Wrangle Report by Tamer Ahmed:

Introduction:

The purpose of this project is to put in practice what I learned in data wrangling data section from Udacity Data Analysis Nanodegree program. The dataset that is wrangled 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. These ratings almost always have a denominator of 10. This report briefly describes my wrangling efforts.

Project details: The tasks of this project are as follows:

- ✓ **Gathering data.**
- ✓ **Assessing data.**
- ✓ **Cleaning data.**
- ✓ **Analyze data.**

Gathering data, the data for this project consist on three different dataset that were obtained as following:

1. Twitter archive file: the twitter_archive_enhanced.csv was provided by Udacity and downloaded manually.
2. The tweet image predictions, i.e., what breed of is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and was downloaded programmatically using the Requests library and URL information
3. Twitter API & JSON: by using the tweet IDs in the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data using Python's Tweepy library and stored each tweet's entire set of JSON data in a file called tweet_json.txt file. I read this .txt file line by line into a pandas dataframe with tweet ID, favorite count, retweet count, followers count, friends count, source, retweeted status and url.

Assessing data:

Once the three tables were obtained, I assessed the data as following:

- o Visually, I used two tools. One was by printing the three entire dataframes separate in Jupyter Notebook and two by checking the csv files in Excel.

- Programmatically, by using different methods (e.g. info, value counts, sample, duplicated, group by, etc.).
- Quality issues and tidiness issues.

Cleaning data:

This part of the data wrangling was divided in three parts:

- ✓ **Define**
- ✓ **Code**
- ✓ **Test the code.**

These three steps were on each of the issues described in the assess section, First and very helpful step was to create a copy of the three original data frames.

- change datatype of tweet_id i tables into object (**Tidiness-1**).
- I wrote the codes to manipulate the copies. If there was an error, I could create a new copy from the original.
- colect all data of three tables into one table (**Tidiness-2**).
- Condensing Dog Type columns (**Tidiness-3**).
- Condensing dog breed predictions (**Tidiness-4**).
- Convert timestamp to datetime object(**quality-1**).
- Remove Retweets and Tweets which does not include image(**quality-2**).
- removing extra columns ['doggo', 'floofer', 'pupper', 'puppo'] (**quality-3**).
- removing the processed columns ['p1', 'p1_conf', 'p1_dog', 'p2', 'p2_conf', 'p2_dog', 'p3', 'p3_conf', 'p3_dog'] (**quality-4**).
- Removing useless columns ['in_reply_to_status_id', 'in_reply_to_user_id'] (**quality-5**).
- Extract Dog Rates and Dog Count(**quality-6**).
- Extract Names(**quality-7**).
- "a", "the" and all non-name words have been removed.(**quality-8**).
- After cleaning, the data was exported to a tidy master CSV file named “twitter_archive_master.csv”.
- Analyze data (Analyze Report).