***1. Description and Datasets used:***

Original Dataset:

<https://www.kaggle.com/datasets/iamsouravbanerjee/cause-of-deaths-around-the-world>

Scrapped data from:

https://en.wikipedia.org/wiki/List\_of\_countries\_by\_air\_pollution

Extra data that helps us fill more years with pollution levels during project 2 from:

https://www.kaggle.com/datasets/kweinmeister/pm25-global-air-pollution-20102017/data

The dataset provides comprehensive information on causes of death worldwide from the year 2000 to 2019 to work only on data from last 2 decades, to avoid working on older data which might cause historical outliers. It includes data on death rates and causes categorized by various countries. The dataset includes a wide range of causes of death including infectious diseases, chronic illnesses, external causes such as accidents and violence, and natural disasters.

The goal of using this dataset is to find patterns in the causes of death to help researchers, policymakers, and healthcare professionals identify areas of importance, evaluate the efficacy of public health campaigns, and develop strategies to lower mortality rates.

We have further divided the data between Chronic, Non-chronic illnesses and other factors, to help pinpoint more trends in our data and help analyze it in a more comprehensive matter.

**Added description:**

In Project 2, we have set a hypothesis that pollution plays a major role in most illness factors and some of the other factors.

We are collecting data that shows the amount of PM2.5, which refers to particulate matter with a diameter of 2.5 micrometres or less. These particles are extremely small and can be composed of various substances such as; dust, dirt, soot, smoke, and liquid droplets. PM2.5 particles are an issue because of their tiny size, which allows them to penetrate deep into the respiratory system when inhaled and affects all aspects of our health.

Furthermore, PM2.5 pollution not only affects human health but also has adverse effects on the environment. It can impair visibility, contribute to acid rain, and harm ecosystems by depositing harmful substances onto soil and water bodies and affect weather patterns and temperature as it contributes to climate change.

The data was already collected during this phase through scrapping and merging the data, and we will be exploring our findings during this document through visualization insights.

***2. Goals for Project 2:***

1. How air pollution (PM2.5) levels relate with each of Chronic, non-chronic and other factors when it comes to causes of death in general?

2. Which region has the most deaths due to Respiratory Infections and Chronic Respiratory Diseases and do they increase or decrease together, which means they are directly related and are they directly related to pollution levels?

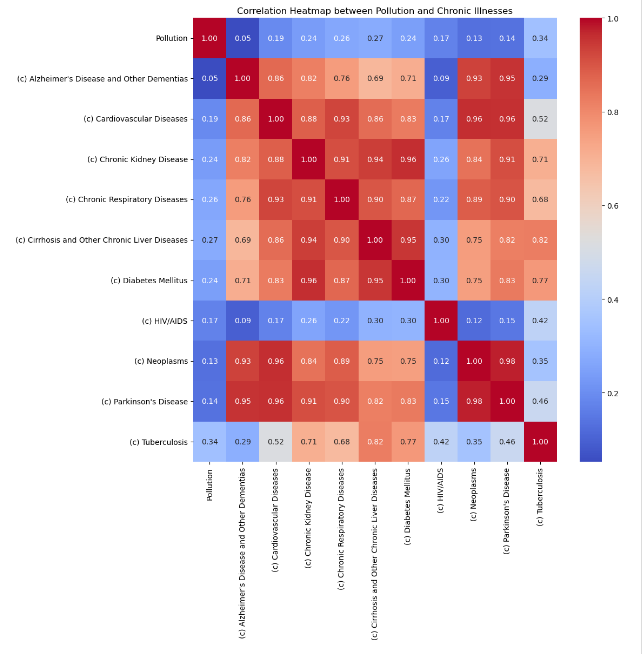
3. Is there any direct correlation between pollution and number of deaths caused by forces of nature for all regions?

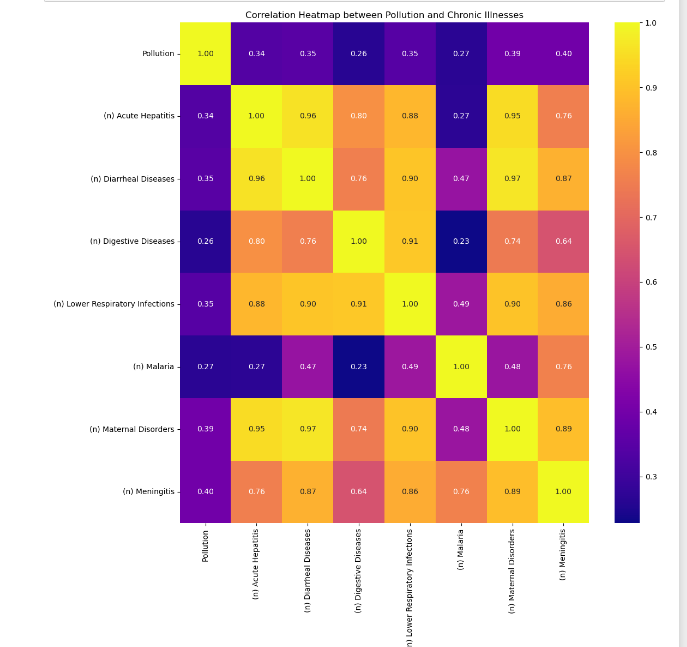
4. Which countries have the highest number of deaths caused by digestive diseases and what regions do they belong to and is there a common region between them and is it directly related to pollution rates?

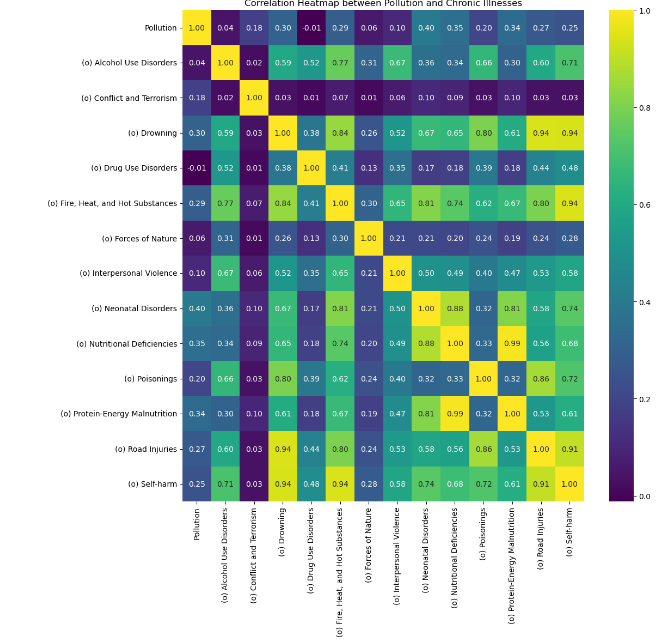
5. Which countries have the highest pollution rates, and is there a direct relationship between pollution levels and the incidence of poisoning cases in these countries?

***3. Insights from Visualization:***

Q1. We will be comparing the pollution with each of the categories alone one by one by using heat maps

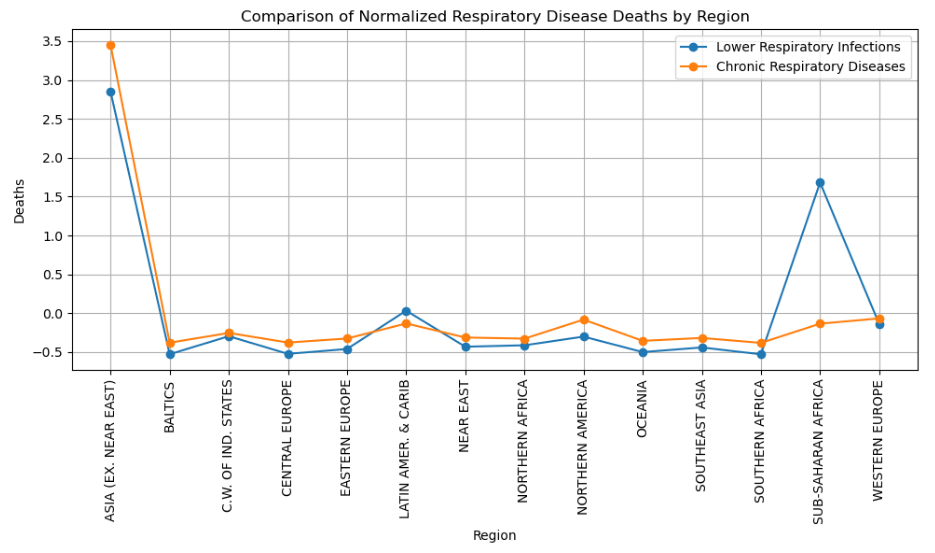




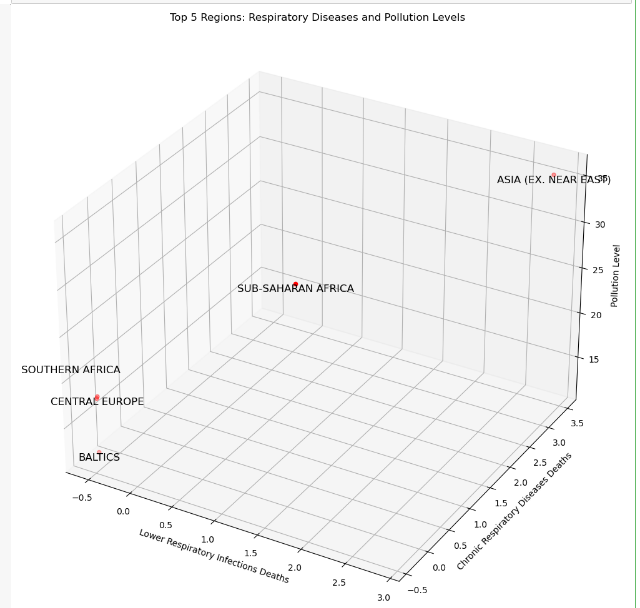


From this question we can see that there is barely any relationship between the pollution and other factors, but this is without considering data from regions or countries which could show a pattern in the rates of pollution and other causes of death. Furthermore, tests such as the T test and Chi-square can help us answer any hypothesis we have about the relation more accurately.

Q2. We will first see if there is any relation between both diseases for each region



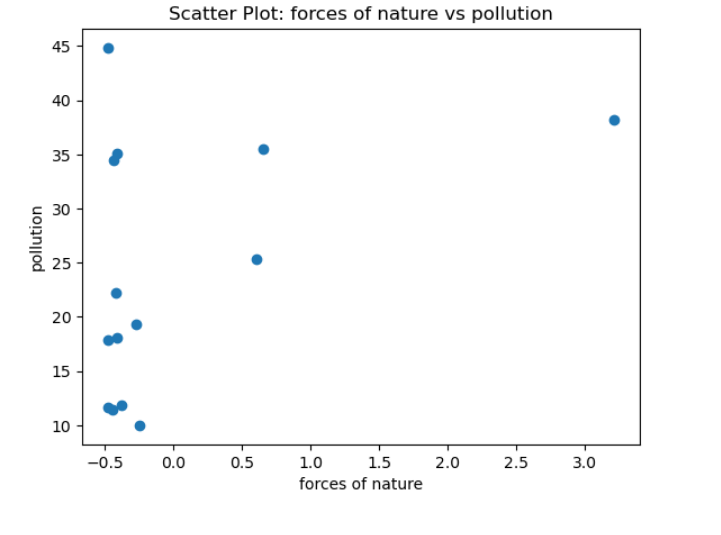
Then we will take the top 5 regions in both as a total and check their pollution numbers



On the other hand, this 3D scatter plot brought out a strong relation between pollution levels and respiratory disease death in all the regions (except the Baltics), considering that the average pollution of each region is above 5 and the recommended amount of PM2.5 is less than 5 by the WHO. The highest pollution levels which are present are linked with lower respiratory infections and chronic respiratory diseases are mostly in the Asia region and Sub-Saharan Africa region. This trend indicates that regions of high pollution also bear the greatest health burden from respiratory conditions, accentuating the potential health dividends of reduced environmental pollution. Moreover, the direct connection of the two respiratory disease types means that interventions directed against one may reap affirmative outcomes for the other. Integrated health approaches should take that into account.

Q3:

This code takes a data frame with regional data on Forces of Nature and Pollution, calculates average values per region, performs Z-score normalization on the Forces of Nature column, and then creates a scatter plot to visualize the relationship between the normalized Forces of Nature and average Pollution values across regions. While the initial impression might be that there's no clear correlation, it's worth noting the outlier on the right side of the plot. This region has high levels of both pollution and deaths caused by forces of nature, suggesting a potential connection that deserves further investigation.



As shown above, data has been grouped by region. Then a scatter plot used to obtain any correlation between forces of nature and pollution that indicates there is no relationship between 2 factors. This relation is obtained by checking the highest pollution levels that exist in regions with little to no death numbers caused by forces of nature and in regions where there is a significant number of deaths by the same incident. However, that one region at the right of the scatter plot graph that has a high number of incidents by death and and pollution may indicate that there maybe a correlation that exists

Q4:

The code calculates regional totals for Digestive Diseases deaths, identifies the top 20 countries/territories with the highest death counts, and then creates a bar chart to visualize those top countries and their respective death counts.

A graph of diseased diseases

Description automatically generated

this code extension uses the filtered data for top countries to create a pie chart that shows the regional distribution of deaths caused by Digestive Diseases within those top countries.

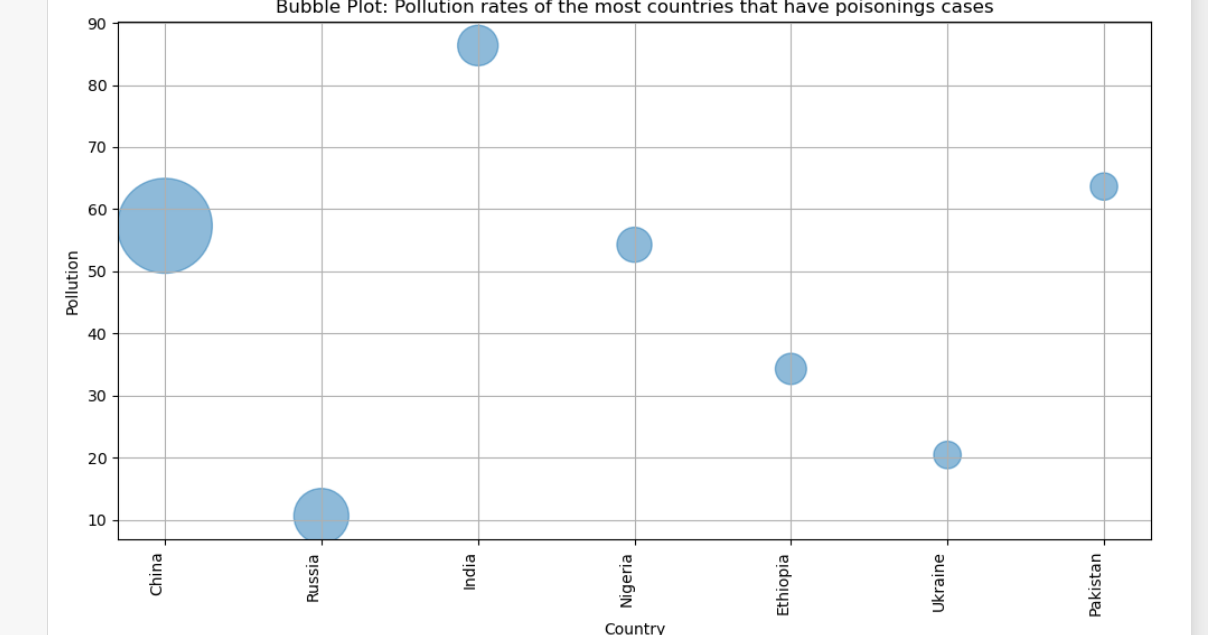
A pie chart with different colored circles

Description automatically generated

this code snippet calculates average deaths from Digestive Diseases for different pollution rates, selects a subset for visualization, and creates a bar chart to explore the potential relationship between pollution and these deaths.

Q5.

we will the sum of the poisonings Group by 'Country/Territory' and then Sort it in descending order, select the top 7 countries, put all of that in a dataframe, then filter it to have the top countries for the previous dataframe to have a new dataframe containing the previous top countries and its pollution rates then merge the two dataframes . after merging, we will create a bubble plot where each bubble represents a country. The x-axis represents the countries, the y-axis represents pollution levels, and the size of each bubble corresponds to the number of poisoning cases.



Highest country that suffers from pollution is India, it doesn't relate with the numbers of Poisonings cases as shown in the visualization above.

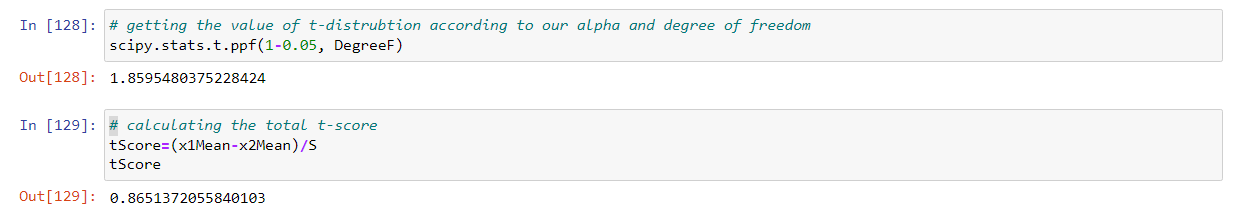
from another point of view Russia is from the countries that has the lowest pollution rates and has a average number of Poisonings cases

in conclusion pollution rates are not directly related with poisoning cases in most of the conditions.

***4.Null Hypothesizes:***

The T test was used in all of our hypothesis testing as it was the most fitting for all 3 with a set alpha of 0.05.

Hypothesis 1: Pollution rates do not have any relationship with the prevalence of digestive diseases leading to mortality in the highest regions that suffer from high numbers of death due to digestive diseases.

******

Our T score = 0.86

The P value by the degree of freedom 8 and alpha of 0.05 = 1.85

As shown above, since our value is lower than the table value so we accept the null hypothesis, which means there is no relationship between pollution rates and death caused by digestive diseases.

However, despite the data showing that we are accepting the null hypothesis, there could be a chance for false negative which is a Type II error, depending on possibility other variables, another explanation is the that impact of the pollution is the overall number deaths is small, or the existence of pollution could be the reason why there a lot of death related to these causes to begin with.

Hypothesis 2: There is no relationship between pollution rates and death caused by respiratory diseases in the top regions affected by respiratory diseases.



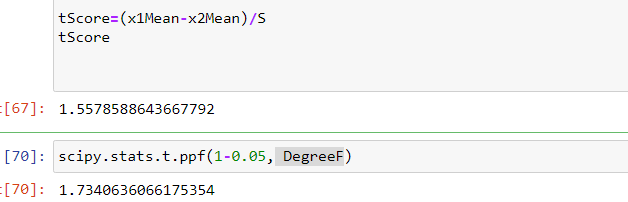
Our T score = 1.008

The P value by the degree of freedom 8 and alpha of 0.05 = 1.85

As shown above, we will accept the null hypothesis again and say there is no relationship between death caused by respiratory diseases and pollution rates as T Score < P table value

As mentioned in the first hypothesis again, despite the data showing that we are accepting the null hypothesis, there could be a chance for false negative which is a Type II error, depending on possibilities that could explain the reason or that pollution itself is the reason why the deaths exist.

Hypothesis 3: There is no relationship between countries which has the highest levels of pollution and the number of deaths by poisoning.



Our T score = 1.55

The P value by the degree of freedom 18 and alpha of 0.05 = 1.75

As shown above, we will accept the null hypothesis again and say there is no relationship between pollution levels and death incidents that is caused by poisonings as T Score < P table value

As mentioned in the two hypotheses again, despite the data showing that we are accepting the null hypothesis, there could be a chance for false negative which is a Type II error, depending on possibilities that could explain the reason or that pollution itself is the reason why the deaths exist.