IMDB Movie Analysis

Project Description

The project involves assuming the role of a data analyst to grasp the statistics related to the IMDB rated movies. Using the fundamental statistical concepts, we'll manipulate the data to perform further analysis. Additionally, we'll explore how Excel functions can efficiently extract information essential for a data analyst's routine tasks. Teaching the practical application of these functions is the primary aim of this project.

Approach

The approach will be to perform the tasks in an organized way, creating separate sheets in the workbook for different tasks performed.
The project analysis follows the below mentioned steps:

Step 1: Data Cleaning

Step 2: Genre Analysis

Step 3: Duration Analysis

Step 4: Language Analysis

Step 5: Director Analysis

Step 6: Budget Analysis

Tech-Stack Used

The analysis for this project has been performed using Microsoft Excel as it offers a user-friendly interface that allows users to work with data, perform calculations, and create visual representations without extensive programming knowledge.

Data Cleaning

The steps involve:

- duplicates removal
- null values detection & imputation
- outlier detection - removal/imputation

45 duplicate data points were removed

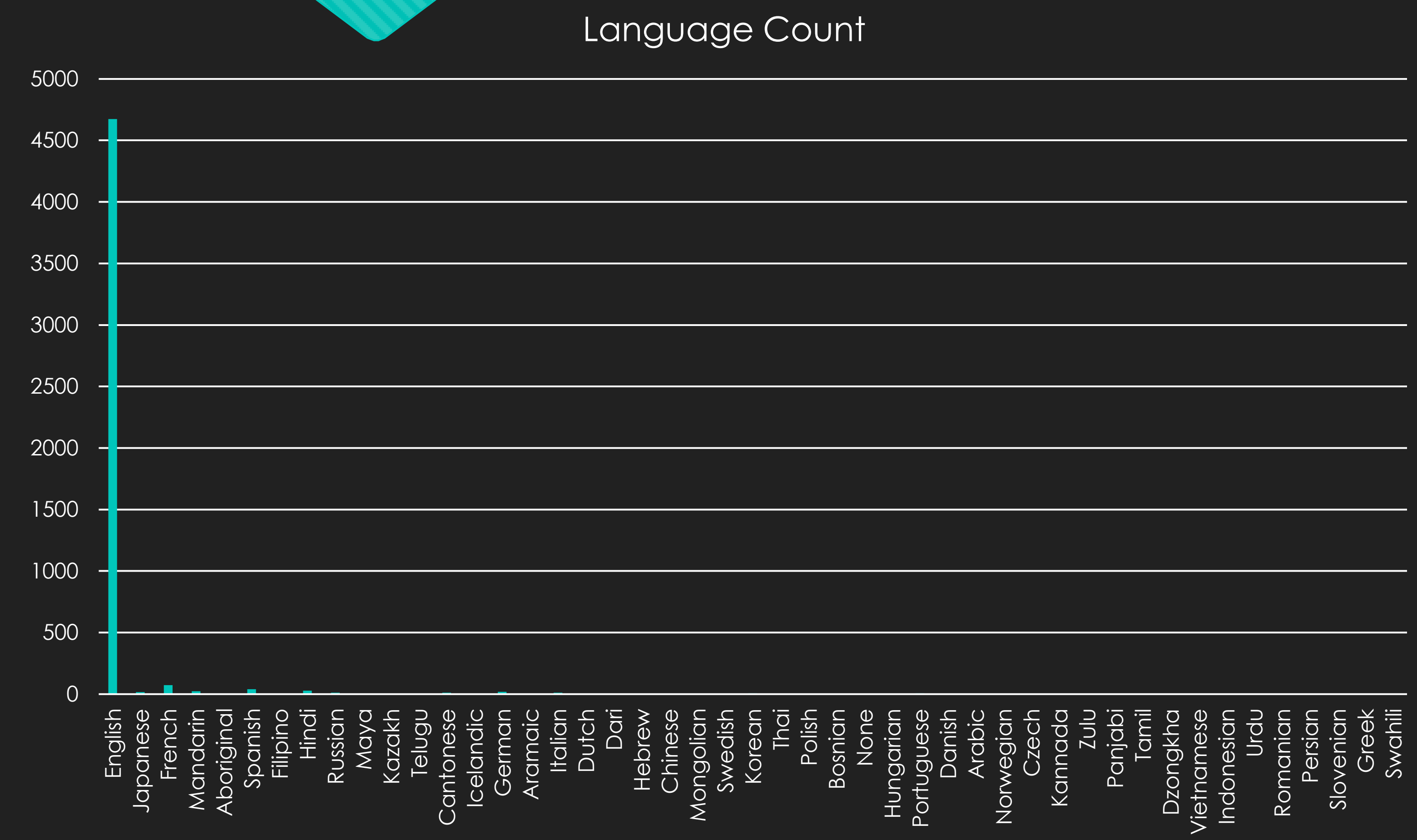
No columns were dropped due to high null values, threshold being 30%

Columns deleted

Columns not required for the analysis were removed

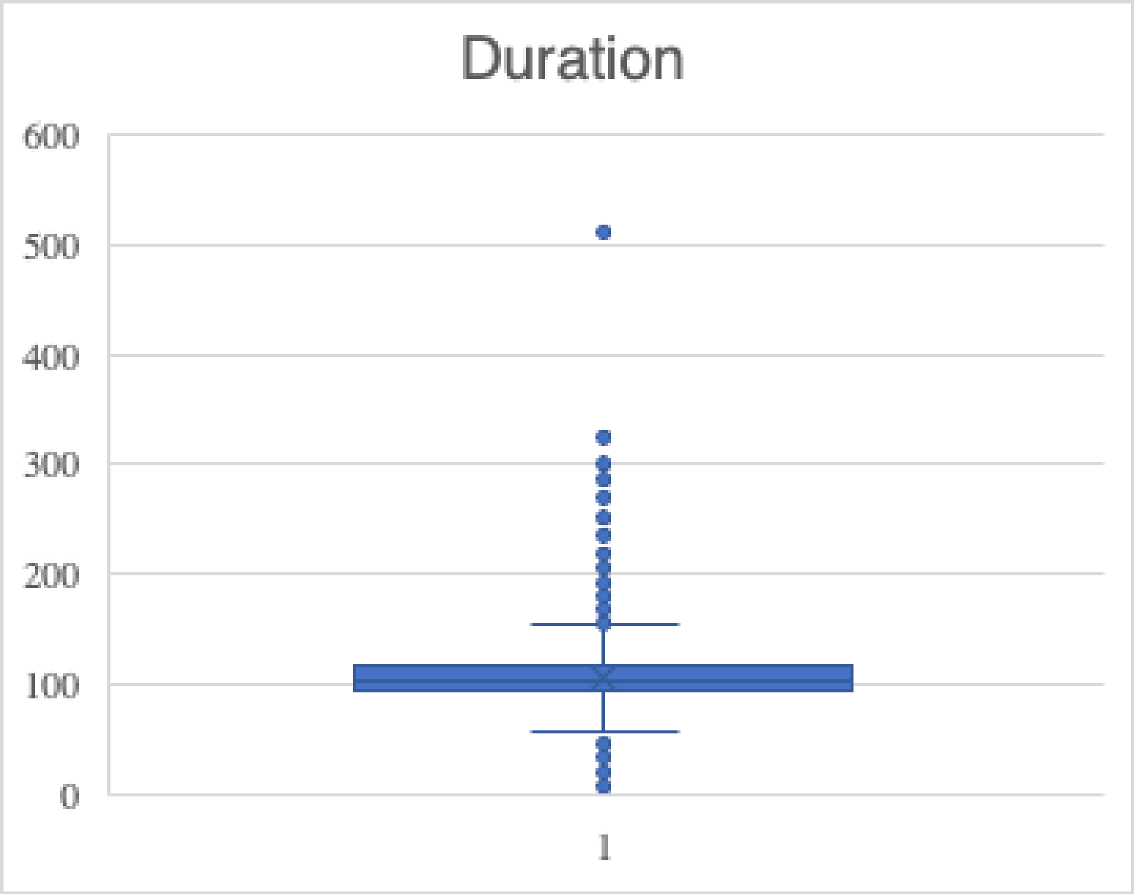
1	Color
2	num_critic_for_reviews
3	director_facebook_likes
4	actor_3_facebook_likes
5	actor_2_name
6	actor_1_facebook_likes
7	actor_1_name
8	num_voted_users
9	cast_total_facebook_likes
10	actor_3_name
11	facenumber_in_poster
12	plot_keywords
13	movie_imdb_link
14	num_user_for_reviews
15	content_rating
16	actor_2_facebook_likes
17	aspect_ratio
18	movie_facebook_likes
19	country
20	title_year

Data Imputation – Language



English has been filled in the null cells due to the highest count

Data Imputation – Duration

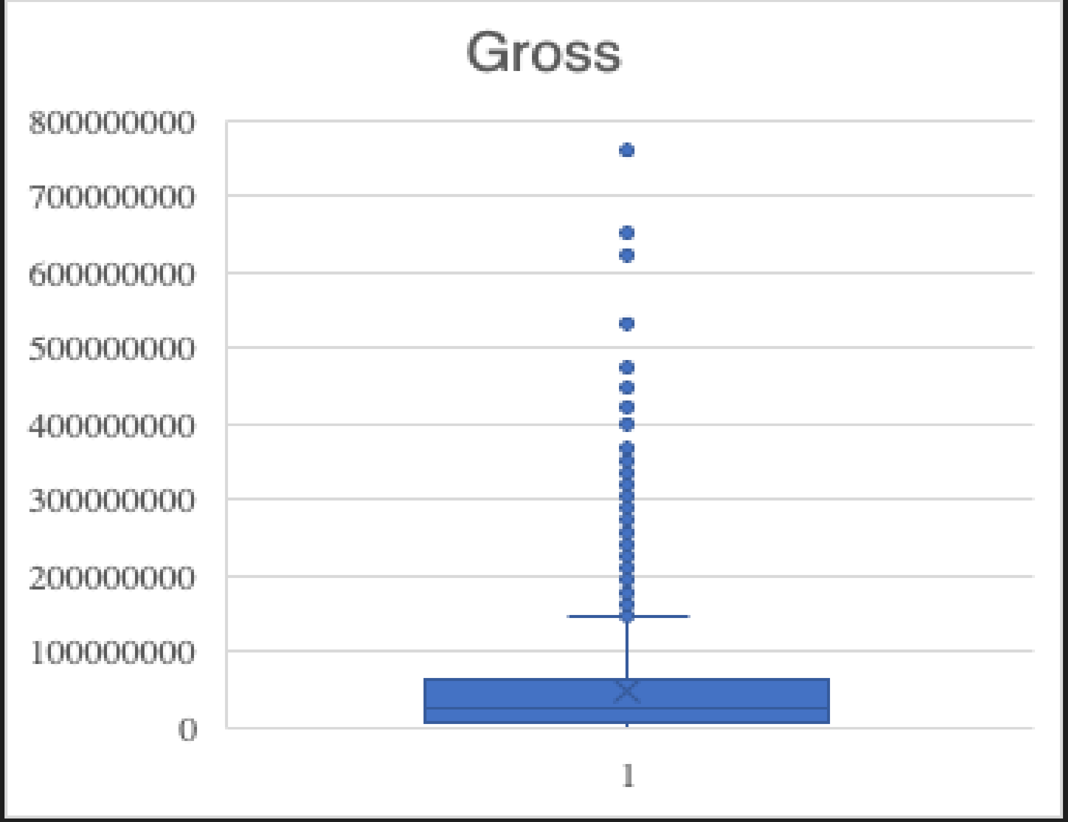


Median value 103 is put in the null value cells

Outlier value is not replaced with median as movie with such duration is a possibility

The data has outliers, so median values is chosen instead of other measures of central tendency

Data Imputation – Gross

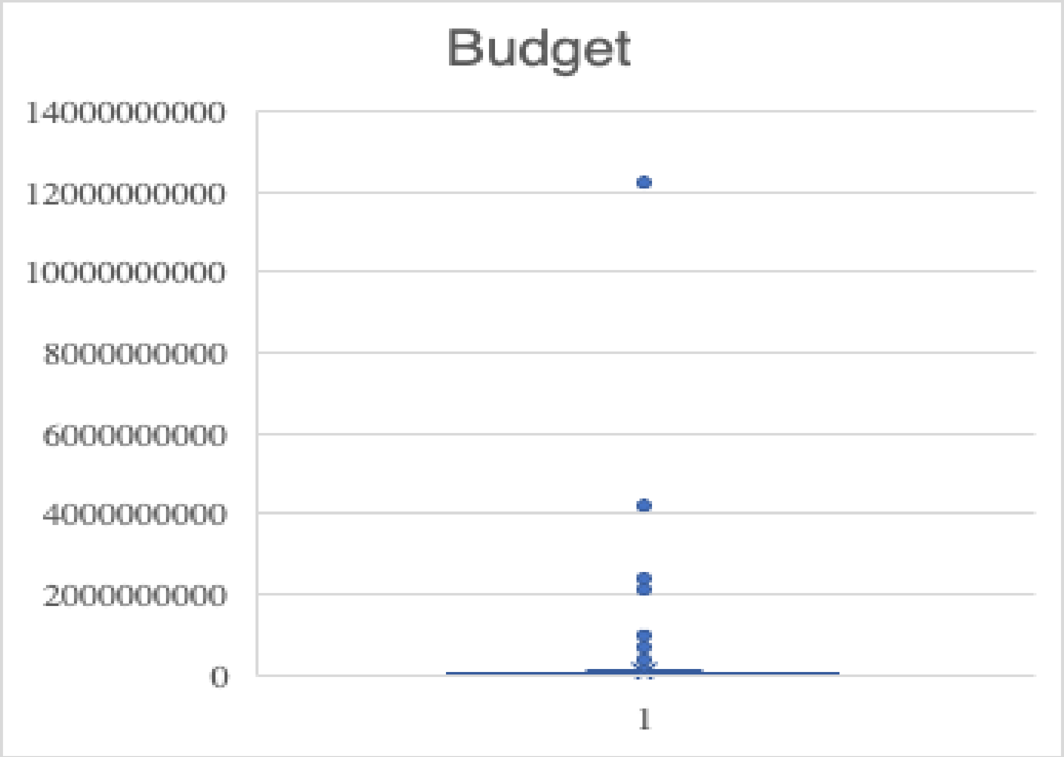


Median value 25440971 is filled in the null value cells

Outlier values have not been replaced as extremes in revenue are possible

The data has outliers, so median values is chosen instead of other measures of central tendency

Data Imputation – Budget



Median value 20000000 is put in the null value cells

Outlier values have not been replaced as extremes in revenue are possible

The data has outliers, so median values is chosen instead of other measures of central tendency

Data Imputation – Director_name

- 103 records were deleted due to null value in director_name column
- Imputation of director’s name is not an ideal choice

INSIGHTS

Genre Analysis

1. **Most Commonly Produced Genres**: Drama, Comedy, and Thriller are the most frequentlyproduced genres in the film industry.

2. **Highest IMDb Score**: The highest IMDb score achieved by any movie belongs to theComedy genre.

3. **Average IMDb Ratings**: Niche categories like Film Noir and News tend to receive thehighest average IMDb ratings despite being less commonly produced.

4. **Standard Deviation**: These niche categories exhibit a relatively low standard deviation in their IMDb ratings. This low variability indicates a more consistent level of quality among these films, which is considered a positive indicator.

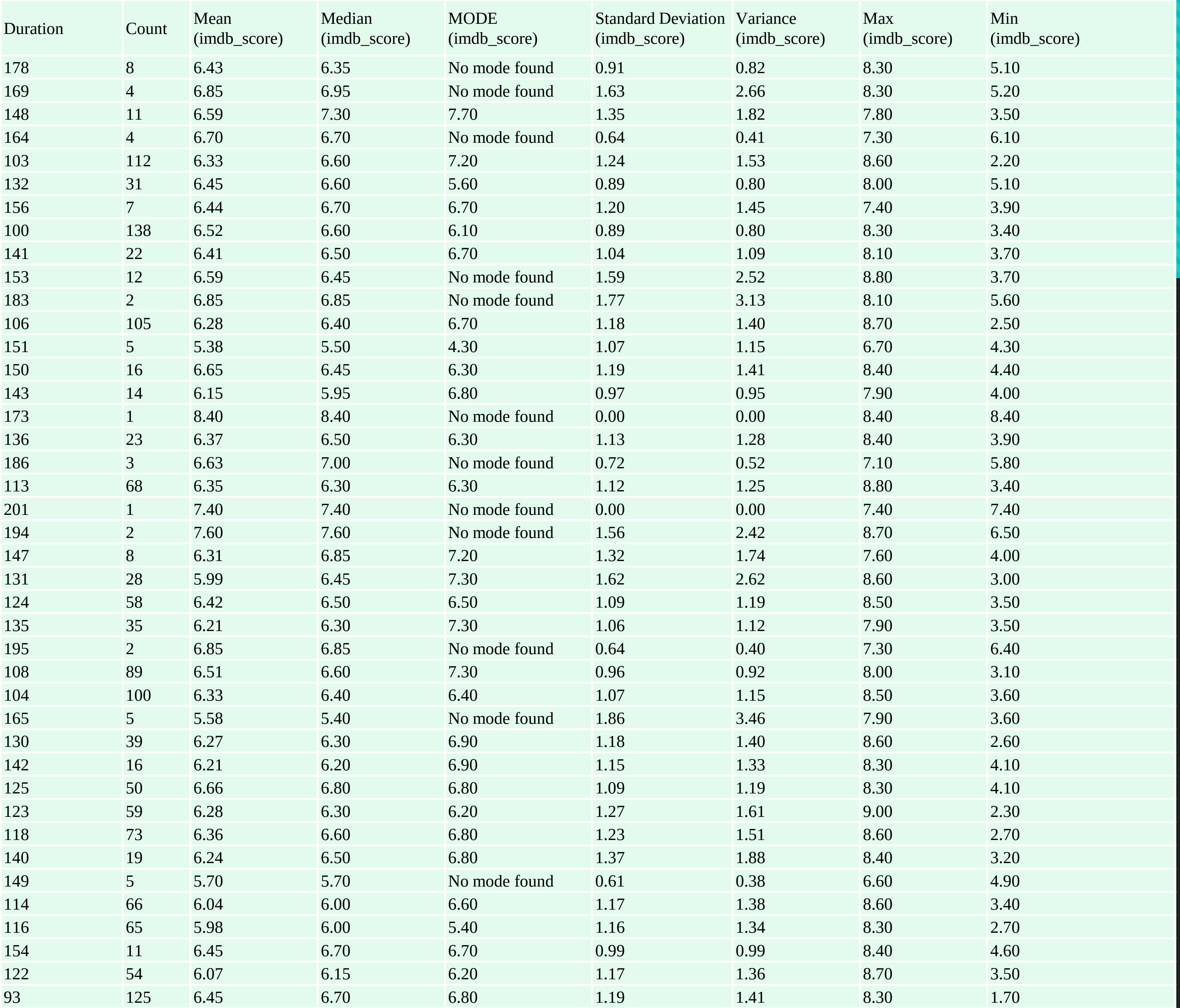
Genre	Count	Mean (imdb_score)	Median (imdb_score)	MODE (imdb_score)	Standard Deviation (imdb_score)	Variance (imdb_score)	Max (imdb_score)	Min (imdb_score)
Action	1143	6.24	6.30	6.10	1.11	1.24	9.10	1.70
Adventure	914	6.44	6.60	6.70	1.13	1.28	8.90	1.90
Animation	242	6.58	6.70	6.70	1.14	1.30	8.60	1.70
Biography	292	7.15	7.20	7.00	0.72	0.52	8.90	4.50
Comedy	1862	6.19	6.30	6.70	1.09	1.19	9.50	1.70
Crime	883	6.56	6.60	6.60	1.03	1.05	9.30	2.40
Documentary	121	7.18	7.40	7.50	1.06	1.12	8.70	1.60
Drama	2571	6.76	6.90	7.20	0.95	0.91	9.30	2.00
Family	544	6.25	6.40	6.70	1.20	1.45	8.70	1.70
Fantasy	604	6.31	6.40	6.70	1.16	1.34	8.90	1.70
Film-Noir	6	7.63	7.65	No mode found	0.43	0.19	8.20	7.10
Game-Show	1	2.90	2.90	No mode found	0.00	0.00	2.90	2.90
History	205	7.08	7.20	7.50	0.89	0.79	8.90	2.00
Horror	556	5.83	5.90	6.20	1.13	1.27	8.70	2.20
Music	324	6.46	6.70	7.10	1.20	1.44	8.50	1.60
Musical	132	6.51	6.70	7.00	1.23	1.50	8.50	2.10
Mystery	493	6.49	6.60	6.60	1.08	1.17	8.60	2.20
News	3	7.53	7.40	No mode found	0.51	0.26	8.10	7.10
Reality-TV	2	4.75	4.75	No mode found	2.62	6.85	6.60	2.90
Romance	1098	6.45	6.50	6.50	1.00	1.00	8.60	2.10
Sci-Fi	611	6.28	6.40	6.70	1.21	1.46	8.80	1.90
Short	5	6.38	6.50	No mode found	0.75	0.56	7.10	5.20
Sport	181	6.60	6.80	7.20	1.10	1.22	8.70	2.00
Thriller	1396	6.31	6.40	6.10	1.05	1.11	9.00	2.20
War	211	7.07	7.10	7.10	0.88	0.77	8.60	2.70
Western	94	6.70	6.80	6.50	1.06	1.11	8.90	3.80

Significance of Genre Analysis:

- **Audience preferences**: Analyzing popular genres can reveal what types of stories or themes resonate with viewers. For instance, if action movies consistently rank high, it indicates a demand for adrenaline-packed narratives.
- **Trends and cycles**: Understanding genre popularity over time can highlight industry shifts and cyclical trends. Certain genres may rise and fall in popularity depending on cultural factors or technological advancements.
- **Market segmentation**: Studios use genre analysis to target specific audience demographics. For example, romantic comedies often cater to a different audience segment than horror films.

Duration Analysis

- **Most Common Duration:** Movies with a duration around 100 minutes are the most commonly produced.
- **Range of Movie Durations:** The majority of movies fall within the range of 85 to 125 minutes in duration.
- **IMDb Ratings:** Movies within this duration range tend to receive ratings near 6 on average.

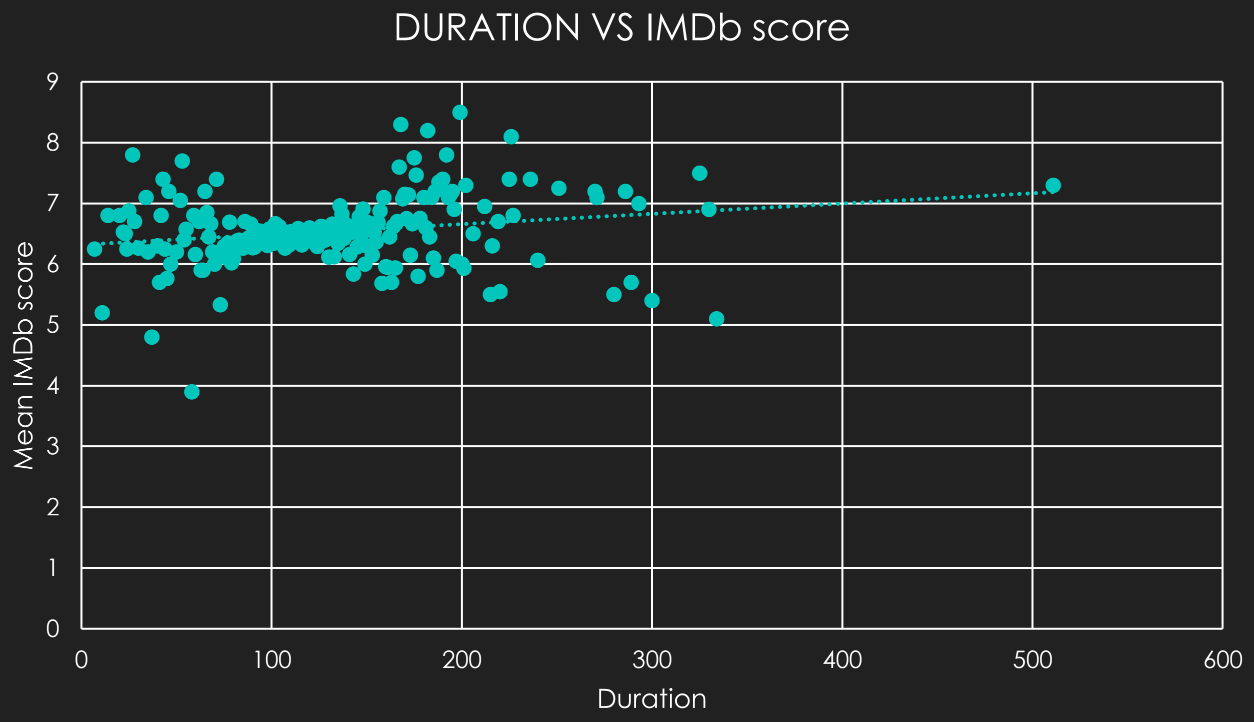


Duration	Count	Mean (imdb_score)	Median (imdb_score)	MODE (imdb_score)	Standard Deviation (imdb_score)	Variance (imdb_score)	Max (imdb_score)	Min (imdb_score)
178	8	6.43	6.35	No mode found	0.91	0.82	8.30	5.10
169	4	6.85	6.95	No mode found	1.63	2.66	8.30	5.20
148	11	6.59	7.30	7.70	1.35	1.82	7.80	3.50
164	4	6.70	6.70	No mode found	0.64	0.41	7.30	6.10
103	112	6.33	6.60	7.20	1.24	1.53	8.60	2.20
132	31	6.45	6.60	5.60	0.89	0.80	8.00	5.10
156	7	6.44	6.70	6.70	1.20	1.45	7.40	3.90
100	138	6.52	6.60	6.10	0.89	0.80	8.30	3.40
141	22	6.41	6.50	6.70	1.04	1.09	8.10	3.70
153	12	6.59	6.45	No mode found	1.59	2.52	8.80	3.70
183	2	6.85	6.85	No mode found	1.77	3.13	8.10	5.60
106	105	6.28	6.40	6.70	1.18	1.40	8.70	2.50
151	5	5.38	5.50	4.30	1.07	1.15	6.70	4.30
150	16	6.65	6.45	6.30	1.19	1.41	8.40	4.40
143	14	6.15	5.95	6.80	0.97	0.95	7.90	4.00
173	1	8.40	8.40	No mode found	0.00	0.00	8.40	8.40
136	23	6.37	6.50	6.30	1.13	1.28	8.40	3.90
186	3	6.63	7.00	No mode found	0.72	0.52	7.10	5.80
113	68	6.35	6.30	6.30	1.12	1.25	8.80	3.40
201	1	7.40	7.40	No mode found	0.00	0.00	7.40	7.40
194	2	7.60	7.60	No mode found	1.56	2.42	8.70	6.50
147	8	6.31	6.85	7.20	1.32	1.74	7.60	4.00
131	28	5.99	6.45	7.30	1.62	2.62	8.60	3.00
124	58	6.42	6.50	6.50	1.09	1.19	8.50	3.50
135	35	6.21	6.30	7.30	1.06	1.12	7.90	3.50
195	2	6.85	6.85	No mode found	0.64	0.40	7.30	6.40
108	89	6.51	6.60	7.30	0.96	0.92	8.00	3.10
104	100	6.33	6.40	6.40	1.07	1.15	8.50	3.60
165	5	5.58	5.40	No mode found	1.86	3.46	7.90	3.60
130	39	6.27	6.30	6.90	1.18	1.40	8.60	2.60
142	16	6.21	6.20	6.90	1.15	1.33	8.30	4.10
125	50	6.66	6.80	6.80	1.09	1.19	8.30	4.10
123	59	6.28	6.30	6.20	1.27	1.61	9.00	2.30
118	73	6.36	6.60	6.80	1.23	1.51	8.60	2.70
140	19	6.24	6.50	6.80	1.37	1.88	8.40	3.20
149	5	5.70	5.70	No mode found	0.61	0.38	6.60	4.90
114	66	6.04	6.00	6.60	1.17	1.38	8.60	3.40
116	65	5.98	6.00	5.40	1.16	1.34	8.30	2.70
154	11	6.45	6.70	6.70	0.99	0.99	8.40	4.60
122	54	6.07	6.15	6.20	1.17	1.36	8.70	3.50
93	125	6.45	6.70	6.80	1.19	1.41	8.30	1.70

Duration Analysis

- **Scatter Plot:**

There seems to be a upward trend between the duration of a movie and its average IMDb ratings and the data points are widely spread.



Significance of Duration Analysis:

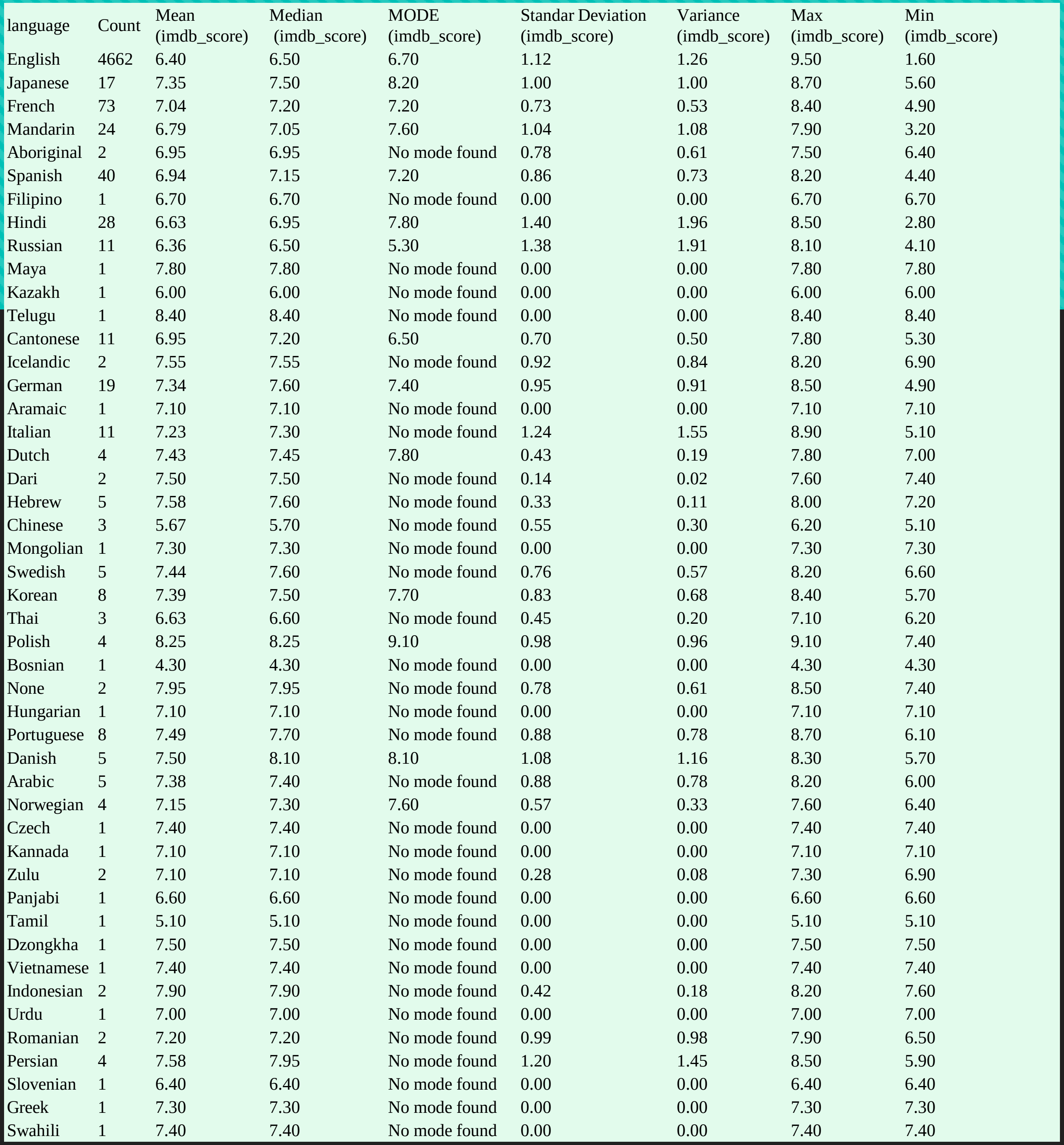
- **Audience attention span**: Patterns in movie duration and IMDb ratings might suggest an optimal length that keeps audiences engaged without overwhelming or underwhelming them.
- **Genre-specific preferences**: Certain genres might lend themselves to different durations.Action movies might have shorter optimal lengths compared to epic dramas.
- **Directorial style**: Certain directors might have a signature duration for their films, which could influence audience expectations.

Language Analysis

1. Commonly Produced Language: English is the most commonly used language for movie production.

2. Average IMDb Ratings: Despite being the most produced language, movies in English do not consistently achieve higher IMDb ratings on average.

3. Highest IMDb Ratings: The highest IMDb ratings are attained by movies in languages such as Telugu and Polish, surpassing the ratings of English language movies.



language	Count	Mean (imdb_score)	Median (imdb_score)	MODE (imdb_score)	Standar Deviation (imdb_score)	Variance (imdb_score)	Max (imdb_score)	Min (imdb_score)
English	4662	6.40	6.50	6.70	1.12	1.26	9.50	1.60
Japanese	17	7.35	7.50	8.20	1.00	1.00	8.70	5.60
French	73	7.04	7.20	7.20	0.73	0.53	8.40	4.90
Mandarin	24	6.79	7.05	7.60	1.04	1.08	7.90	3.20
Aboriginal	2	6.95	6.95	No mode found	0.78	0.61	7.50	6.40
Spanish	40	6.94	7.15	7.20	0.86	0.73	8.20	4.40
Filipino	1	6.70	6.70	No mode found	0.00	0.00	6.70	6.70
Hindi	28	6.63	6.95	7.80	1.40	1.96	8.50	2.80
Russian	11	6.36	6.50	5.30	1.38	1.91	8.10	4.10
Maya	1	7.80	7.80	No mode found	0.00	0.00	7.80	7.80
Kazakh	1	6.00	6.00	No mode found	0.00	0.00	6.00	6.00
Telugu	1	8.40	8.40	No mode found	0.00	0.00	8.40	8.40
Cantonese	11	6.95	7.20	6.50	0.70	0.50	7.80	5.30
Icelandic	2	7.55	7.55	No mode found	0.92	0.84	8.20	6.90
German	19	7.34	7.60	7.40	0.95	0.91	8.50	4.90
Aramaic	1	7.10	7.10	No mode found	0.00	0.00	7.10	7.10
Italian	11	7.23	7.30	No mode found	1.24	1.55	8.90	5.10
Dutch	4	7.43	7.45	7.80	0.43	0.19	7.80	7.00
Dari	2	7.50	7.50	No mode found	0.14	0.02	7.60	7.40
Hebrew	5	7.58	7.60	No mode found	0.33	0.11	8.00	7.20
Chinese	3	5.67	5.70	No mode found	0.55	0.30	6.20	5.10
Mongolian	1	7.30	7.30	No mode found	0.00	0.00	7.30	7.30
Swedish	5	7.44	7.60	No mode found	0.76	0.57	8.20	6.60
Korean	8	7.39	7.50	7.70	0.83	0.68	8.40	5.70
Thai	3	6.63	6.60	No mode found	0.45	0.20	7.10	6.20
Polish	4	8.25	8.25	9.10	0.98	0.96	9.10	7.40
Bosnian	1	4.30	4.30	No mode found	0.00	0.00	4.30	4.30
None	2	7.95	7.95	No mode found	0.78	0.61	8.50	7.40
Hungarian	1	7.10	7.10	No mode found	0.00	0.00	7.10	7.10
Portuguese	8	7.49	7.70	No mode found	0.88	0.78	8.70	6.10
Danish	5	7.50	8.10	8.10	1.08	1.16	8.30	5.70
Arabic	5	7.38	7.40	No mode found	0.88	0.78	8.20	6.00
Norwegian	4	7.15	7.30	7.60	0.57	0.33	7.60	6.40
Czech	1	7.40	7.40	No mode found	0.00	0.00	7.40	7.40
Kannada	1	7.10	7.10	No mode found	0.00	0.00	7.10	7.10
Zulu	2	7.10	7.10	No mode found	0.28	0.08	7.30	6.90
Panjabi	1	6.60	6.60	No mode found	0.00	0.00	6.60	6.60
Tamil	1	5.10	5.10	No mode found	0.00	0.00	5.10	5.10
Dzongkha	1	7.50	7.50	No mode found	0.00	0.00	7.50	7.50
Vietnamese	1	7.40	7.40	No mode found	0.00	0.00	7.40	7.40
Indonesian	2	7.90	7.90	No mode found	0.42	0.18	8.20	7.60
Urdu	1	7.00	7.00	No mode found	0.00	0.00	7.00	7.00
Romanian	2	7.20	7.20	No mode found	0.99	0.98	7.90	6.50
Persian	4	7.58	7.95	No mode found	1.20	1.45	8.50	5.90
Slovenian	1	6.40	6.40	No mode found	0.00	0.00	6.40	6.40
Greek	1	7.30	7.30	No mode found	0.00	0.00	7.30	7.30
Swahili	1	7.40	7.40	No mode found	0.00	0.00	7.40	7.40

Significance of Language Analysis:

- Global appeal: Analyzing the predominant languages in highly-rated movies can provide insights into their international success. Movies in widely spoken languages might have broader global audiences.
- Cultural influences: The language used in a film can indicate its cultural context and intended audience. Subtle nuances or linguistic choices might impact how a movie is received in different regions.

Director Analysis

99th Percentile is 8.3

These directors represent the top 1% in terms of IMDb scores, showcasing their exceptional achievement in consistently receiving high ratings for their work in the film industry.

Top 1% directors based on 99th percentile

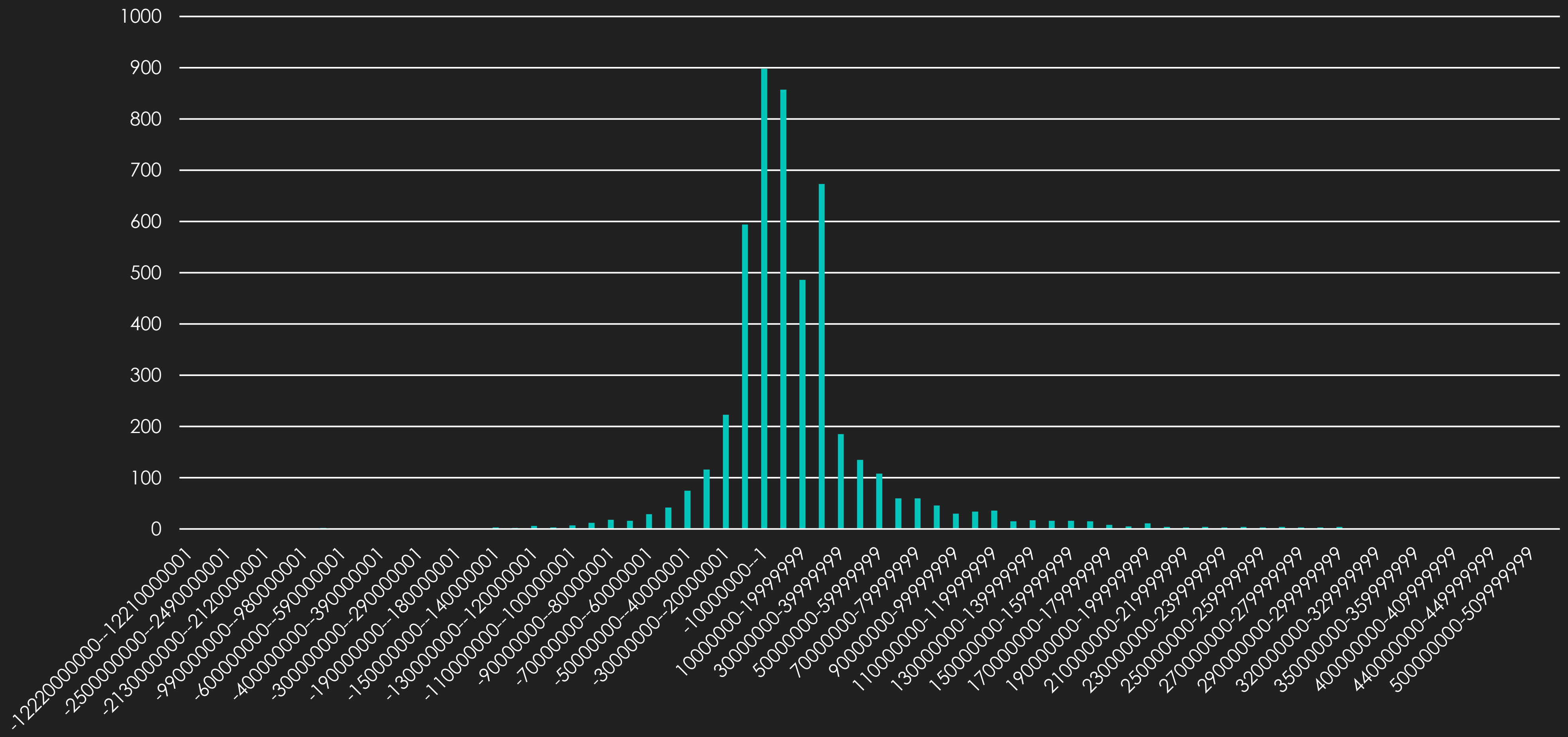
Top 1% directors	Mean (imdb_score)
Christopher Nolan	8.4
S.S. Rajamouli	8.4
Moustapha Akkad	8.4
Richard Marquand	8.4
Sergio Leone	8.5
Catherine Owens	8.4
John Blanchard	9.5
Rakeysh Omprakash Mehra	8.4
Mike Mayhall	8.6
Raja Menon	8.5
Ron Fricke	8.5
Jay Oliva	8.4
Damien Chazelle	8.5
Robert Mulligan	8.4
Mitchell Altieri	8.7
Charles Chaplin	8.6
Sadyk Sher-Niyaz	8.7
Asghar Farhadi	8.4
Marius A. Markevicius	8.4
Majid Majidi	8.5
Cary Bell	8.7
Bill Melendez	8.4

Significance of Director Analysis:

- **Auteur theory**: Examining the works of specific directors can reveal recurring themes, stylistic choices, or storytelling techniques unique to their vision.
- **Consistency in quality**: Directors with consistently high-rated movies may have a strong influence on a film's success, indicating a loyal fan base or a distinct directorial style.
- **Evolution of style**: Tracking a director's filmography over time can show how their style, storytelling, or choice of genres has evolved, impacting audience reception.

Budget Analysis

Distribution of movies by Profit Margin



The data set is normally distributed.

A nearly equal number of movies have made a profit and an equivalent number have incurred losses.

Budget Analysis

Top performing movies (Profit Margin)	Genre
Avatar	Action|Adventure|Fantasy|Sci-Fi
Jurassic World	Action|Adventure|Sci-Fi|Thriller
Titanic	Drama|Romance
Star Wars: Episode IV - A New Hope	Action|Adventure|Fantasy|Sci-Fi
E.T. the Extra-Terrestrial	Family|Sci-Fi
The Avengers	Action|Adventure|Sci-Fi
The Lion King	Adventure|Animation|Drama|Family|Musical
Star Wars: Episode I - The Phantom Menace	Action|Adventure|Fantasy|Sci-Fi
The Dark Knight	Action|Crime|Drama|Thriller
The Hunger Games	Adventure|Drama|Sci-Fi|Thriller

Most of the top 10 highest-profit movies belong to the action and adventure genres, which also represent the genres with the highest production count.

Budget Analysis

Lowest performing movies	Genre
The Host	Action|Adventure|Romance|Sci-Fi|Thriller
Lady Vengeance	Crime|Drama
Fateless	Drama|Romance|War
Princess Mononoke	Adventure|Animation|Fantasy
Steamboy	Action|Adventure|Animation|Family|Sci-Fi|Thriller
Akira	Action|Animation|Sci-Fi
Godzilla 2000	Action|Adventure|Drama|Sci-Fi|Thriller
Tango	Drama|Musical
Kabhi Alvida Naa Kehna	Drama
Kites	Action|Drama|Romance|Thriller

No specific genre trend can be identified among the movies, as many of them also belong to the action and adventure genres.

Significance of Budget Analysis:

- **Return on Investment (ROI)**: Comparing budgets against IMDB ratings can reveal if higher budgets consistently translate to higher ratings or if smaller-budget films can compete in terms of quality.
- **Production value**: Budget analysis can highlight how effectively resources are utilized. It can show if a high budget correlates with better visual effects, star power, or overall production quality.
- **Industry benchmarks**: Understanding budget trends can help studios benchmark their investments against successful movies in similar genres or with similar ratings.

Excel Link

https://docs.google.com/spreadsheets/d/1Pn74y-pIPuyTGPwyWftH45fIlZtfvPCL/edit?usp=sharing&ouid=101949921485202693908&rtpof=true&sd=true

Video link

https://www.loom.com/share/6041f197c2214bf68ee8f0384da52998?sid=49b3889f-315b-4927-b96d-844cdb81b5ac