MOVIE LENS

MOVIE RATINGS AND METADATA REPORT

This report presents an analysis of the dataset gotten from the movie lens video streaming website focusing on feature engineering and exploratory insights that illuminate patterns in user behaviour, genre performance, and rating dynamics.

Engineered features

Some features were gleaned from the dataset and included as new features to give new perspectives.

Dataset summary

Total unique users: 610, Total ratings: 100836, Total unique movies: 9724, Date range of ratings: 1996 – 2018, Release year range: 1902 – 2018

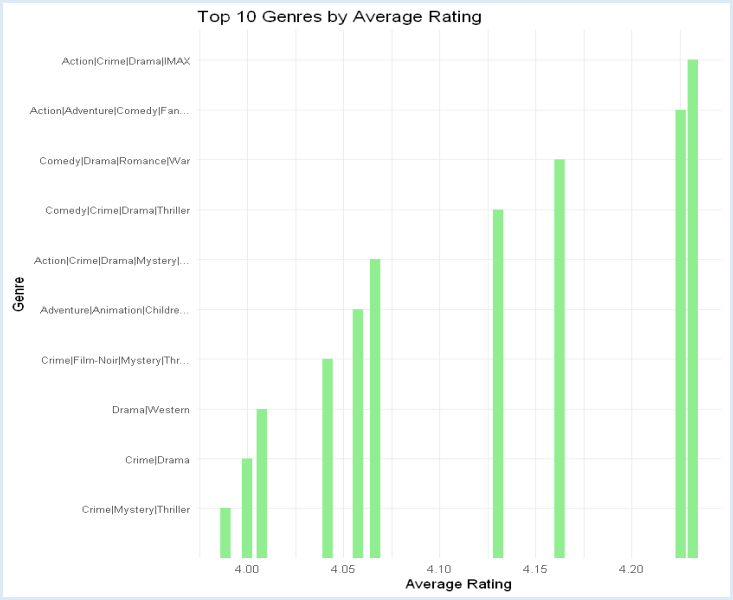
Rating stats: minimum rating: 0.5, median rating: 3.5, mean rating: 3.502, maximum: 5.0

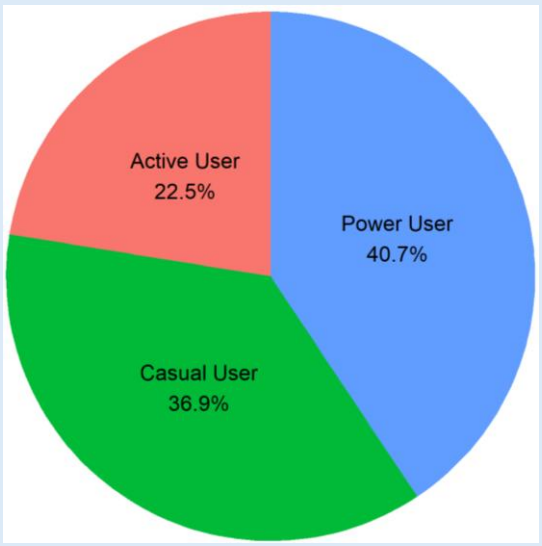
Engineered features

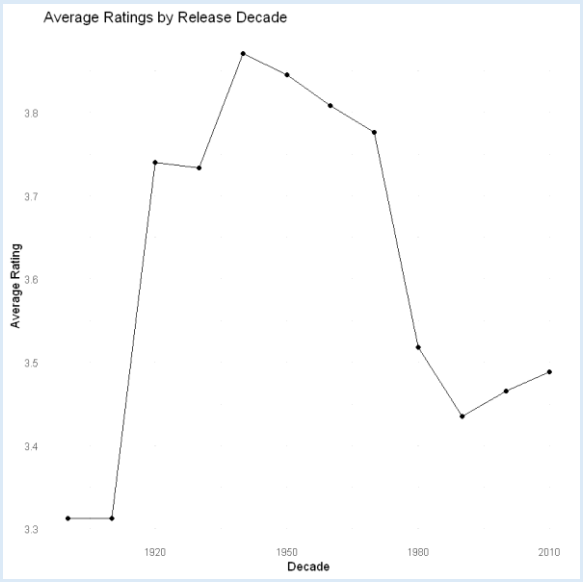
Feature name	Description	Purpose
release_year	Extracted from movie titles	Enables trend detection capture era effects (temporal preferences). Useful for time-aware recommendations
genre_count	Number of genres per movie	Captures genre diversity and complexity, proxy for 'broad appeal' vs niche; informs content-based filtering and similarity measures.
movie_avg_rating	Average rating per movie	Measures overall movie popularity and quality
genre_avg_rating	Average rating per genre	Assesses genre performance and user preferences
user_rating_count	Number of ratings per user	Identifies user engagement levels, proxy for user activity
is_classic	Boolean flag for movies released before 2000	Supports segmentation by era and nostalgia-based recommendations

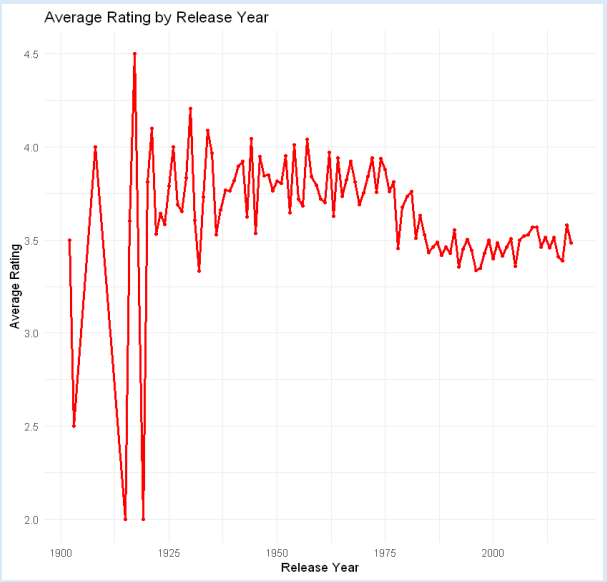
Key Insights Discovered

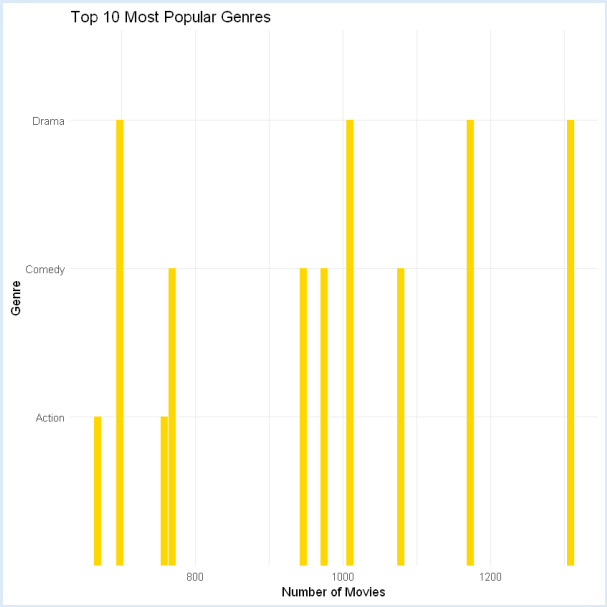
- Top Movies: Shawshank Redemption, Godfather, and Fight Club lead in ratings with substantial review counts.
- Genre Popularity: Drama, Comedy, and Action dominate in volume.
- Genre Performance: Genres like Action|Crime|Drama|IMAX and Action|Adventure|Comedy|Fantasy received the highest average ratings.
- Average Ratings per Decade reveals that movies in the 1900s are generally more liked than movies in 2000s, the 1940's were the golden era of movies (based on this dataset)
- Genre Count vs Rating: there is no definitive indication that movies with moregenres tend to receive higher ratings
- Genre Popularity: Drama, Comedy, and Action dominate in volume.
- Engagement by Movie Type: Classic movies had slightly lower engagement but similar median ratings.
- User Segmentation: Users were classified into Power(>100), Active(50-100), and Casual (<50) based on rating frequency.

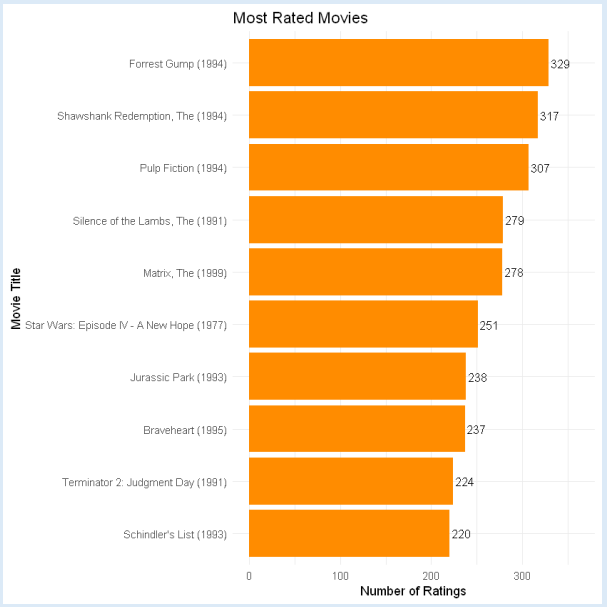


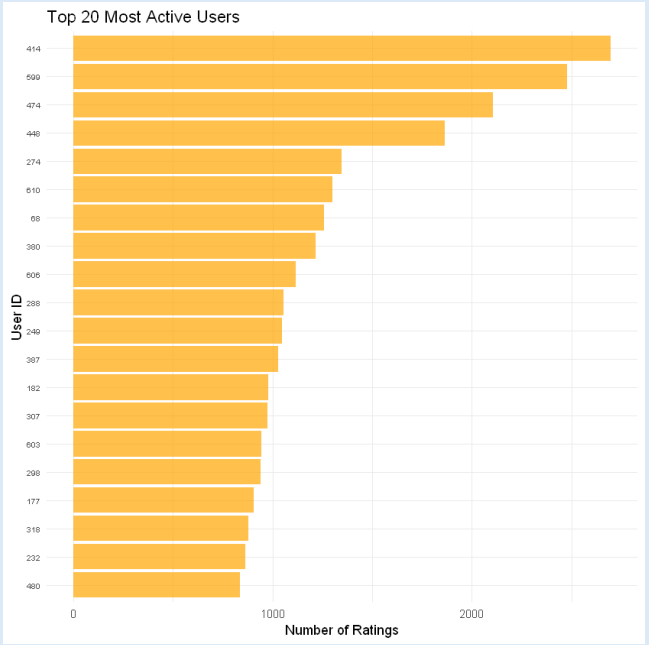


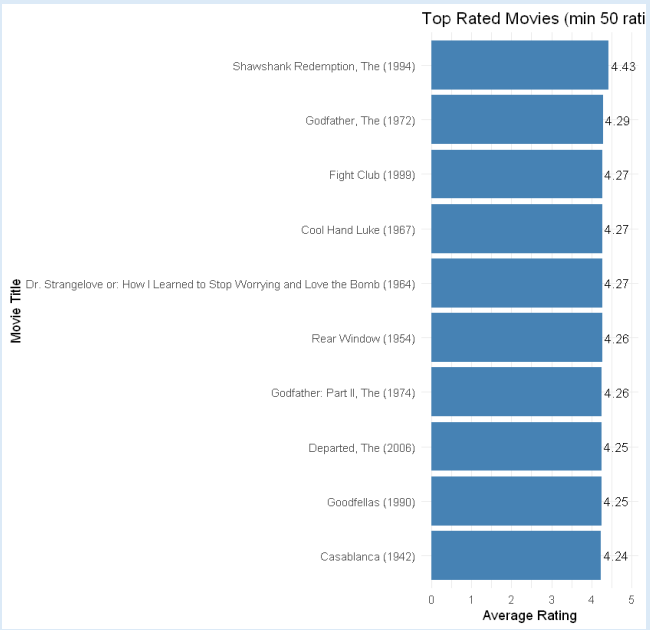

Average Ratings by Release Decade
Average Rating
3.8
3.7
3.6
3.5
3.4
3.3
1920
1950
1980
2010
Decade


Average Rating by Release Year
Average Rating
4.5
4.0
3.5
3.0
2.5
2.0
1900
1925
1950
1975
2000
Release Year


Top 10 Most Popular Genres
Genre
Drama
Comedy
Action
800
1000
1200
Number of Movies


Most Rated Movies
Movie Title
Forrest Gump (1994)
329
Shawshank Redemption, The (1994)
317
Pulp Fiction (1994)
307
Silence of the Lambs, The (1991)
279
Matrix, The (1999)
278
Star Wars: Episode IV - A New Hope (1977)
251
Jurassic Park (1993)
238
Braveheart (1995)
237
Terminator 2: Judgment Day (1991)
224
Schindler's List (1993)
220
0
100
200
300
Number of Ratings


Top 20 Most Active Users
User ID
414
599
474
448
274
610
68
380
606
288
249
387
182
307
603
298
177
318
232
480
0
1000
2000
Number of Ratings


Top Rated Movies (min 50 rati
Movie Title
Shawshank Redemption, The (1994)
4.43
Godfather, The (1972)
4.29
Fight Club (1999)
4.27
Cool Hand Luke (1967)
4.27
Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb (1964)
4.27
Rear Window (1954)
4.26
Godfather: Part II, The (1974)
4.26
Departed, The (2006)
4.25
Goodfellas (1990)
4.25
Casablanca (1942)
4.24
0
1
2
3
4
5
Average Rating

How Do The Features Support a Future Recommendation System?

- **User Profiling**: user_rating_count and rating behavior help tailor suggestions based on each user's engagement level.

- **Content-Based Filtering**: genre_count, genres, and release_year enable matching users to similar movies. Users that prefer certain genres will get recommendations of those genres.

- **Collaborative Filtering**: movie_avg_rating and user_type support clustering users with similar tastes and movie preferences from the clusters can inform recommendations for similar users.

- **Temporal Dynamics**: rating_year and release_year allow for trend-aware recommendations. Ratings often spike around holidays or after major events (e.g., Oscar wins, streaming releases). rating_year allows the system to factor in these temporal spikes and suggest trending content.

- **Cold Start Solutions**: Genre and tag-based features help recommend new or unrated movies. When a movie has no ratings yet, genre and tag-based features allow the system to:
  Compare the new movie to others with similar genres or tags and recommend it to users who have shown interest in those genres or themes.
  New users will answer onboarding questions on preferred genres, this will be used for recommendations

- **Diversity and Novelty**: Insights into genre performance and popularity help balance familiar vs. novel suggestions. By analysing genre performance, we can introduce diverse genres that are still well-rated, expanding user horizons without sacrificing quality.

Conclusion

This analysis of the MovieLens dataset provided a strong foundation for understanding user–item interactions and feature dynamics. Through feature engineering and visual exploration, we identified patterns such as consistent user rating tendencies, genre-based preferences, and temporal trends in movie popularity. These findings support the development of a robust recommendation system by enabling:

- User segmentation for personalized recommendations.
- Trend-aware filtering using temporal features like release_year and rating_year.
- Cold start solutions through genre and tag-based similarity.
- Balanced suggestions that promote both diversity and novelty.

By integrating collaborative filtering with temporal features, future models can deliver more accurate, engaging, and personalized recommendations.