IMDB Movie Analysis

Link for Google Sheets: https://docs.google.com/spreadsheets/d/134qUS6AE1FmzjW2wKUemQOFuQD-CYimsf3JP27GIoPo/edit?usp=sharing

Project Description:

This project aims to analyze a dataset containing information about various movies from the IMDB database. The goal is to gain insights into different aspects of the movies, such as genre, duration, language, directors, and budgets, and their impact on the IMDB scores and financial success. By employing statistics and Excel formulas, we will extract meaningful conclusions to help understand the factors that contribute to a movie's popularity and success.

A. Movie Genre Analysis:
Task 1: Determine the most common genres of movies in the dataset.
Task 2: Calculate descriptive statistics (mean, median, mode, range, variance, standard deviation) of the IMDB scores for each genre.

B. Movie Duration Analysis:
Task 1: Analyze the distribution of movie durations.
Task 2: Visualize the relationship between movie duration and IMDB score.

C. Language Analysis:
Task 1: Determine the most common languages used in movies.
Task 2: Analyze the impact of language on IMDB scores using descriptive statistics.

D. Director Analysis:
Task 1: Identify the top directors based on their average IMDB score.
Task 2: Analyze the contribution of top directors to the success of movies using percentile calculations.

E. Budget Analysis:
Task 1: Analyze the correlation between movie budgets and gross earnings.
Task 2: Identify the movies with the highest profit margin.

By completing the above tasks and analyzing the data using statistics and Excel formulas, we will gain valuable insights into the impact of movie genres, duration, language, directors, and budgets on IMDB scores and financial success. These findings will assist in making informed decisions to improve movie-making strategies and achieve greater popularity and profitability for future films.

Approach:

Data Engineering:

Genre column has multiple genres in the same column and can’t be used directly to find distinct genre count and use each genre effectively. So we first split it to multiple columns by “split text to column” method and then list all distinct genres in a column for future use. We can use “=UNIQUE()” formula to find distinct elements in any column, and will do so in any future analysis.

A. Movie Genre Analysis:

We will use Excel's COUNTIF function to count the occurrences of each genre.

To Calculate descriptive statistics (mean, median, mode, range, variance, standard deviation) of the IMDB scores for each genre. We will first need to manipulate the 'genres' column to separate multiple genres for a single movie. Then, we will use Excel's AVERAGE, MEDIAN, MODE, MAX, MIN, VAR, and STDEV functions to calculate the required statistics for each genre.

B. Movie Duration Analysis:

We will calculate descriptive statistics (mean, median, and standard deviation) for movie durations using Excel's functions.

We will create a scatter plot to visualize the relationship between movie duration and IMDB score. Additionally, we will add a trendline to assess the direction and strength of the relationship.

C. Language Analysis:

We will use Excel's COUNTIF function to count the number of movies for each language.

We will calculate the mean, median, and standard deviation of the IMDB scores for each language using Excel's functions.

D. Director Analysis:

We will calculate the average IMDB score for each director and use Excel's PERCENTILE function to identify the directors with the highest scores.

We will compare the scores of the top directors to the overall distribution of scores to assess their impact.

E. Budget Analysis:

We will calculate the correlation coefficient between movie budgets and gross earnings using Excel's CORREL function.

We will calculate the profit margin (gross earnings - budget) for each movie and use Excel's MAX function to identify the movies with the highest profit margin.

Tech-Stack Used:

For this project, I utilized Google Sheets as the primary software tool. Google Sheets is a spreadsheet application included as part of the free, web-based Google Docs Editors suite offered by Google.

Insights:

A. Movie Genre Analysis:

It is important to analyze the distribution of movies across different genres and understand relationships between genre and IMDB score, to predict what kind of movies audiences prefer. For this descriptive statistics can be used. For doing this following formulas were used.

To find distinct genres from column of genres, we first separated genres into multiple columns and then, made unique list of genres using following formula:

```
=UNIQUE(TOCOL(J:Q))
```

To count the number of movies per genre:

```
=COUNTIF(J:Q,AR3)
```

To find average IMDB Score per genre:

```
=ArrayFormula(AVERAGE(IF((J:J=AR3)+(K:K=AR3)+(L:L=AR3)+(M:M=AR3)+(N:N=AR3)+(O:O=AR3)+(P:P=AR3)+(Q:Q=AR3), AG:AG)))
```

To find median IMDB Score per genre:

```
=ArrayFormula(MEDIAN(IF((J:J=$AR3)+(K:K=$AR3)+(L:L=$AR3)+(M:M=$AR3)+(N:N=$AR3)+(O:O=$AR3)+(P:P=$AR3)+(Q:Q=$AR3), AG:AG)))
```

To find mode IMDB Score per genre:

```
=ArrayFormula(MODE(IF((J:J=$AR3)+(K:K=$AR3)+(L:L=$AR3)+(M:M=$AR3)+(N:N=$AR3)+(O:O=$AR3)+(P:P=$AR3)+(Q:Q=$AR3), AG:AG)))
```

To find minimum IMDB Score per genre:

```
=ArrayFormula(MIN(IF((J:J=$AR3)+(K:K=$AR3)+(L:L=$AR3)+(M:M=$AR3)+(N:N=$AR3)+(O:O=$AR3)+(P:P=$AR3)+(Q:Q=$AR3), AG:AG)))
```

To find maximum IMDB Score per genre:

```
=MAX(MAXIFS(AG:AG,J:J,AR3),MAXIFS(AG:AG,K:K,AR3),MAXIFS(AG:AG,L:L,AR3),MAXIFS(AG:AG,M:M,AR3),MAXIFS(AG:AG,N:N,AR3),MAXIFS(AG:AG,O:O,AR3),MAXIFS(AG:AG,P:P,AR3),MAXIFS(AG:AG,Q:Q,AR3))
```

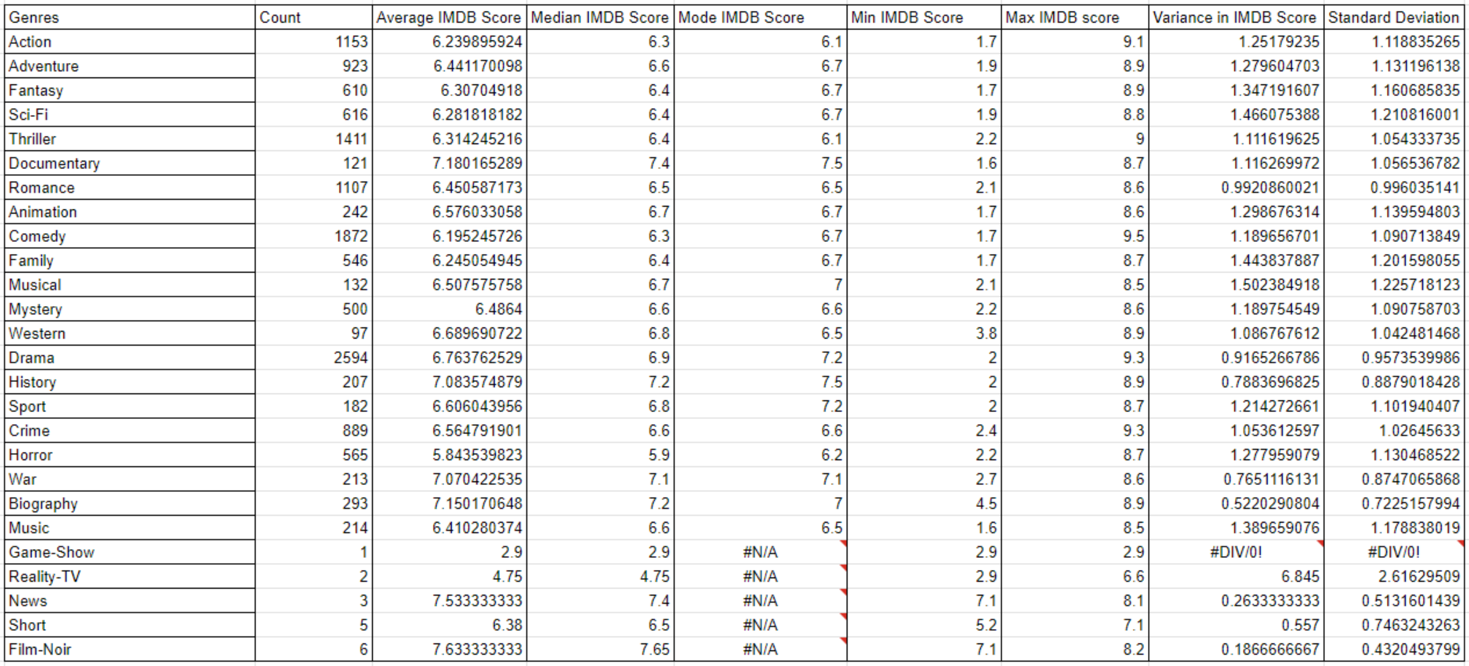
To find variance in IMDB Score per genre:

```
=ArrayFormula(VAR(IF((J:J=$AR3)+(K:K=$AR3)+(L:L=$AR3)+(M:M=$AR3)+(N:N=$AR3)+(O:O=$AR3)+(P:P=$AR3)+(Q:Q=$AR3), AG:AG)))
```

To find standard deviation in IMDB Score per genre:

```
=ArrayFormula(STDEV(IF((J:J=$AR3)+(K:K=$AR3)+(L:L=$AR3)+(M:M=$AR3)+(N:N=$AR3)+(O:O=$AR3)+(P:P=$AR3)+(Q:Q=$AR3), AG:AG)))
```

We can automatically generate complete table for each genre by autofilling formulas in google sheets
Output:



Genres	Count	Average IMDB Score	Median IMDB Score	Mode IMDB Score	Min IMDB Score	Max IMDB score	Variance in IMDB Score	Standard Deviation
Action	1153	6.239895924	6.3	6.1	1.7	9.1	1.25179235	1.118835265
Adventure	923	6.441170098	6.6	6.7	1.9	8.9	1.279604703	1.131196138
Fantasy	610	6.30704918	6.4	6.7	1.7	8.9	1.347191607	1.160685835
Sci-Fi	616	6.281818182	6.4	6.7	1.9	8.8	1.466075388	1.210816001
Thriller	1411	6.314245216	6.4	6.1	2.2	9	1.111619625	1.054333735
Documentary	121	7.180165289	7.4	7.5	1.6	8.7	1.116269972	1.056536782
Romance	1107	6.450587173	6.5	6.5	2.1	8.6	0.9920860021	0.996035141
Animation	242	6.576033058	6.7	6.7	1.7	8.6	1.298676314	1.139594803
Comedy	1872	6.195245726	6.3	6.7	1.7	9.5	1.189656701	1.090713849
Family	546	6.245054945	6.4	6.7	1.7	8.7	1.443837887	1.201598055
Musical	132	6.507575758	6.7	7	2.1	8.5	1.502384918	1.225718123
Mystery	500	6.4864	6.6	6.6	2.2	8.6	1.189754549	1.090758703
Western	97	6.689690722	6.8	6.5	3.8	8.9	1.086767612	1.042481468
Drama	2594	6.763762529	6.9	7.2	2	9.3	0.9165266786	0.9573539986
History	207	7.083574879	7.2	7.5	2	8.9	0.7883696825	0.8879018428
Sport	182	6.606043956	6.8	7.2	2	8.7	1.214272661	1.101940407
Crime	889	6.564791901	6.6	6.6	2.4	9.3	1.053612597	1.02645633
Horror	565	5.843539823	5.9	6.2	2.2	8.7	1.277959079	1.130468522
War	213	7.070422535	7.1	7.1	2.7	8.6	0.7651116131	0.8747065868
Biography	293	7.150170648	7.2	7	4.5	8.9	0.5220290804	0.7225157994
Music	214	6.410280374	6.6	6.5	1.6	8.5	1.389659076	1.178838019
Game-Show	1	2.9	2.9	#N/A	2.9	2.9	#DIV/0!	#DIV/0!
Reality-TV	2	4.75	4.75	#N/A	2.9	6.6	6.845	2.61629509
News	3	7.533333333	7.4	#N/A	7.1	8.1	0.2633333333	0.5131601439
Short	5	6.38	6.5	#N/A	5.2	7.1	0.557	0.7463243263
Film-Noir	6	7.633333333	7.65	#N/A	7.1	8.2	0.1866666667	0.4320493799

Note that Mode is N/A where each element appears only once and variance and standard deviance can’t be calculated for genre with single element.

B. Movie Duration Analysis:

To determine the ideal movie duration, that audience prefer is essential for a successful movie. So to find the relation between movie duration and IMDB Score we can use descriptive statistics to find average, median and standard deviation of movie duration. We can also create a scatterplot to better understand relationship and plot a trendline.

Formulae:

To find average duration of movies:

```
=AVERAGE(D:D)
```

To find median duration of movies:

```
=MEDIAN(D:D)
```

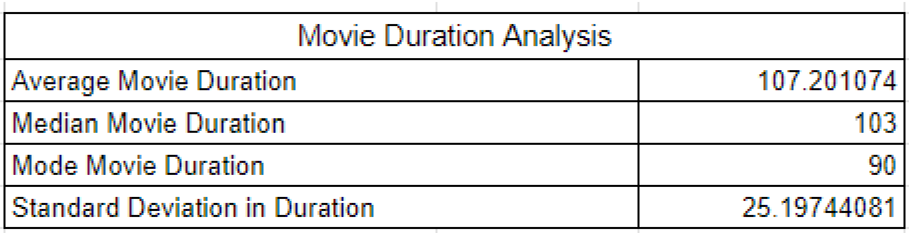
To find mode duration of movies:

```
=MODE(D:D)
```

To find standard deviation in duration of movies:

```
=STDEV(D:D)
```

Output:



Movie Duration Analysis	
Average Movie Duration	107.201074
Median Movie Duration	103
Mode Movie Duration	90
Standard Deviation in Duration	25.19744081

To Create a scatter plot of IMDB Score vs Movie duration we insert a chart and select scatterplot. We select data ranges and visualize the scatter plot. We add a trend line to better understand the relationship.

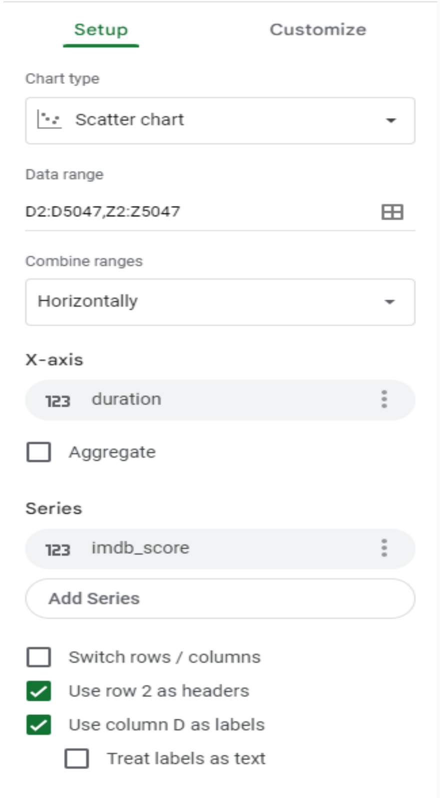


Chart type: Scatter chart

Data range: D2:D5047,Z2:Z5047

Combine ranges: Horizontally

X-axis: 123 duration

□ Aggregate

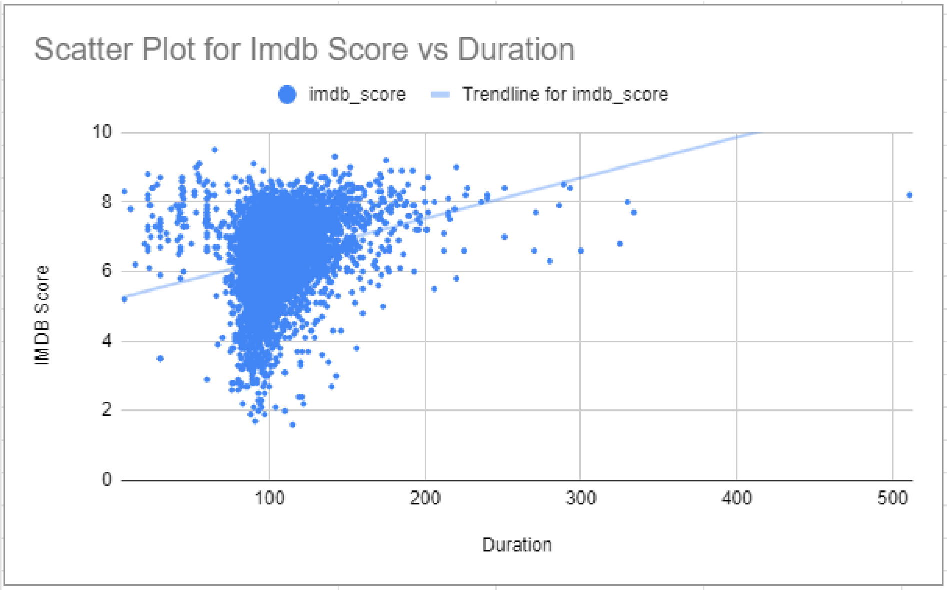
Series: 123 imdb_score

□ Switch rows / columns

☑ Use row 2 as headers

☑ Use column D as labels

□ Treat labels as text



C. Language Analysis:

Determining which language the audience prefer to watch and where the majority of movies are successful is important for making a profit. That's why Descriptive statistics of Language and IMDB score are calculated.

Unique languages are found similarly unique genres are found, by using below formula:

```
=UNIQUE(AA3:AA)
```

To find count of movies in each language following formula is used:

```
=COUNTIF(AA:AA,AU68)
```

To find average IMDB Score per language:

```
=AVERAGEIF(AA:AA,AU68,AG:AG)
```

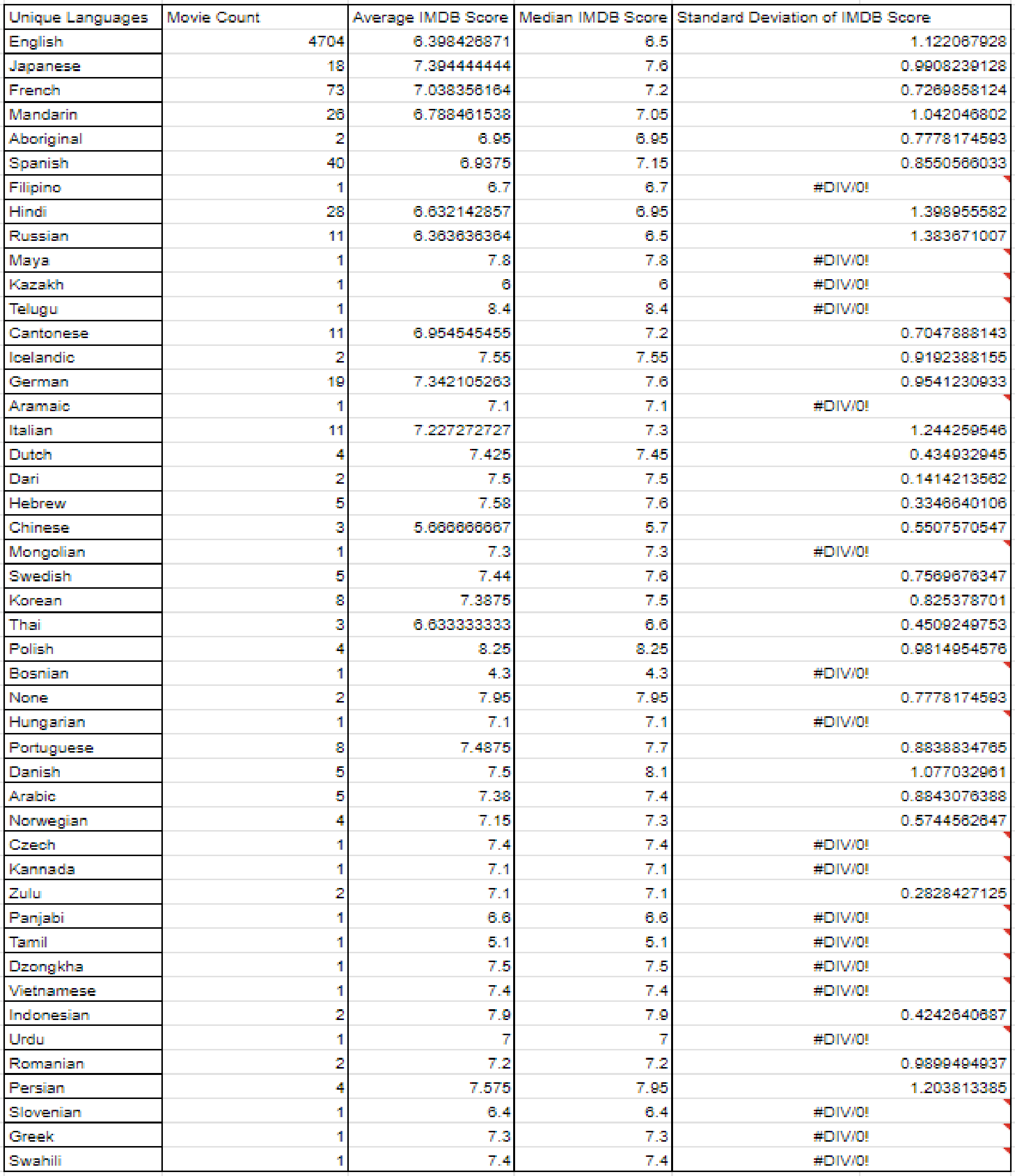
To find median IMDB Score per language:

```
=ArrayFormula(MEDIAN(if(AA:AA=$AU68,AG:AG)))
```

To find standard deviation of IMDB Score per language:

```
=ArrayFormula(STDEV(if(AA:AA=AU68,AG:AG)))
```

Output:



Unique Languages	Movie Count	Average IMDB Score	Median IMDB Score	Standard Deviation of IMDB Score
English	4704	6.398426871	6.5	1.122067928
Japanese	18	7.394444444	7.6	0.9908239128
French	73	7.038356164	7.2	0.7269858124
Mandarin	26	6.788461538	7.05	1.042046802
Aboriginal	2	6.95	6.95	0.7778174593
Spanish	40	6.9375	7.15	0.8550566033
Filipino	1	6.7	6.7	#DIV/0!
Hindi	28	6.632142857	6.95	1.398955582
Russian	11	6.363636364	6.5	1.383671007
Maya	1	7.8	7.8	#DIV/0!
Kazakh	1	6	6	#DIV/0!
Telugu	1	8.4	8.4	#DIV/0!
Cantonese	11	6.954545455	7.2	0.7047888143
Icelandic	2	7.55	7.55	0.9192388155
German	19	7.342105263	7.6	0.9541230933
Aramaic	1	7.1	7.1	#DIV/0!
Italian	11	7.227272727	7.3	1.244259546
Dutch	4	7.425	7.45	0.434932945
Dari	2	7.5	7.5	0.1414213562
Hebrew	5	7.58	7.6	0.3346640106
Chinese	3	5.666666667	5.7	0.5507570547
Mongolian	1	7.3	7.3	#DIV/0!
Swedish	5	7.44	7.6	0.7569676347
Korean	8	7.3875	7.5	0.825378701
Thai	3	6.633333333	6.6	0.4509249753
Polish	4	8.25	8.25	0.9814954576
Bosnian	1	4.3	4.3	#DIV/0!
None	2	7.95	7.95	0.7778174593
Hungarian	1	7.1	7.1	#DIV/0!
Portuguese	8	7.4875	7.7	0.8838834765
Danish	5	7.5	8.1	1.077032961
Arabic	5	7.38	7.4	0.8843076388
Norwegian	4	7.15	7.3	0.5744562647
Czech	1	7.4	7.4	#DIV/0!
Kannada	1	7.1	7.1	#DIV/0!
Zulu	2	7.1	7.1	0.2828427125
Panjabi	1	6.6	6.6	#DIV/0!
Tamil	1	5.1	5.1	#DIV/0!
Dzongkha	1	7.5	7.5	#DIV/0!
Vietnamese	1	7.4	7.4	#DIV/0!
Indonesian	2	7.9	7.9	0.4242640687
Urdu	1	7	7	#DIV/0!
Romanian	2	7.2	7.2	0.9899494937
Persian	4	7.575	7.95	1.203813385
Slovenian	1	6.4	6.4	#DIV/0!
Greek	1	7.3	7.3	#DIV/0!
Swahili	1	7.4	7.4	#DIV/0!

Note: Standard deviation of a single movie in a language can't be calculated.

D. Director Analysis:

Director of a movie plays a major role in the popularity of a movie, so finding popular directors with the most IMDB rating is detrimental in finding which movies are going to make it big in the market. To find top directors we have to first find average IMDB rating per director, and then we find top 1% directors by using “PERCENTILE” function in google sheets

Formulae:

To find all unique directors:

=QUERY(B3:B, "SELECT B WHERE B <> " AND B IS NOT NULL", 0)

To find average IMDB score per directors:

=iferror(AVERAGEIF(B3:B,AR121,AG:AG),AG:AG)

To find value at 99%le of IMDB score:

=PERCENTILE(AS121:AS,99%)

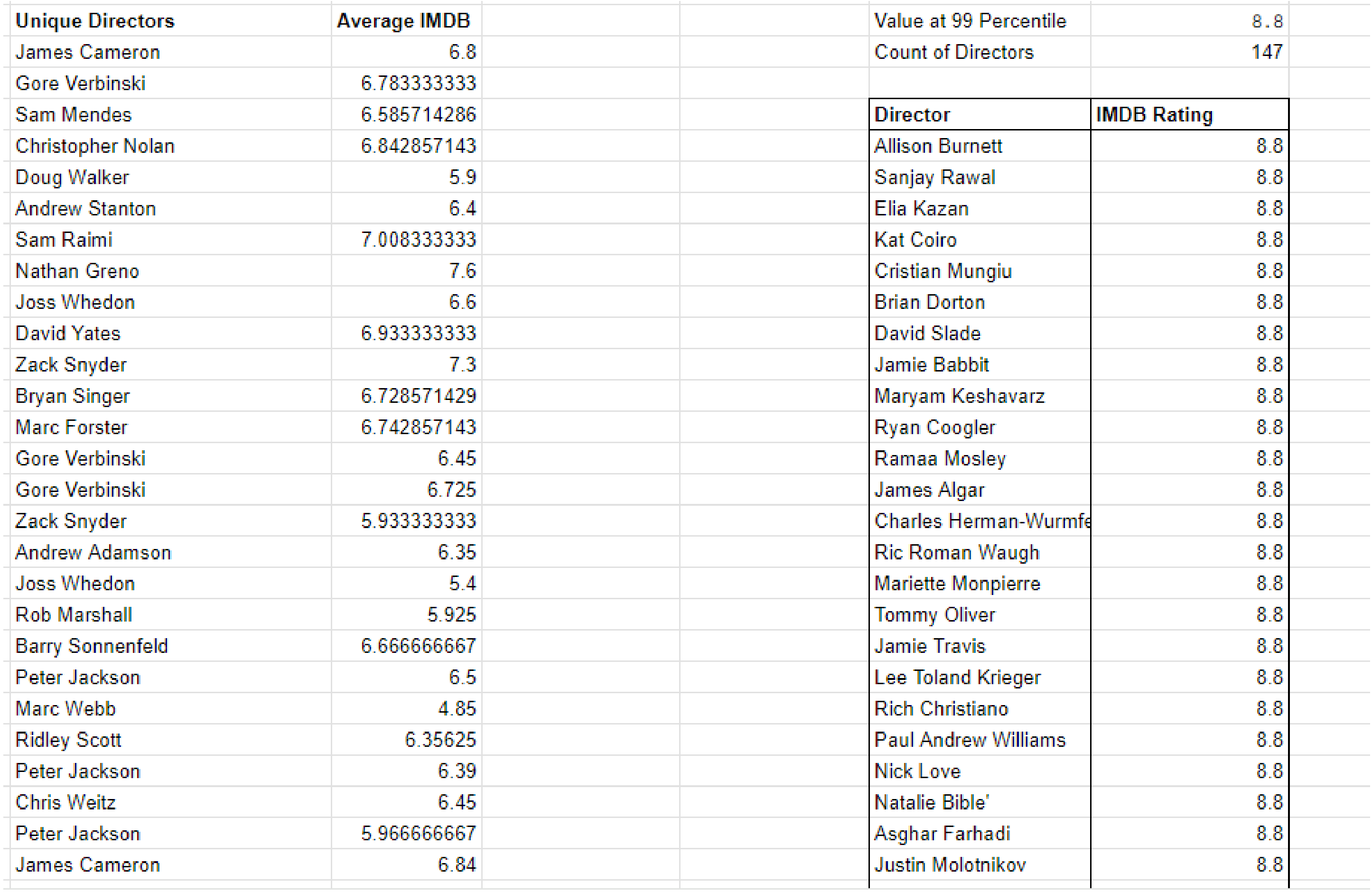
To count of top 1% directors:

=COUNTIF(AS121:AS, ">= "&PERCENTILE(AS121:AS, 99%))

To find list of top 1% directors:

=ARRAYFORMULA(FILTER(AR121:AS, AS121:AS >= AW120))

Output:



Unique Directors	Average IMDB
James Cameron	6.8
Gore Verbinski	6.783333333
Sam Mendes	6.585714286
Christopher Nolan	6.842857143
Doug Walker	5.9
Andrew Stanton	6.4
Sam Raimi	7.008333333
Nathan Greno	7.6
Joss Whedon	6.6
David Yates	6.933333333
Zack Snyder	7.3
Bryan Singer	6.728571429
Marc Forster	6.742857143
Gore Verbinski	6.45
Gore Verbinski	6.725
Zack Snyder	5.933333333
Andrew Adamson	6.35
Joss Whedon	5.4
Rob Marshall	5.925
Barry Sonnenfeld	6.666666667
Peter Jackson	6.5
Marc Webb	4.85
Ridley Scott	6.35625
Peter Jackson	6.39
Chris Weitz	6.45
Peter Jackson	5.966666667
James Cameron	6.84

Value at 99 Percentile	8.8
Count of Directors	147

Director	IMDB Rating
Allison Burnett	8.8
Sanjay Rawal	8.8
Elia Kazan	8.8
Kat Coiro	8.8
Cristian Mungiu	8.8
Brian Dorton	8.8
David Slade	8.8
Jamie Babbit	8.8
Maryam Keshavarz	8.8
Ryan Coogler	8.8
Ramaa Mosley	8.8
James Algar	8.8
Charles Herman-Wurmfe	8.8
Ric Roman Waugh	8.8
Mariette Monpierre	8.8
Tommy Oliver	8.8
Jamie Travis	8.8
Lee Toland Krieger	8.8
Rich Christiano	8.8
Paul Andrew Williams	8.8
Nick Love	8.8
Natalie Bible'	8.8
Asghar Farhadi	8.8
Justin Molotnikov	8.8

These lists are very long and only some part of it is shown in output.

E. Budget Analysis:

The most important factor in determining the success of a movie is if it made a profit or not.
Profit can be defined as the subtraction of budget from gross income it made at box office.
We can also find correlation coefficient to find how likely it is to create a profit by setting an effective budget.
Also It would help to find the movie which made most profit to learn from it, to produce more successful movies.

Formulae:

To find correlation between budget and gross profit:

=CORREL(AD3:AD,I3:I)

To find profit by subtracting budget from gross:

=ArrayFormula((I3:I-AD3:AD))

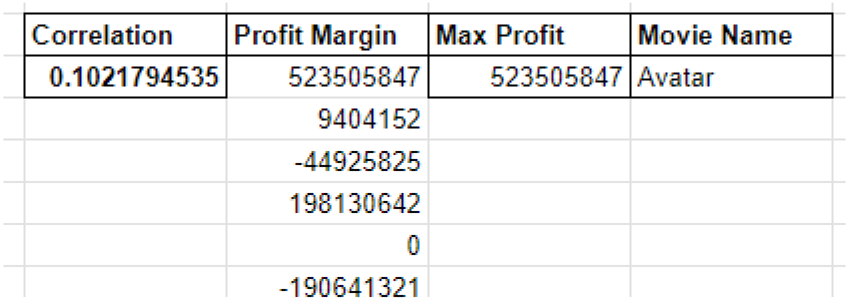
To find most profit made by a movie:

=MAX(AL:AL)

To find movie with most profit:

=if(AM3=AL:AL,S:S)

Output:



Correlation	Profit Margin	Max Profit	Movie Name
0.1021794535	523505847	523505847	Avatar
	9404152		
	-44925825		
	198130642		
	0		
	-190641321		

Note: Output is very large; only the first few lines are shown for the profit margin column.

Results:

While working on this project, I have gained a better understanding of IMDB movie analysis and Advanced Excel methodologies. By analyzing movie data, I was able to provide insights on various aspects such as genre distribution and relation with IMDB score, relation between movie duration and score and visualization, language and IMDB Score ,Impacts of popularity of director on movie, and determining profit from budget and gross income.

This project has helped me enhance my Excel skills, particularly in functions and data visualization to derive meaningful insights. It has also improved my ability to interpret data and provide actionable recommendations based on the analysis.