We Rate Dogs Data Wrangling Report

In this report I will discuss the data wrangling process for the project named WeRateDogs Project. This project requires me to look at twitter account WeRateDogs data. In this project I used three steps of data wrangling process: Gathering Data, Assessing Data, Cleaning Data.

Gathering Data:

This is the first step of data wrangling process. This project required me to gather the required data for the wrangling process:

- We were giving the WeRateDogs Twitter archive, which we had to download manually from Udacity's link.
- The tweet image predictions file. This file (image_predictions.tsv) is hosted on Udacity's servers and was downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
- Additional Data via the Twitter API: Using the tweet IDs in the WeRateDogs Twitter archive, Twitter API was queried for each tweet's JSON data using Python's tweepy library and stored in a file called tweet_json.txt file

All of this data was gathered in Python Jupyter Notebook Using pandas, requests, and tweepy libraries. Additional steps were required from my side to acquire the twitter API, these steps involved me in creating a developer account to access consumer_key, consumer_secret, access_token, access_secret.

Assessing Data:

This is the second step in the data wrangling process. In this step I inspected the datasets for two things: Data Quality and Data Tidiness. The guidelines for this project required to at least find a minimum of 8 Quality issues and 2 Tidiness Issues.
These are the issues that is Assessed in the project.
Quality:

```
- There are 181 retweets.

- There are 78 response tweets.

- Timestamp column is a string instead of Datetime.

- Missing dog names (replaced with 'None')
```

- Incorrect dog names
- json_tweets has column name id.
- rating_numerator & rating_denominator with wrong data type.
- For columns p1, p2 & p3. The breed names are lower case.
- Breeds named '_'.
- Invalid data type for breed columns.
- There are 66 duplicate jpg_urls.

Tidiness:

- Join all 3 data frames.
- Dog stages should be a single column rather than four

Cleaning Data:
Cleaning our data is the last step in the wrangling process. In this step we look at the quality and tidiness issues that we had from our previous step and address these issues.
This process also required me to look at each issue and break it apart into 3 smaller Sections : Define, Code and Test.
This whole process is done programmatically using various python libraries.