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

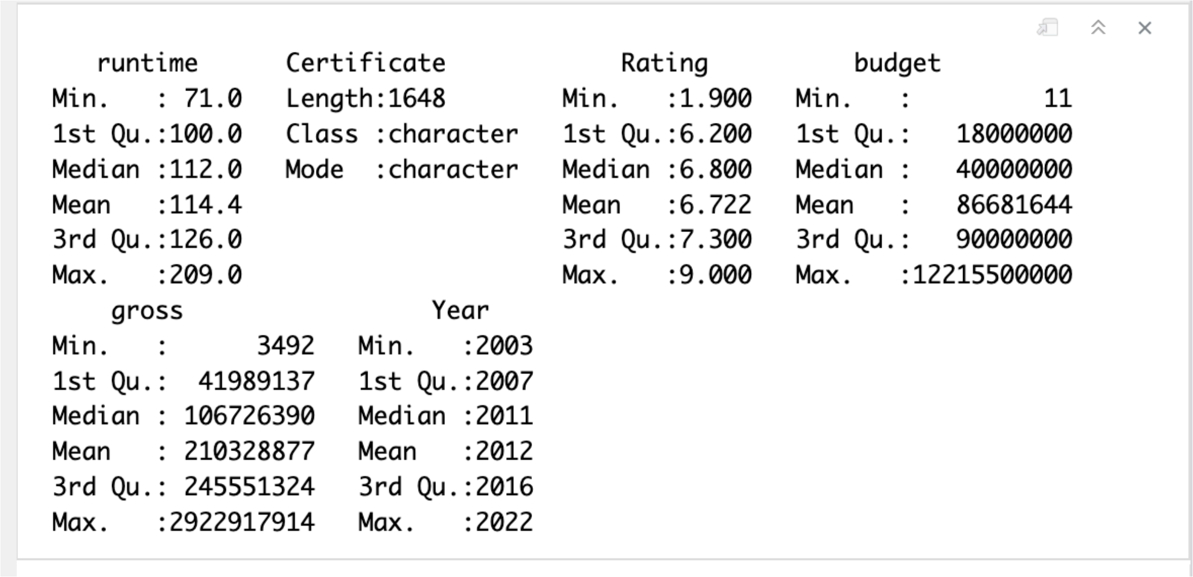
Descriptive Statistics

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.

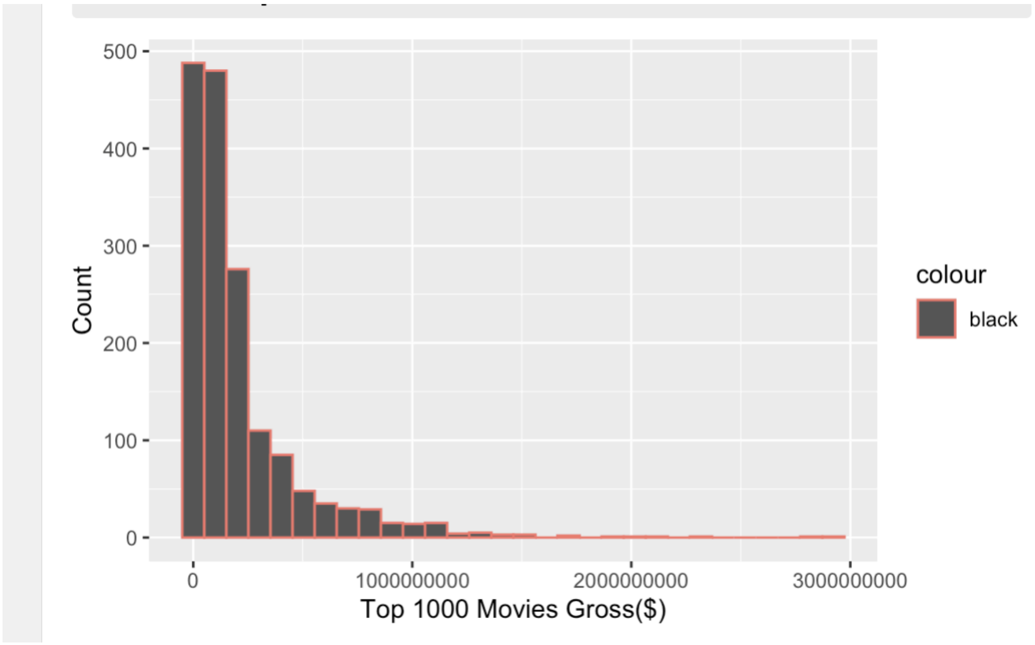


```
    runtime       Certificate           Rating          budget         
 Min.   : 71.0   Length:1648        Min.   :1.900   Min.   :          11  
 1st Qu.:100.0   Class :character   1st Qu.:6.200   1st Qu.:   18000000  
 Median :112.0   Mode  :character   Median :6.800   Median :   40000000  
 Mean   :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***

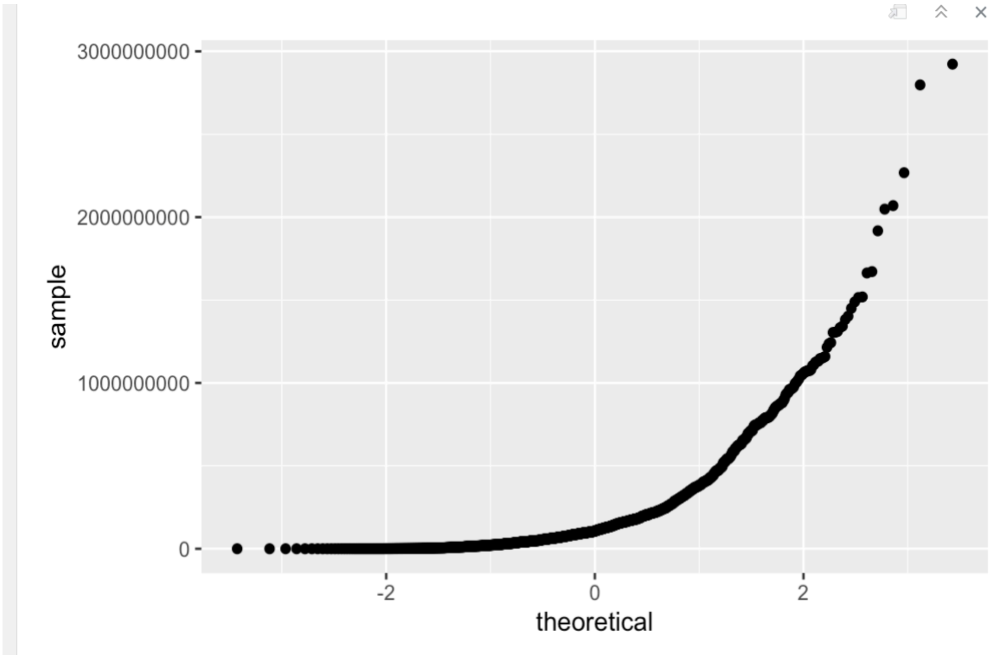
Shows the frequency distribution of box office receipts by plotting a histogram of the distribution of movie box office gross (**Figure1**). Helps us to understand the overall range of box office receipts, concentration trends, and possible outliers. We find that box office receipts show a clear right 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.

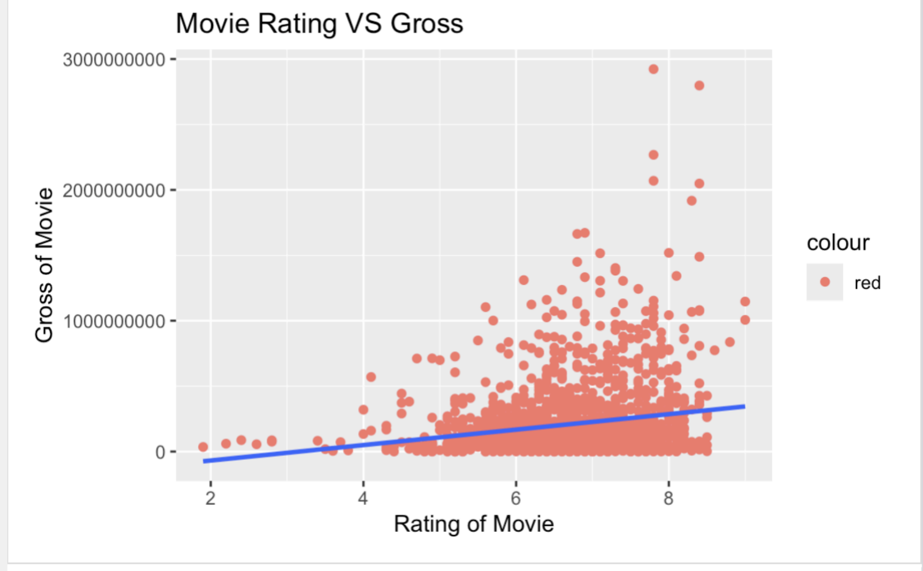
Figure 2 ***Q-Q Plot of the movies gross***



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.

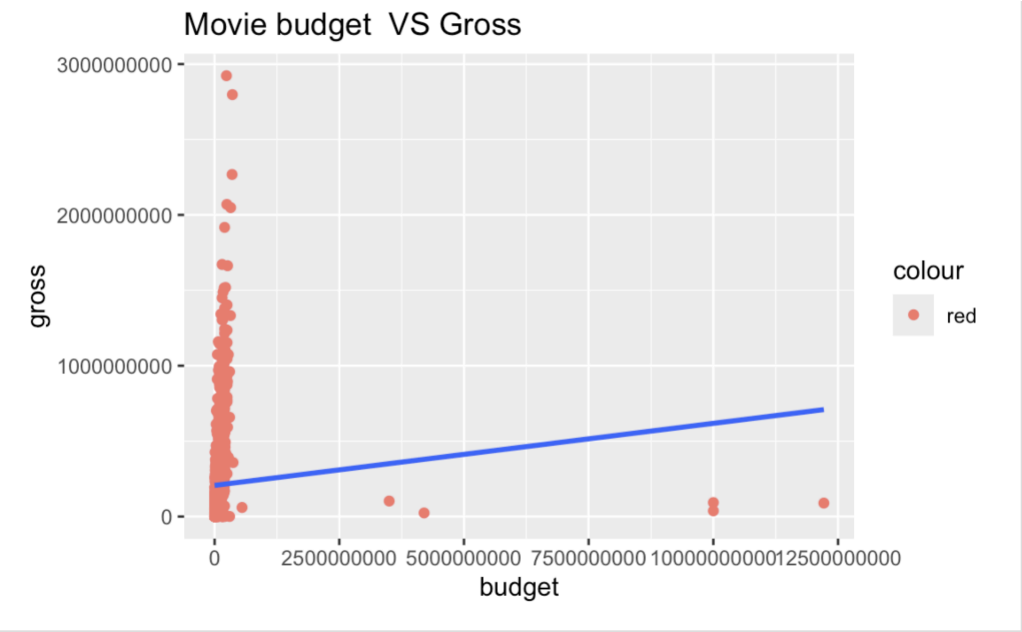
Figure 3 ***Relationship between Rating and Gross of Movies by plotting a scatter plot***



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***

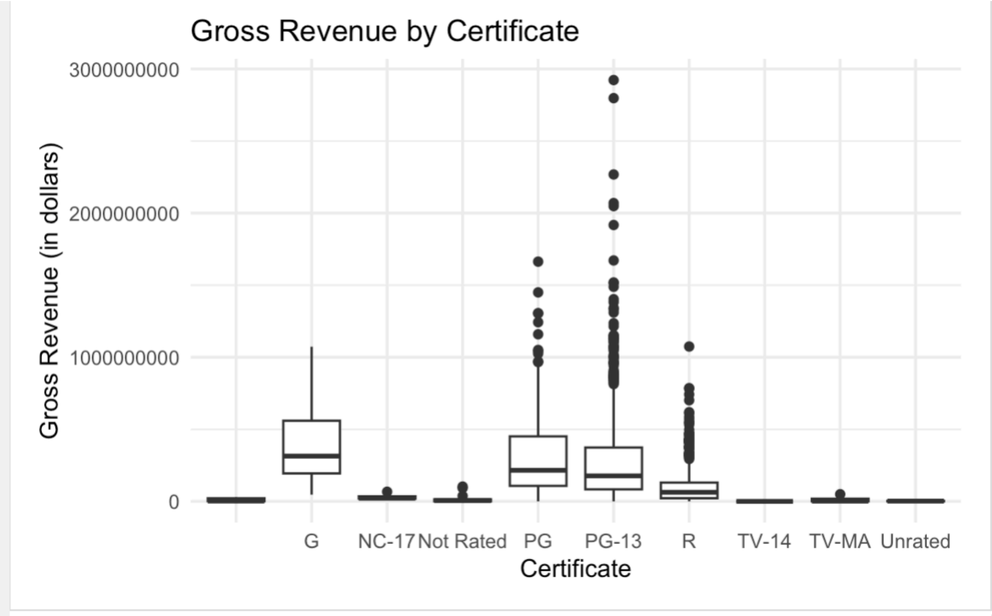


The distribution of Gross for different Certificates is shown by plotting a box-and-line diagram **(Figure5)**. The following are the characteristics and analysis that can be observed from the graph:

- 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.

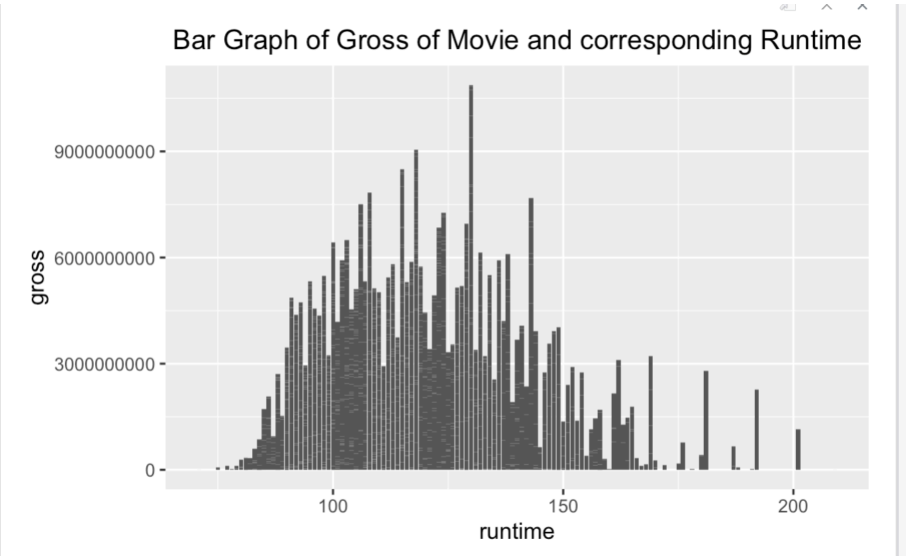
- Rated R (Restricted): The median is relatively low, but there are some outliers that indicate some R-rated movies are strong performers at the box office.
- NC-17, TV-MA, Unrated, etc: Generally low box office receipts and small distribution. Movies with these ratings are usually geared toward a very 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

Transformation

Results and Interpretations

Overall Model Performance

Figure() 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= $e^{\log(\text{gross})}$, where

$\log(\text{gross})=27.469+0.021\cdot\text{runtime}+0.148\cdot\text{Rating}+2.94\times10^{-10}\cdot\text{budget}-0.0077\cdot\text{Year}+\text{Certificate}$

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 "Year" and other 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 (**Pr(>|t|)**: < 0.001) 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

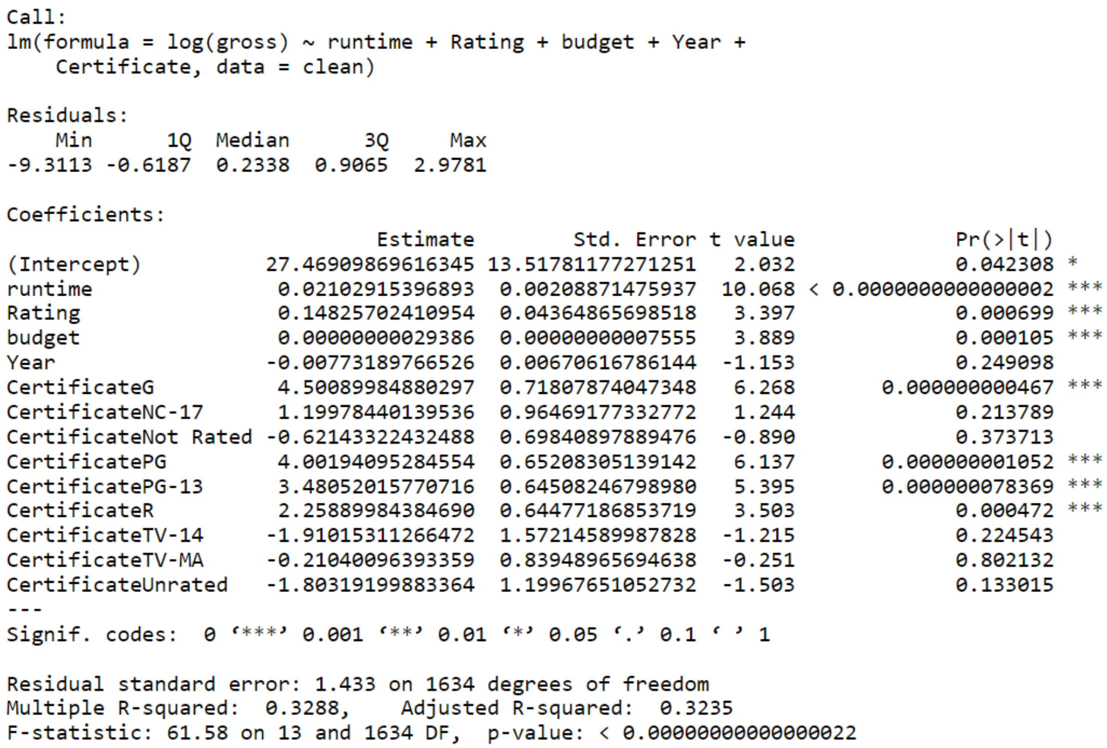
The estimated coefficient for Budget is 2.94×10^{-10}, meaning that for every additional dollar spent on the budget, the gross revenue is predicted to increase by 2.94% while holding other factors constant. P-value of 0.0001 indicates its statistical significance at the 5% level.

Regression Coefficients-Certificate

Movies rated "G" (General) are associated with a substantial increase in gross revenue, with a coefficient of 4.50, indicating a 450% 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 .



```
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

Implication

Work Cited

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