Gary Gregg
INFO 474, Assignment #3
16 February 2018

1. Description of the Data Domain and Storyboard

Selection of Visualization Approach and Dataset

I have long been fascinated with maps and geography, and to me a choropleth map is an attractive way to represent data that varies geographically. In fact, I included a couple of Tableau-generated choropleth maps in assignment #2 for this course. In those maps I used a dataset developed by the National Center for Health Statistics (NCHS) that describes the leading causes of death in the United States, by state, for the years 1999 through 2015. This dataset is available in a variety of formats at:

https://catalog.data.gov/dataset/age-adjusted-death-rates-for-the-top-10-leading-causes-of-death-united-states-2013

In my submission for assignment #2, I examined the data in CSV format, and drilled down to examine the statistics relating to suicide as a cause of death in the various states for the years under consideration. Ultimately, using choropleths and a scatter plot, I noticed that the rate of suicide appeared to be higher in states that are more rural. These are states with lower population density, and this suggests that some suicide victims in these states may have become despondent due to isolation. In my final query for assignment #2, I downloaded population density data for the various states from Wikipedia, formatted it in CSV format, and linked it with suicide rates in the various states. Using a trend line, I was able to show a statistically significant, positive correlation between suicide rate, and lower population density in a state.

The NCHS data contains more data related to cause of death than just suicide. In fact, the data give statistics related to seventeen causes of death overall. For each cause, the data give absolute numbers of victims in each state, along with a death rate that is expressed as a number of victims per 100,000 state residents. I thought that there was more exploration to be done with this NCHS dataset, so I resolved to do an interactive exploration for this assignment. As well, I chose a choropleth map is an attractive and intuitive way to present geographically diverse data.

Selection of Software Toolkit

Javascript and D3 were strongly recommended for this assignment. We were able to work with a basic choropleth map in lab #2 for D3 exploration in this course. This example held the potential for a basis of my code for this assignment. As well – at the D3 website – there is a fine example of a binned, sequential scale choropleth. This choropleth, rendered by Mike Bostock, shows unemployment rates broken down by counties in the U.S. The visualization gives a color-coded legend at the top that shows the threshold scale used for a color change in the resulting choropleth. Mike Bostock's code is given at:

https://bl.ocks.org/mbostock/4060606

I resolved to use either the lab #2 source, or Bostock code as a basis for my own design, but until I began the actual coding I had no information with which to determine which example would be a more useful starting point. See section #2, **Implementation of Design**, for a discussion of this.

I decided that I would again use the NCHS data, and focus only one one cause of death – other than suicide – in the various states for the years under study. The interactivity part would be that the map would update, using buttons, for the year to be viewed. Thus there would be buttons for each year from 1999 through 2015. Alternatively year selection could have been accomplished with a drop-down list.

Storyboard

Whether I opted to start with the lab #2 source, or the Bostock code, I decided to use a red color scheme scale instead of the blue used by Bostock. I thought a red shading would more appropriately convey the exploration of untimely death that is addressed in the visualization. I conceived that the visualization would appear as shown below. Imagine a graded red shading within each state boundary. Additionally, the label for "currently displayed year" would update, along with the map, to reflect the most recently selected year under consideration. The storyboard diagram appears below:

Description of the Visualization, and Thresholds for Sequential Color Scale

Currently displayed year is 1999.

1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011

2012	2013	2014	2015

Why is the Choropleth Appropriate for these Data?

The data supplied by NCHS is obviously specific to states, and use of the choropleth gives the viewer an opportunity to look for trends that may be regional, instead of simply confined to a state. As an alternate, an effective representation of the data might be a bar chart using length to show the quantitative data in an easily comparable form. However, this representation does not allow for that same close comparison by region. The interactive aspect of the visualization allows for easy comparison, by button push, for the different years under consideration.

2. Implementation of Design

Selection of Cause of Death

As mentioned in section #1, **Description of Data Domain and Storyboard**, above, that NCHS data that I selected contains data on seventeen causes of death, given as absolute numbers and rates, and broken down by states and years from 1999 to 2015. To more adequately scope the assignment, I resolved to use only one cause of death for the visualization. I used the R programming language to read in the NCHS data in CSV format, and examined the summary statistics for each cause of death. In particularly, I was looking for the cause that had the most variance. My intuition told me that this would make for the most effective choropleth, as the wider range of values would produce the most contrast in the resulting map. I determined that heart disease had the widest variance of the causes under study, and thought that these would be excellent statistics to present in the map. I used R, again, to isolate the table rows for heart disease, to rename the table columns, and remove rows for the whole of the United States, which could not be shown in this choropleth. I did more in the R code, as noted below. However, I will here note that the R code I wrote to cleanse and modify the data for this assignment is included in the GitHub repository where this document resides. The interested reader may examine it.

Basis for Code

I played with both the lab #2 code and the Bostick code to determine which would be a better basis for the start of this project. The lab #2 code needed to have the information dots removed from the state centers. Alternatively, they could have been retained to present the state name, and exact death rate for the year in question upon mouseover. Ultimately, I chose to remove the dots. The lab #2 code would also need to have a sequential scale inserted so that the color data in the map could be interpreted by the viewer. A more serious difficultly, though, was that the lab #2 code has no example of how to fill a state region with color. The Bostock code, however, did supply that example, albeit at the county level. The Bostock code also supplied an example of color scale legend. With these obvious advantageous, I chose the Bostick code as the basis for this assignment.

Modifications to the Code

It was small change to cause the map to fill color at the state level instead of county. As well, it was reasonably easy to change the color scheme from blue to red. I determined, though, that when the software calls the code to fill a specific state on the map, it does so with an ID code that is without an easily discernible correspondence to the state name. I resolved the encoding by painting each state

red by default. Then, I looked for specific state codes by using a hardcoded ID, and then painting the state blue when the fill listener received that code. By viewing which state was colored blue for a specific ID, I was able to see that the states were encoding in ascending fashion according to their alphabetic ordering, starting at “01” for Alabama. However, there were gaps. I noticed that the codes “03”, “07”, “14”, “43”, and “52” were not used. Including the District of Columbia, therefore, the state codes ended at “56” for Wyoming.

Finalizing the Input Data

I entered the state encoding I had discovered by hand into a new CSV file, indexed by the same state names as used in the NCHS data. Using my R code, again, I merged the two datasets: the first, indexed by state and year giving absolute numbers and death rates for heart disease, the other; indexed by state and giving the state ID used in the D3 map. With the data thus merged, I reordered the columns in the dataset to “id”, “state”, “year”, “count” and “rate”. I then sorted the dataset by “id”, “state” and “year”, and wrote it out to a CSV file for use by the interactive visualization. The interested reader will find the aforementioned CSV files in the same GitHub repository as this document. These CSV files are:

NCHS_-_Leading_Causes_of_Death__United_States.csv	NCHS Data
state_ids.csv	State IDs
heart_disease.csv	Merged File

More Modifications to the Code

After reading in the merged dataset, described above, it took some experimentation to get the map to render correctly. In particular, adjusting the linear scale used by the legend took some work, as it did for the threshold scale used by the legend and the map. I needed to remove data references to unemployment data which were used by the code in its previous life. The title of the legend needed to be adjusted. I debated somewhat how to do that. Since the new data used by the map gives death rates as number per 100,000 persons, I decided to adhere to that statistic in the map, without modification. The index used in the D3 map to store the data needed to be modified to accept the concatenation of state ID with year, and this was reasonably easy to do. For each thus created index, I stored the death rate per 100,000 that is associated with the state ID and year.

I needed to add widgets to the HTML to match the storyboard that I had created. I needed seventeen years from 1999 to 2015, inclusive. Additionally, I needed a label to described the currently displayed year. Upon selection of a year button, I updated a variable I had created to hold the current year, updated the year

display label, and then I repainted the map to show statistics for the current year. I am quite pleased with the informative nature of the resulting interactive visualization, and its aesthetic appeal.

3. Final Writeup

Manifest of Assignment Deliverables

The assignment description asked for a storyboard(s) for the visualization, and that has been provided in part #1, **Description of the Data Domain and Storyboard**. In part #2, **Implementation of Design**, I give a detailed description of the resulting interactive visualization application. The source code for this assignment is given in two identical files contained in this repository: **garygr_info474-a3.html**, and **index.html**. The **index.html** file will allow any web browser directed into a directory containing the files to display the interactive map by default. There are three CSV files included which were discussed in part #2. There is also a JSON file that is part of the code that is entirely borrowed from Bostick for rendering the states. There is the R code that I used to process the data from NCHS, and also to merge my hand-crafted file for associating state IDs to state names. There is the document that described the requirements for this assignment, a basic README file, an MIT license file, and, finally, this document. In order to view the interactive visualization, a user will need to start a web server for the appropriate platform they are using.

I worked alone on this assignment. All work is my own, aside for the javascript basis code borrowed from Mike Bostick as given at the D3 website. Overall, I spent approximately 13 hours on this assignment. The most time intensive tasks involved getting the color fill in the states to occur properly, properly determining a good range for the sequential fill colors, determining state ID codes used by the D3 software, and creating and properly managing the pushbuttons. Lastly, this writeup took approximately five hours.

Differences Between the Storyboard and Implementation

There are few differences between how I initially conceived of this visualization, and the final product. I did discover some difficulty in changing the range of the mapped values from the 0 to 1 scale used in the Bostick code. I noted in the data exploration phase that the highest death rate from heart disease for any state in any year was 347.4. This is to be interpreted as the number of individuals that died of heart disease, per 100,000, for the state and the year in question. It seemed easiest, therefore, to express the unit of measure for the visualization as a fraction of 350 individuals per 100,000. This, then causes the scale of the legend and the colors used in the map to continue to conform to the 0 to 1 range.

I did not plan for this level of detailed thinking when conceiving of the idea for this map.

I originally envisioned that this application would have appropriate frames for the display map, and the labels and pushbutton widgets that would control the map. This feature proved to be too work intensive to implement in the time allotted. Other than these minor differences, the visualization appears exactly the way I intended it.

Findings

Because of the easy comparison of states based on geographic region, it is easy to see – particularly in the earlier years – that death rates from heart disease are highest in the south, and industrial Midwest. I theorize that this may be due to the effects of smoking in the south, and, perhaps the effects of industrial pollution in the Midwest. The rates are lower in the west, except for California and Nevada. Is this too caused by industrial pollutants in the major cities in California, and possibly Las Vegas, in Nevada?

Here is one very heartening finding: As the viewer progresses through the map year by year, without exception there are lower overall death rates from heart disease. This is almost true for each individual state as well. There are some backsliders in the trend, but not many. My theory in this regard is that reduced rates of smoking have been responsible for this change. Some smokers (or former smokers) succumbed to inflictions associated with that habit, including heart disease. As well, some smokers quit the habit, and developed better health. Also, we must consider advancing medical technique for treating this condition. For all of these reasons – attrition of smokers, better health of non-smokers and former smokers, and advances in medicine – the death rate from heart disease over the years is most decidedly down.