Data Analysis Professional Track

Project: Wrangle and Analyze Data

Act Report

Introduction

The dataset used is about the `WeRateDogs®`. I followed data gathering, data assessing, data cleaning and data storing process.

In Data Gathering process, I gathered data from three datasets. The first one I downloaded it manually, the second I downloaded it programmatically, and the third file from the Twitter API.

Based on the data gathered, I have assessed the most evident issues (17 issues in total) and documented it.

In Data Cleaning process I have fixed all identified issues, and I have also merged two files (Archive file and Twitter API file).

The final dataframes were stored as `twitter_archive_master.csv` and `image_predictions_master.csv`.

In the Data Analysis and Visualization, I have posed few questions to guide my analysis like:

1. What's the trend of Retweets and Favorites over Time?
2. What's the 10 most frequent distribution about dog breed inside all levels (Algorithm #1, Algorithm #2 and Algorithm #3)?
3. What's the top 10 breeds that receive the highest/lowest interaction in terms of retweet count average and favorite count average?
4. What's the interaction with different dog stages in terms of retweet count average and favorite count average?
5. What is number of tweets monthly? and etc...

The issues I have faced:

I ignore outliers that are found in rating.

```
In [69]: twitter_archive_df_clean[twitter_archive_df_clean['rating_numerator'] > 15 ].text

Out[69]: 516     Meet Sam. She smiles 24/7 &amp;amp; secretly aspires to be a reindeer. \nKeep Sam smiling by clicking and sharing this lin
         k:\nhttps://t.co/98tB8y7y7t https://t.co/LouL5vdvxx
         979     This is Atticus. He's quite simply America af. 1776/10 https://t.co/GRXwMxLBkh
         2074    After so many requests... here you go.\n\nGood dogg. 420/10 https://t.co/yfAAo1gdeY
         Name: text, dtype: object
```

There are some tweets have low rating because the image of each tweets contains anything except dog.

I read `twitter_archive_master` and `image_predictions_master`.

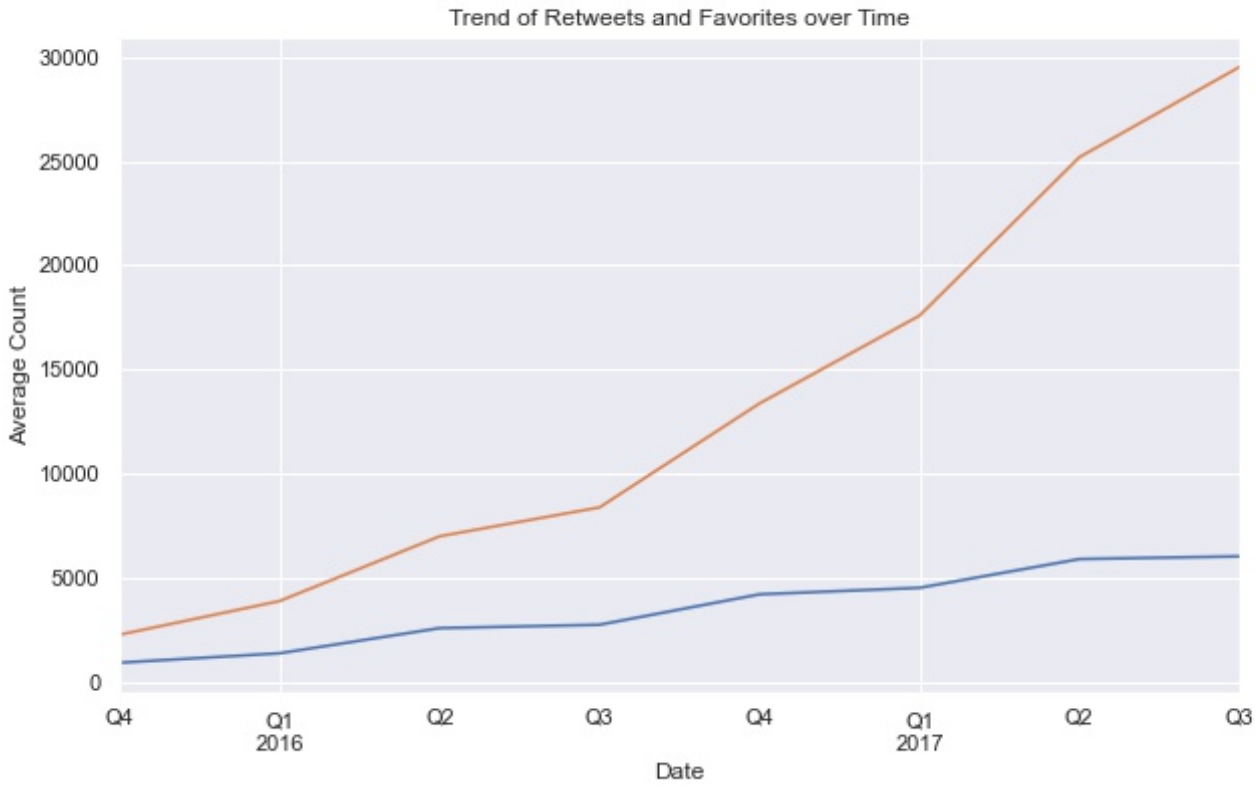
I set the `timestamp` column as an index to `twitter_archive_master_df` dataset.

The most retweeted and the most favorite tweet. Tweet Link

Analysis

Based on a dataframe of several tweets from `WeRateDogs®` (provided by `twitter_archive_master.csv` and `image_predictions_master` file), I would like to investigate:

1. How is the trend of retweets and favorites over time?

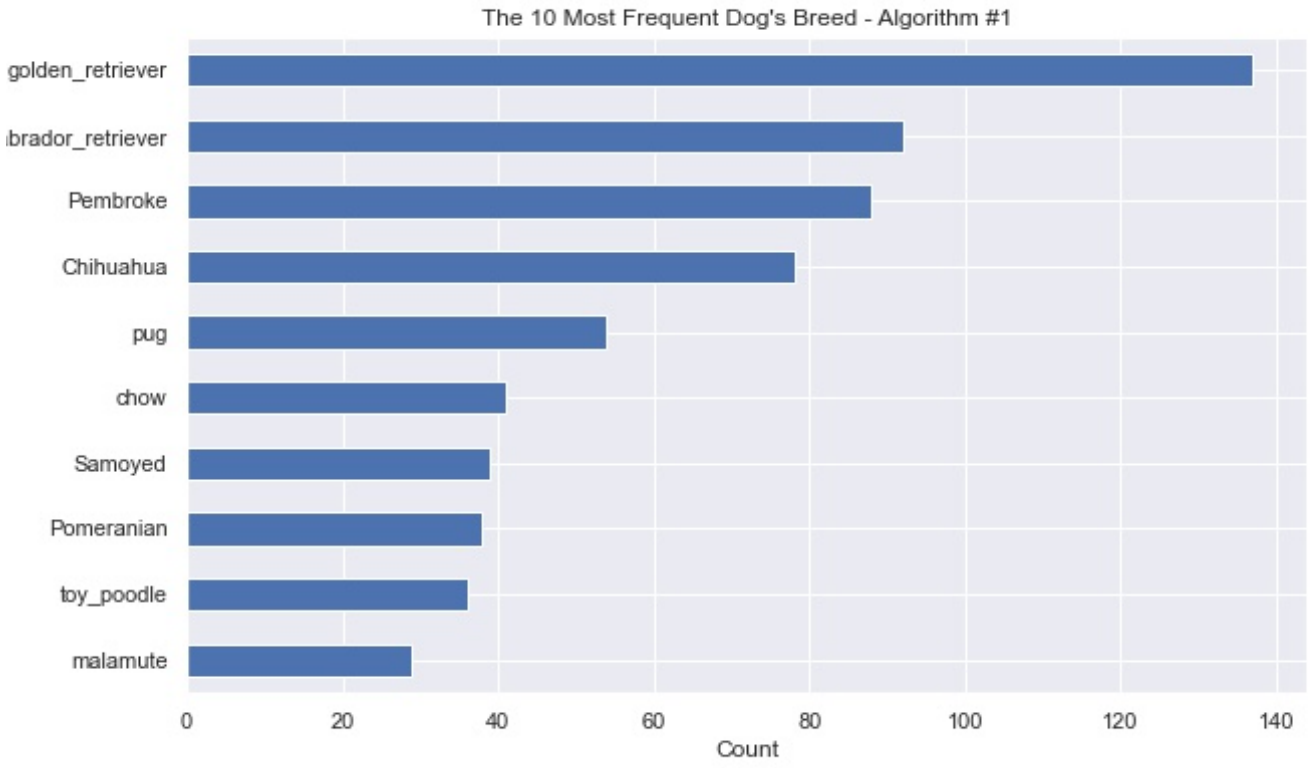


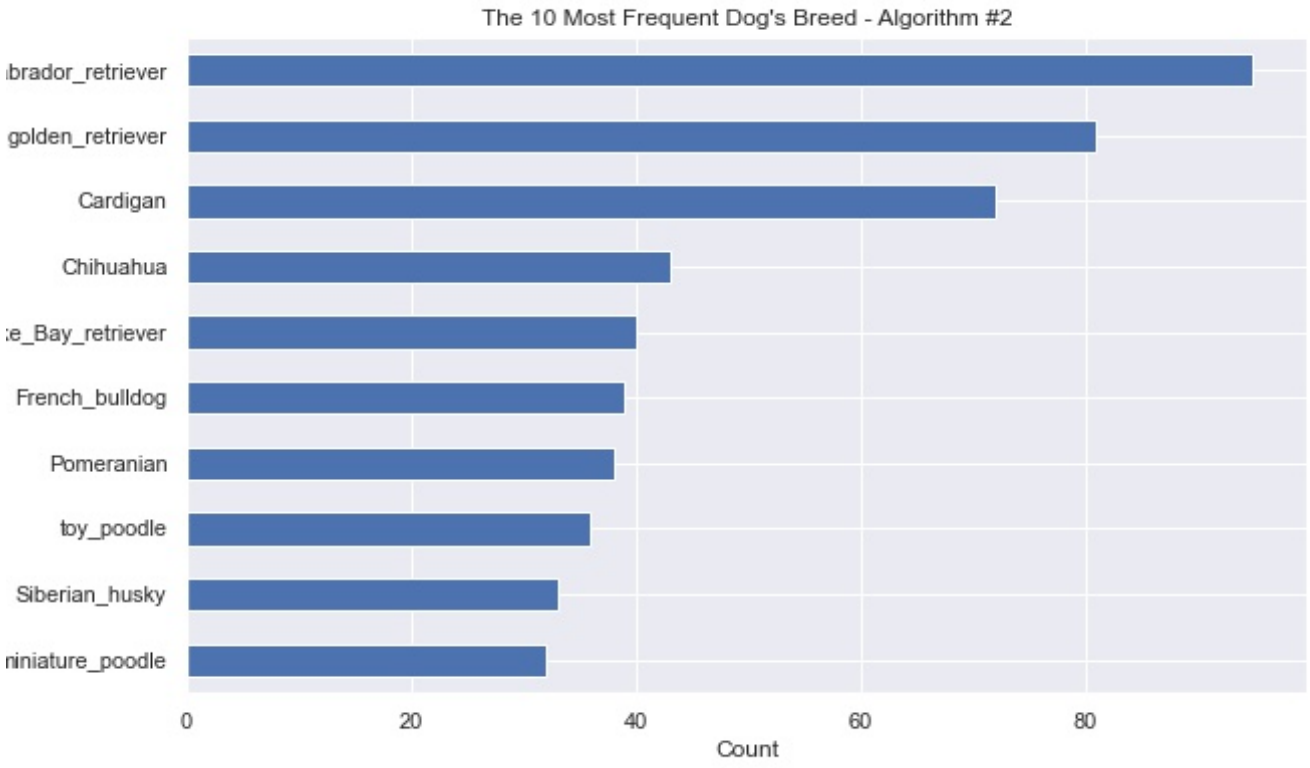
Conclusion

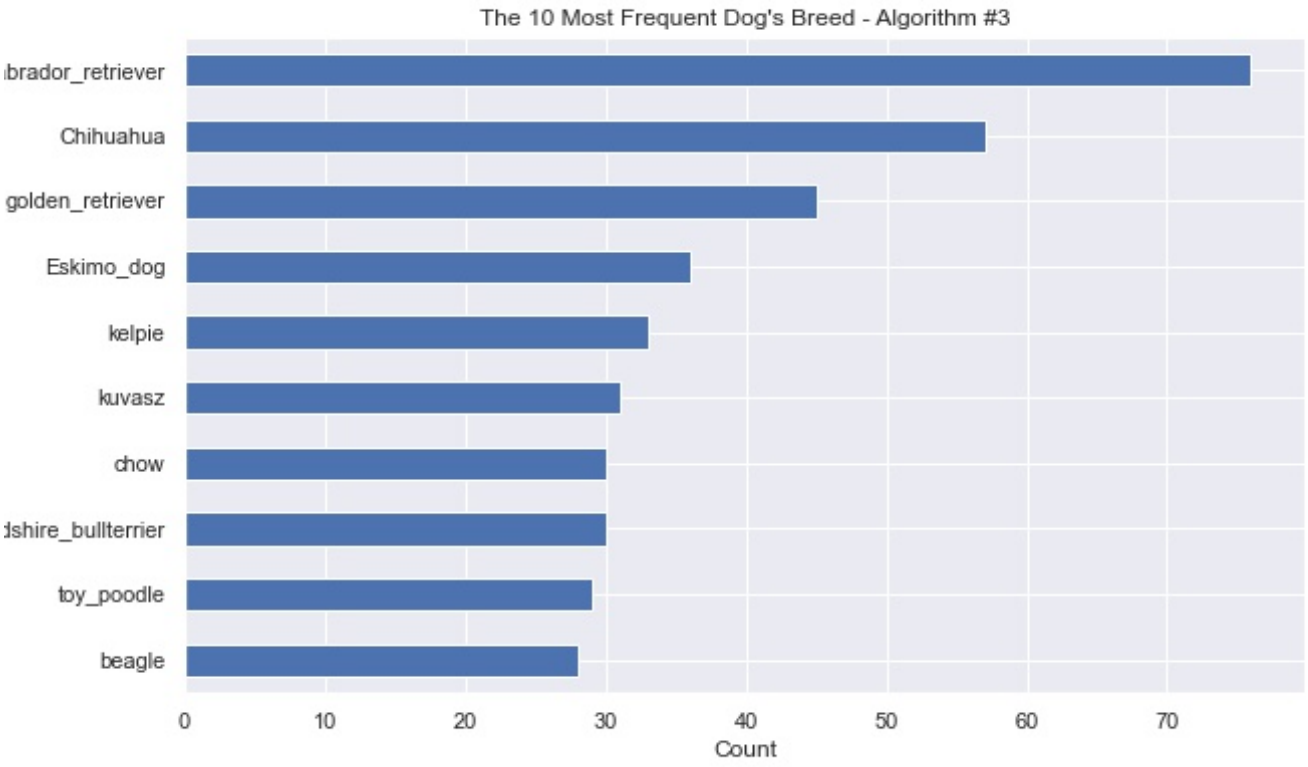
- We see that there are more favorites than retweets. The favorite count increases strongly, the retweet count seems increases slowly.

2. how is the output of each algorithm employed to predict the dog's breed?

```
2.1 What's the first 10 breeds with more appearance?
```







2.2 What's the number of breeds in each algorithm with more than 20?

```
P1: 15 breed.
P2: 18 breed.
P3: 21 breed.
```

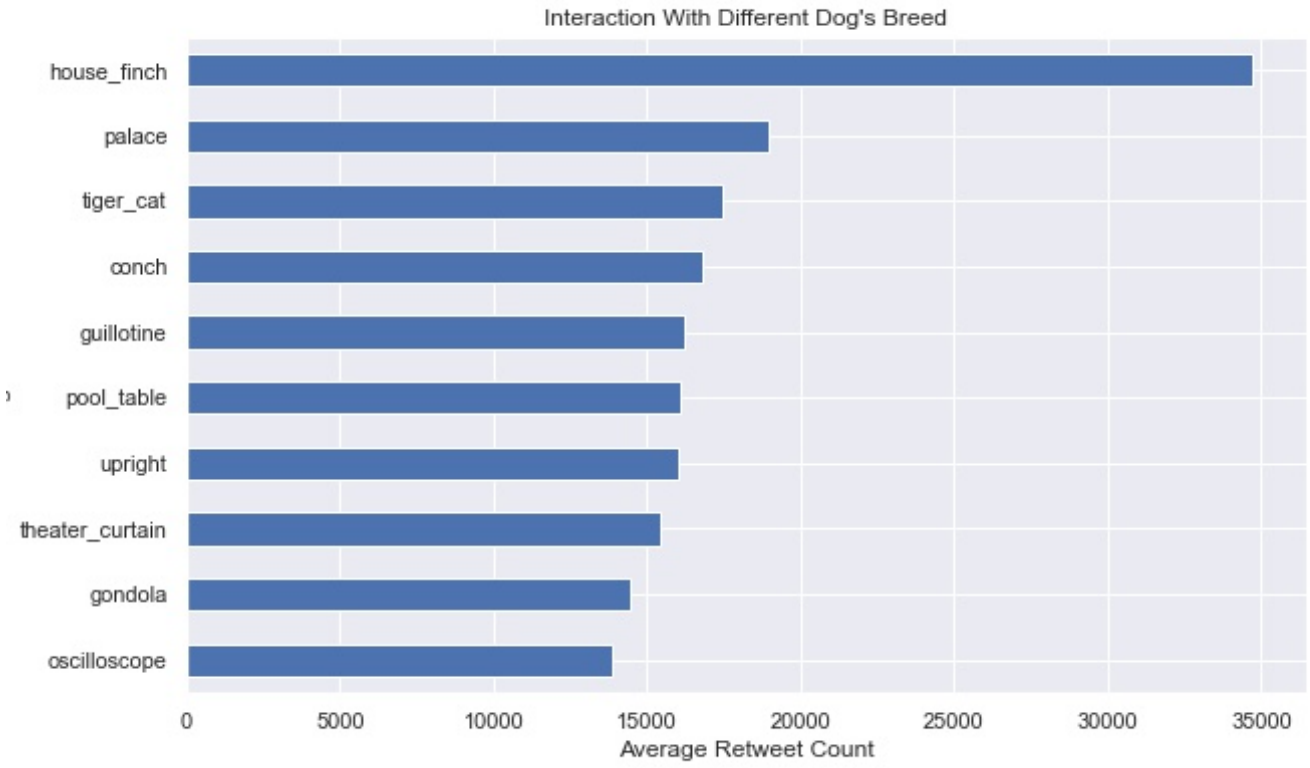
Conclusion

- `Algorithm #1` has fewer breeds and high frequence in some breeds.

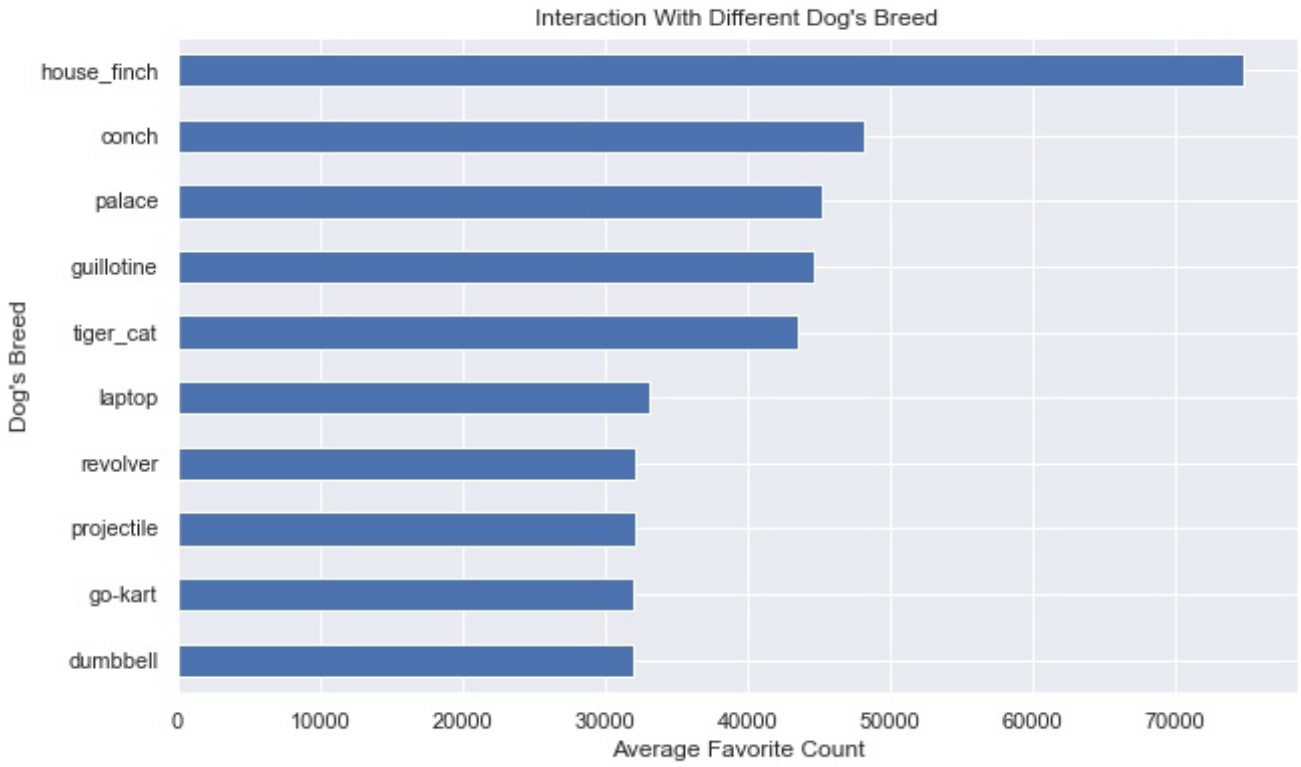
- `Algorithm #3` has more breeds and the dogs are spread in more breeds and also it has less frequence.

3. What's the top 10 breeds that receive the highest/lowest interaction in terms of retweet count average and favorite count average?

```
prediction
oscilloscope       13873.5
gondola            14475.5
theater_curtain    15395.0
upright            16005.0
pool_table         16071.0
guillotine         16185.0
conch              16805.0
tiger_cat          17452.0
palace             18932.0
house_finch        34737.0
Name: retweet_count, dtype: float64
```



```
prediction
dumbbell       31970.0
go-kart        31970.0
projectile     32035.0
revolver       32035.0
laptop         33109.0
tiger_cat      43513.0
guillotine     44579.0
palace         45212.0
conch          48103.5
house_finch    74815.0
Name: favorite_count, dtype: float64
```



```
prediction
trombone              96.0
hair_spray            77.0
piggy_bank            77.0
pitcher               74.0
spotted_salamander    60.0
wing                  53.0
power_drill           45.0
crash_helmet          37.0
toaster               37.0
desk                  32.0
Name: retweet_count, dtype: float64
```

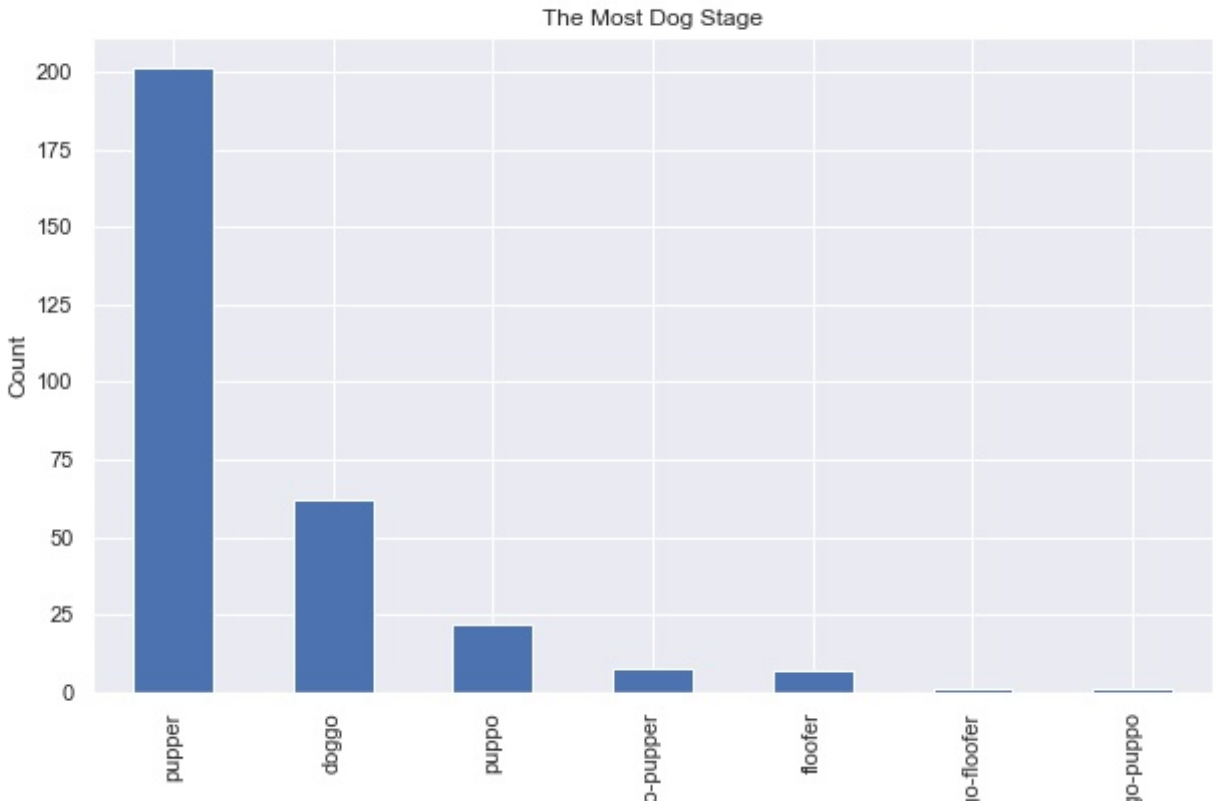
```
prediction
hair_spray      304.0
piggy_bank      304.0
trombone        277.0
French_horn     277.0
cornet          277.0
wing            221.0
toaster         186.0
crash_helmet    186.0
power_drill     151.0
desk            93.0
Name: favorite_count, dtype: float64
```

Conclusion

- `house_finch` breed has highest interaction.
- `desk` breed has lowest interaction.

4. Which dog stage is found most?

```
pupper            201
doggo             62
puppo             22
doggo-pupper      8
floofer           7
doggo-floofer     1
doggo-puppo       1
Name: dog_stages, dtype: int64
```

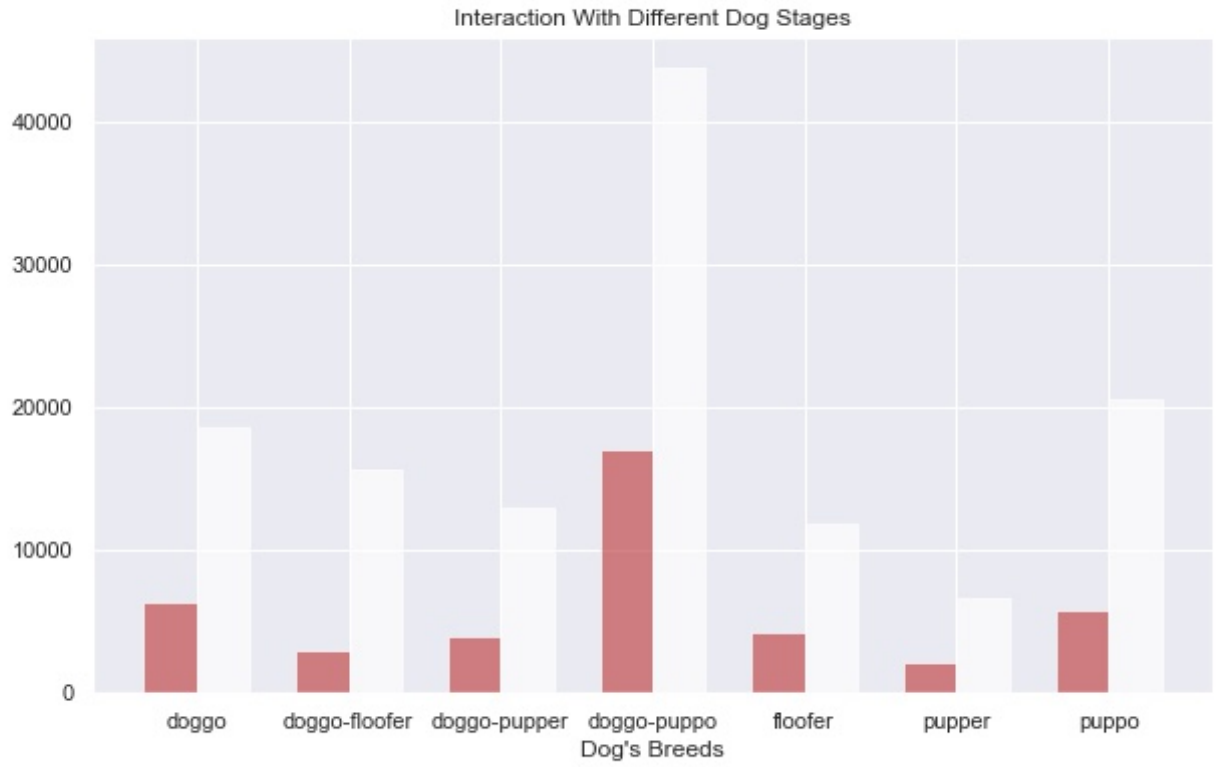


Conclusion

- We can see `pupper` is found maximum (201), followed by `doggo` (62).

5. What's the interaction with different dog stages in terms of retweet count average and favorite count average?

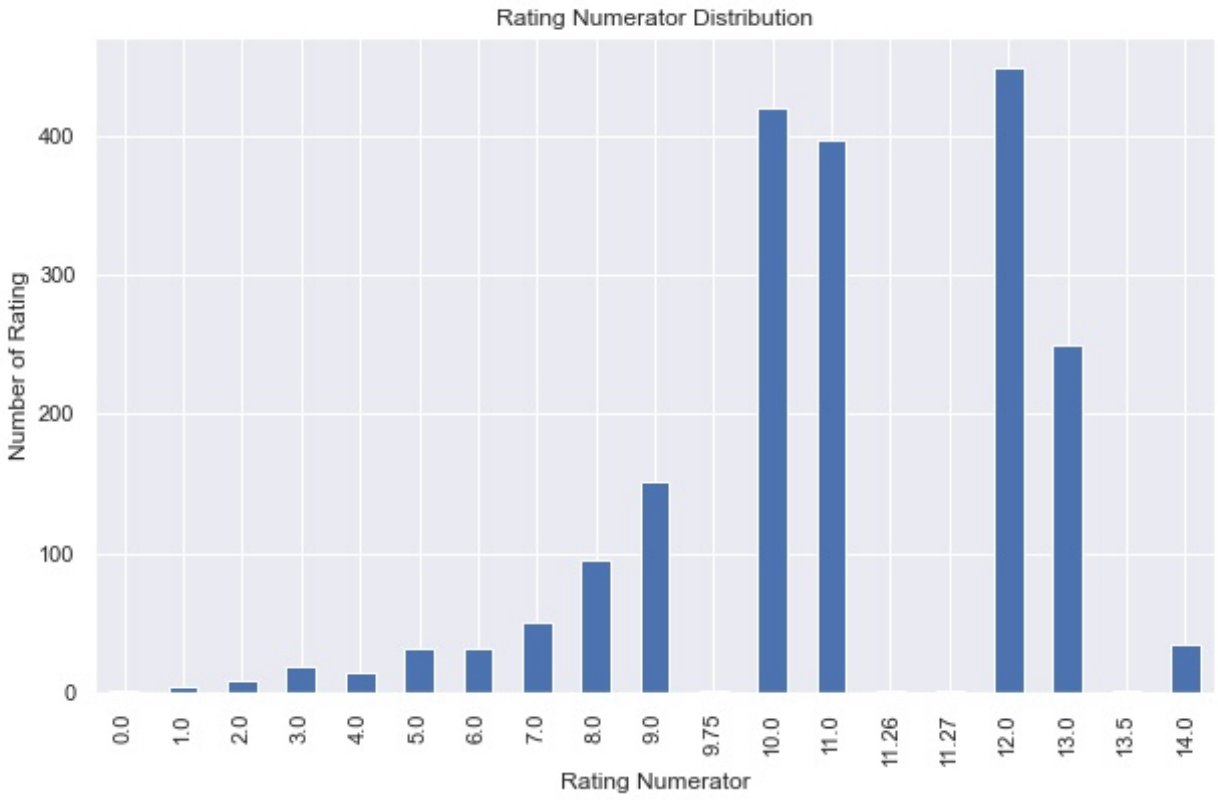
```
dog_stages
doggo            6405.967742
doggo-floofer    2999.000000
doggo-pupper     3962.125000
doggo-puppo      17092.000000
floofer          4267.571429
pupper           2067.497512
puppo            5712.000000
Name: retweet_count, dtype: float64
--------------------------------------------------------------------------------
dog_stages
doggo            18570.903226
doggo-floofer    15604.000000
doggo-pupper     13024.250000
doggo-puppo      43794.000000
floofer          11844.857143
pupper           6601.940299
puppo            20604.590909
Name: favorite_count, dtype: float64
```



Conclusion

- `doggo_puppo` breed is the most retweeted and favorited dog stage.

6. What's the distribution of rating numerator?



Conclusion

- We see the most assigned numerator is `12`.

7. How many tweets rated above 9?

```
1551 tweets are rated above 9.
```

8. How many tweets rated between 10 and 5?

```
362 tweets are rated between 10 and 5.
```

9. How many tweets have low rating?

```
48 tweets are rated under 5.
```

10. How is the change between retweet count and favorite depending on rating numerator?

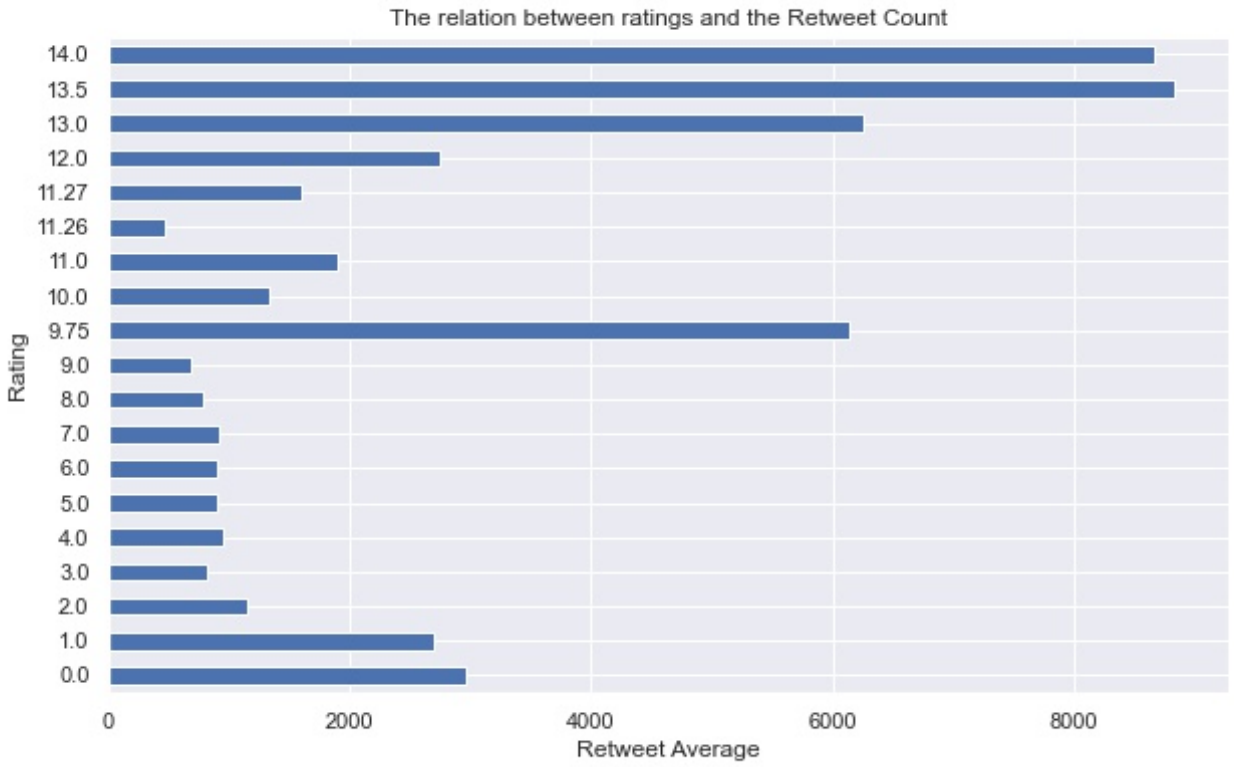
```
Retweet Count mean for rating numerators above 9 is 2841.1798839458415.
Retweet Count mean for rating numerators between 10 to 5 is 794.4337016574585.
Retweet Count mean for rating numerators under 5 is 1117.7291666666667.

Favorite Count mean for rating numerators above 9 is 9708.976144422953.
Favorite Count mean for rating numerators between 10 to 5 is 2316.7127071823206.
Favorite Count mean rating numerators under 5 is 2970.5416666666665.
```

Conclusion

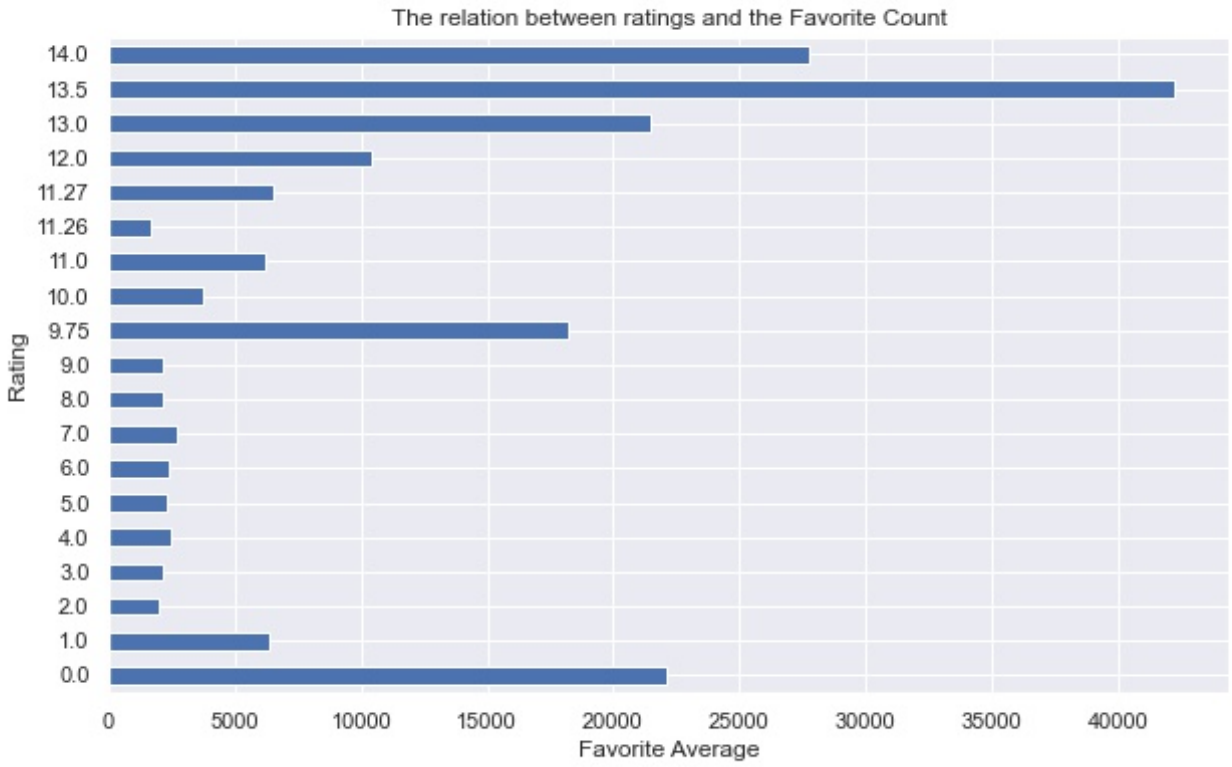
- if your dog got a rating numerator `above 10` there is a good chance your dog will get more likes.

11. What's the relation between rating and retweets and favorites?

```
rating_numerator
0.00      2954.000000
1.00      2693.250000
2.00      1147.444444
3.00      813.421053
4.00      942.800000
5.00      898.843750
6.00      893.031250
7.00      912.882353
8.00      773.736842
9.00      689.079470
9.75      6132.000000
10.00     1325.264916
11.00     1903.085642
11.26     473.000000
11.27     1604.000000
12.00     2738.533482
13.00     6253.372000
13.50     8840.000000
14.00     8668.705882
Name: retweet_count, dtype: float64
```



```
rating_numerator
0.00      22090.000000
1.00      6383.000000
2.00      1995.000000
3.00      2122.578947
4.00      2445.333333
5.00      2334.906250
6.00      2419.625000
7.00      2692.705882
8.00      2185.410526
9.00      2141.178808
9.75      18245.000000
10.00     3767.868735
11.00     6193.256927
11.26     1676.000000
11.27     6524.000000
12.00     10384.301339
13.00     21497.984000
13.50     42268.000000
14.00     27765.617647
Name: favorite_count, dtype: float64
```

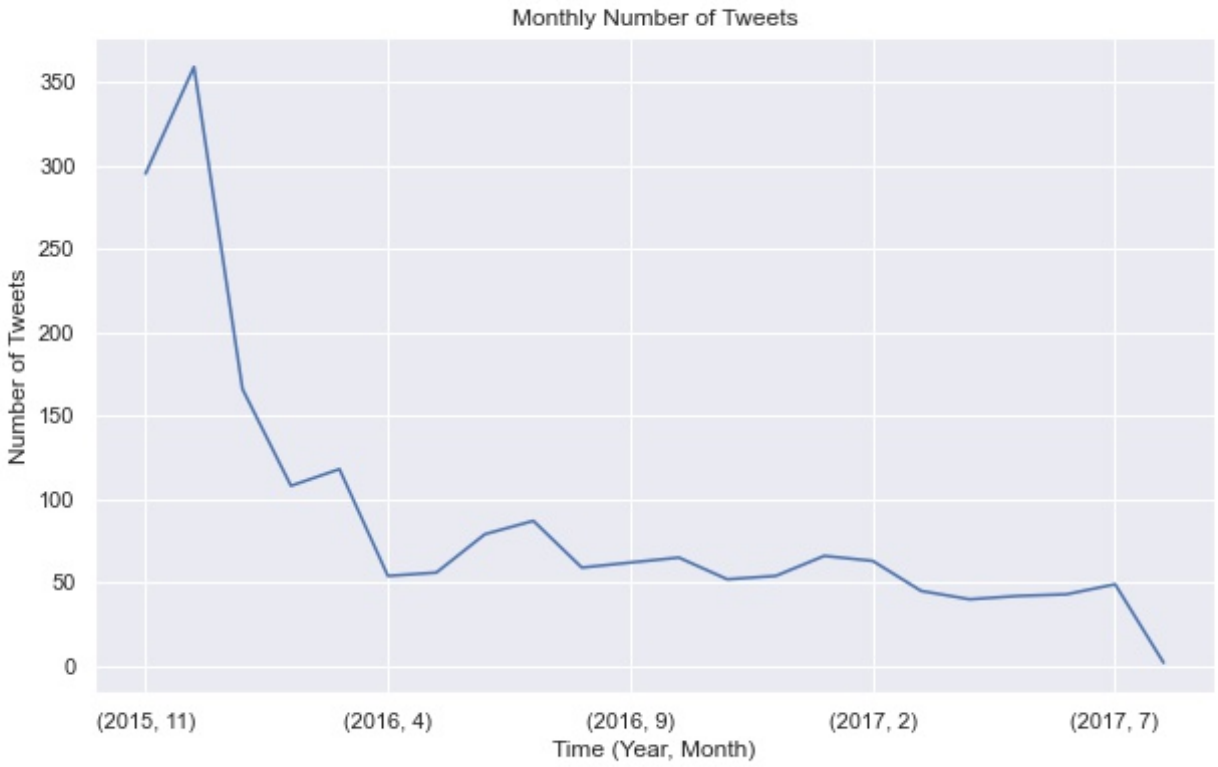


Conclusion

- We see the degree of audience interaction with tweets.
- With `high rating`, we get a good chance your dog will get more likes and retweets.

11. What is number of tweets monthly?

```
timestamp  timestamp
2015       11           295
           12           359
2016       1            166
           2            108
           3            118
           4            54
           5            56
           6            79
           7            87
           8            59
           9            62
           10           65
           11           52
           12           54
2017       1            66
           2            63
           3            45
           4            40
           5            42
           6            43
           7            49
           8            2
Name: tweet_id, dtype: int64
```



Conclusion

We see the most tweets were posted in `December 2015` (359 tweets). The number of tweets decreased rapidly `April 2016` and remained fairly constant since then until `July 2017`.