ACT REPORT

Wrangling Project, 15.09.2019

This report shows the analysis of the *WeRateDogs* page, it is an open Twitter page for dog ratings. This page has an international page used to rate dogs. There are many information about these dogs in form of Data Frames to analyze them. These information include:

Tweets Data Frame:

This table includes the following columns:
Tweet id, Tweet Text, Tweet URL, Ratings, Favorite counts, Retweet Count, source for each tweet (ex, mobile or web) Date and Time for each tweet.

Image Prediction Data Frame:

This table contains some information about dog images. These images are used to be detected by 3 Algorithms. Each algorithm has its results.

This table includes the following columns:
Tweet Id, Image URL, (Prediction, Prediction Confidence, the dog predicted) for each algorithm per Image.

Tools:

In this Analyses Panadas, Matplotlib are used to be able to analyze it.

Analysis and Insights:

I tried to imagine some important question to answer them. These are my questions:

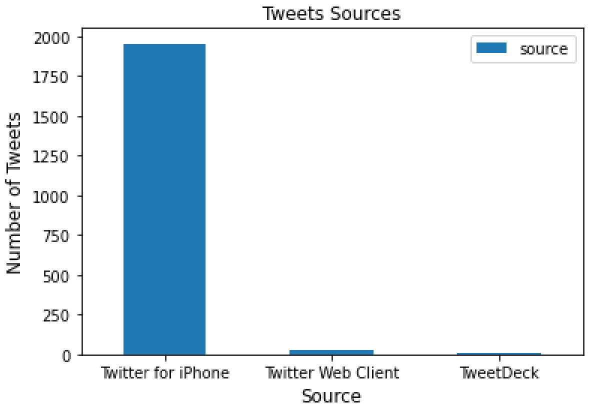
1- How each tweet source is participated in tweets? What is the most used one? What is the mean for retweet count and favorite count grouped by source?
2- Which dog stage is most popular in given tweets? Describe the relation between each dog stage and its favorite count and its retweet count. What do you notice from this description?
3- Which month and year has the most and least favorite count? Which month and year has the most and least retweet count?
4- With name or without name was the retweet counts mean and favorite count mean was higher? Which dog name has the most and least retweet count and favorite count?
5- Which algorithm achiever higher accuracy in the prediction task?

Question 1: How each tweet source is participated in tweets? What is the most used one? What is the mean for retweet count and favorite count grouped by source?

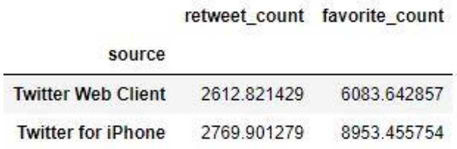
Each source is participated as the following:

```
Twitter for iPhone    1955
Twitter Web Client      28
TweetDeck               11
Name: source, dtype: int64
```

From above analysis, we notice that the most resource used in tweets was the Twitter for iPhone. It has participated with 1955 tweets. The following bar chart show this clearer.



For the second part of question 1 see the following table in the below image:



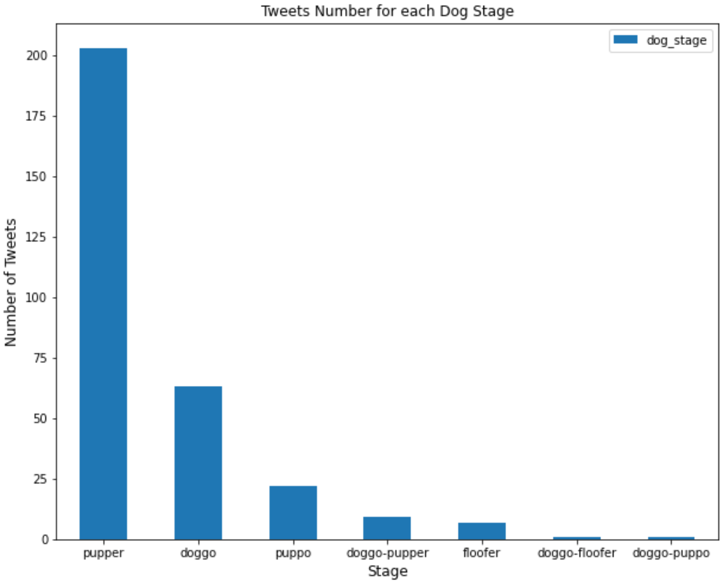
	retweet_count	favorite_count
source		
Twitter Web Client	2612.821429	6083.642857
Twitter for iPhone	2769.901279	8953.455754

From the above image, we can notice that the iPhone has the most retweet count and the most favorite count.

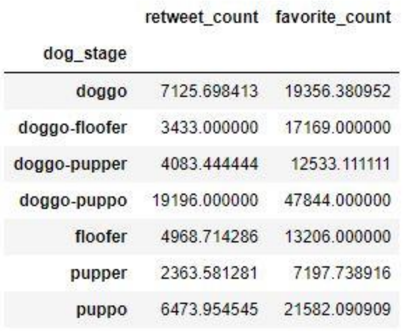
Question 2: Which dog stage is most popular in given tweets? Describe the relation between its favorite count and its retweet count grouped by dog stage. What do you notice from this description?

The most popular dog stage in tweets is Pupper. This is according to the following analysis;

```
pupper           203
doggo             63
puppo             22
doggo-pupper       9
floofer            7
doggo-floofer      1
doggo-puppo        1
```

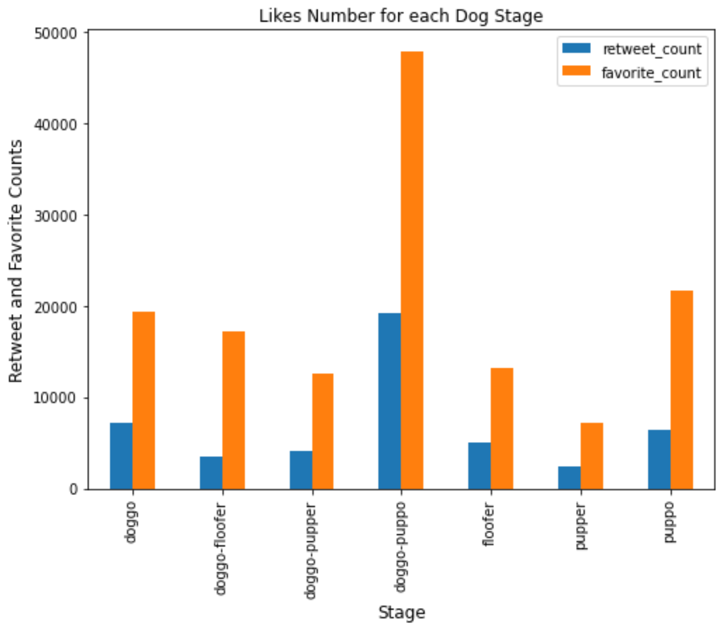


The relation between the retweet count and favorite count grouped by dog stage:



	retweet_count	favorite_count
dog_stage		
doggo	7125.698413	19356.380952
doggo-floofer	3433.000000	17169.000000
doggo-pupper	4083.444444	12533.111111
doggo-puppo	19196.000000	47844.000000
floofer	4968.714286	13206.000000
pupper	2363.581281	7197.738916
puppo	6473.954545	21582.090909

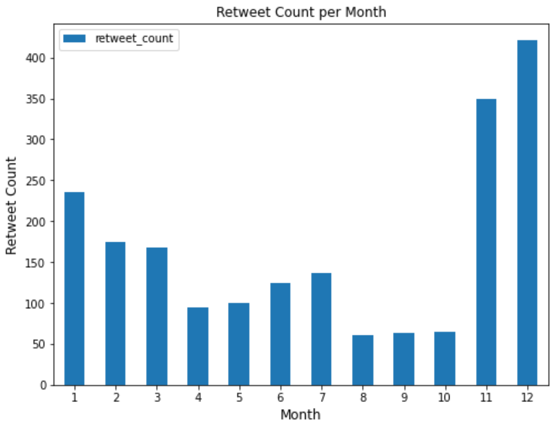
From the above image: the doggo-puppo stage has the most retweet count and the most favorite count. If you like to see them visually, let’s go to the next bar chart.

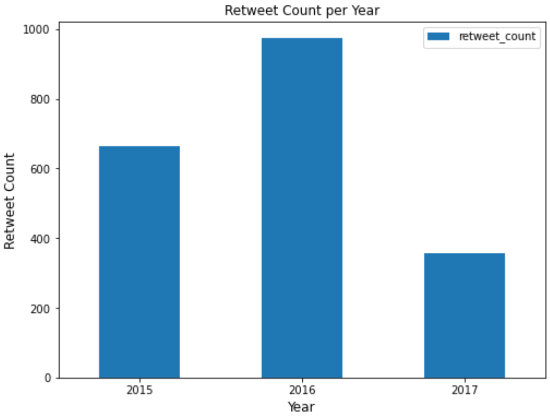


Question3: Which month and year has the most and least favorite count? Which month and year has the most and least retweet count?

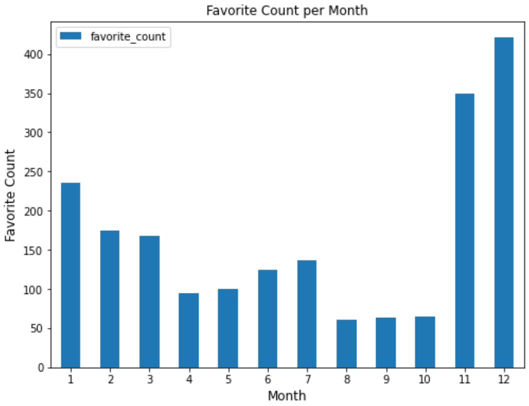
1st : The December has the most retweet count and favorite count and the year 2016 has the most retweet count and favorite count..

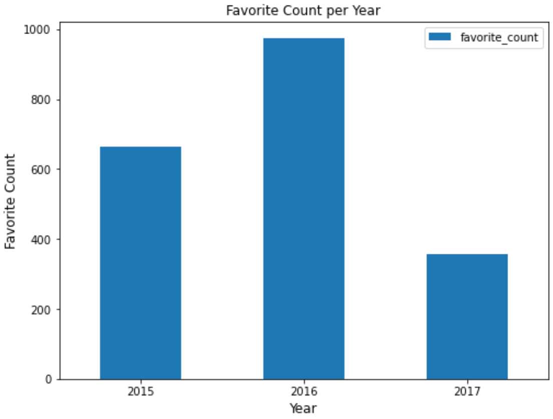
For the retweet counts:



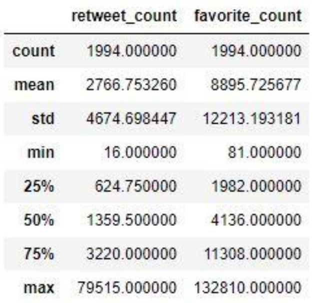


For the favorite counts:





For the description for both:



	retweet_count	favorite_count
count	1994.000000	1994.000000
mean	2766.753260	8895.725677
std	4674.698447	12213.193181
min	16.000000	81.000000
25%	624.750000	1982.000000
50%	1359.500000	4136.000000
75%	3220.000000	11308.000000
max	79515.000000	132810.000000

Question 4: With name or without name was the retweet counts mean and favorite count mean was higher? Which dog name has the most and least retweet count and favorite count?

1st part of question, the mean of the retweet count for tweets without names is higher than tweets with names. But the tweets with names has more mean likes. This is shown in the following result:

```
Mean retweet count for dog with name 2754
Mean retweet count for dog without name     2794

Mean favorite count for dog with name       9414
Mean favorite count for dog without name    7811
```

2nd part of question: The most and least retweet count is None,

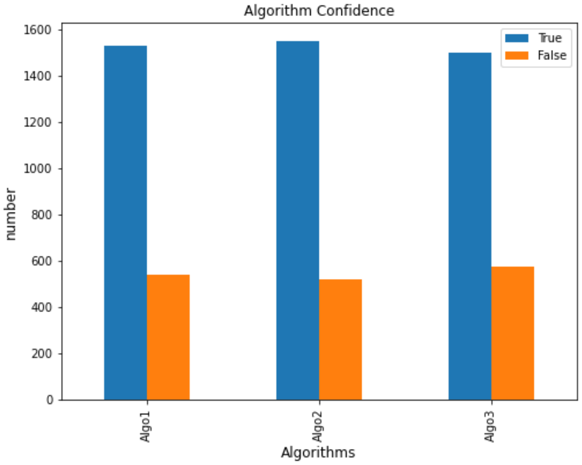
```
The lowest retweet count was for the record number and name as follo
ws: 1977    None
Name: name, dtype: object which received the retweet_count = 16


The highest favorite count was for the record number and name as follow
s: 309    None
Name: name, dtype: object which received the favorite_count = 132810
```

The most and least favorite count are shown in the following:

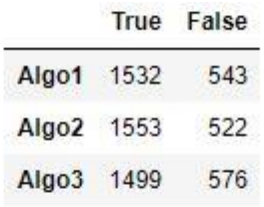
```
The highest favorite count was for the record number and name as fol
lows: 309    None
Name: name, dtype: object which received the favorite_count = 132810

The lowest favorite count was for the record number and name as follows
:
1977    None
Name: name, dtype: object
```

```
which recieved the favorite_count = 81
```

Question 5: Which algorithm achiever higher accuracy in the prediction task?



Since the number of data in rows are equal, so the highest true is the highest performance.



	True	False
Algo1	1532	543
Algo2	1553	522
Algo3	1499	576