MovieLens

SONYA TAHIR, JESSICA SMITH

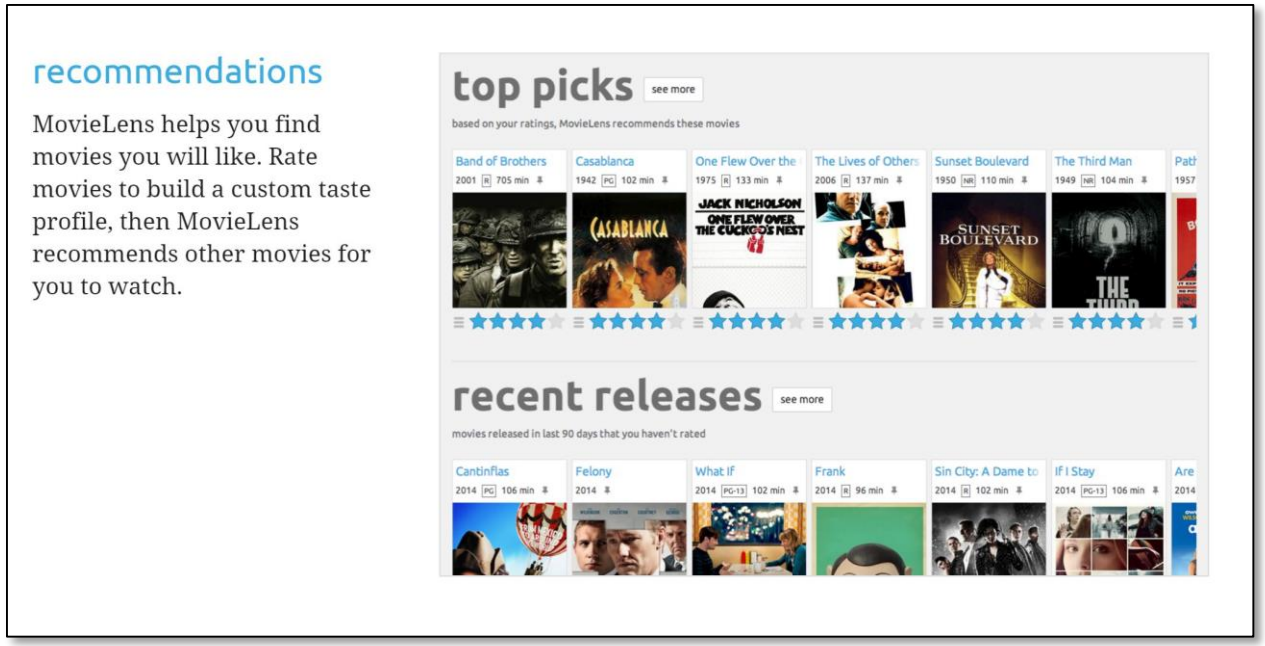
What is MovieLens?

Personalized Movie Recommendation Service

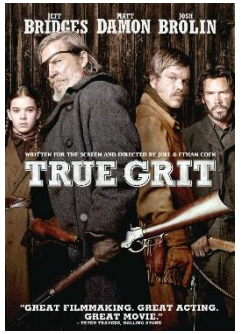
Based on User Preferences

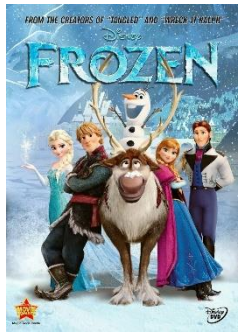
Free Service

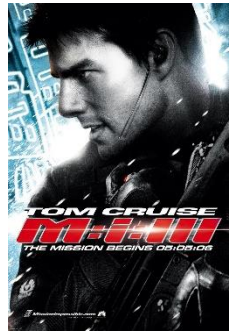
www.movielens.org

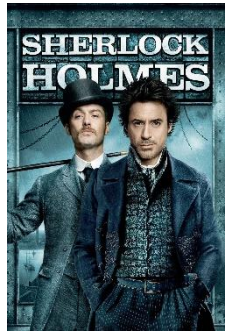


Data









User Information

- Age Range
- Gender
- Occupation
- Zip
- No unique user ID provided

Genre Fields

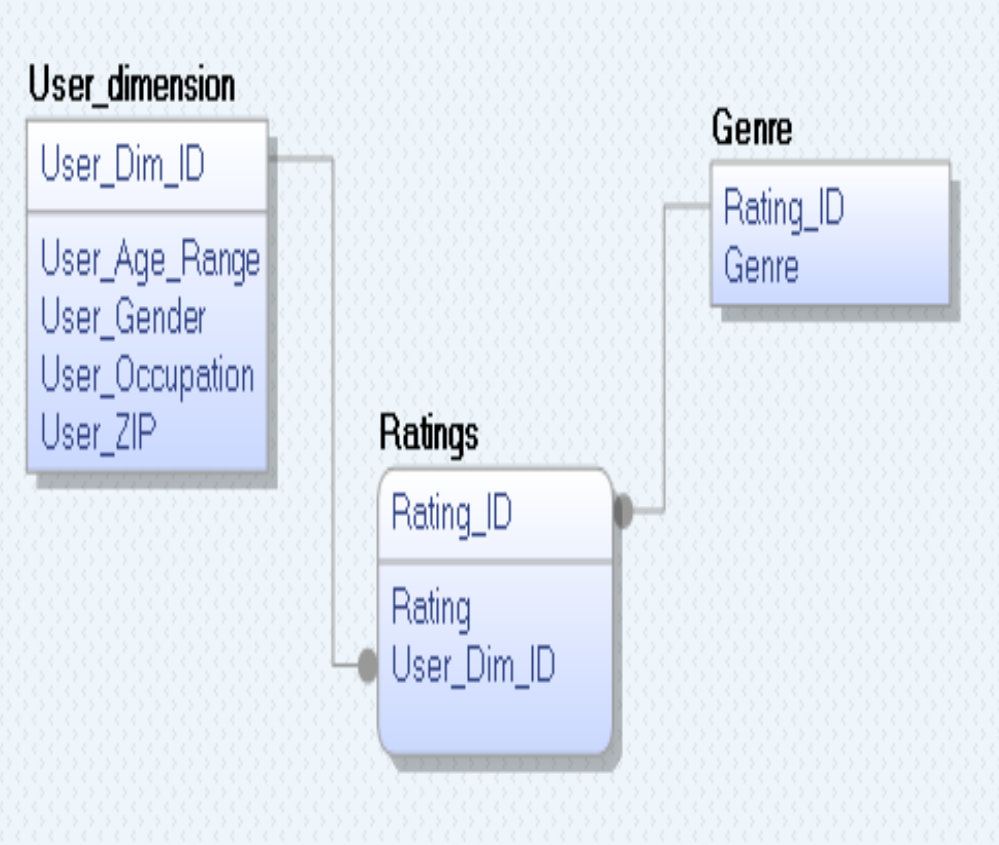
- Action, Adventure, Animation, Children, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War, Western

Rating

- Score between 1 (lowest) and 5 (highest)

Over 1 million records

Transactional Database



The primary goal was to normalize data to make it storage efficient.

User_dimension includes all user attributes.

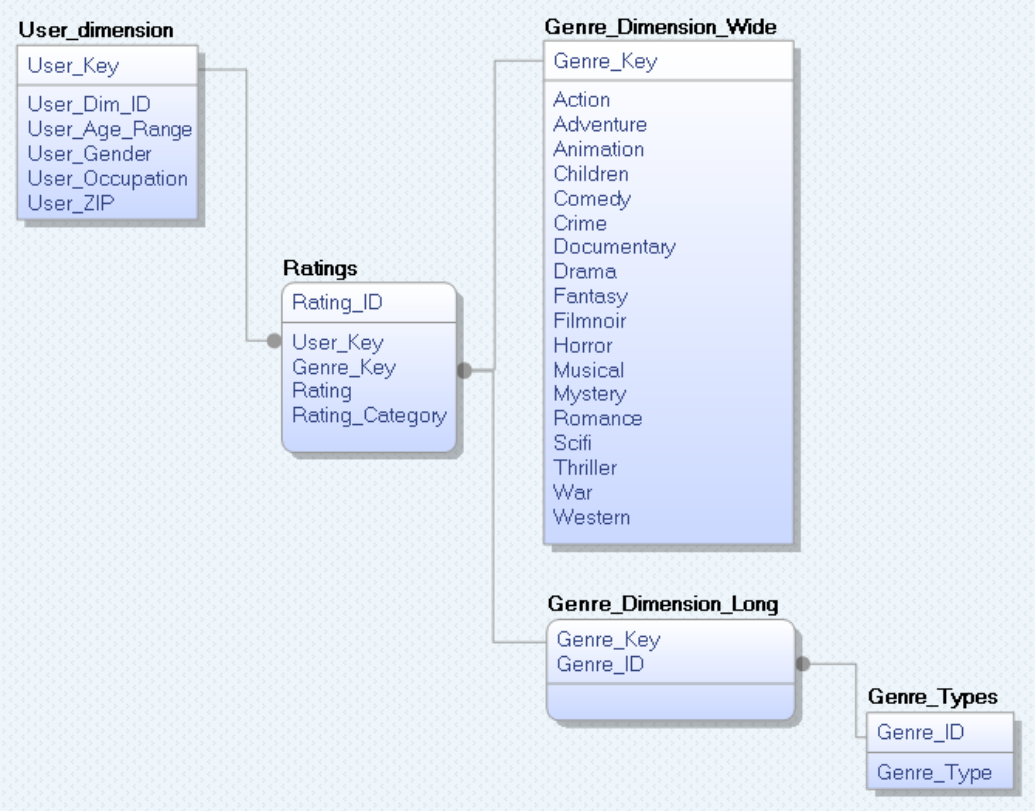
Ratings includes each unique rating.

Genre includes a list of genres associated with each rating.

Reduced the number of rows for the user information

Reduced the number of columns for the genre information

Data Warehouse



The primary goal was to create a star schema.

User dimension includes user attributes like before with an addition of a user key.

Genre is a multi-valued dimensional attribute. We created a bridge so that a single rating may be associated with as many genre fields as needed. Each genre combination has a unique key.

There is also a wide genre dimension table with true/false values to represent genre combinations.

Ratings is the fact table which includes user key, genre key, rating and a calculated field called rating category.

Genre_Dimension_Long

Genre_Key	Genre_ID
1	1
1	2
2	3

Genre_Types

Genre_ID	Genre_Type
1	Action
2	Adventure
3	Animation
4	Children
5	Comedy

Ratings

Genre_Key

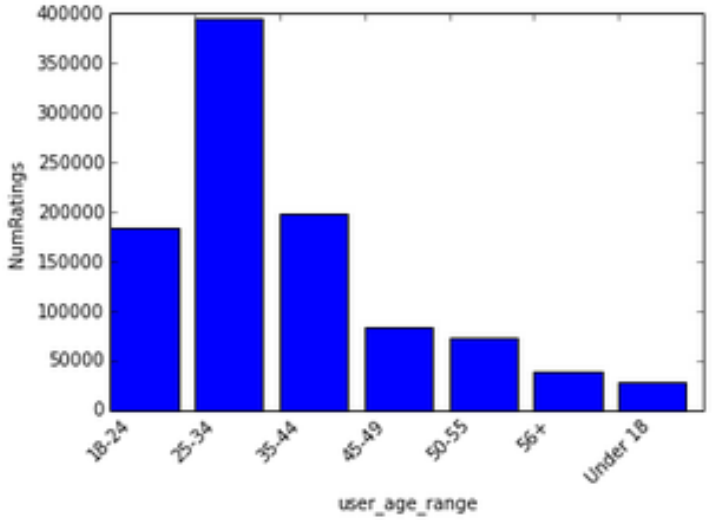
Analysts can query the genre information in long form or in wide form depending on the analytic requirement. To easily group by genre, the long dimension is useful. To easily search for specific combinations of genres, the wide dimension is appropriate.

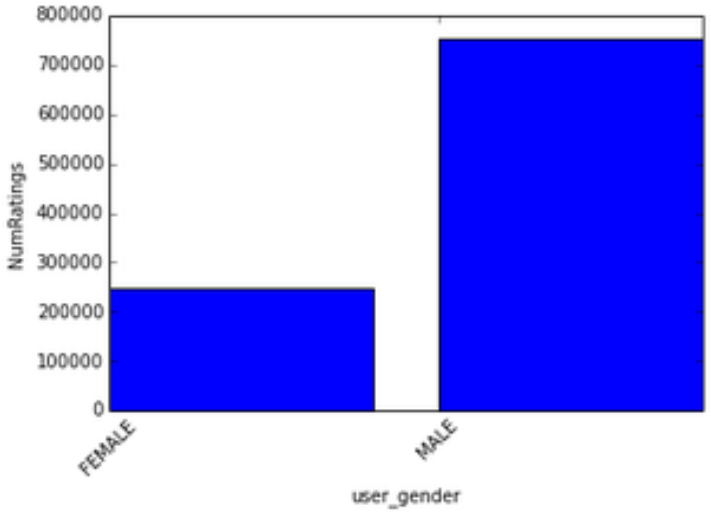
Genre_Dimension_Wide

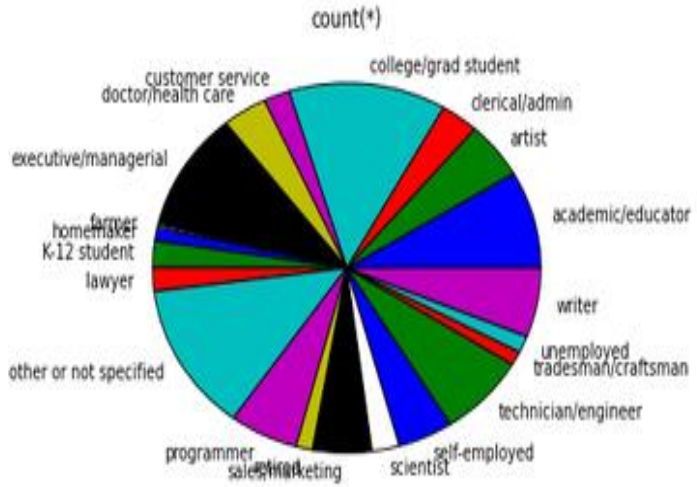
Genre_Key	Action	Adventure	Animation	Children	Comedy	Crime
1	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE
2	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE

Analysis

User



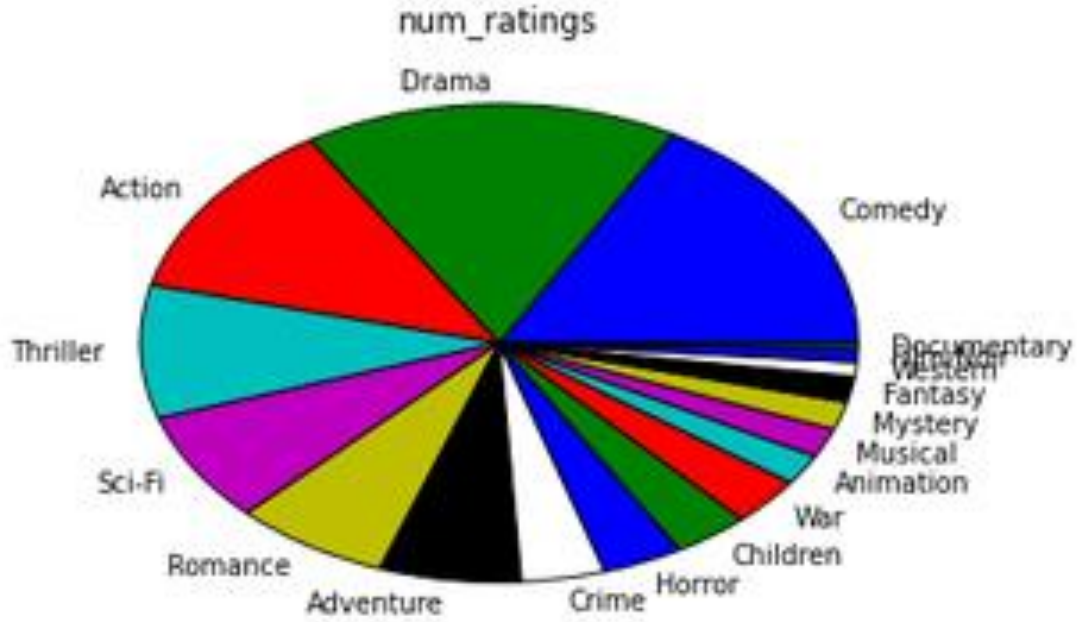


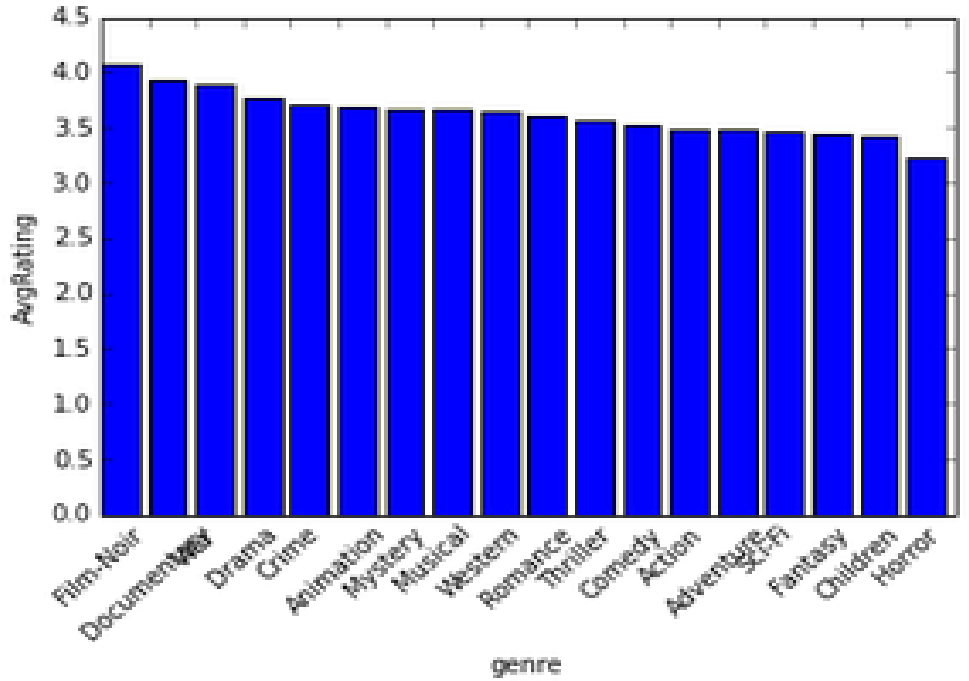


Users are primarily males aged 25-34

Primary occupations include college/grad student, academic/educator, executive/managerial

Genres

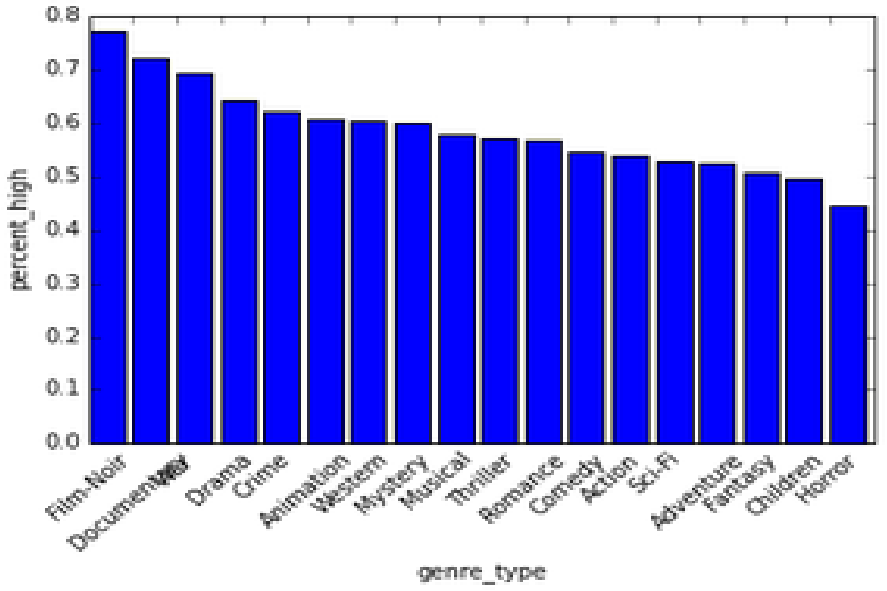




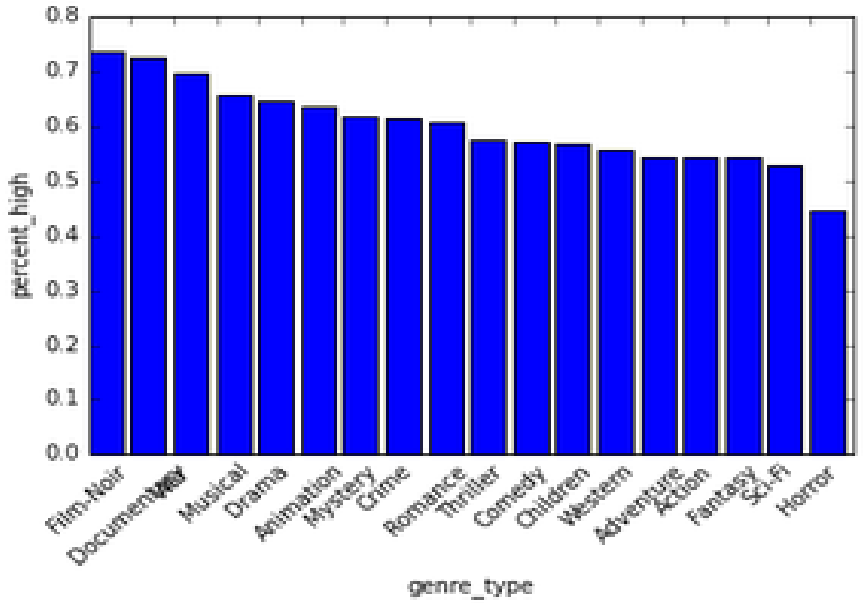
Drama and Comedy have the most movie ratings but Film-Noir and Documentary have the highest average rating.

Gender Preferences

Male



Female



Uses the rating category to determine what portion of ratings are “High” (4 or 5).

Both men and women tend to rate Film-Noir, Documentary and War as High. Men are more likely to rate Crime High whereas women are more likely to rate Musical High.

Genre Type by Age

Action and Adventure

user_age_range	avg_rating	num_ratings
18-24	3.4539	18064
25-34	3.4671	36569
35-44	3.5249	17701
45-49	3.5391	6876
50-55	3.6253	5776
56+	3.6029	2795
Under 18	3.5311	2542

Action and Drama

user_age_range	avg_rating	num_ratings
18-24	3.7787	8666
25-34	3.7118	19693
35-44	3.7809	9787
45-49	3.7494	3986
50-55	3.8498	3734
56+	3.8972	1995
Under 18	3.8284	1055

Do Action movies receive different ratings when they are Action/Adventure vs Action/Drama?

Action/Drama movies receive higher ratings across all age groups.

Conclusion

Limitation

- We initially planned to mimic MovieLens' predictive analysis. However, the lack of unique user IDs and unique movie titles rendered this impossible.

Challenge

- When working with such a large dataset, it was critical to use joins to perform bulk inserts, updates and deletes. Otherwise these tasks took far too long to complete (if at all).
- Dealing with genre classifications was a challenge because a movie can have any combination of 18 values.


title
The End
date
scene
take