wrangle_report

September 6, 2022

0.1 Gathering Data

I gathered three datasets using three methods. The first method was to directly download the twitter archive dataset using pandas library which consisted of downloading the provided csv file of the WeRateDogs twitter archives, I then uploaded the file into my Jupyter Notebook Workspace and read it into a pandas dataframe with a one-line code. In the second method I had to programmatically download the tweet image predictions tsv file from Udacity's servers using a specified url and the Request library, I then read the file into a pandas dataframe. And the third method was to query the twitter API for each tweet's JSON data using the Tweepy library and store the set of JSON data in a file named tweet_json.txt. I could not get access to the Twitter API as I could not get my developer account approved in time, so I downloaded the text file provided, I then read the tweet-json.txt file line by line into a pandas dataframe with three columns tweet_id, retweet_count, and favorite_count.

0.2 Assessing Data

For the visual assessment of the three gathered datasets I just printed the dataframes and scrolled through each of them to recognise data issues to later document them in following markdown cells,I then had to programmatically assess each dataframe.

For the twitter archive dataset I used pandas info tool to get the dimensions of the data and the number of non-null values with the datatype of each column, then I found the sum of null values in each column of the dataframe, I filtered the data frame to see how many entries had incorrect dog names, and finally I looked at the unique values of dog stages.

For the image predictions dataset I used pandas info tool to get the dimensions of the data and the number of non-null values with the datatype of each column, I then filtered the dataframe to see how many entries had at least one of the predictions for the image was not a breed of dog, I again used the same filtering method to find how many entries had all predictions for the image as not dog breeds.

For the tweet API dataset I used pandas info tool to get the dimensions of the data and the number of non-null values with the datatype of each column, then I found the statistical summary for the dataframe using the describe() function.

Documented Data Issues As quality issues I found that the twitter archive dataframe had missing values in In_repy_to_status_id, In_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamps, and expanded urls columns. Two other quality issues was that not all entries had image predictions, the dog names and stages were not all correct. As tidiness issues I found that the Image prediction data contained pictures

of objects not dog types. For example, indian_elephant and orange in the prediction of the images.And the text column in the twitter archive made the dataset untidy and may have served no purpose in the analysis.

0.3 Cleaning Data

Before cleaning the dataset I made copies of the original dataset. To clean the quality issue of missing data I just dropped the columns as they had many null values, for the dog stages I dropped all the columns as they had no use in my analysis, and I then dropped all the incorrect dog name entries. For the image prediction tidiness issue I removed all the entries that had predictions of not a dog breed. I dropped the text column then merged all my datasets to see the entries that didn't have image predictions then I dropped them.

0.4 Storing Data

After I had my merged clean dataframe, I stored in it a csv file named twitter_archive_master.csv as required.

0.5 Analyzing and Visualizing

I discovered three insights from the analysis of the master data using the groupby function to find the averages. The first insight was that the highest rated prediction 1 dog breed is the clumber with a mean rating of 27/10. The second insight was that the most liked(favorite_count) dog breed is the English_springer with 29K mean likes. The final insight was that the most popular(retweet_count) dog breed is the English_springer with 12K mean retweets. I made a bar plot of prediction 1 dog breed versus average favourite counts. The x-axis of the plot was cluttered because our dataset as in real life had many kinds of dog breeds.