We_Rate_Dogs Twitter Archive Wrangle Report

This short report describes the wrangling efforts involved in completing the "WeRateDogs" project as part of Udacity's Data Analysis Nanodegree. The Data Wrangling process consists of:

1.Gathering data 2. Assessing and Cleaning data 3. Conclusion

Gathering :

I gathered data from 3 sources, stored in separate files:

1. WeRateDogs Twitter Enhanced archive, manually downloaded from the Udacity servers.
2. The image predictions file, programmatically downloaded from the Udacity servers.
3. The entire set of each tweets' JSON data, downloaded by querying the Twitter API using the Tweepy library , so Becuose we can't set up a Twitter developer account I using the two files: twitter_api.py and tweet_json.txt The favourite_count and retweet_count were extracted programmatically from" tweet_json" file.

Assessing and Cleaning :

1. twitter_arch table:

- the datatype of the id - columns is integer and should be str
- the datatype of the timestamp - column is object and should be datetime
- some of the dogs are not classified as one of "doggo", "floofer", "pupper" or "puppo" and contain all "None" instead
- some of the dog names are not correct (None, an, by, a, ...)contains retweets
- some of the ratings are not correctly extracted (mostly if there are >1 entries with the pattern "(\d+(.\d+)?\/\d+(.\d+)?)"

- also transforming the ratings to integer created some mistakes (there are also floats)the source column contains html code

2-imge_prediction table:

- the datatype of the id - columns is integer and should be str
- contains retweets (duplicated rows in column jpg_url)
- there are pictures in this table that are not dogs
- the predictions are sometimes uppercase, sometimes lowercase
- also there is a "_" instead of a whitespace in the predictions

3-df_api table

- the datatype of the id - columns is integer and should be str

Tidiness

1-df_twitter table:

- he columns doggo, floofer,pupper and puppo are not easy to analyze and should be in one column

  2-df_predict table:

- the prediction and confidence columns should be reduced to two columns - one for the prediction with the highest confidence (dog)

  3- df_api table :

- display_text_range contains 2 variables

conclusion:

All three tables share the column tweet_id and should be merged together.

Data Cleaning:

1. Merge the tables together
2. Drop the replies, retweets and the corresponding columns and also drop the tweets without an image or with images which don't display doggos
3. Clean the datatypes of the columns

4. Clean the wrong numerators - the floats on the one hand (replacement), the ones with multiple occurence of the pattern on the other (drop)
5. Extract the source from html code
6. Split the text range into two separate columns
7. Remove the "None" out of the doggo, floofer, pupper and puppo column and merge them into one column
8. Remove the wrong names of name column
9. Reduce the prediction columns into two - breed and conf
10. Clean the new breed column by replacing the "_" with a whitespace and make them all lowercase

Conclusion :

Data wrangling provides a clean data frame for future analysis and visualization, in our case we concluded with the “twitter_archive_master.csv’. This file can also be shared with others without having to wrangle the data.