Wrangle Report

1. Introduction

In this project, we analyzed a dataset that contains the tweet contents and other related data from @dog_rates, which is a twitter account that rates people's dogs and gives comments. We followed the Gather → Assess → Clean process to achieve a clean dataset which we can get some useful information from it.

2. Gather

We need three data files in this project and we can gather these data using three different ways.
The first data file is the WeRateDogs twitter archive, we can download the twitter_archive_enhanced.csv directly and use pandas to read it.
The second data file is the tweet image predictions, which contains the dog breeds predictions for the tweet images and the corresponding confidence. We can download it programmatically using the Requests library with the given url.
The third data file contains the retweet and favorite counts of each tweet, we can query the Twitter API to get the JSON data and store it, than we can read it into a pandas DataFrame.

3. Assess

After we got all the data files, we read them and assess them programmatically. In this step, we found some quality and tidiness issues which are listed below.

3.1 Quality Issues

twitter_archive table:

- Tweet id should be a string, not an int value.
- Only original tweets are useful, the records with non-null in_reply_to_status_id and retweeted_status_id are retweets and replys.
- Columns related to reply and retweet are not necessary and most of the values in these columns are NaN.
- The timestamp and retweeted_status_timestamp's data types should be datetime, not string.
- Many dog names are 'None', should be changed into NaN.
- The tweets' texts contain links and rating scores, which should be removed.
- Record No.516 rating missing.
- Record No.1068 rating should be 14/10
- Record No.1165 rating should be 13/10
- Record No.1202 rating should be 11/10
- Record No.1662 rating should be 10/10
- Record No.2335 rating should be 9/10

image_predictions table:

- Tweet id should be a string, not an int value.
- Some dog breeds' names are capitalized, some are not.

tweet_info table:

- Tweet id should be a string, not an int value.
- Favorite and retweet counts should be int values, not strings.

3.2 Tidiness Issues

- The rating_numerator and rating_denominator columns in twitter_archive can be combined into a rating column.
- The three tables can be merged into one table.
- For image_predictions table, only columns tweet_id, p1, p1_conf and p1_dog are useful.

4. Clean

After we identified the quality and tidiness issues, we started to clean the dataset. First, we made copies of all the tables and made all the modifications on these copies. Then we define each issue by indicating what we will do to fix the issue. After that, we used codes to clean the dataset programmatically and also did tests to see if the issue has been fixed.

5. Conclusion

Data wrangling is a useful skill that makes the data easier to be analyzed and visualized. It's very important to follow the data wrangling steps and understand the definition of 'quality' and 'tidiness' in order to get a high quality dataset.