Inferential Statistics

Given that my questions to be answered/addressed in my project are, “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?”, a number of my variables were categorical. This limited the analyses I could conduct during the inferential statistics phase.

That said, using the Twitter data I obtained from the Twitter API, highlighted 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. I utilized bar charts to highlight this as well as dividing the tweets by region and then by sentiment. 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. Linear regression models and plots revealed a slight positive linear relationship between the number of followers a user has and number of people they themselves are following. I utilized 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.

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.