Data Visualization of Project Tycho

Afzal Sufiya(22108480)

Table of contents

1 Introduction

- Health, defined by WHO, includes physical, mental, and social well-being.
- The US ranks 19th globally in COVID-19 vaccination rates. The USA has faced outbreaks like Scarlet fever, Typhoid Mary, H1N1 flu, Measles, and more. Immunization is mandatory in American schools since the 1850s.
- Climate change exacerbates health risks in the US, causing air pollution, wildfires, food and waterborne diseases, and a mental health crisis.
- Case surveillance helps control disease spread. CDC conducts surveillance through NNDSS, with 3,000 health departments contributing data.
- Our data US weekly Nationally Notifiable Disease Surveillance Data covers 1888-2013, including 50 diseases reported by 50 US states and 1284 US cities.
- This data enables analysis of disease trends and research. Our project analyzes this dataset, examining disease prevalence, impact on US states, and casualties.
- We use descriptive statistics and graphical methods to present our findings.

2 Aim

- We have a dataset on disease in the United States from 1888 to 2014.
- The objective of this analysis is to identify and rank the most significant diseases based on their prevalence. It aims to provide a comprehensive understanding of the diseases that have had the highest impact on public health.
- The objectives are as follows:-
  - Explore and preprocess the disease dataset.
  - Calculate the prevalence and morbidity rate for each disease.
  - Analyze disease prevalence on a state-by-state basis.
  - Provide clear visualization of the most affected diseases in different states.
  - Find the prevalence of these three diseases nationwide.
  - Analyze the distribution of deaths and cases for these diseases.
- By achieving these objectives, this analysis aims to contribute to the understanding of disease prevalence and prioritize prevention, control, and management strategies for the most impactful diseases.

3 Methods & Steps

- The R programming language provides us with a variety of convenient tools that allow us to transform our data into visually appealing and informative graphs. These graphs play a crucial role in helping us understand and interpret the data more effectively. Throughout our analysis, we have utilized several types of graphs to visualize the information, making it easier to comprehend.

Bar Plot

Bar plot or Bar Chart in R is used to represent the values in data vector as height of the bars. It is used to display the relationship between a numeric and a categorical variable.

Box Plot

A boxplot, also known as box and whisker plot, is a graphical representation which allows you to summarize the main characteristics of the data and identify the presence of outliers.

Heat-map

A heatmap is a graphical representation of data where the individual values contained in a matrix are represented as colors.

Tree-map

A Tree-map displays hierarchical data as a set of nested rectangles. Each group is represented by a rectangle, which area is proportional to its value.

Line Chart

A line chart or line graph displays the evolution of one or several numeric variables. Data points are connected by straight line segments.

Choropleth map

A choropleth map displays divided geographical areas or regions that are colored in relation to a numeric variable.

3 Methods & Steps

- Apart from that, here we used Morbidity rate as a factor to calculate the severity of diseases.
- The morbidity rate measures the portion of people in a specific geographical location who contracted a particular disease during a specific period. It indicates the frequency of the disease appearing in a population. The formula to calculate respective factor is :

$$Morbidity\ Rate\ (\%) = \frac{Total\ No.of\ case\ of\ disease}{Total\ Population} \text{ X } 100$$

- Additionally we used “Plotly” library for interactive graphs.
- Plotly provides online graphing, analytics, and statistics tools for individuals and collaboration, as well as scientific graphing libraries for Python, R, MATLAB, Perl, Julia, Arduino, JavaScript[1] and REST.

3 Methods & Steps

- Below mentioned R libraries has been used in our project:

  - library(tidyverse)
  - library(readr)
  - library(config)
  - library(data.table)
  - library(treemap)
  - library(treemapify)
  - library(plotly)
  - library(gganimate)
  - library(dplyr)
  - library(ggplot2)
  - library(ggpubr)
  - library(ggthemes)
  - library(ggridges)
  - library(stringr)
  - library(scales)
  - library(tibble)
  - library(tidyr)
  - library(magrittr)
  - library(forcats)

3 Methods & Steps

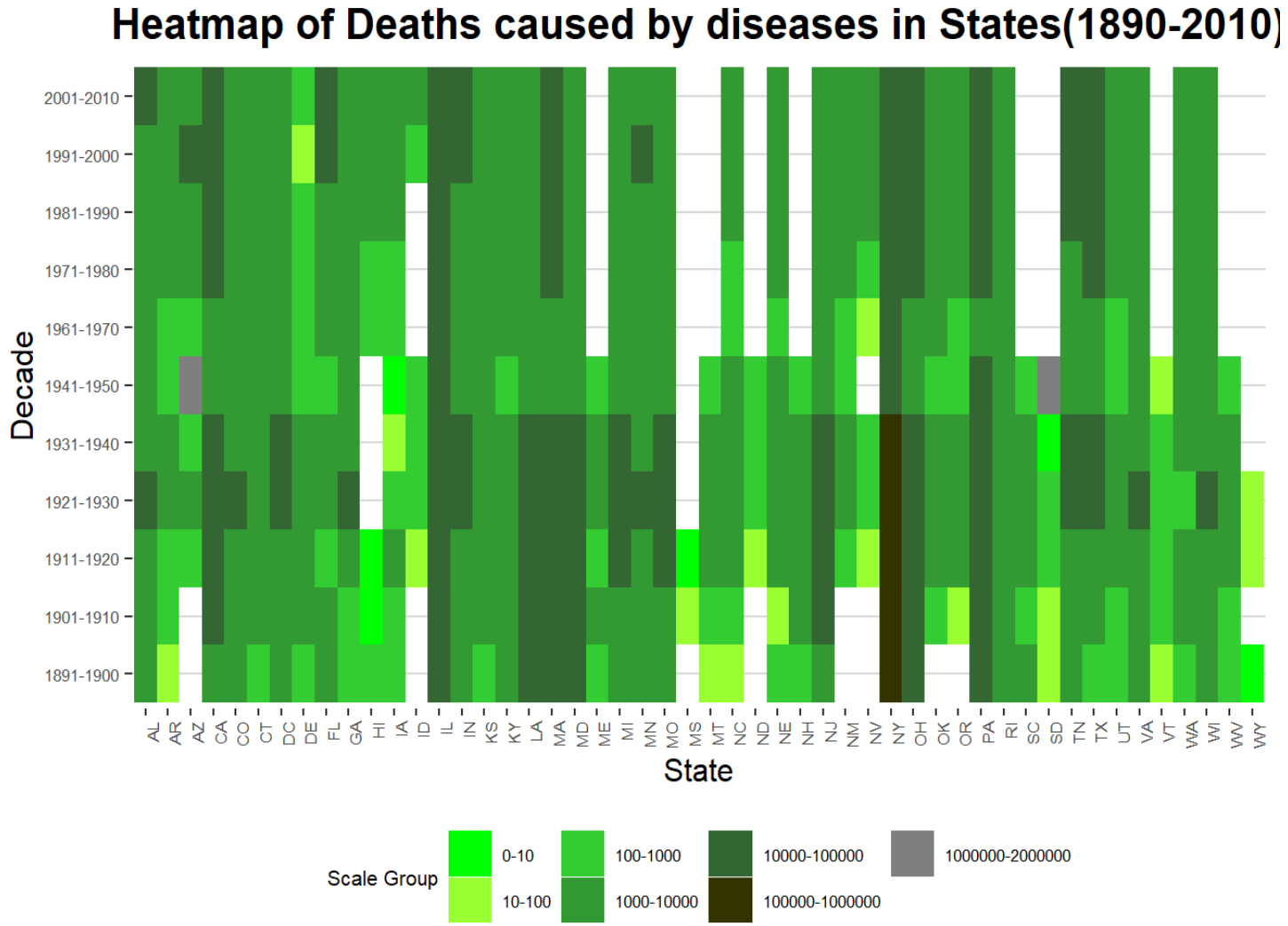
- We achieved the work following important methods of data visualization.

  - ✓ Data Manipulation
    - The vast data from project tycho is manipulated to get required attributes.
  - ✓ Data Analysis
    - With the help of R programming, the manipulated data is analysed thoroughly
  - ✓ Graph Plotting
    - By using various packages and tools available in R, various plots and graphs have been covering all important aspects.

4 Results

The severity of each disease by determining the frequency of occurrence

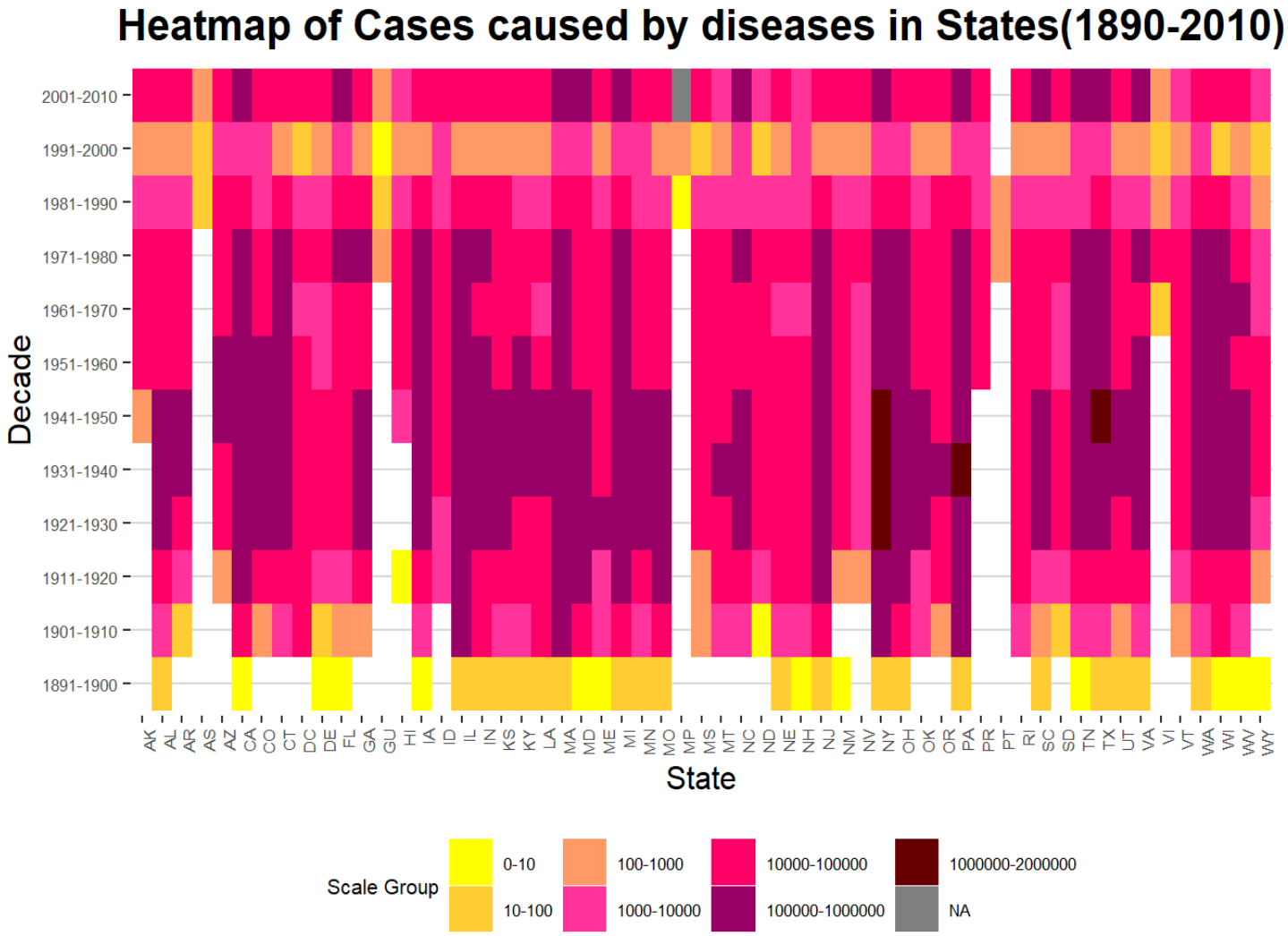
In this, we have created heatmap to analyze the occurrence of each disease in every decades. We create different plot for “DEATHS” and “CASES”.



4 Results

The severity of each disease by determining the frequency of occurrence

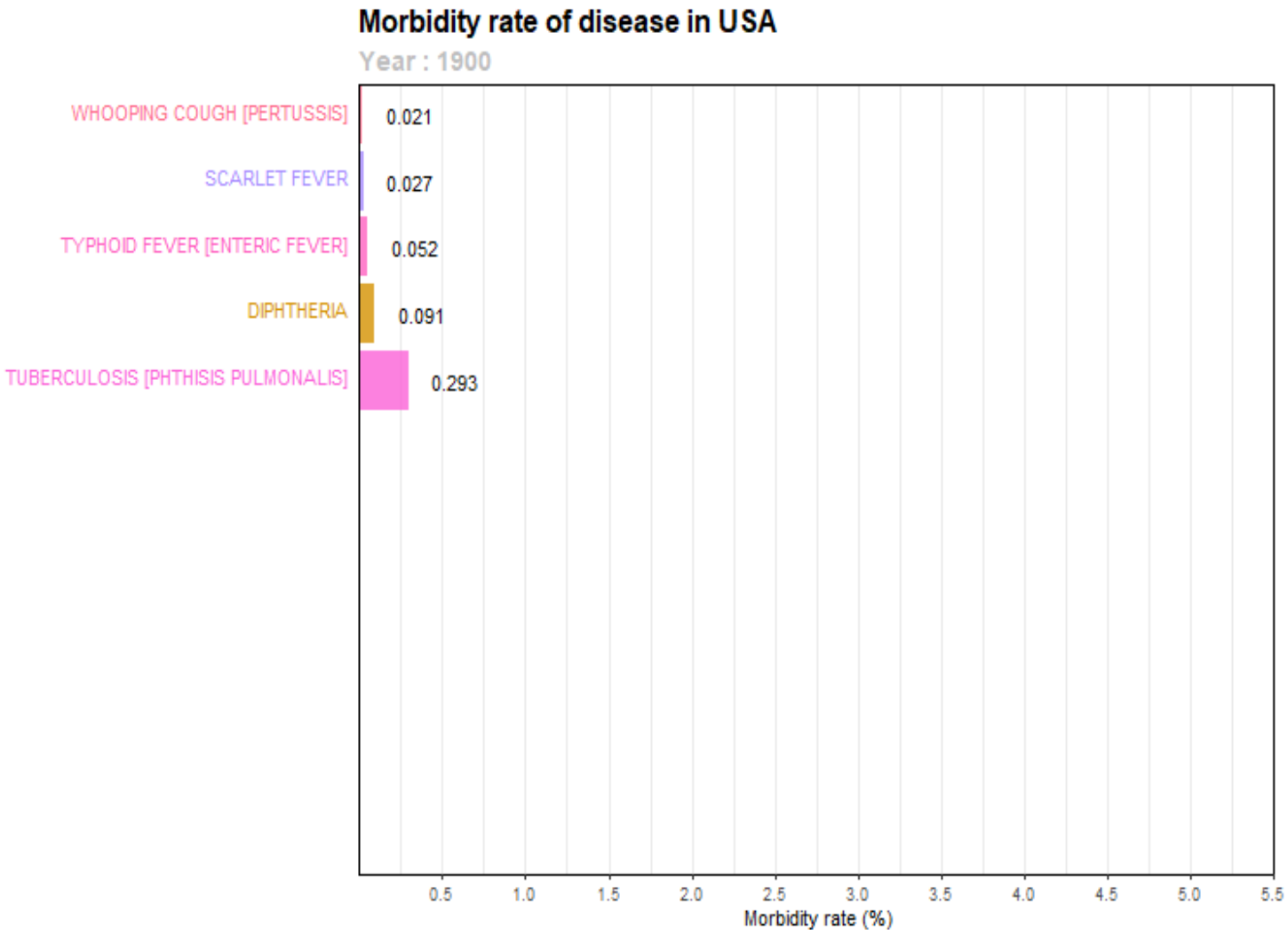
In this, we have created heatmap to analyze the occurrence of each disease in every decades. We create different plot for “DEATHS” and “CASES”.



4 Results

Morbidity rate of all disease with respect to census 1890-2010

In this, we have created animated racing bar plot to analyze the Morbidity rate of all time diseases of USA with different decades. (Census data only available in decades, So here we only can see decade wise plotting)

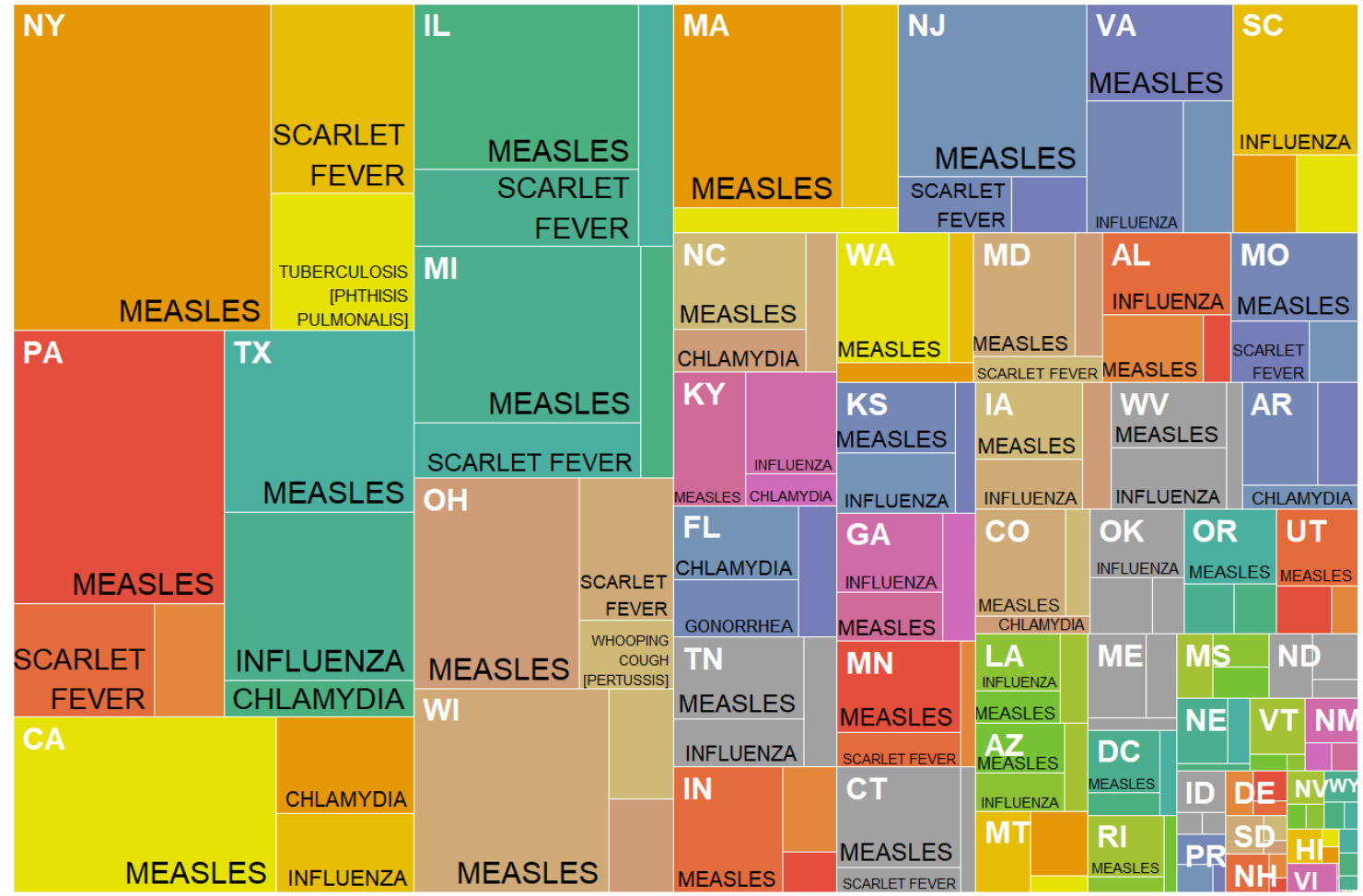


4 Results

Analysis of the top disease prevalence on a state-by-state basis

In this, we have created treemap to find top three most affected disease for all different states of USA. On every portion, top left corner shows state name and bottom right corner shows name of diseases.

All states with most affected three diseases



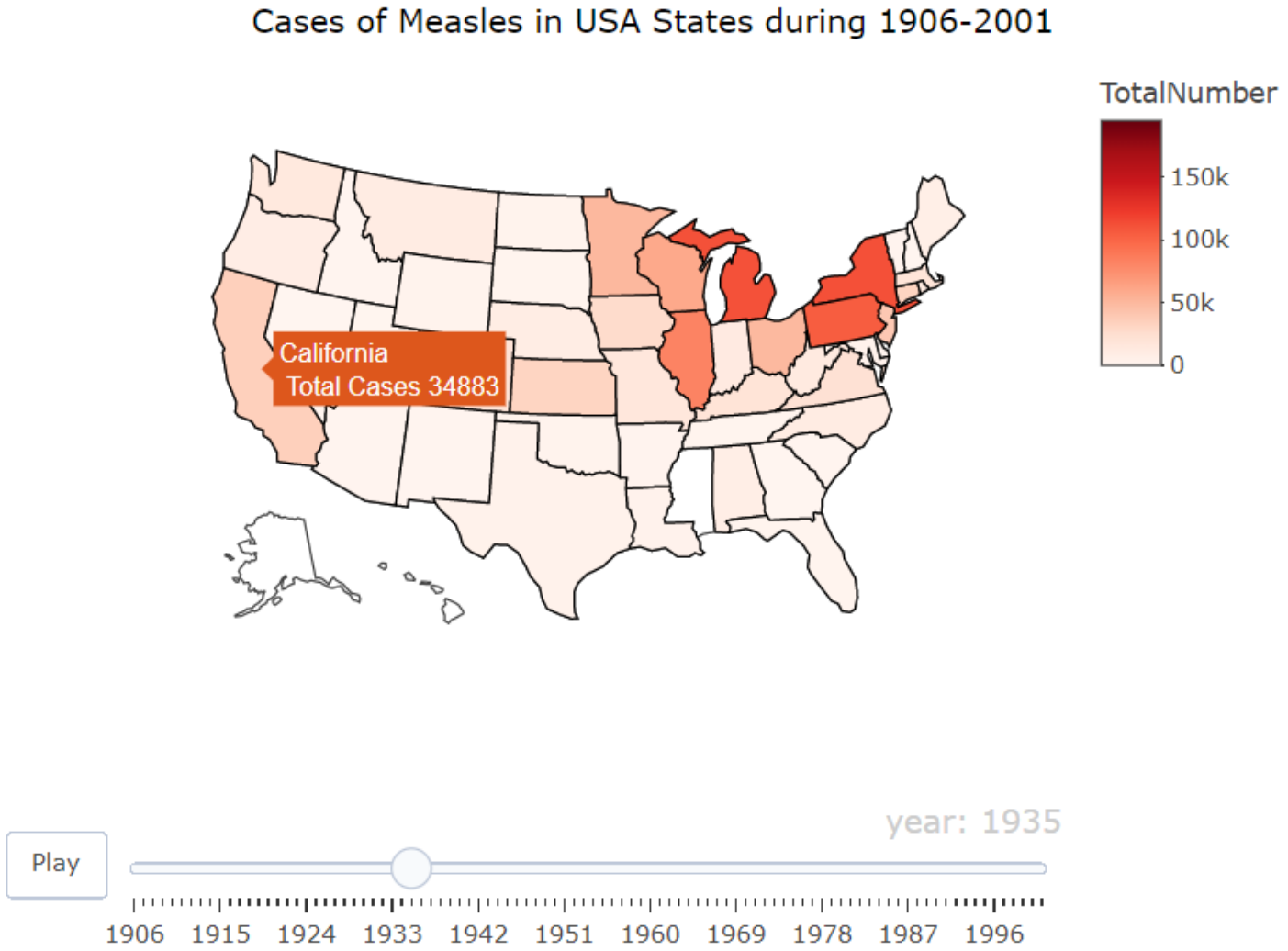
4 Results

Top disease prevalence on a state-by-state basis

By analyzing above data, we get to know that Measles, Scarlet Fever and Influenza were having high impact on USA health history. USA faced epidemic situation caused by these diseases. These diseases are infamous for their outbreaks in USA history.

This map shows Cases for **Measles**(As there are no death casualty by Measles).

In this interactive graph, We can check data for any year using handlebar below map and can get the case number of particular state by hovering cursor on it.
(Note : Cannot embed webpart here because of limitation of PowerPoint)



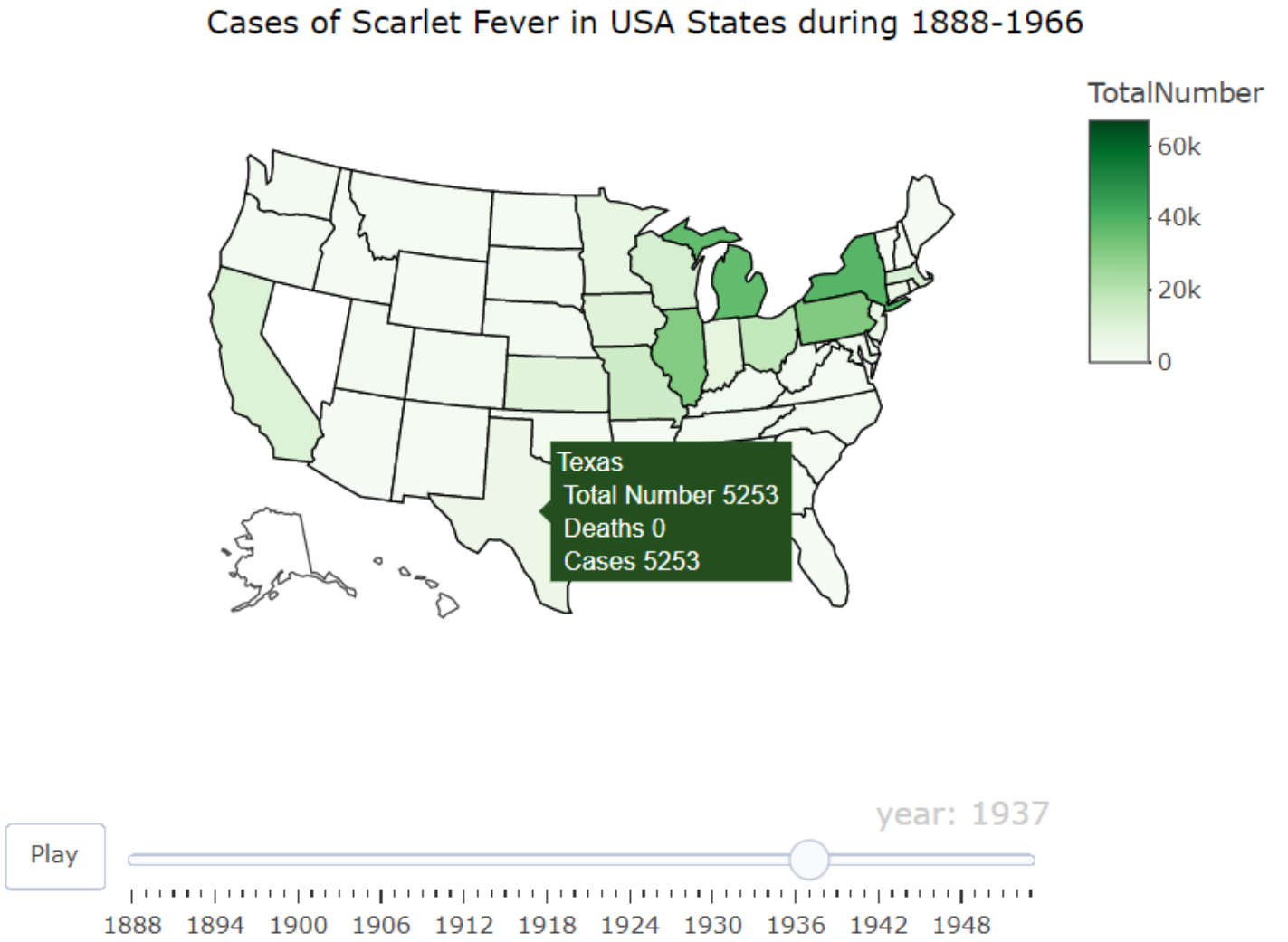
4 Results

Top disease prevalence on a state-by-state basis

By analyzing above data, we get to know that Measles, Scarlet Fever and Influenza were having high impact on USA health history. USA faced epidemic situation caused by these diseases. These diseases are infamous for their outbreaks in USA history.

This map shows Cases & Deaths for **Scarlet Fever**

In this interactive graph, We can check data for any year using handlebar below map and can get the case number of particular state by hovering cursor on it.
(Note : Cannot embed webpart here because of limitation of PowerPoint)



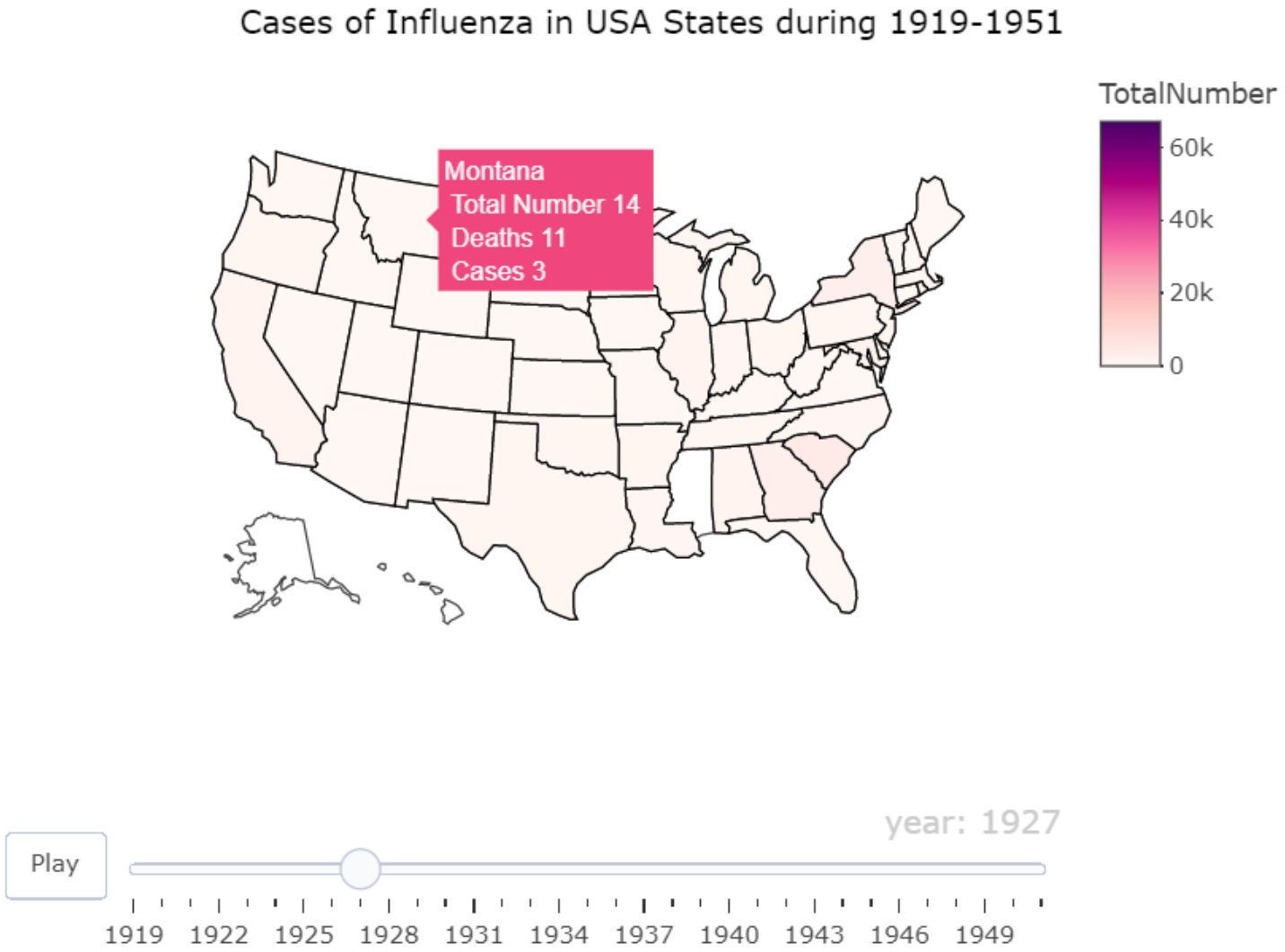
4 Results

Top disease prevalence on a state-by-state basis

By analyzing above data, we get to know that Measles, Scarlet Fever and Influenza were having high impact on USA health history. USA faced epidemic situation caused by these diseases. These diseases are infamous for their outbreaks in USA history.

This map shows Cases & Deaths for **Influenza.**

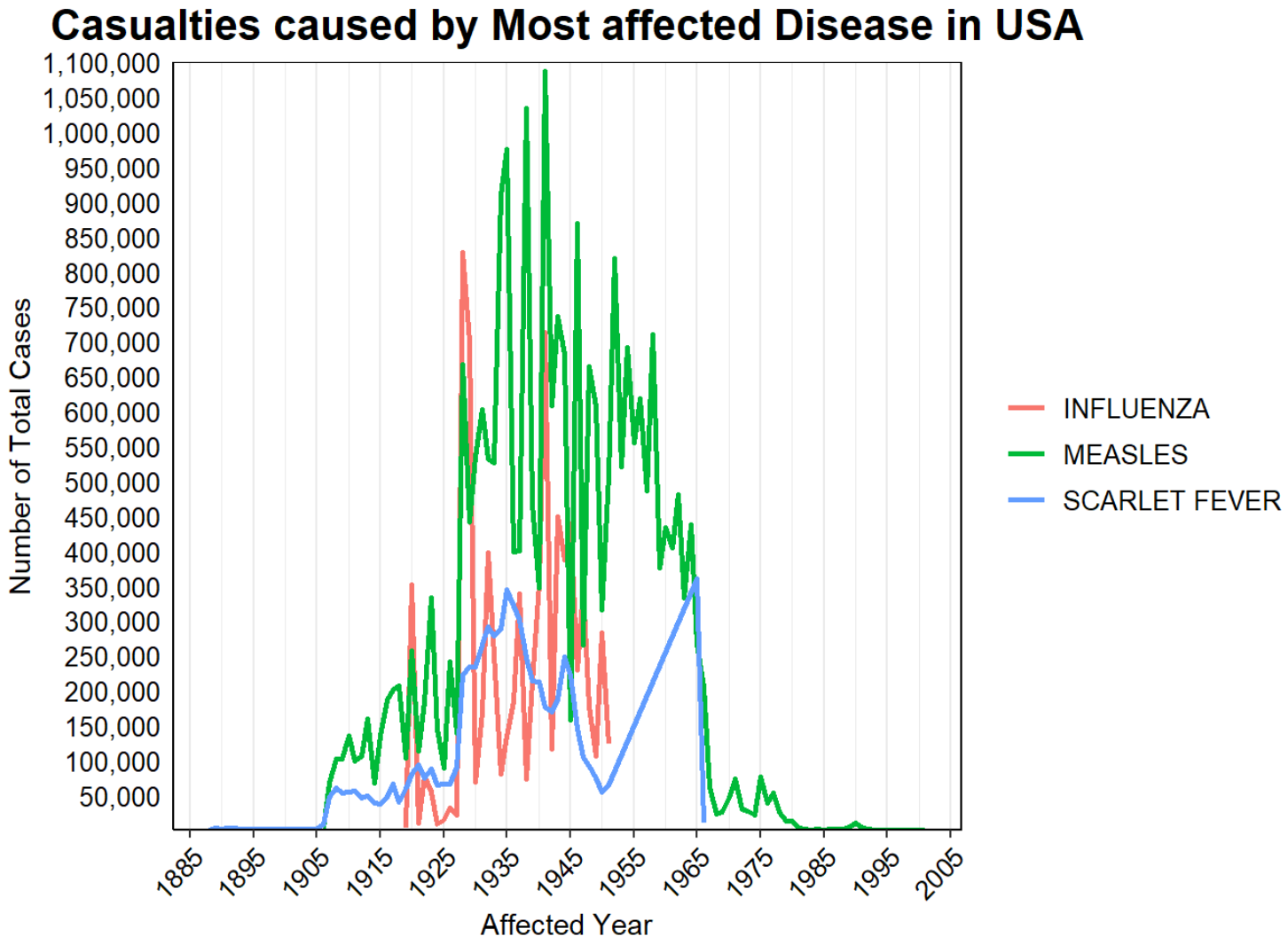
In this interactive graph, We can check data for any year using handlebar below map and can get the case number of particular state by hovering cursor on it.
(Note : Cannot embed webpart here because of limitation of PowerPoint)



4 Results

Prevalence of these three disease in all over USA

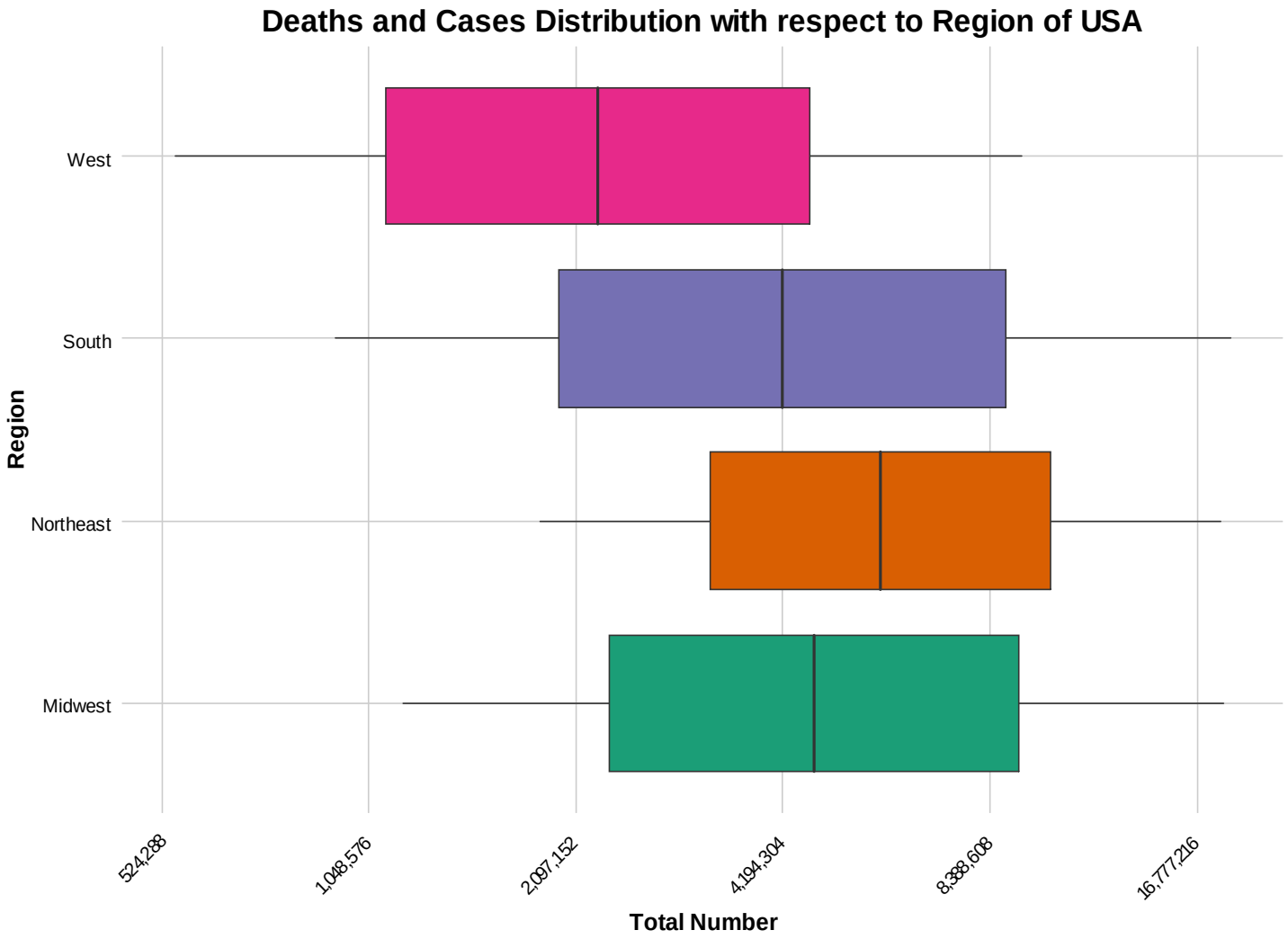
Here, we can see the prevalence of these most affected disease for all time.



4 Results

Distribution of Deaths & Cases of most affected diseases in different region of USA

To depict the distribution of Death and Case for these top diseases in different region of USA(Midwest, Northeast, South & West), we used box plot. Here, we can the distribution of these two unique values with respect to Region and Total Number following log2 trans in scales.

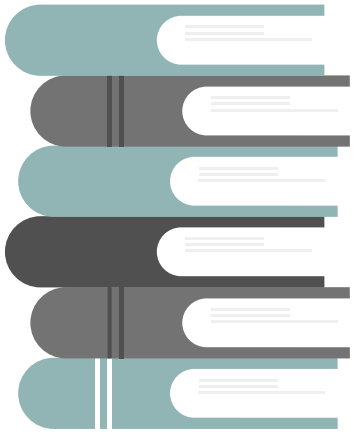


5 Conclusion

- The analysis of disease prevalence in the United States from 1888 to 2014 provides insights into the impact and distribution of health conditions across states. The data reveals trends in disease prevalence and identifies states with higher rates of specific diseases. Additionally, graphical representation simplifies complex data, allowing for accessible visualizations while maintaining accuracy and integrity.
- In conclusion, this study shows how using data visualization can greatly improve data analysis and communication in research. By visually representing information, researchers can better understand and convey complex data in an easier and more effective way. This helps others understand and engage with the research findings more easily. In conclusion, data visualization plays a key role in enhancing research by making data analysis and communication more accessible and impactful.

6 References

- Project Tycho: Home https://www.tycho.pitt.edu/
- U.S. Census Bureau, 2020 Censuses of Population and the population estimate program. https://data.ers.usda.gov/reports.aspx?ID=17827
- National Notifiable Diseases Surveillance System (NNDSS) https://www.cdc.gov/nndss/about/index.html
- Health in the United States (Wikipedia) https://en.wikipedia.org/wiki/Health_in_the_United_States
- Basic Statistics: About Incidence, Prevalence, Morbidity, and Mortality - Statistics Teaching Tools https://www.health.ny.gov/diseases/chronic/basicstat.htm
- United States census (Wikipedia) https://en.wikipedia.org/wiki/United_States_census
- The Worst Outbreaks in U.S. History https://www.healthline.com/health/worst-disease-outbreaks-history
- US States to Abbreviations https://www.kaggle.com/datasets/justinrwong/us-states-to-abbreviations
- gganimate: Getting Started https://gganimate.com/articles/gganimate.html
- Plotly r graphing library in R https://plotly.com/r/
- Morbidity Rate https://corporatefinanceinstitute.com/resources/wealth-management/morbidity-rate/
- US census bureau regions and divisions https://github.com/cphalpert/census-regions/blob/master/us%20census%20bureau%20regions%20and%20divisions.csv
- R Markdown https://rmarkdown.rstudio.com/index.html



Thank You