Data Wrangling Report Project

Objectives

The project main objectives were:

• Perform data wrangling (gathering, assessing and cleaning) on provided thee sources of data.

• Store, analyze, and visualize the wrangled data.

• Reporting on 1) data wrangling efforts and 2) data analyses and visualizations.

Step 1: Gathering Data

In this phase, the three datasets were gathered and represented as pandas dataframes:

• The WeRateDogs Twitter archive (file name: 'twitter-archiveenhanced.csv')

This file was given and manually downloaded from Udacity)

• The tweet image predictions (file name: 'image-predictions.tsv').

This file was downloaded programmatically using the Requests library from a provided URL.

• Each tweet's entire set of JSON data were stored using Twitter API and Python's Tweepy library. (file name: 'tweet-json.txt')

This file was manually downloaded from Udacity as I do not have developer account for twitter.

Each tweet's JSON data was written to its own line with tweet ID, retweet count, and favorite count.

Step 2: Assessing Data

While working with data, a number of observations were made. Below are the observations found in the Assessing Step.

Quality

df_enhanced table

- Erroneous datatypes (timestamp columns)
- Retweets and replies included (redundant records)
- Drop unuseful columns: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp
- Incorrect rating_denominator
    - tweet_id:666287406224695296(2->10)
- Incorrect rating_numerator

    - tweet_id:666287406224695296(1->9)
    - Float number was recognized as integer
- There are 745 None values under name column, and the names with lowercase are invalid names.
- doggo, floofer, pupper, puppo columns contain 'None' value which is not counted as null
- There are records with more than one stages (doggo with one of the floofer, pupper, puppo columns)
    - tweet_id:855851453814013952
    - 854010172552949760
    - 817777686764523521
    - 808106460588765185
    - 802265048156610565
    - 801115127852503040
    - 785639753186217984
    - 781308096455073793
    - 759793422261743616
    - 751583847268179968
    - 741067306818797568
    - 733109485275860992
    - 775898661951791106
    - 770093767776997377
- Calculate rating with values of rating_numerator divided by rating_deniminator

***df_image* table**

- Sometimes lowercase and sometimes uppercase for breed names in p1, p2, p3 columns

Tidiness

- doggo, floofer, pupper, puppo columns should be combined in one column with category data type in the df_enhanced table
- Merge datasets to one

Step 3: Cleaning Data

Taken the action to solve the above issues, as a result, I combined the three datasets into one for data analysis and visualization.

Step 4: Storing, Analyzing, and Visualizing Data

Store the clean DataFrame created above in a CSV file named twitter_archive_master.csv.

In this section I answer the following questions by analyzing and visualizing the data:

- The highest rated dog
- The dog with highest retweet counts
- The dog with highest favorite counts
- What are most popular 10 dogs' names?
- What is the most common dog stage? Is there any difference on rating, retweet counts, favorite counts between dog stages?
- Is there any impact on retweet and favorite counts based on ratings?
- What is the percentage the algorithm can predict a dog breed?