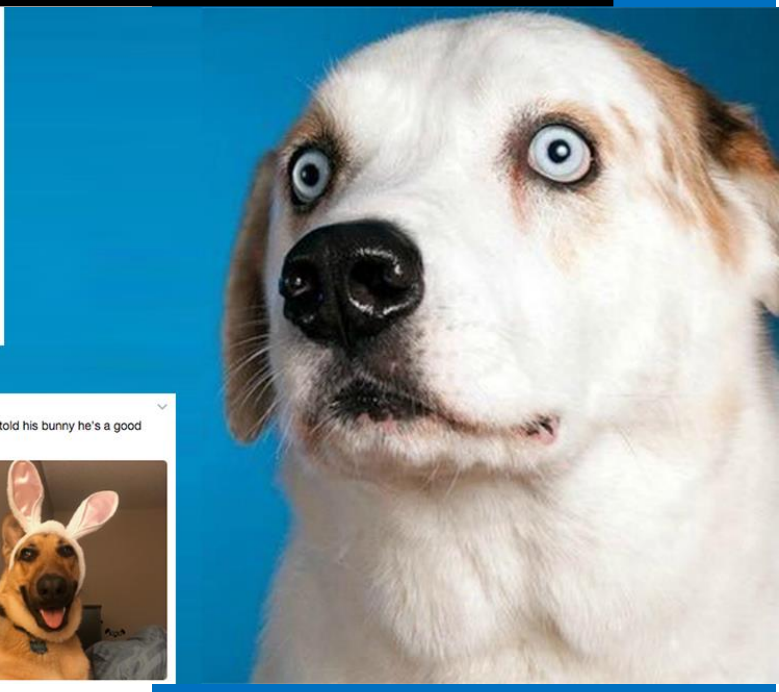
Interesting Facts about WeRateDogs



Introduction

This small article is going to discuss some interesting facts on the filtered dataset collected from WeRateDogs archive mainly, these facts were derived from analysis and visualization of answers for some curious questions.

Which dog stage receives more attention?

As I was check my master dataset, the first question I asked myself was this question. Attention seemed undescriptive enough to tell me what dog stage would people be excited to see and share that dog's tweet to let their friends see it, so in order to give a general claim I need to support it with statistics on how WeRateDogs and followers would react to tweets of different dog stages, and that would be by knowing two things:

1. Which dog stage gets more favorites and retweets on average?
2. Which dog stage gets higher ratings on average?

favorite_count

dog_stage	count	mean	std	min	25%	50%	75%	max
doggo	73.0	19044.164384	21359.955241	2593.0	8295.00	12157.0	20275.00	131075.0
floofer	8.0	13701.375000	10516.926166	2262.0	5560.75	11879.0	18623.25	33345.0
pupper	209.0	7479.019139	10857.577740	693.0	2459.00	3397.0	8282.00	106827.0
puppo	23.0	22723.913043	27931.824108	3277.0	7067.00	15359.0	21977.50	132810.0

retweet_count

max	75%	50%	25%	min	std	mean	count
79515.0	5237.00	3128.0	2072.0	725.0	12183.189954	6941.452055	73.0
18497.0	4130.75	3349.0	2167.0	496.0	5732.138787	4776.750000	8.0
32883.0	2525.00	1258.0	710.0	103.0	3671.067529	2459.622010	209.0
48265.0	7541.50	3220.0	1721.0	716.0	10408.775240	7027.086957	23.0

Figure 1: show some basic statistics for favorite count and retweet count for each dog stage

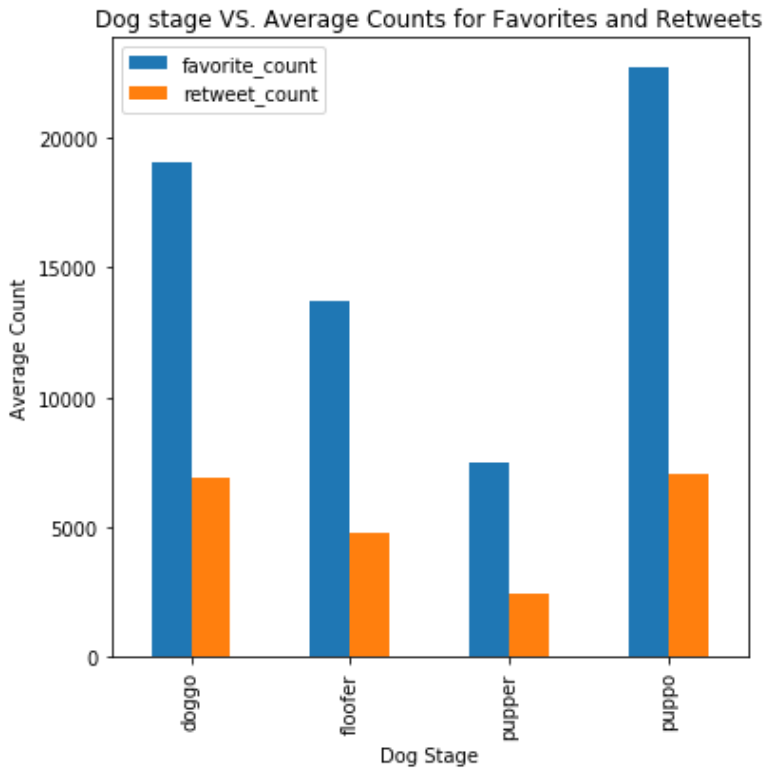


Figure 2: graph shows Dog stage VS. Average Count for Favorites and Retweets

Interesting couple of things in figure 1 and 2. Despite Puppos were second less dog stage in tweets number with only 23 tweets and also they came in second place for maximum retweets count after Doggos, they had the highest average count for both favorites and retweets.

max	75%	50%	25%	min	std	mean	count	dog_stage
14.0	13.0	12.0	11.0	10.0	1.174284	12.057971	69.0	doggo
13.0	13.0	12.0	11.0	10.0	1.125992	11.875000	8.0	floofer
14.0	12.0	11.0	10.0	10.0	1.088604	11.309524	168.0	pupper
14.0	13.0	12.5	12.0	10.0	1.139606	12.181818	22.0	puppo

Figure 3: show some basic statistics rating numerator for each dog stage

Note that here you can find that number of tweets have decreased a little, that's because we applied some filters: on denominator values to be only 10 as there were some outliers, same happened to

nominator values as we only choose our interval to be between [10,14] excluding also nominator outliers as a result only maximum nominator value will be 14 as shown in figure 3.

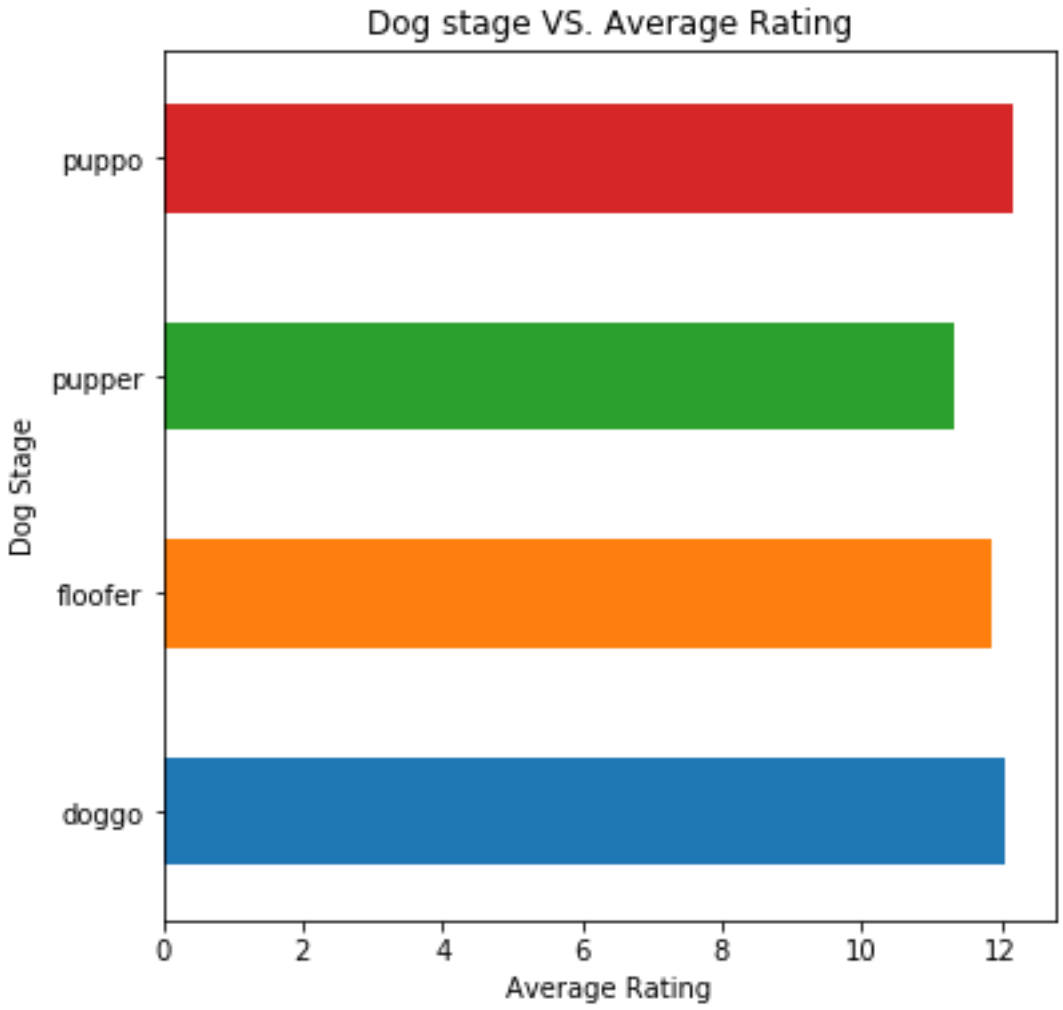


Figure 4: graph shows Dog stage VS. Average Rating

Its clear that Puppos get higher average rating of 12.18% also !

Now we are confident that Puppos tweets receive more attention than any other dog stage, that was really interesting insight for me as majority of people (including me) love cute Puppos!

Asking more questions about the master dataset lead me to wonder, what about tracking WeRateDogs account activity, so I asked the following question.

When does WeRateDogs account usually tweet its tweets?

This question can help us track the account activity which can help to detect when it's going to tweet and why some specific periods have much higher tweets than others which can led to further analysis, to extract right data from timestamp column in the master dataset we need to answer following questions:

1. What hour, day of week, week, and month does WeRateDogs more likely to tweet at/on ?
2. For further curiosity, what year and date had highest tweets percentage?

Here we are dealing with tweets percentage rather that tweets count to give slightly nicer representation for tweeting activity

Starting with HOURS

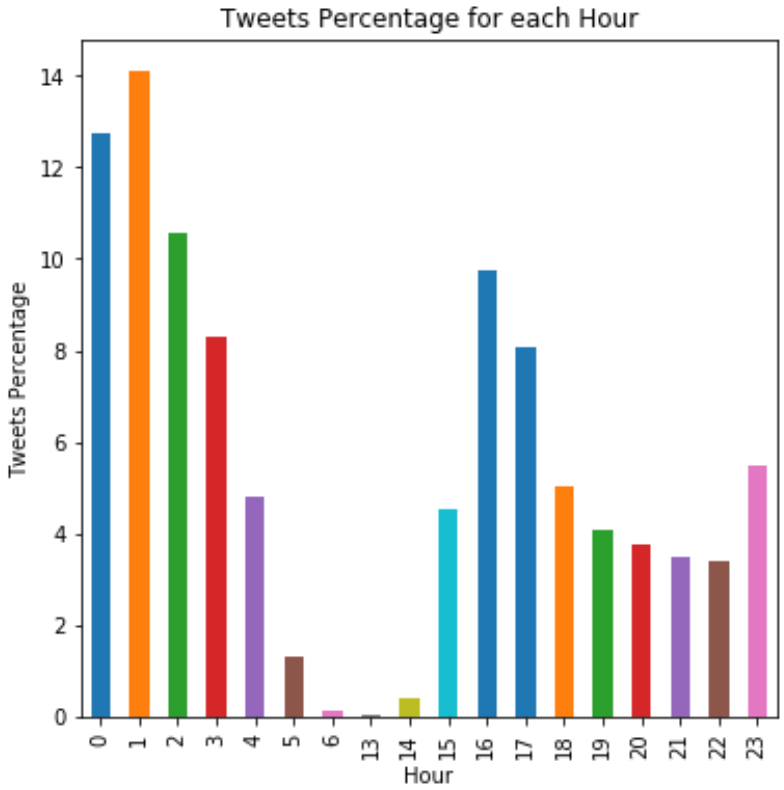


Figure 5: graph shows tweeting activity for each Hour

So based on figure 5, highest tweets percentage found at 1am with 14% , the figure also shows that from 7am to 12pm no tweets where tweeted at all, also we can find that majority of tweets came clearly to two main intervals:

Interval 1 : from 12am to 3am with 45.6% of tweets.

Interval 2 : from 3pm to 6am with 27.4% of tweets.

So these two intervals only have more than 75% of the total tweets.

Now let's take look at DAYS OF WEEK

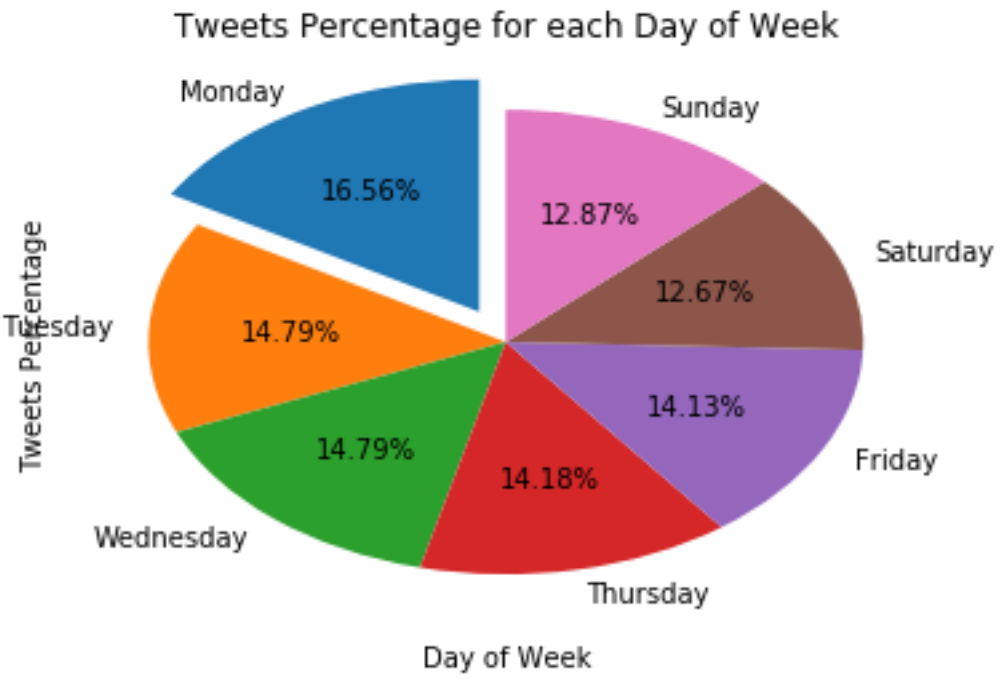


Figure 6: chart shows tweeting activity for each Day of Week

The above pie chart shows that middle days of week are almost having the same tweets percentage, lucky Monday had 16.6% of tweets and as we go down weeks days we reach Saturday and Sunday with 12.7% and 12.9% respectively.

So generally week start with higher tweeting activity and activity goes down until it reaches weekend which found to be having lowest activity .

Now let's consider WEEKS

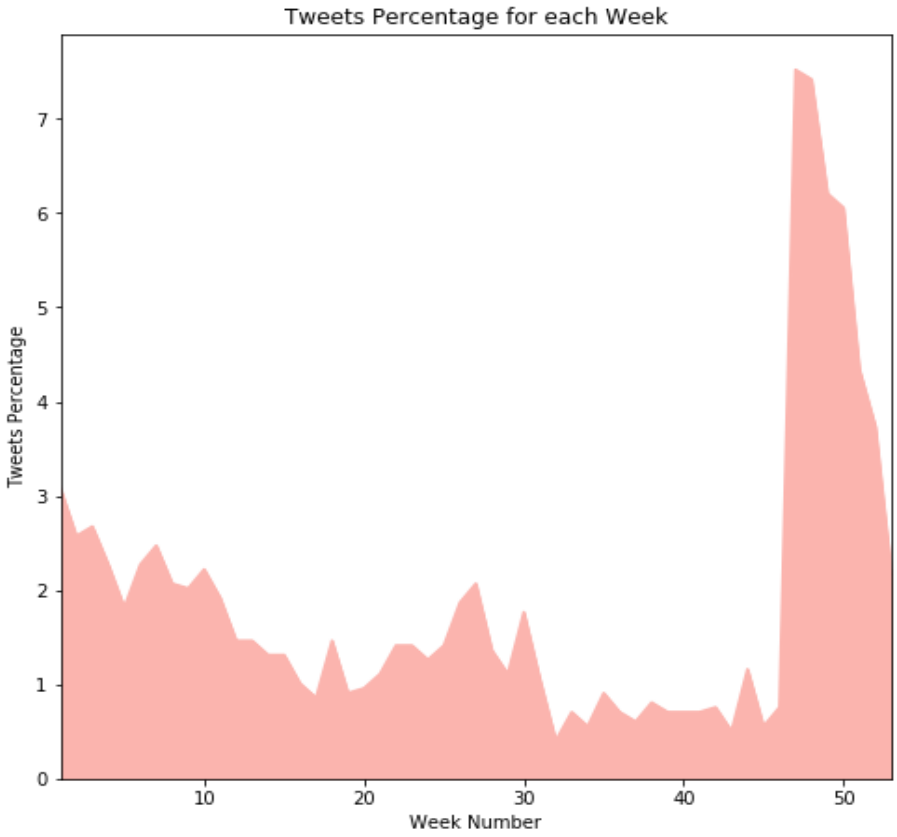


Figure 7: graph shows tweeting activity based on WEEK no.

Wow big spike at last weeks of the year showing that activity suddenly increased from less than 1% for week no. 32 to week no. 45 to more that 7.5% at week no. 47! We need further analysis to explain why would this magnificent change in tweets percentage happened in last weeks.

Let’s consider MONTHS

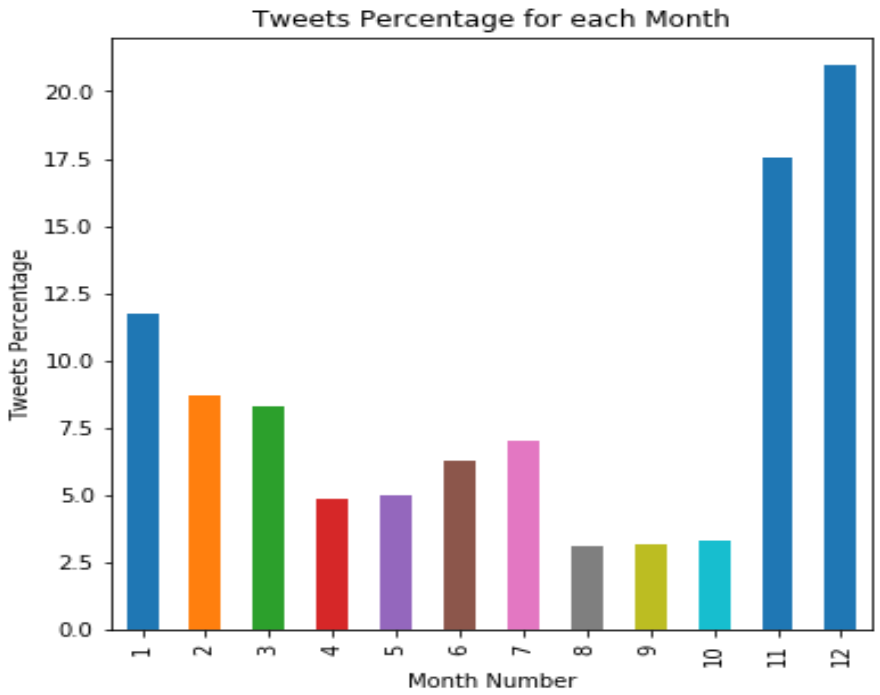


Figure 8: graph shows tweeting activity based on Month no.

Nothing surprising as we already knew that last weeks of the year had mush higher activity but we can ask this; why AUG,SEP,OCT have much lower activity than other months?, Noticing that DEC was the highest month of tweets percentage with almost 20% of all tweets! Is this because of Charismas?!, maybe! I guess as we said before further analysis would be required to answer such questions.

For further curiosity why don’t we also see YEARS and DATES

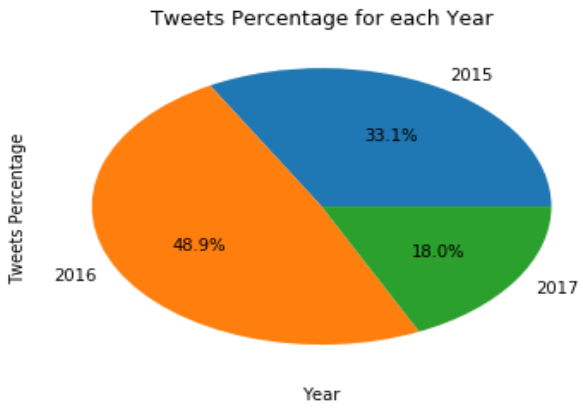


Figure 9: chart shows tweeting activity based on Year no.

1.312468	2015-11-28
1.312468	2015-11-29
1.261989	2015-11-16
1.161030	2015-11-19
1.110550	2015-12-01
1.009591	2015-11-25
1.009591	2015-12-07
1.009591	2015-11-20
1.009591	2015-11-22
1.009591	2015-11-23
0.959112	2015-12-08

Figure 10: table shows Dates with highest activity

So based on the year pie chart, almost 50% of tweets were tweeted in 2016. The table in figure 10 shows dates with highest activity and the interesting insight that two days in a row which are 2015-11-28 and 2015-11-29 had 1.31% of all tweets each!

After surfing our master dataset and finding some interesting insights we concluded from our analysis and visualization, it's time to end our article with a final insight on image-prediction-cleaned dataset with the following two questions.

What is the average confidence level and how likely the prediction is for a dog breed (is_dog = True) for each prediction number?

confidence		
	is_dog	prediction_num
0.540167	False	1
0.613823	True	
0.117090	False	2
0.140470	True	
0.056893	False	3
0.061642	True	

Figure 11: confidence level and if the prediction was for a dog for each prediction num

It’s pretty obvious that predictions for dog breeds is associated with higher confidence levels for each prediction num, for example: predicting a dog breed based on the 1st prediction is about 61% true on average and for the same prediction num predicting other things than dogs is 54%

Now we can wrap up our article with this last question.

What is the easiest dog breed to detect (highest confidence level) for each prediction number?

confidence	prediction	is_dog	prediction_num
1.000000	Zebra	False	1
0.999956	Yorkshire_Terrier	True	
0.488014	Wood_Rabbit	False	2
0.467678	Yorkshire_Terrier	True	
0.255182	Zebra	False	3
0.273419	Yorkshire_Terrier	True	

Figure 12: easiest prediction for each prediction number, whether the image is fora dog or not

So in each prediction number Yorkshire Terrier dog breed is the easiest to predict (highest confidence level), on the other hand for images other than dogs, Zebra is easiest to detect on both prediction number 1 and 3, and wood rabbit for prediction number 2.