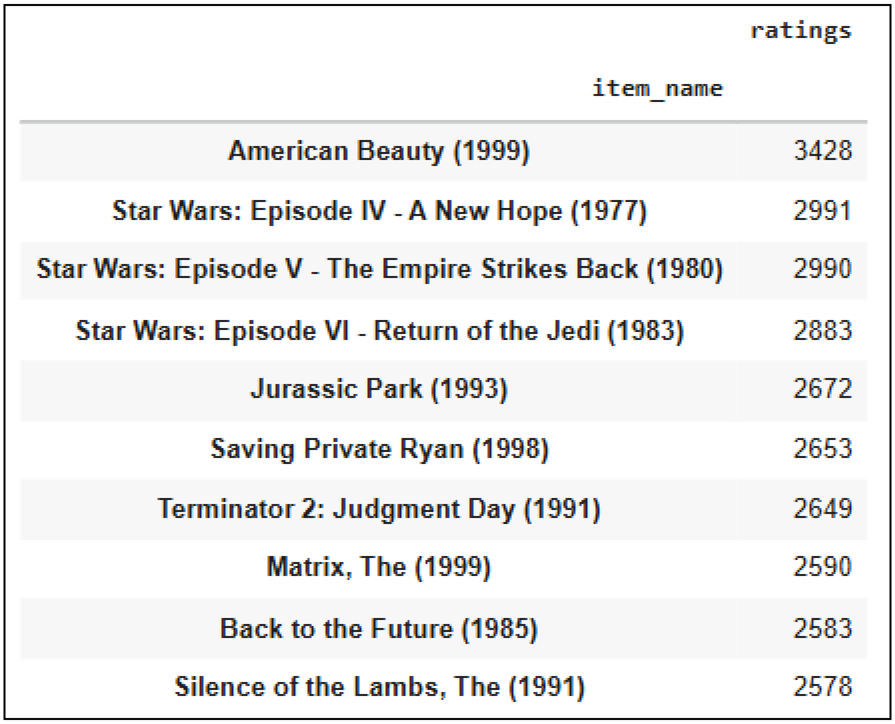
# Case 2- Modern Analytics
# Section C, Team #12

*Ankit Chandekar, Jinghua He, Abigail Miller,  Ricky Han, Shicheng Li*

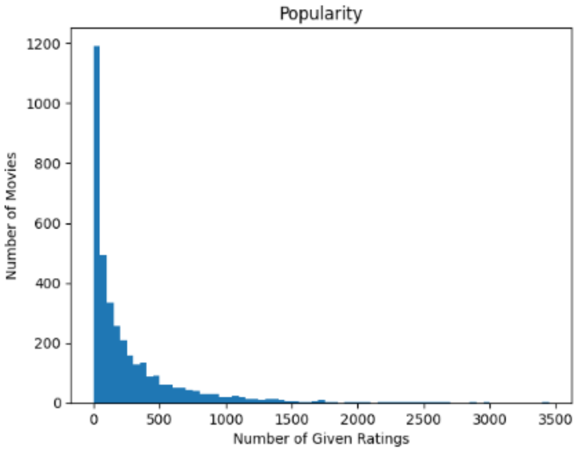
**Find the top 10 most popular movies in the dataset.**



| | ratings |
|---|---|
| item_name | |
| American Beauty (1999) | 3428 |
| Star Wars: Episode IV - A New Hope (1977) | 2991 |
| Star Wars: Episode V - The Empire Strikes Back (1980) | 2990 |
| Star Wars: Episode VI - Return of the Jedi (1983) | 2883 |
| Jurassic Park (1993) | 2672 |
| Saving Private Ryan (1998) | 2653 |
| Terminator 2: Judgment Day (1991) | 2649 |
| Matrix, The (1999) | 2590 |
| Back to the Future (1985) | 2583 |
| Silence of the Lambs, The (1991) | 2578 |

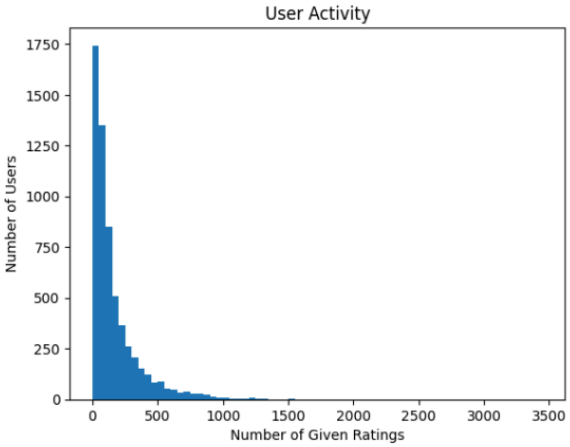
The chart above represents the top ten most popular movies in the dataset. Popularity is calculated by the number of ratings each movie has, so the movies above have the most ratings. The number one most popular movie is American Beauty (1999) with 3,428 ratings. The next three most popular movies are from the Star Wars series.

**Plot the histogram of popularity (x-axis: # of given ratings, y-axis: # of movies with the given # of ratings).**



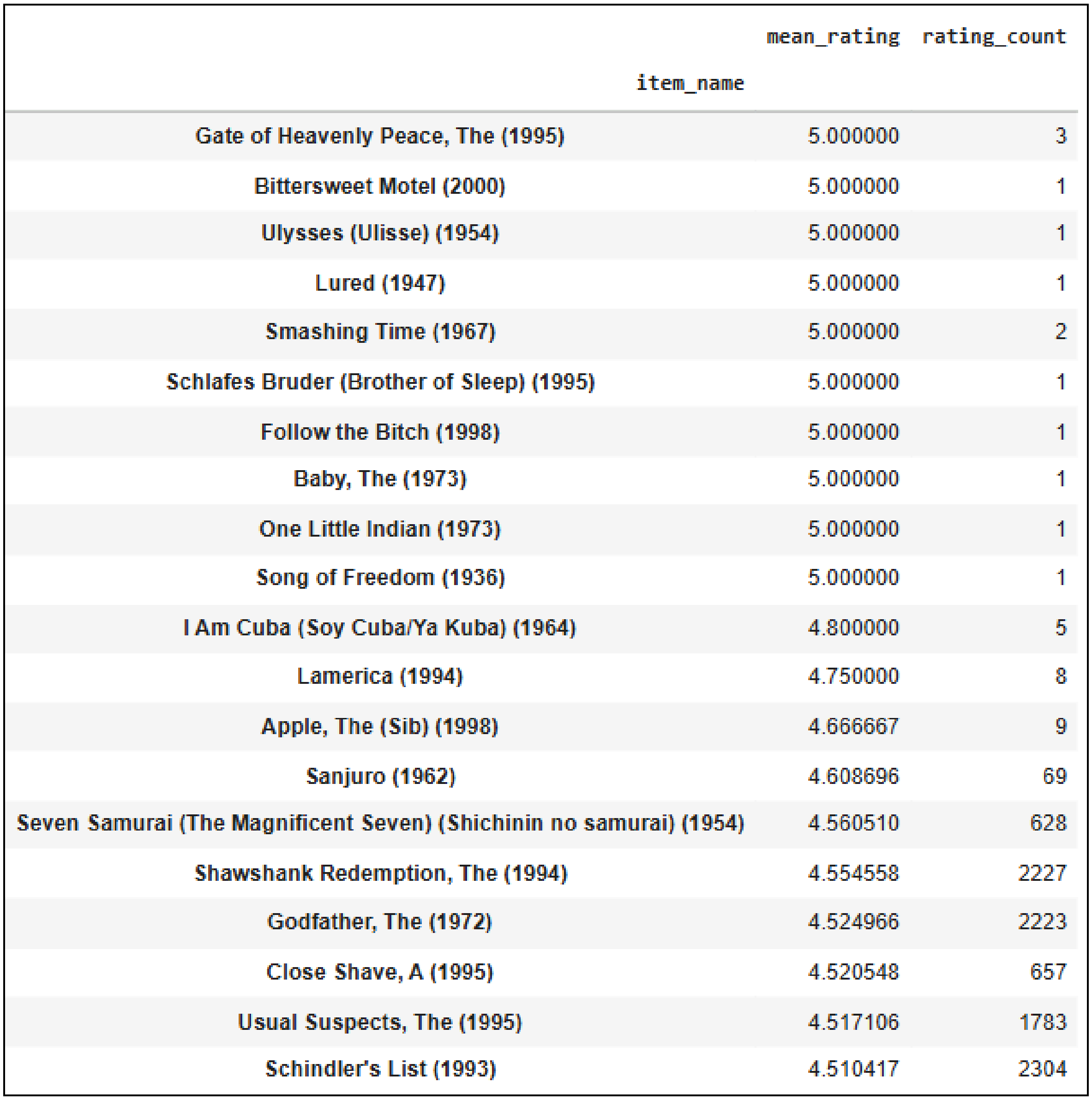
The popularity graph represents the number of movies that have an amount of ratings that fall within each bin. Each bin has a range of 50 ratings. Based on the first bin, about 1,200 movies have less than 50 ratings. The number of movies in each bin rapidly declines after that. Very few movies have a number of ratings beyond 1,500.

**Plot the histogram of user activity**



The user activity graph represents the number of users who have rated an amount of movies that fall within each bin. The bins in this graph are also ranges of 50 ratings. Based on the first bin, about 1,750 users have rated less than 50 movies. This graph also has a steep decline in the number of users in each bin as the number of given ratings increases. As a result, few people have given out more than 1500 ratings.

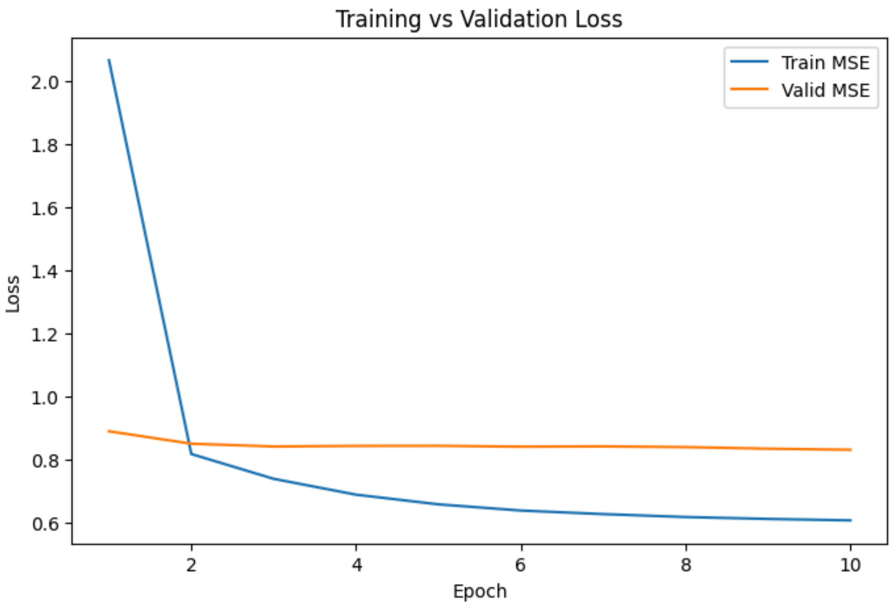
**Compute the average ratings for every movie and find the top 20 highly rated movies. Include the list of 20 highly rated movies and their ratings in your report. Are these highly rated movies popular?**



| item_name | mean_rating | rating_count |
|---|---|---|
| Gate of Heavenly Peace, The (1995) | 5.000000 | 3 |
| Bittersweet Motel (2000) | 5.000000 | 1 |
| Ulysses (Ulisse) (1954) | 5.000000 | 1 |
| Lured (1947) | 5.000000 | 1 |
| Smashing Time (1967) | 5.000000 | 2 |
| Schlafes Bruder (Brother of Sleep) (1995) | 5.000000 | 1 |
| Follow the Bitch (1998) | 5.000000 | 1 |
| Baby, The (1973) | 5.000000 | 1 |
| One Little Indian (1973) | 5.000000 | 1 |
| Song of Freedom (1936) | 5.000000 | 1 |
| I Am Cuba (Soy Cuba/Ya Kuba) (1964) | 4.800000 | 5 |
| Lamerica (1994) | 4.750000 | 8 |
| Apple, The (Sib) (1998) | 4.666667 | 9 |
| Sanjuro (1962) | 4.608696 | 69 |
| Seven Samurai (The Magnificent Seven) (Shichinin no samurai) (1954) | 4.560510 | 628 |
| Shawshank Redemption, The (1994) | 4.554558 | 2227 |
| Godfather, The (1972) | 4.524966 | 2223 |
| Close Shave, A (1995) | 4.520548 | 657 |
| Usual Suspects, The (1995) | 4.517106 | 1783 |
| Schindler's List (1993) | 4.510417 | 2304 |

The chart above shows the top 20 highly rated movies. Based on this, the top 13 rated movies are very unpopular as they have less than ten ratings. Eight of the movies with an average of five stars only have 1 rating. Therefore these numbers are not reliable because it is based on one person’s opinions. The 15th -20th highly rated movies have average ratings above 4.5, and they are fairly popular as they have rating counts between 628 - 2304. This means these ratings are also more reliable because they are averaged over a larger group.

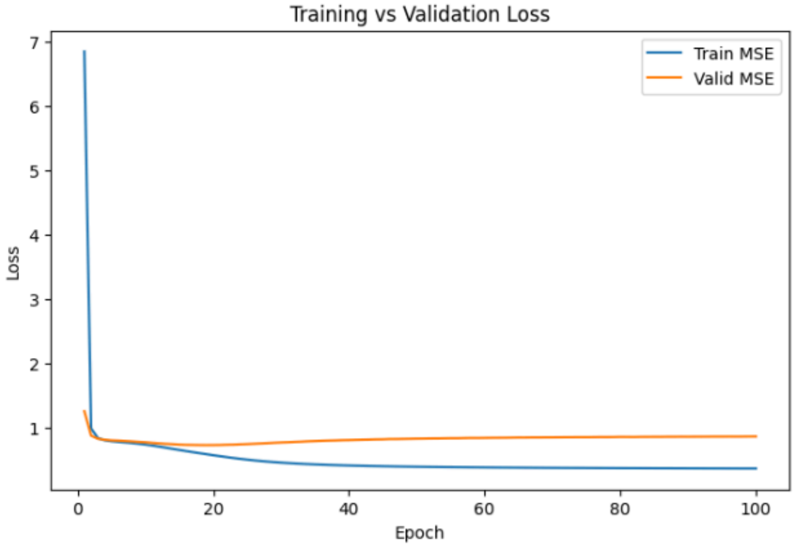
**Train the model and plot the train losses and valid losses over epoches. At which epoch, would you stop the training? Compute the RMSE (square root of the mean squared error) and MAE (mean absolute error) on the test data.**



```
RMSE: 0.9131912139114878, MAE: 0.7142546049186161
```

For our first model, the training and validation losses stabilized after epoch 8, with the minimum validation loss reached at epoch 10 (Val RMSE = 0.9111, MAE = 0.7120). Thus, epoch 10 is the optimal stopping point before overfitting begins. The RMSE for the test data is 0.8884. The MAE for the test data is 0.7407. The hyperparameters for this model were embedding_dim = 32, n_iter = 10, l2 = 1e-5, learning_rate = 1e-2

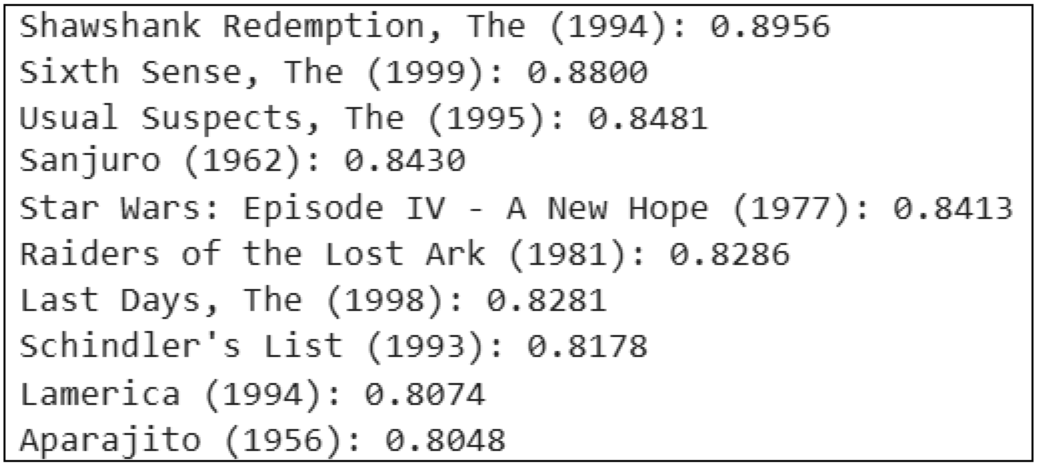
**List the hyperparameters and the RMSE (square root of the mean squared error) and MAE on test data for the optimized model in the report.**



```
RMSE: 0.8569108157188912, MAE: 0.6719570798533303
```

After adjusting the hyperparaments, we landed on a better model (shown above) that has a validation RMSE of .8569.  It also has an MAE of .6719. The hyperparameters for this model were embedding_dim = 40, n_iter = 100, L2 = 1e-5, learning_rate = 1e-3. Compared to the previous model, we increased the embedding dimensions and number of epochs and we decreased the learning rate. After about 40 epochs, both the train and validation MSE flattened out.

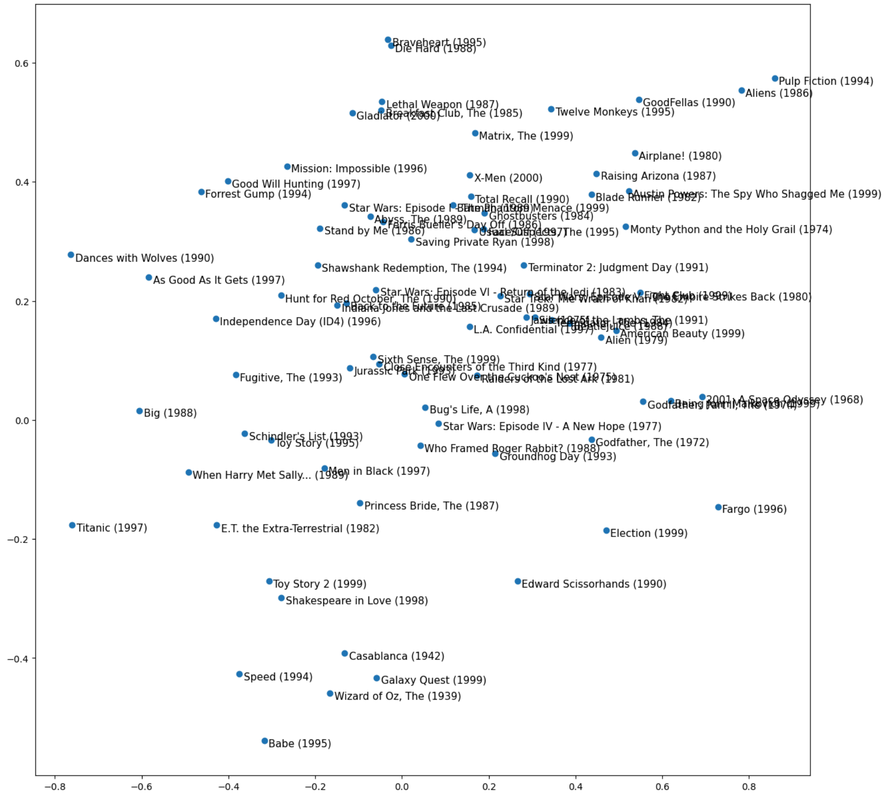
**What are your top 10 movies with the largest values of movie bias? What do you think about this ranking?**



```
Shawshank Redemption, The (1994): 0.8956
Sixth Sense, The (1999): 0.8800
Usual Suspects, The (1995): 0.8481
Sanjuro (1962): 0.8430
Star Wars: Episode IV - A New Hope (1977): 0.8413
Raiders of the Lost Ark (1981): 0.8286
Last Days, The (1998): 0.8281
Schindler's List (1993): 0.8178
Lamerica (1994): 0.8074
Aparajito (1956): 0.8048
```

The list above shows the top 10 movies with the largest values of movie bias, which is consistent with what an item-bias metric measures: films that attract consistently above-average ratings across almost all users. Its concentration on well-established Hollywood classics indicates that the dataset is popularity-skewed toward older, widely seen titles, while more recent or niche works receive insufficient exposure to attain similar biases. The result therefore validates that the bias term has been learned correctly, yet also highlights the need for latent factors to deliver personalised and genre-diverse recommendations beyond these universally favoured films.

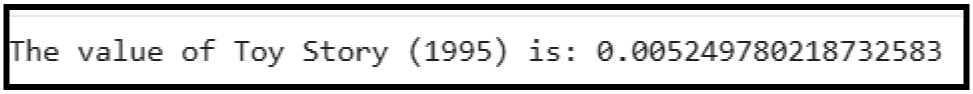
## Do you observe anything interesting from the movie embeddings?



From the scatter plot below of the item–embedding PCA, clear thematic groupings emerge: action-sci-fi hits such as The Matrix, T2 and Blade Runner crowd the upper-right, whereas romantic dramas and comedies like Titanic and When Harry Met Sally are pushed to the lower-left. Comedic cult favourites (Pulp Fiction, Airplane!) stretch far along the first axis, hinting that this component captures humour- or style-driven appeal. Meanwhile, Golden-Age classics (Casablanca, Wizard of Oz) sit apart from 1980-90s blockbusters, so the release era also shapes the space. These patterns confirm the embeddings have learned meaningful genre and era structure, supporting content-aware recommendations rather than just popularity.

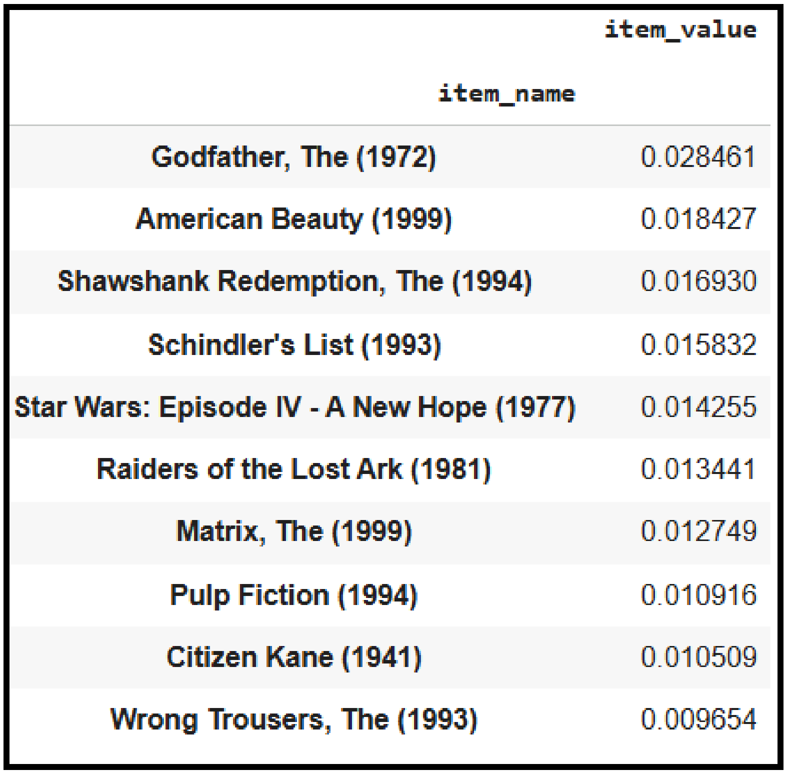
**What is your estimated value of *Toy Story (1995)*?**

The estimated value of Toy Story(1995) is approximately 5.2 Million. This estimated value uses data-driven methods that consider how much time users are likely to spend watching Toy Story, not just how many people have rated it.



```
The value of Toy Story (1995) is: 0.005249780218732583
```

**What are the top 10 mostly valued movies?**

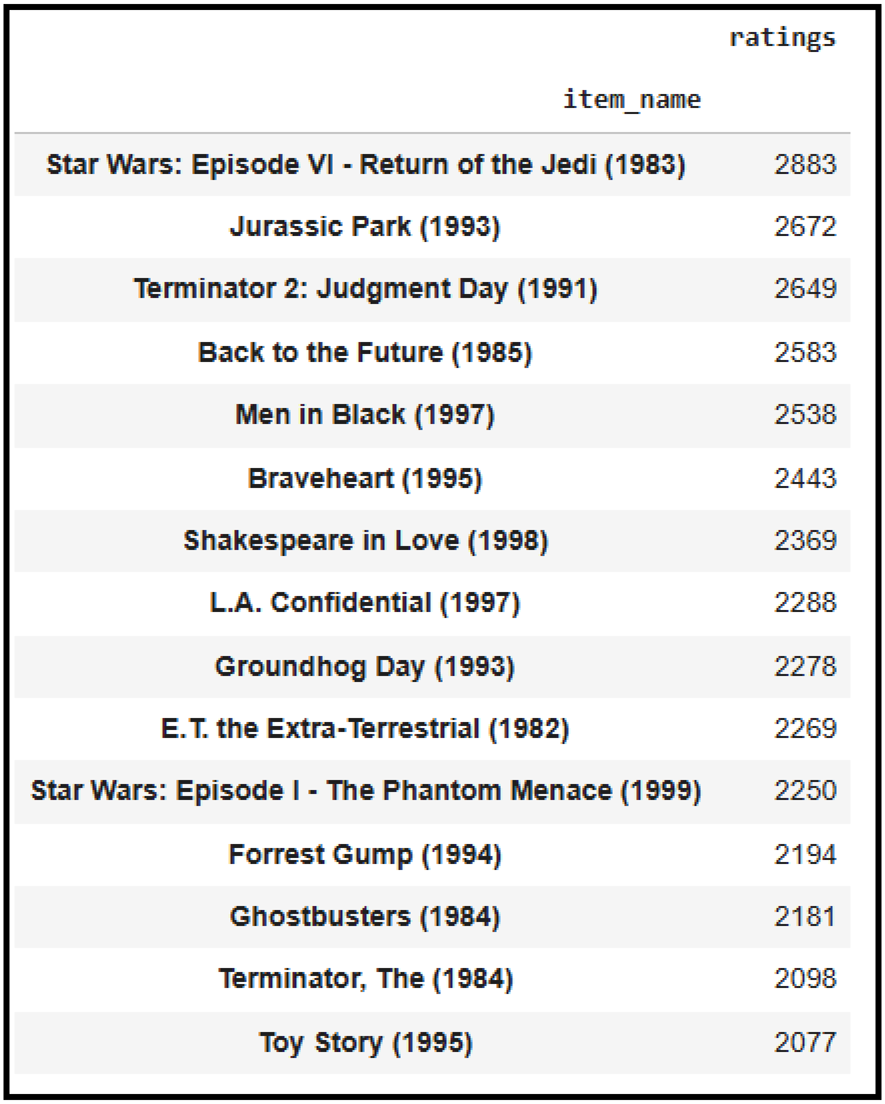


| item_name | item_value |
|---|---|
| Godfather, The (1972) | 0.028461 |
| American Beauty (1999) | 0.018427 |
| Shawshank Redemption, The (1994) | 0.016930 |
| Schindler's List (1993) | 0.015832 |
| Star Wars: Episode IV - A New Hope (1977) | 0.014255 |
| Raiders of the Lost Ark (1981) | 0.013441 |
| Matrix, The (1999) | 0.012749 |
| Pulp Fiction (1994) | 0.010916 |
| Citizen Kane (1941) | 0.010509 |
| Wrong Trousers, The (1993) | 0.009654 |

Here are the top 10 mostly valued movies: The Godfather, American Beauty, Shawshank Redemption, Schindler’s List, Star Wars: Episode IV, Raiders of the Lost Ark, The Matrix, Pulp Fiction, Citizen Kane, and The Wrong Trousers.

The model selected these movies because it predicts that users will spend more time watching them, making them highly valuable for a streaming platform. This ranking doesn’t just rely on popularity but it uses advanced deep learning to understand what users would likely enjoy most.

**What are the movies that are top 30 mostly rated but not in top 30 valued?**



| | ratings |
|---|---|
| item_name | |
| Star Wars: Episode VI - Return of the Jedi (1983) | 2883 |
| Jurassic Park (1993) | 2672 |
| Terminator 2: Judgment Day (1991) | 2649 |
| Back to the Future (1985) | 2583 |
| Men in Black (1997) | 2538 |
| Braveheart (1995) | 2443 |
| Shakespeare in Love (1998) | 2369 |
| L.A. Confidential (1997) | 2288 |
| Groundhog Day (1993) | 2278 |
| E.T. the Extra-Terrestrial (1982) | 2269 |
| Star Wars: Episode I - The Phantom Menace (1999) | 2250 |
| Forrest Gump (1994) | 2194 |
| Ghostbusters (1984) | 2181 |
| Terminator, The (1984) | 2098 |
| Toy Story (1995) | 2077 |

Popular titles like "Star Wars: Episode VI - Return of the Jedi," "Jurassic Park," "Back to the Future," and "Toy Story (1995)" received thousands of ratings from users. However, some of these highly rated movies did not make it into the top 30 most valued list.

This means that while these movies had lots of viewers and received a large number of ratings, the collaborative filtering model estimated that users may not spend as much time watching or re-watching them compared to movies in the top valued list. The valuation process looks beyond simple popularity and takes into account the predicted engagement, which is based on personalized user preferences rather than just rating counts. As a result, a movie can be extremely popular but may not rank as highly in terms of estimated value if the model predicts lower engagement or interest among the targeted users. This distinction highlights how collaborative filtering provides a deeper measure of a movie's true worth for a streaming platform, focusing on genuine user engagement rather than just widespread attention

**What do you think about this approach of movie valuation? How does it compare with the approach based on the popularity of the movies?**

The movie valuation approach using collaborative filtering is much better than popularity based approach because it looks at the user's personal preferences instead of just counting how many people rated a movie. It uses hidden patterns in user behavior to guess which movies a user will like, even if they haven’t watched them yet. This way, it predicts the value of movies based on how much time users might spend watching them, giving a more accurate idea of a movie’s worth to a streaming service.

In contrast, the popularity based approach simply ranks movies by the number of ratings or average scores, which only shows general interest but doesn’t capture individual preferences. So, the collaborative filtering valuation is more precise and personalized, which helps platforms like Netflix and Amazon Prime make better decisions about which movies to invest in or promote, while popularity based metrics provide simpler, broader measures of interest.