Wrangle report for We Rate Dogs

Environment and Tools

The data wrangling process is performed in the Jupyter Notebook. The libraries used in this project are pandas, requests, tweepy, json, default_timer, matplotlib.pyplot, seaborn and so on. %matplotlib inline is added for direct outputs in the notebook.

Data Gathering

The datasets for this projects are from the tweet archive of Twitter user @dog_rates (WeRateDogs).

1. Enhanced Twitter Archive: contains tweet data for all 5000+. Only 2356 records have ratings.
   File name: twitter-archive-enhanced
   Format: csv
   Source: directly download from Udacity website.

2. Image Predictions File: the output from neural network
   File name: image-predictions
   Format: tsv
   Source: get the data from url =
   'https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv'

3. Additional Data via the Twitter API
   File name: tweet_json
   Format: txt
   Source: connect Twitter API to download json format text file and use pandas to read into the notebook.

Data Accessing and Cleaning

Quality

Enhanced Twitter Archive table:

1. Text column has the link for the tweets and ratings at the end.
2. Doggo, floofer, pupper and puppo columns has Non for missing values.
3. Timestamp column is string instead of datetime data type.
4. We are not interested in the retweet and reply ONLY the tweets are needed.
5. Some rows are not related to dogs.
6. Rating_denomination column has values less than 10 and values more than 10 for ratings more than one dog.

Twitter API table:

1. The id column name was inconsistent with the 2 other datasets

Tidiness

Enhanced Twitter Archive table:

1. Removed the rating score and tweet link from the tweets text column.
2. All the dog stages (doggo, floofer, pupper, puppo) should be transposed into one column, dog_stage. If there are more than one stages, the stages will be concatenated and separated by one space.
3. Removed retweet and reply columns from the data
4. Removed any rows with denominator more than 10.

Image Predictions table:

1. All p1, p2, p3, p1_cof, p2_cof, p3_cof, p1_dog, p2_dog, and p3_dog are all about dog breeds and confidence. Therefore, the columns (p1, p1_dog, p1_conf, ...etc) should be just breed and confidence.

General: All three (3) datasets were merged into one

Result

The analzed dataset was integrated with other data and put in a SQlite database

Storing data

1. Store into SQL database, twitter_archive_master.db
2. Stored the cleaned and analysed data in a .csv files, twitter_archive_master.csv

Summary

The most challenging part of the wrangling process is extracting useful information from each tweet. Regular expression is heavily used in this process. After data wrangling, the datasets are ready for analysis and modelling.