We Rate Dogs
Data Wrangling report

Data Gathering:

Data is gathered from 3 different sources as below:

- Enhanced Twitter Archive:
  sent it to Udacity via email exclusively for you to use in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017; Download manually as (twitter_archive_enhanced.csv) from Udacity website and imported into Jupyter note book as a Data Frame.

- Additional Data via the Twitter API:
  Back to the basic-ness of Twitter archives: retweet count and favorite count are two of the notable column omissions. so the data (retweet count and favorite count) for each tweet was queried by tweeter API using the tweet id in the Enhanced Twitter Archive data frame and written to a text file called (tweet_json.txt) which was then read parsed in the notebook as a data frame.

- Image Predictions File:
  Every image in the WeRateDogs Twitter archive is was ran through a neural network that can classify breeds of dogs*. The results: a table full of image predictions (the top three only) alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction (numbered 1 to 4 since tweets can have up to four images); Downloaded programmability using the requests library in python then written to a file called (image-predictions.tsv) and imported into Jupyter note book as a Data Frame.

Data Assessment:

Data has both quality and tidiness issues listed below:

- Quality Issues:

1. There are 181 retweeted tweets
2. There are 78 replied tweets
3. tweet_id is of int type and won't be need for and statistics
4. timestamp is of string type
5. There are some rating_numerator out of normal range (10-15)
6. There are some rating_denominator that are not equal to 10
7. None instead of (NaN) in the following columns (doggo, floofer, pupper, puppo)
8. The prediction data has 3 predictions for dog breed of each picture & we only need one.
9. some rows have False in the 3 predictions which means it's not a dog
10. there is discrepancy in the counts of tweets and the counts image predictions which needs to be resolved
11. After visual assessment of tweet text with abnormal numerators & denominator in a sperate excel sheet ,it appeared that some have problem with the regex that extracted them as it didn't account for decimal values, some toke the first fraction appearance which is not the rating.

- Tidiness Issues:

1. Columns (doggo, floofer, pupper, puppo) are values not variable names
2. After cleaning(removing) non-original tweets the 5 columns (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp) won't be needed.
3. The data related to one tweet is spared across 2 tables twitter-archive-enhanced & image-predictions which will be better if all related data is in one table.
4. The retweet count and favorite count are in a sperate table than the tweets table which need to be merged.

Data Cleaning:

- **Get rid of the 181 retweeted tweets**
- **Get rid of the 78 replied tweets**
- **After cleaning(removing) non-orginal tweets the 5 columns (in_reply_to_status_id, in_reply_to_user_id,retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp) won't be needed.**
- **Timestamp from string DateTime**
- **Convert Tweet ID from int to String**
- **None insted NaN in the following coulmns (doggo, floofer, pupper, puppo)**
- **magrge the dog stage columns (doggo, floofer, pupper, puppo) into 1 (dog_stage)**
- **merging the API DF with tweets DF**
- **we should assgin and actual dog breed for each row insted of the 3 prdictions**
- **get rid of the unnessery columns after comming up with the actual dog breed**
- **getting rid of the images with 3 False predictions as they most certainly not dogs.**
- **merging the predictions data to the tweets archive data to form a complete data set**
- **Merge image predictions for dog breed with their tweets**