WeRateDogs Dataset analysis after wrangling process

by Theodora Koutsothanasi

Do you love dogs?

In this project we try to find some insights and make interesting visualizations from the WeRateDogs Dataset. In case you do not know what is WeRateDogs, you can learn now! WeRateDogs is a twitter account that rates people's dogs and also adds a funny comment about the dog. Dogs are rated on a scale of one to ten, but most of the times the ratings are above 10. In similar cases, like rating we identified some issues with the data, so we cleaned the dataset and now we can analyze it and find some intuitive insights.

Let's Start

First, our dataset contains 1961 tweets and 23 columns with different variables, such as the source of the tweet, the rating of the dog, number of retweets etc.

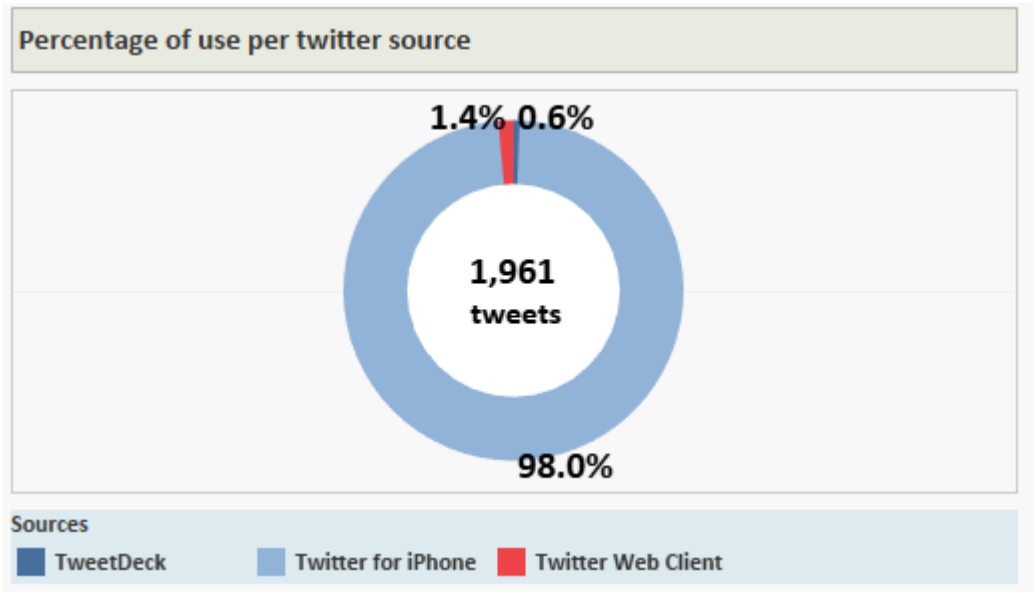
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1961 entries, 0 to 1960
Data columns (total 23 columns):
tweet_id              1961 non-null int64
timestamp             1961 non-null object
source                1961 non-null object
text                  1961 non-null object
expanded_urls         1961 non-null object
rating_numerator      1961 non-null int64
rating_denominator    1961 non-null int64
name                  1961 non-null object
retweet_count         1961 non-null int64
favorite_count        1961 non-null int64
dog_stage             293 non-null object
rating_n              1961 non-null float64
jpg_url               1961 non-null object
img_num               1961 non-null int64
prediction_one        1961 non-null object
confidence_one        1961 non-null float64
isdog_one             1961 non-null bool
prediction_two        1961 non-null object
confidence_two        1961 non-null float64
isdog_two             1961 non-null bool
prediction_three      1961 non-null object
confidence_three      1961 non-null float64
isdog_three           1961 non-null bool
dtypes: bool(3), float64(4), int64(6), object(10)
memory usage: 312.2+ KB
```

Some Basic Statistics for the Dataset

	tweet_id	rating_numerator	rating_denominator	retweet_count	favorite_count	rating_n	img_num	confidence_one
count	1.961000e+03	1961.000000	1961.000000	1961.000000	1961.000000	1961.000000	1961.000000	1961.000000
mean	7.357626e+17	12.228455	10.479857	2769.170321	8907.657828	11.697517	1.202448	0.593877
std	6.751967e+16	41.739741	6.870651	4682.802592	12238.973877	41.010238	0.559987	0.272077
min	6.660209e+17	0.000000	2.000000	16.000000	81.000000	0.000000	1.000000	0.044333
25%	6.758228e+17	10.000000	10.000000	624.000000	1971.000000	10.000000	1.000000	0.362925
50%	7.084699e+17	11.000000	10.000000	1360.000000	4110.000000	11.000000	1.000000	0.587372
75%	7.877176e+17	12.000000	10.000000	3227.000000	11363.000000	12.000000	1.000000	0.846986
max	8.924206e+17	1776.000000	170.000000	79515.000000	132810.000000	1776.000000	4.000000	1.000000

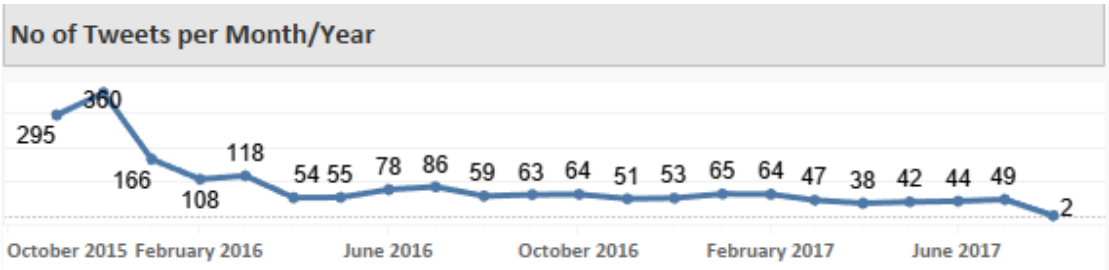
After our first view of the data, let's answer some questions.

1) Which source is mostly used?



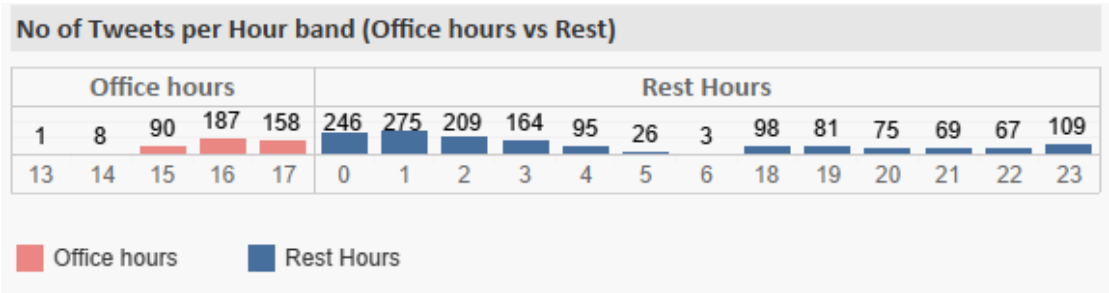
It seems that Twitter for iPhone is the winner, as 98% of the tweets coming from this source.

2) Which is the trend of the tweets throughout the years and months?



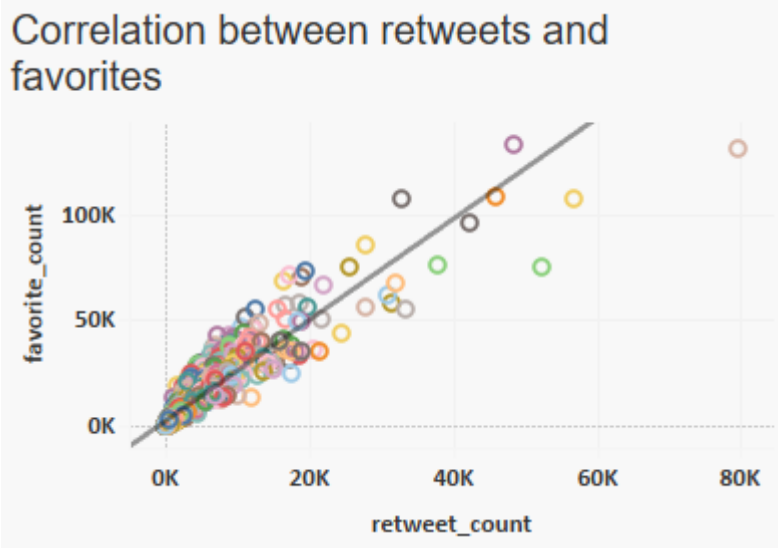
The peak of tweets is on December 2015, and after that the other months have been stable with a declining trend.

3) Which hours of the day do we have greater amount of tweets?



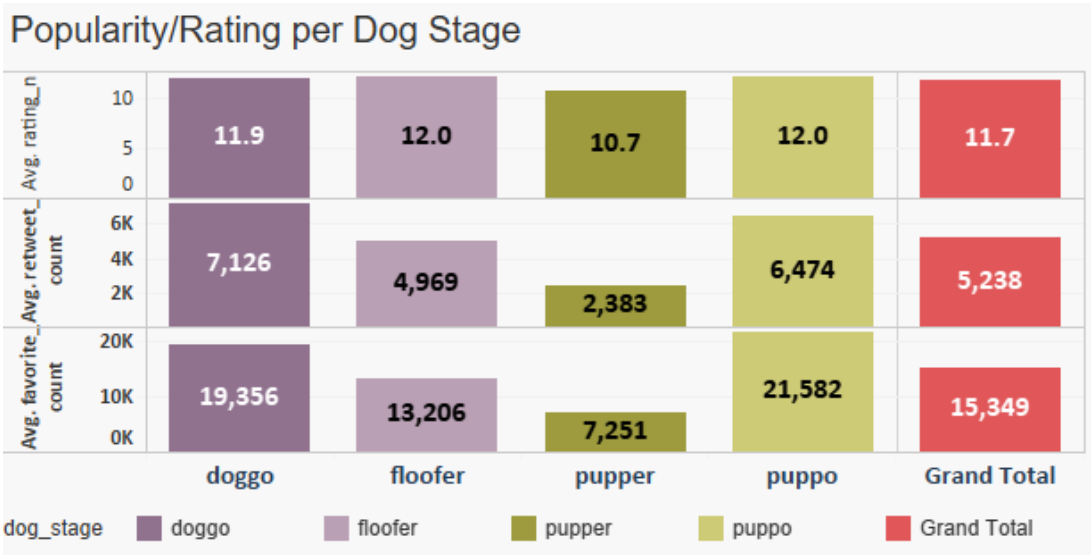
As we expected, in office hour time we have less tweets than the rest hours. 0:00,1:00 and 2:00 AM are the most popular hours.

4) Is there any correlation between retweets and favorites?



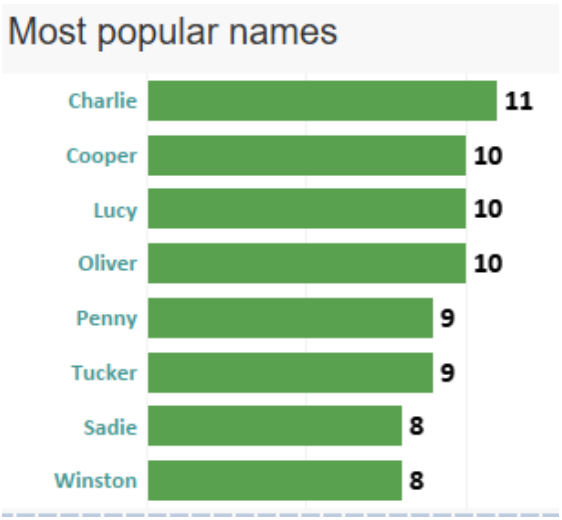
The correlation between retweets and favorites is 91.22%.The figure shows a very strong tendency for number of retweets and number of favorites to both rise above their means at the same time. I expected that rating will have also a positive correlation with retweets or favorites, but finally their correlation is very low. Probably, rating is a subjective matter.

5) Which dog stage is the most popular, according to number of retweets, favorites and rating?



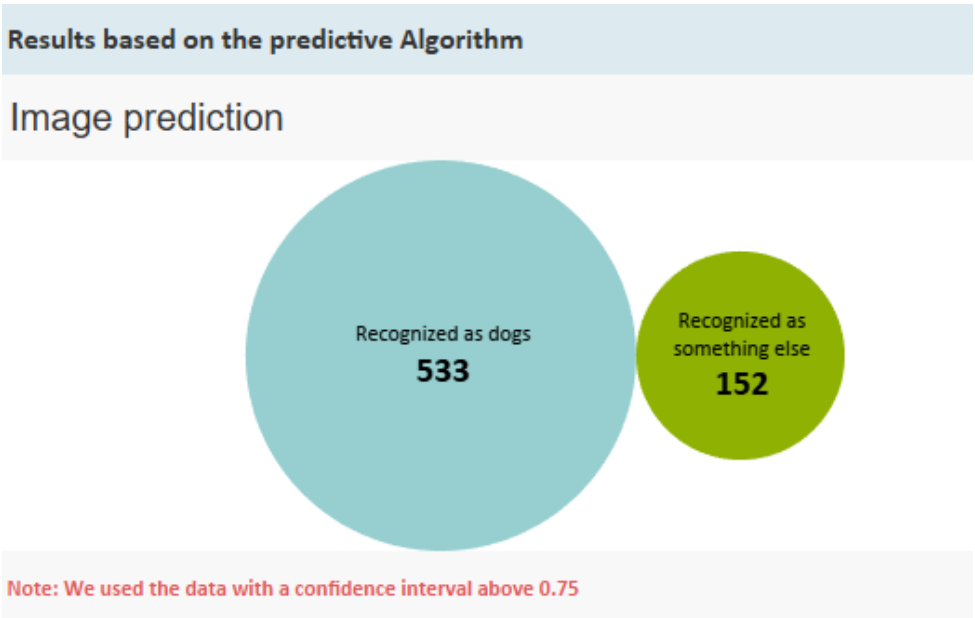
Puppo and floofer stages have the greatest rating among the other dog stages. Puppo has also the greatest number of retweets and favorites, make it the most favorable dog stage.

6) Which are the top 10 names according to their number of tweets?



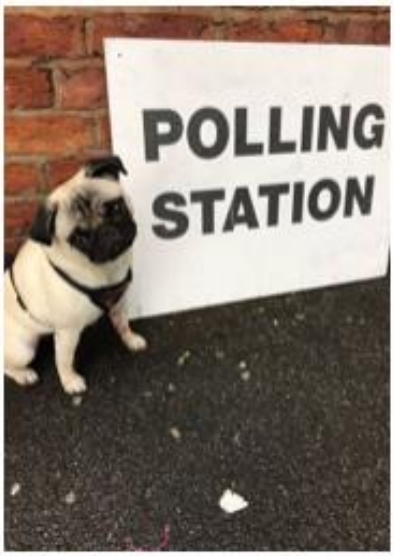
Charlie,Cooper,Lucy and Oliver are the most common dog names in this dataset.

7) How many images predicted dogs in a confidence rate of 75%? Does the algorithm predicts correct if the image contain dog?



In a confidence interval of 0.75, we have 685 observations. 533 recognized that contain dogs and the remaining 152 as something else.

I randomly selected 5 images that the algorithm predicted dog and actually the were dogs!



8) Which dog breeds have better chances to be recognized by the algorithm?

```
prediction_one
Golden_retriever             75
Pembroke                     47
Labrador_retriever           40
Pug                          34
Chihuahua                    25
Samoyed                      22
Pomeranian                   22
Chow                         16
French_bulldog               14
Toy_poodle                   13
Malamute                     10
German_shepherd               9
Maltese_dog                   8
Shetland_sheepdog             8
Chesapeake_bay_retriever      7
Name: tweet_id, dtype: int64
```

Again, we sorted the observations that had confidence interval 0.75 and the algorithm predicted dog. The gold retriever is by far the most recognizable dog breed.

Conclusion

Either you prefer doggos,puppers,floofers or puppos, you have the opportunity to rate your dog. Puppos is the most famous dog stage with a rating at 12. Rate your own dog, add funny images and comments,share them with your beloved people.

Have fun!