Data Wrangling Report

Student Name: Ghaida Altamimi

Data Gathering:

The project involved data gathering from three resources:

- **WeRateDogs Twitter archive**

Twitter archive provided by udacity and it downloaded manually

- **The tweet image predictions**

i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

- **Twitter API & JSON**

Each tweet's retweet count and favorite ("like") count at minimum, and any additional data you find interesting. 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 tweet_json.txt file. Each tweet's JSON data should be written to its own line. Then read this .txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count. The file is provided by udacity in case the API request is rejected.

Data Assessing:

In the data assessing part, we searched for different issues, either quality issues or tidiness issues. The issues are listed below:

Quality Issues:

- **Twitter Archive DataFrame:**

1-Wrong data types.

2-NaN values represented as None.

3-Incorrect dogs' name that is start with a small letter, for example a, quite, not, one, an, my, all, old, the, by, and space.

4- Keep original tweets only, the retweets shall be removed.

5- The text column contains both tweets' text and url.

6- Incorrect rating_numerator and rating_denominator (some denominator values do not equal to 10, and there are some numerator values represented as integer while it is float)

7-Inappropriate columns name.

8- Missing data.

- **Image Predection DataFrame:**

  1- Wrong data type.

  2-Duplicate rows in jpg_url column

- **Twitter JSON DataFrame:**

  1-Missing Data

  2- Inappropriate column name "id" and type.

Tideness Issues:

1- The dog life stages are represented in four columns while it should be represented in a single column.

2- All data frames should be combined, since all of them are about tweets.

Data Cleaning:

In the data cleaning part, is used my knowledge and I also looked over the internet i.e. stackoverflow, youtube, and many other resources to guide me to solve the above issues. After that, I stored the clean data frame into a file. Finally, I analyzed and visualized the data to get insights.