

**Anurag University**

**School of Management**

**MBA-BA**

**BI BATTLE**

**2025**

**Team-10**

**23MG202A19 – K. Ajitha Reddy**

**Domain:** Entertainment and streaming

The streaming industry relies on recommendation engines to enhance user experience. Platforms like Netflix and Amazon Prime use data-driven analytics and machine learning to optimize content discovery, maximize watch time, and improve customer retention.

**Problem statement:** Analyse user watch history to recommend personalized movie and TV show suggestions.

With vast content choices, users struggle to find relevant movies and shows. Traditional recommendation systems often fail to adapt to changing preferences. This project builds an adaptive recommendation system that leverages user history, content metadata, and trends to provide personalized suggestions.

**Data sources:** Kaggle

**Datasets:** Movies and TV Shows

The dataset contains **10,865** records with **30 columns**, covering metadata on movies and TV shows, user watch history, ratings, and recommendation-related attributes. It provides insights into user behaviour, preferences, and engagement on streaming platforms.

**Column Descriptions**

1. **Id** – Unique identifier for each movie or TV show.
2. **Title** – Name of the movie or TV show.
3. **Type** – Specifies whether the content is a "MOVIE" or "SHOW".
4. **Description** – Short synopsis of the movie or TV show.
5. **Release year** – Year the content was released.
6. **Age\_certification** – Age rating (e.g., PG, TV-PG, etc.).
7. **Duration** – Duration of the content (in minutes for movies, episodes for shows).
8. **Genres** – List of genres the content belongs to (e.g., Drama, Comedy, Action).
9. **Production\_countries** – Countries where the content was produced.
10. **Seasons** – Number of seasons (only for TV shows).
11. **Imdb\_id** – Unique IMDb identifier for the content.
12. **Rating** – IMDb rating of the movie or show.
13. **Imdb\_votes** – Number of votes received on IMDb.
14. **Tmdb\_popularity** – Popularity score from The Movie Database (TMDb).
15. **Tmdb\_score** – User rating score from TMDb.
16. **User\_ID** – Unique identifier for each user.
17. **Watch\_History** – List of previously watched content by the user.
18. **Watch\_Duration** – Total time spent watching content.
19. **User\_Rating** – User-provided rating for watched content.
20. **Search\_Click\_History** – Records of movies/shows clicked after searching.
21. **Skipped\_Content** – List of movies/shows the user started but did not finish.
22. **Director** – Director(s) of the content.
23. **Cast** – Main actors starring in the content.
24. **Keywords\_Tags** – Keywords associated with the content.
25. **Streaming\_Availability** – Platform where the content is available (e.g., Amazon Prime, Netflix).
26. **Language** – Primary language of the content.
27. **Trending\_Score** – Popularity score based on real-time trends.
28. **Seasonal\_Popularity** – Popularity classification (Low, Medium, High) based on seasonal trends.
29. **User\_Age** – Age of the user.
30. **Friends\_Watch\_History** – Content watched by the user's friends, useful for social recommendations.

**Objectives:**

 Build a personalized recommendation system for movies and TV shows based on user watch history.

 Utilize machine learning models to analyse viewing behaviour and predict relevant content.

 Implement a hybrid recommendation approach combining collaborative and content-based filtering.

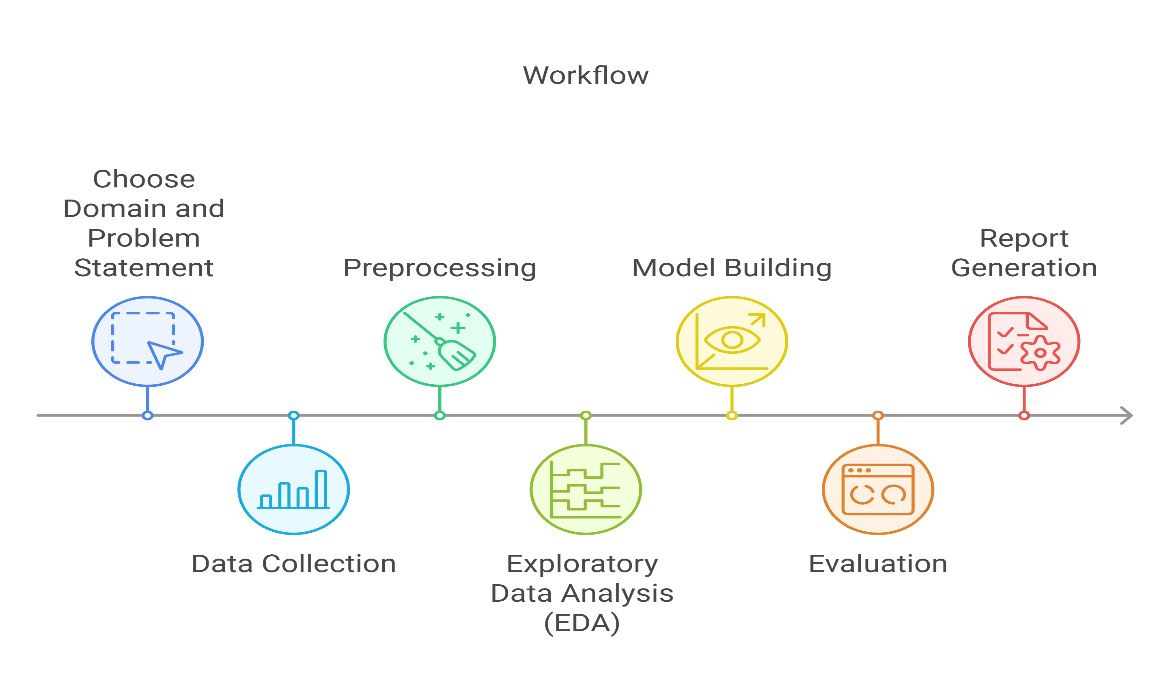
 Leverage multiple OOT datasets to capture evolving user preferences.

 Improve engagement and watch time through optimized recommendations.

 Analyse content metadata (genres, ratings, trending score) to refine suggestions.

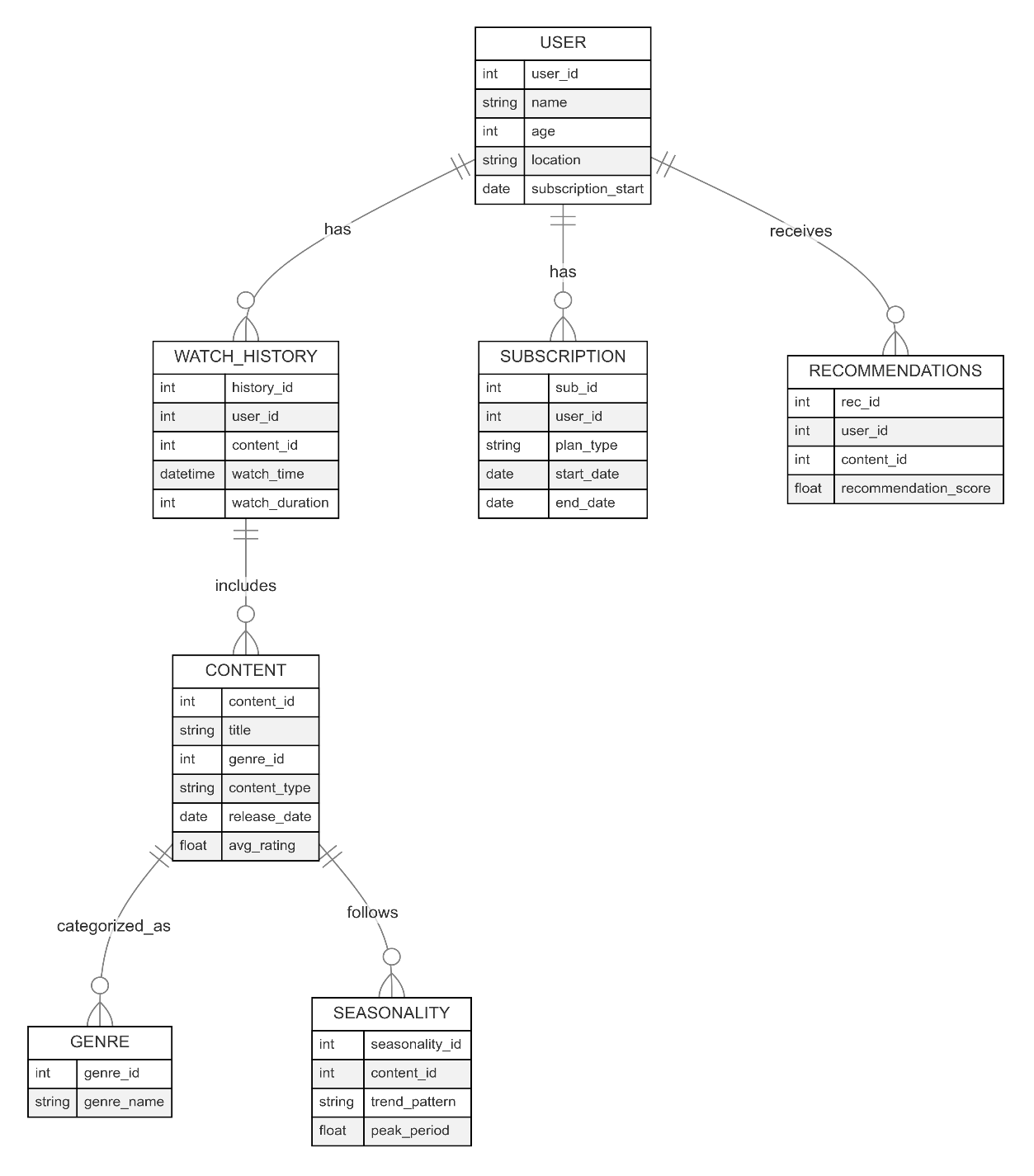
 Use Power BI and Python visualizations to interpret trends and model performance.

**Workflow:**

****

1. **Choose the domain and problem statement**: Define the project scope.
2. **Data Collection**: Gathering watch history data from multiple OOT datasets.
3. **Preprocessing**: Cleaning and structuring data for analysis.
4. **Exploratory Data Analysis (EDA)**: Identifying user behaviour trends, genre preferences, and peak watch times.
5. **Feature Engineering**: Creating user and content embeddings, session-based interactions, and watch-time patterns.
6. **Model Building**: Implementing various machine learning models to develop recommendations.
7. **Evaluation**: Assessing models based on KPIs and fine-tuning hyperparameters for optimal performance.
8. **Report Generation**: Summarizing findings and model performance for insights.

**Data model**:



**Techniques:**

* **Data Cleaning**: Handling missing values, duplicates, and inconsistencies.
* **Data Preprocessing**: Standardizing timestamps, normalizing numerical values, encoding categorical variables.
* **EDA (Exploratory Data Analysis)**: Visualizing trends, distributions, and relationships in the dataset.
* **Feature Engineering**: Extracting and transforming variables for better model performance.
* **Machine Learning Model Building**: Training, validating, and optimizing predictive models.
* **Dashboard Generation**: Using Power BI for interactive insights.

**Tools:**

|  |  |
| --- | --- |
| **Tools** | **Purpose** |
| Excel | Data cleaning |
| Power BI | Data Analysis |
| Jupyter notebook - Python libraries | Data cleaning, Data Analysis, Machine learning |
| Microsoft Word | Documenting |

**Visualizations with Interpretations and insights:**

The following are visualizations with interpretations and insights generated using Jupyter notebook- Python and Power BI.

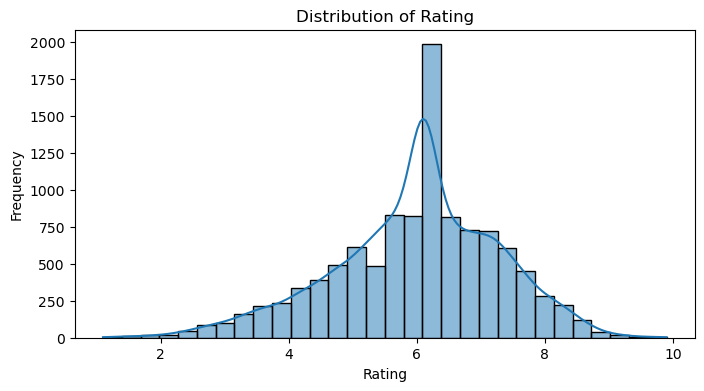
Various types of visualizations like Bar chart, Histogram, Heatmap, Pie chart, Line chart are used.

**Python Visualizations:**

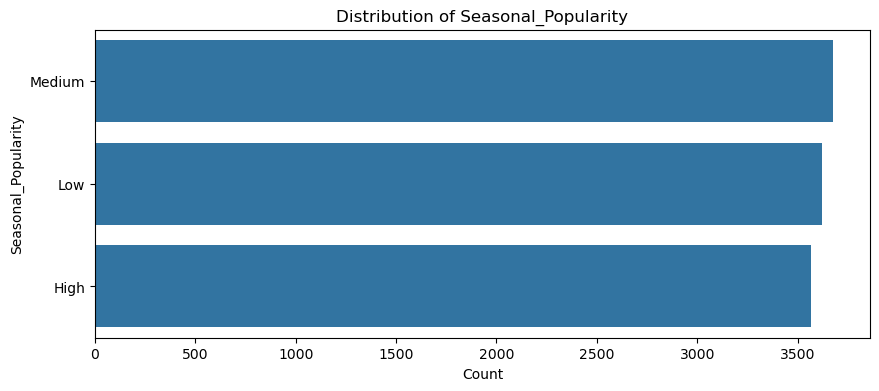
The following are the visualizations and insights generated using various python libraries. The python libraries used are Numpy, Pandas, Seaborn, Matplotlib, Sklearn, Xgboost, Scipy.

**EDA (Exploratory Data Analysis):**

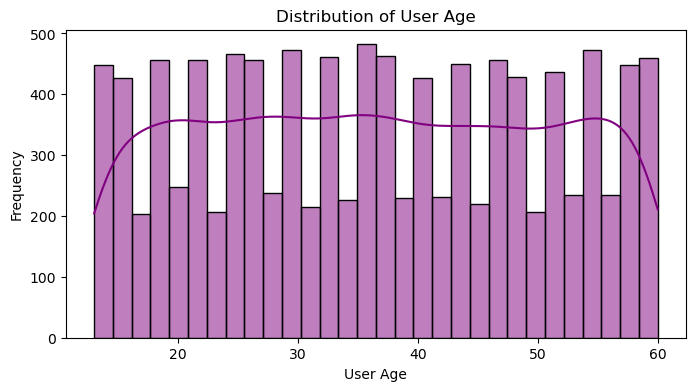
The following are EDA visualizations:



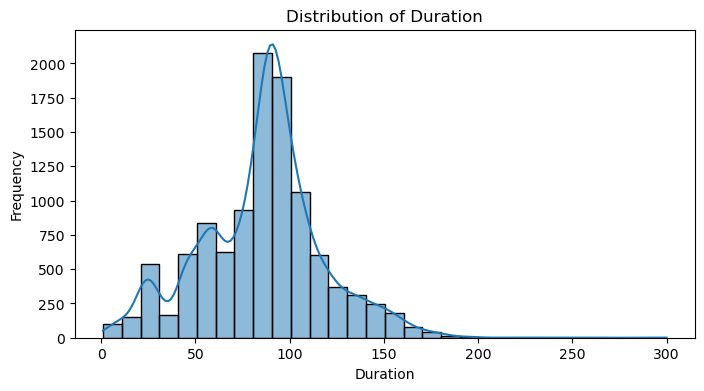
The histogram shows the distribution of user ratings, with most ratings clustering around 6, indicating a central tendency near this value. The presence of a density plot highlights the smooth probability distribution, reinforcing the near-normal shape with a peak at 6. A sharp spike at 6 suggests a possible rounding bias or preference in ratings.



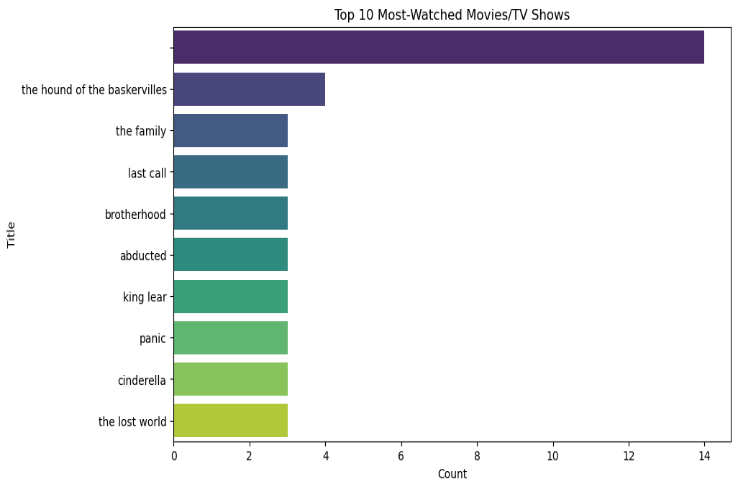
The above bar chart displaying the count of content categorized by seasonal popularity levels (Low, Medium, High). This visualization provides insights into how content popularity varies with seasons.



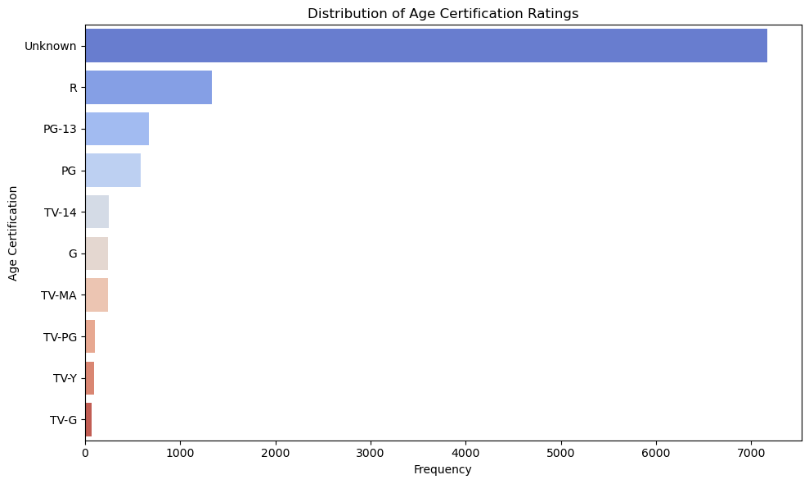
The histogram shows a nearly uniform distribution of user ages, indicating a balanced representation across different age groups. The density plot suggests slight variations, but no strong skewness or concentration in any specific age range.



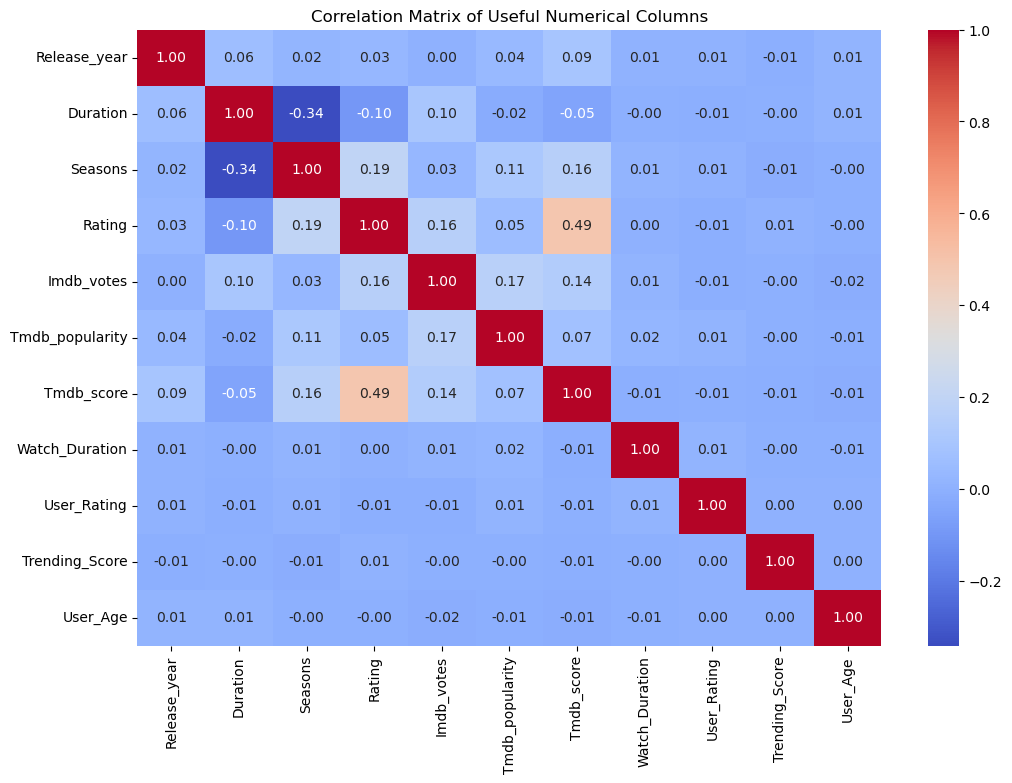
The above is a histogram showing the distribution of watch durations (in minutes). It uses kernel density estimation (KDE) to provide a smooth curve that estimates the probability density function of the watch duration data. This graph helps understand how long users typically watch content in one sitting.



The above bar graph displays the most frequently occurring titles in the dataset, with The Hound of the Baskervilles appearing the most. The presence of multiple adaptations or duplicates suggests potential data cleaning or content consolidation is needed. This insight can refine recommendations by avoiding redundant suggestions.



The majority of content lacks age classification ("Unknown"), reducing the effectiveness of age-based recommendations. R-rated content dominates, while family-friendly categories (TV-Y, TV-G) have limited representation, requiring better data for improved filtering.



The above correlation matrix depicts the relationships between different numerical variables in dataset.

 **Watch Duration - High Ratings** – Users may watch content fully but still rate it poorly. Prioritize ratings over duration.

 **TMDB Score Aligns with Ratings (0.49)** – Higher-rated content generally has better TMDB scores, making it a reliable quality metric.

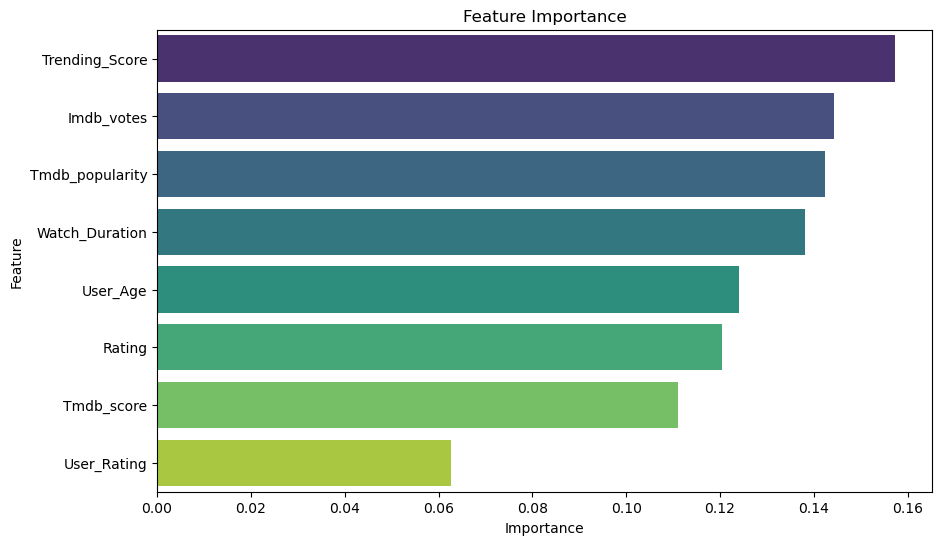
 **Longer Duration - Fewer Seasons (-0.34)** – multi-season shows have shorter episodes, while movies tend to be longer. Match based on preference.

 **Trending Score is Weakly Correlated (~0)** – Trending content doesn’t always align with user taste; blend it with personalized recommendations.

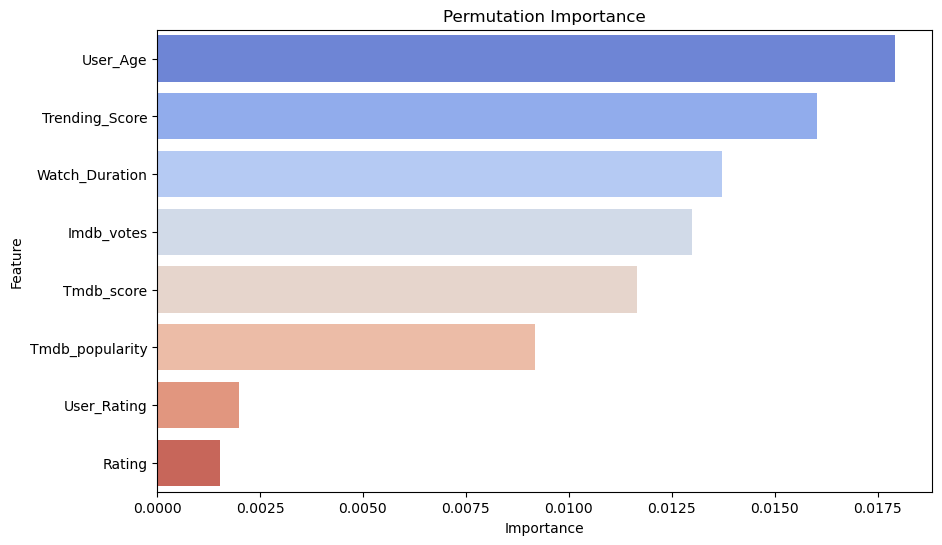
 **IMDb Votes & TMDB Popularity (0.17)** – Content can be popular on one platform but not another. Use multiple sources for ranking.

**Machine learning models:**

Feature importance analysis is done in order to fine tune models for better performance.



Trending Score is the strongest factor, followed by IMDb Votes and TMDB Popularity, highlighting user engagement. Watch Duration and User Age shape personalized recommendations, while Ratings and User Ratings have lower influence.



The graph shows Permutation Importance, ranking features by their impact on model predictions. User\_Age, Trending\_Score, and Watch\_Duration are the most influential, while User\_Rating and Rating contribute the least. This suggests engagement and demographics play a bigger role than individual ratings in recommendations.

Performance of ML models:  
**Classification models:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1-Score |
| **Random Forest Classifier** | **36.22%** | 0.36 | 0.36 | 0.36 |
| **Support Vector Machine (SVM)** | 33.00% | 0.33 | 0.33 | 0.32 |
| **K-Nearest Neighbors (KNN)** | 32.21% | 0.32 | 0.32 | 0.32 |
| **Logistic Regression** | 32.72% | 0.33 | 0.33 | 0.31 |
| **XGBoost** | 33.96% | 0.34 | 0.34 | 0.34 |

Best Performing Model: Random Forest

Best parameters: max\_depth=20, min\_samples\_split=10, n\_estimators=200

**Hybrid Recommendation System:**

How It Works

1. Collaborative Filtering (KNN Model) – Finds similar users based on watch history and ratings, recommending items they watched.
2. Content-Based Filtering (TF-IDF & Cosine Similarity) – Suggests items similar to those the user interacted with based on genres, keywords, and streaming availability.
3. Hybrid Approach – Merges both methods, ranks recommendations by similarity scores, and filters out already-watched content.

Effectiveness

* Personalized & Diverse – Balances user behaviour and content features.
* Handles Cold Start – Recommends items even for new users.
* Improved Accuracy – Reduces biases and enhances relevance.

Using the Hybrid approach-

Top 10 Recommendations for User U1433:

1. Across the Line

2. Red Ribbon Blues

3. No Secrets

4. Wish You Were Here

5. Poola Rangadu

6. Kill Dil

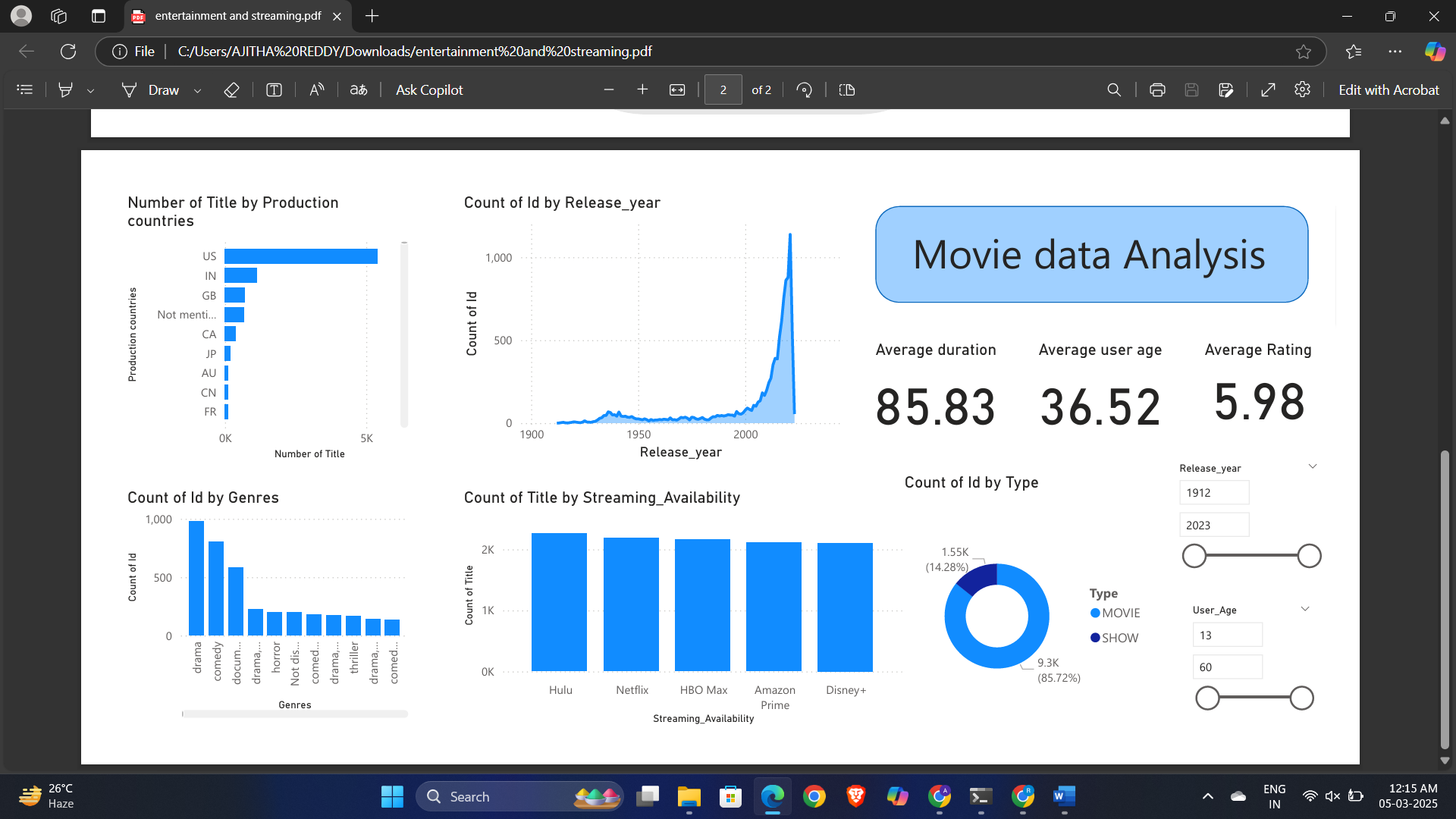
7. Darling

8. Inherit the Viper

9. The 355

10. The Voyeurs

**Power Bi Visualizations:**



This Power BI dashboard provides an analysis of movie and TV show data across different streaming platforms. It includes key metrics and visualizations for understanding trends in genres, release years, user demographics, streaming availability, and content types.

1️. Key Metrics (Top Right)

* Average Duration: 85.83 minutes – The typical duration of movies/shows in the dataset.
* Average User Age: 36.52 years – The average age of users interacting with this content.
* Average Rating: 5.98 – The overall average user rating of movies and shows.

2️. Number of Titles by Production Countries (Top Left)

* Shows which countries produce the most content.
* US has the highest number of movies/shows, followed by India (IN) and Great Britain (GB).

3️. Count of ID by Release Year (Top Middle)

* X-Axis: Release Year (1912–2023)
* Y-Axis: Number of Titles Released
* The graph shows a sharp increase in content production after 2000, peaking in recent years.

4️. Count of Titles by Streaming Availability (Bottom Middle)

* Compares the number of titles available on different streaming platforms:
  + Hulu, Netflix, HBO Max, Amazon Prime, Disney+

5️. Count of ID by Genres (Bottom Left)

* Top Genres:
  + Drama and Comedy dominate the dataset.
  + Documentary, Horror, Thriller, and Action follow.
* This helps in understanding what type of content is most popular.

6️. Count of ID by Type (Bottom Right - Pie Chart)

* Two categories:
  + Movies (85.72%) – Majority of content.
  + Shows (14.28%) – A smaller portion.
* This indicates that most content in the dataset consists of movies rather than TV shows.

7️. Filters (Right Side)

* Release Year Range (1912–2023) – Users can filter content based on the year of release.
* User Age Range (13–60) – Allows filtering based on user age preferences.
* Type Selector (Movie/Show) – Enables switching between movies and TV shows for better insights.

Insights & Use Cases

* Helps in content recommendation based on streaming availability, genres, and user preferences.
* Useful for OTT platforms to identify popular genres, production trends, and audience behaviour.
* Can aid in strategic decision-making for acquiring or promoting content.

**Conclusion:**

This recommendation system enhances content discovery by analysing user behaviour and trends. By integrating multiple OTT datasets and machine learning models, it improves engagement, increases watch time, and personalizes recommendations, benefiting both users and streaming platforms.

**Future Enhancements:**

* Cross-Platform Integration: Expanding recommendations across services.
* Sentiment-Based Analysis: Using user reviews for refined suggestions.
* Context-Aware Recommendations: Incorporating time, mood, and demographics.
* Deep Learning Models: Enhancing predictions using advanced neural networks.
* Privacy-Preserving Methods: Federated learning to ensure data security.