IMDB Movie Analysis

Description:

Problem Statement: The dataset provided is related to IMDB Movies. A potential problem 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 problem 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.

Project Description

The project aimed to investigate the factors influencing the success of movies on IMDb, focusing on determining what attributes correlate with higher IMDb ratings. By analyzing IMDb movie data, the objective was to provide insights to aid movie producers, directors, and investors in making informed decisions for future projects.

Approach

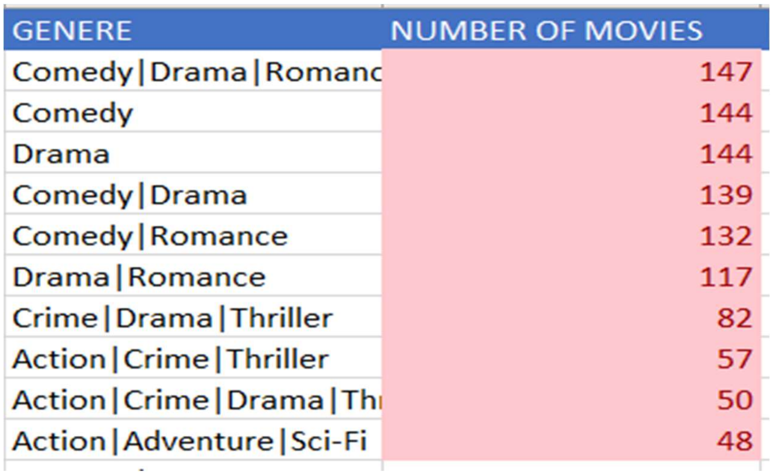
- Data Collection: Describe how the IMDb movie dataset was obtained and its characteristics.
- Data Cleaning: Detail the steps taken for data preprocessing, including handling missing values, duplicates, data type conversion, and feature engineering.
- Data Analysis Techniques: Explain the methods used to explore relationships between variables (correlation analysis, statistical modeling, etc.).

Tech-Stack Used

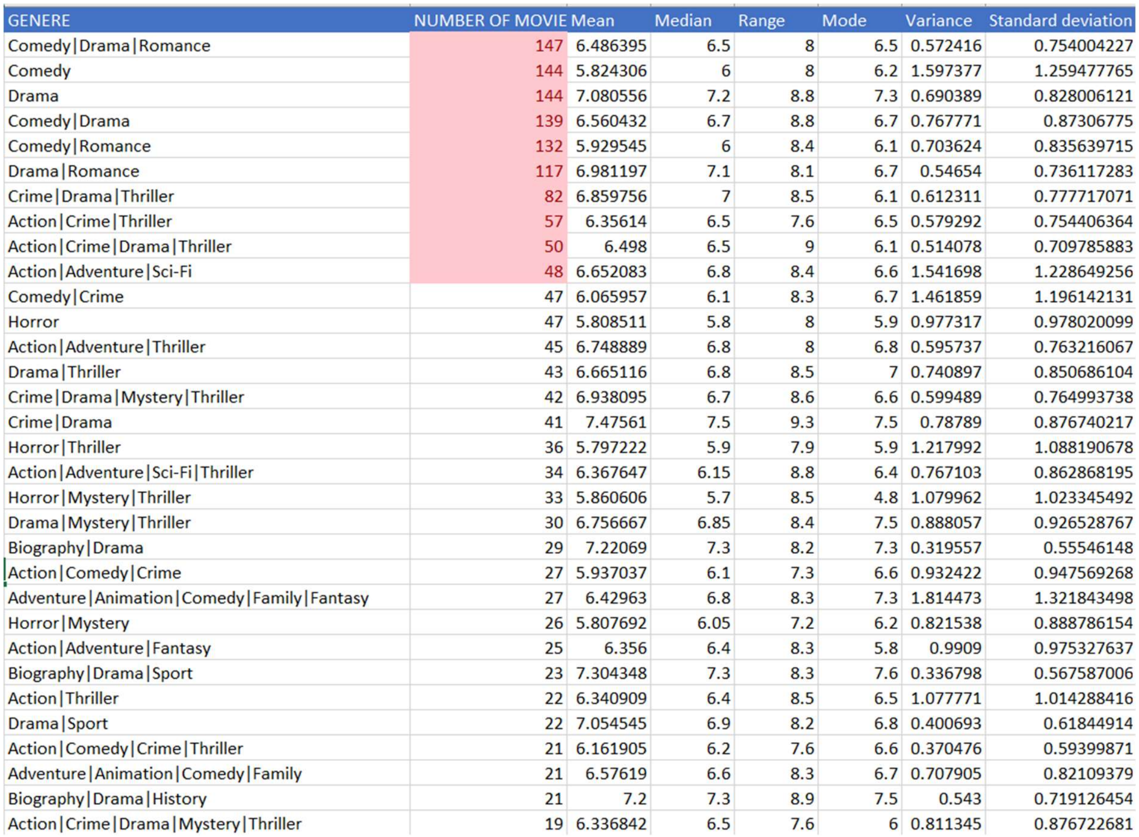
Software: Microsoft excel

Data Analytics Tasks:

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.
   Hint: Use Excel's COUNTIF function to count the number of movies for each genre. You might need to manipulate the 'genres' column to separate multiple genres for a single movie. Use Excel's functions like AVERAGE, MEDIAN, MODE, MAX, MIN, VAR, and STDEV to calculate descriptive statistics. Compare the statistics to understand the impact of genre on movie ratings.
   - Top 10 Common Genres of movies



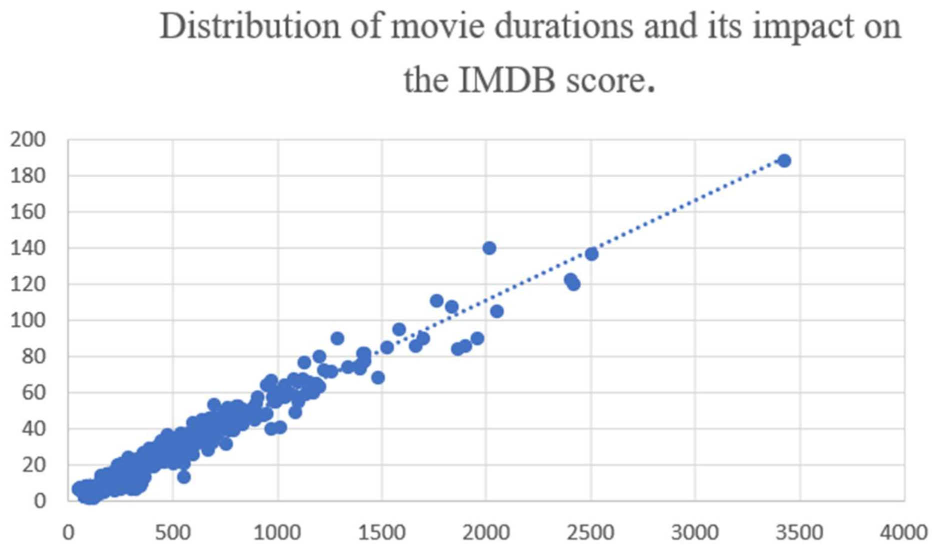
GENERE	NUMBER OF MOVIES
Comedy|Drama|Romanc	147
Comedy	144
Drama	144
Comedy|Drama	139
Comedy|Romance	132
Drama|Romance	117
Crime|Drama|Thriller	82
Action|Crime|Thriller	57
Action|Crime|Drama|Th	50
Action|Adventure|Sci-Fi	48

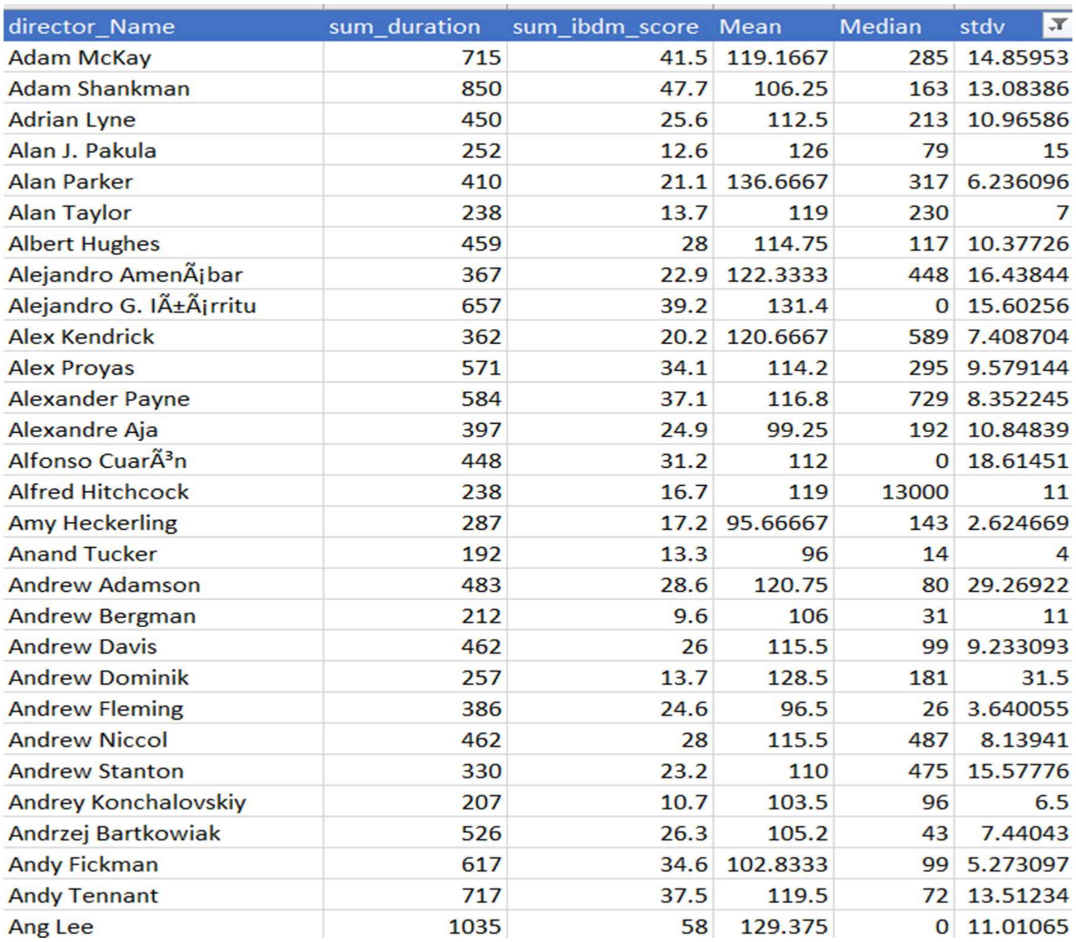


GENERE	NUMBER OF MOVIE	Mean	Median	Range	Mode	Variance	Standard deviation
Comedy|Drama|Romance	147	6.486395	6.5	8	6.5	0.572416	0.754004227
Comedy	144	5.824306	6	8	6.2	1.597377	1.259477765
Drama	144	7.080556	7.2	8.8	7.3	0.690389	0.828006121
Comedy|Drama	139	6.560432	6.7	8.8	6.7	0.767771	0.87306775
Comedy|Romance	132	5.929545	6	8.4	6.1	0.703624	0.835639715
Drama|Romance	117	6.981197	7.1	8.1	6.7	0.54654	0.736117283
Crime|Drama|Thriller	82	6.859756	7	8.5	6.1	0.612311	0.777717071
Action|Crime|Thriller	57	6.35614	6.5	7.6	6.5	0.579292	0.754406364
Action|Crime|Drama|Thriller	50	6.498	6.5	9	6.1	0.514078	0.709785883
Action|Adventure|Sci-Fi	48	6.652083	6.8	8.4	6.6	1.541698	1.228649256
Comedy|Crime	47	6.065957	6.1	8.3	6.7	1.461859	1.196142131
Horror	47	5.808511	5.8	8	5.9	0.977317	0.978020099
Action|Adventure|Thriller	45	6.748889	6.8	8	6.8	0.595737	0.763216067
Drama|Thriller	43	6.665116	6.8	8.5	7	0.740897	0.850686104
Crime|Drama|Mystery|Thriller	42	6.938095	6.7	8.6	6.6	0.599489	0.764993738
Crime|Drama	41	7.47561	7.5	9.3	7.5	0.78789	0.876740217
Horror|Thriller	36	5.797222	5.9	7.9	5.9	1.217992	1.088190678
Action|Adventure|Sci-Fi|Thriller	34	6.367647	6.15	8.8	6.4	0.767103	0.862868195
Horror|Mystery|Thriller	33	5.860606	5.7	8.5	4.8	1.079962	1.023345492
Drama|Mystery|Thriller	30	6.756667	6.85	8.4	7.5	0.888057	0.926528767
Biography|Drama	29	7.22069	7.3	8.2	7.3	0.319557	0.55546148
Action|Comedy|Crime	27	5.937037	6.1	7.3	6.6	0.932422	0.947569268
Adventure|Animation|Comedy|Family|Fantasy	27	6.42963	6.8	8.3	7.3	1.814473	1.321843498
Horror|Mystery	26	5.807692	6.05	7.2	6.2	0.821538	0.888786154
Action|Adventure|Fantasy	25	6.356	6.4	8.3	5.8	0.9909	0.975327637
Biography|Drama|Sport	23	7.304348	7.3	8.3	7.6	0.336798	0.567587006
Action|Thriller	22	6.340909	6.4	8.5	6.5	1.077771	1.014288416
Drama|Sport	22	7.054545	6.9	8.2	6.8	0.400693	0.61844914
Action|Comedy|Crime|Thriller	21	6.161905	6.2	7.6	6.6	0.370476	0.59399871
Adventure|Animation|Comedy|Family	21	6.57619	6.6	8.3	6.7	0.707905	0.82109379
Biography|Drama|History	21	7.2	7.3	8.9	7.5	0.543	0.719126454
Action|Crime|Drama|Mystery|Thriller	19	6.336842	6.5	7.6	6	0.811345	0.876722681

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.
Hint: Calculate descriptive statistics such as mean, median, and standard deviation for movie durations. Use Excel's functions like AVERAGE, MEDIAN, and STDEV.

Create a scatter plot to visualize the relationship between movie duration and IMDB score. Add a trendline to assess the direction and strength of the relationship.


Distribution of movie durations and its impact on the IMDB score.

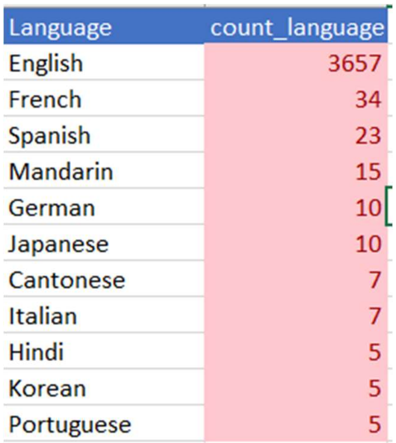


director_Name	sum_duration	sum_ibdm_score	Mean	Median	stdv
Adam McKay	715	41.5	119.1667	285	14.85953
Adam Shankman	850	47.7	106.25	163	13.08386
Adrian Lyne	450	25.6	112.5	213	10.96586
Alan J. Pakula	252	12.6	126	79	15
Alan Parker	410	21.1	136.6667	317	6.236096
Alan Taylor	238	13.7	119	230	7
Albert Hughes	459	28	114.75	117	10.37726
Alejandro AmenÃ¡bar	367	22.9	122.3333	448	16.43844
Alejandro G. IÃ±Ã¡rritu	657	39.2	131.4	0	15.60256
Alex Kendrick	362	20.2	120.6667	589	7.408704
Alex Proyas	571	34.1	114.2	295	9.579144
Alexander Payne	584	37.1	116.8	729	8.352245
Alexandre Aja	397	24.9	99.25	192	10.84839
Alfonso CuarÃ³n	448	31.2	112	0	18.61451
Alfred Hitchcock	238	16.7	119	13000	11
Amy Heckerling	287	17.2	95.66667	143	2.624669
Anand Tucker	192	13.3	96	14	4
Andrew Adamson	483	28.6	120.75	80	29.26922
Andrew Bergman	212	9.6	106	31	11
Andrew Davis	462	26	115.5	99	9.233093
Andrew Dominik	257	13.7	128.5	181	31.5
Andrew Fleming	386	24.6	96.5	26	3.640055
Andrew Niccol	462	28	115.5	487	8.13941
Andrew Stanton	330	23.2	110	475	15.57776
Andrey Konchalovskiy	207	10.7	103.5	96	6.5
Andrzej Bartkowiak	526	26.3	105.2	43	7.44043
Andy Fickman	617	34.6	102.8333	99	5.273097
Andy Tennant	717	37.5	119.5	72	13.51234
Ang Lee	1035	58	129.375	0	11.01065

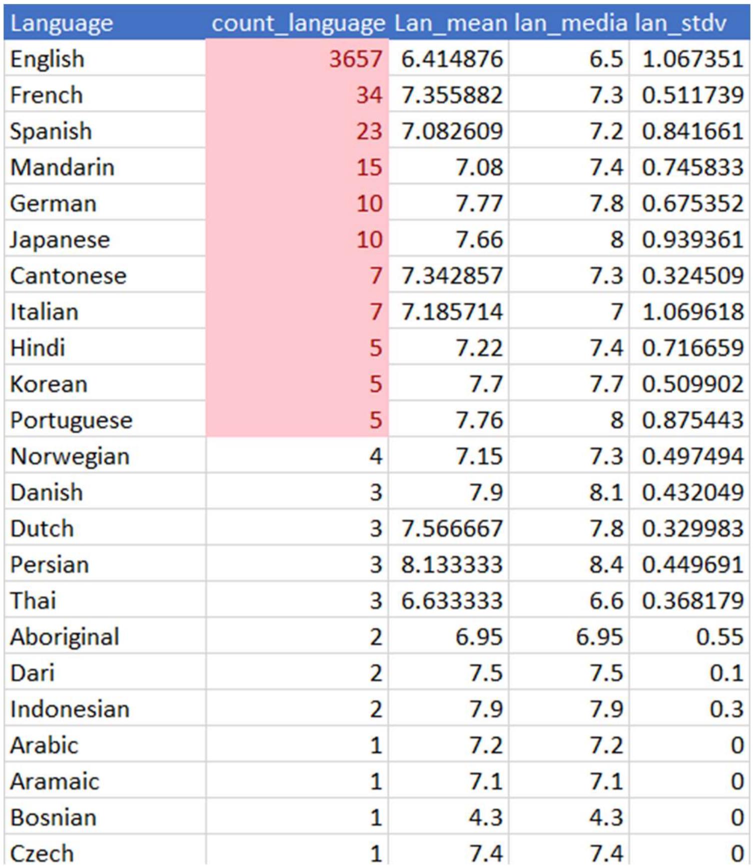
C. Language Analysis: Situation: Examine the distribution of movies based on their language.

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

Hint: Use Excel's COUNTIF function to count the number of movies for each language. Calculate the mean, median, and standard deviation of the IMDB scores for each language. Compare the statistics to understand the impact of language on movie ratings.



Language	count_language
English	3657
French	34
Spanish	23
Mandarin	15
German	10
Japanese	10
Cantonese	7
Italian	7
Hindi	5
Korean	5
Portuguese	5

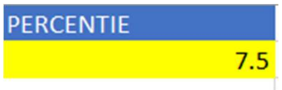


Language	count_language	Lan_mean	lan_media	lan_stdv
English	3657	6.414876	6.5	1.067351
French	34	7.355882	7.3	0.511739
Spanish	23	7.082609	7.2	0.841661
Mandarin	15	7.08	7.4	0.745833
German	10	7.77	7.8	0.675352
Japanese	10	7.66	8	0.939361
Cantonese	7	7.342857	7.3	0.324509
Italian	7	7.185714	7	1.069618
Hindi	5	7.22	7.4	0.716659
Korean	5	7.7	7.7	0.509902
Portuguese	5	7.76	8	0.875443
Norwegian	4	7.15	7.3	0.497494
Danish	3	7.9	8.1	0.432049
Dutch	3	7.566667	7.8	0.329983
Persian	3	8.133333	8.4	0.449691
Thai	3	6.633333	6.6	0.368179
Aboriginal	2	6.95	6.95	0.55
Dari	2	7.5	7.5	0.1
Indonesian	2	7.9	7.9	0.3
Arabic	1	7.2	7.2	0
Aramaic	1	7.1	7.1	0
Bosnian	1	4.3	4.3	0
Czech	1	7.4	7.4	0

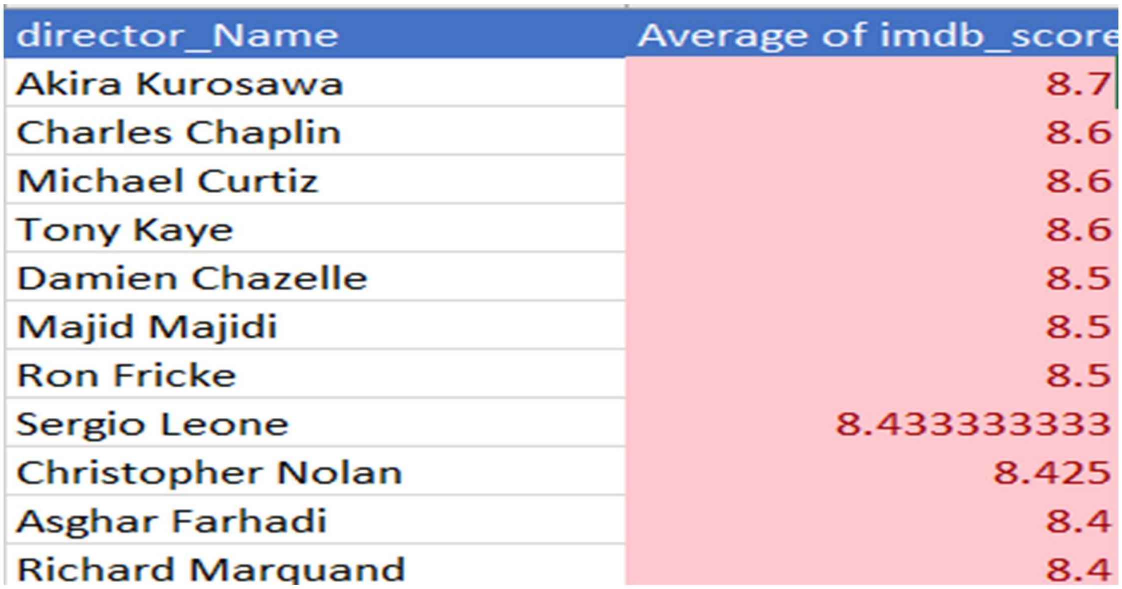
D. Director Analysis: Influence of directors on movie ratings.

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

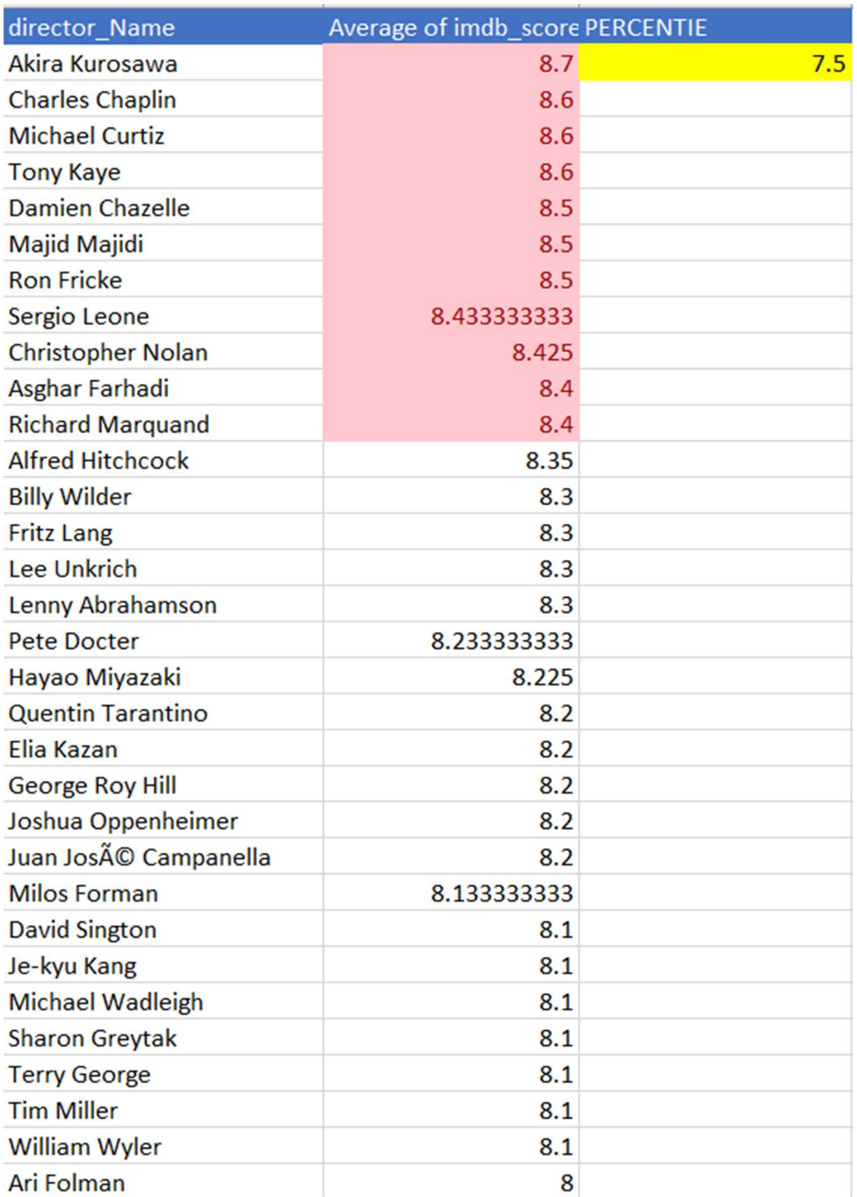
Hint: Calculate the average IMDB score for each director. Use Excel's PERCENTILE function to identify the directors with the highest scores. Compare the scores of these directors to the overall distribution of scores.



PERCENTIE
7.5



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



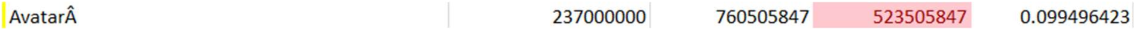
director_Name	Average of imdb_score	PERCENTIE
Akira Kurosawa	8.7	7.5
Charles Chaplin	8.6	
Michael Curtiz	8.6	
Tony Kaye	8.6	
Damien Chazelle	8.5	
Majid Majidi	8.5	
Ron Fricke	8.5	
Sergio Leone	8.433333333	
Christopher Nolan	8.425	
Asghar Farhadi	8.4	
Richard Marquand	8.4	
Alfred Hitchcock	8.35	
Billy Wilder	8.3	
Fritz Lang	8.3	
Lee Unkrich	8.3	
Lenny Abrahamson	8.3	
Pete Docter	8.233333333	
Hayao Miyazaki	8.225	
Quentin Tarantino	8.2	
Elia Kazan	8.2	
George Roy Hill	8.2	
Joshua Oppenheimer	8.2	
Juan JosÃ© Campanella	8.2	
Milos Forman	8.133333333	
David Sington	8.1	
Je-kyu Kang	8.1	
Michael Wadleigh	8.1	
Sharon Greytak	8.1	
Terry George	8.1	
Tim Miller	8.1	
William Wyler	8.1	
Ari Folman	8	

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

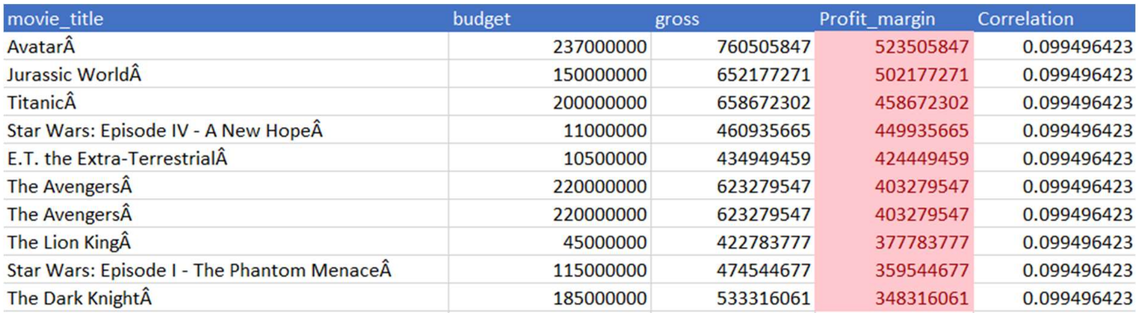
Task: Analyze the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.

Hint: Calculate the correlation coefficient between movie budgets and gross earnings using Excel's CORREL function. Calculate the profit margin (gross earnings - budget) for each movie and identify the movies with the highest profit margin using Excel's MAX function.

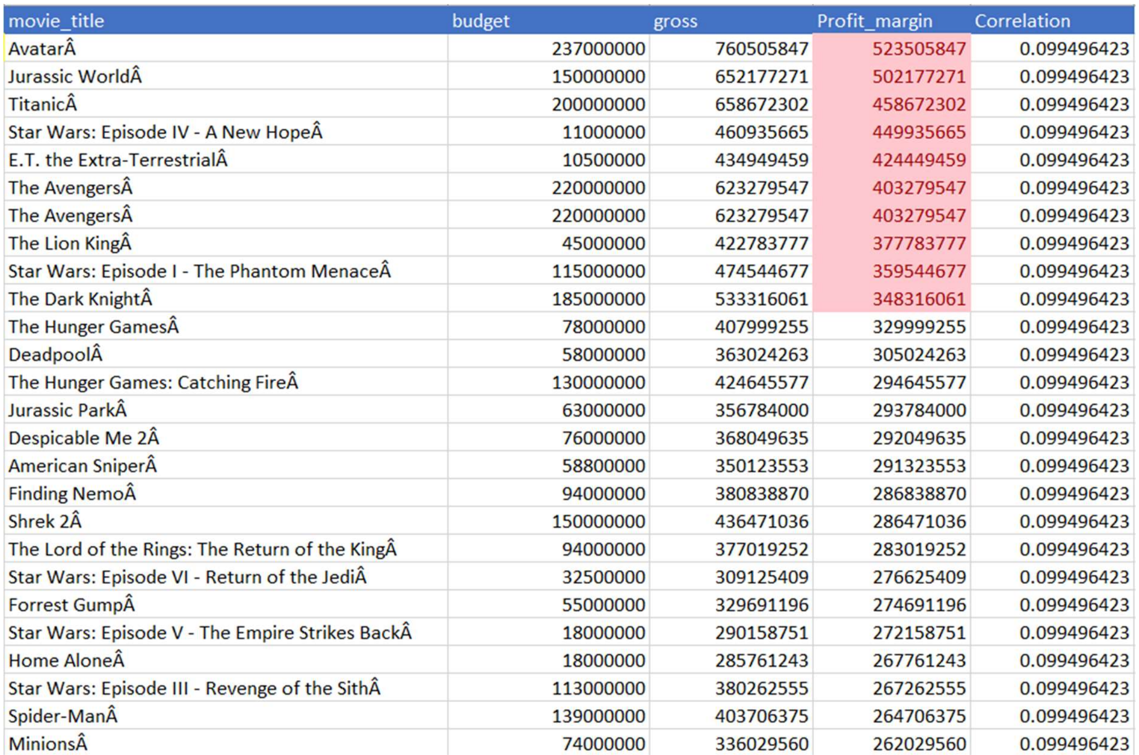
- Movie with the highest profit margin is Avatar



AvatarÂ		237000000	760505847	523505847	0.099496423



movie_title	budget	gross	Profit_margin	Correlation
AvatarÂ	237000000	760505847	523505847	0.099496423
Jurassic WorldÂ	150000000	652177271	502177271	0.099496423
TitanicÂ	200000000	658672302	458672302	0.099496423
Star Wars: Episode IV - A New HopeÂ	11000000	460935665	449935665	0.099496423
E.T. the Extra-TerrestrialÂ	10500000	434949459	424449459	0.099496423
The AvengersÂ	220000000	623279547	403279547	0.099496423
The AvengersÂ	220000000	623279547	403279547	0.099496423
The Lion KingÂ	45000000	422783777	377783777	0.099496423
Star Wars: Episode I - The Phantom MenaceÂ	115000000	474544677	359544677	0.099496423
The Dark KnightÂ	185000000	533316061	348316061	0.099496423



movie_title	budget	gross	Profit_margin	Correlation
AvatarÂ	237000000	760505847	523505847	0.099496423
Jurassic WorldÂ	150000000	652177271	502177271	0.099496423
TitanicÂ	200000000	658672302	458672302	0.099496423
Star Wars: Episode IV - A New HopeÂ	11000000	460935665	449935665	0.099496423
E.T. the Extra-TerrestrialÂ	10500000	434949459	424449459	0.099496423
The AvengersÂ	220000000	623279547	403279547	0.099496423
The AvengersÂ	220000000	623279547	403279547	0.099496423
The Lion KingÂ	45000000	422783777	377783777	0.099496423
Star Wars: Episode I - The Phantom MenaceÂ	115000000	474544677	359544677	0.099496423
The Dark KnightÂ	185000000	533316061	348316061	0.099496423
The Hunger GamesÂ	78000000	407999255	329999255	0.099496423
DeadpoolÂ	58000000	363024263	305024263	0.099496423
The Hunger Games: Catching FireÂ	130000000	424645577	294645577	0.099496423
Jurassic ParkÂ	63000000	356784000	293784000	0.099496423
Despicable Me 2Â	76000000	368049635	292049635	0.099496423
American SniperÂ	58800000	350123553	291323553	0.099496423
Finding NemoÂ	94000000	380838870	286838870	0.099496423
Shrek 2Â	150000000	436471036	286471036	0.099496423
The Lord of the Rings: The Return of the KingÂ	94000000	377019252	283019252	0.099496423
Star Wars: Episode VI - Return of the JediÂ	32500000	309125409	276625409	0.099496423
Forrest GumpÂ	55000000	329691196	274691196	0.099496423
Star Wars: Episode V - The Empire Strikes BackÂ	18000000	290158751	272158751	0.099496423
Home AloneÂ	18000000	285761243	267761243	0.099496423
Star Wars: Episode III - Revenge of the SithÂ	113000000	380262555	267262555	0.099496423
Spider-ManÂ	139000000	403706375	264706375	0.099496423
MinionsÂ	74000000	336029560	262029560	0.099496423

Insights

- Summarize major observations, such as correlations between IMDb ratings and variables like genre, director, budget, actors, release year, etc.
- Highlight any significant trends or patterns discovered during analysis.
- Discuss insights obtained from the 'Five Whys' approach and their implications.

Result

- Describe the insights gained and their potential impact on decision-making for movie stakeholders.
- Explain how the project contributes to understanding the factors influencing movie success on IMDb.

Drive Link

Excel link: https://drive.google.com/file/d/1JsYhlcBSrf8f0G5taOYzf1eAoar3HHQ-/view?usp=sharing