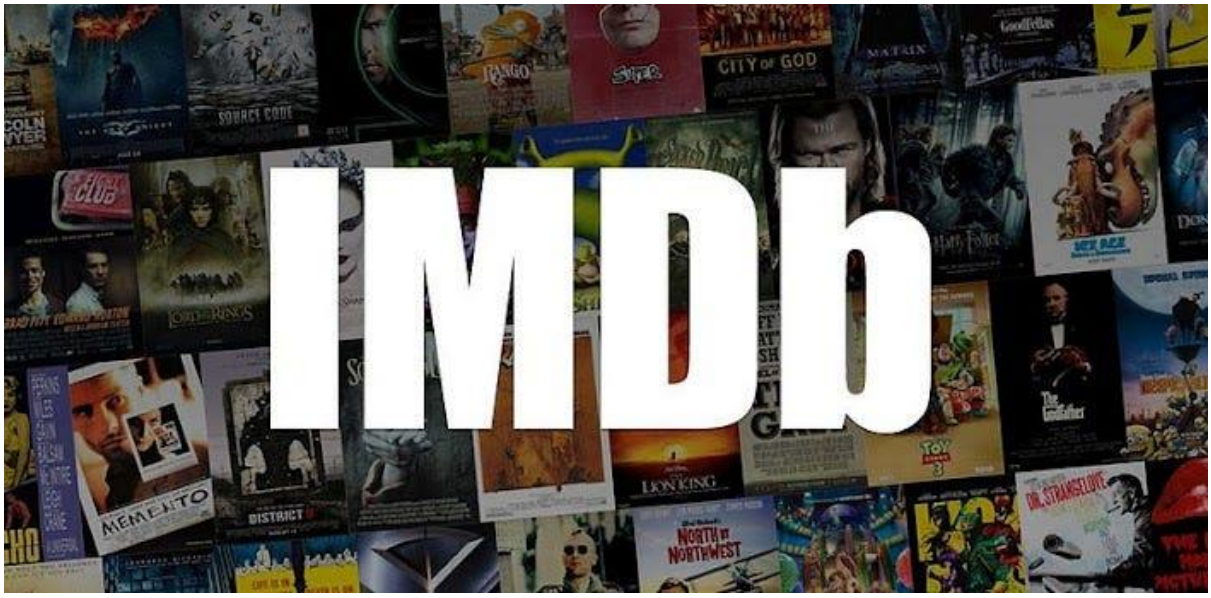
IMDB Movie Analysis



- **Samruddhi Pawar**

Project Description:

A potential question for filmmakers to investigate could be: “What factors influence the success of a movie on IMDB?” Here, success can be defined by high IMDB ratings. The impact of this question is significant for movie producers, directors, and investors who want to understand what makes a movie successful to make informed decisions in their future projects.
This Project is about giving insights into the success of a Movie based on IMDB data provided which will be helpful for filmmakers and other stakeholders during the production of a movie.

The goal of this project is to analyse a dataset of IMDB movies and draw insights from the data. The dataset includes various columns such as movie names, budgets, gross revenue, and IMDB ratings. The tasks include identifying the movie with the highest profit or the top IMDB movies, as well as sharing insights by identifying any problems or trends in the data. The overall objective of the project is to gain a better understanding of the movie industry by analysing the data and drawing meaningful conclusions.

In this project, I will be analysing a dataset containing information about movies. The main objective is to explore and derive insights from the data using various data analysis techniques. This includes cleaning the data by removing null values and dropping unnecessary columns.

In this project, I was required to provide a detailed report for the below data record mentioning the answers to the questions that follow:

- Movie Genre Analysis: Analyse the distribution of movie genres and their impact on the IMDB score.
- Movie Duration Analysis: Analyse the distribution of movie durations and its impact on the IMDB score.
- Language Analysis: Situation: Examine the distribution of movies based on their language.
- Director Analysis: Influence of directors on movie ratings.
- Budget Analysis: Explore the relationship between movie budgets and their financial success.

Approach:

1. *Recognize the data:* I spent some time becoming acquainted with the data before starting the analysis. Take a look at the data's structure to gain an idea of its general substance. This aids in the identification of any possible problems or difficulties that I might encounter while carrying out my analysis.
2. *Inspect for incomplete or missing data:* Verify that your dataset is free of any blank spaces or missing data.

3. *Recognize and deal with outliers:* Data points that differ noticeably from the rest of the data are known as outliers. They have the potential to affect summary statistics and skew your analysis's findings significantly. It's critical to recognize any outliers and determine the best course of action, such as eliminating them from analysis.
4. *Communicate your findings:* After doing your analysis, give a succinct and clear presentation of your results to your audience. Employ visual aids like graphs and charts to help explain your findings. Make sure you describe your technique and the significance of your findings in detail.

Root Cause Analysis: 5 Why's approach

Once we have the problem better defined, we can use the 5 Whys technique to determine its root cause by repeatedly asking the question "Why". It's also called the Root Cause Analysis, developed by Sakichi Toyoda, founder of Toyota Industries. Here's an example of how this technique could be used to figure out the cause of the following problem: A business went over budget on a recent project.

Q: Why last few movies flop?
A: Because they failed to impress both critics and the audience.

Q: Why they failed to impress?
A: Because they gave the audience something that they couldn't be satisfied with.

Q: Why they are not able to satisfy the audience?
A: Because of the miscasting of actors and directors as well as the poor selection of genre and storyline.

Q: Why did they miss-casted and select a poor story?
A: Because they didn't use data-driven decision-making to track audience preference and trends in the film industry.

Q: Why didn't they use data-driven decision-making?
A: Because they didn't have access to these decision-making models.

Impact:
Meaningful insights from data analysis will help in:

- Tracking audience preferences and identifying the target audience
- Trends in the film industry, trending genres, and story
- Determining budget and minimizing the risk of losses

Tech-Stack Used:

1. *Python* - The programming language used for Data Pre-processing
2. *Google Colab* - Interactive platform to write and execute codes in various programming languages (in this case Python).
3. *Microsoft Excel* - A spreadsheet editor software used mainly by professionals to enter data in table format, perform computations, plot graphs, etc. Here Microsoft Excel is used to pre-process the data.

Dataset Overview:

The dataset provided is related to IMDB Movies and contains records of movies from several years and geographical locations.

- *The Dataset details are:*
  - Number of Data Points: 5,043
  - Number of Features: 28
  - Column Details:
  1. **color**: The movie is Coloured or Black and White
  2. **director_name**: Name of the movie's director
  3. **num_critic_for_reviews**: Number of reviews by film critics'
  4. **duration**: Duration of the movie
  5. **director_facebook_likes**: Facebook Likes of the director
  6. **actor_3_facebook_likes**: Facebook Likes of one of the actors
  7. **actor_2_name**: Name of one of the actors
  8. **actor_1_facebook_likes**: Facebook Likes of one of the actors
  9. **gross**: A gross collection of the movie
  10. **genres**: Genres of the movie
  11. **actor_1_name**: Name of one of the actors
  12. **movie_title**: Name of the movie
  13. **num_voted_users**: Number of users voted for the movie
  14. **cast_total_facebook_likes**: Movie cast's total Facebook likes
  15. **actor_3_name**: Name of one of the actors
  16. **facenumber_in_poster**: Number of faces in the movie's poster
  17. **plot_keywords**: Some keywords from the plot of the movie
  18. **movie_imdb_link**: IMDB link of the movie
  19. **num_user_for_reviews**: Number of users who reviewed the movie
  20. **language**: The original language of the movie
  21. **country**: Country of origin of the movie
  22. **content_rating**: Content rating of the movie (Certification tag)
  23. **budget**: The budget for the movie
  24. **title_year**: Year in which the movie was released
  25. **actor_2_facebook_likes**: Facebook Likes of one of the actors

26. **imdb_score**: IMDB Score of the movie
27. **aspect_ratio**: Aspect ratio in which the movie was made
28. **movie_facebook_likes**: Facebook likes of the movie

Source of Data:
https://docs.google.com/spreadsheets/d/1mrNfd4wnbEWiOtgfYvRziHvs6kI_QOnnLCAH000uszE/edit?usp=sharing

Data Pre-Processing

Handling Duplicate Values

- I found some rows where all column values were duplicated. Keeping the first occurrence of each duplicate, dropped the rest of the duplicates.
- On checking rows with duplicate values of movie_title, we observed that except for movies “Out of the Blue” and “The Host”, rest for almost all the movies, the difference between column values for rows with the same movie_title is in columns related to the facebook likes and “num_voted_users”. So except for the two movies mentioned above, we can drop the rest of the duplicate rows leaving just one copy of the row without much effect on the overall analysis.

Handling Missing Values:

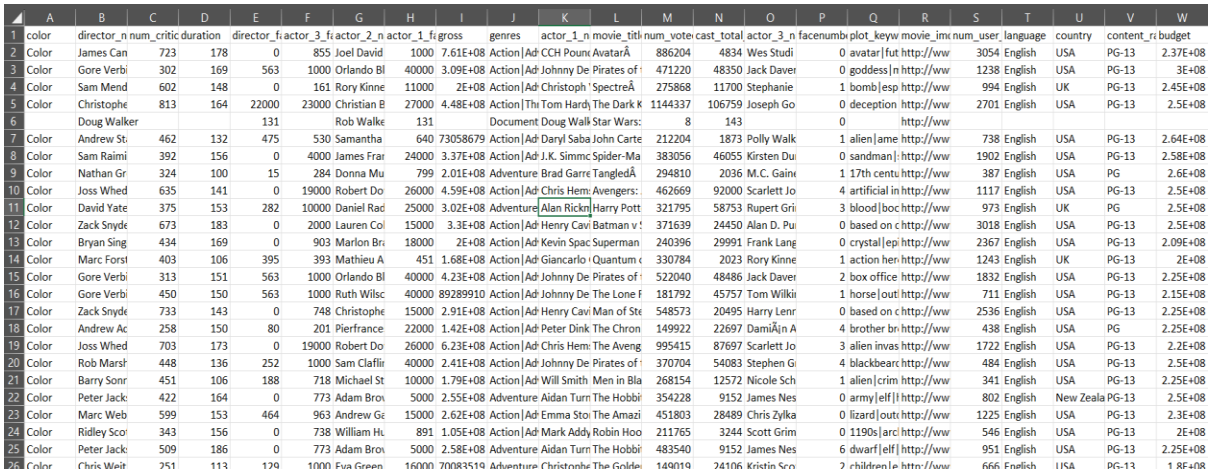
- Checked frequency of row-wise null values and dropped all the rows where the number of null values was greater than 30%
- For null values in the gross column, I calculated the median of the gross value which had a similar IMDB Score of respective null gross value.
- For null values in a budget column, I calculated the median of the budget value which had a similar IMDB Score of respective null budget value.
- The null values of the language column of such rows are replaced with ‘English’ after confirming them on the internet.
- For null values in the duration column, we scrapped the movie’s Wikipedia page to get the duration of the movie and store it in the original data frame.

Handling Outliers:

- For the title_year column, values less than 1916 seem to be outliers which we replaced with correct values of the release year of the movies after a manual check on the internet.
- Replaced values of budget less than 0 with median value.
- Replaced values of gross less than 0 with median value.

Feature Engineering:

- Created a new column profit that has the difference between gross and budget column values.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	color	director_n	num_critic	duration	director_fa	actor_3_fa	actor_2_n	actor_1_fa	gross	genres	actor_1_n	movie_titl	num_vote	cast_total	actor_3_n	facenumb	plot_keyw	movie_ima	num_user	language	country	content_ra	budget
2	Color	James Can	723	178	0	855	Joel David	1000	7.61E+08	Action|Ad	CCH Poun	AvatarÂ	886204	4834	Wes Studi	0	avatar|fut	http://ww	3054	English	USA	PG-13	2.37E+08
3	Color	Gore Verbi	302	169	563	1000	Orlando Bl	40000	3.09E+08	Action|Ad	Johnny De	Pirates of	471220	48350	Jack Daver	0	goddess|n	http://ww	1238	English	USA	PG-13	3E+08
4	Color	Sam Mend	602	148	0	161	Rory Kinne	11000	2E+08	Action|Ad	Christoph	SpectreÂ	275868	11700	Stephanie	1	bomb|esp	http://ww	994	English	UK	PG-13	2.45E+08
5	Color	Christophe	813	164	22000	23000	Christian B	27000	4.48E+08	Action|Th	Tom Hardy	The Dark K	1144337	106759	Joseph Go	0	deception	http://ww	2701	English	USA	PG-13	2.5E+08
6		Doug Walker			131		Rob Walke	131		Document	Doug Walk	Star Wars:	8	143		0		http://ww					
7	Color	Andrew St	462	132	475	530	Samantha	640	73058679	Action|Ad	Daryl Saba	John Carte	212204	1873	Polly Walk	1	alien|ame	http://ww	738	English	USA	PG-13	2.64E+08
8	Color	Sam Raimi	392	156	0	4000	James Frar	24000	3.37E+08	Action|Ad	J.K. Simmc	Spider-Ma	383056	46055	Kirsten Du	0	sandman|	http://ww	1902	English	USA	PG-13	2.58E+08
9	Color	Nathan Gr	324	100	15	284	Donna Mu	799	2.01E+08	Adventure	Brad Garre	TangledÂ	294810	2036	M.C. Gaine	1	17th centu	http://ww	387	English	USA	PG	2.6E+08
10	Color	Joss Whed	635	141	0	19000	Robert Do	26000	4.59E+08	Action|Ad	Chris Hem	Avengers:	462669	92000	Scarlett Jo	4	artificial in	http://ww	1117	English	USA	PG-13	2.5E+08
11	Color	David Yate	375	153	282	10000	Daniel Rad	25000	3.02E+08	Adventure	Alan Rickn	Harry Pott	321795	58753	Rupert Gri	3	blood|boc	http://ww	973	English	UK	PG	2.5E+08
12	Color	Zack Snyde	673	183	0	2000	Lauren Col	15000	3.3E+08	Action|Ad	Henry Cavi	Batman v S	371639	24450	Alan D. Pu	0	based on c	http://ww	3018	English	USA	PG-13	2.5E+08
13	Color	Bryan Sing	434	169	0	903	Marlon Bra	18000	2E+08	Action|Ad	Kevin Spac	Superman	240396	29991	Frank Lang	0	crystal|epi	http://ww	2367	English	USA	PG-13	2.09E+08
14	Color	Marc Forst	403	106	395	393	Mathieu A	451	1.68E+08	Action|Ad	Giancarlo	Quantum o	330784	2023	Rory Kinne	1	action her	http://ww	1243	English	UK	PG-13	2E+08
15	Color	Gore Verbi	313	151	563	1000	Orlando Bl	40000	4.23E+08	Action|Ad	Johnny De	Pirates of	522040	48486	Jack Daver	2	box office	http://ww	1832	English	USA	PG-13	2.25E+08
16	Color	Gore Verbi	450	150	563	1000	Ruth Wilsc	40000	89289910	Action|Ad	Johnny De	The Lone F	181792	45757	Tom Wilkir	1	horse|out	http://ww	711	English	USA	PG-13	2.15E+08
17	Color	Zack Snyde	733	143	0	748	Christophe	15000	2.91E+08	Action|Ad	Henry Cavi	Man of Ste	548573	20495	Harry Lenr	0	based on c	http://ww	2536	English	USA	PG-13	2.25E+08
18	Color	Andrew Ac	258	150	80	201	Pierfrance	22000	1.42E+08	Action|Ad	Peter Dink	The Chron	149922	22697	DamiÃ¡n A	4	brother br	http://ww	438	English	USA	PG	2.25E+08
19	Color	Joss Whed	703	173	0	19000	Robert Do	26000	6.23E+08	Action|Ad	Chris Hem	The Aveng	995415	87697	Scarlett Jo	3	alien invas	http://ww	1722	English	USA	PG-13	2.2E+08
20	Color	Rob Marsh	448	136	252	1000	Sam Claflir	40000	2.41E+08	Action|Ad	Johnny De	Pirates of	370704	54083	Stephen Gi	4	blackbeard	http://ww	484	English	USA	PG-13	2.5E+08
21	Color	Barry Sonr	451	106	188	718	Michael St	10000	1.79E+08	Action|Ad	Will Smith	Men in Bla	268154	12572	Nicole Sch	1	alien|crim	http://ww	341	English	USA	PG-13	2.25E+08
22	Color	Peter Jack	422	164	0	773	Adam Brov	5000	2.55E+08	Adventure	Aidan Turn	The Hobbi	354228	9152	James Nes	0	army|elf|l	http://ww	802	English	New Zeala	PG-13	2.5E+08
23	Color	Marc Web	599	153	464	963	Andrew Ga	15000	2.62E+08	Action|Ad	Emma Sto	The Amazi	451803	28489	Chris Zylka	0	lizard|out	http://ww	1225	English	USA	PG-13	2.3E+08
24	Color	Ridley Sco	343	156	0	738	William Hu	891	1.05E+08	Action|Ad	Mark Addy	Robin Hoo	211765	3244	Scott Grim	0	1190s|arc	http://ww	546	English	USA	PG-13	2E+08
25	Color	Peter Jack	509	186	0	773	Adam Brov	5000	2.58E+08	Adventure	Aidan Turn	The Hobbi	483540	9152	James Nes	6	dwarf|elf|	http://ww	951	English	USA	PG-13	2.25E+08
26	Color	Chris Weit	251	113	129	1000	Eva Green	16000	70083519	Adventure	Christophe	The Golde	149019	24106	Kristin Sco	2	children|e	http://ww	666	English	USA	PG-13	1.8E+08

Before Cleaning

director_name	num_critic_for_review	duration	actor_2_name	gross	genres	actor_1_name	movie_title
Frank Darabont	199	142	Jeffrey DeMunn	28341469	Crime|Drama	Morgan Freeman	The Shawshank RedemptionÂ
Christopher Nolan	645	152	Heath Ledger	5.33E+08	Action|Crime|Drama|Thriller	Christian Bale	The Dark KnightÂ
Christopher Nolan	642	148	Tom Hardy	2.93E+08	Action|Adventure|Sci-Fi|Thriller	Leonardo DiCaprio	InceptionÂ
David Fincher	315	151	Meat Loaf	37023395	Drama	Brad Pitt	Fight ClubÂ
Quentin Tarantino	215	178	Eric Stoltz	1.08E+08	Crime|Drama	Bruce Willis	Pulp FictionÂ
Robert Zemeckis	149	142	Siobhan Fallon Hogan	3.3E+08	Comedy|Drama	Tom Hanks	Forrest GumpÂ
Peter Jackson	297	171	Orlando Bloom	3.14E+08	Action|Adventure|Drama|Fantasy	Christopher Lee	The Lord of the Rings: The Fellowship of the RingÂ
Lana Wachowski	313	136	Marcus Chong	1.71E+08	Action|Sci-Fi	Keanu Reeves	The MatrixÂ
Peter Jackson	328	192	Billy Boyd	3.77E+08	Action|Adventure|Drama|Fantasy	Orlando Bloom	The Lord of the Rings: The Return of the KingÂ
Francis Ford Coppola	208	175	Marlon Brando	1.35E+08	Crime|Drama	Al Pacino	The GodfatherÂ
Christopher Nolan	813	164	Christian Bale	4.48E+08	Action|Thriller	Tom Hardy	The Dark Knight RisesÂ
Peter Jackson	294	172	Orlando Bloom	3.4E+08	Action|Adventure|Drama|Fantasy	Christopher Lee	The Lord of the Rings: The Two TowersÂ
David Fincher	216	127	Brad Pitt	1E+08	Crime|Drama|Mystery|Thriller	Morgan Freeman	Se7enÂ
Joss Whedon	703	173	Robert Downey Jr.	6.23E+08	Action|Adventure|Sci-Fi	Chris Hemsworth	The AvengersÂ
Ridley Scott	265	171	Connie Nielsen	1.88E+08	Action|Drama|Romance	Djimon Hounsou	GladiatorÂ
Christopher Nolan	478	128	Liam Neeson	2.05E+08	Action|Adventure	Christian Bale	Batman BeginsÂ
Quentin Tarantino	765	165	Christoph Waltz	1.63E+08	Drama|Western	Leonardo DiCaprio	Django UnchainedÂ
Christopher Nolan	712	169	Anne Hathaway	1.88E+08	Adventure|Drama|Sci-Fi	Matthew McConaughey	InterstellarÂ
George Lucas	282	125	Peter Cushing	4.61E+08	Action|Adventure|Fantasy|Sci-Fi	Harrison Ford	Star Wars: Episode IV - A New HopeÂ
Jonathan Demme	185	138	Scott Glenn	1.31E+08	Crime|Drama|Horror|Thriller	Anthony Hopkins	The Silence of the LambsÂ
James Cameron	723	178	Joel David Moore	7.61E+08	Action|Adventure|Fantasy|Sci-Fi	CCH Pounder	AvatarÂ
Quentin Tarantino	486	153	Brad Pitt	1.21E+08	Adventure|Drama|War	Michael Fassbender	Inglourious BasterdsÂ
Steven Spielberg	219	169	Vin Diesel	2.16E+08	Action|Drama|War	Tom Hanks	Saving Private RyanÂ
Martin Scorsese	352	151	Matt Damon	1.32E+08	Crime|Drama|Thriller	Leonardo DiCaprio	The DepartedÂ
Steven Spielberg	174	185	Embeth Davidtz	96067179	Biography|Drama|History	Liam Neeson	Schindler's ListÂ

After Cleaning

Insights:

A. ***Movie Genre Analysis:*** Analyse the distribution of movie genres and their impact on the IMDB score.

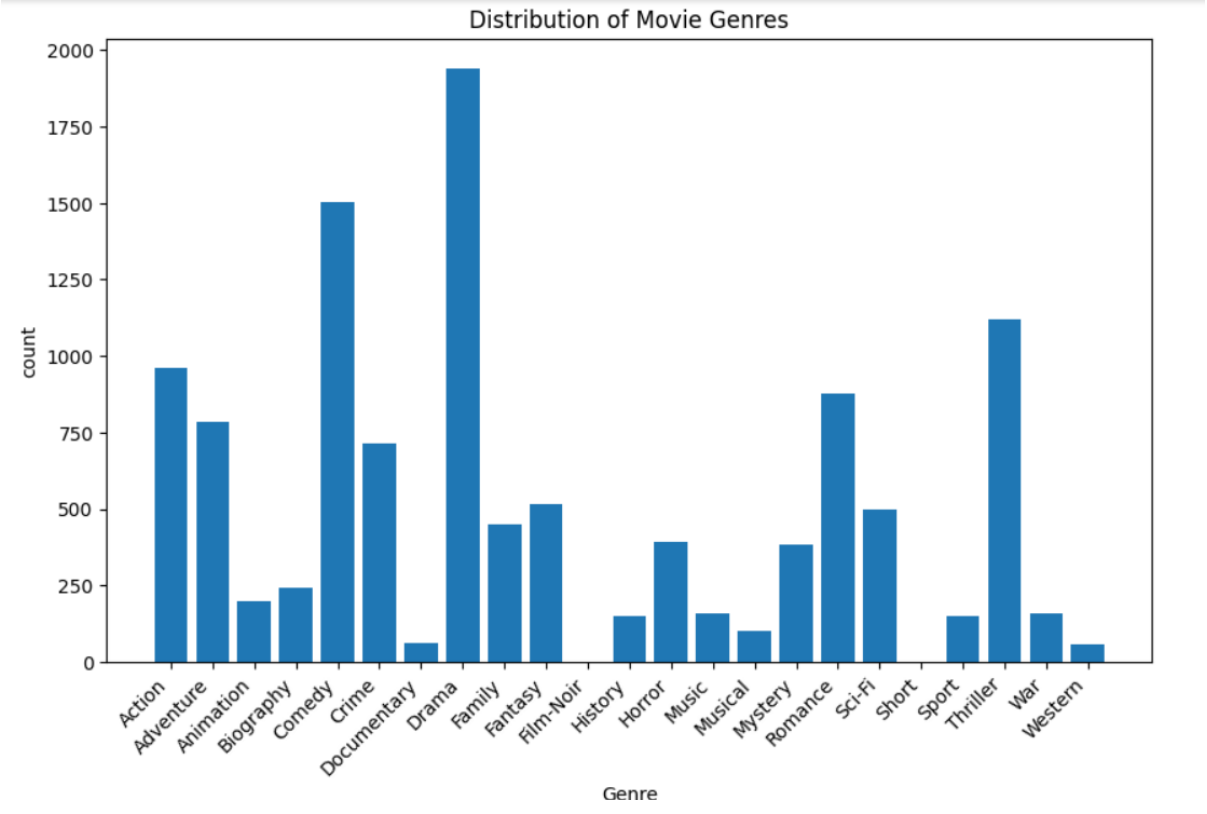
Task: Determine the most common genres of movies in the dataset. Then, for each genre, calculate descriptive statistics (mean, median, mode, range, variance, standard deviation) of the IMDB scores.

Result: The top 10 most common genres are Drama, Comedy, Thriller, Action, Romance, Adventure, Crime, Fantasy, Sci-Fi, and Family. Also, all the top 10 genres' descriptive statistics are almost at the same level.

Below is the descriptive statistics of all genres

index	Genre	count	sum	mean	median	std	var	mode	max	min	range
0	Action	961	6044.8	6.290114464099896	6.3	1.031734021466284	1.0644750910509908	6.1	9.0	2.1	6.9
1	Adventure	786	5075.8	6.457760814249364	6.6	1.1087363892891116	1.2292963809338562	6.7	8.9	2.3	6.600000000000000
2	Animation	199	1333.4	6.700502512562815	6.8	0.9885964690049267	0.977322978529009	6.7	8.6	2.8	5.8
3	Biography	243	1735.4	7.141563786008231	7.2	0.7090040907262799	0.502686800666599	7.0	8.9	4.5	4.4
4	Comedy	1503	9295.3	6.184497671324018	6.3	1.0379832937338016	1.0774093180704716	6.7	8.8	1.9	6.9
5	Crime	714	4673.0	6.544817927170868	6.6	0.9801578616955859	0.9607094338436635	6.6	9.3	2.4	6.9
6	Documentary	63	440.4	6.990476190476190	7.2	1.234153005325108	1.523133640552996	6.6	8.5	1.6	6.9
7	Drama	1940	13165.1	6.786134020618557	6.9	0.8913245861482937	0.7944595178724271	6.7	9.3	2.1	7.200000000000001
8	Family	450	2794.7	6.210444444444444	6.3	1.162576157058177	1.351583320960159	6.7	8.6	1.9	6.699999999999999
9	Fantasy	514	3233.1	6.290077821011673	6.4	1.1298703516226734	1.2766070114759436	6.7	8.9	2.2	6.7
10	Film-Noir	1	7.7	7.7	7.7	NaN	NaN	7.7	7.7	7.7	0.0
11	History	152	1084.7	7.136184210526365	7.2	0.6748607255977633	0.45543699895433964	7.7	8.9	5.5	3.4000000000000004
12	Horror	391	2317.1	5.926086956521739	6.0	0.9970905277979044	0.9941895206243037	5.9	8.6	2.3	6.3
13	Music	159	1013.1	6.371698113207548	6.5	1.2140596233493026	1.4739407690470505	6.5	8.5	1.6	6.9
14	Musical	103	675.6	6.559223300970874	6.7	1.1409865780176591	1.301850371216448	7.1	8.5	2.1	6.4
15	Mystery	384	2485.6	6.472916666666666	6.5	1.0056217090725765	1.0112750217580497	6.6	8.6	3.1	5.5
16	Romance	877	5638.4	6.429190421892816	6.5	0.9672683876719188	0.9356081337894333	6.5	8.5	2.1	6.4
17	Sci-Fi	497	3142.4	6.322736418511067	6.4	1.1560962874904688	1.3365586259492448	6.7	8.8	1.9	6.9
18	Short	2	13.6	6.8	6.8	0.4242640687119283	0.1799999999999998	6.5	7.1	6.5	0.5999999999999996
19	Sport	151	997.1	6.603311258278146	6.8	1.0433546676334011	1.088588962472405	7.2	8.4	2.0	6.4
20	Thriller	1118	7129.9	6.377370304114449	6.4	0.9660102947742104	0.933175889609757	6.5	9.0	2.7	6.3
21	War	160	1128.6	7.053749999999999	7.1	0.807970319835976	0.6528160377358493	7.1	8.6	4.3	4.3
22	Western	58	392.4	6.76551724137931	6.8	0.9985167463192473	0.997035692679976	6.8	8.9	4.1	4.800000000000001

Q1) What is the most and least common genre amongst all genres?
Ans: Drama is the most common genre of all. And Film-noir is the least common genre. 1940 movies have Drama as its genre and only one movie is based on Film-noir as a genre.

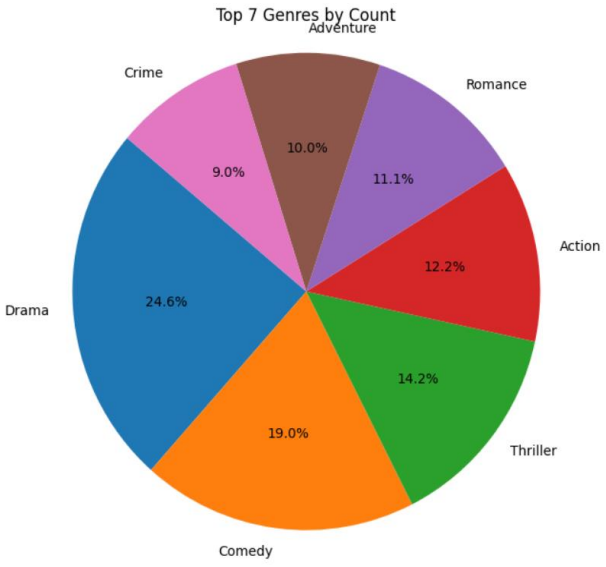


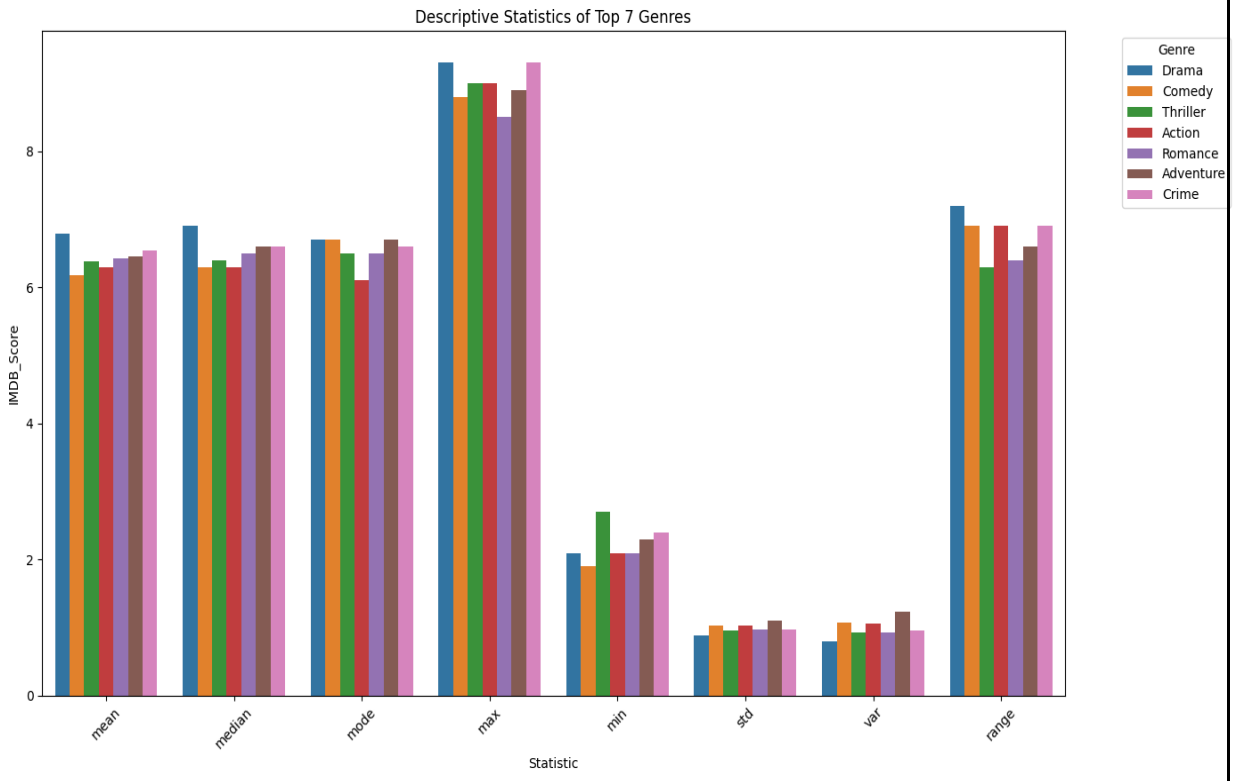
Q2) Explain the distribution of the top 7 genres.
Ans: The top 7 genres include Drama, Comedy, Thriller, Action, Romance, Adventure, and Crime.
- The mean suggests that in all 7 genres, the average of the movies' IMDB scores is

between 6-7.
- The median suggests that in all 7 genres, half of the movies have a score between 6-7 and the other half have a score of above 6-7.
- The mode suggests that in all 7 genres, most of the movies' IMDB score is between 6-7.
- The standard deviation measures the amount of variation or dispersion in a set of values.

A low standard deviation indicates that the IMDb scores are clustered closely around the mean, suggesting consistent quality. A high standard deviation indicates a wide range of ratings, suggesting varying quality. Genres like Drama, Thriller, Romance, and Crime show that the ratings vary moderately around the mean as std is less than 1.
Genres like Comedy, Action, and Adventure show that some ratings are slightly deviated.
- Genres like Comedy, Action, and Adventure have a higher variance as compared to other genres indicating that they have comparatively varying IMDb scores.
- The max IMDb score is of Drama and Crime with a rating of 9.3. The min IMDb score is of Comedy with a rating of 1.9.

	Genre	count	sum	mean	median	std	var	mode	max	min	range
7	Drama	1940	13165.1	6.786134	6.9	0.891325	0.794460	6.7	9.3	2.1	7.2
4	Comedy	1503	9295.3	6.184498	6.3	1.037983	1.077409	6.7	8.8	1.9	6.9
20	Thriller	1118	7129.9	6.377370	6.4	0.966010	0.933176	6.5	9.0	2.7	6.3
0	Action	961	6044.8	6.290114	6.3	1.031734	1.064475	6.1	9.0	2.1	6.9
16	Romance	877	5638.4	6.429190	6.5	0.967268	0.935608	6.5	8.5	2.1	6.4
1	Adventure	786	5075.8	6.457761	6.6	1.108736	1.229296	6.7	8.9	2.3	6.6
5	Crime	714	4673.0	6.544818	6.6	0.980158	0.960709	6.6	9.3	2.4	6.9





B. ***Movie Duration Analysis:*** *Analyse the distribution of movie durations and its impact on the IMDB score.*

Task: Analyse the distribution of movie durations and identify the relationship between movie duration and IMDB score.

Result: The distribution of Movie Durations shows that it closely follows a Normal Distribution. Also, the scatter plot shows that duration and imdb_scores have a positive relationship.

	movie_title	duration	imdb_\nscore
0	The Shawshank RedemptionÂ	142	9.3
1	The Dark KnightÂ	152	9.0
2	InceptionÂ	148	8.8
3	Fight ClubÂ	151	8.8
4	Pulp FictionÂ	178	8.9
...	...	...	...
3842	Time to ChooseÂ	100	7.0
3843	Call + ResponseÂ	86	7.5
3844	The Knife of Don JuanÂ	110	7.2
3845	Born to Fly: Elizabeth Streb vs. GravityÂ	82	6.8
3846	Mi AmericaÂ	125	7.2

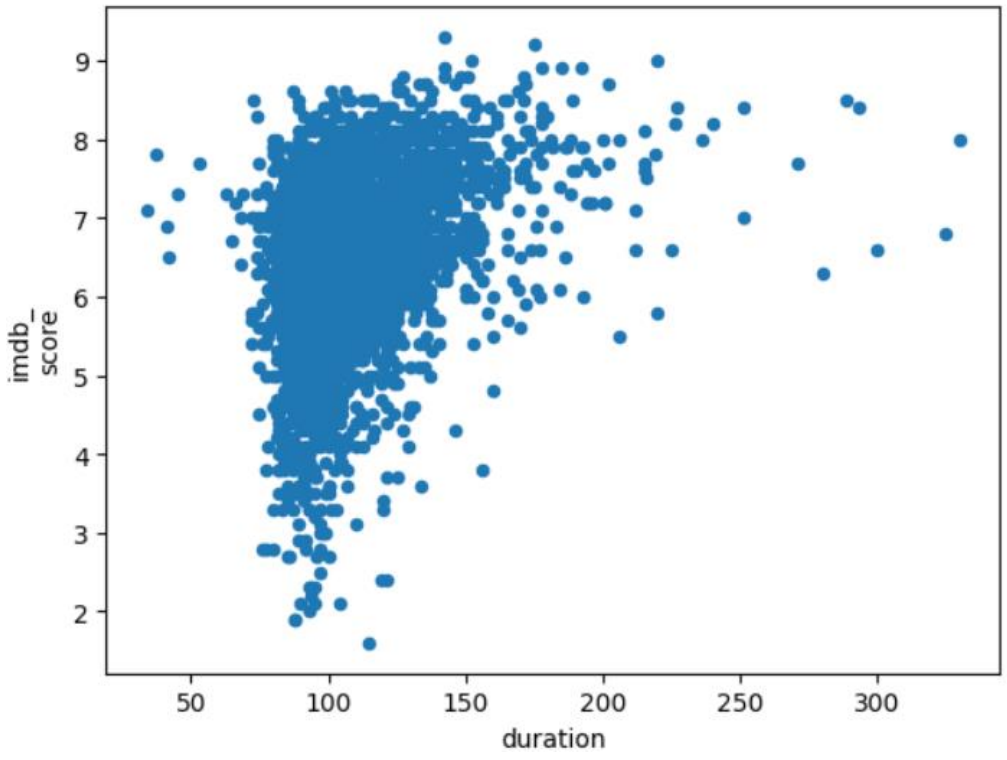
3847 rows × 3 columns

Q1) Explain the descriptive statistics of movie duration across movies.
Ans:

	mean	median	mode	std	var	max	min
movie_duration							
0	109.930075	106.0	101	22.752953	517.696877	330	34

- The mean suggests that the average of the movie's duration is 109 minutes.
- The mode suggests that the duration for most of the movies is 101 minutes.
- The median suggests that across all movies, half of the movies have a duration below 106 mins and the other half have a duration above 106 mins.
- The standard deviation measures the amount of variation or dispersion in a set of values.
Here the std dev is 22.75 which is significantly higher. A high standard deviation indicates that the duration is way farther from the mean.
- The variance is quite high which is 517.69. It indicates it has a highly varying duration in the movie dataset.
- The movie with the longest duration is 'Blood In, Blood Out ' with a duration of 330 minutes. The movie with the shortest duration is 'Marilyn Hotchkiss's Ballroom Dancing and Charm School ' with a duration of 34 minutes.

Q2) Explain the distribution of movie duration across all the movies in the dataset.
Ans:



- Very few movies have a duration of less than 75 minutes.
- As we can see from the scatter plot, most of the movies lie between time duration of 75 mins to 150 mins.
- As duration increases, the number of movies decreases.
- Very movies have a duration of 200-300mins.

C. ***Language Analysis:*** Situation: Examine the distribution of movies based on their language.
Task: Determine the most common languages used in movies and analyse their impact on the IMDB score using descriptive statistics.
Result: From the analysis it clearly shows that English is the most common language used in movies followed by French, Spanish, Hindi, and Mandarin.

	movie_title	language	imdb_\nscore
0	The Shawshank RedemptionÂ	English	9.3
1	The Dark KnightÂ	English	9.0
2	InceptionÂ	English	8.8
3	Fight ClubÂ	English	8.8
4	Pulp FictionÂ	English	8.9
...	...	...	...
3842	Time to ChooseÂ	English	7.0
3843	Call + ResponseÂ	English	7.5
3844	The Knife of Don JuanÂ	Spanish	7.2
3845	Born to Fly: Elizabeth Streb vs. GravityÂ	English	6.8
3846	Mi AmericaÂ	English	7.2

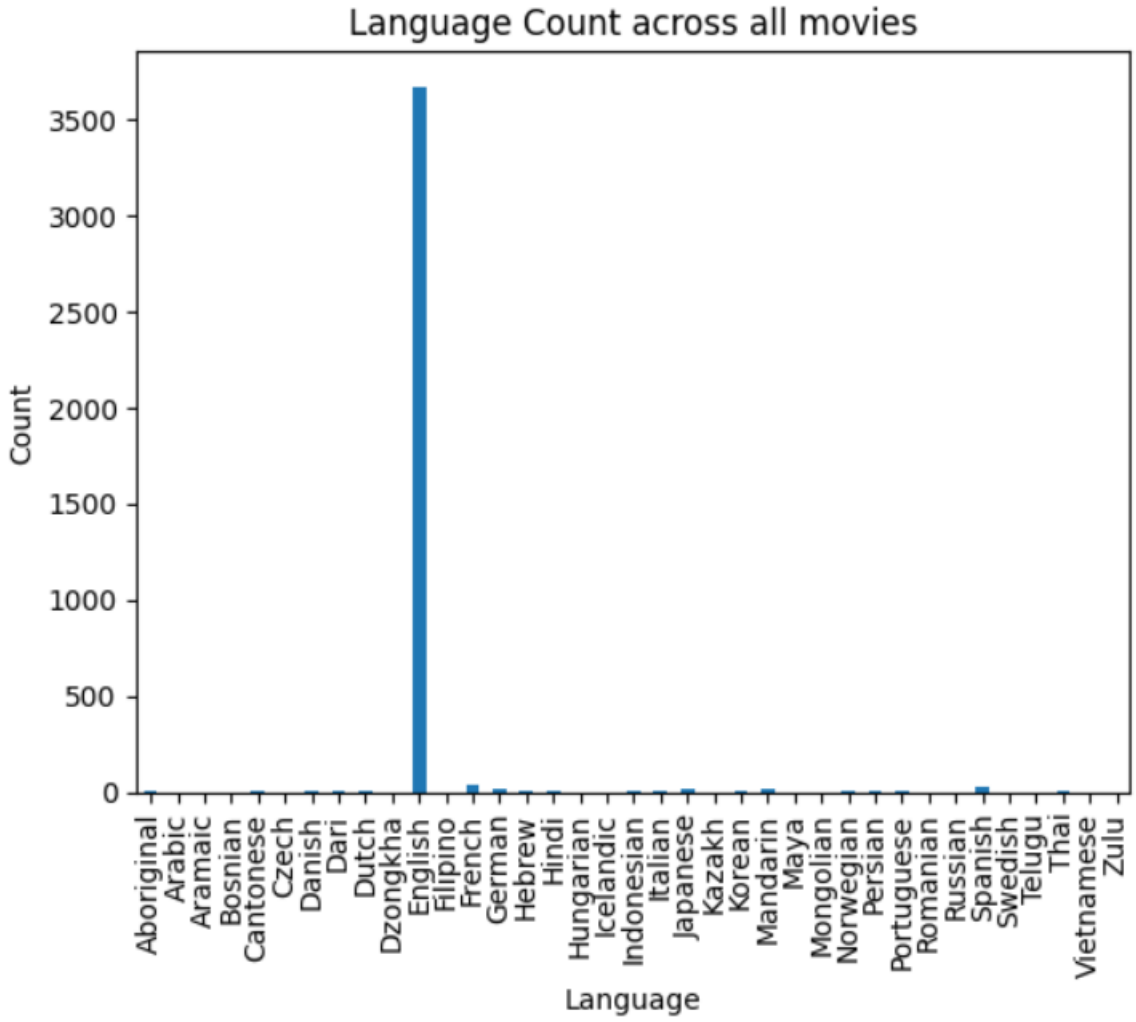
3847 rows × 3 columns

Q1) Explain the descriptive statistics of movie duration across movies.
Ans:
- English is the most used language followed by French, Spanish, Hindi, and Mandarin in the given movies' dataset. 3668 movies have English as their language.
- Telegu has the highest mean of 8.4, followed by Persian. Most of the languages have an overall mean of 7.
- The French language has comparatively higher mean and median but lower standard deviation implying that most of the French language movies have their IMDb score on the higher side.

language	count	mean	median	std
Aboriginal	2	6.950000	6.95	0.777817
Arabic	1	7.200000	7.20	NaN
Aramaic	1	7.100000	7.10	NaN
Bosnian	1	4.300000	4.30	NaN
Cantonese	8	7.237500	7.30	0.440576
Czech	1	7.400000	7.40	NaN
Danish	3	7.900000	8.10	0.529150
Dari	2	7.500000	7.50	0.141421
Dutch	3	7.566667	7.80	0.404145
Dzongkha	1	7.500000	7.50	NaN
English	3668	6.423555	6.50	1.048809
Filipino	1	6.700000	6.70	NaN
French	37	7.286486	7.20	0.561329
German	13	7.692308	7.70	0.640913
Hebrew	2	7.650000	7.65	0.494975
Hindi	10	6.760000	7.05	1.111755
Hungarian	1	7.100000	7.10	NaN
Icelandic	1	6.900000	6.90	NaN
Indonesian	2	7.900000	7.90	0.424264
Italian	7	7.185714	7.00	1.155319
Japanese	12	7.625000	7.80	0.899621
Kazakh	1	6.000000	6.00	NaN
Korean	5	7.700000	7.70	0.570088
Mandarin	14	7.021429	7.25	0.765786
Maya	1	7.800000	7.80	NaN
Mongolian	1	7.300000	7.30	NaN
Norwegian	4	7.150000	7.30	0.574456
Persian	3	8.133333	8.40	0.550757
Portuguese	5	7.760000	8.00	0.978775
Romanian	1	7.900000	7.90	NaN
Russian	1	6.500000	6.50	NaN
Spanish	26	7.050000	7.15	0.826196
Swedish	1	7.600000	7.60	NaN
Telugu	1	8.400000	8.40	NaN
Thai	3	6.633333	6.60	0.450925
Vietnamese	1	7.400000	7.40	NaN

Q2) Why English is the most common language?

Ans: The plot shows that the USA is the most common country in the dataset and most films in the USA are in English language as it is the most spoken language in the country.



D. ***Director Analysis:*** Influence of directors on movie ratings.

Task: Identify the top directors based on their average IMDB score and analyse their contribution to the success of movies using percentile calculations.

Ans:

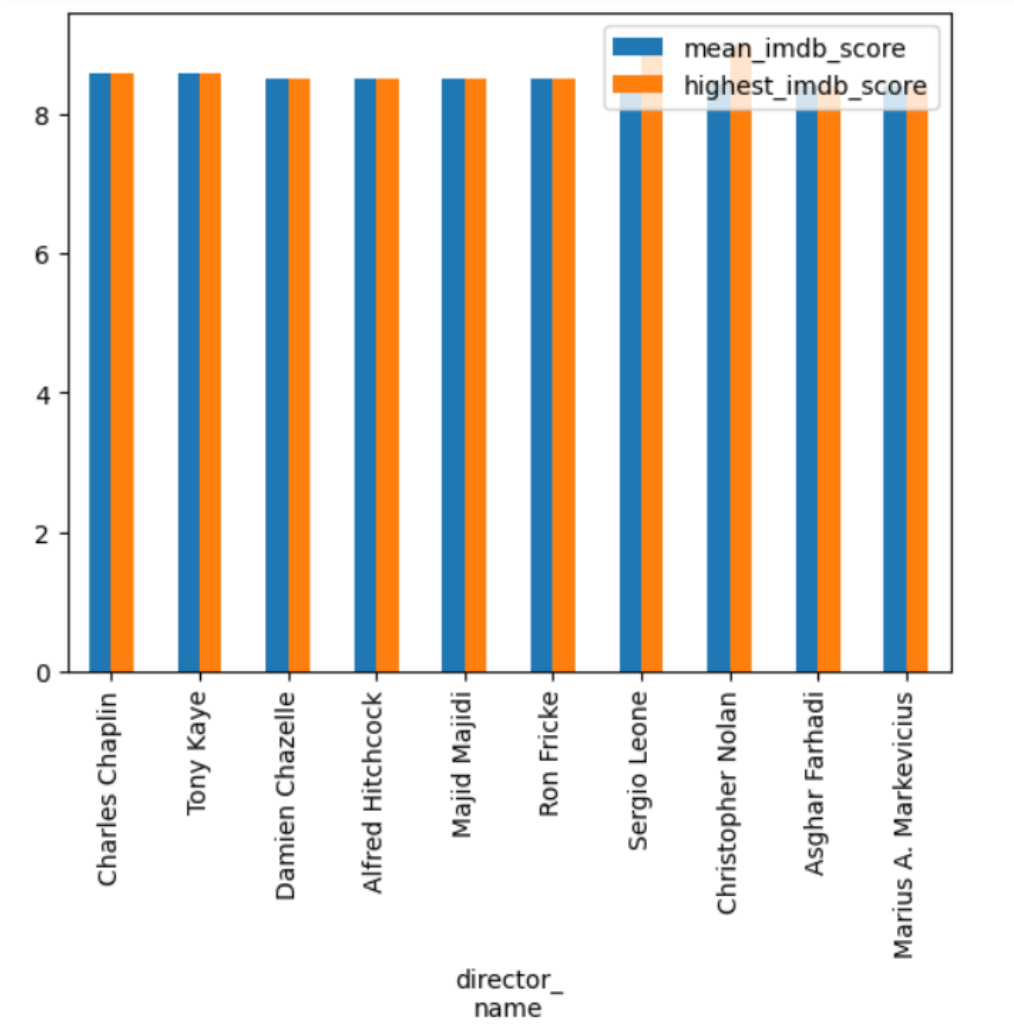
The analysis given below is of the top 10 directors from the given movies' dataset. It gives a correlation between IMDb scores to give a significant outlook on the analysis.

	director_\nname	imdb_\nscore	movie_title	profit
150	Alfred Hitchcock	8.5	PsychoÂ	31193053
1417	Asghar Farhadi	8.4	A SeparationÂ	6598492
820	Charles Chaplin	8.6	Modern TimesÂ	-1336755
1	Christopher Nolan	9.0	The Dark KnightÂ	348316061
170	Damien Chazelle	8.5	WhiplashÂ	9792000
3059	Majid Majidi	8.5	Children of HeavenÂ	745402
3620	Marius A. Markevicius	8.4	The Other Dream TeamÂ	-366222
3173	Ron Fricke	8.5	SamsaraÂ	-1398153
460	Sergio Leone	8.9	The Good, the Bad and the UglyÂ	4900000
35	Tony Kaye	8.6	American History XÂ	-787759

Q1) Define a correlation between IMDb score and Profit.
Ans:
Most of the directors have an average IMDB score between 8 to 9. Christopher Nolan is the only director who has an IMDb score of 9 and also has the highest profitable movies of all. The above analysis shows that in some cases, even if the IMDB score is quite high the movies aren't profitable at the box office. Those directors are Charles Chaplin, Marius A. Markevicius, Ron Fricke, and Tony Kaye.
Other than that most directors have done well with both IMDB scores and at the box office collection.

Q2) Define a correlation between mean IMDb scores and percentile scores of top 10 directors.
Ans:

	director_\nname	mean_imdb_score	highest_imdb_score
0	Charles Chaplin	8.600000	8.6
1	Tony Kaye	8.600000	8.6
2	Damien Chazelle	8.500000	8.5
3	Alfred Hitchcock	8.500000	8.5
4	Majid Majidi	8.500000	8.5
5	Ron Fricke	8.500000	8.5
6	Sergio Leone	8.433333	8.9
7	Christopher Nolan	8.425000	9.0
8	Asghar Farhadi	8.400000	8.4
9	Marius A. Markevicius	8.400000	8.4



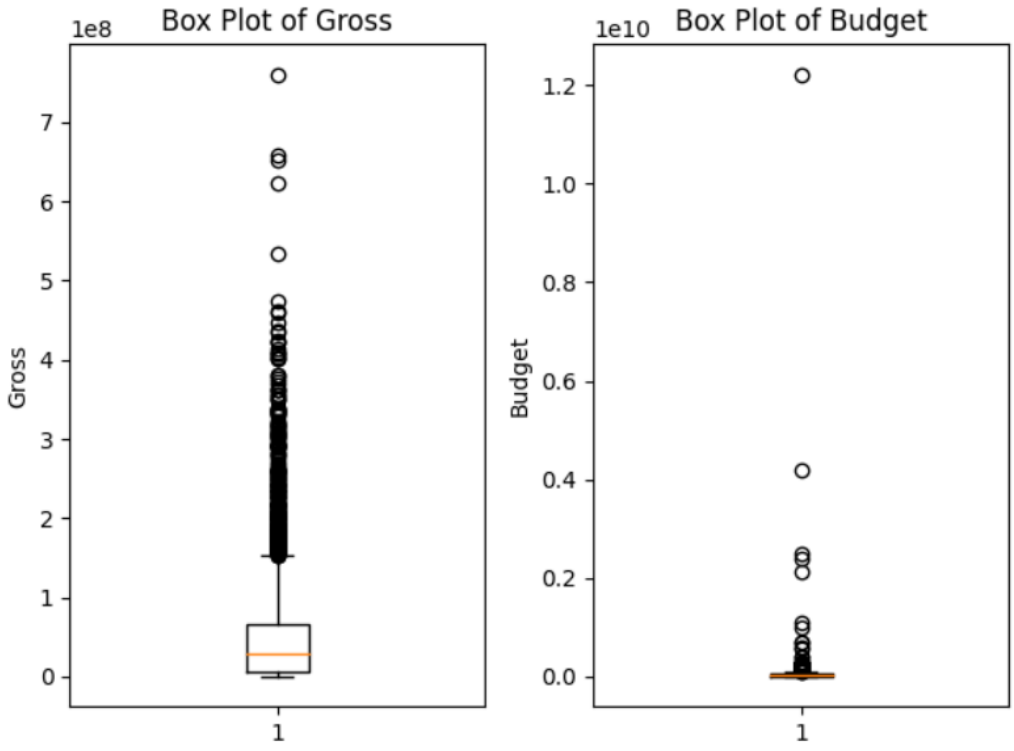
Most of the directors have the same mean IMDb score as their percentile score. This means that they have provided consistent efforts to make a high-quality movie in all of the movies they have directed.
Sergio Leone and Christopher Nolan have a higher percentile IMDb score than their average IMDb score. The above bar graph shows that both the bars (orange and blue) have some rating except for Sergio Leone and Christopher Nolan.

E. ***Budget Analysis:*** Explore the relationship between movie budgets and their financial success.
Task: Analyse the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.
Result:
The below analysis is about the movie budgets and gross earnings and what effect it has on profit margin.

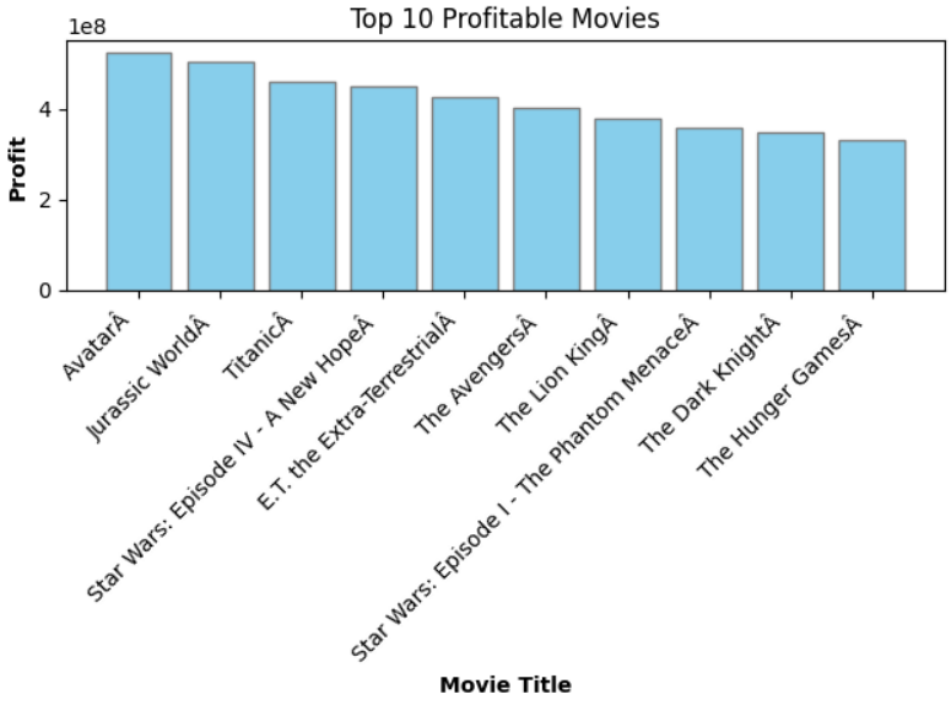
	gross	budget	profit
0	28341469	25000000	3341469
1	533316061	185000000	348316061
2	292568851	160000000	132568851
3	37023395	63000000	-25976605
4	107930000	8000000	99930000
...	...	...	...
3842	29233	3500000	-3470767
3843	215185	200000	15185
3844	3830	1200000	-1196170
3845	21199	500000	-478801
3846	3330	2100000	-2096670

3847 rows × 3 columns

Q1) Explain the correlation between Gross and Budget.
Ans: The box plot shows that the correlation between Gross and Budget is positive. That is, the relationship shows that as the budget of movies increases, there is a very high probability that the gross collection of the movie will also increase.



Q2) Mention the highest profitable movies.
Ans:
To derive this I have created a column named Profit which is the subtraction of Budget and Gross. The highest profitable movie is Avatar with a profit margin of $ 523505847 followed by Jurassic World, and Titanic.



Summary:

- It appears that movies like "Avatar" and "Jurassic Park" have the potential to earn high profits.
- If the goal is to maximize profit, it may be advisable to consider making movies with similar themes or characteristics.
- It appears that the movie "The Shawshank Redemption" has the highest IMDB score among those with a minimum of 25,000 voted users.
- From the top 250 IMDB movies, we can conclude that only 37 of them are not in the English language. This suggests that English is a preferable language for these films.
- Consider working with Christopher Nolan, Tony Kaye and Charles Chaplin as a director on future projects, as their past work has received high ratings from audiences and critics.
- It appears that the Crime|Drama|Fantasy|Mystery genre has the highest average IMDB score, indicating that it is a more preferable genre.

Conclusion:

From this project, I got an understanding of cleaning and modifying data which is very important. Also removing unnecessary data i.e. columns, blank spaces or cells, etc. makes the data much more readable. Identifying and removing outliers changes the results of data analysis. Presenting data with charts and graphs makes it look way more interesting and clearer.

Through this project, I was able to understand the importance of Data Analytics in Movies analysis as it provides valuable insights such as the director's relationship with IMDB Score, the genre's relationship with IMDB Score, the budget's relationship with IMDB Score etc. which helps in making Data-Driven Decisions.

- Both during the pre-production and post-production phases of a movie, analysis is a key component.
- Additionally, it's not a given that the movie with the highest IMDB rating will also make the most money.
- The quantity of tickets sold by theatres around the world is the real basis for profit calculation.
- In conclusion, I'd like to state that, prior to the production of a film, not only movie producers but also a variety of financiers, stakeholders, and cinema outlet owners perform IMDB Movie Analysis or any other similar analysis.