WeRateDogs Act Report

The following insights and visualizations are deduced from three datasets; twitter archive, image predictions and tweet json which were wrangled and merged into a master data.

Insights

1. About 570 dogs that is 29% of dogs have no name.

```
In [90]: masterdata.name.value_counts() / masterdata.shape[0] * 100

Out[90]: None          29.156010
         Charlie        0.511509
         Cooper         0.511509
         Penny          0.511509
         Oliver         0.511509
                          ...
         Rocco          0.051151
         Fido           0.051151
         Emma           0.051151
         Luna           0.051151
         Christoper     0.051151
         Name: name, Length: 912, dtype: float64

In [99]: masterdata.name.value_counts()

Out[99]: None          570
         Charlie        10
         Cooper         10
         Penny          10
         Oliver         10
                       ...
         Rocco           1
         Fido            1
         Emma            1
         Luna            1
         Christoper      1
         Name: name, Length: 912, dtype: int64
```

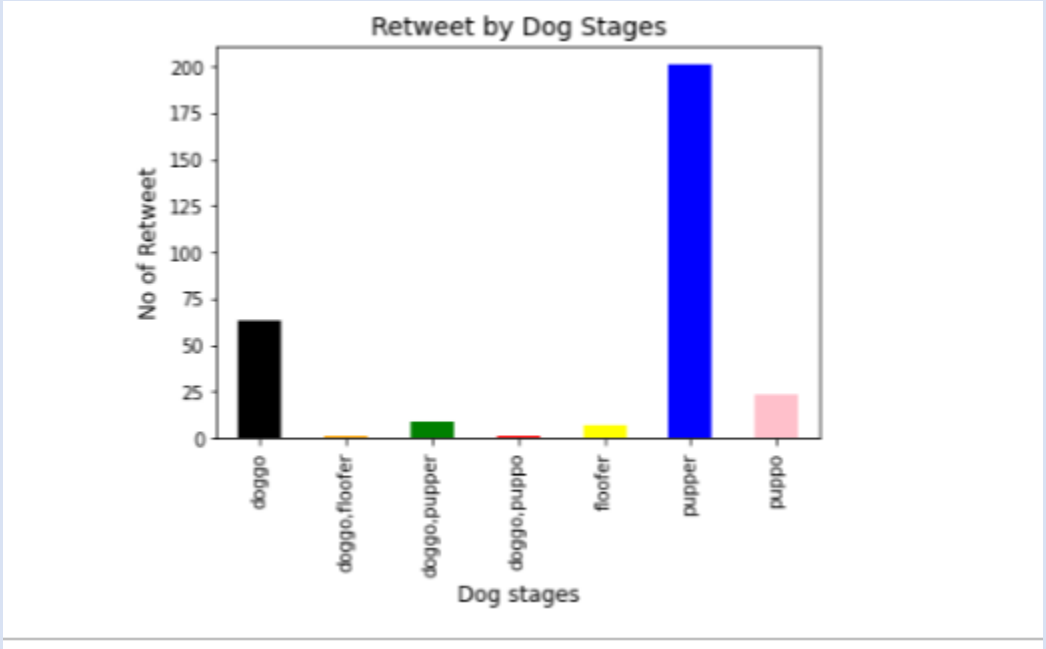
	tweet_id	rating_numerator	rating_denominator	img_num	p1_conf	p2_conf	p3_conf	retweet_count	favorite_count
count	1.887000e+03	1887.000000	1887.000000	1887.000000	1887.000000	1.887000e+03	1.887000e+03	1887.000000	1887.000000
mean	7.373672e+17	12.075782	10.460519	1.207737	0.594639	1.338589e-01	6.018044e-02	2254.919979	7793.210917
std	6.772969e+16	41.353955	6.501114	0.565239	0.272992	1.004429e-01	5.101972e-02	4022.554466	11346.884747
min	6.660209e+17	0.000000	7.000000	1.000000	0.044333	1.011300e-08	1.740170e-10	11.000000	66.000000
25%	6.766051e+17	10.000000	10.000000	1.000000	0.360115	5.388625e-02	1.594050e-02	511.000000	1731.000000
50%	7.099013e+17	11.000000	10.000000	1.000000	0.587764	1.169770e-01	4.909330e-02	1097.000000	3522.000000
75%	7.906404e+17	12.000000	10.000000	1.000000	0.849540	1.941255e-01	9.217055e-02	2580.500000	9769.000000
max	8.924206e+17	1776.000000	150.000000	4.000000	1.000000	4.880140e-01	2.734190e-01	70335.000000	144253.000000

Insights:

2. The merged data has 21 rows and 1887 columns. All columns have no missing entries apart from the dog_stage.
3. Image number 1 is the most prominent
4. The merged dataset has 21 columns and 1955 rows,
5. All the rows except for the dog stage column are completely filed with no missing value.
6. The columns are 'tweet_id', 'tweet_time, 'source', 'text', 'expanded_urls', 'rating_numerator', 'rating_denominator', 'name', 'stage', 'retweet_count', 'favorite_count', 'jpg_url', 'img_num', 'p1', 'p1_conf', 'p1_dog', 'p2', 'p2_conf', 'p2_dog', 'p3', 'p3_conf', 'p3_dog'.

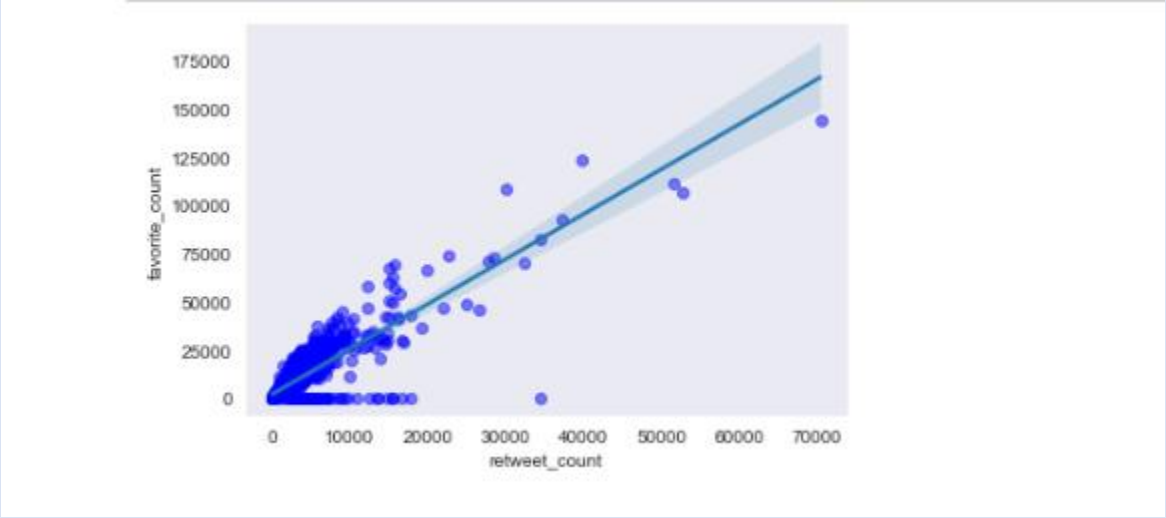
Visualizations

Question 1: What is the most popular dog stage according to retweets?



The graph above shows that the pupper stage (usually a small dog) is the most popular dog stage followed by doggo, then puppo'.

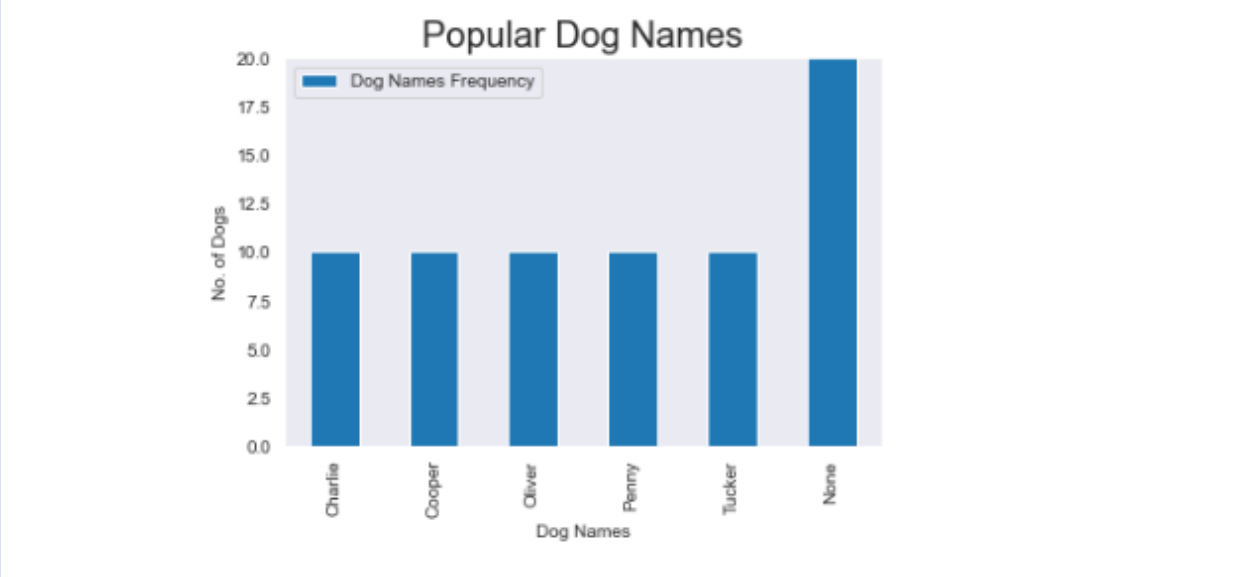
Question 2: Does retweet count positively correlate with favourite count?



From the visual representation above, there is a linear relationship between the two varibales. This does not imply an increase in retweet_count will cause an increase in favorite_count but when you compare both linearly, there is a strong positive linear relationship between retweet_count and favorite_Count.

This obviously shows that retweet and likes contributed to rating of dogs.

Question 3: What is the most popular dog name?



Turns out the most popular dog names in this master dataset are Charlie, Cooper, Oliver, Penny and Tucker.