

# act\_report

February 21, 2019

## 0.1 Act\_Report

Starting with around five thousand rows of data and after the cleaning process ending with around three hundred is a significant drop off. This is sometimes the case though when there is a lot of missing data or very unclear data (where we lost approximately three thousand rows) and data needs to be merged together that is incomplete from one set to another. With the end result being a completely filled dataframe in mind, dropping so many observations was the only way to accomplish this.

[1] Looking at any numerical features in relation to one another, it's obvious there isn't a whole lot of correlation going on between much of the data. Several features offered nothing in regard to insight, like 'breed\_confidence', where almost every relationship it has was a non-descript distribution. The same can be said for the most part for 'tweet\_image\_number' and 'normalized\_rating'.

[2] The two features with the greatest correlation, which is no surprise, is 'favorite\_count' and 'retweet\_count'. It just makes sense that these two go hand in hand as we can tell from this scatter plot with a regression line. Personally I like this graph from Seaborn because besides showing the data points, it plots a regression line and shows the distribution of both features against each axis.

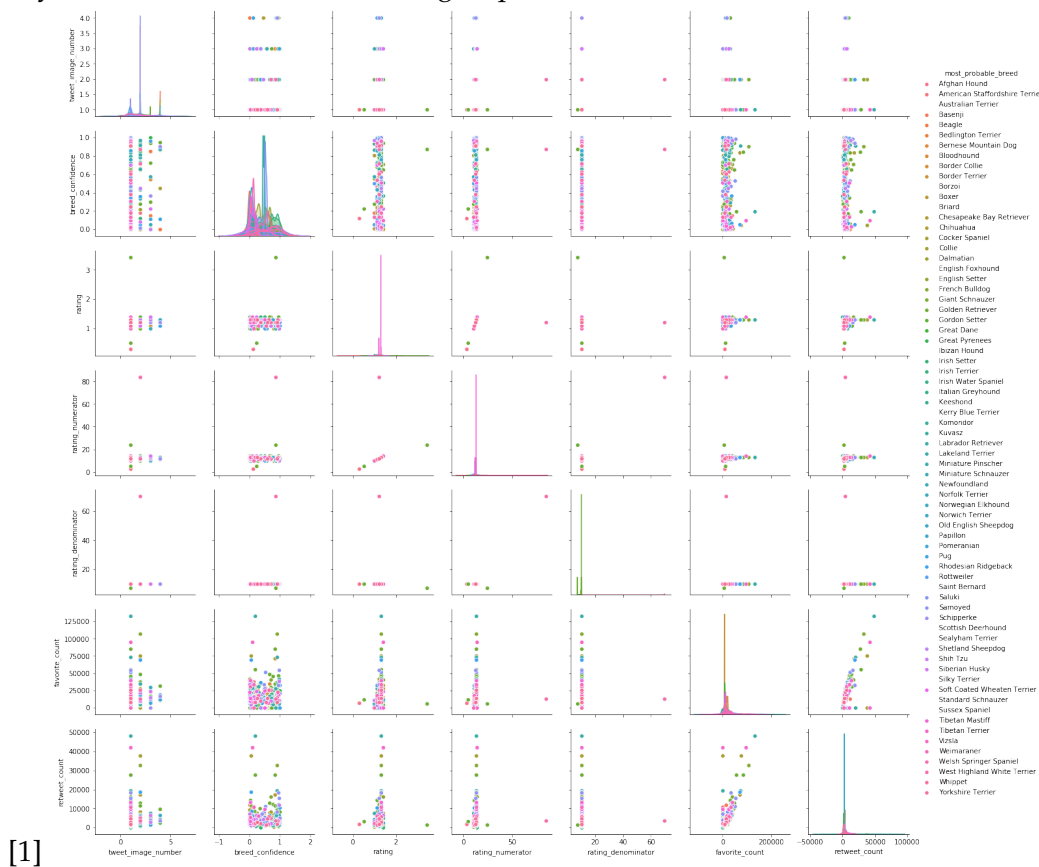
[3] Another (slightly) positive relationship which makes sense is the relationship between 'breed\_confidence' and 'tweet\_image\_number'. I would image that the more certain the algorithm is in making a prediction, the more likely it would be that it would come to this conclusion on the first image. This could be due to several factors; maybe the first image was a higher quality photo, maybe it was a clearer photo, maybe that particular breed is more easily identified. No matter the reason tho it does make sense to see this.

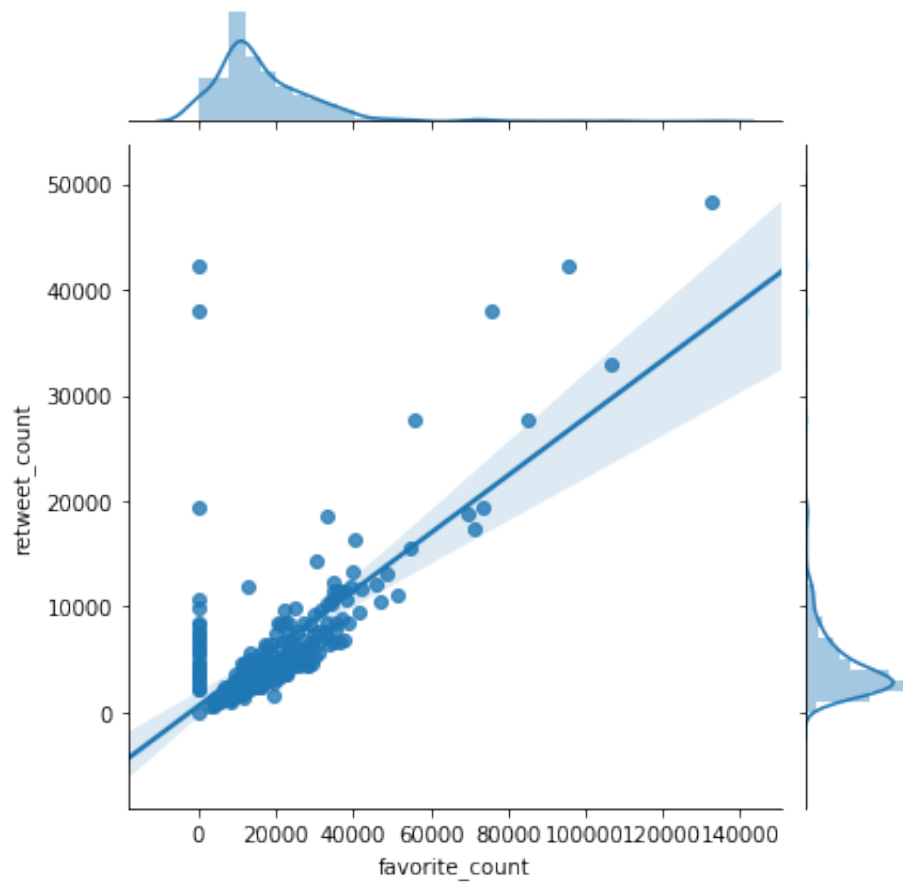
[4 & 5] The next two graphs show some more positive correlations, which again if we think about it is not surprising. The higher the 'normalized\_rating' gets, the more retweets or favorites the dog got. This makes sense, and likely would have been even clearer should the data have been a little better (their rating system not so strange and the so many observations needing to be dropped). Note: some additional observations were dropped for these graphs to exaggerate the regression line. This would have been apparent in the data itself but some outliers prevented the correct distribution.

[6] Another thing I looked at was the distribution of breeds being selected during photos 1-4 of the tweets. While the majority of high confidence levels came during the first photo, no particular breed was more prone to being selected later compared to the other breeds.

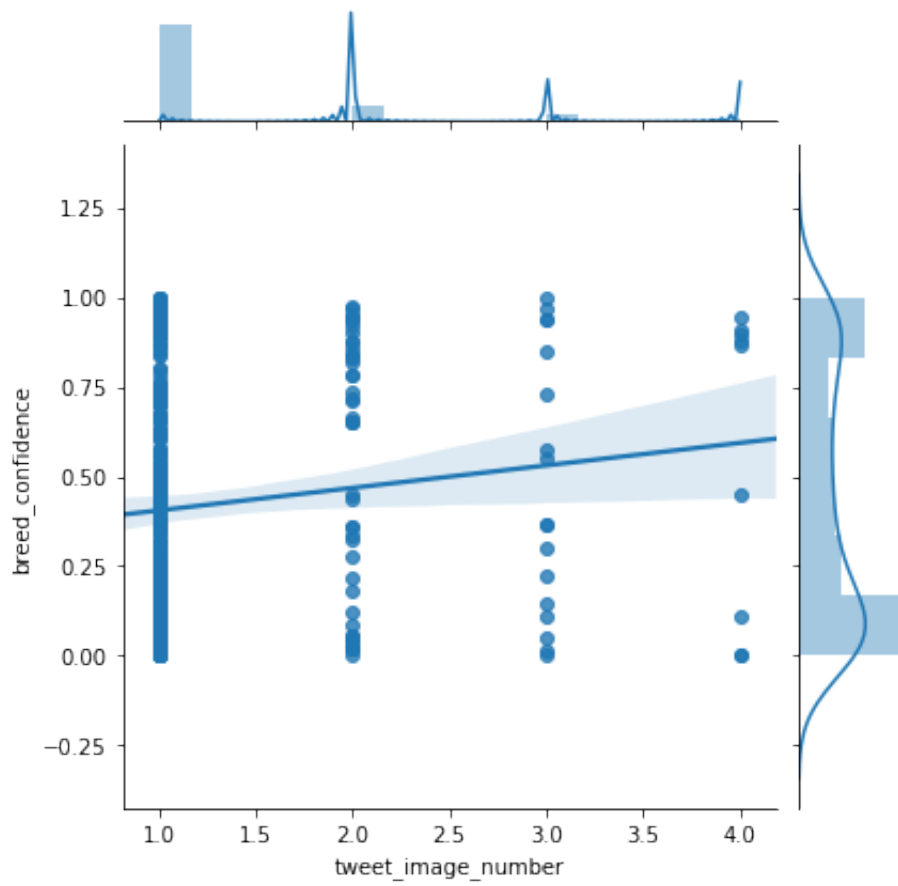
[7] With the limited time we had to explore these data sets and a relatively large amount of data being lost for various reasons I thought the insights that could be drawn from this activity were predictable. I wasn't particularly surprised by anything, although portions of the activity were good (worthwhile) like web scraping to aggregate the data. The last thing we'll look at is the

popularity of an individual breed. Not surprising, America's favorite dog the Golden Retriever was by far the most retweeted in the group.

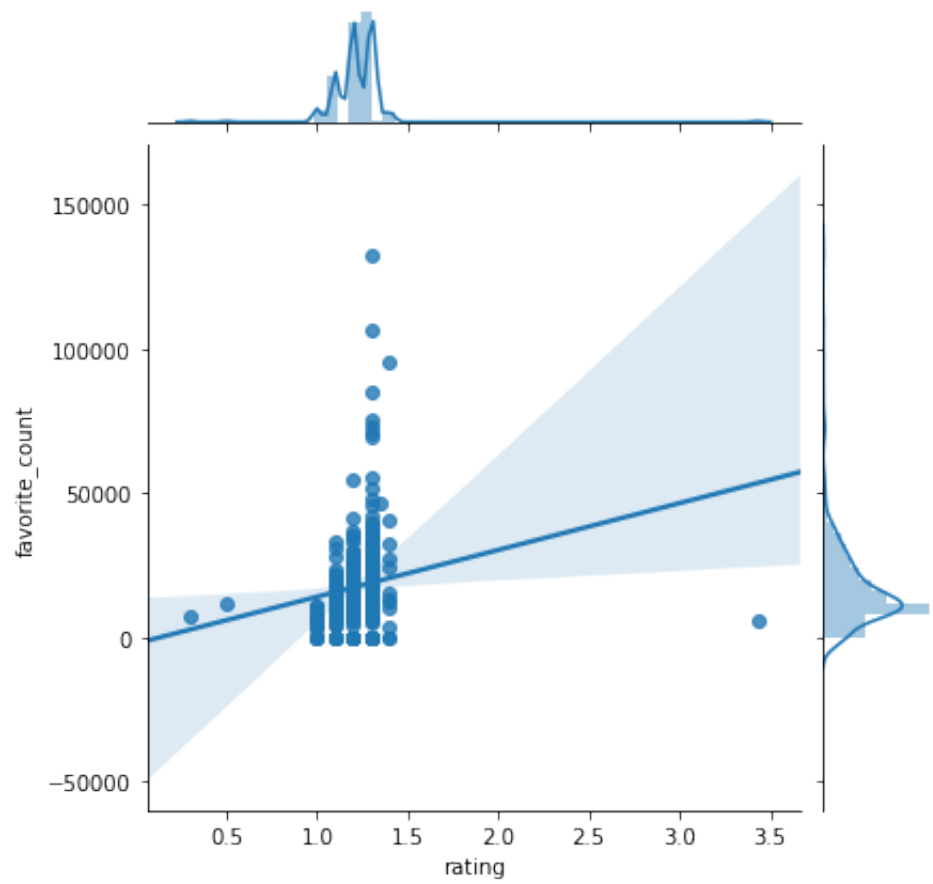




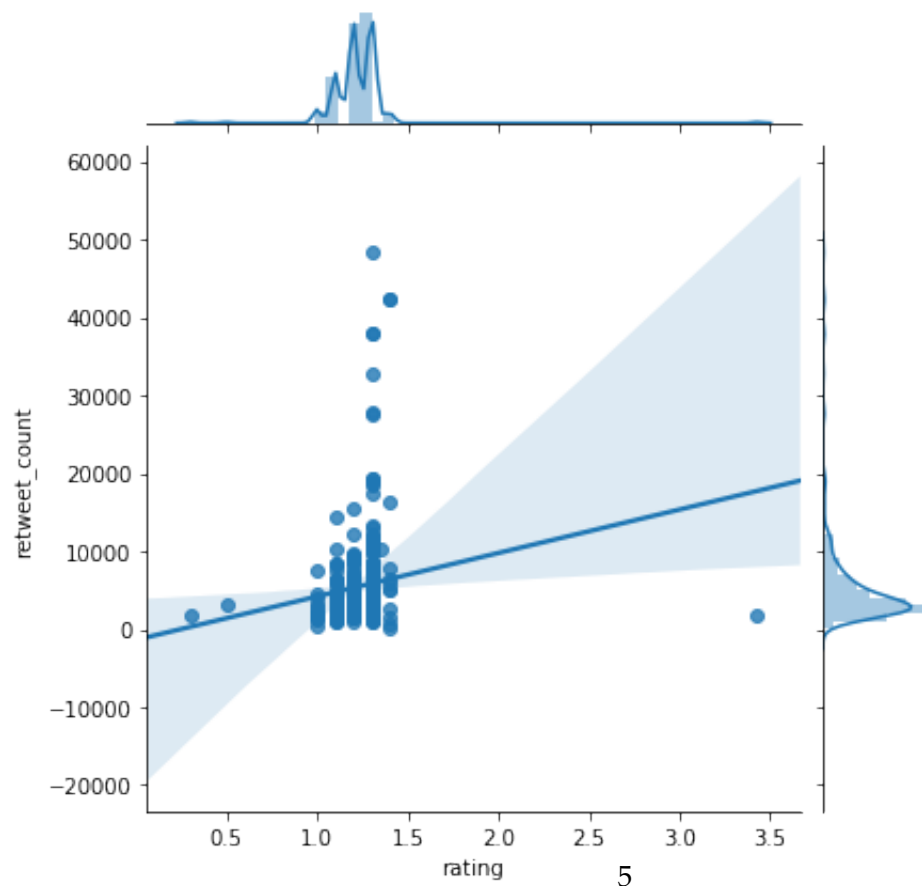
[2]

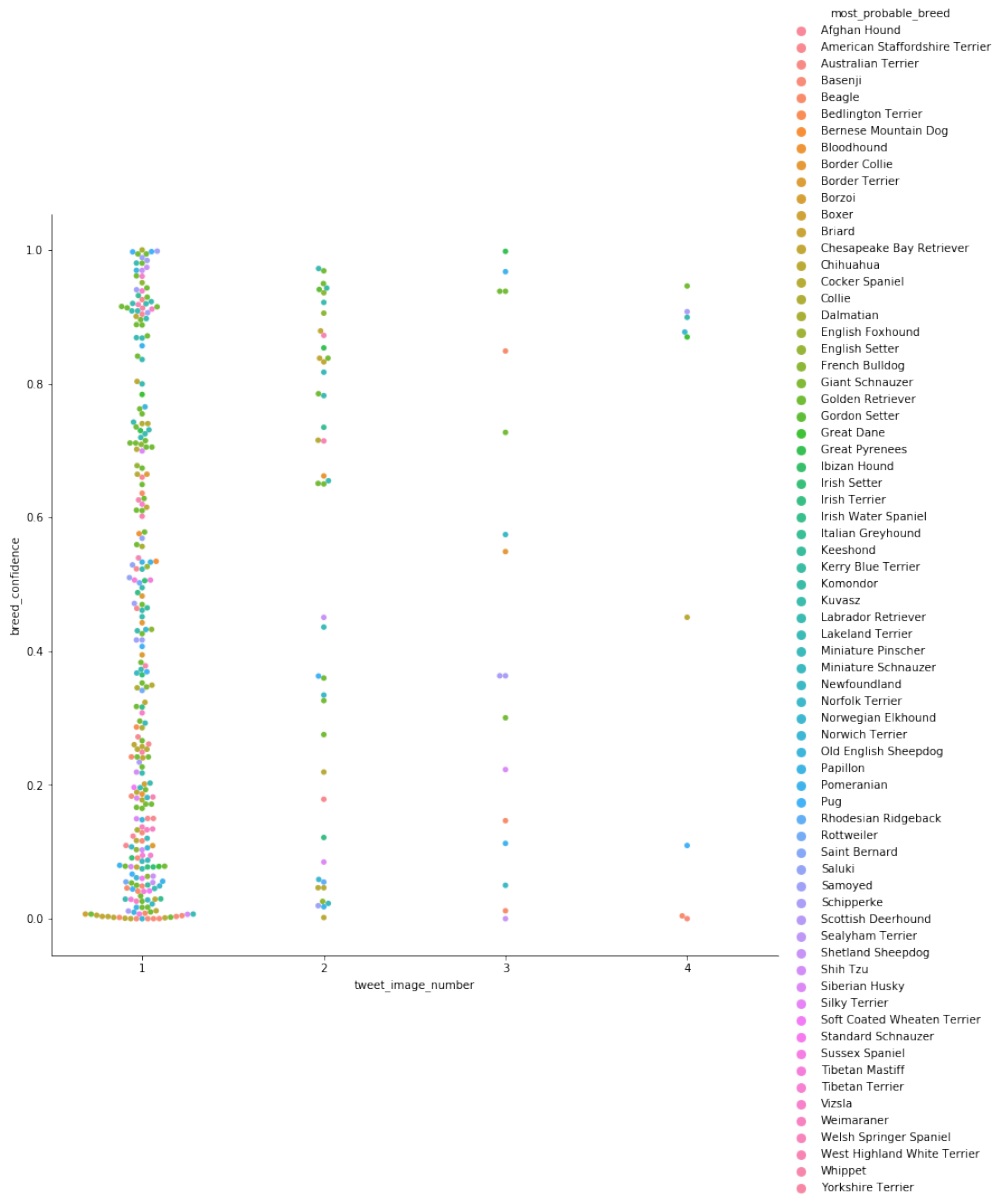


[3]

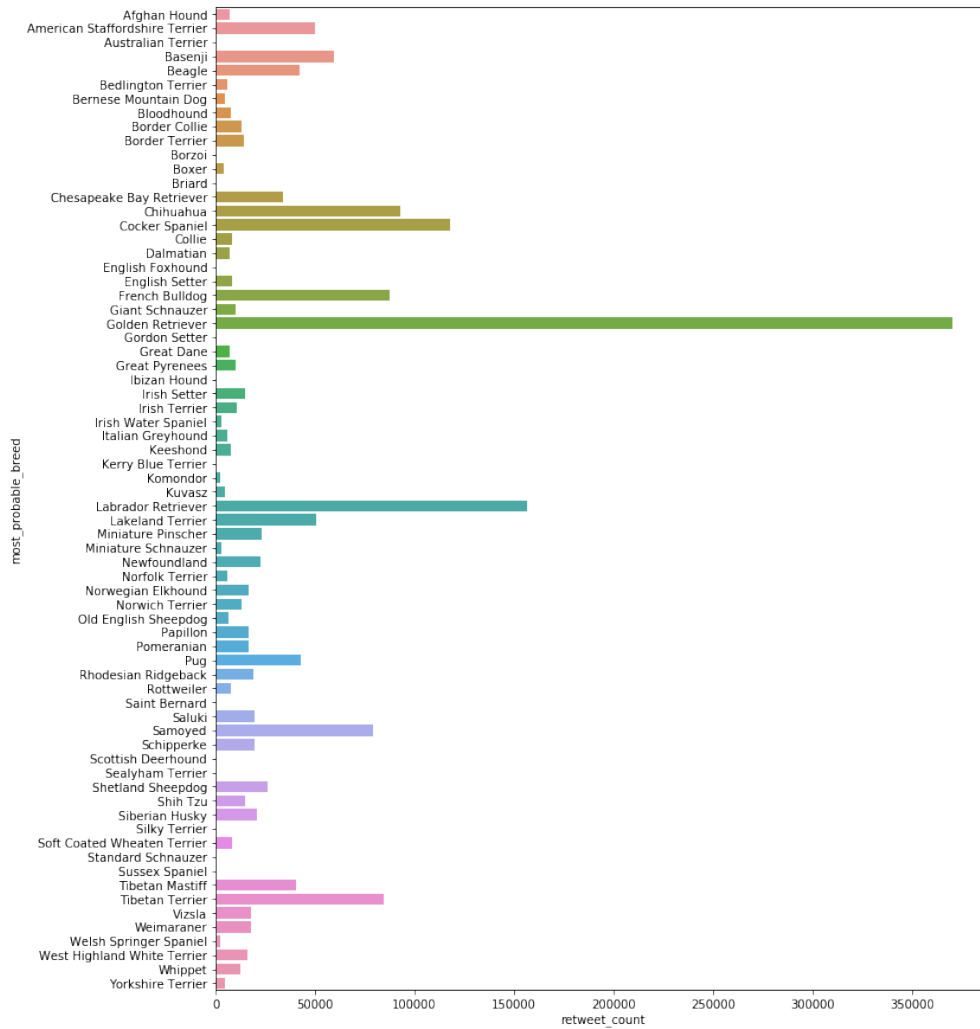


[4 & 5]





[6]



[7]