Wrangle Report –

For this analysis, I wrangled three different datasets. The first was a dataset I downloaded using the requests library from https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv. This was the dataset which included the breed predictions from the images on the weratedogs tweets.

The second dataset used was provided in the project materials and was read into a dataframe from the csv file. This was the twitter archive dataset.

2. The third dataset used was a dataset created from api requests from the Twitter API developer portal. To create this dataset, I created a twitter developer account and used the Tweepy library to make api "df_twitter_archive" columns "doggo", "floofer", "pupper", and "puppo" have "None" listed for all values. These columns hold no value and should just be dropped. (Tidiness issue because columns should store actual/useful variables and these columns store nothing).

requests of all the tweet ids from the other two dataframes.

Due to the request limit from the twitter api, I made a function which paused the requests for fifteen minutes every 900 requests.

I saved these requests in a json format in a .txt file, then created another function which read the file line by line and parsed the json format to extract only the tweet ids, favorite counts, and retweet counts for the dataframe.

I used both visual and programmatic assessment to see what needed to be cleaned in the datasets. I found the following issues and addressed each of them:

1. In all three dataframes, the tweet_id was a numeric value and not a string/object. I converted each of the columns to the string/object datatype because the tweet_ids are only used to identify the tweets and should not be used for numeric calculations.

2. In "df_image_predictions", breed names were not uniform in their casing (some start with uppercase, some lowercase). I used the str.lower() method to change all of the casing to lowercase to be uniform.

3. Columns for "in_reply_to_status_id" and "in_reply_to_user_id" were changed to strings/objects instead of numbers because they are used to identify other tweets, not for numeric calculations.

4. NaN values in "df_twitter_archive" dataframe. I ended up dropping several columns including "in_reply_to_status_id", "in_reply_to_user_id", "retweeted_status_id", "retweeted_status_user_id", and "retweeted_status_timestamp." These columns were ultimately dropped because the vast majority of the rows were NaN, and they were not useful for further analysis across the datasets.

5. I changed the rating numerator and rating denominator columns datatypes to floats in case future ratings have decimals to account for.

6. Looking closer at the text values in the twitter archive, I noticed some text saying, "don't send (x)" (such as "please don't send in pictures without dogs, we only rate dogs, not porches", etc. I dropped these rows since they do not contain ratings of dogs (even if they still contain ratings, they are not the same as the normal weratedogs dog image ratings, so they shouldn't be compared in the dataset).

7. In "df_twitter_archive", "Timestamp" column was just a string/object, not a datetime type. I used the pandas to_datetime method to convert this into a datetime format instead of just a string.

8. In "df_twitter_archive" "name" column - several values were obviously not names. I manually looked through the unique values in the name column and removed any values which were obviously not names (such as “a”, “an”, “unacceptable”, etc). I replaced these values in the name column with the string “NOT FOUND,” as the name was unable to be parsed correctly.

Tidiness issues:

1. In "df_twitter_archive" columns "doggo", "floofer", "pupper", and "puppo" all described the same value which is the how WeRateDogs classifies the dog in the post. These columns were condensed into one column as "dog_class."

2. According to the rules of tidy data, I created a master dataframe by merging the three dataframes and saved the master dataframe into a separate .csv file, twitter_archive_master.csv

This concluded my wrangling and cleaning efforts for these datasets.