Wrangle Report

This report includes a brief overview of the effort put in to wrangle data for this project. The dataset used in this project was obtained from the Twitter archive of user @dog_rates. WeRateDogs is a Twitter account that rates and comments humorously on people's dogs. Almost always, the denominator of these ratings is ten. The numerators are almost always greater than 10. The data was sent by the @dog_rates user to be assessed, cleaned, and analysed.

The steps taken to construct this project are:

1. Gathering data
2. Assessing data
3. Cleaning data
4. Storing data
5. Analysing, and visualizing data

1. Gathering Data:

Data was obtained from three different resources: the Enhanced Twitter Archive, Twitter API data, and the Image Predictions File.

A) Enhanced Twitter Archive:

The WeRateDogs Twitter archive provides basic tweet information for every 5000 tweets. For example, tweet text was used to extract the rating, dog name, and dog stage.

B) Twitter API data:

This data was gathered from Twitter's API. Twitter's API contains the necessary data, which are the tweet id, retweet, and favorite. These data are necessary for the analysis, so they work here as a compulsory complement to Twitter’s archived data.

C) Image Predictions File:

This data was derived from a neural network that can classify breeds of dogs. The image predictions file contains key columns, including image predictions, tweet id, image URL, and image number.

2. Assessing data:

The assessment process is conducted in two ways: visual assessment and programmatic assessment.

- For visual assessment, the data was displayed on the notebook, and the columns were discovered and briefly explained for enhanced readability.
- For programmatic assessment, Pandas' functions and/or coding methods are used to assess the data. For example, the following functions were used to assess the data:

- .info()
- .describe()
- .isnull()
- .duplicated()

3. **Cleaning data**: Several tidiness and quality issues were discovered during the assessing process, including:

A) Tidiness issues (3 issues):

1. Columns in_reply_to_status_id and in_reply_to_user_id are not necessary for later analysis.
2. The four dog stages should be in one column.
3. Data frames will be merged according to the necessity of the analysis.

B) Quality issues:

1. Columns for the master data frame are not clear and descriptive.

2. There are 181 rows that are not original tweets; instead, they are retweets. the given instructions require to analyse only the original retweets.

3. For the prediction images, there are rows that did not predict a dog (given by "false" for three predictions)

4. Names of the breads in Breed_Prediction_1, Breed_Prediction_2, and Breed_Prediction_3 include "_". This need to be removed for good representation and consistency.

5. Some breeds names are not consistency (some are upper case and other are lower case).

6. some names are missing or invalid names such as None or a in one of the rows.

7. Some dog names in column Dog_name is not capitalized.

8. the column Tweet_ID it should be object because no calculations are required for this column.

9. rating_denominator has values less than 10. It should be 10.

10. rating_denominator has values more than 10. it should be 10.

11. There is a rating of 26, but this is not true. Based on the tweet text, the rating is 11.26/10, not 26. Also, the tweet with tweet id (883482846933004288) the rating is 5, but the tweet text says 13.5. The rest of rating of 5 correspond with the tweet text that rated dogs less than 10, so those ratings will be deleted in Issue #14.

12. rating_numerator max value is 1770. It seems an outlier that needs to be fixed (clean).

13. There is a rating of 27, but this is not true. Based on the tweet text, the rating is 11.27/10, not 27. Also, the tweet with tweet id(883482846933004288) the rating is 5, but the tweet text says 13.5. The rest of rating of 5 correspond with the tweet text that rated dogs less than 10, so those ratings will be deleted in Issue #14.

14. rating_numerator min value is 0. it should be more than 10.

4. Storing data:

After cleaning, all of the data was saved in a file called twitter_archive_master.csv.