Project Report

In this project, I wrangled and analysed data collected from WeRateDogs. WeRateDogs is a twitter account with over nine million followers (as at the time this project is carried out). WeRateDogs is popularly known for its weird rating of dogs beyond the rating standard (10).

This report documents the steps taken in gathering, accessing, cleaning and storing data.

Gathering Data

There were three sources or format in which the data were gathered. The first dataset, twitter_archive_enhanced was collected in CSV format provided by Udacity. The second dataset was in a tab separated value (TSV) format. It was collected programmatically through a URL link provided by Udacity. The requests library was used to extract the data from the web. Thirdly, the third dataset was to be collected through twitter API. However, I collected the data from the json file provided by Udacity. The pandas read_json method was used to read the data to the dataframe.

Accessing Data

The data was accessed through programmable and visual means. Below are the issues identified with the data.

Quality Issues

Twitter archive

Issues identified with the Twitter archive data

1. Some of the rating denominators has values other than 10
    - I will be taking 10 as the standard rating value which dog rating values will be measured.
    - I'm only considering one dog at a time not multiple dog at once.

2. The timestamp column is in a string data type instead of datetime

3. The tweet id is in integer format instead of string.

4. Some values start with lowercase in p1, p2 and p3 data.

5. Underscore in p1, p2 and p3 instead of space.

6. Html tag in source column.

7. There are data other than dog data in the dataset.

8. There are dog tweets with no image in the dataset.

9. There are outliers in the rating_numerator column. Values as high as 1776

10. Missing data in the following columns:

    - in_reply_to_status_id
    - in_reply_to_user_id
    - retweeted_status_id
    - retweeted_status_user_id
    - retweeted_status_timestamp
    - expanded_urls

Image prediction

Issues identified with the Image prediction data

1. Tweet id is in integer format instead of string.

Tweet-json

Issues identified with the tweet-json data

1. The id is in integer data type instead of string data type.

2. favourite_retweet is not simple enough to describe the column name. total_retweet will be better.

3. favourite_count is not simple enough to describe the column name. total_likes will be better.

Tidiness issues

1. The dog_stages columns: floofer, doggo, pupper and poppo are in different columns instead of one.
2. In the image prediction dataset, there are too many columns about the strength of the predictions.
    - Cleaning should be performed on these prediction columns (p1, p1_conf, p1_dog p2, etc..) to produce a resultant outcome of fewer columns.

3. Tweet id is repeated in all the datasets.

Cleaning Data

In this stage, I cleaned the data. I ensured that every single issue, both quality and tidiness issues highlighted above were cleaned. But before I started the cleaning process, I made a copy of the three datasets. and named them archive_clean, image_clean and tweet_clean.

After cleaning all data, I merged all the three datasets into a combined form. Then, all the columns that will not be useful for my exploration were dropped. I ended up having a total of 5 columns and 1,666 rows of data.

This master dataset is stored as a CSV file called twitter_archive_master.csv.

Dropped columns after cleaning

The following columns were dropped from the image dataset on tidiness basis after the result of the prediction data (p1, p1_conf, p1_dog, p2, etc.) were used to separate dog records from non-dog records.

- jpg_url
- img_num
- p1
- p1_conf
- p1_dog
- p2
- p2_conf
- p2_dog
- p3
- p3_conf
- p3_dog

Other columns that were dropped include:

- Prediction
- name
- dog_stage
- source
- date

Please, kindly note that I am unable to include every single step of the data cleaning process in this report in order not to go too far from the 600 words limit recommended. However, all the details are included in the jupyter notebook file submitted.