[ReelGood]

[G4]

Data Science Capstone Project
Exploratory Data Analytics Report

Date:

[03/08/2025]

Team Members:

Name: Alireza Hatami

Name: Caitlin Dunne

Name: Jaz Zhou

Name: Precious Orekha

Analysis the basic metrics of variables

The analysis of the basic metrics of variables is provided for both the Movies and Ratings dataframes (Table 1 and 2 in the Appendix):

A) Movies Dataframe:

- Contains 45,020 rows and multiple numerical and categorical variables.
- Numerical variables include budget, revenue, popularity, vote_average, and vote_count.
- Categorical variables include genres, original_language, first_three_actors, directors, production_countries, and title.
- Budget and revenue are continuous variables, while vote_count is discrete.

Some numerical variables, such as production_countries, budget, revenue, and runtime, were only considered for exploratory data analysis (EDA) purposes and will not be included in the construction of recommender systems. However, some of these features may provide interesting insights. A summary of the descriptive statistics for these numerical features is provided below (The statistical table is located in the Appendix, Table 1):

- Budget Statistics

- Mean (Average): 4.26 million
- Standard Deviation (std): 17.50 million → There is a Significant variation in budget values.
- Minimum Budget: 0 (Missing or unknown budgets).
- Maximum Budget: 380 million (Blockbuster movies).
- 25th, 50th (Median), and 75th Percentiles: All $0 → Indicates many missing budget values.

- Revenue Statistics

- Mean Revenue: 11.31 million
- Standard Deviation: 64.63 million → Extreme differences in revenue between movies.
- Minimum Revenue: 0 (Some movies didn't report revenue or made nothing).
- Maximum Revenue: 2.79 billion (Likely a major blockbuster).

- Runtime Statistics

- Mean Runtime: 94.5 minutes
- Minimum Runtime: 0 minutes (Perhaps we have some missing or incorrect values).
- Maximum Runtime: 1,256 minutes (Maybe it's a data error or extended documentary).

- Year Statistics

- The dataset includes movies released between 1874 and 2020.
- Mean Year: The average release year is 1991.
- 50th Percentile (Median): the median release year is 2001 → At least half of the movies were produced after the early 2000s
- Most Movies Produced Around: 2010 (75th percentile).

B) Ratings Dataframe:

- The dataset contains 10,212,750 rows, which consist of users' ratings for various movies.
- Includes numerical attributes such as rating, which is continuous, and userId and movieId, which serve as unique identifiers.
- The timestamp is recorded as an integer but can be converted into categorical features such as year and month.

The statistics table is provided in the Appendix, Table 2. Below is a summary of the ratings and timestamp statistics:

- Rating Statistics:

- The dataset contains over 10 million individual movie ratings.
- Mean (Average Rating): The average movie rating given by users is 3.52, indicating that most movies receive mid-range scores.

- Standard Deviation (std): 1.06 → Ratings show a significant variation, meaning users have diverse opinions on movies.
- Minimum Rating: The lowest rating given is 0.5
- 25th Percentile (Q1): 3.00 → 25% of all ratings are below 3.00, showing that a significant portion of movies receive below-average ratings.
- 50th Percentile (Median, Q2): 3.50 → The median rating is 3.50, meaning half of all ratings are below 3.50, and half are above.
- 75th Percentile (Q3): 4.00 → 75% of movies have a rating below 4.00, indicating that highly-rated movies are relatively fewer.
- Maximum Rating: 5.00 → The highest possible rating is 5.00, meaning some movies are rated perfectly by certain users.

- Timestamp Statistics:

- Mean: The average timestamp corresponds to February 4, 2005. This suggests that, on average, most ratings in the dataset were given around the mid-2000s.
- 25th Percentile (Q1): The first quartile corresponds to February 12, 2000, indicating that 25% of the ratings were given before this date.
- 50th Percentile (Median): The median timestamp converts to December 1, 2004, meaning that half of the ratings were recorded before this date, and the other half were recorded after.
- 75th Percentile (Q3): The third quartile maps to March 19, 2010, meaning that 75% of the ratings were recorded before this date.
- Minimum Timestamp: The earliest recorded rating is from January 9, 1995, marking the beginning of the dataset's time range.
- Maximum Timestamp: The most recent rating in the dataset was recorded on August 4, 2017, suggesting the dataset spans over 22 years of movie ratings.
- Standard Deviation: The standard deviation of timestamps is quite large, approximately 6.4 years, showing that ratings are spread out over a long period rather than concentrated in a specific timeframe.

These findings indicate that the dataset captures a broad time range of ratings, spanning from the mid-90s to late 2017. The concentration of ratings around the mid-2000s suggests that this period might have seen a significant surge in user activity, likely due to the rise of online movie rating platforms.

Non-graphical and graphical univariate analysis

A) Movies Dataframe:

1. Categorical Variable:

In the movies dataset, the categorical variables include genres, original_language, first_three_actors, directors, production_countries, and title. Although production_countries is not included in the final cleaned data, it is added solely for visualization purposes in the EDA phase.

- Genre:

The unique genres are : War, Romance, Documentary, Drama, Music, Crime, Action, History, TV Movie, Short, News, Musical, Biography, Horror, Foreign, Fantasy, Animation, Adult, Sport, Science Fiction, Western, Comedy, Film-Noir, Family, Mystery, Thriller, Sci-Fi, and Adventure.

The dataset reveals that Drama (21,373 movies) and Comedy (13,790 movies) are the most prevalent genres. Thriller, Romance, and Action follow closely, reflecting production interest in suspenseful, emotional, and high-energy films. Niche genres such as Film-Noir (6 movies), News (7 movies), and Sport (32 movies) are the least represented in our dataset.

- **User Ratings and Genre:** The number of ratings users gave for each genre indicates that genres like Drama, Comedy, Thriller, and Action movies are more popular than others.

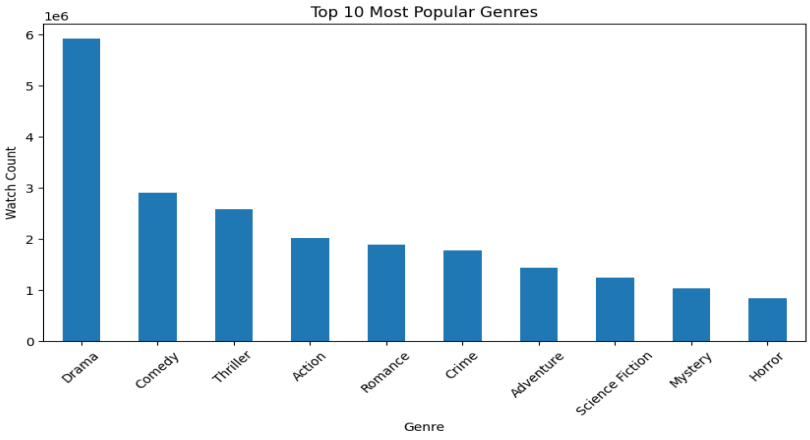


Figure 1

- Actors:

The number of actors extracted from our data set is 48269. By extracting the number of movies that actors appear in, we can rank the actors.

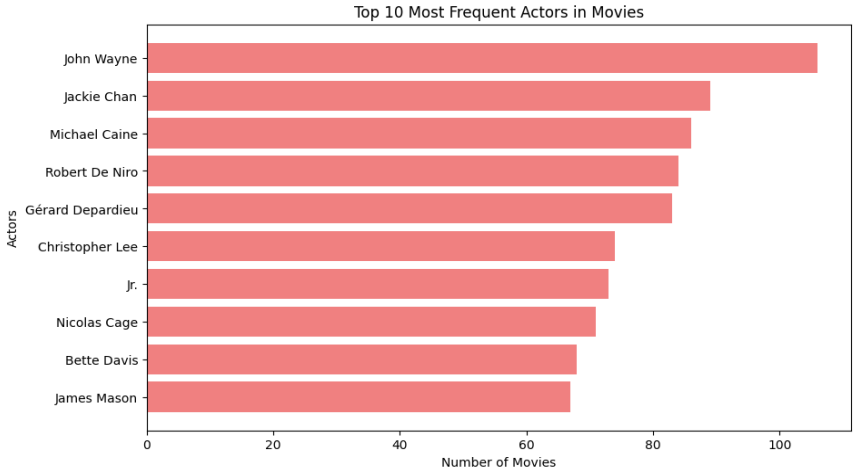


Figure 2

We were able to extract the number of genres in which each actor appears. This information helps us identify the actors who are primarily cast in specific genres. For example, in action movies, Jackie Chan and John Wayne hold the first and second positions, respectively (The bar graph for ranking the top 10 action movie actors is available in the Appendix, Figure 17).

- Language:
Our dataset includes 89 languages for movies. Due to the large number of unique language values, we counted the frequency of each language and visualized the top 10 below. Most movies in the dataset are in English, with French and Italian films following in second and third place, respectively.

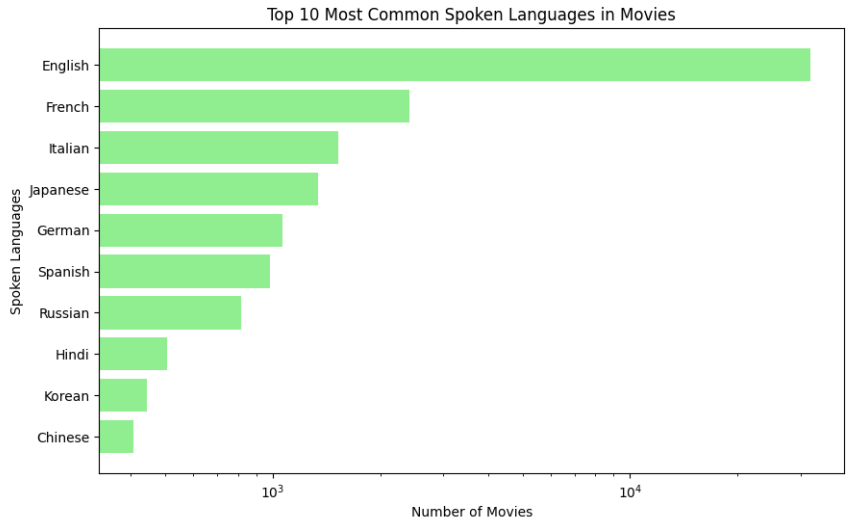


Figure 3

- **Ratings and Language:** The average ratings received by users for each spoken language in movies indicate that although English-language films have the highest frequency, they tend to receive lower average ratings compared to films in other languages. One possible explanation for this phenomenon is that the audience for English movies is much larger than the audience for films in other languages; therefore, it is natural for English movies to have lower average ratings (The visualization can be found in the Appendix, Figure 18).

- Production Countries:
Our dataset includes a total of 161 unique production countries. The United States of America has the highest number of movies, totaling 21061. However, there is a substantial group of movies (6,077) for which no production countries were recorded. Additionally, the United Kingdom ranks second, with 4,067 movies in our dataset. (A list of the ten most common production countries can be found in Appendix, Figure 19).

- Directors:

The total number of unique directors in our dataset is 17884. Additionally, we count how many times each director appears in the dataset, and we use this information to extract the most frequent directors. This will help us find the filmmakers who have created the most films in the dataset. To improve visualization, we extracted the top 10 directors who made the most films due to the large number of directors.

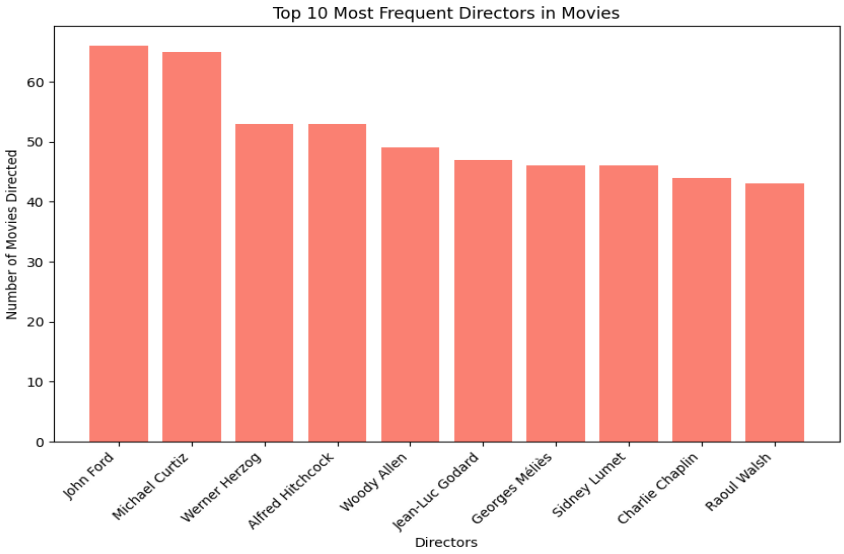


Figure 4

- Title:

To visualize the words that stand out in movie titles, we used a Word Cloud. The most frequent words include Love, Day, Girl, Man, Life, and Night.

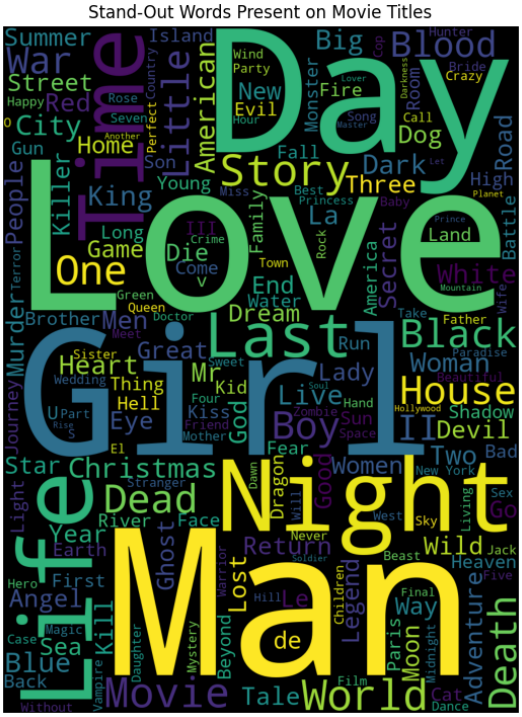


Figure 5

2. Numerical Variables:

The numerical variables in the movies dataset are Year, Budget, Popularity, Revenue, and Runtime. While we will use the Year variable for model development, the other numerical variables—Budget, Popularity, Revenue, and Runtime—will only be considered for exploratory data analysis (EDA) purposes.

- Year:

The dataset includes movies released between 1874 and 2020, with an average release year of 1991. The median release year is 2001, indicating that at least half of the movies were produced after the early 2000s.

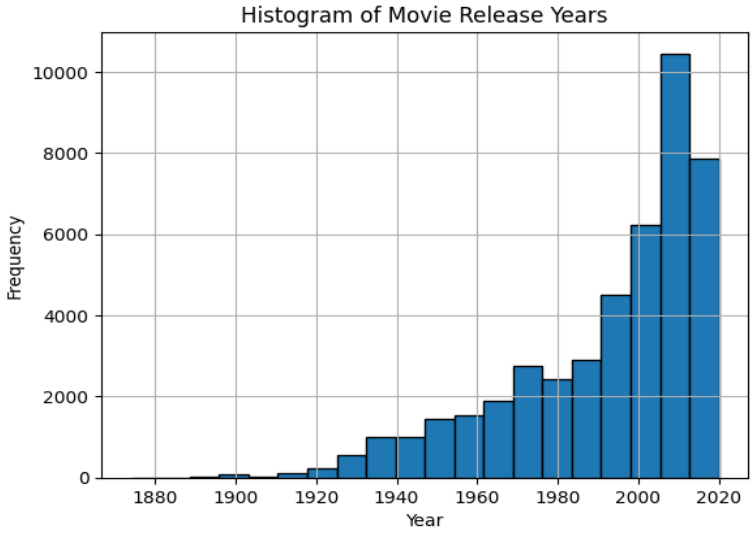


Figure 6

According to the histogram, we can observe a steady increase in movie production over time, with a sharp rise from the 1980s onward, peaking in the 2010s.

Value counts were implemented to identify the years with the highest movie production: 2014 with 1,957 movies, 2015 with 1,891 movies, and 2013 with 1,874 movies. The dataset contains movies from 135 unique release years, further highlighting the diversity of production across different time periods.

- Budget and Revenue:

The dataset contains 45,020 movies, but a significant portion lacks budget data, as 50% of the movies have a recorded budget of zero. While the average budget is approximately $4.26 million, the distribution is highly skewed, with a maximum budget of $380 million, indicating the presence of high-budget blockbusters amid many low-budget or unreported films (The boxplot can be found in the Appendix, Figure 20, 21).

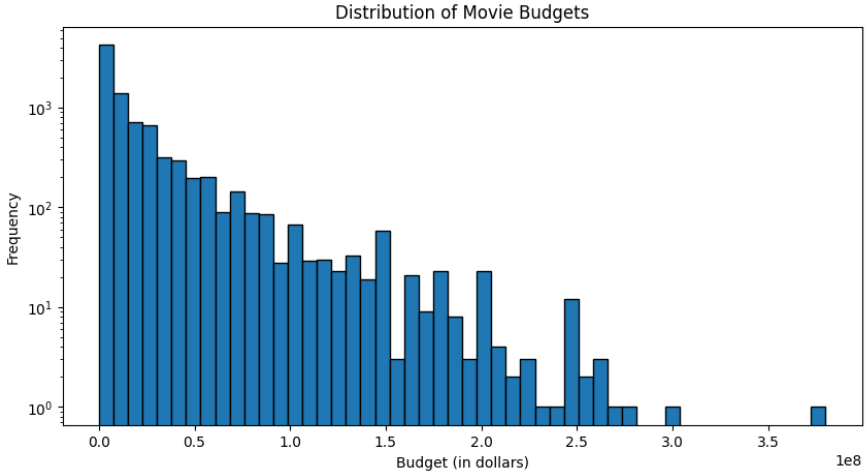


Figure 7

The analysis and visualization for revenue reveal a significant skewness in the data. The average revenue is $11.31 million; however, 50% of the movies report zero revenue. This suggests either missing data or low earnings for many films. While 35957 movies have no recorded revenue, a select few high-grossing blockbusters—with a maximum revenue of $2.79 billion—are responsible for the large variation in overall earnings.

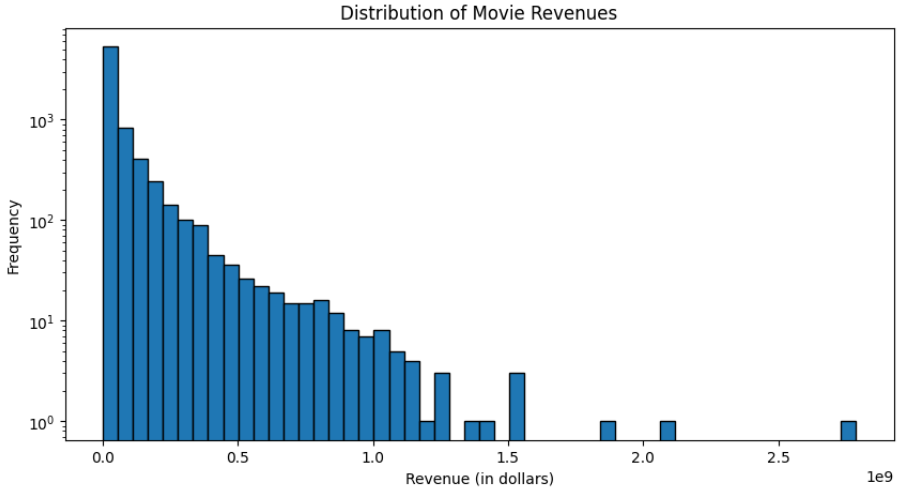


Figure 8

By analyzing both budget and revenue, we were able to calculate the highest and lowest-earning movies in the dataset. The top three highest-grossing movies in our dataset are Avatar, Star Wars, and Titanic. In contrast, the three lowest-grossing films are The Lone Ranger, The Alamo, and Mars Needs Moms.

- Runtime:

The average movie runtime is 94.5 minutes, with most movies ranging between 85 and 107 minutes, which aligns with typical feature film lengths. However, some entries have incorrect or missing data, including a maximum runtime of 1,256 minutes, which is likely a data entry error.

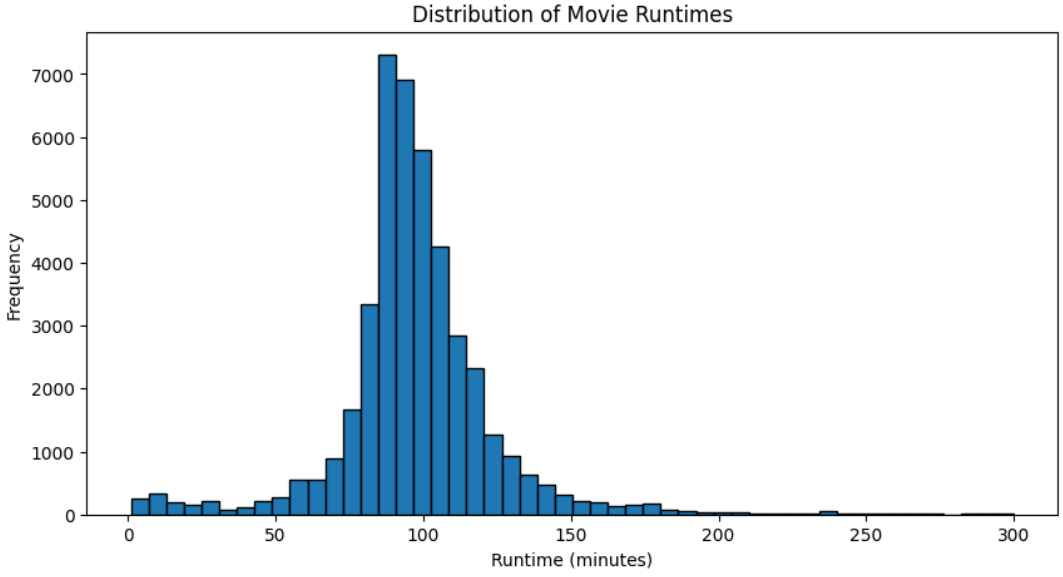


Figure 9

The histogram shows a right-skewed distribution, meaning most movies have shorter runtimes, while fewer movies are much longer. The highest frequency occurs around 90-100 minutes, indicating that most movies follow the traditional feature film length.

- Popularity:

The average popularity score is 3.05, but the distribution is skewed, with 75% of movies having a score below 3.97, indicating that most movies are not widely recognized. A few extreme outliers, such as the highest score of 547.49 (statistic table for movies in the Appendix, Table 1), suggest that only a small number of movies achieve exceptionally high popularity. Some of the most popular movies from the dataset are visualized according to their popularity scores.

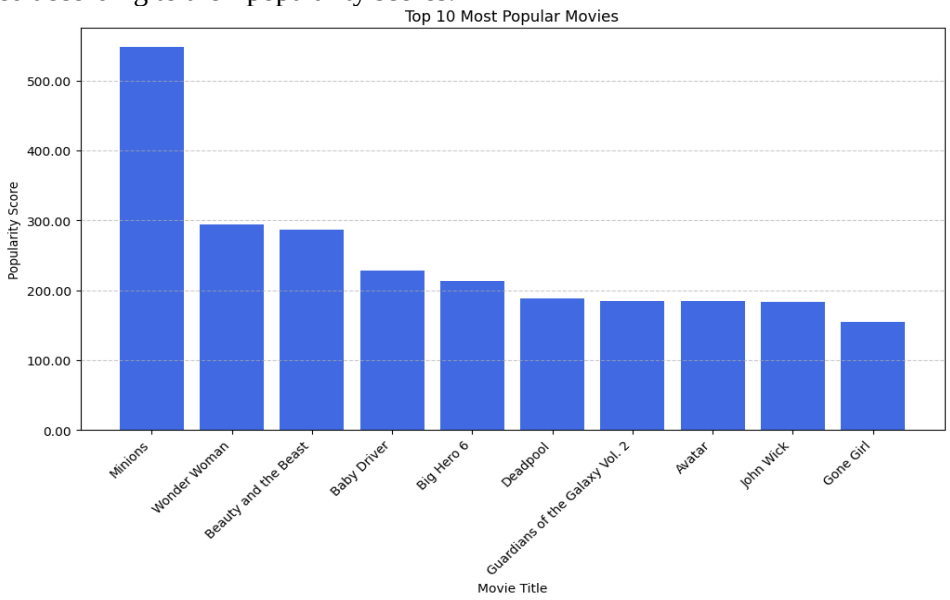


Figure 10

The source of the dataset does not provide details on how the popularity score is calculated, but it is likely based on a combination of factors such as views, searches, ratings, or social media engagement. Without a clear definition, direct comparisons between movies should be interpreted with caution. This feature is analyzed solely for exploratory data analysis and is not included in the final model development features.

B) Ratings Dataset:

- Ratings:

The dataset contains **120,147 unique users** who have rated **7,508 unique movies**, indicating a diverse range of user interactions. The rating distribution in this dataset is right-skewed, with a higher concentration of ratings between 3.0 and 5.0. This suggests that users tend to rate movies more favorably instead of giving very low scores.

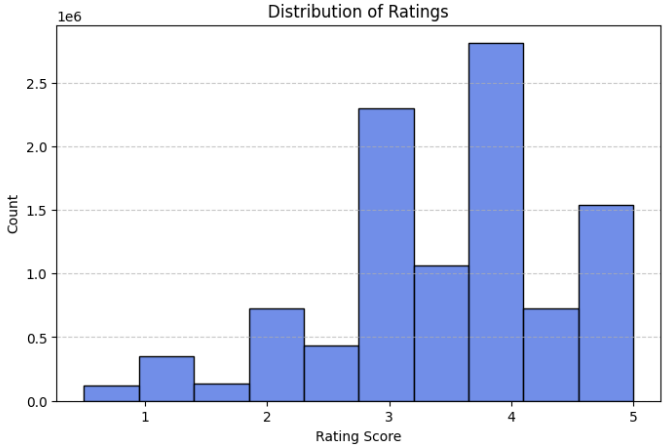


Figure 11

- **Top Movies with Most Number of Ratings:**

The most-rated movies in the dataset include "Terminator 3: Rise of the Machines" (72,611 ratings) and "The Million Dollar Hotel" (67,882 ratings), indicating their high engagement among users. The presence of both mainstream blockbusters and critically acclaimed films in the top-rated list suggests a diverse audience preference in the dataset. The visualization of the top 10 movies with the most ratings is included in the Appendix, Figure 22.

- Timestamps:

The distribution of movie ratings over time shows a significant increase in 1995, likely due to the emergence of online rating platforms where users initially rated both new and old movies. After this early surge, ratings declined before experiencing a resurgence around 2000, possibly influenced by the growing popularity of platforms like IMDb and Rotten Tomatoes. From 2005 to 2015, ratings gradually declined, possibly due to a shift in user behavior, with more focus on rating recent releases rather than older films.

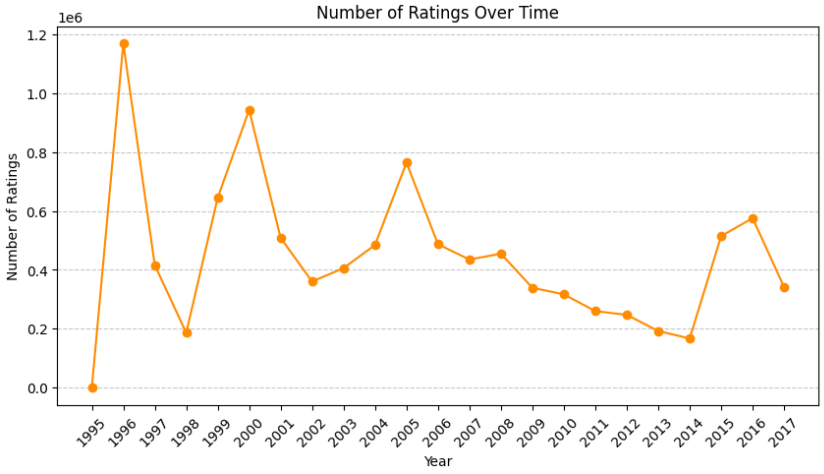


Figure 12

We have also evaluated ratings based on specific times of the year. The analysis of ratings based on the time of the year reveals that Monday and Tuesday receive the highest number of ratings, while activity decreases midweek before rising again on Sunday, suggesting increased engagement at the beginning and end of the week. Similarly, ratings peak in November and December, possibly due to holiday seasons and year-end releases, while September sees the lowest engagement, indicating a seasonal drop in user activity (More visualization can be found in the Appendix, Figure 23).

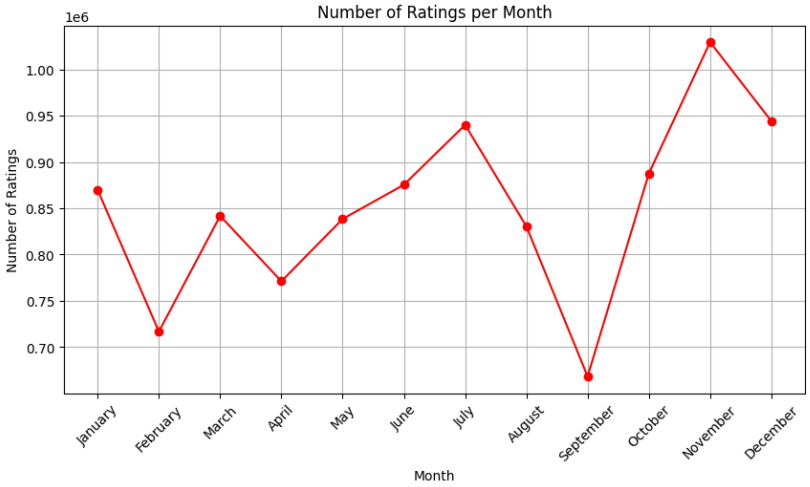


Figure 13

Missing value analysis and outlier analysis

The ratings dataframe contains users' ratings for different movies. The box plot for the ratings distribution depicts the existence of some outliers. The analysis shows that there are 474,464 outliers in the ratings column.

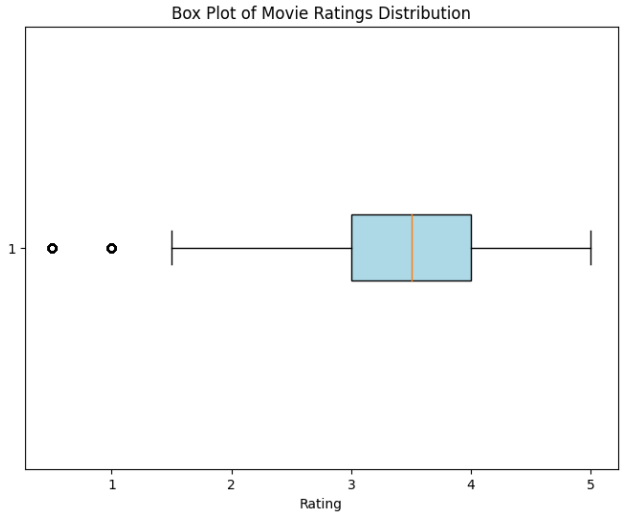


Figure 14

This indicates that many ratings fall outside the expected range, possibly due to inconsistencies in the data. To address this issue, we can consider options such as trimming, capping (Winsorization), or using a weighted rating system. Winsorization is a useful technique for managing extreme values, or outliers, in a dataset. Instead of removing these outliers completely, Winsorization replaces them with the nearest values within a specified percentile range. This approach allows all data points to be retained while minimizing the influence of extreme values on the analysis. For example, in a 5% Winsorization, the lowest 5% of values are replaced with the value at the 5th percentile. After applying this method, all outliers were replaced.

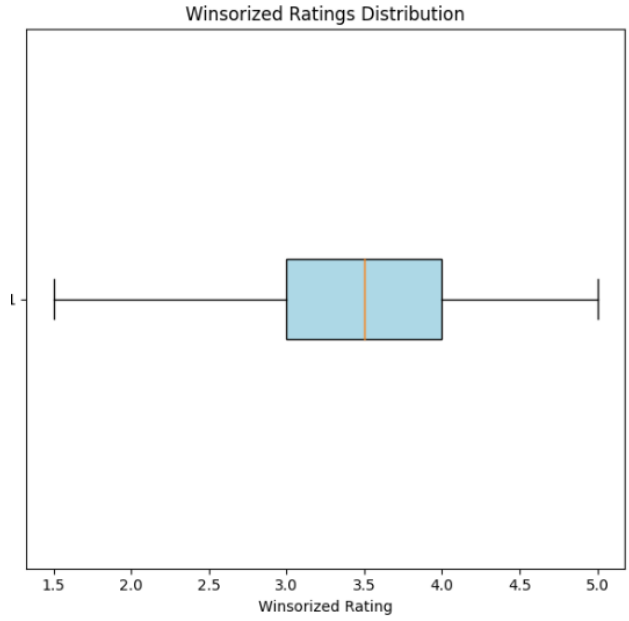


Figure 15

Feature selection justification analysis

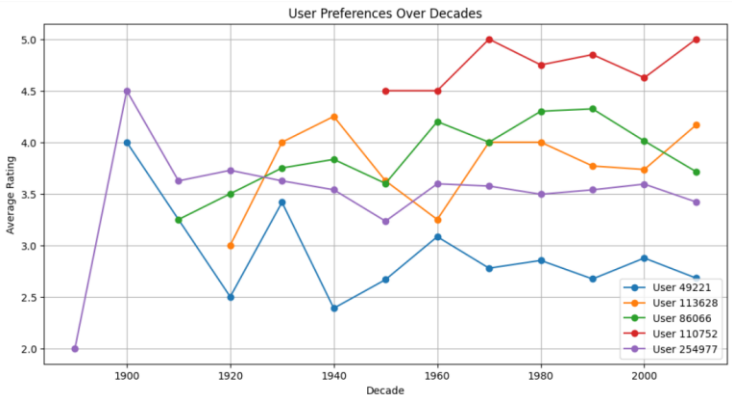
Metadata

We selected **genre**, **first_three_actors**, **director**, **language**, and **year** as key metadata features to enhance the model. To validate this choice, we conducted **some EDA.**

Word Cloud Analysis

- The word clouds below represent **user preferences** based on genre, first_three_actors, director, and language.
- **Larger words** indicate users gave average **higher ratings**, for example, first user gave higher ratings to Crime/Comedy genres, highlighting unique tastes.
- Some words appear frequently across many users, while others are more personalized, confirming that these metadata features play a role in user preferences.

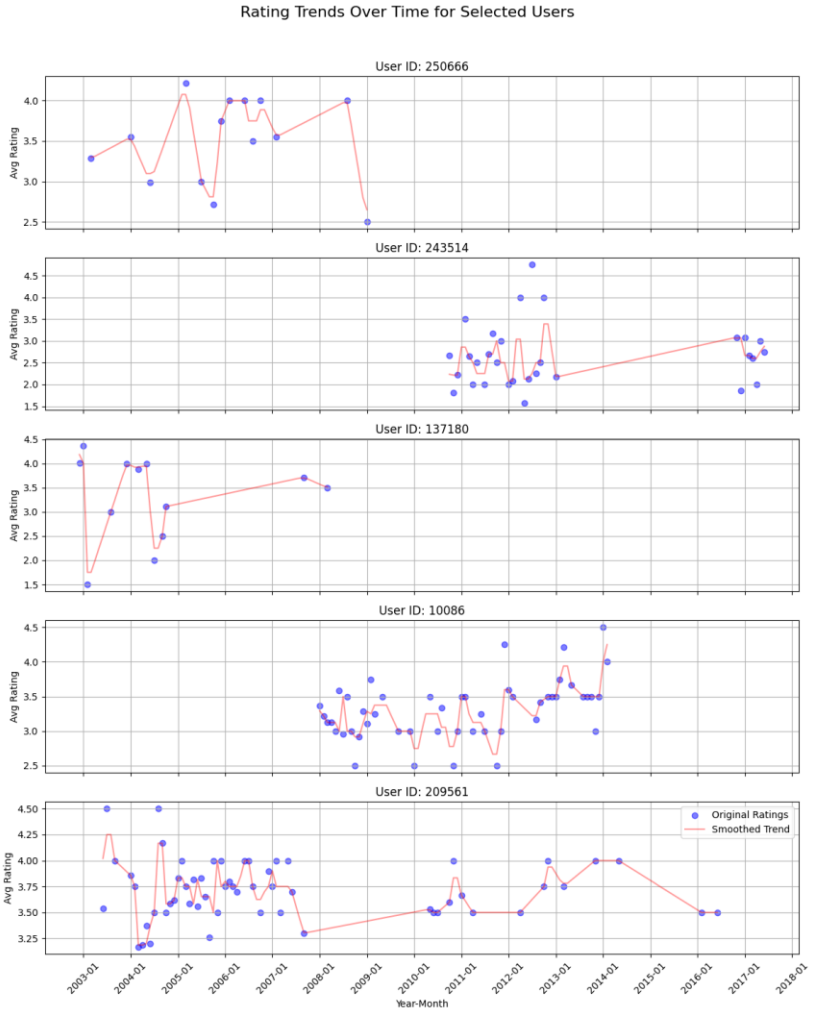
User preferences for movies can vary based on their production year. Some users prefer classic films, while others enjoy modern releases. The analysis below shows that users have different rating patterns for movies from different decades, indicating that production year influences preferences.



Temporal data

User rating behavior changes over time due to various factors. A user who was once strict with ratings may become more generous, or an account shared within a household may reflect different preferences over time.

The plot below shows that a single user's average rating fluctuates over time, indicating that their preferences are not static. Ignoring timestamps may lead to outdated or less relevant recommendations.



Biases

In a rating system, users and movies have different rating patterns. Some users give high ratings to most movies, while others are stricter. Some movies are widely liked, while others are mostly disliked. Without adjusting for these biases, the model may misinterpret ratings and give inaccurate recommendations.

Key Observations from Data

- **Users rate differently:**
  - Generous users (Avg ≥ 4.5): 2,738
  - Strict users (Avg ≤ 2.5): 1,661
- **Movies vary in popularity:**
  - Universally liked movies (Avg ≥ 4.5): 47
  - Universally disliked movies (Avg ≤ 2.5): 1,181
- **Unexpected trends:**
  - Generous users still rated disliked movies high (Avg = 4.18).
  - Strict users rarely rated liked movies at all.

```
Analysis Results:
Number of Generous Users (avg >= 4.5): 2738
Number of Strict Users (avg <= 2.5): 1661
Number of Universally Liked Movies (avg >= 4.5): 47
Number of Universally Disliked Movies (avg <= 2.5): 1181
Avg Rating by Generous Users to Disliked Movies: 4.176566579634465
Avg Rating by Strict Users to Liked Movies: nan
```

Feature engineering and analysis

1. Filtering and Indexing

- **Filtered Movies with Ratings**
  - Retained only movies in `meta_df` that exist in `rate_df`.
- **Indexing User and Movie IDs**
  - Created integer-based IDs (`userIdInt` and `movieIdInt`) for users and movies.
  - Store dictionary mappings from original ids to new integer-based IDs for later retrieval of extra information from raw data, just in case.

2. Handling Categorical Features

- **Language Processing**
  - Defined a threshold of occurrences to classify rare languages as 'other'.
- **Tokenization and Cleaning**
  - Processed `first_three_actors` and `director` columns make each full name a single token:

    Example:

    - **Original:** "Tom Hanks, Tim Allen, Don Rickles"
    - **Processed:** "tomhanks,timallen,donrickles"

3. Encoding Categorical Features

- **One-Hot Encoding (OHE)**
  - Applied OHE to:
    - `genres` column (prefix='G').
    - `original_language` column (prefix='L').
- **Word2Vec Embedding**
  - Trained Word2Vec models to encode:
    - `first_three_actors` as `actor_vec`.
    - `director` as `director_vec`.
  - Represented each movie's actors and director as 20-dimensional vectors.
  - Used mean pooling to obtain fixed-length embeddings.

4. Feature Scaling

  - Applied MinMax scaling to `year` (normalized as `year_norm`).
  - Standardized all features using StandardScaler to avoid feature dominance.

5. Content-Based Similarity Calculation

- **Cosine Similarity Computation**
  - Created a similarity matrix based on the engineered metadata.

6. Bias Computation

- Global Bias (μ): Mean rating across all users and items.
- User Bias (b_u): Deviation of user ratings from the global mean.
- Item Bias (b_i): Deviation of movie ratings from the global mean.

7. Sparse Matrix Representation for CF

- **User-Item Interaction Matrix**
  - Constructed a sparse user-item rating matrix.

8. Shrunken Pearson Similarity Computation

- **Bias-Adjusted Ratings**
  - Adjusted ratings by subtracting the global, user, and item biases.
- **Computed Shrunken Pearson Similarity for Collaborative Filtering***
  - Used Pearson Correlation Coefficient to measure similarity between movies.
  - Applied shrinkage to penalize weak correlations from movies with few common ratings.
  - Converted the similarity matrix into a sparse format for efficient storage.

**Pseudocode can be found in Appendix*

Appendix

Additional EDA

A) Tables of Statistics for Movies and Ratings Datasets:

Table 1

Movies Dataset	id	year	popularity	revenue	runtime	budget
count	45020	45020	45020	4.50E+04	44777	4.50E+04
mean	107470.712	1991.954	2.943525	1.13E+07	94.5184	4.26E+06
std	111978.102	23.91884	6.028728	6.46E+07	37.5312	1.75E+07
min	2	1874	0	0.00E+00	0	0.00E+00
25%	26265.5	1978	0.396847	0.00E+00	85	0.00E+00
50%	59203.5	2001	1.13895	0.00E+00	95	0.00E+00
75%	154682.75	2010	3.732156	0.00E+00	107	0.00E+00
max	469172	2020	547.488298	2.79E+09	1256	3.80E+08

Table 2

Ratings Dataset	userId	movieId	rating	timestamp
count	1.02E+07	1.02E+07	1.02E+07	1.02E+07
mean	1.35E+05	7.20E+03	3.52E+00	1.12E+09
std	7.81E+04	1.84E+04	1.06E+00	2.03E+08
min	4.00E+00	2.00E+00	5.00E-01	7.90E+08
25%	6.73E+04	5.00E+02	3.00E+00	9.51E+08
50%	1.35E+05	1.48E+03	3.50E+00	1.10E+09
75%	2.03E+05	3.06E+03	4.00E+00	1.27E+09
max	2.71E+05	1.76E+05	5.00E+00	1.50E+09

B) Analysis and Visualizations of Categorical Variables:

Unique genre distribution in our dataset:

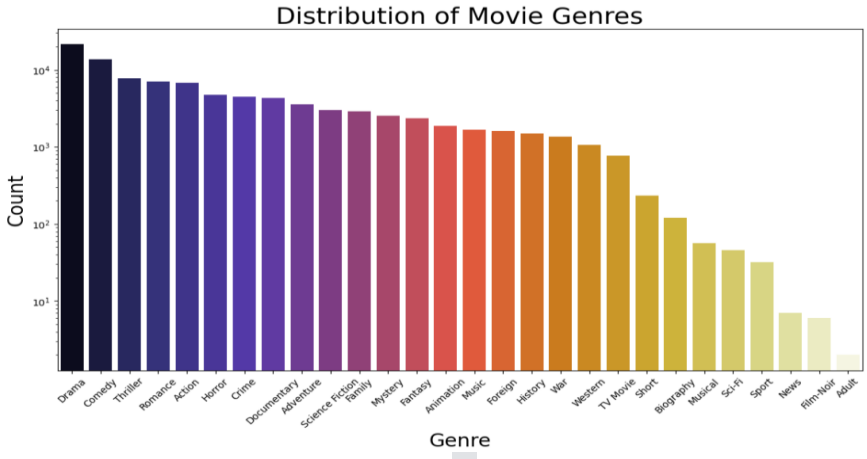


Figure 16

Top 10 actors in action movies:

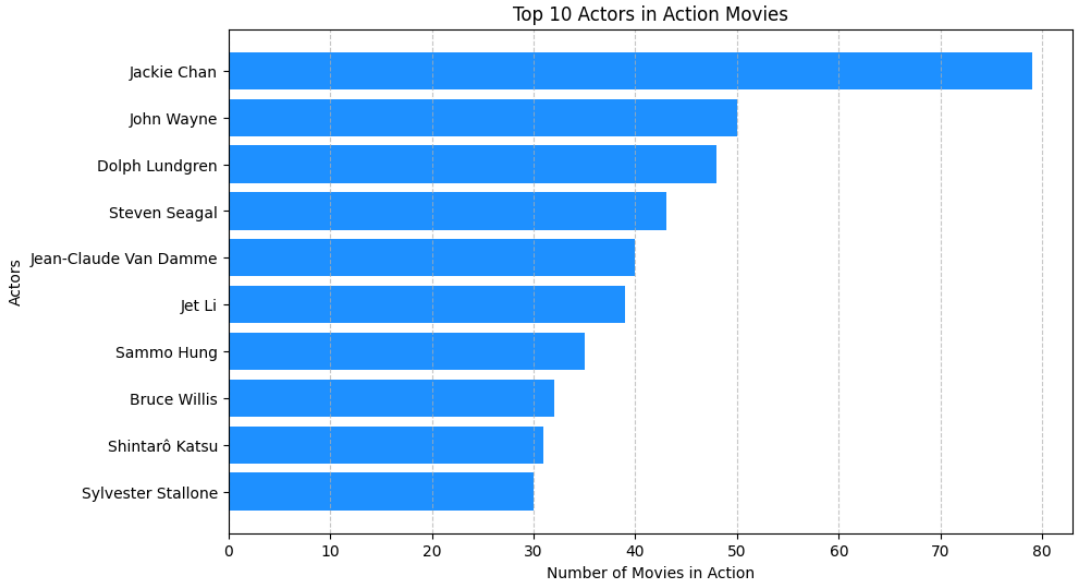


Figure 17

Average Rating by Language:

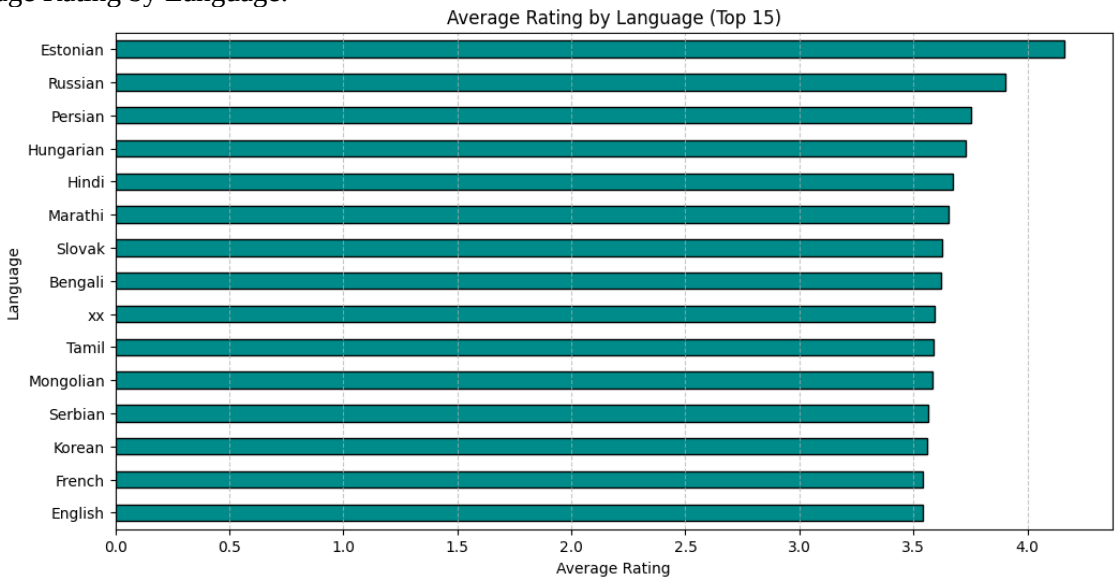


Figure 18

Top 10 Most Common Production Countries:

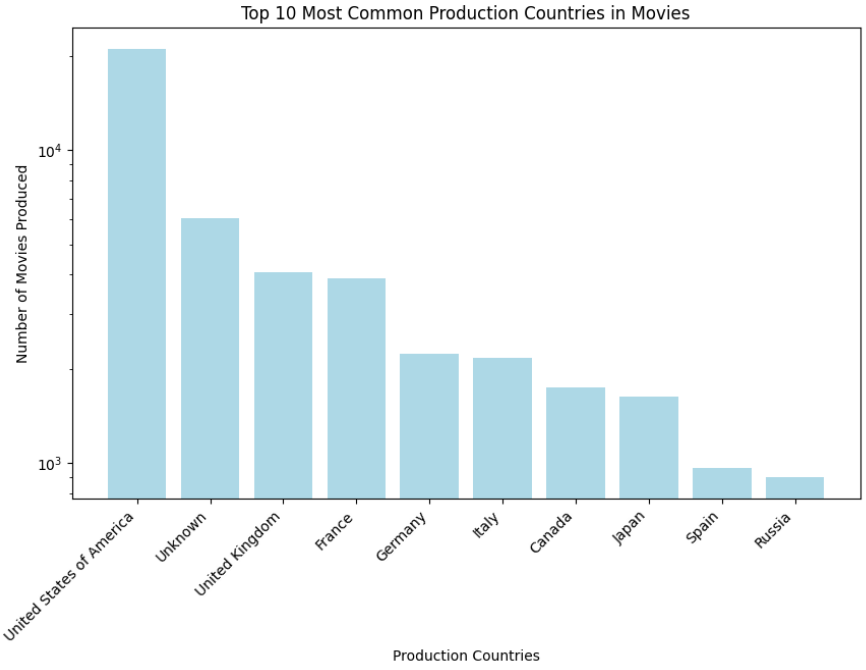


Figure 19

Box Plot for Budget (This feature is only visualized for EDA purposes and is not included among the features for model development; therefore, outlier handling was not required):

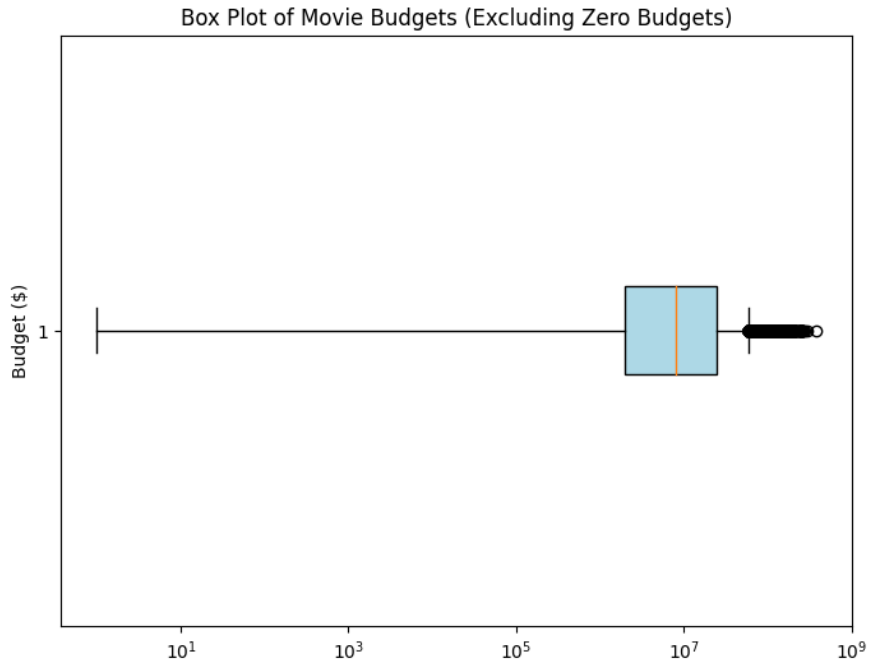


Figure 20

Box Plot for Revenue (This feature is only visualized for EDA purposes and is not included among the features for model development; therefore, outlier handling was not required):

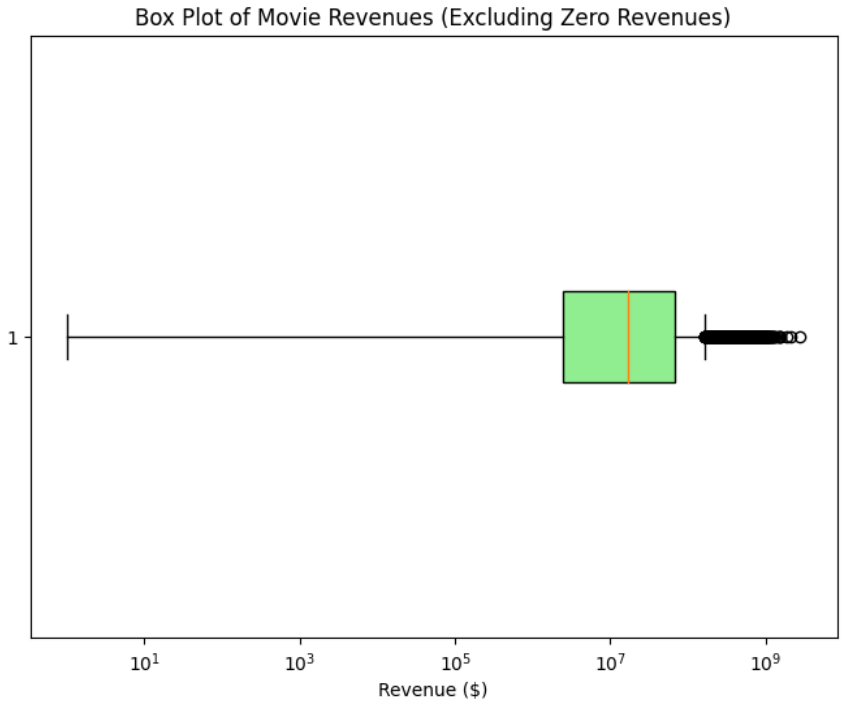


Figure 21

Top 10 Most-Rated Movies:

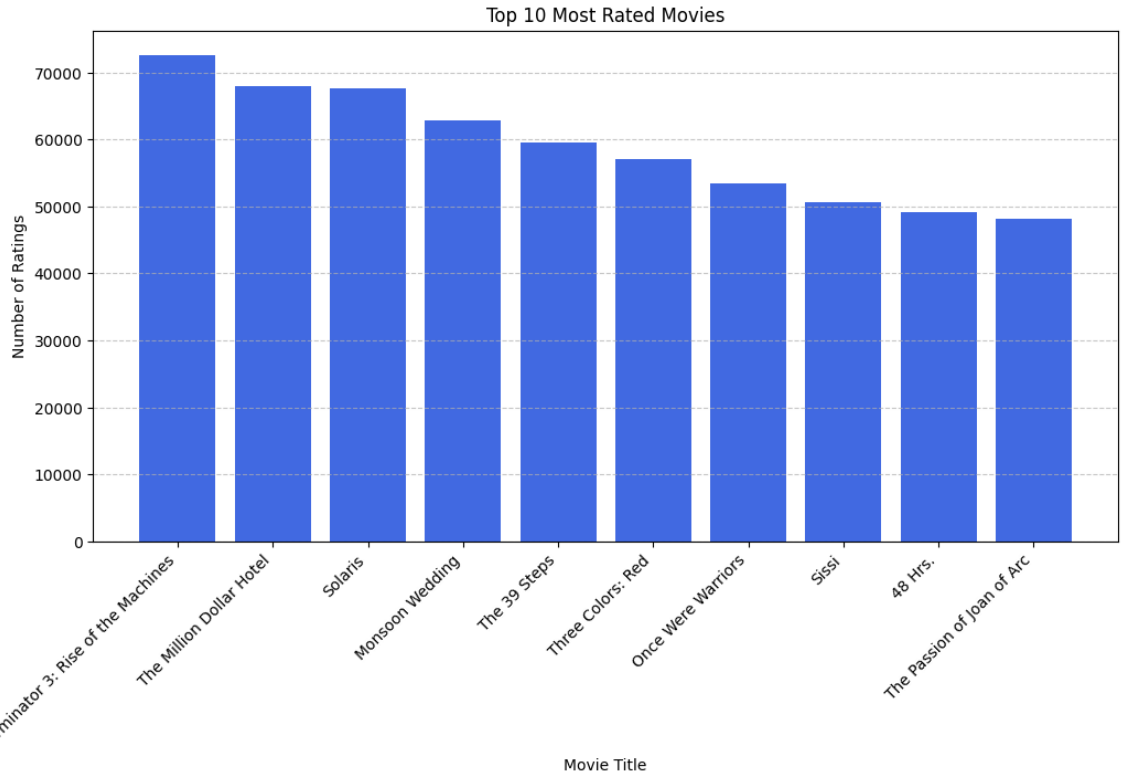


Figure 22

Timestamp: Number of Ratings by Day of the Week:

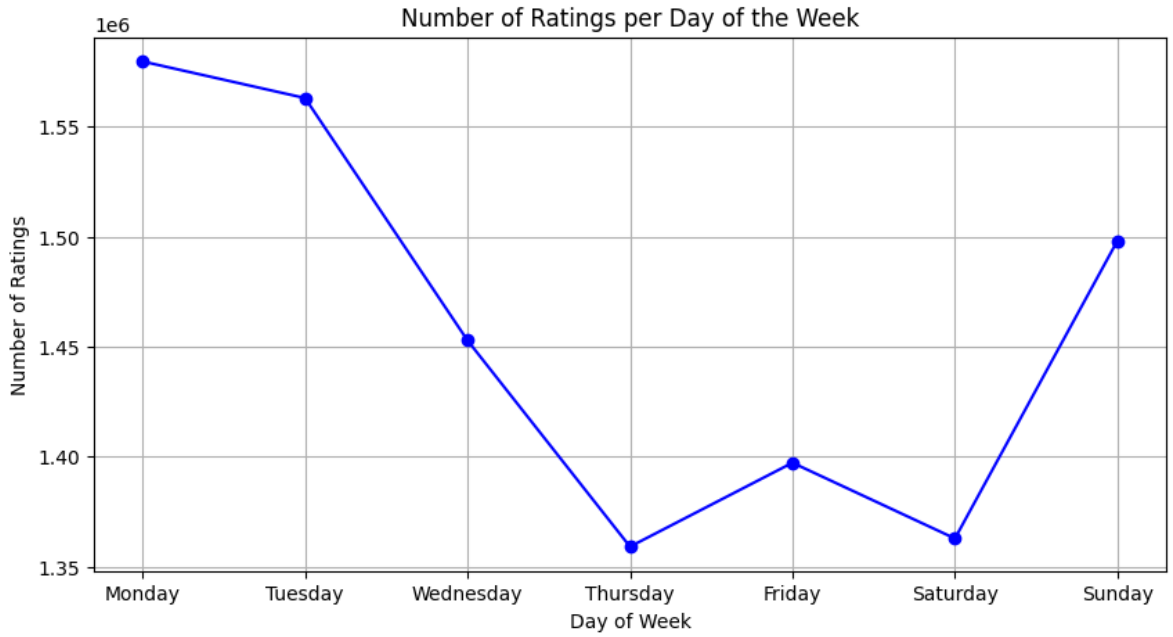


Figure 23

Pseudocode

A) Shrunken Pearson Similarity Computation:

```
INITIALIZE similarity_matrix as zero matrix (movies x movies)

FOR each movie i:
    FOR each movie j where j > i:
        FIND users who have rated both movies i and j
        IF number of co-rated users > 1:
            COMPUTE Pearson correlation between ratings of movie i and movie j
            APPLY shrinkage factor:
                weight = num_com_users / (num_com_users + shrink_value)
                similarity_matrix[i, j] = correlation * weight
                similarity_matrix[j, i] = similarity_matrix[i, j]

CONVERT similarity_matrix to sparse format
SAVE as .npz file
```

B) KNN Model Building Progress:
We have made some progress towards the next phase of model building and training, following is a brief about the progress so far.

Koren Neighborhood Models (Koren-KNN) rely on collaborative filtering (CF) similarity to make recommendations. However, CF alone struggles with cold-start issues and sparsity problems. To enhance recommendation accuracy, our approach modifies Koren-KNN by integrating content-based similarity and learning dynamic similarity weights.

This approach learns a hybrid similarity matrix where a trainable weight γ_{ij} balances CF-based similarity and content-based similarity dynamically for each item pair.

Key Modifications - Hybrid Similarity Computation:

- Instead of relying purely on CF similarity, we introduce a **content-based similarity matrix**.
- We dynamically learn a hybrid similarity score $\widetilde{w_{ij}}$ that combines both similarities:

$$\widetilde{w_{ij}} = \gamma_{ij} w_{ij}^{\text{rating}} + (1 - \gamma_{ij}) w_{ij}^{\text{content}}$$

What's the Same:

- **Learned Weights for Similarity:**
    - Unlike fixed-weight models, w_{ij}^{rating} and w_{ij}^{content} are learned iteratively during training just like the original Koren-KNN model.

- **Bias Adjustment:**
    - Like the original Koren-KNN model, user bias b_u and item bias $\boldsymbol{b_i}$ are optimized during training.
- **Neighbor Selection:**
    - **Top-K neighbors** are selected based on the hybrid similarity matrix.

Initialize To-learn Variables:

```
LOAD precomputed CF-based similarity matrix: cf_sim_matrix
LOAD precomputed content-based similarity matrix: con_sim_matrix
LOAD precomputed global bias miu

INITIALIZE cf_weight_ij, con_weight_ij, hybrid similarity weight γ_ij randomly
INITIALIZE biases b_u and b_i randomly
```

One Training Epoch:

```
FOR each item i:
    FOR each user u with a rating r_ui:
        FIND top-K similar items N_k(i, u) based on combined similarity

        COMPUTE predicted rating

        COMPUTE error between prediction and label

        UPDATE biases b_u and b_i
        UPDATE similarity weights cf_weight_ij and con_weight_ij
        UPDATE hybrid weight γ_ij
```

Training Loop:

```
WHILE not converged:
    RUN one training epoch
    CHECK if error change is below threshold
```

Test and Evaluation:

```
FOR each test user u:
    FOR each movie i:
        FIND top-K neighbors N_k(i, u)
        COMPUTE predicted rating r_ui_hat
        COMPUTE squared error:
                total_squared_error += (r_ui - r_ui_hat)²
        INCREMENT count
COMPUTE RMSE:
    RMSE = sqrt(total_squared_error / count)
```

Table of Contributions

The table below identifies contributors to various sections of this document.

	Section	Writing	Editing
1	**Analysis the basic metrics of variables**	**Alireza Hatami**	
2	**Non-graphical and graphical univariate analysis**	**Alireza Hatami**	**Precious Orekha**
3	**Missing value and outlier analysis**	**Alireza Hatami**	
/	**Feature Selection Justification**	**Jaz Zhou**	
4	**Feature engineering and analysis**	**Jaz Zhou**	**Caitlin Dunne**
5	**Appendix**	**Alireza Hatami** **Jaz Zhou**	

Grading

The grade is given on the basis of quality, clarity, presentation, completeness, and writing of each section in the report. This is the grade of the group. Individual grades will be assigned at the end of the term when peer reviews are collected.