**Oscar Award Winner Predictions**

Course Project – Patrick Prothro

**Abstract**

For my project I decided to see whether I could build a model that could predict the winners of specific Oscar categories based on characteristics of winners from past Academy Awards. At some point I would love to extend the predictions to each and every category, but for the purpose of this project the three categories I decided to predict were; Cinematography, Directing and Best Picture. The following report details the steps I took to gather, clean, and transform the data. It will detail descriptive plots that helped add context to the overarching problem. Last, it shows the results of the models and the features which ended up being the most important in terms of predictive accuracy.

**Statement of Research**

Living in Los Angeles, the entertainment capital of the world, it’s hard not to be to some extent drawn into the buzz of awards season. This past Academy Awards were especially interesting to me personally because I saw the majority of movies up for nominations. Each film is independent of any film to come before and after it in theory, but I questioned whether there were patterns that could be derived from historical data which could be used to make educated guesses on future winners. Others have undergone this same challenge with a high degree of accuracy, but from what I’ve noticed their predictions are heavily influenced by the results of other award ceremonies prior to the Academy Awards (Golden Globes, British Academy Film Awards)[[1]](#footnote-1). There is nothing intrinsically wrong with this, but for my purposes I want to see whether I can make predictions on films without the knowledge of how they have already fared in previous competitions. I’m assuming that among different award ceremonies the criteria for winners is fairly close to homogenous. For example, Leonardo DiCaprio is an amazing actor and has starred in several movies where himself or others have been nominated for awards. Is simply having him in the cast of any such movie raising the probability of a film winning an award? These are the questions I want to see whether I can answer with this project.

To help try and answer this question I decided to grab data for Oscar nominated movies dating back to 1980 and see if I can predict the winners from the 2015-2019 Academy Awards in the categories of Cinematography, Directing, and Best Picture. For reference throughout the rest of the report the Academy Award year is for movies that came out the year specified, not when the award show took place. For example, the year 2019 in the dataset represents movies that released in 2019 even though the award show took place in February of 2020. The models used would be based around classification where the film with the highest probability of winning within a category and a year would be predicted as the winner.

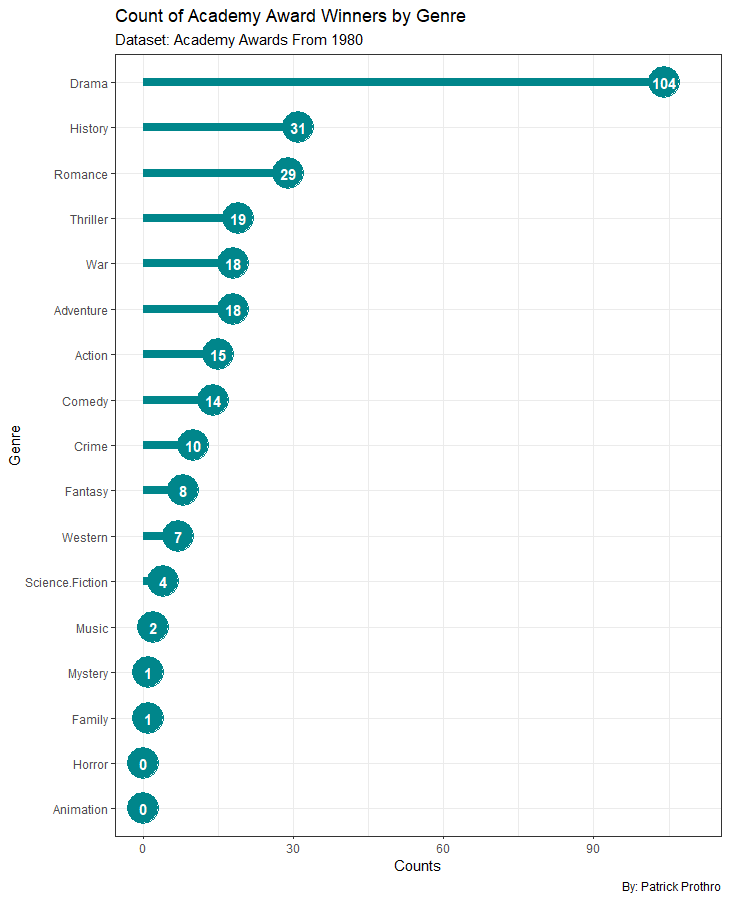
**Data Collection**

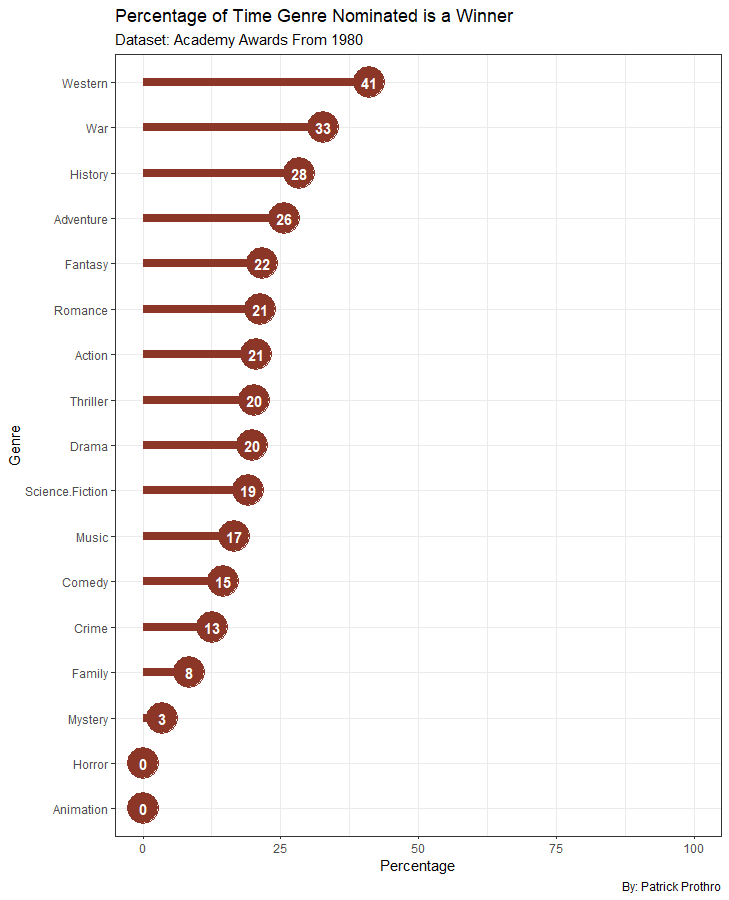
The dataset used for this project came from a variety of sources. First, I needed to get a comprehensive list of Academy Award nominees and winners over the past 40 years. Luckily for me someone had already compiled this list[[2]](#footnote-2). The next part of information I wanted were detail oriented characteristics regarding each film. Characteristics such as, the top 3 actors/actresses in the film, how long was the film, what genre, etc. This information was found using an API from the movie database site[[3]](#footnote-3). This allowed me to programmatically send the award nominees to the API and request an output of information for each movie. This worked seamlessly for roughly 90% of the data, however due to discrepancies in titles, such as the word “and” instead of “&”, it required some manual labor on my part to make sure my data was not only clean, but accurate. The last piece of information I wanted for my dataset is the average review score by critics of each film. I was able to gather this data using a Rotten Tomatoes API[[4]](#footnote-4). Similar to the movie database API, this also required some manual work as far as adjusting movie titles to line up consistently across the API and the list of Academy Award nominees.

At the end of the data cleaning process I had ended up with the following variables for each movie: Top 3 Genres, Top 3 Actors/Actresses, Top 3 Production Companies, Top 3 Cinematographers, Top 3 Directors, Top 3 Producers, MPAA Rating, Original Language, Movie Runtime, Budget, Revenue, and Release Date. Top 3 of any category means that potentially a movie could have more than one in each of those categories, so for simplicity sake I took the top 3 listed for those specific categories.

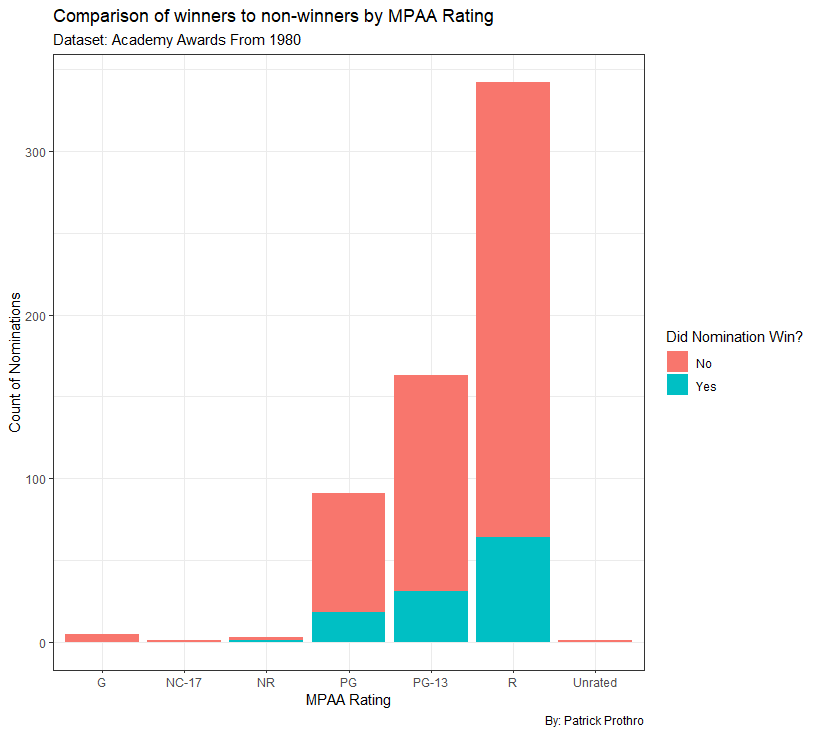
**Descriptive Statistics**

Before diving into the predictive modeling aspect of this project, I decided to get a better sense of the data and understand some of the nuances and relationships between the predictor and the covariates. Below I’ve added some exploratory data analysis visuals that gave me a better understanding of the underlying patterns (if any) associated with the study.

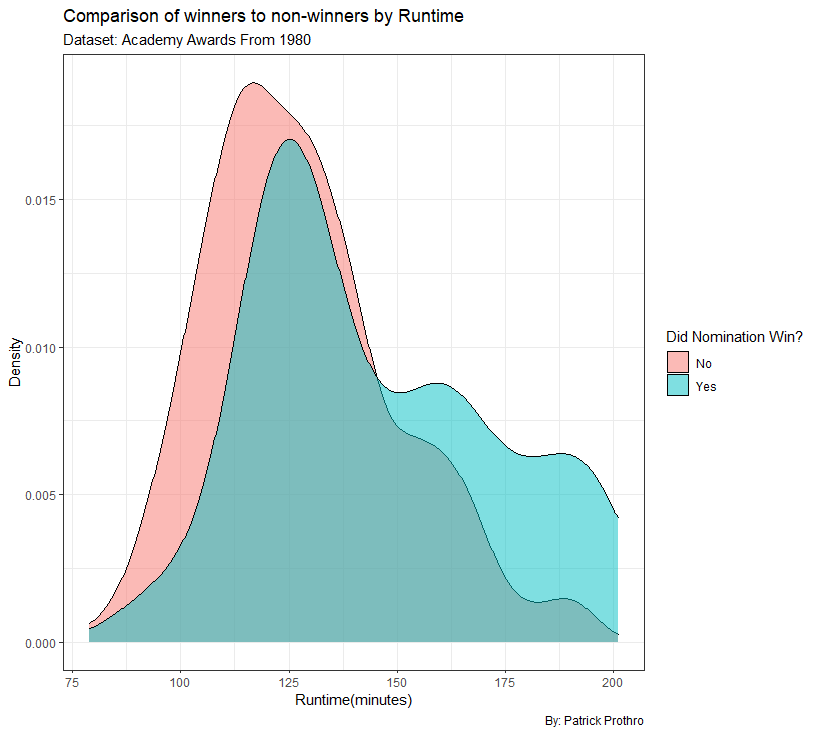


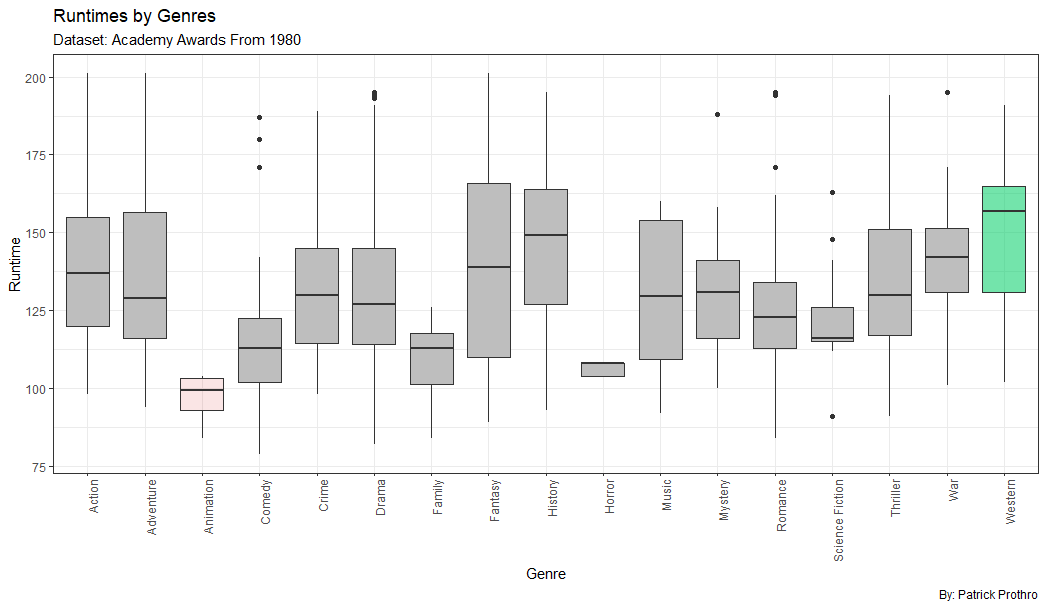


I found the above plots interesting because they give a sense of what type of movies are winning these awards. The first plot gives a raw count of winners by genre. Drama is the overwhelmingly majority winner, which I believe intuitively makes sense. Drama as a genre is fairly ambiguous and can probably be considered a sub-genre within most movie types. What I believe is more interesting is that in the three categories we are predicting on, movies in Animation and Horror have not won once! The second plot is also interesting as it describes when nominated what percentage of the time that genre is winning. Given the large counts of war movies that have been nominated as well as the fact that they win nearly a third of the time when they are nominated, war movies always seem to be a safe bet to be taken seriously at the Academy Awards.

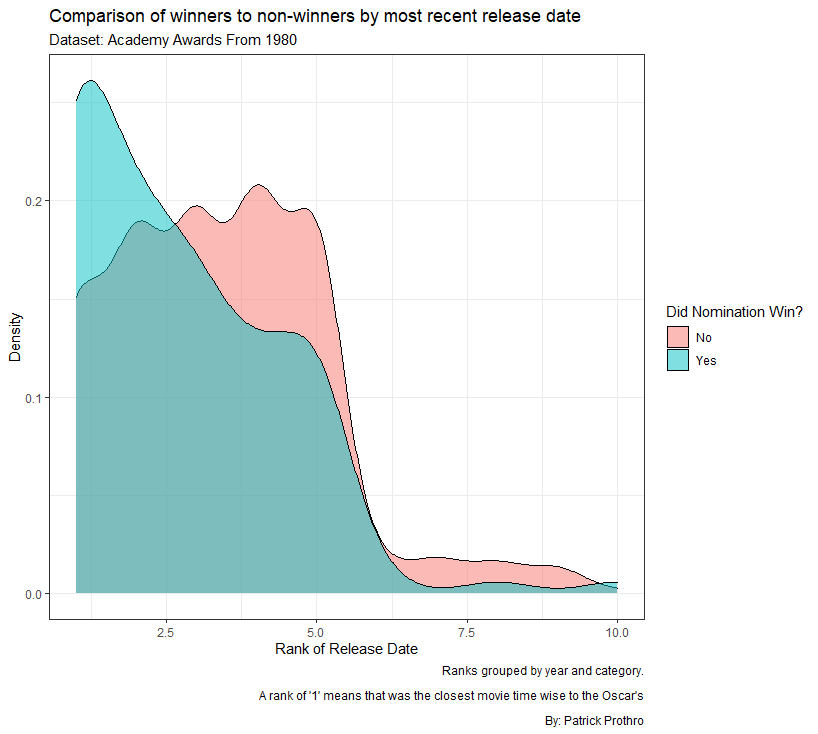


This plot shows that overwhelmingly movies rated R, are nominated and end up winning Academy Awards. At first glance I found this surprising, my guess from a marketing standpoint was that a PG-13 rating would strike the perfect balance of widespread consumption, but still being able to display moderately mature themes that these awards garner. Further research I found, showed that since the MPAA Rating system came into existence that overwhelmingly most movies are rated R (58%)[[5]](#footnote-5). This could very well help explain this chart.





This density chart was extremely interesting because it showed that the longer a movie was the more likely the odds of it winning were. The boxplot also gives some additional context with how long movies typically are by genre. Animation and Horror movies(consistent losers) are generally shorter than your average movies, while Western movies(consistent winner) tend to be longer than most which helped explain some of the findings here.



Another interesting density chart was showing how it appears to be “Oscar bait” to schedule your movie release as close to the Academy Awards as possible. A rank of 1 here means that the movie was released the closest to the Academy Awards in comparison to the other movies nominated in that same year and within the same category.

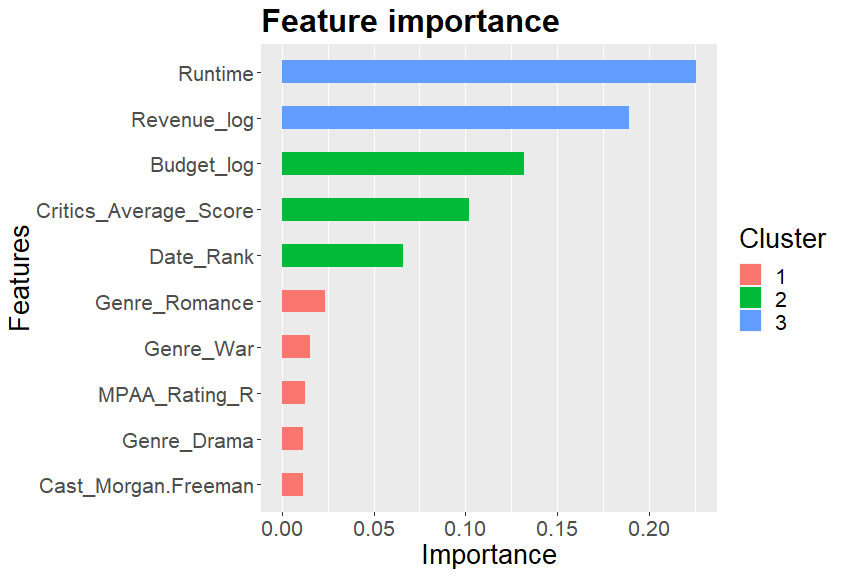
**Modeling**

The best way I believed to model this problem was using classification. The method would be to have those who won the Oscar as a “1” and those who didn’t win as a “0”. From there run a classification algorithm where the predicted “winner” would be the film with the highest probability in a given category for the year. I first tried to create a new model for each category which I believe is the correct approach for this type of problem, however after testing I received better results when using just one model and coding the different categories as their own binary columns. My hypothesis is that a joint model has more data to make conclusions from and so for the sake of this project I used a singular model for all the data. All categorical variables were one-hot-encoded and this created a very expansive dataset of 2185 columns and only 550 rows. Given these constraints I decided to use two models; Ridge Regression and a boosting algorithm, XGBoost. The Ridge model is great for prediction accuracy where you have several variables but need the coefficients of some variables reduced if they are either correlated with another variable or not relevant in prediction accuracy. XGBoost is also great for this use case because it combines both Ridge Regression and Lasso to prevent the model from over-fitting and to ultimately select the most important variables in the dataset. It also allows for hyper parameter turning via cross validation to ensure that you can optimize your model to the best of its ability. XGBoost is an iterative process where it selects a random subset of the variables and tries to predict the Award Winners based on a random subset of data until it achieves a solution that is no longer significantly improving. The models were trained on the Academy Award nominees and winners for the years between 1980-2014 and then tested on data between 2015-2019. In essence, we are trying to test our model on 15 different Oscar awards to see how many of those awards do they accurately give the highest probability to the actual winner of the award.

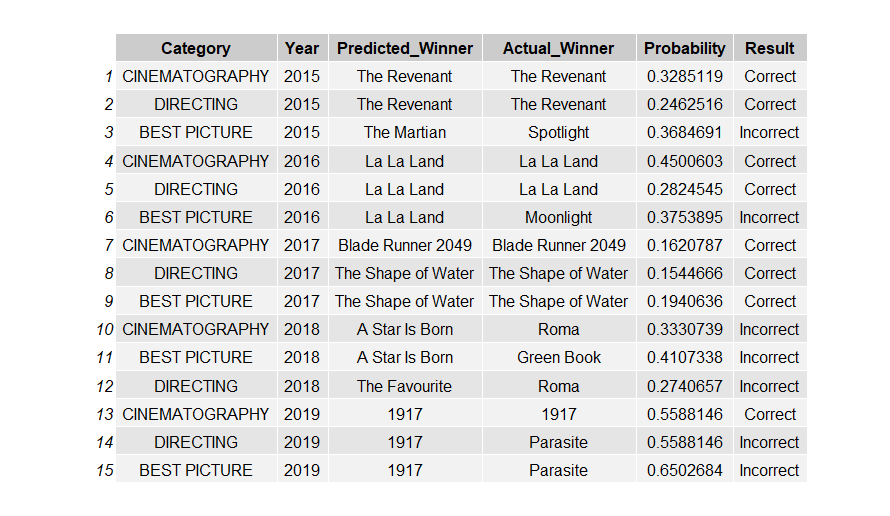
**Results**

The Ridge Regression model was able to consistently predict the winner of the Oscar 6 out of 15 times. What I found interesting from the coefficient values of this model, was that the strongest coefficients were typically actors/actresses and producers that won or loss multiple times for the same movie. For example, the largest coefficient values for this model were Debra Winger and Shirley Maclaine both who were in Terms of Endearment which won both Best Director and Best Picture in 1983. Neither actress was in another movie within this dataset. If they ever were however, on new and unseen data, the model would give that movie a significant boost towards winning the award simply because of those two actresses who up to this point never lose. This intuitively makes sense, but from an interpretability standpoint, it’s hard to decipher what’s truly important in predicting Oscar winners. However, the model still did reasonably well and much better than if a random person decided to pick winners. Mathematically, a person randomly selecting and correctly picking 6 or more Oscar winners from the listed nominees would be less than 1%.

The XGBoost model was able to improve on the Ridge Regression model and predict 8 out of 15 winners correctly. The hyperparameters were turned using cross validation and below is an image of the most important features deemed by the algorithm. The feature importance attribute of XGBoost doesn’t notify you of “how” that specific feature is important (positive or negative) however, given density plots I’ve made from the data (some displayed above) it appears to me that the longer the film is, the more revenue the film generates, and the larger the budget the more likely that film is to win an Oscar. Other significant indicators as displayed in the chart include; how close the movie’s release date is to the Academy Awards, if the critics scored the movie highly and if it was a Romance or War movie.



**XGBoost Model Results**



**Conclusions**

All in all, I was pleasantly surprised by how well the models specifically XGBoost did in predicting Oscar winners. I believe the data brought in for this experiment was a good starting point, however the possibilities of factors that may impact who wins an award are probably endless and worth exploring to see if this can improve predictions. While both models did great in predicting 2015, 2016, and 2017 winners they were terrible in predicting winners in 2018 and 2019. I’m curious whether this is a fault in the models or a shifting in paradigm of those assigned to vote for these movies. By most historical metrics 1917 (a war movie released in late 2019) should have won more awards, however Parasite was actually the big winner, the first South Korean film in history to win an Oscar and a movie that was released in mid-summer. Nowhere near award season. The Academy Awards have been under a lot scrutiny lately due to claimed lack of diversity in their selections of nominees and winners[[6]](#footnote-6). Could this have an influence on recent winners such as Green Book, Roma, and Parasite? Perhaps, if this becomes a repeated pattern, maybe the models will select new important variables to predict future Oscar winners.

1. Oscars 2020: Harvard math grad uses data to predict winners with near-perfect accuracy

   <https://finance.yahoo.com/news/oscars-2020-harvard-math-grad-predicts-winners-with-near-perfect-accuracy-120453261.html> [↑](#footnote-ref-1)
2. List of Academy Award Winners

   <https://datahub.io/rufuspollock/oscars-nominees-and-winners> [↑](#footnote-ref-2)
3. The Movie Database API Link

   <https://developers.themoviedb.org/3/getting-started/introduction> [↑](#footnote-ref-3)
4. Rotten Tomatoes API Link

   <https://pypi.org/project/rotten_tomatoes_client/> [↑](#footnote-ref-4)
5. More than half of all MPAA rated movies have been rated – R

   <https://www.slashfilm.com/more-than-half-of-all-mpaa-rated-movies-have-been-rated-r/> [↑](#footnote-ref-5)
6. The Oscars diversity problem

   <https://www.ft.com/content/ca2e8368-48e6-11ea-aeb3-955839e06441> [↑](#footnote-ref-6)