FLIP ROBO

Project Report on **Cause of Death**

Submitted By

Ramesh Pyru

ACKNOWLEDGEMENTS

I would like to thank FlipRobo team for their direction, assistance, and guidance. In particular Ms. Khushboo Garg (SME Flip Robo), her recommendations and suggestions have been invaluable for the project.

I also wish to thank “Data trained” which provided the opportunity for internship at FlipRobo.

Special thanks to my family & friends who helped me in many ways.

References use in this project:

1. SCIKIT Learn Library Documentation
2. Blogs from towards data science, Analytics Vidya, Medium
3. Andrew Ng Notes on Machine Learning (GitHub)

TABLE OF CONTENTS

Introduction

1.1 Background

We have 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

Expectation→

- We need data analysis which influence Cause of Death (Both Mortality & Morbidity)
- Mortality + Morbidity →Measured by ‘Disability Adjusted Life Years‘

1.2 Problem Definition

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.

- ✓ We need data analysis which influence Cause of Death (Both Mortality & Morbidity)

1.3 Stages of Project

- Data Collection
- Data Cleaning--->Missing Values imputation--->Handling Outliers
  - 2.a. Univariate Analysis(Graphical representation-->Histograms (Bar Charts) , Box Plots, Bar &Group charts, Stem and Leaf Plots)
  - 2.b.Bivariate Analysis(Graphical representation-->Pai Plots, Bar Plot,CrossTab,Scatter Plots)
  - 2.c. Multivariate Analysis(Graphical representation-->HeatMap,LinePlot,PairPlot)
- Choosing the Right Statistical Methods
- Visualizing and Analyzing Results
- Building the Machine Learning Model
- Test the model with Test Data available

The below figure depicts the different steps for EDA analysis

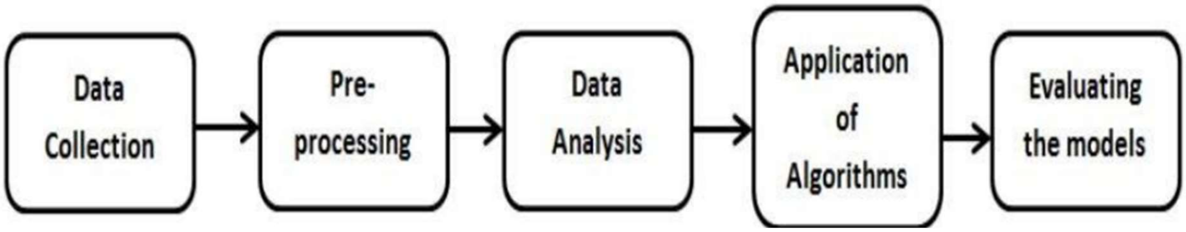


Fig :- High Level Stages

EDA Approach & Outcome

Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.

Steps involved

1) Importing a dataset
2) Understanding the big picture
3) Data Preparation
4) Understanding of variables
5) Study of the relationships between variables
6) Brainstorming

Data Preprocessing Done

Data pre-processing is an integral step in Machine Learning as the quality of data and the useful information that can be derived from it

- Loading the dataset as a dataframe
- Checked the number of rows and columns present in our training dataset
- Checked for missing data using Data Prep Library
- Checked the unique values information in each column to get a gist for categorical data

Assumptions

The assumption part for me was relying strictly on the data provided to me and taking into consideration that the datasets were obtained from real people surveyed for their preferences

Hardware and Software Requirements and Tools Used

Hardware Used:

1. RAM: 16 GB
2. CPU: 11th Gen Intel(R) Core(TM) i5-1135G7 @ 2.40GHz 2.42 GHz
3. GPU: Intel® Iris® Xe Graphics

Software Used:

1. Programming language: Python
2. Distribution: Anaconda Navigator
3. Browser based language shell: Jupyter Notebook

Libraries/Packages Used:

- ❖ Pandas, NumPy, matplotlib, seaborn, scikit-learn

Model/s Development and Evaluation

N/A

Testing of Identified Approaches (Algorithms)

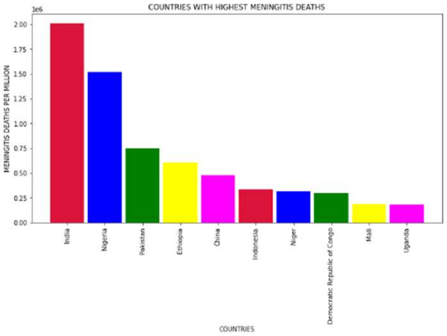
N/A

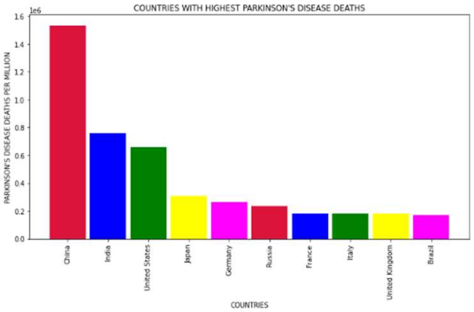
Key Metrics for success in solving problem under consideration

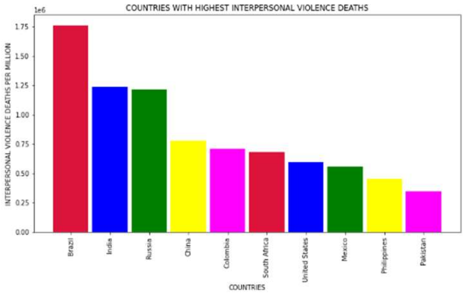
N/A

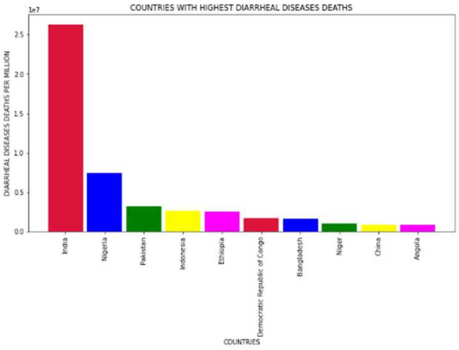
Visualizations(Analysis)

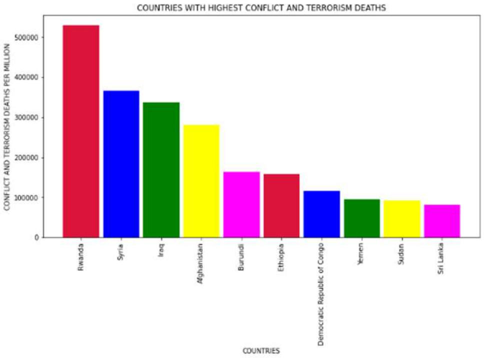
Cause of Death by Country/Territory

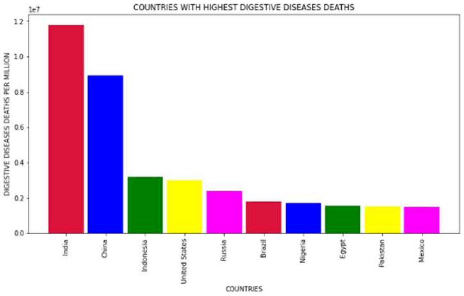




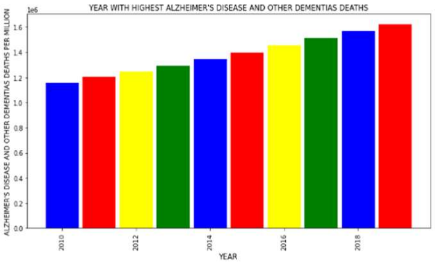


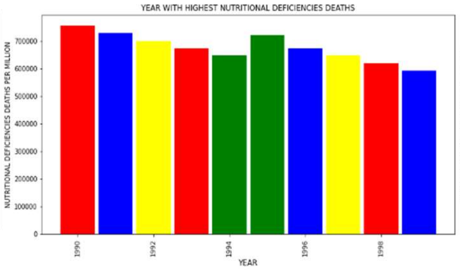


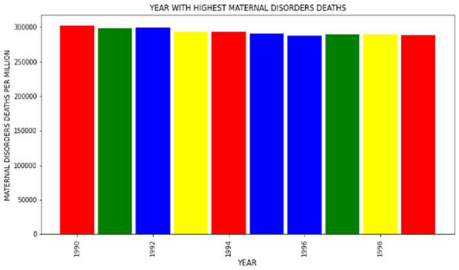


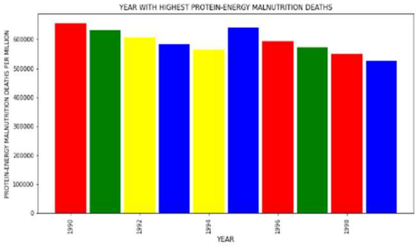


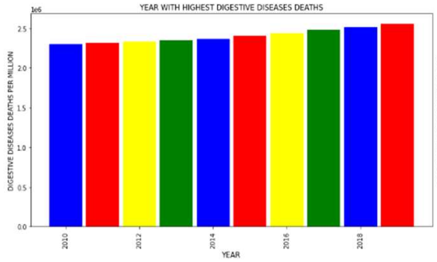
Cause of Death by Year

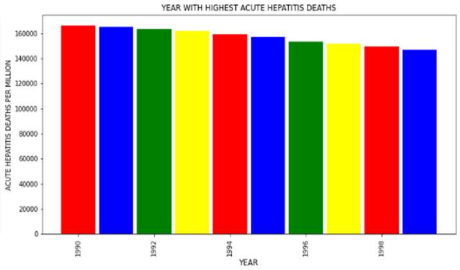


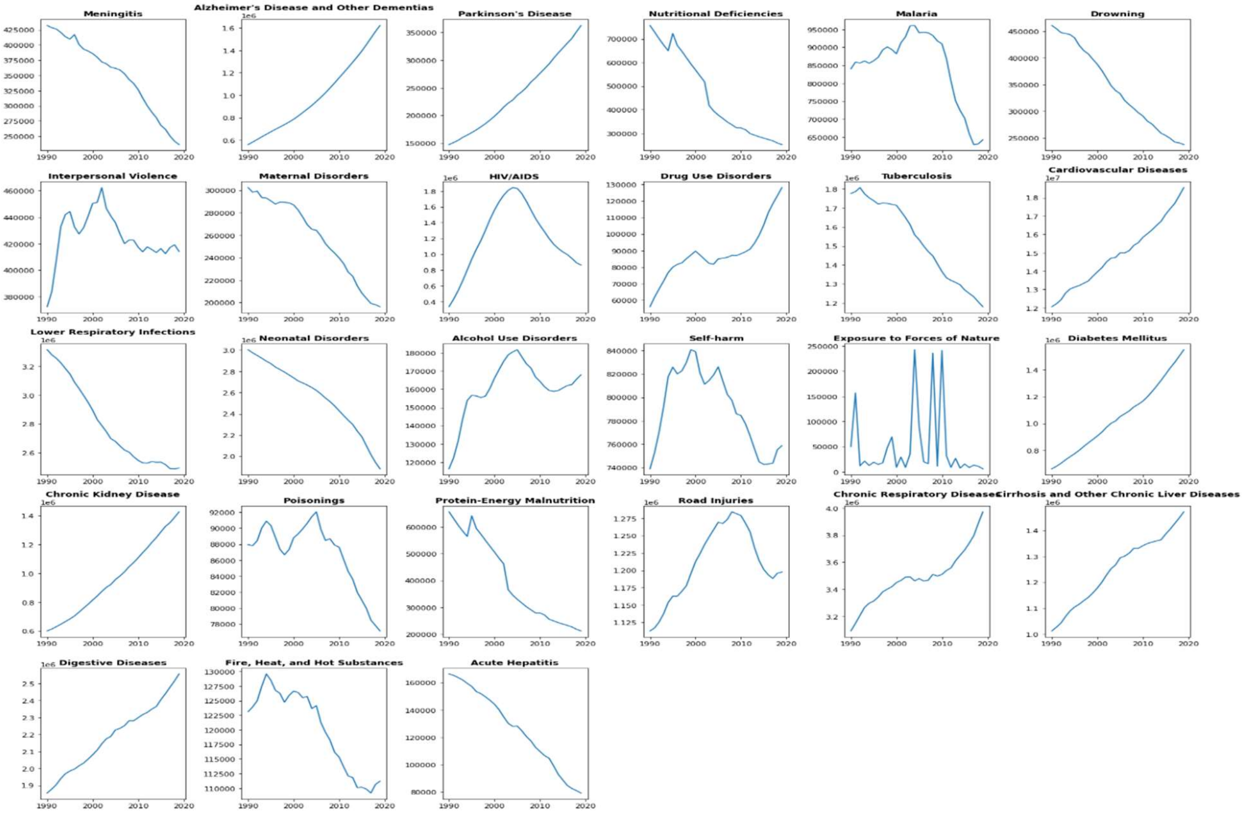










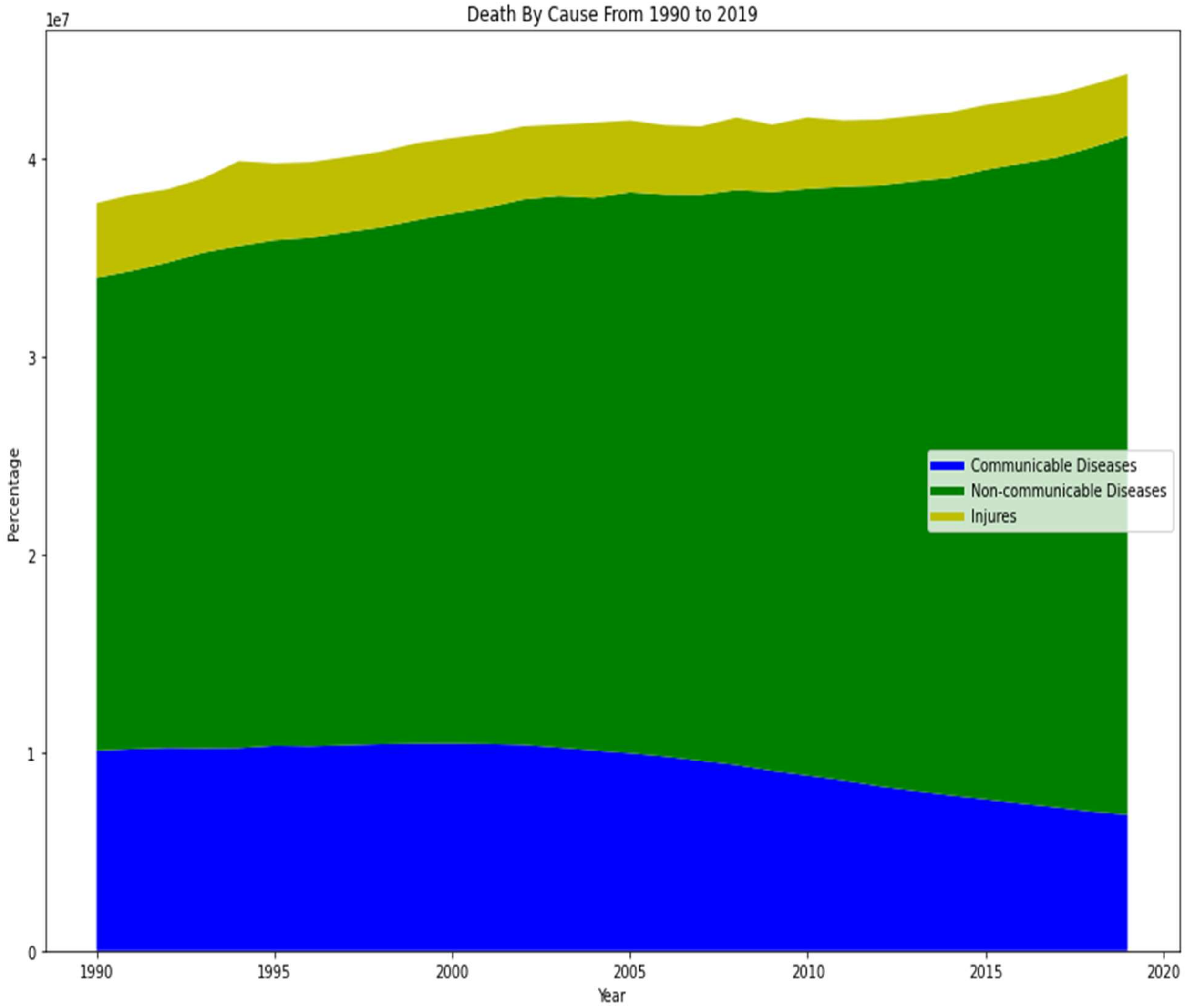


Following are the death causes exhibiting an increasing or no trend over time:

- Dementia, Kidney, Cancer, Parkinson, and Diabetes diseases have been increasing since the beginning of our data span (1990)
- Drug and alcohol use disorders are correlated with Liver disease trend
- HIV/AIDS related deaths rate started to decline since 2005
- Natural disasters show no trend as expected

Divide the causes of death into 3 main categories:

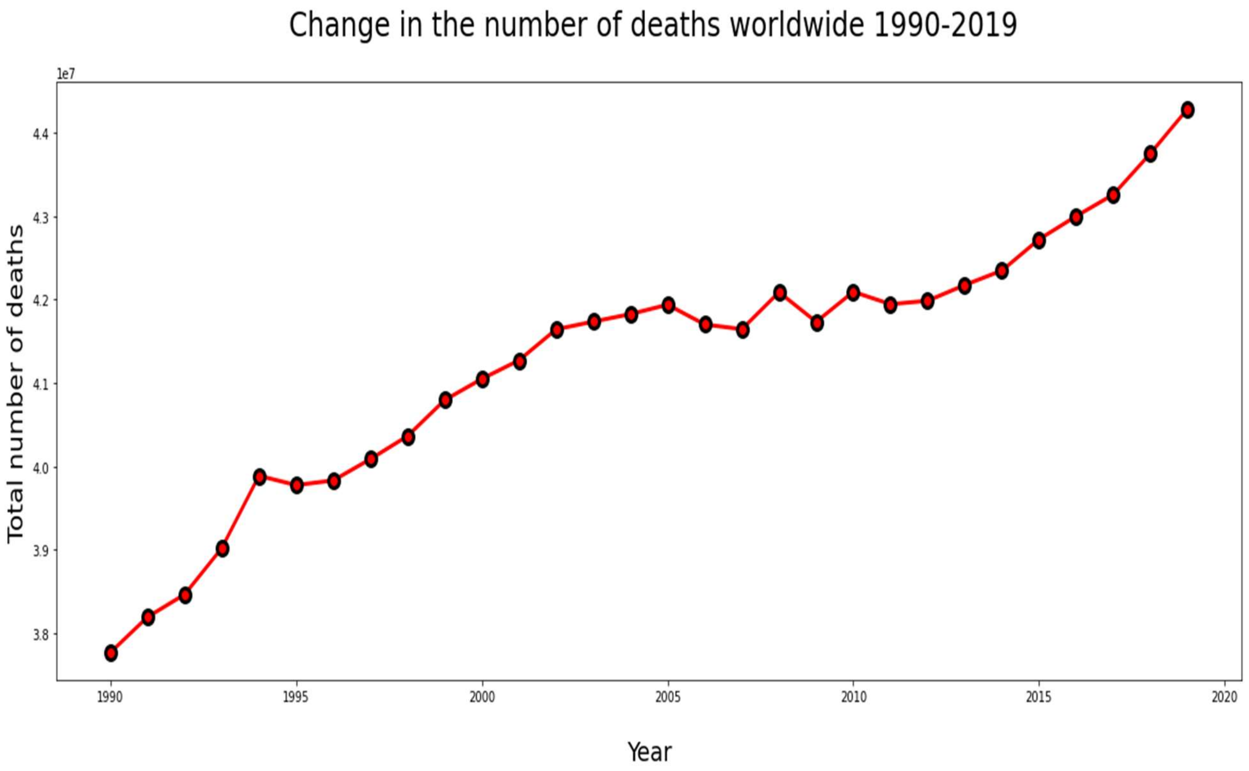
- **Communicable diseases**
- **Non-communicable diseases**
- **Injures**



During the 30 years from 1990 to 2019, the following trends were observed:

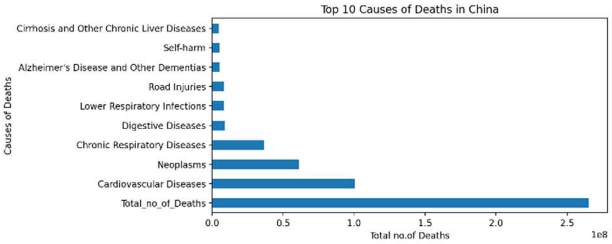
- The number of deaths from non-communicable diseases always accounts for the highest rate and tends to increase gradually.
- The number of deaths from communicable diseases accounts for the lowest rate, and maintains a fairly stable number over the years.
- The number of deaths from injures accounts for a high rate, but tends to decrease.

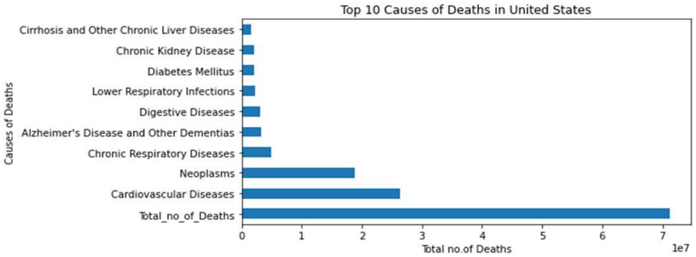
Number of deaths worldwide 1990-2019

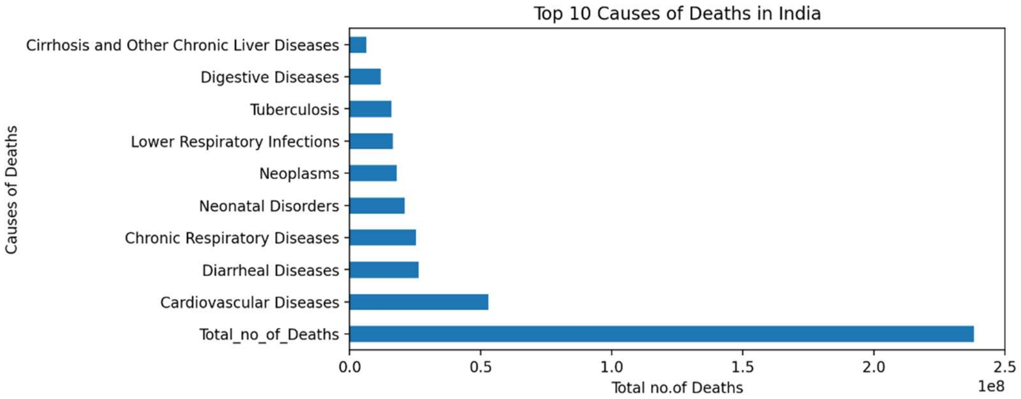


- The number of deaths in the world tends to increase each year, proportional to the population growth. Realizing that countries with a large population have a died and vice versa.
- Leading in the world in the number of deaths is: China, India, United States, Russia, Indonesia (whether in 1990 or 2019, these countries are still at the top of the number of deaths).
- The countries/territories with the highest or lowest number of deaths in 2019 are directly proportional to their population. This is considered reasonable.

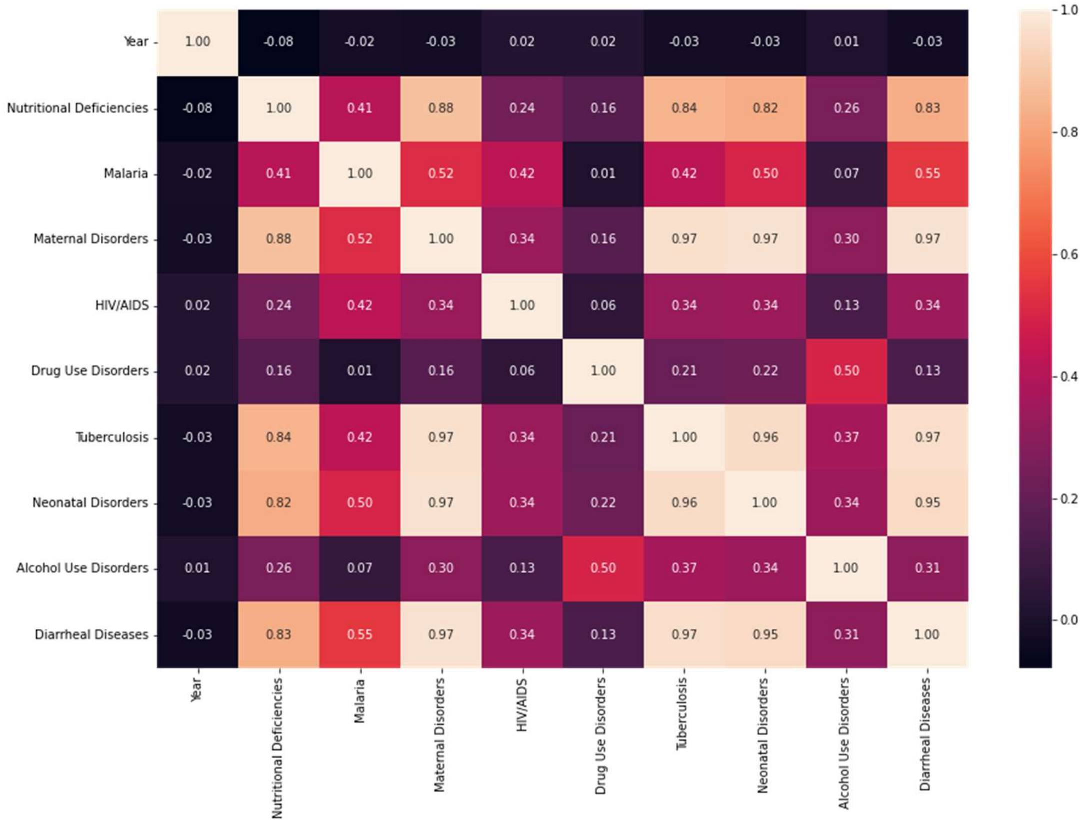
Top 10 Causes of Deaths – (China, India, USA)





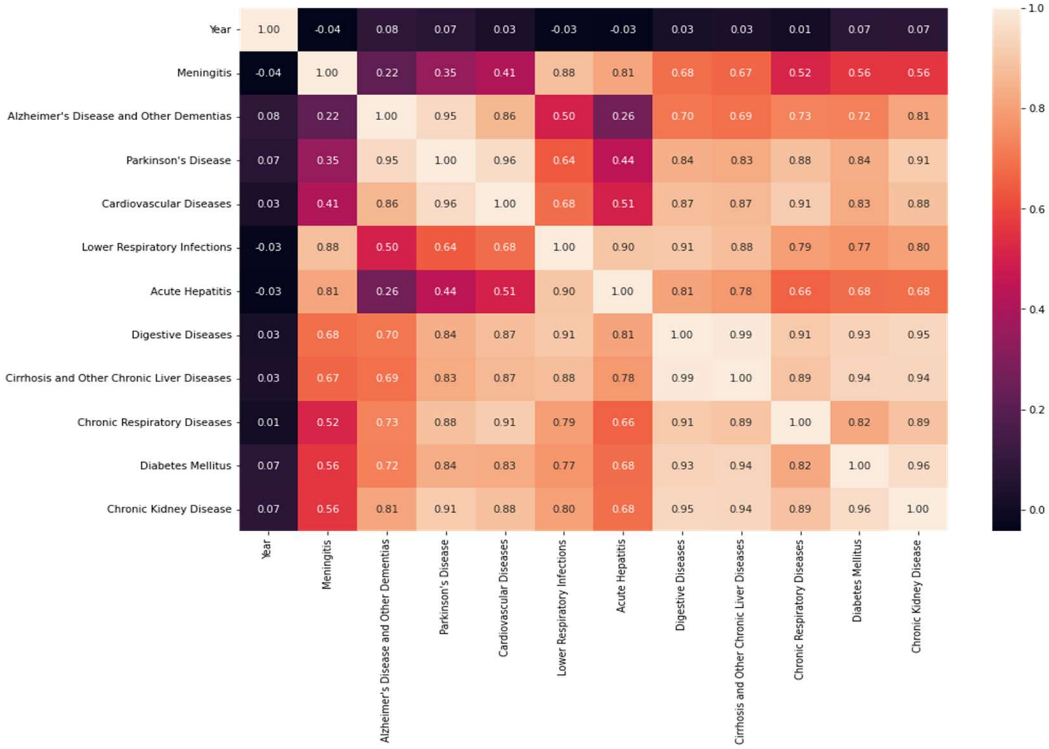


Correlation-communicable diseases



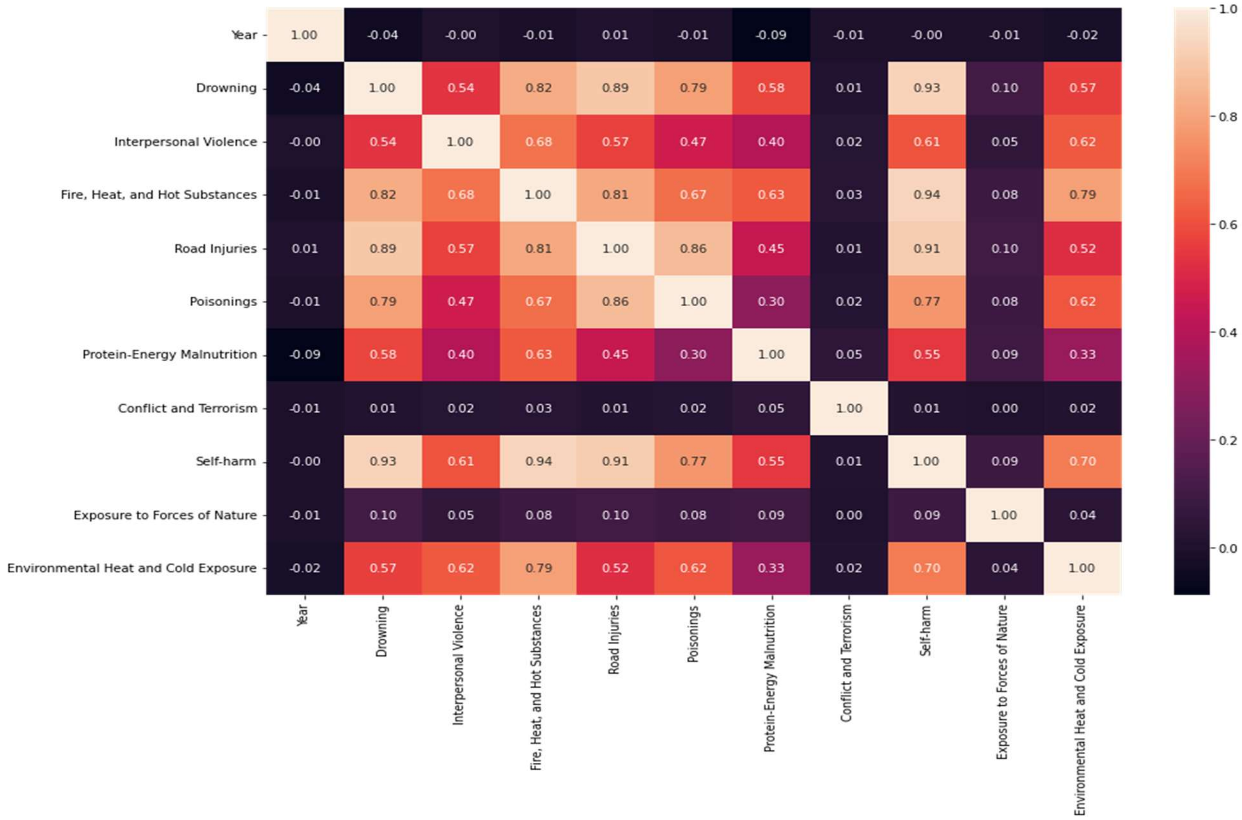
	Year	Nutritional Deficiencies	Malaria	Maternal Disorders	HIV/AIDS	Drug Use Disorders	Tuberculosis	Neonatal Disorders	Alcohol Use Disorders	Diarrheal Diseases
Year	1.00	-0.08	-0.02	-0.03	0.02	0.02	-0.03	-0.03	0.01	-0.03
Nutritional Deficiencies	-0.08	1.00	0.41	0.88	0.24	0.16	0.84	0.82	0.26	0.83
Malaria	-0.02	0.41	1.00	0.52	0.42	0.01	0.42	0.50	0.07	0.55
Maternal Disorders	-0.03	0.88	0.52	1.00	0.34	0.16	0.97	0.97	0.30	0.97
HIV/AIDS	0.02	0.24	0.42	0.34	1.00	0.06	0.34	0.34	0.13	0.34
Drug Use Disorders	0.02	0.16	0.01	0.16	0.06	1.00	0.21	0.22	0.50	0.13
Tuberculosis	-0.03	0.84	0.42	0.97	0.34	0.21	1.00	0.96	0.37	0.97
Neonatal Disorders	-0.03	0.82	0.50	0.97	0.34	0.22	0.96	1.00	0.34	0.95
Alcohol Use Disorders	0.01	0.26	0.07	0.30	0.13	0.50	0.37	0.34	1.00	0.31
Diarrheal Diseases	-0.03	0.83	0.55	0.97	0.34	0.13	0.97	0.95	0.31	1.00

Correlation -noncommunicable diseases



	Year	Meningitis	Alzheimer's Disease and Other Dementias	Parkinson's Disease	Cardiovascular Diseases	Lower Respiratory Infections	Acute Hepatitis	Digestive Diseases	Cirrhosis and Other Chronic Liver Diseases	Chronic Respiratory Diseases	Diabetes Mellitus	Chronic Kidney Disease
Year	1.00	-0.04	0.08	0.07	0.03	-0.03	-0.03	0.03	0.03	0.01	0.07	0.07
Meningitis	-0.04	1.00	0.22	0.35	0.41	0.88	0.81	0.68	0.67	0.52	0.56	0.56
Alzheimer's Disease and Other Dementias	0.08	0.22	1.00	0.95	0.86	0.50	0.26	0.70	0.69	0.73	0.72	0.81
Parkinson's Disease	0.07	0.35	0.95	1.00	0.96	0.64	0.44	0.84	0.83	0.88	0.84	0.91
Cardiovascular Diseases	0.03	0.41	0.86	0.96	1.00	0.68	0.51	0.87	0.87	0.91	0.83	0.88
Lower Respiratory Infections	-0.03	0.88	0.50	0.64	0.68	1.00	0.90	0.91	0.88	0.79	0.77	0.80
Acute Hepatitis	-0.03	0.81	0.26	0.44	0.51	0.90	1.00	0.81	0.78	0.66	0.68	0.68
Digestive Diseases	0.03	0.68	0.70	0.84	0.87	0.91	0.81	1.00	0.99	0.91	0.93	0.95
Cirrhosis and Other Chronic Liver Diseases	0.03	0.67	0.69	0.83	0.87	0.88	0.78	0.99	1.00	0.89	0.94	0.94
Chronic Respiratory Diseases	0.01	0.52	0.73	0.88	0.91	0.79	0.66	0.91	0.89	1.00	0.82	0.89
Diabetes Mellitus	0.07	0.56	0.72	0.84	0.83	0.77	0.68	0.93	0.94	0.82	1.00	0.96
Chronic Kidney Disease	0.07	0.56	0.81	0.91	0.88	0.80	0.68	0.95	0.94	0.89	0.96	1.00

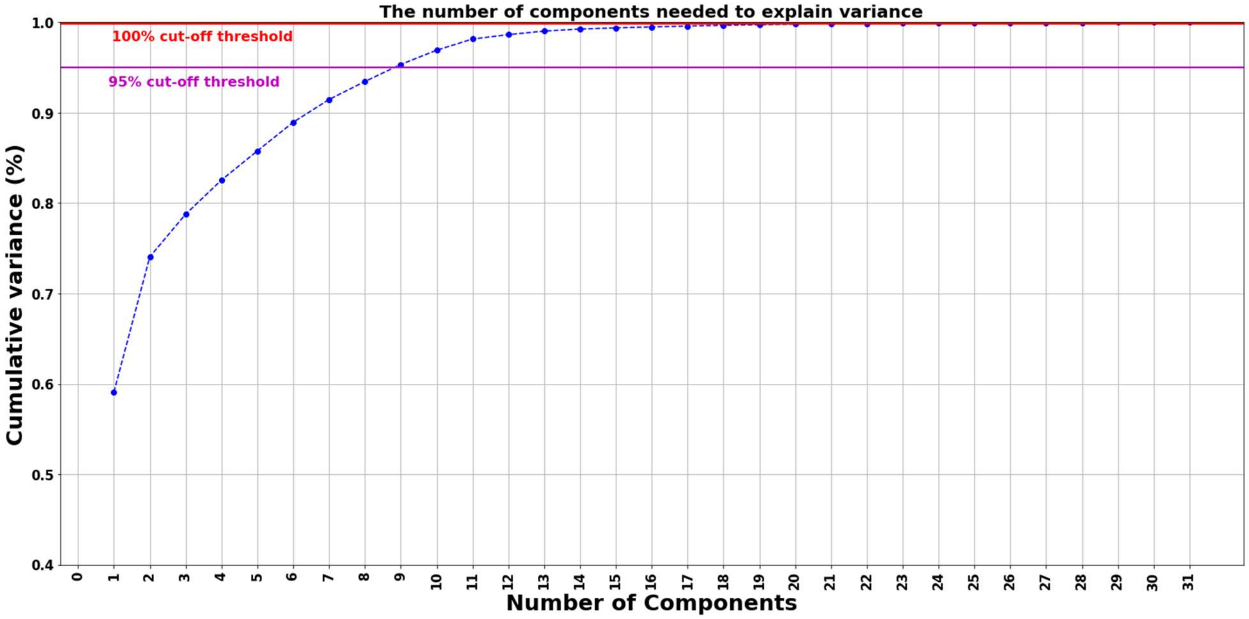
Correlation-injures



VIF & PCA

Check for Multicollinearity with VarianceInflationFactor(VIF) and apply Principal Component Analysis(PCA)→

- Typically, we will remove columns with VIF values > 10 which indicates strong multicollinearity of the features
- Multicollinearity can be addressed with either removing columns with VIF > 10 or using PCA
- As there are multiple features with VIF values > 10, we will apply PCA to reduce the number of features



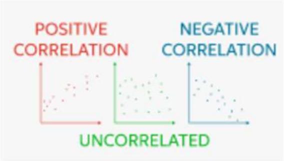
Observations→

- As per the graph, we can see that 8 principal components attribute for 95% of variation in the data. So, we will pick 8 components for our prediction
- We will use 8 features as number of components in PCA to reduce the dimensions

SUMMARY (How we got here!!)

EDA

- Data divided with Datatypes & Top Features
- Univariate Analysis for Features is performed for better insights
- Multivariate analysis is performed across features for Cause of Death



Correlation

- Correlation of Features

Analysis

- Count plot, Scatter plots provided insights
- Insights from combination of Multi features are derived

CONCLUSION

- ✓ Most dreadful disease which causes more than 300k deaths are Neoplasms, Cardiovascular diseases, Tuberculosis and HIV/AIDS.
- ✓ The disease which causes 0 deaths are Exposure to forces of nature and Malaria. India and China have the highest self-harming death rates.
- ✓ CHINA , INDIA AND USA face the largest brunt of deaths due to diseases in the world Cardiovascular diseases , Neoplasms (Malignancy/Cancer) and Lower Respiratory Tract Infections (for example : Pneumonia) are the top 3 killer diseases in the world.

DIRECTIONS OF FUTURE WORK

- ❖ Target variable can be added to data which will enable us 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.