WeRateDogs Twitter Data Project

By Itaman Favour Ofure

STEP 1: GATHERING DATA

After going through the data wrangling course and project instructions given on how to gather the data needed for wrangling and analysis, the steps i used are below:

1. I directly downloaded the twitter archive data titled twitter-archive-enhanced.csv on my pc.
2. I then proceeded to download the image prediction data using request library in my jupyter notebook. It is in a tsv file format.
3. Lastly i was supposed to create a twitter developer account and get the json file named tweet_json.txt using api but i was not able to get my developer account approved in time so i used the alternative provided by udacity by doenloading thr tweet_json file from my udacity class room.

After gathering each data i made them into three(3) dataframes.

Dataframe1: The twitter-archive-enhanced.csv was converted into a dataframe named twitter_archive. It gives the basic tweet information such as tweet_id, source, text, name, etc.

Dataframe2: The image_prediction.tsv file was converted into a dataframe named image_prediction. It contains information about dog prediction images.

Dataframe3: The extracted data from the tweet-json.txt file was converted into a dataframe called api. It contains information such as tweet_id, retweet_count, favorite_count.

STEP 2: ASSESSING DATA

In assessing these data, I made use of the two(2) ways of assessing data that is: Visual assessment and Programmatic assessment. In assessing visually I had to assess all three dataframes using an excel spreedsheet to be able to see through properly. After completing my visual assessment and noting things like missing values, incorrect entries, unwanted columns and splitup columns in the twitter_archive dataframe, I progressed to programmatic assessment using codes like twitter_archive.info(), etc. I moved on to the other dataframes and did the same thing. Below is a list of issues I discovered after assessing. They are grouped into quality issues and tidiness issues.

Quality issues:

1. There are a lot of missing data in the in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id,retweeted_status_user_id, retweeted_status_timestamp, and some missing data in the expanded_urls columns.

2. There are retweeted tweets, we only want original ratings that have images, so all retweeted ratings should be removed.

3. Ratings have inconsistent values.

4. There are unusual names for dogs such as none, a, an, the, are found in the name column.

5. The tweet_id column has a wrong datatype across all three tables, it is supposed to be a str.

6. The timestamp column is has a wrong datatype, it is supposed to be datetime datatype.

7. There are inconsistencies in cases(lower case and upper cases) in the p1,p2 and p3 column.

8. There are 2075 entries of data in the image_prediction data while there are 2356 entries of data in the twitter_archive data, which means there are missing images/tweets with no image.

Tidiness issues:

1. The dog category has four different columns, doggo, floofer, pupper,puppo, it should be made into one column titled dog_stage.

2. All three tables should be merged. Having just one tweet_id column after the merger.

STEP 3: CLEANING DATA

In cleaning all three dataframes of the issues discovered, I followed the following steps:

- First I made copies of all data, the purpose of this is so that if while cleaning an irreversable error is made it will only affect the copy made while the original remains intact. While making the copies, twitter_archive copy was named twitter_clean, image_prediction copy was named image_clean, and api copy was named api_clean.
- Then I proceeded to correcting the wrong datatype in the tweet_id column in all three tabels. Changing the datatype from integer to string.
- Next, I corrected the wrong timestamp column datatype in the twitter_clean dataframe. From object/string to datetime.
- Then, the next pressing issue was the dog stage. There are four(4) dog stages; doggo, floofer, pupper and puppo. They were in four separate columns, which is a major tidiness issue. I corrected this by combining/merging all four columns into one column named dog_stage.
- Next I merged all three tabels into one making sure the final tabel named main_data after merging had just one tweet_id column. I then made a copy of main_data named main_clean.
- I removed columns with missing values and retweets as we only need original ratings. First i removed all retweeted rows , that is all non-null values in the retweeted_status_id column before proceding to remove the columns that needed to be removed, which are; in_reply_to_status_id, in_reply_to_user_id,

retweeted_status_id,retweeted_status_user_id, retweeted_status_timestamp and retweet_count.

- Next, the issue of inconsistent values in the ratings columns, I decided against touching them because I assume they are part of the unique rating system.
- I then changed all unusual dog names such as none, a, an, the, which were not names where change to 'no name'.
- Then the issue of inconsistency in case in the p1, p2 and p3 column was fixed by changing all cases to lowercase.
- Lastly, I removed all rows with no images and incorrect images.

STEP 4: STORING DATA

I stored the final dataframe into a csv file format with the name 'twitter_archive_master.csv' with a final data of 1951 rows and 20 columns.