Cause of Death Project

Submitted by:

Amita Saini

Acknowledgement

It gives me immense pleasure to express gratitude to Flip Robo Technologies who has provided this opportunity of doing this project to showcase my analytical skills. I express my special gratitude to my SME Ms Khushboo Garg, Flip Robo Technologies who has extended all help and support in doing this project.

I also express special thanks to Data Trained who provided opportunity of Internship at Flip Robo Technologies. I also extend my deep gratitude to my family whose support has been vital at every step of my life.

References:

1. Blogs from Analytics Vidya & towardsdatascience
2. SCIKIT Learn Library Documentation
3. Data Science Projects with Python Second Edition by Packt
4. Hands on Machine Learning with Scikit learn and tensor flow by Aurelien Geron

Introduction

Business Problem Framing

A straightforward way to assess the health status of a population is to focus on mortality – or concepts 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 diseases) provides a more encompassing view on health outcomes. This is the topic of this entry. The sum of mortality and morbidity is referred to as the 'burden of disease' and can be measured by a metric called 'Disability Adjusted Life Years' (DALYs). DALYs are measuring 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. Conceptually, 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. The first 'Global Burden of Disease' (GBD) was GBD 1990 and the DALY metric was prominently featured in the World Bank's 1993 World Development Report. Today it is published by both the researchers at the Institute of Health Metrics and Evaluation (IHME) and the 'Disease Burden Unit' at the World Health Organization (WHO), which was created in 1998. The IHME continues the work that was started in the early 1990s and publishes the Global Burden of Disease study.
So, in depth analysis is required to present data on cause of death.

Conceptual Background of the Domain Problem

In this Dataset, we have Historical Data of different cause of deaths for all ages around the World. The key features of this Dataset are: 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.

- ## Review of Literature

  The Steps followed were as per below mentioned Links:

  https://www.kaggle.com/code/spscientist/a-simple-tutorial-on-exploratory-data-analysis
  https://en.wikipedia.org/wiki/Exploratory_data_analysis#:~:text=In%20statistics%2C%20exploratory%20data%20analysis,and%20other%20data%20visualization%20methods

- ## Motivation for the Problem Undertaken

  The project is provided to me by Flip Robo Technologies as a part of Internship. The deepest learning and solving real life case were chief motivators.

Analytical Problem Framing

- ## Mathematical/ Analytical Modeling of the Problem

  Univariate analysis and Bivariate Analysis. Also, further bifurcation of cause of Death and analysis against year for specific country and Disease.

- ## Data Sources and their formats

  The dataset provided by Flip Robo was in the format of CSV (Comma Separated Values). The dimension of data is 6120 Rows and 34 Columns.

Import Data

```
In [3]: df=pd.read_csv('cause_of_deaths dataset.csv')
In [4]: df
```

Out[4]:

	Country/Territory	Code	Year	Meningitis	Alzheimer's Disease and Other Dementias	Parkinson's Disease	Nutritional Deficiencies	Malaria	Drowning	Interpersonal Violence	...	Diabetes Mellitus	Chronic Kidney Disease	Poisonings	Ma
0	Afghanistan	AFG	1990	2159	1116	371	2087	93	1370	1538	...	2108	3709	338	
1	Afghanistan	AFG	1991	2218	1136	374	2153	189	1391	2001	...	2120	3724	351	
2	Afghanistan	AFG	1992	2475	1162	378	2441	239	1514	2299	...	2153	3776	386	
3	Afghanistan	AFG	1993	2812	1187	384	2837	108	1687	2589	...	2195	3862	425	
4	Afghanistan	AFG	1994	3027	1211	391	3081	211	1809	2849	...	2231	3932	451	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
6115	Zimbabwe	ZWE	2015	1439	754	215	3019	2518	770	1302	...	3176	2108	381	
6116	Zimbabwe	ZWE	2016	1457	767	219	3056	2050	801	1342	...	3259	2160	393	
6117	Zimbabwe	ZWE	2017	1460	781	223	2990	2116	818	1363	...	3313	2196	398	
6118	Zimbabwe	ZWE	2018	1450	795	227	2918	2088	825	1396	...	3381	2240	400	
6119	Zimbabwe	ZWE	2019	1450	812	232	2884	2068	827	1434	...	3460	2292	405	

6120 rows × 34 columns

- ## Data Pre processing Done

  The dataset provided is large and it may contain some data error. In order to reach clean, error free data some data cleaning & data pre-processing performed data.

Data Inspection

```
In [6]: # Data information
        df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6120 entries, 0 to 6119
Data columns (total 34 columns):
 #   Column                                  Non-Null Count  Dtype
---  ------                                  --------------  -----
 0   Country/Territory                       6120 non-null   object
 1   Code                                    6120 non-null   object
 2   Year                                    6120 non-null   int64
 3   Meningitis                              6120 non-null   int64
```

Data types is correct but changing year to categorical for better analysis

```
In [7]: df['Year']=df.Year.astype('object')
```

```
In [8]: # Check Duplicated records
        df.duplicated().sum()
```

```
Out[8]: 0
```

There are no duplicated records

```
In [9]: # Check the null values
        df.isna().sum()
```

```
Out[9]: Country/Territory                       0
        Code                                    0
        Year                                    0
        Meningitis                              0
        Alzheimer's Disease and Other Dementias 0
        Parkinson's Disease                     0
        Nutritional Deficiencies                0
        Malaria                                 0
        Drowning                                0
```

There are no null values

```
In [11]: # Check the number of uniques countries
         df['Country/Territory'].nunique()
```

```
Out[11]: 204
```

```
In [12]: # Check the number of uniques code
         df['Code'].nunique()
```

```
Out[12]: 204
```

It seems alot of outliers or skewness are in the data

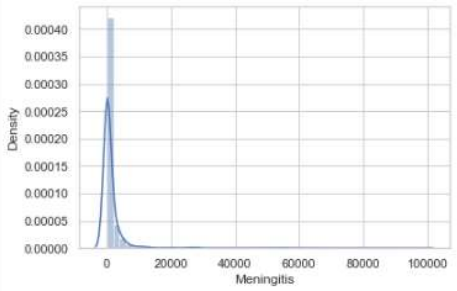
```
In [16]: df1_cat.describe(include=['O'])
```

Out[16]:

	Country/Territory	Year
count	6120	6120
unique	204	30
top	Afghanistan	1990
freq	30	204

```
In [17]: # Check Normality of continous data
for i in df1_cont.columns:
    sns.distplot(df1_cont[i])
    print(kstest(df1_cont[i].values,'norm'))
    plt.show()
```

```
KstestResult(statistic=0.8829688223001868, pvalue=0.0)
```




```
KstestResult(statistic=0.9661337620990889, pvalue=0.0)
```

All of the features are not follow the NDs (Right Skewed) as it shown in both graphs and Statistics

```
In [20]: # The unique Year data in the Dataframe

df1_cat['Year'].unique()
```

```
Out[20]: array([1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000,
       2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011,
       2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019], dtype=object)
```

There are 30 years of statistics in this data set (1990-2019).

```
In [21]: # Creating a new column for 'Total_no_of_Deaths' for individual Country and Year

df2['Total_no_of_Deaths'] = df2.sum(axis=1)
```

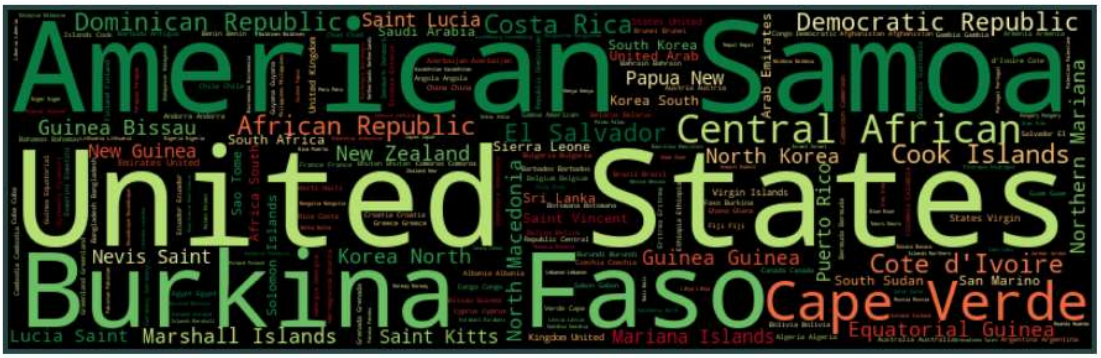
```
In [22]: df2
```

Out[22]:

	Country/Territory	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
0	Afghanistan	1990	2159	1116	371	2087	93	1370	1538	2655	34	93	4661
1	Afghanistan	1991	2218	1136	374	2153	189	1391	2001	2885	41	102	4743
2	Afghanistan	1992	2475	1162	378	2441	239	1514	2299	3315	46	118	4976

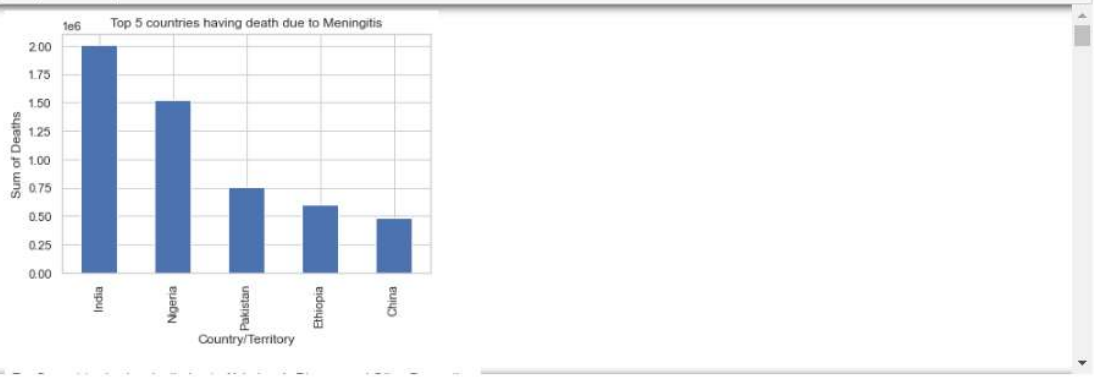
Univariate analysis

In [23]:
```
text = " ".join(subject_titles for subject_titles in df1_cat["Country/Territory"])
wordcloud = WordCloud(width = 800, height = 250,
                background_color ="black",colormap="RdYlGn",max_font_size=100, stopwords =None,repeat= True).generate(text)
plt.figure(figsize = (20, 8),facecolor= "#254441")
plt.imshow(wordcloud)
plt.axis("off")
plt.margins(x=0, y=0)
plt.tight_layout(pad = 0)
plt.show()
```



Bivariate analysis

In [24]:
```
for i in df1_cont.columns:
    df2.groupby('Country/Territory')[i].sum().sort_values(ascending=False).head().plot(kind='bar')
    plt.title('Top 5 countries having death due to '+i)
    plt.ylabel('Sum of Deaths')
    plt.show()
```



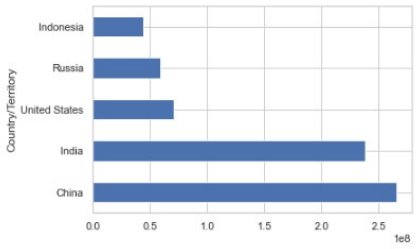
Data Inputs- Logic- Output Relationships

In [25]:
```
# Top countries in different death disease/enviroment  from above graphs
# India---- Meningits, Nutritional Deficiencies, Maternal Disorders, Tuberculosis, Lower Respiratory Infections, Self-harm
# Neonatal Disorders,Diarrheal Diseases, Diabetes Mellitus,Choronic kidney,Protein-energy malnutrition,Digestive,Acute Hepatitis
# Cirrhosis & other chronic liver, Fire/heat & hot substances

# China---- Alzheimer, Parkinson, Drowning, Cardiovascular Diseases, Neoplasms, Poisonings, Road injuries, Chronic respiratory
# Nigeria---- Malaria
# Brazil---- Interpersonal Violence
# South Africa---- HIV/AIDS
# USA---- Drug use Disorders
# Russia---- Alcohol use Disorders, Environmental Heat&cold exposure
# Haiti---- Exposure to forces of nature
# Rwanda---- Conflict & Terrorism
```

In [29]:
```
df2.groupby('Country/Territory')['Total_no_of_Deaths'].sum().sort_values(ascending=False).head().plot(kind='barh')
plt.show()
```

Out[29]: <AxesSubplot:ylabel='Country/Territory'>



- ## Hardware and Software Requirements and Tools Used

  Hardware used –

  1. Processor – Intel i3 processor with 2.4Ghz
  2. RAM – 4GB

  Software used –

  1. Anaconda – Jupyter notebook

  Libraries used

Import Libraries

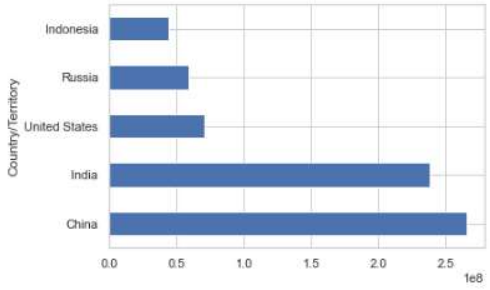
In [1]:
```
import numpy as np
import pandas as pd
from scipy.stats import kstest
import matplotlib.pyplot as plt
%matplotlib inline
import plotly.express as px
from wordcloud import WordCloud
import seaborn as sns
sns.set(style='whitegrid')
import warnings
warnings.filterwarnings('ignore')
from sklearn.preprocessing import PowerTransformer
```

Model/s Development and Evaluation

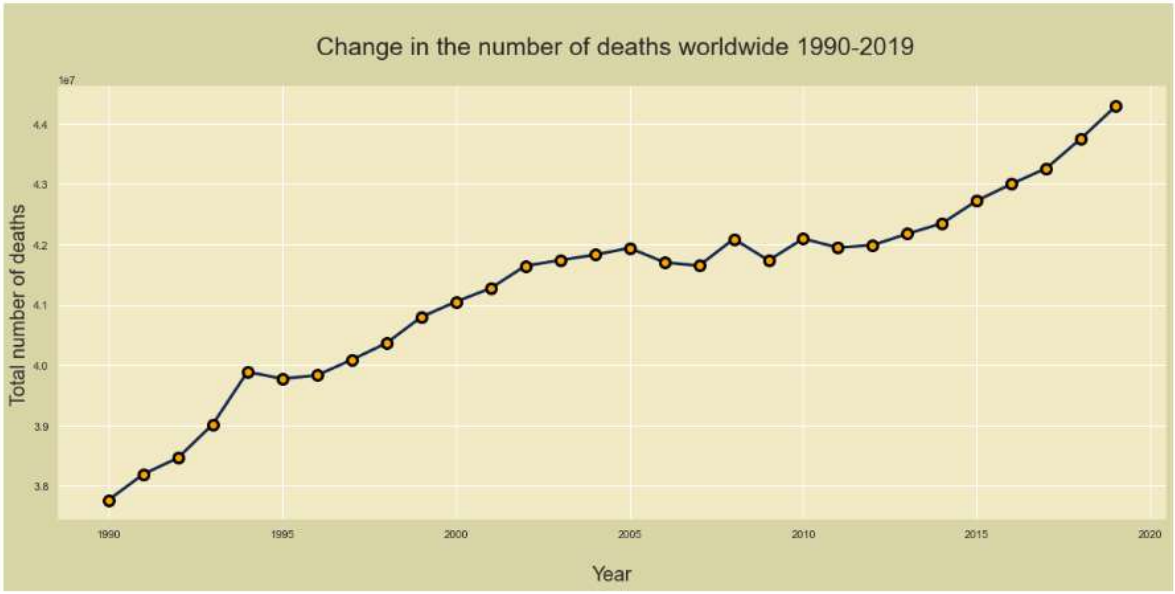
- Identification of possible problem-solving approaches (methods)

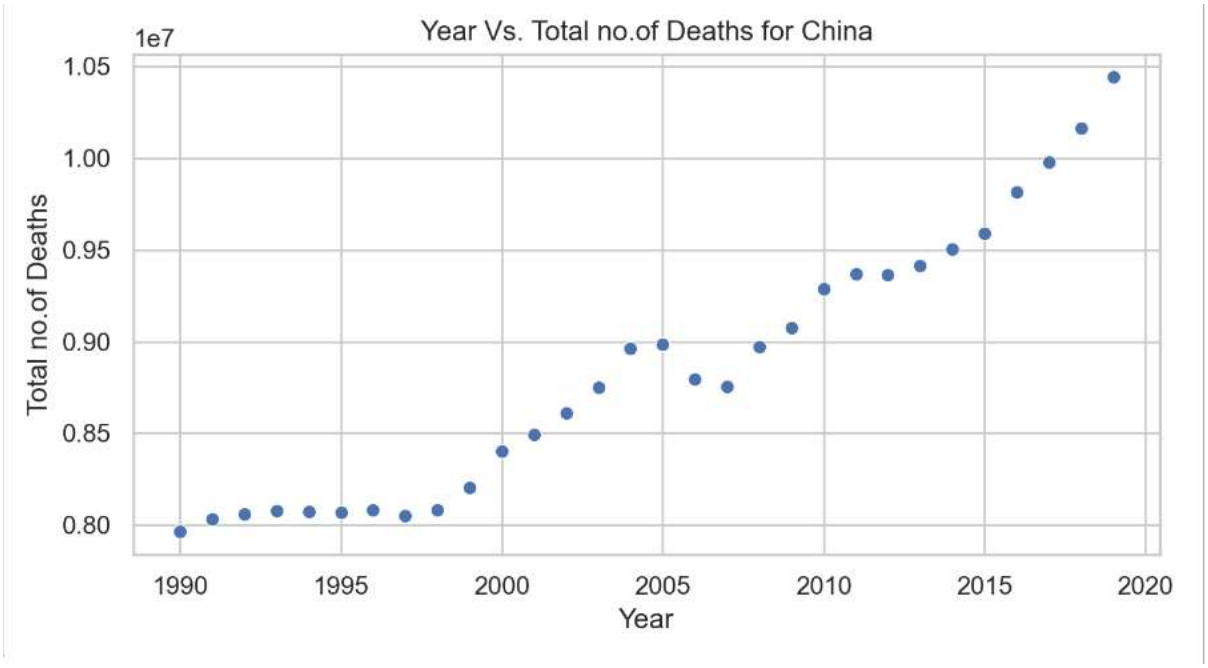
Univariate analysis and Bivariate Analysis. Also, further bifurcation of cause of Death and analysis against year for specific country and Disease.

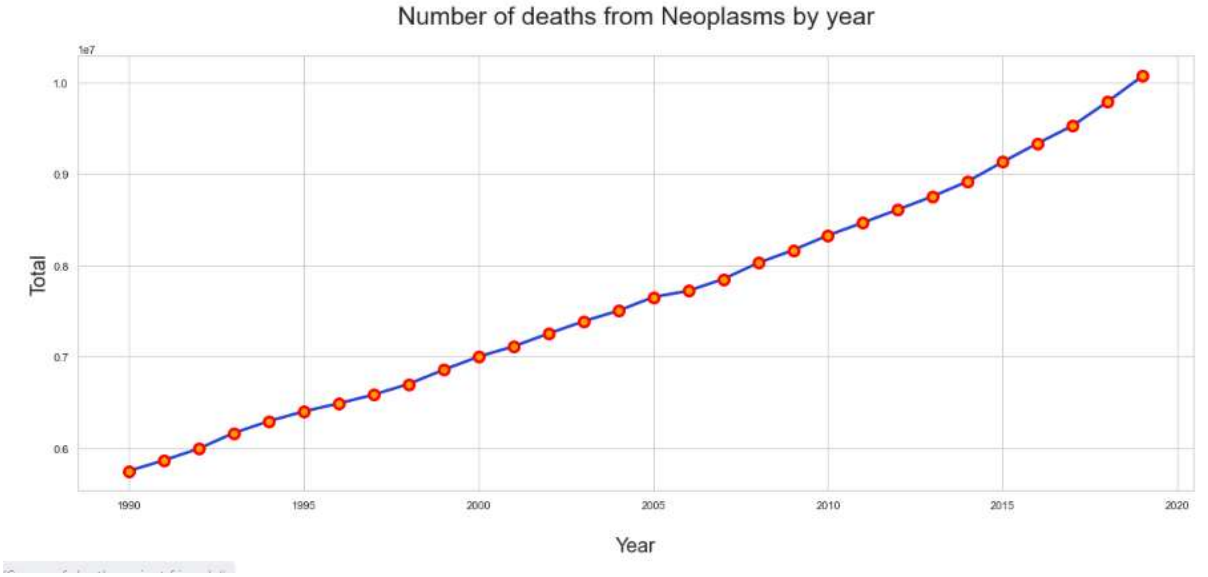
- Visualizations and Interpretation of the Results




```
Total_no_of_Deaths belongs to 'China' and 'India' combined, followed by 'USA' and 'Russia' and finally 'Indonesia'. This
because of the fact "China" and "India" are the countries that stand in top 2 interms of population.
```



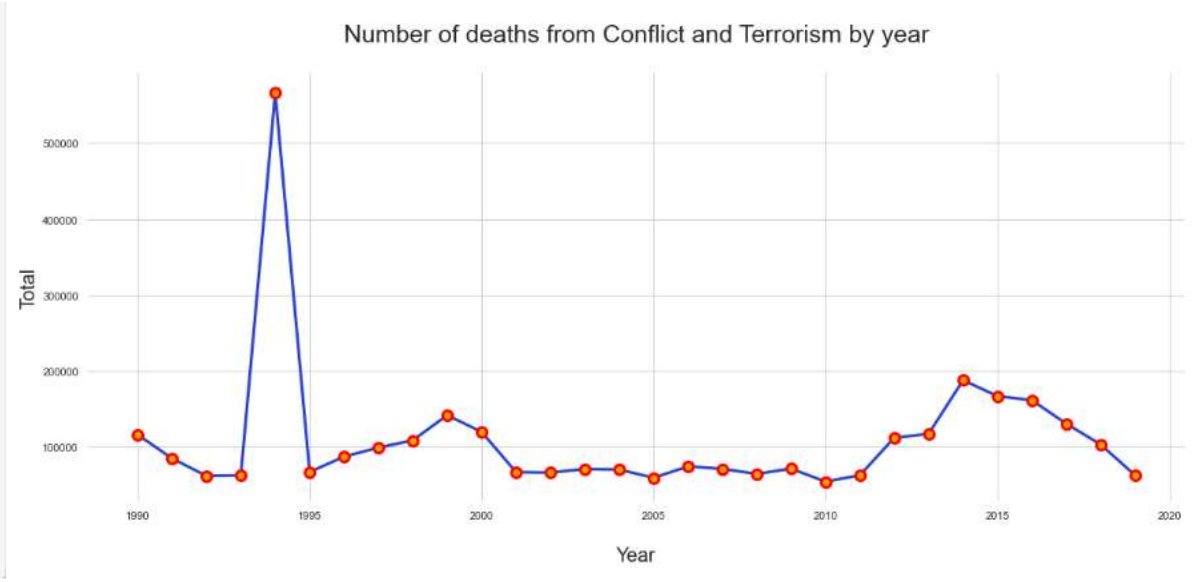




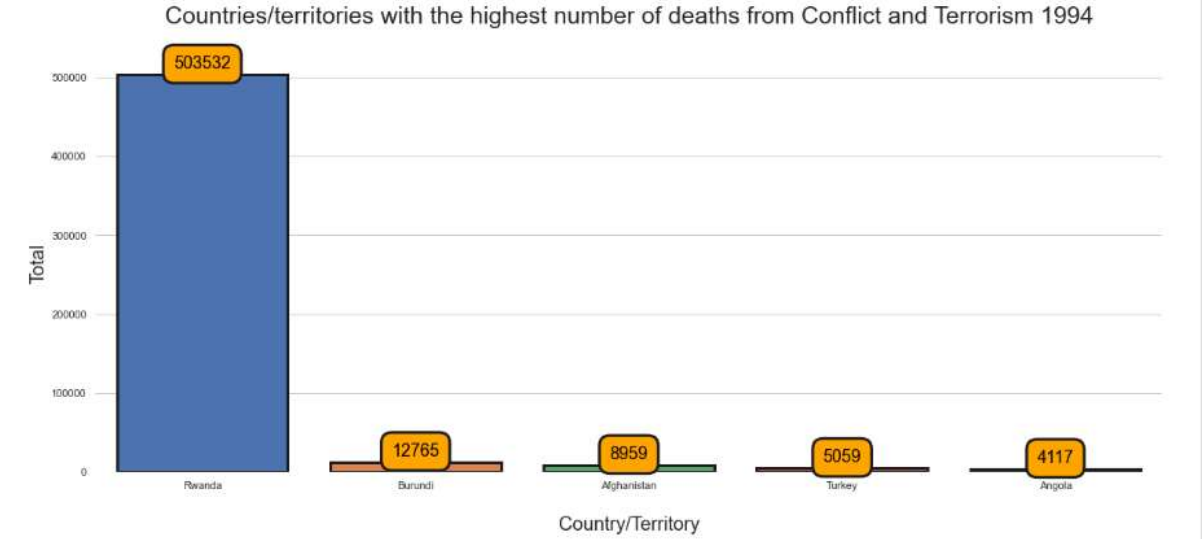
Rates of new cancer and cancer deaths continue to increase year by year. Cancer is not a disease, but a group of diseases. To date, about 200 different types of letters on the human body have been identified.

Cancers are classified by the type of cell that the tumor cells resemble and is therefore presumed to be the origin of the tumor. These types include:

Carcinoma: Cancers derived from epithelial cells. This group includes many of the most common cancers and include nearly all those in the breast, prostate, lung, pancreas and colon.
Sarcoma: Cancers arising from connective tissue (i.e. bone, cartilage, fat, nerve), each of which develops from cells originating in mesenchymal cells outside the bone marrow.
Lymphoma and leukemia: These two classes arise from hematopoietic (blood-forming) cells that leave the marrow and tend to mature in the lymph nodes and blood, respectively.[122]
Germ cell tumor: Cancers derived from pluripotent cells, most often presenting in the testicle or the ovary (seminoma and dysgerminoma, respectively).
Blastoma: Cancers derived from immature "precursor" cells or embryonic tissue.

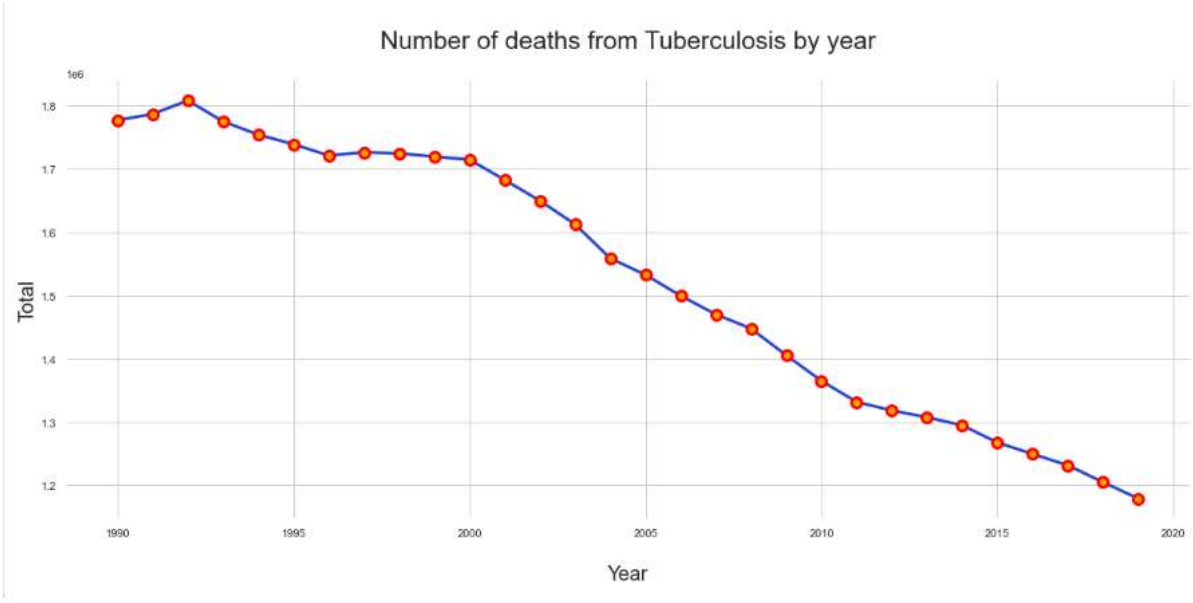


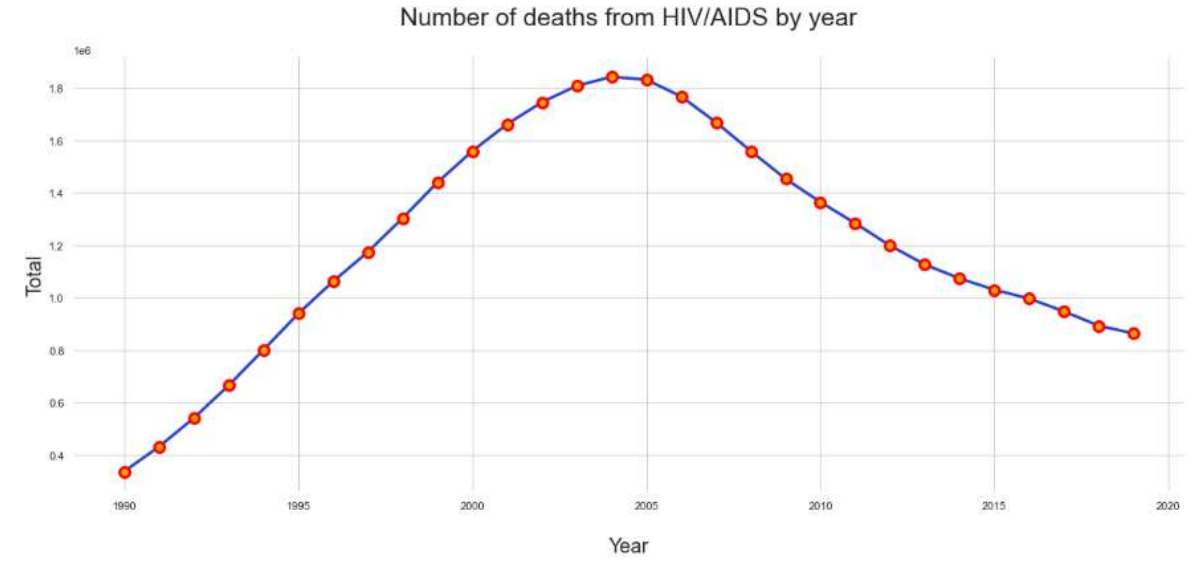

```
Tuberculosis dates back to BC. In the past, due to the lack of understanding of the cause of TB, TB was considered a genetic
disease
```




```
In 1994, the number of deaths from Conflict and Terrorism skyrocketed in Rwanda

Cause: The Rwandan genocide occurred between 7 April and 15 July 1994 during the Rwandan Civil War. During this period of around
100 days, members of the Tutsi minority ethnic group, as well as some moderate Hutu and Twa, were killed by armed Hutu militias.
The most widely accepted scholarly estimates are around 500,000 to 662,000 Tutsi deaths.
```



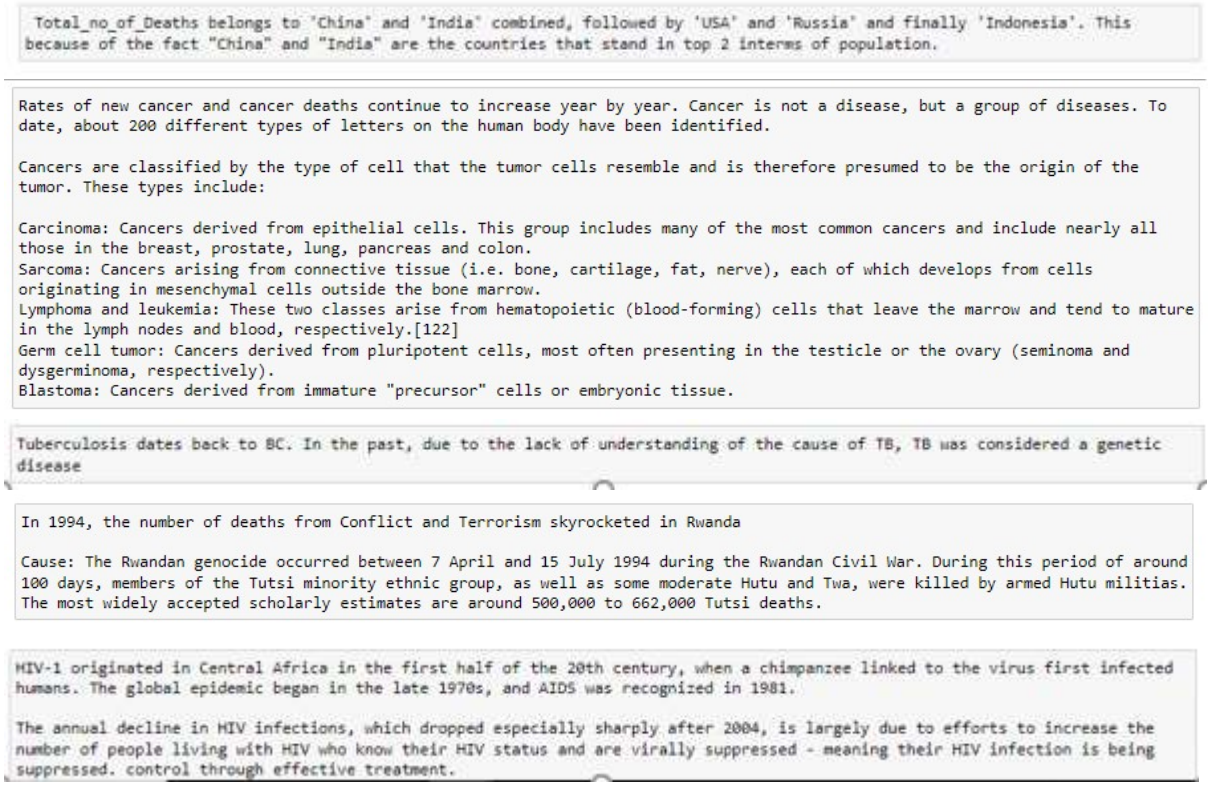


HIV-1 originated in Central Africa in the first half of the 20th century, when a chimpanzee linked to the virus first infected humans. The global epidemic began in the late 1970s, and AIDS was recognized in 1981.

The annual decline in HIV infections, which dropped especially sharply after 2004, is largely due to efforts to increase the number of people living with HIV who know their HIV status and are virally suppressed - meaning their HIV infection is being suppressed. control through effective treatment.

CONCLUSION

- Key Findings and Conclusions of the Study



- Learning Outcomes of the Study in respect of Data Science

  This Study gives inferences through visuals in a quick way.

- Limitations of this work and Scope for Future Work

  This Study is limited to few diseases and year analysis.