Business understanding

Overview

**StreamFlix** is launching a project to develop an advanced movie recommendation system. This hybrid system will blend collaborative filtering with content-based techniques to deliver personalized top 5 movie suggestions based on user ratings, preferences and viewing history. It will also incorporate real-time user interactions and trending content to ensure up-to-date and relevant recommendations. The goals are to enhance user satisfaction, increase watch time and boost customer retention. By offering superior content curation and optimizing content strategies, StreamFlix aims to stand out in the competitive streaming market and strengthen its position as a leader.

Problem Statement

Streamflix is experiencing major issues such as high churn rates and a significant number of users canceling their subscriptions within the first three months. Engagement has decreased as the average watch time per user has declined by 15% over the past year. Users are also struggling with content discovery spending an average of 20 minutes per session browsing before selecting a movie which leads to frustration and reduced satisfaction. Increasing competition from rival streaming services which offer more personalized user experiences is further impacting Streamflix's market share.

Objectives

Main Objective

To develop and deploy a hybrid recommendation system that combines collaborative filtering and content-based filtering techniques to enhance recommendation accuracy by capturing both user-item interactions and movie attributes ensuring a more personalized movie recommendation experience.

Specific Objectives

1. To build a collaborative filtering model using user ratings to generate top 5 movie recommendations, leveraging algorithms such as Singular Value Decomposition (SVD) and k-Nearest Neighbors (k-NN).
2. To address the cold start problem for new users by integrating content-based filtering, utilizing features such as movie genres, directors, and cast.
3. To evaluate the hybrid recommendation system using appropriate metrics like Root Mean Square Error (RMSE), Mean Average Precision (MAP), and Normalized Discounted Cumulative Gain (NDCG) to ensure accuracy and relevance of the recommendations.

Success Metrics

1. Root Mean Square Error (RMSE) < 0.9 for rating predictions
2. Mean Average Precision @5 (MAP@5) > 0.3 for recommended movies
3. Precision@5 of around 0.2 to 0.5
4. Recall@5 of around 0.2 to 0.5
5. F1 Score of around 0.3 to 0.7

Constraints

1. Data Limitations: Small dataset size leading to potential bias due to demographic skew and time period of ratings and lack of additional movie metadata beyond the information provided in the datasets.
2. Data Sparsity: Many users may have rated only a small fraction of available movies creating challenges in finding similar users or items with limited data points.
3. Cold Start Problem: Difficulty providing accurate recommendations for new users or newly added movies, limited effectiveness of collaborative filtering for users with few ratings.
4. A/B Testing Capabilities: Limitations in conducting extensive A/B tests to compare different recommendation algorithms.
5. Ethical Considerations: Ensuring fairness in recommendations across different user groups and balancing business objectives with ethical recommendation practices.
6. Genre and Diversity: Ensuring a balance between accuracy and diversity in recommendations and avoiding over-specialization in user recommendations.
7. Privacy and Data Protection: Ensuring user data privacy and compliance with regulations like GDPR that establishes the general obligations of data controllers and of those processing personal data on their behalf (processors).
8. Evaluation Metrics: Limitations of evaluation metrics in reflecting real-world user satisfaction and a lack of direct user feedback on recommendation quality.

Data Understanding

The data used in this project is the Movielens dataset from GroupLens Research Lab which includes movie ratings from 1902 to 2018. This dataset comprises 100,836 ratings and 3,683 tag applications across 9,742 movies, with each user rating at least 20 films. Although the full dataset includes 1.9 million ratings, we focused on a subset of approximately 100,000 ratings for our current model due to time and resource limitations. This sample size provides a balance between computational efficiency and statistical relevance for developing our recommendation engine. The datasets involved are links.csv, movies.csv, ratings.csv, and tags.csv. For our recommendation system, we utilized the features movieId, userId\_x, rating, title and genres. After merging the datasets on the movieId column, we obtained a DataFrame with 285,783 rows and 11 columns. While this data sufficed for our objectives, additional information such as actors, directors, production studios, runtime and user demographics could have offered more context and insights into user preferences and movie characteristics potentially enhancing the recommendations.

Data Preparation

To prepare the data for our analysis, we started by creating a class specifically designed to explore the datasets. We decided to merge all the datasets in order to get a holistic view of what we would be working with. The first step was to drop columns we deemed not relevant to our model curation process.

The second step was to check the integrity of the data and we found that there were 21 missing values in our merged dataset. We opted to drop the rows with the missing values as we would still have sufficient data for our analysis.

We then examined the properties of the DataFrame columns. Initially, we noticed that the user\_id column was of type float. Given that user IDs are inherently integer values, we converted this column from float to integer to ensure data consistency.

After these adjustments, we used the data\_explorer.merged\_data.info() function to review the structure of our DataFrame. It revealed that we had 100,818 entries and six columns: movieId, title, genres, user\_id, rating, and release\_year. Each column contained non-null values with user\_id having been successfully converted to an integer data type. With these preparations completed, we were ready to proceed to exploratory data analysis confident that our dataset was now well-structured and suitable for further investigation.

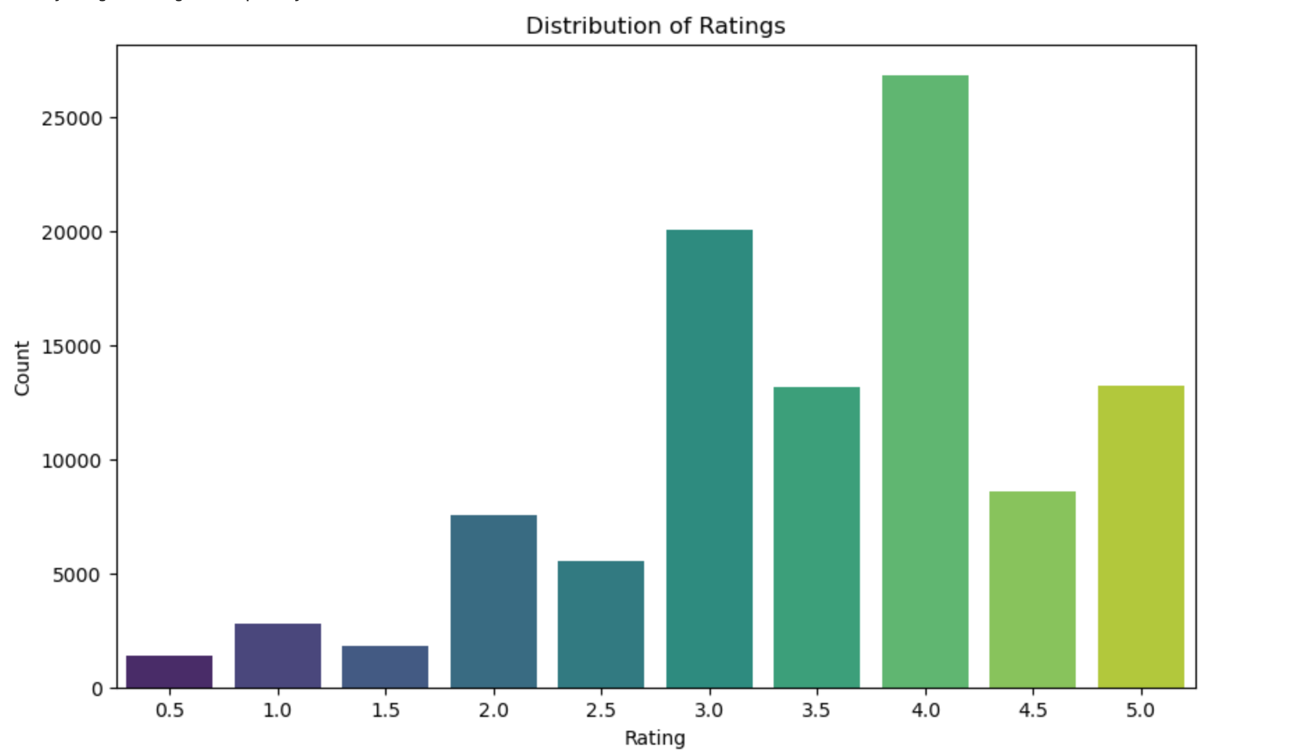
EDA

For our exploratory data analysis (EDA), we began by analyzing various aspects of the dataset to uncover patterns and insights.

Univariate Analysis

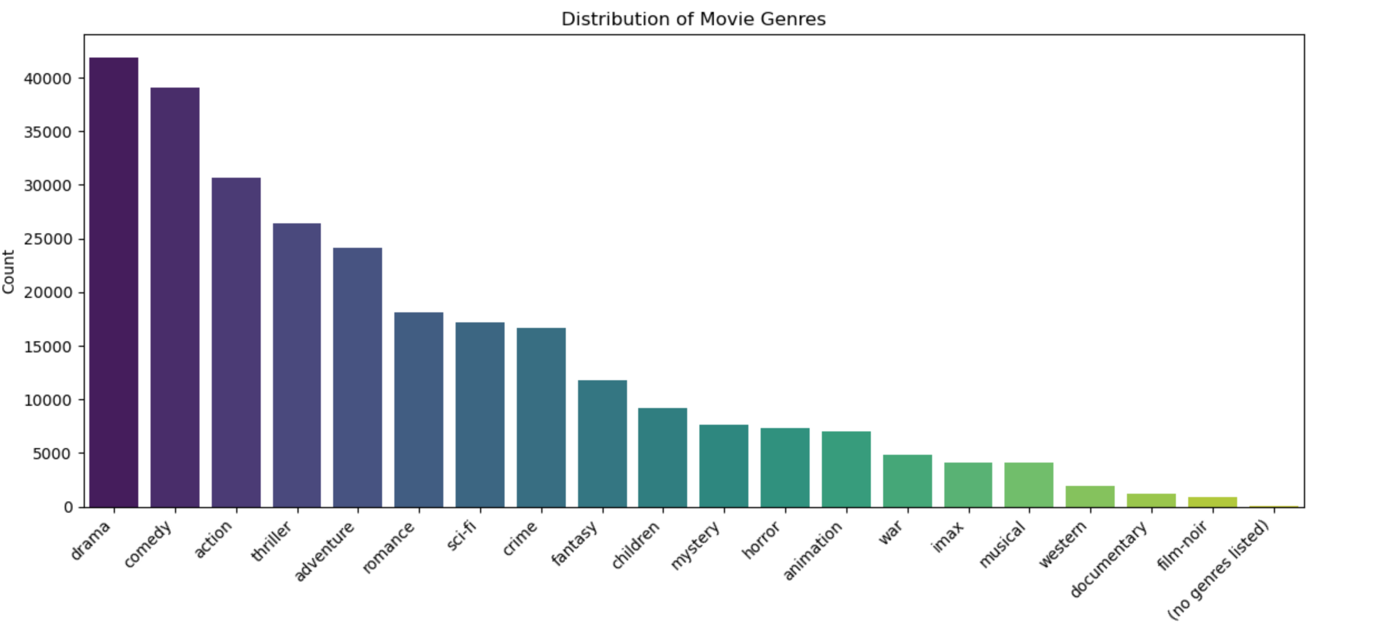
In our univariate analysis, we examined several key aspects of the dataset to gain deeper insights.

1. Distribution of ratings



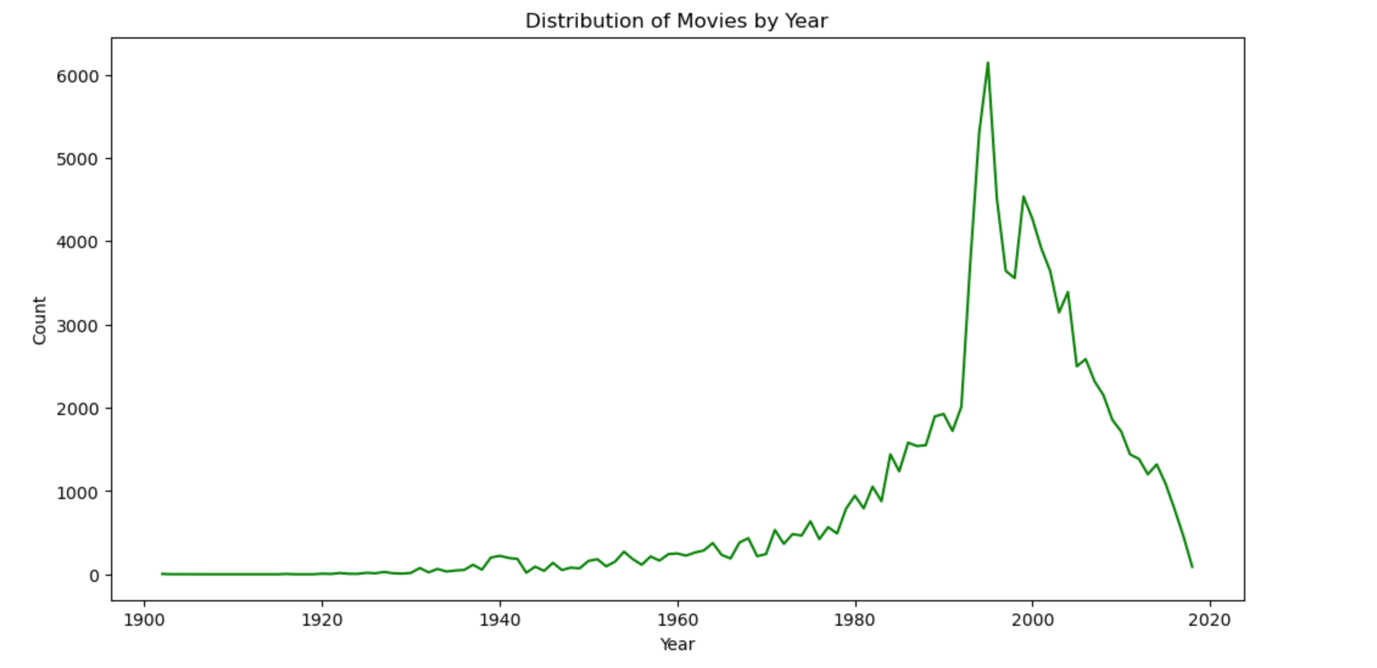
We began with the “distribution of ratings” plot which revealed that the most common rating is 4.0, with around 25,000 movies receiving this score. The distribution is positively skewed, indicating a tendency towards higher ratings. Significant counts are also observed at ratings of 3.0 and 3.5. Although ratings of 2.0, 4.5, and 5.0 are less frequent, they are still notable. Extremely low ratings of 0.5 and 1.0 are rare, suggesting that most movies are generally rated at least moderately well.

1. Distribution of genres



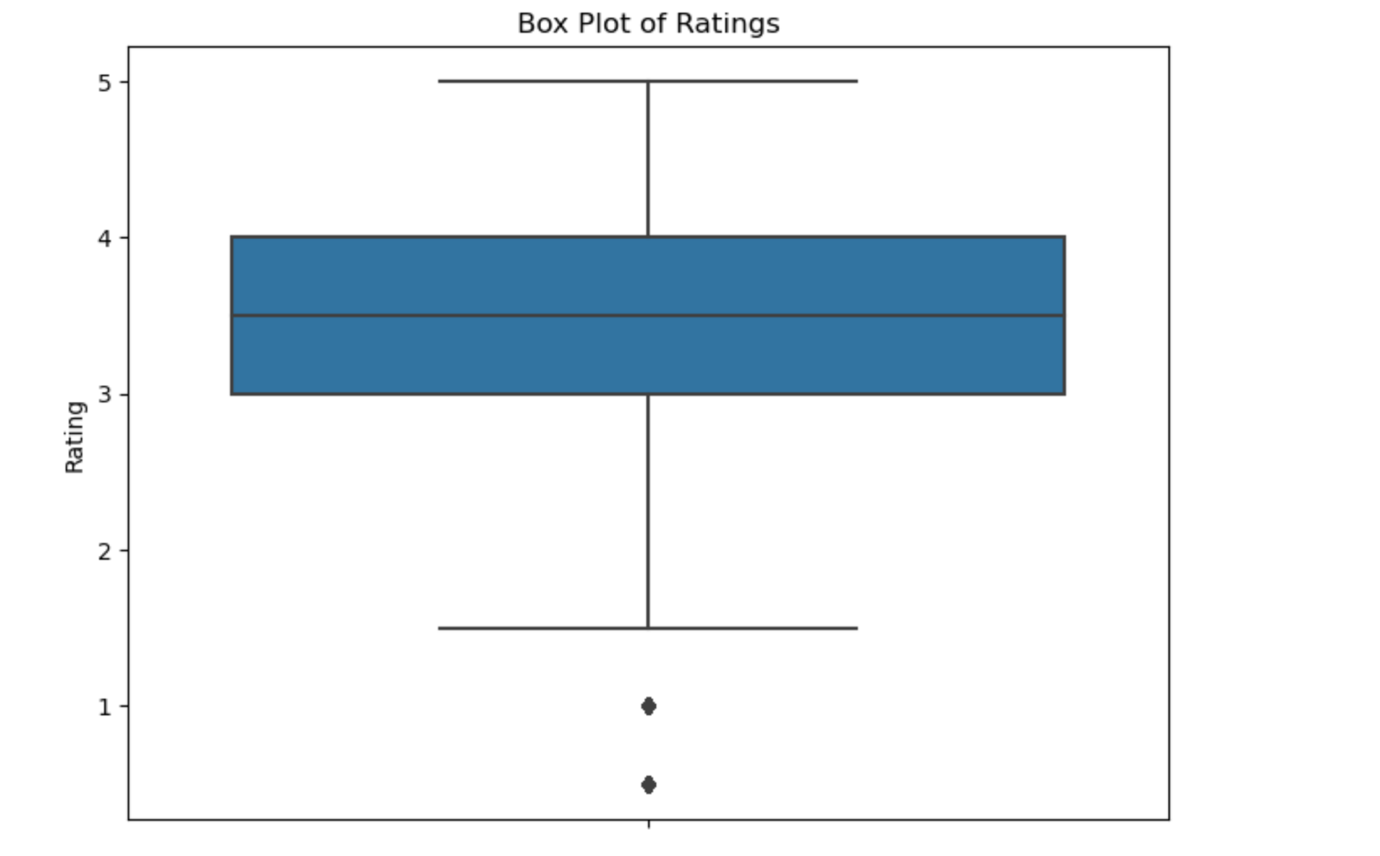
Next, we analyzed the “distribution of genres” plot. Drama emerged as the most prevalent genre, featuring slightly more than 40,000 movies, closely followed by Comedy with just under 40,000 films. Action and Thriller genres also have significant representation, with Action being more popular than Thriller. Genres like Adventure and Romance are moderately popular, each boasting over 20,000 movies, while Sci-Fi and Crime have slightly fewer titles. Less common genres include Fantasy, Children’s, Mystery, and Horror, each with moderate counts but significantly fewer than the top genres. Animation, War, IMAX, Musical, and Western genres have even lower counts, with fewer than 10,000 movies each, indicating they cater to more niche audiences. Documentary and Film-Noir are among the least common genres, with fewer than 5,000 movies each. Additionally, there is a small category for movies without any listed genres.

1. Distribution of movies by year



We then looked at the “distribution of movies by year”, which highlights historical trends in movie production. We observed significant growth in the movie industry starting in the late 1960s, peaking in the late 1990s with a maximum output of over 6,000 movies. This was followed by a sharp decline in production, suggesting potential shifts in industry dynamics such as changes in movie consumption patterns, technological advancements, or economic factors impacting production.

1. Analysis of rating outliers

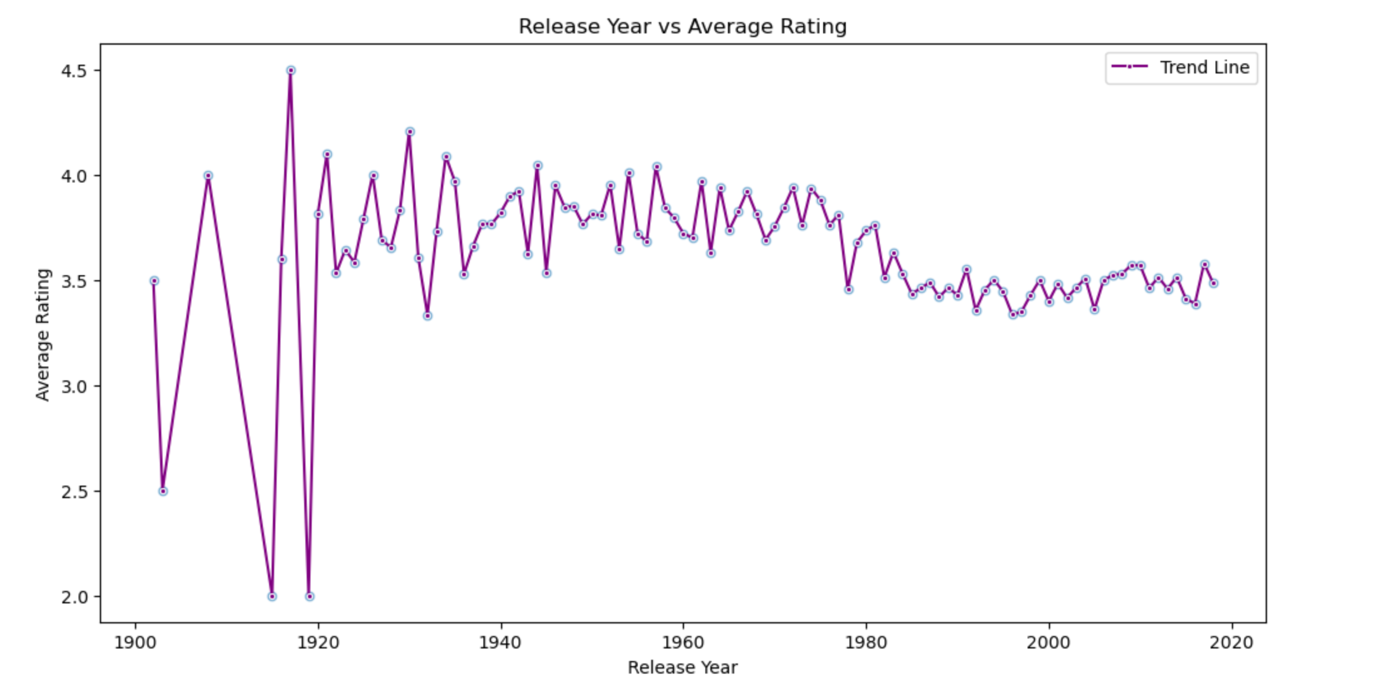


Finally, the “analysis of rating outliers” plot identified some unusual ratings. Despite their presence, we decided to keep these outliers in the dataset as they may be valuable for curating our recommender system.

This univariate analysis provided a detailed understanding of rating patterns, genre distributions, production trends and outliers, laying the groundwork for further analysis and model development.

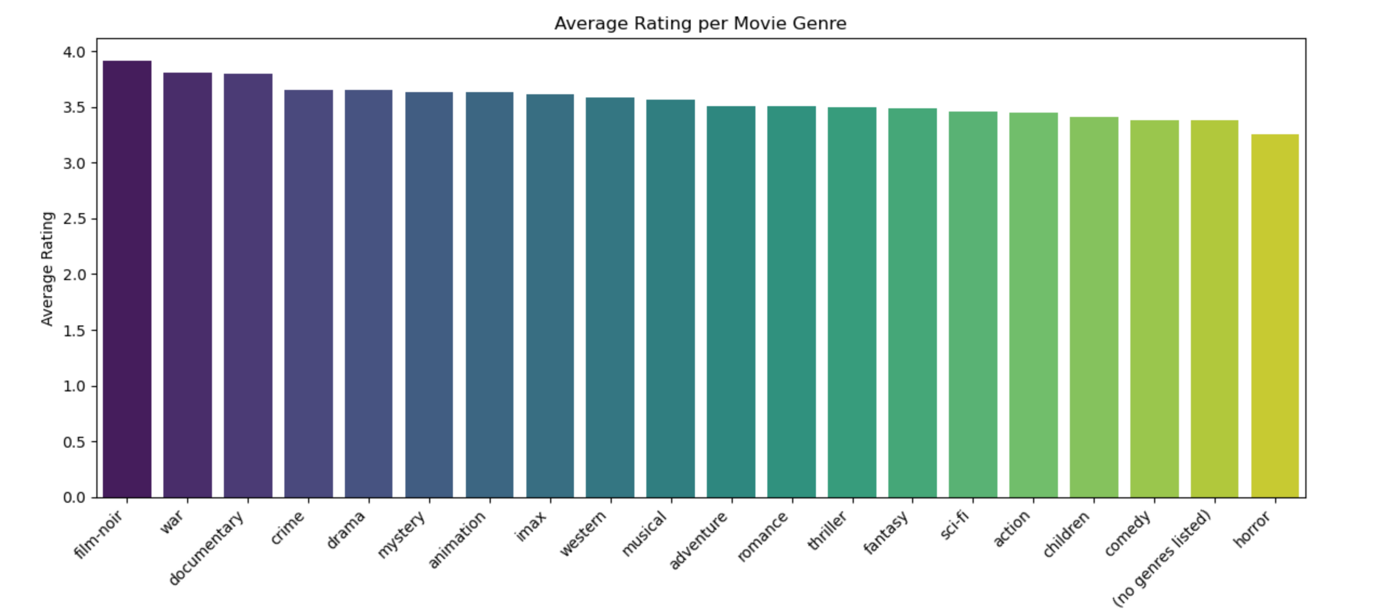
Bivariate Analysis

1. Release Year vs Average Rating



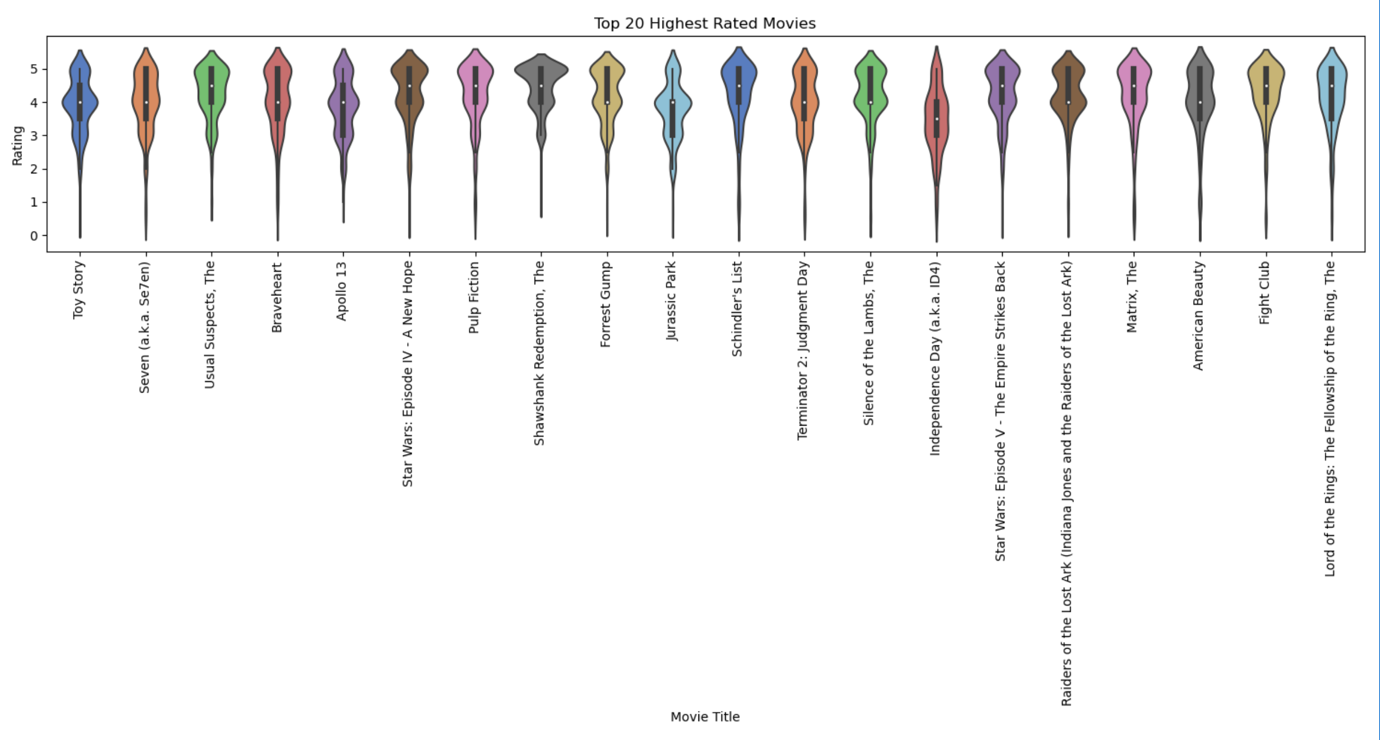
Firstly, we examined the "Release Year vs Average Rating" plot, which illustrated the fluctuations in average ratings over time. In the early years, from 1900 to 1940, the ratings showed considerable variation, likely reflecting the film industry's formative period. This was followed by notable peaks in the early 1900s, a stabilization phase post-1940, and a gradual decline from the 1970s to 2020. From the 1980s onwards, ratings consistently hovered around 3.5, with outliers becoming less pronounced over time.

1. Average Rating Per Movie Genre



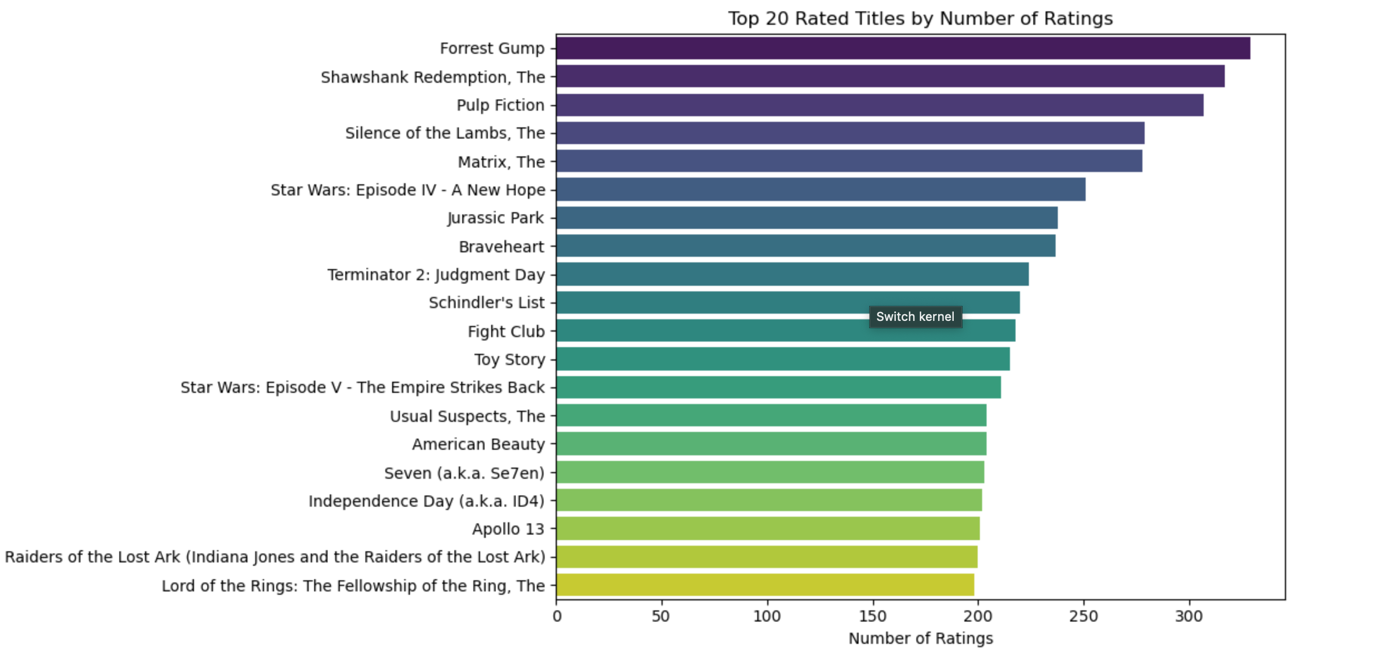
Next, we analyzed the "Average Rating Per Movie Genre" plot, which provided insights into how different genres are rated on average. Genres such as film-noir and war received the highest ratings, approaching 4.0. Documentary, crime, and drama genres also scored highly, around 3.75. Other genres like mystery, animation, IMAX, western, and musical followed closely. Adventure, romance, thriller, and fantasy genres had slightly lower average ratings, around 3.5, while sci-fi, action, children’s, and comedy genres averaged closer to 3.0. The genres with the lowest average ratings were unspecified genres and horror movies, both falling below 3.0. Overall, film-noir and war emerged as the most highly rated genres, while horror films had the lowest average ratings.

1. Top 20 Highest Rated Movies



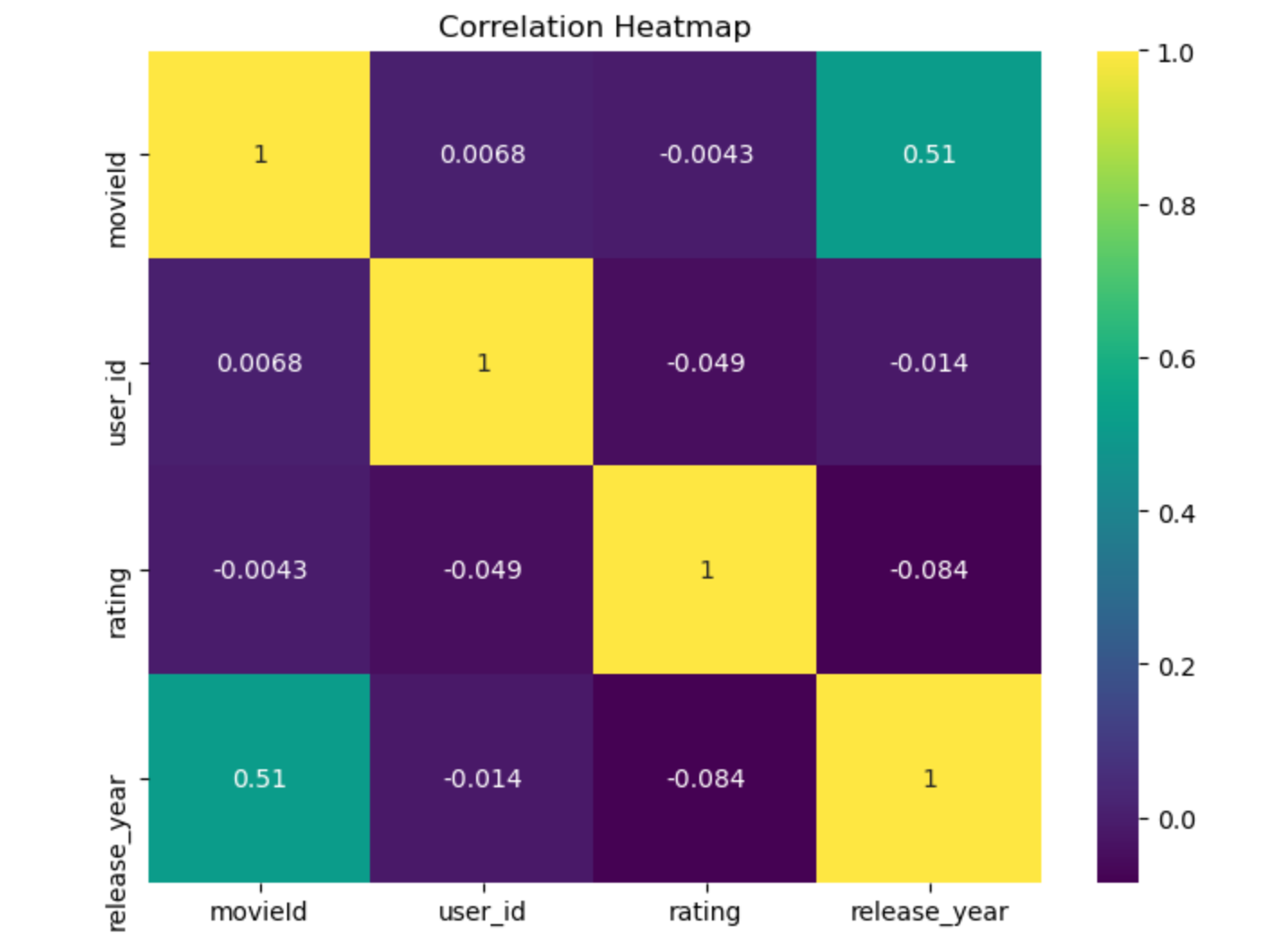
We then reviewed the "Top 20 Highest Rated Movies" plot, which showcased the rating distributions for the top 20 films. Each movie was represented by a vertical violin shape, with the width indicating the frequency of ratings at each score level. Most top-rated movies displayed a concentration of high ratings, with a bulge near the top of the 5-point scale, though there was some variation in rating patterns across different films.

1. Top 20 Rated Titles by Number of Ratings



Lastly, the "Top 20 Rated Titles by Number of Ratings" plot revealed that 'Forrest Gump' led with the highest number of ratings, followed closely by 'The Shawshank Redemption' and 'Pulp Fiction'. The number of ratings ranged from around 200 for 'Lord of the Rings: The Fellowship of the Ring' to nearly 300 for the top-rated films, indicating their popularity and wide review base.

1. Correlation Heatmap



To conclude our analysis, we examined the "Correlation Heatmap," which illustrated the relationships between numeric variables. It revealed a strong positive correlation (0.98) between release year and decade, moderate positive correlations between movieId and both release year (0.51) and decade (0.5), with weak or negligible correlations among other variables. The diagonal values of 1 in the heatmap represented perfect self-correlation.

This comprehensive EDA provided valuable insights into trends and relationships within the dataset, guiding further analysis and model development.