Introduction

The IMDB dataset, obtained from Kaggle, includes information about the top 1000 movies and TV shows. It contains various details such as the link to the poster image used by IMDB (Poster_Link), the title of the movie or TV show (Series_Title), and the year of release (Released_Year). It also includes the certification rating (Certificate), total runtime in minutes (Runtime), genre(s) (Genre), and the rating on the IMDB website (IMDB_Rating). Additionally, the dataset provides a brief summary or mini story (Overview), the score from Metacritic (Meta_score), the name of the director (Director), and the names of the main stars (Star1, Star2, Star3, Star4). Furthermore, it includes the total number of votes received on IMDB (No_of_votes) and the amount of money earned by the movie (Gross). This dataset offers a comprehensive overview of the most popular movies and TV shows, including key details about their release, rating, and cast.

We came to a decision regarding our problem after comparing the performance indicators of two movies; the Shawshank Redemption and Captain Marvel. For all intents and purposes, the movies were identical in all factors except for two: IMDb rating and gross revenue. Whereas the Shawshank Redemption scored a 9.3 on IMDb and grossed $73.3m, Captain Marvel scored a 6.8 on IMDb but grossed over $1.131 billion.

At their core, movies are the products of business, and the fundamental goal of every business is to turn a profit. As such, we began to inquire on the counterintuitive relationship that seemed to exist between IMDb rating and gross revenue. If the Shawshank Redemption were a significantly better movie according to meta-critics, how is it possible that it grossed less than a tenth of what the significantly worse rated Captain Marvel did? If a movie doesn't earn money for being good, then what determines its gross revenue? The question presented itself: what makes a movie successful? Is it possible, through data analysis, to quantify the fiscal success of a movie accurately enough to build a model that can predict the box office success of other movies? Our real problem was therefore centered around the movies' gross revenues and comparing them against "determining" factors/variables.

Data Clean Up

During the data cleaning process, we took several steps to ensure the dataset was ready for analysis.

1. **Identifying missing values:**
   We checked for missing values in the dataset and found that several columns had them.
   All rows had at least one missing value.
2. **Handling missing values:**
   For columns like Gross and Meta_score, we filled in missing values with the median of each column. This helps maintain the data's integrity.
   For the Certificate column, which contains categorical data, we replaced missing values with the mode (most frequent value). This ensures consistency in the data.
3. **Converting categorical data**
   We converted the Certificate column into binary variables to make it easier to analyze.
4. **Correcting data errors**
   We corrected an error in the Released_Year column where "PG" was mistakenly listed, replacing it with "1970".
5. **Feature engineering**
   We created a new column, Runtime_decimal, to convert the total runtime from minutes to decimal hours. This makes it easier to understand.
6. **Dropping unnecessary columns**
   We removed the Overview column as it wasn't needed for our analysis.
7. **Ensuring data type consistency**
   We made sure the Released_Year column had a consistent numeric data type.

Association Analysis

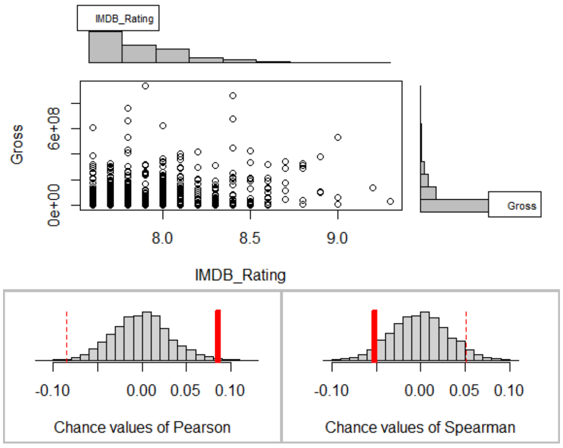
In this section, we explored the associations between the variable of interest, Gross (y-variable), and other related variables in the dataset.

Variables Selected for Analysis
y-Variable: Gross (Money earned by the movie)
(Rating of the movie at IMDB site)
x2: Meta_score (Score earned by the movie)
x3: Runtime_decimal (Total runtime of the movie in decimal hours)
x4: Released_Year (Year at which the movie was released)
x5: No_of_Votes (Total number of votes)
Analysis Methodology

Analysis Methodology

- We conducted association analysis between Gross and each of the selected variables using the associate() function.
- The associations were visualized and statistical tests were performed to determine their significance.
- Due to the presence of heteroscedasticity and many outliers in the data, Pearson's correlation coefficient was not appropriate for assessing statistical significance. Instead, Spearman's rank correlation coefficient was used.

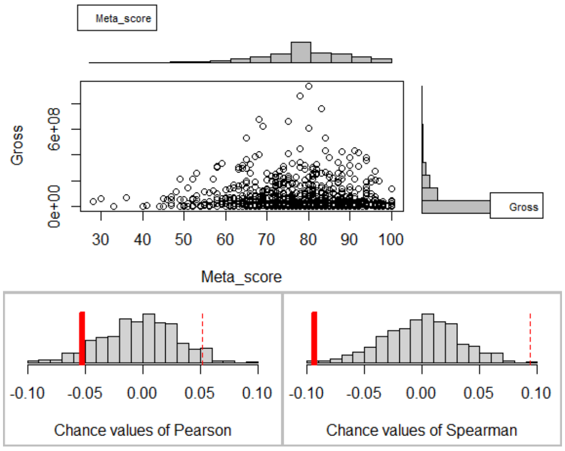
Results



- x1: IMDB_Rating

Association between IMDB_Rating (numerical) and Gross (numerical)
using 1000 complete cases
Permutation procedure:
Value Estimated p-value
Spearman's rank correlation -0.05118562
0.10525
With 4000 permutations, we are 95% confident that:
the p-value of Spearman's rank correlation is
between 0.096 and 0.115

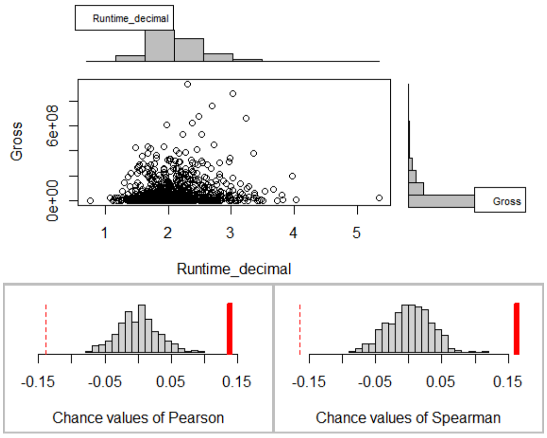
The association with Gross was inconclusive using Spearman's correlation.



- x2: Meta_score

Association between Meta_score (numerical) and Gross (numerical)
using 1000 complete cases
Permutation procedure:
Value Estimated p-value
Spearman's rank correlation -0.09343756
0.004
With 500 permutations, we are 95% confident that:
the p-value of Spearman's rank correlation is
between 0 and 0.014

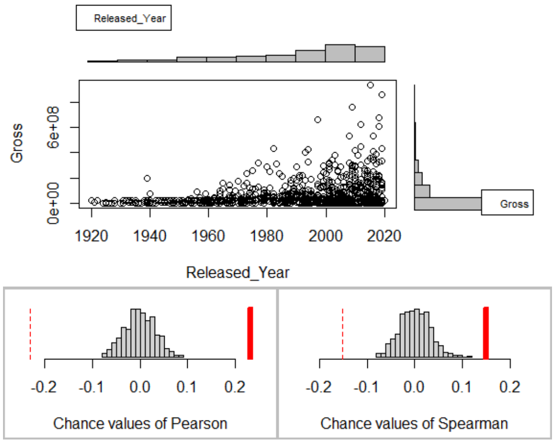
The association with Gross was statistically significant using Spearman's correlation.



- x3: Runtime_decimal:

Association between Runtime_decimal (numerical) and Gross (numerical)
using 1000 complete cases
Permutation procedure:
Value Estimated p-value
Spearman's rank correlation 0.1636677 0
With 500 permutations, we are 95% confident that:
the p-value of Spearman's rank correlation is between 0 and 0.007

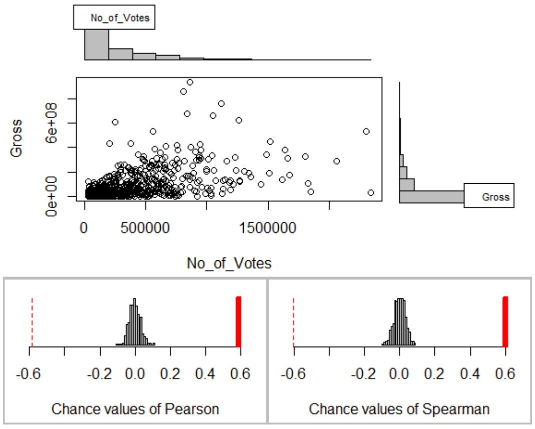
The association with Gross was statistically significant using Spearman's correlation.



- x4: Released_Year

Association between Released_Year (numerical) and Gross (numerical)
using 1000 complete cases
Permutation procedure:
Value Estimated p-value
Spearman's rank correlation 0.1507193 0
With 500 permutations, we are 95% confident that:
the p-value of Spearman's rank correlation is between 0 and 0.007

The association with Gross was statistically significant using Spearman's correlation.



- x5: No_of_Votes

Association between No_of_Votes (numerical) and Gross (numerical)
using 1000 complete cases
Permutation procedure:
Value Estimated p-value
Spearman's rank correlation 0.6034343 0
With 500 permutations, we are 95% confident that:
the p-value of Spearman's rank correlation is between 0 and 0.007

The association with Gross was statistically significant using Spearman's correlation.

Regression Analysis

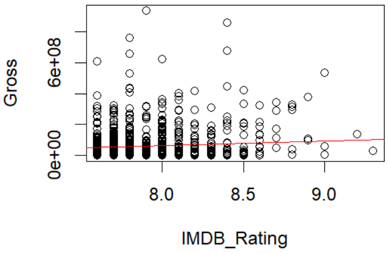
In this section, we conducted regression analysis to predict the y-variable, Gross (Money earned by the movie), using various predictor variables.

Regression models

- Model 1: Gross vs IMDB Rating

  Coefficients:

  Intercept: The estimated intercept is -189,751,053. This represents the predicted Gross when the IMDB Rating is zero. However, this value may not have practical significance as IMDB Ratings typically do not reach zero.

  IMDB_Rating: The estimated coefficient for IMDB Rating is 31,482,602. This indicates that for every one-unit increase in IMDB Rating, the Gross is expected to increase by approximately $31,482,602.

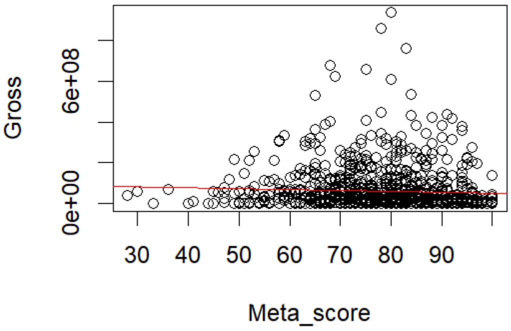
  Statistical Significance:
  Both the Intercept and IMDB_Rating coefficients are statistically significant, as indicated by their p-values (Intercept: p = 0.04017, IMDB_Rating: p = 0.00681). The significance level (alpha) is typically set at 0.05. Therefore, the p-values for both coefficients are indicating a statistically significant relationship between Gross and IMDB Rating.

  Confidence Intervals:
  The 95% confidence interval for the Intercept ranges from -370,975,804 to -8,526,302.
  The 95% confidence interval for the IMDB_Rating coefficient ranges from 8,698,693 to 54,266,512.

- Model 2: Gross vs Meta_score

  Coefficients:

  Intercept: The estimated intercept is 96,682,515. This represents the predicted Gross when the Meta_score is zero.

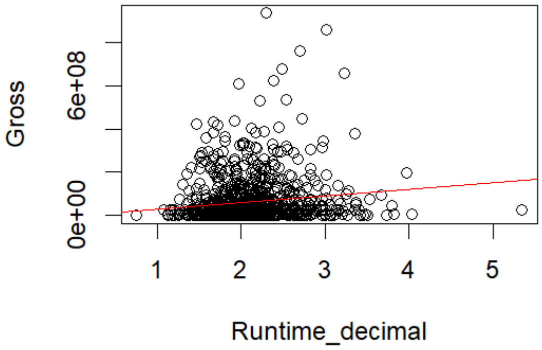
  Meta_score: The estimated coefficient for Meta_score is -462,915. This indicates that for every one-unit increase in Meta_score, the Gross is expected to decrease by approximately $462,915. However, this coefficient is not statistically significant.

Statistical Significance:
The Intercept coefficient is statistically significant (p < 0.001), indicating a significant relationship between the Intercept and Gross.
However, the Meta_score coefficient is not statistically significant (p = 0.101). This suggests that Meta_score may not have a significant impact on Gross earnings.

Confidence Intervals:
The 95% confidence interval for the Intercept ranges from $52,987,615 to $140,377,414.45.
The 95% confidence interval for the Meta_score coefficient ranges from -$1,016,331 to $90,501.35.

- Model 3: Gross vs Runtime_decimal
  Coefficients:

  Intercept: The estimated intercept is -963,983. This represents the predicted Gross when the Runtime_decimal is zero. However, this value may not have practical significance as Runtime_decimal typically does not reach zero.



  Runtime_decimal: The estimated coefficient for Runtime_decimal is 30,016,298. This indicates that for every one-unit increase in Runtime_decimal (one hour increase in runtime), the Gross is expected to increase by approximately $30,016,298. This coefficient is statistically significant.

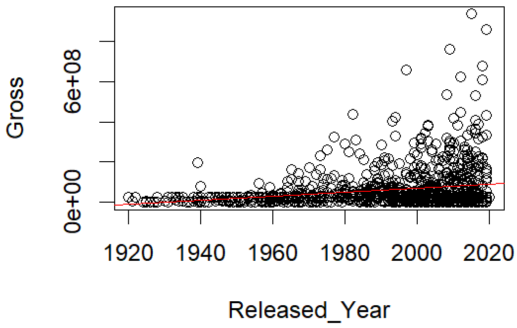
  Statistical Significance:
  The Intercept coefficient is not statistically significant (p = 0.946), indicating that the Intercept may not have a significant impact on Gross earnings.
  However, the Runtime_decimal coefficient is statistically significant (p < 0.001), suggesting that Runtime_decimal is a significant predictor of Gross earnings.

  Confidence Intervals:
  The 95% confidence interval for the Intercept ranges from -$28,957,596 to $27,029,630.
  The 95% confidence interval for the Runtime_decimal coefficient ranges from $16,691,912 to $43,340,684.

- Model 4: Gross vs Released_Year

  Coefficients:

  Intercept: The estimated intercept is -1.945e+09. This represents the predicted Gross when the Released_Year is zero. However, this value may not have practical significance as Released_Year typically does not reach zero.

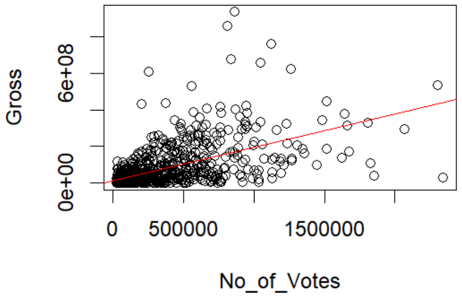
  

  Released_Year: The estimated coefficient for Released_Year is 1.007e+06. This indicates that for every one-year increase in Released_Year, the Gross is expected to increase by approximately $1,007,000. This coefficient is statistically significant.

  Statistical Significance:
  Both the Intercept and Released_Year coefficients are statistically significant (p < 0.001), indicating significant relationships with Gross earnings.

  Confidence Intervals:
  The 95% confidence interval for the Intercept ranges from -$2.468 billion to -$1.420 billion.
  The 95% confidence interval for the Released_Year coefficient ranges from $743,849.5 to $1,270,059.

- Model 5: Gross vs No_of_Votes
  Coefficients:
  Intercept: The estimated intercept is 10,620,000. This represents the predicted Gross when the No_of_Votes is zero.
  No_of_Votes: The estimated coefficient for No_of_Votes is 182.3. This indicates that for every one-unit increase in No_of_Votes, the Gross is expected to increase by approximately $182.3. This coefficient is statistically significant.

  Statistical Significance:
  Both the Intercept and No_of_Votes coefficients are statistically significant (p < 0.001), indicating significant relationships with Gross earnings.
  Confidence Intervals:
  The 95% confidence interval for the Intercept ranges from $3,983,807.23 to $17,258,620.
  The 95% confidence interval for the No_of_Votes coefficient ranges from 166.7339 to 197.8528.

Based on the regression analysis conducted on the dataset, it appears that the model with the lowest Root Mean Squared Error (RMSE) would be considered the better model, as it indicates the smallest average deviation between the observed and predicted values of Gross (Money earned by the movie). From our dataset, it's evident that the model with Gross vs No_of_Votes has the lowest RMSE compared to the others.

Since the Gross vs No_of_Votes model has the lowest RMSE, it suggests that this model provides the best fit for predicting Gross based on the variables considered in the analysis. Therefore, it can be concluded that the Gross vs No_of_Votes model is the better model among the ones evaluated.

Conclusion

Although many of our findings were conclusive and statistically significant, our analysis would suggest that the answer to our question of what makes a movie successful(?) is that **no one variable/factor can accurately determine or predict the box office success of a movie**.

Our analysis left a lot to be desired in terms of tangible relationships between variables with very loose correlations between our main y-variable and all five of our x-variables. The only notable correlation that existed in our dataset was that between No_of_votes and gross revenue - unfortunately, however, the relationship between the number of votes (on IMDb) a movie garners and the box office success are too intuitively related to be indicative of anything unknown or shed light on our real problem. The no_of_votes statistic can virtually be interpreted as “no_of_ticket_sales” or “no_of_viewers”, in which case it is not a revelation to suggest that the more tickets a movie sells the more money it will make (gross revenue).

Interpreting our analysis in the context of our dataset reveals that the answer to our question is actually that **we can not accurately determine or predict a movie’s box office success <u>with the given variables and observations.</u>** Operating on the principle that a movie is the product of a business, the determining factor(s) for a product’s fiscal success are more likely to be found in correlations drawn between income (gross revenue) and **advertising budget**, **production budget**, **marketing reach** etc. A product sells not only because it is a quality product but because it is perceived to be a quality product, has significant investments in its launch, and has a significant marketing presence and social outreach. To that end, it would make sense if movies like the Shawshank Redemption made less gross revenue than movies like Captain Marvel if they weren't as marketed or didn't have as big of a production team/budget despite the difference in their IMDb rating.

Finally, there are tangible and intangible variables that must be considered and are likely to influence the differing performance of movies over time. These will include but are not limited to; *the continued growth of the world economy* (both in terms of purchasing power and inflation), *the continued growth of the population*, *and the continued expansion of social media in the evolution of advertising and marketing.* Although these are variables/data points we did not have immediate access to, they are likely to have notable correlations in explaining why movies like Captain Marvel can make ten times what movies like the Shawshank Redemption make. It is not necessarily that Captain Marvel does anything better than the Shawshank Redemption, but that movies are making more money as a result of the continued development of the industry, growth of the population and inflation (and development) of the economy.