Hannah Dunn
Leopoldo Peña
Natasha Armbrust
INFO/CS 3300
28 April 2016

Project 2 Written Description

Data:

We obtained our data from the University of Wisconsin County Health Rankings (UWCHR) 2016 and the American Community Survey. The first dataset was found at http://www.countyhealthrankings.org/rankings/data. We looked at several variables within the datasets. The UWCHR dataset contained all of the counties in the United States, and for many different health and lifestyle statistics it provided the percent of people in that county affected, what quartile the county was in for that statistic, and often other data such as raw number of people affected or ratio of doctors to people. Our second data set came from the https://www.census.gov/programs-surveys/acs/data.html ACS data. This provided statistic on counties. We used this dataset to acquire income information on all the counties. Our goal was to show what a county's income had to do with various health factors in the county. We used both these datasets by integrating with datausa.io API and made API calls to create csvs with the county info, health stats and county info, income stats from both datasets. The API calls are in a text file in the zipped folder. Once we had the csv files we used d3 to upload them and create two data structures. One was an array on county objects that combined the health stat data from UWCHR and the income data from ACS. The other data structure was an object which had keys of county ids. This allowed us to do constant time lookups once we knew the county id to get all the health statistics and income data.

For the Sankey graph, the data needed to be formatted as a JSON, but also only the 9 wealthiest and poorest counties needed their data included. Using the array of data formatted above, we wrote a javascript script (json_script.js) that would sort and parse the data into a JSON containing nodes and links. To do so, we first parsed the wealthiest and poorest counties into an array. Then, for each health statistic, we created a new sorted array. For each element of each array, we added a node containing the county, the health statistic value, and the geo code to the list that would be turned into a json string. Also, for each array except that last statistic array, we added a link from each element to its corresponding county element in the next array containing the source and target node value. Lastly, we took the array of all nodes and links and ran JSON.stringify in order to produce a JSON string that could be used for the Sankey graph.

Mapping from Data to Visual Elements:

The first major graphs on our visualization consist of a scatterplot and an AlbersUsa map. When a specific health statistic is clicked, the scatterplot will map each county from the data based on income and frequency of that health statistic. For this graph, the y-axis

represents income, so this is mapped on a linear scale ranging from minimum income to maximum income. The x-axis, on the other hand represents the frequency of the health statistic, so it also is mapped on a linear scale but it ranges from the minimum value of the health statistic to the maximum value of the health statistic. Additionally, points are plotted with a different color depending on what the health statistic is and how the county performs in that health statistic. Each health statistic has its own color scheme in order to show a distinction and keep the graph visually interesting. Then, based on each county's position on the x-axis, using a linear scale and an array containing a gradient of colors, it is given a color based on what interval of the x-axis it falls in. When the health statistic buttons are clicked, the United States map also changes color. Each county is given a different color based on the color scheme determined in the scatterplot graph. This way, it is easy to look at the United States map and, for each statistic, find the highest and lowest concentrations.

The scatterplot and the United States map both have compatible interactive elements. When any county in either graph is hovered over, a tooltip containing the county's name, income, and health statistic is displayed. For more interactivity, users can click on the "Explore Levels of Income" or the "Explore Health Stat" button and hover over the scatterplot. This will highlight intervals of either the income or health stat on the scatterplot as the mouse moves around, and it will also make any county that is not in that interval gray on the United States map. This functionality was not included in d3 and thus took a long time to formulate. Upon mouseover of the svg we create a rectangle on the svg. We filtered the counties by income level or health stat depending on if the Explore Health Stat or Explore Income was pressed. We then highlight those counties on the graph and in the scatterplot. Although there is lag time we really wanted to include this feature. This novel approach allows a user to more deeply explore all these health stats. More importantly, it allows the user to explore brackets of income and the relative color of the map for that income bracket. What we wanted to show was the in a majority of the health statistics the lower income hover over shows darker colors than the higher income hover over thus indicating that lower income means poorer health.

To further explore the graphs we added functionality on click of a county on the scatter plot and the USA graph. This appends an svg element with that county and all the health stats plotted against the average for that health stat using bar charts. The y-axis is the min and max of the health statistics value and the x-axis indicates the county bin and average bin. The color indicates the color scale that county is in for that specific health stat. We wanted to include this graph so that a user can explore specific counties. More importantly, we wanted a user to be able to pick out outliers in one health stat and see if a county is an outlier in other health stats as they are able to compare its statistics to the average within each category.

Additionally, we added a Sankey graph at the bottom of the visualization. This graph displays the nine richest and nine poorest counties and displays how they compare to each other based on each of the health statistics. The color scale used for this graph was simple. If a county was wealthy, its nodes and links are displayed as a purple color - the color of wealth. Poor counties were represented with brown nodes and links because brown represents poverty. Lastly, if a county has no value for a health statistic, the node and links on either side of the node for that health statistics were displayed as gray. To interact with this graph, viewers can

hover over any node to see a tooltip showing the corresponding county and health statistic value. Hovering over any node or link also lowers the opacity of other counties so it is easier to see all the nodes of the county in focus.

The Story:

With our visualization, we wanted to explore how income of a county could affect the health of that county. With the scatterplot and the United States map, it is easy to quickly see that some health statistics seem to be highly correlated with income. When this is the case, if the “Explore Levels of Income Button” is clicked, as someone hovers over the bottom of the scatterplot (low income), the United States map starts out with highly concentrated colors that represent a high occurrence of the health statistic. As the viewer moves the mouse higher towards high income, the colors on the map become much less concentrated. In these cases, it is interesting to see how income levels can really affect health. However, we soon realized that this is not always the case, and some health statistics are less highly correlated with income. For this reason, we decided to do a comparison of the richest and poorest counties in the United States, and created a Sankey graph. While this is a less quantitative graph, it makes it clear that richer counties are not always better off. While there is a distinct advantage in some categories, such as having lower rates of diabetes in the richer counties, other categories like rates of alcohol deaths show that even the wealthiest people aren’t immune to having the same problems as those who are in poverty. Between these three graphs, it is easy to see that while income can definitely be correlated with some statistics, it cannot be said to be responsible for all health statistics.

This visualization ended up being surprising, because we initially expected that every single health statistic would be correlated with income. It is often believed that those people who have more money cannot face adverse conditions. This visualization really showed that even the people who are seemingly the most “well off” may be afflicted with a health problem.

At the end of our visualization we added a trends to note section. We added this because we ended up finding some very interesting trends with our visualization that were not solely about income but were revealed because of the way we visualized the data. Check out that section for the trends we found.