IMDB Movie Analysis

Submitted by:
Nadendla Dharani
Hyper Link to the Excel Sheet:
https://docs.google.com/spreadsheets/d/1T2RBqMZfvQClR6ZBuMZ9QdHdjcHgBnyh/edit?usp=sharing&ouid=111691645789497796027&rtpof=true&sd=true

Project Description

The project aims to analyze the factors influencing the success of movies on IMDB. Success, in this context, is defined by high IMDB ratings. The dataset contains information on various aspects of movies, including genres, duration, language, directors, and budgets. The goal is to provide actionable insights for movie producers, directors, and investors to make informed decisions in their future projects.

Approach

The project commences with a meticulous data cleaning phase, ensuring data integrity. Following this, an exploratory data analysis investigates the relationships between movie genres, duration, language, director, and budget with IMDB ratings. Leveraging statistical measures and the Five 'Whys' approach, the study aims to uncover key factors influencing movie success. Specific data analytics tasks, such as genre and duration analysis, language examination, director influence assessment, and budget correlation exploration, provide detailed insights. The final report combines these findings into a cohesive narrative, utilizing visualizations to communicate actionable insights for stakeholders in the film industry. Overall, the project offers a targeted approach to understanding the multifaceted dynamics that contribute to a movie's success on IMDB.

Tech-Stack Used

The project leverages Microsoft Excel as the primary tool for data analysis and visualization. Excel's versatile functions and features facilitate data cleaning, computation of summary statistics, and the creation of various visualizations, including pie charts, bar graphs, and histograms. The familiar interface of Excel allows for efficient handling of tasks such as handling missing values, clubbing categories, and outlier detection. Additionally, standard statistical functions within Excel contribute to deriving meaningful insights, making it a powerful and accessible tech-stack for this data analytics project.

Insights

1.Movie Genre:

- Certain genres may have higher average IMDB scores.
- Producers can target genres with a higher likelihood of success.

2.Movie Duration:

- There might be a sweet spot for movie duration that correlates with higher IMDB ratings.
- Directors can optimize movie length for better audience reception.

3.Language:

- Movies in certain languages may be more positively received.
- Consideration of language diversity can be crucial for global success.

4.Director Influence:

- Identification of top-performing directors can guide hiring decisions.
- Understanding the impact of directors on movie success can shape collaboration strategies.

5.Budget Analysis:

- Correlation between budgets and earnings can help in financial planning.
- Recognition of movies with high profit margins allows for strategic investment decisions.

Results

Data Cleaning: This step involves preprocessing the data to make it suitable for analysis. It includes handling missing values, removing duplicates, converting data types if necessary, and possibly feature engineering.

- From the provided dataset, I have observed that the dataset contains a lot of empty cells, unnecessary columns. So first I deleted all the columns that do not provide any help for the data extraction.
- Next, I found the blank rows and highlighted them, and since there are lot of blank rows I deleted those blank rows.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	director_nam	num_critic_fo	duration	gross	genres	movie_title	num_voted_u	num_user_for	language	country	content_ratin	budget	title_year	imdb_score
277	Christopher Barnard		22		Comedy	10,000 B.C.Â	6							7.2
1499	Tony Kaye				Crime|Drama	Black Water Tran	219		English	USA		23000000	2009	7.2
2245			30		Comedy	Fired UpÂ	114	6	English	USA				6.7
2329			30		Drama|Family	The Doombolt Ch	18		English	UK				7.2
2332	Jane Clark		7		Romance|Short	The TouchÂ	118		English	USA		13000	2007	5.2
2338	Jonathan Jakubowicz		105		Action|Biograph	Hands of StoneÂ	178	1	English	Panama	R	20000000	2016	7.2
2357			43		Comedy|Drama|	Gone, Baby, Gon	29		English		TV-14			6.6
2691	Dan Curtis		99		Fantasy|Romanc	The Love LetterÂ	1465	56	English	USA	Unrated		1998	7.4
2746	John Blanchard		65		Comedy	Towering Inferno	10		English	Canada				9.5
2848	Niels Arden Oplev		88		Action|Crime|M	Del 1 - MÃ¤n son	335		Swedish	Sweden				8.1
3778	Jim Amatulli		90		Drama|Family|M	Flying ByÂ	215	2	English	USA	PG-13		2009	4.5
3783	Eric Bross		87		Crime|Drama|Th	We Have Your Hu	216	3	English	USA	TV-PG	5000000	2011	5.5
3788	Mike Mayhall		88		Family	Running Forever	8		English	USA		5000000	2015	8.6
4057			197		Drama|War	Deadline Gallipol	299	1	English	Australia		15000000		7.4
4099	Lance Kawas		90		Comedy	The DeportedÂ	62	1	English	USA	PG-13	3000000	2009	6.2
4195	Jaco Booyens		90		Drama|Thriller	8 DaysÂ	44	3	English	USA	PG-13	2500000	2014	6.9
4211	Patricia Cardoso		89		Drama	Lies in Plain Sight	544	4	English	USA	TV-PG	2100000	2010	6.3
4289	Simon Napier-Bell		81		Documentary	To Be Frank, Sina	7		English	UK		2000000	2015	7.4
4320	John Laing		86		Drama	AbandonedÂ	333	6	English	New Zealand			2015	6.3
4338	Mitchell Altieri		87		Comedy|Horror|	A Beginner's Gui	13		English	USA			2016	8.7
4401	Jaime Zevallos		93		Comedy|Drama|	Me You and Five	7		English	USA		1500000	2015	7.6
4409	Frank Lotito		102		Comedy|Drama|	Growing Up Smit	108	1	English	USA	PG-13	2000000	2015	8.2
4424	Robert M. Young		105		Western	The Ballad of Gre	39	2	English	USA		1250000	1982	7.1

Data Analytics Tasks:

You are required to provide a detailed report for the below data record mentioning the answers of the questions that follows:

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

- **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.

- **Hint:** Use Excel's COUNTIF function to count the number of movies for each genre. You might need to manipulate the 'genres' column to separate multiple genres for a single movie. Use Excel's functions like AVERAGE, MEDIAN, MODE, MAX, MIN, VAR, and STDEV to calculate descriptive statistics. Compare the statistics to understand the impact of genre on movie ratings.

I first separated the genres using the formula:

```
=IF(ISNUMBER(SEARCH("|",'Cleaned Dataset'!E2)),LEFT('Cleaned Dataset'!E2,SEARCH("|",'Cleaned Dataset'!E2)-1),'Cleaned Dataset'!E2)
```

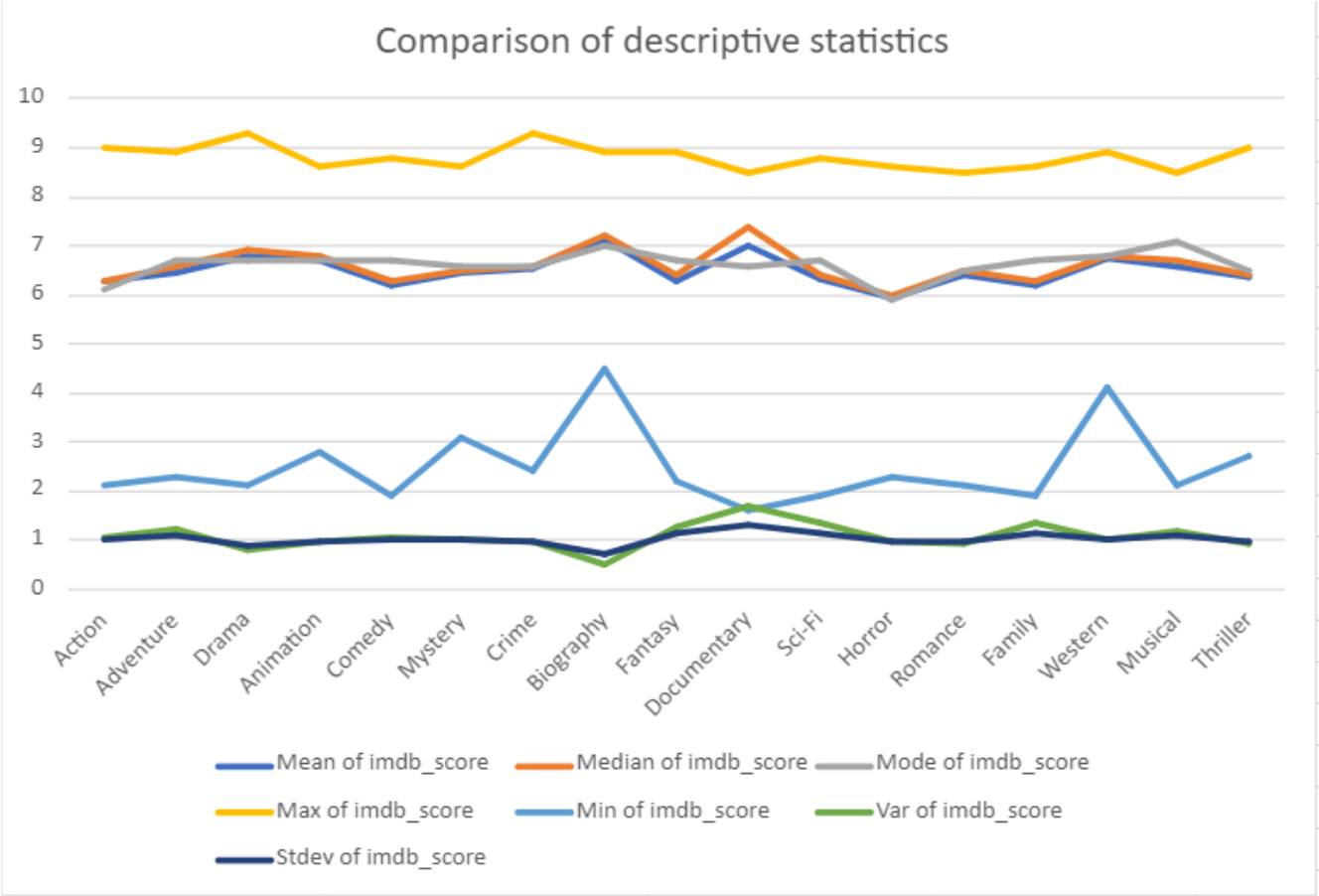
And then, counted the number of movies for each genre:

```
=COUNTIF('Cleaned Dataset'!E2:E3804,"*" & E4 & "*")
```

Later I calculated the needed descriptive statistics:

```
=AVERAGEIF('Cleaned Dataset'!E2:E3804,"*" & E4 & "*",'Cleaned Dataset'!N2:N3804)
=MEDIAN(IF(ISNUMBER(SEARCH("*" & E4 & "*",'Cleaned Dataset'!E2:E3804)),'Cleaned Dataset'!N2:N3804))
=MODE(IF(ISNUMBER(SEARCH("*" & E4 & "*",'Cleaned Dataset'!E2:E3804)),'Cleaned Dataset'!N2:N3804))
=MAXIFS('Cleaned Dataset'!N2:N3804,'Cleaned Dataset'!E2:E3804,"*" & E4 & "*")
=MINIFS('Cleaned Dataset'!N2:N3804,'Cleaned Dataset'!E2:E3804,"*" & E4 & "*")
=VAR(IF(ISNUMBER(SEARCH("*" & E4 & "*",'Cleaned Dataset'!E2:E3804)),'Cleaned Dataset'!N2:N3804))
=STDEV(IF(ISNUMBER(SEARCH("*" & E4 & "*",'Cleaned Dataset'!E2:E3804)),'Cleaned Dataset'!N2:N3804))
```

Unique Genres	Movies Count	Mean of imdb_score	Median of imdb_score	Mode of imdb_score	Max of imdb_score	Min of imdb_score	Var of imdb_score	Stdev of imdb_score
Action	957	6.287565308	6.3	6.1	9	2.1	1.063977016	1.031492616
Adventure	781	6.452112676	6.6	6.7	8.9	2.3	1.22808848	1.108191536
Drama	1918	6.782221064	6.9	6.7	9.3	2.1	0.798311809	0.893482965
Animation	198	6.702525253	6.8	6.7	8.6	2.8	0.981465672	0.990689493
Comedy	1485	6.18013468	6.3	6.7	8.8	1.9	1.074477071	1.036569858
Mystery	382	6.472774869	6.5	6.6	8.6	3.1	1.012327713	1.006144976
Crime	710	6.541408451	6.6	6.6	9.3	2.4	0.963021991	0.981336839
Biography	241	7.142323651	7.2	7	8.9	4.5	0.506284578	0.711536772
Fantasy	507	6.281854043	6.4	6.7	8.9	2.2	1.269314342	1.126638514
Documentary	55	6.998181818	7.4	6.6	8.5	1.6	1.702774411	1.304903985
Sci-Fi	491	6.316089613	6.4	6.7	8.8	1.9	1.34392427	1.159277477
Horror	388	5.922680412	6	5.9	8.6	2.3	0.98940675	0.994689273
Romance	865	6.427398844	6.5	6.5	8.5	2.1	0.927593342	0.963116474
Family	444	6.204054054	6.3	6.7	8.6	1.9	1.357003844	1.164905079
Western	57	6.770175439	6.8	6.8	8.9	4.1	1.013558897	1.006756623
Musical	97	6.575257732	6.7	7.1	8.5	2.1	1.196464777	1.093830324
Thriller	1109	6.374752029	6.4	6.5	9	2.7	0.930787957	0.964773526



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

- Task: Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.
- Hint: Calculate descriptive statistics such as mean, median, and standard deviation for movie durations. Use Excel's functions like AVERAGE, MEDIAN, and STDEV. Create a scatter plot to visualize the relationship between movie duration and IMDB score. Add a trendline to assess the direction and strength of the relationship.

I calculated the descriptive statistics using the formulas:

```
=AVERAGE(B2:B3804)
```

```
=MEDIAN(B2:B3804)
```

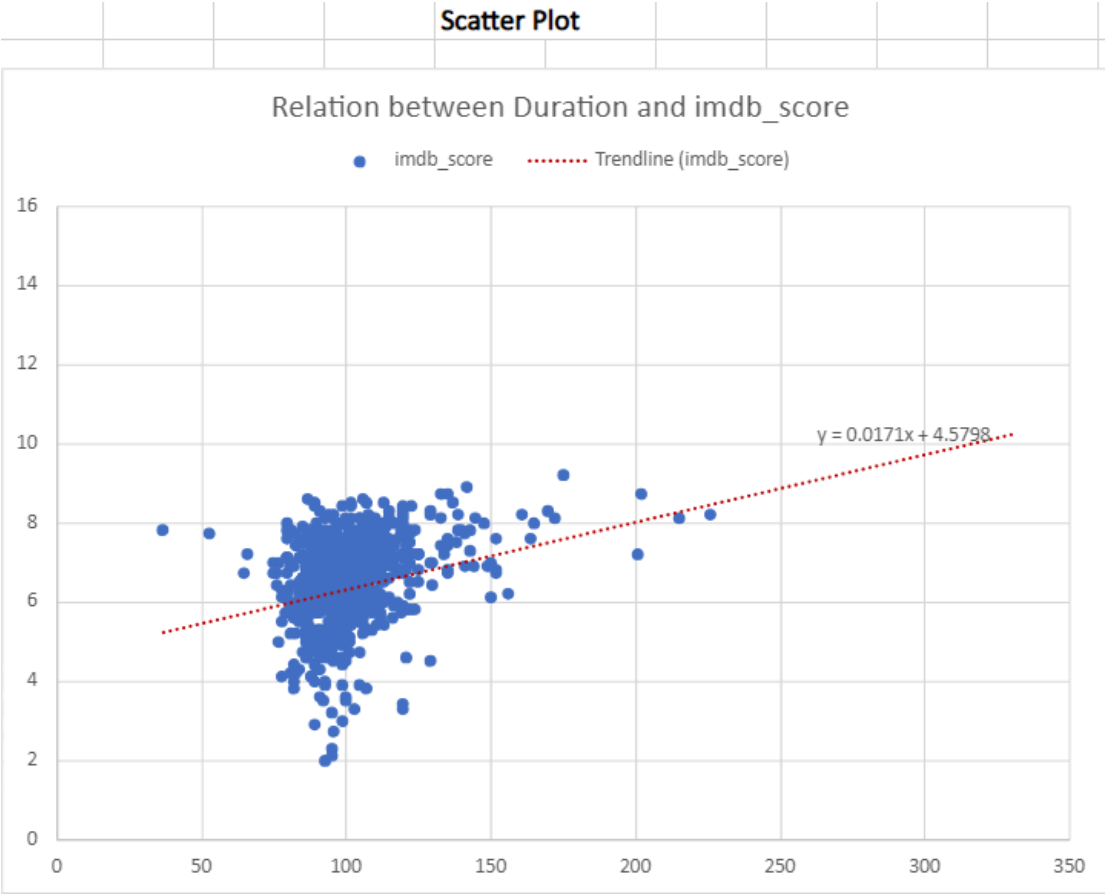
```
=MODE(B2:B3804)
```

```
=STDEV(B2:B3804)
```

```
=VAR(B2:B3804)
```

Movie Duration Analysis

Statistics	Values
Mean of duration	110.010518
Median of duration	106
Mode of duration	101
Standarad Deviation	22.61790418
Varience	511.5695895



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

- **Task:** Determine the most common languages used in movies and analyze their impact on the IMDB score using descriptive statistics.
- **Hint:** Use Excel's COUNTIF function to count the number of movies for each language. Calculate the mean, median, and standard deviation of the IMDB scores for each language. Compare the statistics to understand the impact of language on movie ratings.

First, I extracted the unique languages:

```
=UNIQUE('Cleaned Dataset'!I2:I3804)
```

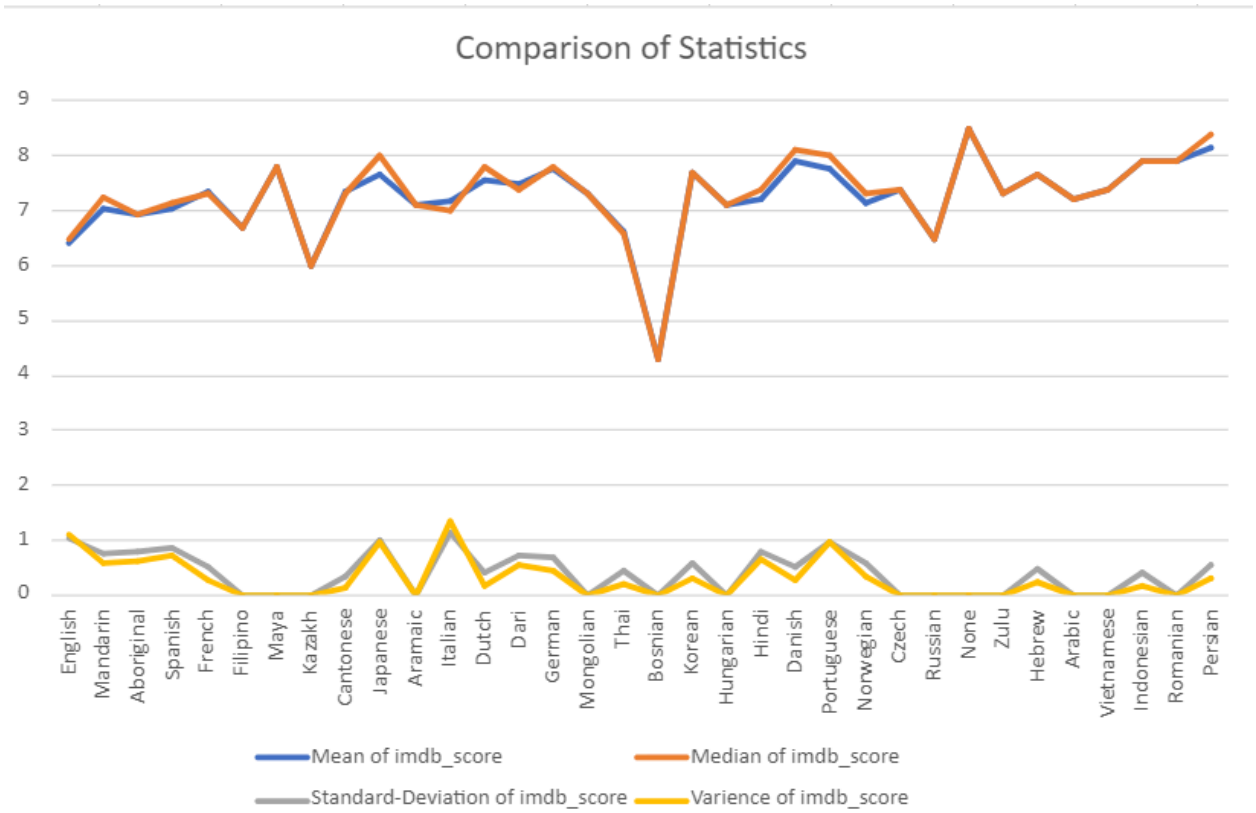
I counted the number of movies for each language:

```
=COUNTIF('Cleaned Dataset'!I2:I3804,"ENGLISH")
```

I calculated the mean, median, and standard deviation of the IMDB scores for each language:

```
=AVERAGEIFS('Cleaned Dataset'!N2:N3804,'Cleaned Dataset'!I2:I3804,"English")
=MEDIAN(IF(ISNUMBER(SEARCH("*" & A2 & "*",'Cleaned Dataset'!I2:I3804)), 'Cleaned Dataset'!N2:N3804))
=STDEV(IF(ISNUMBER(SEARCH("*" & A2 & "*",'Cleaned Dataset'!I2:I3804)), 'Cleaned Dataset'!N2:N3804))
=VAR(IF(ISNUMBER(SEARCH("*" & A2 & "*",'Cleaned Dataset'!I2:I3804)), 'Cleaned Dataset'!N2:N3804))
```

	Language	Number of Movies	Mean of imdb_score	Median of imdb_score	Standard-Deviation of imdb_score	Varience of imdb_score
2	English	3643	6.422426572	6.5	1.049182856	1.100784665
3	Mandarin	14	7.021428571	7.25	0.765786244	0.586428571
4	Aboriginal	2	6.95	6.95	0.777817459	0.605
5	Spanish	24	7.045833333	7.15	0.860727279	0.740851449
6	French	34	7.355882353	7.3	0.519435111	0.269812834
7	Filipino	1	6.7	6.7	#DIV/0!	#DIV/0!
8	Maya	1	7.8	7.8	#DIV/0!	#DIV/0!
9	Kazakh	1	6	6	#DIV/0!	#DIV/0!
10	Cantonese	7	7.342857143	7.3	0.350509833	0.122857143
11	Japanese	10	7.66	8	0.990173947	0.980444444
12	Aramaic	1	7.1	7.1	#DIV/0!	#DIV/0!
13	Italian	7	7.185714286	7	1.155318962	1.334761905
14	Dutch	3	7.566666667	7.8	0.404145188	0.163333333
15	Dari	2	7.5	7.4	0.732319375	0.536291667
16	German	11	7.763636364	7.8	0.675681474	0.456545455
17	Mongolian	1	7.3	7.3	#DIV/0!	#DIV/0!
18	Thai	3	6.633333333	6.6	0.450924975	0.203333333
19	Bosnian	1	4.3	4.3	#DIV/0!	#DIV/0!
20	Korean	5	7.7	7.7	0.570087713	0.325
21	Hungarian	1	7.1	7.1	#DIV/0!	#DIV/0!
22	Hindi	5	7.22	7.4	0.801249025	0.642
23	Danish	3	7.9	8.1	0.529150262	0.28
24	Portuguese	5	7.76	8	0.978774744	0.958
25	Norwegian	4	7.15	7.3	0.574456265	0.33
26	Czech	1	7.4	7.4	#DIV/0!	#DIV/0!
27	Russian	1	6.5	6.5	#DIV/0!	#DIV/0!
28	None	1	8.5	8.5	#DIV/0!	#DIV/0!
29	Zulu	1	7.3	7.3	#DIV/0!	#DIV/0!
30	Hebrew	2	7.65	7.65	0.494974747	0.245
31	Arabic	1	7.2	7.2	#DIV/0!	#DIV/0!
32	Vietnamese	1	7.4	7.4	#DIV/0!	#DIV/0!
33	Indonesian	2	7.9	7.9	0.424264069	0.18
34	Romanian	1	7.9	7.9	#DIV/0!	#DIV/0!
35	Persian	3	8.133333333	8.4	0.550757055	0.303333333



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

- Task: Identify the top directors based on their average IMDB score and analyze their contribution to the success of movies using percentile calculations.
- Hint: Calculate the average IMDB score for each director. Use Excel's PERCENTILE function to identify the directors with the highest scores. Compare the scores of these directors to the overall distribution of scores.

First, I extracted the unique director:

```
=UNIQUE('Cleaned Dataset'!A2:A3804)
```

And calculated the average IMDB score for each director:

```
=AVERAGEIF('Cleaned Dataset'!A2:A3804,A2,'Cleaned Dataset'!N2:N3804)
```

Used Excel's PERCENTILE function:

```
=PERCENTRANK.EXC(B2:B1709,B2)
```

Director	Average of imdb_score	Percentile
James Cameron	7.914285714	0.976
Gore Verbinski	6.985714286	0.729
Sam Mendes	7.5	0.893
Christopher Nolan	8.425	0.994
Andrew Stanton	7.733333333	0.954
Sam Raimi	6.85	0.686
Nathan Greno	7.8	0.959
Joss Whedon	7.866666667	0.97
David Yates	7.2	0.818
Zack Snyder	7.175	0.818
Bryan Singer	7.2875	0.854
Marc Forster	7.228571429	0.848
Andrew Adamson	7.15	0.813
Rob Marshall	6.6	0.564
Barry Sonnenfeld	6.457142857	0.509
Peter Jackson	7.675	0.94
Marc Webb	7.133333333	0.811
Ridley Scott	7.070588235	0.779
Chris Weitz	6.08	0.356
Anthony Russo	7	0.736
Peter Berg	6.666666667	0.603
Colin Trevorrow	7	0.736
Shane Black	7.4	0.879

Top 5 directors

Director	Percentile
Neill Dela Llana	1
Tony Kaye	0.994
Christopher Nolan	0.994
Sergio Leone	0.992
Alfred Hitchcock	0.991
Lee Unkrich	0.991

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

- Task: Analyze the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.
- Hint: Calculate the correlation coefficient between movie budgets and gross earnings using Excel's CORREL function. Calculate the profit margin (gross earnings - budget) for each movie and identify the movies with the highest profit margin using Excel's MAX function.

Calculated the correlation coefficient between movie budgets and gross earnings using Excel's CORREL function:

```
=CORREL(B2:B3804,C2:C3804)
```

Calculated the profit margin (gross earnings - budget) for each movie:

```
=B2-C2
```

movie_title	gross	budget	Profit margin
AvatarÂ	760505847	237000000	523505847
Pirates of the Caribbean: At World's EndÂ	309404152	300000000	9404152
SpectreÂ	200074175	245000000	-44925825
The Dark Knight RisesÂ	448130642	250000000	198130642
John CarterÂ	73058679	263700000	-190641321
Spider-Man 3Â	336530303	258000000	78530303
TangledÂ	200807262	260000000	-59192738
Avengers: Age of UltronÂ	458991599	250000000	208991599
Harry Potter and the Half-Blood PrinceÂ	301956980	250000000	51956980
Batman v Superman: Dawn of JusticeÂ	330249062	250000000	80249062
Superman ReturnsÂ	200069408	209000000	-8930592
Quantum of SolaceÂ	168368427	200000000	-31631573
Pirates of the Caribbean: Dead Man's ChestÂ	423032628	225000000	198032628
The Lone RangerÂ	89289910	215000000	-125710090
Man of SteelÂ	291021565	225000000	66021565
The Chronicles of Narnia: Prince CaspianÂ	141614023	225000000	-83385977
The AvengersÂ	623279547	220000000	403279547
Pirates of the Caribbean: On Stranger TidesÂ	241063875	250000000	-8936125
Men in Black 3Â	179020854	225000000	-45979146
The Hobbit: The Battle of the Five ArmiesÂ	255108370	250000000	5108370
The Amazing Spider-ManÂ	262030663	230000000	32030663
Robin HoodÂ	105219735	200000000	-94780265
The Hobbit: The Desolation of SmaugÂ	258355354	225000000	33355354

	Correlation between Gross and movie budgets
	0.100542