WRANGLE REPORT

The dataset used for wrangling is the tweet archive of Twitter user **@dog_rates**, also known as **WeRateDogs**. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.

Goals for the Project

- Data Wrangling of the Twitter data which consists of:
    1. Gathering the data
    2. Assessing the data
    3. Cleaning the data
- Storing, analysing, and visualizing your wrangled data
- Reporting on data wrangling efforts and data analysis and visualizations

GATHER DATA

- **Twitter Archive Data**: The WeRateDogs is a csv file that has information about the tweet_id, ratings, dog type and dog stage etc.
- **Image Predictions Data**: The tweet image predictions file is a tsv file which should be downloaded programmatically from the url given by udacity. It consists of predictions of what breed the dog is for each tweet.
- **Tweet JSON file**: The tweet’s retweets and favourite count is gathered using Twitter API and python’s Tweepy library.

ASSESSING DATA

The gathered data should be accessed programmatically and visually for quality and tidiness issues.

In this project, I will be assessing eight quality and 2 tidiness issues in wrangle_act.ipynb file.

CLEANING DATA

Twitter Archive data

1. Remove the retweets as we are interested only in original tweets.
2. Drop the unnecessary columns (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp)
3. Change the datatype of timestamp, rating_numerator, rating_denominator columns.
4. The name column in the table has incorrect dog names.
5. The rating_numerator , rating_denominator column has some incorrect ratings.

Image Predictions Data

1. The column 'jpg_url' has duplicate entries. Those should be dropped.
2. The columns p1, p2, p3 has inconsistent names. Change it to Title case.

3. The column names p1_dog, p2_dog, p3_dog can be renamed to meaningful names.

Tweet Json Data

1. The id column name should be changed to 'tweet_id' for consistency.

TIDINESS ISSUES

1. Merge the three dataframes with respect to the tweet_id.
2. The different dog types column in the twitter archive data can be put into a single column.