Analysis of the MovieLens

CIT646: Data Mining Assignment 1

Team Members:

Salma Hisham Mohamed, 231000533

Abdelrahman khaled Ibrahim, 211001429

Norhan Mohamed Swar, 231000486

Supervised by:

Dr.Tamer Arafa

This report Explored user demographics and behaviors through Exploratory Data Analysis (EDA) and implemented MapReduce tasks for insightful data preprocessing in a movie dataset in addition to PageRank for Movie Popularity Analysis and Locality Sensitive Hashing for Similar Movie Discovery

1. Data Loading

- Loaded user, movie, and ratings data from respective .dat files into Pandas DataFrames.
- Assigned appropriate column names to DataFrames based on the data README file.
- Merged the three DataFrames into a single DataFrame (merged_df).
- Dataset: https://grouplens.org/datasets/movielens/1m/

2. Exploratory Data Analysis (EDA)

- Checked for Null values and found none.
- Checked for duplicates in merged_df and found none.
- Explored demographics:
  - Most common occupation: Programming.
  - Age range with the highest movie views: 20 to 30.
  - Gender distribution: Skewed towards men.
  - Top-viewed genre: Comedy.

3. Data Preprocessing Using Map Reduce Jobs

- Top-Rated Movies:
  - Implemented MapReduce to calculate the average rating for each movie.
  - Merged results with movie details, selected top 10 movies with over 500 ratings.

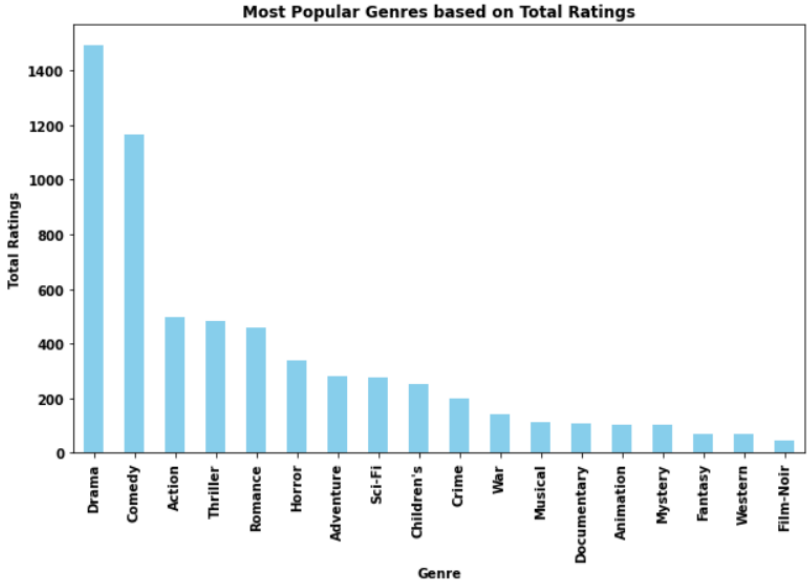
```
Top 10 Movies by Average Rating > 500
```

	MovieID	AvgRating	TotalRatings	Title	Genres
1092	2019	4.560510	628	Seven Samurai (The Magnificent Seven) (Shichin...	Action|Drama
167	318	4.554558	2227	Shawshank Redemption, The (1994)	Drama
669	858	4.524966	2223	Godfather, The (1972)	Action|Crime|Drama
29	745	4.520548	657	Close Shave, A (1995)	Animation|Comedy|Thriller
259	50	4.517106	1783	Usual Suspects, The (1995)	Crime|Thriller
23	527	4.510417	2304	Schindler's List (1993)	Drama|War
535	1148	4.507937	882	Wrong Trousers, The (1993)	Animation|Comedy
127	1198	4.477725	2514	Raiders of the Lost Ark (1981)	Action|Adventure
629	904	4.476190	1050	Rear Window (1954)	Mystery|Thriller
44	260	4.453694	2991	Star Wars: Episode IV - A New Hope (1977)	Action|Adventure|Fantasy|Sci-Fi

Insight: Seven Samurai, Shawshank Redemption, and The Godfather are the top three movies according to the highest AVG rating, which received more than 500 ratings.

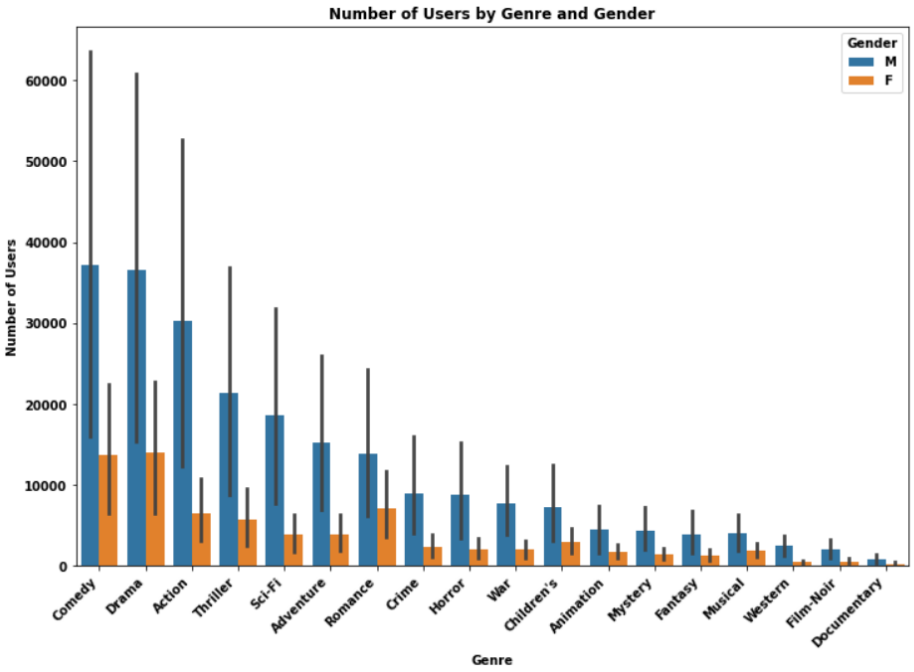
- Most Popular Genres:
  - Utilized MapReduce to count the occurrences of each movie genre.
  - Plotted the most popular genres based on total ratings.

	Genre	TotalRatings
0	Drama	1493
5	Comedy	1163
6	Action	495
11	Thriller	485
4	Romance	459
15	Horror	339
7	Adventure	281
9	Sci-Fi	274
2	Children's	250
12	Crime	201
10	War	141
3	Musical	113
17	Documentary	110
1	Animation	105
13	Mystery	104
8	Fantasy	68
14	Western	67
16	Film-Noir	44



Insight: Comedy was the first EDA genre based on the average rating, but according on the total rating here, drama is the most popular genre overall.

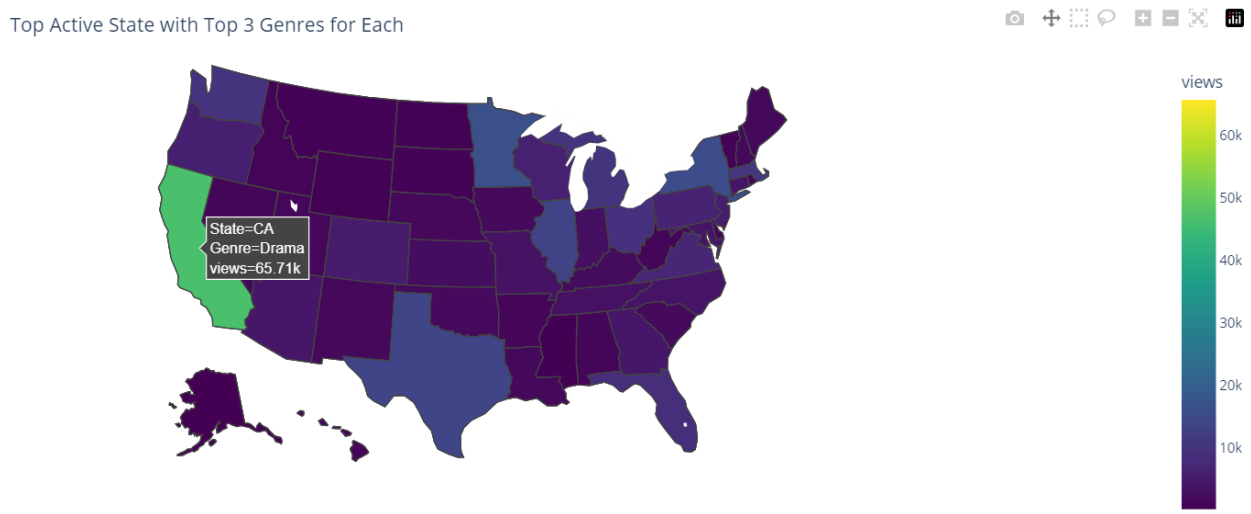
- Gender Preferences:
  - Applied MapReduce to analyze user preferences by gender and genre.
  - Visualized the number of users by genre and gender.



Insight: This graph indicates that males like comedy while women tend to appreciate drama the most.

- User Demographics Distribution:
  - Extracted geo coordinates from zip codes and added them to the user DataFrame.
  - Performed MapReduce to analyze user activity by state and genre.
  - Visualized the top 3 genres for each state using a choropleth map.

	UserID	Gender	Age	Occupation	City	State	Latitude	Longitude
0	1	F	1	10	Royal Oak	MI	42.488735	-83.13752
1	2	M	56	16	Marrero	LA	29.869283	-90.10933
2	3	M	25	15	Saint Paul	MN	44.989065	-93.10666
3	4	M	45	7	Newtonville	MA	42.352996	-71.20907
4	5	M	25	20	Minneapolis	MN	44.971965	-93.23588



Insight: Here we can receive the amount of participants from each state with their preferred genre, as demonstrated in California, which appears to be the top among Drama.

---

4. PageRank for Movie Popularity Analysis

Methodology:

1) Graph Construction:

A graph is constructed , where nodes represent movies.
Edges between nodes are created based on the similarity of average ratings, forming a weighted graph.

2) PageRank Algorithm:

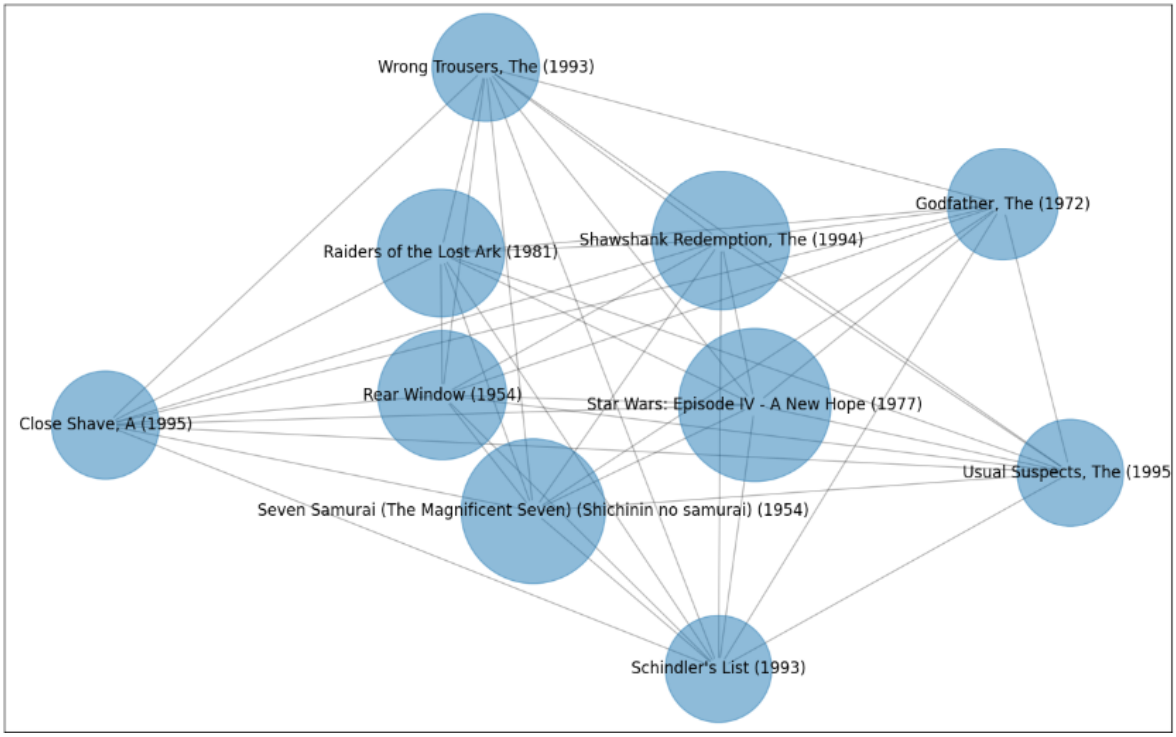
The PageRank algorithm from is applied to calculate the importance of each movie node.
The algorithm considers edge weights, reflecting the similarity of average ratings.

3) Visualization:

Positions for nodes in the graph are determined using the spring layout algorithm.
Node sizes are assigned based on PageRank scores, providing a visual representation of movie importance.

Results:

The implementation successfully analyzes movie popularity based on user ratings and interactions using the modified PageRank algorithm. The visualization visually represents the importance of each movie within the network, with larger nodes indicating higher PageRank scores. This approach allows for the identification of movies with greater user engageme, providing valuable insights into movie popularity within the dataset.



5. Locality Sensitive Hashing

Methodology:

The implementation consists of the following steps:

- Step 1: Shingling
  - The ratings of each movie are converted into a set of shingles. A shingle is a sequence of n consecutive ratings.
  - The `shingling` function takes a list of ratings and returns a set of shingles.
- Step 2: MinHashing
  - The set of shingles for each movie is used to create a MinHash object, which represents the high-dimensional vector for that movie.
  - The `minhashing` function takes a set of shingles and a specified number of permutations and returns a MinHash object.
- Step 3: LSH Clustering
  - The MinHash signatures of movies are clustered using the LSH algorithm.
  - The `lsh_clustering` function performs clustering using LSH on a DataFrame of signatures.
  - Similar movies are grouped together based on their MinHash signatures.

Results:

By implementing this basic LSH algorithm, we can cluster movies with similar rating patterns. The resulting clusters group together movies that have similar user rating preferences, allowing us to discover similar movies and identify patterns in user preferences.

```
Cluster 1650:
      MovieID  AvgRating                        Title
1649     2474    0.32953  Color of Money, The (1986)

Cluster 1651:
      MovieID  AvgRating                         Title
1650     3852   0.680943     Tao of Steve, The (2000)
2233     3093   0.680943  McCabe & Mrs. Miller (1971)

Cluster 1652:
      MovieID  AvgRating                    Title
1651     1172   1.558737  Cinema Paradiso (1988)

Cluster 1653:
      MovieID  AvgRating          Title
1652     3788   0.847124  Blowup (1966)

Cluster 1655:
      MovieID  AvgRating           Title
1654      233   0.555595  Exotica (1994)
```