Wrangle Report

Gathering Data

I have gathered the data from three different sources.

1. Enhanced Twitter Archive
   This data file was given and I have directly downloaded it from the udacity.
2. Image Predictions File
   Image Predictions data file was imported from a given url.
3. Twitter API
   The data was gathered as a JSON file from Twitter API using Tweepy library and then the JSON file was imported to the dataframe.

Assessing Data

The Assessment was done both visually and programmatically using the following methods:

- .info()
- .head()
- .describe()
- .value_counts()

Tidiness

1. Column names 'doggo', 'floofer', 'pupper', and 'puppo' in Twitter Archive are values. This can merge into one column.
2. All three data frames need to be merged into one df.

Quality

1. Tweet Ids should be in string data type instead of int.
2. In Twitter Archive, 'timestamp' should be datetime data type instead of string.
3. In Twitter Archive, row 313 has invalid rating denominator. (row 313, denominator is 0)
4. In Twitter Archive, the records with 'rating_numerator' equals to 0 need to be removed for better analysis. It is not logical to have a 0 rating.
5. In Twitter Archive, 'rating_numerator' and 'rating_denominator' should be in float data type to calculate ratio.
6. In Twitter Archive, there are some invalid dog names. ('a', 'an', 'None', 'very', etc) All the dog names that start with lowercase letters are invalid names.
7. In Twitter Archive, entries that are retweets or replies should be removed.
8. In Twitter Archive, there are unnecessary columns for analysis including reply, retweet, source and expanded_urls.
9. In Image Predictions, if the neural network did NOT recognize a dog at all, we will drop the records. If it did recognize, we will record the highest probable dog breed only.
10. In Image Predictions, after classifying the dog breed, we won't need all the other columns. (will require only 'tweet_id' and new column 'dog_breed')
11. In Image Predictions, underscores for the names should be replaced with spaces. It should also start with uppercase letters.
12. In Twitter API, there is data with the zero favorite when the retweets are at a couple thousands. We assume that these are incorrect data.

Cleaning Data

I have cleaned the data by resolving the issues from the assessment note in order.

Each issue was cleaned in the process of three parts: Define, Code and Test.

For tidiness issues, I have merged the four columns ('doggo', 'floofer', 'pupper', and 'puppo') into one categorical column named 'dog_stage' and also merged all three dataframes into one dataframe.

For quality issues, I have noted below the list of the cleaning data I did.

1. I converted 'tweet_id' data type from int to string
2. I converted 'timestamp' data type to datetime format
3. I removed the rows with the rating numerator or denominator equals to 0.
4. I converted the rating_numerator and rating_denominator data type to float to calculate the ratio. I removed the rating_numerator and rating_denominator after.
5. I dropped all the rows with dog names that start with lowercase letters and the records with dog names that are 'None'.
6. I dropped the rows that have retweeted_status_id or in_reply_to_status_id values.
7. I dropped 'in_reply_to_status_id', 'in_reply_to_user_id', 'source', 'text', 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp', and 'expanded_urls' columns.
8. I dropped the rows with all 'p1_dog', 'p2_dog', and 'p3_dog' values being False. I saved the name that matches with the first True value under 'p1_dog', 'p2_dog', and 'p3_dog' columns to the new column named 'dog_breed'.
9. I dropped all the columns for Image Predictions except for 'tweet_id' and 'dog_breed'.
10. I have replaced underscores with space for column 'dog_breed'. Also I ensured values under the column 'dog_breed' to start with uppercase letters.
11. I dropped the rows with zero favorites.

At the end of the cleaning process, I reordered the columns as the following.
['tweet_id', ‘'name', 'timestamp', 'dog_stage', 'dog_breed', 'rating_ratio', 'retweet_count', 'favorite_count', 'text', 'jpg_url']