Phases of Movie Analysis

Phase 1:

Introduction

The rapid growth of the film industry has led to the generation of vast amounts of data related to movies, including genres, budgets, revenues, ratings, and audience preferences. Analyzing this data provides valuable insights into the factors that influence a movie's success and helps stakeholders such as producers, directors, and streaming platforms make informed decisions.

The **Movies Data Analysis Project** aims to explore and analyze a movie dataset to identify trends, patterns, and relationships between different attributes such as movie genres, release years, ratings, and financial performance. By applying data analysis and visualization techniques, the project seeks to answer key questions like:

What genres are most popular over time?

Is there a relationship between budget and revenue?

How do ratings vary across different genres and years?

This project utilizes data preprocessing, exploratory data analysis (EDA), and statistical insights to transform raw movie data into meaningful information. The results of this analysis can help understand audience behavior, evaluate movie performance, and support data-driven decision-making in the entertainment industry.

In this **Movies Data Analysis Project**, a dataset with a shape of **(3726, 18)** is analyzed, meaning it contains **3,726 movies** described by **18 attributes**. These attributes include movie metadata such as *Title, Release Date, Genre, Language, Country,* and *Color/B&W*, as well as performance-related features like *IMDb Score, Gross Revenue, Budget, Total Reviews,* and *Duration*. The dataset also incorporates social media popularity indicators, including *Lead Actor*

Facebook Likes, Cast Facebook Likes, Director Facebook Likes, and *Movie Facebook Likes*.

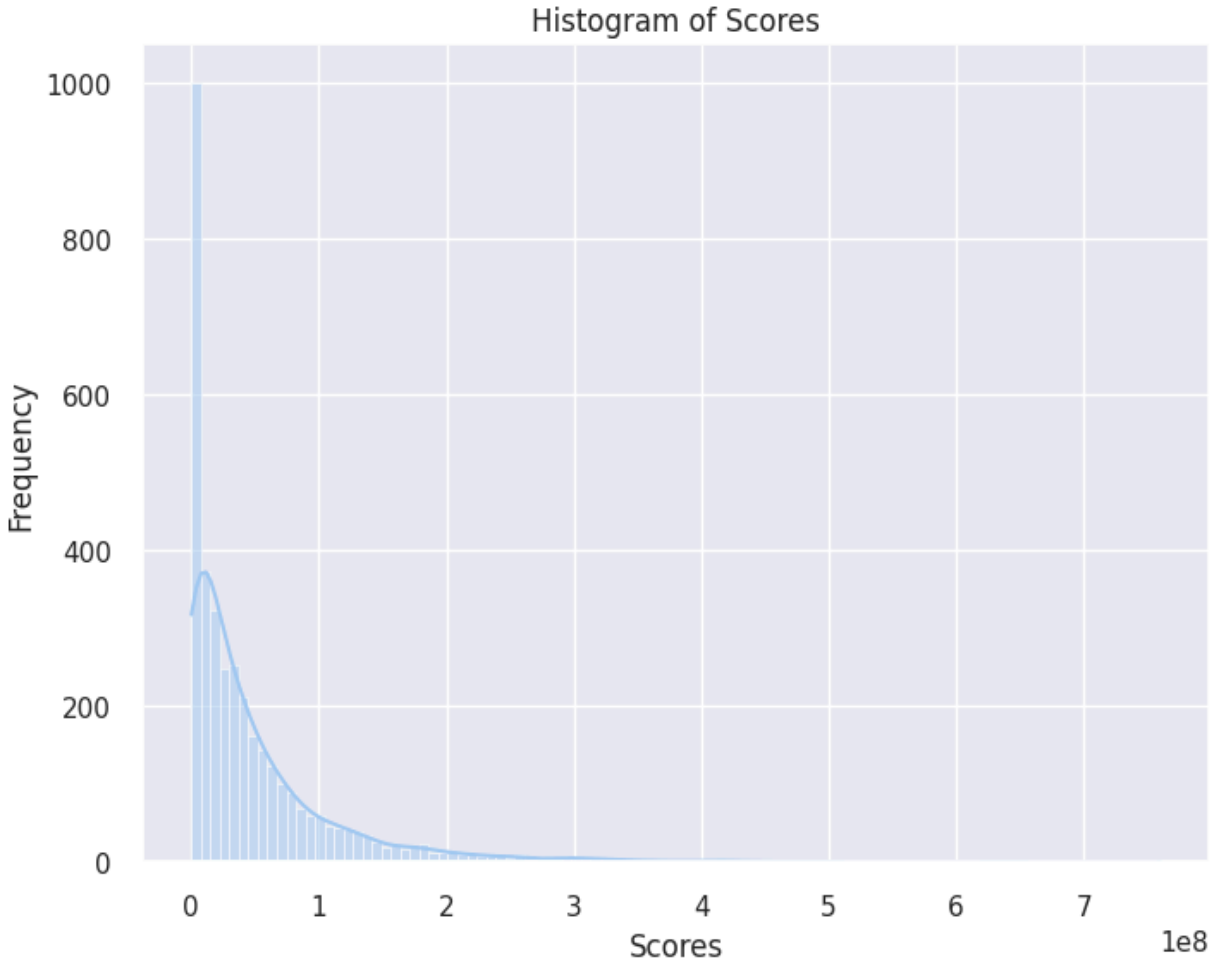
Gross Revenue is the target…The skewness of the **Gross Revenue** variable is **3.07**, which indicates a **strong positive (right) skewness**. This means that the distribution of gross revenue is not symmetric and contains a long right tail.

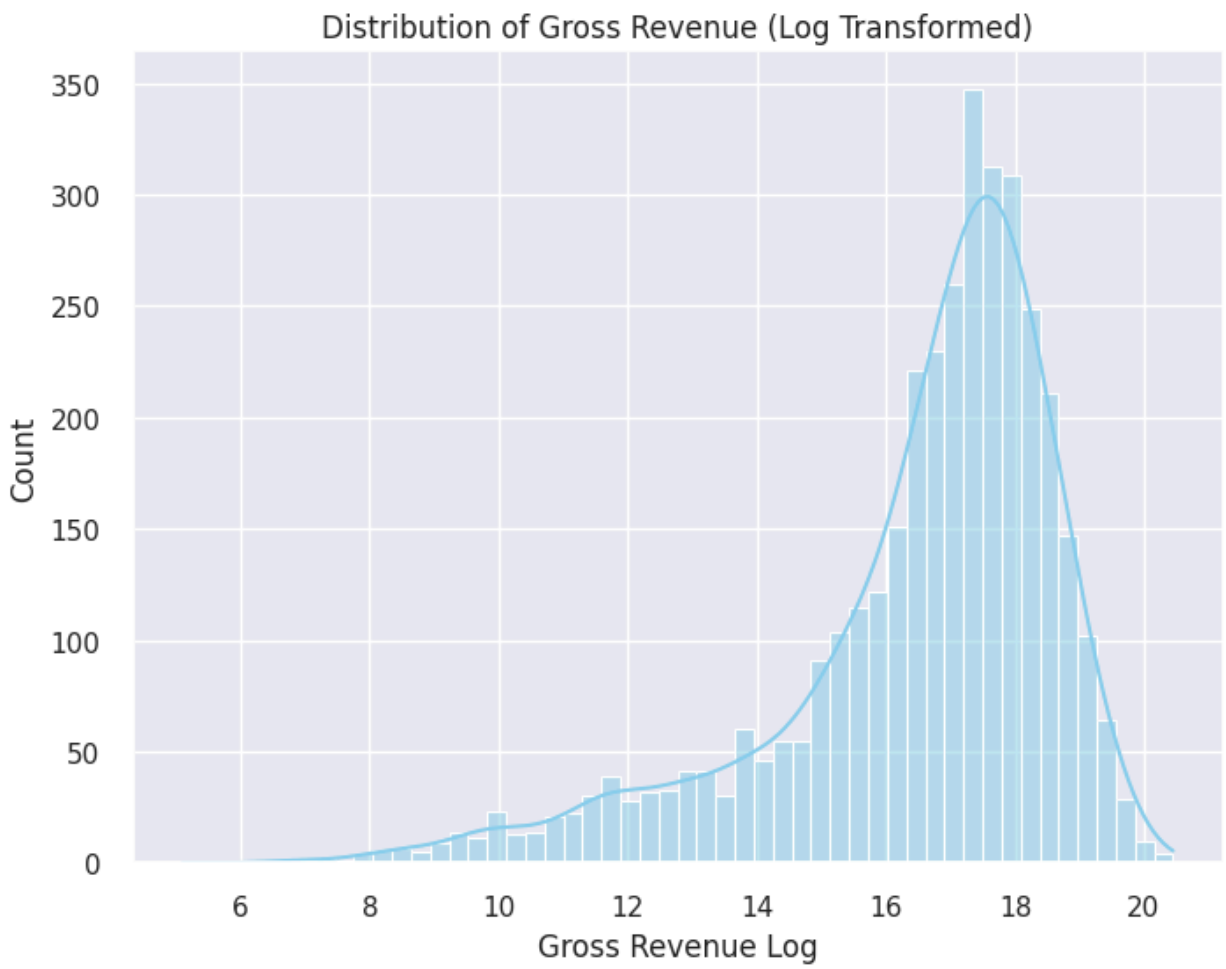
Most movies in the dataset earn **low to moderate revenues**, while a **small number of blockbuster films generate extremely high revenues**, pulling the distribution to the right. This behavior is common in the movie industry, where only a few films achieve exceptionally high financial success.

Due to this high skewness, the **mean gross revenue is significantly higher than the median**, and the data may violate normality assumptions. Therefore, transformations such as **log transformation** or the use of **robust statistics (median, IQR)** are more appropriate for further analysis and modeling involving gross revenue.

We draw the skewness of data as it is (Right_skewed)...So, we make Log transformation to make it almost normal Distribution .

You can see the difference :





Then I make some visualization and I notice Imbalance in Data which we treated it using Smooth and ENN.

Phase 2: Exploratory Data Analysis (EDA) on Movie Dataset:

1) Introduction

Exploratory Data Analysis (EDA) is a crucial step in understanding the underlying patterns, relationships, and quality of a dataset before applying any predictive modeling. In this study, we performed EDA on a movie dataset, which contains features such as **Gross Revenue, Budget, IMDb Score, Genre, Rating, Color/B&W, Lead Actor, Director Name, Release Date,** and **other numerical and categorical variables**. The goal was to identify key insights, examine distributions, detect outliers, and explore relationships among variables using cleaned data

---

2) Data Cleaning and Preprocessing:

a) Handling Missing Values:
   i) Numerical columns (**Gross Revenue, Budget, Total Reviews, Duration,** social media likes, **IMDb Score**) were filled with their respective medians.
   ii) Categorical columns (**Language, Country, Rating, Lead Actor, Director Name**) were filled using the mode of each column.

b) Outlier Detection and Removal:
   i) Outliers in **Gross Revenue** were identified using the **Interquartile Range (IQR)** method. Observations outside **1.5*IQR** from Q1 and Q3 were removed.

   ii) This resulted in a cleaner dataset (**df_cleaned**) suitable for analysis and visualization.

c) Feature Engineering:
   i) **Gross Revenue Log**: Log-transformed **Gross Revenue** to reduce skewness.

   ii) **Release Year, Release Month, Release Day:** Extracted from the **Release Date.**

   iii) **Is Recent**: Binary feature indicating movies released from 2015 onward.

3) Univariate Analysis:

a) Numerical Features:
   i) Histograms and KDE plots were generated for numerical features, including **Gross Revenue, Budget, IMDb Score, Total Reviews, Duration,** and social media likes.

   ii) Boxplots illustrated the distribution of numerical features before and after removing outliers.
   Observation: **Gross Revenue** was right-skewed, which was corrected partially with log transformation.

b) Categorical Features:
   i) Count plots were used to visualize the distribution of categorical features such as **Genre, Rating, Color/B&W, Language, Country, Lead Actor,** and **Director Name**.
      1) **Most common genres:** Identified by frequency counts.
      2) **Color/B&W distribution:** Majority of movies are colored.
      3) **Top actors and directors:** Determined by movie count.

---

4) Bivariate Analysis:

a) Gross Revenue Relationships:
   - **Gross Revenue vs Budget:** Positive trend observed; higher budgets tend to correspond to higher gross revenue.
   - **Gross Revenue vs IMDb Score:** Weak-to-moderate positive correlation; higher-rated movies tend to perform better, but not always.
   - **Gross Revenue vs Genre:** Certain genres such as Action and Adventure have higher median gross revenue.
   - **Gross Revenue vs Release Year:** Recent movies (post-2015) generally show higher revenues.
   - **Gross Revenue vs Color/B&W:** Colored movies dominate high-grossing films.

b) Budget Relationships:

- Budget varies across genres. Boxplots indicated that genres such as Action, Adventure, and Sci-Fi generally have higher budgets.

c) Top Lead Actors:

- The top 10 lead actors were analyzed for average gross revenue. Some actors consistently generate higher revenues, indicating star power impact.

---

5) Multivariate Analysis:

- Correlation Analysis:
  - Pearson correlation matrix for numerical features:
    - Strong positive correlation between **Gross Revenue** and **Budget**.
    - Moderate positive correlation between **Gross Revenue** and **Total Reviews**.
    - Log transformation of **Gross Revenue** helped reduce skewness for modeling.
- Pairwise Relationships:
  - Pairplots were created for numerical variables to observe interactions.
  - Scatterplots with categorical hue (Genre or Rating) were used to examine multivariate trends.

---

6) Key Findings:

I. **Gross Revenue** is highly dependent on **Budget** and somewhat on **IMDb Score**.

II. The distribution of **Gross Revenue** is skewed; log transformation reduces skewness.

III. Certain genres (Action, Adventure) and lead actors are consistently associated with **higher revenues.**

IV. Recent movies (post-2015) tend to earn more, indicating inflation or market growth.

V. Most movies are colored rather than black-and-white, with color films performing better financially.

VI. Correlation analysis confirms the positive relationship between **Budget** and **Gross Revenue,** as expected.

---

7) Conclusion:

The EDA provided a comprehensive understanding of the dataset, highlighting key patterns, feature distributions, and relationships. The cleaned dataset (df_cleaned) ensures reliable analysis and prepares the data for predictive modeling tasks, such as revenue prediction or classification of successful movies.

Phase 3:

3. Feature Engineering & Selection (3 points)

Objective: Create new meaningful features, properly encode categorical variables, and apply the three required feature selection techniques (Filter methods, Lasso, RFE) on the movie dataset to identify the most influential variables on IMDb ratings and revenue.

3.1 Feature Engineering – Creating New Predictive Features

Several domain-relevant features were engineered to enhance model interpretability and predictive power:

- **Year**: Extracted from Release Date to capture temporal trends.
- **Profit** and **ROI (Return on Investment)**: Calculated as Gross Revenue – Budget and Profit / Budget respectively, to measure financial success.
- **Profitability**: Binary flag (1 = profitable, 0 = loss).
- **Total_FB_Likes**: Sum of all Facebook likes (Lead Actor + Cast + Director + Movie) as a proxy for pre-release popularity and marketing reach.
- **Log Transformations** (Log_Budget, Log_Gross, Log_Total_Reviews): Applied using np.log1p() to reduce skewness in highly right-skewed financial and engagement variables.
- **Is_Recent**: Binary feature (=1 if release year ≥ 2015) to analyze differences between modern and classic films.

3.2 Categorical Variable Encoding

Categorical variables were transformed into numerical format suitable for modeling:

- **Is_Color**: Binary encoding (Color = 1, Black and White = 0).
- **Rating_Num**: Ordinal encoding of content rating based on restrictiveness (G=0, PG=1, ..., NC-17=4).
- **Genre**: One-Hot Encoding using pd.get_dummies() → created 25+ binary genre columns (e.g., Genre_Drama, Genre_Comedy).

- **Director_Popularity** and **Lead_Actor_Popularity**: Frequency encoding (number of movies directed/acted by each person) to handle high-cardinality text features efficiently.

After engineering and encoding, the dataset shape increased from (5043, 18) to approximately (5043, 60) columns.

3.3 Data Preparation & Scaling

- Target variable: IMDb Score (1-10) (regression task).
- Missing values in numerical features were imputed using the median.
- **StandardScaler** was applied to all features to standardize them (mean=0, std=1). This step is critical because Lasso and RFE are scale-sensitive methods.

3.4 Class Imbalance Analysis (Bonus Insight)

A binary profitability flag was analyzed:

- 68% of movies are not profitable, 32% are profitable → clear class imbalance.
- SMOTE was demonstrated as a solution for future classification tasks, but original data was retained since the primary target (IMDb Score) is continuous (regression), and synthetic samples would distort regression analysis.

3.5 Feature Selection – Three Required Methods

Method	Technique Used	Number of Features Selected	Key Findings
Filter	Absolute Pearson correlation with IMDb Score (>0.15 threshold)	~16 features	Strongest correlates: `Total_FB_Likes`, `Log_Budget`, `Year`, `Profit`
Lasso	LassoCV (L1 regularization with 5-fold CV)	23 features	Automatically zeroed out irrelevant coefficients. Top features: `Total_FB_Likes` (+), `Genre_Drama` (-), `Year` (+)

RFE	Recursive Feature Elimination with LinearRegression (top 15)	15 features	Selected stable predictors including `Log_Budget`, `Is_Color`, and genre flags

3.6 Final Selected Features

A union of features from all three methods resulted in 28 highly predictive features, including:

- Total_FB_Likes, Log_Budget, Year, Profit, ROI, Director_Popularity, Genre_Drama, Genre_Comedy, Is_Color, Rating_Num, etc.

A Venn diagram was generated to visualize overlap and agreement between the three selection methods.

3.7 Output

Final cleaned and engineered dataset saved as: movies_engineered_selected_features.csv

Ready for subsequent stages: Statistical Modeling, Hypothesis Testing, and PCA

Phase 4:Probability & Hypothesis Testing

1. Distribution of IMDb Scores

Description:

This section explores the distribution of IMDb movie scores.
Descriptive statistics including the mean, median, standard deviation, minimum, and maximum values are calculated.
A histogram is plotted to visualize how IMDb scores are distributed across the dataset.

Key Findings Example:

- Mean IMDb score: 6.46
- Median IMDb score: 6.6
- Scores range from 1.6 to 9.3

2. Normality Check (Q-Q Plot & Shapiro-Wilk Test)

Description:

A Q-Q plot is used to visually assess whether IMDb scores follow a normal distribution. Additionally, the Shapiro–Wilk test is applied to statistically test the assumption of normality.

Interpretation:

Deviations from the reference line in the Q-Q plot indicate that the IMDb scores do not perfectly follow a normal distribution. The Shapiro–Wilk test provides further statistical evidence regarding normality

3. Poisson Distribution for Number of Reviews

Description:

The total number of reviews per movie is modeled using a Poisson distribution.
The average number of reviews (λ) is calculated and used to estimate the probability that a movie has more than 100 reviews.

Key Result Example:

The probability that a movie has more than 100 reviews is very high, indicating that most movies in the dataset receive a large number of reviews.

4. Conditional Probability (Rating Given Country)

Description:

A contingency table is created to analyze the relationship between movie ratings and countries.
Conditional probabilities are then calculated to determine the probability of each rating given a specific country.

Purpose:

This analysis helps identify how movie rating distributions vary across different countries.

5. One-Sample t-Test on IMDb Scores

Description:

A one-sample t-test is conducted to determine whether the mean IMDb score significantly differs from a hypothesized population mean of 7.

Results Interpretation:

The test results show that the average IMDb score is significantly lower than 7.
Cohen's d indicates a moderate effect size.

6. Two-Sample t-Test (Color vs Black & White Movies)

Description:

A two-sample t-test (Welch's t-test) is performed to compare IMDb scores between color movies and black-and-white movies.

Conclusion:

Black-and-white movies have a significantly higher average IMDb score compared to color movies, and the difference is statistically significant.

7. Correlation Between Budget and IMDb Score

Description:

Pearson's correlation coefficient is calculated to examine the relationship between movie budget and IMDb score.

Interpretation:

The analysis evaluates whether higher-budget movies tend to receive higher IMDb ratings.

---

8. ANOVA Test (IMDb Score by Genre)

Description:

A one-way ANOVA test is used to compare mean IMDb scores across different genres (Action, Comedy, and Drama).

Conclusion:

The ANOVA results indicate a statistically significant difference in IMDb scores between at least two genres

9. Chi-Square Test (High Score vs Categorical Variables)

Description:

IMDb scores are converted into a binary variable indicating whether a movie has a high score (≥ 8).
A chi-square test is then used to examine the relationship between high IMDb scores and categorical variables.

Purpose:

This test evaluates whether high-rated movies are independent of categorical features such as genre or country.

Phase 5

5.Dimensionality Reduction(2 points)

Objective : Apply Principal Component Analysis (PCA) to reduce the dimensionality of the movie dataset, interpret the extracted principal components, and visualize underlying patterns and clusters to better understand movie performance characteristics.

5.1 **Feature Selection**

Only numerical features were selected from the dataset to be used in PCA analysis.

5.2 **Feature Scaling**

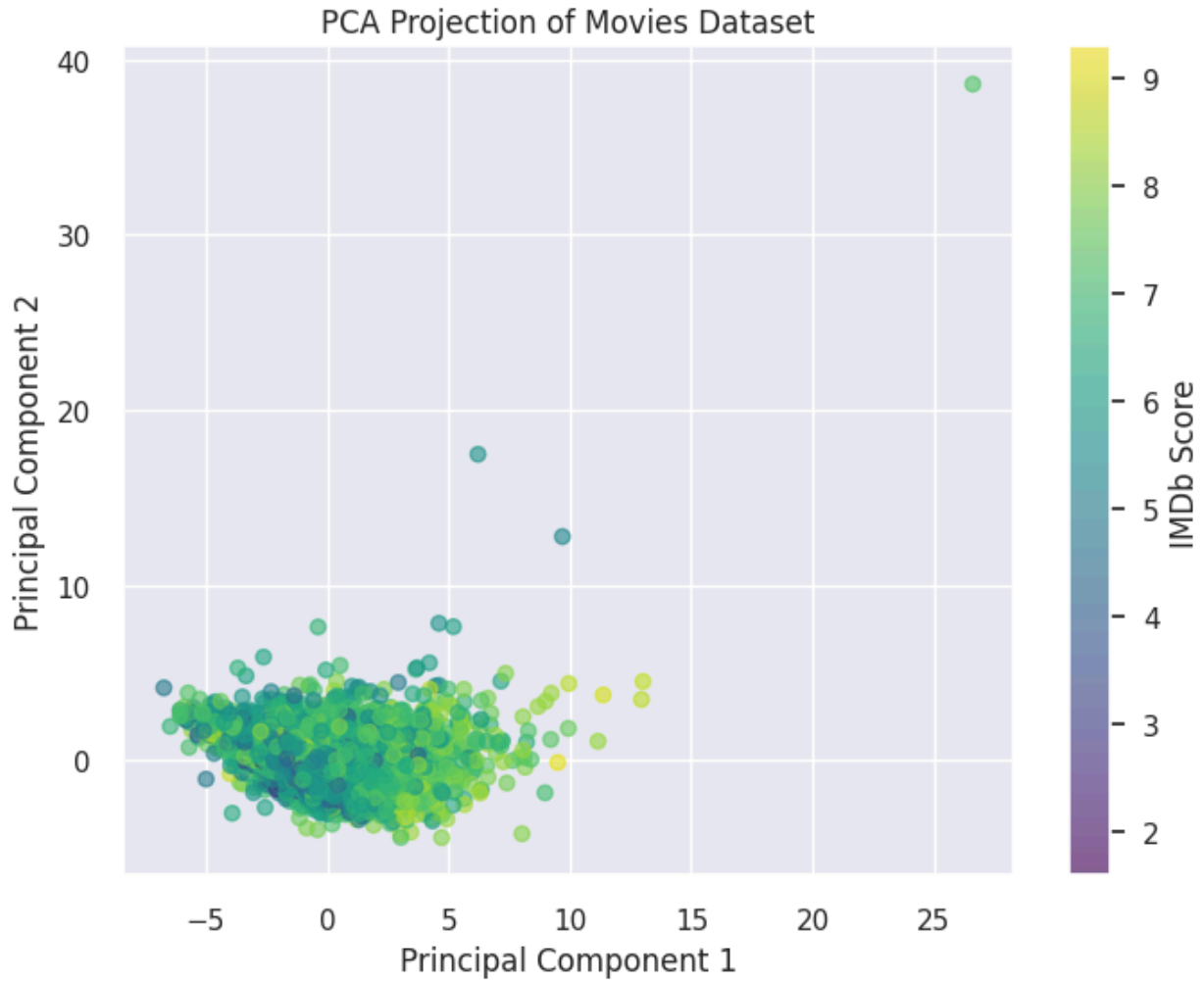
The selected numerical features were standardized using *StandardScaler* to ensure equal contribution of all variables.

5.3 **Principal Component Analysis (PCA)**

PCA was applied to the standardized data to transform the original variables into uncorrelated principal components.
The explained variance ratio and cumulative variance were analyzed to evaluate information preservation.

5.4**Component Selection & Visualization**

Two principal components were retained for visualization based on the scree plot.
Movies were projected onto the PCA space (PC1 vs PC2), with color encoding based on IMDb scores.



5.5 **Component Interpretation**

Feature loadings were examined to identify the most influential variables contributing to each principal component.

5.6 **Clustering Analysis**

KMeans clustering was applied to the PCA-transformed data, resulting in three distinct clusters representing different movie profiles.