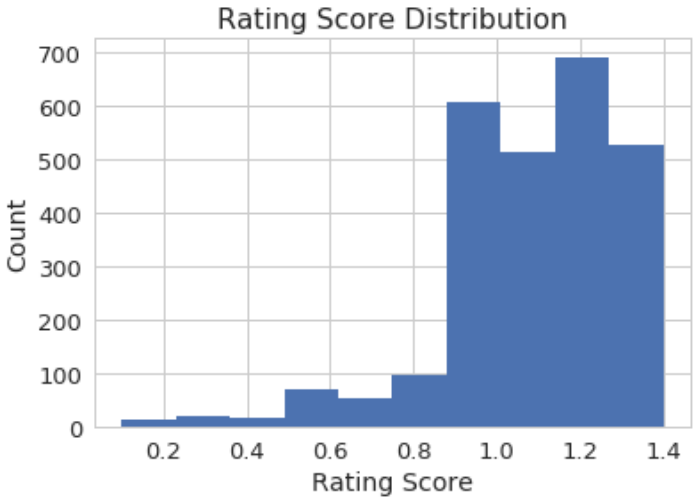
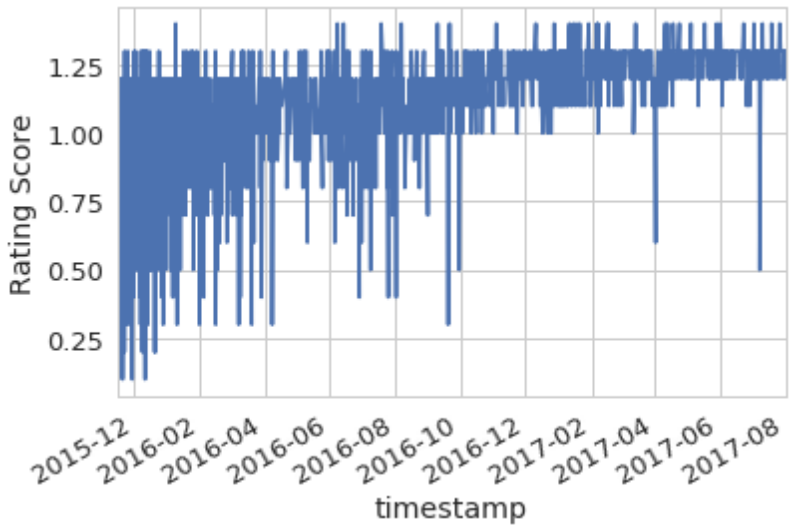
Act report

Bo Fan

I gain a few insights after doing data wrangling and putting the data from 3 sources into a master data frame. The following figure shows the histogram of the rating score, which is the rating numerator divided by the denominator. We find that most of the people like to rate the dogs above 0.9 or more (9/10, 10/10 or more).

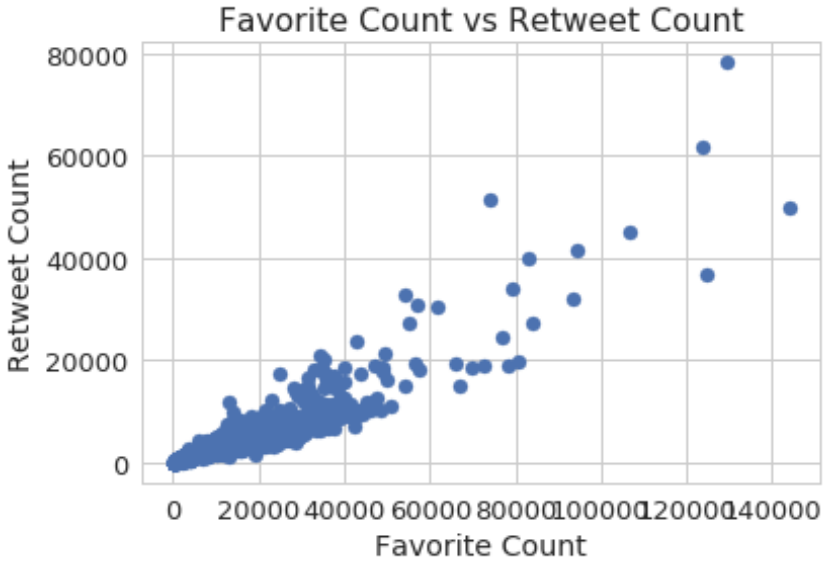


When plotting the rating scores as a function of timestamp, we find that the standard deviation of the scores in a given timeframe becomes smaller in the recent few months or years. In the end of 2015, the score fluctuates in a big range from 0.2 to 1.25. In 2017, however, the range becomes smaller from 0.5 to 1.4. It may be because people are becoming more objective in rating dogs.



It is also intuitive to look into the relationship between retweet count and favorite count, since

dogs with large favorite counts are more likely to be revisited. The following figure also shows the strong positive correlation between favorite count and retweet count. One can see that for favorite count around 20,000 the retweet counts are roughly 10,000. When favorite count becomes 200,000, the retweet count goes up to 60,000.



Additionally, we also provide the distributions of the prediction confidence scores according to the image prediction data set. The predictions are generated by a neural network. As shown in the figure, I use blue, green and red to display the histograms from 3 candidates. For the 1st prediction in blue, the confidence scores are uniformly distributed in the range of 0.2 to 0.8. The 2nd and 3rd place prediction scores are mostly skewed to the left, which satisfies the common sense that the classifier is more confident about the top ranked label.