Insights and Visualizations Report

Data Analyst Nanondegree, Project 4: Wrangle and Analyze Data

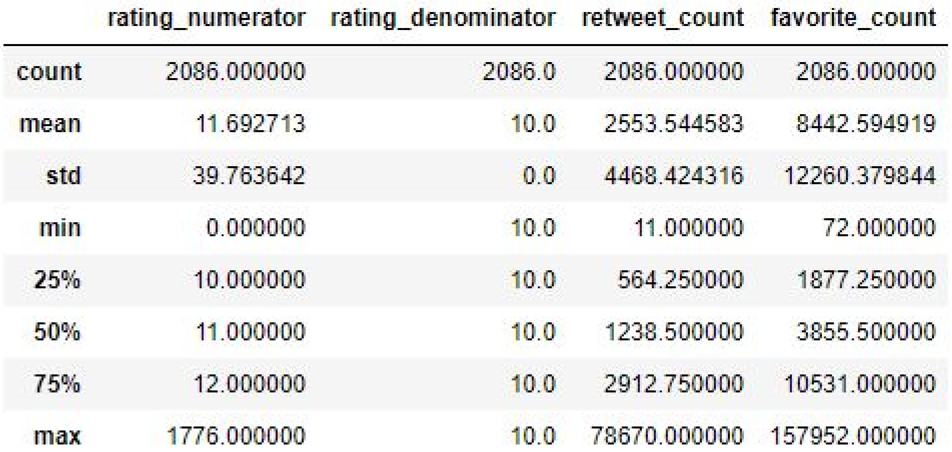
After cleaning the datasets from WeRateDogs, I made some observations and derived some insights with useful visualizations, which I will present here.

Tools Used:

- Python programming language Jupyter Notebook (IPython)
- Pandas, Seaborn, matplotlib

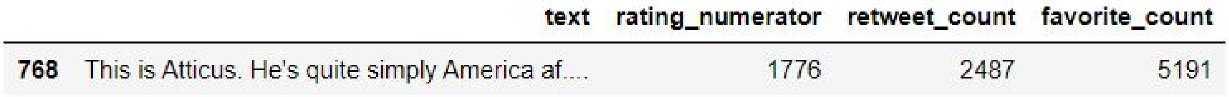
Quick general numerical analysis:

I like to start with simple stats, by calling `.describe` function on the cleaned dataset:



	rating_numerator	rating_denominator	retweet_count	favorite_count
count	2086.000000	2086.0	2086.000000	2086.000000
mean	11.692713	10.0	2553.544583	8442.594919
std	39.763642	0.0	4468.424316	12260.379844
min	0.000000	10.0	11.000000	72.000000
25%	10.000000	10.0	564.250000	1877.250000
50%	11.000000	10.0	1238.500000	3855.500000
75%	12.000000	10.0	2912.750000	10531.000000
max	1776.000000	10.0	78670.000000	157952.000000

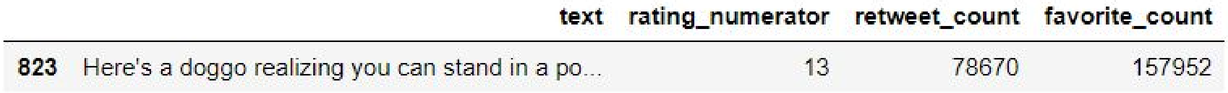
There are several unique features we can infer.
First, have a look at the dog with the highest rating:



	text	rating_numerator	retweet_count	favorite_count
768	This is Atticus. He's quite simply America af....	1776	2487	5191

While it has the highest rating, it doesn't have the highest retweet nor favorite count, not even close!

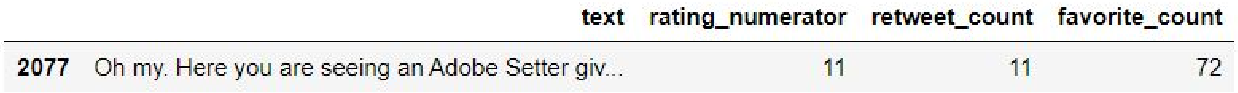
Let's see the dog with the highest retweet_count:



	text	rating_numerator	retweet_count	favorite_count
823	Here's a doggo realizing you can stand in a po...	13	78670	157952

Well, while not a bad rating, it's nowhere close to the one before.

Here's the dog with the lowest favorite and retweet count:

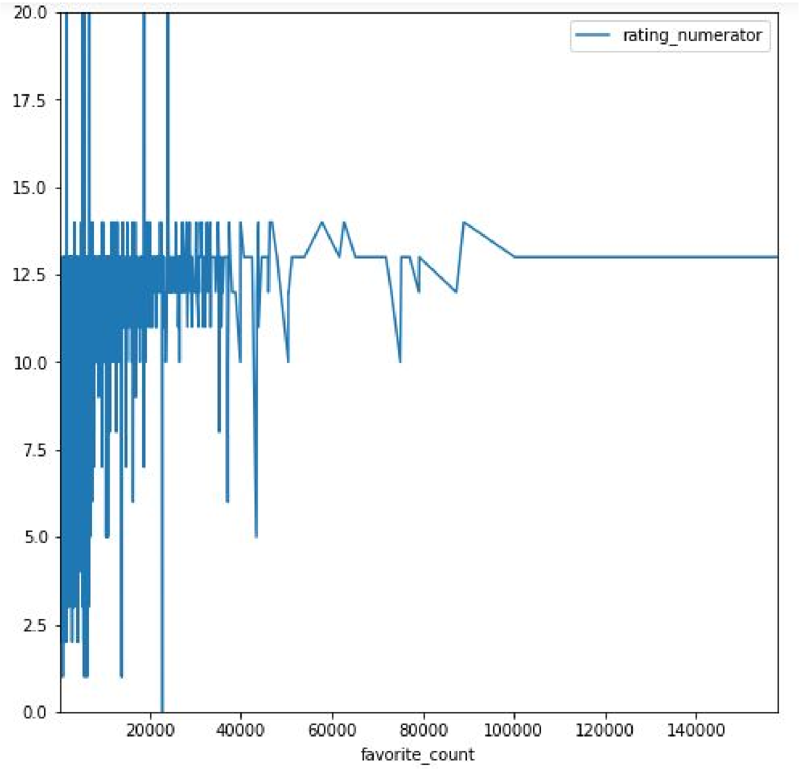


	text	rating_numerator	retweet_count	favorite_count
2077	Oh my. Here you are seeing an Adobe Setter giv...	11	11	72

While retweet and favorite count are both much, much lower, the dog still has a close rating to the most favorite dog.

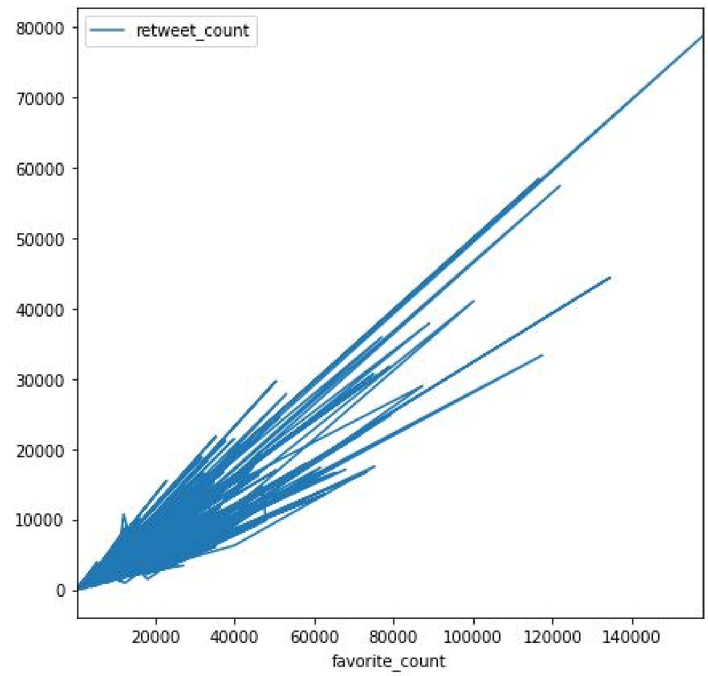
Some Visualizations:

What can we infer? That there isn’t a strong relation between retweet or favorite count and the given rating.
Don’t believe me? Have a look:



There’s only a weak positive relation for small counts, However, as the count gets large, the rating settles around 12 it seems.
I got a similar one for retweet_count as well.

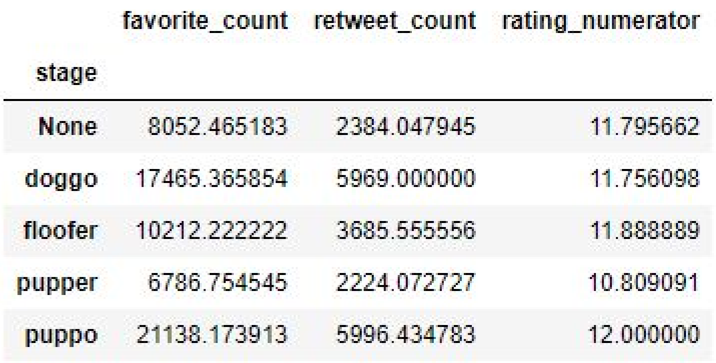
But there’s something we can infer for sure: The higher the retweet_count, the higher the favorite_count as well.



Fascinating

Statistics per dog stage:

For each dog, we have a stage column; it could be a **Doggo**, **Floofer**, **Pupper** or **Puppo** (or None).
Those are like maturity stages, but for dogs.
Anyway, let's see how they compare:



	favorite_count	retweet_count	rating_numerator
stage			
None	8052.465183	2384.047945	11.795662
doggo	17465.365854	5969.000000	11.756098
floofer	10212.222222	3685.555556	11.888889
pupper	6786.754545	2224.072727	10.809091
puppo	21138.173913	5996.434783	12.000000

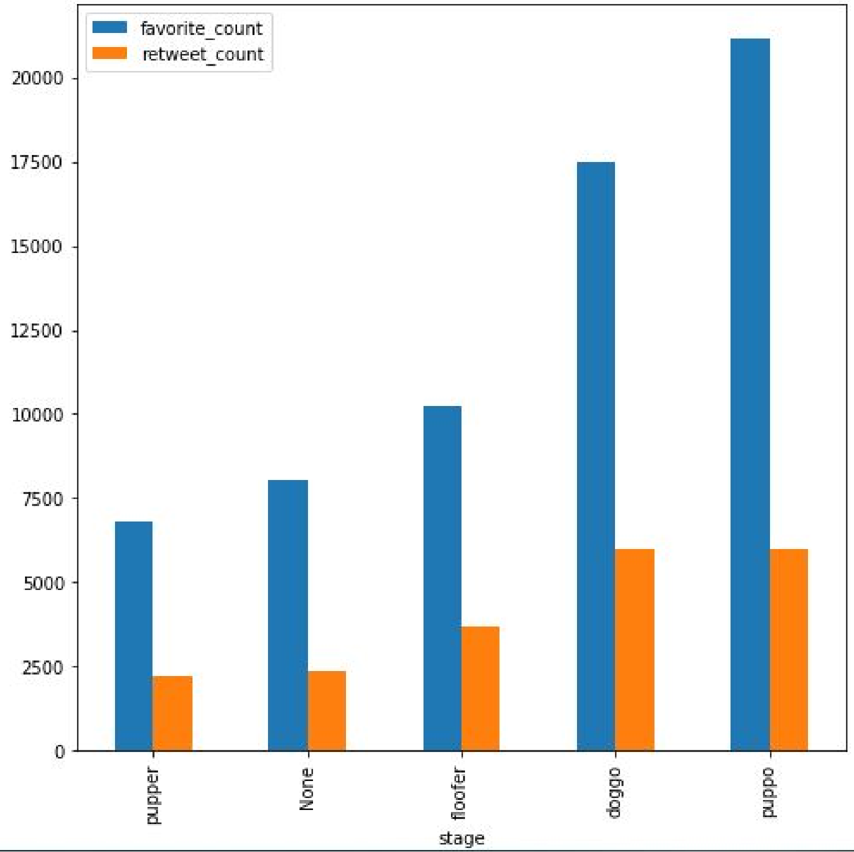
So most people actually prefer dogs at **Puppo** stage, this is evident in the highest *favorite_count*, at 20k, and the highest *rating_numerator* mean, as well as the highest *retweet_count*.

This is followed by **Doggo** then **Floofer** then, surprisingly, **None**, which indicates that *having a stage assigned for a dog isn't necessary for people to like it.*

One must note here that the proportional difference in favorite_count isn't very similar to the rating between stages.

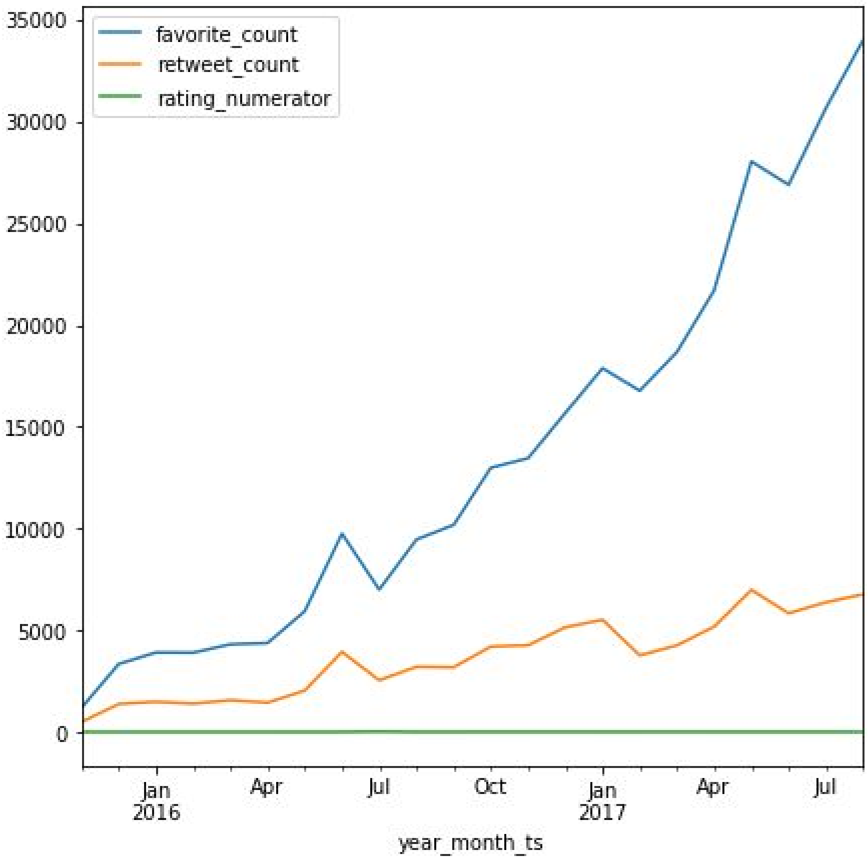
In other words, while dogs without a stage have less *favorite_count* mean than **Duggos**, they still have a *higher average rating*.

All in all, the most reliable measure is the *favorite_count* followed by the *retweet_count*.
To see this more clearly, here's a bar chart:

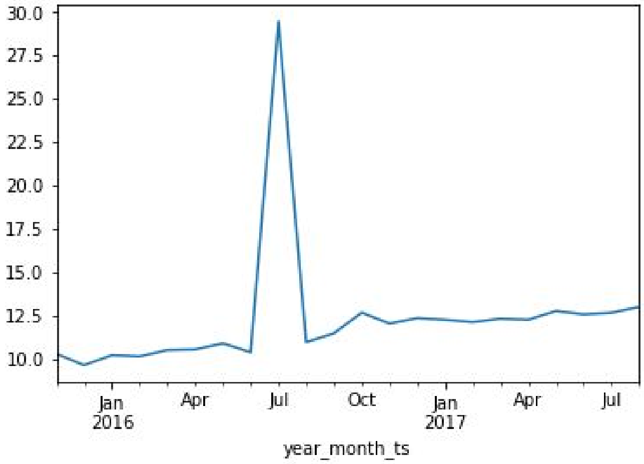


Statistics over time:

I also examined those statistics as time goes by, from the first tweet in 2015 till the last one in august 2017 (as recorded).
Here's an interesting line plot. The mean of each statistic of each month:



This shows that WeRateDogs is certainly becoming more famous, and *not* that dogs are becoming nicer.
Because *rating_numerator* mean almost stays constant, while *favorite_count* is increasing. Which makes sense.
One last thing, that's the line plot for rating alone:



That's all the insights and visualizations I have produced :)