act_report

June 28, 2022

1 Analyzing and Visualizing the data

After cleaning was done and data was merged into one dataset, I did some exploratory data analysis to find insights in the data. Exploratory Data Analysis refers to the exploring of a dataset or datasets using a variety of techniques and tools to get hidden insights. I started by checking the statistics of the data using the describe() function.

```
In [143]: df_mergedclean.describe()
```

```
Out[143]:            tweet_id  rating_numerator  rating_denominator      img_num  \
          count  2.073000e+03       2073.000000         2073.000000  2073.000000
          mean   7.383634e+17         12.265798           10.511819     1.203570
          std    6.780118e+16         40.699924            7.180517     0.561856
          min    6.660209e+17          0.000000            2.000000     1.000000
          25%    6.764706e+17         10.000000           10.000000     1.000000
          50%    7.119681e+17         11.000000           10.000000     1.000000
          75%    7.931959e+17         12.000000           10.000000     1.000000
          max    8.924206e+17       1776.000000          170.000000     4.000000

                      p1_conf       p2_conf       p3_conf  retweet_count  \
          count  2073.000000  2.073000e+03  2.073000e+03    2073.000000
          mean      0.594532  1.346665e-01  6.034005e-02    2976.089243
          std       0.271234  1.006830e-01  5.092769e-02    5054.897526
          min       0.044333  1.011300e-08  1.740170e-10      16.000000
          25%       0.364095  5.390140e-02  1.619920e-02     634.000000
          50%       0.588230  1.186220e-01  4.947150e-02    1408.000000
          75%       0.843911  1.955730e-01  9.193000e-02    3443.000000
          max       1.000000  4.880140e-01  2.734190e-01   79515.000000

                   favorite_count
          count        2073.000000
          mean         8556.718283
          std         12098.640994
          min             0.000000
          25%          1674.000000
          50%          3864.000000
          75%         10937.000000
          max        132810.000000
```

I looked at the value counts of different dogs in the look column just to get a sense of which dog was most popular or which dog term was used the most in this dataset.

```
In [151]: df_mergedclean.look.value_counts()
```

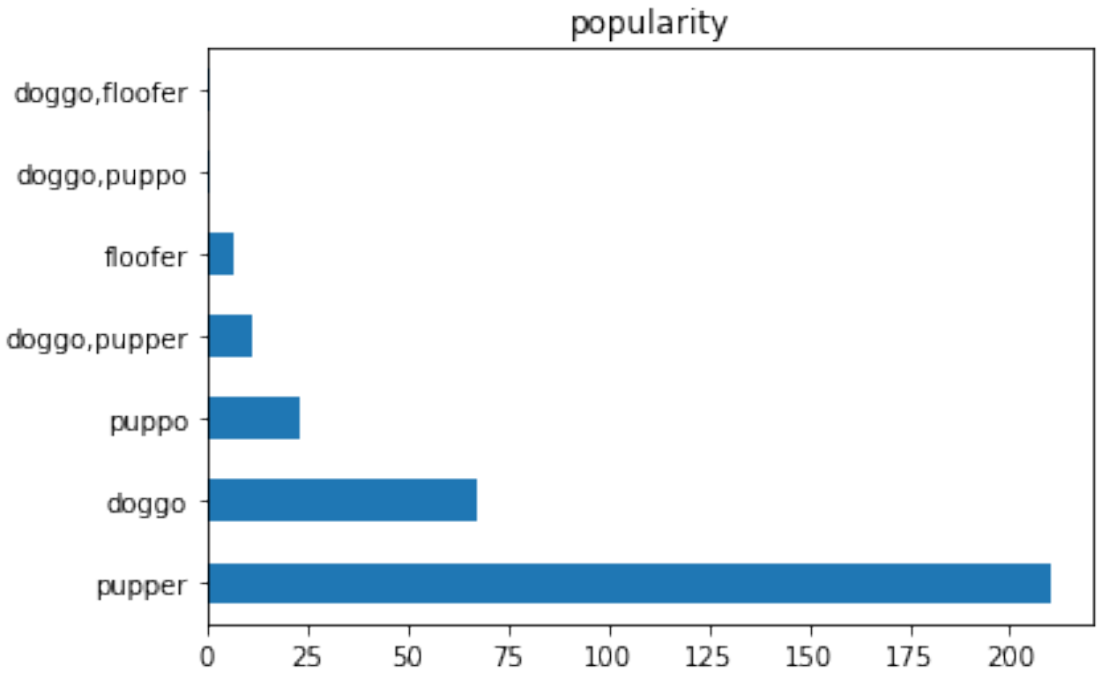
```
Out[151]: pupper    210
          doggo      67
          puppo      23
```

```
doggo,pupper      11
floofer            7
doggo,puppo        1
doggo,floofer      1
Name: look, dtype: int64
```

From the above data, we see that pupper had the most entries showing they had more popularity. Most probably because of how adorable they are. I then plotted a graph to represent this data. Floofer had the least entries in this data meaning the term was probably not as popular as the rest. I named the graph popularity in that it could be the popularity of the pupper (puppys) or the term itself (pupper).

```
In [154]: # Plotting popularity of the dog
          df_mergedclean.look.value_counts().plot(kind='barh',title= 'popularity')
```

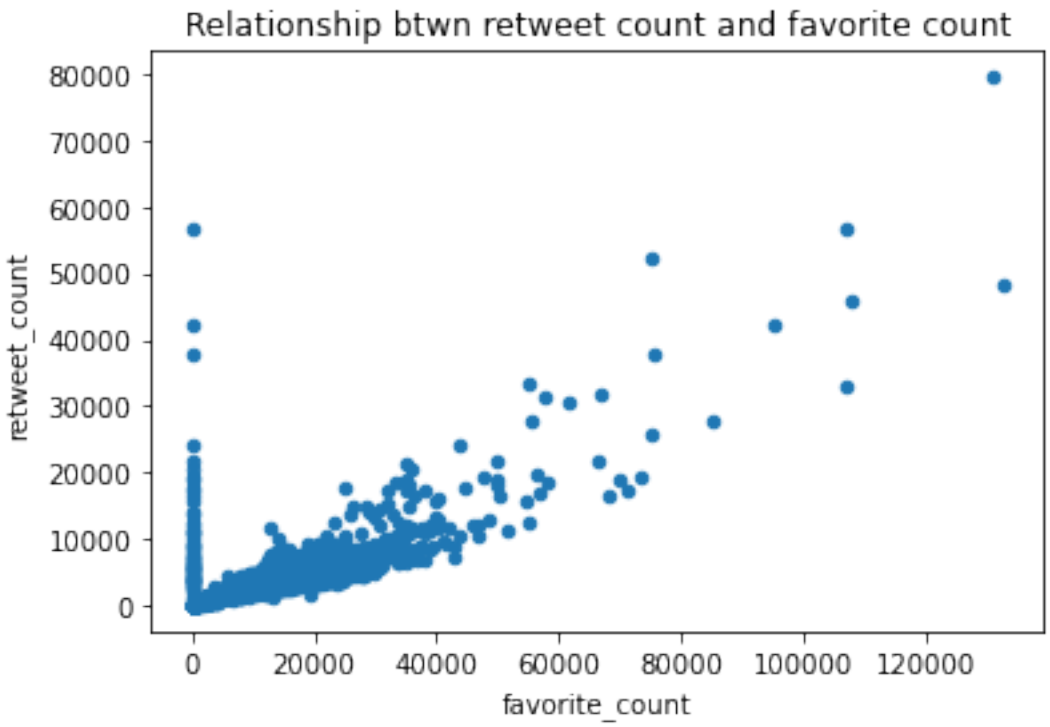
```
Out[154]: <AxesSubplot:title={'center':'popularity'}>
```



The above graph clearly illustrates the popularity of the pupper.

I then decided to plot a graph showing the relationship between retweet count and favorite count.

```
In [149]: # Plotting a graph showing the relationship between retweet count and favorite count
          viz1 = df_mergedclean.plot.scatter(x='favorite_count', y='retweet_count', title='Relat
```



From the graph, I noted there is a positive correlation between retweet count and favorite count.

1.0.1 Insights:

After doing my analysis, these are the insights I came up with:

1. Pupper was the most popular. Pupper is slang for a small dog or puppy. Another possibilty is that it could also mean most prefer referring to dogs as pupper as compared to doggo, puppo and floofer. Pupper is a popular term.
2. Maximum favorite count was 132810 and maximum retweet count was 79515. This means people favorited tweets more than retweeting. We also see that retweet count and favorite count are positively correlated for the majority meaning popularity of a tweet was portrayed in both.
3. A large percentage of the dog's names weren't provided in the names section. This could mean that the data wasn't provided by the actual owners or owners opted not to privide their pets' names in the instances where there was no name info.

1.0.2 Conclusion

The project was quite interesting and engaging. I have learnt plenty when it comes to wrangling data and added some new skills to my data analysis skilset.

In [ ]: