WeRateDogs Project

Table of contents:

Gathering data

The first thing I did was importing all the needed libraries which are (pandas, numpy, requests, tweepy, time, JSON, seaborn, re, matplotlib). Then reading the provided csv file which is (twitter-archive-enhanced.csv), and downloading the image prediction file programmatically using the link provided. After that I created a developer Twitter account to gather the rest of the data needed for the project using twitter API, then stored the acquired data in a JSON file, creating from it a new dataframe called count_list.

Assessing data

I did the assessing phase for each file separately. I started with twitter_archive, then image_prediction, and last with the count_list dataframe that I created from the JSON file. I displayed each file, used .info(), .describe(), .sample(), .duplicated(), and .value_counts(), to find the quality and tidiness issues and I found the following:

Quality

- In name column some names are false
- IDs are float and int although we wont do any calculations on them
- Timestamp column are object
- Source column can not be read easily
- Delete columns that are unnecessary for calculations
- Some dog stages has two entries might be a typing error (found while merging files in the cleaning process)
- p1,p2,p3 columns are not consistently lower or uppercase
- p1,p2,p3 columns words are separated by an underscore
- img_num column is unnecessary

Tidiness

- doggo floofer pupper puppo columns all represent dog stage
- The p1,p2,p3_conf and p1,p2,p3 colums are unnecessary
- All three tables could be merged into one table twitter_archive

Cleaning data

First I made copies of the files that Iam going to perform the cleaning on. Then I started with the tidiness issues first. Merged all three files to one called Tweet_df. Created dog breed prediction and prediction confidence columns to merge all the unnecessary columns (p1,p2,p3_conf and p1,p2,p3). Created dog stage column to add all the (doggo floofer pupper puppo). Then found out after creating the dog stage column that there is a quality issue which was, some dog stages has two entries and I thought this might be a typing error so I iterated and went back to the assessing stage, and fixed that issue by removing the double entries. Then I deleted all the columns that I won’t be needing with my Analyzing and Visualizing stage. Fixed the readability issue with the source column using the (re) library. Then changed the datatype of the (source, tweet_id, timestamp) to (category, str, datetime). In the name column there are names that are false like lowercase letters, 'by', 'all', 'the', and others like, so I located them all and replaced them with None. As for the breed_prediction column I capitalized all the breed names and separated them with ' ' instead of '_'.

Storing data¶

Finally I stored the final product file Tweet_df to a new csv file called 'twitter_archive_master.csv'.