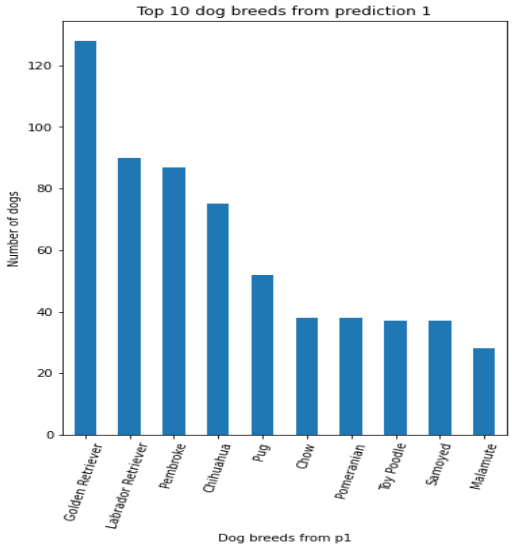
Insight and data visualization of WeRateDogs

Introduction

This report is made from cleaned and combined dataset of twitter-archive-enhanced.csv, image_prediction.tsv and tweeter_jason from WeRateDogs tweeter account.

1. Top 10 dog breeds from three predictions

The top 10 dog breeds are taken from predictions of prediction 1 where prediction confidence level is true. Through prediction 1, the count of top 10 dog breeds are widely spread that the top 10 breeds has standard deviation of 32. This means few breeds have higher prediction chance than most of the breeds.



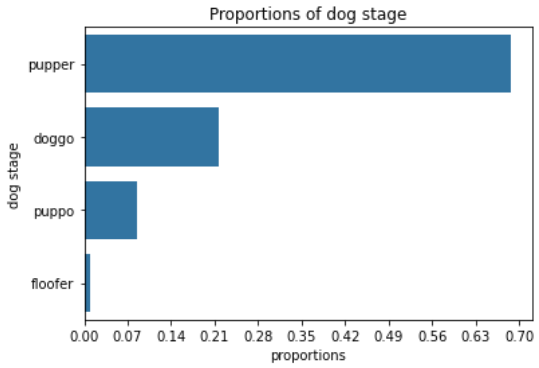
The average confidence level when prediction is true or false is given below.

```
True     61.480154
False    53.444334
```

This shows that prediction made by neutral network has higher confidence level when prediction is true than when prediction is false.

2. Proportions of dog stage

After removing missing dog stage in cleaned dataset, the proportion of dog stage as per the tweeter count is shown in the below figure. This shows that in internet most people like medium to large sized puppies where they name “pupper” and they least like floofer, fluffy dog.

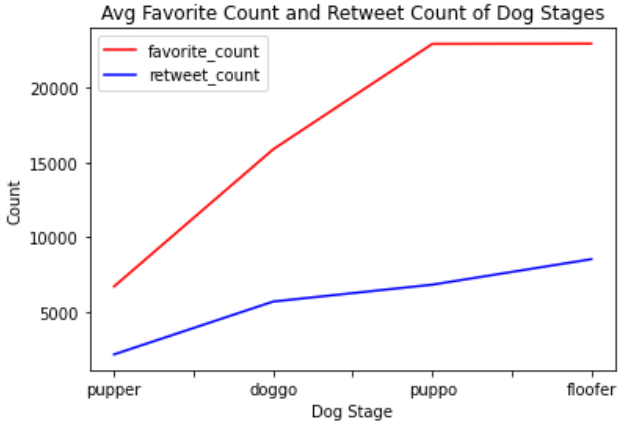


The proportion of dog stage is in order of pupper, doggo, puppo, and floofer. This order is reversed when we order dog stage as per their average rating ratio.

Counts of dog stage	
pupper	218
doggo	69
puppo	27
floofer	3

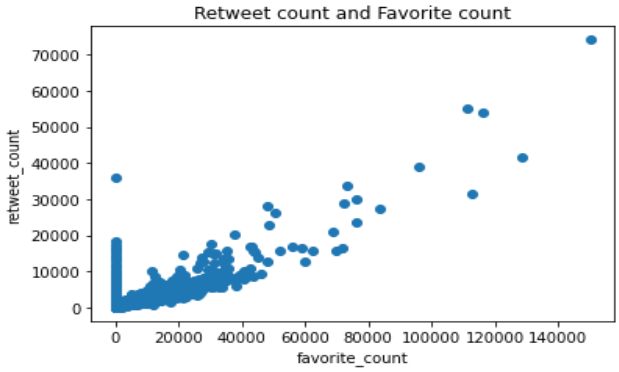
Average rating ratio	
floofer	1.266667
puppo	1.214815
doggo	1.160870
pupper	1.054248

When we counted average favorite counts and average retweet counts of four dog stage, it showed that these average two counts showed same order for all four dog stages. In counts, average favorite counts are higher than average retweet counts.



3. Counts of Retweet and Favorite tweets over time

It is obvious that favorite tweet tends to get more retweet. Even though very few retweet and fav orite tweet have outliers, these two counts are highly correlated that is Pearson correlation of 0. 91.


Retweet count and Favorite count
retweet_count
favorite_count
0
10000
20000
30000
40000
50000
60000
70000
0
20000
40000
60000
80000
100000
120000
140000