Wrangle Report

Introduction

The purpose of this project is to wrangle, analyze, and visualize tweet archive of Twitter user @dog_rates, also known as WeRateDogs.

Process

Gathering Data

1. The WeRateDogs Twitter archive: Download this file by the following link: twitter_archive_enhanced.csv

2. The twitter image predictions, i.e., what breed of dog (or another object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

3. Twitter API & JSON: Each tweet's retweet count and favorite ("like") count at minimum, and any additional data you find interesting. Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file. 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. Note: do not include your Twitter API keys, secrets, and tokens in your project submission.

Assessing Data

1. twitter_archive_enhanced Quality:
    - Columns have null value (including 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_user_id', 'retweeted_status_timestamp', 'expanded_urls')
    - There are some datatypes errors, for example (timestamp, rating_numerator, rating_denominator)
    - Bad source format, not easy to read.
    - rating_numerator and rating_denmoinator columns have invalid values
    - name column has invalid values. For example, a, Mo,Bo).
    - Keep original tweets.
2. twitter_archive_enhanced Tidiness:
    - Only need one column for Dog stage inside of 4.
3. image_predictions Quality:
    - Too many predictions. p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3, p3_conf, p3_dog should be merged into two columns (picture_predictions & confidence_level)
    - Some tweet_ids have the same jpg_url
4. image_predictions Tidiness:
    - Compare p1, p2, and p3 predictions and create a new column save results
    - Merge 'tweet_info' and 'image_predictions' into 'twitter_archive'.
5. tweet_json Quality:
    - Keep original tweets.
    - Bad source format, not easy to read.

Cleaning Data

- According to last two step, I created copy for each of three data frame. I did follow step to get clean data.
- Drop columns have null value (including 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_timestamp', 'expanded_urls')
- Change data type for twitter_archive_enhanced (including timestamp, rating_numerator, and rating_denominator)
- Corrected source format for twitter_archive_enhanced and tweet_json and made them easily to read.
- rating_numerator and rating_denmoinator columns have invalid values. To get correct rating, I use re package to rescan each tweet text and update rating_numerator and rating_denominator columns with right rating.
- name column has invalid values. For example, a, Mo,Bo).
- Keep original tweet from twitter_archive_enhanced.
- Move doggo, floofer, pupper and puppo columns into dogs_stage.
- Merge p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3, p3_conf, p3_dog into two columns (picture_predictions & confidence_level)
- Drop duplicated jpg_url in Image_predictions

- Create two columns (image prediction & confidence level)

Storage

- Merge and save tables into one dataset