Data Wrangling

The question to be answered/addressed in my project is, “Do Americans support U.S. President Donald Trump’s position in 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?” Using Twitter data, I intend to perform sentiment analysis to help answer these questions about the president’s position on immigration.

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. As opposed to focusing on an event that happened a few weeks ago (i.e., the summit with North Korean leader Kim Jong-Un), 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. Because location information is not typically provided/enabled by a Twitter user, I relied on one object in the geolocation metadata called ‘place_id’, which was first accessed by establishing a list of the top 100 cities in the United States (based on population) and querying 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.”