Wrangle Report

Phase 1: Gathering

I began this project by importing my packages and data sets. An enormous challenge for me at the start, was determining the proper code with which to access the twitter API. After several attempts and consultations with the internet, I found that a “for”, “try”, “except” loop was the only method that could extract real data from the source.

Phase 2: Assessing

After gathering the data, I began to assess it through a visual technique before utilizing the programmatic approach. On first glance, there were several inconsistencies in the twitter_df dataframe. I then explored these visual assessments using info(), duplicated(), describe(), an loc() functions.

In the end, I determined the eight quality and two tidiness issues that I would resolve in this project. In tweet_df, I would remove erroneous datatypes, change timestamp column to date time, find missing values for five columns, and remove the additional characters in front of the urls in the source column. In the image predictions dataframe, I needed to lowercase the p1_dog, p2_dog, and p3_dog columns, remove retweets, resolve confidence levels over 1 in the p2 and p3 columns, and strip away the duplicate images.

Phase 3: Cleaning

Upon resolving these issues in the cleaning steps, I was able to create an additional file with which to utilize the cleaned data to generate visualizations. I focused on favorite counts across time, favorite counts vs. retweet counts, representations of the different dog_type, as well as the rating numerators. From these visualizations came several key takeaways from this data set:

1. In fact, as the author tweeted more frequently (across time), the higher the numerator rating became. This could be due to the fact that a high numerator became more characteristic of the twitter user and thus became more of a signature.
2. Also, favorite count and retweet count are extremely positively correlated. Favorite and tweet counts are also positively correlated with tweet_id, suggesting that tweets increased in popularity with the passage of time/tweets.
3. Finally, the data shows that tweets are much more likely to receive a higher numerator rating at in the early and late hours of the evening: 0400, 2000, 2100. Perhaps due to an urge from the user to generate as much attention/shock from twitter followers as possible.