WRANGLE A DATASET REPORT

INTRODUCTION

This project involves wrangling (and analyzing and visualizing) the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comments about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.

The datasets used contain the content shown below:

```
twitter_archive_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2356 entries, 0 to 2355
Data columns (total 17 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   tweet_id                    2356 non-null   int64
 1   in_reply_to_status_id       78 non-null     float64
 2   in_reply_to_user_id         78 non-null     float64
 3   timestamp                   2356 non-null   object
 4   source                      2356 non-null   object
 5   text                        2356 non-null   object
 6   retweeted_status_id         181 non-null    float64
 7   retweeted_status_user_id    181 non-null    float64
 8   retweeted_status_timestamp  181 non-null    object
 9   expanded_urls               2297 non-null   object
 10  rating_numerator            2356 non-null   int64
 11  rating_denominator          2356 non-null   int64
 12  name                        2356 non-null   object
 13  doggo                       2356 non-null   object
 14  floofer                     2356 non-null   object
 15  pupper                      2356 non-null   object
 16  puppo                       2356 non-null   object
dtypes: float64(4), int64(3), object(10)
memory usage: 313.0+ KB
```

```
dog_images.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2075 entries, 0 to 2074
Data columns (total 12 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   tweet_id  2075 non-null   int64
 1   jpg_url   2075 non-null   object
 2   img_num   2075 non-null   int64
 3   p1        2075 non-null   object
 4   p1_conf   2075 non-null   float64
 5   p1_dog    2075 non-null   bool
 6   p2        2075 non-null   object
 7   p2_conf   2075 non-null   float64
 8   p2_dog    2075 non-null   bool
 9   p3        2075 non-null   object
 10  p3_conf   2075 non-null   float64
 11  p3_dog    2075 non-null   bool
dtypes: bool(3), float64(3), int64(2), object(4)
memory usage: 152.1+ KB
```

```
tweets.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2354 entries, 0 to 2353
Data columns (total 3 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   id              2354 non-null   int64
 1   retweet_count   2354 non-null   int64
 2   favorite_count  2354 non-null   int64
dtypes: int64(3)
memory usage: 55.3 KB
```

QUESTIONS

The following were the questions that guided me in my analysis:

1. What is the trend of retweet count and favorite count over time?
2. What is the correlation between rating_numerator,rating_numerator, retweet_count, favorite_count,p1_conf,p2_conf and p3_conf.
3. What are the most common dogs

FINDINGS

The following were the findings observed from the analysis process

1. **What is the trend of retweet count and favorite count over time?**

   There were more favorite counts than retweeted counts over the years.

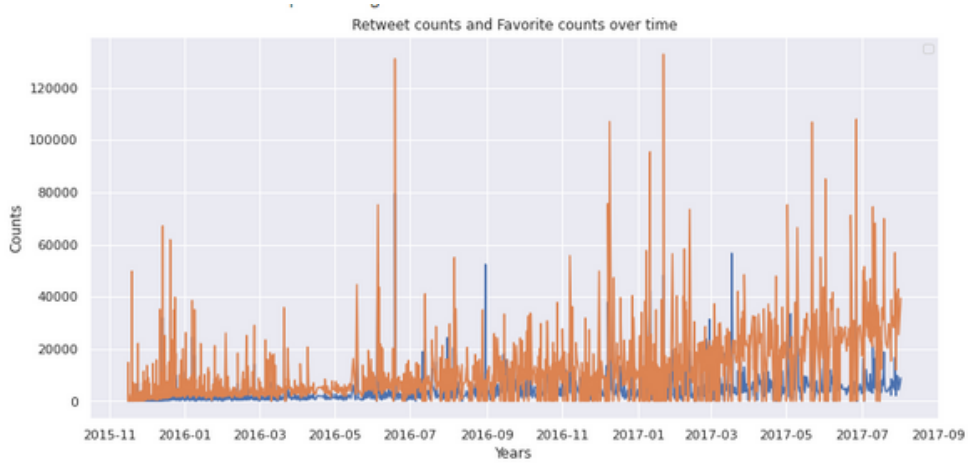

Figure 1.

2. What is the correlation between rating_numerator,rating_numerator, retweet_count, favorite_count,p1_conf,p2_conf and p3_conf

There is a strong correlation between favorite count and ratings. This means that a dog that has more favorite counts has a higher rating

Figure 2.

3. What are the most common dogs

The golden_retriever is the most common dog, followed by labrador_retriever and pembroke

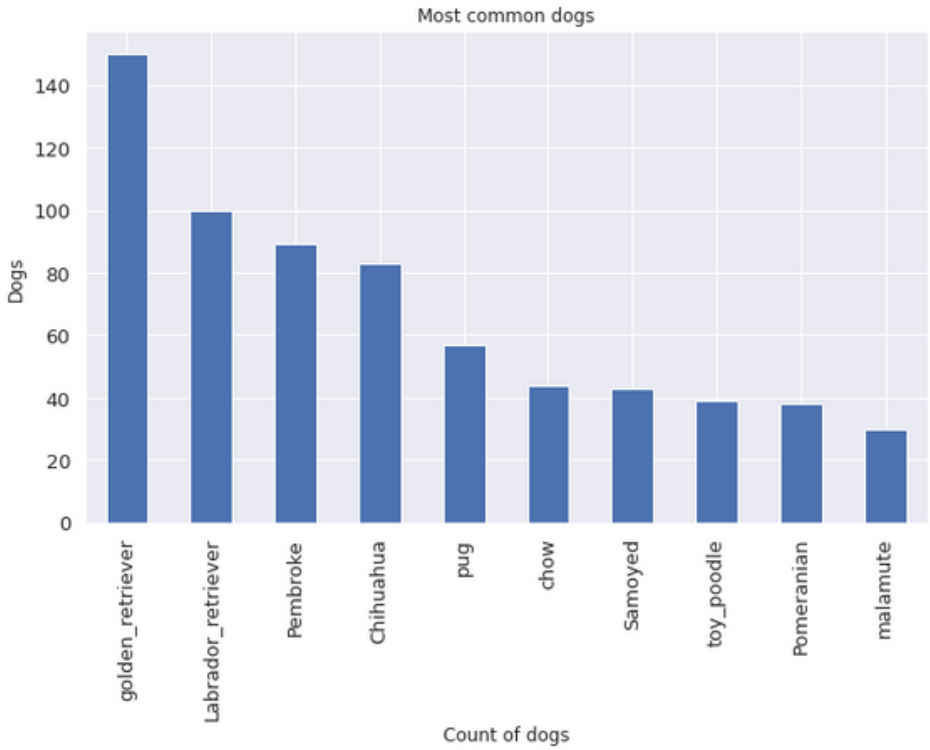

Figure 3.

CONCLUSIONS

The most favorited dog is the golden_retriever.
There is a strong correlation between rating_numerator and favorite_count meaning that dogs that had more favorite counts were highly rated

LIMITATIONS

The dataset had some missing values which affected the quality and tidiness of the data. This made the wrangling process difficult.

REFERENCES

https://matplotlib.org/3.5.0/tutorials/introductory/pyplot.html
https://seaborn.pydata.org/generated/seaborn.heatmap.html
https://seaborn.pydata.org/generated/seaborn.pairplot.html
Udacity course content