STA302: Method of Data Analysis I
Project Part 1: Proposal
October 10th, 2024

Contribution

Group 108:

Ba Minh Dang Le

- Help brainstorming and finding necessary datasets.
- Clean the dataset using Excel and implement the RMD file.
- Revise the final writing components and RMD file.

Dang Trung Kien Nguyen

- Help brainstorming and finding necessary datasets.
- Research related articles and summarize them in depth for future references.
- Revise the final writing components and RMD file.

Abdelrahman Alkhawas

- Help brainstorming and finding necessary datasets.
- Help clean the dataset using SQL for comparison.

Introduction

The financial success of a movie is crucial to decision-makers in the film industry due to the high costs and risks involved. Gross Revenue is a standard measure of financial success and represents the total income a film generates from box office sales, streaming services, and other distribution channels. This study aims to explore the key factors that influence a movie's gross revenue by answering the research question: **"To what extent can IMDb rating, Metascore, release year, movie genre, and production cost predict the gross revenue of movies?"**

The response variable of interest is "gross revenue", a clear and quantifiable measure that is crucial for decision-makers in the film industry. The underlying hypothesis is that these variables, either individually or collectively, can predict the financial success of a movie, as measured by its gross revenue. For example, higher IMDb ratings or larger production budgets might correlate with higher revenue, while genre-specific preferences could result in varied financial performance.

By identifying these predictors, this study aims to provide valuable insights that could help filmmakers make more informed decisions regarding resource allocation,

marketing, and production strategies, ultimately mitigating the financial risks associated with film production.

The chosen peer-reviewed articles show that the financial success of a movie can be influenced by various factors:

- The article ”**Dynamic Effects among Movie Ratings, Movie Revenues and Viewer Satisfaction**” suggests early box office success tends to generate positive feedback (Moon et al., 2010).
- The article “**Predicting Box Office Success: Do Critical Reviews Really Matter?**” concludes that although audience and critic opinions do not always align, positive reviews still enhance a movie's financial performance (Alec K., 2010).
- The article “**The determinants of box office performance in the Film Industry Revisited**” implies factors such as product cost, producer reputation, and awards nominations are key drivers while mentioning the declining influence of genre and critic reviews in predicting revenue (Pangarker et al., 2013).

We believe these studies highlight the relevance of different movie characteristics in predicting box office revenue, forming the basis for this research.

Multiple linear regression is appropriate for this analysis because the model can evaluate both numerical and categorical predictors. This approach quantifies the impact of each variable on gross revenue, offering a clear and interpretable model that highlights the key drivers of a movie's gross revenue.

Data Description

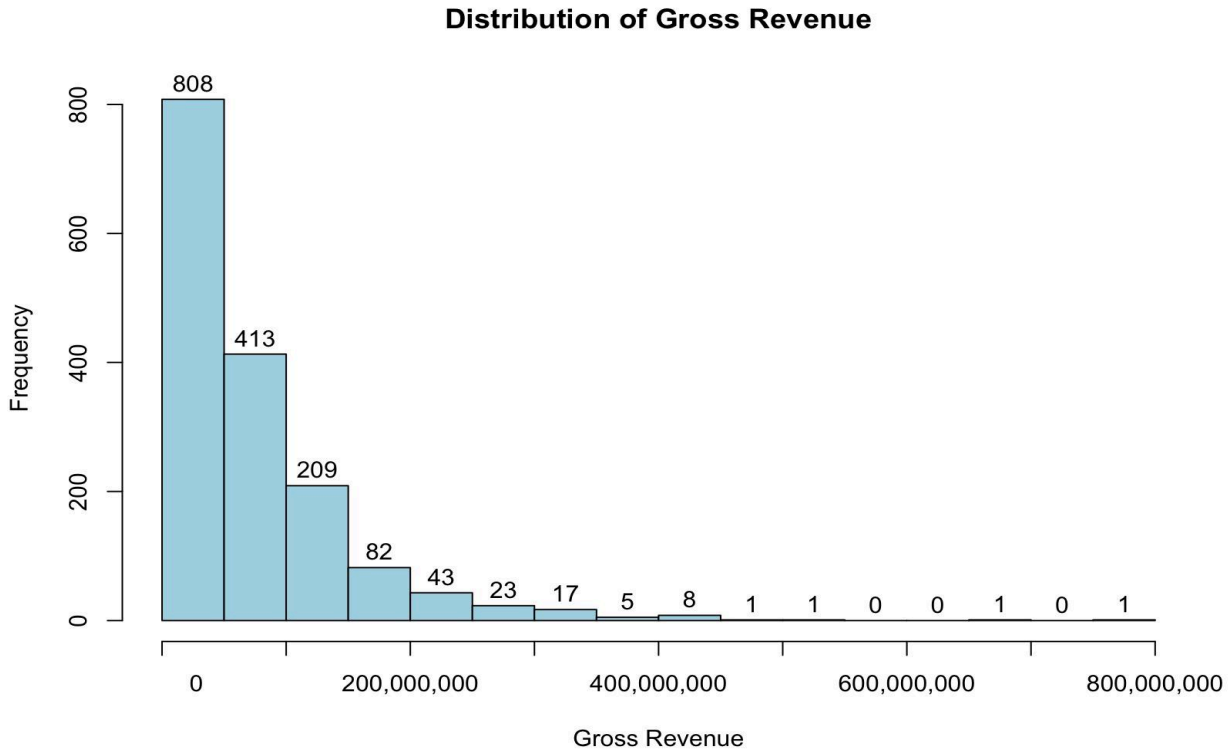
Description of data source

Both datasets are publicly available through Kaggle, a reputable platform for data scientists and machine learning practitioners. The “**IMDB dataset,**” a collection of different statistics from 2000 movies from IMDB database, created by Prisha Sawhney while the “**movie dataset**” was created by Rounak Banik. More information can be found under the bibliography.

Our hypothesis suggests that production costs could be an important extra potential predictor by which we merged the production costs from the “**movie dataset**” to the “**IMDB dataset**” by matching movie names. More detail can be found under the Rmd file.

Response variable summary

The response variable for this regression analysis is Gross Revenue (in dollars). The following statistical summary and histogram are generated using R:



Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum	Standard Deviation
10000	22857500	49675000	71718319	95422500	760510000	75294763

Gross revenue has a right-skewed distribution, showing that it is highly concentrated at lower values. As the revenue increases the number of movies decreases significantly, which is common in the film industry.

Gross revenue is an ideal response variable for the regression model because it is continuous and quantifiable. Further transformation will be implemented if needed.

Predictor variable summaries

The study hypothesizes that various predictors could be related to gross revenue, including:

- IMDB rating (a measure of audience perception)
- Metascore (critical acclaim)
- Release year (capturing changes in market conditions or audience preferences over time)
- Movie genre (a categorical predictor, reflecting genre-based audience preferences)
- Production cost (which directly correlate with the scale of marketing, distribution, and overall production quality)

Statistical Results:

IMDB Rating:

```
> summary(IMDB_CLEANED_DANG$IMDBRating)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.400   6.400   7.000   6.911   7.525   9.300
> sd(IMDB_CLEANED_DANG$IMDBRating)
[1] 0.9077944
```

Metascore:

```
> summary(IMDB_CLEANED_DANG$Metascore)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   9.00   47.00   61.00   60.44   73.00  100.00
> sd(IMDB_CLEANED_DANG$Metascore)
[1] 17.77703
>
```

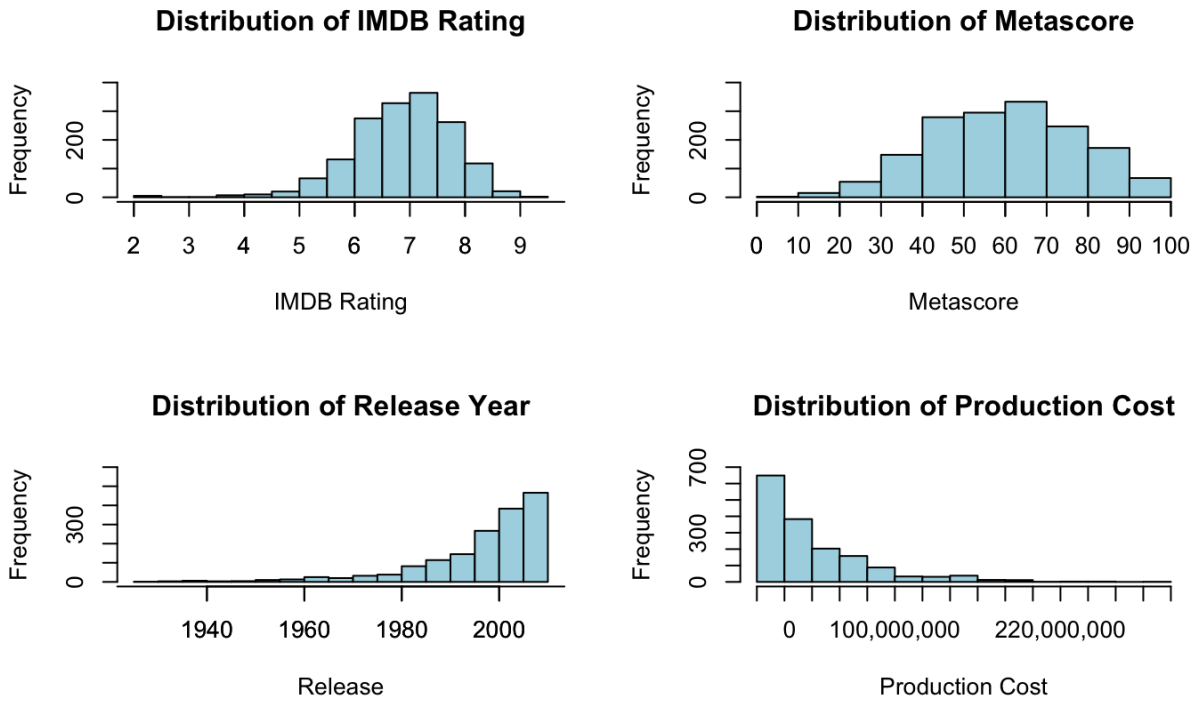
Release Year:

```
> summary(IMDB_CLEANED_DANG$ReleaseYear)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1927    1993    2001    1997    2006    2010
> sd(IMDB_CLEANED_DANG$ReleaseYear)
[1] 12.99977
>
```

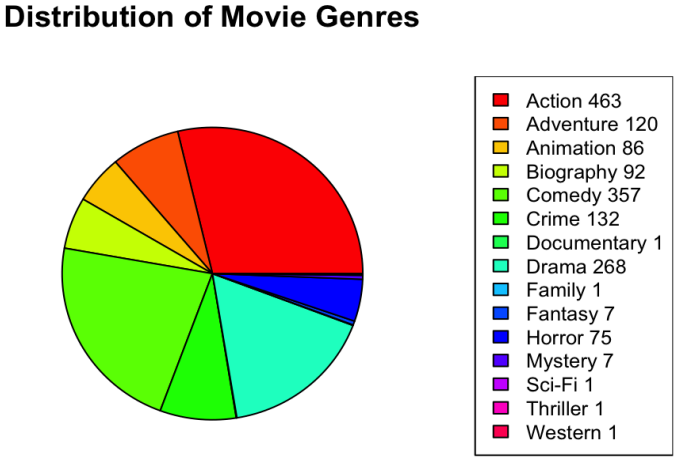
Production Cost:

```
> summary(IMDB_CLEANED_DANG$Budget)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
     6000  12000000  28000000  41644913  60000000 300000000
> sd(IMDB_CLEANED_DANG$Budget)
[1] 41616958
>
```

Histogram of predictor variables:



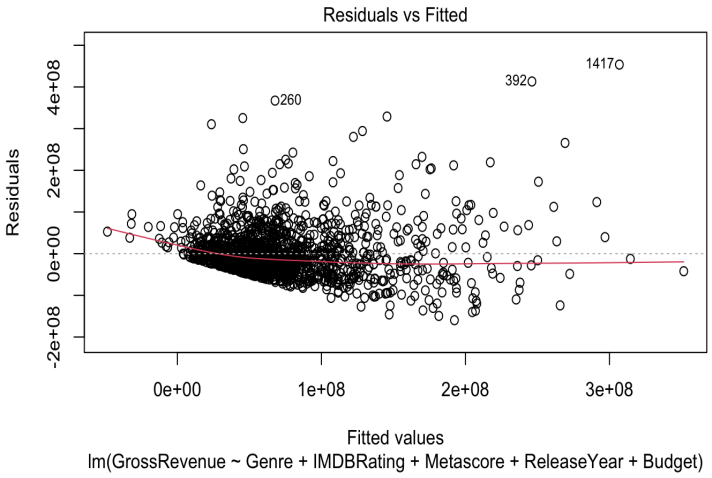
The histogram shows that both IMDB rating and Metascore have an approximate normal distribution while production cost has a right-skewed distribution, reflecting hesitancy in high-cost movie productions. Meanwhile, the Release year shows a left-tail distribution indicating the increasing trends in movie making over time.

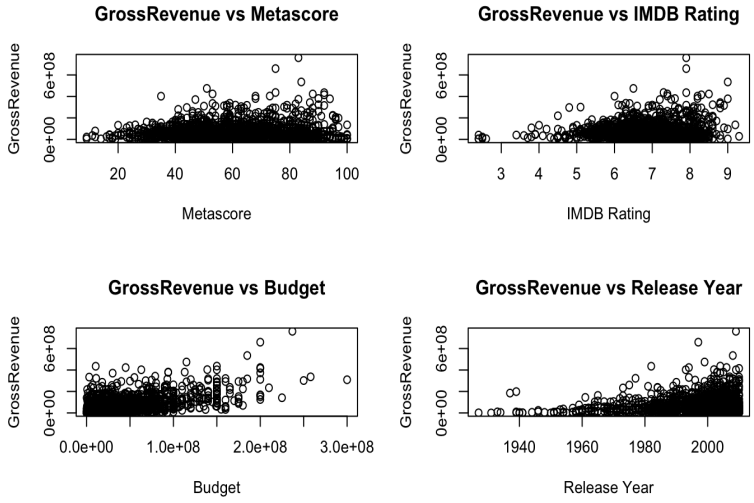


The Pie Chart shows the most popular genres which might be useful insights in predicting consumer preferences. Overall, genre, IMDB rating, Metascore, releasing year, and production cost are important predictors as they reflect audience perception and the scale of movies.

Preliminary Model Results Section

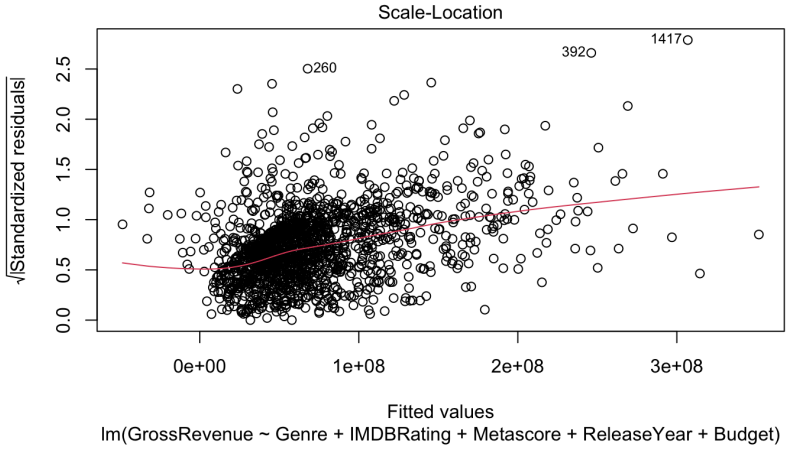
Residual analysis of the preliminary model





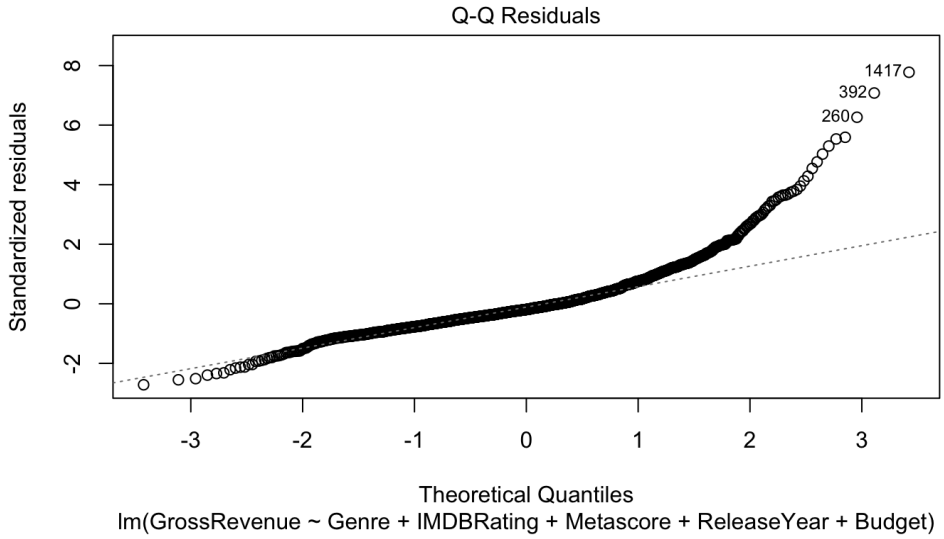
From the residual plot, we can observe a dispersed pattern of residuals, which implies a violation of linearity. Despite the evidence against a linear relationship, the scatter plots suggest that we can perform a transformation on predictor variables to improve the fitness of our model.

Constant Variance, Homoscedasticity Assumption



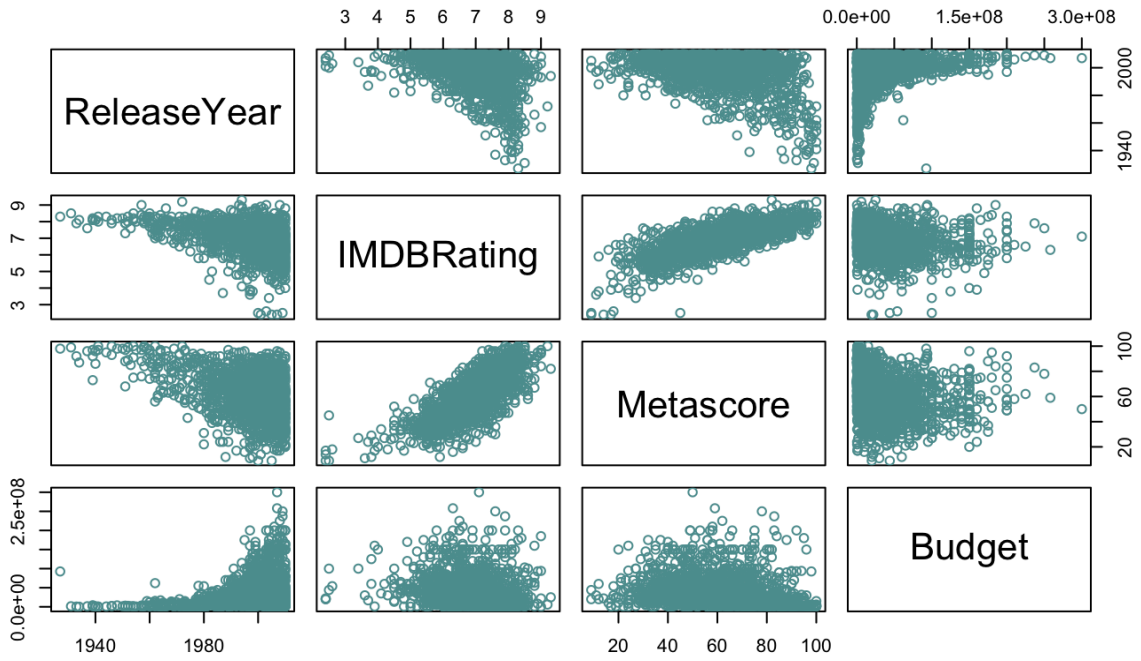
The Scale Location plot suggests a violation of the homoscedasticity assumption due to upward upward-curved smoothed trend line (red line). The data points seem more dispersed at higher fitted values. Such patterns indicate potential heteroscedasticity in the model and imply that the model needs adjustments, such as transforming variables.

Normality of Errors



Based on the QQ plot, the residual at the upper tail shows considerable deviation from the fitted line. This provides evidence against the Normality of Error assumption and implies that the model needs further adjustment such as transforming the response variable.

Multicollinearity Check:



The scatterplot matrix of predictor variables suggests strong evidence to support non-multicollinearity between chosen variables. The only potential correlations are the following pair of variables: IMDB vs Metascore and Release Year vs Budget, as the scatter plots show a curved trend but the spread in data points is wide and thus not too significant. A further predictor consideration will be conducted upon receiving feedback.

Preliminary model discussion

Evaluating the Multiple R-squared, Residual Standard Error, and F-statistics will tell us how well the model performed. The regression model has an R-squared value of 0.3938, which implies that the model is decent and can explain 39.38 percent of the variability of gross revenue. The F-statistic has a p-value of less than 0.05 significant level suggests that the model is statistically significant overall. The positive coefficient and low p-value for IMDB rating and Metascore suggest that the effect of the mentioned variables aligns with expectations from previous research. In addition, the Residual standard error might seem large itself but its interpretation should be relative to the scale of the gross revenue.

Bibliography

Original Datasets:

Sawhney, P. (n.d.). *IMDB Dataset*. Kaggle. Retrieved October 1, 2024, from https://www.kaggle.com/datasets/prishasawhney/imdb-dataset-top-2000-movies
Banik, R. (n.d.). *The Movie Dataset*. Kaggle. Retrieved October 1, 2024, from https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset/data?select=movies_metadata.csv

Related academic paper:

Pangarker, N. A., & Smit, E. v. d. M. (2013). The determinants of box office performance in the Film Industry Revisited. *South African Journal of Business Management*, *44*(3), 47–58. https://doi.org/10.4102/sajbm.v44i3.162

Moon, S., Bergey, P. K., & Iacobucci, D. (2010). Dynamic Effects among Movie Ratings, Movie Revenues, and Viewer Satisfaction. *Journal of Marketing*, *74*(1), 108–121. http://www.jstor.org/stable/20619083

Alec, Kenedy. (2010). Predicting Box Office Success: Do Critical Reviews Really Matter? University of California at Berkeley. https://www.stat.berkeley.edu/~aldous/157/Old_Projects/kennedy.pdf