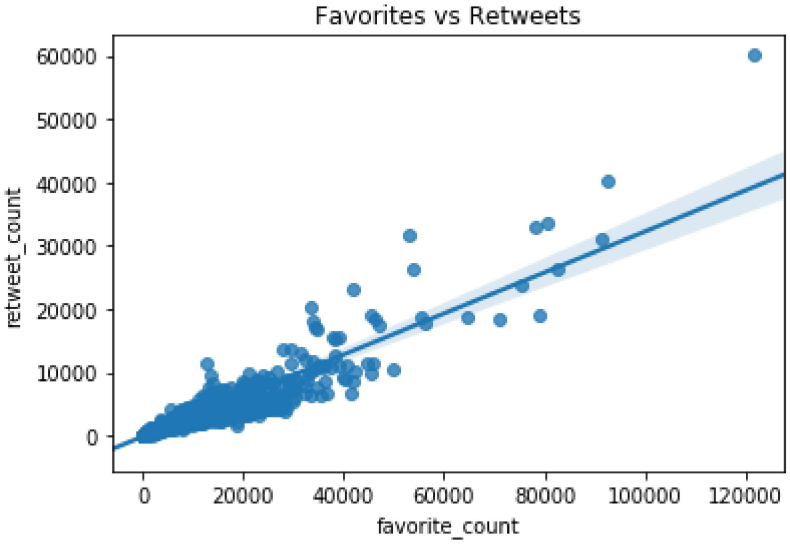
Jed Meng

Insights and Visualizations

I began my analysis just creating a new dataframe with data that I could run statistical analysis on. I ended up with a small dataframe of 5 columns and I began looking into what could be significant. There definitely seemed to be a few outliers since our standard deviation was so high for favorites and retweets. This makes sense because popular posts gain exponentially more popular since more people retweet. I then chose to just look at the most popular tweet since it was such an outlier and we saw that the predictions were accurate and the image was of a Chihuahua. I then decided to take a look at the correlation between favorites and retweets and it turns out it is very high (see graph below). The correlation coefficient was .92 which signifies very high correlation. The mean rating for the dogs was 12.23 which is mostly a joke so running more analysis on this number would not be too important but this number matches joke since most dogs should be around 12-13/10.



The prediction confidence is also very interesting. The algorithms used to gather this data were about 45% more confident in their first prediction than their second. The third prediction was very low with a mean of 6%. It looks like the prediction was actually incorrect with our most popular tweet since it guessed it was a Chihuahua. The number 1 tweet is actually a video of a corgi getting dragged around on a Swiffer. I would assume that in this case, the machine learning algorithms were not that sure about what the video depicted since it guessed with only 50% confidence. I could see the resemblance between corgis and chihuahuas but I did expect a correct outcome. Maybe videos are less easily recognizable? A deeper dive into how the machine learning algorithm worked would be very interesting!