Sahand Niasti

Data wrangling report

Introduction

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. The account was started in 2015 by college student Matt Nelson, and has received international media attention both for its popularity and for the attention drawn to social media copyright law when it was suspended by Twitter for breaking these aforementioned laws [Wikipedia]. Our main goal here is to separate a portion of these tweets to wrangle and analyze data. For this purpose, WeRateDogs led Udacity to have access to its archive tweets with some basic data (such as tweet id, timestamp, replies, retweets and etc.).

As I said before, our sole aim here is to practice wrangling data with WeRateDogs Twitter data and create enjoyable and meaningful analyzing at the end of the report. There are different steps toward us during the project that will be introduced and explain thoroughly.

Steps

Data Gathering: first step of data wrangling, aiming to gather data from 3 different sources here.
Data Assessment: dive deep into tables and data to find any anomaly or misinterpretation in data.
Data Cleaning: carry out three main step (define, code, test) of cleaning data to be ready for analysis.

Gathering Data

Three main sources of data were required to be downloaded for this project, which will be introduced below.

- Twitter Archive: The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets, in a specific period of time. This was the building block of our work but not enough.
- Image Prediction: The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is trained in each tweet according to a neural network. Udacity's servers has provided links for this file and students should download the file programmatically using the Requests library.
- Twitter API Data: Twitter give us access to its database through api's.So, the tweet id's required to be downloaded was selected and requested from twitter api. This file was saved as a text called "tweet-json". Now, to extract required data, each tweet's JSON data should be written to its own line. Then read this .txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count.

Data Assessment

Each table is assessed separately based on tidiness and quality of its data. Below, each step will be explained.

Twitter Archive

❖ Quality

- Dog names are not complete and there are some unrelated words(a,...)
- contains retweets
- the source column contains html code
- data type of timestamp and other columns need to be changed
- some of the dogs are not classified as one of "doggo", "floofer", "pupper" or "puppo" and contain all "None" instead
- some of the ratings are not correctly extracted

❖ Tidiness

- One column for Dog classification(doggo, floofer, pupper or puppo)
- unecessary source column with hardly readable information

Image Prediction

❖ Quality

- 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

❖ Tidiness

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

I could not find any problem with Twitter api table because I somehow generate this table on my own.

Data Cleaning

Cleaning of this data was done in 8 consecutive steps.

1. Merging Tables
2. Dropping unnecessary columns
3. Cleaning data types
4. Extracting the source from html code
5. Removing the "None" out of the doggo, floofer, pupper and puppo column and merge them into one column
6. Removing the wrong names of name column
7. Summarizing the prediction columns into breed and conf
8. making dog breed column lowercase