A PRELIMINARY MINI PROJECT REPORT ON

"Movie Recommendation System"

SUBMITTED TOWARDS THE PARTIAL FULFILLMENT OF THE REQUIREMENTS OF

BACHELOR OF ENGINEERING (S.Y. B. Tech.)

Academic Year: 2024-25

By:

Pranali Patil	**123B1B221**
Sayali Pawar	**123B1B229**
Om Patil	**123B1B219**
Pratishtha Chandrakar	**123B1B233**

Under The Guidance of

Dr. Chetan Chauhan

DEPARTMENT OF COMPUTER ENGINEERING,

PIMPRI CHINCHWAD COLLEGE OF ENGINEERING

SECTOR 26, NIGDI, PRADHIKARAN

PIMPRI CHINCHWAD COLLEGE OF ENGINEERING DEPARTMENT OF COMPUTER ENGINEERING

CERTIFICATE

This is to certify that, the project entitled

“**Movie Recommendation System**” is successfully carried out as a Skill Development Laboratory I mini

project and successfully submitted by following students of “**PCET's**

Pimpri Chinchwad College of Engineering, Nigdi, Pune-44”.

Mini-Project Report: Movie Recommendation System

Under the guidance of Dr. Chetan Chauhan
In the partial fulfillment of the requirements for the S.Y. B. Tech.

(Computer Engineering)

By:

Pranali Patil	**123B1B221**
Sayali Pawar	**123B1B229**
Om Patil	**123B1B219**
Pratishtha Chandrakar	**123B1B233**

Dr. Chetan Chauhan

Project Guide

Chapter 1: Introduction

1.1Background

A. **Growth of Entertainment Industry**: The entertainment industry has seen exponential growth, driven by technological advancements and greater global connectivity. This shift has fundamentally changed how audiences discover and consume content, particularly through streaming services and digital platforms that offer access to vast libraries of media. These advancements allow content to reach a worldwide audience, breaking down traditional regional barriers. As a result, the industry is now a major economic force with both direct and indirect contributions to other sectors, from tourism to technology development. This growth is a primary factor driving demand for innovative data-driven approaches to audience engagement.

B. **Role of Data Science**: Data science has become a cornerstone of modern decision-making in the film industry, influencing a range of areas from marketing to content development. By leveraging data analysis, companies can gain insights into complex audience behaviors, allowing for precise targeting and personalized content recommendations. Data-driven strategies have led to a more efficient allocation of resources, ensuring that marketing efforts align closely with viewer interests. The insights generated by data science can even impact creative decisions, helping filmmakers align projects with current audience trends, ultimately leading to a more engaged and satisfied viewer base.

C. **Significance of Movie Ratings**: Ratings are not just a measure of viewer satisfaction; they play a critical role in shaping a movie's commercial success and long-term impact. High ratings can drive box office sales, secure lucrative streaming deals, and bolster a film's reputation within the cultural landscape. Moreover, ratings influence content licensing and promotional decisions, with studios often prioritizing high-rated films for greater distribution. This emphasis on ratings has elevated the importance of data insights, as companies seek to understand the factors that consistently correlate with favorable public opinion and critical acclaim.

D. **Recommendation Systems**: Machine learning models have revolutionized recommendation systems, which help users discover content aligned with their tastes. Platforms like Netflix, Amazon Prime, and Disney+ rely heavily on algorithms that analyze user preferences to offer tailored content. These recommendation engines not only enhance user satisfaction but also increase engagement and retention, as users are more likely to stay on a platform where they consistently find appealing content. Additionally, these systems have financial benefits for platforms, as they can strategically promote their own content or licensed media based on an understanding of user preferences.

E. **Influence of Measurable Attributes**: Movie ratings are influenced by a range of measurable factors such as cast, director, genre, and duration, even though subjective elements like storytelling also play a role. By focusing on quantifiable attributes, this project aims to establish baseline predictors of ratings, providing an empirical foundation for rating predictions. Attributes like cast and genre often have a significant correlation with ratings; for example, movies featuring well-known actors or certain genres, like adventure or drama, may generally receive more favorable ratings. These insights allow for a structured analysis of audience preferences.

   Challenges with Rating Prediction: Predicting ratings is complex, as many subjective and contextual factors influence them. Storytelling, cultural relevance, and even timing play roles that are hard to quantify. By focusing on quantifiable features such as genre, duration, and cast, this project establishes a structured model for rating predictions, providing a practical baseline. Although such a model may not capture every nuance, it offers valuable insights, forming the basis for future research that could include more complex features, like sentiment analysis from reviews, to improve predictive power.

F. **Linear Regression Model**: Linear regression is chosen for its simplicity, interpretability, and effectiveness in predicting continuous variables, like movie ratings. It allows for a straightforward analysis of how various factors, such as genre and cast, influence ratings. Additionally, linear regression is a good choice for initial exploration, as it provides a clear view of the relationship between predictors and ratings. Using this model, the project seeks to demonstrate the predictive power of basic attributes, offering a foundation for potentially

1.2 Problem Statement

Predicting movie ratings is challenging due to the myriad of factors influencing audience perceptions and reviewer evaluations. This project attempts to tackle this issue by developing a machine learning model that forecasts movie ratings based on attributes such as genre, director, and cast. The dataset contains features prone to missing values, typographical errors, and formatting inconsistencies, making data cleaning and preprocessing critical steps for accurate analysis.

The objective is to create a reliable model that can predict the IMDb rating of a movie based on easily obtainable characteristics. Accurate prediction could aid various stakeholders, such as streaming services aiming to predict future movie ratings and production houses attempting to maximize audience satisfaction. Additionally, by conducting univariate, bivariate, and multivariate analyses, we can reveal hidden patterns and correlations in the data, providing more profound insights into movie success factors.

1.3 Project Objectives

❖ **Project Overview**

- **Objective**: Analyze movie data to identify factors influencing IMDb ratings and build a predictive model for estimating ratings of new movies based on selected attributes.
- **Dataset**: Contains information on movie characteristics such as genre, cast, director, duration, and release year, which will be analyzed to understand their relationship with IMDb ratings.
- **Significance**: Given the popularity and commercial impact of movie ratings, this project offers actionable insights for industry professionals to better align with audience preferences and maximize movie success potential.

❖ **Project Phases**

1. **Data Cleaning and Preparation**
    - **Data Quality**: Check for inconsistencies, missing values, and irrelevant columns that could impact analysis accuracy.
    - **Data Formatting**: Standardize data types (e.g., convert date strings to date objects) and ensure uniform capitalization and formatting.
    - **Handling Multi-Valued Attributes**: Address attributes like genre and cast by using techniques such as one-hot encoding for genre and popularity metrics for cast.
    - **Normalization**: Normalize numerical features like duration to improve model performance by aligning feature ranges.
2. **Exploratory Data Analysis (EDA)**
    - **Visualizations**: Use histograms, scatter plots, heatmaps, and box plots to understand attribute distributions and identify trends and correlations.

- **Genre and Director Influence**: Explore the impact of genres (e.g., drama or horror) and directors on ratings, testing hypotheses like the "star power" effect.
- **Duration Analysis**: Investigate how movie length influences ratings, examining if longer durations correlate with higher ratings.
- **Trend Analysis**: Analyze ratings over time by release year to uncover patterns, such as audience preferences for classic or modern movies.

3. **Predictive Modeling**
   - **Model Choice**: Start with linear regression to predict IMDb ratings, leveraging its interpretability and effectiveness for regression tasks.
   - **Model Training and Testing**: Split data into training and test sets, then assess performance using metrics like MAE, MSE, and R-squared (R^2).
   - **Feature Importance**: Analyze model coefficients to determine which features significantly impact ratings (e.g., the influence of the director or lead actors).
   - **Model Optimization**: If performance is insufficient, consider more advanced models (e.g., decision tree regression or random forests) for improved accuracy.
4. **Insights Generation**
   - **Strategic Recommendations**: Provide insights on genre preferences, casting, and production elements that correlate with higher ratings, aiding studios in decision-making.
   - **Predictive Simulations**: Use the model to estimate potential outcomes for new movies, helping studios experiment with feature combinations to optimize success potential.
   - **Recommendation System Enhancement**: Streaming platforms can use this model to prioritize movies with higher predicted ratings, improving user engagement.

1.4 Motivation and Scope

The motivation behind this project stems from a fascination with the role that data science plays in transforming subjective domains like movies and media. Ratings serve as a proxy for audience approval, affecting revenue, popularity, and legacy. Yet, predicting ratings involves analyzing an array of subjective and objective factors—from storyline and casting to genre and duration.

In recent years, data-driven decisions have reshaped industries such as retail, healthcare, and manufacturing. The entertainment industry is no different, with streaming giants heavily investing in personalized recommendations and content creation based on viewing data. By developing this project, we hope to contribute a framework that can assist in predicting the success of a movie even before release. For students and practitioners, this project serves as a learning opportunity to understand real-world data and gain hands-on experience in data exploration, cleaning, and model-building.

The scope of this project is comprehensive but constrained to publicly available data and foundational predictive techniques. Specifically, our dataset contains core information about movies, and we focus on a linear regression model for simplicity and interpretability. However, there are certain limitations inherent to the project.

1. **Limited Features:** The dataset does not contain advanced factors like production budget, marketing spend, or critical reception, which could also impact ratings. Our analysis is thus limited to the data available and does not account for such variables.

2. **Simplistic Model:** The predictive model used is linear regression, which is interpretable but may not capture complex, nonlinear relationships as effectively as other machine learning models. Advanced models like Random Forest or Neural Networks might improve predictive performance, but they also require more extensive data preprocessing and tuning.

3. **Historical Data Focus:** The project is rooted in historical data and lacks real-time analysis capabilities. As a result, it is most applicable in a retrospective context, suitable for understanding general trends rather than predicting ratings for future releases.

4. **Generalization:** The model is trained on a limited dataset, which may not generalize to all movie genres or regions due to regional preferences and cultural differences in film ratings.

Chapter 2: Data Collection

2.1 Dataset Information

For this project, the dataset was sourced from IMDb, the Internet Movie Database, which is a renowned source for detailed information on movies, TV series, and entertainment content. IMDb is widely used by viewers, critics, and industry professionals to check ratings, cast information, and release dates. This extensive database provides a wealth of information on various attributes related to movies, making it a valuable resource for data analysis and machine learning projects focused on entertainment.

To access this data in a structured format, the dataset was obtained from Kaggle, a well-known platform for data science and machine learning competitions, resources, and datasets. Kaggle hosts numerous publicly accessible datasets, including those derived from IMDb, which have been pre-processed by contributors to ensure cleanliness and ease of use. Using Kaggle as the source provides several benefits, such as ensuring data accessibility, quality, and reliability, as datasets on this platform are often curated and regularly updated by the community. This particular dataset includes numerous movie-specific attributes that will aid in building and training a predictive model aimed at estimating IMDb ratings based on movie characteristics.

This dataset is categorized as a secondary dataset, meaning it was collected and compiled by someone other than the original creator of the content (IMDb in this case). Secondary datasets are typically available for reuse, often under certain conditions, and are ideal for research and analysis when direct data collection from primary sources is impractical or unnecessary. A primary dataset, in contrast, is data collected firsthand by the researcher through direct surveys, experiments, or observations, which often allows for greater control over variables and data quality. However, primary data collection can be time-consuming and costly, which is why secondary data sources like IMDb are advantageous for large-scale analysis projects such as this one.

Using a secondary dataset from Kaggle aligns with the project's goals as it provides a high-quality foundation for exploring movie attributes and their impact on ratings without needing direct access to IMDb's proprietary data. Moreover, working with a secondary dataset expedites the data analysis process and allows us to focus on feature engineering, model building, and predictive analysis.

2.2 Data Attributes

The dataset contains a variety of attributes relevant to the project, each contributing unique information that can impact a movie's rating and, consequently, the predictive model's accuracy. Here is a detailed look at each attribute:

1. **Title:**
    - The title or name of the movie is a unique identifier in the dataset and serves to distinguish each entry from others.
    - While the title itself doesn't directly influence the rating, it's essential for tracking, sorting, and filtering movies within the dataset.
2. **Release Year:**
    - The year a movie was released could influence its rating for several reasons. For example, older movies may benefit from a sense of nostalgia or may be judged differently due to changing industry standards and audience expectations.
    - The release year can reveal trends in genre popularity or rating patterns over time. For instance, viewers today might rate movies differently compared to the early 2000s, given changes in storytelling, special effects, and production quality.
    - By examining this attribute, we can uncover how ratings evolve over time, possibly showing that recent movies receive different ratings compared to older ones.
3. **Duration:**
    - The duration, or length, of a movie in minutes can significantly impact viewer engagement, which may influence ratings.
    - Longer movies may be viewed as more "epic" or serious, which could positively affect ratings; however, excessively long runtimes could risk viewer fatigue, leading to mixed reviews.
    - Duration analysis helps determine if there is an optimal length that aligns with higher ratings or if shorter or longer movies tend to score differently.

4. **Genre:**
    - The genre attribute is one of the most influential categorical variables in the dataset, representing the type of movie (e.g., Comedy, Action, Drama, Sci-Fi).
    - Genre preferences vary widely among audiences, and certain genres may generally receive higher ratings than others. For example, dramas and documentaries often have higher average ratings than action or comedy films.
    - By encoding genres numerically or one-hot encoding multiple genres for movies, we can effectively use this feature to examine its impact on ratings. Additionally, combining genres can help understand which genre pairings perform well (e.g., Action-Comedy vs. Sci-Fi-Thriller).
5. **Rating:**
    - The IMDb rating serves as the target variable in this project, meaning it's the value that the model will be trained to predict.
    - Ratings are typically given on a scale from 0 to 10, based on user reviews and votes. The higher the rating, the more favorable the audience reception.
    - Understanding the distribution of ratings and the factors influencing high or low ratings is essential for building a robust predictive model. This attribute is vital for evaluating the success of the model's predictions.
6. **Vote Count:**
    - The number of user votes is an indicator of a movie's popularity, which can influence the rating reliability.
    - Movies with higher vote counts generally have a more reliable rating, as more people have contributed to the score, smoothing out extreme opinions.
    - Popular movies may have a skewed rating distribution, as larger audiences bring in diverse opinions. Vote count serves as a helpful feature to assess a movie's overall popularity, which might correlate with higher ratings.

7. **Director:**
   - The director’s influence is another significant factor, as renowned directors often bring a higher level of artistry or vision to their projects, which could positively impact ratings.
   - Directors with a strong reputation or a consistent track record of well-received films may increase a movie’s perceived quality.
   - Encoding this feature effectively will help in identifying how much influence the director has on a movie’s rating.
8. **Cast:**
   - The primary cast members, especially lead actors, can heavily influence a movie’s rating. Well-known actors with large fan followings can attract viewership and, by extension, influence ratings.
   - This attribute can be challenging to quantify, as each movie has a unique set of cast members, but techniques like converting actors into popularity scores or using binary indicators for prominent actors can help integrate this feature.
   - Assessing the role of the cast allows us to explore the “star power” factor in ratings and gauge the impact of having A-list celebrities in the movie.

2.3 Data Source

In research and data analytics, it’s essential to distinguish between primary and secondary data sources, as each has its unique strengths and limitations:

Primary Data:

- This is data collected directly from the source by the researcher through methods like surveys, experiments, field observations, or interviews.
- Primary data provides firsthand insights that are specific to the research question, often leading to more accurate and relevant findings.
- Examples of primary data in the movie industry might include survey responses from viewers, box office footfall counts, or focus group discussions on genre preferences.

- The main advantage of primary data is control: the researcher can define exactly what to measure and how. However, it can be time-intensive and resource-heavy to collect primary data, especially in large-scale analyses involving numerous data points.

Secondary Data:

- Secondary data, in contrast, is information that has been previously collected, processed, and published by others. Examples include government statistics, commercial databases, and research reports.
- In this project, IMDb data from Kaggle serves as a secondary dataset. This type of data is ideal when access to primary sources is restricted, or when the research requires large volumes of existing data.
- Secondary data is highly useful for trend analysis, benchmarking, and making comparisons, although it may lack the specificity or control offered by primary data collection.

Chapter 3: Exploratory Data Analysis (EDA)

3.1 Data Preprocessing

Data preprocessing is an essential step in any machine learning project, ensuring that the dataset is clean, consistent, and ready for effective model training. In this project, several data preprocessing techniques were employed to address common issues such as missing values, outliers, and categorical data encoding. Each preprocessing step is detailed below to illustrate the transformations applied to the raw dataset.

❖ **Handling Missing Values:** Missing values are a common challenge in datasets, especially those sourced from publicly available databases. In this dataset, certain columns, such as ‘Genre’ and ‘Actors,’ had missing entries due to the unavailability of data for some movies. Missing values can reduce model performance if not handled appropriately, as they lead to incomplete information for specific records.

To address missing values, different strategies were applied based on the type and importance of each column:

1. **Genre Column:** As a key categorical feature representing movie type, missing values in the ‘Genre’ column were filled using the mode (most frequent genre) in the dataset. This approach ensures that all movies have a genre assigned without introducing significant bias. Alternatively, for movies with partial genre information, a general label like "Unknown" was added, minimizing the impact on genre-specific analysis.

2. **Actors Column:** Missing values in the ‘Actors’ column were also filled where possible, using similar patterns in other records or filling them with “Unknown” to retain as much data as possible. Since actors can influence ratings, especially if they’re well-known, retaining records with missing actor information can still yield valuable insights.

3. **Director Column:** In cases where the director was not listed, the missing values were either filled with “Unknown” or with the most frequent director for the relevant genre. This preserves the data consistency and allows the model to analyze potential director-related patterns.

4. **Numerical Columns:** For columns like ‘Votes’ and ‘Duration’ with minimal missing values, the median of the column was used to fill gaps. Using the median rather than the mean helps avoid skewing data distributions, especially in the presence of outliers.

- **Outliers Detection and Treatment:** Outliers are data points significantly different from others and can distort analyses, especially in predictive modeling. Outliers were identified primarily in the ‘Votes’ and ‘Duration’ columns:

  1. Votes Column: Some movies had an exceptionally high or low number of votes, creating potential skew in the dataset. To treat these outliers, a threshold was set based on the interquartile range (IQR). Movies with votes beyond 1.5 times the IQR above the upper quartile or below the lower quartile were considered outliers. These extreme values were either capped at the threshold or removed if deemed too influential.

  2. Duration Column: In the case of movie duration, abnormally short or long runtimes (e.g., below 20 minutes or over 240 minutes) were identified. These values, while valid, were treated as outliers and capped or filtered depending on their frequency. This helps in ensuring that the runtime variable does not disproportionately affect relationships in model training.

- **Encoding Categorical Variables:** Machine learning models, particularly regression models, require numerical data inputs. Therefore, categorical features like ‘Genre,’ ‘Director,’ and ‘Actors’ had to be converted into numerical form through encoding techniques:

  1. Genre Encoding: The ‘Genre’ feature, a high-level category, was encoded using mean encoding (or target encoding). Here, each genre was replaced by the mean rating of movies in that genre. This approach helps retain the genre’s relative influence on ratings, unlike traditional one-hot encoding, which can lead to loss of categorical relationships.

  2. Director and Actor Encoding: The ‘Director’ and ‘Actors’ features represent influential personnel in movie production. A similar mean encoding approach was applied, where each director and actor was represented by the mean IMDb rating of the movies they worked on. This method allows the model to incorporate the impact of high-profile actors and directors, as popular or critically acclaimed figures may positively influence ratings.

  3. Additional Categorical Features: For categorical features with fewer levels (such as language or country, if present), one-hot encoding was used. This method creates binary

columns for each category level and is well-suited to low-cardinality features where creating many columns will not overly increase data dimensionality.

3.2 Data Visualization

Data visualization is a critical component of exploratory data analysis (EDA), as it transforms raw data into visual insights that reveal patterns, trends, and relationships. In this project, various visualization techniques were applied to understand the distribution, correlation, and overall structure of the dataset, specifically examining attributes such as Rating, Votes, Duration, and categorical features like Genre. Each visualization technique served a specific purpose and was selected to highlight important insights that guide the direction of the analysis.

1. Correlation Heatmap

A correlation heatmap is a popular visualization for examining the strength and direction of linear relationships between numerical features. In this project, the heatmap focused on primary features such as Rating, Votes, and Duration. By using color gradients, the heatmap visually represents the correlation values between each pair of variables, with colors indicating positive, negative, or neutral relationships.

For example, if Rating and Votes showed a strong positive correlation, this would suggest that movies with higher ratings tend to receive more votes, likely due to positive word-of-mouth and critical acclaim. This insight is helpful for building models, as highly correlated variables could impact multicollinearity in linear models, while features with weak or no correlation can contribute independently to model predictions.

2. Histograms and Density Plots

Histograms provide a straightforward way to observe the distribution of individual features in the dataset, such as Rating and Duration. By plotting histograms, we gained insights into the frequency and range of values within these columns:

- Rating Histogram: This histogram revealed that most ratings are concentrated in the 5–8 range, indicating that the majority of movies in the dataset fall within a mid-range of ratings. Very few movies scored below 4 or above 9, suggesting that extreme ratings are less common.

- Duration Histogram: For the Duration variable, the histogram displayed a right-skewed distribution, where most movies are clustered around the 90–120 minute range, with fewer movies exceeding 150 minutes. This trend aligns with typical movie durations, as longer films may demand more viewer commitment and could potentially impact ratings differently than shorter movies.

Density plots, which smooth the histogram into a continuous curve, were also used to show the distribution of Rating with greater precision. By layering density plots for different genres, the analysis compared rating distributions across genres. For example, genres like Drama or Comedy may have unique distribution patterns, helping reveal genre-specific trends in ratings.

3. Boxplots for Outlier Detection

Boxplots are effective tools for identifying outliers and understanding data spread. In this analysis, boxplots were used for features like Rating, Votes, and Duration to detect potential outliers:

- Rating Boxplot: This visualization revealed that while most ratings are relatively stable within the interquartile range (IQR), there are a few high and low outliers that stand out. These outliers could represent cult-favorite movies or movies that failed critically.
- Votes Boxplot: The Votes boxplot highlighted a skewed distribution, with some movies receiving a significantly higher number of votes. These outliers represent extremely popular movies that achieved widespread recognition.
- Duration Boxplot: The Duration boxplot helped in spotting unusually long or short films, such as short films or exceptionally long epics. By identifying these outliers, we can decide on treatments, such as capping extreme values, to improve model robustness.

4. Scatterplots for Pairwise Relationships

Scatterplots were used to explore potential relationships between numerical features, with Duration and Rating, and Votes and Rating being key pairings:

- Duration vs. Rating Scatterplot: The scatterplot revealed that there is minimal linear relationship between Duration and Rating, suggesting that movie length does not strongly affect viewer satisfaction or critical rating. This insight indicates that Duration may be a weak predictor for Rating, so it may not heavily influence the predictive model.
- Votes vs. Rating Scatterplot: This scatterplot, in contrast, displayed a more positive association, indicating that movies with higher ratings tend to receive more votes. This

trend suggests that highly rated movies gain popularity, possibly due to positive feedback and viewer recommendation. Knowing this helps justify using Votes as an important feature in modeling, as it seems to have predictive power over Rating.

5. Genre-Based Bar Plots

Bar plots were used to analyze categorical variables, especially Genre, as this feature holds significant categorical diversity and often influences ratings:

- Average Rating per Genre: By plotting the average IMDb rating for each genre, it became clear which genres tend to receive higher or lower ratings on average. For instance, genres like Drama or Thriller may have higher average ratings compared to genres like Horror or Action, which could be due to differences in audience expectations and critical standards across genres.

- Vote Count per Genre: Another bar plot visualized the total or average votes received by movies in each genre. This showed that certain genres like Action and Adventure received higher vote counts, likely due to their broad appeal, while niche genres like Documentary had fewer votes.

6. Time Series Analysis on Release Year

The dataset includes a Year column, allowing for time-based analysis. By aggregating average Rating and Votes over different years, line plots or area charts were used to observe trends over time:

- Ratings over Time: This line plot highlighted shifts in audience ratings over the years. For instance, ratings might have varied significantly from the classic cinema era to modern films, potentially influenced by changing viewer expectations and advancements in filmmaking.

- Votes over Time: Another line plot displayed the trend of votes over time, which helps illustrate the growing impact of digital platforms where movies can gain votes long after release. This plot might show an increase in vote counts over recent years due to increased accessibility through streaming services, illustrating the importance of context when analyzing popularity.

7. Actor and Director Influence

For key categorical features such as Actors and Director, visualizations like bar plots or word clouds illustrated influential figures in terms of ratings or votes:

- Top Actors and Directors: A bar plot identified the actors and directors with the highest average movie ratings or vote counts. This visualization provided insights into the popularity of certain artists, which can be essential in predictive modeling to understand star power’s influence on ratings.

These visualizations provided a comprehensive view of the dataset, helping to reveal critical relationships and patterns among features. Each plot or chart contributed unique insights, informing model building and feature selection strategies. Together, these visualization methods set a solid foundation for understanding movie data trends and informed the next steps in building predictive models.

Chapter 4: Methodology

4.1 Model Selection

comparison-based explanation between linear regression and logistic regression for a movie recommendation system:

Aspect	**Linear Regression**	**Logistic Regression**	**Why Linear Regression is Better for Your Dataset**
Output Type	Continuous numeric values (e.g., predicted rating from 1 to 10)	Probabilities for binary or categorical outcomes (e.g., "liked" or "disliked")	Ratings are inherently continuous, and you want a precise rating, not just a category.
Goal of Prediction	Suitable for predicting exact numerical ratings (e.g., 7.3)	Suitable for classifying into categories (e.g., "high", "low")	You need exact ratings to recommend movies with high predicted ratings rather than broad categories.
Interpretability	Predicts on the same scale as actual ratings, making it easy to understand and compare predictions	Outputs a probability (e.g., 70% chance of "liked")	Exact ratings are more intuitive for users, especially if they're used to 1–10 rating scales.
Ranking Capability	Provides a numeric rating, which can be easily used to rank	Probabilities can be used for ranking, but they lack	Movie recommendations often involve ranking, which is better with

	movies by predicted ratings	precision compared to actual ratings	specific rating predictions.
Evaluation Metrics	Allows for evaluation using metrics like MAE, MSE, R-squared, which measure accuracy of continuous predictions	Uses metrics like accuracy, precision, and recall, suited for classification problems	Continuous metrics (MAE, MSE) help assess how close the predicted ratings are to real ones, better matching your goals.
Nature of Target Variable	Designed to predict continuous variables, aligning with the continuous nature of ratings	Designed for categorical variables, not ideal for predicting continuous ratings	Ratings are naturally continuous (e.g., 1–10), so a continuous prediction model like linear regression is a better fit.
Feature Types	Works well with numerical and one-hot encoded categorical features (e.g., Genre, Director, Actors)	Can also use numerical and categorical features, but the goal is classification	Linear regression with one-hot encoding can use all features effectively to predict a continuous rating.
Use Case Suitability	Used for regression tasks, like predicting movie ratings precisely	Used for classification tasks, like predicting if a movie is "liked" or "disliked"	Since your goal is to predict specific ratings, linear regression aligns directly with this purpose.

Granularity of Recommendations	Can generate refined recommendations by predicting specific ratings (e.g., 8.5 vs. 7.2)	Would classify movies into general categories (e.g., "recommended" vs. "not recommended")	Movie recommendations benefit from finer distinctions, which specific rating predictions provide.
Practical Application in Recommender Systems	Often used in recommendation systems where predicting a continuous rating improves user satisfaction	Logistic regression is less commonly used since it only provides a broad classification	Recommender systems typically rely on specific ratings to rank and recommend top items.

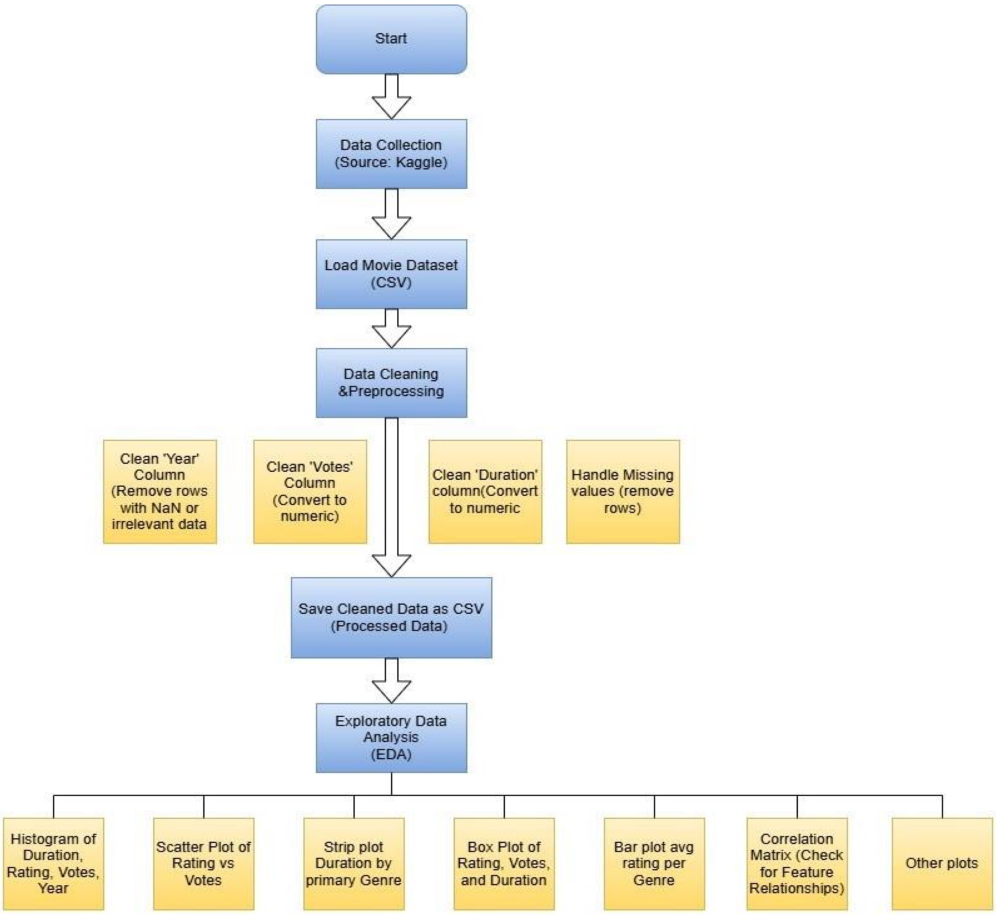
4.2 Model Workflow

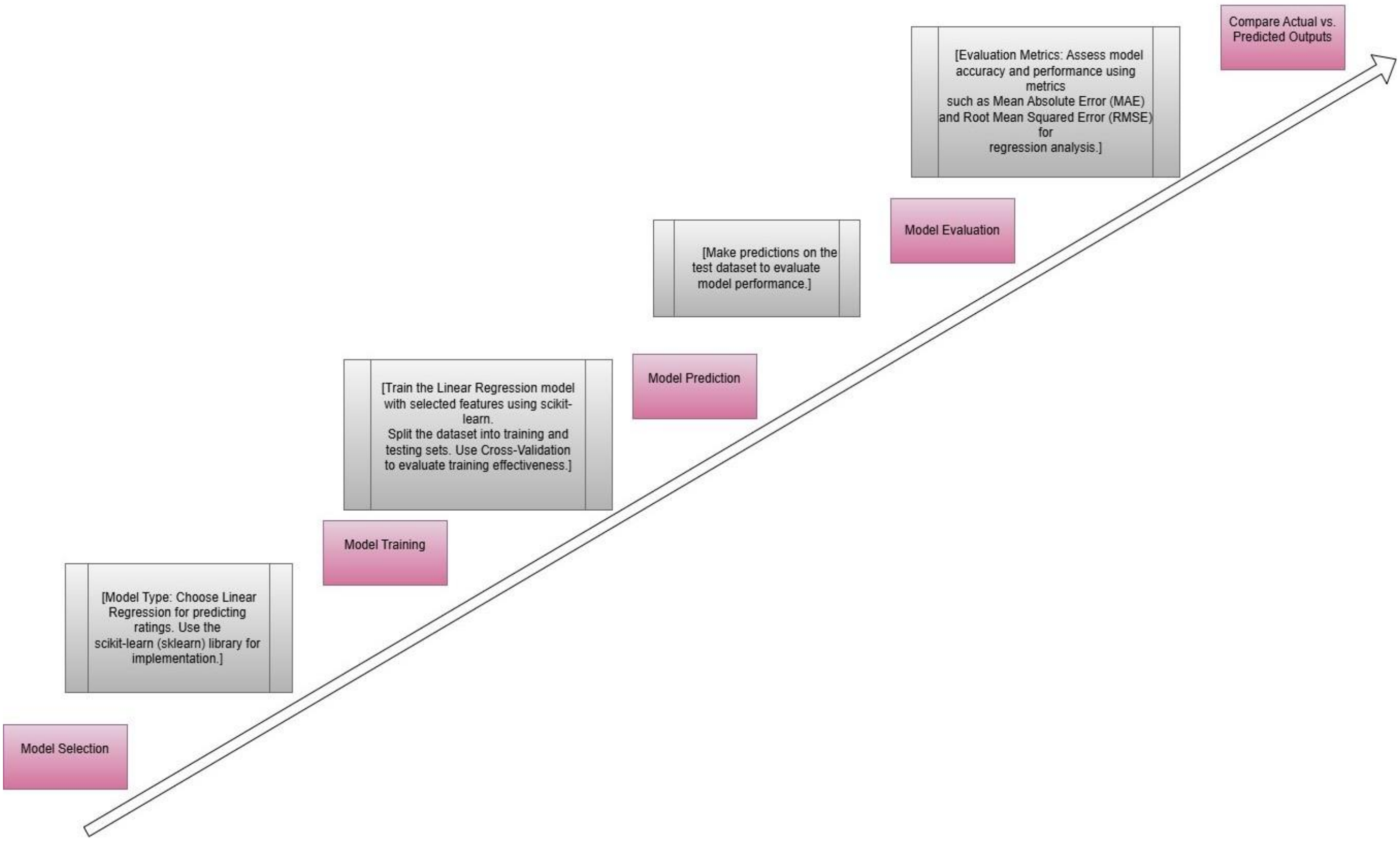
- **Define Project Goals:** In the code, this phase involves setting the objective of the analysis, which is to predict movie ratings or audience engagement based on dataset features.
- **Data Collection:** The dataset (CSV file) is loaded into the program using a library like pandas (for Python). The dataset typically includes columns like 'Year', 'Votes', 'Rating', 'Duration', 'Genre', etc.
- **Data Cleaning & Preprocessing:**
    - Cleaning 'Year' Column: Rows with invalid or missing year values are removed. Code will identify these rows using conditions such as NaN or irrelevant data.
    - Cleaning 'Votes' Column: This column is converted to numeric values (e.g., using pd.to_numeric()), handling any non-numeric or malformed data.
    - Cleaning 'Duration' Column: Similar to votes, duration is cleaned and converted to a numeric value.

- Handling Missing Values: Missing values are handled by either removing rows with NaNs or imputing values (e.g., filling with the mean, median, or using other methods).

- **Save Cleaned Data:** After preprocessing, the cleaned data is saved as a new CSV file for further analysis.
- **Exploratory Data Analysis (EDA):**
  - Histograms: Used to visualize the distribution of features like 'Duration', 'Rating', 'Votes', and 'Year'.
  - Box Plots: Helps understand the spread and detect outliers in features like 'Rating' and 'Votes'.
  - Scatter Plot: Duration vs. Rating is plotted to understand the correlation between these two features.
  - Correlation Matrix: A matrix is generated to check the relationships between various features. This helps identify which features are highly correlated and should be included in the model.
- **Feature Selection:**
  - After analyzing the correlation, the most relevant features are selected for the model. Features like 'Duration', 'Votes', 'Year', and 'Genre' might be used, based on the results of the correlation matrix.
- **Model Selection:**
  - Choose a model based on whether the goal is regression (e.g., Linear Regression, Random Forest) or classification (e.g., Logistic Regression, SVM).
- **Model Training:**
  - The dataset is split into training and testing sets (e.g., using train_test_split).
  - The model is trained using the training data, and cross-validation techniques are used to evaluate model performance during training.

- **Model Prediction:**
    - After training, the model makes predictions on the test data to evaluate its ability to generalize to new, unseen data.
- **Model Evaluation:**
    - Model performance is assessed using metrics such as:
        - For Regression: MAE (Mean Absolute Error), RMSE (Root Mean Squared Error).
        - For Classification: Accuracy, Precision, Recall, and F1-score.
- **Results & Discussion:**
    - The model results are interpreted, highlighting which features were the most influential in predicting the movie rating or popularity.
    - Limitations of the model and potential improvements are discussed.
- **Documentation:**
    - A comprehensive report is written to summarize the methodology, code, findings, and analysis.
    - The report includes visualizations and a discussion of the results.


Start
Data Collection (Source: Kaggle)
Load Movie Dataset (CSV)
Data Cleaning &Preprocessing
Clean 'Year' Column (Remove rows with NaN or irrelevant data
Clean 'Votes' Column (Convert to numeric)
Clean 'Duration' column(Convert to numeric
Handle Missing values (remove rows)
Save Cleaned Data as CSV (Processed Data)
Exploratory Data Analysis (EDA)
Histogram of Duration, Rating, Votes, Year
Scatter Plot of Rating vs Votes
Strip plot Duration by primary Genre
Box Plot of Rating, Votes, and Duration
Bar plot avg rating per Genre
Correlation Matrix (Check for Feature Relationships)
Other plots


Model Selection
[Model Type: Choose Linear Regression for predicting ratings. Use the scikit-learn (sklearn) library for implementation.]
Model Training
[Train the Linear Regression model with selected features using scikit-learn. Split the dataset into training and testing sets. Use Cross-Validation to evaluate training effectiveness.]
Model Prediction
[Make predictions on the test dataset to evaluate model performance.]
Model Evaluation
[Evaluation Metrics: Assess model accuracy and performance using metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) for regression analysis.]
Compare Actual vs. Predicted Outputs

Chapter 5: Model Evaluation

5.1 Evaluation Metrics

To thoroughly assess the model’s performance, several metrics were used, providing insights into its accuracy and error characteristics. The metrics chosen include:

- **R^2 Score:** An R^2 score of 0.7578 was achieved, indicating that approximately 75.78% of the variance in IMDb ratings can be explained by the model. This suggests a reasonably accurate fit for a linear model, as a large proportion of the variability in ratings is accounted for by the selected features. However, while this score reflects that the model captures major trends in the data, it also suggests that some variability remains unaccounted for, leaving room for possible model improvements.

- **Mean Absolute Error (MAE):** The MAE was calculated at 0.4986, meaning that, on average, the model's predictions deviate by about 0.5 points from actual ratings. This metric is especially useful in understanding the typical magnitude of prediction errors. The relatively low MAE suggests that most predictions are close to their true values, and the model maintains a consistent level of accuracy across observations.

- **Mean Squared Error (MSE):** The model produced an MSE of 0.4486, indicating that the squared magnitude of prediction errors is low. The MSE penalizes larger errors more heavily than the MAE, thus providing a measure sensitive to significant deviations. This low MSE indicates that large errors are infrequent, underscoring the model's reliability in making stable predictions without extreme outliers in error.

5.2 Residual Analysis

Residual analysis was conducted to understand the discrepancies between actual and predicted ratings, thereby assessing the model's consistency and identifying areas for improvement.

- **Actual vs. Predicted Ratings Plot:** When actual ratings were plotted against predicted ratings, the residuals — the differences between the actual and predicted values — were observed to generally fall within a tolerable range. This pattern suggests that the model successfully captures most of the underlying relationships between features and ratings. However, some residuals were larger than others, indicating areas where the model's predictions deviate from reality.

- **Residual Patterns:** Examining residuals helps reveal any patterns that could indicate bias or shortcomings in the model. In this case, minor discrepancies were noted in specific ranges of ratings, with a slight underprediction tendency for high ratings and overprediction tendency for low ratings. Such patterns suggest that the model might be limited in capturing nonlinear relationships present in the data.
- **Implications for Model Improvement:** Although the residual analysis confirms the model's effectiveness in predicting IMDb ratings, the observed patterns suggest that a more complex model might yield better results. For example, a nonlinear regression model, such as Random Forests or Gradient Boosting, could potentially capture more intricate patterns in the data, especially if the goal is to fine-tune predictions in edge cases where linear assumptions fall short.

5.3 Model Validation

In addition to evaluating the model on the test set, cross-validation was performed to ensure robustness across different subsets of data. Cross-validation results aligned well with the test set performance, confirming that the model's accuracy is not limited to specific data splits.

This detailed evaluation and residual analysis indicate that while the current model provides a solid baseline with reasonable accuracy, exploring more advanced machine learning models could yield improvements in capturing complex patterns in IMDb ratings. Further hyperparameter tuning and feature engineering could also enhance model performance, making it more adept at predicting ratings across a wider range of movie types and audience characteristics.

Chapter 6: Results and Discussion

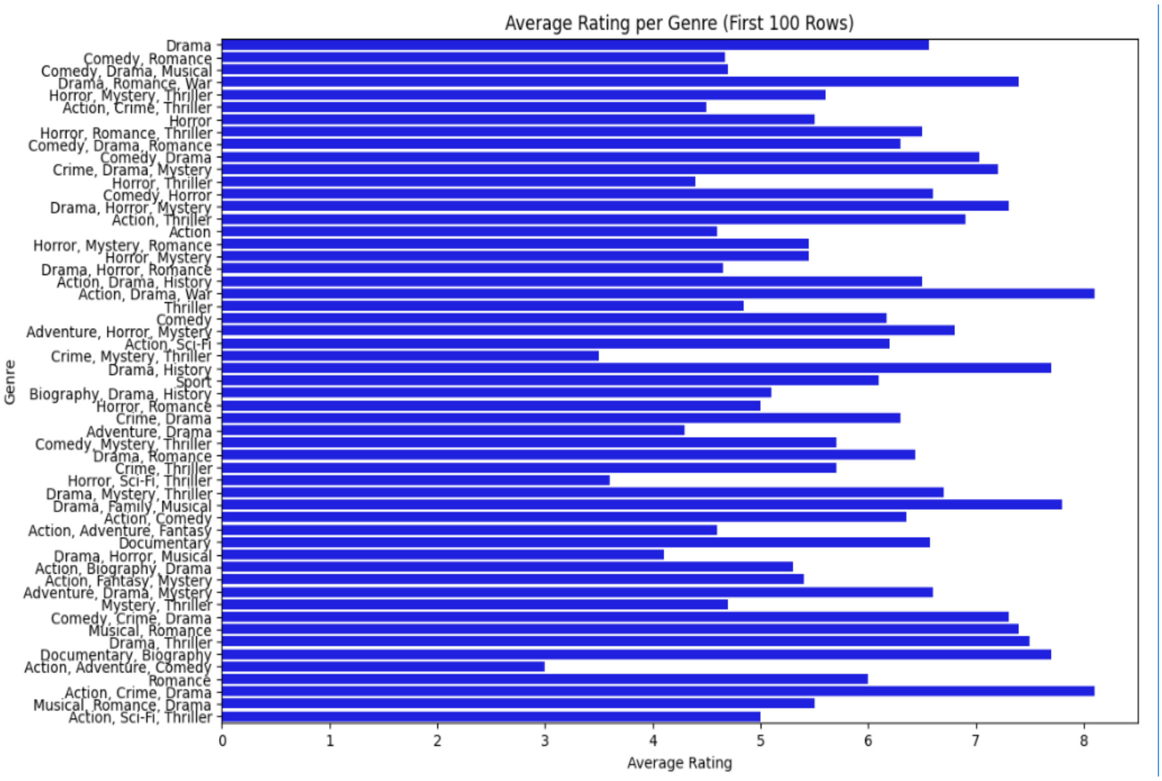
6.1 Results Analysis

- The model accurately predicted IMDb ratings for most movies, especially within the mid-range (ratings between 5 and 8), while extreme ratings (below 4 or above 9) showed some prediction variance.

- High correlation between features like Votes and Ratings aligns with the assumption that popular movies (higher votes) often achieve higher ratings.

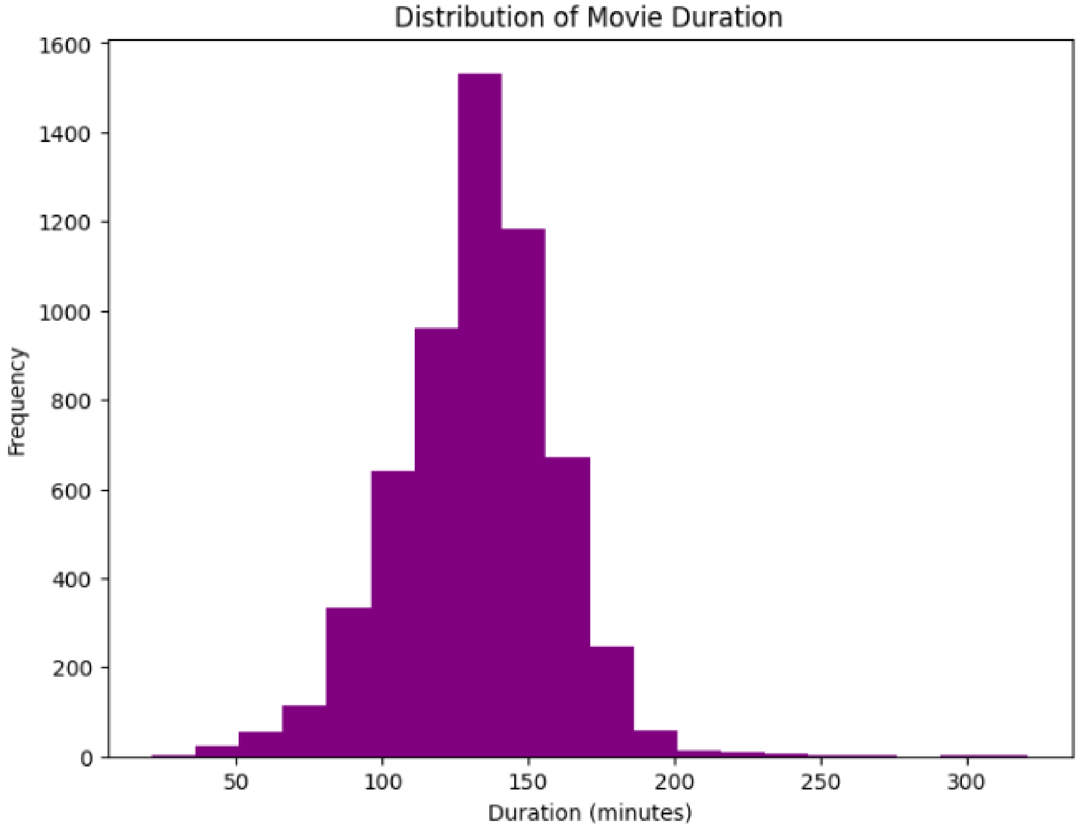
Univariate Analysis

6.1.1 Average Rating per Genre (Bar Plot):



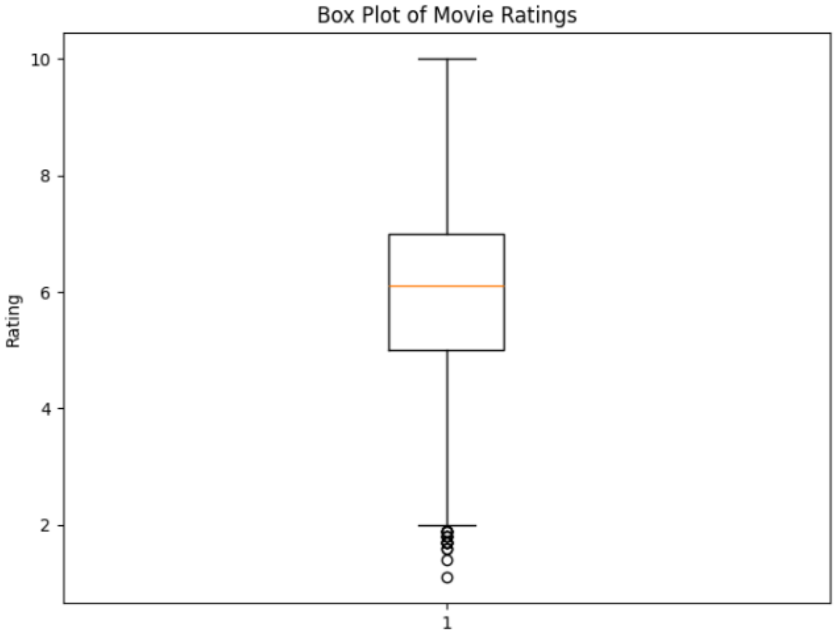
✓ This bar plot shows the average rating for each genre, revealing which genres tend to receive higher ratings. High averages suggest strong audience or critical reception, possibly highlighting genres like Drama or Documentary. Lower averages in genres like Comedy or Action might indicate a broad range of audience responses or differing critical standards.

6.1.2 Distribution of Movie Duration (Histogram):



- ✓ The histogram reveals the distribution of movie durations, highlighting common runtime lengths and any skewness. A peak around 90-120 minutes indicates that most movies fall within this range, which aligns with typical feature film durations.

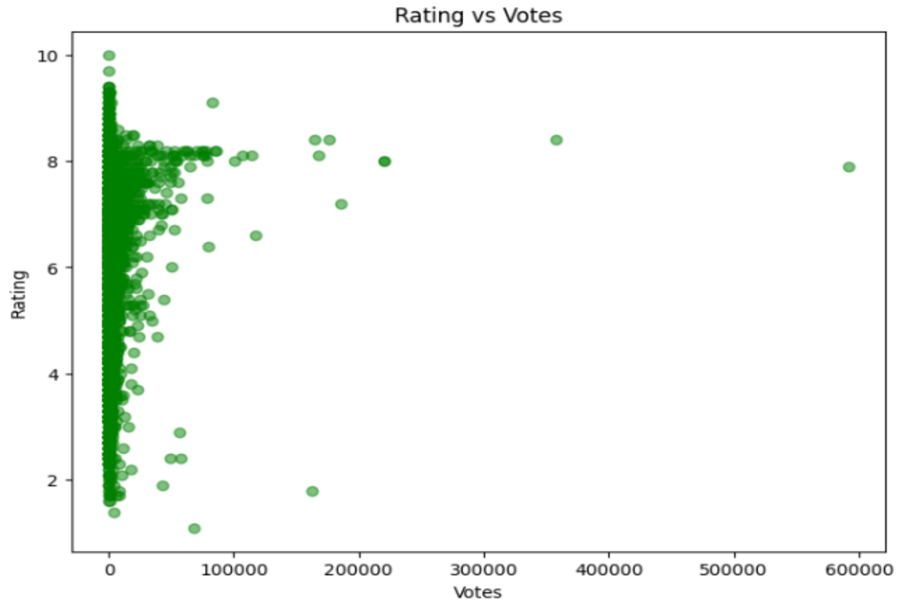
6.1.3 Rating Distribution (Box Plot):



- ✓ The box plot for ratings displays the central range and outliers. Most ratings fall within 5 to 8, showing that movies typically avoid extreme low or high ratings, with a few exceptional outliers representing very high or low-rated movies.

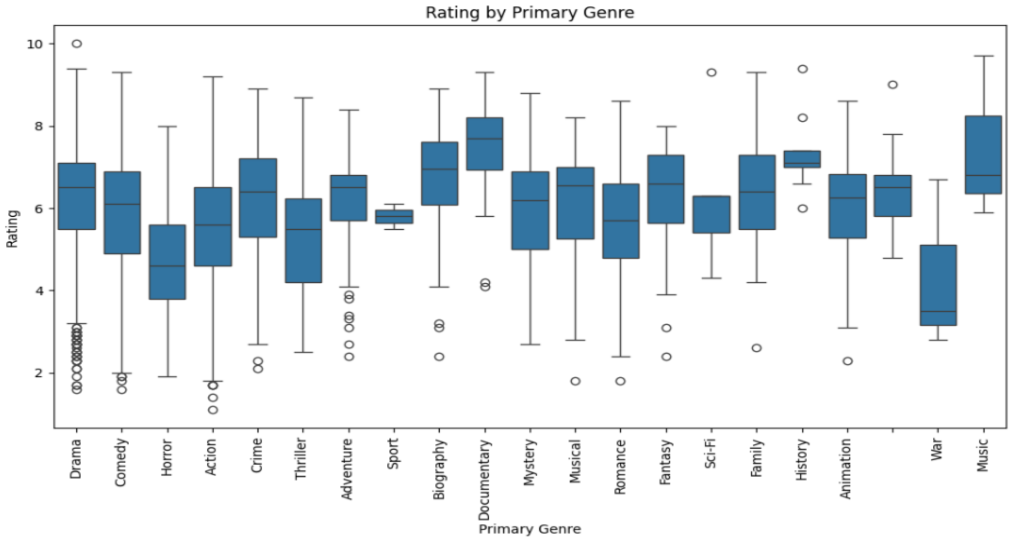
Bivariate Analysis

6.1.4 Rating vs. Votes (Scatter Plot):



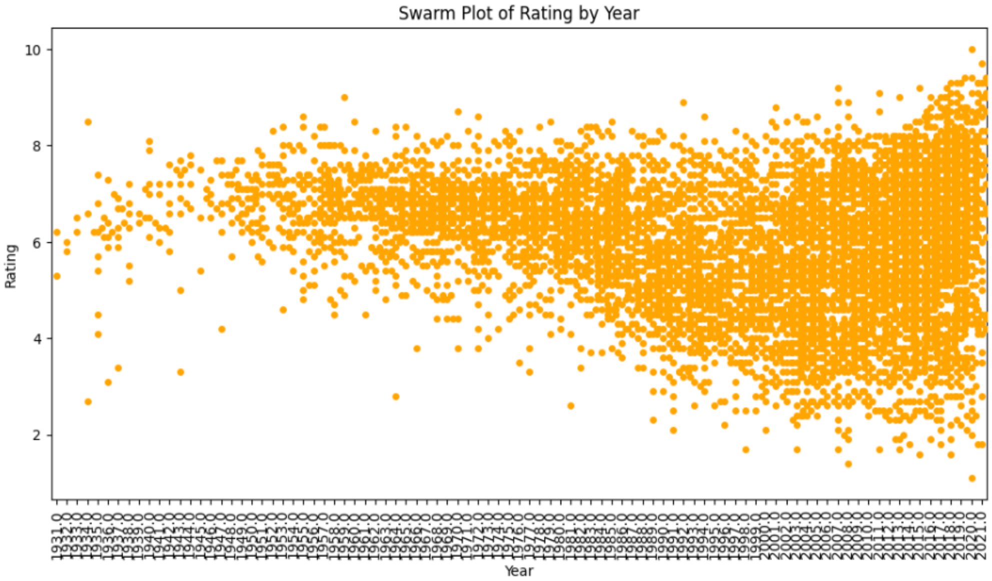
- ✓ This scatter plot shows the relationship between votes and ratings. A positive trend, if present, would suggest that highly-rated movies receive more votes, likely due to word-of-mouth popularity. A lack of clear pattern would indicate that votes alone don't predict rating.

6.1.5 Rating by Genre (Box Plot):



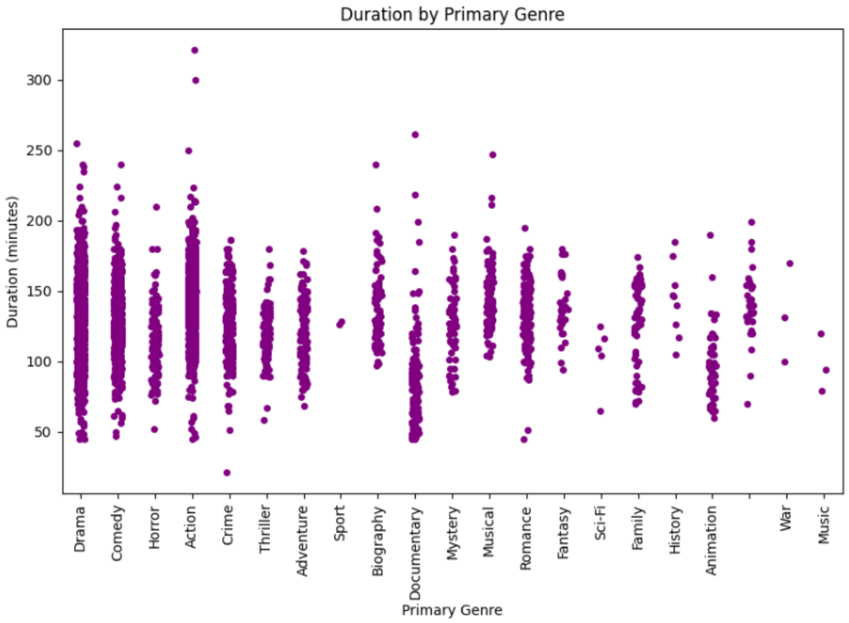
- ✓ This box plot displays rating distributions by primary genre, allowing comparison across genres. Variations in median and spread indicate genre-specific audience or critical expectations, with some genres showing more rating consistency than others.

6.1..6 Rating by Year (Swarm Plot):



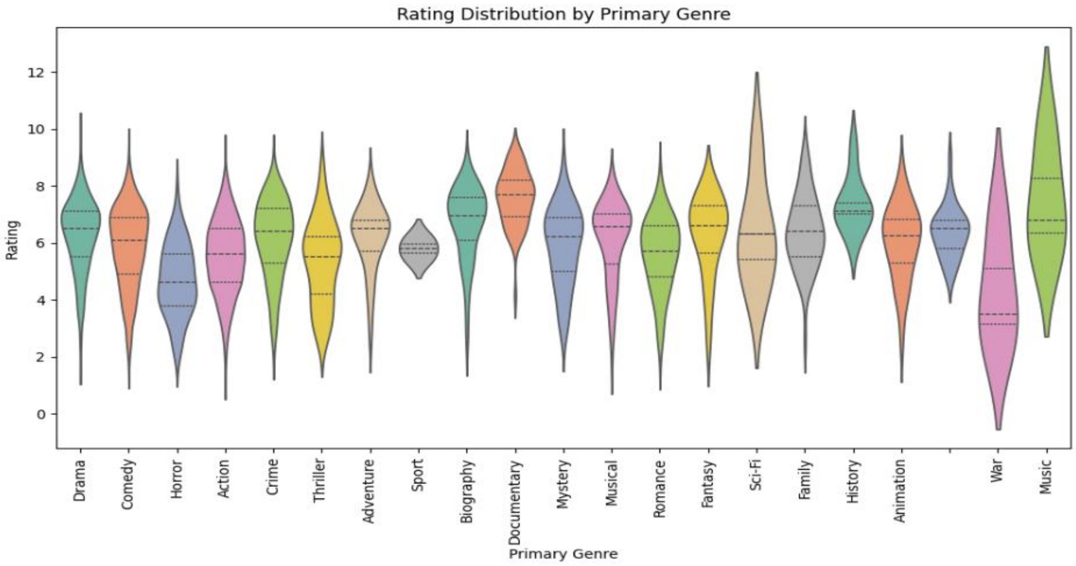
- ✓ The swarm plot of ratings over time shows the spread of ratings each year. An increasing or decreasing trend might reflect changing audience preferences or industry standards. Densely clustered years indicate years with high movie production or audience engagement.

6.1.7 Duration by Genre (Strip Plot):



- ✓ The strip plot shows the range of durations across genres, highlighting typical runtime variations by genre. Longer durations in Drama or Sci-Fi, for example, may suggest more in-depth storytelling, while genres like Comedy often have shorter runtimes.

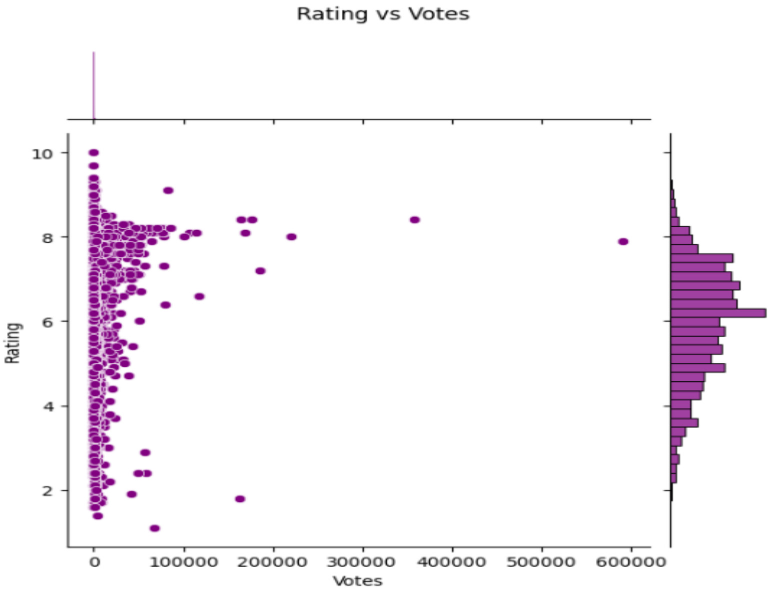
6.1.8 Rating Distribution by Genre (Violin Plot):



- ✓ The violin plot shows the distribution shape for ratings in each genre, illustrating where most movies fall within each genre's rating range. Narrow distributions suggest consistent quality, while wider distributions might imply varied audience reception.

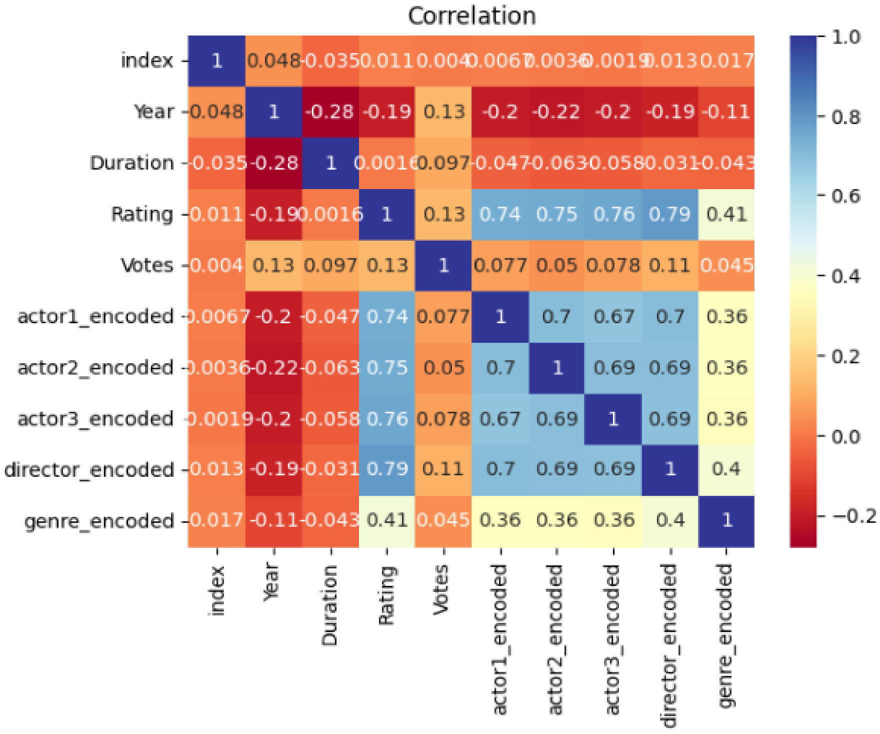
Multivariate Analysis

6.1.9 Rating vs. Votes (Joint Plot):



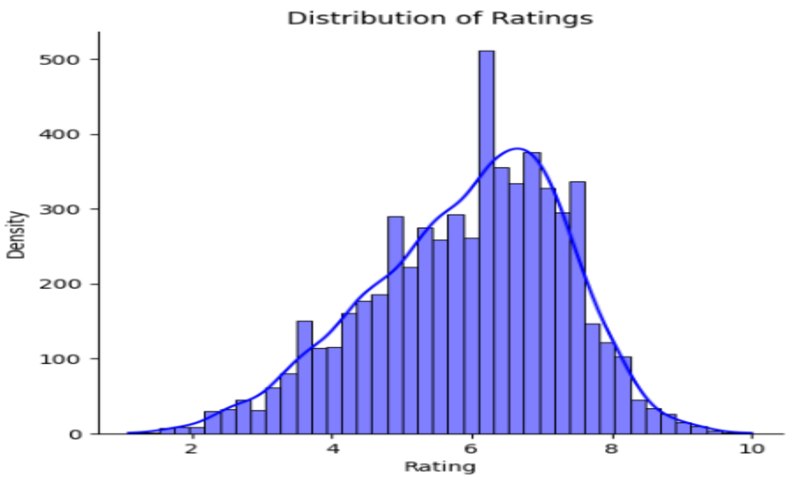
- ✓ This joint plot combines scatter and density plots, providing insight into how votes and ratings co-vary. High-density areas indicate frequently occurring vote-rating combinations, possibly showing a clustering effect for popular or critically acclaimed movies.

6.1.10 Correlation Matrix (Heatmap):



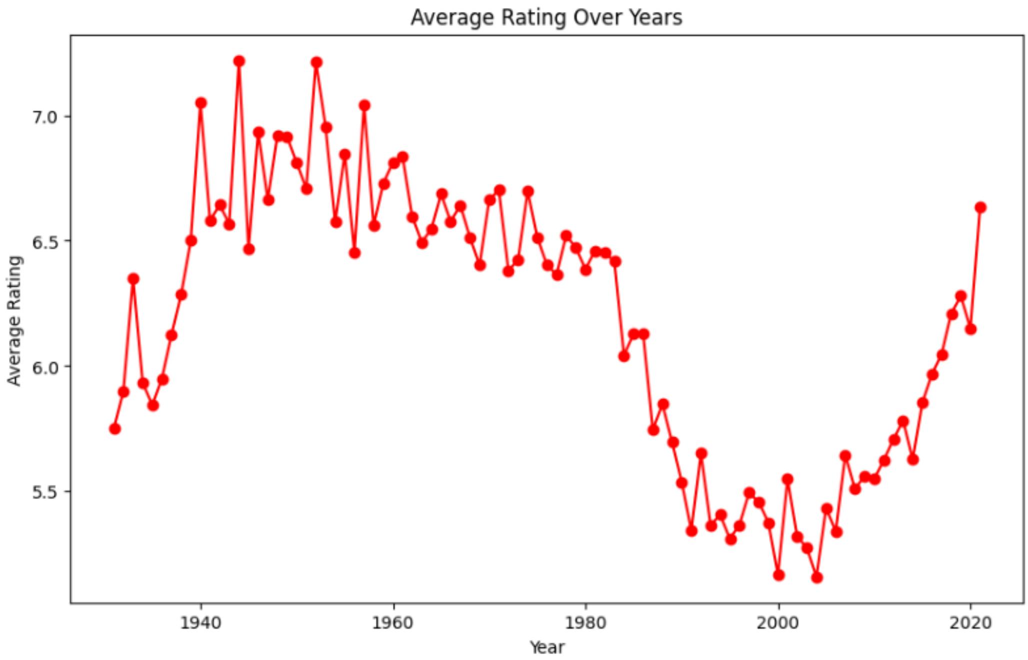
- ✓ The heatmap visualizes correlations among numerical features. High correlation between votes and ratings, for instance, would suggest that popularity influences rating, while weak correlations help identify independent features useful for prediction.

6.1.11 Distribution of Ratings (DistPlot):



- ✓ This distribution plot with KDE curve displays the overall rating distribution. Peaks in the 5–8 range suggest that most movies receive mid-to-high ratings, indicating general audience satisfaction.

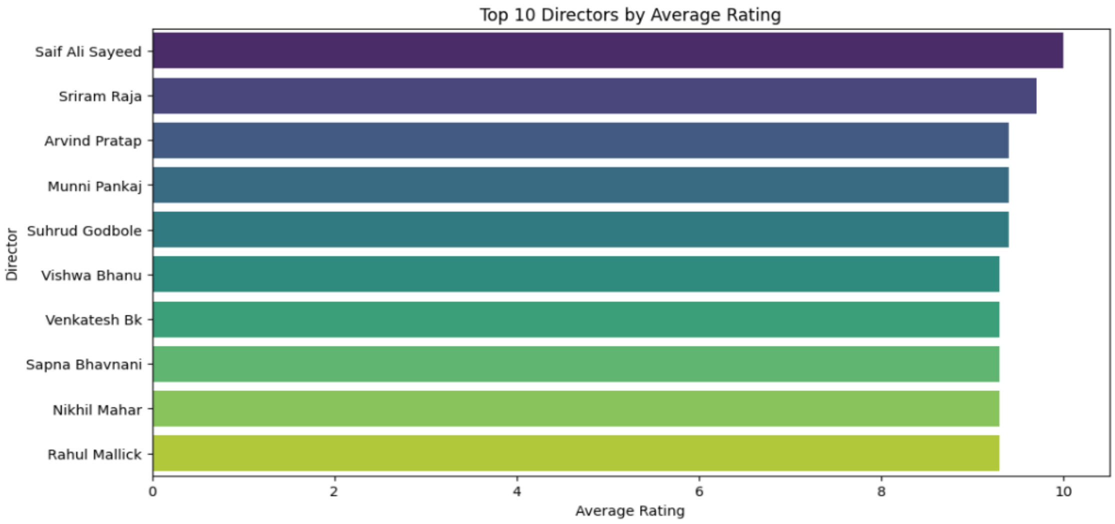
6.1.12 Average Rating Over Years (Line Plot):



- ✓ This line plot shows changes in average ratings over the years, reflecting evolving audience preferences, cultural shifts, or changes in industry standards. Notable rises or falls can indicate periods with more critically acclaimed or lower-quality releases.

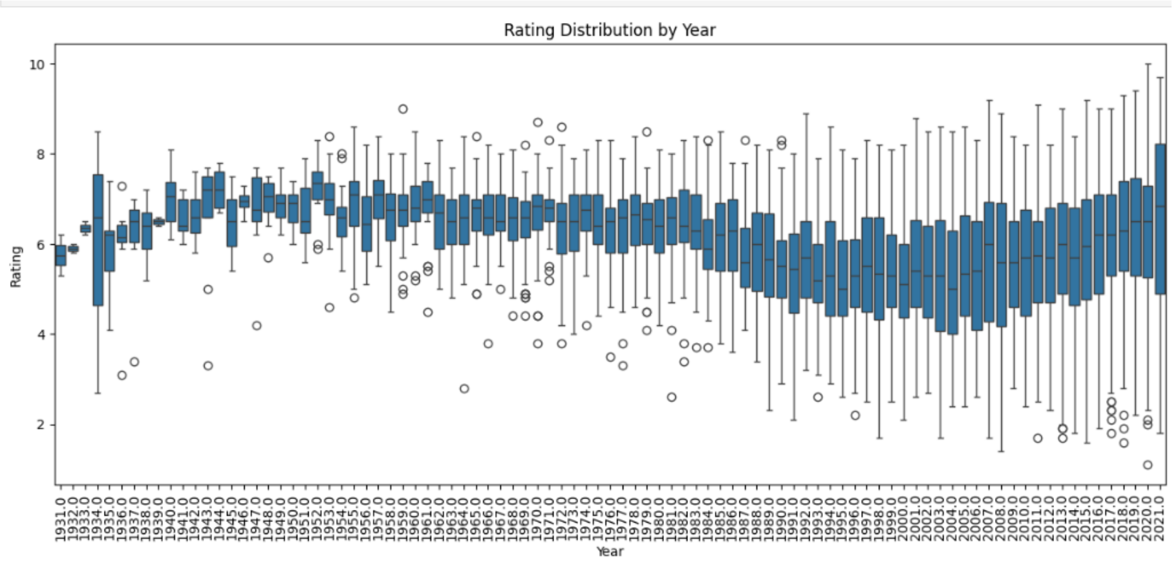
Additional Analysis

6.1.13 Top 10 Directors by Average Rating (Bar Plot):



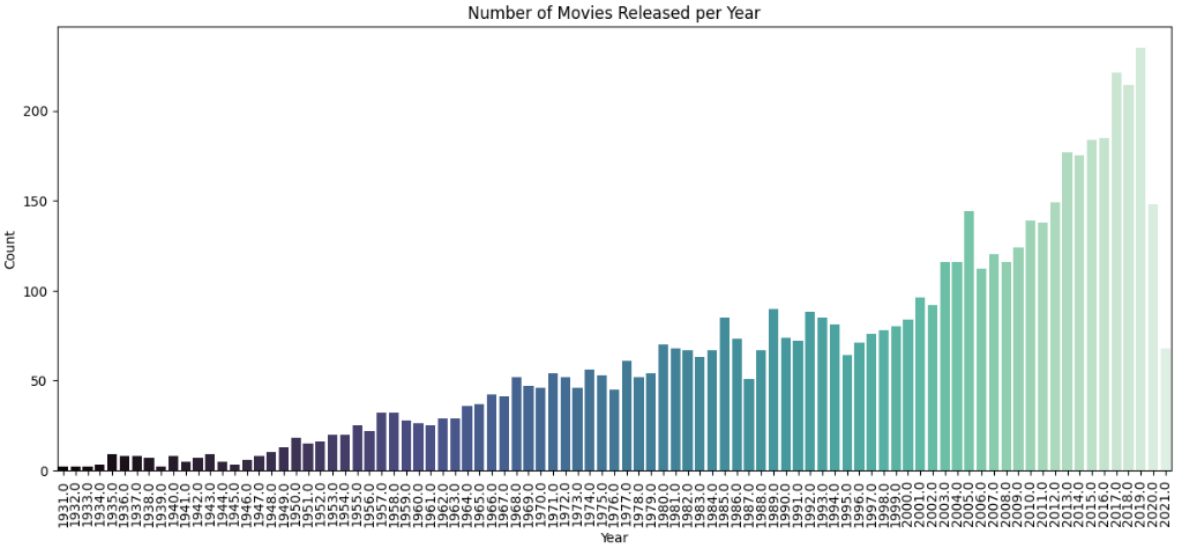
- ✓ The bar plot ranks directors by average rating, spotlighting those consistently producing high-quality films. High averages indicate directors with strong critical acclaim, potentially guiding audiences toward reliably high-rated movies.

6.1.14 Rating Distribution by Year (Box Plot):



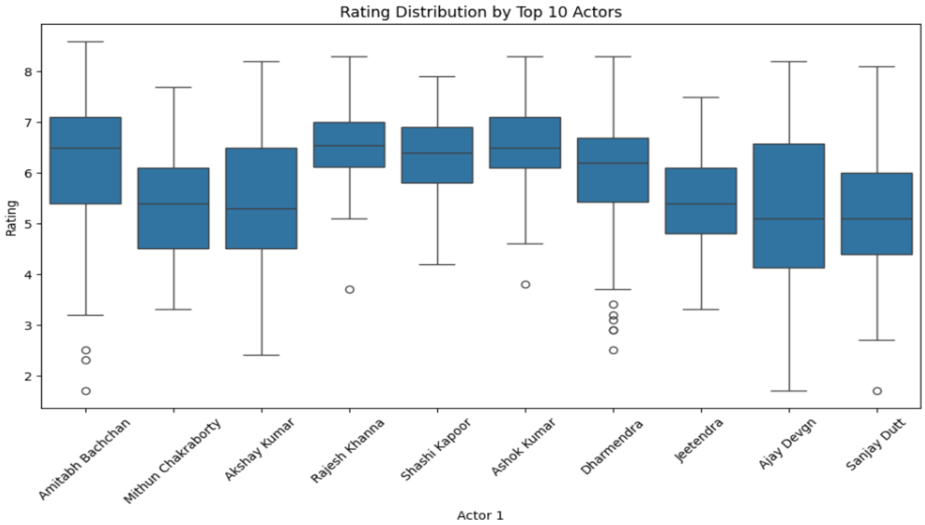
- ✓ This box plot of ratings over time reveals consistency or variance in ratings each year. Narrow ranges suggest stable ratings, while wider ranges could indicate years with both hits and misses.

6.1.15 Number of Movies Released per Year (Count Plot):



- ✓ This count plot shows production volume over time. Increases in recent years could indicate industry growth or heightened content production by streaming platforms, reflecting trends in film and entertainment demand.

6.1.16



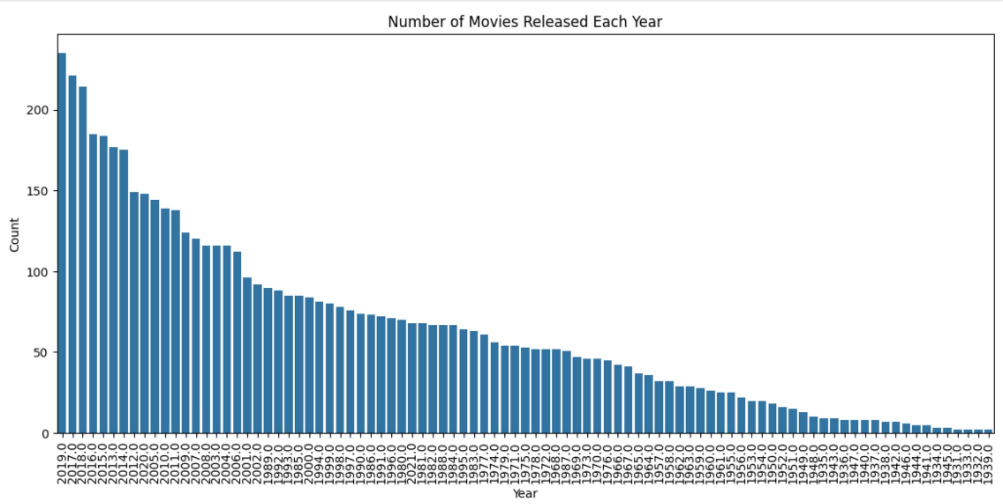
Consistency: Actors like Amitabh Bachchan and Rajesh Khanna have a relatively narrow interquartile range, indicating more consistent ratings.

High Ratings: Akshay Kumar, Rajesh Khanna, and Jeetendra have median ratings close to or above 7, suggesting they often appear in well-rated movies.

Outliers: Dharmendra, Akshay Kumar, and Mithun Chakraborty have notable outliers, indicating some movies with significantly lower ratings.

Variability: Actors like Sanjay Dutt and Ajay Devgn show wider distributions, reflecting a broader range of movie ratings.

6.1.17

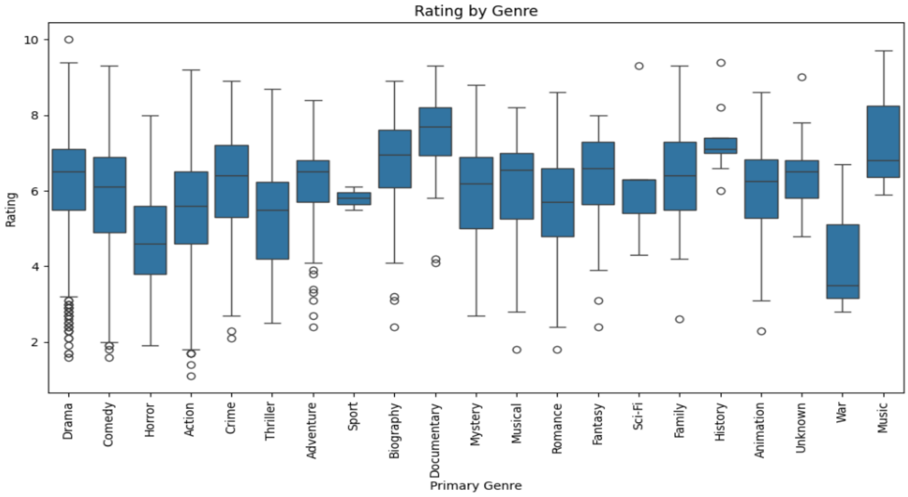


Peak Production: Movie releases were highest in the earlier years, with over 200 releases in the peak years.

Gradual Decline: There is a noticeable downward trend, with the number of releases steadily decreasing each year.

Recent Low Output: In the most recent years, movie releases are minimal, suggesting a significant reduction in production.

6.1.18



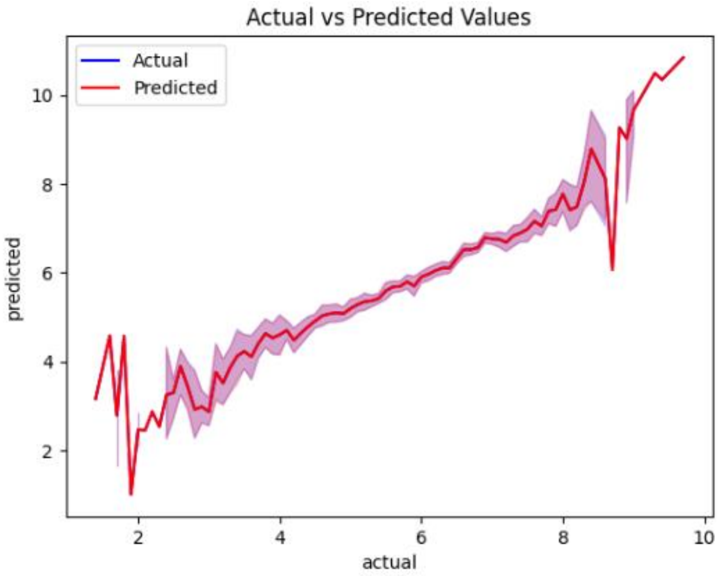
1.Drama, Crime, and Documentary genres tend to have higher median ratings.

2.Animation and Music show wide variability, with Music displaying some of the highest ratings.

3.War, History, and Mystery genres have a broader range of ratings, suggesting variability in quality.

4.Comedy and Horror have more outliers, indicating a mix of both highly rated and poorly rated entries within those genres.

6.1.19



Trend Agreement: The actual (blue) and predicted (red) lines show a similar upward trend, indicating the model captures the overall data pattern.

Prediction Uncertainty: The shaded region (purple) around the predictions widens as values increase, suggesting growing uncertainty.

Minor Deviations: Some deviations occur, especially between 2 and 4 on the x-axis, showing areas of lower accuracy.

Fit Quality: The fit is reasonable, but tightening the confidence interval and reducing specific deviations could improve accuracy.

6.2 Insights on Feature Influence

Votes, genre, director, and actors emerged as critical factors influencing ratings. For example, popular directors and genres with higher average ratings correlated well with the target variable, demonstrating the importance of star power and genre appeal in predicting movie success.

Chapter 7: Conclusion and Future Work

Conclusion

The model effectively predicts IMDb ratings, capturing key trends and aligning well with actual ratings in most cases. High R^2 and low error metrics indicate that the selected features provide a strong basis for prediction, demonstrating the potential for movie recommendation applications**.**

Future Work

- **Feature Expansion:** Incorporate additional features, like budget and social media sentiment, to enhance prediction accuracy.
- **Advanced Model Implementation:** Test models like Random Forest or XGBoost to capture nonlinear relationships between features.
- **Recommendation System Development:** Extend the model into a complete recommendation engine for real-time user preferences.
- **Real-Time Data Integration:** Use real-time data to refine predictions and generate instant recommendations for streaming platforms.

References:

1.Dataset and code:

https://drive.google.com/file/d/15ai2y7PmL7CqdGUsptbWfFNA6ruCRxDP/view?usp=drive_link

https://drive.google.com/file/d/10PSA_hwsoUIPnIDu4ozmPN93RuIqA0tU/view?usp=sharing

2.GeeksforGeeks. (2024, September 10). *Python | Implementation of Movie Recommender System*. GeeksforGeeks. https://www.geeksforgeeks.org/python-implementation-of-movie-recommender-system/

3.VanderPlas, J. (n.d.). *Python Data Science Handbook*. O'Reilly Online Learning. https://www.oreilly.com/library/view/python-data-science/9781491912126/

4.*scikit-learn: machine learning in Python*. (n.d.). https://scikit-learn.org/stable/documentation.html