In this project I wrangle data of the WeRateDogs twitter.

I first gather three pieces of data: 1) manually downloaded ‘The WeRateDogs Twitter archive’, and load the data with ‘read_csv’; 2) downloaded ‘The tweet image predictions’ using the Requests library, and read the tsp file using ‘read_csv’; 3) queried the Twitter API for each tweets using Python’s Tweepy library.

Then I visually and programmatically accessed the data, and found 11 quality issues and 3 tidiness issues:

quality issues:¶

1. Some dogs have abnormal names like a, an, the, such, quite and etc.
2. Missing names should be NaN instead of string 'None'.
3. Rating denominator is not always 10.
4. Rating numerator is not accurate.
5. Retweets need to be removed.
6. Dog stages are not accurate.
7. There are 2075 images predicted but 2356 tweets in twitter_archive_enhanced.
8. Dog stages columns should be of boolean format.
9. datatype of tweet_id is int,it should be str
10. Timestamp should be date-time format.
11. Some tweets are not rating dogs.

tidiness issues:¶

1. Dog stages should be one column.
2. Unnecessary columns which is not useful for analysis.
3. All three pieces of data can be merged into one dataframe.

Next, I define 12 solutions to solve the issues that I found during data accessing:

1. change column 'id' and 'created_at' into 'tweet_id' and 'timestamp' in tweet_api.
2. Merge 3 dataframe into one, and Keep tweets with image prediction
3. change datatype of tweet_id into str.
4. Drop retweets
5. change dog stages into one column.
6. drop unnecessary columns.
7. change datatype of timestamp into data-time.
8. correct rating demonimator.
9. correct rating numerator.
10. correct dog names. And use 'NaN' when names are missing.

Then I made codes to solve each issues and constantly check if everything is correctly and on the right track.

Finally, I got two dataframes and saved them as csv files.
The two dataframes are ‘tweet_dog’ containing dog rating of the dog and their image prediction; ‘tweet_info’ containing information of each tweet, i.e. text, favourite count, and etc.