HOUSING PRICE PREDICTION

Submitted by

P.NagaSuvarchala

ACKNOWLEDGMENT

I would like to express my deep sense of gratitude to my SME (Subject Matter Expert) **Mr. Khushboo Garg** as well as **Flip Robo Technologies** who gave me the golden opportunity to do this data analysis project on **Housing: Price Prediction**, which also helped me in doing lots of research and I came to know about so many new things.

I am very much thankful to **Dr. Deepika, Trainer (DataTrained)**, for her valuable guidance, keen interest, and encouragement at various stages of my training period which eventually helped me a lot in doing this project.

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.

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.

Dataset Glossary (Column-wise)

- 01. Country/Territory - Name of the Country/Territory
- 02. Code - Country/Territory Code
- 03. Year - Year of the Incident
- 04. Meningitis - No. of People died from Meningitis
- 05. Alzheimer's Disease and Other Dementias - No. of People died from Alzheimer's Disease and Other Dementias
- 06. Parkinson's Disease - No. of People died from Parkinson's Disease
- 07. Nutritional Deficiencies - No. of People died from Nutritional Deficiencies
- 08. Malaria - No. of People died from Malaria
- 09. Drowning - No. of People died from Drowning
- 10. Interpersonal Violence - No. of People died from Interpersonal Violence
- 11. Maternal Disorders - No. of People died from Maternal Disorders
- 12. Drug Use Disorders - No. of People died from Drug Use Disorders
- 13. Tuberculosis - No. of People died from Tuberculosis
- 14. Cardiovascular Diseases - No. of People died from Cardiovascular Diseases
- 15. Lower Respiratory Infections - No. of People died from Lower Respiratory Infections
- 16. Neonatal Disorders - No. of People died from Neonatal Disorders
- 17. Alcohol Use Disorders - No. of People died from Alcohol Use Disorders
- 18. Self-harm - No. of People died from Self-harm
- 19. Exposure to Forces of Nature - No. of People died from Exposure to Forces of Nature
- 20. Diarrheal Diseases - No. of People died from Diarrheal Diseases
- 21. Environmental Heat and Cold Exposure - No. of People died from Environmental Heat and Cold Exposure
- 22. Neoplasms - No. of People died from Neoplasms
- 23. Conflict and Terrorism - No. of People died from Conflict and Terrorism
- 24. Diabetes Mellitus - No. of People died from Diabetes Mellitus

- 25. Chronic Kidney Disease - No. of People died from Chronic Kidney Disease
- 26. Poisonings - No. of People died from Poisoning
- 27. Protein-Energy Malnutrition - No. of People died from Protein-Energy Malnutrition
- 28. Chronic Respiratory Diseases - No. of People died from Chronic Respiratory Diseases
- 29. Cirrhosis and Other Chronic Liver Diseases - No. of People died from Cirrhosis and Other Chronic Liver Diseases
- 30. Digestive Diseases - No. of People died from Digestive Diseases
- 31. Fire, Heat, and Hot Substances - No. of People died from Fire or Heat or any Hot Substances
- 32. Acute Hepatitis - No. of People died from Acute Hepatitis

Motivation for the Problem Undertaken

There are several motivations for conducting an Data Analysis project on Causes of Death.

The prime intention of performing this project is Exploratory Data analysis(EDA) model to find Cause of Death with the help of other supporting feature attributes. The sample data is provided by our client database.

Public Health: Understanding the leading causes of deaths can inform public health policies and interventions aimed at reducing mortality rate and improving population health.

Epidemiology: The data can be used to study patterns of disease and death across different populations, identify high- risk groups, and inform the development of targeted preventive measures.

Health Service planning: The analysis can help to identify the need for specific health services, such as access to specialized care, and allocate resources more effectively.

Research: The data can be used as a resource for a academic research, helping to advance the understanding of disease and death and inform the development of new treatments and interventions.

Awareness: The result of the analysis can be used to raise awareness about important public health issues and promote healthy behavior.

Overall, a data analysis project on causes of death can have significant implications of improving population health and reducing mortality rates.

Analytical Problem Framing:

Mathematical/ Analytical Modeling of the Problem:

The provided dataset is in XLs format. We will begin with loading the dataset andreading the dataset from the XLs file using the read_csv() function from the Pandas Python package.

Next, we will perform Non-Graphical Exploratory Data Analysis (EDA) such as checking the data types and missing values using pandas info() function, Then, we will get statistical information about the numeric columns in our dataset using pandas. DataFrame.describe() method. We want to know the mean,the standard deviation, the minimum, the maximum, and the 50th percentile (the median) for each numeric column in the dataset. After that, we will move on to perform graphical EDA to get more insights from our dataset and how the feature attribute affects the target attribute.

Since we have to find Cause of Deaths and patterns.

Data Sources and their formats:

The dataset is being provided by Flip Robo and is in the format of xls Separated. Let’s load it and start the analysis.

Importing Libraries

importing Libraries

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

Loading Dataset:-

.

In [2]:
```
df = pd.read_csv('cause_of_deaths dataset.csv')
df
```

Out[2]:

	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	

Dimension of the data set:-

[3]:
```
number_of_rows,no_of_columns=df.shape
print(f'Number of rows:{number_of_rows}\nNumber of columns:{no_of_columns}')
```

```
Number of rows:6120
Number of columns:34
```

The Data Type Information:-

Data Type of Each Column

```
[4]: 1 column_name=list(df.columns.values)
     2 column_dtype=pd.Series(df[column_name].dtypes)
     3 column_dtype
```

```
[4]: Country/Territory                              object
     Code                                           object
     Year                                            int64
     Meningitis                                      int64
     Alzheimer's Disease and Other Dementias         int64
     Parkinson's Disease                             int64
     Nutritional Deficiencies                        int64
     Malaria                                         int64
     Drowning                                        int64
     Interpersonal Violence                          int64
     Maternal Disorders                              int64
     HIV/AIDS                                        int64
     Drug Use Disorders                              int64
     Tuberculosis                                    int64
     Cardiovascular Diseases                         int64
     Lower Respiratory Infections                    int64
     Neonatal Disorders                              int64
     Alcohol Use Disorders                           int64
     Self-harm                                       int64
     Exposure to Forces of Nature                    int64
     Diarrheal Diseases                              int64
```

Checking missing values/null values:-

```
6]:  1  df.isnull().sum()
```

```
6]: 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
    Interpersonal Violence                     0
    Maternal Disorders                         0
    HIV/AIDS                                   0
    Drug Use Disorders                         0
    Tuberculosis                               0
    Cardiovascular Diseases                    0
    Lower Respiratory Infections               0
    Neonatal Disorders                         0
    Alcohol Use Disorders                      0
    Self-harm                                  0
    Exposure to Forces of Nature               0
    Diarrheal Diseases                         0
    Environmental Heat and Cold Exposure       0
    Neoplasms                                  0
    Conflict and Terrorism                     0
    Diabetes Mellitus                          0
    Chronic Kidney Disease                     0
    Poisonings                                 0
```

There is no missing values in data

Statistical summary of Data:-

```
:  1  df.describe()
```

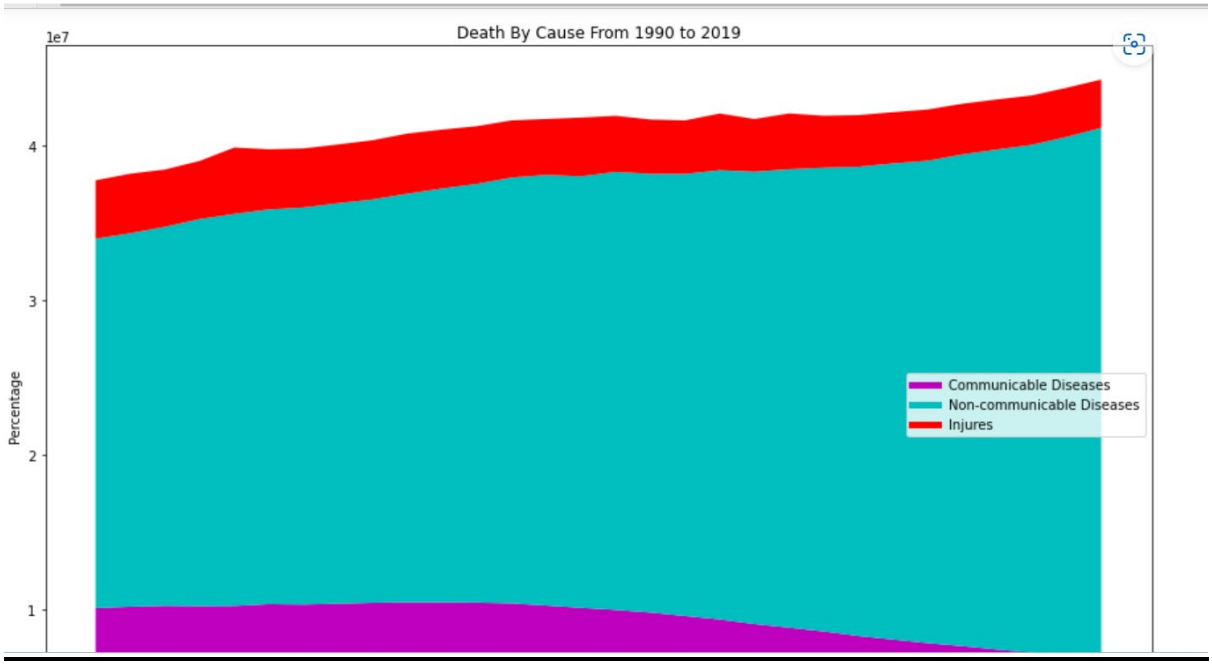
:

	Year	Meningitis	Alzheimer's Disease and Other Dementias	Parkinson's Disease	Nutritional Deficiencies	Malaria	Drowning	Interpersonal Violence	Maternal Disorders	HIV/AIDS	...
count	6120.000000	6120.000000	6120.000000	6120.000000	6120.000000	6120.000000	6120.000000	6120.000000	6120.000000	6120.000000	...
mean	2004.500000	1719.701307	4864.189379	1173.169118	2253.600000	4140.960131	1683.333170	2083.797222	1262.589216	5941.898529	...
std	8.656149	6672.006930	18220.659072	4616.156238	10483.633601	18427.753137	8877.018366	6917.006075	6057.973183	21011.962487	...
min	1990.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...
25%	1997.000000	15.000000	90.000000	27.000000	9.000000	0.000000	34.000000	40.000000	5.000000	11.000000	...
50%	2004.500000	109.000000	666.500000	164.000000	119.000000	0.000000	177.000000	265.000000	54.000000	136.000000	...
75%	2012.000000	847.250000	2456.250000	609.250000	1167.250000	393.000000	698.000000	877.000000	734.000000	1879.000000	...
max	2019.000000	98358.000000	320715.000000	76990.000000	268223.000000	280604.000000	153773.000000	69640.000000	107929.000000	305491.000000	...

8 rows × 32 columns

Divide the causes of death into 3 main categories ¶

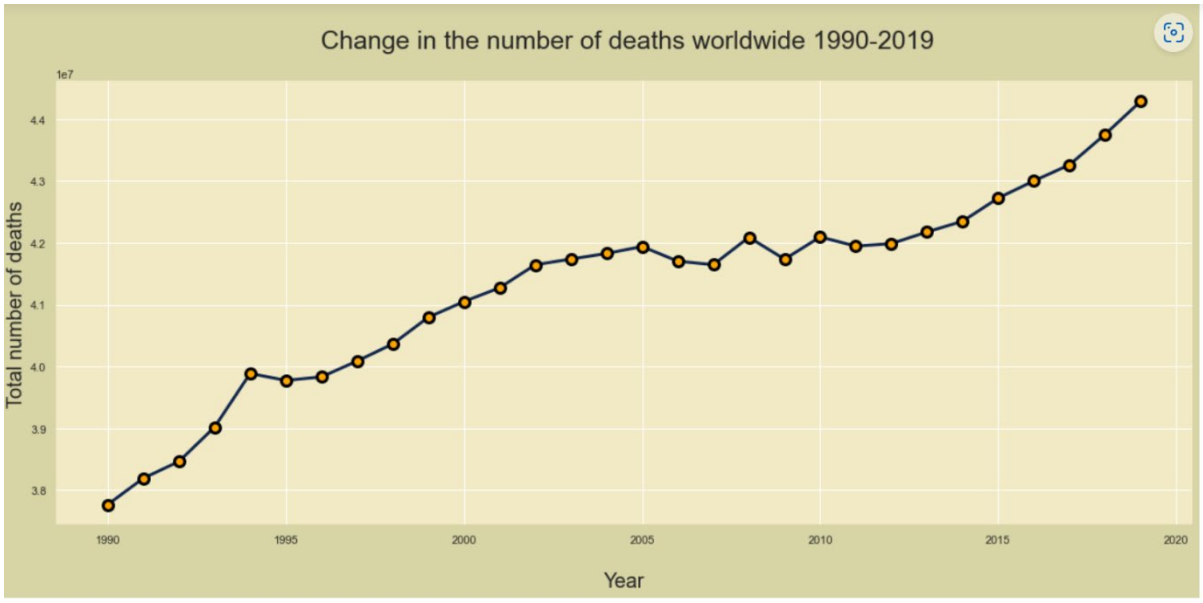
Communicable diseases	Non-communicable diseases	Injures
Nutritional Deficiencies	Meningitis,Alzheimer's Disease and Other Dementias	Drowning
Malaria	Parkinson's Disease	Interpersonal Violence
Maternal Disorders	Cardiovascular Diseases	Fire, Heat, and Hot Substances
HIV/AIDS	Lower Respiratory Infections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



Observations:

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

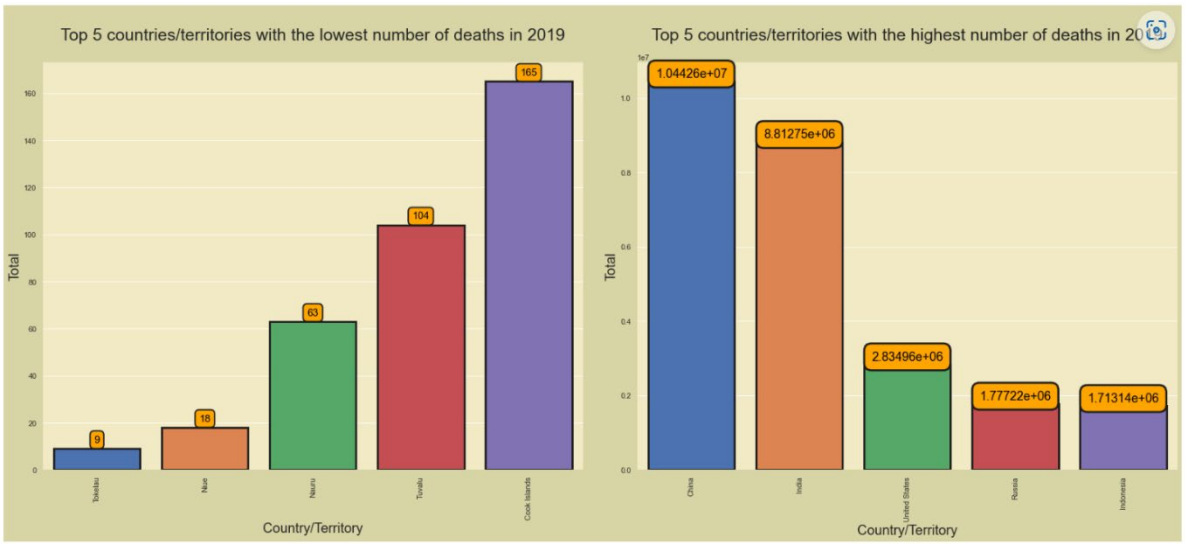
- The number of deaths from non-communicable diseases always accounts for the highest rate and tends to increase gradually.
- The number of deaths from communicable diseases accounts for the lowest rate, and maintains a fairly stable number over the years.
- The number of deaths from injures accounts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



Observations:-

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).



Observations:

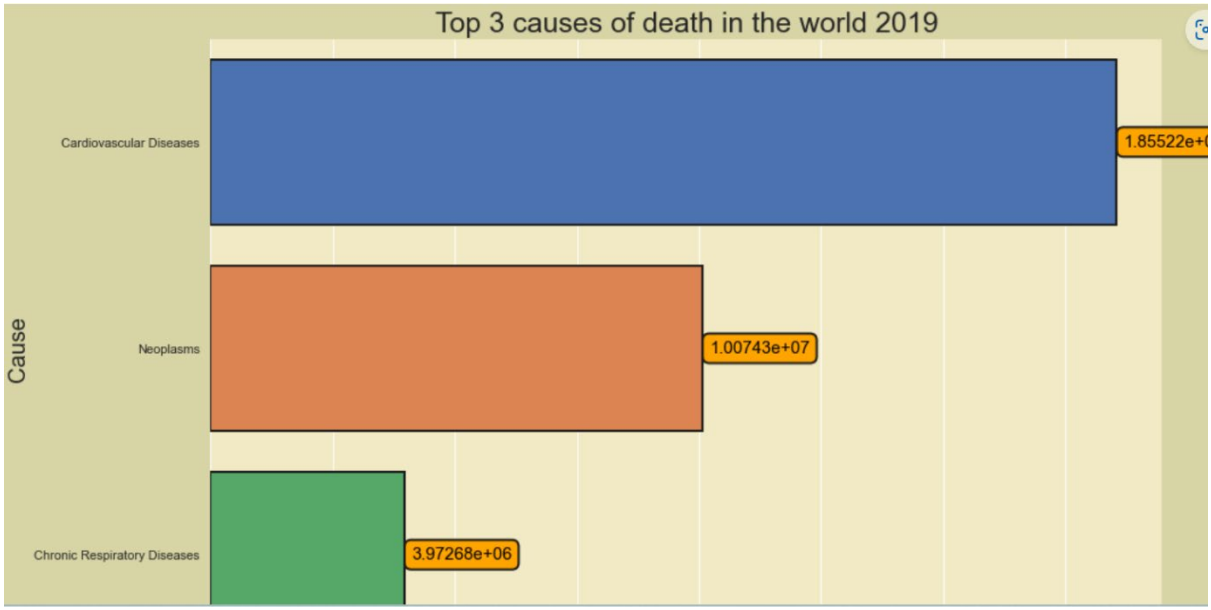
The countries/territories with the highest or lowest number of deaths in 2019 are directly proportional to their population. This is considered reasonable.

Top 5 countries/territories with the highest number of deaths in 2019

Country/Territory	Population	Deads
China	1433783686	10442561
India	1366417754	8812747
United States	328239523	2834964
Russia	145872256	1777223
Indonesia	273523615	1713143

Top 5 countries/territories with the lowest number of deaths in 2019

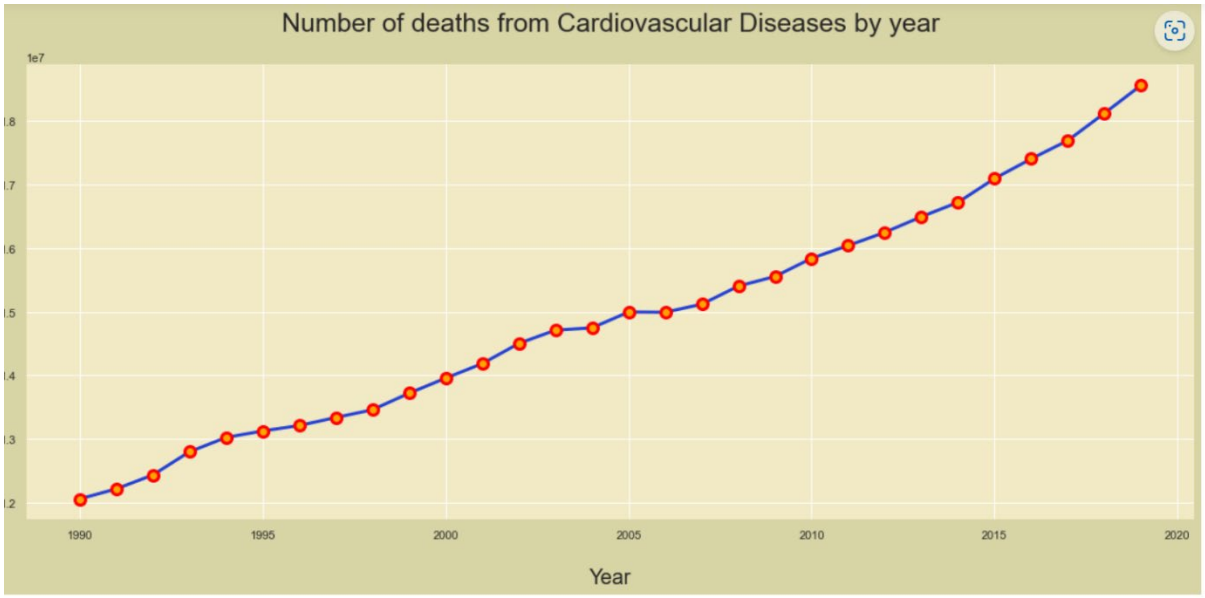
Country/Territory	Population	Deads
Tokelau	1340	9
Niue	1615	18
Nauru	10756	63
Tuvalu	10956	104
Cook Islands	17548	165



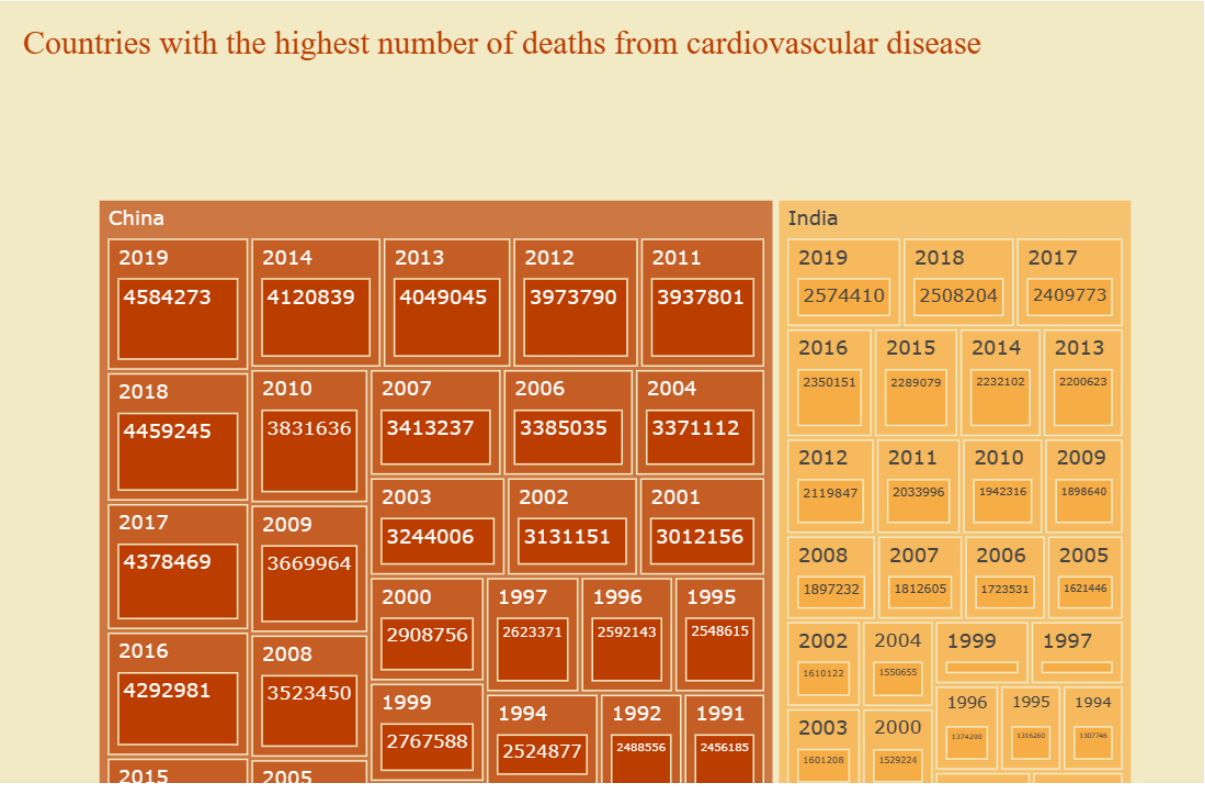
Observations:

In 2019:

The number of deaths from Cardiovascular problems reached 18552218 people, accounting for the highest proportion. Taking the top 2 position is Neoplasms, with 10074275 deaths. Taking the top 3 position is Chronic Respiratory Diseases, with 3972681 deaths.



Top 10 countries with the highest number of deaths from cardiovascular disease over the years



Observations:-

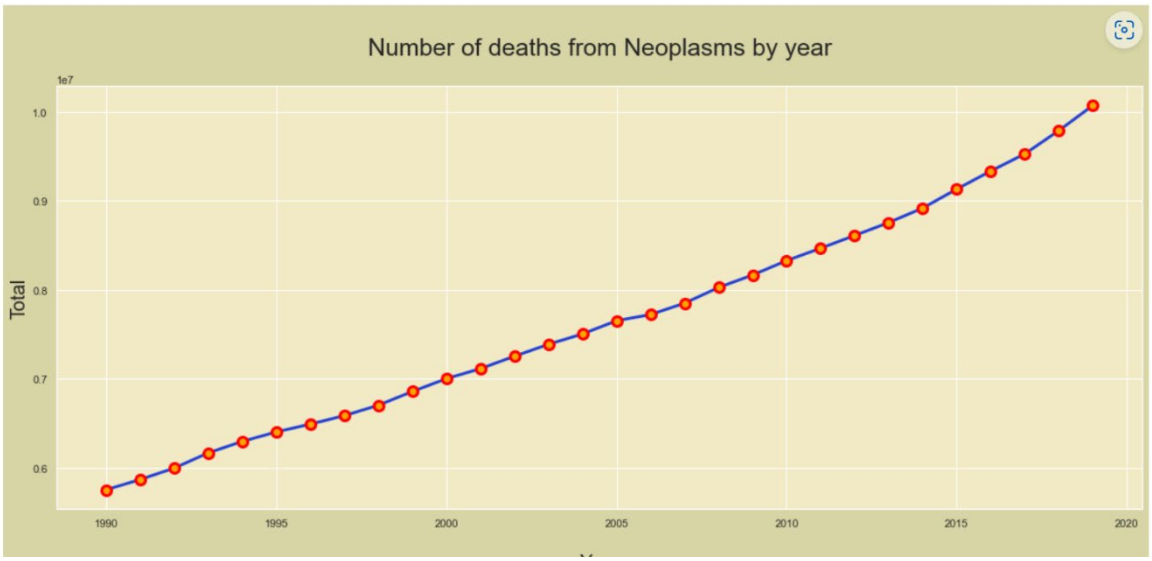
Cardiovascular disease remains the leading burden of disease

The number of deaths related to cardiovascular problems increases year by year, accounting for the highest proportion of all causes. Especially in countries with large populations and developed economies.

According to WHO data, heart disease is the largest cause of death in the world. In which, ischemic heart disease accounted for 16% and stroke accounted for 11% of global deaths. Since 2000, the number of deaths from the disease has increased the most, increasing by more than 2 million to 8.9 million deaths in 2019.

Especially in the current situation of COVID-19 epidemic, the risk of death often focuses mainly on the elderly population, with underlying medical conditions such as hypertension, cardiovascular disease and other chronic diseases. Data from Wuhan (China) show that the mortality rate accounts for 10.5% in people with COVID-19 with heart disease, 7.3% in people with diabetes, 6.3% in people with diabetes. people with respiratory disease and 6% in those with hypertension. On the other hand, worries about the epidemic situation and people's travel restrictions have led

to cardiovascular patients delaying their follow-up visits. This is really dangerous for general chronic illness, which often has no obvious symptoms or signs.



Observations

Neoplasms - Cancer 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.

Below is a list of some of the most common types of cancer:

Lung cancer: accounts for the highest rate in men, including developed and developing countries (accounting for 12.4% of all cancers). Mortality and morbidity rates increase from age 40 and older and peak at age 75. Lung cancer mortality is estimated to be the sum of the four types of colorectal, breast, prostate, and prostate cancers. pancreas.

Stomach cancer: an estimated 934,000 new cancer patients are diagnosed each year. There are many causes of stomach cancer, but an estimated 30% of new cases in developed countries and 47% in developing countries are related to Helicobacter Pylori. Some regions such as Southeast Asia, South America, Eastern Europe... have higher rates of stomach cancer than other regions in the world.

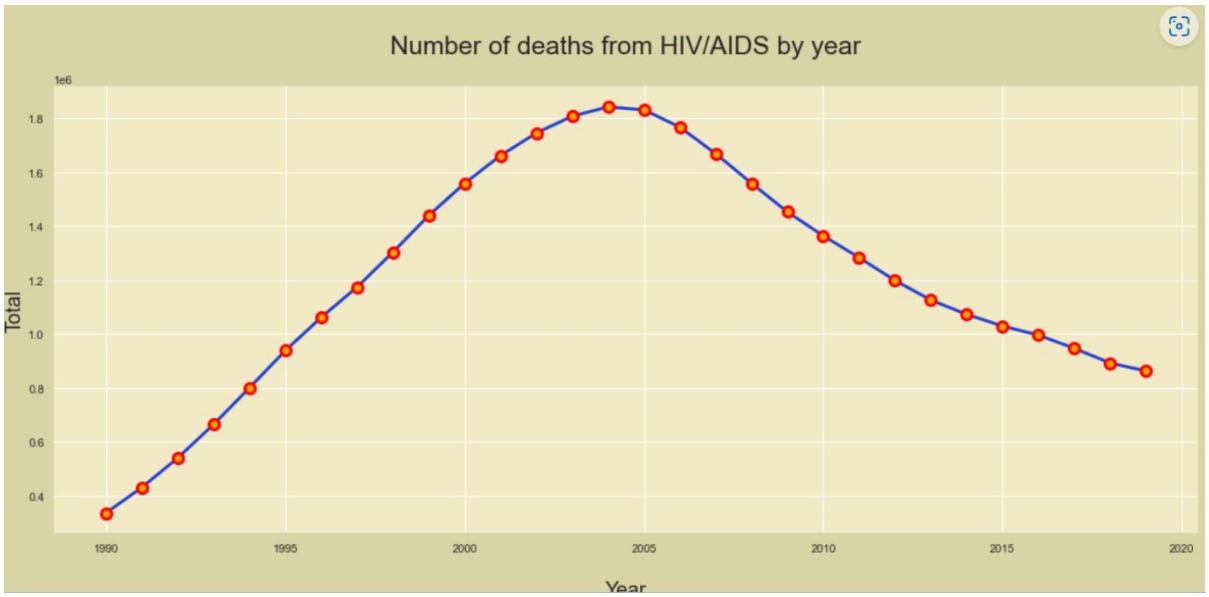
Breast cancer: is the most common cancer in women (accounting for about 23% of all cancers), especially women in developing countries. The rate of breast cancer increases at the age of 50, 60 and peaks at the age of 70. The incidence is expected to be 111 per 100,000 population in the early years of the 21st century.

Colorectal cancer: accounts for about 9.4% of all cancers. This type of cancer is often related to diet and living standards..., the disease is more common in developed countries than in poor countries. The disease has a genetic component. The risk of colorectal cancer is increased in people with a pre-existing history of colitis.

Liver cancer: accounts for about 5.7% of all cancers and is closely related to a history of hepatitis B and C. Asian countries have high rates of liver cancer.

Prostate cancer: the disease is common in the elderly, tends to increase due to the increasing life expectancy. An estimated 679,000 people are newly diagnosed with the disease every year. The prevalence is high in developed countries and lower in developing countries.

Cervical cancer: an estimated 493,000 new cases every year. Human papilloma virus (HPV) is probably one of the risk factors for this disease in poor and developing countries.



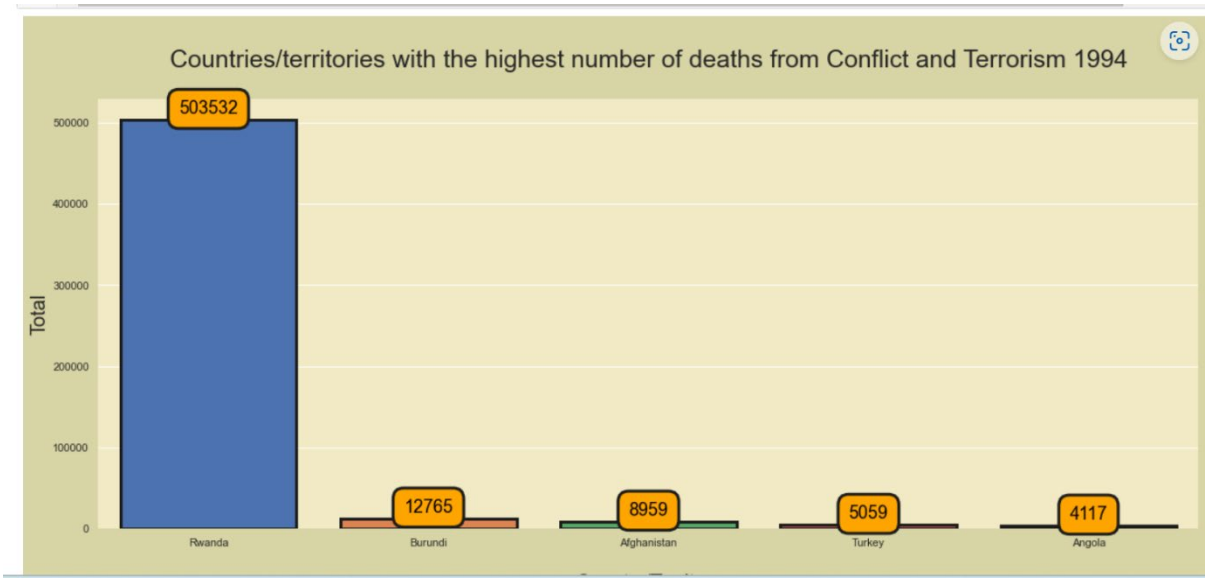
Observation:

HIV/AIDS

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. This is a top public health priority. Studies have shown that, in addition to improving the health of people with HIV, early

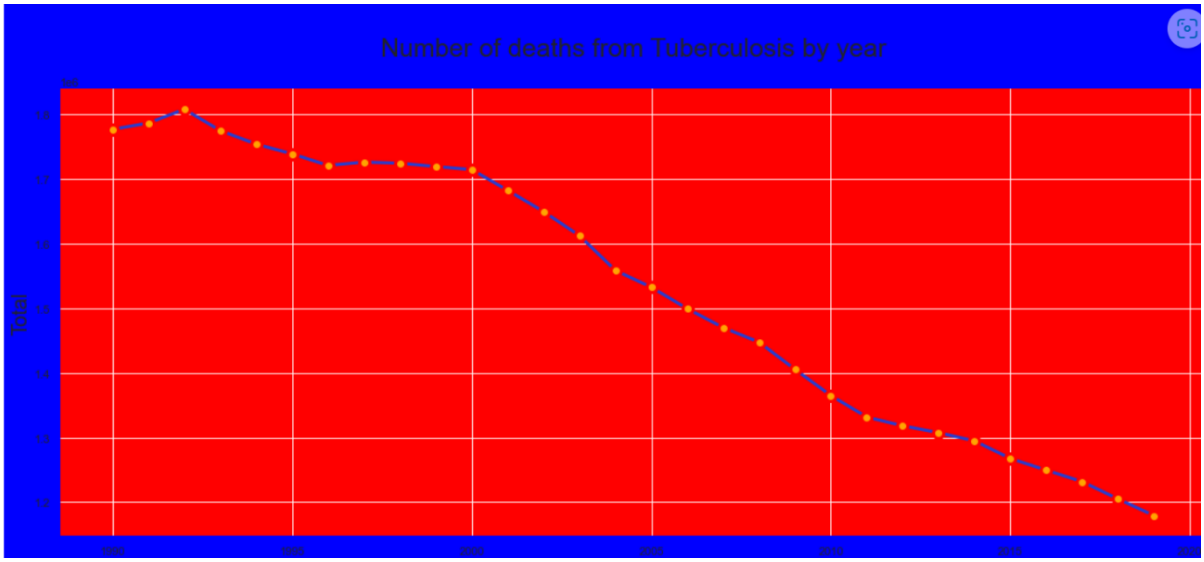
treatment with antiretroviral drugs significantly reduces the risk of transmitting the virus to others.



Observations:

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.



Observations:

Tuberculosis

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. On Sunday, March 24, 1992, a German doctor named Robert Koch announced the discovery of tuberculosis bacteria. At that time TB was ravaging Europe and America with a rate of 1 out of 7 people alive and dying from TB. Thus, on the line graph we see a spike in TB deaths in 1992.

After 19992, ushered in a new era of TB understanding, advances in TB detection, diagnosis and treatment. Looking at the graph, we can see that the number of deaths from TB decreased significantly after 1992.

Software Tools used:

Programming language: Python

3.0Distribution: Anaconda

Navigator

Browser-based language shell: Jupyter Notebook

- Libraries/Packages Used:
- Pandas
- Numpy
- Matplotlib
- Seaborn
- Scipy.stats
- Sklearn

Conclusion:

With the above EDA we understood that 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.

With this EDA, we identified Data of different cause of deaths for all ages around the World.