Assignment 3

**Part b)**

This section shows the geographical distributions of the tweets for various candidates in terms of counts or volumes. Every state is colored depending upon the number of tweets originating from the state within our dataset. The problem however, with such kind of visual analysis is that it draws our attention to highly populous states. We shall observe that the most populous states in the US including California, Texas and New York stand out in visuals corresponding to all the major candidates. Of course, we know that primaries in New York and California are yet to take place.

Bernie Sanders

The map below gives the geographical distribution of the tweets mentioning Bernie Sanders. From the 'frequency' (Freq) legend shown to the right, we can learn that the states colored with a lighter shade of blue correspond to a greater number tweets. At a quick glance, the highly populous Texas, New York and California (the lightest) stand out for obvious reasons. A higher population is most likely correlated with greater aggregate activity on social media. In the assignment, our major question is 'Whether the states in which the primaries were held show a significant increase in tweets as compared to other states'. However, the answer to the question is not clear but there is substantial evidence to the contrary. The states that appear relatively lighter in addition to the three mentioned before (such as Illinois, Michigan and Florida) are relatively populous states as opposed to ones that appear darker.


Geo - distribution of Bernie Sanders tweets
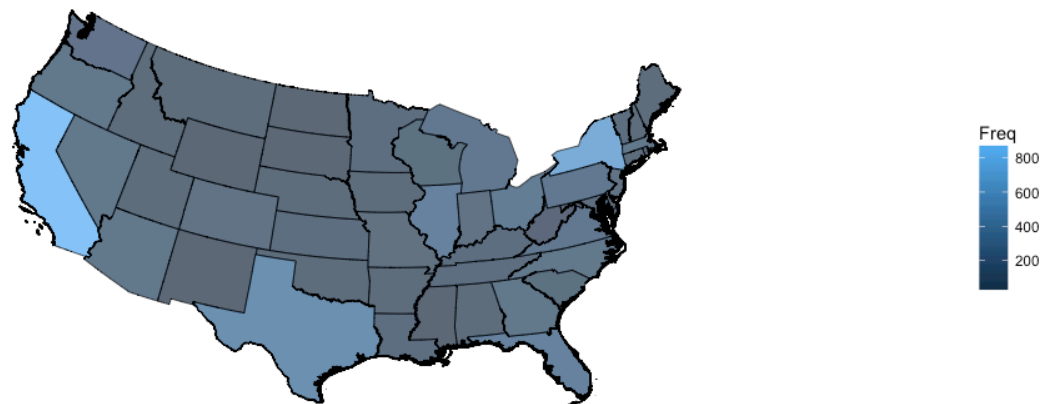
Hillary Clinton:

The map for Hillary Clinton also shows similar patterns in case of the most populous states. There are certain differences though. Florida in Clinton's case seems visibly lighter as compared to Sander's map which is not counter intuitive (Relative strength of Hillary's campaign in Florida). Moreover New York appears to be visibly lighter whereas Texas just appears to be marginally lighter compared to Sander's map. One could hypothesize and attribute the differences to New York being Hillary Clinton's home state and to the high margin victory of Clinton over Sanders in the Texas primaries.

Saad Khalid (sk3968)

Geo - distribution of Hillary Clinton tweets
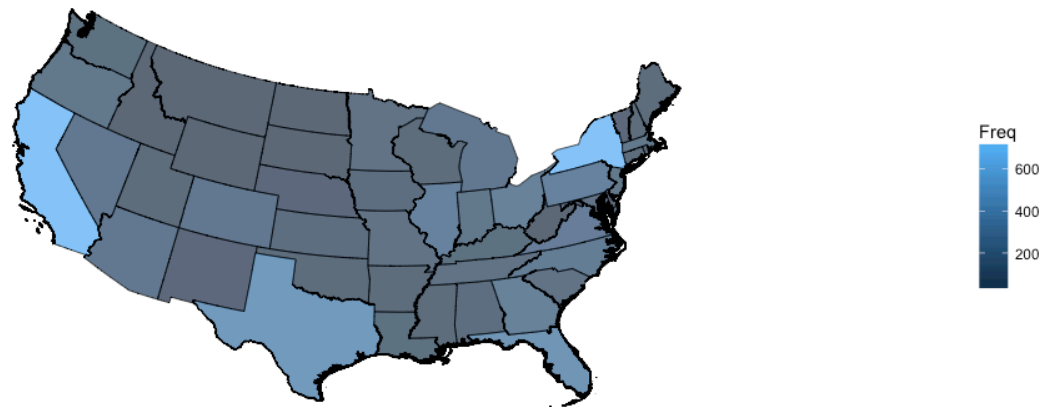
Donald Trump:

Trump's map shows similar patterns for New York, California and Texas. However, compared with the Democratic candidates' maps, Texas for Trump is colored much lighter. This may be explained by the fact that Texas is a conservative state and usually votes for Republicans. This leads to another interesting question; how can we relate this information to Trump's loss in the state to senator Ted Cruz. Could it be explained by 'bi-partisan politics' in the US?

Geo - distribution of Donald Trump tweets

Saad Khalid (sk3968)
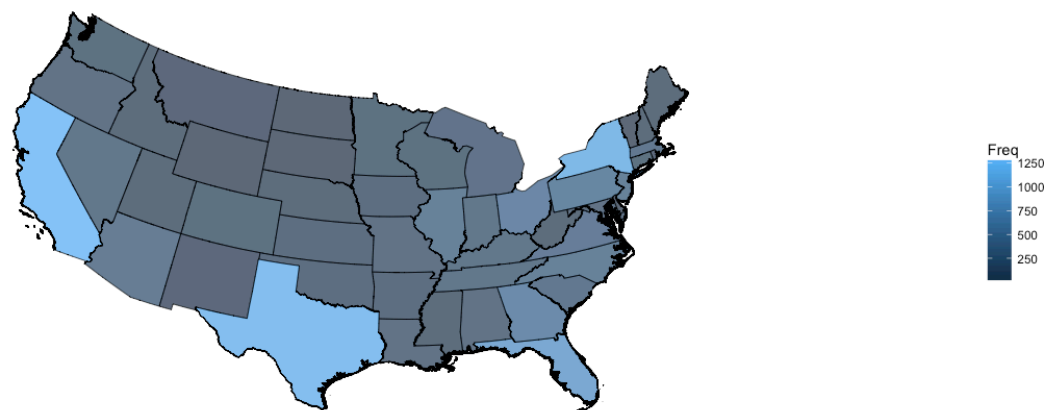
Ted Cruz:

Cruz appears to be the most popular in his home state of Texas (lightest of all candidates)[1] which is corroborated by his high margin win over Trump in the primary. Moreover, we can also observe that he is much less discussed in the more liberal populous states such as California and New York.

Geo - distribution of Ted Cruz tweets
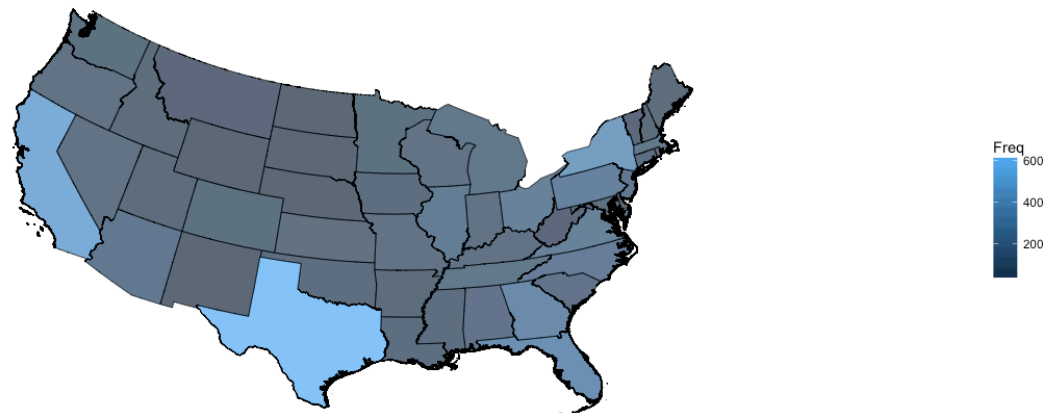
Marco Rubio:

Rubio is most popular in terms of tweet volume in his home state, Florida, California and Texas. Interestingly, Despite having more tweet mentions in his home state, Rubio lost to Trump in Florida

---

[1] However, we can't make definitive conclusions on comparisons between candidates based on the lightness of color as the scale of color varies with candidates (some candidates have more tweets as compared to others)

Geo - distribution of Marco Rubio tweets
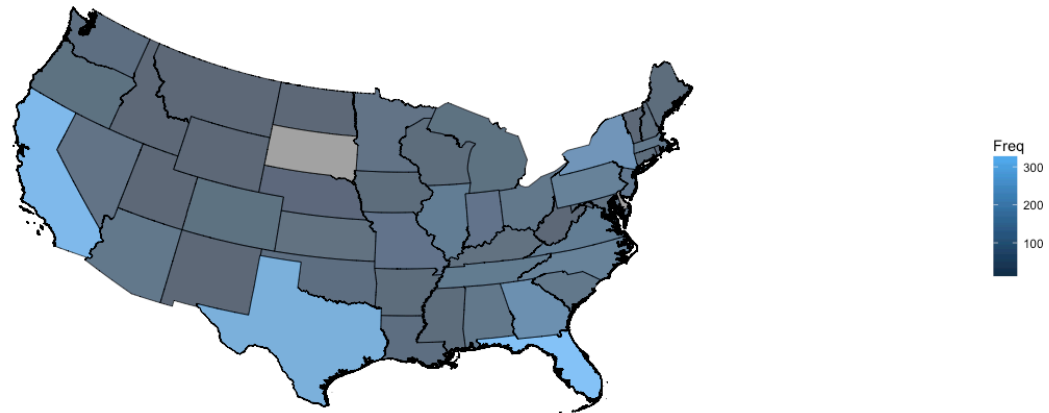
**Tweet Per Voter Population method:**
As discussed above, the varying population within states skews our visual analysis of 'popularity' of candidates. For that reason, I obtained the voter population data of US states (downloaded by my friend and classmate, Adnan Hajizada) and compute a metric called tweets per voter population, so that we do an 'apple apple' comparison. Surprisingly, for most of the states the metric remains stable for almost all candidates except for Nevada which is visibly lighter in all maps. I hypothesized that a possible explanation could be that Nevada is a low population state with most of the population living in urban areas (assuming urban population engage more in social media). Additionally, primaries were also held in Nevada in February.

Another state that stands out in the Bernie Sander's map is Vermont (the Senator's home state) which is much lighter indicating a significantly high tweet per voter population ratio. The explanation to this phenomenon seems to be straightforward; the senator has gained great political rapport and support from the local population during the time he served as mayor of Burlington and a senator for the state.
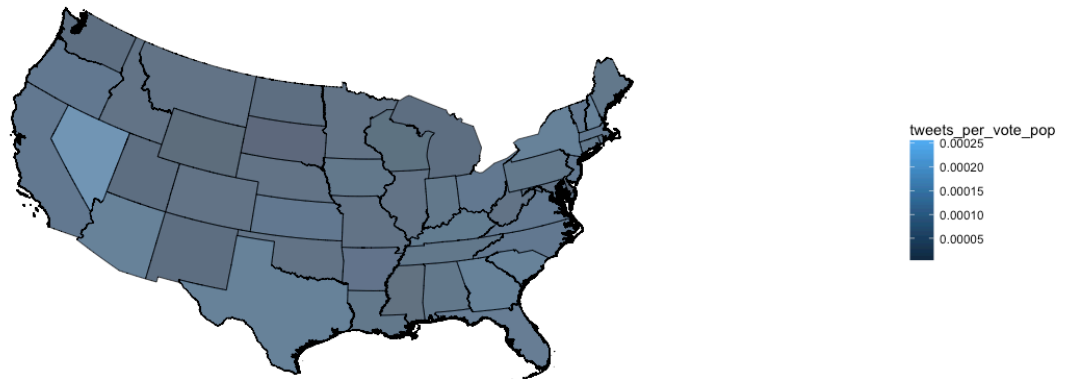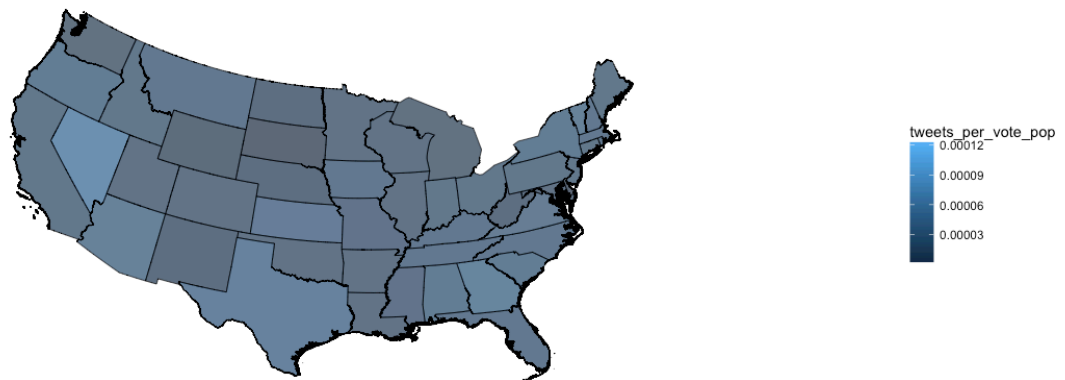
Geo - distribution of Bernie Sanders tweets

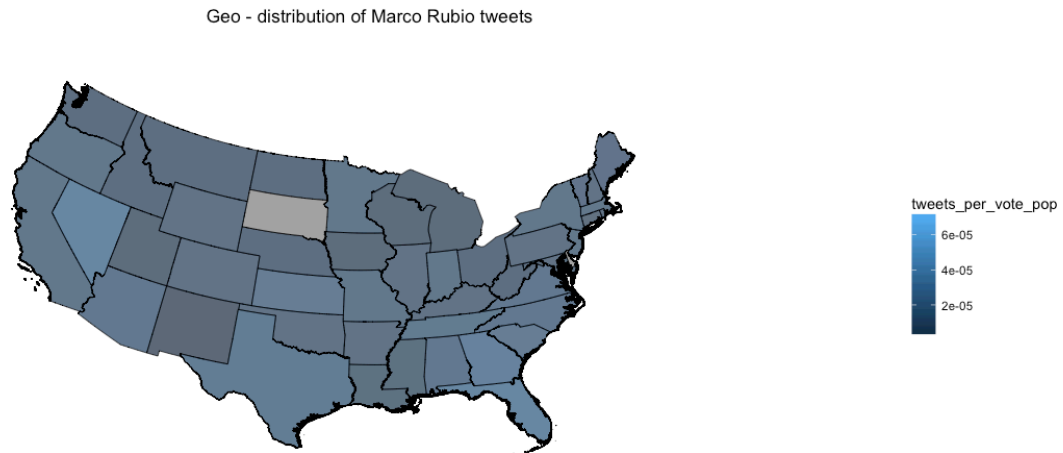

tweets_per_vote_pop

Geo - distribution of Hillary Clinton tweets

tweets_per_vote_pop
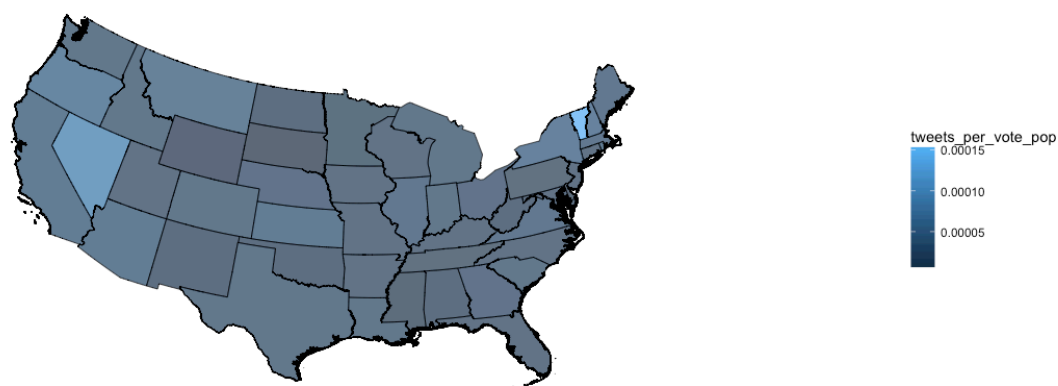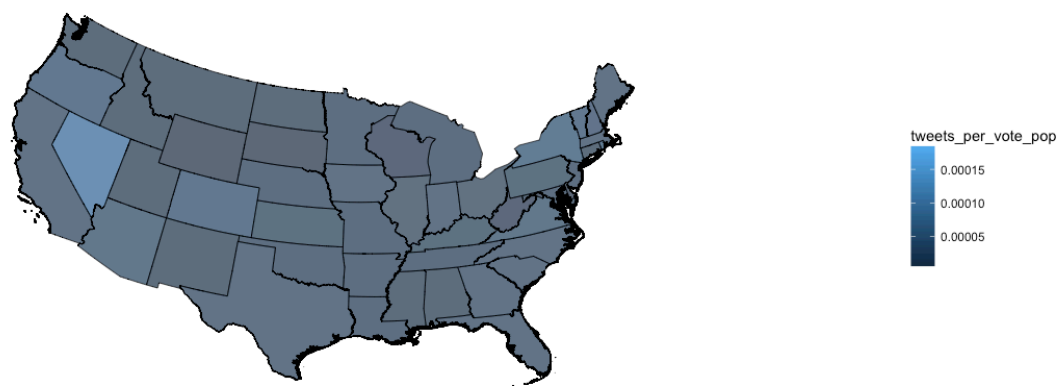
Geo - distribution of Donald Trump tweets

Geo - distribution of Ted Cruz tweets

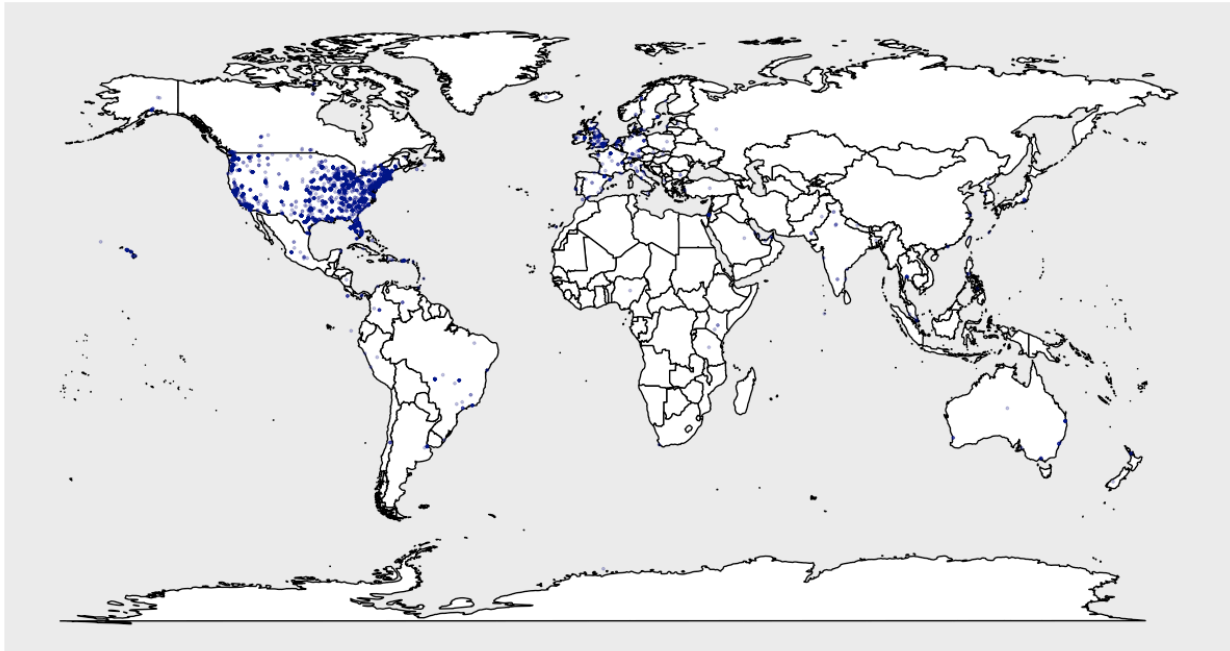Saad Khalid (sk3968)

Geo - distribution of Marco Rubio tweets

Limitations:
In my analysis, I excluded the states of Alaska and Hawaii. Although the exclusion of the states probably did not lead to a significant loss of information but there must be a way to map the non-contiguous states with the contiguous states which I failed to achieve. Moreover, during the analysis for Marco Rubio, the tweets for the state of South Dakota turned out to be 'NA' for some reason which I haven't yet looked into, This explains 'non colored' region for one state in the maps for Rubio.

Part a)
To answer the central question of this part, we do observe tweets from countries other than the US. The important thing to notice is that most of the points represent 'place_lat' and 'place_lon' within the dataset which correspond to the geo location of hometown rather than the actual location of the tweet. Important observations include that most of the tweets originate from the highly populated urban centers of the US. There is also a significant contribution from the UK and Northern/ western Europe which may be indicative of either 'political interest in the region' or more plausibly 'greater access to internet and social media'. We also see points here and there in Latin America that may be because of geographical proximity and possibly represents 'political interest in the region'. On a frivolous note, Donald Trump is probably the most talked about in Mexico (followed by Clinton).
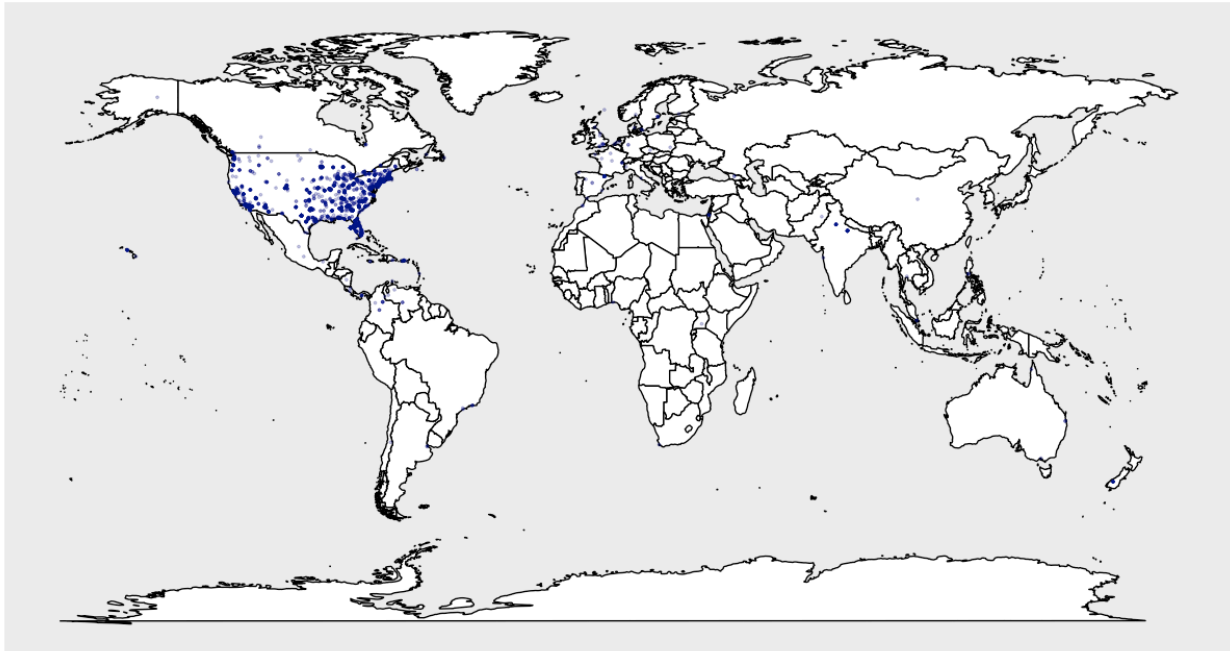
World distribution of Bernie Sanders tweets
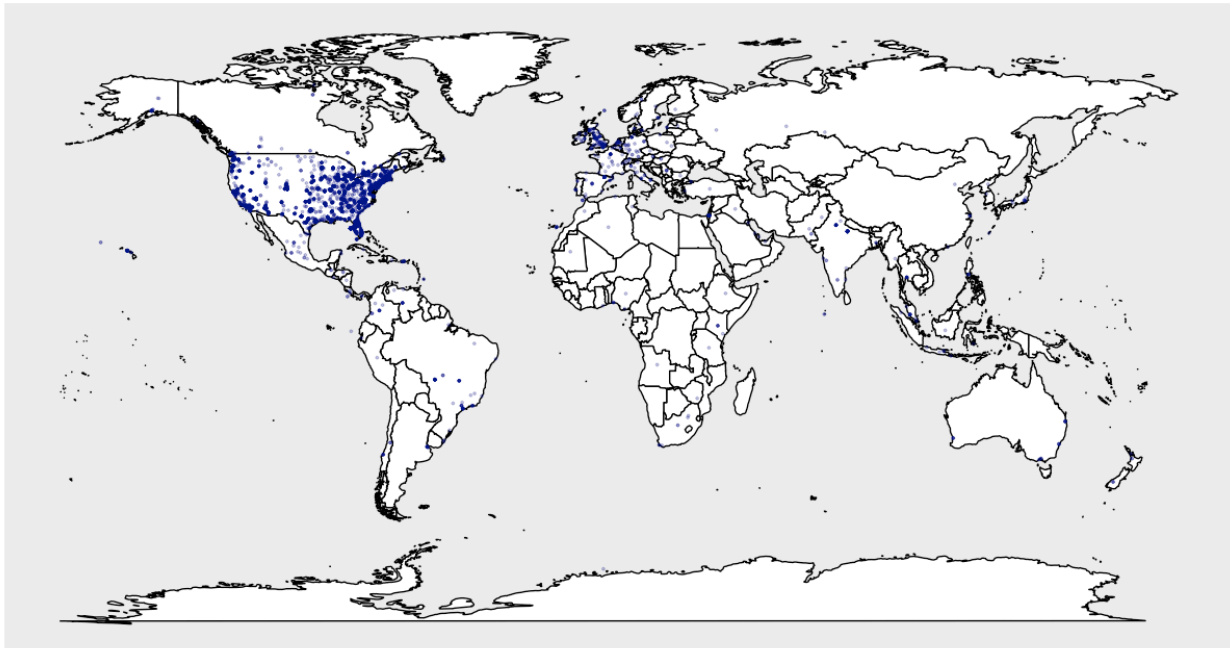

World distribution of Ted Cruz tweets
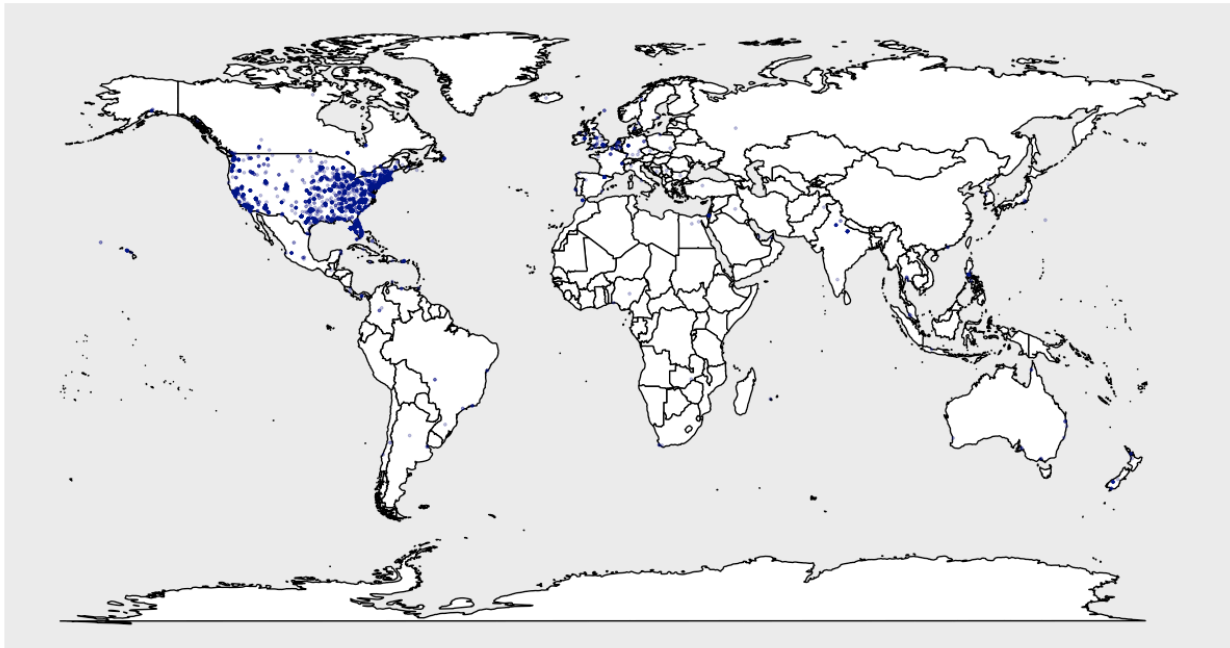
Saad Khalid (sk3968)

World distribution of Marco Rubio tweets

World distribution of Hillary Clinton tweets

World distribution of Donald Trump tweets