IMDB MOVIE ANALYSIS

TRAINITY

BY- ANURAG JOHN PHILLIPS

ROADMAP

PROJECT DESCRIPTION

APPROACH

TECH-STACK USED

INSIGHTS

RESULTS

IMDb

PROJECT DESCRIPTION

The objective of this project is to analyse the IMDB movie dataset and derive insights from it. This dataset has several columns, and we need to manipulate the data using Excel formulas.

Various tasks are required to be completed such as determining the most common genres, analysing the impact of genres on IMDB scores, examining the distribution of movie durations, identifying top directors based on their average IMDB scores, exploring the relationship between movie budgets and gross earnings, and calculate profit margins for each movie.

IMDb

APPROACH

- To execute the project, I first familiarized myself with the dataset and identified relevant columns for analysis.

- I then cleaned the data and removed unnecessary columns and rows which had null values, or which were not relevant for our tasks.

BEFORE CLEANING

COLUMNS – 28
ROWS - 5044

AFTER CLEANING

COLUMNS – 9
ROWS - 3818

CLEANED DATASET LINK

https://docs.google.com/spreadsheets/d/1VHlnwRA8T787v9x_VETPmYuBhoGbM86Jj-bYPnQ_-Dc/edit?usp=sharing

IMDb

WORKING EXCEL FILES

I have used these two excel files to do my analysis, reasons for dividing the Excel files into two files was, my system was getting hanging a lot when I was trying to do the analysis in just one files.

The links of these files are:

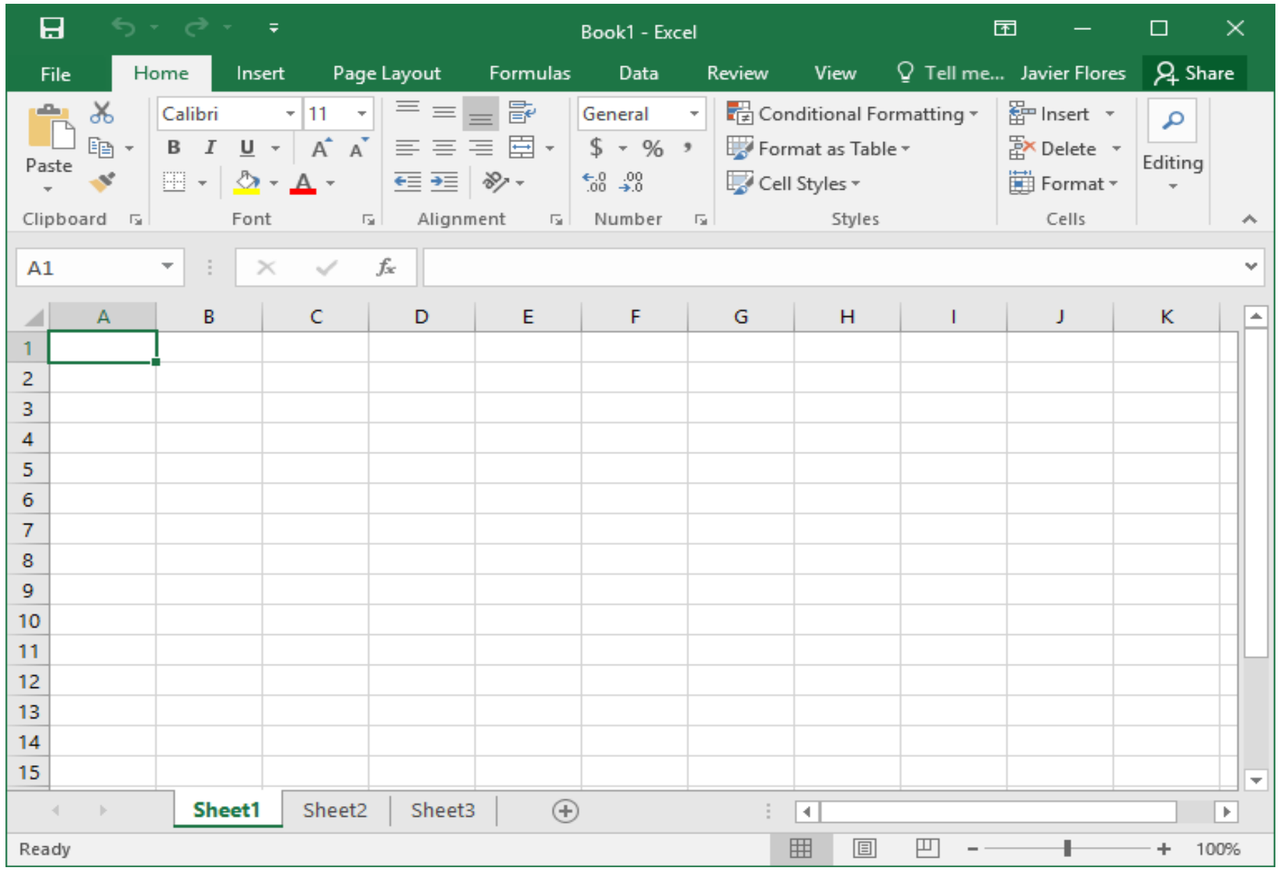
- https://docs.google.com/spreadsheets/d/1AIxcFU-PB3QepWV_ZjZWSRxYgenOf6bv/edit?usp=sharing&ouid=109466755193972209405&rtpof=true&sd=true

- https://docs.google.com/spreadsheets/d/15pi7R1bcRgHN00UxFtUBvBsc4SVm454v/edit?usp=sharing&ouid=109466755193972209405&rtpof=true&sd=true

TECH-STACK USED

Microsoft Excel for Mac Version 16.74

My primary tool for Data Analysis was Microsoft Excel and various functions and features of Excel were used which are mentioned in the respective sheets to calculate descriptive statistics, find correlations, and identify top directors and movies with the highest profit margins



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.

IMDB MOVIE ANALYSIS

Genres	1st	2nd	3rd	4th	5th	6th	7th	8th	IMDB SCORE
Documentary|Music	Documentary	Music							1.6
Comedy|Family|Sci-Fi	Comedy	Family	Sci-Fi						1.9
Comedy	Comedy								1.9
Comedy|Sport	Comedy	Sport							2
Drama|Music|Romance	Drama	Music	Romance						2.1
Comedy|Musical|Romance	Comedy	Musical	Romance						2.1
Action|Sport	Action	Sport							2.1
Comedy|Family|Fantasy	Comedy	Family	Fantasy						2.2
Comedy	Comedy								2.3
Horror|Sci-Fi	Horror	Sci-Fi							2.3
Adventure|Comedy	Adventure	Comedy							2.3
Comedy|Crime|Romance	Comedy	Crime	Romance						2.4
Action|Adventure|Sci-Fi	Action	Adventure	Sci-Fi						2.4
Comedy|Crime|Family|Sci-Fi	Comedy	Crime	Family	Sci-Fi					2.5
Documentary	Documentary								2.7
Action|Adventure|Fantasy|Sci-Fi|T	Action	Adventure	Fantasy	Sci-Fi	Thriller				2.7
Comedy	Comedy								2.7
Comedy|Romance	Comedy	Romance							2.7
Comedy|Music	Comedy	Music							2.8
Action|Crime|Sci-Fi	Action	Crime	Sci-Fi						2.8
Adventure|Animation|Comedy|Fan	Adventure	Animation	Comedy	Family	Fantasy				2.8
Adventure|Family	Adventure	Family							2.8
Adventure|Horror|Thriller	Adventure	Horror	Thriller						2.8
Action|Adventure|Fantasy|Horror	Action	Adventure	Fantasy	Horror					2.9
Comedy|Family	Comedy	Family							2.9
Adventure|Crime|Drama|Romance	Adventure	Crime	Drama	Romance					3
Comedy	Comedy								3
Action|Sci-Fi|Sport	Action	Sci-Fi	Sport						3

IMDb

Here I have divided the Genre Column into different columns based on “ | ” .

I used the function Text to Columns, which is under Data Tab

IMDb

Excel Formula Used

Genres	Count	Mean	Median	Mode	Range	Variance	Standard Deviation
Drama	1922	6.790426639	6.9	6.7	7.2	0.79110039	0.889438244
Comedy	1498	6.187583445	6.3	6.7	6.9	1.07431466	1.036491517
Thriller	1115	6.378295964	6.4	6.5	6.3	0.93313713	0.965990233
Action	957	6.293103448	6.3	6.1	6.9	1.06401097	1.031509072
Romance	874	6.432837529	6.5	6.5	6.4	0.9328609	0.965847243
Adventure	783	6.458109834	6.6	6.7	6.6	1.23305115	1.110428365
Crime	711	6.545428973	6.6	6.6	6.9	0.9585671	0.979064403
Fantasy	512	6.293164063	6.4	6.7	6.7	1.27664594	1.129887578
Sci-Fi	497	6.322736419	6.4	6.7	6.9	1.33655863	1.156096287
Family	448	6.213616071	6.3	6.7	6.7	1.35249875	1.162969798
Horror	391	5.926086957	6	5.9	6.3	0.99418952	0.997090528
Mystery	382	6.478534031	6.5	6.6	5.5	1.01029916	1.005136388
Biography	241	7.151037344	7.2	7	4.4	0.48184267	0.694148881
Animation	198	6.702525253	6.8	6.7	5.8	0.98146567	0.990689493
Music	158	6.463562753	6.7	6.2	6.9	1.41460189	1.189370375
War	155	7.070967742	7.1	7.1	4.3	0.65545036	0.809598886
Sport	151	6.603311258	6.8	7.2	6.4	1.08858896	1.043354668
History	148	7.160135135	7.2	7.7	3.4	0.44282175	0.665448533
Musical	103	6.559223301	6.7	7.1	6.4	1.30185037	1.140986578
Documentary	58	7.017241379	7.3	6.6	6.9	1.5923291	1.261875231
Western	57	6.812280702	6.8	6.8	4.2	0.88573935	0.941137263
Western	57	6.812280702	6.8	6.8	4.2	0.88573935	0.941137263
Short	2	6.8	6.8	#N/A	0.6	0.18	0.424264069
Film-Noir	1	7.7	7.7	#N/A	0	#DIV/0!	#DIV/0!

To count the number of movies in each genre-
=COUNTIF(Table3[[1st]:[8th]],L2)

To calculate Genre Mean –
=AVERAGE(FILTER($J$2:$J$3818, ISNUMBER(SEARCH(L2, $A$2:$A$3818))))

To calculate Genre Median –
=MEDIAN(FILTER($J$2:$J$3818, ISNUMBER(SEARCH(L2, $A$2:$A$3818))))

To calculate Genre Mode –
=MODE(FILTER($J$2:$J$3818, ISNUMBER(SEARCH(L2, $A$2:$A$3818))))

To calculate Genre Range –
=MAX(FILTER($J$2:$J$3818,ISNUMBER(SEARCH(L2,$A$2:$A$3818)))-
MIN(FILTER($J$2:$J$3818,ISNUMBER(SEARCH(L2,$A$2:$A$3818))))

To calculate Genre Variance–
=VAR(FILTER($J$2:$J$3818, ISNUMBER(SEARCH(L2, $A$2:$A$3818))))

To calculate Genre Standard Deviation–
=STDEV(FILTER($J$2:$J$3818, ISNUMBER(SEARCH(L2, $A$2:$A$3818))))

FINDINGS

Drama, Comedy, and Thriller are the most common genres, with a high number of movies in each genre.

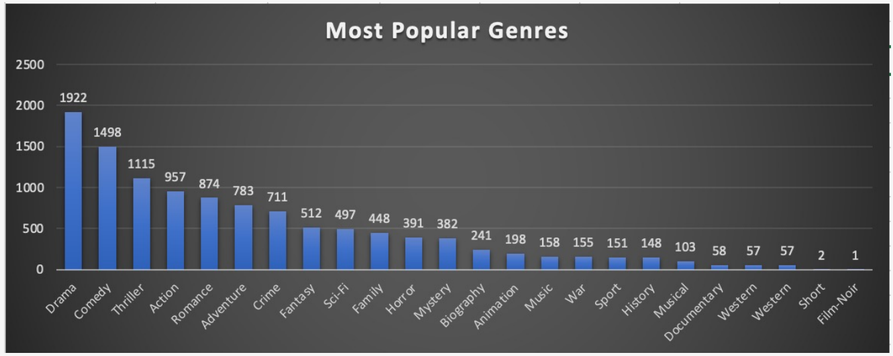
History, Biography, War and Documentary, have higher Mean IMDB scores. This indicates that movies in these genres generally receive better ratings from viewers.

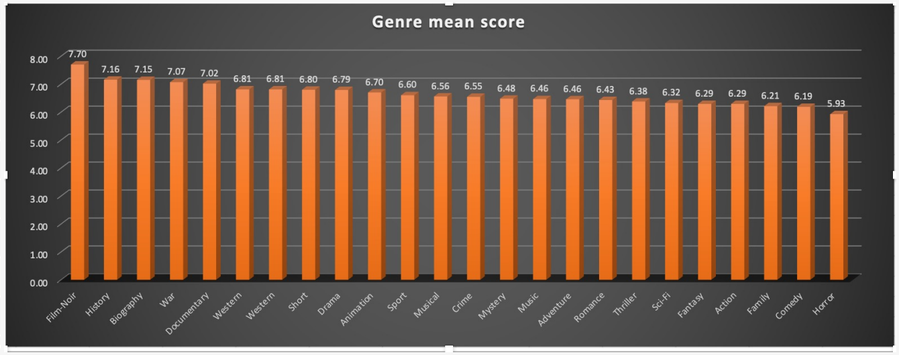
Horror and Comedy have lower Mean IMDB scores, which indicates that movies in these genres might receive relatively lower ratings.

Adventure and Fantasy genres have a wider range of IMDB scores, meaning that some movies in these genres have very high ratings, while others have lower ratings.

Drama and Romance, **have similar mean, median, and mode scores**, showing a relatively balanced distribution of ratings.

Genres	Count	Mean
Drama	1922	6.790426639
Comedy	1498	6.187583445
Thriller	1115	6.378295964
Action	957	6.293103448
Romance	874	6.432837529
Adventure	783	6.458109834
Crime	711	6.545428973
Fantasy	512	6.293164063
Sci-Fi	497	6.322736419
Family	448	6.213616071
Horror	391	5.926086957
Mystery	382	6.478534031
Biography	241	7.151037344
Animation	198	6.702525253
Music	158	6.463562753
War	155	7.070967742
Sport	151	6.603311258
History	148	7.160135135
Musical	103	6.559223301
Documentary	58	7.017241379
Western	57	6.812280702
Western	57	6.812280702
Short	2	6.8
Film-Noir	1	7.7





B.

MOVIE DURATION ANALYSIS

Analyse the distribution of movie durations and identify the relationship between movie duration and IMDB score

IMDB MOVIE ANALYSIS

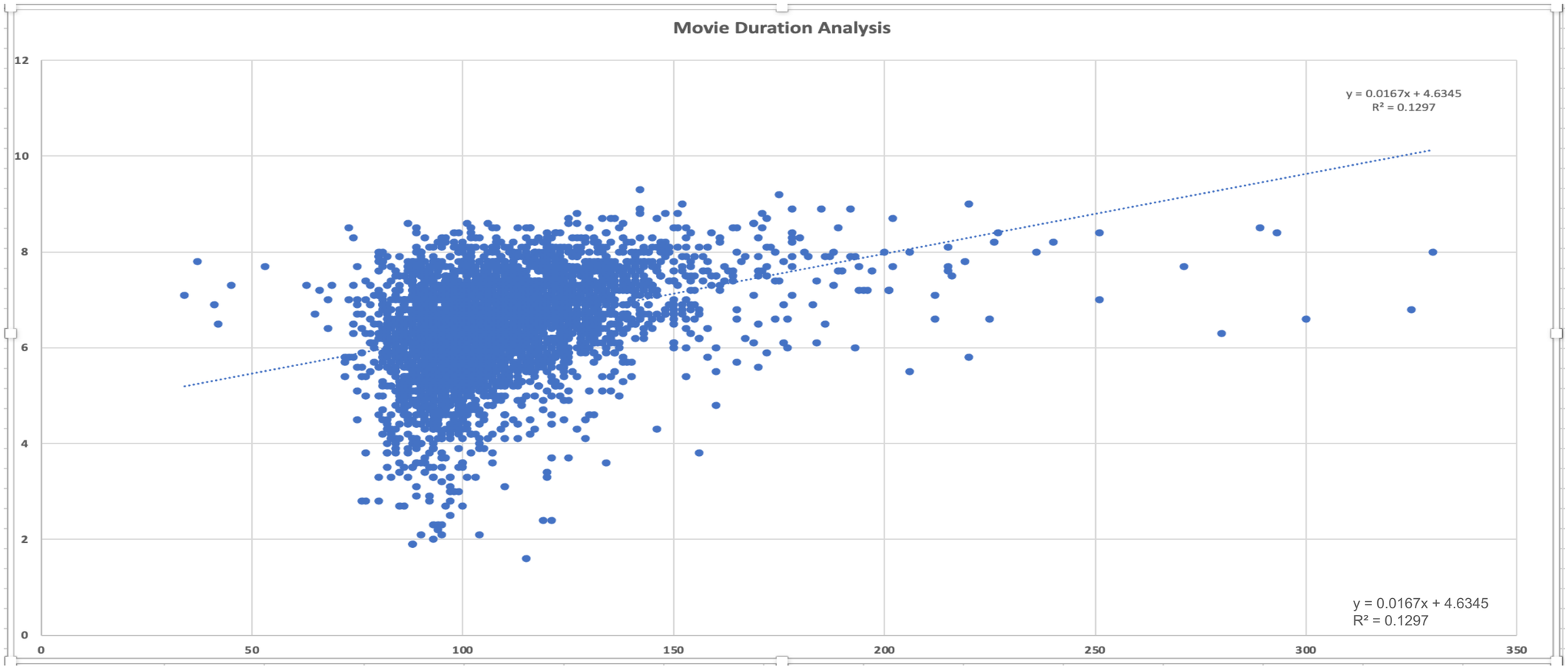
FOR MOVIE DURATION ANALYSIS
DURATION AND **IMDB_SCORE** COLUMNS
FROM THE DATASET
ARE CHOSEN FOR ANALYSIS

duration	imdb_score	movie_title
142	9.3	The Shawshank Redemption¬†
175	9.2	The Godfather¬†
152	9	The Dark Knight¬†
220	9	The Godfather: Part II¬†
192	8.9	The Lord of the Rings: The Return of the King¬†
185	8.9	Schindler's List¬†
178	8.9	Pulp Fiction¬†
142	8.9	The Good, the Bad and the Ugly¬†
148	8.8	Inception¬†
171	8.8	The Lord of the Rings: The Fellowship of the Ring¬†
151	8.8	Fight Club¬†
142	8.8	Forrest Gump¬†
127	8.8	Star Wars: Episode V - The Empire Strikes Back¬†
172	8.7	The Lord of the Rings: The Two Towers¬†
136	8.7	The Matrix¬†
146	8.7	Goodfellas¬†
125	8.7	Star Wars: Episode IV - A New Hope¬†
133	8.7	One Flew Over the Cuckoo's Nest¬†
135	8.7	City of God¬†
202	8.7	Seven Samurai¬†
169	8.6	Interstellar¬†
169	8.6	Saving Private Ryan¬†
127	8.6	Se7en¬†
138	8.6	The Silence of the Lambs¬†
125	8.6	Spirited Away¬†
101	8.6	American History X¬†
106	8.6	The Usual Suspects¬†
87	8.6	Modern Times¬†
164	8.5	The Dark Knight Rises¬†
171	8.5	Gladiator¬†
153	8.5	Terminator 2: Judgment Day¬†
165	8.5	Django Unchained¬†
151	8.5	The Departed¬†
73	8.5	The Lion King¬†

	Mean	Median	Standard Deviation
Duration	110.01	106	22.79
IMDB Score	6.47	6.6	1.05

From the above statistics, we can observe the following:

- The Mean and Median of Duration are relatively close, suggesting that the distribution of movie durations are roughly symmetric.
- The Standard Deviation for movie durations is larger than the standard deviation for IMDb scores, indicating that movie durations have more variability compared to IMDb scores..



To determine a relationship between movie duration and IMDb score, I plotted movie duration (x) and IMDb score (y) on a scatter plot and added the trendline to observe the direction and strength of the relationship visually.

The equation ($y = 0.0167x + 4.6345$) shows a positive linear relationship between movie duration and IMDb score . However, the coefficient of determination- R^2 value (0.1297) indicates that only 12.97% of the variation in IMDb scores can be explained by the variation in movie durations.

This suggests that there is a **weak positive relationship** between duration of the movie and its IMDB Score, what it means it that duration of the movie is not a big factor in the increase or decrease of their IMDB Scores, other factors also influence IMDb scores.

C.
LANGUAGE ANALYSIS

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

IMDB MOVIE ANALYSIS

Language	Count	IMDB Score Mean	IMDB Score Median	IMDB Score Std Dev
English	3641	6.43	6.5	1.048477755
French	37	7.29	7.2	0.561328861
Spanish	24	7.08	7.15	0.841829874
Mandarin	14	7.02	7.25	0.765786244
German	12	7.69	7.75	0.669407246
Japanese	12	7.63	7.8	0.899621132
Hindi	10	6.76	7.05	1.111755369
Cantonese	8	7.24	7.3	0.440575922
Italian	7	7.19	7	1.155318962
Portuguese	5	7.76	8	0.978774744
Korean	5	7.70	7.7	0.570087713
Norwegian	4	7.15	7.3	0.574456265
Persian	3	8.13	8.4	0.550757055
Danish	3	7.90	8.1	0.529150262
Dutch	3	7.57	7.8	0.404145188
Thai	3	6.63	6.6	0.450924975
Indonesian	2	7.90	7.9	0.424264069
Hebrew	2	7.65	7.65	0.494974747
Dari	2	7.50	7.5	0.141421356
Aboriginal	2	6.95	6.95	0.777817459
Telugu	1	8.40	8.4	#DIV/0!
Romanian	1	7.90	7.9	#DIV/0!
Maya	1	7.80	7.8	#DIV/0!
Swedish	1	7.60	7.6	#DIV/0!
Dzongkha	1	7.50	7.5	#DIV/0!
Czech	1	7.40	7.4	#DIV/0!
Vietnamese	1	7.40	7.4	#DIV/0!
Mongolian	1	7.30	7.3	#DIV/0!
Zulu	1	7.30	7.3	#DIV/0!
Arabic	1	7.20	7.2	#DIV/0!
Aramaic	1	7.10	7.1	#DIV/0!
Hungarian	1	7.10	7.1	#DIV/0!
Icelandic	1	6.90	6.9	#DIV/0!
Filipino	1	6.70	6.7	#DIV/0!
Russian	1	6.50	6.5	#DIV/0!
Kazakh	1	6.00	6	#DIV/0!
Bosnian	1	4.30	4.3	#DIV/0!

IMDb

EXCEL FORMULAES USED:

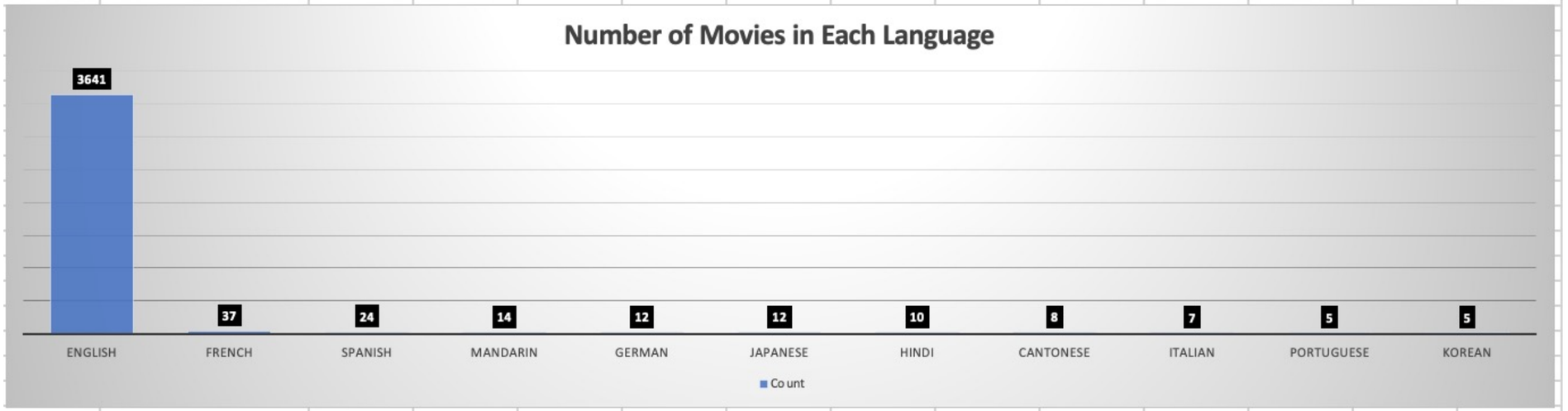
- **To get unique languages from Languages column**
  `=UNIQUE(Table2[language],FALSE,FALSE)`

- **To calculate count for "English"**
  `=COUNTIF(Table2[language],D2)`

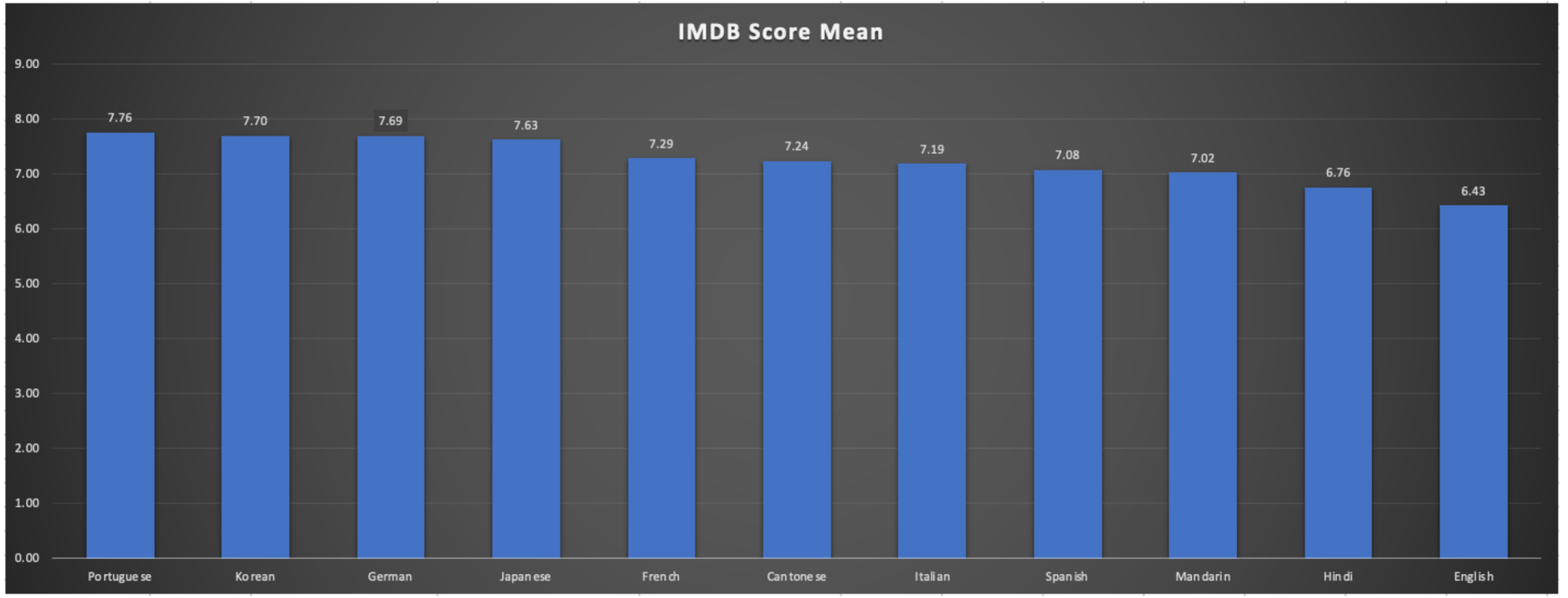
Languages with a larger number of movies have more reliable statistics, as the data is more representative of the overall movie population that is why for our analysis, we have taken only those movies whose language count is equal to or more than 5 movies per language.

Mean and Median of IMDB Score - helps us understand the overall reception of movies in each language. Languages with a high number of movies and relatively high IMDb Score Means and Medians indicates that these languages have consistently produced well-received movies.

Standard Deviation of IMDb Score - higher standard deviations suggests that there are many good and bad movies in these languages



Number of movies in each language



Mean IMDB Score of different Languages

IMDb

FINDINGS

- **English** has the most movies in the dataset, with a mean IMDB score of 6.43 and a median IMDB score of 6.5
- **French** has the second highest count with a mean IMDB score of 7.29 and a median IMDB score of 7.2
- **Spanish** has the third highest count, with a mean IMDB score of 7.08 and a median IMDB score of 7.15
- P**ortuguese, Korean and German** language movies have the highest Mean IMDB Score, but the number of movies in each language are less compared to English

D.
DIRECTOR ANALYSIS

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

IMDB MOVIE ANALYSIS

Director Names	Average IMDB SCORE	Number of Movies	Percentile
Christopher Nolan	8.4	8	95 Percentile
Quentin Tarantino	8.2	8	95 Percentile
James Cameron	7.9	7	95 Percentile
Alejandro G. IÌ±Ì°rritu	7.8	5	95 Percentile
David Fincher	7.8	10	95 Percentile
Martin Scorsese	7.7	16	90 Percentile
Peter Jackson	7.7	12	90 Percentile
Francis Ford Coppola	7.7	9	90 Percentile
Wes Anderson	7.6	7	90 Percentile
Paul Greengrass	7.6	7	90 Percentile
Brad Bird	7.6	5	90 Percentile
Steven Spielberg	7.5	25	90 Percentile
Paul Thomas Anderson	7.5	6	90 Percentile
Sam Mendes	7.5	8	90 Percentile
Darren Aronofsky	7.5	6	80 Percentile
Danny Boyle	7.4	8	80 Percentile
Alexander Payne	7.4	5	80 Percentile
George Lucas	7.4	5	80 Percentile
John Lasseter	7.4	5	80 Percentile
Mike Leigh	7.4	5	80 Percentile
Jean-Pierre Jeunet	7.3	5	80 Percentile
Terry Gilliam	7.3	7	80 Percentile
Richard Linklater	7.3	11	80 Percentile
Edward Zwick	7.3	8	80 Percentile
Robert Zemeckis	7.3	13	80 Percentile
Bryan Singer	7.3	8	80 Percentile
Ang Lee	7.3	8	80 Percentile
Marc Forster	7.2	7	80 Percentile
Clint Eastwood	7.2	19	80 Percentile
James Wan	7.2	7	80 Percentile
Jason Reitman	7.2	6	80 Percentile
Zack Snyder	7.2	8	80 Percentile
David O. Russell	7.2	7	80 Percentile

For our analysis I have taken those directors who have number of movies equal to or greater than 5

EXCEL FORMULAS USED

IMDb

To get Unique Directors Name-

=UNIQUE(Table4[director_name],FALSE,FALSE)

To get Mean IMDB Score of Each Director-

=AVERAGEIF($A$2:$A$3818, F2, $B$2:$B$3818)

To calculate Percentile- -->

=PERCENTILE($B$2:$B$3818, 0.95)

=PERCENTILE($B$2:$B$3818, 0.95)

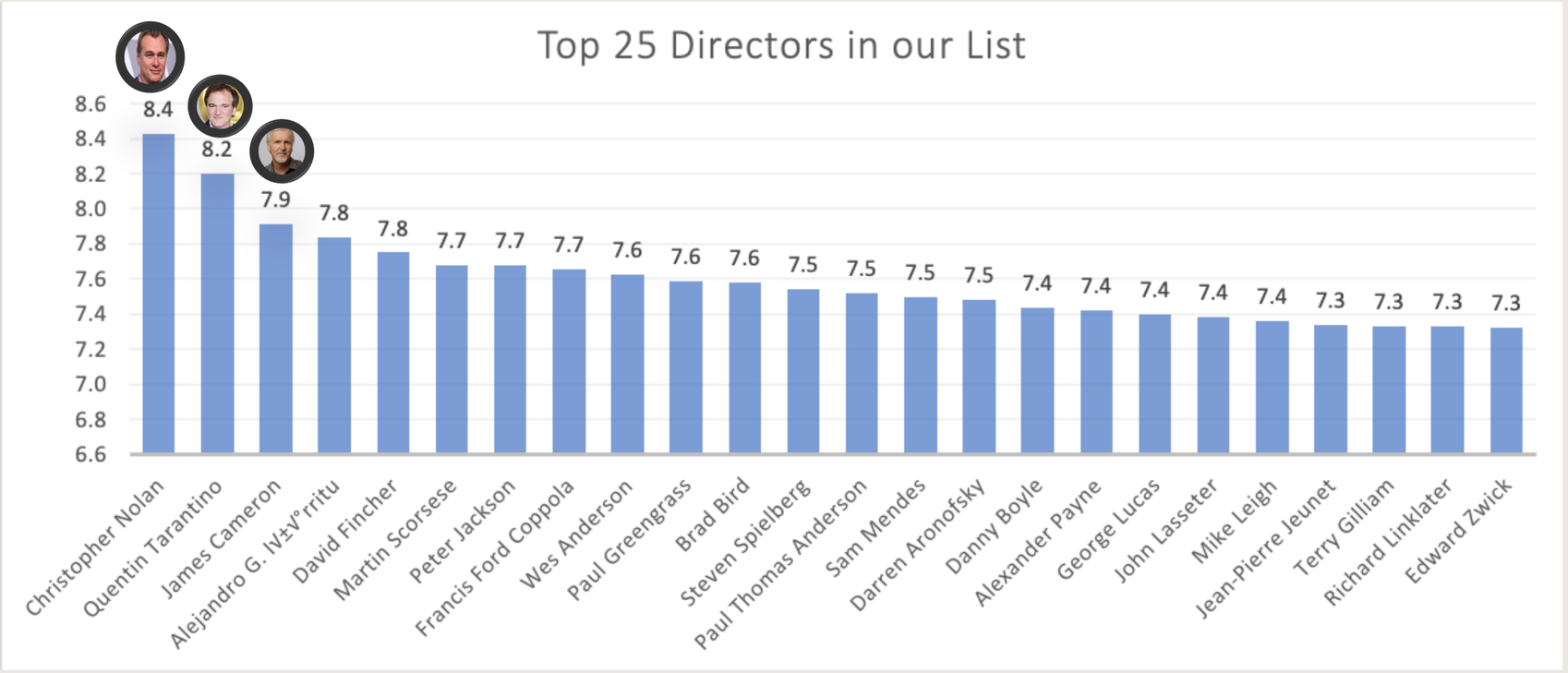
	G	H	I
	Total Directors	1724	
	95 Percentile	7.7	
	90 Percentile	7.5	
	80 Percentile	7.1	
	70 Percentile	6.9	
	60 Percentile	6.7	
	50 Percentile	6.5	
	40 Percentile	6.2	
	30 Percentile	5.95	
	20 Percentile	5.6	
	10 Percentile	5.1	
	5 Percentile	4.4	
	1 Percentile	3.223	
	0.5 Percentile	2.8	

To get number of Movies for each Director-

=COUNTIF(Table4[director_name],F2)

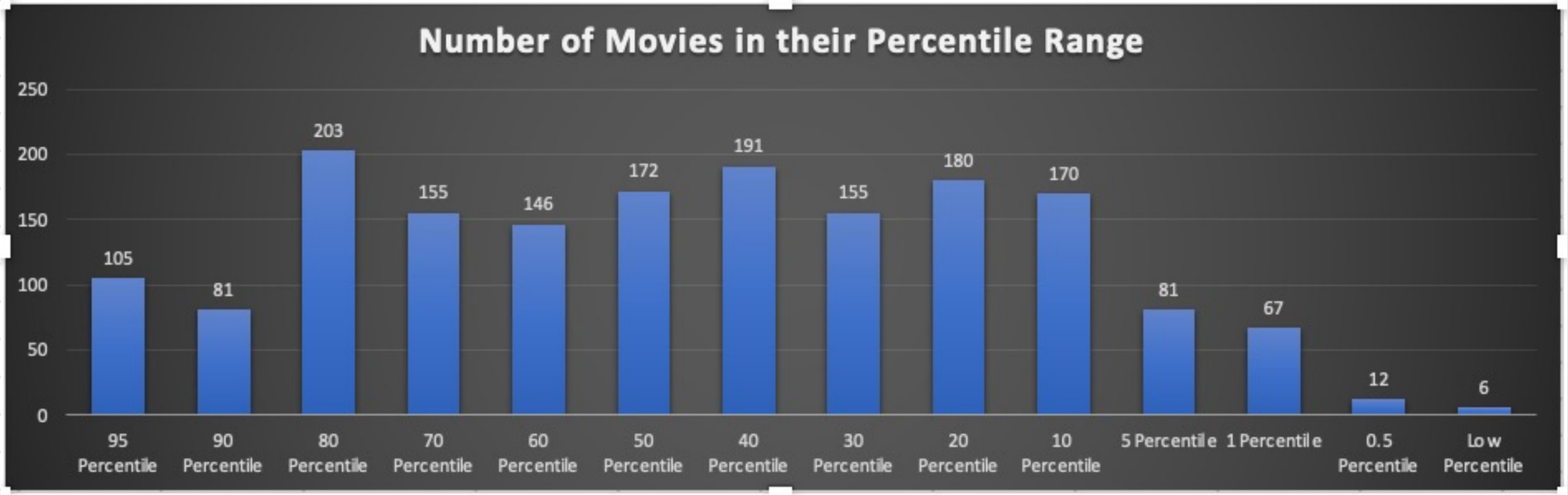
To automatically populate Percentile column according to their Average IMDB Score-

=IF(B2>=$H$2,$G$2,IF(B2>=$H$3, $G$3,IF(B2>=$H$4, $G$4,IF(B2>=$H$5, $G$5,IF(B2>=$H$6, $G$6,IF(B2>=$H$7, $G$7,IF(B2>=$H$8, $G$8,IF(B2>=$H$9, $G$9,IF(B2>=$H$10, $G$10,IF(B2>=$H$11, $G$11,IF(B2>=$H$12, $G$12,IF(B2>=$H$13, $G$13,IF(B2>=$H$14, $G$14,"Low Percentile")))))))))))))



We see the TOP 25 Directors in our list and the number of movies in their Percentile Range in the graph below

Bin Limit	Bin Label	Bin Counts
7.7	95 Percentile	105
7.5	90 Percentile	81
7.1	80 Percentile	203
6.9	70 Percentile	155
6.7	60 Percentile	146
6.5	50 Percentile	172
6.2	40 Percentile	191
5.95	30 Percentile	155
5.6	20 Percentile	180
5.1	10 Percentile	170
4.4	5 Percentile	81
3.223	1 Percentile	67
2.8	0.5 Percentile	12
	Low Percentile	6



E.
BUDGET ANALYSIS

Analyse the correlation between movie budgets and gross earnings and identify the movies with the highest profit margin.

IMDB MOVIE ANALYSIS

movie_title	Budget	gross	Profit	Max Profit
Avatar¬†	237000000	760505847	523505847	523505847
Jurassic World¬†	150000000	652177271	502177271	
Titanic¬†	200000000	658672302	458672302	
Star Wars: Episode IV - A	11000000	460935665	449935665	
E.T. the Extra-Terrestrial	10500000	434949459	424449459	
The Avengers¬†	220000000	623279547	403279547	
The Lion King¬†	45000000	422783777	377783777	
Star Wars: Episode I - Th	115000000	474544677	359544677	
The Dark Knight¬†	185000000	533316061	348316061	
The Hunger Games¬†	78000000	407999255	329999255	
Deadpool¬†	58000000	363024263	305024263	
The Hunger Games: Catc	130000000	424645577	294645577	
Jurassic Park¬†	63000000	356784000	293784000	
Despicable Me 2¬†	76000000	368049635	292049635	
American Sniper¬†	58800000	350123553	291323553	
Finding Nemo¬†	94000000	380838870	286838870	
Shrek 2¬†	150000000	436471036	286471036	
The Lord of the Rings: Th	94000000	377019252	283019252	
Star Wars: Episode VI - R	32500000	309125409	276625409	
Forrest Gump¬†	55000000	329691196	274691196	
Star Wars: Episode V - Th	18000000	290158751	272158751	
Home Alone¬†	18000000	285761243	267761243	
Star Wars: Episode III - R	113000000	380262555	267262555	
Spider-Man¬†	139000000	403706375	264706375	
Minions¬†	74000000	336029560	262029560	
The Sixth Sense¬†	40000000	293501675	253501675	
Jaws¬†	8000000	260000000	252000000	
Frozen¬†	150000000	400736600	250736600	
The Secret Life of Pets¬†	75000000	323505540	248505540	
The Twilight Saga: New I	50000000	296623634	246623634	
The Lord of the Rings: Th	94000000	340478898	246478898	
The Hangover¬†	35000000	277313371	242313371	
My Big Fat Greek Weddi	5000000	241437427	236437427	
The Twilight Saga: Eclips	68000000	300523113	232523113	
Independence Day¬†	75000000	306124059	231124059	
The Blind Side¬†	29000000	255950375	226950375	
Raiders of the Lost Ark¬·	18000000	242374454	224374454	
The Lord of the Rings: Th	93000000	313837577	220837577	

EXCEL FORMULAS USED

To calculate Profit Margin

=[@gross]-[@Budget]

To get highest Profit Margin

=MAX([Profit])

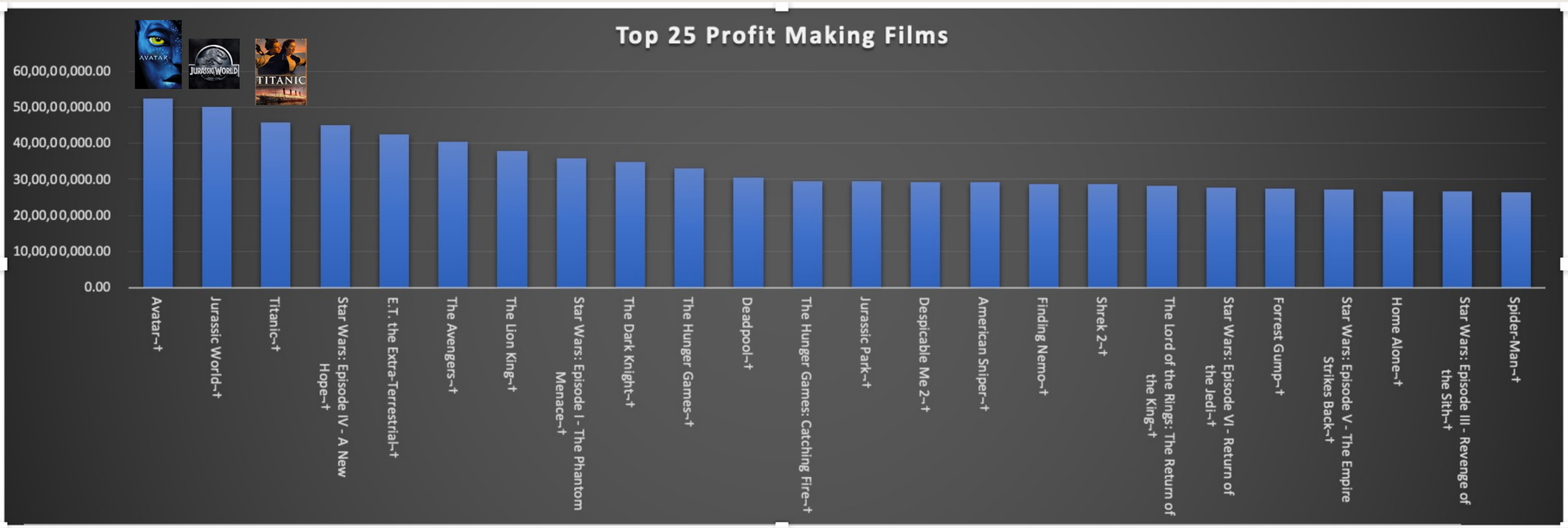
To calculate Correlation Coefficient

=CORREL(Table8[Budget],Table8[gross])

Correlation Coefficient	0.1000617

A correlation coefficient of 0.1000617 indicates a weak positive correlation between Movie Budgets and Gross Earnings.

This means that there is a slight tendency for movies with higher budgets to have higher gross earnings, however there are many movies with high budgets that did not have high gross earnings, and vice versa.



Top 25 Profit - and Loss-Making movies in our datasheet.

IMDb

RESULT

In this project, I successfully analysed various aspects of the IMDB movie dataset and provided meaningful insights into the distribution of genres, their impact on IMDB scores, movie durations, top directors, and the relationship between movie budgets and gross earnings.

The insights which I got are:

- **Movie Genre Analysis-** Drama, Comedy and Thriller are the most common genres
- **Movie Duration Analysis-** Shows that there is a weak positive relationship between the duration of the movie and its IMDB Score and other factors also influence IMDb scores
- **Language Analysis-** English, French and Spanish are the most common languages in our IMDB dataset
- **Director Analysis-** Christopher Nolan, Quentin Tarantino and James Cameron are the top 3 directors who are above 95 percentile range and have the highest Mean IMDB Score
- **Budget Analysis-** Avatar, Jurassic World and Titanic are the top 3 movies with the highest Profit Margins

This analysis has contributed to a better understanding of the factors influencing movie ratings and financial success, and it has provided valuable information for further exploration in the realm of movie analysis.

THANK YOU !