# Capstone Project 1
# Movie Recommender System

Capstone Project 1
Samin Rastgoufard, Ph.D.

Mentor: Rahul Sagrolikar

Springboard
Data Science Career Track

December 2018

# Outline




Capstone Project 1

Overview

Business Objective

Data

Questions

Workflow

MovieLens Dataset

EDA

Recommender System

Simple Recommender

IMDB Weighted Rating Formula

Content-based filtering

Collaborative Filtering

MetaData Dataset

Data Wrangling

EDA

Recommender System

Simple Recommender

IMDB Weighted Rating Formula

Matrix Factorization-based algorithms

Machine Learning

Conclusion

# Overview

- Nowadays, we need recommender systems almost everywhere in our lives.
- Therefore, retailers are become more interested in recommender systems to analyze patterns of user interest in products and provide personalized recommendations.
- The first goal of this project is understanding, analyzing, and correlating the trend in average rating movies of different genres.
- The second goal is building recommender engines to provide recommendations to different users and build different machine learning models to predict the rating of each movie.

# Business Objective

- The recommender system is useful to any business that makes money via recommendations.
- Since we want to work with movie dataset, the client of this project could be Amazon, Netflix, Hulu, HBO, etc.
- Giving good recommendations will help users spend less time searching for their type of movie and having a recommender platform.
- This will help the customer to continue with the service and having a good experience.

# Data

- The first dataset is collected from a free, noncommercial movie recommender as MoviLens.org. (https://grouplens.org/datasets/movielens). This recommender is like Netflix, minus the ability to watch movies. The dataset which was released in April 2015 contains 20 million ratings.
- The second dataset is collected from (https://www.kaggle.com/karrrimba/movie-metadatacsv). We have more features in this dataset like information about revenue, budget, release date, popularity,... .

# Questions

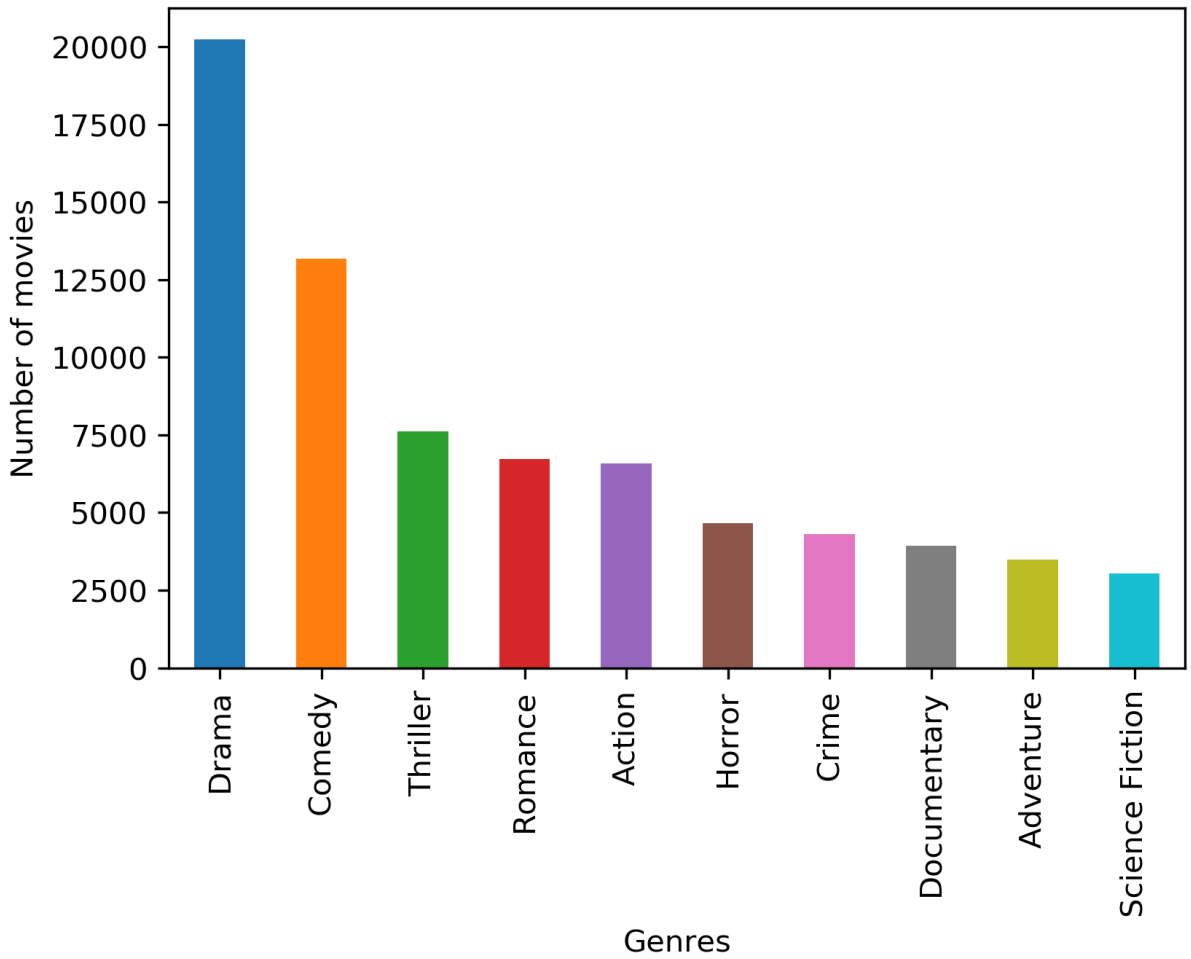
- How to get the insight about how peoples rating change for different genres over the years?
- How to recommend movies?
- How to predict movie ratings?

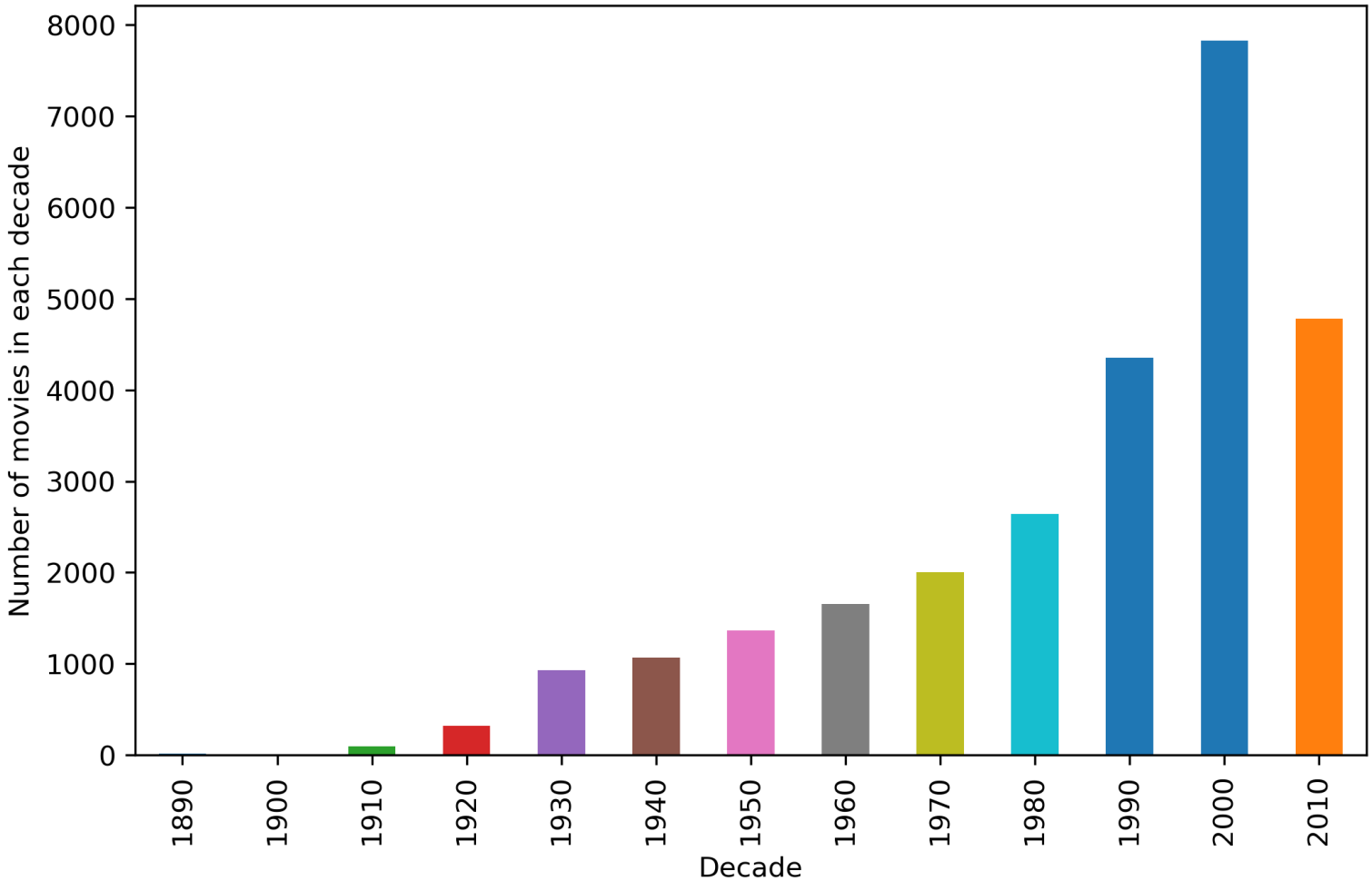
# Workflow

- Collecting data and applying data wrangling methods
- Starting exploratory data analysis to find trends and storytelling
- Conduct further data analysis to identify relationships between different variables
- Perform in-depth analysis using collaborative filtering and machine learning techniques to recommend and predict
- Conclusion and Future works

# First Dataset: MovieLens

# Exploratory Data Analysis (EDA) and Storytelling



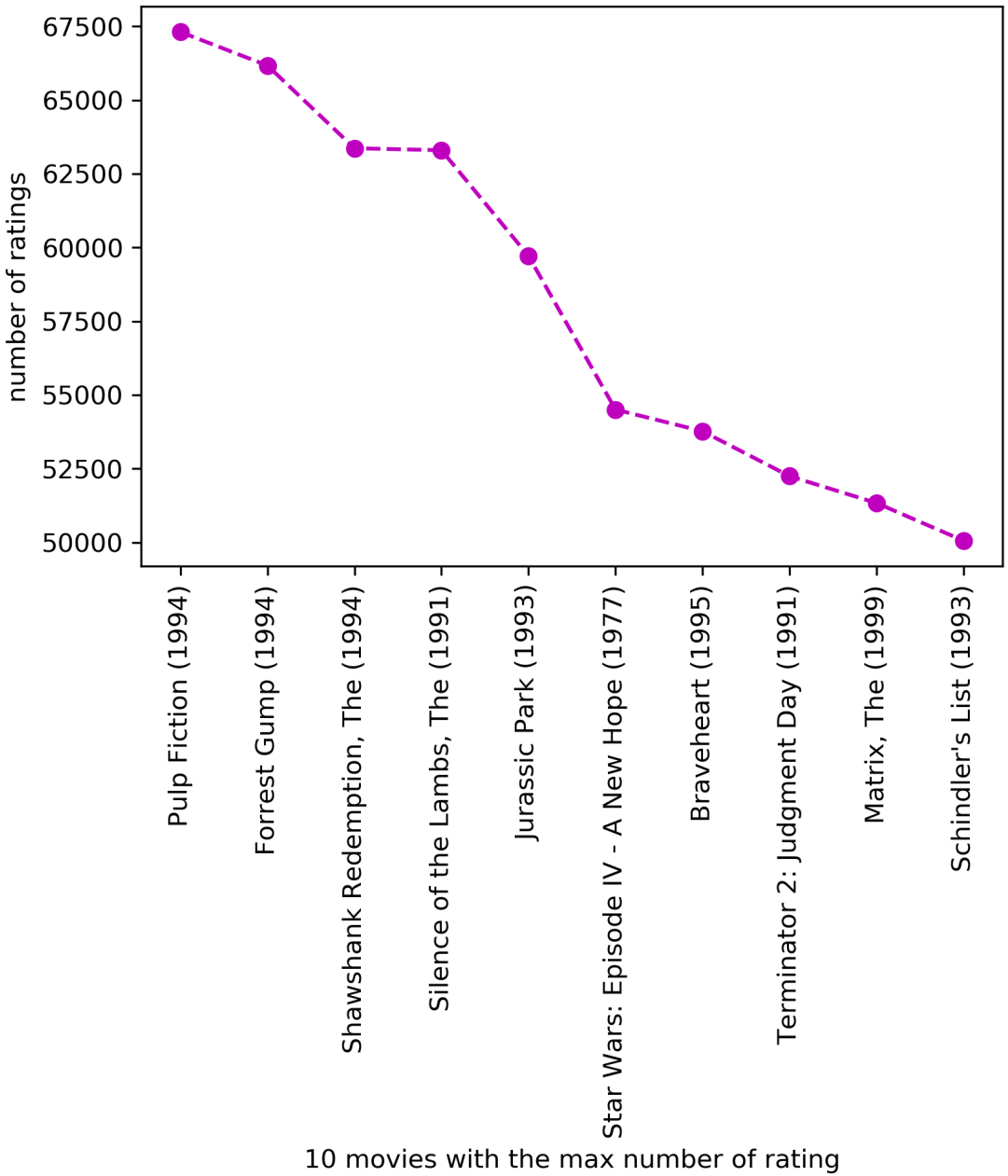


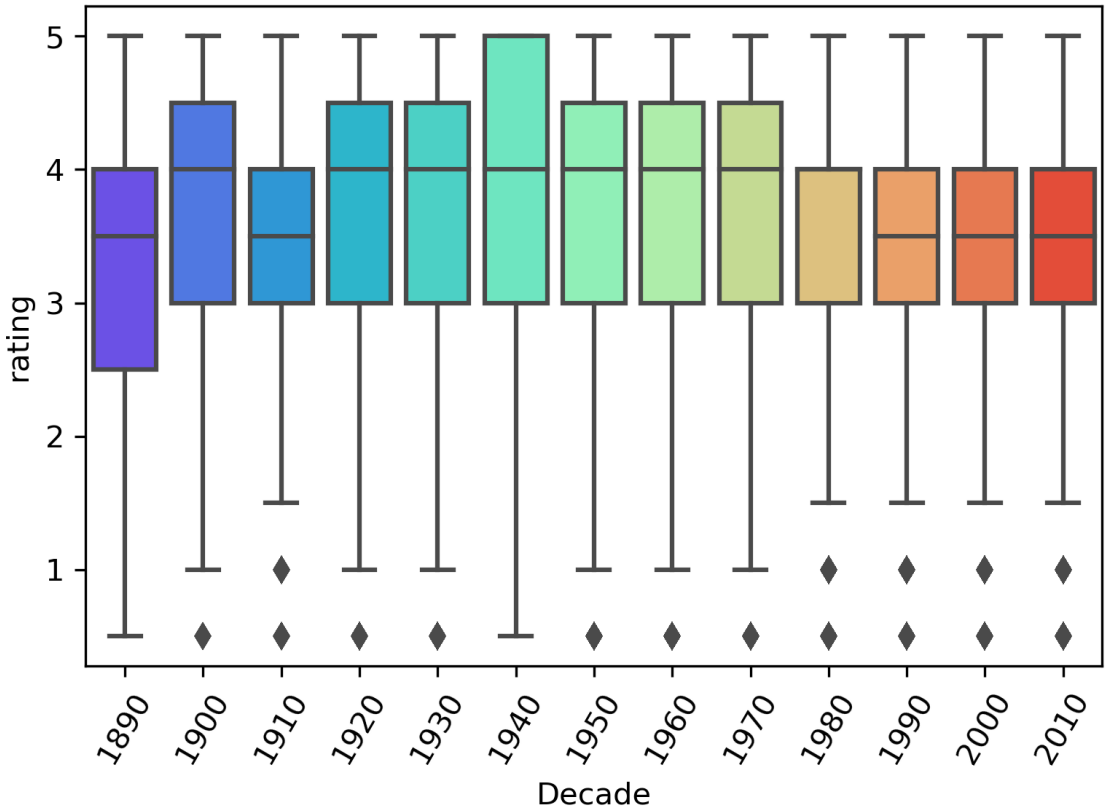
Movies with highest rankings are:

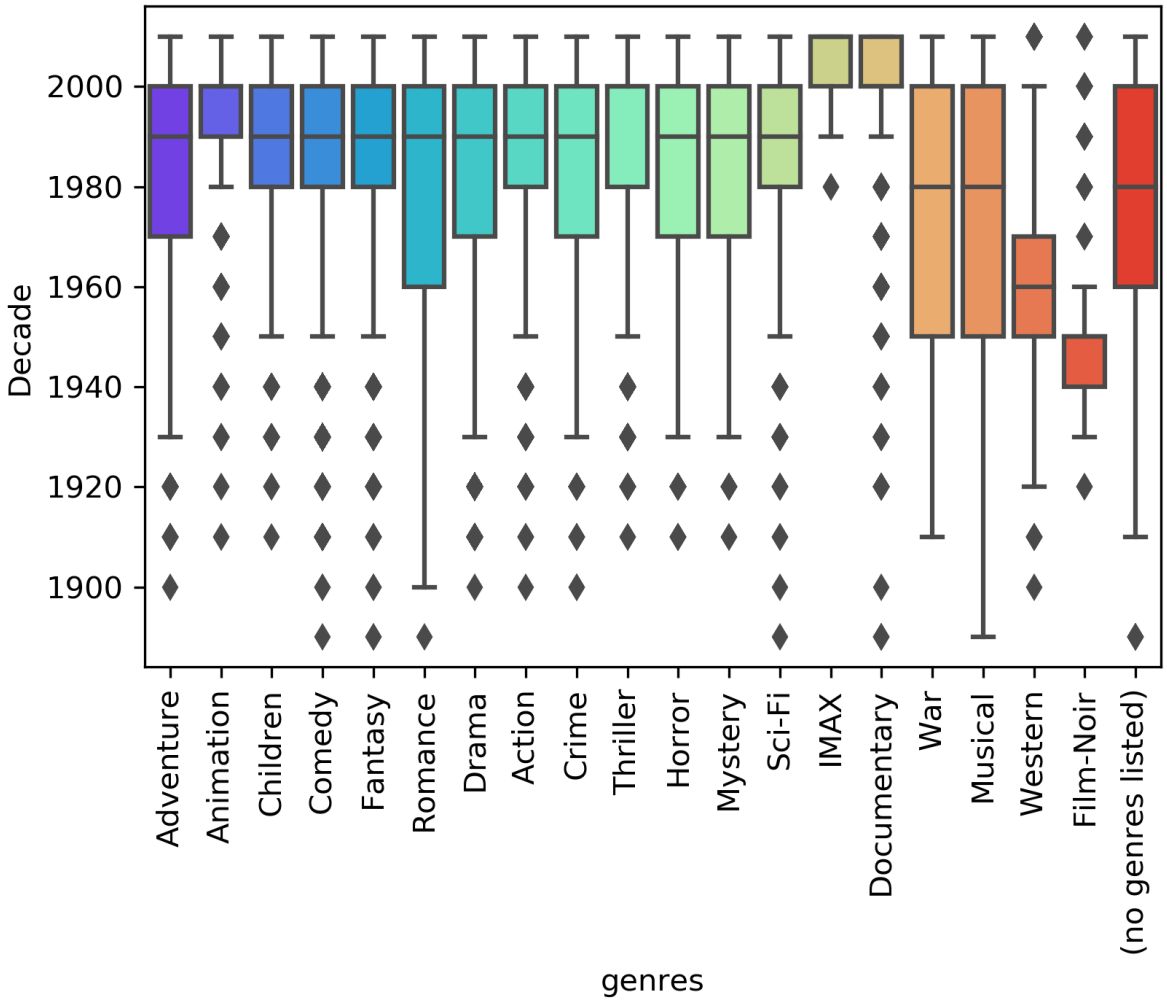
| | |
|---|---|
| Rate: 4.000000 | Silence of the Lambs, The (1991) |
| | Fugitive, The (1993) |
| | Forrest Gump (1994) |
| | Jurassic Park (1993) |
| | Terminator 2: Judgment Day (1991) |
| | Apollo 13 (1995) |
| | Toy Story (1995) |
| | Pulp Fiction (1994) |
| | Twelve Monkeys (a.k.a. 12 Monkeys) (1995) |
| | Braveheart (1995) |

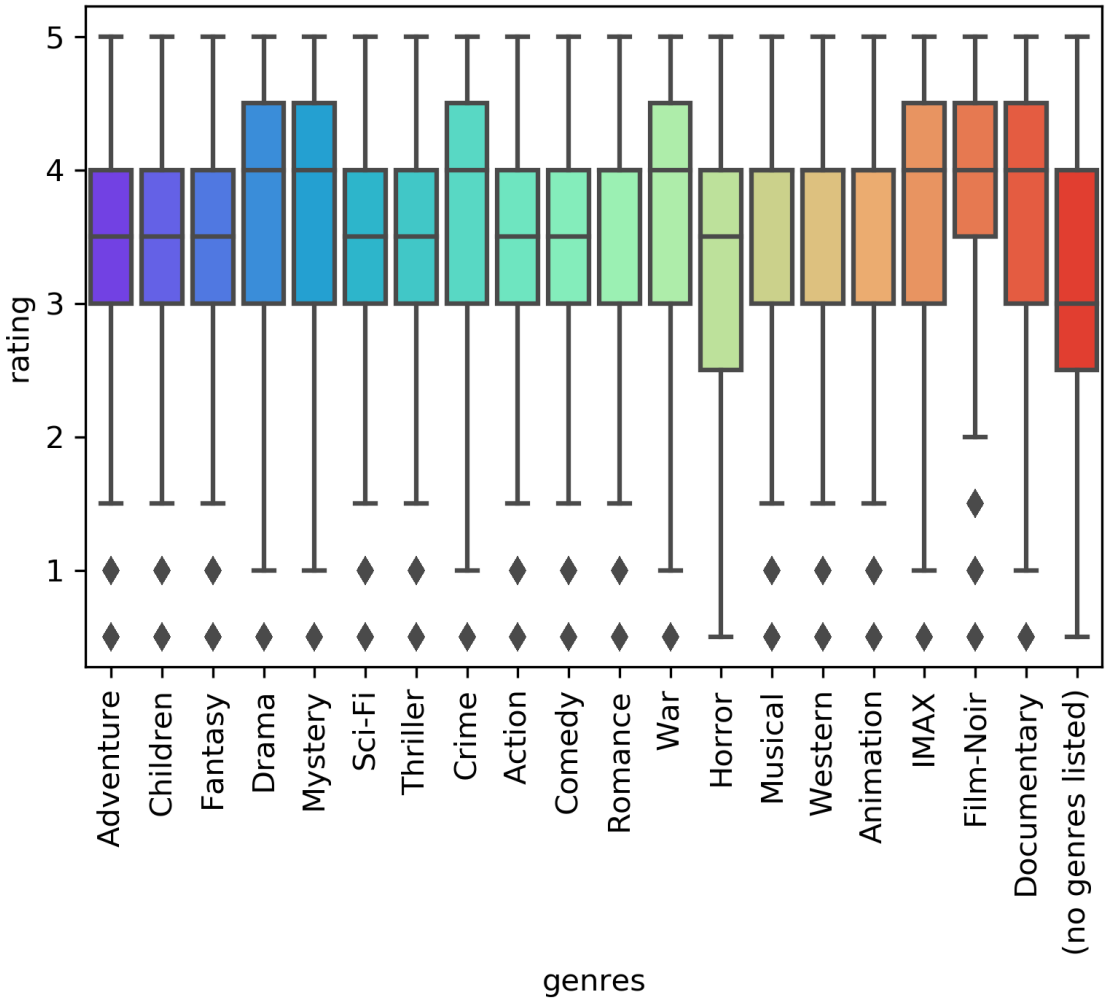
| | |
|---|---|
| Rate: 4.500000 | Shawshank Redemption, The (1994) |
| | Matrix, The (1999) |
| | Pulp Fiction (1994) |
| | Fight Club (1999) |
| | Lord of the Rings: The Fellowship of the Ring,... |
| | Lord of the Rings: The Two Towers, The (2002) |
| | Lord of the Rings: The Return of the King, The... |
| | Usual Suspects, The (1995) |
| | Silence of the Lambs, The (1991) |
| | Forrest Gump (1994) |

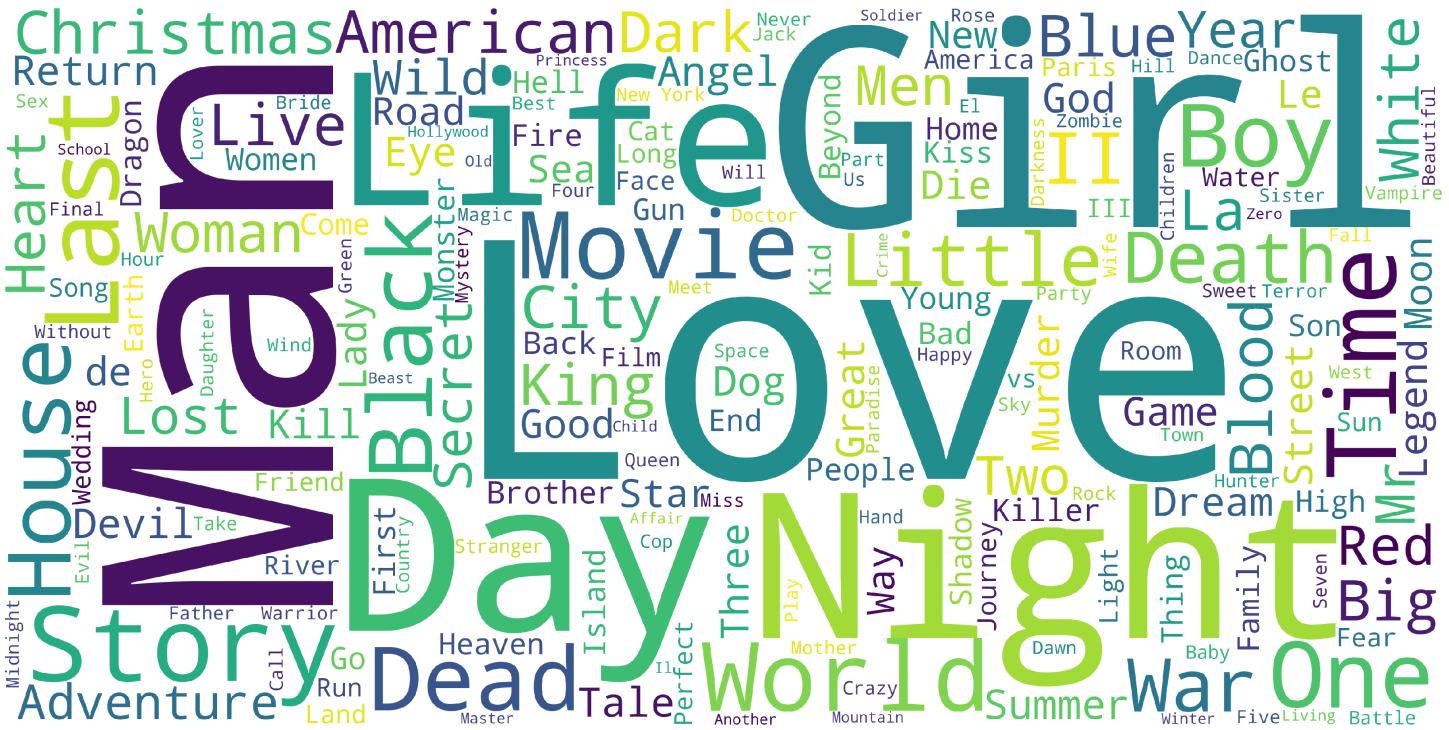
| | |
|---|---|
| Rate: 5.000000 | Shawshank Redemption, The (1994) |
| | Pulp Fiction (1994) |
| | Silence of the Lambs, The (1991) |
| | Schindler's List (1993) |
| | Star Wars: Episode IV - A New Hope (1977) |
| | Forrest Gump (1994) |
| | Godfather, The (1972) |
| | Usual Suspects, The (1995) |
| | Matrix, The (1999) |
| | Braveheart (1995) |











The list of the 10 best Drama movies to recommend are:

```
1 : Pulp Fiction (1994)
2 : Shawshank Redemption, The (1994)
3 : Schindler's List (1993)
4 : American Beauty (1999)
5 : Fargo (1996)
6 : Godfather, The (1972)
7 : Fight Club (1999)
8 : Lord of the Rings: The Return of the King, The (2003)
9 : One Flew Over the Cuckoo's Nest (1975)
10 : Godfather: Part II, The (1974)
```

The list of the 10 best Romance movies to recommend are:

```
1 : Forrest Gump (1994)
2 : Beauty and the Beast (1991)
3 : Princess Bride, The (1987)
4 : Groundhog Day (1993)
5 : Shrek (2001)
6 : Sleepless in Seattle (1993)
7 : Good Will Hunting (1997)
8 : Four Weddings and a Funeral (1994)
9 : Crouching Tiger, Hidden Dragon (Wo hu cang long) (2000)
10 : There's Something About Mary (1998)
```

The list of the 10 best Action movies to recommend are:

```
1 : Star Wars: Episode IV - A New Hope (1977)
2 : Braveheart (1995)
3 : Matrix, The (1999)
4 : Star Wars: Episode VI - Return of the Jedi (1983)
5 : Star Wars: Episode V - The Empire Strikes Back (1980)
6 : Raiders of the Lost Ark (Indiana Jones and the Raiders of the Lost Ark) (1981)
7 : Fight Club (1999)
8 : Saving Private Ryan (1998)
9 : Princess Bride, The (1987)
10 : Lord of the Rings: The Return of the King, The (2003)
```

# IMDB Weighted Rating Formula

| | |
|---|---|
| 0 | Shawshank Redemption, The (1994) |
| 1 | Usual Suspects, The (1995) |
| 2 | Godfather, The (1972) |
| 3 | Schindler's List (1993) |
| 4 | Fight Club (1999) |
| 5 | Pulp Fiction (1994) |
| 6 | Star Wars: Episode IV - A New Hope (1977) |
| 7 | Silence of the Lambs, The (1991) |
| 8 | Matrix, The (1999) |
| 9 | Star Wars: Episode V - The Empire Strikes Back... |
| 10 | North by Northwest (1959) |
| 11 | Princess Bride, The (1987) |
| 12 | Godfather: Part II, The (1974) |
| 13 | Raiders of the Lost Ark (Indiana Jones and the... |
| 14 | Dark Knight, The (2008) |
| 15 | City of God (Cidade de Deus) (2002) |
| 16 | Monty Python and the Holy Grail (1975) |
| 17 | Casablanca (1942) |
| 18 | Lord of the Rings: The Return of the King, The... |
| 19 | Dr. Strangelove or: How I Learned to Stop Worr... |
| 20 | Inception (2010) |

| | |
|---|---|
| 20 | Inception (2010) |
| 21 | Chinatown (1974) |
| 22 | American Beauty (1999) |
| 23 | Life Is Beautiful (La Vita è bella) (1997) |
| 24 | Fargo (1996) |
| 25 | Wallace & Gromit: The Wrong Trousers (1993) |
| 26 | Seven Samurai (Shichinin no samurai) (1954) |
| 27 | Memento (2000) |
| 28 | Rear Window (1954) |
| 29 | Blade Runner (1982) |

# Content-based filtering

Content based filtering Recommend based on the user's rating history.

$$r_{u,i} = \text{aggr}_{i' \in I(u)}[r_{u,i'}]$$

A simple example using the mean as an aggregation function:

$$r_{u,i} = \bar{r}_u = \frac{\sum_{i' \in I(u)} r_{u,i'}}{|I(u)|}$$

RMSE for content base filtering is: 0.9969
Total elapsed time of content base filtering is: 65.319

# Collaborative Filtering

Collaborative Filtering is based on the idea that users similar to a me can be used to predict how much I will like a particular product or service those users have used/experienced but I have not.

**I- Simple Collaborative filtering base on mean**

Recommend based on other user's rating histories.

Generic expression (notice how this is kind of a 'col-based' approach):

$$r_{u,i} = \text{aggr}_{u' \in U(i)}[r_{u',i}]$$

A simple example using the mean as an aggregation function:

$$r_{u,i} = \bar{r}_i = \frac{\sum_{u' \in U(i)} r_{u',i}}{|U(i)|}$$

```
Collaborative mean filter for userId: 1 and movieId: 151 is: [3.31060606]
Actual rating value is: 4.0

Collaborative mean filter for userId: 1 and movieId: 29 is: [3.80232558]
Actual rating value is: 3.5

Collaborative mean filter for userId: 138493 and movieId: 55269 is: [4.06666667]
Actual rating value is: 5.0
```

# Collaborative Filtering

**II- Matrix Factorization-based algorithms**

The idea is basically to take a large (or potentially huge) matrix and factor it into some smaller representation of the original matrix. You can think of it in the same way as we would take a large number and factor it into two much smaller primes. We end up with two or more lower dimensional matrices whose product equals the original one. When we talk about collaborative filtering for recommender systems we want to solve the problem of our original matrix having millions of different dimensions, but our tastes not being nearly as complex.

I use **Surprise** library which has several powerful algorithms like **Singular Value Decomposition (SVD), Non-negative Matrix Factorization (NMF), K Nearest Neighbor (KNN), and CoClusternig** to minimise RMSE (Root Mean Square Error) and give recommendations.

# SVD

```
Evaluating RMSE, MAE of algorithm SVD on 5 split(s).

                  Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)    0.9171  0.9140  0.9113  0.9219  0.9097  0.9148  0.0043
MAE (testset)     0.7087  0.7049  0.7019  0.7094  0.7012  0.7052  0.0034
Fit time          13.00   13.68   12.76   43.84   21.19   20.89   11.89
Test time         0.30    0.39    0.27    1.97    0.73    0.73    0.64
Total Elapsed time with model is: 110.37742877006531


SVD
Total Elapsed time with model is: 9.202
Prediction of a particular user: 3.69487816192052
Actual rating value is: 4.0

Prediction of a particular user: 3.6717143007180093
Actual rating value is: 3.5

Prediction of a particular user: 3.860117330365739
Actual rating value is: 5.0
```

# Alternating Least Squares (ALS)

III- Alternating Least Squares (ALS) Collaborative Filtering

- Alternating Least Squares (ALS) is a the model well use to fit our data and find similarities.
- ALS is an iterative optimization process where we for every iteration try to arrive closer and closer to a factorized representation of our original data.
- We have our original matrix R of size u x i with our users, items and some type of feedback data.
- We then want to find a way to turn that into one matrix with users and hidden features of size u x f and one with items and hidden features of size f x i.
- In U and V we have weights for how each user/item relates to each feature. What we do is we calculate U and V so that their product approximates R as closely as possible: R  U x V.

Similar movies to **Shawshank Redemption**:

| MovieId | Score | Title |
|---|---|---|
| 318 | 0.302835 | Shawshank Redemption, The (1994) |
| 527 | 0.301579 | Schindler's List (1993) |
| 593 | 0.301338 | Silence of the Lambs, The (1991) |
| 110 | 0.300669 | Braveheart (1995) |
| 296 | 0.300460 | Pulp Fiction (1994) |
| 50 | 0.300460 | Usual Suspects, The (1995) |
| 356 | 0.300159 | Forrest Gump (1994) |
| 47 | 0.298370 | Seven (a.k.a. Se7en) (1995) |
| 480 | 0.298036 | Jurassic Park (1993) |
| 589 | 0.296039 | Terminator 2: Judgment Day (1991) |

Recommended movies to **userId=1**:

| MovieId | Score | Title |
|---|---|---|
| 594 | 1.362166 | Snow White and the Seven Dwarfs (1937) |
| 2019 | 1.331825 | Seven Samurai (Shichinin no samurai) (1954) |
| 3703 | 1.304854 | Road Warrior, The (Mad Max 2) (1981) |
| 720 | 1.297422 | Wallace & Gromit: The Best of Aardman Animatio... |
| 1375 | 1.285937 | Star Trek III: The Search for Spock (1984) |
| 1282 | 1.277820 | Fantasia (1940) |
| 551 | 1.234990 | Nightmare Before Christmas, The (1993) |
| 2657 | 1.218386 | Rocky Horror Picture Show, The (1975) |
| 2160 | 1.211304 | Rosemary's Baby (1968) |
| 1748 | 1.211127 | Dark City (1998) |

# Second Dataset: MetaData

# MetaData: Features

**adult**: Indicates if the movie is X-Rated or Adult.
**belongs to collection**: A stringified dictionary that gives information on the movie series the particular film belongs to.
**budget**: The budget of the movie in dollars.
**genres**: A stringified list of dictionaries that list out all the genres associated with the movie.
**homepage**:The Official Homepage of the move.
**id**:The ID of the move.
**imdb id**: The IMDB ID of the movie.
**original language**: The language in which the movie was originally shot in.
**original title**: The original title of the movie.
**overview**: A brief blurb of the movie.
**popularity**: The Popularity Score assigned by TMDB.
**poster path**: The URL of the poster image.
**production companies**: A stringified list of production companies involved with the making of the movie.
**production countries**: A stringified list of countries where the movie was shot/produced in.
**release date**:Theatrical Release Date of the movie.
**revenue**: The total revenue of the movie in dollars.
**runtime**: The runtime of the movie in minutes.
**spoken languages**: A stringified list of spoken languages in the film.
**status**: The status of the movie (Released, To Be Released, Announced, etc.)
**tagline**: The tagline of the movie.
**title**: The Official Title of the movie.
**video**: Indicates if there is a video present of the movie with TMDB.
**vote average**: The average rating of the movie.
**vote count**: The number of votes by users, as counted by TMDB.

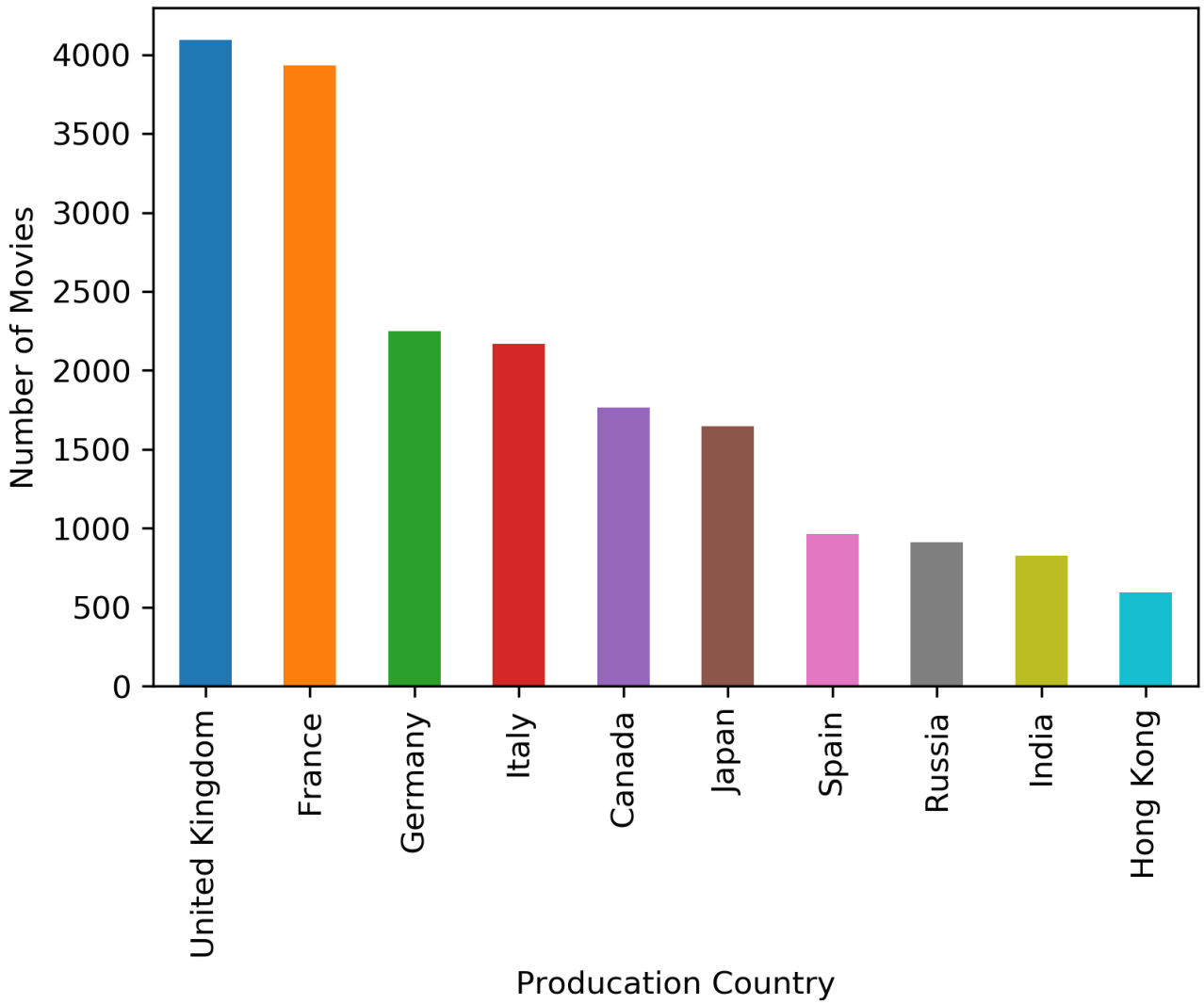
# Data Wrangling

For cleaning the data, I started with:

- dropping the duplicate rows. (13 duplicates)
- Just kept the rows wih False and True in adult column.
- I dropped the columns 'status','adult','homepage','imdb id','original title','poster path','tagline','video','spoken languages', since they are not providing important information for this project
- The budget and popularity columns have been changed to numeric.
- The release date has been changed to datetime format.
- year column is added for future use.
- I used the literal eval, apply and lambda to change the format of production companies, genres, and production countries columns.

# Production countries

USA total number of production is: 21,140



# Highest Earning Production Companies

| | Total | Average | Number |
|---|---|---|---|
| **Warner Bros.** | 6.352519e+10 | 5.082015e+07 | 1250 |
| **Universal Pictures** | 5.525919e+10 | 6.657734e+07 | 830 |
| **Paramount Pictures** | 4.876940e+10 | 4.872068e+07 | 1001 |
| **Twentieth Century Fox Film Corporation** | 4.768775e+10 | 5.704276e+07 | 836 |
| **Walt Disney Pictures** | 4.083727e+10 | 1.552748e+08 | 263 |
| **Columbia Pictures** | 3.227974e+10 | 7.489498e+07 | 431 |
| **New Line Cinema** | 2.217339e+10 | 8.004834e+07 | 277 |
| **Amblin Entertainment** | 1.734372e+10 | 2.282068e+08 | 76 |
| **DreamWorks SKG** | 1.547575e+10 | 1.629027e+08 | 95 |
| **Dune Entertainment** | 1.500379e+10 | 2.308275e+08 | 65 |

# Movies with the max revenue

| title | popularity | vote_average | revenue |
|---|---|---|---|
| Avatar | 185.070892 | 7.2 | 2.787965e+09 |
| Star Wars: The Force Awakens | 31.626013 | 7.5 | 2.068224e+09 |
| Titanic | 26.889070 | 7.5 | 1.845034e+09 |
| The Avengers | 89.887648 | 7.4 | 1.519558e+09 |
| Jurassic World | 32.790475 | 6.5 | 1.513529e+09 |
| Furious 7 | 27.275687 | 7.3 | 1.506249e+09 |
| Avengers: Age of Ultron | 37.379420 | 7.3 | 1.405404e+09 |
| Harry Potter and the Deathly Hallows: Part 2 | 24.990737 | 7.9 | 1.342000e+09 |
| Frozen | 24.248243 | 7.3 | 1.274219e+09 |
| Beauty and the Beast | 287.253654 | 6.8 | 1.262886e+09 |

# Movies with the maximum number of votes

| title | vote_count | year |
|---|---|---|
| Inception | 14075.0 | 2010 |
| The Dark Knight | 12269.0 | 2008 |
| Avatar | 12114.0 | 2009 |
| The Avengers | 12000.0 | 2012 |
| Deadpool | 11444.0 | 2016 |
| Interstellar | 11187.0 | 2014 |
| Django Unchained | 10297.0 | 2012 |
| Guardians of the Galaxy | 10014.0 | 2014 |
| Fight Club | 9678.0 | 1999 |
| The Hunger Games | 9634.0 | 2012 |

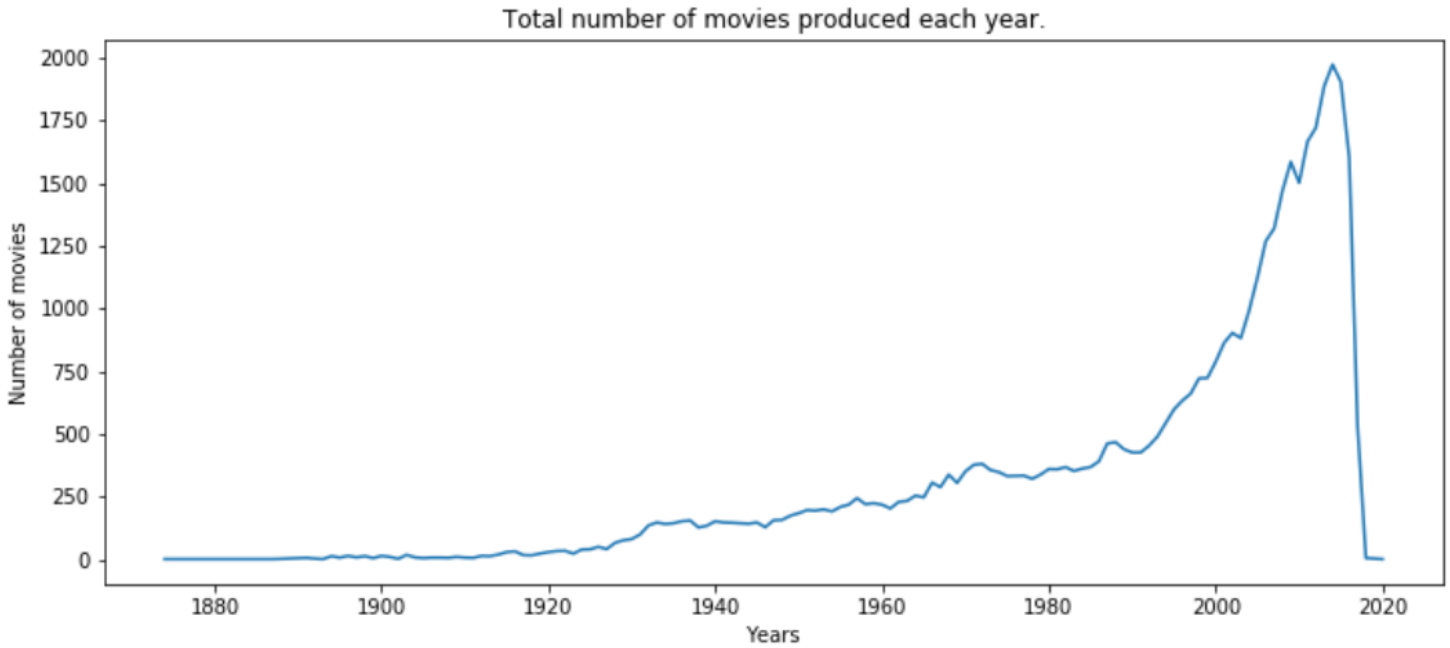
# Most popular movies

| title | popularity | vote_average | revenue |
|---|---|---|---|
| Minions | 547.488298 | 6.4 | 1.156731e+09 |
| Wonder Woman | 294.337037 | 7.2 | 8.205804e+08 |
| Beauty and the Beast | 287.253654 | 6.8 | 1.262886e+09 |
| Baby Driver | 228.032744 | 7.2 | 2.245113e+08 |
| Big Hero 6 | 213.849907 | 7.8 | 6.521054e+08 |
| Deadpool | 187.860492 | 7.4 | 7.831130e+08 |
| Guardians of the Galaxy Vol. 2 | 185.330992 | 7.6 | 8.634161e+08 |
| Avatar | 185.070892 | 7.2 | 2.787965e+09 |
| John Wick | 183.870374 | 7.0 | 8.876166e+07 |
| Gone Girl | 154.801009 | 7.9 | 3.693304e+08 |

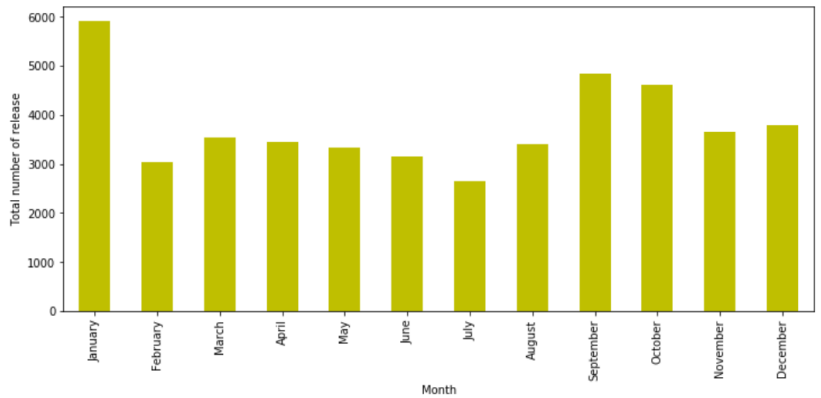
Movies with more than 2000 votes and highest vote average are listed bellow.

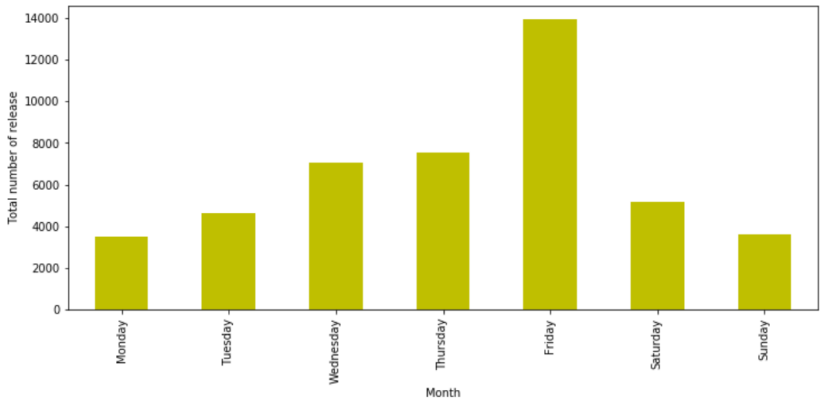
| title | vote_average | vote_count | year |
|---|---|---|---|
| The Shawshank Redemption | 8.5 | 8358.0 | 1994 |
| The Godfather | 8.5 | 6024.0 | 1972 |
| Life Is Beautiful | 8.3 | 3643.0 | 1997 |
| Spirited Away | 8.3 | 3968.0 | 2001 |
| One Flew Over the Cuckoo's Nest | 8.3 | 3001.0 | 1975 |
| Psycho | 8.3 | 2405.0 | 1960 |
| Fight Club | 8.3 | 9678.0 | 1999 |
| The Godfather: Part II | 8.3 | 3418.0 | 1974 |
| The Dark Knight | 8.3 | 12269.0 | 2008 |
| Pulp Fiction | 8.3 | 8670.0 | 1994 |

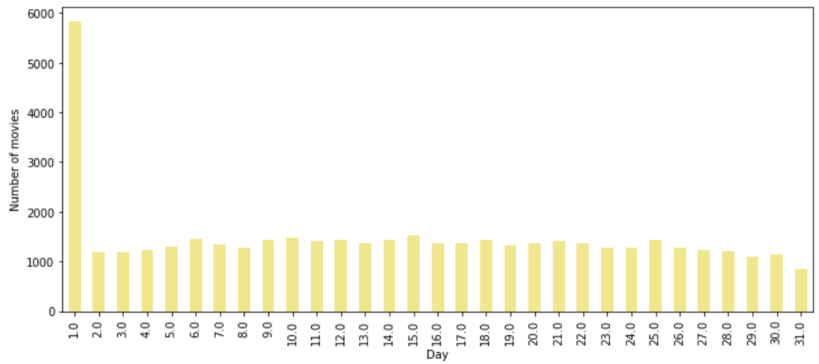
# Release time



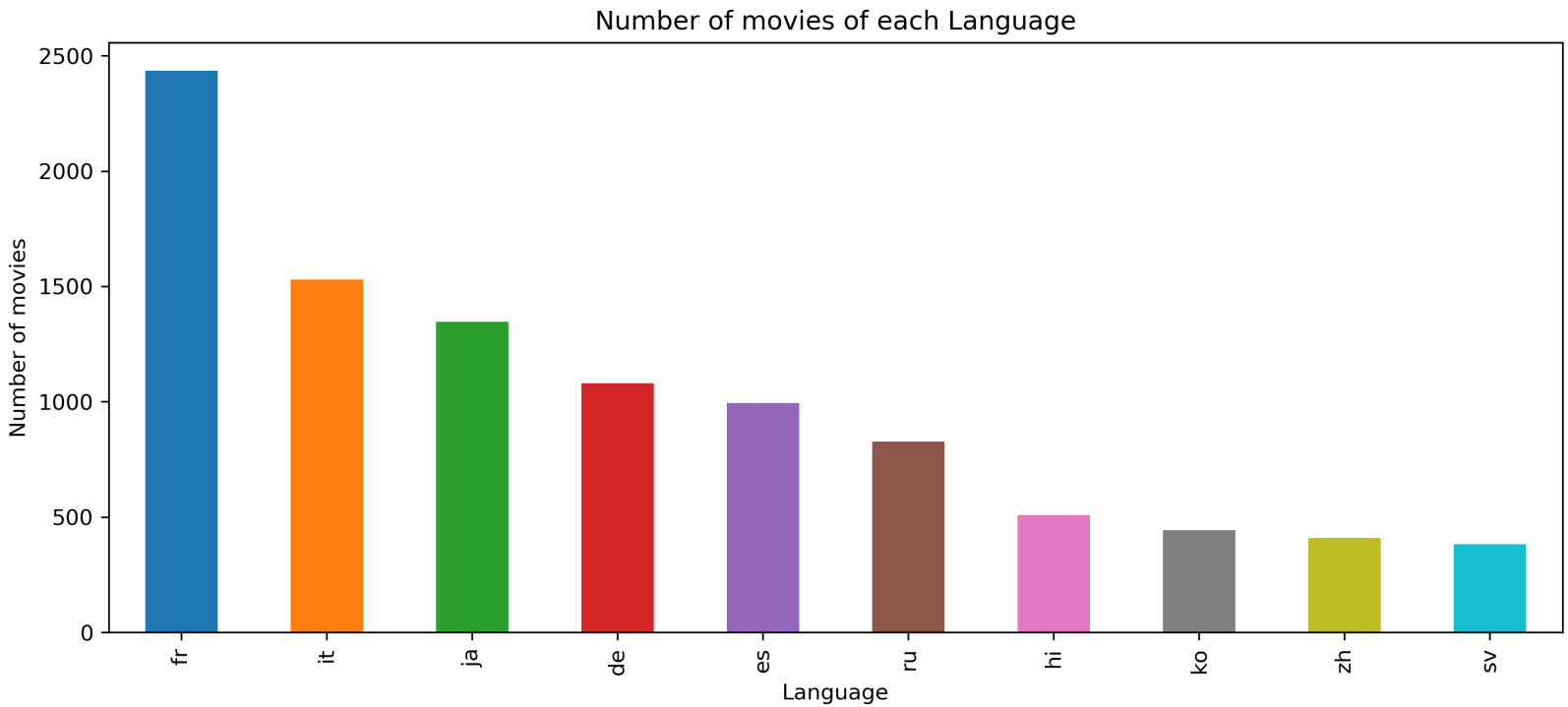
# Release time



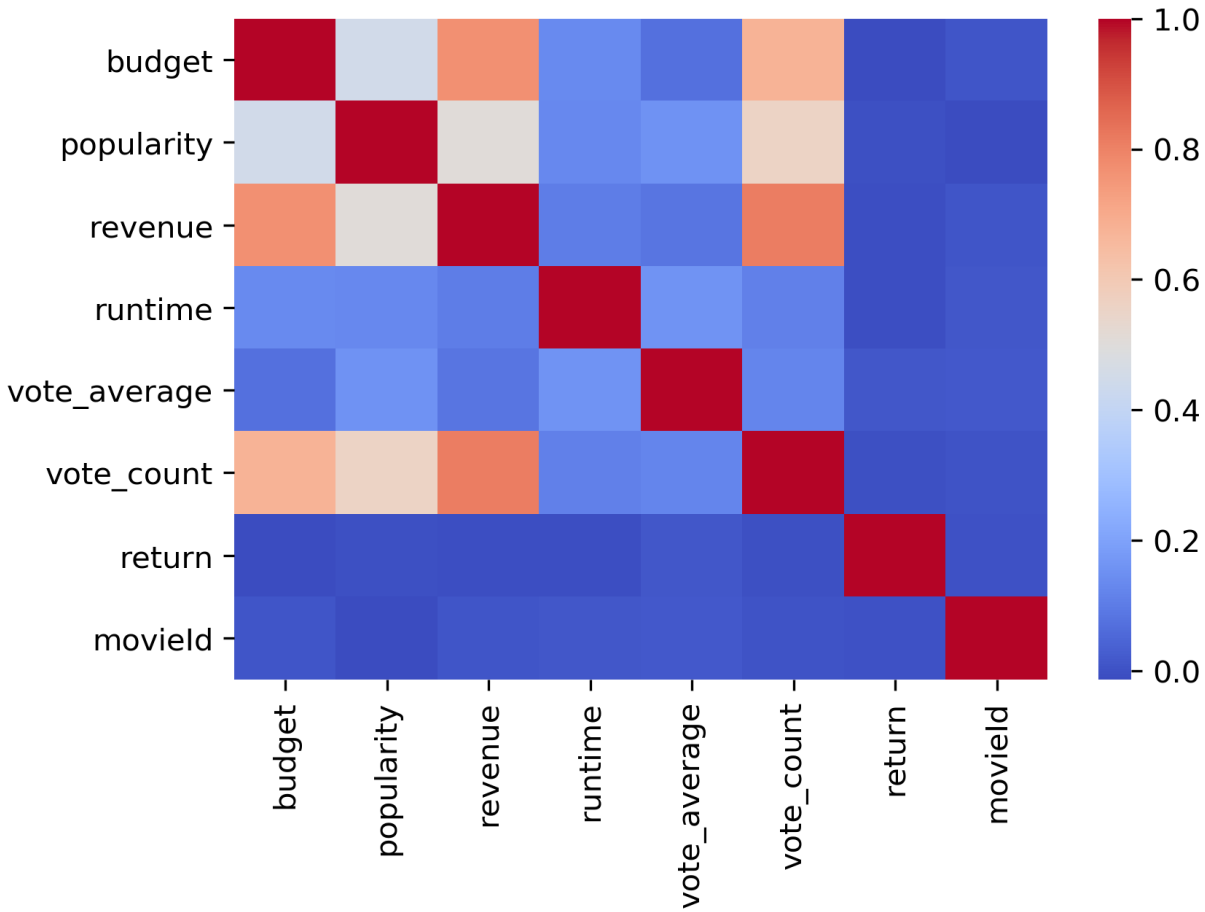



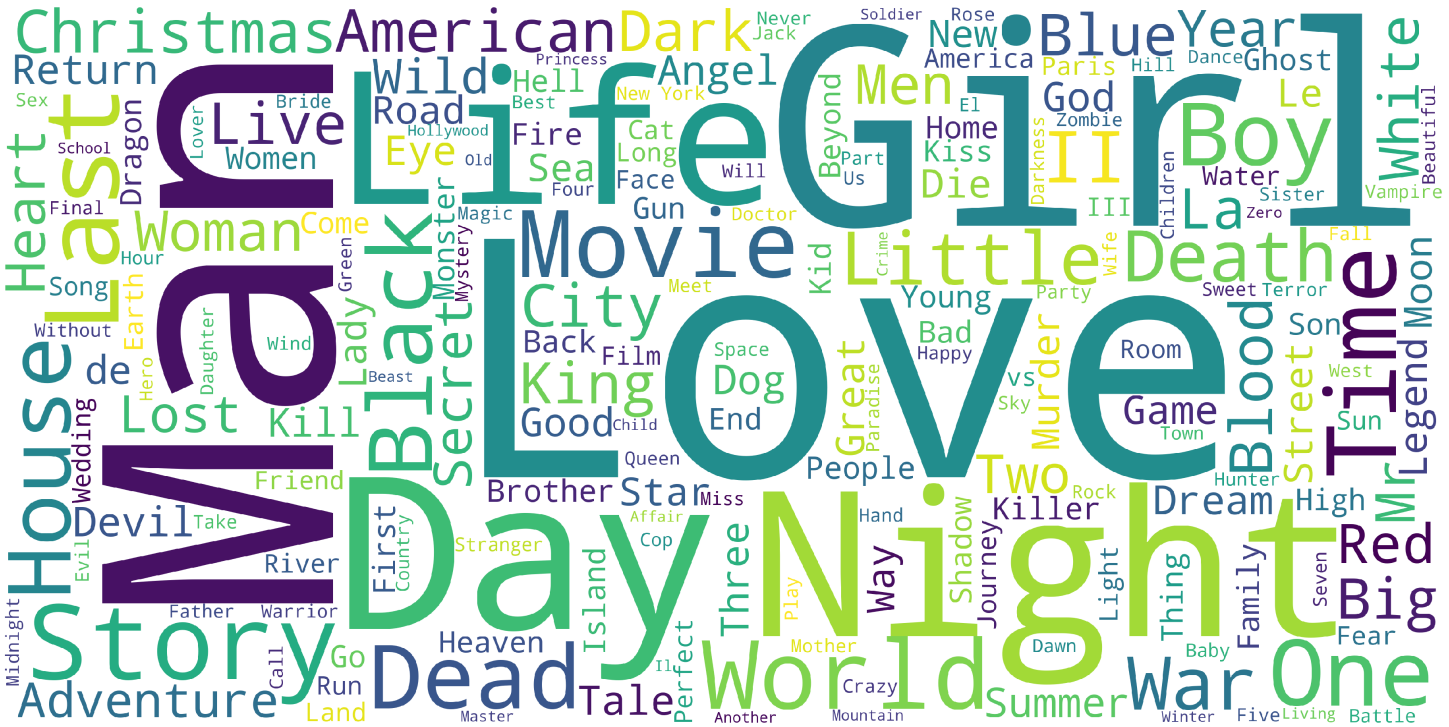
# Language



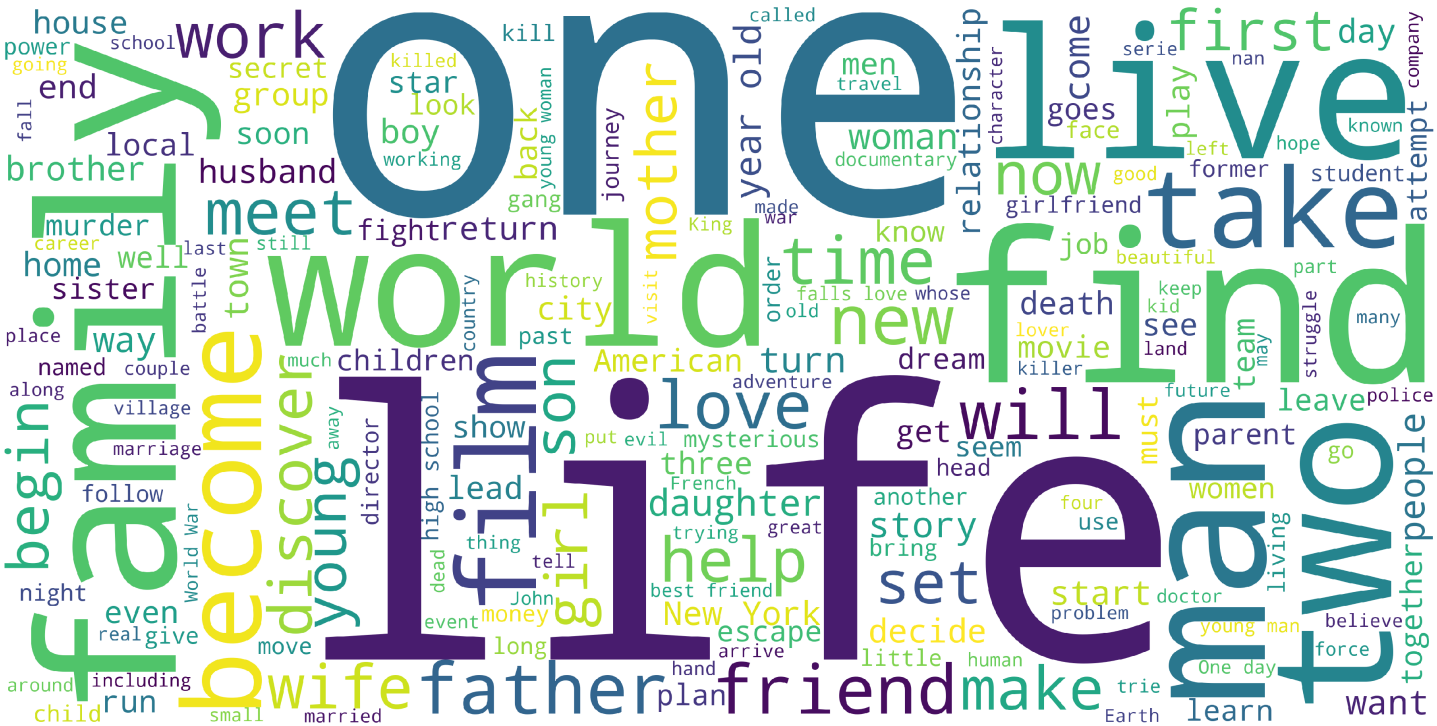
# Correlation



The word Love is the most commonly used word in movie titles. Girl, Day and Man are also among the most commonly occurring words.

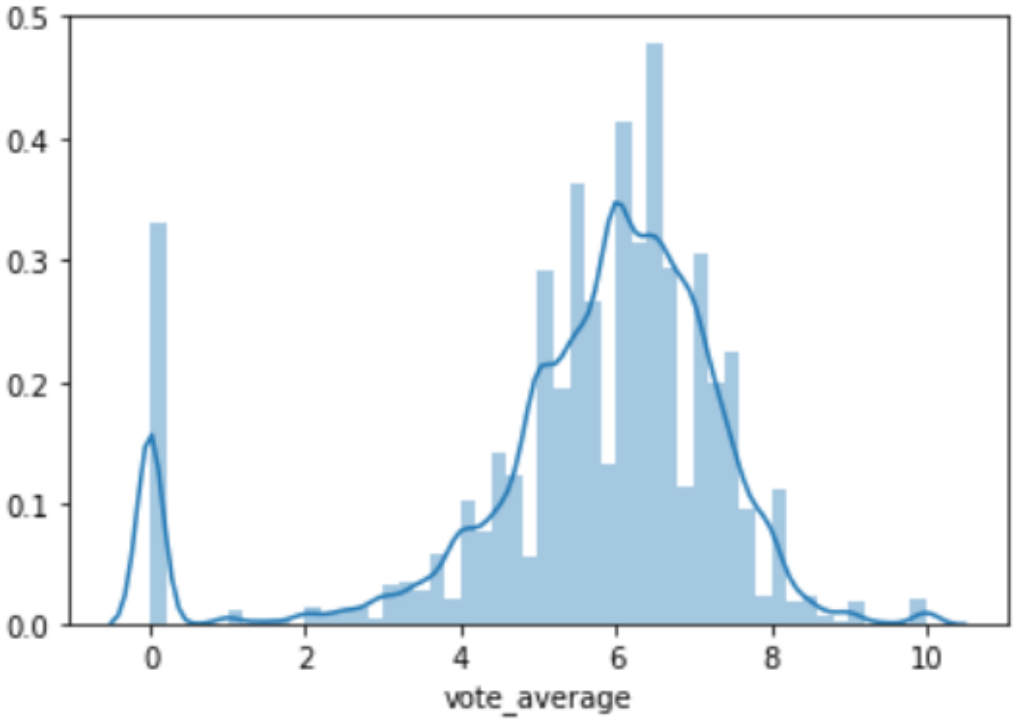


Life is the most commonly used word in Movie titles. One and Find are also popular.



# Vote average

The mean rating is only a 5.6 on a scale of 10. Half of the movies have a rating of less than or equal to 6.



# Simple Recommender

The list of the 15 best Drama and Romantic movies to recommend are:

| | |
|---|---|
| 0 | The Dark Knight |
| 1 | Interstellar |
| 2 | Fight Club |
| 3 | The Shawshank Redemption |
| 4 | Forrest Gump |
| 5 | The Godfather |
| 6 | The Intouchables |
| 7 | Schindler's List |
| 8 | Whiplash |
| 9 | Leon: The Professional |
| 10 | The Green Mile |
| 11 | Life Is Beautiful |
| 12 | The Godfather: Part II |
| 13 | The Usual Suspects |
| 14 | GoodFellas |

| | |
|---|---|
| 0 | Forrest Gump |
| 1 | Titanic |
| 2 | La La Land |
| 3 | Her |
| 4 | The Great Gatsby |
| 5 | The Fault in Our Stars |
| 6 | Eternal Sunshine of the Spotless Mind |
| 7 | Edward Scissorhands |
| 8 | Aladdin |
| 9 | Amélie |
| 10 | The Theory of Everything |
| 11 | The Curious Case of Benjamin Button |
| 12 | The Notebook |
| 13 | A Beautiful Mind |
| 14 | The Perks of Being a Wallflower |

# Simple Recommender

The list of the 15 best Action and Fantasy movies to recommend are:

| | |
|---|---|
| 0 | Avatar |
| 1 | The Lord of the Rings: The Fellowship of the Ring |
| 2 | The Lord of the Rings: The Return of the King |
| 3 | Star Wars: The Force Awakens |
| 4 | The Lord of the Rings: The Two Towers |
| 5 | Pirates of the Caribbean: The Curse of the Bla... |
| 6 | Harry Potter and the Philosopher's Stone |
| 7 | X-Men: Days of Future Past |
| 8 | Harry Potter and the Deathly Hallows: Part 2 |
| 9 | Harry Potter and the Prisoner of Azkaban |
| 10 | Harry Potter and the Chamber of Secrets |
| 11 | Doctor Strange |
| 12 | Harry Potter and the Goblet of Fire |
| 13 | Harry Potter and the Deathly Hallows: Part 1 |
| 14 | Harry Potter and the Order of the Phoenix |

| | |
|---|---|
| 0 | Inception |
| 1 | The Dark Knight |
| 2 | Avatar |
| 3 | The Avengers |
| 4 | Deadpool |
| 5 | Guardians of the Galaxy |
| 6 | Mad Max: Fury Road |
| 7 | The Dark Knight Rises |
| 8 | The Matrix |
| 9 | Iron Man |
| 10 | The Lord of the Rings: The Fellowship of the Ring |
| 11 | The Lord of the Rings: The Return of the King |
| 12 | Star Wars: The Force Awakens |
| 13 | The Lord of the Rings: The Two Towers |
| 14 | Batman Begins |

# IMDB Weighted Rating Formula

Another technique could be IMDB's weighted rating formula which is mathematically represented as follows:

Weighted Rating (WR) $= (\frac{v}{v+m}.R) + (\frac{m}{v+m}.C)$

where,

$v$ is the number of votes for the movie

$m$ is the minimum votes required to be listed in the chart

$R$ is the average rating of the movie

$C$ is the mean vote across the whole report

| title | year | vote_count | vote_average | popularity | genres | wr |
|---|---|---|---|---|---|---|
| Inception | 2010 | 14075 | 8 | 29.108149 | [Action, Thriller, Science Fiction, Mystery, A... | 7.917592 |
| The Dark Knight | 2008 | 12269 | 8 | 123.167259 | [Drama, Action, Crime, Thriller] | 7.905876 |
| Interstellar | 2014 | 11187 | 8 | 32.213481 | [Adventure, Drama, Science Fiction] | 7.897113 |
| Fight Club | 1999 | 9678 | 8 | 63.869599 | [Drama] | 7.881759 |
| The Lord of the Rings: The Fellowship of the Ring | 2001 | 8892 | 8 | 32.070725 | [Adventure, Fantasy, Action] | 7.871793 |
| Pulp Fiction | 1994 | 8670 | 8 | 140.950236 | [Thriller, Crime] | 7.868667 |
| The Shawshank Redemption | 1994 | 8358 | 8 | 51.645403 | [Drama, Crime] | 7.864007 |
| The Lord of the Rings: The Return of the King | 2003 | 8226 | 8 | 29.324358 | [Adventure, Fantasy, Action] | 7.861934 |
| Forrest Gump | 1994 | 8147 | 8 | 48.307194 | [Comedy, Drama, Romance] | 7.860663 |
| The Lord of the Rings: The Two Towers | 2002 | 7641 | 8 | 29.423537 | [Adventure, Fantasy, Action] | 7.851931 |
| Star Wars | 1977 | 6778 | 8 | 42.149697 | [Adventure, Action, Science Fiction] | 7.834213 |
| Back to the Future | 1985 | 6239 | 8 | 25.778509 | [Adventure, Comedy, Science Fiction, Family] | 7.820822 |
| The Godfather | 1972 | 6024 | 8 | 41.109264 | [Drama, Crime] | 7.814857 |
| The Empire Strikes Back | 1980 | 5998 | 8 | 19.470959 | [Adventure, Action, Science Fiction] | 7.814108 |

# Matrix Factorization-based algorithms

In the first part, the vote average is predicted from the vote count and the popularity:

```
Evaluating RMSE, MAE of algorithm SVD on 5 split(s).

                  Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)    1.5976  1.5968  1.6044  1.5851  1.6242  1.6016  0.0129
MAE (testset)     1.3305  1.3212  1.3326  1.3168  1.3542  1.3310  0.0130
Fit time          6.34    5.65    5.73    6.62    5.33    5.94    0.47
Test time         0.25    0.19    0.15    0.24    0.25    0.21    0.04
Total Elapsed time with the SVD is: 31.930691242218018
```

In the second part, the vote average is predicted from the vote count and the budget:

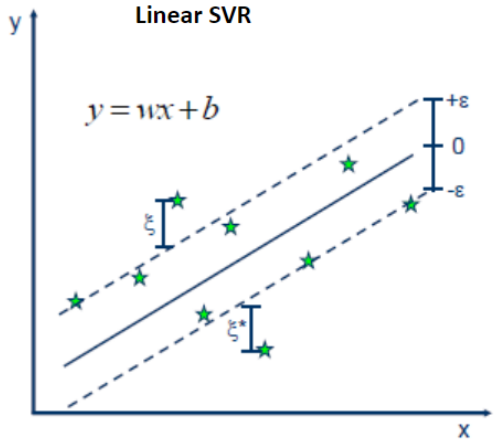
```
Evaluating RMSE, MAE of algorithm SVD on 5 split(s).

                  Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)    1.5954  1.5934  1.6020  1.6071  1.5966  1.5989  0.0050
MAE (testset)     1.3248  1.3302  1.3278  1.3367  1.3285  1.3296  0.0040
Fit time          6.10    9.06    3.51    4.85    3.52    5.41    2.07
Test time         0.31    0.12    0.08    0.12    0.09    0.15    0.08
Total Elapsed time with SVD is: 28.691766500473022
```

# Random Forest

Random forest is a type of supervised machine learning algorithm based on ensemble learning. The random forest algorithm combines multiple algorithm of the same type i.e. multiple decision trees, resulting in a forest of trees, hence the name "Random Forest".

```
Total Elapsed time with model is: 36.87901520729065
RMSE value for n_estimator=  140 is: 1.228152851883824

Total Elapsed time with model is: 35.937907457351685
RMSE value for n_estimator=  150 is: 1.2278656497690312

Total Elapsed time with model is: 40.82821488380432
RMSE value for n_estimator=  160 is: 1.2273117394335689

Total Elapsed time with model is: 42.62268590927124
RMSE value for n_estimator=  170 is: 1.2268291900261832

Total Elapsed time with model is: 44.19658589363098
RMSE value for n_estimator=  180 is: 1.2270316755432094

Total Elapsed time with model is: 47.269429206848145
RMSE value for n_estimator=  190 is: 1.22688515828762
```

# Support Vector Regression




- Minimize:

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*)$$

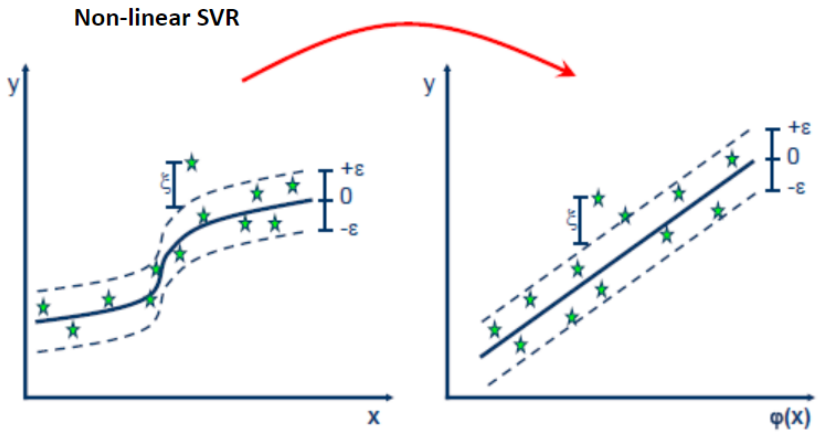
- Constraints:

$$y_i - wx_i - b \le \varepsilon + \xi_i$$

$$wx_i + b - y_i \le \varepsilon + \xi_i^*$$

$$\xi_i, \xi_i^* \ge 0$$

$$y = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)\cdot\langle\varphi(x_i), \varphi(x)\rangle + b$$




$$y = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)\cdot K(x_i, x) + b$$

Polynomial

$$k(\mathrm{x}_i, \mathrm{x}_j) = (\mathrm{x}_i . \mathrm{x}_j)^d$$

Gaussian Radial Basis function

$$k(\mathrm{x}_i, \mathrm{x}_j) = \exp\left(-\frac{\|\mathrm{x}_i - \mathrm{x}_j\|^2}{2\sigma^2}\right)$$

# Support Vector Regression

As it can see from the results, although SVR is a strong technique for prediction but is slow and it is essential to tune the parameters and the RMSE is not as small as random forest.

```
Total Elapsed time with model is: 77.32080340385437
SVR RMSE (gamma=.01, C=100) is: 1.8604387694736897

Total Elapsed time with model is: 66.25891923904419
SVR RMSE(gamma=0.001, C=100) is: 1.8819087974081985

Total Elapsed time with model is: 64.6372721195221
SVR RMSE(gamma=0.001, C=10) is: 1.888942925254787
```

# K-Nearest Neighbors (Regression)

KNN is fast but the RMSE is not as small as random forest.

```
RMSE value for k= 11 is: 1.8038207834268334
RMSE value for k= 13 is: 1.7934131089891299
RMSE value for k= 15 is: 1.78986557123273
RMSE value for k= 17 is: 1.786579572380617
RMSE value for k= 19 is: 1.783226039661783
```

# Conclusion

- This work highlighted the process of data wrangling, storytelling, EDA, data visualization, inferential statistics, recommender systems and machine learning techniques to perform on two movie datasets.
- In this work, several recommended systems are implemented as: **Simple Recommender IMDB Weighted Rating system Content Based Recommender Machine learning**, Random Forest, SVR, KNN
  **Which technique is the best to utilize for the stakeholder?**
- Picking the best technique for utilizing is not very easy.
- There is not one techniques suitable for all data sets. But As it is shown in the previous sections, When we don't have many features (Movielens dataset), SVD could be the easiest and fastest techniques to recommend.

- For datasets with several features (MetaData dataset) we can predict the rating or popularity of movies based with different machine learning techniques. Random Forest worked fine and fast for predicting the target for this dataset.
- **Future Work**
  For future, other recommender techniques like Hybrid techniques or Deep learning methods could be implemented. Also A/B testing technique could be combined with previous techniques to improve the result.

# Thanks for Your Attention
# Any Question?