wrangle_report

December 13, 2020

1 WeRateDogs wrangling report¶

In Udacity Wrangling and Analyze Data project. i analyzed the tweet archive of Twitter user @dog_rates, also known as WeRateDogswill. WeRateDogs is a Twitter account that rates people's dogs. Each tweet include the dog's picture, types, comments and its rating. All dog's rating are overrated with scores higher than 10/10. The datasets will be gathered, assessed and cleaned to give insights and visualizations.

libraries used : pandas numpys os tweepy json seaborn matplotlib requests

there were 3 steps done before analysing the data which are the following :

1.1 1_ Gathering the data

datat were gathered from 3 diffenet sources - first source is Enhanced Twitter Archive: whic is The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets, but not everything. One column the archive does contain though: each tweet's text which was read through pandas into a dataframe - second source of data was Twitter API which i used here for the the first time so created twitter developer account ot get the keys and token which would allow me to use the API , the api was used to pull the numbers of the tweets_id which we had from the first data source the csv file , the data was filtered to elminte the errors and then were stored in a list and then a json file was created from this lis t to create our second data frame - third data source was the Image Predictions File which was a provided to me through a url, the file was downloaded programatically whic was a tsv file and from it i created my third data frame

1.2 2- Assessing the data

in this step i assessed the data visually and progrmatically - visually i noticed that there were a couple of things off for example the tideness , bscically all the dataframes had to be merged toegther , and there were alot of cloumns that had barley any values present in it so i had to get red of them and also the dogs catagories had seperate columns and some of the columns didn't have any meaningful names like p1 p2 etc - progrmatically i fouud that there were some missing values in the second and the third dataframes, and the tweet id and some other columns had a wrong type and timestamp was of value object, none values where present in the name column instead of nan ,also mulitple dog catagories where present in the same row in 14 different index,also rating_numerator and rating_domincator can be combined in one column called rating,capitalization of the p1, p2, and p3 column values uniform, also found _ in p1, p2 and p3 with space and

1.3 3- Cleaning the data

- first i made a copy of all the three data frames and then merged them together in a data frame called df_master and then dropped the index which had nultiple dog catagories in one row and then corrected the wrong types of the columns , and then replaced all the None values in the name column with Nan and also false names were converted to NaN,removed the decimel from the rating_numerator and then combined the rating_denominator and rating_numerator to one column ,i aslo dropped unneeded columns from the dataframe , columns (doggo, floofer, pupper and puppo) were melted togther under a column called stage ,all the names of the dog predections were made lower case and got rid of the _ and - from the dog names , and then changed all of the predections columns names

1.4 4- Analysis

- during the anylisis i started with checking average score of all the dogs ratings and then average score of all the dogs where the first prediction is true and i plotted the results
- then i plotted the top 10 dog breads using matplotlib
- after that i used .corr methhod pearson on the whole dataframe to check the correlation and i found that the correlation between the favorites counts and the retweets counts was postive and the highest so i plotted it and also plotted the correlation betwwen tghe first prediction rate and the 3rd prediction trated as the correlation was negative
- after that i tried to find the most popular dog name by plotting it
- i checked the development of this trend over time so groupped the retweets count by month and year and plooted it and same was done for the favorites counts which made me find out that this trind strated to boom by 2017
- i created a cloud of the words in the tweets

[ ]: