Data Wrangling Report

- ## Introduction:

  In this project we are going to deal with a real-world data. The data we will analysis and wrangle is **WeRateDogs** twitter account data. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog . I will gather the data from different sources, Assess and clean the data, finally observer some insights and visualization.

- ## Data Gathering

  In this project we will gather 3 different data sets from 3 different sources.

  - **Twitter archive data**: It's on hand file provided by Udacity as twitter_archive_enhanced.csv .
  - **Image prediction data:** I used Requests web scraping library in python to download this data set**.**
  - **Twitter JSON API data:** It's the file that contains the tweet data that's is not provided in twitter archive data set like favorite count, retweet count, followers count. etc. I used Tweepy Python library to reach the data and stored it into pandas' data frame.

- ## Data Assessing

  After gathering all 3 data sets. I used python different tools and libraries to assist the data to identify any data quality or tidiness issues. And I found the following:

- **Quality issues**

1. Timestamp should convert to date-time type instead of Object

2. In my opinion, separate the Timestamp column into 2 columns, month and year is going to be more useful for different type of analysis .

3. There are several columns that is not necessary and useful for the analysis, deleting them in better.

4. Missing values in name column stored as None, and some of the names are not correct names of dogs.

5. There is 20 cases where denominators is greater than 10. it can be deleted consider it is a small amount of the data set.

6. 66 Duplicated images need to remove.

7. Many entries are not dogs like mailbox for example.

8. Some entries are retweets which is duplication of the actual tweets and this entries should be removed

9. Wrong data type for this column (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id). It should be integers instead of float.

10. rating_numerator and rating_denominator should be float data type

11. HTML code in source column make it hard to read

Tidiness issues

1. There are 4 separate columns (doggo, floofer,pupper,puppo) should be combined into 1 column

2. Merging the 3 data frames into 1 data frame

3. (p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3, p3_conf, p3_dog) are taken 9 columns where it can be combined into only 2 columns.

- ## Data Cleaning

  I used python tools to fix some of the issues that were discovered in Assessing data process.

  - Merge the three data frames into one data frame to improve data Tidiness

- Combine the 4 columns (doggo,floofer,pupper,puppo) into 1 column "Dog Stage"
- Convert timestamp data type to (datetime)
- Delete the rows where the rating_denominator > 10 since it is only a very small amount of the data set

- **Analyzing and Visualizing Data**

  I will provide a full separate report about this process.