Outline of the Data wrangling effort

In this project we followed the three essential steps of data wrangling which is data gathering, data assessing and data cleaning. These steps will be discussed in detail below.

Data Gathering:

In this project we had to gather data from three separate sources. The first source of data was called twitter_archive_enhanced.csv already given as a csv file that could be downloaded, hence getting this data was just a simple case of importing it using the pandas read_csv function.

The second source of data was from a file hosted on Udacity servers called image-predictions.tsv. This file was downloaded programmatically using the requests library. The data was then imported using the pandas read_csv function

The third source of data was hosted on twitter so I had to use the Twitter API to get the data from Twitter. The library that was used for this was the Tweepy library. The data was compiled and stored in a file called tweet_json.txt

Data assessing.

Data assessing involves checking the data quality and tidiness. In the dataset eight data quality issues were identified which are outlined below:

1. retweets should not be in the dataset as only original dog ratings are needed.

2. retweet-related columns are not needed after removing rows with retweets as these columns become empty.

3. values labeled None in name column should be replaced by np.nan

4. timestamp type is inappropriate, it is string instead of datetime object and also tweet_id type is inaapropriate it is int instead of string

5. the source column should not contain urls, it should just contain twitter device description of where a tweet comes from

6. Column headers should be more descriptive (examples of non-descriptive column header are img_num, p1, p1_conf, p1_dog, etc.)

7. invalid dog names like a, such, getting, which are usually in lower case should be removed entirely.From visual inspection I was able to deduce that those entries are not actually dogs but some other animals.

8. Some of the rating numerators are not correctly extracted. For instance, 9.75 is extracted as 75

Data Tidiness has to do with the structural issues of data. The data tidiness issues are outlined below.

1. The three tables make a single observational unit so there is no need for them to be split
2. The columns doggo, floofer, pupper, puppo should be variables under a column named dog_stages.

Data cleaning

Data cleaning is the final step in the data wrangling process. The steps taken in the data cleaning project is outlined below.

1) Copies of the original data from the different sources were made.
2) The device used to access twitter was extracted from the source column using the beautiful soup package.
3) I removed all retweets from the twitter_archived_enhanced dataset and then dropped the columns in_reply_to_status_id ,in_reply_to_user_id, retweeted_status_id ,retweeted_status_user_id ,retweeted_status_timestamp from the twitter_archive_enhanced dataset.
4) I removed all entries that are not title case in the name column from the twitter_archive_enhanced dataset and then replace all None values in the column with np.nan
5) I renamed columns p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3, p3_conf, p3_dog in the Image_predictions dataset to more suitable names
6) The columns doggo, floofer, pupper, puppo were combined into a single column called dog_stages. Also the term multistage was used for dogs in more than one stage.
7) I used regex to extract decimal values for rating_numerator
8) I also converted tweet_id in the twitter_archive_enhanced dataset and tweet_id in the Image predictions dataset from the int type to string type and timestamp column was converted from string to datetime data type
9) The tables of twitter_archive_enhanced, Image_predictions, and tweet_json.txt were merged together.
10) The rows that did not contain values for favorite_count and retweet_count were dropped

Finally after the cleaning the project was saved as twitter_archive_master.csv file