The Oscar Awards and Consumer Preference

East Coast Regional Datathon - ECR team3

Hojin Lee[1], Tae Yoon Lim[1], Tony Tohme[2], Haoyang Wang[1]
1 {hl3328, tl2968, hw2758}@columbia.edu 2 tohme@mit.edu

September 20, 2020

1 Executive Summary

The Academy Awards, also known as the Oscars, is the most prestigious awards ceremony for people working in the movie and entertainment industry. We were curious about the relationship between the consumer's preference and the winner of the Oscars.

1. Is consumer's preference predictive of which films will be nominated and received the specific academy awards ?

2. How do the consumer's preference change after a movie receives an award?

To answer these questions, we first checked if there is a statistically significant effect on the ratings of the movie before and after the Oscars. Using statistical methods, we identify ratings/reviews as proxies for consumer preference, and find that it does have a significant effect on the Oscar Awards, and is also influenced by the Oscar Awards afterwards.

Then, to further explore how consumer preference contributes to predicting the winners of the Oscar Awards, we construct a novel method to track the consumer's preference on certain kinds of movies, and build a machine learning algorithm to predict the movies that have high chances of being nominated for the awards.The best performing model showed about 90 percent accuracy in detecting non-Oscar material, and around 50 percent accuracy for the actual Oscar-material. Given that the data was severely imbalanced, this is an impressive result. And as indicators of consumer preference, revenue, votes, and ratings are the most important features in our prediction models, strengthening the claims that consumers' preference could influence Oscar awards significantly.

2 Initial Exploration

We first started our analysis by investigating the data in `movie_industry.csv`. Firstly, we wanted to check the distribution of the data. From Figure 1, we realized that there aren't any significant differences in the number of movies in the data set we have. Then, we moved on to consumer preference analysis. It seemed like the run-time of the movie and genre play significant roles in preference. In terms of consumer preference, we could observe disproportionate distribution of movie genres. In addition, the average ratings of the genres over the years are volatile. For further analysis, we performed dimension reduction learning to further analyze the relationship between genre and preference.

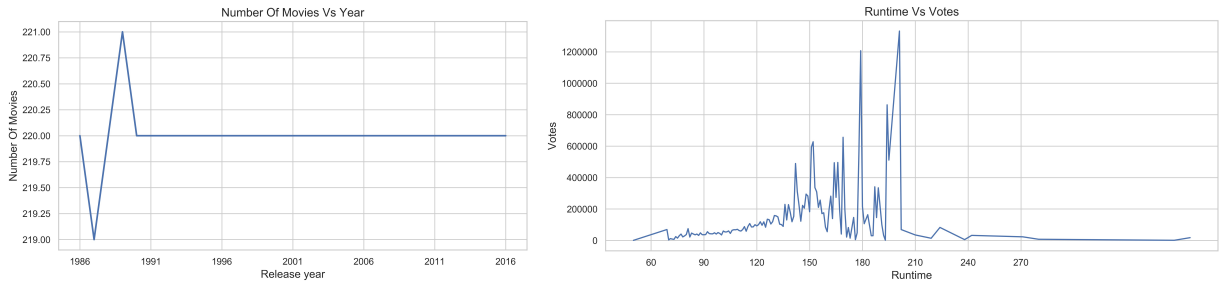

Figure 1: Number of movies per year and movie duration vs votes.

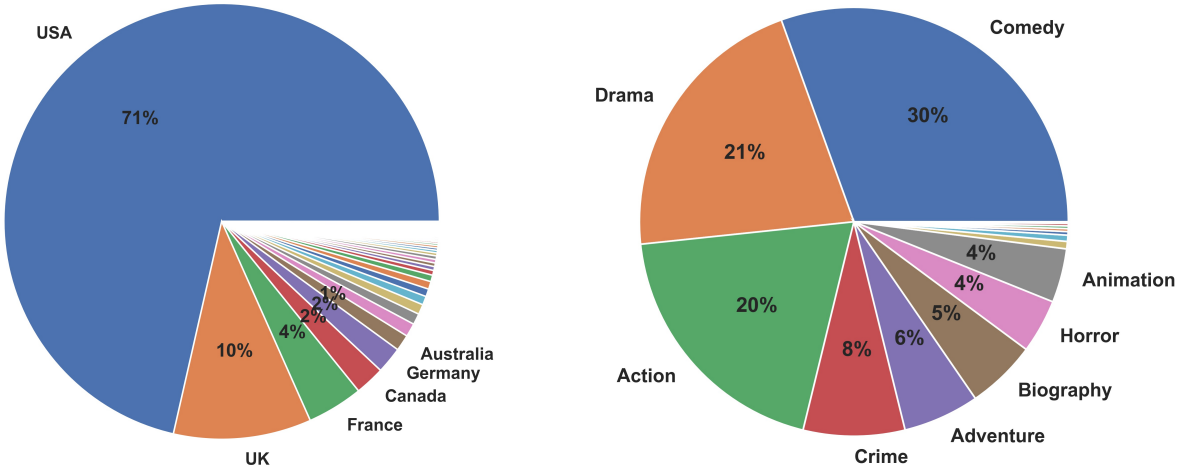

Figure 2: Countries and genres with the highest number of released films.

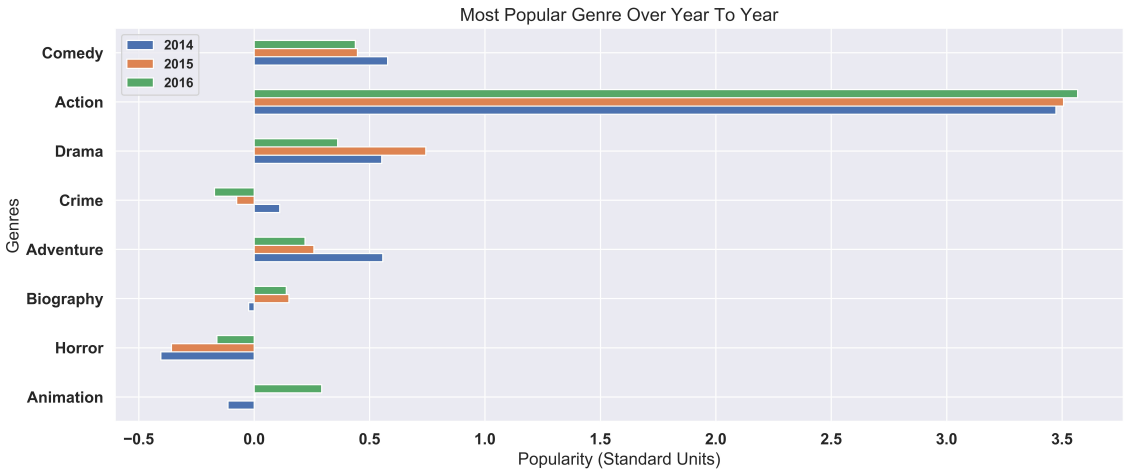

Figure 3: Genre popularity over the years.

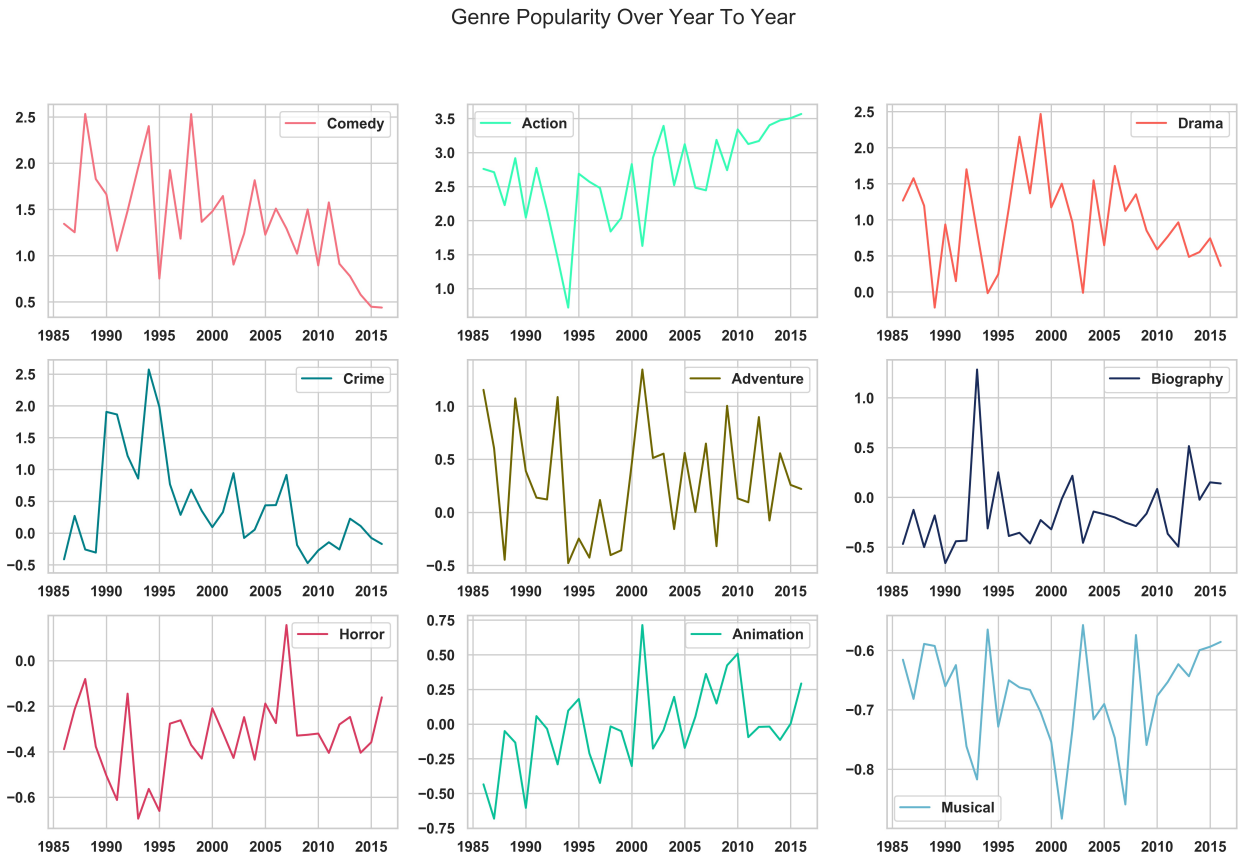

Figure 4: Genre popularity over the years.

3 Oscar Award Winners & Consumer Preference

We first study the dynamics between consumer preference and the Oscar Award winner, to statistically answer the 2 questions mentioned above. Could consumer preference determine which movie will win the Oscar Award? Would consumer preference change once the Oscar Award winners were announced?

3.1 Data Construction

3.1.1 Dataset Used

Because we need 'year_ceremony' (the year of an Oscar Award ceremony) variables for the analysis, we use the data from `the_oscar_award.csv`, i.e., all Academy Awards nominations since 1927. Since it's impossible for a movie to get nominations in multiple years, we don't have time-series data. Hence, we treat these nominated movies as cross sectional data (non-sequential data). Most nominated movies had higher ratings than those without nomination. We will show that even with high ratings, the increase in ratings can still generate a positive effect on the Oscar Award results.

We can see the difference between nominated movies and the entire dataset. Note that 'preRatings_mean' represents the ratings before the Oscar Awards ceremony, the construction method is shown below. For the movies in `movie.csv`, we don't have features recording the year of the Oscar Awards ceremony. Therefore, we treat the year after the movies' release as the year of the Oscars. This is not entirely accurate because some movies might be open in advance and the year for ceremony should 2 years later. Hence, we just post the description

Table 1: Ratings in `movie.csv` `the_oscar_award.csv`

	preRatings_mean		postRatings_mean
count	9568	count	1588
mean	3.096909	mean	3.841628
std	0.748054	std	0.418837
min	0.5	min	0.5
25%	2.714286	25%	3.662162
50%	3.17724	50%	3.886039
75%	3.571429	75%	4.043478
max	5	max	5

for comparison purposes, and do not use them in the regression test.

3.1.2 Proxy for Consumer Preference

If we only use `the_oscar_award.csv`, and `movies.csv` (all the movies in MovieLens) as our dataset, we can only choose movie ratings to determine preference for these movies in terms of quality.

Combining `the_oscar_award.csv`, `movies.csv` and `movie_industry.csv` (popular movies' list), we have multiple variables eligible to be a proxy for consumer preference, such as gross revenue (reflecting the ticket sales for the movies), number of votes in IMDB (showcasing how many people have watched the movies), scores (in IMDB) or ratings (in MovieLens). However, we find that there are strong correlations between these variables. In that case, if we use all the variables in our regression test, it may cause strong multicolinearity problems, making the standard error extremely high and the significance test meaningless.

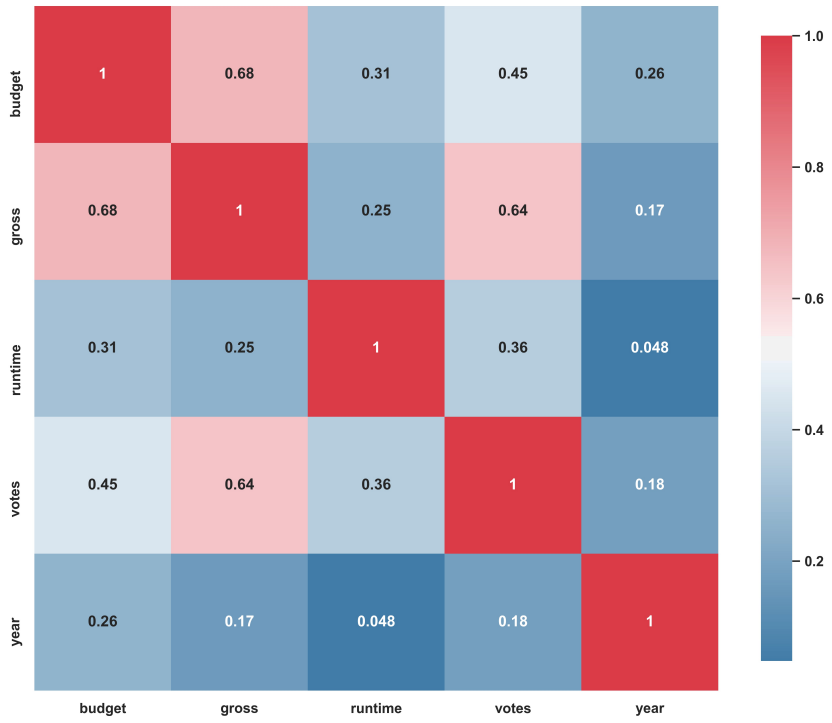

Figure 5: Correlation map.We can see the close relationship between 'gross' and 'votes'.

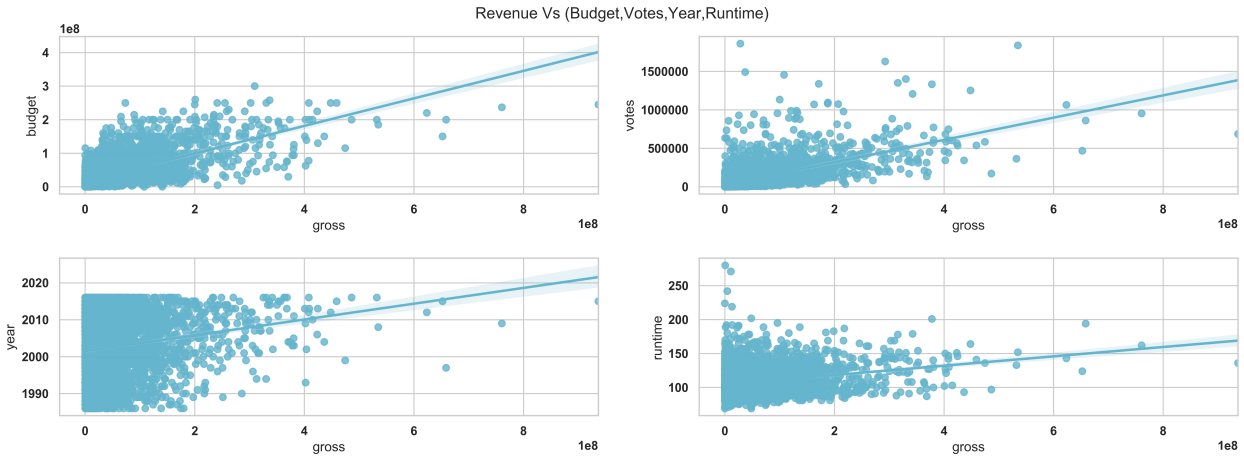

Figure 6: Revenue vs {budget, votes, year, runtime}.

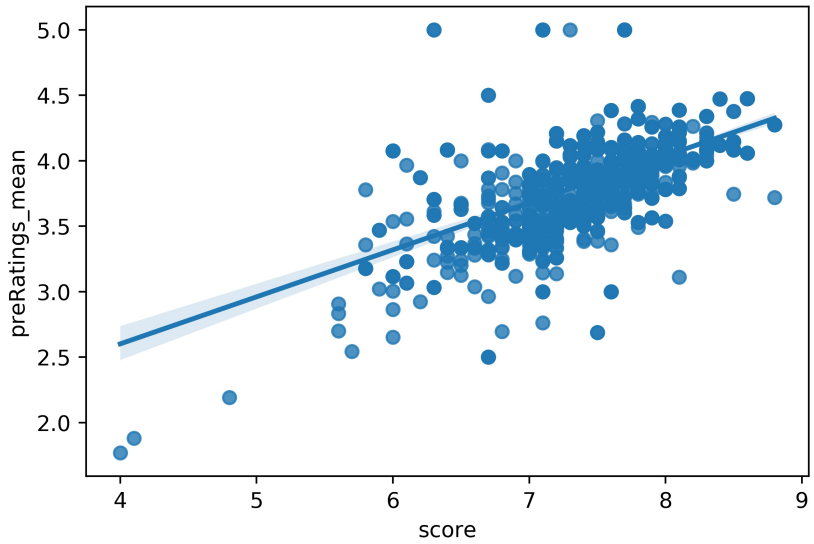

Figure 7: score vs. rating (named 'preRatings_mean'. Correlation is over 0.6.

Therefore, to mitigate such problems, we choose one of them as the proxy for consumer preference, with prior values as the main independent variable (feature) to predict the Oscar Award winners, and posterior values as the dependent variable (target) to rely on Oscar Award winners. Here, we choose the ratings in MovieLens because:

1. most importantly, we have timestamps for ratings in MovieLens. We can only use the ratings before the Oscar Award as prior values to predict the Oscar Award winners, and the ratings afterwards as posterior values used to analyze Oscar Award's effect. It avoids the usage of future data in data science's view and endogenous problems in statistical analysis' view (bi-directional influences between Oscar Award and Consumer Preference will definitely cause endogenous problems).

2. we can use it for analysis with or without consideration for `movie_industry.csv`.

3. 'gross' and 'votes' are highly correlated to 'genre', e.g. Action and Comedy movies will have higher gross and more votes. If we control the 'genre' variable in the regression model, to purify consumer preference's influence on Oscar Award, using these variables will cause multicollinearity problems.

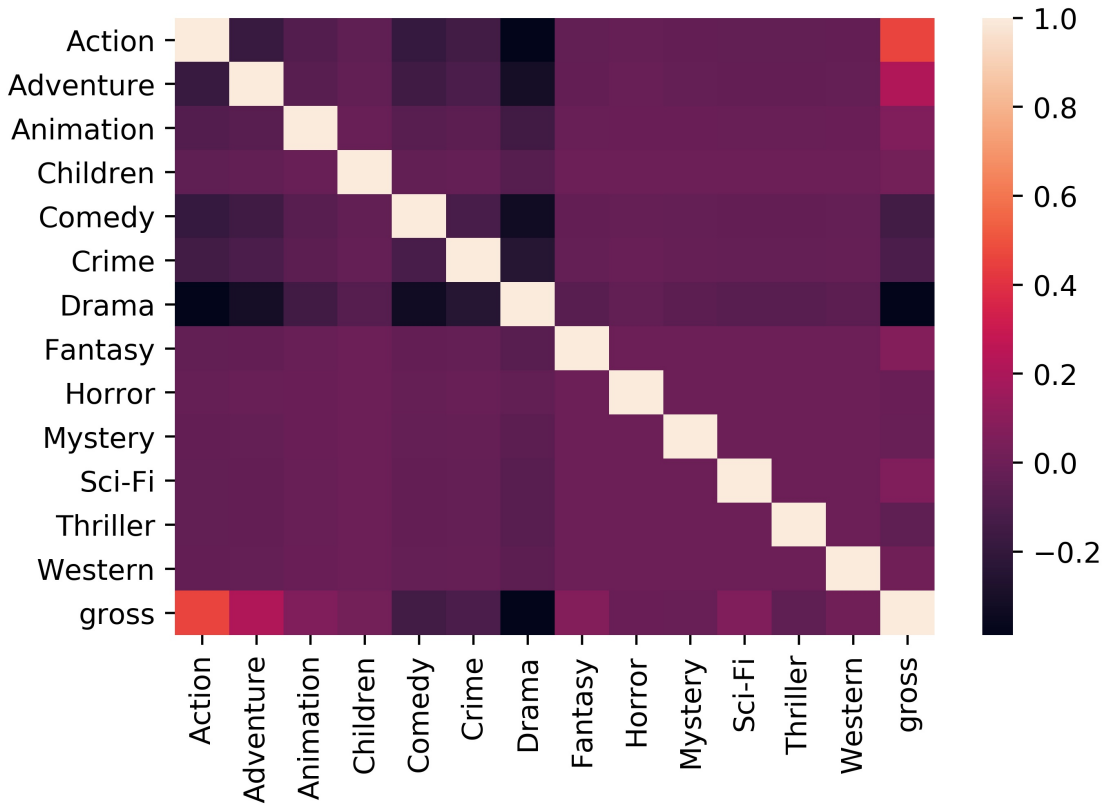

Figure 8: Correlation map. Gross is correlated to Action and Adventure movies.

It's also remarkable that the 'year' variable is highly correlated to these features, meaning that all these features have an upward trend within the same year. Consequently any variable recording the years should not be included in regression analysis.

The Oscar Award ceremonies are held in February or March every year. Therefore, we choose all the ratings before February in 'year_ceremony' for that movie, as the ratings before Oscar Awards, i.e. pre-ratings (named 'preRatings'); and the ratings after March in 'year_ceremony' for that movie as the ratings afterwards, i.e. post-ratings (named 'postRatings').

And in `the_oscar_award.csv`, we only have the ratings from individual users. In order to get the overall ratings for a movie, we can collect all the ratings from different users and calculate the mean (named 'preRatings_mean', 'postRatings_mean'), or the mode (named 'preRatings_mode', 'postRatings_mode'), to represent the overall preference for different movies.

3.1.3 Oscar Award Winners

In `the_oscar_award.csv`, we have different categories of Oscar Awards, e.g. Picture, Actor, Actress, Directing, Writing, etc. If we use one of the categories to represent Oscar Award winners, the TRUE(=winning) values will be much less than FALSE value. Such a highly biased sample is not suitable to the linear regression.

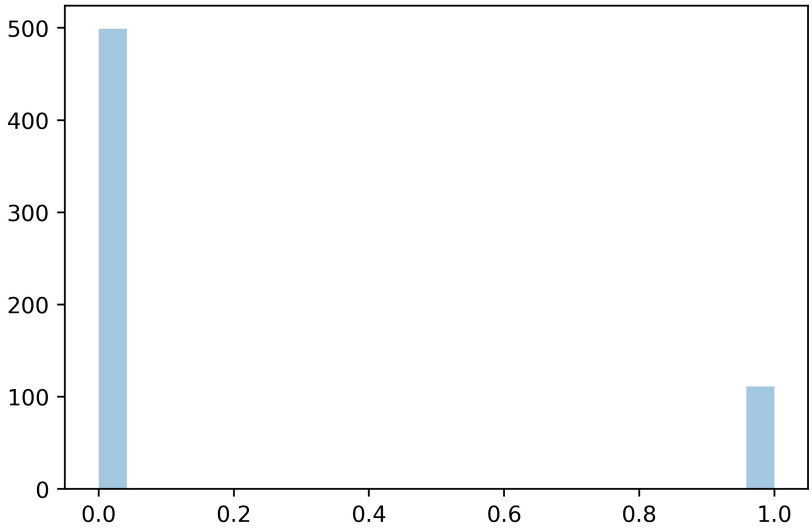

Figure 9: Distribution for Best Picture, with the most TRUE value, but still unfeasible to estimate the density.

So for the analysis below, we used two ways to represent the Oscar Award winners:

1. whether a movie is the Oscar Award winner for at least one category. TRUE = a movie got at least one Oscar awards, FALSE = a movie didn't get any Oscar awards, named 'Oscar winner';

2. the number of Oscar awards a movie got, 'num of awards'.

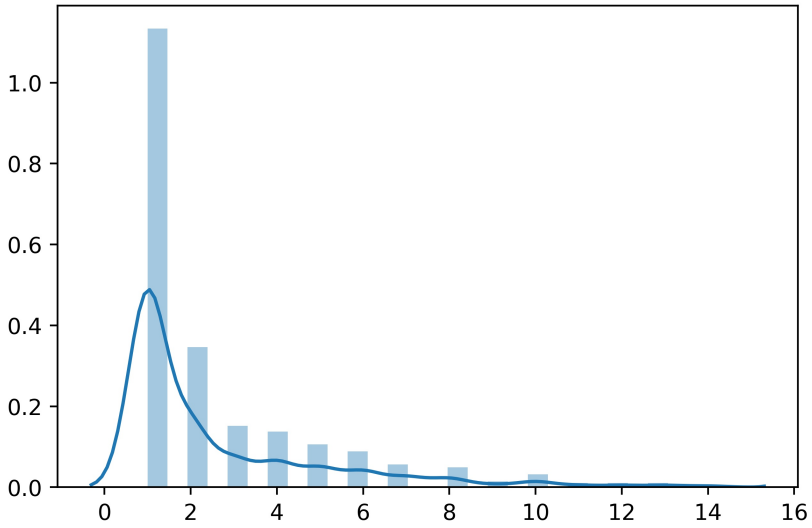

Figure 10: Distribution for 'num of awards'.

If we use Oscar Award winners as the dependent variable (target), the first method will lead to a classification problem, and the second method will lead to a regression problem. We do the analysis in both ways.

3.2 Test Results

In order to statistically show the significance of different variables, we must use linear regression, even for the classification problem. For more accurate results, we built some Machine

Learning models to address the classification task for Oscar Awards. In order to solve the classification task more accurately, we perform different machine learning algorithms below and also prove the significant influence of consumer preference.

Would consumer preference influence Oscar Award winners?
Shown here are the results of examining the influence of consumer preference. We used 'preRatings_mean' described above as the main independent variable and the aforementioned 'Oscar winner' variable as the target variable.

Table 2: Oscar Winner and Pre-ratings

	coef	**std err**	**t**	**P > |t|**	**[0.025**	**0.975]**
preRatings_mean	0.0823	0.005	16.878	0.000***	0.073	0.092

(Note: *** means that it's significant at 1% confidence level)

We can use the 'num of awards' as the target variable. Here are the results.

Table 3: Number of Awards and Pre-ratings

	coef	**std err**	**t**	**P > |t|**	**[0.025**	**0.975]**
preRatings_mean	0.7205	0.026	27.357	0.000***	0.669	0.772

(Note: *** means that it's significant at 1% confidence level)

We can see that 'preRatings_mean' has a significant effect on Oscar Award, regardless of how we interpret the Oscar Award winners. The results are almost identical if we use 'preRatings_mode' as the main independent variable. Therefore, we show the result using 'preRatings_mean' below. This means that higher overall ratings before the Oscar Award ceremony indicates a higher probability of becoming an Oscar Award winner.

We can add some control variables to purify the influence of ratings. By Combining `the_oscar_award.csv` and `movies.csv`, we can add the 'genres' variable (genres labeled in IMDB). Since 'genres' are categorical features, we one-hot encode it and use the one-hot features (dummy variables) in the model. We cannot use all these dummy variables to avoid multicolinearity. We dropped out the one-hot features for 'Action'. We could have also included the 'year_ceremony' variables as a control. However, it is closely related to these consumer preference indicators, as shown in the correlation map above. Furthermore the VIF test shows a strong multicolinearity between 'year_ceremony' and 'preRatings_mean'; therefore, we do not include it. We used 'Oscar winner' as the target. Here are the results:

Table 4: Oscar Winner - Pre-ratings, with Control Variables 1

	coef	**std err**	**t**	**P > |t|**	**[0.025**	**0.975]**
preRatings_mean	0.089	0.012	7.445	0.000***	0.066	0.112
genres_Adventure	-0.0212	0.07	-0.302	0.763	-0.159	0.117
genres_Animation	-0.1028	0.09	-1.139	0.255	-0.28	0.075
genres_Children	-0.0303	0.175	-0.173	0.862	-0.374	0.313
genres_Comedy	-0.0947	0.069	-1.38	0.168	-0.229	0.04
genres_Crime	-0.0279	0.086	-0.324	0.746	-0.197	0.141
genres_Documentary	-0.1051	0.079	-1.33	0.184	-0.26	0.05
genres_Drama	0.0043	0.053	0.081	0.936	-0.1	0.109
genres_Fantasy	0.3456	0.263	1.316	0.189	-0.17	0.861
genres_Horror	0.2115	0.32	0.662	0.508	-0.416	0.839
genres_Mystery	-0.3483	0.263	-1.323	0.186	-0.865	0.169
genres_Sci-Fi	0.3994	0.229	1.742	0.082*	-0.051	0.85
genres_Thriller	-0.0917	0.229	-0.401	0.689	-0.541	0.358
genres_War	-0.3398	0.451	-0.753	0.451	-1.226	0.546
genres_Western	0.1494	0.321	0.466	0.642	-0.481	0.779

(Note: * means that it's significant at 10% confidence level, *** means significant at 1% level)

We can see that 'preRatings_mean' still has a significant effect on who wins the Oscar Awards. Genres have no significant effect on Oscar Award winners. Exceptionally, Sci-Fi films may have a slightly increased probability of winning an Oscar Award. Otherwise, the genre of a movie generally has no effect on whether it will get an Oscar Award. We used 'num of awards' as the target. Here are the results.

Table 5: Number of Awards - Pre-ratings, with Control Variables 1

	coef	**std err**	**t**	**P > |t|**	**[0.025**	**0.975]**
preRatings_mean	0.7934	0.063	12.625	0.000***	0.67	0.917
genres_Adventure	-0.2696	0.37	-0.729	0.466	-0.996	0.457
genres_Animation	-1.6172	0.475	-3.406	0.001***	-2.55	-0.685
genres_Children	0.754	0.919	0.82	0.412	-1.051	2.559
genres_Comedy	-0.4373	0.361	-1.213	0.226	-1.146	0.271
genres_Crime	0.2779	0.453	0.613	0.54	-0.612	1.168
genres_Documentary	-1.8538	0.416	-4.459	0.000***	-2.67	-1.037
genres_Drama	-0.0115	0.28	-0.041	0.967	-0.562	0.539
genres_Fantasy	-0.5296	1.381	-0.384	0.701	-3.241	2.182
genres_Horror	-0.0726	1.68	-0.043	0.966	-3.373	3.228
genres_Mystery	-0.7721	1.384	-0.558	0.577	-3.49	1.946
genres_Sci-Fi	1.6242	1.205	1.348	0.178	-0.743	3.991
genres_Thriller	-0.7968	1.204	-0.662	0.508	-3.161	1.568
genres_War	-2.0295	2.371	-0.856	0.392	-6.686	2.627
genres_Western	3.3738	1.686	2.001	0.046**	0.062	6.686

(Note: ** means that it's significant at 5% confidence level, *** means significant at 1% level)

Ratings still significantly affect the number of the Oscar Awards. In this case, genres have no significant effect on the number of awards. Children's movies and Documentaries

have fewer awards than other genres' movies, while Western movies tend to have more awards.

Furthermore, combining `the_oscar_award.csv`, `movies.csv` and `movie_industry.csv`, we can add 'rating' variable viewership ratings in IMDB, not the ratings in MovieLens). We one-hot encoded it and used dummy variables in the model. To avoid multicolinearity, we dropped out the one-hot features for PG-13'. We used 'Oscar winner' as the target. Here are the results:

Table 6: Oscar Winner - Pre-ratings, with Control Variables 2

	coef	**std err**	**t**	**P > |t|**	**[0.025**	**0.975]**
preRatings_mean	0.0837	0.016	5.203	0.000***	0.052	0.115
genre_Adventure	-0.0415	0.096	-0.431	0.667	-0.231	0.148
genre_Animation	-0.0277	0.124	-0.224	0.823	-0.271	0.216
genre_Biography	0.0592	0.082	0.718	0.473	-0.103	0.221
genre_Comedy	-0.0505	0.086	-0.591	0.555	-0.219	0.118
genre_Crime	0.0442	0.103	0.431	0.667	-0.157	0.246
genre_Drama	0.0505	0.075	0.676	0.499	-0.096	0.197
genre_Fantasy	0.2075	0.329	0.63	0.529	-0.44	0.855
genre_Mystery	-0.3686	0.463	-0.797	0.426	-1.278	0.541
genre_Thriller	0.6738	0.462	1.459	0.145	-0.234	1.581
rating_G	0.0766	0.156	0.492	0.623	-0.229	0.383
rating_NOT RATED	-0.2711	0.461	-0.588	0.557	-1.178	0.636
rating_PG	-0.057	0.091	-0.626	0.532	-0.236	0.122
rating_R	0.002	0.054	0.038	0.97	-0.104	0.108

(Note: *** means that it's significant at 1% confidence level)

Here are the results using 'num of awards' as the target:

Table 7: Num of Awards - Pre-ratings, with Control Variables 2

	coef	**std err**	**t**	**P > |t|**	**[0.025**	**0.975]**
preRatings_mean	0.9118	0.095	9.575	0.00***	0.725	1.099
genre_Adventure	0.0883	0.569	0.155	0.877	-1.031	1.207
genre_Animation	-1.6082	0.733	-2.193	0.029**	-3.049	-0.167
genre_Biography	0.1419	0.488	0.291	0.771	-0.816	1.1
genre_Comedy	-0.6428	0.506	-1.271	0.205	-1.637	0.352
genre_Crime	0.1807	0.607	0.298	0.766	-1.012	1.373
genre_Drama	0.0737	0.442	0.167	0.868	-0.795	0.942
genre_Fantasy	-0.8944	1.949	-0.459	0.647	-4.726	2.937
genre_Mystery	-1.7234	2.737	-0.63	0.529	-7.104	3.657
genre_Thriller	0.7379	2.732	0.27	0.787	-4.632	6.108
rating_G	0.4232	0.921	0.459	0.646	-1.387	2.234
rating_NOT RATED	-1.859	2.73	-0.681	0.496	-7.225	3.507
rating_PG	-0.2973	0.539	-0.552	0.581	-1.356	0.761
rating_R	-0.2681	0.319	-0.84	0.401	-0.896	0.359

(Note: ** means that it's significant at 5% confidence level, *** means significant at 1% level)

Ratings still have a significant effect on the Oscar Awards. Therefore, we can conclude

that the influence of ratings are really stable.

Additionally, to test the effect of dummy variables further, we use the interaction item (dummy variables * 'preRatings_mean') to see whether they cause a structural change in the ratings' influence. Here are the results using 'winner' as the target.:

Table 8: Dummy Variable Test - Interaction Iterms

	coef	**std err**	**t**	**P > |t|**	**[0.025**	**0.975]**
preRatings_mean	0.0635	0.016	3.857	0.000***	0.031	0.096
genre_Adventure*preRatings_mean	0.0081	0.026	0.311	0.756	-0.043	0.059
genre_Animation*preRatings_mean	0.0099	0.034	0.292	0.771	-0.057	0.077
genre_Biography*preRatings_mean	0.0331	0.022	1.483	0.139	-0.011	0.077
genre_Comedy*preRatings_mean	0.007	0.023	0.303	0.762	-0.039	0.053
genre_Crime*preRatings_mean	0.0265	0.027	0.977	0.329	-0.027	0.08
genre_Drama*preRatings_mean	0.0326	0.02	1.606	0.109	-0.007	0.072
genre_Fantasy*preRatings_mean	0.0765	0.095	0.808	0.42	-0.11	0.263
genre_Mystery*preRatings_mean	-0.0705	0.106	-0.667	0.505	-0.278	0.137
genre_Thriller*preRatings_mean	0.1878	0.119	1.577	0.116	-0.046	0.422
rating_G*preRatings_mean	0.0328	0.043	0.766	0.444	-0.051	0.117
rating_NOT RATED*preRatings_mean	-0.0706	0.12	-0.589	0.556	-0.306	0.165
rating_PG*preRatings_mean	-0.0114	0.024	-0.472	0.637	-0.059	0.036
rating_R*preRatings_mean	0.007	0.014	0.482	0.63	-0.021	0.035

(Note: *** means that it's significant at 1% confidence level)

There is no significant effect from the interaction effect, meaning that these dummy variables do not change the influence of ratings.

Summary: Ratings' effect on Oscar Award winners is substantial, and this effect is stable enough. Hence, we can safely say that higher ratings mean a higher likelihood of receiving the Oscar Award. As we mentioned before, ratings are the best indicator for consumer preference, and we conclude that consumer preference has a positive effect on Oscar Award winners.

Does consumer preference increase after a movie receives the Oscar Awards?
Similarly, we mainly use ratings ('postRatigns_mean') and Oscar Awards data in the linear regression models. Also, due to the high similarity between 'postRatigns_mean' and 'postRatigns_mode', we only show the results for 'postRatigns_mean' below.

The table below displays the influence of 'Oscar winner' on 'postRatings_mean' (all these variables are described above):

Table 9: Post-ratings and Oscar Winner

	coef	**std err**	**t**	**P > |t|**	**[0.025**	**0.975]**
Oscar winner	3.5954	0.111	32.251	0.000***	3.377	3.814

(Note: *** means that it's significant at 1% confidence level)

It shows that winning an Oscar has a strong positive effect on the ratings afterwards. However, the ratings before the Oscar Award ceremony is also highly correlated to the ratings afterwards (corralation higher than 0.65). In the previous section, we showed that pre-ratings have a stable and significant effect on Oscar Award winners. Thus, there may exist endogenous problems if we use post-ratings as the dependent variable.

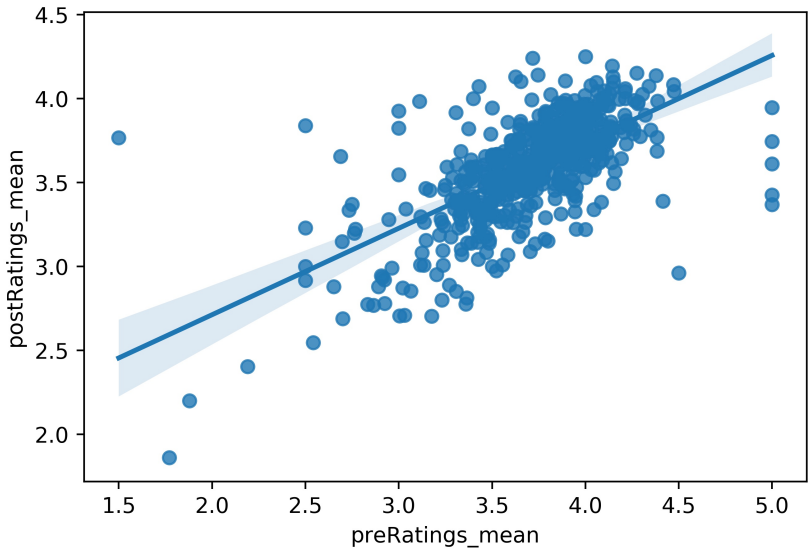

Figure 11: Correlation between Pre-ratings and Post-ratings is over 0.65.

In order to avoid the endogenous problems, we construct the ratings difference (*postRatings_mean*−*preRatings_mean*) as the target variable, and explore whether ratings are influenced by a movie's reception of an Oscar. Described here are rating differences:

Table 10: Description for Rating Difference

	Ratings Difference
count	586
mean	-0.11628
std	0.300631
min	-1.63054
25%	-0.25826
50%	-0.10992
75%	0.002521
max	2.267857

We performed a regression analysis using 'Oscar winner' as the independent variable:

Table 11: Rating Difference and Oscar Winner

	coef	**std err**	**t**	**P > |t|**	**[0.025**	**0.975]**
Oscar winner	-0.1897	0.023	-8.17	0.000***	-0.235	-0.144

(Note: *** means that it's significant at 1% confidence level)

And now we see that Oscar winner has a negative effect on rating difference. If a movie wins the Oscar awards, the rating difference will be lower, which means that post-ratings are more consistent to the pre-ratings. Given the negative rating difference, more consistent ratings mean higher post-ratings. This supports the claim that the reception of an Oscar has a positive effect on post ratings.

We also control the genre and viewership rating, with one-hot encoding and dropout, similar to the analytic before. 'votes', 'gross', 'score', which mentioned above as the indicators for consumer preference, has a high correlation with 'Oscar winner', therefore they cannot be used as control variable. Here's the result:

Table 12: Rating Difference and Oscar Winner, with Control Variables

	coef	**std err**	**t**	**P > |t|**	**[0.025**	**0.975]**
genres_Adventure	0.0009	0.048	0.02	0.984	-0.093	0.095
genres_Animation	-0.0039	0.081	-0.049	0.961	-0.163	0.156
genres_Children	0.1212	0.141	0.86	0.39	-0.156	0.398
genres_Comedy	-0.1809	0.038	-4.704	0.000***	-0.256	-0.105
genres_Crime	-0.0422	0.052	-0.804	0.422	-0.145	0.061
genres_Drama	-0.0544	0.028	-1.918	0.056*	-0.11	0.001
genres_Fantasy	0.1322	0.158	0.837	0.403	-0.178	0.443
genres_Horror	0.0227	0.269	0.084	0.933	-0.507	0.552
genres_Mystery	-0.1078	0.192	-0.563	0.574	-0.484	0.269
genres_Sci-Fi	0.048	0.156	0.308	0.758	-0.258	0.354
genres_Thriller	0.071	0.192	0.371	0.711	-0.306	0.448
genres_Western	-0.0484	0.19	-0.254	0.799	-0.422	0.325
rating_G	-0.1748	0.079	-2.209	0.028**	-0.33	-0.019
rating_NOT RATED	0.1794	0.27	0.663	0.508	-0.352	0.711
rating_PG	-0.1337	0.044	-3.063	0.002***	-0.219	-0.048
rating_R	-0.0177	0.029	-0.606	0.545	-0.075	0.04
Oscar winner	-0.1296	0.027	-4.763	0.000***	-0.183	-0.076

(Note:*: significant at 1% confidence level, **: significant at 5% level, ***: significant at 1% level)

We can see that Oscar winner still have negative effects on rating difference, presenting a positive effect on post-ratings. Comedy movies, G-rated and PG-rated movies also tend to have a lower rating difference.

Summary: Oscar Award winner will negatively influence the rating difference, the different between post-ratings and pre-ratings. Given the negative rating difference, it can

increase the post-ratings significantly. This effect is stable, even with multiple control variables.

Now, in the light of statistical analysis, we can say that consumer preference and Oscar awards can influence each other, and these influences are significant and stable. However, we mentioned that linear regression and statistical analysis are actually not suitable to the classification problems prediction. To further explore the influence of consumer preference on the prediction, we need to use other models specializing in classification.

4 Oscar Award Nominees Prediction

In this section, we try to solve the classification problem more accurately and see the roles of consumer's preference. To utilize the available datasets, We now are curious about whether we could predict the movies that will be nominated for the Oscar awards. Besides the aforementioned consumer preference indicators, i.e. gross, votes, scores/ratings, we want to track better whether there is a change in the consumer's preference over year and the effect on the nomination of the awards. Utilizing movie industry data, genome tag data from movie lens data and customer preference data, we have performed classification supervised learning tasks.

4.1 Methodology

First, we have performed data cleaning and wrangling. For each movie, we have only chosen the top 50 most relevant tags. As there were a total of around 1200 tags for each movie, we had to limit the number of tags for each movie and we chose 50 as our cutoff point using the elbow method.

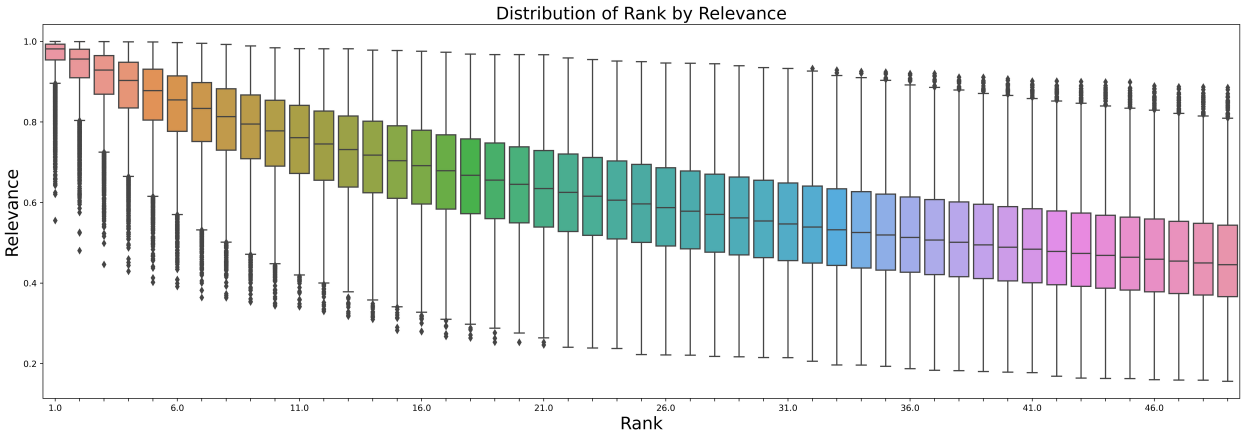

Figure 12: Distribution of rank by relevance.

For the proof of the concept, we have chosen *Toy Story* as my example and chose a top 50 tag, and they were *toys, computer animation, kids and family, cartoon, children, and adventure*. Using only top 50 tags, we have calculated the cosine similarities between the movies. Some of the movies that were similar to *Toy Story* were *Bug's Life, Monster's Inc, Ice Age, Finding Nemo and Ratatouille*. This proves that the top 50 most relevant tags very well represent the genre and sentiment of the movies.

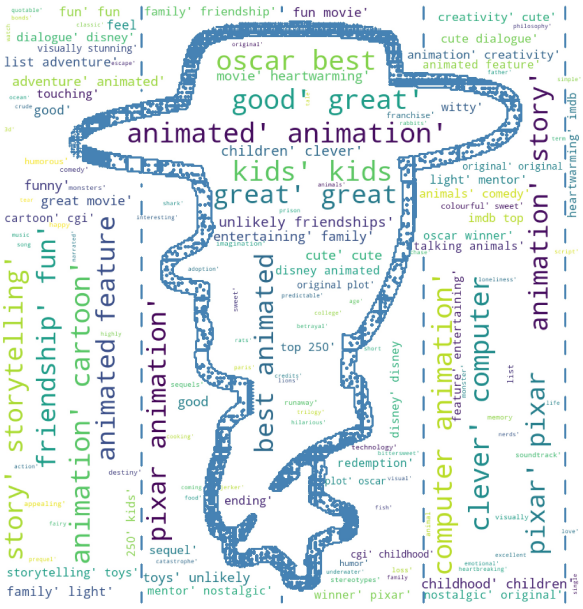

Figure 13: Word-cloud plot for movies similar to Toy Story.

4.2 Quantification of Consumer Preference

And we construct a powerful tool to quantify the consumers' preference by timeline. The `ratings.csv` from `movie_lense` dataset has the ratings of users on movie with the time they rated movie. With the abundance of the data (27 million rows, 28 thousands users), we thought this would be a great point to quantify the consumers' preference.

We have sampled 2000 users from each year (1996-2018) and joined the genre of the movie they have watched. The genre of a movie is one-hot encoded and marked with the value of rating of the movie. Each user's preference is marked by the average of the ratings in each genre (total of 19 genres). In order to classify each user's preference into smaller categories, we have used a clustering algorithm and the below elbow method depicts that either 2 or 3 clusters is a good choice. We have used TSNE to reduce the dimensions for visualization purposes.

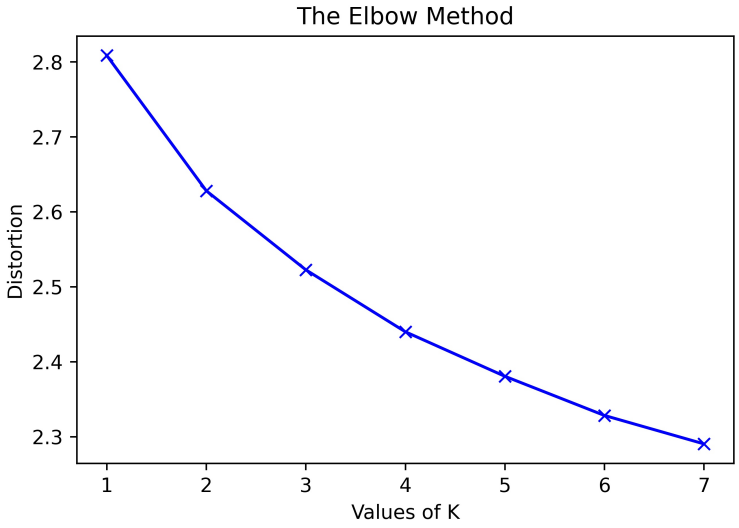

Figure 14: The elbow method.

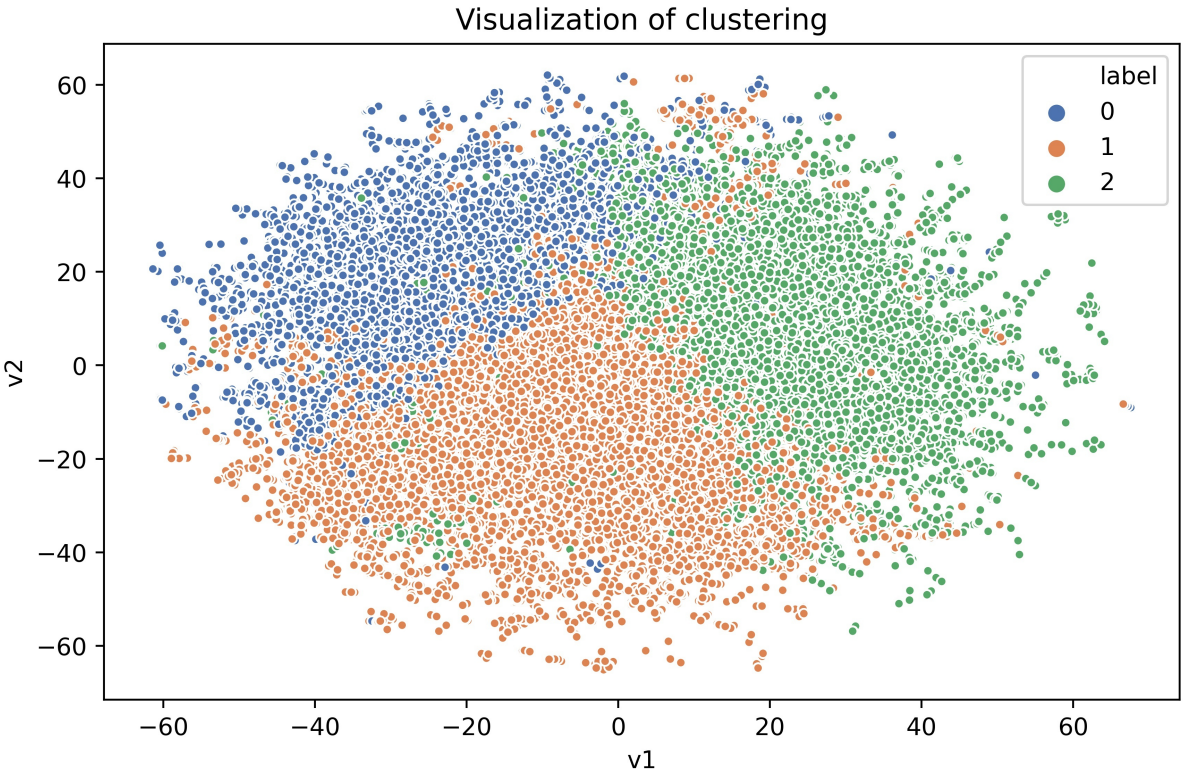

Figure 15: Visualization of clustering.

Despite the fact that it is not perfectly separated, there is a certain distinction or boundary among clusters, considering that clusters are formed with 19 features not 2. Instead of looking for certain boundaries between clusters, we have decided to investigate the center of clusters to find characteristics of each label. Below is the data frame of cluster centers.

label	Action	Adventure	Animation	Children	Comedy	Crime	Documentary	Drama	Fantasy	Film-Noir	Horror	IMAX	Musical	Mystery	Romance	Sci-Fi	Thriller	War	Western
0	2.318116	1.53786	0.176014	0.189996	0.682294	0.791328	0.015086	1.10658	0.398197	0.040168	0.303035	0.284956	0.055764	0.417478	0.341753	1.363494	1.714529	0.16091	0.065177
1	0.722564	0.864279	0.38296	0.502152	1.806984	0.415794	0.040666	1.036229	0.540045	0.024475	0.232032	0.104303	0.224842	0.198556	0.741994	0.443343	0.585469	0.128915	0.062373
2	0.634009	0.409535	0.106349	0.153458	0.973061	0.798541	0.037882	2.835228	0.223682	0.048533	0.175238	0.087395	0.108368	0.341683	0.908992	0.321644	0.90321	0.364559	0.069928

Figure 16: Cluster centers.

If we investigate a genre with max rating, we have Action, Comedy, Drama for label 0,1,2 respectively. However, the genre of movie is composed of multiple genres, instead of single, we believe the label represents the consumer's preference on certain movies. Even though we cannot perfectly interpret the meaning of each cluster, if we can find the implication of this preference, we think it should be the meaningful implication.

After training the clustering algorithm with the whole data, we have labeled each user in each time line and visualize it. It is clear that consumers' preferences change overtime. Therefore, we have decided to use it to predict the nominee of Oscar award including the features and analyze the impact of the consumer preference indicators.

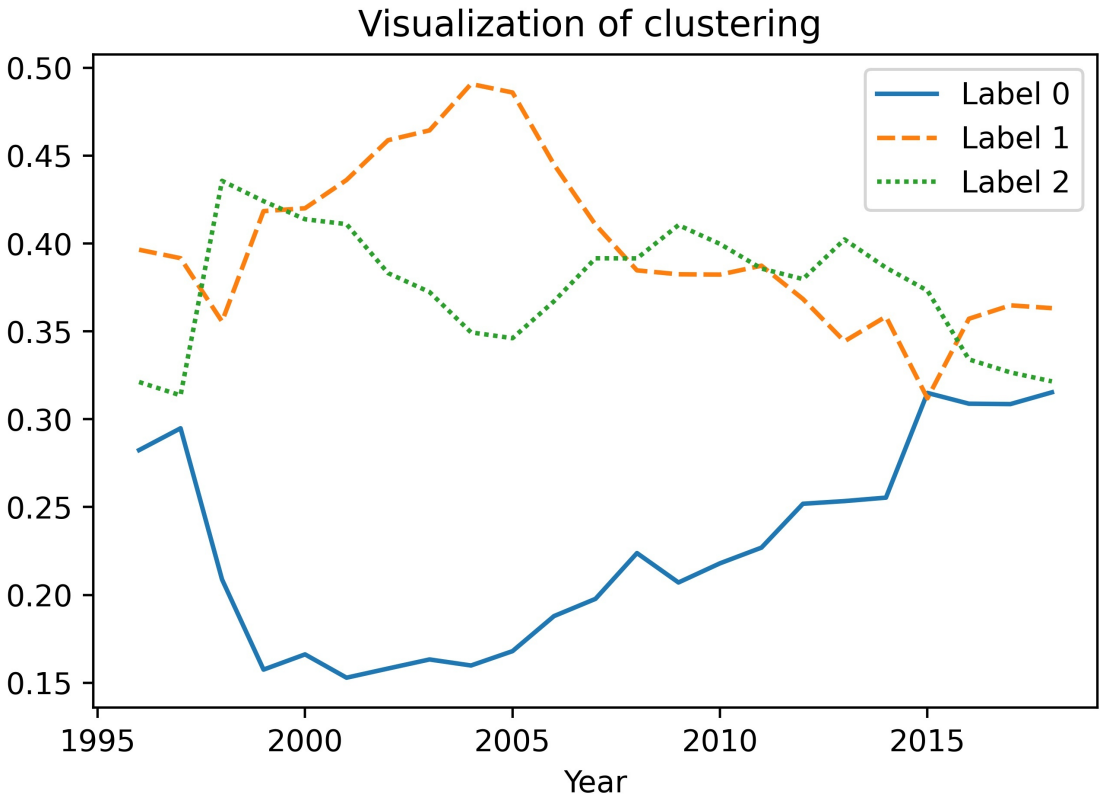

Figure 17: Frequency changes in labels.

Finally, we have dropped columns such as company, country, director, year and title that are less relevant to classifications. Also, to prevent the look-ahead bias, we have removed the Oscar related genome tags from the final dataframe. On the other hand, we have imputed the missing budget by calculating the average profit-loss conversion rate multiplying the conversion ratios to revenue of budget-missing movies. Return-on-Investment (ROI) variable was also added here. Next, we have encoded categorical variables by one-hot encoding and normalized votes, budget, revenue with min-max scaler and score and ROI by standard scalers. After all the imputation and feature selection and extraction, we had to go through one more step before we test ML algorithms. As this was a severely imbalanced data set (total of 492 movies that were nominated and a total of 2305 movies that were not nominated in the Training set), we implemented SMOTE, an oversampling algorithm which populates new data points using nearest neighbors algorithm and creates a balanced data set.

After running various algorithms, including logistic regression, nearest neighbor, naive_bayes, Tree models, support vector machines neural network, the random forest algorithm seemed to perform the best. After cross validation, the model yielded around 90% accuracy rates. However, since this was a severely imbalanced data set, we need to look into precision score, which shows the True Positive prediction rate. The random forest yielded about 50% for the precision prediction.

To interpret the results, this means that our model was able to accurately predict whether a movie will make it to the Oscars nomination by 90% in general, and out of the movies that the model predicted to be nominated, about half of them were actually nominated. As the nomination to Oscar award is deemed as a successful movie, the model being able to predict the result by 90% was very fascinating.

We took a further look at the result and analyzed the feature importance of the model, to analyze the importance of consumer preference. The below is a chart of the top 20 important features. We could see that monetary success, as a good indicator for consumer's preference is a great indicator of the Oscar nomination as we could see how important revenue, ROI, etc. are. Votes are key factors as well. Also scores/ratings are quite important to the prediction. They are also reflecting the consumer's preference Some of the genome tags gave us great hints such as great acting, dramatic, drama, story and emotional. We could tell the genres and the time-varying customer preference were not as important as the sentiment of the movie. Although the constructed time-varying preference isn't powerful enough, the monetary success, votes and scores/ratings still show that consumer preference is a powerful factor to predict the Oscar awards, consistent with the statistical analysis results in the previous section.

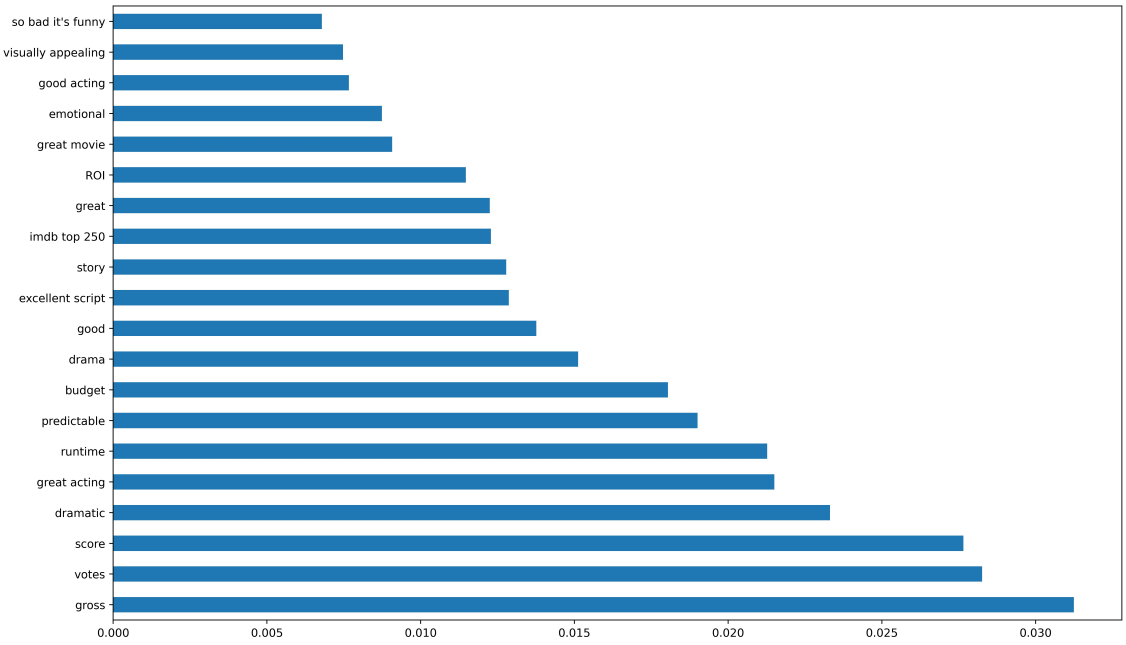

Figure 18: Frequency changes in labels.

5 Conclusion

From the analysis above, we can safely answer the 2 questions we want solve:

1. Consumer preference does influence the Oscar awards, no matter whether we choose scores/ratings to represent the consumer preference, and whether we choose Oscar awards winners/nominees as our objective. This effect are significant and stable, shown from both statistical models and machine learning algorithms.

2. Oscar Awards winner will have a narrower difference between pre-ratings and post-rartings, and therefore a higher ratings afterwards give the negative difference.

So we can conclude that committee may consider the overall preference in the market as one of the cafeteria; and people will follow the Oscar Award, and watch and give a higher credit to the winners.

For the future analysis, we could try to separate the Oscar awards category, and try to predict separately. The dataset will be extremely biased, but some of the categories have very different meanings than the rest of the awards. Even though great make-up and sound

tracks are required for the great movie, often the best picture awarded movie tend to not win the best make-up award or the best-sound track awards. Therefore, making a more granular prediction model will boost the overall accuracy of the model and help us find something more interesting.

6 Appendix

We made further investigation into the dataset in `movie_industry.csv`. Here are our results.

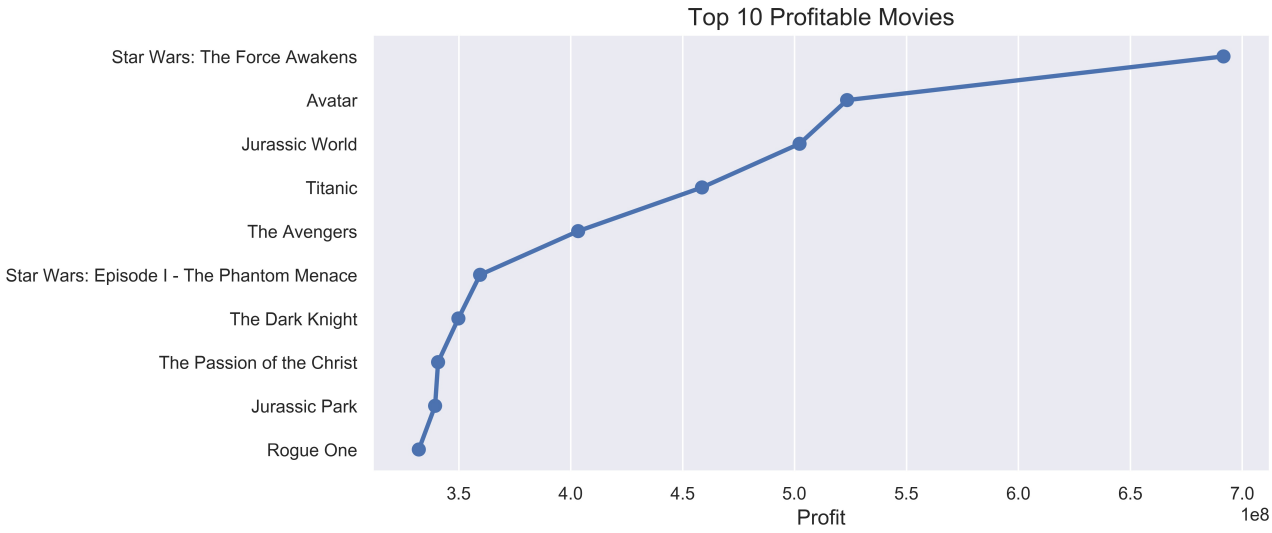

Figure 19: Top 10 profitable movies.

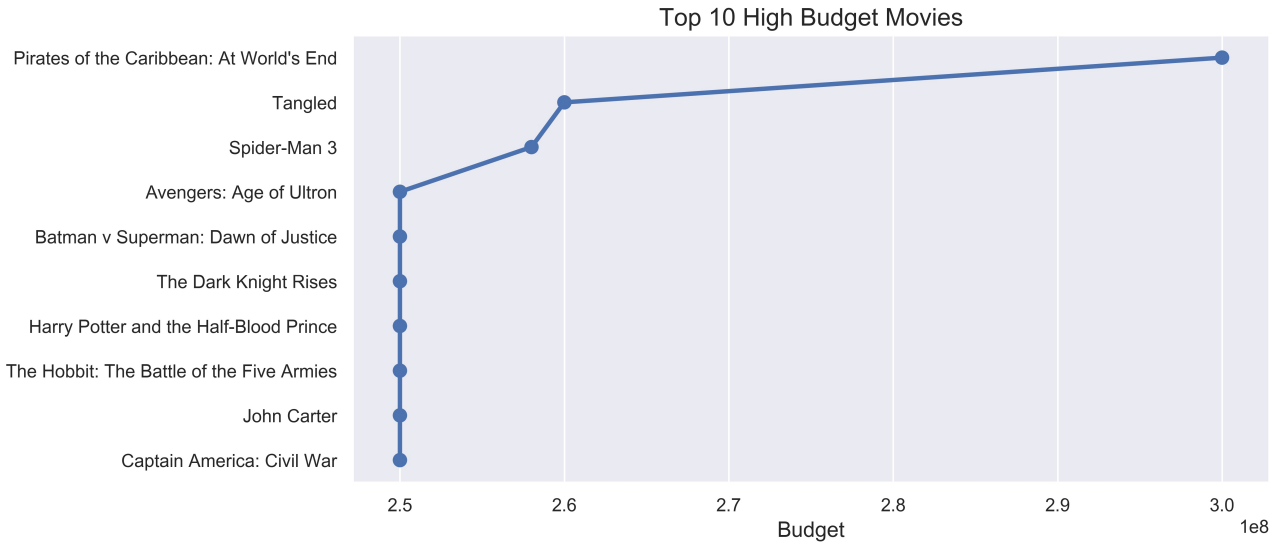

Figure 20: Top 10 high budget movies.

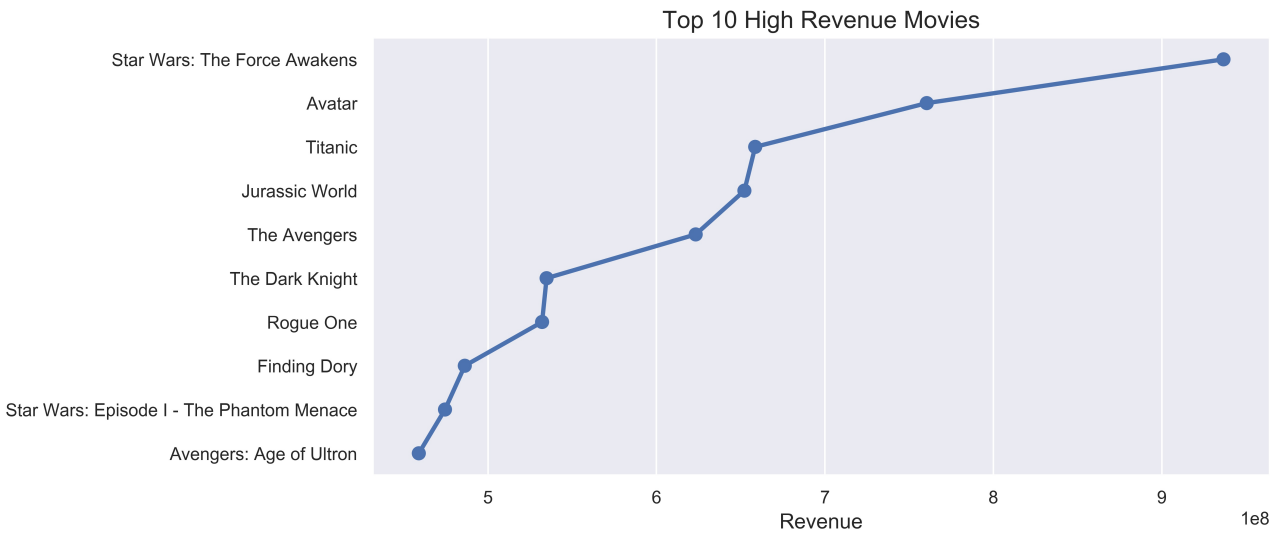

Figure 21: Top 10 high revenue movies.

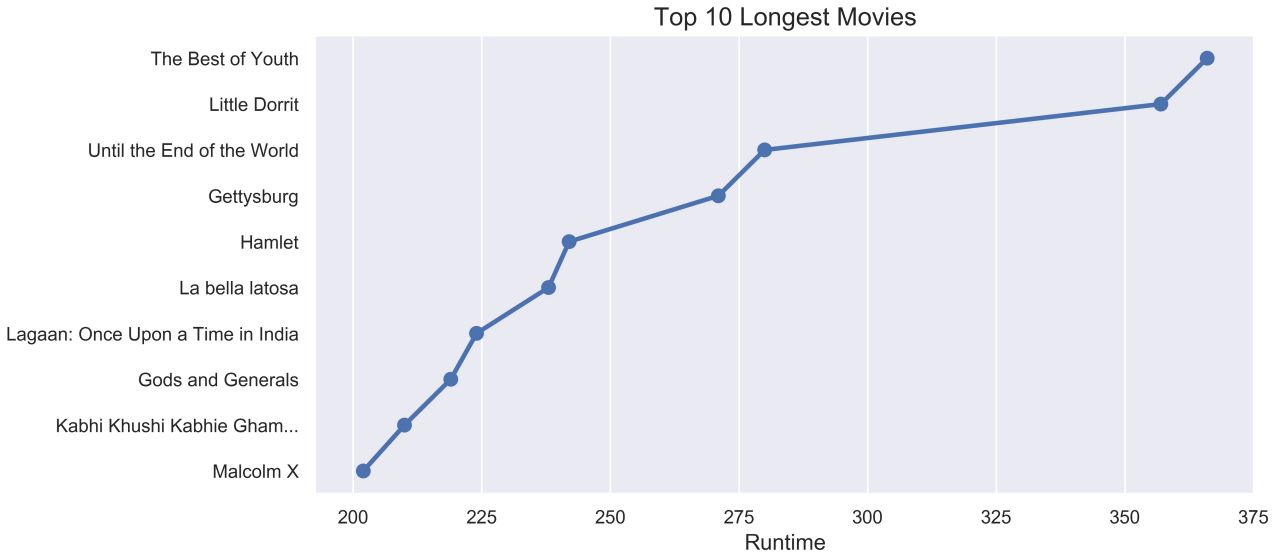

Figure 22: Top 10 longest movies.

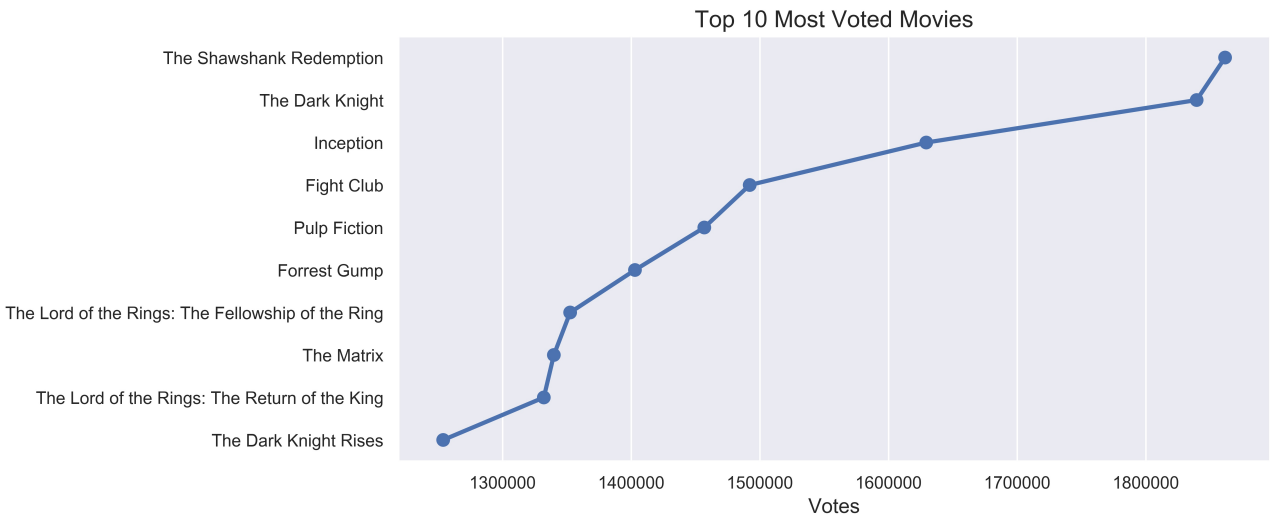

Figure 23: Top 10 most voted movies.

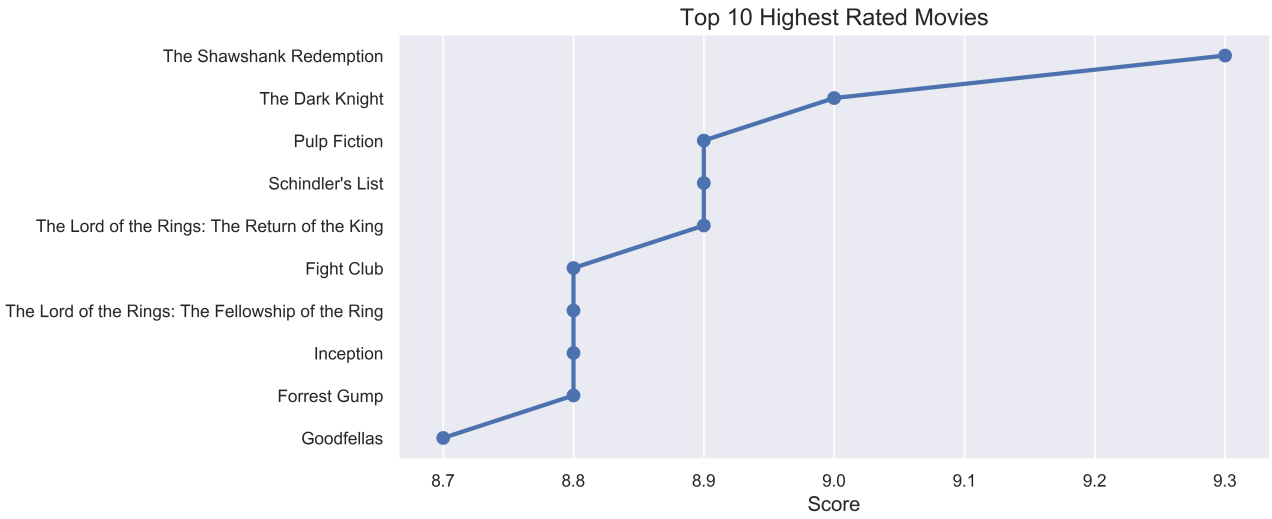

Figure 24: Top 10 most rated movies.

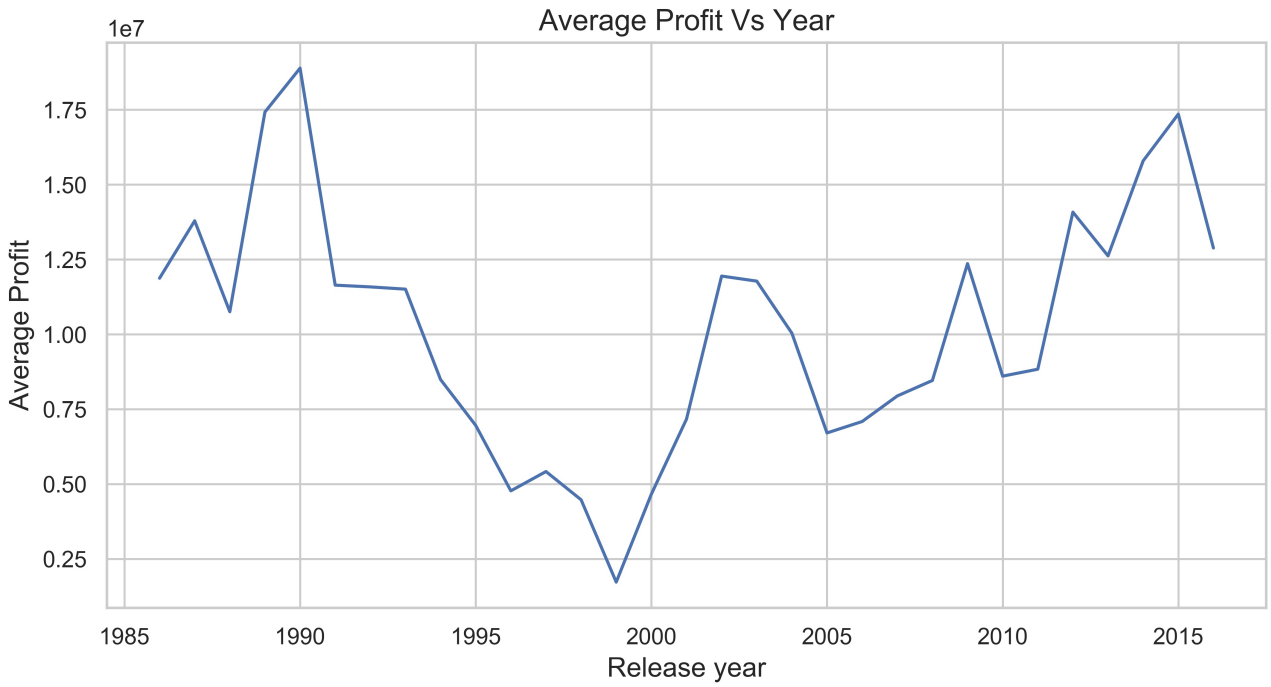

Figure 25: Average profit per year.

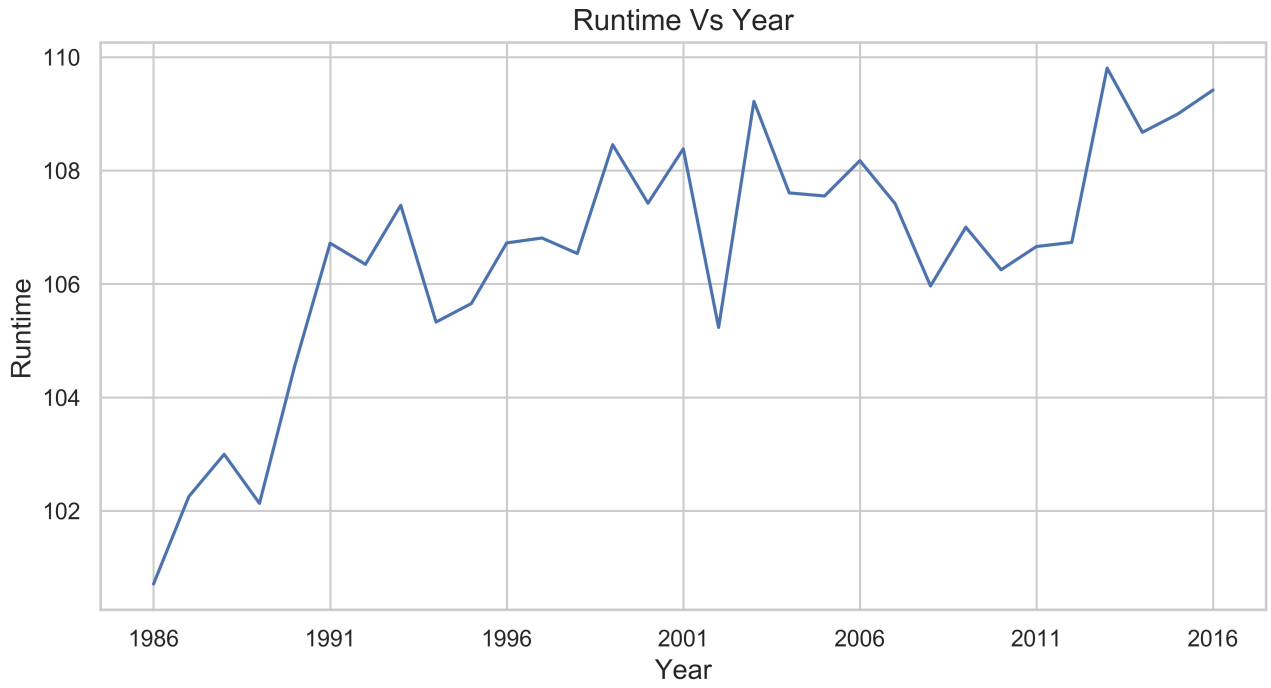

Figure 26: Average runtime of movies from year to year.

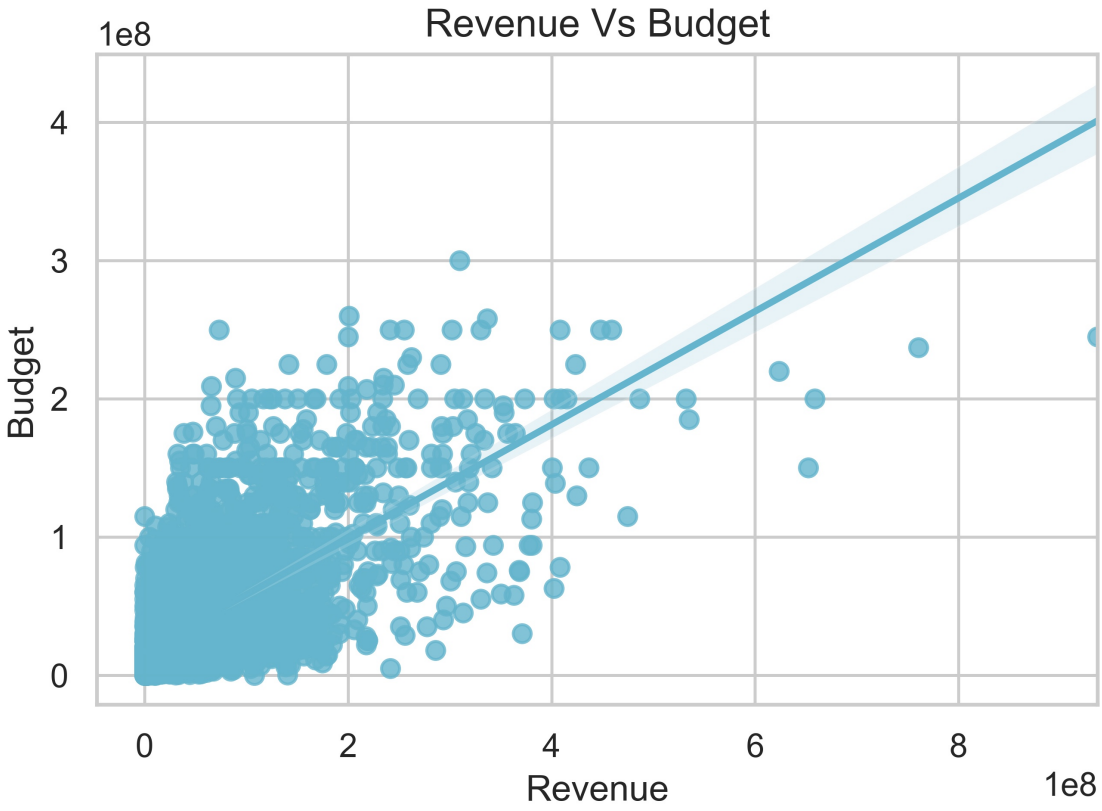

Figure 27: Correlation between revenue and budget: 0.68.

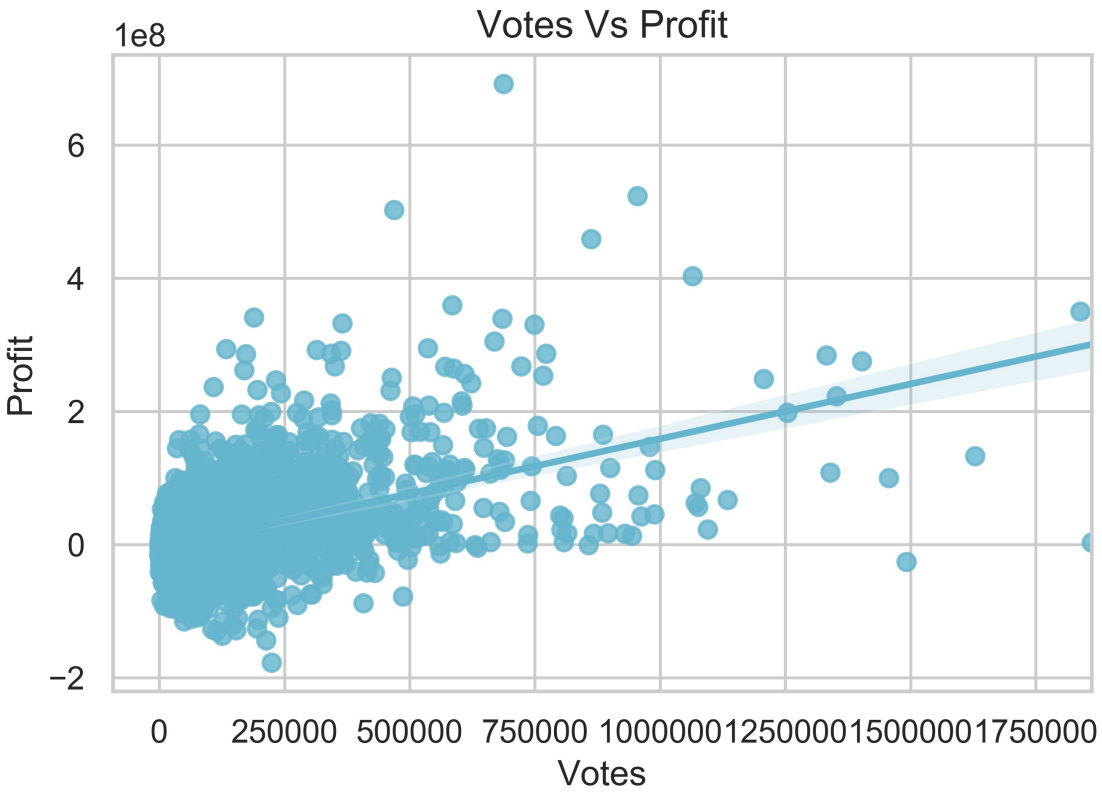

Figure 28: Correlation between votes and profit: 0.50.

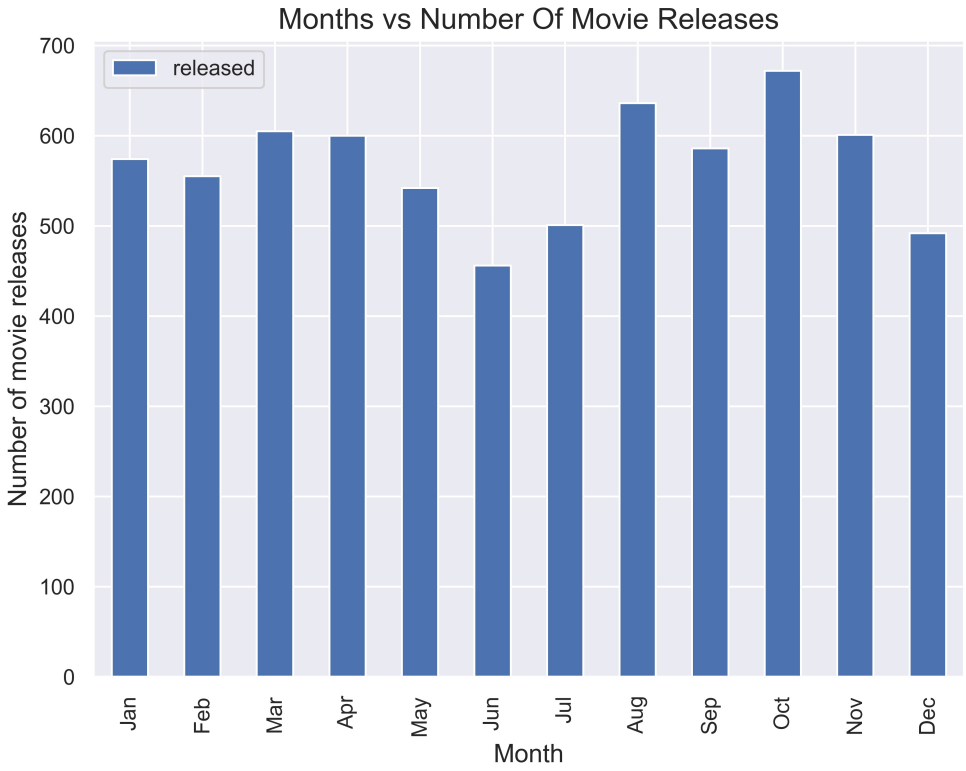

Figure 29: Number of movies released vs month.

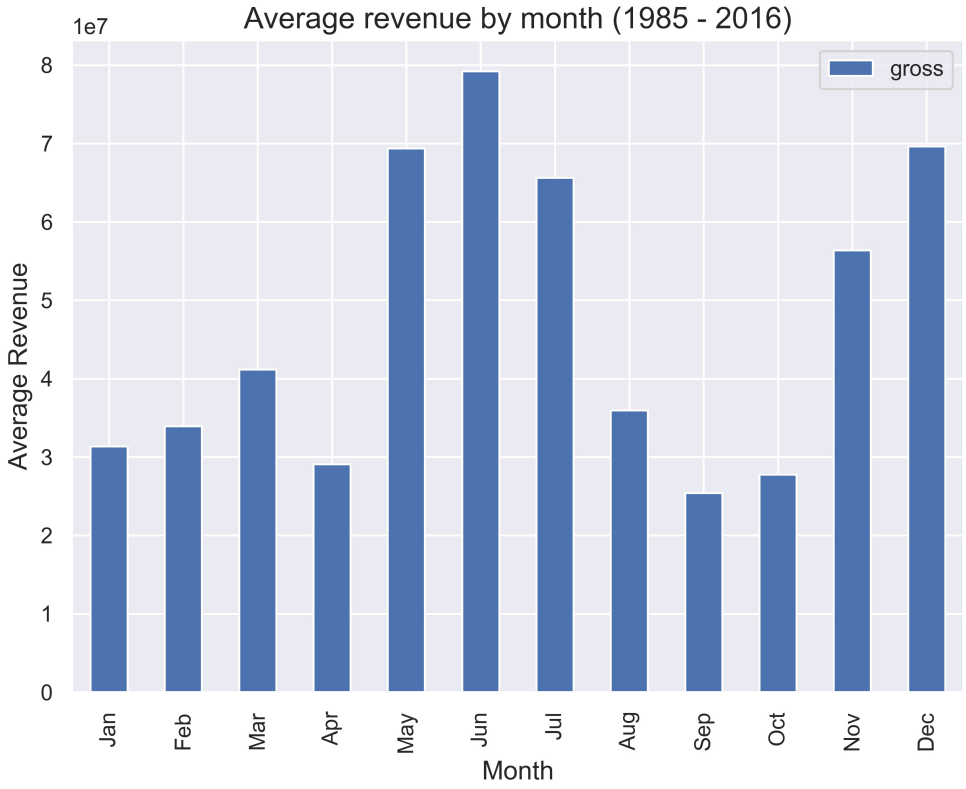

Figure 30: Average revenue per month.

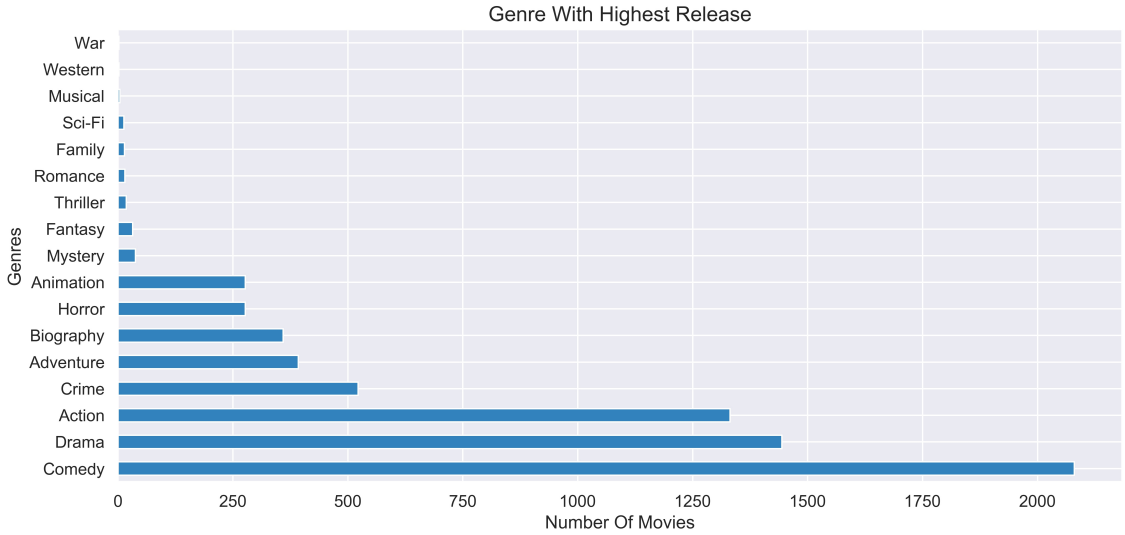

Figure 31: Genre with highest releases.

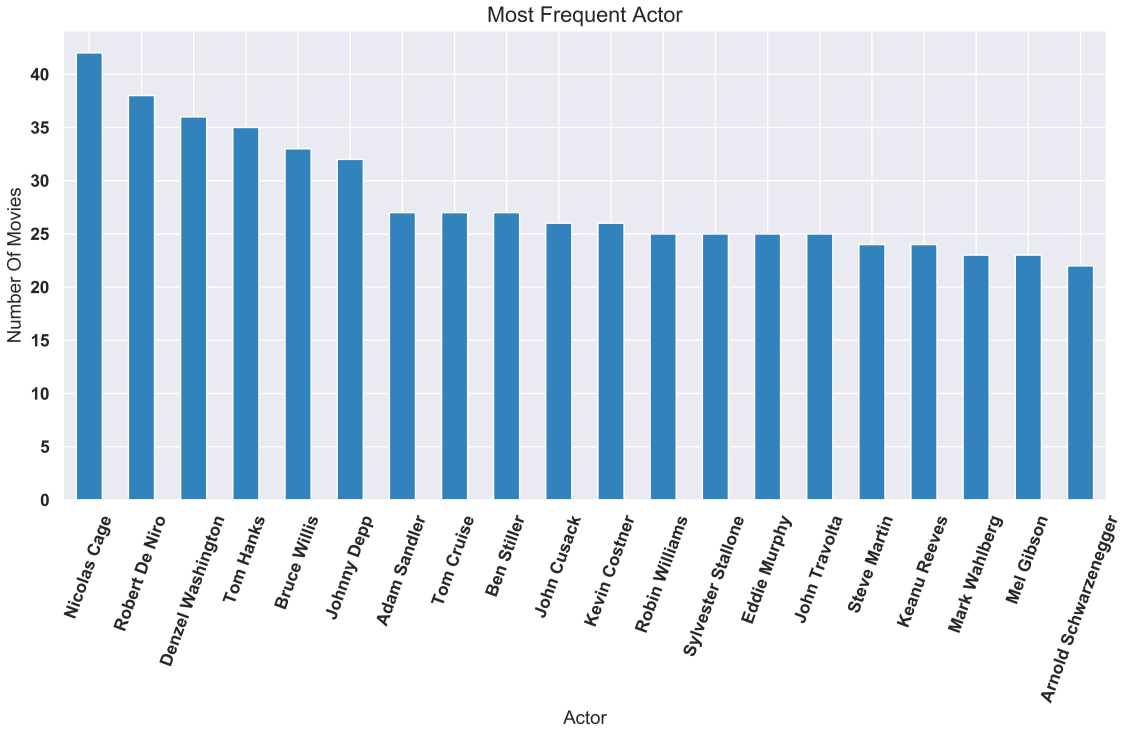

Figure 32: Most frequent actors.

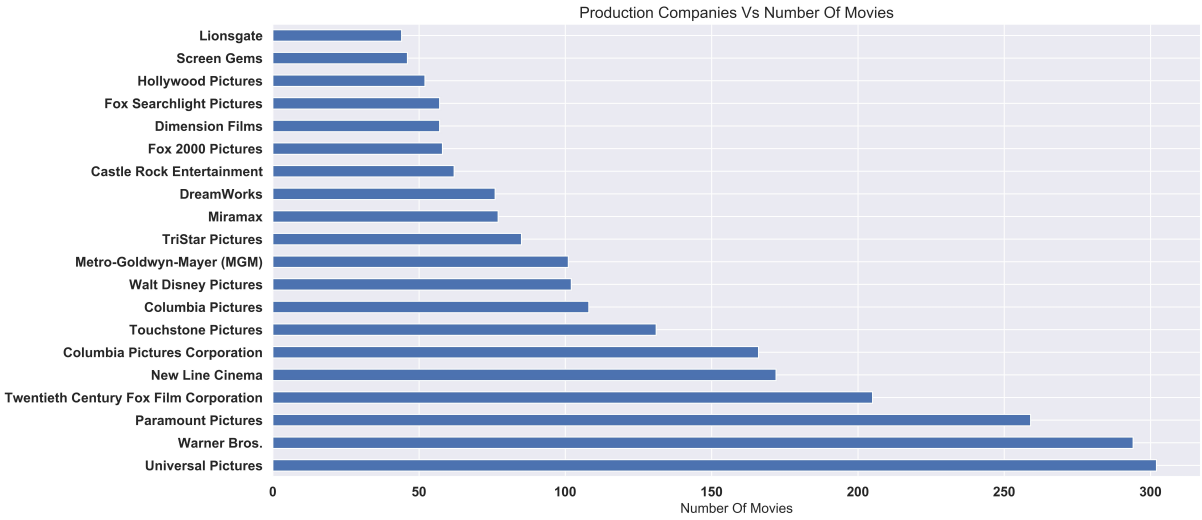

Figure 33: Top 20 production companies with higher number of release.

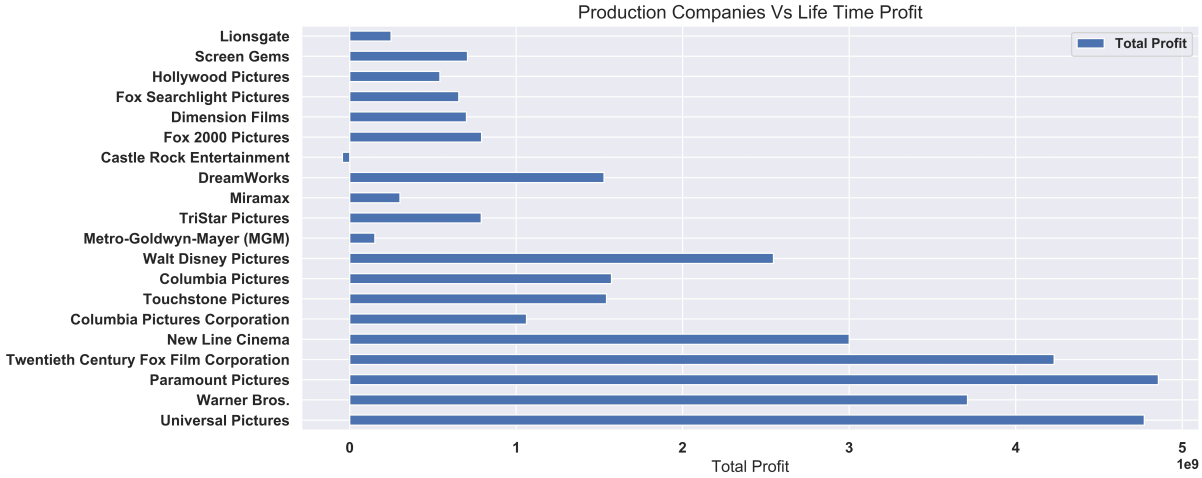

Figure 34: Life time profit earn by each production company.

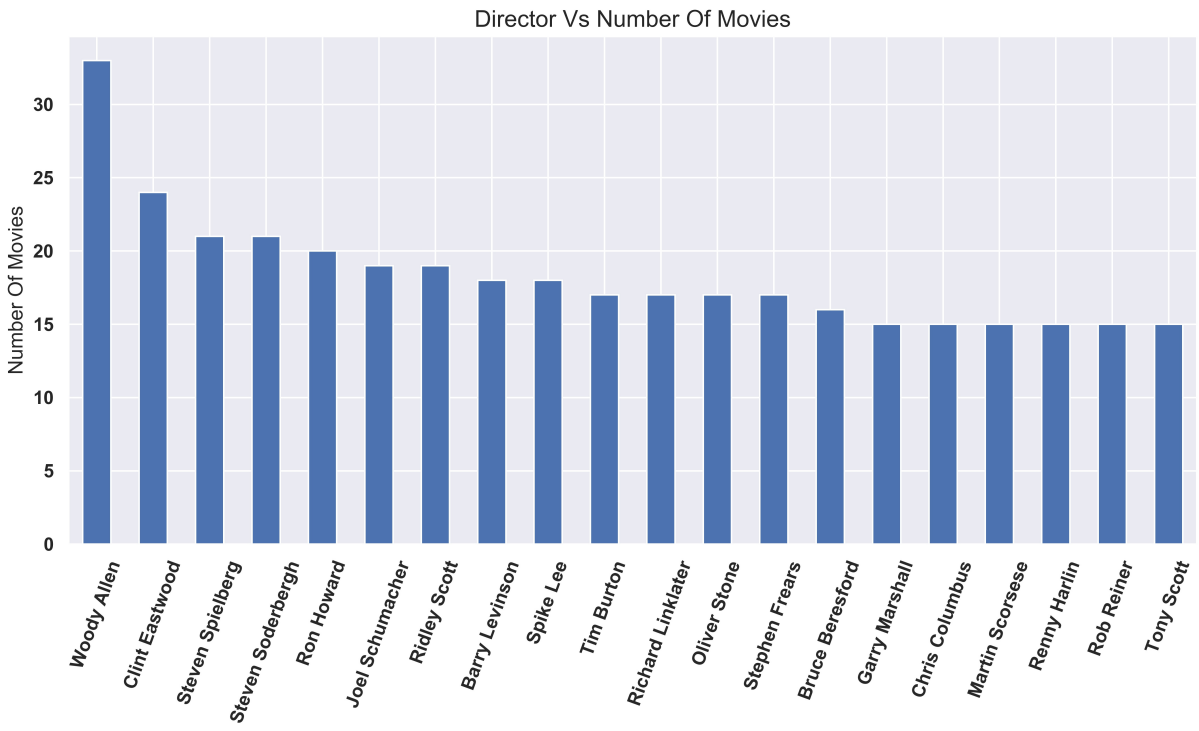

Figure 35: Top 20 directors who direct maximum movies.