IMDb Movie Analysis

By

Stan Pereira

Content

01 Project Description

02 Approach

03 Tech-Stack Used

04 Insights

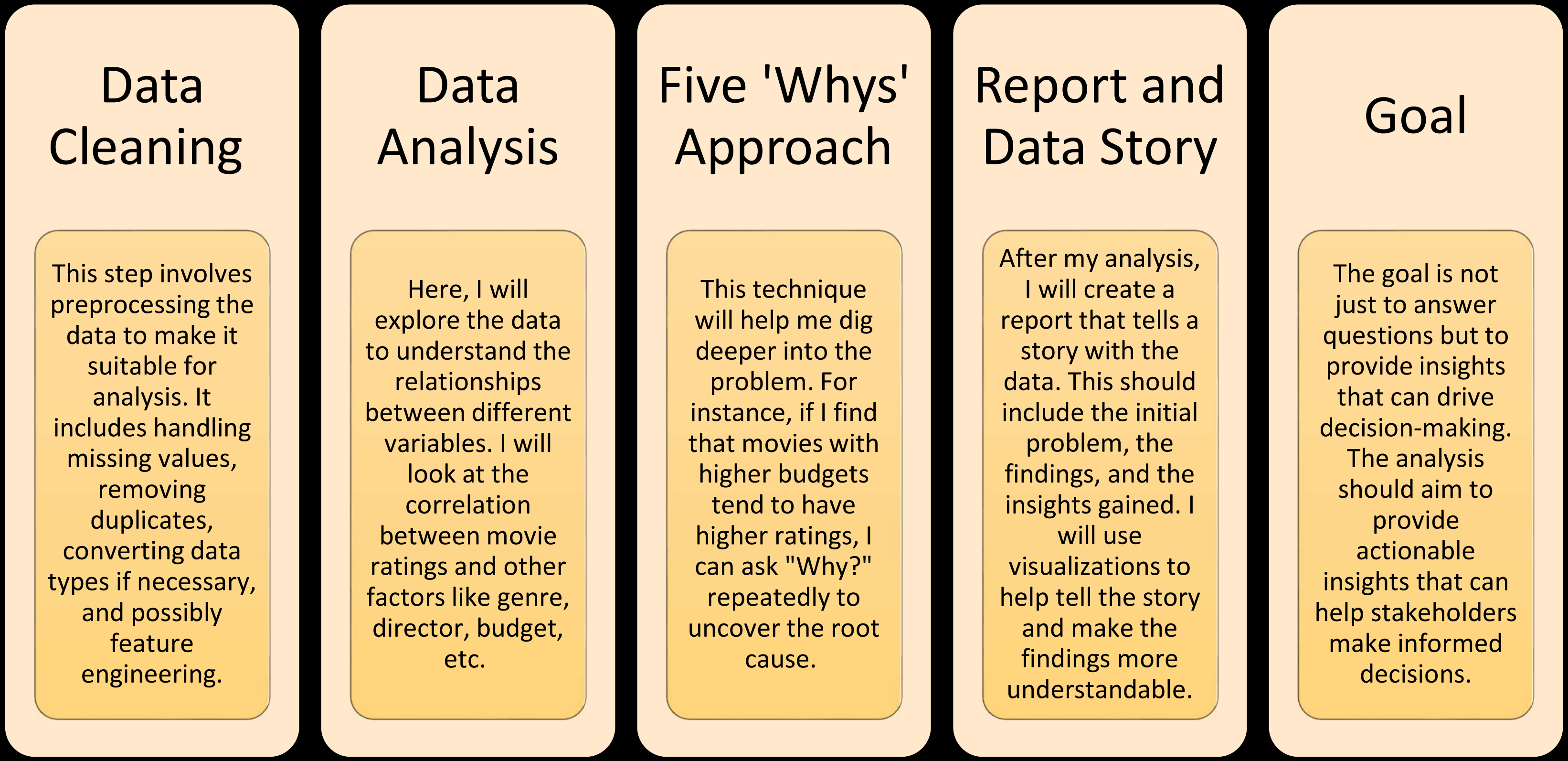
05 Results

Project Description

IMDb (***Internet Movie Database***) is an online database of information related to films, television series, podcasts, home videos, video games, and streaming content online – including cast, production crew and personal biographies, plot summaries, trivia, ratings, and fan and critical reviews.

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.

Approach



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.

Data Analysis

Here, I will explore the data to understand the relationships between different variables. I will look at the correlation between movie ratings and other factors like genre, director, budget, etc.

Five 'Whys' Approach

This technique will help me dig deeper into the problem. For instance, if I find that movies with higher budgets tend to have higher ratings, I can ask "Why?" repeatedly to uncover the root cause.

Report and Data Story

After my analysis, I will create a report that tells a story with the data. This should include the initial problem, the findings, and the insights gained. I will use visualizations to help tell the story and make the findings more understandable.

Goal

The goal is not just to answer questions but to provide insights that can drive decision-making. The analysis should aim to provide actionable insights that can help stakeholders make informed decisions.

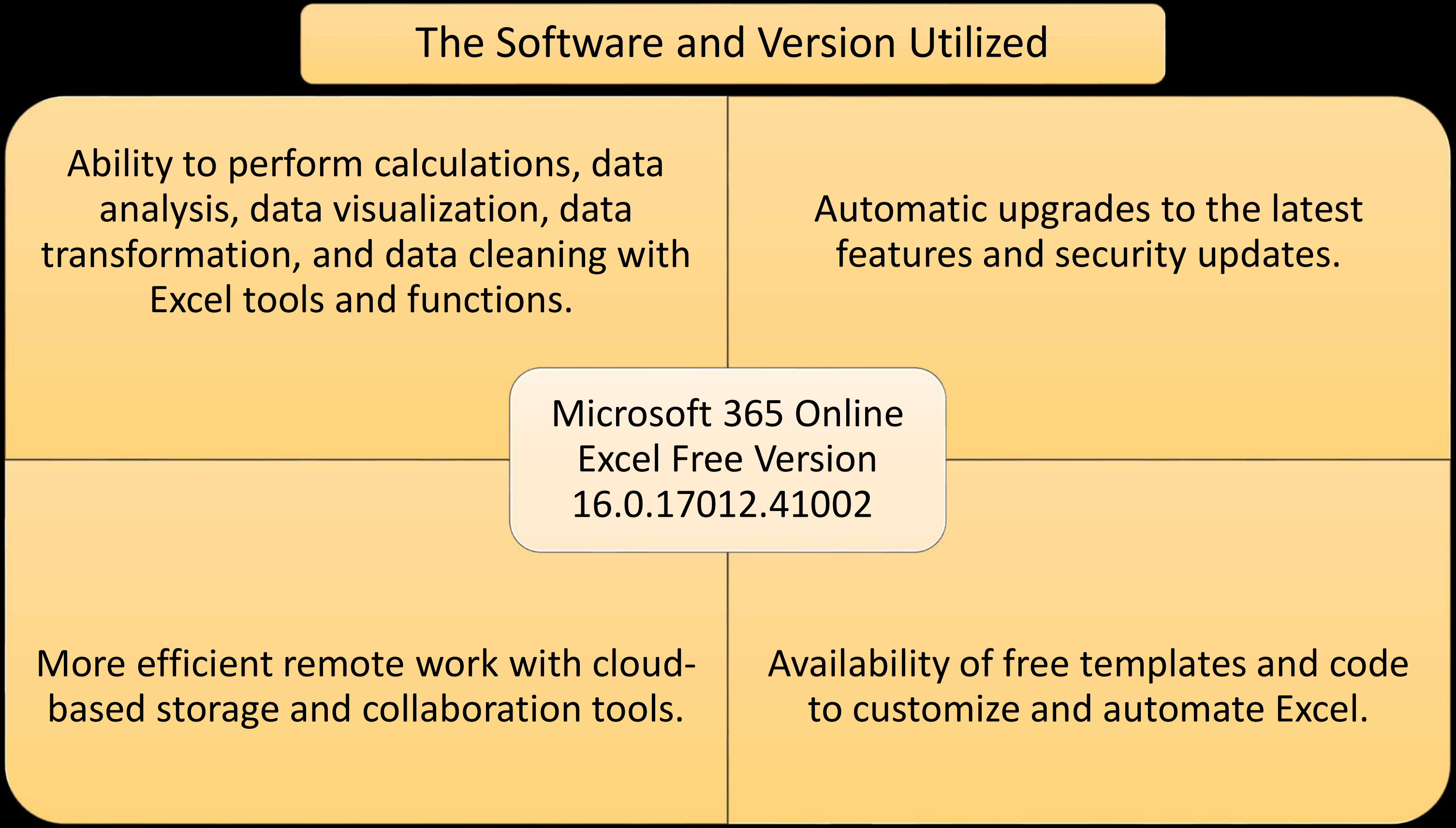
Data Cleaning

As part of the data cleaning, I made the below changes:

- Determined the main columns required : IMDb Score, Movie Title, Genre, Movie Duration, Language, Director Name, Budget & Gross Earnings
- Created a new column "Profit Margin" = ("Gross Earnings"-"Budget")
- Removed rows that had blanks in the required columns
- Removed duplicates in the required parameters – The Movie Title "The Host" was duplicate, but it was two different movies having the same name.

Total Number of Records before "Data Cleaning" was ***5043***
Total Number of Records after "Data Cleaning" became ***3786***

Tech Stack Used



The Software and Version Utilized

Ability to perform calculations, data analysis, data visualization, data transformation, and data cleaning with Excel tools and functions.

Automatic upgrades to the latest features and security updates.

Microsoft 365 Online Excel Free Version 16.0.17012.41002

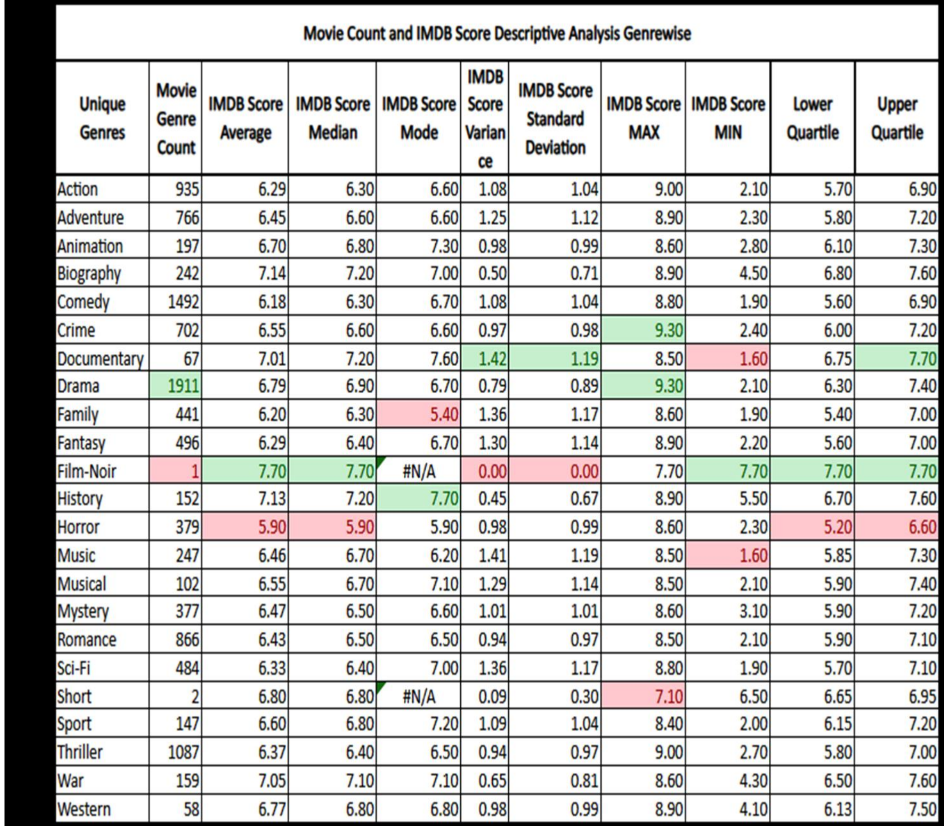
More efficient remote work with cloud-based storage and collaboration tools.

Availability of free templates and code to customize and automate Excel.

Charts & Insights

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

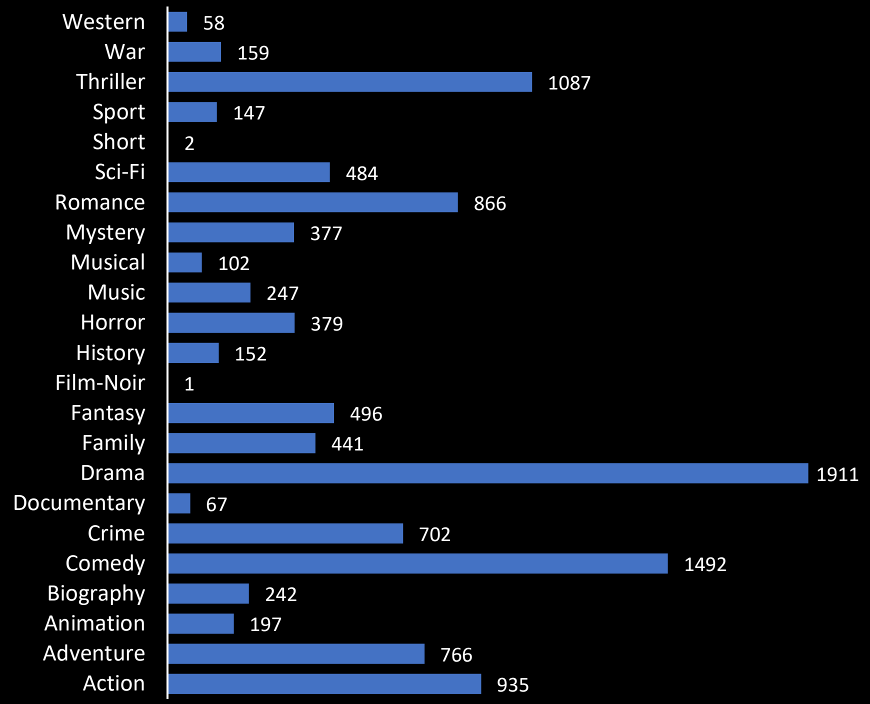
Task: Determine the most common genres of movies in the dataset. Then, for each genre, calculate descriptive statistics (mean, median, mode, range, variance, standard deviation) of the IMDB scores.

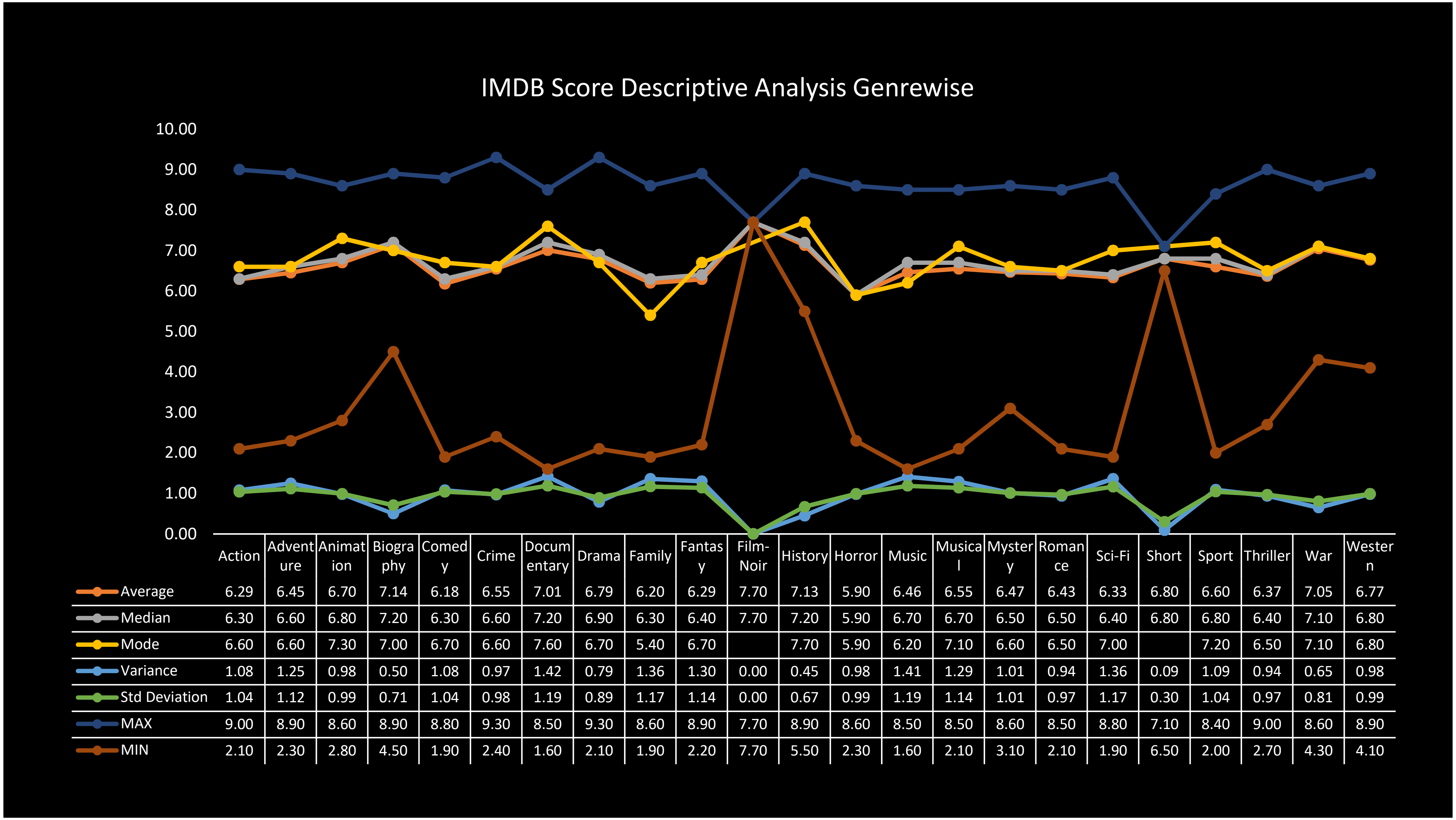


Movie Count and IMDB Score Descriptive Analysis Genrewise

Unique Genres	Movie Genre Count	IMDB Score Average	IMDB Score Median	IMDB Score Mode	IMDB Score Variance	IMDB Score Standard Deviation	IMDB Score MAX	IMDB Score MIN	Lower Quartile	Upper Quartile
Action	935	6.29	6.30	6.60	1.08	1.04	9.00	2.10	5.70	6.90
Adventure	766	6.45	6.60	6.60	1.25	1.12	8.90	2.30	5.80	7.20
Animation	197	6.70	6.80	7.30	0.98	0.99	8.60	2.80	6.10	7.30
Biography	242	7.14	7.20	7.00	0.50	0.71	8.90	4.50	6.80	7.60
Comedy	1492	6.18	6.30	6.70	1.08	1.04	8.80	1.90	5.60	6.90
Crime	702	6.55	6.60	6.60	0.97	0.98	9.30	2.40	6.00	7.20
Documentary	67	7.01	7.20	7.60	1.42	1.19	8.50	1.60	6.75	7.70
Drama	1911	6.79	6.90	6.70	0.79	0.89	9.30	2.10	6.30	7.40
Family	441	6.20	6.30	5.40	1.36	1.17	8.60	1.90	5.40	7.00
Fantasy	496	6.29	6.40	6.70	1.30	1.14	8.90	2.20	5.60	7.00
Film-Noir	1	7.70	7.70	#N/A	0.00	0.00	7.70	7.70	7.70	7.70
History	152	7.13	7.20	7.70	0.45	0.67	8.90	5.50	6.70	7.60
Horror	379	5.90	5.90	5.90	0.98	0.99	8.60	2.30	5.20	6.60
Music	247	6.46	6.70	6.20	1.41	1.19	8.50	1.60	5.85	7.30
Musical	102	6.55	6.70	7.10	1.29	1.14	8.50	2.10	5.90	7.40
Mystery	377	6.47	6.50	6.60	1.01	1.01	8.60	3.10	5.90	7.20
Romance	866	6.43	6.50	6.50	0.94	0.97	8.50	2.10	5.90	7.10
Sci-Fi	484	6.33	6.40	7.00	1.36	1.17	8.80	1.90	5.70	7.10
Short	2	6.80	6.80	#N/A	0.09	0.30	7.10	6.50	6.65	6.95
Sport	147	6.60	6.80	7.20	1.09	1.04	8.40	2.00	6.15	7.20
Thriller	1087	6.37	6.40	6.50	0.94	0.97	9.00	2.70	5.80	7.00
War	159	7.05	7.10	7.10	0.65	0.81	8.60	4.30	6.50	7.60
Western	58	6.77	6.80	6.80	0.98	0.99	8.90	4.10	6.13	7.50

Movie Count Genrewise



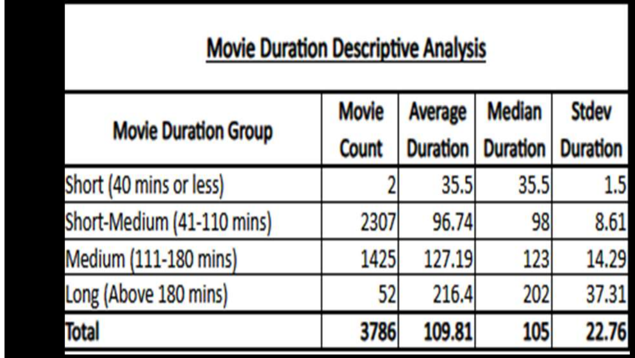


IMDB Score Descriptive Analysis Genrewise

	Action	Adventure	Animation	Biography	Comedy	Crime	Documentary	Drama	Family	Fantasy	Film-Noir	History	Horror	Music	Musical	Mystery	Romance	Sci-Fi	Short	Sport	Thriller	War	Western
Average	6.29	6.45	6.70	7.14	6.18	6.55	7.01	6.79	6.20	6.29	7.70	7.13	5.90	6.46	6.55	6.47	6.43	6.33	6.80	6.60	6.37	7.05	6.77
Median	6.30	6.60	6.80	7.20	6.30	6.60	7.20	6.90	6.30	6.40	7.70	7.20	5.90	6.70	6.70	6.50	6.50	6.40	6.80	6.80	6.40	7.10	6.80
Mode	6.60	6.60	7.30	7.00	6.70	6.60	7.60	6.70	5.40	6.70		7.70	5.90	6.20	7.10	6.60	6.50	7.00		7.20	6.50	7.10	6.80
Variance	1.08	1.25	0.98	0.50	1.08	0.97	1.42	0.79	1.36	1.30	0.00	0.45	0.98	1.41	1.29	1.01	0.94	1.36	0.09	1.09	0.94	0.65	0.98
Std Deviation	1.04	1.12	0.99	0.71	1.04	0.98	1.19	0.89	1.17	1.14	0.00	0.67	0.99	1.19	1.14	1.01	0.97	1.17	0.30	1.04	0.97	0.81	0.99
MAX	9.00	8.90	8.60	8.90	8.80	9.30	8.50	9.30	8.60	8.90	7.70	8.90	8.60	8.50	8.50	8.60	8.50	8.80	7.10	8.40	9.00	8.60	8.90
MIN	2.10	2.30	2.80	4.50	1.90	2.40	1.60	2.10	1.90	2.20	7.70	5.50	2.30	1.60	2.10	3.10	2.10	1.90	6.50	2.00	2.70	4.30	4.10

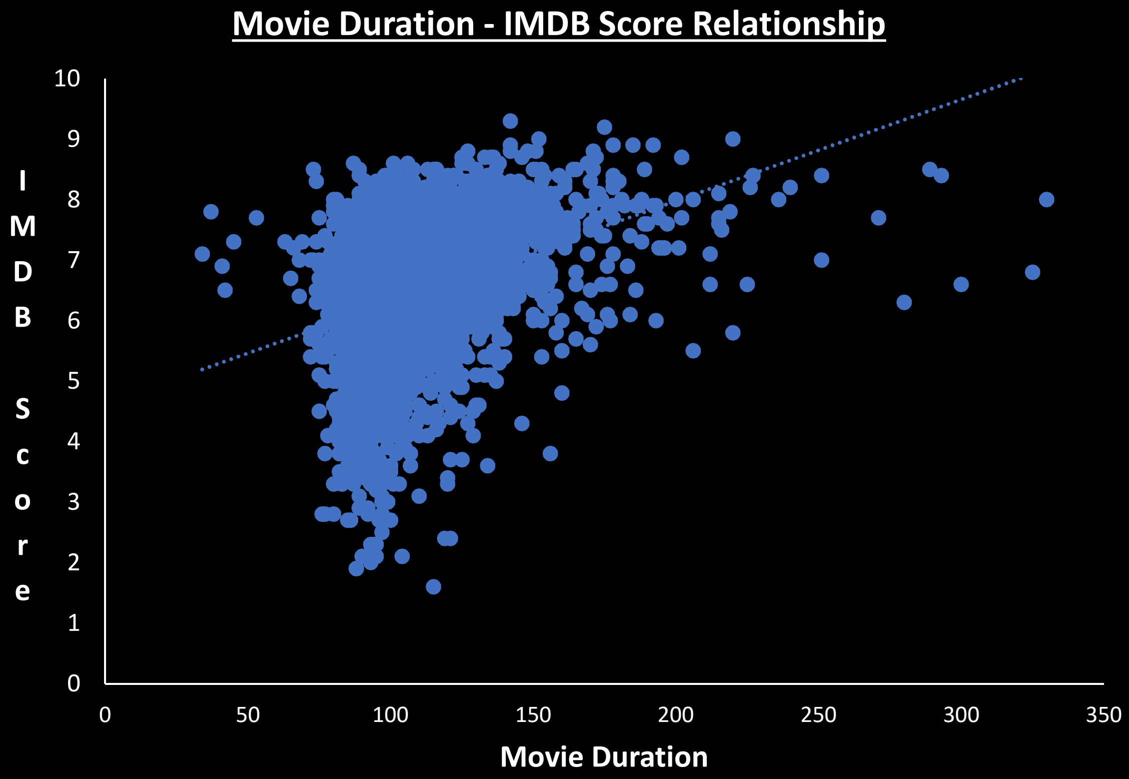
Q. B) Movie Duration Analysis: Analyze the distribution of movie durations and its impact on the IMDB score.

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



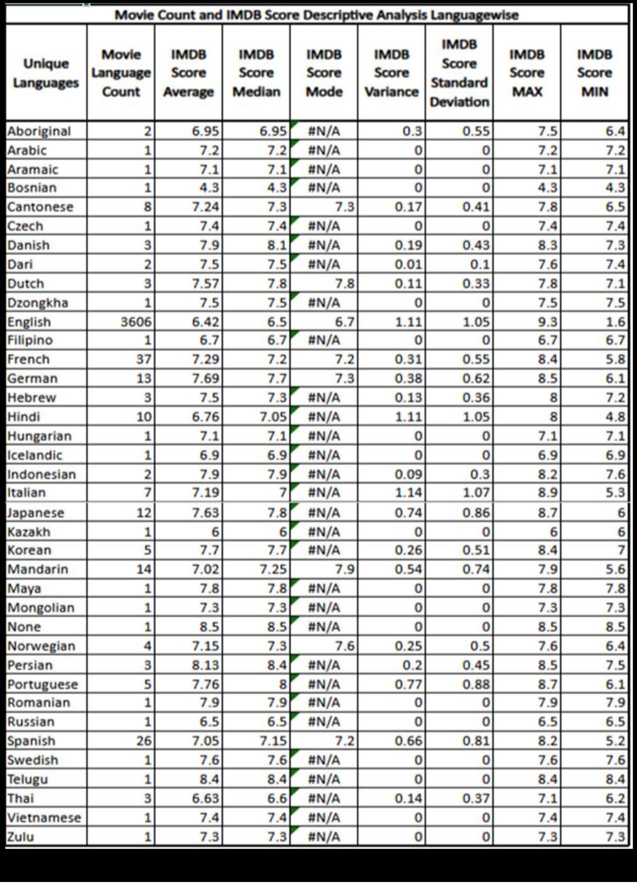
Movie Duration Descriptive Analysis

Movie Duration Group	Movie Count	Average Duration	Median Duration	Stdev Duration
Short (40 mins or less)	2	35.5	35.5	1.5
Short-Medium (41-110 mins)	2307	96.74	98	8.61
Medium (111-180 mins)	1425	127.19	123	14.29
Long (Above 180 mins)	52	216.4	202	37.31
Total	**3786**	**109.81**	**105**	**22.76**



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

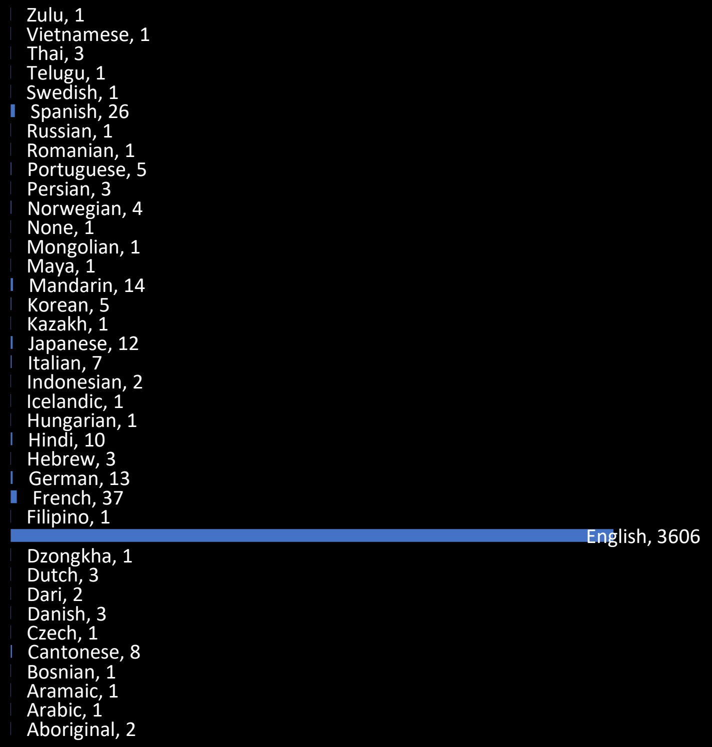
Task: Determine the most common languages used in movies and analyze their impact on the IMDB score using descriptive statistics.



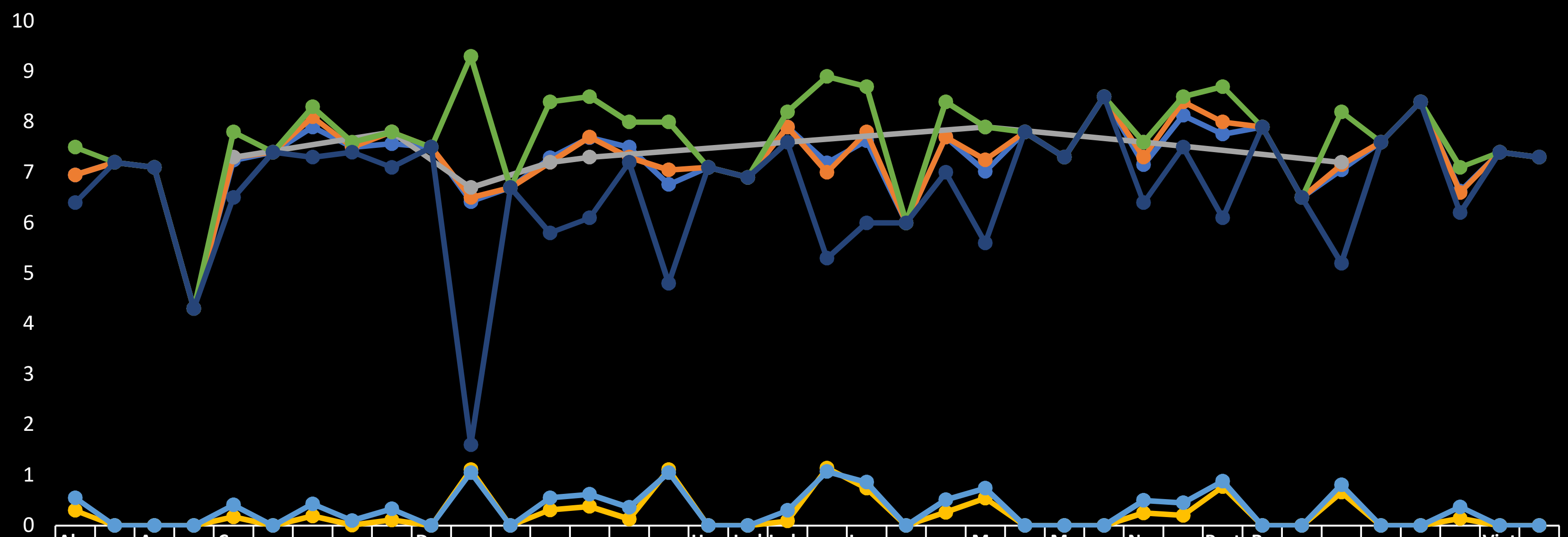
Movie Count and IMDB Score Descriptive Analysis Languagewise

Unique Languages	Movie Language Count	IMDB Score Average	IMDB Score Median	IMDB Score Mode	IMDB Score Variance	IMDB Score Standard Deviation	IMDB Score MAX	IMDB Score MIN
Aboriginal	2	6.95	6.95	#N/A	0.3	0.55	7.5	6.4
Arabic	1	7.2	7.2	#N/A	0	0	7.2	7.2
Aramaic	1	7.1	7.1	#N/A	0	0	7.1	7.1
Bosnian	1	4.3	4.3	#N/A	0	0	4.3	4.3
Cantonese	8	7.24	7.3	7.3	0.17	0.41	7.8	6.5
Czech	1	7.4	7.4	#N/A	0	0	7.4	7.4
Danish	3	7.9	8.1	#N/A	0.19	0.43	8.3	7.3
Dari	2	7.5	7.5	#N/A	0.01	0.1	7.6	7.4
Dutch	3	7.57	7.8	7.8	0.11	0.33	7.8	7.1
Dzongkha	1	7.5	7.5	#N/A	0	0	7.5	7.5
English	3606	6.42	6.5	6.7	1.11	1.05	9.3	1.6
Filipino	1	6.7	6.7	#N/A	0	0	6.7	6.7
French	37	7.29	7.2	7.2	0.31	0.55	8.4	5.8
German	13	7.69	7.7	7.3	0.38	0.62	8.5	6.1
Hebrew	3	7.5	7.3	#N/A	0.13	0.36	8	7.2
Hindi	10	6.76	7.05	#N/A	1.11	1.05	8	4.8
Hungarian	1	7.1	7.1	#N/A	0	0	7.1	7.1
Icelandic	1	6.9	6.9	#N/A	0	0	6.9	6.9
Indonesian	2	7.9	7.9	#N/A	0.09	0.3	8.2	7.6
Italian	7	7.19	7	#N/A	1.14	1.07	8.9	5.3
Japanese	12	7.63	7.8	#N/A	0.74	0.86	8.7	6
Kazakh	1	6	6	#N/A	0	0	6	6
Korean	5	7.7	7.7	#N/A	0.26	0.51	8.4	7
Mandarin	14	7.02	7.25	7.9	0.54	0.74	7.9	5.6
Maya	1	7.8	7.8	#N/A	0	0	7.8	7.8
Mongolian	1	7.3	7.3	#N/A	0	0	7.3	7.3
None	1	8.5	8.5	#N/A	0	0	8.5	8.5
Norwegian	4	7.15	7.3	7.6	0.25	0.5	7.6	6.4
Persian	3	8.13	8.4	#N/A	0.2	0.45	8.5	7.5
Portuguese	5	7.76	8	#N/A	0.77	0.88	8.7	6.1
Romanian	1	7.9	7.9	#N/A	0	0	7.9	7.9
Russian	1	6.5	6.5	#N/A	0	0	6.5	6.5
Spanish	26	7.05	7.15	7.2	0.66	0.81	8.2	5.2
Swedish	1	7.6	7.6	#N/A	0	0	7.6	7.6
Telugu	1	8.4	8.4	#N/A	0	0	8.4	8.4
Thai	3	6.63	6.6	#N/A	0.14	0.37	7.1	6.2
Vietnamese	1	7.4	7.4	#N/A	0	0	7.4	7.4
Zulu	1	7.3	7.3	#N/A	0	0	7.3	7.3

Movie Language Count



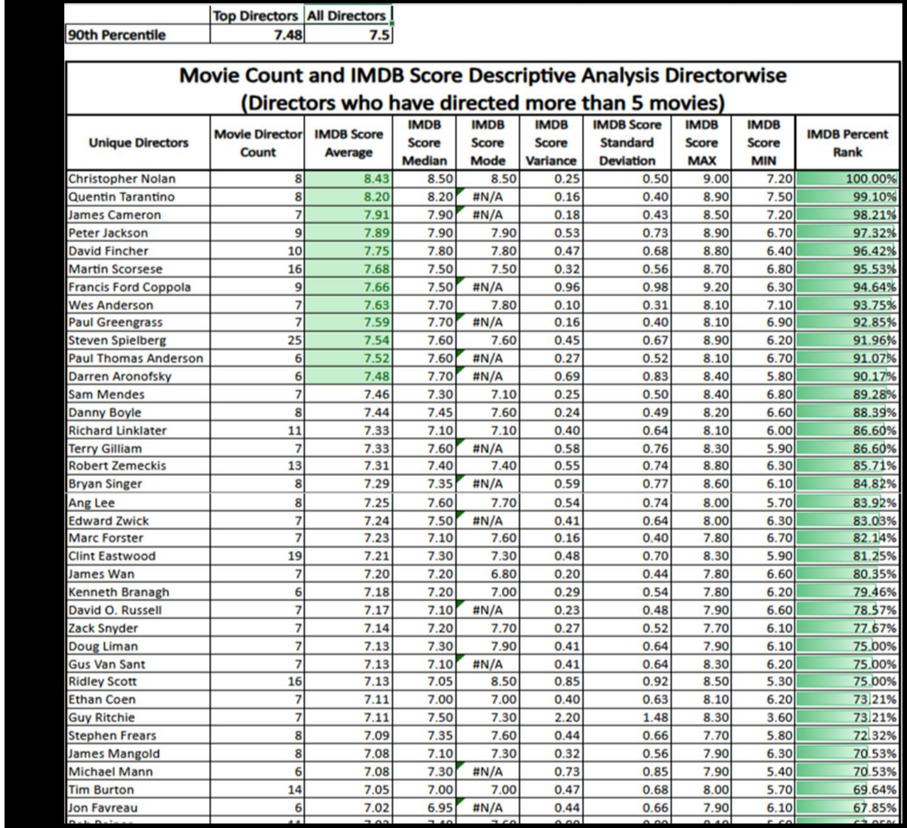
IMDB Score Descriptive Analysis Languagewise



	Aboriginal	Arabic	Aramaic	Bosnian	Cantonese	Czech	Danish	Dari	Dutch	Dzongkha	English	Filipino	French	German	Hebrew	Hindi	Hungarian	Icelandic	Indonesian	Italian	Japanese	Kazakh	Korean	Mandarin	Maya	Mongolian	None	Norwegian	Persian	Portuguese	Romanian	Russian	Spanish	Swedish	Telugu	Thai	Vietnamese	Zulu
Average	6.95	7.2	7.1	4.3	7.24	7.4	7.9	7.5	7.57	7.5	6.42	6.7	7.29	7.69	7.5	6.76	7.1	6.9	7.9	7.19	7.63	6	7.7	7.02	7.8	7.3	8.5	7.15	8.13	7.76	7.9	6.5	7.05	7.6	8.4	6.63	7.4	7.3
Median	6.95	7.2	7.1	4.3	7.3	7.4	8.1	7.5	7.8	7.5	6.5	6.7	7.2	7.7	7.3	7.05	7.1	6.9	7.9	7	7.8	6	7.7	7.25	7.8	7.3	8.5	7.3	8.4	8	7.9	6.5	7.15	7.6	8.4	6.6	7.4	7.3
Mode					7.3				7.8		6.7		7.2	7.3										7.9				7.6					7.2					
Variance	0.3	0	0	0	0.17	0	0.19	0.01	0.11	0	1.11	0	0.31	0.38	0.13	1.11	0	0	0.09	1.14	0.74	0	0.26	0.54	0	0	0	0.25	0.2	0.77	0	0	0.66	0	0	0.14	0	0
Std Dev	0.55	0	0	0	0.41	0	0.43	0.1	0.33	0	1.05	0	0.55	0.62	0.36	1.05	0	0	0.3	1.07	0.86	0	0.51	0.74	0	0	0	0.5	0.45	0.88	0	0	0.81	0	0	0.37	0	0
MAX	7.5	7.2	7.1	4.3	7.8	7.4	8.3	7.6	7.8	7.5	9.3	6.7	8.4	8.5	8	8	7.1	6.9	8.2	8.9	8.7	6	8.4	7.9	7.8	7.3	8.5	7.6	8.5	8.7	7.9	6.5	8.2	7.6	8.4	7.1	7.4	7.3
MIN	6.4	7.2	7.1	4.3	6.5	7.4	7.3	7.4	7.1	7.5	1.6	6.7	5.8	6.1	7.2	4.8	7.1	6.9	7.6	5.3	6	6	7	5.6	7.8	7.3	8.5	6.4	7.5	6.1	7.9	6.5	5.2	7.6	8.4	6.2	7.4	7.3

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

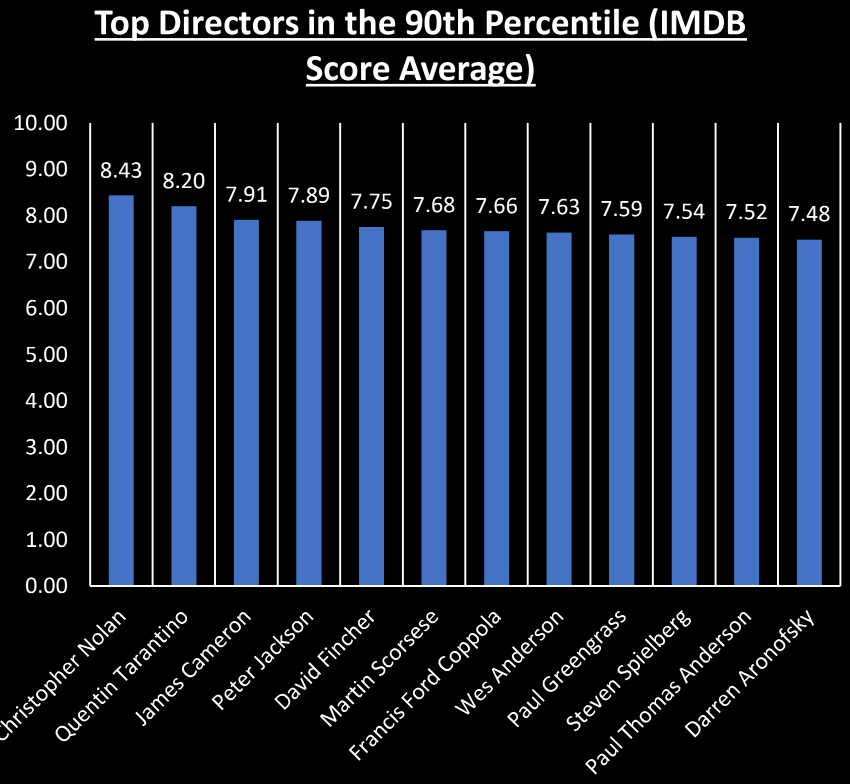
Task: Identify the top directors based on their average IMDB score and analyze their contribution to the success of movies using percentile calculations.



	Top Directors	All Directors
90th Percentile	7.48	7.5

Movie Count and IMDB Score Descriptive Analysis Directorwise (Directors who have directed more than 5 movies)

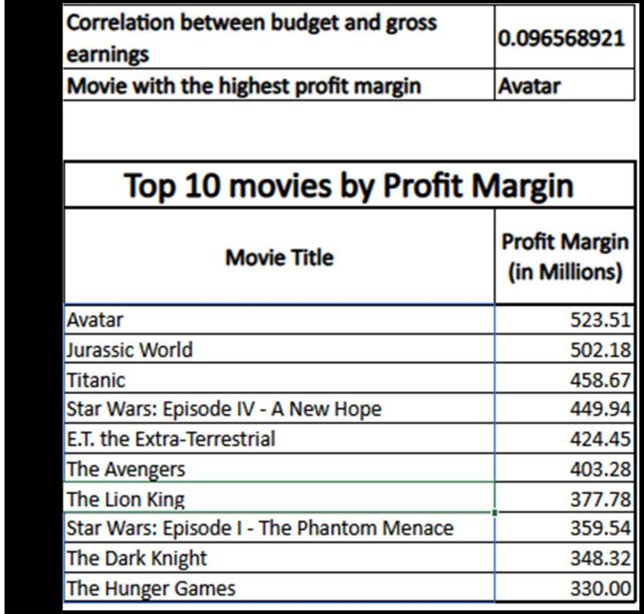
Unique Directors	Movie Director Count	IMDB Score Average	IMDB Score Median	IMDB Score Mode	IMDB Score Variance	IMDB Score Standard Deviation	IMDB Score MAX	IMDB Score MIN	IMDB Percent Rank
Christopher Nolan	8	8.43	8.50	8.50	0.25	0.50	9.00	7.20	100.00%
Quentin Tarantino	8	8.20	8.20	#N/A	0.16	0.40	8.90	7.50	99.10%
James Cameron	7	7.91	7.90	#N/A	0.18	0.43	8.50	7.20	98.21%
Peter Jackson	9	7.89	7.90	7.90	0.53	0.73	8.90	6.70	97.32%
David Fincher	10	7.75	7.80	7.80	0.47	0.68	8.80	6.40	96.42%
Martin Scorsese	16	7.68	7.50	7.50	0.32	0.56	8.70	6.80	95.53%
Francis Ford Coppola	9	7.66	7.50	#N/A	0.96	0.98	9.20	6.30	94.64%
Wes Anderson	7	7.63	7.70	7.80	0.10	0.31	8.10	7.10	93.75%
Paul Greengrass	7	7.59	7.70	#N/A	0.16	0.40	8.10	6.90	92.85%
Steven Spielberg	25	7.54	7.60	7.60	0.45	0.67	8.90	6.20	91.96%
Paul Thomas Anderson	6	7.52	7.60	#N/A	0.27	0.52	8.10	6.70	91.07%
Darren Aronofsky	6	7.48	7.70	#N/A	0.69	0.83	8.40	5.80	90.17%
Sam Mendes	7	7.46	7.30	7.10	0.25	0.50	8.40	6.80	89.28%
Danny Boyle	8	7.44	7.45	7.60	0.24	0.49	8.20	6.60	88.39%
Richard Linklater	11	7.33	7.10	7.10	0.40	0.64	8.10	6.00	86.60%
Terry Gilliam	7	7.33	7.60	#N/A	0.58	0.76	8.30	5.90	86.60%
Robert Zemeckis	13	7.31	7.40	7.40	0.55	0.74	8.80	6.30	85.71%
Bryan Singer	8	7.29	7.35	#N/A	0.59	0.77	8.60	6.10	84.82%
Ang Lee	8	7.25	7.60	7.70	0.54	0.74	8.00	5.70	83.92%
Edward Zwick	7	7.24	7.50	#N/A	0.41	0.64	8.00	6.30	83.03%
Marc Forster	7	7.23	7.10	7.60	0.16	0.40	7.80	6.70	82.14%
Clint Eastwood	19	7.21	7.30	7.30	0.48	0.70	8.30	5.90	81.25%
James Wan	7	7.20	7.20	6.80	0.20	0.44	7.80	6.60	80.35%
Kenneth Branagh	6	7.18	7.20	7.00	0.29	0.54	7.80	6.20	79.46%
David O. Russell	7	7.17	7.10	#N/A	0.23	0.48	7.90	6.60	78.57%
Zack Snyder	7	7.14	7.20	7.70	0.27	0.52	7.70	6.10	77.67%
Doug Liman	7	7.13	7.30	7.90	0.41	0.64	7.90	6.10	75.00%
Gus Van Sant	7	7.13	7.10	#N/A	0.41	0.64	8.30	6.20	75.00%
Ridley Scott	16	7.13	7.05	8.50	0.85	0.92	8.50	5.30	75.00%
Ethan Coen	7	7.11	7.00	7.00	0.40	0.63	8.10	6.20	73.21%
Guy Ritchie	7	7.11	7.50	7.30	2.20	1.48	8.30	3.60	73.21%
Stephen Frears	8	7.09	7.35	7.60	0.44	0.66	7.70	5.80	72.32%
James Mangold	8	7.08	7.10	7.30	0.32	0.56	7.90	6.30	70.53%
Michael Mann	6	7.08	7.30	#N/A	0.73	0.85	7.90	5.40	70.53%
Tim Burton	14	7.05	7.00	7.00	0.47	0.68	8.00	5.70	69.64%
Jon Favreau	6	7.02	6.95	#N/A	0.44	0.66	7.90	6.10	67.85%



Top Directors in the 90th Percentile (IMDB Score Average)

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

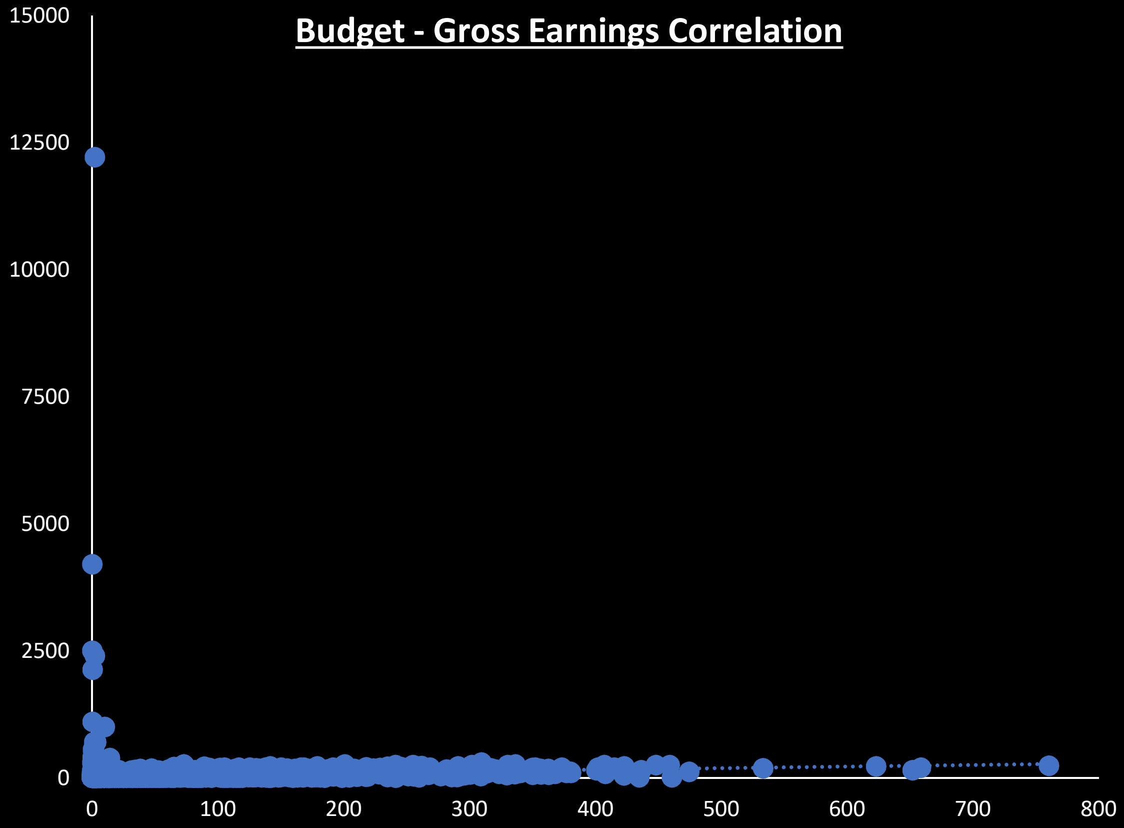
Task: Analyze the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.

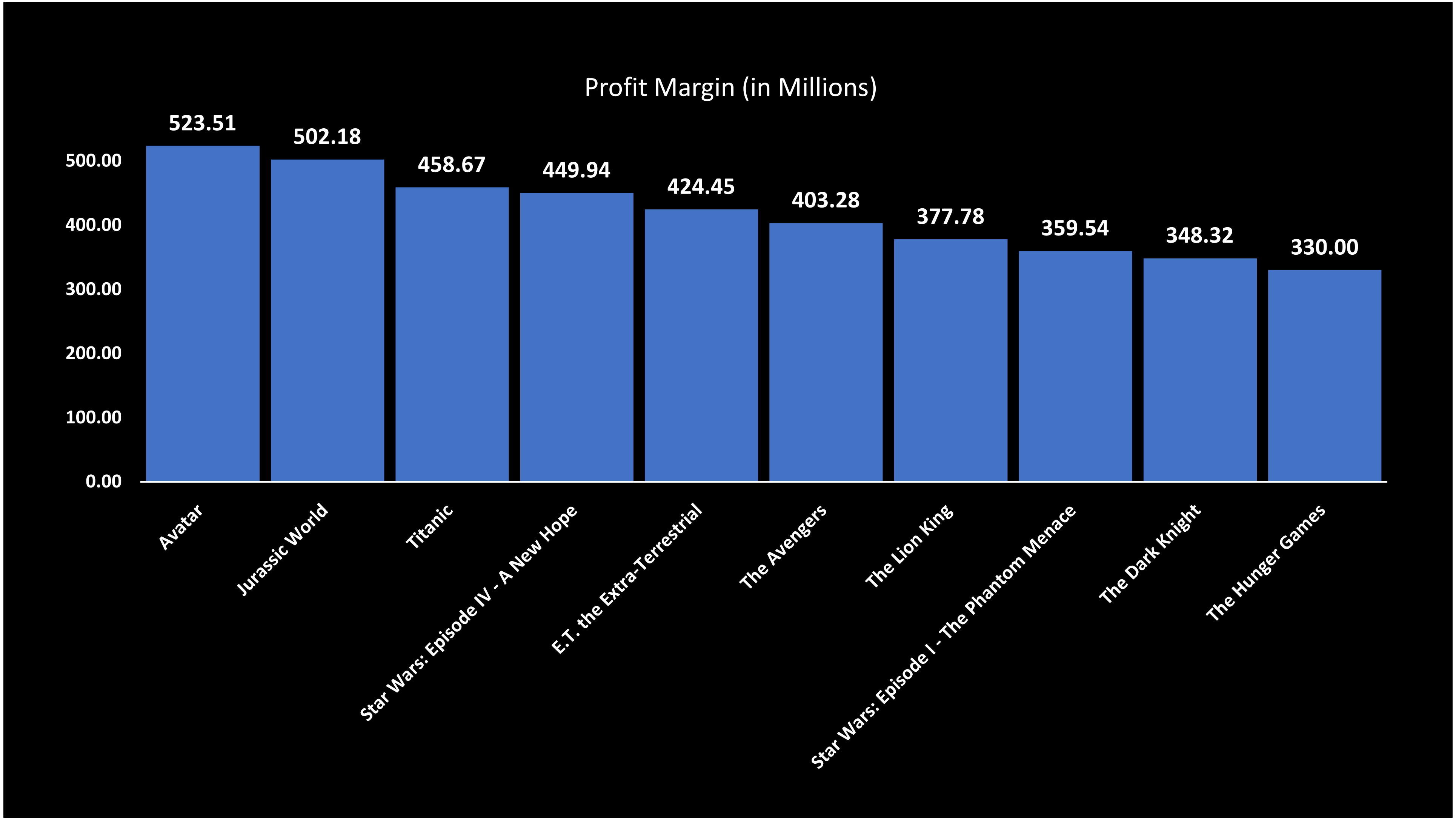


Correlation between budget and gross earnings	**0.096568921**
Movie with the highest profit margin	**Avatar**

Top 10 movies by Profit Margin

Movie Title	**Profit Margin (in Millions)**
Avatar	523.51
Jurassic World	502.18
Titanic	458.67
Star Wars: Episode IV - A New Hope	449.94
E.T. the Extra-Terrestrial	424.45
The Avengers	403.28
The Lion King	377.78
Star Wars: Episode I - The Phantom Menace	359.54
The Dark Knight	348.32
The Hunger Games	330.00




Profit Margin (in Millions)
523.51
502.18
458.67
449.94
424.45
403.28
377.78
359.54
348.32
330.00
500.00
400.00
300.00
200.00
100.00
0.00
Avatar
Jurassic World
Titanic
Star Wars: Episode IV - A New Hope
E.T. the Extra-Terrestrial
The Avengers
The Lion King
Star Wars: Episode I - The Phantom Menace
The Dark Knight
The Hunger Games

Insights

- ***Drama*** is the most common genre, followed by ***Comedy*** and ***Thriller.***
- ***Biography*** has the best IMDB Average (Film-Noir is not taken into account as it has only 1 movie), followed by ***History*** and ***War***.
- The mean, median and mode of all genres are mostly similar.
- Most movies made are in the ***41-110 mins*** duration category.
- The average duration for a movie is ***110 mins***.
- The relationship between movie duration and imdb scores has a ***slight upward*** trend.
- The most popular language is **English**.
- There are ***1751*** unique directors, but only ***113*** have directed over ***5 movies*** and are used to rate the best.
- ***Steven Spielberg*** has directed the most number of movies at ***25***.
- ***Christopher Nolan*** is the most highly rated director at ***8.43***.
- ***12*** directors are in the 90th percentile based on IMDB Score Average which ***above 7.48***.
- Correlation between budget and gross earnings is ***0.096*** which means that there is ***little or no relation*** between Budget and Gross Earnings.
- The movie with the highest profit margin is ***'Avatar'***.

Results

The ***'IMDb Movie Analysis'*** is important for movie producers, directors, and investors who want to understand what makes a movie successful. This allows them to make informed decisions in their future projects.

The project has helped me understand the relationship between various fields, like ***'Genre & IMDb Score'*** and ***'Budget & Earnings'***. It has also helped me improve my knowledge of ***Data Analysis*** using Excel and Statistics.

This project allowed me to ask the question ***'Why?'***, making me dive deeper to finally find the root cause.

Link to IMDB Movie Analysis Excel File

THANK YOU