IMDB Movie Analysis

Project Description: This project aims to analyze the factors that influence the success of a movie on IMDb, with success defined by high IMDb ratings. By investigating variables such as genre, director, cast, budget, release year, and runtime, we can identify key contributors to a movie's performance. The findings will provide valuable insights for producers, directors, and investors to make data-driven decisions for future film projects.

Approach: To complete this project I have undergone the following steps:

DATA PREPARATION: We start by downloading and analyzing the data. Through this, we learn that we do not need all the columns in the data set. So further, we only retain the useful columns such as

director_name

duration

gross

genres

movie_title

language

country

budget

title_year

imdb_score

And by deleting the rows which have null values.

DATA ANALYSIS: After cleaning the data we performed different types of analyses to perform the given tasks such as descriptive analysis, data extraction, and visualizing the relationship between 2 or more variables.

Insights:

A. Movie Genre Analysis: Analyze 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.

- Output:

Genre	IMDB_score
Action	7.9
Adventure	7.9
Fantasy	7.9
Action	7.1
Adventure	7.1
Fantasy	7.1
Action	6.8
Adventure	6.8
Thriller	6.8
Action	8.5
Thriller	8.5
Action	6.6
Adventure	6.6
Sci-Fi	6.6
Action	6.2
Adventure	6.2
Romance	6.2
Adventure	7.8
Animation	7.8
Comedy	7.8
Action	7.5
Adventure	7.5
Sci-Fi	7.5
Adventure	7.5
Family	7.5
Fantasy	7.5

Descriptive Statistics on IMDB Score for each Genre

Genre	Count of Genres	Sum	Mean	Median	Mode	Max	Min	Range	Variance	Standard Deviation
Action	970	6101.9	6.29	6.35	6.6	9	2.1	6.9	1.08	1.04
Adventure	795	5132.3	6.46	6.6	6.7	8.9	2.3	6.6	1.23	1.11
Fantasy	357	2223.4	6.23	6.3	6.7	8.8	2.2	6.6	1.26	1.12
Thriller	601	3883	6.46	6.5	6.4	8.5	2.8	5.7	0.94	0.97
Sci-Fi	331	2109.6	6.37	6.4	6.4	8.8	1.9	6.9	1.33	1.15
Romance	688	4418.5	6.42	6.5	6.5	8.5	2.1	6.4	0.91	0.95
Animation	199	1333.4	6.70	6.8	6.7	8.6	2.8	5.8	0.98	0.99
Comedy	1492	9214.2	6.18	6.3	6.7	8.8	1.9	6.9	1.08	1.04
Family	294	1798.3	6.12	6.2	5.4	8.6	1.9	6.7	1.41	1.19
Western	30	202.7	6.76	6.6	6.5	8.9	4.1	4.8	1.11	1.06
Drama	1914	12990.8	6.79	6.9	6.7	9.3	2.1	7.2	0.80	0.89
Crime	705	4620.6	6.55	6.6	6.6	9.3	2.4	6.9	0.96	0.98
Horror	382	2268.3	5.94	6	5.9	8.6	2.3	6.3	0.98	0.99
History	114	815.7	7.16	7.2	7.7	8.9	5.6	3.3	0.45	0.67
Biography	244	1742.6	7.14	7.2	7	8.9	4.5	4.4	0.50	0.71
Mystery	319	2062.7	6.47	6.5	6.8	8.6	3.3	5.3	0.99	0.99
Sport	109	727.6	6.68	6.8	7.2	8.4	2	6.4	1.18	1.09
War	75	533.4	7.11	7.2	7.7	8.6	4.3	4.3	0.75	0.86
Musical	49	321	6.55	6.8	6.2	8	2.1	5.9	1.46	1.21
Documentary	67	469.8	7.01	7.2	6.6	8.5	1.6	6.9	1.44	1.20
Music	130	834.1	6.42	6.65	6.5	8.5	1.6	6.9	1.44	1.20
Short	1	6.5	6.50	6.5	0	6.5	6.5	0	0.00	0.00
Film-Noir	1	7.7	7.70	7.7	0	7.7	7.7	0	0.00	0.00

- Result: To arrive at the above result, I started with extracting the genre by using the split-by-delimiter function, and then I used the index function to unpivot the separated genre by using the formula, [INDEX($B$2:$D$5,INT((ROW(A1)-1)/3)+1,MOD(ROW(A1)-1,3)+1)] and [=INDEX($E$2:$E$5,INT((ROW(A1)-1)/3)+1)]
- After applying the above formula to get the modified data we move on to conduct the descriptive analysis where we calculate the count, sum of the IMDB Score, mean median, mode, max, min, range, variance, and standard deviation.

  By applying the formulas like,

Count of Genres	{=COUNTIF(A:A,E4)}
Sum	{=SUM(IF(A$2:A$9868=E4, B$2:B$9868))}
Mean	{=AVERAGEIF(A:A,E4,B:B)}
Median	{=MEDIAN(IF(A$2:A$9868=E4, B$2:B$9868))}
Mode	{=MODE(IF(A$2:A$9868=E4, B$2:B$9868))}
Max	{=MAX(IF(A$2:A$9868=E4, B$2:B$9868))}
Min	{=MIN(IF(A$2:A$9868=E4, B$2:B$9868))}
Range	{=K4-L4}
Variance	{=VAR.S(IF(A$2:A$9868=E4, B$2:B$9868))}
Standard Deviation	{=STDEV.S(IF(A$2:A$9868=E4, B$2:B$9868))}

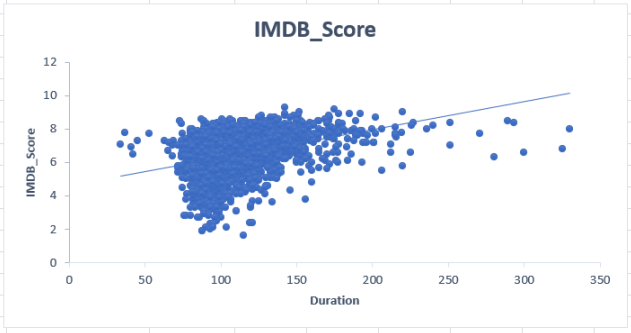
B. Movie Duration Analysis: Analyze the distribution of movie durations and its impact on the IMDB score.

- Task: Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.
- Output:

movie_title	Duration	IMDB Score
AvatarÂ	178	7.9
Pirates of the Caribbean: At World's EndÂ	169	7.1
SpectreÂ	148	6.8
The Dark Knight RisesÂ	164	8.5
John CarterÂ	132	6.6
Spider-Man 3Â	156	6.2
TangledÂ	100	7.8
Avengers: Age of UltronÂ	141	7.5
Harry Potter and the Half-Blood PrinceÂ	153	7.5
Batman v Superman: Dawn of JusticeÂ	183	6.9
Superman ReturnsÂ	169	6.1
Quantum of SolaceÂ	106	6.7
Pirates of the Caribbean: Dead Man's ChestÂ	151	7.3
The Lone RangerÂ	150	6.5
Man of SteelÂ	143	7.2
The Chronicles of Narnia: Prince CaspianÂ	150	6.6
The AvengersÂ	173	8.1
Pirates of the Caribbean: On Stranger TidesÂ	136	6.7
Men in Black 3Â	106	6.8
The Hobbit: The Battle of the Five ArmiesÂ	164	7.5
The Amazing Spider-ManÂ	153	7
Robin HoodÂ	156	6.7
The Hobbit: The Desolation of SmaugÂ	186	7.9
The Golden CompassÂ	113	6.1
King KongÂ	201	7.2
TitanicÂ	194	7.7

Descriptive Statistics of IMDB Score and Duration

Descriptive Statistics	IMDB Score	Duration
Mean	6.4642655	109.902
Median	6.6	106
Mode	6.7	101
Max	9.3	330
Min	1.6	34
Variance	1.1150094	515.8604
Standard Deviation	1.0558042	22.70964



- Result: In the above output we have performed the descriptive analysis on the duration and IMDB score of the movies and a scatter plot between them.

C. Language Analysis: Situation: Examine the distribution of movies based on their language.

- Task: Determine the most common languages used in movies and analyze their impact on the IMDB score using descriptive statistics.
- Output:

movie_title	language	imdb_score
AvatarÂ	English	7.9
Pirates of the Caribbean: At World's EndÂ	English	7.1
SpectreÂ	English	6.8
The Dark Knight RisesÂ	English	8.5
John CarterÂ	English	6.6
Spider-Man 3Â	English	6.2
TangledÂ	English	7.8
Avengers: Age of UltronÂ	English	7.5
Harry Potter and the Half-Blood PrinceÂ	English	7.5
Batman v Superman: Dawn of JusticeÂ	English	6.9
Superman ReturnsÂ	English	6.1
Quantum of SolaceÂ	English	6.7
Pirates of the Caribbean: Dead Man's ChestÂ	English	7.3
The Lone RangerÂ	English	6.5
Man of SteelÂ	English	7.2
The Chronicles of Narnia: Prince CaspianÂ	English	6.6
The AvengersÂ	English	8.1
Pirates of the Caribbean: On Stranger TidesÂ	English	6.7
Men in Black 3Â	English	6.8
The Hobbit: The Battle of the Five ArmiesÂ	English	7.5
The Amazing Spider-ManÂ	English	7
Robin HoodÂ	English	6.7
The Hobbit: The Desolation of SmaugÂ	English	7.9
The Golden CompassÂ	English	6.1
King KongÂ	English	7.2
TitanicÂ	English	7.7

Descriptive Statistics on IMDB Score for each Language

language	Count	Sum	Mean	Median	Max	Min	Range	Variance	Standard Deviation
English	3706	23807.50	6.42	6.50	9.30	1.60	7.70	1.10	1.05
French	37	269.60	7.29	7.20	8.40	5.80	2.60	0.32	0.56
Spanish	26	183.30	7.05	7.15	8.20	5.20	3.00	0.68	0.83
Mandarin	15	106.20	7.08	7.40	7.90	5.60	2.30	0.60	0.77
German	13	100.00	7.69	7.70	8.50	6.10	2.40	0.41	0.64
Japanese	12	91.50	7.63	7.80	8.70	6.00	2.70	0.81	0.90
Hindi	10	67.60	6.76	7.05	8.00	4.80	3.20	1.24	1.11
Cantonese	8	57.90	7.24	7.30	7.80	6.50	1.30	0.19	0.44
Italian	7	50.30	7.19	7.00	8.90	5.30	3.60	1.33	1.16
Korean	5	38.50	7.70	7.70	8.40	7.00	1.40	0.33	0.57
Portugues	5	38.80	7.76	8.00	8.70	6.10	2.60	0.96	0.98
Norwegian	4	28.60	7.15	7.30	7.60	6.40	1.20	0.33	0.57
Dutch	3	22.70	7.57	7.80	7.80	7.10	0.70	0.16	0.40
Thai	3	19.90	6.63	6.60	7.10	6.20	0.90	0.20	0.45
Danish	3	23.70	7.90	8.10	8.30	7.30	1.00	0.28	0.53
Hebrew	3	22.50	7.50	7.30	8.00	7.20	0.80	0.19	0.44
Persian	3	24.40	8.13	8.40	8.50	7.50	1.00	0.30	0.55
Aboriginal	2	13.90	6.95	6.95	7.50	6.40	1.10	0.61	0.78
Dari	2	15.00	7.50	7.50	7.60	7.40	0.20	0.02	0.14
Indonesian	2	15.80	7.90	7.90	8.20	7.60	0.60	0.18	0.42
Filipino	1	6.70	6.70	6.70	6.70	6.70	0.00	0.00	0.00
Maya	1	7.80	7.80	7.80	7.80	7.80	0.00	0.00	0.00
Kazakh	1	6.00	6.00	6.00	6.00	6.00	0.00	0.00	0.00
Telugu	1	8.40	8.40	8.40	8.40	8.40	0.00	0.00	0.00

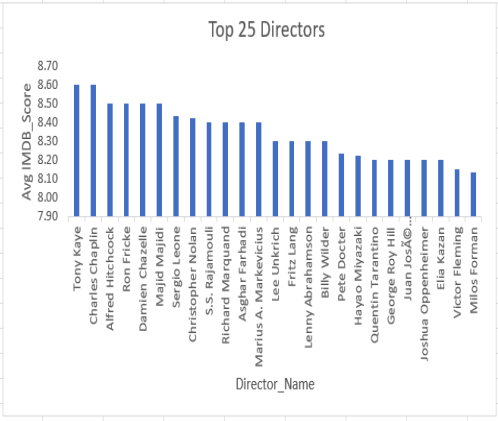
- Result: the language with the highest IMDB Score is English followed by French, Spanish, Mandarin, and German.

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

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

movie_title	director_name	imdb_score
AvatarÂ	James Cameron	7.9
Pirates of the Caribbean: A	Gore Verbinski	7.1
SpectreÂ	Sam Mendes	6.8
The Dark Knight RisesÂ	Christopher Nola	8.5
John CarterÂ	Andrew Stanton	6.6
Spider-Man 3Â	Sam Raimi	6.2
TangledÂ	Nathan Greno	7.8
Avengers: Age of UltronÂ	Joss Whedon	7.5
Harry Potter and the Half-E	David Yates	7.5
Batman v Superman: Dawn	Zack Snyder	6.9
Superman ReturnsÂ	Bryan Singer	6.1
Quantum of SolaceÂ	Marc Forster	6.7
Pirates of the Caribbean: D	Gore Verbinski	7.3
The Lone RangerÂ	Gore Verbinski	6.5
Man of SteelÂ	Zack Snyder	7.2
The Chronicles of Narnia: P	Andrew Adamsor	6.6
The AvengersÂ	Joss Whedon	8.1
Pirates of the Caribbean: O	Rob Marshall	6.7
Men in Black 3Â	Barry Sonnenfeld	6.8
The Hobbit: The Battle of t	Peter Jackson	7.5
The Amazing Spider-ManÂ	Marc Webb	7
Robin HoodÂ	Ridley Scott	6.7
The Hobbit: The Desolation	Peter Jackson	7.9
The Golden CompassÂ	Chris Weitz	6.1
King KongÂ	Peter Jackson	7.2
TitanicÂ	James Cameron	7.7

director_name	count	Mean
Tony Kaye	1	8.60
Charles Chaplin	1	8.60
Alfred Hitchcock	1	8.50
Ron Fricke	1	8.50
Damien Chazelle	1	8.50
Majid Majidi	1	8.50
Sergio Leone	3	8.43
Christopher Nolan	8	8.43
S.S. Rajamouli	1	8.40
Richard Marquand	1	8.40
Asghar Farhadi	1	8.40
Marius A. Markevicius	1	8.40
Lee Unkrich	1	8.30
Fritz Lang	1	8.30
Lenny Abrahamson	1	8.30
Billy Wilder	1	8.30
Pete Docter	3	8.23
Hayao Miyazaki	4	8.23
Quentin Tarantino	8	8.20
George Roy Hill	2	8.20
Juan JosÃ© Campa	1	8.20
Joshua Oppenheimer	1	8.20
Elia Kazan	1	8.20
Victor Fleming	2	8.15
Milos Forman	3	8.13
Akira Kurosawa	2	8.10



- Result: The director with the highest IMDB Score is Tony Kaye with an 8.60 IMDB Score.

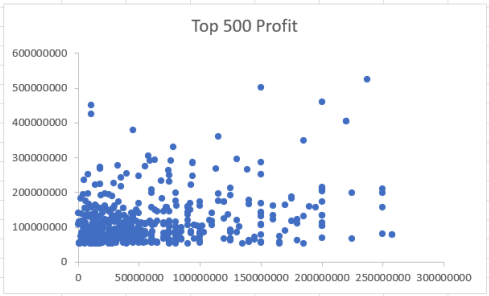
E. Budget Analysis: Explore the relationship between movie budgets and their financial success.

- Task: Analyze the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.
- Output:

	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 King	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 Bac	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
27	MinionsÂ	336029560	74000000	262029560

Correlation 0.102129

This indicates a week positive correlation



- Result: On calculating the correlation between gross and budget, we get a correlation of 0.102129, indicating a weak but positive correlation.

Drive link: IMDB Movie Analysis.xlsx