ANALYSIS REPORT

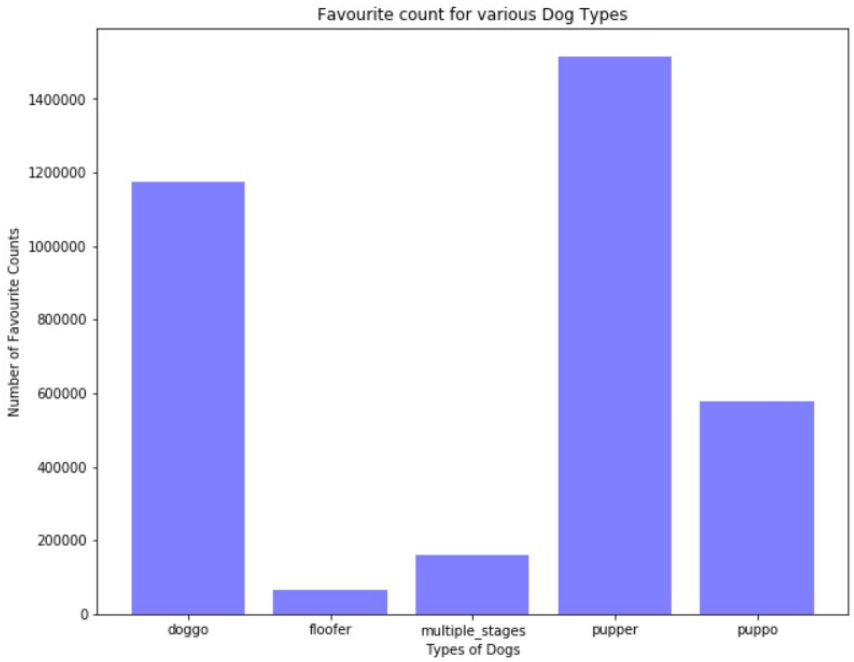
Project: We Rate Dogs - Data Wrangling and Visualization

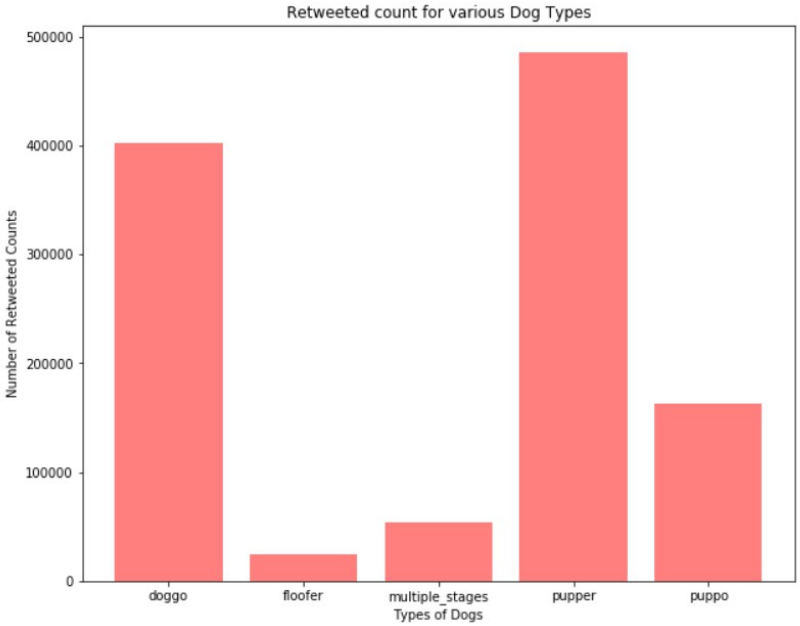
This report summarizes my Exploratory Data Analysis and Findings from the Project “We Rate Dogs – Data Wrangling and Visualization”.

There are so many points to research and many more findings can be made from these datasets, but for me I have taken up the following points to do my exploration analysis:

1. Which dog type is the most and the least popular in terms of favorites and retweets counts?
2. Which dog type is highest rated?
3. Common Names for dogs.
4. Top 5 Highest Predicted Dogs.
5. Relationship between Favorites and Retweets.

Which dog type is the most and the least popular in terms of favorites and retweets counts?

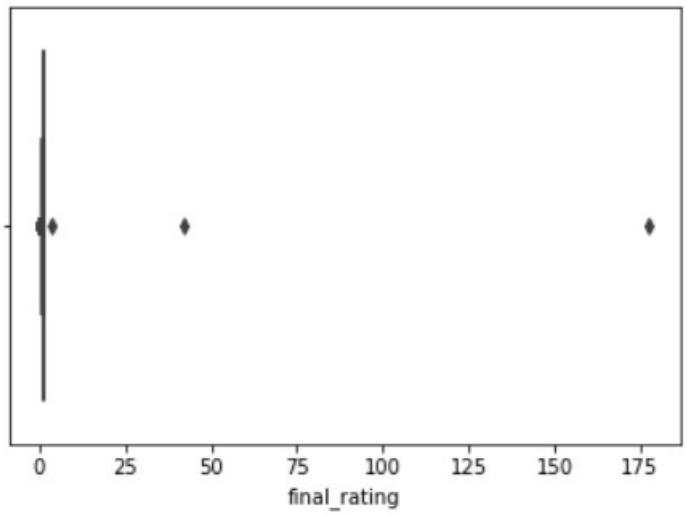




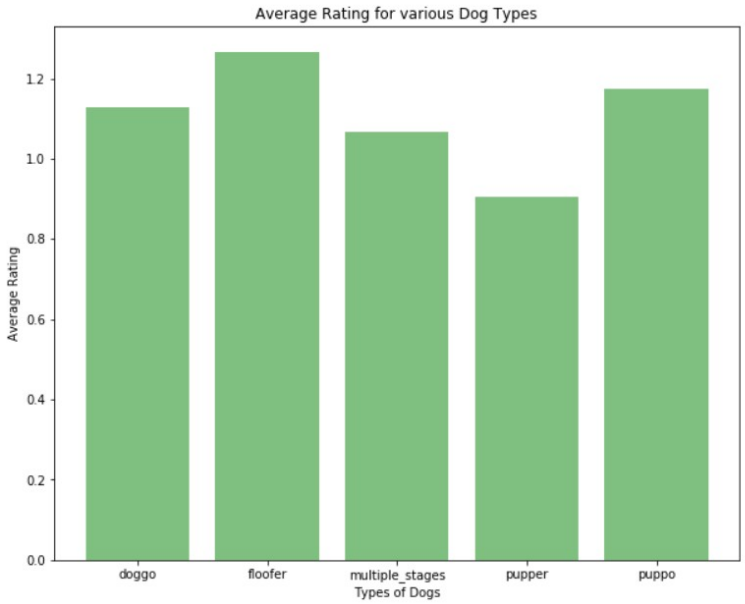
Considering the total number of favorites and retweet counts for each dog type (doggo, floofer, pupper and puppo), pupper with the highest number of both counts (1515783 and 66537) and is the most popular and floofer with the least count (485685 and 24309) is the least popular dog type.

Which dog type is highest rated?

While analyzing this point, there are few outliers I observed.



But I found out that, the corresponding dog_types for these outlier data are NaN, so I am assuming there will be no impact on my analysis due to these outliers.



Calculating the average rating (rating_numerator/rating_denominator) for each dog type and plotting them, we found out that floofer even though is the least popular type, has the highest rating overall (1.266667) and pupper with the lowest average rating (0.904248).

A point to be mentioned related to the above two analysis findings is that, out of 1971 tweets, there are only 322 records that have a valid 'dog_type' information. So, these findings are based on only these 322 records. So, there is high chance that these results can be altered if there are valid data for the remaining records too.

Common Names for dogs:

```
Charlie    16
Oliver     15
Lucy       15
Cooper     14
Daisy      13
Penny      10
Toby       10
Koda       10
Winston     9
Duke        9
Name: name, dtype: int64
```

Above listed is the top 10 most common names.

Top 5 Highest Predicted Dogs:

For this analysis, first I calculated based on the overall prediction level count irrespective of whether it is level1 or level2 or level3. Then, I calculated the highest prediction based on level1 prediction level count only.

Based on Overall Prediction Level Counts:

labrador_retriever	265
golden_retriever	264
chihuahua	178
pembroke	138
cardigan	112

Based on level1 Predicted Level Counts:

golden_retriever	137
labrador_retriever	94
pembroke	88
chihuahua	78
pug	54

So, from the above we can conclude that **labrador_retriever** has been predicted for most number of times considering the overall prediction level (that includes level1, level2 and level3 predictions) and **golden_retriever** is predicted as the most number of level 1 predictions.

Relationship between Favorites and Retweets:

Just by looking at the data values for favorites and retweet counts for the tweets and I was thinking there looks to be some relationship between these two parameters. So, I decided to programmatically and mathematically find if there exists any relationship, if an increase in favorites also increases retweets count and vice versa or anything.

OLS Regression Results

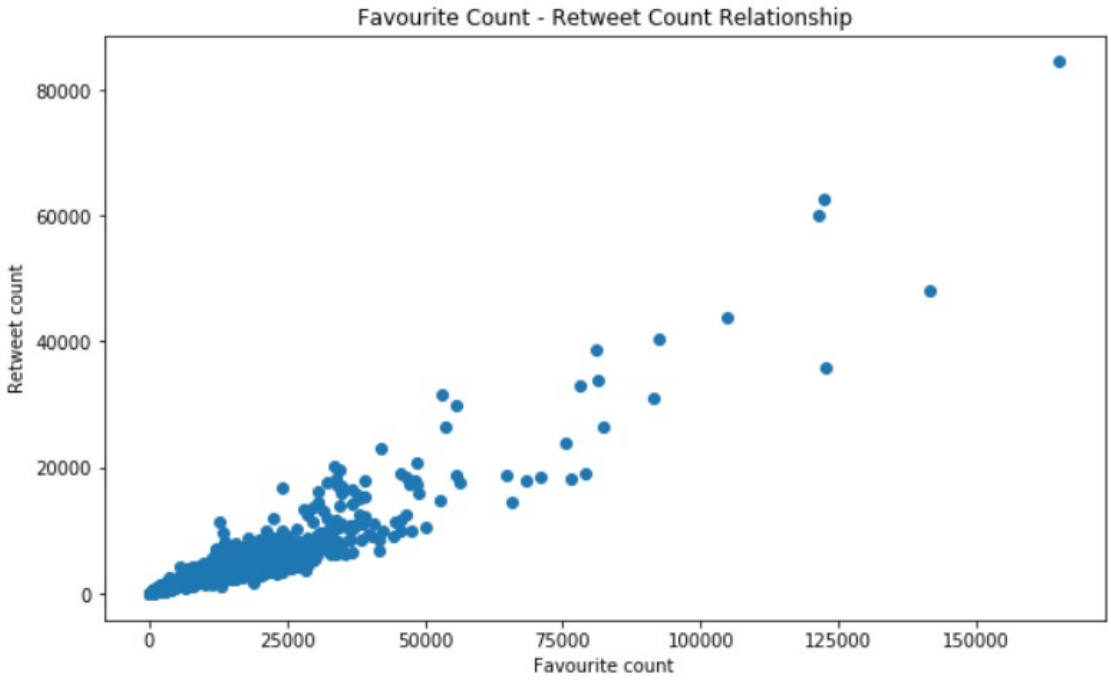
Dep. Variable:	favorite_count	**R-squared:**	0.863
Model:	OLS	**Adj. R-squared:**	0.863
Method:	Least Squares	**F-statistic:**	1.243e+04
Date:	Sat, 20 Oct 2018	**Prob (F-statistic):**	0.00
Time:	18:48:33	**Log-Likelihood:**	-19492.
No. Observations:	1971	**AIC:**	3.899e+04
Df Residuals:	1969	**BIC:**	3.900e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
intercept	2085.8014	123.439	16.897	0.000	1843.716	2327.887
retweet_count	2.5062	0.022	111.497	0.000	2.462	2.550

Omnibus:	515.441	**Durbin-Watson:**	0.766
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	16760.944
Skew:	0.553	**Prob(JB):**	0.00
Kurtosis:	17.243	**Cond. No.**	6.30e+03

So, from the above OLS Regression Results, I can conclude that "**for every increase in 1 retweet count, I can expect the favorite count to be increase by around 2.5 count**."

Scatter plot for these two parameters is shown below:



Again, adding the observation from the scatter plot, I think it is quite evident that there exists a strong relationship between the “Favorites Count” and the “Retweeted Counts”.

Finally, let me present you our highest predicted dog breed “labrador_retriever” with the highest rating.