Data Wrangling - Project Two

Ashinze Emmanuel Chidi

Udacity Nanodegree Program

ACT REPORT

ANALYSIS AND VISUALIZATION

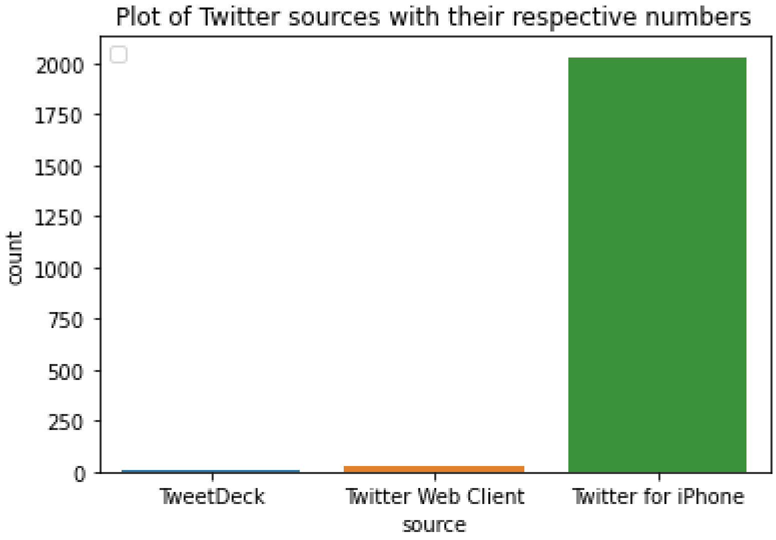
Introduction

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.

WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively to be used in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.Insights will be drawn from this dataset after the data wrangling process. Proper visualizations are used.

Questions

Q1. What Source tends to have the most tweets?



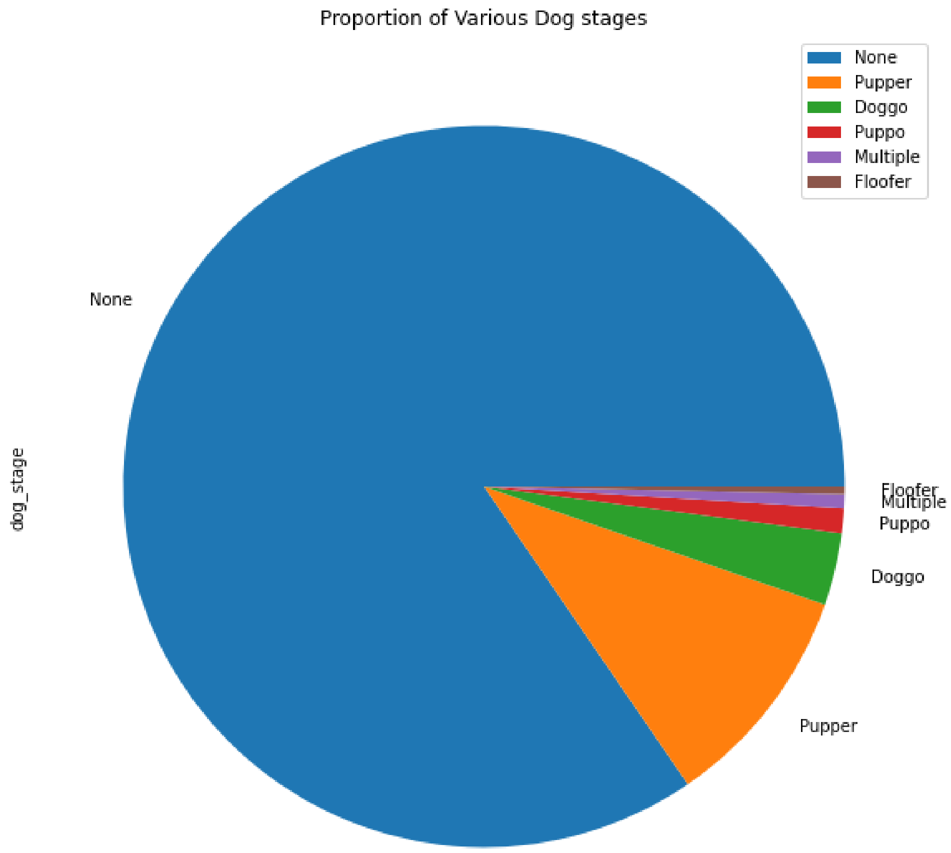
```
Twitter for iPhone    2031
Twitter Web Client      30
Tweet Deck              11
Name: source, dtype: int64
```

iPhone Users tend to send tweets to this page more than other platforms. The difference is so huge that almost all the tweets came from iPhone users. The Tweet Deck platform has the least performance, and this can be as a result of other factors not shown in the dataset.

Q2. Is there any difference in the proportion of the dog stages?

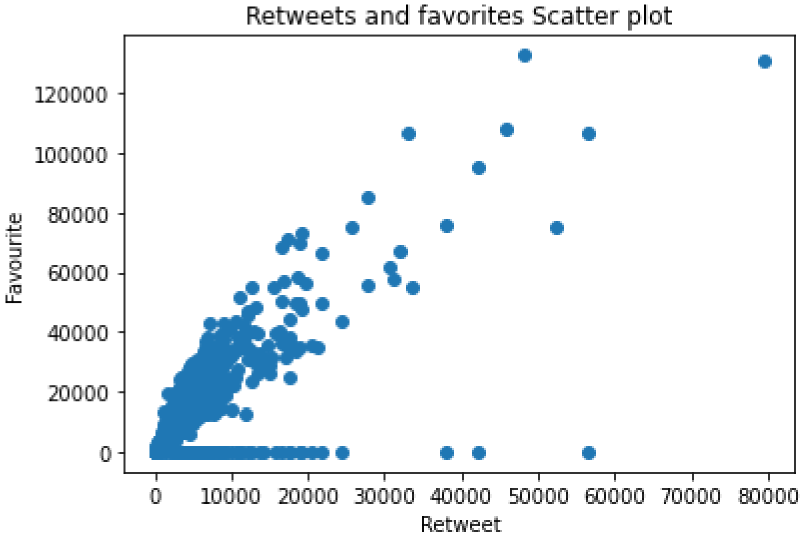
There are various dog stages as follows;

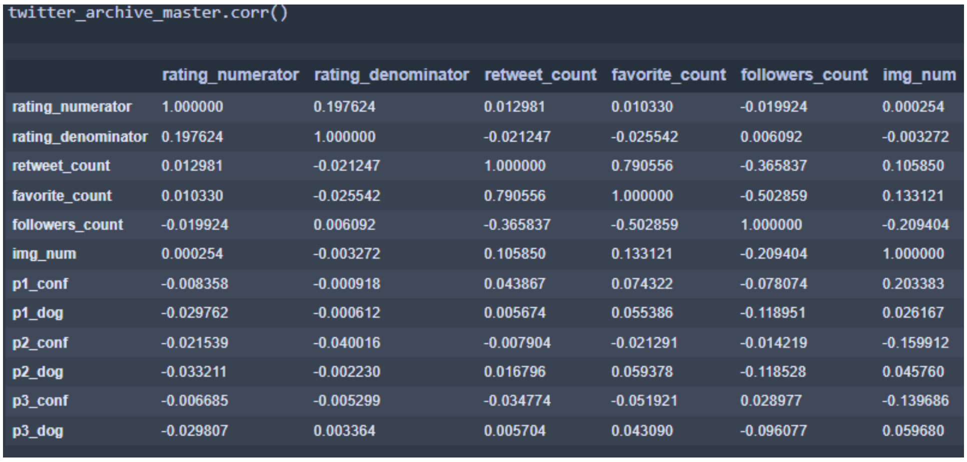
```
None       1752
Pupper      210
Doggo        67
Puppo        23
Multiple     13
Floofer       7


Name: dog_stage, dtype: int64
```



Looking at the visualization, the None datatype has the highest percentage in the dataset and that is as a result of the data quality issues. However, considering the other dog stages, it can be seen that Pupper has the highest percentage and floofer with the lowest percentage when it come to the numbers.

Q3. How does retweet count vary with favourite count?





```
twitter_archive_master.corr()
```

	rating_numerator	rating_denominator	retweet_count	favorite_count	followers_count	img_num
rating_numerator	1.000000	0.197624	0.012981	0.010330	-0.019924	0.000254
rating_denominator	0.197624	1.000000	-0.021247	-0.025542	0.006092	-0.003272
retweet_count	0.012981	-0.021247	1.000000	0.790556	-0.365837	0.105850
favorite_count	0.010330	-0.025542	0.790556	1.000000	-0.502859	0.133121
followers_count	-0.019924	0.006092	-0.365837	-0.502859	1.000000	-0.209404
img_num	0.000254	-0.003272	0.105850	0.133121	-0.209404	1.000000
p1_conf	-0.008358	-0.000918	0.043867	0.074322	-0.078074	0.203383
p1_dog	-0.029762	-0.000612	0.005674	0.055386	-0.118951	0.026167
p2_conf	-0.021539	-0.040016	-0.007904	-0.021291	-0.014219	-0.159912
p2_dog	-0.033211	-0.002230	0.016796	0.059378	-0.118528	0.045760
p3_conf	-0.006685	-0.005299	-0.034774	-0.051921	0.028977	-0.139686
p3_dog	-0.029807	0.003364	0.005704	0.043090	-0.096077	0.059680

The retweet count and favorite count are strongly correlated with a coefficient of almost 0.8. This means that if a dog is part of the favourites, it tends to have more retweets than the other ones.