Blockbuster Insights: Predicting Movie Success Metrics

Nandini Ramakrishnan
Pious Khemka
Pranay Penikalapati
Vishesh Shukla

Movie Agenda

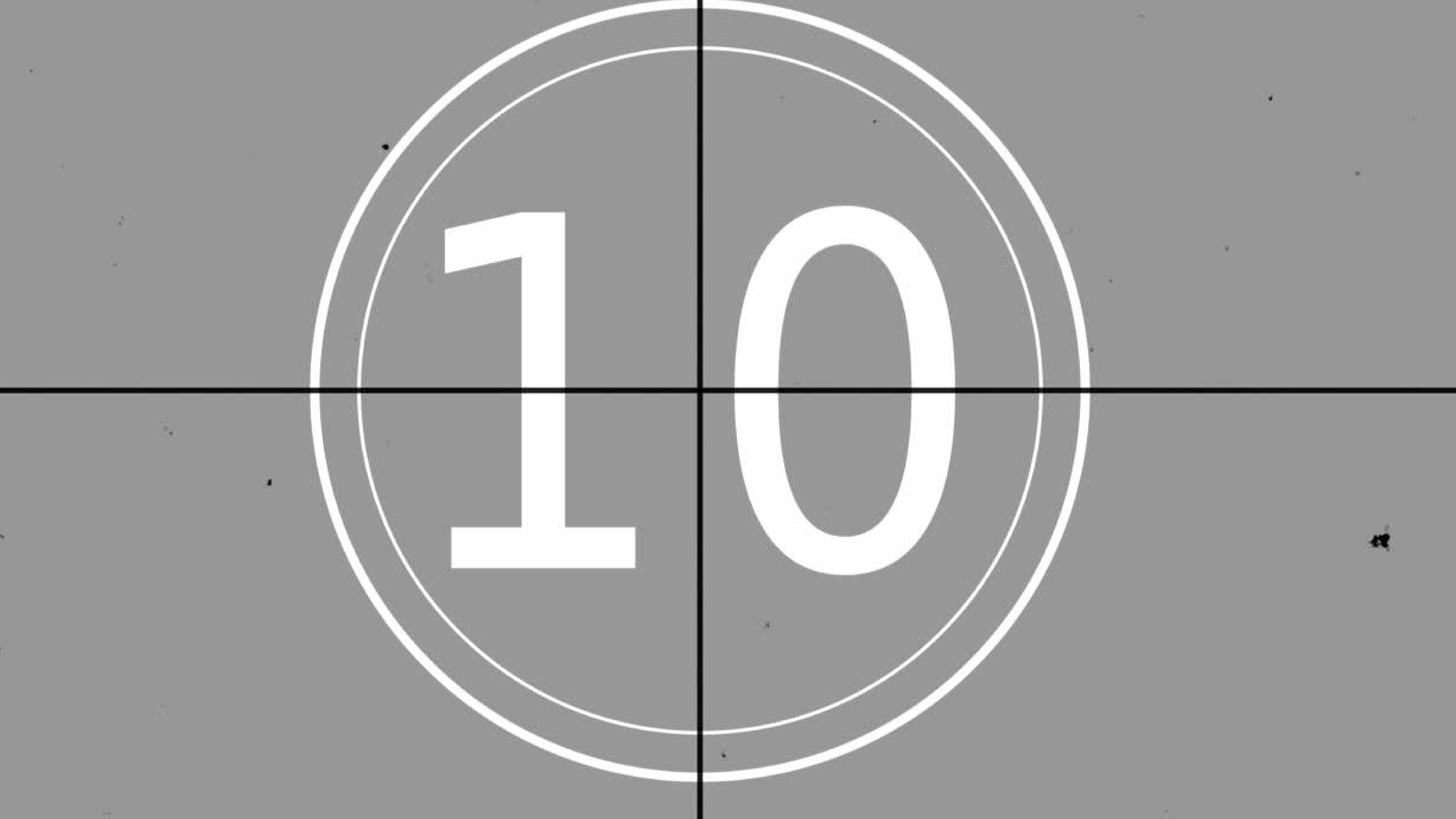
Objective

Introduction

Methodology

Data Analysis

Findings & Conclusion


10

Objective

Develop a predictive model that accurately forecasts key performance metrics for movie

Problem Statement

Predicting Movie Performance Metrics for Critical and Commercial Success (Gross Earnings)

1140

Rows - After Random selection

22

Columns

Numerical & Categorical

Data Type

Methodology

Data Collection
Data Preparation

Data Collection : KAGGLE
https://www.kaggle.com/datasets/carolzhangdc/imdb-5000-movie-dataset

Data Preparation

Removed Missing Data

Removed Outliers - IQR*1.5 times

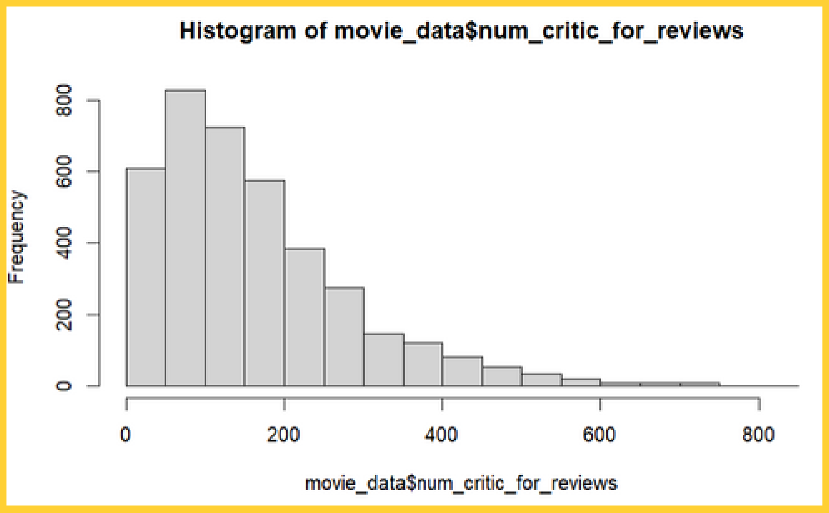
Categorical Variable - Genres

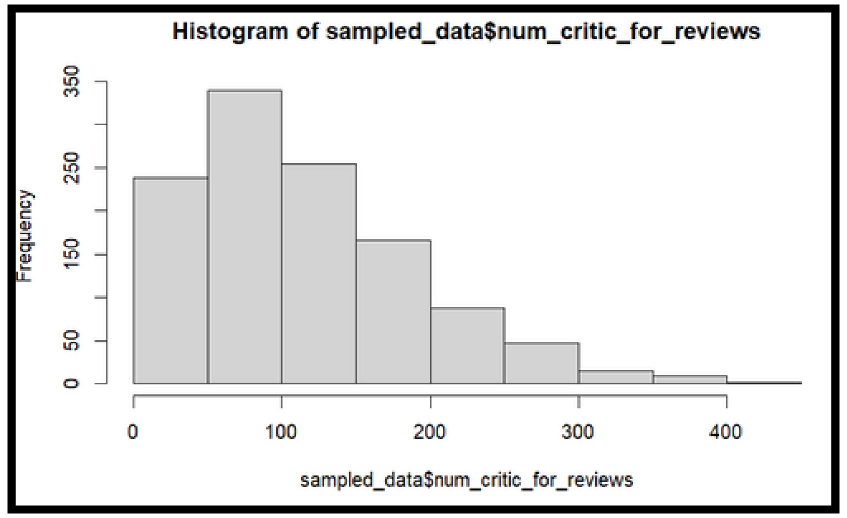
Exploratory Data Analysis

Pre-Cleaning

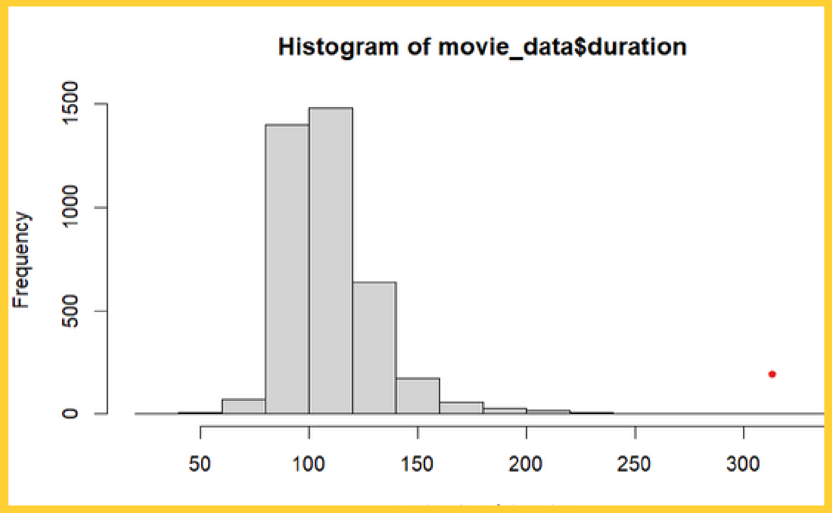
Post-Cleaning (Random Sampling)

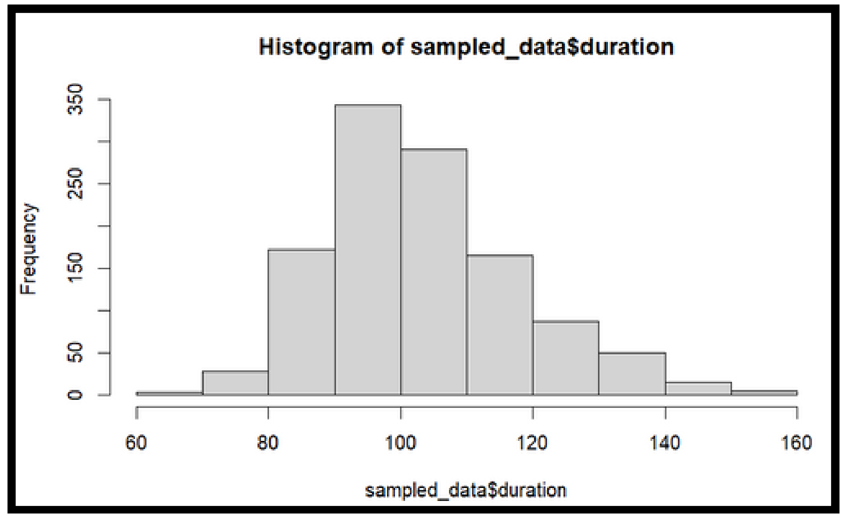
No of critic for review



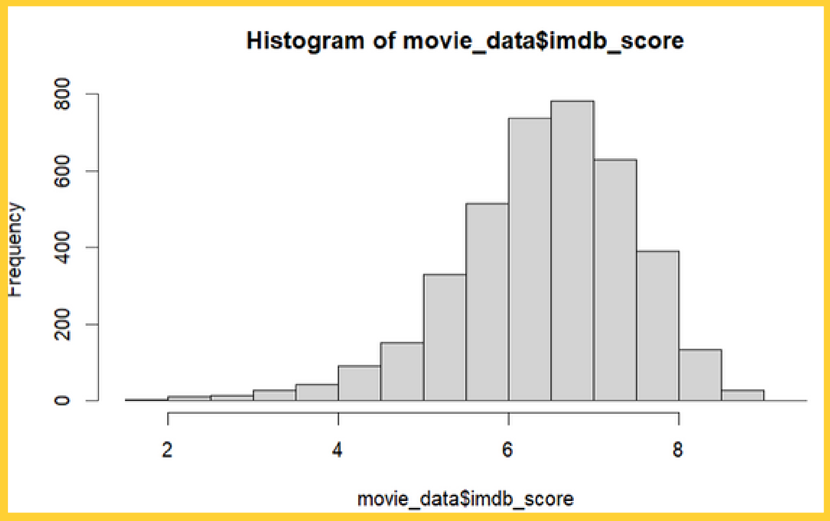


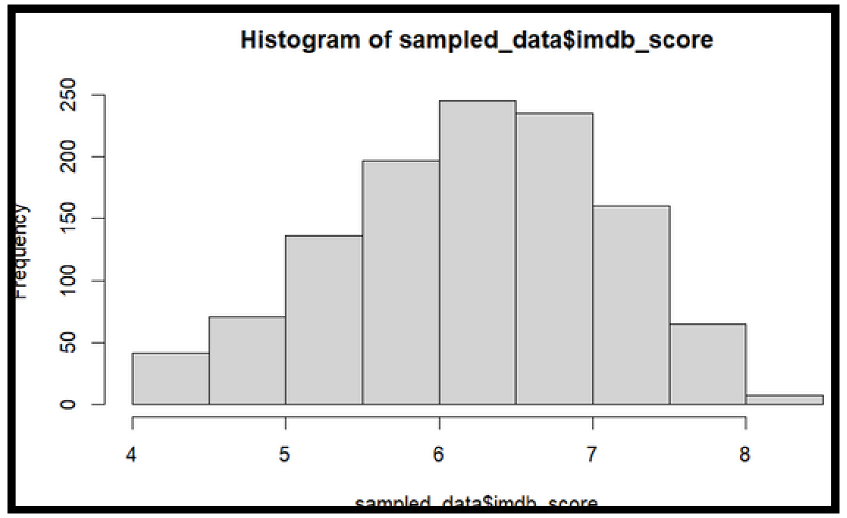
Duration





IMDb Score

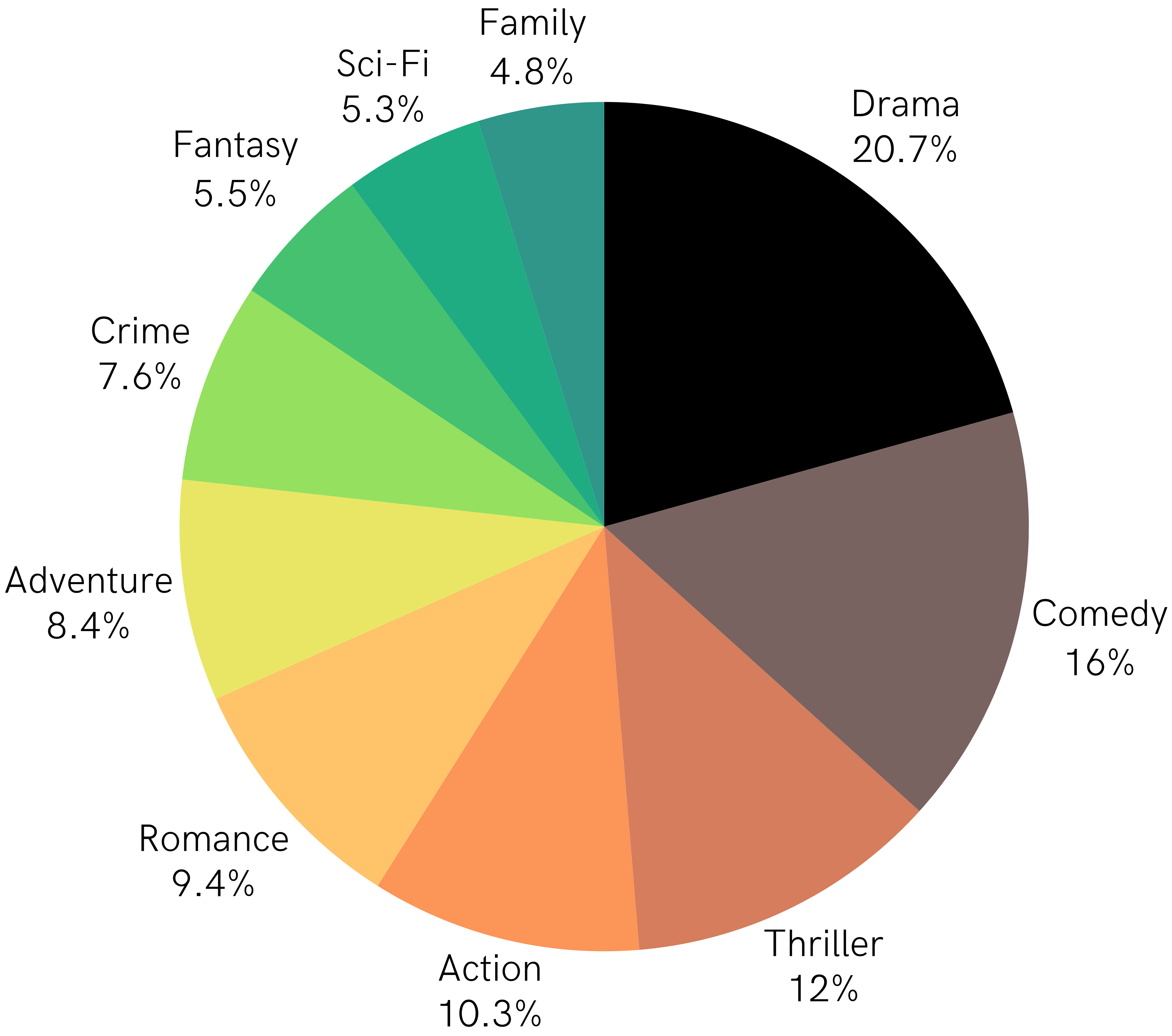




Data Analysis

Genres - Analysis

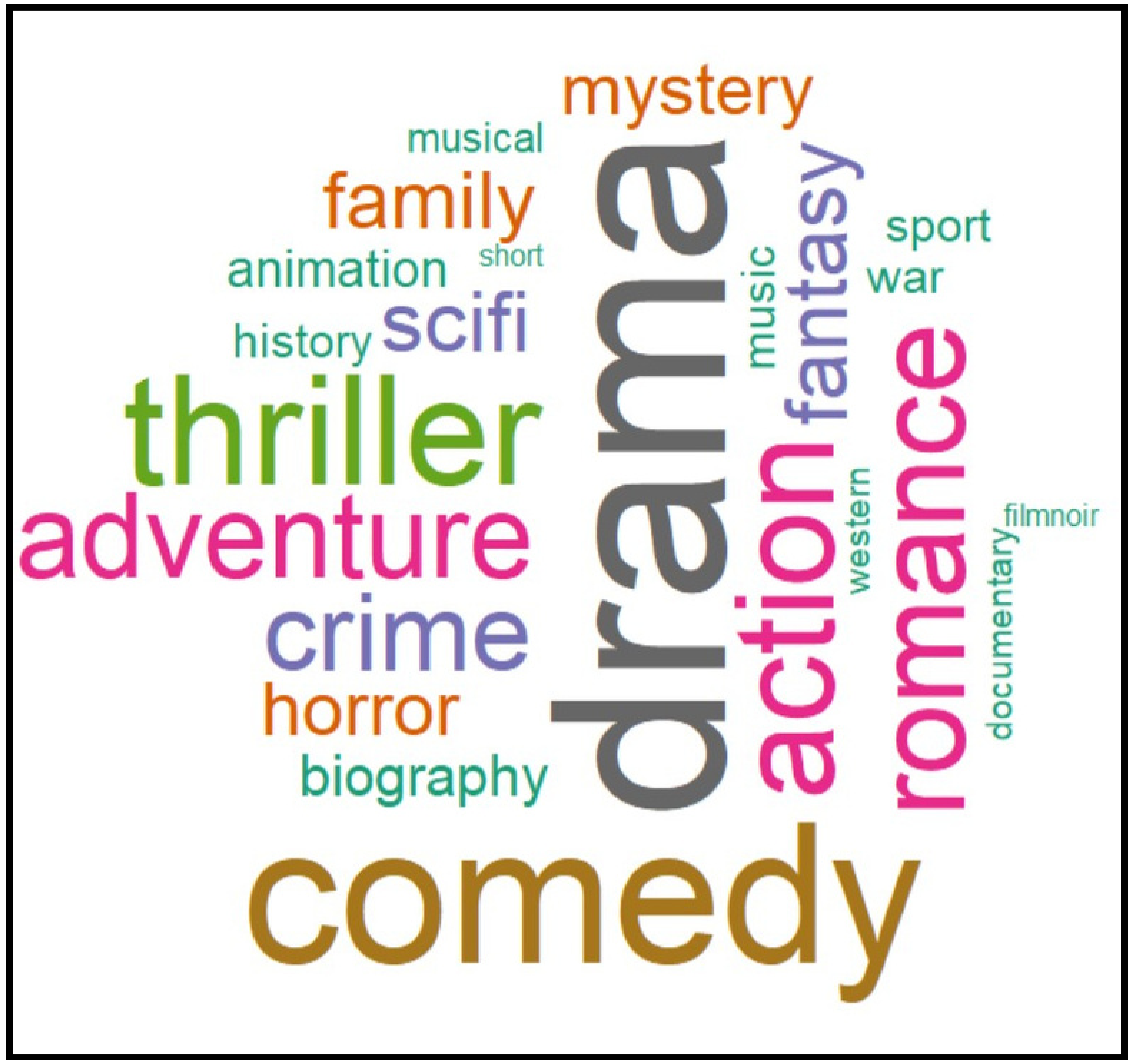
Word Cloud



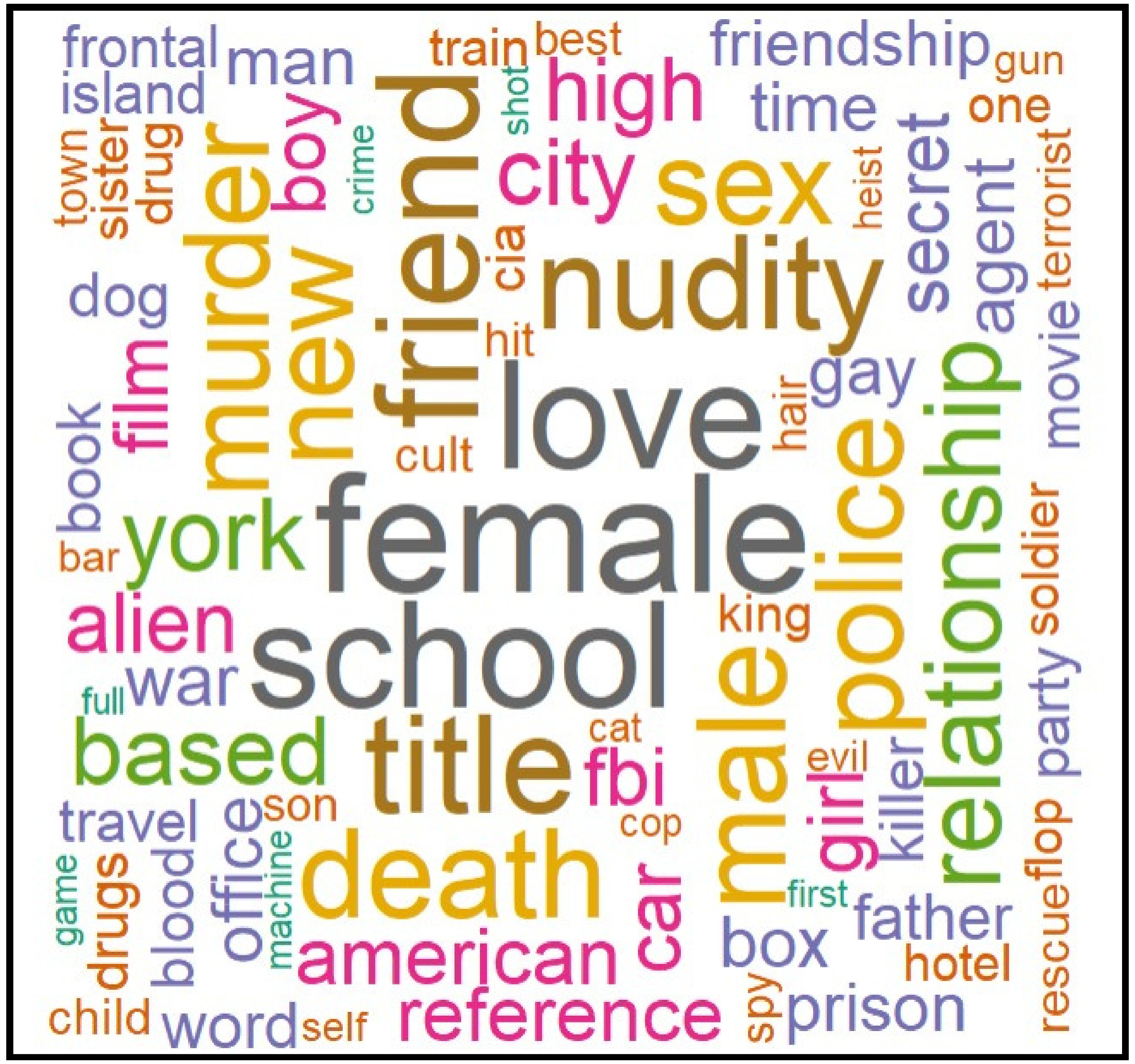
Drama

Comedy

Word Cloud - Analysis



Genres



Plot Key Words

Correlation between 'Gross Earnings' and 'IMDb Rating'

Null Hypothesis (H0): There is no significant correlation between a movie's gross earnings and its IMDb rating.
Alternative Hypothesis (H1): There is a significant correlation between a movie's gross earnings and its IMDb rating.

Correlation - 0.0143415

There is a positive correlation between gross earnings and IMDb ratings

Null Hypothesis Rejected

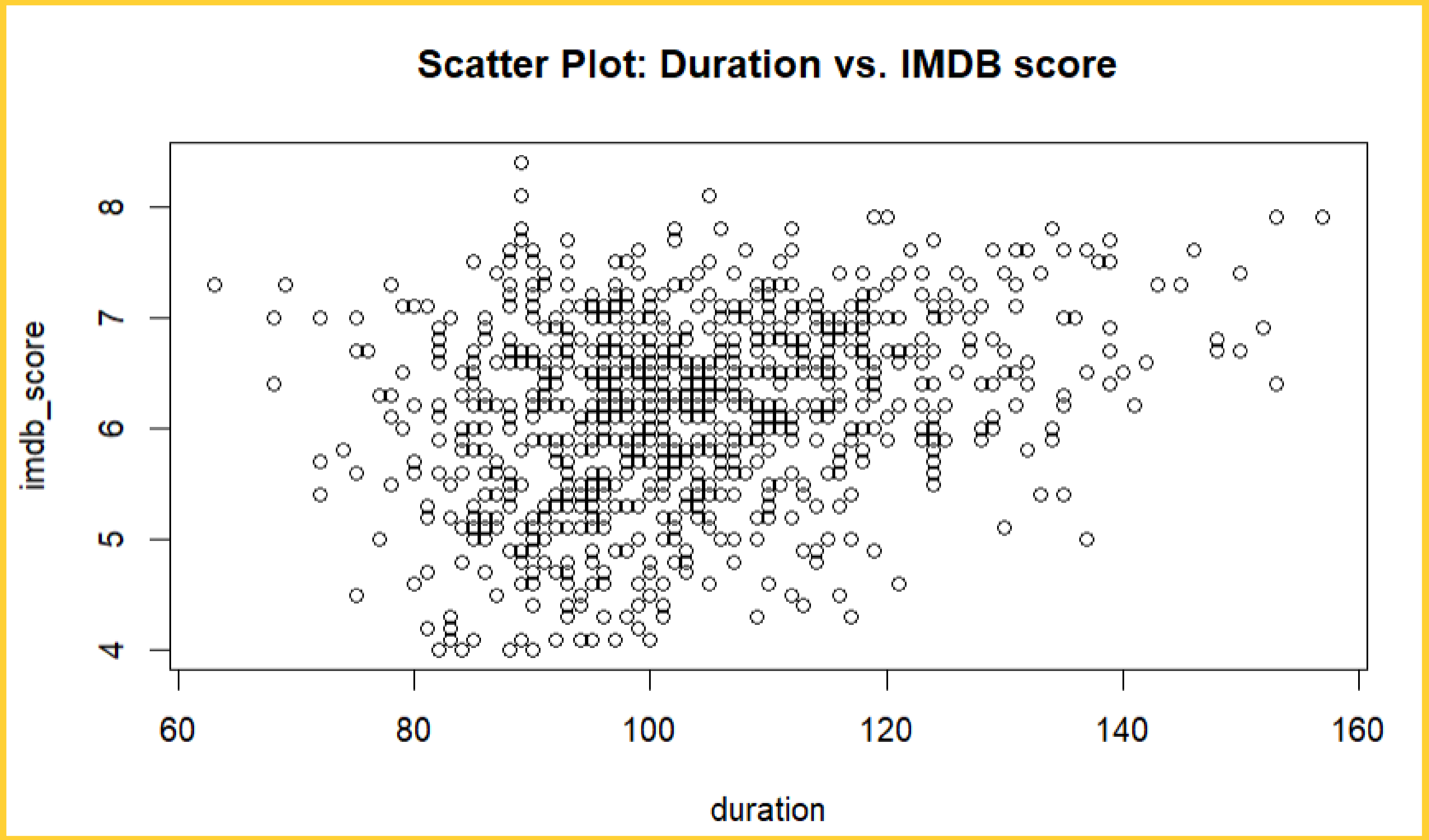
Regression and KNN Analysis

Sample - 858

Movie Country Origin - The United States of America

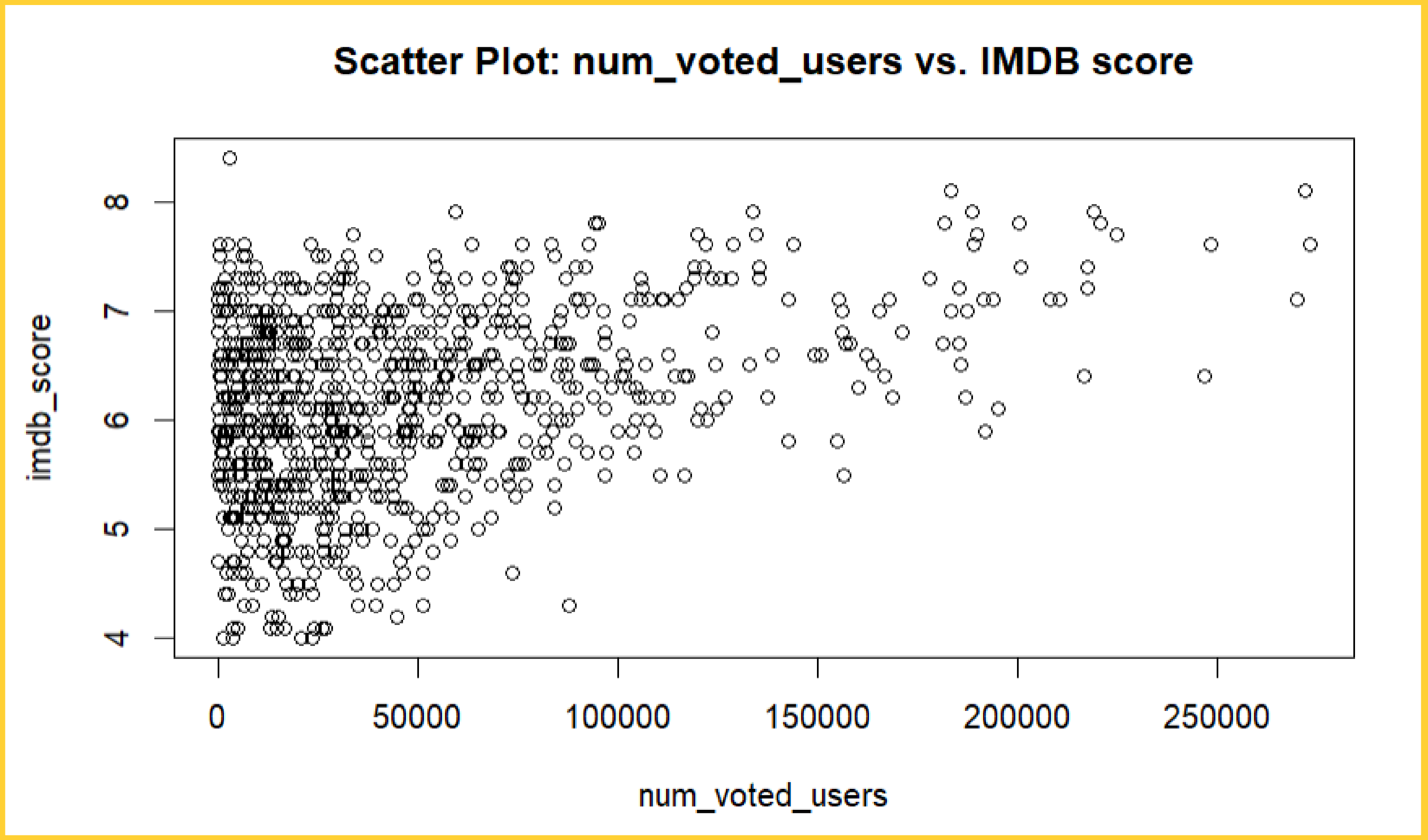
Movie Language - English

Scatter Plot between IMDb score and variables highly correlated with IMDb score

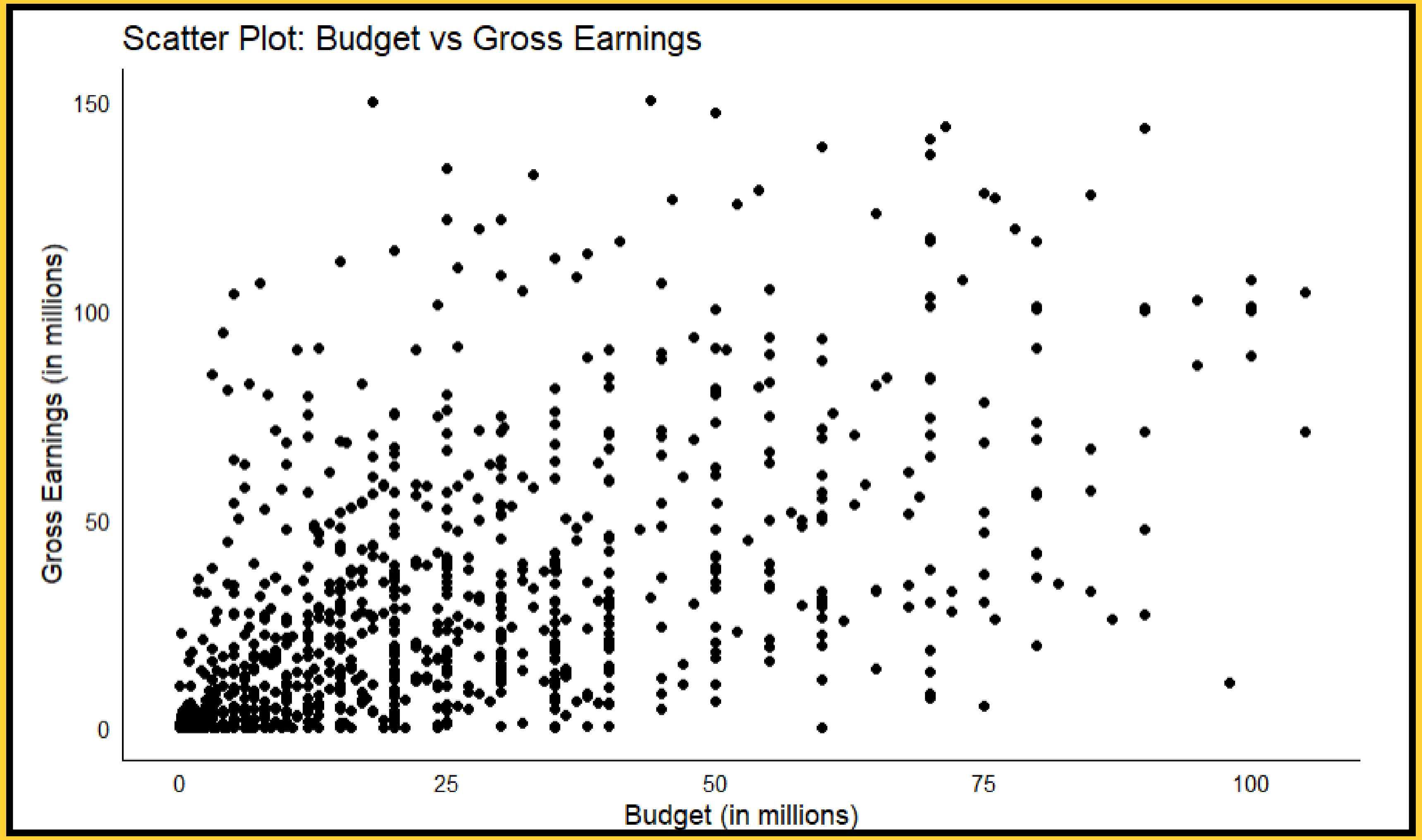


Duration

No of voted users

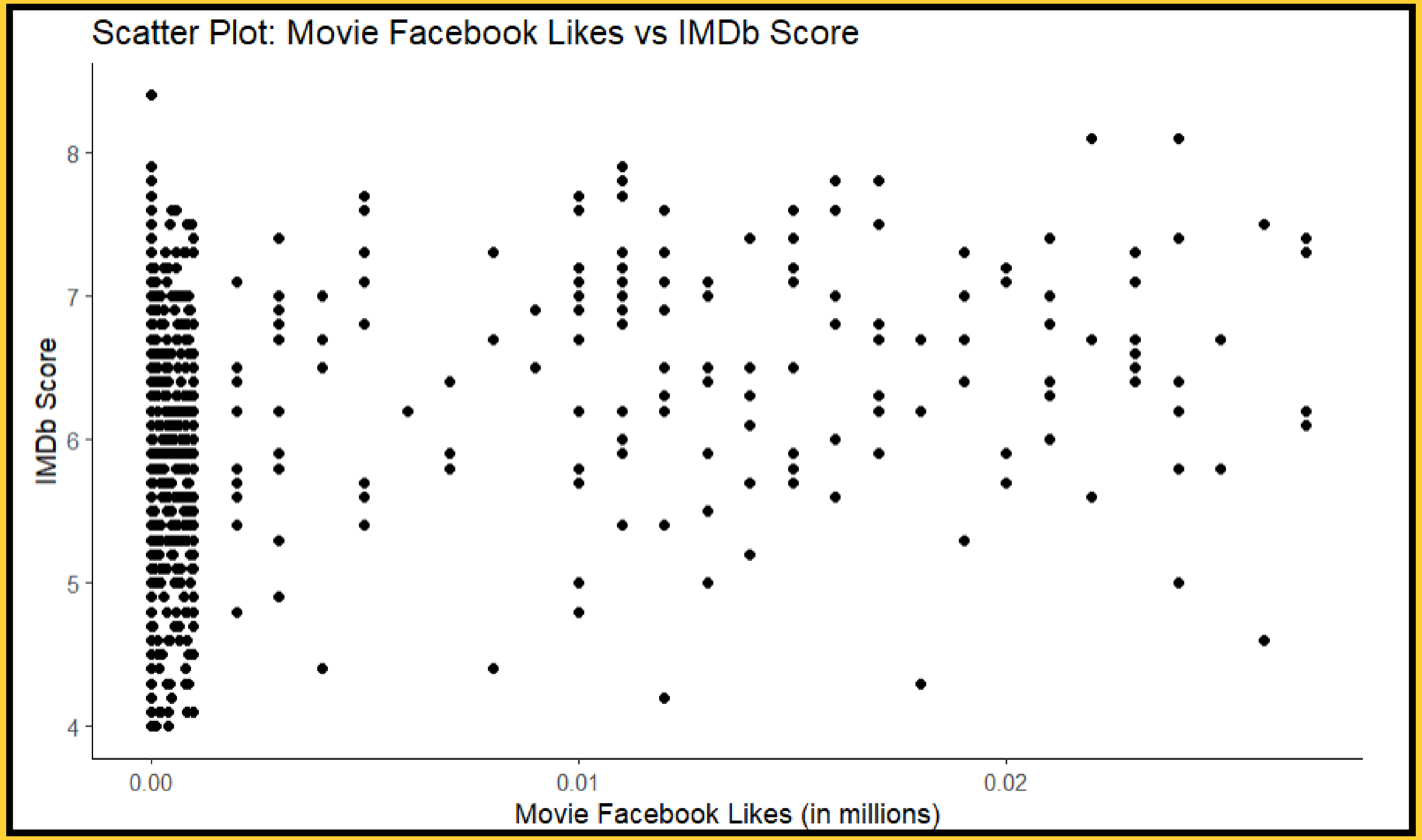


Scatter Plot - Financial & Social Metrics

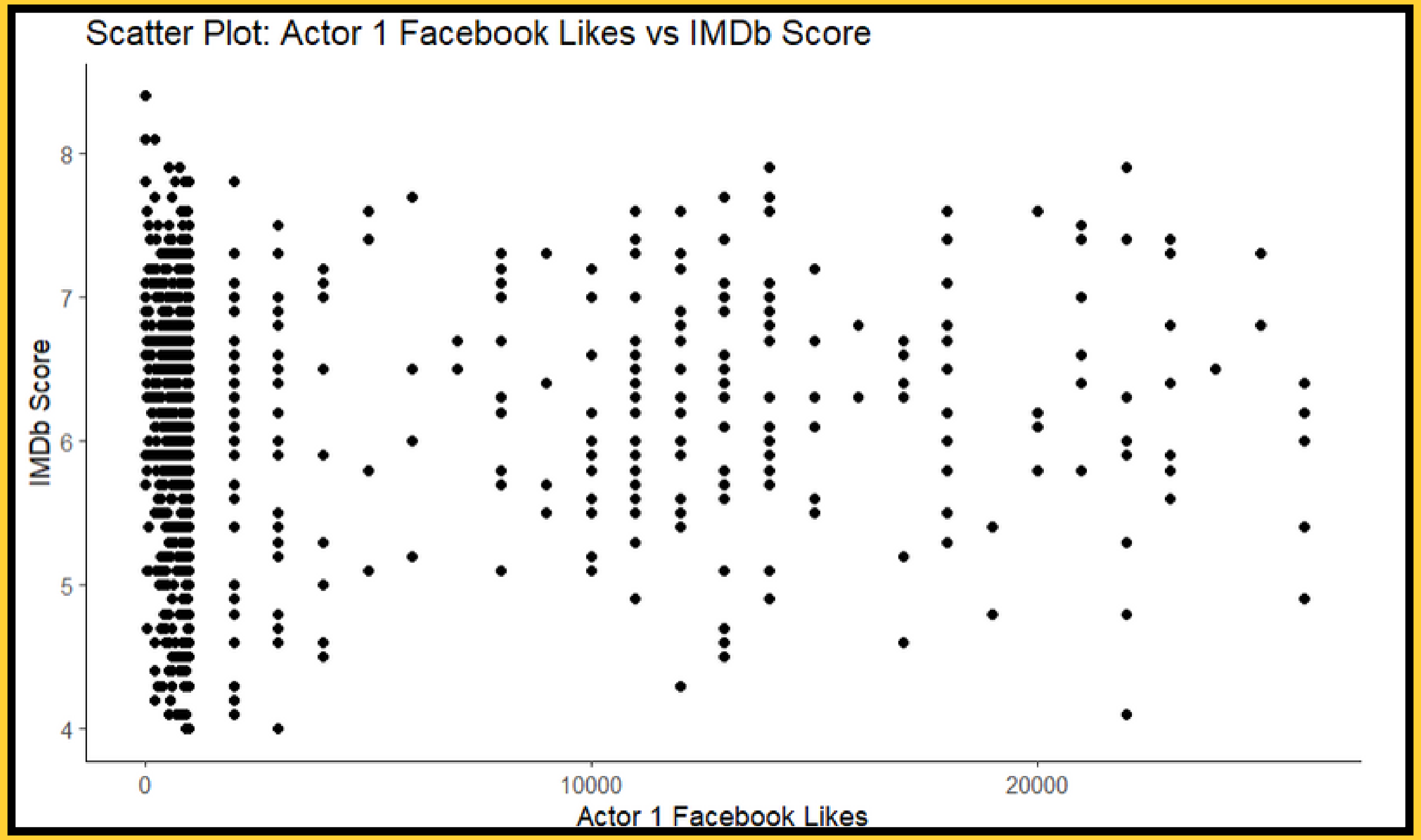


Budget Vs Gross Earning

Movie Facebook likes Vs IMDb Score

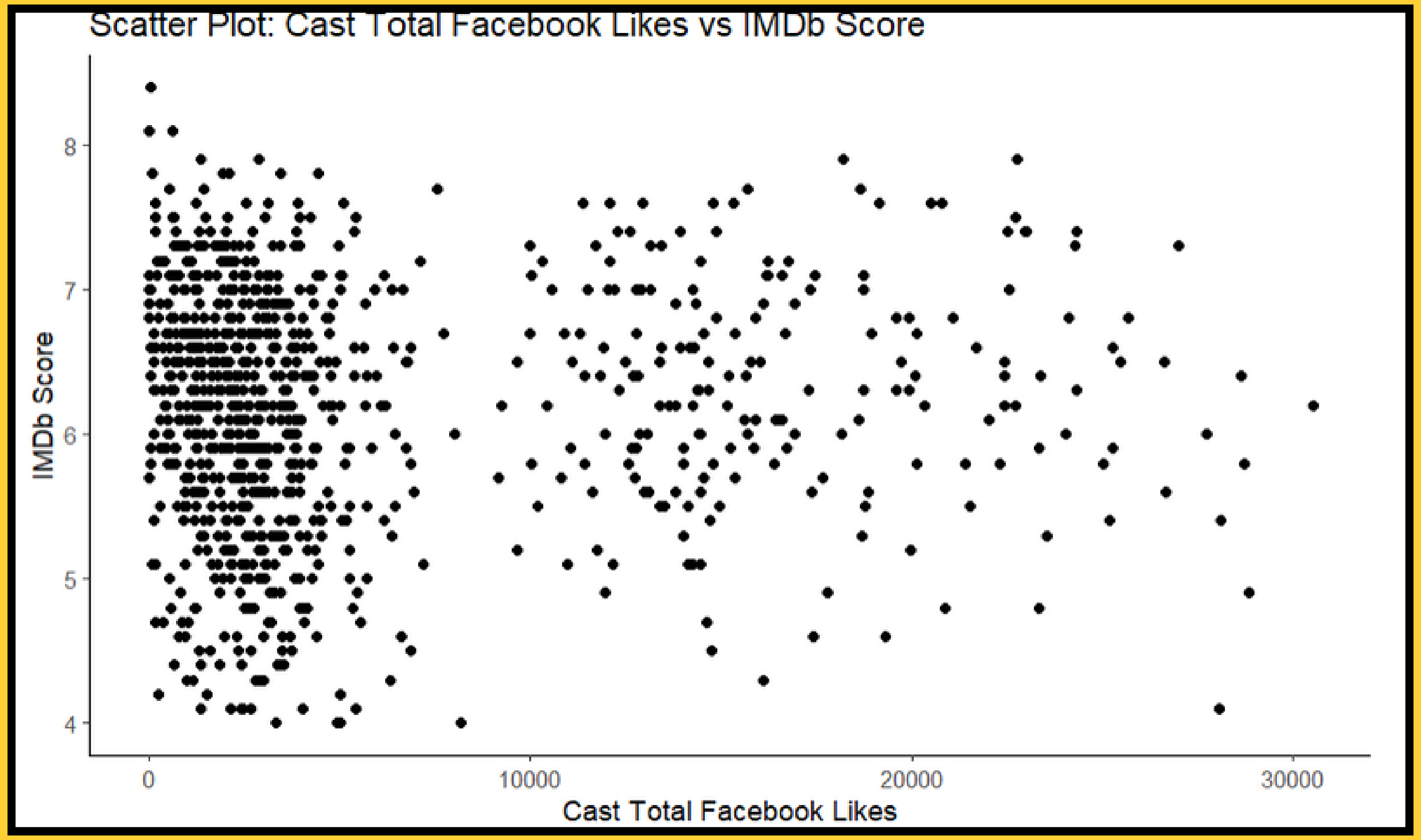


Scatter Plot - Engagement Metrics (Facebook Likes - Now Meta)

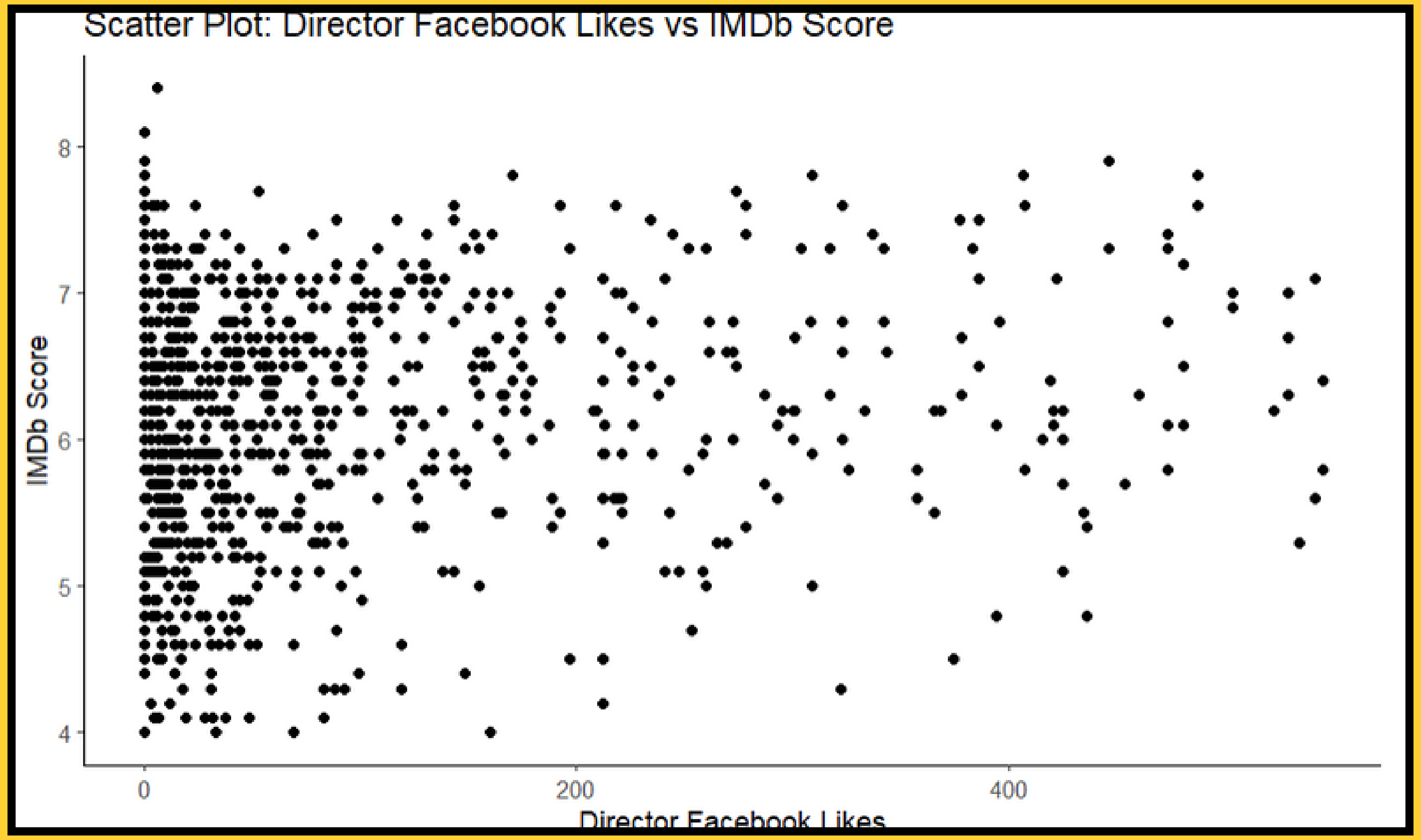


Actor FB likes Vs IMDb Score

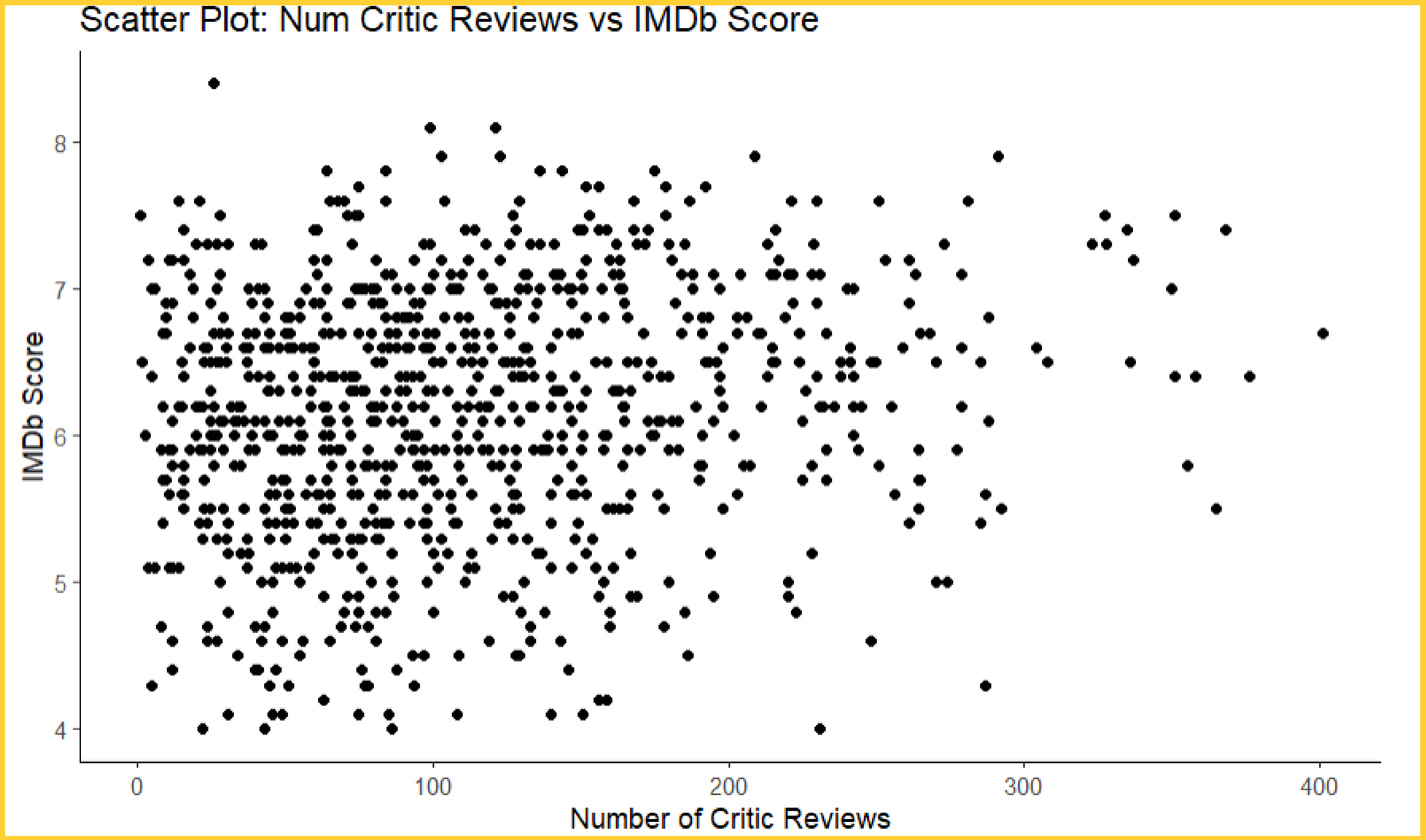
Cast Total FB likes Vs IMDb Score



Director FB likes Vs IMDb Score

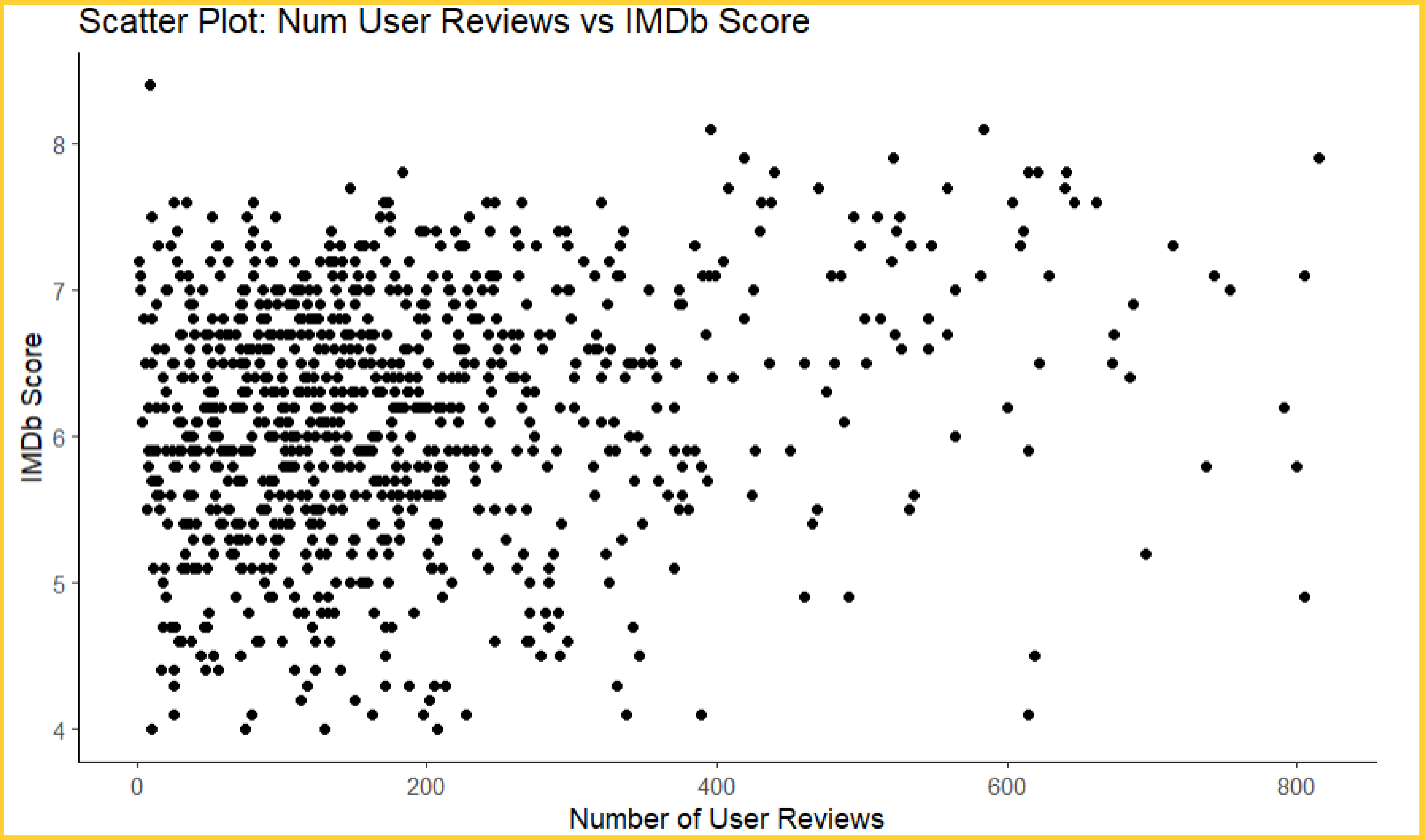


Scatter Plot - Review & Reception Metrics

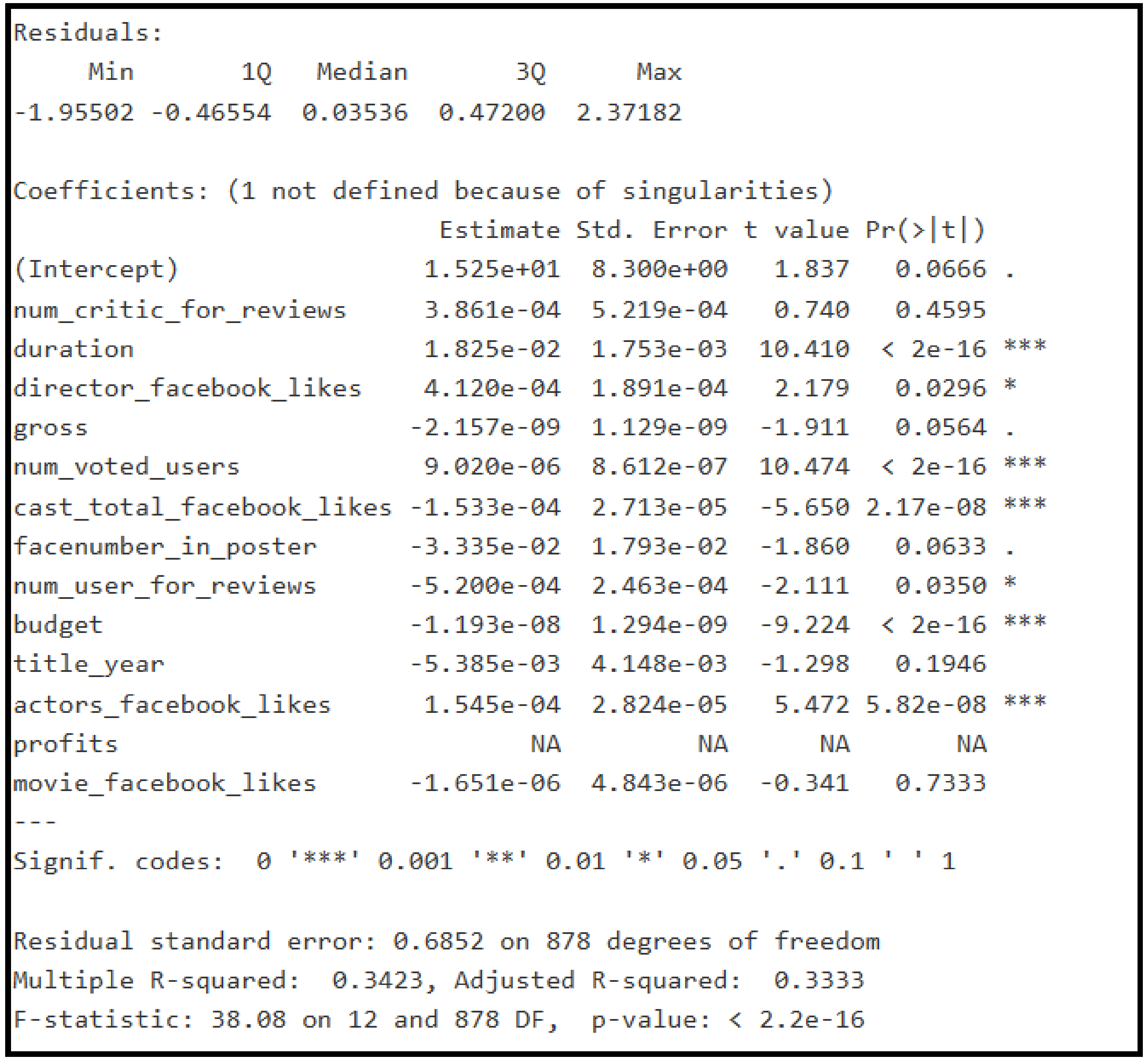


No of Critic Vs IMDb score

No of User Reviews Vs IMDb score



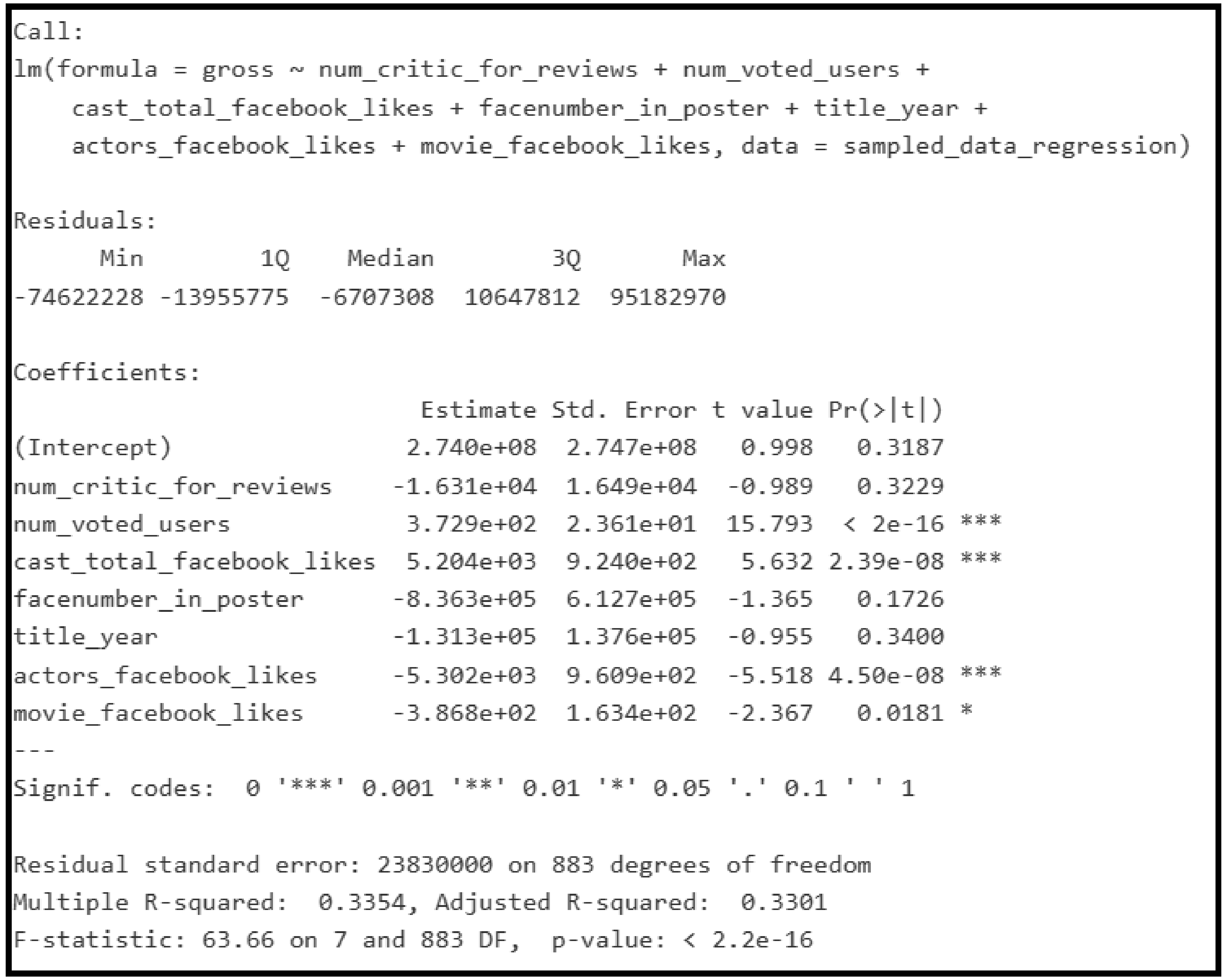
Regression Analysis - IMDb Score



```
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
```

Residual Standard Error - 0.6852
Adjusted R-squared - 33.33%

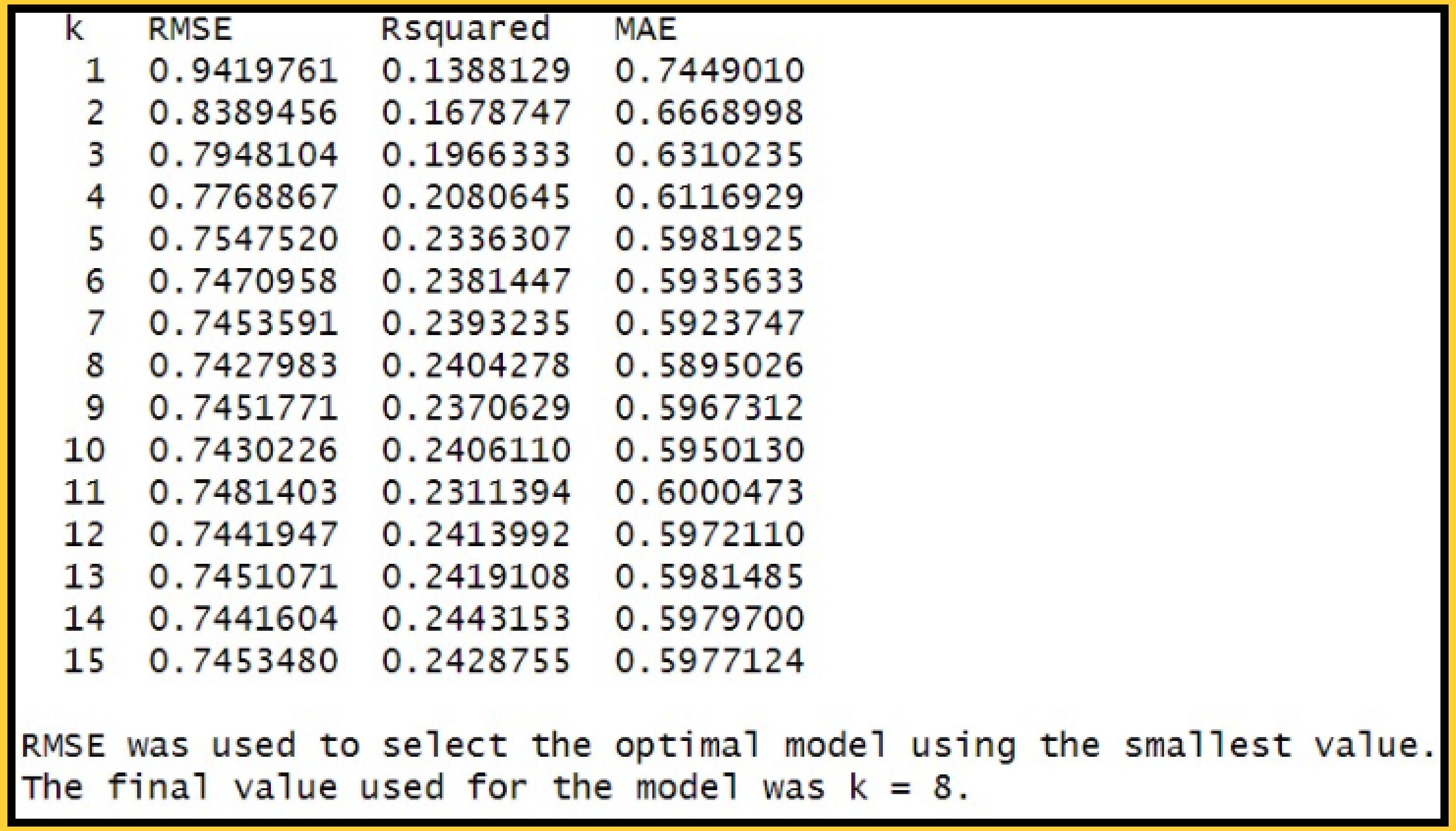
Regression Analysis - Gross Earnings



```
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
```

Residual Standard Error - 23.83 (million dollars)
Adjusted R-squared - 33.01%

KNN Analysis



```
 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.
```

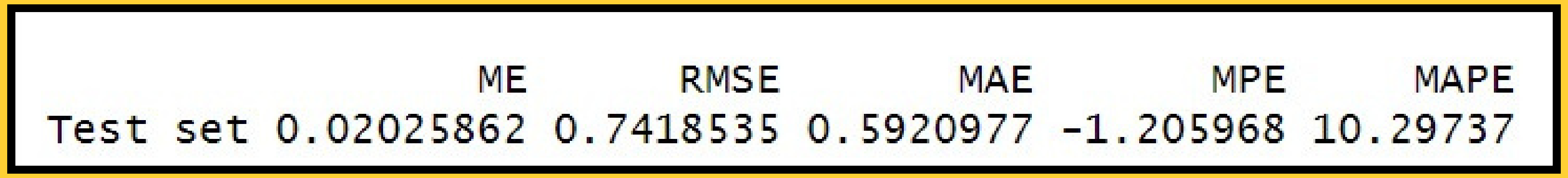
90% - 783 Samples

13 predictors

10 fold cross-validation

k=8

RMSE = 0.74



```
                 ME      RMSE       MAE       MPE     MAPE
Test set 0.02025862 0.7418535 0.5920977 -1.205968 10.29737
```

Conclusion

Thank you!

Merry Chirstmas and a Happy New Year!