wrangle_report

September 5, 2022

0.1 GATHERING

The entirety of this project is a wrangling process to analyse the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. This process involved the gathering, assessing, cleaning of the dataframe. The gathering phase included downloading of WeRateDogs twitter archives, extracting a second dataframe from a URL, and then querying Twitter API to get the third dataset. These three tables were all gathered, assesed and cleaned before merging to form a master dataframe.

However this wrangling process would not have been possile without the functinality of these pyhton libraries; pandas, NumPy, requests, tweepy, json, matplotlib.

0.2 ASSESSING

In this section, i assessed all three tables visually, and programmatically for issues that had to do with tidiness and quality. From assessing the dataframes visually and programmatically, I realized the following snags;

Quality issues

1.column 'retweeted_status_user_id', 'name' is not descriptive enough

2.There are retweets in the dataframe.

3.(in_reply_to_status_id, in_reply_to_user_id, retweeted_status_user_id, retweeted_status_timestamp, source, text, expanded_url, 'jpg_url', 'p2', 'p2_conf', 'p2_dog', 'p3', 'p3_conf', 'p3_dog', 'date_created') are irrelevant.

4.False predictions in the image prediction dataframe indicates predictions contain animals other than dogs

5.p1, p1_conf, p1_dog are not descriptive enough

6.Datas contained in the 'timestamp' not in the right format

7.Name column in the twitter_archive dataset contains incosistent data

8.Dog breed are inconsistent

Tidiness issues

1.The dog "stage" (i.e. doggo, floofer, pupper, and puppo) should be one column

2.Numerator and Denominator should be one column

0.3 CLEANING

To address the issues listed above, I cleaned them using python codes, pandas and numpy funtions

1.column 'retweeted_status_user_id', 'name' was renamed as 'retweet_id'. and 'dog_name' respectively

2.There are retweets in the dataframe, so i masked the 'retweet_id' column to return only null values for retweets.

3.(in_reply_to_status_id, in_reply_to_user_id, retweeted_status_user_id, retweeted_status_timestamp, source, text, expanded_url, 'jpg_url', 'p2', 'p2_conf', 'p2_dog', 'p3', 'p3_conf', 'p3_dog','date_created') columns are irrelevant, so i dropped them

4.False predictions in the image prediction dataframe indicates predictions contain animals other than dogs, used the .loc function to provide rows that had predictions that were True.

5.Rating should be a single column, so i combined the numerator and denominator column to form a single column and renamed it ratings.

6.Datas contained in the 'timestamp' not in the right format, because both the time, and date are joined together, so I splitted the column, and dropped the time segment, and renamed the column 'date'.

7.Name column in the twitter_archive dataset contains incosistent data, i therefore replaced the names represented in small letters with 'None'

8.Dog breeds are inconsistent, so I replaced the serapator with space

9.The dog "stage" (i.e. doggo, floofer, pupper, and puppo) should be one column, and i wanted them to be on a single column rather than having to be scattered into four different columns so i merged the four columns together and renamed the column 'dog_stages'

10.p1, p1_conf, p1_dog are not descriptive enough, so i renamed them to more descriptive words.

0.4 STORING

After the cleaning phase, i merged all three dataframes together to form a single master dataframe 'twitter_archive_master' and saved it