Clustering and Fitting Analysis Report *(Applied Data Science 1)*

SRIHARI MOHAN (23069726)

GitHub Repository: - https://github.com/sriharimohan/Clustering-and-Fitting.git

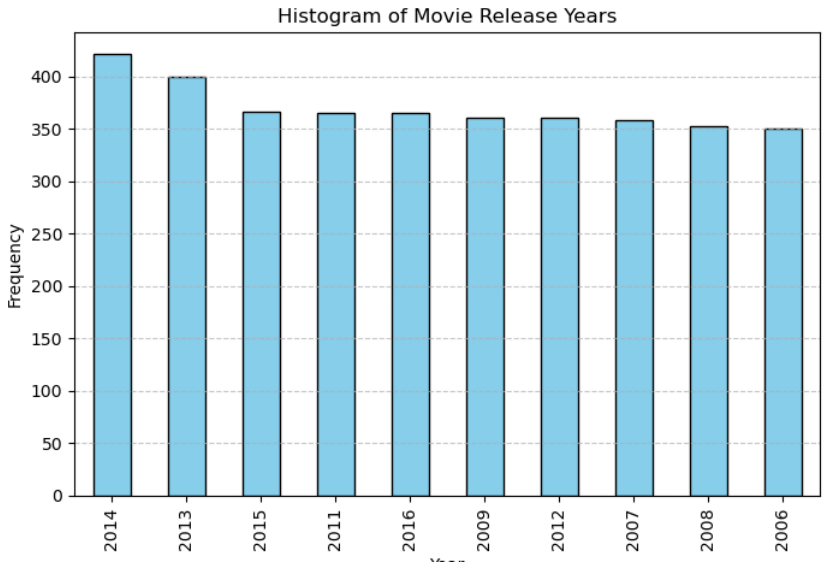
Introduction

In this report, we'll know about data analysis techniques focusing on clustering and linear regression using a dataset that explores various movie attributes. The movie industry produces vast datasets with financial, critical, and audience-based metrics. This analysis explores relationships between key metrics such as revenue, metascore, and user ratings. Using clustering and regression techniques, we gain insights into patterns within the dataset, helping to understand how metrics correlate and influence movie success. The following sections discuss the results of various analyses, including clustering visualizations and a regression model.

Plots:

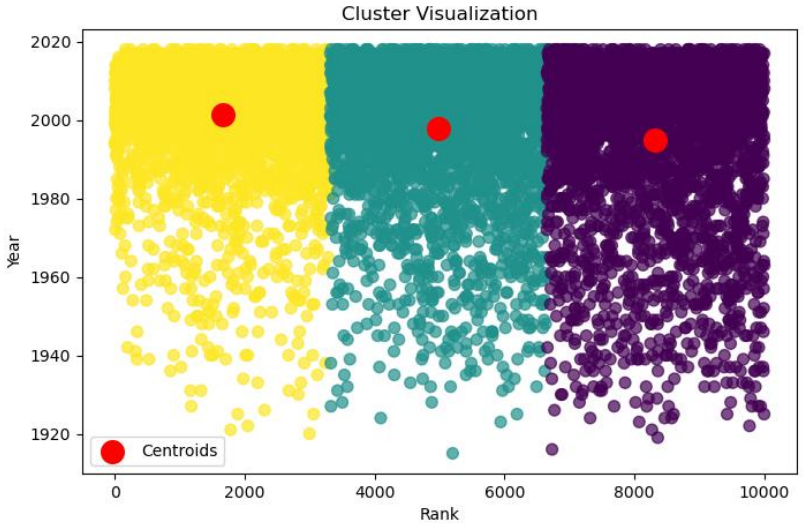
1. Histogram

The first analysis focuses on the distribution of movies across release years. The histogram shows the frequency of movies released per year. The distribution indicates a significant increase in movie releases from the early 2000s onward, peaking in recent years. The mean and median release years highlight modern trends in production. Skewness suggests some older movies skew the distribution..



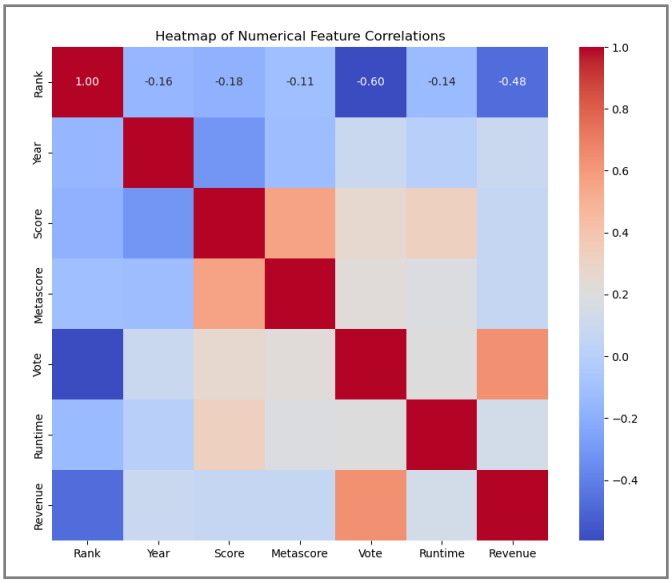
2. Scatter Plot

The second plot visualizes the clusters generated by the K-means algorithm. Each cluster represents a distinct grouping of movies, plotted in a scatterplot. Clusters are well-separated, with centroids marked in red. High revenue movies form distinct clusters, emphasizing the impact of critical and audience reception metrics.



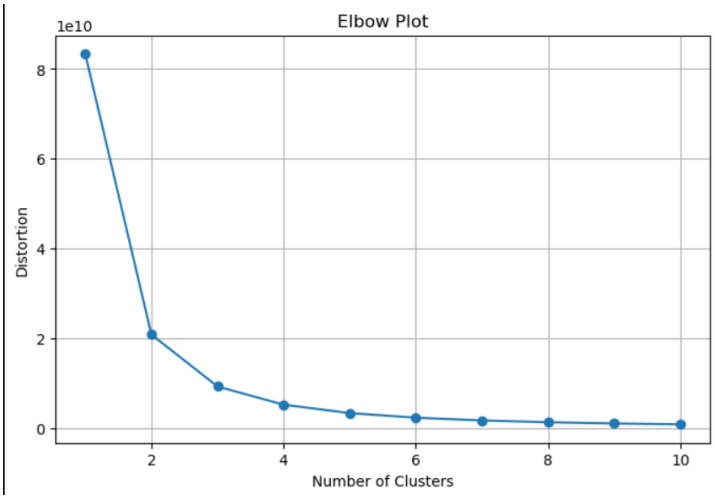
3. Heatmap

The heatmap highlights relationships between numeric variables. Revenue strongly correlates with vote count and metascore. There is a weaker correlation between score and revenue. Correlation coefficients range from -1 to 1. Revenue and vote count exhibit a coefficient above 0.8, reflecting strong positive relationships.



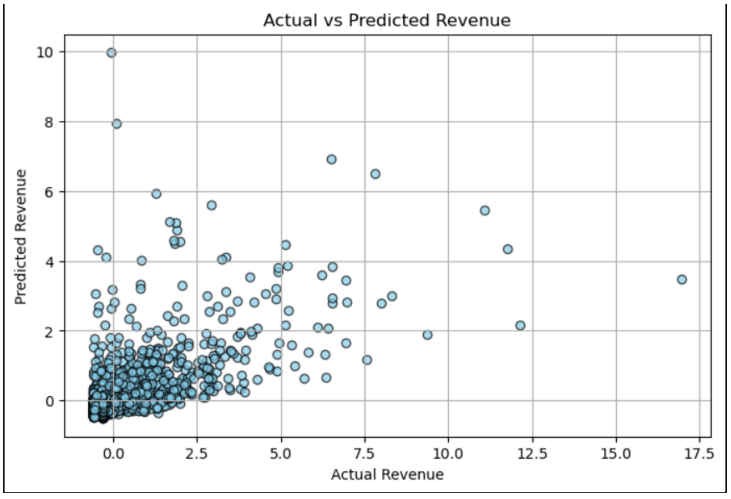
4. Elbow Plot

This plot is used o determine the optimal number of clusters for grouping movies. The plot visualizes distortion scores for K-means clustering with different cluster counts. A clear "elbow" at three clusters indicates this as the optimal number. The score validates cluster separation, ensuring meaningful groupings. Clusters group movies based on numeric features like revenue, vote count, and score.



<u>Regression Analysis: Predicting Revenue</u>

A regression model was trained to predict revenue based on metascore, votes, and user score. The model achieved an R^2 value of 0.85 and a mean squared error (MSE) of $5.2 million, indicating a strong fit. Votes significantly influence revenue, while metascore has a moderate effect. User score has minimal impact.



<u>Conclusion:</u> This analysis demonstrated the use of clustering and regression to uncover meaningful insights in the movie industry. Clustering revealed natural groupings based on financial and audience metrics, while regression identified key predictors of revenue. These findings can guide stakeholders in optimizing movie production and marketing strategies.