**San José State University**

**Department of Applied Data Science**

**DATA 230**

**Data Visualization**

**Fall 2023**

**Section 27**

**Instructor: Andrew H. Bond**

**Term Project: Polluted Air: The Unseen Killer - A Global Data Visualization of Pollution and Death**

Baghavathula Sai Kiran

017404452

**Abstract**

The report explains how air pollution affects death rates around the world. For this project downloaded the dataset from Kaggle which includes columns like country, year, different kinds of deaths caused by air pollution (measured per 100,000 people), population numbers, GDP, urbanization area, energy consumption, air quality, healthcare spending, and policies related to pollution. Data cleaning is done using Pandas in Python.

Various types of graphs are plotted to analyze the data from different aspects. Bar charts show the overall deaths due to air pollution in various regions. Stacked bar charts compare the number of deaths from indoor air pollution, outdoor particulate matter, and outdoor ozone pollution. Line graph tracks how deaths related to air pollution have changed over time in Afghanistan, providing insights into the country's situation.

The project utilizes a multi-line graph to compare death trends from indoor and outdoor pollution, offering an understanding of the risks associated with various pollution sources. Additionally, it investigates the economic aspects of air pollution using basic and bubble scatter plots. These plots analyze how economic variables correlate with health impacts, considering variations in population sizes.

Furthermore, the project uses a choropleth map to illustrate regional variations in the Air Quality Index, giving a spatial view of air pollution intensity. A bubble map is also used to show the relationship between air pollution-related deaths and healthcare expenditures, shedding light on how healthcare efforts impact reducing deaths from pollution.

Moreover, the project includes a pie chart that displays the percentages of various causes of death within the population for the year 2017. In addition, it consists of a dashboard which gives a brief idea.

This project brings to light the significant impact of air pollution on global health, and at the same time, it reveals the complex connections among environmental elements, economic circumstances, healthcare provisions, and policy measures in determining health impacts associated with air pollution.

TABLE OF CONTENTS

ABSTRACT……………………………………………………………………………………2

DATASET DESCRIPTION……………………………………………………………………………..3

DATA CLEANING……………………………………………………………………………4

DATA ANALYSIS AND VISUALIZATION………………………………………………….9

1. BAR GRAPH…………………………………………………………….9
2. STACKED BAR GRAPH………………………………………………………10
3. LINE GRAPH ……………………………………11
4. MULTI-LINE GRAPH ……………………………….12
5. SCATTER PLOT ………………………………………13
6. BUBBLE SCATTER PLOT………………………14
7. CHOROPLETH MAP…………………………………………….15
8. BUBBLE MAP…………………………………………………16
9. PIE CHART …………………………………………………….17
10. Dashboard …………………………………………………….18

CONCLUSION ………………………………………………………………………………………….20

REFERENCES ……………………………………………………………………………………………20

**DataSet Description:**

|  |  |
| --- | --- |
| **Column** **Name** | **Description** |
| Entity | Country Name / Region |
| Code | Code of the country |
| Year | Year |
| Air pollution (total) (deaths per 100,000) | Total Deaths per 100000 |
| Indoor air pollution (deaths per 100,000) | Deaths due to indoor air pollution |
| Outdoor particulate matter (deaths per 100,000) | Deaths due to outdoor particulate matter |
| Outdoor ozone pollution (deaths per 100,000) | Deaths due to outdoor ozone pollution |
| Population | Population of the country |
| GDP | GDP of the country |
| Urbanization Rate | Urbanization rate of the country |
| Energy Consumption | Energy consumption by the country |
| Air Quality Index (AQI) | Air Quality Index of country |
| Healthcare Expenditure | Expenditure spent on healthcare by country |
| Legislation and Policy Index | Policies of country |

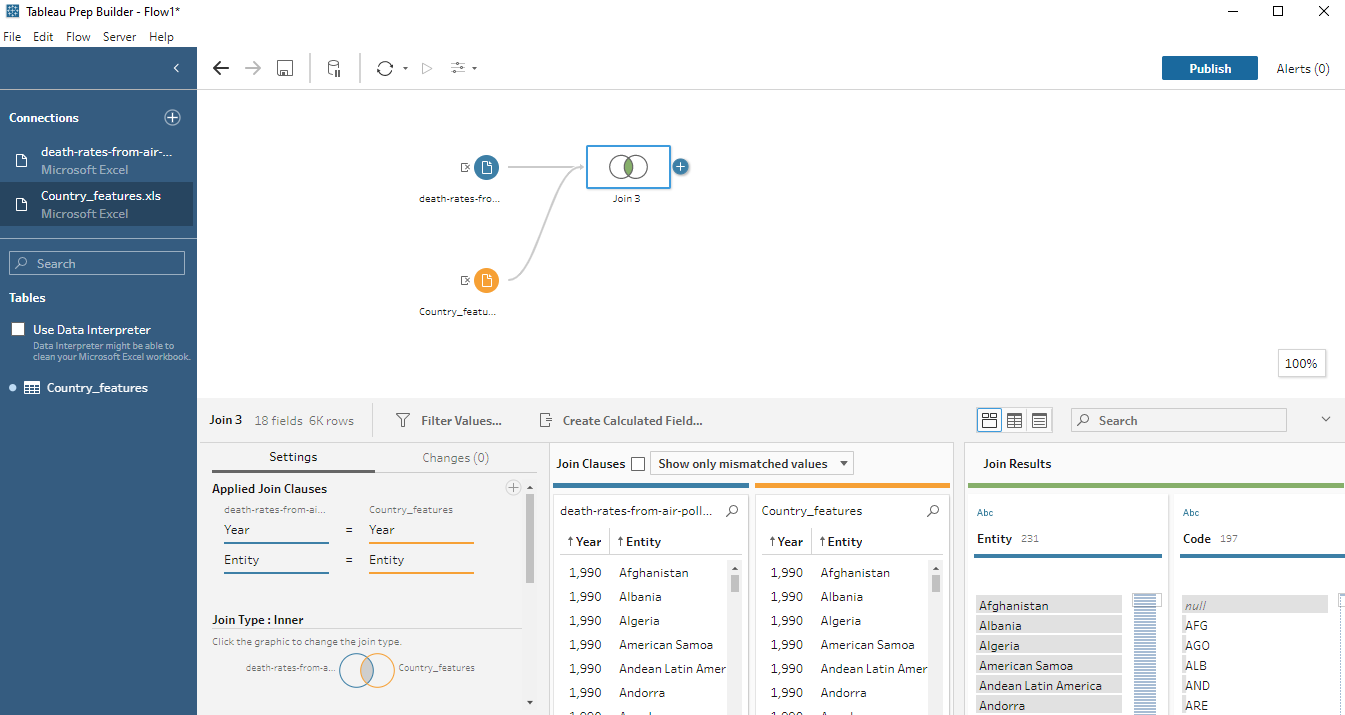
Table 1: Dataset column names and their descriptions

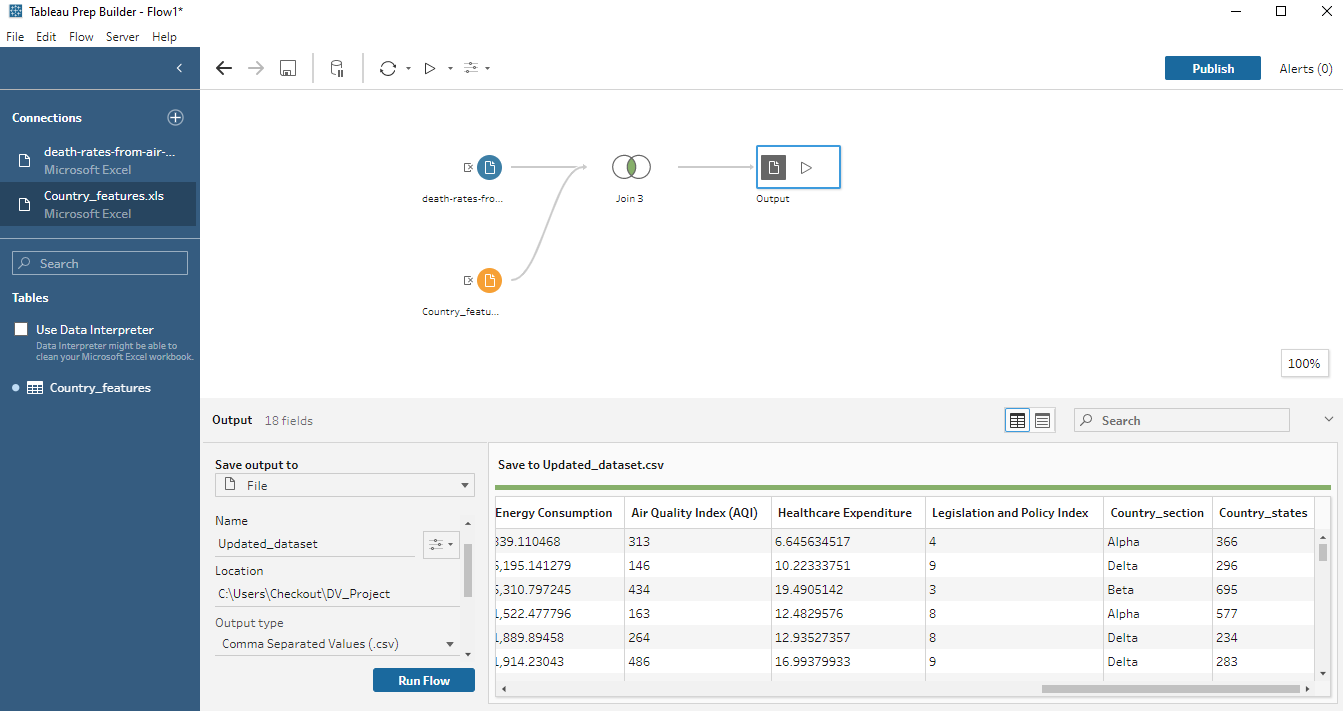
**Data Exploration and Cleaning:**

Downloaded the dataset (death-rates-air-pollution.csv) from the Kaggle and generated additional columns (Country\_features.csv) using ‘mockaroo’.

Combined these two csv files using tableau prep builder and named it as ‘Updated\_dataset’

The below are the images which illustrates the steps for joining the two files in tableau prep builder





Downloaded the updated\_dataset.csv from tableau prep builder

Used Pandas in Python for data cleaning process

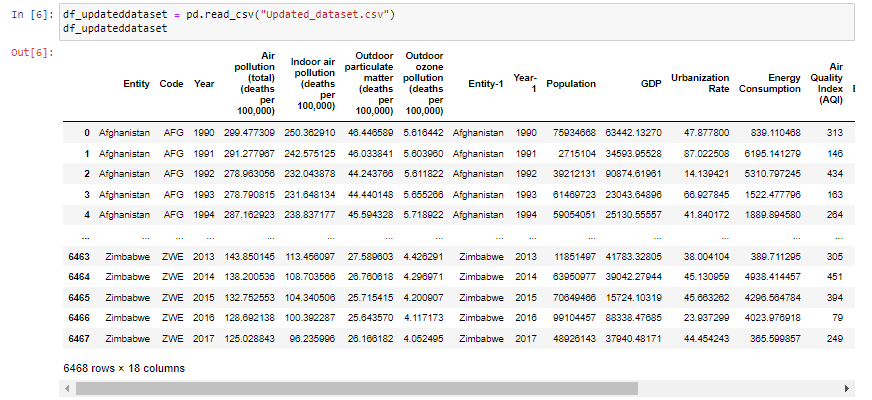
The below are the data cleaning steps:

* First imported packages which required for the procedure

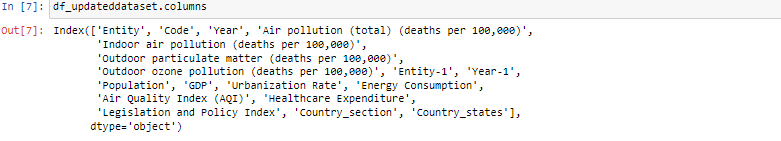
*import pandas as pd*

*import numpy as np*

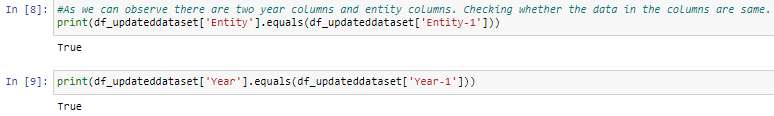
* Reading the csv file and displaying the file in data frame



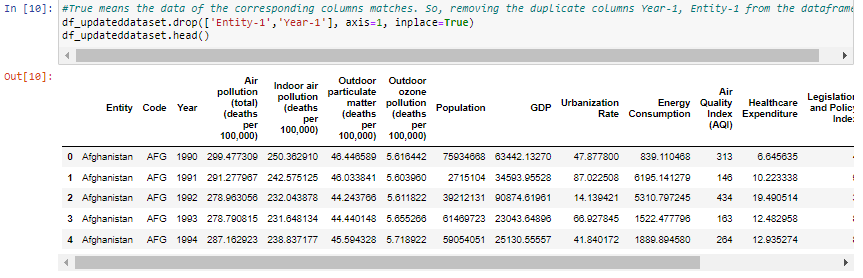
* Printing the column names in the data frame so, if any column names are same we can remove the column name after checking the data of both the columns



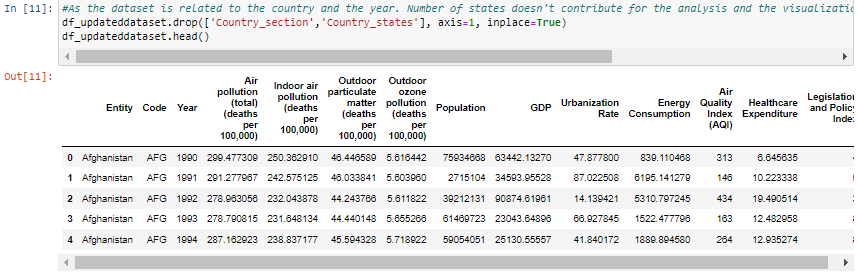
* In the above image there are two columns for the year and entity. So, checking the data of the two columns whether they are same or not



* Here true defines the data in the columns are same. Hence, removing the duplicated columns Year-1, Entity-1



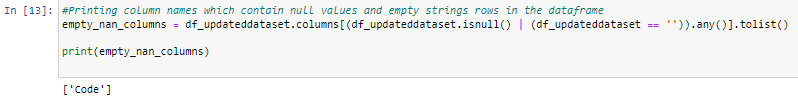
* The project is related to the country/region and pollution. The column names country states, Country section won’t contribute much to the aim of the project. Hence, removed the two columns



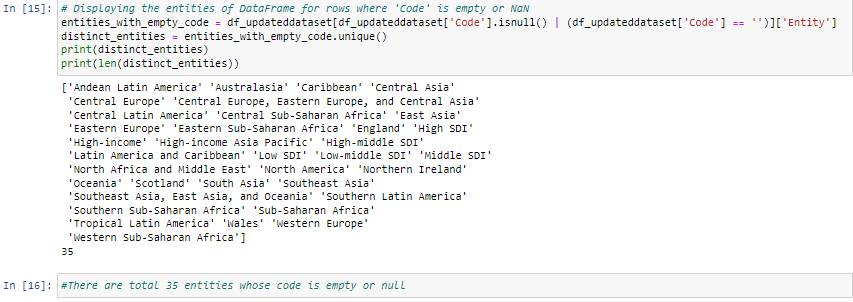
* Checking whether any duplicated rows are present in the dataset



* Checking any columns contains empty values or nan values

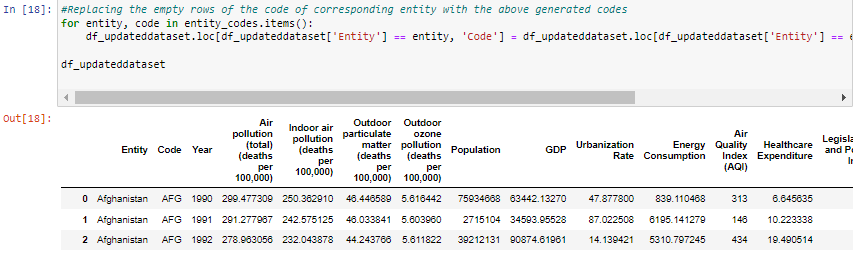


* In the above image it is observed that ‘Code’ column has empty value rows.
* Inserting the codes to the corresponding entities which doesn’t have code

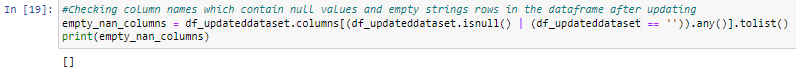




* Updating the data frame with the generated codes

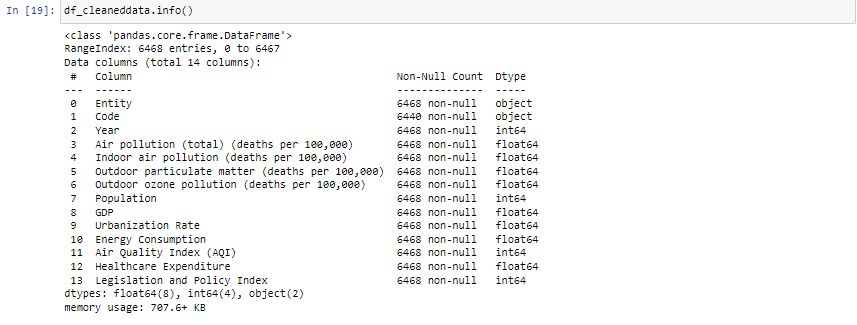


* For verification purpose again checking column names which contain nan or empty string rows



Empty list indicates there are no null values and empty strings in the columns

* Displaying the information about the cleaned dataset

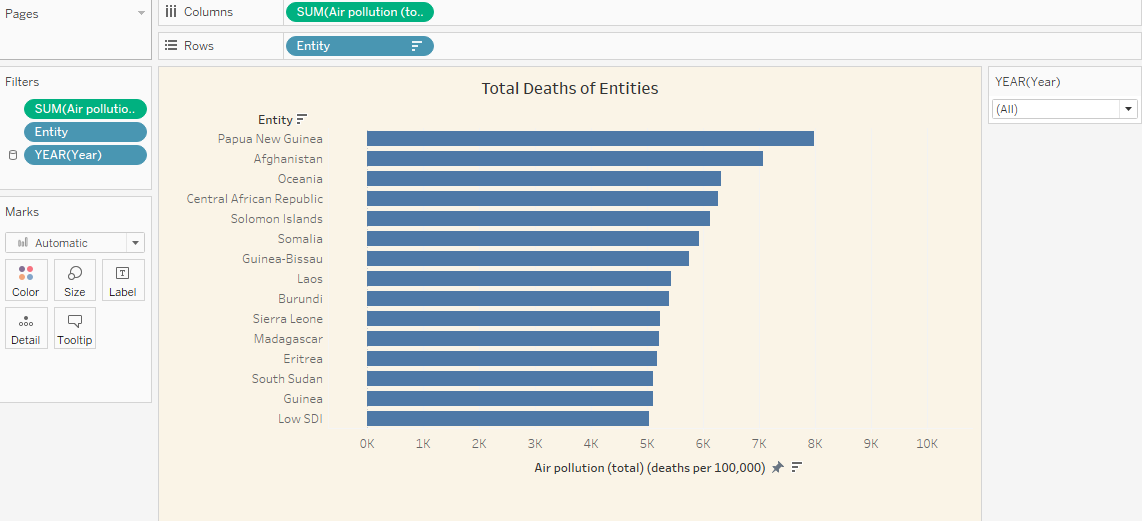


* Downloaded the cleaned data as ‘Cleaneddata.csv’ which is used for the project.



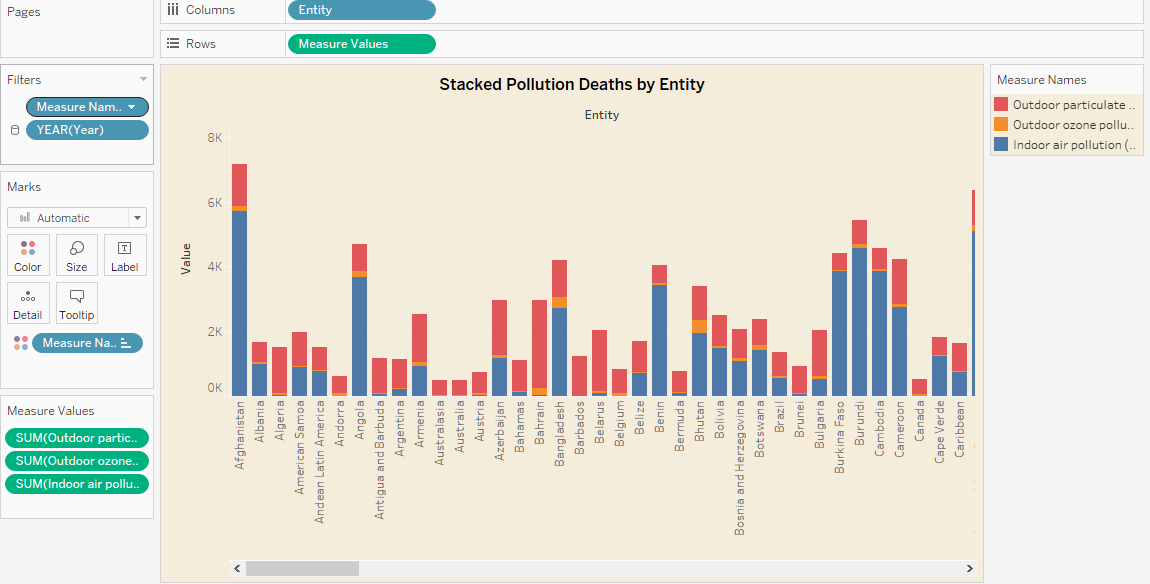
**Data Visualization Patterns:**

**Worksheet1: Total Deaths of Entities**



* To plot the bar graph placing ‘Air pollution Air pollution (total) (deaths per 100,000)’ in the columns shelf, ‘Entity’ in the rows shelf. Also adding ‘Year’ to the filter shelf for the interactive purpose.
* The above bar graph ‘Total Deaths of Entities’ shows the comparison of total deaths per 100000 of the various entities.
* Papua New Guinea, Afghanistan, and Oceania as the top three entities with the highest rates of deaths due to air pollution, each with rates more than 7,000 deaths per 100,000 individuals.
* The Central African Republic, Solomon Islands, and Somalia also exhibit high mortality rates due to air pollution, with each entity has approximately 4,000 to 6,000 deaths per 100,000 individuals.
* Entities like Guinea-Bissau, Laos, Burundi, and Sierra Leone show somewhat lower rates due to air pollution, ranging around 2,000 to 3,000 deaths per 100,000 individuals.
* The data shows there is a considerable variation in the death rates among different countries.

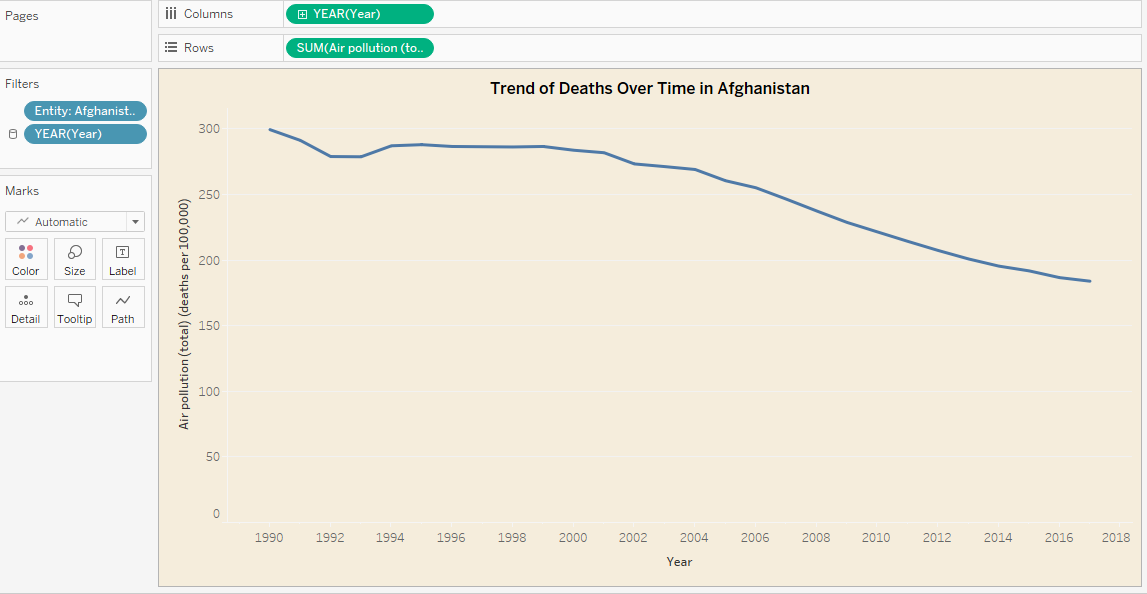
**Worksheet 2: ‘Stacked Pollution Deaths by Entity’**



* To plot the stacked bar graph inserted ‘Entity’ into the columns shelf, ‘Measure Values’ into the rows shelf. And filtered measure values only to the Indoor air pollution (deaths per 100,000), Outdoor particulate matter, Outdoor ozone pollution for stacking purpose.
* Red color in the graph indicates deaths due to outdoor particulate matter, orange color in the graph indicates deaths due to outdoor ozone pollution and blue indicates deaths due to indoor air pollution.
* The tallest bars in the graph indicates the highest rates of deaths, are seen in Afghanistan and Papua New Guinea. This indicates the pollution rate of air in the respective nations is high and needs to take the measures and implement policies to reduce the death rates.
* It is observed from the graph that large portion of deaths in most of the nations are due to the indoor air pollution.

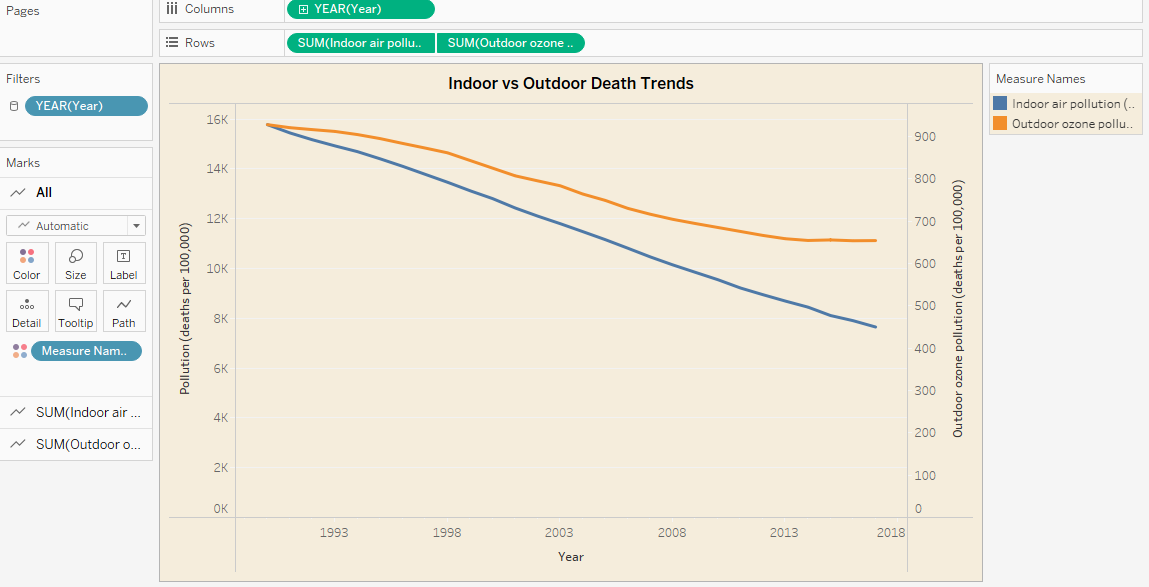
The above graph gives a clear representation of deaths due to different air pollutions and which is also helpful to take policy decisions and implement new rules to reduce the deaths.

**Worksheet 3: Trend of Deaths Over Time in Afghanistan**



* To plot the line chart placed ‘Year’ into columns shelf, Air Pollution (total)(deaths per 100000) into rows shelf.
* Added ‘Entity’ and ‘Year’ to the filter shelf and selected ‘Afghanistan’ for the entity.
* The line chart shows the trends of the deaths over time in Afghanistan.
* The mortality rate of the country changed significantly over the time.
* In 1990, the death rate is very high where it is approximately 300 deaths per 100000 individuals
* From 1996 to 2002 the death rate is almost equal in each year and there is no change in the mortality rate
* The mortality rate has been decreased from 2006, where it is recorded the least mortality rate in 2017 that is 183 deaths per 100000 individuals.

**Worksheet 4: Indoor VS Outdoor Death Trends**



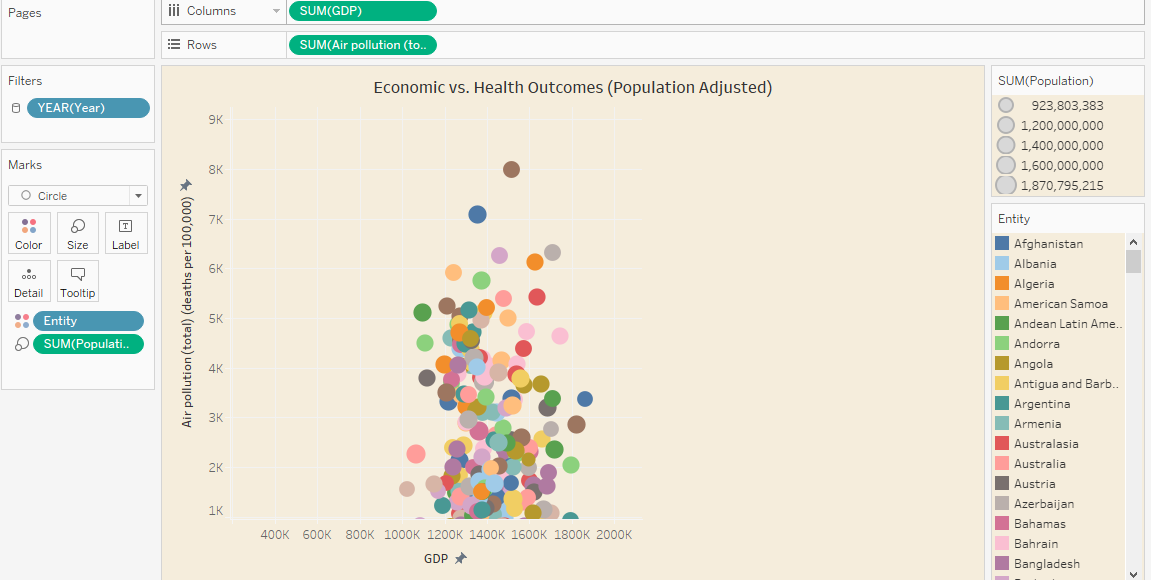
* To plot the multi-line graph, placed the ‘Year’ columns shelf, ‘Indoor air pollution’ and ‘Outdoor ozone pollution’ in rows shelf.
* The blue line indicates the trend of deaths caused by indoor air pollution and orange line indicates the trend of deaths caused by outdoor ozone pollution.
* Year is placed in the filters for the interactive purpose and it allows to choose the years
* Measure Names is placed in Color section of the Mark shelf, which gives the blue and orange color to the lines.
* Both the pollution deaths showing downward trend over the time
* This down trend indicates the overall improvement in the health of the people.
* Indoor air pollution deaths started at 16000 deaths per 100000 individuals, and decreased to 7000 deaths per 100000 individuals in 2017
* This indicates there is an improvement in cooking fuels and ventilators over the time
* Outdoor ozone pollution started at 927 deaths per 100000 individuals, and decreased to 654 deaths per 100000 individuals.
* There is a constant decrease in deaths caused by indoor air pollution over the years.

**Worksheet 5: Economic Impact on Health**



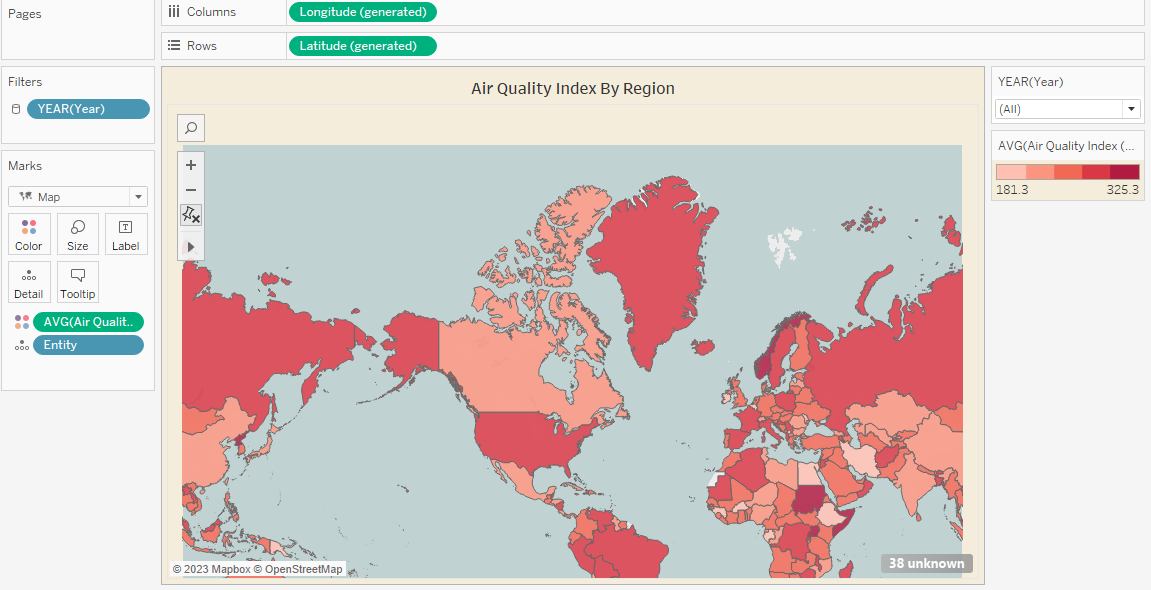
* To plot the scatter plot, placed the ‘GDP’ in columns shelf, placed ‘Air pollution (total)(deaths per 100000)’ in rows shelf .
* Inserted year into filters shelf for the interactive purpose and it allows to choose the year.
* The scatter plot compares the GDP of various entities and the death rates.
* The data are clustered at the lower end in the GDP scale and the wide spread in number of deaths to air pollution.
* As GDP increases, the death rates are decreased due to air pollution are seen by fewer data points in the graph. It indicates wealthier entities can afford better healthcare system, and also making the policies to reduce the deaths due to air pollution
* The graph doesn’t show a linear relationship between GDP and deaths. So, there might be other factors which are also making an impact on the deaths.

**Worksheet 6: Economic VS Health Outcomes (Population Adjusted)**



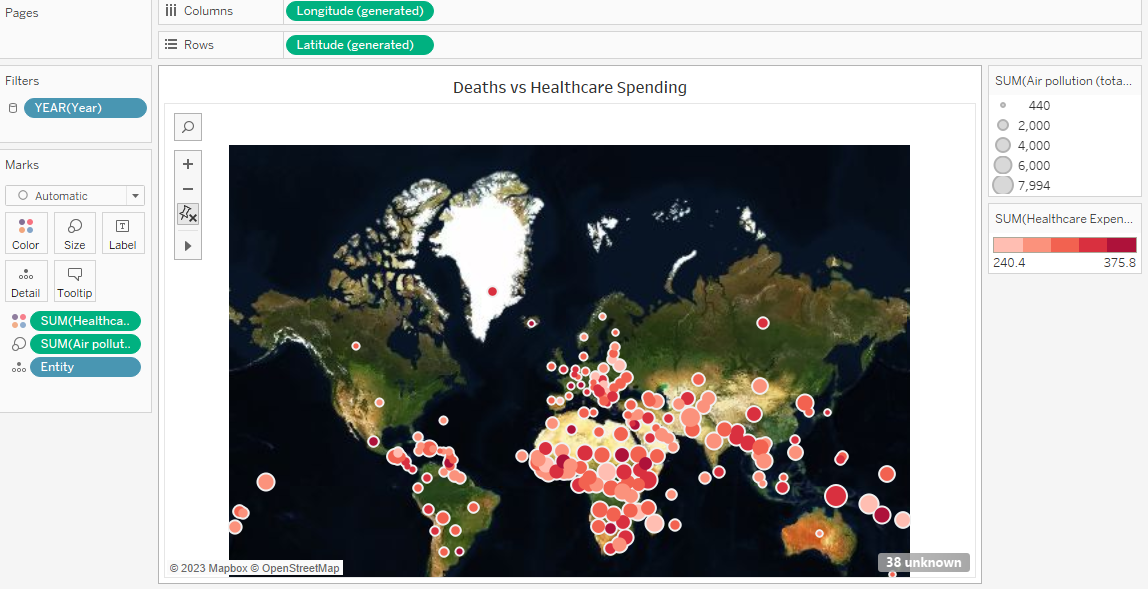
* To plot the graph placed GDP in columns shelf, Air Pollution (total)(deaths per 100000) in rows shelf.
* Added Population to the size section in marks shelf. It displays the size of the bubble based on the population.
* To give different colors to each entity added entity to the color section in marks shelf.
* The size of the bubble indicates the population size of the entity.
* The GDP axis and Air pollution axis indicate that there is no clear linear relationship between economic status of an entity and deaths.
* There are few entities whose GDP is more and has less death rate.
* Few entities have low GDP, more population and high death rate, which indicates the less economically developed areas, so population may be more affected by air pollution, leading to a greater number of deaths.

**Worksheet 7: Air Quality Index by Region**



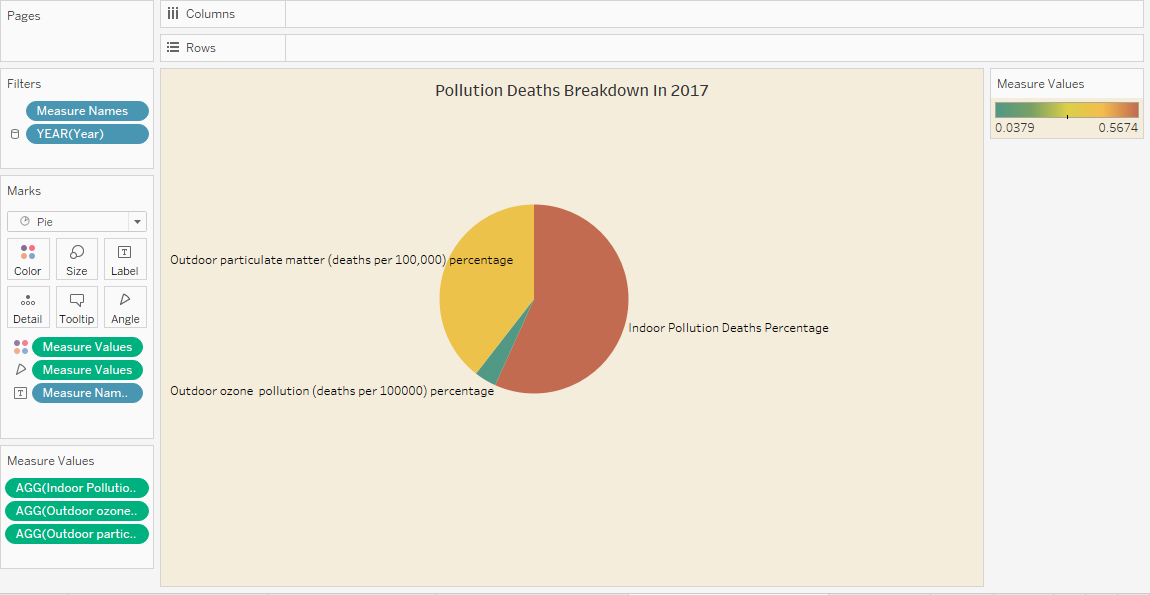
* To plot the graph placed latitude in rows shelf, longitude in columns shelf
* Adding air quality index to the color section of the marks shelf to differentiate the air quality index by country.
* Added year to the filter shelf, to make graph interactive
* The map uses a color gradient ranging from 181.3 to 325.3.
* The darker color indicates the air quality index of particular country is bad and the lighter color indicates the air quality index of particular entity is good or not bad.
* Some regions have high AQI, which indicates the urbanization, industrialization, or less stringent air quality regulations
* This map also provides a global overview, allowing for a comparison between continents and regions.

**Worksheet 8: Deaths VS Healthcare spending**



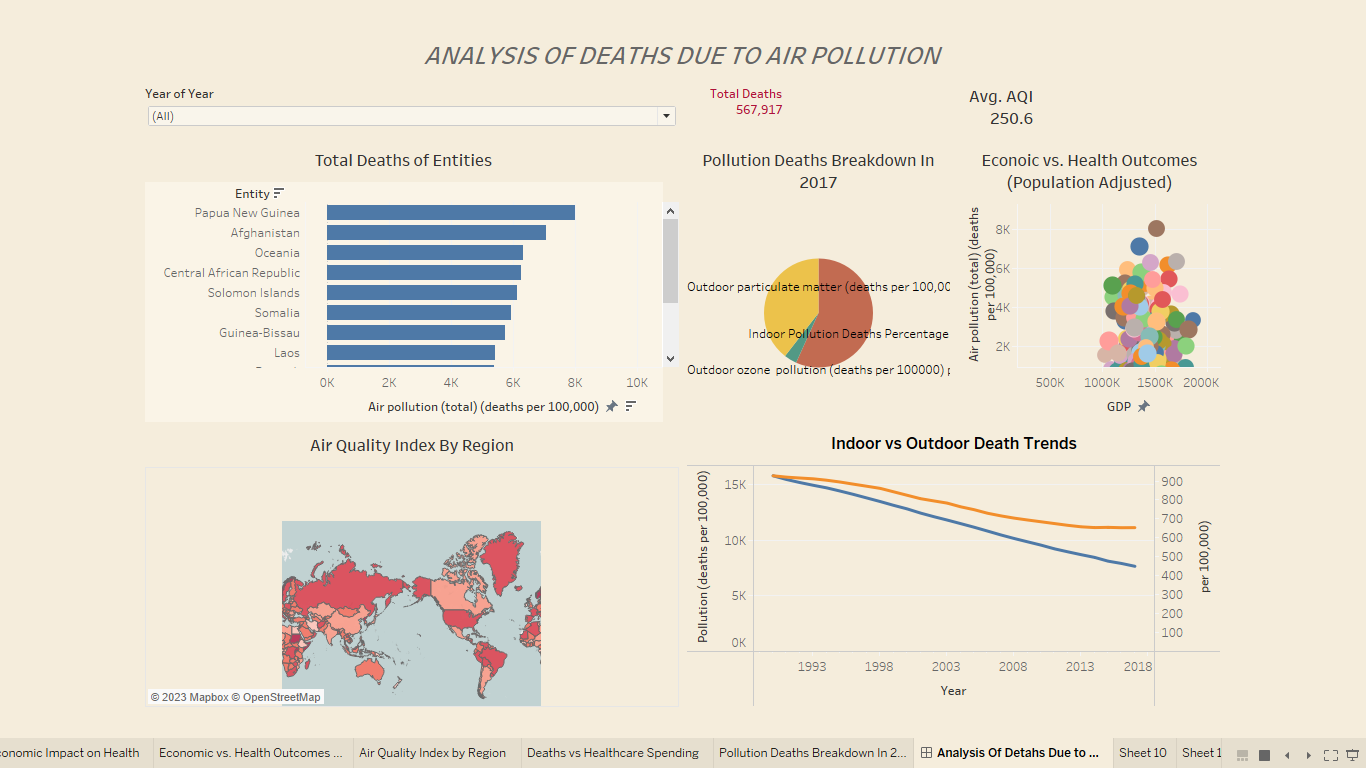
* To plot the graph place longitude in columns shelf and latitude in rows shelf. Health care expenditure is placed in color section of marks shelf and Air pollution (total)(deaths per 100000) in size section in marks shelf.
* The darker color indicates the high health care expenditure and lighter color indicates the low healthcare expenditure.
* Size of the bubble indicates the deaths of the particular entity
* There is no inverse relationship between color of healthcare expenditure and size of the bubble.
* This tells us that higher health care spending does not correlate with the lower air pollution related mortality rates

**Worksheet 9: Pollution Deaths Breakdown in 2017**



* To plot the pie graph placed Measure names, Year in filters shelf and placed measure values in color and angle section in marks shelf.
* Pie graph represents the pollution deaths caused by different air pollution in 2017.
* As per the graph, a greater number of deaths due to the indoor air pollution. Indoor air quality is affected by many factors such as solid fuels used for cooking and heating, which are most common in many developed countries.
* The second largest portion of the pie chart is Outdoor particulate matter. After, indoor air pollution a greater number of deaths are caused by outdoor particulate matter. Outdoor particulate matter originates from sources like transportation, industry, dust storms.
* The deaths caused by outdoor ozone pollution is very low compared to the other two types of pollutions.

**Dashboard: Analysis of Deaths Due to Air Pollution**



* The dashboard “ANALYSIS OF DEATHS DUE TO AIR POLLUTION”, it provides a view of the impact of air pollution on death rates across the world.
* The bar chart in the dashboard represents the total number of deaths due to air pollution for particular entity
* The world map in the dashboard represents the air quality index (AQI) of the different entities.
* Multi line graph shows the trend of indoor and outdoor pollution deaths.
* The scatter plot shows the relationship between GDP of the nation and deaths of the nation.
* There is also a year filter on the top right of the dashboard, it allows to select the desired years and the graphs change accordingly.
* The dashboard also displays the total deaths and average air quality index on the top right corner.

**Use/Test Cases:**

To validate the insights derived from the above visualization graphs, here are some hypothetical test cases.

Test Case A: Air Quality Index by Region (Choropleth map)

* Agencies related to environment and government can monitor the AQI of countries. Take the measures as required.
* Government can educate people about air quality in different regions

Test Case B: Total Deaths of Entities (Bar graph)

* It is used to evaluate the health policies by the international agencies in different countries.
* Can also used to identify which countries need the urgent healthcare interventions.

Test Case C: Stacked Deaths by Entity (Stacked bar chart)

* To guide governments in preparing targeted environmental regulations
* NGOs can raise awareness among the public regarding the particular pollution hazard in different regions

Test Case D: Trends of Deaths Over Time in Afghanistan (Line graph)

* It helps the government to compare the death rates over time and review and change the policies as per the requirement.

Test Case E: Economic Impact on Health (Scatter Plot)

* Economists can draw the correlation between GDP and the public health.
* So, they can recommend the government to allocate the budget for healthcare based on GDP.

Test Case F: Pollution Deaths Breakdown in 2017 (Pie Chart)

* Governments can review and compare the deaths caused by the different air pollution

in 2017. It also helps in planning and distribution of healthcare resources.

**Conclusion**

Air pollution is a serious concern, which is affecting the public’s health silently. By considering array of graphical representations, including bar, line, scatter and map, it effectively analyzes and break down the impact of air pollution on public’s health. It also considered features of the country like GDP, health care expenditure, Policy index which is used to draw correlation between the features and public health. The above visualizations not only show the current state of air pollution – related health issues around the world and also emphasize the necessity of coordinated efforts in healthcare strategies, policy interventions, and environmental management. The above-mentioned use/test cases can take into consideration serves to address air pollution, which is a dangerous threat to the world.

**References**

Github: <https://github.com/SaikiranBSK10/DeathsDueToAirPollution_Visualization>

Kaggle Dataset: <https://www.kaggle.com/datasets/akshat0giri/death-due-to-air-pollution-19902017>

Generated data using mockaroo for population, GDP, Urbanization Rate, Energy Consumption, Air Quality Index (AQI), Healthcare Expenditure, Legislation and Policy Index: <https://www.mockaroo.com/>

Lecture Notes from DATA230