Data Wrangling Report

Data Compilation

There were three overall sources of data. Two, a csv and tsv file, were provided by Udacity and uploaded to the project. The third was collected by querying Twitter API and save into a JSON file. The CSV file, twitter_archive_enhanced.csv was originally @dog_rates twitter archive that was partially cleaned. The tsv file, image_predictions.tsv , was generated from a neural network which provided dog breed predictions for every tweet including a picture. The data extracted from twitter, tweet_json.txt, was a simple text file containing each individual, tweed id, favorite count and retweet count.

Data Assessment

Pandas functions like. describe, .value, counts(), and .info were used for the initial exploratory data investigation. After gaining a comfortable understanding of the size, shape and context of the data, I moved on to the cleaning stage.

Data Quality Issues

1. The columns tweet_id, in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and timestamp, in the tweet_archive dataframe are the wrong data type.
2. 'Tweet_id' is the wrong data type in the additional_tweet and doggo_pred dataframes.
3. Entries which are not pictures of dogs
4. Several tweets have been deleted and are missing from the API data
5. Missing data from the dog breed predictions
6. Incorrect dog names such as 'a', 'the', etc
7. Only looking for original ratings, not tweet replies
8. The variables ‘type’, ‘p1’. P1_dog, p2’, ‘p2_dog’,’ p3’, ‘p3_dog’ are the wrong data type
9. Missing Dog types

Data Structure Issues

1) The additional_tweet_data, tweet_archive and doggo_pred needed to be merged into a single data frame
2) The doggo, pupper, puppo and floofer column are values of the same column.

Data Cleaning

Compared to the assessment portion, data cleansing (or wrangling) was by far the most time-consuming part of the project. While it was frustrating at first, but became more and more enjoyable as I began to improve. I think some of solutions were overly complicated and technical, but it was good learning experience. Similarly, I learned to be more comfortable with ambiguity in data, there are so many exceptions and special cases, that getting everything 100% perfect would be unnecessarily complex and slow.