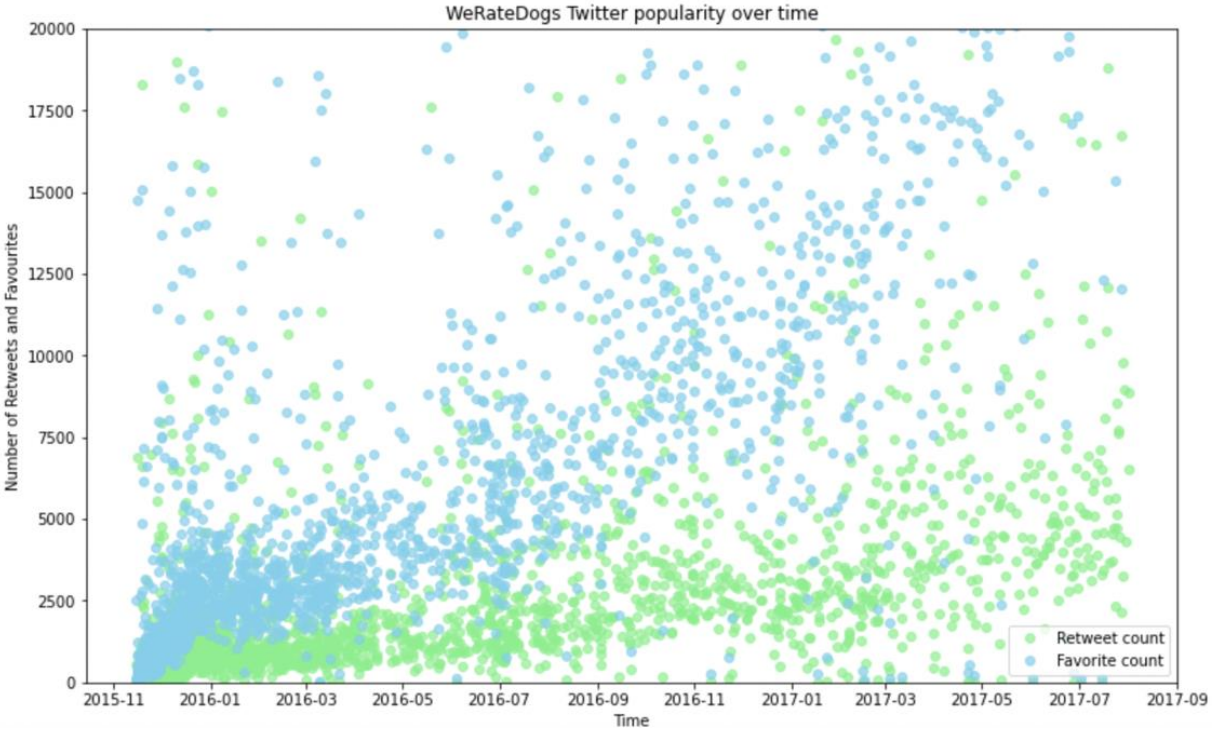
Contents

1. **Section 1** investigates WeRateDogs' Twitter popularity over time through data visualisations of retweets and favourites over time.
2. **Section 2** investigates the potential effect of the 'dog stage' assigned by WeRateDogs on the rating given to the dog.
3. **Section 3** investigates the potential effect of dog breed (predicted from twitter images by neural network) on the number of retweets and favourites.

1. Investigating WeRateDogs Twitter popularity over time

It is clear that WeRateDogs is a popular account as it has over 4 million followers and has received international media coverage. They even published a Dogtionary to explains the various stages of dog: doggo, pupper, puppo, and floof(er) while their book could be purchased on Amazon.

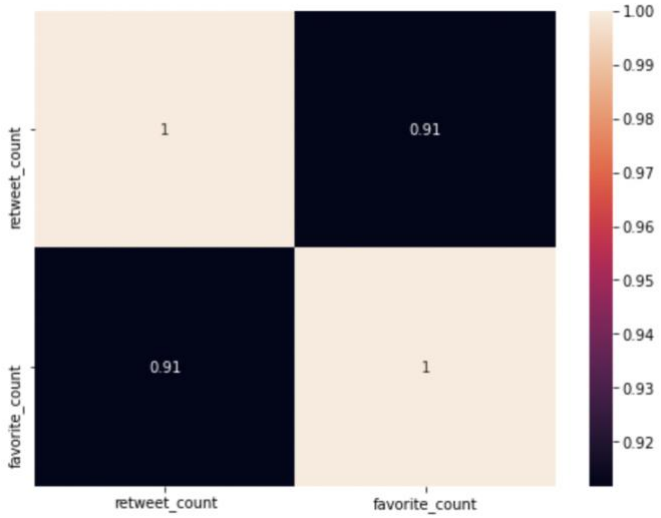
With the data regarding their tweets gathered and cleaned earlier in this project, it will be possible to visualise how the number of retweets favourites grew over time. The figures below show these visualisations:



As can be seen from the two graphs above, WeRateDogs Twitter account has significantly grown over time. My main insights from this visualisations are as follows:

- The number of favourites has almost always remained higher than the number of favourites over the late 2015 to late 2017 period.
- Between 2015-11 and 2016-05, the number of favourites was rarely above 5,000. However, after that, the number of favourites exploded, with tweets often getting more than 10,000 favourites
- A similar pattern could be observed for the retweet count. In the left half of the graph, the number of retweets rarely passes the 2,500 mark. Meanwhile, in the year 2017, it significantly increases, often reaching 5,000 or more retweets.

It could be said that these results were expected due to the fact that as the Twitter account grows in popularity, the number of favourites and retweets will increase. Moreover, it is not surprising that the amount of retweets is not as big as the number of favourites, as this is often the case. Users are more likely to favourite a tweet than to retweet it, and therefore these variations were expected as well.

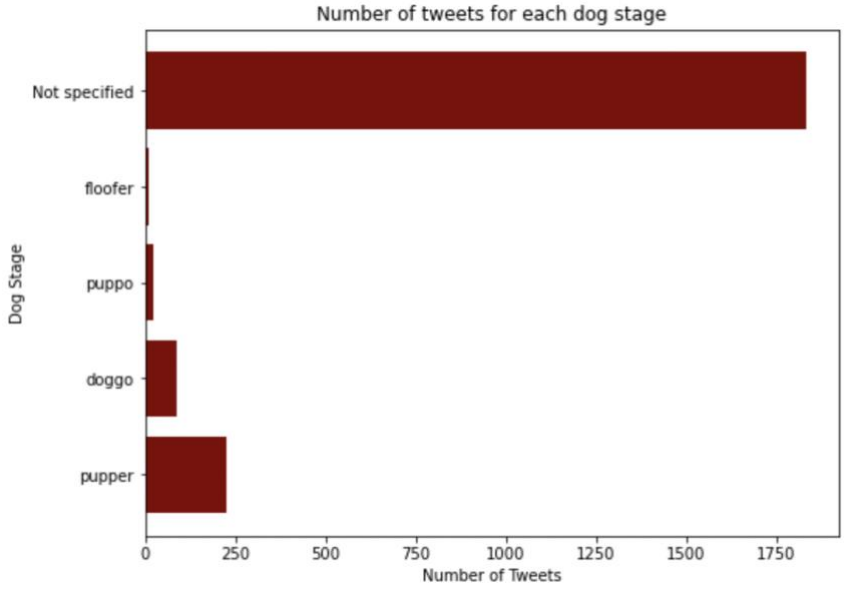


Furthermore, the above matrix shows that the number of retweets is indeed related to the number of favourites. With a correlation of 0.91, it is safe to say that the two metrics are strongly correlated with each other. Along with the previous graph, this further suggests that as one metric grew, so did the other.

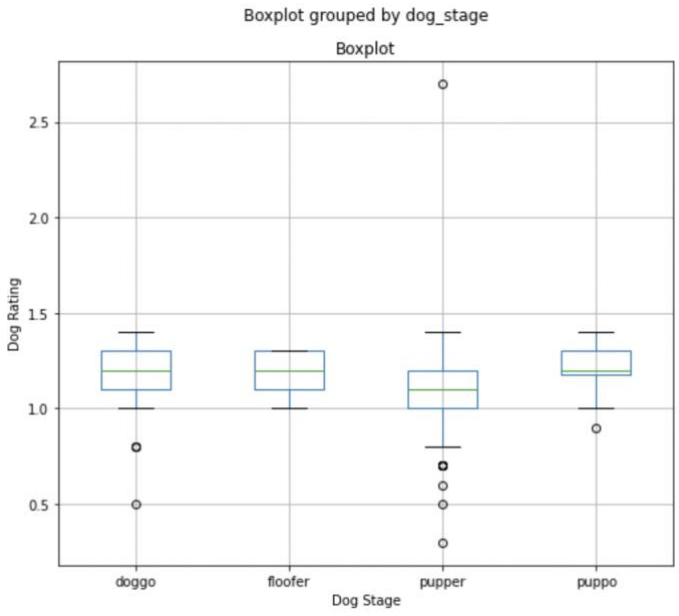
2. Investigating the relationship between dog stage and dog rating

It is clear from the previous section that WeRateDogs' Twitter account is popular. And despite WeRateDogs' rating system being inconsistent and sometimes confusing, it is possible that despite all dog being 'good dogs', some dogs could be more good than others. If this is the case, it could be hypothesised that dog characteristics could influence the rating given to the dog.

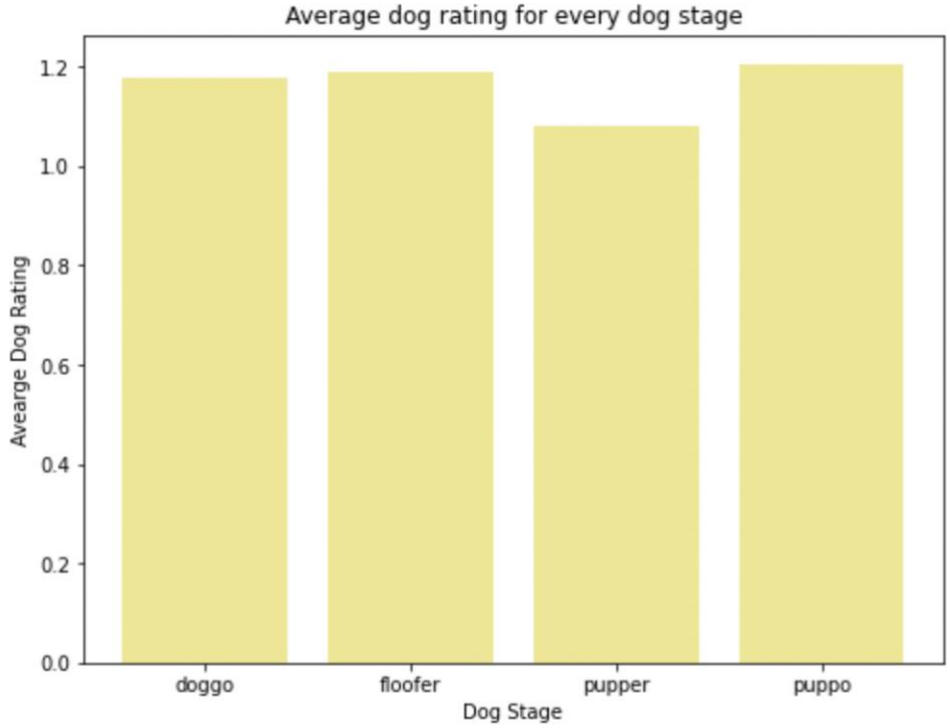
A suitable variable to test this is the 'dog stage'. While this variable is hardly scientific, WeRateDogs often assing dog stages to the dogs they rate and even have a Dogtionary. Since WeRateDogs both rates the dogs and assigns dog stages, it could be hypothesised that a dog's stage could affect its rating.



The graph above shows that most tweets did not assign a dog stage. However, when the dog stage was assigned, the most popular dog stage was 'pupper'. Unfortunately, this also shows that the majority of tweets present in the dataset cannot be used in the analysis and therefore, there is always a possibility that the sample will not be representative of the population.



Visualising the average ratings will help to understand if there are significant variations between every rating for different dog stages.The box plot above shows that there are some outliers in the data. However, taking into account the y-axis scale, these outliers are not very extreme and should be kept. Furthermore, in the case of outliers, due to the skewness of the data, the median is usually used. However, taking into account descriptive statistics, the mean and median for the data are were very similar. Therefore, the traditional approach of using mean will be used.



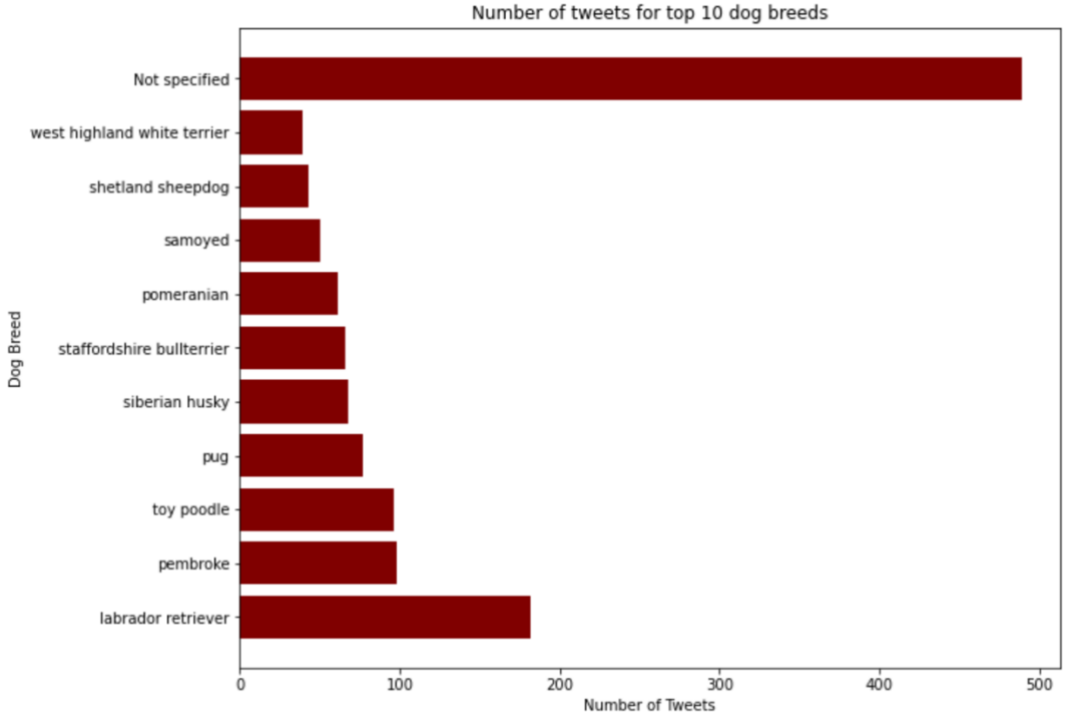
As can be seen from the graph above, the average rating for all 4 dog stages is relatively the same, it is around 1.2 (12/10). Hence, from simply looking at the graph, one could say that different dog stages will not result in different ratings, and thus, dog stages do not affect the dog ratings. However, regression will also be run to further explore this.

	coef	std err	t	P>|t|	[0.025	0.975]
intercept	1.2453	0.099	12.620	0.000	1.052	1.439
doggo	-0.0683	0.466	-0.147	0.883	-0.983	0.846
floofer	-0.0565	1.420	-0.040	0.968	-2.841	2.729
pupper	-0.1645	0.301	-0.547	0.584	-0.754	0.425

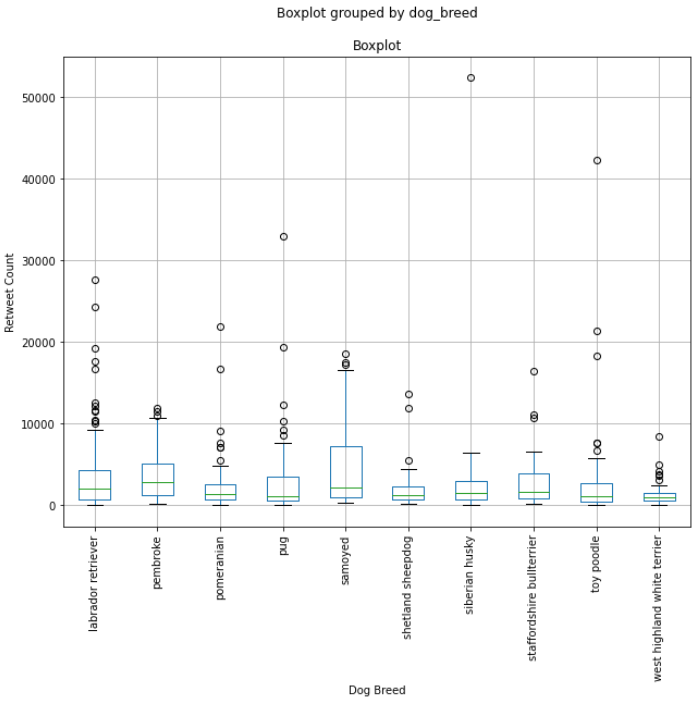
The regression model output shows that the predicted difference in the ratings. For instance, for a doggo stage dog and puppo stage dog, holding other variables constant, the predicted difference is expected to be -0.0683. This means that the doggo stage dog's rating is predicted to be 0.0683 less than the puppo stage dog. However, as can be seen from P>|t| column, none of these findings are statistically significant. All of those p-values are higher than the accepted alpha value under 95% confidence. Therefore it could be stated that no relationship was found between dog stage and dog rating.

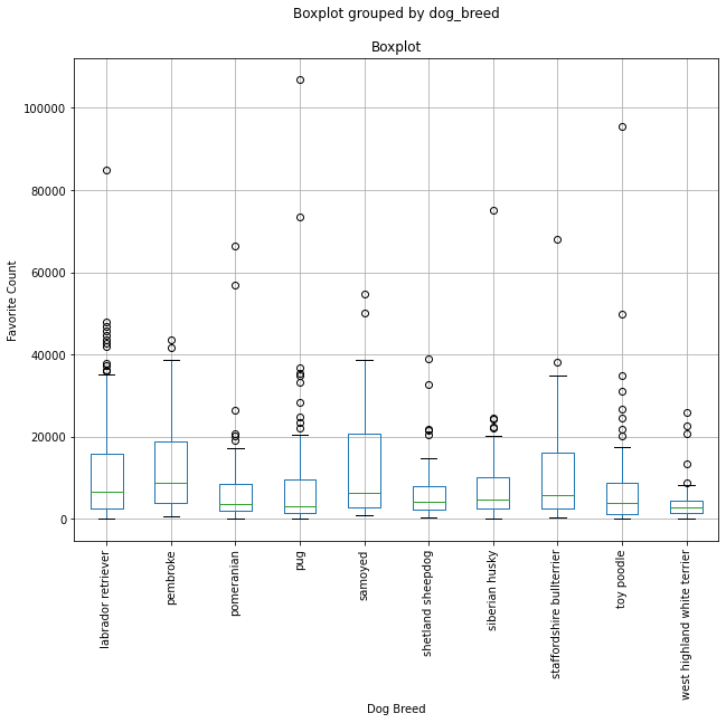
3. Investigating the relationship between dog breed (predicted from twitter images by neural network) and the number of retweets and favourites.

Following the logic from the previous section, it is possible that certain dog breeds will be more popular than others. However, instead of taking rating into account, the view of users will be considered. Twitter users are likely to be less subjective than the team behind WeRateDogs and could only favourite and retweet the breeds they like. Therefore it could be hypothesised that a dog's breed in the tweet could determine the number of retweets and favourites the tweet will get. To ease the visualisation process, only the top 10 most popular dog breeds were used in this analysis.

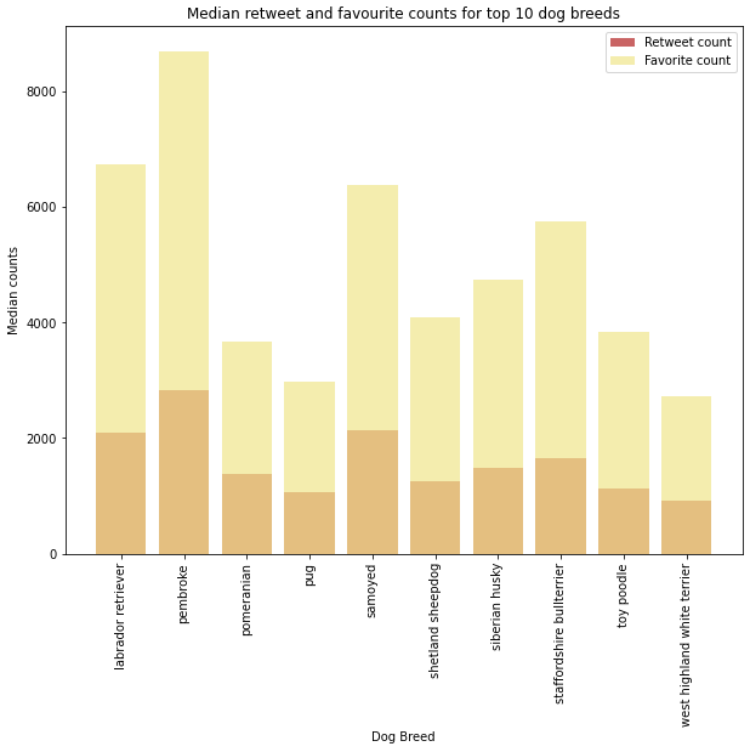


The graph above shows that most tweets did not have a predicted dog breed. However, when the dog breed was assigned, the most popular dog breed, by far, was 'labrador retriever'. Unfortunately, just like with the previous section 6.2, this also shows that the majority of tweets present in the dataset cannot be used in the analysis.





As can be seen from both box plots above, for both the retweet and favourite counts, there are significant outliers in the dataset. Moreover, taking into account the descriptive statistics, there are significant differences between the mean and the median. This suggests the mean could be seen as unreliable in this case, and therefore the median statistic will be used instead.



The bar graph above shows that there are indeed observable differences between dog breeds and the median retweet and favourite counts. However, this is not true for all breeds. For instance, samoyed and Labrador retriever have almost identical counts. Pembroke is the dog breed with the highest median favourite count (8696) and median

retweet count (2828.5). Similarly, west highland white terrier dog breed had the lowest median favourite count (2730) and median retweet count (917). Therefore a regression will be used to further investigate this.

	coef	std err	t	P>|t|	[0.025	0.975]
intercept	1440.6667	678.534	2.123	0.034	108.670	2772.664
labrador retriever	1925.0256	747.709	2.575	0.010	457.237	3392.815
pembroke	2097.6190	802.267	2.615	0.009	522.728	3672.510
pomeranian	1071.9235	868.774	1.234	0.218	-633.524	2777.371
pug	1392.4372	832.828	1.672	0.095	-242.446	3027.321
samoyed	3142.6133	905.278	3.471	0.001	1365.508	4919.719
shetland sheepdog	569.9147	937.011	0.608	0.543	-1269.484	2409.314
siberian husky	1226.0833	851.156	1.440	0.150	-444.779	2896.946
staffordshire bullterrier	1245.4545	855.844	1.455	0.146	-434.610	2925.519
toy poodle	1061.4375	804.643	1.319	0.188	-518.116	2640.991

The above table shows the results from a regression model constructed for retweet count variable is structured in the same way as the previous section. It also illustrates interesting results. Firstly, under the 95% confidence level, not all of the findings are statistically significant. However, the p-values for labrador retriever, pembroke and samoyed breeds have a lower p-value than the accepted alpha value of 0.025 and therefore could be considered statistically significant. Taking this into consideration, it could be said that these particular breeds, could be used to predict the number of retweets. However, only 10 breeds were considered within the regression model, and hence it is still hard to say if the dog breed indeed affects the retweet count. A larger dataset with more breeds should be used in future work.

	coef	std err	t	P>|t|	[0.025	0.975]
intercept	4768.9487	1934.513	2.465	0.014	971.404	8566.493
labrador retriever	6543.6117	2131.729	3.070	0.002	2358.922	1.07e+04
pembroke	7554.4186	2287.277	3.303	0.001	3064.380	1.2e+04
pomeranian	2864.4775	2476.890	1.156	0.248	-1997.780	7726.735
pug	4588.4279	2374.407	1.932	0.054	-72.652	9249.507
samoyed	8133.3913	2580.962	3.151	0.002	3066.835	1.32e+04
shetland sheepdog	2215.2373	2671.433	0.829	0.407	-3028.918	7459.393
siberian husky	3531.6101	2426.661	1.455	0.146	-1232.045	8295.265
staffordshire bullterrier	5935.8695	2440.025	2.433	0.015	1145.979	1.07e+04
toy poodle	2734.9159	2294.050	1.192	0.234	-1768.418	7238.250

The above table shows the results from a regression model constructed for the favourite count variable. Interestingly, just like with the previous regression model, most breeds have a p-value larger than the accepted alpha value with labrador retriever, pembroke and samoyed breeds have a lower p-value than the accepted alpha value of 0.025 and therefore could be considered statistically significant. Interestingly, staffordshire bullterrier, contrary to the previous model, could also be used to predict the number of favourites. However, once again, only 10 breeds were considered within the regression model, and hence it is still hard to say if the dog breed indeed affects the retweet count. A larger dataset with more breeds should be used in future work.