N328 Final Project Documentation

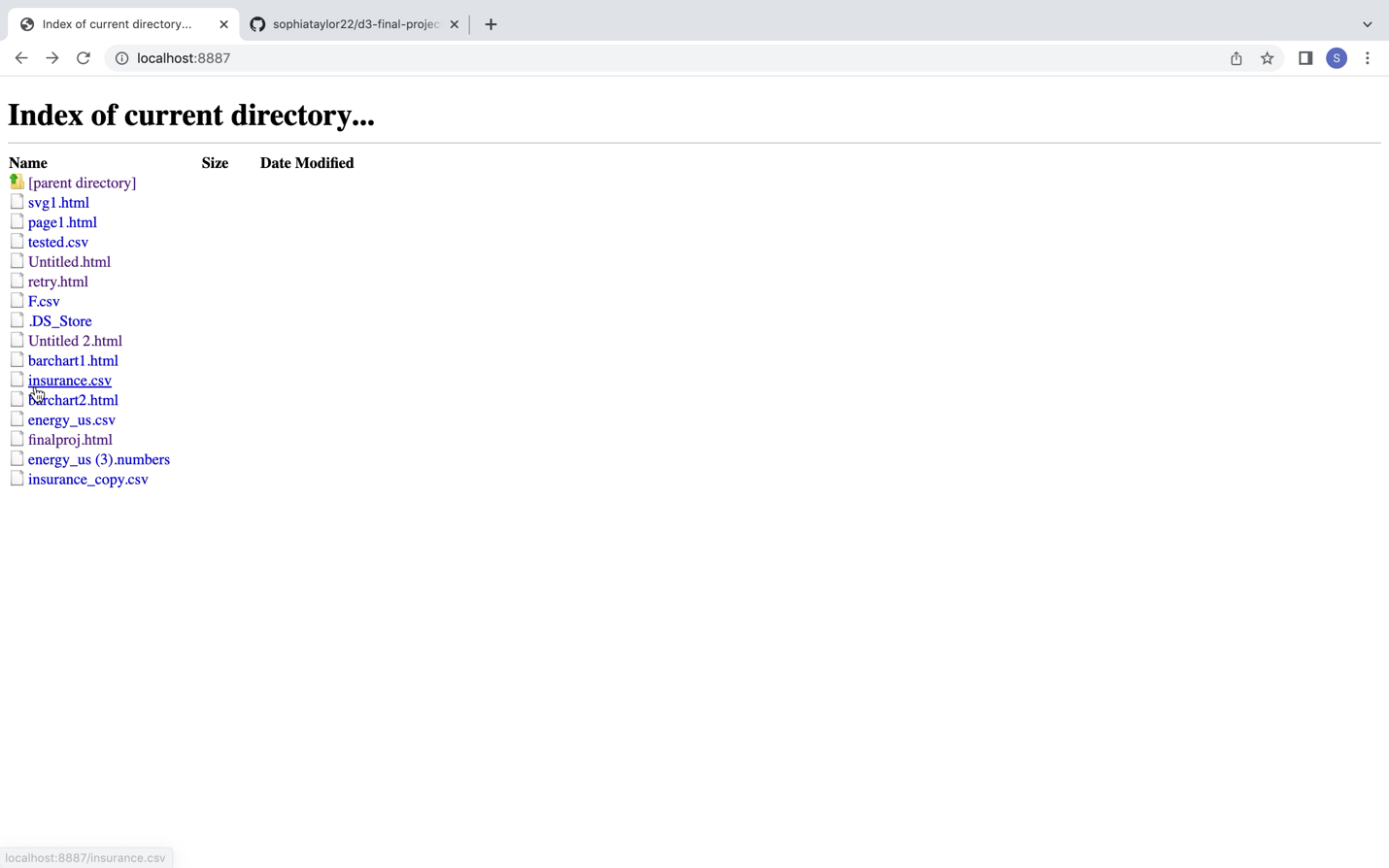
Sophia Taylor

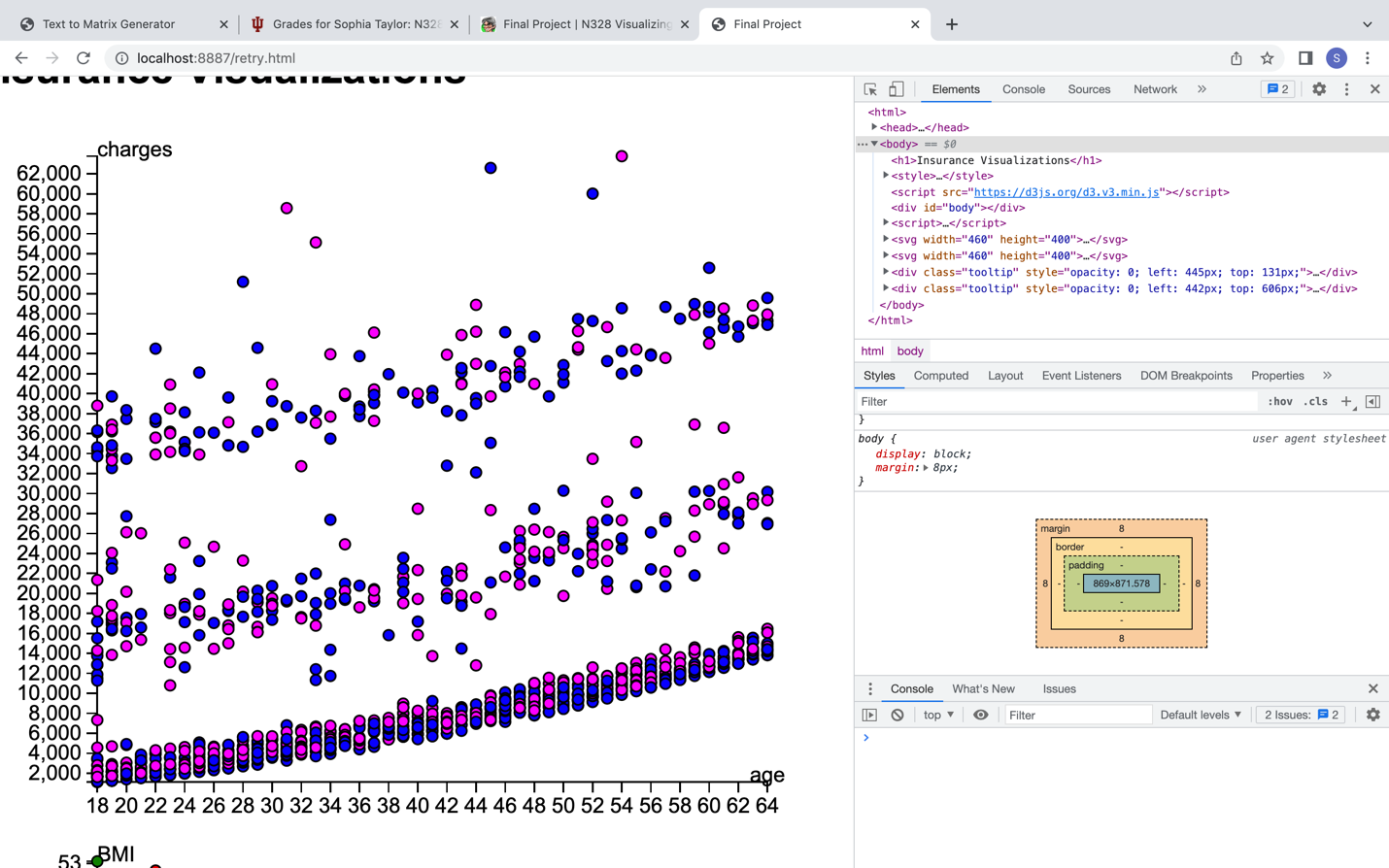
For my final project, I found a dataset about insurance and the demographics of the primary holder. It included the age of the holder of the insurance, the gender of the holder, the holder’s BMI, the number of children belonging to the holder, whether the holder is a smoker or not, the region where the holder is located, and the amount of money charged to the insurance company. I decided in order to include the 5-10 attributes, I would make a couple different visualizations dealing with different attributes. When I first started this project, I thought I would make a bar chart, but I quickly learned that would not represent the data well. After some discussion, I settled on making scatter plots to describe the data. I thought I could use the thousands of rows of data to make a bar chart; however, that was simply too much data to represent it properly. The scatter plot allows many more points to be seen, and it’s easier to see trends in the data. I also went through several datasets before finding this one. This one had a lot less attributes than the other ones, and it seemed simpler to work with. Another problem I had with this project was getting the data to be interactive within the graphs. I settled on using the mouseover interactivity on the points in the scatterplot. I thought I would be able to connect another graph to each of the visualizations (like male and female), but I couldn’t figure out how to do that. Instead, I included different colors on both graphs to represent a third attribute, along with the x and y axes. I had trouble figuring out how to add a legend to the graphs as well, so I will explain what the colors on each graph mean in the video walkthrough.

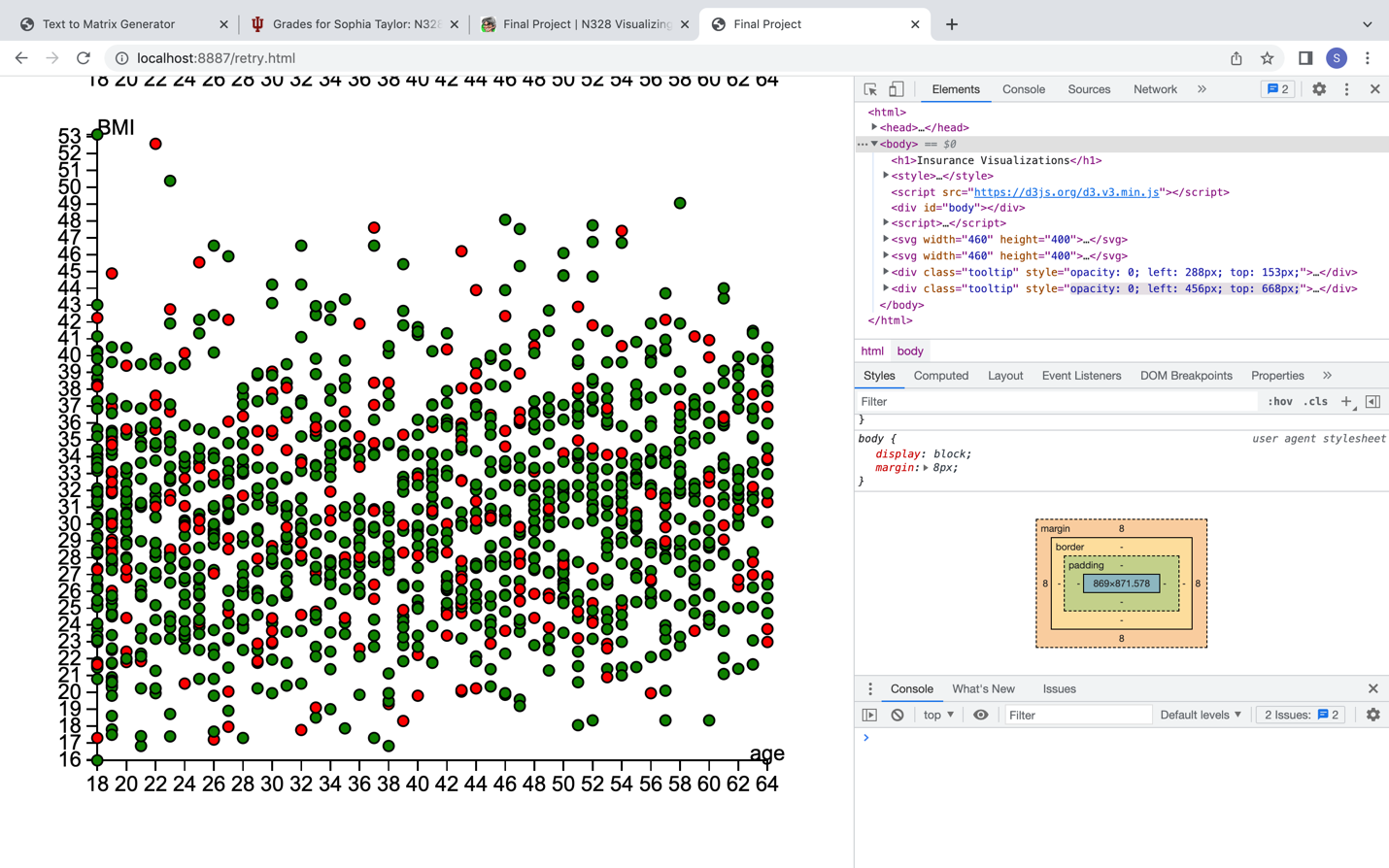
The first question I chose to answer was “Do older people charge more to insurance?”. For this question, I selected the variables “age”, “charges”, and “sex”. The age attribute is on the x-axis and the charges attribute is on the y-axis. Male and female is represented by the blue and pink data points on the scatter plot. I chose blue and pink to represent male and female because those are the common colors associated with those genders. I wanted to look at the gender of the policy holders as well, which is why I included the male and female colors. There even appears to be close to an equal number of males and females represented in the data. It doesn’t appear that just men or just women charge more to insurance either. I chose the variables age and charges because I had a general idea of the shape of the plot. I thought there would be a linear association between age and charges. From the visualization, you can see there definitely is a relationship between the two. There are three separate lines formed across the graph. One starts at the bottom, which I would say are the people not charging much to insurance. All these data points are tightly packed and form a straight line. The second line forms across the middle of the graph, and it appears these people charge more to insurance than the first group. The data points in this group are not as close together and don’t form as straight of a line as the first group. The third line is above both the others and has many outliers above it. This group charges a large amount to insurance and the data points aren’t as close as the first two groups. I chose a scatterplot for this data because it allowed each datapoint to be seen. It allowed trends to be seen easier as well, and the point of this graph was to see if there was a relationship between age and charges.

The second question I chose to answer was “Do older people have a higher BMI than younger people?”. For this question, I selected the variables “age”, “BMI”, and “smoker”. The age attribute is on the x-axis again and the BMI is on the y-axis. Whether a person smokes or not is represented by the red and green dots on the graph. If a person is a smoker, the datapoint is red. If a person doesn’t smoke, the datapoint is green. I figured these color options make sense because smoking is considered “bad”, which I associated with red. If a person doesn’t smoke, that’s “good”, and I associated that with green. This graph is a lot different than the previous graph. There wasn’t much correlation between the two variables I selected, and I was surprised by that. There were a lot less smokers than non-smokers so I can’t say for sure if smokers had higher BMIs than non-smokers. Most of the datapoints were close to the low 20s to the mid 30s. There were some below in the teens and some up over 50. It appears age isn’t a factor of BMI, which I learned from an internet search. I’d say this visualization was a lot more surprising than the previous because of the lack of a relationship between the variables. I’m also surprised that smoking wasn’t associated with higher or lower BMIs as well. I associate smoking with being unhealthy, which can mean a very low BMI or a very high BMI. This visualization has been informative in the sense that you can’t judge BMI based off a person’s age.

As I’ve said previously, I chose to make scatterplots to represent my data. I tried making a bar chart, but it didn’t make much sense and wouldn’t have shown the correlation as nicely as a scatterplot would. With over a thousand data points, it’s easier to see each point on a scatterplot. I listed age on the x-axis both times because it was the independent variable, while charges and BMI were both the dependent variables. I included a lot of tick marks to see correlation between the two variables on the graphs. More ticks allowed the data to be more spread out so the points weren’t all sitting on each other. I also included the function “mouseover” to see what each individual point represents on both graphs. The first graph will list the age, charges, and sex of the person. The second graph will list the age, BMI, and if the person smokes or not. The “mouseover” display will appear on the bottom of the first graph for both graphs. You must roll your mouse over a datapoint, and the information will be displayed based on what graph the point belongs to. The blue and pink represent the genders in the first graph, male and female. The red and green represent whether a person smokes or not, with the red meaning the person smokes and the green meaning the person doesn’t smoke. The red is associated with “bad”, and the green is associated with “good”. There is one graph on the left and one graph on the right. I didn’t have a preference on where they were positioned on the page, so I placed them next to each other. I put the header as “insurance visualizations” above the graphs. I couldn’t figure out how to add a legend to the graphs, so I don’t have one on either of them. I will explain what the colors mean in the video. I also had to do a lot of website searching and YouTube video watching to put these visualizations together, so I will include links to some of the sources I used to make the graphs. My video is the frame right below this as well (double click to play).







Links:

<https://d3-graph-gallery.com/graph/scatter_grouped.html>

<https://www.youtube.com/watch?v=5wpHXf8qqGA&t=178s>

<https://bl.ocks.org/d3noob/5987480>

<https://d3-graph-gallery.com/graph/scatter_basic.html>