Wrangle Report

By Steven Coleman
Udacity Data Analyst Nano Degree

To wrangle the data for this project I first imported the required libraries for this exercise and then added the provided code to setup Tweepy to access the Twitter api. After entering in my credentials, I set about gathering the necessary data.

For the first dataset, I downloaded and uploaded it to the notebook directory. The file was then read in and converted into a dataframe. The second dataset containing the predictions data was retrieved using the requests library and converted into a pandas dataframe. The third dataset containing retweet and favorites data was obtained after and extended period of making calls to Twitter's api and writing that information to a text file. Once all of the information was saved, the text file information was read in as a pandas dataframe.

After gathering all the information, I set about examining it. I assessed the data visually and programmatically using pandas functions such as head and info. Additionally I defined each column header, explaining what type of data each column contained.

Upon completing my assessment of the data I set to work finding and recording a list of quality and tidiness concerns I would be addressing with the dataset including missing information, inconsistent formatting, excessive columns, etc. When this step was completed I created copies of the original dataframes in order to make changes I needed to without altering the original data. Then I set about addressing each concern making sure to clearly document what dataframes I was working, defining the problem, enacting my solution, and then testing it. Some actions included removing rows with insufficient information, changing character casing to establish consistency, combining columns, and merging dataframes.

Once these steps were finished, I stored the cleaned dataframes in new csv's in the notebook directory before beginning my data insights.