act_report

February 14, 2019

1 Act Report

1.1 Analysis on the Wrangling Project

I have analysed tweets from **@WeRateDogs** and would like to share my findings with you.

After gathering, assessing and cleaning the information on tweets up to 1st of August 2017, I have started an exploratory analysis to see where that would take me. I thought a historical trend of the retweet and favorite counts could start giving some background:

Apart from a clear peak during the summer of 2016, it looks like the trend is increasing, yet very volatile. As it also seems that, generally, a high favorite count corresponds to a high retweet count, I have plotted a scatter plot to look at the correlation between the two variables:

Favorite and retweet counts seem to be directly proportional. Hence, on average, a post that is liked is also retweeted, which makes complete sense. But, for the specific case of @WeRateDogs, what gives the post a higher chance of being liked or retweeted? I have decided to dive into the new dog jargon and see whether there is a relationship between the just mentioned variables and the dog type.

The tweets I looked at show the following four dog types: - Doggo: mature pupper - Pupper: younger, more immature and more inexperienced than doggo - Puppo: dog equivalent of a teenager - Floofer: any dog or dog with excessive fur

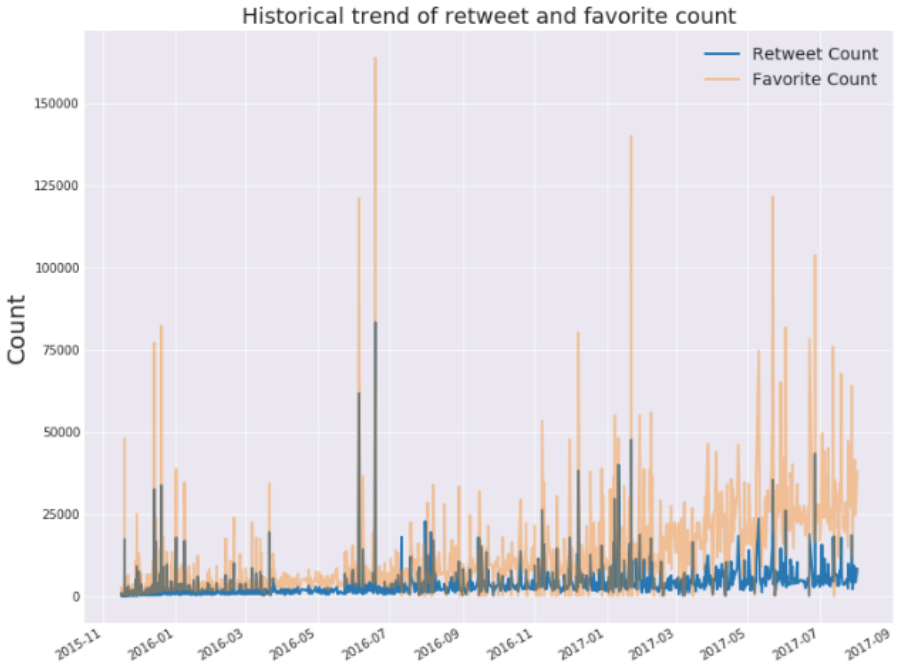
Despite the fact that most of the tweets do not have this specification and we therefore cannot draw confident conclusions, I was curious to see whether we could get some more insights regarding this.

It looks like doggo and puppo have a better chance to get a high retweet count and favorite count. However, as a reminder we have quite little data on the dog type. Hence, we can't draw any conclusions with high confidence.

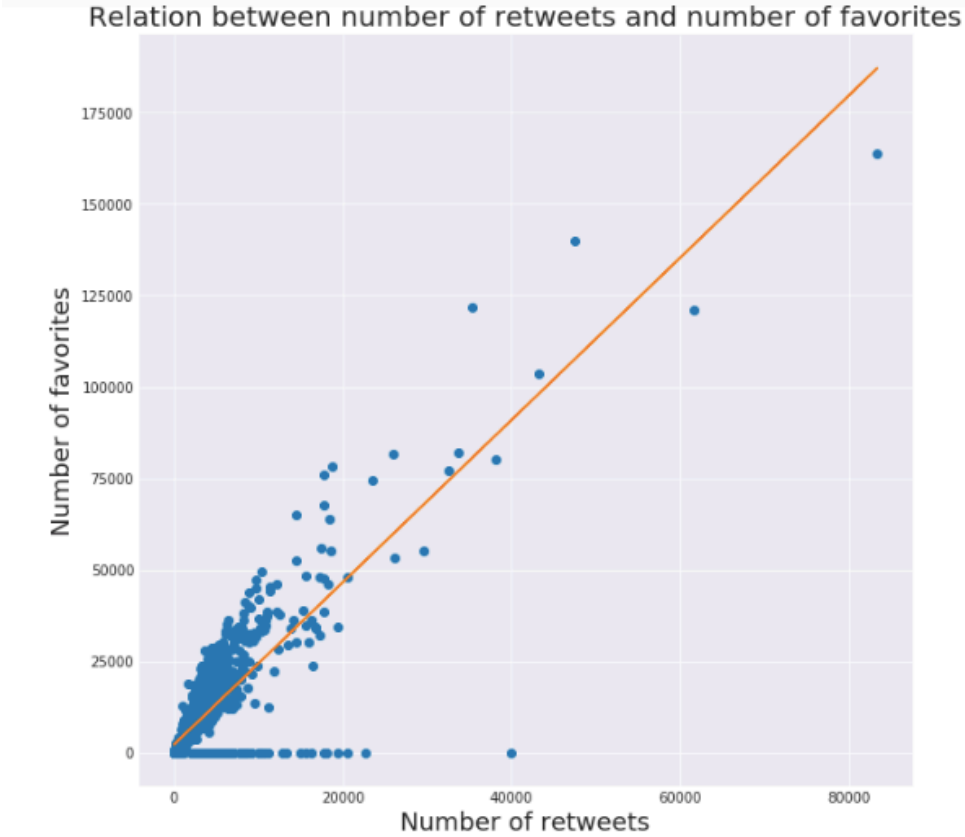
I was finally wondering, what about the rating? Would a post that shows a dog with a rating of 13/10 have more likes and retweets than one with a rating of 1/10? Or is it the opposite?

I have decided to draw a heatmap with buckets of likes in the x-axis and buckets of retweets in the y-axis. The colour scale depends on the rating fraction, that is the quotient between the rating numerator and the rating denominator.

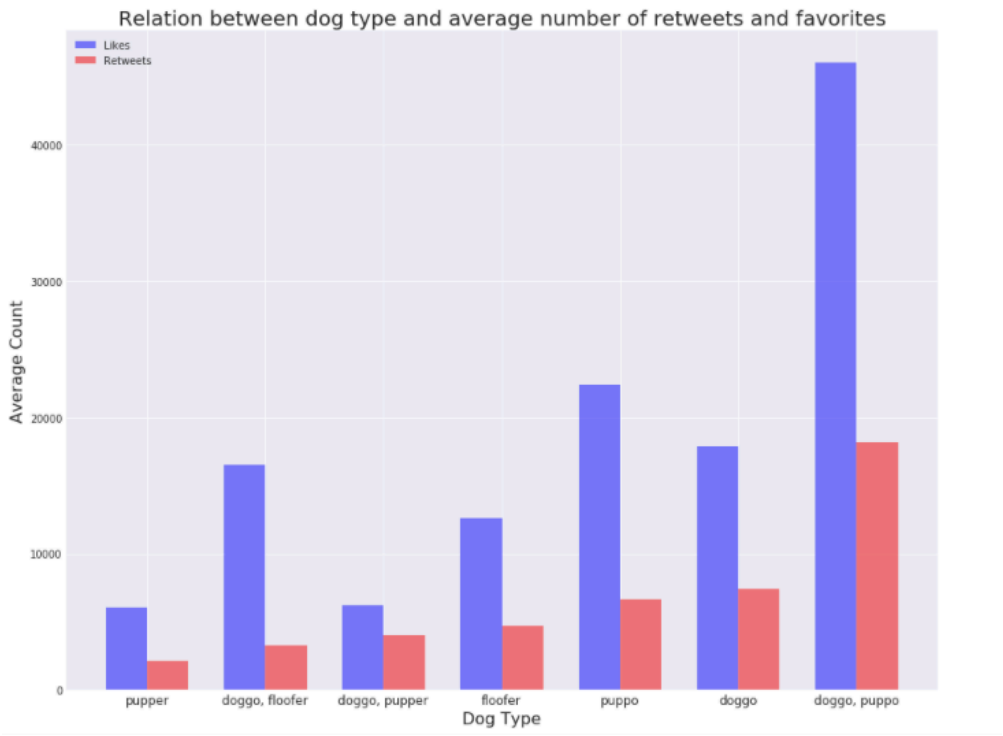
Many observations can be drawn from the heatmap above - all the posts with 8000+ retweets have 10000+ likes - looking at the 10000+ favorites, they all have a quite high rating (with rating here considered as quotient between numerator and denominator) - also the RHS of the heatmap, with likes ranging from 6000 to 10000 have all quite high ratings - the highest ratings can be found in posts having 7001-8000 likes and 4001-6000 retweets - the lowest ratings can be found in posts having 3001-4000 likes and 201-400 retweets



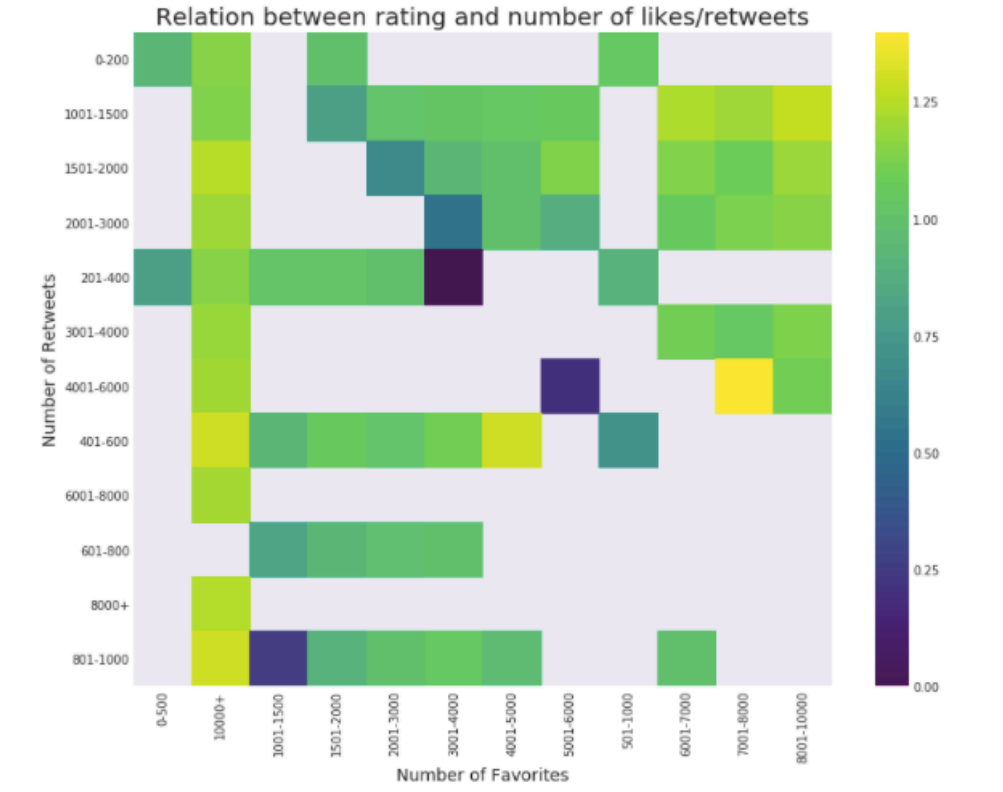
alt text



alt text



alt text



alt text

These are the key points I have drawn from the heatmap but definitely many more observations can be made and investigated!