[Sept 15th 2020]

[Wrangle report] by Amira Saad

Summary

This Report is about wrangling efforts done in the datasets, WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog and has over 8 million followers also has received international media coverage, these data wrangling include:

- Gathering data (from three different sources)
- Assessing data
- Cleaning data

Gathering data

The data is divided into three different files from three different sources:

1- The enhanced twitter archive
2- The image prediction file
3- Twitter JSON file (tweet_json.txt)

Assessing and Cleaning

Enhanced Twitter Archive

Is the first CSV file which was downloaded manually, after assessing this file visually by Excel and programmatically the following issues were noticed and cleaned:

No	Data Assessing and Cleaning description
A-Quality Issues:	
1	Changed the Timestamp column to be a Datetime type instead of object
2	I found that the rating numbers have some incorrect values in the denominator and numerator columns, some rating have unrealistic values
3	As per requirements in the project, original ratings (no retweets) that have images are to be included in the wrangling process, so I removed all retweets in the twitter archive
4	Dropped all un-necessary columns that include ('source', 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp') and most of these columns have null values.
5	Dog stage columns have 'none' values, I changed those to nan

6	Also the names of the dogs are not extracted accurately e.g (a), also some of the values in the name column are missing i.e. 'None'
B- Tidiness issues	
1	According to tidiness standards, dog "stage" columns (i.e. doggo, floofer, pupper, and puppo) should be in one column with name 'stage' , so I used str.cat to merge them together.

Image Prediction file:

Was downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

the following issues were noticed and cleaned:

No	Data Assessing and Cleaning description
A-Quality Issues:	
1	Removed all retweets also from the image prediction file to include the same values as in twitter archive file
2	P1 & P1_config columns should have indicative names (p1 dog breeds the algorithm's #1 prediction for the image in the tweet, p1_conf is how confident the algorithm is in its #1 prediction) and the rest of the columns P2 and P3, so I changed the names of the columns

Note: I did not merge the image prediction file with the original twitter archive, because image prediction table I think has different observational unit.

Twitter API file:

Extracted from the file called tweet_json.txt. Each tweet's JSON data was written to its own line. And was read line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count, the file had the following issues which was cleaned:

No	Data Assessing and Cleaning description
A-Quality Issues:	
1	Changed column name 'id' into 'tweet_id' for merging purpose
B- Tidiness issues	
1	Merged twitter API file with twitter archive file, because they both contain relevant data the tweet id, favorite counts, and retweets count.

Finally, Storing the data:

The data was stored into 2 files **:**

`twitter_archive_master.csv`

`img_pred_master.csv`