CAUSE OF DEATHS PROJECT

Submitted by:

Satya Jyothi. T

Flip Robo

Internship Batch – 34

ACKNOWLEDGMENT

I would like to express my deep and sincere gratitude towards Flip Robo Technologies for providing me the internship opportunity and a great chance for learning and professional development.

A big special thanks to my SME Ms. Khushboo Garg for providing necessary help in solving the problems and for providing clarifications on time throughout.

My sincere thanks to “Data Trained” who are the reason behind my internship at Flip Robo.

Last but not least my all-well-wishers including parents/ spouse/ friends who have been my backbone in every step of my life.

INTRODUCTION

Business Problem Framing

- There is a historical data of different cause of deaths for all ages around the World where Disability Adjusted Life Years (DALYs) are measuring lost health & a standardized metric that allow for direct comparisons of disease burdens of different diseases across countries, between different populations, and over time.

- One DALY is the equivalent of losing one year in good health because of either premature death or disease or disability.

- One DALY represents one lost year of healthy life

- Objective is to perform the data analysis which influence Cause of Death (Both Mortality & Morbidity)

- Mortality + Morbidity → Measured by 'Disability Adjusted Life Years

Conceptual Background of the Domain Problem

Sum of mortality and morbidity → Measured by metric called 'Disability Adjusted Life Years' (DALYs).We need to assess health outcomes by both mortality & morbidity (the prevalent diseases) which provides a more encompassing view on health outcomes.

DALYs Measure lost health and are a standardized metric that allow for direct comparisons of disease burdens of different diseases across countries, between different populations, and over time.

Need data analysis which influence Cause of Death (Both Mortality & Morbidity)

Motivation for the Problem Undertaken

This project was given by Flip Robo Technologies as a part of the internship program. This opportunity gives the exposure to real world data and using my skillset in solving a real time problem has been the primary motivation.

In this presentation, the focus will be on

- How to analyse the dataset of Cause of Deaths
- Overall data analysis on the problem

Analytical Problem Framing

Mathematical/ Analytical Modeling of the Problem

The main objective of doing this project is to apply the analytical skills to get findings and conclusions in detailed data analysis of Cause of Deaths dataset

There is no target variable in the dataset hence this study is limited to only data analysis and no machine learning model is developed

Data Sources and their formats

Data set provided by Flip Robo was in the format of CSV (Comma Separated Values). There is one data set that is given.

cause_of_deaths dataset.csv file will be used for detailed data analysis. The dimension of data is 6120 rows and 34 columns.

{int64: ['Year', 'Meningitis', 'Alzheimer's Disease and Other Dementias', 'Parkinson's Disease', 'Nutritional Deficiencies', 'Malaria', 'Drowning', 'Interpersonal Violence', 'Maternal Disorders', 'HIV/AIDS', 'Drug Use Disorders', 'Tuberculosis', 'Cardiovascular Diseases', 'Lower Respiratory Infections', 'Neonatal Disorders', 'Alcohol Use Disorders', 'Self-harm', 'Exposure to Forces of Nature', 'Diarrheal Diseases', 'Environmental Heat and Cold Exposure', 'Neoplasms', 'Conflict and Terrorism', 'Diabetes Mellitus', 'Chronic Kidney Disease', 'Poisonings', 'Protein-Energy Malnutrition', 'Road Injuries', 'Chronic Respiratory Diseases', 'Cirrhosis and Other Chronic Liver Diseases', 'Digestive Diseases', 'Fire, Heat, and Hot Substances', 'Acute Hepatitis'], object: ['Country/Territory', 'Code']}

Data Pre-processing Done

Data integrity is checked by checking for duplicate values, white spaces and missing values. There are no duplicate values, white spaces found in the dataset.

There are no missing values present in the dataset.

There are 2 data types in the data set. a) Object b) int64

Statistical summary is checked for both Object and Numerical data to draw the insights of descriptive statistics (count, mean, min, max, std deviation and IQR values) in a simple manner.

Data Inputs- Logic- Output Relationships

There is no output variable in the dataset.

However, this dataset helps to draw multiple insights like below

- ✓ Top countries with more deaths due to different diseases
- ✓ Top diseases which are causing more deaths in the world
- ✓ Diseases with least deaths
- ✓ Countries with least deaths
- ✓ Dataset helps the countries to take necessary steps to reduce the no. of deaths due to different diseases

State the set of assumptions (if any) related to the problem under consideration

Didn't drop any column from the dataset as all the features are important for the data analysis.

Column "Code" can be dropped however it is not used in the data analysis instead used column "Country/Territory" for the data analysis.

Hardware and Software Requirements and Tools Used

Software Used:

Programming language: Python

Distribution: Anaconda Navigator

Browser based language shell: Jupyter Notebook

Libraries/Packages Used:

Pandas, NumPy, matplotlib, seaborn, scikit-learn

Hardware Used -

Processor — Intel i7 processor 8th Generation

RAM — 32 GB

GPU — 4GB NVIDIA Graphics card

Model/s Development and Evaluation

Identification of possible problem-solving approaches (methods)

- There are many features in dataset. Used various statistical and analytical techniques to solve the problem.
- Performed univariate, bivariate and multivariate graphical analysis to draw the key insights from the dataset
- Used Pearson's coefficients of correlation to find out the relation among multiple features
- Multi-collinearity between features is checked using VIF (Variance Inflation Factor). Typically, the columns with VIF>10 need to be dropped which indicates strong multicollinearity of the features however there are many features with VIF>10. VIF can also be addressed with PCA (Principal Component Analysis) without dropping the features. In this project, used PCA to reduce the no. of features by capturing 95% of variance of the data
- Outliers are checked for the X numerical data. In z-score method, removed outliers if z-score > 3 and < -3. In IQR method, removed outliers beyond Inter Quartile Range (Q3-Q1). Calculated %data loss with both methods. z-score method resulted in data loss with less than 10% (8.1%) which is acceptable whereas IQR method resulted in the data loss of 50.5% hence finally used z-score method for outliers removal.
- Skewness is checked for X numerical data. Applied multiple transformation techniques (Power Transformer, Quantile Transformer) to check the skewness reduction. Considered -0.5 to +0.5 range as fairly symmetrical. Quantile Transformer has reduced the skewness within this range, hence considered the data for further processing from Quantile Transformer.
- Scaled the X data using Standard Scaler.

Testing of Identified Approaches (Algorithms)

- No ML model is developed as there is no target variable in the dataset.

Run and Evaluate selected models

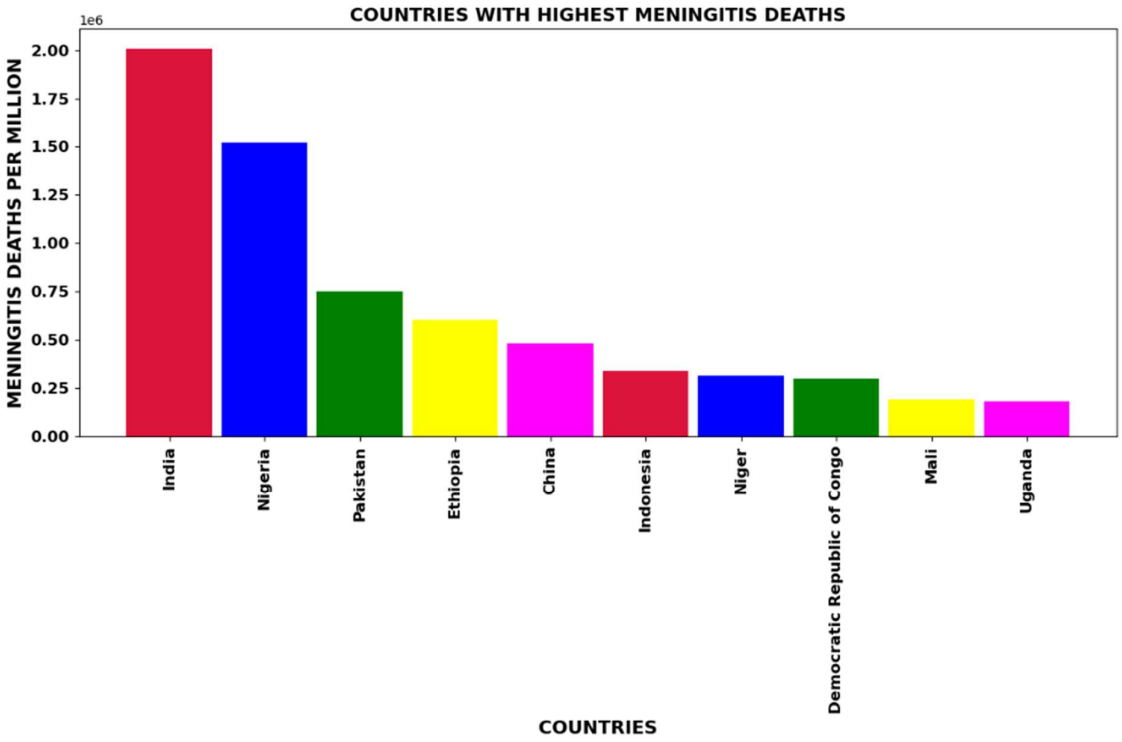
- No ML model is developed as there is no target variable in the dataset.

Key Metrics for success in solving problem under consideration

- No ML model is developed as there is no target variable in the dataset.

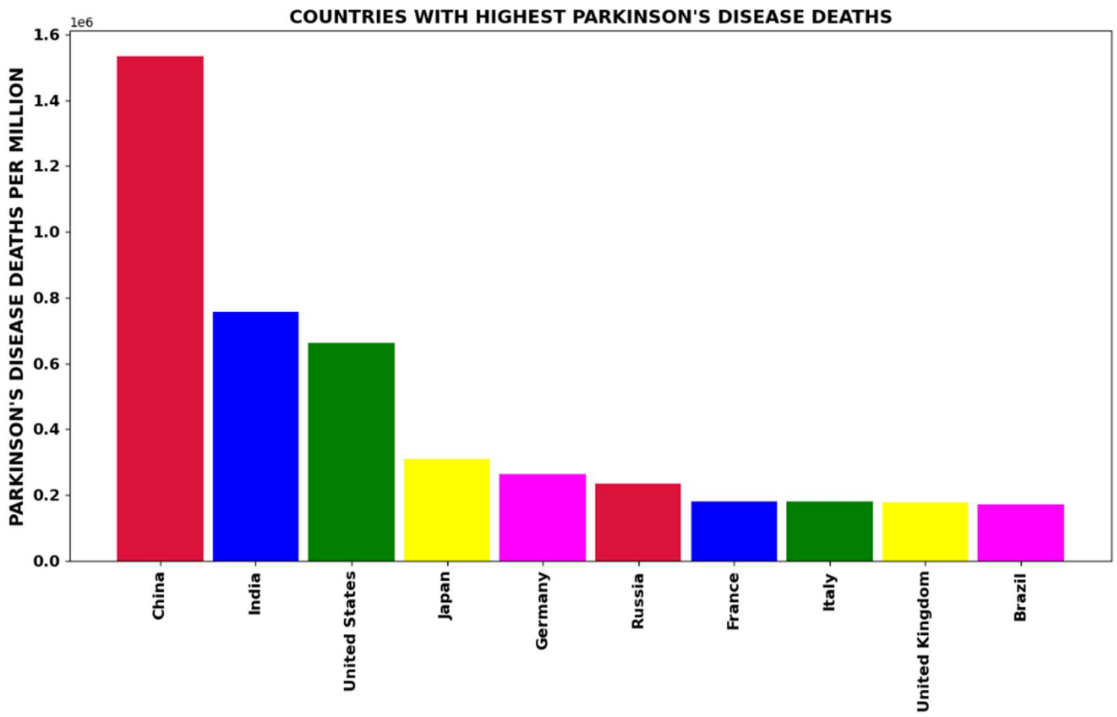
Visualizations

TOP 10 COUNTRIES WITH MORTALITY FOR EACH DISEASE



Observations for MENINGITIS:

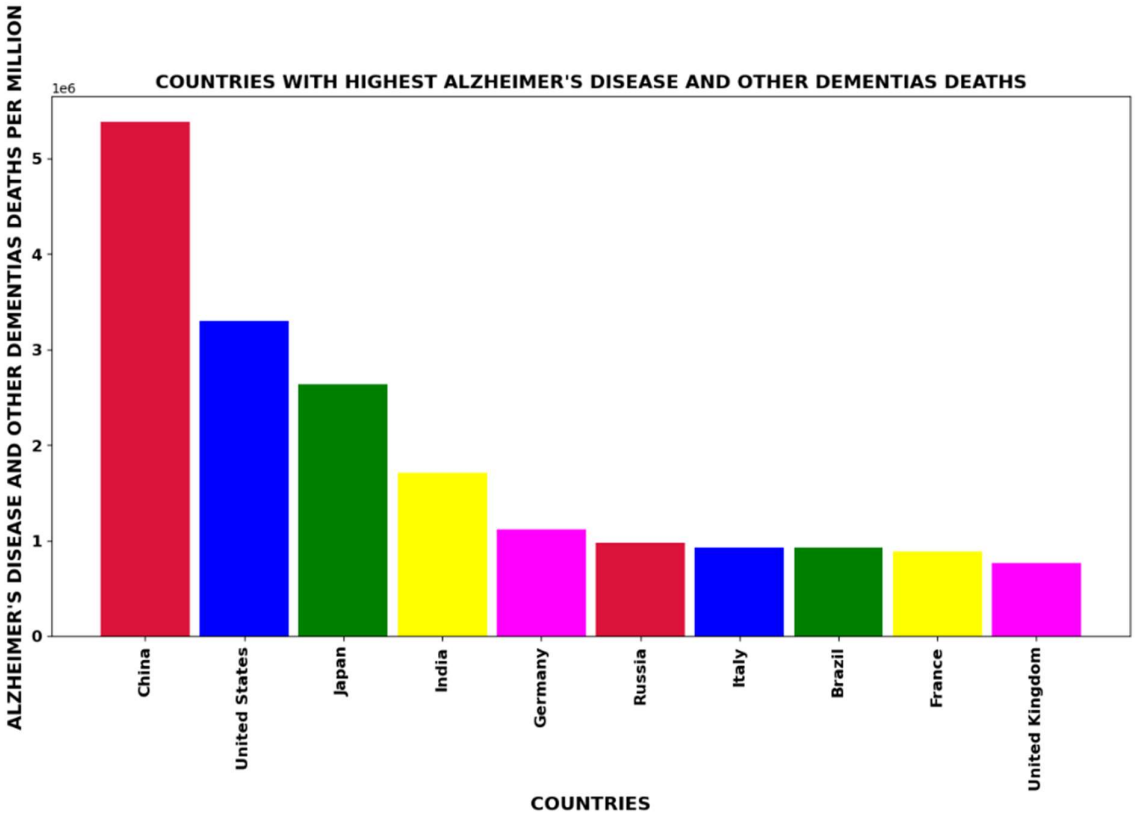
- The highest number of deaths due to MENINGITIS occurring in country INDIA
- The least number of deaths due to MENINGITIS occurring in country PALAU



Observations for PARKINSON'S DISEASE:

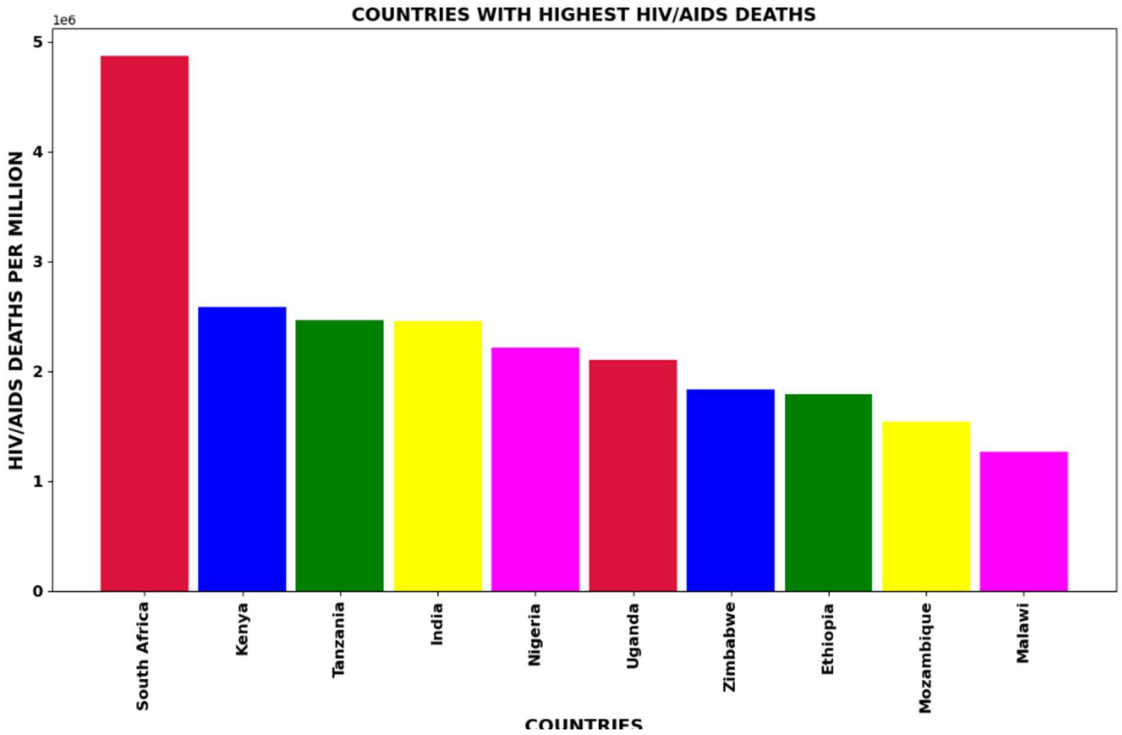
- The highest number of deaths due to PARKINSON'S DISEASE occurring in country CHINA

- The least number of deaths due to PARKINSON'S DISEASE occurring in country NAURU



Observations for ALZHEIMER'S DISEASE AND OTHER DEMENTIAS:

- The highest number of deaths due to ALZHEIMER'S DISEASE AND OTHER DEMENTIAS occurring in country CHINA
- The least number of deaths due to ALZHEIMER'S DISEASE AND OTHER DEMENTIAS occurring in country TOKELAU

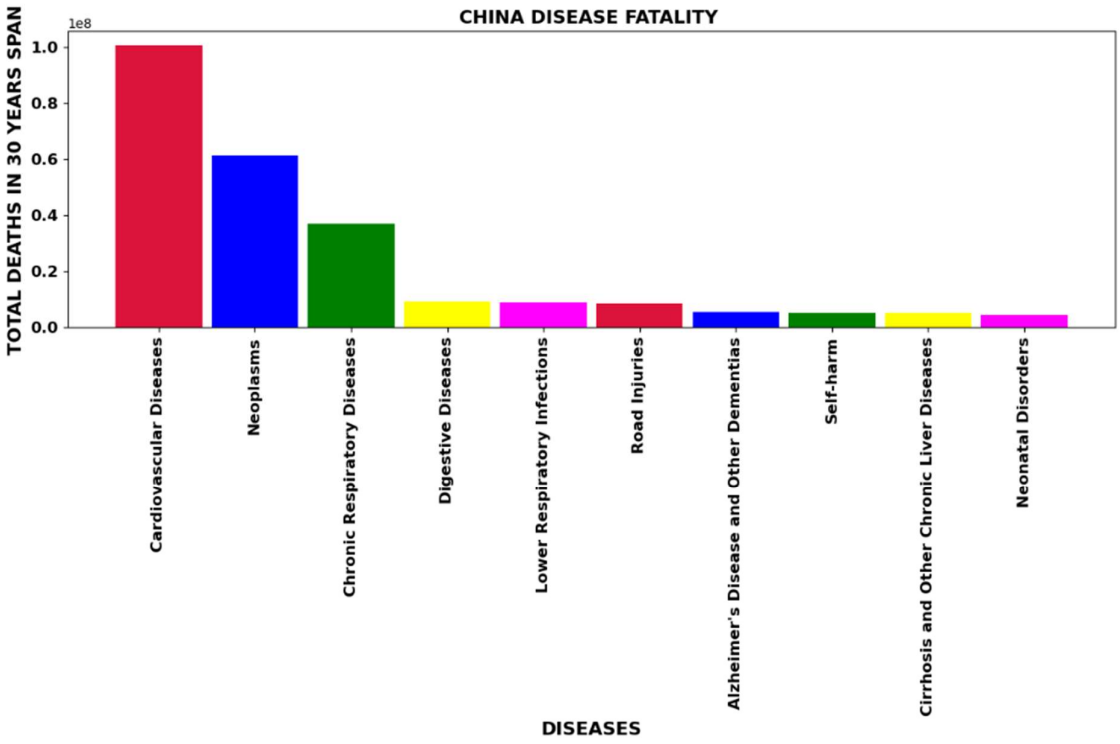


Observations for HIV/AIDS:

- The highest number of deaths due to HIV/AIDS occurring in country SOUTH AFRICA

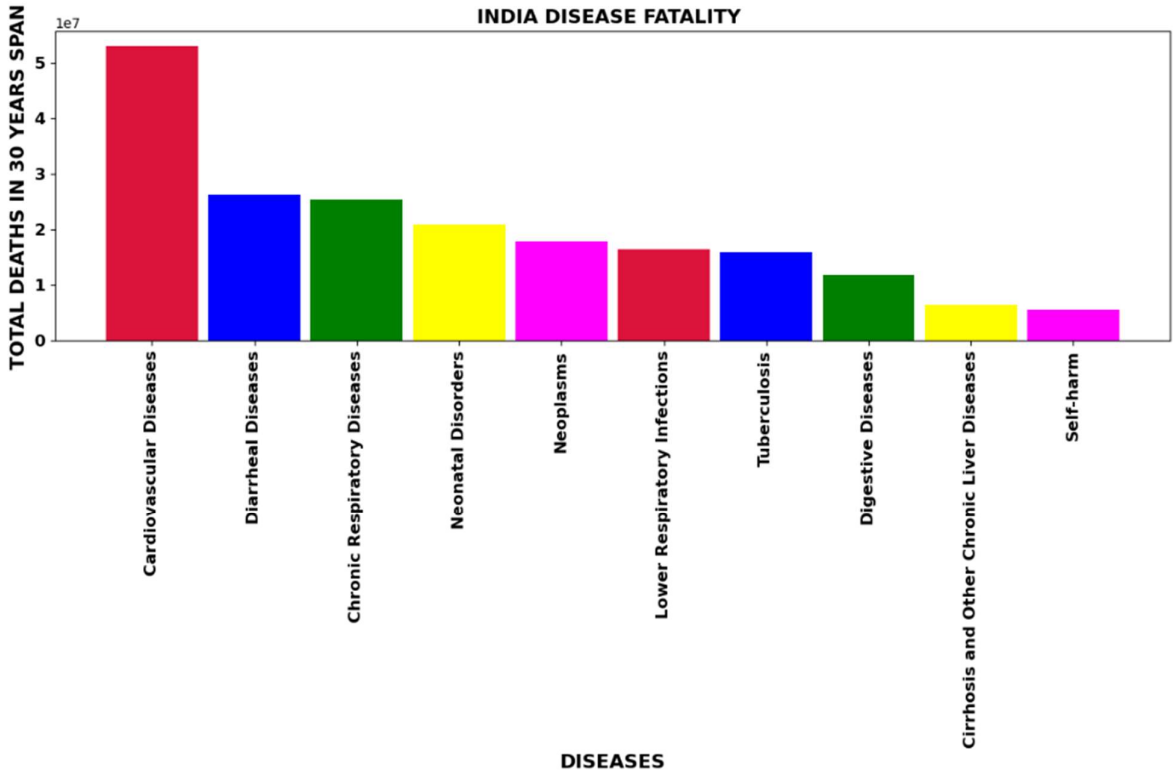
- The least number of deaths due to HIV/AIDS occurring in country NIUE

COUNTRY WISE TOP 10 DISEASES WITH MORTALITY



Observations for CHINA:

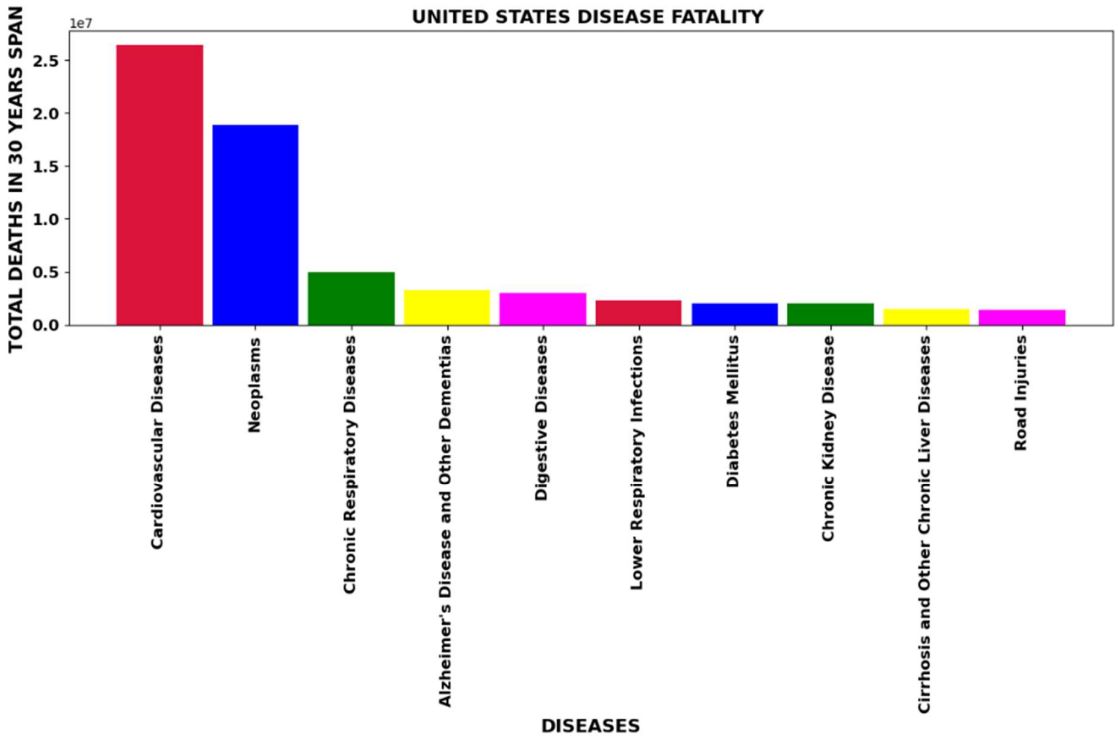
- In CHINA, most deaths are due to CARDIOVASCULAR DISEASES
- In CHINA, least deaths are due to CONFLICT AND TERRORISM



Observations for INDIA:

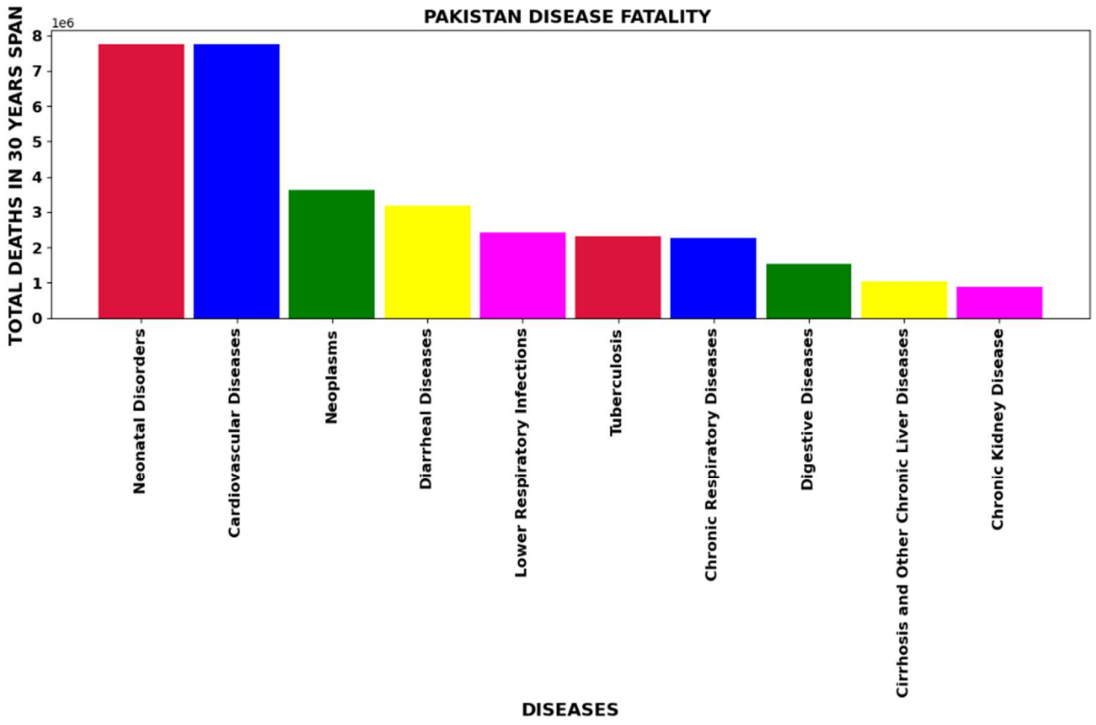
- In INDIA, most deaths are due to CARDIOVASCULAR DISEASES

- In INDIA, least deaths are due to CONFLICT AND TERRORISM



Observations for UNITED STATES:

- In UNITED STATES, most deaths are due to CARDIOVASCULAR DISEASES
- In UNITED STATES, least deaths are due to MALARIA

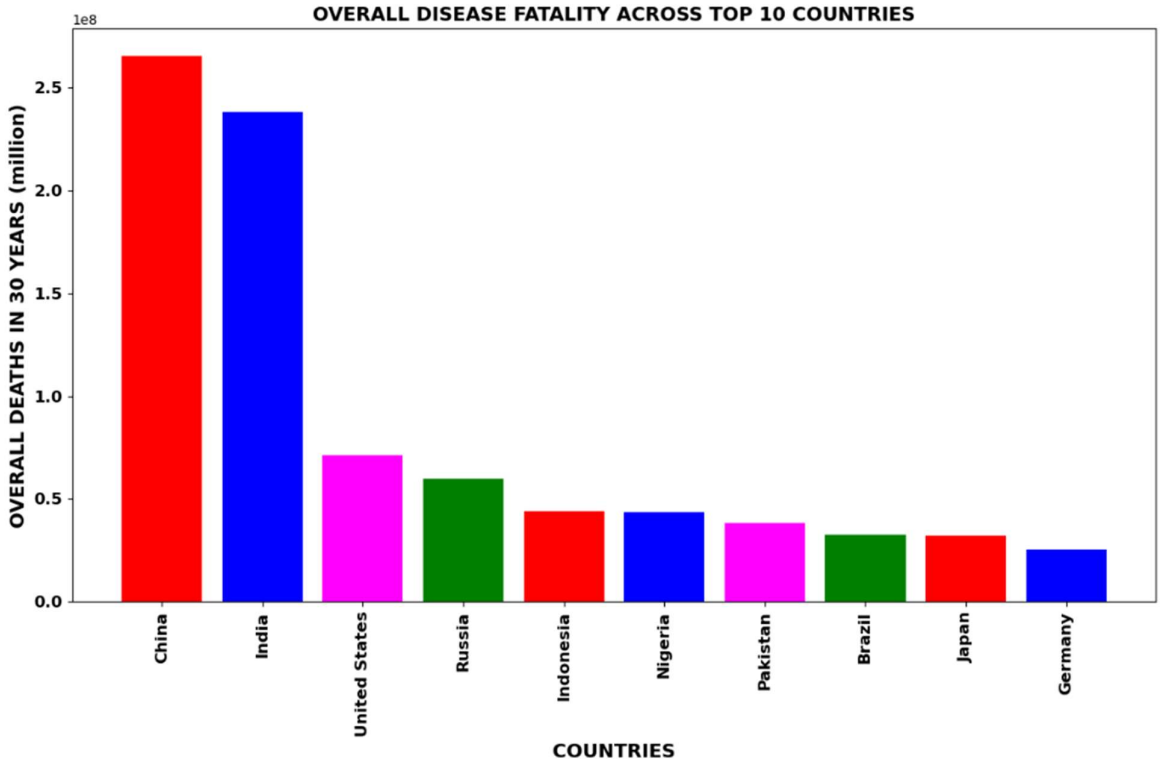


Observations for PAKISTAN:

- In PAKISTAN, most deaths are due to NEONATAL DISORDERS

- In PAKISTAN, least deaths are due to ENVIRONMENTAL HEAT AND COLD EXPOSURE

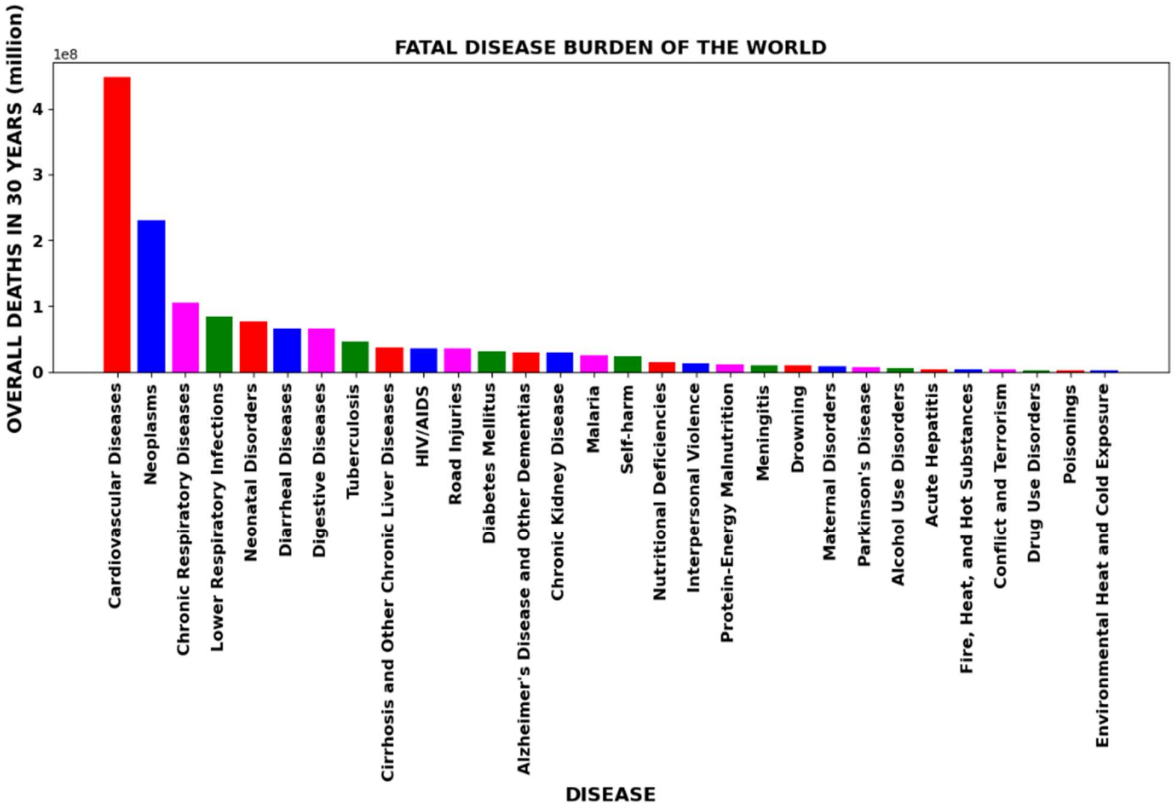
COUNTRIES WITH HIGHEST DEATHS DUE TO DISEASES



Observations:

- ✓ CHINA and INDIA are top countries having most deaths compared to other countries in 30 years span
- ✓ This could be due to large population in these two countries
- ✓ UNITED STATES is the country having more no. of deaths after CHINA and INDIA

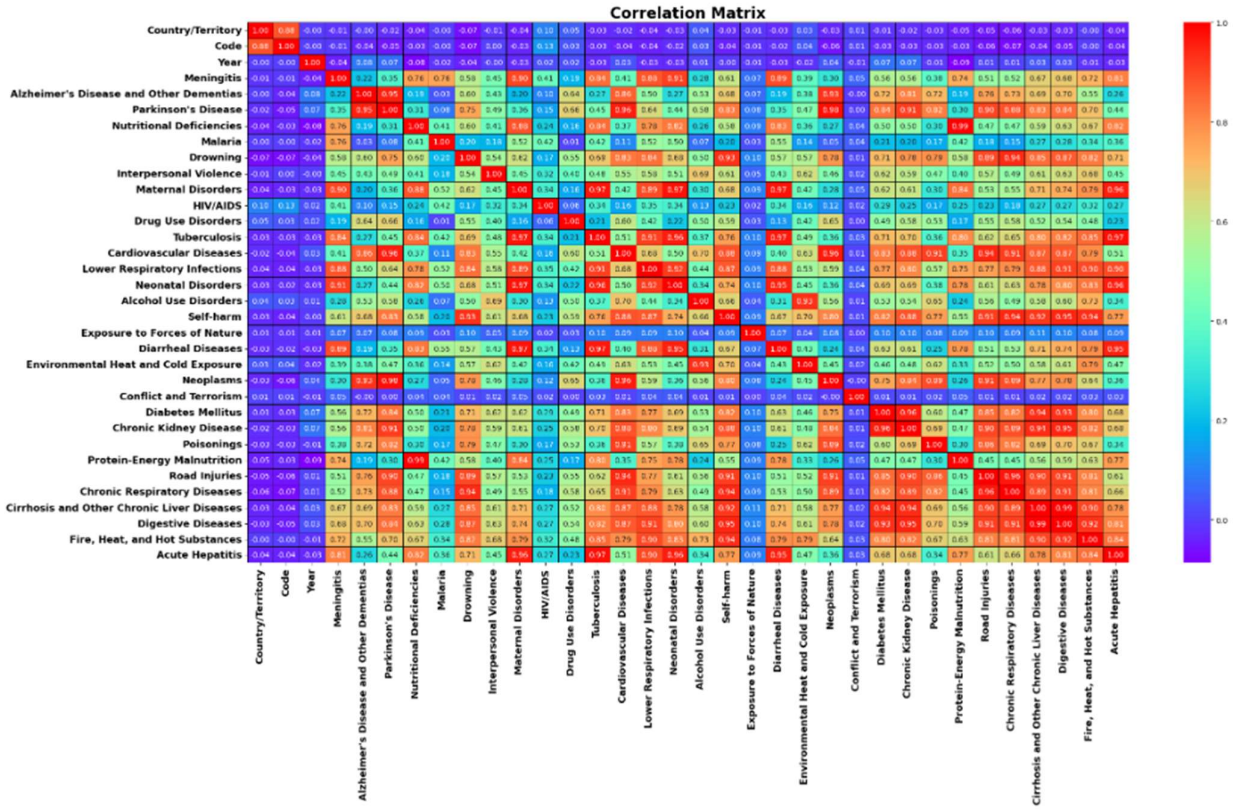
TOP KILLER DISEASES IN THE WORLD



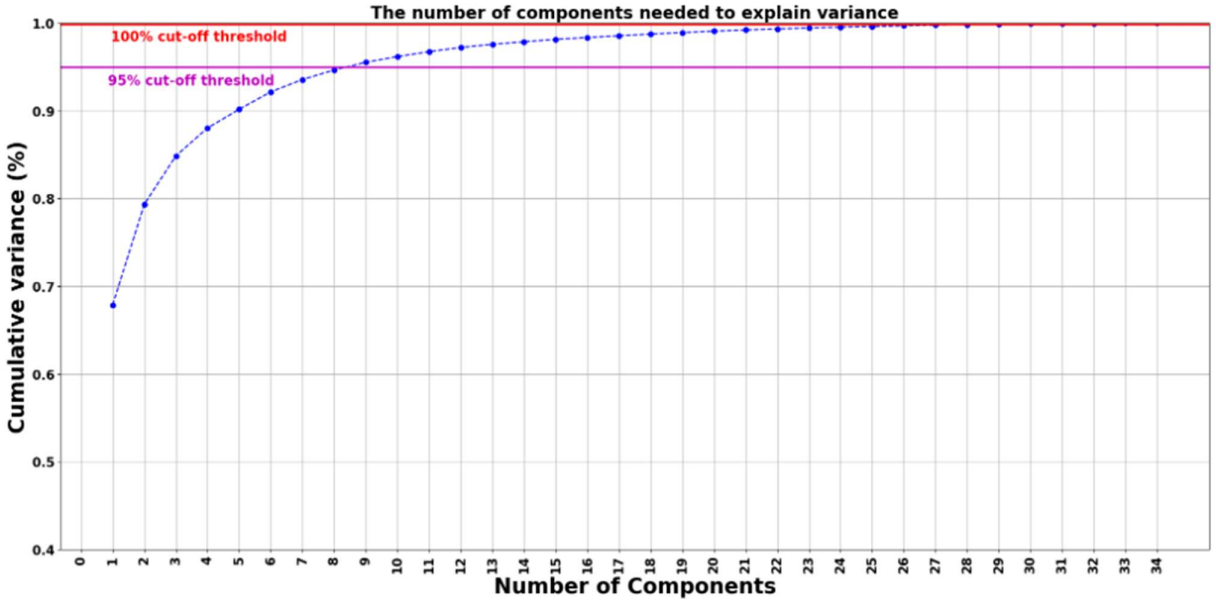
Observations:

- ✓ Cardiovascular Diseases is the most killing disease in the world
- ✓ The lowest no. of deaths occurred due to Environmental Heat and Cold Exposure in the world
- ✓ CHINA , INDIA AND USA face the largest brunt of deaths due to diseases in the world Cardiovascular diseases , Neoplasms and Lower Respiratory Infections are the top 3 killer diseases in the world

COEFFICIENTS OF CORRELATION



PCA (Principal Component Analysis)



PCA:

- ✓ As the dataset is having many X columns and as there is multi-collinearity, used PCA technique to reduce the number of dimensions in a dataset while preserving the most important information in it
- ✓ Initially, calculated the number of components needed to explain the variance
- ✓ Based on that, chosen the no. of columns required be used in PCA
- ✓ As per the graph, we can see that 8 principal components attribute for 95% of variation in the data. Hence, I picked 8 components
- ✓ I will use 8 features as no. of components in PCA to reduce the dimensions

Interpretation of the Results

- As there is no output variable, no ML model is developed however drawn key insights from the dataset which influence the cause of deaths in different countries of the world

CONCLUSION

Key Findings and Conclusions of the Study

- Cardiovascular Diseases is the most killing disease in the world
- The lowest no. of deaths occurred due to Environmental Heat and Cold Exposure in the world
- CHINA, INDIA AND USA face the largest brunt of deaths due to diseases in the world. The most %of deaths in CHINA and INDIA could be due to large population
- Cardiovascular diseases, Neoplasms and Lower Respiratory Infections are the top 3 killer diseases in the world

Learning Outcomes of the Study in respect of Data Science

Data Analysis – This study helps to apply analytical skills to get findings and conclusions with the detailed data analysis of dataset which helped to identify the top diseases which caused more no. of deaths in the world in different countries. This study will help all the countries to take necessary steps to avoid or reduce the no. of deaths due to various diseases and special focus on top diseases causing more deaths.

Limitations of this work and Scope for Future Work

Target variable can be added to data which will helps to perform supervised machine learning model

Performing unsupervised learning can be done using clustering, anomaly detection, neural networks

More features & data can be available to help in best recommendation