Movie Dataset Analysis

Group_10

2025-03-22

Introduction

This report analyzes a movie dataset, exploring trends in ratings, genres, and yearly movie counts. Various statistical models are applied to examine factors influencing movie ratings.Namely: exploring Movie Trends, rating distribution, genre popularity, success prediction.

Load necessary libraries for data manipulation, visualization, and statistical analysis

Load Dataset

The dataset consists of 7 variables: film_id (Unique identifier for each movie), year (The year the movie was released), length (Duration of the movie in minutes), budget (Budget of the movie (in million dollars)), votes (Number of votes received on IMDb), genre (The genre(s) of the movie), rating (IMDb rating of the movie).

```
data <- read.csv("dataset10.csv")
str(data)
```

```
'data.frame':   1495 obs. of  7 variables:
 $ film_id: int  49834 53923 30020 46364 19967 41873 39319 42809 32883 47313 ...
 $ year   : int  1963 1984 1992 2000 1964 2003 1983 1989 1940 1906 ...
 $ length : int  107 NA 32 NA 87 95 96 92 81 8 ...
 $ budget : num  11.4 9.7 15.4 11.5 9.4 13.3 9.6 12 16 11 ...
 $ votes  : int  225 59 6 69 34 22 10 200 7 5 ...
 $ genre  : chr  "Romance" "Comedy" "Documentary" "Action" ...
 $ rating : num  3.1 2.3 7.7 2.3 5.5 4.9 2.9 4.2 5.3 7.7 ...
```

```
summary(data)
```

```
    film_id           year          length           budget          votes
 Min.   :   40   Min.   :1903   Min.   :  2.00   Min.   : 3.2   Min.   :    5
 1st Qu.:14059   1st Qu.:1958   1st Qu.: 74.75   1st Qu.:10.1   1st Qu.:   12
 Median :28849   Median :1984   Median : 90.00   Median :11.9   Median :   33
 Mean   :29134   Mean   :1977   Mean   : 82.29   Mean   :12.0   Mean   :  790
 3rd Qu.:44299   3rd Qu.:1998   3rd Qu.:100.00   3rd Qu.:13.9   3rd Qu.:  115
 Max.   :58764   Max.   :2005   Max.   :288.00   Max.   :22.1   Max.   :84488
                                NA's   :59
    genre               rating
 Length:1495        Min.   :0.800
 Class :character   1st Qu.:3.700
 Mode  :character   Median :4.600
                    Mean   :5.346
                    3rd Qu.:7.800
                    Max.   :8.900
```

Data Cleaning

By using na.omit() function, the data was cleaned, namely sells with "N/A" were deleted in a whole row.

```
data1 <- na.omit(data)
```

Exploratory Data Analysis (EDA)

Rating Distribution

The histogram shows the distribution of IMDb movie ratings, revealing a bimodal pattern with peaks around 4.0 and 7.5. This suggests that movies tend to be rated either poorly or highly, with fewer falling in the mid-range (5.5–7.0). The x-axis represents ratings, while the y-axis shows the count of movies in each range. The gap in the middle indicates differences in movie quality, budget, or audience perception, where some films receive strong praise while others are widely criticized. Distribution is not distributed normally at all, however, by categorizeing them into binomial, in GLM the data will be observed only that, which has rating more than 7.

```
ggplot(data1, aes(x = rating)) +
  geom_histogram(binwidth = 0.5, fill = "blue", color = "black", alpha = 0.7)
   ↪  +
  labs(title = "Distribution of Rating", x = "Rating", y = "Count")
```

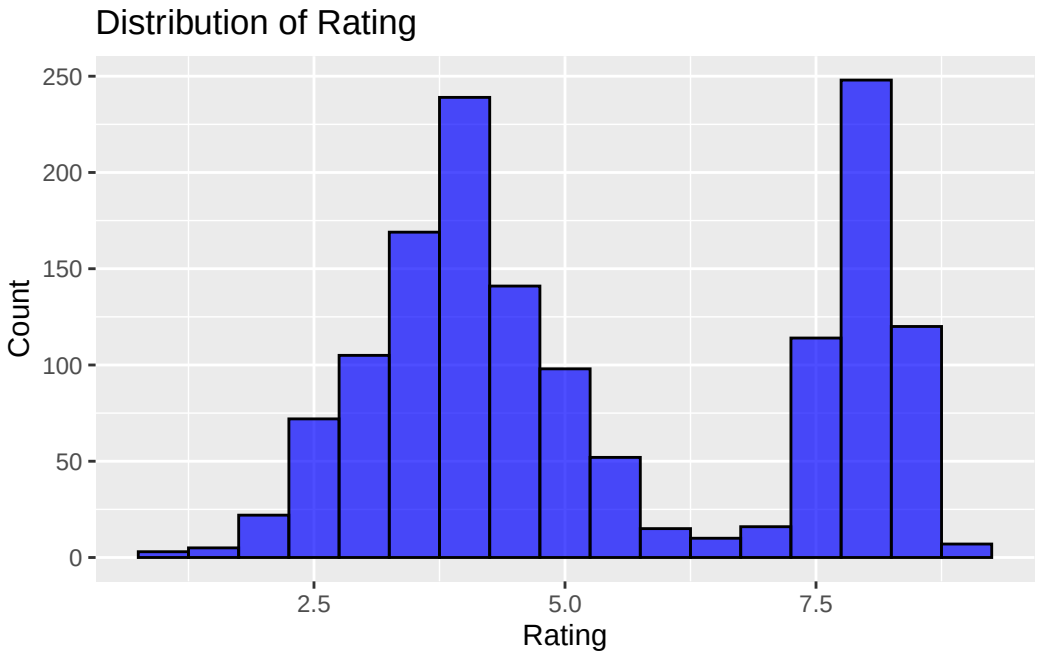

Figure 1: Distribution of Rating

Top 10 Movie Genres

Barplot below shows top-10 most common movie genres. The first two places were taken by "drama" and "action" genres with more than 435, and the lowest number of films in "romance" genre, that are about 16.

```
genre_count <- data %>%
  separate_rows(genre, sep = "\\|") %>%
  count(genre, sort = TRUE) %>%
  head(10)

ggplot(genre_count, aes(x = reorder(genre, n), y = n)) +
  geom_bar(stat = "identity", fill = "orange") +
  coord_flip() +
```

```
  labs(title = "Top 10 most common movie genres", x = "Genre", y = "Count")
```

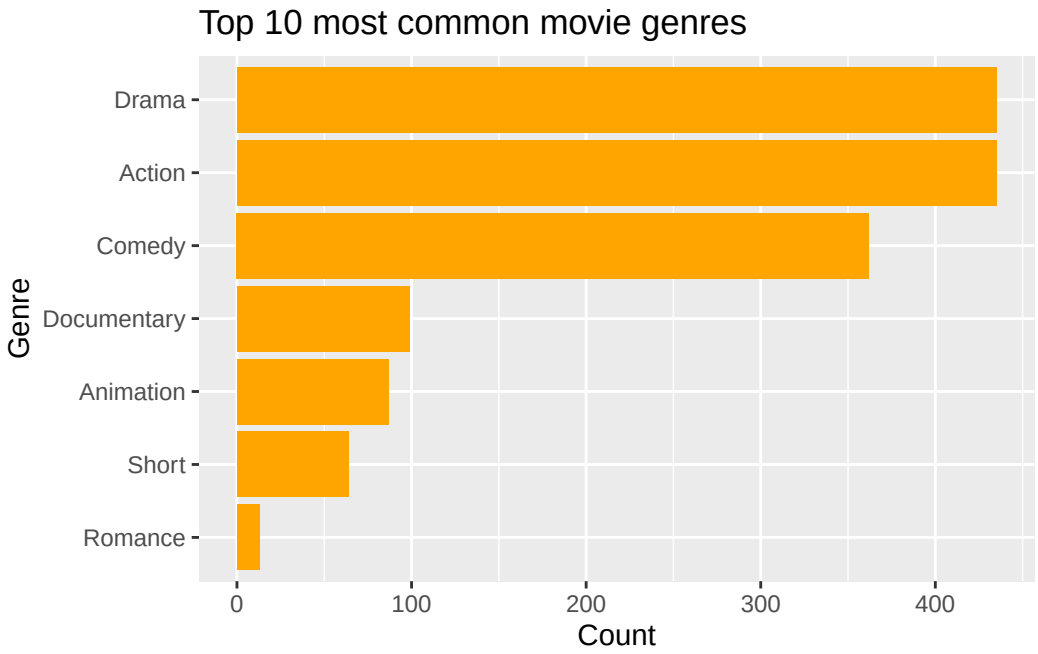

Figure 2: Top 10 most common movie genres

Movies Released Per Year

The line graph illustrates the number of movies released per year from the early 1900s to the early 2000s. Initially, movie production was minimal, but it steadily increased throughout the 20th century, with occasional fluctuations. A noticeable surge occurs around the 1970s and beyond, with a significant peak in the late 1990s and early 2000s, reflecting the rise of the global film industry, advancements in technology, and increased accessibility to filmmaking. The sharp decline at the end could be due to incomplete data for recent years. The red dots highlight individual data points, while the blue line connects them, showing an overall upward trend in movie production over time.

```
movies_per_year <- data %>%
  group_by(year) %>%
  summarize(count = n())

ggplot(movies_per_year, aes(x = year, y = count)) +
  geom_line(color = "blue") +
```

```
  geom_point(color = "red") +
  labs(title = "Number of movies released per year", x = "Year", y = "Count")
```

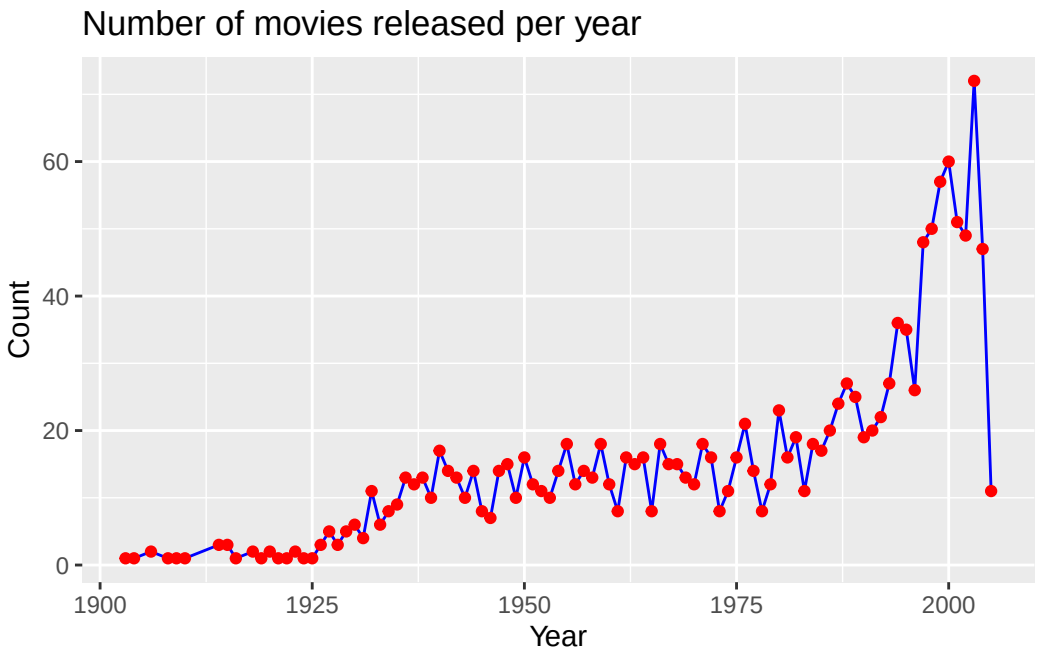

Figure 3: Number of movies released per year

Binary Rating Classification

Since the main idea asks to observe all films, that has rating more than 7 in the dataset, and dataset's rating spreads widely, the main idea of bunary rating calssification was taken. All films, that were rated more than 7 have "1" as binomial category and all other are "0". Do show the distribution, histogram was visualized below.

```
data1$rating_binary <- ifelse(data1$rating >= 7, 1, 0)

ggplot(data1, aes(x = factor(rating_binary), fill = factor(rating_binary))) +
  geom_bar() +
  theme_minimal() +
  labs(title = "Distribution of Binary Rating", x = "Binary Rating", y =
   ↪  "Count")
```

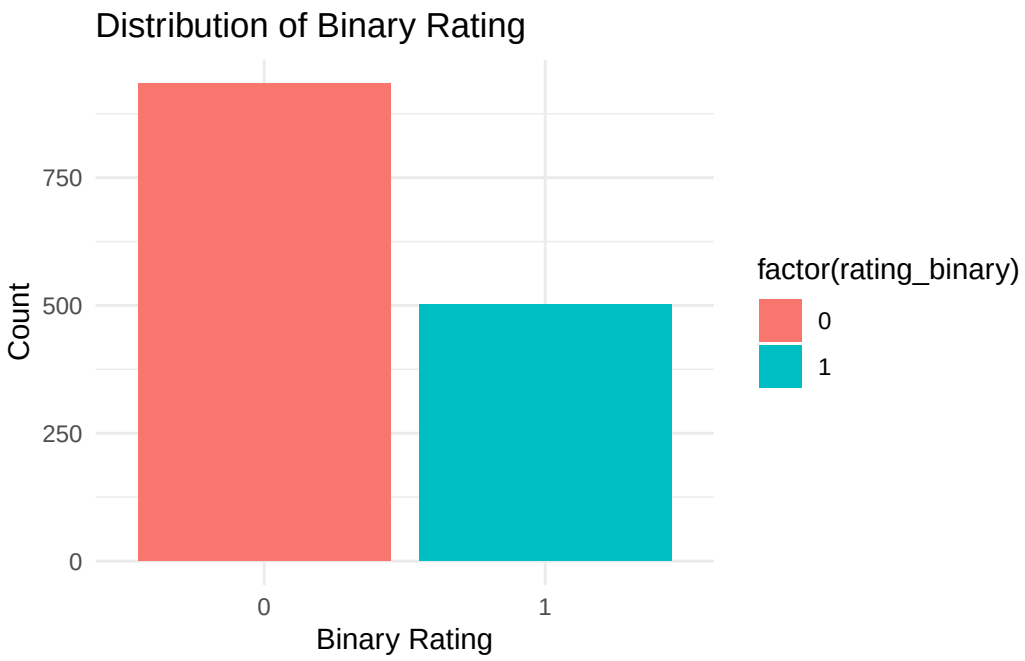

Figure 4: Distribution of Binary Rating

Standardization

The min_max_norm function scales numerical values so that the minimum value becomes 0 and the maximum becomes 1, which helps in standardizing data for machine learning models. lapply(data1[,2:5], min_max_norm) applies this function to columns 2 to 5 of data1, modifying them in place. Normalization ensures that all numerical features have the same scale, preventing any single variable from dominating analysis due to larger values.

```
#x-min/max-min standard methods
min_max_norm=function(x){
  return((x-min(x))/(max(x)-min(x)))
}
#standardize
data1[,2:5]=lapply(data1[,2:5],min_max_norm)
```

Regression Analysis

Linear Model

Model 1: Linear Regression for Movie Rating Prediction

This linear regression model predicts a movie's IMDb rating based on **year**, length, budget, votes, and genre. The Intercept (4.5230) suggests that, on average, a movie with all predictors at their baseline would have a rating of **4.52**. The budget (**3.5523**) and votes (**1.6732**) have a positive impact, indicating that movies with higher budgets and more votes tend to have higher ratings. Conversely, length (**-4.6849**) has a negative coefficient, suggesting that longer movies tend to receive lower ratings. **Year (-0.1382)** is included in the model but is not statistically significant ($\mathbf{p = 0.43751}$), indicating that the release year does not have a strong influence on ratings in this dataset.

Among the genre variables, comedies (**1.6255**), documentaries (**2.6033**), animations (**1.0747**), and short films (**2.1542**) have significant positive effects, indicating these genres tend to have higher ratings than the baseline genre. Dramas (**-0.4226**) and romances (**-1.2728**) have negative effects, implying that these genres are associated with lower ratings on average.

The model explains **48.87%** of the variance in movie ratings ($\mathbf{R^2 = 0.4887}$), which suggests that while the included predictors have a strong impact, other unexplored factors (such as director, cast, screenplay quality, and audience demographics) may also significantly influence ratings. The F-statistic ($\mathbf{136.2}$, $\mathbf{p < 2.2e\text{-}16}$) indicates that the overall model is highly significant. However, the residual standard error (**1.499**) suggests some variability in predictions, and the residual plot (from the provided image) shows non-random patterns, indicating that the model might not fully capture all relationships. This suggests that non-linear models or additional features might improve prediction accuracy.

```r
data1$genre_num <- as.factor(data1$genre)
model1 <- lm(rating ~ year+length + budget + votes + genre_num, data = data1)
summary(model1)
```

```
Call:
lm(formula = rating ~ year + length + budget + votes + genre_num,
    data = data1)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7995 -0.9983  0.0002  1.0404  4.4763

Coefficients:
```

```
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)             4.5230     0.2224  20.336  < 2e-16 ***
year                   -0.1382     0.1780  -0.777  0.43751
length                 -4.6849     0.4341 -10.793  < 2e-16 ***
budget                  3.5523     0.2597  13.677  < 2e-16 ***
votes                   1.6732     0.7613   2.198  0.02814 *
genre_numAnimation      1.0747     0.2078   5.171 2.66e-07 ***
genre_numComedy         1.6255     0.1104  14.730  < 2e-16 ***
genre_numDocumentary    2.6033     0.1767  14.731  < 2e-16 ***
genre_numDrama         -0.4226     0.1042  -4.054 5.30e-05 ***
genre_numRomance       -1.2728     0.4231  -3.008  0.00267 **
genre_numShort          2.1542     0.2342   9.198  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.499 on 1425 degrees of freedom
Multiple R-squared:  0.4887,    Adjusted R-squared:  0.4851
F-statistic: 136.2 on 10 and 1425 DF,  p-value: < 2.2e-16
```

Residual Plot for Linear Model

The residual plot of the linear model shows the residuals (errors) on the y-axis and the fitted values (predicted ratings) on the x-axis. Ideally, residuals should be randomly scattered around the red dashed line at zero, indicating that the model's errors are uniformly distributed. However, the plot reveals a clear pattern, suggesting that the model might not fully capture the relationship between predictors and ratings. The funnel shape (wider spread for lower ratings) hints at heteroscedasticity, meaning the variance of errors is not constant across predictions. Additionally, the curved trend implies possible non-linearity, meaning a linear model might not be the best fit.

```r
ggplot(data1, aes(x = model1$fitted.values, y = model1$residuals)) +
  geom_point(color = "blue", alpha = 0.6) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  labs(title = "Residual Plot of Linear Model", x = "Fitted Values", y =
   ↪  "Residuals")
```

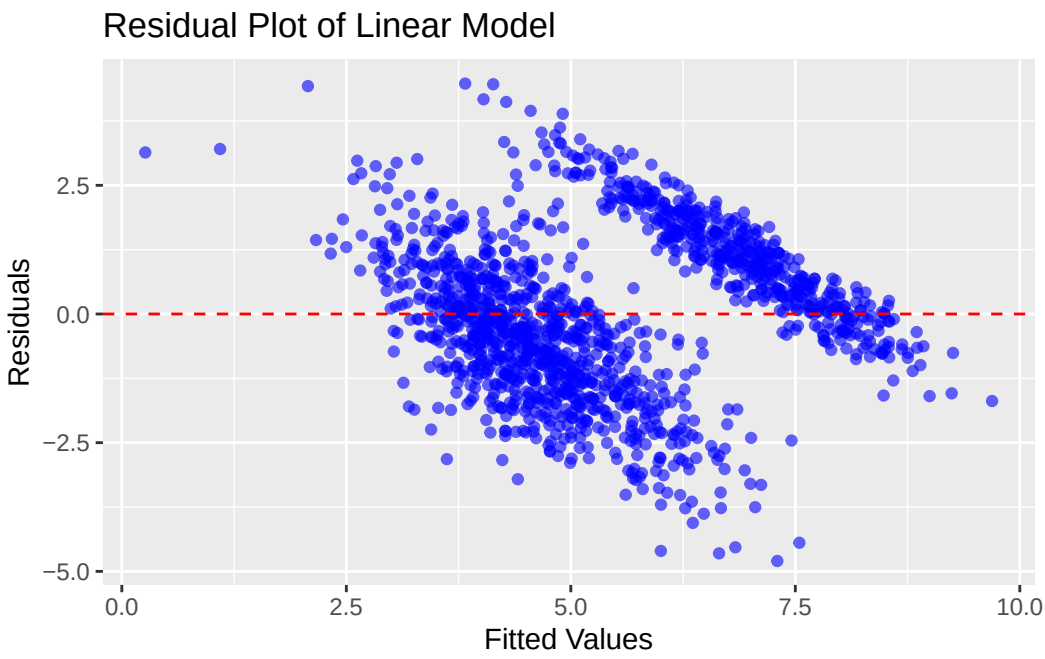

Figure 5: Residual Plot of Linear Model

Logistic Regression Models

Model 2: Logistic Regression for Binary Rating Classification

This logistic regression model predicts whether a movie's rating falls into a high or low category (binary classification) using various features such as year, length, budget, votes, and genre. The Intercept value of -2.2215 suggests that, without any predictor variables, the log-odds of a high rating are negative, indicating a lower likelihood. The budget (11.2481) and votes (4.8041) have strong positive effects, meaning that movies with higher budgets and more votes are more likely to receive a high rating. Conversely, length (-19.5733) and drama genre (-2.0779) have significant negative impacts, implying that longer movies and dramas are less likely to be rated highly.

The p-values indicate the significance of each variable: budget, length, votes, comedy, documentary, and short films have highly significant effects ($p < 0.05$), whereas year, animation, and romance genres are not statistically significant. The null deviance (1858.7) vs. residual deviance (711.5) shows a substantial reduction, indicating that the model explains a significant portion of variability. The AIC (733.5) suggests the model's overall quality, with lower values indicating better fit. However, the large standard error for the romance genre (562.2956) suggests instability, possibly due to sparse data in that category. Overall, the model performs well

but may benefit from feature engineering or alternative classification techniques to improve accuracy.

```
model2 <- glm(rating_binary ~ year + length + budget + votes + genre_num,
 ↪  data = data1, family = binomial)
summary(model2)
```

```
Call:
glm(formula = rating_binary ~ year + length + budget + votes +
    genre_num, family = binomial, data = data1)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           -2.2215     0.5814  -3.821 0.000133 ***
year                   0.3890     0.4421   0.880 0.378955
length               -19.5733     1.7256 -11.343  < 2e-16 ***
budget                11.2481     0.8620  13.050  < 2e-16 ***
votes                  4.8041     1.7406   2.760 0.005778 **
genre_numAnimation    -0.8668     0.5398  -1.606 0.108331
genre_numComedy        3.2749     0.2635  12.427  < 2e-16 ***
genre_numDocumentary   5.5619     0.5524  10.068  < 2e-16 ***
genre_numDrama        -2.0779     0.3459  -6.007 1.89e-09 ***
genre_numRomance     -15.4464   562.2956  -0.027 0.978085
genre_numShort         3.5357     1.1431   3.093 0.001980 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1858.7  on 1435  degrees of freedom
Residual deviance:  711.5  on 1425  degrees of freedom
AIC: 733.5

Number of Fisher Scoring iterations: 15
```

Model 3: Logistic Regression for Binary Rating Classification (Without Year Variable)

This logistic regression model is a variation of Model 2, where the year variable has been removed. The model still predicts whether a movie's rating falls into a high or low category, based on length, budget, votes, and genre. The Intercept (-2.0297) remains negative,

suggesting that a movie with average values for all predictors is more likely to receive a low rating.

The budget (11.2226) and votes (4.8916) continue to have a strong positive impact, indicating that movies with larger budgets and higher vote counts are more likely to receive high ratings. Conversely, length (-19.2431), and drama genre (-2.0757) retain their negative,effects, meaning that longer movies and dramas are less likely to receive high ratings. The romance genre still shows an insignificant effect (p = 0.9777) with a high standard error (556.2689), reinforcing potential data sparsity issues in that category.

The removal of the year variable did not significantly impact model performance, as indicated by the residual deviance (712.27, close to Model 2's 711.5) and AIC (732.27 vs. 733.5 in Model 2). The significance levels remain nearly identical for all variables, suggesting that year did not contribute much predictive power. This model is slightly simpler than Model 2 and may be preferred if interpretability and feature selection are prioritized. However, alternative modeling approaches, such as non-linear models or interactions, could be explored for further improvements.

```
model3<-glm(rating_binary ~ length + budget + votes + genre_num, data =
 ↪  data1, family = binomial)
summary(model3)
```

```
Call:
glm(formula = rating_binary ~ length + budget + votes + genre_num,
    family = binomial, data = data1)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           -2.0297     0.5379  -3.773 0.000161 ***
length               -19.2431     1.6783 -11.466  < 2e-16 ***
budget                11.2226     0.8603  13.046  < 2e-16 ***
votes                  4.8916     1.7233   2.838 0.004534 **
genre_numAnimation    -0.8426     0.5387  -1.564 0.117774
genre_numComedy        3.2833     0.2638  12.446  < 2e-16 ***
genre_numDocumentary   5.6177     0.5474  10.262  < 2e-16 ***
genre_numDrama        -2.0757     0.3442  -6.031 1.63e-09 ***
genre_numRomance     -15.5254   556.2689  -0.028 0.977734
genre_numShort         3.6338     1.1388   3.191 0.001418 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 1858.72  on 1435  degrees of freedom
Residual deviance:  712.27  on 1426  degrees of freedom
AIC: 732.27

Number of Fisher Scoring iterations: 15
```

Model Comparison

The difference between Model 2 and Model 3 is that Model 3 adds the extra variable year. However, the regression results show that the coefficient of year is not significant ($p = 0.379$), which means that it does not contribute much to the prediction of rating_binary and cannot significantly improve the classification ability. In addition, year may be related to budget and votes, which may lead to the problem of multicollinearity and affect the stability of the model. Therefore, in order to ensure the interpretability of the model, we remove the year variable, which simplifies the model and maintains the good prediction effect.

Regarding the categorical variable genre_num, there are obvious differences in the impact of different film genres on the ratings. Comedies, documentaries and short films are more likely to get high ratings, especially documentaries have the strongest positive effect. Drama, on the other hand, has a significant negative impact on ratings, indicating that it is more difficult for films of this genre to obtain high ratings. Meanwhile, animated films and romances do not have a significant effect, which may be related to the uneven distribution of the sample or the diversity of rating characteristics.

In the end, we chose Model 3 (without year) as the final logistic regression model to ensure the stability and interpretability of the model, while retaining the analysis of key factors such as film genre, budget, duration and number of votes.

```
AIC(model2, model3)
```

```
       df      AIC
model2 11 733.4958
model3 10 732.2727
```

```
BIC(model2, model3)
```

```
       df      BIC
model2 11 791.4616
model3 10 784.9689
```

Wald Test

The results of the z-tests for both Model 2 and Model 3 reveal that the most significant predictors of the outcome are movie length, budget, votes, and the presence of certain genres such as comedy, documentary, and drama, with p-values consistently showing high significance. In contrast, theyear and romance genre show no significant impact in either model, as evidenced by their non-significant p-values. While the coefficients for animation in both models have somewhat higher p-values, indicating no strong effect, the overall significance of the other predictors suggests that factors like budget and genre play a crucial role in the model's prediction, while year and romance appear to be less influential. These findings suggest that the models are largely consistent, with key predictors being reliably identified across both specifications.

```
wald_test1 <- coeftest(model2)
wald_test1
```

```
z test of coefficients:

                     Estimate Std. Error  z value  Pr(>|z|)
(Intercept)          -2.22150    0.58143  -3.8207 0.0001331 ***
year                  0.38899    0.44212   0.8798 0.3789547
length              -19.57326    1.72561 -11.3428 < 2.2e-16 ***
budget               11.24811    0.86195  13.0496 < 2.2e-16 ***
votes                 4.80411    1.74055   2.7601 0.0057782 **
genre_numAnimation   -0.86685    0.53984  -1.6057 0.1083314
genre_numComedy       3.27486    0.26352  12.4272 < 2.2e-16 ***
genre_numDocumentary  5.56190    0.55241  10.0684 < 2.2e-16 ***
genre_numDrama       -2.07790    0.34589  -6.0074 1.885e-09 ***
genre_numRomance    -15.44637  562.29563  -0.0275 0.9780847
genre_numShort        3.53572    1.14306   3.0932 0.0019801 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
wald_test2 <- coeftest(model3)
wald_test2
```

```
z test of coefficients:

                     Estimate Std. Error  z value  Pr(>|z|)
(Intercept)          -2.02974    0.53791  -3.7734  0.000161 ***
```

```
length                -19.24314    1.67831 -11.4658 < 2.2e-16 ***
budget                 11.22259    0.86026  13.0456 < 2.2e-16 ***
votes                   4.89156    1.72333   2.8384  0.004534 **
genre_numAnimation     -0.84262    0.53870  -1.5642  0.117774
genre_numComedy         3.28325    0.26380  12.4460 < 2.2e-16 ***
genre_numDocumentary    5.61770    0.54742  10.2621 < 2.2e-16 ***
genre_numDrama         -2.07571    0.34419  -6.0307 1.632e-09 ***
genre_numRomance      -15.52537  556.26887  -0.0279  0.977734
genre_numShort          3.63377    1.13877   3.1910  0.001418 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residuals Plot

The residual plots for Models 1, 2, and 3 highlight key issues in their respective fits. Model 1 (Linear Regression) shows a clear pattern in the Pearson residuals, indicating a violation of the linearity assumption, while the funnel shape suggests heteroscedasticity. Model 2 (Logistic Regression with Year) displays Pearson residuals clustered near the extremes, suggesting that certain observations are predicted with high certainty, but others may not fit well. Model 3 (Logistic Regression without Year) follows a similar pattern to Model 2, with residual clustering at the boundaries. These results indicate that while logistic regression is expected to show non-normal residuals, improvements such as feature engineering, interaction terms, or alternative modeling approaches (e.g., decision trees or ensemble methods) could enhance prediction performance and model reliability.

```r
# Function to plot residuals
plot_residuals <- function(model, model_name) {
  residuals_data <- data.frame(
    Fitted_Values = fitted(model),
    Pearson_Residuals = residuals(model, type = "pearson"),
    Deviance_Residuals = residuals(model, type = "deviance")
  )

  # Pearson Residuals Plot
ggplot(residuals_data, aes(x = Fitted_Values, y = Pearson_Residuals)) +
    geom_point(alpha = 0.5) +
    geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
    labs(title = "Pearson Residuals ", x = "Fitted Values", y = "Pearson
     ↪  Residuals") +
    theme(plot.title = element_text(size = 12, hjust = 0.5))

}
```

```
# Generate residual plots for Model 1, 2, and 3
plot_residuals(model1, "Model 1")
```

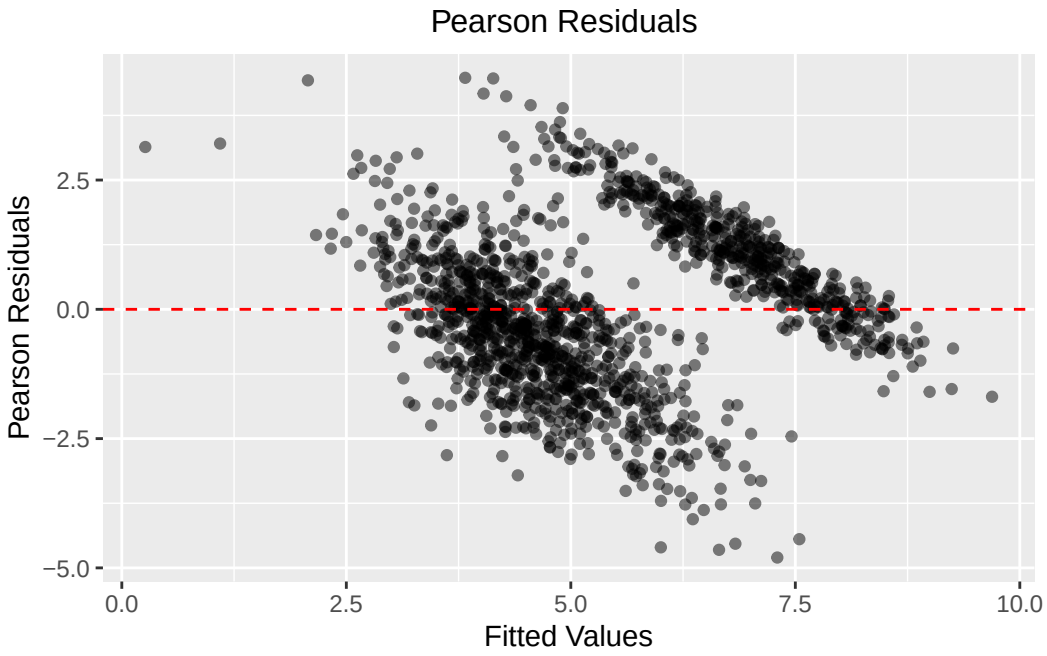

Figure 6: Pearson Residuals(model1)

```
plot_residuals(model2, "Model 2")
```

Figure 7: Pearson Residuals(model2)

```
plot_residuals(model3, "Model 3")
```

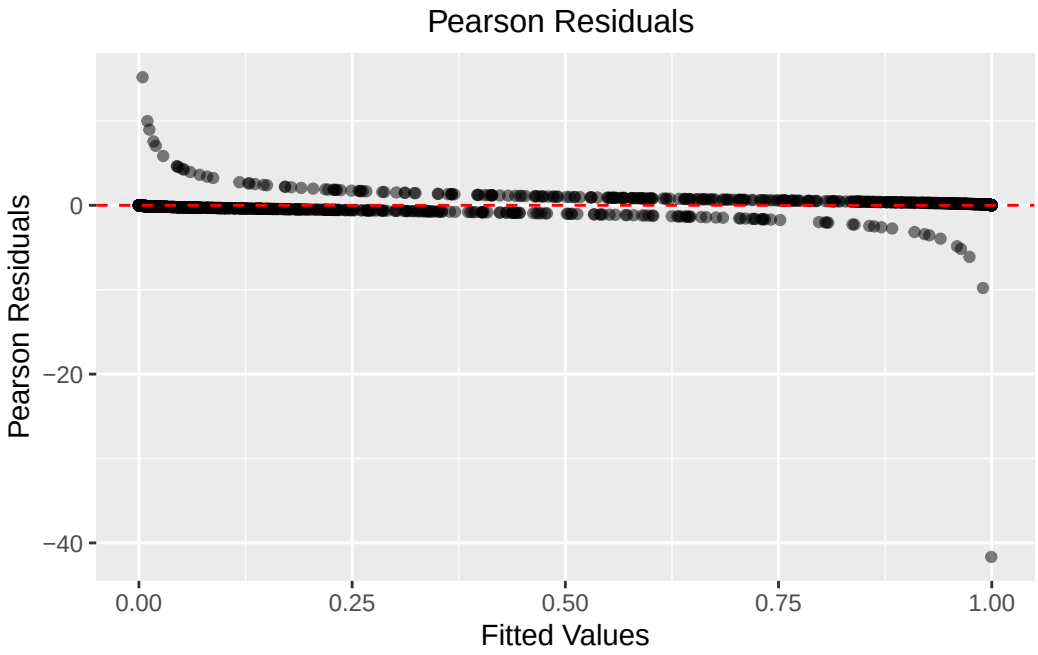

Figure 8: Pearson Residuals(model3)

Visualization of Model Predictions

The histogram illustrates the predicted probabilities of a binary movie rating (0 or 1) from Model 2. The x-axis represents the predicted probability, while the y-axis shows the count of observations. The red bars correspond to movies classified as 0 (low rating), and the blue bars correspond to 1 (high rating). The plot reveals a strong separation, with most low-rated movies having predicted probabilities near 0, and high-rated movies concentrated around 1, suggesting that the model effectively differentiates between the two classes. However, there is a small region of overlap in the middle, indicating some misclassification or uncertainty in predictions. This suggests that while the model performs well overall, it may benefit from further refinements such as adding interaction terms or addressing potential data imbalances.

```
data1$predicted_probs <- predict(model2, type = "response")
ggplot(data1, aes(x = predicted_probs, fill = factor(rating_binary))) +
  geom_histogram(binwidth = 0.05, alpha = 0.7, position = "identity") +
  labs(title = "Predicted Probabilities of Binary Rating", x = "Predicted
   ↪  Probability", y = "Count")
```

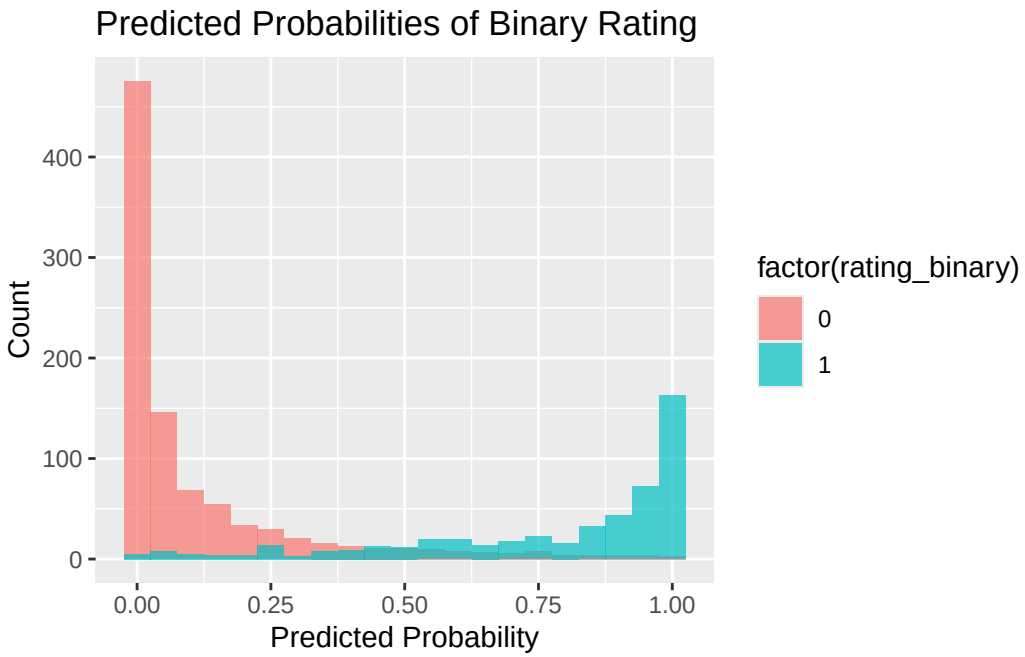

Figure 9: Predicted Probabilities of Binary Rating

Conclusion

The analysis of the logistic regression models for predicting binary movie ratings provided valuable insights. Model 1, which included all predictors, and Model 2, which excluded the year variable, both demonstrated strong predictive power with similar residual deviance and AIC values. Model 3, which further simplified the model by removing year, showed only a slight increase in residual deviance, suggesting that the year variable may not be a crucial predictor. Across all models, budget, votes, and movie length were highly significant, with genre also playing a role, particularly for Comedy, Documentary, and Drama films. Interestingly, some genres, such as Romance, did not show strong statistical significance, which may indicate either a weaker relationship with rating success or data limitations.

Residual plots for all three models revealed some issues, particularly in the Pearson residuals for Model 1, which displayed a distinct pattern, indicating potential model misfit. Additionally, the predicted probability histogram for Model 2 illustrated a strong separation between low-rated and high-rated movies, with most low-rated movies having probabilities near zero and high-rated movies clustering around one. However, some overlap between the two classes suggests possible misclassification and areas where the model could be improved.

Based on the model selection criteria, we choose **Model 3**, which does not include the `year` variable, as it has the lowest AIC value. The analysis also indicates that the `year` variable and

the movie identifier are **not statistically significant** (p-value > 0.05), meaning they do not meaningfully contribute to predicting the binary rating outcome. Removing these variables simplifies the model without losing predictive power. Future work may explore additional predictors or alternative modeling approaches to further enhance predictive performance. »»»> Stashed changes

Overall, while the models successfully classify movies based on key attributes, there are opportunities for refinement. Alternative modeling approaches, such as non-linear transformations, interaction terms, or more advanced machine learning techniques like decision trees and ensemble methods, could improve predictive performance. Additionally, checking for multicollinearity, considering additional relevant predictors, or experimenting with different threshold values for classification may further enhance the model's accuracy. Despite some residual concerns, these models provide a solid foundation for predicting movie success and offer a useful tool for understanding the factors influencing audience ratings.