wrangle_report

April 16, 2020

1 Wrangle Report

1.1 Introduction

I was given the task through Udacity Data Analyst Nanodegree program to wrangle data from Twitters WeRateDogs. In this report I will describe my wrangling efforts and the issues I ran into.

1.2 Gathering

Gathering is the first step in the process. The data for this project consisted of three pieces of data.

The WeRateDogs Twitter Archive that was provided to me through Udacity. I manually downloaded it to my computer my clicking the link provided to me. twitter-archive-enhanced.csv

The tweet-image-predictions file that was provided to us on a neural network This file (image_predictions.tsv) is hosted on Udacity's servers and was downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

I was unable to utilize Twitter's API to access the final tweet data. I utilized the shortcut provided to me the tweet_json.txt file. Then read the tweet_json.txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count.

1.3 Accessing

Accessing is the second step in this process and after I finished gathering all the data, I was able to visually and programmatically look for quality and tidiness issues.

1.3.1 Quality

twitter dataset

- 59 null values in the expanded_url column
- timestamp should be in a datetime format.
- In several columns null objects are non-null (None to NaN).
- Name column has invalid names i.e 'None', 'a', 'an'.
- We only want original ratings (no retweets) that have images.

- Some columns are in reply to a tweet and can be dropped since they are not needed. (in_reply_to_status_id, in_reply_to_user_id)
- Text column is not displaying full text

image dataset

- Missing values from images dataset (2075 rows instead of 2356)
- jpg_url has 66 duplications
- p1,p2,p3 have inconsistant capitalization and names as None

tweet dataset

- tweet id is labeled as 'id'

1.3.2 Tidiness

- Four columns: doggo, floofer, pupper, puppo when we only need one for general dog_stage
- p1,p2,p3 can be combined into one prediction column
- p1_conf,p2_conf,p3_conf can be combined into one confidence column
- Join 'image_clean' and 'tweet_clean' to 'twitter_clean'

1.4 Cleaning

Cleaning the data is the third process in the data wrangling process. This step is where I fixed all the quality and tidiness issues I described in the accessing step. I started off by making copies of the original data frames. I would Define, Code, and Test each problem and ensure the corrections were made. Initially I did not find all the quality and tidiness issues I stated and had to iterate back to the accessing step. Not all the Quality and Tidiness issues were found only enough to suffice for the project. After it was fixed I saved the new data frame.

```
In [ ]:
```