MOVIE DATA ANALYSIS

# PREDICTING AUDIENCE SCORES: A Linear Regression Analysis of Movie Features

Ariella Asti Cahyani
02113069

START SLIDE

# RESEARCH QUESTION :

WHAT FACTORS SIGNIFICANTLY INFLUENCE THE AUDIENCE SCORES OF MOVIES?

## 🧠 Why This Question?

🎥 Audience score is a key indicator of how well a movie is received by the public.

💡 Understanding what drives audience ratings helps filmmakers improve content

📊 Shifts focus from critics to actual viewers

🔍 Reveals key factors like IMDb rating, genre, and more

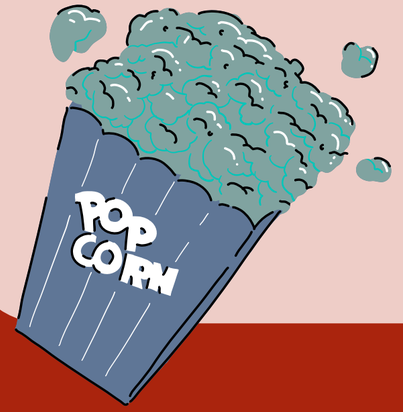
# DATA OVERVIEW

- DATA: 651 RANDOMLY SAMPLED MOVIES RELEASED BEFORE 2016
- VARIABLES: 32 MOVIE-RELATED VARIABLES
- SOURCE: ROTTEN TOMATOES & IMDB APIs
- CAUSALITY: OBSERVATIONAL DATA – NO CAUSAL CONCLUSIONS
- VARIABLE SELECTION: GENRE, RUNTIME, MPAA_RATING, IMDB_RATING, CRITICS_SCORE, SOURCE
- DATA QUALITY: ADDRESSED MULTICOLLINEARITY & MISSING VALUES FOR VALID RESULTS

# VARIABLES IN THE MODEL

## RESPONSE VARIABLE (OUTCOMES):

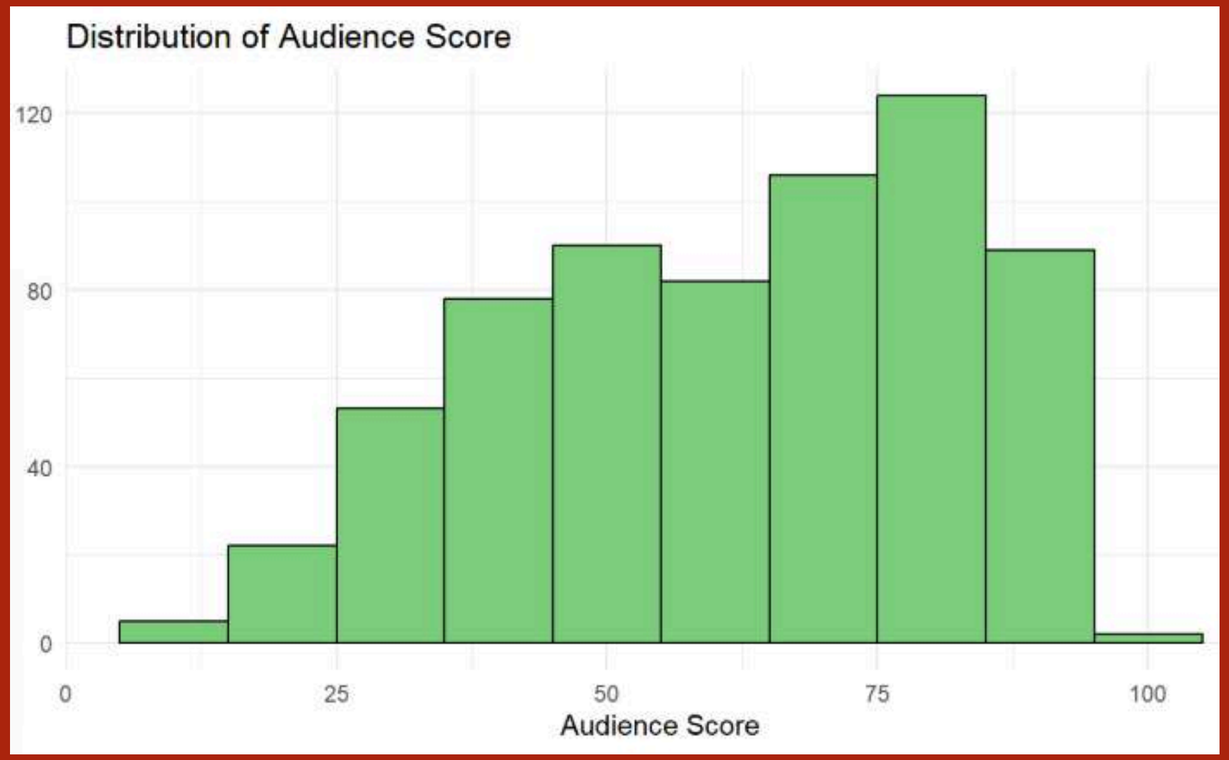
- Audience Score → reflects general viewer reception



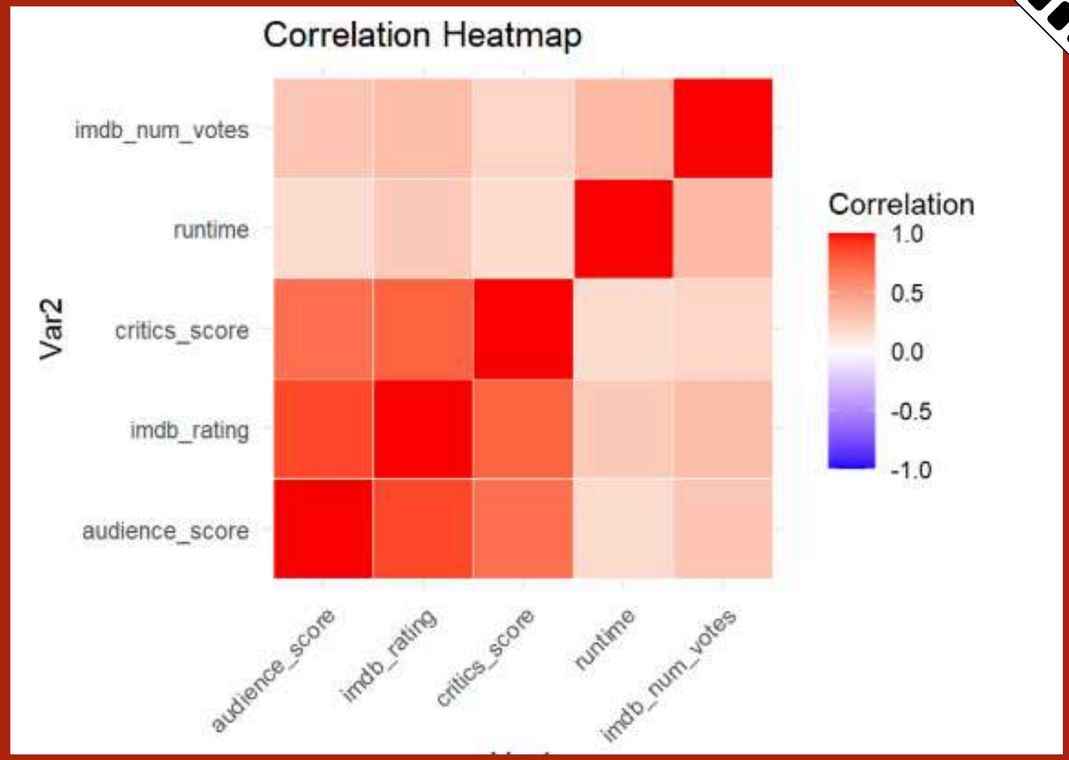

## EXPLANATORY VARIABLE (PREDICTOR):

- Genre → includes categories like Animation, Horror, etc.
- Runtime → movie length in minutes
- MPAA Rating → G, PG, PG-13, R
- IMDb Rating → viewer-based rating from IMDb
- Critics Score → critic-based score from Rotten Tomatoes and IMDB APIs

# EXPLANATORY DATA ANALYSIS

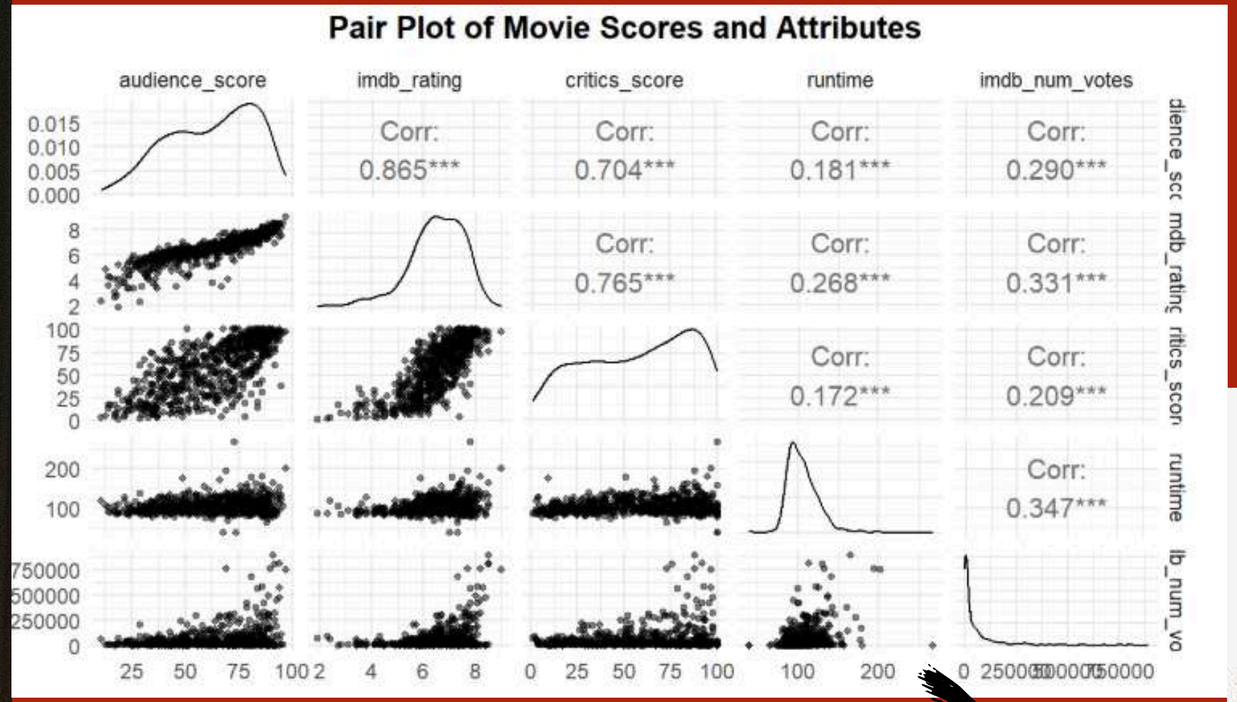



- Most movies have high audience scores, peaking between 75–87.5.
- Slight left skew, with fewer low-scoring films pulling the average down.



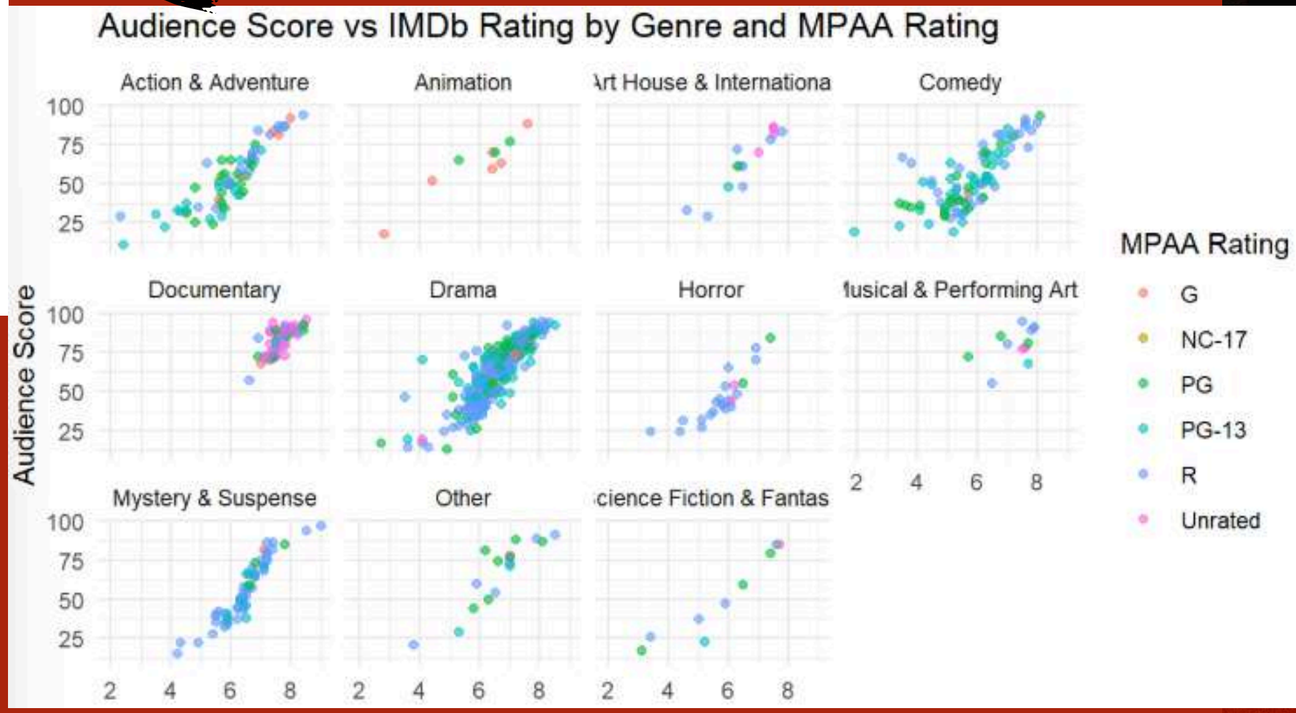

- Strong positive correlations between audience score, IMDb rating, and critics score.
- Runtime shows weak correlation, while vote count moderately aligns with scores.

# EXPLANATORY DATA ANALYSIS

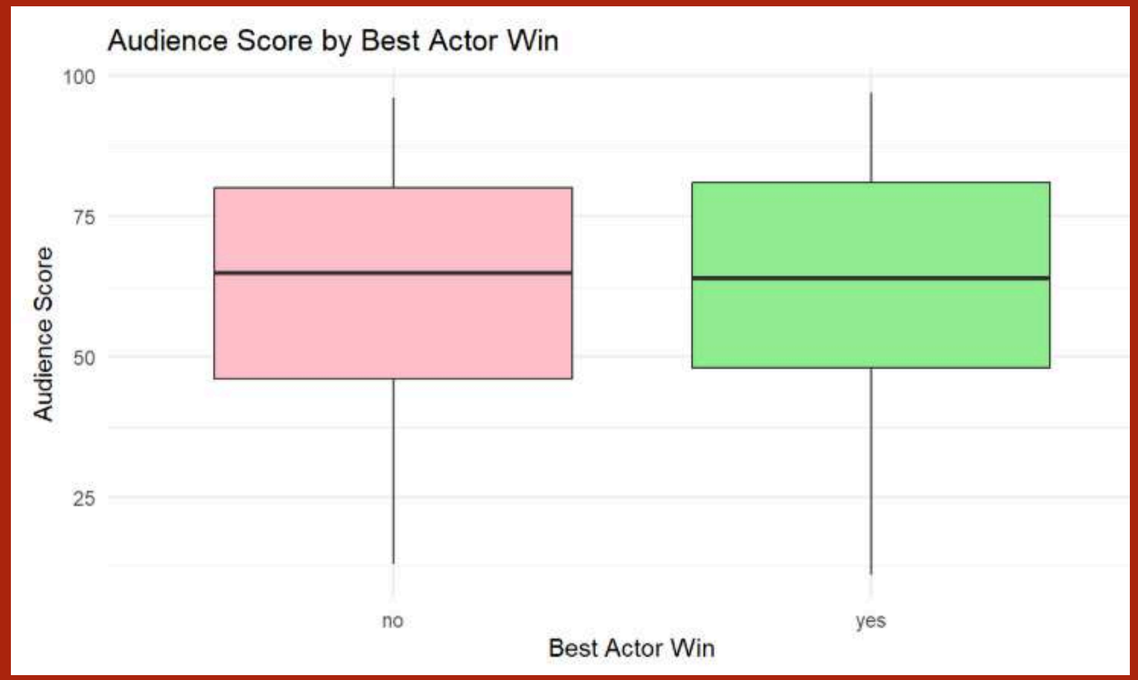


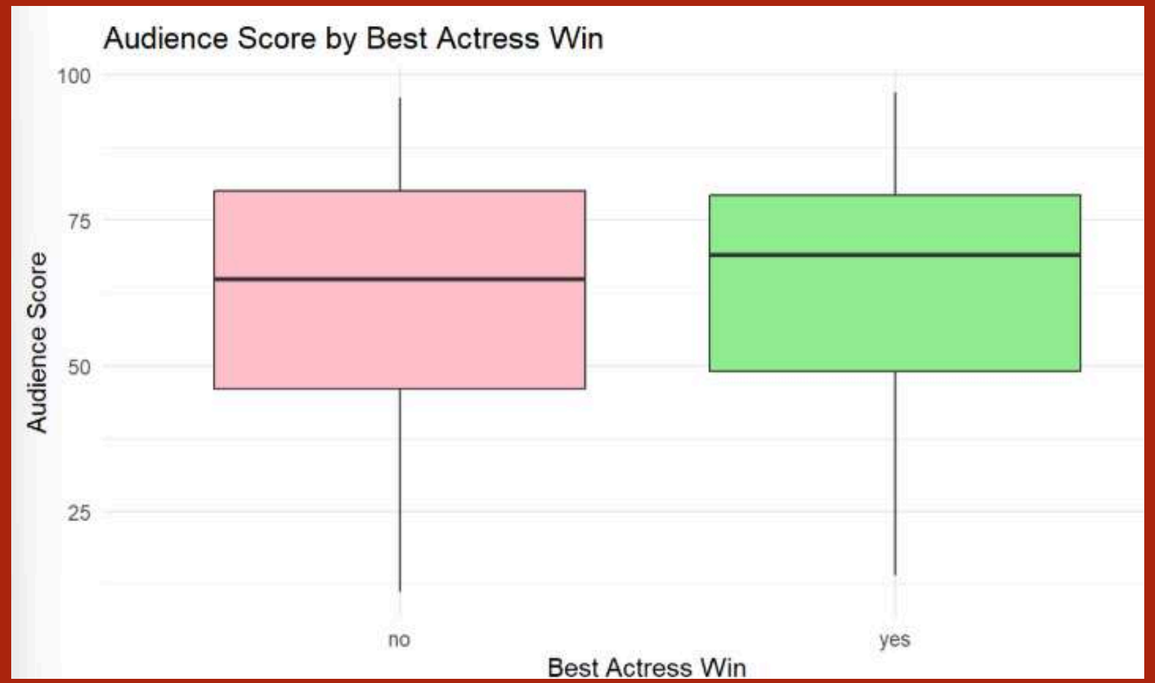

- Audience, IMDb, and critics scores show strong positive correlations, suggesting consistency in perceived movie quality.
- Runtime and number of votes show weaker yet significant relationships with scores.

- IMDb ratings and audience scores show a strong positive relationship across most genres, especially in Action, Documentary, and Sci-Fi.
- Genres like Comedy and Horror show more variation, while Animation tends to receive consistently high ratings on both metrics.



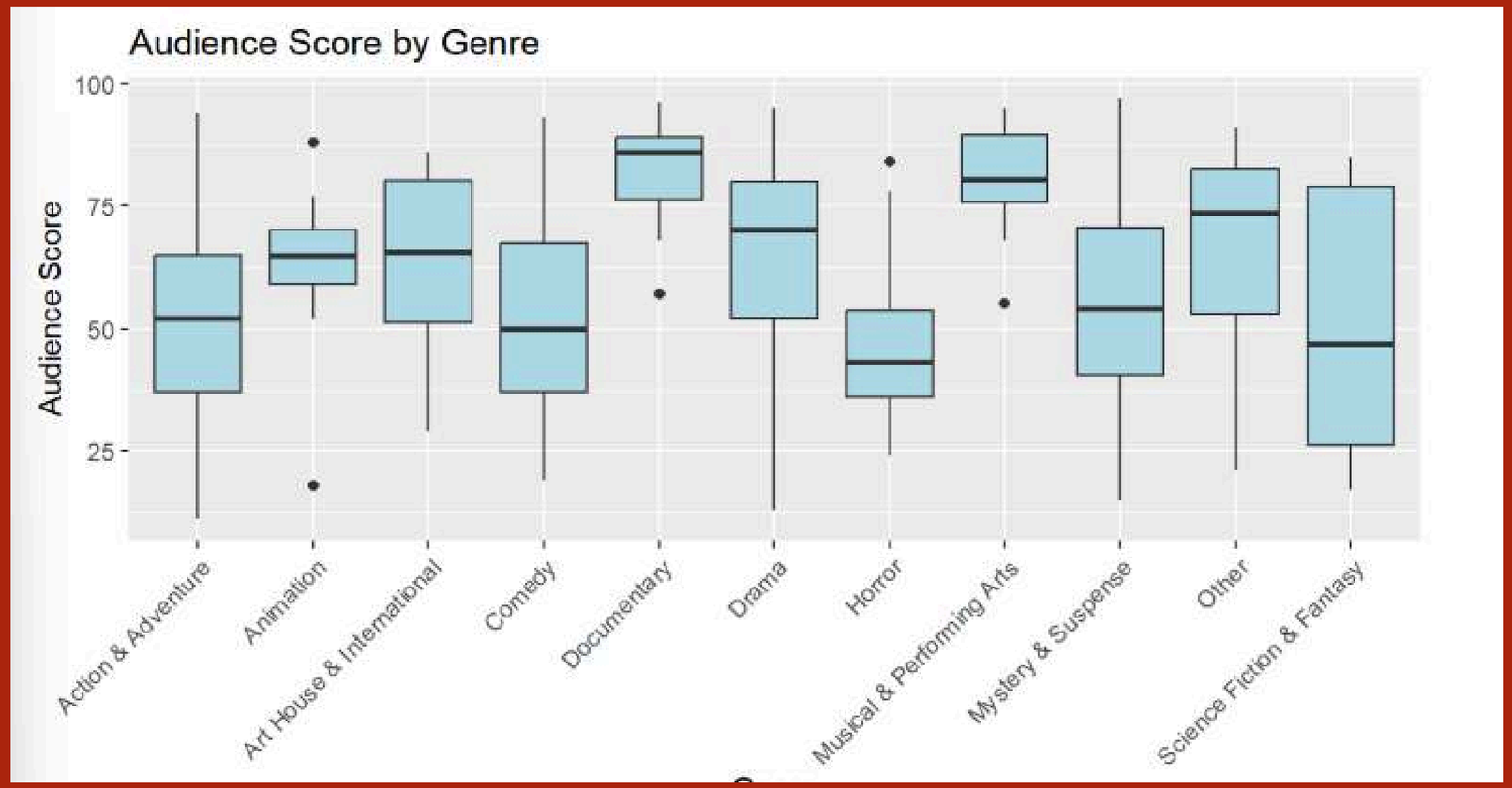
# EXPLANATORY DATA ANALYSIS




- Audience scores are similar regardless of Best Actor wins, but movies with a Best Actress win show a slight upward shift in typical scores.

# EXPLANATORY DATA ANALYSIS




- Documentary and Musical genres have the highest and most consistent audience scores, while Horror and Sci-Fi/Fantasy show lower and more varied scores.
- Sci-Fi/Fantasy has the widest range of audience reactions, with some films scoring very high or very low compared to other genres.

# MODELING

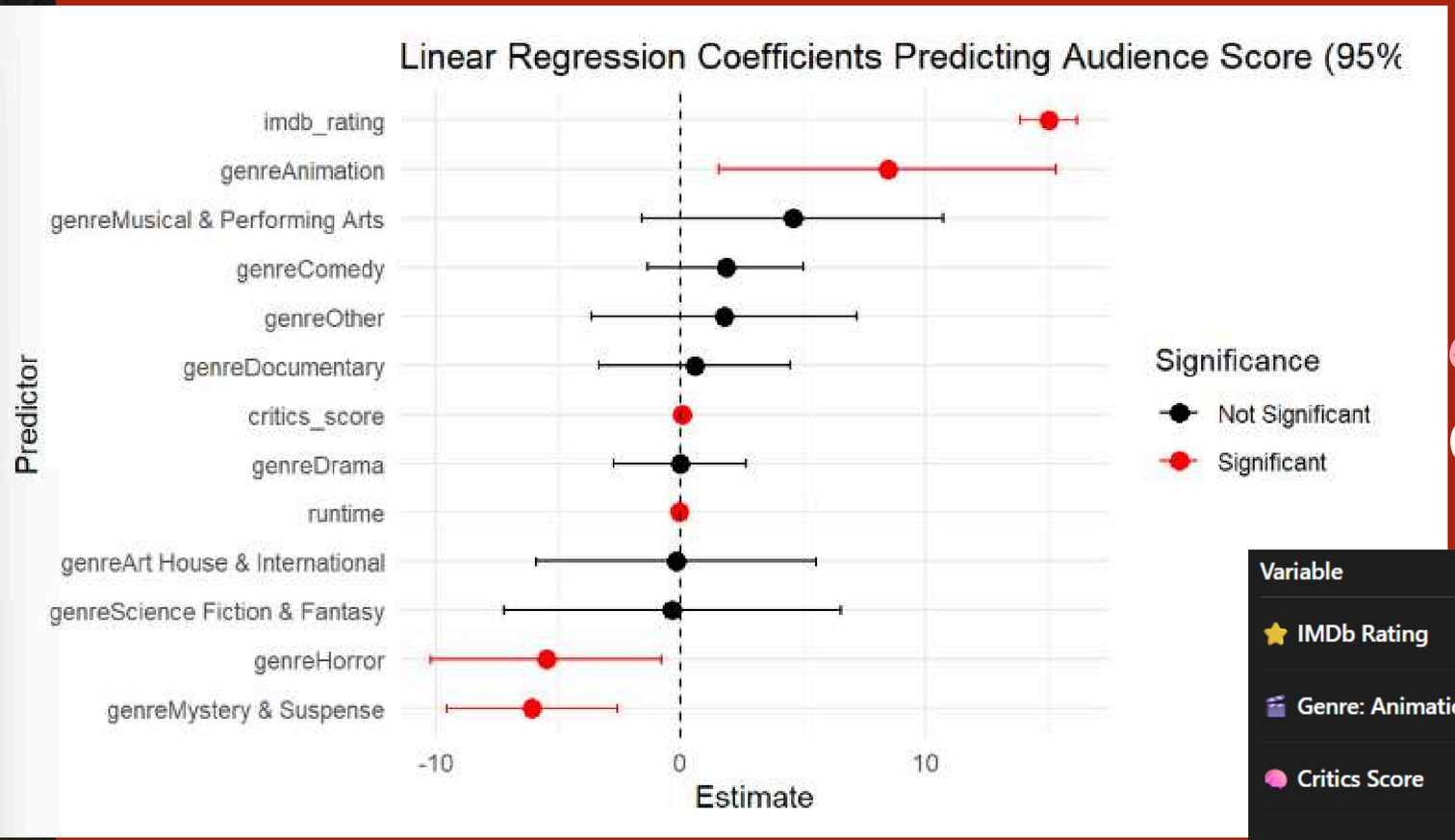
```r
# Fit a linear regression model
model <- lm(audience_score ~ imdb_rating + critics_score + runtime + genre + mpaa_rating, data = movies)
summary(model)
```

```
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -34.59073    4.21562  -8.205 1.29e-15 ***
## imdb_rating                     15.02730    0.58902  25.512  < 2e-16 ***
## critics_score                    0.06320    0.02178   2.902  0.00384 **
## runtime                         -0.04209    0.02229  -1.888  0.05948 .
## genreAnimation                   8.49679    3.83345   2.216  0.02701 *
## genreArt House & International  -0.22976    2.97398  -0.077  0.93844
## genreComedy                      1.93987    1.63586   1.186  0.23613
## genreDocumentary                 0.25499    2.25255   0.113  0.90991
## genreDrama                       0.12498    1.41760   0.088  0.92977
## genreHorror                     -5.38764    2.45102  -2.198  0.02830 *
## genreMusical & Performing Arts   4.47095    3.16247   1.414  0.15793
## genreMystery & Suspense         -5.86328    1.82390  -3.215  0.00137 **
## genreOther                       1.59775    2.77788   0.575  0.56538
## genreScience Fiction & Fantasy  -0.36507    3.50848  -0.104  0.91716
## mpaa_ratingNC-17                -3.75396    7.44514  -0.504  0.61429
## mpaa_ratingPG                    1.12300    2.71223   0.414  0.67898
## mpaa_ratingPG-13                -0.08796    2.79094  -0.032  0.97487
## mpaa_ratingR                     0.18145    2.68803   0.068  0.94620
## mpaa_ratingUnrated               0.99141    3.07018   0.323  0.74686
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.822 on 632 degrees of freedom
## Multiple R-squared:  0.7706, Adjusted R-squared:  0.7641
## F-statistic:   118 on 18 and 632 DF,  p-value: < 2.2e-16
```

**Model predicts audience scores well (76% explained).**

- IMDb rating = strongest positive effect (+15 points per rating).
- Critics score = small positive effect (+0.06 points).
- Animation genre boosts scores; Horror & Mystery lower them.
- Longer runtime slightly decreases scores.
- MPAA rating and other genres not significant.

# MODELING



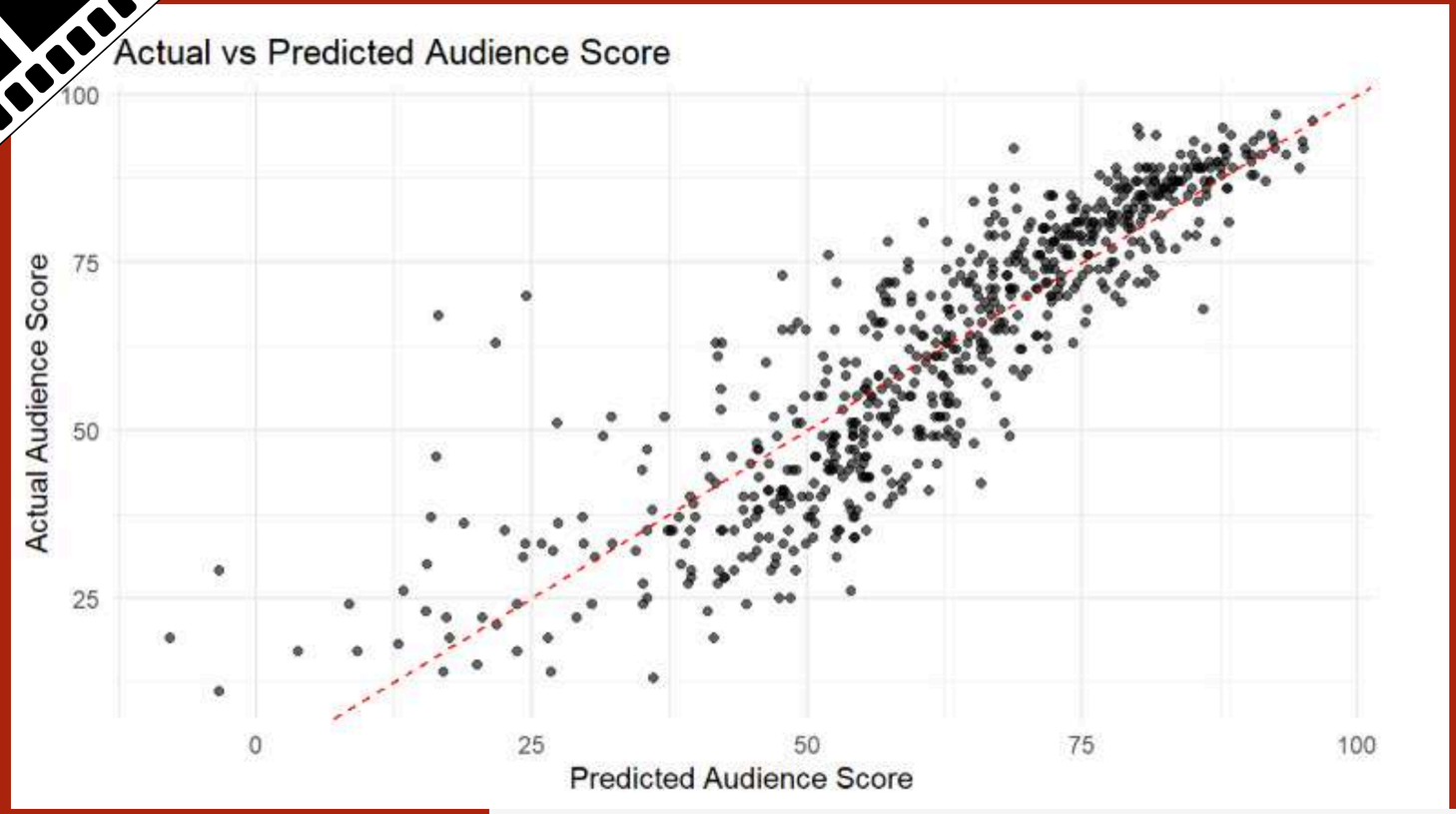

## 🧠 Which Variables are Significant?

| Variable | Effect on Audience Score | Notes |
|---|---|---|
| ⭐ IMDb Rating | 🔺 Strong positive (+15 points) | Most influential factor |
| 🎬 Genre: Animation | 🔺 Positive (+8 points) | Animated films are favored |
| 🧠 Critics Score | 🔺 Slight positive (+0.06 points) | Higher critic scores help |
| ⏱️ Runtime | 🔻 Slight negative (–0.04 points) | Longer films slightly hurt |
| 👻 Genre: Horror | 🔻 Negative (–5 points) | Horror is less liked |
| 🔪 Genre: Mystery & Suspense | 🔻 Negative (–6 points) | Less appealing to audiences |

## ❌ Not Significant Variables

- Genres like Comedy, Drama, Documentary, etc.
- MPAA Ratings (PG, PG-13, R, etc.)
- ➤ These variables do not show a meaningful effect on Audience Score in this model.

# MODELING




✅ Positive correlation: Predicted scores align well with actual scores — the model captures overall trends accurately.

❗ Prediction errors exist: Some deviations from the perfect line indicate room for improvement in precision.

```r
# Predict Audience Score using the model (optional)
movies$predicted_audience_score <- predict(model, newdata = movies)

# Plot actual vs predicted Audience Score
ggplot(movies, aes(x = predicted_audience_score, y = audience_score)) +
  geom_point(alpha = 0.6) +
  geom_abline(slope = 1, intercept = 0, color = "red", linetype = "dashed") +
  labs(title = "Actual vs Predicted Audience Score",
       x = "Predicted Audience Score",
       y = "Actual Audience Score") +
  theme_minimal()
```

# PREDICTION

```r
# Prepare new data
new_movies <- data.frame(
  title = c("Zootopia", "Deadpool", "La La Land", "Before the
Flood", "The Witch"),
  genre = factor(c("Animation",
                   "Action & Adventure",
                   "Musical & Performing Arts",
                   "Documentary",
                   "Horror"),
                 levels = levels(movies$genre)),
  runtime = c(108, 108, 128, 96, 92),
  mpaa_rating = factor(c("PG", "R", "PG-13", "PG", "R"),
                       levels = levels(movies$mpaa_rating)),
  imdb_rating = c(8.0, 8.0, 8.0, 8.2, 6.9),
  critics_score = c(98, 85, 91, 75, 90),
  critics_source = c("Rotten Tomatoes",
                     "Metacritic",
                     "IMDb Metascore",
                     "Geo National",
                     "Rotten Tomatoes"),
  audience_score = c(92, 96, 81, 85, 55)
)
```

This is a new dataset created to test the model's prediction performance on unseen movies.

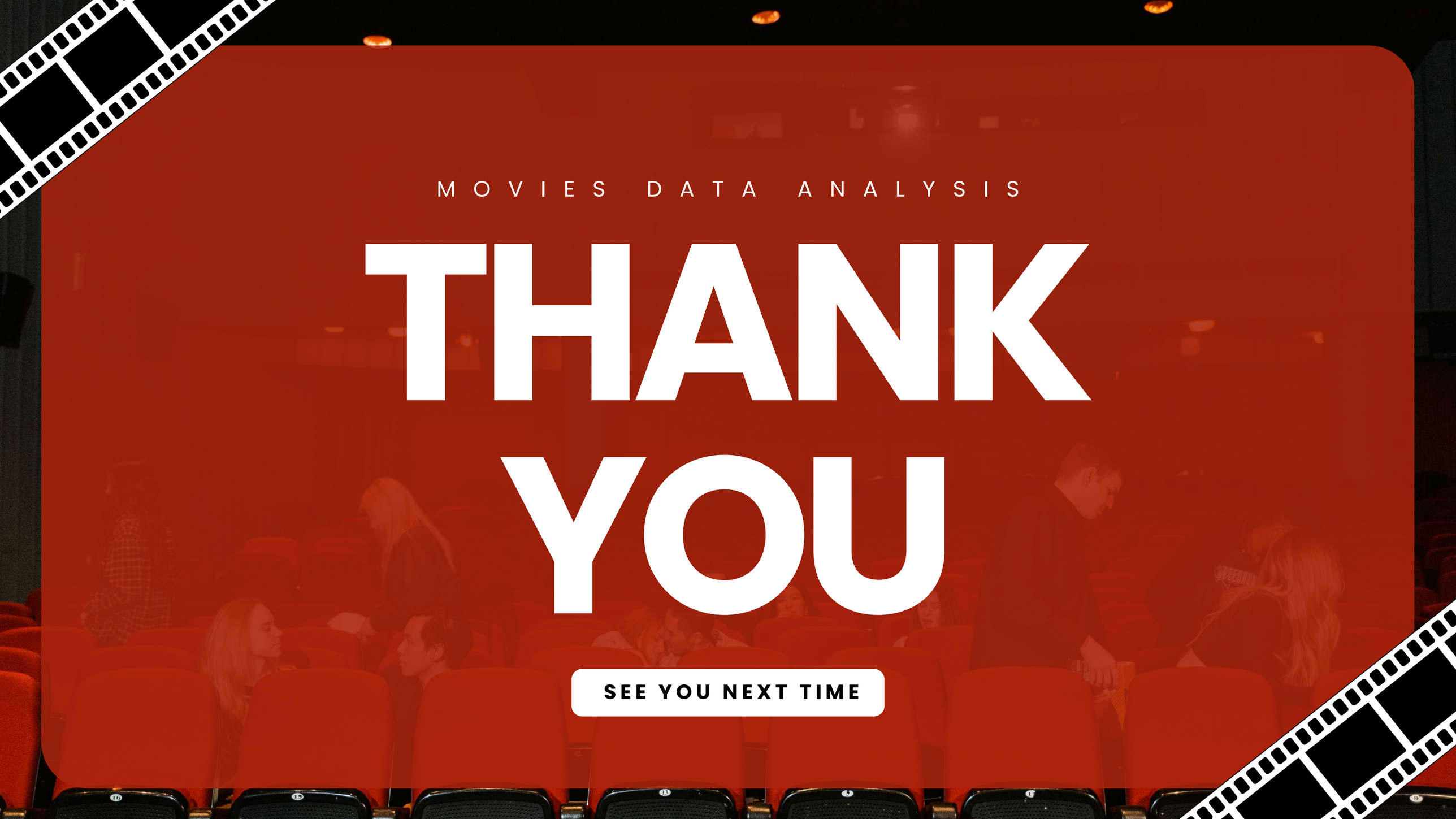
| Title | Genre | Runtime | MPAA Rating | IMDb Rating | Critics Score | Critics Source | Audience Score |
|---|---|---|---|---|---|---|---|
| Zootopia | Animation | 108 | PG | 8.0 | 98 | Rotten Tomatoes | 92 |
| Deadpool | Action & Adventure | 108 | R | 8.0 | 85 | Metacritic | 96 |
| La La Land | Musical & Performing Arts | 128 | PG-13 | 8.0 | 91 | IMDb Metascore | 81 |
| Before the Flood | Documentary | 96 | PG | 8.2 | 75 | Geo National | 85 |
| The Witch | Horror | 92 | R | 6.9 | 90 | Rotten Tomatoes | 55 |

# PREDICTION & ACTUAL

| Title | Actual Score | Predicted Score | Difference |
|---|---|---|---|
| Zootopia | 92 | 96.18 | +4.18 |
| Deadpool | 96 | 86.90 | -9.10 |
| La La Land | 81 | 90.99 | +9.99 |
| Before the Flood | 85 | 90.32 | +5.32 |
| The Witch | 55 | 65.92 | +10.92 |

# CONCLUSION

- This linear regression model explains 76% of the variation in audience scores.
- IMDb rating is the strongest predictor, followed by critics score and certain genres.
- Some genres like Animation boost scores, while Horror lowers them.
- Runtime and MPAA ratings have minimal impact.
- The model shows good predictive power but has some issues with outliers and unequal variance.
- Further improvements could include robust methods or data transformations.


MOVIES DATA ANALYSIS
THANK YOU
SEE YOU NEXT TIME