CSDS 133 Final Project: Predicting Oscar Winners Based on Associated Variables

Kelly Gorman, Ravi Corrie, Ana Gómez Corral, Cilla Nee

Abstract

In this project, we attempted to determine a variable that could be used to predict the nominated film most likely to win the Best Picture category of the Oscars. We examined several potential predictors, including the films' revenues, budgets, performances at other award shows, critic and audience scores, and number of nominations in other Oscar categories. We found that revenue, budget, audience score, and critic score could not be used as predictors of being the Oscar winner. Variables that could be used as predictors of Oscar winners included the films' performances at other award shows and the films' number of nominations in other Oscar categories. The strongest predictor of the nominee being the winner, based on our analysis, was the films' performances at other award shows. Thus, if we were to create a predictive model for future Best Picture nominees, we would likely use this variable.

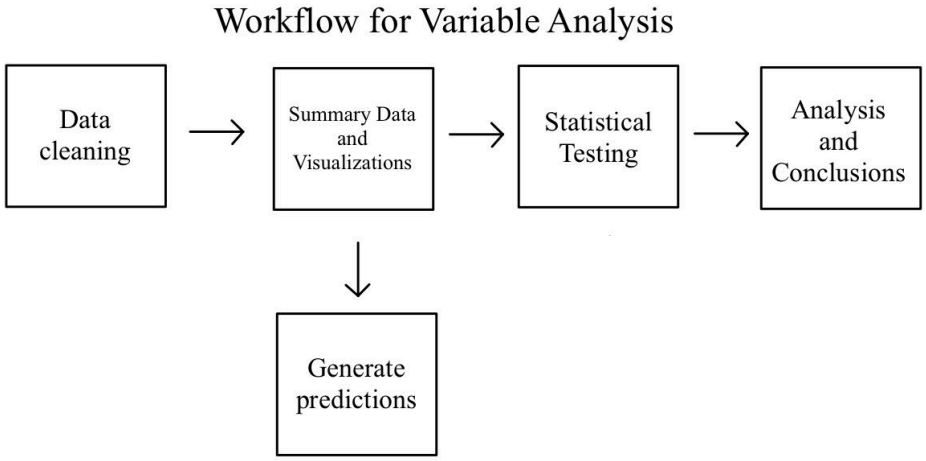
1. Introduction

Award shows tend to seem random in nature, rewarding films with the intangible "best" qualities. In this project, we attempted to determine if there is a reliable variable that can be used to quantifiably determine a given nominee in the Best Picture category's likelihood of winning. First, we defined several variables that could potentially have a correlation with being the winning nominee, including revenue (international and domestic), budget, performance at other award shows, critic and audience scores, and the number of non-Best Picture Oscar nominations the film received. We utilized the availability of data relating to our variables, found on the website Kaggle. We then cleaned this data and used it to perform statistical testing to draw conclusions about correlation with winning status. These statistical tests included t-tests and Pearson correlations to determine variables' associations. We also accounted for the possibility that our variables could influence each other, as the variables we observed were all related to the same set of Oscar nominated films and did not exist in a vacuum. We attempted to gain insights from these comparisons, considering the possibilities for correlation, or lack of correlation, between variables.

2. Methods

We conducted our analyses individually, each focusing on a specific variable to determine its capacity to serve as a predictor. Our processes differed slightly by datasets, but generally, we began by creating a comprehensive dataset for our variable by merging Oscar-specific datasets with data related to our variable. All individual datasets were found on the website Kaggle. Then, we cleaned our datasets in order to find general summary data, such as mean and standard deviation. We used this information and assumptions related to our variables in order to generate hypotheses, and we created visualizations for our variables using our cleaned datasets. We analyzed our visualizations and used our datasets to perform statistical testing in order to draw accurate conclusions about our variables' abilities to predict the winning nominees. When creating a single visualization that compared winners and non-winners, we utilized separate

y-axes, making the y-axis for the nominees exactly four times that of the winners. This was done to compare overall trends between winners and non-winners, disregarding the magnitude, as there will always be exactly four non-winning nominees for each winner.



2.1 Data Cleaning

The audience score and critic score datasets were generated and cleaned in similar ways. A dataset of Rotten Tomatoes certified critic ratings for over seventeen thousand films was merged with a dataset of Oscar data for about ten thousand films [6]. The rows were merged based on film title, which led to possible omissions based on differences in punctuation, spelling, or spacing. This process was repeated using a dataset of audience scores [5]. In order to make the dataset more digestible and focus on data relevant to our analysis, only Oscar ceremonies after the year 2000 were included, and all categories except Best Picture were excluded. We did not exclude outliers, as there could be legitimate explanations for scores outside the expected range.

In order to clean data relating to Oscar nominations for other categories, we again utilized a dataset of all Oscar data from 1928 to 2023 [1] along with a dataset for Oscar nominations found on Kaggle [3]. All data from before 2000 was once again eliminated from the dataset. Next, we separated our data in our dataset between winners and losers. This was done so by sorting based on the entry in the "win" column (either "TRUE" value or "FALSE" value). Our data was then sorted into separate datasets based on specific category and winning status.

To clean the datasets for total nominations and wins at other award shows, we first eliminated films from before the year 2000 from the dataset found on Kaggle [4]. The dataset that we were observing had a column for the total nominations at other shows and a column for total wins at other shows. We used this data, along with our large Oscar dataset, in order to separate films based on winning status in the Best Picture category. This combined dataset was easy to use in order to generate visualizations and summary data to analyze this variable.

Data relating to the films' budget and revenue was cleaned in a similar manner, eliminating films from before the year 2000 and merging our Kaggle dataset [2] with the larger Oscar-related

dataset. There were limitations in these datasets, though. Only films with high budgets and revenues were featured, and the dataset included less than two hundred films after cleaning. Because of this, there is a possibility that trends were misrepresented during statistical analysis and visualization.

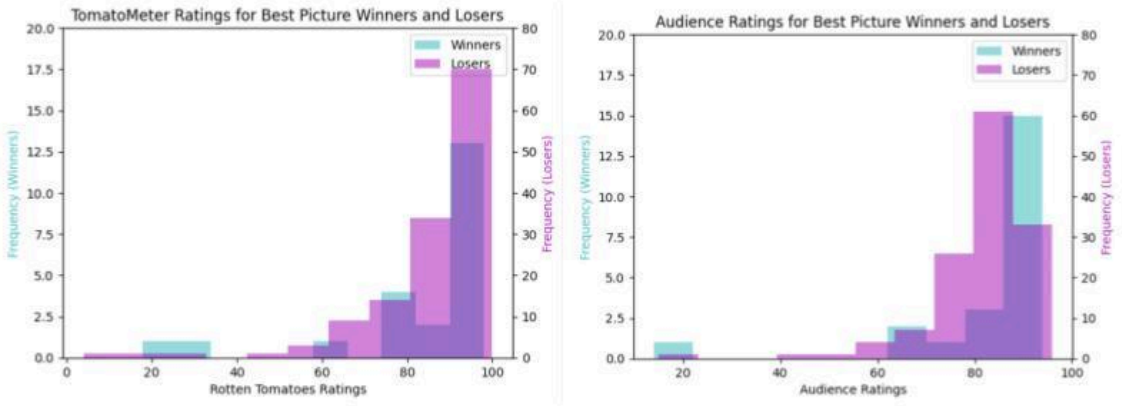
In order to generate visualizations and perform statistical testing related to relationships between our variables, we first needed to combine our individual datasets specific to our individual variables. This task proved challenging, as many film titles had small differences in the ways they were entered in our datasets, leading to exclusions. This occurred in differences in spacing, spelling, punctuation, and other small, easy-to-miss differences. Additionally, because our datasets were each focused on unique variables, many films included in one set were excluded from others. Generally, attempting to merge more than five datasets based on film title led to many exclusions and duplicates, making the final dataset unusable. Instead, when observing variable relationships, we constructed smaller datasets by merging just the individual variables' datasets that we were interested in. This allowed for fewer exclusions, and we performed a similar cleaning process to that of individual variables.

3. Results and Discussion

3.1 Visualization

Before we began statistical analysis, we created visualizations for each variable, as well as a variety of visualizations to observe variable relationships. These visualizations were used to qualitatively assess trends in our variables to make preliminary observations regarding if variables match predicted trends, and to form hypotheses for statistical testing.

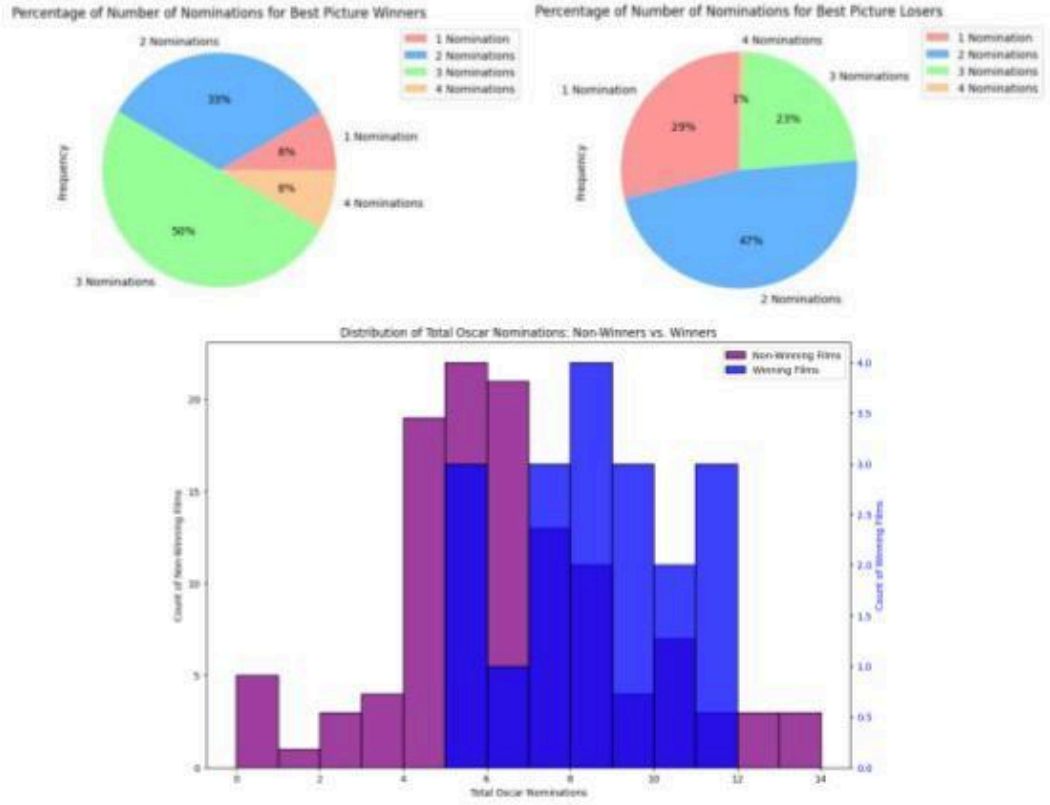
Figure 1: Critic and Audience Ratings for Winners vs. Non-Winners



We found the winning and non-winning films had similar trends regarding audience and critic scores. Each visualization in Figure 1 showed a left-skewed distribution, with average scores in the eighties. This is an expected result, as one would assume films nominated and winning Oscar categories would have generally higher audience and critic scores. The critic score visualization indicated the mean and standard error was 82.59 ± 3.67 for winners and 81.47 ± 0.89 for losers. The audience score visualization indicated the mean and standard error was 82.14 ± 4.50 for

winners and 85.94 $\pm$ 1.27 for losers. For both variables, the winners had a higher standard error, indicating a greater variance in scores than the non-winning nominee.

Figure 2:
Number of Oscar Nominations (Major Categories) for Best Picture Winners vs. Losers (Top) and Number of Oscar Nominations (Total) for Best Picture Winners vs. Losers (Bottom)



For winning versus non-winning films, we found that there was qualitatively a difference in the number of nominations received in the other Oscar categories, based on our visualizations in Figure 2. Based on our top visualizations, when rounding to the nearest percentage, we found that 50% of Best Picture winning films had three nominations in the four major categories (which we defined as the Best Picture, Best Actor, Best Actress, and Best Director categories). For non-winning Best Picture nominees, only 23% received nominations in three out of four categories. From summary data based on our dataset, we found that the mean number of nominations from the four major categories that we observed for winners was 2.58 $\pm$ 0.759 and the mean for losers was 1.95 $\pm$ 0.738. The standard deviations for winner and non-winners was extremely similar, meaning the data had similar variance in its distribution. Based on our bottom visualization, though both winner and non-winner data were normally distributed, we found that the winning nominees had a slightly higher average number of nominations across all Oscar categories.

Figure 3: Number of Nominations and Wins (Other Award Shows) for Best Picture Winners vs. Losers

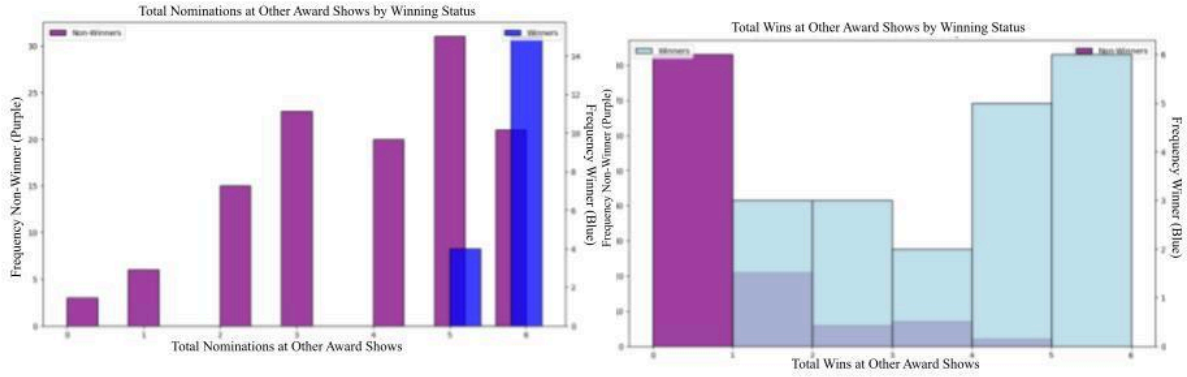
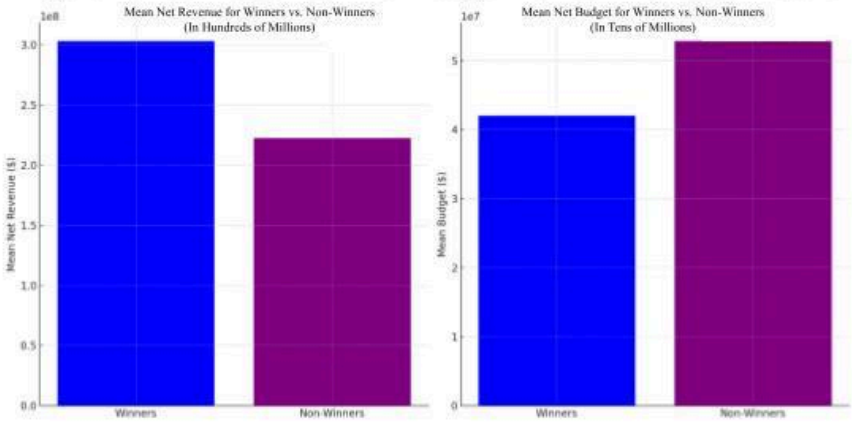


Figure 3's visualizations display the total nominations in other award shows by winning status and the total wins in other award shows by winning status. Based on our visualizations, we can assert there is a clear difference in distribution between the winners and non-winners in total wins at other award shows. The data for winners presents a slight left skew, while the data for non-winners exhibits a distinct right skew. This suggests that Best Picture winners generally win more awards at other award shows than Best Picture non-winners, which supports our predicted outcome. Summary data from the dataset supported this assessment, as Best Picture winners had an average 3.623 ± 1.77 wins at other award shows, compared to an average of 0.538 ± 1.02 wins at other award shows for non-winners. Though the difference is less pronounced for nominations at other award shows, Best Picture winners still generally receive a higher average number of nominations at 5.8 ± 0.49, compared to 3.9 ± 1.592 nominations for non-winners.

Figure 4: Net Revenue and Budget for Winners vs. Losers



These visualizations in Figure 4 highlight the difference in mean net revenue and budget for winners vs. losers for nominees in the Best Picture category. We found that while winners had a higher average net revenue compared to non-winners, the non-winners had a higher average net budget. These values were presented in hundreds of millions of dollars and tens of millions of

dollars for revenue and budget respectively, as revenue tended to far exceed budget, and we wanted to clearly visually demonstrate the trend.

3.2 Statistical Testing

We performed an independent sample t-test to find the correlation between critic score and being the Oscar-winning film. We performed the same test to find the relationship between the audience score and being the winning film. For these tests, our null hypotheses were there was no relationship between scores and being the winning film, and our alternative hypotheses were that there was a positive correlation between scores and being the winning film. These tests yielded p-values of 0.671 and 0.294 respectively. Because these exceeded our alpha values of 0.05, we failed to reject our null hypotheses at 95% confidence level. This suggests that there is no correlation between increased audience and critic scores and the likelihood of being the winning film.

Next, we performed a two sample t-test to determine if there was a statistically significant difference between the mean number of Oscar nominations in other major categories for winners vs. non-winners. Our null hypothesis was there was no correlation between having more nominations in other major categories and being the Best Picture winner, and our alternative hypothesis was there was a positive correlation. By performing this test, the p-value was determined to be $1.67\ x\ 10^{-4}$, meaning that the data is statistically significant, as it fell below our pre-determine alpha value of 0.05. We can reject our null hypothesis, concluding that there is a statistically significant difference between the mean number of nominations for winners compared to the mean number of nominations for losers.

We ran a Mann-Whitney-U test for total nominations and total wins at other award shows, since the data is not normally distributed. Our null hypotheses were that there would be no relationship between nominations and wins at other award shows and winning status in the Best Picture category. For each test, we determined our alpha value was 0.05. Testing the correlation total nomination at other award shows and Best Picture winning status, we computed a p-value of 1.78×10^{-7}. This p-value allows us to reject our null hypothesis, finding a positive association between total nominations at other award shows and winning status at a 95% confidence level. This suggests Best Picture winners generally are nominated for more awards at other award shows than Best Picture non-winners, which supports our predicted outcome. Performing the same test for the total wins at other award shows variable, we computed 3.09×10^{-12} as a p-value, which also means we reject the null hypothesis and say that there is a statistically significant difference between the wins at other awards shows for films that won the Best Picture category compared to non-winners.

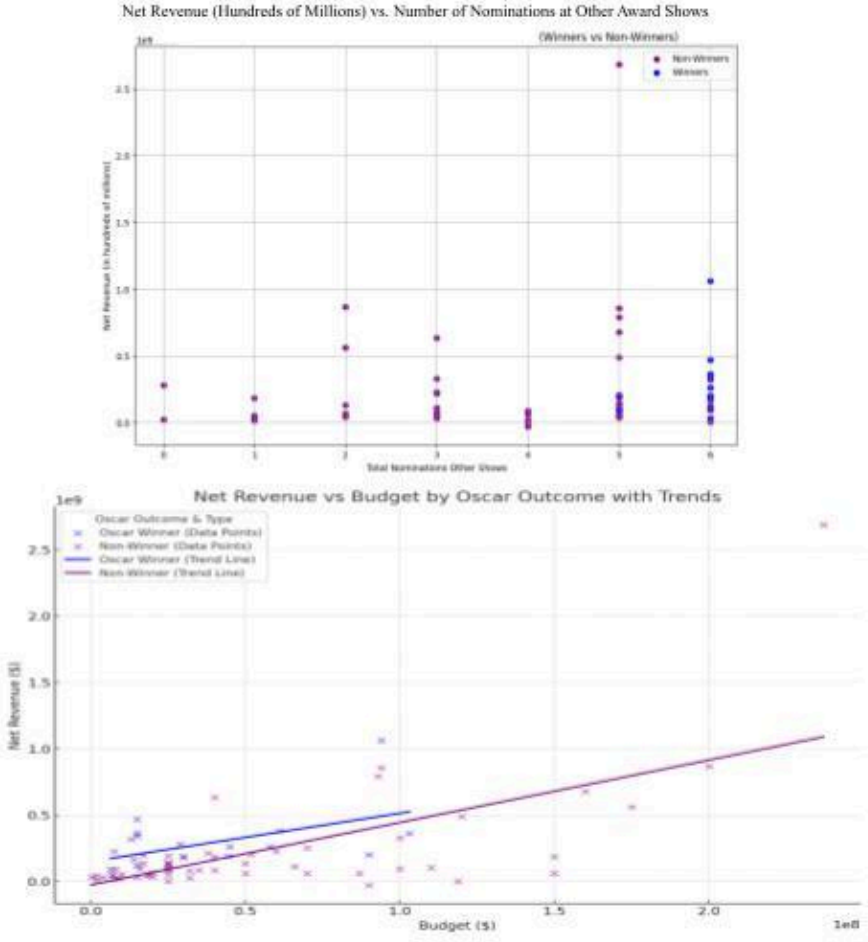
To test the budget and revenue variables' relationships with winning status, we once again performed two sample t-tests for each variable. Our null hypotheses were that there would be no correlation between being the winning film and having an increased budget or revenue, with our alternative hypotheses being a positive correlation would be observed. At a 95% confidence level, we failed to reject our null hypotheses for both variables. We found a p-value of 0.473 for our budget variable and 0.486 for revenue, exceeding our alpha value and suggesting there is no discernible correlation. It is worth noting once again that the dataset used to test this variable was

extremely small, so to make a definitive claim about these variables' association with winning status, we would ideally have access to more data.

3.3 Relationships Between Variables

Because the variables had the potential to interact, as they are all related to the same set of films, we briefly explored potential ways the variables could be related to each other. First, we observed the relationship between audience score and critic score, as we predicted they would be strongly correlated. Next, we observed the correlation between budget and revenue, predicting a strong correlation between those variables as well. Finally, we observed the relationship between net revenue and number of nominations at other award shows. These variables had a less intuitive relationship, so we attempted to observe if a positive correlation would still occur, as we predict both would be positively correlated with Best Picture winning status.

Figure 5: Variable Interactions
Net Revenue vs. Nominations at Non-Oscar Award Shows (Top) and Net Revenue vs. Budget (Bottom)



These visualizations in Figure 5 suggest a positive, linear relationship of moderate strength for both relationships. Additionally, these visualizations demonstrate that trends for relationships between variables differ based on winning status, once again suggesting that these variables could potentially serve as predictors of winning status.

The net revenue and nominations at other award shows variables had a correlation coefficient of 0.30, indicating that a moderate positive correlation exists between the variables. This correlation disregards the distinction between winners and non-winners, instead assessing the overall trend regardless of winning status.

The net revenue and budget had a moderate positive correlation, with an r-squared value of 0.394 overall. Additionally, we observed the difference in correlation for the data based on winning status. The winning films had an r-squared value of 0.231, and the non-winning films had an r-squared of 0.427. This indicates that the non-winning films had a stronger correlation between net revenue and budget than winning films.

Finally, though we did not generate visualizations for the relationship between audience score and critic score, we also tested the strength of this relationship. We performed an independent sample t-test, with a null hypothesis that there would be no correlation between audience score and critic score. We found that, with a p-value of 0.016, we could reject the null hypothesis. This demonstrates that there is a correlation between audience and critic score, as we predicted.

Based on these visualizations and tests, we generally found that our variables have at least some relationship with each other. This must be considered when evaluating the best possible variables for predicting winning nominees. Though one variable may appear to have a significantly higher correlation with winning status than the other, the variables are often associated with each other and could possibly be interdependent, leading to some of our results and observations.

Conclusion

We initially predicted that Best Picture winners would have higher budgets, higher revenues, higher audience scores, higher critic scores, more nominations in other Oscar categories, more nominations at other award shows, and more wins at other award shows. Based on statistical testing, we found that there is no relationship between increased budget, revenue, audience score, or critic score and winning status. However, performance in other categories and other award shows was positively correlated with winning status. This means films with more wins and nominations at other award shows and more nominations in other Oscar categories tended to be the winning Best Picture nominee.

Our visualizations demonstrated the greatest observable distinction between winners and non-winners in Figures 2 and 3, which related to the nominations in other categories, nominations at other award shows, and wins at other award shows. In visualizations for other variables, including Figures 1 and 4, which related to the critic score, audience score, budget, and revenue variables, the winners and non-winners had similar trends and distributions related to the variables' frequency or magnitude. These observations aligned with the data were derived from summary statistics, independent t-tests, and other statistical testing.

The variable with the strongest observed relationship with winning status, based on our statistical testing, was the number of wins at other award shows. We also found that the variable with the weakest observed relationship with winning status was the critic score. Based on this, if we were to apply these findings in order to create a predictive model to predict the winner in future nominees, the best variable for this model to utilize would be the films' wins at other award shows. In practicality, this could be difficult, as many award shows determine their winners or air after the Oscars. This additionally introduces the confounding variable that Oscar outcomes could influence other award shows, meaning these two variables are not completely independent, as we assumed in our statistical testing.

Generally, our project demonstrated an attempt to determine a variable correlated with winning status to predict future winners in the Oscar category. Though some of our predicted variables exhibited no relationship with winning status, our statistical analysis proved that there is a possibility the winning Best Picture nominee could be forecasted using a predictive model that analyzes the number of awards and nominations at other award shows each nominee has received, as this variable demonstrated the strongest correlation.

References

[1]"The Oscar Award, 1927 - 2024," www.kaggle.com. https://www.kaggle.com/datasets/unanimad/the-oscar-award?resource=download (accessed May 08, 2024).

[2]"Top 500 Movies by Production Budget," www.kaggle.com. https://www.kaggle.com/datasets/mitchellharrison/top-500-movies-budget (accessed May 08, 2024).

[3] "Oscar Academy Award-winning films 1927-2022," www.kaggle.com. https://www.kaggle.com/datasets/pushpakhinglaspure/oscar-dataset (accessed May 08, 2024).

[4] "Oscars Major Award Nominees & Winners 1928-2020," www.kaggle.com. https://www.kaggle.com/datasets/darinhawley/oscars-major-award-nominees-winners-19282020 (accessed May 08, 2024).

[5] "Top 100 Rotten Tomatoes Movies by Genre," www.kaggle.com. https://www.kaggle.com/datasets/prasertk/top-100-rotten-tomatoes-movies-by-genres (accessed May 08, 2024).

[6] "Rotten Tomatoes Movies and Critic Reviews," www.kaggle.com. https://www.kaggle.com/datasets/stefanoleone992/rotten-tomatoes-movies-and-critic-reviews-dataset (accessed May 08, 2024).

Source Code