

Case Study: Comprehensive Data Analytics on MovieLens Dataset

Problem Statement

You are tasked with building a robust data pipeline to process, analyze, and generate actionable insights from the MovieLens dataset. The focus is on performing advanced data operations such as joins, aggregations, windowing, and other critical transformations, while optimizing the pipeline for performance and scalability. The case study will explore in-depth user behavior, movie trends, and genre analysis through sophisticated data operations.

Dataset Description

You will use the MovieLens 20M Dataset, which contains:

- ratings.csv: Contains user ratings for movies.
- movies.csv: Contains movie titles and genres.
- tags.csv: Contains user-assigned tags for movies.
- links.csv: (Optional) Maps movies to external metadata sources.

1. Data Loading and Preprocessing

Load the MovieLens dataset into Spark DataFrames and perform essential preprocessing.

```
+-----+-----+-----+
|movieId|      title|      genres|
+-----+-----+-----+
| 1| Toy Story (1995)|Adventure|Animati...
| 2| Jumanji (1995)|Adventure|Childre...
| 3|Grumpier Old Men ...| Comedy|Romance| |
| 4|Waiting to Exhale...|Comedy|Drama|Romance|
| 5|Father of the Bri...| Comedy|
| 6| Heat (1995)|Action|Crime|Thrill...
| 7| Sabrina (1995)| Comedy|Romance|
| 8| Tom and Huck (1995)| Adventure|Children|
| 9| Sudden Death (1995)| Action|
|10| GoldenEye (1995)|Action|Adventure|...
|11|American Presiden...|Comedy|Drama|Romance|
|12|Dracula: Dead and...| Comedy|Horror|
|13| Balto (1995)|Adventure|Animati...
|14| Nixon (1995)| Drama|
|15|Cutthroat Island ...|Action|Adventure|...
|16| Casino (1995)| Crime|Drama|
|17|Sense and Sensibi...| Drama|Romance|
|18| Four Rooms (1995)| Comedy|
|19|Ace Ventura: When...| Comedy|
|20| Money Train (1995)|Action|Comedy|Cri...
+-----+-----+-----+
only showing top 20 rows
```

```
+-----+-----+-----+
|userId|movieId|rating|      timestamp|
+-----+-----+-----+
| 1|    2| 3.5|2005-04-02 23:53:47|
| 1|   29| 3.5|2005-04-02 23:31:16|
| 1|   32| 3.5|2005-04-02 23:33:39|
| 1|   47| 3.5|2005-04-02 23:32:07|
| 1|   50| 3.5|2005-04-02 23:29:40|
| 1|  112| 3.5|2004-09-10 03:09:00|
| 1|  151| 4|2004-09-10 03:08:54|
| 1|  223| 4|2005-04-02 23:46:13|
| 1|  253| 4|2005-04-02 23:35:40|
| 1|  260| 4|2005-04-02 23:33:46|
| 1|  293| 4|2005-04-02 23:31:43|
| 1|  296| 4|2005-04-02 23:32:47|
| 1|  318| 4|2005-04-02 23:33:18|
| 1|  337| 3.5|2004-09-10 03:08:29|
| 1|  367| 3.5|2005-04-02 23:53:00|
| 1|  541| 4|2005-04-02 23:30:03|
| 1|  589| 3.5|2005-04-02 23:45:57|
| 1|  593| 3.5|2005-04-02 23:31:01|
| 1|  653| 3|2004-09-10 03:08:11|
| 1|  919| 3.5|2004-09-10 03:07:01|
+-----+-----+-----+
only showing top 20 rows
```

```
+-----+-----+-----+
|userId|movieId|      tag|      timestamp|
+-----+-----+-----+
|   18|  4141|Mark Waters|2009-04-24 18:19:40|
|   65|  208|dark hero|2013-05-10 01:41:18|
|   65|  353|dark hero|2013-05-10 01:41:19|
|   65|  521|noir thriller|2013-05-10 01:39:43|
|   65|  592|dark hero|2013-05-10 01:41:18|
|   65|  668|bollywood|2013-05-10 01:37:56|
|   65|  898|screwball comedy|2013-05-10 01:42:40|
|   65| 1248|noir thriller|2013-05-10 01:39:43|
|   65| 1391|mars|2013-05-10 01:40:55|
|   65| 1617|neo-noir|2013-05-10 01:43:37|
|   65| 1694|jesus|2013-05-10 01:38:45|
|   65| 1783|noir thriller|2013-05-10 01:39:43|
|   65| 2022|jesus|2013-05-10 01:38:45|
|   65| 2193|dragon|2013-05-10 02:01:54|
|   65| 2353|conspiracy theory|2013-05-10 02:01:06|
|   65| 2662|mars|2013-05-10 01:40:55|
|   65| 2726|noir thriller|2013-05-10 01:39:43|
|   65| 2840|jesus|2013-05-10 01:38:45|
|   65| 3052|jesus|2013-05-10 01:38:46|
|   65| 5135|bollywood|2013-05-10 01:37:56|
+-----+-----+-----+
only showing top 20 rows
```

```
+-----+-----+-----+
|movieId|imdbId|tmdbId|
+-----+-----+-----+
|     1|114709|  862|
|     2|113497|  8844|
|     3|113228| 15602|
|     4|114885| 31357|
|     5|113041| 11862|
|     6|113277|  949|
|     7|114319| 11860|
|     8|112302| 45325|
|     9|114576|  9091|
|    10|113189|  710|
|    11|112346|  9087|
|    12|112896| 12110|
|    13|112453| 21032|
|    14|113987| 10858|
|    15|112760|  1408|
|    16|112641|  524|
|    17|114388|  4584|
|    18|113101|  5|
|    19|112281|  9273|
|    20|113845| 11517|
+-----+-----+-----+
only showing top 20 rows
```

Handle missing or invalid values in all datasets.

Normalize genres in movies.csv to create a one-movie-per-genre structure.

```
+-----+-----+-----+
|movieId|      title|  genres|
+-----+-----+-----+
| 1| Toy Story (1995)|Adventure|
| 1| Toy Story (1995)|Animation|
| 1| Toy Story (1995)| Children|
| 1| Toy Story (1995)| Comedy|
| 1| Toy Story (1995)| Fantasy|
| 2| Jumanji (1995)|Adventure|
| 2| Jumanji (1995)| Children|
| 2| Jumanji (1995)| Fantasy|
| 3|Grumpier Old Men ...| Comedy|
| 3|Grumpier Old Men ...| Romance|
| 4|Waiting to Exhale...| Comedy|
| 4|Waiting to Exhale...| Drama|
| 4|Waiting to Exhale...| Romance|
| 5|Father of the Bri...| Comedy|
| 6|      Heat (1995)| Action|
| 6|      Heat (1995)| Crime|
| 6|      Heat (1995)| Thriller|
| 7|     Sabrina (1995)| Comedy|
| 7|     Sabrina (1995)| Romance|
| 8| Tom and Huck (1995)|Adventure|
+-----+-----+-----+
only showing top 20 rows
```

2. Data Enrichment and Joins

Enrich the ratings data with movie metadata and user-generated tags.

- Perform an inner join between ratings.csv and movies.csv on movieId.
- Left join the result with tags.csv on movieId and userId.
- Use links.csv to associate movies with external metadata, if applicable.

```

compilation: disabled (not enough contiguous free space left)
+-----+-----+-----+-----+-----+-----+
|userId|movieId|rating|ratingsDF_timestamp|      title| genres|      tag|  tagsDF_timestamp| imdbId|tmdbId|
+-----+-----+-----+-----+-----+-----+
| 51266|    10|     4|2004-08-30 09:41:59| GoldenEye (1995)| Thriller| Bond|2006-04-19 09:57:56| 113189|   710|
| 51266|    10|     4|2004-08-30 09:41:59| GoldenEye (1995)| Adventure| Bond|2006-04-19 09:57:56| 113189|   710|
| 51266|    10|     4|2004-08-30 09:41:59| GoldenEye (1995)| Action| Bond|2006-04-19 09:57:56| 113189|   710|
| 66635| 100163| 3.5|2013-04-18 03:25:02|Hansel & Gretel: ...|   IMAX| action|2013-04-18 03:26:02|1428538| 60304|
| 66635| 100163| 3.5|2013-04-18 03:25:02|Hansel & Gretel: ...|   IMAX| Gemma Arterton|2013-04-18 03:26:02|1428538| 60304|
| 66635| 100163| 3.5|2013-04-18 03:25:02|Hansel & Gretel: ...|   IMAX| Jeremy Renner|2013-04-18 03:26:02|1428538| 60304|
| 66635| 100163| 3.5|2013-04-18 03:25:02|Hansel & Gretel: ...|   IMAX| steampunk|2013-04-18 03:26:02|1428538| 60304|
| 66635| 100163| 3.5|2013-04-18 03:25:02|Hansel & Gretel: ...|   IMAX| Witches|2013-04-18 03:26:02|1428538| 60304|
| 66635| 100163| 3.5|2013-04-18 03:25:02|Hansel & Gretel: ...| Horror| action|2013-04-18 03:26:02|1428538| 60304|
| 66635| 100163| 3.5|2013-04-18 03:25:02|Hansel & Gretel: ...| Horror| Gemma Arterton|2013-04-18 03:26:02|1428538| 60304|
| 66635| 100163| 3.5|2013-04-18 03:25:02|Hansel & Gretel: ...| Horror| Jeremy Renner|2013-04-18 03:26:02|1428538| 60304|
| 66635| 100163| 3.5|2013-04-18 03:25:02|Hansel & Gretel: ...| Horror| steampunk|2013-04-18 03:26:02|1428538| 60304|
| 66635| 100163| 3.5|2013-04-18 03:25:02|Hansel & Gretel: ...| Horror| Witches|2013-04-18 03:26:02|1428538| 60304|
| 66635| 100163| 3.5|2013-04-18 03:25:02|Hansel & Gretel: ...| Fantasy| action|2013-04-18 03:26:02|1428538| 60304|
| 66635| 100163| 3.5|2013-04-18 03:25:02|Hansel & Gretel: ...| Fantasy| Gemma Arterton|2013-04-18 03:26:02|1428538| 60304|
| 66635| 100163| 3.5|2013-04-18 03:25:02|Hansel & Gretel: ...| Fantasy| Jeremy Renner|2013-04-18 03:26:02|1428538| 60304|
| 66635| 100163| 3.5|2013-04-18 03:25:02|Hansel & Gretel: ...| Fantasy| steampunk|2013-04-18 03:26:02|1428538| 60304|
| 66635| 100163| 3.5|2013-04-18 03:25:02|Hansel & Gretel: ...| Fantasy| Witches|2013-04-18 03:26:02|1428538| 60304|
| 66635| 100163| 3.5|2013-04-18 03:25:02|Hansel & Gretel: ...| Action| action|2013-04-18 03:26:02|1428538| 60304|
| 66635| 100163| 3.5|2013-04-18 03:25:02|Hansel & Gretel: ...| Action| Gemma Arterton|2013-04-18 03:26:02|1428538| 60304|
+-----+-----+-----+-----+-----+-----+
only showing top 20 rows

```

Used Caching for Memory Management.

3. Advanced Aggregations

Compute in-depth metrics to understand user behavior, movie popularity, and genre trends.

- Calculate average rating, total number of ratings, and rating variance for each movie.

```
- Calculate average rating, total number of ratings, and rating variance for each movie.

+-----+-----+-----+-----+
|movieId| average_rating|total_ratings| rating_variance|          title|      genres|
+-----+-----+-----+-----+
| 1090| 3.94311377245509|       334| 0.7460034885184588|    Platoon (1986)| Drama|War| |
| 2162|           3.75|       18| 0.06617647058823528|NeverEnding Story...|Adventure|Childre...|
| 296| 4.457757296466974|       7812| 0.6702163497419152| Pulp Fiction (1994)|Comedy|Crime|Dram...|
| 4032|           3.5|        1| null|Everlasting Piece...|      Comedy|
| 6731| 2.987741935483871|       186| 0.9638186573670443|Day of the Dead (...|Horror|Sci-Fi|Thr...|
| 89864| 4.12807881773399|       406| 0.405775345131665|      50/50 (2011)|      Comedy|Drama|
| 1372| 3.9779411764705883|       204| 0.8381809620399883|Star Trek VI: The...|Action|Mystery|Sc...|
| 3826| 2.2127659574468086|       141| 0.9829787234042552| Hollow Man (2000)|Horror|Sci-Fi|Thr...|
| 58559| 4.446294307196563|       3724| 0.7293201346868955|Dark Knight, The ...|Action|Crime|Dram...|
| 8092|           3.0|        15| 0.0|Frankenstein Unbo...| Drama|Horror|Sci-Fi|
| 90439|3.9541666666666666|       240| 0.5271792189679221| Margin Call (2011)|      Drama|Thriller|
| 2696|           4.75|        22| 0.16071428571428564|Dinner Game, The ...|      Comedy|
| 2700| 4.242647058823529|       612| 0.8776054202368341|South Park: Bigge...|Animation|Comedy|...|
| 3949| 4.246062992125984|       762| 0.8963379618612065|Requiem for a Dre...|      Drama|
| 919| 4.2835820895522385|       1072| 0.4591886505846118|Wizard of Oz, The...|Adventure|Childre...|
| 101577|           3.6875|        24| 0.3220108695652172| Host, The (2013)|Action|Adventure|...|
| 1953| 3.8055555555555554|       162| 0.26475155279503104|French Connection...|Action|Crime|Thr...|
| 2550| 4.438461538461539|       130| 0.5233154442456768|Haunting, The (1963)|      Horror|Thriller|
| 32657| 4.818181818181818|       22| 0.15584415584415587|Man Who Planted T...|Animation|Drama|
| 3993| 4.262711864406779|       118| 0.430392582934956| Quills (2000)|      Drama|Romance|
+-----+-----+-----+-----+
only showing top 20 rows
```

- Identify the top 10 movies with the highest average rating (minimum 50 ratings).

```
- Identify the top 10 movies with the highest average rating (minimum 50 ratings).

+-----+-----+-----+
|movieId|          title| average_rating|total_ratings|
+-----+-----+-----+
| 40033|Adventures of Pri...|       5.0|       64|
|  8580|Into the Woods (1...|       5.0|       56|
| 33801|Godzilla: Final W...|       5.0|       88|
|  4160|Widow of St. Pier...|       4.98|       50|
|  2585|Lovers of the Arc...| 4.894736842105263|       76|
|  7505|Kingdom, The (Rig...| 4.887096774193548|       93|
| 101243|Klip (Clip) (2012)|       4.875|       72|
|  97957|Excision (2012)| 4.866666666666666|       60|
|  45183|Protector, The (a...| 4.863636363636363|       88|
| 26159|Tokyo Drifter (Tô...|       4.825|       60|
+-----+-----+-----+
```

-Aggregate data by genre to compute total ratings, average ratings, and most popular genre.

```
aggregate data by genre to compute total ratings, average ratings, and most popular genre.
-----+-----+-----+
genres|total_ratings|  average_rating|total_users|
-----+-----+-----+-----+
Drama |      200632|3.890947107141433|      200632|
-----+-----+-----+
```

Aggregate all the data and showing only the popular genre data by using limit operator.

- Calculate average rating per year using the parsed timestamp column.

```
alculate average rating per year using the parsed timestamp column.
-----+-----+
year|  average_rating|
-----+-----+
2012|3.7347949560081957|
2014| 3.673793873892598|
2013| 3.668952064743723|
2005| 4.057298168955472|
2000| 4.201560468140442|
2002|3.9157303370786516|
2009|3.8131947842359626|
2006|3.8188488118905615|
2004|4.0836175626658004|
2011|3.8275800167546508|
2008|3.7822049941404488|
1999| 4.05088161209068|
2007|3.7131246438985155|
2015|3.8719307827900393|
2001| 4.015463917525773|
2010| 3.855782930576896|
2003| 4.013817480719794|
1998| 3.390885750962773|
1997| 4.006756756756757|
-----+-----+
```

- Determine the top 5 most active users based on the number of ratings.

```

• - Determine the top 5 most active users based on the number of ratings.

+-----+-----+
|userId|num_ratings|
+-----+-----+
| 58612|     24178|
|130827|     21471|
|124998|     20942|
| 88738|     20156|
| 10616|     14812|
+-----+-----+

```

4. Windowing Operations

Use window functions to gain time-based insights.

- Rank movies by rating count and average rating within each genre.

```

Rank movies by rating count and average rating within each genre.

+-----+-----+-----+-----+-----+
|movieId|      title|      genres|rating_count|  average_rating|rank|
+-----+-----+-----+-----+-----+
| 131031|La liga no es cos...|(no genres listed)|      8|      3.0|   1|
| 130532|I giorni contati ...|(no genres listed)|      7|      2.5|   2|
| 129530|Slingshot Hip Hop...|(no genres listed)|      3|      5.0|   3|
| 114725|Study in Choreogr...|(no genres listed)|      3|      5.0|   3|
| 123939|Women Aren't Funn...|(no genres listed)|      3|      4.0|   5|
| 126593|    Kocken (2005)|(no genres listed)|      3|      3.0|   6|
| 125535|Fist of Jesus (2012)|(no genres listed)|      3|      3.0|   6|
| 123439|Wuthering Heights...|(no genres listed)|      3|      2.0|   8|
| 114723|      At Land (1944)|(no genres listed)|      2|      5.0|   9|
| 131082|    Playground (2009)|(no genres listed)|      2|      4.5|  10|
| 117192|Doctor Who: The T...|(no genres listed)|      2|      4.0|  11|
| 125910| Stars Above (2012)|(no genres listed)|      1|      4.0|  12|
| 127005|A Year Along the ...|(no genres listed)|      1|      4.0|  12|
| 130298|    Pot v raj (2014)|(no genres listed)|      1|      3.5|  14|
| 130878|Doppelgänger Paul...|(no genres listed)|      1|      3.0|  15|
| 130296| A Fight For (2014)|(no genres listed)|      1|      3.0|  15|
| 126552|    Fort McCoy (2014)|(no genres listed)|      1|      2.5|  17|
|  83829|Scorpio Rising (1...|(no genres listed)|      1|      0.5|  18|
|  2959|    Fight Club (1999)|      Action| 1702|4.662162162162162|   1|
|  79132|    Inception (2010)|      Action| 1525|4.371147540983607|   2|
+-----+-----+-----+-----+-----+
only showing top 20 rows

```

- Calculate rolling average ratings for each movie over a 30-day window.

```
Calculate the average rating for each movie over the last 30 days.
+-----+-----+-----+-----+-----+
| movieId | title | genres | userId | rating | timestamp | rolling_avg_rating |
+-----+-----+-----+-----+-----+
| 1 | Toy Story (1995) | Adventure | Animati... | 99851 | 4 | 822853800 | 4.0 |
| 1 | Toy Story (1995) | Adventure | Animati... | 124035 | 5 | 823165403 | 4.5 |
| 1 | Toy Story (1995) | Adventure | Animati... | 46380 | 4 | 823235514 | 4.33333333333333 |
| 1 | Toy Story (1995) | Adventure | Animati... | 113947 | 5 | 823244802 | 4.5 |
| 1 | Toy Story (1995) | Adventure | Animati... | 121731 | 5 | 823494197 | 4.6 |
| 1 | Toy Story (1995) | Adventure | Animati... | 8050 | 5 | 824114137 | 4.66666666666667 |
| 1 | Toy Story (1995) | Adventure | Animati... | 99961 | 5 | 824981864 | 4.714285714285714 |
| 1 | Toy Story (1995) | Adventure | Animati... | 22528 | 4 | 824986853 | 4.625 |
| 1 | Toy Story (1995) | Adventure | Animati... | 107537 | 4 | 824990660 | 4.55555555555555 |
| 1 | Toy Story (1995) | Adventure | Animati... | 63203 | 2 | 825060356 | 4.3 |
| 1 | Toy Story (1995) | Adventure | Animati... | 81468 | 5 | 825353770 | 4.363636363636363 |
| 1 | Toy Story (1995) | Adventure | Animati... | 63308 | 5 | 825471878 | 4.454545454545454 |
| 1 | Toy Story (1995) | Adventure | Animati... | 20990 | 5 | 825932188 | 4.444444444444445 |
| 1 | Toy Story (1995) | Adventure | Animati... | 137727 | 5 | 825992544 | 4.5 |
| 1 | Toy Story (1995) | Adventure | Animati... | 7546 | 4 | 826273652 | 4.4 |
| 1 | Toy Story (1995) | Adventure | Animati... | 94179 | 5 | 826417952 | 4.454545454545454 |
| 1 | Toy Story (1995) | Adventure | Animati... | 7334 | 5 | 826446113 | 4.5 |
| 1 | Toy Story (1995) | Adventure | Animati... | 13261 | 5 | 826623271 | 4.538461538461538 |
| 1 | Toy Story (1995) | Adventure | Animati... | 135464 | 5 | 826741034 | 4.538461538461538 |
| 1 | Toy Story (1995) | Adventure | Animati... | 3338 | 5 | 826942928 | 4.571428571428571 |
+-----+-----+-----+-----+-----+
only showing top 20 rows
```

- Analyze user activity trends based on the number of ratings over time.

```

- Analyze user activity trends based on the number of ratings over time.

Daily Activity Trend
+-----+-----+
|userId|ratingsDF_timestamp|total_ratings|
+-----+-----+
| 79366|1997-09-15 21:13:53|       6|
| 79366|1997-09-15 21:46:17|       2|
| 79366|1997-09-15 22:00:10|       2|
| 79366|1997-09-15 22:09:27|       4|
| 79366|1997-09-15 22:17:29|       2|
| 79366|1997-09-15 22:32:38|       3|
| 79366|1997-09-15 23:01:27|       2|
| 79366|1997-09-16 00:43:28|       1|
| 79366|1997-09-16 00:57:04|       2|
| 79366|1997-09-17 23:04:46|       5|
| 79366|1997-10-02 00:50:41|       6|
| 79366|1997-10-02 00:50:42|       2|
| 1835|1997-10-15 06:01:46|       3|
| 1835|1997-10-15 06:33:24|       3|
| 1835|1997-10-16 08:06:41|       2|
| 1835|1997-10-16 08:08:18|       3|
| 1835|1997-10-16 08:10:52|       3|
| 79366|1997-10-28 00:15:42|       1|
| 83726|1997-11-10 14:44:23|       3|
| 83726|1997-11-10 15:10:31|       7|
+-----+
only showing top 20 rows

```

```

Monthly Activity Trend
+-----+-----+
|userId| month|total_ratings|
+-----+-----+
| 79366|1997-09|        29|
| 1835|1997-10|        14|
| 79366|1997-10|         9|
| 83726|1997-11|        82|
| 79366|1997-11|         2|
| 95895|1997-11|         6|
| 83726|1997-12|         2|
| 23333|1997-12|         4|
| 83726|1998-01|         2|
| 42259|1998-01|        20|
| 58612|1998-01|      695|
| 33323|1998-02|        28|
| 83726|1998-03|         3|
| 58612|1998-04|       140|
| 1835|1998-04|        23|
| 79366|1998-04|         2|
| 16687|1998-05|         4|
| 79366|1998-05|         8|
| 58612|1998-05|        15|
| 106308|1998-05|       71|
+-----+
only showing top 20 rows

```

5. Pivoting and Complex Transformations

Perform complex transformations to analyze trends and anomalies.

- Create a pivot table showing average ratings per genre by year.

Pivoting and Complex Transformations																			
genre	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Crime	4.0	3.83	4.38	4.22	4.14	4.1	4.08	4.15	4.11	3.94	3.84	3.87	3.92	4.0	3.96	3.88	3.77	3.79	4.07
Romance	3.51	3.02	3.98	4.32	4.1	3.84	3.82	4.07	3.98	3.76	3.69	3.86	3.78	3.81	3.8	3.77	3.8	3.73	3.86
Thriller	4.0	3.12	3.89	3.93	4.03	3.78	3.89	4.01	4.0	3.81	3.71	3.82	3.77	3.83	3.79	3.71	3.67	3.68	3.94
Adventure	4.17	3.55	4.15	4.17	3.95	3.85	4.1	4.09	4.01	3.83	3.71	3.69	3.79	3.78	3.75	3.64	3.54	3.62	3.73
Drama	4.21	3.52	4.05	4.21	4.07	4.15	4.04	4.15	4.13	3.89	3.75	3.85	3.88	3.93	3.86	3.84	3.83	3.98	
War	4.67	2.14	4.22	4.25	4.33	4.43	4.2	3.96	4.21	3.92	3.81	3.81	3.92	3.81	3.9	3.94	3.78	3.81	4.15
Documentary	null	2.0	3.35	3.85	4.0	4.12	4.09	4.24	3.9	3.97	3.87	3.64	3.66	3.91	3.83	3.87	3.67	3.79	4.0
Fantasy	2.67	3.48	3.87	4.25	4.1	3.88	4.19	4.18	4.18	3.84	3.71	3.73	3.85	3.77	3.84	3.61	3.58	3.55	3.91
Mystery	4.25	3.43	3.83	4.33	4.21	4.11	4.06	4.19	4.21	3.91	3.78	3.93	3.91	3.98	3.88	3.83	3.82	3.85	4.01
Musical	3.71	3.39	3.91	4.02	3.98	3.77	4.17	4.15	4.01	3.81	3.62	3.79	3.95	3.98	3.88	3.82	3.83	3.7	4.0
Animation	3.8	3.7	4.14	4.79	3.45	3.92	4.15	4.27	4.25	3.94	3.98	3.89	3.9	3.94	3.99	3.91	3.75	3.59	3.74
Film-Noir	4.5	4.11	3.87	4.27	4.58	4.25	4.2	4.18	4.27	4.16	3.8	4.04	4.05	3.97	3.89	3.75	4.07	3.91	4.46
(no genres listed)	null	0.5	null	4.71	3.2														
IMAX	4.0	5.0	5.0	3.0	4.75	2.91	3.91	4.24	3.89	4.01	3.54	3.82	3.96	3.71	3.73	3.51	3.53	3.74	
Horror	4.2	2.89	4.36	4.37	3.96	3.83	3.78	4.04	4.06	3.52	3.39	3.55	3.64	3.67	3.65	3.44	3.42	3.38	3.7
Western	null	4.5	3.9	4.32	4.18	3.98	4.22	4.13	3.83	3.78	3.71	3.78	3.9	3.71	3.75	3.69	3.99		
Comedy	3.8	3.35	4.1	4.32	3.87	3.84	3.94	4.06	4.0	3.73	3.67	3.72	3.75	3.78	3.78	3.72	3.59	3.68	3.79
Children	3.8	3.58	4.22	4.32	4.04	3.85	4.17	4.12	3.74	3.63	3.65	3.8	3.82	3.89	3.64	3.6	3.55	3.62	
Action	4.21	3.43	4.01	4.08	3.95	3.86	3.96	3.91	3.93	3.74	3.63	3.69	3.72	3.78	3.73	3.64	3.5	3.51	3.77
Sci-Fi	4.23	3.56	4.19	4.09	3.97	3.7	4.1	4.04	4.02	3.79	3.77	3.73	3.83	3.89	3.75	3.7	3.62	3.6	3.75

- Detect anomalies in ratings for movies within each genre.

```
Detect anomalies in ratings for movies within each genre.

+-----+-----+-----+-----+
| movieId | title | genres | rating | z_score |
+-----+-----+-----+-----+
| 1227 | Once Upon a Time ... | Crime | 0.5 | -3.5610187452155215 |
| 1573 | Face/Off (1997) | Crime | 1 | -3.039306390424684 |
| 1573 | Face/Off (1997) | Crime | 1 | -3.039306390424684 |
| 1573 | Face/Off (1997) | Crime | 1 | -3.039306390424684 |
| 1573 | Face/Off (1997) | Crime | 1 | -3.039306390424684 |
| 20 | Money Train (1995) | Crime | 1 | -3.039306390424684 |
| 2383 | Police Academy 6:... | Crime | 0.5 | -3.5610187452155215 |
| 420 | Beverly Hills Cop... | Crime | 0.5 | -3.5610187452155215 |
| 420 | Beverly Hills Cop... | Crime | 0.5 | -3.5610187452155215 |
| 4246 | Bridget Jones's D... | Romance | 0.5 | -3.23874235187354 |
| 4246 | Bridget Jones's D... | Romance | 0.5 | -3.23874235187354 |
| 4246 | Bridget Jones's D... | Romance | 0.5 | -3.23874235187354 |
| 4246 | Bridget Jones's D... | Romance | 0.5 | -3.23874235187354 |
| 736 | Twister (1996) | Romance | 0.5 | -3.23874235187354 |
| 736 | Twister (1996) | Romance | 0.5 | -3.23874235187354 |
| 420 | Beverly Hills Cop... | Thriller | 0.5 | -3.1750545040075036 |
| 420 | Beverly Hills Cop... | Thriller | 0.5 | -3.1750545040075036 |
| 736 | Twister (1996) | Thriller | 0.5 | -3.1750545040075036 |
| 736 | Twister (1996) | Thriller | 0.5 | -3.1750545040075036 |
| 3704 | Mad Max Beyond Th... | Adventure | 0.5 | -3.0760620981433253 |
+-----+-----+-----+-----+
only showing top 20 rows
```

```
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| userId | movieId | rating | ratingsDF.timestamp | title | genres | tag | tagsDF.timestamp | imdbId | tmdbId | rolling_avg | rolling_stddev | z_score |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| 79713 | 592 | 3 | 1999-12-11 16:13:53 | Batman (1989) | Crime | dark hero | 2013-05-11 12:38:32 | 96895 | 268 | 4.866666666666666 | 0.5163977794943222 | -3.614784456460255 |
| 23333 | 1245 | 4 | 1999-12-30 05:04:59 | Miller's Crossing... | Crime | neo-noir | 2013-05-16 00:36:57 | 100158 | 379 | 4.933333333333334 | 0.2581988897471611 | -3.614784456460257 |
| 136517 | 4684 | 1 | 2001-08-26 02:35:59 | Beverly Hills Cop... | Crime | crappy sequel | 2006-06-01 06:08:32 | 92644 | 96 | 4.4 | 1.1212238211627763 | -3.032408789508772 |
| 7723 | 1206 | 4 | 2002-02-16 20:04:12 | Clockwork Orange,... | Crime | Futuristmovies.com | 2006-01-15 02:36:04 | 66921 | 185 | 4.933333333333334 | 0.2581988897471611 | -3.614784456460257 |
| 30739 | 1396 | 2 | 2002-05-14 15:58:11 | Sneakers (1992) | Crime | ensemble cast | 2013-05-11 16:52:31 | 1095435 | 2322 | 4.6 | 0.8280786712108249 | -3.139798295087712 |
| 96370 | 4677 | 1.5 | 2003-05-15 18:04:34 | Turner & Hooch (1...) | Crime | dogs | 2006-01-16 01:13:11 | 98536 | 6951 | 4.033333333333333 | 0.8338093878327917 | -3.0582643447057904 |
| 1741 | 4719 | 0.5 | 2003-05-15 19:51:24 | Osmosis Jones (2001) | Crime | ostensibly for ki... | 2007-05-02 22:06:55 | 181739 | 12610 | 4.1 | 1.0555973258234952 | -3.4103913603528304 |
| 3149 | 1729 | 2 | 2003-07-06 02:15:30 | Jackie Brown (1997) | Crime | Tarantino | 2006-02-20 03:46:06 | 119396 | 184 | 3.8 | 0.5916079783099616 | -3.0425553170226594 |
| 1741 | 6662 | 3 | 2003-08-13 01:00:15 | Pink Panther, The... | Crime | Inspector Clouseau... | 2007-04-29 12:05:01 | 57413 | 936 | 4.5 | 0.4999999999999999 | -3.0000000000000004 |
| 107572 | 1783 | 1.5 | 2003-09-01 14:26:18 | Palmetto (1998) | Crime | boring | 2006-09-08 17:35:41 | 120792 | 30949 | 3.933333333333333 | 0.7761320457119086 | -3.135205338046275 |
| 58612 | 185 | 1.5 | 2004-02-27 17:07:36 | Net, The (1995) | Crime | computers | 2010-07-04 18:08:30 | 113957 | 1642 | 4.033333333333333 | 0.789816132912923 | -3.2074975779364494 |
| 123297 | 4406 | 4.5 | 2004-03-30 23:13:41 | Man Who Shot Libe... | Crime | BD-R | 2011-09-22 19:52:24 | 56217 | 11697 | 4.033333333333333 | 0.12909944487358055 | 3.614784456460257 |
| 123297 | 2731 | 5 | 2004-04-03 19:03:18 | 400 Blows, The (L...) | Crime | DVD-Video | 2007-02-26 21:43:18 | 53198 | 1471 | 4.033333333333333 | 0.2968084198523318 | 3.256870769190453 |
| 58612 | 7523 | 0.5 | 2004-05-07 13:09:48 | Desperate Hours (... | Crime | home invasion | 2006-08-12 07:55:16 | 99469 | 31676 | 4.033333333333333 | 1.1412190641506792 | -3.0961043714800884 |
| 101891 | 296 | 0.5 | 2004-08-07 07:44:14 | Pulp Fiction (1994) | Crime | organized crime | 2009-05-21 18:53:35 | 110912 | 680 | 3.8 | 0.9783367810436532 | -3.373071588374387 |
| 51266 | 56 | 3 | 2004-08-30 09:41:39 | Usual Suspects, T... | Crime | unreliable narrators | 2008-01-01 19:27:00 | 114814 | 629 | 4.6 | 0.50789255283711 | -3.155242559864612 |
| 103872 | 8370 | 4 | 2004-10-16 16:14:26 | Blind Swordsman: ... | Crime | Takeshi Kitano | 2006-05-16 11:01:27 | 1363226 | 246 | 4.466666666666667 | 0.12909944487358055 | -3.614784456460257 |
| 131086 | 367 | 3 | 2005-03-23 23:39:00 | Mask, The (1994) | Crime | carrey decline be... | 2006-02-07 21:56:54 | 118675 | 854 | 4.5 | 0.4999999999999999 | -3.0000000000000004 |
| 131910 | 6874 | 2 | 2005-04-14 21:31:58 | Kill Bill: Vol. 1... | Crime | Bibliothek | 2006-01-30 22:26:21 | 266697 | 24 | 4.8 | 0.7745966692414834 | -3.614784456460256 |
| 93004 | 2379 | 1 | 2005-07-25 17:47:11 | Police Academy 2:... | Crime | idiot plot | 2014-01-05 22:37:55 | 89822 | 10157 | 5.733333333333334 | 0.8208590157935307 | -3.3298450534663364 |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
only showing top 20 rows
```

- Calculate rating distributions for each genre (e.g., percentage of 5-star ratings).

```

Calculate rating distributions for each genre (e.g., percentage of 5-star ratings).
+-----+-----+-----+-----+
|      genres|rating|rating_count|total_ratings|    percentage|
+-----+-----+-----+-----+
|(no genres listed)| 0.5|      1|        46|2.1739130434782608|
|(no genres listed)| 2|      3|        46| 6.521739130434782|
|(no genres listed)| 2.5|      8|        46|17.391304347826086|
|(no genres listed)| 3|     16|        46| 34.78260869565217|
|(no genres listed)| 3.5|      1|        46|2.1739130434782608|
|(no genres listed)| 4|      7|        46|15.217391304347828|
|(no genres listed)| 4.5|      2|        46|4.3478260869565215|
|(no genres listed)| 5|      8|        46|17.391304347826086|
|      Action| 0.5| 1707| 107325|1.5904961565338924|
|      Action| 1| 2524| 107325| 2.351735383181924|
|      Action| 1.5| 2644| 107325| 2.463545306312602|
|      Action| 2| 5285| 107325| 4.924295364546937|
|      Action| 2.5| 6550| 107325| 6.102958304216166|
|      Action| 3| 12173| 107325|11.342184952247845|
|      Action| 3.5| 16446| 107325| 15.3235499650594|
|      Action| 4| 25217| 107325| 23.49592359655253|
|      Action| 4.5| 16179| 107325| 15.07477288609364|
|      Action| 5| 18600| 107325|17.330538085255064|
|      Adventure| 0.5| 1047| 77199|1.3562351843935803|
|      Adventure| 1| 1577| 77199| 2.042772574774285|
+-----+-----+-----+-----+
only showing top 20 rows

```

6. User Behavior Analysis

Deep dive into user preferences and rating patterns.

- Cluster users based on their rating behavior (e.g., predominantly high ratings, diverse ratings).

```

User Behavior Analysis
+-----+-----+-----+-----+
|cluster|avg_rating_in_cluster|variance_in_cluster|rating_count_in_cluster|user_count|
+-----+-----+-----+-----+
| 1| 3.5949968893718514| 0.8587160816933747| 9854.6| 30|
| 3| 2.5925698192587037| 0.0890947860440152| 7.61243523316622| 965|
| 2| 3.6715647043749544| 0.9719422810322198| 287.54859218891914| 2202|
| 0| 4.548772735817733| 0.16285936072597526| 36.69232710752146| 3962|
+-----+-----+-----+-----+

```

- Identify the most tagged movies and the most common tags across all users.

```

    . - Identify the most tagged movies and the most common tags across all users.

+-----+-----+
|movieId|tag_count|
+-----+-----+
| 79132|    10675|
|  296|     7812|
| 2959|     6808|
| 4878|     4576|
| 32587|     4330|
| 2571|     4218|
|  356|     3912|
| 72998|     3844|
| 60684|     3822|
| 58559|     3724|
+-----+-----+

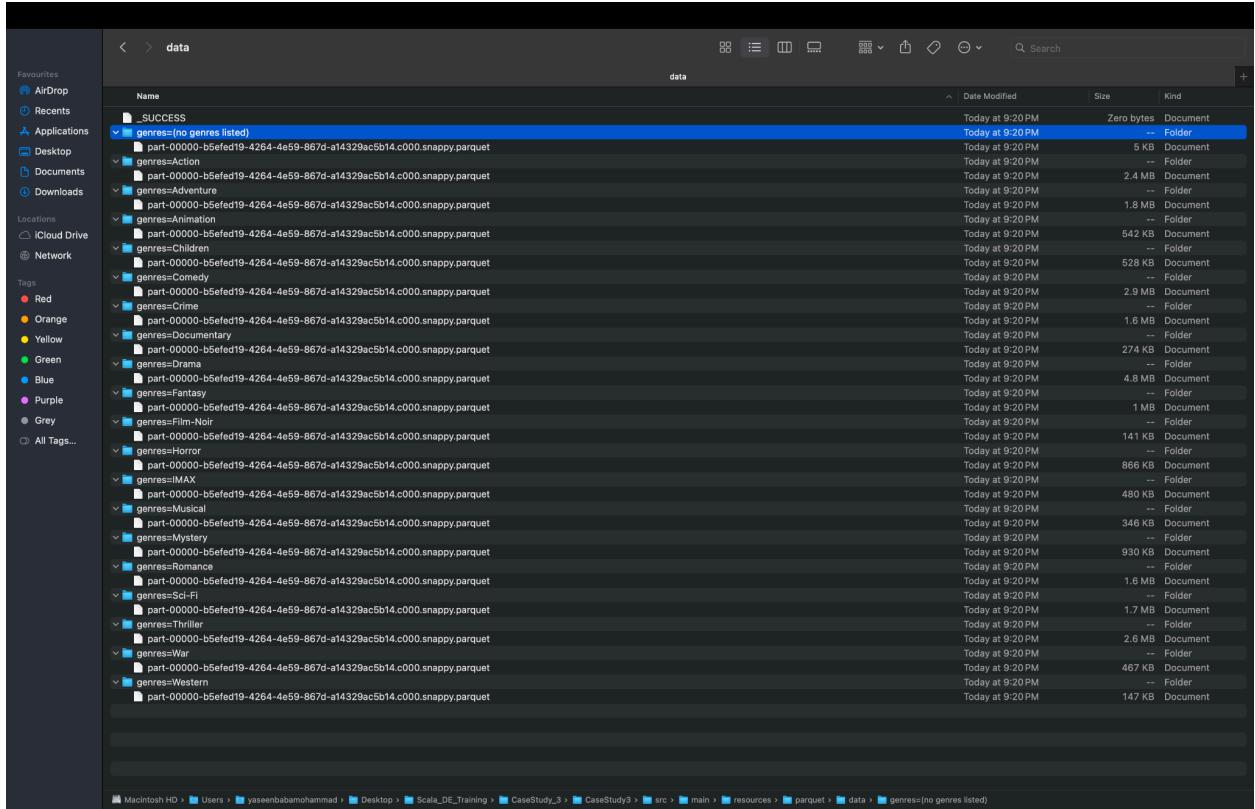
+-----+-----+
|      tag|tag_count|
+-----+-----+
|  sci-fi|    11151|
|  action|     8439|
|atmospheric|     8150|
|   surreal|     7936|
|twist ending|     6919|
|  dystopia|     6800|
|based on a book|     6733|
|   comedy|     6316|
|  stylized|     5879|
|    funny|     5770|
+-----+-----+

```

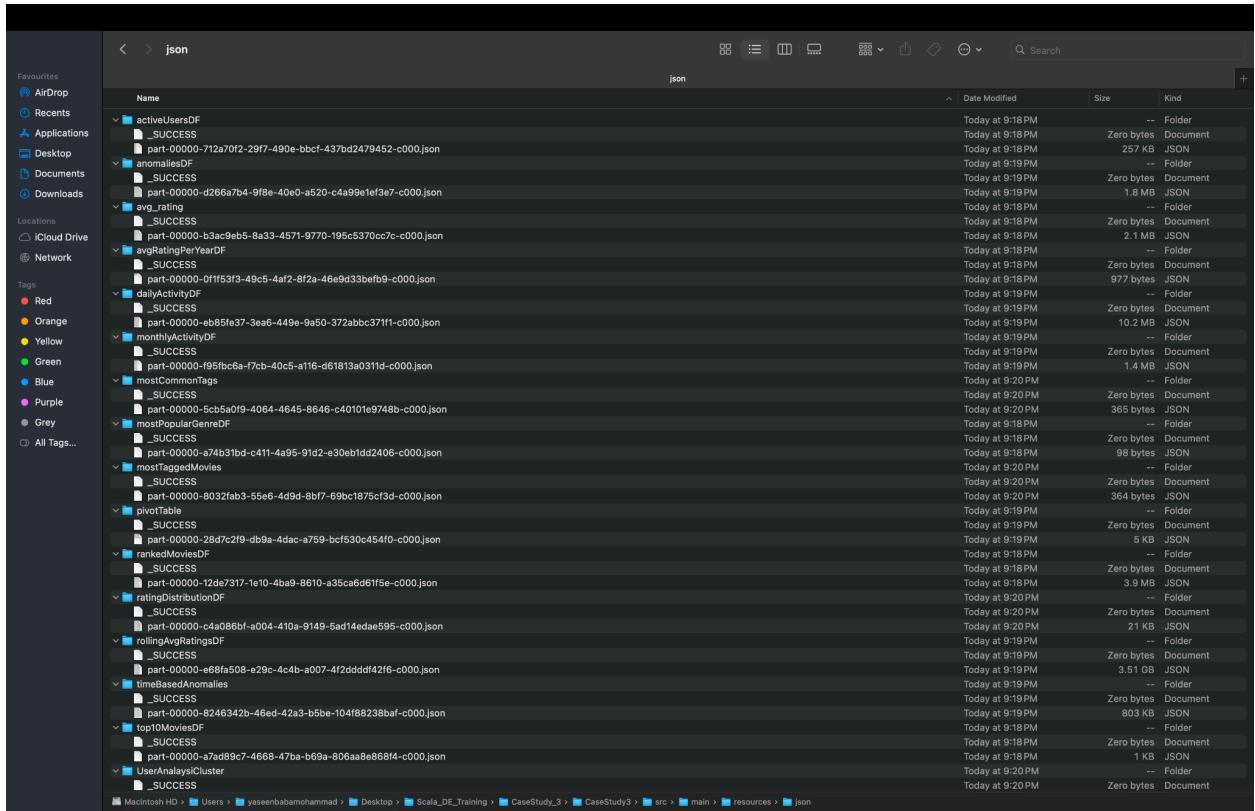
7. Storage Optimization

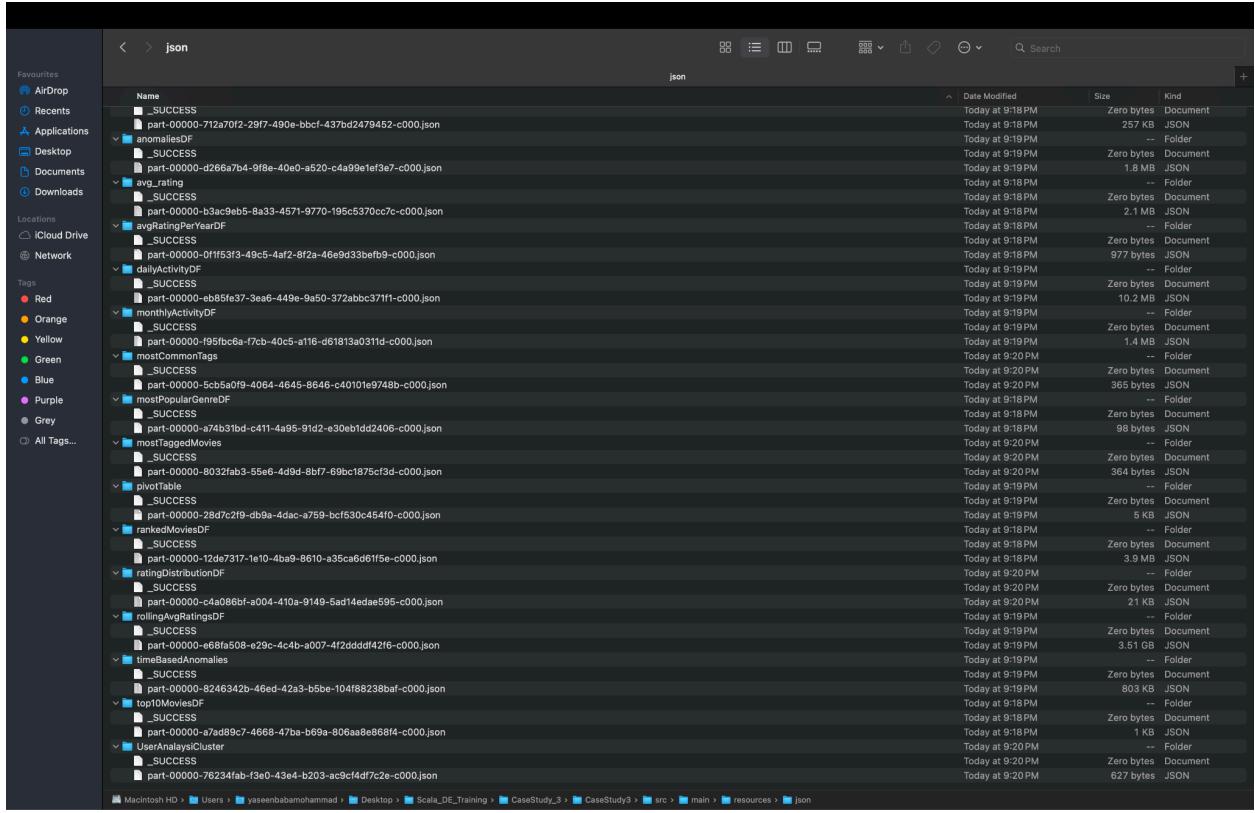
Optimize the storage of processed data for future use.

- Store enriched datasets in Parquet format, partitioned by genre.



- Save aggregated metrics in JSON format for visualization.





8. Visualization and Reporting

Generate reports and visualizations based on the aggregated data.

- Create summary reports for genre trends, user activity, and top-rated movies.

```

Summary Reports for the User Activity
+-----+-----+-----+
|userId|movies_rated|      avg_rating|      rating_stddev|
+-----+-----+-----+
| 58612|     24178| 3.505811067912979|0.8475891165291282|
|130827|     21471| 3.156536723953239|0.9662294988137566|
|124998|     20942| 2.67407601948238| 1.145374902386008|
| 88738|     20156|3.6148045247072833|1.3099118789934774|
| 10616|     14812| 3.632088846880907|0.9725077185543893|
| 52814|     13613| 3.376478366267538|0.7077583352172978|
| 27898|     11732| 3.568232185475622|0.9115531655378883|
| 70201|     10702|3.0117735002803214|0.5675158579624405|
| 68558|     10320| 3.309980620155039|0.7489290086247757|
| 77463|     9601|3.7662222685136966|0.5465836968616014|
| 1741|     9516|3.2311370323665405|1.0578070543816132|
| 25737|     8548|3.8304281703322416|1.2761804095367304|
| 11081|     8309| 2.729510169695511|0.7679268298996731|
| 4450|     8099| 3.000432152117545|0.8756501650724755|
| 9815|     7986| 4.331329827197596|0.6031955477151367|
|123297|     7784| 4.044321685508736|0.5508580225705941|
| 28906|     7766| 3.366340458408447|0.7214192740294245|
|120937|     7533|3.9754413912120006|0.5569596189820307|
| 57434|     7289| 3.620112498285087|0.8565938075415146|
|119367|     7059| 4.19974500637484|1.0340518276894535|
+-----+-----+-----+
only showing top 20 rows

```

```

Summary Reports for the Top Rated Movies
+-----+-----+-----+
|movieId|      avg_rating|num_ratings|
+-----+-----+-----+
| 77658| 4.739669421487603|     121|
| 42422|       4.71875|     128|
| 32460| 4.712765957446808|     188|
| 2959| 4.662162162162162|     6808|
| 4973| 4.661599099099099|    1776|
| 7669| 4.654929577464789|     142|
| 1233| 4.593896713615023|     639|
| 293| 4.56312625250501|    1996|
| 318|4.5556835637480795|    2604|
| 27156| 4.526315789473684|     285|
| 5618|4.5188953488372094|    2064|
| 55721|        4.5|     468|
| 27728|        4.5|     240|
| 4226| 4.496536796536796|    2310|
| 7502|4.4950495049504955|     303|
| 1916| 4.492063492063492|     126|
| 1188| 4.491803278688525|     122|
| 2139| 4.489130434782608|     184|
| 1221| 4.487623762376238|     404|
| 86781|4.4868421052631575|     114|
+-----+-----+-----+
only showing top 20 rows

```

Summary Reports for the Genres Trends				
	genres	avg_rating	unique_users	total_ratings
	Drama	3.890947107141433	5495	200632
	Comedy	3.74152734795479	4384	118912
	Thriller	3.76777100842162	4282	115061
	Action	3.680801304449103	3961	107325
	Sci-Fi	3.7531929824561403	3137	78375
	Adventure	3.7280923328022384	3343	77199
	Romance	3.797961542964177	3435	68336
	Crime	3.912818109936451	3570	68294
	Fantasy	3.7539541727070125	2644	46217
	Mystery	3.9145623985126963	2657	42493
	Horror	3.559419730273863	2111	36259
	IMAX	3.702177044981591	1770	24988
	Animation	3.880011842327863	1666	23644
	Children	3.7440638380692874	1602	20552
	War	3.888863357843137	1824	19584
	Musical	3.8690921585095825	1299	13097
	Documentary	3.8139393268379096	926	9032
	Film-Noir	3.9944802989130435	816	5888
	Western	3.8154984840378097	761	5607
	(no genres listed)	3.369565217391304	15	46

- Generate plots for average ratings per year and rolling averages over time.

