wrangle_report

September 15, 2023

1 Wrangling efforts

Using the available information in the project I have summarised my wrangling efforts as below- " Enhanced Twitter Archive

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 I used to extract rating, dog name, and dog "stage" (i.e. doggo, floofer, pupper, and puppo) to make this Twitter archive "enhanced." Of the 5000+ tweets, I have filtered for tweets with ratings only (there are 2356). "

Most faviourted Dog- https://pbs.twimg.com/media/C2tugXLXgAArJO4?format=webp&name=medium

1.1 What we know-

Some ratings are not correct Dog names could be incorrect Dog stages are incorrect Original ratings and no retweets no need to go beyond August 1st, 2017

1.1.1 Effort

I have made use of libraries and methods taught as part of the first module. I have heavily relied on using pandas apply method to filter out and fix necessary issues that are document below. I have additionally read the pandas documentation to fix timestamp issues using the method 'to_datetime'. I have consolidated the final cleaned data in a csv file called 'twitter_archive_master.csv' and read the same in the master data frame called df_master

While identifying the stage of the dogs i had to use a specific order in the list created i.e. dog_stage = ['pupper', 'floofer', 'puppo', 'doggo']. This helped me correctly identify the dog stages, although this is still not completely reliable however the order does produce better results if otherwise changed.

While using the melt function I had to drop the column regarding stages as it led to duplication of data.

1.1.2 Gathering of data:

- Twitter archive
- image_predictions
- extended tweets data

I tried playing with the twitter APIs, however no free version of APIs support this use-case hence had to resort back to using the data available in the project

1.1.3 Quality issues

df_twitter:- - 1) tweet_id is int - 2) timestamp is object - 3) expanded urls has missing data - 4) tweet_id with 810984652412424192 doesn't have rating and 666287406224695296 has incorrect i.e. 9/10 instead of 1/2 - 5) Name for dogs 'None' counted as object - 6) Incomplete data for Dog names - 7) Incomplete stages for dogs - 8) source column can be standardised i.e. iphone, webclient, vine etc

df_tweet:- - 9) extended_entities colum in df_tweets has media_urls

df_image:- - 10) 2075 entries for data(sort missing data) - 11) column names for p1 algorithm, p1_conf and p1_dog as user friendly names

1.1.4 Tidiness issues

df_twitter:- - 1) Dog stages mentioned as columns - 2) dog images and dog stages should be combined - 3) Merging the data together for a master table

1.1.5 Final files

1. twitter_archive_master.csv
2. Data Frames:

   - df_master
   - df_tweets_data
   - df_dogs_data

1.1.6 Conclusion

I had a lot of fun doing this project, i learned a great deal about pandas dataframes as well as pandas series. I made use of lot of pandas methods, also realised that using pandas aggregate functions can be handy when sql is not the first choice.

```
In [ ]:
```