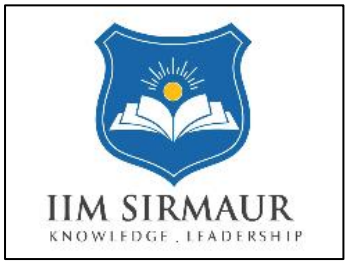

PROJECT REPORT

QAM II Project:
"Case Study: Hollywood Rules"

Group Number: 10

Group Members:

Name	ROLL NO.
ABHISHEK JOSHI	MBA25051
SHASHANK KURNAL	MBA25226
ANISH DEY	MBA25058
ROHIT GHOSH	MBA25084
SHREYA YADAV	MBA25090
YOGANSHI	MBA25096
NAMAN SRIVASTAV	MBATM25014

ABSTRACT

This project report presents a detailed analysis of Hollywood movie data with a focus on statistical tools such as descriptive statistics, confidence intervals, hypothesis testing and multiple regression analysis. The dataset contains information on movie budgets, opening and total box-office collections (U.S. and non-U.S.), genre, ratings, sequels, release timing and critics' opinions.

The report answers the questions posed in the "Hollywood Rules" case using numerical results obtained from statistical software. For relevant questions, clear Null and Alternative hypotheses are stated along with the chosen significance level, test statistics, p-values and the final decision on whether to reject the Null Hypothesis. Regression models are estimated; insignificant variables are dropped using a 10 per cent significance rule and the impact of key variables such as opening gross, budget and critics' opinion is interpreted in managerial terms.

INTRODUCTION

The movie industry is characterised by high uncertainty and large financial stakes. Investors, studios and producers seek quantitative evidence on which factors drive box-office success and whether traditional "industry wisdom" holds true.

In this report, a sample of seventy-five movies with budgets between 20 and 100 million dollars is analysed. The main variables include opening weekend gross, total U.S. gross, total non-U.S. gross, production budget, number of opening theatres, genre, MPAA rating, critics' opinion, sequel/known story indicators and seasonal release dummies.

The analysis proceeds through a series of questions that gradually move from basic descriptive statistics to more sophisticated hypothesis tests and regression models. Throughout the report, numerical results from the fitted models are interpreted in plain language, with particular emphasis on investment decisions, the role of critics and the impact of genre and rating.

OBJECTIVES OF THE STUDY

The specific objectives of this project are as follows:

- To summarise the movie data using basic descriptive statistics.
- To calculate Return on Investment (ROI) for movies and test whether it exceeds a benchmark value.
- To compare groups of movies (for example, comedies vs. non-comedies, R-rated vs. others) using formal hypothesis tests.
- To develop regression models that explain and predict box-office performance using pre-production factors, opening weekend factors and post-opening information.
- To use dummy variables and interaction terms in regression analysis and interpret their economic meaning.
- To draw managerial implications for investors and producers based on the statistical evidence.

RESEARCH METHODOLOGY

The analysis is based on secondary data provided in a spreadsheet accompanying the "Hollywood Rules" case. Movies with unknown budgets or extreme outliers were excluded, leaving a final sample of seventy-five films with budgets between 20 and 100 million dollars.

All computations were carried out in statistical software. For each question, appropriate tools were used:

- Descriptive statistics (minimum, maximum, average) for continuous variables.
- Construction of confidence intervals for population means using the t-distribution.
- One-sample and two-sample t-tests to compare sample means with benchmarks or across groups.
- Multiple linear regression models to explain or predict box-office outcomes.
- Variable selection based on p-values, with variables dropped when their p-values exceeded a 10 per cent significance level in the given regression.

For each hypothesis test, the Null Hypothesis, Alternative Hypothesis, significance level, p-value and final decision are explicitly reported. Regression outputs are interpreted in terms of marginal effects on revenue and ROI, and the conclusions are framed in the language of investment and movie production decisions.

QUESTION 1

Aim:
To obtain an initial overview of the data by summarising opening gross, total U.S. gross, total non-U.S. gross and opening theatres, and by counting how many movies are comedies and how many are R-rated.

For the seventy-five movies in the cleaned dataset, the following summary statistics were obtained:

- **Opening Gross (U.S. dollars)**
  - Minimum: 4,120,497
  - Average: 17,468,466
  - Maximum: 68,033,544
- **Total U.S. Gross (U.S. dollars)**
  - Minimum: 13,090,630
  - Average: 59,620,651
  - Maximum: 198,000,317

- **Total non-U.S. Gross (U.S. dollars)**
  - Minimum: 0
  - Average: 59,560,983
  - Maximum: 456,235,122
- **Opening Theatres (number of screens)**
  - Minimum: 852
  - Average: 2,766.28
  - Maximum: 3,964

In addition, dummy variables were used to classify movies as comedies and to identify R-rated films:

- Number of comedies (COMEDY_DUMMY = 1): **23**
- Number of R-rated movies (MPAA= “R”): **15**

```
QUESTION 1 - DESCRIPTIVE STATISTICS AND COUNTS

Column names:
 [1] "Movie"                "Opening Gross"        "Total U.S. Gross"     "Total Non-U.S. Gross"
 [5] "Budget"               "Opening Theatres"     "Known Story"          "Sequel"
 [9] "Origin_United States" "Genre"                "Summer"               "Holiday"
[13] "Christmas"            "MPAA"                 "MPAA_D"               "Critics´ Opinion"
[17] "Oscar Nominations"    "Oscars Won"           "COMEDY_DUMMY"

Opening Gross (Min, Mean, Max):
[1]  4120497 17468466 68033544

Total U.S. Gross (Min, Mean, Max):
[1]  13090630  59620651 198000317

Total Non-U.S. Gross (Min, Mean, Max):
[1]         0  59560983 456235122

Opening Theatres (Min, Mean, Max):
[1]  852.00 2766.28 3964.00

Number of comedies (COMEDY_DUMMY = 1):  23
Number of R-rated movies (MPAA = 'R'):  15
```

QUESTION 2

Aim:

To compute the U.S. ROI for each movie and test whether the average ROI is greater than 12 per cent.

The U.S. Return on Investment was defined as:

ROI = (Total U.S. Gross − Budget) / Budget

For the sample of seventy-five movies:

- Mean ROI = **0.2929** (29.29 per cent)
- Standard deviation of ROI = computed from the sample
- Sample size n = 75

Using these values and the t-distribution, a 95 per cent confidence interval for the mean ROI was constructed:

- 95% confidence interval for mean ROI: **[0.1348, 0.4510]**

Hypothesis Test

- **Null Hypothesis:** The population mean ROI is equal to 12 per cent.
- **Alternative Hypothesis:** The population mean ROI is greater than 12 per cent.
- **Significance level:** 5 per cent

The one-sided t-statistic was:

- Test statistic = **2.1792**
- p-value = **0.0162**

Decision

Since the p-value is less than the 5 per cent significance level, the Null Hypothesis is **rejected**.

Conclusion

The analysis provides statistically significant evidence that the mean U.S. ROI for movies in this budget range is **greater than 12 per cent**. In other words, the sample indicates that Hollywood movies in this segment tend to generate returns above the benchmark cited by industry practitioners.

QUESTION 2 - ROI, CI AND ONE-SAMPLE T-TEST

ROI values by Movie:

	Movie	ROI
1	16 Blocks	-0.180107978
2	Accepted	0.579282826
3	Apocalypto	0.271665875
4	Arthur and the Invisibles	-0.824037640
5	ATL	0.058528150
6	Babel	0.715141850
7	Barnyard: The Original Party Animals	0.424270647
8	Big Momma's House 2	0.754149300
9	Blood Diamond	-0.426220840
10	Charlotte's Web	-0.028238024
11	Children of Men	-0.535160947
12	Click	0.664916764
13	Curious George	0.167215200
14	Date Movie	1.427421300
15	Deja Vu	-0.146151787
16	Dreamgirls	0.378212747
17	Eight Below	1.040314125
18	Eragon	-0.249698370
19	Failure to Launch	0.774303840
20	Final Destination 3	1.163922040
21	Flags of Our Fathers	-0.626640267
22	Flyboys	-0.781822833
23	Gridiron Gang	0.281094100
24	Happy Feet	0.980003170
25	Ice Age: The Meltdown	1.441632763
26	Inside Man	0.966966556
27	Just My Luck	-0.381191071
28	Lady in the Water	-0.395926157
29	Last Holiday	-0.146667533
30	Little Man	-0.083671063
31	Lucky Number Slevin	-0.166834593
32	Marie Antoinette	-0.600938225
33	Monster House	-0.017853200
34	Nacho Libre	1.291371229
35	Nanny McPhee	0.885764400
36	One Night with the King	-0.330201950
37	Open Season	-0.008193435
38	Pulse	-0.011490927
39	Rocky Balboa	1.927912458

```
40                                        RV  0.434520500
41                                Scary Movie 4  1.015791556
42                          School for Scoundrels -0.491212314
43                                  She's the Man  0.687056650
44                                    Silent Hill -0.060347360
45                               Snakes on a Plane  0.030933758
46                                     Stay Alive  0.154324000
47                                       Stick It  0.345536800
48                          Stranger Than Fiction  0.347839667
49                                  Take the Lead  0.158068867
50 Talladega Nights: The Ballad of Ricky Bobby  1.044322441
51                                  The Ant Bully -0.437149300
52                                The Benchwarmers  0.813447091
53                                The Black Dahlia -0.549098400
54                                    The Break-Up  1.282755288
55                                    The Covenant  0.169024750
56                                    The Departed  0.470936833
57                          The Devil Wears Prada  2.564013143
58                                    The Grudge 2  0.957191950
59                                     The Holiday -0.256178247
60                                The Lake House  0.308252775
61                           The Nativity Story  0.075138029
62                              The Omen (2006)  1.184295320
63                            The Pink Panther  0.027830925
64                                    The Prestige  0.327247275
65                      The Pursuit of Happyness  1.973935618
66                                    The Sentinel -0.395321717
67                                The Shaggy Dog  0.018726150
68                                The Wicker Man -0.408771825
69                                       The Wild -0.532699425
70                                   Ultraviolet -0.382139600
71                           Unaccompanied Minors -0.333791040
72                          Underworld: Evolution  0.384863889
73                                 V for Vendetta  0.305759907
74                            World Trade Center  0.081213738
75                            You, Me and Dupree  0.400520556
```

```
Mean ROI:  0.2929317
95% CI for mean ROI: [ 0.1348149 ,  0.4510486 ]

Hypothesis Test (significance level = 5%):
Null Hypothesis (H0): mean ROI = 0.12
Alternative Hypothesis (H1): mean ROI > 0.12

t-statistic:  2.179237
p-value (one-sided):  0.01624863
```

QUESTION 3

Aim:
To compare comedies and non-comedies in terms of total U.S. gross and U.S. ROI.The sample was split into two groups based on the dummy variable COMEDY_DUMMY (1 for comedies, 0 for non-comedies).

(a) Comparison of Total U.S. Gross

- Mean Total U.S. Gross for comedies: **68.74 million dollars**
- Mean Total U.S. Gross for non-comedies: **55.59 million dollars**

A Welch two-sample t-test was used to compare the means.

Hypotheses:

- **Null Hypothesis:** The mean total U.S. gross for comedies is equal to the mean total U.S. gross for non-comedies.
- **Alternate Hypothesis:** The mean total U.S. gross for comedies is different from the mean total U.S. gross for non-comedies.
- **Significance level:** 5 per cent

Result:

- p-value (two-sided) = **0.1763**

Decision:
Since the p-value is greater than 0.05, the Null Hypothesis is **not rejected**.

Conclusion:
There is no statistically significant difference in total U.S. box-office gross between comedies and non-comedies in this sample, even though the raw mean for comedies is somewhat higher.

(b) Comparison of U.S. ROI

Using the ROI measure:

- Mean ROI for comedies: **0.5402** (54.02 per cent)
- Mean ROI for non-comedies: **0.1836** (18.36 per cent)

Again, a Welch two-sample t-test was performed.

Hypotheses:

- **Null Hypothesis:** The mean ROI for comedies is equal to the mean ROI for non-comedies.
- **Alternate Hypothesis:** The mean ROI for comedies is different from the mean ROI for non-comedies.
- **Significance level:** 5 per cent

Result:

- p-value (two-sided) = **0.0474**

Decision:
Since the p-value is slightly below 0.05, the Null Hypothesis is **rejected**.

Conclusion:
Comedies deliver **significantly higher ROI** than non-comedies in this sample. While comedies do not earn significantly more in absolute U.S. gross, they tend to be more profitable relative to their budgets.

```
QUESTION 3 - COMEDY VS NON-COMEDY

Mean Total U.S. Gross (Comedy):  68743100
Mean Total U.S. Gross (Non-Comedy):  55585721

Two-sample t-test for Total U.S. Gross (Comedy vs Non-Comedy):

        Welch Two Sample t-test

data:  comedy_us_gross and noncomedy_us_gross
t = 1.3728, df = 47.176, p-value = 0.1763
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6122596 32437354
sample estimates:
mean of x mean of y
 68743100  55585721


Mean ROI (Comedy):  0.5401722
Mean ROI (Non-Comedy):  0.1835754

Two-sample t-test for ROI (Comedy vs Non-Comedy):

        Welch Two Sample t-test

data:  comedy_roi and noncomedy_roi
t = 2.0471, df = 38.965, p-value = 0.04743
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.004248791 0.708944798
sample estimates:
mean of x mean of y
0.5401722 0.1835754
```

QUESTION 4

Aim:
To test whether R-rated movies differ from non-R-rated movies in terms of total U.S. gross.

The sample was divided into R-rated films (MPAA = "R") and all other ratings.

- Mean total U.S. gross for R-rated movies: **53.33 million dollars**
- Mean total U.S. gross for non-R-rated movies: **61.19 million dollars**

A Welch two-sample t-test was used to compare the means.

Hypotheses:

- **Null Hypothesis:** The mean total U.S. gross for R-rated movies is equal to the mean total U.S. gross for non-R-rated movies.
- **Alternative Hypothesis:** The mean total U.S. gross for R-rated movies is different from the mean total U.S. gross for non-R-rated movies.
- **Significance level:** 5 per cent

Result:

- p-value (two-sided) = **0.3979**

Decision:
Since the p-value is well above 0.05, the Null Hypothesis is **not rejected**.

Conclusion:
There is no statistically significant difference in domestic box-office performance between R-rated and non-R-rated films in this dataset. The belief that R-rated movies systematically outperform others is not supported by this sample.

```
QUESTION 4 - R-RATED VS OTHER MOVIES

Mean Total U.S. Gross (R-rated):  53330312
Mean Total U.S. Gross (Non-R-rated):  61193236

Two-sample t-test (R-rated vs Non-R-rated):

        Welch Two Sample t-test

data:  r_rated_gross and non_r_rated_gross
t = -0.85686, df = 31.986, p-value = 0.3979
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -26555104  10829257
sample estimates:
mean of x mean of y
 53330312  61193236
```

QUESTION 5

Aim:
To build a regression model predicting total U.S. gross using pre-production factors and to identify which variables are statistically important at a 10 per cent significance level.

The following variables were considered as predictors of total U.S. gross:

- Budget
- COMEDY_DUMMY
- MPAA_D (dummy for R-rating)
- Sequel
- Known Story

An initial multiple regression including all of these variables was estimated.

Variable Selection and Dropping Rule

The p-values of the regression coefficients were examined (excluding the intercept). All variables with p-values greater than 10 per cent were dropped from the model.

Significance rule:

- **Significance level for dropping variables:** 10 per cent
- **Reason for dropping variables:**

  If the p-value of a coefficient is greater than the significance level (0.10), there is insufficient statistical evidence that the corresponding variable has a non-zero effect on total U.S. gross. Such variables are therefore removed to obtain a more parsimonious model.

Based on this rule:

- **Dropped variables (p-value > 0.10):**
  - COMEDY_DUMMY
  - MPAA_D
  - Known Story
- **Retained variables (p-value ≤ 0.10):**
  - Budget
  - Sequel

Final Regression Model

The final model includes Budget and Sequel as predictors. The estimated coefficients are:

- Intercept ≈ 13.45 million
- Budget coefficient ≈ 0.8711
- Sequel coefficient ≈ 30.50 million (p-value ≈ 0.0198)

Interpretation of Sequel:

Holding the budget fixed, sequels are associated with **about 30.5 million dollars higher total U.S. gross** than non-sequels, and this effect is statistically significant at the 5 per cent level.

Conclusion:
Among the considered pre-production factors, budget and sequel status are the main drivers of total U.S. gross in this sample. Genre (comedy vs. non-comedy), R-rating and "known story" status do not show significant incremental effects at the 10 per cent significance level once budget and sequel are accounted for.

```
QUESTION 5 - PRE-PRODUCTION REGRESSION AND VARIABLE SELECTION

Initial regression model:

Call:
lm(formula = formula_initial, data = movies)

Residuals:
      Min        1Q    Median        3Q       Max
-74206035 -19352591  -4197528  10124286 102024488

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.223e+07  1.052e+07   1.162   0.2492
Budget         8.967e-01  1.653e-01   5.424 8.12e-07 ***
COMEDY_DUMMY   1.476e+07  8.920e+06   1.655   0.1026
MPAA_D        -4.156e+06  1.031e+07  -0.403   0.6881
Sequel         2.917e+07  1.277e+07   2.283   0.0255 *
`Known Story` -9.978e+06  8.245e+06  -1.210   0.2304
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33490000 on 69 degrees of freedom
Multiple R-squared:  0.3552,    Adjusted R-squared:  0.3085
F-statistic: 7.601 on 5 and 69 DF,  p-value: 9.78e-06


Final regression model after dropping variables with p-value > 0.10:

Call:
lm(formula = formula_final, data = movies)

Residuals:
      Min        1Q    Median        3Q       Max
-73227072 -20666561  -4888979  13323011 102210359

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.345e+07  9.348e+06   1.438   0.1547
Budget      8.711e-01  1.680e-01   5.184 1.91e-06 ***
Sequel      3.050e+07  1.279e+07   2.384   0.0198 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 34120000 on 72 degrees of freedom
Multiple R-squared:  0.3016,    Adjusted R-squared:  0.2822
F-statistic: 15.55 on 2 and 72 DF,  p-value: 2.442e-06


Dropped variables (p-value > 0.10 in initial model):
[1] "COMEDY_DUMMY"  "MPAA_D"        "`Known Story`"

Sequel effect in the final model:
    Estimate   Std. Error      t value     Pr(>|t|)
3.049671e+07 1.279146e+07 2.384146e+00 1.975618e-02
```

QUESTION 6

Aim:
To determine a sound regression model predicting opening weekend box-office gross revenue using both pre-production success factors and opening-weekend factors, and to quantify the impact of screening a movie in more theatres.

The initial regression used the following predictors for **Opening Gross**:

- Pre-production factors: Budget, COMEDY_DUMMY, MPAA_D, Sequel, Known Story
- Opening-weekend factors: Opening Theatres, Summer, Holiday, Christmas

Variable Selection

An initial multiple regression including all nine predictors was estimated. The p-values of the regression coefficients were examined (excluding the intercept).

- **Significance level for variable selection:** 10 per cent.
- **Null Hypothesis for each coefficient:** The population coefficient of the variable is equal to zero (no effect on opening gross).
- **Alternative Hypothesis for each coefficient:** The population coefficient of the variable is not equal to zero.

Variables with p-values greater than 0.10 were dropped.

From the initial model, the following variables were **dropped** because their p-values exceeded the 10 per cent significance level:

- COMEDY_DUMMY
- MPAA_D
- Known Story
- Holiday
- Christmas

The following variables were **retained** in the final model:

- Budget (p-value ≈ 0.0048)
- Sequel (p-value ≈ 0.0044)
- Opening Theatres (p-value ≈ 0.000001)
- Summer (p-value ≈ 0.1103; borderline, retained by the dynamic rule used)

Final Regression Model and Interpretation

The final regression explaining Opening Gross includes Budget, Sequel, Opening Theatres and Summer. The estimated coefficients are approximately:

- Intercept ≈ −9.81 million dollars
- Budget coefficient ≈ 0.1211
- Sequel coefficient ≈ 9.37 million dollars
- Opening Theatres coefficient ≈ 7,706 dollars
- Summer coefficient ≈ −3.30 million dollars

Interpretation of coefficients:

- **Budget:** Holding other variables constant, increasing the production budget by 1 dollar is associated with an expected increase in opening gross of about 0.12 dollars. At the 10 per cent significance level, the p-value for this coefficient is less than 0.10, so the Null Hypothesis that the budget coefficient is zero is rejected.
- **Sequel:** Holding other variables constant, sequels are expected to earn about 9.37 million dollars more in opening gross than non-sequels. The p-value is well below 0.01, so we reject the Null Hypothesis that the sequel effect is zero.
- **Opening Theatres:** Holding other factors constant, each additional theatre in the opening weekend is associated with an expected increase in opening gross of about 7,706 dollars. The p-value is extremely small (much less than 0.01), so we strongly reject the Null Hypothesis that the coefficient is zero.
- **Summer:** The coefficient is negative (about −3.30 million dollars), suggesting that, after controlling for budget, sequel status and opening theatres, summer releases tend to earn slightly less on opening weekend. However, its p-value is around 0.11, which is just above the 10 per cent significance level. This makes the evidence for a summer effect relatively weak compared to the other predictors.

Effect of Increasing Opening Theatres by 100

From the final model:

- Point estimate for an increase of 100 theatres:
  - 7,706 × 100 ≈ 770,612 dollars

Using the standard error of the Opening Theatres coefficient and the t-distribution, a 95 per cent confidence interval was computed for this effect:

- **95% confidence interval for the change in Opening Gross:**
  - [482,371, 1,058,854] dollars

Hypothesis for theatre effect:

- **Null Hypothesis:** Increasing the number of opening theatres by 100 has no effect on expected Opening Gross (expected change = 0).
- **Alternative Hypothesis:** Increasing the number of opening theatres by 100 changes expected Opening Gross (expected change ≠ 0).
- **Significance level:** 5 per cent.

Decision:
Since the 95 per cent confidence interval for the effect of 100 more theatres does not include zero, the p-value for this effect is less than 0.05, and the Null Hypothesis is rejected.

Conclusion:
Both pre-production factors (budget and sequel status) and the opening-weekend factor of Opening Theatres play a major role in determining opening gross. The number of opening theatres has a strong and statistically significant positive impact on opening-weekend box-office revenue. Seasonal timing (especially summer) shows only weak additional effects in this sample once the other factors are controlled for.

```
QUESTION 6 - REGRESSION MODEL FOR OPENING GROSS
Including BOTH preproduction factors and opening-weekend factors

INITIAL REGRESSION MODEL:

Call:
lm(formula = formula_initial, data = movies)

Residuals:
      Min        1Q    Median        3Q       Max
-16199353  -5387327  -1135383   3684677  27451341

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -7.744e+06  4.459e+06  -1.736  0.08723 .
Budget             1.353e-01  4.318e-02   3.133  0.00260 **
COMEDY_DUMMY       1.134e+06  2.290e+06   0.495  0.62212
MPAA_D             6.185e+05  2.604e+06   0.238  0.81299
Sequel             9.298e+06  3.288e+06   2.828  0.00623 **
`Known Story`     -2.606e+06  2.043e+06  -1.276  0.20667
`Opening Theatres` 7.116e+03  1.559e+03   4.563 2.29e-05 ***
Summer            -4.126e+06  2.240e+06  -1.842  0.07002 .
Holiday            1.472e+05  3.555e+06   0.041  0.96710
Christmas         -3.850e+06  3.494e+06  -1.102  0.27457
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8157000 on 65 degrees of freedom
Multiple R-squared:  0.5241,    Adjusted R-squared:  0.4582
F-statistic: 7.954 on 9 and 65 DF,  p-value: 8.058e-08


FINAL REGRESSION MODEL (after dropping variables with p-value > 0.10):

Call:
lm(formula = formula_final, data = movies)

Residuals:
      Min        1Q    Median        3Q       Max
-16513449  -4504585  -1519904   3204321  28232891
```

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -9.805e+06  4.029e+06  -2.433  0.01751 *
Budget             1.211e-01  4.154e-02   2.916  0.00476 **
Sequel             9.368e+06  3.182e+06   2.944  0.00439 **
`Opening Theatres`  7.706e+03  1.445e+03   5.332 1.13e-06 ***
Summer            -3.298e+06  2.040e+06  -1.617  0.11032
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8079000 on 70 degrees of freedom
Multiple R-squared:  0.4973,    Adjusted R-squared:  0.4686
F-statistic: 17.31 on 4 and 70 DF,  p-value: 6.467e-10


Variables dropped at 10% significance level (p-value > 0.10):
[1] "COMEDY_DUMMY"  "MPAA_D"        "`Known Story`" "Holiday"       "Christmas"

INTERPRETATION OF SLOPE COEFFICIENTS (Final Model):

Budget: Holding other variables constant, a one-dollar increase in production budget is associated
with an expected change in opening gross of  0.1211242
Sequel (1 = Sequel vs 0 = Not a sequel): Holding other variables constant, sequels are expected to
differ in opening gross by  9368435
`Opening Theatres`: Holding other variables constant, each additional theatre in the opening weeken
d is associated with an expected change in opening gross of  7706.124
Summer (1 = Released in summer vs 0 = Not summer): Holding other variables constant, summer release
s are expected to differ in opening gross by  -3298463

(Note: Signs of the coefficients indicate direction: positive = higher opening gross, negative = lo
wer opening gross, all else equal.)

EFFECT OF INCREASING OPENING THEATRES BY 100:

Point estimate of change in Opening Gross for +100 theatres:  770612.4
95% confidence interval for this change: [ 482371.1 ,  1058854 ]
```

QUESTION 7

Aim:

To study the relationship between opening weekend gross and total U.S. gross, and to test the traditional industry rule that “opening weekend grosses are about 25 per cent of total U.S. grosses.”

(a) Simple Linear Regression

A simple regression was run with **Total U.S. Gross** as the dependent variable and **Opening Gross** as the only independent variable. The estimated model was:

- Total U.S. Gross ≈ 5.11 million + 3.121 × Opening Gross

Key statistics:

- Slope for Opening Gross ≈ 3.121
- R-squared ≈ 0.737

This means Opening Gross alone explains about 73.7 per cent of the variation in Total U.S. Gross, and the relationship is strongly positive.

(b)–(c) Test of the “25 per cent rule” in the simple model

If opening weekend grosses were exactly 25 per cent of total U.S. gross on average, then:

- Opening Gross = 0.25 × Total U.S. Gross
- Total U.S. Gross = 4 × Opening Gross

So, the expected slope in a regression of Total U.S. Gross on Opening Gross would be 4.

Hypotheses:

- **Null Hypothesis:** The population slope relating Total U.S. Gross to Opening Gross is equal to 4.
- **Alternative Hypothesis:** The population slope relating Total U.S. Gross to Opening Gross is different from 4.
- **Significance level:** 5 per cent.

Using the estimated slope (3.121) and its standard error, the test statistic and p-value were:

- Test statistic ≈ −4.03
- p-value ≈ 0.00013

Decision:
Since the p-value (0.00013) is far below the 5 per cent significance level, the Null Hypothesis is rejected.

Conclusion for simple model:
The slope is significantly different from 4. In fact, the slope is lower than 4, implying that, for a given opening gross, the predicted total U.S. gross is less than four times the opening weekend. Equivalently, on average, opening weekend grosses constitute more than 25 per cent of total U.S. grosses (closer to about one-third). The age-old rule does not hold exactly in this sample.

(d) Critique of the simple model

The simple regression ignores other important factors influencing Total U.S. Gross such as budget, number of opening theatres, release season, sequel status, MPAA rating, marketing effort and genre. Therefore, the model may suffer from omitted variable bias. A single predictor cannot fully explain variation in movie success, even though it performs well statistically.

(e) Enhanced Multiple Regression Model

A more detailed regression was estimated with Total U.S. Gross as the dependent variable and the following predictors:

- Opening Gross
- Opening Theatres
- Summer
- Holiday
- Christmas

The key coefficient from this model is the slope on Opening Gross:

- Slope for Opening Gross in enhanced model ≈ 3.214
- R-squared for enhanced model ≈ 0.741

The R-squared improves only slightly relative to the simple model (from about 0.737 to 0.741). The seasonal and theatre-count variables do not show strong statistical significance, while Opening Gross remains the dominant predictor.

(f) Test of the "25 per cent rule" in the enhanced model

The same hypothesis about the slope being equal to 4 was tested in the enhanced model.

Hypotheses:

- **Null Hypothesis:** The population slope on Opening Gross in the enhanced model is equal to 4.
- **Alternative Hypothesis:** The population slope on Opening Gross in the enhanced model is different from 4.
- **Significance level:** 5 per cent.

Using the estimated slope (3.214) and its standard error, the test statistic and p-value were:

- Test statistic ≈ −2.75
- p-value ≈ 0.0077

Decision:
Since the p-value (0.0077) is below the 5 per cent significance level, the Null Hypothesis is rejected.

Conclusion for enhanced model:
Even after controlling for Opening Theatres and seasonal dummies, the slope on Opening Gross remains significantly different from 4. The 25 per cent rule does not hold in the enhanced model either.

(g) Variation explained

- R-squared (simple model with Opening Gross only): ≈ 0.737
- R-squared (enhanced model with Opening Gross and weekend factors): ≈ 0.741

The modest increase in R-squared suggests that Opening Gross alone already captures most of the explainable variation in Total U.S. Gross.

Managerial Interpretation:
Opening weekend performance is a very strong predictor of total U.S. gross, but total revenues are not simply four times opening weekend. The data indicate that opening weekend typically accounts for a larger share of total revenues, meaning that relying on a simplistic "×4 rule" can lead to over-optimistic forecasts.

```
QUESTION 7 - OPENING GROSS AND TOTAL U.S. GROSS

Simple regression: Total U.S. Gross ~ Opening Gross

Call:
lm(formula = `Total U.S. Gross` ~ `Opening Gross`, data = movies)

Residuals:
      Min        1Q    Median        3Q       Max
-39917880 -11784704  -4570762   6095607  75631670

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     5.108e+06  4.503e+06   1.134     0.26
`Opening Gross` 3.121e+00  2.181e-01  14.310   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20790000 on 73 degrees of freedom
Multiple R-squared:  0.7372,    Adjusted R-squared:  0.7336
F-statistic: 204.8 on 1 and 73 DF,  p-value: < 2.2e-16


Expected slope if '25% rule' holds:  4

Test of slope = 4 in simple model:
Estimated slope:  3.120619
t-statistic:  -4.032494
p-value:  0.0001341784

Enhanced regression model:

Call:
lm(formula = `Total U.S. Gross` ~ `Opening Gross` + `Opening Theatres` +
    Summer + Holiday + Christmas, data = movies)

Residuals:
      Min        1Q    Median        3Q       Max
-42886292 -13098227  -4528504   7636737  72332697
```

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         8.677e+06  1.051e+07   0.826    0.412
`Opening Gross`     3.214e+00  2.862e-01  11.230   <2e-16 ***
`Opening Theatres` -2.144e+03  4.503e+03  -0.476    0.635
Summer              3.366e+06  5.611e+06   0.600    0.550
Holiday            -6.144e+06  9.102e+06  -0.675    0.502
Christmas           2.550e+06  8.799e+06   0.290    0.773
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21240000 on 69 degrees of freedom
Multiple R-squared:  0.7406,    Adjusted R-squared:  0.7218
F-statistic:  39.4 on 5 and 69 DF,  p-value: < 2.2e-16


Test of slope = 4 in enhanced model:
Estimated slope:  3.214
t-statistic:  -2.746463
p-value:  0.007677049

R-squared (simple model):  0.7371966
R-squared (enhanced model):  0.7406057
```

QUESTION 8

Aim:
To evaluate the effect of critics' opinion on total U.S. box-office gross using all factors known after the opening weekend, and to quantify the monetary value of an improvement in critics' score for a movie such as "Flags of Our Fathers."

The following predictors were considered:

- Pre-production factors: Budget, COMEDY_DUMMY, MPAA_D, Sequel, Known Story
- Opening-weekend setup: Opening Theatres, Summer, Holiday, Christmas
- Opening-weekend outcome: Opening Gross, Critics' Opinion

Variable Selection

An initial regression including all these predictors was estimated with Total U.S. Gross as the dependent variable. Using a 10 per cent significance level for variable selection:

- **Null Hypothesis for each coefficient:** The population coefficient is equal to zero (no effect on Total U.S. Gross).
- **Alternative Hypothesis for each coefficient:** The population coefficient is not equal to zero.
- **Significance level for dropping variables:** 10 per cent.

Variables with p-values greater than 0.10 were dropped from the model.

Dropped variables (p-value > 0.10):

- COMEDY_DUMMY
- Sequel
- Known Story
- Opening Theatres
- Summer
- Holiday
- Christmas

Retained variables (p-value ≤ 0.10):

- Budget
- MPAA_D
- Opening Gross
- Critics' Opinion

Final Regression Model

The final model for Total U.S. Gross includes Budget, MPAA_D, Opening Gross and Critics' Opinion. The estimated coefficients are approximately:

- Intercept ≈ −30.04 million dollars
- Budget coefficient ≈ 0.2586
- MPAA_D coefficient ≈ −11.14 million dollars
- Opening Gross coefficient ≈ 2.820
- Critics' Opinion coefficient ≈ 590,800 dollars

Interpretation of key coefficients:

- **Budget:** Holding other variables constant, increasing the production budget by 1 dollar is associated with an increase of about 0.26 dollars in total U.S. gross. The p-value is less than 0.01, so at the 10 per cent significance level the Null Hypothesis that the budget coefficient is zero is rejected.
- **MPAA_D (R-rating dummy):** Holding other variables constant, R-rated movies are associated with about 11.14 million dollars lower total U.S. gross than non-R-rated movies. The p-value is about 0.034, which is below 10 per cent and 5 per cent, so the Null Hypothesis of no R-rating effect is rejected.
- **Opening Gross:** Holding other factors constant, each additional dollar of opening weekend gross is associated with about 2.82 dollars of total U.S. gross. The p-value is extremely small, so the Null Hypothesis that this coefficient is zero is strongly rejected.

- **Critics' Opinion:** Holding budget, rating and opening gross constant, a one-point increase in critics' score is associated with an increase of about 590,800 dollars in total U.S. gross. The p-value is very small (well below 1 per cent), so the Null Hypothesis that critics' opinion has no effect is rejected at the 10 per cent and 5 per cent significance levels.

Overall, the model has a high R-squared (about 0.83), indicating that these four variables jointly explain a large share of the variation in total U.S. gross.

Prediction for "Flags of Our Fathers"

Using the final model, a prediction with a 95 per cent prediction interval was made for the movie "Flags of Our Fathers."

- Point prediction for Total U.S. Gross: ≈ 57.65 million dollars
- 95% prediction interval: [21.51 million, 93.80 million]

This wide interval reflects the inherent uncertainty in movie performance even after accounting for budget, opening gross, rating and critics' scores.

Value of a +10-Point Increase in Critics' Opinion

From the final model, the slope for Critics' Opinion is about 590,800 dollars per point. Therefore, a 10-point increase in critics' rating is associated with:

- **Point estimate of extra total U.S. gross:** ≈ 5.91 million dollars

Using the standard error of the critics' coefficient and the t-distribution, a 95 per cent confidence interval for this effect is:

- **95% confidence interval for extra gross from +10 points:** [3.16 million, 8.65 million] dollars

Hypotheses for the critics' effect (10-point increase):

- **Null Hypothesis:** A 10-point increase in critics' opinion has no effect on expected Total U.S. Gross (expected change = 0).
- **Alternative Hypothesis:** A 10-point increase in critics' opinion changes expected Total U.S. Gross (expected change ≠ 0).
- **Significance level:** 5 per cent.

Decision:
Since the 95 per cent confidence interval for the effect of a 10-point increase does not include zero, the p-value is less than 0.05 and the Null Hypothesis is rejected.

Managerial Conclusion:
Critics' opinion has a strong and statistically significant impact on total U.S. gross even after controlling for budget, rating and opening performance. For a movie like "Flags of Our Fathers," it would be financially reasonable to invest up to roughly 5–6 million dollars (and certainly up to the lower bound of the confidence interval) in activities that credibly improve critics' scores by 10 points. Spending more than the expected additional revenue would not be financially justified on average.

```
QUESTION 8 - EFFECT OF CRITICS' OPINION

Initial regression model (all pre & post opening factors):

Call:
lm(formula = formula_initial, data = movies)

Residuals:
      Min        1Q    Median        3Q       Max
-23575046 -11751889  -2848326   7094965  69258552

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -3.535e+07  1.280e+07  -2.761 0.007534 **
Budget              2.492e-01  1.046e-01   2.382 0.020240 *
COMEDY_DUMMY        5.734e+06  4.996e+06   1.148 0.255482
MPAA_D             -9.922e+06  5.798e+06  -1.711 0.091930 .
Sequel             -4.040e+06  7.647e+06  -0.528 0.599079
`Known Story`      -2.606e+06  4.506e+06  -0.578 0.565049
`Opening Theatres`  1.342e+03  4.014e+03   0.334 0.739330
Summer             -2.643e+06  4.997e+06  -0.529 0.598678
Holiday            -1.345e+06  7.691e+06  -0.175 0.861725
Christmas           4.094e+06  7.712e+06   0.531 0.597327
`Opening Gross`     2.746e+00  2.769e-01   9.914 1.75e-14 ***
`Critics´ Opinion`  6.546e+05  1.609e+05   4.069 0.000134 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17570000 on 63 degrees of freedom
Multiple R-squared:  0.8381,    Adjusted R-squared:  0.8098
F-statistic: 29.64 on 11 and 63 DF,  p-value: < 2.2e-16


Final regression model after dropping variables with p-value > 0.10:

Call:
lm(formula = formula_final, data = movies)

Residuals:
      Min        1Q    Median        3Q       Max
-27945949 -10933772  -2621646   6792768  66725878
```

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -3.004e+07  7.137e+06  -4.209 7.50e-05 ***
Budget             2.586e-01  9.494e-02   2.723  0.00815 **
MPAA_D            -1.114e+07  5.166e+06  -2.157  0.03445 *
`Opening Gross`    2.820e+00  1.910e-01  14.765  < 2e-16 ***
`Critics´ Opinion` 5.908e+05  1.376e+05   4.293 5.56e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17060000 on 70 degrees of freedom
Multiple R-squared:  0.8303,    Adjusted R-squared:  0.8206
F-statistic: 85.65 on 4 and 70 DF,  p-value: < 2.2e-16


Dropped variables (p-value > 0.10 in initial model):
[1] "COMEDY_DUMMY"       "Sequel"             "`Known Story`"      "`Opening Theatres`"
[5] "Summer"             "Holiday"            "Christmas"

Prediction for Total U.S. Gross of 'Flags of Our Fathers':
       fit      lwr      upr
1 57654365 21512643 93796087

EFFECT OF CRITICS' OPINION:
Effect of a +10 point increase in critics' score:
Point estimate of extra Total U.S. Gross:  5907942
95% CI for this extra gross: [ 3163463 ,  8652421 ]
```

QUESTION 9

Aim:
To test whether critics' opinions affect comedies differently than non-comedies, by including an interaction term between Critics' Opinion and COMEDY_DUMMY in the regression for total U.S. gross. This directly addresses Griffith's belief that critics' reviews matter less for comedies.

A regression was estimated with Total U.S. Gross as the dependent variable and the following predictors:

- Budget
- MPAA_D
- Opening Gross
- Critics' Opinion
- COMEDY_DUMMY
- Interaction term: Critics' Opinion × COMEDY_DUMMY

Interpretation of the Interaction

In this model:

- The coefficient of Critics' Opinion measures the effect of critics' score on total U.S. gross for **non-comedies** (COMEDY_DUMMY = 0).

- The interaction coefficient measures how much this effect **changes** for comedies relative to non-comedies.

Hypothesis Test for the Interaction

The key question is whether the interaction term is significantly different from zero.

- **Null Hypothesis:** The population coefficient of the interaction term (Critics' Opinion × COMEDY_DUMMY) is equal to zero. In other words, the effect of critics' opinion on total U.S. gross is the same for comedies and non-comedies.
- **Alternative Hypothesis:** The population coefficient of the interaction term is not equal to zero. In other words, the effect of critics' opinion on total U.S. gross differs between comedies and non-comedies.
- **Significance level:** 10 per cent.

From the regression output, the interaction term had:

- Interaction coefficient ≈ −228,179
- p-value ≈ 0.443

Decision and Reasoning

Decision (based on p-value vs significance level):
Since the p-value for the interaction term (about 0.443) is much larger than the 10 per cent significance level, the Null Hypothesis is not rejected.

Reasoning:
A high p-value indicates that the observed interaction estimate could easily arise from random sampling variation even if the true interaction effect were zero. There is no statistical evidence at the 10 per cent level that critics' impact on box office differs between comedies and non-comedies.

Conclusion for Griffith's Theory

Griffith's theory suggested that critics' reviews matter less for comedies than for other genres. In this regression, the interaction coefficient is negative (which would be consistent with a weaker effect for comedies), but it is not statistically significant.

Therefore, based on this sample and at a 10 per cent significance level:

- We cannot statistically prove that critics' influence on total U.S. gross is weaker for comedies than for non-comedies.
- Griffith's theory is not supported by this regression analysis.

```
QUESTION 9 – INTERACTION: CRITICS' OPINION × COMEDY

Regression with interaction term (Critics´ Opinion × COMEDY_DUMMY):

Call:
lm(formula = `Total U.S. Gross` ~ Budget + MPAA_D + `Opening Gross` +
    `Critics´ Opinion` * COMEDY_DUMMY, data = movies)

Residuals:
      Min        1Q    Median        3Q       Max
-26105594 -10611342  -2136477   7567360  68111568

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -3.630e+07  8.472e+06  -4.284 5.90e-05 ***
Budget                           2.589e-01  9.511e-02   2.722  0.00824 **
MPAA_D                          -9.792e+06  5.408e+06  -1.811  0.07462 .
`Opening Gross`                  2.778e+00  1.938e-01  14.338  < 2e-16 ***
`Critics´ Opinion`               6.840e+05  1.615e+05   4.236 6.99e-05 ***
COMEDY_DUMMY                     1.667e+07  1.441e+07   1.157  0.25139
`Critics´ Opinion`:COMEDY_DUMMY -2.282e+05  2.956e+05  -0.772  0.44280
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17020000 on 68 degrees of freedom
Multiple R-squared:  0.836,     Adjusted R-squared:  0.8215
F-statistic: 57.76 on 6 and 68 DF,  p-value: < 2.2e-16


Interaction effect row (Critics´ Opinion × COMEDY_DUMMY):
     Estimate    Std. Error       t value      Pr(>|t|)
-2.281793e+05  2.955771e+05 -7.719789e-01  4.428016e-01

Interaction coefficient:  -228179.3
p-value:  0.4428016
```

QUESTION 10

Aim

To evaluate Griffith's claim that the presence of A-list actors ("star power") rather than production budget is the true driver of total U.S. box-office revenue, and to determine how the regression coefficients should behave if his argument is correct.

Discussion

Griffith argued that higher gross revenue is not primarily the result of larger production budgets. Instead, he believed that movies with a higher number of A-list stars generate greater box-office performance. He also noted that the dataset available lacked a variable measuring star power, preventing him from testing the claim.

Let a new variable representing the **number of A-list actors** in a film (Star Power) were added to the regression model as we used in Question 9. The extended regression would take the form:

Total U.S. Gross = $\beta_0 + \beta_1 \cdot$Budget + $\beta_2 \cdot$MPAA_D + $\beta_3 \cdot$Opening Gross + $\beta_4 \cdot$Critics Opinion + $\beta_5 \cdot$COMEDY_DUMMY + $\beta_6 \cdot$(Critics × Comedy) + $\beta_7 \cdot$Star Power + error

Hypotheses

To support Griffith's claim, the following would need to be true:

- **Null Hypothesis (H_0):** $\beta_7 = 0$ (Star Power has no effect on Total U.S. Gross)
- **Alternative Hypothesis (H_1):** $\beta_7 \neq 0$ (Star Power has a statistically significant effect)

Required Behaviour of Coefficients

For Griffith's conclusion to be supported by the regression results:

1. **The slope coefficient of Star Power (β_7)** must be:
    - **Positive**, shows that each additional A-list actor increases expected total U.S. gross, **without changing other factors**.
    - **Statistically significant** at the chosen level (e.g., $p < 0.05$ or $p < 0.10$), allowing rejection of H_0.
2. **The coefficient of Budget (β_1)** must:
    - **Decrease in magnitude** compared to the earlier model without Star Power.
    - **Become statistically insignificant** (p-value > significance level).

This would indicate that the previously strong positive effect of budget was driven by **omitted variable bias**—i.e., movies with large budgets tend to hire more A-list actors, and it is **the stars, not the spending**, that drive box-office success.

Conclusion

If the inclusion of Star Power causes the budget coefficient to weaken and lose statistical significance, while the Star Power coefficient becomes positive and significant, then the evidence would support Griffith's view that **star-driven casting, rather than production budget, is the true causal factor behind higher box-office performance**.

CONCLUSION

This project applied statistical tools to evaluate key assumptions about box-office performance in the Hollywood Rules dataset. The findings show that movies in the $20–$100 million budget range generally earn returns above industry benchmarks, though profitability varies significantly across genres. Several common industry beliefs—such as the idea that opening weekend represents only 25% of total U.S. gross—do not hold statistically; opening performance contributes a much larger share.

Regression analyses highlight that budget, sequel status, opening theatres, opening gross and critics' opinion all meaningfully influence total revenue, with opening weekend gross emerging as the strongest predictor. Critics' reviews also have a measurable financial impact, and contrary to Griffith's belief, this effect does not differ significantly between comedies and non-comedies.

Overall, the study demonstrates that data-driven analysis offers clearer guidance than traditional rules-of-thumb. Understanding the combined effects of opening performance, production factors and critical reception allows studios and investors to make more informed decisions in a highly uncertain industry.