Leading Causes of Death in the United States, 1999-2017

CSIT 553: EXPLORATORY DATA ANALYSIS

- CECILIA KAMAU
- NOAH MENGICH
- TINOTENDA MUCHENJE

Dataset Description

- **Source:**
  - National Center for Health Statistics
- Dataset name;
  - Leading Causes of Death in the United States, 1999-2017
  - Retrieved from data.CDC.gov;
    - https://www.cdc.gov/nchs/data-visualization/mortality-leading-causes/index.htm
- **Description**:
  - The number of deaths and age-adjusted death rates for the 10 leading causes of death
  - All the causes of death combined, in the United States and by state for 1999-2017
  - Size of the dataset; 10869 * 6 (Rows * Columns)
- The scope of the project includes;
  - Exploratory data analysis
  - Map visualization
  - Aggregation visualization
  - Interactive visualization
  - Python program codes used in the project

Objectives

The research project seeks to analyze and visualize number of deaths and age-adjusted death rates for the 10 leading causes of death. All the causes of death are combined in the United States and organized by state for 1999-2017.

- To discover patterns and relationship between number of deaths and age-adjusted death rates for the 10 leading causes of death
- To conduct data importation, preprocessing, and the exploratory data analysis (EDA) using Python; analyzing and visualizing the relationship of variables using maps, aggregation, and interactive visualizations in the US

Importing libraries

```
#Import libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

import geopandas as gpd
import folium
from folium.plugins import HeatMap
from geopy.geocoders import Nominatim

from shapely.geometry import Polygon
import matplotlib.colors as mcolors

import plotly.express as px

import dash
import dash_table
import dash_html_components as html
import dash_core_components as dcc
import dash_bootstrap_components as dbc
from dash.dependencies import Input, Output
```

Loading and Understanding Dataset

```
#Load dataset
#From National Center for Health Statistics
path = r'/content/NCHS.csv'
df = pd.read_csv(path)

#Understanding the dataset
df.head(200)
```

	Year	113 Cause Name	Cause Name	State	Deaths	Age-adjusted Death Rate
0	2017	Accidents (unintentional injuries) (V01-X59,Y8...	Unintentional injuries	United States	169,936	49.4
1	2017	Accidents (unintentional injuries) (V01-X59,Y8...	Unintentional injuries	Alabama	2,703	53.8
2	2017	Accidents (unintentional injuries) (V01-X59,Y8...	Unintentional injuries	Alaska	436	63.7
3	2017	Accidents (unintentional injuries) (V01-X59,Y8...	Unintentional injuries	Arizona	4,184	56.2
4	2017	Accidents (unintentional injuries) (V01-X59,Y8...	Unintentional injuries	Arkansas	1,625	51.8
...	...	...	...	...	...	...
195	2017	Cerebrovascular diseases (I60-I69)	Stroke	Pennsylvania	6,700	36.5
196	2017	Cerebrovascular diseases (I60-I69)	Stroke	Rhode Island	425	29.4
197	2017	Cerebrovascular diseases (I60-I69)	Stroke	South Carolina	2,691	44.9
198	2017	Cerebrovascular diseases (I60-I69)	Stroke	South Dakota	414	36.7
199	2017	Cerebrovascular diseases (I60-I69)	Stroke	Tennessee	3,519	45.0

Dataset loading, Understanding dataset and Cleaning

```
df.info()
#Checking any missing values
missing_values = df.isnull().sum()
print("Missing values in each
column:")
print(missing_values)
```

```
#Check for duplicate rows
duplicate_rows = df[df.duplicated()]
print("Duplicate rows found:")
print(duplicate_rows)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10868 entries, 0 to 10867
Data columns (total 6 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Year                    10868 non-null  int64
 1   113 Cause Name          10868 non-null  object
 2   Cause Name              10868 non-null  object
 3   State                   10868 non-null  object
 4   Deaths                  10868 non-null  object
 5   Age-adjusted Death Rate 10868 non-null  object
dtypes: int64(1), object(5)
memory usage: 509.6+ KB
Missing values in each column:
Year                      0
113 Cause Name            0
Cause Name                0
State                     0
Deaths                    0
Age-adjusted Death Rate   0
dtype: int64
```

```
Duplicate rows found:
Empty DataFrame
Columns: [Year, 113 Cause Name, Cause Name, State, Deaths, Age-adjusted Death Rate]
Index: []
```

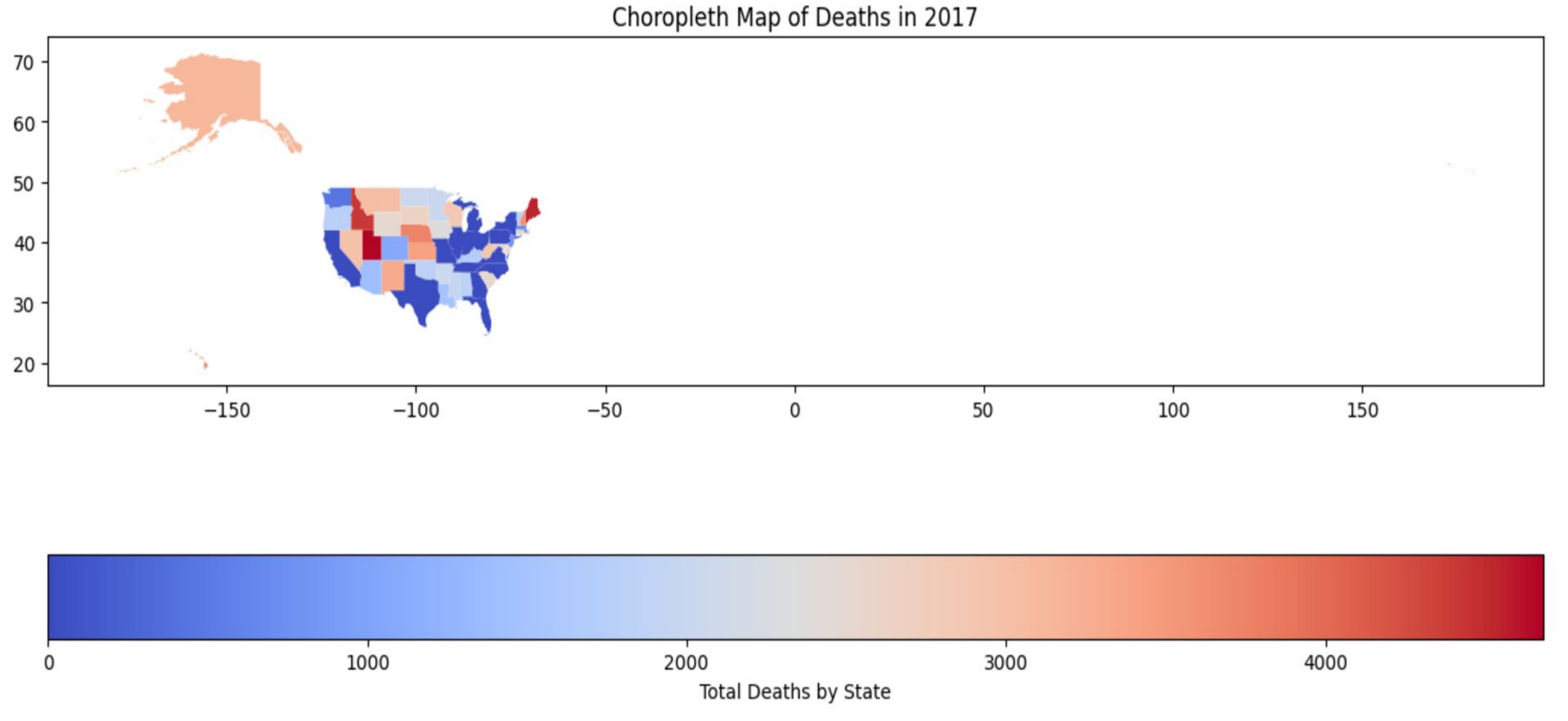
...Dataset loading, Understanding dataset and Cleaning

```
#Convert 'Deaths' and 'Age-adjusted Death Rate' to numeric,
coercing errors to NaN (useful for invalid values)
df['Deaths'] = pd.to_numeric(df['Deaths'], errors='coerce')
df['Age-adjusted Death Rate'] =
pd.to_numeric(df['Age-adjusted Death Rate'], errors='coerce')

# Check if there are any NaN values after conversion
print(df[['Deaths', 'Age-adjusted Death Rate']].isna().sum())

# Drop rows with NaN values in 'Deaths' and 'Age-adjusted
Death Rate'
df_cleaned = df.dropna(subset=['Deaths', 'Age-adjusted Death
Rate'])

# Check the data info again after cleaning
df_cleaned.info()
```

```
Deaths                     6992
Age-adjusted Death Rate      28
dtype: int64
<class 'pandas.core.frame.DataFrame'>
Index: 3876 entries, 2 to 10867
Data columns (total 6 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Year                   3876 non-null   int64
 1   113 Cause Name         3876 non-null   object
 2   Cause Name             3876 non-null   object
 3   State                  3876 non-null   object
 4   Deaths                 3876 non-null   float64
 5   Age-adjusted Death Rate  3876 non-null   float64
dtypes: float64(2), int64(1), object(3)
memory usage: 212.0+ KB
```

Map Visualization

1. Choropleth Map - Death Rate by State



… Map Visualization

2. Density Map

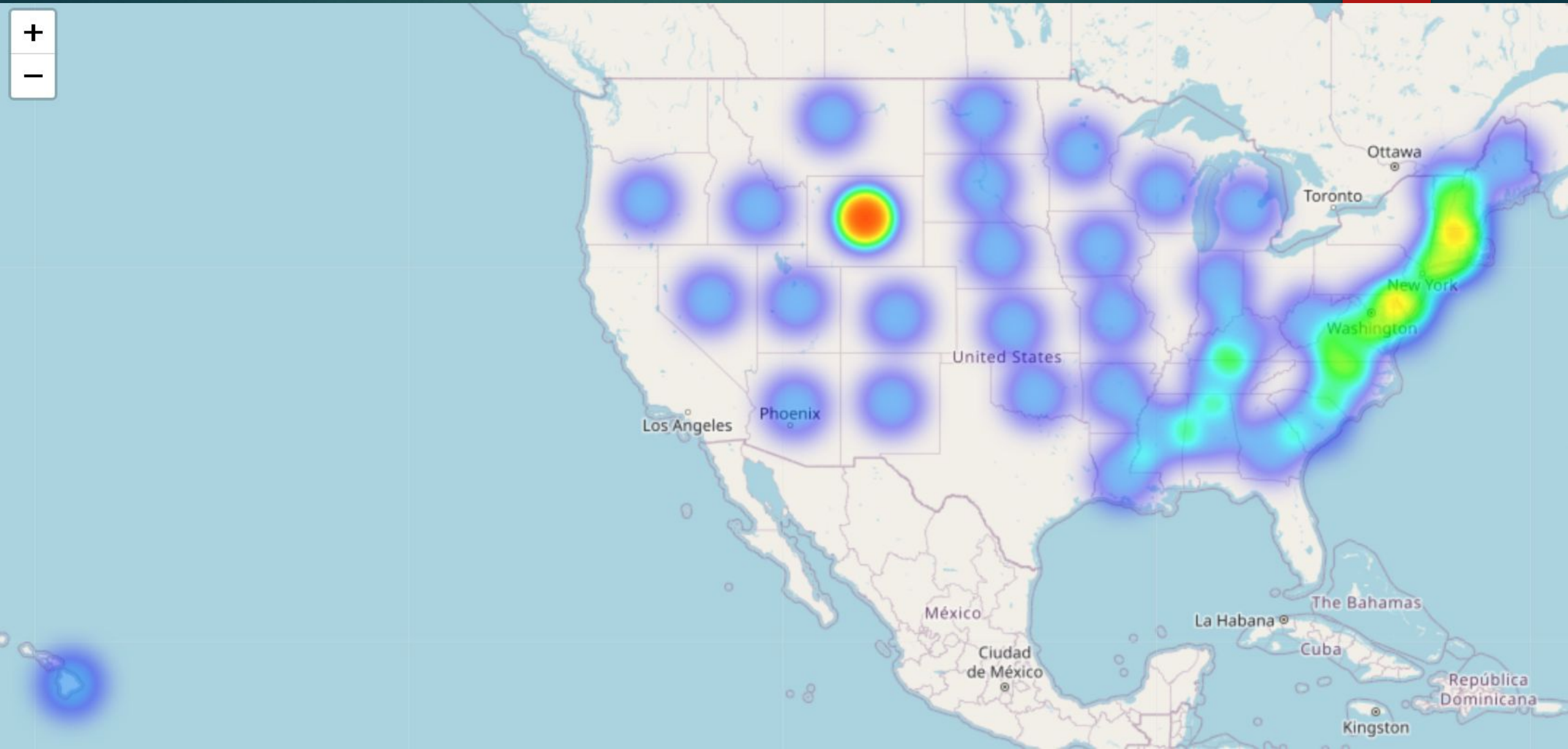
```
Rows with missing geocoding: 0
Map has been saved as 'death_density_map.html'
```

```
df.head()
```

	Year	113 Cause Name	Cause Name	State	Deaths	Age-adjusted Death Rate	Lat	Lon
0	2017	Accidents (unintentional injuries) (V01-X59,Y8...	Unintentional injuries	United States	NaN	49.4	39.783730	-100.445882
1	2017	Accidents (unintentional injuries) (V01-X59,Y8...	Unintentional injuries	Alabama	NaN	53.8	33.258882	-86.829534
2	2017	Accidents (unintentional injuries) (V01-X59,Y8...	Unintentional injuries	Alaska	436.0	63.7	64.445961	-149.680909
3	2017	Accidents (unintentional injuries) (V01-X59,Y8...	Unintentional injuries	Arizona	NaN	56.2	34.395342	-111.763275
4	2017	Accidents (unintentional injuries) (V01-X59,Y8...	Unintentional injuries	Arkansas	NaN	51.8	35.204888	-92.447911

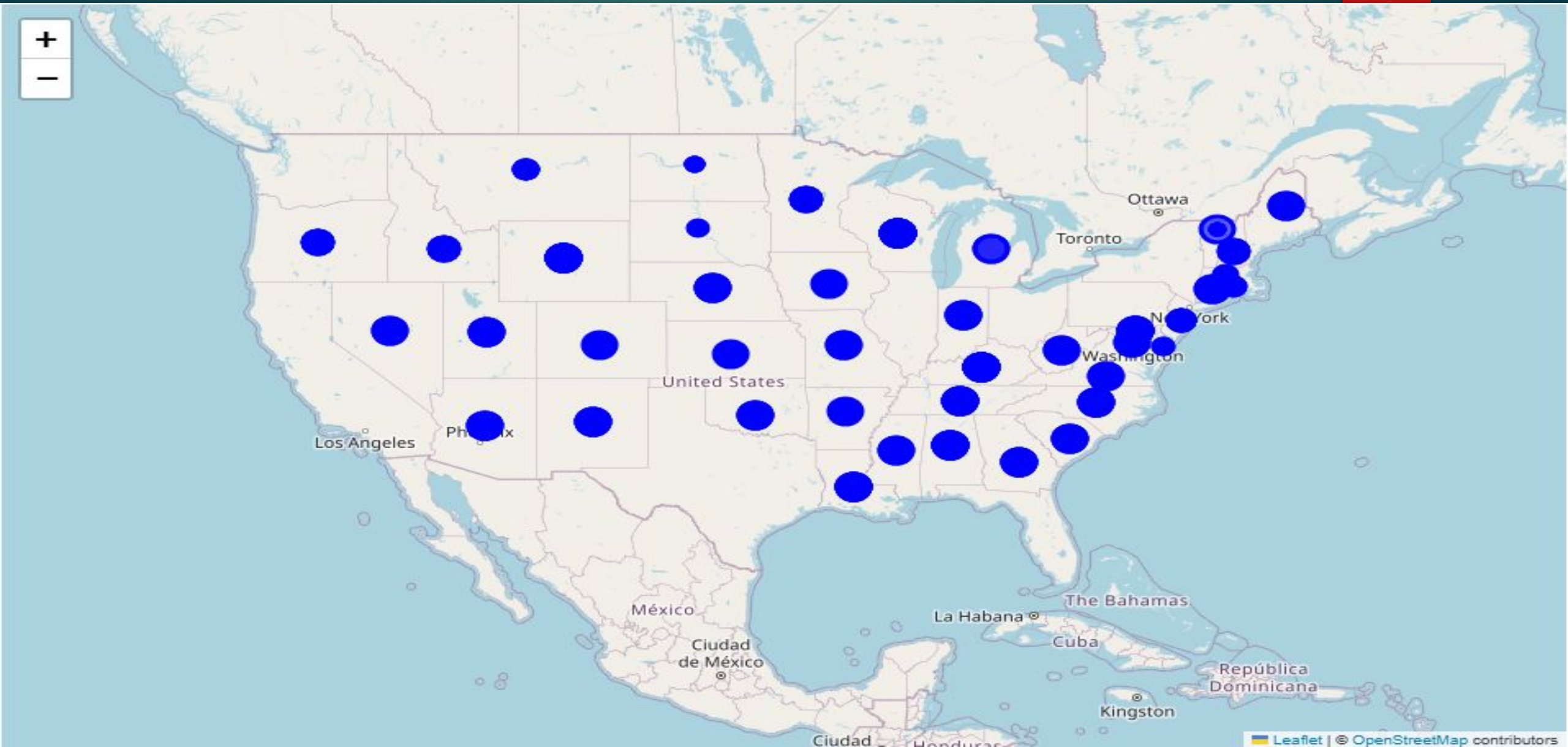
… Map Visualization

2. Density Map



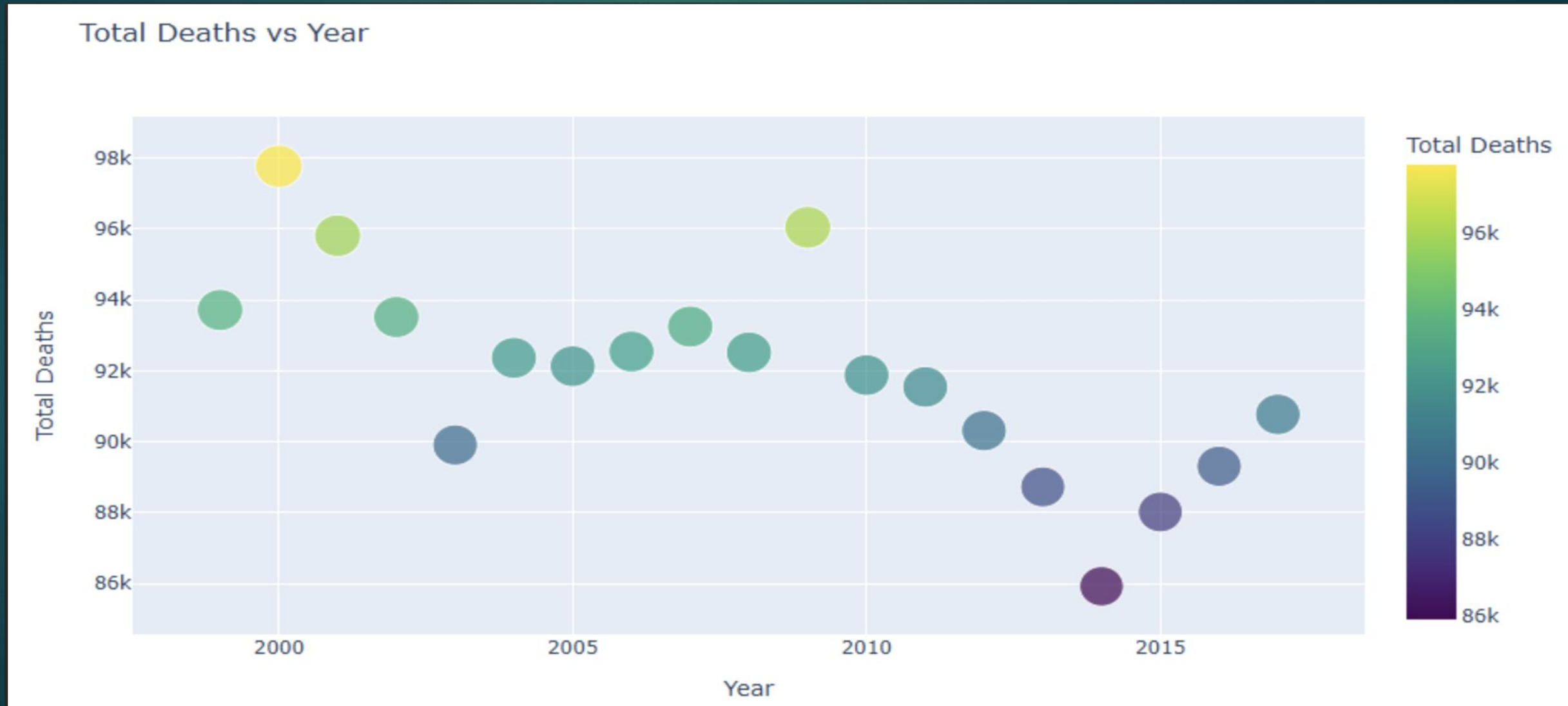
… Map Visualization

3. Bubble Map



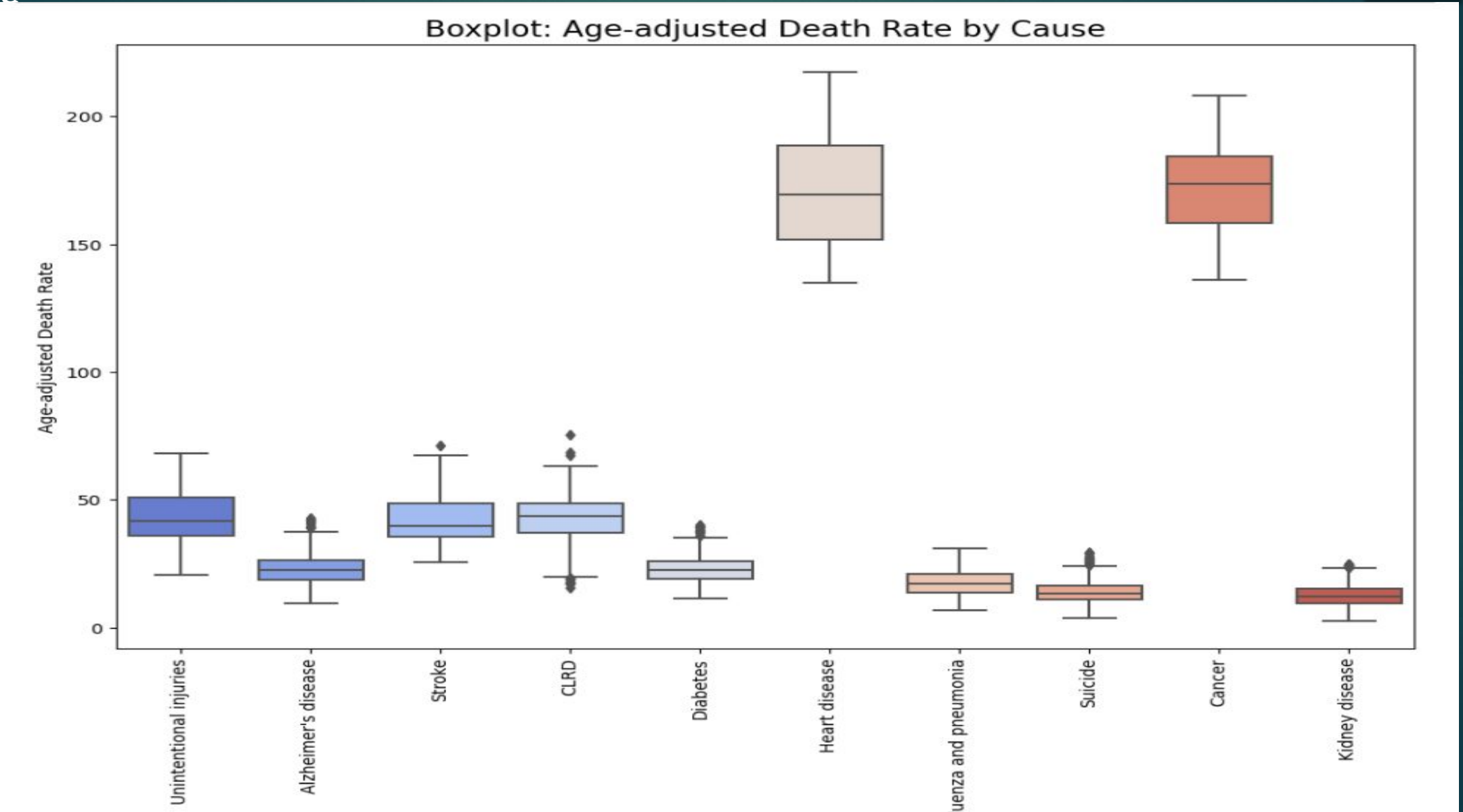
… Aggregation Visualization

1. Scatter Plot



… Aggregation Visualization

2.. Box Plot for Age-adjusted Death Rate by Cause



… Aggregation Visualization

3. Violin plot deaths by state

Interactive Visualization

1. Tabular Visualization (Interactive Table)

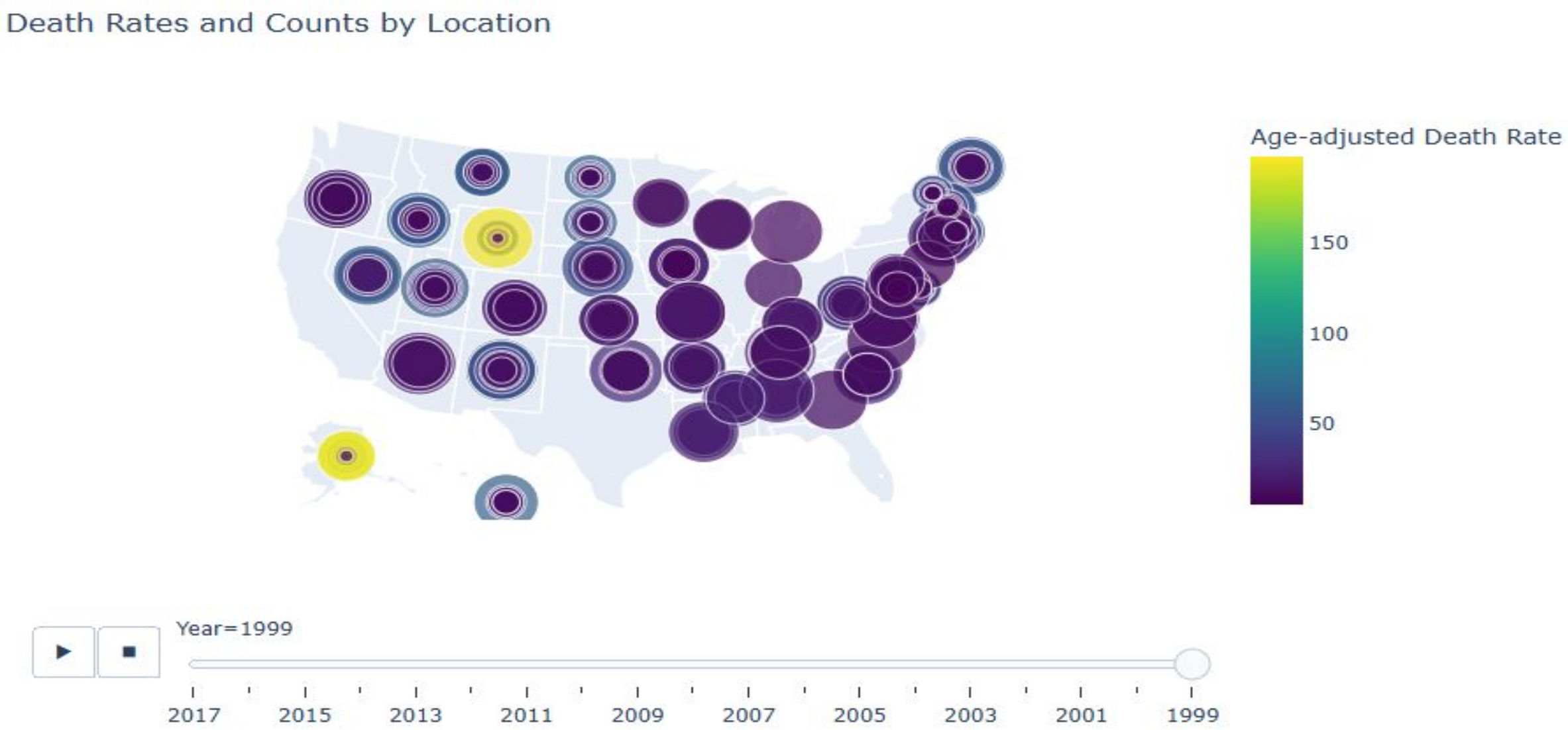
Leading Causes of Death - Data Table

Year	113 Cause Name	Cause Name	State	Deaths
filt				
2017	Accidents (unintentional injuries) (V01-X59,Y85-Y86)	Unintentional injuries	Alaska	436
2017	Accidents (unintentional injuries) (V01-X59,Y85-Y86)	Unintentional injuries	Delaware	608
2017	Accidents (unintentional injuries) (V01-X59,Y85-Y86)	Unintentional injuries	District of Columbia	427
2017	Accidents (unintentional injuries) (V01-X59,Y85-Y86)	Unintentional injuries	Hawaii	585
2017	Accidents (unintentional injuries) (V01-X59,Y85-Y86)	Unintentional injuries	Idaho	876
2017	Accidents (unintentional injuries) (V01-X59,Y85-Y86)	Unintentional injuries	Maine	990
2017	Accidents (unintentional injuries) (V01-X59,Y85-Y86)	Unintentional injuries	Montana	579
2017	Accidents (unintentional injuries) (V01-X59,Y85-Y86)	Unintentional injuries	Nebraska	811
2017	Accidents (unintentional injuries) (V01-X59,Y85-Y86)	Unintentional injuries	New Hampshire	907
2017	Accidents (unintentional injuries) (V01-X59,Y85-Y86)	Unintentional injuries	North Dakota	339

« ‹ 1 / 388 › »

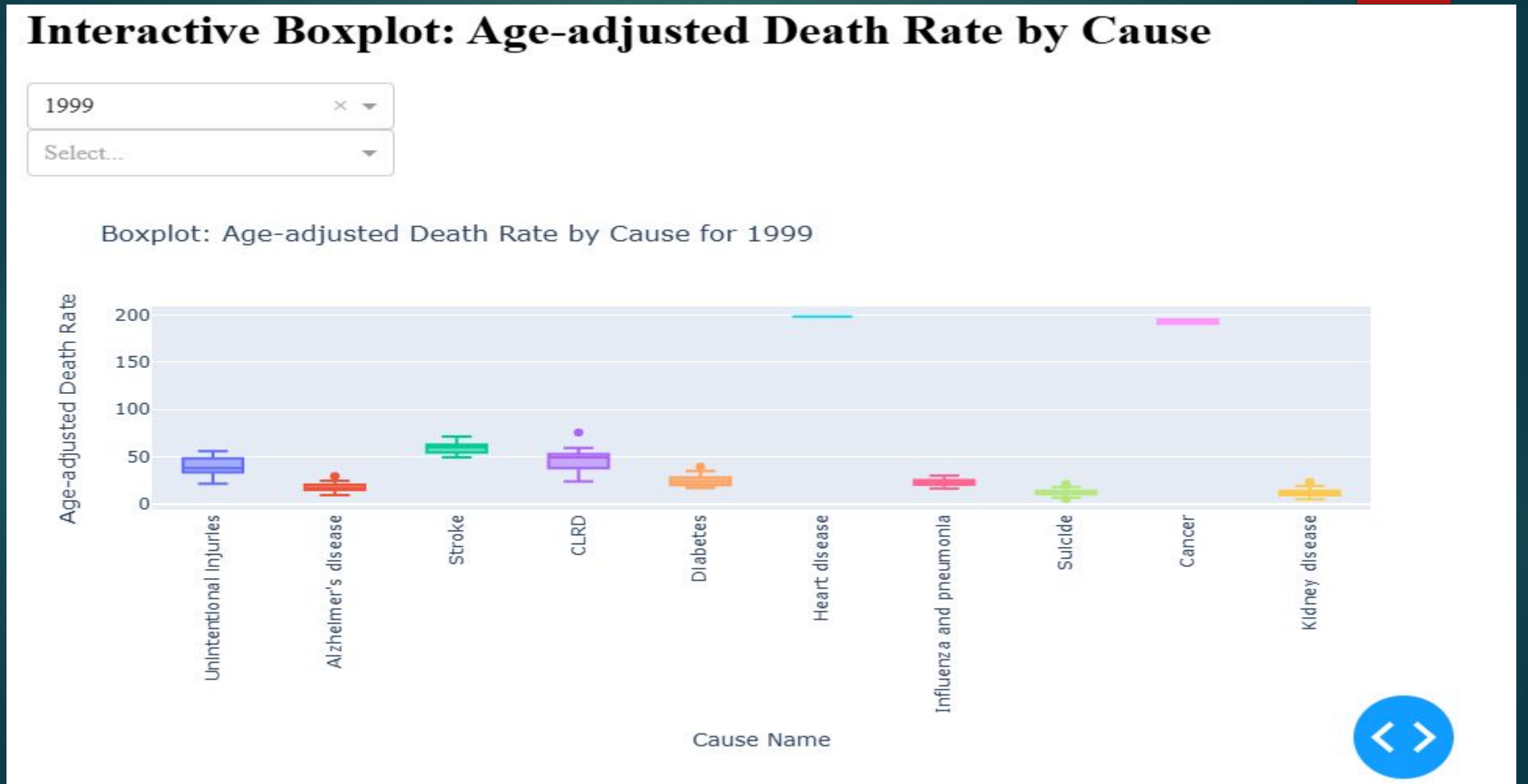
… Interactive Visualization

2. Interactive Choropleth Map



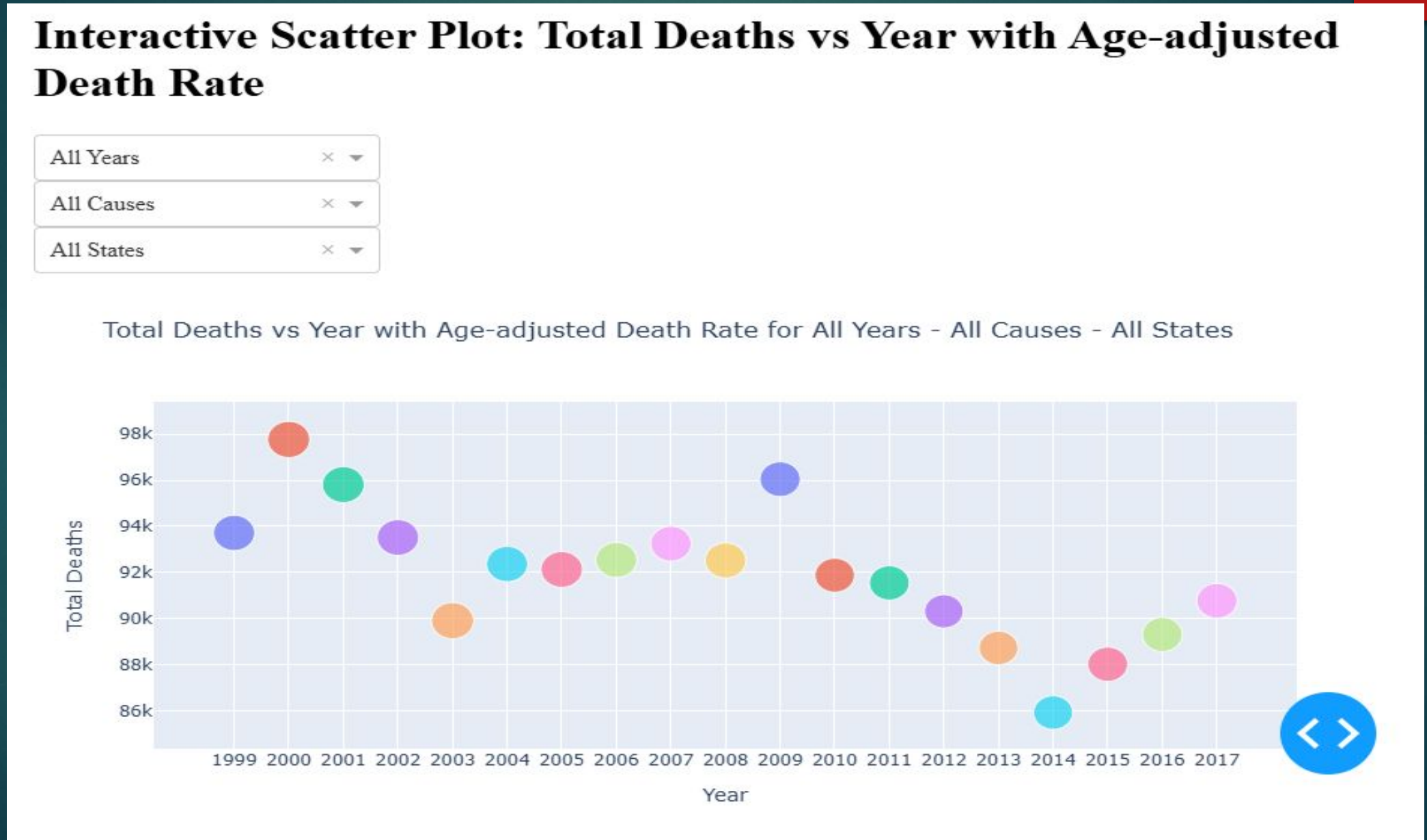
… Interactive Visualization

3. Interactive Box Plot



… Interactive Visualization

4. Interactive Scatter Plot



CONCLUSION

The project analyzed and visualized the leading causes of death in the United States (1999–2017) using advanced visualization techniques.

The project identified significant trends in cause of deaths over time and geographic regions, utilized different visualization techniques to make complex data insights accessible and intuitive, and applied interactive platform for deeper insights into National Center for Health Statistics data.

Techniques Implemented:

1. Map visualization

- Choropleth map: Displayed state-wise variations in leading causes of death and mortality rates.
- Density and Bubble maps: Highlighted geographic clusters and distribution of causes of death.

2. Aggregation visualization

- Scatter plots: To show the relationships based on the annual total death rate, interactively revealing annual patterns across different regions and causes of death.
- Box/Violin plot: Compared state-level variations in mortality rates for selected causes.

3. Interactive visualization

- Interactive plots and map: Enabled exploration of mortality rates and causes across states and years.
- Tabular visualization: Provided tools for sorting, searching tables for detailed analysis of key datasetmetrics.


THANK YOU