AMERICAN INTERNATIONAL UNIVERSITY- BANGLADESH

Project: IMDB 5000 Movie Dataset

Course Title: INTRODUCTION TO DATA SCIENCE.

Section: C **Date of Submission: 16-08-2025**

Semester: Summer, 2024-2025 **Course Teacher: DR. ASHRAF UDDIN**

Group No: 9

No	Name	ID	Program
01	MD TANJIL TASHRIK ZIM	22-48021-2	BSc in CSE
02	MD. AL-IMRN SAYEM	22-48023-2	BSc in CSE
03	MD. ABRAR RAFID SHARIAR	22-48055-2	BSc in CSE
04	S.M. RASEL	22-48039-2	BSc in CSE

Faculty use only		
FACULTYCOMMENTS	**Marks Obtained**	
	Total Marks	

1. Introduction

This project analyzes the IMDB 5000 Movie Dataset, focusing on key attributes such as IMDb scores, budget, gross revenue, duration, and content ratings. The objective is to uncover patterns and trends that influence both a movie's financial success and critical reception. To achieve this, the dataset was cleaned and preprocessed to ensure accuracy, followed by statistical analysis, visualizations, and predictive modeling techniques. Through descriptive statistics, correlation analysis, and regression models, this study provides valuable insights into how different factors interact within the film industry. The findings are intended to support filmmakers, researchers, and industry enthusiasts in understanding the dynamics behind successful movies.

Key column:

- imdb_score (numeric): IMDB rating of the movie.
- budget (numeric): Movie production budget in USD.
- gross (numeric): Gross revenue in USD.
- duration (numeric): Movie runtime in minutes.
- content_rating (categorical): MPAA or equivalent rating .
- genres (categorical): List of genres (e.g., Drama, Comedy).
- num_user_for_reviews, num_critic_for_reviews (numeric): Engagement metrics.
- num_voted_users (numeric): Total number of user votes.
- movie_facebook_likes (numeric): Facebook popularity.
- title_year, language, country (categorical): Production metadata.

2. Data Preprocessing Steps

Data Inspection

1. Dataset size: **5,043 rows × 28 columns**.
2. Data types checked: Numeric (imdb_score, budget, gross, duration) and Categorical (content_rating, genres, language).
3. Initial review showed missing values in several important columns.

Handling Missing Values

- Checked using colSums(is.na(movie_data)).
- Missing values found:
  - **gross** = 884
  - **budget** = 492
  - **aspect_ratio** = 329
  - **content_rating** = 303
- Action: Removed rows with NA values using na.omit().

- Justification: Ensures clean, complete dataset for reliable analysis.

Removing Duplicates

- Code removed duplicate rows using !duplicated().
- Justification: Prevents repeated records from biasing results.

Exploratory Data Analysis (EDA)

Univariate Analysis (single variable):

- **IMDb Score:** Mean = 6.44, Median = 6.60, Mode = 6.7, SD = 1.13, IQR = 1.40.
- **Budget:** Strongly skewed; most budgets under $50M.
- **Gross Revenue:** Right-skewed, dominated by few blockbusters.
- **Profit:** Wide variance; some low-budget movies achieved massive ROI.
- **Duration:** Mean ≈ 107 minutes; most clustered between 90–120 min.
- Histograms plotted for IMDb score, budget, and other numeric variables.

Multivariate Analysis (multiple variables):

- Scatterplot matrix: Relationships among IMDb score, budget, gross, and duration.
- Correlation heatmap:
  - num_voted_users ↔ imdb_score
  - num_critic_for_reviews ↔ imdb_score
  - budget ↔ gross weak

Data Wrangling

- Filtered only movies with gross > $100M (~12%).
- Selected key attributes (title, imdb_score, budget, gross).
- Created **profit** column (gross – budget).

Normalization & Scaling

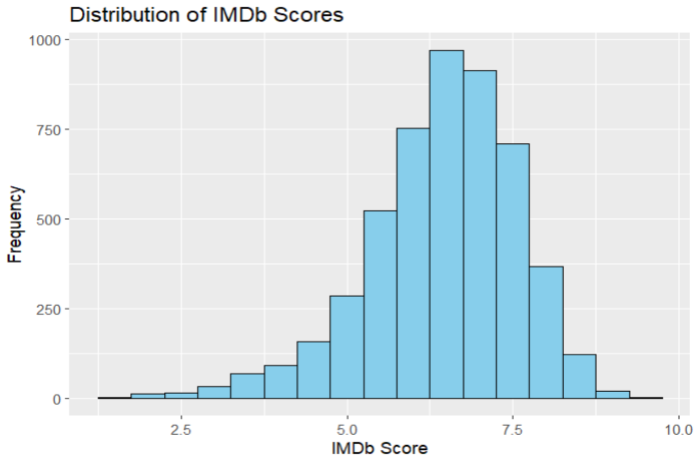
- Scaled numeric features (imdb_score, budget, gross) for consistency.
- Justification: Ensures fair comparison and prepares dataset for machine learning.

3. Key Findings & Visualizations

Finding 1: IMDb Score Distribution

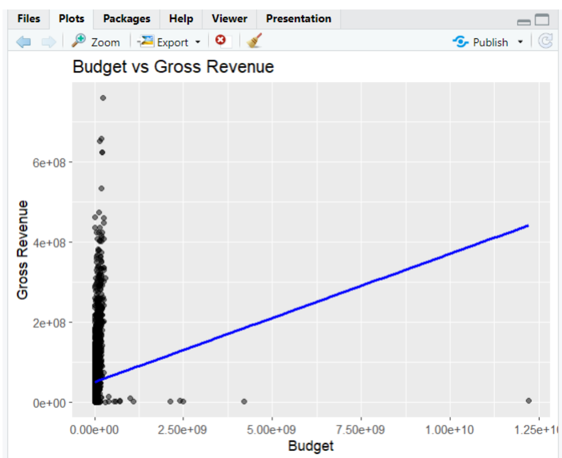
- Histogram shows majority clustered between **6–7**.

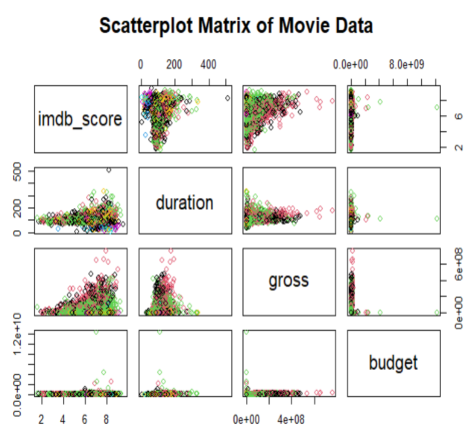
- Insight: Most movies are moderately rated; very high or very low scores are rare.



Finding 2: Budget vs Gross Revenue

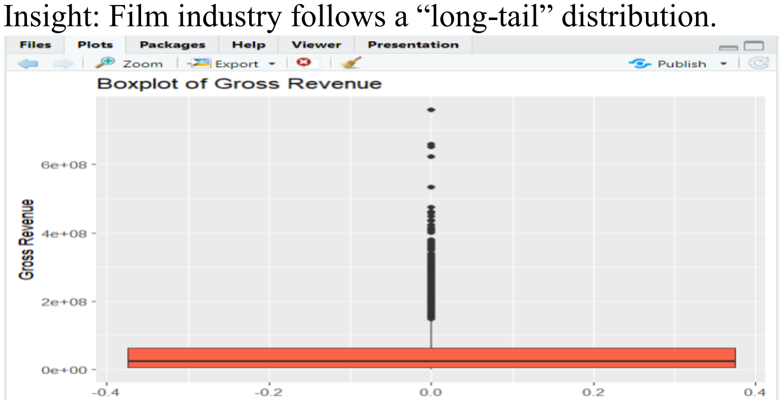
- Scatterplot with regression line shows weak relationship ($R^2 \approx 0.01$).
- Insight: Large budgets do **not** guarantee commercial success.





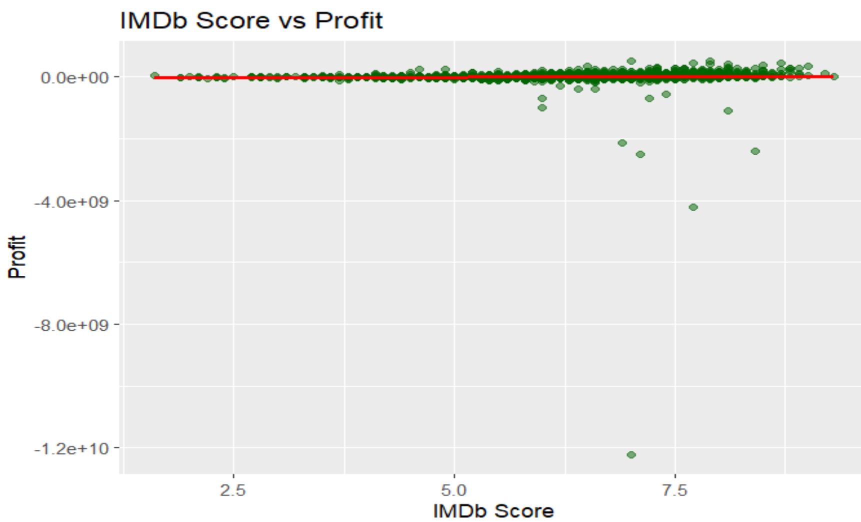
Finding 3: Boxplot of Gross Revenue

- Shows strong right-skew; a few blockbusters dominate revenues.
- Insight: Film industry follows a "long-tail" distribution.



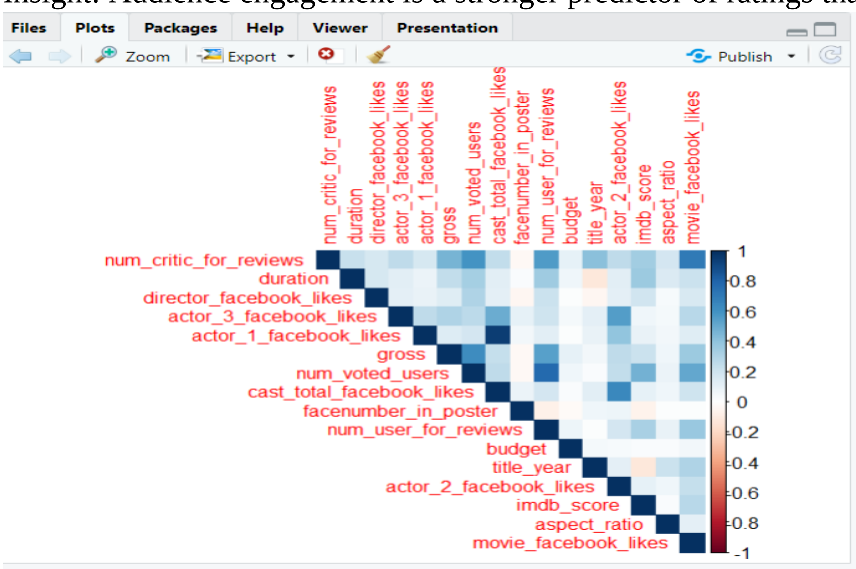
Finding 4: IMDb Score vs Profit

- Scatterplot shows a weak but noticeable trend that higher IMDb scores link to better profitability.
- Suggests critical reception may influence financial outcomes indirectly.



Finding 5: Correlation Heatmap

- Positive correlations between engagement metrics (votes, critic reviews) and IMDb scores.
- Insight: Audience engagement is a stronger predictor of ratings than budget is of revenue.



Finding 6: High-Grossing Movies (> $100M)

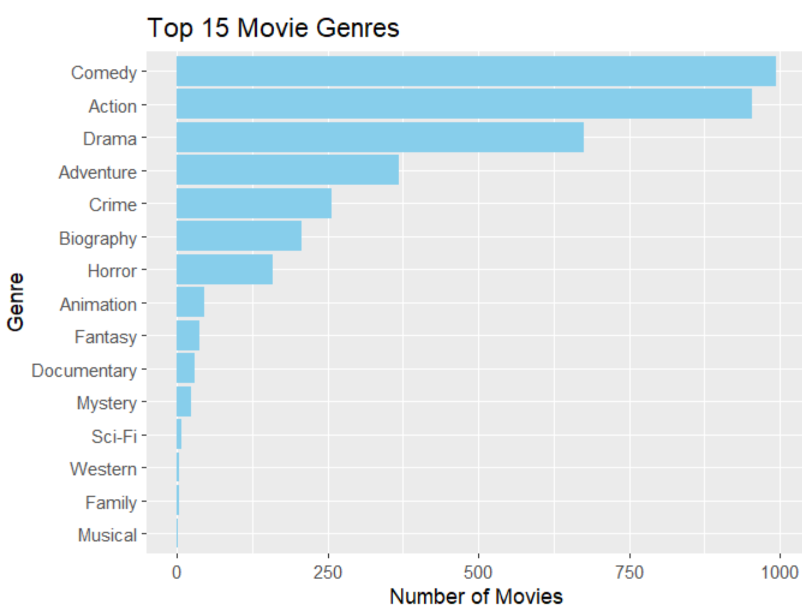
- Only about **12%** of movies cross this mark.
- Derived **profit** metric highlights some low-budget films with very high ROI.

```
> movies_filtered <- movie_clean %>%
+   filter(gross > 1e8)
> movies_filtered
   color     director_name num_critic_for_reviews duration director_facebook_likes actor_3_facebook_likes
1  Color     James Cameron                    723      178                       0                    855
2  Color    Gore Verbinski                    302      169                     563                   1000
3  Color        Sam Mendes                    602      148                       0                    161
4  Color Christopher Nolan                    813      164                   22000                  23000
5  Color         Sam Raimi                    392      156                       0                   4000
6  Color      Nathan Greno                    324      100                      15                    284
7  Color       Joss Whedon                    635      141                       0                  19000
8  Color       David Yates                    375      153                     282                  10000
9  Color       Zack Snyder                    673      183                       0                   2000
10 Color      Bryan Singer                    434      169                       0                    903
11 Color      Marc Forster                    403      106                     395                    393
12 Color    Gore Verbinski                    313      151                     563                   1000
13 Color       Zack Snyder                    733      143                       0                    748
14 Color    Andrew Adamson                    258      150                      80                    201
15 Color       Joss Whedon                    703      173                       0                  19000
16 Color      Rob Marshall                    448      136                     252                   1000
17 Color  Barry Sonnenfeld                    451      106                     188                    718
18 Color     Peter Jackson                    422      164                       0                    773
19 Color         Marc Webb                    599      153                     464                    963
20 Color      Ridley Scott                    343      156                       0                    738
21 Color     Peter Jackson                    509      186                       0                    773
22 Color     Peter Jackson                    446      201                       0                     84
23 Color     James Cameron                    315      194                       0                    794
24 Color     Anthony Russo                    516      147                      94                  11000
25 Color    Colin Trevorrow                   644      124                     365                   1000
26 Color        Sam Mendes                    750      143                       0                    393
27 Color         Sam Raimi                    300      135                       0                   4000
28 Color       Shane Black                    608      195                    1000                   3000
29 Color        Tim Burton                    451      108                   13000                  11000
30 Color      Brett Ratner                    334      104                     420                    560
31 Color       Dan Scanlon                    376      104                      37                    760
32 Color       Michael Bay                    366      150                       0                    464
33 Color       Michael Bay                    378      165                       0                    808
34 Color         Sam Raimi                    525      130                       0                  11000
35 Color         Marc Webb                    495      142                     464                    825
```

```
35 Color          Marc Webb                    495      142
          actor_2_name actor_1_facebook_likes     gross
1     Joel David Moore                   1000 760505847
2        Orlando Bloom                  40000 309404152
3         Rory Kinnear                  11000 200074175
4       Christian Bale                  27000 448130642
5         James Franco                  24000 336530303
6         Donna Murphy                    799 200807262
7    Robert Downey Jr.                  26000 458991599
8     Daniel Radcliffe                  25000 301956980
9         Lauren Cohan                  15000 330249062
10       Marlon Brando                  18000 200069408
11     Mathieu Amalric                    451 168368427
12       Orlando Bloom                  40000 423032628
13  Christopher Meloni                  15000 291021565
14 Pierfrancesco Favino                 22000 141614023
15   Robert Downey Jr.                  26000 623279547
16         Sam Claflin                  40000 241063875
17   Michael Stuhlbarg                  10000 179020854
18          Adam Brown                   5000 255108370
19     Andrew Garfield                  15000 262030663
20        William Hurt                    891 105219735
21          Adam Brown                   5000 258355354
22  Thomas Kretschmann                   6000 218051260
23         Kate Winslet                  29000 658672302
24  Scarlett Johansson                  21000 407197282
25          Judy Greer                   3000 652177271
26        Helen McCrory                   883 304360277
27        James Franco                  24000 373377893
28         Jon Favreau                  21000 408992272
29        Alan Rickman                  40000 334185206
30      Kelsey Grammer                  20000 234360014
31         Tyler Labine                 12000 268488329
32          Kevin Dunn                    894 402076689
33         Sophia Myles                   974 245428137
34          Mila Kunis                  44000 234903076
35     Andrew Garfield                  15000 202853933
                                                        genres
1                              Action|Adventure|Fantasy|Sci-Fi
```

Finding 6: Genre Distribution

- Drama and Comedy dominate the dataset.
- Top 15 genres visualized; multi-genre combinations are common.



4. Conclusion

This project explored the Movie Metadata dataset through data cleaning, statistical analysis, and visualization. At the beginning, missing values and duplicate records were removed to improve data quality. New variables such as profit and return on investment were created to better understand the financial performance of movies.

Descriptive statistics (mean, median, mode, standard deviation, and interquartile range) were calculated for key variables including budget, gross revenue, profit, IMDb score, and duration. These measures provided insights into the overall trends and variability in the dataset.

Univariate and bivariate visualizations revealed important patterns. Histograms showed the distribution of IMDb scores, budgets while scatterplots highlighted the relationships between budget and gross revenue, IMDb score and profit, as well as ROI and IMDb score. Correlation analysis further demonstrated strong links between budget, revenue, and profit.

The genre analysis identified the top 15 most common genres, showing which types of movies are most frequently produced. This information, along with IMDb scores and profitability measures, helps to understand both audience preferences and market performance across different genres.

In conclusion, this project demonstrates how data science techniques can be effectively applied to real-world datasets. It highlights the importance of thorough data cleaning, descriptive analysis, and visualization in generating meaningful insights that support informed decision-making.