Udacity - Analysis and Insights Documentation

Introduction

To take a dip into the analysis phase after wrangling is always a rewarding experience where the data which we have wrangled is now made ready to answer some of the complex questions regarding the dataset. These finding give out a brand new insights and also makes us question fields and metrics which we didn't know it existed when just looking at the same data via Tweeter, e.g., via scrolling through the WeRateDogs Hash tag we wouldn't be able to judge as to how many different types of Dog breeds are being published in that forum or what is the average or total like or retweet count for that particular dog breed.

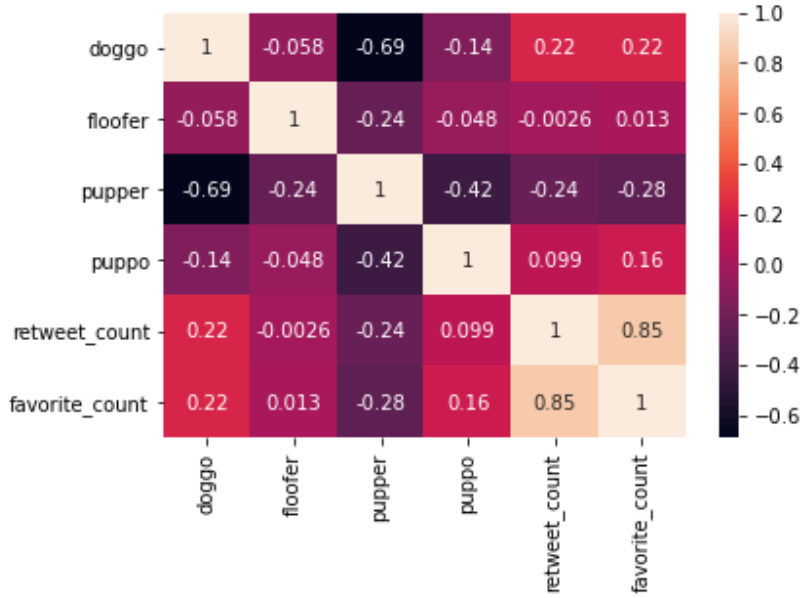
For analysis for this project we will be looking into the below questions.

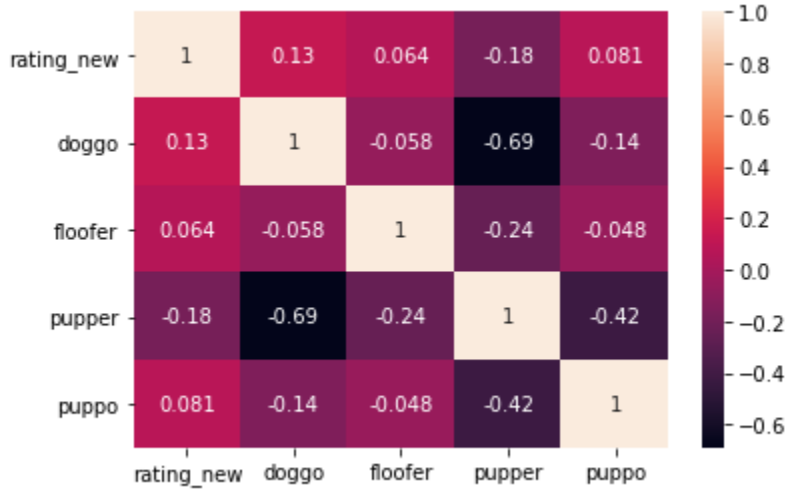
Is there any co-relation between usages of the Dog terms to the rating?

In this analysis we are first creating a column called rating_new which will basically help us identify if the given tweet ID is categorized as either Good or Bad rating, where False denotes Bad and True denotes good.

Also we shall only be considering only those tweets which has dog terms usage in them and shall be ignoring if they don't have it categorized.

	tweet_id	rating_new	rating_numerator	rating_denominator	doggo	floofer	pupper	puppo	retweet_count	favorite_count
9	890240255349198849	True	14	10	1	0	0	0	6210.0	28467.0
12	889665388333682689	True	13	10	0	0	0	1	8523.0	42826.0
14	889531135344209921	True	13	10	0	0	0	1	1915.0	13579.0
29	886366144734445568	True	12	10	0	0	1	0	2687.0	18891.0
42	884162670584377345	True	12	10	1	0	0	0	2536.0	18253.0

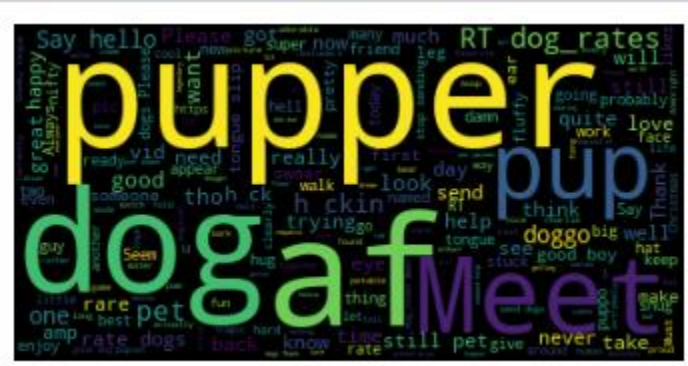




From the above mentioned correlation graph we can have the below conclusions

- There is a very weak relation between Usages of dog terms to the rating.
    1. As the word “puppo” is only having a 8% correlation with the rating metric.
    2. The rest of the dog terms is having a range between -2% to 1.3%
- Usage of dog terms can be compared with effect it has on retweet and favorite count
- Drawback – the Dog term categorized column is only subjected to only about 374 records out of 2297 records which is very low.

What is the most common word in the tweet for both negative and positive rated tweets?



We see that the usage of the word "*pupper"* is the most frequent in tweets. Hence a lot of users like to address cute or small dogs with the term pupper which looks to be a common trend in the channel.

Which is the favorite dog breed among the users?

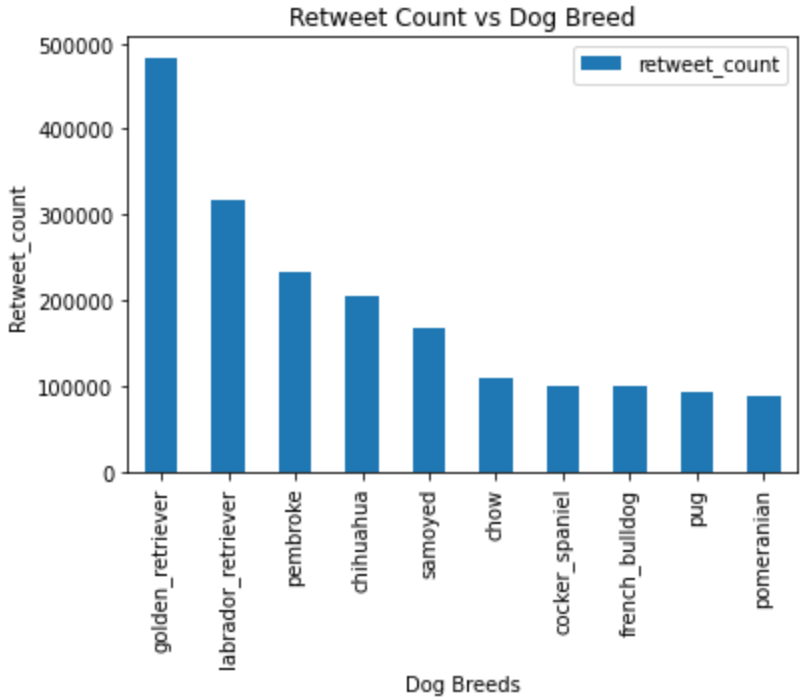
There are 118 different unique Dog breeds in the given data sets!
For this we shall first be making use of the image prediction data and classify only those records where the prediction was made true where the image consisted of a Dog. It also gives us the Dog breed if the identified image is a dog. This is particularly useful in identifying the Dog breed for each Tweet ID.

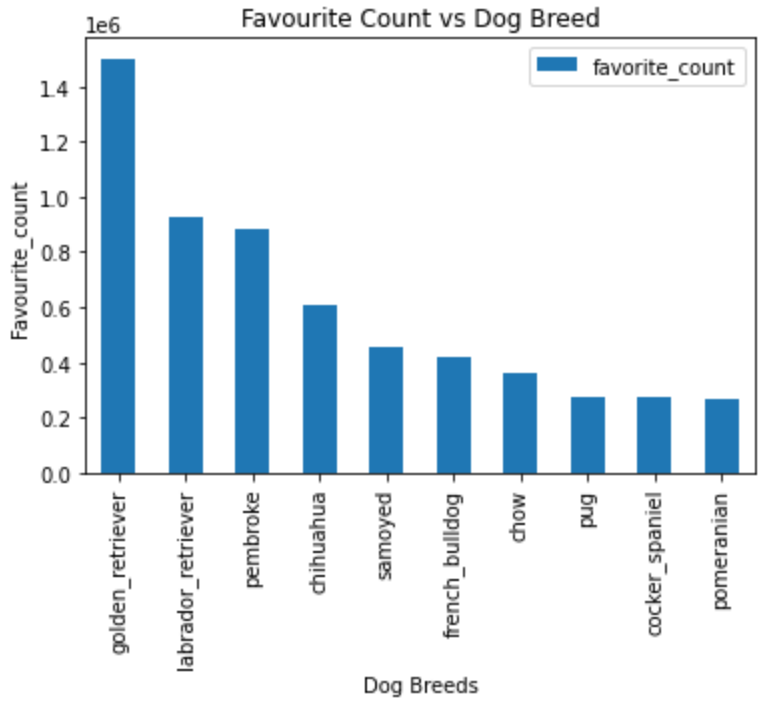
```
In [93]: df[['p1_conf', 'p2_conf','p3_conf']].describe()
```

Out[93]:

	p1_conf	p2_conf	p3_conf
count	2075.000000	2.075000e+03	2.075000e+03
mean	0.594548	1.345886e-01	6.032417e-02
std	0.271174	1.006657e-01	5.090593e-02
min	0.044333	1.011300e-08	1.740170e-10
25%	0.364412	5.388625e-02	1.622240e-02
50%	0.588230	1.181810e-01	4.944380e-02
75%	0.843855	1.955655e-01	9.180755e-02
max	1.000000	4.880140e-01	2.734190e-01

It is safe to take p1 as p2 and p3 max prediction is just 48% and on average give between 6-13% prediction score, whereas p1 has 100% max and on average of 59% prediction score.



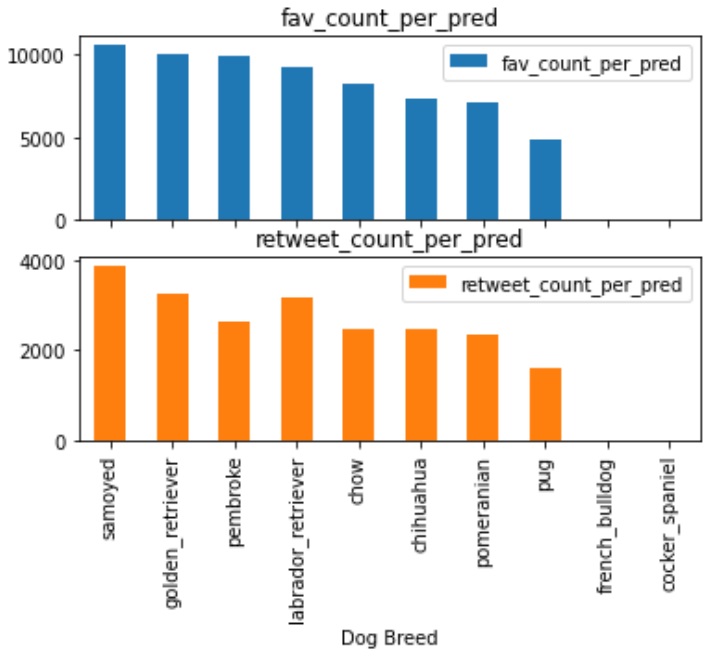


From the above we find that the Golden Retriever has the highest count of both Favourite and also the most retweeted from the channel.

Which Dog breed has the best favourited/retweeted based on its occurrence in the tweet?

To get the result for this data let us look into the number of occurances for that particular dog breed using the tweet ID and the p1 columns where we find the number of times that particular dog breed has appeared. Then we shall then normalize the favourite count and the retweeted count, to identify which Dog breed has a really good popularity amongst the user of that channel.

p1	favorite_count	retweet_count	No_of_dogs	fav_count_per_pred	retweet_count_per_pred
golden_retriever	1502814.0	483586.0	150.0	10018.760000	3223.906667
labrador_retriever	927478.0	316244.0	100.0	9274.780000	3162.440000
pembroke	883411.0	232296.0	89.0	9925.966292	2610.067416
chihuahua	609837.0	204885.0	83.0	7347.433735	2468.493976
samoyed	455909.0	166870.0	43.0	10602.534884	3880.697674
french_bulldog	415944.0	100202.0	NaN	NaN	NaN
chow	360064.0	108790.0	44.0	8183.272727	2472.500000
pug	276438.0	91612.0	57.0	4849.789474	1607.228070
cocker_spaniel	273111.0	100447.0	NaN	NaN	NaN
pomeranian	269329.0	89022.0	38.0	7087.605263	2342.684211



From the previous question we found that the Golden Retriever had the best metrics in terms of total favorite and retweet counts. But surprisingly we find that it is in second place when judged by the appearance of it. We see that Samoyed dog breed has the best ratio of popularity metric to its

occurrence. Samoyed beats the Golden Retriever in both favourite count/retweeted count per occurance by approximately ~600.

Then again we see that the Samoyed Dog breed is only mentioned/identified by the image detection model in only 43 tweets in total with respect to the 150 of the Golden Retriever prediction count. Overall the conclusion being, Samoyed Dog breed has a lot of potential to actually dethrone the supposed Golden Retriever's popularity in the channel as it looks to be talked/posted about it less in the community. But then