WeRateDogs - wrangle_report

Gathering data

The goal for this part is to gather data from three different places. The first one is an Archive file provided by Udacity, second, the image prediction data, we need to download programmatically using the Requests library from the URL that Udacity provided. The last one, we create an API object that is used to gather Twitter data. After querying each tweet ID, we write its JSON data to the required tweet_json.txt file with each tweet's JSON data on its own line. You will then read this file, line by line, then we collect the data we need such as favorite count and retweet count to create a pandas DataFrame that you will soon assess and clean. There are 25 tweet_id can't be downloaded automatically.

Assessing data and Cleaning data

I checked three datasets and there are some issues regarding quality and tidiness as below

Quality

In this analysis, I want to focus on dogs only, thus I filtered out the rating lower than 10. In some cases, there are some weird ratings such as 1776 for Atticus. I did some time on WeRateDogs, I found out the highest score is around 15. Therefore, I also filtered out the rating more than 15. Moreover, the rating_denominator should always be 10. I corrected the records that are not equal to 10. I changed the erroneous data types to the correct one.

Tidiness

I noticed that there is a mixed information in "text" column. Thus, I separated comment and url. Image prediction p1,p2,p3 should be in same column as "dog_prediction", pi_conf, p2_conf,p3_conf should be in same column as "conf". Moreover, four columns (doggo, floofer, pupper, puppo) should be in one column as "doggolingo". In the end, I merged the img and json tables to archive table.

Improvement

- Incorrect dog names and default dog name 'a'

- Lots of record is missing dog stages

I may need to assess and clean these columns for better analysis and visualization.