The report of the wrangle efforts

1.1. Data Gathering

The three csv files uploaded successfully, including the WeRateDogs Twitter achieve data, image predictions and the tweets via Twitter API. The main effort was related to the API and the waiting time to upload the tweets with the help of API.

1.2. Data Assessing

Both programmatically and visually data assessing was performed and the cleaning actions were defined as below:

Tidiness issues

- three dataframes needs to be merged into one
- dogs' stages column (doggo, floofer, pupper and puppo) needs to be aggregated into one column

Quality issues

- change the timestamp column's datatype: datatype is object, not a datetime
- drop retweeted_status related columns; retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, in_reply_to related columns; in_reply_to_status_id and in_reply_to_user_id columns which are redundant
- remove the urls from the text column
- format p1, p2 and p3 columns have inconsistency of their string format, having lower/upper case
- duplicate rows control whether there are duplicate rows or not
- name, doggo, pupper, puppo columns have 'None' values but programmatically this format does not seem as NULL.
- clean the rating_numerator values above 15
- detect invalid names in the "name" column such as "a"

1.3. Data Cleaning

The actions based on the assessment were taken.

1.4. Storing Data

The cleaned version of the csv was saved in the local environment.

1.5. Analyzing and Visualizing Data

A copy of the stored csv was generated and first some exploratory analysis was made by using, describe, info methods. Then the focus was on the timestamp column to be able to come up with some trend insights. To do that, additional columns were added such as time and month by using the timestamp column. Then, new data frames were generated for the visualization purposes. The line chart of matplotlib was chosen as one of the best options to visualize the trend analysis.