What are the Leading Causes of Death in the United States?

Wiley Winters

Regis University Anderson College of Business and Computing

MSDS 670 Data Visualization

Mr. John Koenig

March 10, 2024

Abstract

This paper describes research on the leading causes of death in the United States from 2000 to 2017. The data was sourced from the U.S. Department of Health and Human Services and was published by the Centers for Disease Control and Prevention. The scope of this research is to identify the top causes of death and not the underlying social, economic, or behavior factors that may contribute to the death rates.

What are the Leading Causes of Death in the United States?

Research Question

During the 2020 pandemic, COVID-19 became the third leading cause of death in the United States. The final overall death rate rose from 715 deaths per 100,000 in 2019 to 835 per 100,000 in 2020 due to Covid (CDC, 2022). While COVID-19 caused a lot of deaths, it still did not become the number one or two cause of death in the United States. The research conducted for this paper will answer the question: "What are the leading causes of death in the United States?"

Data

All major data sets were obtained from the U.S. National Center for Health Statistics (NCHS) and the U.S. Census Bureau. In addition, a state spelling to abbreviation dataset was manually constructed from resources found on the internet. Data was downloaded as either a CSV or xlsx formatted file. If the file was xlsx formatted, it was converted to CSV to reduce its size and make it easier to manually edit. The author if this paper had difficulty finding continuous U.S. State population statistics on the Census Bureau's web site. Two CSV files were downloaded. One covers the years from 2000 to 2010 and another from 2011 to 2020. The last two years of census data were not used in this analysis. The Census Bureau files were cleaned up and concatenated into a single CSV file. The State population data was required for calculating the *Crude Death Rate* for each U.S. State.

The main data set for this study was downloaded from the NCHS's website in xlsx format and converted to a CSV. This made it easier to manually edit and work with the file if required. The original file consisted of seven columns and 137,700 rows. There are a number of blank or *NaN* values in the dataset, but these were not imputed or removed. According to the data set's web site, blank values do not indicate the data is missing, it just has not been reported and recorded at time of publication and will be added later. For this analysis, missing values can be tolerated.

How Death Rates are Measured

There are three measures used when reporting death rates in literature. The first is the raw number which is just a count of the number of deaths and their causes. The next is *age-adjusted rate* computed using a direct method by applying age-specific rates in a population of interest to a standardized age distribution. Age-adjusted rates are calculated as in:

$$\sum ni = 1 r_i \times (p_i / P)$$

where

r_i = rate in age group *i* in the population of interest

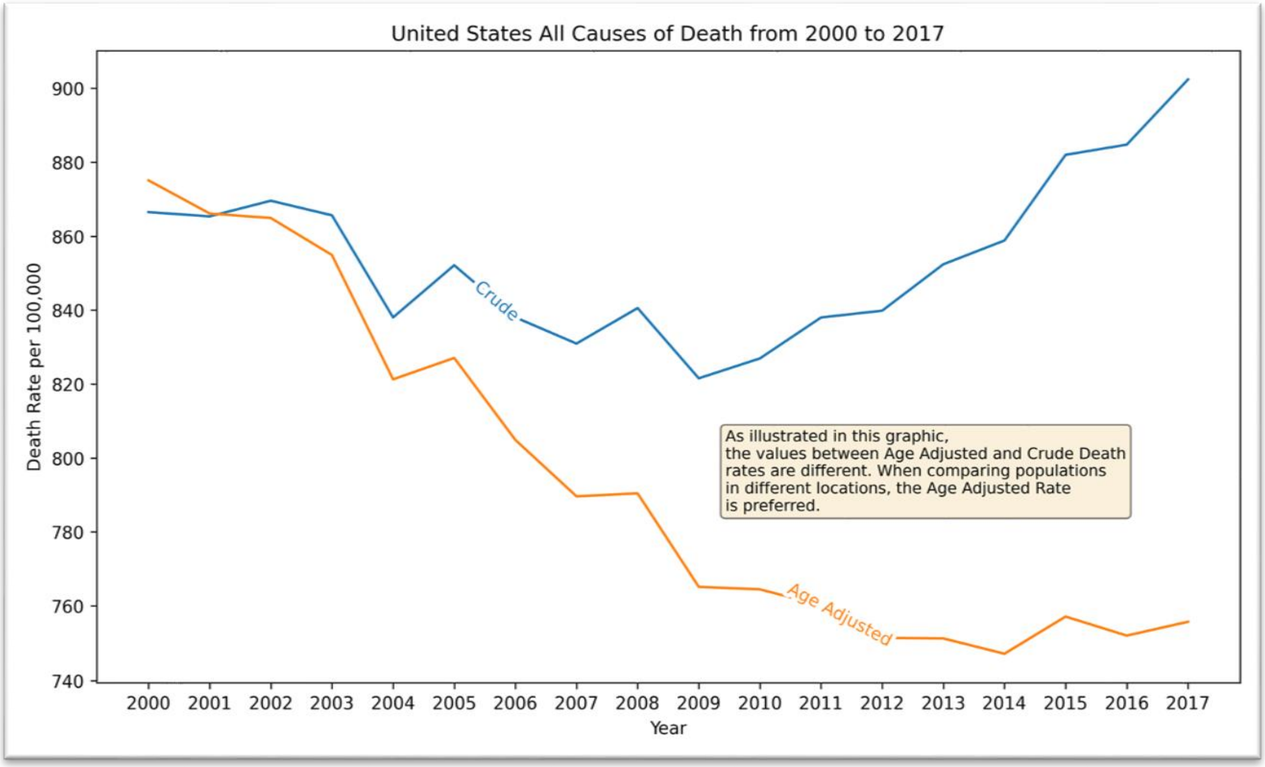
p_i = standard population in age group *i*

$P = \sum ni = 1 p_i$

n = total number of age groups over the age range of the age-adjusted rate (CDC, 2022).

The CDC recommends using the age-adjusted rate when comparing populations of different geographic areas. However, it does warn that age-adjusted rates should be viewed as relative indexes rather than actual measures of risk (CDC, 2022). The final measure used in reporting death rates is the *Crude Death Rate*. These are useful when a person wants to map or observed state or country wide death statistics. (CDC, 2018). The formula for calculating the crude death rate is *Number of deaths / population * 100,000.* This will give the crude death rate per 100,000 people (Spears, 2024).

While the crude death and age-adjusted rates are often used in reporting mortality rates, they are not the same. The plot below illustrates how they differ when comparing all deaths in the United States. The age-adjusted rate is trending lower than the crude rate.


United States All Causes of Death from 2000 to 2017

As illustrated in this graphic,
the values between Age Adjusted and Crude Death
rates are different. When comparing populations
in different locations, the Age Adjusted Rate
is preferred.

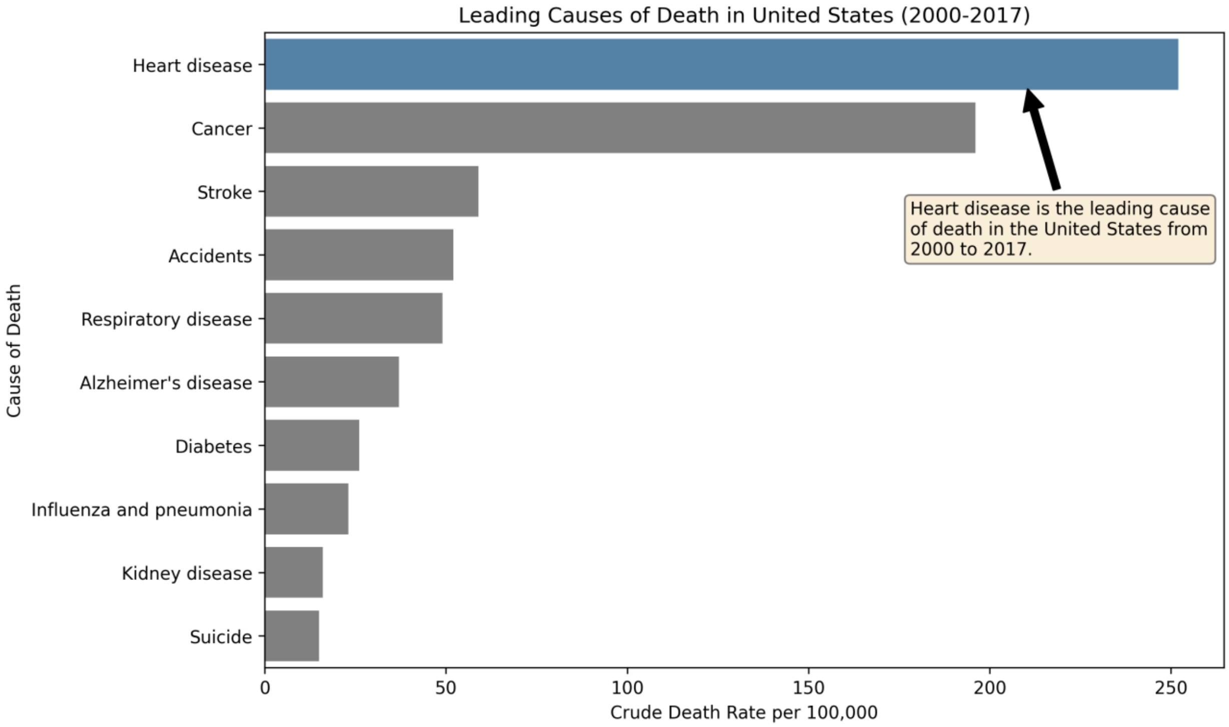
Method

A review of relevant literature was conducted, and sources were selected for this analysis. Early in the process the author decided to use datasets from trusted sources such as the NCHS and U.S. Census Bureau. Both organizations provide access to their data sets and the quality of them is sufficient for conducting analysis and reporting. After the datasets were selected and downloaded, they were manually evaluated using a spreadsheet application, converted to CSV as required, and loaded into *Pandas* data frames using Jupyter Lab. The analysis required the use of data from the NCHS and U.S. Census Bureau. This required the merging of three data frames into one. Merging was accomplished by using the pandas.merge() method. Data frames were left joined on the state field.

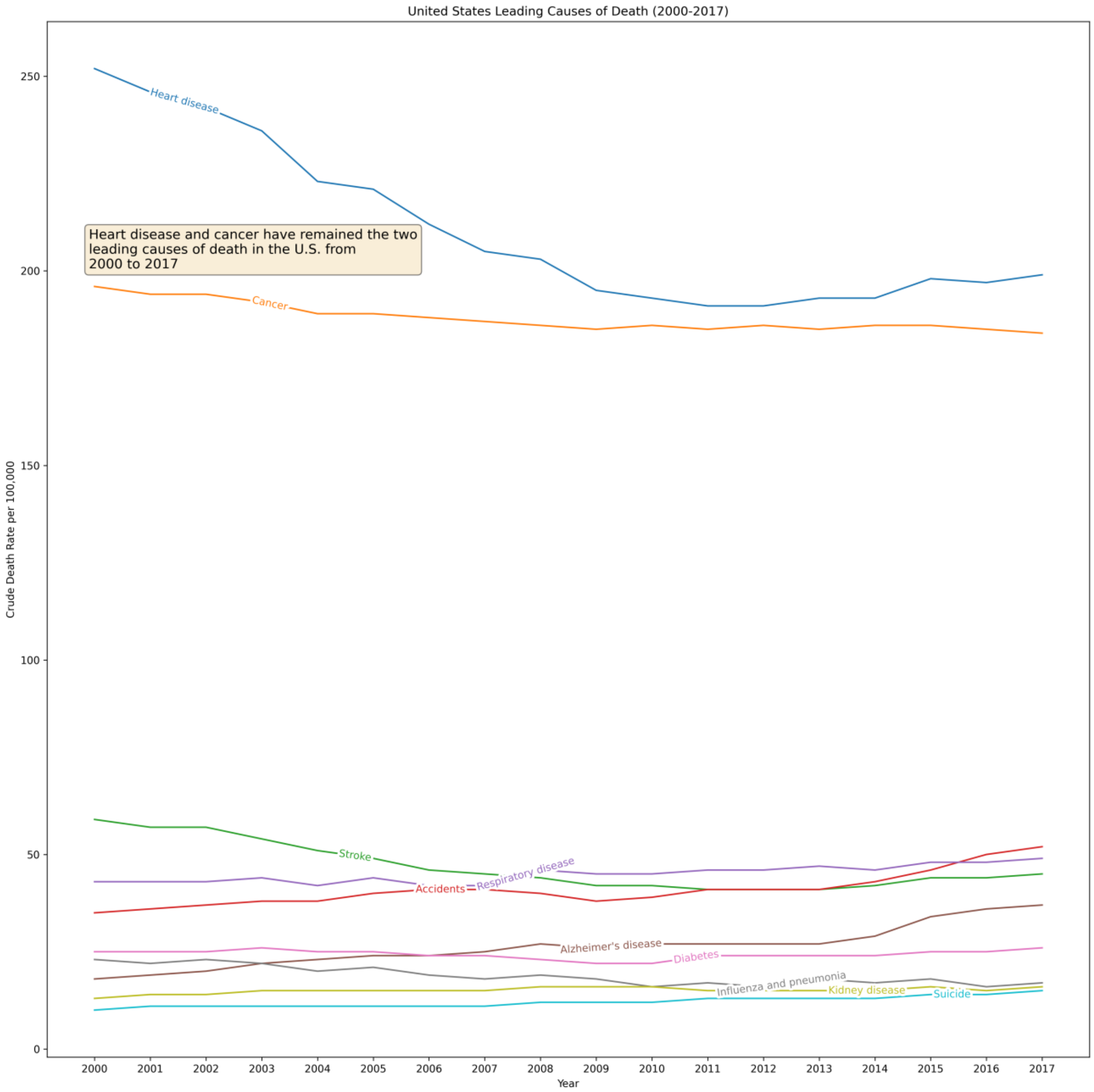
After creating the main data frame, the crude death rate was calculated for the whole United States and for each state. In addition, a data set was created with state names and abbreviations. This was also merged into the main data for use in creating a choropleth map with death rates displayed on it. Some basic EDA was conducted to ensure the data frame was sound enough for performing an analysis.

Findings

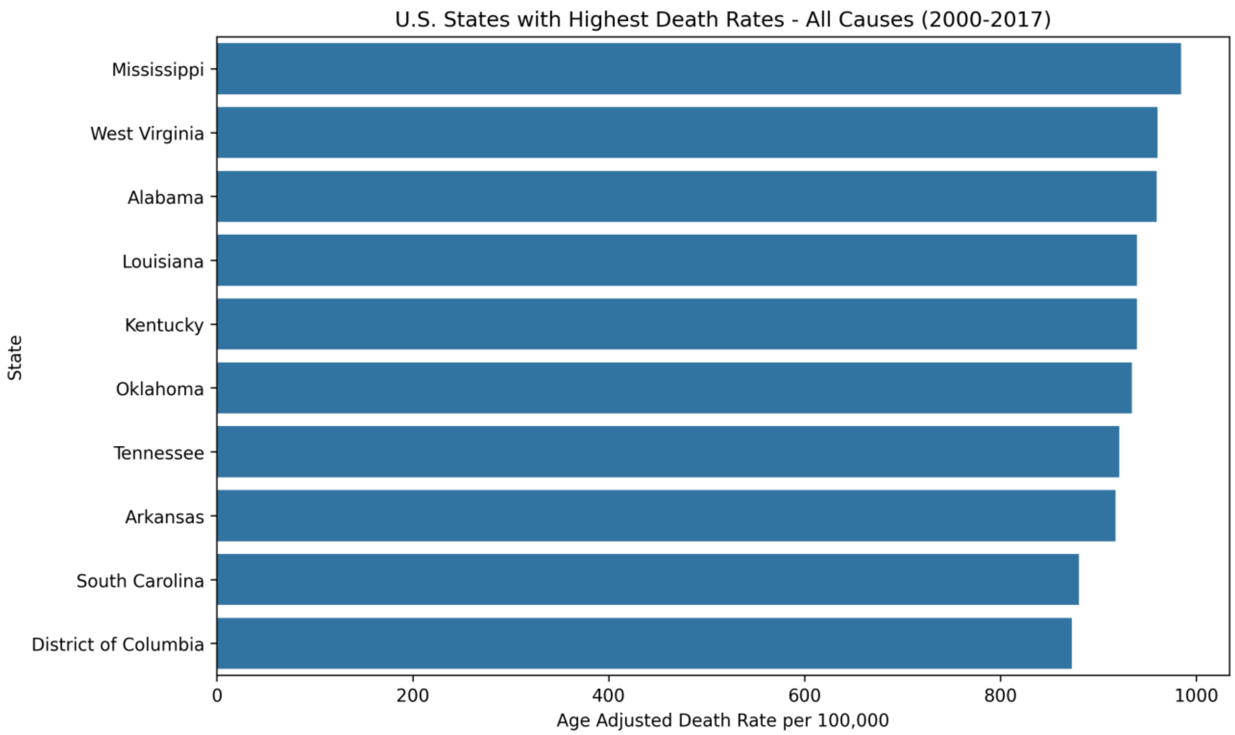
The goal of this research is to identify the top causes of death in the United States. Data was filtered to include only relevant information to determine what the leading causes are. Using the crude and age-adjusted death rates for the U.S. it was determined that heart disease is the leading cause of death from 2000 to 2017. The top 10 causes are listed in the figure below.



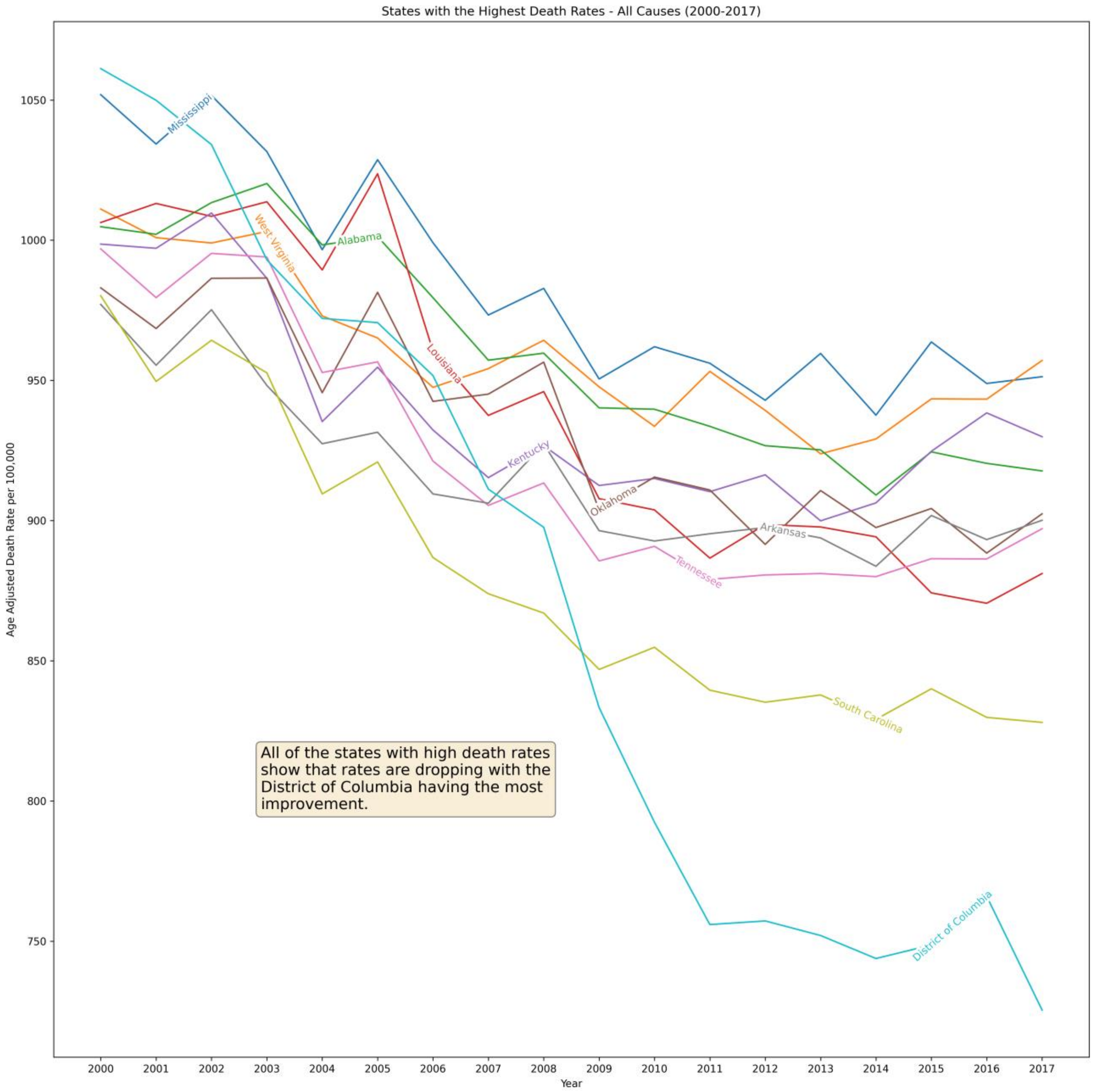
The two leading causes of death in the U.S. are heart disease and cancer. The next eight causes listed are significantly less than the first two. Only providing a snapshot of deaths rates in a bar plot does not illustrate variations in rates for the timeframe of this analysis. Therefore, the author created a time series plot to display how the rates have varied from 2000 to 2017. The line chart below displays the general trend for heart disease and cancer. One item that is clearly illustrated is heart disease and cancer death rates are much higher than the other causes of death.



After finding what are the leading causes of death in the U.S., the analysis was extended to determine what states have the highest death rates. This was done using the mean of the *Age-Adjusted* rate. In this data set the District of Columbia is counted as a state.

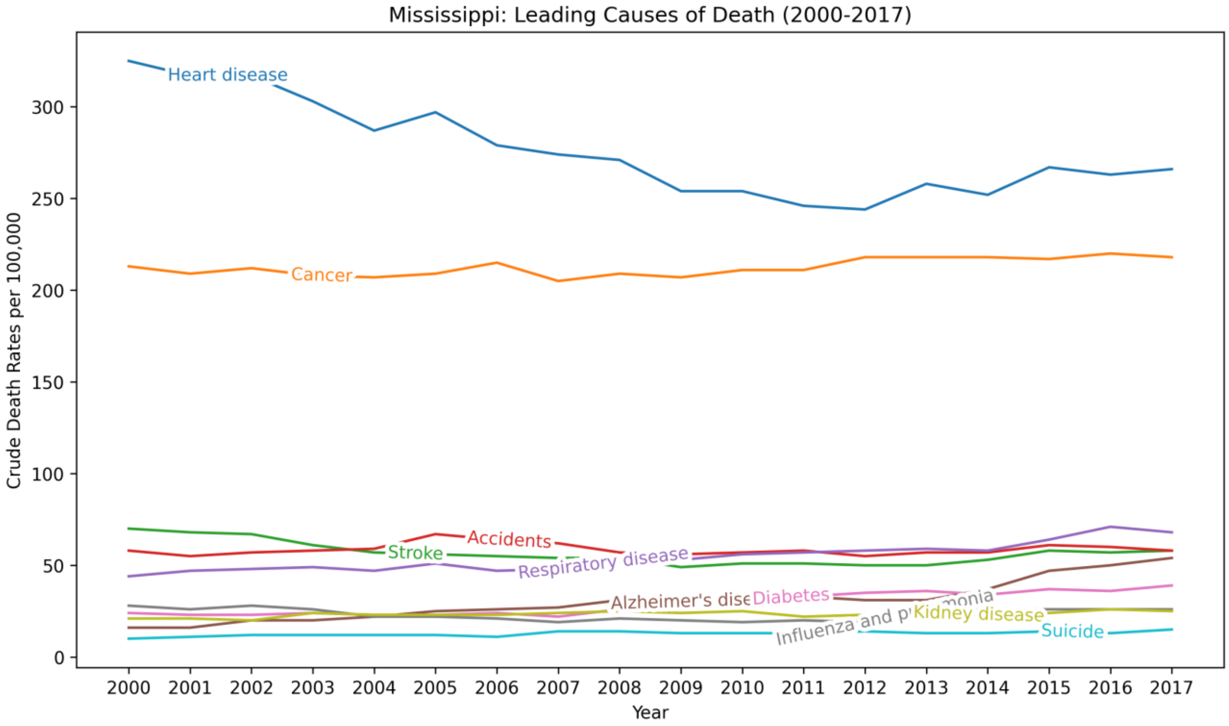


From the data, it appears that southern states have the highest death rates. The scope of this analysis does not explore the social, economic, or other reasons for the higher rates. However, the data shows that while the rates are higher than other states, they are dropping with the District of Columbia having the most improvement.

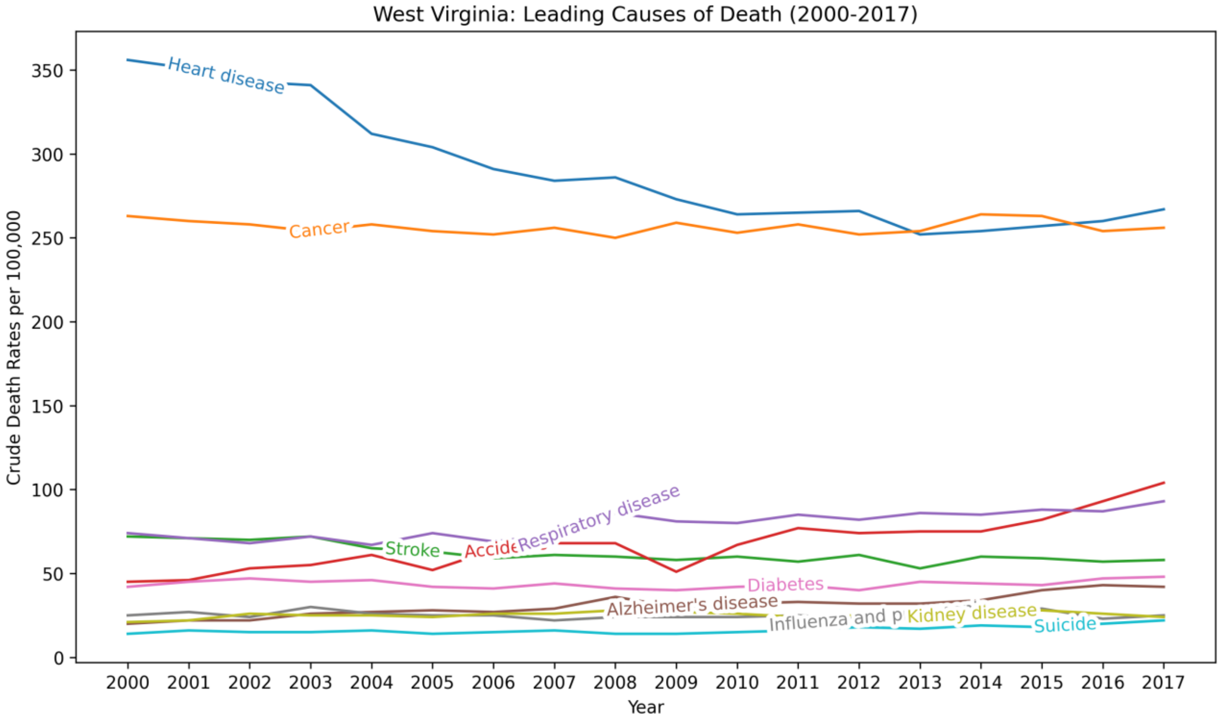


All of the states with high death rates show that rates are dropping with the District of Columbia having the most improvement.

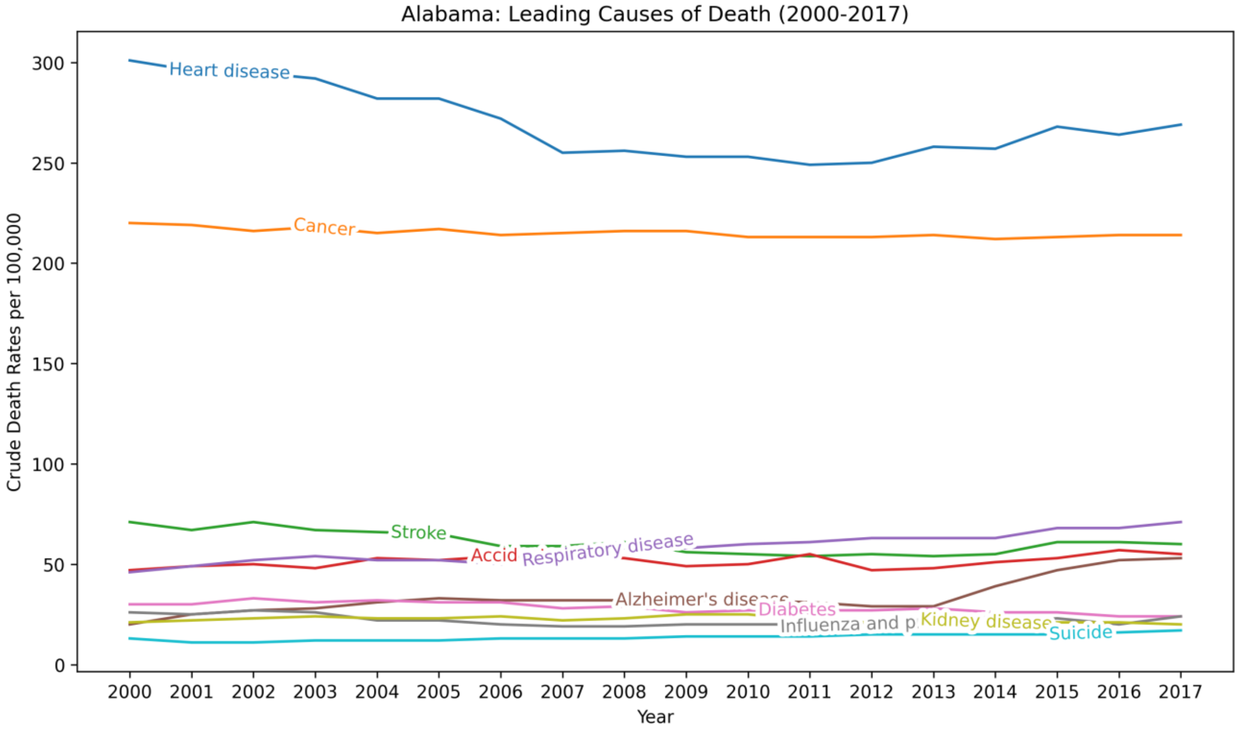
To further analyze death rates, each of these states were analyzed individually, and the results indicate that heart disease and cancer are the leading causes of death. None the less, as the century progressed many states displayed a reduction of these deaths.



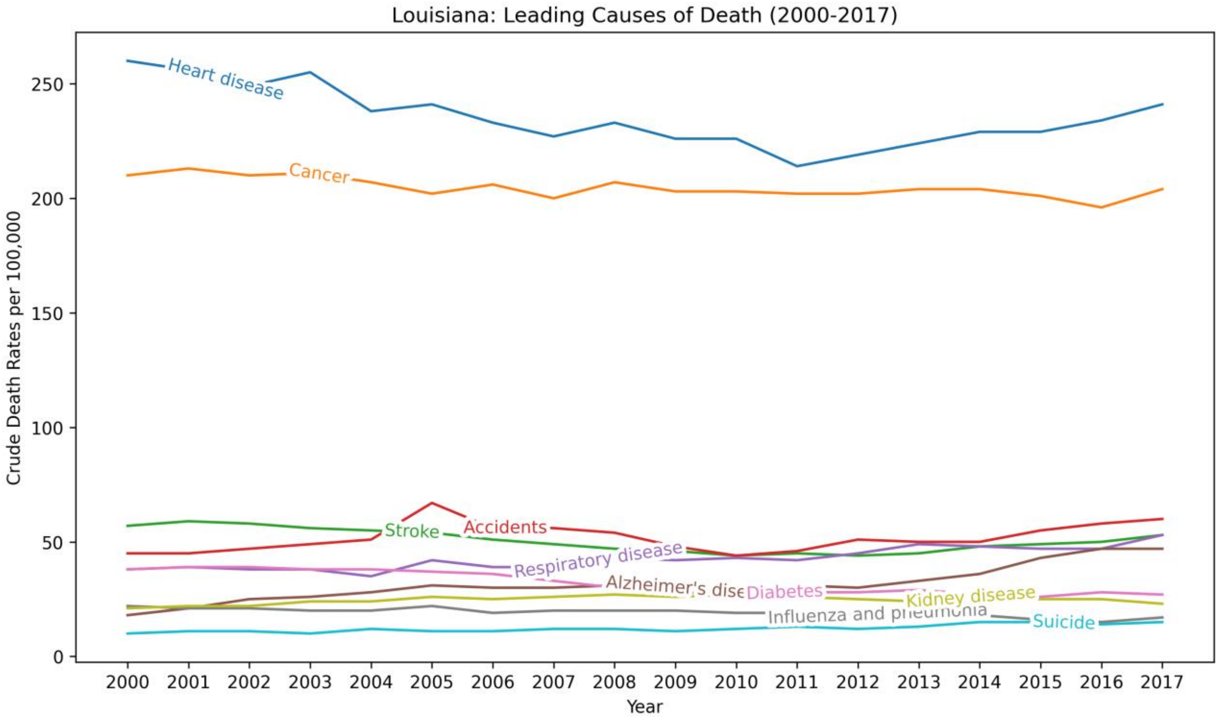
Mississippi has reduced heart disease, but the other causes remain almost constant between 2000 and 2017.



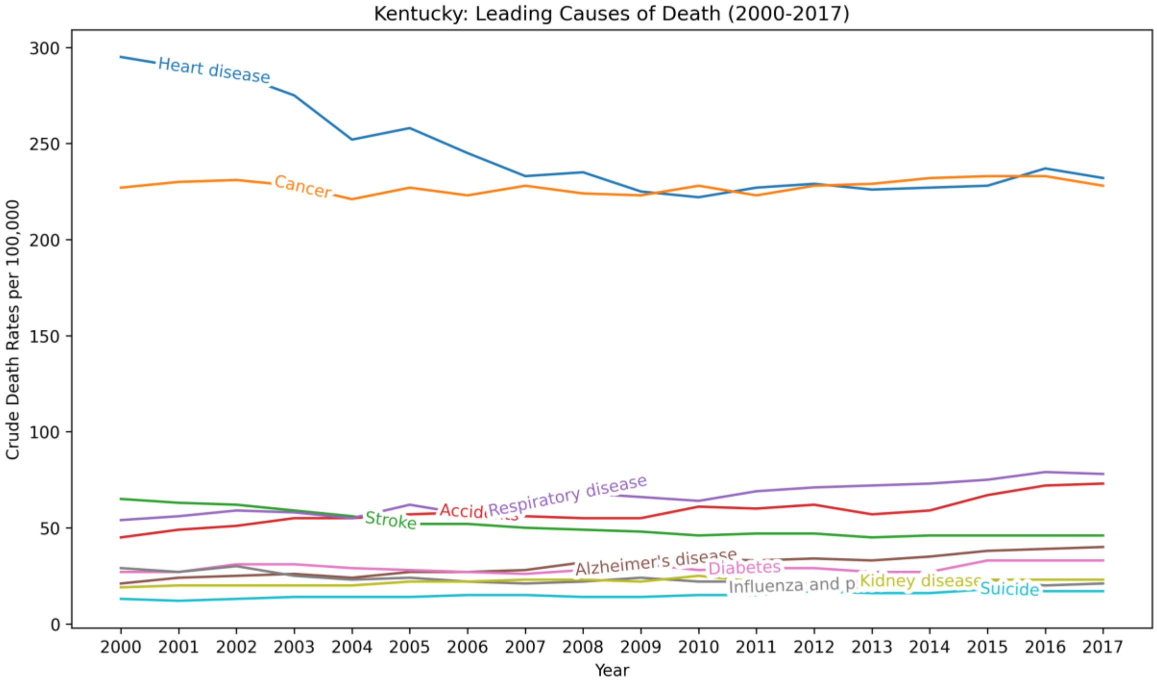
West Virginia has reduced the crude death rate for heart disease but increases in deaths from accidents and respiratory disease.



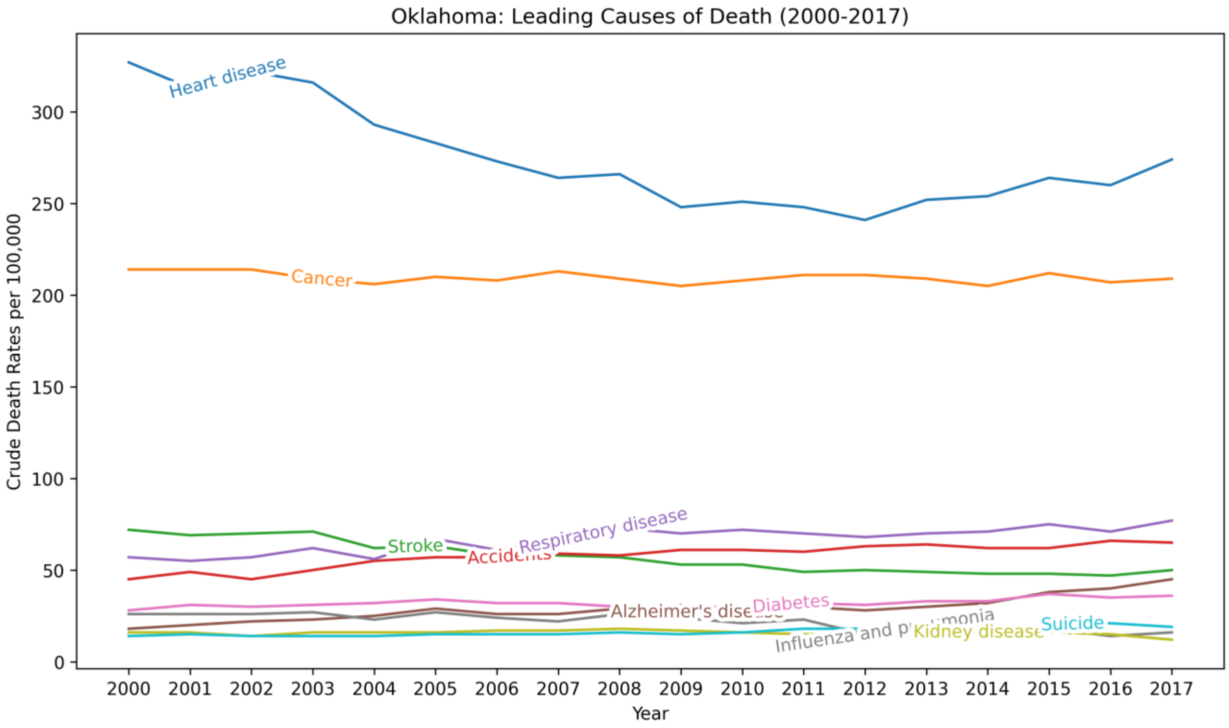
Alabama's results are interesting. From 2000 to 2008 there was a decline in heart disease, then from 2013 to 2017 it started to rise again. Cancer has remained constant and there were rises in respiratory disease, Alzheimer's, and accidents toward the 2016 to 2017 timeframe.



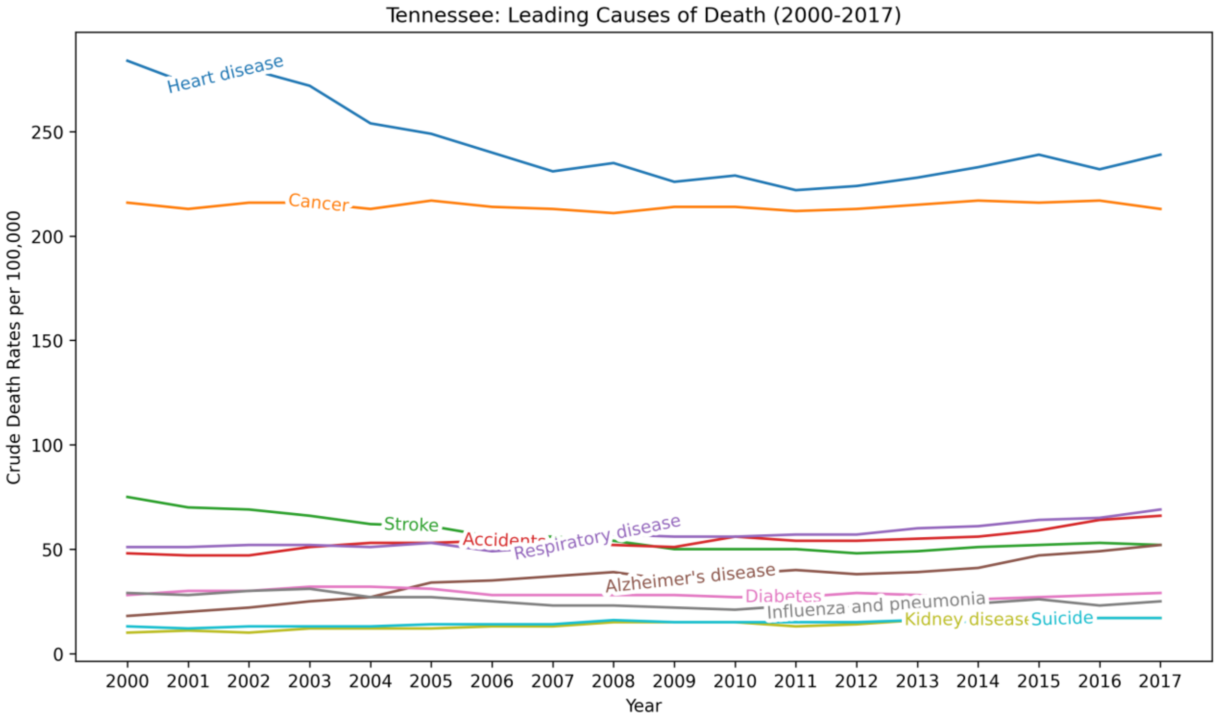
Louisiana had a slight decline in the heart disease death rate but finished 2017 with it almost matching what it was in 2000. Cancer also has a rise in its rate starting in 2016. Toward 2017 the other top causes of death also increased.



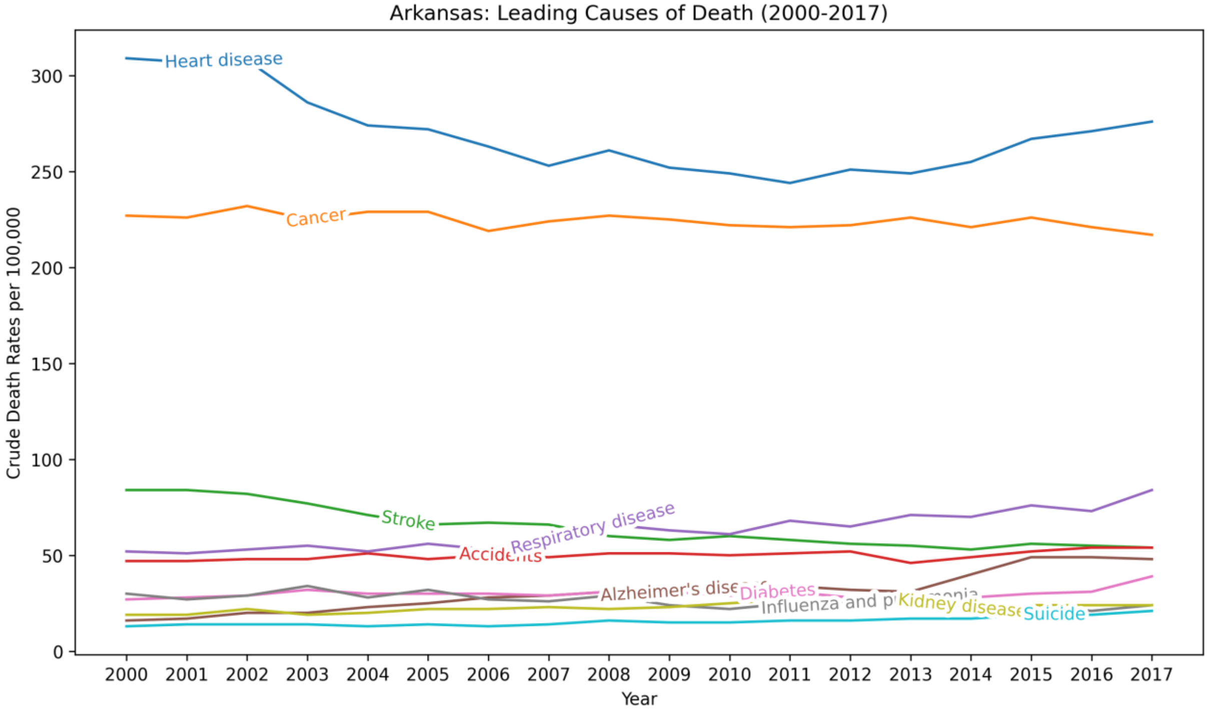
Kentucky started the century with a decline in heart disease and cancer, but both have leveled out and slightly increased in 2016. Like other states, respiratory disease and accidental deaths have risen.



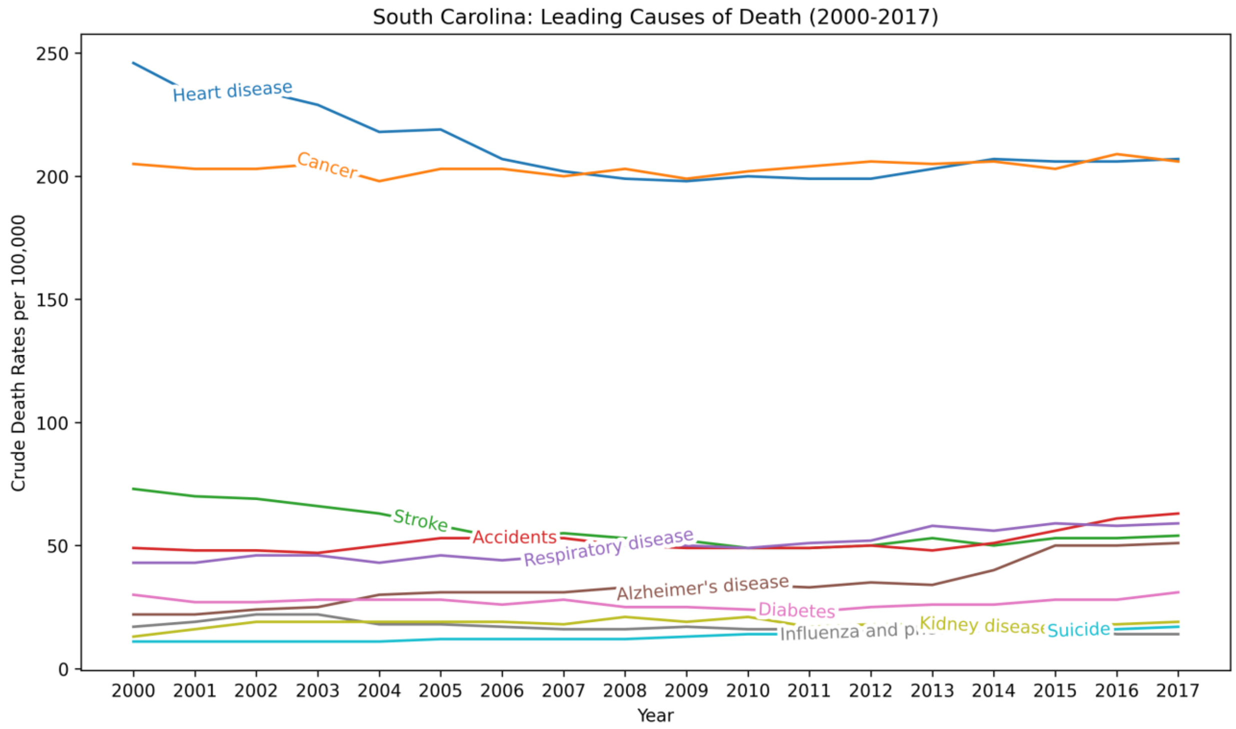
Oklahoma started off with the death rates for heart disease decreasing until 2012 where it started to rise again. Cancer death rate has been constant over the 18-year period. Here again, respiratory disease and accidents have increased.



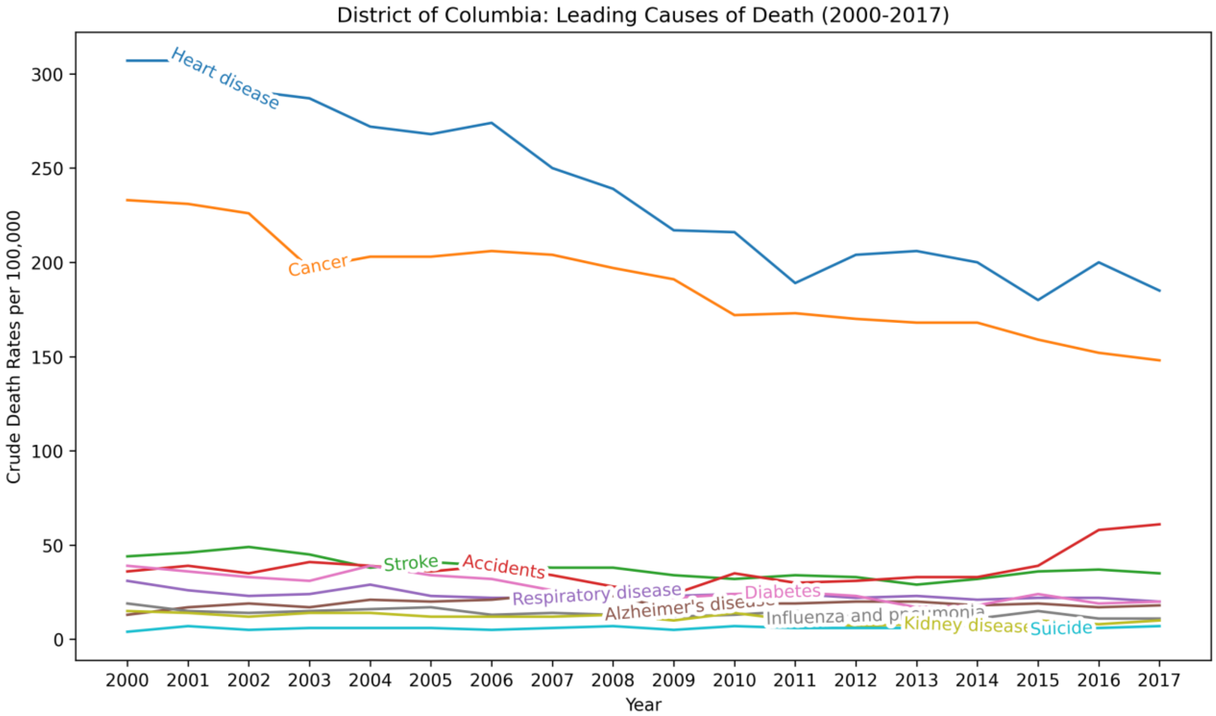
Tennessee from 2000 to 2017 hasn't had much change in its death rates. Heart disease has decreased, but it appears to be trending up around 2016.



Again, heart disease and cancer have the highest death rates for Arkansas. Respiratory disease is on the rise along with diabetes.



Heart disease and cancer have the highest rates in South Caroline. The death rates for respiratory disease and accidents are trending up as the century progresses.

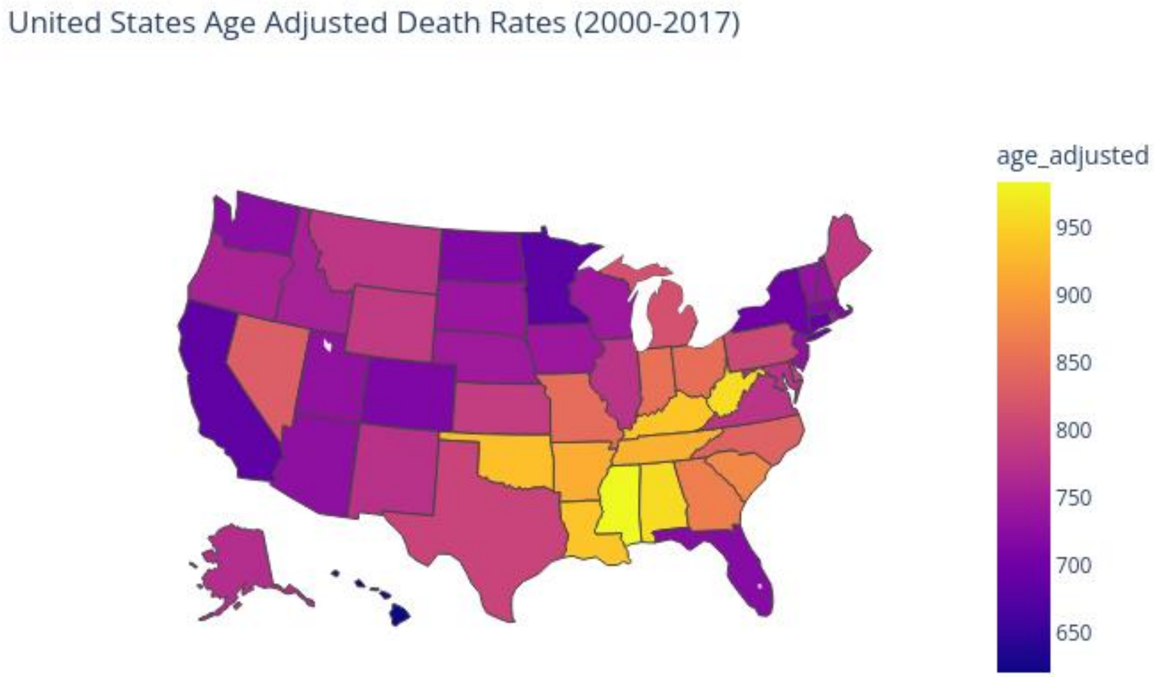


The District of Columbia has reduced its death rates for heart disease and cancer, but accidental deaths are on the rise.

Conclusions

This analysis didn't discover any new information, but it did identify heart disease and cancer as being the two biggest causes of death in the United States. One interesting trend that was identified is the states with the highest death rates do not follow the rest of the country's average rates when it comes to the 10 largest causes of death. U.S. wide the leading causes of death from the highest to the lowest are *heart disease, cancer, stroke, accidents, respiratory disease, Alzheimer's disease, diabetes, influenza and pneumonia, kidney disease,* and *suicide*. The leading causes of death in the ten states with the highest rates are: heart disease, cancer, ***accidents***, and ***respiratory disease***. The other causes generally follow the national trend.

All the states within the top ten death rate group are in the southern United States. The plot below illustrates this trend.



The more yellow in color the state is, the higher its death rate is.

References

Centers for Disease Control and Prevention. (2022, January 7). *2020 Final Death Statistics: Covid-19 as an Underlying Cause of Death vs. Contributing Cause.* https://www.cdc.gov/nchs/pressroom/podcasts/2022/20220107/20220107.htm

Centers for Disease Control and Prevention. (2022, August 12). *Age adjustment - Health, United States.* https://www.cdc.gov/nchs/hus/sources-definitions/age-adjustment.htm

Spears, B. (2024, February 27). *Libguides: Publicly Available Sources of Data for Health & Social Determinants of Health: Rates & Formulas*. Rates & Formulas - Publicly Available Sources of Data for Health & Social Determinants of Health - LibGuides at Health Sciences Library System. https://hsls.libguides.com/health-data-sources/rates-formulas

Data Sets

U.S. Census Bureau (2019). *Population, Population Change, and Estimated Components of Population Change: April 1, 2010, to July 1, 1999* [Data set]. https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-total.html

U.S. Census Bureau (2010). State Intercensal Tables: 2000-2010 [Data set]. https://www.census.gov/data/tables/time-series/demo/popest/intercensal-2000-2010-state.html

U.S. National Center for Health Statistics (2022). *Leading Causes of Death: United States* [Data set]. https://catalog.data.gov/dataset/nchs-leading-causes-of-death-united-states