Final Paper: Analysis of IMDB Movies Dataset

Group Leader: Kyle Aoki
Members: Emily Callahan, Brian McGrinder, Gabe Stocker, Andy Henze
Plots created by Kyle Aoki.

PART 1: INTRODUCTION (Henze):

Two easy ways to measure a cinematic production's success are to consider the feedback of its viewers and to assess the amount of revenue it produces. A plethora of factors may affect both a viewer's opinion about a movie and a movie's ability to earn money. This analysis aims to address two specific questions involving these measures of success. The questions are as follows: "What factors determine the IMDb score of a movie?" and "What factors determine the profitability of a movie?" We analyzed movie data to assess how some aspects of a movie may or may not affect its Internet Movie Database (IMDb) Score and its profitability. IMDb is an online service that provides online streaming for movies and television programs and a unique opportunity for individuals to rate movies based on a one-to-ten scale. This rating is what is referred to as an "IMDb Score."

The entertainment industry is vast and complex; these questions may provoke interest in those concerned with visual entertainment. Regular movie-goers or television binge-watchers could benefit from knowing the extent, if any, to which factors beyond the context of what appears on a screen affects an audience's reaction to what they experience. Assuming that an IMDb score is a reliable measurement of viewer-response, being able to identify and utilize a predictive relationship between production studios and IMDb scores may help audiences specify their choice of source for visual entertainment. It may interest producers or investors to learn during which season or month a movie has the highest chance of being released and returning a profit. It may be equally as interesting to be able to gauge to what extent making a profit may or may not be feasible during a calendrical period of time. Knowing how IMDb scores and profitability may be affected by things unbeknownst to the average audience member deserves interest from both private entertainment seekers and economic analysts.

Owning the movie data set without possessing a sufficient understanding of its internal relationships benefits the owner to no functional degree. With a proper understanding of how external and internal production factors affect the outcome of a product's performance, the owner of this information can make much smarter and safer decisions about how to produce his or her product. In the context of this project, the owner of this information would be assessing how to best produce movies and a movie's performance would be measured by its IMDb score and profitability. These two measurements alone more than sufficiently warrant a surface-level

investigation into the relationships between production factors and product performance. From an artistic perspective, it is important for a producer to effectively deliver a message or cinematic experience with which their target audience will resonate. From a business perspective, it is common practice for a producer of any type of good to make informed monetary decisions that will maximize a firm's profits. Considering both these perspectives, an in-depth investigation into what makes a movie successful in terms of audience satiability and net profitability makes the owner of such data a well-equipped artist and well-informed businessperson.

PART 2: DATA (Aoki):

The dataset for this project was acquired from Kaggle.com. It consists of approximately 5,000 observations and 50+ variables. Each row is an observation and each column is a variable. Each observation is of a movie produced in the time period between 1920 - 2017. The movie data was collected from IMDB.com through the use of a web scraping algorithm. A sample of 5,000 movies was scraped, as there are more than 5,000 movies on IMDB.com. Some variables contain dollar values; the dollar values are adjusted for inflation based on 2017 US Dollars.

FIGURE 0.1:

Genre	Count	Actor	Count	Production Company	Count
Drama	2207	Robert De Niro	46	Universal Pictures	310
Comedy	1629	Bruce Willis	35	Warner Bros.	250
Action	1120	Matt Damon	35	Fox Film Corporation	219
Thriller	986	Nicolas Cage	34	Touchstone Pictures	100
Romance	762	Samuel L. Jackson	34	Metro-Goldwyn-Mayer (MGM)	99
Adventure	748	Johnny Depp	32	Columbia Pictures Corporation	96
Crime	607	Brad Pitt	30	Relativity Media	70
Horror	490	Morgan Freeman	29	Regency Enterprises	67
Science Fiction	407	Tom Hanks	29	Lionsgate	55
Family	402	Ben Stiller	28	TriStar Pictures	55
Fantasy	350	Eddie Murphy	28	Dune Entertainment	52
Mystery	287	Mark Wahlberg	28	Summit Entertainment	52
Animation	223	Tom Cruise	28	Amblin Entertainment	47
History	169	George Clooney	27	Dimension Films	47
Music	152	Harrison Ford	27	Fox 2000 Pictures	47

Figure 0.1 above is a table of some of the categorical variables in the dataset. There are over 50 categorical dummy variables in the dataset. Dummy variables are variables which take either 1 or 0 as their value. For example, if “Drama” is a genre descriptor of a movie in the dataset, the “Drama” categorical dummy variable will have a 1. Otherwise, the value will be 0. The figure above shows several categorical variables in the dataset: genre dummy variables, actor dummy variables, and production company dummy variables. For the genre dummy variables, all genres were used in the regressions, as their counts were sufficiently high to merit their inclusion.

Correspondingly, the standard errors for the genre dummy variables were low, resulting in a majority of the genre dummy variables being significant at alpha = 0.05 in the regressions we conducted. The actor dummy variables were used as well, however, only the top 6 actors were included, as the rest of the actors had insufficiently high counts to merit being included. All actor categorical variables were insignificant at alpha = 0.05, which means our final models do not include actor dummy variables. Finally, we also tested production company dummy variables for significance. Many of the production company dummy variables had high counts, and thus low enough standard errors to be significant. However, few of them made it through the entire modeling process.

FIGURE 0.2:

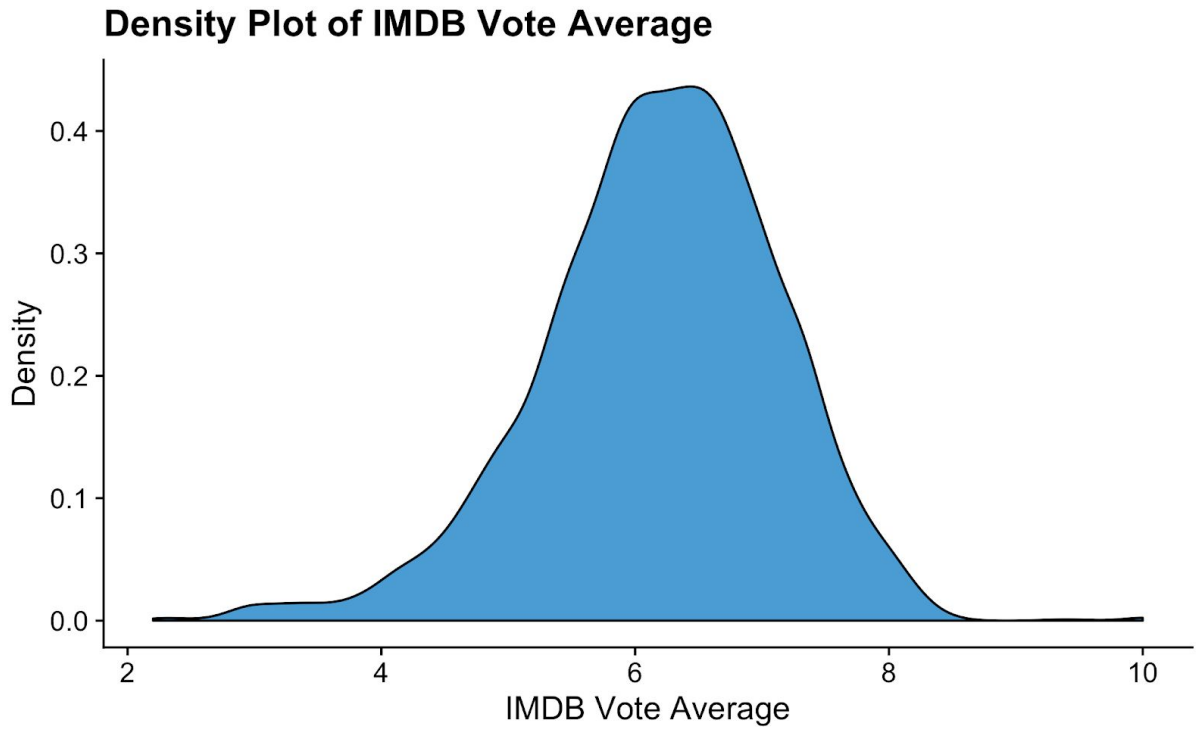


Figure 0.2 shows the distribution of one of our most important variables: IMDB vote average. Every movie in the dataset has an IMDB vote average, which is a statistic generated from averaging the votes of thousands of IMDB users who chose a rating between 0 and 10. Another variable in our dataset, vote count, tells us how many users voted for the movie. The distribution of vote average is centered at approximately 6.12, with standard deviations of 1.15. As one can see from Figure 0.2, vote average is approximately normally distributed, where few movies receive a very low or very high IMDB score. Thus, after having explored the data in the EDA, predicting vote average with other variables became a question of interest to us.

FIGURE 0.3:

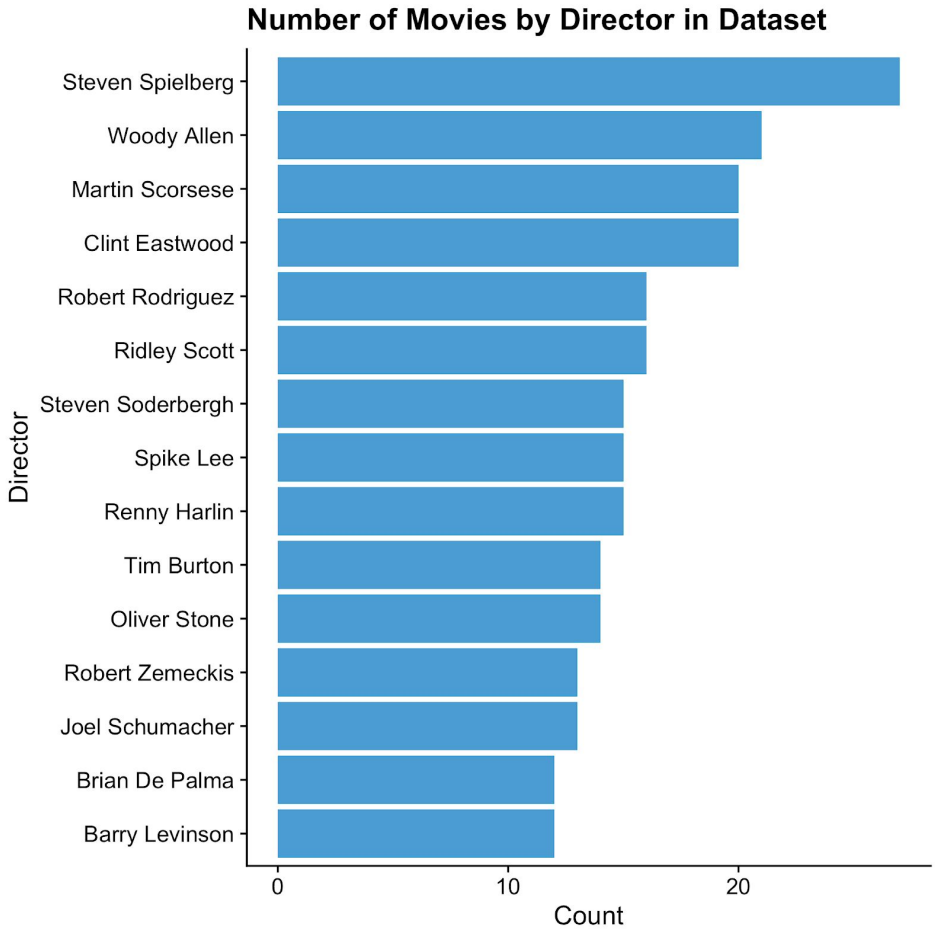
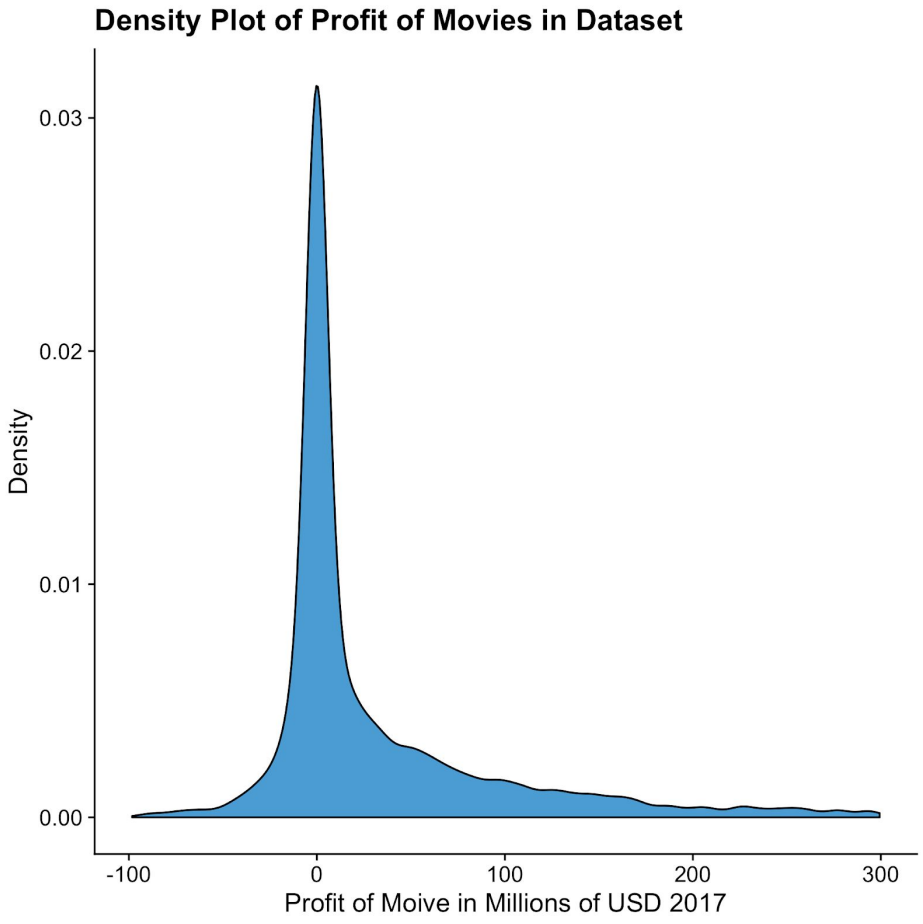


Figure 0.3 depicts a bar plot of director count in our dataset. We included director dummy variables in the regressions we performed. As directors can have a large impact on the outcome of a movie, we wanted to test the degree to which directors affect the IMDB and profitability of a movie. However, the counts for nearly all the directors were too low so as to produce low enough standard errors where the coefficient can be significant. Consequently, few director dummy variables survived the modeling process, however, some of them did.

Figure 0.4:

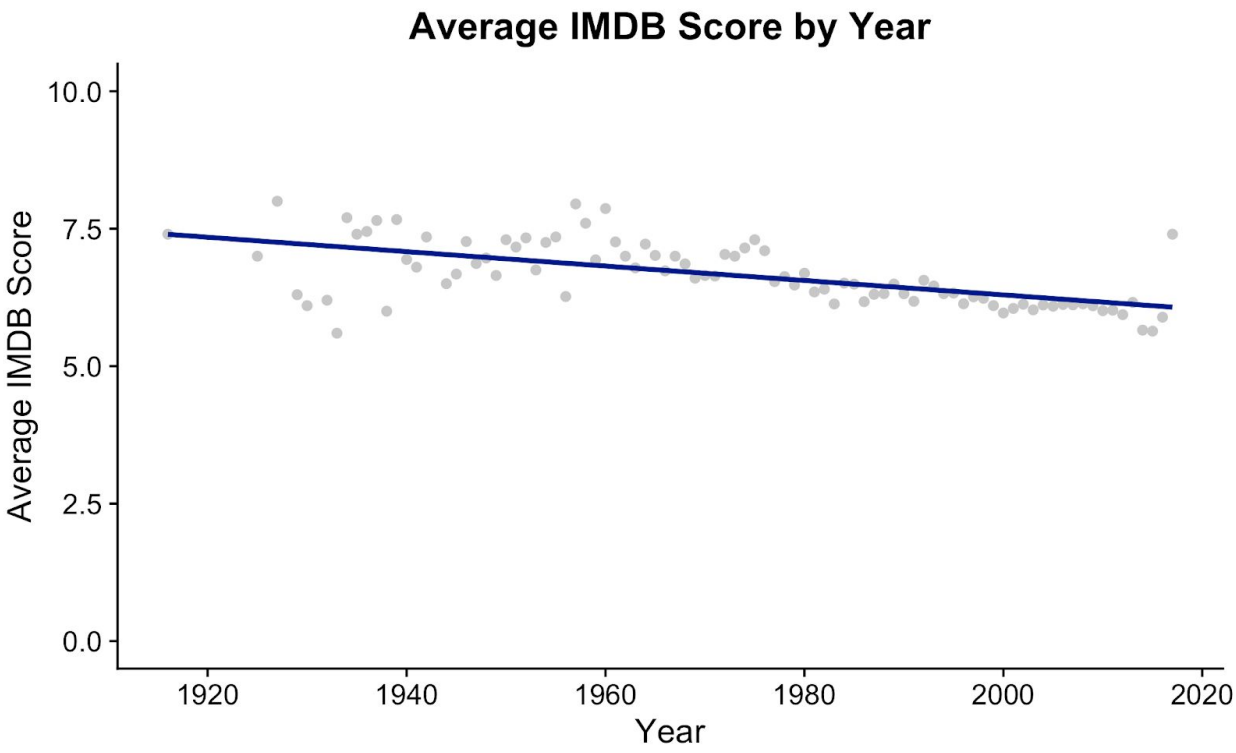


Another important variable in the dataset is the variable "profitable." This variable takes on either a 1 or a 0 depending on whether the profit variable is positive or negative. The profit variable is calculated by subtracting budget from revenue. Then, if the profit is greater than zero, the profitable variable takes on a 1. If the profit is zero or smaller, the profitable variable takes on a 0. The budget, revenue, and profit variables are adjusted for inflation based on 2017 USD. Figure 0.4 above depicts a density plot of the profit variable. As one may tell from the plot, approximately 47% of the movies in the dataset are not profitable and 53% are profitable. This means that 47% of the movies have either zero or lower as profit, and 53% have greater than zero as profit.

PART 3: RESULTS:

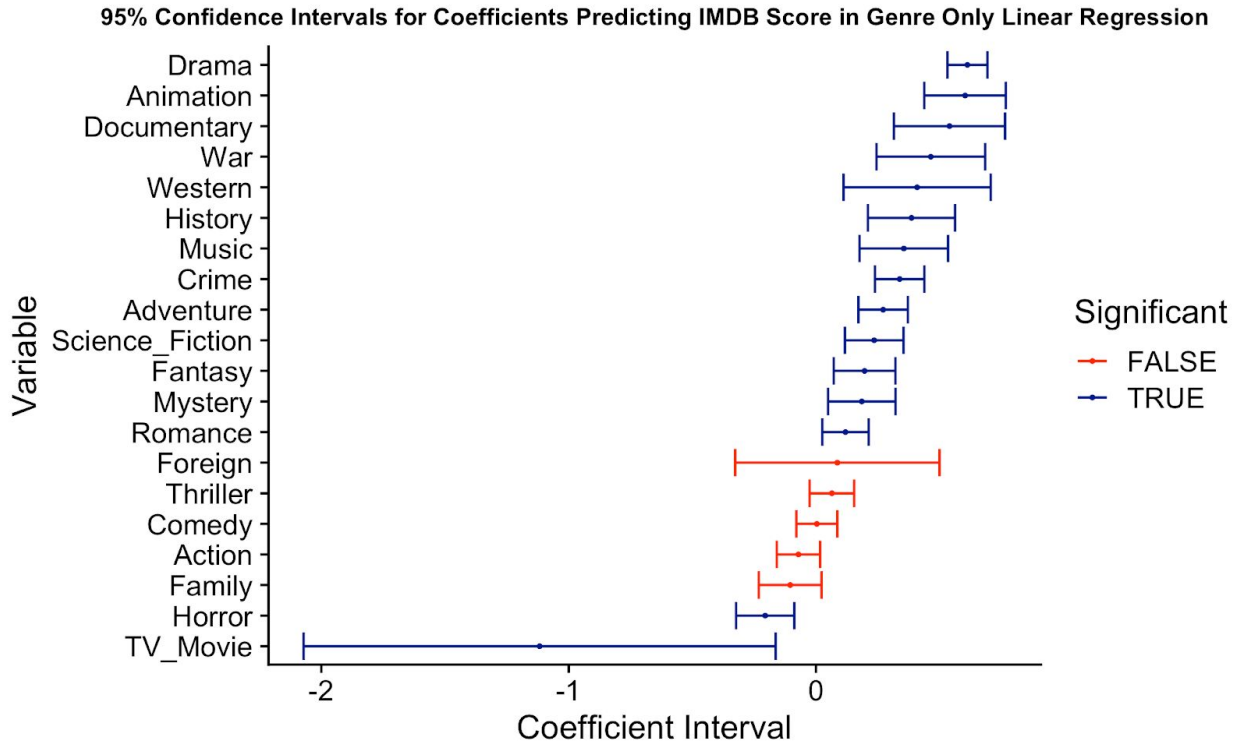
QUESTION 1 (McGrinder): What factors determine the IMDB Score of a Movie?

FIGURE 1.1:



Each movie in our dataset was given an IMDb score. This score originates from qualified users on IMDB.com that cast votes ranging from 1 to 10 on movies and TV shows from before 1920 to current day. IMDb originated in 1990, so many movies throughout history were retroactively given grades. The scores of each movie were put in bins based on the year they were released and then an average for each year was calculated. Figure 1.1 displays how the average IMDb score has changed over time. From the graph, it can be concluded that time and average IMDb score are at least somewhat negatively correlated. It appears that especially over the last 30 or so years the steady decline is even more consistent. The erratic values of average IMDb prior to 1960 may be due to the smaller sample size of movies in each year as well as possible differing opinions from people today about movies from over 80 years ago. Here it is also important to note that there is no omitted variable bias and IMDB score by release date still has a decent correlation. We know there is no omitted variable bias because the full regression of IMDB vote average did not cause release date to become insignificant.

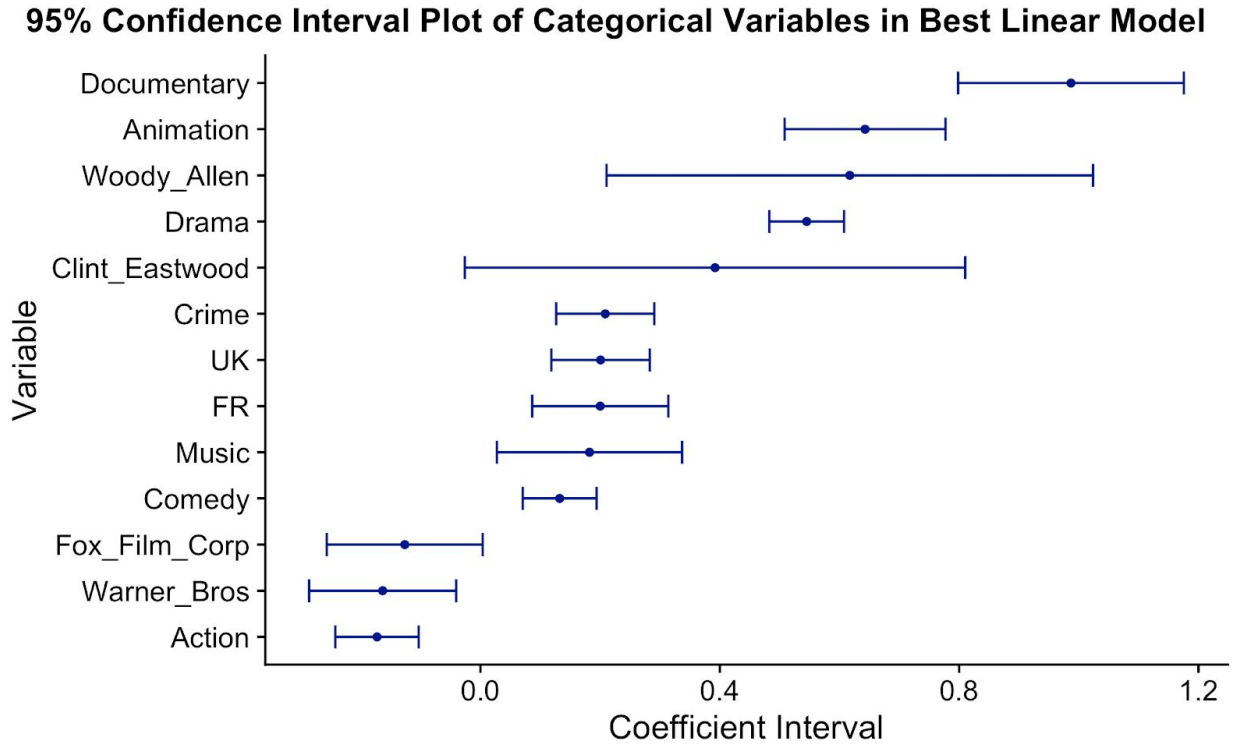
FIGURE 1.2:



Obviously, none of the previous findings matter if the data is not found to be statistically significant. The above confidence interval plot takes into account each genre's impact on the overall IMDB score of a movie. To calculate this, dummy columns were created in which a movie that had the genre desired was given a "1" and if it did not have the genre desired it was given a "0" and not included in the pool for the genre calculations. Of these variables, it appears that a movie under the scope of a TV not only has a statistically significant impact on the IMDB score of a movie, but has a fairly negative coefficient, even on the right tail of the interval. This makes sense as typically these movies are labeled "cheesy" due to poor production quality and do not have a large company to fund advertising for them. The movie with the highest positive coefficients and a statistically significant result were movies within the genres of animations and documentaries. This is an interesting result as advertisements would have a person believe that action and thriller movies would be the best movies. This result is likely due to the fact that documentaries and animated films take a significant amount of time and manpower, which allows for a lot of fine tuning to the film. Which could lead to this positive correlation with IMDB score. Moreover, even though many genres had statistically significant results, their

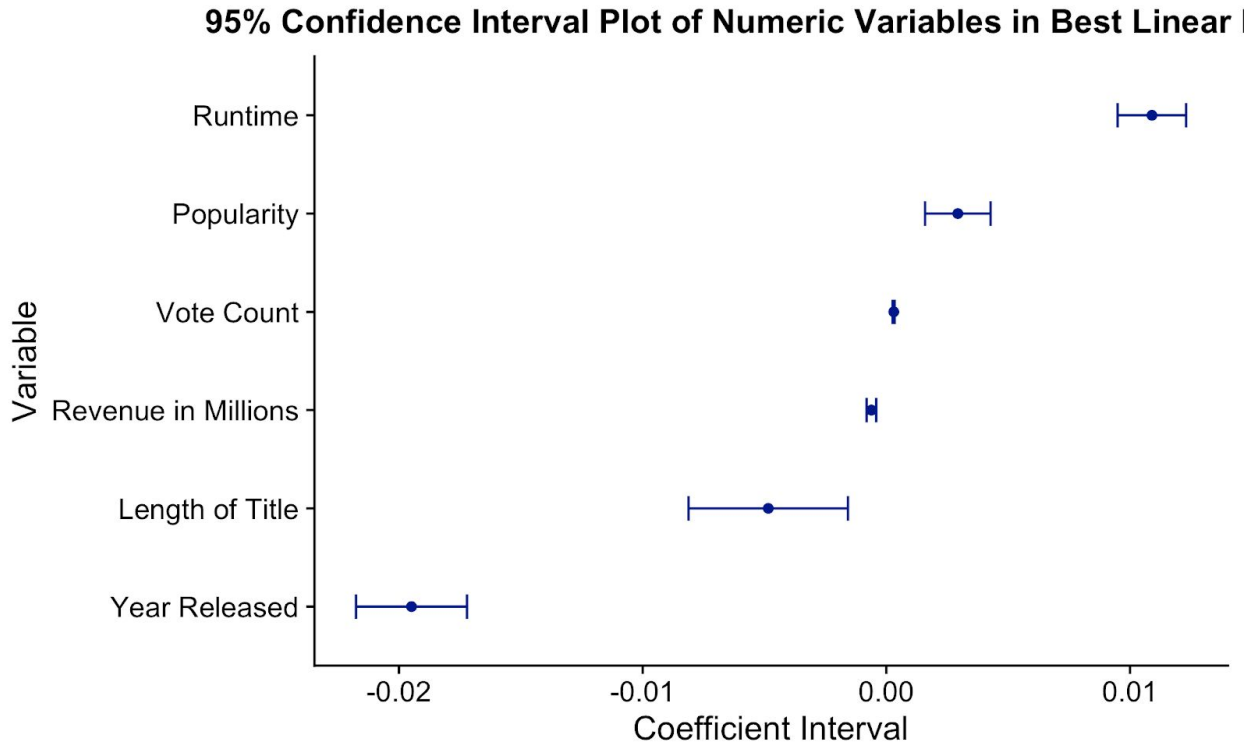
coefficients were so close to zero that their impact is not strong enough to draw any large conclusions from, besides that genre does not impact IMDB too much.

FIGURE 1.3:



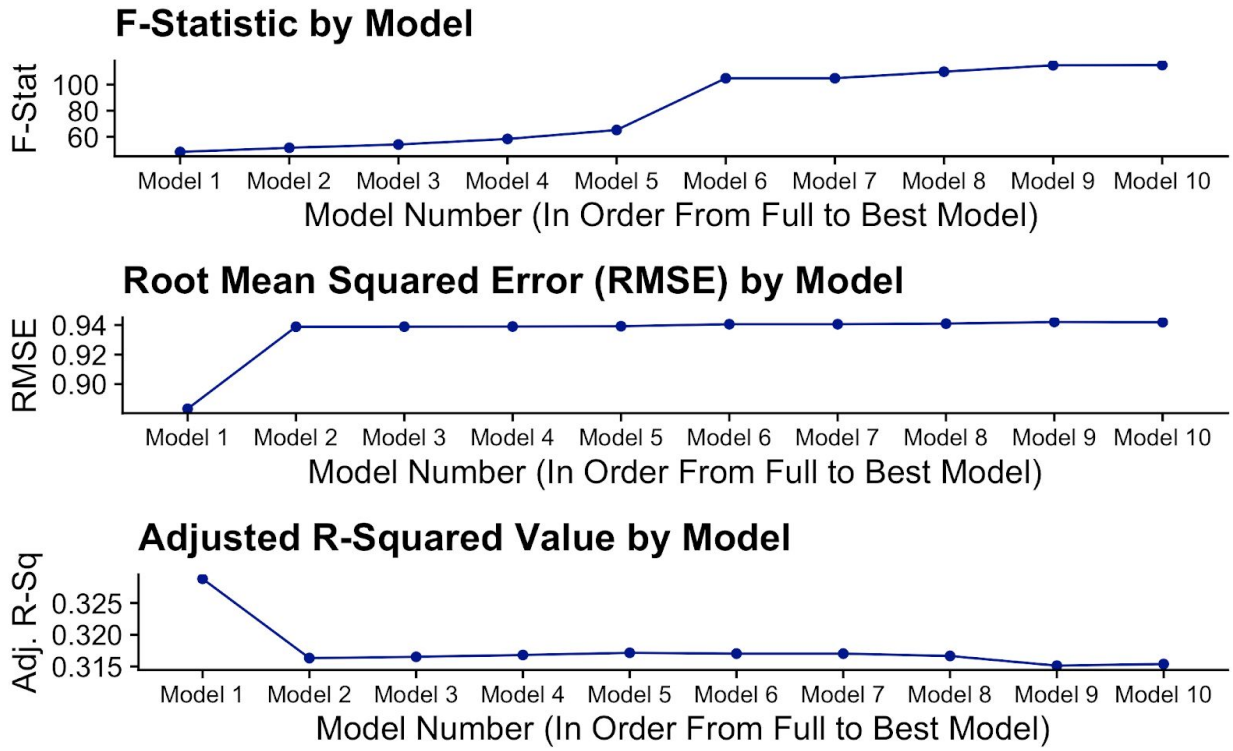
Following our analysis of genre, it was decided to run a larger scale examination of statistical significance within the categorical variables in our data set. Once again, dummy variables were created and movies were given a score of "1" if they had the desired categorical variable and a "0" if they did not. These bins were used for the calculation for the confidence intervals in our multiple regression model. Out of the confidence interval created, this model highlights results that were both statistically significant and had a fairly significant coefficient interval. One thing to note here is that "Clint_Eastwood" is not movies that the actor had acted in, but rather movies he has directed. Taking a look at the plot, it appears that documentaries once again shined through with the most positive correlation coefficient on IMDB score. Some other interesting takeaways were that movies directed by Woody Allen and Clint Eastwood had a massive coefficient interval because their standard errors were relatively large. Whereas other variables had hundreds of instances, the director dummy variables only about 30 instances in the dataset. Crime and Comedy, while not being massively positive in terms of coefficient, have a fairly small confidence interval and as a result appear to be fairly predictable in terms of how they scale the IMDB ladder. Overall, it appears that some categorical variables did have a fairly large impact on the success of a movie among IMDB critics.

Figure 1.4:



Another aspect of the data was numerical variables such as revenue in millions, popularity and vote count. As with the previous models, companies with fewer than 30 instances (movies made) in our data set were excluded. Dummy variables were not necessary in this case however, besides some to account for omitted variable bias (ex. Top 6 Production companies dummy variable) as each variable had a singular column. Looking at the model it is clear that numeric variables do not impact IMDB score as much as categorical variables do. This is because the coefficient interval range is much closer to 0 (.03 range) as compared to that of the (~1.6 range) of the categorical plot. However, since the numeric variables can take on large values, this result makes sense. Examining this plot further it can be seen that year released has a negative coefficient that is statistically significant. This is in line with what Figure 1.1 indicates, which was this somewhat negative relationship. On the other end of the spectrum, it appears that movies that IMDB score had a positive coefficient with relation to the variable runtime. This makes sense, as many films that are considered "quality" by critics are typically longer in nature such as at Sundance Film Festival. Besides year released however, it appears that these results would only be marginal in terms of their effect on the IMDB score of a movie.

Figure 1.5:

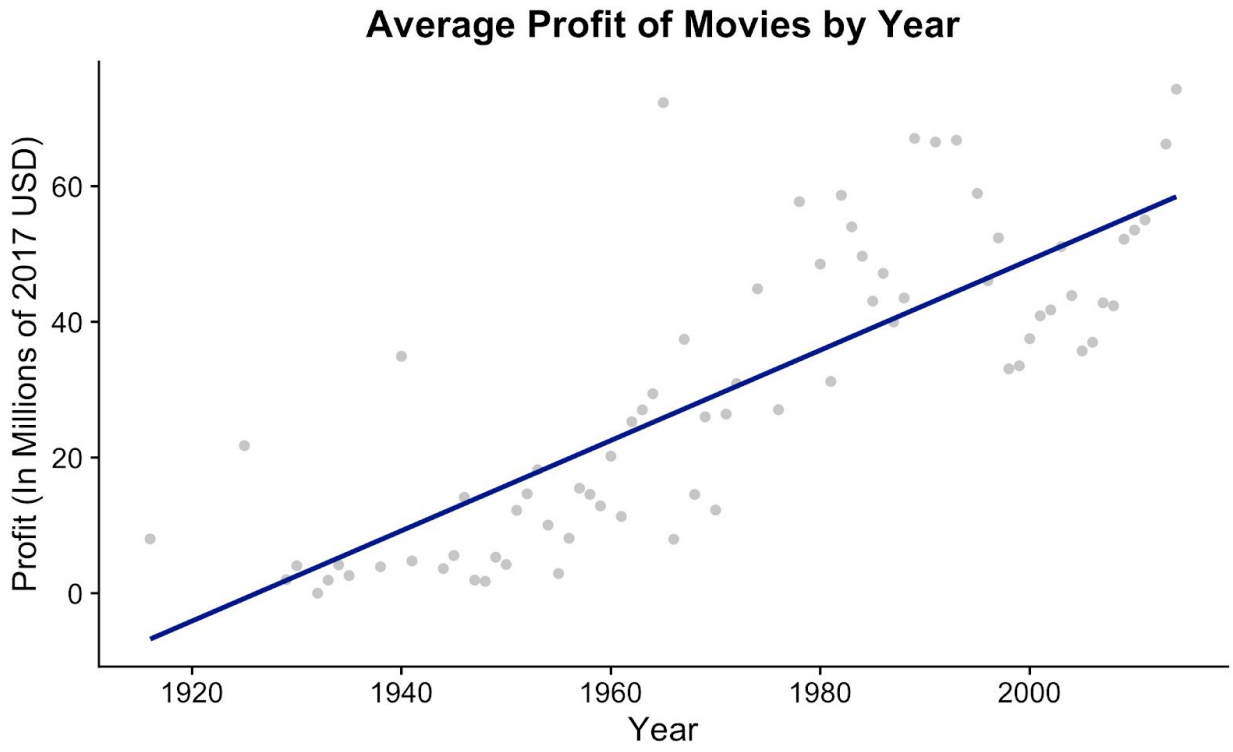


For Question 1, we chose multiple linear regression as our method of data analysis. The modeling process began with the creation of a "Full Model", which is a model that included all 50 of the variables in our dataset. The "lm" class of model object in R was used to accomplish this. Variables were removed in batches of 2 - 3 according to their p-value. The Model 1 points in Figure 1.5 above represent the Full Model. The three methods used to evaluate the models were: (1) the F-Statistic, (2) the Root Mean Squared Error (RMSE), and (3) the Adjusted R-Squared score. First, the data was separated in a 70-30 ratio with 70% of the data being used for training and 30% of the data used for testing. The 30% used for testing was randomly sampled from the entire dataset. Figure 1.5 shows the metrics derived from testing each model. The F-Statistic plot shows the F-statistic increasing as we removed insignificant variables from the model. The null hypotheses for each regression was that all coefficients were equal to zero, and the alternative hypotheses for each regression was that at least one of the coefficients was not equal to zero. The null was rejected in all models. The RMSE stayed relatively constant throughout the modeling process. Even after having removed over 20 variables, the RMSE hovered around 0.94, which suggests that the removal of those variables were correct decisions. Finally, the Adjusted R-Squared was also considered. The Adjusted R-squared is the percent of variation in IMDB score that is explained by the variation in the regressors. For most models, the

Adjusted R-Squared remained around 32% - 31%. This confirms that removing insignificant variables from earlier models was correct.

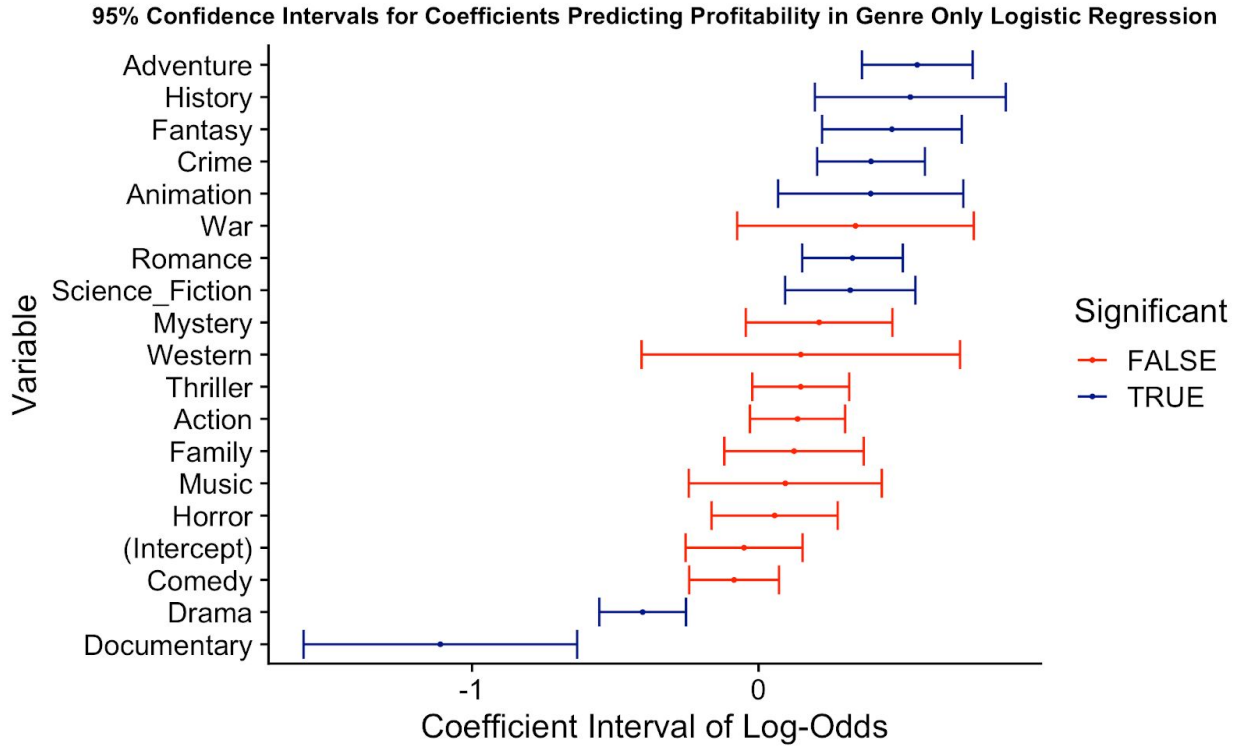
QUESTION 2 (Stocker): What factors determine the profitability of a movie?

FIGURE 2.1:



Within our dataset there were several indicators of numeric (monetary) value, the one we chose to analyze in relation to year released was profit. This decision was made as profit is adjusted for inflation within our data set and gross profit had more significant results than that of return (which was in a percentage). Similarly to the first plot in question one, the profits of each movie were placed in bins based on the year the movie was released. Each year's pool was then averaged out to find what the mean profit of movies released in said year was. Looking at the model it is clearly visible that the line of best fit shows a strong positive correlation between profitability and the year it was released. This means that on average, a movie that was released today will have a profit significantly higher than that of one released in the 1940's. That being said a caveat with this analysis is that it does not adjust for population growth, so while it may be true that the year a movie was released has an impact on the profit of a movie, it could also be related to global population exploding in the last one-hundred years. Once again, as the first model did, we accounted for omitted variable bias by including many other variables in our logistic regression.

FIGURE 2.2:

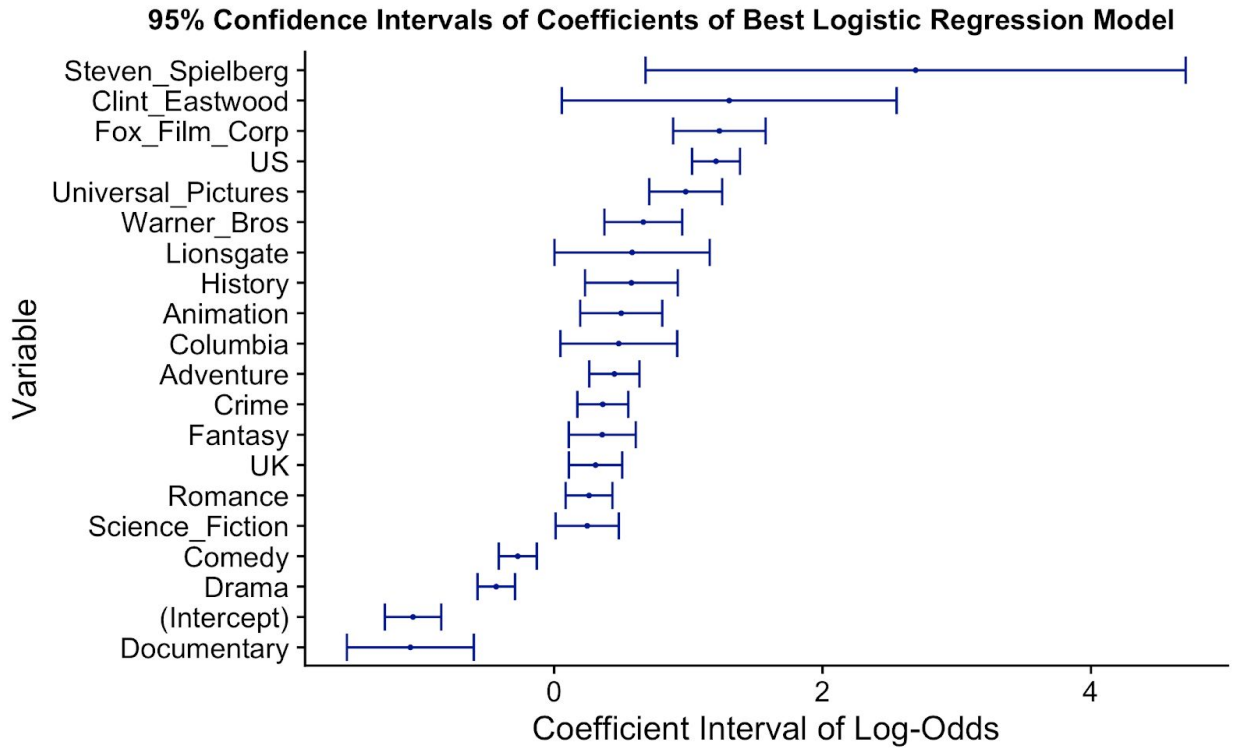


PARAGRAPH #2:

First, we conducted a logistic regression on the variable "profitable" using only genre variables. The "profitable" variable takes on a 1 if the movie in question grossed positive profit, and a 0 if it grossed zero or negative profit. Figure 2.2 portrays the logistic regression coefficients of each of the different genres that appeared 30 or more times in our movies dataset. Each interval corresponds with the effect that the genre has, whether positive or negative, on the log-odds the movie has of being profitable or not. The color of the line indicates whether or not

the genres are significant at a .05 significance level. It should not come as a surprise that the documentary genre had the most negative effect on whether a movie was profitable or not. This is because many documentaries are not-for-profit and are produced for the purpose of education rather than to generate money. What is notable about the figure is the genres that are most positively related in predicting a movies profitability. The genres that provided the greatest positive effect on profitability were Adventure, History, and Fantasy. However, 9 of the 18 genres we regressed, including the intercept, were insignificant.

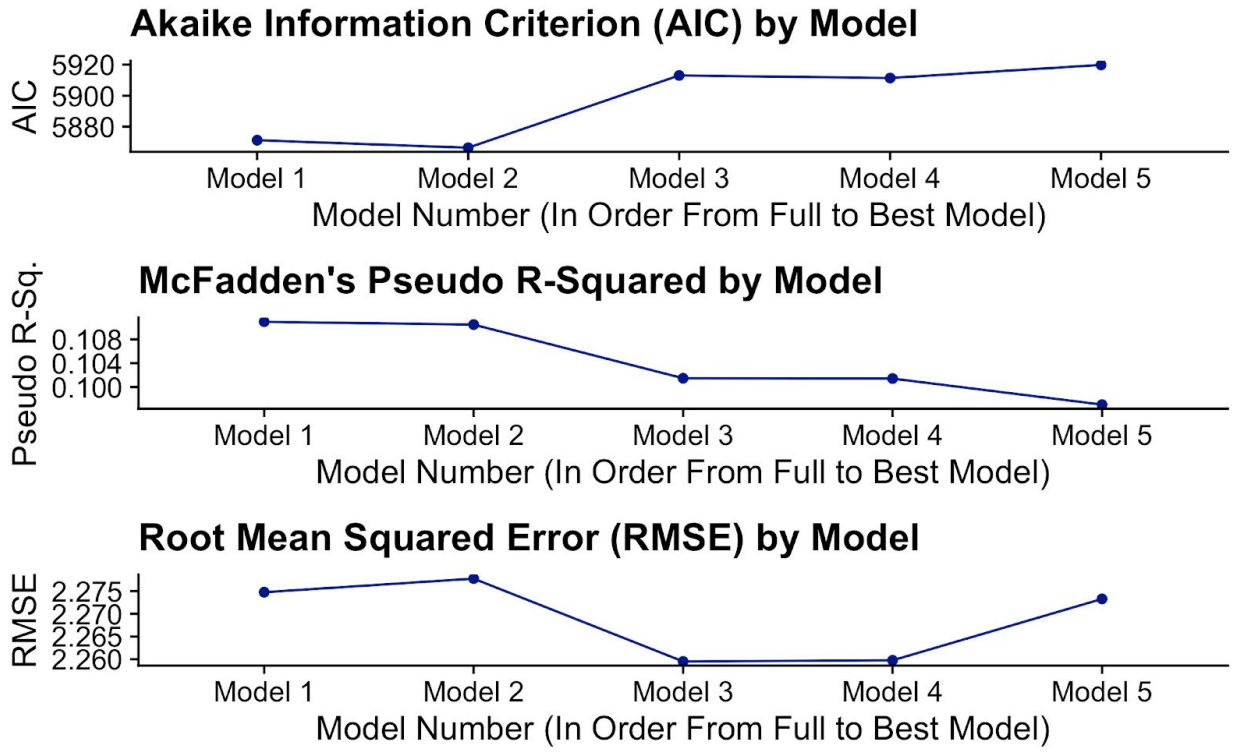
FIGURE 2.3:



PARAGRAPH #3:

After having conducted the genre only regression, we conducted a full regression which included all of our categorical variables. Figure 2.3 provides a visualization of the actors, production companies, genres, countries, and directors that proved to be significant when predicting whether a movie will be profitable or not. The two most influential are movies in which Steven Spielberg or Clint Eastwood were directors. This was surprising simply due to how many other directors were represented in the dataset. It was also interesting that no actors were significant when predicting movie profitability. This indicated that the director may be more important than the actors when predicting whether or not a movie will make a profit. Figure 2.3 additionally shows the influence of the large production companies in predicting profitability as they accounted for four out of the top seven predictors. The Comedy genre, which was insignificant in our genre only regression, became significant in the full regression.

Figure 2.4:



For Question 2, we chose multiple logistic regression as our method of data analysis. The modeling process began by building a "Full Model", which includes all of the categorical variables in our dataset. The numeric variables were not used, as many of them were available only after the movie was released. It is not useful to a company to build a model that predicts profitability if that model uses information about the movie available only after the movie is released. Therefore, only variables available before the release of the movie in question were used. The data was randomly split into a 70-30 ratio for training and testing. Figure 2.4 above shows the numbers derived from testing the models on the 30% of the dataset set aside for testing. Three criteria were used to evaluate the models: (1) Akaike Information Criterion (AIC), (2) McFadden's Pseudo R-Squared (pR2), and (3) Root Mean Squared Error (RMSE). For AIC, higher values are worse and lower values are better. This is because AIC represents the amount of information lost to the modeling process. The AIC stayed constant around 5900, despite the removal of many insignificant variables. The pR2 was used to asses that percentage of variation in profitability that is explained by the variation in the regressors. The pR2 remained relatively constant throughout. Finally, the RMSE was considered. It also hovered around 2.26, which suggests that the removal of insignificant variables did not compromise the model.

PART 4 (Callahan): CONCLUSION:

Throughout this analysis, our group found a combination of results that were expected and unexpected. For our first question, we examined the factors that determine the IMDB score of a movie. We found an overall negative correlation between release date and IMDB score, which was unexpected. We expected to see a positive correlation because the group assumed that IMDB scores would increase over time as film quality and production techniques have subsequently improved with time. The group also expected for numeric variables, like revenue in millions and popularity, to have a significant impact on IMDb score. However, after running our analysis, we found this assumption to be false as well. Another interesting conclusion we were able to draw from this first question was that animations and documentaries have the highest positive coefficients, meaning they have a strong positive correlation with high IMDb scores.

For our second question, we looked at what factors determine the profitability of a movie. The results from this analysis showed that there is a positive correlation with release date, which the group originally anticipated. However, after examining categorical variables impact on profitability we found that directors are more important than actors when predicting profit, which our group had thought would be the other way around. We also interestingly found that large production companies and the genres adventure, history, and fantasy all have highly positive effects on profitability.

The conclusions we were able to draw from this dataset are somewhat important in terms of the real world. While some of this information could be valuable for movie-goers and television binge watchers, it is most important for movie creators, producers and investors.Our results from Question 2 are more beneficial for movie production teams and investors because they are most likely concerned with which factors will impact how much money they will be making. Our analysis found several interesting variables that help predict profitability such as production company size, director and genres. By knowing which variables boost profitability, investors can develop models that allow them to be more confident that they are investing in profitable movies. However, our results from Question 1 seem to be less beneficial for movie production teams and investors because an IMDb score is given after the movie has been created and the money had been invested. Also, our Question 1 results found that IMDb scores are decreasing over time, which could potentially be due to the fact that there are differing opinions from people today about movies from over 80 years ago.

Although we had a semester to complete this analysis, much more time could be devoted to this investigation to discover deeper and richer insights. In our team, we had some members where it was there first time utilizing R. Our modeling could have been improved by utilizing more sophisticated methods such as k-nearest neighbors, which separates the data into classes in order to predict classifications of new points. We could have also improved our modeling by using neural networks to improve prediction in both the training and testing samples. Our analysis could have also been improved by using a dataset that scraped the entire IMDb website

as opposed to a random sample of 5000. In conclusion, we believe this research has the potential to be furthered and used in the real world.