A picture containing calendar

Description automatically generated

**AMERICAN INTERNATIONAL UNIVERSITY–BANGLADESH (AIUB)**

**Dept. of Computer Science**

**Faculty of Science and Technology**

**CSC4180: INTRODUCTION TO DATA SCIENCE**

**Summer 2024-2025**

**Section: D**

**Mid-Term Project Report**

**Supervised By**

**Dr. Ashraf Uddin**

**Submitted By:**

|  |  |
| --- | --- |
| **Name** | **ID** |
| **1. ABDULLAH ADNAN ABUL KALAM** | **22-47846-2** |
| **2. MD. NAFIS ISLAM** | **22-47908-2** |

**Introduction**

The film industry blends creativity, commerce, and culture, captivating global audiences. For this midterm project in the Introduction to Data Science course, our group analyzed the IMDB 5000 Movie Dataset from Kaggle, covering ~5,043 movies from the early 20th century to the 2010s. It includes 28 variables: quantitative metrics like budget (USD), gross revenue, IMDb scores (1-10), duration (minutes), review counts, and Facebook likes; qualitative attributes like genres, content ratings, director/actor names, plot keywords and release years, primarily U.S. films with some international entries. The dataset reveals insights into movie success but requires cleaning due to missing financial data, duplicates, zeros as placeholders, and title inconsistencies. Conducted in RStudio, our analysis loaded data from a GitHub URL for reproducibility, followed by cleaning, dplyr-based wrangling, and EDA to identify trends. We addressed these key questions:

1. **What are the distributions of IMDb scores and durations, including central tendencies and outliers?**
2. **How do budgets relate to gross revenue, and how does ROI vary by genre?**
3. **How has the average IMDb score evolved over release years and decades?**
4. **What descriptive statistics emerge for key numerical variables like budget, gross, and reviews, and how do outliers affect them?**

These questions demonstrate foundational data science skills while uncovering drivers of cinematic success.

**Data Cleaning Process**

We cleaned the IMDB 5000 Movie Dataset (~5,043 rows, 28 columns) in RStudio using dplyr, tidyr, and stringr to handle missing values, duplicates, and inconsistencies for reliable EDA. Key steps and rationale follow.

1. Initial Inspection: Loaded via read\_csv() from a GitHub URL; used str(), summary(), and glimpse() to check structure and types. Identified missing values (e.g., 884 in gross, 492 in budget) to prioritize, as they could bias financial analyses.
2. Formatting: Trimmed trailing spaces in movie\_title with str\_trim(); converted zero budget/gross to NA using na\_if(), as zeros likely meant missing data, avoiding skewed metrics.
3. Duplicates: Removed 4 duplicates via distinct() on movie\_title and title\_year to prevent over-representation; a full distinct() confirmed no exact duplicates, ensuring integrity.
4. Filtering: Dropped ~150 rows missing imdb\_score or duration with filter(), as these core variables were essential for rating and runtime analyses.
5. Feature Engineering: Created budget\_million and gross\_million (scaled by /1e6) for interpretable plots; primary\_genre via str\_extract() to parse multi-genres; decade from title\_year for temporal trends. These enabled genre and time-based insights.
6. Imputation: Filled missing budget\_million\_imp and gross\_million\_imp (new columns) with medians (~$20M) using if\_else(), selected over means for skewed data, to support ROI calculations without data loss. Then computed roi\_imp as ((gross\_million\_imp - budget\_million\_imp)/budget\_million\_imp) using imputed values for comprehensive financial analysis.
7. Outliers: Flagged via IQR (e.g., extreme budgets) but retained, as they represent valuable blockbusters like Pirates of the Caribbean.

This yielded a cleaned dataset of ~4,889 rows for robust movie trend exploration.

**Key Findings and Visualizations**

Our EDA on the cleaned IMDB 5000 Movie Dataset uncovered insights into ratings, durations, finances, and trends using ggplot2. Five plots with titles, labels, and Viridis scales address our questions, highlighting success factors.

A graph of a bar chart

AI-generated content may be incorrect.

**Figure 1: Distribution of IMDb Scores**

This histogram below (x: IMDb score 1-10, y: count, binwidth 0.5) shows a near-normal distribution peaking at 6-7, with left skew and few scores >8 (e.g., The Dark Knight). Caption: Scores cluster at 6-7, indicating most films earn moderate ratings, with rarities at the top.

A graph of a movie duration

AI-generated content may be incorrect.  
**Figure 2: Distribution of Movie Durations**

This histogram (x: minutes, y: count, binwidth 10) is right-skewed, peaking at ~100 minutes (90-120 range); outliers like Titanic (194 minutes) suggest epic formats. Caption: Most movies span 90-120 minutes, aligning with audience preferences; longer ones are uncommon.

A graph with a red line

AI-generated content may be incorrect.

**Figure 3: Budget vs. Gross Revenue (USD, Millions)**

This scatter plot (x/y: imputed millions USD) with red regression line reveals weak positive correlation; high-budget films like Avatar ($237M) vary in returns. Caption: Budget weakly correlates with gross, implying other factors (e.g., marketing, cast) drive revenue.

A graph with numbers and a graph

AI-generated content may be incorrect.

**Figure 4: ROI by Top 10 Genres**

This boxplot (x: genre, y: ROI) for common genres shows higher median ROI in Comedy and Thriller, with Action's greater variability from big budgets. Caption: Comedy/Thriller offer efficient returns; Action yields diverse outcomes.

A graph showing the growth of the year

AI-generated content may be incorrect.

**Figure 5: Average IMDb Score by Release Year**

This scatter plot (x: year, y: average score) with red regression line displays stable scores over time, with minor fluctuations. Caption: Average IMDb scores remain consistent across decades, suggesting steady audience perceptions or rating norms. These visualizations highlight moderate ratings, standard runtimes, variable finances, genre profitability, and temporal stability.

**Conclusion**

Our EDA of the IMDB 5000 Movie Dataset revealed key movie trends. IMDb scores peak at 6-7, with few exceeding 8; durations cluster at 90-120 minutes. Budgets weakly correlate with gross, not guaranteeing returns; Comedy/Thriller genres show higher median ROI, while Action varies. Average scores stay stable over decades with minor shifts. Limitations include imputed missing values (884 gross, 492 budget), potentially biasing skewed finances; Hollywood focus limits global applicability; primary genre extraction overlooks multi-genres. Overall, this analysis underscores budget, genre, and other drivers of success, laying groundwork for deeper studies.