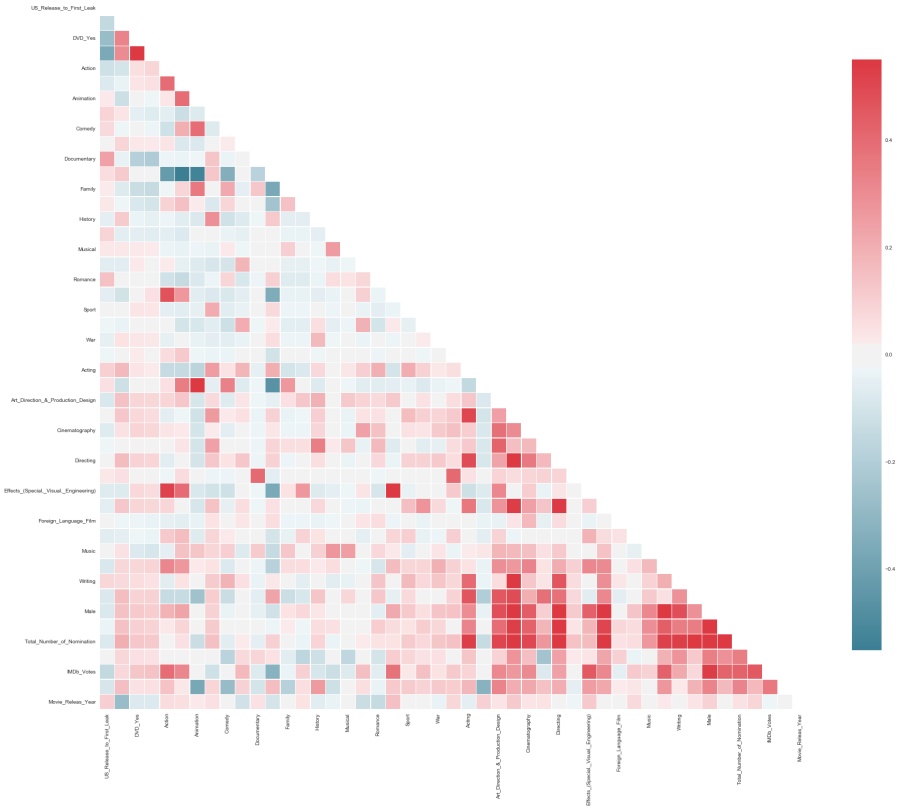
MVP Document- Rohan Shah

Problem Statement

Piracy has become an expensive cost to Hollywood and the movie business. A lot of times hackers are able to get their hands on copies of movies in either a blu ray, HQ, DVD or Oscar film script format and are able to release these films on the internet days after or sometimes even before cinematic release. A lot of times such a release may lead to significant loss in revenue for a film making it pretty important to predict exactly when (no. of days after or before cinematic release) a given film maybe pirated and released for public consumption.

Method

The data collected includes piracy information on 600 films which have all been nominated for an Oscar in a certain category. Because of the high value and prominence of these films they may be more likely to be targeted by internet pirates than other. This data was joined with dataset on movie related information for each film in order to gain more details on the same. The following table was used to chart and understand correlations between the variables in order to perform regressions.



After understanding correlations, I will be predicting the “No. of days after or before US release” variable to understand variables that may lead to higher chance of piracy release and when the release may happen in comparison to cinematic release. Since we have nearly 45 variables it will be important to remove insignificant variables and improve Adj. R square values

manually before using more advanced techniques like Lasso and Ridge to improve accuracy scores.