Wrangle Report

Intro

This is a Data Wrangling Project to clean the WeRateDogs tweet data into a functional and effective data set. I will also be using this data to create a few visuals and insights into the analysis.

Gather

1. The WeRateDogs Twitter archive; I imported the "twitter_archive" into a data frame using pd.read_csv()
2. The tweet image predictions; the "image_predictions" data contains a variety of dog breeds and I downloaded it programmatically using the URL provided by Udacity
3. Twitter API for each tweet's JSON data using Python's Tweepy library; this gave me a little trouble at the beginning but was able to figure it out. Initially, I had to query Twitter API for each tweet in the Twitter archive and save the JSON in a text file. From here I queried the JSON data into tweet_json.txt and used json.dump. From here, I created an empty data frame to load the .txt file line by line. I checked to ensure my code worked using tweets_df.head(5).

Assess

When assessing the data, I looked for two main key things, Quality and Tidiness Issues. Some of the quality issues I looked for involved any possible missing or wrong data. Whereas tidiness issues involve any organizational or visual problems when observing the data set. I ran a couple informational lines of code (.head(), .tail(), .info(), .value_counts, etc) to take a look at each data set as a whole.

Cleaning

From here, I identified tidiness and quality issues to be cleaned.

Tidiness Issues

1. In twitter_archive_df, the separate columns of 'doggo', 'floofer', 'pupper', 'puppo' should be combined into one
2. To make the data cleaner and easier to read, I'm going to combine tweets_df and twitter_archive_df

Quality Issues

Twitter_archive/tweets

1. A few dog names are wrong and should be fixed
2. To keep formatting the same, we'll have the dog names always have a capital letter to start
3. The standard denominator is 10 for the ratings; fix the ones that aren't.
4. There are some abnormal numerators that need to be fixed, as well.

5. Some columns have a lot of missing data, for example, "in_reply_to_status_id", "in_reply_to_user_id", "retweeted_status_id", "in_reply_to_user_id", "retweeted_status_id", "retweeted_status_user_id", "retweeted_status_timestamp". I don't really need these pieces of info so I'll just remove them.
6. The dtype for "timestamp" is wrong and should be fixed
7. The extremely long URL's should be shortened
8. The "expanded_urls" has few missing values. These should be void when looking at the rating

Image Predictions

1. The names such as p1, p2, p3 could be confusing. I'll clean these up to make it very easy to interpret
2. Make dog breed naming convention consistent; have capital letters at start of each name

Before moving on to cleaning, I made a copy of each data set to ensure a fresh start and for consistency