act_report

July 19, 2022

0.0.1 WeRateDog Analysis Act Report

This report is the summary of the data analysis process of this data wrangling project.

In this project, I made use of three different datasets that were obtained using three different methods.

- The first dataset was provided by Udacity named `twitter_archive_enhanced.csv` and it was downloaded manually and saved as `twitter_df`. It contained basic information about 2356 tweets.
- The second dataset was a tsv file named `image_predictions.tsv` which was already hosted on Udacity's servers.Then, I downloaded programmatically using python's requests and os libraries. I read it into a pandas dataframe named as `imagepre_df`. It contained 2075 predictions made by a neural network that can classify dog breeds.
- The third dataset was a `tweet-json.txt` file which I downloaded and then read the file line by line to obtain basic information such as tweets, favorite count, retweet count which were a total of 2354.

While assessing, I detected 8 quality issues and 2 tidiness issues which I cleaned using variety of python and pandas methods.

After gathering, assessing and cleaning, I saved the master dataset as a csv file named `twitter_archive_master.csv`.

Here are the insights and visualizations I obtained after analyses.

First, I imported the pandas library and loaded the master dataset.

```
In [1]: import pandas as pd
        df= pd.read_csv('twitter_archive_master.csv')
```

```
In [2]: df.head()
```

```
Out[2]:              tweet_id            timestamp              source  \
        0  890240255349198849  2017-07-26 15:59:51  Twitter for iPhone
        1  884162670584377345  2017-07-09 21:29:42  Twitter for iPhone
        2  872967104147763200  2017-06-09 00:02:31  Twitter for iPhone
        3  871515927908634625  2017-06-04 23:56:03  Twitter for iPhone
        4  871102520638267392  2017-06-03 20:33:19  Twitter for iPhone

                                                     text  rating_numerator  \
0  This is Cassie. She is a college pup. Studying...                14
1  Meet Yogi. He doesn't have any important dog m...                12
2  Here's a very large dog. He has a date later. ...                12
3  This is Napolean. He's a Raggedy East Nicaragu...                12
4  Never doubt a doggo 14/10 https://t.co/AbBLh2FZCH                14

   rating_denominator      name dog_stages  favorite_count  retweet_count  \
0                  10    Cassie      doggo         32467.0         7711.0
1                  10      Yogi      doggo         20771.0         3128.0
2                  10      None      doggo         28031.0         5669.0
3                  10  Napolean      doggo         20730.0         3628.0
4                  10      None      doggo         21461.0         5764.0

                                              tweets
0  {'created_at': 'Wed Jul 26 15:59:51 +0000 2017...
1  {'created_at': 'Sun Jul 09 21:29:42 +0000 2017...
2  {'created_at': 'Fri Jun 09 00:02:31 +0000 2017...
3  {'created_at': 'Sun Jun 04 23:56:03 +0000 2017...
4  {'created_at': 'Sat Jun 03 20:33:19 +0000 2017...
```

Then using the describe function, I got more information about the dataset.

```
In [3]: df.describe()
```

```
Out[3]:            tweet_id  rating_numerator  rating_denominator  favorite_count  \
        count  2.347000e+03       2347.000000              2347.0     2345.000000
        mean   7.431992e+17         12.232211                10.0     8141.895522
        std    6.863351e+16         40.900209                 0.0    11873.823039
        min    6.660209e+17          0.000000                10.0        0.000000
        25%    6.784049e+17         10.000000                10.0     1415.000000
        50%    7.210012e+17         11.000000                10.0     3627.000000
        75%    8.000798e+17         12.000000                10.0    10192.000000
        max    8.924206e+17       1776.000000                10.0   132810.000000

               retweet_count
        count    2345.000000
        mean     3189.313433
        std      5309.440551
        min         0.000000
        25%       631.000000
        50%      1489.000000
        75%      3652.000000
        max     79515.000000
```

In order to gain insights, I asked the data some questions.

0.0.2 Questions

1. Which of the dog stages is the most popular?