

Data Exploration

Introduction

As stated earlier in the wrangling report, one question with the WeRateDogs data was the ratings. These ratings are carried out on a scale of 10 but ratings inputted by users are almost always greater than 10. 11/10, 12/10, 13/10, etc. Why?

Objective

The data exploration was aimed towards getting an insight with respect to the following;

- The dogs stage
- The dogs rating
- The dogs breed prediction result

Result Summary

Below is the summary of the result and observation made from the data exploration with respect to the different objective:

- The dogs stage investigation

Result

stage	count	avg_likes	avg_retweets	avg_rating	avg_followers
doggo	64	11024.218750	2884.046875	11.578125	8.824233e+06
floofer	6	8923.166667	2368.000000	11.833333	8.824230e+06
pupper	197	5905.055838	1836.406091	10.598985	8.824227e+06
puppo	20	13392.750000	3285.850000	11.850000	8.824233e+06

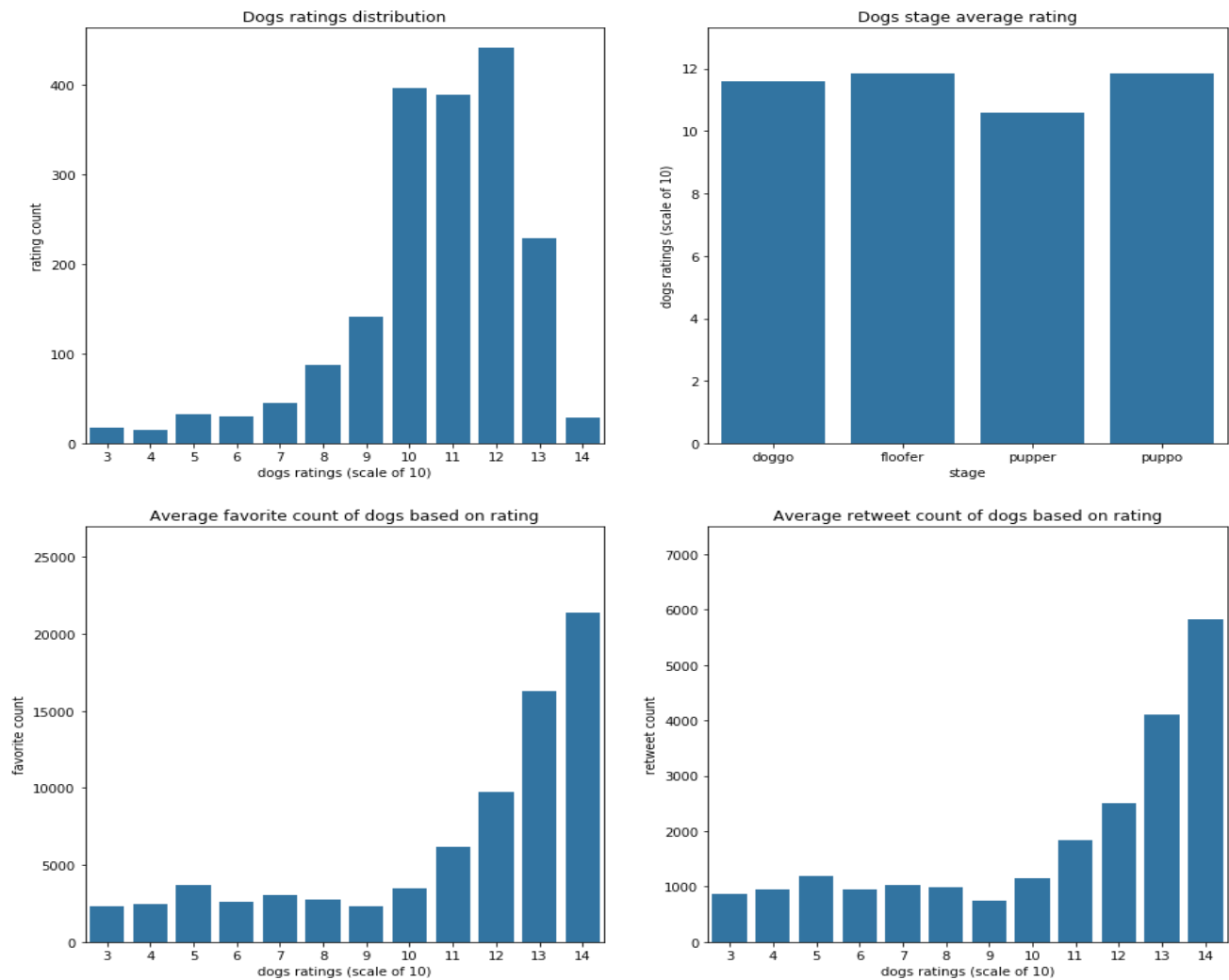
Dogs stage investigation table

Observations

1. The pupper dogs forms large proportion of the dogs
2. On average, the dogs in the puppo stage has more likes than others
3. Also, the puppo dogs has the highest average retweet in comparison to the other stage of dogs
4. Puppo better rated on an average than other stage of dogs
5. Finally, there is no significant difference between the average followers of the dogs

- Investigation on the dogs rating

Result



Observations

1. Most dogs were allocated a rating of 12 by the users
2. On average, dogs in the floofer stage has the better rating compared to others
3. The shape of the favorite and retweet count distribution based on rating is similar. This implies a direct relationship between the favorite and retweet count
4. Dogs with rating of 14 has the highest favorite and retweet count
5. In general, there is a trend relating the favorite and retweet count to the ratings but an increasing trend is observed from a rating of 9 to 14

- **Investigation of the dogs bred predictions result**

Result

Img_num	pred_count	correct_pred	pred_accuracy(%)	pred_avg_conf
1	1379	1172	84.989123	0.593710
2	1391	146	10.496046	0.135534
3	1346	50	3.714710	0.060285

Classifiers result summary tablez22

p1	pred_count	correct_pred	pred_accuracy(%)	pred_avg_conf
golden_retriever	130.0	98.0	75.384615	0.721899
Labrador_retriever	90.0	80.0	88.888889	0.644050
Pembroke	82.0	71.0	86.585366	0.721344
Chihuahua	73.0	61.0	83.561644	0.574581
pug	52.0	46.0	88.461538	0.768736
chow	39.0	32.0	82.051282	0.612089
Samoyed	38.0	34.0	89.473684	0.735467
Pomeranian	36.0	30.0	83.333333	0.756210
toy_poodle	34.0	28.0	82.352941	0.610915
malamute	27.0	24.0	88.888889	0.589129

Deep dive into classifier 1 table

Observations

1. Most prediction of the dog breed was image number 1
2. Classifier 1 has the best accuracy and also average prediction coefficient
3. With respect to classifier 1:
 - Majority of the predictions and correct predictions were golden retriever
 - Samoyed breed had the best accuracy and also average prediction confidence

Conclusion

With view to my result and observation, it could be concluded that:

- People are attracted to and rated dogs in puppo stage better than others
- Dogs with rating higher than 10 are most likely to be better and attractive than those with lesser rating
- Finally, the classifier 1 algorithm for the dogs image predictions is better than the other algorithm