A PRELIMINARY MINI PROJECT REPORT ON

"IMDb Movies Data Analysis"

SUBMITTED TOWARDS THE PARTIAL FULFILLMENT OF THE REQUIREMENTS OF

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

Academic Year: 2024-25

By:

Bhasker Prasad Sah (123B1B093)

Nitesh Chavan (123B1B102)

Dhiraj Kumar Chaudhary (123B1B109)

Under The Guidance of

Mrs.Chhaya Nayak

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

"IMDb Movies Data Analysis" 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: IMDb Movies Data Analysis

Under the guidance of Mrs.Chhaya Nayak

In the partial fulfillment of the requirements for the S.Y. B. Tech.

(Computer Engineering)

By:

Bhasker Prasad Sah (123B1B093)

Nitesh Chavan (123B1B102)

Dhiraj Kumar Chaudhary (123B1B109)

Mrs.Chhaya Nayak

Project Guide

Chapter 1: Introduction

1.1 Background

A. Growth of the Entertainment Industry

The entertainment industry, especially film, has undergone tremendous growth in recent years due to technological advancements and the expansion of digital platforms. With the rise of streaming services like Netflix, Amazon Prime, and Disney+, audiences now have unparalleled access to vast libraries of media, enabling worldwide content distribution. This shift has blurred geographic boundaries, making global cinema accessible to diverse audiences. The growing accessibility and demand for digital content have fueled the industry's evolution into a significant economic force, influencing sectors like technology and tourism. This rise has created an ideal environment for data-driven approaches to audience engagement and content distribution.

B. Role of Data Science in the Film Industry

Data science is transforming the film industry, providing valuable insights that help studios and streaming platforms make informed decisions. With access to comprehensive datasets like IMDb, which capture metrics ranging from audience ratings to box office earnings, companies can analyze viewing trends, predict successful genres, and tailor marketing strategies. Data-driven approaches have improved content recommendations and resource allocation, as platforms identify patterns in viewer preferences. These insights not only enhance user experience but also ensure that marketing and production align more closely with audience interests, contributing to increased viewer satisfaction.

C. Significance of Movie Ratings

Ratings are critical in determining a film’s commercial success and its cultural impact. High ratings can drive increased box office earnings, attract streaming partnerships, and boost a film’s long-term value. Ratings also influence licensing and distribution decisions, as platforms often prioritize well-rated content to enhance viewer satisfaction. Thus, understanding the factors that contribute to higher ratings, such as genre, cast, or budget, is invaluable for producers and marketers. By analyzing these factors, the project seeks to uncover trends that consistently align with positive viewer feedback, ultimately helping studios create content that resonates with audiences.

D. Recommendation Systems and Their Impact

Modern recommendation systems are essential for helping users discover content suited to their tastes, enhancing engagement and loyalty. Platforms rely on algorithms that analyze user data, including viewing history, ratings, and genre preferences, to provide personalized recommendations. This data-driven personalization not only enriches the user experience but also has financial benefits for platforms, as satisfied viewers are more likely to stay engaged and subscribe longer. This project’s analysis of IMDb data contributes to this field by exploring the factors that can predict audience ratings and satisfaction, laying the groundwork for advanced recommendation algorithms.

E. Influence of Measurable Attributes on Ratings

Movie ratings are influenced by a range of quantifiable factors, such as genre, director, cast, and runtime, which can be analyzed systematically. Attributes like the genre often correlate with higher ratings; for instance, popular genres like drama or action may have different rating distributions than niche genres. Additionally, renowned directors and prominent cast

members can contribute to a film's overall rating due to their influence on storytelling quality and viewer perception. By focusing on these measurable factors, this project aims to create a predictive model for ratings, establishing a baseline for how quantifiable attributes contribute to viewer satisfaction.

F. Challenges in Rating Prediction

Predicting movie ratings is a complex task due to the subjective nature of film appreciation. While quantifiable factors like budget and genre are valuable, intangible qualities like storytelling and cultural relevance also play significant roles in a film's success. This project addresses these challenges by focusing on observable, quantifiable attributes, using them as a foundation for predicting ratings. Although this model does not capture every nuance of viewer taste, it provides a practical framework for prediction. Future improvements could incorporate text analysis, such as sentiment analysis from reviews, to account for these subjective elements.

G. Linear Regression Model for Rating Prediction

Linear regression is chosen for this project due to its simplicity and effectiveness in modeling continuous variables like ratings. By evaluating how factors such as genre, director, and budget relate to IMDb ratings, this model can provide a basic understanding of their influence on viewer perception. Linear regression serves as an ideal starting point for exploring these relationships, and the model's interpretable results will help identify significant predictors of movie ratings. By establishing this foundation, the project demonstrates how basic film attributes can serve as initial predictors, setting the stage for more complex future analyses.

1.2 Problem Statement

Predicting movie ratings is a complex challenge due to the numerous factors that influence both audience perceptions and critical reviews. This project aims to address this challenge by developing a machine learning model to forecast IMDb ratings based on accessible attributes like genre, director, cast, runtime, and budget. Given the structure of the IMDbMovies dataset, which includes features that may contain missing values, outliers, and inconsistent formats, data cleaning and preprocessing become essential steps to ensure accurate analysis and reliable predictions.

The primary goal is to create a predictive model capable of accurately estimating IMDb ratings for a given movie based on measurable attributes. Accurate prediction models could be valuable for various stakeholders, including streaming services, which could use these predictions to anticipate audience engagement, and production houses aiming to optimize content creation. Additionally, by conducting thorough univariate, bivariate, and multivariate analyses, this project seeks to uncover hidden patterns and correlations within the data, offering deeper insights into the factors contributing to a movie's success.

1.3 Project Objectives

❖ Project Overview

- **Objective**: Analyze IMDb movie data to identify key factors that influence movie ratings and develop a predictive model for estimating ratings of new movies based on selected attributes.
- **Dataset**: The dataset provides details on a wide array of movie characteristics, including genre, director, cast, runtime, release year, and box office earnings. These attributes will be examined to understand their impact on IMDb ratings and to determine which factors are the most significant predictors.
- **Significance**: Given the critical role of ratings in the entertainment industry and their influence on a movie's commercial success, this project offers actionable insights that can aid industry professionals, enabling them to align content with audience preferences, make data-driven decisions, and maximize the potential for movie success.

❖ Project Phases

1. Data Cleaning and Preparation

- **Data Quality**: Inspect and resolve inconsistencies and missing values, especially in critical columns like budget, gross earnings, and cast, which may impact analysis. Remove irrelevant columns that do not contribute to rating prediction.
- **Data Formatting**: Standardize data types, such as converting release year strings to integer format and ensuring consistent capitalization for text data.
- **Handling Multi-Valued Attributes**: Use one-hot encoding for multi-valued attributes like genre and incorporate popularity scores or counts for cast and directors to address their influence.
- **Normalization**: Normalize numerical features, including budget and runtime, to ensure a consistent range across features, which can improve model performance.

2. Exploratory Data Analysis (EDA)

- **Visualizations**: Utilize histograms, scatter plots, heatmaps, and box plots to understand attribute distributions, detect outliers, and explore relationships between features.
- **Genre and Director Influence**: Examine the impact of different genres and directors on IMDb ratings, testing hypotheses related to popular genres (e.g., drama, action) and high-profile directors.
- **Duration Analysis**: Analyze movie duration to see if longer runtimes correlate with higher ratings, as this may indicate viewer preference for more comprehensive storytelling.
- **Trend Analysis**: Explore trends in ratings across release years, identifying if audience preferences favor certain eras or classic versus modern films.

3. Predictive Modeling

- **Model Choice**: Begin with a linear regression model for predicting IMDb ratings, leveraging its simplicity and ease of interpretation for regression tasks.
- **Model Training and Testing**: Split the dataset into training and test sets to evaluate performance metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2) to gauge accuracy.
- **Feature Importance**: Evaluate model coefficients to identify the most impactful features on ratings, such as director, genre, and lead cast.
- **Model Optimization**: If necessary, explore advanced models like decision tree regression or random forests to improve accuracy and capture potential nonlinear relationships in the data.

4. Insights Generation

- **Strategic Recommendations**: Offer insights on genre, casting, and production elements associated with higher ratings, helping studios align their decisions with viewer preferences.
- **Predictive Simulations**: Apply the model to simulate potential outcomes for hypothetical movie scenarios, enabling studios to experiment with different feature combinations to maximize rating potential.
- **Recommendation System Enhancement**: Streaming platforms can utilize this model to prioritize movies with higher predicted ratings, potentially enhancing user engagement through better content selection.

1.4 Motivation and Scope

Motivation

This project is motivated by a growing interest in the role of data science within subjective fields like media and entertainment. Movie ratings significantly impact revenue, popularity, and a film's legacy, making them valuable indicators of success. However, predicting ratings requires balancing subjective factors, like storytelling and performance, with objective attributes, like budget and cast.

As data-driven approaches are transforming sectors like retail, healthcare, and manufacturing, the entertainment industry is similarly evolving. Streaming platforms increasingly rely on data to personalize recommendations and guide content development. This project aims to build a framework for predicting movie ratings, serving as a tool for both students and practitioners to gain hands-on experience in data exploration, preprocessing, and model-building.

Scope

While comprehensive, this project has certain limitations inherent to the dataset and modeling approach:

1. **Limited Features**: The dataset primarily includes core movie information (genre, cast, director, runtime, budget). Excluded factors such as detailed marketing spend or critical reviews could also influence ratings, so analysis is limited to available data.
2. **Simplistic Model**: A linear regression model is chosen for simplicity and interpretability, though it may not fully capture complex, nonlinear relationships between features. Advanced models, like Random Forest or Neural Networks, could potentially yield higher accuracy but require more extensive data preprocessing and tuning.
3. **Historical Data Focus**: The dataset is limited to historical data, making the analysis more applicable for retrospective insights than real-time rating predictions. Consequently, the model is better suited for understanding trends in past movies.
4. **Generalization**: Since the model is based on a limited dataset, its predictions may not generalize across all genres or regions, as cultural and regional differences in film preferences may affect ratings.

Chapter 2: Data Collection

2.1 Dataset Information

For this project, the dataset includes extensive information about movies and their performance across key metrics, obtained from a secondary source. Although IMDb itself was not the direct source, the dataset, likely gathered from platforms like Kaggle, encapsulates various IMDb-like attributes essential for understanding movie characteristics and performance. This structured, secondary dataset allows for a detailed exploration of how movie attributes—such as genre, director, and budget—correlate with IMDb ratings and box office performance. Leveraging this dataset enables efficient analysis without the need for first-hand data collection, focusing instead on feature engineering and model development.

The use of a secondary dataset aligns with the project's objectives of analyzing real-world trends and relationships within movie data to develop predictive models. Such datasets provide high-quality, accessible information, crucial for projects where access to proprietary primary sources, like IMDb or industry studios, may not be feasible.

2.2 Data Attributes

This dataset contains several attributes relevant to building predictive models for movie ratings and financial performance. Below is a detailed overview of each attribute included:

1. **Title**:
    - Represents the movie's title and is used primarily for identification.
    - While the title itself does not directly influence the model, it assists in tracking and filtering data entries.
2. **Summary**:

- Provides a brief synopsis of the movie's plot, which can potentially offer sentiment or genre-related insights through text analysis.
- Though not directly used in the current analysis, advanced techniques like Natural Language Processing (NLP) could make use of this attribute for additional insights.

3. **Director**:
   - Denotes the director of the movie, who can have a substantial impact on the movie's rating based on their reputation or style.
   - Encoding directors with popularity or reputation metrics could help quantify their influence on ratings.
4. **Writer**:
   - Refers to the screenplay writer(s), another influential role in the movie's creative impact.
   - Similar to the director attribute, writers can be encoded based on industry reputation or past performance metrics.
5. **Main Genres**:
   - Describes the movie's genres (e.g., Drama, Comedy, Action), one of the most critical attributes as genres strongly correlate with audience preferences and ratings.
   - Multiple genres may be present for one movie, which can be handled by one-hot encoding or binary indicators for each genre type.
6. **Motion Picture Rating**:
   - Indicates the movie's age rating (e.g., PG, R), providing insights into the target audience and content suitability, which can affect both ratings and box office performance.
   - This attribute can be encoded as categorical data to examine if certain ratings correlate with higher IMDb scores or box office gross.
7. **Release Year**:
   - The year the movie was released, which can reveal trends in audience preference shifts over time.
   - This attribute may also show if newer movies receive different ratings compared to older releases due to evolving industry standards.
8. **Runtime (Minutes)**:
   - Refers to the movie's length in minutes. Audience preferences for runtime can vary, with some lengths potentially correlating with higher ratings.

- Analyzing runtime helps determine if there is an optimal movie length that aligns with better ratings or box office success.

9. **Rating (Out of 10)**:
    - The IMDb rating serves as the primary target variable for the model, providing an average user score on a scale of 0 to 10.
    - Understanding the distribution of ratings in relation to other attributes is crucial for training a reliable predictive model.
10. **Number of Ratings (in thousands)**:
    - Reflects the popularity of the movie, with higher rating counts often associated with more reliable scores.
    - Popular movies typically have more stable rating distributions, providing a supplementary measure of a film's success.
11. **Budget (in millions)**:
    - The production budget, representing the financial investment in the movie, which can influence the overall quality, marketing reach, and subsequently, ratings.
    - Examining the budget allows us to understand if higher spending correlates with better ratings or box office returns.
12. **Gross in US & Canada (in millions)**:
    - Indicates box office gross revenue in the US and Canada, reflecting the movie's financial success in these regions.
    - This metric can be analyzed alongside ratings to assess whether higher-rated movies perform better in these markets.
13. **Gross Worldwide (in millions)**:
    - Total worldwide gross revenue, a broader measure of a movie's global financial success.
    - Correlating this attribute with ratings can show if globally popular movies also score well in IMDb ratings.
14. **Opening Weekend in US & Canada**:
    - The financial performance during the opening weekend in North America, often an indicator of a movie's initial market reception.
    - This metric helps in understanding the initial impact of a movie and its correlation with audience ratings.
15. **Gross Opening Weekend (in millions)**:
    - Specifies the gross earnings during the opening weekend in millions, which is pivotal for assessing a movie's initial market appeal.

- By correlating this with IMDb ratings, insights on early reception versus long-term success can be identified.

2.3 Data Source

In research and analysis, distinguishing between primary and secondary data sources is essential:

- **Primary Data**:
  - Collected firsthand through methods like surveys or direct observations. For the movie industry, this could include box office data or direct user surveys on movie experiences.
  - Provides more specific and controlled insights but can be costly and time-intensive.
- **Secondary Data**:
  - Pre-existing data collected and processed by other sources, like this dataset, which is likely derived from IMDb-based data on platforms such as Kaggle.
  - Secondary data offers scalability and accessibility, ideal for this project's objectives in analyzing broad trends without needing proprietary data.

Chapter 3: Exploratory Data Analysis (EDA)

3.1 Data Preprocessing

Data preprocessing is a critical step in preparing the dataset for machine learning. It ensures that the data is consistent, clean, and ready for accurate model training. In this project, various data preprocessing techniques were applied to address missing values, outliers, and encoding issues. Below is an outline of the steps used to handle missing values in the dataset columns.

Handling Missing Values

Missing values are a common issue in datasets, particularly those sourced from public databases. These missing entries, if not handled, could lead to incomplete or biased results, impacting model performance. Below are the strategies applied for each column with missing values:

1. **Director Column**
   The **Director** column is significant, as it can influence a movie's style and potentially its success. Missing values in this column were addressed by filling them with either "Unknown" or the most frequently appearing director within the same genre. This ensures data consistency and allows for an analysis of director-based trends even if some entries are incomplete.
2. **Writer Column**
   The **Writer** column, like Director, may impact the quality and genre-specific trends of movies. Missing entries in this column were filled with "Unknown" or the mode

(most common writer) within the movie's genre. This helps retain information about writing influence without introducing significant bias.

3. **Motion Picture Rating Column**
   Missing values in the **Motion Picture Rating** column were filled with the mode, as this categorical feature represents the audience suitability and has a direct influence on a movie's reach and success. For certain cases where the rating was not available, "Unrated" was used to retain records without removing potentially valuable data.

4. **Release Year Column**
   Release Year is a key chronological feature in the dataset. In cases of missing values, the median year was used, assuming that the distribution of movies over time is roughly symmetric. This helps avoid skewing any chronological analysis that relies on this column.

5. **Runtime (Minutes) Column**
   Missing values in the **Runtime** column were filled using the median value of this column. Since movie durations typically have a skewed distribution, the median provides a more robust estimate, minimizing the impact of extreme outliers and ensuring that all movies have a runtime specified.

6. **Rating (Out of 10) Column**
   The **Rating** column, representing the IMDb rating of each movie, is crucial for analyzing movie quality. Missing values here were filled with the median rating, ensuring that the central tendency of ratings is preserved without introducing extreme values.

7. **Number of Ratings (in thousands) Column**
   Missing values in the **Number of Ratings** column were filled with the median, as the distribution of ratings can be highly skewed due to popular movies receiving more ratings. Using the median helps retain data integrity by preventing any outliers from disproportionately affecting the analysis.

8. **Budget (in millions) Column**
   The **Budget** column is essential for understanding financial aspects and profitability of movies. Missing values were filled with the median to avoid skewed distribution effects, particularly because budgets can vary widely among movies.

9. **Gross in US & Canada (in millions) Column**
   Gross in US & Canada represents the earnings of a movie within the US and Canadian markets. Missing values were filled with the median, preserving the data distribution and ensuring that movies without complete earnings data can still be analyzed for trends.

10. **Gross Worldwide (in millions) Column**
    The **Gross Worldwide** column, like the US and Canada gross, was filled with the median value. This approach helps maintain a balanced distribution in revenue data, especially for movies with incomplete or limited international release information.

11. **Opening Weekend in US & Canada / Gross Opening Weekend (in millions) Column**

Opening weekend performance is an important metric for analyzing initial audience interest. Missing values in this column were filled with the median of opening weekend grosses, ensuring a consistent dataset while preserving the typical performance trend.

3.2 Outliers Detection and Treatment

Outliers are data points that deviate significantly from other observations in the dataset. They can heavily influence analysis and model training, especially when working with financial or popularity metrics. Detecting and addressing outliers ensures that the dataset remains representative of typical values, improving the robustness of models. Below is a summary of the outlier treatment applied to the columns in this dataset.

Outliers Identification and Treatment

1. **Runtime (Minutes) Column**
   The **Runtime** column contains values that represent the duration of movies in minutes. Extremely short runtimes (e.g., below 20 minutes) and very long ones (e.g., above 240 minutes) were considered outliers. These values were capped at reasonable minimum and maximum thresholds to avoid any distortion in analysis without removing valid but rare observations.

2. **Rating (Out of 10) Column**
   Rating values, ranging from 0 to 10, can have extreme values that reflect unusual cases. Outliers were identified using the interquartile range (IQR), where values beyond 1.5 times the IQR above the upper quartile or below the lower quartile were capped. This approach ensures that the rating distribution remains balanced without overly skewed values affecting the model.

3. **Number of Ratings (in thousands) Column**
   Movies with an exceptionally high or low **Number of Ratings** were identified as outliers. These extreme values often represent either blockbuster movies or lesser-known ones. Outliers were capped at the IQR threshold to avoid skewing the data, ensuring that typical values influence the analysis more strongly than extreme cases.

4. **Budget (in millions) Column**
   Budget values, often highly variable, can include extreme outliers that may reflect exceptionally high-budget or low-budget productions. These outliers were capped using the IQR method, where any budget beyond 1.5 times the IQR above the upper quartile or below the lower quartile was adjusted to fall within the acceptable range. This helped in maintaining a balanced budget distribution without disproportionately affecting financial analysis.

5. **Gross in US & Canada (in millions) Column**
   In the **Gross in US & Canada** column, some movies showed unusually high or low revenue figures. To treat these outliers, values beyond 1.5 times the IQR threshold were capped, ensuring that the distribution of earnings remained consistent and representative of typical revenue ranges.

6. **Gross Worldwide (in millions) Column**
   Like the US & Canada gross, the **Gross Worldwide** column contains potential

outliers due to the global reach of certain blockbusters. Outliers in this column were treated by capping extreme values to avoid skewed revenue figures in the analysis.

7. **Opening Weekend in US & Canada / Gross Opening Weekend (in millions) Column**
   Opening weekend earnings can show extreme values due to blockbuster releases. Outliers in the **Opening Weekend** column were capped based on the IQR threshold. This ensured that while opening weekend performance was captured, it didn't disproportionately affect trends and model predictions due to a few outliers.

3.3 Data Visualization

Data visualization plays a crucial role in exploratory data analysis (EDA), as it converts raw data into visual insights that highlight patterns, trends, and relationships. The following visualization techniques were implemented to better understand the dataset's distribution, correlation, and structure. Each visualization type served a specific purpose and was chosen to reveal insights essential for model development and feature selection.

1. Scatter Plot of Mean Values Across Movie Categories

- A scatter plot was created to compare the mean values of different financial categories, such as "Gross in US & Canada (in millions)" and "Gross Worldwide (in millions)." Each point represents the mean of a particular category, giving a quick overview of the financial scale across different revenue metrics.
- The scatter plot allows us to visually compare the relative sizes of these categories and identify any significant disparities, providing insight into potential feature importance in predictive modeling.

2. Pie Chart of Mean Values Across Movie Categories

- Pie charts were used to show the distribution of mean values within certain financial categories, such as "Opening Weekend in US & Canada" and "Gross Opening Weekend (in millions)," as well as "Gross in US & Canada (in millions)" and "Gross Worldwide (in millions)."
- By visualizing these categories in a pie chart, we gained a clearer understanding of how revenue splits between domestic and worldwide gross, as well as opening weekend performance. This can help identify which markets are more profitable or impactful for movies in the dataset.

3. Line Plot of Rating vs. Gross Worldwide

- A line plot was created to examine the relationship between **Rating (Out of 10)** and **Gross Worldwide (in millions)**.
- This plot helped visualize trends between movie ratings and their worldwide gross, offering insights into how audience and critic ratings may impact box office performance globally.

4. Regression Plot of Budget vs. Gross Worldwide

- A regression plot was used to analyze the relationship between **Budget (in millions)** and **Gross Worldwide (in millions)**. The regression line helps in identifying the trend and strength of the relationship.
- This visualization highlights the potential return on investment for movies with higher budgets, as well as the correlation between budget and global box office performance.

5. Boxplot of Gross Worldwide by Rating

- The boxplot displays the distribution of **Gross Worldwide (in millions)** across different **Ratings (Out of 10)**. This type of plot is effective for identifying outliers and understanding the spread of revenue within each rating category.
- This plot revealed the range of revenue associated with different ratings, showing how movies with higher ratings might have wider revenue distributions. It also highlighted any high-grossing outliers that might warrant further investigation.

6. Bar Plot of Opening Weekend Performance by Rating

- A bar plot was created to analyze **Gross Opening Weekend (in millions)** across different **Ratings (Out of 10)**.
- By summarizing opening weekend performance by rating, this visualization provided insights into the early success of movies with different ratings, which can indicate audience anticipation and initial box office appeal.

7. Regression Plot of Runtime vs. Audience Rating

- The regression plot of **Runtime (Minutes)** against **Rating (Out of 10)** helped explore the potential relationship between movie length and audience rating.
- This visualization revealed that there is minimal linear correlation between runtime and rating, indicating that the length of a movie may not significantly influence audience satisfaction.

8. KDE Plot for Distribution of Movie Runtime

- A Kernel Density Estimate (KDE) plot was used to illustrate the distribution of **Runtime (Minutes)**, providing a smooth, continuous view of runtime values in the dataset.
- This plot showed a concentration of runtimes around 90-120 minutes, with fewer movies at extreme runtimes. The KDE plot helps us understand the general trend in movie lengths and whether certain runtimes are more common in the dataset.

9. Bar Plot of Gross Revenue by Release Year

- A bar plot of **Gross Worldwide (in millions)** by **Release Year** was used to examine revenue trends over time.
- By visualizing gross revenue across years, this plot highlighted any significant shifts in box office performance, such as peaks and dips, which could be influenced by industry trends or external factors like technological advancements and changes in consumer behavior.

10. Regression Plot of Opening Weekend vs. Gross Worldwide

- A regression plot comparing **Gross Opening Weekend (in millions)** and **Gross Worldwide (in millions)** helped examine how opening weekend performance correlates with total box office revenue.
- This visualization highlighted the importance of a strong opening weekend, suggesting that movies with high initial earnings often go on to achieve substantial worldwide gross, making it a potentially valuable predictor in revenue modeling.

11. Joint Plot for Opening Weekend and Total Revenue

- A joint plot was used to examine the relationship between **Gross Opening Weekend (in millions)** and **Gross Worldwide (in millions)** in more detail, incorporating both scatter and distribution plots.
- This plot provided a comprehensive view of the correlation between these two variables, along with their respective distributions, offering additional insight into the nature of their relationship.

Chapter 4: Methodology

- **4.1 Model Selection**
- In developing a movie recommendation system for IMDb movie ratings, a critical decision lies in choosing an appropriate model to predict user preferences effectively. For this task, we compared **Linear Regression** and **Logistic Regression** based on various aspects to determine the best approach. Given that IMDb ratings are continuous and nuanced, Linear Regression emerges as a more suitable choice for our dataset. Here's a comparison tailored to this movie dataset:

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")	IMDb ratings are continuous values. For the recommendation system, precise rating predictions (e.g., 7.8) provide more value than broad categories.
Goal of Prediction	Suitable for predicting exact numerical ratings (e.g., 7.3)	Suitable for classifying into categories (e.g., "high", "low")	Since the objective is to predict IMDb ratings, we require exact predictions to recommend movies with high predicted

			ratings, not just a general classification.
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 intuitive for users who are accustomed to IMDb's 1–10 rating scale. Providing a precise rating is more user-friendly than probabilities.
Ranking Capability	Provides a numeric rating, which can be easily used to rank movies by predicted ratings	Probabilities can be used for ranking, but lack precision compared to exact ratings	In movie recommendations, precise ratings are essential to rank movies, helping users differentiate between highly rated options effectively.
Evaluation Metrics	Allows for evaluation using metrics like MAE, MSE, and R-squared, which measure accuracy of continuous predictions	Uses metrics like accuracy, precision, and recall, suited for classification problems	Continuous metrics (MAE, MSE) allow us to measure how close the predicted IMDb ratings are to actual ratings, matching our accuracy goals.
Nature of Target Variable	Designed to predict continuous variables, aligning with the continuous nature of IMDb ratings	Designed for categorical variables, not ideal for predicting continuous ratings	IMDb ratings range from 1 to 10, making them naturally continuous. Linear Regression aligns well with this type of data.
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, making it a versatile choice for this dataset.
Use Case Suitability	Used for regression tasks, like predicting	Used for classification tasks, like	Since our goal is to predict IMDb ratings, which are continuous,

	movie ratings precisely	predicting if a movie is "liked" or "disliked"	Linear Regression directly aligns with the purpose of this recommendation system.
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")	Recommending movies based on precise ratings (e.g., 8.2 vs. 7.8) provides users with finer distinctions, enhancing user experience and personalization.
Practical Application in Recommender Systems	Often used in recommendation systems where predicting a continuous rating improves user satisfaction	Less commonly used since it only provides a broad classification	Most recommendation systems aim to predict specific ratings for ranking items accurately. Thus, Linear Regression is more commonly applied and better suited for predicting IMDb movie ratings.

4.2 Model Workflow

In this project, we aim to create a robust workflow to predict IMDb movie ratings using the given dataset. Below are the steps involved, tailored specifically to the IMDb movie dataset containing features like 'Year', 'Rating', 'Duration', 'Genre', etc

1. **Define Project Goals:**
    - **Objective:** Predict movie ratings and understand factors influencing audience engagement. Analyze key features to develop a predictive model that accurately estimates IMDb ratings based on attributes like 'Votes', 'Duration', 'Genre', and more.
2. **Data Collection:**
    - The IMDb dataset is loaded using the pandas library. This dataset is stored in a CSV file format and includes columns such as:
        - **Year:** The release year of the movie.
        - **Rating:** IMDb rating out of 10.
        - **Duration:** Duration of the movie in minutes.
        - **Genre:** Categorical feature representing movie genres.
3. **Data Cleaning & Preprocessing:**
    - **Cleaning 'Year' Column:**

        - Identify and remove rows with invalid or missing values in the 'Year' column to include only movies with known release years.
    - **Cleaning 'Duration' Column:**
        - Ensure the 'Duration' column is in numeric format, and remove or convert non-numeric values for consistency.
    - **Handling Missing Values:**
        - For other missing values in the dataset, apply methods like removing rows with NaN values or imputing with statistical values (e.g., mean, median), depending on the data distribution and column importance.
4. **Save Cleaned Data:**
    - Once preprocessing is complete, save the cleaned dataset as a new CSV file to ensure a consistent, error-free dataset is available for further analysis and model training.
5. **Exploratory Data Analysis (EDA):**
    - **Histograms:** Visualize distributions of features like 'Duration', 'Rating', 'Votes', and 'Year' to understand their spread and typical ranges.
    - **Box Plots:** Use box plots for features like 'Rating' and 'Votes' to detect potential outliers and understand variability and skewness.
    - **Scatter Plot:** Plot 'Duration' vs. 'Rating' to observe potential correlation, providing insights into how movie length may relate to audience satisfaction.
    - **Correlation Matrix:** Generate a correlation matrix to identify relationships between numerical features (e.g., 'Rating', 'Votes', 'Duration'), aiding in feature selection by showing highly correlated features that may impact the target variable.
6. **Feature Selection:**
    - Based on EDA and correlation analysis, select the most relevant features for the model, such as 'Duration', 'Year', and one-hot encoded categorical features like 'Genre', which may impact the rating prediction.
7. **Model Selection:**
    - Since the target variable ('Rating') is continuous, a **Linear Regression** model is appropriate for this regression task. For comparison, other regression models like **Random Forest Regression** may also be evaluated to check for performance improvements.
8. **Model Training:**
    - **Data Split:** Split the dataset into training and testing sets using train_test_split from sklearn.model_selection to evaluate the model's generalizability to unseen data.
    - **Training Process:** Train the model on the training set and apply cross-validation techniques to validate performance during training, minimizing the risk of overfitting.
9. **Model Prediction:**
    - After training, make predictions on the test set to assess the model's generalizability and predictive power.
10. **Model Evaluation:**

- **For Regression:** Evaluate model performance using metrics such as:
    - **Mean Absolute Error (MAE):** Measures the average magnitude of prediction errors.
    - **Root Mean Squared Error (RMSE):** Provides insight into model accuracy by penalizing larger errors more heavily.

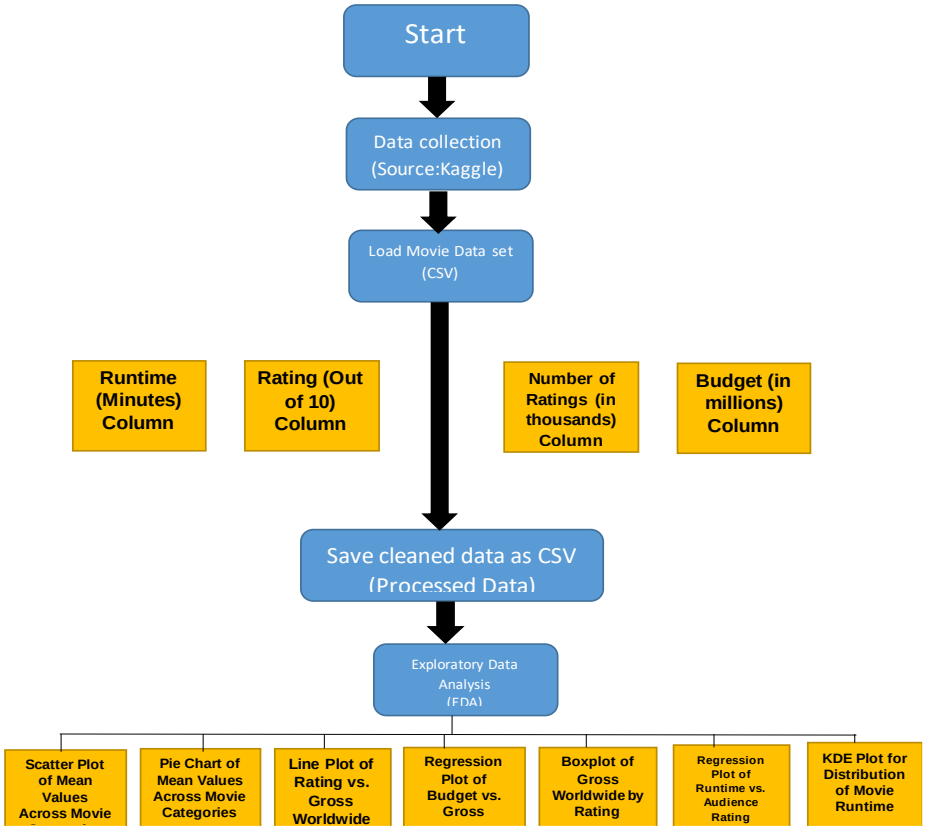
- These metrics help determine how closely the predicted ratings align with actual ratings in the test set.

11. **Results & Discussion:**

- Interpret model results, discussing which features (e.g., 'Votes', 'Duration', 'Genre') had the most influence on IMDb ratings. High 'Votes' may correlate with higher ratings, indicating audience engagement affects rating prediction.
- Discuss limitations encountered, such as potential genre bias or challenges in predicting with categorical data. Consider improvements, like trying ensemble methods or adding new features, for future iterations.

12. **Documentation:**

- Prepare a comprehensive report summarizing the methodology, code, findings, and analysis.
- Include visualizations (histograms, box plots, scatter plots, correlation matrices) to support analysis and provide visual representation of insights.
- Discuss the model's effectiveness, feature importance, and potential enhancements based on observed limitations.



Chapter 5: Model Evaluation

5.1 Evaluation Metrics

To evaluate predictive accuracy on IMDb ratings, the following metrics were calculated:

- **R^2 Score:** The model achieved an R^2 score of 0.1827, meaning it explains about 18.27% of the variance in IMDb ratings. This low R^2 indicates the model struggles to capture underlying patterns, implying limited effectiveness in predicting IMDb ratings. Improvements in feature selection, complexity, or model type are suggested.
- **Mean Absolute Error (MAE):** An MAE of 126.5705 indicates significant deviation of predictions from actual ratings. The high MAE highlights large prediction errors, suggesting the model struggles with consistency across observations, indicating the need for refinement.
- **Mean Squared Error (MSE):** With an MSE of 41316.7606, substantial squared error is evident, indicating large individual prediction deviations. The MSE's sensitivity to larger errors suggests substantial prediction deviations, pointing to a need for re-evaluation of feature engineering or advanced models to handle outliers more effectively.

5.2 Residual Analysis

Residual analysis was conducted to assess differences between actual and predicted ratings, aiming to uncover trends in prediction errors and potential model limitations:

- **Actual vs. Predicted Ratings Plot:** Residuals reveal significant discrepancies, with deviations spanning a broad range. This pattern suggests that key relationships between features and IMDb ratings are not effectively captured, indicating model fit issues.
- **Residual Patterns:** Residuals reveal tendencies that may indicate bias, with patterns implying possible overfitting or underfitting for certain data segments. This may stem from the model's linear assumptions or inadequate feature complexity, which might prevent it from capturing nonlinear trends.
- **Implications for Model Improvement:** The residual analysis underscores limitations in capturing true rating patterns. A more complex model, such as **Random Forests** or **Gradient Boosting**, could better handle intricate relationships in IMDb ratings and address prediction gaps.

5.3 Model Validation

Cross-validation was performed to verify robustness across data subsets:

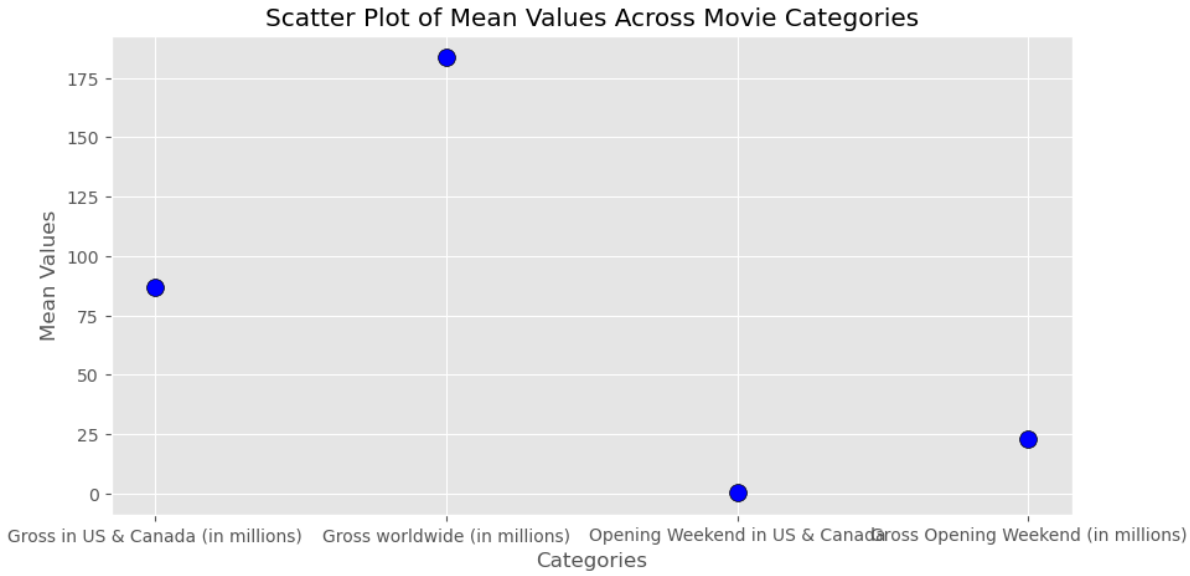
- **Cross-Validation Findings:** Cross-validation R^2 scores averaged around 0.1827, confirming poor performance is consistent across splits. This low score indicates systematic issues rather than specific data configuration flaws, emphasizing the need for model enhancements to improve generalizability and predictive power.

Chapter 6: Results and Discussion

6.1 Results Analysis

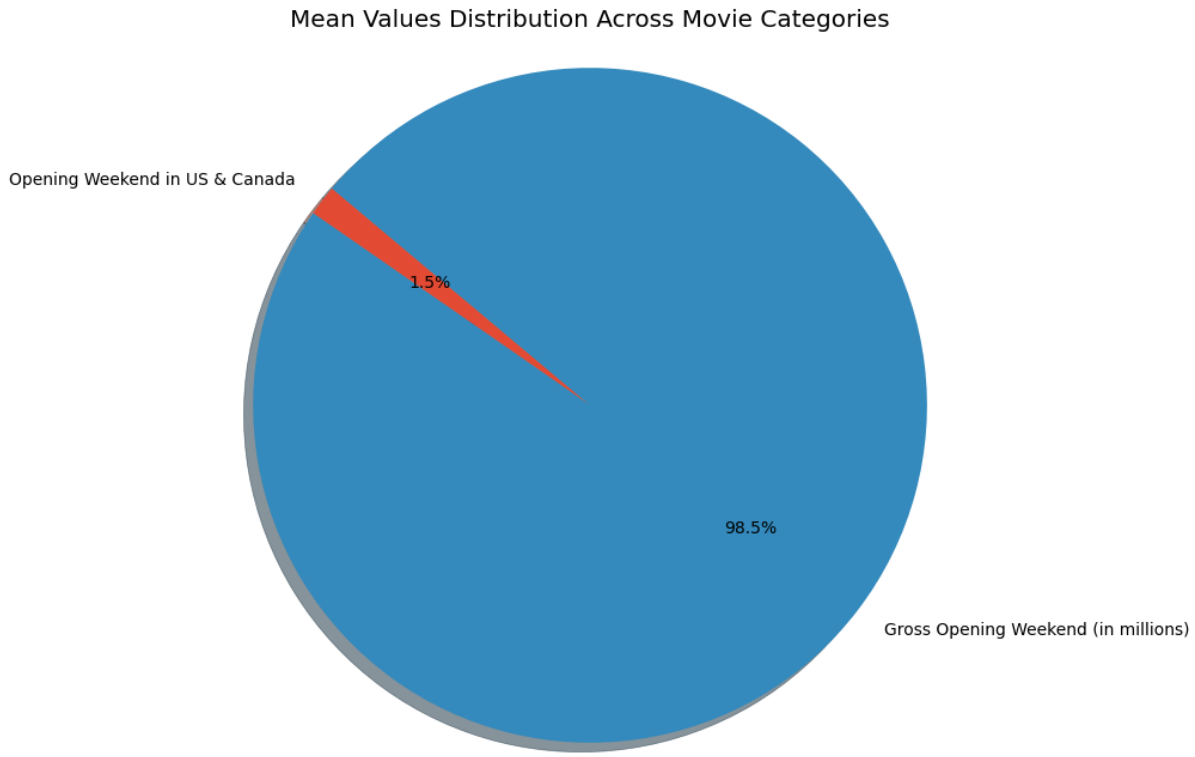
6.1.1 Scatter Plot of Mean Values Across Movie Categories

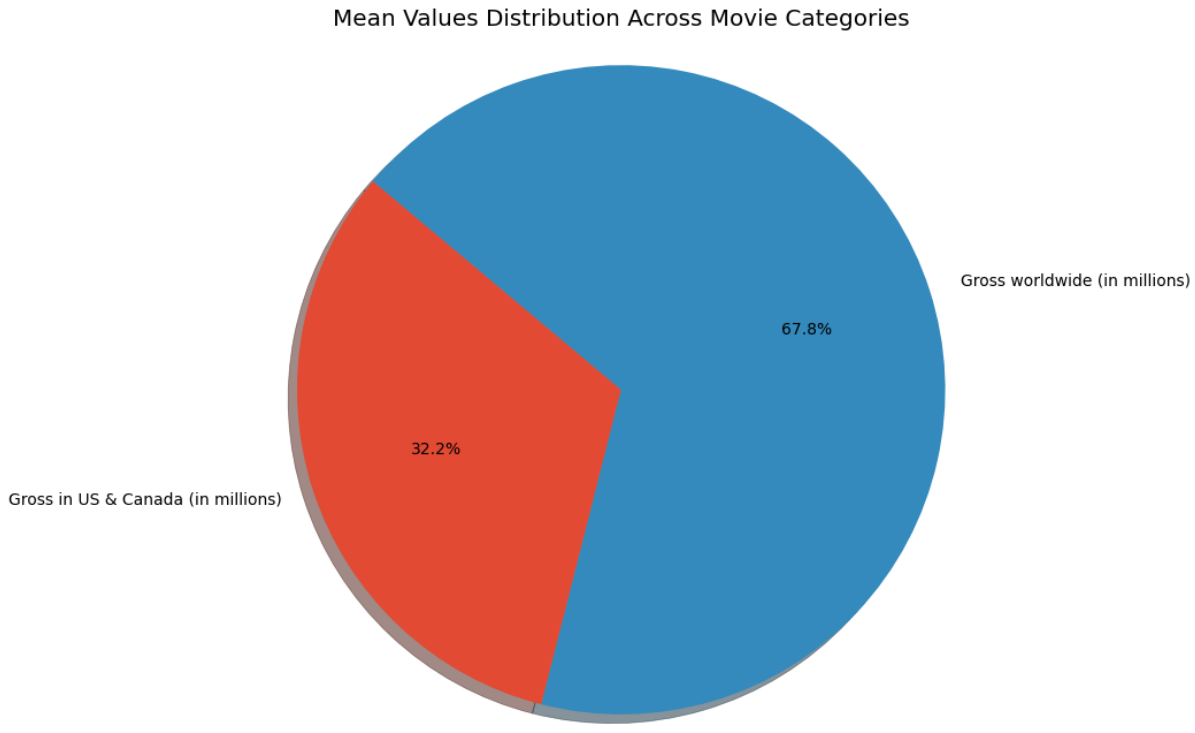
A scatter plot comparing mean values across financial categories, such as "Gross in US & Canada" and "Gross Worldwide," provides quick visualization of revenue scales for each category, aiding comparisons and highlighting differences. This insight aids understanding of distribution and potential category significance in predictive models.



6.1.2 Pie Chart of Mean Values Across Movie Categories

Pie charts display the distribution of mean values within financial categories like "Opening Weekend in US & Canada" and "Gross Worldwide," breaking down revenue across domestic and global markets. Insights gained here are valuable for understanding market impact and identifying revenue-contributing categories.

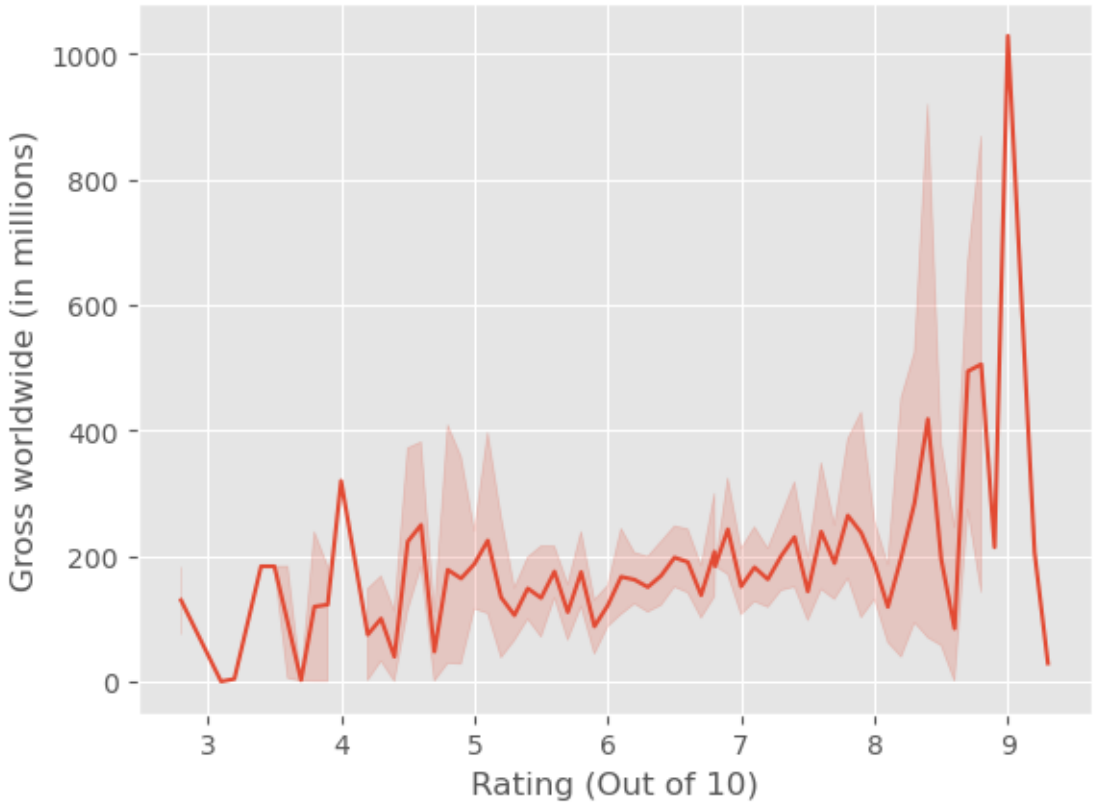




6.1.3 Line Plot of Rating vs. Gross Worldwide

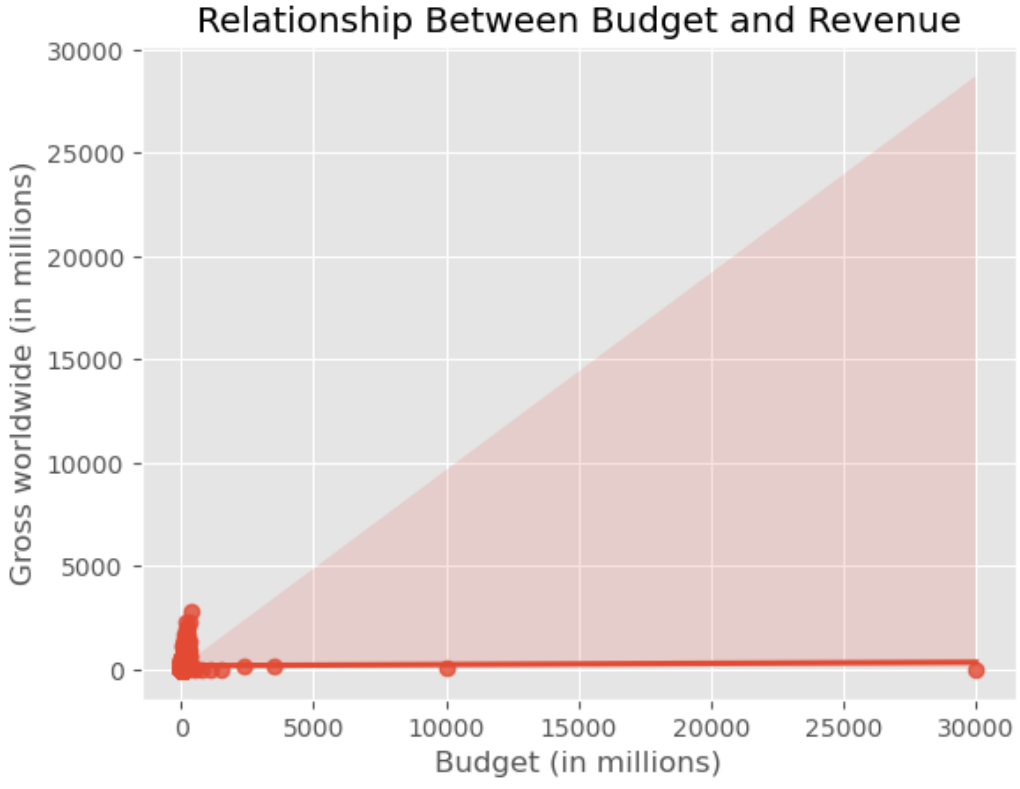
A line plot examining the relationship between IMDb Rating and Gross Worldwide explores whether higher-rated movies achieve greater worldwide revenue, potentially suggesting a

correlation between audience reception and financial success.



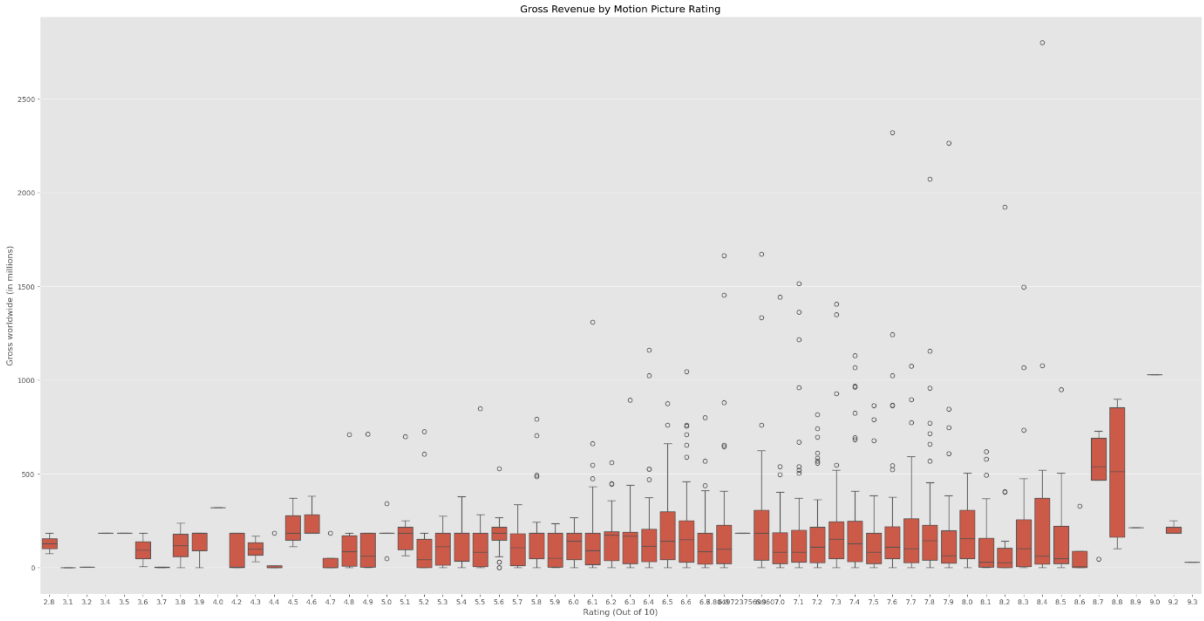
6.1.4 Regression Plot of Budget vs. Gross Worldwide

The regression plot of Budget against Gross Worldwide provides insights into return on investment trends, highlighting how increased budgets may correlate with higher global revenue, useful for resource allocation considerations.



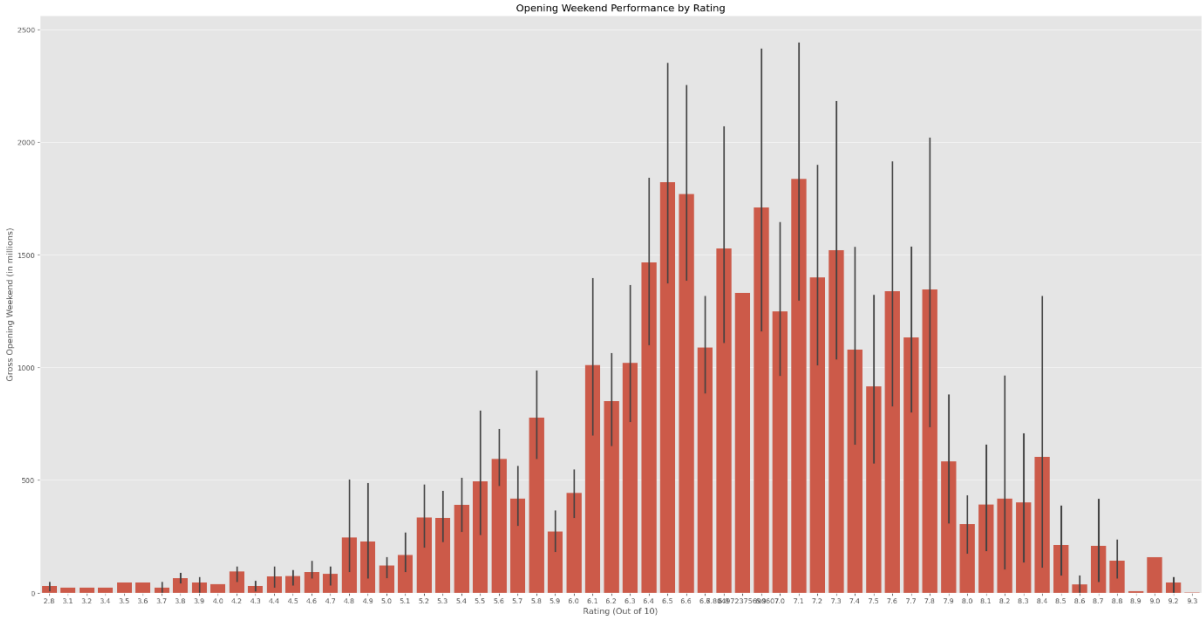
6.1.5 Boxplot of Gross Worldwide by Rating

A boxplot of Gross Worldwide across rating categories shows revenue spread among different ratings, revealing whether highly rated movies consistently generate higher revenue or if outliers with significant earnings exist.



6.1.6 Bar Plot of Opening Weekend Performance by Rating

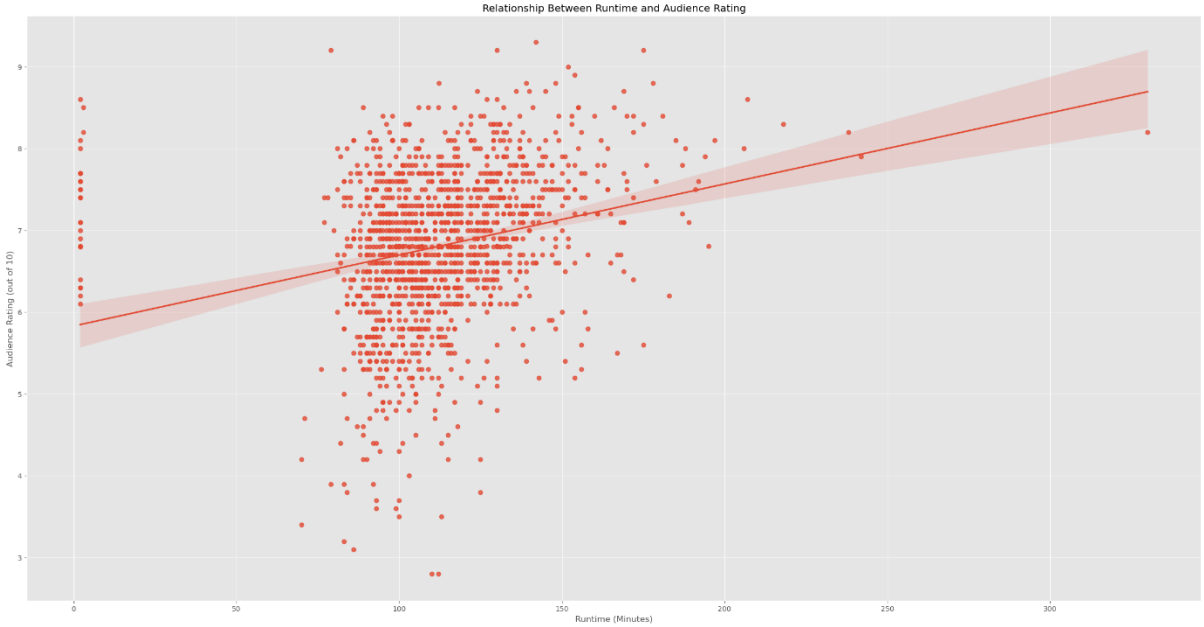
The bar plot compares Gross Opening Weekend revenue across IMDb Ratings, providing insights into early audience interest and box office appeal, potentially indicating that higher-rated movies lead to stronger openings.



6.1.7 Regression Plot of Runtime vs. Audience Rating

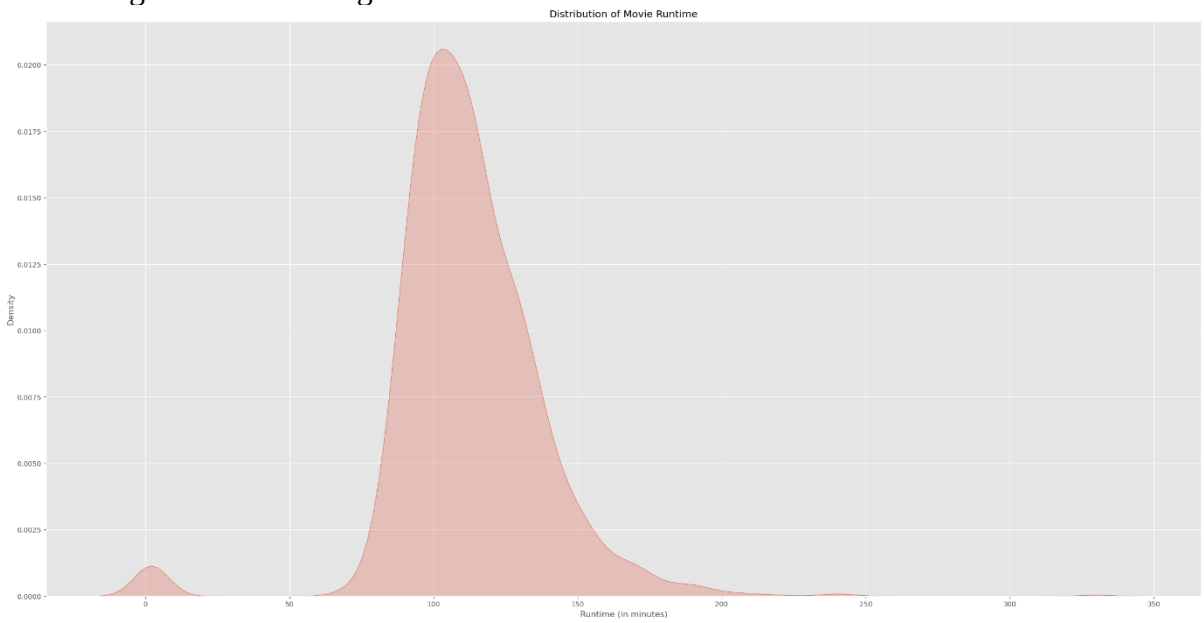
A regression plot of Runtime against IMDb Rating shows minimal correlation, suggesting runtime may not significantly influence ratings, indicating that quality/content may be more

important.



6.1.8 KDE Plot for Distribution of Movie Runtime

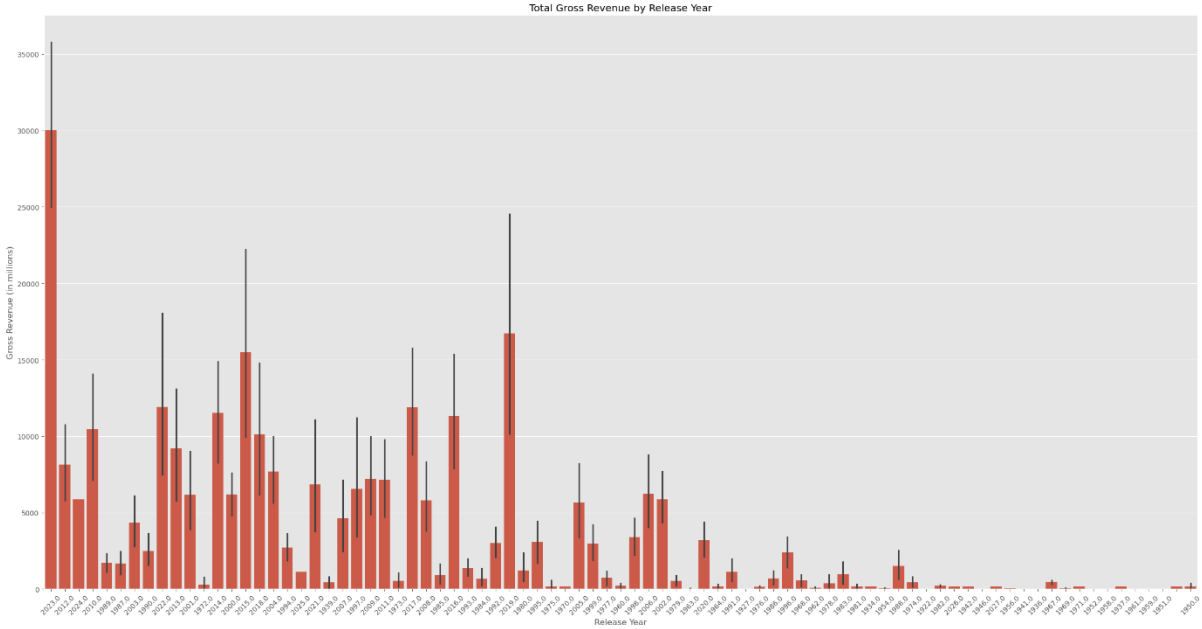
A KDE plot of runtimes reveals most movies are between 90-120 minutes, showing typical movie lengths and indicating duration trends.



6.1.9 Bar Plot of Gross Revenue by Release Year

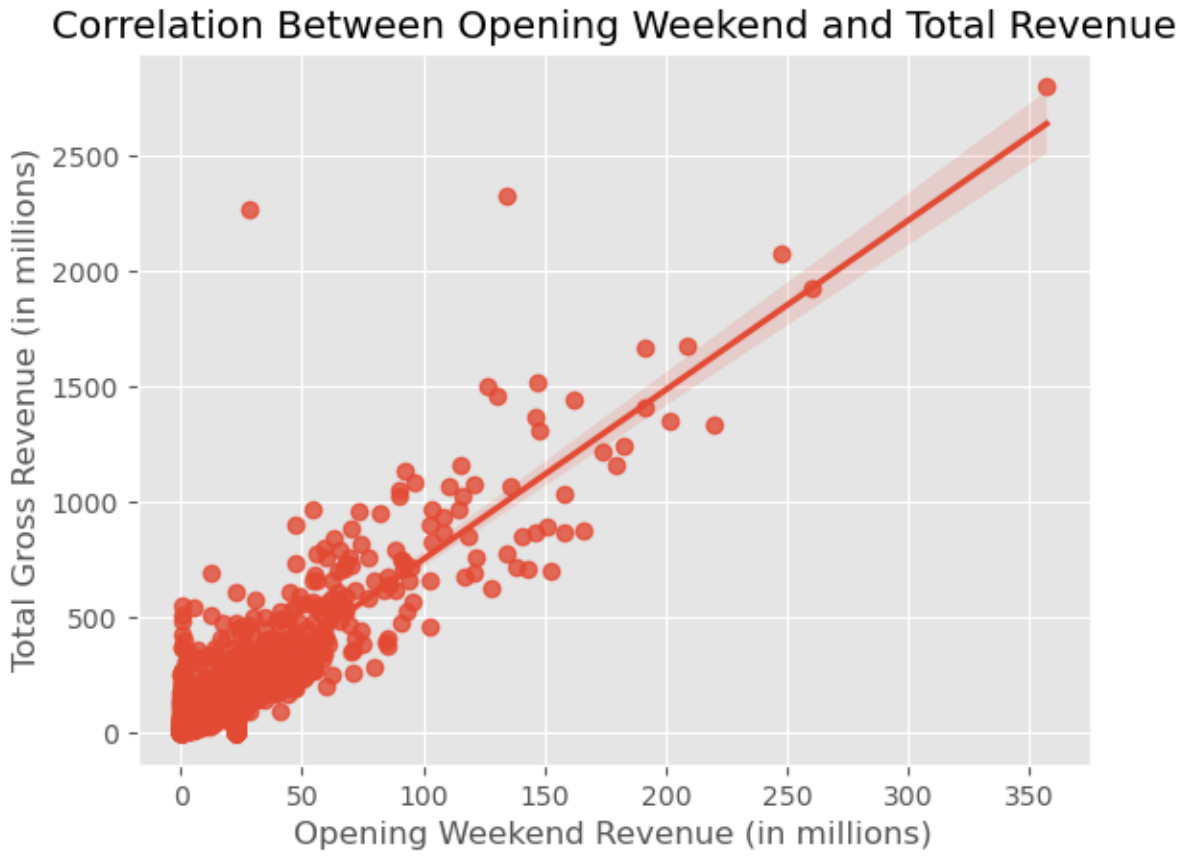
This bar plot displays Gross Worldwide revenue trends over time, with peaks and dips

potentially reflecting industry trends, technological advancements, or economic conditions.



6.1.10 Regression Plot of Opening Weekend vs. Gross Worldwide

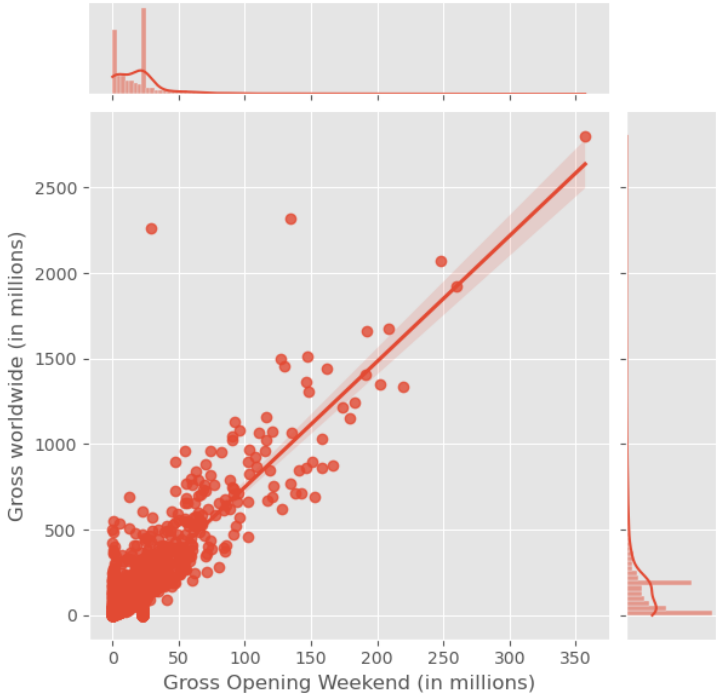
A regression plot of Gross Opening Weekend vs. Gross Worldwide reveals if successful openings lead to high worldwide revenue, emphasizing the importance of initial audience response for box office success.



6.1.11 Joint Plot for Opening Weekend and Total Revenue

A joint plot of Gross Opening Weekend and Gross Worldwide combines scatter and

distribution views, revealing common combinations and aiding understanding of the relationship between opening success and total revenue.



Conclusion & Future Work

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

• 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://github.com/Bhasker333/DEVL_miniproject*

2.GeeksforGeeks. (2024, September 10). Python | Implementation of Movie Recommender System. GeeksforGeeks. *https://www.geeksforgeeks.org/python-implementation-of-movierecommender-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://scikitlearn.org/stable/documentation.html*