COURSE PROJECT

Analysis for IMDb Top 1000 Movies

Application Development (CSF-510)

Made By
Syed Muhammad Meesum Abbas (31486)
Syed Daniyal Hussain (31487)

Table of Contents

Data Analysis Report

Chapter A: System Study / Domain Analysis

1. Business Process and Potential Analytics

Business Process

The dataset represents top-rated IMDb movies, offering valuable insights into the film industry. This data is crucial for stakeholders like film producers, distributors, marketers, streaming platforms, and researchers. Understanding trends in movie ratings, genres, audience preferences, and revenue can aid decision-making processes in multiple domains, such as:

- **Film Production**: Identifying successful genres, runtime patterns, or certifications to optimize future productions.
- **Marketing Strategies**: Analyzing star power and director influence to plan promotions and collaborations.
- **Streaming Platforms**: Selecting popular genres or top-rated movies to attract viewers.
- **Audience Analysis**: Understanding preferences based on IMDb ratings and votes to recommend movies.

Potential Analytics and Their Usefulness

- **Genre Popularity Analysis**: Determine the most common and highest-rated genres.
- **Director and Star Influence**: Evaluate how directors or lead actors correlate with higher ratings.
- **Revenue Patterns**: Analyze gross revenue trends by genre, runtime, or release year.
- **Rating Correlations**: Explore relationships between IMDb ratings, Metascore, and audience votes.
- **Trend Analysis**: Examine how movie characteristics (e.g., runtime, genres) have evolved over decades.

Such analyses can:

- Enhance investment decisions in movies with high potential for success.
- Provide tailored recommendations on streaming platforms.
- Offer insights for academic research or industry reports on movie trends.

2. Dataset Description

a. Dataset Type

This dataset can be categorized as a **clustering and regression dataset**, depending on the analysis:

- **Clustering**: Grouping movies based on attributes like genre, runtime, or ratings.
- **Regression**: Predicting variables such as `Gross` revenue based on features like `IMDB_Rating`, `No_of_Votes`, and `Meta_score`.

b. Data Balance

Certain attributes like `IMDB_Rating` and `No_of_Votes` appear continuous and balanced. However, `Certificate` and `Genre` might show class imbalance:

- **Imbalanced Classes**: Genres or certificates may have fewer representatives, skewing analyses.
- **Consequences**: Imbalances can:
  - Bias machine learning models.
  - Limit insights into underrepresented genres or ratings.

Approaches like oversampling or undersampling might be required to address this issue.

c. Data Composition

Below is the dataset's composition summary:

Attribute Name	*Type*	*Missing Values*	*Feature Importance*
Poster_Link	Categorical	None	Drop
Series_Title	Categorical	None	High
Released_Year	Categorical	None	High
Certificate	Categorical	101	Average
Runtime	Categorical	None	High
Genre	Categorical	None	High
IMDB_Rating	Numerical	None	High
Overview	Text	None	Low
Meta_score	Numerical	157	Average
Director	Categorical	None	Average
Star1	Categorical	None	High
Star2	Categorical	None	Average
Star3	Categorical	None	Average
Star4	Categorical	None	Low
No_of_Votes	Numerical	None	High
Gross	Numerical	169	High

Key considerations include:

- Dropping `Poster_Link` due to low analytical value.
- Prioritizing attributes like `IMDB_Rating`, `No_of_Votes`, and `Gross` for revenue-related insights.
- Treating missing values in `Certificate`, `Meta_score`, and `Gross` to improve analysis quality.

3. Data Source

The dataset for this analysis was sourced from Kaggle. (https://www.kaggle.com/datasets/harshitshankhdhar/imdb-dataset-of-top-1000-movies-and-tv-shows)

Chapter B: Data Cleaning

1. Data Type Analysis and Transformation

- Converted `Released_Year` to numeric type.
- Extracted numeric values from `Runtime`.
- Removed commas from `Gross` and converted it to numeric.

2. Missing Value Handling

- **Numerical Columns**: Missing values in `Meta_score`, `Gross`, and `Runtime` were filled with the median value to avoid skewing the data.
- **Categorical Columns**: Missing values in `Certificate` were filled with the mode (most frequent value).

Result

All missing values were handled, and the cleaned dataset was saved for further processing.

Chapter C: Exploratory Data Analysis

1. Univariate Analysis

- **Boxplots and Distribution Plots** were created for numerical variables (`IMDB_Rating`, `Meta_score`, `Gross`, `Runtime`, `No_of_Votes`).
- **Normality Tests** were conducted using the Shapiro-Wilk test.

Observations:

- `IMDB_Rating` and `Meta_score` showed nearly normal distributions.
- `Gross` and `Runtime` exhibited skewed distributions.

2. Bivariate Analysis

- **Scatterplots and Correlation Analysis** were conducted between numerical variables and `IMDB_Rating`.

Results:

- `No_of_Votes` showed the strongest correlation with `IMDB_Rating` (0.495).
- `Meta_score` and `Runtime` had weak correlations.
- `Gross` showed minimal correlation with `IMDB_Rating`.

Chapter D: Dashboard Creation

A dashboard was created using Dash with the following functionalities:

a. Landing / Welcome Page

- The landing page provides an overview of the business use case and allows users to upload their dataset.
- Displays the dataset in a grid format for easy inspection.

b. Univariate Analysis Page

- Users can select any numerical attribute to visualize its distribution through boxplots and histograms.
- Options for viewing normality test results.

c. Bivariate Analysis Page

- Allows users to select pairs of attributes to visualize their relationships.
- Scatterplots and correlation coefficients are dynamically updated based on user input.

The dashboard integrates interactive elements for seamless exploration of univariate and bivariate analysis.

Chapter E: Preprocessing

1. Discretization

- `Gross` and `Runtime` were discretized into 3 uniform bins using KBinsDiscretizer.

2. Normalization

- Remaining numerical attributes (`IMDB_Rating`, `Meta_score`, `No_of_Votes`, `Released_Year`) were normalized using Min-Max scaling.

3. One-Hot Encoding

- Categorical variables (`Certificate`, `Genre`, `Director`) were one-hot encoded.

4. Train-Test Split

- The dataset was split into training (70%) and testing (30%) sets.

Conclusion

The cleaned and processed dataset provided actionable insights into IMDb top-rated movies. The dashboard offers a user-friendly interface for exploring trends and relationships within the data, while the preprocessing steps ensure readiness for further machine learning applications or analysis.