Analyzing the Data

Overview:

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. As it might be obvious, because "they're good dogs." WeRateDogs has over 4 million followers and has received international media coverage.

WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively to be used in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.

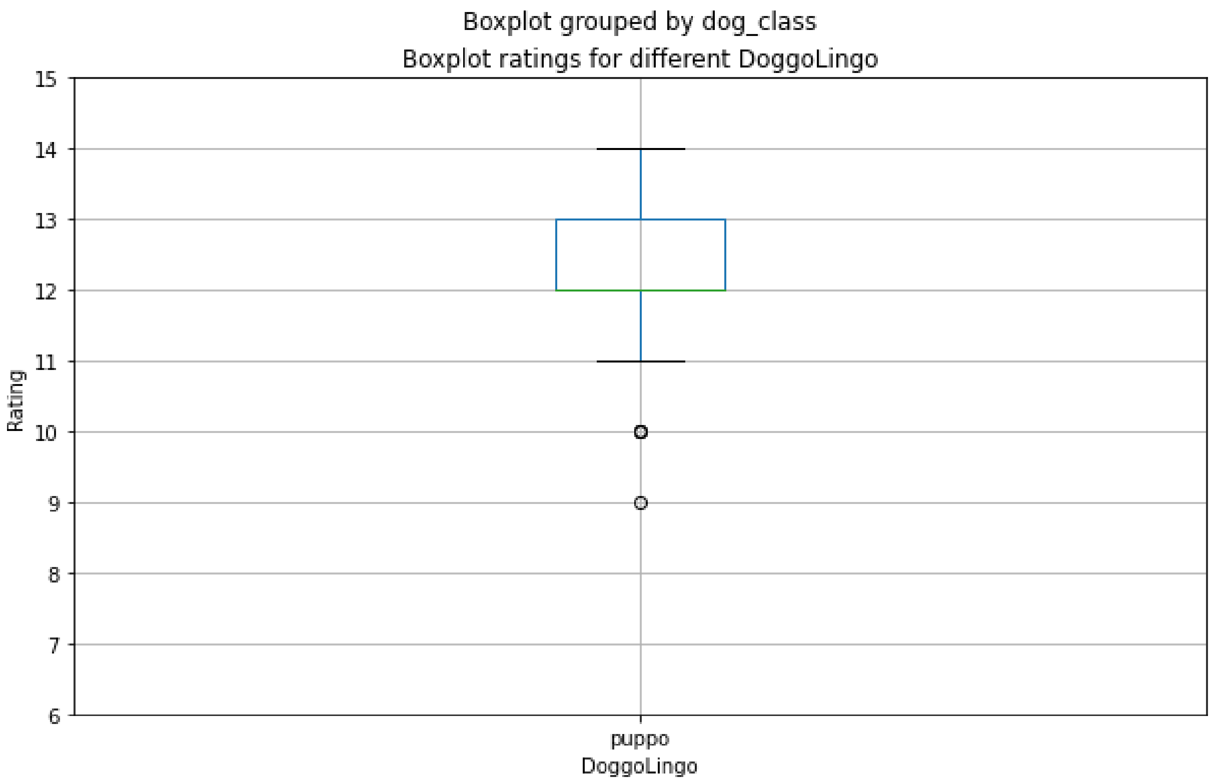
Tweepy package and tweeter API was used to collect the current status (retweet and favorite status) for the provided tweet IDs. At this stage, it is important to acknowledge that tweeter enabled me a developer account and provided me with needed keys to use tweeter APIs. A software was used to predict the dog breed in the tweets. The predictions were supplied in a separate data file.

In the previous data wrangling efforts, data were gathered, assessed, cleaned and finally combined in one master document to be used in data analysis and visualization. This project will try to answer these questions:

- Does one class of DoggoLingo receive more ratings than the others?
- Is dog class (DoggoLingo) „doggo, puppo... etc“ is associated with higher retweets and favourites?
- Is there a specific breed that has more retweets and favourite records than the others?
- Is there a relationship between ratings, favourite counts and retweets?

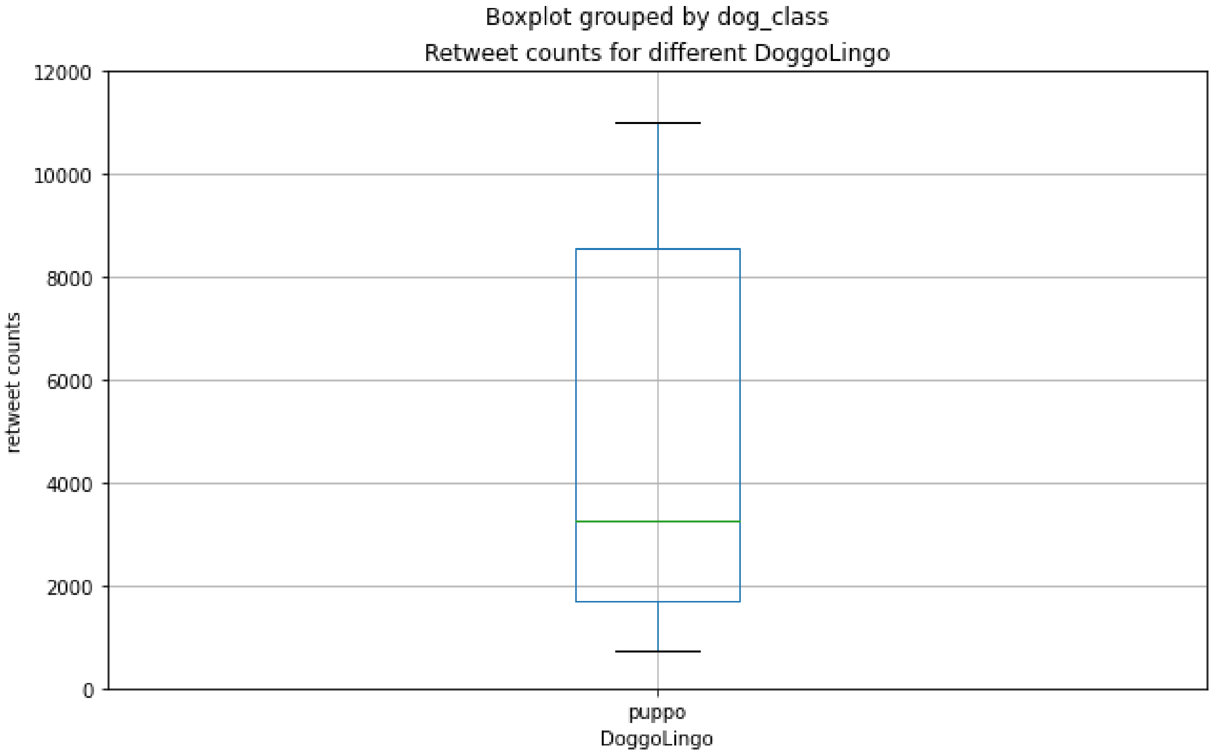
Plot the DoggoLingo with ratings

```
Out[3]: Text(0, 0.5, 'Rating')
```

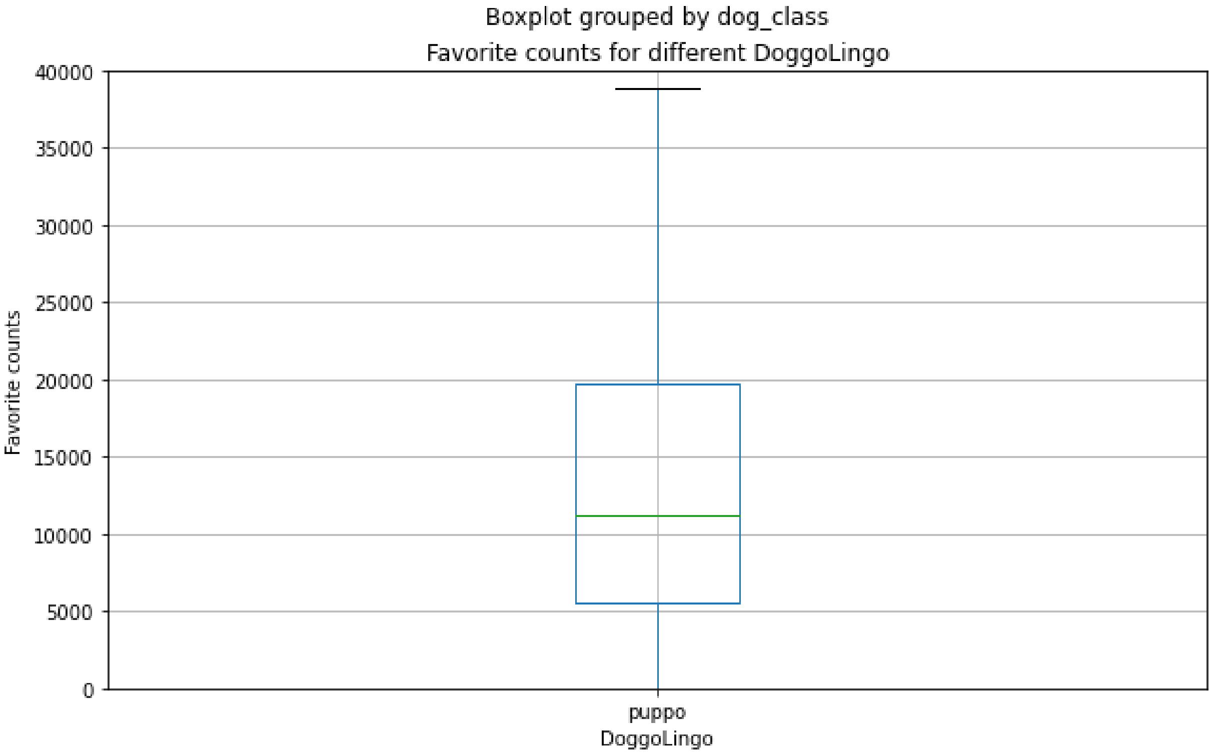


The box plot indicates that there is more consistency in rating puppo dogs high. There is also more consistency in rating the pupper dogs less than the others.

Plot the DoggoLingo with retweets

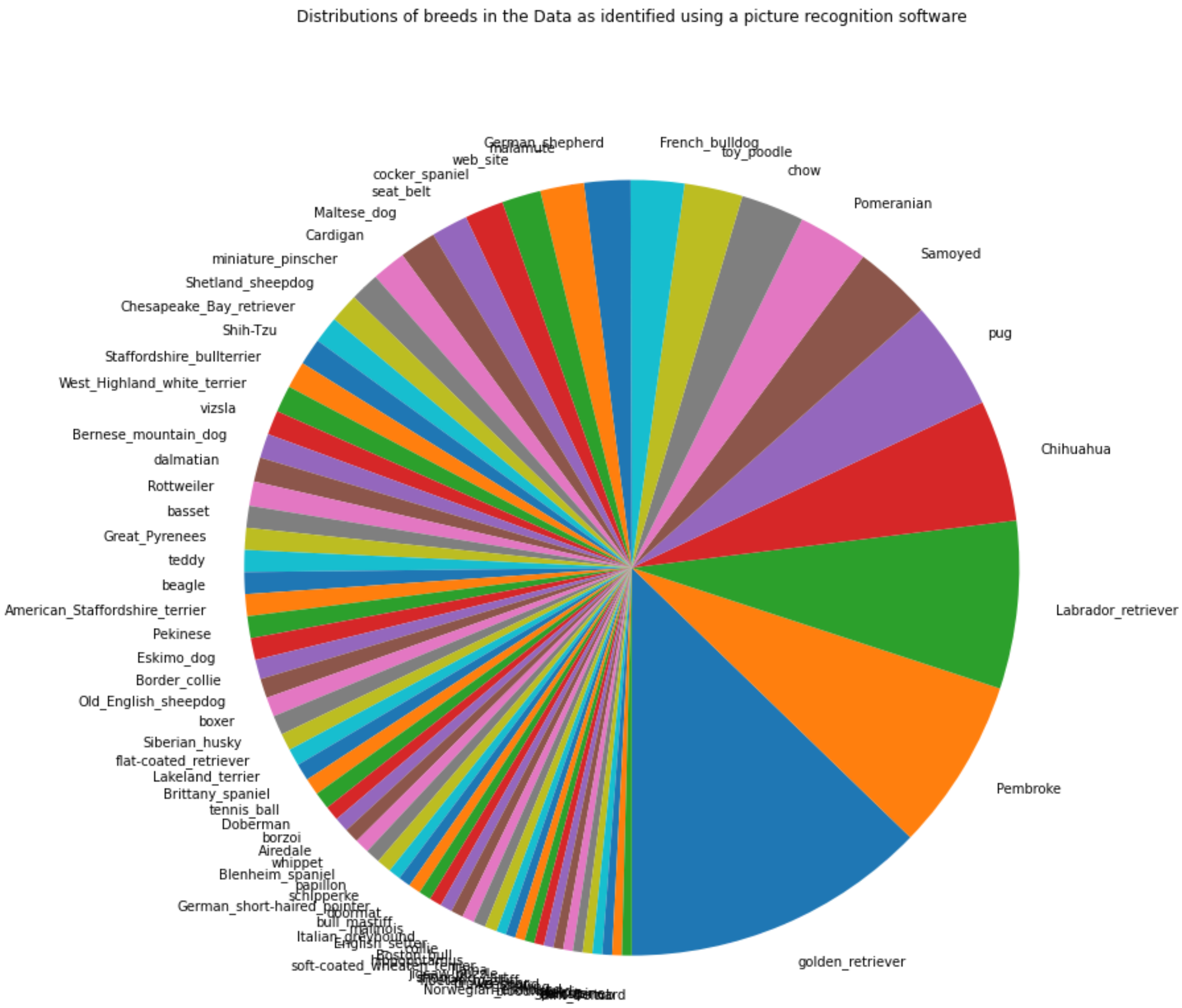


Plot the DoggoLingo with favourite counts



Puppos received consistently more favorite than the others.

Is there a specific breed that has more retweets and favourite record than the others?¶



Golden retriever, pembroke, Labrador retriever and Chihuahua are the most common dog breeds. Some strange values like Seat-Belt still exist in the data which need to be cleaned.

There are 73 different breeds in this data (Note: that appeared more than 3 times). Some strange values like "Seat Belt" still exist in the data and need to be cleaned.

Is there a specific breed that received more favorite than the others?¶

Out[8]:

	favorite_count							
	count	mean	std	min	25%	50%	75%	max
breed_probability1								
Labrador_retriever	69.0	13143.072464	19556.336568	0.0	2169.00	6569.0	17063.00	131075.0
Chihuahua	50.0	8544.140000	16490.536087	0.0	1226.25	3077.0	10564.00	107015.0
French_bulldog	22.0	18099.363636	23878.067337	341.0	3315.75	9335.5	25477.00	106827.0
golden_retriever	126.0	11490.714286	12975.141273	0.0	3246.00	7479.5	16029.75	85011.0
Eskimo_dog	9.0	17148.000000	23181.141026	550.0	3809.00	7908.0	17379.00	75163.0

Out[9]:

	favorite_count						
	count	mean	std	min	25%	50%	75%
breed_probability1							
swing	4.0	26663.000000	24442.574973	8157.0	11134.50	18350.5	33879.0
hippopotamus	5.0	20152.800000	15549.216112	0.0	11252.00	20275.0	28996.0
French_bulldog	22.0	18099.363636	23878.067337	341.0	3315.75	9335.5	25477.0
Eskimo_dog	9.0	17148.000000	23181.141026	550.0	3809.00	7908.0	17379.0
Chesapeake_Bay_retriever	11.0	15104.272727	20345.771836	227.0	4324.00	7335.0	17379.5

A dog from the Labrador_retriever breed received the maximum number of favourites (164825). On average, Swings received higher favourites (average = 31424) than the others. In the second place for favourites comes the Eskimo_dog (average = 21996).

Is there a specific breed that received more retweets than the others?

Out[10]:

	retweet_count							
	count	mean	std	min	25%	50%	75%	max
breed_probability1								
Labrador_retriever	69.0	5214.434783	10491.457715	96.0	907.00	2269.0	5134.00	79515.0
Chihuahua	50.0	4398.340000	11005.679309	52.0	552.50	1328.0	3569.25	56625.0
Eskimo_dog	9.0	7826.555556	16803.662511	163.0	1176.00	2243.0	3220.00	52360.0
French_bulldog	22.0	4933.954545	7415.783261	123.0	847.75	2705.5	4571.00	32883.0
swing	4.0	10243.750000	13672.163627	2873.0	3300.50	3680.0	10623.25	30742.0

Out[11]:

	favorite_count						
	count	mean	std	min	25%	50%	75%
breed_probability1							
swing	4.0	26663.000000	24442.574973	8157.0	11134.50	18350.5	33879.0
hippopotamus	5.0	20152.800000	15549.216112	0.0	11252.00	20275.0	28996.0
French_bulldog	22.0	18099.363636	23878.067337	341.0	3315.75	9335.5	25477.0
Eskimo_dog	9.0	17148.000000	23181.141026	550.0	3809.00	7908.0	17379.0
Chesapeake_Bay_retriever	11.0	15104.272727	20345.771836	227.0	4324.00	7335.0	17379.5

A dog from the Labrador_retriever breed received the maximum number of retweets (84021). On average, Swings received higher retweets (average = 10910) than the others. In the second place for retweets comes the hippopotamus (average = 10206).

Is there a specific breed that received more retweets than the others?¶

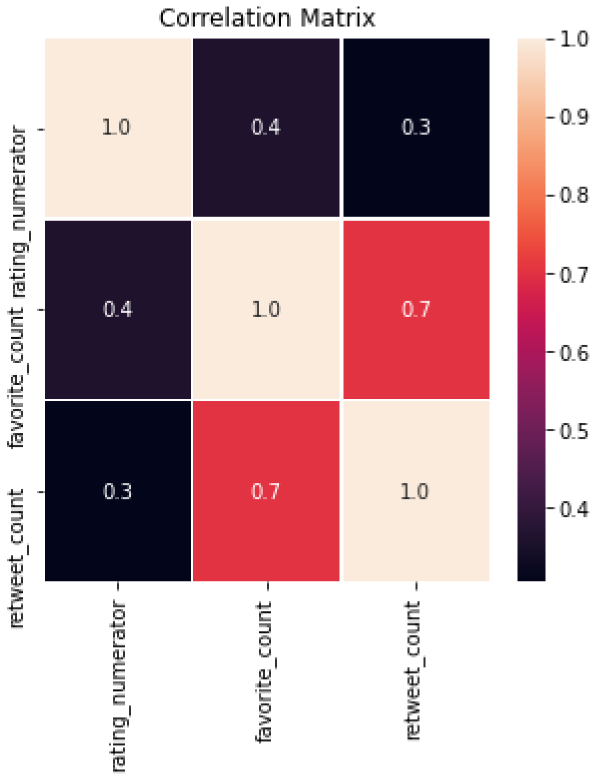
Out[12]:

	retweet_count							
	count	mean	std	min	25%	50%	75%	max
breed_probability1								
Labrador_retriever	69.0	5214.434783	10491.457715	96.0	907.00	2269.0	5134.00	79515.0
Chihuahua	50.0	4398.340000	11005.679309	52.0	552.50	1328.0	3569.25	56625.0
Eskimo_dog	9.0	7826.555556	16803.662511	163.0	1176.00	2243.0	3220.00	52360.0
French_bulldog	22.0	4933.954545	7415.783261	123.0	847.75	2705.5	4571.00	32883.0
swing	4.0	10243.750000	13672.163627	2873.0	3300.50	3680.0	10623.25	30742.0

Out[13]:

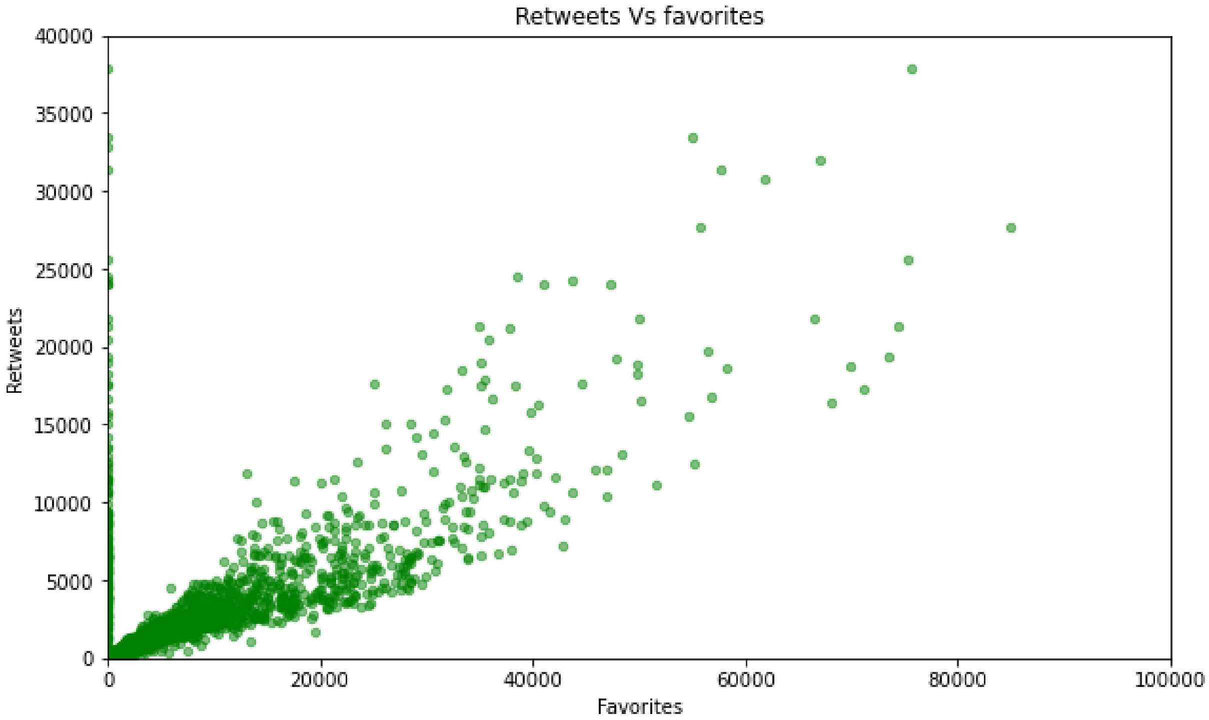
	retweet_count							
	count	mean	std	min	25%	50%	75%	max
breed_probability1								
hippopotamus	5.0	10779.400000	4193.247882	5174.0	7445.00	12882.0	14198.00	14198.0
brown_bear	4.0	10453.250000	11601.325883	298.0	460.75	10507.5	20500.00	20500.0
swing	4.0	10243.750000	13672.163627	2873.0	3300.50	3680.0	10623.25	30742.0
Eskimo_dog	9.0	7826.555556	16803.662511	163.0	1176.00	2243.0	3220.00	52360.0
Tibetan_mastiff	4.0	6332.250000	4360.785967	1035.0	3618.75	7193.5	9907.00	9907.0

Is there a relationship between ratings, favourite count and retweets?¶



Out[15]:

	rating_numerator	favorite_count	retweet_count
rating_numerator	1.000000	0.361302	0.305336
favorite_count	0.361302	1.000000	0.702881
retweet_count	0.305336	0.702881	1.000000



It is normal to expect that not all favorited tweets were retweeted and vice versa. This caused the line of dots at the favorites value of zero. There is a linear relationship between the retweet counts and the favourite counts. The regression coefficient for this relationship is strong (r= 0.797). This relationship might be already expected as people will favourite what they retweet or the other way around. Interesting, the rating have a weak significant correlation (r = 0.371) with favorits but not that significant correlation with retweets. This indicates that people might favour high rated dogs. This causality relationship could be justified because the dogs are been rated by WeRateDogs first and before people react (by favouring or retweeting) to them.

To sum up

There is more consistency in rating puppo dogs high. There is also more consistency in rating the pupper dogs less than the others. Puppos, on average, were retweeted more than the others and received more favorites as well.

Golden retriever, pembroke, Labrador retriever and Chihuahua are the most common dog breeds in WeRateDogs data. Some strange values like Seat-Belt still exist in the data which need to be re-cleaned. A dog from the Labrador_retriever breed received the maximum number of favourites and retweets. On average, Swings received higher retweets (average = 10910) than the others. In the second place for retweets comes the Brown Bears (average = 9947).

It is normal to expect that not all favorited tweets were retweeted and vice versa. However, there is a linear relationship between the retweet counts and the favourite counts. The regression coefficient for this relationship is strong (r= 0.797). This relationship might be already expected as people will favourite what they retweet or the other way around.

Interesting, the rating have a weak significant correlation (r = 0.371) with favorits but not that significant correlation with retweets. This indicates that people might favour high rated dogs. This causality relationship could be justified because the dogs are been rated by WeRateDogs first and before people react (by favouring or retweeting) to them.