Module 4: Assessing dog breeds for maximum social media outreach

I will devise a model to predict how many favorite_counts a dog of a given breed is likely to get based on the breed, confidence level of predictions and rating. I'll also do a similar analysis for retweet_count.
On cleaning the data, I can conclude that any kind of model I build on top of these predictions would not be a good one since we don't have significant number of observations for a particular breeds/predictions. We can however, try analysing the top predictions appearing in the dataframe. Through this we can get an estimate of what breeds tend to get submitted to WeRateDogs. We can use this to build a flawed yet functional estimator for the metric we desire to obtain. This is the billionth time I wish we had more observations and less variability!!

I will take toy_poodle and Unidentified as the baseline variables. The features planned to be included in our analysis are:

1. absolute_ratings which is rating_numerator/rating_denominator
2. Dog breeds with toy poodle as the baseline variable
3. Categories that the dog is presumed to belong to.

1. Fitting Linear Models to the Data

On fitting a linear regression to the data, we find the following coefficients:

OLS Regression Results

Dep. Variable:	favorite_count	R-squared:	0.159
Model:	OLS	Adj. R-squared:	0.138
Method:	Least Squares	F-statistic:	7.352
Date:	Mon, 21 Jun 2021	Prob (F-statistic):	7.24e-16
Time:	16:28:09	Log-Likelihood:	-6862.0
No. Observations:	637	AIC:	1.376e+04
Df Residuals:	620	BIC:	1.383e+04
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Chihuahua	2560.9704	2334.648	1.097	0.273	-2023.806	7145.747
Labrador_retriever	3723.4424	2281.231	1.632	0.103	-756.434	8203.319
Pembroke	3316.1377	2295.036	1.445	0.149	-1190.848	7823.124
Pomeranian	740.4927	2709.206	0.273	0.785	-4579.840	6060.826
Samoyed	4461.3787	2703.463	1.650	0.099	-847.675	9770.433
chow	1599.6771	2665.695	0.600	0.549	-3635.208	6834.563
golden_retriever	2857.4568	2184.752	1.308	0.191	-1432.953	7147.867
malamute	2163.8879	2906.585	0.744	0.457	-3544.056	7871.832
pug	1471.3051	2506.525	0.587	0.557	-3451.002	6393.613
pupper_1	1.181e+04	2768.912	4.266	0.000	6374.096	1.72e+04
doggo_1	2640.6055	5324.957	0.496	0.620	-7816.531	1.31e+04
puppo_1	-2509.2087	1480.441	-1.695	0.091	-5416.494	398.077
floofer_1	4395.7144	4472.080	0.983	0.326	-4386.545	1.32e+04
pupper_2	-6219.9595	5899.288	-1.054	0.292	-1.78e+04	5365.048
puppo_2	-1.048e+04	1.2e+04	-0.872	0.384	-3.41e+04	1.31e+04
absolute_ratings	2.251e+04	2718.453	8.280	0.000	1.72e+04	2.78e+04
intercept	-1.898e+04	3573.523	-5.311	0.000	-2.6e+04	-1.2e+04

Omnibus:	600.486	Durbin-Watson:	1.586
Prob(Omnibus):	0.000	Jarque-Bera (JB):	30560.655
Skew:	4.053	Prob(JB):	0.00
Kurtosis:	35.950	Cond. No.	40.5

Most of the p values are above the alpha value (assumed 5% in this case). This means we cannot reject the null hypothesis that breeds other than toy_poodle, categories other than Unidentified's result in greater favorite counts. We can however roughly gauge which features are negatively and positively correlated with

the dependent variable with respect to the baseline. This inabilility to draw conclusions stems from the fact that we don't have enough data

OLS Regression Results

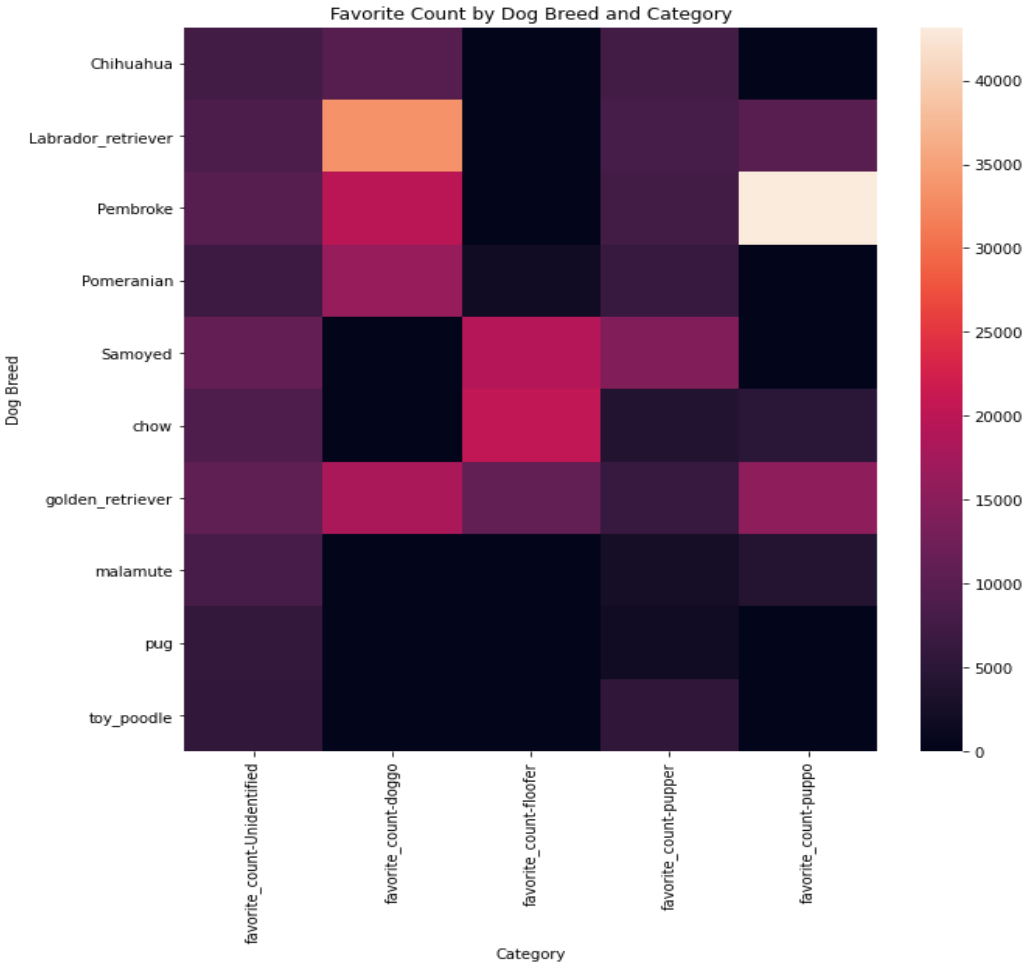
Dep. Variable:	retweet_count	**R-squared:**	0.100
Model:	OLS	**Adj. R-squared:**	0.077
Method:	Least Squares	**F-statistic:**	4.318
Date:	Mon, 21 Jun 2021	**Prob (F-statistic):**	4.51e-08
Time:	16:28:47	**Log-Likelihood:**	-6251.4
No. Observations:	637	**AIC:**	1.254e+04
Df Residuals:	620	**BIC:**	1.261e+04
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Chihuahua	471.4157	895.240	0.527	0.599	-1286.654	2229.485
Labrador_retriever	877.3844	874.756	1.003	0.316	-840.460	2595.229
Pembroke	338.7462	880.050	0.385	0.700	-1389.494	2066.986
Pomeranian	36.3127	1038.867	0.035	0.972	-2003.811	2076.437
Samoyed	1257.2360	1036.664	1.213	0.226	-778.563	3293.035
chow	9.2441	1022.182	0.009	0.993	-1998.115	2016.603
golden_retriever	300.2546	837.760	0.358	0.720	-1344.937	1945.447
malamute	204.3838	1114.553	0.183	0.855	-1984.373	2393.140
pug	7.7364	961.147	0.008	0.994	-1879.762	1895.234
pupper_1	4861.3571	1061.761	4.579	0.000	2776.273	6946.442
doggo_1	1746.9002	2041.897	0.856	0.393	-2262.972	5756.773
puppo_1	-368.7116	567.687	-0.649	0.516	-1483.533	746.110
floofer_1	197.1897	1714.855	0.115	0.908	-3170.438	3564.817
pupper_2	-2893.2706	2262.129	-1.279	0.201	-7335.634	1549.093
puppo_2	-5207.0652	4609.864	-1.130	0.259	-1.43e+04	3845.774
absolute_ratings	5850.1718	1042.412	5.612	0.000	3803.085	7897.259
intercept	-4431.7698	1370.296	-3.234	0.001	-7122.754	-1740.786

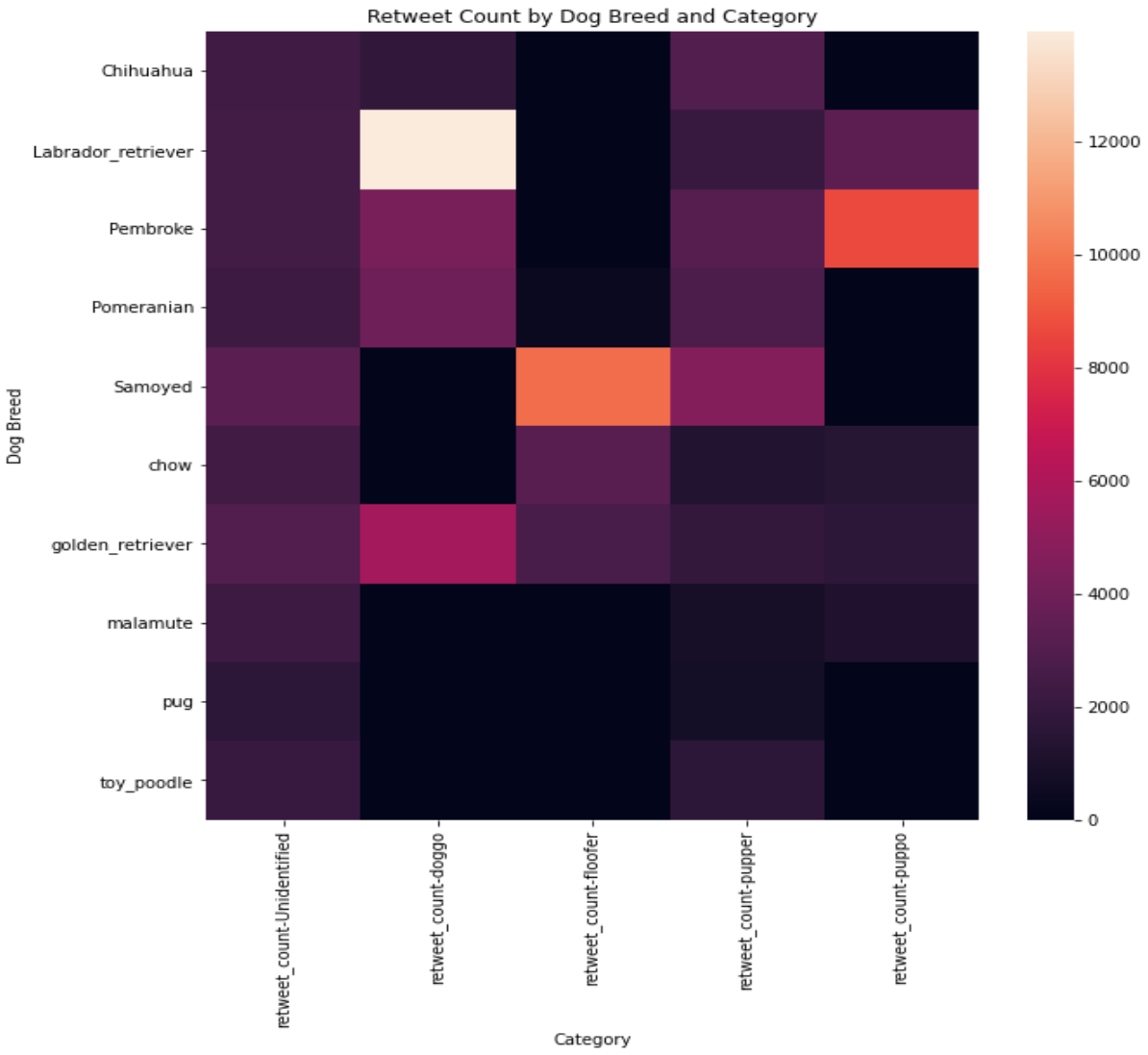
Omnibus:	920.850	**Durbin-Watson:**	1.894
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	248730.478
Skew:	7.789	**Prob(JB):**	0.00
Kurtosis:	98.544	**Cond. No.**	40.5

Here again, results are not statistically significant due to the same reasons

2. Visualizing Heatmap of Data Based on Dog Breed and Category



Here we can observe that Pembroke in the puppo stage receives maximum number of likes. Labrador Retriever in doggo stage receives considerable likes. The safest option to go by is the Golden Retriever since it receives consistent favorites throughout its stages



Similar stats are observed for retweet_count suggesting that retweet_count and favorite_count are positively correlated (pretty intuitive). Here as well Golden Retriever does consistently well across all its stages and Labrador retriever and Pembroke achieve high status in the pupper and puppo stages respectively.

One may attribute the drop in performance of other dog breeds in specific stages to a lack of data for those breeds in those stages. This is a perfectly valid observation.