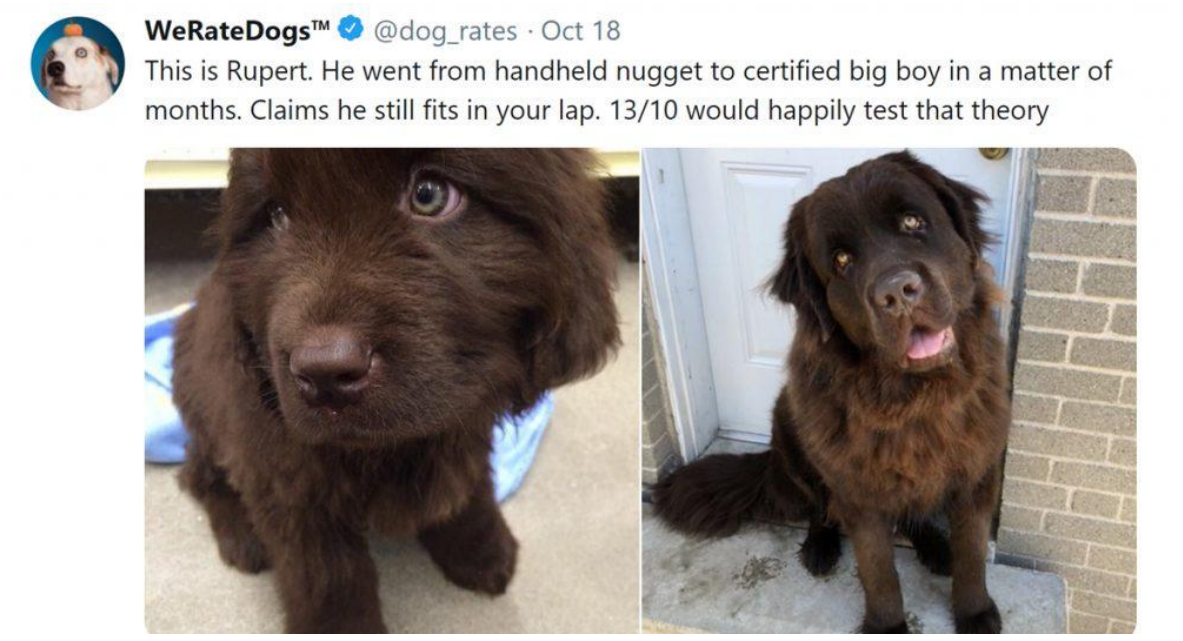
We Rate Dogs Report

Name : Islam Ashraf Muhammed

Email : es-eslamashraf@alexu.edu.eg



This report is about the code done about the three steps of data wrangling done in the project which are:

- Gathering
- Assissing
- Cleaning

I will walk you through the three steps briefly.

Gathering:

The data used in this project gathered from three places, the first one was a csv file called “twitter-archive-enhanced” downloaded from the project page.
The second one called “image-predictions” was downloaded programmatically from udacity url.
And Finally the third one was scrapped from twitter API.

Assissing:

Assissing For first dataframe:
Quality Issues:
- Missing Values of Calssifications Like doggo, floofer, pupper, puppo.
- The missing values are written as none and not NaN.
- Wrong names like the name a.
- None values for names.

- Some rating numerators are Wrong, some are very large and some are below 10.
- Three ids like twiteer id should be strings.
- Dogs classifications like doggo and floofer should be categories.
- some rating denominators does not equal 10.
- data with no images should be deleted (Expanded URL = NaN)
- Deleting the rows that retweets and replies have values in it.*

Tideness Issues:

- Deleting columns of retweets and replies and source and expanded url.
- There are four columns and it should be only one named classification.
- merge three data frames into one.

These were the main points I found in the first file. These problems were found visually and programmatically using .info() and .value_counts() functions.

Sure there are other problems but I did not write them for the shortage of time.

Cleaning:

- Delet the IDs that are in first dataframe and not in second one.

As the data frames will be combined we need first to know the IDs common in both files.

- Scrap missing names from text if not there make it as NaN.

As there was some missing names and wrong ones I tried to find them in text if not there, I replaced them with NaN as they are easy to handle.

- Delet rows that has values in retweeted_status_idcolumns
- Delet rows that have no values in expanded_urls.

As the retweeted are not needed so we deleted them.

- Replace all None Values in the Doggo, floofer ,.. columns with empty character
- make all columns as one column called dog_class
- replace all empty character with NaN values

I combined the dog Classes into one Column.

- convert the ratings to float
- get the number of dogs
- calculate the average

I got the average number of dogs in photos and divided the numerator by this number to calculate the average.
- Making all IDs alike in the three dataframes
Making all rows for the three data frames alike.
- Resetting all indexes
- Merge three DFs to one big DF
- Convert tweet_ID to string
- change time to datetime.
P.S: I apologize for the poor quality of coding as I have exams and I did not have too much time.