Wrangling Report

Introduction

This report briefly describes the work done on this project. The datasets wrangled and analyzed in this project belongs to @dog_rates from their #WeRateDogs results archive.

The project included the following:

- *Data Wrangling process*
    - *Gather*
    - *Assess*
    - *Clean*
- *Storing the wrangled data*
- *Analyzing and visualizing the data's findings and reporting them*

Data Wrangling process

Gather

I managed to gather the following datasets

- *Twitter_archive_enhanced.csv* was provided by Udacity for all students enrolled. Containing all tweets from 2015 till 2017.
- The tweet image predictions file was pulled off the provided link using the Requests functionality to save it as a TSV file format (and then reading it through pandas . and the name of the saved file was *image_predictions.tsv*.
- Finally, the JSON file was downloaded from the project's page as provided, because I could not establish my dev account for Twitter in time, this part will be tackled later. The file was tweet-json.txt. I opened the file on the Jupyter notebook and converted the json list import into a DataFrame named *json_tweets*.

Assess

After the gathering process, I started assessing the data.

Quality Issues (9)

tweet dataset

- retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp show 181 records, this means that these records are retweets from our original tweets gathered (i.e. a duplicate of the original tweet).
- in_reply_to_status_id, in_reply_to_user_id have many null data, also the source column not needed.
- timestamp format is an object not a datetime.

- rating_numerator, rating_denominator both have illogical values. The numerator has very high values up to 1776, while there are rows that have the denominator more than 10 (which is not correct according to the info in hand).
- rating_numerator, rating_denominator should be combined as the twitter rating, this also means that these data are categorial.
- Records have invalid names; they are easily detected as they start with a lower-case letter. These naming should be replaced with a more appropriate description as 'None'.
- The 4 variables should be formatted as a category (This is related to the first tidiness issue also) and making sure so after combining them into one.

image_predictions dataset

- Inconsistency in the 3 breed predictors in terms of casing.
- Some breed naming is invalid, as they contain naming like 'desk' or 'web_site'.

Tidiness issues (2)

- Variables doggo, floofer, pupper, puppo should be in one column.
- The *json_tweets* dataset can be merged to the *tweet* dataset. Same goes for the *image_predictions* dataset.

Clean

After the assessment of my datasets, I executed the cleaning process. The first thing is that I copied all 3 datasets to a new one with a suffix '_clean'. The cleaning process structure consists of Define, Clean, Test for each assessed issue found and can be seen clearly in the notebook, with supported links.

The final look of the combined dataset (combining all 3 together) was named *twitter_archive_master*.

Storing the wrangled data

I stored the data on 2 parts, the first is to store the 3 cleaned datasets into an SQLite database with their respective names, and the combined data as *twitter_archive_master.csv*.

Analyzing and visualizing the data's findings and reporting them

The final piece of the whole project was to provide an Analysis of the findings in this dataset and publish a report about it. I presented 3 insights and 2 visualizations with the following titles

How well did the image predictor do? (insight)

Who is twitter's popular dog? (insight)

Which dog should you buy? (insight)

How generous is @dog_rates in their ratings? (Visualization)

Is there a corelation between the retweets and favorite counts (Visualization)