As a first stab at this it is clear that my data needs some re-factoring and some re-imagination. My initial thoughts were to use directors and actors and genres as features to indicate strong returns on investment. Unfortunately, these are categorical variables, and turning them into features is proving to be more challenging and involved than I initially anticipated. To turn the categorical variables into something a linear model can understand, I took every actor, and calculated the mean ROI for movies that they have appeared in, This seems to work fine for actors that have a high number of appearances, but breaks down when there are fewer performances. I repeated this process over the other categorical variables I have in my feature set. I'm not entirely convinced that linear regression is good model for what I'm trying to do here. I know I have pretty good data, and I have plenty of observations. It's clear to me though that my categorical data either needs to be more effectively translated into something a linear model can understand, or I'll need guidance in selecting a model that might work well. I have yet to do a train test split.