WeRateDogs Tweets

Hanin Falatah | Nanodegree Data Analyst | 12/30/19

Introduction

The dataset that I will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also wikipedia known as in WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because they're good dogs Brent. WeRateDogs has over 4 million followers and has received international media coverage.

> WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively for you to use in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.

Data Gathering

The data that need gathering are the below:

1. The WeRateDogs Twitter archive, the csv file of tweet archived that was given to us from the Udacity from @dog_rates twitter-archive-enhanced.csv.
2. The tweet image predictions that is provided by Udacity and the image-predictions.tsv as requested to be downloaded programmatically from the URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
3. The Twitter API the requests for the json text file is rejected I did upload to Jupyter, that is provided by Udacity, then read the JSON data from the txt file to extract tweet id, favorite count, and retweet count and put that data in DataFrame.

Data Assessing

Assessing all the gathered data for issues in the tidiness and quality.

Tidiness Issues

1. The df_archive_tweet dataframe the columns doggo, floofer, pupper, and puppo should all be one column called stage.
2. All the dataframe should be in one dataframe, the df_tweet_data, df_predict, and df_archive_tweet should be one datafram.

Quality Issues

1. The source contexts should be easy to read.
2. The df_archive_tweet dataframe contain of the wrong data, for the following columns: tweet_id, in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and timestamp.
3. In the df_predict and df_tweet_data dataframe the tweet_id wrong data type.

4. The retweeted tweets and the replying on other tweet are not needed the needed are only The Original Ratings tweets.
5. The names for the dogs couldn't be. (a, the, an, ...) as in the table.
6. In the df_archive_tweet some of tweets with rating less than 10 does not have images.

Data Cleaning

Here the issues that were discovered in the assessing step were fixed here, the work in the data cleaning is through:

1. Define: defining the assessing issues one by one.
2. Code: put the definition to code.
3. Test: testing that the code worked in cleaning the issues.

In cleaning I started with tidiness issues the is melting the columns of dog's stages into the one column called stage, then merging the DataFrams df_archive_tweet, the df_predict, the df_tweet_data into one DataFram, then the quality issues of the source contexts should be easy to read, then the converting the wrong data types of df_archive_tweet dataframe, for the following columns: tweet_id, in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and timestamp, then removing the tweets except the Original Ratings tweets, then fixing the dogs names.

Then the other cleaning, cleaned the tweets with rating less than 10 or the prediction 1, and 2 say that the picture does not contain a dog, cleaning the none and the two values in stages

The DataFram after all that is ready for visualiztion