IMDb Data Analysis

A data-driven approach to exploring IMDb Data on Movies

Authors:

Sarah Li (10174020)

Vardaan Bhatia (30181243)

Gavin Lau (10151742)

Prem Patel (30192278)

Table of Contents

Introduction

Significance of IMDb

Movies have the unique ability to transport us to other worlds, cultures, and times, and evoke a wide range of emotions. For many people, including everyone in this group, movies hold a special place in their hearts and have been a source of entertainment, inspiration, and comfort throughout their lives. For many movie enthusiasts, their passion goes beyond just watching movies. They may collect movie memorabilia, attend film festivals, participate in online discussions about their favorite films, or even create their own movies. Some may also study film history or theory, analyzing the cultural and social significance of different movies and the impact they have had on the world. Whatever form it takes, a passion for movies can be a deeply personal and meaningful aspect of a person's life. It can bring people together, create shared experiences and memories, and inspire creativity and imagination. It's no wonder that so many people are drawn to the world of cinema and the many stories it has to tell.

From the classic black-and-white films of the 1920s to the latest blockbusters of today, movies have continued to captivate audiences and become an integral part of our cultural fabric. As we all have an intense passion for movies, we are all intrigued by the inner workings of the film industry and the factors that contribute to the success or failure of a movie. This motivated us to explore the world of data analysis and its application to movie data, particularly the vast dataset provided by IMDb. By utilizing the power of data analysis, we hope to gain insights into the movie industry and uncover the factors that contribute to the success of a movie, and in turn, contribute to the improvement of the industry as a whole.

IMDb (Internet Movie Database) is one of the most comprehensive and widely used online databases for information about movies, television shows, and other content. The information on IMDb includes details about the cast, crew, plot summary, release date, and ratings, among other things. The IMDb dataset provides a wealth of information that can be used to analyze various aspects of the film industry, including trends in box office revenues, the popularity of different genres, and the impact of factors like budget and runtime on a film's success.

By performing data analysis on the IMDb dataset, we can gain insights into various aspects of the film industry, such as:

- Predicting a movie's success: By analyzing the impact of factors like budget, runtime, and ratings on a movie's success, we can develop predictive models that can help us forecast a movie's potential success at the box office.
- Identifying popular movie genres: By analyzing the ratings and other information about movies in the dataset, we can identify which genres are most popular among movie-goers. This information can be useful for movie studios and production companies when deciding which types of movies to produce.
- Understanding movie ratings: By analyzing the ratings and other information in the dataset, we can gain insights into which movies are typically given which types of ratings. This can help audiences gain insight when choosing the movie to watch, to understand the likelihood of its' appropriateness for the intended audience type.
-

Overall, the IMDb dataset provides a rich source of information that can be used to gain insights into various aspects of the film industry, and the data analysis on this dataset can provide valuable insights for filmmakers, movie studios, and others in the industry.

Objectives

The objective of this data analysis is to explore the IMDb dataset and understand the relationship between various movie attributes with the available dataset. Through this analysis, we aim to develop models that can predict IMDb score ratings, good/bad movies, gross (i.e. revenue) of the movie, and MPAA (Motion Picture Association of America) ratings based on the available attributes.

To achieve this objective, we will use various statistical methods within the framework of 4 exploratory questions. Further details on the objectives and significance of each exploratory question can be found below.

1. Can we predict the IMDB Score Rating based on budget, votes, gross, and genre as our predictors?

Predicting IMDb score ratings can have different motivations depending on the context and the stakeholders involved. Here are some of the reasons why predicting the IMDb score rating can help potential stakeholders:

- Movie studios may want to predict the IMDb score rating of their upcoming movies to estimate their potential commercial success and adjust their marketing and distribution strategies accordingly.
- Filmmakers and producers may want to predict the IMDb score rating of their completed movies to assess the quality of their work and evaluate how well it was received by the public.
- Film critics and enthusiasts may want to predict the IMDb score rating of upcoming movies to get a sense of how they will be perceived by the wider public and use this information to inform their reviews and recommendations.
- Researchers in the field of movie analytics may want to predict the IMDb score rating of movies to study the factors that contribute to their success or failure, such as the impact of budget, genre, or cast.
- In general, the motive behind IMDb score prediction is to gain insights into the factors that influence the reception of movies by audiences and to use this knowledge to inform decision-making, research, or critique.

2. Using logistic regression to model good/bad movies and multinomial regression to model genre based on the other variables

By using logistic regression to model good/bad movies, the aim is to identify the important factors that distinguish between successful and unsuccessful movies. The predictors used in this model may provide insights into the aspects of a movie that are most important for audience reception and critical acclaim.

On the other hand, the use of multinomial regression to model genre based on other variables can help in understanding how different factors are associated with different movie genres. This may be useful in identifying patterns and trends in movie production and audience preferences, which can inform marketing and production decisions.

a) Using Linear Regression to model Gross Income

One of the most significant metrics for determining a movie's success is the gross income. This is typically

defined as a combination of box-office success and subsequent sales. Because of the continuous nature of this metric, linear regression seems most appropriate. If a model can accurately predict a movie's gross income based on its other parameters, we would be able to better understand what factors contribute to a movie's success.

b) Using Logistic Regression, LDA and QDA to predict star gender.

Over the course of the project, we discovered an interesting relationship between the gender of the star and the movie's gross, namely that the gender of the star was a significant contributor to the movie's gross income. There is a massive gender disparity in Hollywood, from directors to stars to writers. We try to predict whether or not a movie's star will be male or female, based on the parameters of the movie.

3. Can movie ratings be predicted? If so, with which predictors?

The purpose of this question is to explore the possibility of predicting Motion Picture Association (MPA) film ratings based on certain predictors. The MPA film ratings system provides information about the suitability of a film for different age groups and is an important tool for parents, educators, and others in making informed decisions about which movies are appropriate for different audiences. If it is possible to predict MPA film ratings using certain predictors in our dataset, it could help filmmakers and studios in making decisions about which films to produce and how to market them, as well as helping viewers to make informed choices about which movies to watch.

Dataset

- **Dataset source**: Kaggle (Grijalva, 2020)
- **Dataset Description**: Information on 7,600 movies spanning from 1986 to 2020 (when the dataset was last updated). The data was scrapped from IMDb.
- **Dataset dimensions**: 15 columns and 7,668 rows
- **Dataset Breakdown**:
    - Categorical variables:
        - Name: Title of the film
        - Rating: Rating of the movie (i.e. MPA rating)
        - Genre: Main genre of the movie
        - Director: Director of the movie
        - Writer: Main writer of the movie
        - Star: Main actor/actress of the movie
        - Country: Country of origin for the movie
        - Company: Production company of the movie
    - Numerical variables:
        - Budget: Available budget information for the movie
        - Gross: Revenue for the movie
        - Runtime: Duration of the movie
        - Year: Year of the movie release
        - Released: The exact release date for the movie
        - Score: IMDb user rating for the movie
        - Votes: The number of user votes for the movie

Workload Distribution

The workload breakdown was by exploratory question, as per follows:

- Analysis 1: Prem
- Analysis 2: Vardaan

- Analysis 3: Gavin
- Analysis 4: Sarah

Each individual was responsible for the completion of the analysis, the code, as well as the write-up. All members contributed to the writing of this report and the final presentation.

Results

The results for each of the analyses will be summarized below. Note that detailed RStudio outputs are not included for brevity and conciseness of the report. Please reference the associating Rmd file for detailed code and outputs.

Analysis 1: Can we predict the IMDB Score Rating based on budget, votes, gross, and genre as our predictors?

IMDb (Internet Movie Database) is one of the most popular movie databases on the internet, and its rating system is widely used to evaluate the quality and popularity of movies. IMDb ratings are based on a scale from 1 to 10, with higher scores indicating better ratings.

The IMDb rating for a movie is determined by the average of the ratings given by registered IMDb users. IMDb users can rate movies on a scale from 1 to 10, with the option to give half-point ratings.

It's important to note that IMDb ratings are not a definitive measure of a movie's quality or success. Ratings can be influenced by a variety of factors, such as the movie's budget, marketing, and critical reception. Additionally, IMDb ratings can be subject to manipulation, such as vote brigading or review bombing.

Thus, the first analysis aims to predict the IMDb score rating using a set of predictor variables that includes budget, votes, gross, and genre.

By attempting to answer this question, one aims to understand the relationship between these predictor variables and the target variable, IMDb score rating. This information can be used to build a predictive model that can be used to predict the IMDb score rating of a new movie based on its budget, votes, gross, and genre.

Sample of Dataset for this question:

row names	name	rating	genre	year	released	score	votes
5645	[Rec]²	R	Horror	2009	October 2, 2009 (Spain)	6.5	70000
1100	*batteries not included	PG	Comedy	1987	December 18, 1987 (United States)	6.7	32000
6907	10 Cloverfield Lane	PG-13	Action	2016	March 11, 2016 (United States)	7.2	300000
3453	10 Things I Hate About You	PG-13	Comedy	1999	March 31, 1999 (United States)	7.3	309000
394	10 to Midnight	R	Crime	1983	March 11, 1983 (United States)	6.3	7200
5345	10,000 BC	PG-13	Action	2008	March 7, 2008 (United States)	5.1	127000
2856	101 Dalmatians	G	Adventure	1996	November 27, 1996 (United States)	5.7	105000
3723	102 Dalmatians	G	Adventure	2000	November 22, 2000 (United States)	4.9	36000

IMDb Score Rating

Response Variable: IMDb score
Predictors:

- Votes
- Genre
- Budge
- Gross
- Runtime

IMDb (Internet Movie Database) is one of the most popular movie databases on the internet, and its rating system is widely used to evaluate the quality and popularity of movies. IMDb ratings are based on a scale from 1 to 10, with higher scores indicating better ratings.

The IMDb rating for a movie is determined by the average of the ratings given by registered IMDb users. IMDb users can rate movies on a scale from 1 to 10, with the option to give half-point ratings.

It's important to note that IMDb ratings are not a definitive measure of a movie's quality or success. Ratings can be influenced by a variety of factors, such as the movie's budget, marketing, and critical reception. Additionally, IMDb ratings can be subject to manipulation, such as vote brigading or review bombing.

Let’s introduce our first analytical question in which we will build a model to predict IMDB Score Rating.

Since most of our response variable is continuous and numerical it made sense to first use linear regression to predict our IMDB rating. First step would be to check if our dataset met the assumptions to implement linear regression. We started by checking for normality, as seen below in Figure 1.

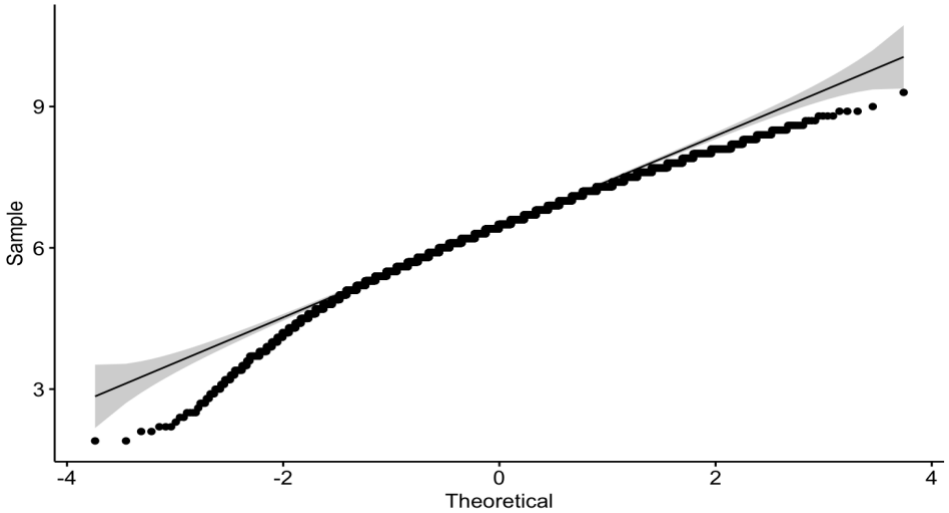


Figure 1: Normality Review for Analysis 1

Examining the following normality graph we can clearly see that all the points fall approximately along this reference line, so we can assume normality. Our first assumption is met. Now let’s check to see if our predictors

are independent or highly co-related. For this we will need to run a VIF test on our data, the results of this analysis can be seen in Figure 2.

	GVIF	Df	GVIF^(1/(2*Df))
votes	1.734552	1	1.317024
budget	2.624243	1	1.619952
gross	2.970591	1	1.723540
runtime	1.505703	1	1.227071
factor(genre)	1.653491	14	1.018123

Figure 2: VIF outcomes for Analysis 1

From the test we can examine that each of the predictor that we will use in our model has a VIF value lower than 2 which means they are variable and not highly co-related to each other, and it meets the standard to perform a linear regression.

First Order Model

Now that we have checked for our assumptions let's build our First Order Model with all the predictors.

$$\begin{aligned}\widehat{\text{score}} = {} & 4.29 + 0(\text{votes}) + 0(\text{budget}) + 0(\text{gross}) + \\ & 0.02(\text{runtime}) + 0.08(\text{factor(genre)}_{\text{Adventure}}) + 0.91(\text{factor(genre)}_{\text{Animation}}) + 0.58(\text{factor(genre)}_{\text{Biography}}) + \\ & 0.12(\text{factor(genre)}_{\text{Comedy}}) + 0.33(\text{factor(genre)}_{\text{Crime}}) + 0.34(\text{factor(genre)}_{\text{Drama}}) + 0.31(\text{factor(genre)}_{\text{Family}}) - \\ & 0.04(\text{factor(genre)}_{\text{Fantasy}}) - 0.24(\text{factor(genre)}_{\text{Horror}}) - 0.08(\text{factor(genre)}_{\text{Mystery}}) + 0.42(\text{factor(genre)}_{\text{Romance}}) + \\ & 0.22(\text{factor(genre)}_{\text{Sci-Fi}}) - 0.04(\text{factor(genre)}_{\text{Thriller}}) + 0(\text{factor(genre)}_{\text{Western}})\end{aligned}$$

The p-value for each variable indicates the probability of observing a coefficient as extreme as the estimated one. For the predictors that have p-value less than the significance level (0.05), we can reject the null hypothesis and conclude that the predictor variable has a statistically significant effect on the IMDb score rating. In our case that will be votes, budget, gross, runtime, and certain categorical values from genre.

We achieved a R-squared value of 0.4047 indicates that the model explains 40.47% of the variance in the IMDb score ratings. The adjusted R-squared value of 0.4028 takes into account the number of predictor variables in the model and is slightly lower than the multiple R-squared value.

The F-statistic tests the overall significance of the model by comparing the regression sum of squares to the residual sum of squares. The extremely low p-value (< 0.00000000000000022) indicates that the model is statistically significant and fits the data better than the null model (i.e., a model with only the intercept).

The residual standard error of 0.7443 indicates the average difference between the actual IMDb score ratings and the predicted IMDb score ratings by the model.

Now that we have built our initial model it's time to check interaction terms. I believe that the effect of one predictor variable on the response variable may depend on the level of another predictor variable. Specifically, I want to check for interaction terms since I have included multiple predictor variables in my model and there is reason to believe that the effect of one predictor on the response variable may differ at different levels of another predictor.

For example, we are predicting movie ratings based on both the budget, genre, gross, and runtime of the movie. The effect of budget on ratings may depend on the genre of the movie, and respectively for other factors. As result we will check for any interaction term between all the predictors in my linear model.

$$\begin{aligned}\widehat{\text{score}} = {} & 3.1 + 0(\text{votes}) + 0(\text{budget}) + 0(\text{gross}) + 0.54(\text{genre}_{\text{Adventure}}) + 1.76(\text{genre}_{\text{Animation}}) + \\ & 2.39(\text{genre}_{\text{Biography}}) + 0.72(\text{genre}_{\text{Comedy}}) + 0.79(\text{genre}_{\text{Crime}}) + 1.27(\text{genre}_{\text{Drama}}) + 2.25(\text{genre}_{\text{Family}}) + 0.68(\text{genre}_{\text{Fantasy}}) + \\ & 0.15(\text{genre}_{\text{Horror}}) + 1.3(\text{genre}_{\text{Mystery}}) + 0.44(\text{genre}_{\text{Romance}}) + 1.55(\text{genre}_{\text{Sci-Fi}}) - 3.06(\text{genre}_{\text{Thriller}}) + 1.71(\text{genre}_{\text{Western}}) + \\ & 0.03(\text{runtime}) + 0(\text{votes} \times \text{budget}) + 0(\text{votes} \times \text{gross}) + 0(\text{votes} \times \text{genre}_{\text{Adventure}}) + 0(\text{votes} \times \text{genre}_{\text{Animation}}) + 0(\text{votes} \times \text{genre}_{\text{Biography}}) + \\ & 0(\text{votes} \times \text{genre}_{\text{Comedy}}) + 0(\text{votes} \times \text{genre}_{\text{Crime}}) + 0(\text{votes} \times \text{genre}_{\text{Drama}}) + 0(\text{votes} \times \text{genre}_{\text{Family}}) + 0(\text{votes} \times \text{genre}_{\text{Fantasy}}) + 0(\text{votes} \times \text{genre}_{\text{Horror}}) + \\ & 0(\text{votes} \times \text{genre}_{\text{Mystery}}) + 0(\text{votes} \times \text{genre}_{\text{Romance}}) + 0(\text{votes} \times \text{genre}_{\text{Sci-Fi}}) + 0(\text{votes} \times \text{genre}_{\text{Thriller}}) + 0(\text{votes} \times \text{genre}_{\text{Western}}) + 0(\text{votes} \times \text{runtime}) + \\ & 0(\text{budget} \times \text{gross}) + 0(\text{budget} \times \text{genre}_{\text{Adventure}}) + 0(\text{budget} \times \text{genre}_{\text{Animation}}) + 0(\text{budget} \times \text{genre}_{\text{Biography}}) + 0(\text{budget} \times \text{genre}_{\text{Comedy}}) + 0(\text{budget} \times \text{genre}_{\text{Crime}}) + \\ & 0(\text{budget} \times \text{genre}_{\text{Drama}}) + 0(\text{budget} \times \text{genre}_{\text{Family}}) + 0(\text{budget} \times \text{genre}_{\text{Fantasy}}) + 0(\text{budget} \times \text{genre}_{\text{Horror}}) + 0(\text{budget} \times \text{genre}_{\text{Mystery}}) + 0(\text{budget} \times \text{genre}_{\text{Romance}}) + \\ & 0(\text{budget} \times \text{genre}_{\text{Sci-Fi}}) + 0(\text{budget} \times \text{genre}_{\text{Thriller}}) + NA(\text{budget} \times \text{genre}_{\text{Western}}) + 0(\text{budget} \times \text{runtime}) + 0(\text{gross} \times \text{genre}_{\text{Adventure}}) + 0(\text{gross} \times \text{genre}_{\text{Animation}}) + \\ & 0(\text{gross} \times \text{genre}_{\text{Biography}}) + 0(\text{gross} \times \text{genre}_{\text{Comedy}}) + 0(\text{gross} \times \text{genre}_{\text{Crime}}) + 0(\text{gross} \times \text{genre}_{\text{Drama}}) + 0(\text{gross} \times \text{genre}_{\text{Family}}) + 0(\text{gross} \times \text{genre}_{\text{Fantasy}}) + \\ & 0(\text{gross} \times \text{genre}_{\text{Horror}}) + 0(\text{gross} \times \text{genre}_{\text{Mystery}}) + 0(\text{gross} \times \text{genre}_{\text{Romance}}) + 0(\text{gross} \times \text{genre}_{\text{Sci-Fi}}) + 0(\text{gross} \times \text{genre}_{\text{Thriller}}) + NA(\text{gross} \times \text{genre}_{\text{Western}}) + \\ & 0(\text{gross} \times \text{runtime}) - 0.01(\text{genre}_{\text{Adventure}} \times \text{runtime}) - 0.01(\text{genre}_{\text{Animation}} \times \text{runtime}) - 0.02(\text{genre}_{\text{Biography}} \times \text{runtime}) + 0(\text{genre}_{\text{Comedy}} \times \text{runtime}) + 0(\text{genre}_{\text{Crime}} \times \text{runtime}) - \\ & 0.01(\text{genre}_{\text{Drama}} \times \text{runtime}) + NA(\text{genre}_{\text{Family}} \times \text{runtime}) - 0.01(\text{genre}_{\text{Fantasy}} \times \text{runtime}) + 0(\text{genre}_{\text{Horror}} \times \text{runtime}) - 0.01(\text{genre}_{\text{Mystery}} \times \text{runtime}) + 0(\text{genre}_{\text{Romance}} \times \text{runtime}) - \\ & 0.01(\text{genre}_{\text{Sci-Fi}} \times \text{runtime}) + 0.02(\text{genre}_{\text{Thriller}} \times \text{runtime}) + NA(\text{genre}_{\text{Western}} \times \text{runtime})\end{aligned}$$

After adding the interaction terms, the Adjusted R^2 value improved even more after adding our interaction terms to the model. The R^2 value jumped from 0.40 (First Model) to 0.45 (Interaction model).

Even though we achieved a R-squared value of 0.45, I was still not satisfied. For that reason I decided to check the if my data had many outliers and if that impacted our accuracy. To test this theory I used the cook's distance to find out those outlier and remove them from my data – see Figure 3.

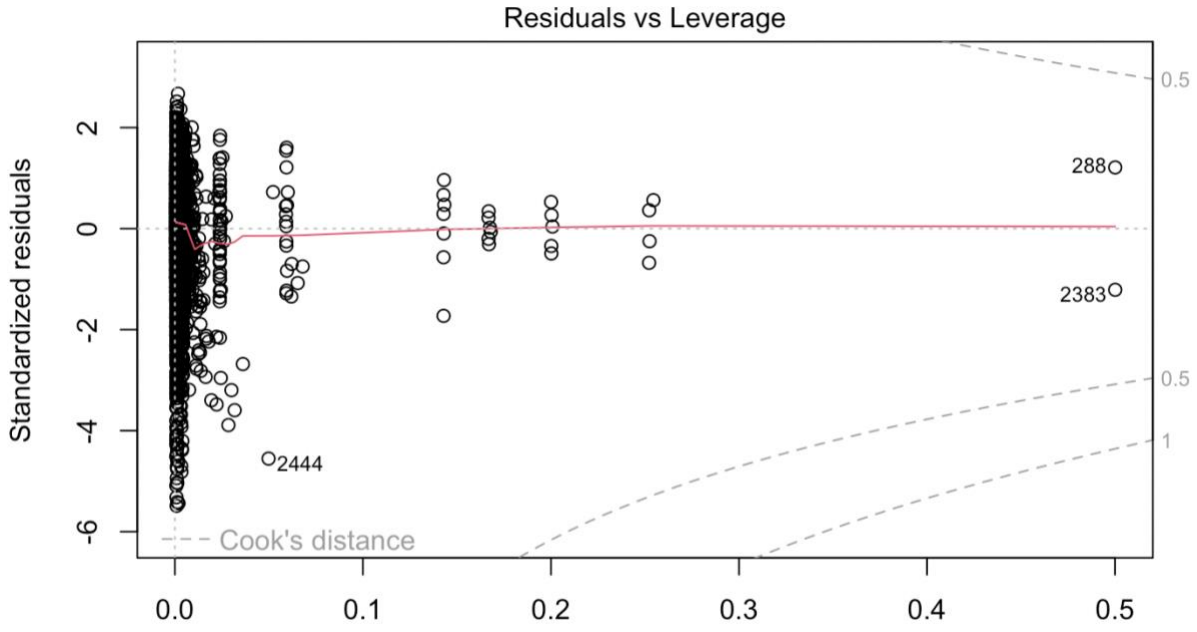


Figure 3: Outlier review using Cook's distance for Analysis 1

Analyzing my cook's distance graph, it shows several observations with large values of Cook's distance, these observations are outliers or have a disproportionate impact on the regression model. In this case, it was important to examine these observations carefully and should be excluded from the analysis. After removing these outliers I build my final model and here are the results.

$$\widehat{\text{score}} = 3.59 + 0(\text{votes}) + 0(\text{budget}) + 0(\text{gross}) + 0.02(\text{runtime}) + 0.21(\text{genre}_{\text{Adventure}}) + 1.05(\text{genre}_{\text{Animation}}) + 0.54(\text{genre}_{\text{Biography}}) + 0.12(\text{genre}_{\text{Comedy}}) + 0.33(\text{genre}_{\text{Crime}}) + 0.32(\text{genre}_{\text{Drama}}) + 0.35(\text{genre}_{\text{Family}}) - 0.06(\text{genre}_{\text{Fantasy}}) - 0.24(\text{genre}_{\text{Horror}}) - 0.18(\text{genre}_{\text{Mystery}}) + 0.3(\text{genre}_{\text{Romance}}) + 0.24(\text{genre}_{\text{Sci-Fi}}) + 0.16(\text{genre}_{\text{Thriller}}) + 0(\text{votes} \times \text{budget}) + 0(\text{votes} \times \text{runtime}) + 0(\text{budget} \times \text{gross}) + 0(\text{votes} \times \text{gross})$$

After removing the outliers we can see that our Adjusted R2 value improved further. The R2 value jumped from 0.45 (Interaction Model) to 0.50 (final model). This means that we achieved a R-squared value of 0.5 which indicates that the model explains 50% of the variance in the IMDb score ratings.

Regression Tree

Even though we improved our linear model, regression trees can also be a useful alternative to linear regression in certain cases, including predicting IMDb score ratings. Here are a few reasons why I considered using a regression tree instead of linear regression:

Non-linearity: Linear regression assumes a linear relationship between the predictors and the response variable. However, this assumption may not always hold in practice. Regression trees are a non-parametric method, meaning they make no assumptions about the underlying distribution of the data. This allows them to capture non-linear relationships between the predictors and the response variable.

Interactions: Linear regression assumes that the effect of each predictor on the response variable is independent of the other predictors. However, in practice, there may be interactions between predictors that affect the response variable. Regression trees can capture interactions between predictors by splitting the data based on combinations of predictor values.

Easy to interpret: The output of a regression tree is a decision tree, which is easy to interpret and can provide insight into the relationships between the predictors and the response variable. In contrast, the coefficients of a linear regression model can be difficult to interpret, particularly if there are interactions or non-linear relationships between the predictors and the response variable.

Robustness: Regression trees are relatively robust to outliers and missing data, whereas linear regression can be sensitive to these issues.

Overall, while linear regression is a powerful and widely used method, there are situations where regression trees may be a better choice for predicting IMDb score ratings. As a result built a regression tree model to predict IMDb score ratings.

First, we divided our data into 80/20 train test split. Then we used the tree function in R to build our regression tree.

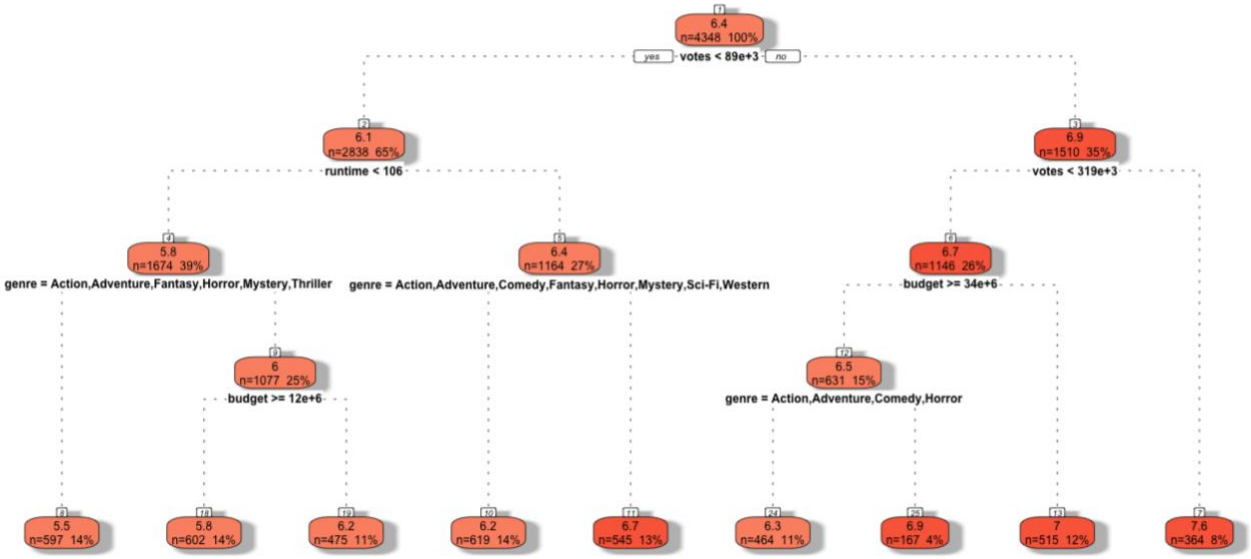


Figure 4: Regression tree for Analysis 1

The regression tree model uses the predictor variables "votes," "runtime," “genre,” and "budget" to create a tree structure that predicts IMDb score ratings for new movies based on their values for those variables. The splits are based on the values of predictor variables. Each line contains a condition that determines whether a given observation will follow the left or right branch of the tree.

Next I calculated the MSE of my tree model and achieved rate of about 0.60. This means that the tree model which suggests that the regression tree model has a moderate level of error in its predictions.

In order to decide if I should prune my tree, I used cross-validation to estimate the performance of the tree at different levels of pruning. I used the cv.tree function in R, which performs k-fold cross-validation on a decision tree and provides a measure of the prediction error at each level of pruning.

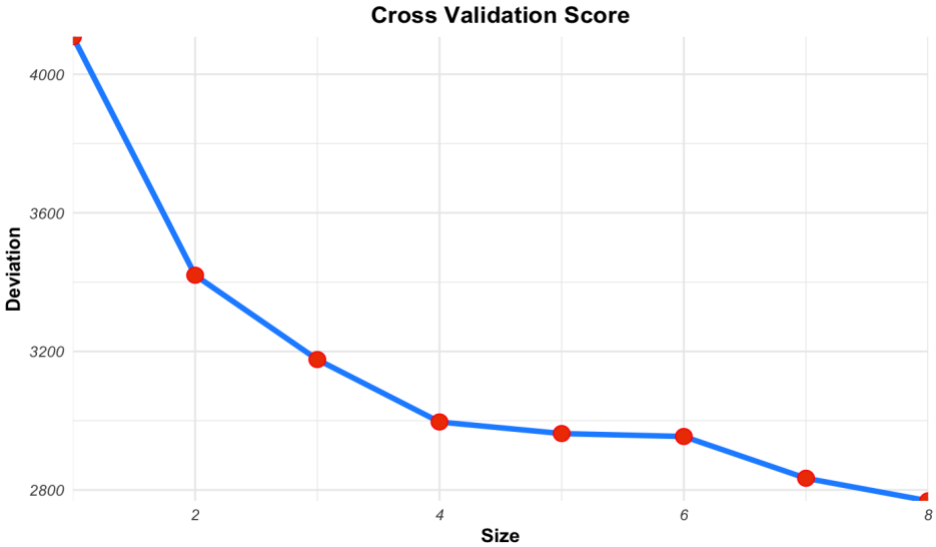


Figure 5: Cross-validation error as a function of the number of terminal nodes for Analysis 1

Looking at the cross-validation graph (Figure 5) we need to prune our tree at node 8.

From our pruned tree we calculated an MSE of 0.6 which means our pruning did not have a significant impact of our tree model.

K-Fold Cross Validation

The k-fold cross-validation method is used for model evaluation in machine learning. It involves dividing a dataset into k subsets, or "folds", of approximately equal size. The model is then trained on k-1 of these folds and evaluated on the remaining fold. This process is repeated k times, with each fold serving as the evaluation set once.

I wanted to use this method to compare the performance of both the linear regression and regression tree methods, to see which one performed better. I used K=10 to subset the data in to 10 folds of equal size. The error metric obtained from the 10-fold cross-validation, for my Linear Regression model gave me a value of 0.71 indicates the average prediction error across the 10 folds. This means that on average, the model's predictions deviate from the true values by 0.71 units. And for Regression tree it gave me a value of 0.78. As a result our linear regression model will predict the IMDb score rating much accurately then my regression tree model.

Practical Application

In order to challenge my model I decided to predict a movie IMDb score that came out recently in 2021 (not included in our original dataset). The movie I chose was ***Fast & Furious 9***, and below (Table 1) are the prediction results.

Table 1: Prediction vs Actual for Analysis 1

	Actual	*Predict*
Linear	***5.2***	***6.0***
Tree	***5.2***	***6.5***

The actual IMDB Score for Fast and Furious 9 is 5.2 and for our linear model predicted IMDb score of 6.2 while our Regression Tree predicted a score of 6.5. From these results we can certainly tell that our Linear Model outperformed our tree model by a moderate margin of 7%.

Analysis 2: Using logistic regression to model good/bad movies and multinomial regression to model genre based on the other variables

Logistic Regression to predict whether a movie is Good or Bad

Logistic regression is a popular statistical technique used in machine learning to analyze and model binary classification problems. In the case of predicting good and bad movies, logistic regression can be a powerful tool for identifying the key factors that contribute to a movie's success or failure.

The objective of a logistic regression model for predicting good and bad movies is to predict the probability of a movie being good or bad based on a set of input variables. These variables could include factors such as genre, budget, critical reception, and box office performance.

The logistic regression model aims to estimate the probability that a movie belongs to the "good" category based on the values of the input variables. This probability can then be used to make a binary prediction - if the probability is above a certain threshold, the model will predict the movie as "good", and if it is below the threshold, the model will predict the movie as "bad".

The Data

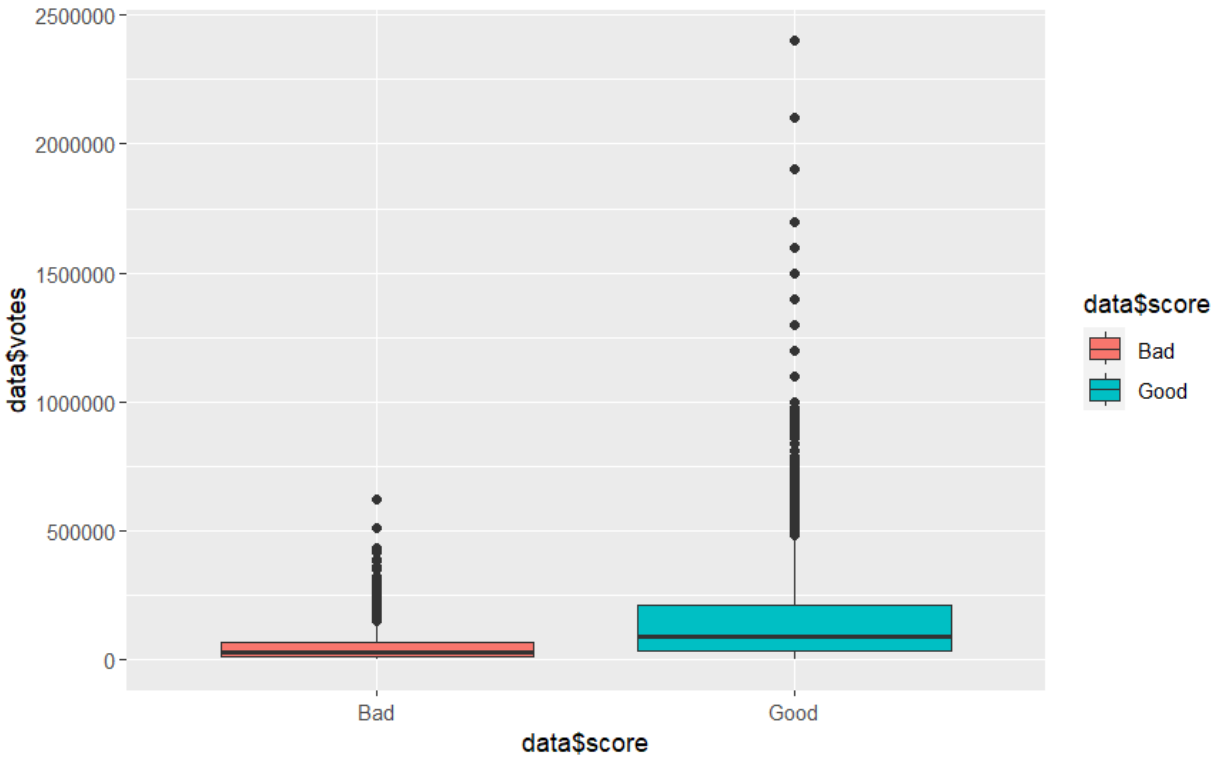


Figure 6: Boxplot comparing the distribution of votes for "Good" and "Bad" movies for Analysis 2

Based on the boxplots comparing the distribution of votes received for good and bad movies, it appears that good movies tend to receive significantly more votes than bad movies. This could be interpreted as an indication that good movies are more popular and have a wider audience than bad movies. It may also suggest that good movies are more likely to be seen by a larger number of people, which could be due to a number of factors such as positive reviews, effective marketing, and strong word-of-mouth recommendations.

This disparity in the number of votes received between good and bad movies could have implications for filmmakers, investors, and movie studios. It may suggest that in order to create a successful movie that resonates with a wide audience, it is important to focus on producing high-quality content that is well-received by critics and

audiences alike. This may involve investing in better scripts, hiring talented directors and actors, and providing adequate marketing and distribution resources to ensure that the movie reaches a wide audience.

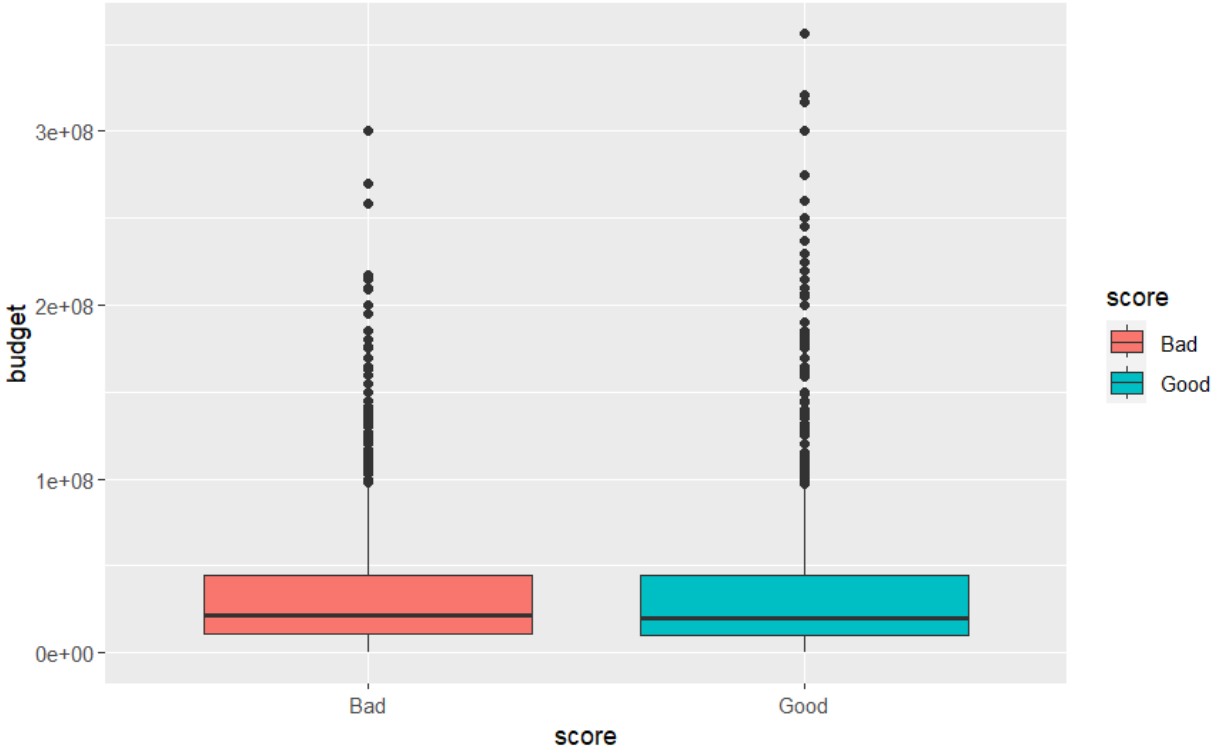


Figure 7: Boxplot comparing the distribution for budget for "Good" and "Bad" movies for Analysis 2

On the other hand, it appears that there is not a significant difference between the two groups. This could be interpreted as an indication that the budget of a movie may not be a significant factor in determining whether it will be successful or not.

There are several possible explanations for this finding. One possibility is that there are many other factors besides budget that can contribute to a movie's success, such as the quality of the script, the acting, the direction, and the marketing. Another possibility is that the budget of a movie may not be a direct predictor of its quality or success, as some lower-budget movies have achieved critical and commercial success, while some high-budget movies have failed to impress audiences.

Assumptions

```
           GVIF Df GVIF^(1/(2*Df))
rating 1.975956 10        1.034639
genre  2.421290 14        1.032086
votes  2.625761  1        1.620420
budget 2.698334  1        1.642661
gross  3.019564  1        1.737689
```

Figure 8: VIF outcomes for Analysis 2

The output from the multicollinearity test (see Figure 8) using Variance Inflation Factor (VIF) suggests that there is some level of multicollinearity among the predictor variables in your model. Specifically, the VIF values for some of the variables are greater than 2, which indicates that there may be a significant correlation between these variables.

In particular, the variables "budget" and "gross" have the highest VIF values, which suggests that they may be highly correlated with each other. This is not surprising, as budget and gross revenue are both measures of financial performance for a movie, and they are likely to be related to each other in some way.

The VIF values for "genre" and "votes" are also somewhat high, although they are lower than the values for "budget" and "gross". This suggests that there may be some level of correlation between these variables as well.

However, the VIF value for "rating" is below 2, which suggests that this variable is not highly correlated with any of the other predictor variables in the model.

Overall, the results of the VIF test suggest that there may be some level of multicollinearity among the predictor variables in your model. However, since the VIF value is less than 5 for all the variables we are going ahead with the following variables for the next assumption test.

Chi-Squared Test for Independence

Null hypothesis (H_0): There is no significant association between the two categorical variables.
Alternative hypothesis (H_a): There is a significant association between the two categorical variables.

```
        Pearson's Chi-squared test

data:  table(residuals, data$scorepred)
X-squared = 5435, df = 5428, p-value = 0.4707
```

Figure 9: Chi-Square test outcome for Analysis 2

In this case, a p-value of 0.4707 suggests that there is no significant association between the two categorical variables being tested at the 0.05 significance level. In other words, there is not enough evidence to reject the null hypothesis that there is no association between the two variables in favor of the alternative hypothesis that there is an association between the two variables. Since the p-value is 0.47, there is not enough evidence to suggest a significant association between the independent and dependent variables. But we proceeded with the model regardless in an attempt to see the metrics from our model.

Cooks Distance

Cook's distance is a statistical method used to identify influential points, or outliers, in a regression analysis. In the context of a movie dataset, Cook's distance can be used to detect observations that have a disproportionate impact on the results of a logistic regression model, potentially skewing the model's predictions. By removing these influential points, we can improve the accuracy of our model and make more reliable predictions about which movies are likely to be successful or unsuccessful. Cook's distance works by measuring the change in the regression coefficients when each observation is removed from the analysis and identifying those observations that cause the greatest change. By removing these influential points, we can obtain a more accurate representation of the underlying relationship between the predictors and the outcome variable, and produce a more reliable logistic regression model for predicting the success of movies.

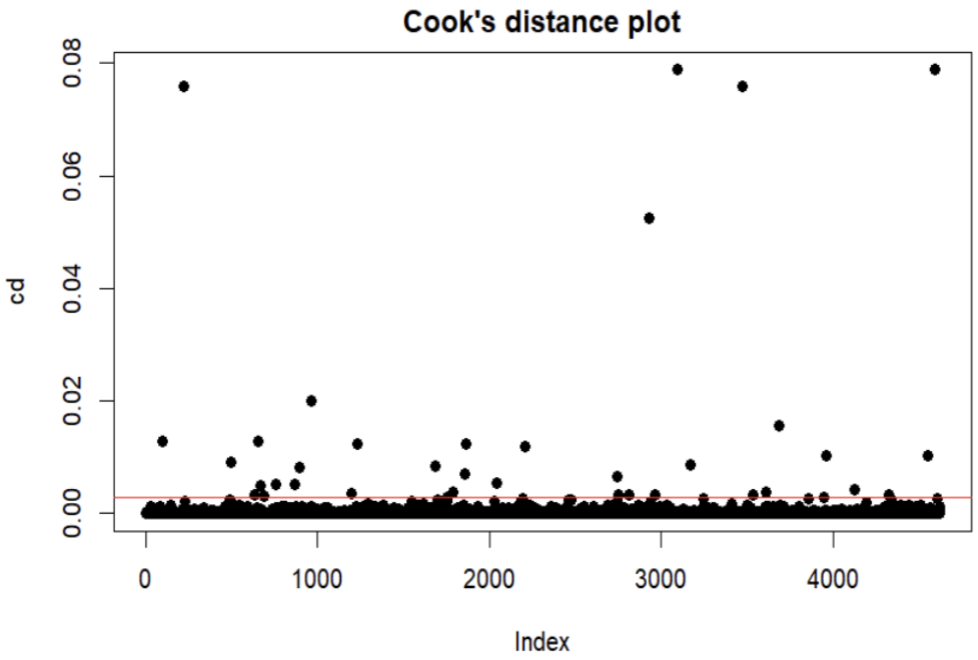


Figure 10: Outlier analysis using Cook's distance for Analysis 2

The red line at 0.003 mark represents the threshold for the Cook's distance to remove the outliers. By removing these influential points, we can improve the accuracy and reliability of our logistic regression model and produce more accurate predictions about which movies are likely to be successful or unsuccessful.

Initial Model

For my initial model we are utilizing predictors variables as Genre, Budget, Gross, Rating and votes. We are not utilizing variables like directors, writers and stars in the logistic regression model due to the possibility of factors in unknown factors in the test that does not have an equivalent training set data point.

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \text{rating} + \beta_2 \text{genre} + \beta_3 \text{votes} + \beta_4 \text{budget} + \beta_5 \text{gross}$$

From the above model, we were getting a misclassification rate of 25.98%, but based on the P values provided from the summary() function for the model showcases the rating variable to be insignificant along with all of its interaction terms, so we decided to enhance the model and remove the rating variable from our analysis.

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \text{genre} + \beta_2 \text{votes} + \beta_3 \text{budget} + \beta_4 \text{gross}$$

The enhanced model described by the snippet above gave us a misclassification rate of 22.8% which is a slight improvement over our initial model.

Multinomial Regression to predict the Genres for a movie

In order to create a multinomial logistic regression model to predict the genre of a movie, we first need to prepare our data. One important step is to remove any missing values in the dataset, as these can cause errors in the model. We can use the na.omit() function in R to remove any rows with missing values.

Next, we may want to remove any genres that have a small number of data points, as these may not provide enough information to accurately predict the genre. We can use the table() function in R to count the number of movies in each genre, and then filter out any genres that have a low count. The specific threshold for what constitutes a "small" number of data points may depend on the size of the dataset and the number of genres present.

After cleaning the data, we can use the multinom() function in the nnet package to fit a multinomial logistic regression model to the data. This function allows us to model the relationship between the predictor variables and the outcome variable (the movie's genre).

In R, the multinom() function from the nnet package is a popular choice for fitting multinomial logistic regression models. This function uses the base logit function as the link function to model the relationship between the predictor variables and the outcome variable, which is the categorical variable representing the different movie genres.

We are specifically using multinom function due to the size of our model. Considering the large amount of interaction with multiple genres with each of the predictor variables, the model weight was a bit to large to compute and it was running out of memory, to fix this we utilized the multinom function with the MaxWeight parameter which allows us to tweak the limits of the weights set by default for the model by the function.

$$\ln\left(\frac{P(\text{genre}i)}{P(\text{genre}K)}\right) = \beta 0i + \beta 1i\text{runtime} + \beta_{2i}\text{budget} + \beta_{3i}\text{gross} + \beta_{4i}\text{rating} + \beta_{5i}\text{score}$$

```
# weights:  120 (98 variable)
initial  value 9459.379573
iter  10 value 8852.384898
iter  20 value 8502.042604
iter  30 value 8058.407876
iter  40 value 7824.286971
iter  50 value 7634.823772
iter  60 value 7593.853193
iter  70 value 7591.749241
iter  70 value 7591.749193
iter  80 value 7591.202490
final  value 7591.202204
converged
```

Figure 11: Convergence for Analysis 2 using the multinom function

Our model started with an initial value of 9459 and converged at a relatively high value of 7591 (see Figure 11). This could suggest that the model has a lot of complexity to handle and is potentially not the best predictor variable for the model. But since the model converged, we proceeded with further analysis.

After predicting the data that was split in our test set, we received the following prediction table, seen in Figure 12.

	Action	Adventure	Animation	Biography	Comedy	Crime	Drama	Horror
Action	124	13	20	10	54	18	27	3
Adventure	0	0	0	0	0	0	0	0
Animation	0	0	4	0	0	0	0	0
Biography	0	0	0	0	0	0	0	0
Comedy	81	27	9	33	166	54	111	36
Crime	0	0	0	0	0	0	0	0
Drama	3	1	0	0	5	1	1	1
Horror	0	0	0	0	0	0	0	0

Figure 12: Prediction table for Analysis 2 using the multinom function

Based on the prediction table we were getting a misclassification rate of 63.2%. The one inference we got to realize as soon as we ran the prediction was the disparity between the predictions in genres like Action and Comedy compared with all other Genres. To check the reason for this difference, we got the counts of our genres. After getting the counts we found that Action and Comedy constitute more than 40% of the data described by our dataset.

To try to enhance the model, we decided to remove all Genres with less than 1000 rows of data, this left us with three genres- Action, Comedy, and Drama. After remodeling the data, the following was observed:

```
# weights:  42 (26 variable)
initial  value 3529.841283
iter  10 value 3062.198676
iter  20 value 2924.638898
iter  20 value 2924.638897
iter  30 value 2891.336520
iter  40 value 2874.791508
iter  50 value 2861.937862
iter  60 value 2859.977644
iter  70 value 2859.634944
iter  70 value 2859.634918
final  value 2859.603094
converged
```

Figure 13: Convergence, second attempt, for Analysis 2 using the multinom function

The value for each of our iterations started at a much lower value and converged with lesser iterations than our basic model. From the enhanced model we got a prediction table as the following:

	Action	Comedy	Drama
Action	116	36	29
Comedy	87	155	58
Drama	24	27	36

Figure 14: Prediction table, second attempt, for Analysis 2 using the multinom function

From Figure 14, we found the predictions to be better defined than the initial prediction table from the basic model. The misclassification for the enhanced model came out as 45.9%. This was much better than our initial model but considering the lack of all genres along with choosing the best predictor variables for this model, the misclassification rate is on the higher end.

In the attempt to describe the data through a different modeling technique with proceeded with a classification tree to check if the data we are modeling can be described with a tree compared to the multinomial model.

After building our tree on the dataset, we get the following tree structure:

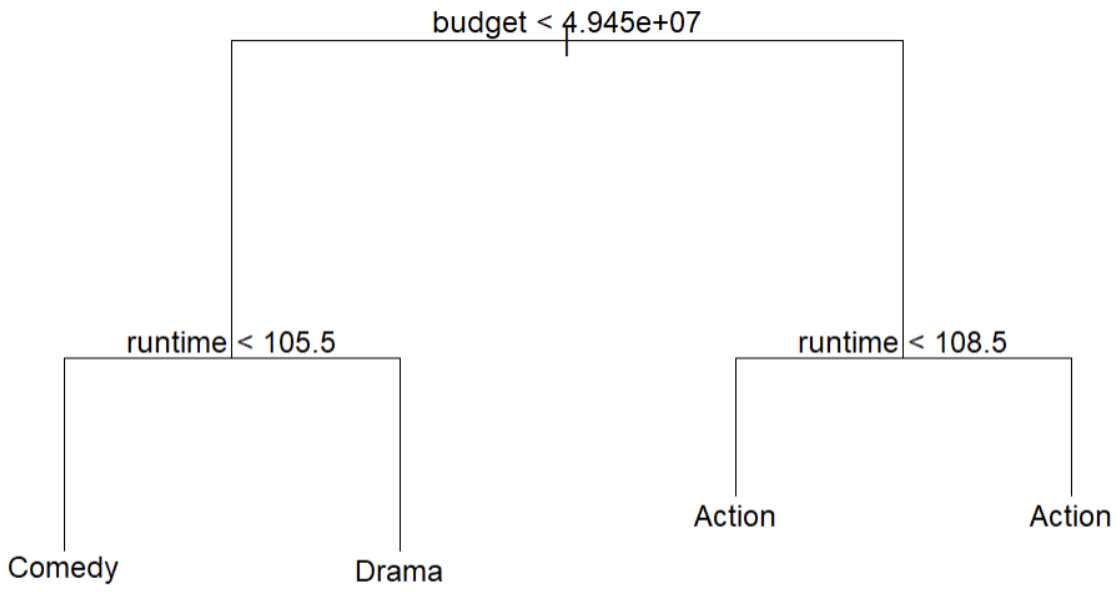


Figure 15: Regression tree for Analysis 2

From the tree, we can see a similar trend in terms of the terminal nodes and how only the genres with the larger number of records are described in this tree. Moreover, out of all the variables used for the prediction only two are considered by the tree as significant for the prediction process.

Finally, the misclassification rate from this model comes out as 63.4% which is very similar to our initial model utilizing multinomial regression. This suggests that neither of the two modeling techniques was able to account for the variability of all the genres based on the predictor variable that we are using. Additionally, neither of the two models is able to account for the imbalance in the datasets very well. Overall, while our enhance model predicts the model at a better rate compared to the other models, this dataset is not very well suited for the type of predictions we are trying to make.

Analysis 3a: Can we predict gross income from the data?

For predicting movie gross income, we opted to first try to convert some of the other variables that we had into meaningful factors. In the data, the columns “star” and “director” are both just columns with the names of the various stars and directors of the movies; data that is useless for our purposes of constructing models.

We used web scraping to find the respective genders of all the directors and stars and merged it with the main data frame.

Linear regression seemed the most suitable for predicting gross income.

```
Call:
lm(formula = gross ~ ., data = combinedSet)

Residuals:
     Min       1Q    Median       3Q      Max
-878057577 -38040742  -1360152  29536995 1863514949
```

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -2.171e+08  3.117e+08  -0.696 0.486210
ratingApproved     5.249e+07  1.142e+08   0.460 0.645843
ratingG            1.920e+07  3.352e+07   0.573 0.566826
ratingNC-17       -1.489e+07  4.384e+07  -0.340 0.734118
ratingNot Rated    6.055e+06  3.578e+07   0.169 0.865633
ratingPG           1.703e+07  3.188e+07   0.534 0.593256
ratingPG-13       -6.255e+06  3.178e+07  -0.197 0.844003
ratingR           -1.994e+07  3.170e+07  -0.629 0.529388
ratingTV-MA        2.941e+08  8.371e+07   3.514 0.000446 ***
ratingUnrated     -1.428e+07  4.132e+07  -0.346 0.729609
ratingX           -3.071e+07  1.140e+08  -0.269 0.787638
year               1.044e+05  1.544e+05   0.676 0.498784
score              6.689e+06  1.920e+06   3.484 0.000499 ***
votes              3.690e+02  1.041e+01  35.452  < 2e-16 ***
budget             2.638e+00  4.703e-02  56.081  < 2e-16 ***
runtime           -3.912e+05  1.003e+05  -3.898 9.81e-05 ***
star_genderMale    -2.005e+07  3.547e+06  -5.652 1.66e-08 ***
director_genderMale -2.568e+06  6.467e+06  -0.397 0.691352
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 109500000 on 5452 degrees of freedom
Multiple R-squared:  0.6609,             Adjusted R-squared:  0.6598
F-statistic:   625 on 17 and 5452 DF,  p-value: < 2.2e-16
```

This first model showed we could remove ‘year’ and ‘score’ as well as ‘director_gender’ from the equation.

First, we checked for multicollinearity with a ‘vif’ call.

```
                GVIF Df GVIF^(1/(2*Df))
rating          1.307902 10       1.013512
year            1.216896  1       1.103130
score           1.562433  1       1.249973
votes           1.668304  1       1.291629
budget          1.714124  1       1.309246
runtime         1.467633  1       1.211459
star_gender     1.083913  1       1.041111
director_gender 1.050769  1       1.025070
```

Since none of the values are particularly high (all values are < 5), we can assume there is no significant multicollinearity.

Next, we created the next iteration of the linear regression model:

```
Call:
lm(formula = gross ~ rating + score + votes + budget + runtime +
   star_gender, data = combinedSet)

Residuals:
     Min       1Q    Median       3Q      Max
-879481815  -37631469   -1271467   29184691 1862265312

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)    -1.033e+07  3.436e+07  -0.301 0.763728
ratingApproved   5.038e+07  1.142e+08   0.441 0.659019
ratingG         1.877e+07  3.350e+07   0.560 0.575306
ratingNC-17    -1.492e+07  4.384e+07  -0.340 0.733585
ratingNot Rated  6.988e+06  3.576e+07   0.195 0.845073
ratingPG        1.679e+07  3.187e+07   0.527 0.598403
ratingPG-13    -5.916e+06  3.178e+07  -0.186 0.852314
ratingR        -1.981e+07  3.169e+07  -0.625 0.531987
ratingTV-MA     2.961e+08  8.365e+07   3.539 0.000405 ***
ratingUnrated   -1.431e+07  4.131e+07  -0.346 0.729034
ratingX        -3.195e+07  1.140e+08  -0.280 0.779223
score          6.751e+06  1.918e+06   3.519 0.000437 ***
votes          3.694e+02  1.038e+01  35.567  < 2e-16 ***
budget          2.646e+00  4.486e-02  58.995  < 2e-16 ***
runtime        -3.965e+05  9.987e+04  -3.970 7.28e-05 ***
star_genderMale -2.065e+07  3.455e+06  -5.978 2.40e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 109500000 on 5454 degrees of freedom
Multiple R-squared:  0.6608,            Adjusted R-squared:  0.6599
F-statistic: 708.5 on 15 and 5454 DF,  p-value: < 2.2e-16
```

Our new model has an adjusted R-squared of 0.6599 and a RMSE of 109500000. We plotted the residuals to check our assumptions of homoscedasticity.

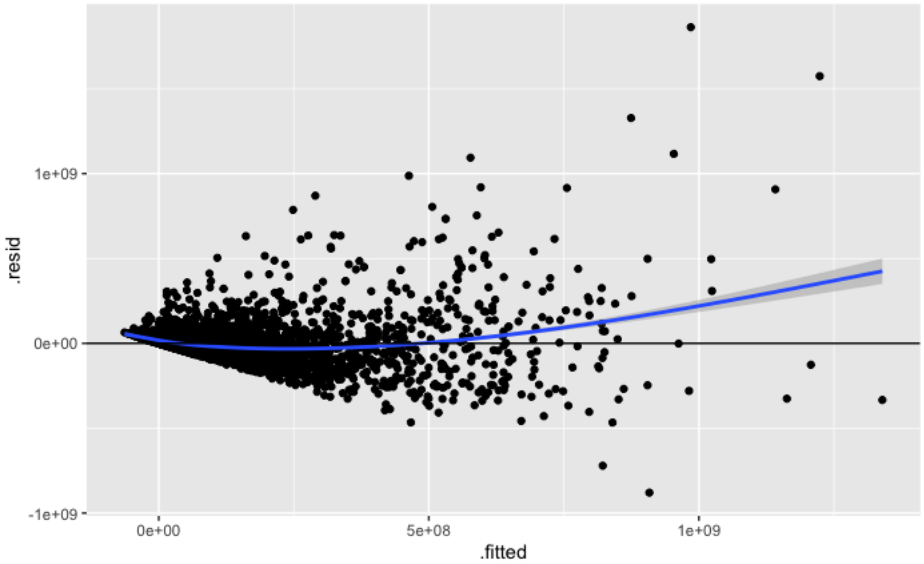


Figure 16: Homoscedasticity review for Analysis 3

Since the points are cone-shaped, our model does not pass the assumptions of homoscedasticity. This is further corroborated by the non-normal distribution of the residuals.

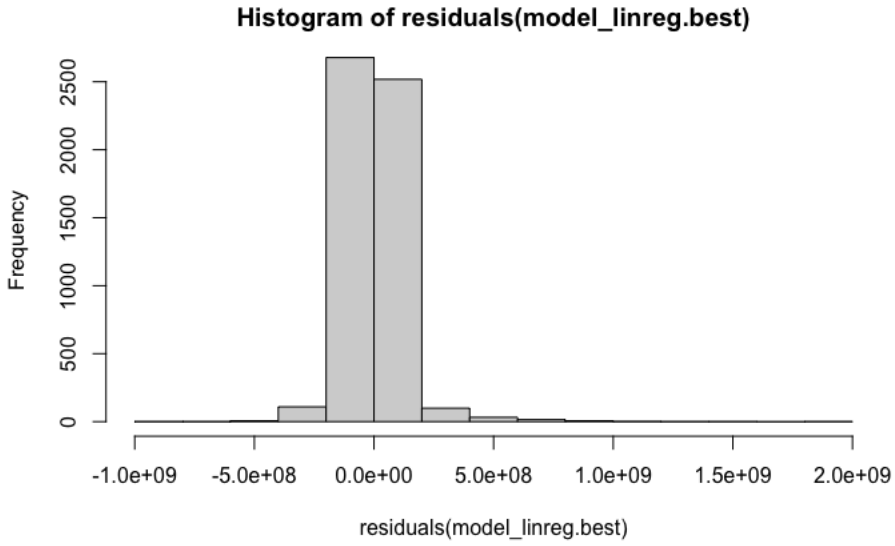


Figure 17: Residuals review for Analysis 3

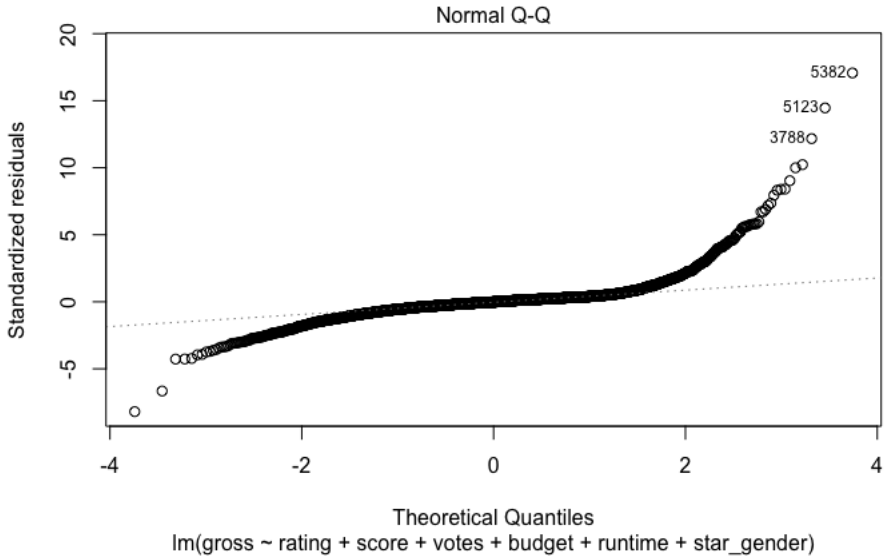


Figure 18: Q-Q plot for Analysis 3

Finally, we can see that the Q-Q plot of the residuals does not conform to the line, suggesting also that our assumption of normality is also violated.

To further investigate, we plotted the Cook's distances to observe any outliers.

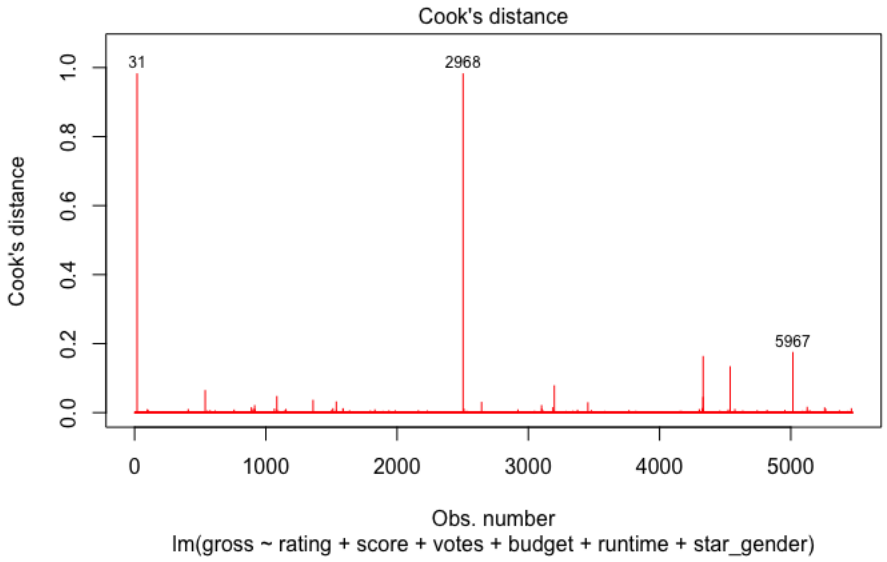


Figure 19: Outlier review using Cook's distance for Analysis 3

Although there appears to be two outliers, they both have a Cook's distance < 1, so we keep it in the data. At any rate, the removal or addition of two data points, will not fix the assumptions our model has failed.

Overall, although we could create a model with an acceptable adjusted R-squared, because our data does not conform at all to the assumptions required, we do not recommend modelling this data with linear regression technique.

Analysis 3b: Can we predict star_gender?

In the previous section, we converted the 'star' and 'director' column into their respective genders. After creating the linear regression model, we saw that 'star_gender' was a significant predictor of the gross income of the movie, but not 'director_gender', which led us to think about ways that we could investigate the difference.

We started by splitting our data 75-25, training to testing data.

```
set.seed(2023)
train_index=sample(1:nrow(combinedSet),3/4*nrow(combinedSet))
test=combinedSet[-train_index,]
train = combinedSet[train_index,]
```

Logistic Regression

From this training data, we started with creating a logistic regression model.

```
Call:
glm(formula = factor(director_gender) ~ ., family = binomial,
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3441   0.2073   0.2540   0.3448   0.9506

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  8.082e+01  5.401e+02   0.150 0.881045
```

```
ratingApproved   7.217e-01  1.552e+03   0.000 0.999629
ratingG        -1.227e+01  5.399e+02  -0.023 0.981865
ratingNC-17      1.973e-01  6.755e+02   0.000 0.999767
ratingNot Rated -1.267e+01  5.399e+02  -0.023 0.981279
ratingPG       -1.274e+01  5.399e+02  -0.024 0.981179
ratingPG-13     -1.287e+01  5.399e+02  -0.024 0.980984
ratingR        -1.246e+01  5.399e+02  -0.023 0.981581
ratingTV-MA      6.683e-01  1.160e+03   0.001 0.999540
ratingUnrated   -1.338e+01  5.399e+02  -0.025 0.980221
ratingX        -7.672e-01  1.552e+03   0.000 0.999606
year           -3.290e-02  6.907e-03  -4.763  1.9e-06 ***
score          -4.072e-02  9.045e-02  -0.450 0.652587
votes           2.507e-06  9.616e-07   2.607 0.009138 **
budget          1.412e-08  3.792e-09   3.725 0.000195 ***
gross          -1.220e-09  8.235e-10  -1.482 0.138402
runtime        -6.227e-03  5.089e-03  -1.224 0.221091
star_genderMale  1.466e+00  1.402e-01  10.458  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1888.8  on 4101  degrees of freedom
Residual deviance: 1668.8  on 4084  degrees of freedom
AIC: 1704.8

Number of Fisher Scoring iterations: 14
```

From there, we reduced the number of predictors in the model.

```
Call:
glm(formula = star_gender ~ year + score + votes + budget + gross +
    director_gender, family = binomial, data = train)

Deviance Residuals:
    Min      1Q  Median      3Q     Max
-2.6994  -0.7696   0.6450   0.7629   1.7183

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       6.704e+01  7.539e+00   8.893  < 2e-16 ***
year             -3.427e-02  3.764e-03  -9.106  < 2e-16 ***
score             1.331e-01  4.465e-02   2.981  0.00287 **
votes             1.975e-06  4.293e-07   4.600 4.22e-06 ***
budget            1.272e-08  1.650e-09   7.711 1.25e-14 ***
gross            -2.006e-09  3.605e-10  -5.565 2.62e-08 ***
director_genderMale  1.490e+00  1.390e-01  10.715  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4662.2  on 4101  degrees of freedom
Residual deviance: 4324.1  on 4095  degrees of freedom
AIC: 4338.1

Number of Fisher Scoring iterations: 5
```

Now we tested this model by using the test data we had set aside, which produced this confusion matrix.

```
    Actual
Predict Female Male
   0    34   23
   1   301 1010
```

This resulted in a misclassification rate of 0.2368.

We checked the distribution of each of the variables within the subpopulations (male vs. female) to see if the data was normally distributed.

For 'year' (female and male):

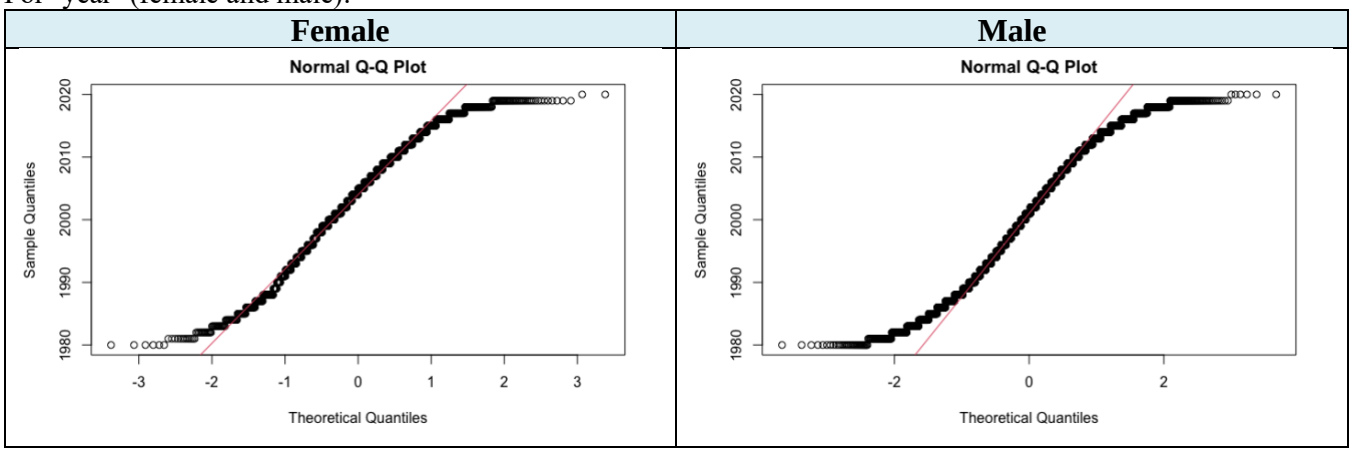


Figure 20: Q-Q Plots for Analysis 3b, for the variable year

For 'score' (female and male):

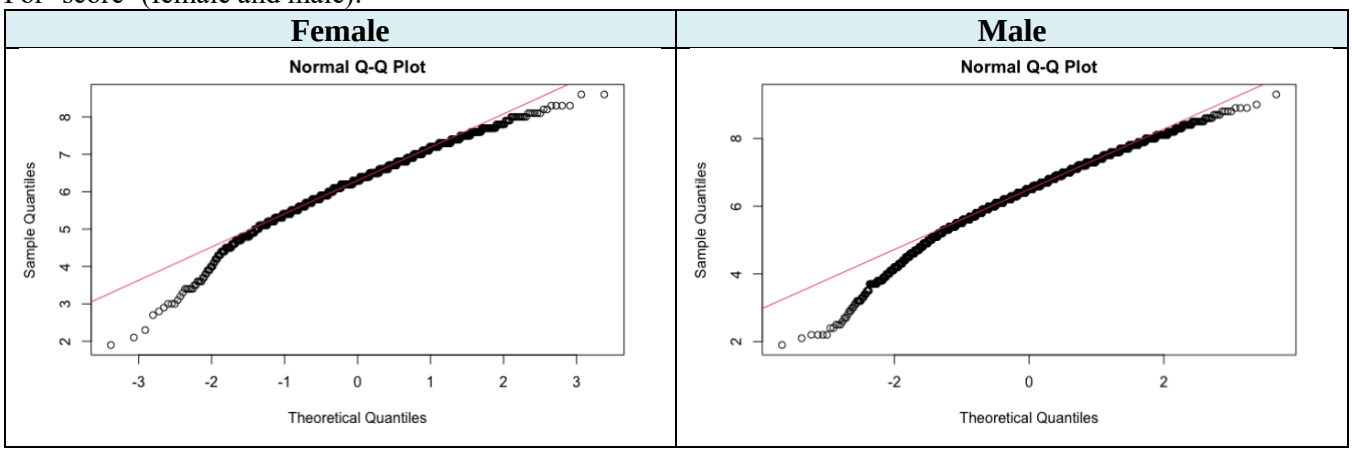


Figure 21: Q-Q Plots for Analysis 3b, for the variable score

For ‘votes’ (female and male):

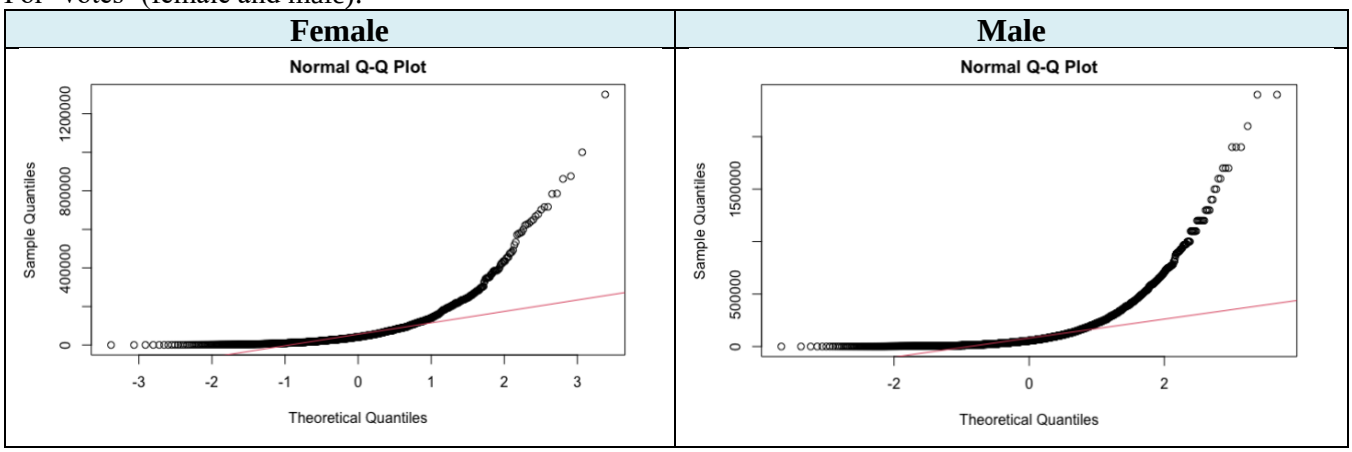


Figure 22: Q-Q Plots for Analysis 3b, for the variable votes

For ‘budget’ (female and male):

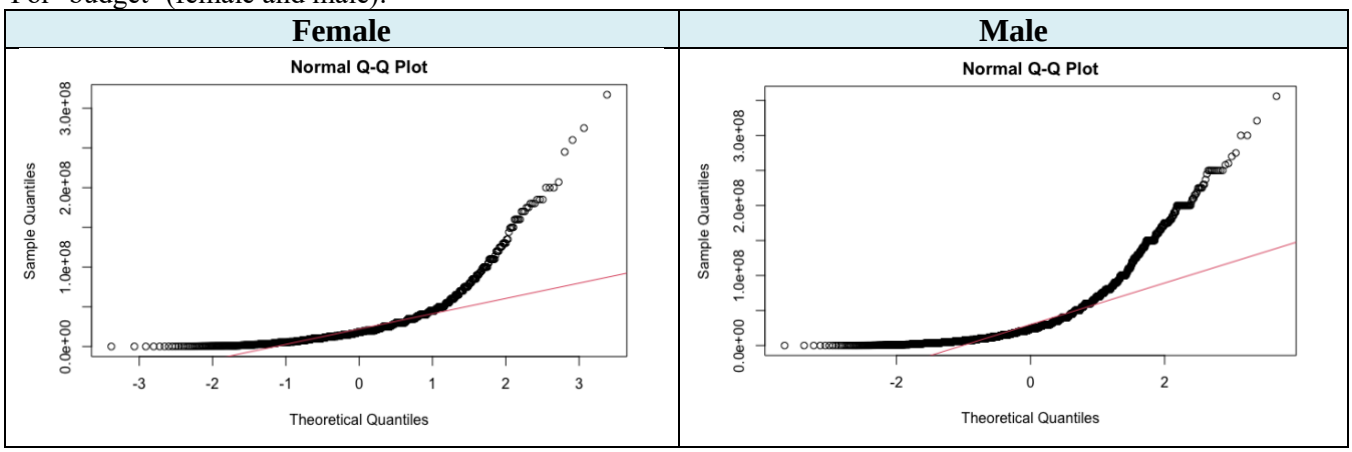


Figure 23: Q-Q Plots for Analysis 3b, for the variable budget

For ‘gross’ (female and male):

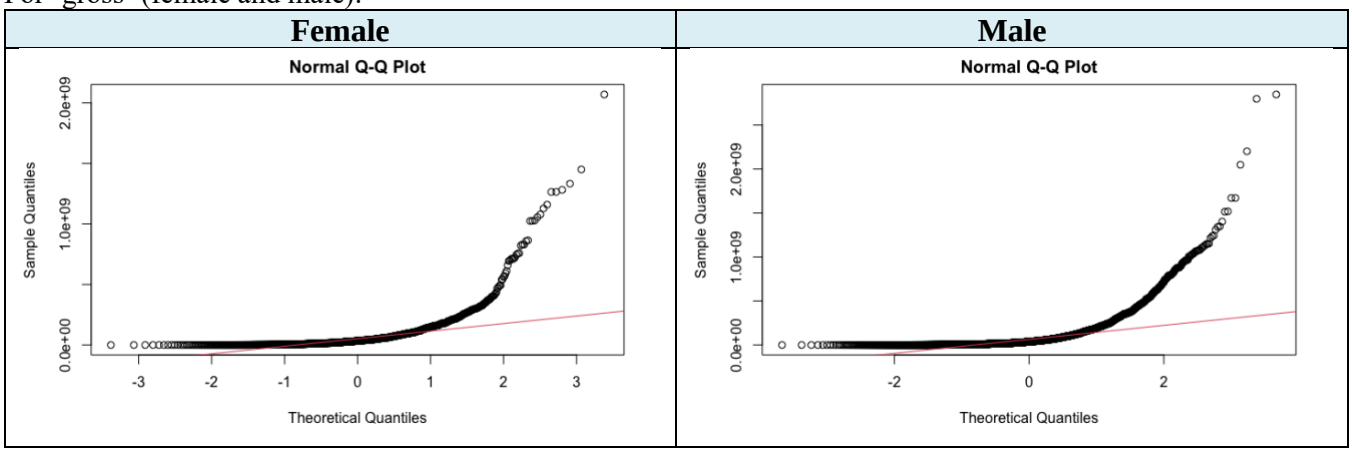


Figure 24: Q-Q Plots for Analysis 3b, for the variable gross

For ‘runtime’ (female and male):

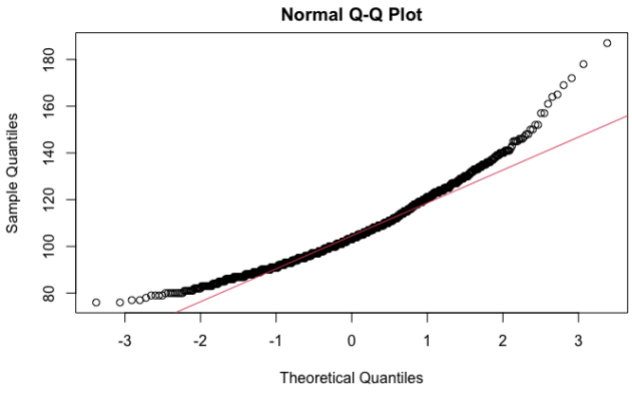
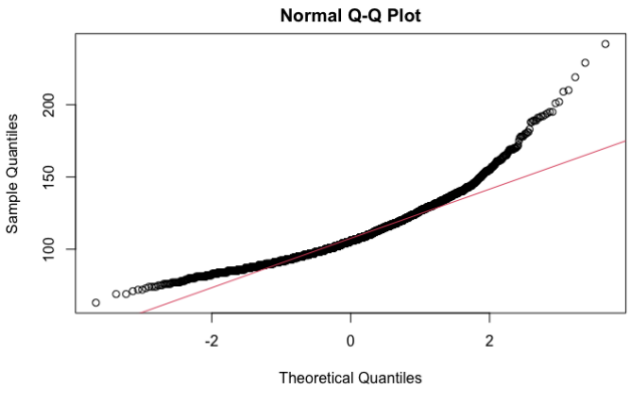
Female	Male
	

Figure 25: Q-Q Plots for Analysis 3b, for the variable runtime

The graphs show that only ‘runtime’, ‘score’ are really the only ones that are normally distributed, while the other variables do not conform at all.

We confirmed this with a Shapiro-Wilks test for each of the parameters. All of the tests showed significance (we’ve only included one for demonstrative purposes), suggesting none of the parameters are normally distributed.

```
        Shapiro-Wilk normality test

data:  star_female$votes
W = 0.95888, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  star_male$votes
W = 0.96583, p-value < 2.2e-16
```

LDA

After constructing the Logistic Regression model, we now construct a LDA model.

```
lda.star_gender=lda(star_gender~.,data=train)
lda.star_gender.best = lda(star_gender~year+score+votes+budget+gross+director_gender,data=train)
```

The misclassification rate of this model was as follows:

```
[1] "class"     "posterior" "x"

       Female Male
 Female     35   35
 Male      300  998
Misclassification Rate:  0.2485163
```

QDA

Finally, we construct a QDA model as well.

```
qda.star_gender = qda(star_gender~year+score+votes+budget+gross+director_gender,data = train)
qda.class = predict(qda.star_gender,test)$class
table(qda.class,test$star_gender)
```

The misclassification rate of this model was as follows:

```
qda.class Female Male
  Female    38   42
  Male     297  991
Misclassification Rate:  0.247807
```

Analysis 4: Can movie ratings be predicted? If so, with which predictors?

For the last investigation, the feasibility of predicting movie ratings was explored. All applicable variables were utilized to understand which ones are best in exploring movie rating predictions.

Prior to modelling, the dataset was reviewed for specifically for the context of this question. There are two instances where ratings can be combined:

1) Not Rated and Unrated: These ratings are equivalent in the Cinema world (Nettle, 2018).
2) X and NC-17: "X" was transitioned to "NC-17" (Taylor, 2004).

Due to running issues with computational times in RStudio, columns unnecessary for the analysis were dropped to create a new dataframe to work with and any rows with missing rating information was also removed. The columns dropped include: title of the film, year the film was made, and release date of the film. The title of the film was dropped as it is a unique categorical input for each film – therefore, it could not be used to categorize a group of data. A similar concern was present for the release date of the film, as the exact date is listed and thus, likely it is a unique input for each film and therefore, cannot be use to categorize the films present. More importantly, information relating to dates were removed as the methods being applied are not suitable for time series data; this is why the year the film was made was removed as well. Therefore, out of the remaining variables, they can be split up such that:

1) Categorical variables: genre, director, writer, star, country, company, updated rating (variable aimed to be predicted)
2) Numerical variables: score, votes, budget, gross, runtime

Rows with missing rating information were dropped as well. The finalized count of the number of observations per rating can be seen in Table 2.

Table 2: Count of the number of observations per rating for Analysis 4

Rating	Number of Observations
Approved	1
G	153
NC-17	26

Rating	Number of Observations
Not rated	335
PG	1252
PG-13	2112
R	3697
TV-14	1
TV-MA	9
TV-PG	5

Finalized Approach: Contingency Table

When incorporating director, writer, star, country, or company, the table was exceptionally large and introduced predictors with many different inputs and thus resulted in being unable to show correlation to rating. Therefore, only genre was used as a predictor. In addition, this was the original hypothesis when generating this exploratory question since it was suspected that certain genres would be more likely to have certain ratings (e.g. horror films are more likely to be rated R).

First, the contingency and proportion tables are created, see below in Table 3 and Table 4.

Table 3: Full contingency table for Analysis 4

	Action	Adventure	Animation	Biography	Comedy	Crime	Drama	Family	Fantasy	Horror	Music
Approved	0	1	0	0	0	0	0	0	0	0	0
G	1	21	100	1	16	0	11	3	0	0	0
NC-17	0	0	0	1	4	4	15	0	0	2	0
Not Rated	49	5	10	14	61	41	140	0	2	9	0
PG	181	205	185	65	428	7	155	7	3	7	0
PG-13	623	91	23	135	738	46	390	0	9	44	0
R	843	102	13	223	984	447	767	0	29	256	1
TV-14	1	0	0	0	0	0	0	0	0	0	0
TV-MA	1	1	2	1	1	0	3	0	0	0	0
TV-PG	0	0	2	0	1	0	2	0	0	0	0

	Musical	Mystery	Romance	Sci-Fi	Sport	Thriller	Western
Approved	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0
NC-17	0	0	0	0	0	0	0
Not Rated	1	1	1	1	0	0	0
PG	0	0	3	2	0	3	1
PG-13	0	4	2	3	1	3	0
R	0	15	2	3	0	10	2
TV-14	0	0	0	0	0	0	0
TV-MA	0	0	0	0	0	0	0
TV-PG	0	0	0	0	0	0	0

Table 4: Proportional table, determined as a conditional on the row variable for Analysis 4

	Action	Adventure	Animation	Biography	Comedy	Crime
Approved	0.0000000000	1.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
G	0.0065359477	0.1372549020	0.6535947712	0.0065359477	0.1045751634	0.0000000000
NC-17	0.0000000000	0.0000000000	0.0000000000	0.0384615385	0.1538461538	0.1538461538
Not Rated	0.1462686567	0.0149253731	0.0298507463	0.0417910448	0.1820895522	0.1223880597
PG	0.1445686901	0.1637380192	0.1477635783	0.0519169329	0.3418530351	0.0055910543
PG-13	0.2949810606	0.0430871212	0.0108901515	0.0639204545	0.3494318182	0.0217803030
R	0.2280227211	0.0275899378	0.0035163646	0.0603191777	0.2661617528	0.1209088450
TV-14	1.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
TV-MA	0.1111111111	0.1111111111	0.2222222222	0.1111111111	0.1111111111	0.0000000000
TV-PG	0.0000000000	0.0000000000	0.4000000000	0.0000000000	0.2000000000	0.0000000000

	Drama	Family	Fantasy	Horror	Music	Musical
Approved	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
G	0.0718954248	0.0196078431	0.0000000000	0.0000000000	0.0000000000	0.0000000000
NC-17	0.5769230769	0.0000000000	0.0000000000	0.0769230769	0.0000000000	0.0000000000
Not Rated	0.4179104478	0.0000000000	0.0059701493	0.0268656716	0.0000000000	0.0029850746
PG	0.1238019169	0.0055910543	0.0023961661	0.0055910543	0.0000000000	0.0000000000
PG-13	0.1846590909	0.0000000000	0.0042613636	0.0208333333	0.0000000000	0.0000000000
R	0.2074655126	0.0000000000	0.0078441980	0.0692453341	0.0002704896	0.0000000000
TV-14	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
TV-MA	0.3333333333	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
TV-PG	0.4000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000

	Mystery	Romance	Sci-Fi	Sport	Thriller	Western
Approved	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
G	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
NC-17	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
Not Rated	0.0029850746	0.0029850746	0.0029850746	0.0000000000	0.0000000000	0.0000000000
PG	0.0000000000	0.0023961661	0.0015974441	0.0000000000	0.0023961661	0.0007987220
PG-13	0.0018939394	0.0009469697	0.0014204545	0.0004734848	0.0014204545	0.0000000000
R	0.0040573438	0.0005409792	0.0008114688	0.0000000000	0.0027048959	0.0005409792
TV-14	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
TV-MA	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
TV-PG	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000

Since it's not a 2 by 2 contingency table, the following tests cannot be applied: tests of difference, relative risk, and odds ratio. Instead, the Chi-square for hypothesis testing and if a warning is received regarding sample size, then Fisher's Exact Test will be applied instead.

The hypotheses being tested when performing these tests are:

$$H_0: \theta = 1$$
$$H_a: \theta \neq 1$$

Where θ is defined as the odds ratio, which can be found defined below:

$$\theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$$

Unfortunately, the chi-square approximation error was received, as seen below in Figure 26.

```
Warning: Chi-squared approximation may be incorrect
        Pearson's Chi-squared test

data:  cont.tab
X-squared = 3104.5, df = 153, p-value < 0.00000000000000022
```

Figure 26: Chi-Square test output for contingency table initial attempt for Analysis 4

When running the Fisher's Exact Test, a warning message indicates that the p-value needs to be simulated. This is required due to the computational power needed to calculate the p-value based off the given data is too large. Therefore, RStudio has recommended simulated p-values instead. When simulating the p-values, the following output in Figure 27 was obtained.

```
        Fisher's Exact Test for Count Data with simulated p-value (based on 2000 replicates)

data:  cont.tab
p-value = 0.0004998
alternative hypothesis: two.sided
```

Figure 27: Fisher's Exact test output for contingency table initial attempt for Analysis 4

Due to the very low p-value (less than an α of 0.05), the null hypothesis is rejected and thus, the two variables show dependence to one another. Since the p-value was much lower than the threshold of 0.05, there are no issues or concerns with using a simulated p-value approach.

The Chi-Square test was attempted again, this time, with categories with higher number of observations only. The categories with low observations are removed ("Approved", "NC-17", "TV-14", "TV-MA", "TV-PG", and "X"). The chi-square approximation error is observed and upon attempting Fisher's Exact Test, the p-values needed to be simulated as well, indicating that removing the categories with low observations could not eliminate the need for Fisher's exact test nor simulated p-values.

Nominal-Ordinal

The Nominal-Ordinal approach was also applied as the ratings are an ordinal categorical variable with a natural ranking: "G", "PG", "PG-13", "R", "NC-17", "Not Rated" (Motion Picture Association, 2023). However, note that this is the cinema rankings and there are some films within this dataset that uses TV rankings. Most of these films appear to be foreign films so it is likely they were assigned a TV ranking instead of a cinema ranking. The TV rankings are classified on a different scale and can be ranked as: "TV-PG", "TV-14", "TV-MA" (TV Parental Guidelines, n.d.). The dataset will be split into TV and cinema ratings to attempt this method.

The rankings are assigned such that more mature content will be given a higher score. The ranking updates can be found in Table 5 for TV ratings.

Table 5: Updated ranking for TV ratings for Analysis 4

Original rating	Assigned ranking
TV-PG	1
TV-14	2
TV-MA	3

The ranking updates for cinema ratings can be found in Table 6.

Table 6: Updated ranking for cinema ratings for Analysis 4

Original rating	Assigned ranking
G	1
PG	2
PG-13	3
R	4
NC-17	5
Not Rated	6

The only rating which has been left out of this ranking is “Approved”; this is since this rating was what was given prior to the current "scale" references, where previously, films were "Approved" or "Not Approved" (IMDb, 2013) (John, 2016). Ideally, this could've been modelled using a binomial distribution but there is no information on movies that were "Not approved". Thus, this category will not be included in this section of analysis.

ANOVA tests will be conducted with the following hypotheses:

$$H_0: Rating\ and\ Genre\ are\ independent$$
$$H_a: Rating\ and\ Genre\ are\ NOT\ independent$$

TV Rankings

The outcome of the ANOVA test is as follows:

```
             Df Sum Sq Mean Sq F value Pr(>F)
Genre         5  1.633  0.3267    0.26  0.924
Residuals     9 11.300  1.2556
```

Figure 28: ANOVA test output for TV Rankings for Analysis 4

Since the p-value is greater than an α value of 0.05, the null hypothesis cannot be rejected, indicating both variables are independent to one another. Although this is different from what was expected, it is likely due to the small sample size for TV ratings.

Cinema Rankings

The outcome of the ANOVA test is as follows:

```
              Df Sum Sq Mean Sq F value             Pr(>F)
Genre         17   1398   82.23   104.7 <0.0000000000000002 ***
Residuals   7557   5934    0.79
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 29: ANOVA test output for cinema Rankings for Analysis 4

Since the p-value is less than an α value of 0.05, the null hypothesis is rejected, indicated both variables are not independent to one another. Thus, the Genre is related to the rating given.

The proportions for rating by genre are regenerated for cinema ratings only, as seen below in Figure 30.

Genre	cinema.Rating 1	2	3	4	5	6
Action	0.0005892752	0.1066588097	0.3671184443	0.4967589864	0.0000000000	0.0288744844
Adventure	0.0495283019	0.4834905660	0.2146226415	0.2405660377	0.0000000000	0.0117924528
Animation	0.3021148036	0.5589123867	0.0694864048	0.0392749245	0.0000000000	0.0302114804
Biography	0.0022779043	0.1480637813	0.3075170843	0.5079726651	0.0022779043	0.0318906606
Comedy	0.0071716719	0.1918422232	0.3307933662	0.4410578216	0.0017929180	0.0273419991
Crime	0.0000000000	0.0128440367	0.0844036697	0.8201834862	0.0073394495	0.0752293578
Drama	0.0074424899	0.1048714479	0.2638700947	0.5189445196	0.0101488498	0.0947225981
Family	0.3000000000	0.7000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000
Fantasy	0.0000000000	0.0697674419	0.2093023256	0.6744186047	0.0000000000	0.0465116279
Horror	0.0000000000	0.0220125786	0.1383647799	0.8050314465	0.0062893082	0.0283018868
Music	0.0000000000	0.0000000000	0.0000000000	1.0000000000	0.0000000000	0.0000000000
Musical	0.0000000000	0.0000000000	0.0000000000	0.0000000000	0.0000000000	1.0000000000
Mystery	0.0000000000	0.0000000000	0.2000000000	0.7500000000	0.0000000000	0.0500000000
Romance	0.0000000000	0.3750000000	0.2500000000	0.2500000000	0.0000000000	0.1250000000
Sci-Fi	0.0000000000	0.2222222222	0.3333333333	0.3333333333	0.0000000000	0.1111111111
Sport	0.0000000000	0.0000000000	1.0000000000	0.0000000000	0.0000000000	0.0000000000
Thriller	0.0000000000	0.1875000000	0.1875000000	0.6250000000	0.0000000000	0.0000000000
Western	0.0000000000	0.3333333333	0.0000000000	0.6666666667	0.0000000000	0.0000000000

Figure 30: Proportions for Cinema rankings for Analysis 4

To confirm what is observed with the ANOVA test for the cinema ratings subset of the data, the chi-square test was applied, which resulted in the approximation error warning. Therefore, Fisher's Exact Test with approximated p-values was used instead and resulted in the same conclusion, supporting the findings of the ANOVA test. A screenshot of the outcome of Fisher's Exact Test is seen below in Figure 31.

```
        Fisher's Exact Test for Count Data with simulated p-value (based on 2000 replicates)

data:  cinema.tab
p-value = 0.0004998
alternative hypothesis: two.sided
```

Figure 31: Fisher's Exact Test with simulated p-values for cinema ratings for Analysis 4

Methods which were not used but yielded results

These methods were not pursued due to high misclassification rates and the restriction of being unable to incorporate categorical variables as predictors. This is especially concerning as from the contingency table, it was shown that rating and genre are dependent; thus, genre should be included in any model for rating.

Categorical Tree

None of the categorical variables remaining are ordinal – thus, they cannot be included within the categorical tree analysis. This analysis was conducted with the numerical variables.

Prior to all other steps, all cells with missing information were removed from the dataframe.

First, the sample will be split into training and validation sets. The method of sampling chosen was stratified sampling by genre. A review of the number of observations per rating indicated that TV-MA and Approved have observations of 2 and 1, respectively. This is too low of a sample size to allow for sampling then testing and thus, both genres were removed. For this method and question, the 80-20 convention was chosen – where 80% of the data is used for training while the remaining 20% is used for validation (Joseph, 2022).

After the data was split, the resulting classification tree can be seen below in Figure 32.

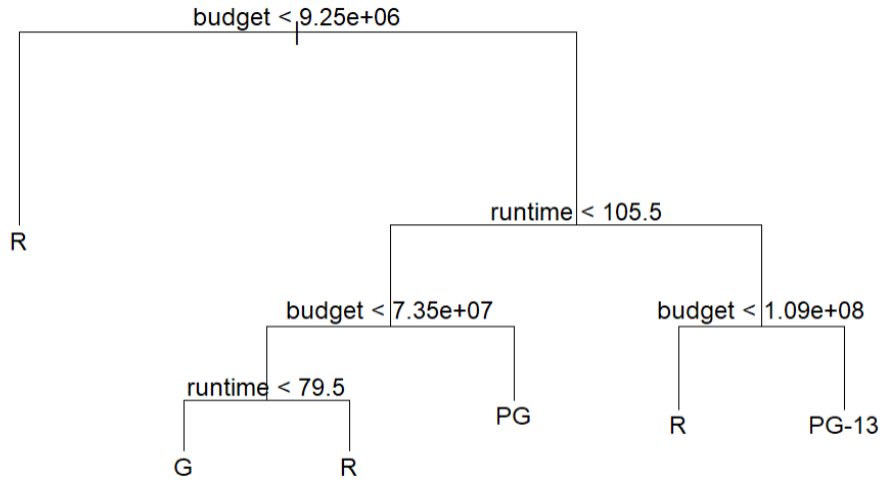


Figure 32: Classification tree for Analysis 4

Cross-validation was used to determine if pruning the tree would be appropriate. A figure of cross-validation error and the number of terminal nodes is generated below in Figure 33.

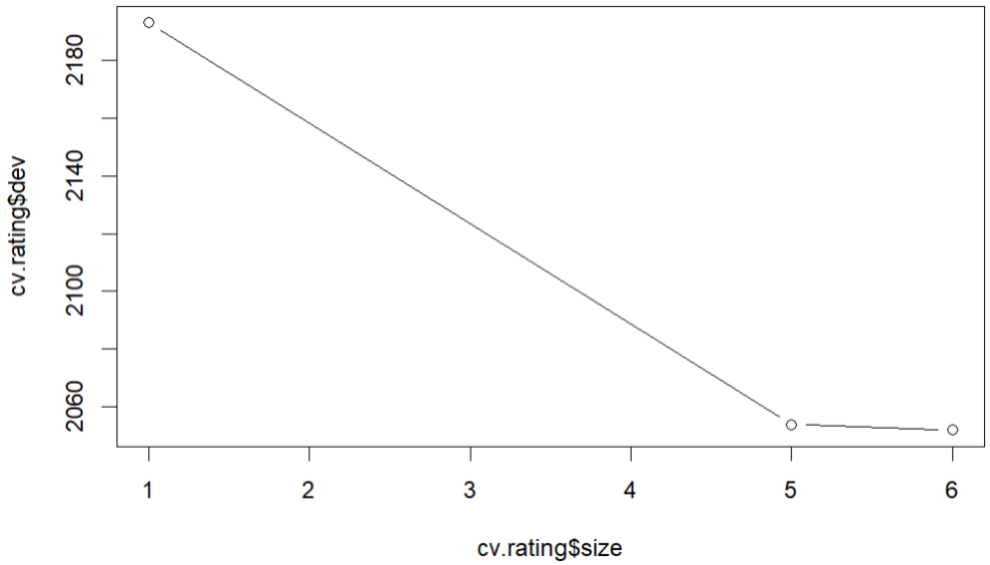


Figure 33: Cross-validation error as a function of the number of terminal nodes for Analysis 4

As seen above, the appropriate number of terminal nodes to minimize the cross-validation error while reducing the risk of over-fitting the model would be 5. The pruned tree was generated with the criteria of 5 terminal nodes (see Figure 34).

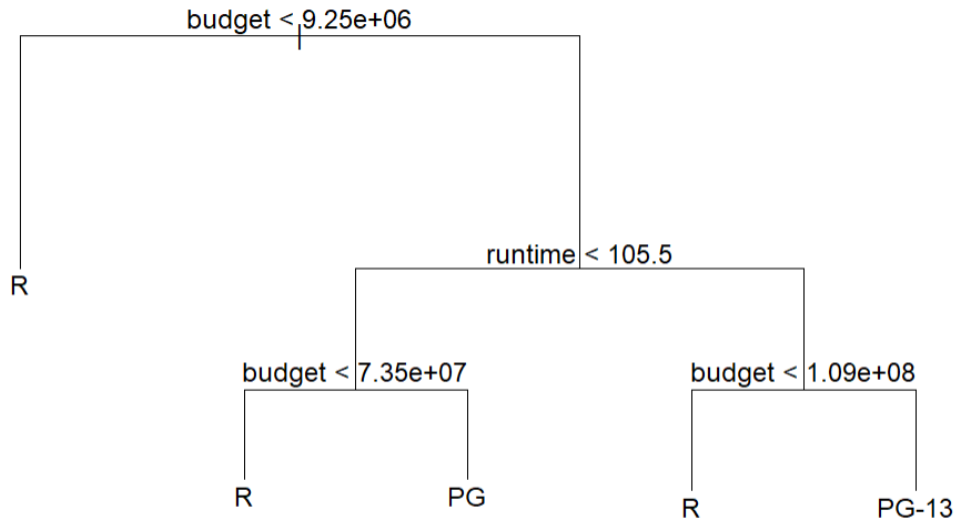


Figure 34: Pruned tree for Analysis 4

The misclassification rate for the original and pruned tree can be seen below in Table 7.

Table 7: Misclassification rate for categorization tree approach for Analysis 4

	Original tree	Pruned tree
Misclassification rate (%)	46.3100	46.4022

Note that all values have been rounded to 4 decimal places.

The misclassification rate is slightly higher for the pruned tree; however, this model would be proposed as it is a small increase in the misclassification rate for a lower cross-validation error. Overall, the misclassification rate is still quite high.

LDA/QDA

The assumptions required for this model were reviewed first.

The first assumption being reviewed is the equal variance assumption. This was determined using the Breusch-Pagan (BP) test.

$$H_0: Equal\ variance\ is\ present$$
$$H_a: Equal\ variance\ is\ NOT\ present$$

The outcome of the BP test can be seen below:

```
        studentized Breusch-Pagan test

data:  lda.assumptions
BP = 91.757, df = 4, p-value < 0.00000000000000022
```

Figure 35: BP Test outputs for assumption checking for Analysis 4

As the p-value of the model is smaller than 0.05, we will reject the null hypothesis and conclude that the homoscedasticity (i.e. equal variance) does not hold. Thus, the equal variance assumption fails. This indicates that

LDA is not a suitable method for modelling.

The second assumption being reviewed is the normality assumption. As previously stated, both Approved and TV-MA are not included in the final modelling process as the sample size is too low for a training and validation set to be created. A Q-Q (Quantile-Quantile) plot for the entire dataset can be seen in Figure 36.

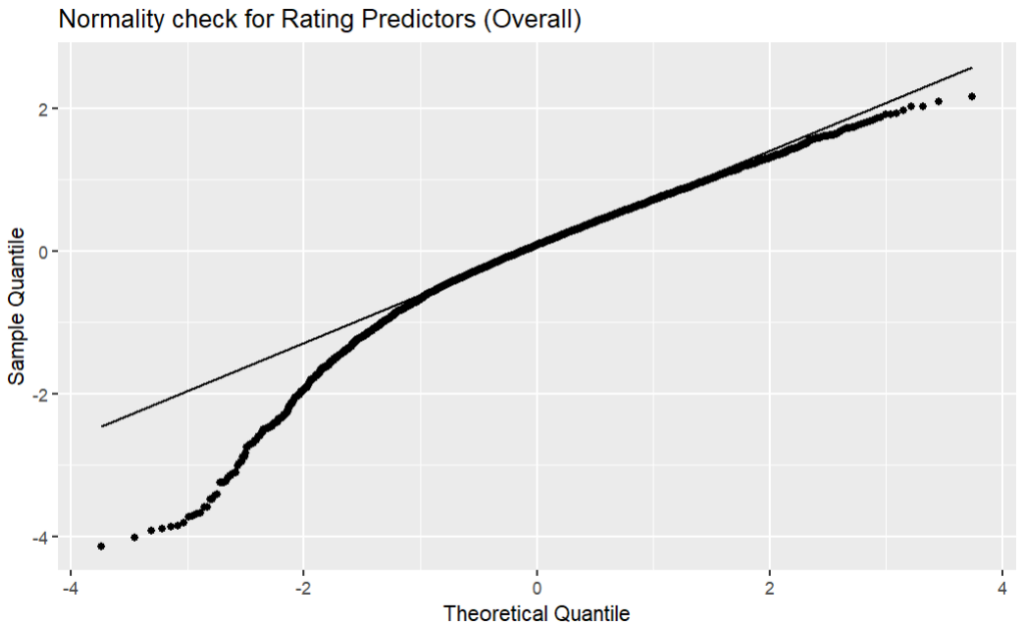


Figure 36: Normality check for overall dataset (Q-Q plot) for assumption checking for Analysis 4

Unfortunately, it does not appear as if the data meets the normality assumption. The Shapiro-Wilk test could not be applied on the entire dataset as a warning message was received for the size of the dataset. However, further analysis will be conducted for the data available for each category in the variable of interest (i.e. each rating).

The hypotheses being tested as part of the Shapiro-Wilk test are:

$$H_0: The\ sample\ data\ are\ significantly\ normally\ distributed$$
$$H_a: The\ sample\ data\ are\ NOT\ significantly\ normally\ distributed$$

The outcome of the Shapiro-Wilk test can be seen in Table 8. Note all p-values are rounded to 4 decimal places.

Table 8: Shapiro-Wilk test outcomes for Analysis 4

Rating	Shapiro-Wilk p-value	Normally distributed?
G	0.0037	No
NC-17	0.8332	Yes
Not Rated	0.8011	Yes
PG	<0.0001	No
PG-13	<0.0001	No
R	<0.0001	No

As an extra precautionary measure, the Q-Q plots are also generated for each rating, as seen in Figure 37.

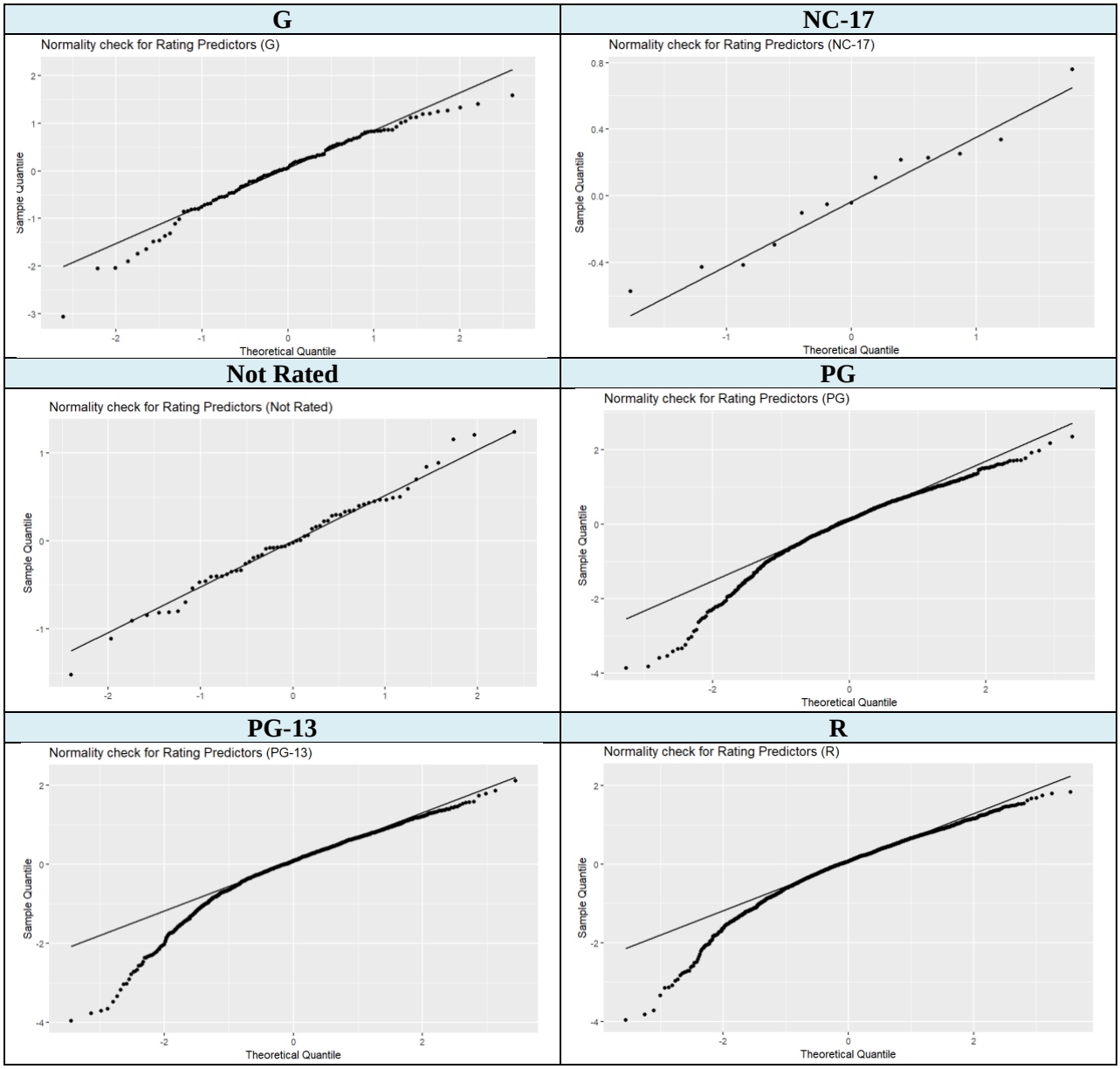


Figure 37: Q-Q plots for each rating category for Analysis 4

It appears that large sample sizes do not fit the normality assumption. Since the larger samples are driving the overall dataset to not meet the normality assumption, it can be concluded that normality is not met.

For the sake of understanding both methods of analyses, the LDA/QDA model will continue to be applied.

The misclassification rate for both LDA and QDA methods can be seen in Table 9.

Table 9: Misclassification rate for LDA/QDA approaches for Analysis 4

	LDA	QDA
Misclassification rate (%)	43.8192	44.0959

A plot of the LDA analysis can be seen in Figure 38.

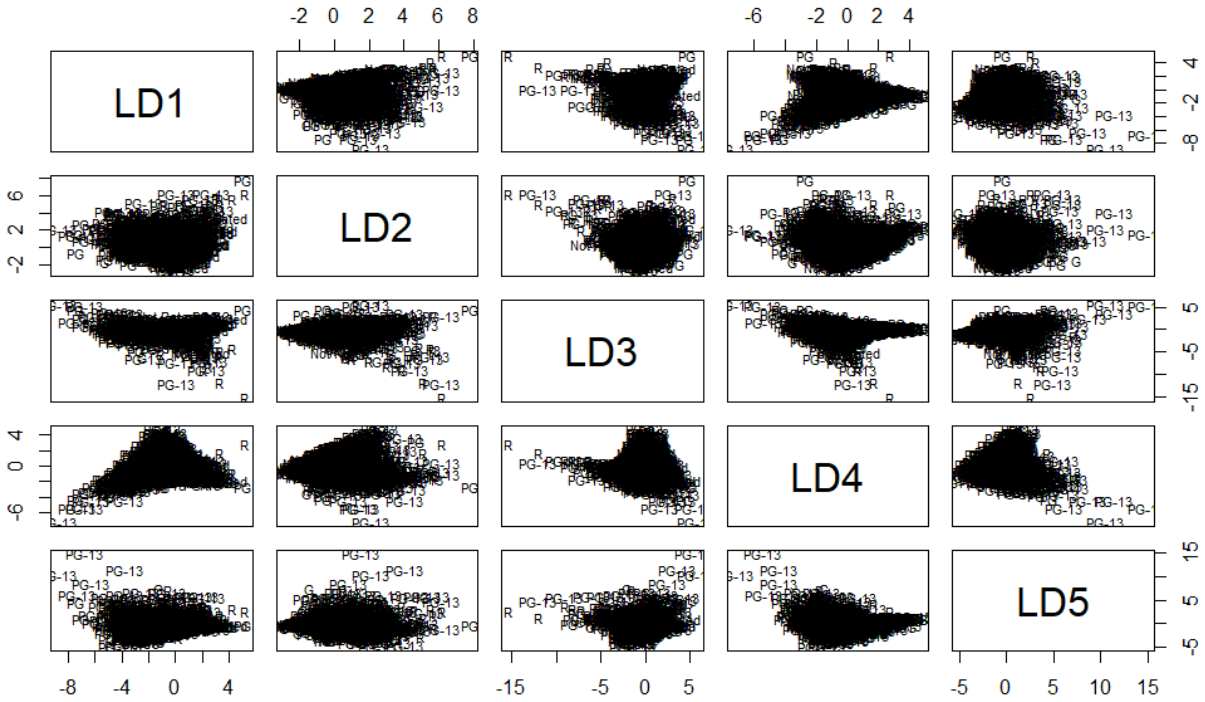


Figure 38: LDA plot for Analysis 4

The above figure indicates that perhaps there are too many categories for the model to be able to accurately predict the outcomes. And it showcases the inaccuracy of the LDA model.

Method which did not yield results: Multinomial

Ideally, after the Contingency Table method determined that the two variables are correlated, the probabilities can be determined through the Multinomial method.

The cumulative logit or continuation ratio method were applied first as rating has already been established as an ordinal categorical variable. Unfortunately, both methods yielded a goodness-of-fit of 0. The cumulative logit method was attempted again with reduced rows to determine if this would help increase the goodness-of-fit. Therefore, any genres with less than 50 observations were removed. This still unfortunately resulted in a goodness-of-fit of 0.

The baseline category method is attempted to see if it has a better goodness-of-fit. A warning message was received indicating the table was too large for a proper computation. Thus, the reduced table created for the second attempt at cumulative logit is used. The same warning message was received – however, the model was generated and thus, the goodness-of-fit still could be calculated. Even with the baseline category method, it returned a value of 0. Thus, after consultation with professor Jiang, this method was discontinued.

Conclusion

In the first analysis to predict IMDb scores, a linear regression model is proposed which incorporates the following predictors with a meaningful coefficient (β) term: genre and runtime.

In the second analysis to determine which variables contribute to if a movie is “good” or “bad”, a logistic regression model is proposed which incorporates the following predictors: genre, budget, gross, and votes. Within the same analysis, to predict the genre for a movie, various analyses were carried out but the ultimate conclusion

is that this dataset is not suitable for the type of modelling attempted.

In the first part of the third analysis, we tried to create a model for gross income. Overall, although we could create a model with an acceptable adjusted R-squared value, because our data does not conform at all to the assumptions required, we do not recommend modelling this data with the linear regression technique. In the second part, we tried to effectively predict the gender of the star given certain parameters of the movie. While this was quite successful (misclassification rates <.25), we could not fulfill all of the assumptions needed for the model to be valid.

In the fourth and last analysis to determine if predicting the rating of a movie is possible, the final conclusion is that it is dependent on genre but unfortunately could not be modelled using any of the attempted techniques.

References

1. ATLAS Cinemas. (2023). *The Film Rating System*. ATLAS Cinemas. https://atlascinemas.net/ratings.html#:~:text=If%20a%20film%20has%20not,(UR)%20are%20often%20used
2. Grijalva, D. (2020). *Movie Industry*. Kaggle. https://www.kaggle.com/datasets/danielgrijalvas/movies
3. IMDb. (2013). What does the "Approved" rating mean in IMDb? *IMDb*. https://community-imdb.sprinklr.com/conversations/data-issues-policy-discussions/what-does-the-approved-rating-mean-in-imdb/5f4a792c8815453dba747080#:~:text=Giancarlo_Cairella,%22moral%22%20or%20%22immoral
4. John. (2016, December 23). *What are the meanings of the terms "Passed" and "Approved" with regards to a movie title?* Stackexchange. https://movies.stackexchange.com/questions/65430/what-are-the-meanings-of-the-terms-passed-and-approved-with-regards-to-a-mov
5. Joseph, V. R. (2022). Optimal ratio for data splitting. *Statistical Analysis and Datamining: The ASA Data Science Journal*, *15*(4), 531–538. https://doi.org/10.1002/sam.11583
6. Motion Picture Association. (2023). *Film Ratings*. Motion Picture Association. https://www.motionpictures.org/film-ratings/
7. Naveen. (2022, June 18). *R – Replace String with Another String or Character*. SparkByExamples. https://sparkbyexamples.com/r-programming/replace-string-with-another-string-in-r/
8. Nettle. (2018, March 15). *Replace all occurrences of a string in a data frame*. Stackoverflow. https://stackoverflow.com/questions/29271549/replace-all-occurrences-of-a-string-in-a-data-frame
9. Schork, J. (2022, June 13). *NA Omit in R | 3 Example Codes for na.omit (Data Frame, Vector & by Column)*. Statistics Globe. https://statisticsglobe.com/na-omit-r-example/
10. Taylor, R. (2004, July 17). X to NC-17: The Evolution of Film's Most Controversial Rating. *Not Coming to a Theater Near You*. http://www.notcoming.com/features/nc17/
11. TV Parental Guidelines. (n.d.). *UNDERSTANDING THE TV RATINGS AND PARENTAL CONTROLS*. TV Parental Guidelines. Retrieved February 1, 2023, from http://www.tvguidelines.org/resources/TV_Parental_guidelines_Brochure.pdf