IMDb Movie Analysis - 10,000 Movies

Shoshana Farber

DATA 602 - Spring 2023

Abstract

The world of movies is ever-evolving, with fresh releases hitting the scene week after week. Yet, not every film finds success, and some are deemed "flops" based on viewer reviews. This analysis aims to discover popular movies and to uncover patterns and insights related to movie popularity and ratings. The data, comprising 10,000 movies, includes features such as release year, rating, genre, certification, and critical production details.

Exploratory data analysis (EDA) is employed to understand the distributions of variables such as rating, metascore, and gross revenue for movies, and to determine how to best “wrangle” the data. Further analysis is done to determine factors relating to high movie ratings. Some notable findings include romance movies as the top rated genre, followed closely by mystery, drama, and crime. A model was created using release year, movie length, votes, genre, and certification to predict movie ratings. The model accounted for about 47% of the variability in movie ratings and could be further improved with additional movie information.

Additionally, the project introduces a basic recommender system based on IMDb's weighted rating formula and a content based recommender system using the movie description, genres, and other production information. Using these recommender systems, one could filter the data to come up with the top rated movies for their specifications or they can find similar movies to those they’ve enjoyed.

Introduction

The dataset selected for this project encompasses IMDb movies and their associated attributes, including release year, rating, genre, certification, and key production details such as director, stars, runtime, and more. IMDb is an internet movie database featuring information on thousands of movies, TV shows, video games, and more worldwide. From timeless classics to the latest releases, IMDb spans the entire spectrum of entertainment. Registered users can rate movies, helping others discover popular and noteworthy films.

This project aims to see which movies/types of movies are most popular on IMDb and to identify features which may influence film popularity. I will also be attempting to come up with a few basic recommender systems based on the rating and content of each movie.

Required Libraries

General:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import re
```

Modeling:

```
import sklearn.linear_model as lm
from sklearn.model_selection import train_test_split as tts
from sklearn.metrics import mean_squared_error
```

Recommender:

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import linear_kernel
```

Dataset and Variables

Movie Name - name of the movie

Year of Release - year movie was released

Run Time in minutes - length of the movie

Movie Rating - rating of movie

Votes - number of users who rated the movie

MetaScore - metascore of the movie (metric of rating based on movie critics)

Gross - how much the movie made

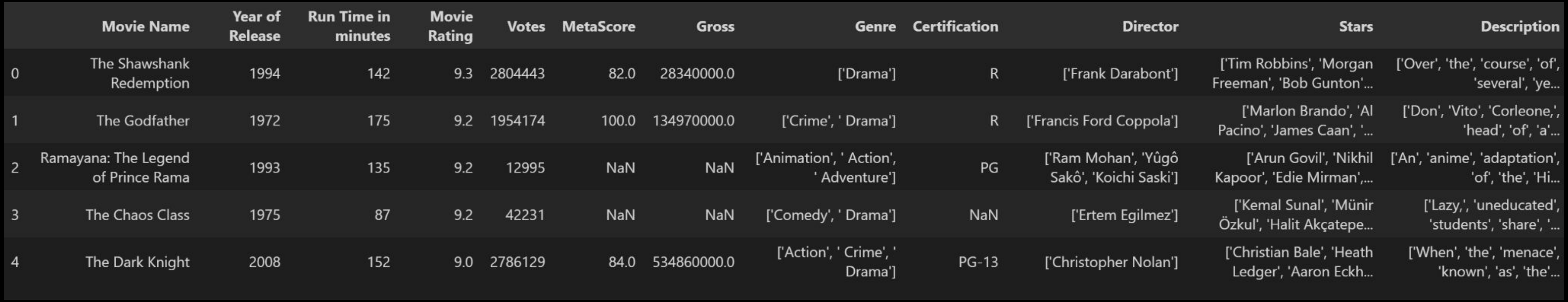
Genre - genre(s) of movie

Certification - content rating of movie

Director - director(s) of movie

Stars - main actor(s) in movie

Description - movie bio



	Movie Name	Year of Release	Run Time in minutes	Movie Rating	Votes	MetaScore	Gross	Genre	Certification	Director	Stars	Description
0	The Shawshank Redemption	1994	142	9.3	2804443	82.0	28340000.0	['Drama']	R	['Frank Darabont']	['Tim Robbins', 'Morgan Freeman', 'Bob Gunton'...	['Over', 'the', 'course', 'of', 'several', 'ye...
1	The Godfather	1972	175	9.2	1954174	100.0	134970000.0	['Crime', ' Drama']	R	['Francis Ford Coppola']	['Marlon Brando', 'Al Pacino', 'James Caan', '...	['Don', 'Vito', 'Corleone,', 'head', 'of', 'a'...
2	Ramayana: The Legend of Prince Rama	1993	135	9.2	12995	NaN	NaN	['Animation', ' Action', ' Adventure']	PG	['Ram Mohan', 'Yûgô Sakô', 'Koichi Saski']	['Arun Govil', 'Nikhil Kapoor', 'Edie Mirman',...	['An', 'anime', 'adaptation', 'of', 'the', 'Hi...
3	The Chaos Class	1975	87	9.2	42231	NaN	NaN	['Comedy', ' Drama']	NaN	['Ertem Egilmez']	['Kemal Sunal', 'Münir Özkul', 'Halit Akçatepe...	['Lazy,', 'uneducated', 'students', 'share', '...
4	The Dark Knight	2008	152	9.0	2786129	84.0	534860000.0	['Action', ' Crime', ' Drama']	PG-13	['Christopher Nolan']	['Christian Bale', 'Heath Ledger', 'Aaron Eckh...	['When', 'the', 'menace', 'known', 'as', 'the'...

DATA WRANGLING

Column Manipulation

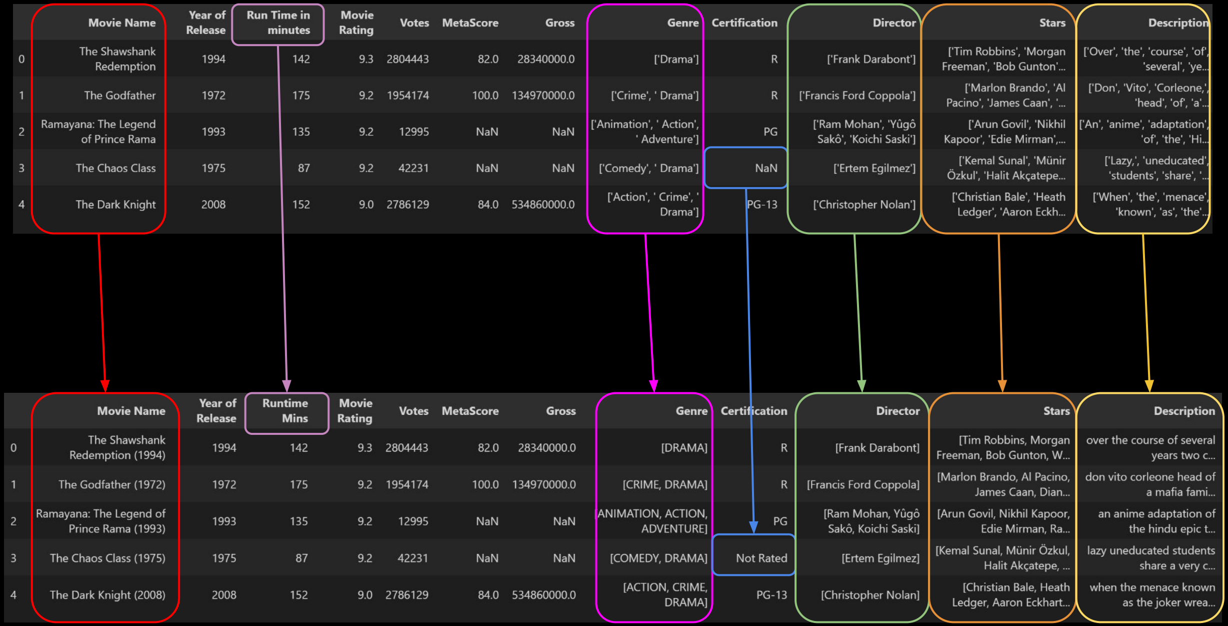
- Change **Run Time in minutes** to **Runtime Mins**

- Clean up and turn **Genres, Directors, Stars, Description** columns into lists

- Join words in **Description** columns list

- Add **Year of Release** into **Movie Name**

- Fill any NaN **Certification** with “Not Rated”



	Movie Name	Year of Release	Run Time in minutes	Movie Rating	Votes	MetaScore	Gross	Genre	Certification	Director	Stars	Description
0	The Shawshank Redemption	1994	142	9.3	2804443	82.0	28340000.0	['Drama']	R	['Frank Darabont']	['Tim Robbins', 'Morgan Freeman', 'Bob Gunton'...	['Over', 'the', 'course', 'of', 'several', 'ye...
1	The Godfather	1972	175	9.2	1954174	100.0	134970000.0	['Crime', ' Drama']	R	['Francis Ford Coppola']	['Marlon Brando', 'Al Pacino', 'James Caan', '...	['Don', 'Vito', 'Corleone,', 'head', 'of', 'a'...
2	Ramayana: The Legend of Prince Rama	1993	135	9.2	12995	NaN	NaN	['Animation', ' Action', ' Adventure']	PG	['Ram Mohan', 'Yûgô Sakô', 'Koichi Saski']	['Arun Govil', 'Nikhil Kapoor', 'Edie Mirman',...	['An', 'anime', 'adaptation', 'of', 'the', 'Hi...
3	The Chaos Class	1975	87	9.2	42231	NaN	NaN	['Comedy', ' Drama']	NaN	['Ertem Egilmez']	['Kemal Sunal', 'Münir Özkul', 'Halit Akçatepe...	['Lazy,', 'uneducated', 'students', 'share', '...
4	The Dark Knight	2008	152	9.0	2786129	84.0	534860000.0	['Action', ' Crime', ' Drama']	PG-13	['Christopher Nolan']	['Christian Bale', 'Heath Ledger', 'Aaron Eckh...	['When', 'the', 'menace', 'known', 'as', 'the'...

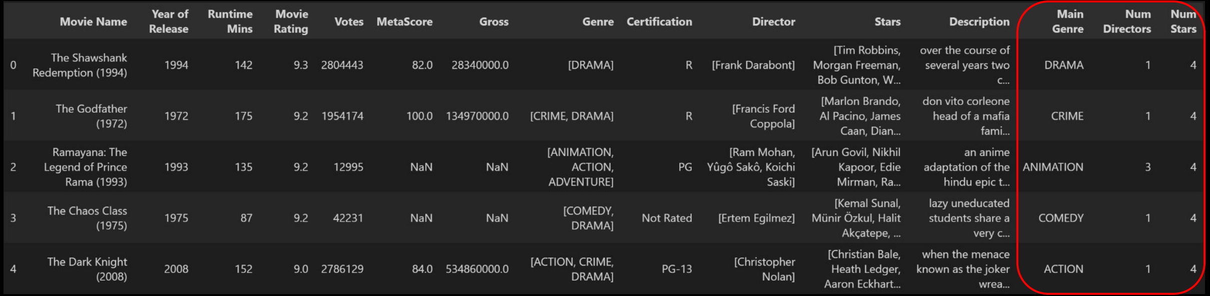
	Movie Name	Year of Release	Runtime Mins	Movie Rating	Votes	MetaScore	Gross	Genre	Certification	Director	Stars	Description
0	The Shawshank Redemption (1994)	1994	142	9.3	2804443	82.0	28340000.0	[DRAMA]	R	[Frank Darabont]	[Tim Robbins, Morgan Freeman, Bob Gunton, W...	over the course of several years two c...
1	The Godfather (1972)	1972	175	9.2	1954174	100.0	134970000.0	[CRIME, DRAMA]	R	[Francis Ford Coppola]	[Marlon Brando, Al Pacino, James Caan, Dian...	don vito corleone head of a mafia fami...
2	Ramayana: The Legend of Prince Rama (1993)	1993	135	9.2	12995	NaN	NaN	[ANIMATION, ACTION, ADVENTURE]	PG	[Ram Mohan, Yûgô Sakô, Koichi Saski]	[Arun Govil, Nikhil Kapoor, Edie Mirman, Ra...	an anime adaptation of the hindu epic t...
3	The Chaos Class (1975)	1975	87	9.2	42231	NaN	NaN	[COMEDY, DRAMA]	Not Rated	[Ertem Egilmez]	[Kemal Sunal, Münir Özkul, Halit Akçatepe, ...	lazy uneducated students share a very c...
4	The Dark Knight (2008)	2008	152	9.0	2786129	84.0	534860000.0	[ACTION, CRIME, DRAMA]	PG-13	[Christopher Nolan]	[Christian Bale, Heath Ledger, Aaron Eckhart...	when the menace known as the joker wrea...

Column Creation

- **Main Genre** - first genre in list

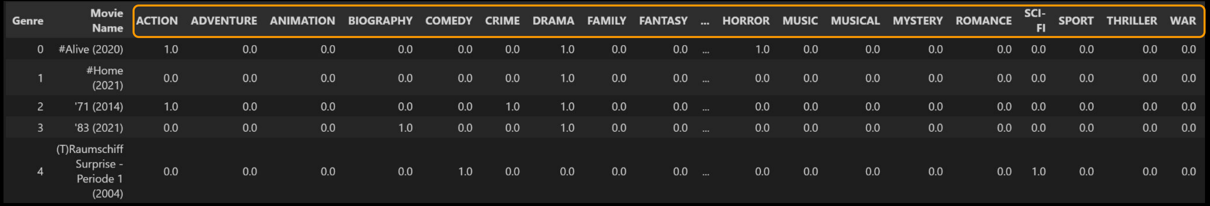
- **Num Directors** - number of directors

- **Num Stars** - number of main stars



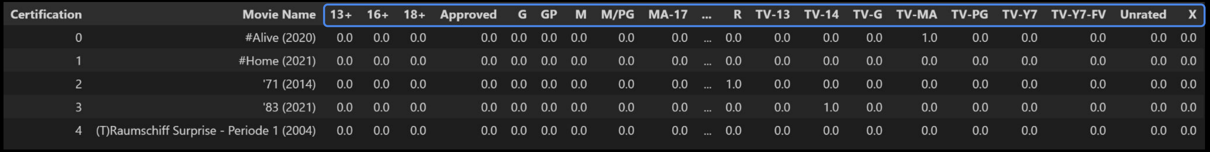
	Movie Name	Year of Release	Runtime Mins	Movie Rating	Votes	MetaScore	Gross	Genre	Certification	Director	Stars	Description	Main Genre	Num Directors	Num Stars
0	The Shawshank Redemption (1994)	1994	142	9.3	2804443	82.0	28340000.0	[DRAMA]	R	[Frank Darabont]	[Tim Robbins, Morgan Freeman, Bob Gunton, W...	over the course of several years two c...	DRAMA	1	4
1	The Godfather (1972)	1972	175	9.2	1954174	100.0	134970000.0	[CRIME, DRAMA]	R	[Francis Ford Coppola]	[Marlon Brando, Al Pacino, James Caan, Dian...	don vito corleone head of a mafia fami...	CRIME	1	4
2	Ramayana: The Legend of Prince Rama (1993)	1993	135	9.2	12995	NaN	NaN	[ANIMATION, ACTION, ADVENTURE]	PG	[Ram Mohan, Yûgô Sakô, Koichi Saski]	[Arun Govil, Nikhil Kapoor, Edie Mirman, Ra...	an anime adaptation of the hindu epic t...	ANIMATION	3	4
3	The Chaos Class (1975)	1975	87	9.2	42231	NaN	NaN	[COMEDY, DRAMA]	Not Rated	[Ertem Egilmez]	[Kemal Sunal, Münir Özkul, Halit Akçatepe, ...	lazy uneducated students share a very c...	COMEDY	1	4
4	The Dark Knight (2008)	2008	152	9.0	2786129	84.0	534860000.0	[ACTION, CRIME, DRAMA]	PG-13	[Christopher Nolan]	[Christian Bale, Heath Ledger, Aaron Eckhart...	when the menace known as the joker wrea...	ACTION	1	4

- Dummy variables for **Genre**



Genre	Movie Name	ACTION	ADVENTURE	ANIMATION	BIOGRAPHY	COMEDY	CRIME	DRAMA	FAMILY	FANTASY	...	HORROR	MUSIC	MUSICAL	MYSTERY	ROMANCE	SCI-FI	SPORT	THRILLER	WAR
0	#Alive (2020)	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	...	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	#Home (2021)	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	'71 (2014)	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	'83 (2021)	0.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	(T)Raumschiff Surprise - Periode 1 (2004)	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0

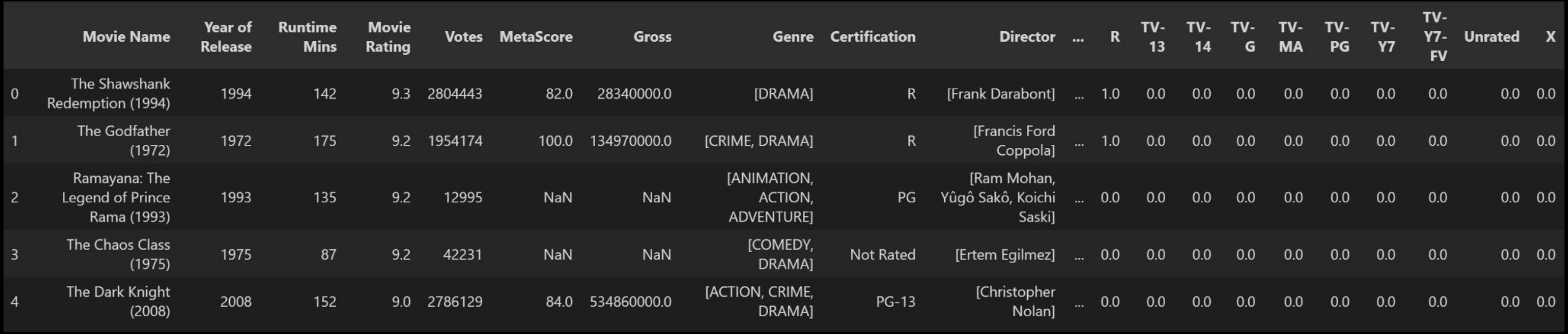
- Dummy variables for **Certification**



Certification	Movie Name	13+	16+	18+	Approved	G	GP	M	M/PG	MA-17	...	R	TV-13	TV-14	TV-G	TV-MA	TV-PG	TV-Y7	TV-Y7-FV	Unrated	X
0	#Alive (2020)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
1	#Home (2021)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	'71 (2014)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	'83 (2021)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	(T)Raumschiff Surprise - Periode 1 (2004)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Final Data Frame

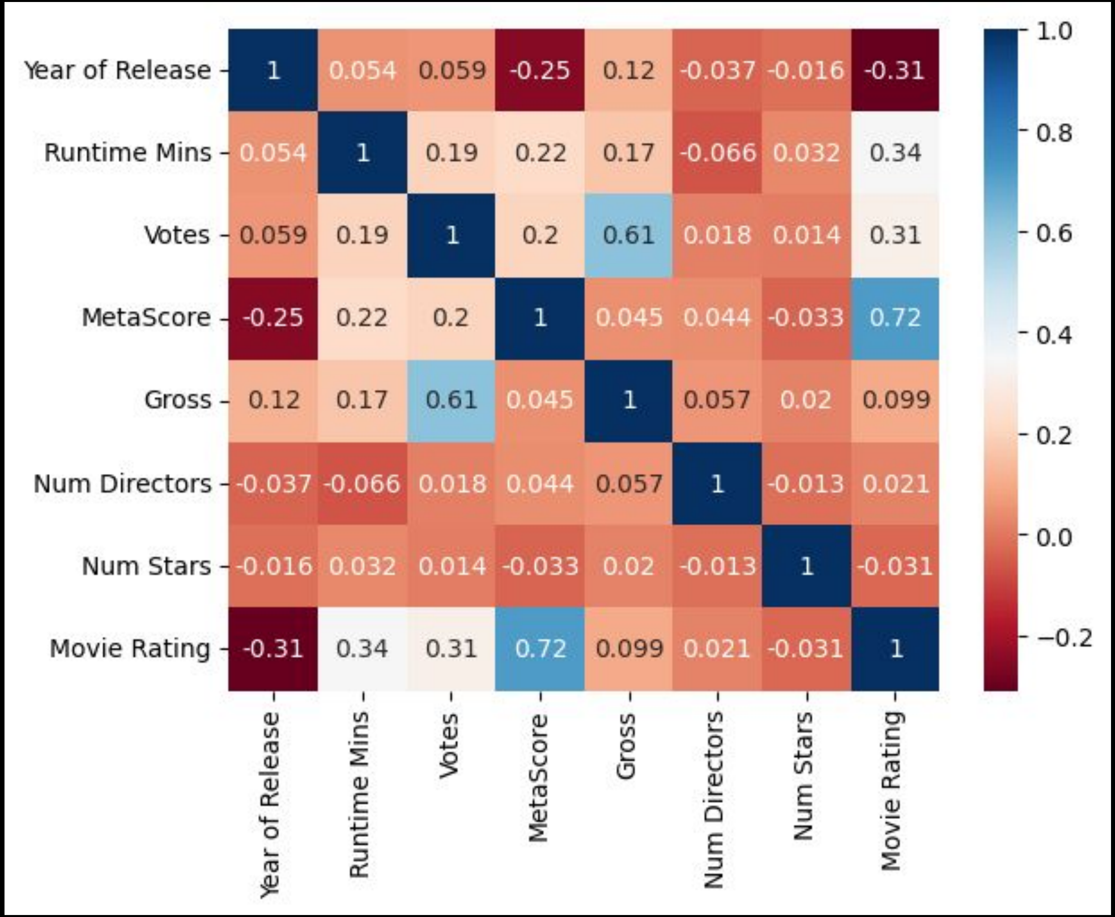
- Join main data frame with dataframes of dummy variables



	Movie Name	Year of Release	Runtime Mins	Movie Rating	Votes	MetaScore	Gross	Genre	Certification	Director	...	R	TV-13	TV-14	TV-G	TV-MA	TV-PG	TV-Y7	TV-Y7-FV	Unrated	X
0	The Shawshank Redemption (1994)	1994	142	9.3	2804443	82.0	28340000.0	[DRAMA]	R	[Frank Darabont]	...	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	The Godfather (1972)	1972	175	9.2	1954174	100.0	134970000.0	[CRIME, DRAMA]	R	[Francis Ford Coppola]	...	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	Ramayana: The Legend of Prince Rama (1993)	1993	135	9.2	12995	NaN	NaN	[ANIMATION, ACTION, ADVENTURE]	PG	[Ram Mohan, Yûgô Sakô, Koichi Saski]	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	The Chaos Class (1975)	1975	87	9.2	42231	NaN	NaN	[COMEDY, DRAMA]	Not Rated	[Ertem Egilmez]	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	The Dark Knight (2008)	2008	152	9.0	2786129	84.0	534860000.0	[ACTION, CRIME, DRAMA]	PG-13	[Christopher Nolan]	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

The Godfather
Pulp Fiction
Schindler's List
12 Angry Men
Inception
Jack Nicholson
One Flew Over the Cuckoo's Nest
Lord of the Rings
Seven Samurai
Fight Club
Star Wars
GoodFellas
Casablanca
Cidade de Deus
Once Upon a Time in the West
James Stewart
Rear Window
Indiana Jones
Matrix
Psycho
The Silence of the Lambs
Toy Story 3
It's a Wonderful Life
Memento
Sunset Boulevard
Tom Hanks
Forrest Gump
Leon
Dr. Strangelove
Apocalypse Now
Orson Welles
Citizen Kane
North by Northwest
American Beauty
American History X
Terminator 2
Saving Private Ryan
James Stewart
Kim Novak
Vertigo
Amélie
Alien
Wall-E
Spirited Away
The Shining
Paths of Glory
Lawrence of Arabia
Black Swan
Double Indemnity
The Pianist
A Clockwork Orange
Charlie Chaplin
City Lights
The Lives of Others
M
The Departed
Requiem for a Dream
Das Boot
The Third Man
Reservoir Dogs
Chinatown
Charlie Chaplin
Modern Times
The Treasure of the Sierra Madre
Back to the Future
The Prestige
Raging Bull
Cinema Paradiso
Singin' in the Rain
Gene Kelly
Donald O'Connor
Debbie Reynolds
Some Like It Hot
The Bridge on the River Kwai
Rashomon
Amadeus
All About Eve
The Green Mile
Full Metal Jacket
2001: a space odyssey
Charlie Chaplin
The Great Dictator
Braveheart
Downfall
Up
Gran Torino
Metropolis
Gladiator
Bogart
Astor
The Maltese Falcon
Elephant Man
올드보이
Mr. Smith Goes to Washington
Sin City
On the Waterfront
Rebecca
Return of the Jedi
Indiana Jones
Die Hard
The Great Escape
Princess Mononoke
The King's Speech
Det sjunde inseglet
Jaws
Hotel Rwanda
Fargo
The General
Slumdog Millionaire
No Country for Old Men
Heat
The Wizard of Oz
For a Few Dollars More
Touch of Evil
Yojimbo
Ran
Witness for the Prosecution
Grave of the Fireflies
Snatch
The Sixth Sense
Annie Hall
The Deer Hunter
Donnie Darko
District 9
Paul Newman
Cool Hand Luke
Strangers on a Train
It Happened One Night
High Noon
Kill Bill
The Lion King
Platoon
Into the Wild
There Will Be Blood
Avatar
Toy Story
Notorious

ANALYSIS



	Year of Release	Runtime Mins	Votes	MetaScore	Gross	Num Directors	Num Stars	Movie Rating
Year of Release	1	0.054	0.059	-0.25	0.12	-0.037	-0.016	-0.31
Runtime Mins	0.054	1	0.19	0.22	0.17	-0.066	0.032	0.34
Votes	0.059	0.19	1	0.2	0.61	0.018	0.014	0.31
MetaScore	-0.25	0.22	0.2	1	0.045	0.044	-0.033	0.72
Gross	0.12	0.17	0.61	0.045	1	0.057	0.02	0.099
Num Directors	-0.037	-0.066	0.018	0.044	0.057	1	-0.013	0.021
Num Stars	-0.016	0.032	0.014	-0.033	0.02	-0.013	1	-0.031
Movie Rating	-0.31	0.34	0.31	0.72	0.099	0.021	-0.031	1

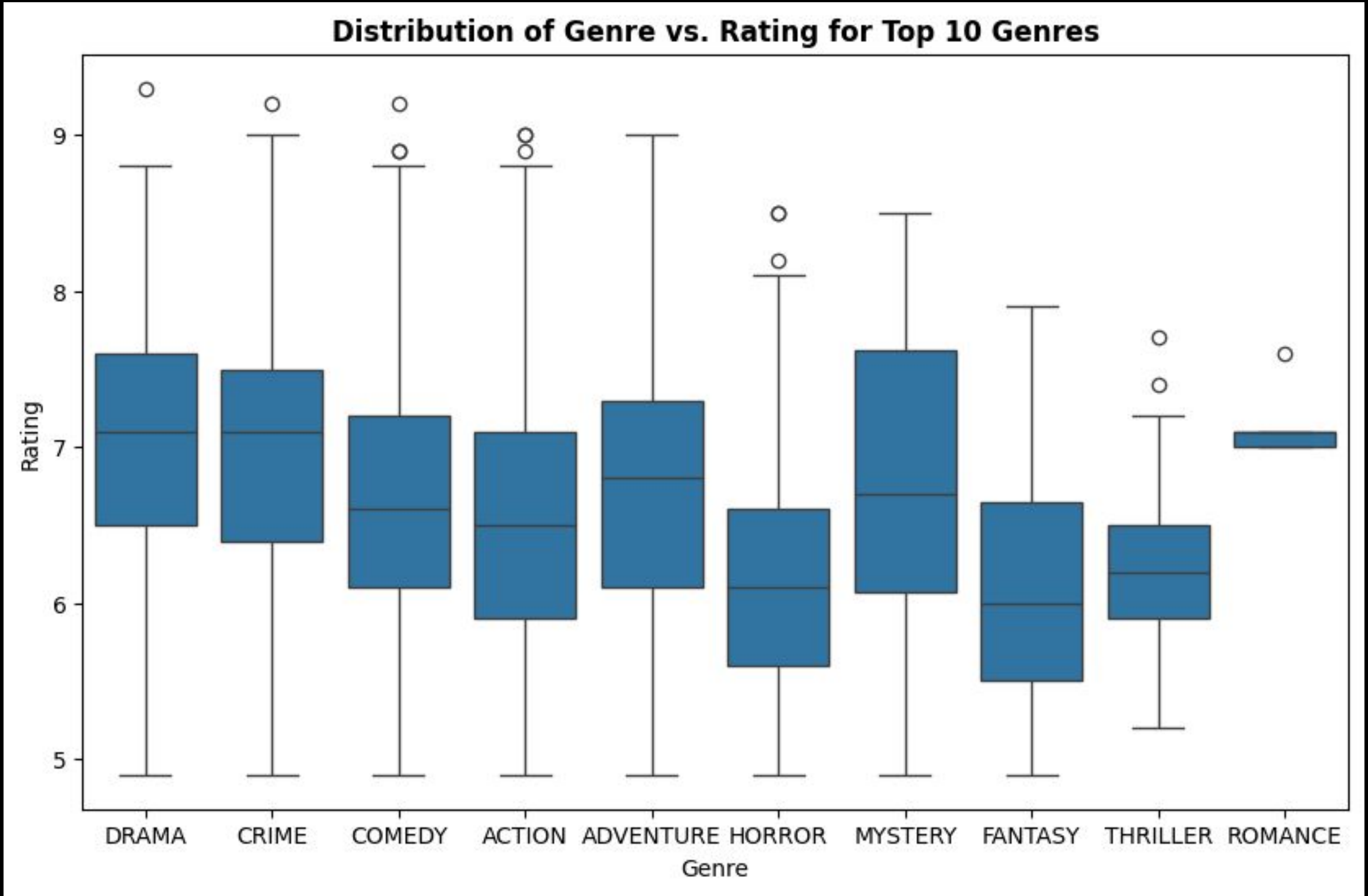
Numerical Correlations

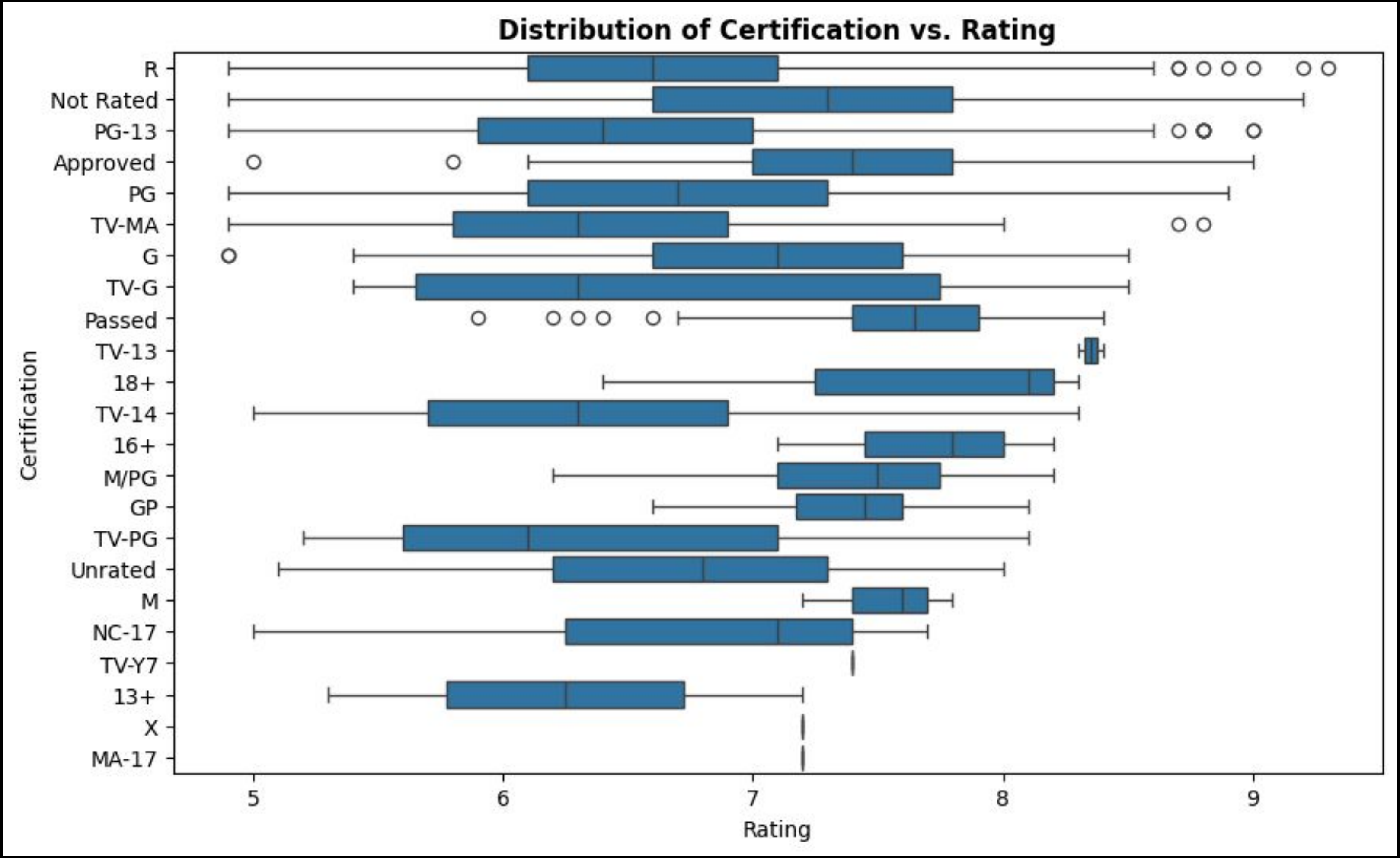
- Large positive correlation between **MetaScore** and **Movie Rating**

- Slight positive correlation between **Runtime Mins** and **Movie Rating**, **Votes** and **Movie Rating**

- Very small positive correlation between **Gross** and **Movie Rating**

- There is a slight negative correlation between **Year of Release** and **Movie Rating**


Distribution of Genre vs. Rating for Top 10 Genres
Rating
9
8
7
6
5
DRAMA
CRIME
COMEDY
ACTION
ADVENTURE
HORROR
MYSTERY
FANTASY
THRILLER
ROMANCE
Genre


Distribution of Certification vs. Rating
Certification
R
Not Rated
PG-13
Approved
PG
TV-MA
G
TV-G
Passed
TV-13
18+
TV-14
16+
M/PG
GP
TV-PG
Unrated
M
NC-17
TV-Y7
13+
X
MA-17
Rating
5
6
7
8
9

Modeling

```
# variables of interest
variables = full[['Movie Rating', 'Year of Release', 'Runtime Mins', 'Votes', 'ACTION', 'ADVENTURE',
'ANIMATION', 'BIOGRAPHY', 'COMEDY', 'CRIME', 'DRAMA', 'FAMILY', 'FANTASY', 'FILM-NOIR', 'HISTORY',
'HORROR', 'MUSIC', 'MUSICAL', 'MYSTERY', 'ROMANCE', 'SCI-FI', 'SPORT', 'THRILLER', 'WAR', 'WESTERN',
'13+', '16+', '18+', 'Approved', 'G', 'GP', 'M', 'M/PG', 'MA-17', 'NC-17', 'Not Rated', 'PG', 'PG-13',
'Passed', 'R', 'TV-13', 'TV-14', 'TV-G', 'TV-MA', 'TV-PG', 'TV-Y7', 'TV-Y7-FV', 'Unrated', 'X']]

# define X, y and split into train:test data
X = variables.drop(columns=['Movie Rating']).values
y = variables['Movie Rating'].values
X_train, X_test, y_train, y_test = tts(X, y, test_size=0.3, random_state=23)

# initiate model and fit data
model = lm.LinearRegression()
model.fit(X_train, y_train)
```

Modeling

```
# predict
y_pred = model.predict(X_test)

# check accuracy
r2 = model.score(X_test, y_test)
mse = mean_squared_error(y_test, y_pred)

print(f'R-squared: {r2}')
print(f'Mean Squared Error: {mse}')

>>>
R-squared: 0.4671394895179628
Mean Square Error: 0.3639993450869121
>>>
```

RECOMMENDER SYSTEM

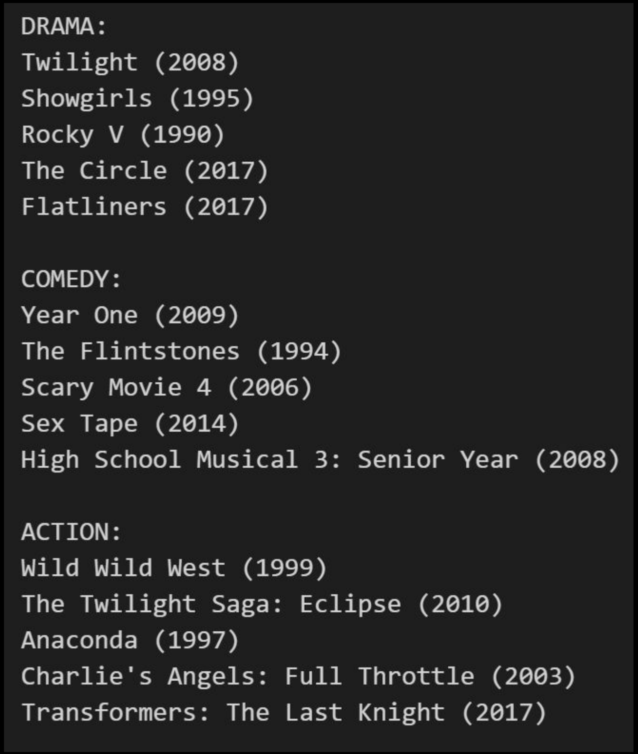
Basic Recommender - Weighted Rating

$$WR = \frac{v}{v+m} * R + \frac{m}{v+m} * C$$

- R = average rating for movie
- v = number of votes for movie
- m = minimum votes to be listed in Top 250 (25,000 votes)
- C = mean vote across dataset

```
# function for weighted rating
def weighted_rating(df, m=25000):
    R = df['Movie Rating']
    v = df['Votes']
    C = df['Movie Rating'].mean()

    return (v/(v+m) * R) + (m/(v+m) * C)
```



```
DRAMA:
Twilight (2008)
Showgirls (1995)
Rocky V (1990)
The Circle (2017)
Flatliners (2017)

COMEDY:
Year One (2009)
The Flintstones (1994)
Scary Movie 4 (2006)
Sex Tape (2014)
High School Musical 3: Senior Year (2008)

ACTION:
Wild Wild West (1999)
The Twilight Saga: Eclipse (2010)
Anaconda (1997)
Charlie's Angels: Full Throttle (2003)
Transformers: The Last Knight (2017)
```

Content Based Recommender

Step 1: Combine information into one column (**Description, Directors, Actors**)

Step 2: Vectorize words in new column

```
tfidf = TfidfVectorizer(stop_words='english')

# make TF-IDF matrix
tfidf_matrix = tfidf.fit_transform(movies_df['Content'])
```

Step 3: Use cosine similarity scores to compare similar information

```
# cosine similarity matrix
cosine_sim = linear_kernel(tfidf_matrix, tfidf_matrix)

# reverse map of indices and movie titles
indices = pd.Series(movies_df.index, index=movies_df['Movie Name'])
```

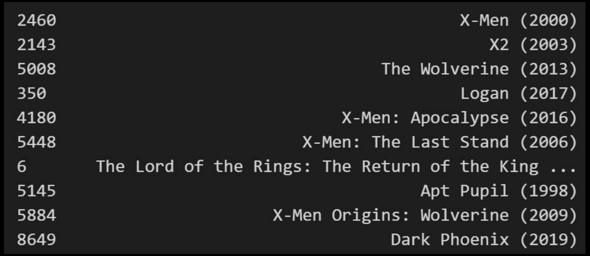
Content Based Recommender

Step 4: Function to get top 10 recommendations based on similarity scores

```
# function that takes in movie title as input and outputs most similar movies
def get_recommendations(title, cosine_sim=cosine_sim):
    index = indices[title]  # get index of movie
    sim_scores = list(enumerate(cosine_sim[index])) # get similarity scores
    sim_scores = sorted(sim_scores, key=lambda x: x[1], reverse=True)   # sort based on similarity
scores
    sim_scores = sim_scores[1:11]   # scores of 10 most similar movies
    movie_indices = [i[0] for i in sim_scores]  # movie indices for similar movies
    return movies_df['Movie Name'].iloc[movie_indices]  # return top 10 similar movies
```

Step 5: Implement function

```
get_recommendations('X-Men: Days of Future Past (2014)')
```



```
2460                                          X-Men (2000)
2143                                             X2 (2003)
5008                                  The Wolverine (2013)
350                                           Logan (2017)
4180                           X-Men: Apocalypse (2016)
5448                        X-Men: The Last Stand (2006)
6          The Lord of the Rings: The Return of the King ...
5145                                      Apt Pupil (1998)
5884                      X-Men Origins: Wolverine (2009)
8649                                   Dark Phoenix (2019)
```

Conclusions

This analysis of IMDb movie data provides insights into some of the factors determining movie popularity and ratings. Through exploratory data analysis, we identified the top genres, with romance movies emerging as the highest rated, closely followed by mystery, drama, and crime. The distribution of ratings for different genres was visualized, revealing interesting patterns and variations. The correlation analysis highlighted factors such as movie length and the number of votes, indicating their influence on higher ratings. A model was created to predict movie rating based on release year, movie length, votes, genre, and certification. The model accounted for 47% of the variation in the data and had a relatively high mean square error but could be improved using other demographic or user information which was not sourced for this project.

A basic recommender system was implemented to suggest movies based on highest weighted rating and a content based recommender system was implemented to suggest movies based on previously watched movies.

In conclusion, this project provides a comprehensive exploration of IMDb movie data, offering valuable insights for movie enthusiasts, analysts, and industry professionals. The findings contribute to a deeper understanding of the dynamics that drive movie ratings and popularity on the IMDb platform.

Sources

1. Data: https://www.kaggle.com/datasets/dk123891/10000-movies-data?select=data.csv

2. IMDb Weighted Ratings: https://www.quora.com/How-does-IMDbs-rating-system-work

3. Recommender Systems Tutorial: https://www.datacamp.com/tutorial/recommender-systems-python