PROJECT

IMDB MOVIE ANALYSIS

The project is about finding out the various insights in IMDB Movies dataset. We analyze this data and some following questions:

1. Analyze the distribution of movie genres and their impact on the IMDB score.

2. Analyze the distribution of movie durations and its impact on the IMDB score.

3. Examine the distribution of movies based on their language.

4. Influence of directors on movie ratings.

5. Explore the relationship between movie budgets and their financial success

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.

IMDB registered users can cast a vote (from 1 to 10) on every released title in the database. Individual votes are then aggregated and summarized as a single IMDB ratings, this rating describes the popularity of a movie in the public.

CLEANING THE DATA

First, I analysed the dataset in MS Excel to find out that there are 5043 rows, 28 columns and 1 header row containing the column names.

Then, after looking at the given questions, we find out that most columns are not required to find out the solution. So, we remove the following columns: color, num_critic_for_reviews, director_facebook_likes, actor_3_facebook_likes, actor_2_name, actor_1_facebook_likes, num_voted_users, cast_total_facebook_likes, actor_3_name, facenumber_in_poster, plot_keywords, movie_imdb_link, num_user_for_reviews, content_rating, actor_2_facebook_likes, aspect_ratio, movie_facebook_likes. After that I cleared all the rows containing any null values as we don't need them. Now we are left with 3884 rows on the basis of which we made our analysis. After that I removed the movies that were duplicate using remove duplicates function.

APPROACH

- When conducting a IMDB Movies analysis project in Excel, I defined the objectives clearly.
- Then I collected the data by downloading the files provided to us.
- Then I cleaned the data and organized it.
- Then I used various formulas to find solutions to various questions put to us by management.
- Finally I analysed the solutions and represented them in graphs and drawn the conclusions and made a report for the management.

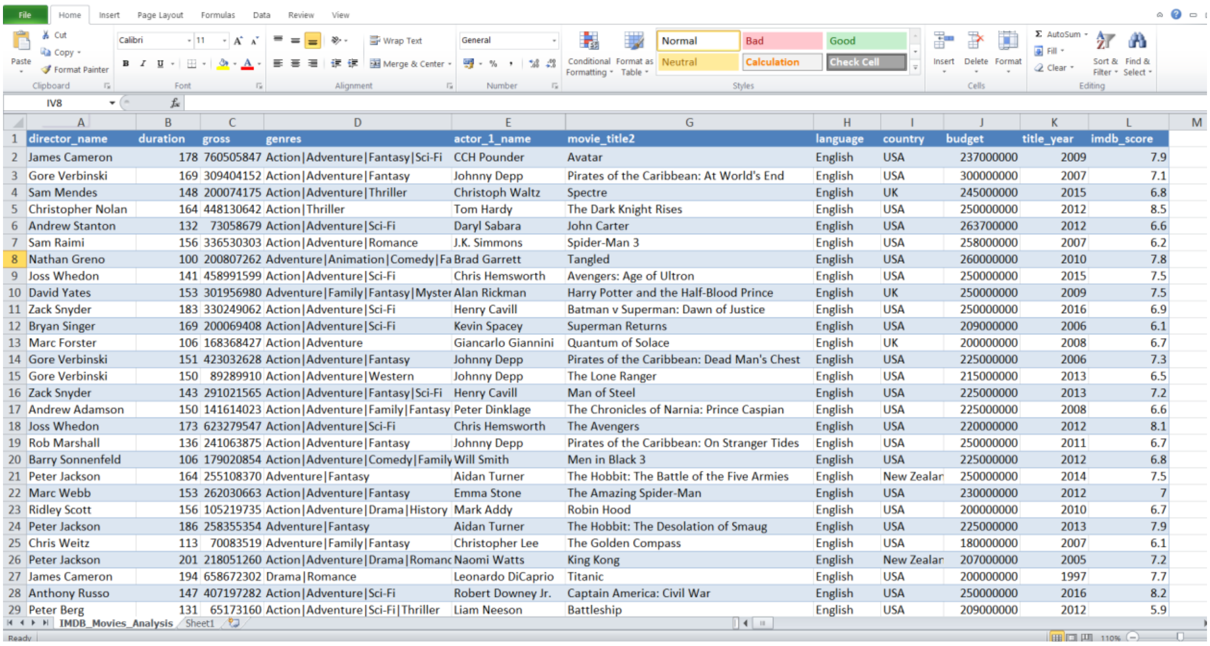
TECH-STACK USED

For this project I used Microsoft Excel to run my queries. Microsoft Excel is a spreadsheet developed by Microsoft for Windows, MacOS, Android and iOS. It features calculation or computation capabilities, graphing tools, pivot tables, and a macro programming language called Visual Basic for Applications (VBA). Excel forms part of the Microsoft Office suite of software.

I used the Excel sheet provided and ran multiple functions to get the desired answers.

This project helped me in understanding the Excel Table at a much detailed manner and helped to improve my strength in extracting data from tables and visualize it in the forms of different graphs.

DATASHEET

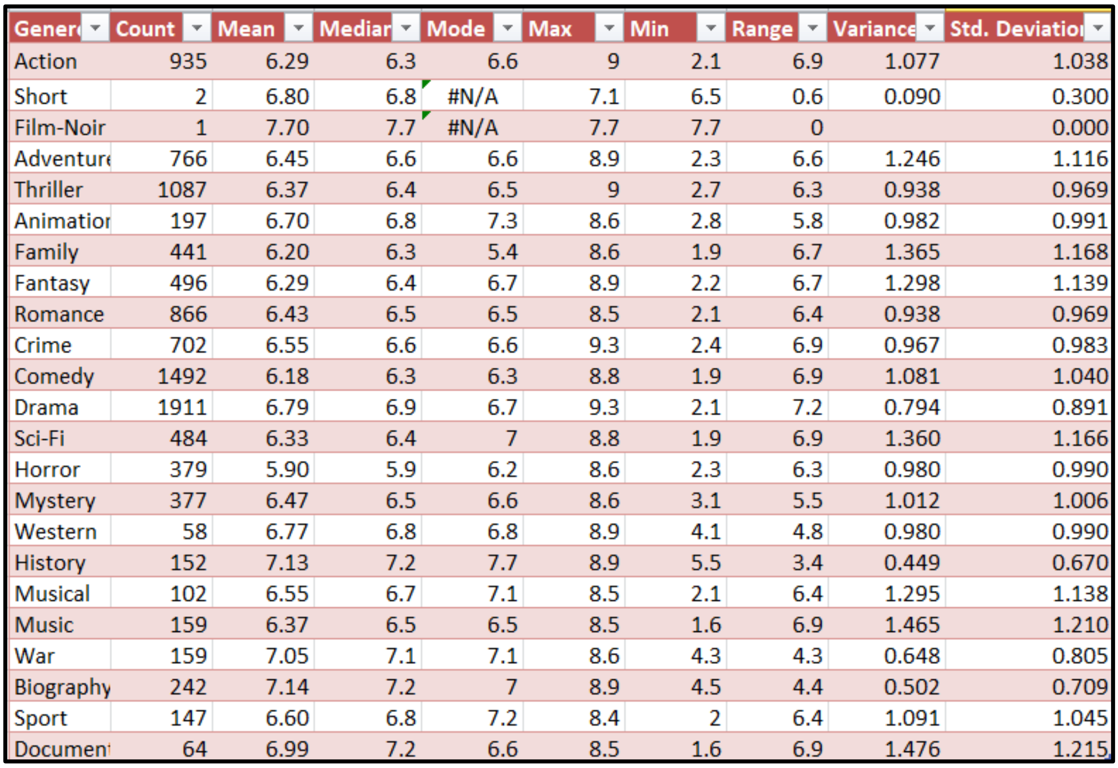


	A	B	C	D	E	G	H	I	J	K	L
1	director_name	duration	gross	genres	actor_1_name	movie_title2	language	country	budget	title_year	imdb_score
2	James Cameron	178	760505847	Action|Adventure|Fantasy|Sci-Fi	CCH Pounder	Avatar	English	USA	237000000	2009	7.9
3	Gore Verbinski	169	309404152	Action|Adventure|Fantasy	Johnny Depp	Pirates of the Caribbean: At World's End	English	USA	300000000	2007	7.1
4	Sam Mendes	148	200074175	Action|Adventure|Thriller	Christoph Waltz	Spectre	English	UK	245000000	2015	6.8
5	Christopher Nolan	164	448130642	Action|Thriller	Tom Hardy	The Dark Knight Rises	English	USA	250000000	2012	8.5
6	Andrew Stanton	132	73058679	Action|Adventure|Sci-Fi	Daryl Sabara	John Carter	English	USA	263700000	2012	6.6
7	Sam Raimi	156	336530303	Action|Adventure|Romance	J.K. Simmons	Spider-Man 3	English	USA	258000000	2007	6.2
8	Nathan Greno	100	200807262	Adventure|Animation|Comedy|Fa	Brad Garrett	Tangled	English	USA	260000000	2010	7.8
9	Joss Whedon	141	458991599	Action|Adventure|Sci-Fi	Chris Hemsworth	Avengers: Age of Ultron	English	USA	250000000	2015	7.5
10	David Yates	153	301956980	Adventure|Family|Fantasy|Myster	Alan Rickman	Harry Potter and the Half-Blood Prince	English	UK	250000000	2009	7.5
11	Zack Snyder	183	330249062	Action|Adventure|Sci-Fi	Henry Cavill	Batman v Superman: Dawn of Justice	English	USA	250000000	2016	6.9
12	Bryan Singer	169	200069408	Action|Adventure|Sci-Fi	Kevin Spacey	Superman Returns	English	USA	209000000	2006	6.1
13	Marc Forster	106	168368427	Action|Adventure	Giancarlo Giannini	Quantum of Solace	English	UK	200000000	2008	6.7
14	Gore Verbinski	151	423032628	Action|Adventure|Fantasy	Johnny Depp	Pirates of the Caribbean: Dead Man's Chest	English	USA	225000000	2006	7.3
15	Gore Verbinski	150	89289910	Action|Adventure|Western	Johnny Depp	The Lone Ranger	English	USA	215000000	2013	6.5
16	Zack Snyder	143	291021565	Action|Adventure|Fantasy|Sci-Fi	Henry Cavill	Man of Steel	English	USA	225000000	2013	7.2
17	Andrew Adamson	150	141614023	Action|Adventure|Family|Fantasy	Peter Dinklage	The Chronicles of Narnia: Prince Caspian	English	USA	225000000	2008	6.6
18	Joss Whedon	173	623279547	Action|Adventure|Sci-Fi	Chris Hemsworth	The Avengers	English	USA	220000000	2012	8.1
19	Rob Marshall	136	241063875	Action|Adventure|Fantasy	Johnny Depp	Pirates of the Caribbean: On Stranger Tides	English	USA	250000000	2011	6.7
20	Barry Sonnenfeld	106	179020854	Action|Adventure|Comedy|Family	Will Smith	Men in Black 3	English	USA	225000000	2012	6.8
21	Peter Jackson	164	255108370	Adventure|Fantasy	Aidan Turner	The Hobbit: The Battle of the Five Armies	English	New Zealar	250000000	2014	7.5
22	Marc Webb	153	262030663	Action|Adventure|Fantasy	Emma Stone	The Amazing Spider-Man	English	USA	230000000	2012	7
23	Ridley Scott	156	105219735	Action|Adventure|Drama|History	Mark Addy	Robin Hood	English	USA	200000000	2010	6.7
24	Peter Jackson	186	258355354	Adventure|Fantasy	Aidan Turner	The Hobbit: The Desolation of Smaug	English	USA	225000000	2013	7.9
25	Chris Weitz	113	70083519	Adventure|Family|Fantasy	Christopher Lee	The Golden Compass	English	USA	180000000	2007	6.1
26	Peter Jackson	201	218051260	Action|Adventure|Drama|Romanc	Naomi Watts	King Kong	English	New Zealar	207000000	2005	7.2
27	James Cameron	194	658672302	Drama|Romance	Leonardo DiCaprio	Titanic	English	USA	200000000	1997	7.7
28	Anthony Russo	147	407197282	Action|Adventure|Sci-Fi	Robert Downey Jr.	Captain America: Civil War	English	USA	250000000	2016	8.2
29	Peter Berg	131	65173160	Action|Adventure|Sci-Fi|Thriller	Liam Neeson	Battleship	English	USA	209000000	2012	5.9

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

Our 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.*

Results:



Gener	Count	Mean	Mediar	Mode	Max	Min	Range	Variance	Std. Deviatio
Action	935	6.29	6.3	6.6	9	2.1	6.9	1.077	1.038
Short	2	6.80	6.8	#N/A	7.1	6.5	0.6	0.090	0.300
Film-Noir	1	7.70	7.7	#N/A	7.7	7.7	0		0.000
Adventur	766	6.45	6.6	6.6	8.9	2.3	6.6	1.246	1.116
Thriller	1087	6.37	6.4	6.5	9	2.7	6.3	0.938	0.969
Animatior	197	6.70	6.8	7.3	8.6	2.8	5.8	0.982	0.991
Family	441	6.20	6.3	5.4	8.6	1.9	6.7	1.365	1.168
Fantasy	496	6.29	6.4	6.7	8.9	2.2	6.7	1.298	1.139
Romance	866	6.43	6.5	6.5	8.5	2.1	6.4	0.938	0.969
Crime	702	6.55	6.6	6.6	9.3	2.4	6.9	0.967	0.983
Comedy	1492	6.18	6.3	6.3	8.8	1.9	6.9	1.081	1.040
Drama	1911	6.79	6.9	6.7	9.3	2.1	7.2	0.794	0.891
Sci-Fi	484	6.33	6.4	7	8.8	1.9	6.9	1.360	1.166
Horror	379	5.90	5.9	6.2	8.6	2.3	6.3	0.980	0.990
Mystery	377	6.47	6.5	6.6	8.6	3.1	5.5	1.012	1.006
Western	58	6.77	6.8	6.8	8.9	4.1	4.8	0.980	0.990
History	152	7.13	7.2	7.7	8.9	5.5	3.4	0.449	0.670
Musical	102	6.55	6.7	7.1	8.5	2.1	6.4	1.295	1.138
Music	159	6.37	6.5	6.5	8.5	1.6	6.9	1.465	1.210
War	159	7.05	7.1	7.1	8.6	4.3	4.3	0.648	0.805
Biography	242	7.14	7.2	7	8.9	4.5	4.4	0.502	0.709
Sport	147	6.60	6.8	7.2	8.4	2	6.4	1.091	1.045
Documen	64	6.99	7.2	6.6	8.5	1.6	6.9	1.476	1.215

FORMULA USED:

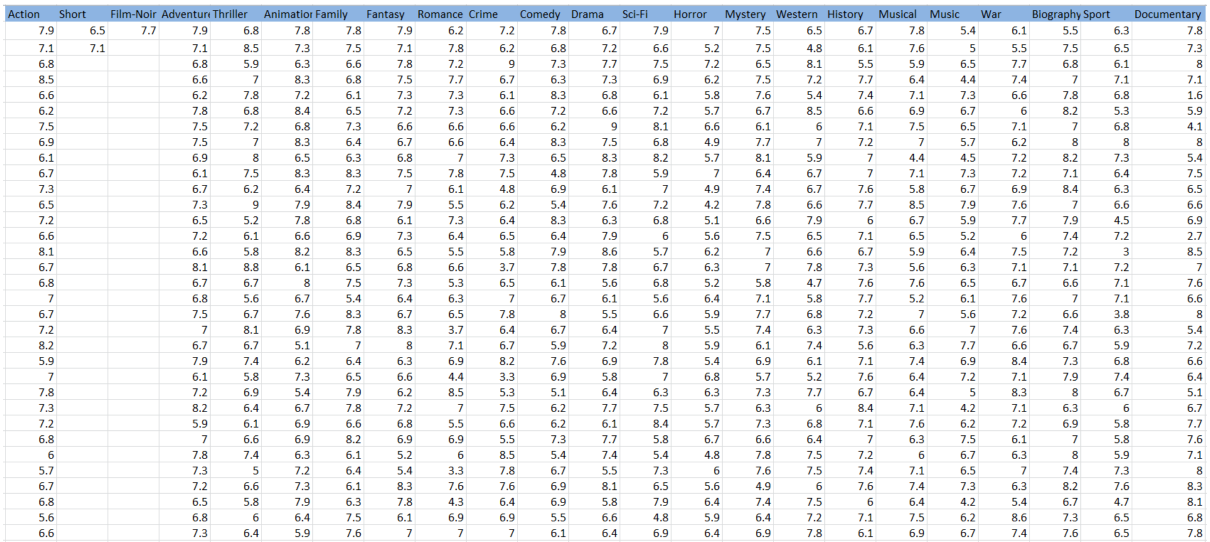
First of all I used text to columns function in "Data tab" to separate the multiple genres being written in a single column - to different columns. It has given me 7 columns with separated genres. Now I have created a separate table containing names of genres in first rows and IMDB ratings list corresponding to those genres.

Using the following formula:

=IF(Table2[@Column1]=AA$1,Table1[@[imdb_score]],IF(Table2[@Column2]=AA$1,Table1[@[imdb_score]],IF(Table2[@Column3]=AA$1,Table1[@[imdb_score]],IF(Table2[@Column4]=AA$1,Table1[@[imdb_score]],IF(Table2[@Column5]=AA$1,Table1[@[imdb_score]],IF(Table2[@Column6]=AA$1,Table1[@[imdb_score]],IF(Table2[@Column7]=AA$1,Table1[@[imdb_score]])))))))

Then I have replaced all the false values with blanks and removed the blanks. Similarly, for all the genres.

This has given a table as follows:



Action	Short	Film-Noir	Adventur	Thriller	Animatior	Family	Fantasy	Romance	Crime	Comedy	Drama	Sci-Fi	Horror	Mystery	Western	History	Musical	Music	War	Biography	Sport	Documentary
7.9	6.5	7.7	7.9	6.8	7.8	7.8	7.9	6.2	7.2	7.8	6.7	7.9	7	7.5	6.5	6.7	7.8	5.4	6.1	5.5	6.3	7.8
7.1	7.1		7.1	8.5	7.3	7.5	7.1	7.8	6.2	6.8	7.2	6.6	5.2	7.5	4.8	6.1	7.6	5	5.5	7.5	6.5	7.3
6.8			6.8	5.9	6.3	6.6	7.8	7.2	9	7.3	7.7	7.5	7.2	6.5	8.1	5.5	5.9	6.5	7.7	6.8	6.1	8
8.5			6.6	7	8.3	6.8	7.5	7.7	6.7	6.3	7.3	6.9	6.2	7.5	7.2	7.7	6.4	4.4	7.4	7	7.1	7.1
6.6			6.2	7.8	7.2	6.1	7.3	7.3	6.1	8.3	6.8	6.1	5.8	7.6	5.4	7.4	7.1	7.3	6.6	7.8	6.8	1.6
6.2			7.8	6.8	8.4	6.5	7.2	7.3	6.6	7.2	6.6	7.2	5.7	6.7	8.5	6.6	6.9	6.7	6	8.2	5.3	5.9
7.5			7.5	7.2	6.8	7.3	6.6	6.6	6.6	6.2	9	8.1	6.6	6.1	6	7.1	7.5	6.5	7.1	7	6.8	4.1
6.9			7.5	7	8.3	6.4	6.7	6.6	6.4	8.3	7.5	6.8	4.9	7.7	7	7.2	7	5.7	6.2	8	8	8
6.1			6.9	8	6.5	6.3	6.8	7	7.3	6.5	8.3	8.2	5.7	8.1	5.9	7	4.4	4.5	7.2	8.2	7.3	5.4
6.7			6.1	7.5	8.3	8.3	7.5	7.8	7.5	4.8	7.8	5.9	7	6.4	6.7	7	7.1	7.3	7.2	7.1	6.4	7.5
7.3			6.7	6.2	6.4	7.2	7	6.1	4.8	6.9	6.1	7	4.9	7.4	6.7	7.6	5.8	6.7	6.9	8.4	6.3	6.5
6.5			7.3	9	7.9	8.4	7.9	5.5	6.2	5.4	7.6	7.2	4.2	7.8	6.6	7.7	8.5	7.9	7.6	7	6.6	6.6
7.2			6.5	5.2	7.8	6.8	6.1	7.3	6.4	8.3	6.3	6.8	5.1	6.6	7.9	6	6.7	5.9	7.7	7.9	4.5	6.9
6.6			7.2	6.1	6.6	6.9	7.3	6.4	6.5	6.4	7.9	6	5.6	7.5	6.5	7.1	6.5	5.2	6	7.4	7.2	2.7
8.1			6.6	5.8	8.2	8.3	6.5	5.5	5.8	7.9	8.6	5.7	6.2	7	6.6	6.7	5.9	6.4	7.5	7.2	3	8.5
6.7			8.1	8.8	6.1	6.5	6.8	6.6	3.7	7.8	7.8	6.7	6.3	7	7.8	7.3	5.6	6.3	7.1	7.1	7.2	7
6.8			6.7	6.7	8	7.5	7.3	5.3	6.5	6.1	5.6	6.8	5.2	5.8	4.7	7.6	7.6	6.5	6.7	6.6	7.1	7.6
7			6.8	5.6	6.7	5.4	6.4	6.3	7	6.7	6.1	5.6	6.4	7.1	5.8	7.7	5.2	6.1	7.6	7	7.1	6.6
6.7			7.5	6.7	7.6	8.3	6.7	6.5	7.8	8	5.5	6.6	5.9	7.7	6.8	7.2	7	5.6	7.2	6.6	3.8	8
7.2			7	8.1	6.9	7.8	8.3	3.7	6.4	6.7	6.4	7	5.5	7.4	6.3	7.3	6.6	7	7.6	7.4	6.3	5.4
8.2			6.7	6.7	5.1	7	8	7.1	6.7	5.9	7.2	8	5.9	6.1	7.4	5.6	6.3	7.7	6.6	6.7	5.9	7.2
5.9			7.9	7.4	6.2	6.4	6.3	6.9	8.2	7.6	6.9	7.8	5.4	6.9	6.1	7.1	7.4	6.9	8.4	7.3	6.8	6.6
7			6.1	5.8	7.3	6.5	6.6	4.4	3.3	6.9	5.8	7	6.8	5.7	5.2	7.6	6.4	7.2	7.1	7.9	7.4	6.4
7.8			7.2	6.9	5.4	7.9	6.2	8.5	5.3	5.1	6.4	6.3	6.3	7.3	7.7	6.7	6.4	5	8.3	8	6.7	5.1
7.3			8.2	6.4	6.7	7.8	7.2	7	7.5	6.2	7.7	7.5	5.7	6.3	6	8.4	7.1	4.2	7.1	6.3	6	6.7
7.2			5.9	6.1	6.9	6.6	6.8	5.5	6.6	6.2	6.1	8.4	5.7	7.3	6.8	7.1	7.6	6.2	7.2	6.9	5.8	7.7
6.8			7	6.6	6.9	8.2	6.9	6.9	5.5	7.3	7.7	5.8	6.7	6.6	6.4	7	6.3	7.5	6.1	7	5.8	7.6
6			7.8	7.4	6.3	6.1	5.2	6	8.5	5.4	7.4	5.4	4.8	7.8	7.5	7.2	6	6.7	6.3	8	5.9	7.1
5.7			7.3	5	7.2	6.4	5.4	3.3	7.8	6.7	5.5	7.3	6	7.6	7.5	7.4	7.1	6.5	7	7.4	7.3	8
6.7			7.2	6.6	7.3	6.1	8.3	7.6	7.6	6.9	8.1	6.5	5.6	4.9	6	7.6	7.4	7.3	6.3	8.2	7.6	8.3
6.8			6.5	5.8	7.9	6.3	7.8	4.3	6.4	6.9	5.8	7.9	6.4	7.4	7.5	6	6.4	4.2	5.4	6.7	4.7	8.1
5.6			6.8	6	6.4	7.5	6.1	6.9	6.9	5.5	6.6	4.8	5.9	6.4	7.2	7.1	7.5	6.2	8.6	7.3	6.5	6.8
6.6			7.3	6.4	5.9	7.6	7	7	7	6.1	6.4	6.9	6.4	6.9	7.8	6.1	6.9	6.7	7.4	7.6	6.5	7.8

The imdb rating for all the genres are now classified.

Then calculates descriptive statistics by using following:

For the count =COUNTIF(Table2,CA2)

For the mean =AVERAGE(BA:BA)

For the median =MEDIAN(BA:BA)

For the mode =MODE.MULT(BA:BA)

For the maximum =MAX(BA:BA)

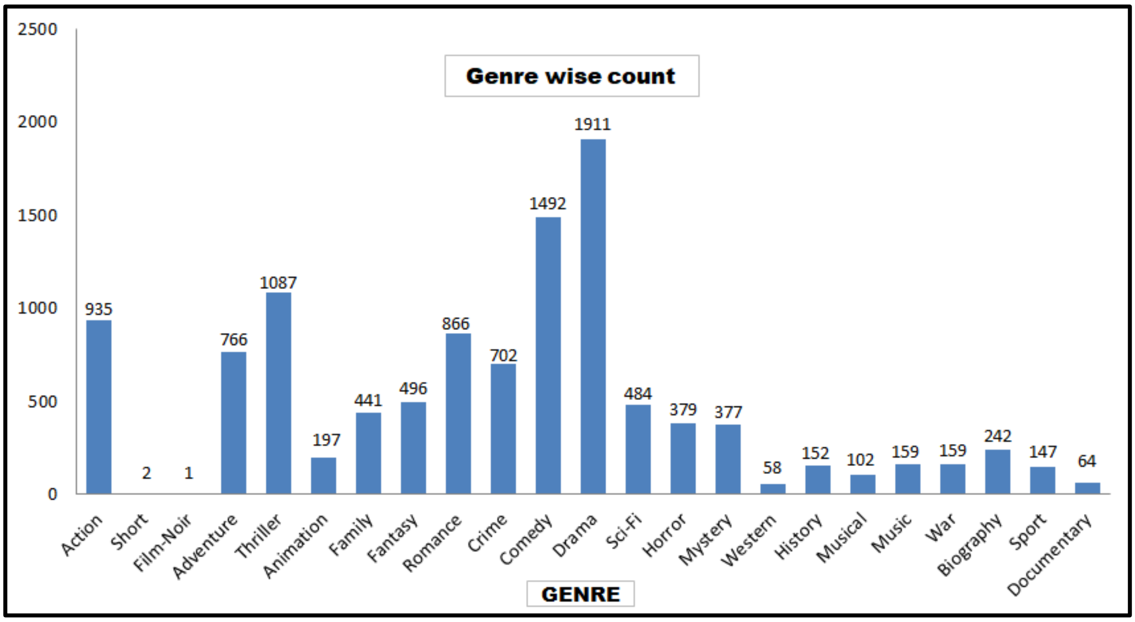
For the minimum =MIN(BA:BA)

For the range = (maximum – minimum) =CF2-CG2

For the variance =VAR.P(BA:BA)

For the std deviation =STDEV.P(BA:BA)

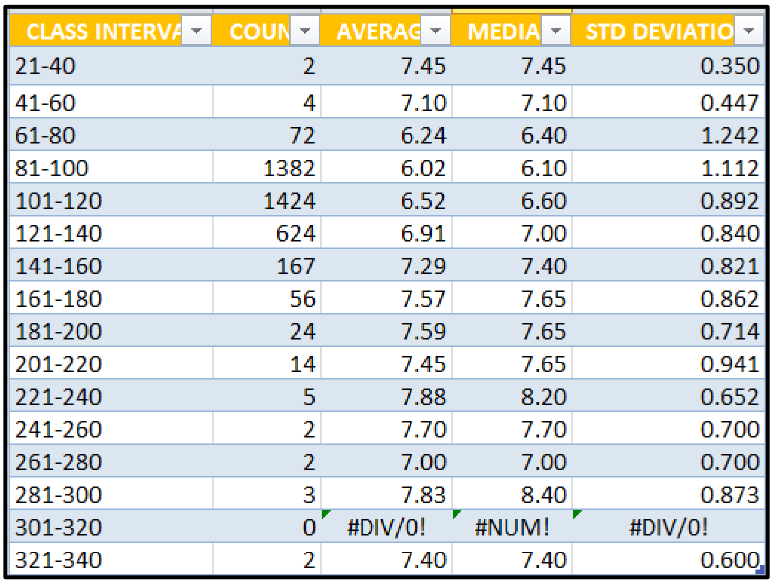
We found that he maximum number of movies belong to Drama genre.



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

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

Results:



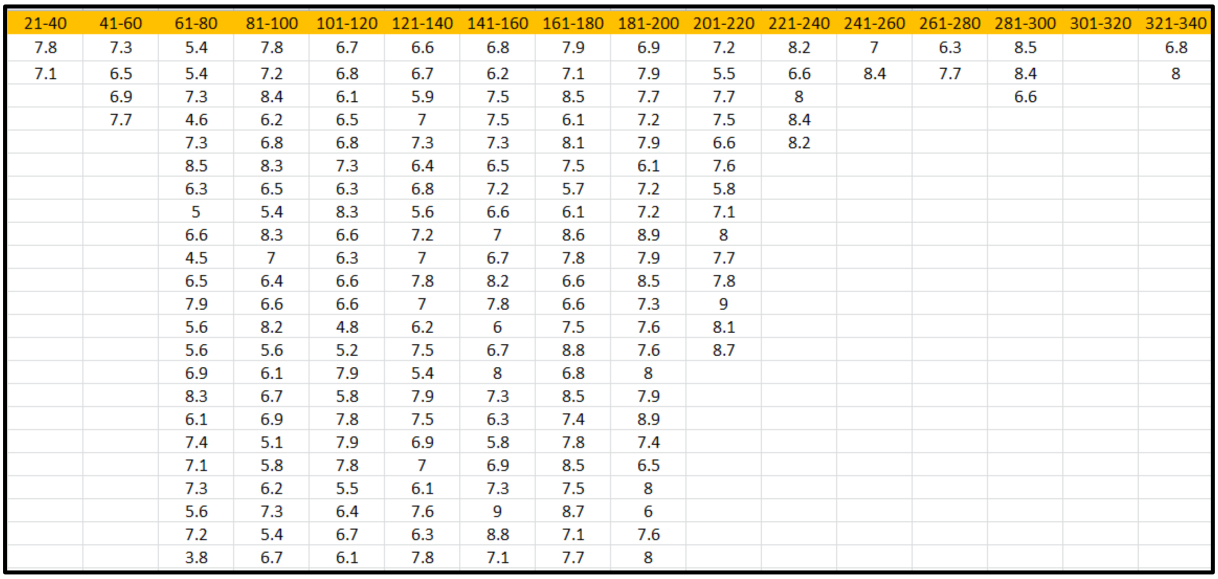
CLASS INTERVA	COUN	AVERAG	MEDIA	STD DEVIATIO
21-40	2	7.45	7.45	0.350
41-60	4	7.10	7.10	0.447
61-80	72	6.24	6.40	1.242
81-100	1382	6.02	6.10	1.112
101-120	1424	6.52	6.60	0.892
121-140	624	6.91	7.00	0.840
141-160	167	7.29	7.40	0.821
161-180	56	7.57	7.65	0.862
181-200	24	7.59	7.65	0.714
201-220	14	7.45	7.65	0.941
221-240	5	7.88	8.20	0.652
241-260	2	7.70	7.70	0.700
261-280	2	7.00	7.00	0.700
281-300	3	7.83	8.40	0.873
301-320	0	#DIV/0!	#NUM!	#DIV/0!
321-340	2	7.40	7.40	0.600

FORMULA USED:

Instead of calculating statistics for each movie I created class intervals to analyse the case better.

In this also I have created a separate table showing class interval in minutes in top rows and list of the IMBD ratings belonging to that class interval by using the following formula : =IF((B2>=21)*(B2<=40),L2) i.e. =if((duration>=21)*(duration<=40),give me IMDB score)

Then I have replaced all the false values with blanks and removed the blanks.



21-40	41-60	61-80	81-100	101-120	121-140	141-160	161-180	181-200	201-220	221-240	241-260	261-280	281-300	301-320	321-340
7.8	7.3	5.4	7.8	6.7	6.6	6.8	7.9	6.9	7.2	8.2	7	6.3	8.5		6.8
7.1	6.5	5.4	7.2	6.8	6.7	6.2	7.1	7.9	5.5	6.6	8.4	7.7	8.4		8
	6.9	7.3	8.4	6.1	5.9	7.5	8.5	7.7	7.7	8			6.6		
	7.7	4.6	6.2	6.5	7	7.5	6.1	7.2	7.5	8.4					
		7.3	6.8	6.8	7.3	7.3	8.1	7.9	6.6	8.2					
		8.5	8.3	7.3	6.4	6.5	7.5	6.1	7.6						
		6.3	6.5	6.3	6.8	7.2	5.7	7.2	5.8						
		5	5.4	8.3	5.6	6.6	6.1	7.2	7.1						
		6.6	8.3	6.6	7.2	7	8.6	8.9	8						
		4.5	7	6.3	7	6.7	7.8	7.9	7.7						
		6.5	6.4	6.6	7.8	8.2	6.6	8.5	7.8						
		7.9	6.6	6.6	7	7.8	6.6	7.3	9						
		5.6	8.2	4.8	6.2	6	7.5	7.6	8.1						
		5.6	5.6	5.2	7.5	6.7	8.8	7.6	8.7						
		6.9	6.1	7.9	5.4	8	6.8	8							
		8.3	6.7	5.8	7.9	7.3	8.5	7.9							
		6.1	6.9	7.8	7.5	6.3	7.4	8.9							
		7.4	5.1	7.9	6.9	5.8	7.8	7.4							
		7.1	5.8	7.8	7	6.9	8.5	6.5							
		7.3	6.2	5.5	6.1	7.3	7.5	8							
		5.6	7.3	6.4	7.6	9	8.7	6							
		7.2	5.4	6.7	6.3	8.8	7.1	7.6							
		3.8	6.7	6.1	7.8	7.1	7.7	8							

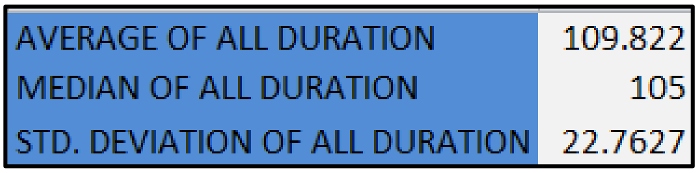
To calculate count: =COUNT(DE2:DE3)

To calculate average: =AVERAGE(DE2:DE3)

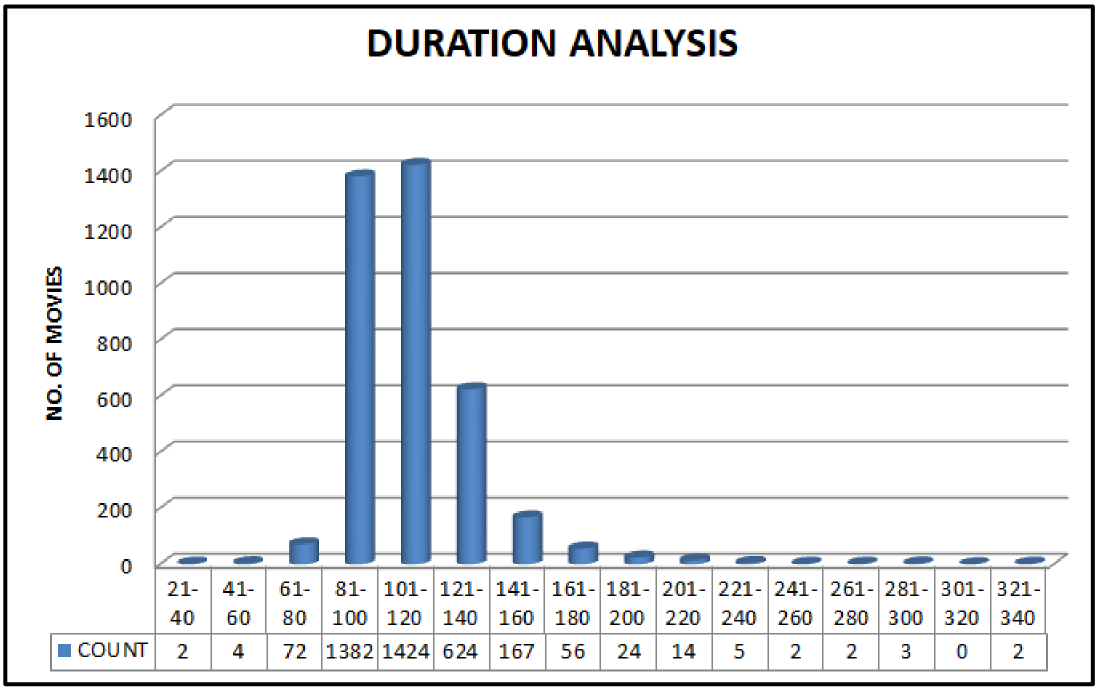
To calculate median: =MEDIAN(DE2:DE3)

To calculate standard deviation: =STDEV.P(DE2:DE3)

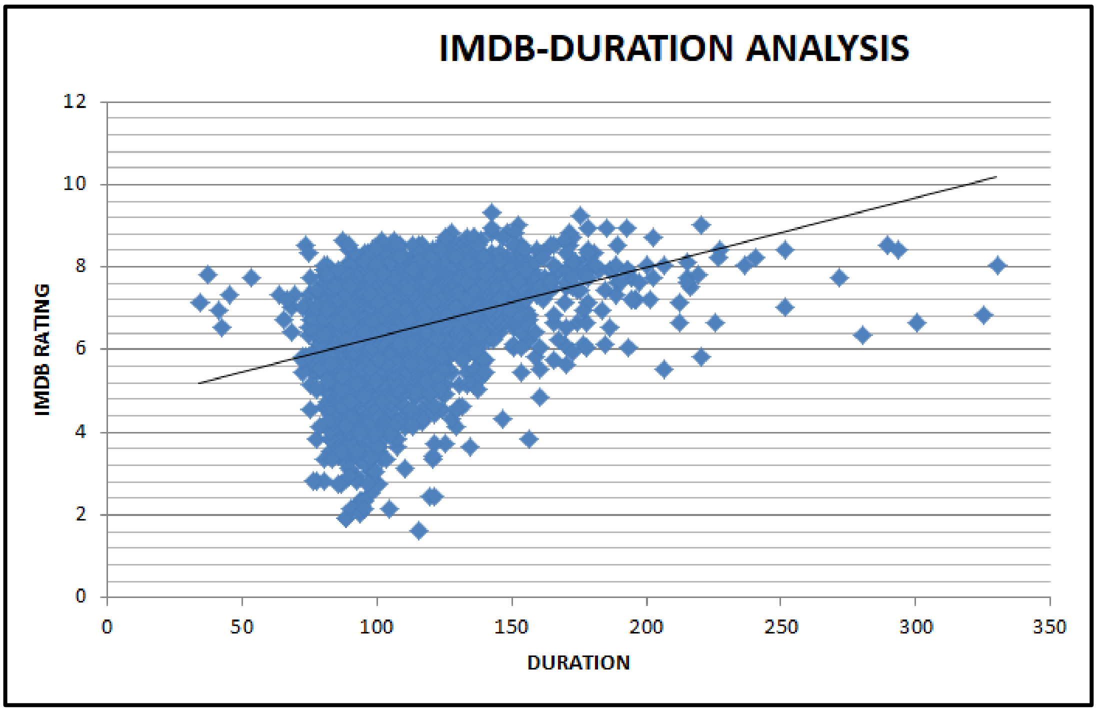
However the averages and median and standard deviation for all the duration is as follows:



AVERAGE OF ALL DURATION	109.822
MEDIAN OF ALL DURATION	105
STD. DEVIATION OF ALL DURATION	22.7627



	21-40	41-60	61-80	81-100	101-120	121-140	141-160	161-180	181-200	201-220	221-240	241-260	261-280	281-300	301-320	321-340
■ COUNT	2	4	72	1382	1424	624	167	56	24	14	5	2	2	3	0	2

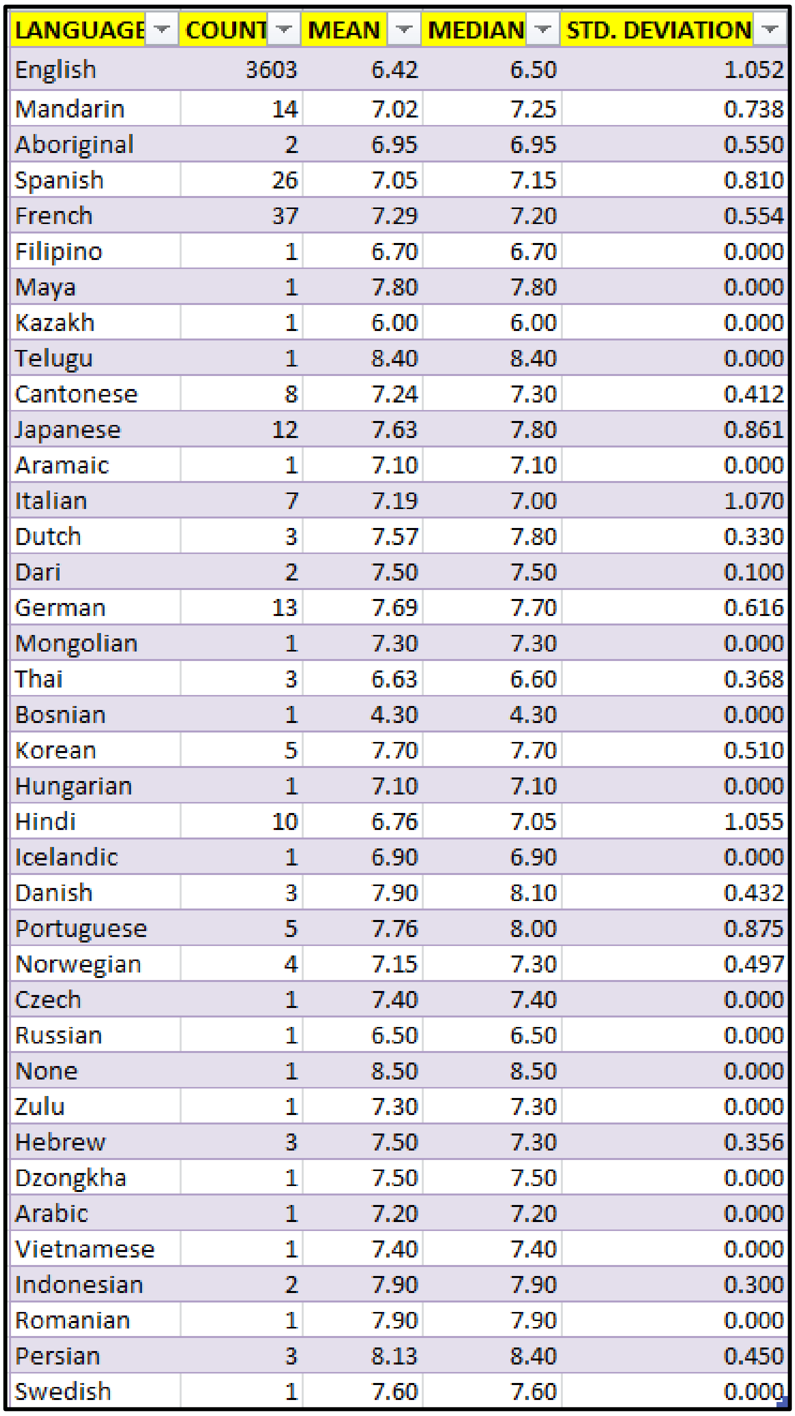


Conclusion: We found that movie with duration 100-120 minutes has maximum number of movies, but the average IMDB is highest for 221-240 minutes. The highest median IMDB is in 281-300 minutes. The trend line shows that most of the IMDB ratings are between 75-150 minutes.

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

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

Results:

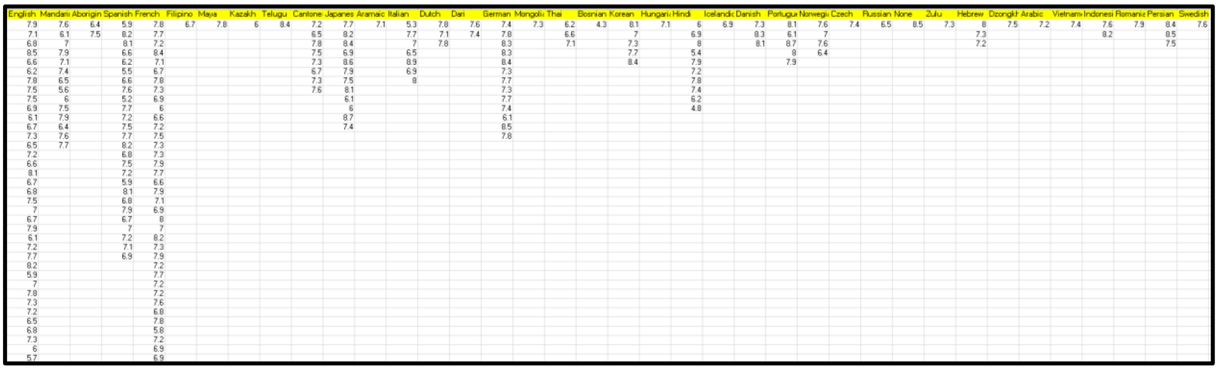


LANGUAGE	COUNT	MEAN	MEDIAN	STD. DEVIATION
English	3603	6.42	6.50	1.052
Mandarin	14	7.02	7.25	0.738
Aboriginal	2	6.95	6.95	0.550
Spanish	26	7.05	7.15	0.810
French	37	7.29	7.20	0.554
Filipino	1	6.70	6.70	0.000
Maya	1	7.80	7.80	0.000
Kazakh	1	6.00	6.00	0.000
Telugu	1	8.40	8.40	0.000
Cantonese	8	7.24	7.30	0.412
Japanese	12	7.63	7.80	0.861
Aramaic	1	7.10	7.10	0.000
Italian	7	7.19	7.00	1.070
Dutch	3	7.57	7.80	0.330
Dari	2	7.50	7.50	0.100
German	13	7.69	7.70	0.616
Mongolian	1	7.30	7.30	0.000
Thai	3	6.63	6.60	0.368
Bosnian	1	4.30	4.30	0.000
Korean	5	7.70	7.70	0.510
Hungarian	1	7.10	7.10	0.000
Hindi	10	6.76	7.05	1.055
Icelandic	1	6.90	6.90	0.000
Danish	3	7.90	8.10	0.432
Portuguese	5	7.76	8.00	0.875
Norwegian	4	7.15	7.30	0.497
Czech	1	7.40	7.40	0.000
Russian	1	6.50	6.50	0.000
None	1	8.50	8.50	0.000
Zulu	1	7.30	7.30	0.000
Hebrew	3	7.50	7.30	0.356
Dzongkha	1	7.50	7.50	0.000
Arabic	1	7.20	7.20	0.000
Vietnamese	1	7.40	7.40	0.000
Indonesian	2	7.90	7.90	0.300
Romanian	1	7.90	7.90	0.000
Persian	3	8.13	8.40	0.450
Swedish	1	7.60	7.60	0.000

FORMULA USED:

In this also first of all I created a table with language names of the top and list of imbd belonging to that language.

I used =IF(Table1[@language]=EO1,Table1[@[imdb_score]])) i.e., =if(language=English, imdb rating) formula to get the imdb rating and Then I have replaced all the false values with blanks and removed the blanks.



English	Mandari	Aborigin	Spanish	French	Filipino	Maya	Kazakh	Telugu	Cantone	Japanes	Aramaic	Italian	Dutch	Dari	German	Mongoli	Thai	Bosnian
7.9	7.6	6.4	5.9	7.8	6.7	7.8	6	8.4	7.2	7.7	7.1	5.3	7.8	7.6	7.4	7.3	6.2	4.3
7.1	6.1	7.5	8.2	7.7					6.5	8.2		7.7	7.1	7.4	7.8		6.6	
6.8	7		8.1	7.2					7.8	8.4		7	7.8		8.3		7.1	
8.5	7.9		6.6	8.4					7.5	6.9		6.5			8.3			
6.6	7.1		6.2	7.1					7.3	8.6		8.9			8.4			
6.2	7.4		5.5	6.7					6.7	7.9		6.9			7.3			
7.8	6.5		6.6	7.8					7.3	7.5		8			7.7			
7.5	5.6		7.6	7.3					7.6	8.1					7.3			
7.5	6		5.2	6.9						6.1					7.7			
6.9	7.5		7.7	6						6					7.4			
6.1	7.9		7.2	6.6						8.7					6.1			
6.7	6.4		7.5	7.2						7.4					8.5			
7.3	7.6		7.7	7.5											7.8			
6.5	7.7		8.2	7.3														
7.2			6.8	7.3														
6.6			7.5	7.9														
8.1			7.2	7.7														
6.7			5.9	6.6														
6.8			8.1	7.9														
7.5			6.8	7.1														
7			7.9	6.9														
6.7			6.7	8														
7.9			7	7														
6.1			7.2	8.2														
7.2			7.1	7.3														
7.7			6.9	7.9														
8.2				7.2														
5.9				7.7														
7				7.2														
7.8				7.2														
7.3				7.6														
7.2				6.8														
6.5				7.8														
6.8				5.8														
7.3				7.2														
6				6.9														
5.7				6.9														

Korean	Hungari	Hindi	Icelandic	Danish	Portugu	Norwegi	Czech	Russian	None	Zulu	Hebrew	Dzongkh	Arabic	Vietnam	Indonesi	Romania	Persian	Swedish
8.1	7.1	6	6.9	7.3	8.1	7.6	7.4	6.5	8.5	7.3	8	7.5	7.2	7.4	7.6	7.9	8.4	7.6
7		6.9		8.3	6.1	7					7.3				8.2		8.5	
7.3		8		8.1	8.7	7.6					7.2						7.5	
7.7		5.4			8	6.4												
8.4		7.9			7.9													
		7.2																
		7.8																
		7.4																
		6.2																
		4.8																

Now calculating mean median std deviation becomes very easy for each language.

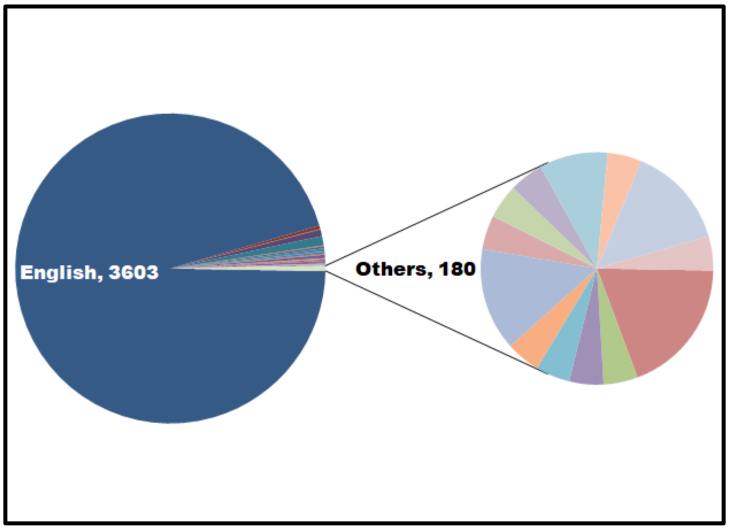
For count : =AVERAGE(GC2:GC3604)

For mean: =AVERAGE(GC2:GC3604)

For median: =MEDIAN(GC2:GC3604)

For standard deviation: =STDEV.P(GC2:GC3604)

Since the proportion of English language is highest, pie chart would be appropriate.

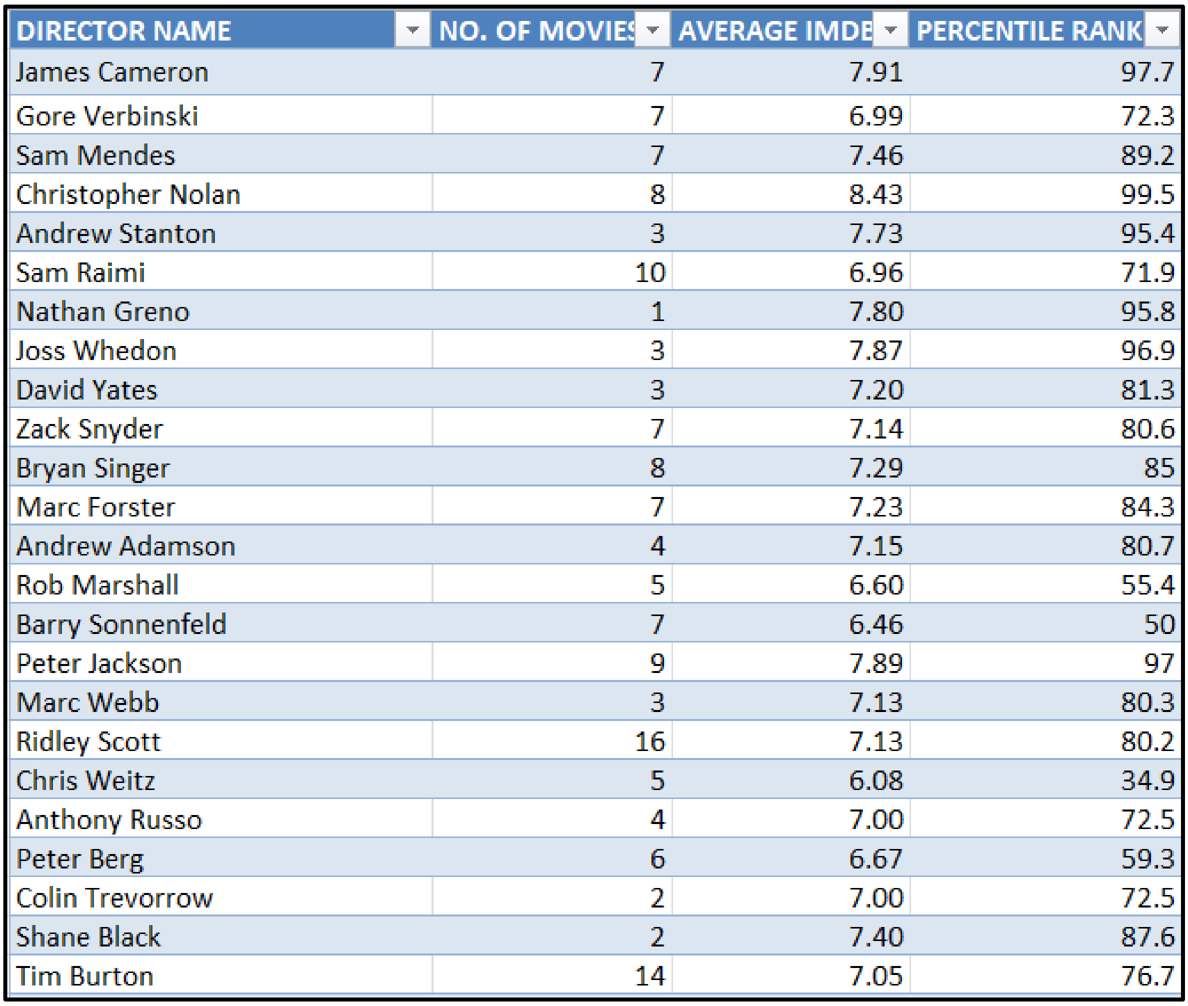


Conclusion: The language in which majority of movies are made is English i.e., 3603. This amounts to 95% of the movies under analysis. The maximum standard deviation is of Italian movies.

D.) Director Analysis: *Influence of directors on movie ratings.*

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

Results:



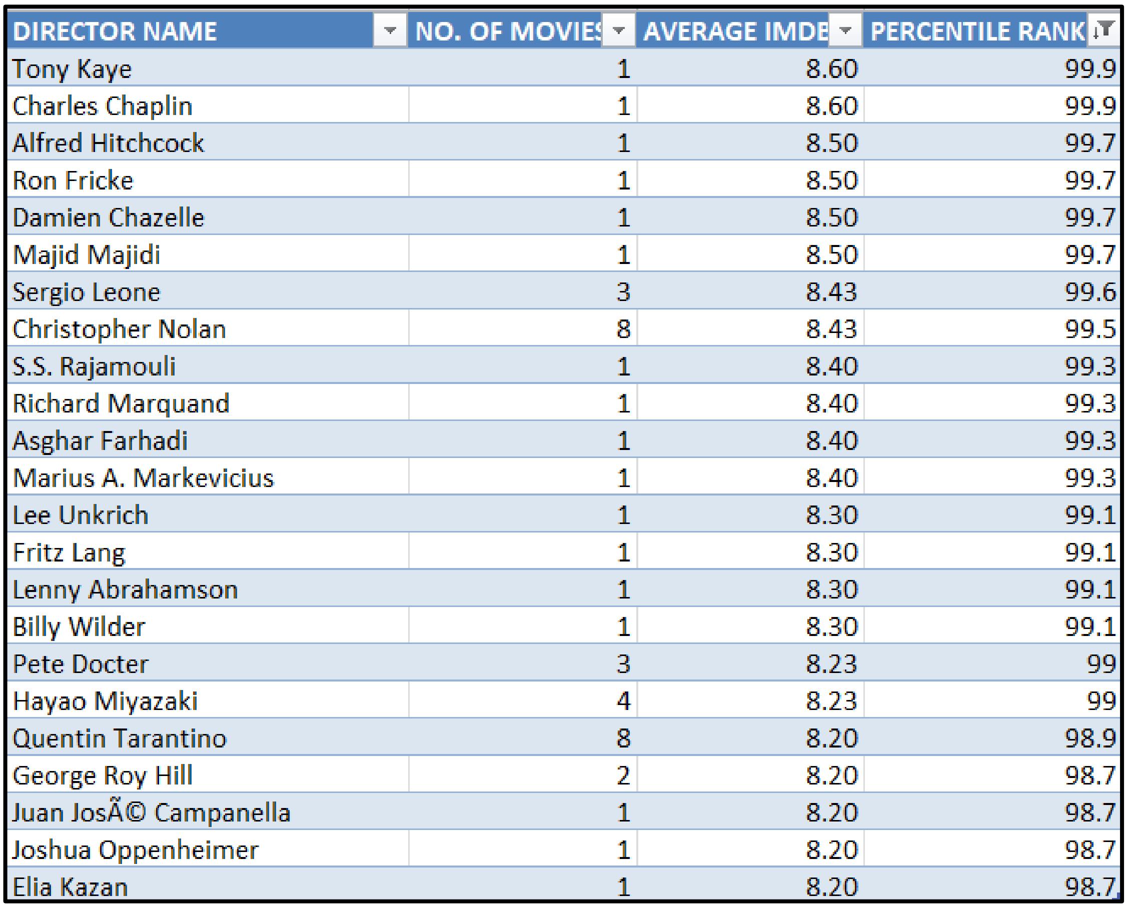
DIRECTOR NAME	NO. OF MOVIES	AVERAGE IMDB	PERCENTILE RANK
James Cameron	7	7.91	97.7
Gore Verbinski	7	6.99	72.3
Sam Mendes	7	7.46	89.2
Christopher Nolan	8	8.43	99.5
Andrew Stanton	3	7.73	95.4
Sam Raimi	10	6.96	71.9
Nathan Greno	1	7.80	95.8
Joss Whedon	3	7.87	96.9
David Yates	3	7.20	81.3
Zack Snyder	7	7.14	80.6
Bryan Singer	8	7.29	85
Marc Forster	7	7.23	84.3
Andrew Adamson	4	7.15	80.7
Rob Marshall	5	6.60	55.4
Barry Sonnenfeld	7	6.46	50
Peter Jackson	9	7.89	97
Marc Webb	3	7.13	80.3
Ridley Scott	16	7.13	80.2
Chris Weitz	5	6.08	34.9
Anthony Russo	4	7.00	72.5
Peter Berg	6	6.67	59.3
Colin Trevorrow	2	7.00	72.5
Shane Black	2	7.40	87.6
Tim Burton	14	7.05	76.7

FORMULA USED:

For calculating number of movies of each director: =COUNTIF($A$2:$A$3784,IJ2)

For average of imbd director wise: =AVERAGEIF($A$2:$A$3784,IJ2,$L$2:$L$3784)

For calculating percentile rank: =PERCENTRANK.INC($IL$2:$IL$1749,IL2)*100



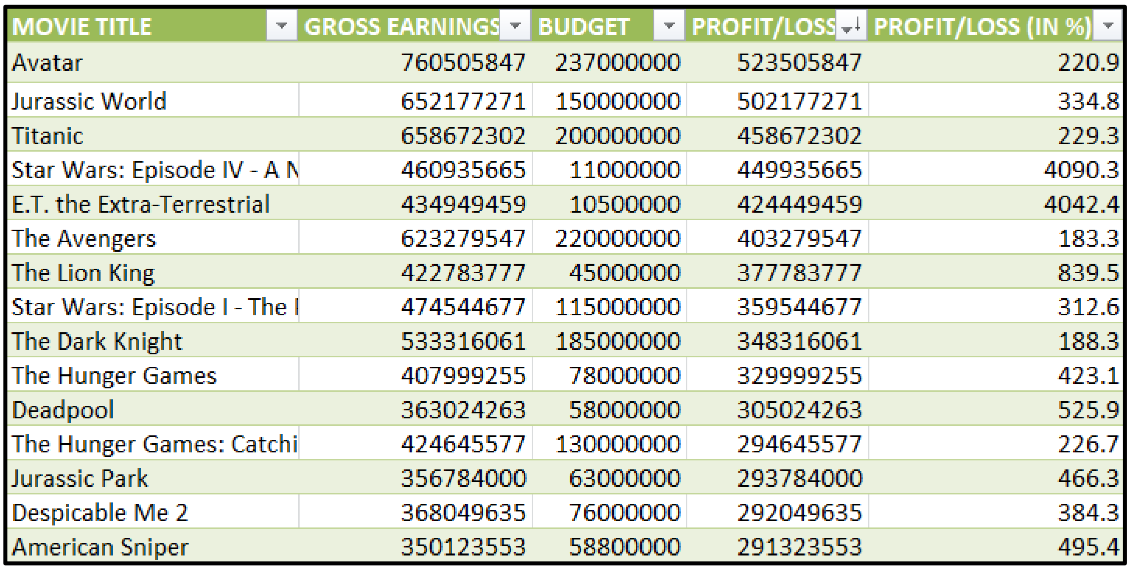
DIRECTOR NAME	NO. OF MOVIES	AVERAGE IMDE	PERCENTILE RANK
Tony Kaye	1	8.60	99.9
Charles Chaplin	1	8.60	99.9
Alfred Hitchcock	1	8.50	99.7
Ron Fricke	1	8.50	99.7
Damien Chazelle	1	8.50	99.7
Majid Majidi	1	8.50	99.7
Sergio Leone	3	8.43	99.6
Christopher Nolan	8	8.43	99.5
S.S. Rajamouli	1	8.40	99.3
Richard Marquand	1	8.40	99.3
Asghar Farhadi	1	8.40	99.3
Marius A. Markevicius	1	8.40	99.3
Lee Unkrich	1	8.30	99.1
Fritz Lang	1	8.30	99.1
Lenny Abrahamson	1	8.30	99.1
Billy Wilder	1	8.30	99.1
Pete Docter	3	8.23	99
Hayao Miyazaki	4	8.23	99
Quentin Tarantino	8	8.20	98.9
George Roy Hill	2	8.20	98.7
Juan JosÃ© Campanella	1	8.20	98.7
Joshua Oppenheimer	1	8.20	98.7
Elia Kazan	1	8.20	98.7

Conclusion: The maximum number of movies were made by “steven Spielberg” I.e., 25 movies. However, the percentile rank of average imdb is highest of “tony kaye”. But, since he has made only one movie so this won’t be an appropriate means of analysing.

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

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

Results:



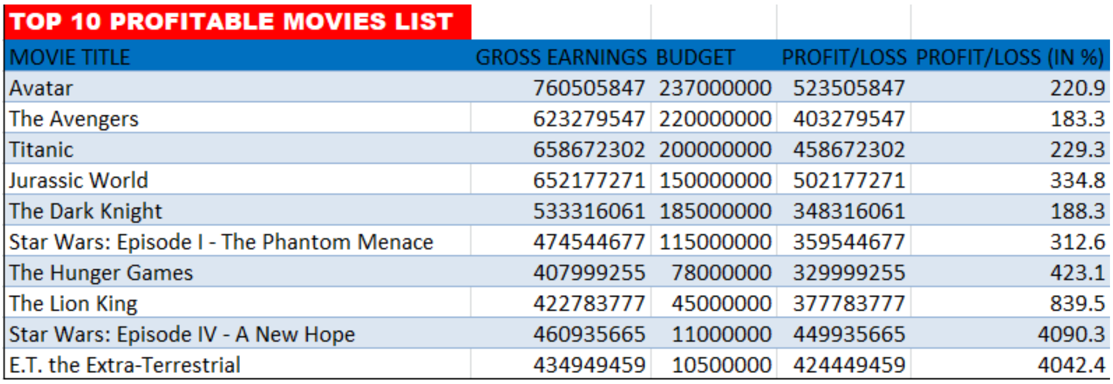
MOVIE TITLE	GROSS EARNINGS	BUDGET	PROFIT/LOSS	PROFIT/LOSS (IN %)
Avatar	760505847	237000000	523505847	220.9
Jurassic World	652177271	150000000	502177271	334.8
Titanic	658672302	200000000	458672302	229.3
Star Wars: Episode IV - A N	460935665	11000000	449935665	4090.3
E.T. the Extra-Terrestrial	434949459	10500000	424449459	4042.4
The Avengers	623279547	220000000	403279547	183.3
The Lion King	422783777	45000000	377783777	839.5
Star Wars: Episode I - The I	474544677	115000000	359544677	312.6
The Dark Knight	533316061	185000000	348316061	188.3
The Hunger Games	407999255	78000000	329999255	423.1
Deadpool	363024263	58000000	305024263	525.9
The Hunger Games: Catchi	424645577	130000000	294645577	226.7
Jurassic Park	356784000	63000000	293784000	466.3
Despicable Me 2	368049635	76000000	292049635	384.3
American Sniper	350123553	58800000	291323553	495.4

FORMULA USED:

For profit: =IQ2-IR2

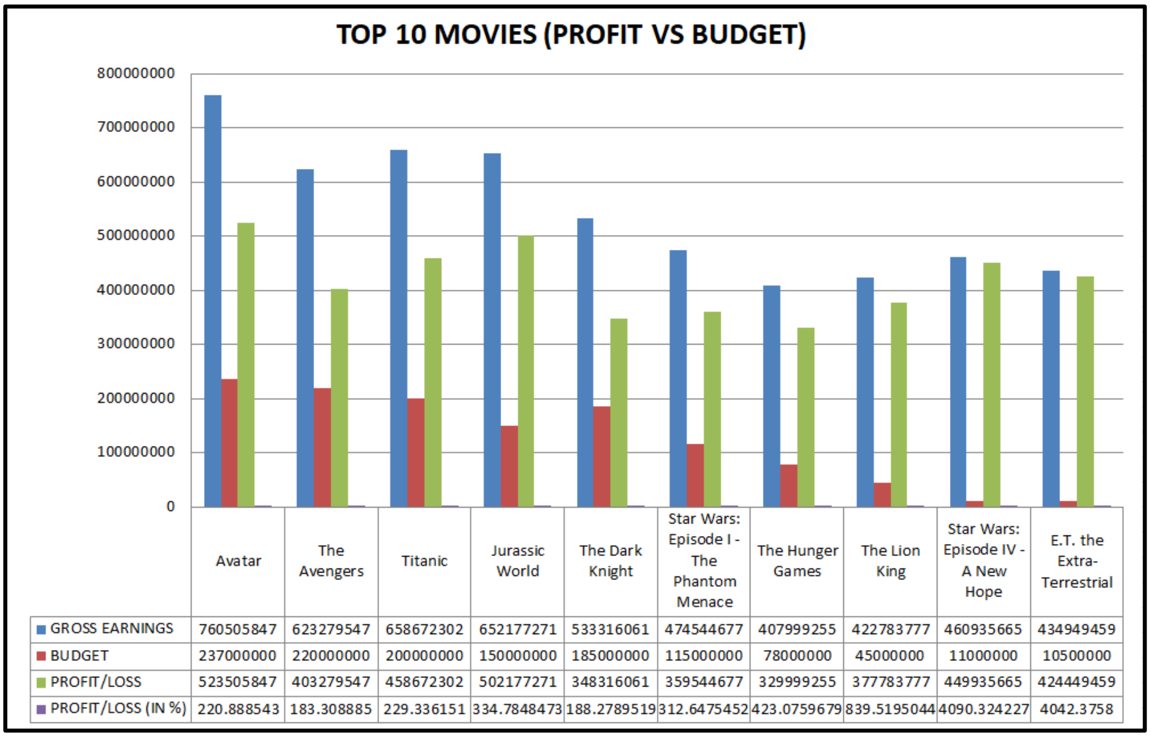
For profit percentage: =(IS2/IR2)*100

For correlation coefficient: =CORREL([GROSS EARNINGS],[BUDGET])



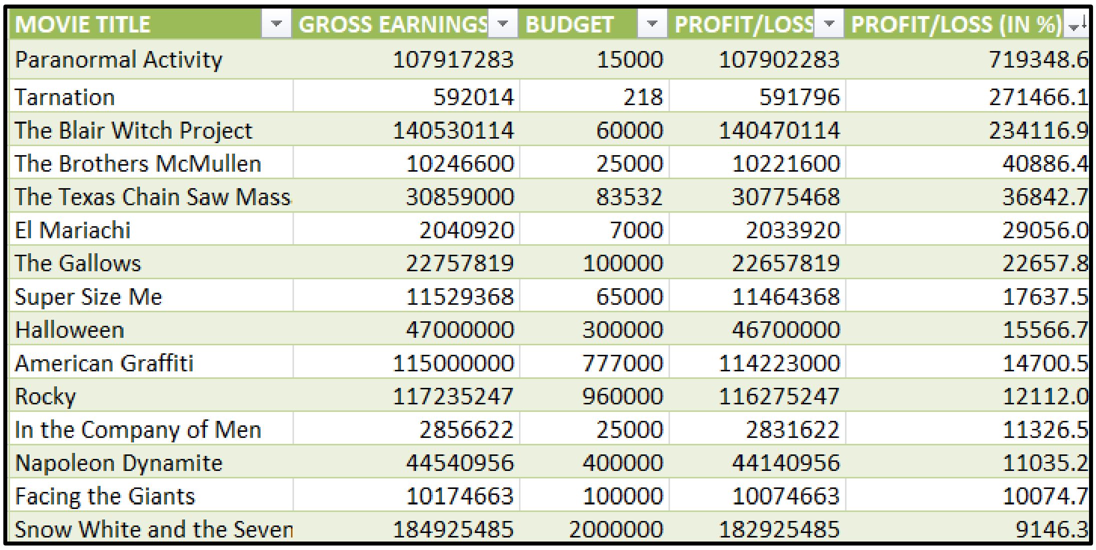
TOP 10 PROFITABLE MOVIES LIST

MOVIE TITLE	GROSS EARNINGS	BUDGET	PROFIT/LOSS	PROFIT/LOSS (IN %)
Avatar	760505847	237000000	523505847	220.9
The Avengers	623279547	220000000	403279547	183.3
Titanic	658672302	200000000	458672302	229.3
Jurassic World	652177271	150000000	502177271	334.8
The Dark Knight	533316061	185000000	348316061	188.3
Star Wars: Episode I - The Phantom Menace	474544677	115000000	359544677	312.6
The Hunger Games	407999255	78000000	329999255	423.1
The Lion King	422783777	45000000	377783777	839.5
Star Wars: Episode IV - A New Hope	460935665	11000000	449935665	4090.3
E.T. the Extra-Terrestrial	434949459	10500000	424449459	4042.4



	Avatar	The Avengers	Titanic	Jurassic World	The Dark Knight	Star Wars: Episode I - The Phantom Menace	The Hunger Games	The Lion King	Star Wars: Episode IV - A New Hope	E.T. the Extra-Terrestrial
🟦 GROSS EARNINGS	760505847	623279547	658672302	652177271	533316061	474544677	407999255	422783777	460935665	434949459
🟥 BUDGET	237000000	220000000	200000000	150000000	185000000	115000000	78000000	45000000	11000000	10500000
🟩 PROFIT/LOSS	523505847	403279547	458672302	502177271	348316061	359544677	329999255	377783777	449935665	424449459
🟪 PROFIT/LOSS (IN %)	220.888543	183.308885	229.336151	334.7848473	188.2789519	312.6475452	423.0759679	839.5195044	4090.324227	4042.3758

Sorting according to percentile ranking:



MOVIE TITLE	GROSS EARNINGS	BUDGET	PROFIT/LOSS	PROFIT/LOSS (IN %)
Paranormal Activity	107917283	15000	107902283	719348.6
Tarnation	592014	218	591796	271466.1
The Blair Witch Project	140530114	60000	140470114	234116.9
The Brothers McMullen	10246600	25000	10221600	40886.4
The Texas Chain Saw Mass	30859000	83532	30775468	36842.7
El Mariachi	2040920	7000	2033920	29056.0
The Gallows	22757819	100000	22657819	22657.8
Super Size Me	11529368	65000	11464368	17637.5
Halloween	47000000	300000	46700000	15566.7
American Graffiti	115000000	777000	114223000	14700.5
Rocky	117235247	960000	116275247	12112.0
In the Company of Men	2856622	25000	2831622	11326.5
Napoleon Dynamite	44540956	400000	44140956	11035.2
Facing the Giants	10174663	100000	10074663	10074.7
Snow White and the Seven	184925485	2000000	182925485	9146.3

Conclusion: The maximum profit is of “Avatar” movie which is Rs.52,35,05,847. However, the profit percent is highest of the “paranormal activity” i.e., 719348.6%. The data seems to be suspicious as the movie is made in only 15,000. Also the correlation coefficient between earnings and budget is 0.096477

INSIGHTS

- While doing this project learnt a lot about data cleaning and approach you need to take and also its importance by lots of trial and error.
- Also learnt a lot about tables and its pros and usage.
- And lastly learnt a lot about dealing with databases and things need to be done for certain desired outcomes.

For viewing the detailed excel sheet for this project:
https://docs.google.com/spreadsheets/d/1UN19zmQB3WLiGms-8e3Y92dpWs62eqj2/edit?usp=sharing&ouid=105345698502572804950&rtpof=true&sd=true

THANK YOU

Name : Abhay Saxena

Email ID : abhaysaxena700@gmail.com

Linkedin : https://www.linkedin.com/in/abhay-saxena-060496196/