**ANALYSIS OF LEADING CAUSES OF DEATH IN THE UNITED STATES**



**Department of Applied Data Science, San Jose State University**

**DATA 230: Data Visualization**

**Professor: Andrew H. Bond**

**By:**

**Jugal Kishore Ruvva**

**016587519**

**TABLE OF CONTENTS**

[ABSTRACT 4](#_Toc121678090)

[INTRODUCTION 5](#_Toc121678091)

[OBJECTIVES 6](#_Toc121678092)

[DATASET 7](#_Toc121678093)

[**Related work:** 7](#_Toc121678094)

[**About the dataset:** 7](#_Toc121678095)

[**Data Collection:** 7](#_Toc121678096)

[**Column Description:** 8](#_Toc121678097)

[METHODS 11](#_Toc121678098)

[**Data Cleaning and Wrangling:** 11](#_Toc121678099)

[**Visualization Design:** 12](#_Toc121678100)

[**Tableau Operation:** 13](#_Toc121678101)

[RESULTS 14](#_Toc121678102)

[DISCUSSION 22](#_Toc121678103)

[**What I learned and how I used it:** 22](#_Toc121678104)

[**Conclusion:** 26](#_Toc121678105)

[FUTURE WORK 27](#_Toc121678106)

[REFERENCES 28](#_Toc121678107)

# **ABSTRACT**

For a very long time, we might have heard the term “Health is wealth,” and nowadays, irrespective of age and gender, many people face different kinds of health issues. I found an exciting dataset containing data related to Health Problems with columns like Age, Sex, different types of diseases and chances of getting affected by that disease, and other essential columns. In this project, we collect data regarding the health conditions of people of different ages, races, and gender and analyze the same to support decision-making. First, we clean up the collected data by filtering out insufficient data, filling in missing values, and reformatting the data. Then we analyze the data and get insights. I expect to find out what age group of people is affected and which gender is affected most. After that, we will create meaningful visualizations in Tableau and derive valuable conclusions. The above analysis results show that the healthcare systems can make critical decisions regarding their insurance and health plans.

# **INTRODUCTION**

According to the report “Leading Causes of deaths in the USA” by the National Centre of Health Statistics, over 153 million people died from the year 1999 to 2015 in various states of the United States by multiple causes, broadly divided into sixteen categories. Going into the future, the number of deaths and reasons can only be seen to increase as the standard of living goes down, the food habits of people get worse, and due to outbreaks of different viruses, bacteria and pandemics.

Insurance health plans don't cover all body parts, and the ones that cover are so costly, and the cost of treatment is too high, and diseases the death toll keeps increasing.

Hence, there is a need to analyze and visualize the factors and causes of these deaths. So, let’s create different visualization using different available data attributes. Use the visualization to analyze the trends in the graphs, recognize the leading causes and draw insights that might help reform public health decisions. People expect better health and survival if the health sector prioritizes reforming the health and insurance plans and strengthens its response system.

# **OBJECTIVES**

* Data Wrangling and transforming the data
* Geo Spatial visualization regarding deaths by each cause
* Total number of deaths in each state by various causes per year
* Forecasting the number of deaths per each cause
* Finding the leading cause of death
* Exploring more about the leading cause of deaths
* Analysis of the number of deaths at different ages
* Finding leading causes of death in each state of the USA
* Number of deaths recorded in each year
* Record the number of deaths by each cause
* Cause-of-death statistics for each state
* Trend analysis of age-adjusted death rate
* Reforming health plans and insurance plans based on insights
* Publishing the visualization in GitHub and Tableau Public

# **DATASET**

**Related work:**

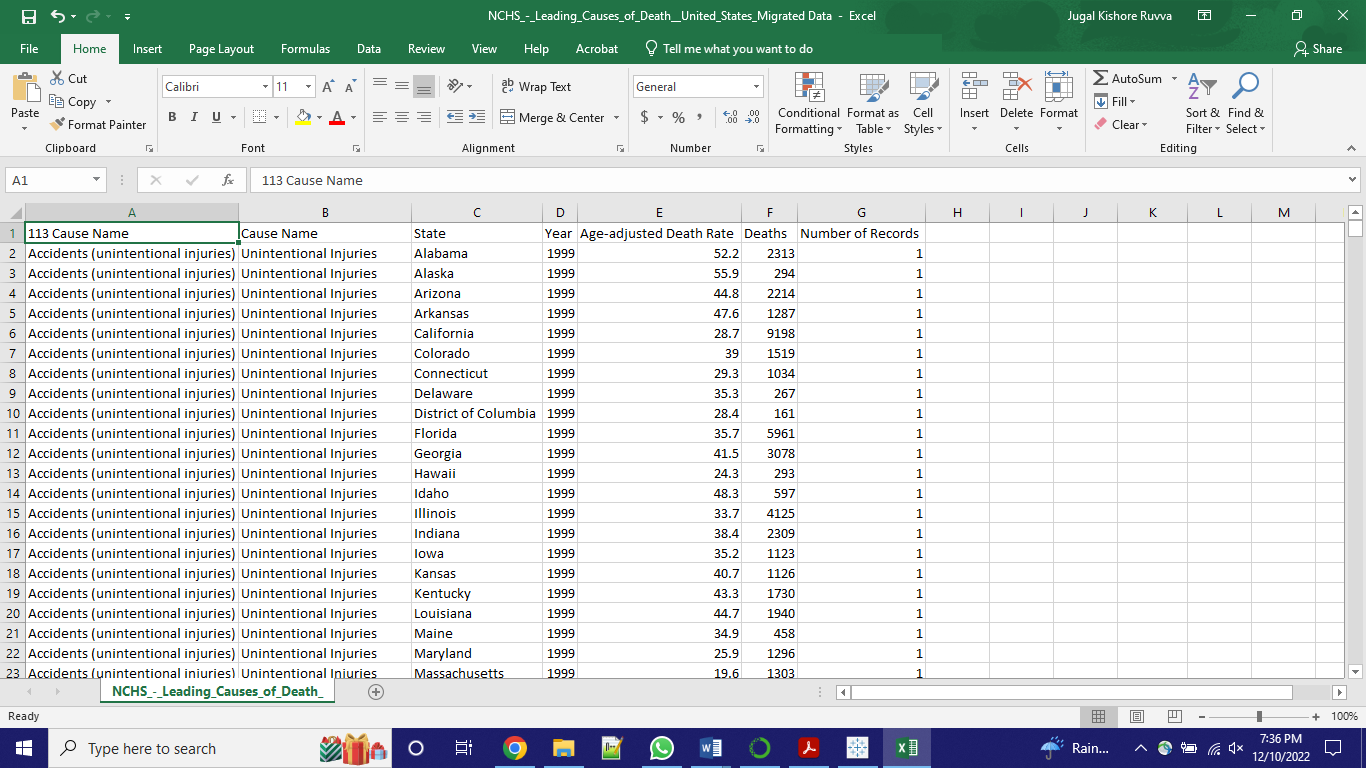
When planning the objectives, I want a geospatial graph to find which state has more deaths. So, I referred to some Tableau visualizations to understand how visualization techniques can help analyze the leading causes of death and came across a visualization done by [Hyerim Hwang](https://studentwork.prattsi.org/infovis/author/hhwang5/). The author used global data to plot the graph and visualize which country has more deaths. I took inspiration from his visualization and followed the suggestions that he mentioned in his blog to find the dataset.

**About the dataset:**

Data exploration and analysis require selecting the appropriate dataset to make predictions, which is a crucial step. After consulting the relevant research papers, I discovered that the analysis would not be insightful and valuable if the dataset were small. Therefore, finding a pertinent dataset for our study was a crucial task. For this project, I can find the “Leading Causes of Death in the United States” dataset, a public dataset from the National Center for Health Statics. It has detailed mortality tables showing deaths, causes of death, age-adjusted death rates, and the number of deaths by year and state.

**Data Collection:**

The oldest and most compelling example of inter-governmental data sharing in public health is the National Vital Statistics System. The NCHS collects and disseminates the official vital statistics for the country using shared relationships, standards, and practices. Contracts between NCHS and vital registration systems run by the various jurisdictions legally in charge of registering life-changing events like births, deaths, marriages, divorces, and fetal deaths result in the provision of these data. Online access to essential statistics data is also available. Legal responsibility for registering these events in the United States rests with each of the 50 States.



**Fig 1. Excel sheet downloaded from NCHS website**

**Column Description:**

The dataset draws data from one file with seven columns describing the cause of death, age-adjusted death rate, and the number of deaths in each state, with descriptions as follows.

**Generalized cause name:**

Type: String

Nullable: No

This column describes the cause of death in the broad category and has a mortality matrix code given by the health organization.

**Cause name:**

Type: String

Nullable: No

Some causes of death, including some with significant public health implications, are not given individual rankings because they fall under a larger category. For instance, lung cancer is covered under "cancer," and auto accidents are covered under "unintentional injuries."

**State:**

Type: String

Nullable: No

The data in this column describes each state in the United States of America.

**Year:**

Type: Whole Number

Nullable: No

This column describes the Years that are ranging from 1999 to 2015.

**Age-Adjusted Death Rate:**

Type: Decimal Number

Nullable: No

Age-adjusted death rate =  **X 1,000**

Compared to the normal population, age-adjusted death rates for various geographic areas or population groups are comparable.

**Deaths:**

Type: Whole Number

Nullable: No

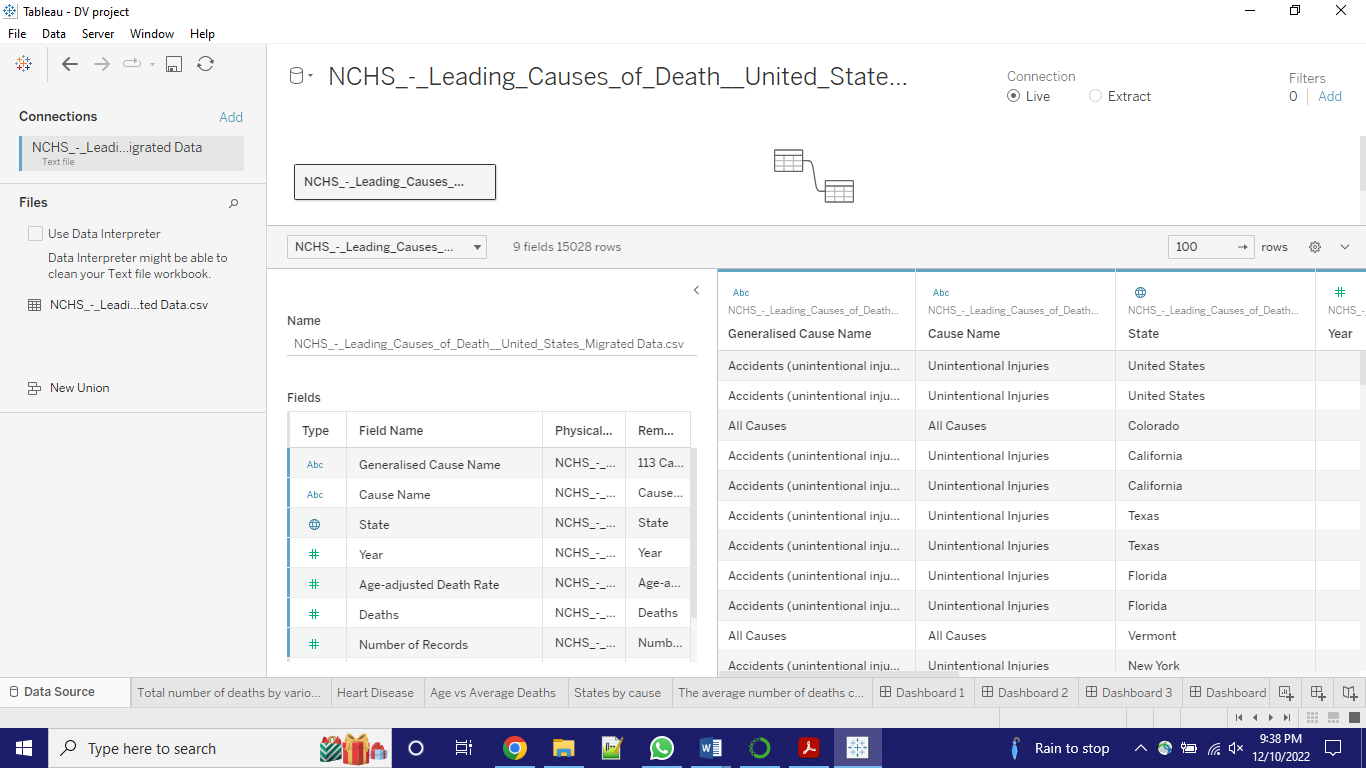
This column has data about the total number of deaths by each cause.

**Number of Records:**

Type: Whole Number

Nullable: Yes

The data in this column describes the number of appearances of this record in the dataset.



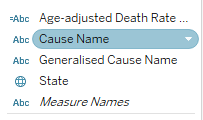
**Fig 2. The dataset loaded in Tableau**

# **METHODS**

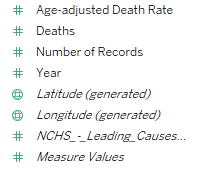
**­­­ Data Cleaning and Wrangling:**

There are seven columns with various datatypes, as seen in the data set above. In some cases, their values may need to be completed or relevant to our analysis. Data preparation and cleaning are crucial to obtain accurate results from the visualization.

I uploaded the data into Tableau and looked at Fig 3 and 4 for dimensions and measures respectively.



**Fig 3. Dimensions**



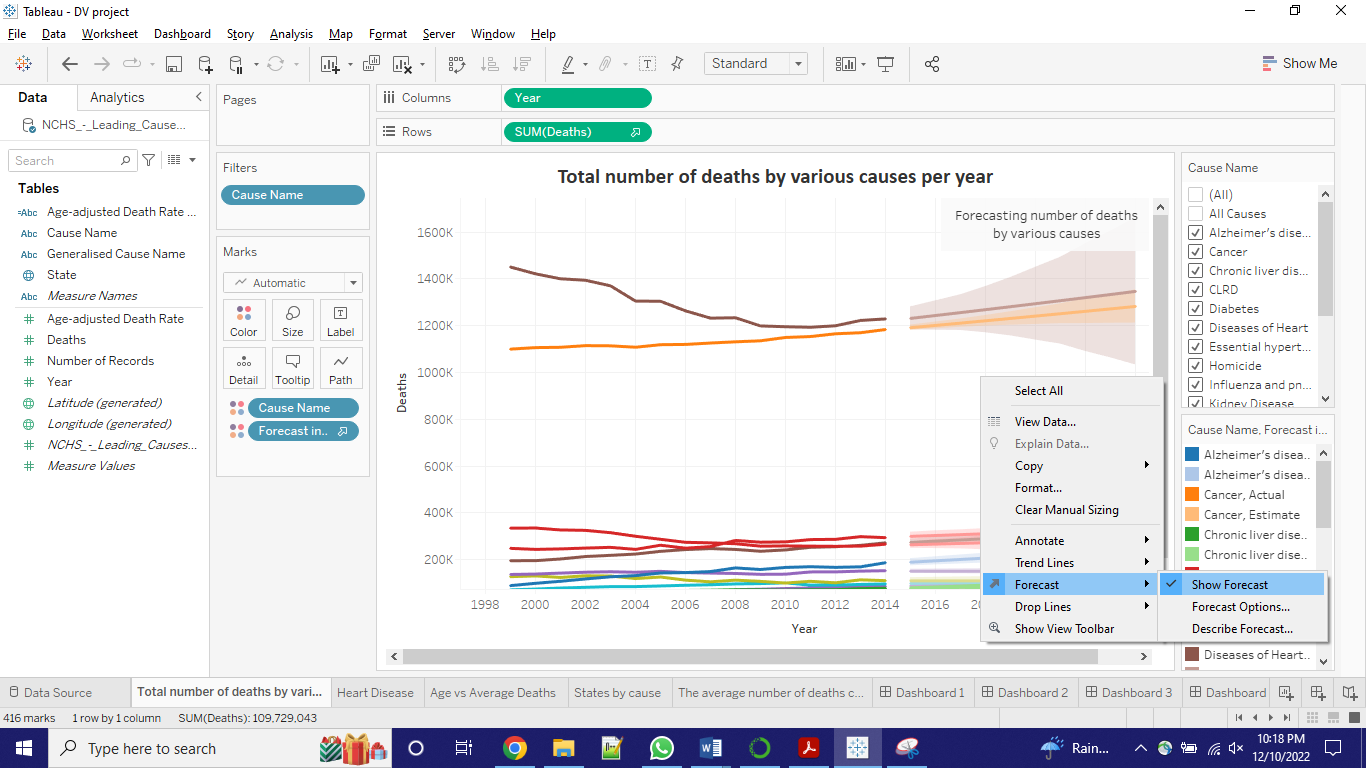
**Fig 4. Measures**

Firstly, I identified and eliminated columns with null values, and columns like ‘Number of Records’ can be skipped as this column is less critical for analysis. In Tableau, I only need to hide those attributes from the Measures. The Year column needs to duplicate into dimensions, as that will help in some visualizations. The other parts of this dataset are straightforward and well-prepared. I used them directly for our visualization.

**Visualization Design:**

The issue around the cause, number of deaths, location, and year is complicated. To address this issue from all possible angles, I divided this project into several parts: year-wise, geographical distribution, and perspective of the state.

For forecasting, Tableau has a function that helps to forecast from existing values. First, I need to plot Deaths vs. Years to get a line graph that describes the number of deaths that happened over time because of different diseases. Then I can use the forecasting option (see Fig 5) to predict the death toll in coming years.



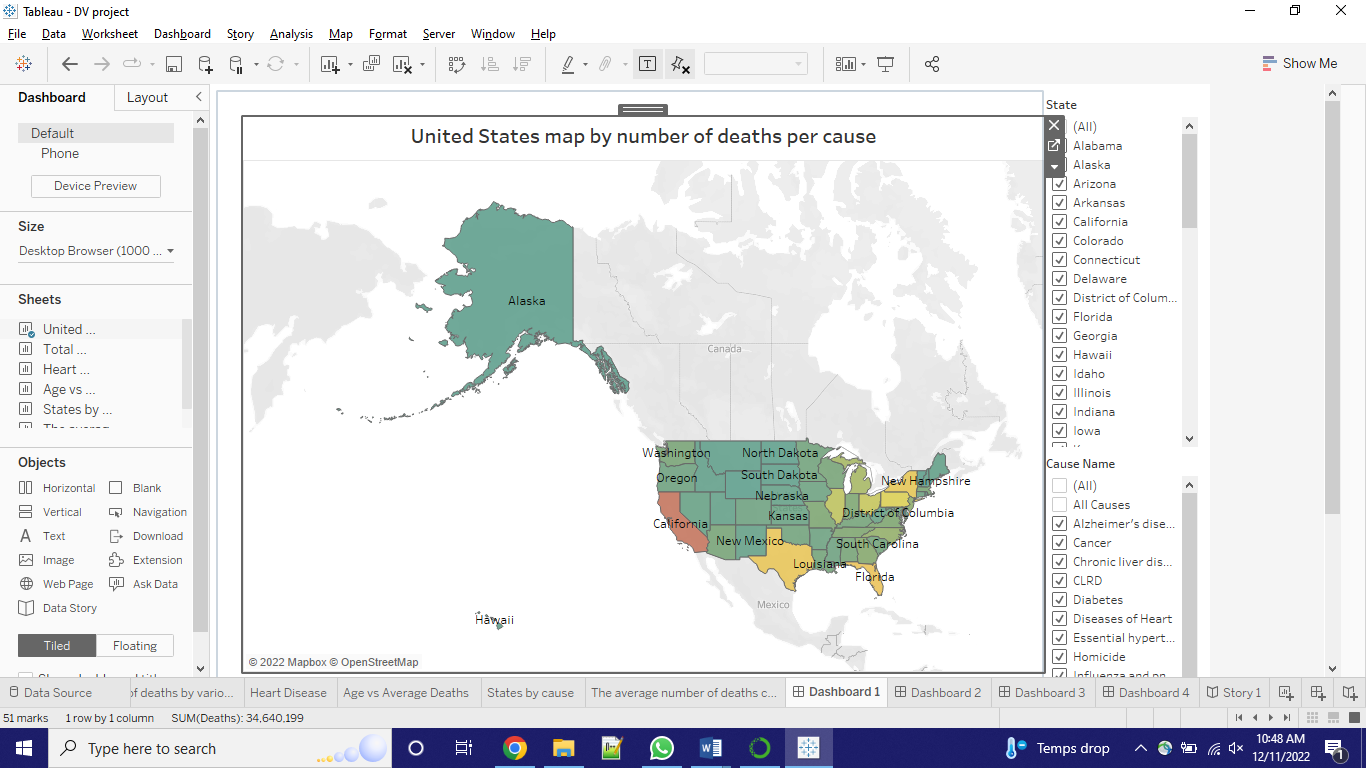
**Fig 5. Using Forecasting function**

**Tableau Operation:**

Creating an interactive dashboard was the project's initial goal for data visualization. I must create multiple dashboards because there are so many factors that need to be considered. Tableau's Story successfully met our needs. Creating a single sheet is the first step. Here, I primarily cleaned and filtered the data. For instance, I didn't include null values in the graphs.

Grouping related sheets onto a single dashboard is the next step. For users to see correlations or cause-and-effect relationships, I wanted to compile related information onto a single page. For improved visualization, I also carefully considered the style and arrangement of the graph.

Making a storyline to connect all the dashboards is the final step. Our dashboards each have a single feature. We could weave them into a compelling narrative that would guide the user to the conclusion. When users delve further and further, this is the best way for them to learn.

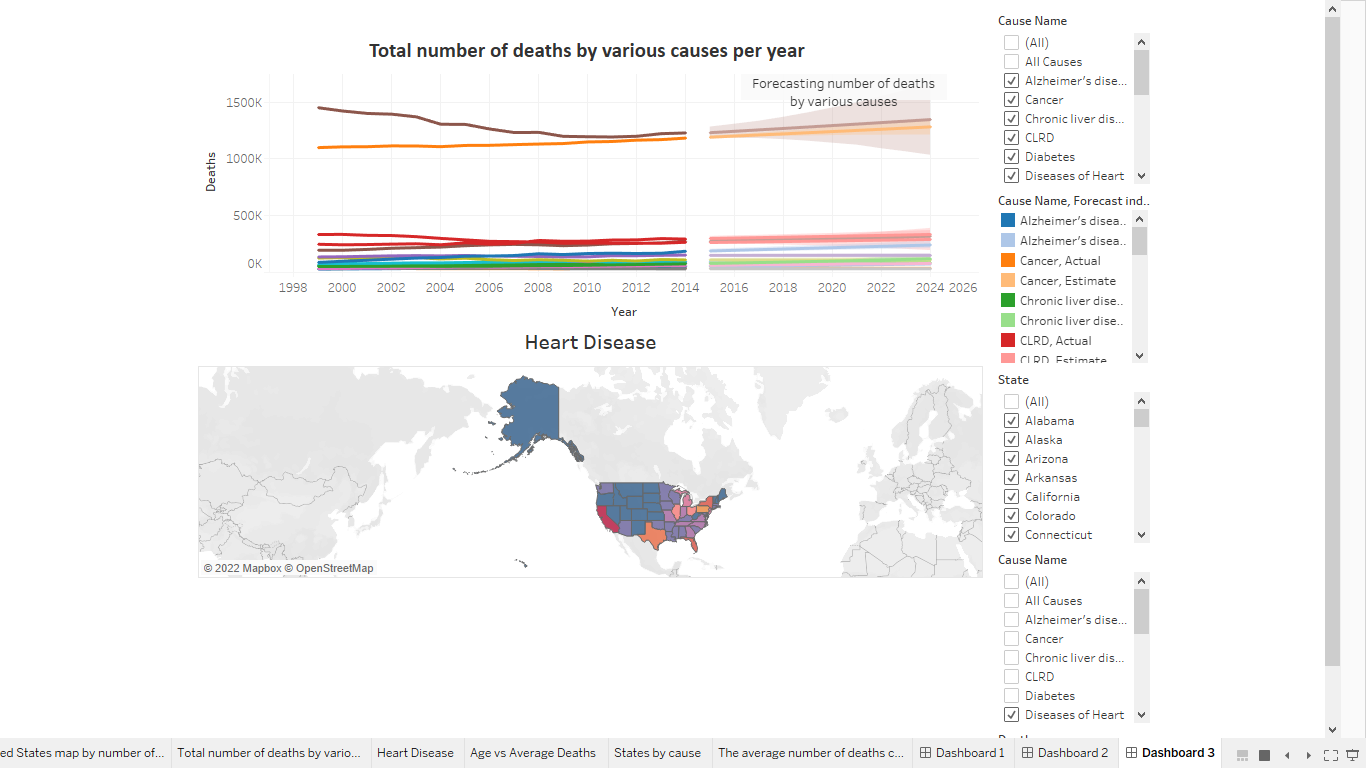


**Fig 6. Showing multiple dashboards and a story**

# **RESULTS**

My result in the Tableau story contains six worksheets that address all the problems from top to bottom.

1. Geographic death trend
2. The U.S death trend and forecasting
3. Exploring the leading cause
4. Death age analysis
5. State-wise death rate analysis
6. Time analysis

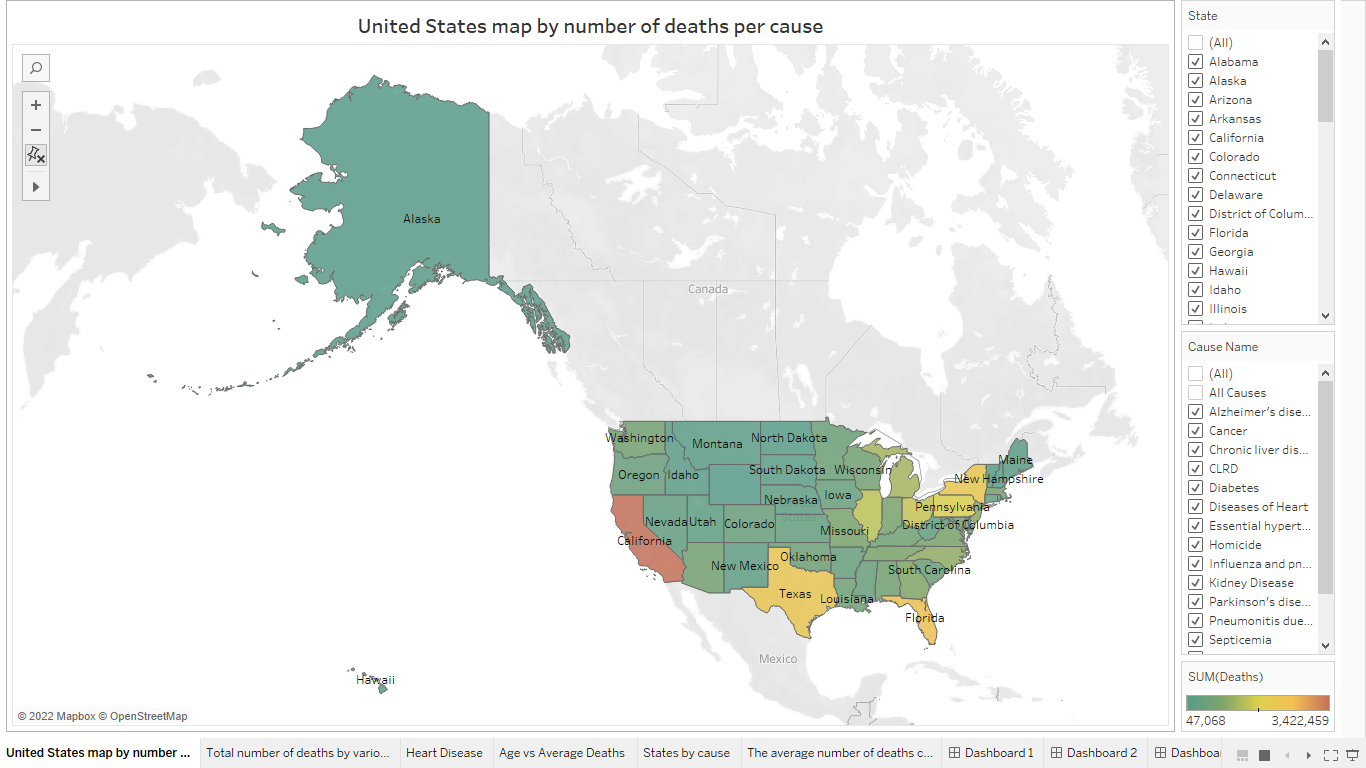


**Fig 7. All six worksheets, four dashboards, and a story**

1. Geographic death trend

The below figure, i.e., Fig 8, shows the first worksheet that gives an overview of deaths in the United States. I used a geo-spatial map to show the density of deaths from 1999 to 2015 in different states of the United States. To visualize the sum of fatalities in each state, I used temperature color grading, red being the severe ones (more deaths), yellow being average, and green being lax (fewer deaths) which can see at the bottom of the workbook.

The total number of deaths in California is around 3.5 million, which is very high when compared to the remaining states, so it is wrapped in red color. Alaska state is in a dark grade of green as the total number of deaths in that state is meager, as much as 47 thousand.



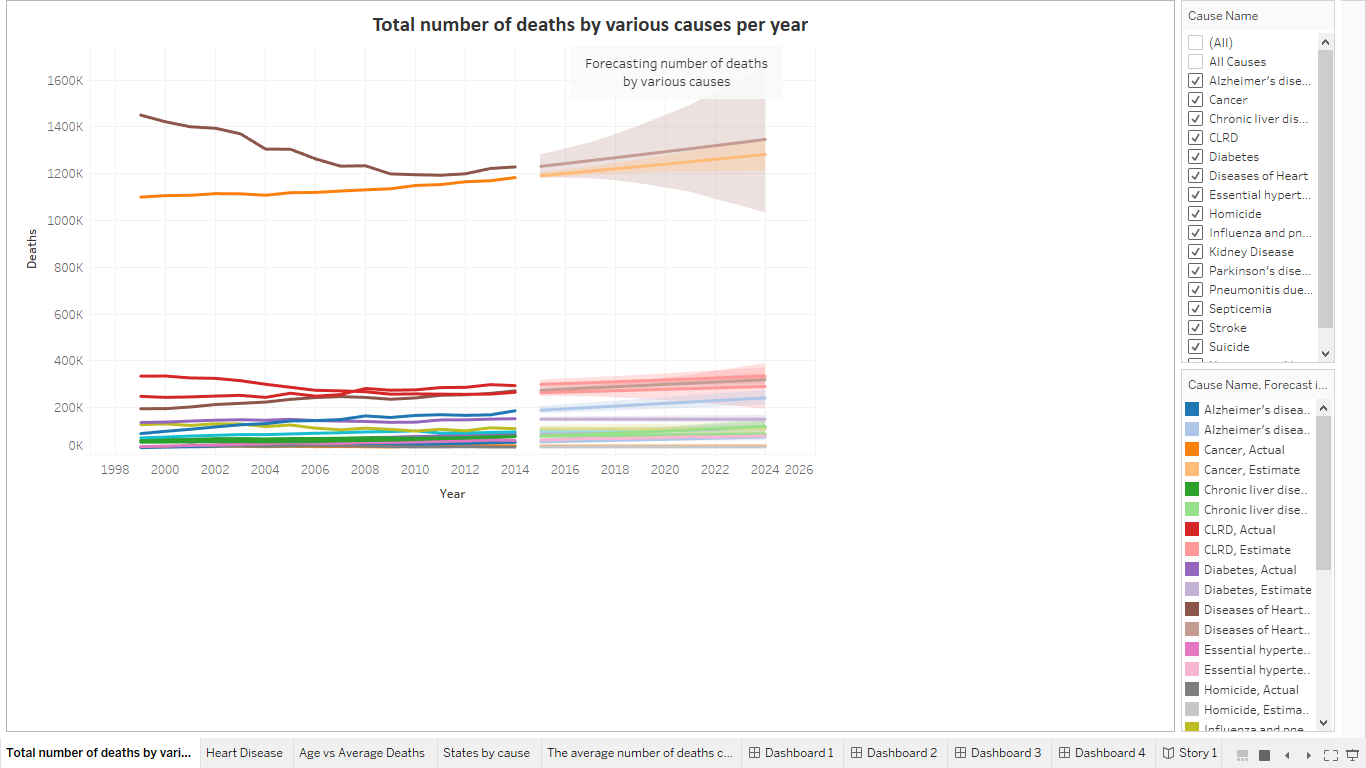
**Fig 8. Geographic death trend**

1. The U.S death trend and forecasting

In the following figure i.e., Fig 9, I used a line graph to visualize the total number of deaths by various causes in each year and forecast the death toll for coming years. I plotted Years on X-axis and Deaths on Y-axis and filtered using the cause name.

Then right click 🡪 Forecasting 🡪 Forecasting options 🡪 Set forecast length as until 10 periods, so that we can see the forecasting until 2024.

The number of deaths by disease of heart and cancer is way more high compared to other causes, and out of them both, heart disease victims are more in number. The forecast predicts that the number will increase, and most will die because of heart disease and cancer.

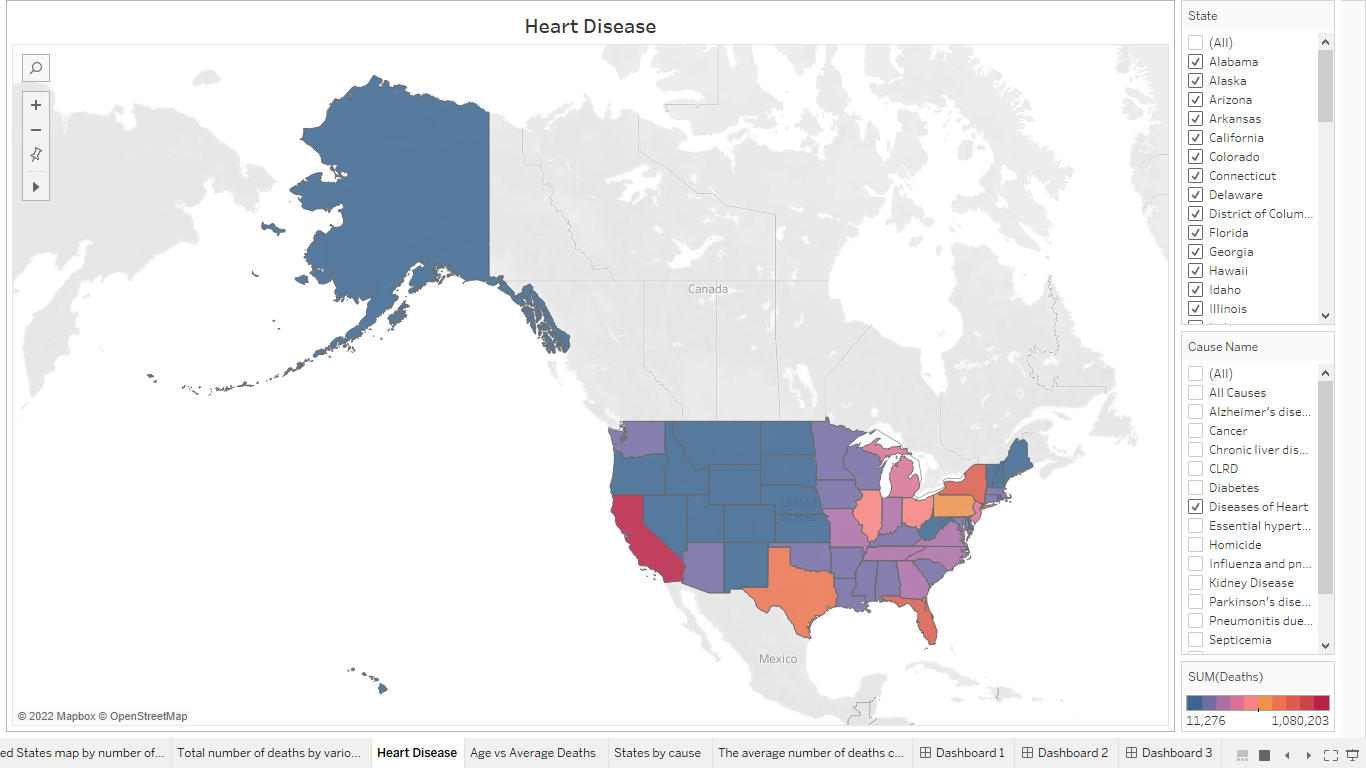


**Fig 9. The U.S death trend and forecasting**

1. Exploring the leading cause

As we can see from the previous worksheet, one of the leading causes of death is heart disease, so let’s explore more about it in this worksheet.

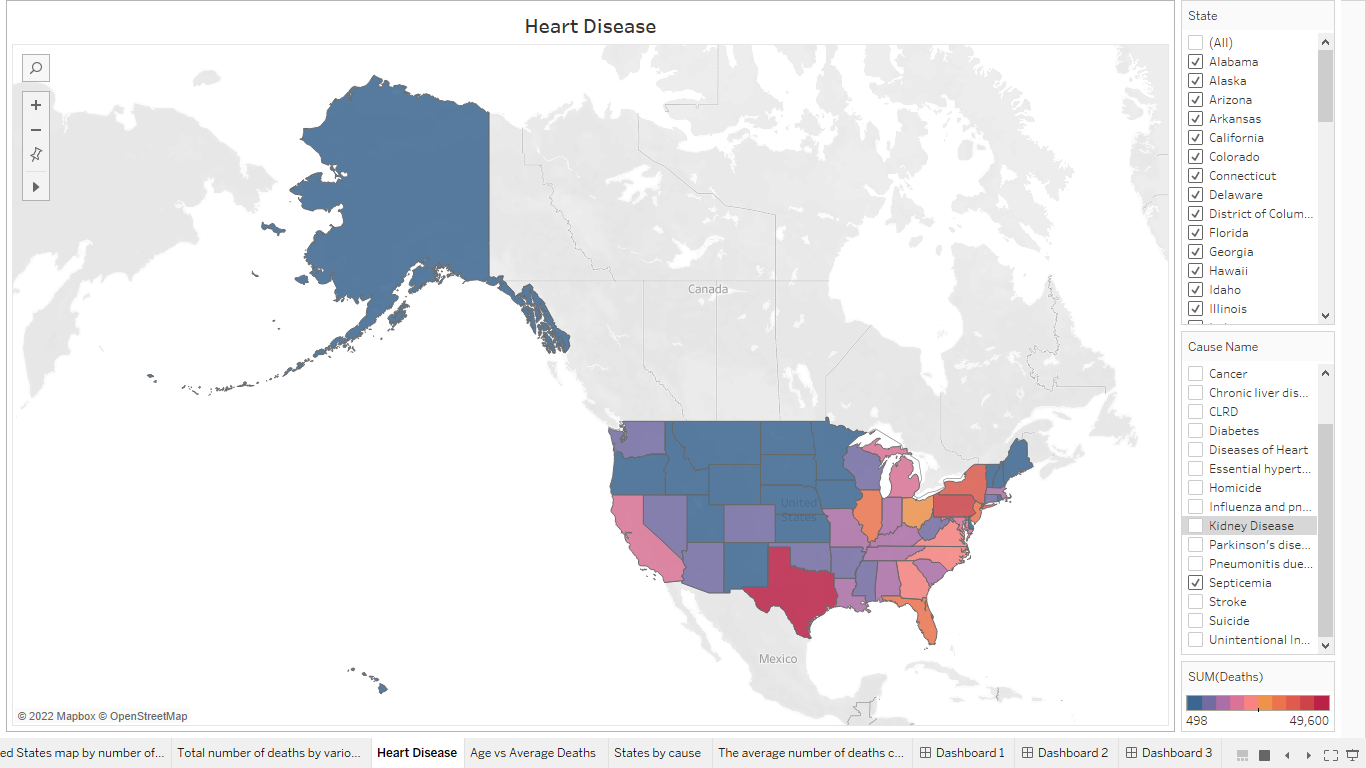
As shown in the below figure, i.e., Figure 10, this third sheet gives an overview of the geographic distribution of the disease of the heart. I dressed the visualization in color grading, with red being a high death toll and blue being a smaller number of deaths. Users can select different diseases to see the death toll distribution around other states of the United States.



**Fig 10. Effect of heart diseases**

It’s visible that the number of deaths because of heart disease is highest in California and higher in states in the west and south. Maybe this is because of the food habits or the lifestyle they lead.

If the user wants to see the effect of other causes, the user can select the other disease. In this worksheet, the user can select multiple diseases at the same time, to see their effect in different states of the United States of America.



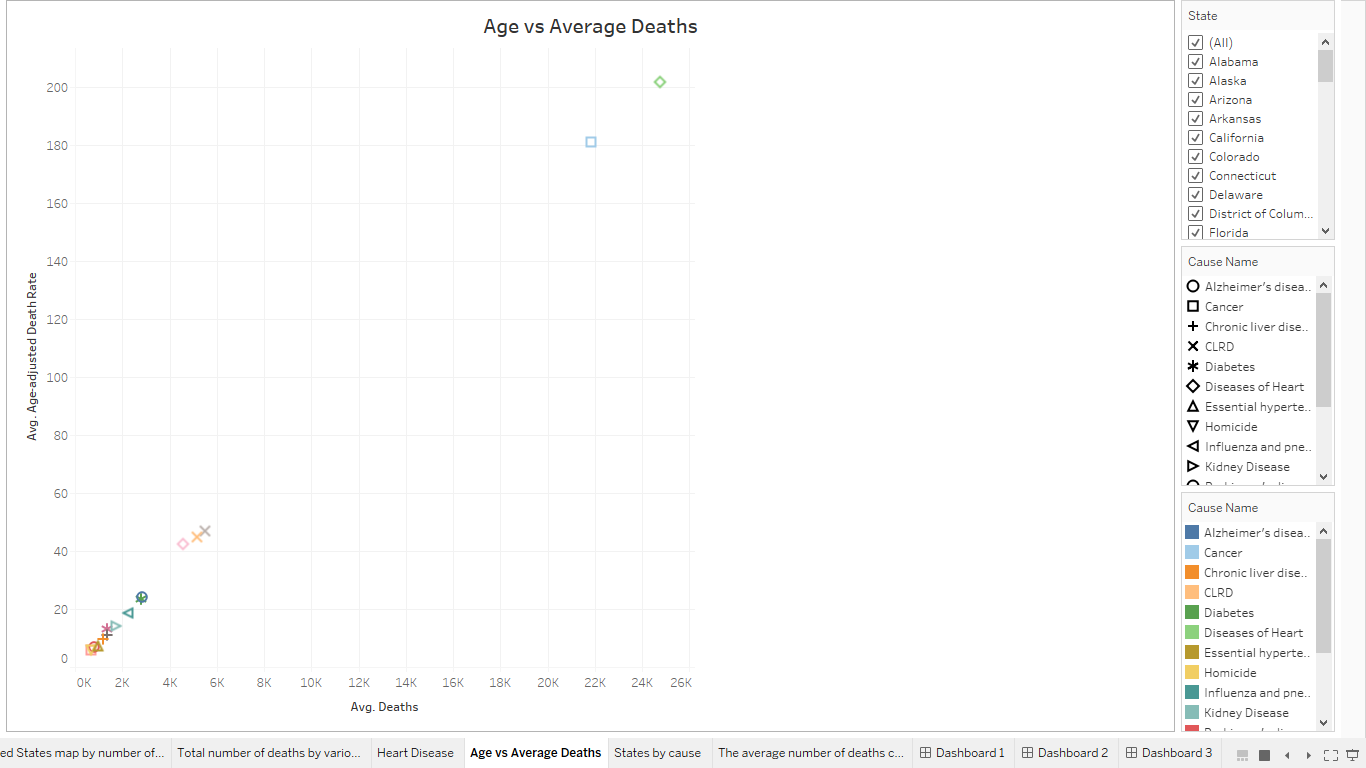
**Fig 11. Effect of Septicemia**

Observe figure 11. Here I selected the cause as Septicemia, and the color grade of each state in the graph is updated. As said earlier, a darker color indicates a high death toll, and Texas has a more significant number of deaths because of this disease than any other state.

1. Death age analysis

This workbook is about Age vs. Deaths analysis. I plotted this graph using a scatter plot expecting to find at what age people die because of different causes. I expected to see the effect of various reasons on different aged people. Each cause is represented in different shapes and colors in this graph. As this age is not an actual age but a calculated one, we can get an analysis report about only the death rate.

Age-adjusted death rate =  **X 1,000**.



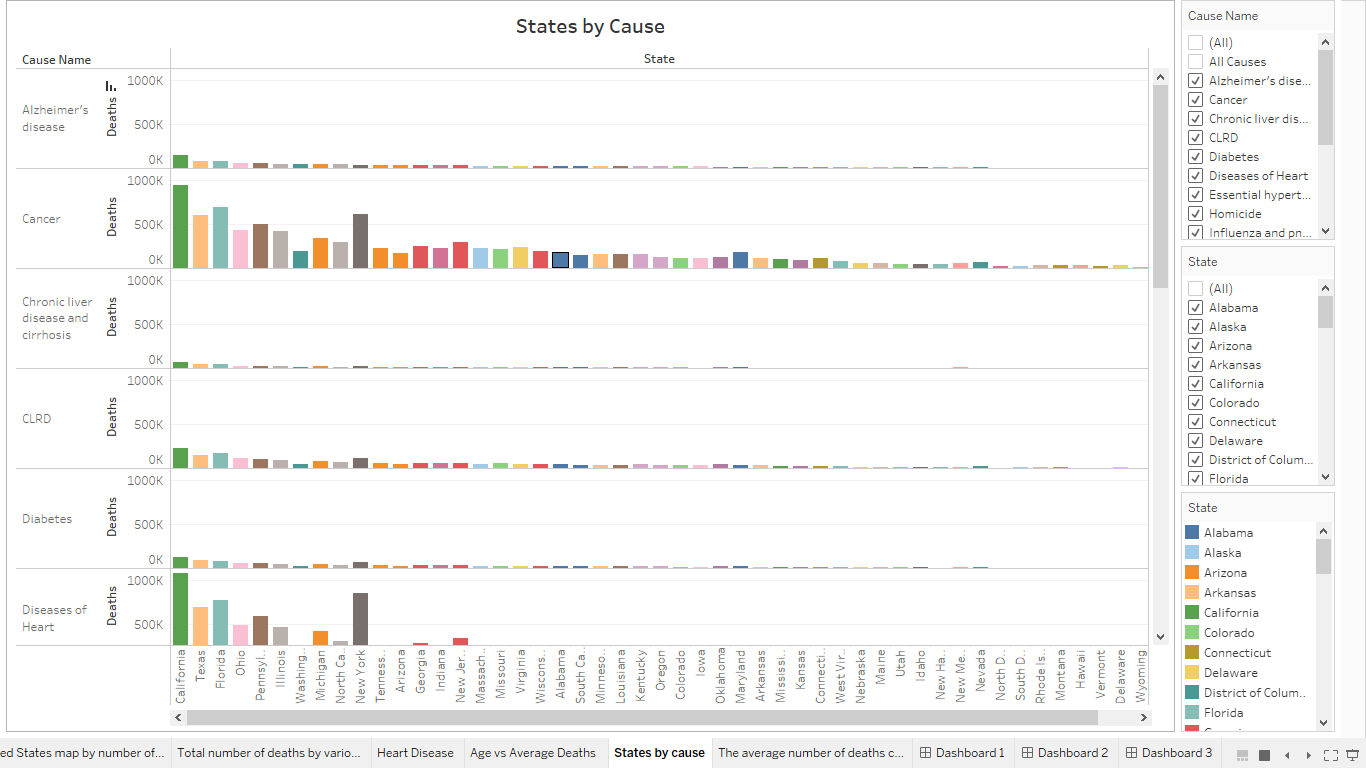
**Fig 12. Death Age analysis**

1. State-wise death rate analysis

In this worksheet, I went into a deep analysis of the effect of each disease in each state over the years. In this graph, we can see the leading causes of death in each state.

I plotted the state as columns; for rows, I used the Cause name and sum of deaths to get a bar plot. Each cause is displayed in different colors and can be seen at the bottom right of the screen.

Here, I sorted the graph by the number of deaths from high to low. In the below figure 13, after sorting the top 3 states with the highest death toll are California, Texas, Florida, and New York.

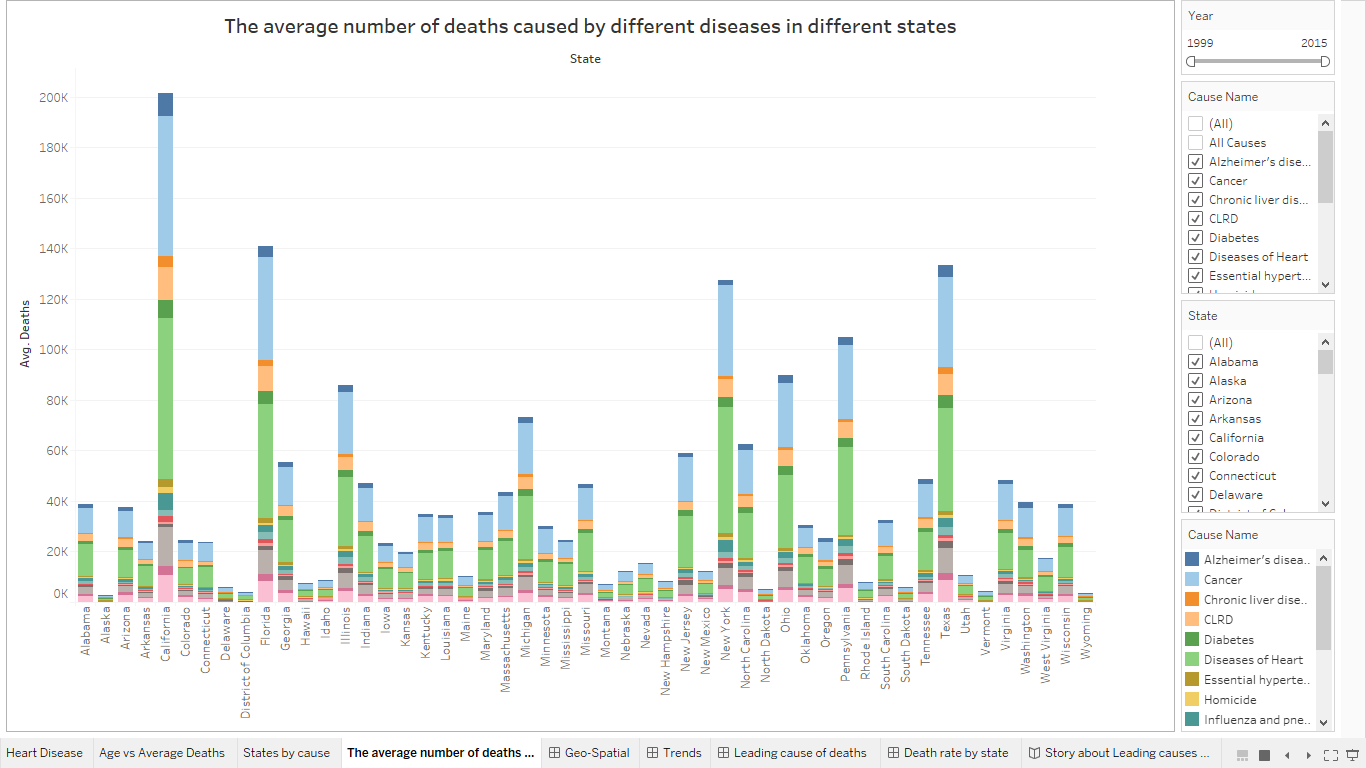


**Fig 13. The state-wise death rate analysis**

1. Time analysis

As we go to the last worksheet, I did a time-scale analysis of the deaths that happened from the year 1999 to the year 2015. I hope to see whether the number of deaths by various will go up or down at a distinct year constraint. So, I used year as a filter to see the trend in deaths by various causes across different states of the United States. By changing the filter, we can see the death trend.

I used the whole dataset (Years ranging from 1999 to 2015) for this worksheet. This bar graph helps to show the trend between the number of deaths in each state by various causes and the year.



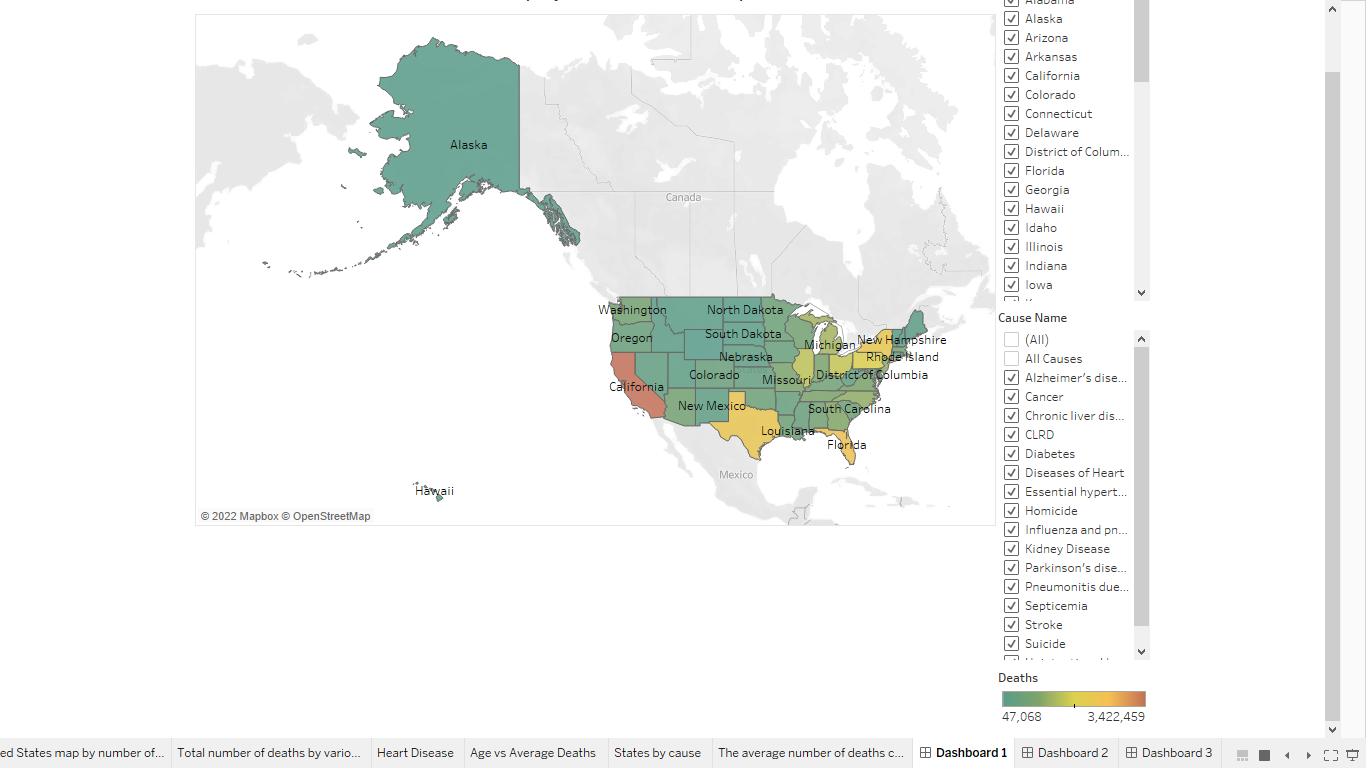
**Fig 14. Time Analysis**

# **DISCUSSION**

**What I learned and how I used it:**

Here are the dashboards that I developed for this project based on the knowledge I gained from this class.

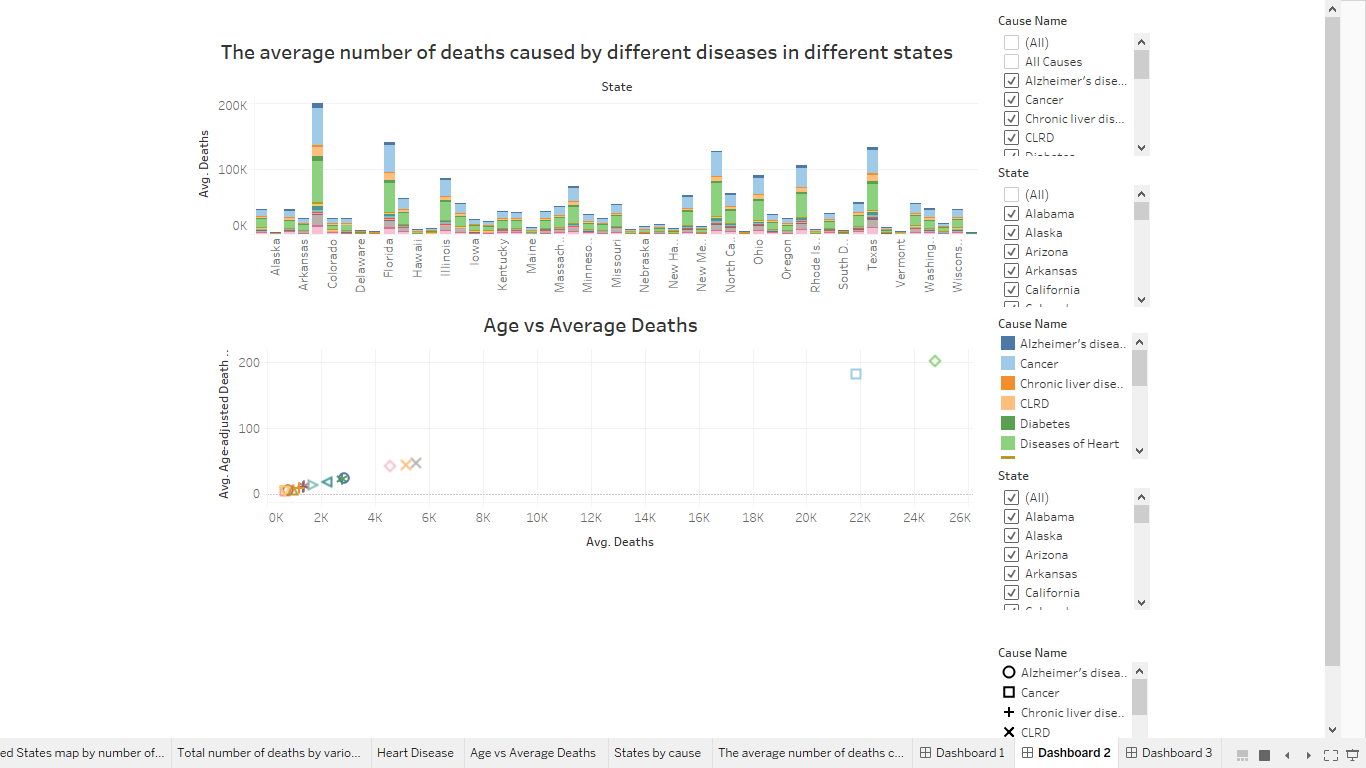
**Dashboard 1: Geo Spatial Dashboard**



**Fig 15. Geo-Spatial Dashboard**

*Distribution of the number of deaths across the USA*

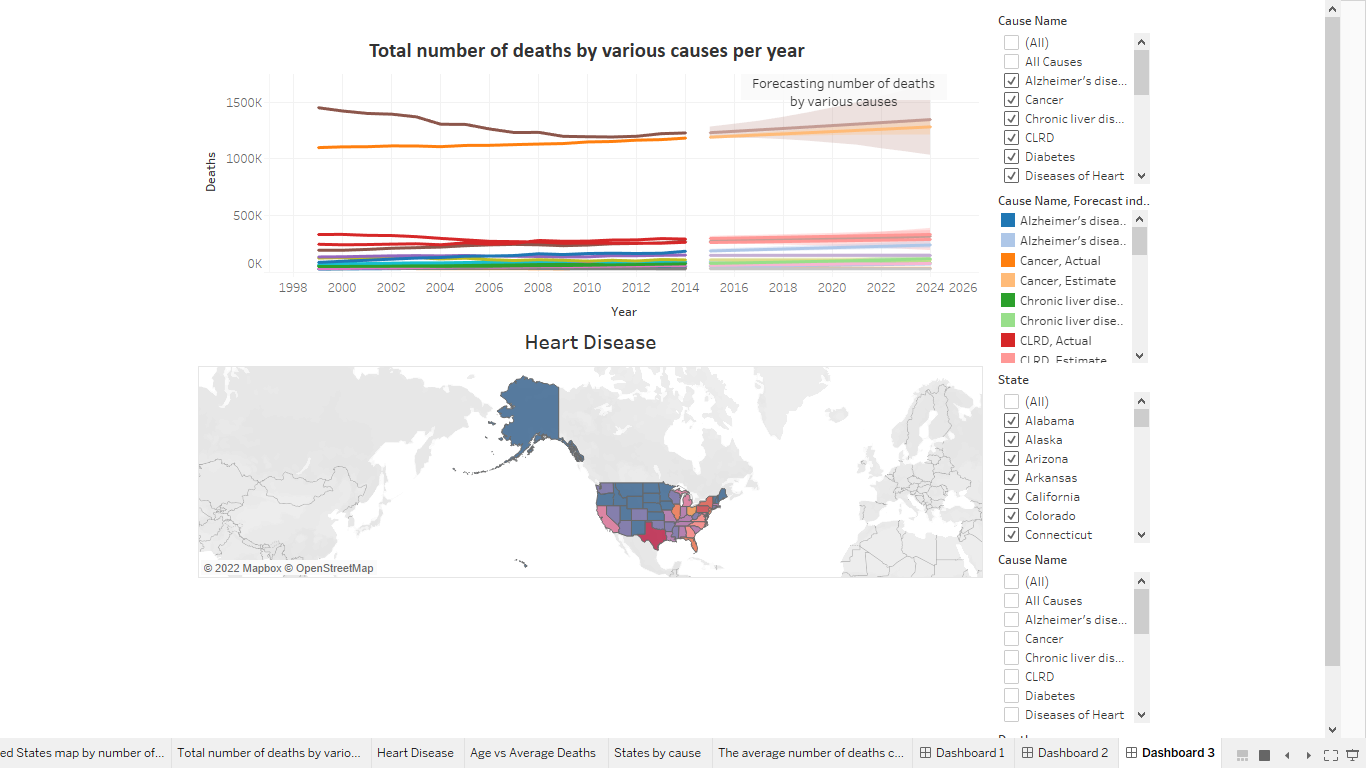
**Dashboard 2: Trend Analysis**



**Fig 16. Trend Analysis**

*(Exploring more about the factors that may have an impact)*

**Dashboard 3: Leading causes of death**



**Fig 16. Leading causes of death**

*(Total number of deaths by various causes and diving deep into the main leading cause)*

**Dashboard 4: Death rate by State**



**Fig 17. Death rate by state**

*(Distribution of the number of deaths by various causes in each state from 1999 to 2015)*

**Conclusion:**

I expect the audience to be informed of the leading causes of death in the United States. They should be aware of the importance of reducing the chance and severity of diseases. I also hope my suggestions can help the health department to reform the health and insurance plans. From the above graphs, we can conclude that the leading causes of death in the United States are heart disease and cancer. So, the people and health care providers should work hand in hand to develop better living conditions and treatments.

Other factors like geographic location, type of food they eat, age, and genes also affect life expectancy.

The major diseases in the world are cancers, cardiovascular diseases, and chronic respiratory diseases, which still are gradually climbing.

**Maybe a cost-effective, personalized health plan and insurance plans based on geographic location, family history, and type of diet will help to put a lid on this increasing number of deaths.**

**GitHub:** <https://github.com/kishoreruvva/Data230>

# **FUTURE WORK**

I briefly analyzed and visualized leading causes of death in the U.S. dataset with Tableau in this project. Using more advanced algorithms, we can provide further and deeper information through this large dataset with other analysis software such as Python. For example, we could use a regression model to calculate the risk index based on each factor we considered.

Even though we used forecasting which is just limited to forecasting the number of deaths in coming years, a future prediction may be another extension we could do based on this visualization project. For example, predicting which state, which year, and which disease will create havoc and become a pandemic killing most of the population. Thus, our health department could pay more attention to those circumstances, provide advanced warnings, and take preventive measures.

We can also create web platforms, such as a smartphone app, with the appropriate analysis and visualization to provide real-time updates. For further implementation of this project, we can dive deep into finding out why people from one state are dying more than beside states because of a cause.

# **REFERENCES**

1. Data source - National Centre for Health Statistics

<https://www.cdc.gov/nchs/nvss/leading-causes-of-death.htm#publications>

1. Dataset and related documentation – National Centre for Health Statistics

<https://www.cdc.gov/nchs/nvss/mortality_methods.htm>

1. Hyerim Hwang, CAUSE OF DEATH DATA VISUALIZATION

<https://studentwork.prattsi.org/infovis/visualization/cause-of-death-world-2017/>

1. Hyerim Hwang, World Map by Death Rates, 2017

<https://public.tableau.com/app/profile/hyerim.hwang/viz/CauseofDeath-V/CauseofDeath>

1. Anirudh Pillai, Leading Death Causes

<https://public.tableau.com/app/profile/anirudh1102>