Reel Good

Movie Recommender System

G4
Alireza Hatami
Caitlin Dunne
Jaz Zhou
Precious Orekha

Image Source: Go on a thrill ride with action-packed movies like Extraction, The Gray Man, RRR, and The Old Guard. . Accessed on [Feb 6 2025].
Available at [https://dnm.nflximg.net/api/v6/2DuQlx0fM4wd1nzqm5BFBi6ILa8/AAAAQRC29H19twWKcTZ9Zpg4biJbGNaHF2GGlYNeZ6fvwugUJbuKxTjjMFPCS-y5P3ZePL57rupDtSkyUIJhv3P8leMJGMzszuG2CHNd65NwWPu5LeKxQkRNfNMHmxAwt7tmQZFk1VlrBd1aXr2AR5DM.jpg?r=5b1]

Project Recap

- Our project is creating a recommender system for movies (similarly to Netflix)
- We are only using collaborative filtering for this project
- Out dataset has both reviews and associated metadata for over 45,000 movies released prior to August 2017
- We have selected k-nearest neighbors and SVD++ as the algorithms to build our recommender system with

ANALYSIS OF THE BASIC METRICS OF VARIABLES

Movies Dataset:

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

ANALYSIS OF THE BASIC METRICS OF VARIABLES

- Budget Statistics

- Mean (Average): 4.26 million dollars
- Standard Deviation (std): 17.50 million dollars → There is a Significant variation in budget values.
- Minimum Budget: 0 (Missing or unknown budgets).
- Maximum Budget: 380 million dollars (Blockbuster movies).

- Revenue Statistics

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

- Year Statistics

- The dataset includes movies released between 1874 and 2020.
- 50th Percentile (Median): The median release year is 2001 → At least half of the movies were produced after the early 2000s
- 75% of the movies in the dataset were released in the year 2010 or earlier

ANALYSIS OF THE BASIC METRICS OF VARIABLES

Ratings Dataset:

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

ANALYSIS OF THE BASIC METRICS OF VARIABLES

- Rating Statistics:

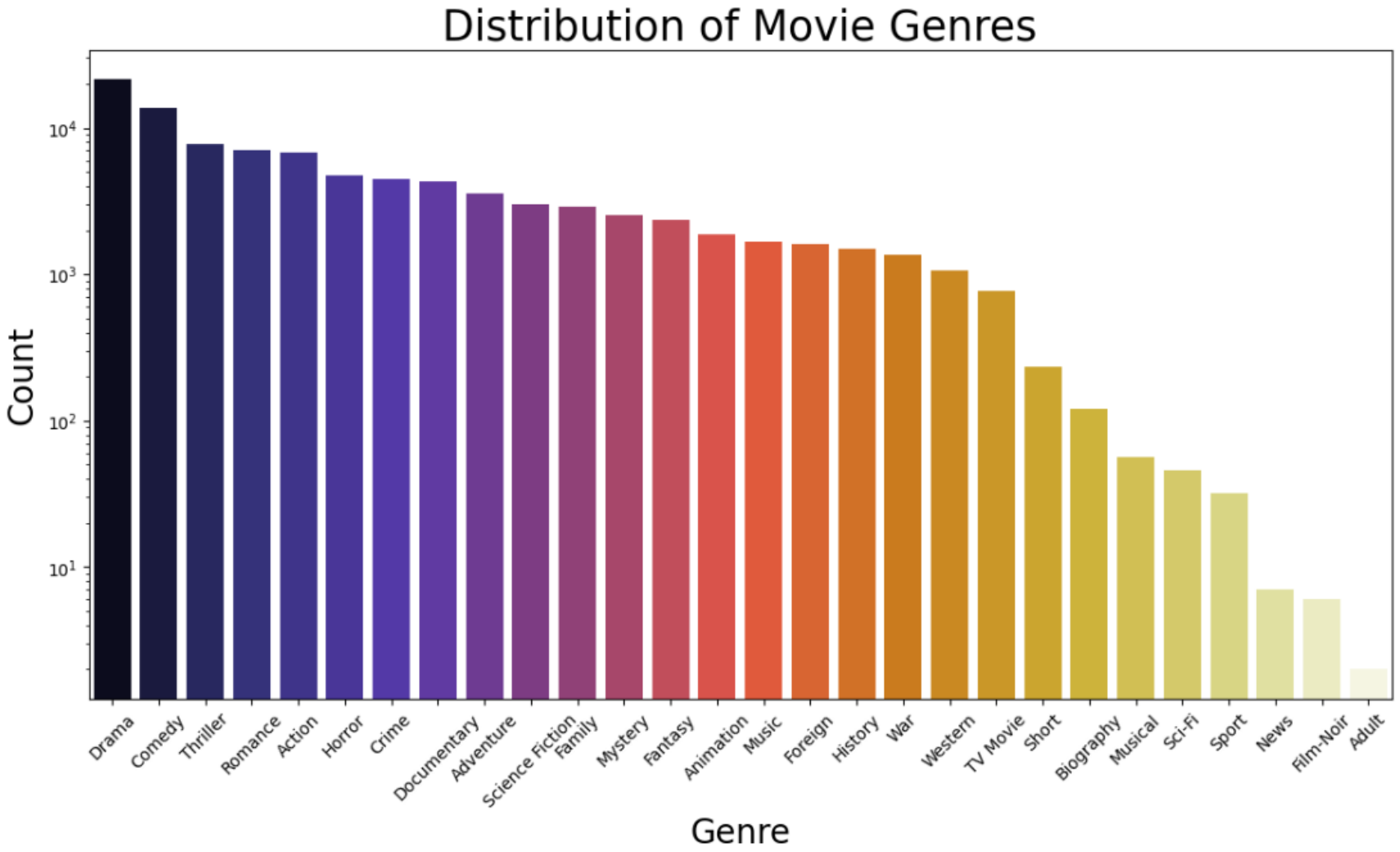
- The dataset contains over 10 million ratings.
- Average Rating: 3.52 → Most movies receive mid-range scores.
- Standard Deviation (std): 1.06 → Users have diverse opinions about movies.
- Minimum and Maximum Rating → The minimum is 0.5 / The maximum is 5
- 50th Percentile (Median, Q2): 3.50 → Half of all ratings are below 3.50, and half are above.

- Timestamp Statistics:

- Mean: The average timestamp corresponds to February 2005. → Most ratings in the dataset were given around the mid-2000s.
- 50th Percentile (Median): December 2004 → Half of the ratings were recorded before this date, and the other half were recorded after.
- Minimum Timestamp: January 1995 → The beginning of the dataset's time range.
- Maximum Timestamp: August 2017 → The dataset spans over 22 years of movie ratings.

CATEGORICAL VARIABLES IN MOVIES DATASET

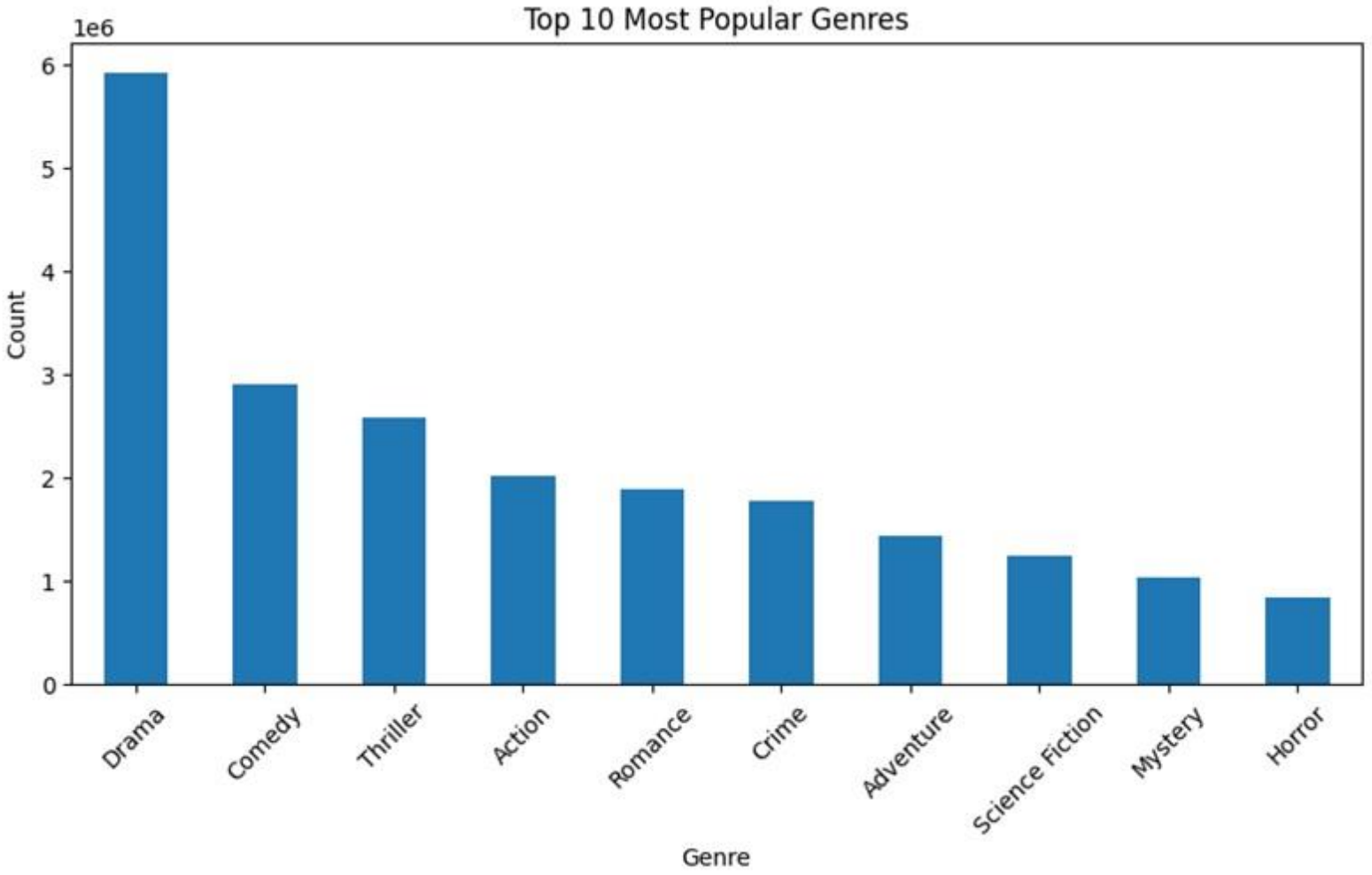
- **Genre:**
  - Drama (21,373 movies)
  - Comedy (13,790 movies)
  - Film-Noir (6 movies), News (7 movies), and Sport (32 movies) are the least represented



CATEGORICAL VARIABLES IN MOVIES DATASET

- User Ratings and Genre:

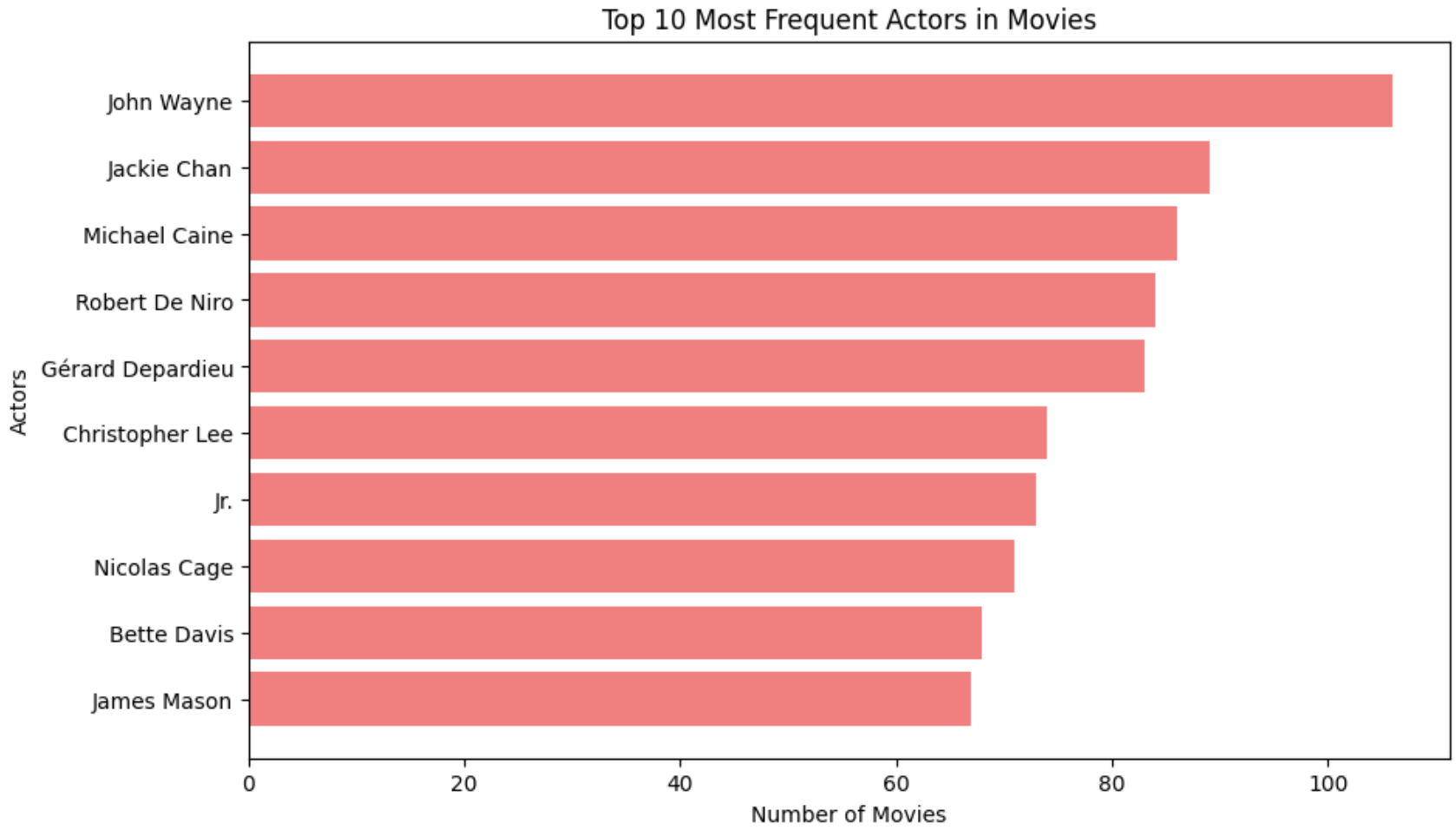
- ✓ The figure illustrates the number of ratings users gave for genres
- ✓ **Drama, Comedy**, **Thriller**, and **Action** movies are more popular than others.



CATEGORICAL VARIABLES IN MOVIES DATASET

- Actors:

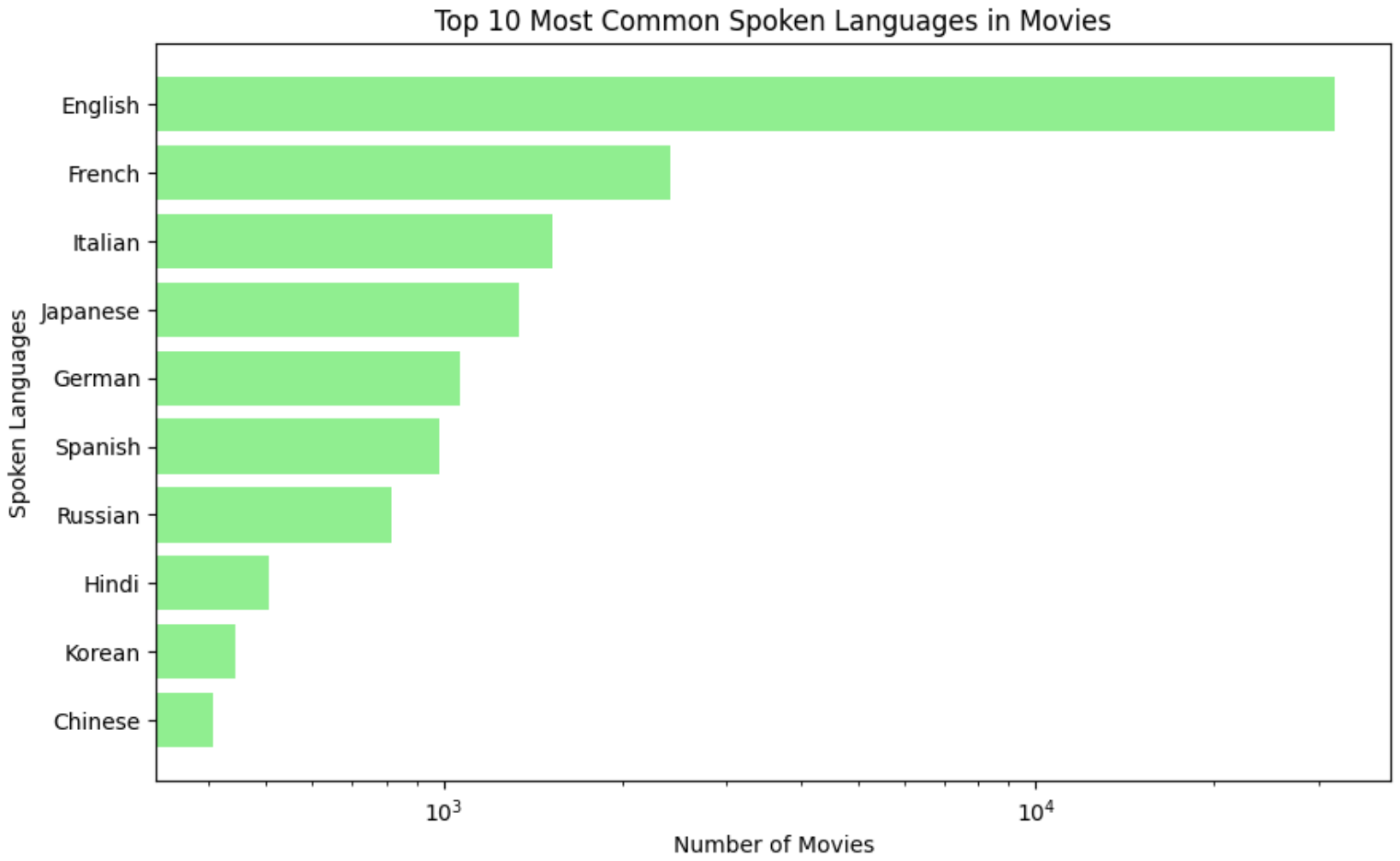
- ✓ The number of actors extracted from our dataset is 48269.
- ✓ We ranked the actors based on the number of movies they appeared in.



CATEGORICAL VARIABLES IN MOVIES DATASET

- Language:

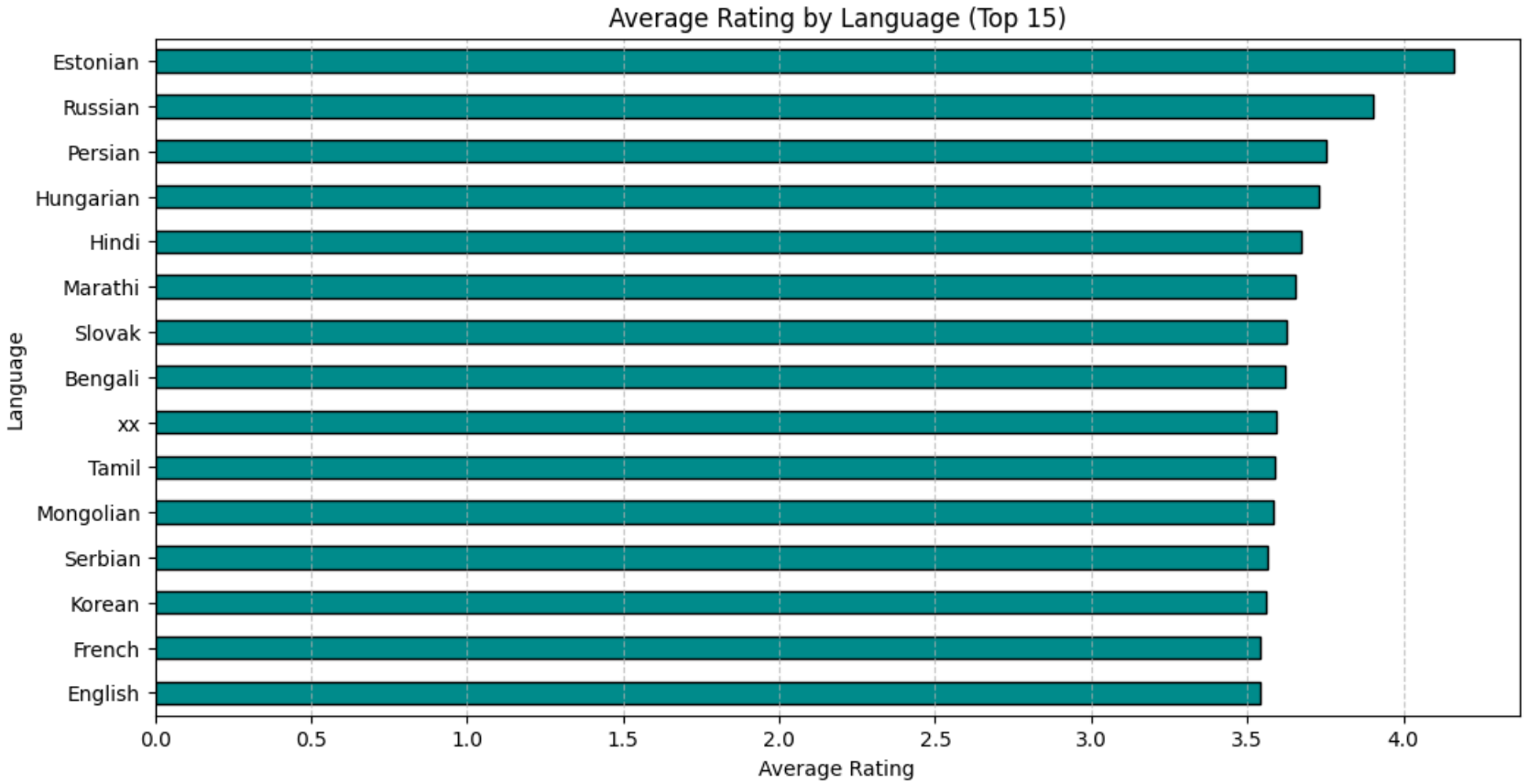
- ✓ Our dataset includes 89 languages for the movies.
- ✓ Due to the large number of languages, we visualized the top 10.
- ✓ Most of the movies → Are in **English**, **French**, and **Italian**.



CATEGORICAL VARIABLES IN MOVIES DATASET

- Ratings and Language:

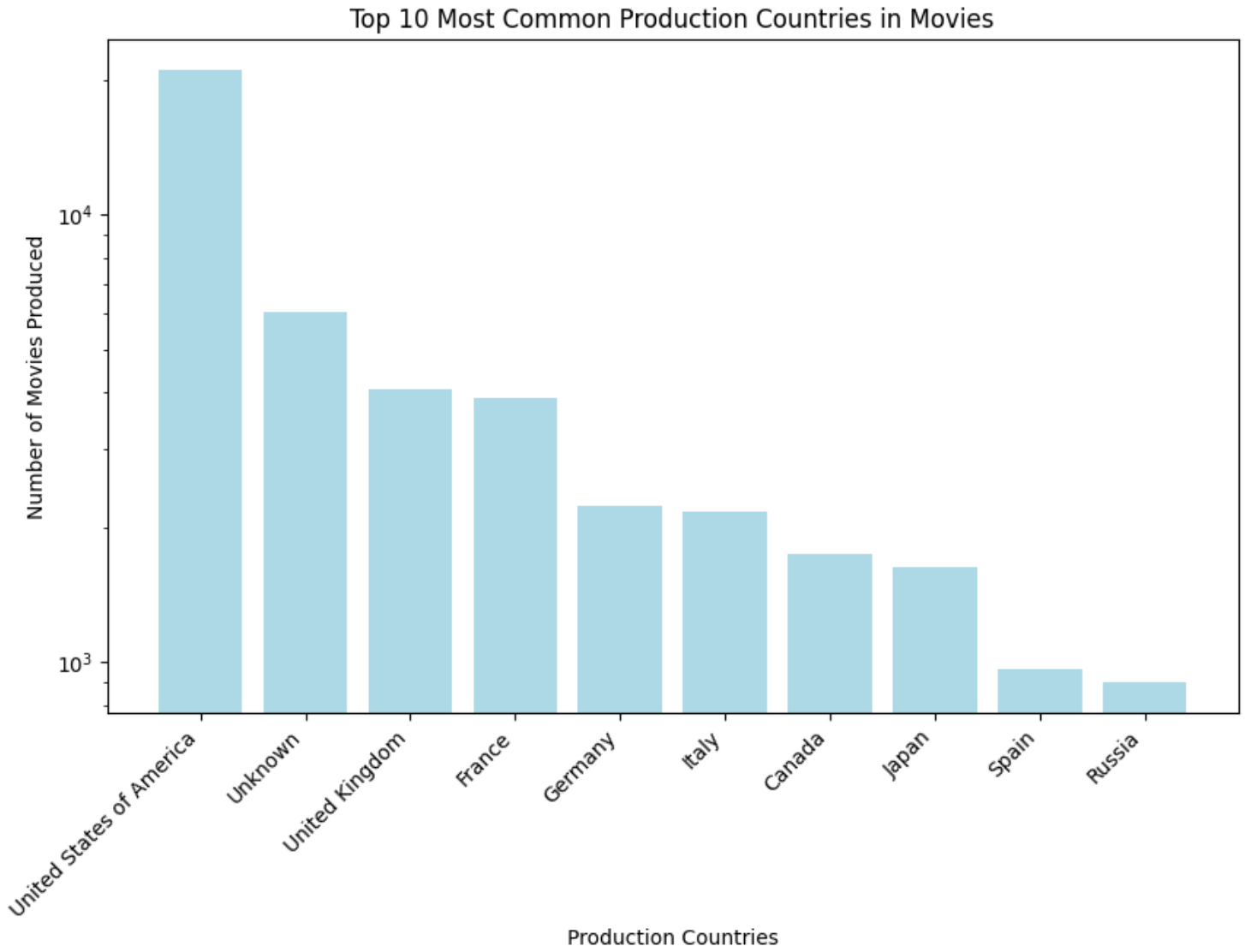
✓ The figure illustrates the average ratings received by users for each spoken language



CATEGORICAL VARIABLES IN MOVIES DATASET

- Production Countries:

- ✓ Our dataset includes a total of 161 unique production countries.
- ✓ The **United States of America** has the highest number of movies (21061).
- ✓ However, there is a substantial group of movies (6,077) for which no production countries were recorded.

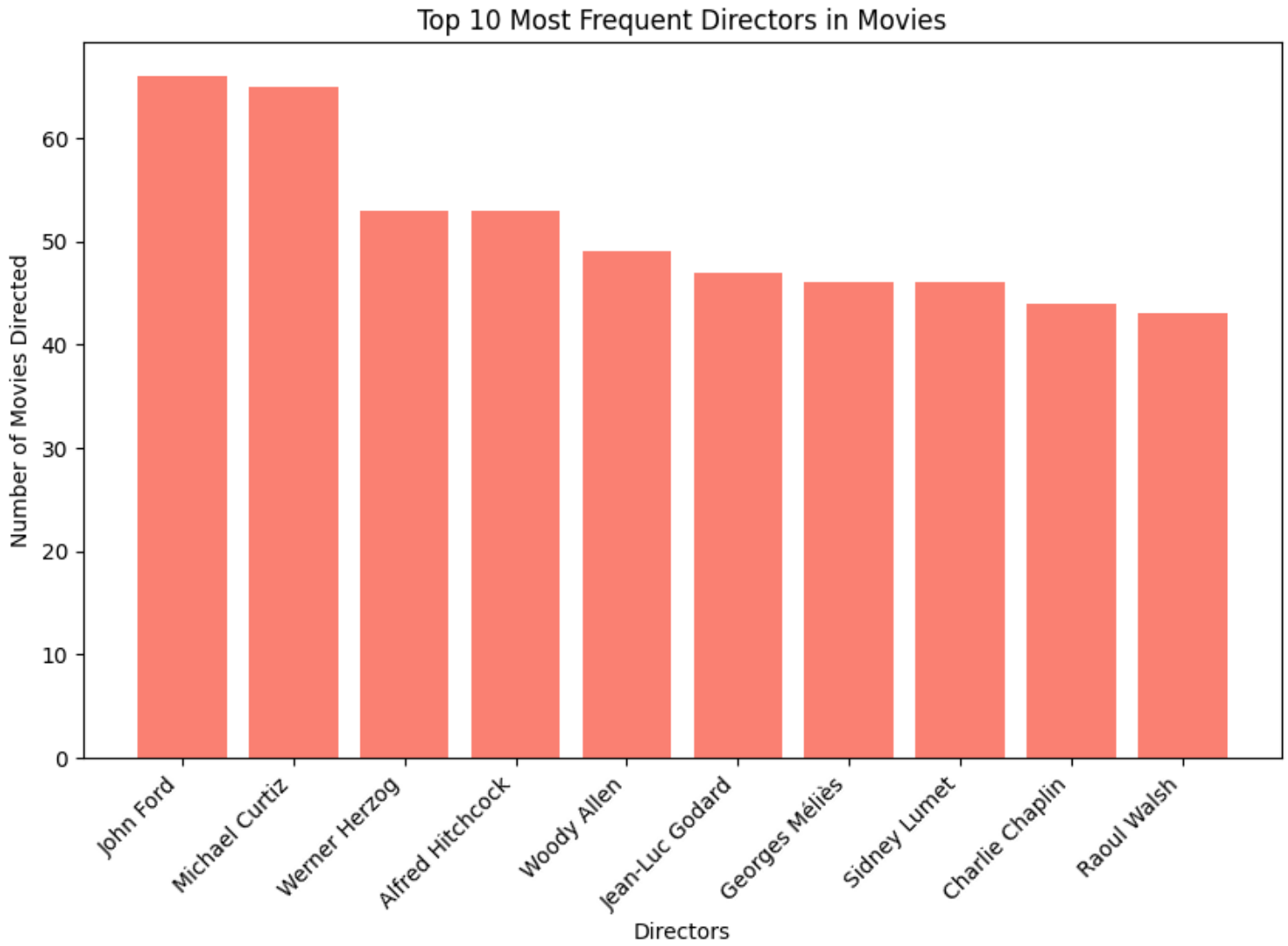


CATEGORICAL VARIABLES IN MOVIES DATASET

- **Directors:**

✓ The total number of directors in our dataset is 17884.

✓ The figure visualized the top 10 directors who made the most films



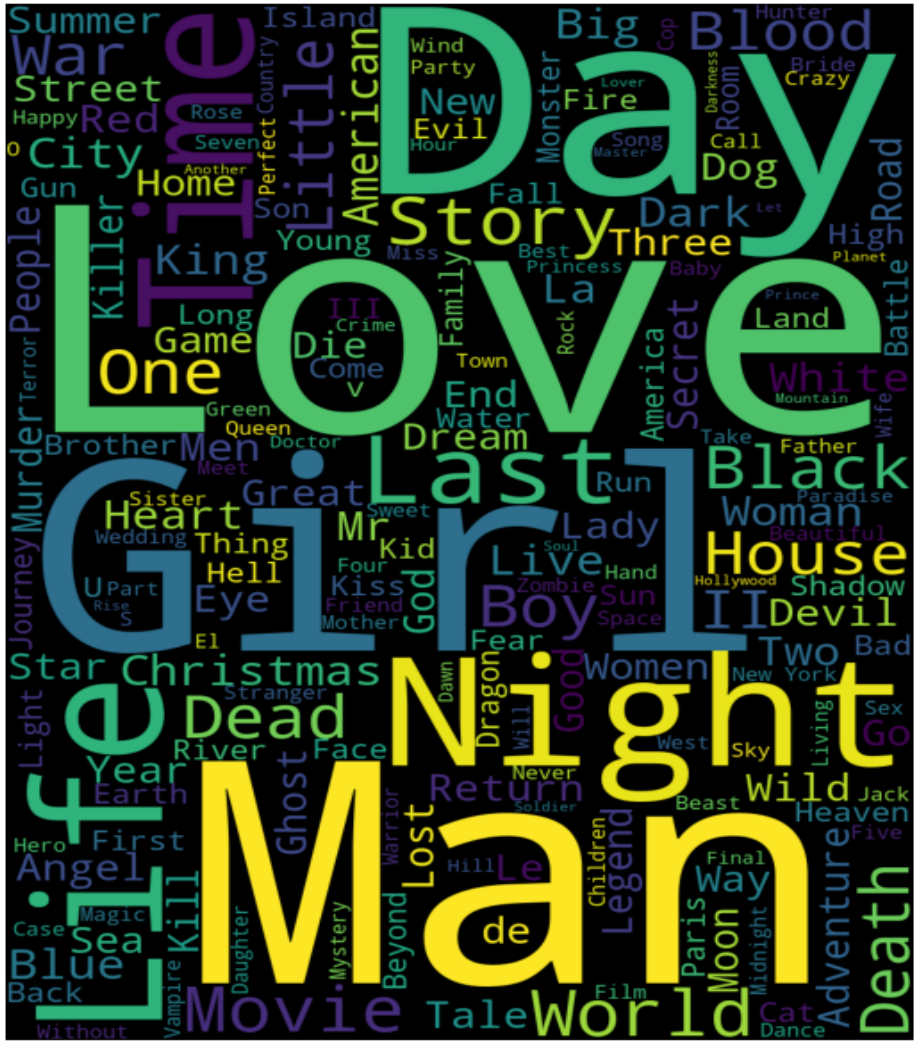
CATEGORICAL VARIABLES IN MOVIES DATASET

- Title:

✓To visualize the words that stand out in movie titles, we used a WordCloud.

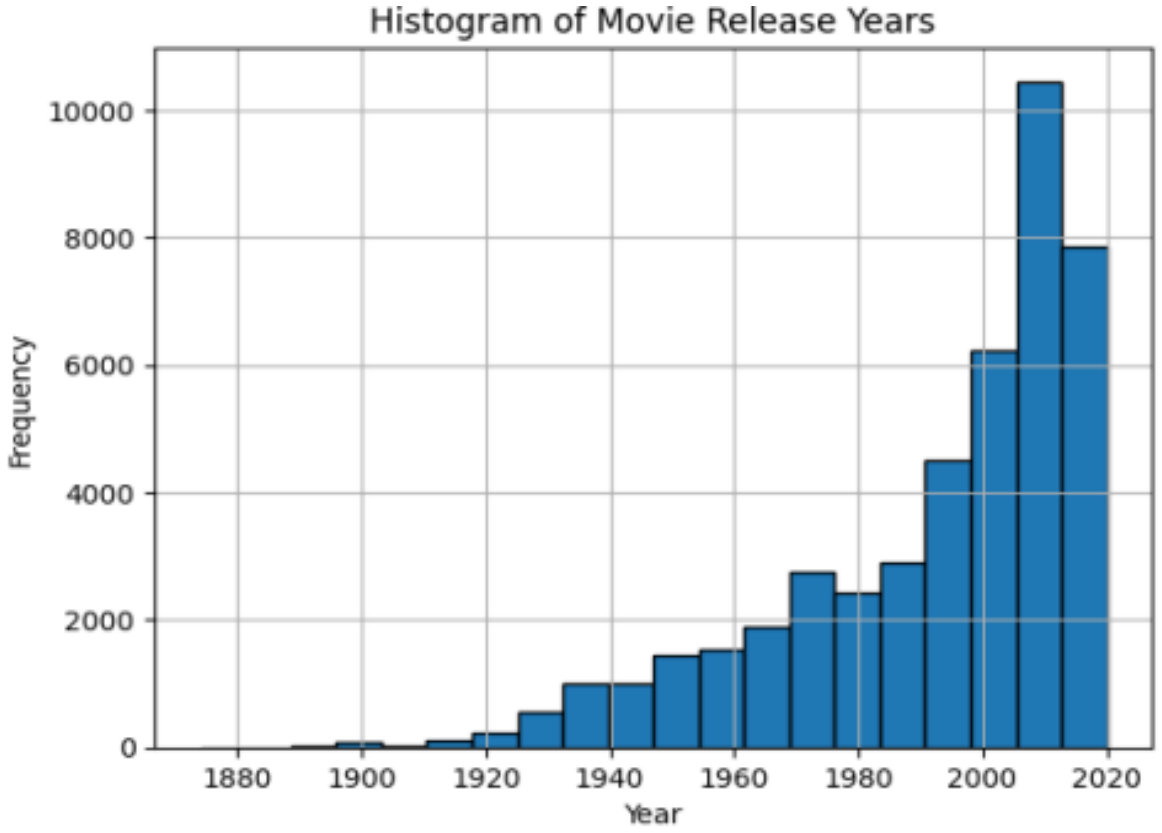
✓The most frequent words include Love, Day, Girl, Man, Life, and Night.

Stand-Out Words Present on Movie Titles



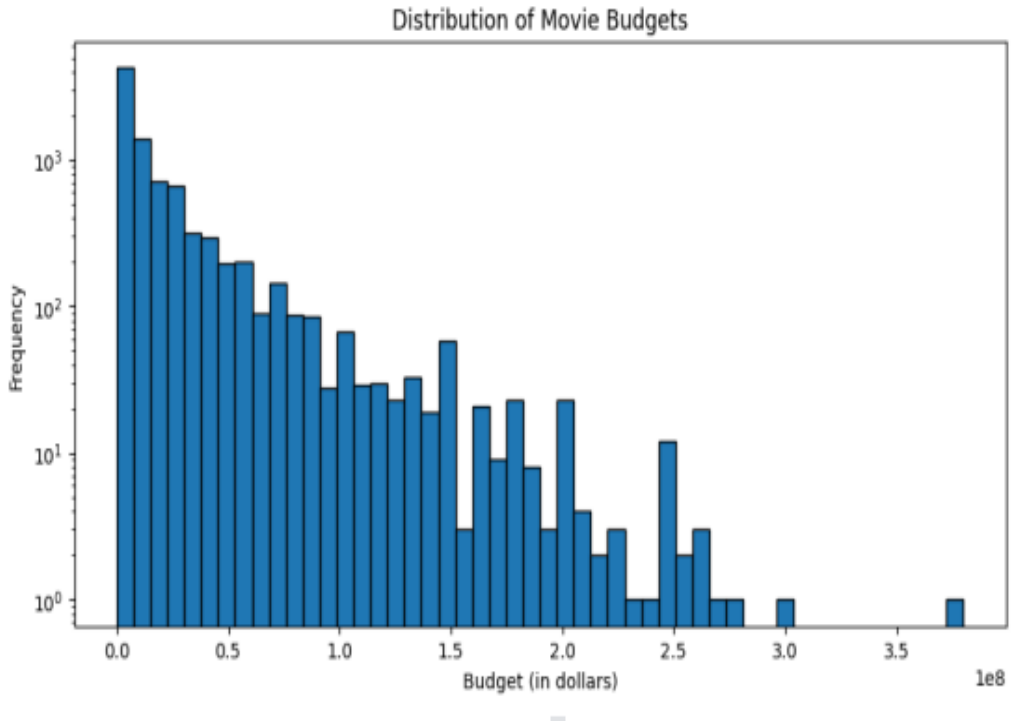
NUMERICAL VARIABLES IN MOVIES DATASET

- 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
- Steady increase in movie production over time, with a sharp rise from the 1980s onward, peaking in the 2010s



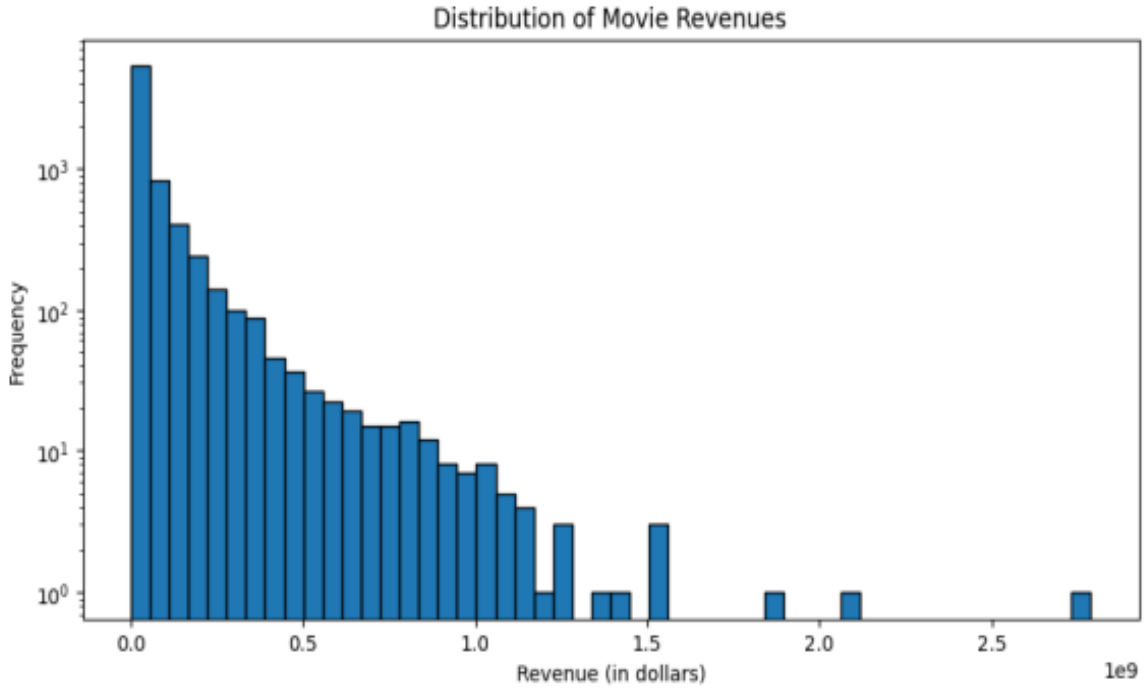
NUMERICAL VARIABLES IN MOVIES DATASET

- The dataset contains 45,020 movies, but a significant portion i.e. 50% of the movies have a recorded budget of zero.
- The distribution is highly skewed, with an average budget of $4.2 million, and a maximum budget of $380 million
- The budget distribution highlight the presence of high-budget blockbusters amid many low-budget or unreported films



NUMERICAL VARIABLES IN MOVIES DATASET

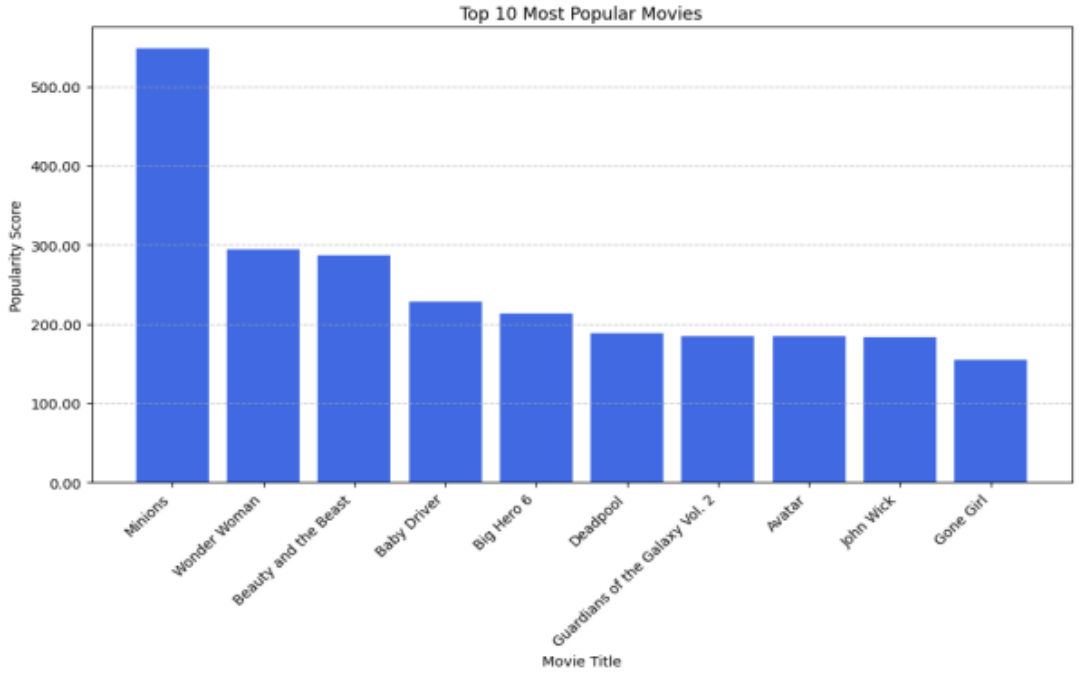
- The average revenue was $11.31 million; however, 50% of the movies reported zero revenue. This suggested either missing data or low earnings for many films
- Top 3 highest grossing: **Avatar, Star Wars, and Titanic**
- Top 3 lowest grossing: **The Lone Ranger, The Alamo,** and **Mars Needs Moms**



NUMERICAL VARIABLES IN MOVIES DATASET

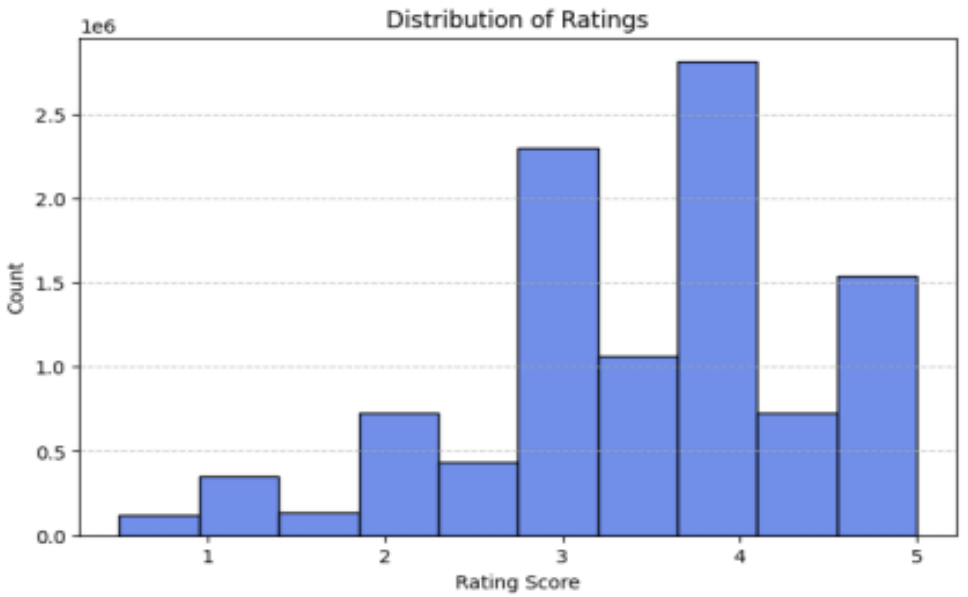
- The average popularity score is 3.05, but the distribution is skewed, with 75% of movies having a score below 3.97, indicating that most movies are not widely recognized

- The popularity score is calculated based on a combination factors such as views, searches, ratings, or social media engagement



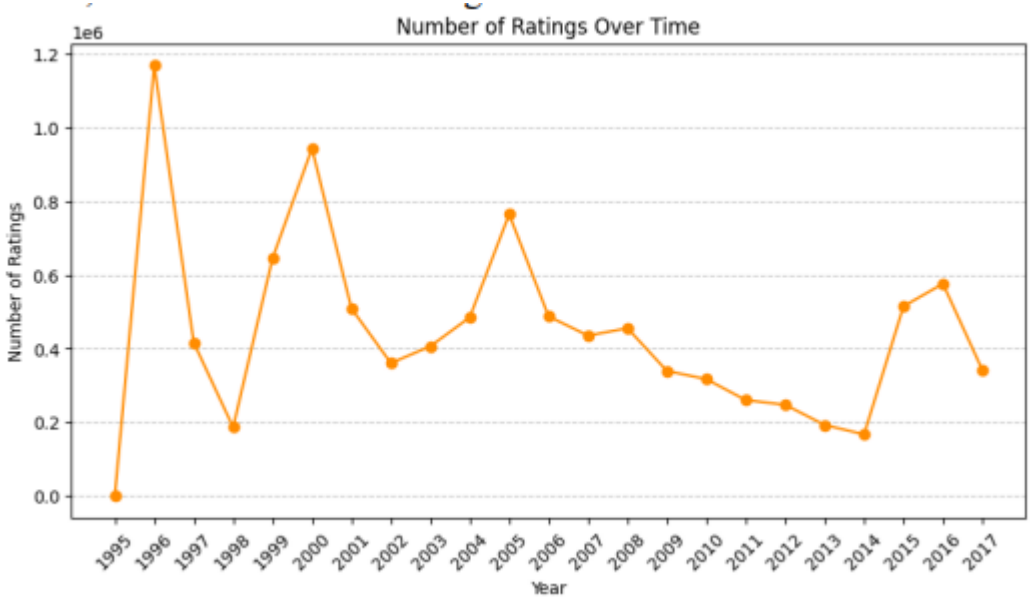
NUMERICAL VARIABLES IN MOVIES DATASET

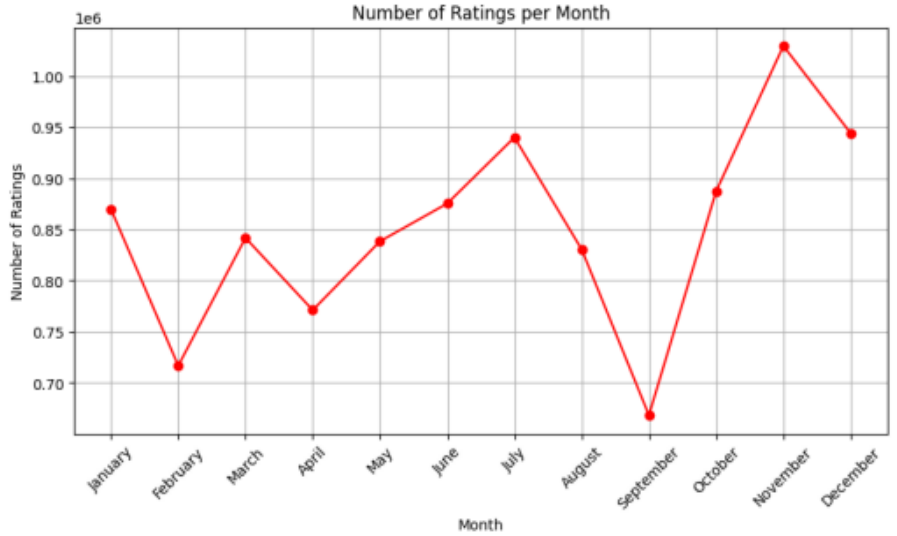
- Dataset contains a diverse range of user interactions with 120,147 unique users who have rated 7,508 unique movies
- Most-rated movies include "Terminator 3: Rise of the Machines" (72,611 ratings) and "The Million Dollar Hotel" (67,882 ratings), indicating their high engagement among users



NUMERICAL VARIABLES IN MOVIES DATASET

- Ratings experienced a decline between 2005 to 2015, possibly due to a shift in user behavior
- Monday and Tuesday received 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





Feature Engineering

Encoding

Genre	Language	Year	Actors	Director

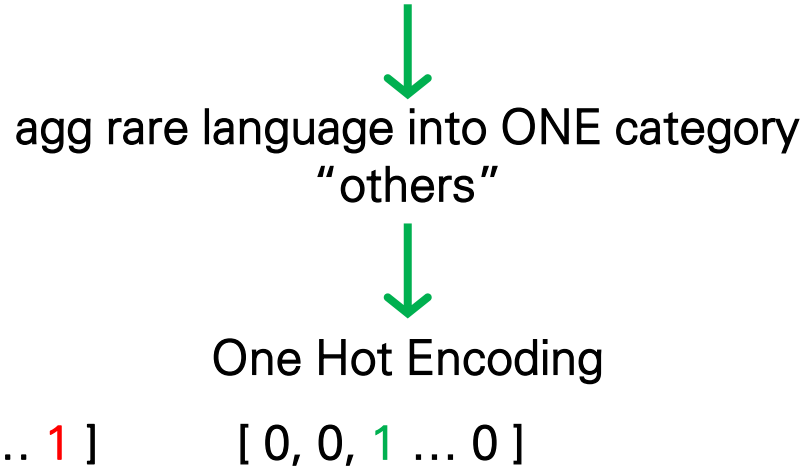
One Hot Encoding

[0, 1 ,0 … 1]

Feature Engineering

Encoding

Genre	Language	Year	Actors	Director



Feature Engineering

Encoding

Genre	Language	Year	Actors	Director

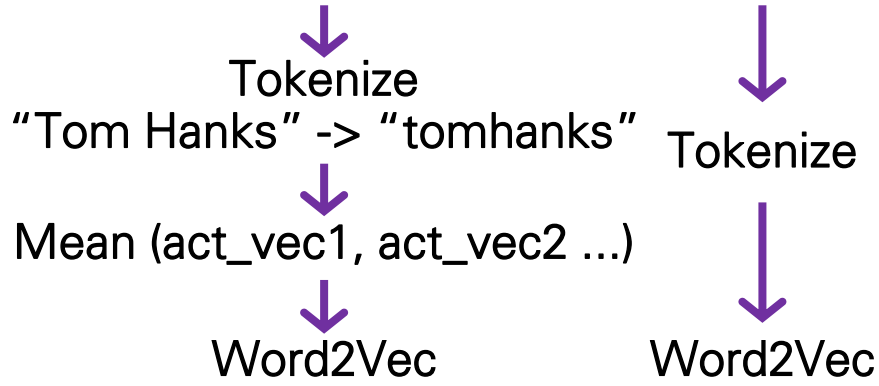
MinMaxScaler

[0, 1 ,0 … 1] [0, 0, 1 … 0] 0.826772

Feature Engineering

Encoding

Genre	Language	Year	Actors	Director

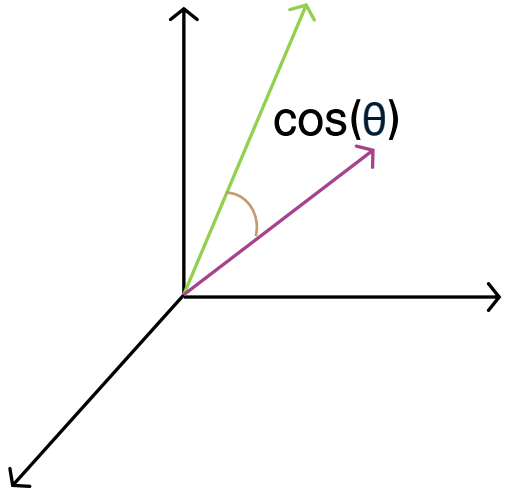


[0, 1 ,0 … 1] [0, 0, 1 … 0] 0.826772 [-0.42, 0.13, ... 0.31] [-0.27, 0.16, ... 0.32]

Feature Engineering

Precomputed Features to Speed up Training

Genre	Language	Year	Actors	Director
[0, 1 ,0 … 0]	[0, 0, 1 … 0]	0.826772	[-0.42, 0.13, … 0.31]	[-0.27, 0.16, … 0.32]
[0, 1 ,0 … 0]	[0, 0, 1 … 0]	0.826772	[-0.42, 0.13, … 0.31]	[-0.27, 0.16, … 0.32]



Content-based Similarity Matrix

$$\begin{bmatrix} 1 & 0.28 & -0.21 & \dots & -0.02 \\ 0.28 & 1 & -0.09 & \dots & -0.05 \\ -0.21 & -0.09 & 1 & \dots & 0.05 \\ \dots & \dots & \dots & \dots & \dots \\ -0.02 & -0.05 & 0.05 & \dots & 1 \end{bmatrix}$$

Shape: (7508, 7508)

Feature Engineering

Precomputed Features to Speed up Training

(Avg ≥ 4.5): 2,738

(Avg ≤ 2.5): 1,661

4.18

(Avg ≥ 4.5): 47

(Avg ≤ 2.5): 1,181

Feature Engineering

Precomputed Features to Speed up Training

(Avg ≥ 4.5): 2,738

(Avg ≤ 2.5): 1,661

(Avg ≥ 4.5): 47

NaN

(Avg ≤ 2.5): 1,181

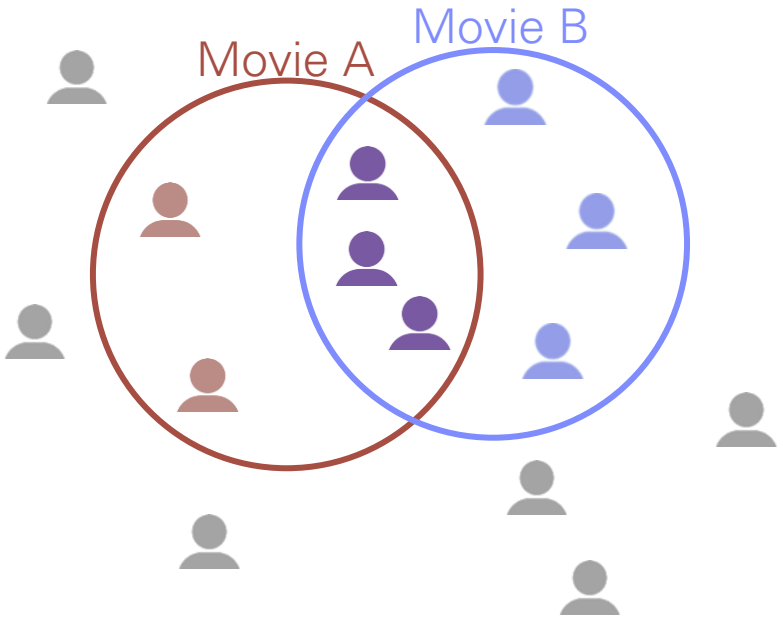
Global bias

User bias

Item bias

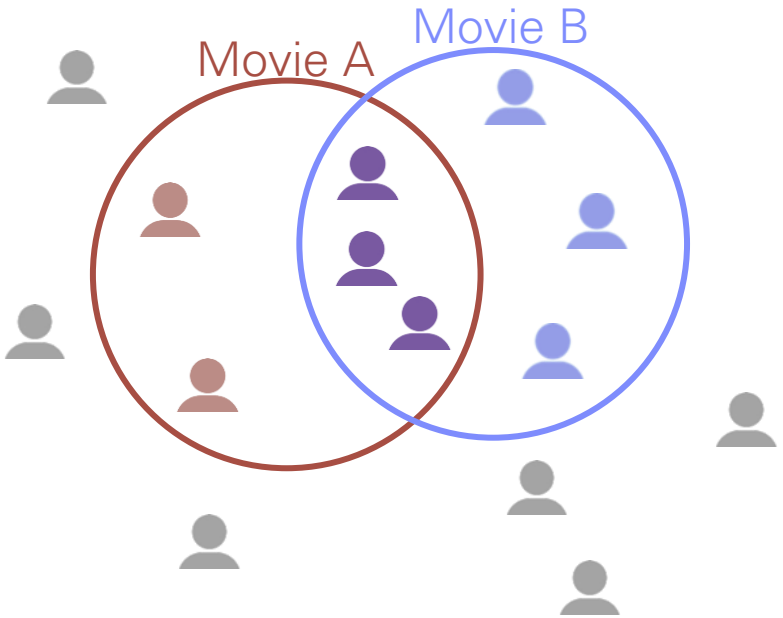
Feature Engineering

Precomputed Features to Speed up Training



Feature Engineering

Precomputed Features to Speed up Training



Shrunken-Pearson similarity matrix

$$s_{ij} \overset{\text{def}}{=} \frac{n_{ij}}{n_{ij} + \lambda_2} \rho_{ij}$$

ρ_{ij} Pearson correlation value

n_{ij} Number of items that user i and j both rated: 3

λ_2 Shrink factor: 100

Finalized Data

- Encoded Metadata

Genre	Language	Year	Actors	Director
[0, 1 ,0 … 0]	[0, 0, 1 … 0]	0.826772	[-0.42, 0.13, … 0.31]	[-0.27, 0.16, … 0.32]

- Biases

- [Similarity Matrix Metadata] [Similarity Matrix Rating]

- Train/ Test Data

Next Steps and Plans

- **Build the Models**
- **Test the Models** – Use a small dataset to check if the model works.
- **Train the Models** – Use the full dataset and fine-tune it.
- **Adjust Dataset (If Needed)** – Reduce data size if training is too slow.
- **Evaluate Results** – Check model accuracy and performance.
- **Improve Data (If Needed)** – Adjust preprocessing to get better features.
- **Repeat as Needed** – Keep improving until the model performs well.

Key Lessons Learned