Project 05 - IMDB Movie Analysis

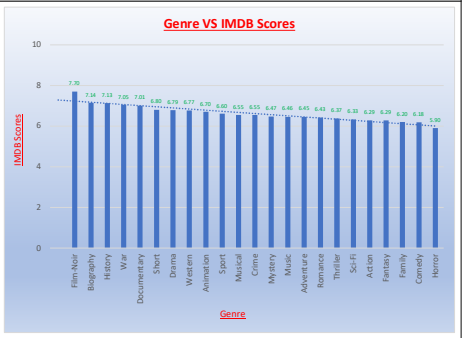
Data Cleaning

Before	After
Total Columns = 28	Total Columns = 7 Removed the unnecessary columns which are not useful for our analysis
Total Rows = 5044	Total Rows = 4922 Removed 122 duplicate rows
Total Rows = 4922 Column 'director_name' has 102 blank cells Column 'duration' has 15 blank cells Column 'language' has 12 blank cells Column 'budget' has 485 blank cells Columb 'gross' has 624 blank cells	Total Rows = 3787 Removed all rows that had blank cells

A. Movie Genre Analysis: 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.

Answer:

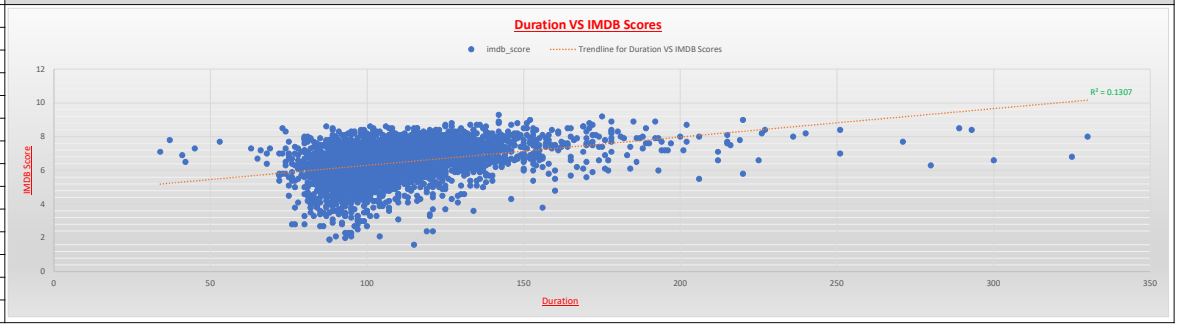
Genre	Count of Movies	Mean	Median	Mode	Range		Variance	Standard Deviation	Graphical Representation
					Min	Max			
Film-Noir	1	7.70	7.7	#N/A	7.7	7.7	#DIV/0!	#DIV/0!	
Biography	242	7.14	7.2	7	4.5	8.9	0.504237338	0.71009671	
History	152	7.13	7.2	7.7	5.5	8.9	0.451578947	0.671996241	
War	159	7.05	7.1	7.1	4.3	8.6	0.652386753	0.80770462	
Documentary	67	7.01	7.2	6.6	1.6	8.5	1.439855269	1.199939694	
Short	2	6.80	6.8	#N/A	6.5	7.1	0.18	0.424264069	
Drama	1911	6.79	6.9	6.7	2.1	9.3	0.793996652	0.891064898	
Western	58	6.77	6.8	6.8	4.1	8.9	0.997035693	0.998516746	
Animation	197	6.70	6.8	7.3	2.8	8.6	0.987295659	0.993627525	
Sport	147	6.60	6.8	7.2	2	8.4	1.09876526	1.048220043	
Musical	102	6.55	6.7	7.1	2.1	8.5	1.307672297	1.143535	
Crime	702	6.55	6.6	6.6	2.4	9.3	0.968463042	0.984105199	
Mystery	377	6.47	6.5	6.6	3.1	8.6	1.014838309	1.007391835	
Music	247	6.46	6.7	6.2	1.6	8.5	1.413359666	1.188848041	
Adventure	766	6.45	6.6	6.6	2.3	8.9	1.247524378	1.116926308	
Romance	866	6.43	6.5	6.5	2.1	8.5	0.938953731	0.968996249	
Thriller	1087	6.37	6.4	6.5	2.7	9	0.939112803	0.969078327	
Sci-Fi	484	6.33	6.4	7	1.9	8.8	1.362318841	1.16718415	
Action	935	6.29	6.3	6.6	2.1	9	1.078186788	1.038357736	
Fantasy	496	6.29	6.4	6.7	2.2	8.9	1.30054464	1.140414241	
Family	441	6.20	6.3	5.4	1.9	8.6	1.367909091	1.169576458	
Comedy	1492	6.18	6.3	6.3	1.9	8.8	1.081431552	1.039919012	
Horror	379	5.90	5.9	6.2	2.3	8.6	0.982127152	0.991023285	



B. Movie Duration Analysis:	Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.

Answer:	For Movie Durations	
	Mean	109.808505
	Median	105
	Standard Deviation	22.763201

Graphical Representation

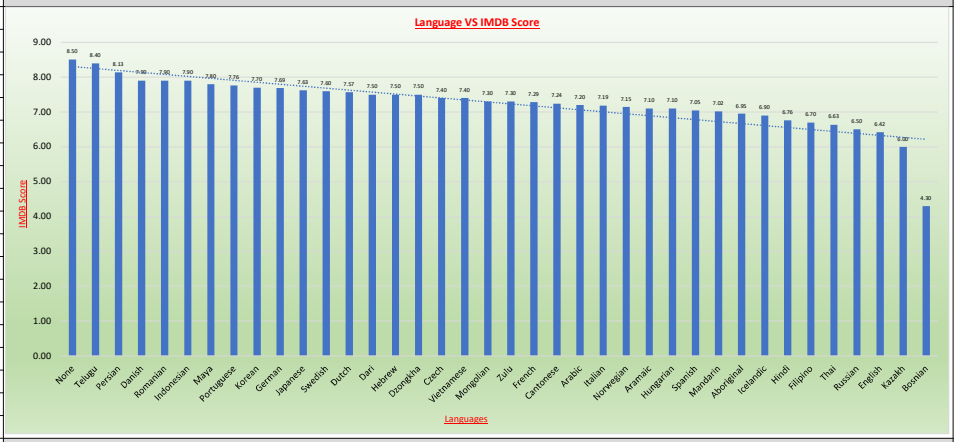


C. Language Analysis: Determine the most common languages used in movies and analyze their impact on the IMDB score using descriptive statistics.

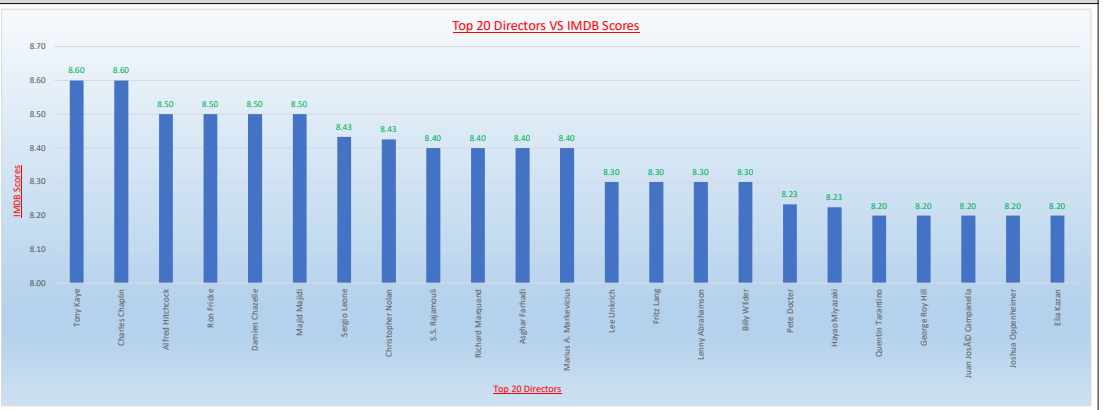
Answer:

Languages	Count of Movies	Mean	Median	Standard Deviation
None	1	8.50	8.4	#DIV/0!
Telugu	1	8.40	8.4	0.550757055
Persian	3	8.13	8.1	0.529150262
Danish	3	7.90	7.9	#DIV/0!
Romanian	1	7.90	7.9	0.424264069
Indonesian	2	7.90	7.8	#DIV/0!
Maya	1	7.80	8	0.978774744
Portuguese	5	7.76	7.7	0.570087713
Korean	5	7.70	7.7	0.640912811
German	13	7.69	7.8	0.899621132
Japanese	12	7.63	7.6	#DIV/0!
Swedish	1	7.60	7.8	0.404145188
Dutch	3	7.57	7.5	0.141421356
Dari	2	7.50	7.3	0.435889894
Hebrew	3	7.50	7.5	#DIV/0!
Dzongkha	1	7.50	7.4	#DIV/0!
Czech	1	7.40	7.4	#DIV/0!
Vietnamese	1	7.40	7.3	#DIV/0!
Mongolian	1	7.30	7.3	#DIV/0!
Zulu	1	7.30	7.2	0.561328861
French	37	7.29	7.3	0.440575922
Cantonese	8	7.24	7.2	#DIV/0!
Arabic	1	7.20	7	1.155318962
Italian	7	7.19	7.3	0.574456265
Norwegian	4	7.15	7.1	#DIV/0!
Aramaic	1	7.10	7.1	#DIV/0!
Hungarian	1	7.10	7.15	0.826196103
Spanish	26	7.05	7.25	0.765786244
Mandarin	14	7.02	6.95	0.777817459
Aboriginal	2	6.95	6.9	#DIV/0!
Icelandic	1	6.90	7.05	1.111755369
Hindi	10	6.76	6.7	#DIV/0!
Filipino	1	6.70	6.6	0.450924975
Thai	3	6.63	6.5	#DIV/0!
Russian	1	6.50	6.5	1.052498903
English	3606	6.42	6	#DIV/0!
Kazakh	1	6.00	4.3	#DIV/0!
Bosnian	1	4.30	#NUM!	#DIV/0!

Graphical Representation



D. Director Analysis:				Identify the top directors based on their average IMDB score and analyze their contribution to the success of movies using percentile calculations.				
Answer:				Director Name	Average IMDB Score	Rank	PercentRank	Graphical Representation
				Tony Kaye	8.60	1	0.999	
				Charles Chaplin	8.60	1	0.999	
				Alfred Hitchcock	8.50	3	0.997	
				Ron Fricke	8.50	3	0.997	
				Damien Chazelle	8.50	3	0.997	
				Majid Majidi	8.50	3	0.997	
				Sergio Leone	8.43	7	0.996	
				Christopher Nolan	8.43	8	0.996	
				S.S. Rajamouli	8.40	9	0.993	
				Richard Marquand	8.40	9	0.993	
				Asghar Farhadi	8.40	9	0.993	
				Marius A. Markevicius	8.40	9	0.993	
				Lee Unkrich	8.30	13	0.991	
				Fritz Lang	8.30	13	0.991	
				Lenny Abrahamson	8.30	13	0.991	
				Billy Wilder	8.30	13	0.991	
				Pete Docter	8.23	17	0.99	
				Hayao Miyazaki	8.23	18	0.99	
				Quentin Tarantino	8.20	19	0.989	
				George Roy Hill	8.20	20	0.987	
				Juan JosÃ© Campanella	8.20	20	0.987	
				Joshua Oppenheimer	8.20	20	0.987	
				Elia Kazan	8.20	20	0.987	



The functions used and all the directors list and their average are shown in a separate sheet (D. Director Analysis (Detail))

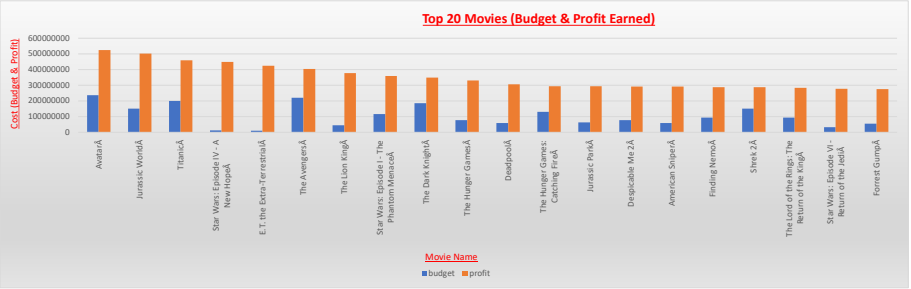
E. Budget Analysis: | Analyze the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.

Answer:

movie_title	budget	gross	profit	Rank
AvatarÂ	237000000	760505847	523505847	1
Jurassic WorldÂ	150000000	652177271	502177271	2
TitanicÂ	200000000	658672302	458672302	3
Star Wars: Episode IV - A New HopeÂ	11000000	460935665	449935665	4
E.T. the Extra-TerrestrialÂ	10500000	434949459	424449459	5
The AvengersÂ	220000000	623279547	403279547	6
The Lion KingÂ	45000000	422783777	377783777	7
Star Wars: Episode I - The Phantom MenaceÂ	115000000	474544677	359544677	8
The Dark KnightÂ	185000000	533316061	348316061	9
The Hunger GamesÂ	78000000	407999255	329999255	10
DeadpoolÂ	58000000	363024263	305024263	11
The Hunger Games: Catching FireÂ	130000000	424645577	294645577	12
Jurassic ParkÂ	63000000	356784000	293784000	13
Despicable Me 2Â	76000000	368049635	292049635	14
American SniperÂ	58800000	350123553	291323553	15
Finding NemoÂ	94000000	380838870	286838870	16
Shrek 2Â	150000000	436471036	286471036	17
The Lord of the Rings: The Return of the KingÂ	94000000	377019252	283019252	18
Star Wars: Episode VI - Return of the JediÂ	32500000	309125409	276625409	19
Forrest GumpÂ	55000000	329691196	274691196	20

Max Profit	523505847
Movie With Highest Profit	AvatarÂ
Correlation Coefficient Between Movie Budgets and Gross Earnings	0.101645695

Graphical Representation



The functions used and all the directors list and their average are shown in a separate sheet (E. Budget Analysis (Detail))

FINAL SUMMARY			
Sr No	**Question**	**Answer**	
1	Project Description	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. This step involves preprocessing the data to make it suitable for analysis. It includes handling missing values, removing duplicates, converting data types if necessary, and possibly feature engineering. Here, you'll explore the data to understand the relationships between different variables. You might look at the correlation between movie ratings and other factors like genre, director, budget, etc. You might also want to consider the year of release, the actors involved, and other relevant factors.	
2	Approach	1. Understood the dataset provided 2. Have done the data cleaning as described in the data cleaning table 3. Used appropriate functions and formulas to get the requried answers for each questions	
2(A)	Movie Genre Analysis	Text to column & Removing duplicates	Extracted the genre list from the given data set
		=COUNTIF(Table1[genres],"**"&F14&"**")	Applied "COUNTIF" function to get the movies count for each genre
		=AVERAGEIF(Table1[genres],"**"&F14&"**",Table1[imdb_score])	Applied "AVERAGEIF" function to get the mean IMDB score for each genre
		{=MEDIAN(IF(ISNUMBER(SEARCH("**"&F14&"**",Table1[genres])),Table1[imdb_score]))}	Applied "MEDIAN" function to get the median IMDB score for each genre
		{=MODE(IF(ISNUMBER(SEARCH("**"&F14&"**",Table1[genres])),Table1[imdb_score]))}	Applied "MODE" function to get the mode IMDB score for each genre
		=MINIFS(Table1[imdb_score],Table1[genres],"**"&F14&"**")	Applied "MINIFS" function to get the minimum IMDB score for each genre
		=MAXIFS(Table1[imdb_score],Table1[genres],"**"&F14&"**")	Applied "MAXIFS" function to get the maximum IMDB score for each genre
		{=VAR(IF(ISNUMBER(SEARCH("**"&F14&"**",Table1[genres])),Table1[imdb_score]))}	Applied "VAR" function to get the variance value of IMDB scores for each genre
		{=STDEV(IF(ISNUMBER(SEARCH("**"&F14&"**",Table1[genres])),Table1[imdb_score]))}	Applied "STDEV" function to get the standard deviation value of IMDB scores for each genre
		Graphical Representation	Visualized the Genres and their IMDB Scores
2(B)	Movie Duration Analysis	=AVERAGE(Table1[duration])	Applied "AVERAGE" function to get the average duration
		=MEDIAN(Table1[duration])	Applied "MEDIAN" function to get the median value of duration
		=STDEVA(Table1[duration])	Applied "STDEV" function to get the standard deviation value of duration
		Graphical Representation	Visualized the Duration and their IMDB scores in a scatter plot and added a trendline to understand the relationship between them

2(C)	Language Analysis	Removed duplicates	Extracted the language list from given data set
		=COUNTIF(Table1[language],F57)	Applied "COUNTIF" function to get the movies count for each language
		=AVERAGEIF(Table1[language],F57,Table1[imdb_score])	Applied "AVERAGEIF" function to get the mean IMDB score for each language
		{=MEDIAN(IF(Table1[language]=F58,Table1[imdb_score]))}	Applied "MEDIAN" function to get the median IMDB score for each language
		{=STDEV(IF(Table1[language]=F58,Table1[imdb_score]))}	Applied "STDEV" function to get the standard deviation value of IMDB scores for each language
		Graphical Representation	Visualized the languages and their IMDB Scores
2(D)	Director Analysis	Removed duplicates	Extracted the directors list from given data set
		=AVERAGEIF(Table1[director_name],C4,Table1[imdb_score])	Applied "AVERAGEIF" function to get the mean IMDB score for each director
		=RANK(D4,$D$4:$D$1754)	Applied "RANK" function to get the rank for each director based on their average IMDB score and to know the top 20 directors
		=PERCENTRANK($D$4:$D$1754,D4)	Applied "PERCENTRANK" function to get the percentage of rank for each director
		Graphical Representation	Visualized the top 20 directors and their IMDB scores
2(E)	Budget Analysis	Profit calculation	Calculated profit for each movies (gross earnings - budget)
		=RANK(E3,$E$3:$E$3788)	Applied "RANK" function to get the rank for each movies based on their profit value and to know the top 20 movies
		Movie with highest profit	Found the movie with highest profit after applying the rank
		Correlation coefficient	Calculated correlation coefficient between the movie budhets and their gross earnings
		Graphical Representation	Visualized the top 20 movies with their budget and profits
3	Tech-Stack Used	Microsoft Office 2019	
4	Insights	In this projects I have used different excel functions, formulas, charts to extract the answers for all the questions which has helped me to improve the way of thinking while working on excel and selecting appropriate functions according to the questions.	
5	Result	This project has helped me to improve my skills on excel functions and charts.	