Act report

Note: I had less time for this part of the assignment (since spent a lot of time on the wrangling part, which turned out to be a bit harder than I thought – especially storing the Twitter API results because of different top level json keys).

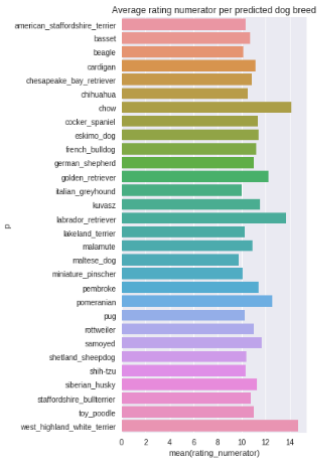
1. Investigating the dataset

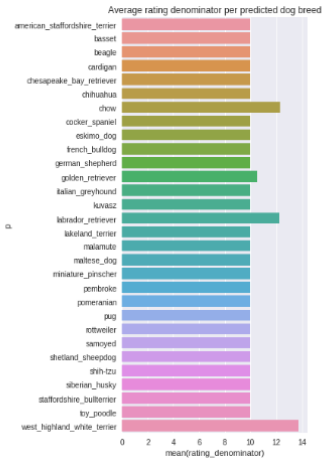
I wanted to investigate the following questions:

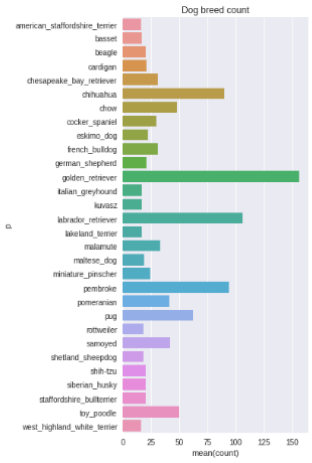
- how number of tweets, average amount of retweets & favorite count was developing in time using time series plots;
- how many predictions were made for most common breeds, what is an average number of retweets per dog breed prediction & favorite count
- what is average confidence, numerator, denominator values per dog breed prediction
- what is the number of tweets per source (what kind of application users of WeRateDogs are using)

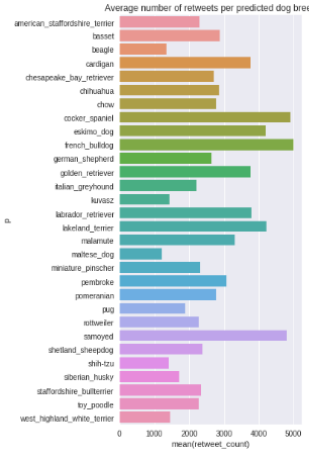
The insights I got:

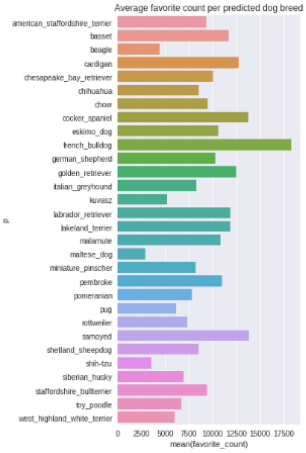
- Most WeRateDogs users use Twitter for iphone to do the tweets
- Golden retriever is the most common dog breed (according to the predictions), also breed 'golden retriever' is predicted with high confidence on average.
- French buldog (which does not occur very often in the tweets) has the highest average favorite count & retweet count.

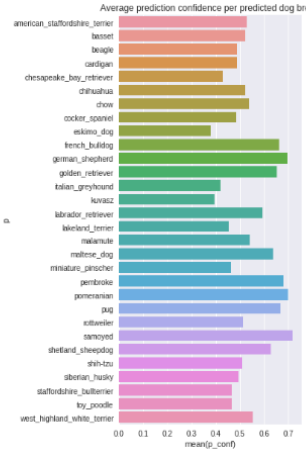












There are some other things I could investigate (for instance, find most common names per dog breed, look at time of posts, make time series plots on a day level rather than a week level). I also could spend time on creating the plots which are easier to read for a human. Now it is hard to get any conclusions from looking at the plots.