Wrangle report

I have wrangled the WeRateDogs Twitter data in order to create analyses and visualizations. The process consisted of me using three different datasets. The first one was the provided dataset *twitter-archive-enhanced*, which contains tweets from the WeRateDogs group on Twitter. The second dataset was programmatically downloaded using the requests package and a provided url. This dataset contains machine learning predictions on what is actually seen in the tweeted pictures. The third dataset was downloaded, using a Twitter API package *Tweepy* and it was necessary for me to create a developer account with Twitter to get access to the data. The tweet id's from the first dataset was used to download more information on these, using the API.

All three datasets were assessed for quality and tidiness errors. Several were identified and a list of these were made before getting into cleaning. In the assess phase I used both visual and programmatic assessment. The visual assessment mainly uses the .head and .sample functions to get an overview of the content. The programmatic assessment uses the .info, value_counts and .describe functions to investigate the datasets. Several problems were identified and some were significant for the analysis process. E.g. the rating numerators seemed to be wrong when compared to the associated tweet text and the index in the downloaded API file did also need to be changed.

An overview of all the identified issues are listed here:

Quality

- Provided dataset
    - There is HTML code in the *source* column. The HTML code should be removed if the source needs to be interpreted
    - Not all URLs in the *source* column are refering to twitter. There are examples of gofundme and vin.co references.
    - The column *rating_numerator* has values below 10 which is not adhering to the general idea. It should be 10 or above.
    - The *rating_denominator* has values different from 10, with a range between 0-170. Generally the idea is that this should be 10.
    - Index 1002 has the dog name set to *a* and the numerator to *8*. The name does not seem realistic and the numerator is supposed to be 10.
    - The *name* column is in several instances missing a dog name.
    - Some fields within the *expanded_urls* seems to be missing.
    - The algorithm to extract grades into the *rating_numerator* column seems to be wrong.
    - Retweets will be removed otherwise tweets will be present several times in the dataset. The column *retweeted_status_id* will be used to identify these.
    - All replies to tweets will also be removed otherwise data will be present several times. The column *in_reply_to_status_id* will be used to identify these.
- Machine learning dataset
    - None found
- Downloaded from Twitter dataset
    - There is HTML code in the *source* column. The HTML code should be removed if the source needs to be interpreted.
    - The columns *contributors*, *coordinates* and *geo* have zero entries.
    - The column *place* has 1 entry.

Tidiness

- Provided dataset
  - The columns *in_reply_to_status_id*, *retweet_status_id*, *retweeted_status_user_id* all seems to be represented in scientific notation as the type is float64. If used they could be converted to int64.
  - The dog rating is included in both the *text* and *rating_numerator* columns.
  - The columns *timestamp*, *retweeted_status_timestamp* are both objects and should be converted to datetime if needed for analysis.
  - The columns *doggo*, *floofer*, *pupper* and *puppo* will be merged into one column
- Machine learning dataset
  - None found
- Downloaded from Twitter dataset
  - The column with tweet id is named *id* and should be renamed to *tweet_id*.
  - The first column has been read wrong.
  - Another index should be used.
  - The columns *id* and *id_str* are similar.
  - Drop some columns as per project description.
  - The columns *in_reply_to_status_id*, *in_reply_to_status_id_str*, *quoted_status_id*, *quoted_status_str* are all float64 and should be converted to int64

The cleaning of the three datasets were based on the findings in the assess phase. A significant part of the columns in the three datasets were deleted as part of the cleaning process as they either did not have any values or they did not provide any data or information valid for analysis. Any problems identified with these columns in the assess phase were not taken care of as they would be removed. The identified problems were programmatically fixed one dataframe at a time. After cleaning all the datasets they were merged using an inner join to only use the tweets with data in all three datasets. This clean dataframe was then saved into a cleaned .csv file which can be used for analysis and visualization.