IMDB Movie Analysis

Project Description

In this project I have to analyze the data provided by IMDB on various movies containing attributes such as movie name, director, actors, budget, etc. Using the data provided I have derived the necessary insights

Approach

For this project, I used the dataset provided by the Trainity team and loaded it into Excel. Then I used the various inbuilt formulas and data transformation techniques of Excel to derive the necessary insights. I have also used various graphs and charts for the visualization of data

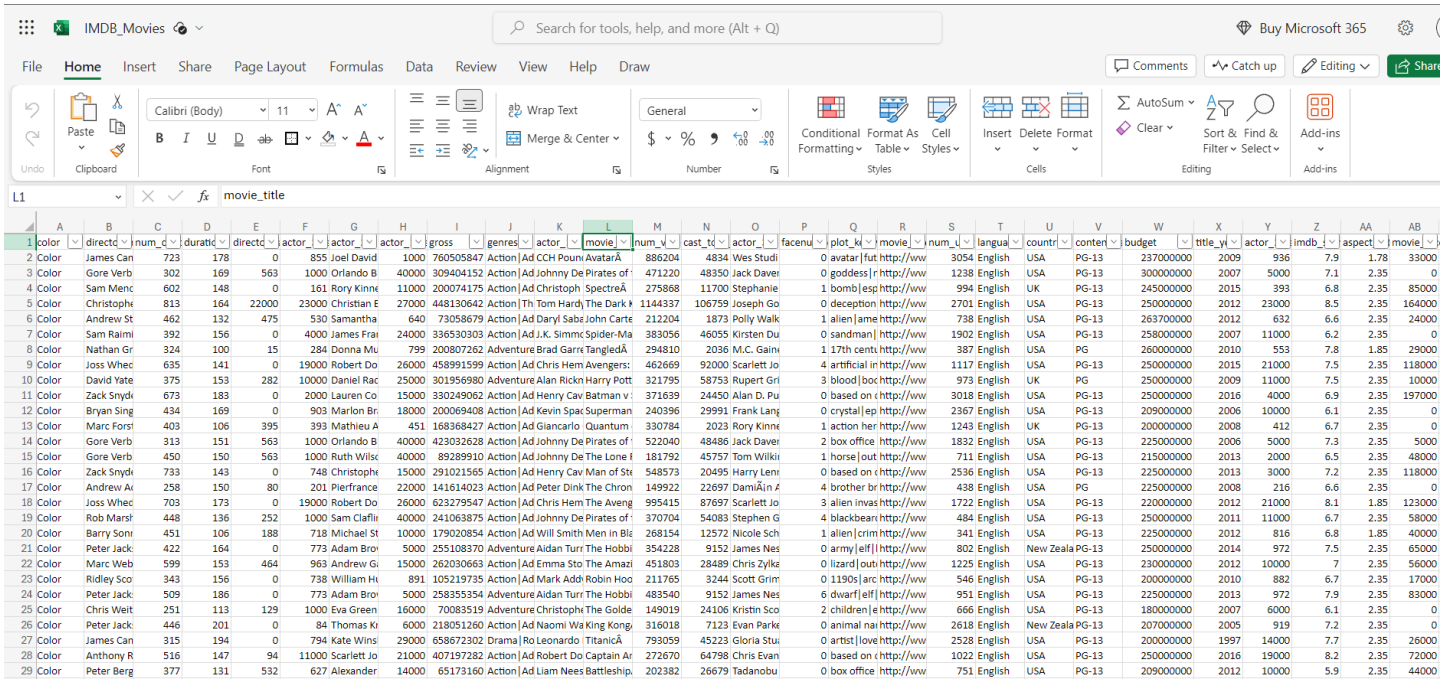
Tech-Stack Used

For this project, I have chosen Microsoft Excel as it is a powerful tool that offers numerous benefits for data analysis, business management, and personal use. Excel provides a wide range of built-in functions and formulas for mathematical, statistical, financial, and logical calculations. The use of filters and sorting mechanisms makes deriving insights easier

Insights

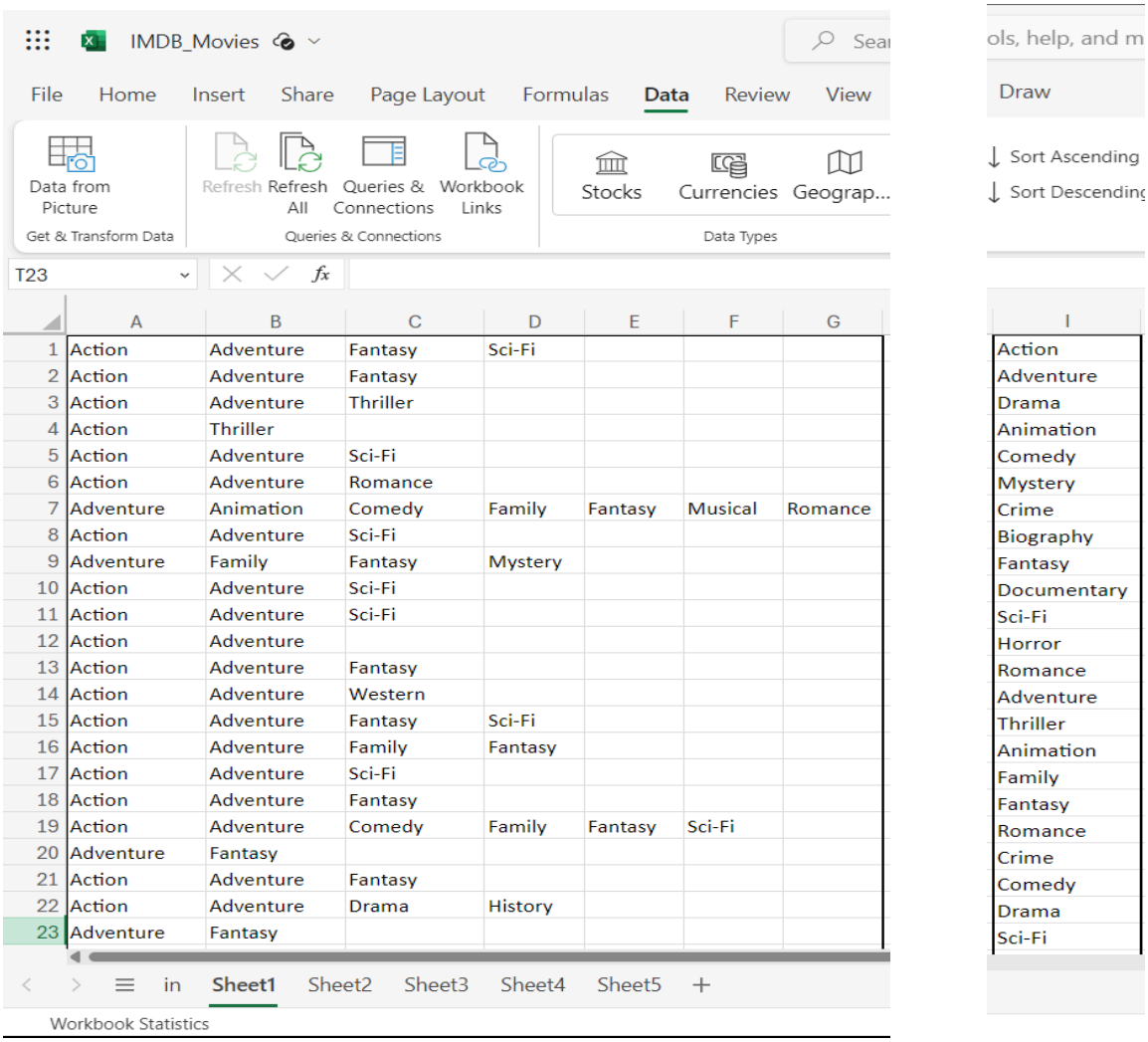
Loading and cleaning the data

- Cleaning the data is done by using the filter option to filter out and “blank” values and delete the row containing them
- Number of rows before cleaning = 5044
  Number of rows after cleaning = 4171



Task 1 : Determine the most common genres of movies in the dataset. Then, for each genre, calculate descriptive statistics of the IMDB scores.

Splitting the various genres into different columns using the split text into column function and finding the the unique values from columns A, B, C, D, E, F, G and compiling them in column I



	A	B	C	D	E	F	G
1	Action	Adventure	Fantasy	Sci-Fi			
2	Action	Adventure	Fantasy				
3	Action	Adventure	Thriller				
4	Action	Thriller					
5	Action	Adventure	Sci-Fi				
6	Action	Adventure	Romance				
7	Adventure	Animation	Comedy	Family	Fantasy	Musical	Romance
8	Action	Adventure	Sci-Fi				
9	Adventure	Family	Fantasy	Mystery			
10	Action	Adventure	Sci-Fi				
11	Action	Adventure	Sci-Fi				
12	Action	Adventure					
13	Action	Adventure	Fantasy				
14	Action	Adventure	Western				
15	Action	Adventure	Fantasy	Sci-Fi			
16	Action	Adventure	Family	Fantasy			
17	Action	Adventure	Sci-Fi				
18	Action	Adventure	Fantasy				
19	Action	Adventure	Comedy	Family	Fantasy	Sci-Fi	
20	Adventure	Fantasy					
21	Action	Adventure	Fantasy				
22	Action	Adventure	Drama	History			
23	Adventure	Fantasy					

I
Action
Adventure
Drama
Animation
Comedy
Mystery
Crime
Biography
Fantasy
Documentary
Sci-Fi
Horror
Romance
Adventure
Thriller
Animation
Family
Fantasy
Romance
Crime
Comedy
Drama
Sci-Fi

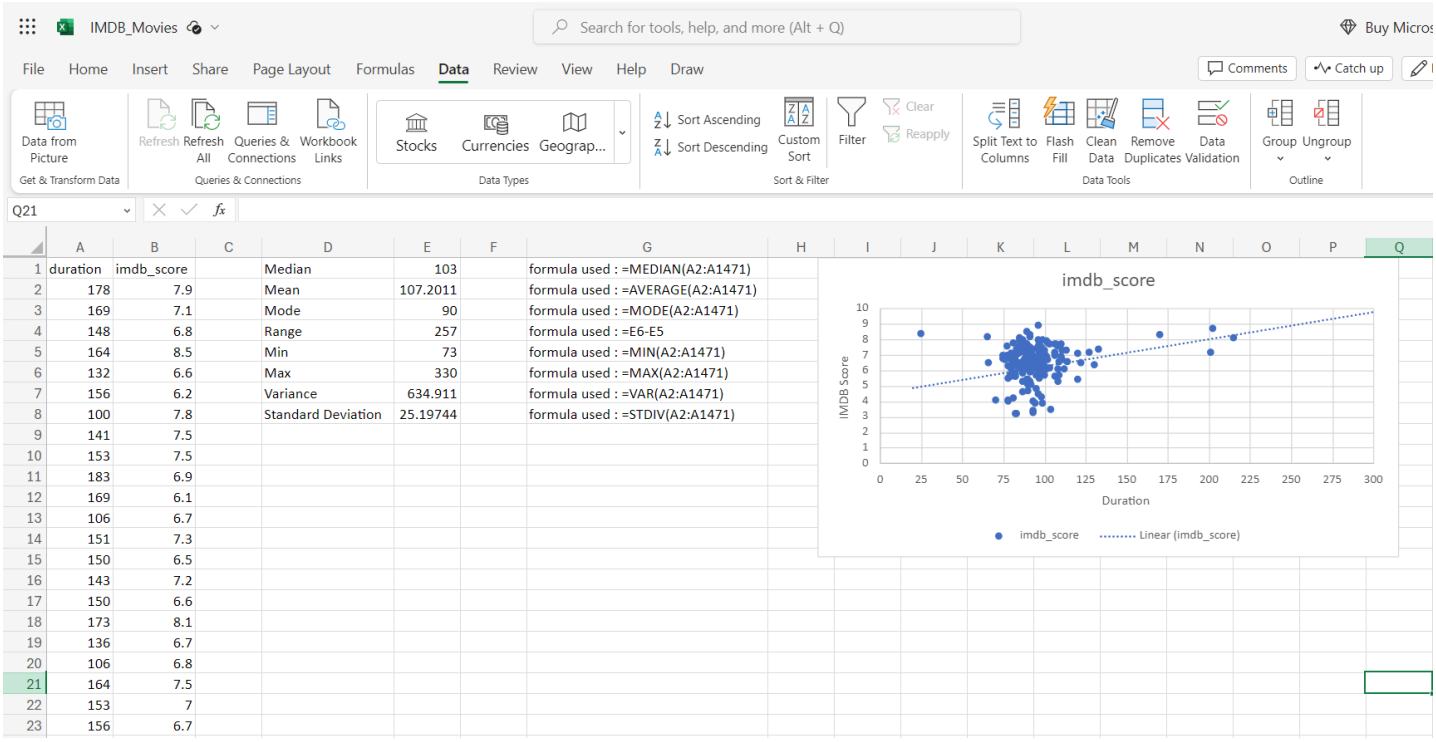
Finding mean, median, mode, max, min and standard deviation using the inbuilt excel formulas

Formula for mean : =AVERAGEIF(K2:K1471, K2, L2:L1471)
Formula for median : =MEDIAN(IF(K2:K1471, K2, L2:L1471))
Formula for mode : =MODE(IF(K2:K1471, K2, L2:L1471))
Formula for max : =MAX(IF(K2:K1471, K2, L2:L1471))
Formula for min : =MIN(IF(K2:K1471, K2, L2:L1471))
Formula for variance : =VAR(IF(K2:K1471, K2, L2:L1471))
Formula for standard deviation : =STDEV(IF(K2:K1471, K2, L2:L1471))

Genre	Count	mean	median	mode	max	min	variance	standard deviation
Action	1043	6.290619	6.35	6.6	9	2.1	1.077487	1.038021
Adventure	842	6.555291	6.6	6.7	8.9	2.3	1.230883	1.109452
Drama	2106	6.814429	6.9	6.7	9.3	2.1	0.804202	0.896773
Animation	205	6.763043	6.8	6.7	8.6	2.8	0.977323	0.988596
Comedy	1565	6.946435	6.3	6.7	8.8	1.9	1.083204	1.040771
Mystery	425	6.608333	6.5	6.6	8.6	3.1	1.037044	1.018353
Crime	769	6.944788	6.6	6.6	9.3	2.4	0.964087	0.981879
Biography	259	7.153846	7.2	7	8.9	4.5	0.500632	0.707554
Fantasy	545	105.0879	6.4	6.7	8.9	2.2	1.290507	1.136005
Documentary	51	6.914286	7.2	6.6	8.4	1.6	1.427424	1.194749
Sci-Fi	551	5.867548	6.4	6.7	8.8	1.9	1.343345	1.159028
Horror	477	5.850909	6	5.9	8.6	2.3	0.996767	0.998382
Romance	935	5.872262	6.5	6.5	8.5	2.1	0.931579	0.965184
Thriller	1226	6.324647	6.4	6.5	9	2.7	0.941511	0.970315
Family	467	6.937267	6.3	5.4	8.6	1.9	1.347412	1.160781
Western	81	5.482568	6.75	6	8.9	4.1	0.974065	0.986947
History	180	6.879993	7.2	7.7	8.9	5.5	0.458758	0.677317
Musical	109	6.7	6.7	7.1	8.5	2.1	1.30185	1.140987
Music	163	5.987457	6.5	6.5	8.5	1.6	1.440547	1.200228
War	181	7.234957	7.1	7.1	8.6	4.3	0.651705	0.807282
Sport	156	6.035858	6.8	7.2	8.4	2	1.083722	1.04102

Task 2 : Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.

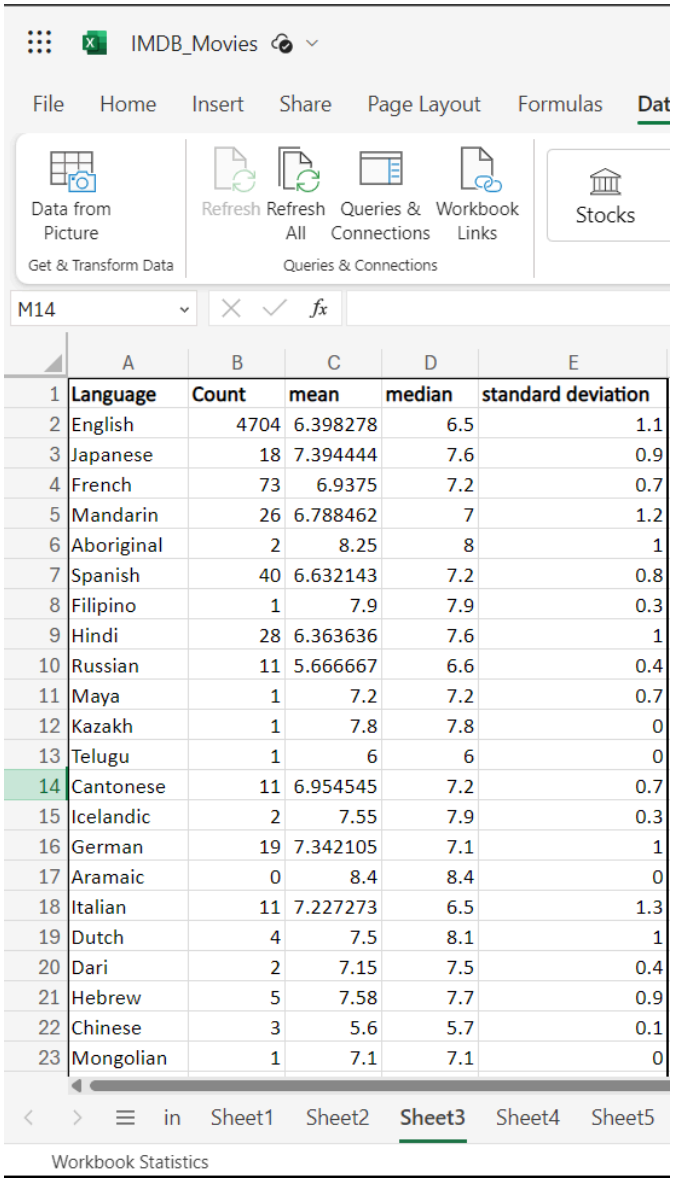
Therefore we can analyze that movies with an average duration between 75-100 mins have higher IMDB scores compared to the others



	A	B	C	D	E	F	G
1	duration	imdb_score		Median	103		formula used : =MEDIAN(A2:A1471)
2	178	7.9		Mean	107.2011		formula used : =AVERAGE(A2:A1471)
3	169	7.1		Mode	90		formula used : =MODE(A2:A1471)
4	148	6.8		Range	257		formula used : =E6-E5
5	164	8.5		Min	73		formula used : =MIN(A2:A1471)
6	132	6.6		Max	330		formula used : =MAX(A2:A1471)
7	156	6.2		Variance	634.911		formula used : =VAR(A2:A1471)
8	100	7.8		Standard Deviation	25.19744		formula used : =STDIV(A2:A1471)
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					

Task 3 : Determine the most common languages used in movies and analyze their impact on the IMDB score using descriptive statistics.

Formula used : =AVERAGEIF()
Formula used : =MEDIAN(IF())
Formula used : =STDIV(IF())



	A	B	C	D	E
1	**Language**	**Count**	**mean**	**median**	**standard deviation**
2	English	4704	6.398278	6.5	1.1
3	Japanese	18	7.394444	7.6	0.9
4	French	73	6.9375	7.2	0.7
5	Mandarin	26	6.788462	7	1.2
6	Aboriginal	2	8.25	8	1
7	Spanish	40	6.632143	7.2	0.8
8	Filipino	1	7.9	7.9	0.3
9	Hindi	28	6.363636	7.6	1
10	Russian	11	5.666667	6.6	0.4
11	Maya	1	7.2	7.2	0.7
12	Kazakh	1	7.8	7.8	0
13	Telugu	1	6	6	0
14	Cantonese	11	6.954545	7.2	0.7
15	Icelandic	2	7.55	7.9	0.3
16	German	19	7.342105	7.1	1
17	Aramaic	0	8.4	8.4	0
18	Italian	11	7.227273	6.5	1.3
19	Dutch	4	7.5	8.1	1
20	Dari	2	7.15	7.5	0.4
21	Hebrew	5	7.58	7.7	0.9
22	Chinese	3	5.6	5.7	0.1
23	Mongolian	1	7.1	7.1	0

Task 4 : Identify the top directors based on their average IMDB score and analyze their contribution to the success of movies using percentile calculations.

Making a list of all the movie directors and their average IMDB score

Formula used : =UNIQUE(A2:A1471)

Formula used : =AVERAGEIF(A$2:A$1471, D2, B$2:B$1471)

director_name	imdb_score
James Cameron	7.9
Gore Verbinski	7.1
Sam Mendes	6.8
Christopher Nolan	8.5
Andrew Stanton	6.6
Sam Raimi	6.2
Nathan Greno	7.8
Joss Whedon	7.5
David Yates	7.5
Zack Snyder	6.9
Bryan Singer	6.1
Marc Forster	6.7
Gore Verbinski	7.3
Gore Verbinski	6.5
Zack Snyder	7.2
Andrew Adamson	6.6
Joss Whedon	8.1
Rob Marshall	6.7
Barry Sonnenfeld	6.8
Peter Jackson	7.5
Marc Webb	7
Ridley Scott	6.7

Director	avg imdb score
James Cameron	7.05
Gore Verbinski	7.05
Sam Mendes	7.34
Christopher Nolan	8.414285714
Andrew Stanton	7.733333333
Sam Raimi	6.74
Nathan Greno	7.8
Joss Whedon	7.925
David Yates	7.05
Zack Snyder	7.1
Bryan Singer	7.1
Marc Forster	6.85
Andrew Adamson	7.15
Rob Marshall	6.45
Barry Sonnenfeld	6.3
Peter Jackson	7.957142857
Marc Webb	6.85
Ridley Scott	7.125
Chris Weitz	5.35
Anthony Russo	7.2
Peter Berg	6.74
Colin Trevorrow	7

Finding the top 10 percentile directors

Formula used : =PERCENTILE(E2:E1471, 0.90)

90 percentile		Top director	imdb score
8.3		John Blanchard	9.5
		Mitchell Altieri	8.7
		Sadyk Sher-Niyaz	8.7
		Cary Bell	8.7
		Mike Mayhall	8.6
		Charles Chaplin	8.6
		Raja Menon	8.5
		Damien Chazelle	8.5
		Majid Majidi	8.5
		Sergio Leone	8.475
		Christopher Nolan	8.425
		S.S. Rajamouli	8.4
		Moustapha Akkad	8.4
		Richard Marquand	8.4
		Catherine Owens	8.4
		Rakeysh Omprakash Mehra	8.4
		Jay Oliva	8.4
		Robert Mulligan	8.4
		Asghar Farhadi	8.4
		Marius A. Markevicius	8.4
		Bill Melendez	8.4
		Lee Unkrich	8.3

Task 5 : Analyze the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.

We can find the profit by subtracting the budget from the gross earnings of the movie. The data is then sorted in descending order

Therefore the movie with the highest profit margin is Avatar

Therefore the correlation between movie budget and gross earnings is 0.54%

movie_title	budget	gross	profit		max profit		Correlation
AvatarÂ	237000000	760505847	523505847		523505847		0.545498079
Jurassic WorldÂ	150000000	652177271	502177271				
TitanicÂ	200000000	658672302	458672302				
The AvengersÂ	220000000	623279547	403279547				
The AvengersÂ	220000000	623279547	403279547				
The Lion KingÂ	45000000	422783777	377783777				
Star Wars: Episode I - The Phantom MenaceÂ	115000000	474544677	359544677				
The Dark KnightÂ	185000000	533316061	348316061				
The Hunger GamesÂ	78000000	407999255	329999255				
DeadpoolÂ	58000000	363024263	305024263				
The Hunger Games: Catching FireÂ	130000000	424645577	294645577				
Jurassic ParkÂ	63000000	356784000	293784000				
Despicable Me 2Â	76000000	368049635	292049635				
American SniperÂ	58800000	350123553	291323553				
Finding NemoÂ	94000000	380838870	286838870				
Shrek 2Â	150000000	436471036	286471036				
The Lord of the Rings: The Return of the KingÂ	94000000	377019252	283019252				
Star Wars: Episode VI - Return of the JediÂ	32500000	309125409	276625409				
Forrest GumpÂ	55000000	329691196	274691196				
Star Wars: Episode III - Revenge of the SithÂ	113000000	380262555	267262555				
Spider-ManÂ	139000000	403706375	264706375				
MinionsÂ	74000000	336029560	262029560				

Result

Hence we were able to leverage Microsoft Excel’s data analytics fuctionalities for IMDB Movie Analysis and extracted the required results