Wrangle Report – WeRateDogs Data

1. Introduction

This project was written for Udacity's Data Analyst nanodegree program. The primary goal is to wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. This dataset is the tweet archive (from November 2015 to August 2017) of Twitter user @dog_rates, also known as WeRateDogs, which is an account that rates people's dogs with a fun comment about each pet.

They also deal with this rating system in a humorous way. Even though the rates should be from 1 to 10, they almost always have a denominator of 10, but numerators greater than 10 (e.g., 11/10 or 13/10). That happens because they believe that all dogs deserve at least a 10 and sometimes even more. For each tweet, there is a picture of a dog that is going to be rated. To analyze the WeRateDogs archive, data will be gathered from a variety of sources and in a variety of formats, assessed in its quality and tidiness, then cleaned. This whole process is called data wrangling.

2. Gathering data

There were 3 main sources of data:

1) WeRateDogs Twitter archive (.csv): Udacity provided this file, which was downloaded manually from their website (twitter_archive_enhanced.csv). This dataset contains basic tweet data for all 5000+ of WeRateDogs tweets that were extracted programmatically. Udacity also used the column with each tweet's text to extract rating, dog name, and dog "stage" (i.e., doggo, floofer, pupper, and puppo) to "enhance" this Twitter archive." Out of the 5000+ tweets, they selected only those tweets with ratings (there are 2356).

2) Twitter API and JSON data: additional data from the WeRateDogs archive was gathered using the Twitter API – in special, data about retweet counts and favorite ("like") counts. Using the tweet IDs in the original dataset, the Twitter API was queried for each tweet's JSON data through python's tweepy library. Then, each tweet's set of JSON data was written in a ".txt" file ("tweet_json.txt"), with each tweet's JSON data on its own line. At last, this .txt file was read, line by line, to create a pandas DataFrame.

3) Tweet image prediction (.tsv): this file (image_predictions.tsv) was also provided by Udacity. They ran every image from the WeRateDogs tweets in the archive through a neural network that was able to classify breeds of dogs. This resulted in a table full of image predictions alongside each tweet ID and image URL. The table also provides columns informing: i) how confident the algorithm is in every prediction; ii) whether or not the prediction is a breed of dog (i.e., true or false); iii) and the image number that corresponded to the most confident prediction (numbered 1 to 4 since tweets can have up to four images). The "image_predictions.tsv" file was hosted on Udacity's servers and was downloaded programmatically using the Requests library from the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

3. Assessing data

Once all the data was gathered, the three datasets were assessed. The goal was to inspect them for data quality issues and lack of tidinesst. This project requires to assess and clean at least 8 quality issues and at least 2 tidiness issues. Therefore, not all issues were adressed. In this process, the assessment was done both visually and programatically. The following tidiness and quality issues were identified:

Tidiness issues:

i) There are 4 columns (doggo, floffer, pupper and puppo) in twitter_archive refering to the dog's stage;

ii) Merge the three tables through the "tweet_id" columns;

Quality issues:

- twitter_archive:

i) The "timestamp" column should be converted to datetime64 datatype;

ii) There are multiple dog stages classifications for the same row;

iii) It is necessary to remove the retweets;

iv) It is necessary to remove columns related to retweets: "in_reply_to_status_id", "in_reply_to_user_id", "retweeted_status_id", "retweeted_status_user_id", "retweeted_status_timestamp";

v) The "name" column has dog's names that were wrongly extracted;

vi) There are wrong values extracted for ratings;

- tweet_json:

vii) The "created_at" column should be converted to datetime64 datatype;

viii) It is necessary to remove the retweets;

- img_pred:

ix) Dog breed names composed by two words are separated by underscores;

x) There are cases of dog breed names in lowercase.

4. Cleaning data

First, all datasets were copied to preserve the original ones. The datasets were cleaned based on the issues above described. This process followed the same ordem for all itens: i) describing the identified issue; ii) coding for solving the issue; and iii) testing the results, checking the desired changes.

5. References

The references used to solve some of the issues described are listed below:

- https://knowledge.udacity.com/questions/242616
- https://knowledge.udacity.com/questions/332090
- https://knowledge.udacity.com/questions/127161
- https://knowledge.udacity.com/questions/386360
- https://knowledge.udacity.com/questions/212738
- https://knowledge.udacity.com/questions/455113
- https://knowledge.udacity.com/questions/389519
- https://knowledge.udacity.com/questions/287523
- https://stackoverflow.com/questions/54611750/dataframe-replace-underscore-with-blank-not-working
- https://www.ti-enxame.com/pt/python/compare-duas-colunas-usando-pandas/1050144517/
- https://stackoverflow.com/questions/47612822/how-to-create-pandas-dataframe-from-twitter-search-api
- https://stackoverflow.com/questions/27900451/convert-tweepy-status-object-into-json
- https://stackoverflow.com/questions/7082345/how-to-set-the-labels-size-on-a-pie-chart-in-python
- https://stackoverflow.com/questions/30765455/why-is-my-plt-savefig-is-not-working
- https://stackoverflow.com/questions/17582137/ipython-notebook-svg-figures-by-default
- https://stackoverflow.com/questions/36622237/jupyter-notebook-inline-plots-as-svg/53719336 -https://youtu.be/C8MT-A7Mvk4