Act Report

Intro:

In this project I gathered, assessed and analyzed sets of data from Udacity and Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. Software used include: Python(pandas, NumPy, requests, tweepy, json), Google Docs.

Files provided ahead include: 'twitter-archive-enhanced-2.csv', image-predictions.tsv, tweet-json.txt. The archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017. 'image-predictions' is a file contains prediction result of whether the object in the picture is a dog or not as well as dog breed. It was powered by a neural network that can classify breeds of dogs and ran through every image in the WeRateDogs Twitter. 'tweet-json' was provided as additional file that contains all information in every tweet.

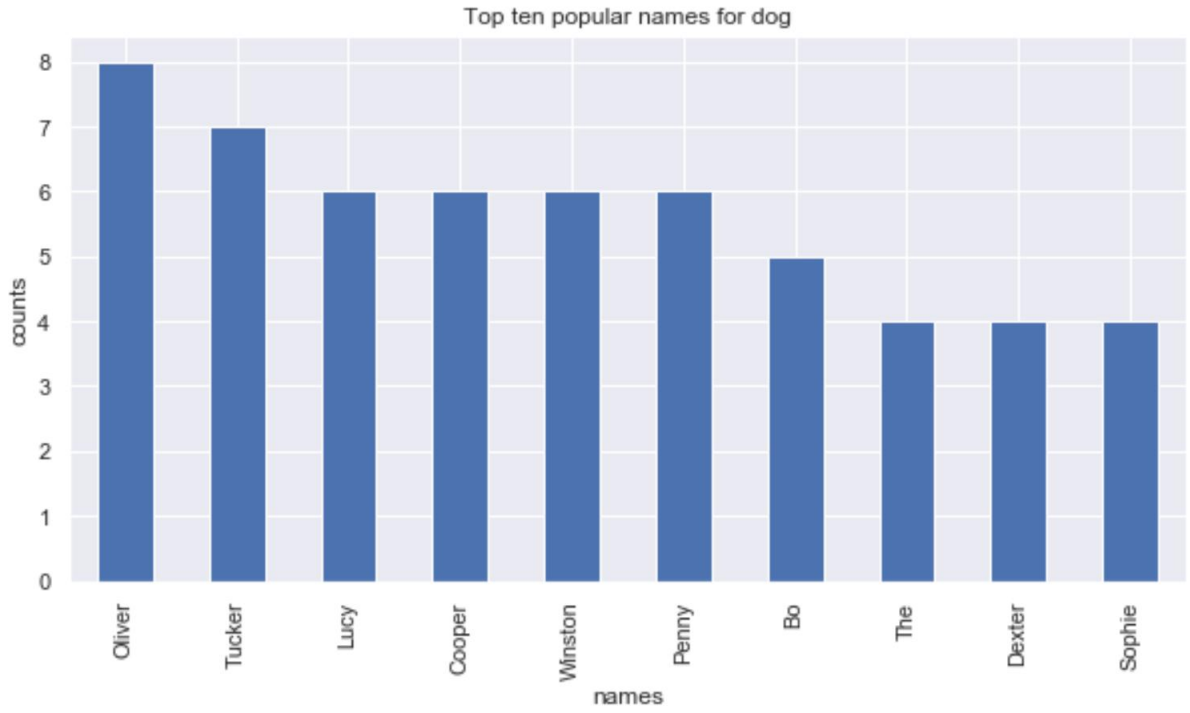
Insight:

We can have an overall insight about statistical results here. On average, those tweets was retweeted 2637 times, was 'liked' 8747 times, the average quotient, although giving the fact 'those are good dogs' and expect quotients would be way above 1, there are few were rated under 1, make up the average quotient round 1. This also make sense consider 50% quotient is 1.1 or less.

	tweet_id	retweet_count	favorite_count	retweeted_status_user_id	rating_numerator	rating_denominator	quotient
count	1.140000e+03	1140.000000	1140.000000	4.000000e+01	1140.000000	1140.000000	1140.000000
mean	7.386192e+17	2911.939474	8440.280702	4.311139e+09	11.352211	10.481579	1.082615
std	6.733339e+16	4750.252658	11430.204431	5.896539e+08	8.014061	6.949499	0.192088
min	6.660293e+17	16.000000	0.000000	4.196984e+09	1.000000	2.000000	0.200000
25%	6.766059e+17	620.500000	1662.000000	4.196984e+09	10.000000	10.000000	1.000000
50%	7.140304e+17	1440.500000	3946.500000	4.196984e+09	11.000000	10.000000	1.100000
75%	7.909661e+17	3443.250000	10555.750000	4.196984e+09	12.000000	10.000000	1.200000
max	8.918152e+17	56625.000000	107015.000000	7.832140e+09	165.000000	150.000000	3.428571

Q: What is the most common name for dogs? What are the top 10 names?

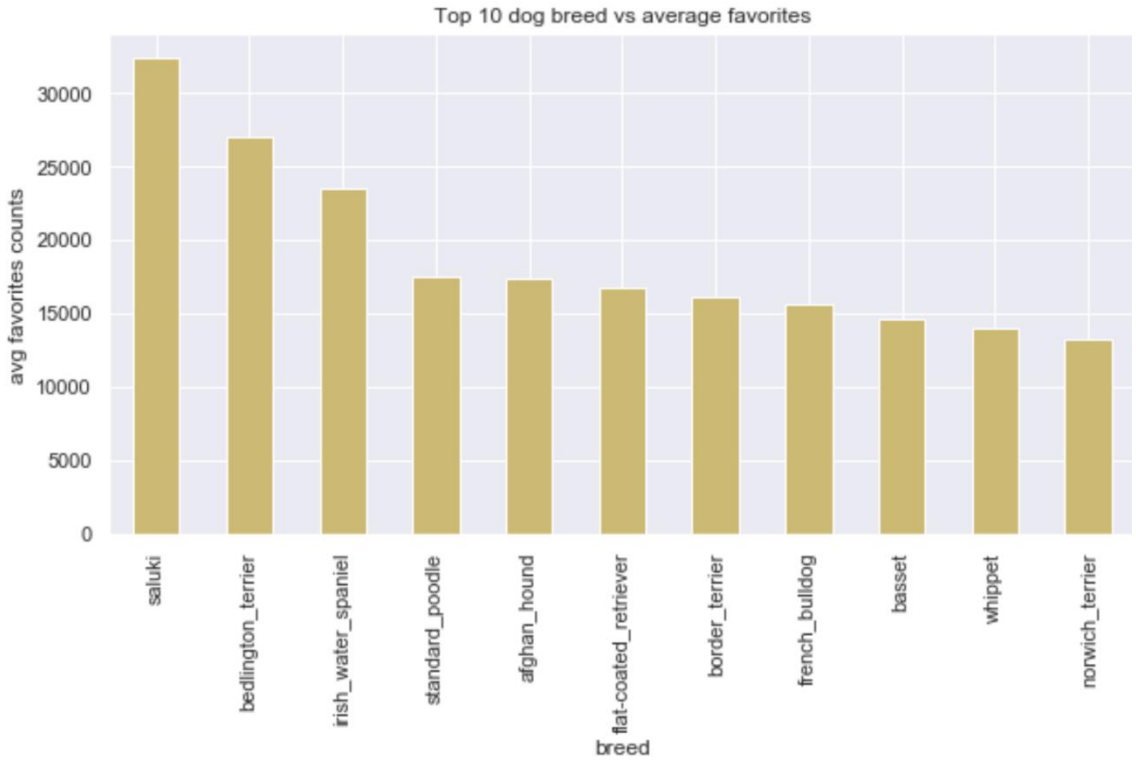
A: Regardless invalid names, the most common name people like to name their dogs seems to be 'Oliver', followed by ‘Tucker’, ’Lucky’.

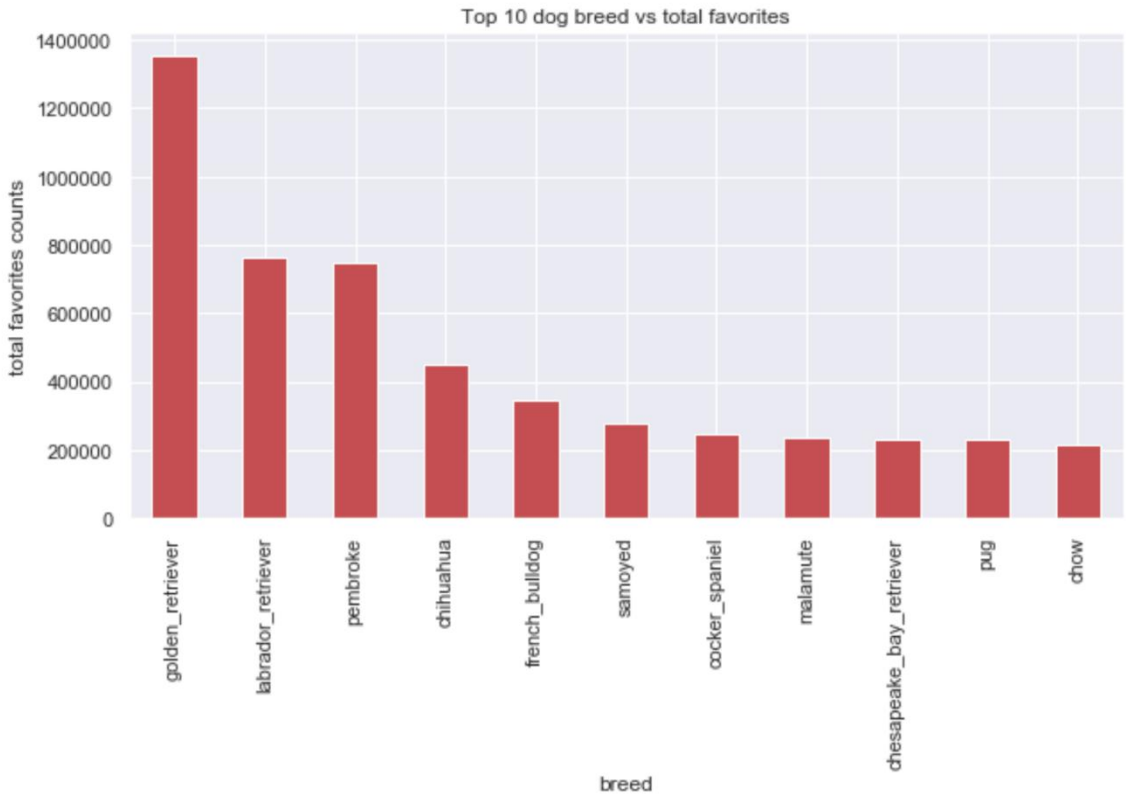


Q: Breed distribution vs average/total favorites?

A:

Here we can see the difference of 'top 10 most favorite dog breeds' when we use different metrics. On average, every saluki dog post earned 32444 'favorites' as the most popular dog breed, while golden retriever seems to be the most popular dog breed when we compare the total ‘favorites’ earned over time which is at 1429448 ‘favorites’. This could due to the fact the denominator: total number of saluki dog post is very small, drove the result to be higher given the fact saluki is not as commonly seen as some other breeds. Some interesting facts and compares I found on dog-learn.com. Which one is your favourite?


Top 10 dog breed vs average favorites
avg favorites counts
30000
25000
20000
15000
10000
5000
0
saluki
bedlington_terrier
irish_water_spaniel
standard_poodle
afghan_hound
flat-coated_retriever
border_terrier
french_bulldog
basset
whippet
norwich_terrier
breed


Top 10 dog breed vs total favorites
total favorites counts
1400000
1200000
1000000
800000
600000
400000
200000
0
golden_retriever
labrador_retriever
pembroke
chihuahua
french_bulldog
samoyed
cocker_spaniel
malamute
chesapeake_bay_retriever
pug
chow
breed

Golden Retriever Saluki

https://www.dog-learn.com/breed-vs-breed/golden-retriever-vs-saluki/

The fact that different metric resulted differently lead me to a hypothesis: there is a potential correlationship between retweet_count, favorite_count and possibly different dog stages as well?

Q: Correlationship between retweet_count, favorite_count and different dog stages?

A:

	coef
intercept	1852.5463
retweet_count	2.3897
puppo	3471.5877
doggo	1013.1004
floofer	277.4183

From the model we can tell retweet_count and favorite_count has a very strong positive correlationship. Every additional retweet is likely to result more than 2 'favorite'. With everything else being equal, we expect dogs at 'puppo' stage receive 3471 more 'favorites' than dogs at 'pupper' stage(baseline), 'doggo' receive 1013 more 'favorites' than 'pupper', 'floofer' receive 277 more 'favorites' than 'pupper'.

Limitation:

sample size is not large (189 counts), leads to inaccurate prediction. On the other hand, p-value and scatter plot indicate different dog stage don't have strong correlation with number of 'favorite'.Limitation: sample size is not large (197 counts), leads to inaccurate prediction. On the other hand, p-value and scatter plot indicate different dog stage don't have strong correlation with number of 'favorite'.