TRAINITY TASK-5

IMDB MOVIE ANALYSIS

PRESENTED BY- ADITYA PALANDE

Email id - adityap.works@gmail.com

CONTENTS

01 DATA CLEANING

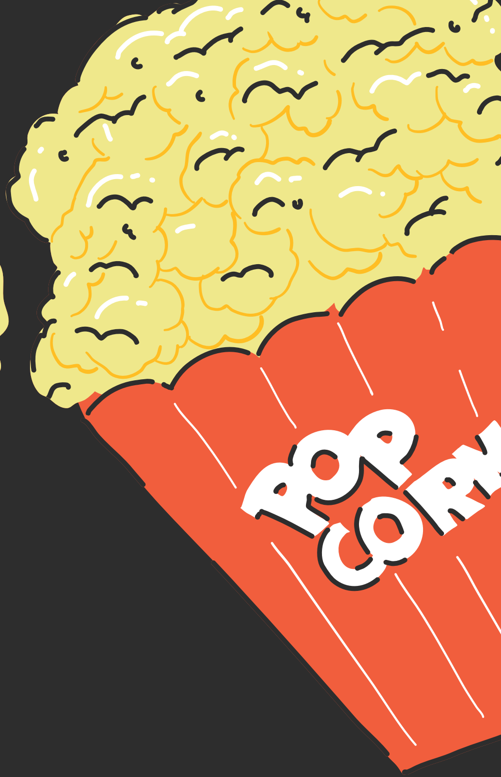
02 MOVIE GENRE ANALYSIS

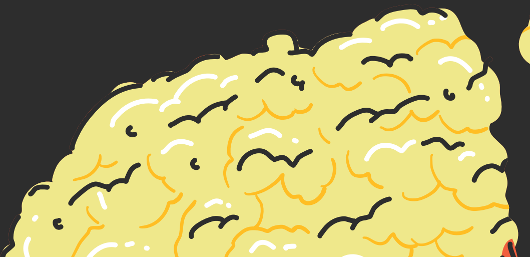
03 MOVIE DURATION ANALYSIS

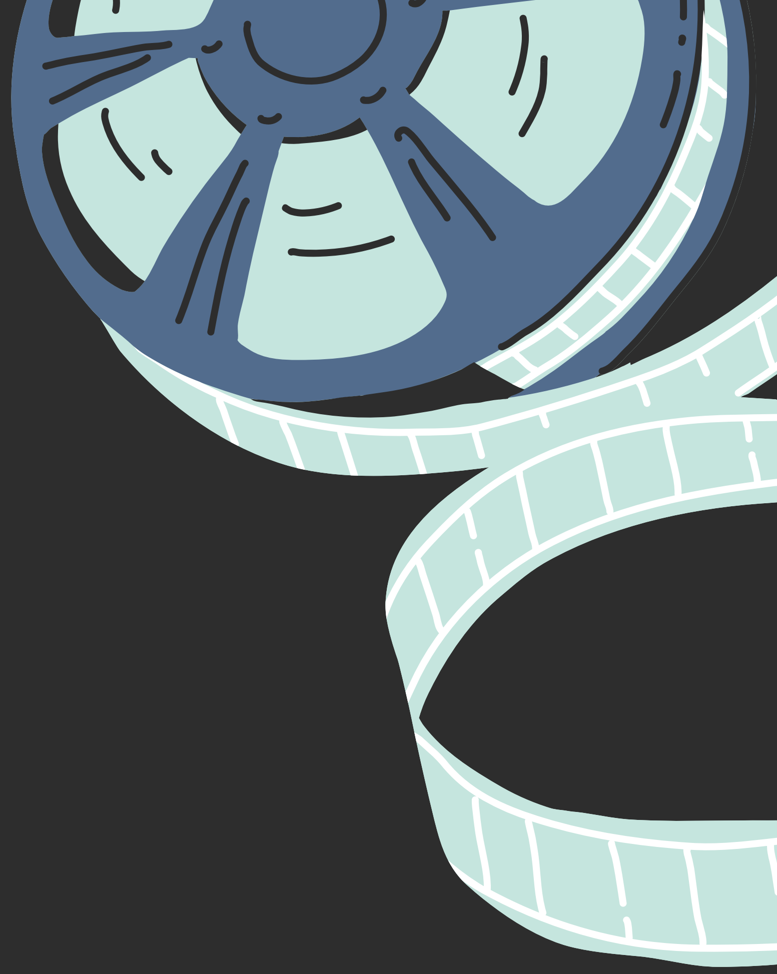
04 LANGUAGE ANALYSIS

05 DIRECTOR ANALYSIS

06 BUDGET ANALYSIS







DATA CLEANING

Data cleaning was done in order to remove blanks from the dataset. The blanks in the dataset were removed following the steps mentioned below:

1. Convert the dataset into a table
2. Filter each column to show only those rows that contain blanks
3. Delete all the rows

MOVIE GENRE ANALYSIS

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.

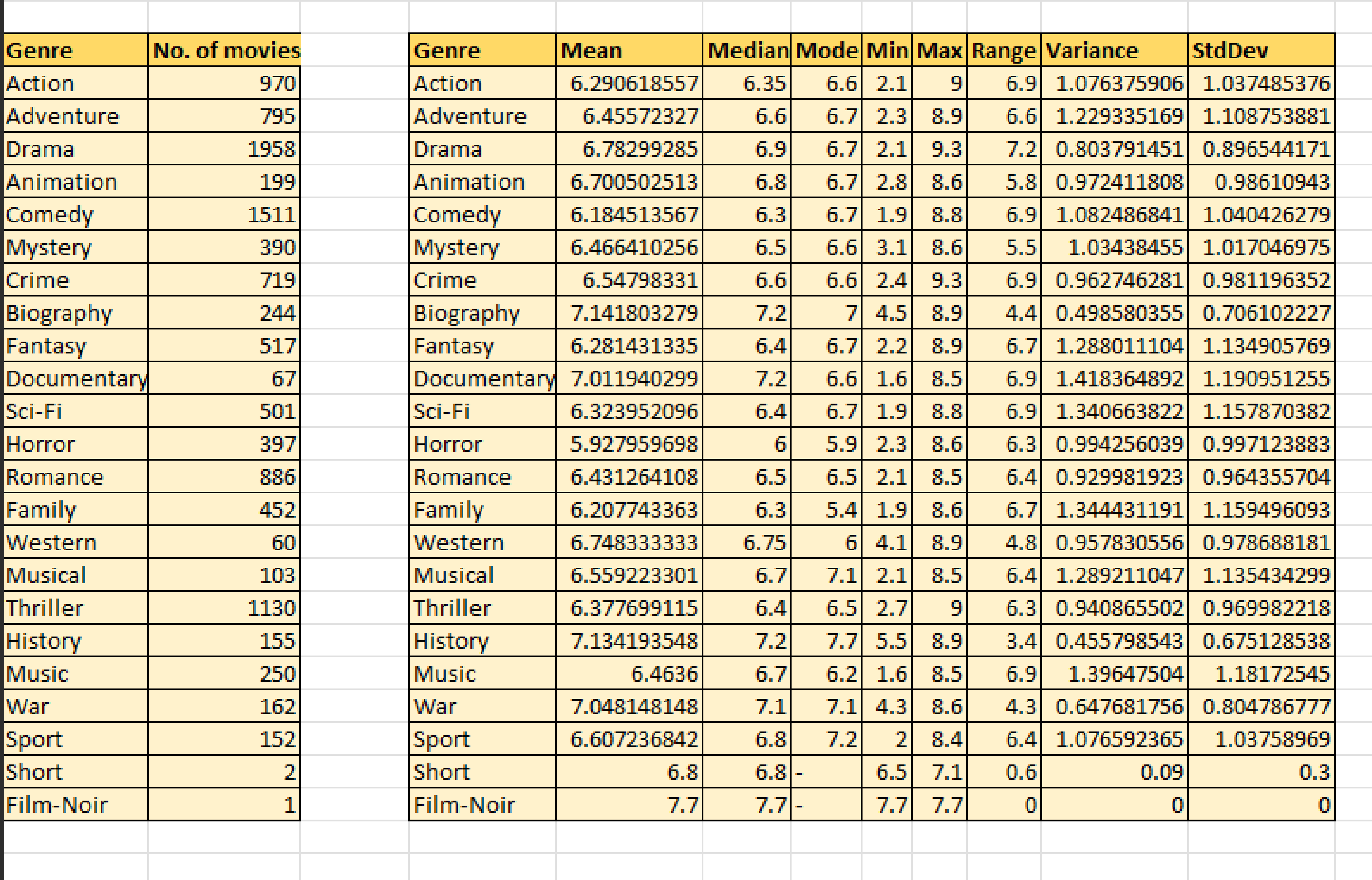
Approach:

1. Manipulation of Genre Column:

used “Text to Column“ (delimiter- “|“)

2. Creation of columns : Genre, No. of Movies, Measn, Median, Mode, Min, Max, Range, Variance and StdDev

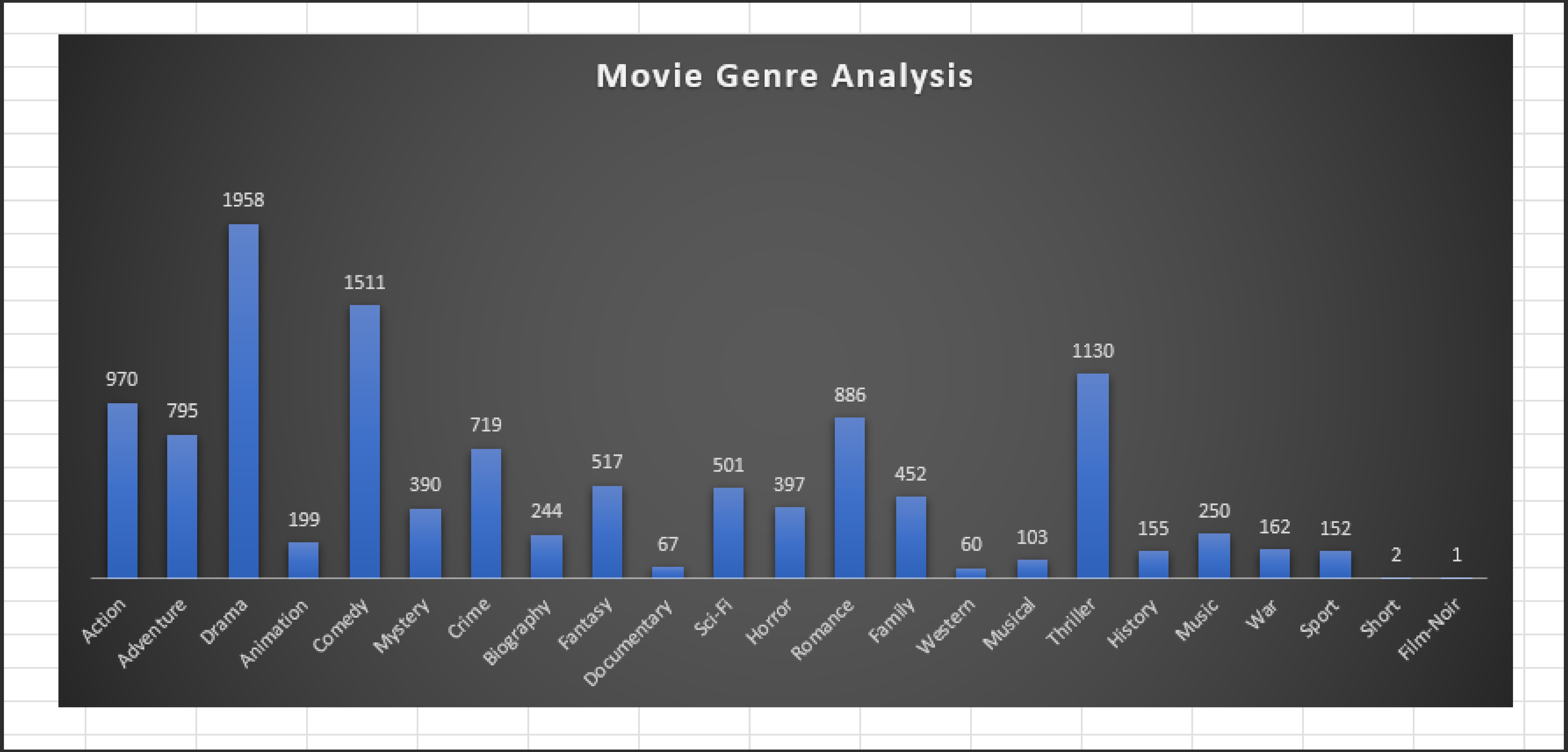
3. Functions used : UNIQUE(), COUNTIF(), MEDIAN(), AVERAGEIF(), MODE.MULT(), MIN(), MAX(), VAR.P(), SQRT(), IF(), ISNUMBER(), SEARCH(), IFERROR()



Genre	No. of movies
Action	970
Adventure	795
Drama	1958
Animation	199
Comedy	1511
Mystery	390
Crime	719
Biography	244
Fantasy	517
Documentary	67
Sci-Fi	501
Horror	397
Romance	886
Family	452
Western	60
Musical	103
Thriller	1130
History	155
Music	250
War	162
Sport	152
Short	2
Film-Noir	1

Genre	Mean	Median	Mode	Min	Max	Range	Variance	StdDev
Action	6.290618557	6.35	6.6	2.1	9	6.9	1.076375906	1.037485376
Adventure	6.45572327	6.6	6.7	2.3	8.9	6.6	1.229335169	1.108753881
Drama	6.78299285	6.9	6.7	2.1	9.3	7.2	0.803791451	0.896544171
Animation	6.700502513	6.8	6.7	2.8	8.6	5.8	0.972411808	0.98610943
Comedy	6.184513567	6.3	6.7	1.9	8.8	6.9	1.082486841	1.040426279
Mystery	6.466410256	6.5	6.6	3.1	8.6	5.5	1.03438455	1.017046975
Crime	6.54798331	6.6	6.6	2.4	9.3	6.9	0.962746281	0.981196352
Biography	7.141803279	7.2	7	4.5	8.9	4.4	0.498580355	0.706102227
Fantasy	6.281431335	6.4	6.7	2.2	8.9	6.7	1.288011104	1.134905769
Documentary	7.011940299	7.2	6.6	1.6	8.5	6.9	1.418364892	1.190951255
Sci-Fi	6.323952096	6.4	6.7	1.9	8.8	6.9	1.340663822	1.157870382
Horror	5.927959698	6	5.9	2.3	8.6	6.3	0.994256039	0.997123883
Romance	6.431264108	6.5	6.5	2.1	8.5	6.4	0.929981923	0.964355704
Family	6.207743363	6.3	5.4	1.9	8.6	6.7	1.344431191	1.159496093
Western	6.748333333	6.75	6	4.1	8.9	4.8	0.957830556	0.978688181
Musical	6.559223301	6.7	7.1	2.1	8.5	6.4	1.289211047	1.135434299
Thriller	6.377699115	6.4	6.5	2.7	9	6.3	0.940865502	0.969982218
History	7.134193548	7.2	7.7	5.5	8.9	3.4	0.455798543	0.675128538
Music	6.4636	6.7	6.2	1.6	8.5	6.9	1.39647504	1.18172545
War	7.048148148	7.1	7.1	4.3	8.6	4.3	0.647681756	0.804786777
Sport	6.607236842	6.8	7.2	2	8.4	6.4	1.076592365	1.03758969
Short	6.8	6.8	-	6.5	7.1	0.6	0.09	0.3
Film-Noir	7.7	7.7	-	7.7	7.7	0	0	0

Table generated after performing the task



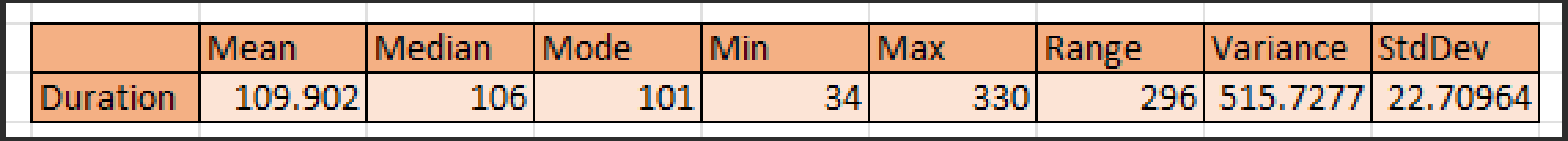
From the above chart it can be seen that movies made in "Drama" genre have the highest mean IMDB ratings followed by Comedy and Thriller. Film-Noir, Short have the least

MOVIE DURATION ANALYSIS

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

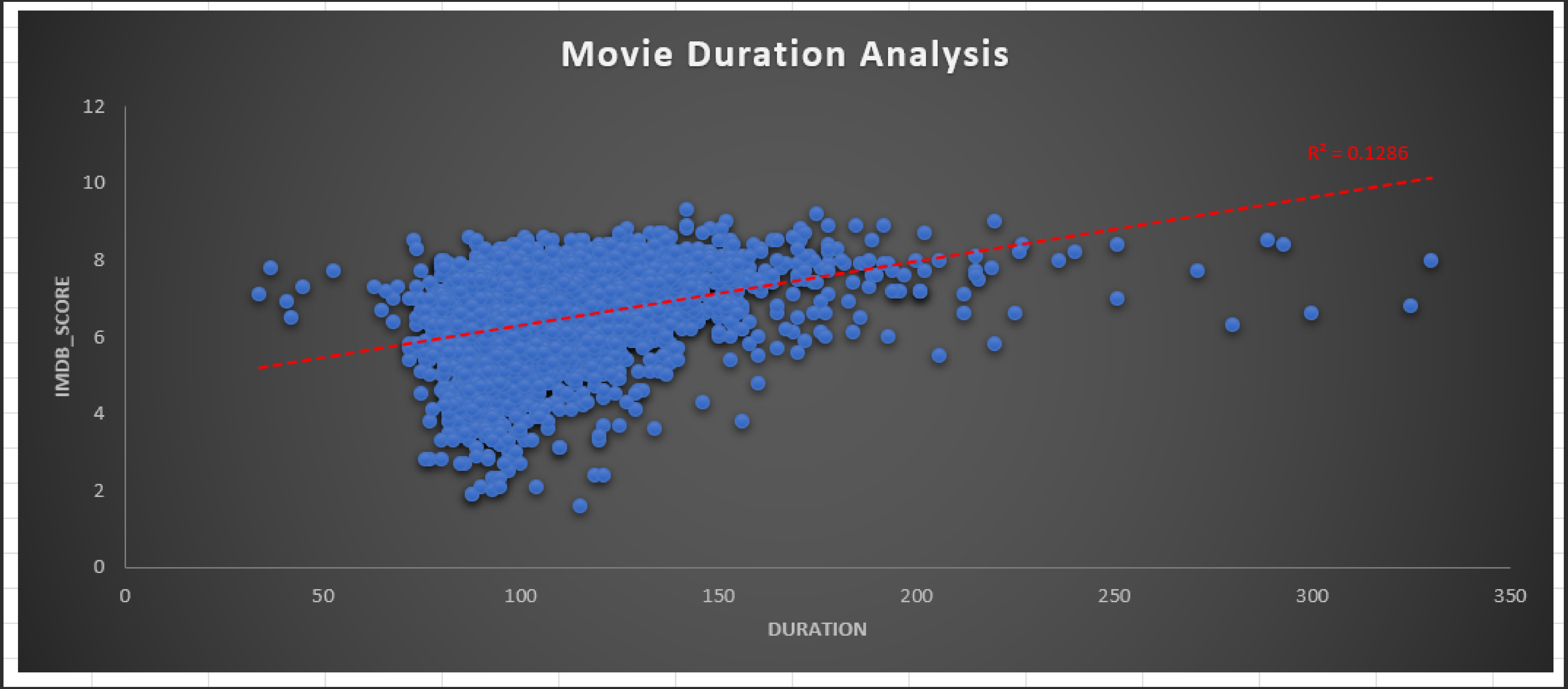
Approach:

1. Duration analysis :



	Mean	Median	Mode	Min	Max	Range	Variance	StdDev
Duration	109.902	106	101	34	330	296	515.7277	22.70964

3. Functions used : MEDIAN(), AVERAGEIF(), MODE.MULT(), MIN(), MAX(), VAR.P(), SQRT(), IF(), ISNUMBER(), SEARCH(), IFERROR()



The above scatter plot chart gives us the relationship between Duration of movies and the IMDB scores. A trendline has been added whose R_squared values is around 0.13. Most movies made were around 70 to 150 mins long.

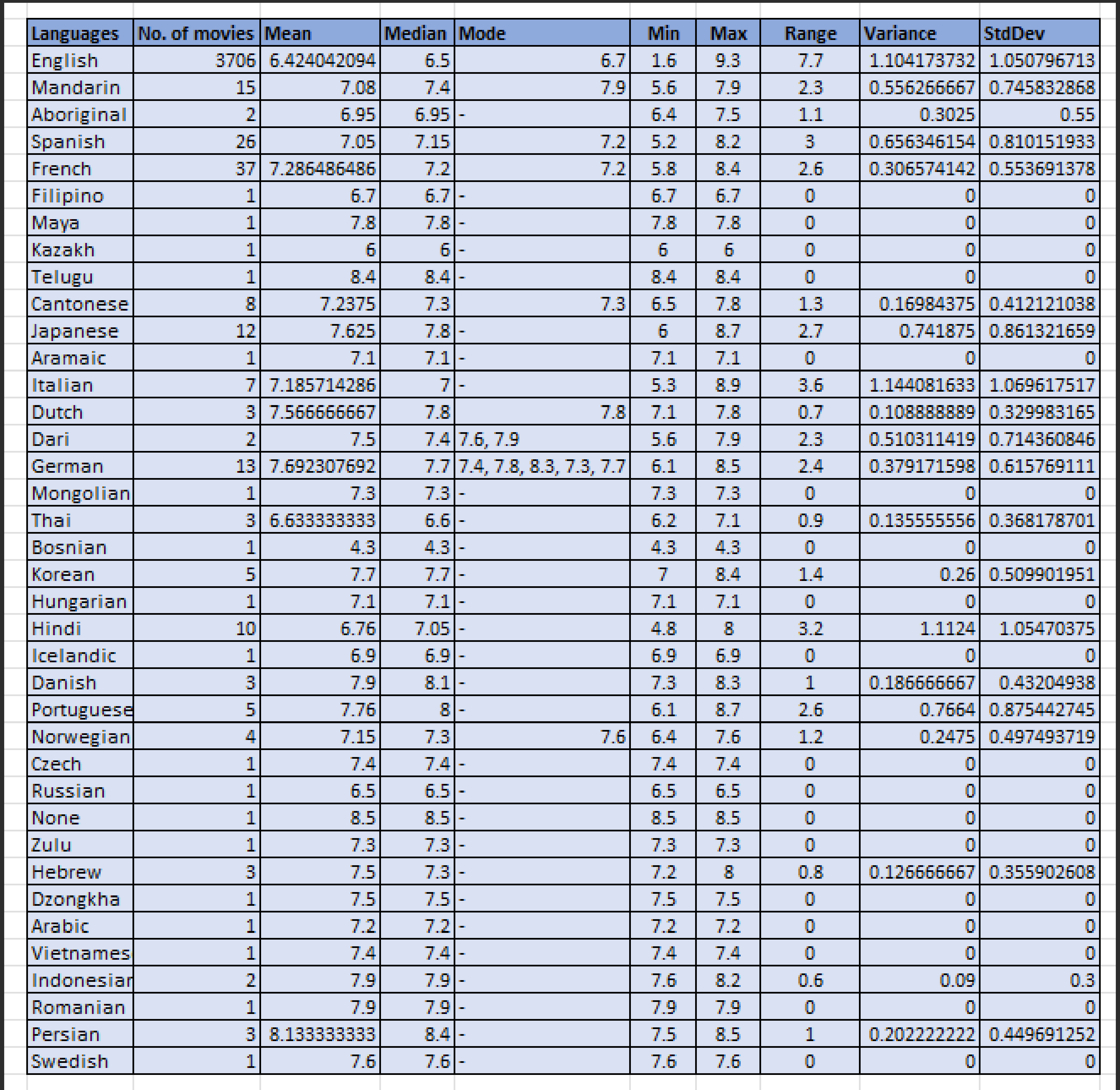
LANGUAGE ANALYSIS

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

Approach:

1. Pivot table for analysis

2. Functions used : UNIQUE(), COUNTIF(), MEDIAN(), AVERAGEIF(), MODE.MULT(), MIN(), MAX(), VAR.P(), SQRT(), IF(), ISNUMBER(), SEARCH(), IFERROR()



Languages	No. of movies	Mean	Median	Mode	Min	Max	Range	Variance	StdDev
English	3706	6.424042094	6.5	6.7	1.6	9.3	7.7	1.104173732	1.050796713
Mandarin	15	7.08	7.4	7.9	5.6	7.9	2.3	0.556266667	0.745832868
Aboriginal	2	6.95	6.95	-	6.4	7.5	1.1	0.3025	0.55
Spanish	26	7.05	7.15	7.2	5.2	8.2	3	0.656346154	0.810151933
French	37	7.286486486	7.2	7.2	5.8	8.4	2.6	0.306574142	0.553691378
Filipino	1	6.7	6.7	-	6.7	6.7	0	0	0
Maya	1	7.8	7.8	-	7.8	7.8	0	0	0
Kazakh	1	6	6	-	6	6	0	0	0
Telugu	1	8.4	8.4	-	8.4	8.4	0	0	0
Cantonese	8	7.2375	7.3	7.3	6.5	7.8	1.3	0.16984375	0.412121038
Japanese	12	7.625	7.8	-	6	8.7	2.7	0.741875	0.861321659
Aramaic	1	7.1	7.1	-	7.1	7.1	0	0	0
Italian	7	7.185714286	7	-	5.3	8.9	3.6	1.144081633	1.069617517
Dutch	3	7.566666667	7.8	7.8	7.1	7.8	0.7	0.108888889	0.329983165
Dari	2	7.5	7.4	7.6, 7.9	5.6	7.9	2.3	0.510311419	0.714360846
German	13	7.692307692	7.7	7.4, 7.8, 8.3, 7.3, 7.7	6.1	8.5	2.4	0.379171598	0.615769111
Mongolian	1	7.3	7.3	-	7.3	7.3	0	0	0
Thai	3	6.633333333	6.6	-	6.2	7.1	0.9	0.135555556	0.368178701
Bosnian	1	4.3	4.3	-	4.3	4.3	0	0	0
Korean	5	7.7	7.7	-	7	8.4	1.4	0.26	0.509901951
Hungarian	1	7.1	7.1	-	7.1	7.1	0	0	0
Hindi	10	6.76	7.05	-	4.8	8	3.2	1.1124	1.05470375
Icelandic	1	6.9	6.9	-	6.9	6.9	0	0	0
Danish	3	7.9	8.1	-	7.3	8.3	1	0.186666667	0.43204938
Portuguese	5	7.76	8	-	6.1	8.7	2.6	0.7664	0.875442745
Norwegian	4	7.15	7.3	7.6	6.4	7.6	1.2	0.2475	0.497493719
Czech	1	7.4	7.4	-	7.4	7.4	0	0	0
Russian	1	6.5	6.5	-	6.5	6.5	0	0	0
None	1	8.5	8.5	-	8.5	8.5	0	0	0
Zulu	1	7.3	7.3	-	7.3	7.3	0	0	0
Hebrew	3	7.5	7.3	-	7.2	8	0.8	0.126666667	0.355902608
Dzongkha	1	7.5	7.5	-	7.5	7.5	0	0	0
Arabic	1	7.2	7.2	-	7.2	7.2	0	0	0
Vietnames	1	7.4	7.4	-	7.4	7.4	0	0	0
Indonesiar	2	7.9	7.9	-	7.6	8.2	0.6	0.09	0.3
Romanian	1	7.9	7.9	-	7.9	7.9	0	0	0
Persian	3	8.133333333	8.4	-	7.5	8.5	1	0.202222222	0.449691252
Swedish	1	7.6	7.6	-	7.6	7.6	0	0	0

The table alonside shows all the languages the movies were made in and their descriptive analysis. 3706 movies were made in English language. The movie that recieved the highest ratings was in English language. The second most popular language was seen to be French.

DIRECTOR ANALYSIS

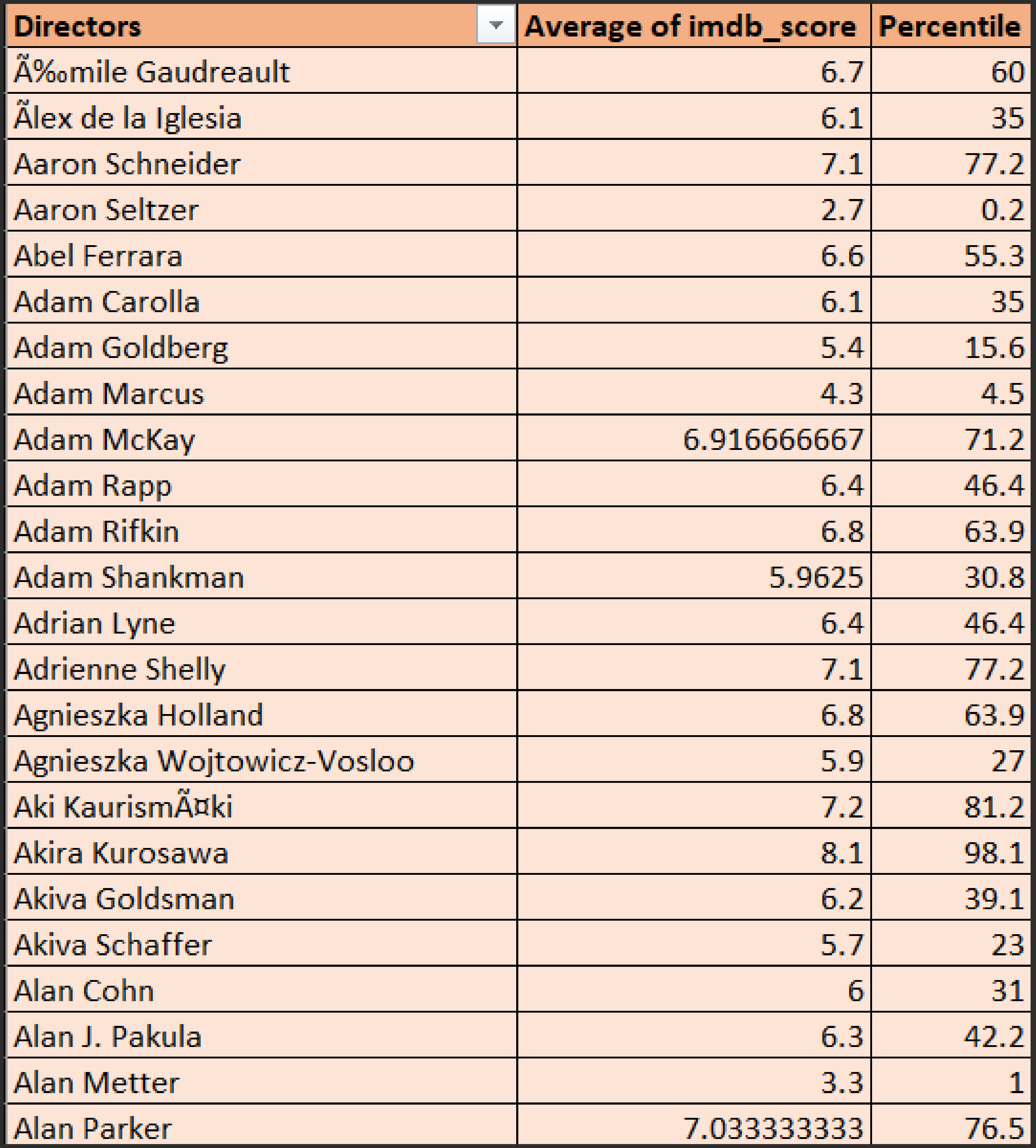
Task: Identify the top directors based on their average IMDB score and analyze their contribution to the success of movies using percentile calculations.

Approach:

1. Pivot table for analysis

2. Columns : Directors, Average of imdb_scores, percentile

3. Functions used : PERCENTRANK.INC()



Directors	Average of imdb_score	Percentile
Ã‰mile Gaudreault	6.7	60
Ãlex de la Iglesia	6.1	35
Aaron Schneider	7.1	77.2
Aaron Seltzer	2.7	0.2
Abel Ferrara	6.6	55.3
Adam Carolla	6.1	35
Adam Goldberg	5.4	15.6
Adam Marcus	4.3	4.5
Adam McKay	6.916666667	71.2
Adam Rapp	6.4	46.4
Adam Rifkin	6.8	63.9
Adam Shankman	5.9625	30.8
Adrian Lyne	6.4	46.4
Adrienne Shelly	7.1	77.2
Agnieszka Holland	6.8	63.9
Agnieszka Wojtowicz-Vosloo	5.9	27
Aki KaurismÃ¤ki	7.2	81.2
Akira Kurosawa	8.1	98.1
Akiva Goldsman	6.2	39.1
Akiva Schaffer	5.7	23
Alan Cohn	6	31
Alan J. Pakula	6.3	42.2
Alan Metter	3.3	1
Alan Parker	7.033333333	76.5

The table contains three columns viz, Directors, Average of imdb_scores, percentile. The most rated director of all comes out to be Charles Chaplin, Tony Kaye both having the average rating of 8.6

BUDGET ANALYSIS

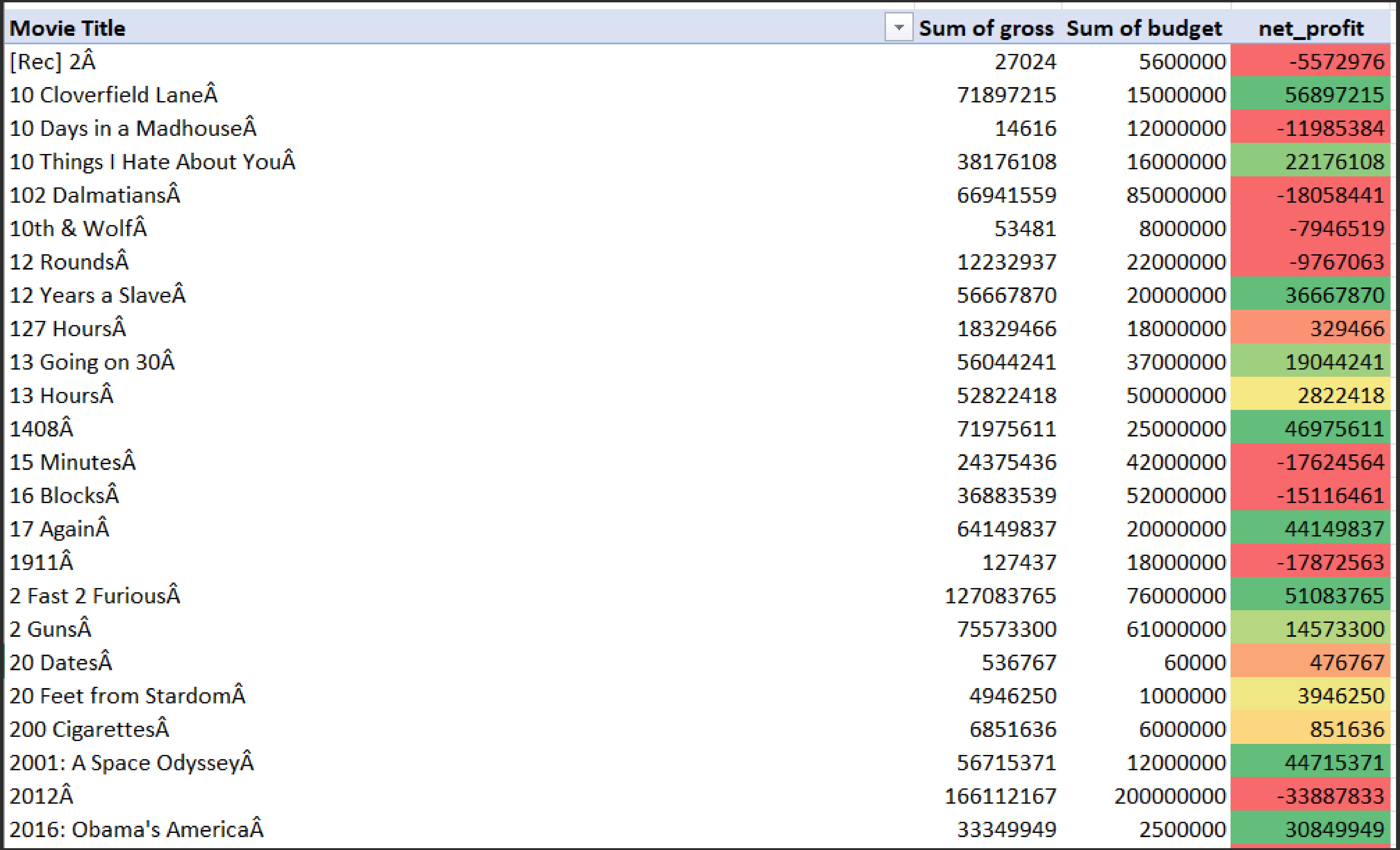
Task: Analyze the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.

Approach:

1. Pivot table for analysis

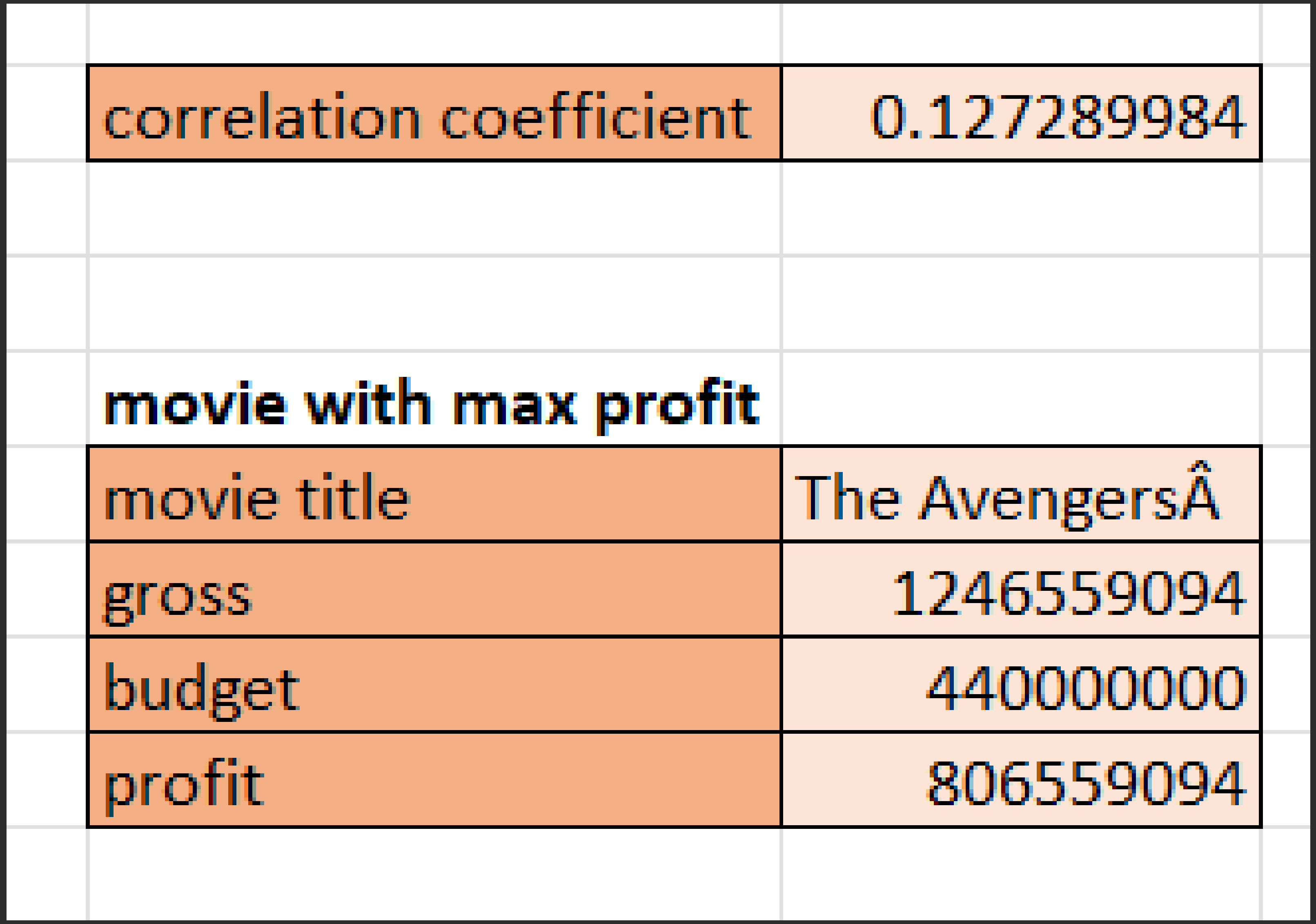
2. Conditional Formatting : Profit margins have been formatted using color scaling for easy and quick insightful understanding.

3. Functions used : CORREL(), MAX(), INDEX(), MATCH()



Movie Title	Sum of gross	Sum of budget	net_profit
[Rec] 2Â	27024	5600000	-5572976
10 Cloverfield LaneÂ	71897215	15000000	56897215
10 Days in a MadhouseÂ	14616	12000000	-11985384
10 Things I Hate About YouÂ	38176108	16000000	22176108
102 DalmatiansÂ	66941559	85000000	-18058441
10th & WolfÂ	53481	8000000	-7946519
12 RoundsÂ	12232937	22000000	-9767063
12 Years a SlaveÂ	56667870	20000000	36667870
127 HoursÂ	18329466	18000000	329466
13 Going on 30Â	56044241	37000000	19044241
13 HoursÂ	52822418	50000000	2822418
1408Â	71975611	25000000	46975611
15 MinutesÂ	24375436	42000000	-17624564
16 BlocksÂ	36883539	52000000	-15116461
17 AgainÂ	64149837	20000000	44149837
1911Â	127437	18000000	-17872563
2 Fast 2 FuriousÂ	127083765	76000000	51083765
2 GunsÂ	75573300	61000000	14573300
20 DatesÂ	536767	60000	476767
20 Feet from StardomÂ	4946250	1000000	3946250
200 CigarettesÂ	6851636	6000000	851636
2001: A Space OdysseyÂ	56715371	12000000	44715371
2012Â	166112167	200000000	-33887833
2016: Obama's AmericaÂ	33349949	2500000	30849949

Conditional Formatting has been used to color scale the profit margin sa that losses and profits can be seen instantly.



correlation coefficient	0.127289984

movie with max profit

movie title	The AvengersÂ
gross	1246559094
budget	440000000
profit	806559094

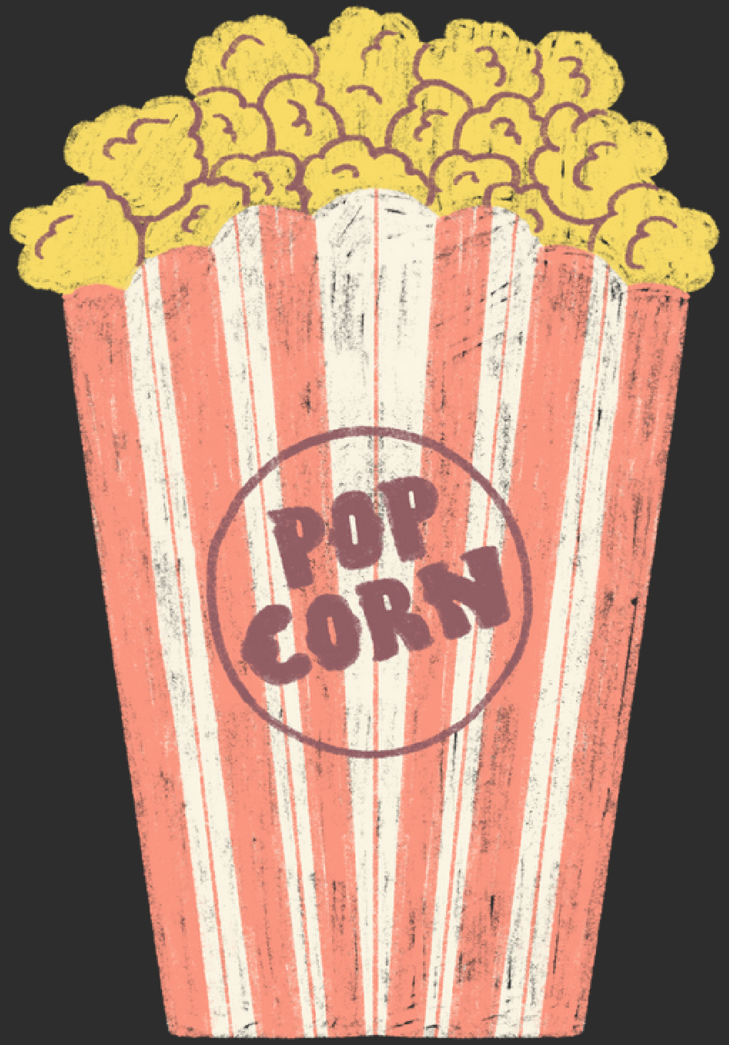
The correlation coefficient was found using the CORREL() function and has a value of 0.128 approx. The movie that made the maximum profit was “The AvengersA” .

Link to my working excel sheet and video presentation:

Excel Sheet

Video

CONCLUSION



The tasks were performed using Microsoft Excel. The tasks not only helped in understanding excel tools but also allowed me to get hands-on experience by solving real-life examples. Through these tasks insights could be drawn and strategies couls be made.