Wrangling Efforts

I first gathered data from 3 difference sources. Twitter data in cvs file provided by Udacity. The image prediction data from Udacity servers. Http request's get method was used to collect the data from the servers. The last data was additional twitter data. I have batch-requested 100 tweet-ids at a time using Tweepy's statuses_lookup API to collect additional data. They were imported into a separate dataframe.

I assessed each data visually and programmatically using python. There were 2 types of issues I was trying identify, quality and tidiness issues.

I identified following 15 quality issues.

twitter_archive table

- Dataset contains retweets in addition to original tweets.
- 'retweeted_status_id', 'retweeted_status_user_id', and 'retweeted_status_timestamp' columns are not needed.
- Missing data : 'in_reply_to_status_id', 'in_reply_to_user_id', and 'expanded_urls'
- Source column contains unnecessary and extra information
- Name column contains words like 'just', 'the', 'a'.
- Some tweets have more than 1 dog life stages.
- Erroneous datatypes : 'in_reply_to_status_id' and 'in_reply_to_user_id' are float.
- Erroneous datatypes : 'timestamp' is string(object).
- Erroneous datatypes : 'doggo', 'floofer', 'pupper', and 'puppo' only containts its own value and 'None'.
- Extract numerator and denomintor from text columns and replace 'rating_numerator' and 'rating_denominator' with correct ones.
- Text column contains text, numerator, denominator, and short-url.
- twitter id 835246439529840640 has wrong numerator and denominator. They should be 13/10 instead of 960/0.
- twitter id 786709082849828864 has wrong numerator and denominator. They should be 10/10 instead of 75/10.

image_predictions table

- The format of predicted dog breeds in column p1, p2, and p3 are not consistent. Some are all lower cases, some use dash instead of underscore, etc.

additional_data table

- There is 1 tweet which has retweet_count 0 and 170 tweets which have favorite_count 0. They could be valid data.

I identified following 3 tidiness issues.

twitter_archive table

- Dog life stage variable is in 4 columns.

image_predictions table

- 'pg_url', 'img_num', 'p1', 'p1_conf', 'p1_dog', 'p2', 'p2_conf', 'p2_dog', 'p3', 'p3_conf', and 'p3_dog' should be part of the twitter_archive table.

additional_data table

- 'retweet_count' and 'favorite_count' should be part of the twitter_archive table

Then, I started thinking in what order it makes sense to tackle the issues.

It was clear that merging of ***image_predictions table*** and ***additional_data table*** to ***twitter_archive*** should be the last steps after all other issues were resolved.

Also, it was clear that filtering out of retweets and only keeping original tweets in ***twitter_archive*** was the first thing needed to be done.

There were other issues that it made sense to do it in certain order. Ex. “Erroneous datatypes : 'doggo', 'floofer', 'pupper', and 'puppo' only containts its own value and 'None'.” doesn’t need to be addressed if “Dog life stage variable is in 4 columns” issue is resolved first.

Once I came up with a plan, I made copies of the original dataframes and cleaned and resolved the issues.

In the end, I only had to tackle 11 quality issues as 4 out of 15 quality issues were resolved as part of previous data cleaning.

Most of cleaning data were pretty straight forward after watching instruction videos and doing jupyter notebook quizzes. However, for “Dog life stage variable is in 4 columns” and “extracting name and numerator/denominator” issues, I really had to think how to approach the problem and worked at them. It was tough, but I really enjoyed the challenge.

Once data was cleaned and tested to make sure the issues were resolved at each step, all 3 tables were merged to one master file 'twitter_archive_master.csv' for further analysis and visualization.