wrangle_report

November 1, 2022

1 DATA WRANGLING PROJECT

We rate dogs data.

1.1 Data Gathering

- Fisrt read the 'twitter-archive-enhanced.csv' file in the workspace to dataframe pandas. It is the first data resource I have.
- Then, get the 'image-predictions.tsv' from an url through requests library, after that create a folder name 'image_predictions', write the content of 'image-predictions.tsv' to a file in that folder. Read the 'image-predictions.tsv' file in the workspace to dataframe pandas.
- The third resource we need get it from tweetpy API, I already sign in to tweeter and request, but they tell me to wait and not receive any result even several days. So I decide to use file 'tweet-json.txt'. Open it and read line each line, load the json to object and append some info we need to a list. After that create a dataframe to store the data.

1.2 Assessing Data

After some exploration, Some quality and tidiness issues were identified for the three tables. Details of the issues identified and solutions are below:

1.2.1 Quality Issues

1.We don't need retweeted tweet so all the tweet have retweeted is redundant.

2.In the twitter_archive_df we just keep `tweet_id`, `timestamp`, `source`, `text`, `rating_numerator`, `rating_denominator`, `name`, `doggo`, `floofer`, `pupper`, `puppo` we don't need some columns so we don't have to care about missing value in that column.

3.tweet_id is an int64 instead of str in 2 first table, and timestamp is an object

4.Source column is in HTML-formatted string, not a normal string.

5.Incorrect dog name "a".

6.Incorrect rating dominator 0.

7.We just need a column predict_breed from table image_predictions_df instead of all current colum

8.Sometimes the dog breeds listed is all lowercase, sometimes it is written in Sentence Case, sometimes there is an underscore (we will fix this after fix all the tidiness issues because we don't need all column in the prediction table)

1.2.2 Tidiness Issues

1.The last four columns in twitter_archive_df all relate to the same variable stage of dog (dogoo, floofer, pupper, puppo)

2.We have 3 table but we have the same observation is tweet

1.3 Cleaning Data

After identify issues with the 3 table, I'm going to clean all of it.

1.3.1 Quality Issues

1.Keep tweet that have no retweeted.

2.Delete redundant columns, we just keep tweet_id, timestamp, source, text, rating_numerator, rating_denominator, name, doggo, floofer, pupper, puppo.

3.Fix wrong data type.

4.Extract content from source column, and make it category type

5.Change incorrect dog name "a" to "None".

6.Change the wrong rating in denominator 960/00 -> 13/10

7.Create predict_breed, value is follow sequence equal p1 if p1_dog is True, if not it equal p2 if p2_dog is True... And delete redundant columns not use

8.Standardized dog breeds in predict column

1.3.2 Tidiness Issues

1.We have 3 table but we have the same observation So we have to joining all the dataframes based on the tweet_id

2.Convert the dog stage into one column instead of the multiple columns. And delete redundant column when done convert

1.4 Storing Cleaned Data

Now the data set is clean and ready for analysis. I saved the master table to twitter_archive_master.csv. Then I started my analysis.