wrangle_report

October 11, 2022

0.1 Reporting: wragle_report

0.1.1 Introduction

This report demonstrates the data wrangle and analysis process for WeRateDogs. My goal for this project is to demonstarate the data analysis process, work flow, and to create interesting and trustworthy analyses and visualizations. In this report I will provide a breaf introduction of the data wrangle process that is used to gether, access, clean, and analyzed the these datasets.

Dataset description The dataset that we will be working with is the tweet archive of Twitter user @dog_rates, also known as 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

0.1.2 Data Gathering

Gethering Data, three pieces of datasets was required in these section. Each file is gethered a differen method as follows

The `Twitter_archive_enhanced.csv` is directly downloaded from the WeRateDogs Twitter archive data. this file contains basic tweet data for all 5000+ of their tweets.

The `Tweet image prediction (image_predictions.tsv)` is downloaded from Udacity using python request library and the URl provided.

The `Tweet_json.txt` I downloaded Tweeter API using the Tweepy library and my own twitter developer credentials.

0.1.3 Assess Data

After obtaining all the required datasets, I moved onto the next step of the data wrangle which is 'Assessing data'. In this step my task was to act as detective and inspect the data quality issues and lack of tidiness using python pandas library to evaluate the data visually and programatically assesments. Upon exploring the data the following quality and and data tidness were observed.

0.1.4 Quality issues:

Inhanced twitter archived table `Enhanced twitter archive table` 1. Delete retweets and replies and keep the original ratings, 2. The `timestamp` The timestamp column is in incorrect datatype format 3. The column `Name` has missing values & invalid names 4. the `source` column has a useless html structure that need to be fixed. 5. in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id are float, should all be str

Prediction table

6. p1_dog, p2_dog, & p3_dog are not all lowercase.
7. Remove entries that have p1_dog, p2_dog, & p3_dog values set to false. they real dogs.
8. Remove duplicate jpg_url entrie

tweet_json table

8. Convert data type of tweet_id to object string data type for merging

Tidiness

1. twitter_archive: doggo, floofer, pupper, puppo are all stages of dog, should be in one column
2. Merge the three datasets into one Master dataset.

0.1.5 Cleaning Data

In this section I have cleaned the quality and tidiness issues I observed above using python and pandas library and combined all the datasets into a master dataset.

0.1.6 Conclusions and limitations

In this report I have beriefly explained the wrangling data analysis steps and analized the dataset at it is best. I have enjoyed going through the wrangling and data analysis i have learned in this course and solving the quality issues and tidiness we have observed inproject.

Limitations Finding the data quality issues, tidiness and cleaning were the hardest and time-consuming part for me, but at the end i was able to overcome this struggle. after cleaning the data and storing the master dataset there was some limitation due to the number missing value we have in dataset, for instance the dog_stage column was missing over 80% tweets don't provide dog stage info in there tweets, so for this analysis may not be 100% guarenteed to be proof error solution.

In [ ]: