Blockbuster Insights: Predicting Movie Success Metrics

by

Nandini Ramakrishnan - 862465172
Pious Khemka - 862467916
Pranay Penikalapati - 862466092
Vishesh Shukla - 862464003

MGT 256: Business Analytics for Management
Mr. Adem Orsdemir
Classroom: SSC 308
December 04, 2023

1. Abstract:

The "Blockbuster Insights" project employs data science methodologies to predict crucial success metrics for movies, encompassing features, audience engagement, and critical reception. Through categorical-to-numerical transformation and exploratory data analysis, the project unveils patterns and relationships, setting the stage for robust predictive modeling. Regression analysis identifies influential factors impacting box office revenue, IMDb ratings, and audience reception. Utilizing K-Nearest Neighbors, the project further segments movies into distinct groups based on specific characteristics. Expected outcomes include precise predictive models, and identification of unique movie segments. This research stands to guide decision-making in the film industry, offering actionable insights for stakeholders and contributing to the optimization of future movie releases. In the intersection of art and science, "Blockbuster Insights" unravels the intricacies that shape the success of movies in the dynamic landscape of entertainment.

2.Introduction:

In today's ever-evolving film industry, making informed decisions stands as the linchpin for success. The "Blockbuster Insights" project emerges as a pivotal exploration, delving deep into the realm of movies through the lens of advanced data science techniques. With the relentless growth of data-driven decision-making, this project aims to predict critical success metrics for movies, a pursuit essential for filmmakers, studios, and industry stakeholders.

At its core, this project seeks to decode the complex factors influencing the success of movies—specifically, their critical acclaim and box-office performance. By leveraging a comprehensive dataset spanning diverse facets of the film industry, this endeavor sets the stage for predictive modeling, paving the way to uncover the key determinants that impact box office

revenue, IMDb ratings, and audience reception.

The overarching goal is to furnish actionable insights and recommendations derived from meticulous regression analyses and machine learning models. Through precision-driven predictions of IMDb scores, this project aims to illuminate the drivers shaping audience perceptions. Moreover, employing segmentation analysis via K-Nearest Neighbors unveils distinct movie clusters, unraveling the multifaceted landscape of the film industry.

This research harbors immense potential, poised to revolutionize decision-making processes within the entertainment domain. By providing invaluable insights and strategic guidance for future movie releases, "Blockbuster Insights" endeavors to bridge the gap between creativity and analytics, offering a roadmap to decipher the intricate patterns underlying movie success in today's dynamic cinematic landscape.

3.Objective:

The "Blockbuster Insights: Predicting Movie Success Metrics" project is designed with two primary objectives, each aimed at deciphering key elements in the film industry:

Prediction of Critical Success Metrics: This objective entails employing advanced data science methodologies to forecast crucial determinants contributing to a movie's critical success. By analyzing a diverse dataset encompassing various facets of the film domain, the project aims to unravel the factors influencing critical acclaim, as measured by IMDb ratings and audience reception. Through regression analysis, the endeavor seeks to identify influential variables that significantly impact audience perceptions and contribute to a movie's critical acclaim.

Prediction of Gross Earnings: The second objective revolves around predicting a movie's box office success. Utilizing comprehensive data analytics and predictive modeling techniques, the project aims to uncover the underlying factors driving box office revenue. By delving into the

dataset's rich information, the study seeks to elucidate the pivotal elements that correlate with a movie's financial success. Regression analyses and segmentation analyses will be leveraged to provide insights into the dynamics that contribute to a movie's commercial triumph.

These objectives stand as pillars guiding the comprehensive analysis within the project, aiming to provide actionable insights and strategic recommendations for filmmakers, studios, and industry stakeholders. Through these endeavors, "Blockbuster Insights" endeavors to bridge the gap between data science and the film industry, offering invaluable insights to enhance decision-making processes and optimize future movie releases.

4.Methodology

a. **Data Collection from IMDB via Kaggle:** In pursuit of comprehensive and reliable data, our research led us to a dataset on Kaggle, a renowned platform for data science collaboration. This dataset comprised of 5000 best-rated movies of all time from IMDB. The dataset had one file in csv format with around 29 columns.

b. **Initial Exploratory Data Analysis (EDA - Pre-Data Cleaning):** An exploratory analysis of the dataset was performed to understand outliers, distributional trends, and a summary of the 5000 records.

c. **Data Cleaning Process:** Missing numbers and discrepancies in the dataset were addressed. Categorical variables were converted into numerical representations suitable for analysis. Random sample techniques were used to reduce the dataset, dropping it to 1000 values for better model fitting.

d. **Exploratory Data Analysis (EDA - Post Data Cleaning):** Following data cleaning, a second step of exploratory analysis was performed to analyze patterns, trends, and outliers within the improved dataset. Relationships between variables were visualized to

find potential predictors connected with movie success metrics.

e. Regression-Analysis:

Regression analytic approaches were used to understand the relationships between predictor variables and movie performance measures such box office revenue, IMDb ratings, and audience reception. Identified and assessed major indicators that have a significant impact on movie success measures.

f. Hypothesis-Analysis:

Conducted a hypothesis analysis to investigate the potential relationship between IMDb ratings and gross earnings. Explored if there exists a correlation or impact between these two crucial success metrics.

g. Segmentation Analysis using K-Nearest Neighbors (KNN):

For dataset segmentation and prediction, the K-Nearest Neighbors (KNN) technique was used. Investigated separate clusters of movies based on specific features to establish distinct groupings and their distinguishing characteristics. KNN was used to create a predictive model for predicting movie imdb ratings.

These methodological steps aimed to prepare and analyze the dataset comprehensively, facilitating the extraction of meaningful insight.

5.Data Analysis:

In this section, we are going to preprocess the data and want to know the trends of the different variables before preprocessing and after preprocessing the data. Histograms:

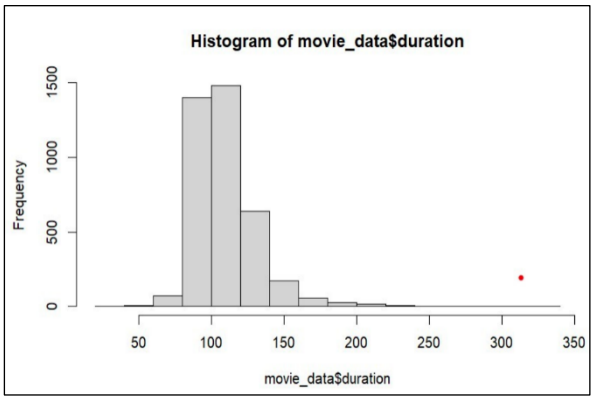


Fig5.1

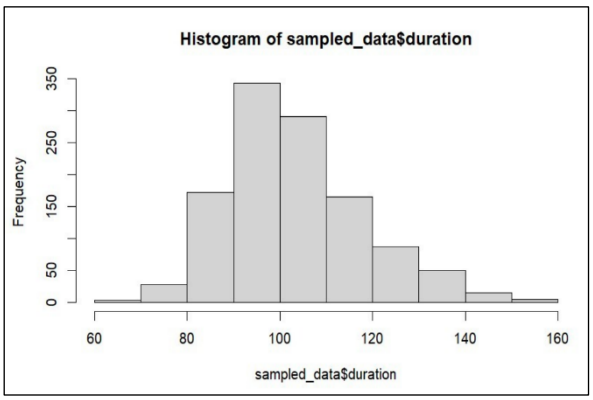


Fig 5.2

Fig 5.1 shows the distribution of movie duration from the original dataset with 5000 records. There are many outliers in the dataset. Fig 5.2 shows the same distribution after data cleaning. We observe that the movie duration ranges between 80 mins to 120 mins for most of the movies.

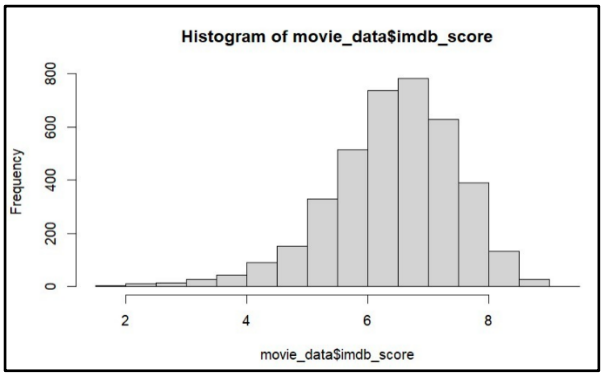


Fig 5.3

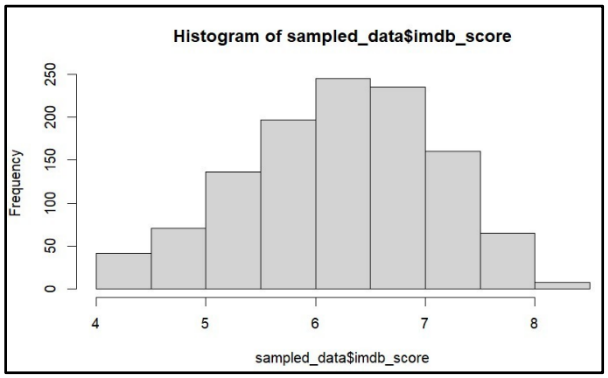
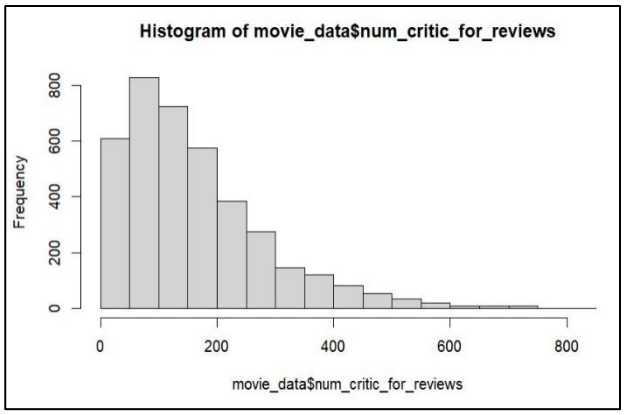


Fig 5.4

Fig 5.3 shows the distribution of movie imdb scores from the original dataset with 5000 records. There are some outliers in the dataset. Fig 5.4 shows the same distribution after data cleaning. We observe that the movie imdb score ranges between 5 mins to 7.5 for most of the movies.



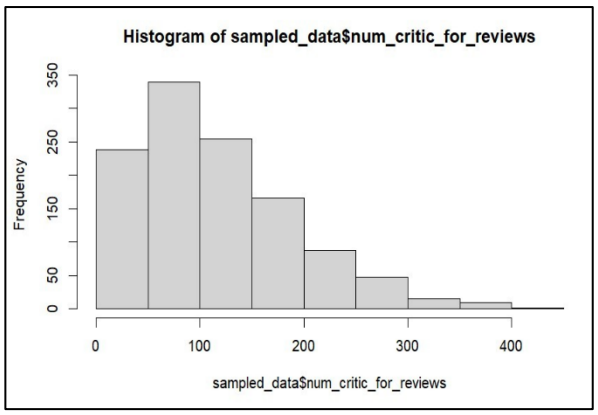


Fig5.5 Fig5.6

Fig 1.5 shows the distribution of number of critics for reviews from the original dataset with 5000 records. There are many outliers in the dataset. Fig 1.6 shows the same distribution after data cleaning. We observe that the number of critics for reviews ranges between 0 mins to 150 for most of the movies.

Word cloud:

Fig 5.7

Genre Word Cloud- The dominance of "drama" as the most prominent genre in the word cloud analysis among the 5000 best-rated movies from IMDB signifies its enduring significance in the film industry. This finding underscores the genre's strong resonance with audiences and critics, highlighting its ability to engage viewers through compelling

storytelling and emotional depth. The prevalence of drama suggests that narratives exploring human emotions and societal complexities are highly valued, contributing to movies' critical success metrics. Filmmakers and industry professionals can leverage this insight to emphasize compelling storytelling and emotional depth in their productions, aligning with audience preferences for impactful cinematic experiences.

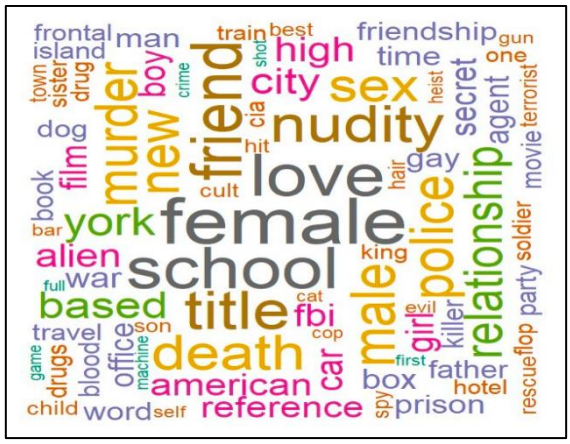


Fig 5.8

Plot Key Word- Word Cloud: The prominent keywords highlighted in the word cloud analysis, including "love," "female," "school," "sex," "nudity," "friend," "police," "male," and "relationship," collectively reveal prevalent themes and focal points within the plots of the top-rated movies in the dataset. These keywords suggest a diverse range of themes and narratives explored in these films.

The prominence of "love" and "relationship" underscores the significant role of interpersonal connections and romantic dynamics within these movies. This indicates a strong emphasis on portraying complex relationships and emotional entanglements as pivotal aspects of the plots.

Additionally, the presence of keywords such as "female," "male," "sex," and "nudity" suggests a focus on gender dynamics and themes of intimacy within the narratives. The inclusion of themes related to schools, friends, and police indicates the incorporation of settings and

situations integral to storytelling, potentially highlighting diverse backdrops and plotlines.

The coexistence of these varied themes - from love and relationships to gender dynamics, school settings, and law enforcement - reflects the multifaceted nature of the movies analyzed. It signifies a storytelling approach that encompasses a spectrum of human experiences, societal contexts, and interpersonal relationships, aiming to engage audiences through diverse and compelling plotlines. Understanding the prevalence of these keywords can guide filmmakers and industry professionals in crafting narratives that resonate with audiences, capturing the complexities of human relationships and societal dynamics while staying attuned to themes that evoke emotional connections and audience engagement.

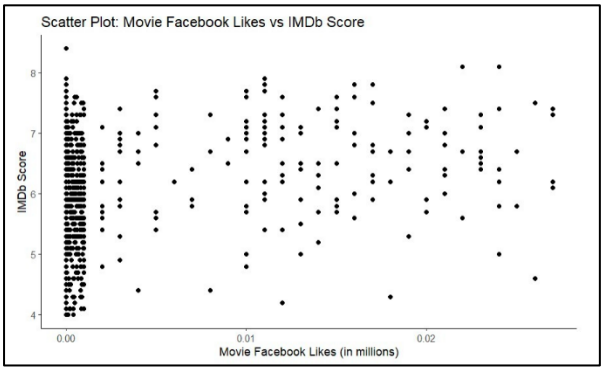


Fig 5.9

Fig. 5.9 is a scatter plot between Movie Facebook Likes Vs IMDB Score. So, we can see that there is less correlation between the variables but as the Facebook likes increases the IMDb score is also increasing.

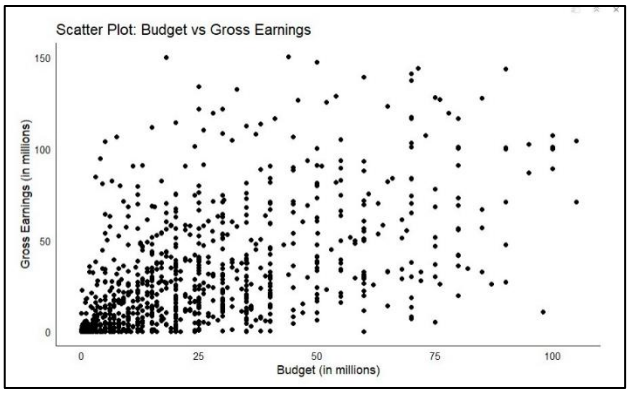


Fig 5.10

The plot Fig 5.10 shows a very faint positive linear relationship between the Gross Earnings and the Budget spent for the movies. We can also see that most data points are clustered around the bottom left of the plot, that is movies with less budget end up earning a less amount of money.

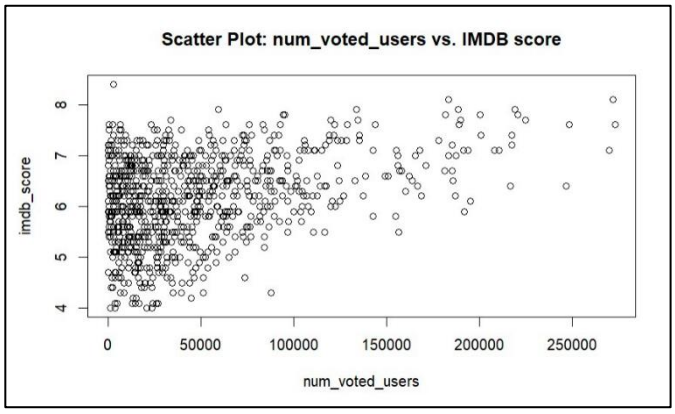


Fig 5.11

Fig 5.11 is a scatter plot between number of voted users Vs IMDB Score. We can see there is a positive correlation between the variables.

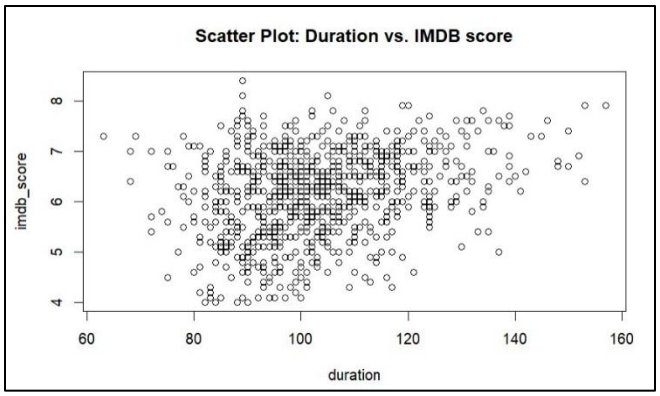


Fig 5.12

Fig 5.12 shows that there is a minor relation between Duration and IMDB score.

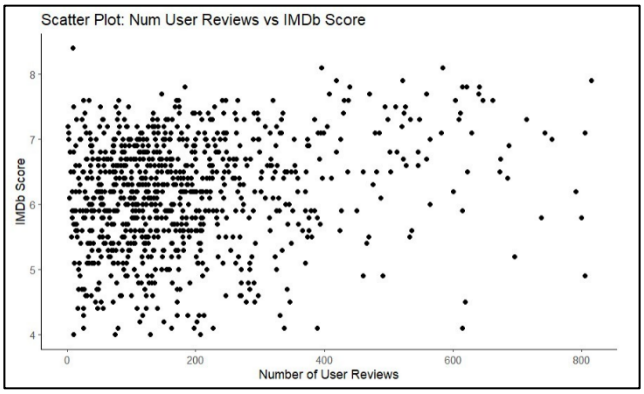


Fig 5.13

Fig 5.13 is a scatter plot between Num User Reviews Vs IMDB Score. So, we can see that there is no significant correlation between the variables.

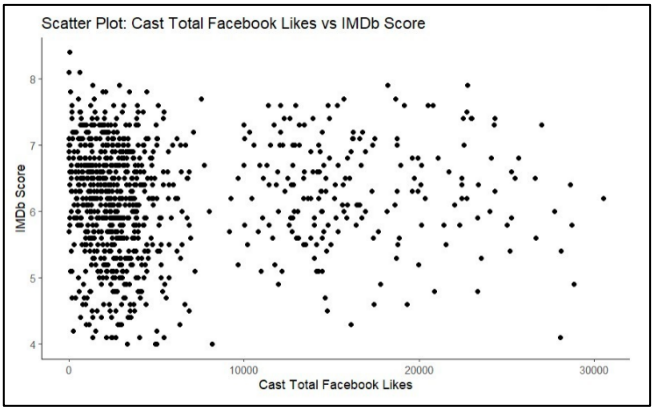


Fig 5.14

Fig 5.14. resembles there is minimal correlation between cast total Facebook likes and IMDb Score. The graph clearly tells us that Facebook likes or cast popularity does not ensure movie success every time.

Linear Regression:

In a linear regression analysis, a dataset comprising 783 samples with 13 predictors were utilized to derive insights into the relationship between the predictors and the target variable. On running multiple linear regression on the dataset with various permutation and combinations we got the following results

Regression model for IMDb rating as dependent variable:

- Model 1: Multiple R-squared: 0.3511, Adjusted R-squared: 0.3409
- Model 2: Multiple R-squared: 0.2319, Adjusted R-squared: 0.2235
- Model 3: Multiple R-squared: 0.1663, Adjusted R-squared: 0.1592

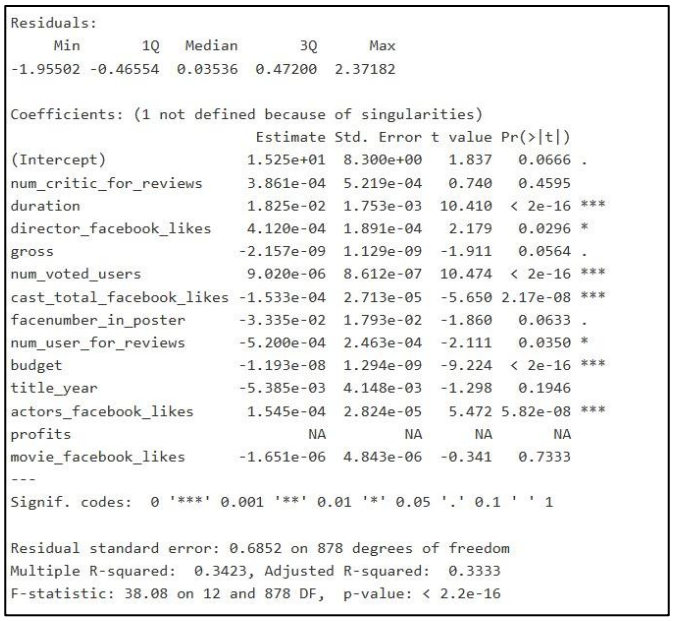
Regression model for gross earnings as dependent variable:

- Model 4: Multiple R-squared: 1, Adjusted R-squared: 1
- Model 5: Multiple R-squared: 1, Adjusted R-squared: 1
- Model 6: Multiple R-squared: 0.3837, Adjusted R-squared: 0.3785

Based on the Adjusted R-squared values:

- Out of the Models for IMDb ratings. Model 3 has the highest R-squared and adjusted R-squared value.
- Model 4 and Model 5 have the highest Adjusted R-squared value of 1, which indicates a perfect fit. However, a perfect fit is unusual and may indicate overfitting.
- Model 6 has the next highest Adjusted R-squared value of 0.3785, indicating a good fit.

Model 3 results and insights: Model for IMDb scores as dependent variable.



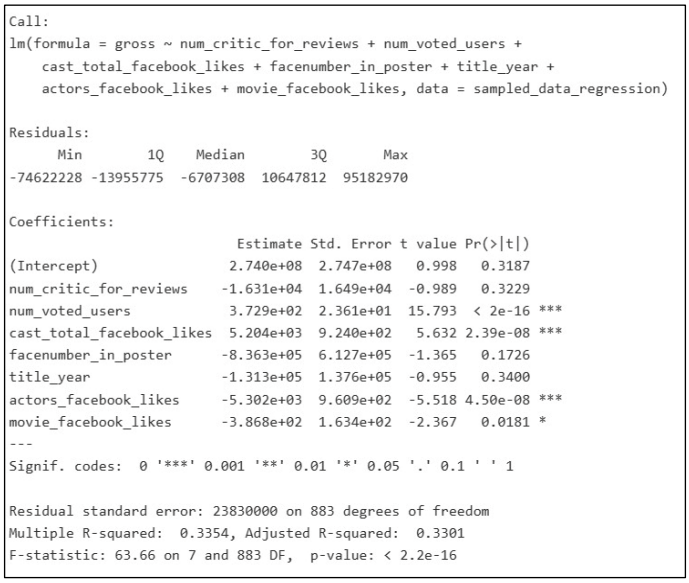
```
Residuals:
     Min       1Q   Median       3Q      Max
-1.95502 -0.46554  0.03536  0.47200  2.37182

Coefficients: (1 not defined because of singularities)
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                1.525e+01  8.300e+00   1.837   0.0666 .
num_critic_for_reviews     3.861e-04  5.219e-04   0.740   0.4595
duration                   1.825e-02  1.753e-03  10.410  < 2e-16 ***
director_facebook_likes    4.120e-04  1.891e-04   2.179   0.0296 *
gross                     -2.157e-09  1.129e-09  -1.911   0.0564 .
num_voted_users            9.020e-06  8.612e-07  10.474  < 2e-16 ***
cast_total_facebook_likes -1.533e-04  2.713e-05  -5.650 2.17e-08 ***
facenumber_in_poster      -3.335e-02  1.793e-02  -1.860   0.0633 .
num_user_for_reviews      -5.200e-04  2.463e-04  -2.111   0.0350 *
budget                    -1.193e-08  1.294e-09  -9.224  < 2e-16 ***
title_year                -5.385e-03  4.148e-03  -1.298   0.1946
actors_facebook_likes      1.545e-04  2.824e-05   5.472 5.82e-08 ***
profits                           NA         NA      NA       NA
movie_facebook_likes      -1.651e-06  4.843e-06  -0.341   0.7333
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6852 on 878 degrees of freedom
Multiple R-squared:  0.3423, Adjusted R-squared:  0.3333
F-statistic: 38.08 on 12 and 878 DF,  p-value: < 2.2e-16
```

Fig 5.15

Fig 5.15 are the results for the regression Model 3. Here we can see that the number of critics for reviews, director Facebook likes, number of voted users have positive impact on IMDb scores. Whereas number of faces on poster, number of users for reviews and title year have a negative impact on IMDb scores. Other variables have less impact on IMDb scores.



```
Call:
lm(formula = gross ~ num_critic_for_reviews + num_voted_users +
    cast_total_facebook_likes + facenumber_in_poster + title_year +
    actors_facebook_likes + movie_facebook_likes, data = sampled_data_regression)

Residuals:
      Min        1Q    Median        3Q       Max
-74622228 -13955775  -6707308  10647812  95182970

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               2.740e+08  2.747e+08   0.998   0.3187
num_critic_for_reviews   -1.631e+04  1.649e+04  -0.989   0.3229
num_voted_users           3.729e+02  2.361e+01  15.793  < 2e-16 ***
cast_total_facebook_likes 5.204e+03  9.240e+02   5.632 2.39e-08 ***
facenumber_in_poster     -8.363e+05  6.127e+05  -1.365   0.1726
title_year               -1.313e+05  1.376e+05  -0.955   0.3400
actors_facebook_likes    -5.302e+03  9.609e+02  -5.518 4.50e-08 ***
movie_facebook_likes     -3.868e+02  1.634e+02  -2.367   0.0181 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23830000 on 883 degrees of freedom
Multiple R-squared:  0.3354, Adjusted R-squared:  0.3301
F-statistic: 63.66 on 7 and 883 DF,  p-value: < 2.2e-16
```

Fig 5.16

Fig 5.16 the results for the regression Model 6. Here we can see that the number of voted

users, total likes on casts Facebook pages have a positive impact on Gross earnings. Whereas number of faces in poster, actors Facebook likes, and Movie Facebook likes have a negative impact on gross earnings. Other variables have less impact on gross earnings.

Therefore, based on the Adjusted R-squared values, Model 3 and Model 6 would be the optimal model selection from the given set of models for IMDb scores and gross earnings, respectively.

KNN Analysis:

The k-Nearest Neighbors (KNN) model analysis on a dataset comprising 783 samples with 13 predictors yielded insightful results. Here is a breakdown of the analysis. Pre-processing steps involved centering and scaling all 13 predictor variables.

Resampling Technique:

The model's performance was evaluated using 10-fold cross-validation, ensuring robustness by dividing the data into 10 equal parts for training and validation.

Optimal Model Selection:

```
   k   RMSE       Rsquared   MAE
   1  0.9419761  0.1388129  0.7449010
   2  0.8389456  0.1678747  0.6668998
   3  0.7948104  0.1966333  0.6310235
   4  0.7768867  0.2080645  0.6116929
   5  0.7547520  0.2336307  0.5981925
   6  0.7470958  0.2381447  0.5935633
   7  0.7453591  0.2393235  0.5923747
   8  0.7427983  0.2404278  0.5895026
   9  0.7451771  0.2370629  0.5967312
  10  0.7430226  0.2406110  0.5950130
  11  0.7481403  0.2311394  0.6000473
  12  0.7441947  0.2413992  0.5972110
  13  0.7451071  0.2419108  0.5981485
  14  0.7441604  0.2443153  0.5979700
  15  0.7453480  0.2428755  0.5977124

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was k = 8.
```

Fig 5.17

From Fig 5.17 we can see that model with k = 8 was selected based on the lowest RMSE value among the tested k values. A lower RMSE signifies improved predictive accuracy,

suggesting that for this dataset, the model's predictions were closest to the actual values when k was set to 8.

The final k-Nearest Neighbors model, with k = 8, displayed superior predictive performance in terms of RMSE compared to other tested k values. This suggests the model's predictions were more accurate when considering the eight nearest neighbors for this dataset. This analysis provides valuable guidance on the optimal selection of k for achieving the most accurate predictions.

RMSE (Root Mean Squared Error): 0.7418535

MAE (Mean Absolute Error): 0.5920977

MPE (Mean Percentage Error): -1.205968

MAPE (Mean Absolute Percentage Error): 10.29737

The lower RMSE and MAE values suggest that the model's predictions were closer to the actual values, indicating a higher level of precision. The negative MPE implies an overall underestimation of predictions, while the MAPE provides insights into the average percentage difference between predicted and actual values.

6.Finding & Conclusion:

It is important to note that despite the rigorous methodologies employed, the "Blockbuster Insights" project faced challenges due to the limitations of the dataset. The dataset, sourced from Kaggle and comprising 5000 best-rated movies from IMDb, presented issues such as outliers and discrepancies. The data cleaning process aimed to address these challenges, but inherent inaccuracies and inconsistencies in the dataset may have influenced the accuracy of the final conclusions. The presence of outliers in variables like movie duration, IMDb scores, and the number of critics for reviews posed challenges in achieving a completely accurate analysis.

These limitations emphasize the critical need for high-quality, reliable datasets in data science endeavors. Factors like Market sentiment, limitations, and constraints on movie release from the government and movie boards have significant influence on gross earnings. We feel a little more information about the market could enable us to create a better prediction model. Despite these challenges, the project provided valuable insights and laid the foundation for future research in the dynamic intersection of data science and the film industry.

7.Data Source:

Source - https://www.kaggle.com/datasets/carolzhangdc/imdb-5000-movie-dataset