Introduction

Internet Movie Database, better known as IMDb, was first launched in 1990 and has been a subsidiary of Amazon since 1998 (*What Is IMDb?*, n.d.). As the best-known online movie database, IMDb has data about millions of movies and TV programs. In addition to providing basic information about movies and TV programs like cast, director, and plot, IMDb allows users to rate the movies they have watched out of 10 (*How Do I Submit My Rating on IMDb?*, n.d.). Once a movie or TV program has received a significant number of ratings, IMDb will show the rating on the movie's page. Through aggregating user ratings, IMDb can better recommend different movies and TV programs to users and rank the top movies.

A common challenge for film producers is determining what type of movies will perform well and receive high ratings. As such, it is beneficial to analyze trends in top movies to determine what factors contribute to a high movie rating. Additionally, our team wants to look at how specific variables such as movie genre, age certification, budget, and run time impact the ratings of movies. Overall, we want to address the following business questions:

What factors contribute to a high movie rating?

- Which movie genre performs the best?
- Does the age certification affect how well the movie does?
- Does a higher budget result in a higher rated movie?
- Does the run time of the movie affect how well it does?

Data

Our team explored Kaggle for relevant datasets and two caught our attention. We chose to use the Movie Gross and Ratings dataset as it contained more movies. The dataset consisted of the top 20 movies of each year from 1989 to 2014 for a total of 510 observations (Sharaff, 2023). While 20 movies cannot be fully representative of all the movies that get released yearly, our team decided that the dataset was still better as there was a wide range of ratings in contrast to the other dataset, where the minimum rating was 8.0.

The data processing steps involved taking out certain variables and modifying others enhances the quality and reliability of our analysis. By removing variables like movie ID, gross, release date, and rating count, we can focus on the most important variables that directly impact

movie ratings. Eliminating movies with missing box office data ensures data completeness and accuracy, avoiding potential biases in our analysis. Excluding movies without runtime or budget information (such as *The Boat*) maintains the integrity of our revenue analysis.

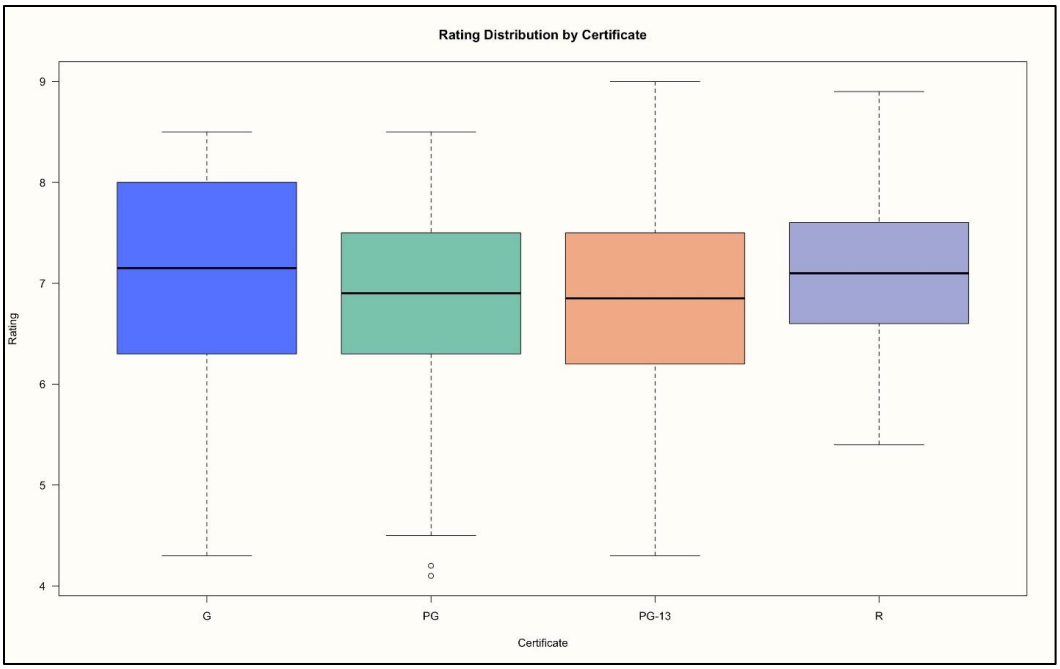
Converting unrated entries to "not rated" ensures consistency in our rating scale according to the entertainment standards of the 21st century and helps with meaningful comparisons across movies. Utilizing only the first genre listed for films with multiple genres simplifies categorization, providing clarity in genre-specific analyses. Standardizing age certifications (e.g., TV-MA, X, 18+ to R; GP to G; 13+ to PG-13) ensures uniformity in our findings. Removing certifications such as “Passed” and “Approved” in the dataset helps us to be more relevant and focuses on factors that accurately influence movie ratings.

Overall, these data processing steps are important in ensuring data accuracy, consistency, and relevance in our analyses and findings.

Name	**Model Role**	**Measurement Level**	**Description**
rating	Target	continuous	rating of film out of 10
budget	input	continuous	budget of film
genre	input	nominal	genre of film
certification	input	ordinal	rating of film (R, PG-13, G, etc.)
run_time	Input	continuous	duration of film

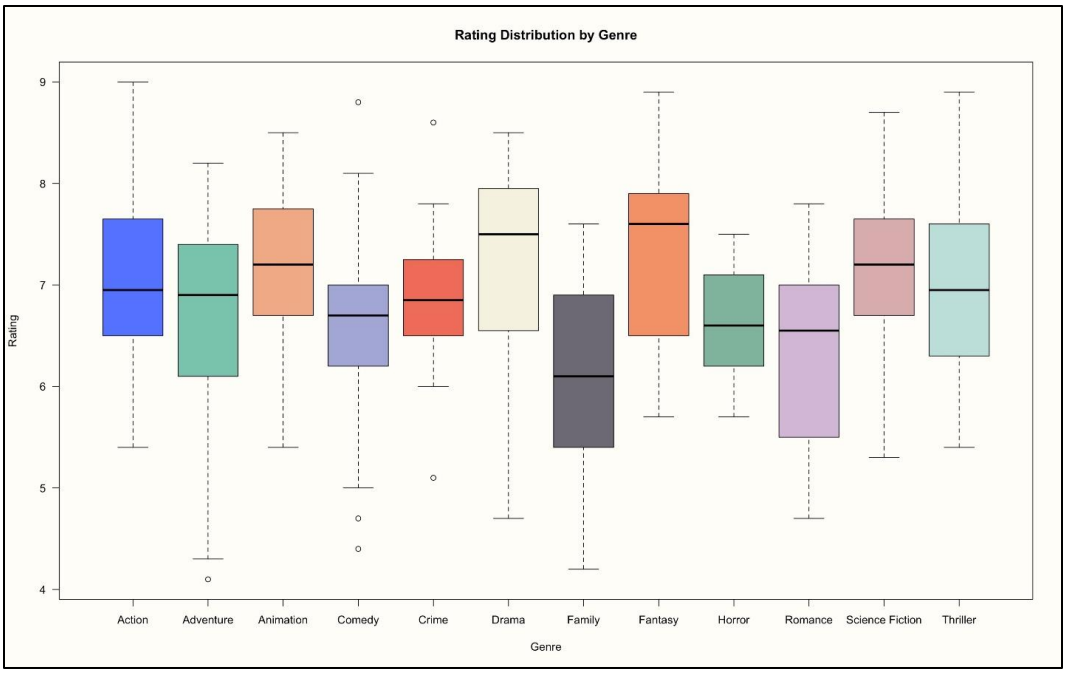
Analyses and Findings

Finding 1: Everyone Wants to be Scared



The impact of certification on rating has little variation, but it is still good for companies to analyze for reference. We see that rated movies have the highest median with 7.2, followed very closely by R-rated movies. A company would not be wrong to produce one over the other as it has little impact by 1/10 of a decimal. Although the distribution of each certificate is similar, R has a much higher minimum than the rest of the certifications. PG-13 has the lowest minimum, but also the highest maximum, considering it to have the largest range, therefore a movie's rating could vary greatly in this genre. G-rated movies have the highest average, so one could say that movies from this genre tend to have a higher average rating than the other certifications. We conclude that R-rated movies are the safest genre to produce from. G rated movies have the highest median. PG-13 is considered a "high risk, high reward" genre due to its range.

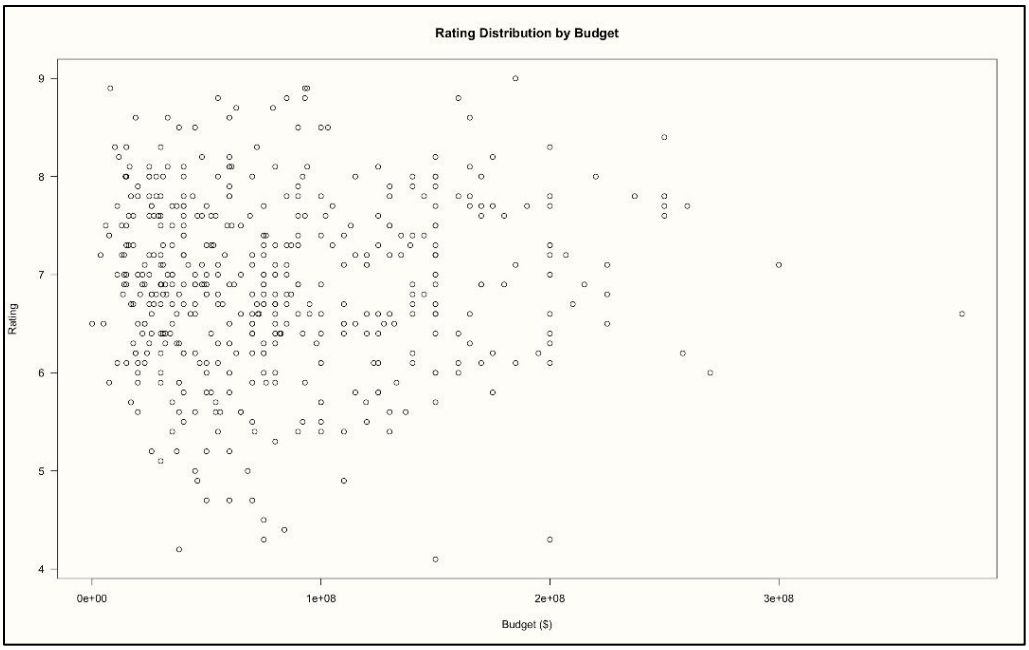
Finding 2: More Magic, More Drama



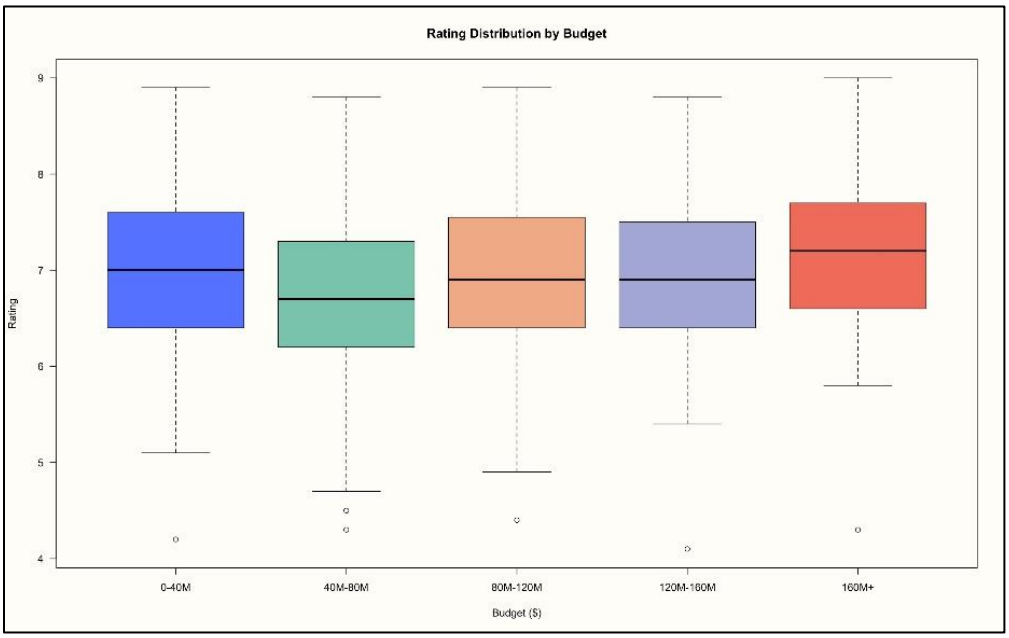
The second variable we looked at was genre. We saw that fantasy and drama movies perform the best on average, having median movie ratings of 7.6 and 7.5, respectively. Family movies perform significantly worse than any other genre, having a median rating of 6.0. Action movies have the highest maximum, but perform average compared to others. From these analyses, we can see that fantasy and drama movies are the best genres for companies to produce while family movies should be avoided due to their low ratings.

Finding 3: Higher Budget Does Not Equal High Rating

The third variable we investigated was the impact of budget on the rating of movies. We generated a scatter plot to visualize the spread of movies in relation to their budget and rating. Additionally, we generated a box plot to better view how different ranges of budget are rated.



In examining the scatter plot, we can easily see that most movies were made with a budget of less than $100 million. Additionally, the movies are mostly rated around a 7.0 out to 10.0. As the budget increases, most of the lowest ratings increase. However, the higher ratings are not significantly higher than the higher ratings of movies with a lower budget.



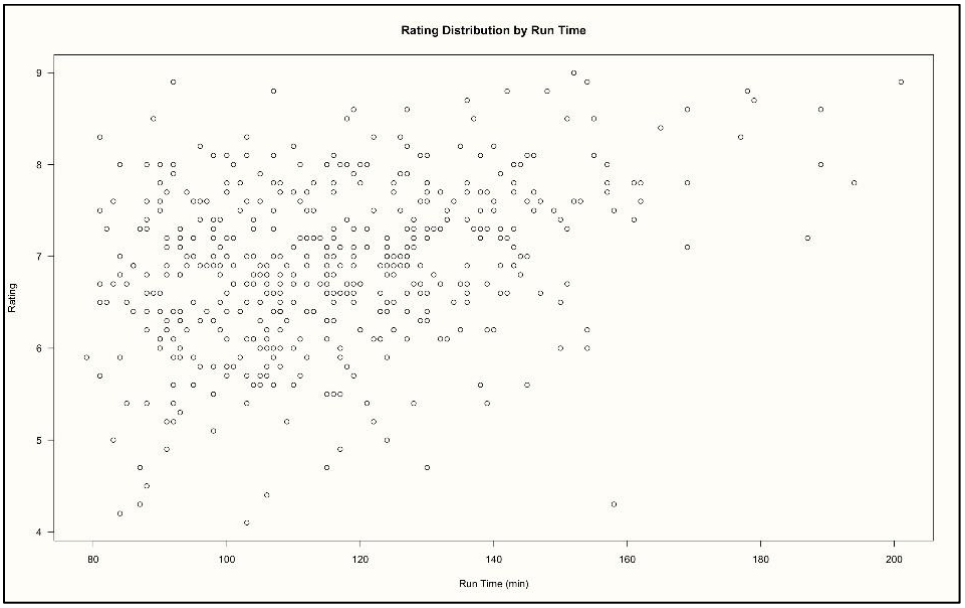
With the box plot, we can see that the range of budgets with the highest median rating are movies with a budget over $160 million with a median rating of 7.2. While the movies are the highest rated and have the highest median rating, this is not a general trend that is followed. The

second highest median rating comes from movies with a budget of less than $40 million. Additionally, movies with a budget between $40 million and $80 million are rated the worst with the lowest minimum and maximum rating.

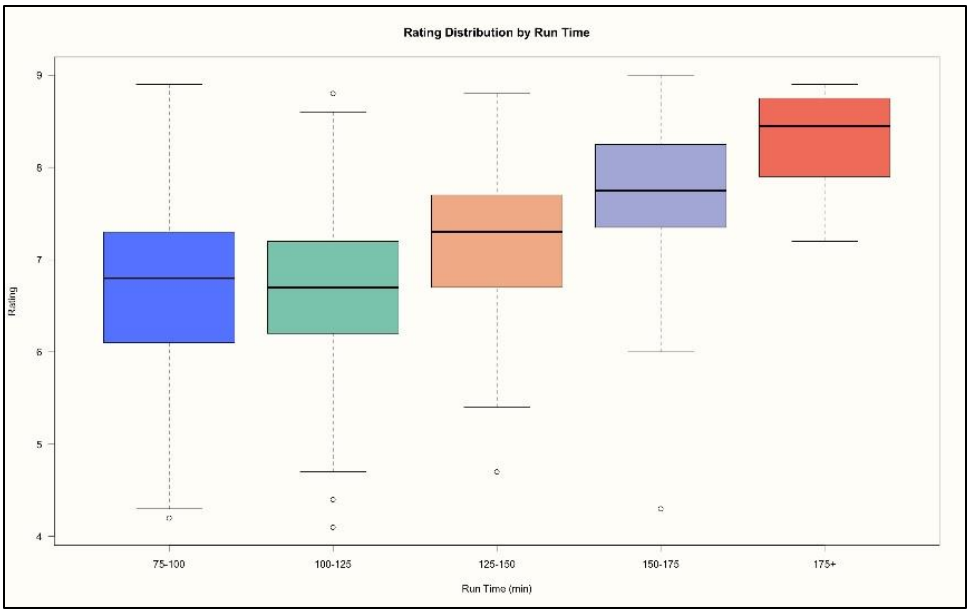
Overall, all the budget ranges have a similar maximum rating, so movies with any budget have the potential to be rated highly if the other variables are executed well. Therefore, anyone who wants to produce a highly rated film should not worry about not having a high budget. However, producers should not think funding the movie more will automatically make it successful.

Finding 4: Audiences Want More

The fourth variable we investigated was run time. Like with our investigation into budget, we utilized a scatter plot and box plot. The two plots allow us to look at individual movies and how they are rated in respect to run time as well as a general overview of runtimes.



From examining the scatter plot, we can see a positive correlation between run time and ratings. The lowest rated movies are all under 120 minutes besides one outlier. While movies with a shorter run time have the potential of being rated highly, movies with a longer run time tend to be rated highly.



With the box plot, the positive trend between run time and rating is even more clear. Movies with a run time between 100 and 125 minutes had the worst median rating compared to the other run times. Another range of run times that were rated poorly compared to the other run times is movies between 75 and 100 minutes. Those movies have the lowest ratings and the second worst median rating of 6.8.

While it can be easy to assume movies with long run times would be rated low, especially in the age of 30 second content, our investigations found that is not the case. Rather, viewers tend to favor movies with longer run time. Therefore, producers should not worry about movies being too long.

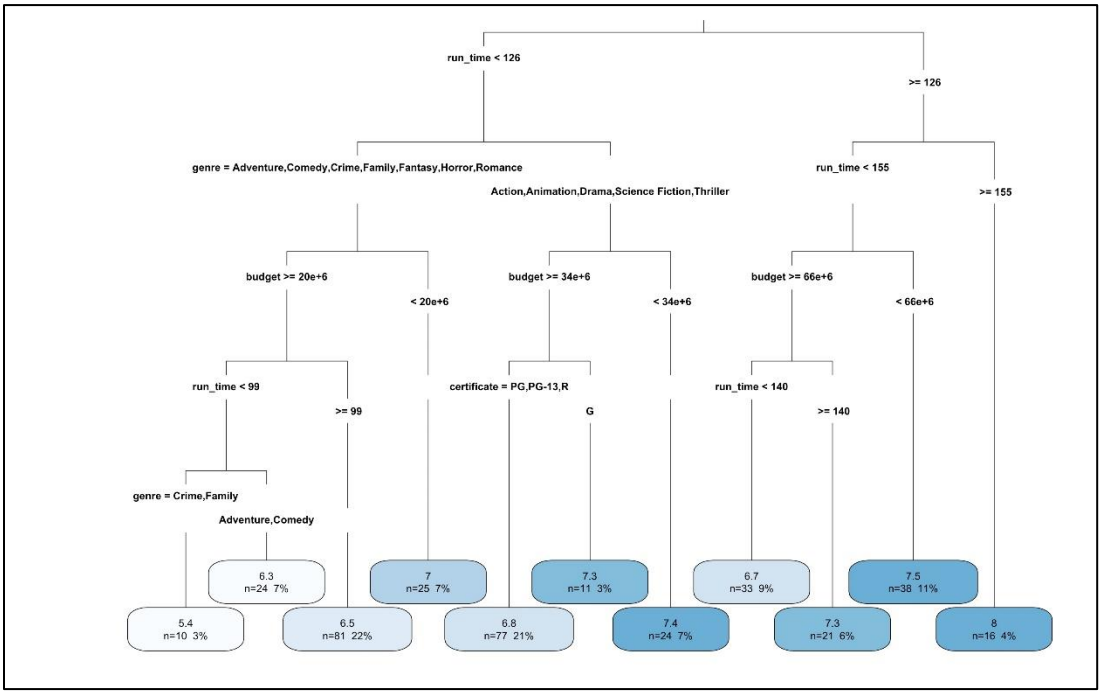
Linear Regression Model

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.63242 -0.44952  0.00469  0.48573  2.50722

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          5.007e+00  3.194e-01  15.678  < 2e-16 ***
run_time             1.964e-02  2.181e-03   9.004  < 2e-16 ***
genreAdventure      -4.494e-01  1.728e-01  -2.601  0.00958 **
genreAnimation       8.068e-01  1.757e-01   4.593 5.64e-06 ***
genreComedy         -2.583e-01  1.303e-01  -1.983  0.04801 *
genreCrime          -3.396e-01  2.151e-01  -1.579  0.11503
genreDrama          -7.869e-02  1.433e-01  -0.549  0.58309
genreFamily         -6.413e-01  1.940e-01  -3.306  0.00102 **
genreFantasy         1.374e-01  1.919e-01   0.716  0.47426
genreHorror         -4.435e-01  2.744e-01  -1.616  0.10671
genreRomance        -7.495e-01  1.796e-01  -4.173 3.59e-05 ***
genreScience Fiction 1.579e-01  1.614e-01   0.978  0.32857
genreThriller       -7.776e-02  1.584e-01  -0.491  0.62374
budget              -3.599e-09  7.744e-10  -4.647 4.39e-06 ***
certificatePG       -1.016e-01  1.719e-01  -0.591  0.55485
certificatePG-13    -9.024e-03  1.966e-01  -0.046  0.96342
certificateR         3.276e-02  2.058e-01   0.159  0.87359
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7525 on 464 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.3013,    Adjusted R-squared:  0.2772
F-statistic:  12.5 on 16 and 464 DF,  p-value: < 2.2e-16
```

We decided to use a linear regression model to show the relationship between our independent variables and the target variable. Our target variable being rating. At the top of the model, the residuals are shown. The median is 0.00469. This is close to 0, meaning the model accurately predicts the target variable, on average. We can then look at the coefficients. Some interpretations we can look at are for run_time and budget, since these are two statistically significant variables. We can see that for each minute increase in run_time, there is a 0.01924 increase in rating, holding all other variables constant. This shows us that movies tend to have higher ratings for having a longer run time Also, for every $1 increase in budget, there is a 0.000000003599 decrease in rating. This is strange, as you would think that a higher budget would directly correlate with a higher rated movie, but that is not exactly the case. That said, it is by a small amount. Since it is so small, this indicates that there is probably not a very strong linear relationship between budget and rating. We can then look at the statistical significance of each variable. We can see that run_time, budget, genreAnimation, and genreRomance are the most statistically significant variables. They are significant at the .001 level. There are also genreAdventure and genreFamily that are significant at the .01 level as well as genreComedy that is significant at the .05 level. This tells us that the findings for those variables are less due to chance in our experiment. The findings for those variables can be used in other experiments and

for other datasets. Lastly, we will look at the r-squared value. We will use the adjusted r-squared value, as it adjusts for the number of predictors in the model. We can see that the adjusted r-squared value is 0.2772. This means that 27.72% of the variation in rating can be explained by the variation in the independent variables. This is not a very high r-squared value, telling us that the model is only useful for explaining a small portion of the variability in rating.

Decision Tree Diagram



We used a decision tree to further interpret our model. On our decision tree, our root node is run_time. This was chosen as the variable that best splits the data into subsets. Following the branches and interval nodes, you can see which rating each of the variables leads to. As you can see, the movies rated the highest are those that are more than 155 minutes (about 2 and a half hours) long. The lowest rated movies are those that are less than 99 minutes (about 1 and a half hours) long, have a budget of over $20 million, and are either the family or crime genre. This combination is predicted to have an IMDB rating of 5.4. That said, a crime movie with a budget of under $20 million is predicted to have an IMDB rating of 7.0. The same genre is predicted to have a different rating depending on other variables.

Implications

Predicting how well a movie will perform is a difficult task. There is no one way to make a successful movie, otherwise everyone would do it. The main thing that producers and directors should take away from this project is that longer movies tend to do well. That said, it must be full of content that viewers would enjoy. If the movie is very high quality and appeals to the audience, they will want more of it. That is why the best movies tend to be long. Viewers get more of what they like. After that, you must make sure you are producing a movie with a genre people want to see. Based on our findings, that tends to be drama or fantasy. Any genre can perform well but these two have the highest ratings on average. Anyone using this project should keep in mind that not all our variables are statistically significant. There is a chance that some of our findings are particular to this specific dataset and would not be the same if this project were to be replicated with another dataset. That is also something to consider when using these results.

Conclusion

Overall, our project was successful. Although we were only able to explain about 28% of the variation in movie rating, we were able to see which variables were statistically significant, given the variables we included. We were able to get a good idea as to which type of movies perform well in the ratings based on run time, budget, genre, and certificate. You cannot use this to perfectly predict how well a movie will be rated, but it is a good start to understand which movies do well based on certain variables. Our findings show that G and R rated movies typically get the best ratings. Also, fantasy and drama genres perform the best. If your budget is over $160M then the movie is more likely to get a higher rating, but if it is under $160M then the specific budget does not matter as much. Finally, the longer the movie is the better it typically performs.

References

How do I submit my rating on IMDb? (n.d.). Retrieved April 29, 2024, from https://help.imdb.com/article/imdb/track-movies-tv/how-do-i-submit-my-rating-on-imdb/G9R8NF943K39DQDT?ref_=helpsect_pro_2_4#

Sharaff, Y. (2023). *Movie Gross and Ratings* (Version 2) [dataset]. https://www.kaggle.com/datasets/thedevastator/movie-gross-and-ratings-from-1989-to-2014

What is IMDb? (n.d.). Retrieved April 29, 2024, from https://help.imdb.com/article/imdb/general-information/what-is-imdb/G836CY29Z4SGNMK5?ref_=helpart_nav_1#