

# Visual Analytics of Causes of Death in the United States

Greeshma Sai Sree Duggireddy

## INTRODUCTION

There have been several causes of death over the years. As we all know, covid has been the major cause of death in the previous two to three years. So, from 1999 to 2017, the data on deaths was collected and reported by the government in various locations in the United States. This study assists in identifying U.S. populations in various locations who are at a higher risk of death from diseases and injuries. The benefit of this process is to access medical treatment and know the frequency of certain risk factors.

### 1 INSIGHT NEEDS

The insight needs for this visualization project are associated with understanding the causes of death and mortality rates throughout the United States which is essential for public health and protection stakeholders, such as policymakers, healthcare professionals, researchers, government officials, and police department as it helps them identify priority areas for interventions and allocate resources accordingly. By having an accurate understanding of past and present, we can identify the most common causes of death, and the stakeholders can develop effective prevention and treatment strategies to reduce mortality rates and improve health outcomes.

The first insight need for this visualization project is to identify the most common causes of death in the US and their trends over time. The second insight need is to identify significant differences in causes of death and mortality rates between different demographic regions within the US. The third insight need is to provide a comprehensive understanding of the states in the US that have higher numbers of deaths, the causes of death that have higher death rates, and the overall deaths in the US between the years 1999 and 2017. The fourth insight need shows which causes of death were most common and which states had the most deaths between 1999 and 2017, along with their ranks.

#### 1.1 Stakeholder Analysis

The stakeholders identified for this visual analytics project are people involved in public health and protection stakeholders, such as policymakers, healthcare professionals, researchers, government officials, and the police department. The insights from the visualizations can help stakeholders create tailored strategies that target the particular risk factors causing increased mortality rates and implement them in a targeted and effective way by identifying high-risk areas of occurrence. This may result in public health initiatives that are more successful and significant, ultimately enhancing the health and well-being of US communities.

## 2 DATA ACQUISITION

The dataset for my visual analytics project is taken from the website – data.gov<sup>[1]</sup>. Data.gov is a website operated by the government of the United States which provides free access to a wide range of data on various topics such as education, health, deaths, agriculture, finance, etc. The dataset I chose is - NCHS-Leading-Causes-of-death-united-states<sup>[2]</sup>. The dataset is in the .csv file type.

### 2.1 Description of Data

The dataset I chose from the data.gov website is a huge dataset with 10868 entries. The columns in the dataset are – Year, Cause Name, State, Deaths, and Age-adjusted Death Rate. The timeframe set for the dataset is from the year 1999 to 2017. The Cause Name is the cause of death like Cancer, Heart Disease, Suicide, and others. The

state is one of the 50 states in the United States. The deaths are the total number of deaths for that particular year, the cause, and the state. The age-adjusted death rate is a computation that accounts for changes in age distributions among populations to make fair comparisons of mortality rates possible.

## 3 ANALYSIS METHODS

After obtaining the dataset from data.gov, it is important to check for any missing data or deficiencies, which I accomplished using Python in Jupyter Notebook; then, I utilized SQL queries to filter the data for use in creating compelling visualizations in tableau and powerBI.

### 3.1 Analysis Methods

After utilizing Python and SQL to analyse the dataset, the next step involves creating visual analytics using tableau and powerBI, which includes a variety of visualizations such as temporal, geospatial, topical, and network analyses that provide valuable insights to stakeholders. These visual analytics methods assist in addressing questions regarding when, where, what, and with whom issues that arise for stakeholders, providing clear and concise insights to better understand and devise effective solutions to problems.<sup>[3]</sup>

## 4 VISUALIZATIONS

The temporal, geospatial, topical, and network visual analytics methods assist in addressing questions regarding when, where, what, and with whom issues respectively that arise for stakeholders, providing clear and concise insights to better understand and devise effective solutions to problems. To answer stakeholder insight needs, the visualizations used are - Line graphs, geo maps, word clouds, stacked bar graphs, and radar graphs.

### 4.1.1 Prototype - 1

The below line graph to show the overall number of deaths from 1999 to 2017. In this prototype, the graphic symbols and variables used are geometric symbols – Line and Linguistic Symbols – Text and Numerals. The X-axis represents the year, and the y-axis represents the total deaths.

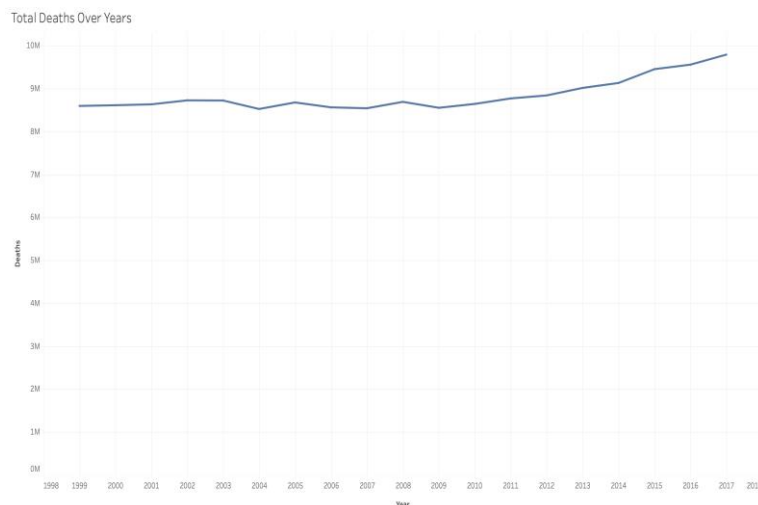


Fig. 1. Figure

#### 4.1.2 Prototype 2

To depict numerous causes of death a second line graph is created. Different hues represent various causes of death. In this prototype, the graphic symbols and variables used are geometric symbols – Line and Linguistic Symbols – Text and Numerals. The X-axis shows the year, while the Y-axis reflects the number of deaths. This graph shows the number of deaths from a certain cause over time in various locations in the United States. The color hue represents various causes of death.

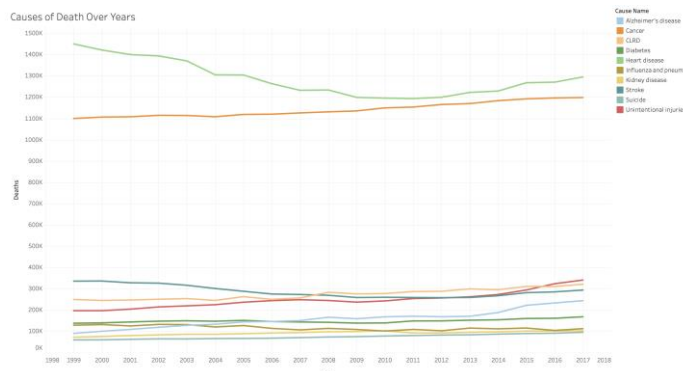


Fig. 2. Figure

### 4.1.3 Prototype 3

The total mortality rate in the United States is shown on this geo map. In this prototype, the graphic symbols and variables used are geometric symbols – Surface and Linguistic Symbols – Text and Numerals. The color saturation represents the total number of deaths for each cause in each state. The yellow hue represents the cause of death with the lowest mortality rate, while the red color represents the cause of death with the greatest mortality rate. The surface layout is the map of the United States.

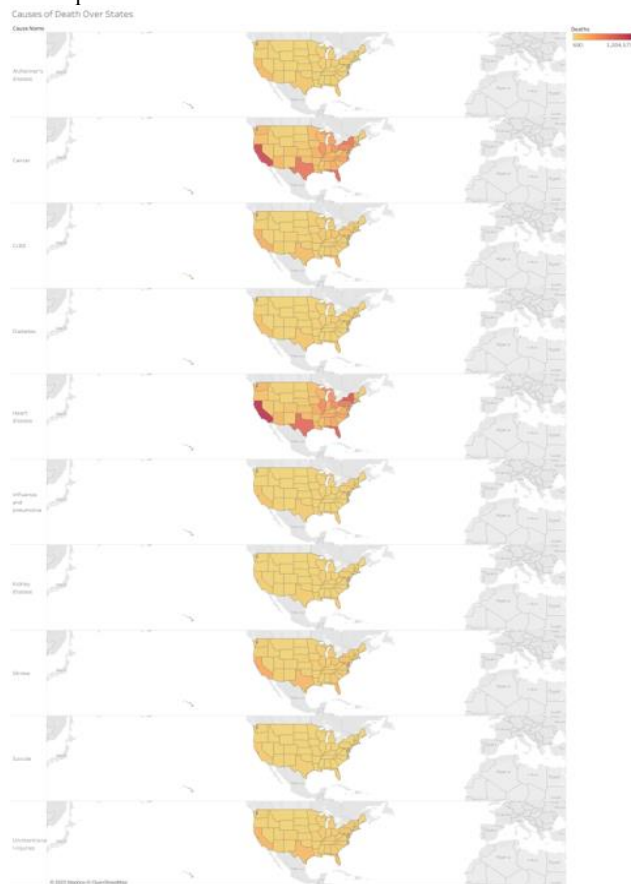


Fig. 3. Figure

#### 4.1.4 Prototype 4

The word cloud shows the causes of death in the US in proportion to the number of deaths due to that cause. In this prototype, the graphic symbols and variables used are Linguistic Symbols – Text and Numerals. The color saturation represents the total number of deaths due to that particular cause.

## Causes of Death



Fig. 4. Figure

#### 4.1.5 Prototype 5

The word cloud shows the states in the US in proportion to the number of deaths in that state. In this prototype, the graphic symbols and variables used are Linguistic Symbols – Text and Numerals. The color saturation represents the total number of deaths in each state.

Deaths in United States



Fig. 5. Figure

#### 4.1.6 Prototype 6

The stacked bar chart below shows the death count against the years along with the classification of the causes of death. In this prototype, the graphic symbols and variables used are Linguistic Symbols – Text and Numerals, Geometric Symbols – Rectangular bars. The color hue represents the causes of death, and the magnitude of the bars represents the number of deaths. The X-axis represents the year, and the Y-axis represents the number of deaths.

Year vs. Death Count

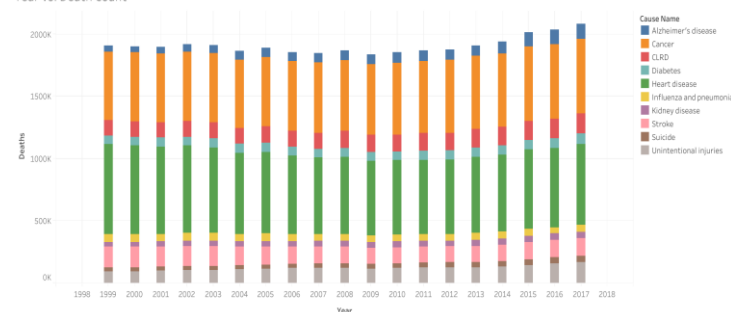


Fig. 6. Figure

#### 4.1.7 Prototype 7

The radar graph shows the causes of death vs. the number of deaths. In this prototype, the graphic symbols and variables used are

Geometric symbols – Point and surface, Linguistic symbols – Text. The text is associated with causes of death and the surface area connects the points which represents the number of deaths due to the cause.

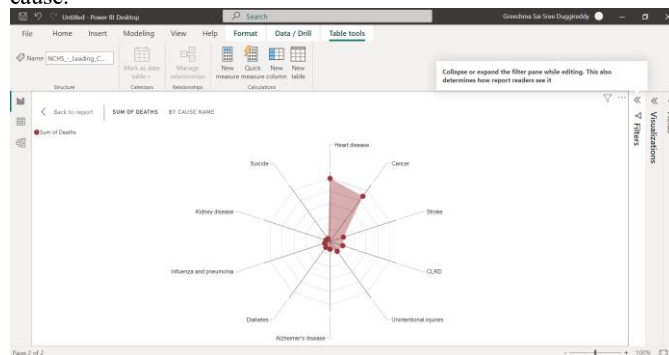


Fig. 7. Figure

#### 4.1.8 Prototype 8

The radar graph shows the States vs. the number of deaths. In this prototype, the graphic symbols and variables used are Geometric symbols – Point and surface, Linguistic symbols – Text. The text is associated with each state and the surface area connects the points which represents the number of deaths in that state.

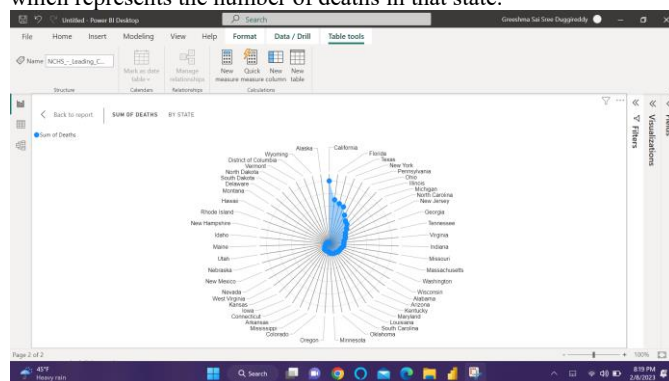


Fig. 8. Figure

## 5 INTERPRETATION OF RESULTS

Prototype 1 shows that the death rate in the USA increased from 1999 to 2003, with a dip in 2004 resulting in the lowest recorded rate. From 2004 to 2010, the death rate remained relatively stable with slight fluctuations. However, since 2010, the death rate has been increasing.

Prototype 2 indicates that heart disease is the most common cause of death, followed by cancer. The number of deaths due to strokes, CLRD, and injuries fluctuated over the years but remained relatively consistent. Kidney disease and suicide were the least common causes of death.

Prototype 3 highlights that California has the highest death rate across all categories.

Prototype 4's word cloud depicts that heart disease and Cancer are leading causes of death in the United States.

Prototype 5's word cloud reveals that California, Florida, New York, and Texas have the highest mortality rates, while Hawaii, Idaho, Utah, and a few other states have the lowest rates.

Prototype 6's stacked bar chart shows a steady increase in death count from 2012 to 2017, with fluctuations in previous years. Cancer and Heart Disease remain the leading causes of death. Additionally, it appears that no particular disease has seen a significant decrease in death count over the years, indicating a lack of effective preventive measures.

Prototype 7's radar graph ranks the causes of death, showing that suicide, kidney disease, diabetes, influenza, and pneumonia have

similar numbers with a lower death count than Cancer and Heart Disease.

Prototype 8's radar graph shows state rankings, with California, Florida, Texas, New York, Pennsylvania, and others having the highest number of deaths, and Alaska, Wyoming, the District of Columbia, and others having the lowest.

Heart disease is the most common cause of death in the United States, followed by cancer. California has the highest death rate across all categories, with the highest number of deaths due to heart disease and cancer. The leading causes of death have been equally proportional through the years 1999 to 2017, indicating that preventive measures for specific diseases have not been effective in reducing death counts. These insights can help the stakeholders change the public health policies and interventions which will aim at reducing death rates and improving health outcomes in the United States. Overall, these visualizations provide a clear understanding of death rates and leading causes of death in the United States.

## ACKNOWLEDGMENTS

The author wishes to thank Professor Andreas Bueckle and Professor Michael Ginda, along with the wonderful and helpful TAs - Saber Sheybani, Athulya Anand, Luis Mestre, and Britain Taylor.

## REFERENCES

- [1] Usdatagov @usdatagov, and Usdatagov. Data.gov, 14 Sept. 2020, <https://data.gov/>.
- [2] "NCHS - Leading Causes of Death: United States." Catalog, Publisher Centers for Disease Control and Prevention, 21 Apr. 2022, <https://catalog.data.gov/dataset/nchs-leading-causes-of-death-united-states>.
- [3] School of Informatics Indiana University, School of Informatics ... - IU. [https://cns.iu.edu/docs/handouts/Atlas\\_of\\_Knowledge\\_Flyer\\_Io.pdf](https://cns.iu.edu/docs/handouts/Atlas_of_Knowledge_Flyer_Io.pdf).