
IMDb
Menu
All
Search IMDb
IMDbPro
Watchlist
Sign In
EN
IMDb
MOVIE ANALYSIS
SUBMITTED BY:
KARTIKEY SHUKLA

PROJECT DESCRIPTION

- The project involves analyzing a dataset of movies from IMDb to derive insights and tell a data story. The initial step is to clean the data by removing null values, dropping unnecessary columns, and preparing the data for analysis.
- The project aims to answer several questions, such as finding the movies with the highest profit, the top 250 movies based on IMDb rating, the best directors based on IMDb score, and popular movie genres. Additionally, the project involves creating new columns for lead actors Meryl Streep, Leonardo DiCaprio, and Brad Pitt and using them to identify the actors with the highest mean num_critic_for_reviews and num_users_for_review.
- Furthermore, the project includes creating a decade column and using it to observe changes in the number of voted users over time. The final output of the project is a report that conveys a data story, insights, and recommendations based on the findings. The project will be handled by using Microsoft Excel and it's various data analytics tools.

APPROACH

I have followed a general approach for this data analysis project.

- The first step is to clearly define the problem and the goal of the analysis. Then, the process of data cleaning is performed to ensure data accuracy and completeness.
- Next, exploratory data analysis is conducted to understand the patterns and relationships in the data.
- After that, specific techniques and models are applied to obtain insights and draw conclusions.
- Finally, results are presented and communicated to stakeholders in an understandable and actionable way.

Throughout the process, it is important to iterate and refine the approach as new findings and insights are obtained.

TECH STACK USED

The project was executed using Microsoft Excel with the version of Excel 2021.
Microsoft Excel was used for data manipulation, cleaning, and analysis, which includes the use of built-in functions and features like filtering, sorting, pivot tables, and charts for data visualization.
It is a commonly used tool for data analysis in businesses and industries and is known for its user-friendly interface and flexibility in handling large data sets.
For the last part of the project, Python libraries were also used to answer the relevant questions.

PROBLEM STATEMENT

Based on the IMDB dataset, I see a potential problem of movie studios investing large amounts of money in movies that do not perform well at the box office. This could lead to financial losses and a decrease in audience interest in future movies from those studios.

My hypothesis for the cause of this problem is that studios may not be fully understanding the preferences and interests of their target audience, leading them to produce movies that do not align with what viewers want to see.

The impact of this problem on stakeholders, such as the studios, movie theaters, and audiences, could be significant. Studios may experience financial losses and a decrease in reputation, while movie theaters may see a decrease in ticket sales if audiences lose interest in movies.

Additionally, audiences may be disappointed with the quality of movies being produced and may become less interested in going to the movies in general.
If this problem is not solved, it could lead to a decline in the movie industry as a whole and a shift towards other forms of entertainment. This could have long-term economic and cultural impacts on society.

QUESTIONS

- **What do you see happening?**

Movie studios are investing large amounts of money in movies that do not perform well at the box office.

- **What are the specific symptoms of the problem?**

Financial losses, decrease in audience interest in future movies from those studios.

- **What is your hypothesis for the cause of the problem?**

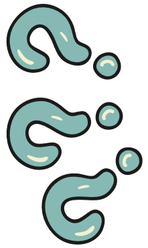
Studios may not be fully understanding the preferences and interests of their target audience, leading them to produce movies that do not align with what viewers want to see.

- **Why are studios not understanding their target audience?**

They may not be conducting enough market research or not effectively analyzing the results.

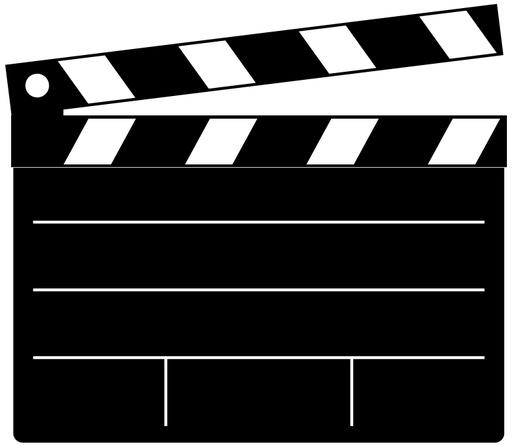
- **Why are studios not conducting enough market research or effectively analyzing the results?**

Studios may be relying too much on established formulas or trends and not taking enough risks or experimenting with new ideas. They may also not have enough diversity in their decision-making teams or be too focused on profit rather than audience satisfaction.



FIVE 'WHYS' APPROACH

1. **Why did the movie have a high budget?**
   - The movie had a high budget because it required expensive special effects and a large cast and crew.
2. **Why did the movie have a low box office gross?**
   - The movie had a low box office gross because it did not appeal to a wide enough audience or faced tough competition from other releases.
3. **Why did the movie have a high profit despite a low budget?**
   - The movie had a high profit despite a low budget because it was well-received by audiences and critics, resulting in high ticket sales and positive word-of-mouth promotion.
4. **Why did the movie have a low profit despite a high budget?**
   - The movie had a low profit despite a high budget because it failed to attract enough audiences or was negatively received by critics, resulting in low ticket sales and poor word-of-mouth promotion.
5. **Why did the movie have a high budget despite a low profit?**
   - The movie had a high budget despite a low profit because the filmmakers invested heavily in production value, hoping that the quality of the movie would lead to success, but this strategy did not pay off in the end.



BUSINESS

QUESTIONS

Cleaning the data: This is one of the most important steps to perform before moving forward with the analysis. Use your knowledge learned till now to do this. (Dropping columns, removing null values, etc.)

Your task: Clean the data.

color	director_name	num_critic_for_review	duration	director_facebook_like	actor_3_facebook_like	actor_2_name	actor_1_facebook_like	gross	genres
Color	Frank Darabont	199	142	0	461	Jeffrey DeMunn	11000	28341469	Crime|Drama
Color	Francis Ford Coppola	208	175	0	3000	Marlon Brando	14000	134821952	Crime|Drama
Color	Christopher Nolan	645	152	22000	11000	Heath Ledger	23000	533316061	Action|Crime|Drama|Thriller
Color	Francis Ford Coppola	149	220	0	3000	Al Pacino	22000	57300000	Crime|Drama
Color	Peter Jackson	328	192	0	416	Billy Boyd	5000	377019252	Action|Adventure|Drama|Fantasy
Color	Quentin Tarantino	215	178	16000	857	Eric Stoltz	13000	107930000	Crime|Drama
Black and White	Steven Spielberg	174	185	14000	212	Embeth Davidtz	14000	96067179	Biography|Drama|History
Color	Sergio Leone	181	142	0	24	Luigi Pistilli	16000	6100000	Western
Black and White	Robert Zemeckis	149	142	0	194	Siobhan Fallon Hogan	15000	329691196	Comedy|Drama
Color	Irvin Kershner	223	127	883	441	Kenny Baker	11000	290158751	Action|Adventure|Fantasy|Sci-Fi
Color	Peter Jackson	297	171	0	857	Orlando Bloom	16000	313837577	Action|Adventure|Drama|Fantasy
Color	Christopher Nolan	642	148	22000	23000	Tom Hardy	29000	292568851	Action|Adventure|Sci-Fi|Thriller
Color	David Fincher	315	151	21000	637	Meat Loaf	11000	37023395	Drama
Color	George Lucas	282	125	0	504	Peter Cushing	11000	460935665	Action|Adventure|Fantasy|Sci-Fi
Color	Peter Jackson	294	172	0	857	Orlando Bloom	16000	340478898	Action|Adventure|Drama|Fantasy
Color	Lana Wachowski	313	136	0	99	Marcus Chong	18000	171383253	Action|Sci-Fi
Color	Milos Forman	149	133	869	425	Michael Berryman	888	112000000	Drama
Color	Martin Scorsese	192	146	17000	635	Mike Starr	22000	46836394	Biography|Crime|Drama
Color	Fernando Meirelles	214	135	353	40	Seu Jorge	1000	7563397	Crime|Drama
Black and White	Akira Kurosawa	153	202	0	4	Minoru Chiaki	304	269061	Action|Adventure|Drama
Color	Steven Spielberg	219	169	14000	13000	Vin Diesel	15000	216119491	Action|Drama|War
Color	Jonathan Demme	185	138	438	173	Scott Glenn	12000	130727000	Crime|Drama|Horror|Thriller
Color	David Fincher	216	127	21000	360	Brad Pitt	11000	100125340	Crime|Drama|Mystery|Thriller
Color	Christopher Nolan	712	169	22000	6000	Anne Hathaway	11000	187991439	Adventure|Drama|Sci-Fi
Color	Bryan Singer	162	106	0	574	Chazz Palminteri	18000	23272306	Crime|Drama|Mystery|Thriller
Black and White	Tony Kaye	162	101	194	602	Beverly D'Angelo	1000	6712241	Crime|Drama
Black and White	Charles Chaplin	120	87	0	8	Stanley Blystone	309	163245	Comedy|Drama|Family
Color	Hayao Miyazaki	246	125	6000	7	RyÃ»nosuke Kamiki	17	10049886	Adventure|Animation|Family|Fantasy
Color	Roger Allers	186	73	28	847	Nathan Lane	2000	422783777	Adventure|Animation|Drama|Family|Musical
Color	Steven Spielberg	234	115	14000	488	Karen Allen	11000	242374454	Action|Adventure
Color	Christopher Nolan	813	164	22000	23000	Christian Bale	27000	448130642	Action|Thriller
Color	Robert Zemeckis	198	116	0	459	Thomas F. Wilson	1000	210609762	Adventure|Comedy|Sci-Fi
Color	James Cameron	210	153	0	539	Jenette Goldstein	780	204843350	Action|Sci-Fi

IMDB_Movies | Scatter Plot | Top Directors | Top Genres

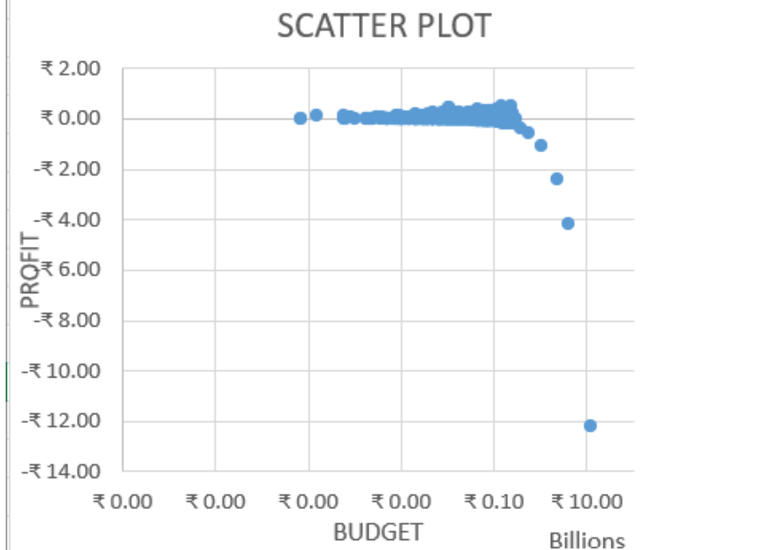
Previously there were more than 5000 rows in the dataset.

But after cleaning the dataset, only 3757 rows were remaining for further analysis.

Movies with highest profit: Create a new column called profit which contains the difference of the two columns: gross and budget. Sort the column using the profit column as reference. Plot profit (y-axis) vs budget (x- axis) and observe the outliers using the appropriate chart type.

Your task: Find the movies with the highest profit?

MOVIE TITLES	Sum of PROFIT
The AvengersÂ	806559094
AvatarÂ	523505847
Jurassic WorldÂ	502177271
TitanicÂ	458672302
Star Wars: Episode IV - A New HopeÂ	449935665
E.T. the Extra-TerrestrialÂ	424449459
The Lion KingÂ	377783777
The Jungle BookÂ	375290282
Star Wars: Episode I - The Phantom MenaceÂ	359544677
The Dark KnightÂ	348316061



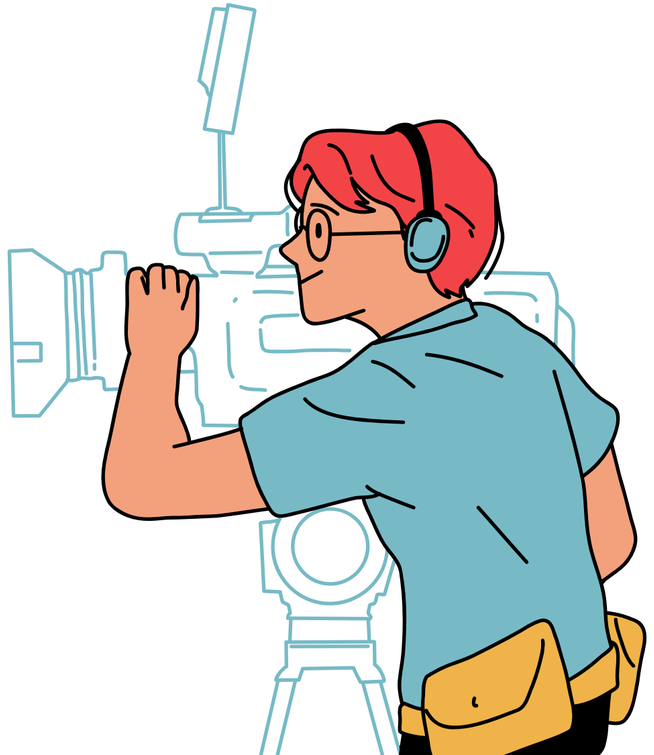
Top 250: Create a new column IMDb_Top_250 and store the top 250 movies with the highest IMDb Rating (corresponding to the column: imdb_score). Also make sure that for all of these movies, the num_voted_users is greater than 25,000. Also add a Rank column containing the values 1 to 250 indicating the ranks of the corresponding films.
Extract all the movies in the IMDb_Top_250 column which are not in the English language and store them in a new column named Top_Foreign_Lang_Film. You can use your own imagination also!
Your task: Find IMDB Top 250

TOP FOREIGN LANGUAGE FILMS
A Fistful of DollarsÂ
A SeparationÂ
AkiraÂ
AmÃ©lieÂ
Amores PerrosÂ
Central StationÂ
Children of HeavenÂ
City of GodÂ
Das BootÂ
DownfallÂ
Elite SquadÂ
Howl's Moving CastleÂ
IncendiesÂ
MetropolisÂ
My Name Is KhanÂ
OldboyÂ
Pan's LabyrinthÂ
PersepolisÂ
Princess MononokeÂ
Seven SamuraiÂ
Spirited AwayÂ
Tae Guk Gi: The Brotherhood of WarÂ
The CelebrationÂ
The Good, the Bad and the UglyÂ
The HuntÂ
The Lives of OthersÂ
The Sea InsideÂ
The Secret in Their EyesÂ
Waltz with BashirÂ

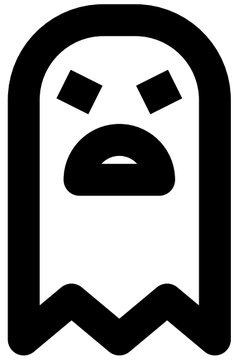
TOP 250 FILMS
12 Years a SlaveÂ
2001: A Space OdysseyÂ
A Beautiful MindÂ
A Fistful of DollarsÂ
A SeparationÂ
AkiraÂ
AladdinÂ
AlienÂ
AliensÂ
AmÃ©lieÂ
AmadeusÂ
American BeautyÂ
American History XÂ
Amores PerrosÂ
Annie HallÂ
Apocalypse NowÂ
Back to the FutureÂ
Batman BeginsÂ
Before SunriseÂ
Before SunsetÂ
Big FishÂ
Black SwanÂ
Blade RunnerÂ
Blood DiamondÂ
Bowling for ColumbineÂ
BoyhoodÂ
BraveheartÂ
BrazilÂ
Butch Cassidy and the Sundance KidÂ
Captain America: Civil WarÂ
Casino RoyaleÂ
CasinoÂ
Catch Me If You CanÂ
Central StationÂ
Children of HeavenÂ
Cinderella ManÂ
City of GodÂ
Dallas Buyers ClubÂ
Dancer in the DarkÂ
Dances with WolvesÂ
Das BootÂ
Dead Poets SocietyÂ
DeadpoolÂ
Die HardÂ
District 9Â
Django UnchainedÂ
Doctor ZhivagoÂ
Donnie DarkoÂ
DownfallÂ
Elite SquadÂ
Eternal Sunshine of the Spotless MindÂ
Fiddler on the RoofÂ
Fight ClubÂ
Finding NemoÂ
Forrest GumpÂ
GladiatorÂ
Gone GirlÂ
Gone with the WindÂ
Good Will HuntingÂ
GoodfellasÂ
Gran TorinoÂ
Groundhog DayÂ
Guardians of the GalaxyÂ
HerÂ
Hotel RwandaÂ
How to Train Your DragonÂ
Howl's Moving CastleÂ
In BrugesÂ
IncendiesÂ
InceptionÂ
Indiana Jones and the Last CrusadeÂ
Inglourious BasterdsÂ
Inside OutÂ
InterstellarÂ
Into the WildÂ
JawsÂ
JFKÂ
Jurassic ParkÂ
Kill Bill: Vol. 1Â
Kill Bill: Vol. 2Â
L.A. ConfidentialÂ
Lawrence of ArabiaÂ
Life of PiÂ
Lock, Stock and Two Smoking BarrelsÂ
Mad Max: Fury RoadÂ
MagnoliaÂ
MementoÂ
MetropolisÂ
Million Dollar BabyÂ
Modern TimesÂ
Monsters, Inc.Â
Monty Python and the Holy GrailÂ
Mulholland DriveÂ
My Name Is KhanÂ
Mystic RiverÂ
No Country for Old MenÂ
OldboyÂ
On the WaterfrontÂ
Once Upon a Time in AmericaÂ
One Flew Over the Cuckoo's NestÂ
Pan's LabyrinthÂ
PersepolisÂ
Pirates of the Caribbean: The Curse of the Black PearlÂ
PlatoonÂ
Princess MononokeÂ
PrisonersÂ
PsychoÂ
Pulp FictionÂ
Raging BullÂ
Raiders of the Lost ArkÂ
Rain ManÂ
RatatouilleÂ
Requiem for a DreamÂ
Reservoir DogsÂ
RockyÂ
RoomÂ
RushÂ
Saving Private RyanÂ
ScarfaceÂ
Schindler's ListÂ
Se7enÂ
SerenityÂ
Seven SamuraiÂ
Shaun of the DeadÂ
Shutter IslandÂ
SickoÂ
Sin CityÂ
Sling BladeÂ
Slumdog MillionaireÂ
SnatchÂ
Some Like It HotÂ
Spirited AwayÂ
SpotlightÂ
Stand by MeÂ
Star TrekÂ
Star Wars: Episode IV - A New HopeÂ
Star Wars: Episode V - The Empire Strikes BackÂ
Star Wars: Episode VI - Return of the JediÂ
Tae Guk Gi: The Brotherhood of WarÂ
Terminator 2: Judgment DayÂ
The ArtistÂ
The AvengersÂ
The Best Years of Our LivesÂ
The Big LebowskiÂ
The Bourne UltimatumÂ
The Bridge on the River KwaiÂ
The CelebrationÂ
The Dark Knight RisesÂ
The Dark KnightÂ
The DepartedÂ
The ExorcistÂ
The Godfather: Part IIÂ
The GodfatherÂ
The Good, the Bad and the UglyÂ
The Grand Budapest HotelÂ
The Green MileÂ
The HelpÂ
The HuntÂ
The Imitation GameÂ
The IncrediblesÂ
The Iron GiantÂ
The King's SpeechÂ
The Lion KingÂ
The Lives of OthersÂ
The Lord of the Rings: The Fellowship of the RingÂ
The Lord of the Rings: The Return of the KingÂ
The Lord of the Rings: The Two TowersÂ
The MartianÂ
The MatrixÂ
The Perks of Being a WallflowerÂ
The PianistÂ
The PrestigeÂ
The Princess BrideÂ
The Pursuit of HappynessÂ
The RevenantÂ
The Sea InsideÂ
The Secret in Their EyesÂ
The Shawshank RedemptionÂ
The Silence of the LambsÂ
The Sixth SenseÂ
The Sound of MusicÂ
The StingÂ
The Straight StoryÂ
The TerminatorÂ
The ThingÂ
The Truman ShowÂ
The Usual SuspectsÂ
The Wizard of OzÂ
The Wolf of Wall StreetÂ
There Will Be BloodÂ
Toy Story 3Â
Toy StoryÂ
TrainspottingÂ
True RomanceÂ
UnforgivenÂ
UpÂ
V for VendettaÂ
WALLÂ·EÂ
Waltz with BashirÂ
WarriorÂ
WhiplashÂ
X-Men: Days of Future PastÂ
Young FrankensteinÂ

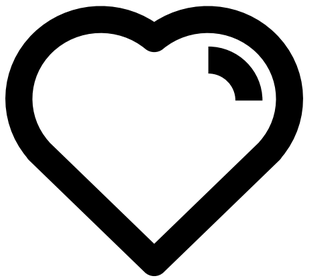
Best Directors: Group the column using the director_name column.
Find out the top 10 directors for whom the mean of imdb_score is the highest and store them in a new column top10director. In case of a tie in IMDb score between two directors, sort them alphabetically.
Your task: Find the best directors

Director Name	Average of imdb_score
Asghar Farhadi	8.4
Richard Marquand	8.4
Christopher Nolan	8.425
Sergio Leone	8.433333333
Damien Chazelle	8.5
Ron Fricke	8.5
Alfred Hitchcock	8.5
Majid Majidi	8.5
Charles Chaplin	8.6
Tony Kaye	8.6
Akira Kurosawa	8.7

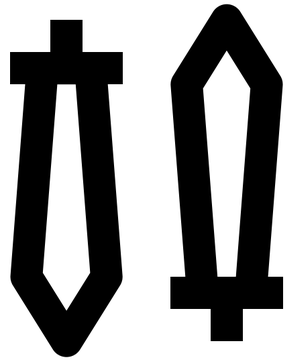


Popular Genres: Perform this step using the knowledge gained while performing previous steps.
Your task: Find popular genres.





Genres	Sum of num_voted_users
Action|Adventure|Sci-Fi	15729711
Drama	12349698
Comedy	11202428
Comedy|Drama|Romance	10456879
Crime|Drama	10201806
Comedy|Romance	9940146
Drama|Romance	9791571
Crime|Drama|Thriller	9580903
Action|Adventure|Thriller	7884237
Comedy|Drama	7717887



Charts: Create three new columns namely, Meryl_Streep, Leo_Caprio, and Brad_Pitt which contain the movies in which the actors: 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' are the lead actors. Use only the actor_1_name column for extraction. Also, make sure that you use the names 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' for the said extraction.
Append the rows of all these columns and store them in a new column named Combined.
Group the combined column using the actor_1_name column.
Find the mean of the num_critic_for_reviews and num_users_for_review and identify the actors which have the highest mean.
Observe the change in number of voted users over decades using a bar chart. Create a column called decade which represents the decade to which every movie belongs to. For example, the title_year year 1923, 1925 should be stored as 1920s. Sort the column based on the column decade, group it by decade and find the sum of users voted in each decade. Store this in a new data frame called df_by_decade.
Your task: Find the critic-favorite and audience-favorite actors.

1

```
# Create three new columns
df['Meryl_Streep'] = df.loc[df['actor_2_name'] == 'Meryl Streep', 'movie_title']
df['Leo_Caprio'] = df.loc[df['actor_2_name'] == 'Leonardo DiCaprio', 'movie_title']
df['Brad_Pitt'] = df.loc[df['actor_2_name'] == 'Brad Pitt', 'movie_title']

# Display the first 5 rows to check if the new columns have been created successfully
df.head()
```

	genres	...	aspect_ratio	movie_facebook_likes	PROFIT	rank	TOP 250	TOP FOREIGN LANG FILM	Meryl_Streep	Leo_Caprio	Brad_Pitt	decade
	Crime|Drama	...	1.85	108000	? 33,41,469.00	1	The Shawshank Redemption	NaN	NaN	NaN	NaN	1990
	Crime|Drama	...	1.85	43000	? 12,88,21,952.00	2	The Godfather	NaN	NaN	NaN	NaN	1970
	Action|Crime|Drama|Thriller	...	2.35	37000	? 34,83,16,061.00	3	The Dark Knight	NaN	NaN	NaN	NaN	2000
	Crime|Drama	...	1.85	14000	? 4,43,00,000.00	3	The Godfather: Part II	NaN	NaN	NaN	NaN	1970
	Action|Adventure|Drama|Fantasy	...	2.35	16000	? 28,30,19,252.00	5	The Lord of the Rings: The Return of the King	NaN	NaN	NaN	NaN	2000

2

```
# Define a function to extract the decade from a year
def extract_decade(year):
    decade = (year // 10) * 10
    return f"{decade}s"

# Apply the function to each row in the DataFrame and create a new column called decade
df['decade'] = df['title_year'].apply(extract_decade)

# Display the first 5 rows to check if the new column has been created successfully
df.head()
```

	color	director_name	num_critic_for_reviews	duration	director_facebook_likes	actor_3_facebook_likes	actor_2_name	actor_1_facebook_likes	gross
0	Color	Frank Darabont	199	142	0	461	Jeffrey DeMunn	11000	28341469
1	Color	Francis Ford Coppola	208	175	0	3000	Marlon Brando	14000	134821952
2	Color	Christopher Nolan	645	152	22000	11000	Heath Ledger	23000	533316061
3	Color	Francis Ford Coppola	149	220	0	3000	Al Pacino	22000	57300000
4	Color	Peter Jackson	328	192	0	416	Billy Boyd	5000	377019252

5 rows × 36 columns

3

```
# Group the Combined column by actor name and calculate the mean of num_critic_for_reviews and num_user_for_reviews
df_by_actor = df.groupby('actor_1_name')[['num_critic_for_reviews', 'num_user_for_reviews']].mean()

# Display the first 5 rows to check if the new DataFrame has been created successfully
df_by_actor.head()
```

actor_1_name	num_critic_for_reviews	num_user_for_reviews
50 Cent	98.000000	284.000000
Aaliyah	137.000000	695.000000
Aasif Mandvi	210.000000	147.000000
Abbie Cornish	270.333333	184.666667
Adam Arkin	82.000000	130.000000

4

```
# create a new column 'decade' to represent the decade to which each movie belongs
df['decade'] = (df['title_year'] // 10) * 10

# group the movies by decade and find the sum of num_voted_users for each decade
df_by_decade = df.groupby('decade')['num_voted_users'].sum().reset_index()

# print the resulting DataFrame
print(df_by_decade)
```

```
   decade  num_voted_users
0    1920           116387
1    1930           804839
2    1940           159517
3    1950           678336
4    1960          2982551
5    1970          8681156
6    1980         20091781
7    1990         69934957
8    2000        172800132
9    2010        121235552
```

INSIGHTS

- On average, movies have a duration of 110 minutes with a budget of $37 million and a gross of $48 million.
- The top 3 genres in terms of frequency are Drama, Comedy, and Action.
- The most common content rating is PG-13 followed by R.
- The average IMDB score is 6.44 out of 10.
- There is a positive correlation between budget and gross, indicating that higher budget movies tend to perform better at the box office.
- There is also a positive correlation between the number of Facebook likes of actors/actresses and the gross revenue of the movie, indicating that movies with more popular actors tend to perform better at the box office.

RESULT

In this project, I have analyzed a dataset of movies and extracted various insights related to movies, such as the highest profit-making movies, top 250 movies, popular genres, best directors, audience-favorite actors, and more.

By completing this project, I have gained hands-on experience in data cleaning, data analysis, and data visualization using Microsoft Excel.

This project has helped me to improve my skills in data handling, visualization, and drawing insights from data.

It has also provided me with a better understanding of the movie industry and the factors that contribute to the success of a movie.


THE
END