IMDB Movie Analysis

Project Description:

The goal of this project is to analyse factors that influence the success of a movie on IMDB, with success defined by high IMBD ratings. The investigation is crucial for a movie producer's. Directors, and investors seeking insights into what makes a movie successful, enabling them to make informed decisions for future projects. By understanding the key factors contributing to a movie's success on IMDB, stakeholders can allocate resources more effectively. Tailor marketing strategies, and enhance the quality of their productions.

Approach:

Data collection and Pre-processing: Gather IMDB movie data, including attributes such as genre, budget, runtime, cast, director, release year, and IMDB rating. Pre-process the data by handling missing values, outliers and data formatting inconsistencies.

Exploratory Data Analysis (EDA): Perform descriptive statistics to understand the distribution of IMDB ratings and other relevant attributes. Identify potential correlations and patterns using scatter plots, histograms, box plots, and correlation matrices.

Feature Engineering: Create new features or transform existing ones that might better capture the essence of a movie's success. For example, create calculated variables like profitability ratios, categorize directors, genres on ratings etc.

Interpretation and Insights: Identify key drivers of movie success based on important scores from the analysis. Provide actionable insights and recommendations for movie industry professionals based on the findings.

Reporting and Recommendations: Providing actionable recommendations for optimizing the hiring process, reducing time to-to-fill, improving candidate experience, and enhancing overall recruitment effectiveness.

Tech Stack use:

MS Excel- A spreadsheet editor software used by professionals and businesses to enter data in a table format, perform data manipulations, computations, modelling, advanced analytics, plot graphs, etc

Hyperlink

https://drive.google.com/file/d/1BFzkTav6dVhMY0op_e9rMiiXW06ylFyb/view?usp=sharing

Insights

Initial Data Cleaning

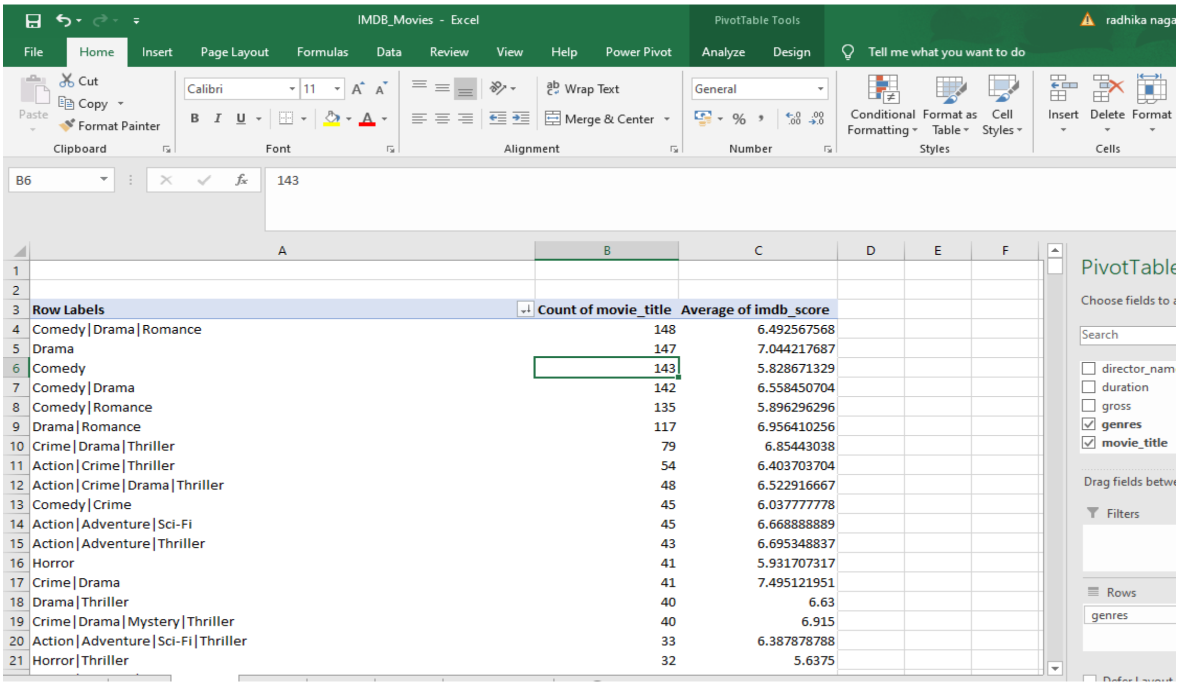
Handling Empty and duplicate values: Firstly check for blank values by find and select option and delete the blank values, after deleting blank check for duplicate values and delete them, before deleting we had 5043 rows, after cleaning we have 3737 unique values Remain.

Data manipulation: we create new calculated columns for better analysis of key metrics. We create profit by subtracting budget by gross earnings and profit margin by diving profit by budget and converting it to percentages.

Deleting unwanted columns: here we are analysing data related to budget , movie duration, genre, directors and language, hence any columns which is not related to these are deleted like actors facebook likes, actor names, links, aspect ration and many more and keep only required columns for analysis.

Task 1: Movie Genre Analysis

From the pivot table analysis the common genre in most of the films is Comedy, Drama and romance. Among the top five results by the count of movies, these three genres are in the top of the list.

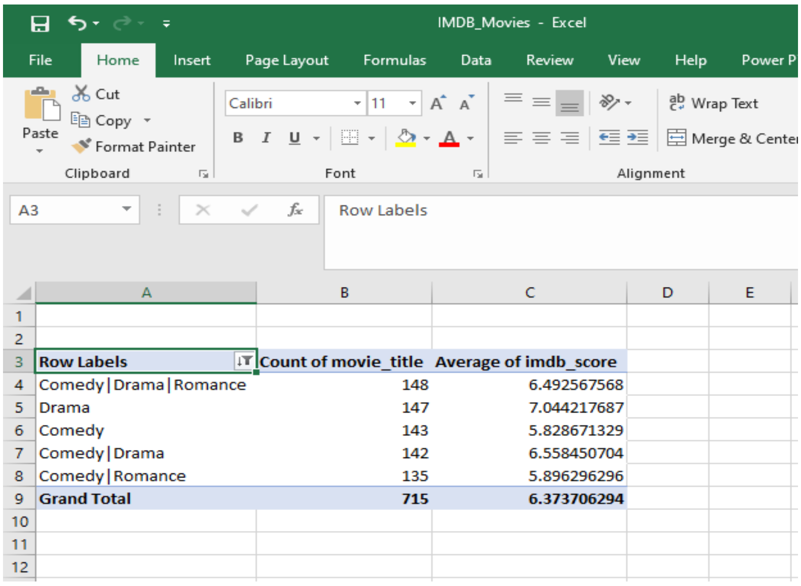


☐ director_nam
☐ duration
☐ gross
☑ genres
☑ movie_title

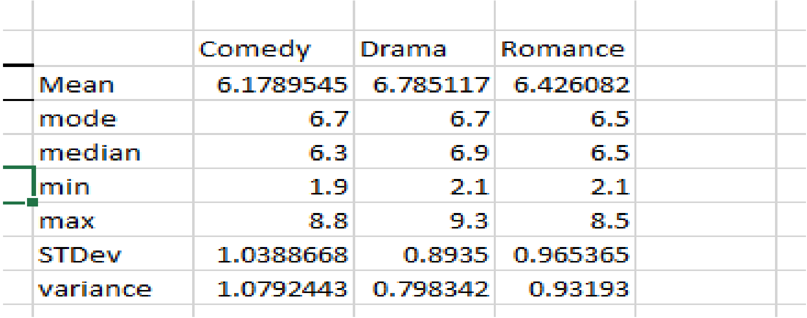
	Row Labels	Count of movie_title	Average of imdb_score
4	Comedy|Drama|Romance	148	6.492567568
5	Drama	147	7.044217687
6	Comedy	143	5.828671329
7	Comedy|Drama	142	6.558450704
8	Comedy|Romance	135	5.896296296
9	Drama|Romance	117	6.956410256
10	Crime|Drama|Thriller	79	6.85443038
11	Action|Crime|Thriller	54	6.403703704
12	Action|Crime|Drama|Thriller	48	6.522916667
13	Comedy|Crime	45	6.037777778
14	Action|Adventure|Sci-Fi	45	6.668888889
15	Action|Adventure|Thriller	43	6.695348837
16	Horror	41	5.931707317
17	Crime|Drama	41	7.495121951
18	Drama|Thriller	40	6.63
19	Crime|Drama|Mystery|Thriller	40	6.915
20	Action|Adventure|Sci-Fi|Thriller	33	6.387878788
21	Horror|Thriller	32	5.6375

By doing top 5 filtering, we can see the top 5 genres with their average IMDB scores, and average IMDB scores of these top genres is 6.37 which is very close to the overall average IMDB score of all the movies that is 6.46.

So it is clear that the top IMDB scores are for the movies of the genre comedy, romance and Drama. And people are most likely to accept and appreciate these kind of movies in future.



Row Labels	Count of movie_title	Average of imdb_score
Comedy|Drama|Romance	148	6.492567568
Drama	147	7.044217687
Comedy	143	5.828671329
Comedy|Drama	142	6.558450704
Comedy|Romance	135	5.896296296
Grand Total	715	6.373706294



	Comedy	Drama	Romance
Mean	6.1789545	6.785117	6.426082
mode	6.7	6.7	6.5
median	6.3	6.9	6.5
min	1.9	2.1	2.1
max	8.8	9.3	8.5
STDev	1.0388668	0.8935	0.965365
variance	1.0792443	0.798342	0.93193

The descriptive statistics using excel advance functions for the top 3 genres indicates that average values lie around the overall average imdb score and standard deviation being close to 1 indicates values lie close to the average and variance being close to 1 indicates values do vary too much from each other and lie close to average indicating imdb score depend on the genre and people are most likely to watch such movies

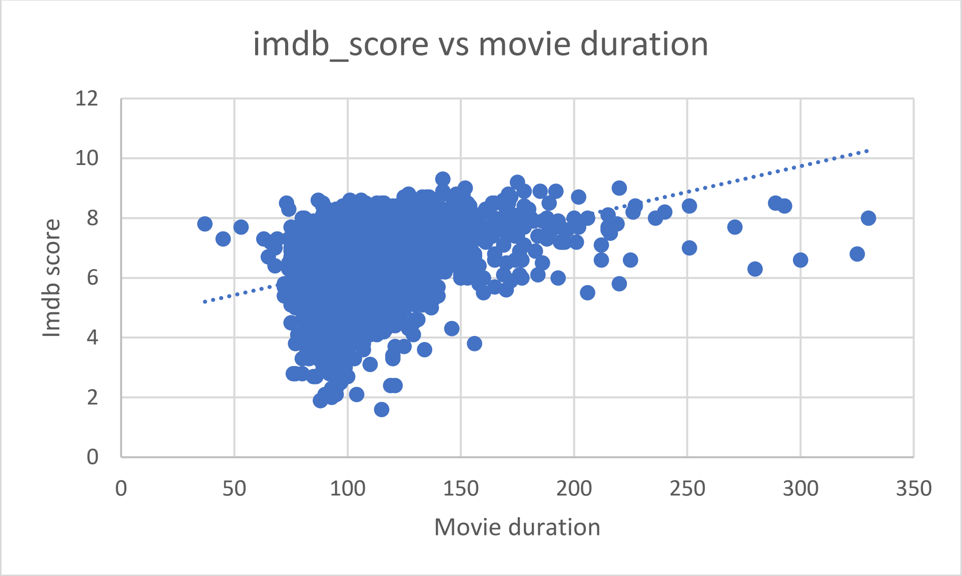
Below are the functions/formulas used for calculating statistics for genre “Comedy”

Mean=AVERAGEIF(D2:D3738,"*Comedy*",K2:K3738)
Mode=MODE(IF(ISNUMBER(SEARCH("Comedy",$D$2:$D$3738)),$K$2:$K$3738))
Median=MEDIAN(IF(ISNUMBER(SEARCH("Comedy",$D$2:$D$3738)),$K$2:$K$3738))
Min=MIN(IF(ISNUMBER(SEARCH("Comedy",$D$2:$D$3738)),$K$2:$K$3738))
Max=MAX(IF(ISNUMBER(SEARCH("Comedy",$D$2:$D$3738)),$K$2:$K$3738))
Stdev=STDEV(IF(ISNUMBER(SEARCH("Comedy",$D$2:$D$3738)),$K$2:$K$3738))
Variance=VAR(IF(ISNUMBER(SEARCH("Comedy",$D$2:$D$3738)),$K$2:$K$3738))

Task 2 : Movie Duration Analysis

The correlation coefficient between movie duration and imdb score is **=CORREL(B2:B3738,K2:K3738)** and the value is **0.37** which is very close to 0 meaning to say there is no relation between the two variables, and the imdb score does not depend on the length of the movie,

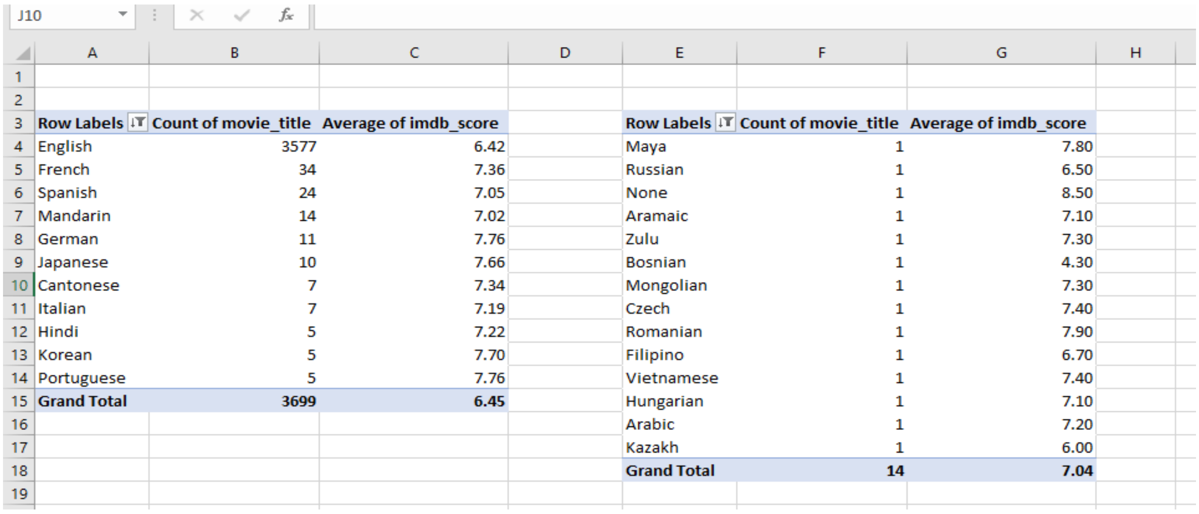
from the analysis and chart the most of the movies duration is around **100 minutes.**



Task 3 : Language Analysis

Very clearly nearly more than half of the movies in the dataset is from the language English and the average imdb score(**6.42)** is also very close to the **overall imdb score average**(**6.46**)

There are more than 10 languages where there is only one movie for each language and their IMDB score is listed below, even though there is only one film and imdb score frm that movie is more than the average overall imdb score for most of the movies, which can imply language doesn't matter but the content of the movie matters for good imdb score



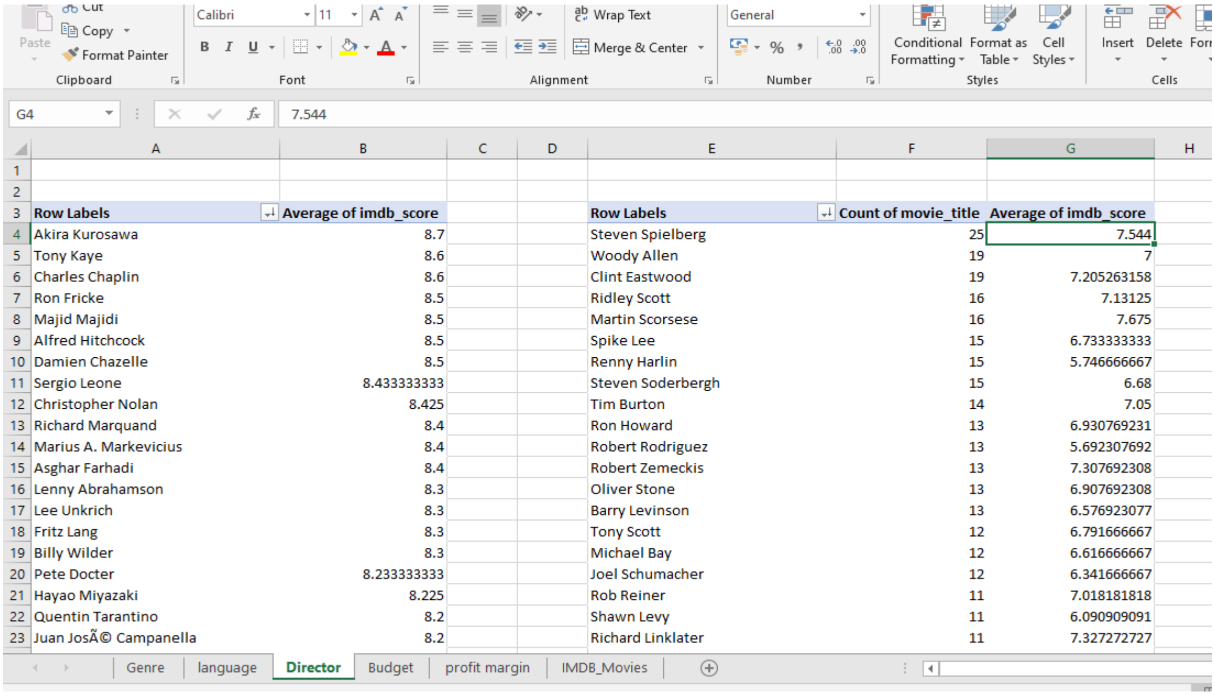
	A	B	C	D	E	F	G	H
1								
2								
3	**Row Labels**	**Count of movie_title**	**Average of imdb_score**		**Row Labels**	**Count of movie_title**	**Average of imdb_score**	
4	English	3577	6.42		Maya	1	7.80	
5	French	34	7.36		Russian	1	6.50	
6	Spanish	24	7.05		None	1	8.50	
7	Mandarin	14	7.02		Aramaic	1	7.10	
8	German	11	7.76		Zulu	1	7.30	
9	Japanese	10	7.66		Bosnian	1	4.30	
10	Cantonese	7	7.34		Mongolian	1	7.30	
11	Italian	7	7.19		Czech	1	7.40	
12	Hindi	5	7.22		Romanian	1	7.90	
13	Korean	5	7.70		Filipino	1	6.70	
14	Portuguese	5	7.76		Vietnamese	1	7.40	
15	**Grand Total**	**3699**	**6.45**		Hungarian	1	7.10	
16					Arabic	1	7.20	
17					Kazakh	1	6.00	
18					**Grand Total**	**14**	**7.04**	
19								

Row Labels	Count of movie_title	Average of imdb_score
Comedy|Drama|Romance	**148**	**6.492567568**
Arabic	1	7.2
English	142	6.464788732
French	2	7.25
Hebrew	1	7.3
Hindi	1	7.4
Italian	1	6.5
Drama	**147**	**7.044217687**
Aramaic	1	7.1
Danish	2	8.2
Dari	2	7.5
English	130	7.014615385
French	2	6.5
German	1	7.7
Hindi	1	6
Italian	1	7.7
Persian	1	7.5
Portuguese	2	7.05
Romanian	1	7.9
Spanish	2	6.9
Vietnamese	1	7.4
Comedy	**143**	**5.828671329**
English	143	5.828671329
Comedy|Drama	**142**	**6.558450704**
English	134	6.513432836
French	3	7.2
Japanese	1	6.1
Norwegian	1	7.6
Portuguese	1	7.9
Spanish	2	7.65
Comedy|Romance	**135**	**5.896296296**
English	133	5.869924812
French	2	7.65
Grand Total	**715**	**6.373706294**

If we consider the top genre with languages, we can see English is seen in all the genres in the above list, indicating language can influence imdb score, English being the universal language, there are viewers who know the language and would prefer movies in English language but, also languages in the above list comes from the bottom most movies with respect of "imdb score for languages" clearly indicating language does not matter for good viewer experience and few people would prefer watching movies in regional languages and if we consider imdb averages, the bottom 10 languages have greater imdb scores showing genre matters more than language.

Task 4: Director Analysis

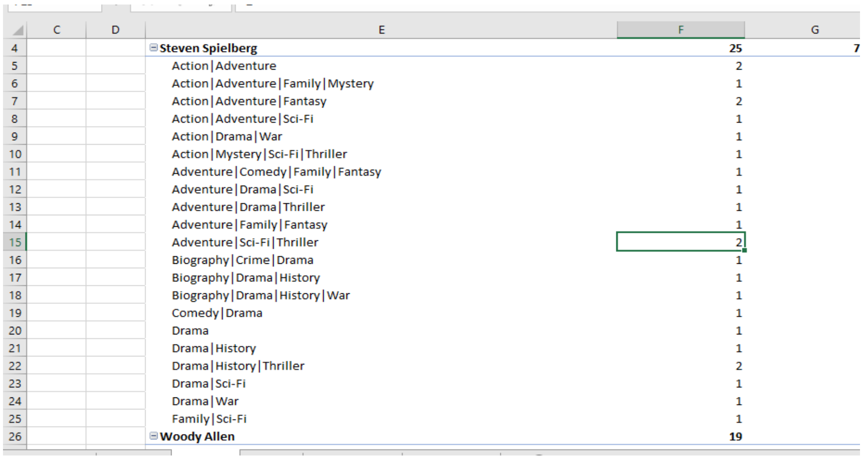
The highest imdb score is for the director Akira Kurosawa, but if we look into the count of movies made by the directors, Steven Spielberg tops the list and the average imdb score for their movies is 7.5 which belongs to the 90th percentile of the imdb score.
=PERCENTILE(G4:G1713,90%)



7.544

	A	B	C	D	E	F	G
1							
2							
3	Row Labels	Average of imdb_score			Row Labels	Count of movie_title	Average of imdb_score
4	Akira Kurosawa	8.7			Steven Spielberg	25	7.544
5	Tony Kaye	8.6			Woody Allen	19	7
6	Charles Chaplin	8.6			Clint Eastwood	19	7.205263158
7	Ron Fricke	8.5			Ridley Scott	16	7.13125
8	Majid Majidi	8.5			Martin Scorsese	16	7.675
9	Alfred Hitchcock	8.5			Spike Lee	15	6.733333333
10	Damien Chazelle	8.5			Renny Harlin	15	5.746666667
11	Sergio Leone	8.433333333			Steven Soderbergh	15	6.68
12	Christopher Nolan	8.425			Tim Burton	14	7.05
13	Richard Marquand	8.4			Ron Howard	13	6.930769231
14	Marius A. Markevicius	8.4			Robert Rodriguez	13	5.692307692
15	Asghar Farhadi	8.4			Robert Zemeckis	13	7.307692308
16	Lenny Abrahamson	8.3			Oliver Stone	13	6.907692308
17	Lee Unkrich	8.3			Barry Levinson	13	6.576923077
18	Fritz Lang	8.3			Tony Scott	12	6.791666667
19	Billy Wilder	8.3			Michael Bay	12	6.616666667
20	Pete Docter	8.233333333			Joel Schumacher	12	6.341666667
21	Hayao Miyazaki	8.225			Rob Reiner	11	7.018181818
22	Quentin Tarantino	8.2			Shawn Levy	11	6.090909091
23	Juan JosÃ© Campanella	8.2			Richard Linklater	11	7.327272727

100th percentile	90th percentile	80th percentile	70th percentile	60th percentile	50th percentile
9.3	7.7	7.4	7.1	6.828	6.67



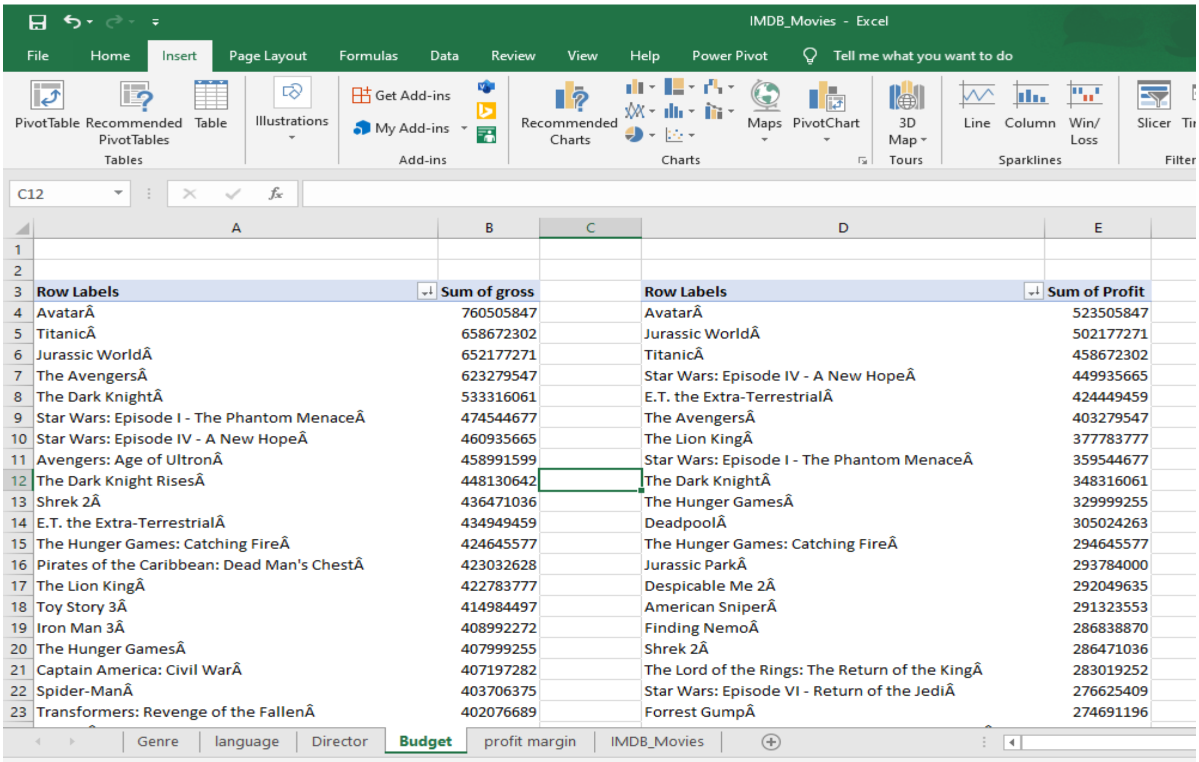
	C	D	E	F	G
4			⊟Steven Spielberg	25	7
5			Action|Adventure	2	
6			Action|Adventure|Family|Mystery	1	
7			Action|Adventure|Fantasy	2	
8			Action|Adventure|Sci-Fi	1	
9			Action|Drama|War	1	
10			Action|Mystery|Sci-Fi|Thriller	1	
11			Adventure|Comedy|Family|Fantasy	1	
12			Adventure|Drama|Sci-Fi	1	
13			Adventure|Drama|Thriller	1	
14			Adventure|Family|Fantasy	1	
15			Adventure|Sci-Fi|Thriller	2	
16			Biography|Crime|Drama	1	
17			Biography|Drama|History	1	
18			Biography|Drama|History|War	1	
19			Comedy|Drama	1	
20			Drama	1	
21			Drama|History	1	
22			Drama|History|Thriller	2	
23			Drama|Sci-Fi	1	
24			Drama|War	1	
25			Family|Sci-Fi	1	
26			⊟Woody Allen	19	

From the percentile distribution, and the count of movies we can see the directors belonging to different percentiles from the overall average imdb score. Also the director Steven Spielberg has a history of having worked on all kind of genres and can be easily considered the best director for future projects with his good imdb score for his movies.

Task 5: Budget Analysis

The highest grossing movie with respect to profit is **avatar**, but if we calculate the profit margin the movie **paranormal activity** tops the list, avatar might have grossed the highest among the movies in the dataset, but with respect to its budget its **221%** whereas paranormal activity with **low budget** has **700k%** profit margin meaning the profit isn't directly related to budget.

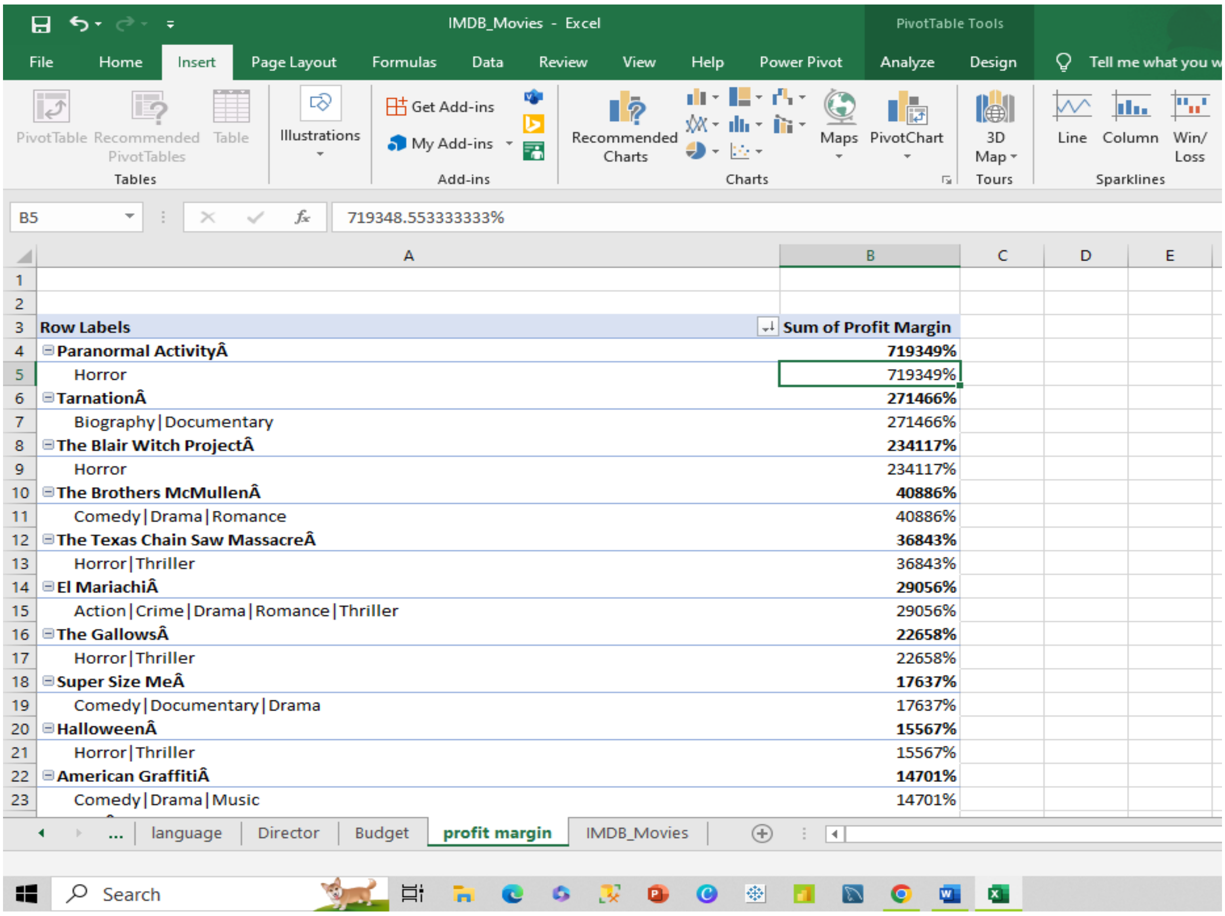
The correlation coefficient of budget vs gross earnings =**CORREL(I2:I3738,C2:C3738)** which is **0.10** which is close to 0. meaning to say not all highly budget movies make good profit and the correlation coefficient says both are not related at all.



	A	B	C	D	E
1					
2					
3	**Row Labels**	**Sum of gross**		**Row Labels**	**Sum of Profit**
4	AvatarÂ	760505847		AvatarÂ	523505847
5	TitanicÂ	658672302		Jurassic WorldÂ	502177271
6	Jurassic WorldÂ	652177271		TitanicÂ	458672302
7	The AvengersÂ	623279547		Star Wars: Episode IV - A New HopeÂ	449935665
8	The Dark KnightÂ	533316061		E.T. the Extra-TerrestrialÂ	424449459
9	Star Wars: Episode I - The Phantom MenaceÂ	474544677		The AvengersÂ	403279547
10	Star Wars: Episode IV - A New HopeÂ	460935665		The Lion KingÂ	377783777
11	Avengers: Age of UltronÂ	458991599		Star Wars: Episode I - The Phantom MenaceÂ	359544677
12	The Dark Knight RisesÂ	448130642		The Dark KnightÂ	348316061
13	Shrek 2Â	436471036		The Hunger GamesÂ	329999255
14	E.T. the Extra-TerrestrialÂ	434949459		DeadpoolÂ	305024263
15	The Hunger Games: Catching FireÂ	424645577		The Hunger Games: Catching FireÂ	294645577
16	Pirates of the Caribbean: Dead Man's ChestÂ	423032628		Jurassic ParkÂ	293784000
17	The Lion KingÂ	422783777		Despicable Me 2Â	292049635
18	Toy Story 3Â	414984497		American SniperÂ	291323553
19	Iron Man 3Â	408992272		Finding NemoÂ	286838870
20	The Hunger GamesÂ	407999255		Shrek 2Â	286471036
21	Captain America: Civil WarÂ	407197282		The Lord of the Rings: The Return of the KingÂ	283019252
22	Spider-ManÂ	403706375		Star Wars: Episode VI - Return of the JediÂ	276625409
23	Transformers: Revenge of the FallenÂ	402076689		Forrest GumpÂ	274691196

Profit margin can be calculated by diving profit by budget and converting it to percentage and by profit margin the top grossing movie is paranormal activity=**MAX(M2:M3738),** also the top three highest grossing belongs to genre horror and documentary, also looking into the pivot table we can see comedy , drama , romance genre is in the top grossings who also have the highest no of films and good average imdb score, which we have seen in the previous analysis.

This can influence investors into choosing correct genre and allocate budget effectively for their upcoming movies.



719348.553333333%

Row Labels	Sum of Profit Margin
⊟Paranormal ActivityÂ	**719349%**
Horror	719349%
⊟TarnationÂ	**271466%**
Biography|Documentary	271466%
⊟The Blair Witch ProjectÂ	**234117%**
Horror	234117%
⊟The Brothers McMullenÂ	**40886%**
Comedy|Drama|Romance	40886%
⊟The Texas Chain Saw MassacreÂ	**36843%**
Horror|Thriller	36843%
⊟El MariachiÂ	**29056%**
Action|Crime|Drama|Romance|Thriller	29056%
⊟The GallowsÂ	**22658%**
Horror|Thriller	22658%
⊟Super Size MeÂ	**17637%**
Comedy|Documentary|Drama	17637%
⊟HalloweenÂ	**15567%**
Horror|Thriller	15567%
⊟American GraffitiÂ	**14701%**
Comedy|Drama|Music	14701%

Result:

Data cleaning and preprocessing: the importance of data cleaning and preprocessing in preparing the dataset for analysis, including handling missing values, removing duplicates, converting calculated columns.

Understanding Factors influencing movie success: Gain insights into the various factors that contribute to the success of a movie on IMDB, such as genre, director, budget and other factors.

Correlation Analysis: learn how to perform correlation analysis to identify relationships between different variables and understand their impact on movie ratings.

Decision Support: learn how to translate data analysis findings into actionable recommendations to support decision-making for movie producers, directors, and investors, such as allocating budgets effectively, selecting genres or directors and prioritizing factors that enhance viewer experience.

Overall, the analysis aims to provide a comprehensive understanding of the dynamics behind movie success on IMDB, leveraging data-driven insights to inform decision-making in the film industry.