WeRateDogs Twitter Archive - Act Report

-By Piyush Kumar

Gathering of Data:

There are 3 datasets which are extracted:

1. WeRateDog twitter enhanced archived is downloaded manually from the link provided in the section "Project Details" in the udacity servers and saved with the name “twitter-archive-enhanced.csv”.
2. Then we programmatically download the "image-predictions.tsv" file from the udacity servers.
3. Then we extracted the Json data of tweets using twitter api and some specific data is extracted such as "tweet_id, retweets, favourites".

Accessing and Cleaning of Data:

Quality Issues

twitter_archive:

1. Missing data in the following columns:
   a. In_reply_to_status_id
   b. in_reply_to_user_id
   c. retweeted_status_id
   d.retweeted_status_user_id
   e. Retweeted_status_timestamp
   f. Expanded_urls

2. Timestamp and retweeted_status_timestamp is an object
3. Source columns have HTML tags
4. This dataset includes retweets, which means there is duplicated data
5. Some of the rows have invalid strings in the name column, e.g. "a", "an", "in".

image_pred:

1.dog breeds are not consistently in p1,p2,p3 columns

tweet_info:

1. id column name to be changed from id

Tidiness Issues

1. Image_predictions and twwet_json files both has to be joined with Twitter_archive file.
2. Dog "stage" variable in four columns: doggo, floofer, pupper, puppo