WeRateDogs Project
Udacity Nanodegree Data Analyst

Act Report

1. Introduction:
2. Observations
3. Remarks

1. Introduction:

Within the project there the data of tweet archive of Twitter user @dog_rates, also known as WeRateDogs was wrangled. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

The Act Report is separated in the Observations and Remarks Part, which are are described in the following sections with regard to the WeRateDogs Project.

2. Observations:

Overall the following can be observed:

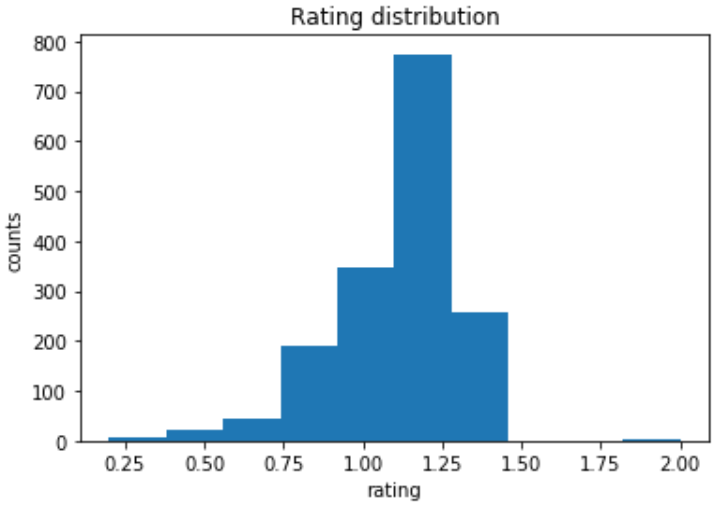
	rating_numerator	rating_denominator	rate_ratio	dog_conf	retweet_count	favorite_count
count	1639.000000	1639.000000	1639.000000	1639.000000	1639.000000	1639.000000
mean	11.466748	10.482611	1.089844	0.561027	2927.587553	8732.450275
std	7.521964	6.504120	0.181762	0.291238	4691.574738	11962.519243
min	1.000000	2.000000	0.200000	0.050512	16.000000	0.000000
25%	10.000000	10.000000	1.000000	0.319108	664.500000	1860.500000
50%	11.000000	10.000000	1.100000	0.565981	1489.000000	4144.000000
75%	12.000000	10.000000	1.200000	0.825674	3463.500000	11107.000000
max	165.000000	150.000000	2.000000	0.999956	56625.000000	132810.000000

Obersavation

- rating numerator is on average above 10 and technically on average out of its limitations
- on average a high amount of retweets can be oberseved (~3000)
- on average a high amount of likes can be oberseved (~8700)

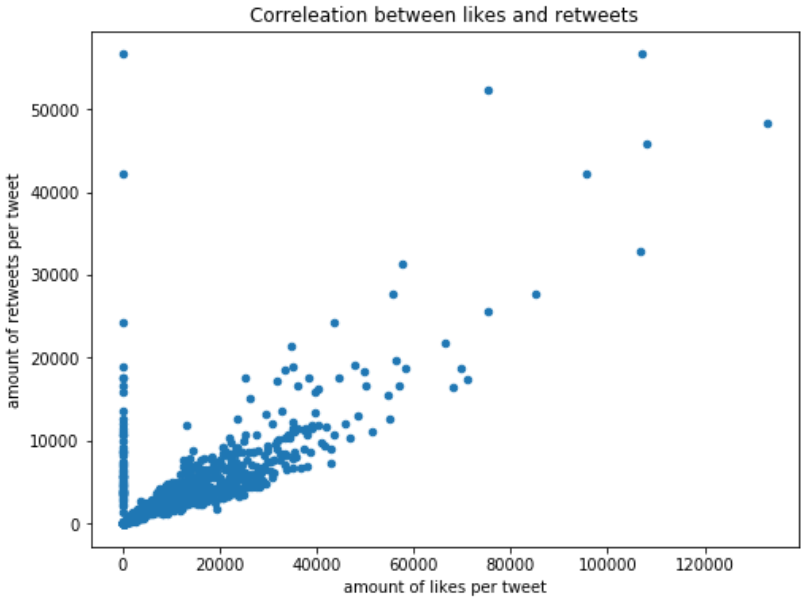
The rate_ratio from the text column **can hardly be used for further analysis**, since its more used for fun and obviously almost everybody who loves his dog and write such tweet, would rate it very high.

This becomes apparent with a look at the rating distribution. This distribution follows somewhat a gaussian bell around the “meme”-rating 13/10.



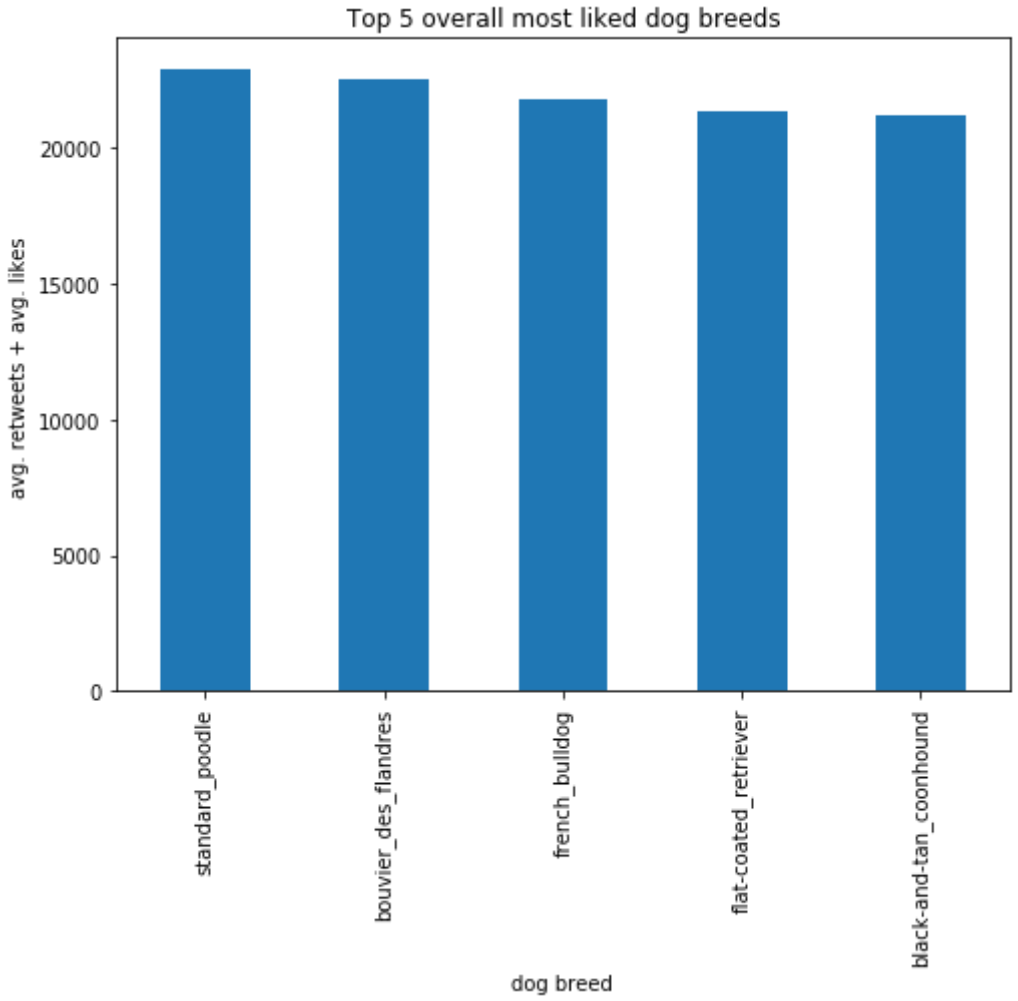
A much better metric for the love of an audiance towards a specific dog breed or dog name should be the **avg. retweets and the avg. likes per tweet**.

To combine these two metrics it makes sense to have a look at the correlation of them.



There is a **visual strong linear correlation** between the amount of avg. likes and avg. retweets per tweet. The correlation coefficient is close enough to one, so that considering **the sum of both metric can be justified**.

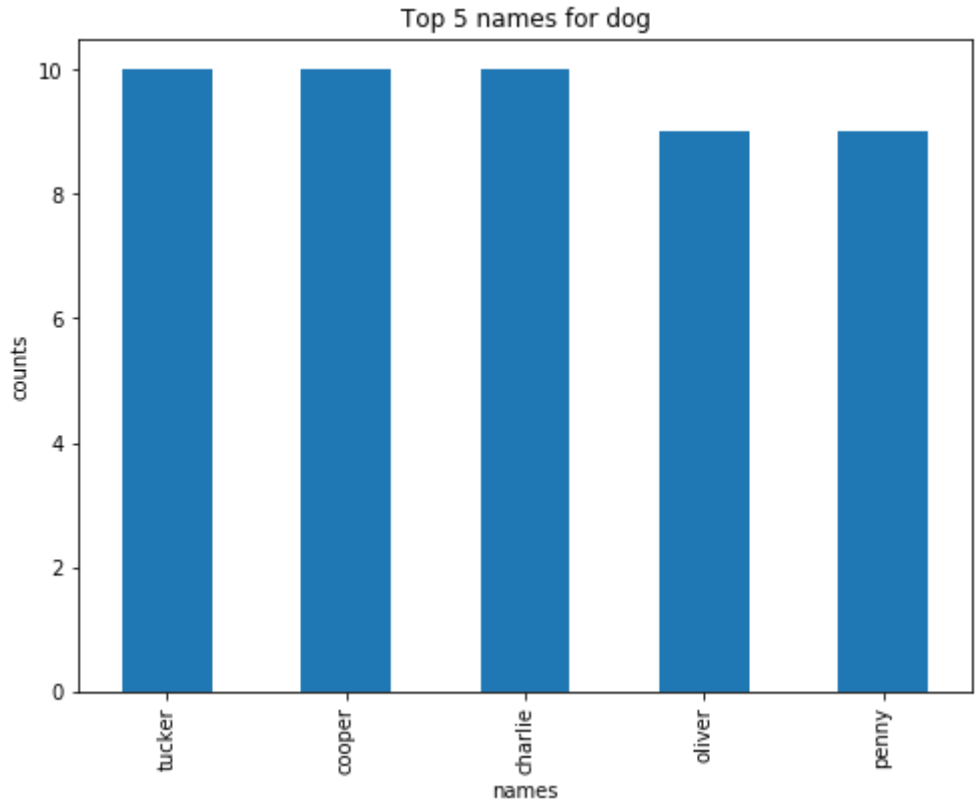
The most liked dog breed



Using these two metrics to determine the most liked dog breed, it can **be concluded that the "standard_poodle" is the most liked dog breed for the given dataset**. (Which is justified since he is adorable)

https://en.wikipedia.org/wiki/Poodle#/media/File:Bo_the_poodle_retrieving_a_duck.jpg

The most used dog names

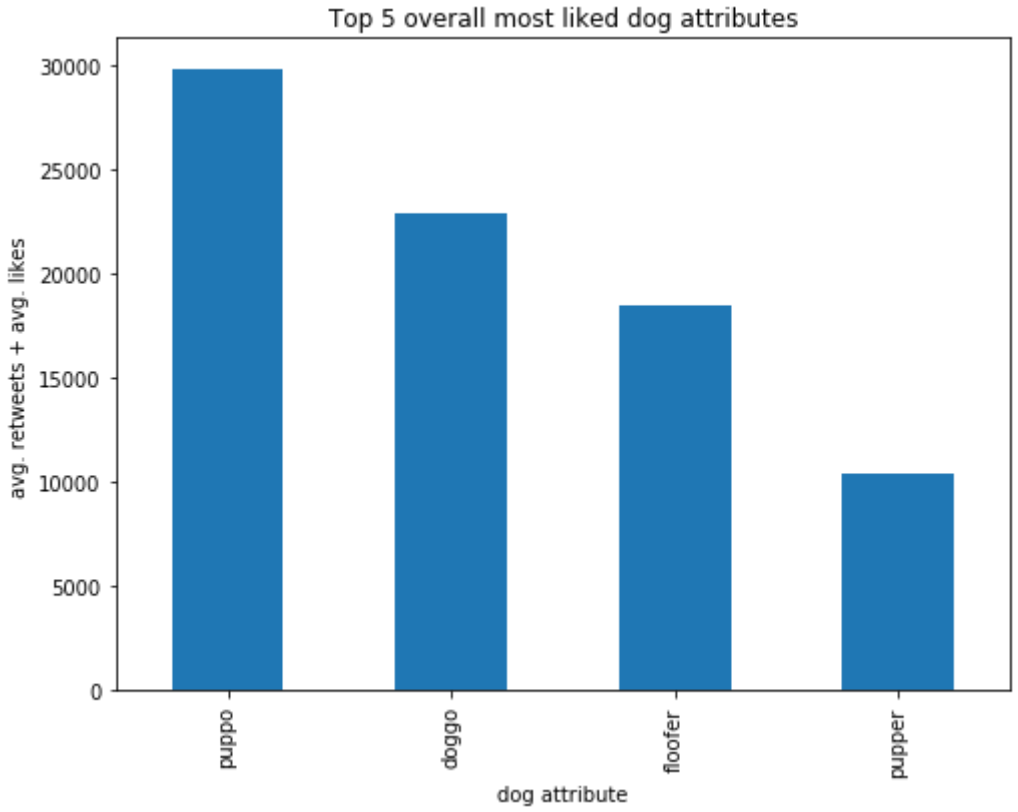


Using only the count metric to determine which names are very popular, it can **be concluded that the there is a tie.** The following three names seem to be in particular popular:

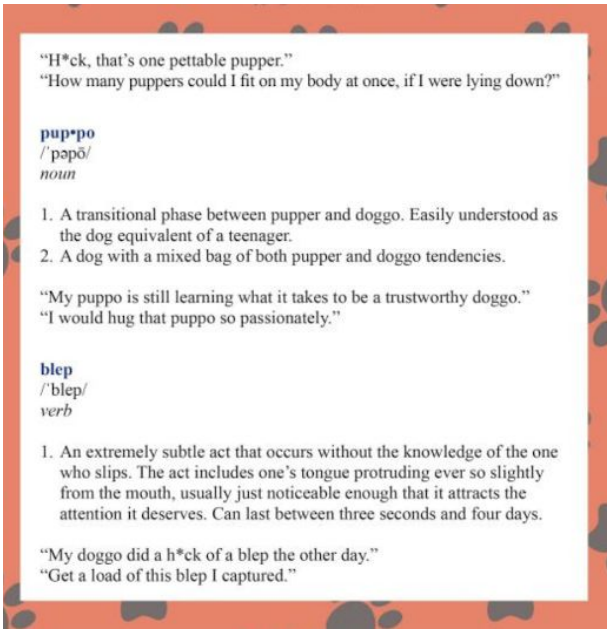
- tucker
- cooper
- charlie

(**or** the owner who name their dogs like this, could be very active on twitter)

The most liked dog attribute.



Using these two metrics to determine the most liked dog breed, it can **be concluded that the "puppo" is the most liked dog attribute for the given dataset**.
This dog attribute is explained here:


"H*ck, that's one pettable pupper."
"How many puppers could I fit on my body at once, if I were lying down?"

pup•po
/ˈpəpō/
noun

1. A transitional phase between pupper and doggo. Easily understood as the dog equivalent of a teenager.
2. A dog with a mixed bag of both pupper and doggo tendencies.

"My puppo is still learning what it takes to be a trustworthy doggo."
"I would hug that puppo so passionately."

blep
/ˈblep/
verb

1. An extremely subtle act that occurs without the knowledge of the one who slips. The act includes one's tongue protruding ever so slightly from the mouth, usually just noticeable enough that it attracts the attention it deserves. Can last between three seconds and four days.

"My doggo did a h*ck of a blep the other day."
"Get a load of this blep I captured."

3. Remarks:

Two things are important to mention:

I. When it comes to the most liked dog breed, technically the winner would be the **Saluki.** However it was decided to exclude this breed since it only 4 entries were found in the master dataset, which is not a sufficient sample size.

```
# check samplesize for winner
len(df.query('dog_breed == "saluki"'))
```

```
4
```

- remove outlier due to small samplesize

II. It could make sense to perform a linear regression analysis on the different dog attributes, since the dataframe was already structured in the appropriate columns. This would however require a bigger sample size, since eg. the “floofer” was mentioned less than 25 times in the master dataframe.