TOPIC

Cause Of Death

By: Atish Kalangutkar

Data Science Intern: Flip Robo Technologies

Introduction:

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.

Content:

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.

What I have understood about this project:

So this project is based on causes of death because of different type of diseases or factors in respect to the countries and years. So in this dataset 204 countries are there and 30 years data is there that is from year 1990 to year 2019.so each country data of number of deaths from 1990 to 2019 is present in this dataset in respect to different types of diseases.

About Columns/Features:

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

So this are the names of the columns and what information that particular column is providing is given after the name of the columns.

Exploratory Data Analysis:

- So in this project firstly I tried to understand the features. After that I imported the required libraries.
- I used pandas and numpy to load the dataset, also to find shape of the dataset meaning to know the rows and columns present in the data set, than to find the data types of the each of every column, than to find if there are any null values present in the data set, than to check the names of the column,than to find the duplicates, than to find the unique values of a particular column and the count of that unique values in the column.
- Than used seaborn and matploblib to plot the several types of plots for visualization.
- So firstly loaded the data set, and by using data.head() code I checked the overview of the data set.
- Than I checked for the shape of the data set, meaning how many rows and columns are there in data set, So there are 6120 rows and 34 columns.
- After that I checked if there are any null values present in the data set or not as I didn't find any null values present.
- After that I tried to find the data types of the columns.So in that only 32 columns has integer data type and 2 columns has string data type meaning object.

- Than I checked the each and every names of the columns .
- After that I checked if there are any duplicates or not, so it came to my notice that duplicates are not present in the dataset.
- After that I checked for unique values for country/territory, so I found that there are about 204 countries.
- Than I checked for column year, in that I found that 30 years data is there starting from year 1990 to year 2019.
- After that I tried to plot distribution plot for continuous data columns to check the distribution of data and to check whether skewness is present or not.
- So I find out that except column, rest all the column with continuous data type was having skewness.and the data was not distributed normally.
- so after that I plotted count plot as per to country and each diseases so to find out which country has highest number of deaths.
- So after plotting I found that :
- most of the people died from India because of Meningitis.
- most of the people died from China because of Alzheimer's Disease and Other Dementias.
- most of the people died from China because of Parkinson's Disease
- most of the people died from India because of Nutritional Deficiencies.
- most of the people died from Nigeria because of Malaria.
- most of the people died from China because Drowning.
- most of the people died from Brazil because of Interpersonal Violence.
- most of the people died from India because of Maternal Disorders.
- most of the people died from South Africa because of HIV/AIDS.
- most of the people died from China because of Drug use Disorders.
- most of the people died from India because of Tuberculosis.
- most of the people died from China because of Cardiovascular Diseases.

- most of the people died from India because of Lower Respiratory Infections.
- most of the people died from India because of Neonatal Disorders.
- most of the people died from Russia because of Alcohol Use Disorders.
- most of the people died from India because of self harm.
- most of the people died from Haiti because of Exposure to Forces of Nature.
- most of the people died from India because of Diarrheal Diseases.
- most of the people died from Russia because of Environmental Heat and Cold Exposure.
- most of the people died from Rwanda because of Conflict and Terrorism.
- most of the people died from India because of Diabetes Mellitus.
- most of the people died from India because of Chronic Kidney Disease.
- most of the people died from China because of Poisonings.
- most of the people died from India because of Protein-Energy Malnutrition.
- most of the people died from China because of Road Injuries.
- most of the people died from China because of Chronic Respiratory Diseases.
- most of the people died from India because of Cirrhosis and Other Chronic Liver Diseases.
- most of the people died from India because of Digestive Diseases.
- most of the people died from India because of Fire, Heat, and Hot Substances.
- most of the people died from India because of Acute Hepatitis.
- After that i plotted count plot as per to year and each disease so find out whether the number of deaths increased or decreased over a period of 30 years that is from 1990 to 2019, so I found that:
- number of deaths caused due to Meningitis have been decreasing year by year, in 1990 it was more than 2000 and in 2019 it has decreased to less than 1500
- number of deaths caused due to Alzheimer's Disease and Other Dementias have been increasing year by year, in 1990 it was less than 3000 and in 2019 it has increased to more than 6000.

- number of deaths caused due to Parkinson's Disease have been increasing year by year, in 1990 it was less than 1000 and in 2019 it has increased to more than 1500.
- number of deaths caused due to Nutritional Deficiencies have been decreasing year by year, in 1990 it was more than 3500 and in 2019 it has decreased to less than 2000.
- number of deaths caused due to Malaria was the same from year 1990 to 2010, as there was slight fluctuations meaning in some yeas there was slight increase or decrease, and than from year 2011 it has been decreasing.
- number of deaths caused due to Drowning have been decreasing year by year, in 1990 it was more than 2000 and in 2019 it has decreased to less than 1500.
- number of deaths caused due to from year 1990 to Interpersonal Violence have been increasing from 1990 to 1995 and from 1995 to 2019 it is decreasing and increasing slightly.
- number of deaths caused due to Maternal disorders have been decreasing year by year, in 1990 it was close to 1500 and in 2019 it has decreased to less than 1300.
- number of deaths caused due to HIV/AIDS have been increasing drastically from year 1990 to 2004, and from year 2005 it has been decreasing.
- number of deaths caused due to Drug Use Disorders have been increasing year by year, in 1990 it was less than 300 and in 2019 it has increased to more than 500.
- number of deaths caused due to Tuberculosis have been decreasing year by year, in 1990 it was more than 8000 and in 2019 it has decreased to less than 7500.
- number of deaths caused due to Cardiovascular Diseases have been increasing year by year, in 1990 it was less than 60000 and in 2019 it has increased to more than 80000.
- number of deaths caused due to Lower Respiratory Infections have been decreasing year by year, in 1990 it was more than 15000 and in 2019 it has decreased to less than 15000.
- number of deaths caused due to Neonatal Disorders have been decreasing year by year, in 1990 it was more than 14000 and in 2019 it has decreased to less than 12000.
- number of deaths caused due to Alcohol Use Disorders have been increasing from year 1990 to 2005, than from year 2006 to 2013 it was decreasing till year 2014, but again from year 2015 it stared increasing.

- number of deaths caused due to self-harm have been same all this years that is from 1990 to 2019, there are slight fluctuations can be seen.
- number of deaths caused due to Exposure to Forces of Nature were in the years 2004,2008 and 2012 that was more than 1000, in recent years it has been decreasing.
- number of deaths caused due to Diarrheal Diseases have been decreasing year by year, in 1990 it was more than 14000 and in 2019 it has decreased to less than 10000.
- number of deaths caused due to Environmental Heat and Cold Exposure have been increasing and decreasing, from year 1990 to 1995 it was increasing, from year 1996 to 1999 it started decreasing, than again from year 2000 it started increasing till year 2003, and from year 2004 it has been decreasing till 2019,also there are slight fluctuations that is in mid years that is from 2004 to 2019 sometimes it slightly increases.
- number of deaths caused due to Neoplasms been increasing year by year, in 1990 it was less than 30000 and in 2019 it has increased to more than 40000.
- highest number of deaths caused due to Conflict and Terrorism was in year 1994 that was more than 2500,in mid years it was under 500 till year 2011,than in recent years that is from 2012 it again started increasing till 2014 and from year 2015 it is decreasing.
- number of deaths caused due to Diabetes Mellitus have been increasing year by year, in 1990 it was less than 4000 and in 2019 it has increased to more than 6000.
- number of deaths caused due to Chronic Kidney Disease have been increasing year by year, in 1990 it was less than 4000 and in 2019 it has increased to more than 6000.
- number of deaths caused due to Poisonings have been increasing and decreasing, that is from year 1990 to 1994 it was increasing than from year 1994 to 1998 is started decreasing than again from year 1998 to 2005 it started increasing and from year 2005 to 2019 it is decreasing.
- number of deaths caused due to Protein-Energy Malnutrition have been decreasing year by year, in 1990 it was more than 3000 and in 2019 it has decreased to less than 2000.
- number of deaths caused due to Road Injuries have been increasing and decreasing, that is from year 1990 to 2008 it was increasing slightly and from year 2008 to 2019 it have been decreasing.

- number of deaths caused due to Chronic Respiratory Diseases have been increasing year by year, in 1990 it was 15000 and in 2019 it has increased to more than 16000.
- number of deaths caused due to Cirrhosis and Other Chronic Liver Diseases have been increasing year by year, in 1990 it was less than 5000 and in 2019 it has increased to more than 6000.
- number of deaths caused due to Digestive Diseases have been increasing year by year, in 1990 it was less than 9000 and in 2019 it has increased to more than 11000.
- number of deaths caused due to Fire, Heat, and Hot Substances have been the same that is from year 1990 to 2019, little bit fluctuations are there but it looks same.
- number of deaths caused due to Acute Hepatitis have been decreasing year by year, in 1990 it was more than 800 and in 2019 it has decreased to less than 500.
- After that I added one more column to the dataset that is total number of deaths according to the year and country.
- After that I tried to find out the highest number of deaths according to year and country.
- So after doing some analysis I came to know that highest number of deaths were reported in year 2019 in China.
- Most of the deaths as per to year wise reported in China only and than followed by India.
- So in top 50, in-terms of total number of deaths only two countries names come into picture that is china and India.
- We can see the names of India and china may be because of high population.
- After that I wanted to check the highest number of deaths according to the diseases year wise.
- So I came to know that:
- From year 1990 to year 2019 highest number of deaths were caused due to Cardiovascular Diseases.
- Since when I tried to find out the top 50 number of deaths according to country and year that time find out that in top 50 only two countries were there,they were China and India.
- So I wanted to check in China and India the diseases which is causing highest number of deaths in this countries.
- So after doing some analysis I found that highest number of deaths in China and India were caused to Cardiovascular Diseases.

- So after that I wanted to check the top 3 countires with highest number of deaths across all 30 years.
- So after doing analysis I found that highest number of deaths were reported in country:

  1:China

  2:India

  3:United States
- So after that I wanted to check the top 10 causes of deaths in each of this countries that is China,India and United States.

- So in china top 10 causes of deaths were:

  1:cardiovascular Diseases

  2:Neoplasms

  3:Chronic Respiratory Diseases

  4:Digestive Diseases

  5:Lower Respiratory Infections

  6:Road Injuries

  7:Alzheimer's Disease and Other Dementias

  8:Self-harm

  9:Cirrhosis and Other Chronic Liver Diseases

  10:Neonatal Disorders

- So in India top 10 causes of deaths were:

  1:Cardiovascular Diseases

  2:Diarrheal Diseases

  3: Chronic Respiratory Diseases

  4: Neonatal Disorders

  5:Neoplasms

  6: Lower Respiratory Infections

  7:Tuberculosis

  8: Digestive Diseases

  9: Cirrhosis and Other Chronic Liver Diseases

  10: Self-harm

- So in United States top 10 causes of deaths were:

  1:Cardiovascular Diseases

  2: Neoplasms

  3:Chronic Respiratory Diseases

  4: Alzheimer's Disease and Other Dementias

  5:Digestive Diseases

  6:Lower Respiratory Infections

  7:Diabetes Mellitus

  8:Chronic Kidney Disease

  9: Cirrhosis and Other Chronic Liver Diseases

  10:Road Injuries

- So after that I deleted the unwanted column that is code which depicted the 3 words cod e of the particular country.
- After that I used label encoder to convert column Country/territory which was having obj ect data type into integer type.
- After that I checked the correlation of all columns with column years.
- So I found out that Alzheimer's Disease and Other Dementias,Parkinson's Disease,Diabet es Mellitus and Chronic Kidney Disease are highly positively correlated to the year.where as Nutritional Deficiencies and Protein-Energy Malnutrition are highly negatively correlat ed to the year.
- After that I plotted heat map to check if there is any multicollinearity problem between t he columns.
- So in that I found that that most of diseases are multi-correlated with each other,so will be checking by using vif.
- After that I checked if there are any outliers present or not in the continuous data colum ns,so it came into my notice that except column year rest all the columns were having ou tliers.so it deleted the outliers by using z-score method.
- Than I checked if there is any skewness present or not,so it came into my notice that exc ept column year rest all the columns was having skewness,so I treated the skewness by u sing power transform method.