Final Project-1

IMDB Movie Analysis

Project Description

The project is about analyzing the IMDb movie dataset to gain insights about the movie industry. The dataset consists of information about movies like the budget, gross, IMDb rating, director name, actor name, and genre, etc. The aim of the project is to :

- perform data cleaning
- find movies with the highest profit
- identify the top 250 movies with the highest IMDb rating
- find the best directors
- popular genres
- audience and critic favorite
- Etc.

Tech-Stack Used

The project was performed using Microsoft Excel. The version used was Microsoft Excel 2019. Excel was used due to its user-friendly interface and its ability to perform data analysis efficiently.

I also used MS PowerPoint to make presentation.

Python: I have used Python to find popular genres.

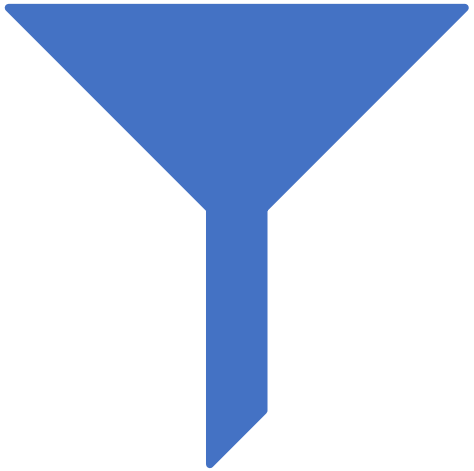
Approach

Step by step approach to solve the given problem and get desired results are as follows:

- Clean the data
- Find the movies with the highest profit
- Find IMDB Top 250
- Find the best directors
- Find popular genres
- Find the critic-favorite and audience-favorite actors

A.Cleaning the data

The first step was to clean the dataset by removing null values and unwanted columns.



B.Movies with highest profit

- Then a new column called profit was created by subtracting the budget from the gross. Movies were sorted based on the profit column.
- Top 25 profitable Movies

	A	B	C	D
1	movie_title	gross	budget	Profit
2	AvatarÂ	760505847	237000000	523505847
3	Jurassic WorldÂ	652177271	150000000	502177271
4	TitanicÂ	658672302	200000000	458672302
5	Star Wars: Episode IV - A New HopeÂ	460935665	11000000	449935665
6	E.T. the Extra-TerrestrialÂ	434949459	10500000	424449459
7	The AvengersÂ	623279547	220000000	403279547
8	The AvengersÂ	623279547	220000000	403279547
9	The Lion KingÂ	422783777	45000000	377783777
10	Star Wars: Episode I - The Phantom MenaceÂ	474544677	115000000	359544677
11	The Dark KnightÂ	533316061	185000000	348316061
12	The Hunger GamesÂ	407999255	78000000	329999255
13	DeadpoolÂ	363024263	58000000	305024263
14	The Hunger Games: Catching FireÂ	424645577	130000000	294645577
15	Jurassic ParkÂ	356784000	63000000	293784000
16	Despicable Me 2Â	368049635	76000000	292049635
17	American SniperÂ	350123553	58800000	291323553
18	Finding NemoÂ	380838870	94000000	286838870
19	Shrek 2Â	436471036	150000000	286471036
20	The Lord of the Rings: The Return of the KingÂ	377019252	94000000	283019252
21	Star Wars: Episode VI - Return of the JediÂ	309125409	32500000	276625409
22	Forrest GumpÂ	329691196	55000000	274691196
23	Star Wars: Episode V - The Empire Strikes BackÂ	290158751	18000000	272158751
24	Home AloneÂ	285761243	18000000	267761243
25	Star Wars: Episode III - Revenge of the SithÂ	380262555	113000000	267262555
26	Spider-ManÂ	403706375	139000000	264706375

C.IMDB Top 250

The next step was to find the top 250 movies with the highest IMDb rating and create a rank column for them.

movie_title	num_voted_users	language	imdb_score	IMDB TOP 250
The Shawshank RedemptionÂ	1689764	English	9.3	The Shawshank RedemptionÂ
The GodfatherÂ	1155770	English	9.2	The GodfatherÂ
The Dark KnightÂ	1676169	English	9	The Dark KnightÂ
The Godfather: Part IIÂ	790926	English	9	The Godfather: Part IIÂ
The Lord of the Rings: The Return of the KingÂ	1215718	English	8.9	The Lord of the Rings: The Return of the KingÂ
Pulp FictionÂ	1324680	English	8.9	Pulp FictionÂ
Schindler's ListÂ	865020	English	8.9	Schindler's ListÂ
The Good, the Bad and the UglyÂ	503509	Italian	8.9	The Good, the Bad and the UglyÂ
Forrest GumpÂ	1251222	English	8.8	Forrest GumpÂ
Star Wars: Episode V - The Empire Strikes BackÂ	837759	English	8.8	Star Wars: Episode V - The Empire Strikes BackÂ
The Lord of the Rings: The Fellowship of the RingÂ	1238746	English	8.8	The Lord of the Rings: The Fellowship of the RingÂ
InceptionÂ	1468200	English	8.8	InceptionÂ
Fight ClubÂ	1347461	English	8.8	Fight ClubÂ
Star Wars: Episode IV - A New HopeÂ	911097	English	8.7	Star Wars: Episode IV - A New HopeÂ
The Lord of the Rings: The Two TowersÂ	1100446	English	8.7	The Lord of the Rings: The Two TowersÂ
The MatrixÂ	1217752	English	8.7	The MatrixÂ
One Flew Over the Cuckoo's NestÂ	680041	English	8.7	One Flew Over the Cuckoo's NestÂ
GoodfellasÂ	728685	English	8.7	GoodfellasÂ
City of GodÂ	533200	Portugues	8.7	City of GodÂ
Seven SamuraiÂ	229012	Japanese	8.7	Seven SamuraiÂ

Movies in the IMDb_Top_250 column which are not in the English language. I have stored them in a new column named Top_Foreign_Lang_Film

language	imdb_score	Top_Foreign_Lang_Film
Italian	8.9	The Good, the Bad and the UglyÂ
Portuguese	8.7	City of GodÂ
Japanese	8.7	Seven SamuraiÂ
Japanese	8.6	Spirited AwayÂ
German	8.5	The Lives of OthersÂ
Persian	8.5	Children of HeavenÂ
Persian	8.4	A SeparationÂ
Korean	8.4	OldboyÂ
German	8.4	Das BootÂ
French	8.4	AmÃ©lieÂ
Japanese	8.4	Princess MononokeÂ
Danish	8.3	The HuntÂ
German	8.3	MetropolisÂ
German	8.3	DownfallÂ
Spanish	8.2	Pan's LabyrinthÂ
Spanish	8.2	The Secret in Their EyesÂ
French	8.2	IncendiesÂ
Japanese	8.2	Howl's Moving CastleÂ

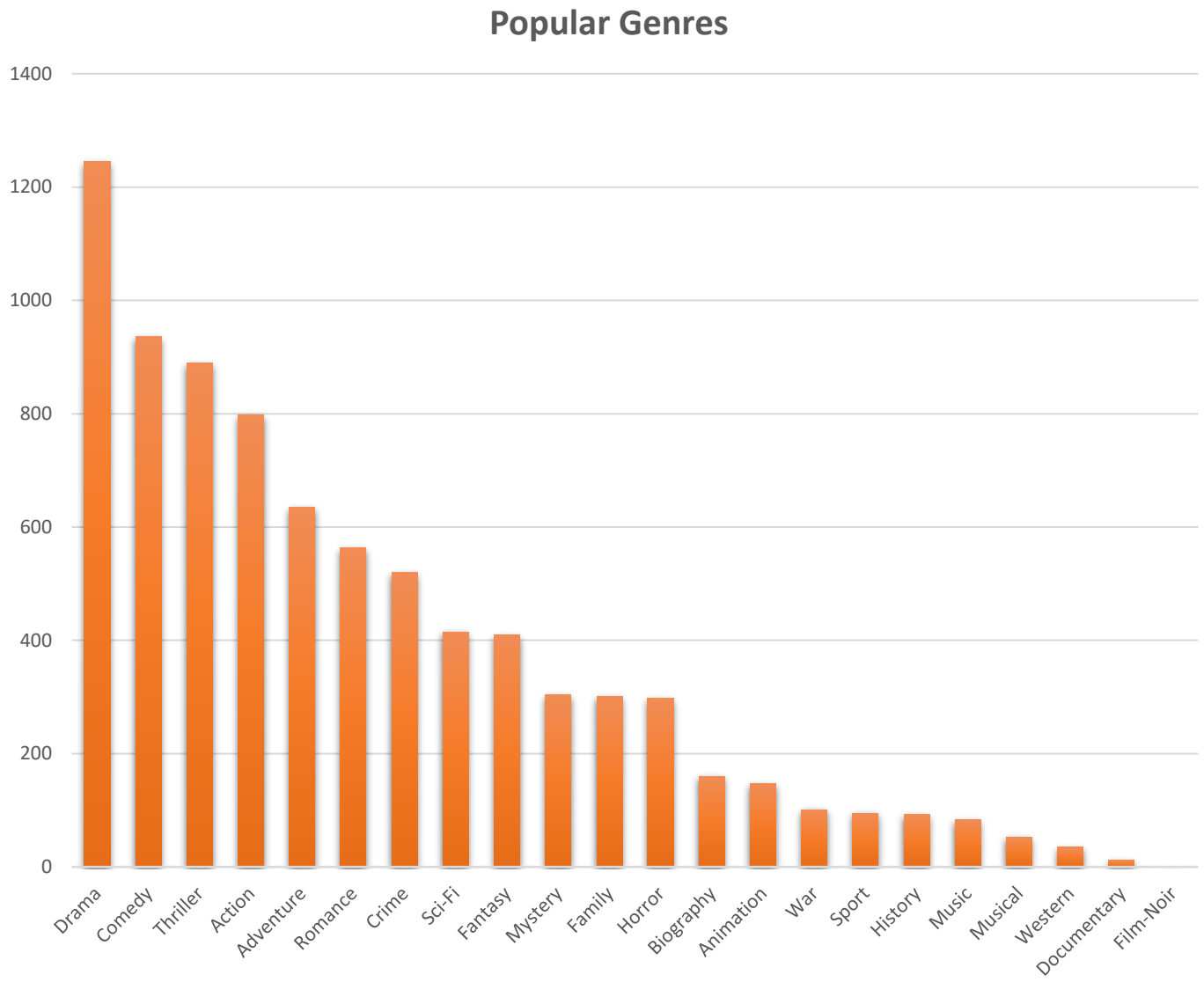
D.Best Directors

In the next step, the dataset was grouped based on the director_name column, and the top 10 directors were identified based on the mean IMDb rating of their movies. If there was a tie between two directors, they were sorted alphabetically.

	A	B	C	D	E
1	director_name	imdb_score			top10director
2	Frank Darabont	9.3			Frank Darabont
3	Francis Ford Coppola	9.2			Francis Ford Coppola
4	Christopher Nolan	9			Christopher Nolan
5	Francis Ford Coppola	9			Francis Ford Coppola
6	Peter Jackson	8.9			Peter Jackson
7	Quentin Tarantino	8.9			Quentin Tarantino
8	Sergio Leone	8.9			Sergio Leone
9	Steven Spielberg	8.9			Steven Spielberg
10	Christopher Nolan	8.8			Christopher Nolan
11	David Fincher	8.8			David Fincher
12	Irvin Kershner	8.8			
13	Peter Jackson	8.8			

E. Popular Genres

The popular genres were identified by grouping the dataset based on the genre column. First, I used “Text To Column” feature to split the genres, then I used count feature. I also took help of Python for this task.



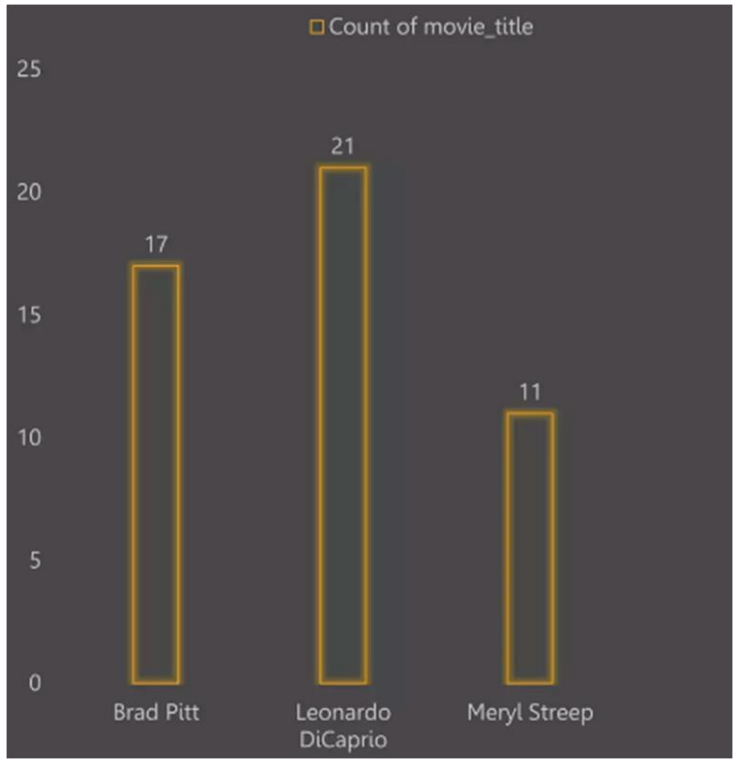
F. Charts

Three new columns were created to identify the movies in which the actors 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' were lead actors. These columns were appended to create a new column called Combined, which was then grouped based on the actor_1_name column. The mean of the num_critic_for_reviews and num_users_for_review was calculated to identify the actors with the highest mean.

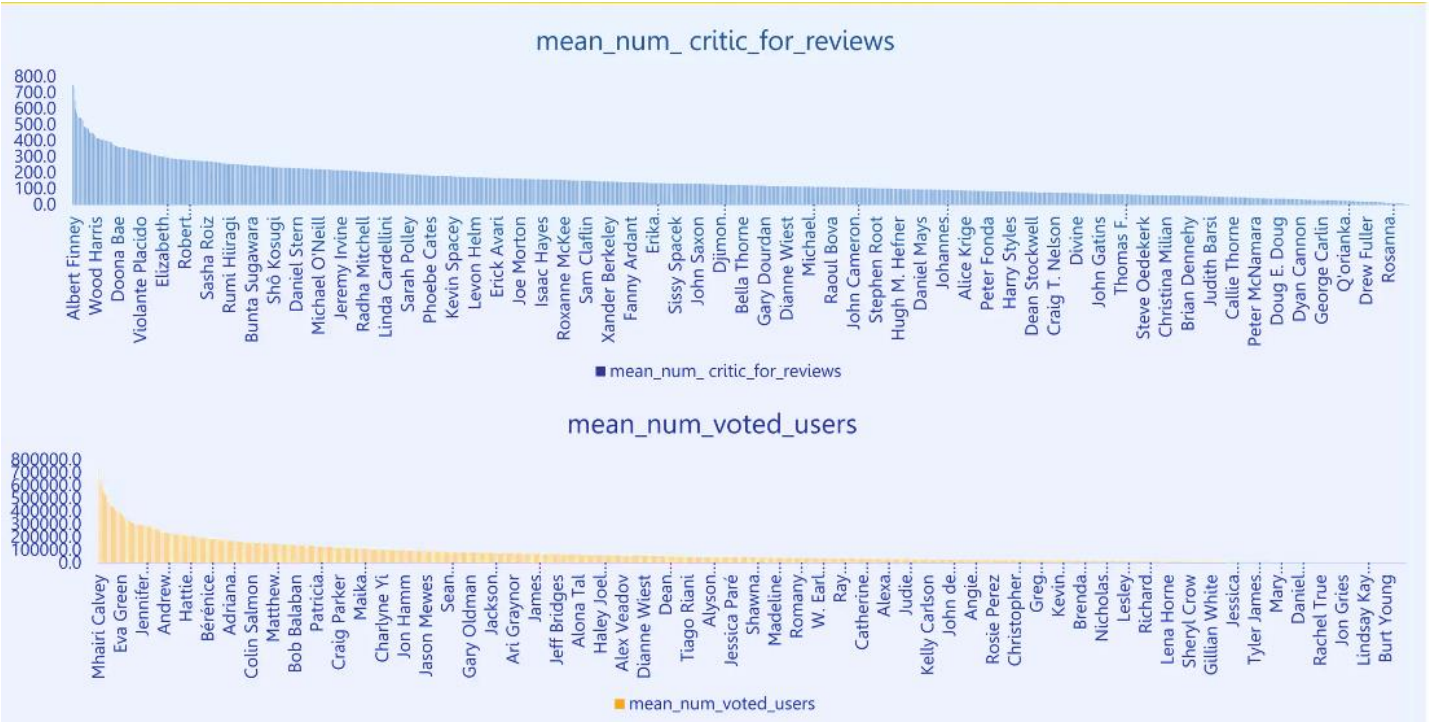
Finally, a new column called decade was created to identify the decade to which each movie belongs to. The dataset was then sorted based on the decade column, and a bar chart was created to analyze the change in the number of voted users over decades.

Count of movies in which 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' are the lead actors

	A	B	C	D	E
1	Meryl_Streep	Leo_Caprio	Brad_Pitt		
2	The River WildÂ	TitanicÂ	True RomanceÂ		
3	The Iron LadyÂ	The Wolf of Wall StreetÂ	TroyÂ		
4	The HoursÂ	The RevenantÂ	The Tree of LifeÂ		
5	The Devil Wears PradaÂ	The Quick and the DeadÂ	The Curious Case of Benjamin ButtonÂ		
6	Out of AfricaÂ	The Man in the Iron MaskÂ	The Assassination of Jesse James by the Coward Robert FordÂ		
7	One True ThingÂ	The Great GatsbyÂ	Spy GameÂ		
8	Lions for LambsÂ	The Great GatsbyÂ	Sinbad: Legend of the Seven SeasÂ		
9	Julie & JuliaÂ	The DepartedÂ	Seven Years in TibetÂ		
10	It's ComplicatedÂ	The BeachÂ	Ocean's TwelveÂ		
11	Hope SpringsÂ	The AviatorÂ	Ocean's ElevenÂ		
12	A Prairie Home CompanionÁ	Shutter IslandÂ	Mr. & Mrs. SmithÂ		
13		Romeo + JulietÂ	Killing Them SoftlyÂ		
14		Revolutionary RoadÂ	Interview with the Vampire: The Vampire ChroniclesÂ		
15		Marvin's RoomÂ	FuryÂ		
16		J. EdgarÂ	Fight ClubÂ		
17		InceptionÂ	By the SeaÂ		
18		Gangs of New YorkÂ	BabelÂ		
19		Django UnchainedÂ			
20		Catch Me If You CanÂ			
21		Body of LiesÂ			
22		Blood DiamondÂ			
23					

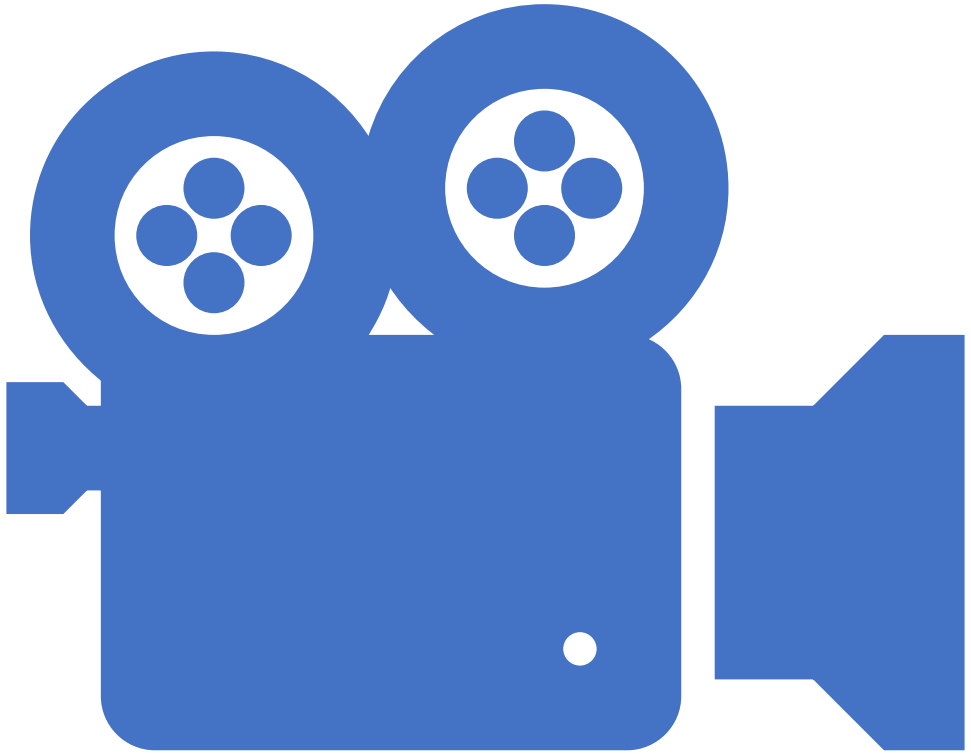


the mean of the num_critic_for_reviews and num_users_for_review and identify the actors which have the highest mean.



Insights

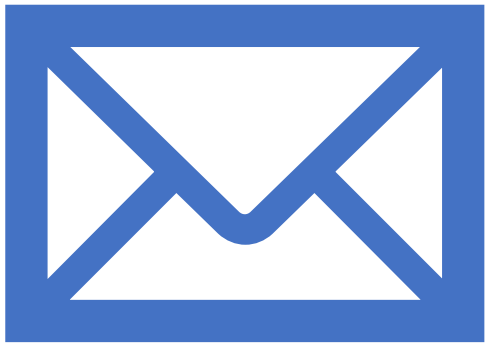
Through this project, we gained insights about the movie industry. We were able to identify the movies with the highest profit, the top 250 movies with the highest IMDb rating, the best directors, popular genres, audience and critic favorite actors, and the change in the number of voted users over decades. We also gained insights into data cleaning, data grouping, and the creation of new columns to extract meaningful information.



Result

The project helped us gain a better understanding of the movie industry and the factors that affect the success of a movie. We were able to identify the top-performing movies, directors, and actors, which can help stakeholders in the industry make informed decisions. The project also helped us gain hands-on experience in data analysis using Excel, which can be applied to other domains as well. Overall, the project was a success, and we were able to achieve our objectives.

Thank You



Prashant Kumar

iprashantkr1@gmail.com