ADNAN HAJIZADA (ah3326)

Assignment 3, Data Processing & Visualization QMSS G4063

Navid Hassanpour, nh2519@columbia.edu

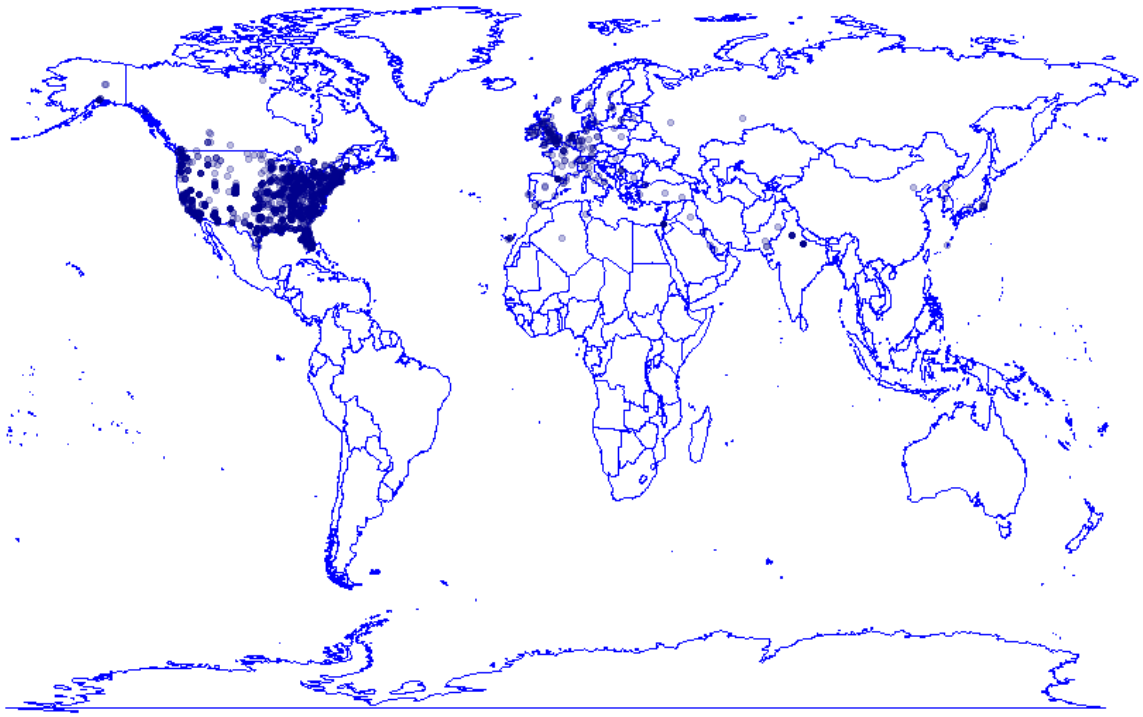
Due Thu. March 24, 2016

Counting Geolocated Tweets in U.S. States, (5×2 visualizations, 1000 words, optional link to online app):

The json files were too big for my laptop to handle, so I did all the parsing and filtering on the lab computers. So most of those code is not in my R file. After filtering the data down to geo located tweets, and separating them by candidate I saved the data frames into a *.Rdata files and transferred them to my laptop through load function.

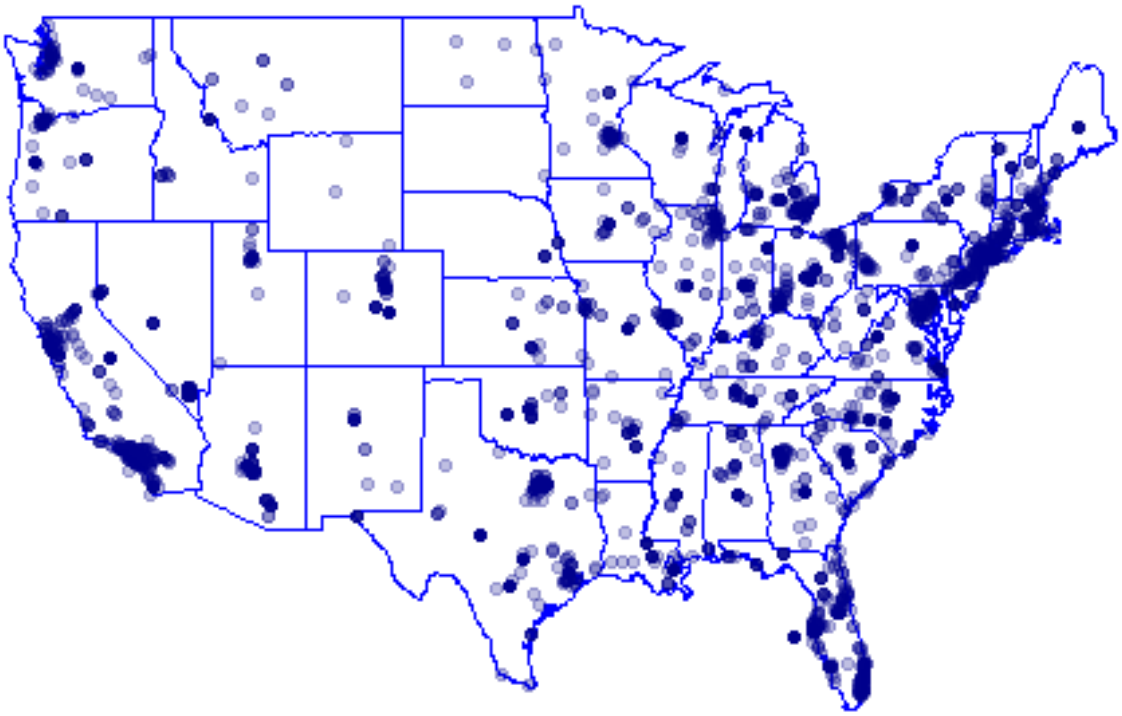
- Generate five maps showing the origins of tweets on each of the candidates. Do most of them come from inside the U.S.?

- ***Hillary Clinton.*** Most of the tweets clearly come from the USA. However it is necessary to note that UK has a considerable amount of tweets also. Continental Europe trails behind UK.

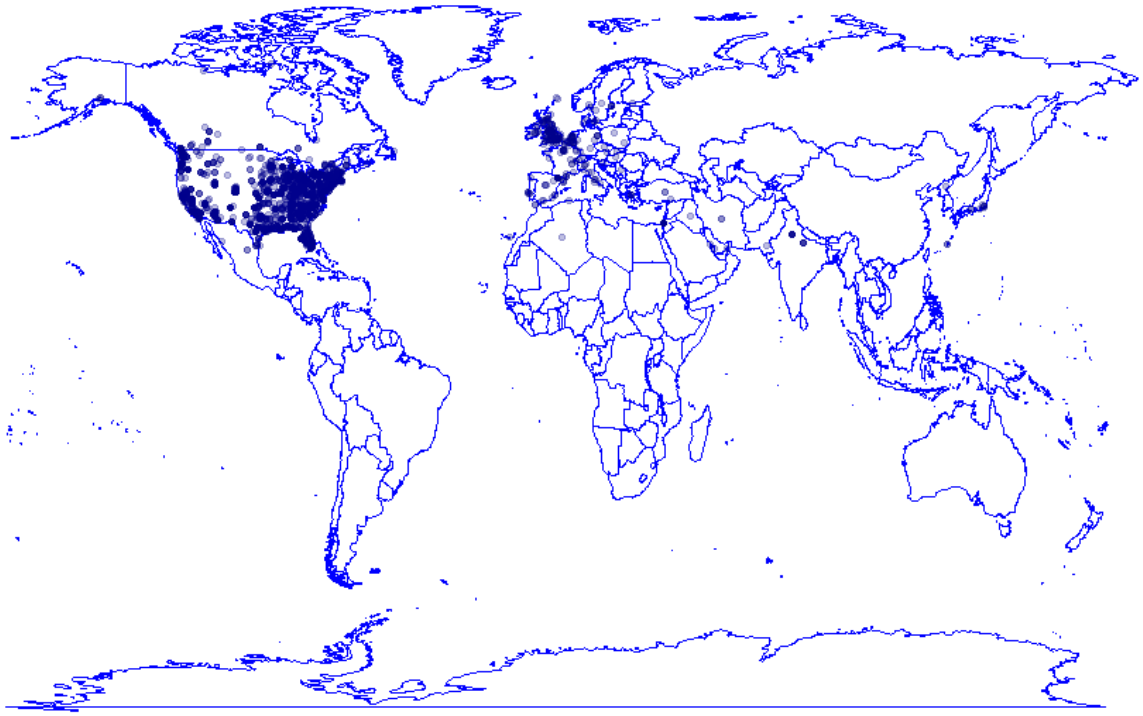


If we look at the Hillary Plot for the United States. The Democratic primaries that fall within our json files time frames are:

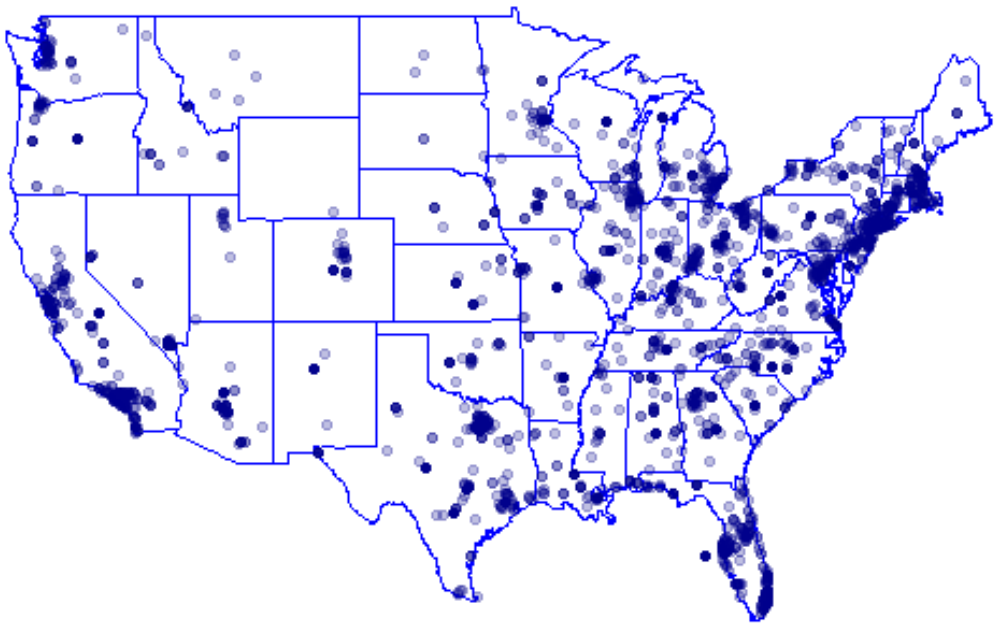
New Hampshire, Nevada, South Carolina, Alabama, Arkansas, Colorado, Georgia, Massachusetts, Minnesota, Oklahoma, Tennessee, Texas, Vermont, Virginia, Kansas, Louisiana, Nebraska, Maine, Michigan, Mississippi, Florida, Illinois, Missouri, North Carolina, South Carolina, Ohio, Arizona, Idaho, Utah. If we look at below map, there is very little tweets from South and North Dakota and Montana. These are the states where primaries did not take place. So that can explain the low activity of tweets. At the same time we can hypothesize that these three states have low twitter usage to begin with. So it is hard to draw conclusions. California has a lot of tweets, but it did not have the primaries yet. Considering the fact that less than 2% of tweets were geo tagged, this data is not representative of the whole situation.



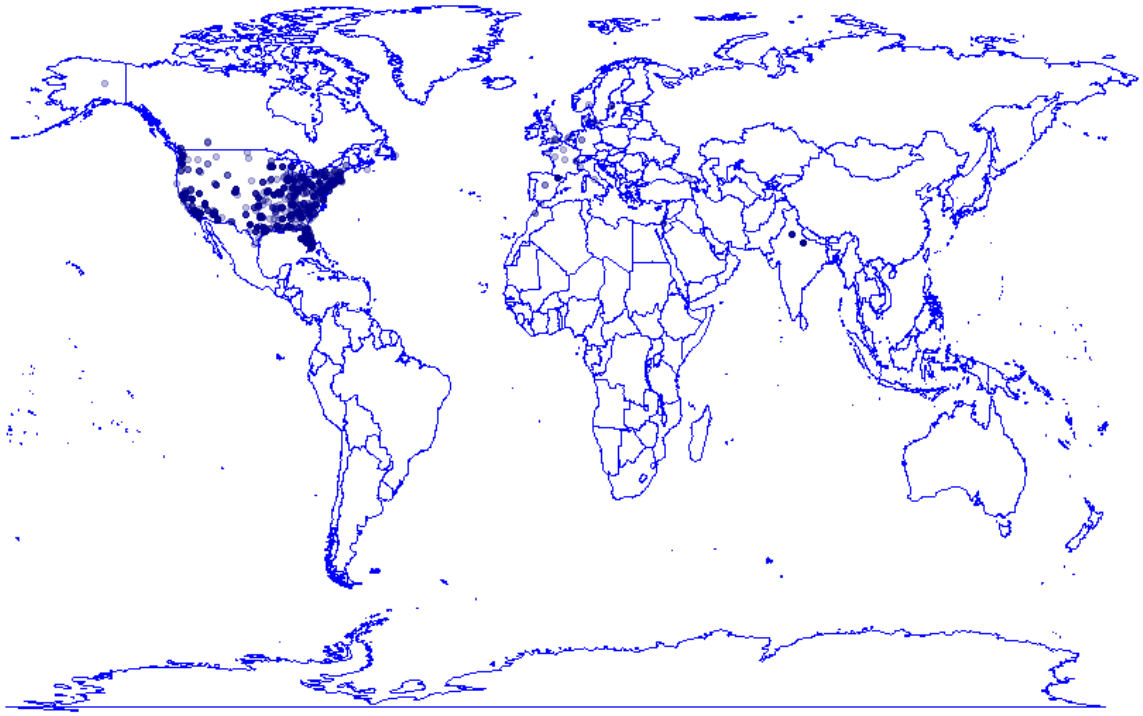
- ***Donald Trump.*** Similar situation



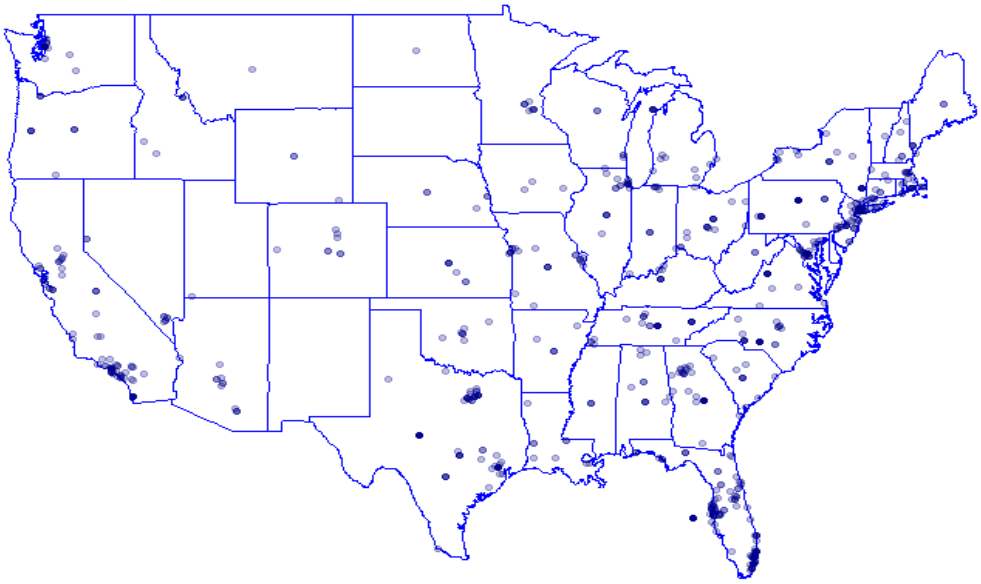
Donald Trump in America Plot. The coasts are more active than the middle of the country.



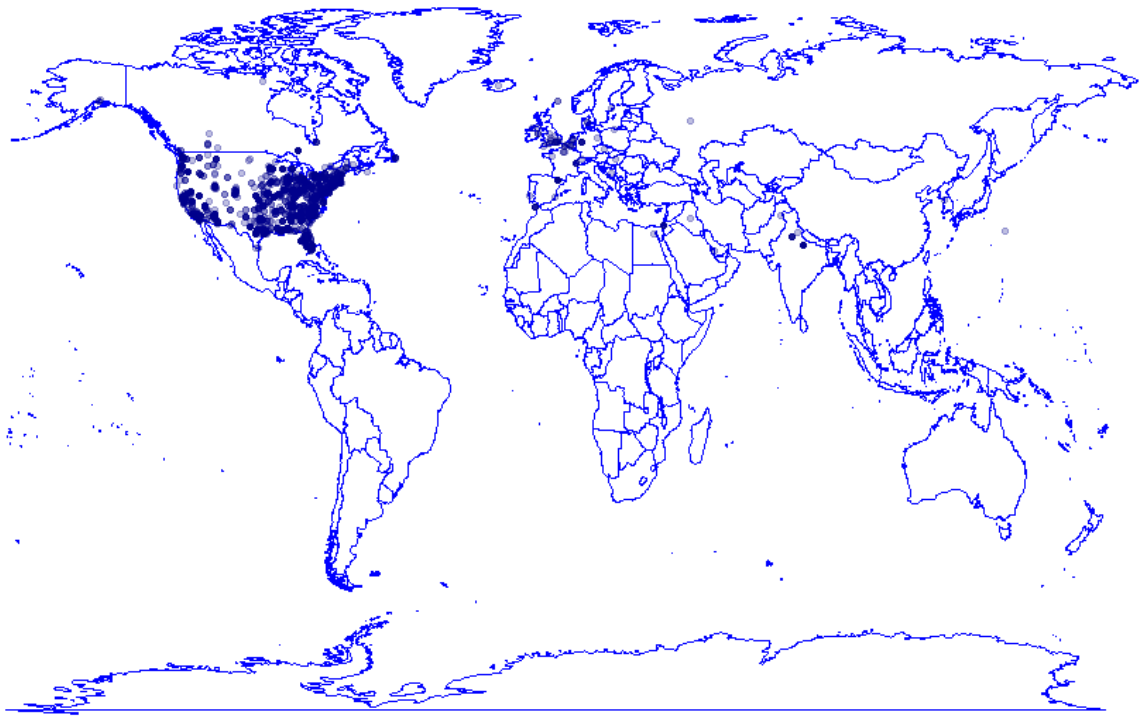
- ***Marco Rubio.*** Is tweeted about mainly from inside of America



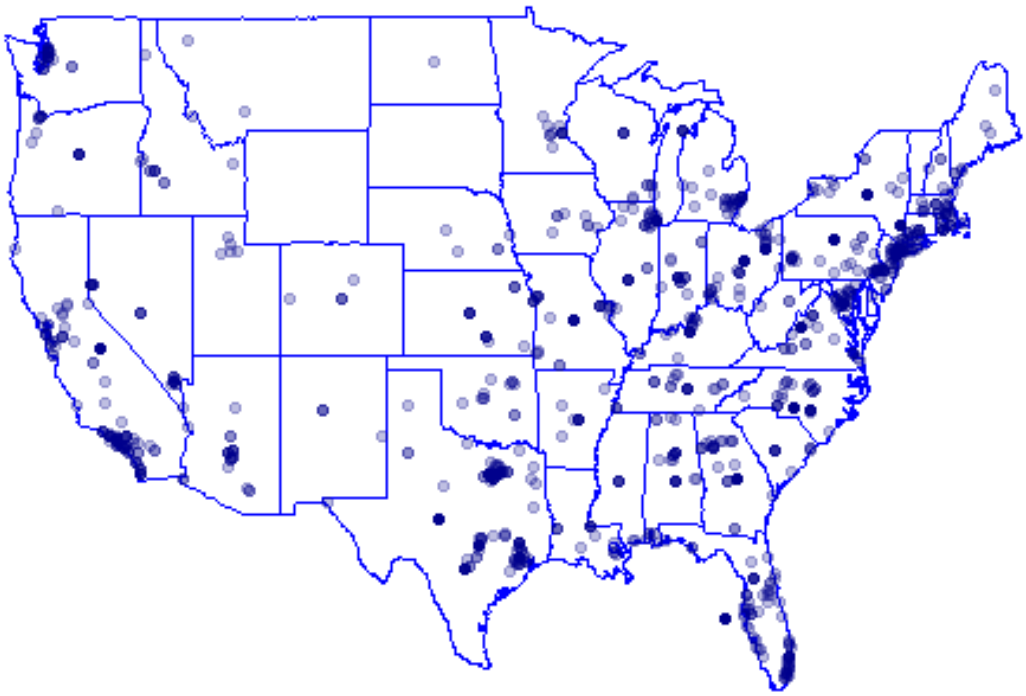
Marco Rubio in America: We can clearly see that Rubio is the point of discussion in his home state of Florida.



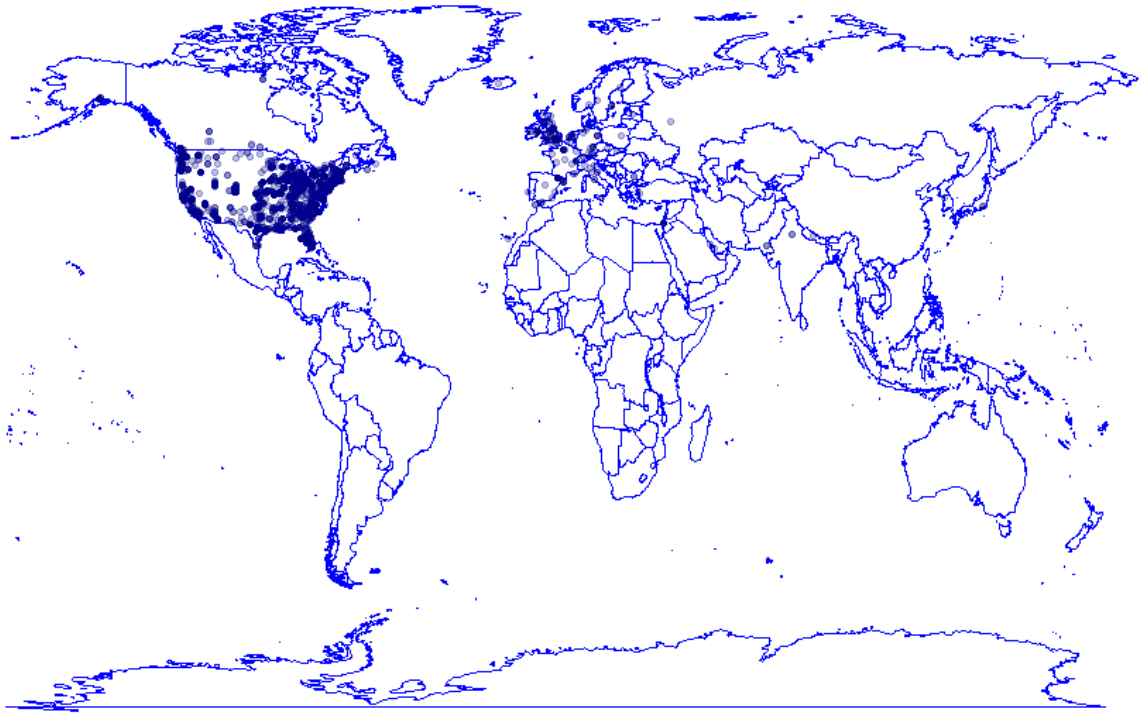
- ***Ted Cruz*** is mainly tweeted about from within United States. With a little representation in Europe.



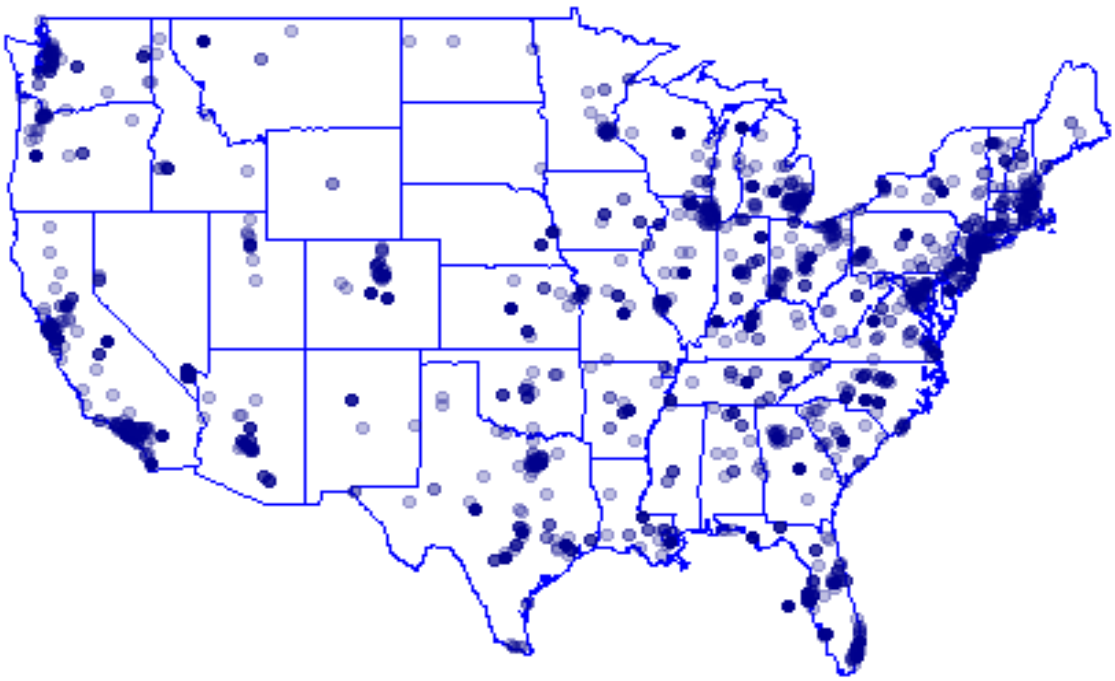
Ted Cruz in America Plot. Shows a lot of Activity in his home state of Texas.



- ***Bernie Sanders.*** It mainly tweeted about from the USA and UK.



Bernie Sanders in America plot. Liberal New York and California both talk about this candidate a lot in their tweets.

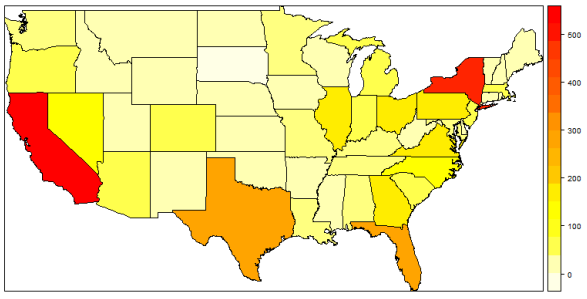
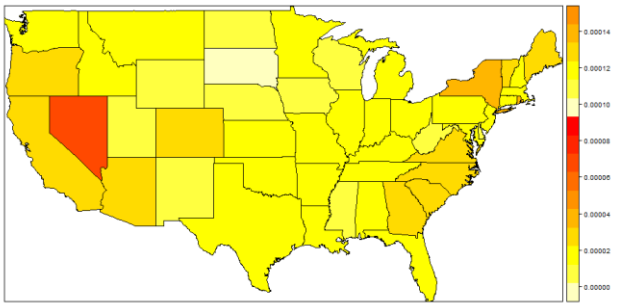


- Count the number of geolocated tweets from each U.S. state on each of the five candidates. Based on the counts produce five U.S. maps, one for each candidate, with states color-coded for showing the proportion of tweets coming from each state:

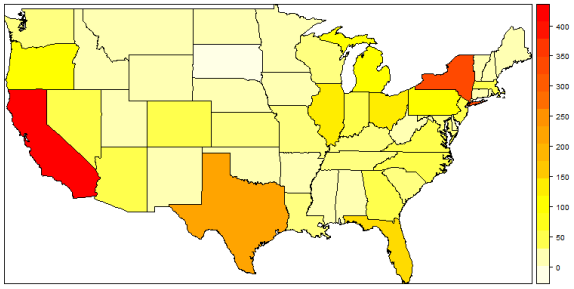
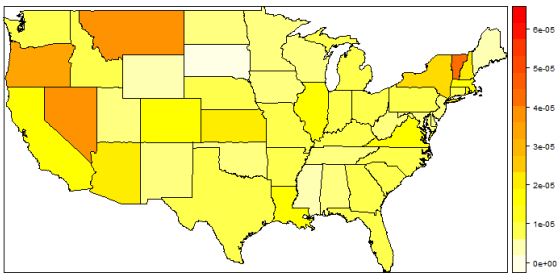
Heat Maps:

I created a heat map with the red being the highest number of tweets for a given candidate in the US and the lowest number of states are shown in light yellow and white. After looking at my plots I realized that the simple tweet count just highlights the states with bigger population and it is necessary to adjust for the number of voters in each state to have a better picture of tweet activity. For each state I divide the number of tweets by the voting population (2008 census data that I feed into R through a CSV file that I created) to get an estimate of "tweets per capita" which gave a better sense of the activity from each state.

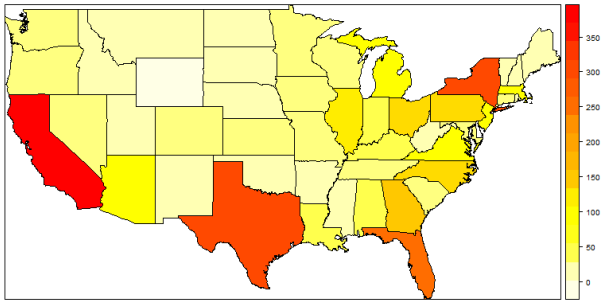
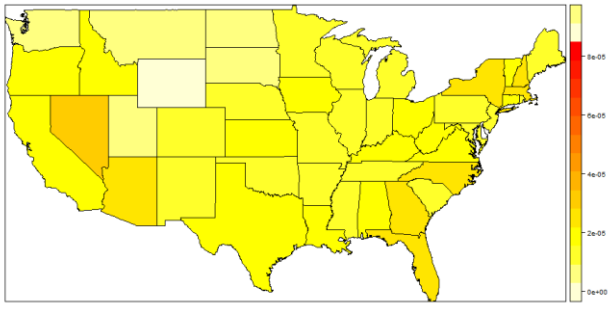
Hillary Clinton. Note that Hillary's home state of New York (she has been a senator from there) tweeted the most about her, even though the primaries did not happen here yet. This might show that the high level of tweets are not necessarily connected to primaries but more to the interest of populace towards specific candidate. At the same time we have to realize that this plot is not adjusted to population. So the plot on the right is the amount of tweets per capita in each state. Looks like primary state Nevada has the more tweets per capita. Note that Virginia, North Carolina, South Carolina and Georgia are more active than other states. It is necessary to note that the best wat to go about this would be to plot tweets by day and check whether on a given day of primaries given state produced more tweets.

Hillary Clinton	
Total count of tweets	Tweets per capita (voter Population)
	

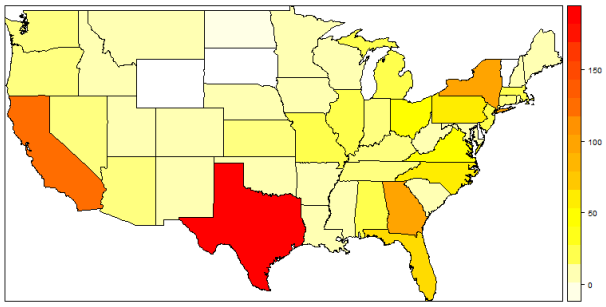
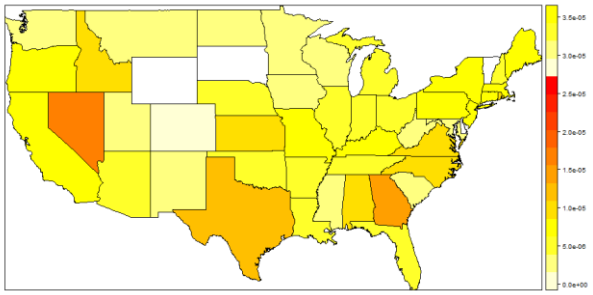
Bernie Sanders. Note that Bernie's birth state of New York has the most tweets about him. Also liberal California tweets more about Bernie. At the same time we have to realize that this plot is not adjusted to population. So the plot on the right is the amount of tweets per capita in each state. Here we can see that Bernie Sanders' home state of Vermont has been tweeting about him a lot, also there was a primary there. The primary states of Nebraska and Nevada are also tweeting a lot.

Bernie Sanders	
Total count of tweets	Tweets per capita (voter Population)
	

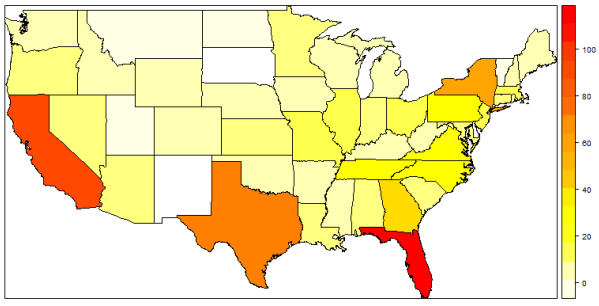
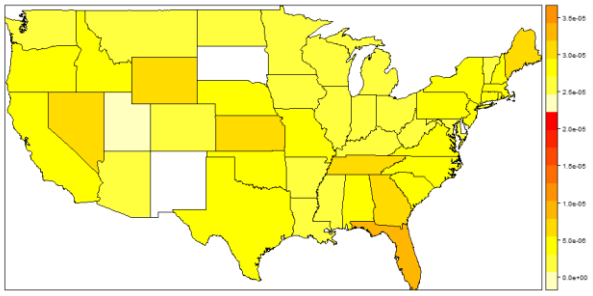
Donald Trump. With Donald Trump. A lot of highly populated states are tweeting about him. He is a point of discussion. If we look at tweets per capita, we can clearly see that Nevada, Florida, North Carolina, Georgia and other primary states have more tweets.

Donald Trump	
Total count of tweets	Tweets per capita (voter Population)
	

Ted Cruz. Texas, which is a home state of Ted Cruz tweets the most about him. If we look at the tweets per capita we also see that Nevada and Georgia also have high amount of tweets due to primaries.

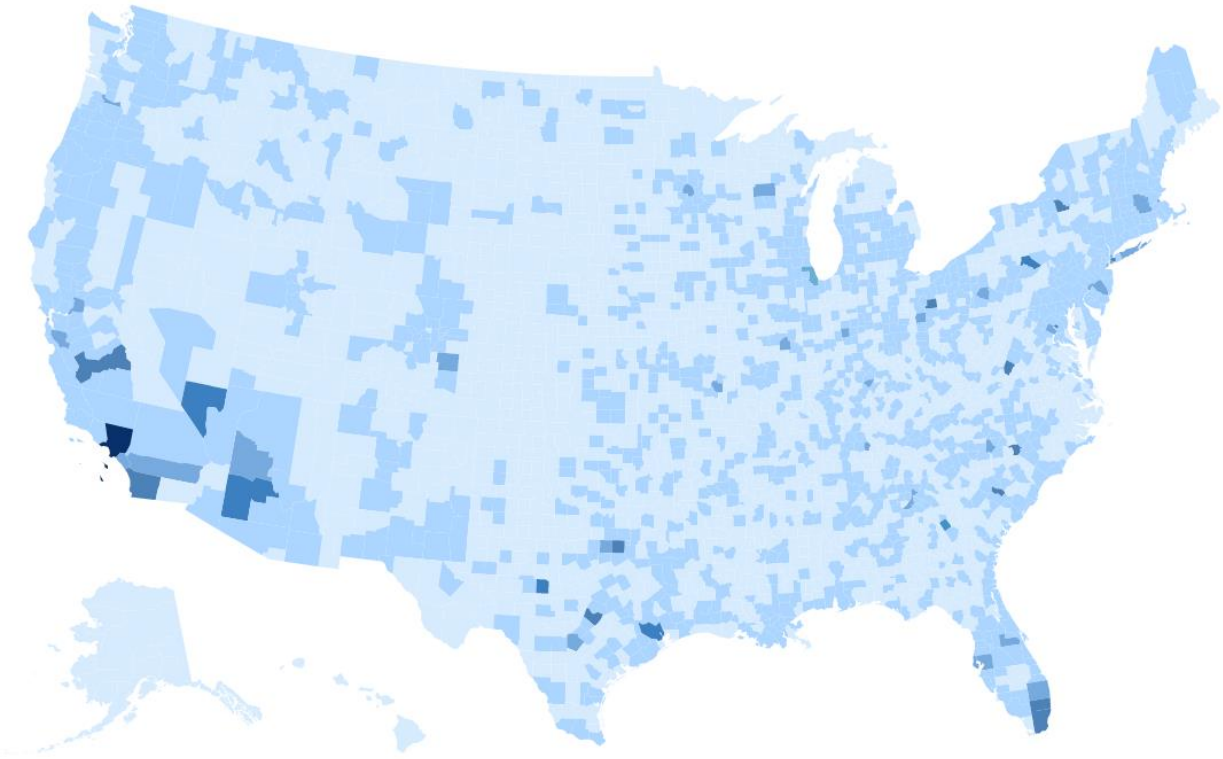
Ted Cruz	
Total count of tweets	Tweets per capita (voter Population)
	

Marco Rubio. Note that Marco Rubio's home state of Florida is tweeting about him the most. If we look at per capita tweets the Florida is confirmed. Also Georgia and Nevada tweeted more due to Primaries.

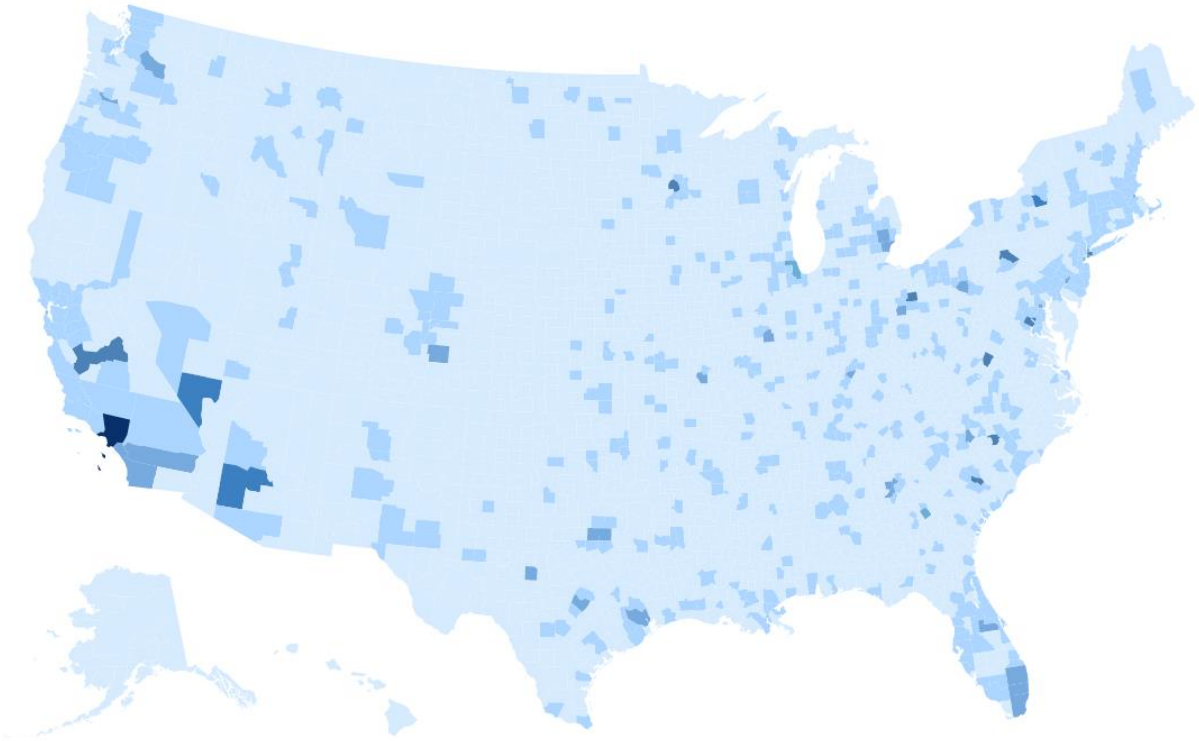
Marco Rubio	
Total count of tweets	Tweets per capita (voter Population)
	

Extra points: Use the D3 examples in the following two links (Link1 and Link2) to generate a static tweet production *choropleth* for each of the five candidates.

- This is the plot of all tweets in accordance to the counties. The darker the color, means more tweets came from that county.



- Similar Plot for the tweets about Hillary Clinton. Same counties are the most active across the board. Chances are this is due to population.



- Donald Trump Counties Plot