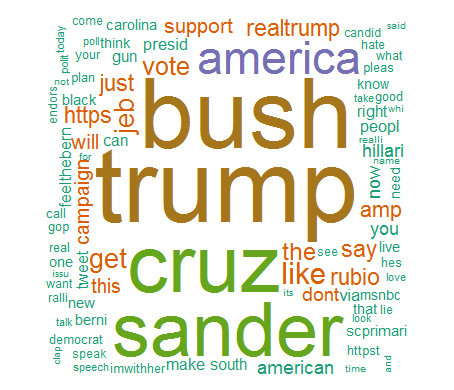
Wang, Xinran

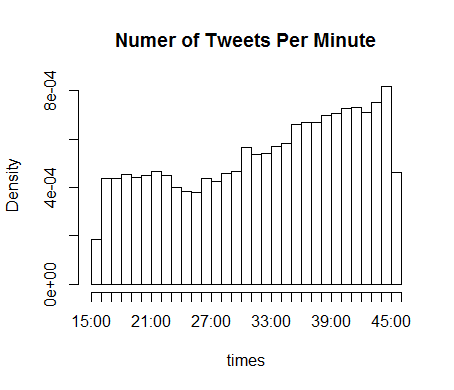
QMSS 4063 Assignment 1

Due 02/16/2016

Question 1

Wordcloud and histogram are the two methods used for visualizing the tweets collected in a thirty-minute time frame. Before executing with the visualization, different keywords for the same candidate are unified into one such as ‘Hillary Clinton’ and ‘Hillaryclinton’ are both represented by ‘Clinton’. The histogram shows the distribution of the number of tweets in the 30-minute period. Y-axis represents the counts in a scale of 10^4 to 8\*10^4 and x-axis represents the time from 4:15 to 4:45 when the tweets were collected. There is no specific pattern and the tweets are almost uniformly distributed. A slight increasing tendency appears from 4:24 to 4:44. With the wordcloud, we see that ‘Bush’ and ‘Trump’ are the most frequently mentioned words. ‘Sander’ and ‘Cruz’ come next. ‘America’ is also of relatively high frequency. The different colors of the wordcloud allow identification of the different categories that the words belong to. I’m surprised to see that ‘Clinton’ does not appear as a major word which means that it is less mentioned in the thirty minute tweets that I collected.



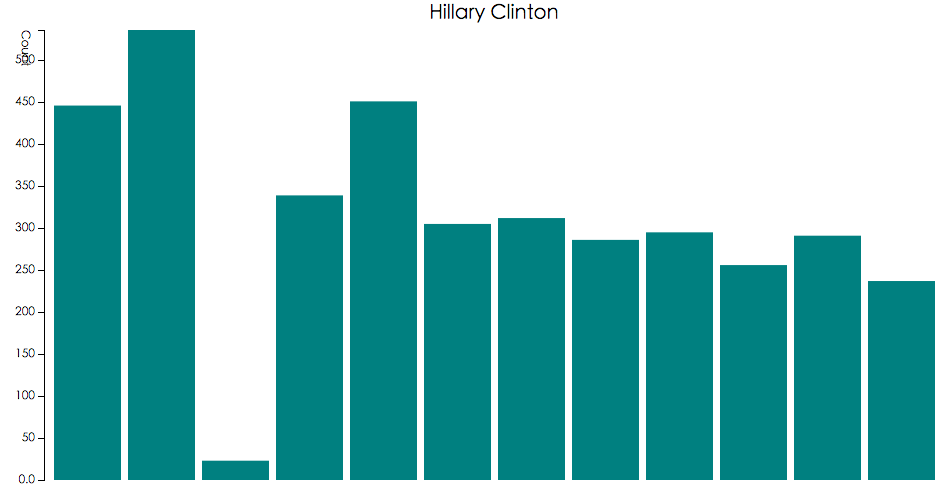


Question 2:

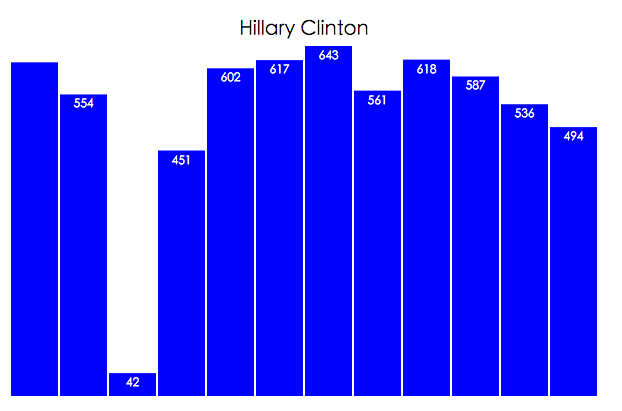
1) I parsed the tweets for four different candidates: Hillary Clinton, Bernie Sanders, Jeb Bush and Marco Rubio and then combined the similar keywords so that “Hillary Clinton” and “Hillaryclinton” are all captured as “Clinton”, similar for the other three candidates. Then I adopt a 12-hour time interval instead of 24, and the tweets are parsed into twelve categories. freq\_HC, freq\_BS, freq\_JB and freq\_MR are the parsed tweets in the data frame format and each has a dimension of 12 rows and 2 columns. The first column corresponds to time duration and the second column corresponds to count, i.e., the number of tweets containing this candidate within the given duration.

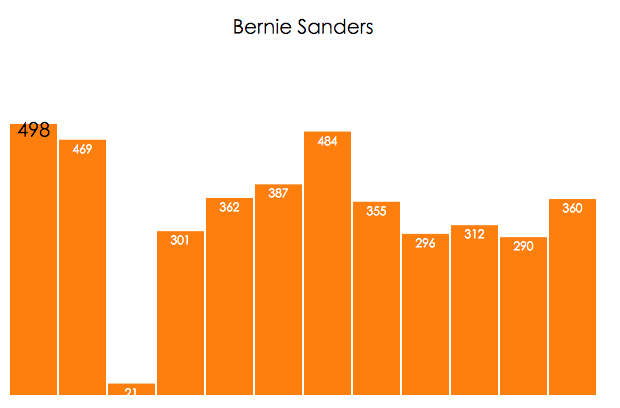
For the first part of the visualization, I made the frequency plot for each candidate. The x-axis corresponds to duration and y-axis corresponds to count. The scale of each graph adjusts to the size of the data of each candidate and each candidate has a different scale. From a rough investigation of the scale, Hillary Clinton and Bernie Sanders have the greatest scale with a minimum of 0 and maximum of 500. This means that these two candidates are mostly mentioned in tweets. The count scale of Jeb Bush is from 0 to 90, which implies that he is not as popular as the other candidates.

There is no obvious shape or node for the graphs. This makes sense because the count of tweets in each time duration is somewhat arbitrary and we should not identify any pattern in them. However, it is obvious that the third bar is significantly shorter than the other bars in all four graphs. This bar corresponds to the time period of 2-3 in the morning in which most people should be sleeping and refrain from tweeting activities. Right before that time period of time is a relatively long bar and indicate very active tweet activity. This fact corresponds to our common habits as most people play with their cell phone and check their twitter before going to bed. For Jeb Bush, Bernie Sanders and Marco Rubio, their graph patterns are more similar. The number of tweets starts off high in the beginning of the time and then drops to the lowest around 3. It starts to go up again until reaching a peak in the middle of the day. Afterwards, the number decreases gradually and has a tendency of increasing at the end. For Hillary Clinton, the pattern is somewhat different by reaching a high point in the fifth bar and decreasing all the way after that.

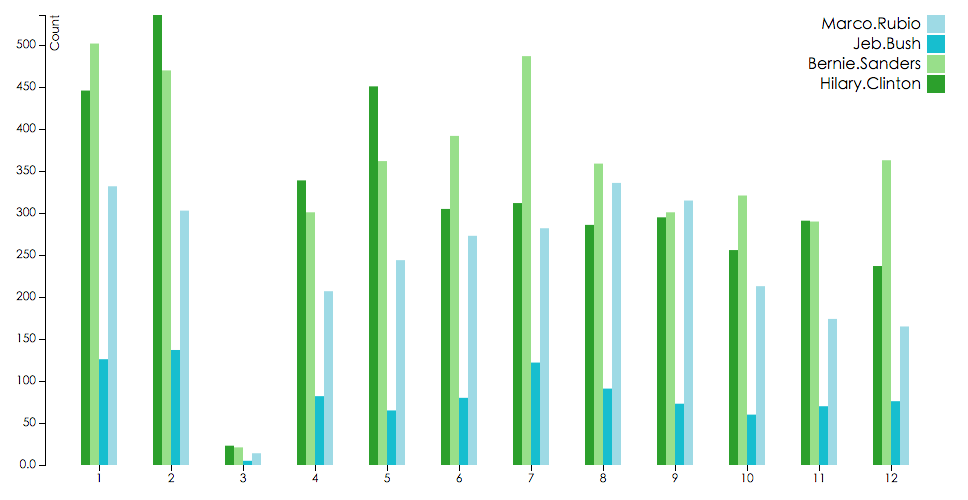


2) The interactive bar plot shows a transition between Hillary Clinton and Bernie Sanders. Starting with Clinton, the bar plot displays the same pattern as depicted in the previous question. The update is that now it has the exact number labeled on each bar. The maximum bar has a count of 643 and the minimum is 42. The range of 601 is noticeably large considering the total number of tweets I parsed. Except for the third bar, the shape of the graph is almost uniformly distributed which means that’s Hillary Clinton is actively discusses as a candidate throughout the day. Upon click and with the change of the color from blue to orange, the bar plot shows the distribution of Bernie Sanders. An obvious change is that the length of each bar is smaller so that the total number of tweets that Bernie Sanders receives is smaller than Hillary’s. There are two nodes at the first and seventh bar, correspond to midnight and noon at the time line. A possible explanation of this is that people are most free at these time of their days and they tend to use twitter. Hillary’s bar plot also demonstrates similar peaks.





3)



The third diagram is a group bar chart, using the total frequencies per candidate. The plot allows us to directly compare the count of each candidate during each given period. The overall graph displays a similar pattern to the individual plot – number of tweets is at the highest at the start and middle of the day, and decreases significantly at the third bar. I’m glad to observe this pattern since it allows me to conclude people’s tweeting habit even with multiple subsections. With the color diagram, it is obvious to see that Jeb Bush is least mentioned in tweets among the four candidates. The height of his bars is less than one half to the more popular candidates Clinton and Bernie’s. I mentioned the difference in pattern between Clinton and Bernie’s tweets graph in question 1, which can be verified with this graph. At the second bar, Clinton’s tweets exceeds Bernie’s whereas at the sixth and seventh bar, Bernie’s tweets are the maximum among all candidates. Total frequency chart display the information of different candidates in one graph using the same scale and is a good tool for cross comparison and pattern observation.