Report on Cause of Death of Persons From All Over The World From 1990 to 2019

By: Satu Vinayak Pole

Intern - Flip Robo Technologies

**About Dataset**

Context

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, Diarrhoeal 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, and finally Acute Hepatitis.

These features mentioned above provides number of persons who suffered/died from this diseases or events with respect to the different countries from all over the world. The data generated or provided to us is from years 1990-2019 i.e. for 30 years.

Exploratory Analysis for data provided:

As mentioned above the dataset consists of 34 features and without a label (dependant variable) suggesting that this is Unsupervised machine learning problem. The dataset consists of names of countries, unique abbreviations for countries, years from 1990-2019 and names of diseases and/or factors due to which deaths occurred. This features/columns were provided with number of suffered this outcomes.

All the data is continuous in nature. The dataset was evaluated using Jupyter notebook using panda, numpy, seaborn and matplotlib libraries.

The shape of the dataset is 6120 rows and 34 columns. This dataset as mentioned earlier does not contain label column. The 6120 rows contains data of 30 years for 1 country. This means that by calculating we can say that the data provided to us is of 204 countries/territories.

The dataset was further checked for null values. It seems that while making this data file no fields were left empty meaning nu null values were found.

Furthermore the dataset was checked if there were any duplicate values and it was found that dataset did not contain any duplicates.

The barplot graphs were then plotted for Data of diseases/instances that caused with respect to 30 years. The following are the observations that are seen after plotting graphs.

1. plot of Year with respect to Meningities disease suggests that number of deaths was highest in 1990 and subsequently decreased in 2019.
2. plot of Year with respect to Alzheimer's Disease and Other Dementias disease suggests that number of deaths was lowest in 1990 and subsequently increased in 2019.
3. plot of Year with respect to Parkinson's Disease suggests that number of deaths was lowest in 1990 and subsequently increased in 2019
4. plot of Year with respect to Nutritional Deficiencies suggests that number of deaths was highest in 1990 and subsequently decreased in 2019.with slight increase in 1995 and then moving down.
5. plot of Year with respect to Malaria suggests that number of deaths were almost same from 1990 to 2001. Then deaths increased slightly until 2010 and then it started lowering with lowest in 2019.
6. plot of Year with respect to Drowning suggests that number of deaths was highest in 1990 and subsequently decreased in 2019 implying less floods/drowning suicides by people in recent years.
7. plot of Year with respect to Interpersonal Violence suggests that number of deaths was lowest in 1990 and subsequently started increasing with almost same for all this years.
8. plot of Year with respect to Maternal Disorders suggests that number of deaths was highest in 1990 and subsequently started decreasing with least deaths in 2019.
9. plot of Year with respect to HIV/AIDS suggests that number of deaths was lowest in 1990 and subsequently started increasing until 2004(highest deaths).Afterwards the deaths from HIV started decreasing with least in 2019.
10. plot of Year with respect to Drug Use Disorders suggests that number of deaths was lowest in 1990 and subsequently started increasing with highest deaths in 2019.
11. plot of Year with respect to Tuberculosis suggests that number of deaths was highest in 1990 and subsequently started decreasing with lowest deaths in 2019.
12. plot of Year with respect to Cardiovascular Diseases suggests that number of deaths was lowest in 1990 and subsequently started increasing with highest deaths in 2019.
13. plot of Year with respect to Lower Respiratory Infections suggests that number of deaths was highest in 1990 and subsequently started decreasing with lowest deaths in 2019.
14. plot of Year with respect to Neonatal Disorders suggests that number of deaths was highest in 1990 and subsequently started decreasing with least deaths in 2019.
15. plot of Year with respect to Alcohol Use Disorders suggests that number of deaths was lowest in 1990 and subsequently started increasing with highest deaths in 2019.
16. plot of Year with respect to Self-harm suggests that number of deaths are almost same from 1990- 2019.
17. plot of Year with respect to Exposure to Forces of Nature suggests that number of deaths were highest in 2004,2008 and 2010 and then kept on decreasing with reaching minima in 2019.
18. plot of Year with respect to Diarrheal Diseases suggests that number of deaths was highest in 1990 and subsequently started decreasing with least deaths in 2019.
19. plot of Year with respect to Environmental Heat and Cold Exposure suggests that there is not trend seen with respect to deaths noted.
20. plot of Year with respect to Neoplasms suggests that number of deaths was lowest in 1990 and subsequently started increasing with highest deaths in 2019.
21. plot of Year with respect to Conflict and Terrorism suggests that number of deaths have been lowest from 1990-2019 with average deaths of 500 people dying but in 1994 deaths were reported for >2500 people.
22. plot of Year with respect to Diabetes Mellitus suggests that number of deaths was lowest in 1990 and subsequently started increasing with highest deaths in 2019.
23. plot of Year with respect to Chronic Kidney Disease suggests that number of deaths was lowest in 1990 and subsequently started increasing with highest deaths in 2019.
24. plot of Year with respect to Poisonings suggests that number of deaths havee been at an average pf 400 people yearwise for all these years from 1990-2019.
25. plot of Year with respect to Protein-Energy Malnutrition suggests that number of deaths was highest in 1990 and subsequently started decreasing with least deaths in 2019.
26. plot of Year with respect to Road Injuries suggests that number of deaths have been at an average of >5000 people yearwise for all these years from 1990-2019.
27. plot of Year with respect to Chronic Respiratory Diseases suggests that number of deaths was lowest in 1990 and subsequently started increasing with highest deaths in 2019.
28. plot of Year with respect to Cirrhosis and Other Chronic Liver Diseases suggests that number of deaths was lowest in 1990 and subsequently started increasing with highest deaths in 2019.
29. plot of Year with respect to Digestive Diseases suggests that number of deaths was lowest in 1990 and subsequently started increasing with highest deaths in 2019.
30. plot of Year with respect to Fire, Heat, and Hot Substances suggests that number of deaths have been at an average of 600 people yearwise for all these years from 1990-2019.
31. plot of Year with respect to Chronic Kidney Disease suggests that number of deaths were highest in 1990 and subsequently started decreasing with least deaths in 2019.

The above plots showed us the frequency of deaths that happened in a span of 30 years. For ex. Incase of Diabetes mellitus, the deaths/morbidity due to Diabetes mellitus was low in 1990 compared to in 2019. One reason according to me can be eating of junk food/oily food which people avoided or people did not have access to.But as soon as people were familiar with junk food, the cases for diabetes mellitus kept on increasing.

Further the deaths/morbid features were plotted against country/territories. This provided the data such as which country had the most cases/deaths with respect to specific diseases mentioned.

Further I thought that abbreviations column for countries names having column name as ‘Code’ was dropped as I think that country name and its abbreviations are both same, hence keeping one of them was a good choice. Further the country/territory column was label encoded as it was composed of string data.

The data was described which using data.describe() method which provided us with mean of features,min and max alongwith interquantile ranges of datapoints. After seeing the interquantile ranges it was known that data contained outliers and also the data is skewed.

After describing the data, a heatmap for dataset was plotted which showed us that many features were showing multicollinearity. For ex. Cardivascular disease is related to Alzheimer disease,Parkinson's disease and Drowning.

Further data was grouped for year with respect to disease/features and their sum was calculated. This code data provided us with the total number of deaths that occurred due to different diseases/outcomes in that particular year.

The data was further analysed and boxplots were plotted for all columns. This provided us with idea that all features have outliers present in them. Next, all the outlier present were removed using Z-Score method. After removing outliers it was seen that 8.81% of data was lost as they were outside the range of zscore of +-3.

Further it was seen that all the features have skewness values more than +-0.5. To confirm this skewness displots were plotted for all the 33 features and was confirmed that most of the features were right skewed. Only Year and Country name columns had skewness values within range. Since this skewness had to be lowered down within the specified range, Power Transformer was applied using yeo-johnson method. After applying this it was seen that skewness was lowered and was seen that it is within range of +-0.5.

This model is a kind of Unsupervised Machine learning model since label/dependant variable is not provided. Hence further model building will be done.