**Act Report**

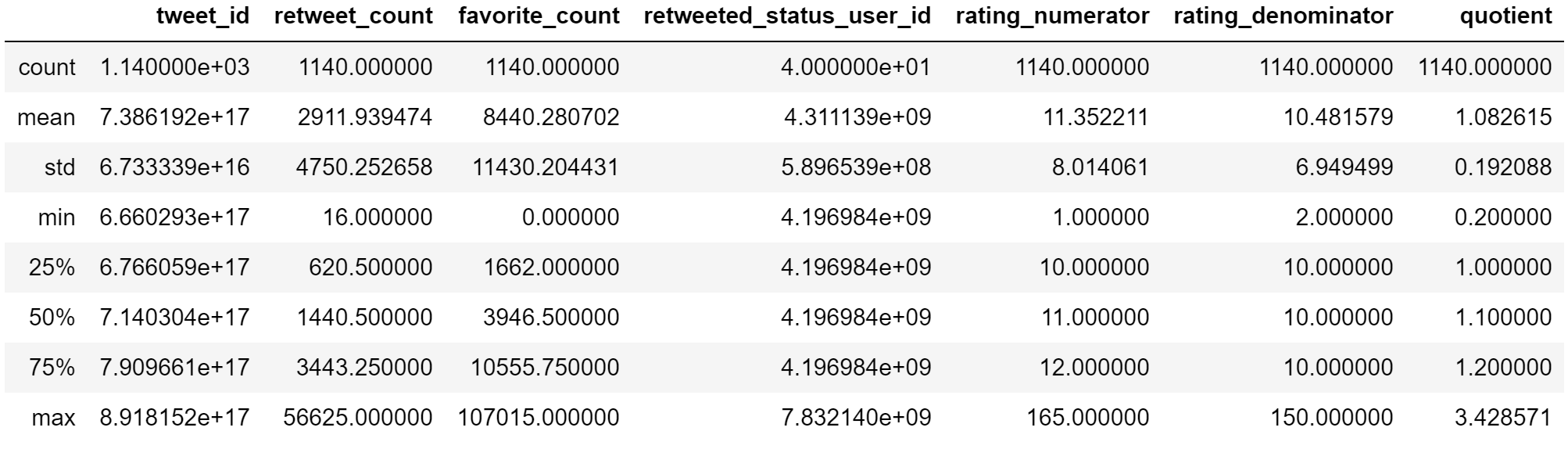
**Intro:**

In this project I gathered, assessed and analyzed sets of data from Udacity and Twitter user [@dog\_rates](https://twitter.com/dog_rates" \t "https://classroom.udacity.com/nanodegrees/nd002/parts/af503f34-9646-4795-a916-190ebc82cb4a/modules/14d9f5f1-9e7b-4bfb-97f3-bcdbf4a3699c/lessons/a8085857-3e28-4fc7-aeb8-da64ccbc2e20/concepts/_blank), also known as [WeRateDogs](https://en.wikipedia.org/wiki/WeRateDogs" \t "https://classroom.udacity.com/nanodegrees/nd002/parts/af503f34-9646-4795-a916-190ebc82cb4a/modules/14d9f5f1-9e7b-4bfb-97f3-bcdbf4a3699c/lessons/a8085857-3e28-4fc7-aeb8-da64ccbc2e20/concepts/_blank). WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.  Software used include: Python(pandas, NumPy, requests, tweepy, json), Google Docs.

Files provided ahead include: ‘twitter-archive-enhanced-2.csv’, image-predictions.tsv, tweet-json.txt. The archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017. ‘image-predictions’ is a file contains prediction result of whether the object in the picture is a dog or not as well as dog breed. It was powered by a [neural network](https://www.youtube.com/watch?v=2-Ol7ZB0MmU" \t "https://classroom.udacity.com/nanodegrees/nd002/parts/af503f34-9646-4795-a916-190ebc82cb4a/modules/14d9f5f1-9e7b-4bfb-97f3-bcdbf4a3699c/lessons/a8085857-3e28-4fc7-aeb8-da64ccbc2e20/concepts/_blank) that can classify breeds of dogs and ran through every image in the WeRateDogs Twitter. ‘tweet-json’ was provided as additional file that contains all information in every tweet.

**Insight:**

*We can have an overall insight about statistical results in the following graph. On average, those tweets was retweeted 2912 times, was 'liked' 8440 times, the average quotient, although giving the fact 'those are good dogs' and expect quotients would be way above 1, there are few were rated under 1, make up the average quotient round 1. This also make sense consider 50% quotient is 1.1 or less.*

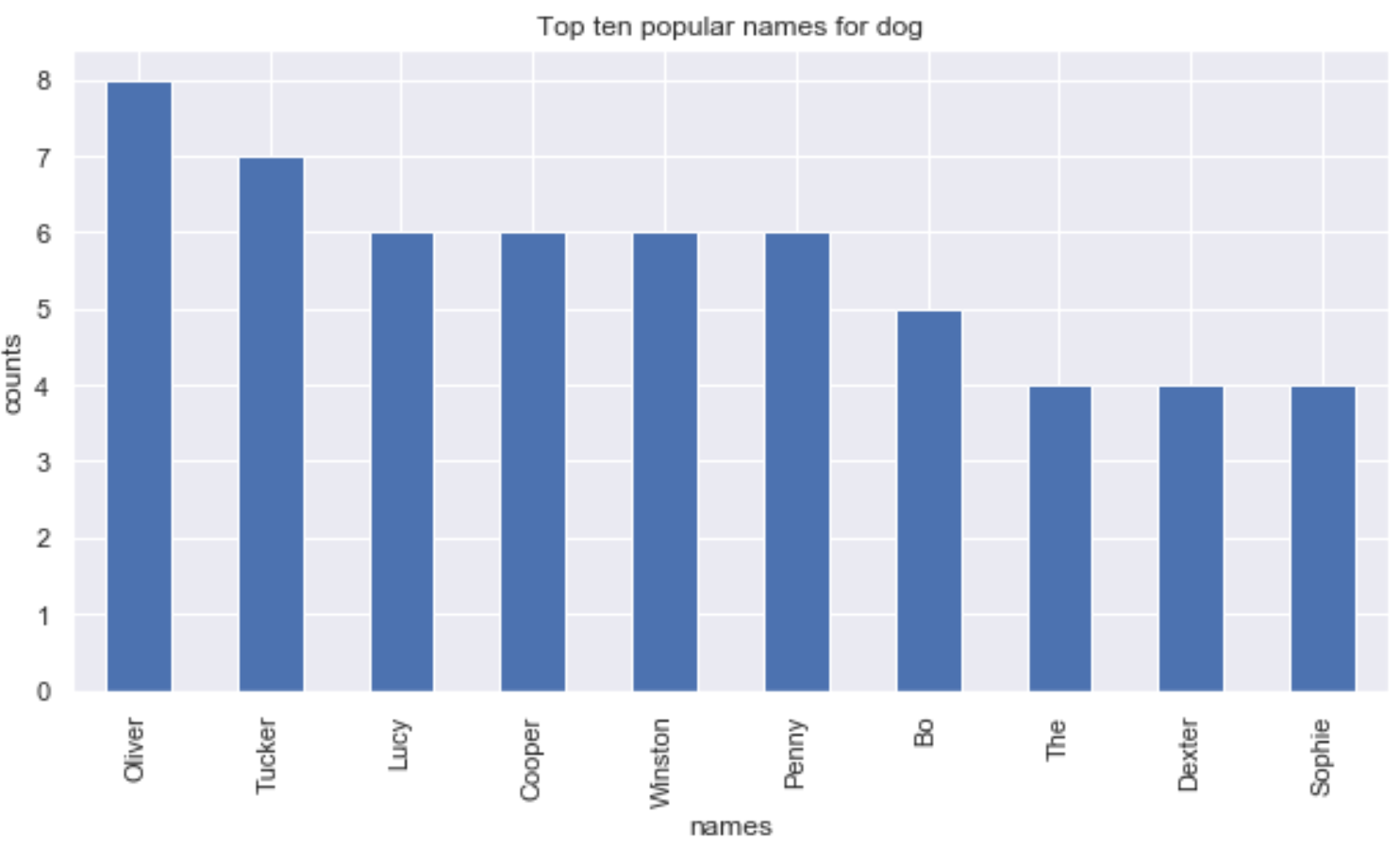
**

I made a quotient column during wrangling process and attempt to use as one of key metrics, but as I dived deeper, I found out data quality issues:

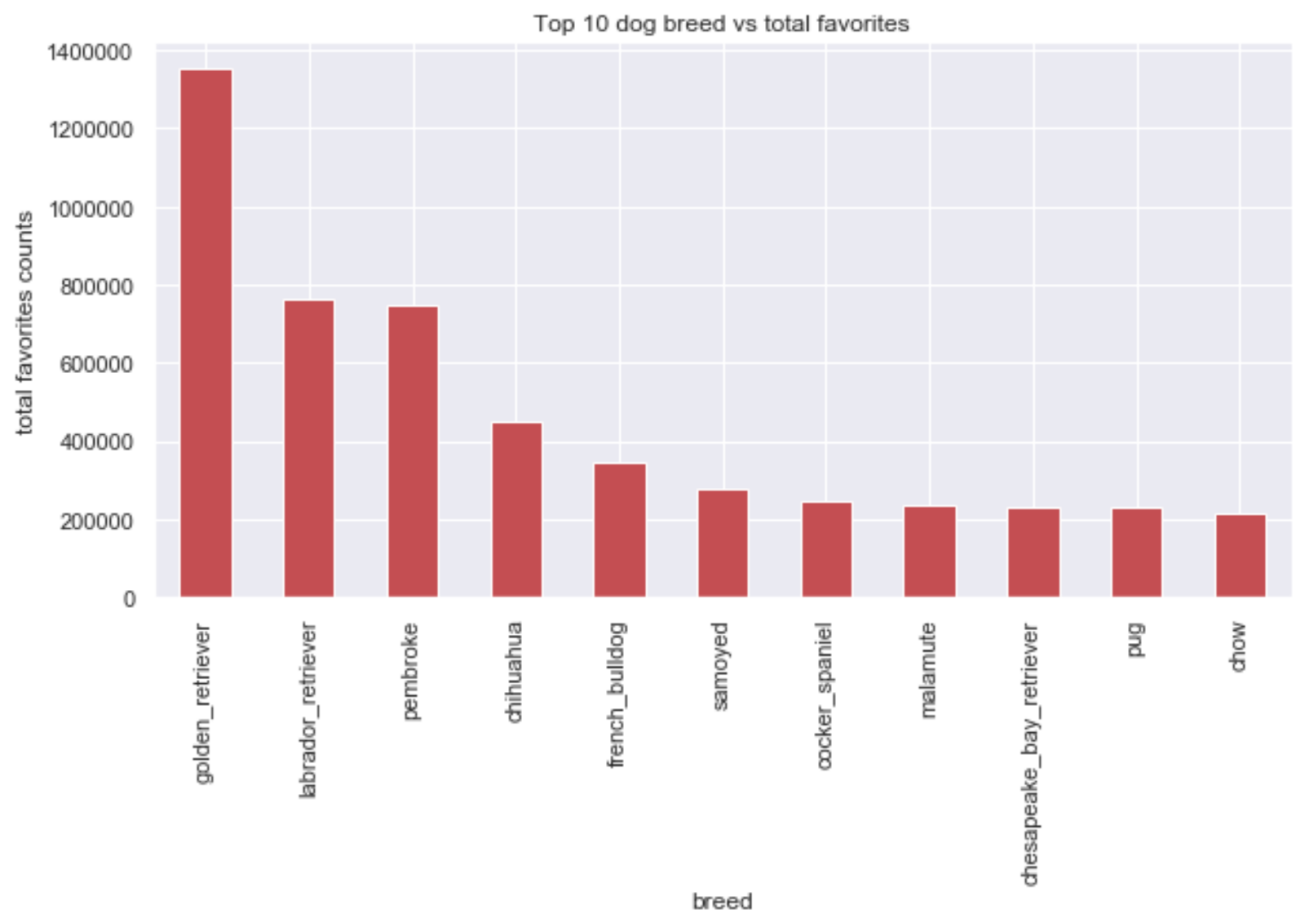
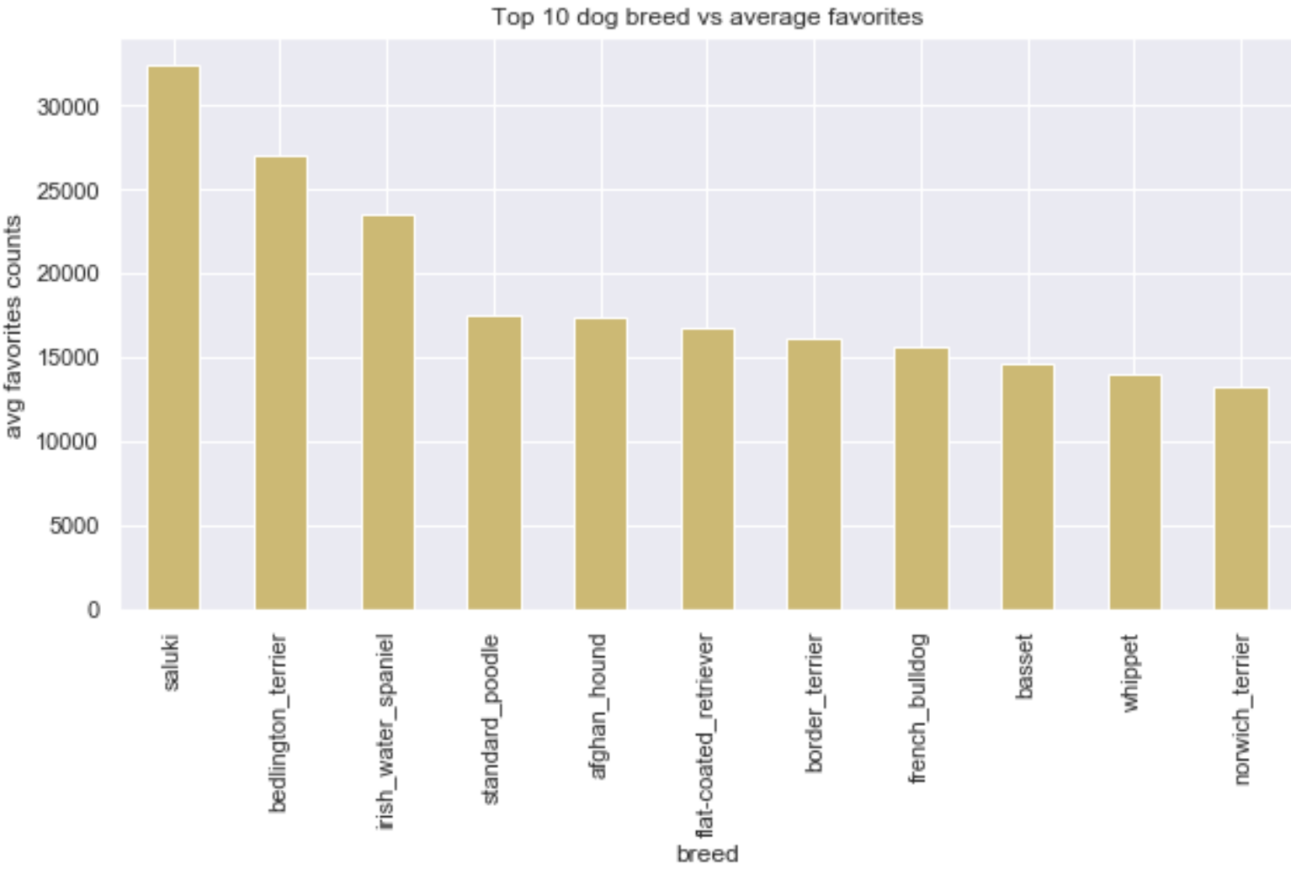


Consider tweet id 722974582966214656 as an example. The text is :”*Happy 4/20 from the squad! 13/10 for all [https://t.co/eV1diwds8a.](https://t.co/eV1diwds8a’.)”* the correct extraction would be 13/10, but regex will capture both 4/20 and 13/10. We can choose to extract the second group manually, but in other text the correct match might be the first group, there is no good way to extract match correctly except going through one by one manually, which is not practical. Therefor, I decided not to take quotient in as a metric.

Some interesting insight about dog names:

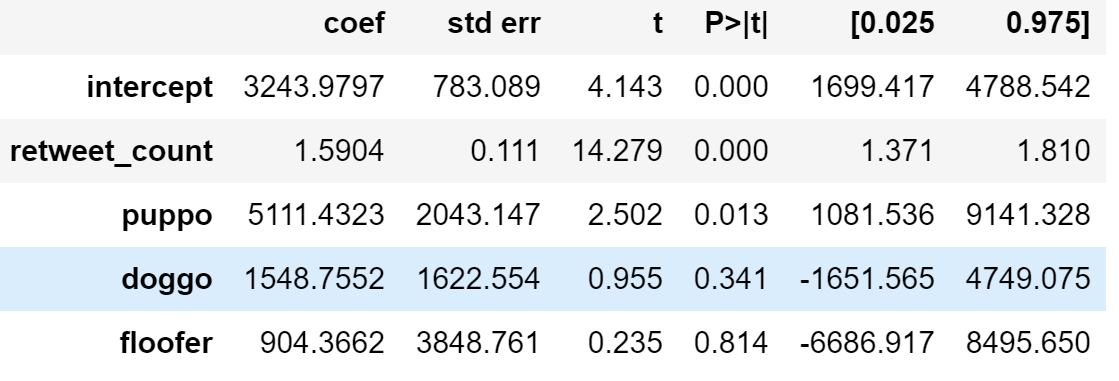


Here we can see the difference of 'top 10 most favorite dog breeds' when we use different metrics. On average, every saluki dog post earned 32444 'favorites' as the most popular dog breed, while golden retriever seems to be the most popular dog breed when we compare the total ‘favorites’ earned over time.



Those two graphics leads me to a hypothesis: there is a potential correlationship between retweet\_count, favorite\_count and possibly different dog stages as well. I decided to build a linear regression model to test out the hypothesis. There were some more wangling needed, steps were:

1. Make mask for records with valid dog stage.
2. Investigate ones with multiple stages.
3. Split records with two dog stage to two dataframes, each with one dog stage
4. Append two dataframes together.
5. Drop records with multiple stage from dfstage then append df1 to dfstage.
6. Set ‘intercept’= 1 and choose a ‘baseline’: ‘pupper’.
7. Fit model and interpret result.



From the model we can tell retweet\_count and favorite\_count has a very strong positive correlationship. Every additional retweet is likely to result more than 1 'favorite'. With everything else being equal, we expect dogs at 'puppo' stage receive 5111 more 'favorites' than dogs at 'pupper' stage(baseline), 'doggo' receive 1549 more 'favorites' than 'pupper', 'floofer' receive 904 more 'favorites' than 'pupper'

Limitation: sample size is not large (197 counts), leads to inaccurate prediction. On the other hand, p-value and scatter plot indicate different dog stage don't have strong correlation with number of 'favorite'.