Wrangle Report

Firstly, I began gathering the data by downloading the on hand file *twitter_archive_enhanced.csv* to the Jupyter Notebook. I then wrote code using Python's requests library to programatically download the image_predictions file that is hosted on Udacity's servers. Finally, for the gathering phase, I used tweepy and queried the Twitter API for each tweet's retweet count and favorite count. I had my code write this data into a file called tweet_json.txt, however I added formatting to that of a .csv file, making it easy to use the Pandas library to use the *read_csv* function. I then used this function to read in all the data to Pandas DataFrames.

Next, I called the head() function on the DataFrames and began visual assessment for any quality or tidiness issues in each data set. I initially found several issues in the *twitter_archive_enhanced* data set, which I documented before moving onto using code to test for further issues. Using code, I verified my suspicions that the dates in the data set were not DateTime objects, but just strings. I also used code to test what was actually in the “doggo, floofer, pupper and puppo” columns, as all of the entries I could visually see were just “None”. I then decided these columns could be condensed down to a single column, as each column only had it's name as an entry besides the “None”s. Satisfied with documenting these issues among other smaller ones, I moved onto the cleaning phase of wrangling.

For the cleaning phase, I first changed all of the datatypes to DateTimes and strings as necessary, as this is the easiest and quickest fix. I then created a new column for dog keywords (puppo, floofer etc.) and concatenated each entry from the relevant columns, replacing the “None” entries with empty strings. Finally, with help from my mentor, I was able to use regular expressions to extract the actual source of tweet access from the source column. After this, I renamed the images columns to be more descriptive. Satisfied with my cleaning, I then prepared for the visualization process.