REPORT ON DATA WRANGLING

16.03.2021

— BY

OKEBIORUN SARAH

Overview

This project involves wrangling data obtained from three sources, all of which relate to the famous WeRateDogs (@dog_rates) Twitter account. WeRateDogs is a Twitter account that tweets images of dogs their owners send in, along with a funny caption and a rating that almost always exceeds 10/10

Goals

1. Gathering our data
2. Accessing our data
3. Cleaning our data
4. Storing, analysis and visualizing our data

Gathering our data

The thing I did to gather my data are:

1. The WeRateDogs Twitter archive file was giving to me by Udacity.
2. The tweet image predictions, i.e., what breed of dog (or other objects, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers for me to download programmatically.
3. I also went to my anaconda to install 'tweepy' library.

Accessing my data

1. QUALITY ISSUES

a. Twitter-archive table:

   1. The timestamp should convert to timestamp datatype.
   2. Columns such as in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id ,retweeted_status_timestamp should be removed
   3. tweet_ id should convert to str.

4. Dog names: some dogs have 'None' as a name, or 'a', or 'an.'

b. Image_predictions.tsv:

5. p1, p2, and p3 columns aren't consistent when it comes to capitalization: sometimes the dog breed listed is all lowercase, sometimes it is written in Sentence Case
6. p1, p2 and p3 columns have invalid data...why would the algorithm labeled a dog photo as a starfish, boathouse, or mailbox

c. tweet_json:

7. Missing Some Data
8. favorite and retweets columns should convert to int datatype

2. TIDDINESS ISSUES

a. twitter-archive-enhanced-2.csv:

1. The last four columns all relate to the same variable (dogoo, floofer,pupper, puppo).

b. Tweet_json:

2. This data set is also part of the same observational unit - one table with all basic information about the dog ratings.

Cleaning my data

I did the following in cleaning my data:

1. Removing columns that are no longer needed
2. Merge the clean versions of df_twitter, image_predictions, and tweet_json dataframes Correct the dog types
3. Delete retweets
4. Creating one column for the various dog types: doggo, floofer, pupper, puppo Remove columns no longer needed.
5. Change tweet_id to string from integer
6. Timestamp to correct datetime format
7. Naming issues
8. Creating a new dog_breed column using the image prediction data