Final Project Report:
Estimation of Movie's Worldwide Gross($)

Group 9: Yiyang Jiao, Jiayi Guo, Xuming Huang, Hao Ni, Yongwei Zhu
INF1344: Introduction to Statistics for Data Science
Professor Tao Wang

Motivation

The global box office revenue of a film is a critical indicator of its commercial success and cultural impact. It also underscores the film industry's significant role in the global economy and its influence on cultural trends. Understanding the factors that drive a movie's financial performance is essential for filmmakers, producers, and marketers aiming to create content that resonates with diverse audiences. This research seeks to analyze the determinants of box office success, including production budgets, genres, release strategies, and marketing efforts. By examining these elements, the study aims to provide insights into audience preferences and the evolving dynamics of the film industry. Such knowledge is invaluable for industry stakeholders striving to align their creative and business strategies with market demands. Moreover, the findings of this research have broader implications beyond the entertainment sector. Industries such as merchandising, streaming services, and tourism are closely linked to the success of films. For instance, the rise of streaming platforms has transformed content distribution and consumption patterns, influencing box office revenues and audience engagement(Hall & Pasquini, 2020). Understanding the factors that contribute to a film's success can inform strategies across these related sectors, fostering a more integrated approach to meeting consumer expectations.

Review of Similar Research

There are many statistical studies published to demonstrate determining factors of movie’s commercial success. An empirical study used Hollywood movies released in North

America from 1999 to 2003 as samples, and screened and analyzed the factors that affect consumers' purchase decisions for new movies in two distribution methods: platform distribution and theatrical distribution. In the statistical analysis of platform movies, researchers introduced the concepts of unpredictable appeal and prior beliefs, which refer to the use of advertising marketing in the early stage of movie release to give consumers some prior influence on new movies and to reduce the unobservable appeal of the movie itself before it is fully watched. After sample modeling, it shows that 70% of the sample movies will achieve better commercial results if they are released as large-scale theatrical releases(Chen et al., 2012). In contrast, the unobservable appeal of relatively high-profit movies (-7.28) is lower than that of low-profit movies (-6.98)(Chen et al., 2012). Specifically, unobservable appeal is judged by a series of indicators such as Oscars and film investment. In addition, in terms of analyzing advertising investment, movies released on platforms help to improve the audience's prior perception of the unobservable appeal of movies, although platform releases accelerate the decline of movie appeal after comparison (Chen et al., 2012). Another study measured the relationship between star effect and movie box office by integrating box office data from nine different countries and regions and the number of visits to the star's homepage on IMDB. The study mainly used dummy variables such as Oscar nominations or homepage visits to measure the influence of stars. In a sample of 2,858 observations, the model ultimately showed that the participation of a top star can increase the box office revenue of a movie by between $5.22 million and $28 million (Nelson & Glotfelty, 2012). Therefore, the economic benefits brought by stars have been proven to have a significant impact on the commercial success of a movie. Existing research both supports that

multiple categorical variables are crucial in terms of a movie's box office, and demonstrates strong correlation between star powers, promotion platforms and movie gross.

Research Question

This report will study key factors that influence movie's worldwide gross($), and further use significant predictors to estimate it. More specifically, this report focuses on the relationship of a movie's runtime, budget, IMDb ratings, certificate, release year and its worldwide gross($).

Data

Data source

To analyze the determinants of movie revenues, we chose a dataset from Kaggle, an open data platform, entitled "Top 100 Movies from 2003 to 2022," published in 2023. We also considered other datasets from various sources related to the movie and entertainment sector but did not include them in the study because of their limited temporal coverage (data for specific years rather than a broader time span), lack of detailed metadata (e.g., lack of records on directors, production budgets, etc.), and lack of up-to-date data (data that has not been updated in the last ten years).

The dataset was originally created for a film industry analysis project that dug into the trends in the popularity of movies based on IMDb data. The dataset contains major metrics like ratings, box office receipts, and audience trends, which turn out to be a rich source of reference information needed when one is analyzing the various factors (genres, production budget, reviews, etc.) that go into making movie revenues.

Data Gathering and Cleaning

Since Kaggle requires a registered account to download the dataset, we downloaded this data in CSV format using our account. This dataset aggregates the data from 100 popular movies between the years 2003 and 2022, including variables such as rating, year, runtime, and genre, hence giving an overview of the film industry for two decades.

We then preprocessed the dataset for analysis by scrutinizing the values of the data for missing or duplicated values and reviewing the variable types for consistency. We also addressed potential outliers that could bias the results. Using tools such as Python and RStudio, we did not find any missing values or errors in the dataset; , Below, we focus on the relevant metrics needed for analysis in the dataset, including “budget,” “rating,” and “runtime” variables, as well as “certificate” variables used to analyze revenue influencing factors, such as “movie type” and other “certificate” variables.

For example:

- Budget
- Year: Year the movie was released
- Income: Revenue from movies.
- Rating: Rates the audience on a scale of 1 to 10.
- Runtime: The length of the film, measured in minutes.
- Genre: Movie genre, categorizes movies into Action, Drama, Comedy, etc. It helps to find out the box office trend of each type of movie.
- Certificate: Movie ratings are a censorship system that categorizes and rates movies as age-appropriate based on their content.

We processed the data differently to clean it up and make it more accurate and consistent. The “Income”, “Budget”, and “Runtime” columns were cleaned by removing non-numeric characters from them and then converting those columns to numeric types. We then proceeded to remove rows containing missing values to maintain the integrity of the data. We standardized the currency units of the budget by extracting the currency symbols in the budget columns and converting the various currencies to U.S. dollar units based on a predefined table of exchange rates. We then simplify the structure of the dataset by extracting columns that are relevant to the analysis, such as runtime, certificates, ratings, budget, gross, and year. Unusual or unrecognizable currency symbols are handled by extracting non-numeric symbols in the budget column. This greatly improves data availability and provides a solid foundation for further analysis and modeling.

Methodology

Descriptive Statistics

The descriptive statistics are implemented on the dataset; RStudio is used to summarize the variability and key features of the movies from 2003 to 2022. From this point forward, we calculate measures such as minimum, maximum, mean, median, and quartiles for key variables **(Table 1)** that are very essential in understanding the trends and distributions of the dataset.

The lengths of movies, as represented in this dataset, ranged from 71 to 209 minutes with a mean length of approximately 114 minutes. Ratings ranged from 1.9 to 9.0 with an average of 6.72, meaning most movies get moderately to highly positive audience responses. Finally, budget values range from $11 to over $1.22 billion with a median of $40 million, epitomizing the huge financial divide in movie production. The gross revenue varies between $3,492 and almost $2.93

billion with a mean revenue of $210.3 million, which shows the blockbuster nature of many films in the dataset.

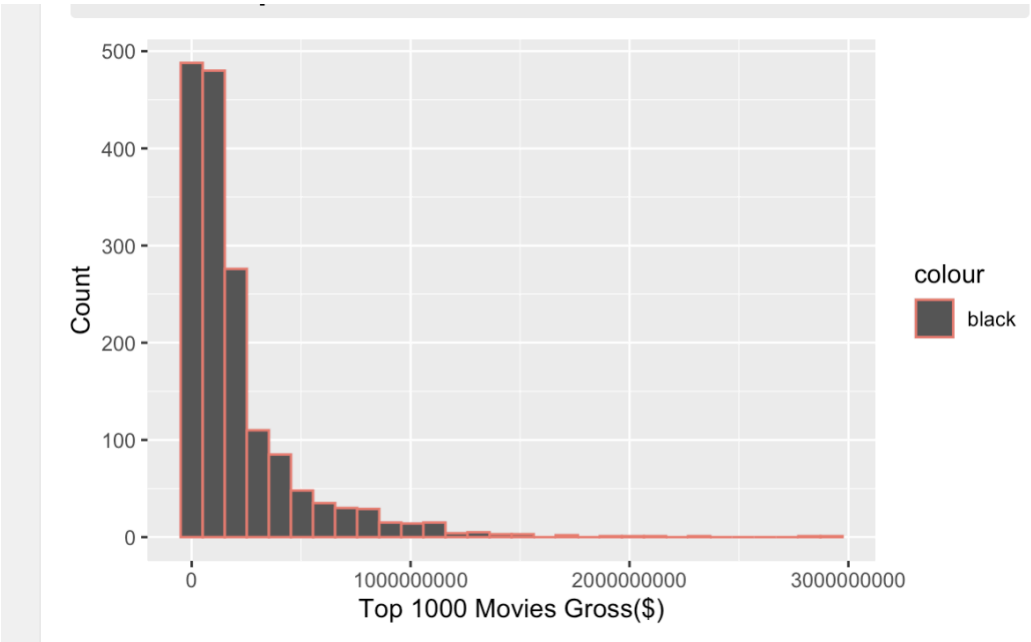
These statistics present the quantitative base for exploring other factors that affect movies' revenues, such as runtime, budget allocation, and audience ratings, which allow more analytical modeling and trend analysis.

```
   runtime      Certificate           Rating          budget
Min.   : 71.0   Length:1648        Min.   :1.900   Min.   :         11
1st Qu.:100.0   Class :character   1st Qu.:6.200   1st Qu.:   18000000
Median :112.0   Mode  :character   Median :6.800   Median :   40000000
Mean   :114.4                      Mean   :6.722   Mean   :   86681644
3rd Qu.:126.0                      3rd Qu.:7.300   3rd Qu.:   90000000
Max.   :209.0                      Max.   :9.000   Max.   :12215500000
    gross                Year
Min.   :      3492   Min.   :2003
1st Qu.:  41989137   1st Qu.:2007
Median : 106726390   Median :2011
Mean   : 210328877   Mean   :2012
3rd Qu.: 245551324   3rd Qu.:2016
Max.   :2922917914   Max.   :2022
```

Table 1 Descriptive Statistics of Key Variable

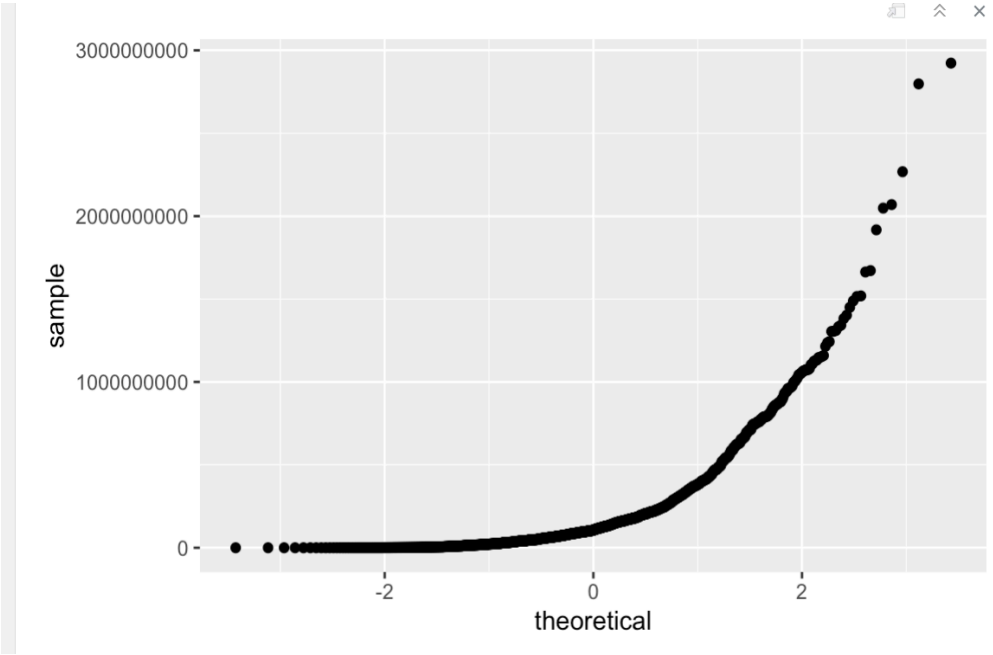
Shows the frequency distribution of box office receipts by plotting a histogram of the distribution of movie box office gross (**Figure1**). Helps us to understand the overall range of box office receipts, concentration trends, and possible outliers. We find that box office receipts show a clear right skewed distribution (most movies are concentrated in the lower range). The long right tail of the histogram suggests that there are a small number of high grossing movies (e.g., movies that grossed over $1 billion or even $2 billion), which are likely to be “blockbusters” that are big hits around the world.

Figure1 Histogram of the distribution of movie gross



By plotting the Q-Q Plot of the box office revenue data **(Figure 2)**, we found that the box office revenue data is deviated from the normal distribution The deviation of the middle point of the plot from the theoretical diagonal line is very obvious, especially in the tail of the distribution. This indicates that the box office receipts of the movie do not conform to the normal distribution and that the data distribution is strongly skewed. Box office receipts have a long-tailed distribution, concentrated in the lower revenue range, with the presence of a few very high revenue movies. Therefore, it may be necessary to use an asymmetric distribution model when analyzing box office receipts. Perform a logarithmic transformation on the box office revenue data to reduce the effects of skewness and bring the distribution closer to a normal distribution, thus making it easier to analyze statistically.

Figure 2 Q-Q Plot of the movies gross

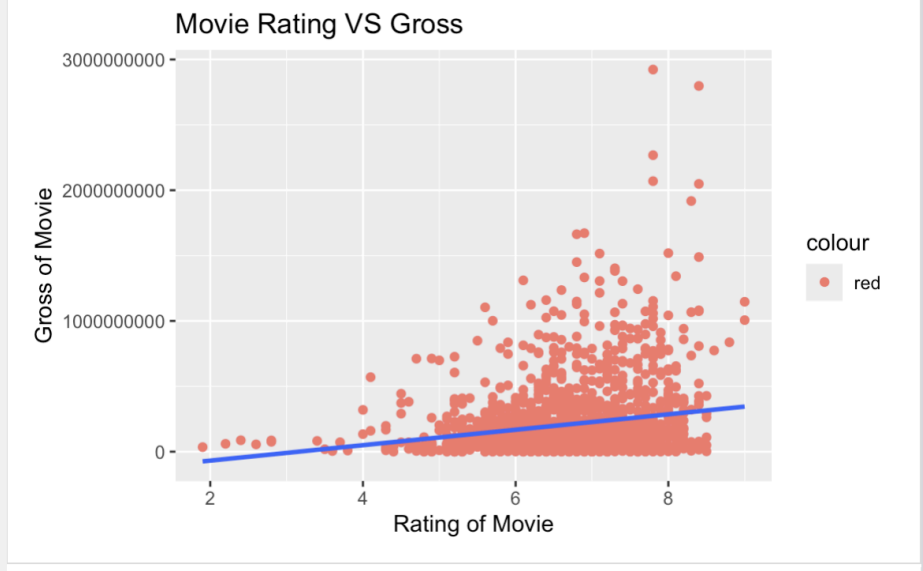


We show the relationship between Rating and Gross of Movies by plotting a scatter plot **(Figure3)**. The blue regression line shows a positive correlation between ratings and box office receipts, i.e., the higher the rating, the higher the box office receipts are likely to be.

However, the distribution of scatter points indicates that this relationship is weaker and that it is not the case that movies with high ratings necessarily have high box office gross. Most data points are concentrated in the range of ratings 6-8 and lower box office gross. A small number of higher rated movies (8 and above) correspond to very high box office receipts (over $1 billion), suggesting that there are some “blockbusters” or phenomena among the highly rated movies. At the same time, we also find some outliers, as certain movies with lower ratings (e.g., below 5) also show higher box office receipts (over $500 million), which may be related to specific factors (e.g., well-known actors, or franchise franchises). This suggests that there is

some positive correlation between ratings and box office, but the correlation is weak, suggesting that ratings are not the only factor influencing box office. In the follow-up, we combined ratings with other variables (e.g., budget, genre, etc.) in a multivariate regression analysis to quantify the actual impact of ratings on the box office.

Figure 3 Relationship between Rating and Gross of Movies by plotting a scatter plot

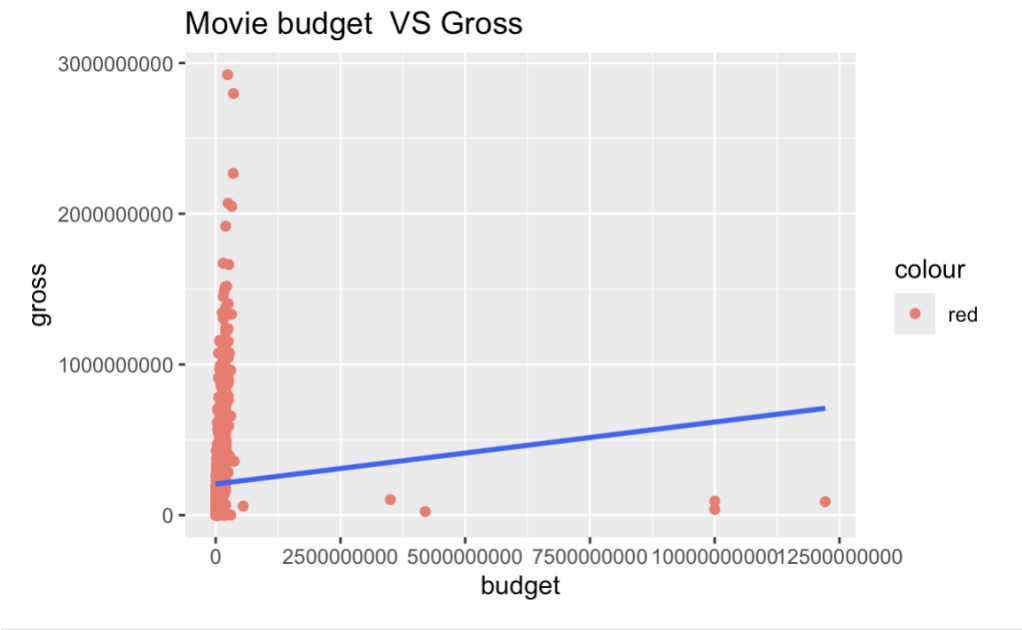


We show the relationship between Budget and Gross by plotting a scatter plot **(Figure4)**. The blue regression line shows that there is a positive correlation between Budget and Box Office Revenue: the higher the budget of a movie, the higher the Box Office Revenue tends to be. However, the distribution of points is more dispersed, indicating that the relationship between budget and box office is not completely linear.

There are some high box office but low budget movies in the graph, which may be due to their market positioning or production strategy. Some of the high-budget movies did not earn commensurately high box office revenues, which may be related to marketing or audience

acceptance. Overall, budget is a significant factor in box office, with movies with higher budgets typically having higher box office revenue potential..

Figure4 The relationship between Budget and Gross by plotting a scatter plot

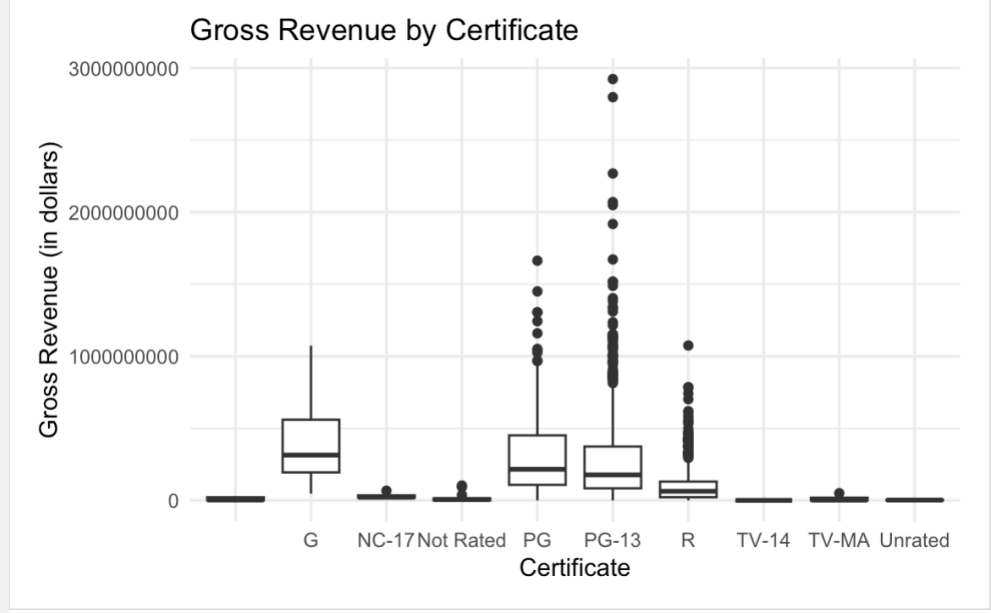


The distribution of Gross for different Certificates is shown by plotting a box-and-line diagram **(Figure5)**. The following are the characteristics and analysis that can be observed from the graph:

- G-rated (General Audiences, suitable for all): Higher median and narrower distribution indicate that these movies usually have stable box office performance.
- PG (Parental Guidance): Higher median, wider distribution, and includes many high-grossing movies;
- PG-13: Widest distribution and contains many high-paying movies.

- Rated R (Restricted): The median is relatively low, but there are some outliers that indicate some R-rated movies are strong performers at the box office.
- NC-17, TV-MA, Unrated, etc: Generally low box office receipts and small distribution. Movies with these ratings are usually geared toward a very limited group of viewers.
- Outliers: There are significant outliers (ultra-high grossing movies) in the PG and PG-13 ratings, such as some animated or superhero blockbusters, and some high-grossing outliers in the R rating, which may be driven by specific blockbusters.

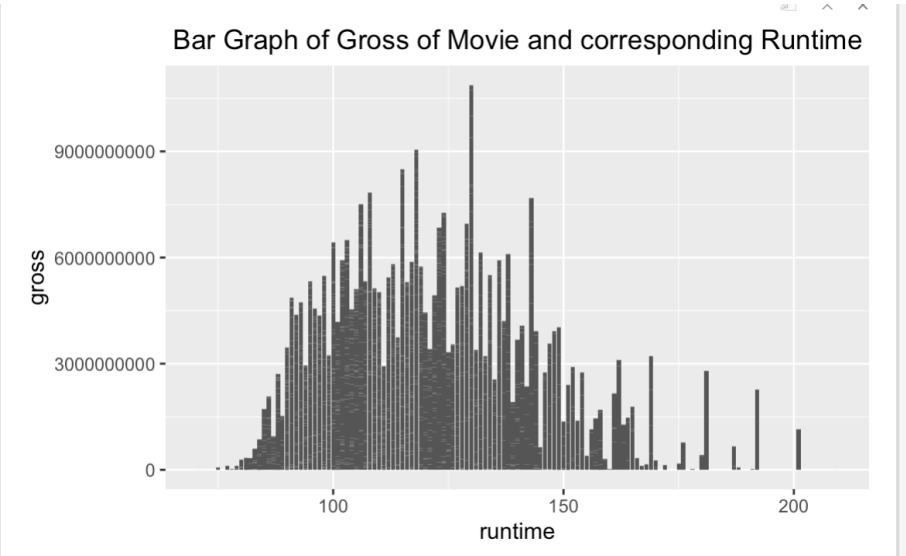
Figure5 The distribution of Gross for different Certificates



The bar chart **(Figure6)** shows the relationship between gross and the corresponding runtime of a movie, and we find that box office revenues are relatively higher and more densely distributed for movies in the 90 to 120 minute range. This suggests that most high grossing movies are in this range. Movies that are less than 90 minutes long generate less box office revenue, and movies that are longer than 150 minutes also generate relatively less box office

revenue. There are some exceptions, such as movies with shorter or longer durations that still generate higher box office gross.

Figure6 Relationship between gross and the runtime of movies



Assumption

Initial Linear Model Fit

We fitted a linear regression model using key predictors, such as movie runtime, IMDb rating, budget, release year, and certificate category, in order to understand the factors that affect a movie's global gross revenue.

Model Summary and Interpretation

The fitted linear model is:

Gross = −10589220072.46 + 4014938.36×Runtime + 38745478.35×Rating + 0.02525×Budget + 4876689.97×Year + Certificate Effects

Table 2: Summary of the Initial Linear Regression Model

```
Call:
lm(formula = gross ~ runtime + Rating + budget + Year + Certificate,
    data = clean)

Residuals:
       Min         1Q     Median         3Q        Max
-556651639 -141219374  -36666753   72499417 2409111359

Coefficients:
                            Estimate         Std. Error t value             Pr(>|t|)
(Intercept)          -10589220072.46271   2310850168.91368  -4.582          0.000004946 ***
runtime                    4014938.35800       357062.73587  11.244 < 0.0000000000000002 ***
Rating                    38745478.35402      7461674.12766   5.193          0.000000233 ***
budget                           0.02525            0.01292   1.955              0.05078 .
Year                       4876689.97165      1146409.59616   4.254          0.000022201 ***
CertificateG             553489884.61391    122754511.35266   4.509          0.000006980 ***
CertificateNC-17         -14222078.65127    164912648.94250  -0.086              0.93129
CertificateNot Rated      44458467.75262    119391994.35428   0.372              0.70966
CertificatePG            445828875.14908    111472644.74384   3.999          0.000066324 ***
CertificatePG-13         347138089.04751    110275905.24131   3.148              0.00167 **
CertificateR             155452323.01298    110222808.71877   1.410              0.15863
CertificateTV-14         171669667.75721    268756044.20122   0.639              0.52307
CertificateTV-MA          29642590.91671    143509530.10545   0.207              0.83638
CertificateUnrated        73641710.97234    205082946.38265   0.359              0.71958
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 245000000 on 1634 degrees of freedom
Multiple R-squared:  0.2753,    Adjusted R-squared:  0.2695
F-statistic: 47.74 on 13 and 1634 DF,  p-value: < 0.00000000000000022
```

According to the p-values from the original linear regression model, the year, runtime, IMDb rating, and specific certificate categories (like G, PG, and PG-13) are the significant factors affecting movie gross revenue. While other predictors, such as budget and some certificate categories, exhibit weak or no significant effects.

Model Fit Summary

- **R-squared**: The R-squared value of 0.2753 indicates that about 27.5% of the variance in gross revenue is explained by the model, suggesting that there are other important factors not captured here.

- **Adjusted R-squared**: The adjusted R-squared is similar, that is 0.2695, indicating that the model's explanatory power is limited.
- **F-statistic**: The overall F-statistic is highly significant (p-value < 0.0000000000000022), indicating that the model is statistically significant.

Conclusion on Model Fit

It is clear from the previous analysis that there is right skewness in the gross revenue (Gross). A linear regression model's ability to accurately fit the data may be hampered by this type of skewness, which could affect the model's predictive capabilities. To better manage this skewness, it is advised to think about using a log transformation.

Figure 7: Y vs. Y_hat Plot Analysis

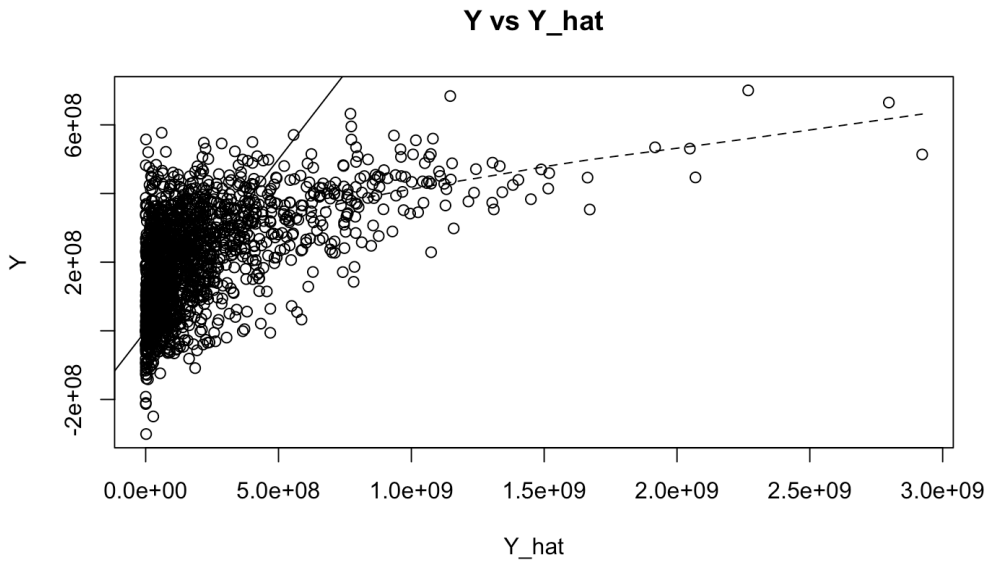


Figure 7 shows the scatter plot of actual gross revenue (Y) versus the predicted gross revenue (Y_hat) to evaluate the model's fit.

Analysis

- Most of the points are clustered in the lower left area of the plot, indicating that the model performs relatively well in predicting lower gross revenues.
- **Points Deviating from the Diagonal Line**: The model struggles to predict the gross revenue of high-grossing films, as evidenced by the significant deviations from the diagonal, especially in regions with higher gross revenue.
- **Trend Line**: The LOWESS line deviates greatly from the ideal diagonal line, particularly for higher gross revenues. This implies that the model does not fully capture the variability in high-grossing movies.

Conclusion

This plot illustrates the model's overall predictive ability. While it performs reasonably well for lower-grossing movies, it exhibits significant errors in predicting high-grossing movies.

Figure 8: Fitted vs. Residuals Plot Analysis

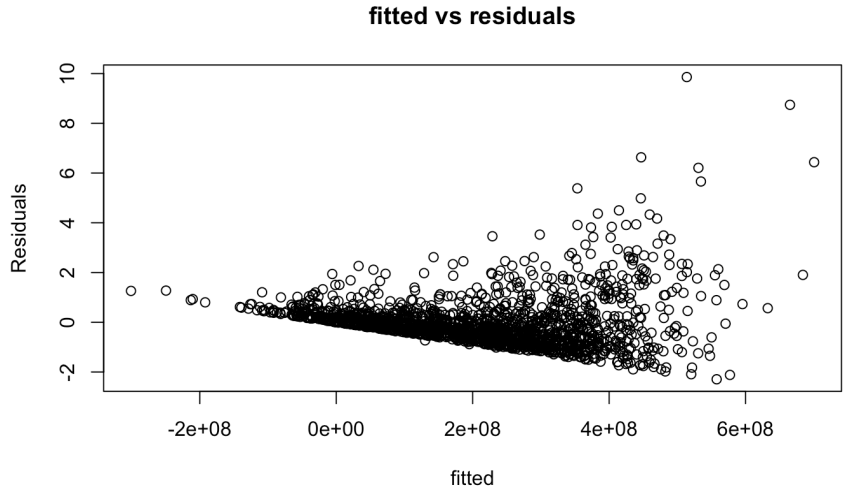


Figure 8 shows a scatter plot of the standardized residuals versus the fitted values of the linear regression model.

Analysis

- In the plot, there appears to be a **funnel-shaped pattern**, with residuals spreading out more as the fitted values increase. This suggests **heteroscedasticity**, meaning that the variability of residuals is not constant across all levels of predicted values.
- The increasing spread indicates that the model may not be capturing all relevant variability, especially for higher gross revenue predictions. This could lead to less reliable estimates and inferences in those areas.

Conclusion

The **heteroscedasticity** observed here violates the Homoscedasticity of linear regression assumption.

Figure 9: Runtime vs. Residuals Plot Analysis

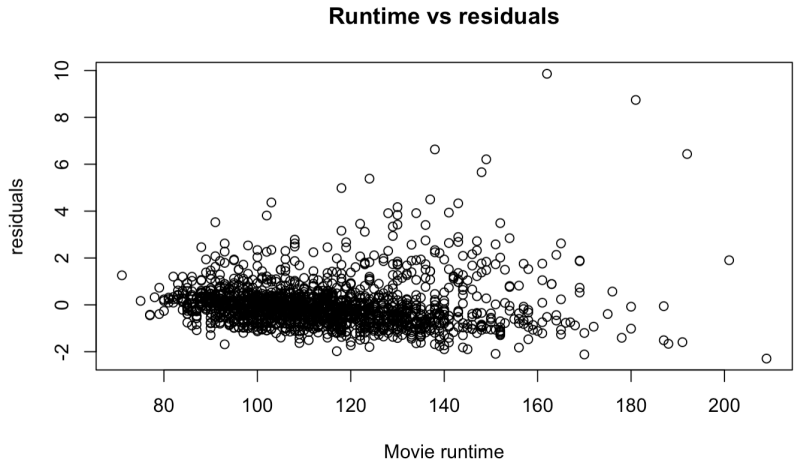


Figure 9 displays the standardized residuals plotted against the movie runtime.

Analysis

- From the plot, we observe that the residuals are generally clustered around zero, with a slight upward spread as runtime increases.
- There is a **fanning-out effect** visible for movies with a runtime greater than 120 minutes, indicating potential **heteroscedasticity**. This suggests that the variability of residuals is larger for movies with longer runtimes.
- Additionally, there are several **outliers** with high residual values, particularly in the longer runtime range (>150 minutes), which indicates that the model struggles to accurately predict the gross revenue for these longer movies.

Conclusion

The presence of increased residual variability for longer runtimes indicates a violation of the homoscedasticity assumption, suggesting that the model does not capture the entire relationship between runtime and gross revenue.

Figure 10: Rating vs. Residuals Plot Analysis

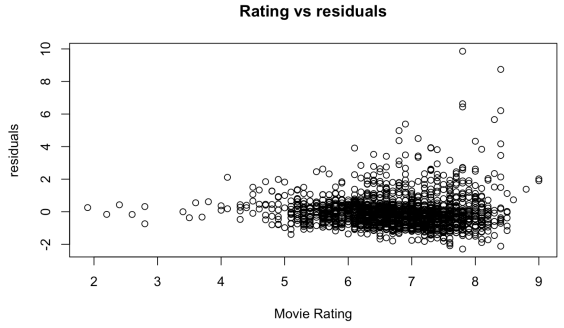


Figure 10 shows a scatter plot of the standardized residuals plotted against the IMDb rating.

Analysis

- From the plot, we observe that the **residuals are clustered densely between IMDb ratings of 5 to 8**, indicating that most of the data points fall within this range.
- There is a slight **fanning-out effect** for higher ratings (above 7), suggesting potential **heteroscedasticity**—the spread of residuals seems to increase with higher ratings. This implies that the model may not capture the variability in movies with higher ratings very well.
- The residuals are also relatively well-distributed around zero for the majority of ratings, but there are a few **outliers**, especially for movies with high ratings (greater than 8), indicating that the model tends to either overestimate or underestimate the gross revenue for these high-rated movies.

Conclusion

The observed pattern and increasing spread of residuals at higher ratings suggest **heteroscedasticity**, which is a violation of the homoscedasticity assumption in linear regression.

Figure 11: Budget vs. Residuals Plot Analysis

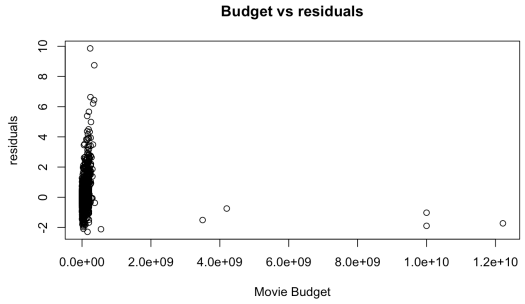


Figure 11 shows the standardized residuals plotted against the movie budget.

Analysis

- The plot reveals that most of the **data points are concentrated at lower budget values**, suggesting that the majority of movies have relatively modest budgets.
- The **spread of residuals** is fairly consistent for lower budget levels, indicating that the model captures the variability reasonably well for these cases.
- The **outliers** at higher budget values are prominent, with residuals showing large deviations from zero. This could imply that the model struggles to make accurate predictions for high-budget movies, possibly due to their unique characteristics (e.g., blockbuster marketing or production quality).

Conclusion

The residual pattern suggests that the model performs well for the majority of movies with **lower to moderate budgets**, but it struggles with **high-budget movies**, which also show significant outliers.

Figure 12: Year vs. Residuals Plot Analysis

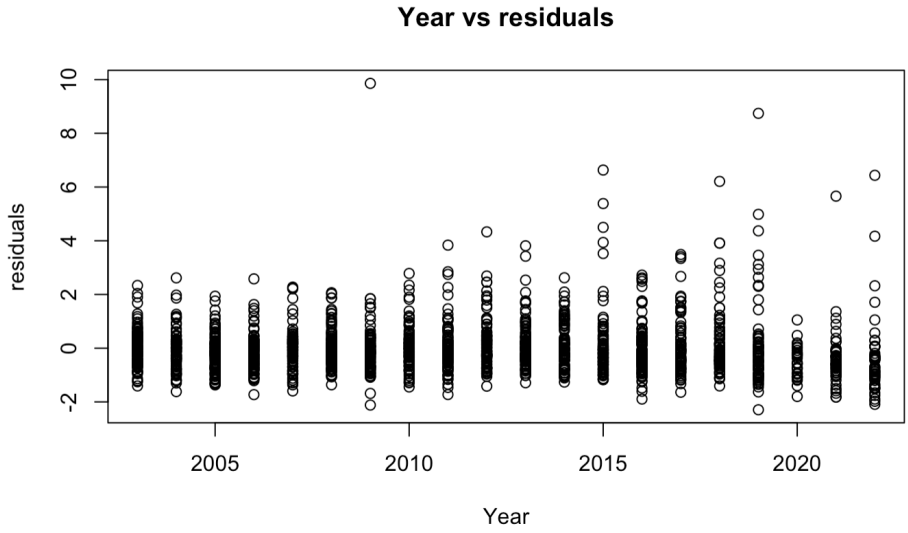


Figure 12 shows the standardized residuals plotted against the year of movie release.

Analysis

- The residuals are fairly evenly distributed across different years, with no strong upward or downward trend visible over time.
- However, there are several **outliers** in recent years, especially from around 2015 onwards, where the residual values are much higher. This may indicate that the model struggles with accurately predicting the gross revenue of movies in recent years.
- **Variability**: The spread of residuals appears consistent across the years, which is a good indication that the model is generally fitting consistently over different time periods.

Conclusion

There is no major trend in residuals over time, suggesting that the model fits reasonably well across different years. However, the presence of outliers in recent years highlights areas where the model might need improvement.

Figure 13: Certificate vs. Residuals Plot Analysis

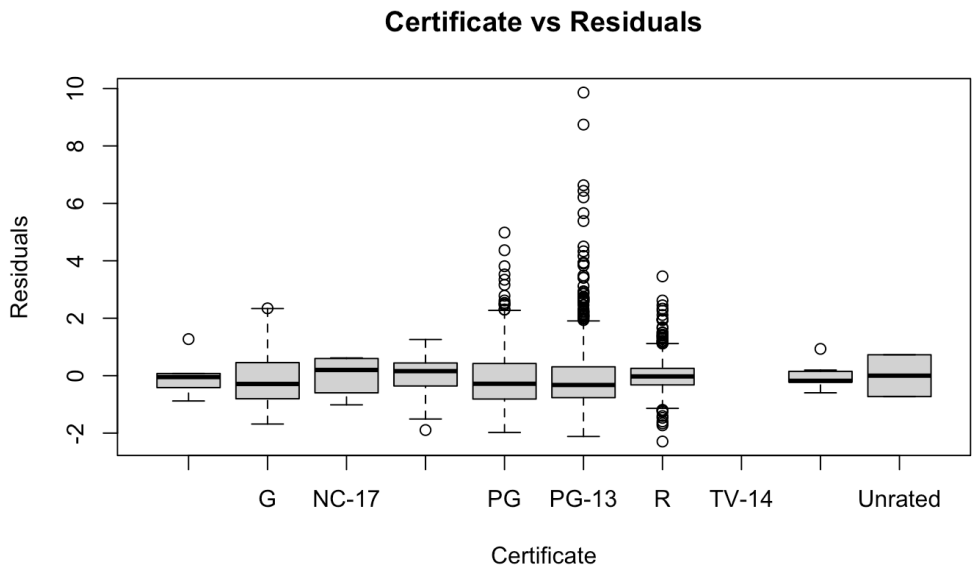


Figure 13 displays a boxplot of standardized residuals against different movie certificate categories.

Analysis

- The **residuals** are distributed relatively evenly across different certificate categories, with **medians** close to zero for most of the categories.

- **Outliers**: There are many outliers in the categories PG, PG-13, and R. This suggests that the model has more difficulty accurately predicting gross revenue for movies in these categories.

Conclusion

The presence of many outliers, especially for popular categories like PG, PG-13, and R, highlights the need for additional features in the model that may better capture the variability in these types of movies.

Figure 14: Normal Q-Q Plot of Residuals

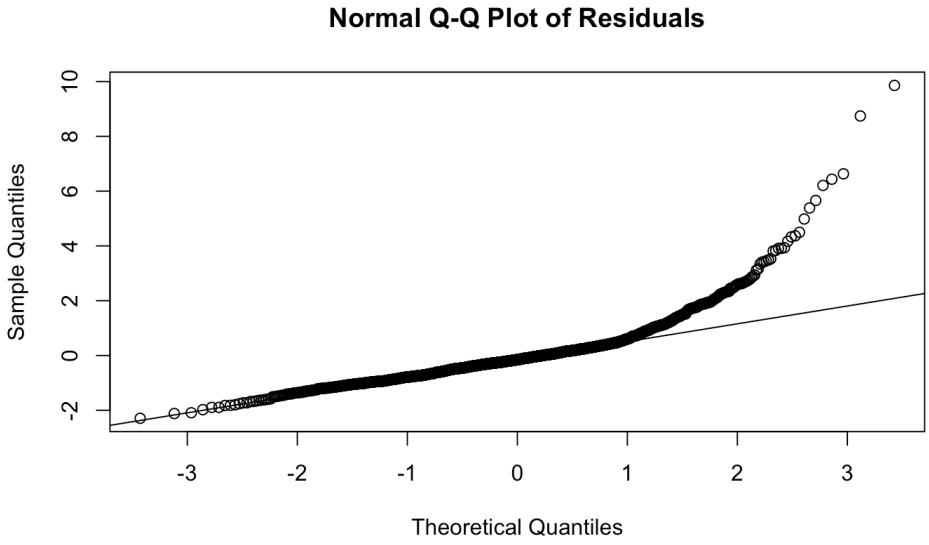


Figure 14 shows a **Normal Q-Q Plot** of the standardized residuals from the linear regression model.

Analysis

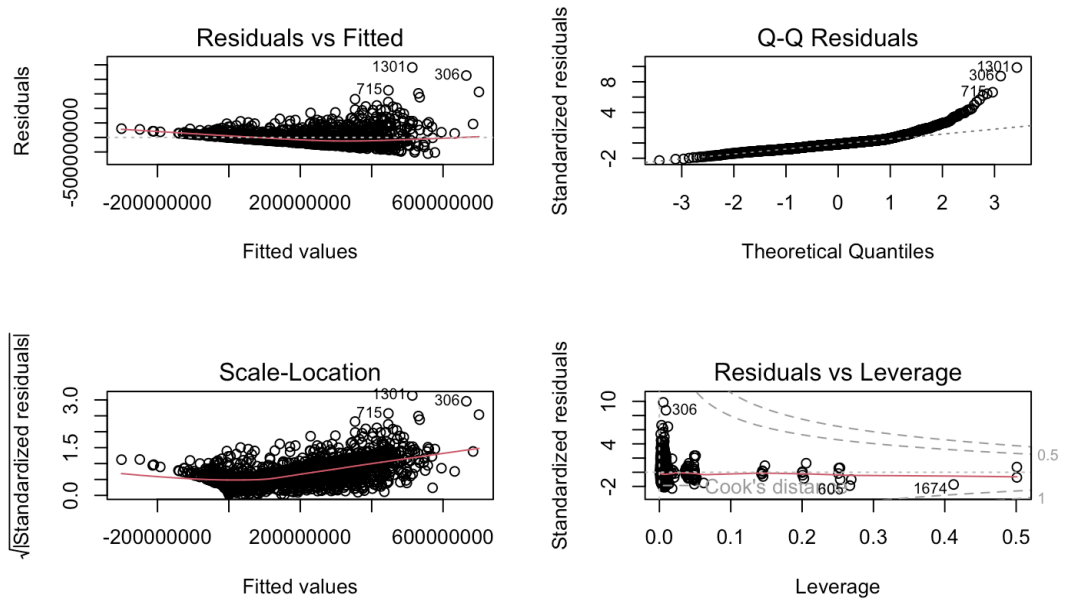
- The plot reveals that the **residuals deviate from the straight line**, especially in the **tails**. The points in the middle portion of the plot align relatively well with the line, indicating that the majority of the residuals are approximately normally distributed.
- However, the **deviation at the upper end** of the plot suggests that there are **more extreme positive residuals** than expected under a normal distribution. This could indicate the presence of outliers or skewness in the data, suggesting that the normality assumption of the residuals may not be fully satisfied.

Conclusion

The deviation from normality at the tails of the plot suggests that the **normality assumption** for residuals is violated.

Transformation

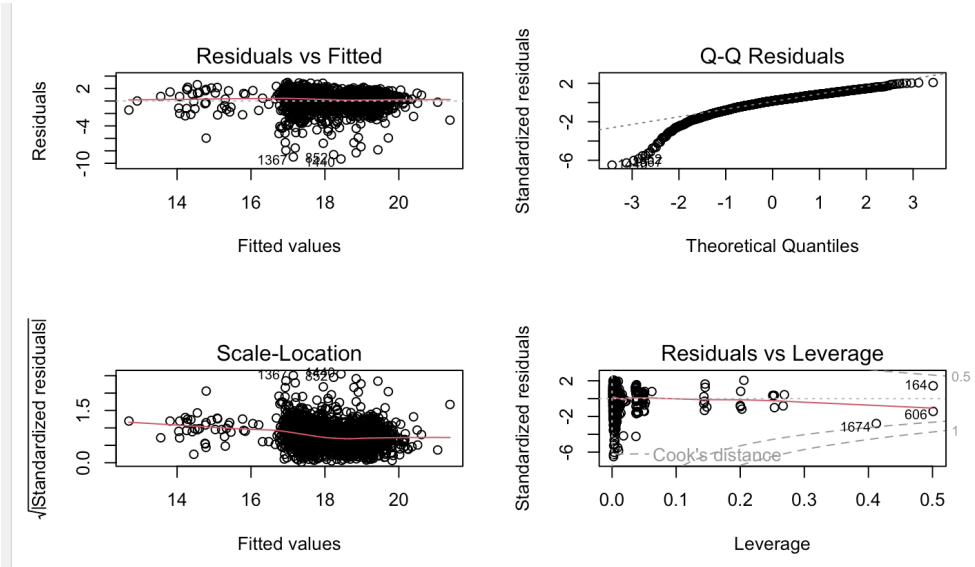
Figure 15 (Residual Plots and QQ Plots)



In Figure 15, we show a set of regression diagnostic plots to identify some issues with the model assumptions: the relationship between residuals and fitted values suggests possible nonlinearity (Residuals vs Fitted), the normality of residuals is violated by deviations at the tails (Q-Q Plot), and the Scale-Location Plot indicates slight heteroscedasticity. Furthermore, the Residuals vs Leverage Plot highlights several high-leverage points (e.g., 605 and 1674), which could affect the model's accuracy and stability.

Thus, we decided to apply a log transformation to the dependent variable gross, changing it to log(gross). The updated model is model_loggross <- lm(log(gross) ~ runtime + Rating + budget + Year + Certificate, data = clean). We then used the same method to review the regression diagnostic plots for the transformed model, as shown in Figure 16.

Figure 16 (Transformed Residual Plots and QQ Plots)



By comparing the original model with the model after applying a log transformation to gross, it is clear that the log transformation is effective in improving the model's assumptions. First, in the Residuals vs Fitted plot, the residuals of the original model show a clear systematic pattern,

especially in the region with high fitted values. The residuals are not randomly distributed, indicating that the original model failed to fully capture the relationships in the data, possibly due to nonlinearity. After applying the log transformation to the dependent variable, the residuals' distribution is significantly improved, becoming more evenly and randomly scattered around 0, with the systematic patterns largely reduced. This suggests that the log transformation successfully addresses potential nonlinearity in the dependent variable and improves the model's fit, making it align better with the linear regression assumptions. In the Scale-Location plot, the original model shows clear signs of heteroscedasticity, where the variance of the standardized residuals increases as the fitted values grow. After the log transformation, the variance of the residuals becomes more uniform, and the red smoothing line is closer to being flat. This indicates that the issue of heteroscedasticity has been reduced. The log transformation effectively minimizes the impact of extreme values on the model's variance, which makes the model assumptions more reliable.

In the Q-Q Plot, the normality check for the original model reveals a significant departure from the theoretical normal distribution, especially at the tails, suggesting the presence of heavy-tailed residuals or the influence of outliers. After applying the log, the residuals align more closely with the theoretical normal distribution. Although some deviations still exist at the tails, the extent of these deviations is reduced, and the impact of outliers on the model has also lessened. Also, in the Residuals vs Leverage plot, the original model shows several high-leverage points (e.g., points near the Cook's distance threshold), which could have a large impact on the model and weaken its robustness. After the log transformation, although some high-leverage

points still exist, the number is reduced, such as point 605 is not shown. Most of the data points are now concentrated in the lower leverage region, indicating that the model robustness has improved.

Overall, applying the log transformation to gross has improved the model's performance in multiple aspects, addressing key issues in linear regression. These improvements make the transformed model better aligned with the assumptions of linear regression, thereby enhancing its interpretability and predictive accuracy.

Model multicollinearity test

Figure 17 VIF Testing Table

	GVIF	Df	GVIF^(1/(2*Df))
runtime	1.312389	1	1.145596
Rating	1.204260	1	1.097388
budget	1.054782	1	1.027026
Year	1.055825	1	1.027533
Certificate	1.188295	9	1.009631

As shown in Figure 17, the result of the Generalized Variance Inflation Factor (GVIF) analysis shows that the problem of multicollinearity among the variables is very slight. The adjusted GVIF values of all the variables are close to 1. For example, the adjusted value of runtime is 1.1456, Rating is 1.0974, and the adjusted value of the factor variables such as Certificate is 1.0096. These results indicate that each variable in the model has a low linear relationship with the other explanatory variables, and the problem of multiple covariance can be ignored. Therefore, each variable is suitable for regression analysis without further adjustment or removal of variables.

Results and Interpretations

Overall Model Performance

Figure 18 shows the statistical summary of our final model of estimation of movie's worldwide gross. Overall, we have successfully generated formula for estimating a movie's gross: Gross= $e^{\log(gross)}$, where

$\log(gross)=27.469+0.021 \cdot runtime+0.148 \cdot Rating+2.94 \times 10^{-10} \cdot budget+$Certificate Coefficients

Where Certificate Coefficients are specific to the movie's certification:

- G: +4.501
- PG: +4.002
- PG-13: +3.481
- R: +2.259

The residual standard error represents the average deviation of the predicted gross revenue from the observed values. A high residual standard error indicates substantial unexplained variation in the model. Based on R-square value, 32.88% of the variability in gross revenue is explained by the predictors. An adjusted R-square of 0.3235 indicates that the model is only slightly penalized for including additional variables. Based On the F-statistic(61.58) and p-value (< 0.00000000000000022), the overall model is highly statistically significant, meaning the predictors collectively explain gross revenue better than a model with no predictors.

Regression Coefficients-Intercept

An intercept of 27.47 represents the gross revenue of a movie when all predictors are zero. Since most predictors cannot realistically be zero, this value has no practical interpretation but serves as a baseline for the model.

Regression Coefficients-Runtime

The estimated coefficient for Runtime is 0.021, which represents that for each additional minute of runtime, the gross revenue is predicted to increase by around 2.1%,, holding other variables constant. P-value for runtime indicates this relationship is highly statistically significant.

Regression Coefficients-Rating

A coefficient of 0.148 suggests that higher ratings are associated with a 14.8% increase in gross revenue, holding other factors constant. This is statistically significant (p=0.00069).

Regression Coefficients-Budget

The estimated coefficient for Budget is 2.94×10^{-10}, meaning that for every additional dollar spent on the budget, the gross revenue is predicted to increase minimally while holding other factors constant. P-value of 0.0001 indicates its statistical significance at the 5% level.

Regression Coefficients-Certificate

Movies rated "G" (General) are associated with a substantial increase in gross revenue, with a coefficient of 4.50, indicating a 4.50% increase compared to the baseline. Similarly, "PG" and "PG-13" movies have coefficients of 4.00 and 3.48, respectively, showing significant positive impacts on gross revenue ($p<0.001$). In addition, Movies rated "R" also show a significant positive effect with a coefficient of 2.26 (p=0.0005).

Insignificant Variables

Based on a p-value of 0.249, Year is not significantly correlated with movie's global revenue. Moreover, other certifications, such as "NC-17," "Not Rated," "TV-14," "TV-MA," and "Unrated," do not exhibit statistically significant effects with corresponding p-values that are all greater than 0.05 .

Figure 18 Summary Table of Final Regression Model

```
Call:
lm(formula = log(gross) ~ runtime + Rating + budget + Year +
    Certificate, data = clean)

Residuals:
    Min      1Q  Median      3Q     Max
-9.3113 -0.6187  0.2338  0.9065  2.9781

Coefficients:
                            Estimate        Std. Error t value             Pr(>|t|)
(Intercept)          27.46909869616345 13.51781177271251   2.032             0.042308 *
runtime               0.02102915396893  0.00208871475937  10.068 < 0.0000000000000002 ***
Rating                0.14825702410954  0.04364865698518   3.397             0.000699 ***
budget                0.00000000029386  0.00000000007555   3.889             0.000105 ***
Year                 -0.00773189766526  0.00670616786144  -1.153             0.249098
CertificateG          4.50089984880297  0.71807874047348   6.268       0.000000000467 ***
CertificateNC-17      1.19978440139536  0.96469177332772   1.244             0.213789
CertificateNot Rated -0.62143322432488  0.69840897889476  -0.890             0.373713
CertificatePG         4.00194095284554  0.65208305139142   6.137       0.000000001052 ***
CertificatePG-13      3.48052015770716  0.64508246798980   5.395       0.000000078369 ***
CertificateR          2.25889984384690  0.64477186853719   3.503             0.000472 ***
CertificateTV-14     -1.91015311266472  1.57214589987828  -1.215             0.224543
CertificateTV-MA     -0.21040096393359  0.83948965694638  -0.251             0.802132
CertificateUnrated   -1.80319199883364  1.19967651052732  -1.503             0.133015
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.433 on 1634 degrees of freedom
Multiple R-squared:  0.3288,    Adjusted R-squared:  0.3235
F-statistic: 61.58 on 13 and 1634 DF,  p-value: < 0.00000000000000022
```

Discussion

This study highlights important factors that influence a movie's worldwide gross revenue.Budget, runtime, ratings, and certification were found to have significant impacts, providing useful insights for filmmakers, producers, and marketers.

The analysis reveals that higher budgets generally lead to higher revenues, though simply increasing spending does not guarantee success. Strategic resource allocation, such as investing in high-quality production, casting, and marketing, is essential. Movies with runtimes between 90–120 minutes typically perform better, as audiences seem to favor this range. Longer movies, especially those exceeding 150 minutes, tend to show diminishing returns, emphasizing the importance of balancing runtime with audience preferences.

Higher IMDb ratings also positively affect box office performance, reinforcing the value of producing quality content that appeals to audiences and critics alike. Movies with better ratings often benefit from positive word-of-mouth and repeat viewership. Certification plays a crucial role as well, with G, PG, and PG-13 rated movies earning significantly higher revenues than R-rated or niche films. This indicates that family-friendly or broadly accessible content attracts larger audiences and is more likely to succeed financially.

While these findings are insightful, the model has some limitations. It explains only 32.88% of the variability in gross revenue, suggesting that other factors, such as marketing strategies, audience demographics, and cultural trends, also play an important role. Additionally, the year of release does not significantly affect revenue, indicating that factors like budget, ratings, and certification are more critical to a movie's success.

These results have practical implications for industry stakeholders. Producers can use this information to allocate budgets effectively and ensure their movies align with audience expectations. Marketers can focus on promoting high ratings and leveraging certifications with

broad appeal, such as PG or PG-13, to attract larger audiences. Moreover, family-friendly and blockbuster genres appear to have a consistent edge in generating higher revenues.

Future research could expand on these findings by including additional variables such as marketing expenditures, regional audience preferences, and long-term revenue streams like merchandising and streaming. This would provide a more comprehensive understanding of what drives box office success, helping stakeholders navigate the evolving entertainment industry with greater precision.

Implication

This study provides important insights for movie industry stakeholders, with practical implications not only in revealing key factors affecting movie gross, but also in suggesting potential strategic directions that can help producers, investors, and marketers make more forward-looking decisions. First, the study shows that there is a significant positive correlation between budget and box office, but this result also reveals room for optimizing resource allocation. The specific direction of allocation may be more critical to producers in budget planning than the total budget. Big-budget movies usually invest a large amount of money in star actors and special effects production, and the results of the study show that high IMDb ratings have a more significant impact on box office. Therefore, instead of focusing too much on the size of the production, more resources should be allocated to content development, such as improving the quality of the script or choosing a more artistically creative director, to ensure that the final product will receive a higher audience rating. In addition, producers need to strike a dynamic

balance between production and promotion, and explore more flexible budget allocation strategies to maximize return on investment.

The family-friendly ratings (G, PG, PG-13) in the study showed a significant boost to box office, but this further suggests how producers can better tailor their designs to audience needs through innovation. Family-friendly films can appeal to both child and adult audiences through "cross-adaptive" content design strategies. For example, by incorporating simple and interesting plots into the narrative to attract children, while providing enough depth to satisfy the aesthetic needs of adult viewers, the movie experience can be enhanced and the likelihood of repeat visits can be increased. Moreover, producers can pay more attention to the digital needs of family viewing in their content design, for example, by developing short film series or chapter narratives to adapt the fragmented watching behavior in modern audiences.

The results of this study also indirectly reveal a potential link between marketing and box office performance. Ratings have the greatest positive impact on box office, indicating that marketing teams can further enhance box office performance by reinforcing word-of-mouth communication on social media. For example, through big data analytics and artificial intelligence tools, marketing teams can predict audience behavior and develop precise promotional strategies to identify and target a movie's core audience. Additionally, the impact of grading on box office provides a clear direction for promotion. For example, family-friendly movies can expand their reach by partnering with schools or communities, while PG-13 movies targeting teenage audiences can use social media platforms such as TikTok to spread and attract more young viewers. A precise and personalized marketing strategy can help a film maximize its box office potential

References

Chen, X., Chen, Y., & Weinberg, C. B. (2013). Learning about movies: the impact of movie release types on the nationwide box office. *Journal of Cultural Economics, 37*(3), 359-386. https://doi.org/10.1007/s10824-012-9189-z

Hall, S. B., & Pasquini, S. (2020). *Can there be a happy ending for Hollywood after covid-19?*. World Economic Forum. https://www.weforum.org/stories/2020/07/impact-coronavirus-covid-19-hollywood-global-film-industry-movie-theatres/

Nelson, R. A., & Glotfelty, R. (2012). Movie stars and box office revenues: an empirical analysis. *Journal of Cultural Economics, 36*(2), 141-166. https://doi.org/10.1007/s10824-012-9159-5

Scutelnicu, G. (2023a, March 15). *Top 100 popular movies from 2003 to 2022 (imdb)*. Kaggle. https://www.kaggle.com/datasets/georgescutelnicu/top-100-popular-movies-from-2003-to-2022-imdb