Data Wrangling Report

Introduction:

The dataset I was working on data wrangling is the tweet historical data(user name @dog_rates) provided by Twitter. To complete this project, I needed to wrangle 3 datasets before analyzing the data:

- **Archive file:** It contains 5000+ extracted tweet data of @dot_rates, which includes tweet ID, ratings and dog stages.
- **Image prediction file:** This table contains image predictions of each image through a neural network, which includes tweet ID, predicted breed and image URL.
- Additional data via the **Twitter API**: I needed to query the Twitter API to get more information, such as retweet count and favorite count by tweet ID, which will be useful while analyzing the data.

Wrangling process:

1. Gathering Data

All of three files I mentioned above are from different sources, and I needed to gather these data using different methods:

- **Archive file**: The archive file was given by Udacity, so I was able to download the file manually.
- **Image prediction file**: The file is hosted on Udacity's server, so I used requests library using provided URL.
- Additional data via the **Twitter API**: I queried the Twitter API using Tweepy library and Tweet IDs provided at archive file, and saved the JSON data which was written line by line as tweet_json.txt. After getting these data, I read the data using pandas and only selected tweet ID, retweet count and favorite count.

2. Assessing Data

After gathering all datasets, I started assessing data visually and programmatically. I was focusing on the following questions:

- **Quality:**
    - Is there missing values ?
    - Is there duplications?
    - Are these data correct?
- **Tidiness:**

- Does each row represent an observation?
- Does each row represent a variable?
- Does each table represent an observational unit?

3. Cleaning Data

The first thing I did before cleaning was making a copy of original files so that I don't need to redo gathering process once I removed some important information accidently. Then I started to clean the data based on the observations which I got in assessing process. Each cleaning case was divided by three parts: define, code and test.

- **Quality**:
  - Removed rows that were not about dog's ratings from image prediction table since I only wanted to analyze dog's ratings.
  - Renamed column p1 and p1_conf of image prediction table.
  - Only kept useful columns in image prediction table
  - Deleted retweet and reply tweets of archive table since I only wanted to analyze original tweets.
  - Removed rows that had denominators other than 10.
  - Fixed some incorrect numerators. For example, because of some numerators were with decimals, the extracted data was incorrect.
  - Removed rows that had dog stages more than one type in archive table.
  - Removed tweet IDs of archive table that were not in image prediction table.
  - Only kept useful columns in archive table.
- **Tidiness**:
  - Combined columns of doggo, floofer, pupper and puppo, since all of these represent dog stages.
  - 3 tables should be able to join each others using primary key, and merged 3 tables to 1 table, and saved it as twitter_archive_master.csv.