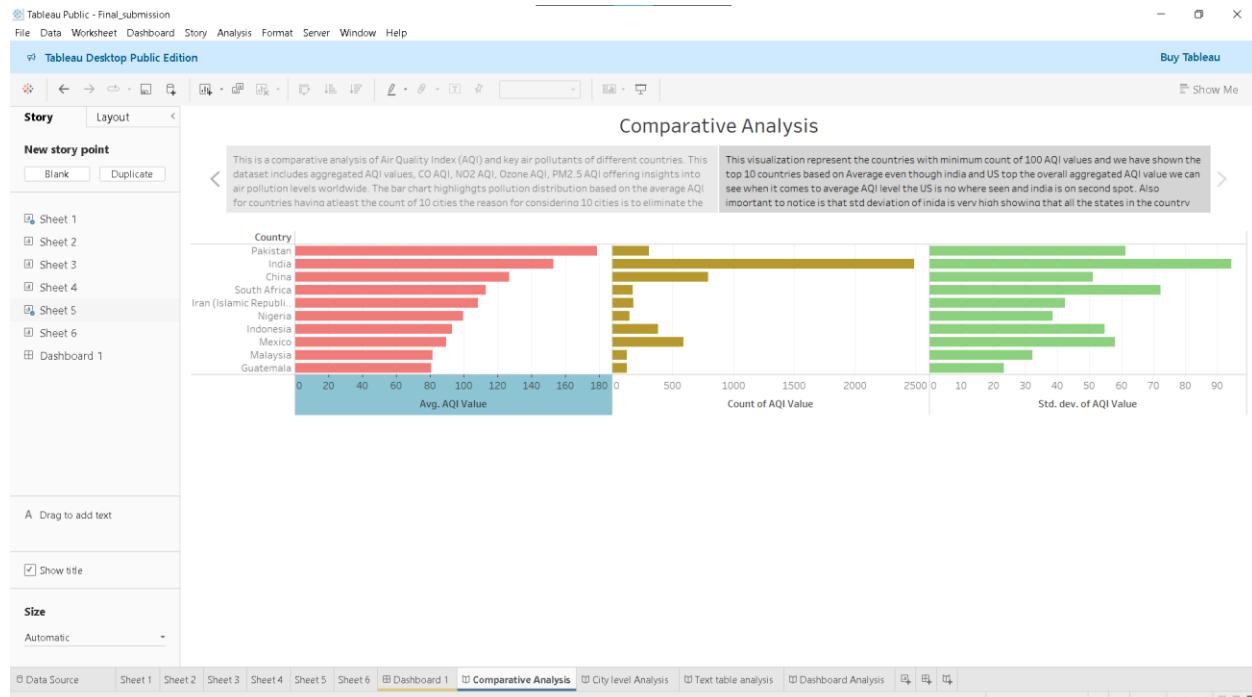


Figure 6.2

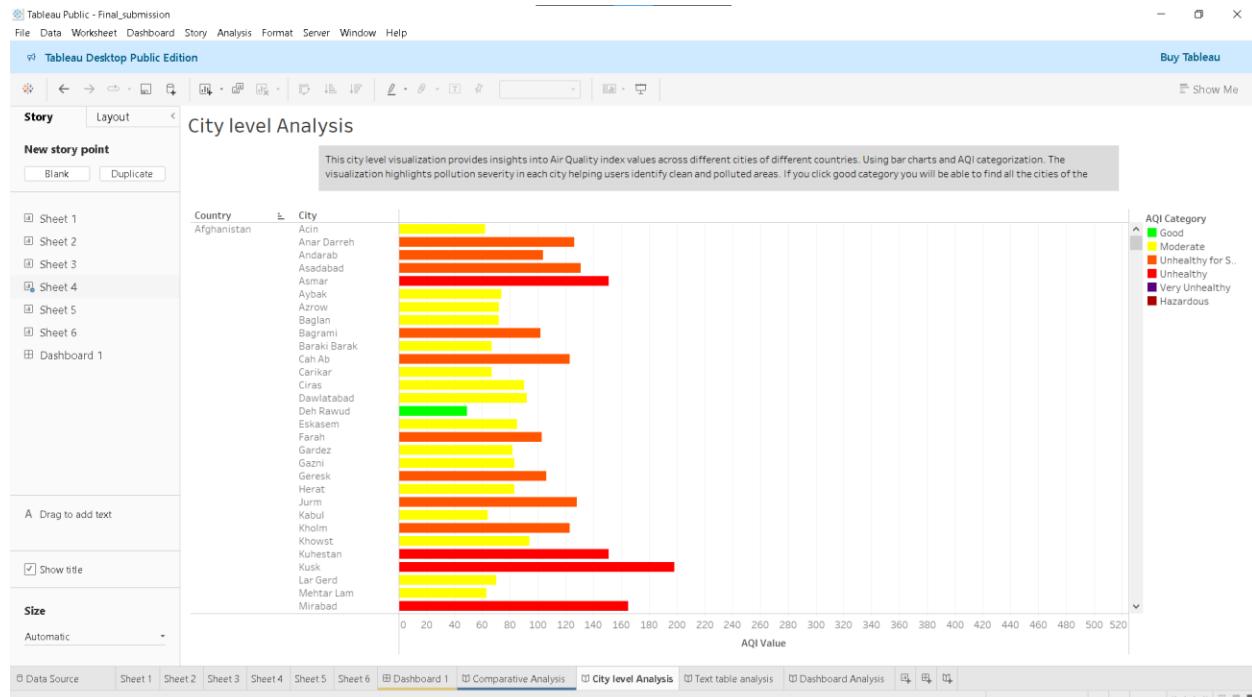


Figure 6. 3

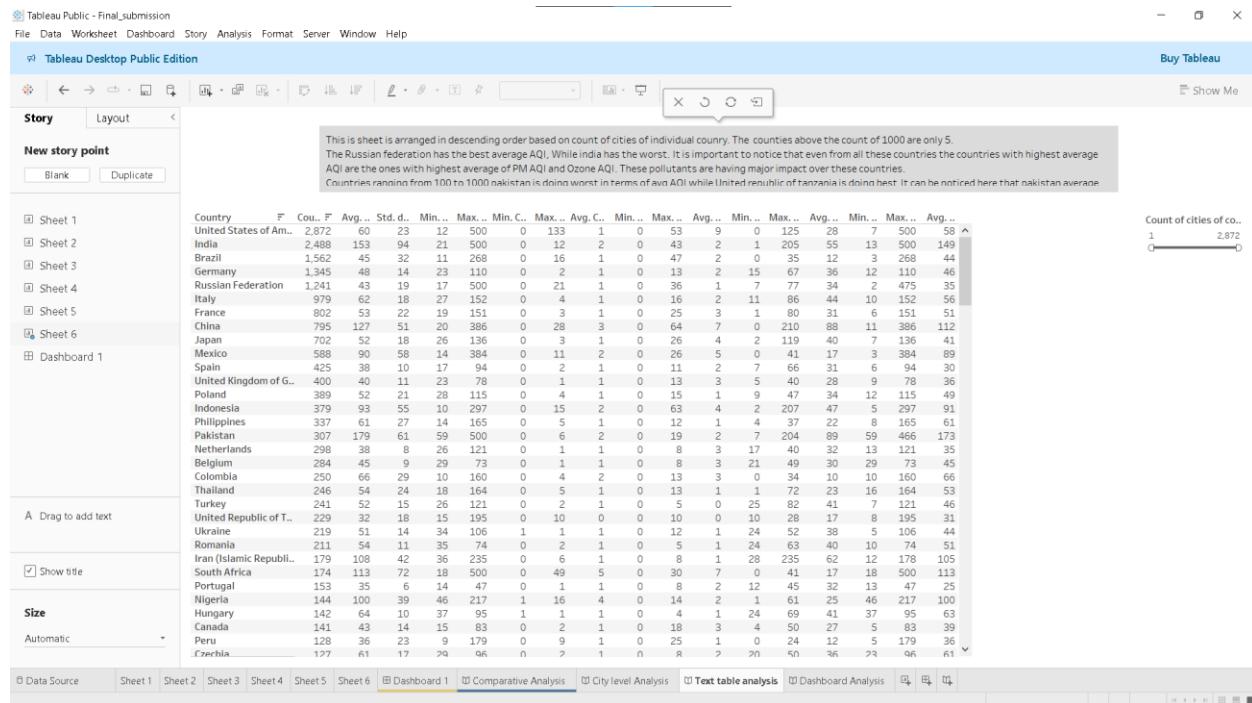


Figure 6. 4

7. Presentation

Presentation is provided assignment.ppt