Richa Rambhia
November 13, 2021

Final Project – Milestone 1
Exploratory Data Analysis

Introduction

Dataset:

The dataset for the Final Project – Milestone 1 was chosen from Kaggle and is a dataset for the number of deaths by risk factor. This dataset has 6468 rows of data and 32 data fields. The data fields include entity, code, year, alcohol use, drug use, air pollution, no access to handwashing facility, smoking, outdoor air pollution, iron deficiency, etc. The entity attributed basically is the country in which we measure the death rate, the year attribute would be the year of a particular country, and other attributes like the alcohol use, smoking, drug use, etc. are the risk factors on which the death rate in a particular country depends.

This dataset contains numeric as well as categorical data and is helpful in determining which risk factor in a particular country would affect the death rate. The reason for choosing this dataset is because while analyzing the dataset it can be found that which risk factor would highly affect the death rate and what is the death rate due to these factors in each country.

Data Cleaning:

Data cleaning is an important phase in the data analysis process and after understanding the dataset, data cleaning was performed. Here, the null values were checked for and removed. The dataset was pretty much clean already and no other work was needed to be done on this data.

Data Analysis

Descriptive Analysis:

Descriptive analysis includes describing the data and computing the statistical values of the numeric values of the dataset. At first, the entire dataset description was done which includes the summary of the dataset, the structure of the dataset, class of the variables of the dataset, etc. Later, after data cleaning, the statistical values of the attribute of the dataset were computed which helps in understanding the values of each variable.

The statistical values computed were minimum, maximum, mean, median, standard deviation, range, and summary. The values computed for Alcohol use are as follows:

Minimum: -2315
Maximum: 2842854
Mean: 50203
Median: 2803
Sd: 195822
Range: -2315 to 2842854

Similarly, values for other attributes were also computed which was helpful in understanding the values of each variable separately. The table below shows the statistical values computed for other attributes.

Statistical Value / Risk Factor	*Min*	*Max*	*Mean*	*Median*	*Standard Deviation*	*Range*
Air Pollution	8.52	4895476	95735.51	6125.09	390933.5	8.52 - 4895476
Smoking	11.707	7099111	133548.3	5935.789	529931.5	11.707 -7099111
Drug Use	1.240	585348.2	8890.242	408.5863	35415.12	1.240 - 585348.2
No access to handwash facility	0.0779	1239519	18933.05	252.4991	89810.37	0.0779 - 1239519

Data Visualization:

Data visualization step was performed once the descriptive analysis was done. But before beginning with the process of data visualizations, subsets of the dataset were created which would help in analysis as the actual dataset contained huge volume of data. In visualizations, various charts were created to better understand the data and further analyze the data. The graphs created were based on the entity attribute vs the different risk factors attribute. The graph would thereby represent the country where the risk factor attribute is lower or higher which would help in knowing which country is affected by a particular risk factor.

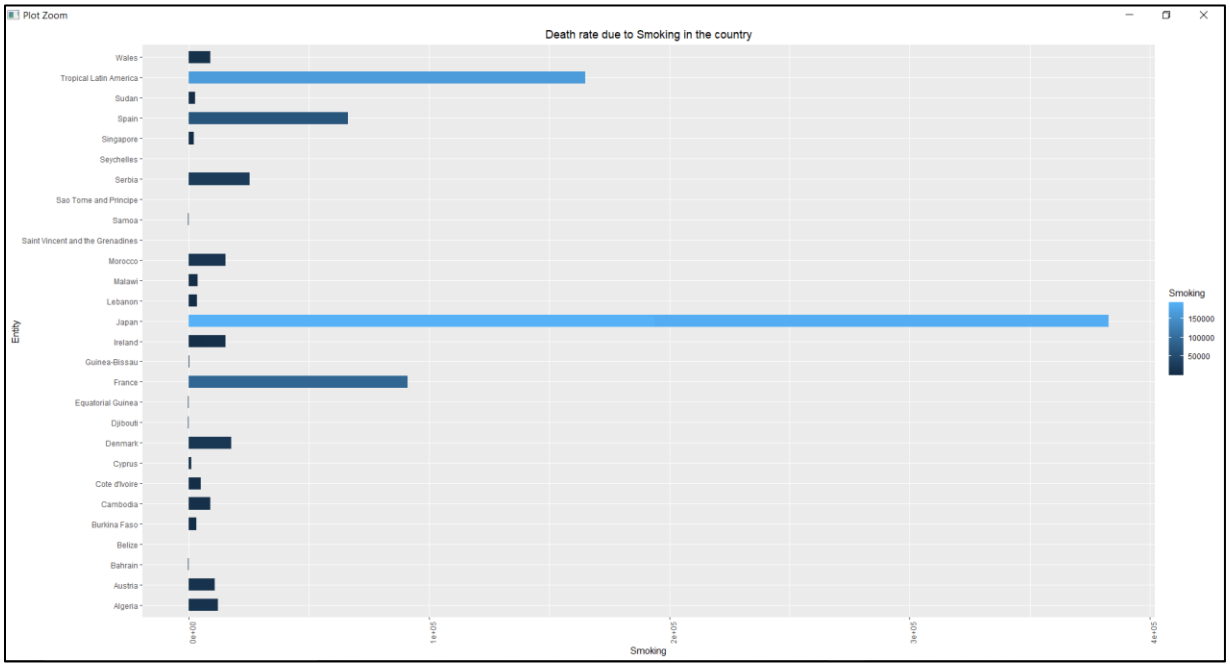
The first graph created was the death rate due to smoking in the country. Since the dataset was too large enough to analyze, a subset of the dataset was created which had some records selected at random and it was found that Japan had the highest death rate due to the risk factor of smoking followed by Tropical Latin America. Similarly, Middle SDI had the highest death rate due to alcohol use and Sudan had the highest death rate due to no access to hand wash facility.

The last graphs created were based on the filtered dataset where the graph for one particular country was created which had the death rate due to alcohol use and drug use in the year. The entity selected for visualizations were Afghanistan and India where the alcohol use and drug use in the year was understandable.

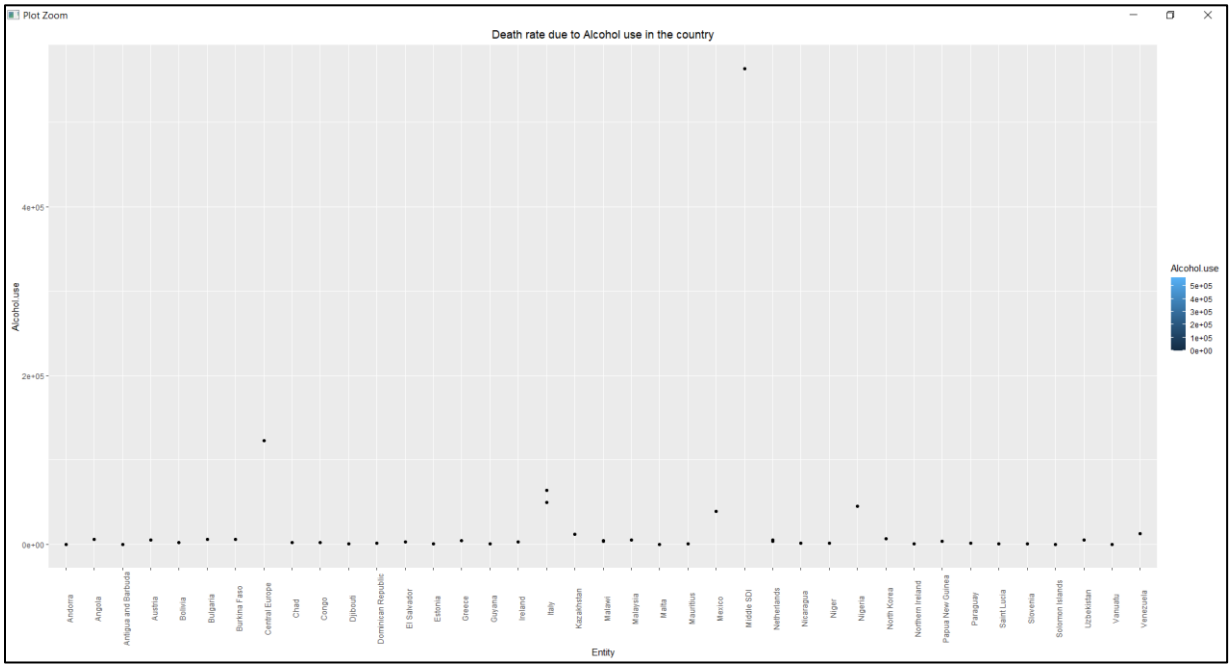
Visualizations:

(Since the data for visualizations is randomly selected, at each run of the code it would select a random data from the dataset which may produce different graphs based on the data randomly selected.)

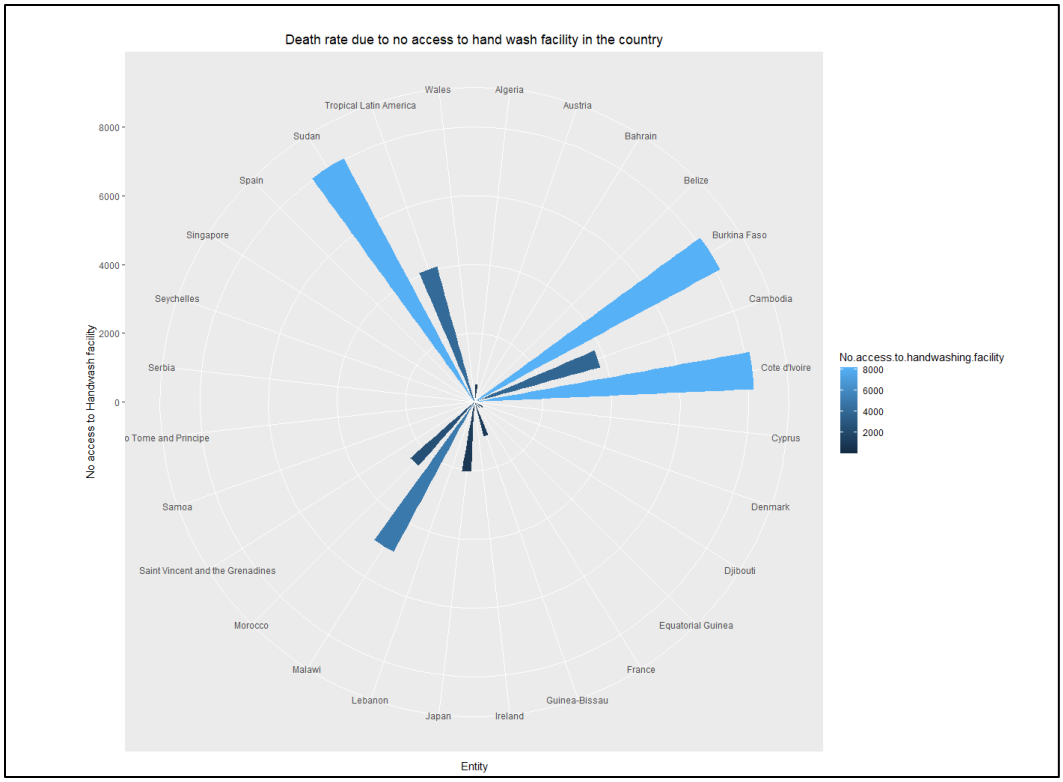
Graph 1: Death rate due to Smoking in the country



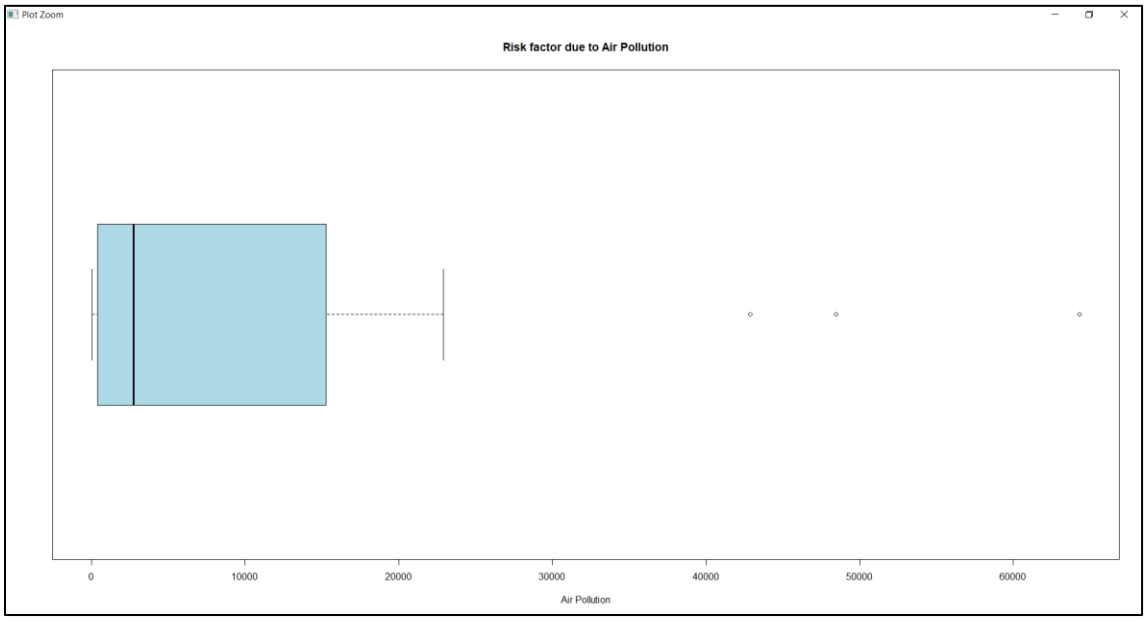
Graph 2: Death rate due to Alcohol use in the country



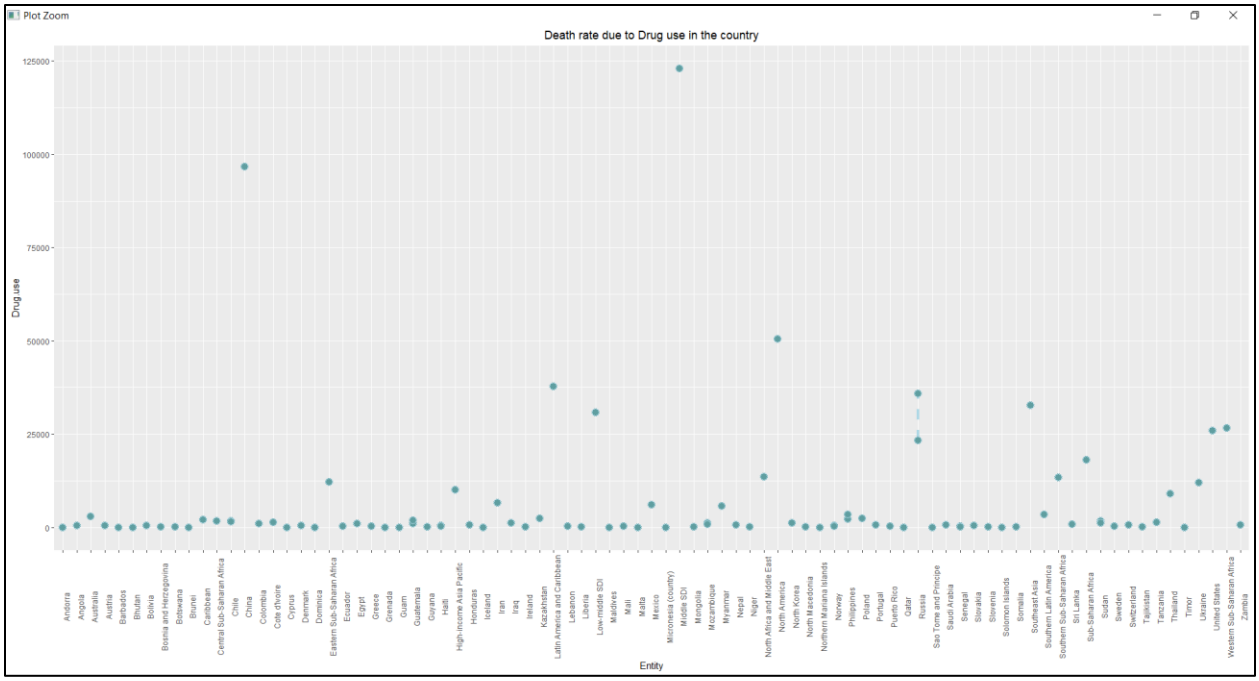
Graph 3: Death rate due to no access to hand wash facility in the country



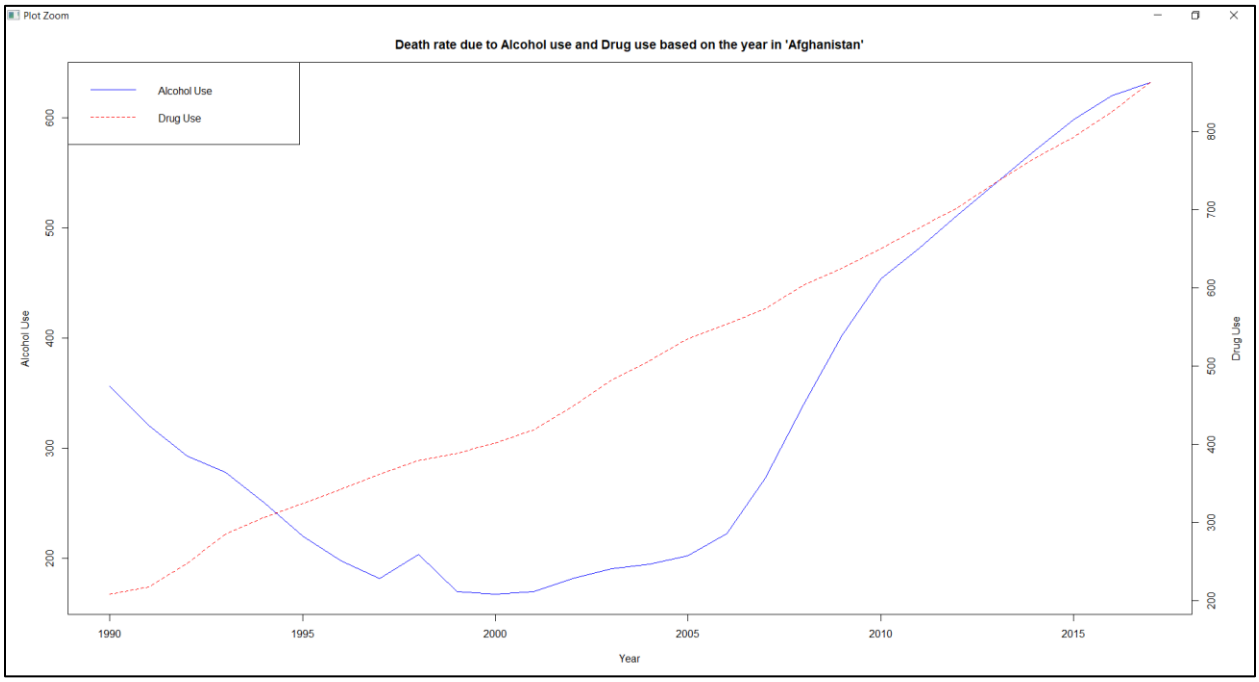
Graph 4: Risk factor due to Air pollution



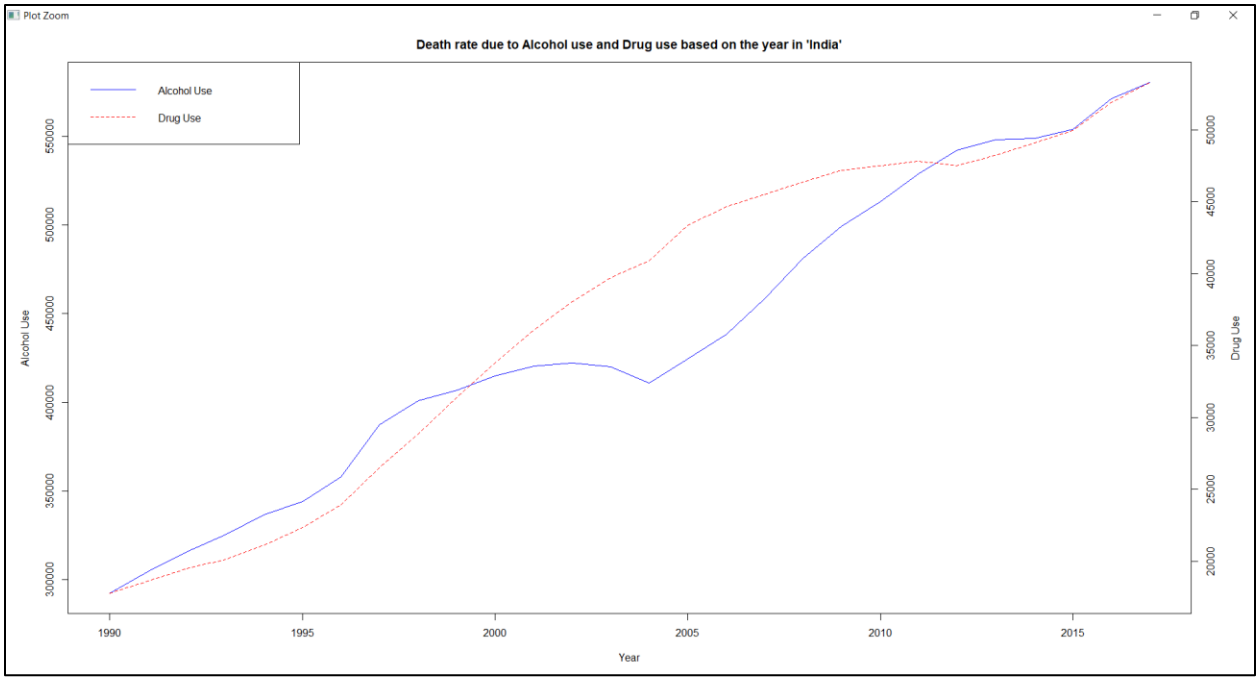
Graph 5: Death rate due to Drug use in the country



Graph 6: Death rate due to Alcohol use and Drug use based on the year in ‘Afghanistan’



Graph 7: Death rate due to Alcohol use and Drug use based on the year in ‘India’



Summary

The analysis of the number of death rate based on the risk factors helps in analyzing which country is affected by which of the risk factor. This analysis therefore will help us in computing and knowing the country affected most and its death rate which will further help in controlling it after knowing its reason.

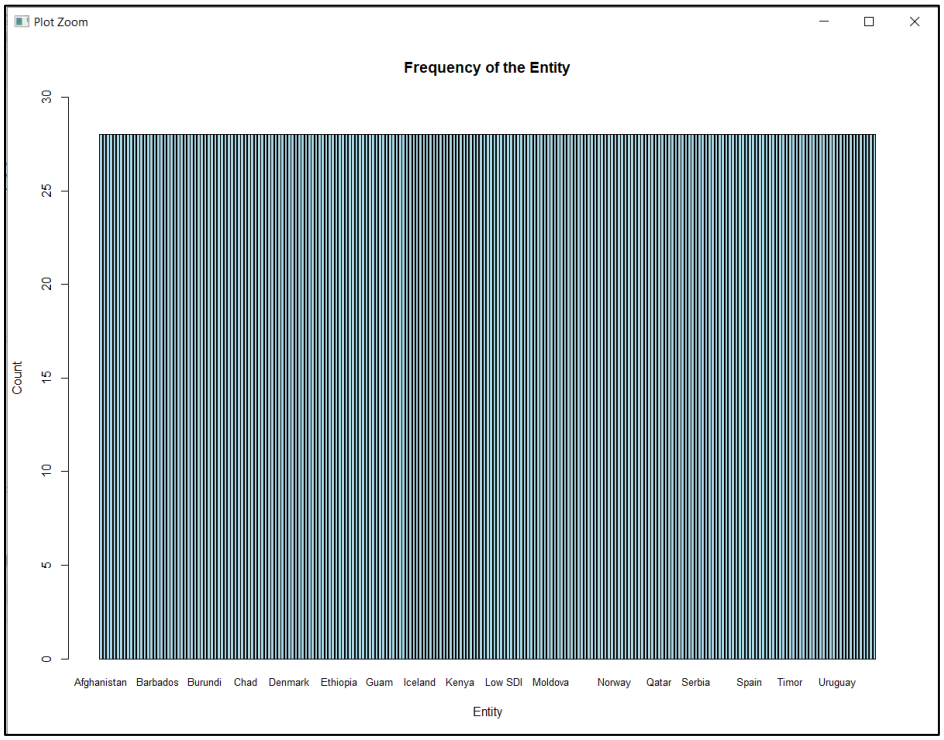
The descriptive analysis and the visualizations created helps in the analysis process as it helps in understanding the data and also in analyzing the countries affected by risk factor leading to the death rate in the year. Since the dataset has a large volume of data, it became easier to create subsets of the datasets and then plotting the risk factors for the same which helps to analyze the data even better. The questions answered throughout the analysis could be, which country is most affected by the risk factor of alcohol use, drug use and smoking. Similarly, which country is most affected and needs attention on by the risk factor of air pollution, outer air pollution, no access to hand wash facility etc. There are other attributes too like unsafe water source, unsafe sanitation, iron deficiency, etc. which will also help in the analysis process in understanding the risk factor a country is affected by.

Further analysis and explosion of this dataset could be done based on the majority of the risk factor one particular country is affected by i.e., by having multiple risk factors plotted for the entity attribute. Also, just like an entity was plotted for alcohol use and drug use in the year, in the same way other entities could be selected and visualizations for the various risk factors can be plotted.

Appendix

The graphs created before getting into the process of data analysis to understand the dataset better are as follows:

1. *Graph A: Frequency graph of the entity*



2. *Graph B: Death rate in a particular year*