Wrangling Report

Gathering

Initially the twitter_archive_enhanced.csv file which acted as the file on hand was stored in a folder named ‘Source_Files’. Then with the use of requests library, image_predictions.tsv file is programmatically downloaded and stored in the same folder. Finally using Tweepy library, we store each tweet’s JSON data (favourite count, retweet count) in a text file named tweet_json.txt which is also store stored in Source_Files folder.

Assessing

The assessing began with the creation of data frames in pandas. Each source file twitter_archive_enhanced.csv, image_predictions.tsv and tweet_json.txt were loaded into data frames twitter_archive, image_predictions and tweet_archive respectively.

twitter_archive

- It is observed that some columns has irregular values which can be corrected by converting its data type.
- It is observed that the columns rating_numerator and rating_denominator were extracted from the column text. We can see that whenever there are multiple instances of expression x/y it chooses the first one even though the second one is the score. Some times the value of rating_numerator are in float. Since rating_numerator’s assigned data type is int, the value after the decimal point is omitted.
- By using value_counts function we found out that some rating_denominator’s values are less than 10.
- It is observed that some tweets doesn’t have ratings.
- It is observed that columns (doggo, puppo, upper, floofer) are just different dog stages and must come under a single column,
- Retweets of tweet in data frame also had rating.

Image_predictions

- By using duplicated function we found out that some entries in jpg_url columns are duplicated.

tweet_archive

- It is observed that some tweets doesn’t have a tweet id.

Cleaning

1. Cleaning started with finding the tweet IDs of the retweets. It was achieved by converting the data type of retweeted_status_id column to integer and deleting the rows which didn’t have a retweeted_status_id. Then the columns associated with retweets were no longer required and they were dropped.
2. The data in the columns doggo, puppo, pupper, floofer were combined and put in an entirely new column dog_stage. Then it is separated according to the column values and one’s without values are named None.
3. The data type of some columns in twitter_archive dataframe are changed.
4. Rows with tweets without rating are deleted.

5. With the use of str.extract function on column text, rating_numerator and rating_denominator are extracted. By converting data type of rating_numerator to float, we solve the earlier problem of having decimal values in rating_numerator.
6. Rows with duplicated images in image_predictions dataframe are removed.
7. Rows without tweet IDs in tweet_archive dataframe are removed.
8. tweet_archive is merged with twitter_arvchive dataframe.
9. img_num colum in image_predictions dataframe is deleted since it isn’t required for analysis.
10. Some columns in twitter_archive dataframe are dropped too since they aren’t required for analysis.
11. p2, p2_conf and p2_dog columns of image_predictions dataframe are merged with twitter_archive dataframe since p2_dog column has more hits in predicting the dog breed when compared with p1_dog and p3_dog.
12. Finally the two dataframes are exported to Source_Files folder as csv files (twitter_archive_master.csv and image_predictions_master.csv)