Analysis of User Interactions on Twitter

MSDS 622 Final Project

Sarah Melancon

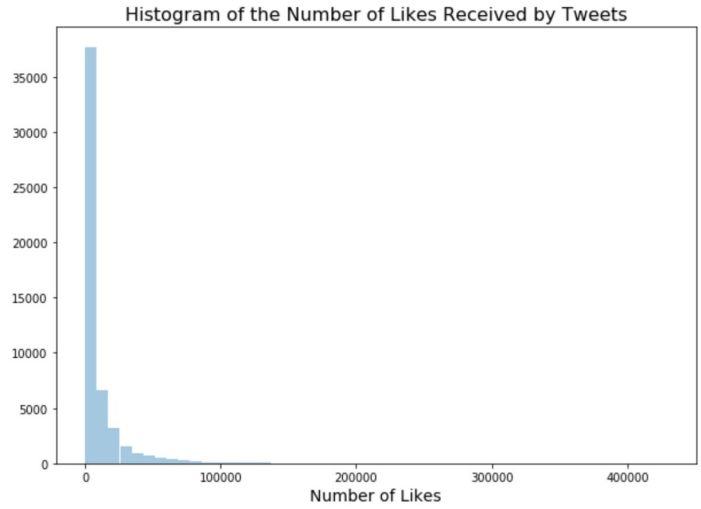
Summary of data: The data is taken from two tables. One is a graph dataset of retweets. There is one column for the user who retweeted, one column for the user that was retweeted, and one column indicating the number of times they retweeted.

The other dataset is a table of different tweets with information about them including the user, language, number of likes, number of retweets, and time of day.

Unfortunately, it is very difficult to obtain information about location for Twitter users because users are not required to disclose their location. For the sake of this project, the location data was obtained by mapping the user's language to country with the most people who speak that language. This is clearly a extremely naive approach. For example, all English speakers are mapped the United States, and all Spanish speakers are mapped to Mexico. However, it gives us some kind of location data to work with.

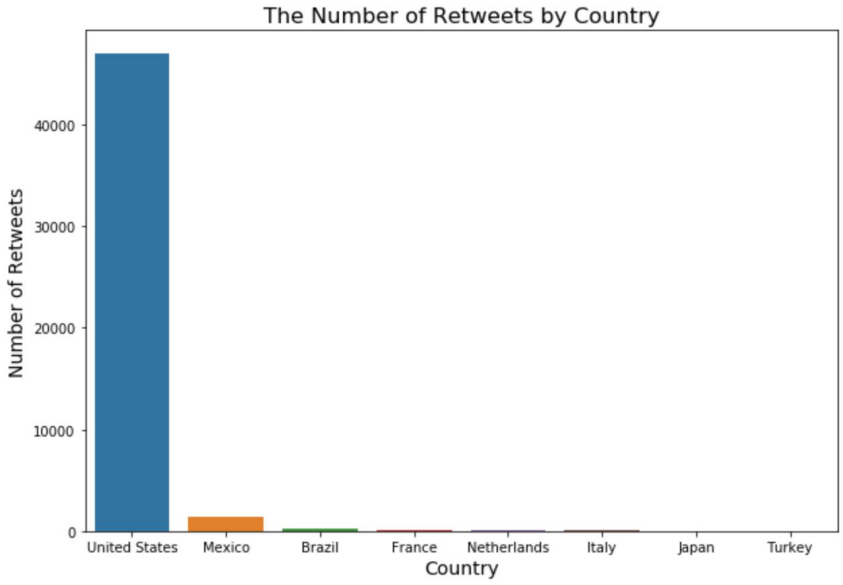
Histogram

Here I created a histogram of the number of likes received by tweets in the dataset. We can see that the vast majority of tweets received very few likes, while a few received as many as 400,000 likes.



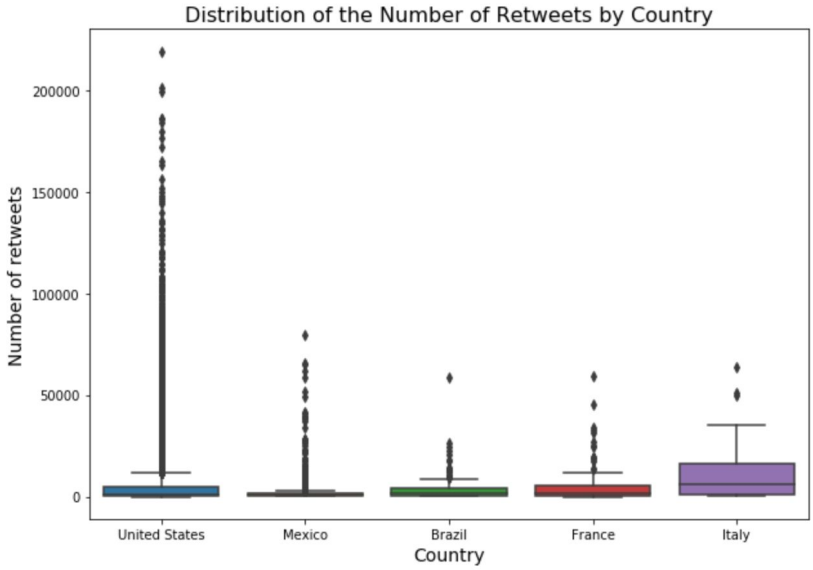
Barplot

I grouped the data by country and calculated the total number of retweets for each country. This is a bar plot of the 8 countries with the largest number of retweets. By far the most retweets were from the United States.



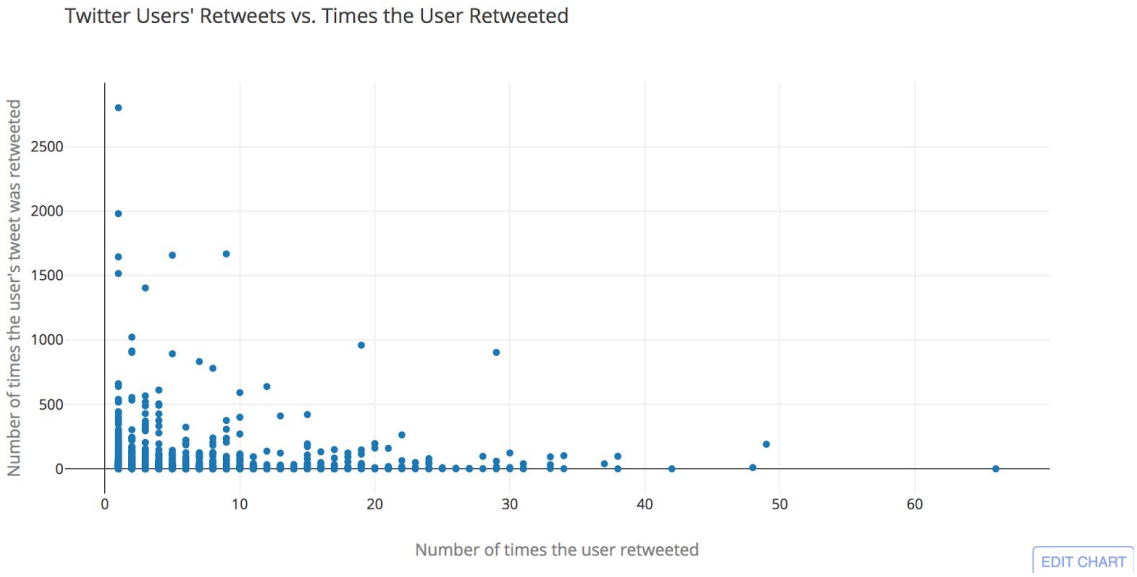
Boxplot

This is a box plot showing the distribution of the number of retweets a tweet receives in each country. I only showed the five countries with the most retweets. There is a much longer upper tail for the United States than for other countries, indicating that there are some Twitter users in the US getting a lot of retweets.



Scatterplot

I created a scatterplot of number of times a user was retweeted versus the number of times they retweeted someone else. Notice that users who retweeted a lot often did not receive as many retweets, and users who were retweeted frequently did not retweet others as often.

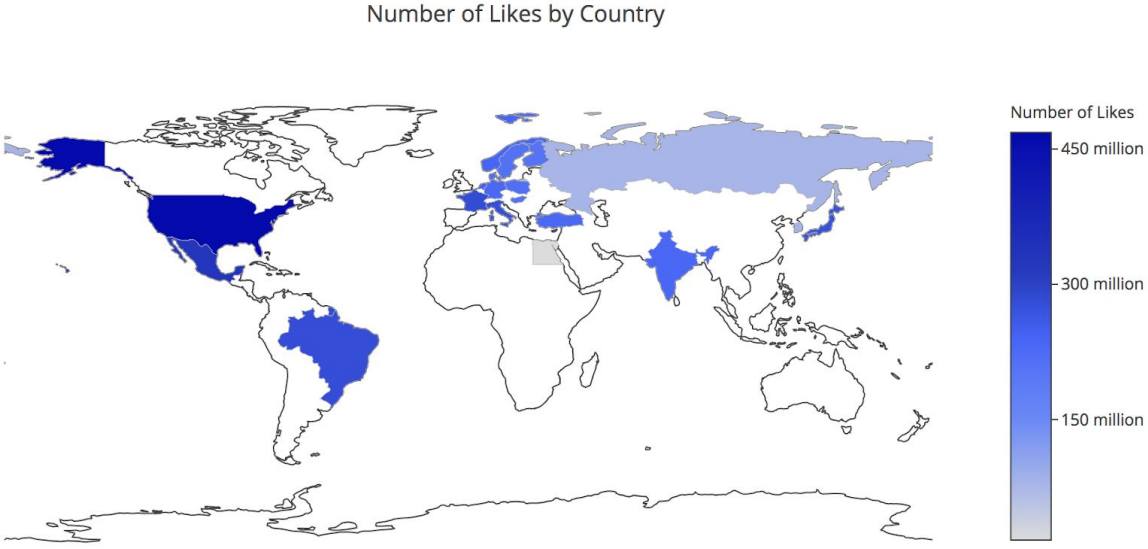


Bubble Map

Here is a bubble map of the total number of retweets by country. The size of the bubble corresponds to the number of times tweets originating in that country were retweeted. Tweets from the US were retweeted the most.

Chloropleth Map

Here is a chloropleth map of the total number of likes by country. The color of the country corresponds to the number of times tweets originating in that country were liked. Once again, tweets from the US received the most likes.

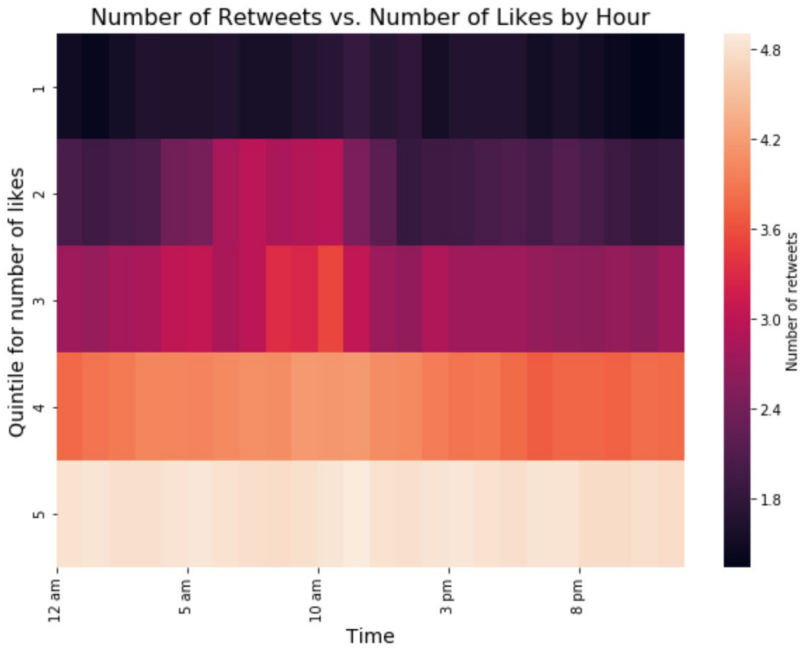


Connection Map

See storyline below.

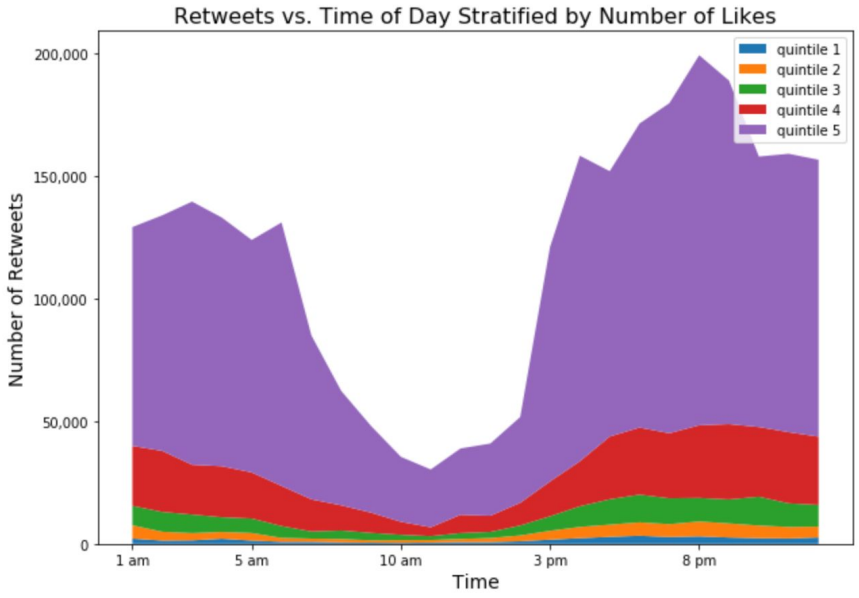
Heatmap

This is a visualization of the number of retweets received by tweets grouped by hour and the number of likes. Since the number of likes is not a categorical variable, I split the data into quintiles.



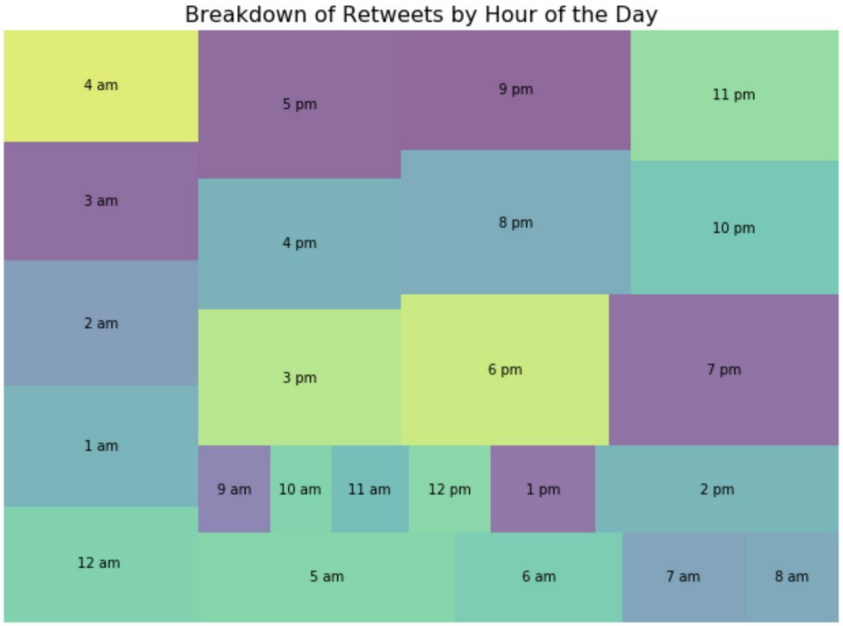
Stacked Area Graph

Here we see the number of retweets vs. the time the user tweeted. The visualization is stratified into quintiles by number of likes. Across all strata, there is a dip in number of retweets between 5am and 3pm.

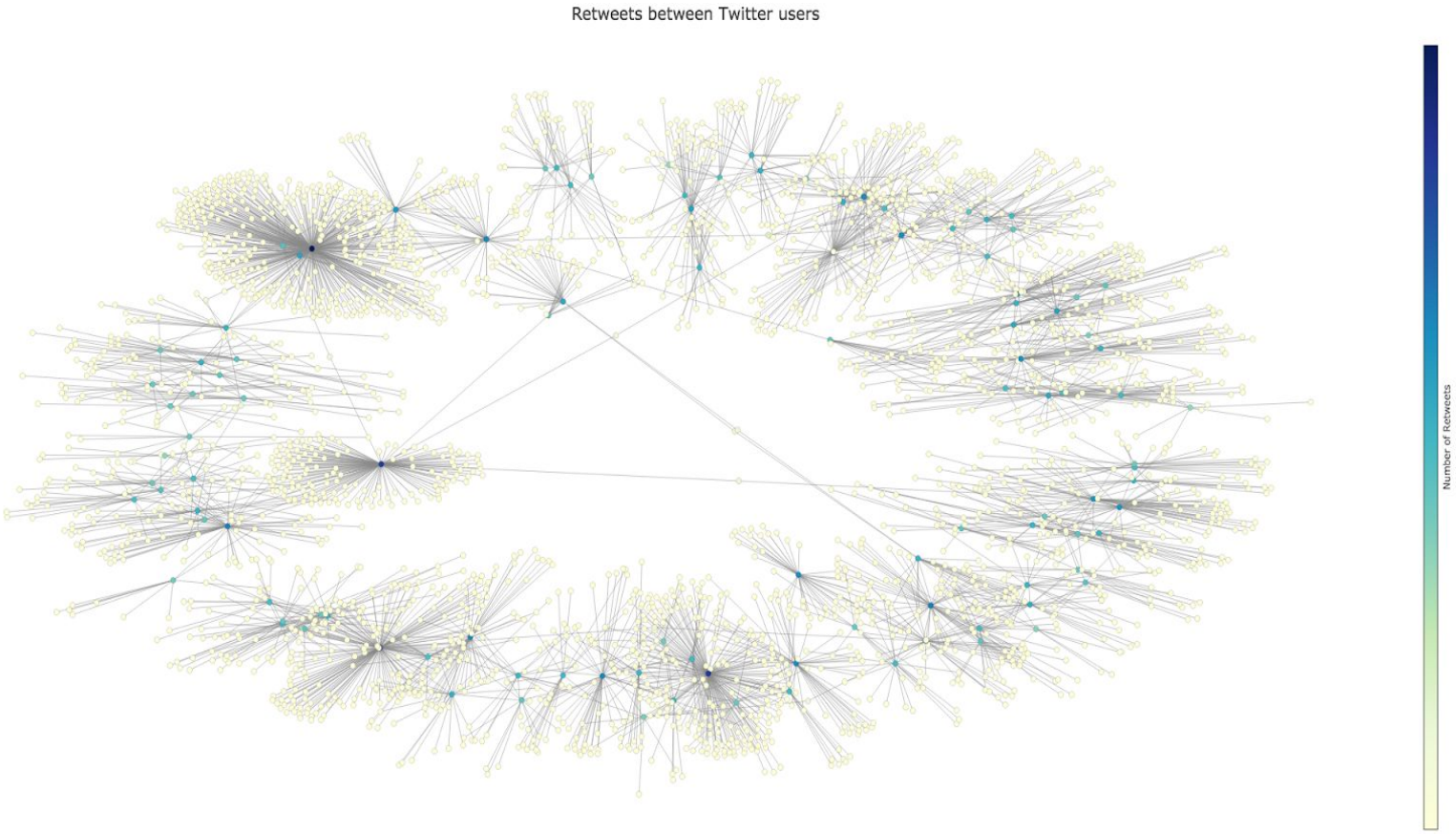


Treemap

This is a treemap breaking down retweets by hour of the day. The size allocated to each hour corresponds to the number of retweets during that hour.



Storyline:



Full-size interactive plot:
https://github.com/smelancon/datavis_project/blob/master/visualizations/twitter-network.png
https://github.com/smelancon/datavis_project/blob/master/code/twitter_graph.ipynb

This is a graph visualization of Twitter users. To ensure that we focus on active Twitter users, users who do not appear in the dataset more than 10 times are filtered out. Every node is a user, and every edge is a retweet. The nodes are colored by the number of retweets that user received, with dark blue indicating the most retweets. The plot is interactive and you can hover over nodes to view the number of retweets.

There are some clear network effects at play here. Mainly, we can see dark blue nodes, the influencers, scattered throughout the plot surrounded by retweeters. In fact, except for the dark blue nodes, the vast majority of Twitter users are rarely retweeted in the dataset.

Unlike Facebook where there is relationships between users are two-sided, Twitter relationships are one-sided in nature. Influencers define the network effects on Twitter. The masses of users are following relatively few, and their main connection to each other is through following the same influencers. This type of relationship can be clearly seen in the visualization.

Results/Summary/Conclusions:

Overwhelmingly, the main thing I found digging into this data is that Twitter is a different beast to other social networks. Whereas with other social networks, relationships are mutual, Twitter promotes more of an influencer/follower relationship. A few people get retweeted a lot, while the majority of the users are primarily there to consume content. The only way they engage with the social network is by retweeting influencers.

We can see this phenomenon in other areas of the above analysis. Users from the US are getting the most retweets by far (although it should be noted that they also make up the largest group of users). Across all countries, the distribution of number of retweets has an extremely long tail. We can see from the histogram that many users receive few likes and a few users receive many likes.

The takeaway from this analysis is that people don't use Twitter as a social network in the traditional sense, but rather to follow people they don't know.

Appendix Containing All Code:

https://github.com/smelancon/datavis_project/tree/master/code

Link to github page: https://github.com/smelancon/datavis_project

Citations:

Graph dataset: http://snap.stanford.edu/data/higgs-twitter.html

Tabular dataset:
https://dataverse.harvard.edu/file.xhtml?persistentId=doi:10.7910/DVN/JBXKFD/F4FULO&version=2.2

Python Graph Gallery: https://python-graph-gallery.com/