Tweepy Data Wrangling Act Report

Once I had my data cleaned, there were several areas of the data that I wanted to explore in order to find trends and correlations. Initially, and since I had a variable with the timestamp for each tweet, I was interested to see if the twitter feed had increased in popularity over time, in terms of average retweet or favorite count per tweet. I also, wanted to see if different dog keywords ('floofer', 'puppo', etc.) corresponded to different average retweet and favorite counts, and if ratings lower than 1 (9/10, 5/10, 1/10, etc.) were less popular in terms of average retweet and favorite counts.

When I started graphing the retweet count and favorite count for each tweet in chronological order, the data was quite volatile and it was difficult to make out any sort of trend or correlation. However, thinking back to Unit 1, I decided to use a 90 tweet moving average for retweet and favorite average graphing, and the results were much easier to interpret. The graph for both showed a strong positive correlation over time, which makes sense as the twitter feed probably did grow followers and viewers over this time period.

It was difficult to figure out how I was going to graph the retweet_count and favorite_count means for each dog keyword, since I was having trouble using the groupby function to get all of the means in one graph. I settled on using a function that returned the retweet_count and favorite_count means for each keyword in a DataFrame, then graphed that function in use. While tweets with the word 'puppo' in them seem to perform better than others on average, it's difficult to tell whether this is because of random variation or true statistical significance.

Finally, I used the value_counts function in combination with a mask to create a subset of the data for tweets with ratings below 1. I then used the .describe() function to look at some of the key statistics of each DataFrame, and I found that the average tweet in the dataset had over 2x more retweets than a tweet in the low ratings subset (2767 vs. 1035), and more than 3x favorites (8897 vs. 2854).

This project has been tremendous in helping me understand the data wrangling process, and has taught me many things about gathering, assessing, and cleaning, which is crucial to data analysis as a whole. Using Twitter's API and a neural network has been very interesting and informative as well, as each have posed additional ideas about gathering data, and just proves that even in these professional and generally well- organized datasets, cleaning is often a difficult but necessary process to make sure data is appropriately projected.