Wrangle Report

This report is briefly describing the data wrangling process executed for this project.

The dataset used for this project to be wrangled, analyzed and visualized is the tweet archive of a twitter user @dog_rates, aka *WeRateDogs* which rates people's dogs with a humorous comment about the dog. These ratings have the denominator of 10, however, the nominator could be greater than 10 like 11/10, 12/10, etc. because they are good dogs Brent. WeRateDogs has over 4m followers and has received international media coverage.

This project has been completely done on the Udacity Project Workspace; however, the reports were created using MS Word and exported as PDFs.

The wrangling process has been executed through three main stages:

a. Gather Data
b. Assess Data
c. Clean Data

Let's take a brief look on each stage:

a. Gather Data

The data used were gathered from three different sources:

1. Enhanced Twitter Archive: which contains data from tweet data sent by WeRateDogs to Udacity via email to be exclusively used for this project. The file provides dog names, ratings, dog stages and other related information.
   The file was manually downloaded, then uploaded to Jupyter notebook using this link: https://video.udacity-data.com/topher/2018/November/5bf60fbf_twitter-archive-enhanced/twitter-archive-enhanced.csv

2. Image Predictions TSV file: which produced by running every image in WeRateDogs Twitter archive through neutral network that classifies breeds of dogs. This process resulted in a table full of image protections (the top three only) alongside with each tweet ID, image URL, and the image number that corresponded to the most confident prediction numbered from one to four cents tweets can have up to four images. This file is hosted on Udacity servers and was downloaded programmatically using Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
3. Tweets data using Twitter APIs: the ready-made version was used in this project and was read line by line into a Pandas data frame with tweet ID, retweet count and favorite count.

b. Assess Data

After gathering data from various resources, we assessed them visually and programmatically to check for quality and tidiness issues.

We have found out the following issues:

- *Tidiness:*

1. Dog stages were separated in four columns.
2. The three data frames needed to be combined into one data frame.

- *Quality:*

** Enhanced Twitter Archive" -- twitter_arch data frame*

1. Change the "id" column label in tweets_data_df_cleaned to "tweets_id"
2. tweet_id doesn't need to be an integer. Instead, it should be an object.
3. There are 181 retweets which need to be eliminated.
4. timestamp is an object which need to be converted into datetime.

5. Row 313 has a denominator of 0 which is invalid.
6. Remove rows that doesn't contain images.

** Image Predictions Data" -- predictions_df data frame*

7. tweet_id is an integer which need to be converted into an object.

** Tweets Data from Twitter API" -- tweets_data data frame*

8. There are two rows missing in the data (2356 instead of 2354)

c. <u>Clean Data</u>

The issues mentioned above were cleaned as appropriate resulting a high quality and tidy master pandas Data Frame.