Merna Ghaffar

Gathering the Data

Data was collected from three resources: an existing file twitter-archive-enhanced.csv was downloaded and saved as *df.* A second file 'image_prediction.tsv' was downloaded programmatically from Udacity servers using the Requests library and was stored as *image_predict.* A third dataset with retweet count and favorite_count was retrieved from Twitter servers in the form of JSON entries, using the Tweepy and saved as *tweet_info.*

Assessing the Data

Assessing the data involves examining both data quality and tidiness. The following highlights some of the issues that I come across:

Quality Issues, which relates to content issues (completeness, validity, accuracy, consistency) of our data

1. Tweet_id is an integer and should be a string in all three datasets
2. Some tweet_id are NAN and should be dropped
3. We only need retweeted_status_id is NAN, drop everything else
4. 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_statis_user_id' are columns that are not useful to us and need to be dropped.
5. Timestamp is set as a string and should be changed into datetime
6. Drop ratings without images
7. Most possible breed column with a prediction confidence column should be created to replace the several breed prediction columns
8. Replace the 'doggo','floofer','pupper','puppo' columns with a 'stage' column

Tidiness, which relates to structural issues in the dataset

1. All three databases should be merged together (df, image_predict and tweet_info) into one database
2. Rating numerator and denominator should be in one rating column (numerator/denominator)

Cleaning the Data

The mentioned points were addressed and the cleaned data was stored as 'twitter_archive_master.csv'.