Cinema Revenue Analysis: Executive Report

Azra Narvel

2025-12-15

Introduction

This report analyzes factors that predict the revenue generated by a movie, with a focus on whether increased production budgets are associated with higher revenue. In the film industry, production budgets reach tens or hundreds of millions of dollars, while successful films can generate hundreds of millions or even billions in worldwide revenue. As a result, decisions about how much to invest in a project carry substantial financial risk.

Using statistical modeling, this analysis examines the relationship between movie revenue and several key variables, including production budget, viewer ratings, and genre indicators. The report includes exploratory data analysis, model selection, diagnostic evaluation, and interpretation of results, with the goal of assessing the expected return to increased production spending.

Data Description

The data set consists of approximately 100 motion pictures directed by 30 different directors. For each movie, the data include worldwide gross revenue, production budget, viewer ratings, run-time, production date, director name, director birth year, and three associated genres. Gross revenue and budget are recorded in nominal U.S. dollars, while viewer ratings are measured on a 1–10 scale. A summary of all variables, including definitions and units of measurement, is provided in Table 1.

Variable	Description
Movie Title	The name of the movie.
Production Date	The official date the movie was produced.
Genres	The three main genres associated with the movie. All holding equal weight.
Run-time (minutes)	The length of the movie.
Director Name	The name of the director of the movie.
Director Birth Year	The birth year of the director of the movie.
Average Rating	The quality rating given to the movie by viewers on a 1-10 scale, with 10 being best.
Budget (dollars)	The movie production budget.
Gross (dollars)	The movie's worldwide gross revenue.

Table 1: Description of variables in data set.

Director birth year and production date were not included in the analysis, as they were not directly relevant to the revenue relationships examined.

Exploratory Data Analysis

Exploratory data analysis was conducted to examine the distributions of key variables and explore preliminary relationships between movie characteristics and revenue. Univariate distributions indicate that production budget and gross revenue are highly right-skewed, with a small number of films earning substantially more than the rest of the sample. After log transformation, both variables show more symmetric distributions and linear relationships meeting the assumptions of normality (Appendix B2 and Appendix C).

Viewer ratings and run-time are more tightly distributed than budget and gross revenue. Ratings are concentrated around the middle of the scale, and run-times are grouped around standard film durations of 90-120 minutes. In contrast to the wide dispersion observed in budget and gross revenue, this indicates less variation in these variables (Appendix B).

Bivariate analysis reveals a strong positive relationship between log-transformed production budget and log-transformed gross revenue, indicating that higher-budget films tend to generate higher revenue on average (Appendix D). A positive association is also observed between viewer ratings and revenue, though this relationship appears weaker and more variable. In contrast, run-time shows no clear relationship with revenue, suggesting limited predictive value. Formal statistical testing supported these exploratory findings, and run-time was therefore excluded from the final model. Scatterplots illustrating these bivariate relationships are shown in Figure 1.

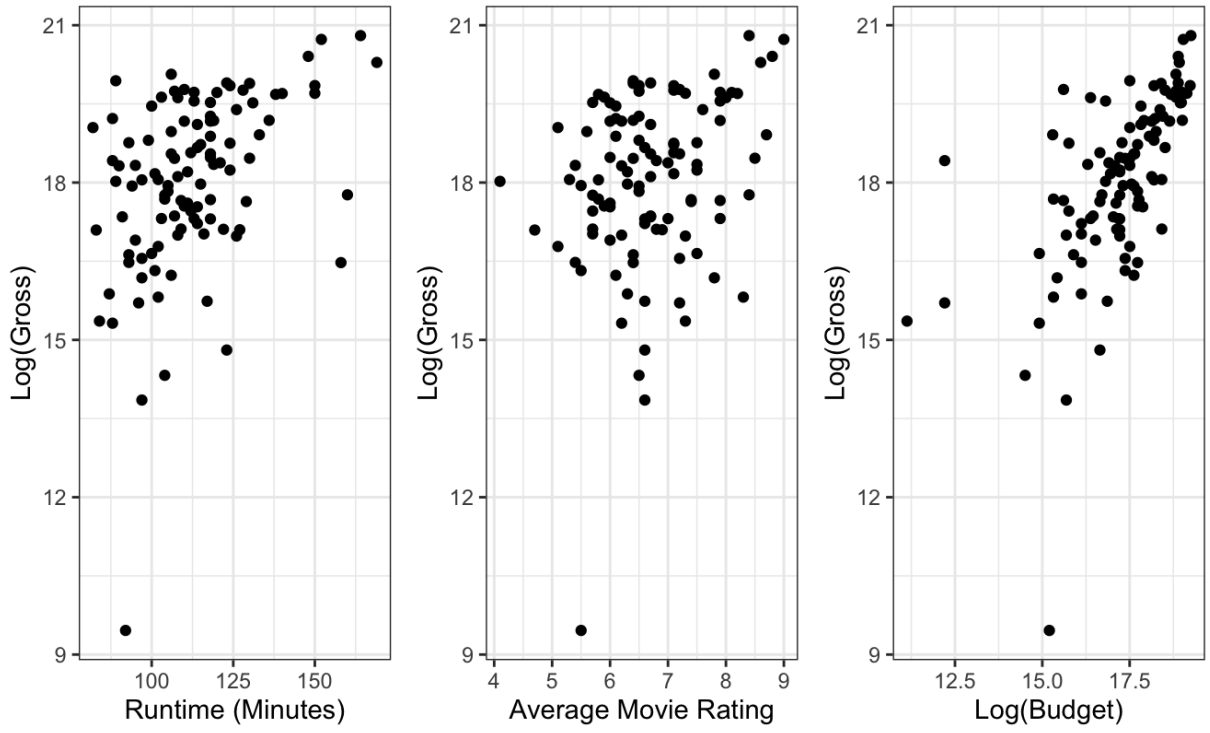


Figure 1: Bivariate relationships between gross revenue and selected predictors (Runtime, Average Movie Rating, Log(Budget).

Genre-based exploratory analysis began by examining the most frequently occurring genres in the dataset. The highest-frequency genres were identified and visualized to assess their prevalence across films (Appendix E1). Although drama was the most common genre, exploratory and modeling results indicated a negative association with revenue once other factors were considered. As a result, attention shifted to the next most prevalent genre, action. Action films tend to earn higher revenue than non-action films on average and is

evident in both distributional comparisons and subsequent modeling results, motivating the inclusion of an action genre indicator in the regression analysis (Appendix E1).

Director-level exploration was also conducted to assess whether specific directors were associated with higher revenue outcomes. The director with the highest average gross revenue, Christopher Nolan, was identified and analyzed using an indicator variable (Appendix E2). However, incorporating this variable did not meaningfully improve model fit. Most directors in the dataset were represented by only a small number of films. Including director-specific indicators under these conditions would lead to unstable estimates and an increased risk of over fitting, reducing the model's ability to generalize beyond the observed sample.

Overall, the exploratory analysis informed several key modeling decisions, including the use of log transformations, the exclusion of run-time and director variables, and the inclusion of the action genre indicator.

Model Selection and Specification

Model selection was guided by strong relationships observed during exploratory analysis and supported by adjusted R^2, AIC, and nested F-tests (Appendix F). A baseline model including production budget and movie average rating was first considered. Additional variables were evaluated individually to assess whether they meaningfully improved model fit. Run-time and director indicators did not improve model performance and were excluded. In contrast, including an action genre indicator improved model fit, as reflected by a lower AIC and a significant F-test, and was therefore retained.

The final regression model estimates the relationship between movie revenue and production budget, viewer rating, and action genre classification, with gross revenue modeled on the logarithmic scale. The fitted model explains approximately 47% of the variability in log-transformed gross revenue, indicating that a substantial portion of revenue variation is captured by the included predictors (Appendix G).

$$log(Gross) = \beta_0 + \beta_1 log(Budget) + \beta_2(AverageRating) + \beta_3(ActionIndicator)$$

The estimated coefficients provide insight into how the key predictors are associated with revenue:

- **Production Budget (log-transformed):** The estimated coefficient of approximately 0.62 indicates that, holding viewer rating and genre constant, a 1% increase in production budget is associated with an expected 0.62% increase in gross revenue. This suggests diminishing returns to increased spending, as revenue increases at a slower rate than budget.
- **Movie Average Rating:** A one-unit increase in average viewer rating is associated with an expected 63% increase in gross revenue, holding budget and genre constant, calculated as exp(.49) - 1. This indicates that higher-rated films tend to earn substantially more revenue, even after independent of production budget.
- **Action Genre Indicator:** Holding budget and rating constant, action films are estimated to earn approximately 77% more revenue than non-action films, calculated by as exp(.57) - 1. This reflects the strong financial performance associated with action movies.
- **Intercept:** The intercept represents the expected log gross revenue for a non-action film with a production budget of one dollar and an average rating of zero. It is is not meaningful on its own and rather serves as a baseline for the model.

Model diagnostics were examined to assess whether the assumptions of linear regression were satisfied (Appendix G). The residuals versus fitted values plot shows no strong violations of linearity or constant variance on the log scale (Figure 2). A normal Q–Q plot of standardized residuals indicates approximate normality.

While some increased variability is observed and few outliers are present, these features are typical for revenue data and do not indicate major violations of model assumptions. Overall, the diagnostics support the validity of the final model for inference and prediction.

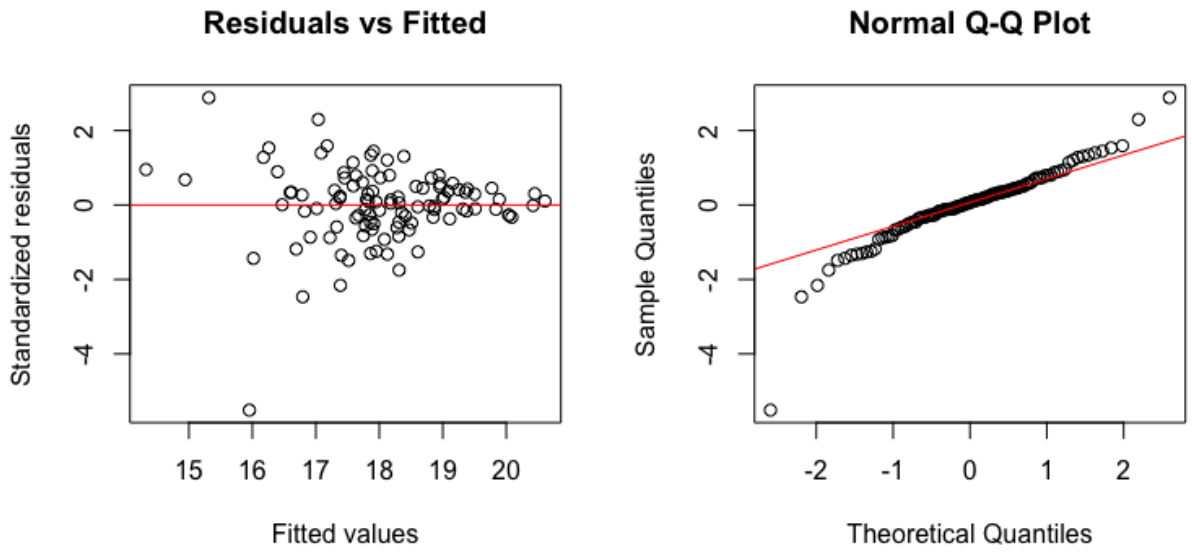


Figure 2: Diagnostic plots for the final regression model, including residuals versus fitted values and a normal Q–Q plot.

Prediction for a New Film

Using the final regression model, revenue was predicted for a hypothetical new movie with a production budget of $10 million, an average viewer rating of 7.0, and a non-action genre classification. The model predicts an expected gross revenue of approximately **$31 million**. A 95% prediction interval for this estimate ranges from roughly **$2.8 million to $349 million**, reflecting substantial uncertainty in individual movie outcomes (Appendix H).

The wide prediction interval highlights the inherent variability in movie revenue. Even after accounting for production budget, viewer ratings, and genre, a large portion of variation in movie revenue remains unexplained. This reflects the influence of factors not included in the dataset, such as marketing expenditure, franchise status, competition at release, and broader market conditions. As a result, while the model provides a reasonable estimate of expected revenue, precise predictions for individual films remain highly uncertain.

Budget Justification

The estimated effect of production budget provides important insight into whether increased spending is justified. The positive coefficient on log-transformed budget indicates that higher production budgets are associated with higher expected gross revenue, holding viewer rating and genre constant. However, the estimated coefficient of approximately 0.62 suggests diminishing returns, as percentage increases in budget are associated with smaller percentage increases in revenue.

This implies that while increasing a film's budget may raise expected revenue on average, each additional increase in spending produces a less-than-proportional gain in revenue. Larger investments therefore carry

greater financial risk, and higher budgets alone do not guarantee strong revenue generation. Budget decisions should be made alongside broader market factors rather than budget alone.

The estimated effect of production budget may be subject to bias due to unobserved factors correlated with higher-budget films. The observed relationship likely reflects not only production spending itself, but also other advantages that tend to accompany larger budgets, such as increased marketing efforts, wider distribution, and more favorable release strategies. As a result, the estimated budget coefficient may overstate the direct impact of production spending on revenue. Incorporating additional information on marketing expenditures, franchise status, and other market factors would allow for a more precise assessment of the return to increased budget.

Conclusion

This analysis demonstrates that production budget, viewer ratings, and genre are all meaningfully associated with movie revenue. Higher budgets are linked to higher expected revenue, but with diminishing returns and substantial uncertainty at the individual film level.

While increased production budgets may be justified on average, they do not eliminate the inherent risk in movie production. Revenue outcomes vary widely, and success depends on more than budget alone. Budget increases should be evaluated as part of a broader decision-making framework.

Appendix

Appendix A: Data Preparation

```
# Load packages
library(ggplot2)
library(tidyverse)
library(patchwork)

# Import data
raw_data <- read_csv("dataset79.csv")

# Check for missing values and duplicate rows
sum(is.na(raw_data))
```

```
## [1] 0
```

```
any(duplicated(raw_data))
```

```
## [1] FALSE
```

```
# Create log-transformed variables
raw_data$logBudget <- log(raw_data$budget)
raw_data$logGross <- log(raw_data$gross)
```

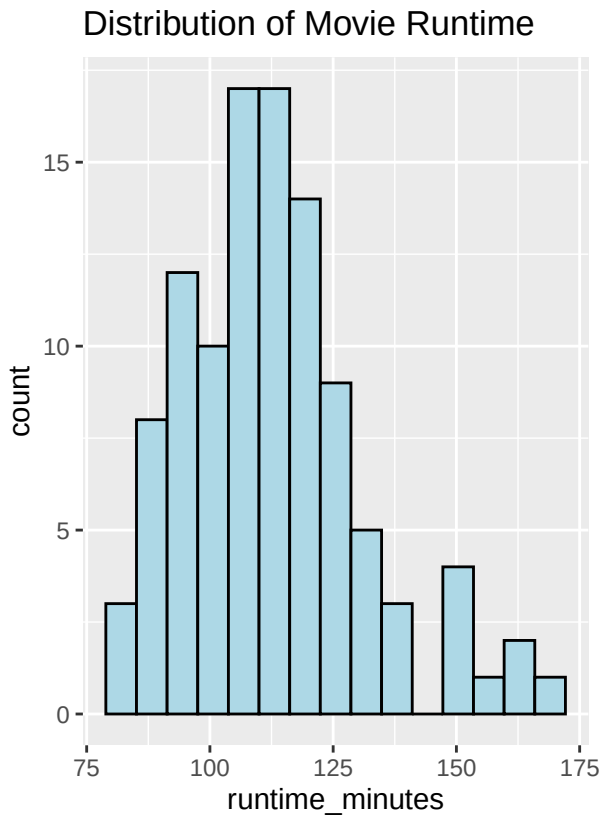
Appendix B: Univariate Analysis

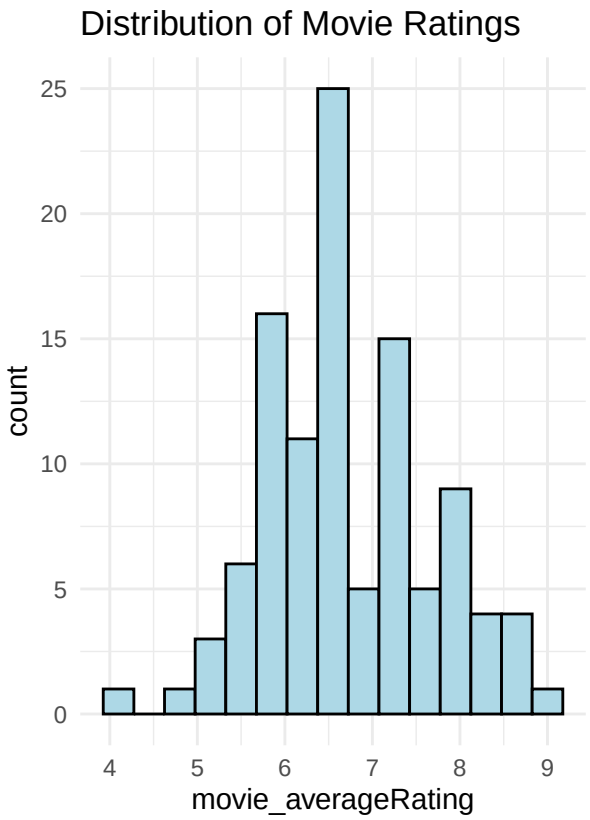
Appendix B1: Run-time and rating

```
p_runtime <- ggplot(raw_data, aes(x = runtime_minutes)) +
  geom_histogram(bins = 15, color="black", fill="lightblue") +
  labs(title = "Distribution of Movie Runtime")

p_rating <- ggplot(raw_data, aes(x = movie_averageRating)) +
  geom_histogram(bins = 15, color="black", fill="lightblue") +
  labs(title = "Distribution of Movie Ratings") +
  theme_minimal()
```

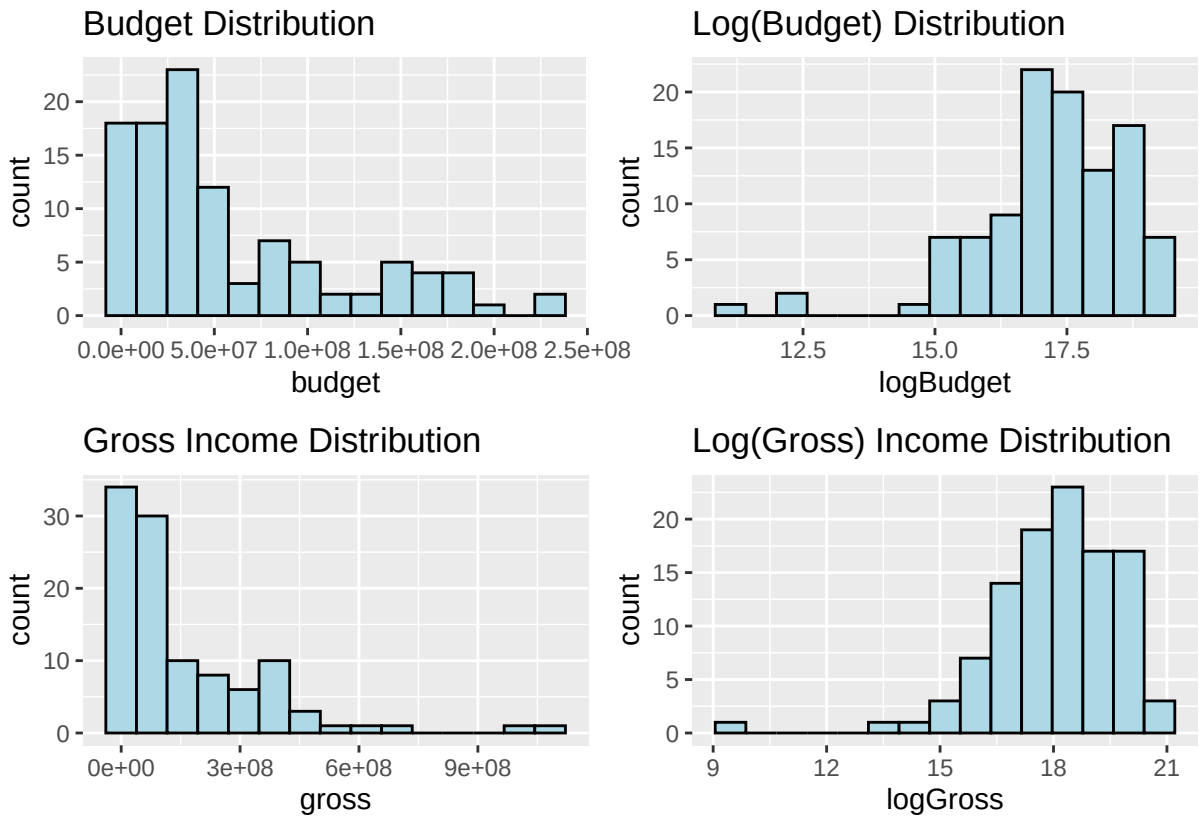
Appendix B2: Budget and gross (raw vs log)

```
# Appendix B2: Budget and gross (raw vs log)
p_gross <- ggplot(raw_data, aes(x = gross)) +
  geom_histogram(bins = 15, color="black", fill="lightblue") +
  labs(title = "Gross Income Distribution")

p_loggross <- ggplot(raw_data, aes(x = logGross)) +
  geom_histogram(bins = 15, color="black", fill="lightblue") +
  labs(title = "Log(Gross) Income Distribution")

p_budget <- ggplot(raw_data, aes(x = budget)) +
  geom_histogram(bins = 15, color="black", fill="lightblue") +
  labs(title = "Budget Distribution")

p_logbudget <- ggplot(raw_data, aes(x = logBudget)) +
  geom_histogram(bins = 15, color="black", fill="lightblue") +
  labs(title = "Log(Budget) Distribution")

(p_runtime | p_rating)
```



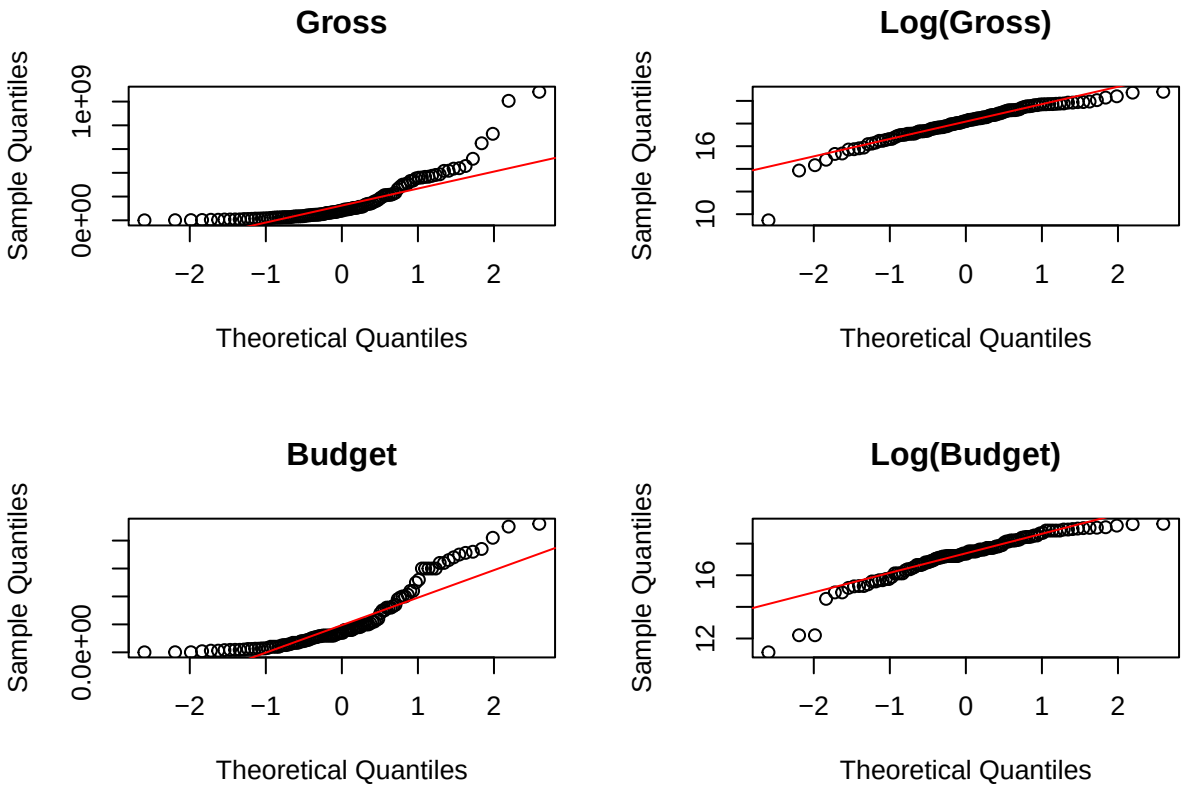



```
(p_budget | p_logbudget) / (p_gross | p_loggross)
```



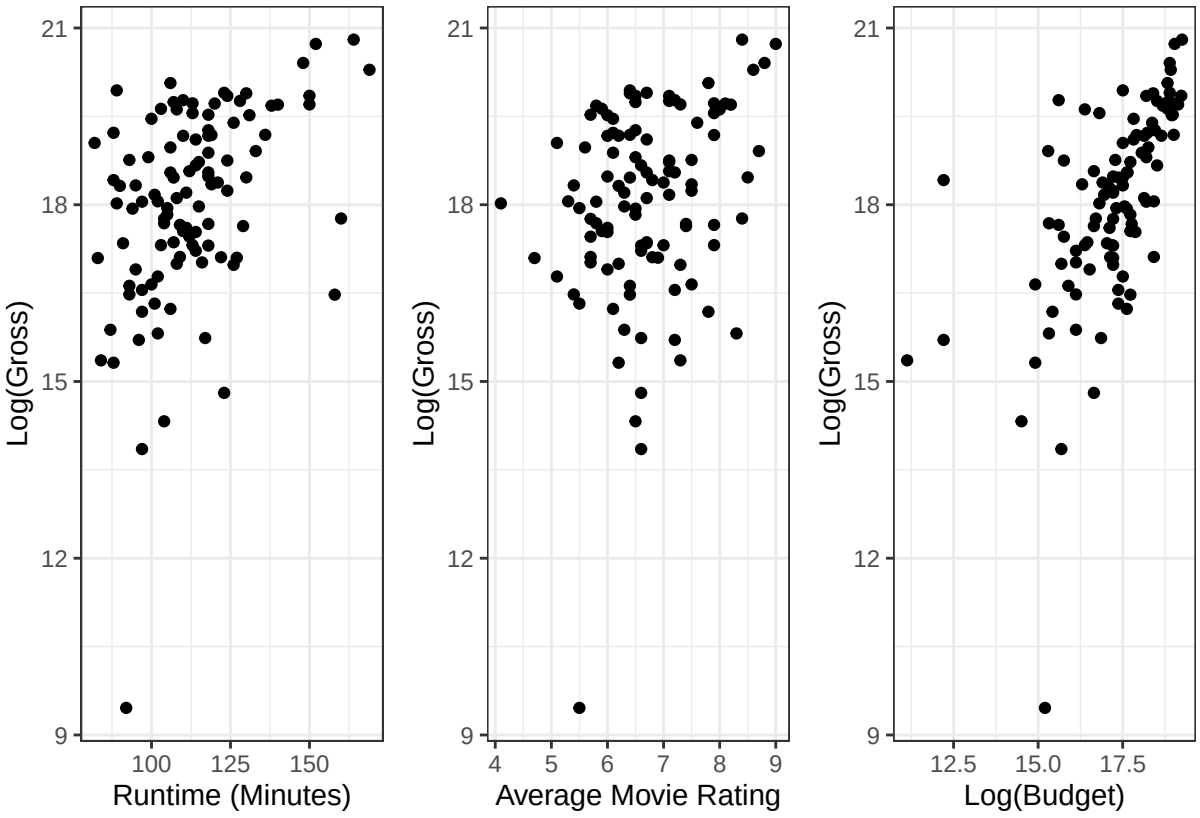
Appendix C: Transformation Justification

```
par(mfrow = c(2, 2))

qqnorm(raw_data$gross, main = "Gross")
qqline(raw_data$gross, col = "red")

qqnorm(raw_data$logGross, main = "Log(Gross)")
qqline(raw_data$logGross, col = "red")

qqnorm(raw_data$budget, main = "Budget")
qqline(raw_data$budget, col = "red")

qqnorm(raw_data$logBudget, main = "Log(Budget)")
qqline(raw_data$logBudget, col = "red")
```



Appendix D: Bivariate Analysis

```
p_runtime_gross <-
ggplot(raw_data, aes(x = runtime_minutes, y = logGross)) +
  geom_point() +
  theme_bw() +
  labs(x = "Runtime (Minutes)", y = "Log(Gross)")

p_rating_gross <-
ggplot(raw_data, aes(x = movie_averageRating, y = logGross)) +
  geom_point() +
  theme_bw() +
  labs(x = "Average Movie Rating", y = "Log(Gross)")

p_budget_gross <-
ggplot(raw_data, aes(x = logBudget, y = logGross)) +
  geom_point() +
  theme_bw() +
  labs(x = "Log(Budget)", y = "Log(Gross)")

(p_runtime_gross | p_rating_gross | p_budget_gross)
```

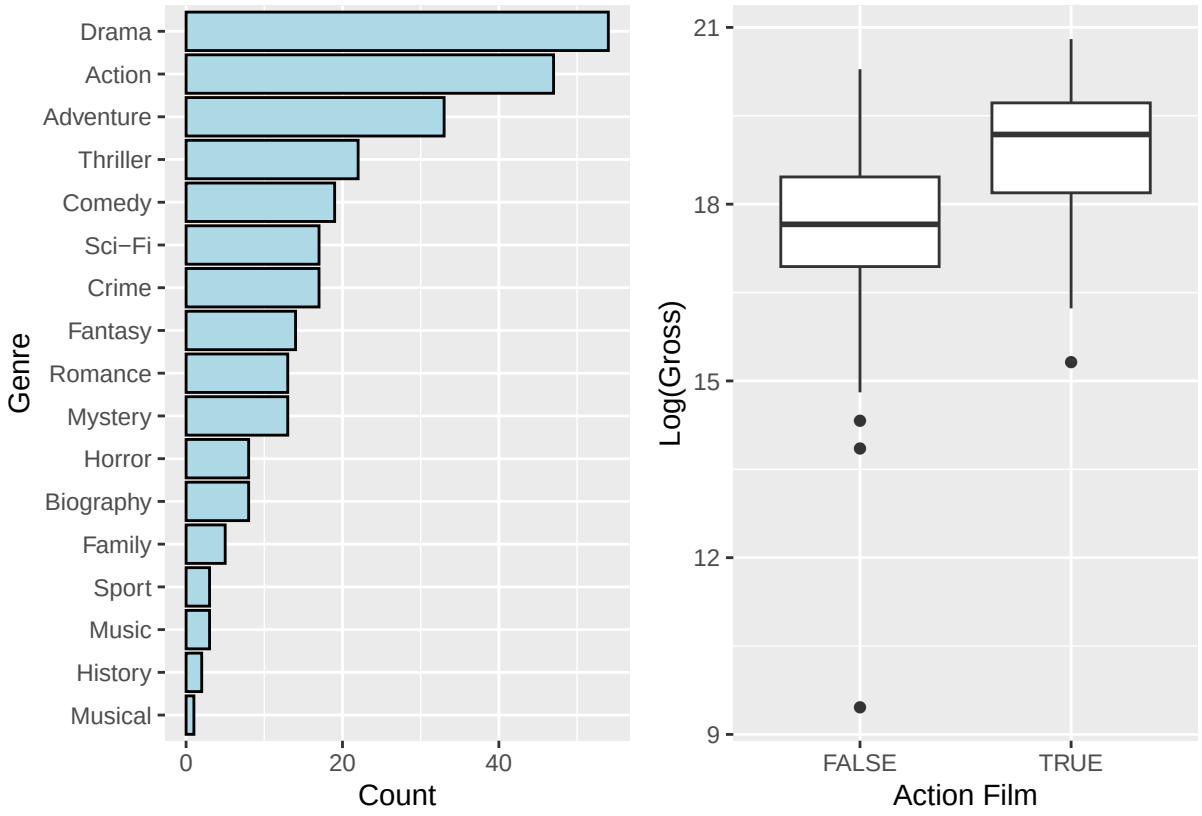


Appendix E: Categorical Variable Construction

Appendix E1: Genre Indicators

```
# Genre counts
all_genres_long <- raw_data %>%
  separate_rows(genres, sep = ",") %>%
  mutate(genres = str_trim(genres))

genre_counts <- all_genres_long %>%
  group_by(genres) %>%
  summarise(count = n()) %>%
  arrange(desc(count))

p_genre_counts <-
ggplot(genre_counts, aes(x = reorder(genres, count), y = count)) +
  geom_col(color="black", fill="lightblue") +
  coord_flip() +
  labs(x = "Genre", y = "Count")

# Action indicator
raw_data <- raw_data %>%
  mutate( is_action = str_detect(genres, "Action"))
```

```
p_action_box <-
ggplot(raw_data, aes(x = is_action, y = logGross)) +
  geom_boxplot() +
  labs(x = "Action Film", y = "Log(Gross)")

p_genre_counts | p_action_box
```



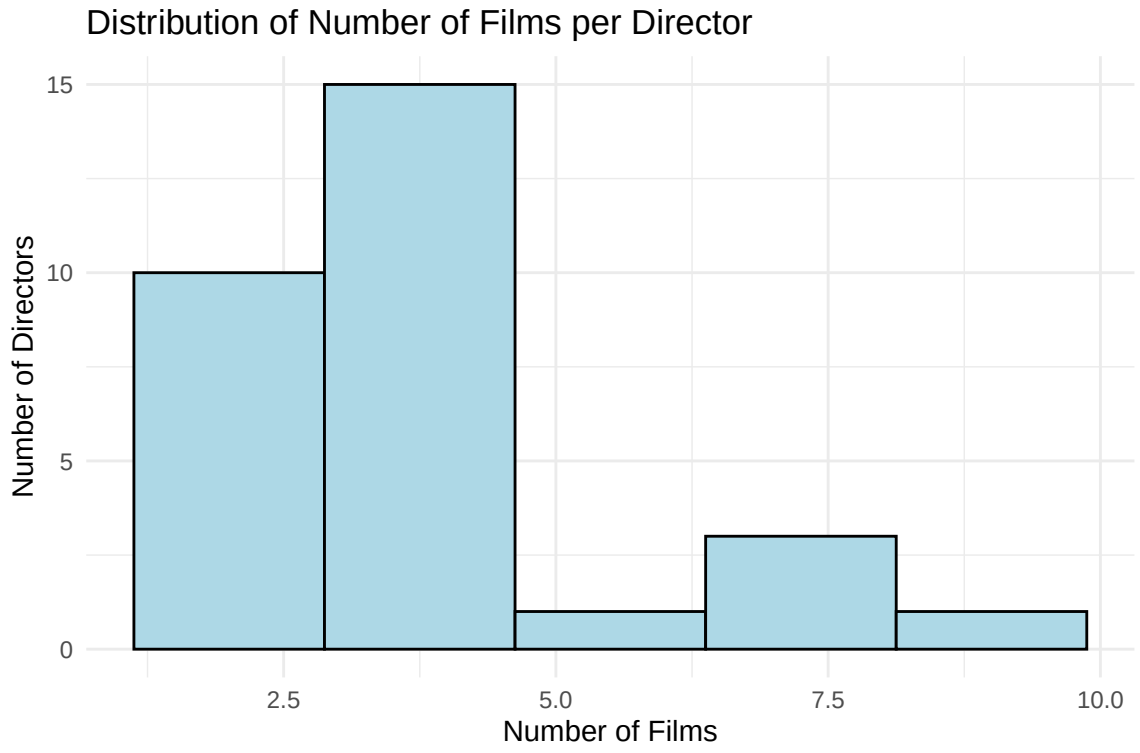
Appendix E2: Director Representation

```
director_count <- raw_data %>%
  count(director_name) %>% arrange(desc(n))

raw_data %>%
  group_by(director_name) %>%
  summarise(avg_gross = mean(logGross, na.rm = TRUE)) %>%
  arrange(desc(avg_gross))
```

```
## # A tibble: 30 x 2
##    director_name     avg_gross
##    <chr>                 <dbl>
##  1 Christopher Nolan      19.9
##  2 Yimou Zhang            19.0
##  3 Len Wiseman            19.0
```

```
##  4 Tarsem Singh           18.9
##  5 George Miller          18.7
##  6 Andrew Adamson         18.6
##  7 Jason Moore            18.5
##  8 Stephen Sommers        18.5
##  9 Gavin Hood             18.5
## 10 Stephen Chbosky        18.4
## # i 20 more rows
```

```r
raw_data <- raw_data %>%
  mutate(is_top_director = director_name == "Christopher Nolan")

ggplot(director_count, aes(x = n)) +
  geom_histogram(bins = 5, color = "black", fill = "lightblue") +
  labs( title = "Distribution of Number of Films per Director",
        x = "Number of Films",
        y = "Number of Directors" ) +
  theme_minimal()
```



Appendix F: Model Selection and Comparison

```r
# Base and alternative models
m_base <- lm(logGross ~ logBudget + movie_averageRating, data = raw_data)
m_runtime <- lm(logGross ~ logBudget + runtime_minutes + movie_averageRating, data = raw_data)
m_action_genre <- lm(logGross ~ logBudget + movie_averageRating + is_action, data = raw_data)
m_director <- lm(logGross ~ logBudget + movie_averageRating + is_top_director, data = raw_data)
```

```r
# ANOVA Comparisons
anova(m_base, m_runtime)
```

```
## Analysis of Variance Table
##
## Model 1: logGross ~ logBudget + movie_averageRating
## Model 2: logGross ~ logBudget + runtime_minutes + movie_averageRating
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    103 155.06
## 2    102 155.04  1  0.018975 0.0125 0.9113
```

```r
anova(m_base, m_action_genre)
```

```
## Analysis of Variance Table
##
## Model 1: logGross ~ logBudget + movie_averageRating
## Model 2: logGross ~ logBudget + movie_averageRating + is_action
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    103 155.06
## 2    102 148.21  1    6.8491 4.7136 0.03224 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(m_base, m_director)
```

```
## Analysis of Variance Table
##
## Model 1: logGross ~ logBudget + movie_averageRating
## Model 2: logGross ~ logBudget + movie_averageRating + is_top_director
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    103 155.06
## 2    102 155.03  1  0.033342 0.0219 0.8825
```

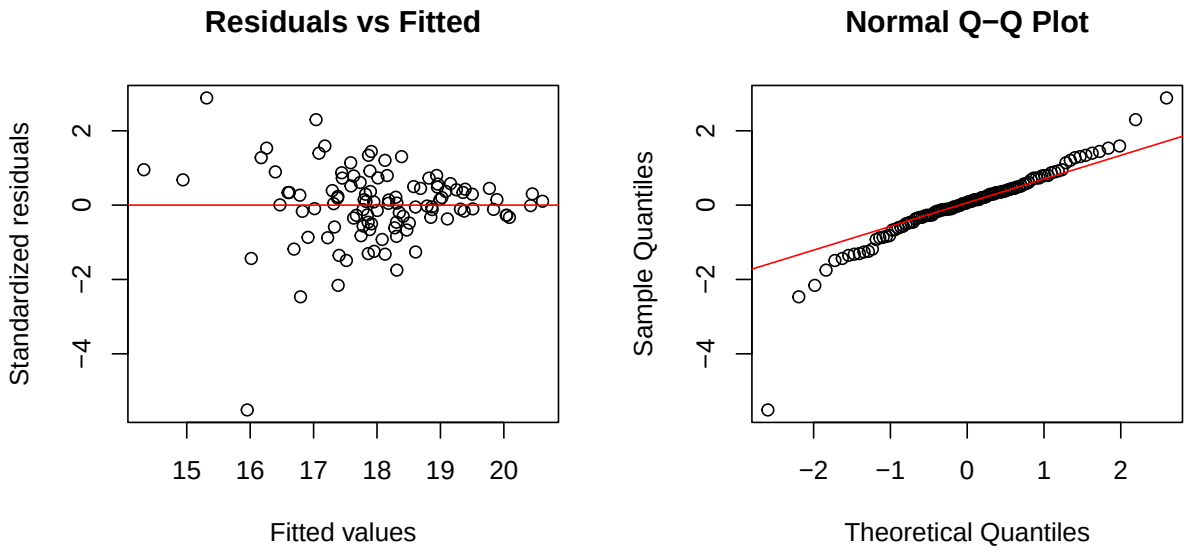
```r
# Information Criteria
AIC(m_base, m_action_genre, m_runtime, m_director)
```

```
##                df      AIC
## m_base          4 349.1335
## m_action_genre  5 346.3448
## m_runtime       5 351.1205
## m_director      5 351.1107
```

Appendix G: Model Diagnostics

```r
# Final model
m_final <- lm(logGross ~ logBudget + movie_averageRating + is_action, data = raw_data)

# Residual diagnostics
```

```r
par(mfrow=c(1,2))

r_standardized <- rstandard(m_final)
plot(fitted(m_final), r_standardized,
     main = "Residuals vs Fitted",
     xlab = "Fitted values", ylab = "Standardized residuals")
abline(h = 0, col = "red")

qqnorm(rstandard(m_final))
qqline(rstandard(m_final), col = "red")
```




```r
summary(m_final)
```

```
##
## Call:
## lm(formula = logGross ~ logBudget + movie_averageRating + is_action,
##     data = raw_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4936 -0.4346  0.0828  0.5870  3.1040
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.87005    1.73412   2.232 0.027822 *
## logBudget            0.61710    0.08961   6.886  4.8e-10 ***
## movie_averageRating  0.49139    0.12423   3.955 0.000141 ***
## is_actionTRUE        0.57035    0.26270   2.171 0.032243 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.205 on 102 degrees of freedom
```

```
## Multiple R-squared:  0.4848, Adjusted R-squared:  0.4696
## F-statistic: 31.99 on 3 and 102 DF,  p-value: 1.168e-14
```

Appendix H: Prediction and Uncertainty

```r
# New movie specification
new_movie <- data.frame(
  logBudget = log(10000000),
  movie_averageRating = 7.0,
  is_action = FALSE
)

# Prediction interval
pred <- predict(m_final, newdata = new_movie, interval = "prediction", level = 0.95)

# Back-transform from log to dollar scale
exp(pred)
```

```
##        fit     lwr       upr
## 1 31206866 2791674 348847459
```