WHICH DOG HAD THE TOP-MOST RATING?

Image from *Boston magazine*

From three (3) data sets; Twitter archive enhanced, tweet_json, and image predictions. All about a Twitter account by name – *WeRateDogs*. I carried out a data analysis to find out which dog had the highest rating, its breed, stage and if at all its rating influenced the number of counts it got for retweets and favorite.

WeRateDogs is a Twitter account that specialize in rating dog selfies uploaded by various users from different part of the world. Though their numerical rating terms seems weird, that has been one of their outstanding trademarks.

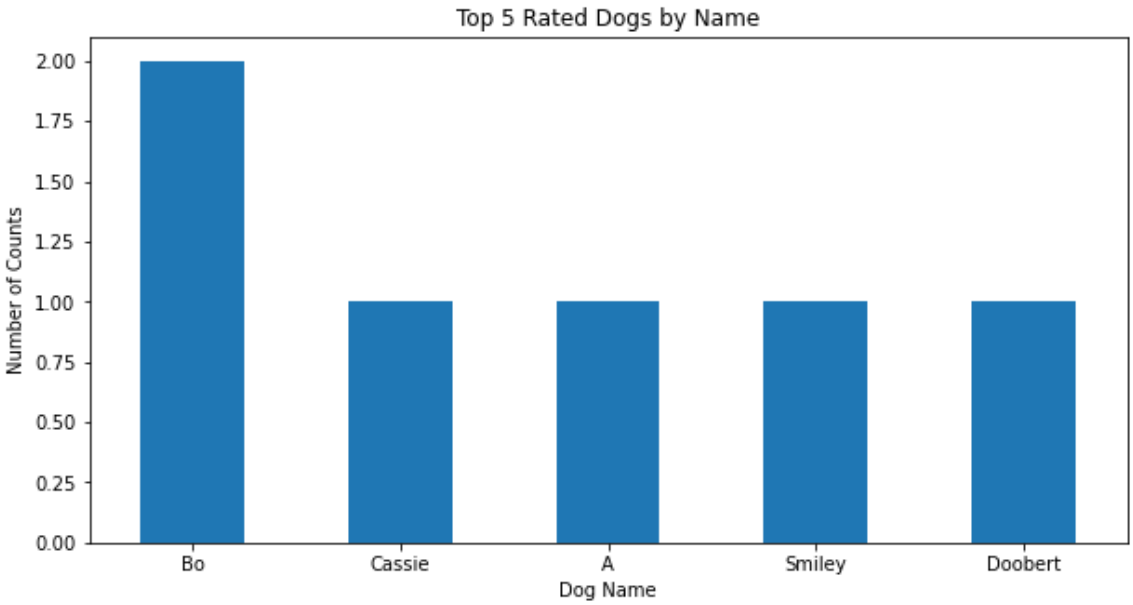
In other to set a clear path for my analysis, I looked through the available variables in each data set, and settled to gain insights from the gathered tweet data by seeking answers to the following questions:

i. Which dog by name had the top-most rating.
ii. Which dog by breed had the highest rating? Also comparing the confidence level of its image prediction.
iii. What stage was the dog with the highest rating in?
iv. Was there a correlation with the dog rating and number of retweeted tweet count and favorite count?

Upon completion of my exploratory data analysis on the dataframes, I came up with the following insights:

INSIGHTS:

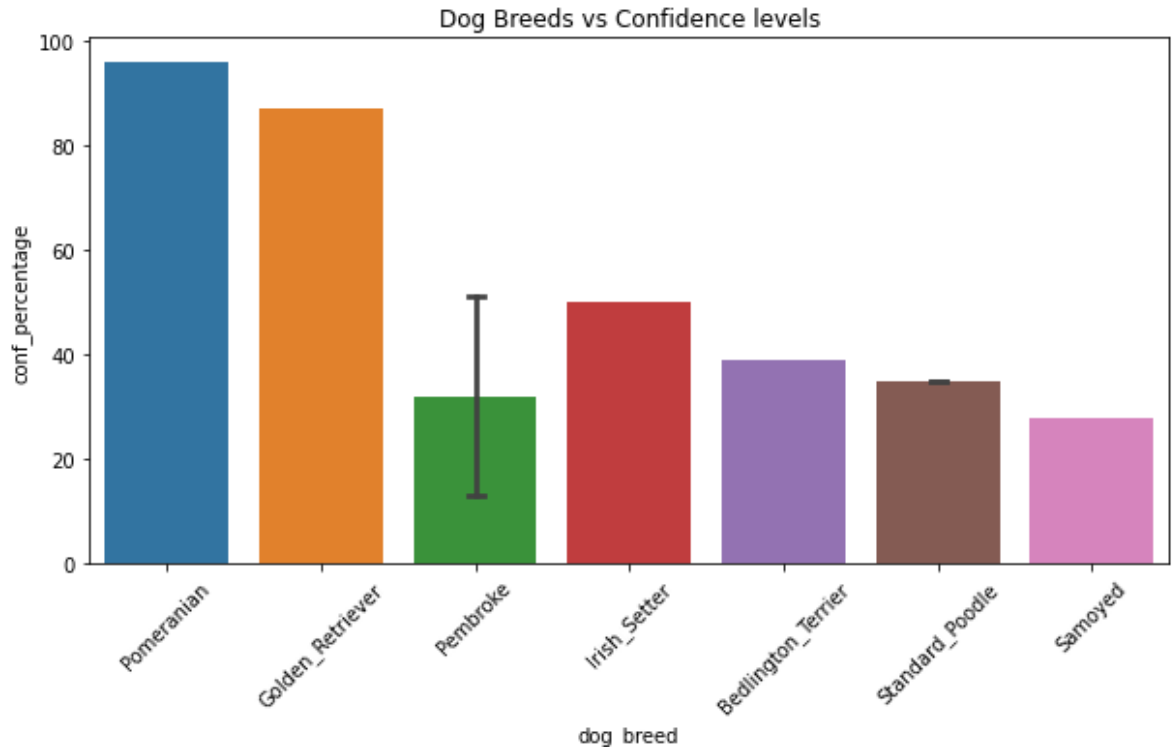
i. The name of the top-most-rated dog is Bo.



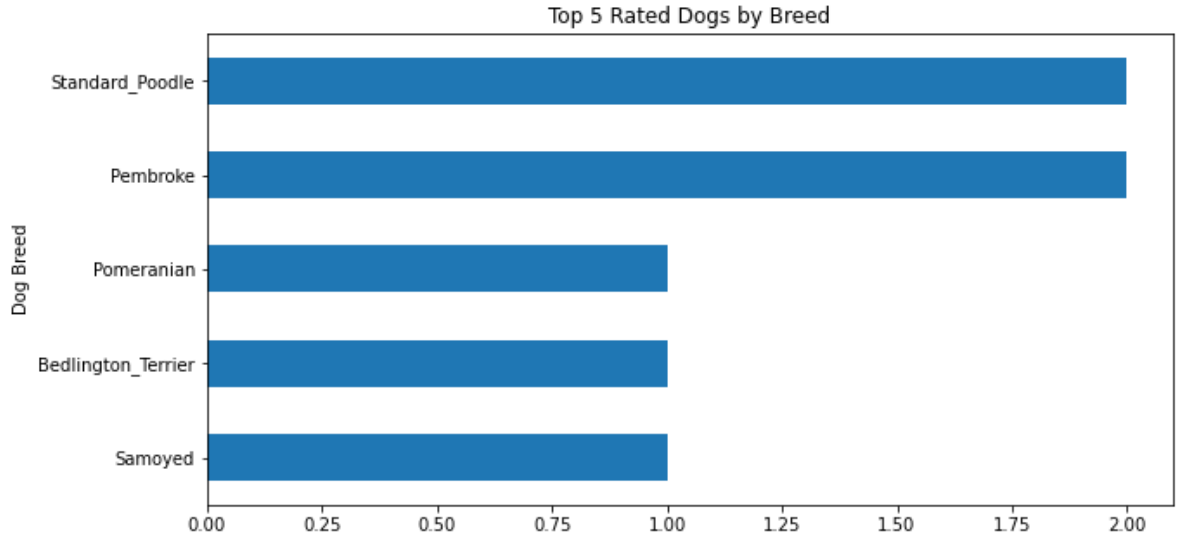
From my analysis, the top 3 highest rating values were – 1.5, 1.4, and 1.3. Only one dog had a rating of 1.5 with no identifiable breed name. This suggest that this could be an outlier value as both the information provided and count were not significant enough to draw any compelling conclusions.

Using the next value – 1.4, several dogs with recorded names, stage and breed were identified. However, to know which dog came top, I used the number of value count for each dog in this category. And the dog named 'Bo' came first with 2 counts.

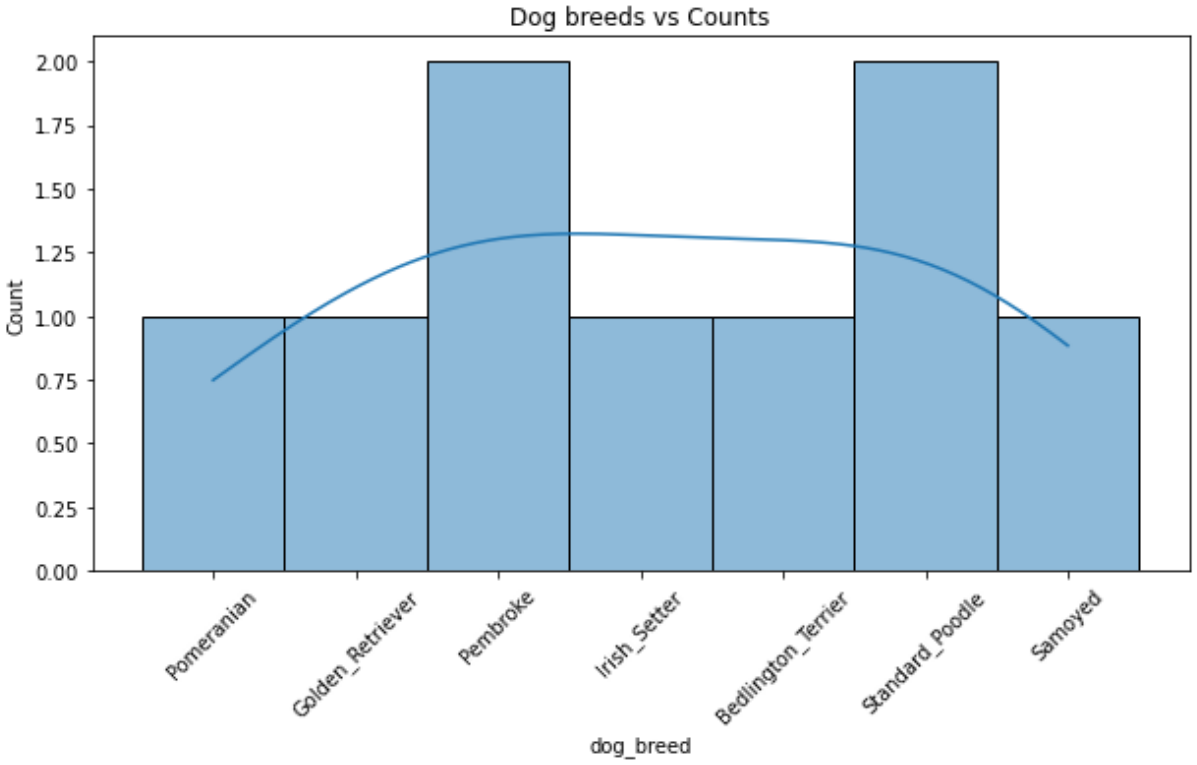
ii. The dog breed with the highest rating was the Standard_Poodle.



Although, the Pomeranian, Golden_Retriever, and Pembroke came 1st, 2nd, and 3rd from the visual analysis of dog breeds by the confidence level of picture prediction. They all had one missing value or other. And none had the highest rating count.



Plotting the dog breeds by the highest rating counts, two dog breeds emerged top – Standard_poodle and Pembroke.

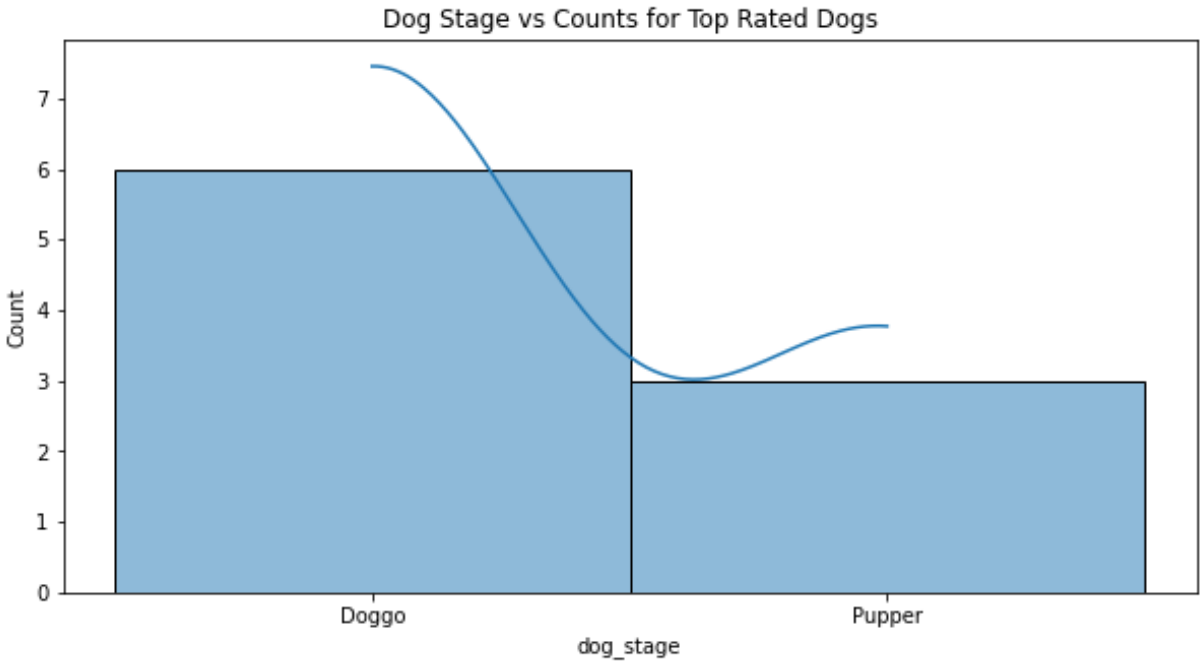


So, to find out which of the two breeds our dog named Bo, belonged to, I filter our dataframe with his credentials and got.

	tweet_ID	rating_num	rating_den	name	dog_stage	rating	img_num	dog_breed	conf_level	conf_percer
355	819015337530290176	14	10	Bo	Doggo	1.4	1	Standard_Poodle	0.351308	
358	819004803107983360	14	10	Bo	Doggo	1.4	1	Standard_Poodle	0.351308	

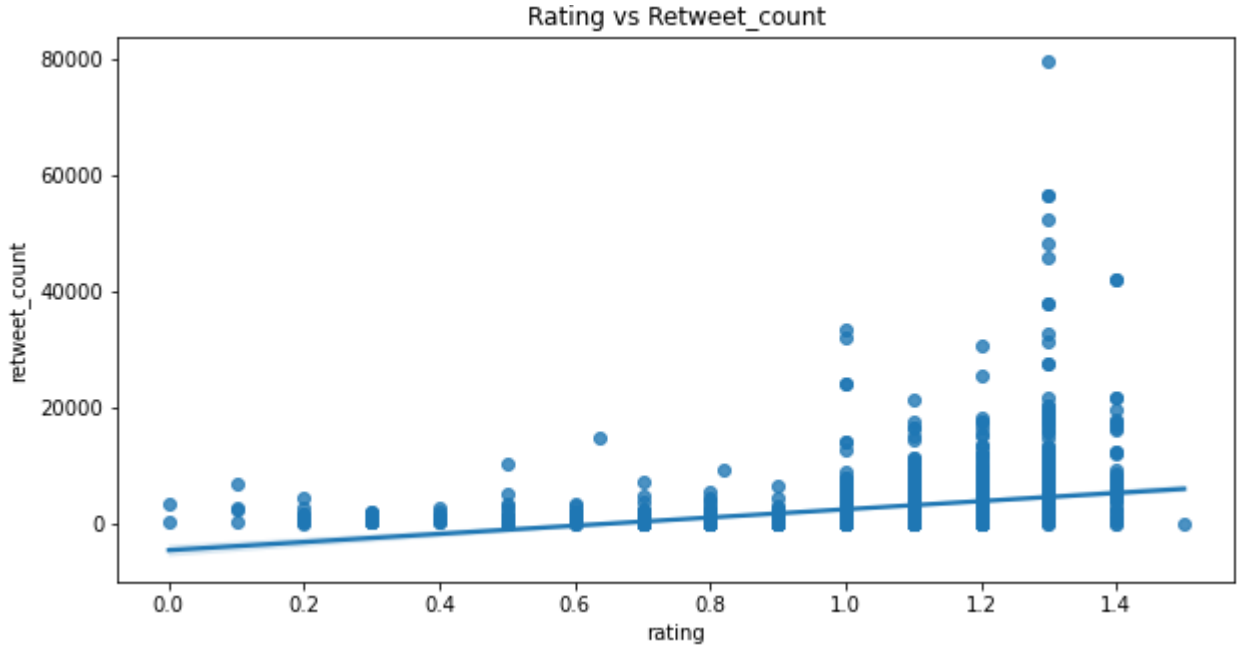
Although, the confidence level of the prediction is low, Standard_Poodle was the dog breed Bo, belonged to.

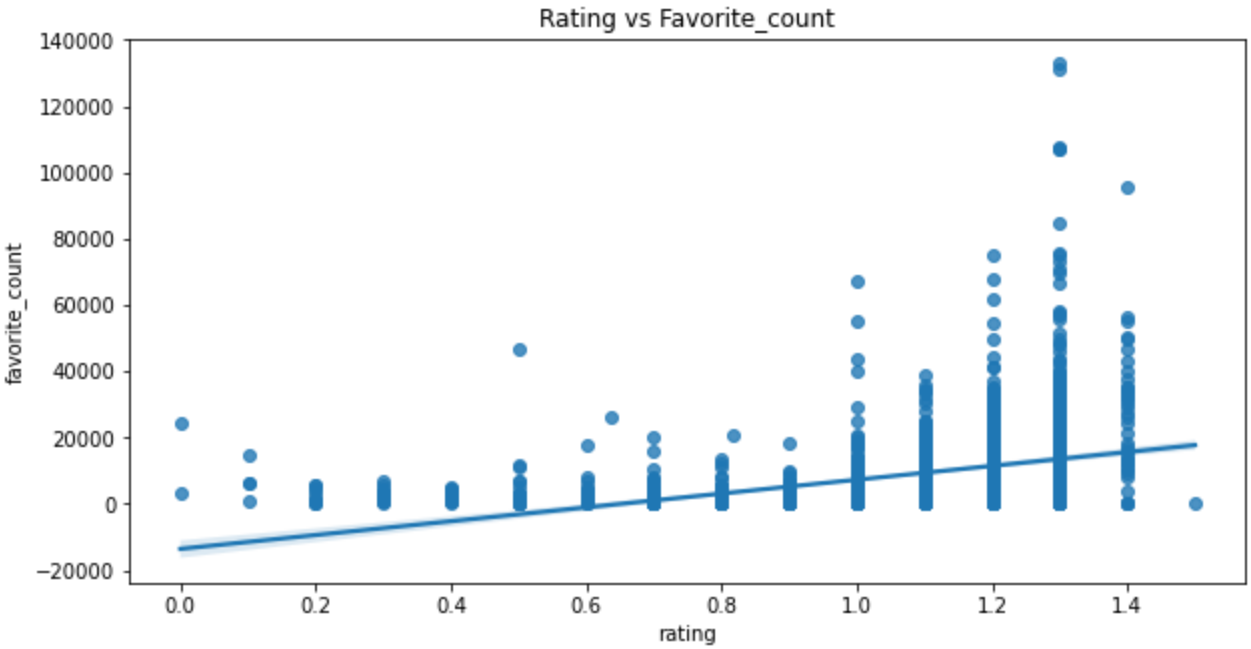
iii. Bo, was in the doggo stage.



With a gap of about 50%, most top-rated dogs were in the doggo stage. From the dogtionary provide in the Udacity class room, this refers to a big pupper, capable of taking/following instructions.

iv. There is a likelihood that dogs that get high rating have a greater chance of being retweeted or marked as favorites.





There is visibly a progressive positive correlation between favorite and retweet count with respect to rating. This implies that dogs were less likely to be marked as favorite or retweet if their ratings were low and vice versa. It's a no-brainer though.

Limitations:

i. The rating system is unconventional and leaves room for ambiguity.

ii. Some columns had large number of missing values (ex. Retweeted_timestamp – over 2000).

iii. Unrelatable names for dog stages.

A Cute White Fluffy Standard Poodle *[a-z-animals.com]*