Cause of Deaths

Submitted by:

Mansi Nagpal

ACKNOWLEDGMENT

I'd like to extend my gratitude to my mentor Ms. Khushboo Garg for giving me this opportunityto work upon this project. Below are all the details of the project which I consumed while preparing and drafting the project –

The references, research papers, data sources, professionals, etc are majorly referred from "Flip Robo Technologies" study collaterals and data repository. Also, some of the other resources like "Research Gate" were explored to gain a deep understanding on the assigned subject.

Above stated details stands correct to the best of my knowledge and I hereby acknowledge the same.

INTRODUCTION

• Business Problem Framing

Many of the scientific discussions and studies in biomedical and healthcare domains address tasks whose end goal is to prevent death or diseases. Since the emergence of the big data science, numerous machine learning based techniques and technologies have been proposed and applied to improve human health by solving different computational challenges that we face today. A less obvious question, that remains to be extensively explored by researchers, is whether Big Data science can contribute to our understanding of factors leading to death or diseases, via analysis of multiple-cause mortality data. In fact it is widely believed that counting the dead is a significant investment to reduce the premature mortality [2]. There have been a number of studies that have proven to offer profound impacts on our understanding of the major causes of death using the statistical analysis of recorded death data. In light of these studies, we were interested to investigate the feasibility of this emerging field in learning hidden complex patterns that are available in the haystack of mortality datasets.

• Conceptual Background of the Domain Problem

Data science comes as a vital tool to analyze the data where we know the exact ratio of cause of death & where the ratio is too high and what is the reason here. Moreover, it also help to know the real cause and gave answers of all these question

Predictive modelling, Market mix modelling, recommendation systems are some of the machine learning techniques globally used for achieving the reasons of cause of deaths actually.

Currently the assigned project utilises Predictive Modelling algorithm to solve the business statement.

- Review of Literature

The research project was for a Cause of death based on 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.

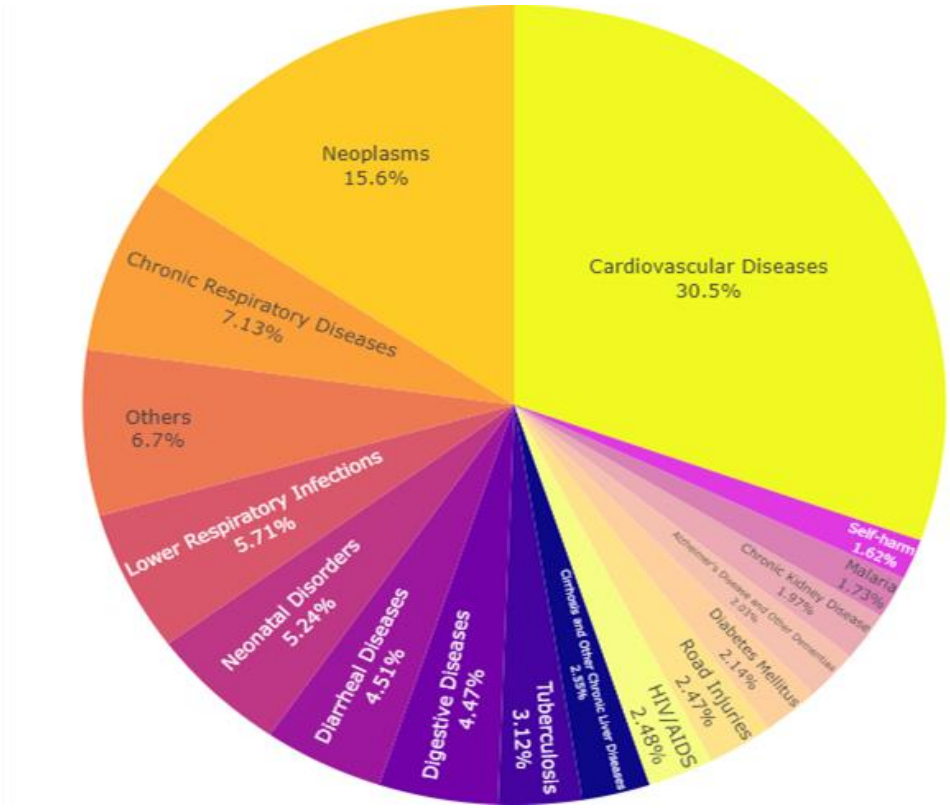
Being a Data Scientist, I have used the RegressionModel using multiple algorithms to design and optimise the results.

Some of the classes which I explored from Scikit-learn libraries were – Statistics, Analytical Modelling, Hyper Tuning Method, CV Ratings, Predictive Modelling and Regression Model, etc.

In totality, my project comprises of different test cases or regression models where the objective was to train for accuracy and test for accuracy using distribution plot to best understand the linearity of clusters.

Data provided through the Column basis where we need to analyze 32 different columns of different disease.

As a result of the above analytical modelling, I have managed to achieve below percentages of causes, which is –



- Motivation for the Problem Undertaken

Despite the importance of the subject, only a handful of studies have so far conducted research seeking to relate multiple causes of death to each other or to other factors. These studies are often limited to classical statistical. 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.

This project guided me to baseline each aspect carefully and be concrete on the decision-making process regardless it's an individual or an entity like an organization.

In order to cater to the above project, my current knowledge and skill set has aided me a lot which I'd explore on this exponentially further.

- Data Sources and their formats

Here the data set is all about related to different countries and Territory. Here 32 types of different cause we covered accordingly.

Here I can show you the exact analysis which I used for it:-

- Exploratory Data Analysis

01. Country/Territory

```
In [9]: df['Country/Territory'].describe()

Out[9]: count                6120
        unique                204
        top           Afghanistan
        freq                   30
        Name: Country/Territory, dtype: object
```

. Country/Territory contain nominal data in text format

```
In [11]: #checking unique of variable
         print(df['Country/Territory'].unique())
         #counting the uniques
         print(df['Country/Territory'].value_counts())
```

```
['Afghanistan' 'Albania' 'Algeria' 'American Samoa' 'Andorra' 'Angola'
 'Antigua and Barbuda' 'Argentina' 'Armenia' 'Australia' 'Austria'
 'Azerbaijan' 'Bahamas' 'Bahrain' 'Bangladesh' 'Barbados' 'Belarus'
 'Belgium' 'Belize' 'Benin' 'Bermuda' 'Bhutan' 'Bolivia'
 'Bosnia and Herzegovina' 'Botswana' 'Brazil' 'Brunei' 'Bulgaria'
 'Burkina Faso' 'Burundi' 'Cambodia' 'Cameroon' 'Canada' 'Cape Verde'
 'Central African Republic' 'Chad' 'Chile' 'China' 'Colombia' 'Comoros'
 'Congo' 'Cook Islands' 'Costa Rica' "Cote d'Ivoire" 'Croatia' 'Cuba'
 'Cyprus' 'Czechia' 'Democratic Republic of Congo' 'Denmark' 'Djibouti'
 'Dominica' 'Dominican Republic' 'Ecuador' 'Egypt' 'El Salvador'
 'Equatorial Guinea' 'Eritrea' 'Estonia' 'Eswatini' 'Ethiopia' 'Fiji'
 'Finland' 'France' 'Gabon' 'Gambia' 'Georgia' 'Germany' 'Ghana' 'Greece'
 'Greenland' 'Grenada' 'Guam' 'Guatemala' 'Guinea' 'Guinea-Bissau'
 'Guyana' 'Haiti' 'Honduras' 'Hungary' 'Iceland' 'India' 'Indonesia'
 'Iran' 'Iraq' 'Ireland' 'Israel' 'Italy' 'Jamaica' 'Japan' 'Jordan'
 'Kazakhstan' 'Kenya' 'Kiribati' 'Kuwait' 'Kyrgyzstan' 'Laos' 'Latvia'
 'Lebanon' 'Lesotho' 'Liberia' 'Libya' 'Lithuania' 'Luxembourg'
 'Madagascar' 'Malawi' 'Malaysia' 'Maldives' 'Mali' 'Malta'
 'Marshall Islands' 'Mauritania' 'Mauritius' 'Mexico' 'Micronesia'
 'Moldova' 'Monaco' 'Mongolia' 'Montenegro' 'Morocco' 'Mozambique'
 'Myanmar' 'Namibia' 'Nauru' 'Nepal' 'Netherlands' 'New Zealand'
 'Nicaragua' 'Niger' 'Nigeria' 'Niue' 'North Korea' 'North Macedonia'
 'Northern Mariana Islands' 'Norway' 'Oman' 'Pakistan' 'Palau' 'Palestine'
 'Panama' 'Papua New Guinea' 'Paraguay' 'Peru' 'Philippines' 'Poland'
 'Portugal' 'Puerto Rico' 'Qatar' 'Romania' 'Russia' 'Rwanda'
 'Saint Kitts and Nevis' 'Saint Lucia' 'Saint Vincent and the Grenadines'
 'Samoa' 'San Marino' 'Sao Tome and Principe' 'Saudi Arabia' 'Senegal'
 'Serbia' 'Seychelles' 'Sierra Leone' 'Singapore' 'Slovakia' 'Slovenia'
 'Solomon Islands' 'Somalia' 'South Africa' 'South Korea' 'South Sudan'
 'Spain' 'Sri Lanka' 'Sudan' 'Suriname' 'Sweden' 'Switzerland' 'Syria'
 'Taiwan' 'Tajikistan' 'Tanzania' 'Thailand' 'Timor' 'Togo' 'Tokelau'
 'Tonga' 'Trinidad and Tobago' 'Tunisia' 'Turkey' 'Turkmenistan' 'Tuvalu'
 'Uganda' 'Ukraine' 'United Arab Emirates' 'United Kingdom'
 'United States' 'United States Virgin Islands' 'Uruguay' 'Uzbekistan'
 'Vanuatu' 'Venezuela' 'Vietnam' 'Yemen' 'Zambia' 'Zimbabwe']
Afghanistan         30
Papua New Guinea    30
Niue                30
North Korea         30
North Macedonia     30
                    ..
Greenland           30
Grenada             30
Guam                30
Guatemala           30
Zimbabwe            30
Name: Country/Territory, Length: 204, dtype: int64
```

In the year column contains ordinal data it is equally distrubuted we have data from year 1990 to 2019 = 30 year of death records we have

04. Meningitis - No. of People died from Meningitis

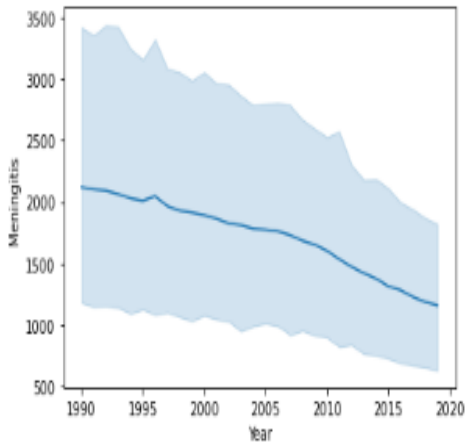
No.of People died from Meningitis

```
In [13]: df['Meningitis'].describe()
```

```
Out[13]: count     6120.000000
         mean      1719.701307
         std       6672.006930
         min          0.000000
         25%         15.000000
         50%        109.000000
         75%        847.250000
         max      98358.000000
         Name: Meningitis, dtype: float64
```

```
In [14]: sns.lineplot(data=df, x="Year", y="Meningitis")
```

```
Out[14]: <AxesSubplot:xlabel='Year', ylabel='Meningitis'>
```



. Year- Year of the Incident

```
In [12]: df['Year'].describe()
```

```
Out[12]: count    6120.000000
         mean     2004.500000
         std         8.656149
         min      1990.000000
         25%      1997.000000
         50%      2004.500000
         75%      2012.000000
         max      2019.000000
         Name: Year, dtype: float64
```

- Data Preprocessing Done

I’ve dropped the attributes from the datasets as these columns have Meningitides almost 64 float number in respective columns and these are 'Alley','PoolQC','Fence','MiscFeature'.

Replaced all the nan values of the numerical columns and the categorical columns with mean and most_frequent(mode) value of the respective columns respectively.
Use Label Encoder to convert all the categorical variables value into numeric form the datasets.

Dropped all the feature variables those were giving the same amount of information to the target variable.

I haven’t remove outliers as almost maximum attributes are consisting of outliers and if I consider the same the dataset could be decrease and our prediction can’t be good enough then.

Hardware and Software Requirements and Tools Used

Listing down the hardware and software requirements along with the tools, libraries and packages used. Describe all the software tools used along with a detailed description of tasks done with those tools.

I have used Python IDE (Integrated Development environment) as a dedicated software throughout solving this project.

Python Libraries that I've used throughout the process are-

- Numpy- It is use for linear algebra
- Pandas- data analysis/manipulation library
- Scipy- Utility function for optimization
- Matplotlib-Data visualization and plotting library
- Seaborn- Data visualization and statistical plotting library
- Sklearn- Machine Learning Tool
- Imblearn- Deal with classification problems of Imbalanced classes
- Statsmodels-Deal with advanced statistics

Classes-

- Label Encoder-Encoding the categorical variables into numbercategory
- Simple Imputer-Replacing the null values with mean,median or mode
- variance_inflation_factor-Calculate multicollinearity
- Power Transformer-remove skewness
- StandardScaler-normalize the feature variables
- Principle Component Analysis- Reduce the dimension of the dataframe
- Cross_val_score- CV score
- GridSearchCV- Find out the best parameters for the model

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

Statistical method-
When I use describe function then I find out that most of the variables mean are less than that of median and the interquartile range b/w the variables are varying too much and it shows that datasets are skewing left hand side and it indicates that the variables are not normally distributed.

Also, I have used correlation method to check what are the variables that are giving strong correlation w.r.t Target variable 'Sale Price'.
Analytical Method-

I've uses boxplot and distribution plots to check the outliers and skewness of the variables respectively through the plotting's.

- Testing of Identified Approaches (Algorithms)

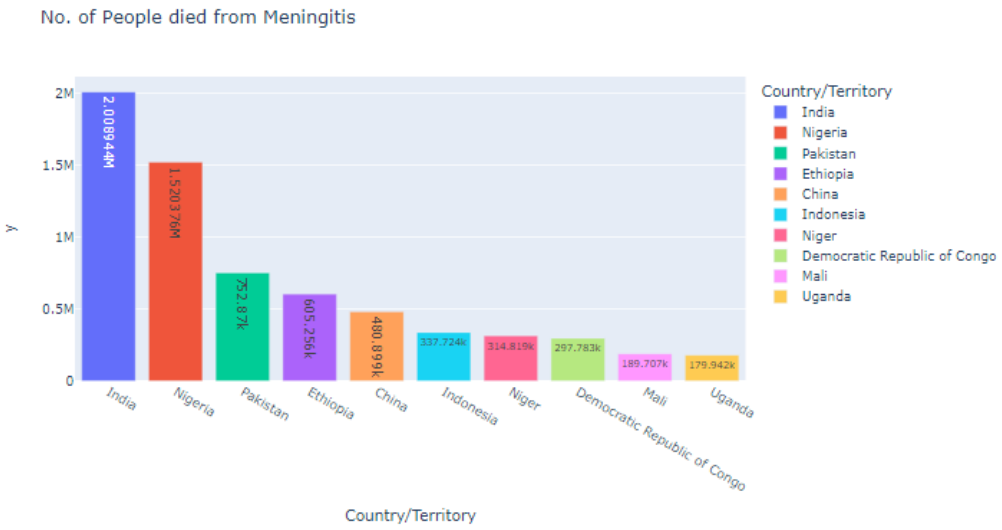
Listing down all the algorithms used for the training and testing.

```
Lr=LinearRegression()
ls=Lasso()
rd=Ridge()
en=ElasticNet()
sgd=SGDRegressor
()
rf=RandomForestRegressor()
ad=AdaBoostRegressor()
grd=GradientBoostingRegressor(
)
```

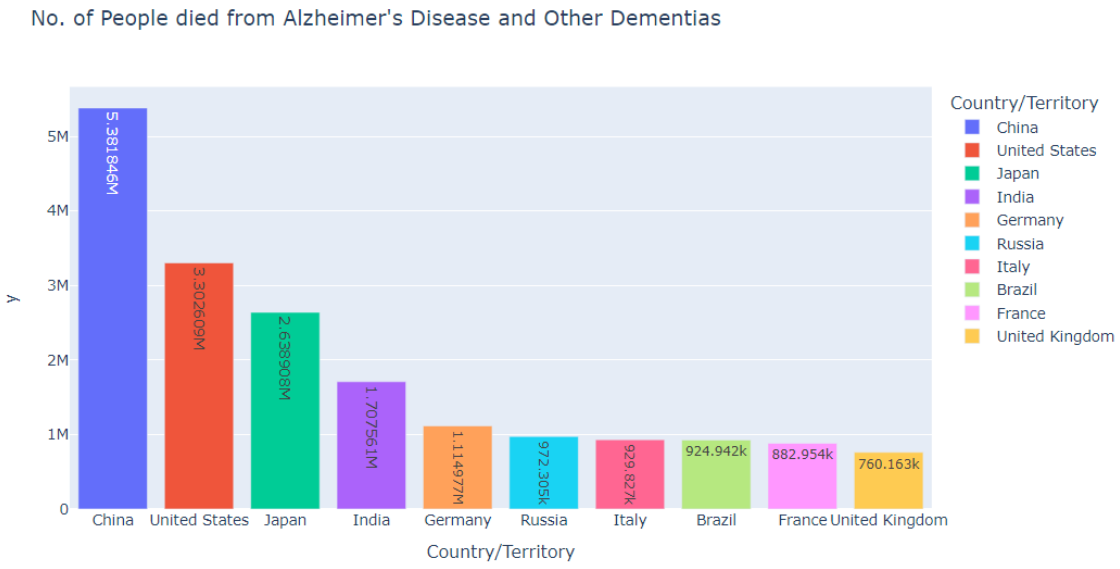
- # Run and Evaluate selected models

Describe all the algorithms used along with the snapshot of their code and what were the results observed over different evaluation metrics.

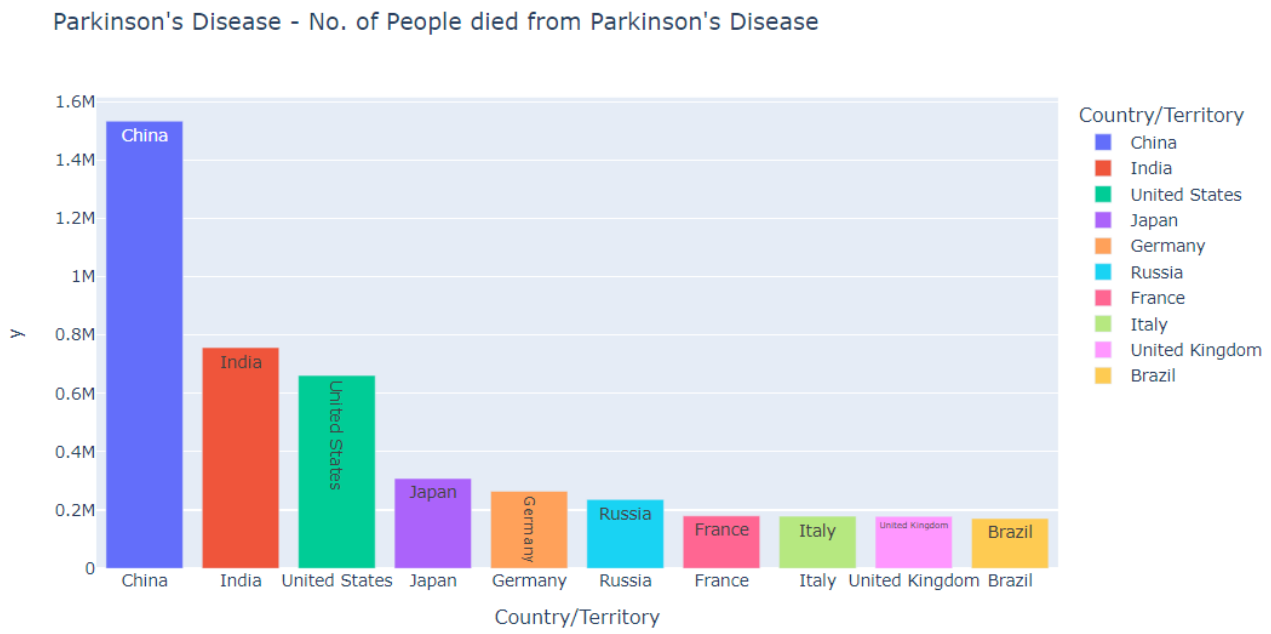
Meningitis - No. of People died from Meningitis



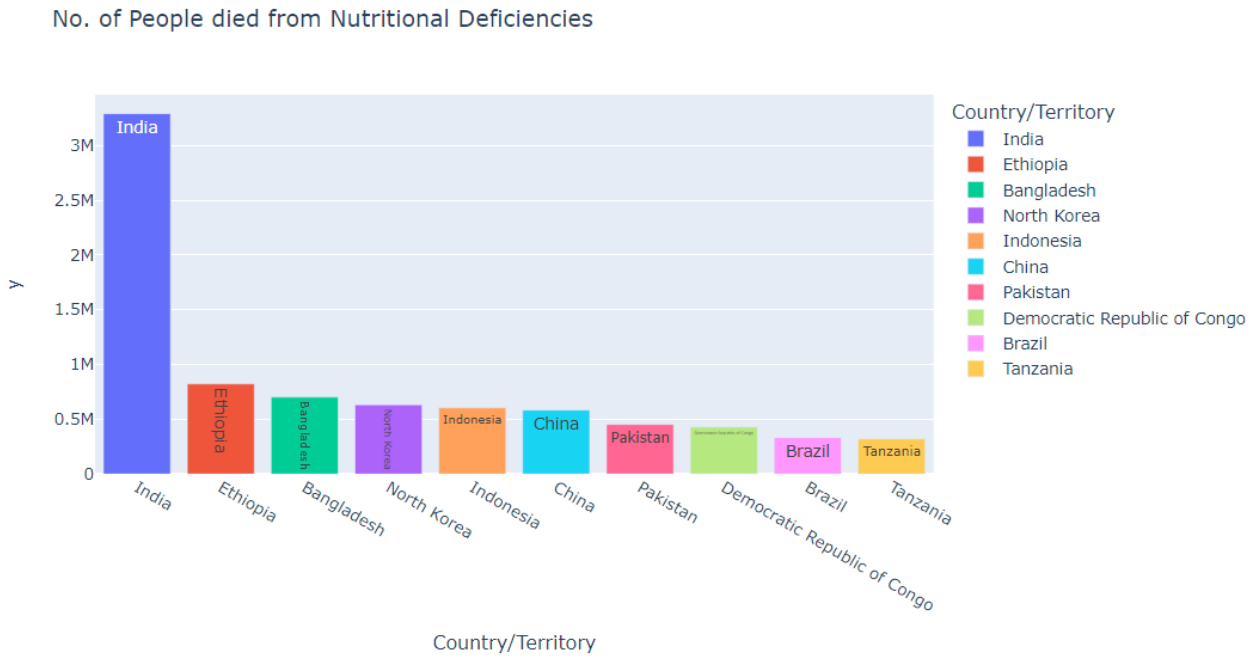
Alzheimer's Disease and Other Dementias - No. of People died from Alzheimer's Disease and Other Dementias



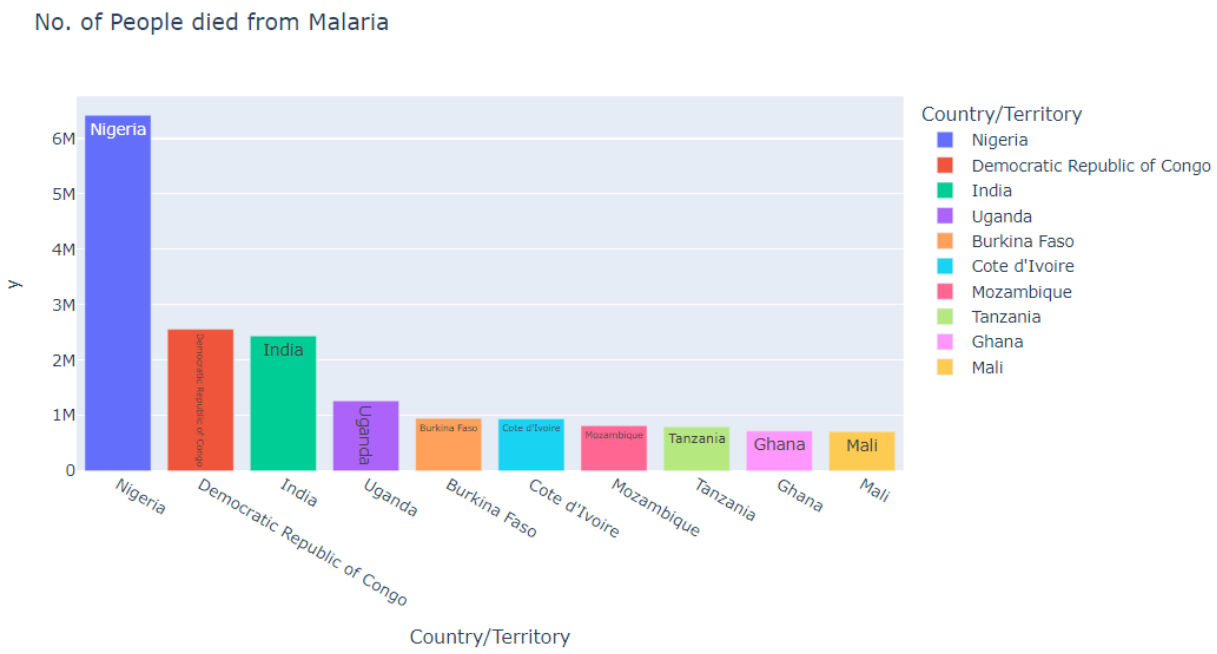
Parkinson's Disease - No. of People died from Parkinson's Disease



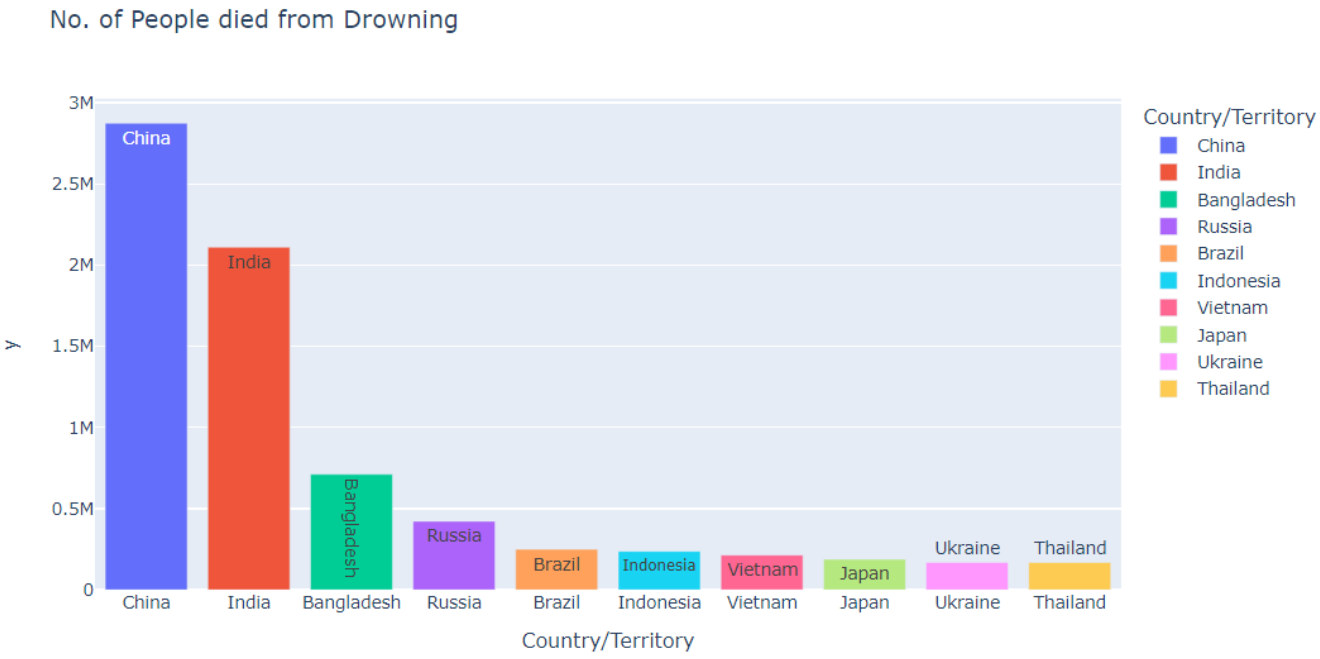
Nutritional Deficiencies - No. of People died from Nutritional Deficiencies



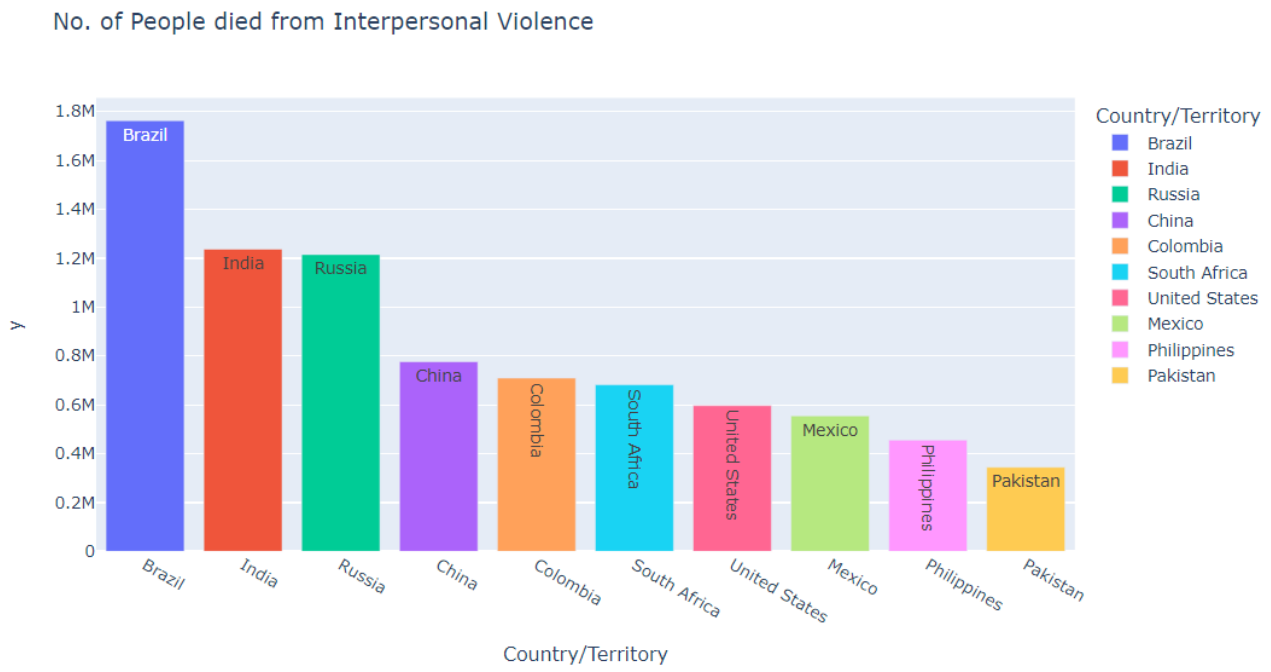
Malaria - No. of People died from Malaria



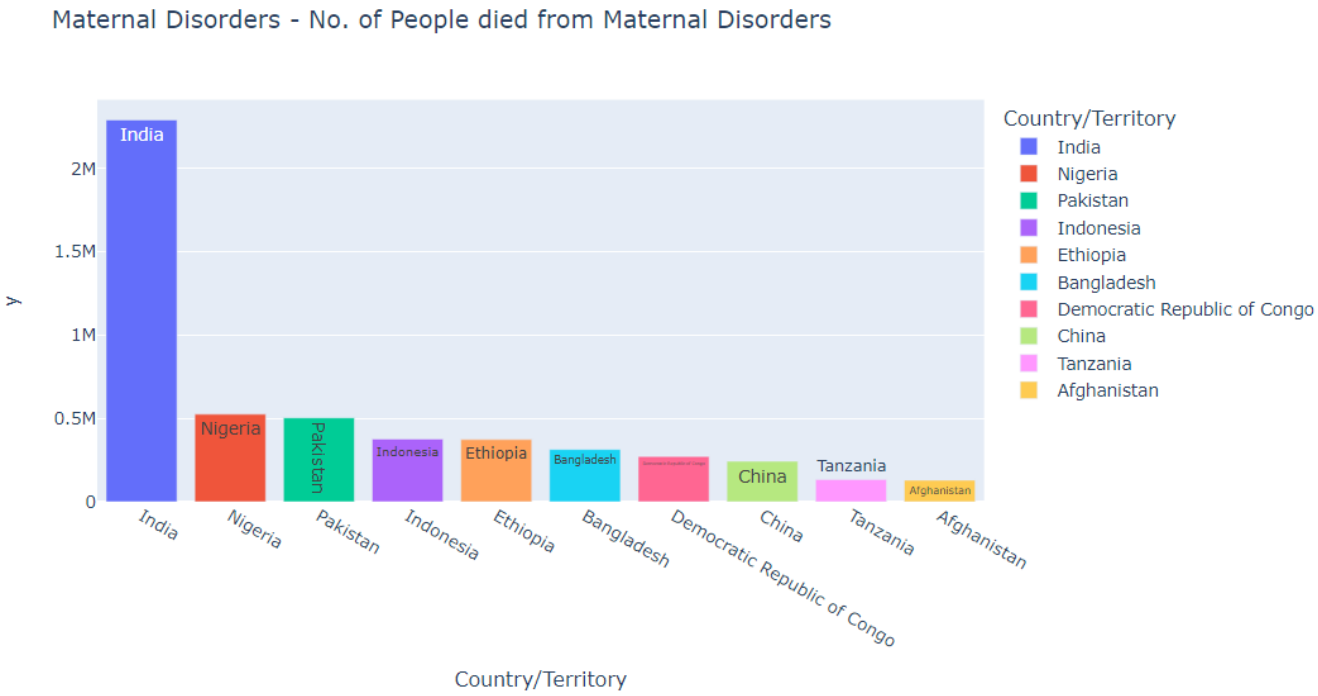
Drowning - No. of People died from Drowning



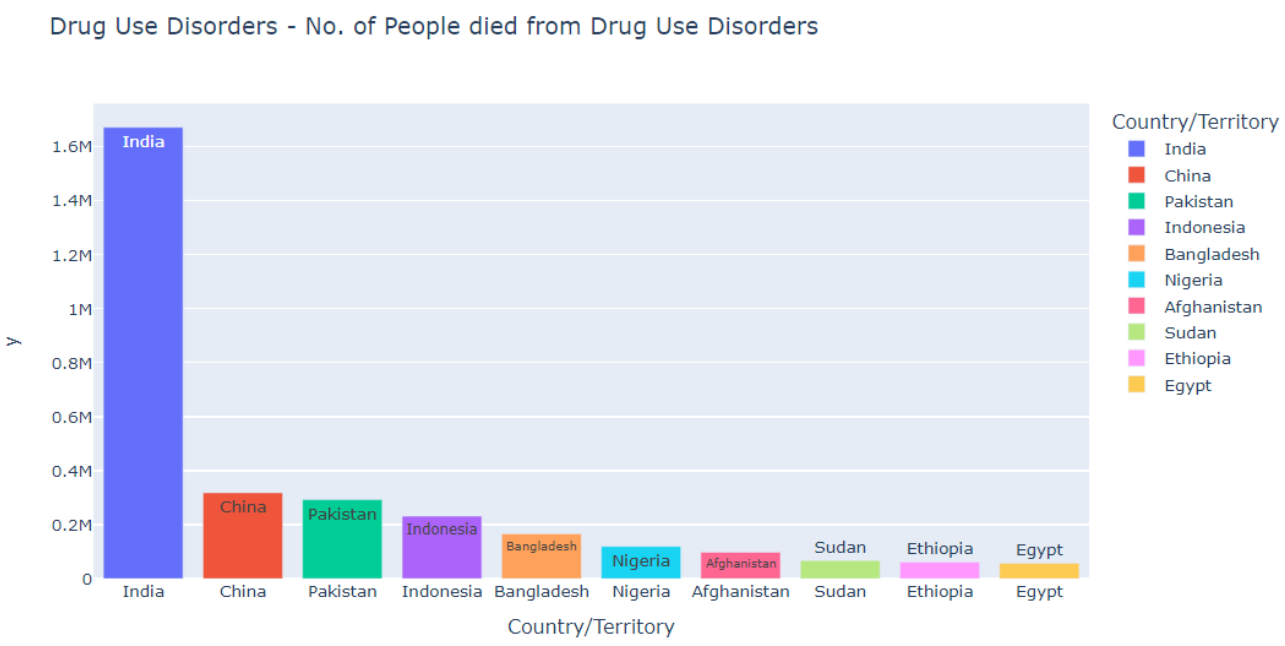
Interpersonal Violence - No. of People died from Interpersonal Violence



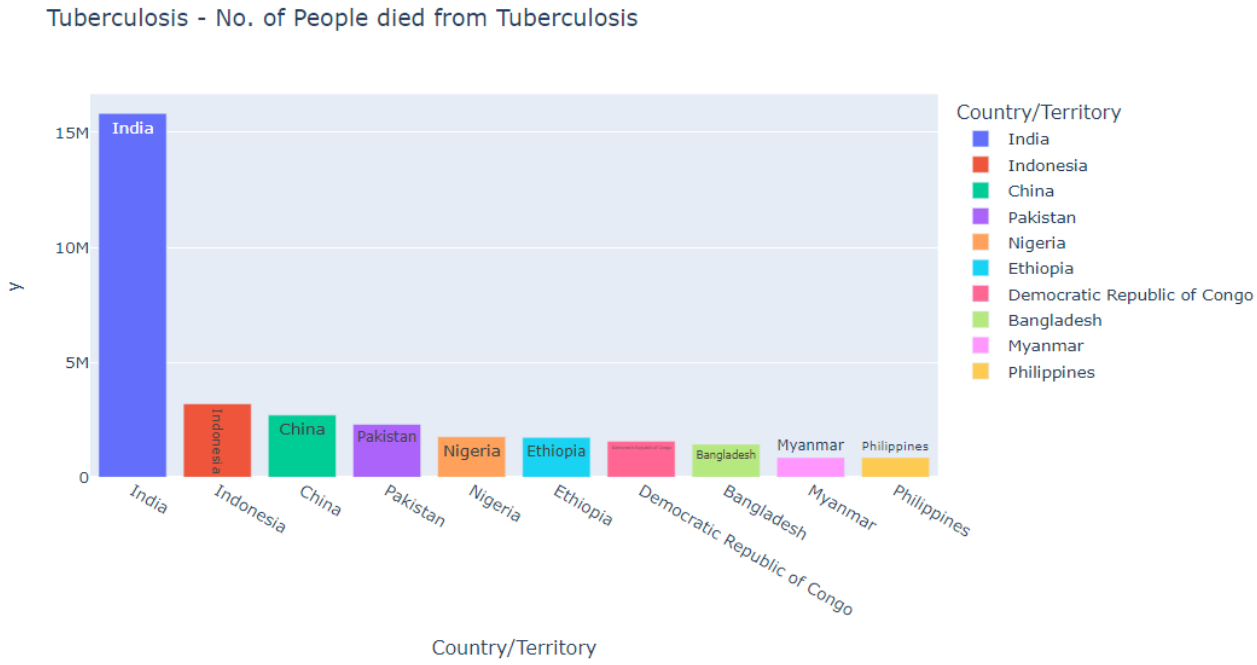
Maternal Disorders - No. of People died from Maternal Disorders



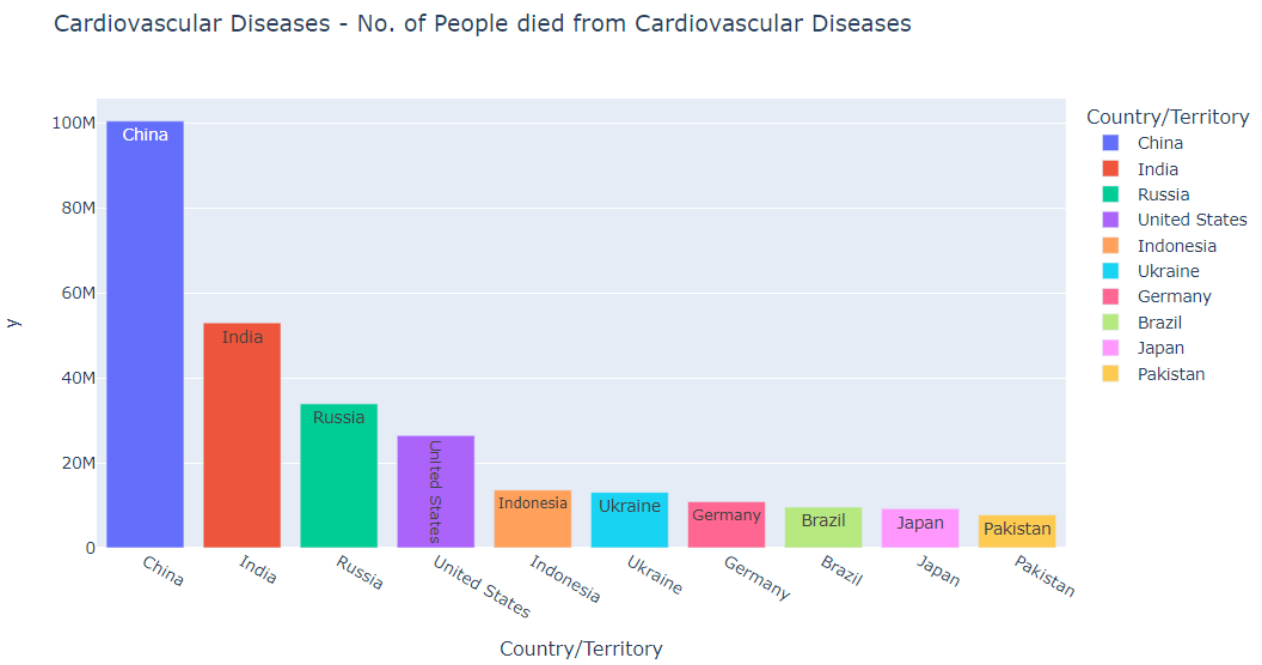
Drug Use Disorders - No. of People died from Drug Use Disorders



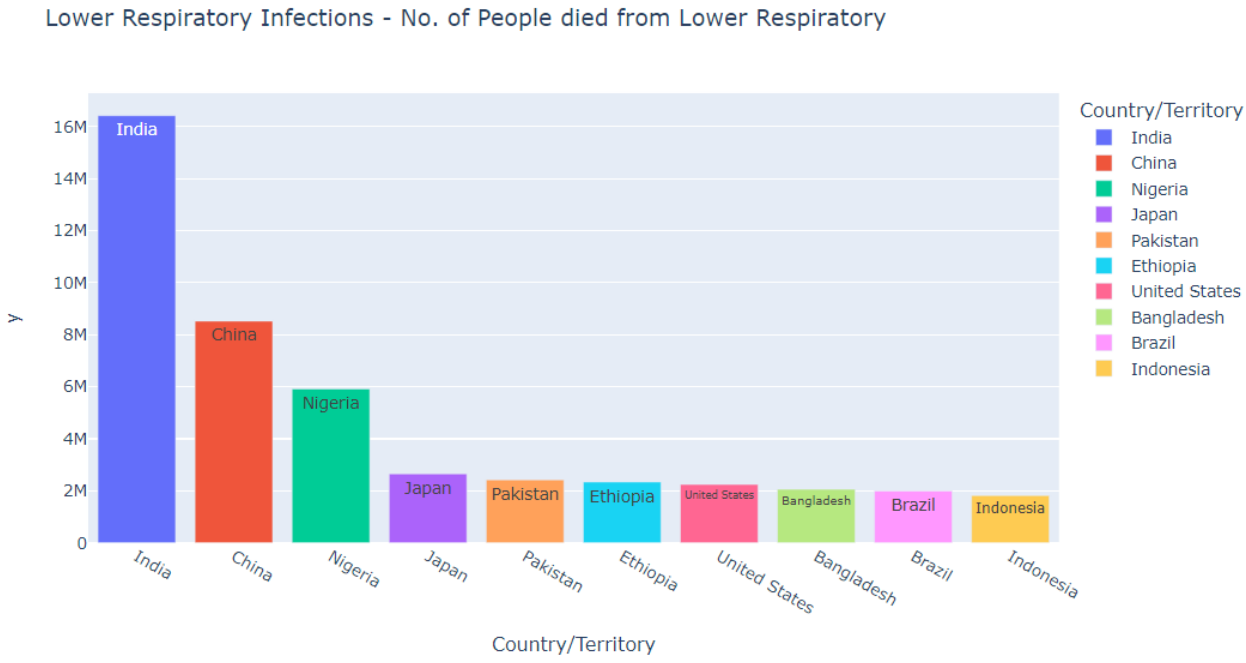
Tuberculosis - No. of People died from Tuberculosis



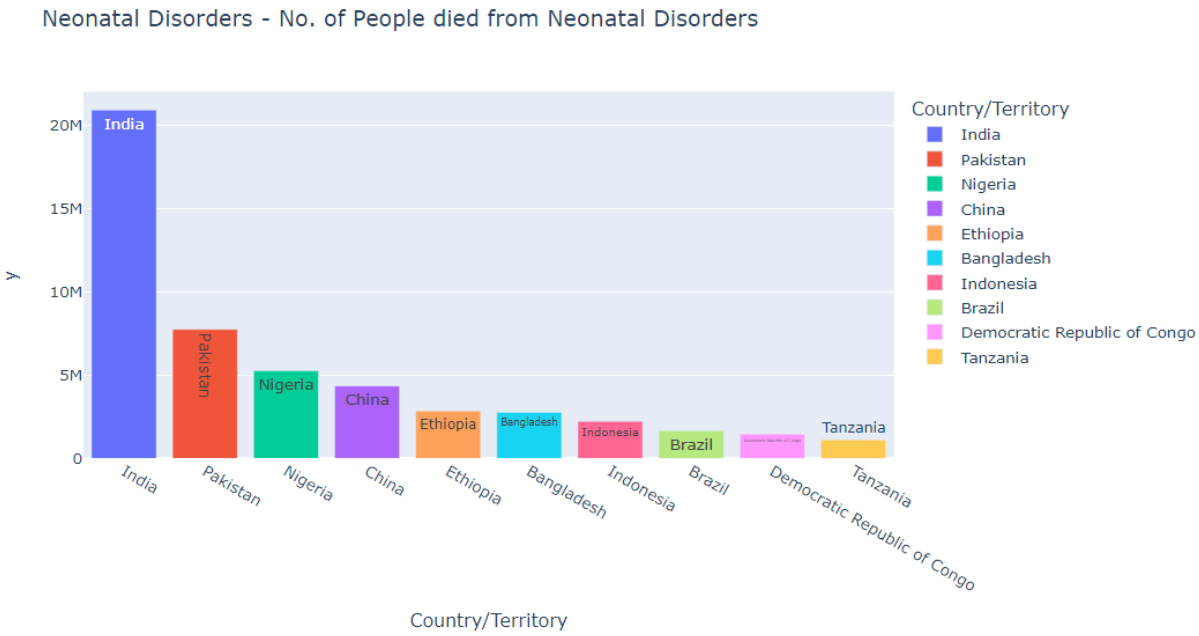
Cardiovascular Diseases - No. of People died from cardiovascular diseases



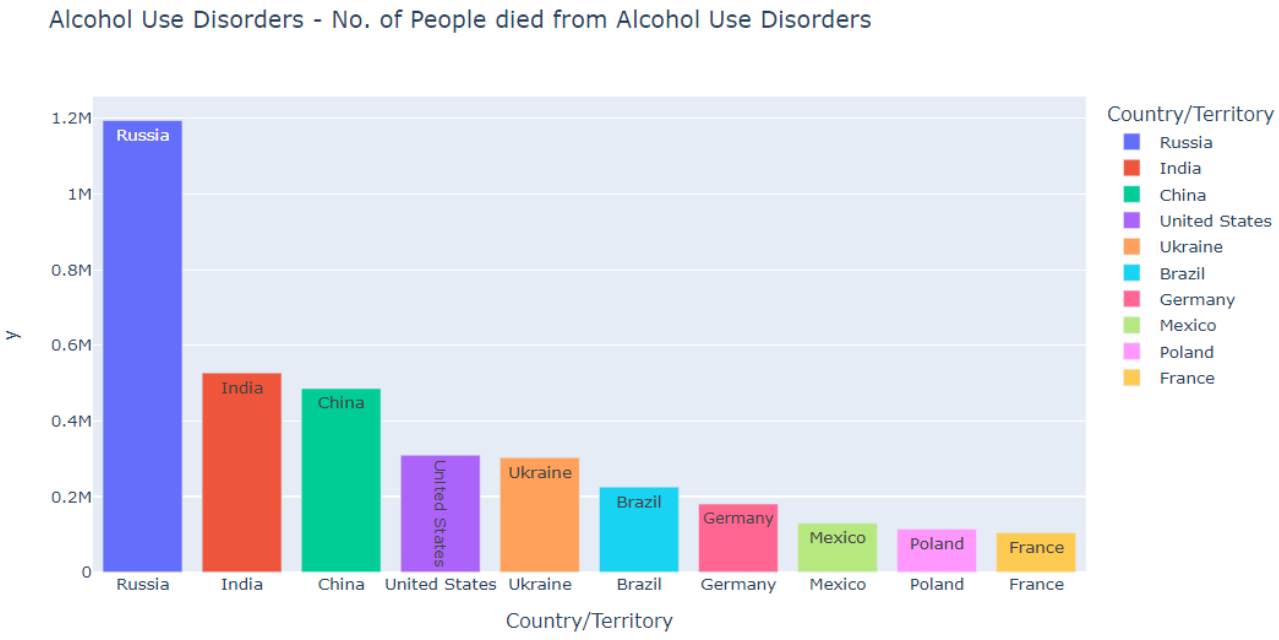
Lower Respiratory Infections - No. of People died from Lower Respiratory Infections



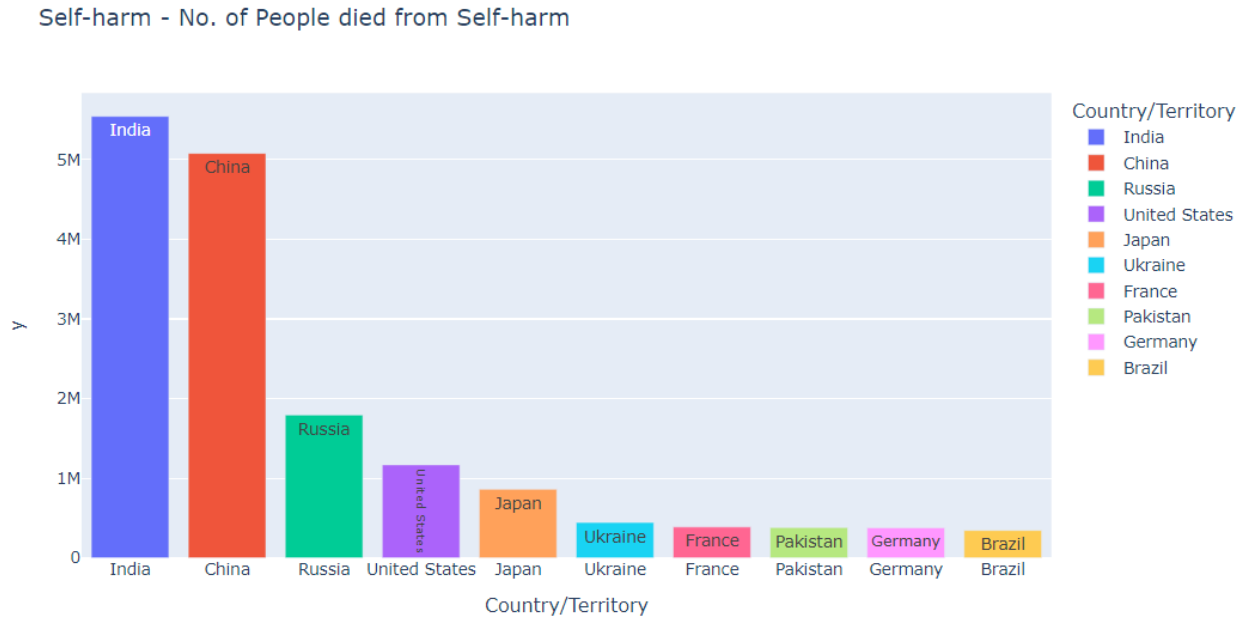
Neonatal Disorders - No. of People died from Neonatal Disorders



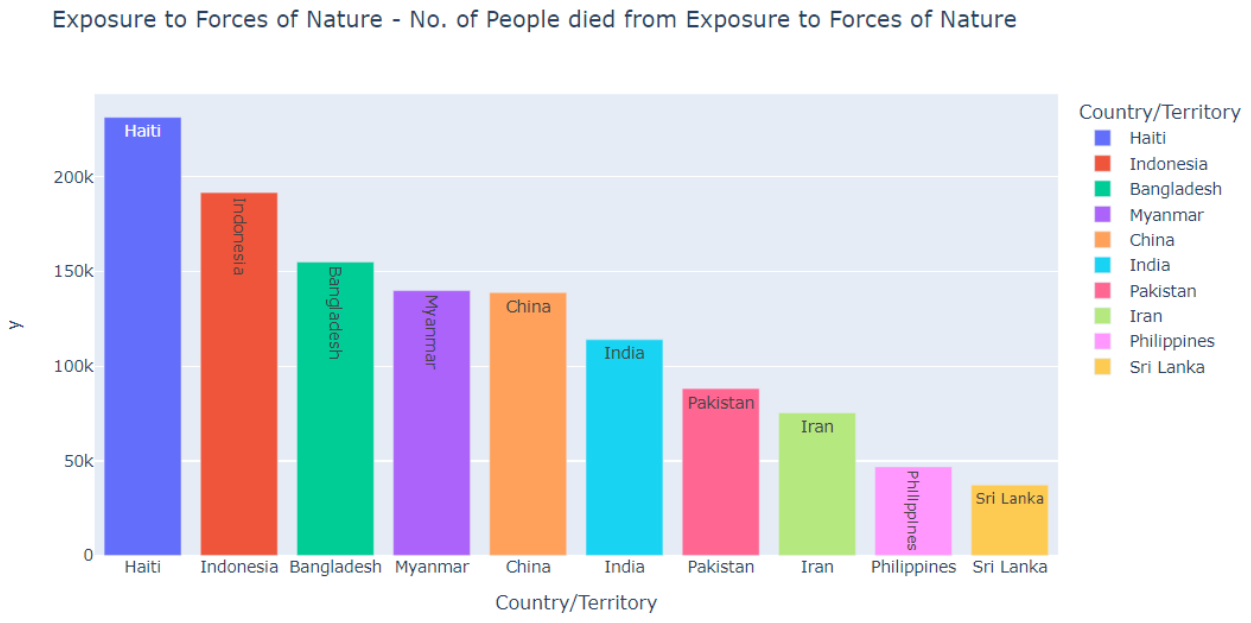
Alcohol Use Disorders - No. of People died from Alcohol Use Disorders



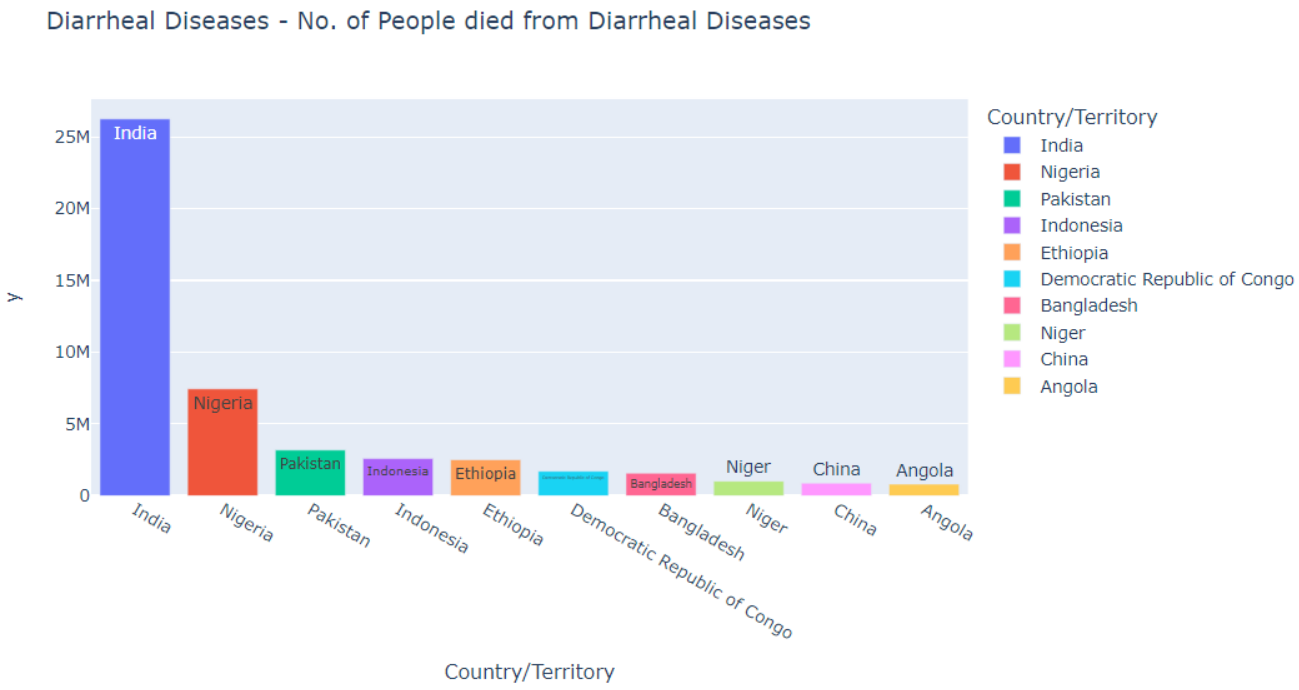
Self-harm - No. of People died from Self-harm



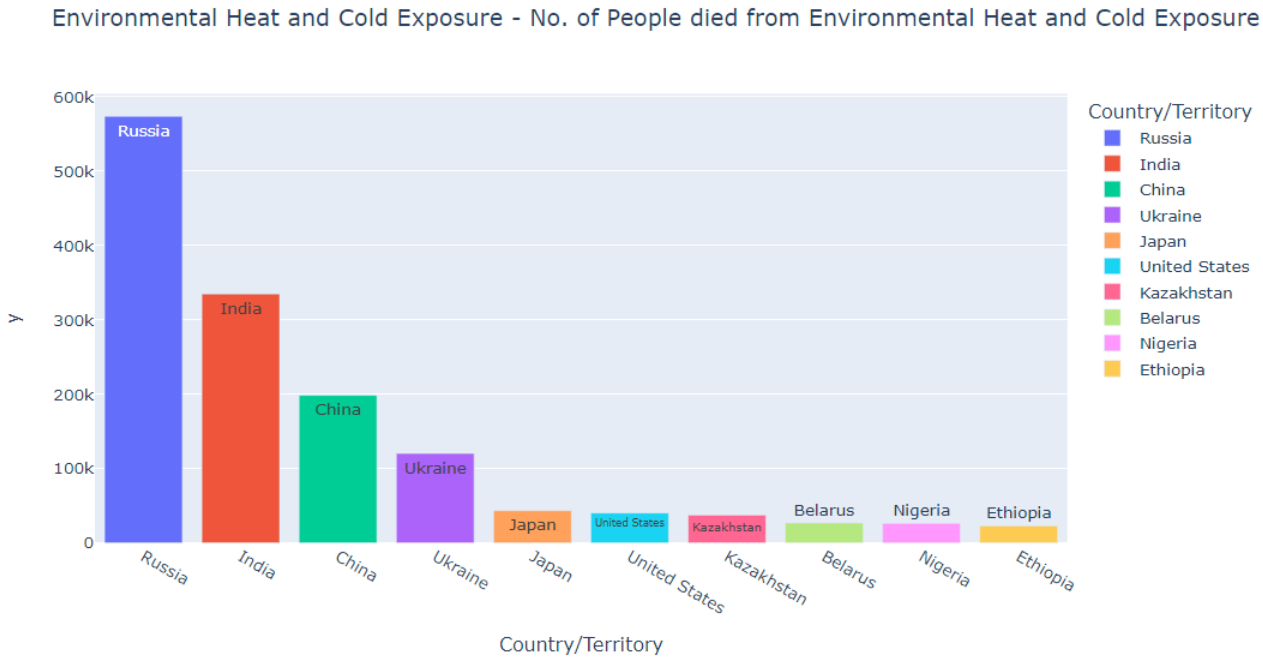
Exposure to Forces of Nature - No. of People died from Exposure to Forces of Nature



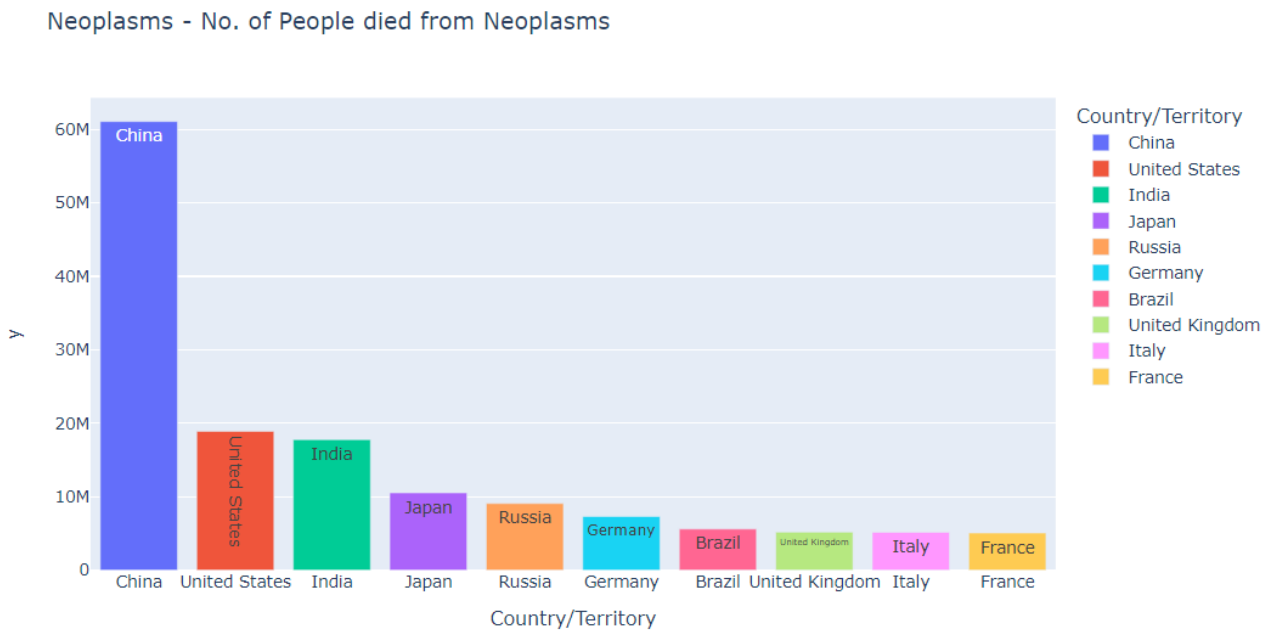
Diarrheal Diseases - No. of People died from Diarrheal Diseases



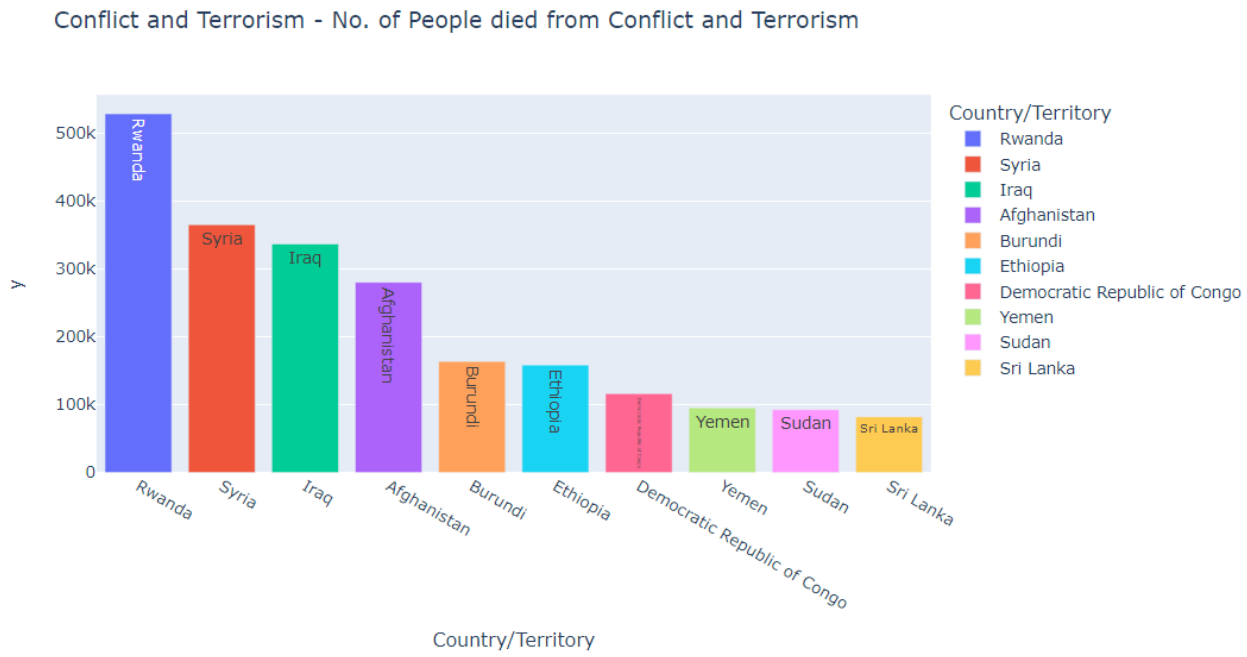
Environmental Heat and Cold Exposure - No. of People died from Environmental Heat and Cold Exposure



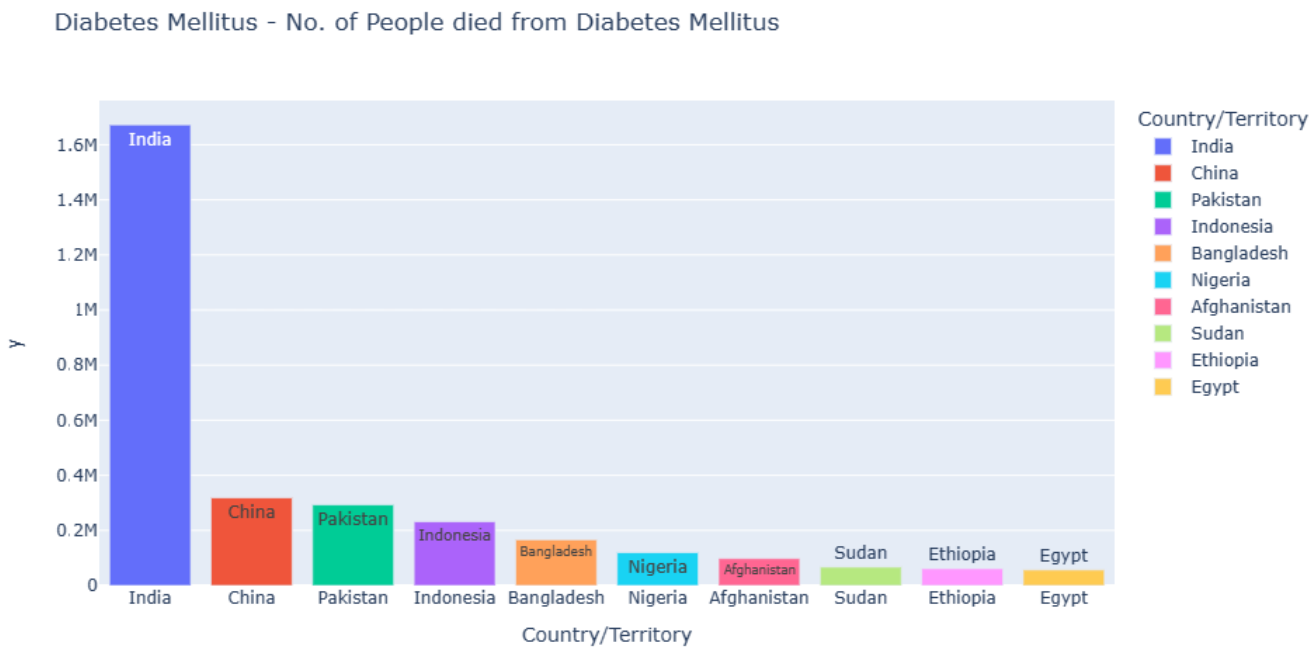
Neoplasms - No. of People died from Neoplasms



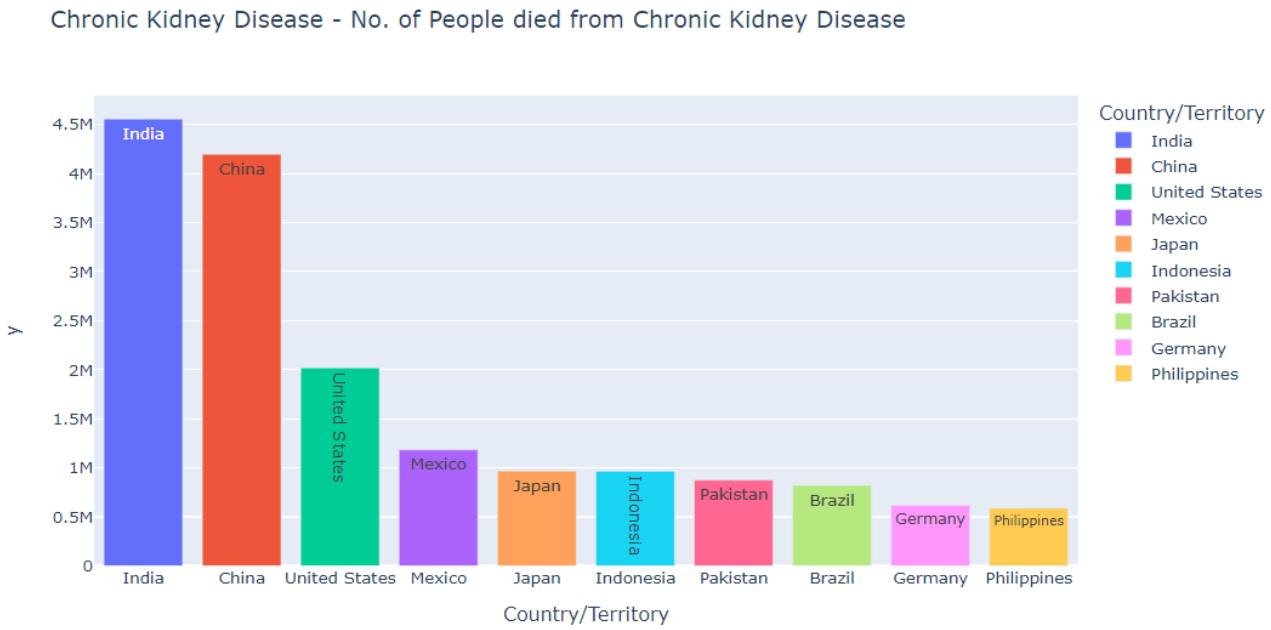
Conflict and Terrorism - No. of People died from Conflict and Terrorism



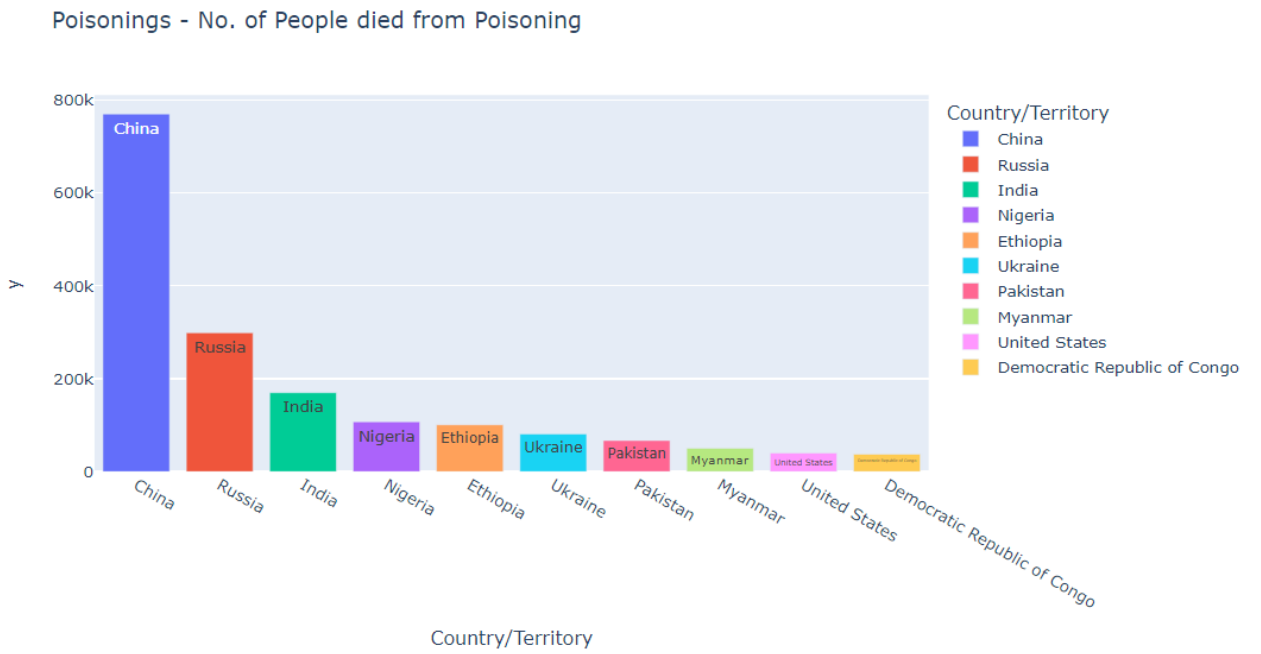
Diabetes Mellitus - No. of People died from Diabetes Mellitus



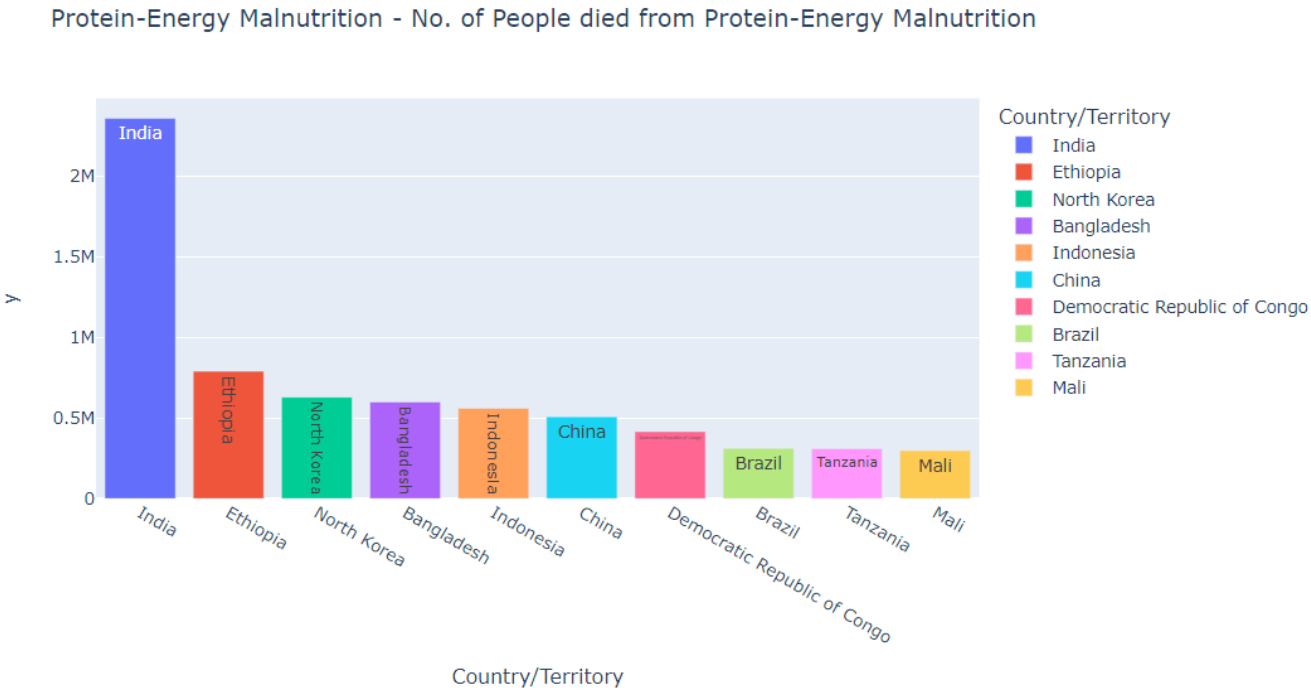
Chronic Kidney Disease - No. of People died from chronic kidney disease



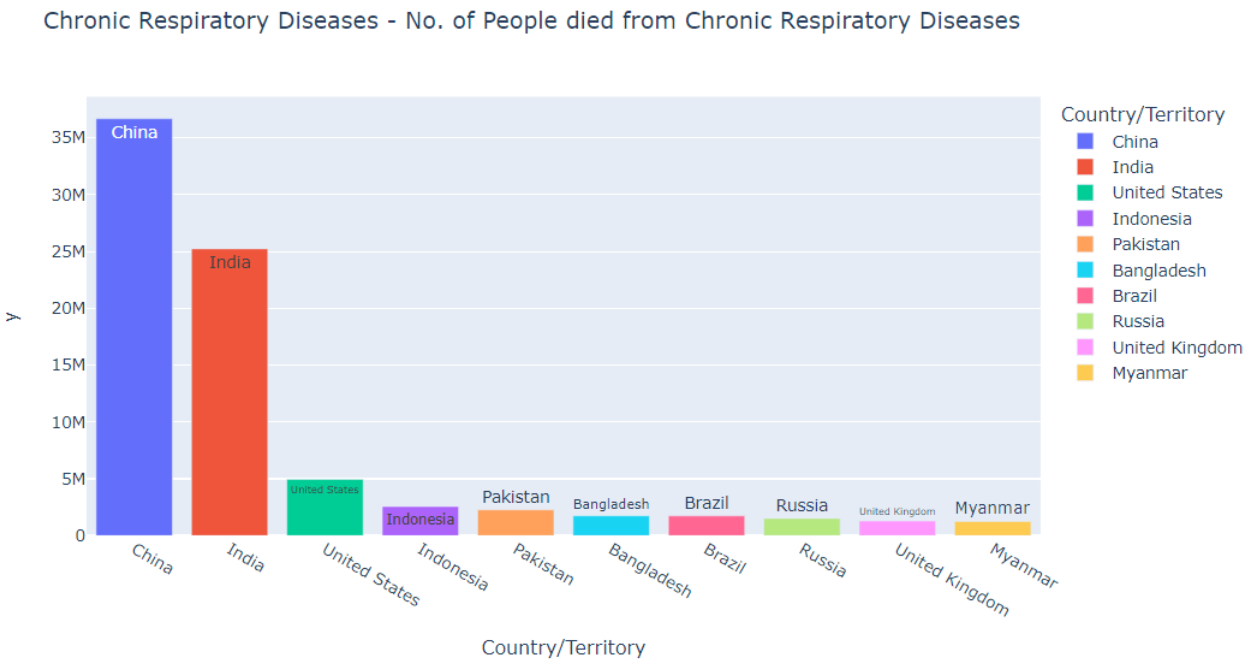
Poisonings - No. of People died from Poisoning



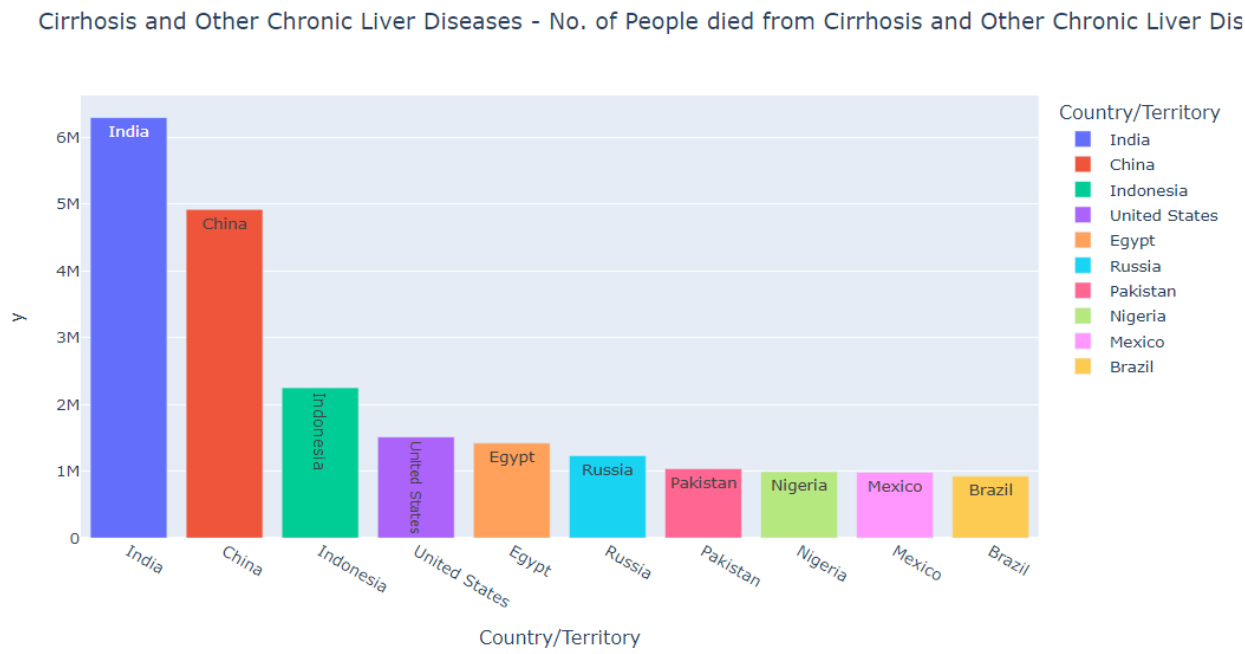
Protein-Energy Malnutrition - No. of People died from Protein-Energy Malnutrition



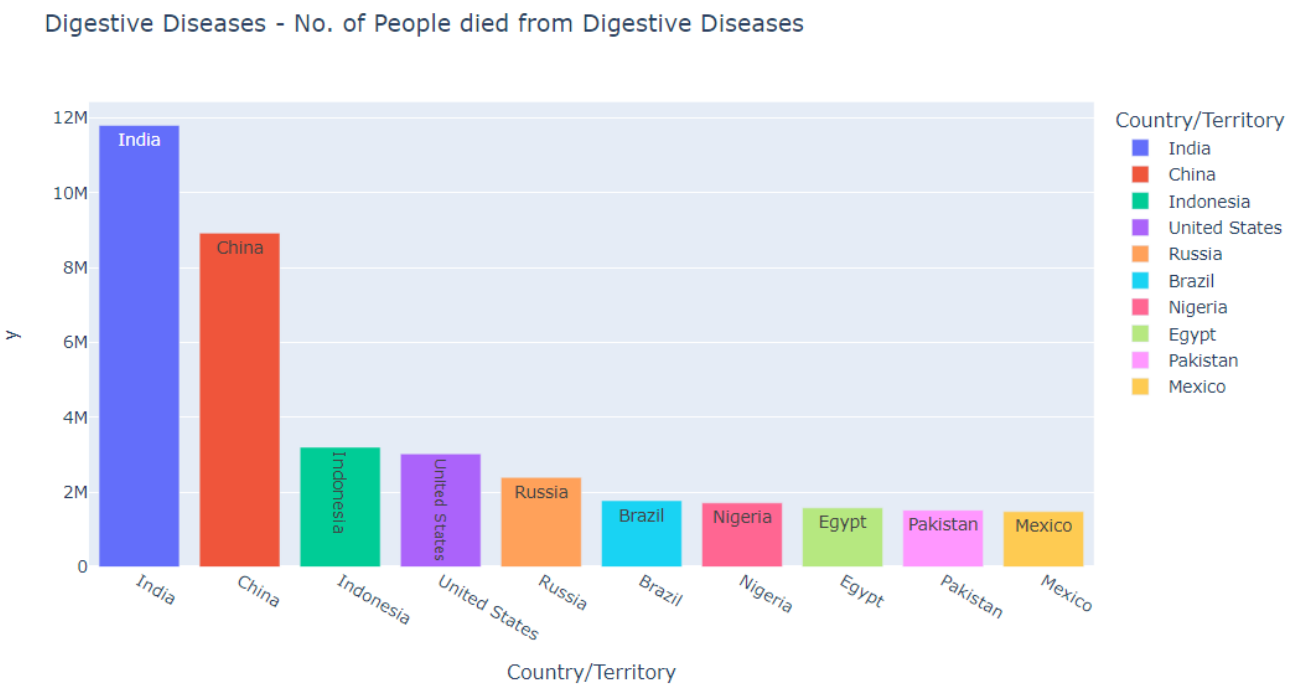
Chronic Respiratory Diseases - No. of People died from Chronic Respiratory Diseases



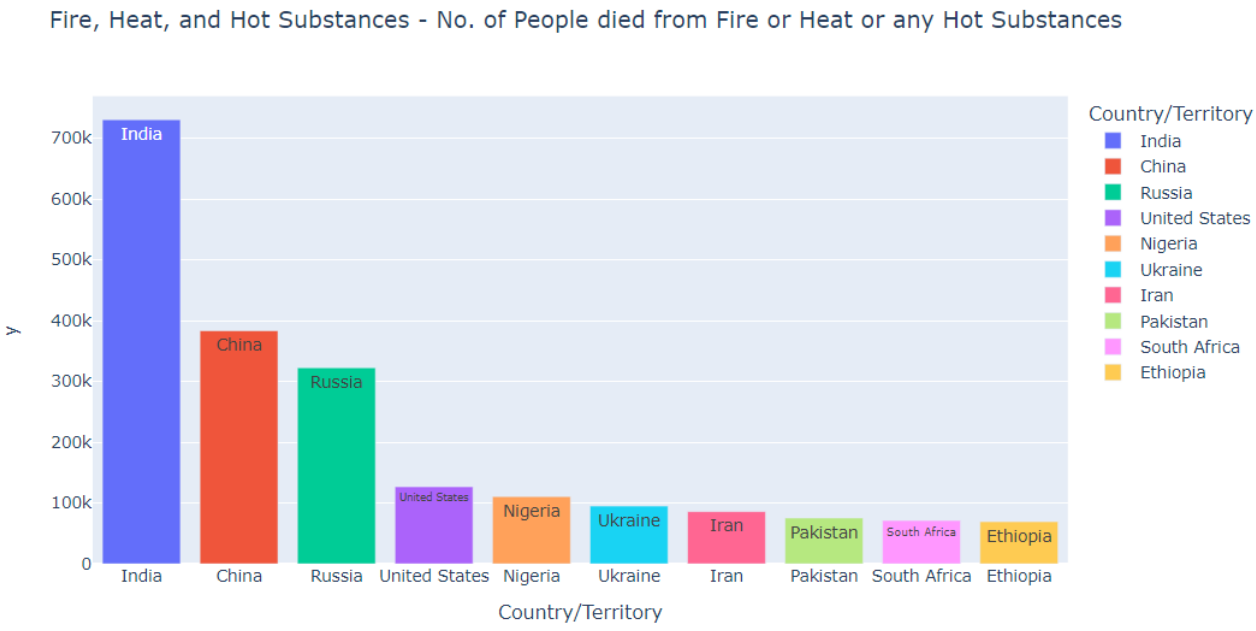
Cirrhosis and Other Chronic Liver Diseases - No. of People died from Cirrhosis and Other Chronic Liver Diseases



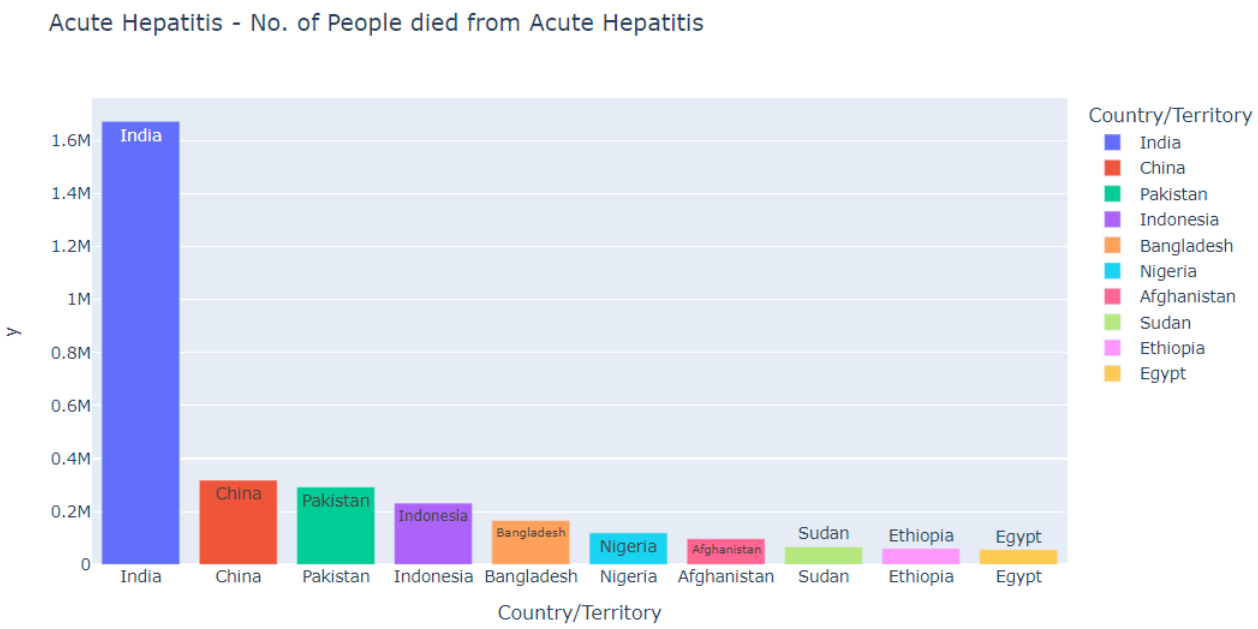
Digestive Diseases - No. of People died from Digestive Diseases



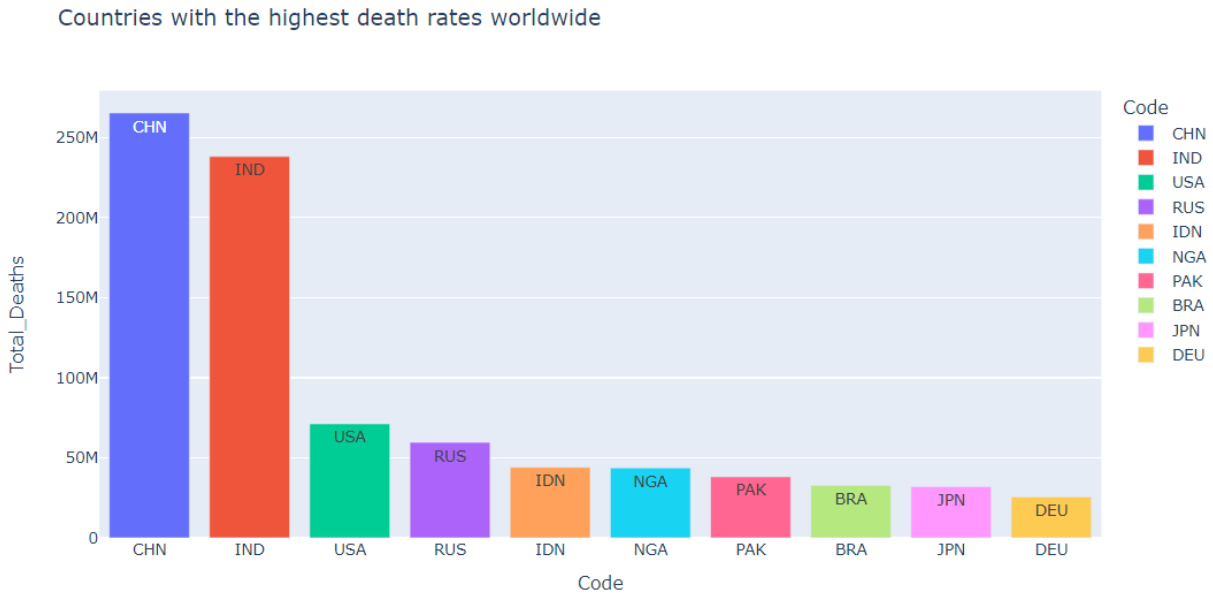
Fire, Heat, and Hot Substances - No. of People died from Fire or Heat or any Hot Substances



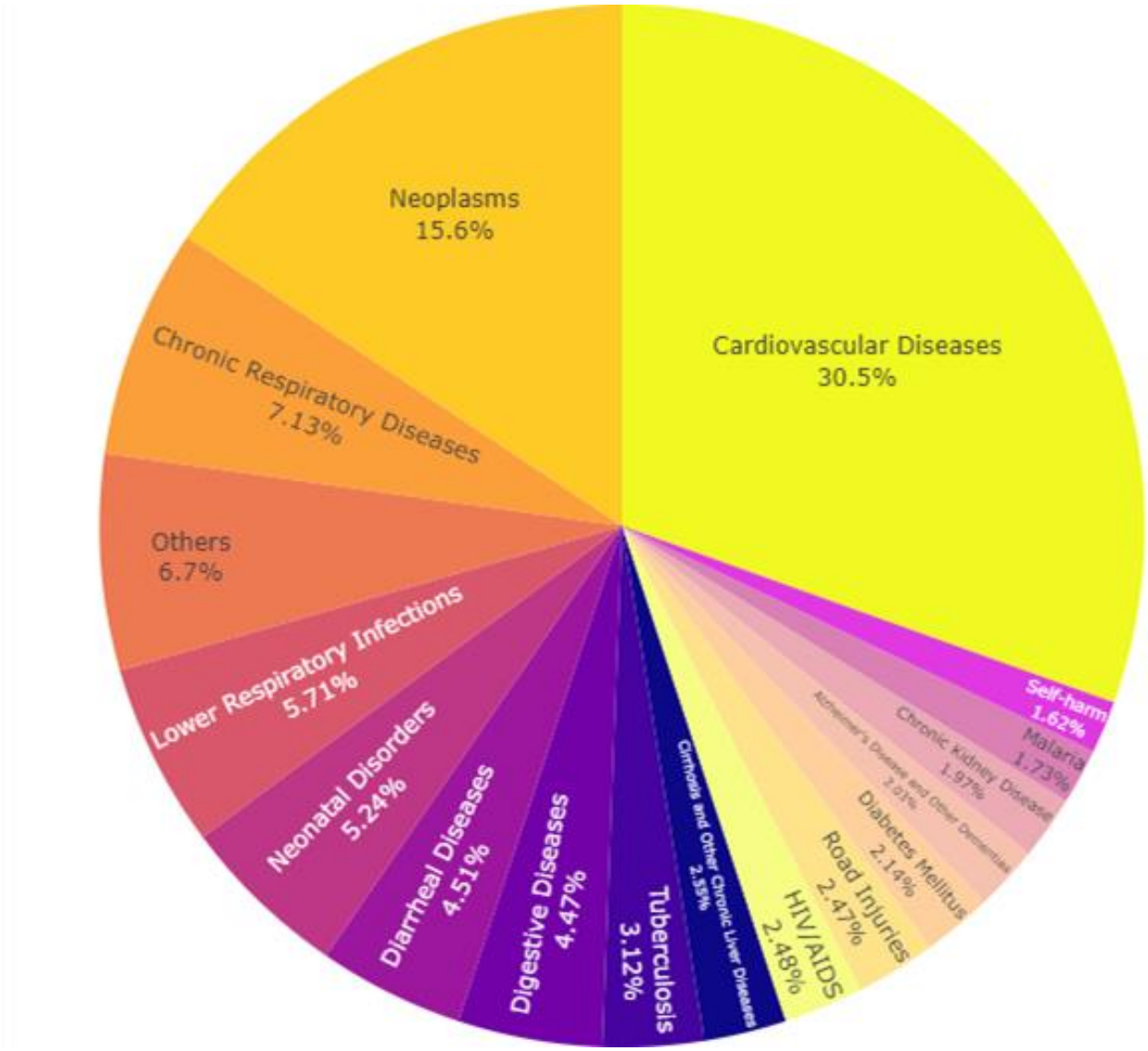
Acute Hepatitis - No. of People died from Acute Hepatitis



Countries with the highest death rates worldwide



Biggest deadliest diseases in the world



CONCLUSION

I have tested out the prediction CHINA , INDIA & USA has highest death rates due to Cardiovascular, Neoplasms ,and Respiratory are The top global causes of death.

Hence GradientBoostingRegressor is my top accurate model in predicting the percentage of death according to the Territory.

Also as you can see in the cardiovascular, Nepolasms are the top prior cause of death. It shows that the air quality and diet which we are taking inside having less quality dure to clauses.

- Learning Outcomes of the Study in respect of Data Science

I have used the Regression Model using multiple algorithms to design and optimise the results. Some of the classes which I explored from Scikit-learn libraries were – Statistics, Analytical Modelling, Hyper Tuning Method, CV Score, Predictive Modelling and RegressionModel, etc.

I condensed to 32 columns keeping in mind the Principle Component Analysis. This resulted in condensed data being trained in a much shorter span of time with utmost accuracy.
The visualizations helped in better and quick understanding of the outliers and skewness present in the data sets

With Mathematical Modeling heped me to find-out the corresponding mean, median, mode and relationship among the variables.
Statistical Modeling helped in Correlation for understanding the relationship among the variables

One of the key challenge which I faced was that the data set and if the data set were in the same file, then I could have analyze more in-depth accuracy. Root mean square error could have been tested if the data was provided to us in a same file access.

- Limitations of this work and Scope for Future Work

Data was limited and more raw data is required to be concrete on the decision making and statistical analysis.

The same analysis which was done for this project can’t be used for another death related projects as other variables or extra identifiers hasn’t been factored in.

Refrences

[1] E. G. Brooks and K. D. Reed, "Principles and pitfalls: a guide to death certification," Clinical medicine & research, vol. 13, pp. 74-82, 2015.
[2] P. Jha, "Counting the dead is one of the world's best investments to reduce premature mortality," Hypothesis, vol. 10, p. e3, 2012.
[3] T. L. Chorba, R. C. Holman, M. J. Clarke, and B. L. Evatt, "Effects of HIV infection on age and cause of death for persons with hemophilia A in the United States," American journal of hematology, vol. 66, pp. 229-240, 2001.
[4] W. C. Hooper, R. C. Holman, M. Clarke, and T. L. Chorba, "Trends in nonhodgkin lymphoma (NHL) and HIVassociated NHL deaths in the United States," American journal of hematology, vol. 66, pp. 159-166, 2001.
[5] D. M. Mannino, C. Brown, and G. A. Giovino, "Obstructive lung disease deaths in the United States from 1979 through 1993: an analysis using multiple-cause mortality data," American journal of respiratory and critical care medicine, vol. 156, pp. 814-818, 1997.
[6] C. Gordon, "Australian Bureau of Statistics, Multiple cause of death analysis. Publication 3319.0. 55.001," ed, 2003.
[7] S. Bah and M. M. Rahman, "Measures of multiple-cause mortality: a synthesis and a notational framework," Genus, vol. 65, pp. 29-43, 2009.
[8] W. H. Organization, International statistical classification of diseases and related health problems vol. 1: World Health Organization, 2004.
[9] R. A. Israel, H. M. Rosenberg, and L. R. Curtin, "Analytical potential for multiple cause-of-death data," American journal of epidemiology, vol. 124, pp. 161-81, 1986.
[10] P. F. Brown, P. V. Desouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai, "Class-based n-gram models of natural language," Computational linguistics, vol. 18, pp. 467-479, 1992.
[11] A. Pak and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," in LREc, 2010.
[12] B. J. Marafino, J. M. Davies, N. S. Bardach, M. L. Dean, R. A. Dudley, and J. Boscardin, "N-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit," Journal of the American Medical Informatics Association, vol. 21, pp. 871-875, 2014.
[13] V. Kešelj, F. Peng, N. Cercone, and C. Thomas, "N-gram-based author profiles for authorship attribution," in Proceedings of the conference pacific association for computational linguistics, PACLING, 2003, pp. 255-264.
[14] A. Tomović, P. Janičić, and V. Kešelj, "n-Gram-based classification and unsupervised hierarchical clustering of genome sequences," Computer methods and programs in biomedicine,

[15] B. Wajid and E. Serpedin, "Review of general algorithmic features for genome assemblers for next generation sequencers," Genomics, proteomics & bioinformatics, vol. 10, pp. 58-73, 2012.
[16] J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters," Communications of the ACM, vol. 51, pp. 107-113, 2008.
[17] Y. Park and M. Kellis, "Deep learning for regulatory genomics," Nat Biotechnol, vol. 33, pp. 825-6, 2015.
[18] H. R. Hassanzadeh and M. D. Wang, "DeeperBind: Enhancing Prediction of Sequence Specificities of DNA Binding Proteins," presented at the Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on, 2016.
[19] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, "Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning," Nature biotechnology, vol. 33, pp. 831-838, 2015.
[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, pp. 1735-1780, 1997.
[21] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," Journal of Machine Learning Research, vol. 9, pp. 2579-2605, 2008.

Thank You