PROJECT REPORT

on

"MOVIE RATING ANALYSIS"

Submitted to

KIIT Deemed to be University

BACHELOR'S DEGREE IN INFORMATION TECHNOLOGY

BY

ASHISH SHARMA	21052402
SAURABH SUSHANT SRIVASTAVA	21051335
NISHCHAL KUMAR SINGH	21051147

UNDER THE GUIDANCE OF

Prof.Abinas Panda

SCHOOL OF COMPUTER ENGINEERING

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

BHUBANESWAR, ODISHA - 751024

MARCH 2024

ACKNOWLEDGEMENT

I would like to express my sincere gratitude towards my college KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY(KIIT) for providing the opportunity to work on this insightful project.

I extend my thanks to the Kaggle community for making the IMDb dataset available, which served as the foundation for my analysis.

I am also grateful to my professor Mr.Abinas Panda for their guidance, support, and valuable feedback throughout the project. His expertise and encouragement was instrumental in shaping my approach and refining my analysis techniques.

Finally, I would like to acknowledge the broader community of data enthusiasts and industry professionals whose insights and resources have enriched my understanding of data analysis and its applications in real-world scenarios.

Thank you to everyone involved in making this project possible.

ABSTRACT

This report presents a comprehensive analysis of movie ratings using the IMDb dataset obtained from Kaggle. The dataset includes information about various attributes of movies such as their title, genre, run-time, ratings, meta-scores, certificates, and gross revenue.

The analysis begins with data pre-processing steps, including handling missing values, cleaning categorical variables, and exploring the distribution of different attributes. Visualizations such as histograms, box plots, and scatter plots are utilized to understand the distribution and relationships among variables.

Key insights are derived from the analysis, including trends in movie ratings over different decades, the impact of run-time on ratings, and the correlation between meta-scores and IMDb ratings. Furthermore, the report explores the distribution of ratings across different genres and identifies the top-rated movies in each genre.

The findings of this analysis provide valuable insights into the factors influencing movie ratings and audience preferences. This information can be leveraged by filmmakers, production studios, and streaming platforms to understand audience preferences better, optimize content creation strategies, and enhance user engagement.

Overall, this report serves as a valuable resource for anyone interested in understanding the dynamics of movie ratings and their underlying factors in the entertainment industry.

INTRODUCTION

Movie ratings play a crucial role in the entertainment industry, influencing audience choices, critical acclaim, and box office success. Understanding the factors that contribute to movie ratings is essential for filmmakers, production studios, and streaming platforms to create content that resonates with audiences and maximizes engagement.

In this report, we conduct a comprehensive analysis of movie ratings using the IMDb data-set sourced from Kaggle. IMDb (Internet Movie Database) is one of the most comprehensive and authoritative sources for movie information, providing data on a wide range of attributes for thousands of movies.

The primary objective of this analysis is to uncover insights into the factors that influence movie ratings, including genre, run-time, critical acclaim (meta-scores), and commercial success (gross revenue). By exploring these factors, we aim to provide valuable insights that can inform content creation strategies, marketing decisions, and audience targeting efforts in the entertainment industry.

Through data pre processing, exploratory data analysis, and visualization techniques, we delve into the distribution, relationships, and trends among different variables in the data-set. Additionally, we identify top-rated movies across various genres and examine how ratings vary across different decades.

This report serves as a comprehensive resource for stakeholders in the entertainment industry, offering actionable insights that can drive informed decision-making and enhance the overall movie-watching experience for audiences worldwide.

PROBLEM STATEMENT

The objective of this project is to perform a comprehensive analysis of movie ratings using IMDb data. The dataset contains information about various aspects of movies, including genres, ratings, runtime, release year, and more. The goal is to explore trends, patterns, and relationships within the data to gain insights into factors influencing movie ratings.

Major task performed are:

- Analyze the distribution of movie ratings and identify any trends over time.
- Investigate the relationship between movie genres and ratings to determine which genres tend to receive higher ratings.
- Explore the impact of runtime, release year, and other variables on movie ratings.
- Identify top-rated movies and genres based on various metrics such as average rating, revenue, and popularity.
- Visualize the findings using appropriate plots and charts to facilitate interpretation and presentation of results.

DATASET OVERVIEW

- The dataset used for analysis is sourced from Kaggle and contains information about movies collected from IMDb.
- It comprises various attributes such as movie titles, genres, ratings, runtime, release year, revenue, and Metascore.
- The dataset encompasses a wide range of movies across different genres, release years, and revenue levels, making it a comprehensive source for movie rating analysis.
- The dataset contains:

  i.Time Period: 1920 to 2020

  ii.Number of Movies: 1000

  iii.Included Columns: Title, Genre, Rating, Runtime, Metascore, Revenue, Year, Certificate

Dataset:

	Rank	Movie_name	Year	Certificate	Runtime_in_min	Genre	Metascore	Gross_in_$_M	Rating_from_10
0	1	The Shawshank Redemption	1994	R	142	Drama	81.0	28.34	9.3
1	2	The Godfather	1972	R	175	Crime, Drama	100.0	134.97	9.2
2	3	The Dark Knight	2008	PG-13	152	Action, Crime, Drama	84.0	534.86	9.0
3	4	The Lord of the Rings: The Return of the King	2003	PG-13	201	Action, Adventure, Drama	94.0	377.85	9.0
4	5	Schindler's List	1993	R	195	Biography, Drama, History	94.0	96.9	9.0
...	...	...	...	...	...	...	...	...	...
995	996	Sabrina	1954	Passed	113	Comedy, Drama, Romance	72.0	69.25089901477833	7.6
996	997	From Here to Eternity	1953	Passed	118	Drama, Romance, War	85.0	30.5	7.6
997	998	Snow White and the Seven Dwarfs	1937	Approved	83	Animation, Adventure, Family	96.0	184.93	7.6
998	999	The 39 Steps	1935	Approved	86	Crime, Mystery, Thriller	93.0	69.25089901477833	7.6
999	1,000	The Invisible Man	1933	TV-PG	71	Horror, Sci-Fi	87.0	69.25089901477833	7.6

1000 rows × 9 columns

DATA PREPROCESSING

- Upon loading the dataset, initial preprocessing steps were undertaken to ensure data quality and consistency.
- Missing values were addressed through appropriate imputation techniques, with mode and mean being used for categorical and numerical columns, respectively.In my dataset columns such as certificate,metascore,and gross revenue were having missing values so we filled those missing values with mean and mode as per requirement of column.
- Data types were adjusted to ensure accurate representation and analysis, with numeric data being converted to the appropriate format.
- Redundant columns, such as 'Rank', were removed to streamline the dataset and focus on relevant attributes for analysis.
- Removing duplicates : There were no duplicates present in our dataset.
- Inconsistent data : There were some inconsistent data present in years column of our dataset which we transformed into standard form.
- Dealing with categorical data : In order to deal with categorical data , combined genre column was splitted into separate columns using one hot encoding.

EXPLORATORY DATA ANALYSIS (EDA)

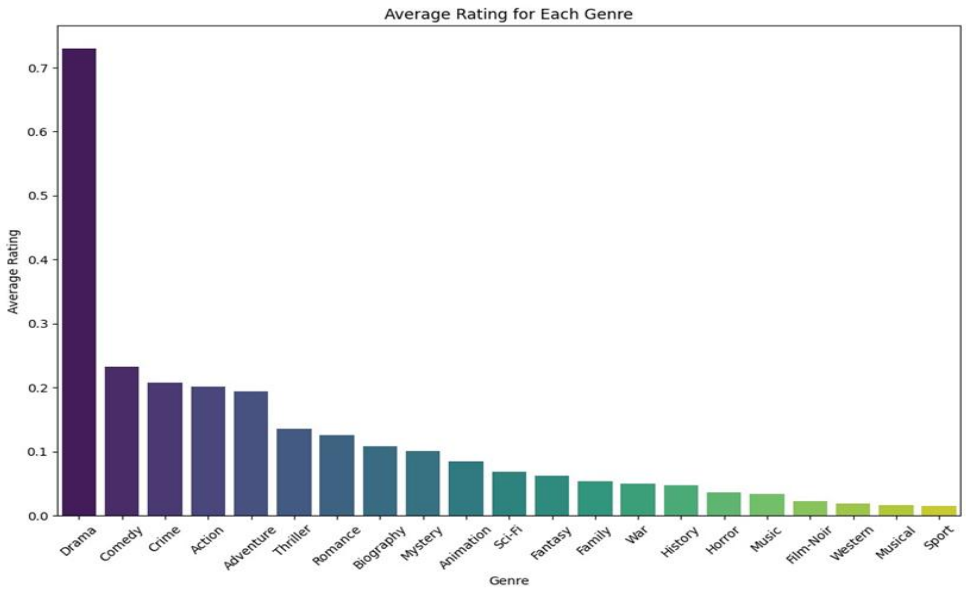
- EDA was conducted to gain insights into the distribution, relationships, and patterns within the dataset.
- Descriptive statistics and visualizations such as histograms, box plots, and scatter plots were employed to explore various attributes and uncover trends.
- Key findings from the EDA phase included the distribution of movie ratings, prevalence of different genres, correlation between runtime and ratings, and the impact of certificates on ratings

GENRE ANALYSIS

- Genres were encoded as binary variables using one-hot encoding to facilitate analysis.

- The count of movies and average ratings for each genre were visualized to identify trends and preferences among audiences.

- Drama emerged as the top-rated genre, indicating its popularity and critical acclaim among viewers, followed by genres like Comedy, Action, and Adventure.

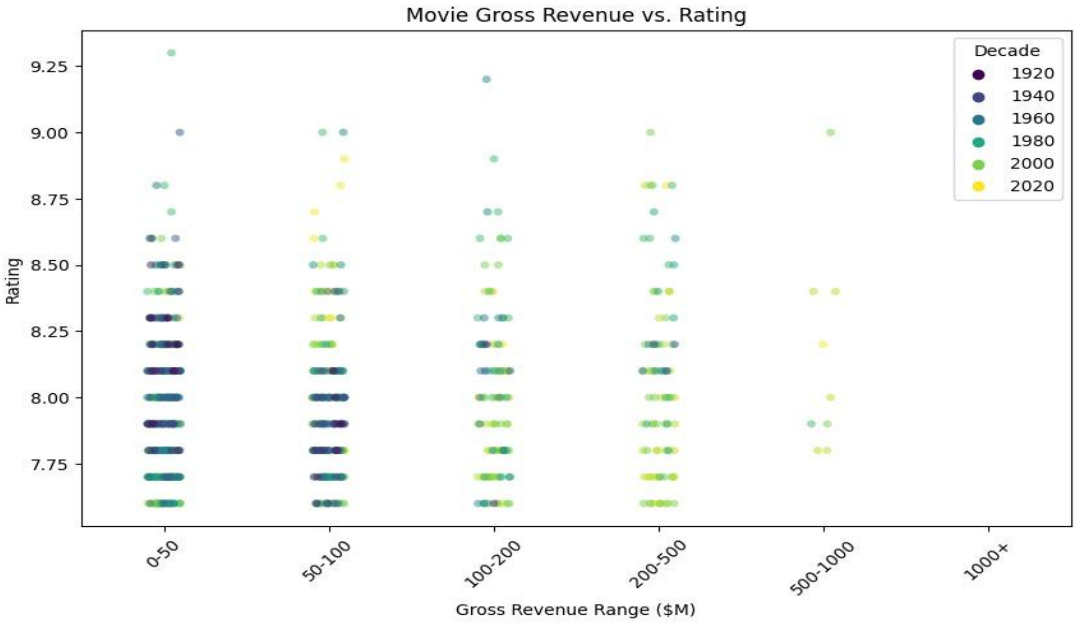


REVENUE ANALYSIS

- Revenue data was categorized into ranges to analyze its relationship with ratings and other attributes.

- Decade-wise analysis provided insights into changing trends and preferences over time, with the 2000s and 2010s showing significant growth in movie revenue.

- Scatter plots were utilized to examine the interplay between movie runtime, revenue, and ratings, revealing potential patterns and correlations that could inform business strategies and production decisions.

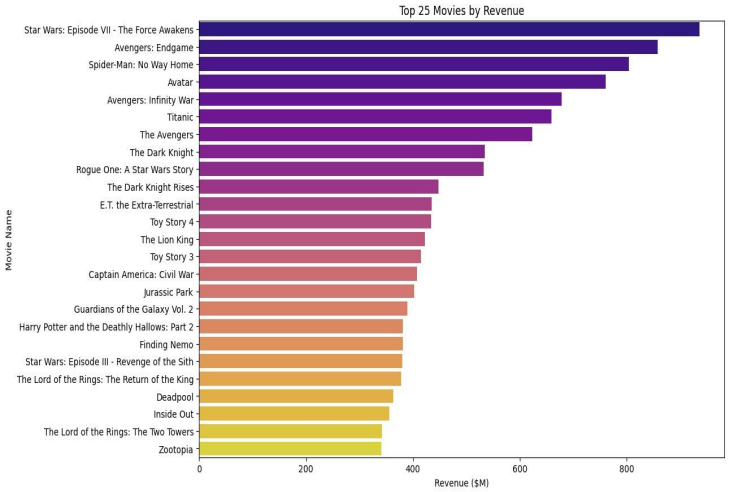


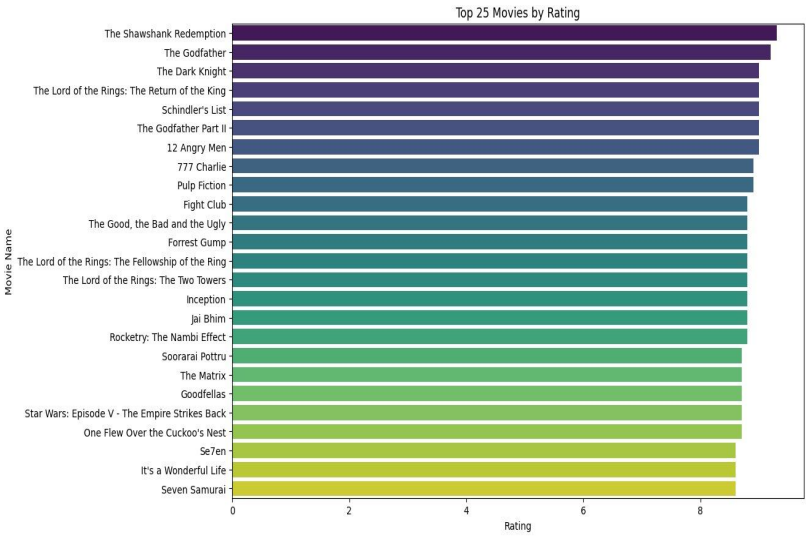
TOP MOVIES ANALYSIS

- Top-performing movies were identified based on ratings, revenue, and Metascore, providing insights into the most successful and acclaimed movies in the dataset.

- Visualizations showcased the top movies across various metrics, highlighting their impact and success within the industry and offering valuable benchmarks for aspiring filmmakers and production studios.

- Genre-specific analyses unveiled the top-rated movies in each genre, providing a diverse perspective on audience preferences and cinematic achievements across different genres and themes





CONCLUSION AND RECOMMENDATIONS

- The analysis provided valuable insights into the factors influencing movie ratings and success, including genre, runtime, revenue, and critical reception.

- Stakeholders in the film industry can leverage these insights to inform content creation, marketing strategies, and audience engagement efforts, thereby maximizing the chances of success for their projects.

- Recommendations may include focusing on genres with high average ratings, optimizing runtime based on audience preferences, and strategizing revenue generation tactics for maximum impact and profitability.

FUTURE DIRECTIONS

- Future analysis could delve deeper into sentiment analysis of user reviews to understand audience preferences and sentiments towards specific movies and genres.

- Predictive modeling of movie ratings and box office performance could be explored to forecast the success of upcoming movies and guide investment decisions.

- Incorporating additional data sources such as social media trends, audience demographics, and industry reports could enhance the depth and accuracy of analysis, providing more actionable insights for stakeholders and industry professionals.