Project #1 Milestone Report

Problem Statement/Challenge

The questions to be answered/addressed in the project are, "Do Americans support U.S. President Donald Trump's position on immigration and, specifically, what many cite as the separation of families as they try to enter/cross the U.S. border? And do those opinions vary by geographic region?" Answers to these questions would be very useful and relevant to political allies and spokespeople—as well as opponents and critics—of the President. In addition to increasing our understanding of how Twitter data provides reflections of these sentiments across the country, this information could be used to shape advertising and marketing campaigns, political platforms, policy formulation, etc.

Project Details and Description of the Data

To address these questions, I used Twitter data obtained directly from the Twitter API. My initial goal was to gather all of the tweets on this topic, perform various analyses to find patterns and identify key variables and statistics, and ultimately plot the tweets on a map color-coded by sentiment, thereby visually demonstrating how Americans feel about the President's position on immigration and to determine if those sentiments varied by region or state.

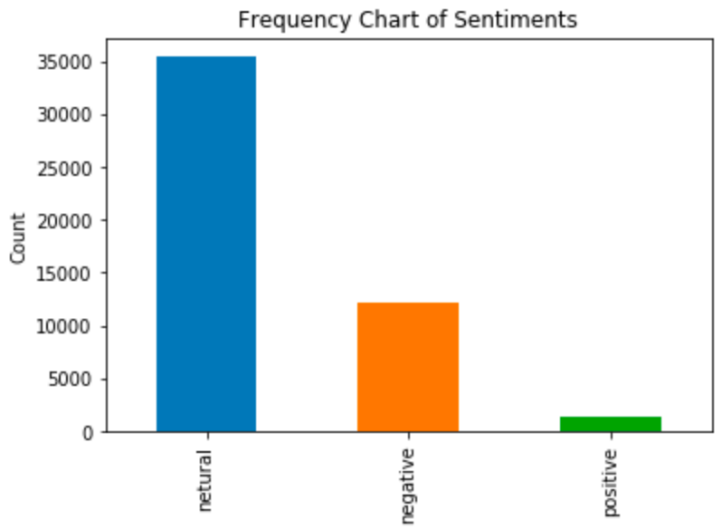
First, I needed to authorize a Twitter API client. As I explored the Twitter API, I realized that standard application access would only allow me to see (and query) tweets for the last seven days. Therefore, I changed topics from my initial proposal. As opposed to focusing on an event that happened a few weeks ago (i.e., the summit with North Korean leader Kim Jong-Un

as initially proposed), I focused on a current topic such as the aforementioned presidential position on immigration and how to deal with families as they try to enter the United States. This felt like a minor change and the code (and subsequent analyses) allows for the simple insertion of any topic. Next, I learned that specific location information is not typically provided/enabled by the average Twitter user. For the degree of fidelity I wanted (specifically geographic coordinate), a Twitter user has to actively and purposefully enable those services. Therefore, I was forced to rely on one object in the geolocation metadata called 'place_id'. I first established a list of the top 100 cities in the United States (based on population) and then queried the Twitter API for the place_id codes for each city. I then converted the results into DataFrames, created new columns to associate the cities with states and regions, and finally merged them to establish my final dataset.

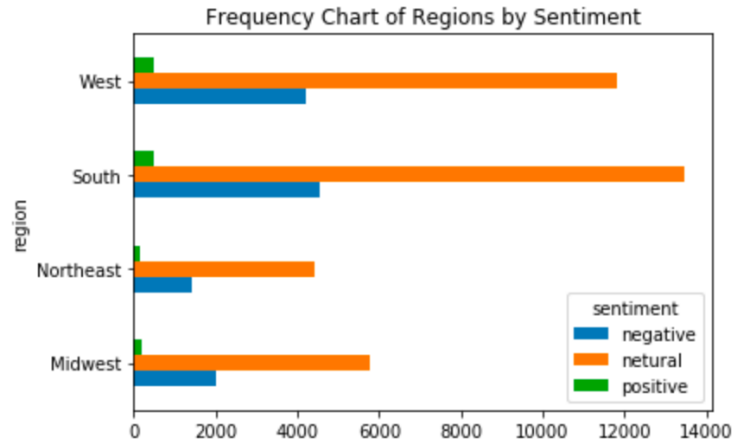
Another overcoming this hurdle, the data was relatively clean and required little to no additional wrangling. Finally, I created a 'sentiment' column with the results of my sentiment analysis that labeled each tweet as "positive", "negative", or "neutral."

Initial Findings from Exploratory Analysis:

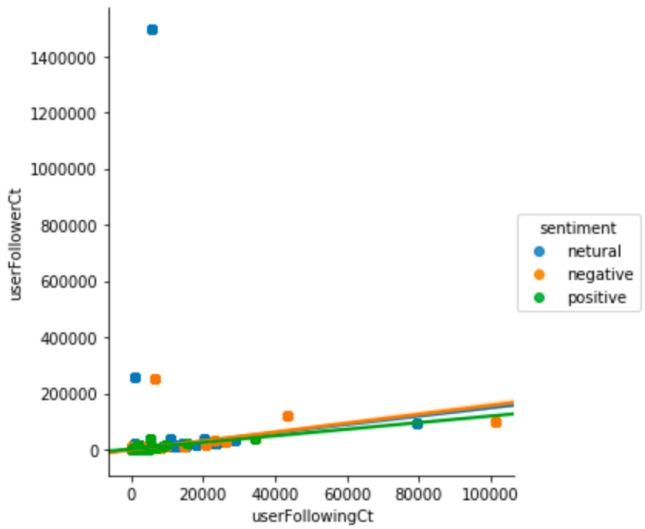
In analyzing the Twitter data, I obtained directly from the Twitter API, I discovered that the number and percentage of "positive", "negative", and "neutral" tweets were pretty interesting. Specifically, there were far more "neutral" tweets (72%) than "negative" (25%) and "positive" (3%) tweets, as one can see in the following bar chart:



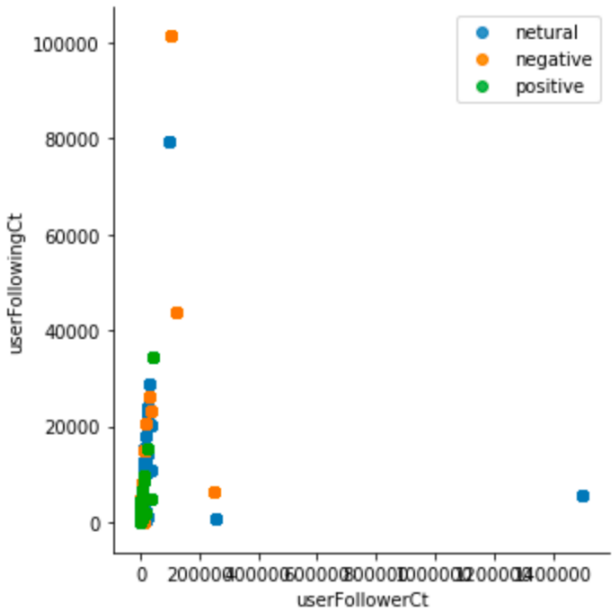
I utilized bar charts to highlight this as well as dividing the tweets by region and then by sentiment. There were more overall tweets from the South and West. The percentages of “neutral”, “negative”, and “positive” by region was consistent with the overall percentages mentioned above. By comparing within region percentages, one can see that the West region had the highest percentages of “negative” tweets and “positive” tweets with the lowest percentage of “neutral” tweets. The Midwest region had the lowest percentage of “positive” tweets.



I utilized a number of Seaborn strip plots to learn how the sentiments vary across, for example, number of retweets and a user's number of followers. Additionally, the number of Twitter followers is positively correlated with the number being followed and were evidenced by linear regression models and plots. The relationship is even stronger among those with "negative" sentiments about the President's immigration position.



I created an empirical cumulative distribution function graph to see that the distribution of the number of user's followers is very similar to the distribution of the number of people they are following. The following scatterplot reveals this general relationship, subdivided by sentiment.**:**



Additionally, I found that very few of the variables were correlated with each other. In fact, only the number of times a tweet was liked/favorited and the number of user followers were strongly correlated. Their correlation coefficient was 0.92.

Moving forward, I will continue to explore these relationships between the other Twitter metadata elements; particularly with the sentiment variable. I will also explore graphing options to begin plotting the tweets by state and region.