Wrangle Report

Greg Moffatt

For this analysis, we looked at the WeRateDogs (@dog_rates) Twitter account. The data included an archive of tweets up to 8/1/2017. This is the 'twitter_archive_enhanced.csv' file that was directly downloaded to my computer and saved in the appropriate folder. The second file is the 'image_predictions.tsv' file that was downloaded using the Requests library. The final file was the 'tweet_json.txt' file. This was downloaded directly from Udacity and placed in the appropriate folder like the first csv file. I was unable to get the Twitter API to work correctly.

I loaded the pandas, numpy, tweepy (which wasn't ever used after my issues with the API), requests, matplotlib, and json libraries. These libraries included the commands I would need to use to gather, assess, and clean, and analyze the data.

Step 1: Gather data

The 'twitter-enhanced-archive' file was read into Jupyter using the pd.read_csv command, giving it the name of df_arch. The 'image-predictions' file was read in a similar manner, but we needed to add the "sep='\t'" option since it was a tsv, or tab-separated value file, that had a different delimiter than a csv (comma-separated file) does. This was given the name df_image. Finally, the 'tweet_json" text file was read in line by line and given the name df_json.

Step 2: Assess data

The data was assessed in 2 different ways. First, I did a visual assessment. I loaded each dataframe, looked at the head, tail, samples , and info of the dataframes. Then I did a programmatic assessment of the dataframes, looking at duplicates, and counts that are 0 that shouldn't be.

Step 3: Clean data

After coming up with the list of items that needed to be cleaned, I set out to clean the data doing the following:

- Changing datatypes of columns of data
- Removing rows with quantity of 0 that shouldn't be 0
- Removing rows that have null values
- Dropping columns that are no longer needed
- Removing rows that have no type of dog in any of the 3 rows that try to identify the breed of dog by the image posted
- Removing duplicate rows
- Melting multiple columns into 1
- Combining all datasets into one file

Step 4: Storing data

Once the data has been cleaned and combined into one dataset, it was saved as a new csv file named “twitter_archive_master.csv”. This file was used for the next step.

Step 5: Analyze and visualize the data

I looked at the data to find the following pieces of information:

1. What is the dog with the highest numerator?
2. What is the most common rating for a dog?
3. What are the top 5 most popular dog breeds (by number of favorites)?

I found that a Labrador Retriever had a rating of 165/150. This is the highest rating by numerator with the cleaned data.

The most common numerator for a dog rating was 12.

And the top 5 dog breeds by number of favorites are Golden Retrievers, Pembrokes, Labrador Retrievers, Chihuahuas, and French Bulldogs