WE RATE DOGS: TWITTER RATING ANALYSIS WRANGLE REPORT

Shrinidhi Manakiwala

May 2021

In this report I outline the wrangling efforts to gather and clean the data required for We Rate Dogs twitter account analysis.

Data Gathering:

I have gathered data from two sources:

1. **Twitter archive file**: Provided by Udacity. There are 17 columns and 2356 rows in the data set.
2. **Twitter download**: I had issues with getting access. Hence, used the JSON file provided by Udacity.

Assessment and Cleaning:

I first assessed the twitter archive file. My assessment findings are as under:

- there are 181 retweets (out of scope for this project)
- there are 78 replies (out of scope for this project)
- rating_numerators are not uniform/consistant they range from 0 to 1776. Maximum is 10 but people rate more
- rating_denominator of 0 will cause the rate to be infinity
- some names are missing, entered as "none" and some of them are entered as "a" , "an" , "the"
- doggo, floofer, pupper, and puppo have "None" values
- dog stage as a tidness issue as there should be a single column having these values doggo, floofer, pupper, or puppo
- there should be seprate columns for day, month and year
- Erronous data types:
    - tweet_id is an int
    - timestamp is a str
    - in_reply_to_status_id is float
    - in_reply_to_user_id is float
    - retweeted_status_id is float

    - retweeted_status_user_id is float
    - retweeted_status_timestamp is str
- url has null values

Then I assessed the twitter download and my findings are as under:

- tweet_id is an int

After assessing the data, I cleaned the data in both the files. I started with the **twitter archive** dataframe and changed the 'tweet_id' column to string and changed the format of the 'timestamp' column to datatime format. I capped the values of 'rating numerator' column to 15 as they ranged from 0 to 1700 which would have affected our analysis. After that, I dropped 181 retweets and 78 replies as they were out of scope for the purpose of analysis. I dropped the 'in_reply_to_user_id', 'in_reply_to_status_id', 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp', 'rating_denominator' , 'expanded_urls' as they were not used for the purpose of analysis. I added weekday column as it was later to be used for the purpose of analysis. The cleaned data was then saved to **'twitter_archive.csv'.**

In **tweet download** dataframe, I changed the 'tweet_id' column to string and removed the outliers from the 'favorite_counts' and 'retweet counts' column. The cleaned data was then saved to **'twitter_download.csv'.**