Hannah Dunn, Lucas Reisch, Mary Regan
INFO/CS 3300 Project 1 Report

Our data:

For our data we chose to use a set of data from https://www.kaggle.com that contained all of the tweets from the first GOP debate. Included with the standard properties of a tweet, each entry in the dataset contained what candidate the tweet was about, the subject matter, as well as whether or not the tweet was positive or negative.

Attached to each tweet was the location of the user. We decided to use this information to see how these tweets would reflect the outcome of an actual primary. In order to sort through real locations from fake locations, we ran a python script that removed any entry that did not have a state name or abbreviation in the location. While this wasn't perfect, it removed all fake locations and gave us a much better view of the dataset.

While all this information was interesting we needed to simplify our data into something we could show. We decided we needed all of the numbers for positive, negative, and neutral tweets in each state that has had a primary election up to this point. After reformatting our data to contain the name of the state where the primary was held, the candidate we are looking at, and the total positive, negative, and neutral tweets about all of them. Using this we were able to begin our visualization.

Flag Visualization:

The underlying data for the flag visualization is very simple; it is an associative array with each candidate and the number of delegates they won. The intent of the flag visualization is to see the portion of the delegates that each candidate has achieved. In other words, a measure of their overall success. Each candidate's number of delegates won was divided by the total number of delegates (within the states that have already participated in the primary. We then created a linear scale function to apply this decimal to the svg flag. The range was set to be from 0 to .6. This upper value is larger than even the largest candidates value (Trump has about .45 of the proportion) but was used so that the text appended to the end of the bars could fit. The domain was simply the pixel width of the svg (0-600). We then used this data to transform the flag into a bar chart. Each candidate's relative proportion is represented by the opaque bar to the right of their name. The transparent area has no statistical significance, it was kept to maintain the shape of the flag.

Circular Visualization:

The intent of the circular sentiment visualizations is to show that sentiments by state about the GOP debates on Twitter do not necessarily correlate with the outcomes of the state primaries. The data used for the visualization of sentiments and wins by state for each candidate is made up of two arrays. The first array contains the percentages of positive, neutral, and

negative sentiments in order by state. The second array contains each state, the winner of the primary in that state, and the abbreviation of each state. Each visualization was made using two overlapping pie charts. A transformation was used so that the origin of the pie chart would be located where we wanted the center of the graph to be. The base pie chart which is dedicated to state wins and losses, used the data from the second array, divided the circle into even slices based on the number of states, and maps a light grey to slices that represent wins and a dark grey to slices that represent losses. No other scales were needed for the base graph. The overlapping pie chart represents the sentiment percentages. Again, each slice is given an equal area, but this time a scale is used to vary the outer radius of each slice. This scale uses a domain of 0 to 1, to represent the full range of possible percentages. The range is from 90 to 180. The value 90 was an arbitrary value, chosen so that bars would be visible behind the caricatures of each candidate. The value of 180 was chosen because the outer radius of the base chart is 200, and this allowed for a buffer for the state name. To reduce noise, it there was 0% of any sentiment, we set its fill to none in order to declutter our visualization. Additionally, the color was varied between red, white, and blue in order to represent positive, neutral, and negative sentiments respectively. These colors were motivated by their patriotic nature, and also because red represents Republicans and would seem positive in this context, and blue is commonly used as a sad color.

What story are we trying to tell?

With our visualizations of our data, we are trying to show that Twitter sentiments about debates do not necessarily correlate to primary outcomes. The media often reports on who won each debate, often based on reactions from social media. For example, after one of the first debates, many media outlets reported that Carly Fiorina was winning based on Twitter, but now she is no longer in the race. For this reason, we expected some discrepancy between Twitter sentiments and primary outcomes, but it was surprising to realize that it is incredibly hard to predict primary outcomes based on Twitter. For example, Donald Trump had overwhelmingly negative sentiments in many of the states that he won, and he is still leading with percentage of delegates, and Marco Rubio received mostly positive sentiments but has only won one state.