# Project Report

## On

# Movie Recommendation and Analytics

Submitted in partial fulfillment for the award of**Post Graduate Diploma in Big Data Analytics** from **C-DAC Kharghar (Mumbai)**


**Guided by**

**Mr. Parikshit Chaudhari**

Presented by:

| | |
|---|---|
| **Varsha Nanaware (PL)** | **PRN: 220940325086** |
| **Anjana Anilkumar** | **PRN: 220940325011** |
| **Mrinal Gawade** | **PRN: 220940325043** |
| **Samyak Tantarpale** | **PRN: 220940325060** |
| **Richa Pathak** | **PRN: 220940325055** |

**Centre of Development of Advanced Computing (CDAC), Mumbai**

# CERTIFICATE

**This is to certify that,**

**Varsha Nanaware (PL)**

**Anjana Anilkumar**

**Mrinal Gawade**

**Samyak Tantarpale**

**Richa Pathak**

**Have successfully completed their project on**

## Movie Recommendation and Analytics

**Under the guidance of Mr. Parikshit Chaudhari**

**Project Guide** **Project Supervisor**

**HOD CDAC KHARGHAR**

**Dr. CP Johnson.**

# ACKNOWLEDGEMENT

We had a great learning experience working on the project "Movie Recommendation and Analytics" and are submitting our work to the Advanced Computing Training School (CDAC) in Kharghar, Mumbai.

We are delighted to acknowledge Mr. Parikshit Chaudhary's valuable guidance in helping us overcome various obstacles and intricacies throughout the project work.

Our sincere gratitude goes to Dr. CP Johnson, Senior Director of C-DAC Mumbai, Kharghar, for his guidance and support throughout our Post Graduate Diploma in Big Data Analytics (PG-DBDA) course.

We would also like to extend our heartfelt thanks to Mrs. Vineeta Singh, the Course Coordinator for PG-DBDA, who provided all the necessary support and coordination, including the required hardware, internet facility, and extra lab hours, to help us complete the project and the course up to the last day here at C-DAC Kharghar.

**From:**
Varsha Nanaware (220940325086)
Anjana Anilkumar(220940325011)
Mrinal Gawade (220940325043)
Samyak Tantarpale (220940325060)
Richa Pathak (220940325055)

## TABLE OF CONTENTS

# 1. Abstract

The proliferation of streaming services has made it challenging for users to navigate vast movie catalogs and find films they enjoy. A movie recommendation system can be a viable solution that offers users personalized movie suggestions based on their viewing history and preferences. In this project, we propose a Movie Recommendation System using PySpark, Pandas, SparkSQL, Matplotlib, PowerBI.

Our recommendation system leverages the IMDB dataset, which contains movie ratings. We preprocess the data and build four different models, namely, Popularity-based Filtering Algorithm, Content-based Filtering Algorithm, Actor-based Filtering Algorithm and Actress-based Filtering Algorithm, for the recommendation system.

We utilize the Scikit-learn library to implement Algorithms, which enables us to scale our system for large datasets. Our system provides users with top-rated movies and similar movies.

# 2. Introduction and Overview of Project

The basic concept behind a Movie Recommendation System is quite simple. There are two main elements in every recommender system: users and items. The system generates movie predictions for its users, while items refer to the movies themselves.

The primary goal of movie recommendation systems is to filter and predict only those movies that a corresponding user is most likely to want to watch. For this recommendation, four different algorithms, namely, Popularity-based Filtering Algorithm, Content-based Filtering Algorithm, Actor-based Filtering Algorithm and Actress-based Filtering Algorithm have been built.

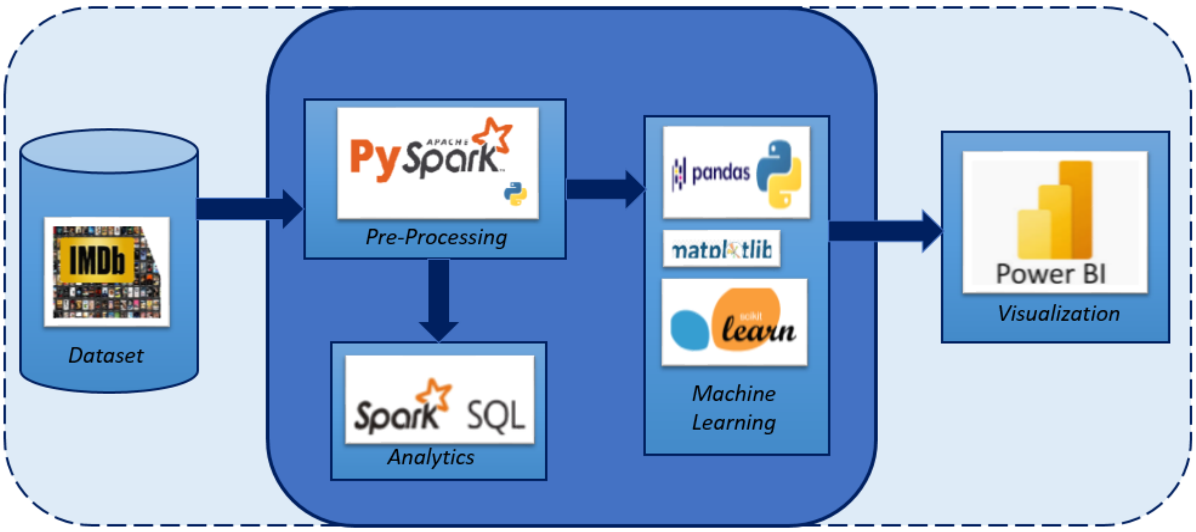
The recommendation system analyzes the past preferences of the concerned user, and then uses this information to find similar movies. This information is available in the database, such as actors, genres, etc. After that, the system provides movie recommendations for the user.

In this project, our aim is to build a movie recommendation system that can provide personalized and relevant movie recommendations to users using PySpark, Pandas, SparkSQL, Scikit-learn, and Power BI.

# 3. Problem Statement

Nowadays, there is a vast collection of movies available on various platforms, making it difficult for users to find movies that match their preferences. To personalize the movie experience and increase user engagement with the platform, a recommendation system is preferred. This system can also help the platform improve its revenue generation in the present competitive environment. The goal of our recommendation system is to suggest movies to users based on their preferences using a dataset of movies with their features.

# Architecture of Movie Recommender System



# 4. Dataset Description

The IMDb (Internet Movie Database) dataset comprises data about movies and TV movies. It encompasses details about movies and TV movies, including title, release date, genre, cast and crew and ratings. The dataset is presented in tabular format, where each row represents a unique movie or TV movie, and each column contains a different attribute of the movie or TV movie.

This dataset is often used for data analysis, data visualization, and machine learning projects related to movies and TV movies. It can be employed to answer several questions, such as which actors or directors are most popular, which genres are most in demand, and which movies or TV movies have the highest ratings.

All in all, the IMDb dataset is a comprehensive and substantial source of information on movies and TV shows, and it provides an invaluable resource for data-driven analysis and machine learning projects.

**<u>Raw Dataset:</u>** There are 4 tables of Raw data that we used. They are as follows:

1. **df_principals:** Contains the principal cast/crew for titles

| **Name** | **Description** |
|---|---|
| tconst (string) | alphanumeric unique identifier of the title |
| ordering (integer) | a number to uniquely identify rows for a given titleId |
| nconst (string) | alphanumeric unique identifier of the name/person |
| category (string) | the category of job that person was in |
| job (string) | the specific job title if applicable, else '\N' |
| characters (string) | the name of the character played if applicable, else '\N' |

| | tconst | ordering | nconst | category | job | characters |
|---|---|---|---|---|---|---|
| **0** | tt0000001 | 1 | nm1588970 | self | \N | ["Self"] |
| **1** | tt0000001 | 2 | nm0005690 | director | \N | \N |
| **2** | tt0000001 | 3 | nm0374658 | cinematographer | director of photography | \N |
| **3** | tt0000002 | 1 | nm0721526 | director | \N | \N |
| **4** | tt0000002 | 2 | nm1335271 | composer | \N | \N |
| **...** | ... | ... | ... | ... | ... | ... |
| **54852106** | tt9916880 | 4 | nm10535738 | actress | \N | ["Horrid Henry"] |
| **54852107** | tt9916880 | 5 | nm0996406 | director | principal director | \N |
| **54852108** | tt9916880 | 6 | nm1482639 | writer | \N | \N |
| **54852109** | tt9916880 | 7 | nm2586970 | writer | books | \N |
| **54852110** | tt9916880 | 8 | nm1594058 | producer | producer | \N |

**Fig. Dataset**

## 2. **df_titleBasics:** Contains the following information for titles

| **Name** | **Description** |
|---|---|
| tconst (string) | alphanumeric unique identifier of the title |
| titleType (string) | the type/format of the title (e.g. movie, tvseries etc) |
| primaryTitle (string) | the more popular title / the title used by the filmmakers on promotional materials at the point of release |
| originalTitle (string) | original title, in the original language |
| isAdult (boolean) | 0: non-adult title; 1: adult title |
| startYear (YYYY) | represents the release year of a title. In the case of TV Series, it is the series start year |
| endYear (YYYY) | TV Series end year. '\N' for all other title types |
| runtimeMinutes | primary runtime of the title, in minutes |
| genres (string array) | includes up to three genres associated with the title |

| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | tconst | titleType | primaryTitle | originalTitle | isAdult | startYear | endYear | runtimeM | genres | | |
| 2 | tt0000001 | short | Carmencita | Carmencita | 0 | 1894 | \N | 1 | Documentary,Short | | |
| 3 | tt0000002 | short | Le clown et ses chiens | Le clown et ses chiens | 0 | 1892 | \N | 5 | Animation,Short | | |
| 4 | tt0000003 | short | Pauvre Pierrot | Pauvre Pierrot | 0 | 1892 | \N | 4 | Animation,Comedy,Romance | | |
| 5 | tt0000004 | short | Un bon bock | Un bon bock | 0 | 1892 | \N | 12 | Animation,Short | | |
| 6 | tt0000005 | short | Blacksmith Scene | Blacksmith Scene | 0 | 1893 | \N | 1 | Comedy,Short | | |
| 7 | tt0000006 | short | Chinese Opium Den | Chinese Opium Den | 0 | 1894 | \N | 1 | Short | | |
| 8 | tt0000007 | short | Corbett and Courtney Be | Corbett and Courtney Before the Kinetograph | 0 | 1894 | \N | 1 | Short,Sport | | |
| 9 | tt0000008 | short | Edison Kinetoscopic Rec | Edison Kinetoscopic Record of a Sneeze | 0 | 1894 | \N | 1 | Documentary,Short | | |
| 10 | tt0000009 | movie | Miss Jerry | Miss Jerry | 0 | 1894 | \N | 45 | Romance | | |
| 11 | tt0000010 | short | Leaving the Factory | La sortie de l'usine LumiÃ¨re Ã  Lyon | 0 | 1895 | \N | 1 | Documentary,Short | | |
| 12 | tt0000011 | short | Akrobatisches Potpourri | Akrobatisches Potpourri | 0 | 1895 | \N | 1 | Documentary,Short | | |
| 13 | tt0000012 | short | The Arrival of a Train | L'arrivÃ©e d'un train Ã  La Ciotat | 0 | 1896 | \N | 1 | Documentary,Short | | |
| 14 | tt0000013 | short | The Photographical Con | Le dÃ©barquement du congrÃ¨s de photograph | 0 | 1895 | \N | 1 | Documentary,Short | | |
| 15 | tt0000014 | short | The Waterer Watered | L'arroseur arrosÃ© | 0 | 1895 | \N | 1 | Comedy,Short | | |
| 16 | tt0000015 | short | Autour d'une cabine | Autour d'une cabine | 0 | 1894 | \N | 2 | Animation,Short | | |
| 17 | tt0000016 | short | Boat Leaving the Port | Barque sortant du port | 0 | 1895 | \N | 1 | Documentary,Short | | |
| 18 | tt0000017 | short | Italienischer Bauerntanz | Italienischer Bauerntanz | 0 | 1895 | \N | 1 | Documentary,Short | | |
| 19 | tt0000018 | short | Das boxende KÃ¤nguruh | Das boxende KÃ¤nguruh | 0 | 1895 | \N | 1 | Short | | |
| 20 | tt0000019 | short | The Clown Barber | The Clown Barber | 0 | 1898 | \N | \N | Comedy,Short | | |
| 21 | tt0000020 | short | The Derby 1895 | The Derby 1895 | 0 | 1895 | \N | 1 | Documentary,Short,Sport | | |
| 22 | tt0000022 | short | Blacksmith Scene | Les forgerons | 0 | 1895 | \N | 1 | Documentary,Short | | |
| 23 | tt0000023 | short | The Sea | Baignade en mer | 0 | 1895 | \N | 1 | Documentary,Short | | |
| 24 | tt0000024 | short | Opening of the Kiel Cana | Opening of the Kiel Canal | 0 | 1895 | \N | \N | News,Short | | |
| 25 | tt0000025 | short | The Oxford and Cambric | The Oxford and Cambridge University Boat Race | 0 | 1896 | \N | \N | News,Short,Sport | | |
| 26 | tt0000026 | short | The Messers. LumiÃ¨re a | Partie d'Ã©cartÃ© | 0 | 1896 | \N | 1 | Documentary,Short | | |
| 27 | tt0000027 | short | Cordeliers' Square in Lyc | Place des Cordeliers Ã  Lyon | 0 | 1895 | \N | 1 | Documentary,Short | | |
| 28 | tt0000028 | short | Fishing for Goldfish | La pÃªche aux poissons rouges | 0 | 1895 | \N | 1 | Documentary,Short | | |
| 29 | tt0000029 | short | Baby's Meal | Repas de bÃ©bÃ© | 0 | 1895 | \N | 1 | Documentary,Short | | |
| 30 | tt0000030 | short | Rough Sea at Dover | Rough Sea at Dover | 0 | 1895 | \N | 1 | Documentary,Short | | |

data

**Fig. Dataset**

3. **df_ratings:** Contains the IMDb rating and votes information for titles

| **Name** | **Description** |
|---|---|
| tconst (string) | alphanumeric unique identifier of the title |
| averageRating | weighted average of all the individual user ratings |
| numVotes | number of votes the title has received |

| | A | B | C |
|---|---|---|---|
| 1 | tconst | averageRating | numVotes |
| 2 | tt0000001 | 5.7 | 1956 |
| 3 | tt0000002 | 5.8 | 263 |
| 4 | tt0000003 | 6.5 | 1789 |
| 5 | tt0000004 | 5.6 | 179 |
| 6 | tt0000005 | 6.2 | 2593 |
| 7 | tt0000006 | 5.1 | 177 |
| 8 | tt0000007 | 5.4 | 812 |
| 9 | tt0000008 | 5.4 | 2096 |
| 10 | tt0000009 | 5.3 | 204 |
| 11 | tt0000010 | 6.9 | 7073 |
| 12 | tt0000011 | 5.3 | 364 |
| 13 | tt0000012 | 7.4 | 12110 |
| 14 | tt0000013 | 5.7 | 1866 |
| 15 | tt0000014 | 7.1 | 5446 |
| 16 | tt0000015 | 6.2 | 1069 |
| 17 | tt0000016 | 5.9 | 1485 |
| 18 | tt0000017 | 4.6 | 323 |
| 19 | tt0000018 | 5.3 | 591 |
| 20 | tt0000019 | 5.1 | 31 |
| 21 | tt0000020 | 4.8 | 355 |
| 22 | tt0000022 | 5.1 | 1086 |
| 23 | tt0000023 | 5.7 | 1424 |
| 24 | tt0000024 | 4.2 | 110 |
| 25 | tt0000025 | 3.8 | 46 |
| 26 | tt0000026 | 5.6 | 1527 |
| 27 | tt0000027 | 5.6 | 1143 |
| 28 | tt0000028 | 5.1 | 1070 |
| 29 | tt0000029 | 5.9 | 3326 |
| 30 | tt0000030 | 5.2 | 842 |

**Fig. Dataset**

**4.df_nameBasics:** Contains the following information for names.

| **Name** | **Description** |
|---|---|
| nconst (string) | alphanumeric unique identifier of the name/person |
| primaryName (string) | name by which the person is most often credited |
| birthYear | in YYYY format |
| deathYear | in YYYY format if applicable, else '\N' |
| primaryProfession (array of strings) | the top-3 professions of the person |
| knownForTitles (array of tconsts) | titles the person is known for |

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | nconst | primaryName | birthYear | deathYear | primaryProfession | knownForTitles | |
| 2 | nm0000001 | Fred Astaire | 1899 | 1987 | soundtrack,actor,miscellaneous | tt0050419,tt0053137,tt0045537,tt0072308 | |
| 3 | nm0000002 | Lauren Bacall | 1924 | 2014 | actress,soundtrack | tt0037382,tt0071877,tt0038355,tt0117057 | |
| 4 | nm0000003 | Brigitte Bardot | 1934 | \N | actress,soundtrack,music_department | tt0057345,tt0056404,tt0054452,tt0049189 | |
| 5 | nm0000004 | John Belushi | 1949 | 1982 | actor,soundtrack,writer | tt0078723,tt0080455,tt0077975,tt0072562 | |
| 6 | nm0000005 | Ingmar Bergman | 1918 | 2007 | writer,director,actor | tt0083922,tt0050976,tt0050986,tt0060827 | |
| 7 | nm0000006 | Ingrid Bergman | 1915 | 1982 | actress,soundtrack,producer | tt0038109,tt0034583,tt0038787,tt0036855 | |
| 8 | nm0000007 | Humphrey Bogart | 1899 | 1957 | actor,soundtrack,producer | tt0037382,tt0043265,tt0042593,tt0034583 | |
| 9 | nm0000008 | Marlon Brando | 1924 | 2004 | actor,soundtrack,director | tt0068646,tt0078788,tt0047296,tt0070849 | |
| 10 | nm0000009 | Richard Burton | 1925 | 1984 | actor,soundtrack,producer | tt0087803,tt0057877,tt0059749,tt0061184 | |
| 11 | nm0000010 | James Cagney | 1899 | 1986 | actor,soundtrack,director | tt0031867,tt0035575,tt0029870,tt0042041 | |
| 12 | nm0000011 | Gary Cooper | 1901 | 1961 | actor,soundtrack,stunts | tt0035896,tt0034167,tt0044706,tt0027996 | |
| 13 | nm0000012 | Bette Davis | 1908 | 1989 | actress,soundtrack,make_up_departm | tt0056687,tt0042192,tt0031210,tt0035140 | |
| 14 | nm0000013 | Doris Day | 1922 | 2019 | soundtrack,actress,producer | tt0049470,tt0045591,tt0048317,tt0053172 | |
| 15 | nm0000014 | Olivia de Havilland | 1916 | 2020 | actress,soundtrack | tt0029843,tt0040806,tt0031381,tt0041452 | |
| 16 | nm0000015 | James Dean | 1931 | 1955 | actor,miscellaneous | tt0039123,tt0049261,tt0048545,tt0048028 | |
| 17 | nm0000016 | Georges Delerue | 1925 | 1992 | composer,soundtrack,music_departm | tt0096320,tt8847712,tt0069946,tt0091763 | |
| 18 | nm0000017 | Marlene Dietrich | 1901 | 1992 | soundtrack,actress,music_department | tt0051201,tt0052311,tt0055031,tt0021156 | |
| 19 | nm0000018 | Kirk Douglas | 1916 | 2020 | actor,producer,soundtrack | tt0054331,tt0050825,tt0049456,tt0043338 | |
| 20 | nm0000019 | Federico Fellini | 1920 | 1993 | writer,director,actor | tt0056801,tt0053779,tt0071129,tt0050783 | |
| 21 | nm0000020 | Henry Fonda | 1905 | 1982 | actor,producer,soundtrack | tt0051207,tt0032551,tt0082846,tt0050083 | |
| 22 | nm0000021 | Joan Fontaine | 1917 | 2013 | actress,soundtrack,producer | tt0032976,tt0034248,tt0040536,tt0035751 | |
| 23 | nm0000022 | Clark Gable | 1901 | 1960 | actor,soundtrack,producer | tt0031381,tt0026752,tt0023382,tt0025316 | |
| 24 | nm0000023 | Judy Garland | 1922 | 1969 | soundtrack,actress | tt0032138,tt0037059,tt0047522,tt0055031 | |
| 25 | nm0000024 | John Gielgud | 1904 | 2000 | actor,writer,director | tt0045943,tt0071877,tt0082031,tt0117631 | |
| 26 | nm0000025 | Jerry Goldsmith | 1929 | 2004 | music_department,soundtrack,compo | tt0112715,tt0119488,tt0077269,tt0117731 | |
| 27 | nm0000026 | Cary Grant | 1904 | 1986 | actor,soundtrack,producer | tt0038787,tt0053125,tt0034248,tt0056923 | |
| 28 | nm0000027 | Alec Guinness | 1914 | 2000 | actor,soundtrack,writer | tt0041546,tt0050212,tt0051739,tt0076759 | |
| 29 | nm0000028 | Rita Hayworth | 1918 | 1987 | actress,soundtrack,producer | tt0036723,tt0040525,tt0038559,tt0035103 | |
| 30 | nm0000029 | Margaux Hemingway | 1954 | 1996 | actress,miscellaneous | tt0077800,tt0110138,tt0102122,tt0074802 | |

**Fig. Dataset**

# Dataframes created:

## 1. movie_Recommender_df:

| Name | Description |
|---|---|
| movieID | The ID of the movie and TV movie |
| movieTitle | The title of the movie and TV movie |
| year | The year in which the movie or TV movie was released |
| genres | The genres of the movie or TV movie |
| directorId | The director ID of the director of movie or TV movie |
| directorName | The director name of movie or TV movie |
| averageRating | The average user rating of the movie or TV movie |
| numVotes | The number of user votes for the movie or TV movie |
| category | The category of job that person was in |

| | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | movie_id | category | director_id | director_name | average_R | num_Vote | titleType | movie_title | year | genres | | |
| 2 | tt0000630 | director | nm0143333 | Mario Caserini | 2.8 | 26 | movie | Amleto | 1908 | Drama | | |
| 3 | tt0000675 | director | nm0194088 | Narciso CuyÃ s | 4.2 | 20 | movie | Don Quijote | 1908 | Drama | | |
| 4 | tt0000862 | director | nm0878467 | Emanuel Tvede | 4.4 | 17 | movie | Faldgruben | 1909 | \N | | |
| 5 | tt0000941 | director | nm0550220 | Alberto Marro | 4.5 | 24 | movie | Locura de amor | 1909 | Drama | | |
| 6 | tt0000941 | director | nm0063413 | Ricardo de BaÃ±os | 4.5 | 24 | movie | Locura de amor | 1909 | Drama | | |
| 7 | tt0001112 | director | nm0143333 | Mario Caserini | 3.8 | 43 | movie | Amleto | 1910 | Drama | | |
| 8 | tt0001790 | director | nm0135052 | Albert Capellani | 6.2 | 51 | movie | Les misÃ©rables - Ã% | 1913 | Drama | | |
| 9 | tt0001911 | director | nm0519315 | Raymond Longford | 3.6 | 24 | movie | Sweet Nell of Old Dru | 1911 | Biography,Drama,History | | |
| 10 | tt0002026 | director | nm0259235 | Adam Eriksen | 4.5 | 14 | movie | Anny - en gatepiges ro | 1912 | Drama,Romance | | |
| 11 | tt0002375 | director | nm0135052 | Albert Capellani | 5.7 | 12 | movie | La mort du duc d'Engh | 1912 | \N | | |
| 12 | tt0002423 | director | nm0523932 | Ernst Lubitsch | 6.6 | 928 | movie | Madame DuBarry | 1919 | Biography,Drama,Romance | | |
| 13 | tt0002588 | director | nm0419327 | Victorin-Hippolyte Ja | 5.9 | 44 | movie | Zigomar contre Nick C | 1912 | Crime,Thriller | | |
| 14 | tt0002591 | director | nm0296193 | Carl Froelich | 6.2 | 10 | movie | Zu spÃ¤t | 1913 | \N | | |
| 15 | tt0002669 | director | nm0316794 | Charles Giblyn | 6.7 | 39 | movie | The Battle of Gettysbu | 1913 | Drama,War | | |
| 16 | tt0002669 | director | nm0408436 | Thomas H. Ince | 6.7 | 39 | movie | The Battle of Gettysbu | 1913 | Drama,War | | |
| 17 | tt0002844 | director | nm0275421 | Louis Feuillade | 6.9 | 2358 | movie | FantÃ´mas - Ã€ l'ombr | 1913 | Crime,Drama | | |
| 18 | tt0002885 | director | nm0938041 | Frank E. Wolfe | 6 | 110 | movie | From Dusk to Dawn | 1913 | Drama | | |
| 19 | tt0003037 | director | nm0275421 | Louis Feuillade | 6.9 | 1601 | movie | Juve contre FantÃ´ma | 1913 | Crime,Drama | | |
| 20 | tt0003131 | director | nm0532622 | Alfred Machin | 6.7 | 167 | movie | Maudite soit la guerre | 1914 | Drama,War | | |
| 21 | tt0003241 | director | nm0532349 | Norval MacGregor | 5 | 21 | movie | One Hundred Years of | 1913 | Drama,History | | |
| 22 | tt0003330 | director | nm0296193 | Carl Froelich | 6.3 | 117 | movie | Richard Wagner | 1913 | Biography,Drama,History | | |
| 23 | tt0003330 | director | nm0915270 | William Wauer | 6.3 | 117 | movie | Richard Wagner | 1913 | Biography,Drama,History | | |
| 24 | tt0003565 | director | nm0533048 | Max Mack | 6 | 37 | movie | Wo ist Coletti? | 1913 | Comedy,Crime | | |
| 25 | tt0003668 | director | nm0281621 | Caryl S. Fleming | 5.6 | 23 | movie | Beating Back | 1914 | Adventure,Biography,Western | | |
| 26 | tt0003816 | director | nm0877783 | Otis Turner | 5.8 | 39 | movie | Damon and Pythias | 1914 | Drama | | |
| 27 | tt0004336 | director | nm0360617 | Howell Hansel | 6 | 40 | movie | The Million Dollar Mys | 1914 | Adventure,Music,Mystery | | |
| 28 | tt0004363 | director | nm0373614 | Thomas N. Heffron | 7.2 | 19 | movie | Mrs. Black Is Back | 1914 | Comedy | | |
| 29 | tt0004398 | director | nm0205986 | J. Searle Dawley | 1.4 | 20 | movie | The Next in Commanc | 1914 | Adventure | | |
| 30 | tt0004630 | director | nm0132324 | Colin Campbell | 6.1 | 87 | movie | The Spoilers | 1914 | Drama,Western | | |

**Fig. Dataset**

## 2.Actor_df_main:

| Name | Description |
|---|---|
| movieID | The ID of the movie and TV movie |
| movieTitle | The title of the movie and TV movie |
| year | The year in which the movie or TV movie was released |
| genres | The genres of the movie or TV movie |
| Actor_id | The Actor ID of the Actor of the movie or TV movie |
| Actor_Name | The Actor’s name of the movie or TV movie |
| averageRating | The average user rating of the movie or TV movie |
| numVotes | The number of user votes for the movie or TV movie |
| category | The category of job that person was in |

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | movie_id | category | Actor_id | Actor_name | average_R | num_Vote | titleType | movie_title | year | genres |
| 2 | tt0000862 | actor | nm0386036 | Carl Hintz | 4.4 | 17 | movie | Faldgruben | 1909 | \N |
| 3 | tt0000862 | actor | nm0511080 | SchiÃ¸ler Linck | 4.4 | 17 | movie | Faldgruben | 1909 | \N |
| 4 | tt0000862 | actor | nm5188470 | Carl Johan Lundkvist | 4.4 | 17 | movie | Faldgruben | 1909 | \N |
| 5 | tt0000862 | actor | nm5289829 | Hr. Andreasen | 4.4 | 17 | movie | Faldgruben | 1909 | \N |
| 6 | tt0000862 | actor | nm5289318 | O. Poulsen | 4.4 | 17 | movie | Faldgruben | 1909 | \N |
| 7 | tt0000941 | actor | nm0034453 | JosÃ© ArgelaguÃ©s | 4.5 | 24 | movie | Locura de amor | 1909 | Drama |
| 8 | tt0000941 | actor | nm0140054 | JoaquÃ-n Carrasco | 4.5 | 24 | movie | Locura de amor | 1909 | Drama |
| 9 | tt0000941 | actor | nm0243918 | JosÃ© Durany | 4.5 | 24 | movie | Locura de amor | 1909 | Drama |
| 10 | tt0001112 | actor | nm0135493 | Dante Cappelli | 3.8 | 43 | movie | Amleto | 1910 | Drama |
| 11 | tt0001531 | actor | nm0738202 | Alfred Rolfe | 4.6 | 15 | movie | Captain Starlight, or Gentlem | 1911 | \N |
| 12 | tt0001531 | actor | nm0627427 | Augustus Neville | 4.6 | 15 | movie | Captain Starlight, or Gentlem | 1911 | \N |
| 13 | tt0001531 | actor | nm0909492 | Stanley Walpole | 4.6 | 15 | movie | Captain Starlight, or Gentlem | 1911 | \N |
| 14 | tt0001790 | actor | nm0959921 | Henri Ã‰tiÃ©vant | 6.2 | 51 | movie | Les misÃ©rables - Ã‰poque | 1913 | Drama |
| 15 | tt0001790 | actor | nm0470307 | Henry Krauss | 6.2 | 51 | movie | Les misÃ©rables - Ã‰poque | 1913 | Drama |
| 16 | tt0001812 | actor | nm0294276 | Theo Frenkel | 5.5 | 14 | movie | Oedipus Rex | 1911 | Drama |
| 17 | tt0001911 | actor | nm0167411 | Stewart Clyde | 3.6 | 24 | movie | Sweet Nell of Old Drury | 1911 | Biography,Drama,History |
| 18 | tt0001911 | actor | nm0492661 | Charles Lawrence | 3.6 | 24 | movie | Sweet Nell of Old Drury | 1911 | Biography,Drama,History |
| 19 | tt0001911 | actor | nm0627427 | Augustus Neville | 3.6 | 24 | movie | Sweet Nell of Old Drury | 1911 | Biography,Drama,History |
| 20 | tt0002026 | actor | nm0064944 | EugÃ¨ne Bech | 4.5 | 14 | movie | Anny - en gatepiges roman | 1912 | Drama,Romance |
| 21 | tt0002026 | actor | nm0115982 | Ole Brun Lie | 4.5 | 14 | movie | Anny - en gatepiges roman | 1912 | Drama,Romance |
| 22 | tt0002026 | actor | nm0959066 | Waldemar Zwinge | 4.5 | 14 | movie | Anny - en gatepiges roman | 1912 | Drama,Romance |
| 23 | tt0002026 | actor | nm0027708 | Johan Andersson | 4.5 | 14 | movie | Anny - en gatepiges roman | 1912 | Drama,Romance |
| 24 | tt0002375 | actor | nm0503415 | RenÃ© Leprince | 5.7 | 12 | movie | La mort du duc d'Enghien | 1912 | \N |
| 25 | tt0002375 | actor | nm0135053 | Paul Capellani | 5.7 | 12 | movie | La mort du duc d'Enghien | 1912 | \N |
| 26 | tt0002375 | actor | nm0959921 | Henri Ã‰tiÃ©vant | 5.7 | 12 | movie | La mort du duc d'Enghien | 1912 | \N |
| 27 | tt0002375 | actor | nm0578805 | Daniel Mendaille | 5.7 | 12 | movie | La mort du duc d'Enghien | 1912 | \N |
| 28 | tt0002423 | actor | nm0509573 | Harry Liedtke | 6.6 | 929 | movie | Madame DuBarry | 1919 | Biography,Drama,Romance |
| 29 | tt0002423 | actor | nm0417837 | Emil Jannings | 6.6 | 929 | movie | Madame DuBarry | 1919 | Biography,Drama,Romance |
| 30 | tt0002423 | actor | nm0903235 | Eduard von Winterstein | 6.6 | 929 | movie | Madame DuBarry | 1919 | Biography,Drama,Romance |
| 31 | tt0002588 | actor | nm1979952 | Charles Krauss | 5.9 | 44 | movie | Zigomar contre Nick Carter | 1912 | Crime,Thriller |

**Fig. Dataset**

# 3. actress_df_main:

| Name | Description |
|---|---|
| movieID | The ID of the movie and TV movie |
| movieTitle | The title of the movie and TV movie |
| year | The year in which the movie or TV movie was released |
| genres | The genres of the movie or TV movie |
| actress_id | The Actress ID of the Actress of the movie or TV movie |
| actress_Name | The Actress's name of the movie or TV movie |
| averageRating | The average user rating of the movie or TV movie |
| numVotes | The number of user votes for the movie or TV movie |
| category | The category of job that person was in |

| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | movie_id | category | actress_id | actress_name | average_R | num_Vote | titleType | movie_title | year | genres | |
| 2 | tt0000630 | actress | nm0624446 | Fernanda Negri Pouget | 2.8 | 26 | movie | Amleto | 1908 | Drama | |
| 3 | tt0000862 | actress | nm0264569 | Kate Fabian | 4.4 | 17 | movie | Faldgruben | 1909 | \N | |
| 4 | tt0000941 | actress | nm0294022 | Elvira Fremont | 4.5 | 24 | movie | Locura de amor | 1909 | Drama | |
| 5 | tt0001112 | actress | nm0143332 | Maria Caserini | 3.8 | 43 | movie | Amleto | 1910 | Drama | |
| 6 | tt0001115 | actress | nm0630641 | Marie Niedermann | 4.6 | 20 | movie | Ansigttyven I | 1910 | Crime | |
| 7 | tt0001498 | actress | nm0768187 | Laura Sawyer | 8 | 13 | movie | The Battle of Trafalgar | 1911 | War | |
| 8 | tt0001531 | actress | nm0198972 | Lily Dampier | 4.6 | 15 | movie | Captain Starlight, or Gentleman c | 1911 | \N | |
| 9 | tt0001531 | actress | nm0528022 | Lottie Lyell | 4.6 | 15 | movie | Captain Starlight, or Gentleman c | 1911 | \N | |
| 10 | tt0001790 | actress | nm0893346 | Maria Ventura | 6.2 | 51 | movie | Les misÃ©rables - Ã‰poque 1: J | 1913 | Drama | |
| 11 | tt0001790 | actress | nm0592965 | Mistinguett | 6.2 | 51 | movie | Les misÃ©rables - Ã‰poque 1: J | 1913 | Drama | |
| 12 | tt0001812 | actress | nm0207207 | Suzanne de Baere | 5.5 | 14 | movie | Oedipus Rex | 1911 | Drama | |
| 13 | tt0001911 | actress | nm0829692 | Nellie Stewart | 3.6 | 24 | movie | Sweet Nell of Old Drury | 1911 | Biography,Drama,History | |
| 14 | tt0002026 | actress | nm0526167 | Gunlaug Lund | 4.5 | 14 | movie | Anny - en gatepiges roman | 1912 | Drama,Romance | |
| 15 | tt0002026 | actress | nm0418086 | Julie Jansen-Fuhr | 4.5 | 14 | movie | Anny - en gatepiges roman | 1912 | Drama,Romance | |
| 16 | tt0002026 | actress | nm0959065 | Fru Zwinge | 4.5 | 14 | movie | Anny - en gatepiges roman | 1912 | Drama,Romance | |
| 17 | tt0002026 | actress | nm0348052 | Aagot Gundersen | 4.5 | 14 | movie | Anny - en gatepiges roman | 1912 | Drama,Romance | |
| 18 | tt0002153 | actress | nm1069712 | Agnes Lorentzen | 6 | 76 | movie | DÃ¸dsspring til hest fra cirkuskup | 1912 | Drama | |
| 19 | tt0002153 | actress | nm0385541 | Alma Hinding | 6 | 76 | movie | DÃ¸dsspring til hest fra cirkuskup | 1912 | Drama | |
| 20 | tt0002199 | actress | nm0310155 | Gene Gauntier | 5.8 | 609 | movie | From the Manger to the Cross | 1912 | Biography,Drama | |
| 21 | tt0002199 | actress | nm0391220 | Alice Hollister | 5.8 | 609 | movie | From the Manger to the Cross | 1912 | Biography,Drama | |
| 22 | tt0002375 | actress | nm0180078 | Nelly Cormon | 5.7 | 12 | movie | La mort du duc d'Enghien | 1912 | \N | |
| 23 | tt0002406 | actress | nm0606530 | Flora Morris | 4.8 | 24 | movie | Oliver Twist | 1912 | Drama | |
| 24 | tt0002406 | actress | nm0851953 | Alma Taylor | 4.8 | 24 | movie | Oliver Twist | 1912 | Drama | |
| 25 | tt0002406 | actress | nm0587610 | Ivy Millais | 4.8 | 24 | movie | Oliver Twist | 1912 | Drama | |
| 26 | tt0002423 | actress | nm0624470 | Pola Negri | 6.6 | 929 | movie | Madame DuBarry | 1919 | Biography,Drama,Romance | |
| 27 | tt0002588 | actress | nm0218469 | Olga Demidoff | 5.9 | 44 | movie | Zigomar contre Nick Carter | 1912 | Crime,Thriller | |
| 28 | tt0002588 | actress | nm0029029 | Josette Andriot | 5.9 | 44 | movie | Zigomar contre Nick Carter | 1912 | Crime,Thriller | |
| 29 | tt0002591 | actress | nm0029806 | Martha Angerstein-Licho | 6.2 | 10 | movie | Zu spÃ¤t | 1913 | \N | |
| 30 | tt0002669 | actress | nm0514517 | Ann Little | 6.7 | 39 | movie | The Battle of Gettysburg | 1913 | Drama,War | |

**Fig. Dataset**

## 5. Data Pre-processing and Cleaning

- Importing Data: We have imported 4 tsv files and merged them to create our dataset.
- Feature Selection: We manually selected features using basic domain knowledge.
- Missing Data: There is no missing data in the dataset.
- Data Type: We have two types of data in our dataset: Categorical and Numerical.
- The label "movieTitle" serves as the target variable, and our objective is to generate movie recommendations based on the user's interests.

# 6. Exploratory Data Analysis

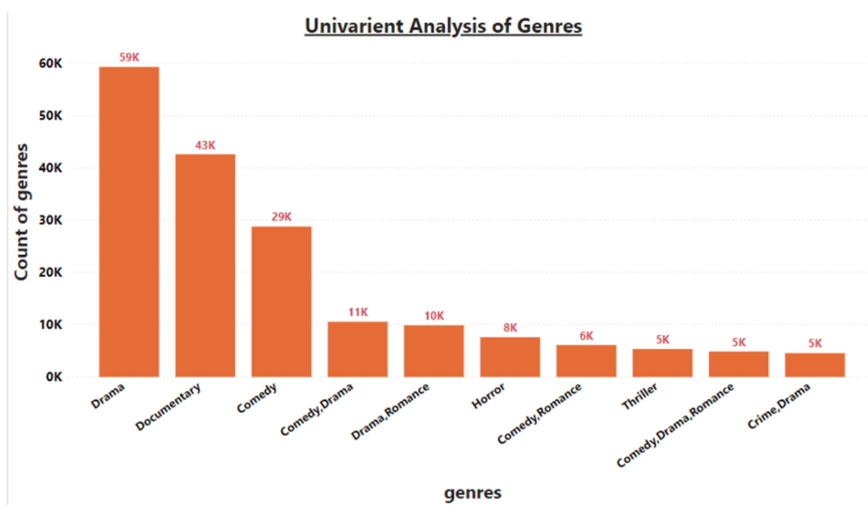
Exploratory Data Analysis (EDA) is a crucial step in analyzing data where an initial investigation is conducted to uncover any underlying patterns, identify anomalies, test hypotheses, and validate assumptions. This is achieved through the use of summary statistics and graphical visualizations, to analyze the data's distribution, relationships, and trends. The primary goal of EDA is to gain a comprehensive understanding of the data, which can inform further analysis and modeling.

**Univariate Data Analysis:** Univariate analysis is a statistical analysis technique that focuses on examining one variable at a time. Univariate analysis can provide useful insights into the characteristics of a single variable, such as the range of values it takes, its central tendency, and how spread out the data is. The following are the Univariate Analysis we conducted:

### 1. What is the count of Genres (top 10)?

In our dataset independent variable is 'Genres'. We are counting each number of genres.

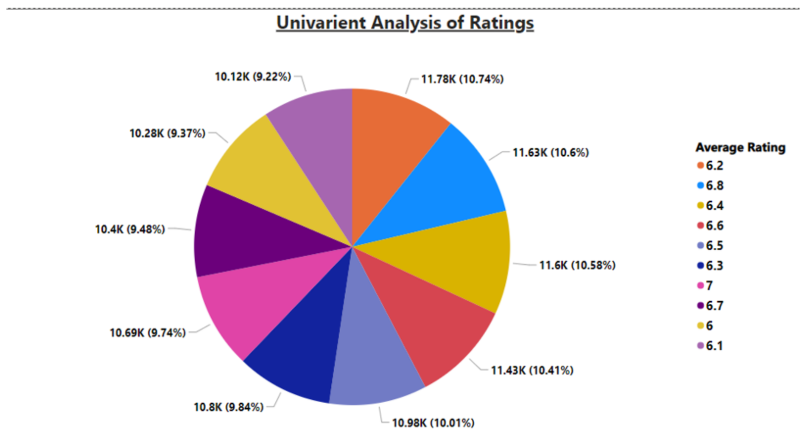
| genres | Count of genres |
|---|---|
| Comedy | 28714 |
| Comedy,Drama | 10532 |
| Comedy,Drama,Romance | 4859 |
| Comedy,Romance | 6099 |
| Crime,Drama | 4552 |
| Documentary | 42523 |
| Drama | 59207 |
| Drama,Romance | 9861 |
| Horror | 7597 |
| Thriller | 5344 |
| **Total** | **179288** |




**Conclusion:** From above graph, we come to know that the count of movies in genre 'Drama' is around 59K, followed by genres 'Documentary' and 'Comedy'. We can infer that movies in the genre 'Drama' are made and released more worldwide, followed by the genres 'Documentary' and 'Comedy'.

## 2. Which rating has been given the maximum number times to movies?

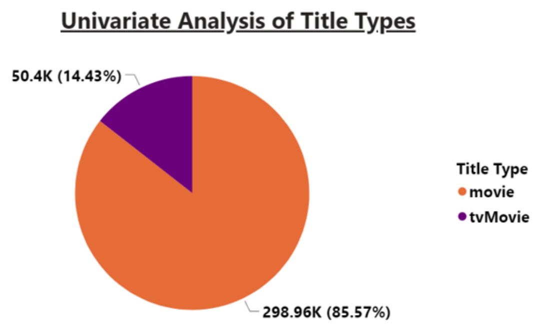
| | averageRating | count(averageRating) |
|---|---|---|
| **0** | 6.2 | 11779 |
| **1** | 6.8 | 11629 |
| **2** | 6.4 | 11604 |
| **3** | 6.6 | 11425 |
| **4** | 6.5 | 10982 |
| **5** | 6.3 | 10796 |
| **6** | 7.0 | 10689 |
| **7** | 6.7 | 10404 |
| **8** | 6.0 | 10284 |
| **9** | 6.1 | 10119 |




**Conclusion:** Based on the graph above, we can observe that a large number of movies have been rated an average rating of 6.2. This suggests that 6.2 is the most common rating that movies receive. Therefore, we can infer that the majority of movies in the dataset are rated around 6.2.

3. **What percentage of titles in the dataset are categorized as movies, and what percentage are categorized as TV movies?**

| titleType | Count of titleType | %GT Count of titleType |
|---|---|---|
| movie | 298956 | 85.57% |
| tvMovie | 50404 | 14.43% |
| **Total** | **349360** | **100.00%** |

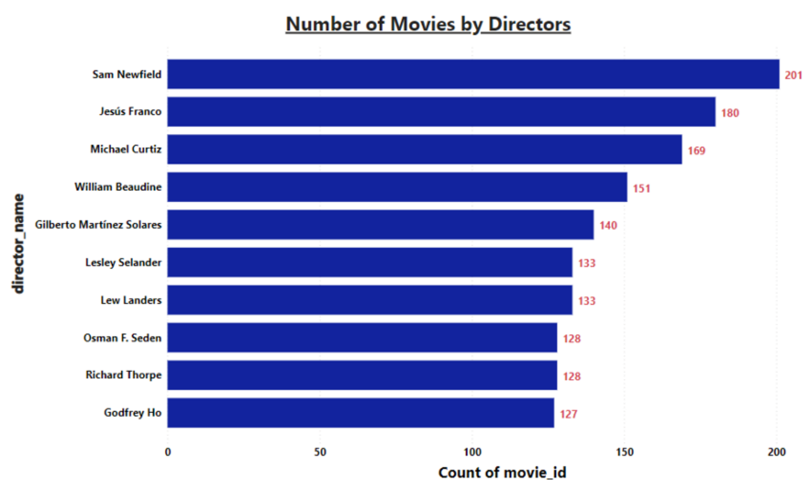



**Conclusion:** Based on the graph above, we can conclude that the majority of movies in the dataset are of the 'movie' title type, accounting for 85.57% of the total movies. In contrast, the 'tvMovie' title type accounts for only 14.43% of the movies. Therefore, we can infer that most movies in the dataset are made in the 'movie' format, rather than the 'tvMovie' format.

**<u>Bivariate analysis</u>:** Bivariate analysis is a statistical analysis technique that enables us to investigate the connection between two variables. Its objective is to analyze the potential correlations, patterns, and trends that exist between the two variables. Graphical representations, such as scatterplots, line graphs, and bar graphs, can also be used to visualize the relationship between the two variables.

## 1 How many movies are directed by each Director (top 10)?

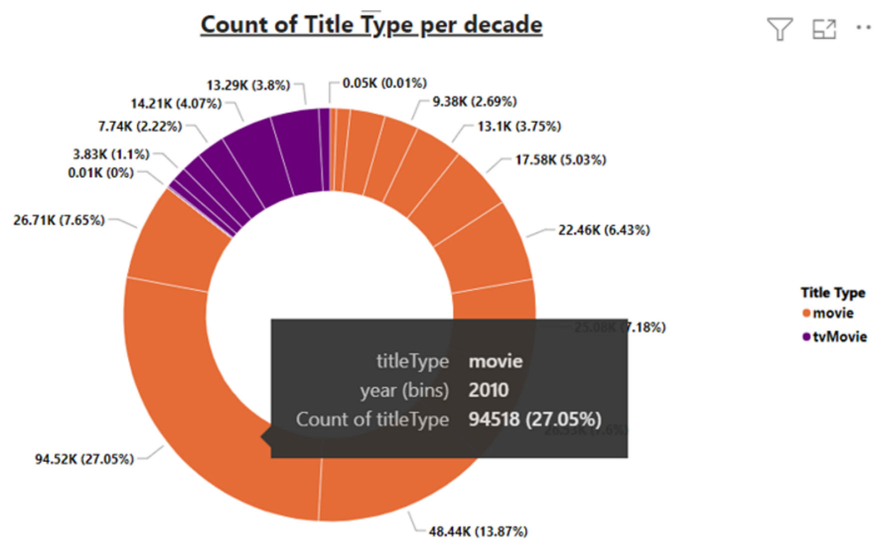
| | directorName | count(movieId) |
|---|---|---|
| 0 | Sam Newfield | 201 |
| 1 | Jesús Franco | 180 |
| 2 | Michael Curtiz | 169 |
| 3 | William Beaudine | 151 |
| 4 | Gilberto Martínez Solares | 140 |
| 5 | Lew Landers | 133 |
| 6 | Lesley Selander | 133 |
| 7 | Richard Thorpe | 128 |
| 8 | Osman F. Seden | 128 |
| 9 | Godfrey Ho | 127 |




**Conclusion:** Based on the chart above, we can observe that Sam Newfield is the director with the highest number of movies, having directed a total of 201 movies. The second highest number of movies is directed by Jesus Franco with 180 movies, and Michael Curtiz follows closely with 169 movies. Therefore, we can conclude that Sam Newfield has directed the most movies, followed by Jesus Franco and Michael Curtiz.

## 2. Show the number of Title types (movie, TV movie) per decade.

| titleType | Count of titleType | year (bins) |
|---|---|---|
| movie | 52 | |
| movie | 2 | 1890 |
| movie | 61 | 1900 |
| movie | 1727 | 1910 |
| movie | 3764 | 1920 |
| movie | 9547 | 1930 |
| movie | 9381 | 1940 |
| movie | 13101 | 1950 |
| movie | 17576 | 1960 |
| movie | 22461 | 1970 |
| movie | 25080 | 1980 |
| movie | 26534 | 1990 |
| movie | 48441 | 2000 |
| movie | 94518 | 2010 |
| movie | 26711 | 2020 |
| tvMovie | 5 | |
| tvMovie | 1 | 1920 |
| tvMovie | 9 | 1930 |
| tvMovie | 17 | 1940 |
| tvMovie | 460 | 1950 |
| tvMovie | 2387 | 1960 |
| tvMovie | 3830 | 1970 |
| tvMovie | 5467 | 1980 |
| tvMovie | 7742 | 1990 |
| tvMovie | 14207 | 2000 |
| tvMovie | 13292 | 2010 |
| tvMovie | 2987 | 2020 |
| **Total** | **349360** | |

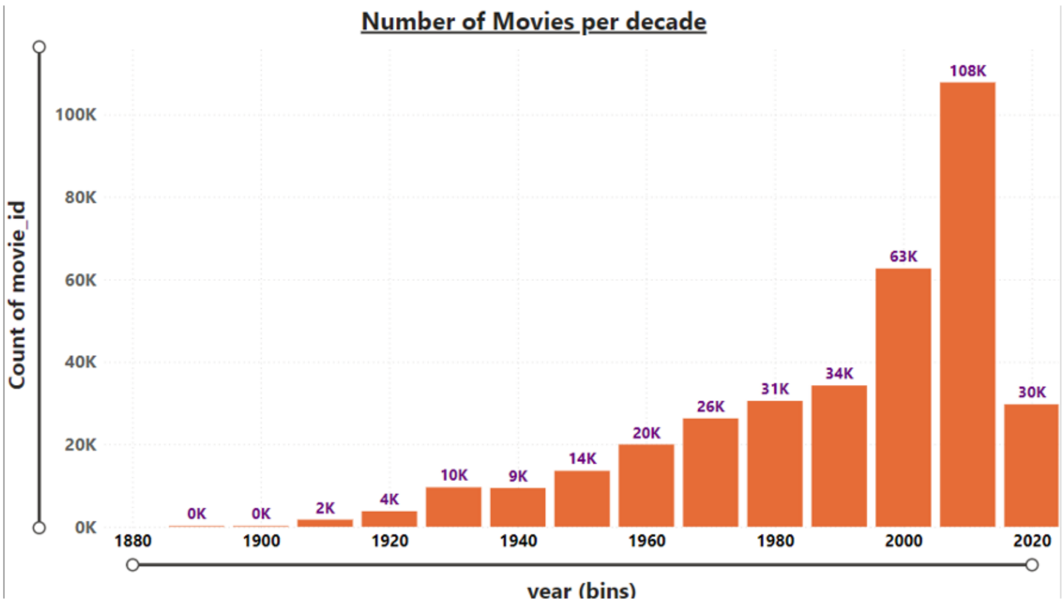



**Conclusion:** Based on the chart above, we can see that the highest number of movies released in any decade is 94,518, and they all have a title type of 'movie'. These movies were released during the decade from 2010 to 2020. Therefore, we can conclude that the decade from 2010 to 2020 saw the highest number of movie releases, having a title type of 'movie'.

According to the chart above, it is evident that the decade between 2000 to 2010 witnessed the maximum number of movies released with the title type 'tvMovie' - a total of 14,207. Hence, we can infer that this decade saw the most significant number of 'tvMovie' releases as compared to any other decade.

### 3. Show the Number of movies in each decade?

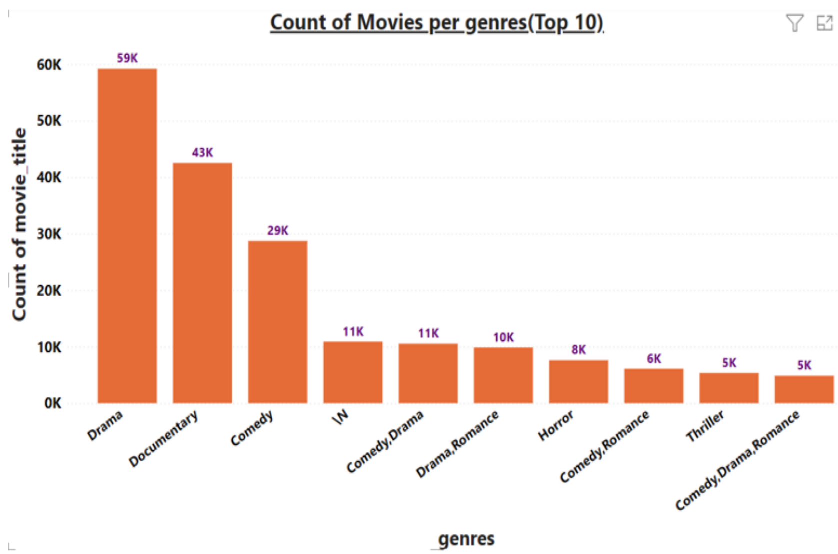
| year (bins) | Count of movie_id |
|---|---|
| | 57 |
| 1890 | 2 |
| 1900 | 61 |
| 1910 | 1727 |
| 1920 | 3765 |
| 1930 | 9556 |
| 1940 | 9398 |
| 1950 | 13561 |
| 1960 | 19963 |
| 1970 | 26291 |
| 1980 | 30547 |
| 1990 | 34276 |
| 2000 | 62648 |
| 2010 | 107810 |
| 2020 | 29698 |
| **Total** | **349360** |




**Conclusion**: Based on the chart above, we can see that the decade between 2010 to 2020 witnessed the highest number of movie releases, with a total count of 107,810 movies. Therefore, we can conclude that the decade from 2010 to 2020 saw the maximum number of movie releases as compared to any other decade.

## 4. Show the number of Movies present in each Genre (Top 10).

| genres | count(movieId) |
|---:|---:|
| Drama | 59207 |
| Documentary | 42523 |
| Comedy | 28714 |
| \N | 10901 |
| Comedy,Drama | 10532 |
| Drama,Romance | 9861 |
| Horror | 7597 |
| Comedy,Romance | 6099 |
| Thriller | 5344 |
| Comedy,Drama,Romance | 4859 |



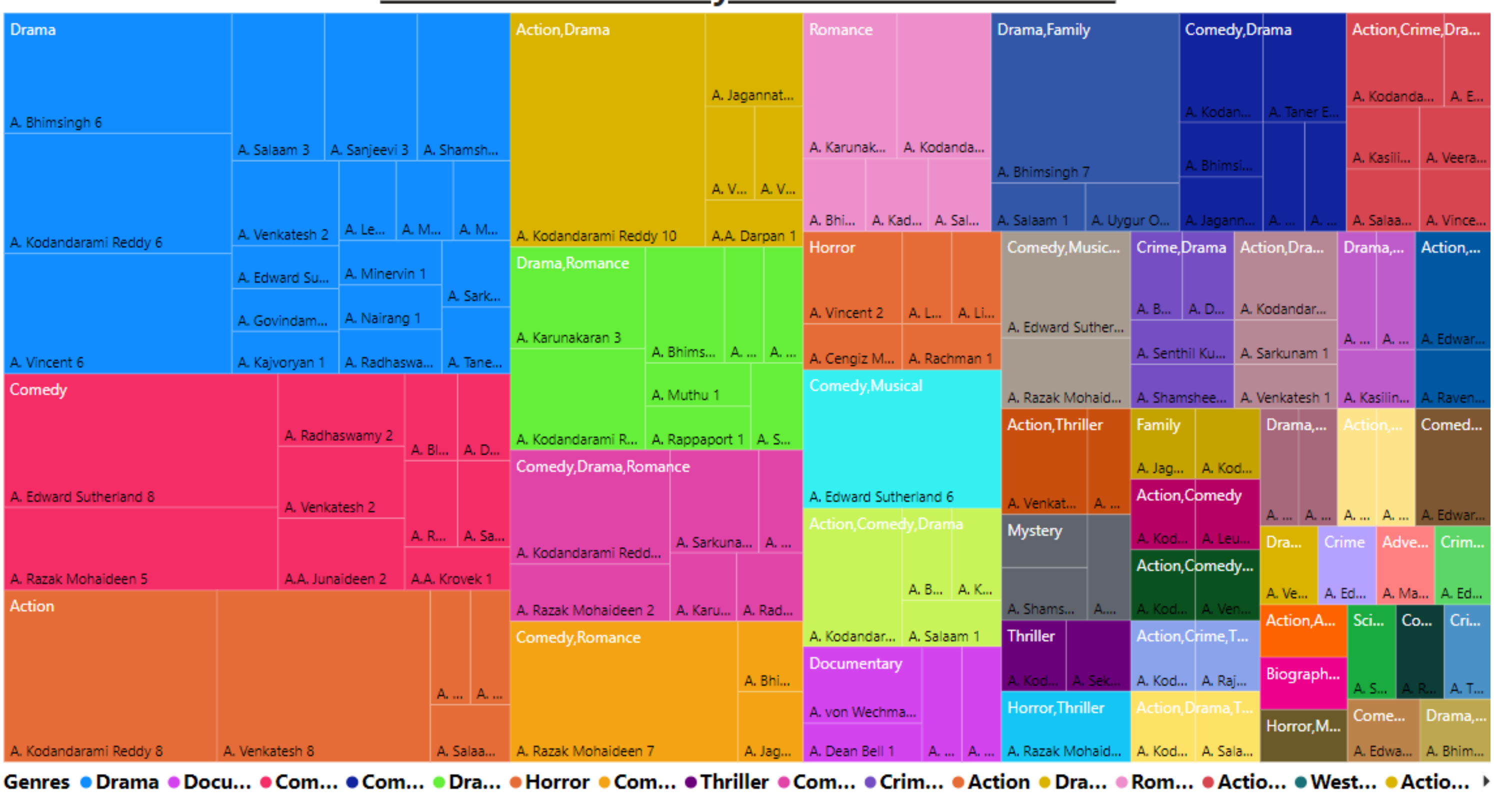

**Conclusion:** Based on the chart above, it can be observed that the genre with the highest number of movies is 'Drama', with a count of approximately 59,000. Following 'Drama' is the genre 'Documentary', with around 43,000 movies, and the genre 'Comedy', with around 29,000 movies. Therefore, we can conclude that 'Drama' is the most commonly made genre, followed by 'Documentary' and 'Comedy'.

It is also evident from the data that there is a set of movies, comprising around 10,901 titles, that have not been classified into any particular genre.

**Multivariate Data Analysis:** Multivariate analysis is a statistical analysis technique that involves the study of the relationship between multiple variables. It examines how multiple variables are interrelated and how they impact each other. The goal of multivariate analysis is to identify patterns and relationships between variables that may not be apparent in a bivariate analysis.

**Show the number of Movies directed by each Director for each Genre.**

Number of Movies by director for each Genre




**Conclusion:** It can be inferred from the chart that in the genre 'Drama', which has the highest number of movies, Directors A. Bhimsingh, A. Kodandarami Reddy, and A. Vincent have each directed 6 movies.

# 7. Model Building

## 1. Popularity-Based Filtering Algorithm:

A Popularity-based filtering algorithm for movie recommendations is one of the simplest and most widely used approaches in recommendation systems. This algorithm ranks movies based on their popularity or the number of ratings they have received, under the assumption that popular movies are generally appealing to a larger audience and therefore likely to be enjoyed by new users. This method is widely used and considered simple yet effective.

**Goal**:

As the name suggests, a popularity-based filtering algorithm works with trends, using movies that are popular or trending over time. For example, if a movie is frequently watched by many people, the system will recognize it as popular, and recommend it to new users who sign up. This increases the likelihood that the new users will watch the recommended movie as well. The algorithm's goal is to recommend movies that are popular, increasing the chances that users will enjoy the movies and find them relevant to their interests.
The measure of a movie's popularity can be determined using different metrics, which may include the number of views or the average rating it has received.

```
dataset_path = movie_Recommender_df
num_recommendations = 10

recommended_movies = popularity_based_recommendations(dataset_path, num_recommendations)

print(recommended_movies)
```

```
['The Silence of Swastika', 'Threat Level Midnight: The Movie', 'The Shawshank Redemption', 'Madhi', 'The Godfather', 'Hababam Sinifi', 'Viratapura Viraagi', 'Nee Jathaga', 'Ramayana: The Legend of Prince Rama', 'Ramayana: The Legend of Prince Rama']
```

## 2. Content Based Filtering Algorithm:

Content-based filtering is a recommendation algorithm that examines the attributes of movies to provide users with suggestions for similar movies. This approach uses the characteristics of the movies, such as genre and director, to generate personalized recommendations for users. The underlying concept behind this algorithm is that users who have enjoyed a particular movie are likely to be interested in other movies with similar content.

**Goal:**

This model will allow us to sort movies that are similar based on their genre, and filter out movies with similar types of content. For example, if a user enjoys movies with content related to action, sci-fi, drama, fantasy, and adventure, the model will recommend other movies with similar genres. The algorithm aims to suggest movies that align with the user's interests and preferences.

For ex: If a user likes movies such as "Avengers: Age of Ultron," then the model will recommend only those movies which are of the same genre. In this example, the user watches "Avengers: Age of Ultron," which is a sci-fi, adventure, and action movie. The model will recommend movies with similar genres or content, based on the user's past preferences and the information available in our dataset.

```
dataset_path = movie_Recommender_df
movie_title = 'Avengers: Age of Ultron'  # Enter a movie title, system will show similar to that movie
num_recommendations = 9

recommended_movies = content_based_recommendations(dataset_path, movie_title, num_recommendations)

print(recommended_movies)
```

```
['Planet of the Apes', 'Bumblebee', 'The Watchers: Revelation', 'Raiders of the Sun', 'Oblivion', 'Legendary', 'Dünyayi Kurtara
n Adam', 'Universal Ninjas', 'Star Trek: Temporal Anomaly']
```

### 3. Actor based filtering algorithm:

Actor-based filtering is a recommendation algorithm that focuses on the actors who appear in movies to suggest similar movies to users. This approach assumes that users who enjoy watching a particular actor are likely to enjoy other movies that feature the same performer.

For example, if a user has watched a movie in which Sidharth Malhotra plays the lead role, the actor-based filtering algorithm will recommend other movies that feature Sidharth Malhotra. As we can see in the screenshot, if the user inputs "Student of the Year," the model suggests movies like "Shershaah," "Aiyaary," "Mission Majnu," "Thank God," and so on.

```
dataset_path = Actor_df_main
movie_title = 'Student of the Year'   # Enter a movie title, system will show similar to that movie
num_recommendations = 10

recommended_movies = actor_based_recommendations(dataset_path, movie_title, num_recommendations)

print(recommended_movies)
```

```
['Brothers', 'Aiyaary', 'Shershaah', 'Mission Majnu', 'Ittefaq', 'Ek Villain', 'Marjaavaan', 'Thank God', 'Aankhen 2', 'A Gentl
eman']
```

By using this model, our target i.e., user, will stay engaged on the platform and will watch more movies.

### 4. Actress based filtering algorithm:

The actress-based filtering algorithm is a recommendation algorithm for movies that suggests movies to users based on the actresses who have starred in them. The algorithm works on the assumption that users who enjoy watching movies featuring a particular actress will also enjoy other films that star the same performer in a leading or supporting role.

For example, if a user has watched a movie in which Alia Bhatt has played the lead role, the actress-based filtering algorithm will suggest other movies featuring Alia Bhatt. For instance, as shown in the screenshot below, if the input is "Student of the Year," the model will recommend movies like "Gully Boy," "Raazi," "Darlings," "Dear Zindagi," and so on.

```
dataset_path = actress_df_main
movie_title = 'Student of the Year'   # Enter a movie title, system will show similar to that movie
num_recommendations = 10

recommended_movies = actress_based_recommendations(dataset_path, movie_title, num_recommendations)

print(recommended_movies)
```

```
['Raazi', 'Udta Punjab', 'RRR (Rise Roar Revolt)', 'Humpty Sharma Ki Dulhania', 'Sadak 2', 'Gully Boy: Live In Concert', 'Gangu
bai Kathiawadi', 'Darlings', 'Dear Zindagi', 'Gully Boy']
```

# 8. Conclusion

In summary, the Movie Recommendation and Analytics project was developed using technologies such as PySpark, SparkSQL, Machine Learning, and Data Visualization (Power BI). Through the use of these technologies, we created four algorithms: Popularity-based algorithm, Content-based algorithm, Actor-based algorithm, and Actress-based algorithm. The goal was to gain insights into movie trends, user preferences, and provide personalized recommendations.

# 9. Future Scope:

The project on Movie Recommendation and Analytics has a vast scope for future development and improvements.
One potential area for future work is incorporating more advanced recommendation algorithms. The project implemented Popularity-based algorithm, Content-based algorithm, Actor-based algorithm and Actress-based algorithm for recommendations. In the future, more advanced algorithms such as collaborative filtering, matrix factorization, and deep learning-based algorithms can be implemented to improve the accuracy of recommendations. Another potential area for future work is storing the dataset in HDFS or Cloud.

# 10. References

**Dataset link:** https://datasets.imdbws.com
name.basics.tsv.gz
title.basics.tsv.gz
title.principals.tsv.gz
title.ratings.tsv.gz

**Models:**
**knn model:** sklearn.neighbors.KNeighborsClassifier — scikit-learn 1.2.2 documentation