Report: act_report

After assessing and cleaning the twitter account WeRateDogs archive datasets, We are ready to make some analysis and visulizations.

Analyzing and Visualizing Data

We did this by answering the following questions:

Q1 : What is the most favorite and retweeted tweet ?

Q2 : What are the top 10 rated tweet names?

Q3 : What are the top 10 rated breads ?

Q4 : Is rating correlated to favorite_count and retweet_count ?

Q5 : What are the top 10 most used dog names ?

Q6 : What is the most used source of tweets ?

Q1 : What is the most favorite and retweeted tweet ?

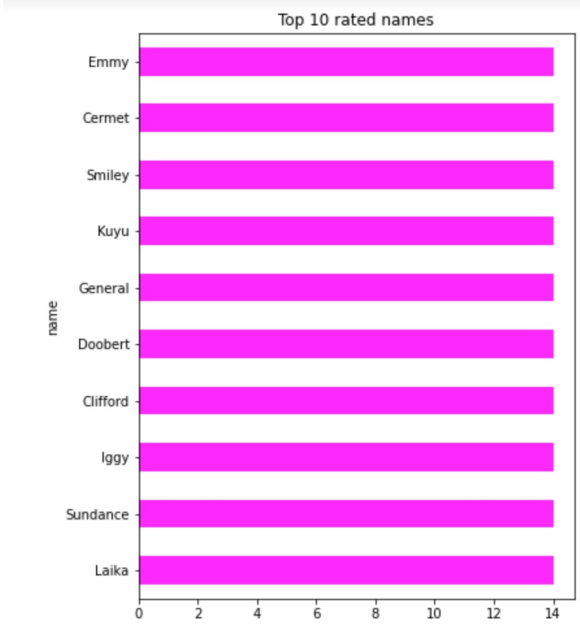
The favorite and most retweeted tweet from the dataset has the tweet_id **744234799360020481** posted on an `iPhone` , on the 2016-06-18 at 18:26:18, the dog was at the `doggo` age_stage and was rated **13/10**.

	tweet_id	tweet_date	source	text	name	retweet_count	favorite_count	bread	age_stage	rating
762	744234799360020481	2016-06-18 18:26:18+00:00	iPhone	Here's a doggo realizing you can stand in a po...	NaN	70365	144303	Labrador_retriever	doggo	1.3

Q2 : What are the top 10 rated tweet names?

The top 10 rated names in our dataset are: `Sundance` , `Laika` , `Clifford` , `Iggy` , `Smiley` , `General` , `Emmy` , `Kuyu` , `Cermet` , `Doobert` with an average rating of `14/10` each.

It is more clear in this horizontal bar chart.



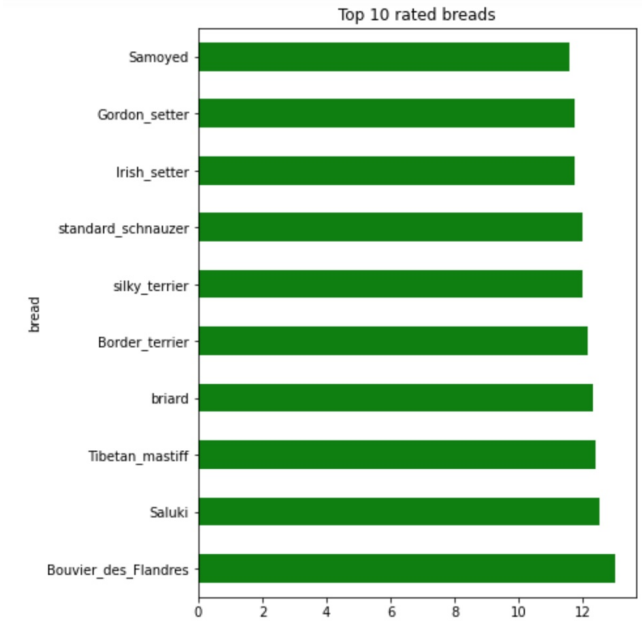
Here is an example of tweet text for `Laika` posted on the `2016-11-03` at `15:51:10`, with tweet id `794205286408003585`:

*'This is Laika. She was a space pupper. The first space pupper actually. Orbited earth like a h*ckin boss. 14/10 hero af https://t.co/trSjgY3h4g'*

This tweet was retweeted `3102` times with a favorite count of `8741`.

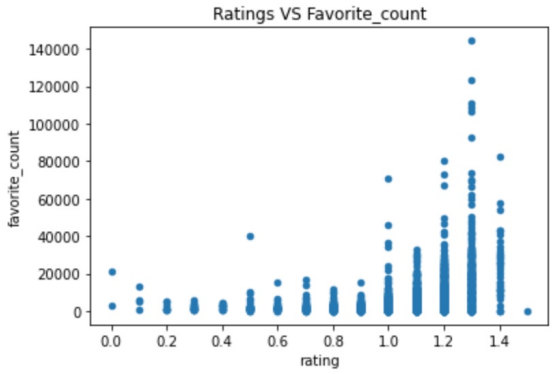
Q3 : What are the top 10 rated breads ?

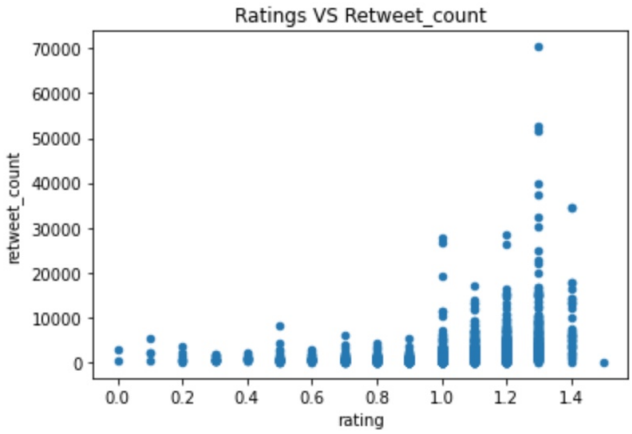
The top 10 rating breads are `Bouvier_des_Frandes`, `Saluki`, `Tibetan_mastiff`, `briard`, `Border_terrier`, `silky_terrier`, `standard_schnauzer`, `Irish_setter`, `Gordon_setter`, and `Samoyed` with rating from `13/10` to `12/10` respectively as shown in this horizontal bar plot.



Q4 : Is rating correlated to favorite_count and retweet_count ?

We can see on these scatter plots, that there is a positive but not high correlation between `rating` and `favorite_count`, then `rating` and `retweet_count`.





The less correlation between these variables can also be explained with the insight of the first question, We discover that the tweet with the id `744234799360020481` was the favorite and the most retweeted tweet, but not the highest rating score.

Q5 : What are the top 10 most used dog names ?

The top 10 most used names in tweet are `Charlie`, `Cooper`, `Oliver`, `Tucker`, `Lucy`, `Penny`, `Winston`, `Sadie`, `Lola`, and `Toby`

```
Charlie    10
Cooper     10
Oliver     10
Tucker      9
Lucy        9
Penny       9
Winston     8
Sadie       8
Lola        7
Toby        7
Name: name, dtype: int64
```

All tweets with dog's name `Charlie` were posted on iPhone with rating numerators greater or equal to 10.

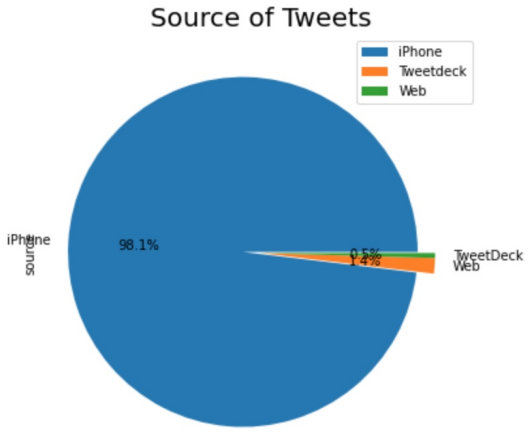
	tweet_id	tweet_date	source	text	name	retweet_count	favorite_count	bread	age_stage	rating
193	844580511645339650	2017-03-22 16:04:20+00:00	iPhone	This is Charlie. He wants to know if you have ...	Charlie	2807	15264	NaN	NaN	1.1
198	842846295480000512	2017-03-17 21:13:10+00:00	iPhone	This is Charlie. He's wishing you a very fun a...	Charlie	3311	14191	Labrador_retriever	NaN	1.3
241	833826103416520705	2017-02-20 23:50:09+00:00	iPhone	Meet Charlie. She asked u to change the channe...	Charlie	3483	17047	Chihuahua	NaN	1.3
249	832369877331693569	2017-02-16 23:23:38+00:00	iPhone	This is Charlie. He fell asleep on a heating v...	Charlie	2914	16046	kelpie	NaN	1.1
283	827199976799354881	2017-02-02 17:00:17+00:00	iPhone	This is Charlie. He wins every game of chess h...	Charlie	2056	9975	Great_Dane	NaN	1.3
583	771102124360998913	2016-08-31 21:47:27+00:00	iPhone	This is Charlie. He works for @TODAYshow. Supe...	Charlie	1324	5848	Labrador_retriever	NaN	1.2
1040	703382836347330562	2016-02-27 00:55:11+00:00	iPhone	This is Charlie. He's a West Side Niddlewog. M...	Charlie	1026	3216	golden_retriever	NaN	1.2
1111	697255105972801536	2016-02-10 03:05:46+00:00	iPhone	Meet Charlie. He likes to kiss all the big mil...	Charlie	1052	2842	Great_Dane	NaN	1.0
1321	681891461017812993	2015-12-29 17:36:07+00:00	iPhone	Say hello to Charlie. He's scholarly af. Quite...	Charlie	744	2267	Chihuahua	pupper	1.0
1664	671504605491109889	2015-12-01 01:42:28+00:00	iPhone	This is Charlie. He was just informed that dog...	Charlie	3114	6220	toy_poodle	NaN	1.1

The tweet text for tweet with dog's name `Charlie` posted the `2017-02-20` at `23:50:09` with tweet id `833826103416520705` and bread `Chihuahua` is :

'Meet Charlie. She asked u to change the channel to Animal Planet at least 6 times. Now taking matters into her own paws. 13/10 assertive af https://t.co/WTzhtfevKY'

Q6 : What is the most used source of tweets ?

The most used source of tweet is `iPhone` with more than 98%, like shown in the pie chart below:



Conclusion

We analyze and visualize the cleaned dataset of WeRateDogs, and discover that, most (more than 98%) of tweet where made via an iPhone, `rating` and `retweet_count`, then `rating` and `favorite_count` are positively but not highly correlated and many other insights.

The dataset anlyzed here was assessed and cleaned following some considerations which are:

- Considering original ratings (no retweets) that have images.
- Assessing and cleaning at least 8 quality issues and at least 2 tidiness issues in this dataset. Not all qualities and tidiness issues were assessed and cleaned.
- The fact that the rating numerators are greater than the denominators does not need to be cleaned.

- No gathering the tweets beyond August 1st, 2017.
- Droping rows with rating denominator different to 10, and rating numerator bigger than 15.

With differents considerations, the insights might not be the same.