Data Wrangling report

Project objectives:

- Perform data wrangling gathering data from three different sources
- assessing those datasets
- cleaning
- Analyze and visualize the information deducted from this wrangled data.

Step 1: Data Gathering

WeRateDogs-related data is gathered from three different sources:
1- WeRateDogs Twitter archive: a CSV file containing basic information (tweet_id, timeStamp, text,..) about 5 000 tweets; downloaded manually using the pandas' library.
2- Supplementary data gathered via Twitter API using Tweepy library thus opened through a JSON file.
3- Tweet image prediction file 'image-predictions.tsv': a tsv file containing predictions about the dog races based on their images; downloaded using the Requests library.

Step 2: Data Assessing

Once gathered, the data get assessed visually and programmatically.
Various quality and tidiness issues get found:

Quality issues

1. the wrong data type for the column timeStamp
2. links of the column 'source' are contained in an HTML anchor tag ()
3. useless informations: columns in_reply_to_status_id , retweeted_status_user_id, retweeted_status_timestamp and lines corresponding to retweets
4. different values for the denominator, but it should always be equal to 10

5. There are dogs with no classification
6. There are duplicated images
7. unclear column names
8. The column 'id' name is incompatible with its equivalent in the other tables.

Tidiness issues

1. Doggo, Puppo, Floofer, and Pupper columns are hard to extract information from them in the archive table
2. The three tables don't have the same number of entries (which causes a problem while merging them)

Step 3: Data Cleaning

Once assessed, the identified issues in the datasets got cleaned.

1. Merge the three datasets in one “Dogs_data”
2. Drop the retweets lines as we only need the original tweets in our analysis process.
3. Drop the unnecessary columns
4. Rename some of the conserved columns to be more meaningful
5. Correct the data type in the ‘timestamp’ column from object to date
6. Drop lines with uncorrect denominator (not equal to 10), to assure the efficiency of comparing ratings
7. Drop rows where dogs have no classification
8. Drop duplicated images
9. Correct the format of ‘source’ column by extracting the needed information from the anchor tag
10. Merge the four columns ‘puppo’ ‘pupper’ ‘doggo’ and ‘floofer’ being different values for one variable ‘classification’

Thus, we obtained one clean dataset: Dogs_data that we stored in a csv file called 'twitter_archive_master.csv'.