A

PROJECT

ON

IMDB MOVIE ANALYSIS

By – Vansh

Mail ID – aroravansh11@gmail.com

Tasks

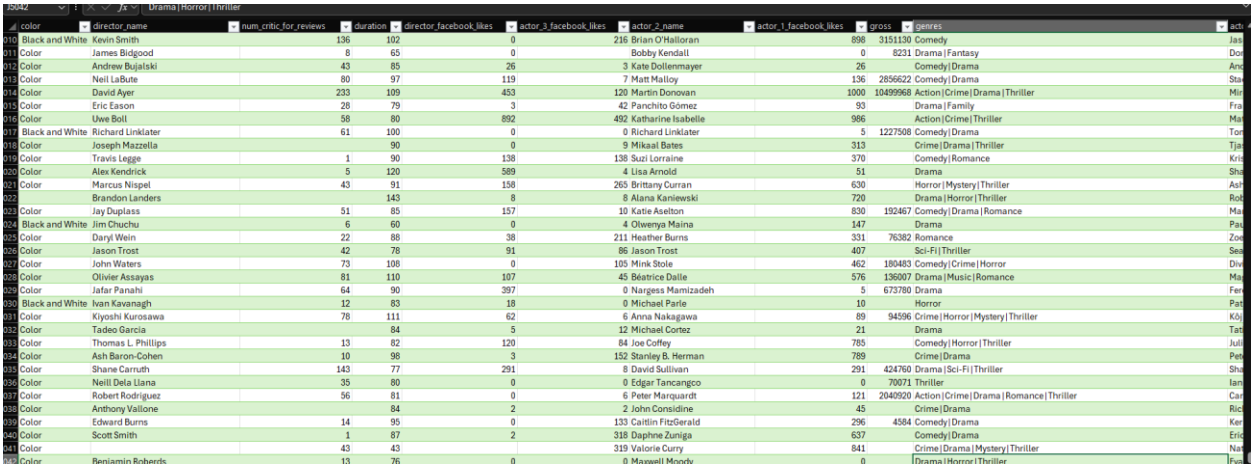
A) Movie Genre Analysis: Analyze the distribution of movie genres and their impact on the IMDB score.

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.

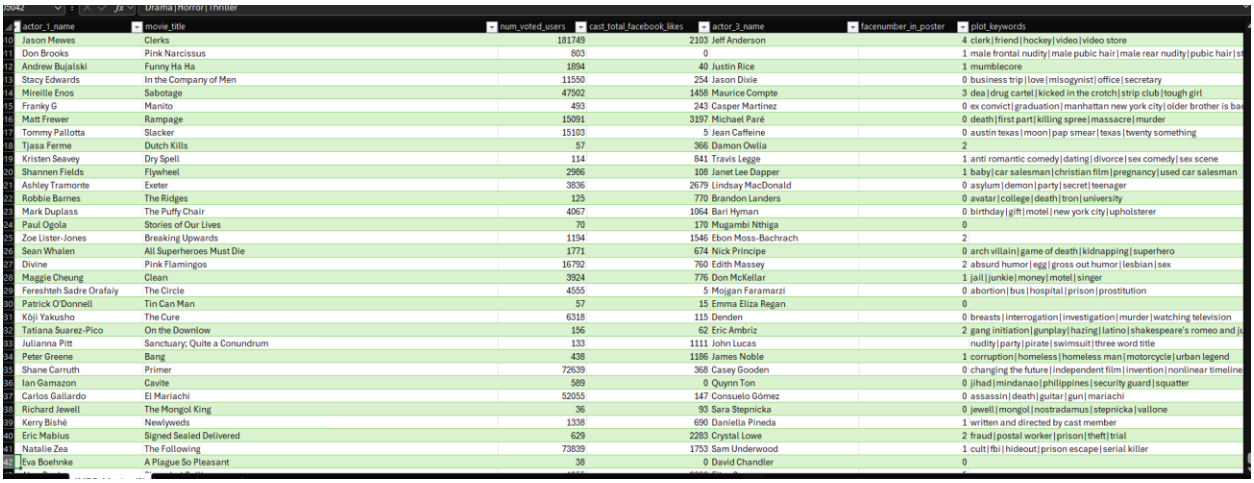
Solution:

Excel file Link -- [..\IMDB excel files\IMDB_Movies TASK A.xlsx]

Provided Excel Dataset:



color	director_name	num_critic_for_reviews	duration	director_facebook_likes	actor_3_facebook_likes	actor_2_name	actor_1_facebook_likes	gross	genres
Black and White	Kevin Smith	136	102	0	216	Brian O'Halloran	898	3151130	Comedy
Color	James Bidgood	8	65	0		Bobby Kendall	0	8231	Drama|Fantasy
Color	Andrew Bujalski	43	85	26	3	Kate Dollenmayer	26		Comedy|Drama
Color	Neil LaBute	80	97	119	7	Matt Malloy	136	2856622	Comedy|Drama
Color	David Ayer	233	109	453	120	Martin Donovan	1000	10499968	Action|Crime|Drama|Thriller
Color	Eric Eason	28	79	3	42	Pancho Gómez	93		Drama|Family
Color	Uwe Boll	58	80	892	492	Katharine Isabelle	986		Action|Crime|Thriller
Black and White	Richard Linklater	61	100	0	0	Richard Linklater	5	1227508	Comedy|Drama
Color	Joseph Mazzella		90	0	9	Mikaal Bates	313		Crime|Drama|Thriller
Color	Travis Legge	1	90	138	138	Suzi Lorraine	370		Comedy|Romance
Color	Alex Kendrick	5	120	589	4	Lisa Arnold	51		Drama
Color	Marcus Nispel	43	91	158	265	Brittany Curran	630		Horror|Mystery|Thriller
	Brandon Landers		143	8	8	Alana Kaniewski	720		Drama|Horror|Thriller
Color	Jay Duplass	51	85	157	10	Katie Aselton	830	192467	Comedy|Drama|Romance
Black and White	Jim Chuchu	6	60	0	4	Olwenya Maina	147		Drama
Color	Daryl Wein	22	88	38	211	Heather Burns	331	76382	Romance
Color	Jason Trost	42	78	91	86	Jason Trost	407		Sci-Fi|Thriller
Color	John Waters	73	108	0	105	Mink Stole	462	180483	Comedy|Crime|Horror
Color	Olivier Assayas	81	110	107	45	Béatrice Dalle	576	136007	Drama|Music|Romance
Color	Jafar Panahi	64	90	397	0	Nargess Mamizadeh	5	673780	Drama
Black and White	Ivan Kavanagh	12	83	18	0	Michael Parle	10		Horror
Color	Kiyoshi Kurosawa	78	111	62	6	Anna Nakagawa	89	94596	Crime|Horror|Mystery|Thriller
Color	Tadeo Garcia		84	5	12	Michael Cortez	21		Drama
Color	Thomas L. Phillips	13	82	120	84	Joe Coffey	785		Comedy|Horror|Thriller
Color	Ash Baron-Cohen	10	98	3	152	Stanley B. Herman	789		Crime|Drama
Color	Shane Carruth	143	77	291	8	David Sullivan	291	424760	Drama|Sci-Fi|Thriller
Color	Neill Dela Llana	35	80	0	0	Edgar Tancangco	0	70071	Thriller
Color	Robert Rodriguez	56	81	0	6	Peter Marquardt	121	2040920	Action|Crime|Drama|Romance|Thriller
Color	Anthony Vallone		84	2	2	John Considine	45		Crime|Drama
Color	Edward Burns	14	95	0	133	Caitlin FitzGerald	296	4584	Comedy|Drama
Color	Scott Smith	1	87	2	318	Daphne Zuniga	637		Comedy|Drama
Color		43	43		319	Valorie Curry	841		Crime|Drama|Mystery|Thriller
Color	Benjamin Roberds	13	76	0	0	Maxwell Moody	0		Drama|Horror|Thriller



actor_1_name	movie_title	num_voted_users	cast_total_facebook_likes	actor_3_name	facenumber_in_poster	plot_keywords
Jason Mewes	Clerks	181749	2103	Jeff Anderson	4	clerk|friend|hockey|video|video store
Don Brooks	Pink Narcissus	803	0		1	male frontal nudity|male pubic hair|male rear nudity|pubic hair|st
Andrew Bujalski	Funny Ha Ha	1894	40	Justin Rice	1	mumblecore
Stacy Edwards	In the Company of Men	11550	254	Jason Dixie	0	business trip|love|misogynist|office|secretary
Mireille Enos	Sabotage	47502	1458	Maurice Compte	3	dea|drug cartel|kicked in the crotch|strip club|tough girl
Franky G	Manito	493	243	Casper Martinez	0	ex convict|graduation|manhattan new york city|older brother is ba
Matt Frewer	Rampage	15091	3197	Michael Paré	0	death|first part|killing spree|massacre|murder
Tommy Pallotta	Slacker	15103	5	Jean Caffeine	0	austin texas|moon|pap smear|texas|twenty something
Tjasa Ferme	Dutch Kills	57	366	Damon Owlia	2	
Kristen Seavey	Dry Spell	114	841	Travis Legge	1	anti romantic comedy|dating|divorce|sex comedy|sex scene
Shannen Fields	Flywheel	2986	108	Janet Lee Dapper	1	baby|car salesman|christian film|pregnancy|used car salesman
Ashley Tramonte	Exeter	3836	2679	Lindsay MacDonald	0	asylum|demon|party|secret|teenager
Robbie Barnes	The Ridges	125	770	Brandon Landers	0	avatar|college|death|tron|university
Mark Duplass	The Puffy Chair	4067	1064	Bari Hyman	0	birthday|gift|motel|new york city|upholsterer
Paul Ogola	Stories of Our Lives	70	170	Mugambi Nthiga	0	
Zoe Lister-Jones	Breaking Upwards	1194	1546	Ebon Moss-Bachrach	2	
Sean Whalen	All Superheroes Must Die	1771	674	Nick Principe	0	arch villain|game of death|kidnapping|superhero
Divine	Pink Flamingos	16792	760	Edith Massey	2	absurd humor|egg|gross out humor|lesbian|sex
Maggie Cheung	Clean	3924	776	Don McKellar	1	jail|junkie|money|motel|singer
Fereshteh Sadre Orafaiy	The Circle	4555	5	Mojgan Faramarzi	0	abortion|bus|hospital|prison|prostitution
Patrick O'Donnell	Tin Can Man	57	15	Emma Eliza Regan	0	
Kôji Yakusho	The Cure	6318	115	Denden	0	breasts|interrogation|investigation|murder|watching television
Tatiana Suarez-Pico	On the Downlow	156	62	Eric Ambriz	2	gang initiation|gunplay|hazing|latino|shakespeare's romeo and ju
Julianna Pitt	Sanctuary: Quite a Conundrum	133	1111	John Lucas		nudity|party|pirate|swimsuit|three word title
Peter Greene	Bang	438	1186	James Noble	1	corruption|homeless|homeless man|motorcycle|urban legend
Shane Carruth	Primer	72639	368	Casey Gooden	0	changing the future|independent film|invention|nonlinear timeline
Ian Gamazon	Cavite	589	0	Quynn Ton	0	jihad|mindanao|philippines|security guard|squatter
Carlos Gallardo	El Mariachi	52055	147	Consuelo Gómez	0	assassin|death|guitar|gun|mariachi
Richard Jewell	The Mongol King	36	93	Sara Stepnicka	0	jewell|mongol|nostradamus|stepnicka|vallone
Kerry Bishé	Newlyweds	1338	690	Daniella Pineda	1	written and directed by cast member
Eric Mabius	Signed Sealed Delivered	629	2283	Crystal Lowe	2	fraud|postal worker|prison|theft|trial
Natalie Zea	The Following	73839	1753	Sam Underwood	1	cult|fbi|hideout|prison escape|serial killer
Eva Boehnke	A Plague So Pleasant	38	0	David Chandler	0	

movie_imdb_link	num_user_for_reviews	language	country	content_rating	budget	title_year	actor_2_facebook_likes	imdb_score	aspect_ratio	movie_facebook_likes
http://www.imdb.com/title/tt0109445/?ref_=fn_tt_tt_1	615	English	USA	R	230000	1994	657	7.8	1.37	0
http://www.imdb.com/title/tt0067580/?ref_=fn_tt_tt_1	16	English	USA	Not Rated	27000	1971	0	6.7	1.37	85
http://www.imdb.com/title/tt0327753/?ref_=fn_tt_tt_1	61	English	USA			2002	6	6.4	1.37	108
http://www.imdb.com/title/tt0119361/?ref_=fn_tt_tt_1	197	English	Canada	R	25000	1997	108	7.3	1.85	489
http://www.imdb.com/title/tt1742334/?ref_=fn_tt_tt_1	212	English	USA	R	35000000	2014	206	5.7	1.85	10000
http://www.imdb.com/title/tt0298050/?ref_=fn_tt_tt_1	21	English	USA		24000	2002	46	7	1.78	61
http://www.imdb.com/title/tt1337057/?ref_=fn_tt_tt_1	129	English	Canada	R		2009	918	6.3	2.35	0
http://www.imdb.com/title/tt0102943/?ref_=fn_tt_tt_1	80	English	USA	R	23000	1991	0	7.1	1.37	2000
http://www.imdb.com/title/tt2759066/?ref_=fn_tt_tt_1	2	English	USA		25000	2015	25	4.8		33
http://www.imdb.com/title/tt2375036/?ref_=fn_tt_tt_1	3	English	USA		22000	2013	184	3.3	1.78	200
http://www.imdb.com/title/tt0425027/?ref_=fn_tt_tt_1	49	English	USA		20000	2003	49	6.9	1.85	725
http://www.imdb.com/title/tt1945044/?ref_=fn_tt_tt_1	33	English	USA	R		2015	512	4.6	1.85	0
http://www.imdb.com/title/tt1781935/?ref_=fn_tt_tt_1	8	English	USA		17350	2011	19	3		33
http://www.imdb.com/title/tt0436689/?ref_=fn_tt_tt_1	71	English	USA	R	15000	2005	224	6.6		297
http://www.imdb.com/title/tt3973612/?ref_=fn_tt_tt_1	1	Swahili	Kenya		15000	2014	19	7.4		45
http://www.imdb.com/title/tt1247644/?ref_=fn_tt_tt_1	8	English	USA		15000	2009	212	6.2	2.35	324
http://www.imdb.com/title/tt1836212/?ref_=fn_tt_tt_1	35	English	USA	Unrated	20000	2011	91	4	2.35	835
http://www.imdb.com/title/tt0069089/?ref_=fn_tt_tt_1	183	English	USA	NC-17	10000	1972	143	6.1	1.37	0
http://www.imdb.com/title/tt0388838/?ref_=fn_tt_tt_1	39	French	France	R	4500	2004	133	6.9	2.35	171
http://www.imdb.com/title/tt0255094/?ref_=fn_tt_tt_1	26	Persian	Iran	Not Rated	10000	2000	0	7.5	1.85	697
http://www.imdb.com/title/tt1235811/?ref_=fn_tt_tt_1	1	English	Ireland		10000	2007	5	6.7	1.33	105
http://www.imdb.com/title/tt0123948/?ref_=fn_tt_tt_1	50	Japanese	Japan		1000000	1997	13	7.4	1.85	817
http://www.imdb.com/title/tt0390323/?ref_=fn_tt_tt_1	3	English	USA			2004	20	6.1		22
http://www.imdb.com/title/tt2049518/?ref_=fn_tt_tt_1	8	English	USA		200000	2012	98	5.4	16	424
http://www.imdb.com/title/tt0109266/?ref_=fn_tt_tt_1	14	English	USA			1995	194	6.4		20
http://www.imdb.com/title/tt0390384/?ref_=fn_tt_tt_1	371	English	USA	PG-13	7000	2004	45	7	1.85	19000
http://www.imdb.com/title/tt0428303/?ref_=fn_tt_tt_1	35	English	Philippines	Not Rated	7000	2005	0	6.3		74
http://www.imdb.com/title/tt0104815/?ref_=fn_tt_tt_1	130	Spanish	USA	R	7000	1992	20	6.9	1.37	0
http://www.imdb.com/title/tt0430371/?ref_=fn_tt_tt_1	1	English	USA	PG-13	3250	2005	44	7.8		4
http://www.imdb.com/title/tt1880418/?ref_=fn_tt_tt_1	14	English	USA	Not Rated	9000	2011	205	6.4		413
http://www.imdb.com/title/tt3000844/?ref_=fn_tt_tt_1	6	English	Canada			2013	470	7.7		84
http://www.imdb.com/title/tt2071645/?ref_=fn_tt_tt_1	359	English	USA	TV-14			593	7.5	16	32000
http://www.imdb.com/title/tt2107644/?ref_=fn_tt_tt_1	3	English	USA		1400	2013	0	6.3		16

et1 Sheet2

Cleaned Dataset:

E1 Count_of_genres

	B	C	D	E	F	G	H	I	J	K
1	imdb_score	genres	Unique_genres	Count_of_genres	Mean_of_uni_genres	Median_of_uni_genre	Mode_of_uni_genre	Range_of_uni_genre	Variance_of_uni_genre	Stdev_of_uni_genre
2	7.9	Action|Adventure|Fantasy|Sci-Fi	Action	1153	6.24	6.3	6.1	7.4	1.25	1.12
3	7.1	Action|Adventure|Fantasy	Documentary	121	7.18	7.4	7.5	7.1	1.12	1.06
4	6.8	Action|Adventure|Thriller	Adventure	923	6.44	6.6	6.7	7	1.28	1.13
5	8.5	Action|Thriller	Drama	2594	6.76	6.9	7.2	7.3	0.92	0.96
6	7.1	Documentary	Animation	242	6.58	6.7	6.7	6.9	1.3	1.14
7	6.6	Action|Adventure|Sci-Fi	Comedy	1872	6.2	6.3	6.7	7.8	1.19	1.09
8	6.2	Action|Adventure|Romance	Mystery	500	6.49	6.6	6.6	6.4	1.19	1.09
9	7.8	Adventure|Animation|Comedy|Family|Fantasy|Musical|Romance	Fantasy	610	6.31	6.4	6.7	7.2	1.35	1.16
10	7.5	Action|Adventure|Sci-Fi	Crime	889	6.56	6.6	6.6	6.9	1.05	1.03
11	7.5	Adventure|Family|Fantasy|Mystery	Biography	293	7.15	7.2	7	4.4	0.52	0.72
12	6.9	Action|Adventure|Sci-Fi	Sci-Fi	616	6.28	6.4	6.7	6.9	1.47	1.21
13	6.1	Action|Adventure|Sci-Fi	Horror	565	5.84	5.9	6.2	6.5	1.28	1.13
14	6.7	Action|Adventure	Romance	1107	6.45	6.5	6.5	6.5	0.99	1
15	7.3	Action|Adventure|Fantasy	Thriller	1411	6.31	6.4	6.1	6.8	1.11	1.05
16	6.5	Action|Adventure|Western	Game-Show	1	2.9	2.9	No Mode	0	N/A	N/A
17	7.2	Action|Adventure|Fantasy|Sci-Fi	Family	546	6.25	6.4	6.7	7	1.44	1.2
18	6.6	Action|Adventure|Family|Fantasy	Music	326	6.46	6.7	7.1	6.9	1.44	1.2
19	8.1	Action|Adventure|Sci-Fi	Western	97	6.69	6.8	6.5	5.1	1.09	1.04
20	6.7	Action|Adventure|Fantasy	Musical	132	6.51	6.7	7	6.4	1.5	1.23
21	6.8	Action|Adventure|Comedy|Family|Fantasy|Sci-Fi	Film-Noir	6	7.63	7.65	No Mode	1.1	0.19	0.43
22	7.5	Adventure|Fantasy	History	207	7.08	7.2	7.5	6.9	0.79	0.89
23	7	Action|Adventure|Fantasy	War	213	7.07	7.1	7.1	5.9	0.77	0.87
24	6.7	Action|Adventure|Drama|History	Sport	182	6.61	6.8	7.2	6.7	1.21	1.1
25	7.9	Adventure|Fantasy	Reality-TV	2	4.75	4.75	No Mode	3.7	6.85	2.62
26	6.1	Adventure|Family|Fantasy	Short	5	6.38	6.5	No Mode	1.9	0.56	0.75
27	7.2	Action|Adventure|Drama|Romance	News	3	7.53	7.4	No Mode	1	0.26	0.51
28	7.7	Drama|Romance								
29	8.2	Action|Adventure|Sci-Fi								
30	5.9	Action|Adventure|Sci-Fi|Thriller	total_unique_movies							
31	7	Action|Adventure|Sci-Fi|Thriller	4917							
32	7.8	Action|Adventure|Thriller								
33	7.3	Action|Adventure|Fantasy|Romance								
34	7.2	Action|Adventure|Sci-Fi								
35	6.5	Adventure|Family|Fantasy								
36	6.8	Action|Adventure|Fantasy|Sci-Fi|Thriller								
37	7.3	Adventure|Animation|Comedy|Family|Fantasy								
38	6	Action|Adventure|Sci-Fi								

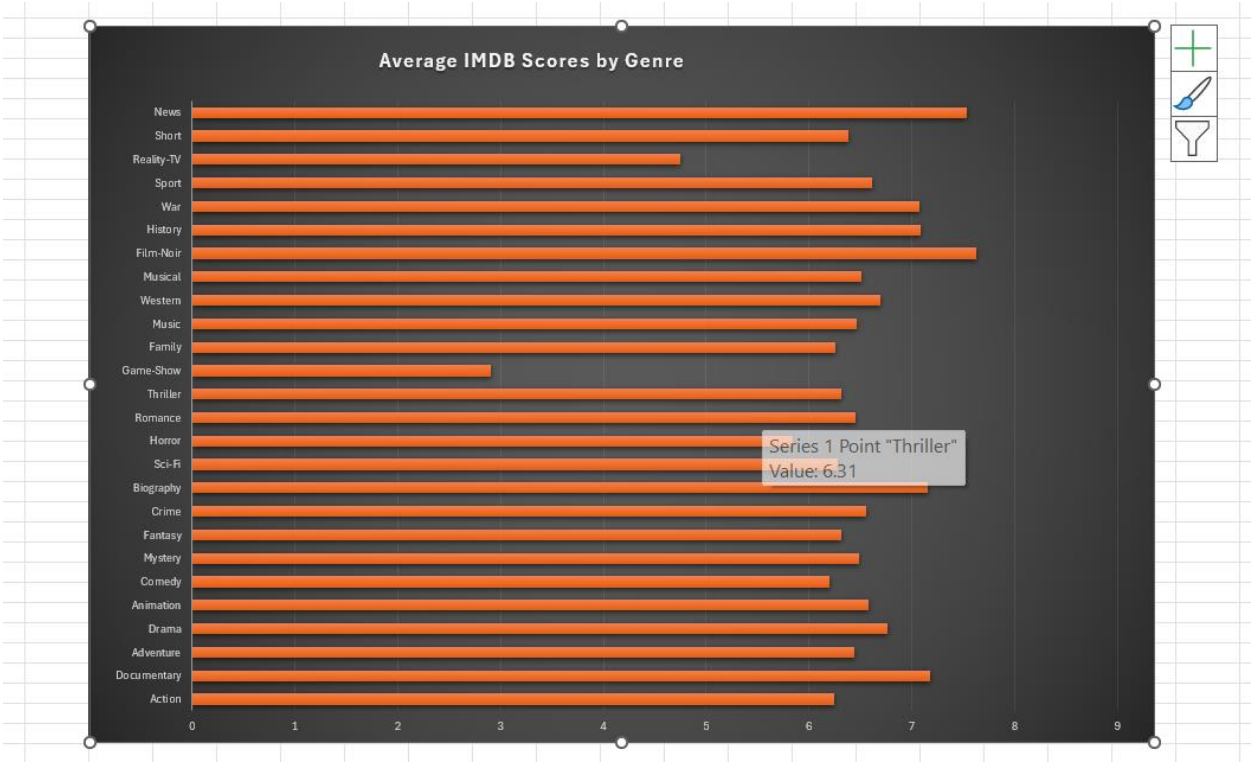
Explanation:

- Firstly, as the provided dataset was too big and messy to see I **cleaned** the data according to use of that data in my task, in this I selected **genres column**, **IMDB_score** column and copied these columns to another sheet.
- Then as genre column was having many genres in a single cell so I used **text to columns** feature which is present under **data** tab to separate the genres then I uniquely

defined all the genres in a different column and removed all the duplicates using "**remove duplicates**" feature present in "**data**" tab

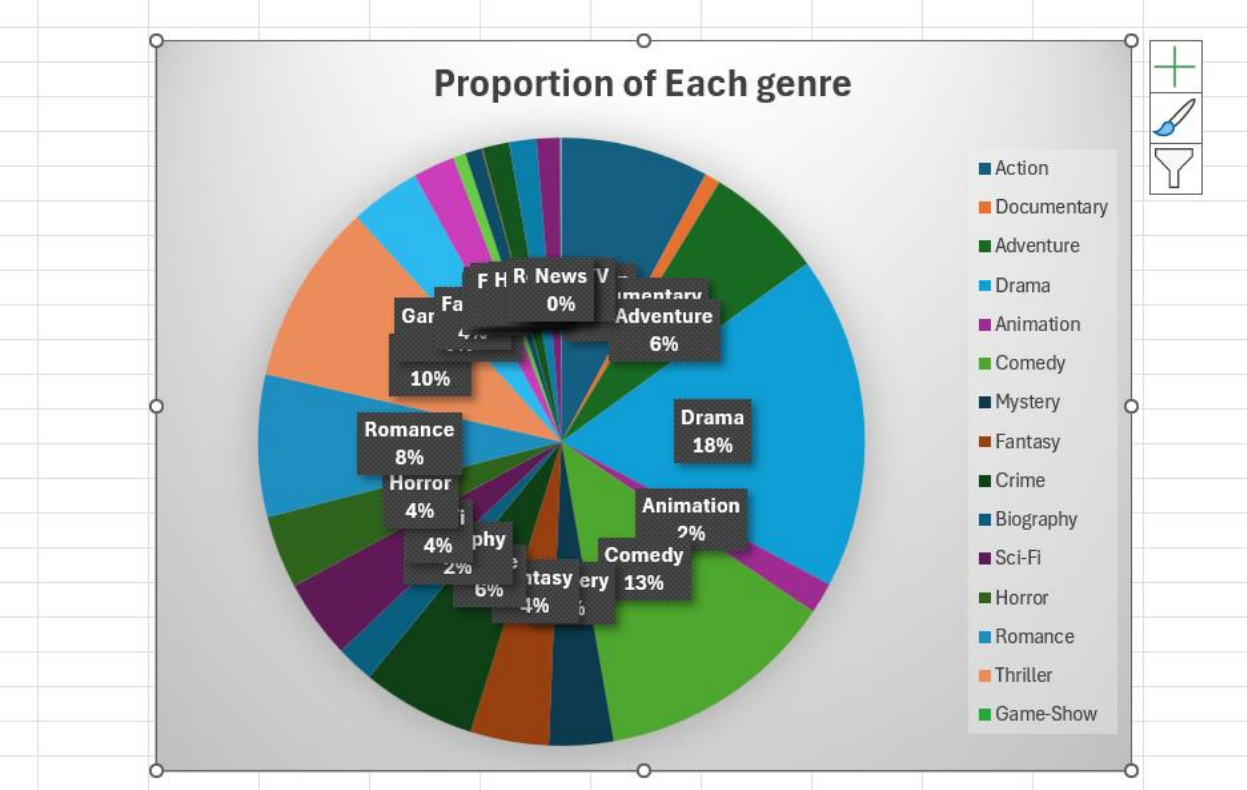
- After this I used **Countif** function to count how many times each genre appears in the dataset which I represented in column "count_of_genres"
- Then I calculated Descriptive statistics for IMDB Scores which include **mean, median, mode, range, variance, standard** deviation using built-in excel functions.
- Now after calculations I **visualized** the impact of IMDB scores using excel charts, etc.

Showing Impact of Genre on Movie Ratings using charts:

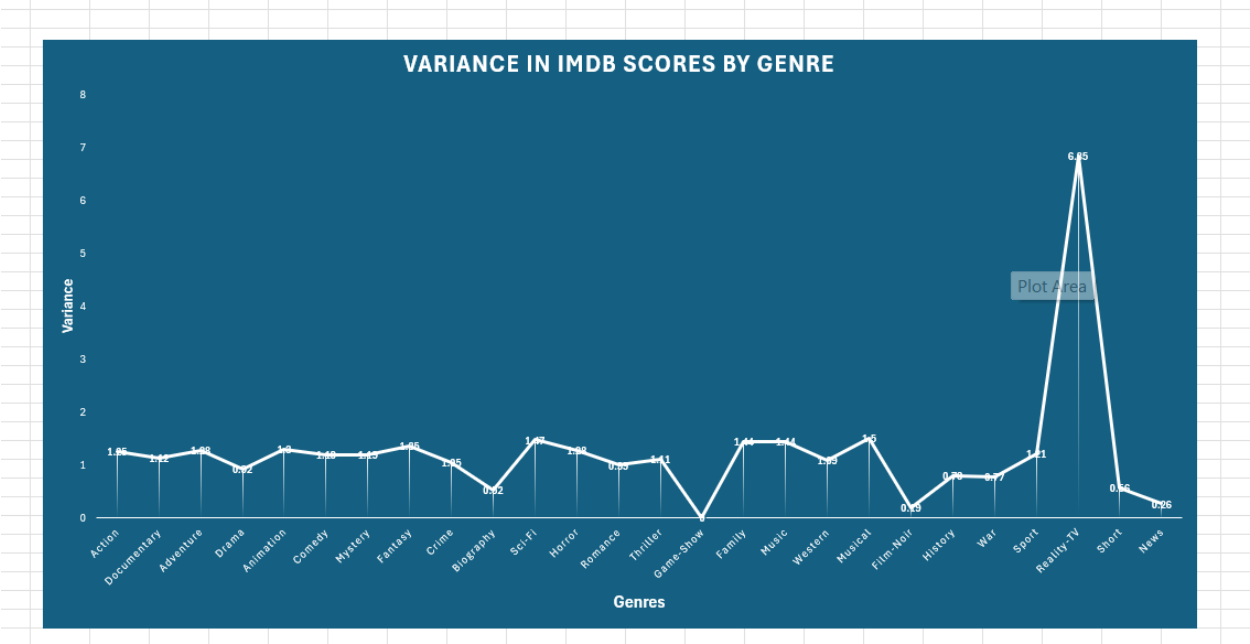


This Bar chart I used for comparing which genres have higher average ratings and which have lower. It consists of genres presented on **X-axis** and Mean Scores presented on **Y-axis**

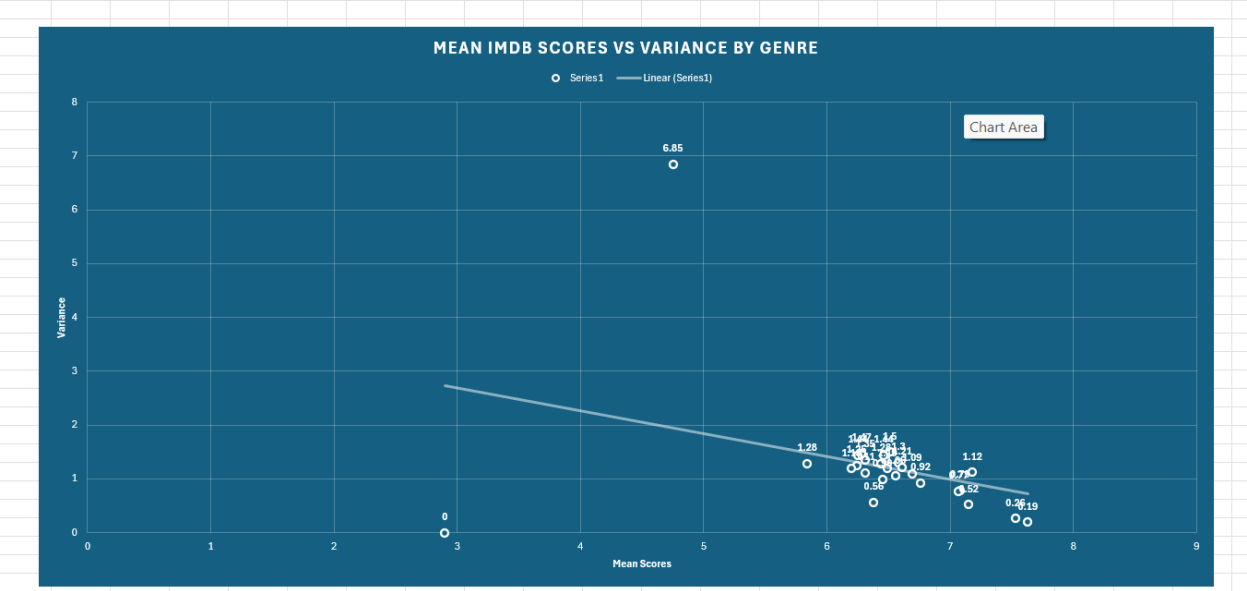
- Here **Film-Noir** genre is having **Highest Rating**.
- And **Game Show** genre is having the **Lowest Rating**.



This **Pie Chart** I used for showing the Distribution of each genre within the dataset.



This above shown **Line Chart** I used for comparing the **variance of IMDB Scores across different genres**
It tells that which genres have the most **consistent** or variable ratings.



This above Shown **Scatter Plot** I used for comparing the relationship between mean IMDB score and Variance across Genres.
After all this analysis I will recommend

For Filmmakers: Consider Focusing on genres with high average scores and low variance for more consistent success.

For Studios: Invest in genres that show consistent high ratings or explore improving genres with high variability**.**

So, at last Genres with High Ratings generally have the strong emotional responses, provide high entertainment value, or offer artistic and narrative depths.

And variable ratings genres always show a mix of high- and low-quality films, leading to more inconsistent ratings.

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

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

Solution:

Excel File Link -- ..\IMDB excel files\IMDB_Movies TASK B.xlsx

Excel table: Cleaned Dataset

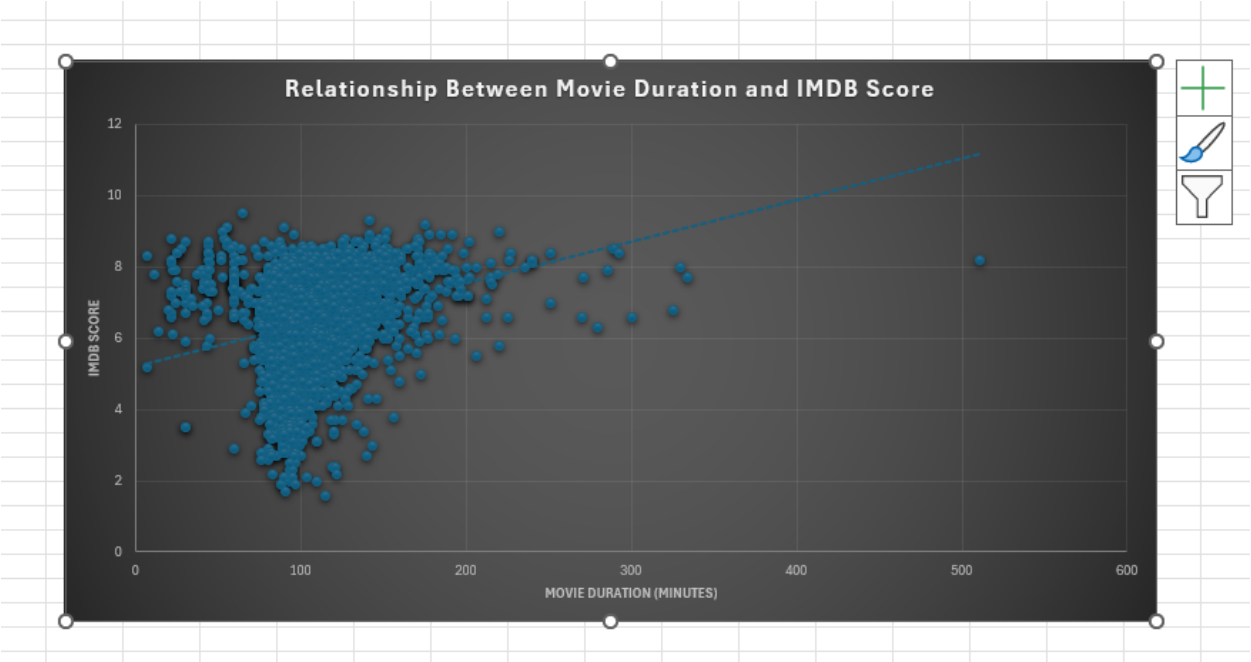
	A	B	C	D	E	F
1	duratio	imdb_score	Mean_MovieDuration	Median_MovieDuration	Stdev_MovieDuration	
2	178	7.9	107.2	103	25.2	
3	169	7.1				
4	148	6.8				
5	164	8.5				
6	132	6.6				
7	156	6.2				
8	100	7.8				
9	141	7.5				
10	153	7.5				
11	183	6.9				
12	169	6.1				
13	106	6.7				
14	151	7.3				
15	150	6.5				
16	143	7.2				
17	150	6.6				
18	173	8.1				
19	136	6.7				
20	106	6.8				
21	164	7.5				
22	153	7				
23	156	6.7				
24	186	7.9				
25	113	6.1				
26	201	7.2				
27	194	7.7				
28	147	8.2				
29	131	5.9				

IMDB_Movies (2) | Sheet1 | +

Explanation:

Here,

- Firstly, As the dataset provided is too big and messy, so I **cleaned** the dataset according to my task requirements and gathered two columns that is **duration, IMDB_score** in another sheet.
- After that I found some **missing values** in the duration column so to **handle missing values**, I **removed** the blank rows
- Then I calculated the **mean, median and standard deviation as**
  Mean Duration – provides the average length of movies in my dataset
  Median Duration – shows the middle value, which can indicate the typical movie length.
  Standard Deviation – Reveals the variability in movie durations.
- After these calculations I **visualized** the data using **Scatter plot and Trendline** which shows the positive correlation between movie duration and IMDB scores **as trendline moves in upward direction** which clearly defines that longer movies have **higher IMDB scores.**



→So according to my analysis Movie Duration has an outstanding impact on IMDB scores.

(P.T.O)

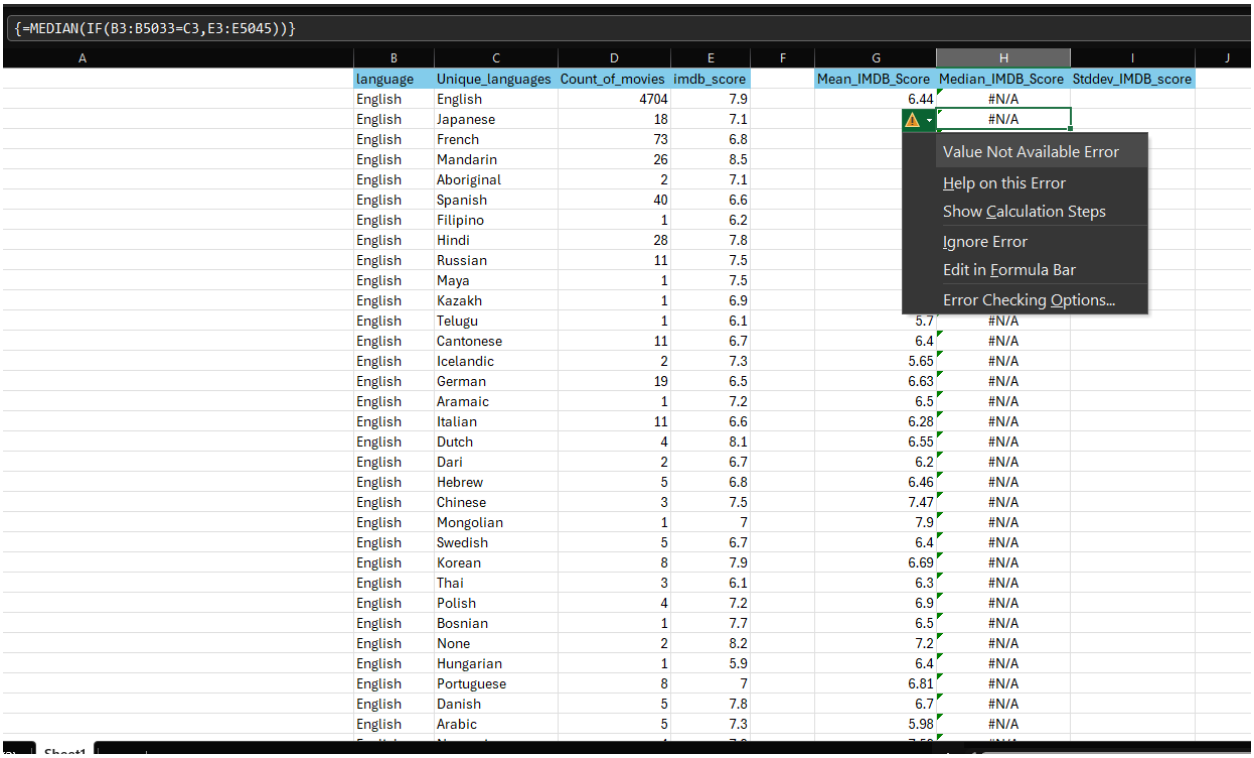
C) Language Analysis: Situation: Examine the distribution of movies based on their language.

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

Solution:

Excel File Link –..\IMDB excel files\IMDB_Movies TASK C.xlsx

Excel Dataset:



language	Unique_languages	Count_of_movies	imdb_score	Mean_IMDB_Score	Median_IMDB_Score	Stddev_IMDB_score
English	English	4704	7.9	6.44	#N/A	
English	Japanese	18	7.1		#N/A	
English	French	73	6.8			
English	Mandarin	26	8.5			
English	Aboriginal	2	7.1			
English	Spanish	40	6.6			
English	Filipino	1	6.2			
English	Hindi	28	7.8			
English	Russian	11	7.5			
English	Maya	1	7.5			
English	Kazakh	1	6.9			
English	Telugu	1	6.1	5.7	#N/A	
English	Cantonese	11	6.7	6.4	#N/A	
English	Icelandic	2	7.3	5.65	#N/A	
English	German	19	6.5	6.63	#N/A	
English	Aramaic	1	7.2	6.5	#N/A	
English	Italian	11	6.6	6.28	#N/A	
English	Dutch	4	8.1	6.55	#N/A	
English	Dari	2	6.7	6.2	#N/A	
English	Hebrew	5	6.8	6.46	#N/A	
English	Chinese	3	7.5	7.47	#N/A	
English	Mongolian	1	7	7.9	#N/A	
English	Swedish	5	6.7	6.4	#N/A	
English	Korean	8	7.9	6.69	#N/A	
English	Thai	3	6.1	6.3	#N/A	
English	Polish	4	7.2	6.9	#N/A	
English	Bosnian	1	7.7	6.5	#N/A	
English	None	2	8.2	7.2	#N/A	
English	Hungarian	1	5.9	6.4	#N/A	
English	Portuguese	8	7	6.81	#N/A	
English	Danish	5	7.8	6.7	#N/A	
English	Arabic	5	7.3	5.98	#N/A	

Explanation:

- Here as usual, first I cleaned the dataset and extracted the **language, IMDB score column** from the original dataset to another sheet.

- Then after removing the duplicate languages, I found the unique languages using **remove duplicate** feature which is there under **data** tab in excel
- After this I counted number of movies for each language using **countif** function
- Now when I started calculating the **descriptive statistics** for imdb scores I easily found the mean but when I tried to calculate the median the excel shown me error **“value not available error”** and same with the standard deviation.
- I tried many other ways to calculate these median and standard deviation, but I failed to get the results
- So, I am attaching my excel file link for better clarification

(P.T.O)

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.

Solution:

Excel File Link – ..\IMDB excel files\IMDB_Movies TASK D.xlsx

Excel Table Dataset:

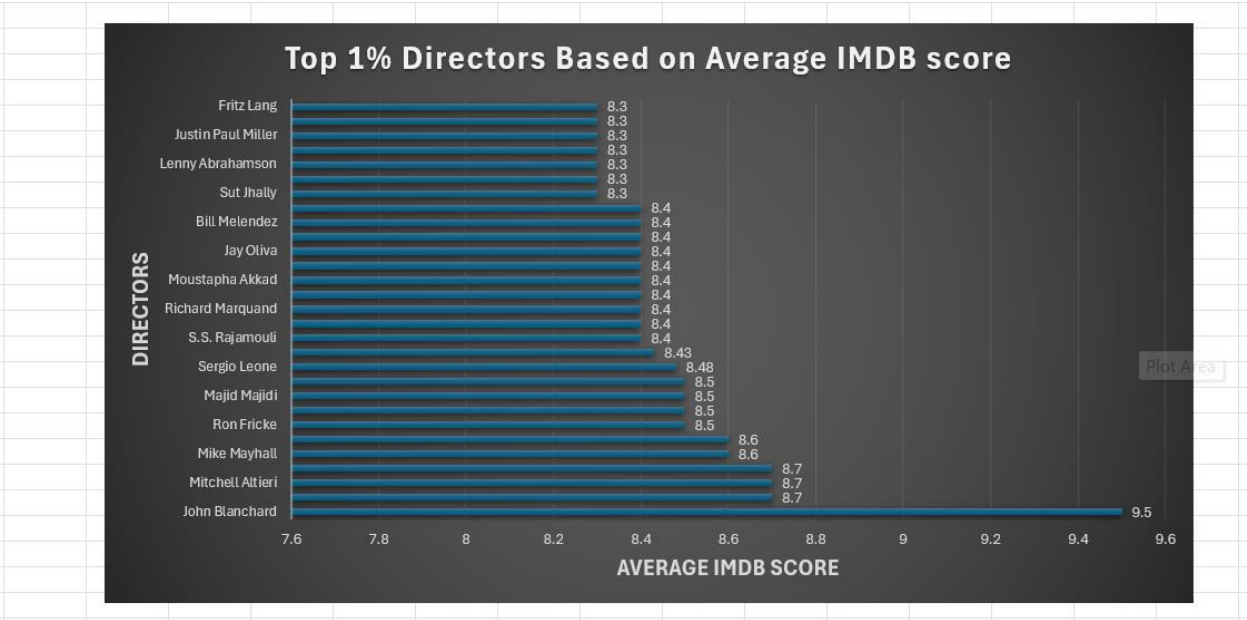
	A	B	C	D	E
1	director_name	imdb_score	Unique_director_name	Average IMDB score	Percentile ra
2	Jerry Jameson	4.7	John Blanchard	9.5	100
3	Tony Scott	7	Sadyk Sher-Niyaz	8.7	99.8
4	Clint Eastwood	7.4	Mitchell Altieri	8.7	99.8
5	Mike Bigelow	4.6	Cary Bell	8.7	99.8
6	Joe Wright	6.7	Mike Mayhall	8.6	99.7
7	Russell Crowe	7.1	Charles Chaplin	8.6	99.7
8	Peter Berg	6.7	Ron Fricke	8.5	99.6
9	David Pastor	6	Raja Menon	8.5	99.6
10	David Fincher	7.8	Majid Majidi	8.5	99.6
11	Alan Parker	7.3	Damien Chazelle	8.5	99.6
12	Luc Besson	7.7	Sergio Leone	8.48	99.5
13	Gavin O'Connor	5.8	Christopher Nolan	8.43	99.5
14	Carlos Saldanha	7	S.S. Rajamouli	8.4	99.1
15	Nancy Meyers	6.7	Robert Mulligan	8.4	99.1
16	Ron Howard	8.2	Richard Marquand	8.4	99.1
17	Don Bluth	6.6	Rakeysh Omprakash Mehra	8.4	99.1
18	Kenneth Branagh	6.2	Moustapha Akkad	8.4	99.1
19	Brian Helgeland	7.1	Marius A. Markevicius	8.4	99.1
20	Kevin Rodney Sullivan	5.9	Jay Oliva	8.4	99.1
21	Jessie Nelson	7.6	Catherine Owens	8.4	99.1
22	Kevin Rodney Sullivan	5.5	Bill Melendez	8.4	99.1
23	Garry Marshall	5.9	Asghar Farhadi	8.4	99.1
24	Roland Emmerich	6.4	Sut Jhally	8.3	98.8
25	James Cameron	7.2	Stanley Donen	8.3	98.8
26	Mike Nichols	5.6	Lenny Abrahamson	8.3	98.8
27		8.5	Lee Unkrich	8.3	98.8
28	Sydney Pollack	6.8	Justin Paul Miller	8.3	98.8
29	Clint Eastwood	6.9	John Sturges	8.3	98.8
30	Martha Coolidge	5.9	Fritz Lang	8.3	98.8
5063					

Explanation:

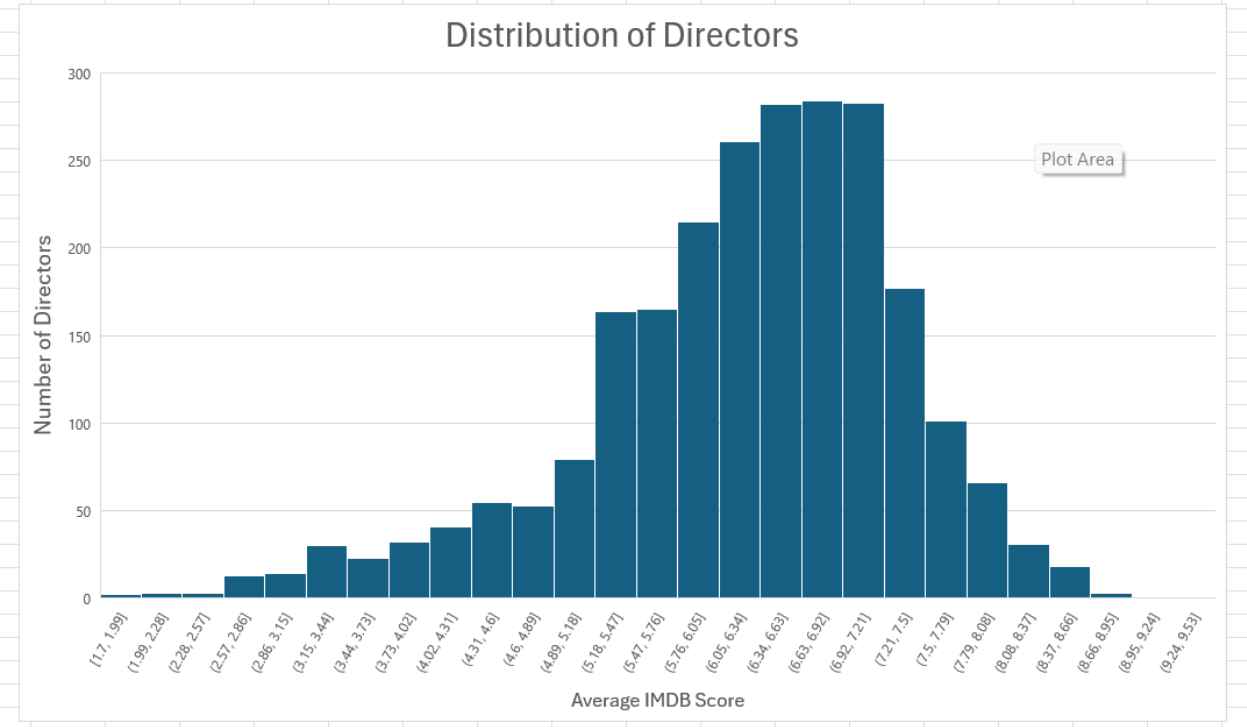
Here,

- I started **cleaning** the dataset first according to task requirements, copied director names and IMDB score column to another sheet for clear calculations.
- Then after copying the data, I removed the duplicate director names from the director's name column using **remove duplicates** feature present under the **data** tab.
- After that I calculated the **average IMDB score** for each director using the **averageif** function.
- Now the time for percentile calculation comes, for calculating the percentile rank of each director's average score I used **percentrank.inc*100** function.
- Then I applied filter to the percentile rank column and **identified the top 1% directors**.
- Now For Visualization of the data I used a **bar chart, histogram, and scatter plot**

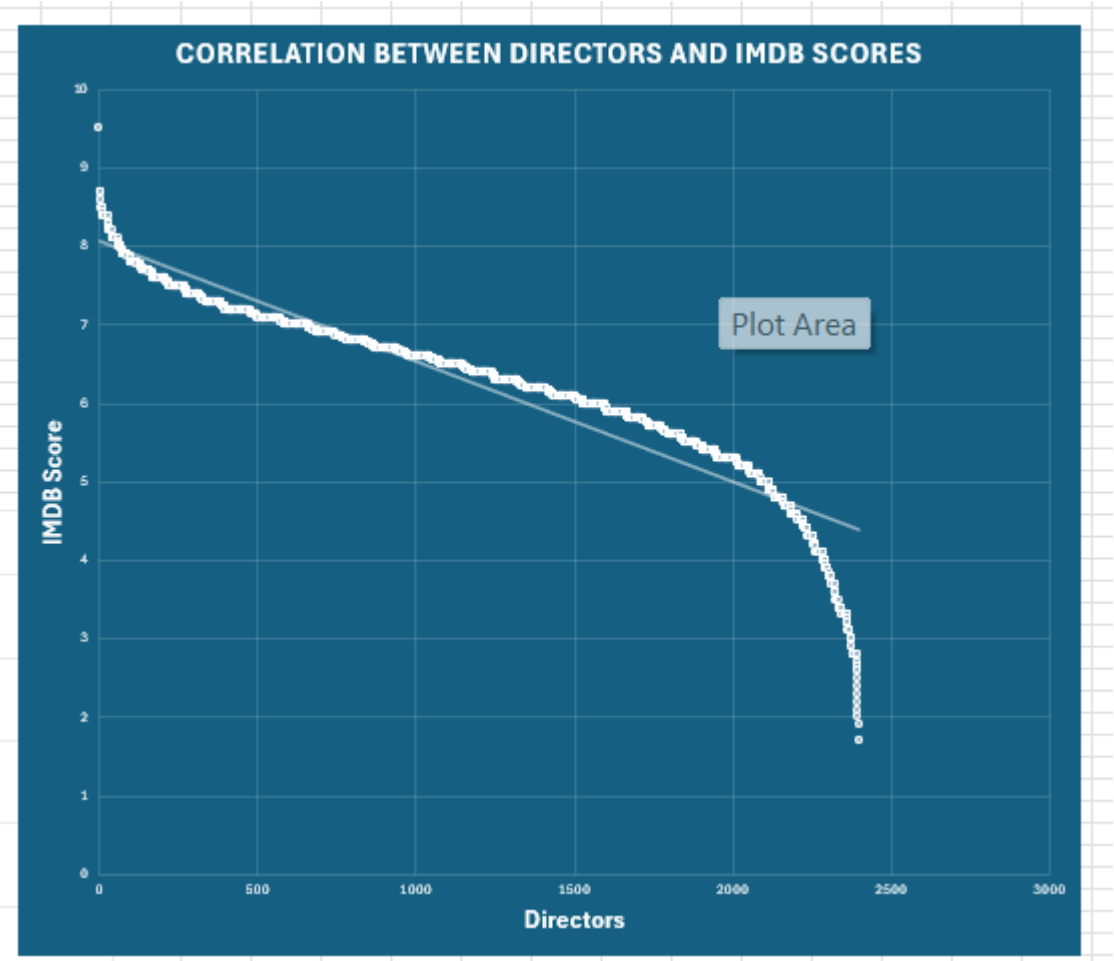
(P.T.O)



This **bar chart** clearly shows that directors with higher bars are those who consistently delivered movies with higher IMDB ratings.



This **Histogram** shows how many directors come across similar average scores or also used to show the wide variations.



So, this **scatter plot** helps me to visualize whether there is a direct correlation between a director and the IMDB scores of the movies they direct.

It shows a **downward trend** which indicates that as we move from left to right the IMDB score generally decreases this tells that leftmost directors produce movies with high IMDB scores compare to right side ones.

At last This Scatter plot clearly suggests that director plays a critical role in determining the success of a movie. Directors with higher ranks are associated with higher rated movies.

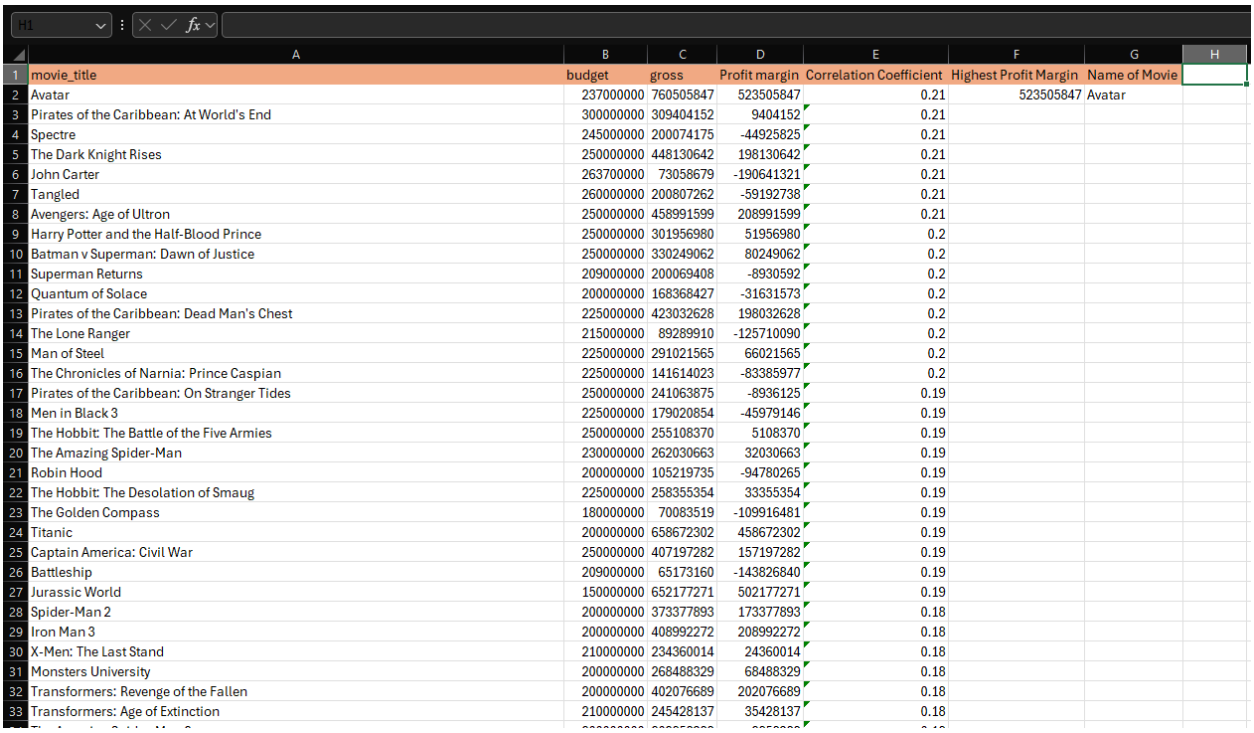
E) Budget Analysis: Explore the relationship between movie budgets and their financial success.

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

Solution:

Excel File Link –..\IMDB excel files\IMDB_Movies TASK E.xlsx

Excel Table Dataset:



	A	B	C	D	E	F	G	H
1	movie_title	budget	gross	Profit margin	Correlation Coefficient	Highest Profit Margin	Name of Movie	
2	Avatar	237000000	760505847	523505847	0.21	523505847	Avatar	
3	Pirates of the Caribbean: At World's End	300000000	309404152	9404152	0.21			
4	Spectre	245000000	200074175	-44925825	0.21			
5	The Dark Knight Rises	250000000	448130642	198130642	0.21			
6	John Carter	263700000	73058679	-190641321	0.21			
7	Tangled	260000000	200807262	-59192738	0.21			
8	Avengers: Age of Ultron	250000000	458991599	208991599	0.21			
9	Harry Potter and the Half-Blood Prince	250000000	301956980	51956980	0.2			
10	Batman v Superman: Dawn of Justice	250000000	330249062	80249062	0.2			
11	Superman Returns	209000000	200069408	-8930592	0.2			
12	Quantum of Solace	200000000	168368427	-31631573	0.2			
13	Pirates of the Caribbean: Dead Man's Chest	225000000	423032628	198032628	0.2			
14	The Lone Ranger	215000000	89289910	-125710090	0.2			
15	Man of Steel	225000000	291021565	66021565	0.2			
16	The Chronicles of Narnia: Prince Caspian	225000000	141614023	-83385977	0.2			
17	Pirates of the Caribbean: On Stranger Tides	250000000	241063875	-8936125	0.19			
18	Men in Black 3	225000000	179020854	-45979146	0.19			
19	The Hobbit: The Battle of the Five Armies	250000000	255108370	5108370	0.19			
20	The Amazing Spider-Man	230000000	262030663	32030663	0.19			
21	Robin Hood	200000000	105219735	-94780265	0.19			
22	The Hobbit: The Desolation of Smaug	225000000	258355354	33355354	0.19			
23	The Golden Compass	180000000	70083519	-109916481	0.19			
24	Titanic	200000000	658672302	458672302	0.19			
25	Captain America: Civil War	250000000	407197282	157197282	0.19			
26	Battleship	209000000	65173160	-143826840	0.19			
27	Jurassic World	150000000	652177271	502177271	0.19			
28	Spider-Man 2	200000000	373377893	173377893	0.18			
29	Iron Man 3	200000000	408992272	208992272	0.18			
30	X-Men: The Last Stand	210000000	234360014	24360014	0.18			
31	Monsters University	200000000	268488329	68488329	0.18			
32	Transformers: Revenge of the Fallen	200000000	402076689	202076689	0.18			
33	Transformers: Age of Extinction	210000000	245428137	35428137	0.18			

Explanation:

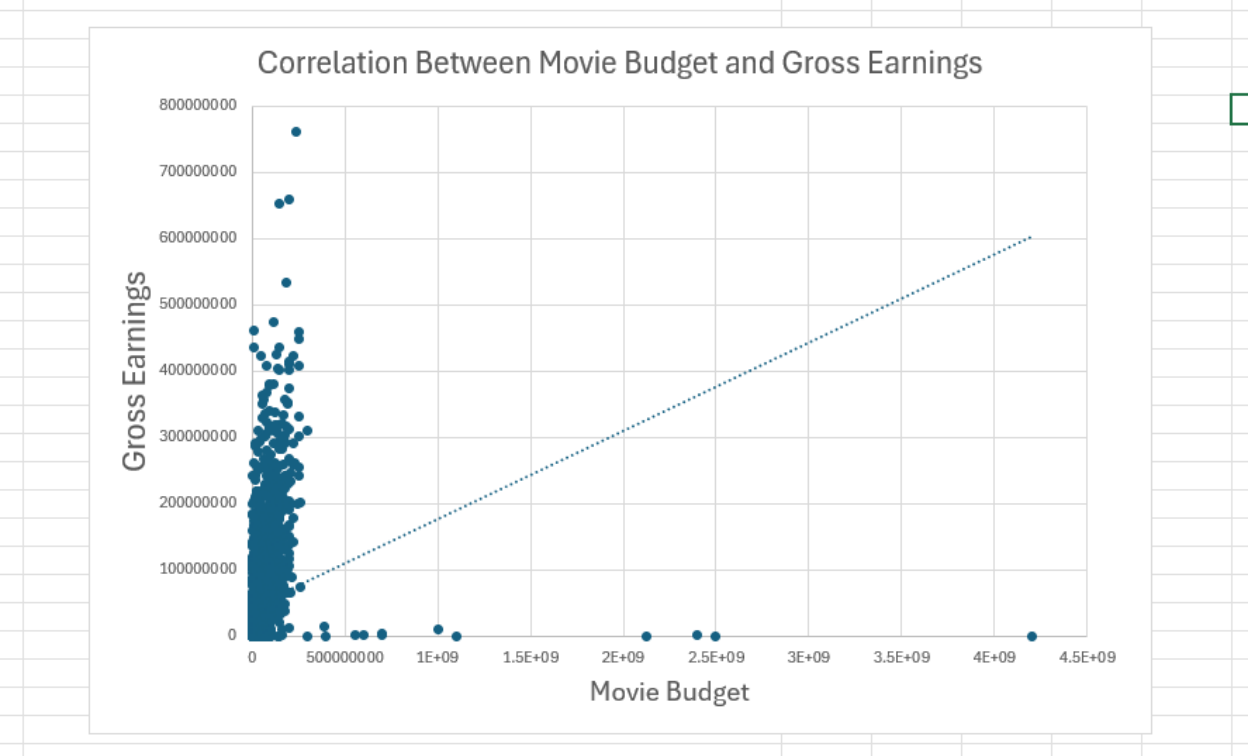
Again Starting with same **cleaning** procedure as the dataset provided is too big and messy –

- Firstly, I extracted **movies title, budget, gross earnings** columns from the original dataset to another sheet

- Then I removed the blanks and duplicate rows using the **remove duplicates** feature which is under the **data** tab and removed the blanks using filter feature.
- After that I calculated **the profit margin** for each movie by **subtracting the budget from the gross earnings**.
- Now I calculated the **Correlation Coefficient** which measures the strength and direction of the **linear relationship between movie budgets and gross earnings**

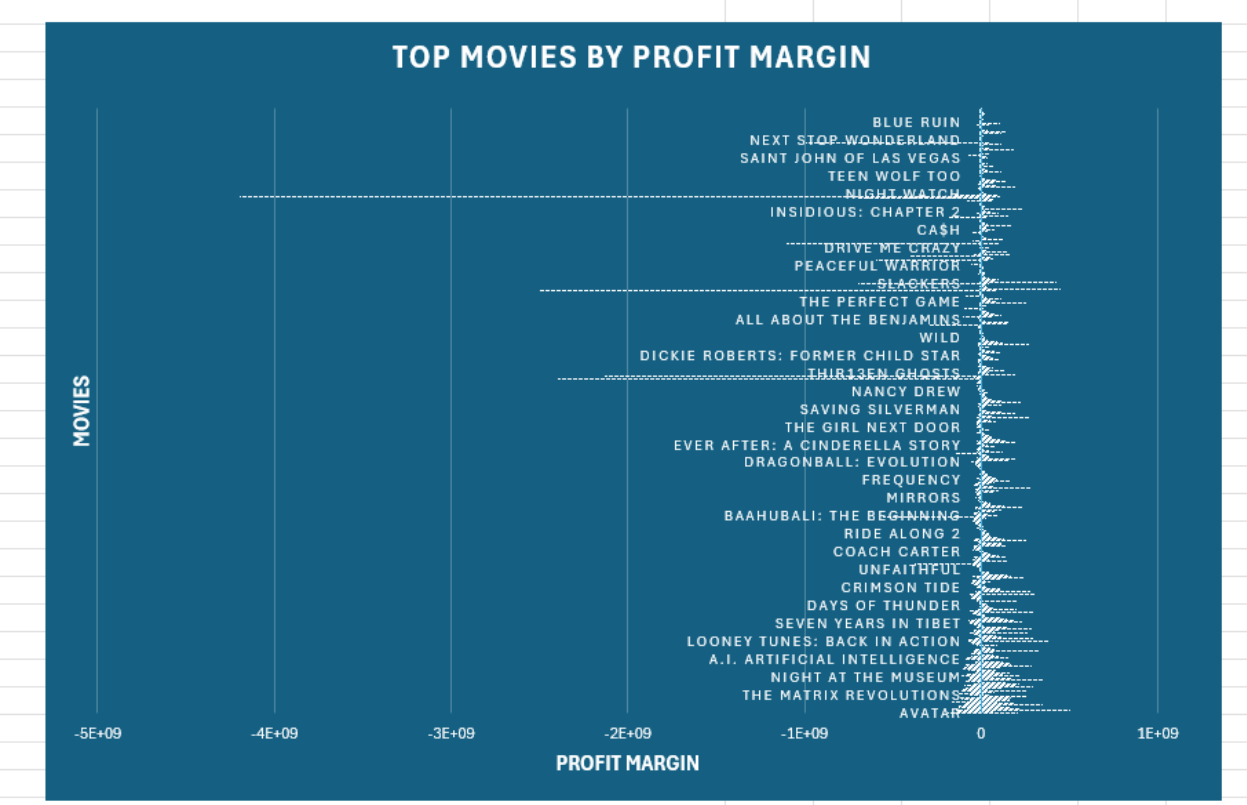
  In this,
  - **1 value** indicates a perfect positive correlation
  - **0 value** indicates no correlation
  - **-1 value** indicates a perfect negative correlation.
- After this I identified the movie with **the highest profit margin**

  in this,
  - First, I found the highest profit margin using **max function**
  - **Then using index function,** I found the name of the movie with highest profit margin
- Then at last I **visualized** the data using bar chart and scatter plot



This above shown **Scatter Plot** clearly highlights the correlation between movie budget and gross earnings as the **trendline is in upward direction** which means **Higher budget movies gives higher gross earnings.**

(P.T.O)



This above shown bar chart clearly defines that Movies achieved the highest financial success relative to their budget

In this **avatar movie** is having highest profit margin**.**

STORY TELLING

Introduction:

After the pandemic situation, which is lockdowns held over the world, **Top Gun: Maverick** became a massive box office hit, earning over $1.4 billion worldwide. Because its story, which proves to be an excellent romance story, sounded very clear to the audience.

Problem Statement:

Given the unpredictable nature of the film industry, understanding the factors that contribute to a movie's success is crucial for filmmakers, studios, and investors. This analysis seeks to uncover the elements that influence both critical acclaim and commercial performance.

Research Objectives:

This study aims to—

- Identify the most popular genres and analyze their impact on audience reception and commercial success.
- Determine the optimal movie duration to maximize viewer engagement and box office returns.
- Explore the influence of language on a film's global appeal and performance.
- Analyze the relationship between directorial talent and movie success.

- Investigate the Correlation between production budget and financial outcome.

Dataset Overview:

The dataset contains information on 4917(Approx) movies, **sourced from a variety of sources. Key variables include movie title, genre, duration, language, director, budget, gross earnings, and IMDB score mainly.**

Starting From

Genre Analysis:

The film industry is a vast ecosystem with a variety of genres for audience attention. My analysis reveals that around **27 distinct genres** are represented in the dataset. The most popular genres include **Drama, Action, Thriller, Romance, Comedy etc., collectively accounting for 50 to 60 % of the total movies.**

Genre, Ratings and Commercial Success:

There is an interesting pattern between **genre and audience** appreciation which is measured by IMDB scores. Some genres like Drama, Comedy etc. **consistently get highest average IMDB scores.** Which **indicates** that the audience loves to watch other genres also, but they find others less attractive.

To understand the **financial performance** of different genres, I analyzed their box office returns. Unsurprisingly, Drama genre is generating the highest average gross earning as most of the box office receipts are the proof of this genre.

Duration Analysis:

The length of the film is a critical factor influencing the viewer's experience. My analysis reveals that the **average movie duration in the dataset is 107 minutes** with **a standard deviation of 25 minutes**. This indicates a wide range of movie lengths, from short thriller movie duration of 70 minutes to epic dramas exceeding 107 minutes.

Duration, Audience, and Commercial Success

To understand how movie length impacts audience perception, I examined the correlation between duration and IMDB scores. Interestingly, **I found a positive correlation between these variables.** So, it suggests that longer length doesn't significantly impact audience satisfaction, as measured by IMDB ratings.

The relationship between movie length and box office performance is complex as some blockbuster films exceed two hours duration and there are films which achieved the success in shorter durations. My analysis indicates that longer/shorter movie length doesn't significantly impact a film's box office potential.

Language Analysis:

In the film industry there is a wide array of languages. My dataset reveals around **47 distinct languages** with **English language** being the most popular, accounting for **80 %** of the total movies.

Language, Audience and Market:

My analysis of IMDB scores across different language groups reveals that films in **English language** tend to have **higher average scores** compare to those in other languages. This suggests that cultural factors, audience preferences, or other reasons may **influence how films in different languages are adopted by viewers.**

Director Analysis:

The director is subjected as a men/woman behind a film's success. To assess the impact of directors on movie outcomes, I **calculated the average IMDB score** for each director with at least 4917 films in the dataset.

Director and Genre:

Some directors are known for their **expertise** in specific genres. For example, **john Blanchard is known for comedy films**, etc. This **specialization** often contributes to their success, as they have a deep understanding of their target audience.

Budget Analysis:

The film industry is characterized by a wide range of budget sizes, **from low budget to high budget.**

Budget and Revenues:

According to my analysis the **relationship between budget and revenues shows** that a films gross earning doesn't totally depends on the budget as shown in the visualization charts

Conclusion

Finally, my analysis has interpreted the factors influencing a movie's success. Genre, duration, language, directorial talent, and budget all contribute to a film's **critical acclaim and commercial performance.**

The data suggests that while big budget films often dominate the box office, it is necessary to consider the concept of **return on investment**. Low budget films with strong storytelling and targeted marketing can give notable success.

Understanding these trends is crucial for filmmakers, studios, and investors seeking to observe the complex area of the film industry.

(P.T.O)

Project Description

This project investigates the factors that influence the success of a movie on IMDB, with success defined by IMDB ratings. By analyzing an IMDB Movies dataset, my aim is to provide insights that can guide movie producers, directors, and investors in making informed decisions for future projects.

I started analysis with data cleaning, which includes handling missing values, removing duplicates, and preparing the data for deeper analysis.

In this project I had to explore the relationship between IMDB ratings and various factors such as genre, director, budget, year of release, and actors.
This exploration helped me to identify the key elements that contribute to movie's success.

Approach

My approach to the project is very clear and simple to understand

- In the first step I understood the task and what must be done then I cleaned the data and preprocessed the data for further analysis.
- Secondly, I started calculating the descriptive statistics as in most of the tasks I have to do the same calculations but with different columns.

- And then after all the calculations, I reached the results, Now I visualized the results using many types of charts present in excel.

Tech-Stack Used

--- Microsoft excel 365 (16.0.17…) enterprise addition

Insights

I Captured so many insights from this project which are

1. Genre has a great impact on the IMDB ratings, as if some genre is holding the market, then that genre has a high chance of getting high IMDB ratings**.**

2. Movie duration also influences the IMDB ratings, audience likes a certain time limit movies and if a movie is longer than that time frame audience gets bored or if a movie is too much short then also audience give bad ratings means movie duration also matters for high ratings.

3. Language also plays a crucial role in the success of a movie as if most of the audience likes to watch in English, and if movie is in any other language, then obviously it will get less ratings.

4. In a movie's success not only these genres, language, movie duration matters but also the **director and his vision**, if a director is good in some specific genre, then he can be the only person who can bring success for a movie.

5. In my analysis I revealed a positive correlation between movie budgets and gross earnings, which indicates that higher investment often leads to higher returns.

Result

This project gave me the knowledge of correlation between these multiple factors, and it filled me with a great IMDB analysis knowledge as earlier I think I properly didn't know that these many factors are involved in the success of a movie.

Overall, the success of a movie depends on multiple factors including genre, language, duration, and the director's involvement.

While the high budgets give the higher earnings, other elements like creative director, storytelling, and audience preferences play a crucial role in determining a movie's rating and overall success.

Video Presentation link--

https://drive.google.com/file/d/1JN4bbskNHItaTdmnFzJTKuqRy9ItdCHf/view?usp=sharing