Divy Tripathy
CS 5630

Process Book

Overview:

Access to healthcare is something many Americans take for granted. However, as many as 8.4% of Americans do not have access to healthcare. I am an economics student and access to healthcare is a hot topic in the field of economics. I want to know how health outcomes are related to the number of hospitals. I think this will be a good project because it will give me a chance to query government databases to procure data. It will give me ETL experience, and display the outcomes on a map (something I don't have experience with yet). A benefit is that the visualizations that come to mind are intuitive to understand (map, barchart, etc.).

Related Work:

I am pursuing this project due to the inspiration from my colleague Mutiu from the economics department. His research in the area with his findings lead me to believe that this work may greatly benefit from visualization.

Project Objectives:

The primary question that I'm trying to answer with the visualization is: how does the number of Hospitals (or some proxy of) impact the health outcomes of a community. Since health outcomes is a broad notion, we can try and approximate health outcomes by looking at life expectancy. Life expectancy would be a good proxy because we would expect healthier people to live longer. I would like to see whether there are patterns across the country depicting the number of hospitals and life expectancy. I also want to know whether or not there are distinct outliers in some parts of the countries. Where the data for life expectancy may not be available, I may be able to use other proxies such as child mortality. Despite the fact that we can use a simple hypothesis test or build a linear regression model to answer our question, I think that a visualization would really help show the nuance in the data very effectively.

The benefits would be that we can very clearly see trends in different parts of the country with a quick glance at the visualization (currently plan to use a map with a colormap). This would be easier to quickly decipher than looking at the output from a linear regression, and we get additional information by using this map instead of using the outputs from linear regression (which can take a while to process because of the ceteris paribus assumption). Looking at the map can be helpful because we can quickly identify the areas that have bad health outcomes in case we want to draft policy to target aid to these areas. We might also see areas that have fewer hospitals and better health outcomes. In this case, we can use these areas as case studies to answer how other areas can have better health outcomes with fewer health resources like hospitals.

Data:

The Bureau of Labor Statistics (BLS) collects state level data measuring the ratio of people per hospital, hospital establishments in the third quarter of 2019, and population on July 1, 2019. All of this data is quantitative data. The link to the data is https://www.bls.gov/opub/ted/2020/number-of-hospitals-and-hospital-employment-in-each-state-in-2019.htm**.** I can either scrape the data or manually write it down.

The Center for Disease Control collects data on life expectancy by state. The data is tabular, and has the attributes of the location, and life expectancy. The location attribute is categorical while the life expectancy variable is quantitative. This is the link to the CDC data is https://www.cdc.gov/nchs/pressroom/sosmap/life_expectancy/life_expectancy.htm. There is a link on the page to download the data as a csv.

In addition, since the data has a spatial component in it (U.S. states), I was able to acquire a topoJson file with the information to draw U.S. states. This information was from https://github.com/topojson/us-atlas. This data had already applied a projection, so I could skip

that part. This data is supposed to help the bar chart by better showing which states clearly have a higher ratio of people per hospital employee by state.

I was also able to get a GeoJSON file drawing Utah counties from the state of Utah's website. This is the link where I got this data: Utah Counties Shape File | Utah Open Data. In addition, I used American Hospital Directory - Individual Hospital Statistics for Utah (ahd.com) to find the number of hospitals in Utah by the city. In addition, I used the state of Utah's website to scrape the data for the cities and counties in utah. Here is the link to this website: Alphabetical Listing of All Utah Cities/Towns. Finally, I got data for the population of Utah by county from Utah Counties by Population (utah-demographics.com).

Data Processing:

I didn't do much data cleaning/ processing for the U.S. choropleth. This is because I already found good sources to take data from where the data was "clean". The data that wasn't readily available, I scraped and organized it into a clean format. I may do minor operations such as the concatenation of data frames to combine the data I would collect from different sources. One big operation was to convert the CDC data from postal abbreviations to full state names by using a pandas apply function. The data was from State and Territory Abbreviations | Boy Scouts of America (scouting.org).

I did a lot more processing for the Utah level data. Most of this was to get all the sources into one dataframe that I would feed into my visualizations. I did things such as rename columns, groupby's and sums, extracting parts of column strings to make it work better with my geoJson file. All of this work can be found here: https://colab.research.google.com/drive/1ZwwHdXa_eUnyxi36n84rps2UIWpy41se?usp=sharing .

EDA:

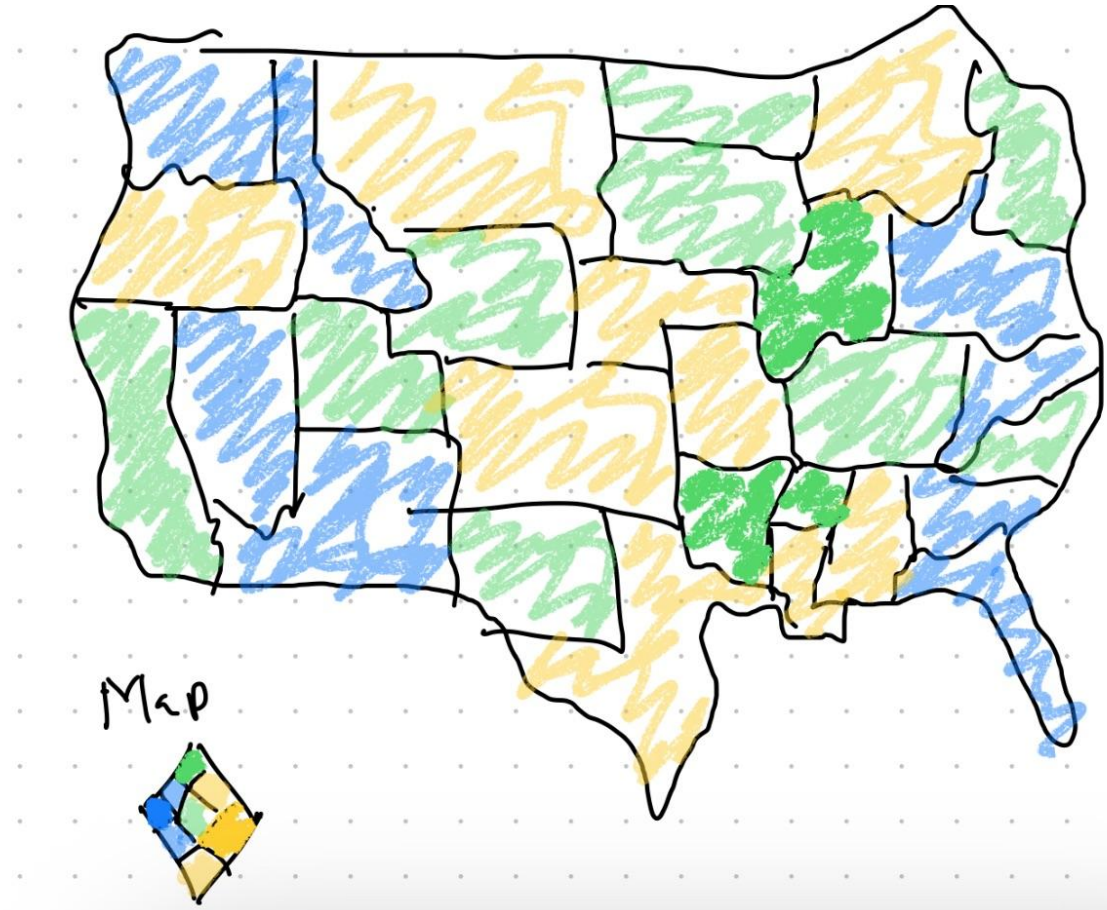
Most of the EDA was done through a combination of making the visualizations and making the data in the right format in python. This can all be seen in my google collab link and my final

project website. Again, the collab link is https://colab.research.google.com/drive/1ZwwHdXa_eUnyxi36n84rps2UIWpy41se?usp=sharing .

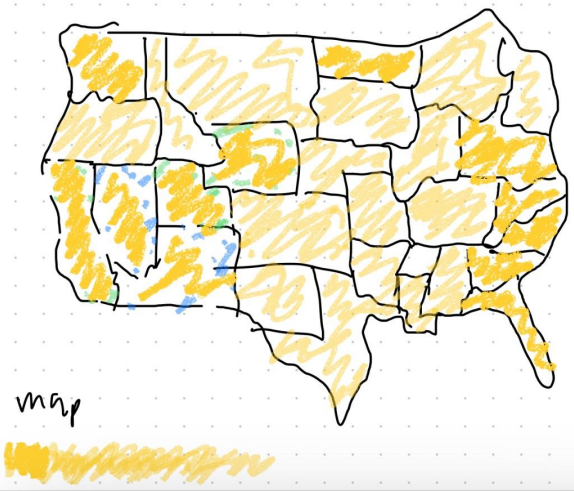
Design Evolution

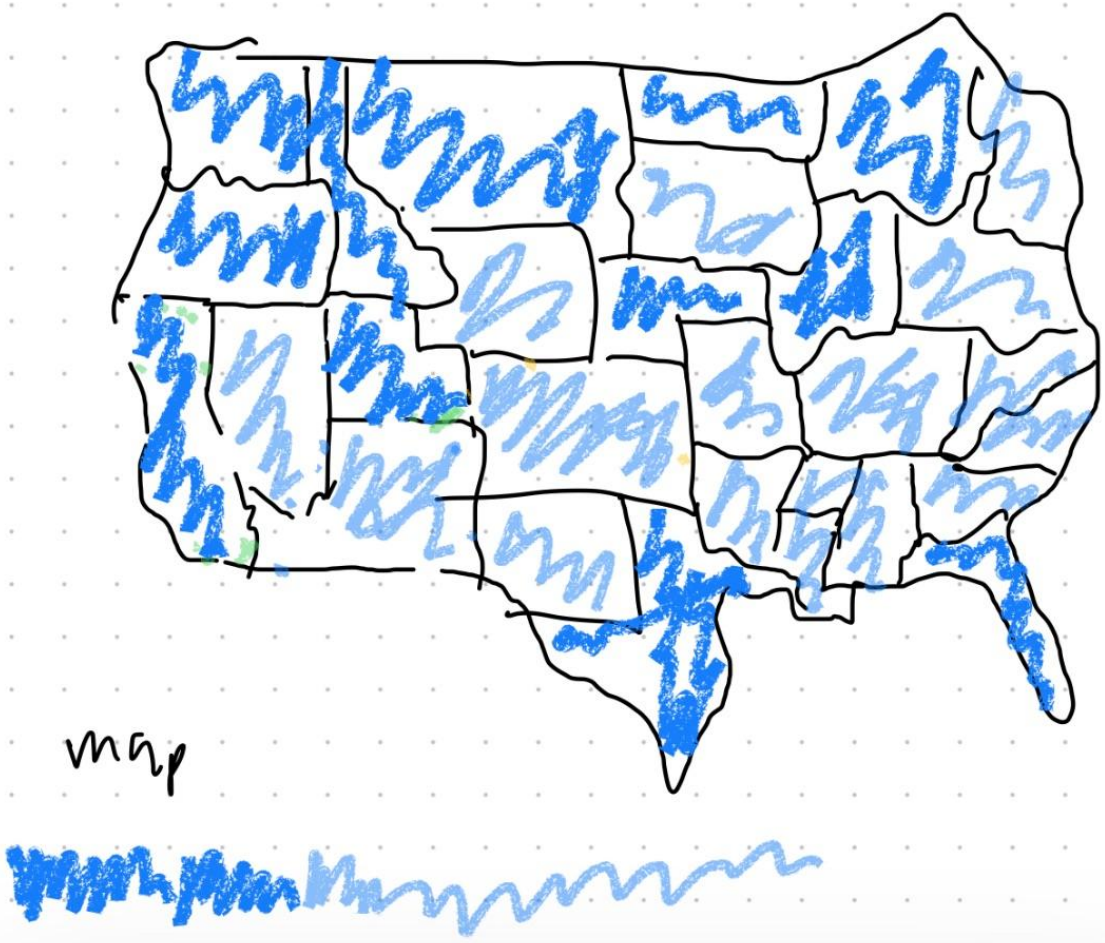
The current designs I have currently used are a bar chart and a choropleth complementing the bar chart. Below are other designs that I've considered/ plan to use.

I considered using a Bivariate choropleth map of the U.S. to make my visualization. I want to use the color blue to represent higher life expectancy. This choice is because a lot of people associate health and wellbeing with the color blue. This will aid with a more intuitive understanding of my visualization. I want to use the color yellow to represent the number of hospitals. The more saturated the yellow, the higher the amount of hospitals present. I want to use the color yellow to represent the number of hospitals because many people associate yellow with happiness. I want to help the viewer make the intuitive connection between more hospitals being a good thing. https://en.wikipedia.org/wiki/Multivariate_map



The design that I went with is two separate maps. The benefit is that we would be able to see the individual spots better instead of conflating the hue of the map to different attributes. I used blue for the number of hospitals and orange for the life expectancy for the previously mentioned reasons. Below is a sketch of some prototypes.





In addition, I considered using a Point map. This map can focus on areas that are outliers and areas doing poorly. For example, if Alabama is doing far worse than the nation's average life expectancy, we can show a large red circle. A benefit to this visualization could be that it would get rid of what a lot of people might consider clutter when they look at the map.

Aside from choropleths, I also plan to use a bar chart. It would display the same information as the choropleth, but it would show different aspects of the data such as cumulative values as opposed to relative values. I found in practice that choropleths and bar charts complement each other quite well.

Finally, the best option may be to use both the bivariate choropleth map and the individual maps or the two choropleths and two barcharts. This would have the benefit of showing the interaction between life expectancy and the number of hospitals, and we would better be able to see where there are more/less hospitals and where there is more/less life expectancy.

As a final result, I ended up using both choropleths and bar charts. The combination of these visualizations produced more information than using a bivariate choropleth. This is because the bivariate choropleth was just a regurgitation of the two individual choropleths.

Implementation:

So far, I have created a bar graph depicting the ratio of people per hospital employee by state. The data that I needed to procure for this step was obtained by the bureau of labor statistics. They are known for the quality of their data collection. The data was not readily available for download, so I had to manually write the data down into an excel sheet and then, I had to export the data into a csv file so that I could use it for analysis. Once I did this, I was able to create a bar chart. I chose a bar chart because it would be simple to show the ratio of people per hospital employee by state. If my data was categorical, it may have made sense to use other types of visualizations, but this chart seemed to be a good way to visualize my data.

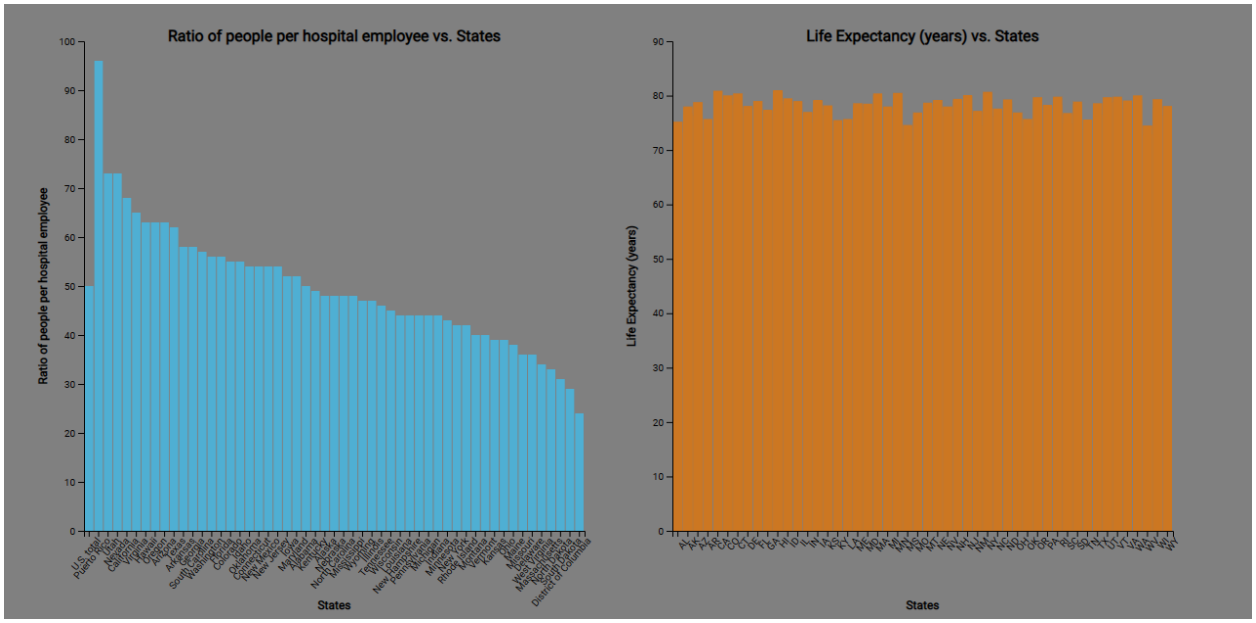
I also created a bar graph showing life expectancy by state. I was able to use the CDC data as is by state to make the bar graph. I placed this bar graph next to the one described above as a juxtaposition.

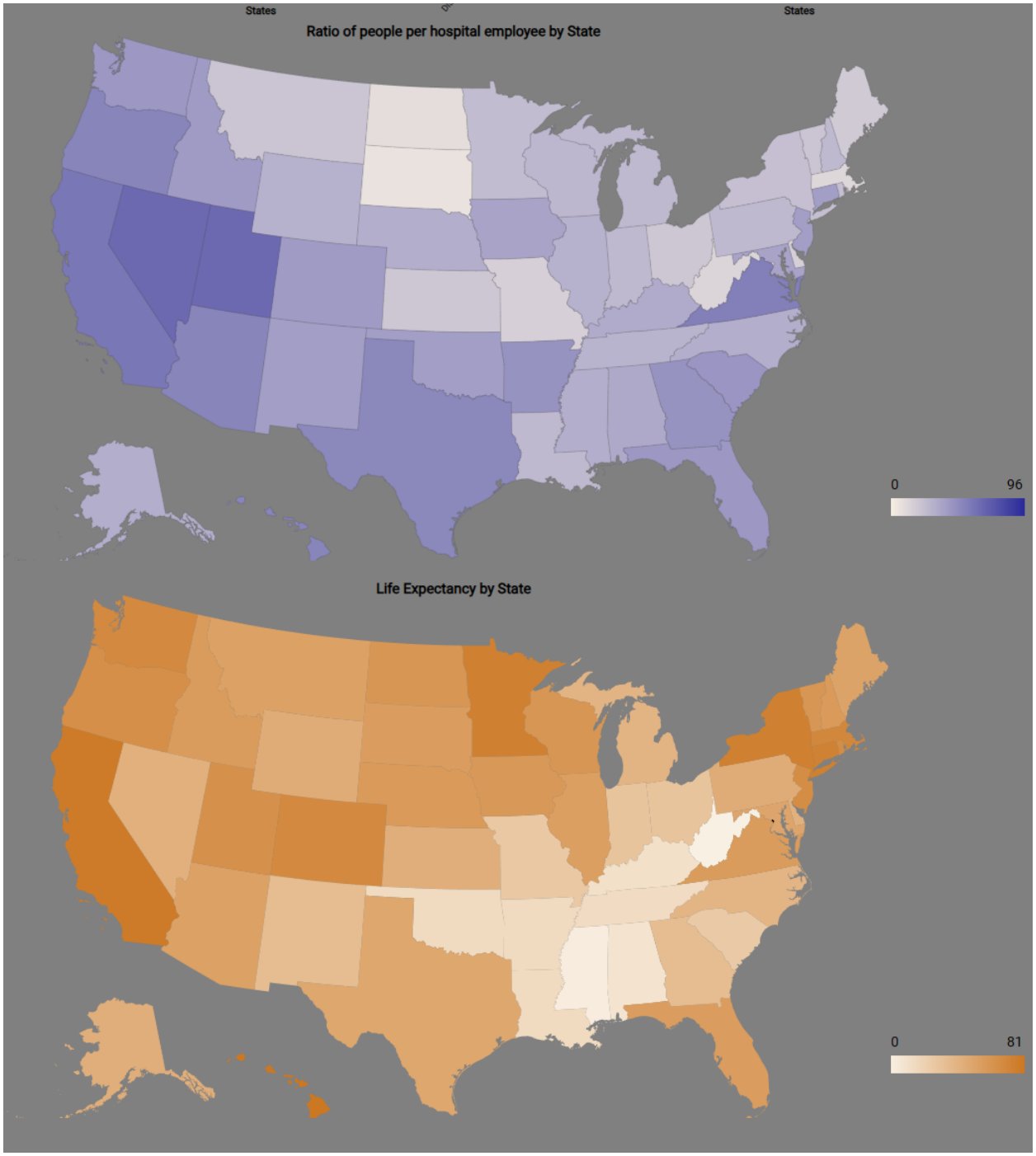
In addition, since the data has a spatial component in it (U.S. states), I was able to acquire a topoJson file with the information to draw U.S. states. This information was from https://github.com/topojson/us-atlas. This data had already applied a projection, so I could skip that part. This data is supposed to help the bar chart by better showing which states clearly have a higher ratio of people per hospital employee by state. The data in the bar chart is a lot when looking for the value of a single state, the map alleviates the problems created by the bar graph. It is easy to look for specific states and see which states have a higher ratio of people

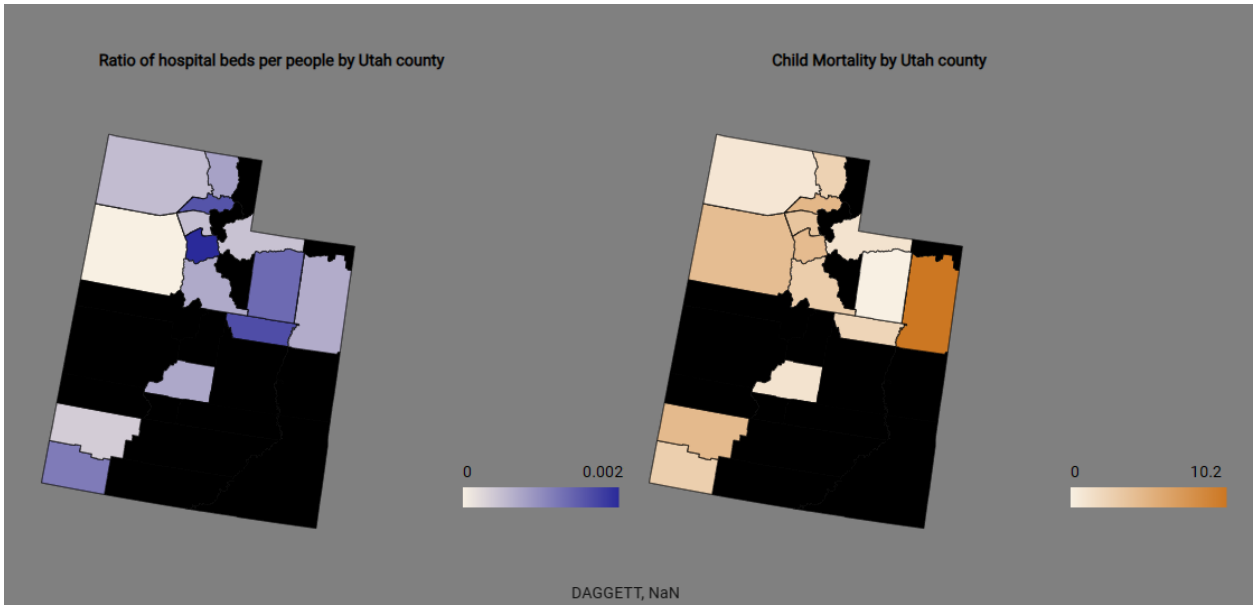
per hospital employee by state and which ones have less. I also added a gradient to be able to tell what the range of values are. In addition, I added an interaction on hover that tells you the name of the state.

I also placed a second choropleth of the US below depicting life expectancy by state. The design is similar to the one above including the interaction. I chose to put the choropleth below since the US choropleth is quite large.

Finally, I made two Utah choropleths below to see a Utah level analysis. One of them shows the number of hospital beds by people by county and the other one shows child mortality by county. There is an interaction that tells you the value for the county below the visualization. The counties with missing data are in black.




States
Ratio of people per hospital employee by State
States
0
96
Life Expectancy by State
0
81



Evaluation:

I learned that there is an imperfect correlation but positive correlation between the states that have a high ratio of people per hospital and the life expectancy. This implies that as we have fewer hospital workers by the number of people, the life expectancy tends to decrease.

On a state level, I noticed that the counties that have a higher level of hospital bed for its population tend to have a lower number of child mortality. This implies that a higher number of hospital beds/ health care availability is linked to higher health outcomes. Both the national and the state level data supports this claim.