CA4022: Data at Speed and Scale
Assignment 1: Pig and Hive on HADOOP MapReduce

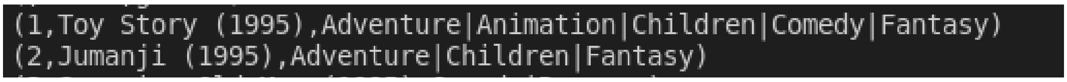
Niamh Murphy - 18301373

https://github.com/murphn88/ca4022-assignment-1

1. Data Cleaning Using PIG

Load Data

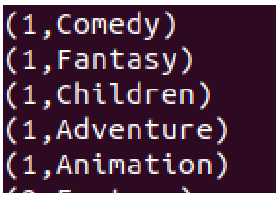
For data cleaning purposes, it was only necessary to load the movies table. The ratings table was sufficiently clean for our analysis, and the links and tags table did not contain any useful information for our analysis. To avoid issues with the delimiter I used the CSVExcelStorage class to read in the movies table.



```
(1,Toy Story (1995),Adventure|Animation|Children|Comedy|Fantasy)
(2,Jumanji (1995),Adventure|Children|Fantasy)
```

Split Genres and Create New Table

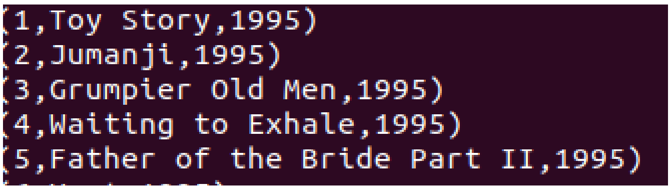
Having a list of genres in one cell violates the First Normal Form rule of relational databases. To address this, I split genres the genre column on pipes ('|'), and created a new table called genres that contains each movieId and genre pair. I then removed the genre field from the movies table.



```
(1,Comedy)
(1,Fantasy)
(1,Children)
(1,Adventure)
(1,Animation)
```

Separate Title and Year

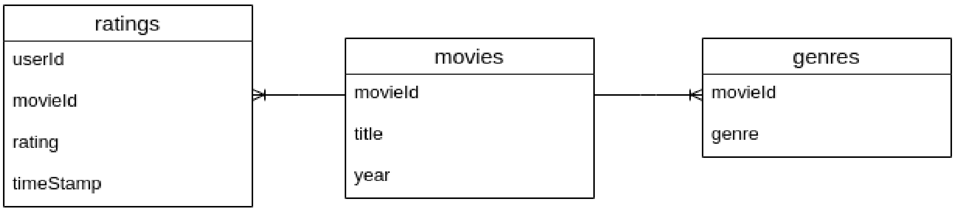
Additionally the title field was violating the First Normal Form rule, as it was storing the title and year in the same field. To split them into separate columns, I used REGEX_EXTRACT. As some movies did not have an associated year, to extract the title, my regular expression first attempts to match characters that precede brackets containing a year, but if there are no brackets containing a year, it extracts the full string.



```
(1,Toy Story,1995)
(2,Jumanji,1995)
(3,Grumpier Old Men,1995)
(4,Waiting to Exhale,1995)
(5,Father of the Bride Part II,1995)
```

Save Cleaned Data

The final data cleaning step was to save the cleaned movies table and genres table, so they could be used for analysis. To achieve this, I used PigStorage('\t'), so there would be no delimiter issues to fix when loading in the data either again in pig, or in hive.

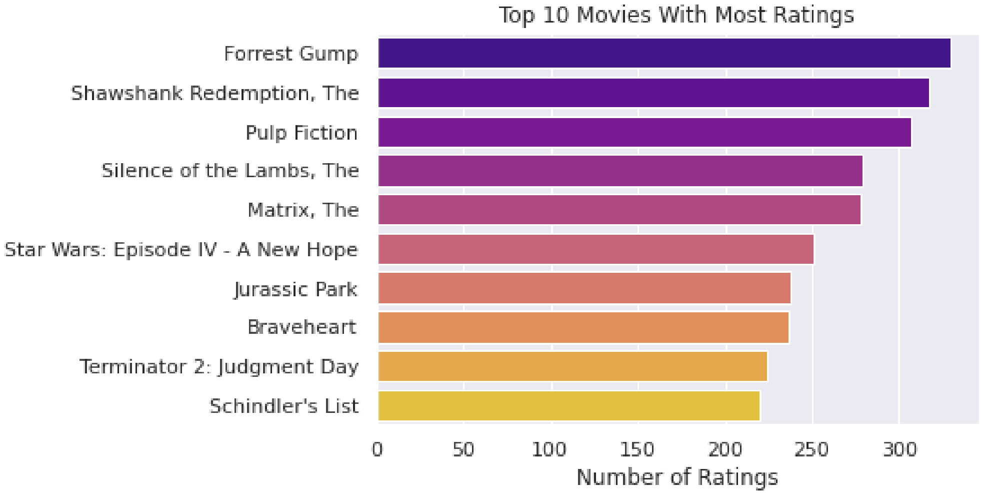


2. Hive and Pig Analysis

Queries 1-3 were done in both Pig and Hive, while the remaining, more complex queries were only executed in Hive. The visualisations have been created using the saved output and Python.

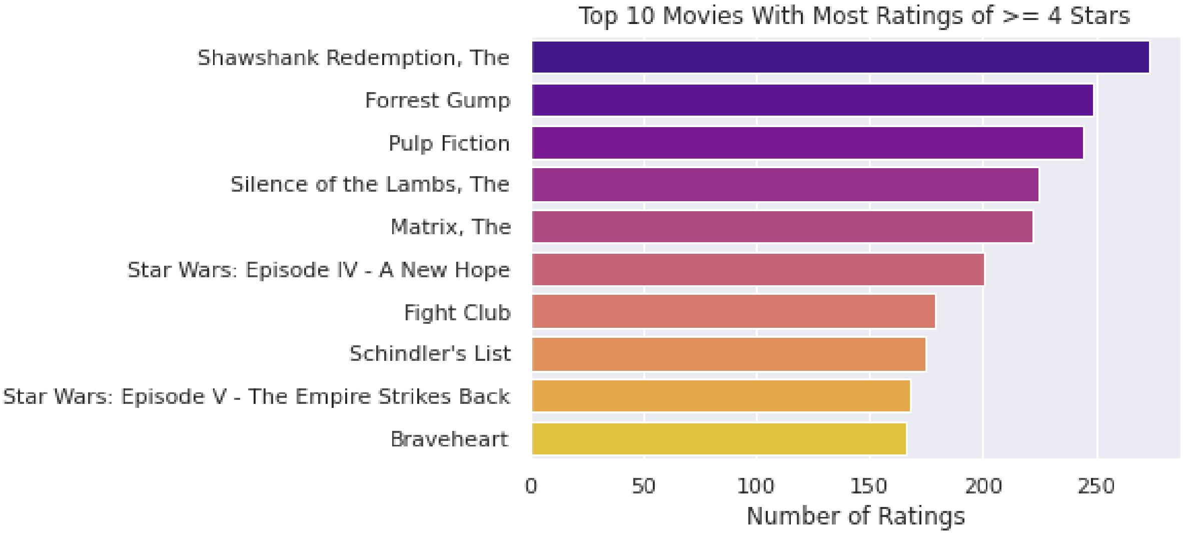
Query 1: What is the title of the movie with the most ratings?

Approach: Group ratings table by movieId and perform a count of ratings, then join to movies table on movieId.
Answer: Forrest Gump, 329 ratings.



Query 2: What is the title of the most liked movie?

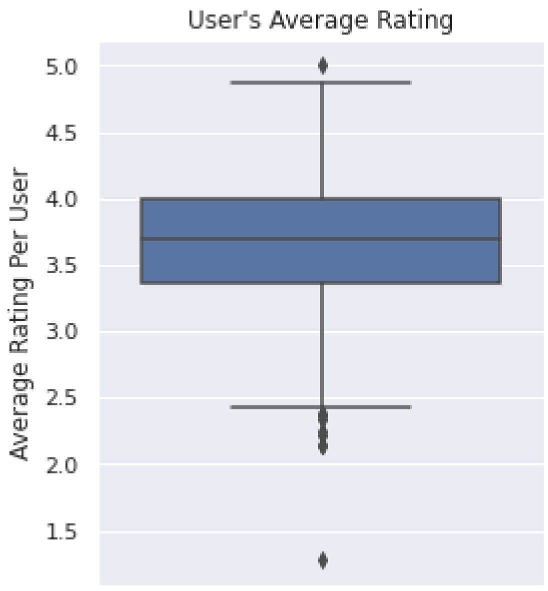
Approach: Filter ratings table by rating >= 4.0, group by movieId, perform count of ratings and join to movies table on movieId.
Answer: Shawshank Redemption, 274 rating of 4 or more stars.



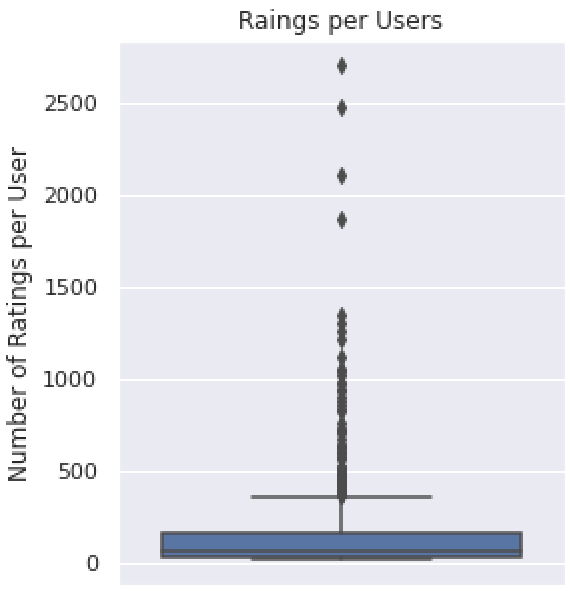
Query 3: What is the User with the highest average rating?

Approach: Group ratings table by userId and get average rating.

Answer: User 53, with an average rating of 5.0.



User's Average Rating



Raings per Users

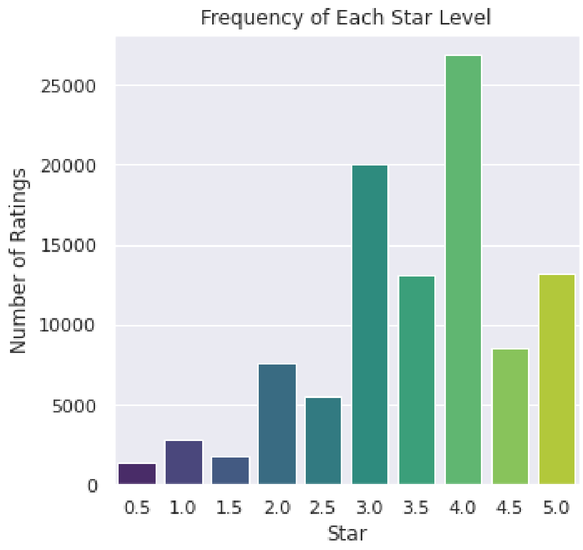
Discussion: The most generous user, user 53, can be easily identified as a slight outlier in the left boxplot of user's average ratings. 50% of users have an average rating of between 3.3 and 4 stars.

From the right boxplot, we can see that the median number of ratings per user is 70.5, with 75% of users having less than 168 ratings. However, there are some extreme outliers, with the most dedicated user having 2698 ratings

Query 4 & 5: Count the number of ratings for each star level & what is the most popular rating?

Approach: Group ratings table by rating and perform count.

Answer: 4 stars is the most popular rating, with 26818 ratings.

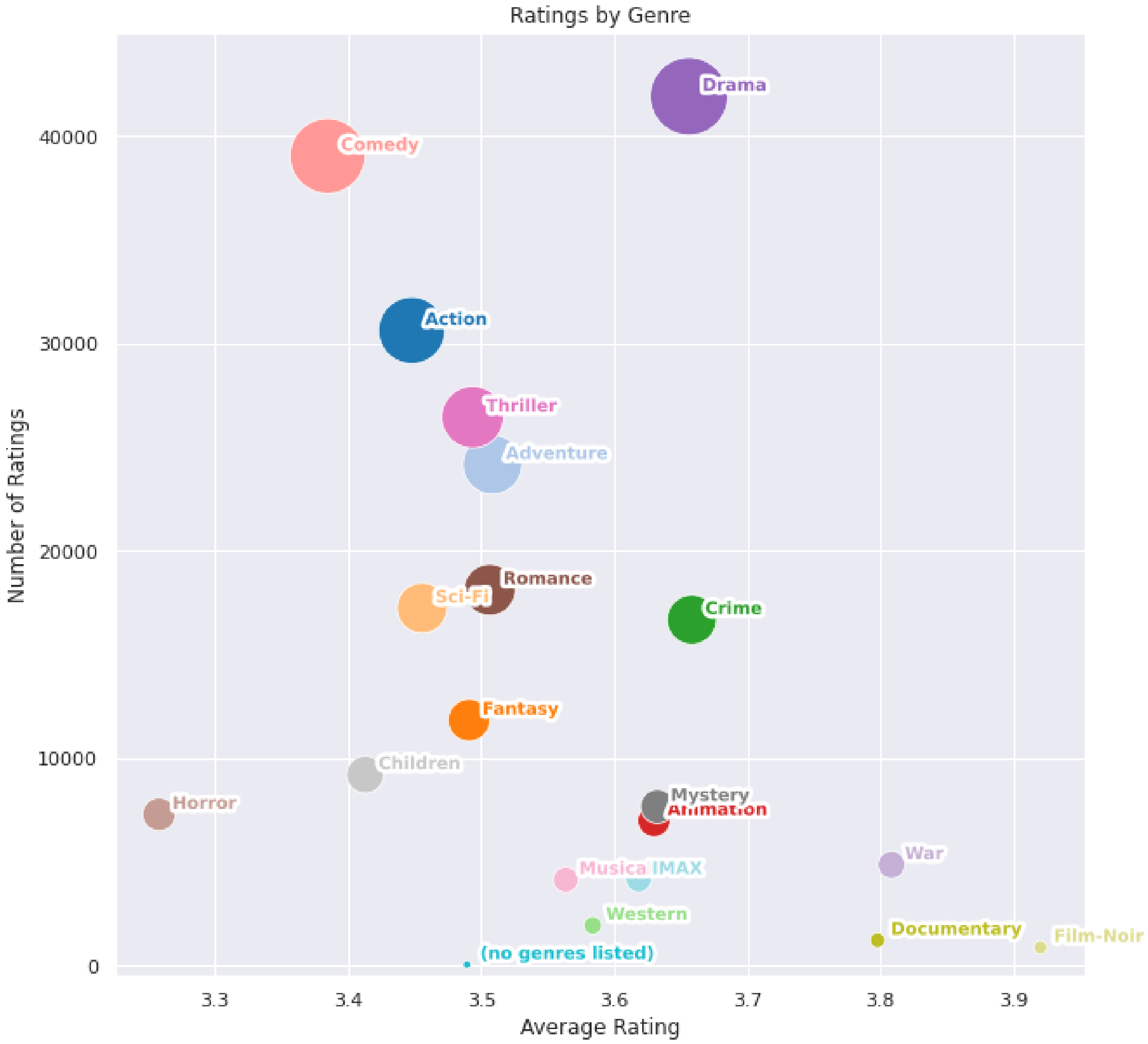


Frequency of Each Star Level

Query 6: How are ratings distributed by genre?

Approach: Join genres table and ratings table on movieId, group by genre, get count of ratings and average rating.

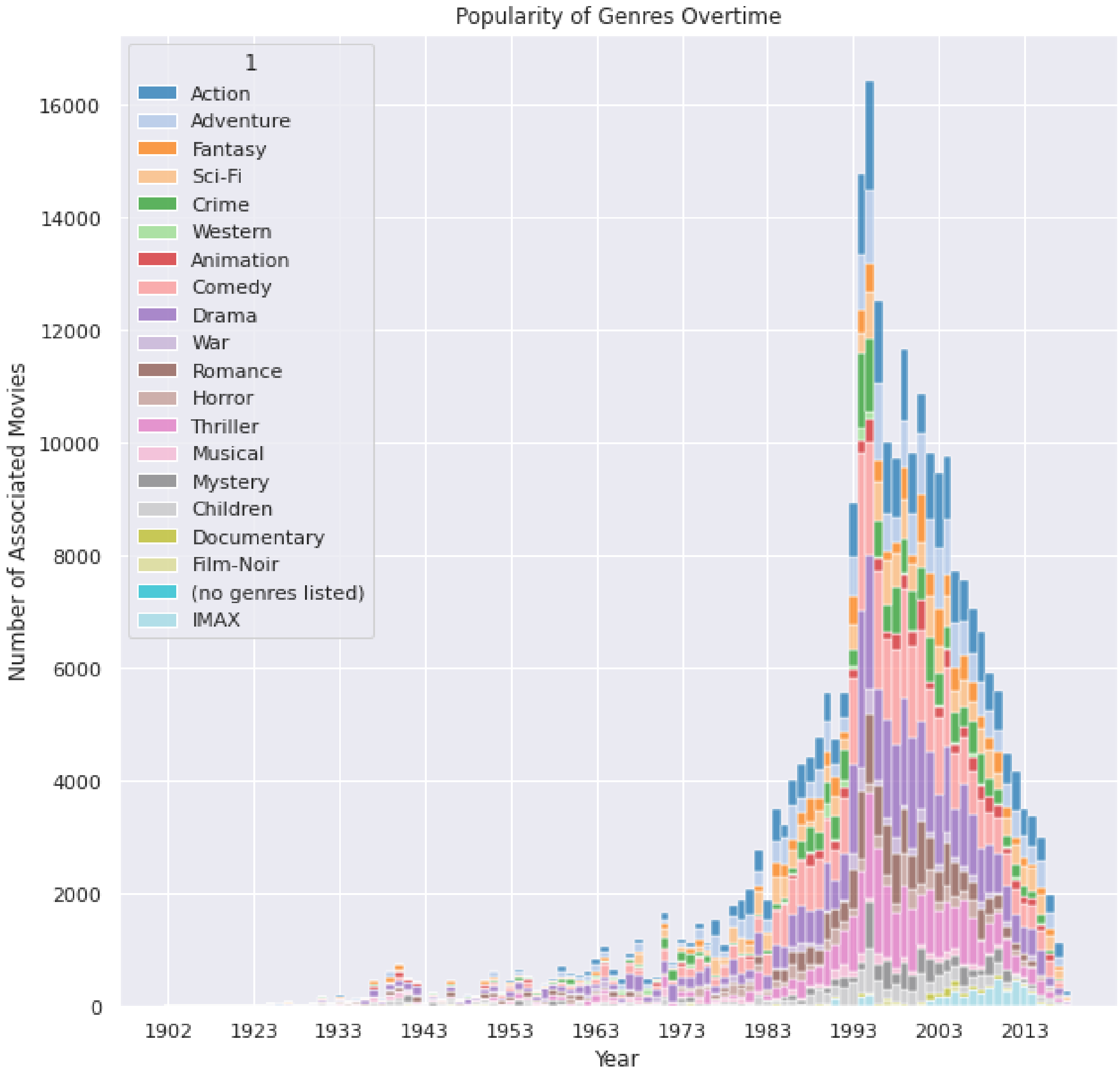
Note: I decided to calculate the average of all ratings per genre, rather than the average rating per movie in each genre. Hence, not giving an unequal weighting to ratings of movies with fewer ratings.



Discussion: This graph captures the average rating and the number of ratings for each genre. The area of the bubbles are influenced by the number of ratings, making it easy to spot the most and least popular genres. Film-Noir is the genre with the highest average rating, however, it has only received 870 ratings, indicating that it is an unpopular or niche genre. This suggests that Film-Noir movies have a small market segment, but are well-made and popular amongst that market. Drama, comedy and action appear to be the three most popular genres, having received the highest number of ratings.

Query 7: Popularity of Genres Overtime

Approach: Join genres table, ratings table, and movies table on movieId, group by year and genre and perform count of ratings.



Discussion: This plot shows how many movies are associated with each genre per year. It is useful for conveying the proportion of genres each year, but also for spotting trends. For example, we can see the emergence and growth of IMAX films. They are first visible in the early 90s, then disappear and remerge and grow from the early 2000s. This is presumably correlated to the initial development of high resolution cameras and theatres and then their reduction in cost and availability in the early 2000s. Drama, comedy, adventure and action movies have remained popular genres since the beginning.