FLIP ROBO

PROJECT REPORT

ON

CAUSE OF DEATHS

Submitted by

ABHISHEK MISHRA

ACKNOWLEDGMENT

As a reference I would like to mention the data source which is provided by the Flip Robo and the dataset which is provided in the form of csv file. I would like to thanks mentors and my SME who support and guided me to completion of this project and not forget to mention some seniors who helped me on GitHub and linkden for this project.

Date-09-02-2023

INTRODUCTION

- **Cause Of Deaths**: -

  The purpose of a cause of death dataset is to provide information on the various factors in the form of disease and other factors that contribute to mortality and the mortality patterns by year in different countries.

  A straightforward way to assess the health status of a population is to focus on mortality or concept like child mortality or life expectancy which are based on mortality estimates. A focus on mortality however does not take into account that the burden of diseases is not only that they kill people but that they cause suffering to people who live with them. Assessing health outcomes by both mortality and morbidity (the prevalent disease) provides a more encompassing view on health outcomes. This is the topic of this entry. The sum of the mortality and morbidity is referred to as the “burden of disease” and can be measured by a metric called “Disability Adjusted Life Years” (DALYs).

- Conceptual Background of the Domain Problem

  This type of dataset is important for understanding the patterns and trends in mortality and can be used by a public health and researchers’ epidemiologists and policymakers to inform strategies for reducing deaths and improving population health.

  The data can be used to identify leading cause of the death and disparities in the mortality among different population. It can be also used a monitor and overtime and assess the impact of interventions aims at reducing death from specific causes.

- Review of Literature

  As per the comprehensive summary we can clarify that the given problem statement is related to the Unsupervised Learning in which the data has only key features not target columns so we can say that this is the clustering related problem. But in this project, we had been asked just to Analise the data rather than machine model prediction. In the dataset there is lots of causes are given as the independent variable of the dataset which is increasing year by year with the population across the countries. So I have analyse the data for the all causes of death year by year and which countries having the largest numbers of deaths by several causes and which one having least numbers of deaths.

- **Motivation for the Problem Undertaken**

  The motivation behind this problem undertaken is to observe the cause of death and analysis the data by various process which may be helpful to prevent the death from the mentioned causes in the dataset. When we observed and analyses the dataset we had motivated enough to do and prevent the causes and able to reduce the death from all the causes.

- **Analytical Problem Framing**

- Mathematical/ Analytical Modelling of the Problem

  These are some important analytical things on which I have done the analysis and treat the things by using the correct methods.

  Non-Linearity:- Non linearity is a statistical term which is used to describe the relation between independent variable and dependent variable.

  By using the plots we find out the non-linearity between the features.

  Skewness:-When the mean value of the column is greater than its median value then the data is skewed. We need to treat the skewness by using some techniques. Here we use the "Log Transform" method to treat the skewness.

  Outliers:- when the data goes across -3 to +3 then those datas or values are called the outliers which affect the model so we need to remove these outliers by some techniques. We use the Z-Transforme technique to treat the outliers.

  PCA technique:- Done the PCA technique to achieve the 100% variance.

- Data Sources and their formats

  We had been provided the data source by Flip robo in the form of CSV file In the dataset there is 6120 rows and 34 column in the dataset. There are some columns which are categorical in nature while some columns are continuous in nature. According to the summary statics there is some skewness present in some columns and outliers are also present so need the treatment of the dataset. The columns in which their mean value is greater than the median value, must be present the skewness. Meanwhile by the plots we can find out the outliers in the columns. We are using the scatterplot to visualize the outliers of the columns.

- Data Pre-processing Done

To handle the missing values in the dataset I have used the mean value of the rest data present in the column. Detect the skewness in the column by using the function ‘df.skew()’. Finally, to remove the skewness I have used the log transformation method. Further visualise the data with boxplot to detect the outliers in the dataset and to treat these outliers from the dataset I am going to use the Z-transform method.

By treating the data with skewness and outliers, it is assumed that my data is now cleaned and ready for the further process.

- Data Inputs- Logic- Output Relationships

Describe the relationship between independent and dependent variables we are using the Bivariate analysis by plotting the different plots like histogram and scatterplots.

Finding the correlation between the two varriables.Plot the heatmap to find out that how much the variables are co related with each other. Remove the multicollinearity if any after this analysis.

- Hardware and Software Requirements and Tools Used

To build this project we need some important libraries which all are Listing down as-

Numpy- Treatment of skewness.

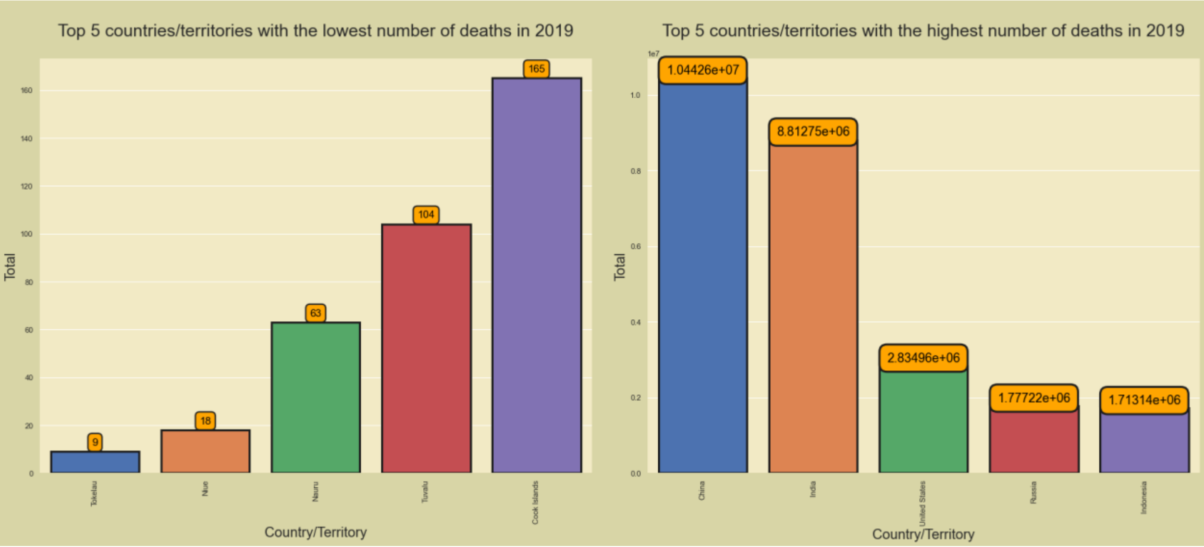
Pandas- To create the data in dataframe.

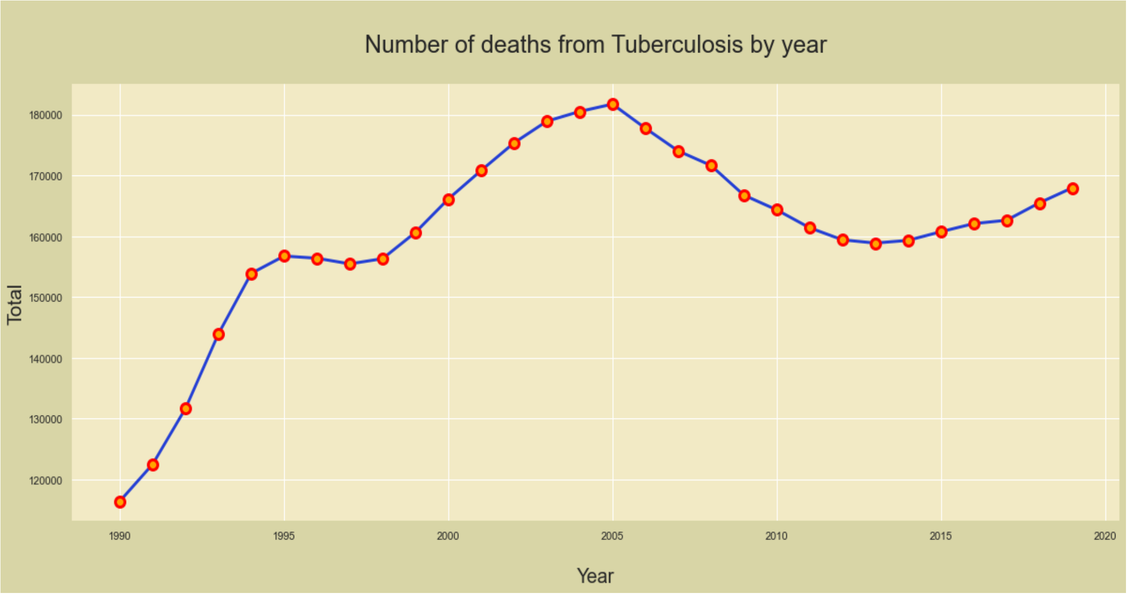
Seaborn- for data visualaization

Scipy- Mathmatical computation,

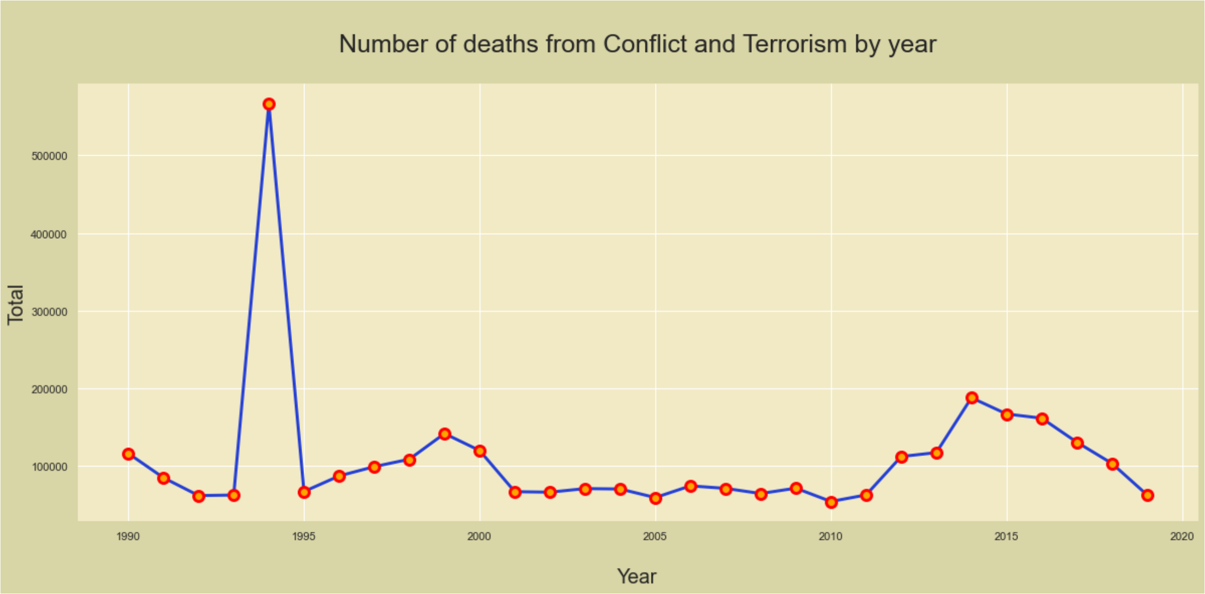
- **Data Analysis**:-

  I had analysed the data after completing the whole EDA steps in which I have explained through the graph that in which year there is more deaths and due to which cause there is highest number of deaths. All these things analysed by the data visualisation and some statistics methods.
  Here I am attaching the graphs by which we can observe clearly that which country have highest numbers of deaths and lowest number of deaths. In the graph I have also shown the graph trend in which the rate of molarity is shown year by year.

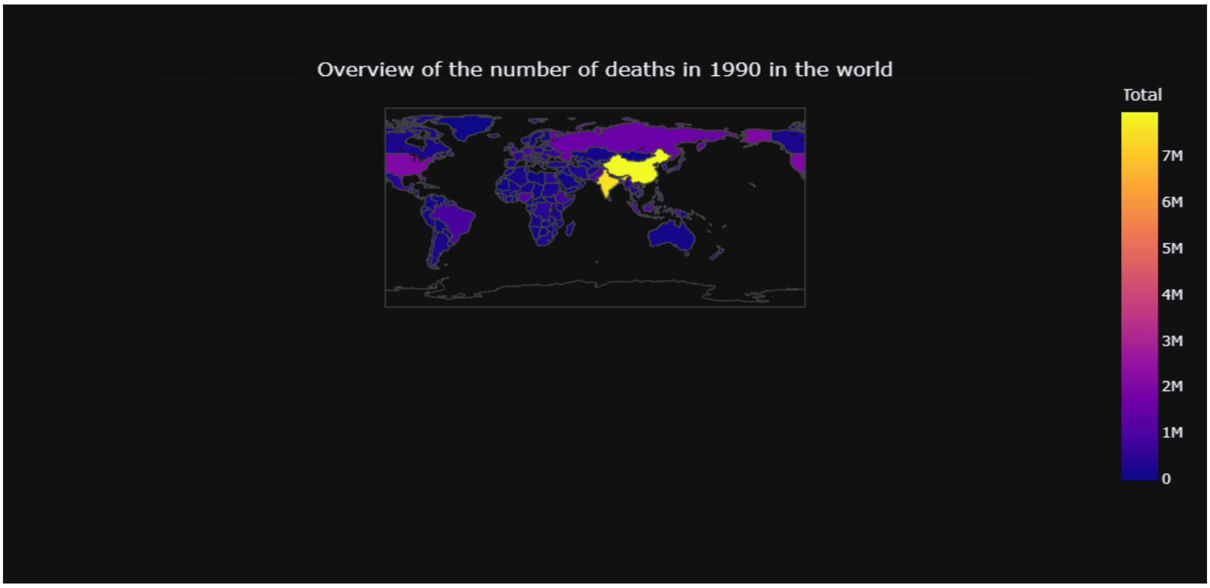




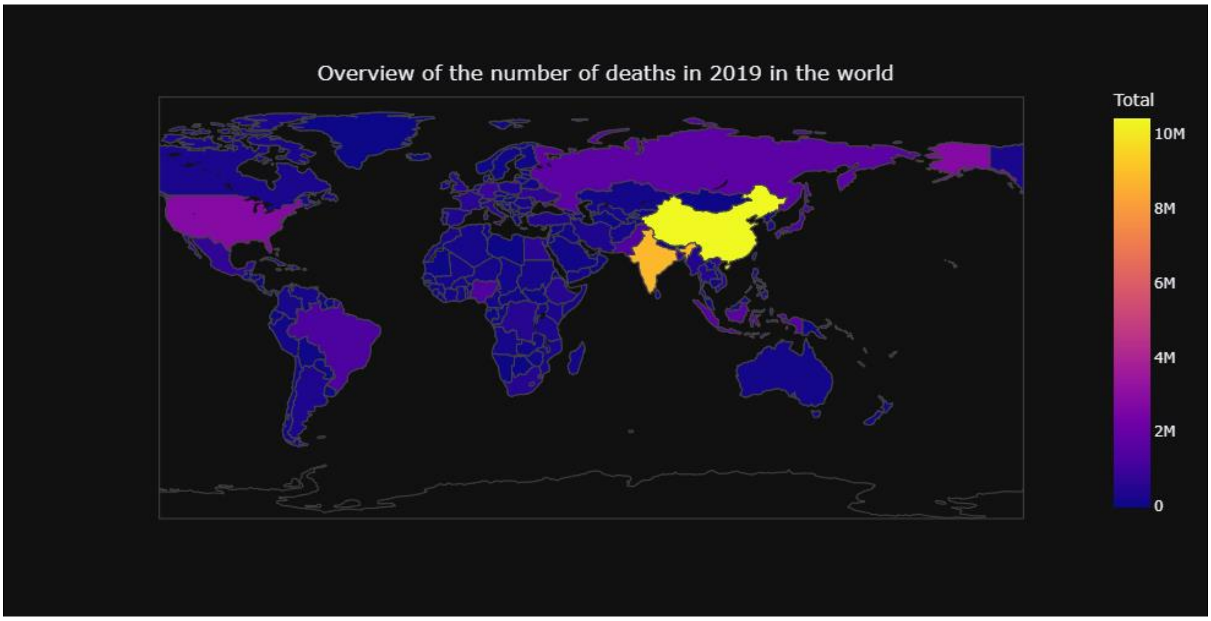
In the above graph we can see the number of deaths from tuberculosis by year.



- In the above graph we can observe the deaths due to conflict and terrorism by year.



- In the above graph we can observe the deaths in 1990 in the world



- The number of Deaths in the year 2019 in the world can be seen in the graph.

CONCLUSION

- Key Findings and Conclusions of the Study

From The above project I have learnt a lot that how to perform well as Data Analyst. The analysis of the Cause of Death dataset gives the reason and stats about the death due to various causes. So we can prevent the deaths and reduce the causes after analyses the data through data analysis method.

- Learning Outcomes of the Study in respect of Data Science

I had faced some issues during the problem statement analysis . I have used the log transform for skewness and outliers.one more important thing I want to discuss that I was stucked a little while analysing the dataset to choose what to analysis because I was not going to predict the model so was it necessary to split the data or not.