Nikhitha Lingutla, Xiao Tan, and Michael Winkler

Dr. Daniel Carr

STAT 663: Statistical Graphics and Data Exploration I

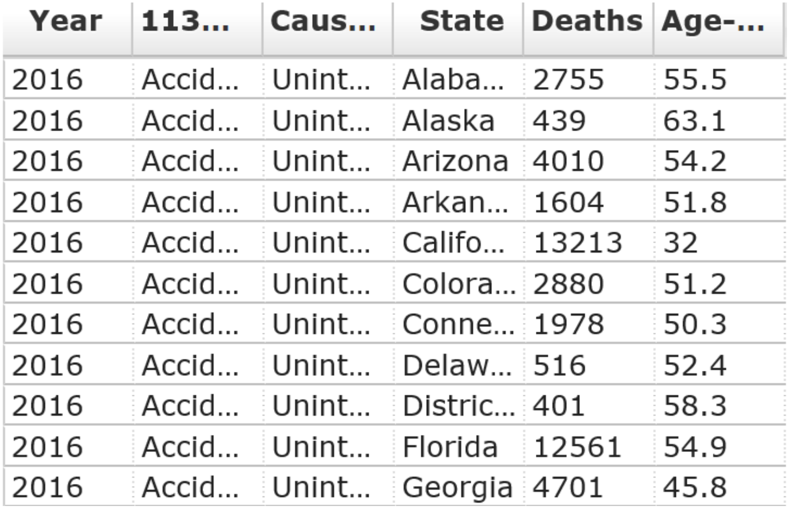
October 31st, 2018

Group Redesign Project: The 10 Most Common Causes of Death in the United States

Section I: Introduction - The Data

Since 1999 The Center for Disease Control and Prevention (CDC) has kept a data set of the ten most common causes of death in the United States. When a resident dies and a death certificate is filed with a state or Washington, D.C., the CDC collects the data and sorts it based on demographic and medical characteristics, including a generic cause of death and an International Classification of Disease, Tenth Revision (ICD-10) Cause of Death. The ICD-10 is a more detailed and scientific classification with a standardized code for each condition. In addition, the CDC publishes the total number of deaths for each state. This shows the raw number of deaths and does not adjust for different populations in different stats. Additionally, there is an age-adjusted death rate which is the weighted average of the age-specific mortality rates per 100,000 people. The population adjustment is done using data from 2000 Decennial Census and is adjusted yearly based on postcensal estimates. Clearly, a higher average age of a population will lead to higher probabilities of certain kinds of death, and this adjustment allows us to compare populations across states. In addition to this the fact that it is standardized to deaths per 100,000 allows for states to be compared more accurately. Without this adjustment we could only conclude that California and New York are large states with large populations. Below is a screenshot of the CSV file of the data.

Figure 1:



Year	113...	Caus...	State	Deaths	Age-...
2016	Accid...	Unint...	Alaba...	2755	55.5
2016	Accid...	Unint...	Alaska	439	63.1
2016	Accid...	Unint...	Arizona	4010	54.2
2016	Accid...	Unint...	Arkan...	1604	51.8
2016	Accid...	Unint...	Califo...	13213	32
2016	Accid...	Unint...	Colora...	2880	51.2
2016	Accid...	Unint...	Conne...	1978	50.3
2016	Accid...	Unint...	Delaw...	516	52.4
2016	Accid...	Unint...	Distric...	401	58.3
2016	Accid...	Unint...	Florida	12561	54.9
2016	Accid...	Unint...	Georgia	4701	45.8

Since this data is coming from the CDC, an agency in the United States Government, it is not surprising that there were not any missing values and that the data was relatively clean. Unfortunately, there were some problems in the data that should be noted. First, the CDC does note that the estimates of death rate and age-adjusted death rate change since they are estimated from the postcensal population estimates change. Although this is not ideal, the data that is used for analysis could change at any point, it is justifiable since all shifts in the estimate should be minimal. The data also includes the names of individual states as opposed to their state code. This is a relatively small issue, however; in order to use Plotly and Linked MicromapST, we did have to change the name to the state code. Additionally, we noticed that the data has an inconsistent organization which can cause serious problems in analysis. For 2016, all of the data is grouped together. The data is organized by year, then by cause of death (alphabetically), and finally by state in alphabetical order. Unfortunately, this is the only year that it is organized this way. In every other year, it is organized by cause (in alphabetical order), followed by state, and finally by date. Either way of organizing the data is acceptable; however, having multiple organizations in the same data is not acceptable. For the redesign project our preference was to keep the state names and add an additional column with the state code, unfortunately, this was not plausible with the organization of the data. Instead, we replaced the state names with the state code.

Section II: Why the Redesign is Necessary and the Plan for the Redesign

The dataset is a relatively large data set with eleven causes of death (ten generic and one total cause), 17 years of data, and 52 localities (50 states, Washington, D.C, and each having their own entry. Despite how much data there is the data is not obviously conducive to linear models. Doing a graphical redesign, we can see trends in the data and we can try to identify relationships that we would not have seen otherwise.

This data has many variables and we want to look at each of the variables by itself. In addition, since we have several years of data and anticipate that there will be quite a bit of variability over time, we would like to see how plots change over time. In order to accomplish this, we will look to use the Shiny package to make two interactive web applications.

First, we will present a line chart that will plot the death rates by individual causes of death against a time index. This plot can be used as a baseline to view the trends in causes of death and by itself will be just an interesting plot. In conjunction with the other plots it will help us to determine if a state or a group of states has a particular death rate growing faster than the national average.

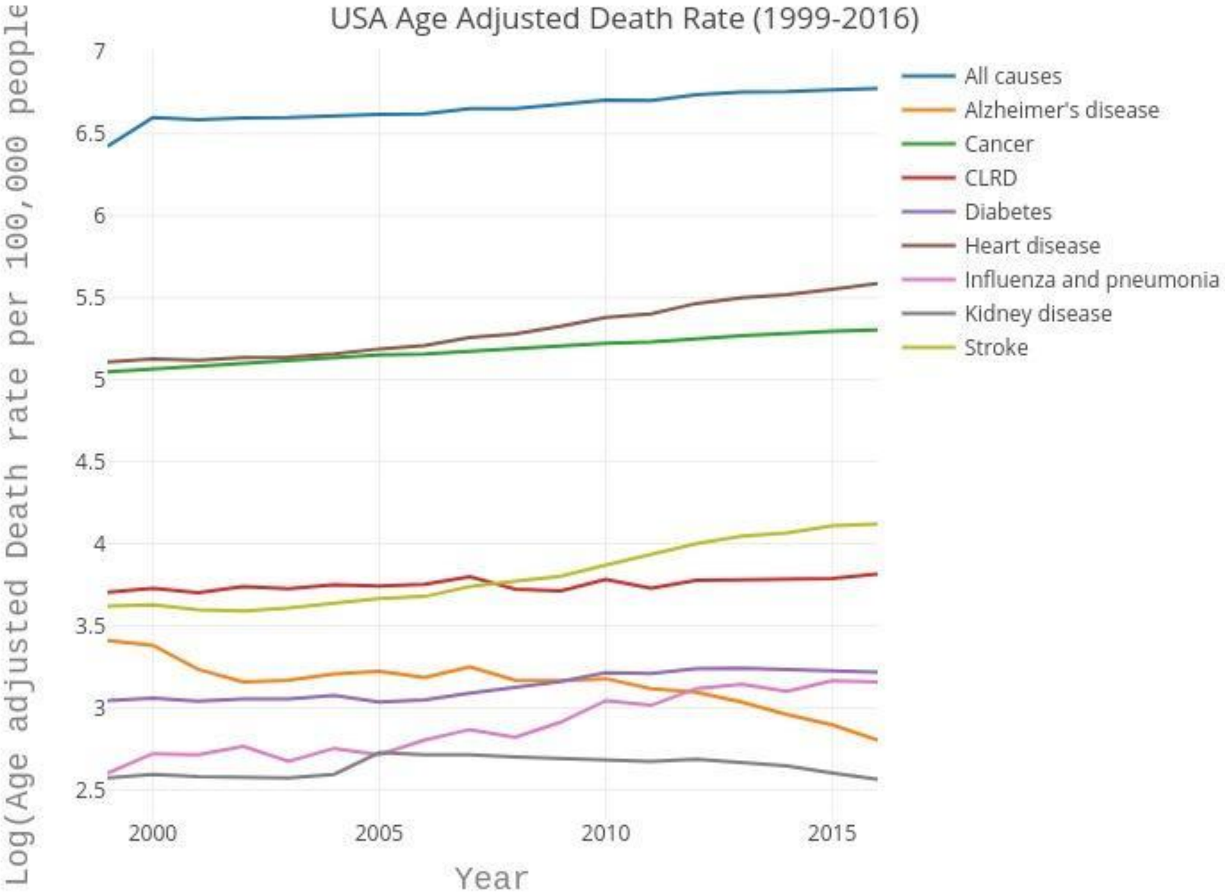
Next, the first application will us Shiny and Plotly to look for trends in the age non-adjusted data using to make a choropleth map of the United States. This will give us a map of the United States with different areas that are darker depending or lighter depending on the death rates in the state.

Finally, in the second Shiny application, we will look the age-adjusted data through linked Micro Maps. Using the Linked Micromaps we will be able to easily see where the states rank against each other in terms of adjusted death rates, while still seeing if there is a geographical connection between the states.

Section III – Time Series Line Plot

Using Plotly, we created a time series of the log adjusted deaths per 100,000 data. It is important to note that the log adjustment is used for two reasons, suppression of the fluctuation in data and suppression of the scale. For this particular dataset, there is not a great deal of fluctuation. So this is not as important. But since we are only concerned about the trend it will not hurt that the log adjustment will suppress extraneous fluctuations. Hence, we would not have needed to do the log adjustment if we did not need to suppress the y-axis. Since it is so important that an analyst is able to reference the total death rate, but since is the sum of all the other death rates the log adjustment will be crucial for suppressing the y-axis to make those comparisons. If this point of reference is not present, comparisons can only be made to individual causes of death. Although this comparison is useful, it is not as useful as it could be. If we were to plot with a normal deaths per 100,000 scale each line would be more compressed and there would be a lot wasted space in the middle of the plot. Both of these factors would make it much less inviting to look and considerably harder to read. However, when using the log scale, the y-axis is very compressed allowing us to see a comparison.

Figure 2: Time Series of Log Adjusted Deaths per 100,000 Residents

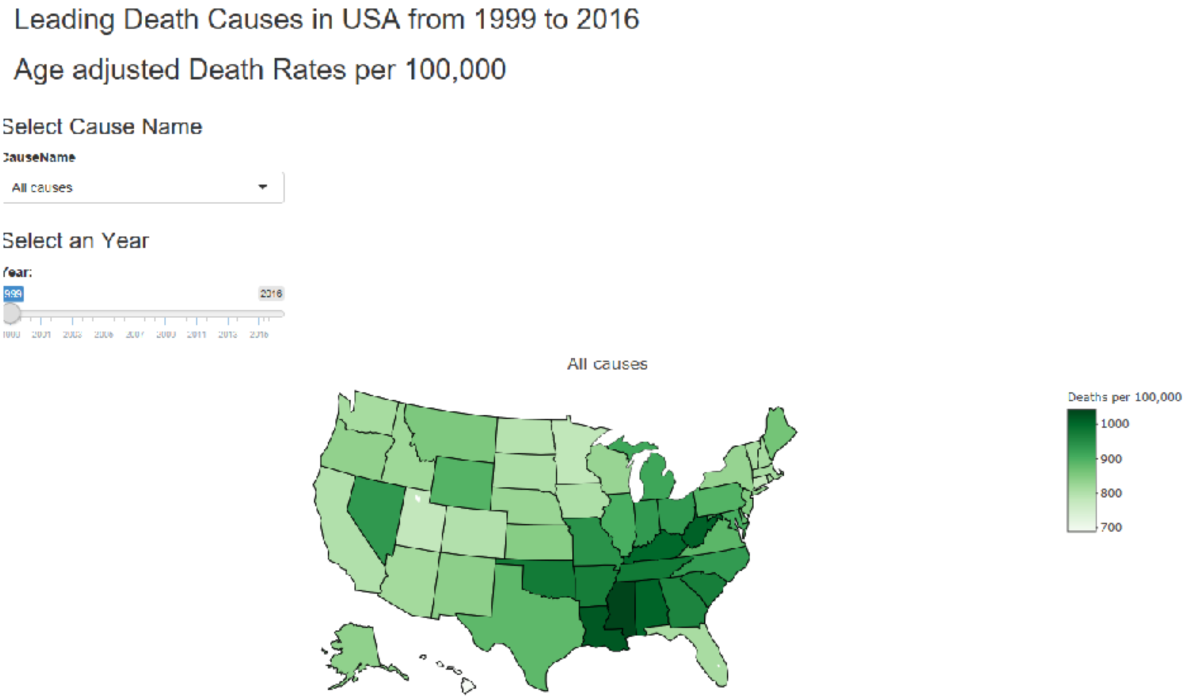


Line graphs are particularly good at allowing a viewer to draw comparisons in the trends over time. In order to help facilitate this, we clearly labeled the x-axis as years and y-axis as a log adjusted death rate per 100,000 and we use subtle grid lines to help the viewer get precise comparisons, even if the lines are further away from either axis. We chose rich colors to make sure the lines are clearly visible against the white background and to add visual appeal. Since many of the groups are close together. it was natural to choose legend over labels over the lines. One approach that we considered was to separate the lines into separate graphs so we could utilize visual groupings of five. This made it more difficult to read as scales were compressed. Additionally, there were no obvious groupings since there are six together and two that are separate with the total death rate which we thought was particularly important to emphasize.

From an interpretation perspective, we can see that the log-adjusted death rate is steadily increasing. This appears to be driven by an increased in log adjusted death rate of heart disease, influenza and pneumonia, and stroke. Interestingly, Chronic Lower Respiratory Disease (CLDR) and Alzheimer's Disease have had a sharp decrease. Most of the variables have changes in rate, but cancer deaths appear to be increasing an incredibly stable rate.

Section IV: Choropleth Representation of Causes of Death in the United States (Age Adjusted)

Figure 3: Choropleth of Causes of Death in the United States (Screen Shot from Shiny)



Leading Death Causes in USA from 1999 to 2016

Age adjusted Death Rates per 100,000

Select Cause Name

CauseName: All causes

Select an Year

Year: 1999 – 2016

All causes

Figure 3 shows a screenshot of our choropleth map application. This choropleth map shows the age adjusted deaths per 100,00 people in the state. Since there were eleven options for the cause of death we decided it was best to not clutter the side menu with radio buttons and instead we chose to use a drop-down menu. We chose to use the slider for time since it is natural to think of time as linear. In general, shades of color are not the best visual encoding; however, the choice of and the choropleth's interactive features allow a viewer to get an exact value for the death rate. This presents a very simple yet powerful representation of the causes of death in the United States. It is a very clean plot and it is very easy to read. It displays that year and cause of death in the top right corner and on the left has a legend with the units of measurement.

Despite the fact that the data is adjusted to deaths per 100,000 residents, we see an interesting phenomenon in that it appears that the most populated states are still the states with the highest death rates. It is also interesting that the death rates for the total population move very little over the course of the 17-year period. There is much more variation in death rates among the individual categories, likely due to the fact that they have much smaller rates.

Section V: Age-Adjusted Linked Micro Map

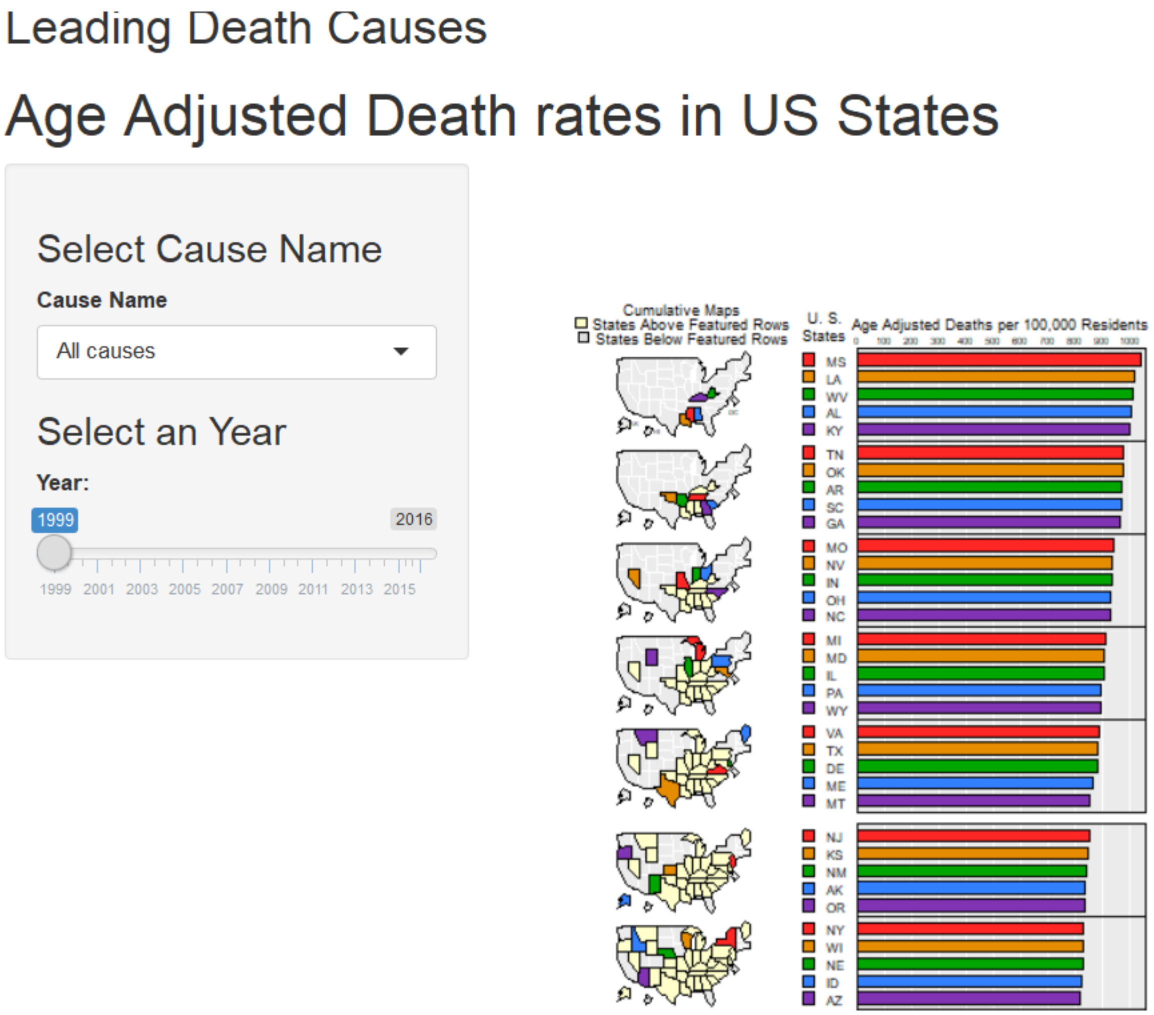
In the final portion of our redesign, we used Shiny to make an interactive Linked Micromap we application. Using drop-down menus, users are able to specify and year and a cause of death and see the states ranked by age-adjusted deaths per 100,000 residents. It was most appropriate to list states in descending order (from highest age-adjusted deaths per 100,000 to the lowest) since we are trying to identify states that have the biggest problems.

The Linked Micromap interface presents quite a bit of information and does is it using small groupings of 5 and is remarkably easy to read. The groupings allow a user to graphically see if there appears to be any geographical connection with the age-adjusted death rate. Linked Micromaps also have a distinct advantage over other types of plots because they draw users in. In addition to the attractiveness due to the grouping, Linked Micromaps have a bright color scheme, they are clean (having not wasted space), and are very approachable (maps are things that most people are very familiar with). The structure of our specific Linked Micromap is set up as follows: the farthest left column is the geographical location for the state, while the middle column is the state code, and the rightmost column shows the age-adjusted death rates. Our preference was to include a time series of the rate of change in the rate, unfortunately, the time series graph is too condensed and it makes it difficult for the user to read.

Another important feature of the Linked Micromap is the Cumulative Map function. It provides light yellow shading in states that are in a higher grouping. It really helps emphasize geographical patterns and helps to the reader to visualize what is going on. Additionally, more importantly it does not require the reader to remember what is going on, they can reference what is going on in a stated based entirely on color (yellow if it is in a higher grouping, colored if it is that grouping, or grey if it is in a lower grouping.

While the Linked Micromap plot is the strongest of the three plots, it does have one major drawback. In order to make the Linked Micromap easy to read, it needs to be relatively large (especially since there is so much information being displayed). Unfortunately, when you make the plot large, it does not fit on one page forcing a user to scroll to see the entire plot. Despite this drawback, it is still the strongest and most descriptive plot for this data because all of the information can be easily seen for each state at one time.

Figure 4: Screenshot of Age-Adjusted Linked Micro Map



Leading Death Causes

Age Adjusted Death rates in US States

Select Cause Name

Cause Name: All causes

Select an Year

Year: 1999

We can also see from our Linked Micromap that there appears to be some connection between eastern and southern states and higher age-adjusted rates. There is quite a bit quite a bit of variation among where the states rank from year to year; however, southern and eastern states still seem to have the highest age-adjusted death rates.

Section VI: Conclusions

The use of line graphs and mapping tools greatly improves the readability of this table of data. Looking at a CSV file or spreadsheet with over 1,000 entries it is very daunting, so trying to extract any information would be impossible. Using standard tools that are emphasized in this course (such as a general descriptive statistic; a linear regression model; or random forests) are not appropriate for this particular data set. Using a graphical redesign is the only logical way to gather trends and look for information. This redesign created an interactive and extremely approachable representation of the data.

There are some very interesting trends that we saw when we were scrolling through the data. First, the number of people per 100,000 is decreasing or several of the top ten indicators (including Alzhiemer's). But while there many that are decreasing, several including Stroke, Influenza and Pneumonia, and Heart Disease are growing at a much faster rate, leading to a consistently positive trend in the death rate. Second, it appears that deaths from chronic illnesses appear to be much more prevalent in the South, while mental health seems to be a much bigger issue on the west coast. We would conjecture that more research would show that there is a correlation between income levels and chronic illness, likely explaining why the south (especially Alabama and Louisiana) has such high rates of death. Third, although some types of death are very concentrated to one geographical area, some, such as influenza and pneumonia, seem to be almost random by year.

Section VII: Improvements to be Made to Continue this Redesign

Although we felt we did have a very successful redesign, we did feel that there were areas that we would have like to go further in the redesign. First, Washington, DC was not included because currently Plotly does not have it in their mapping function. We elected to not include it in the Linked MicroMap as well, although on some level this made it feel incomplete. For the Time Series Line graph, we felt that in a future redesign we could use Shiny to make a web application that could change scales and variables. It would be very useful to be able to be able to choose which variables you would like to compare. For example, if a user wanted to compare the death rates of Diabetes and Alzheimer's this functionality would allow you to compare those two diseases on a larger plot without any unnecessary clutter from other variables. It also would be nice to allow the user to change the scale on the x-axis such as giving the user the ability to look at percentage change. There is more white space than we thought there would Linked MicroMap application, so it would have been nice to include a second column potentially either a rate of change with arrows so people can see how much and which direction the change came from. The biggest improvement that we would have liked to make is with the choropleth. Using Shiny for choropleth and the Linked MicroMap has some risk from the stand point that when people change the bar there could be some degree of change blindness. Being able to have 7 maps on the screen with two of them showing positive changes and two showing negative would be incredibly useful for this particular data set.

Section VIII: Links to the Data and Plots

Data: https://healthdata.gov/dataset/nchs-leading-causes-death-united-states

Line Plot: https://plot.ly/~xiaotan1993/1/#/

Choropleth: https://redesign.shinyapps.io/redesign2/

Linked MicroMap: https://redesign.shinyapps.io/663_Redesign_Project/