Data Analysis

Prepared by: Muhammad Usman Siddiqui

Introduction:

This report discusses the analyses conducted to draw insights from data related to the Twitter account WeRateDogs. Visual and statistical methods are used to draw conclusions.

Analysis and Discussion:

One part of the data contains predictions from an algorithm predicting the dog breed where p1 is the first prediction, p2 is the second prediction, and p3 is the third. The histograms of the confidence the algorithm has in its predictions are plotted below.

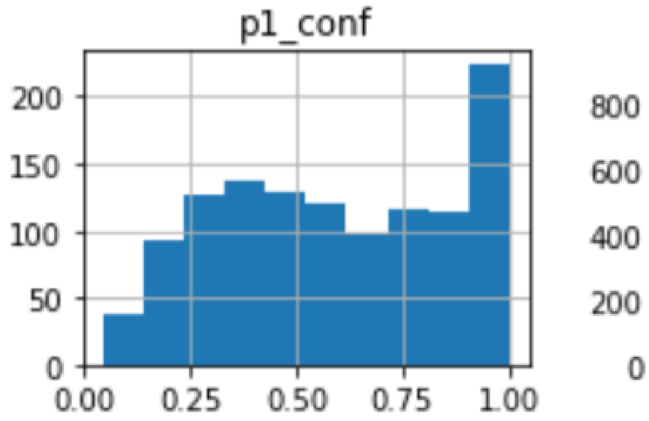


Figure 1 : p1 confidence

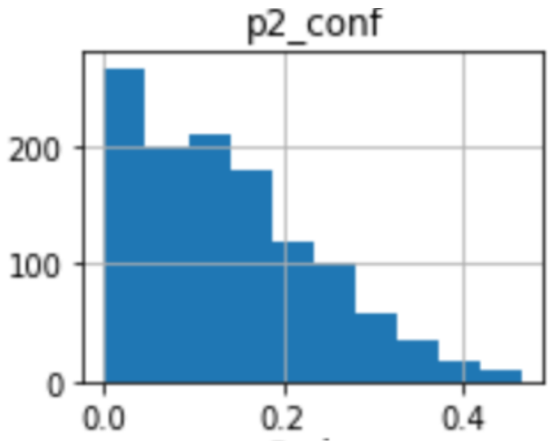


Figure 2: p2 confidence

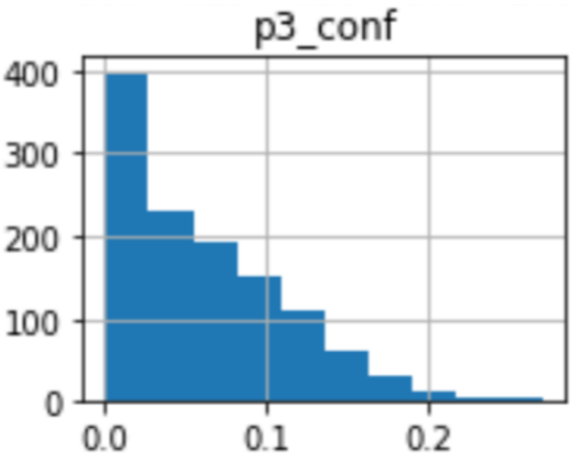
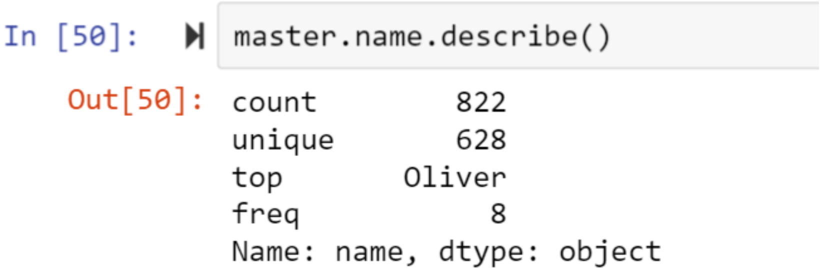


Figure 3: p3 confidence

In the p1 confidence histogram, the highest proportion of values are situated to the right which means most values have a high confidence level. The histograms for both the p2 and p3 confidence levels are skewed to the right. This means that most of the second and third predictions have a low confidence level. The mean confidence is 0.59 for p1, 0.14 for p2, and 0.06 for p3. This again proves that most times the algorithm is most confident in its first prediction.

What is the most common name?

The most common name is found in the data set using the code below.



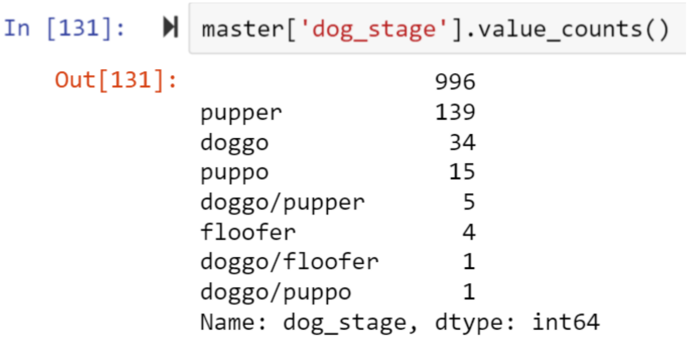
```
In [50]:    master.name.describe()

Out[50]: count         822
         unique        628
         top        Oliver
         freq            8
         Name: name, dtype: object
```

This shows that 76.4%, the vast majority, of the dogs have unique names. In the names that are repeated, Oliver is the most common occurring 8 times.

What is the most common dog stage classification?

The most common dog stage in the data set is determined using the code below.



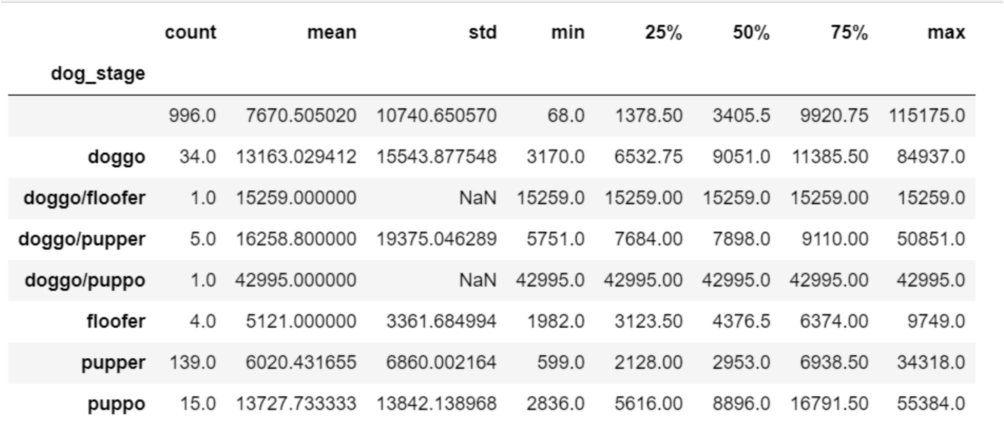
```
In [131]:  master['dog_stage'].value_counts()

Out[131]:                  996
           pupper          139
           doggo            34
           puppo            15
           doggo/pupper      5
           floofer           4
           doggo/floofer     1
           doggo/puppo       1
           Name: dog_stage, dtype: int64
```

This shows that the most common dog stage classification was pupper on WeRateDogs.

Which dog stage do users adore the most?

The table below shows the relation between the dog stage and the favorite count.

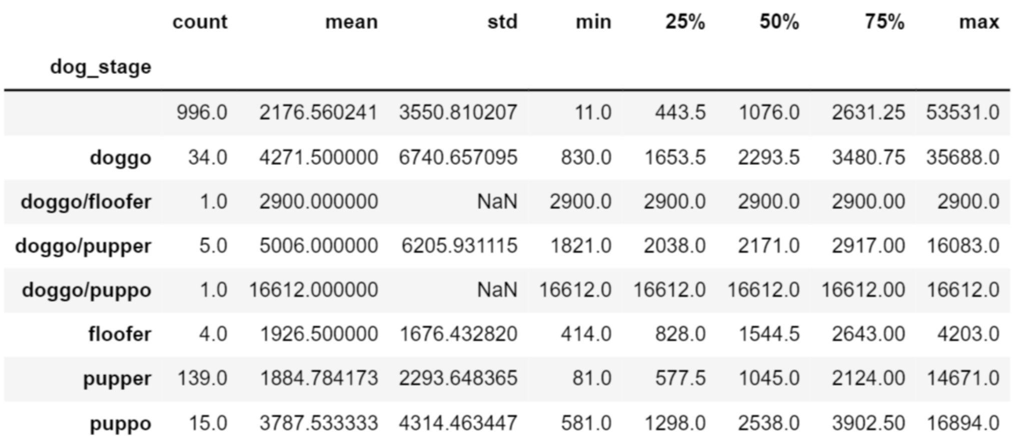


	count	mean	std	min	25%	50%	75%	max
dog_stage								
	996.0	7670.505020	10740.650570	68.0	1378.50	3405.5	9920.75	115175.0
doggo	34.0	13163.029412	15543.877548	3170.0	6532.75	9051.0	11385.50	84937.0
doggo/floofer	1.0	15259.000000	NaN	15259.0	15259.00	15259.0	15259.00	15259.0
doggo/pupper	5.0	16258.800000	19375.046289	5751.0	7684.00	7898.0	9110.00	50851.0
doggo/puppo	1.0	42995.000000	NaN	42995.0	42995.00	42995.0	42995.00	42995.0
floofer	4.0	5121.000000	3361.684994	1982.0	3123.50	4376.5	6374.00	9749.0
pupper	139.0	6020.431655	6860.002164	599.0	2128.00	2953.0	6938.50	34318.0
puppo	15.0	13727.733333	13842.138968	2836.0	5616.00	8896.0	16791.50	55384.0

Table 1: Dog stage and Favourite count

The mean favorite count is the highest for puppo. Note that the dogs with two stages reported are ignored in this analysis because a dog cannot have two stages and its actual stage is not known.

The table below shows the relation between the dog stage and the retweet count.



	count	mean	std	min	25%	50%	75%	max
dog_stage								
	996.0	2176.560241	3550.810207	11.0	443.5	1076.0	2631.25	53531.0
doggo	34.0	4271.500000	6740.657095	830.0	1653.5	2293.5	3480.75	35688.0
doggo/floofer	1.0	2900.000000	NaN	2900.0	2900.0	2900.0	2900.00	2900.0
doggo/pupper	5.0	5006.000000	6205.931115	1821.0	2038.0	2171.0	2917.00	16083.0
doggo/puppo	1.0	16612.000000	NaN	16612.0	16612.0	16612.0	16612.00	16612.0
floofer	4.0	1926.500000	1676.432820	414.0	828.0	1544.5	2643.00	4203.0
pupper	139.0	1884.784173	2293.648365	81.0	577.5	1045.0	2124.00	14671.0
puppo	15.0	3787.533333	4314.463447	581.0	1298.0	2538.0	3902.50	16894.0

Table 2: Dog stage and Retweet count

The mean retweet count is the highest for doggo. It appears that puppo is favorited the most and doggo is retweeted the most. Retweets translate to more engagement. All good posts get favorited, but the best tweets get retweeted. Therefore, the users adored doggo the most with puppo being a close second.

What is the relation between the favorite count and the retweet count?

A scatterplot of the favorite count and the retweet count is plotted.

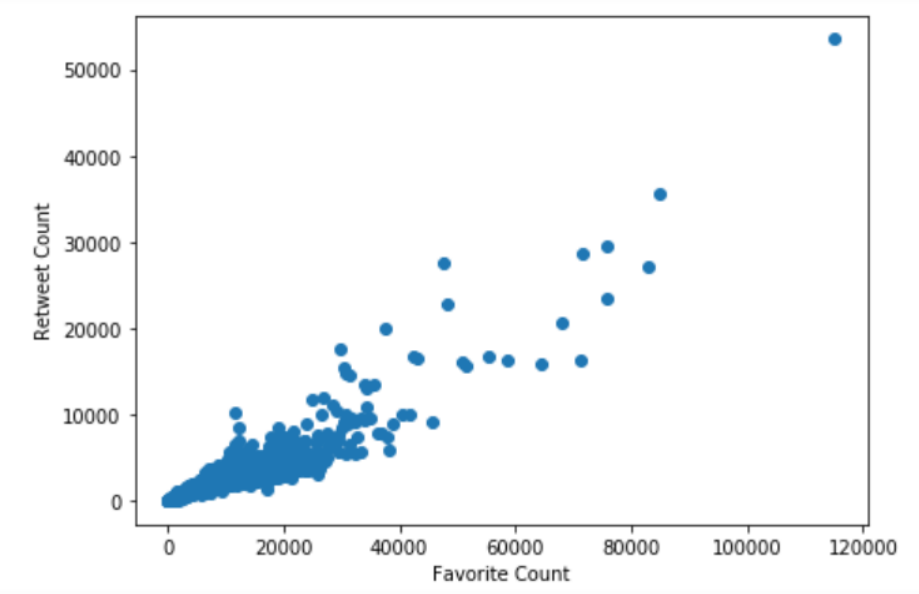


Figure 4: Relation between retweet count and the favorite count

There is a positive correlation between the retweet count and the favorite count. If a tweet has a high favorite count it will likely have a high retweet count as well. It is worth nothing that in general the retweet count is much less than the favorite tweet. This makes sense; a good tweet will get favorited more, and will have a higher chance to get retweeted.