Wrangle & Analyze Data Project

In this project, datasets that, I have wrangled and analyzed, belong to 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.

All works approach wrangling process, which consists of:

- Gathering data.
- Assessing data.
- Cleaning data

Starting with gathering data from three different resources, then assessing these data that identify and categorize common data quality issues with data content and tidiness issues with data structure. Finally, cleaning data that have been addressed in the assessing stage after copying dataset to keep original dirty datasets save if need review later.

After completed the wrangling process, merge theses clean datasets to master dataset.

Gathering Data:

Gathering data from three different resources:

1. WeRateDogs Twitter archive, by reading CSV file to Python Pandas.
2. Tweet image predictions, by retrieving tsv file from URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv .
3. Tweet's retweet count and favorite, by utilizing Twitter APIs and using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's *Tweepy* library and store each tweet's entire set of JSON data in a file called then convert JSON file to a data frame.

Assessing Data:

Two types of assessment Visual assessment and Programmatic assessment have been used for assessing datasets. Start with visual assessment by scrolling through the data with *Microsoft Excel*, then programmatic assessment by using codes to view specific portions and summaries of the data like (*head, sample, tail, info and describe)* methods .other method also used to exploring values in interested columns like *value_counts* method. In the Twitter archive table, there are 8 quality issues have been observed and 1 quality issues for each Tweet image predictions table and Tweet table. Also, there are 2 Tidiness Issues (one from Tidiness Issues and the other to merge all tables in a master table).

Cleaning Data:

The data cleaning process has been approached programmatically through 3 steps:

- **Define:** State the action to be performed according to the issues that have addressed in assessing stage.
- **Code**: Running codes to resolved these issues that defined**.**
- **Test**: Testing the data programmatically after cleaning to be sure that cleaning has resolved the issues and worked properly.

Finally, store the clean master dataset in a CSV file with the main one named twitter_archive_master.csv.