Student Name: Aravind sakinala
Student ID: 23086372
Git Hub:https://github.com/Aravindyadav2705/Aravind-sakinala-.git

Introduction

The IMDB Top 2000 Movies Dataset can be used as a goldmine for beginner data scientists experimenting with regression and recommendation system projects. It contains information such as movie ratings, genres, and release years, among other things. This dataset is useful for analyzing trends, rating predictions, and even for building personalized movie recommendation engines through hands-on machine learning techniques.

Distribution of Movie Genres

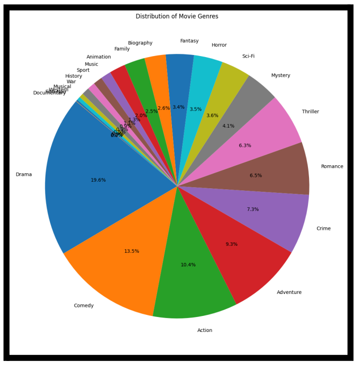


Figure 1: pie chart

The pie chart is a representation of movie genres in the dataset distribution, giving an overview of genre popularity. Each slice, therefore, represents a genre's percent of the total dataset as calculated after splitting multi-genre entries. This chart aids in identifying dominant genres while offering insights into viewer preferences, which helps in carrying out trend analysis and deciding on movie-related applications.

IMDB Rating vs Gross Earnings

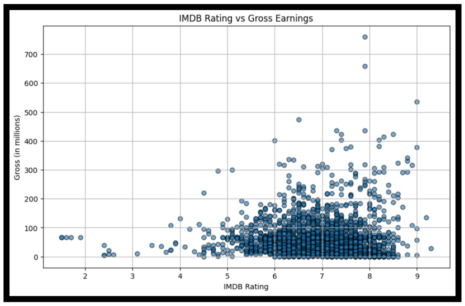


Figure 2: Scatter plot

The scatter plot represents the association between IMDB ratings and the gross earnings of movies. Every point on the scatter plot corresponds to a movie, representing its rating and earnings. Trends such as whether higher-rated movies earn more can be identified from the visualization. Patterns or outliers can identify the audience's preferences, thereby making financial predictions and guiding decisions in the industry.

Correlation Heat map

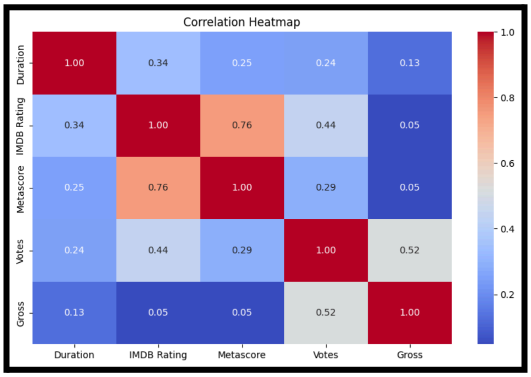


Figure 3: Heat map

It graphically represents the correlation matrix of the data, showing numerical relationships like IMDB ratings, gross earnings, and runtime between features. This tool assists in selecting the most significant features while it shows dependency and which features are affecting predictive models for an accurate representation or depicting dynamics in datasets.

Elbow Method for Optimal K

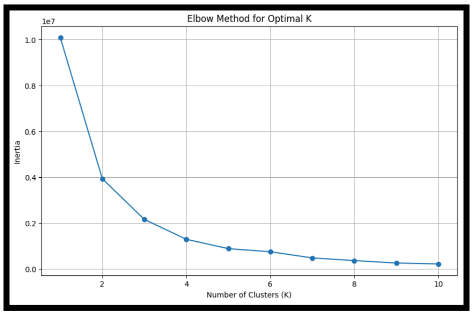


Figure 4: Elbow plot

The Elbow Method plot will help determine the optimal number of clusters (K) to use for K-means clustering. It plots the inertia (sum of squared distances) against different values of K. The "elbow" point is where returns are diminishing, and this point is the ideal K for balancing model accuracy and efficiency, which is critical to effective clustering in data analysis.

K-Means Clustering

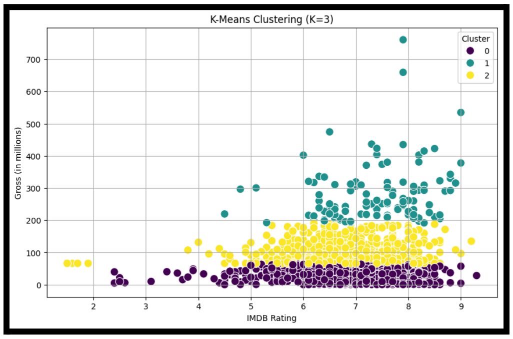


Figure 5: K-Means Clustering

In a visualization, K-Means clusters movie rows of information, classifying movies into three groups based on their ratings from IMDB and gross revenue. Each colored group represents a set of movies with common characteristics; hence, the approach of this method will reveal any hidden patterns or relationships of data, giving insights toward segmentation, trend analysis, and targeted decision-making.

Linear Regression: IMDB Rating vs Gross

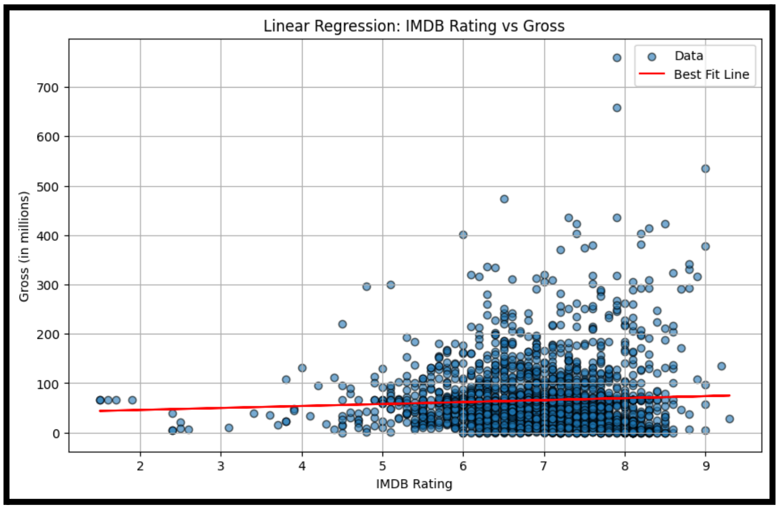
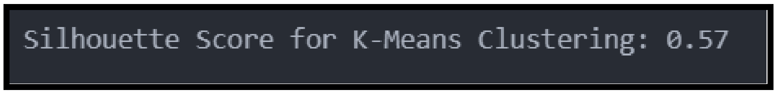


Figure 6: Linear regression

The linear regression plot illustrates the relationship of IMDB ratings with the gross earnings. The scatter plot contains the data points while the red line is used to denote the best fit. Analysis through this plot tends to emphasize trends, showing how ratings tend to influence earnings, hence allowing for predictions that offer useful strategic planning in movies.

Silhouette Score



```
Silhouette Score for K-Means Clustering: 0.57
```

Figure 7: Silhouette Score

The Silhouette Score is the quality of K-Means clustering, which calculates how well data points fit within their clusters compared to others. A score of 0.57 indicates a moderate level of clustering performance and shows that the clusters are well-separated with reasonable compactness. This metric refines clustering approaches for better segmentation and insights from data.