Analysis of WeRateDogs

What is WeRateDogs?

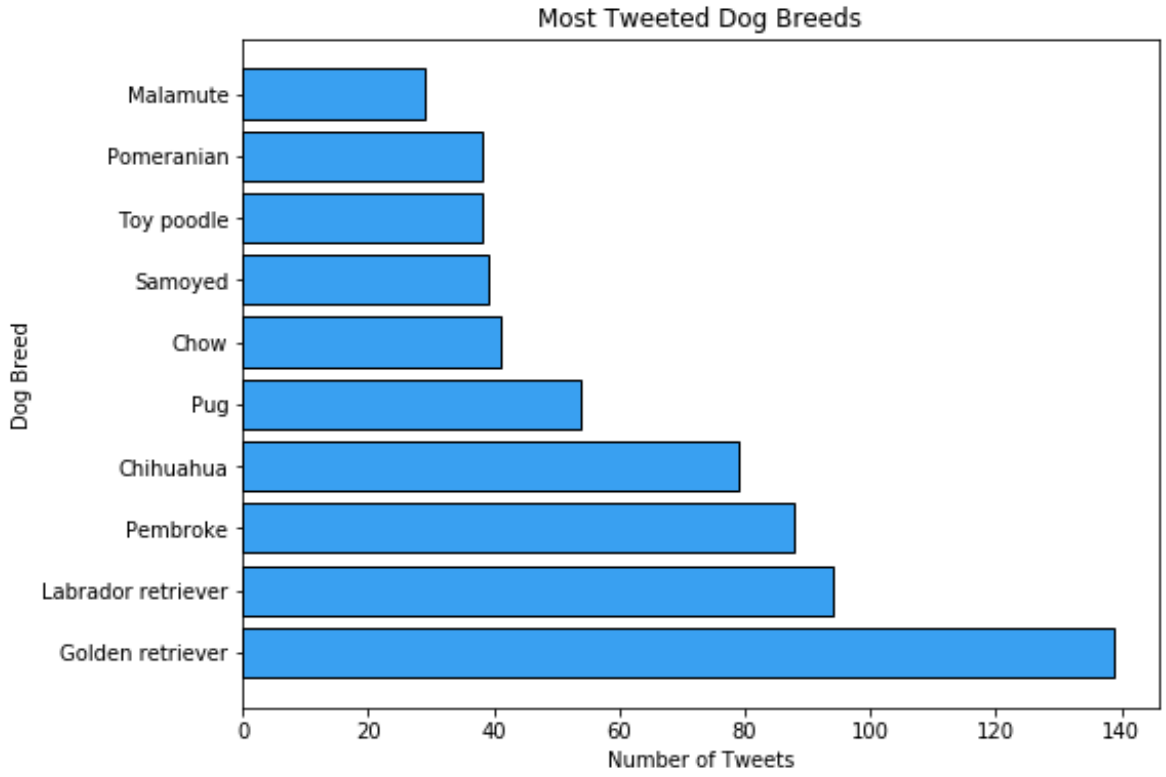
WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. WeRateDogs has over 4 million followers and has received international media coverage.

Analysis: Part I

After cleaning all the three files and curating it into one master dataset, we were equipped with information on 1683 tweets spanning from user details such as when they were tweeted, by who were they tweeted to dog details such as stage of dog, it's rating and its possible breed.

Our first analysis involved answering the very first and obvious question that came to our mind.
Which is the most famous dog breed among Twitter ?
Also, does more number of tweets targeting a certain breed lead to more amount of popularity among the Twitter population as well?

For this, we first looked at the Total sum of tweets made per Breed

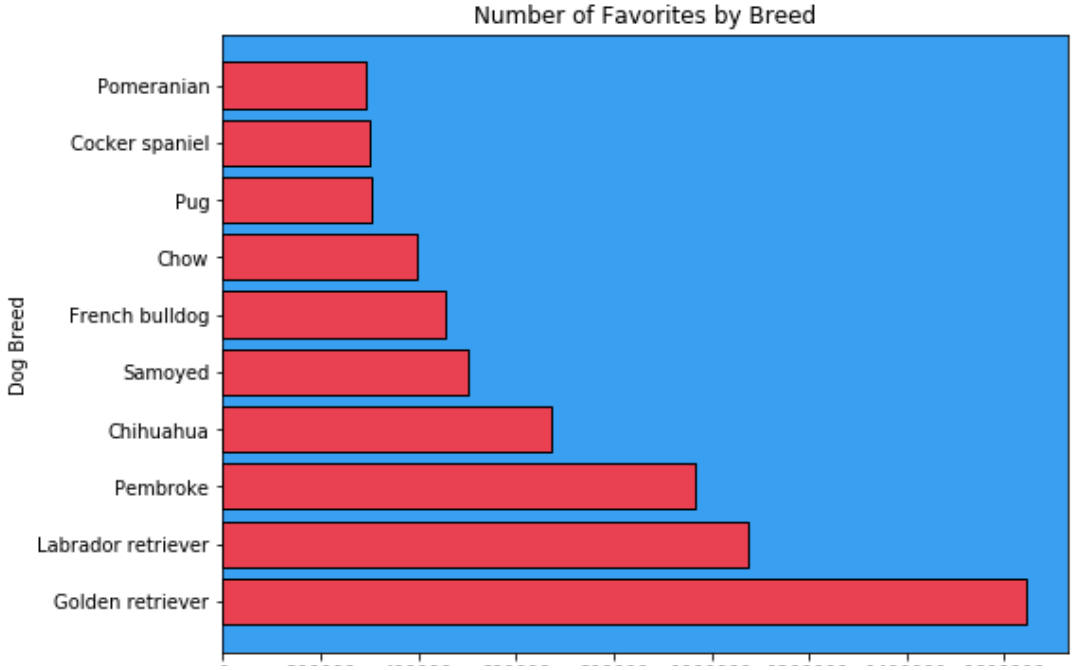


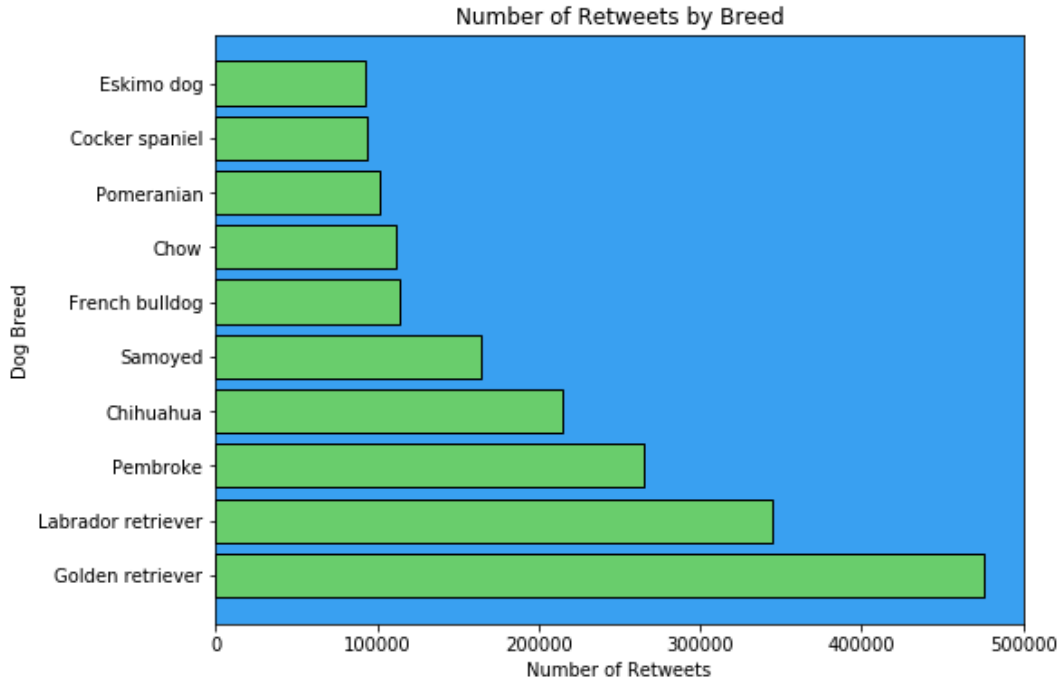
As we can see these are the most tweeted dog breeds when it comes to the @WeRateDogs page, and since this page is globally known (>4 million followers) and does not primarily focus on one type of breed, we can assume that the insights we have gained can hold true for the global population.

CONCLUSION

- The graph shows the Top 10 most tweeted about dog breeds. Golden Retrievers and Labrador Retrievers top the chart at 1st and 2nd place respectively
- If more number of tweets leads to more love for the breed among the twitter population, then atleast 50% of the dog breeds should remain on the top chart when we will be analysing the response to the tweets

The Metrics we used for measuring popularity of a breed among Twitter was the number of retweets , and the total number of favourites obtained by the breed.





CONCLUSION

- Every breed listed in the top 10 favorites list had atleast over 200K favorites and peaked at over 1.6 Million (Golden Retriever)

- Incase of Retweets, they ranged from 100K to just below 500K retweets

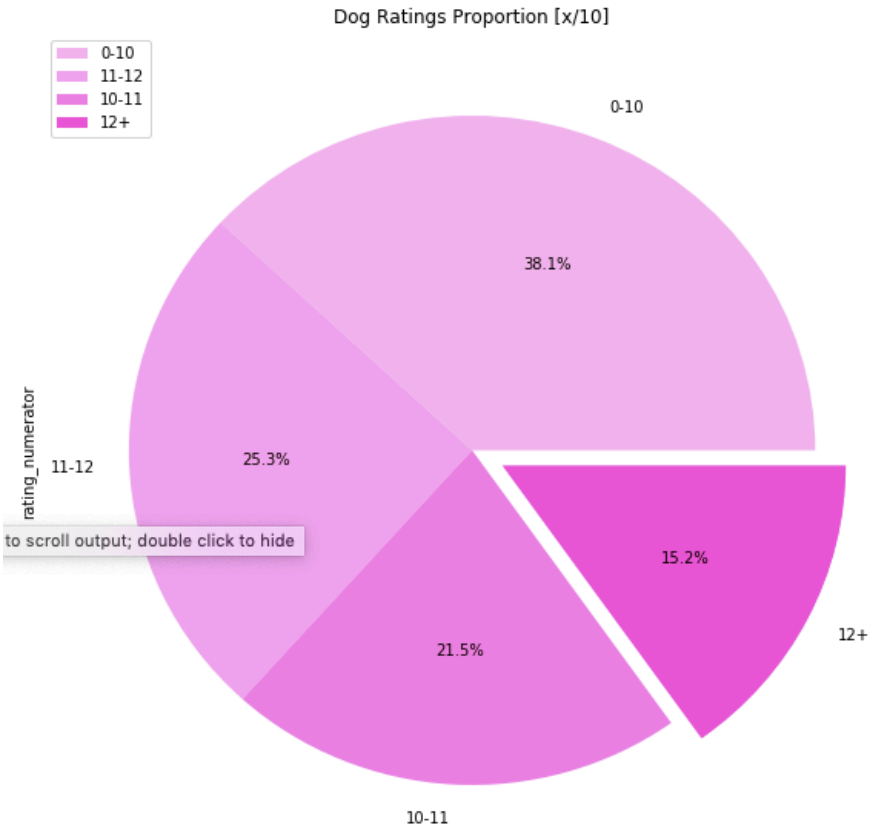
- Golden Retriever topped the chart in both metrics

- The intersection of breeds between all three top 10 (Tweets,Retweets,Favorites) retained 70% of the dog breeds mentioned in the Top 10 most tweeted dogs

[**Note**: The Colors used in the graph match with the colors used by Twitter on their website; Blue - logo, Green - Retweet , Pink - Favourite]

Analysis Part II

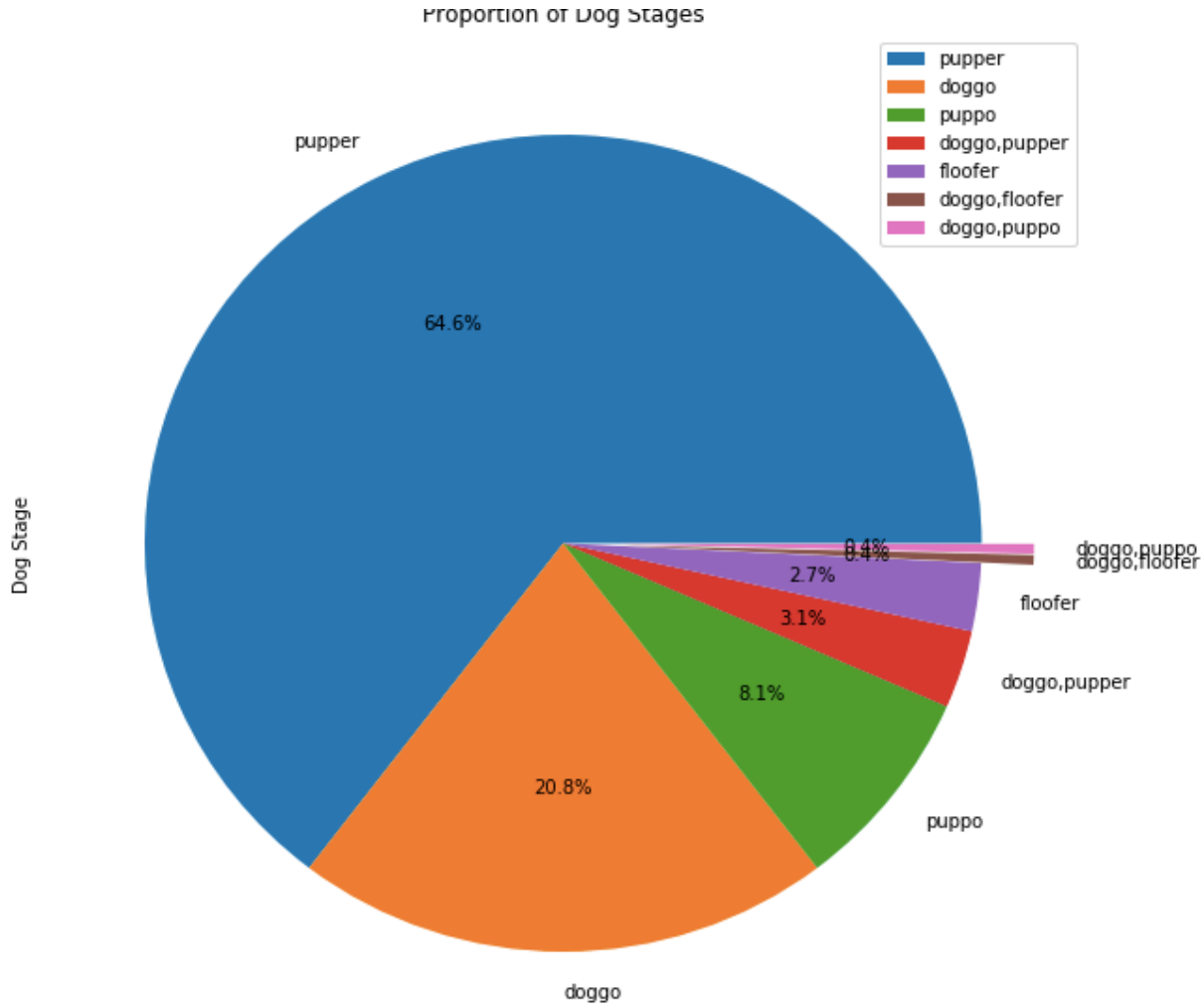
Our Second part of analysis revolved around studying our dataset rather than focusing on solving some populations,
Before we started our Project, we were given a brief overview about the project in which we were informed that WeRateDogs follow a unique rating system as mentioned above.
But knowing how humans are, it's hard to believe that some part of the population wouldn't be giving ratings within the bounds.
And if there is such a part of the population, how big is it?
For this, we first divided our dataset into 4 parts by finding out the quartiles and the divided the dataset into 4 different labels



CONCLUSION

- We can see that approximately 15.2% of total ratings were above 12/10
- A total of 62% of ratings were equal or above 10/10 which shows that rating out of bounds is a common phenomenon in this place
- But still this doesn't stop people from being realistic a rating the dogs within the bounds of 0-10 which accounts for a total of 38.1% of the ratings

Also, we wanted to know more about the vaguely described dog stages which was special to WeRateDogs, and see how they were distributed



CONCLUSION

- More than half of it was covered by puppers: A doggo that is inexperienced , unfamiliar or in anyway unprepared for the responsibilities associated with being a doggo
- 1/5th of the population was covered by the more mature Doggo population
- Even though the terms are pretty much vague, there is only 0.8% of the population where there is more than one dog stage entered

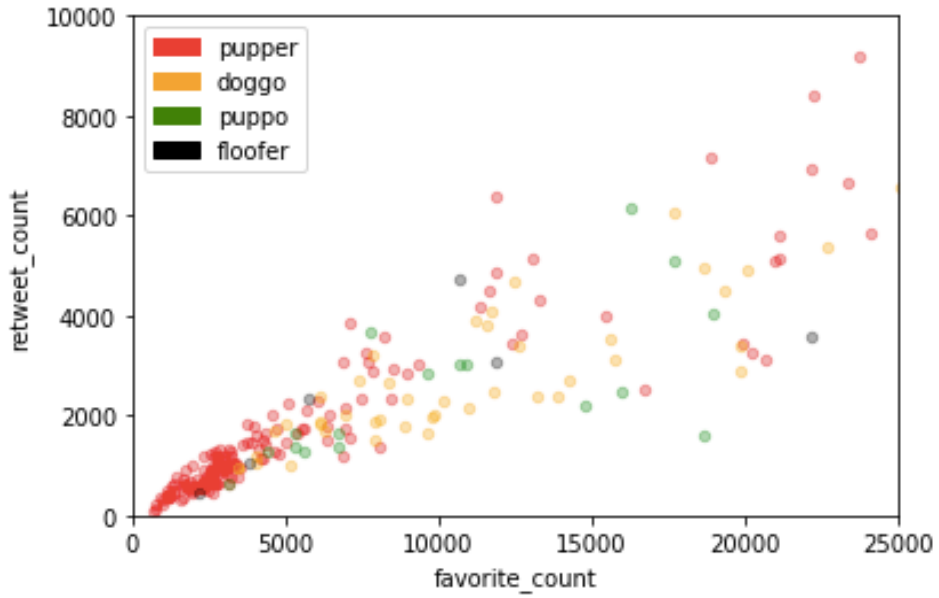
Analysis Part III

We next moved on to try our hand in creating a line of regression which could help us in estimating the retweet counts.

Favourite counts seems like a good factor to take under consideration as a factor. But since we have over 20 columns, let's try using more factors to help use build a better fitting line.

Lets see if there is a difference in retweet_counts and favourite_counts when we makes consider different dog stages

```
<matplotlib.legend.Legend at 0x11be63ef0>
```



In order to see if we should use dog_stages as a factor to determine the retweet counts in our multiple regression, we need to visualise the relation between different dog_stages and the retweet_counts

- We can see that there is difference in ratings w.r.t dog_stages, such as doggos have much greater rating in general than the puppers who are mostly concentrated towards the origins.

- puppos seem to be very spread out and have to be reconsidered when making the model

After Creating dummy variables of dog_stages and building an intercept column, we tried different combinations of variables to see which line fitted the best

Dep. Variable:	retweet_count	**R-squared:**	0.914
Model:	OLS	**Adj. R-squared:**	0.912
Method:	Least Squares	**F-statistic:**	675.4
Date:	Tue, 12 Mar 2019	**Prob (F-statistic):**	2.38e-134
Time:	22:33:00	**Log-Likelihood:**	-2391.5
No. Observations:	260	**AIC:**	4793.
Df Residuals:	255	**BIC:**	4811.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
intercept	-2232.8120	463.096	-4.821	0.000	-3144.792	-1320.832
doggo	1025.0052	543.923	1.884	0.061	-46.149	2096.159
floofer	1773.7999	1011.944	1.753	0.081	-219.033	3766.633
pupper	1562.6756	483.146	3.234	0.001	611.211	2514.140
favorite_count	0.4055	0.008	50.149	0.000	0.390	0.421

CONCLUSION

- When we tried to fit a linear, we obtained a regression line which could explain around 91.4% of variance but the condition number is so large, which indicates there might be a strong collinearity which is probably between favorite counts and retweets
- This is the best fit model which covers the most variance, have a low p-value and decreased the condition number significantly