IMDB Movie Analysis

Imdb_movie_analysis

Sheet1

color,director_name,num_critic_for_reviews,duration,duration_intervals,director_facebook_likes,actor_2_name,gross,genres,g2,g3,g4,g

https://docs.google.com/spreadsheets/d/1O4o2c1ecv2xWnP6lFVSH8ggYQ_5yIFo0S79XXOmaOWg/edit?usp=sharing

Project Description:

This data analysis project focuses on discovering the key factors that influence the success of movies on IMDb, measured by the IMDb ratings, box office collections etc. For stakeholders like producers, directors, and investors in the film industry, being aware of these metrics is crucial for making data-driven decisions in their future projects. By exploring patterns and trends within the data, the actionable insights that can shape the strategies behind successful films.

Approach:

step 1: Understanding the dataset and choosing the right tool for the analysis.

step 2: Prepare the data for analysis. Check for data quality, completeness and accuracy.

step 2: Identifying the business tasks and check if the data you have is sufficient or not.

step 3: Modify and rearrange the dataset, and add new columns if needed to suit your analysis.

step 4: Using the data, analyze and answer the key questions and solve the business tasks.

step 5: Provide insights to the stakeholders, supported by data visualization if needed.

Data Cleaning:

In the imdb movie dataset, 45 duplicate rows were removed.

There were missing data in some other columns, but the dataset was large enough for the analysis.

Calculated fields for duration intervals and profit margin were added to the dataset.

Tech-Stack Used:

Google Sheets as a spreadsheet tool to format and analyze data.

Notion: It is a versatile tool that can be used for a variety of tasks, including project management, data analysis, and report generation as pdf.

Insights:

Business tasks:

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.

Genre	sum	Number of movies	Mean	median	mode	max	min	var	stdev
Action	7128.4	1143	6.24	6.3	6.1	9.1	1.7	1.24	1.11
Adventure	5887.4	914	6.44	6.6	6.7	8.9	1.9	1.28	1.13
Animation	1591.4	242	6.58	6.7	6.7	8.6	1.7	1.30	1.14
Biography	2087.8	292	7.15	7.2	7	8.9	4.5	0.52	0.72
Comedy	11534.5	1862	6.19	6.3	6.7	9.5	1.7	1.19	1.09
Crime	5793.3	883	6.56	6.6	6.6	9.3	2.4	1.05	1.03
Documentary	868.8	121	7.18	7.4	7.5	8.7	1.6	1.12	1.06
Drama	17392.6	2571	6.76	6.9	7.2	9.3	2	0.91	0.95
Family	3398.6	544	6.25	6.4	6.7	8.7	1.7	1.45	1.20
Fantasy	3811.9	604	6.31	6.4	6.7	8.9	1.7	1.34	1.16
Film-Noir	45.8	6	7.63	7.65	#N/A	8.2	7.1	0.19	0.43
Game-Show	2.9	1	2.90	2.9	#N/A	2.9	2.9	#DIV/0!	#DIV/0!
History	1452.4	205	7.08	7.2	7.5	8.9	2	0.79	0.89
Horror	3243	556	5.83	5.9	6.2	8.7	2.2	1.27	1.13
Music	1358.1	212	6.41	6.6	6.5	8.5	1.6	1.40	1.18
Musical	859	132	6.51	6.7	7	8.5	2.1	1.50	1.23
Mystery	3198.9	493	6.49	6.6	6.6	8.6	2.2	1.17	1.08
News	22.6	3	7.53	7.4	#N/A	8.1	7.1	0.26	0.51
Reality-TV	9.5	2	4.75	4.75	#N/A	6.6	2.9	6.85	2.62
Romance	7079.8	1098	6.45	6.5	6.5	8.6	2.1	1.00	1.00
Sci-Fi	3835.5	611	6.28	6.4	6.7	8.8	1.9	1.46	1.21
Short	31.9	5	6.38	6.5	#N/A	7.1	5.2	0.56	0.75
Sport	1195.1	181	6.60	6.8	7.2	8.7	2	1.22	1.10
Thriller	8812.8	1396	6.31	6.4	6.1	9	2.2	1.11	1.05
War	1492.8	211	7.07	7.1	7.1	8.6	2.7	0.77	0.88
Western	630.1	94	6.70	6.8	6.5	8.9	3.8	1.11	1.06

Comedy, drama and thriller movies were the most popular genres in terms of number of movies.

Considering the imdb ratings, and also considering a minimum number of 200 movies,

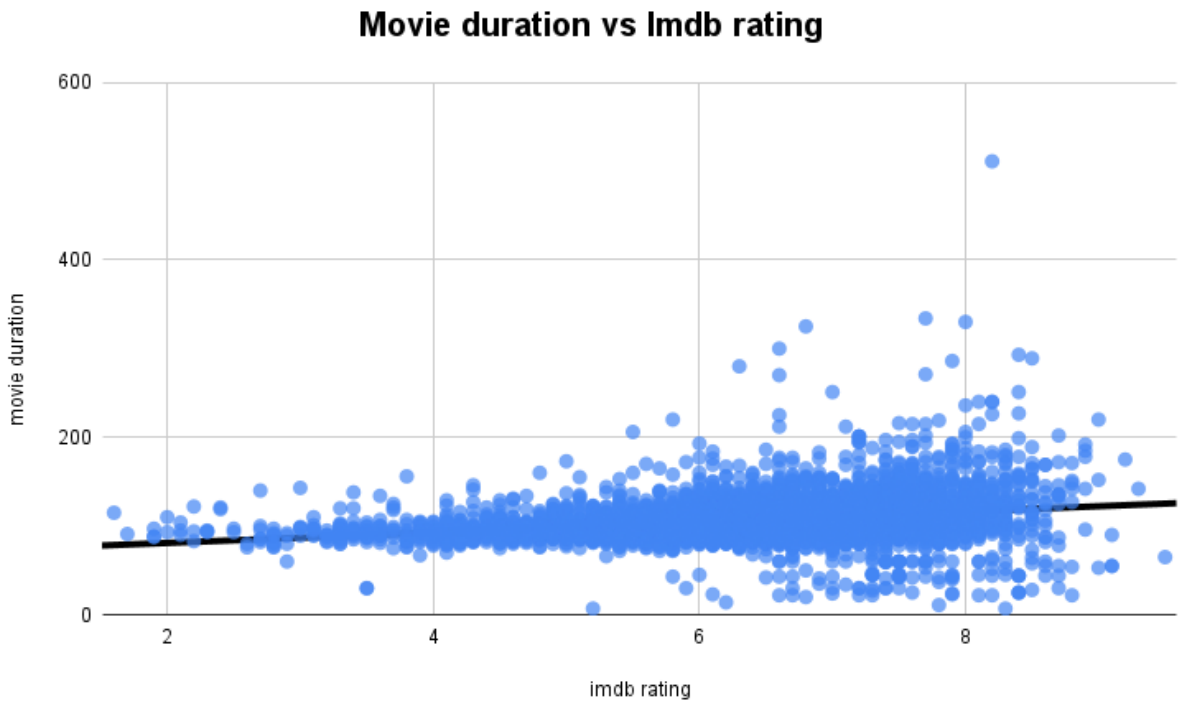
war, history and biography genres have 7+ rating average.

Variance and standard deviation suggests that the values are closely clustered around the mean.

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.

Duration in minutes	Number of films	Average imdb rating	median	stdev
60-120	3762	6.24	6.40	1.10
120-180	1071	7.01	7.10	0.91
20-60	78	7.53	7.50	1.01
180-240	52	7.60	7.70	0.77
240+	16	7.66	7.95	0.74
<20	4	6.88	7.00	1.43



Movies were divided based on the duration to different intervals to suit the analysis.

Most number of the movies were 60 to 120 mins long with an average imdb rating of 6.24.

Standard deviation suggests that the values are closely clustered around the mean.

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.

language	number of films	Mean	Median	Std deviation
English	4662	6.40	6.5	1.12
French	73	7.04	7.2	0.73
Spanish	40	6.94	7.15	0.86
Hindi	28	6.63	6.95	1.40
Mandarin	24	6.79	7.05	1.04
German	19	7.34	7.6	0.95
Japanese	17	7.35	7.5	1.00
Russian	11	6.36	6.5	1.38
Italian	11	7.23	7.3	1.24
Cantonese	11	6.95	7.2	0.70
Portuguese	8	7.49	7.7	0.88
Korean	8	7.39	7.5	0.83
Swedish	5	7.44	7.6	0.76
Hebrew	5	7.58	7.6	0.33
Danish	5	7.50	8.1	1.08
Arabic	5	7.38	7.4	0.88
Polish	4	8.25	8.25	0.98
Persian	4	7.58	7.95	1.20
Norwegian	4	7.15	7.3	0.57
Dutch	4	7.43	7.45	0.43
Thai	3	6.63	6.6	0.45
Chinese	3	5.67	5.7	0.55
Zulu	2	7.10	7.1	0.28
Romanian	2	7.20	7.2	0.99
None	2	7.95	7.95	0.78
Indonesian	2	7.90	7.9	0.42
Icelandic	2	7.55	7.55	0.92
Dari	2	7.50	7.5	0.14
Aboriginal	2	6.95	6.95	0.78

As one might expect, English is the widely popular language in film industry with an average rating of 6.4.

French, Spanish and Hindi makes it to the top 4.

Since there’s a huge gap in the number of movies between English and other languages, we can’t conclude anything about the average rating.

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.

director_name	Number of movies	Average imdb rating	Mode	Std deviation
Christopher Nolan	8	8.43	8.5	0.54
Quentin Tarantino	8	8.20	8.2	0.42
Stanley Kubrick	7	8.00	8.1	0.44
David Fincher	10	7.75	7.8	0.72
Peter Jackson	12	7.68	7.45	0.77
Martin Scorsese	20	7.66	7.5	0.60
Wes Anderson	7	7.63	7.7	0.34
Sam Mendes	8	7.50	7.5	0.52
Steven Spielberg	26	7.48	7.6	0.74
Danny Boyle	8	7.44	7.45	0.52
Francis Ford Coppola	11	7.42	7.5	1.27
Alfred Hitchcock	8	7.35	7.35	0.75
Terry Gilliam	7	7.33	7.6	0.82
Richard Linklater	11	7.33	7.1	0.67
Edward Zwick	8	7.33	7.6	0.68
Robert Zemeckis	13	7.31	7.4	0.77
Bryan Singer	8	7.29	7.35	0.82
Ang Lee	8	7.25	7.6	0.79
Clint Eastwood	20	7.23	7.3	0.70
Zack Snyder	8	7.18	7.3	0.52
Marc Forster	8	7.15	7.05	0.46
Lasse Hallström	9	7.11	7	0.57
Kenneth Branagh	8	7.09	7	0.54
James Mangold	8	7.08	7.1	0.60
Ridley Scott	17	7.07	7	0.96
Stephen Frears	10	7.07	7.2	0.63
Rob Reiner	11	7.02	7.4	0.94
Woody Allen	22	7.01	6.95	0.53
Oliver Stone	14	6.95	7.15	0.74
Antoine Fuqua	8	6.94	6.95	0.55
F. Gary Gray	8	6.94	7.15	0.72
Tim Burton	16	6.93	7	0.75
Ron Howard	13	6.93	6.7	0.92

As expected, Christopher Nolan movies has the highest imdb rating average of 8.43.

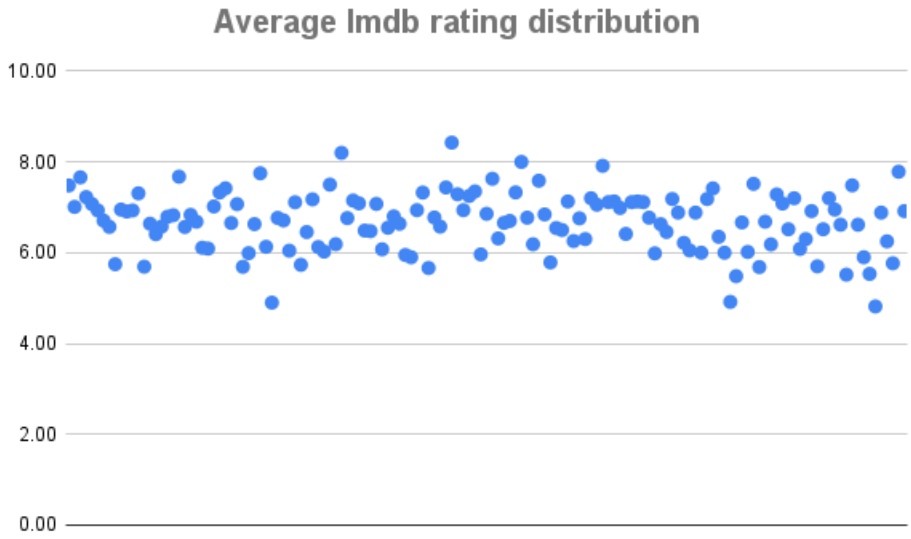
Quentin Tarantino and Stanley Kubrick also has 8+ average imdb rating.

Steven Spielberg has done 26 movies and still maintains 7.48 average rating which is appreciable.

90 percentile	75 percentile	50 percentile	25 percentile
7.73	7.2	6.6	5.8

Above are the percentile calculations of the average imdb rating of Directors.

It provides a metric for us to determine the best directors.



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.

movie_imdb_link	Profit margin	Movie
http://www.imdb.com/title/tt0499549/?ref_=fn_tt_tt_1	523505847	Avatar (2009)
http://www.imdb.com/title/tt0369610/?ref_=fn_tt_tt_1	502177271	Jurassic World (2015)
http://www.imdb.com/title/tt0120338/?ref_=fn_tt_tt_1	458672302	Titanic (1997)
http://www.imdb.com/title/tt0076759/?ref_=fn_tt_tt_1	449935665	Star Wars: Episode IV - A New Hope (1977)
http://www.imdb.com/title/tt0083866/?ref_=fn_tt_tt_1	424449459	E.T. the Extra-Terrestrial (1982)
http://www.imdb.com/title/tt0848228/?ref_=fn_tt_tt_1	403279547	The Avengers (2012)
http://www.imdb.com/title/tt0110357/?ref_=fn_tt_tt_1	377783777	The Lion King (1994)
http://www.imdb.com/title/tt0120915/?ref_=fn_tt_tt_1	359544677	#N/A
http://www.imdb.com/title/tt0468569/?ref_=fn_tt_tt_1	348316061	The Dark Knight (2008)
http://www.imdb.com/title/tt1392170/?ref_=fn_tt_tt_1	329999255	The Hunger Games (2012)
http://www.imdb.com/title/tt1431045/?ref_=fn_tt_tt_1	305024263	Deadpool (2016)
http://www.imdb.com/title/tt1951264/?ref_=fn_tt_tt_1	294645577	The Hunger Games: Catching Fire (2013)
http://www.imdb.com/title/tt0107290/?ref_=fn_tt_tt_1	293784000	Jurassic Park (1993)
http://www.imdb.com/title/tt1690953/?ref_=fn_tt_tt_1	292049635	Despicable Me 2 (2013)
http://www.imdb.com/title/tt2179136/?ref_=fn_tt_tt_1	291323553	American Sniper (2014)
http://www.imdb.com/title/tt0266543/?ref_=fn_tt_tt_1	286838870	Finding Nemo (2003)
http://www.imdb.com/title/tt0298148/?ref_=fn_tt_tt_1	286471036	Shrek 2 (2004)
http://www.imdb.com/title/tt0167260/?ref_=fn_tt_tt_1	283019252	The Lord of the Rings: The Return of the King (2003)
http://www.imdb.com/title/tt0086190/?ref_=fn_tt_tt_1	276625409	Star Wars: Episode VI - Return of the Jedi (1983)

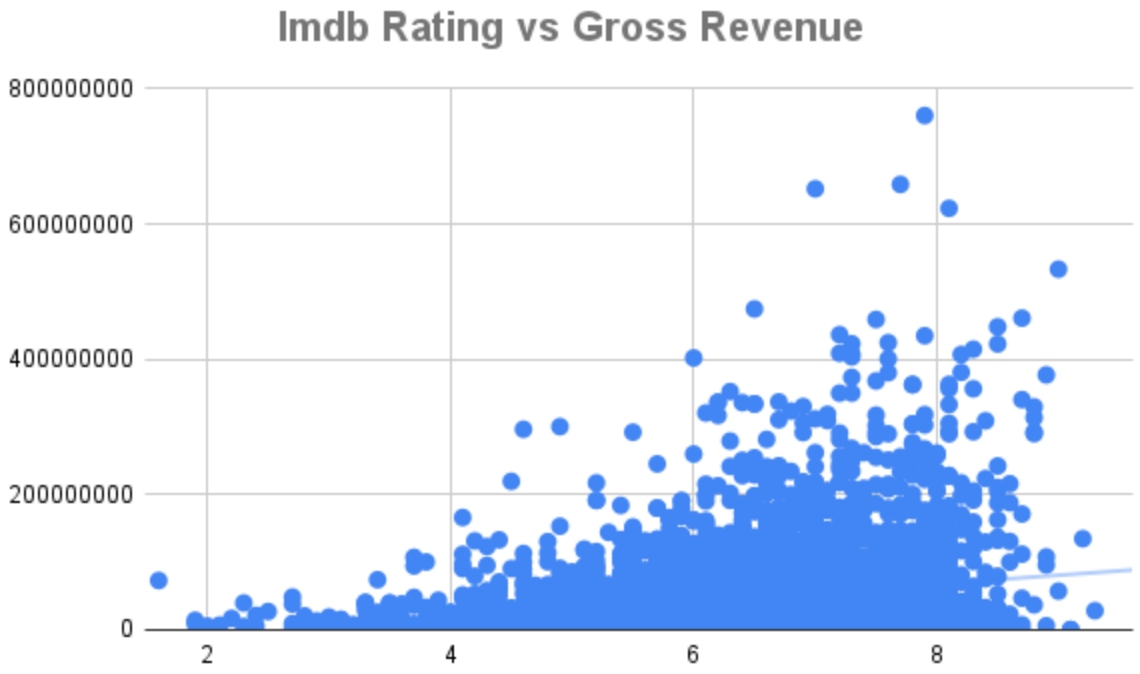
Profit margin is calculated as the difference between budget and the gross revenue.

Above list shows movies with the highest gross profit.The movie Avatar ranks 1 in the list.

Correlation coeff
0.1010334778

A correlation coefficient of 0.1010334778 suggests a weak positive linear relationship between budget and gross revenue.The positive sign indicates that when budget increases, gross revenue tends to increase slightly, but the relationship is not strong.

- **Comparing imdb ratings to the gross revenue to see if highly rated movies are always successful in terms of revenue.**



The correlation coefficient in this case is 0.19 which doesn't show a strong relation between the variables.

Result:

- Comedy, drama, and thriller are the most popular genres.
- War, history, and biography genres achieved an average IMDb rating above 7
- Most movie makers prefer film duration of 60-120mins.

- English language is the widely popular one in film industry.
- Christopher Nolan has the highest IMDb rating average, followed by Quentin Tarantino and Stanley Kubrick.
- A weak positive correlation coefficient indicates a slight increase in revenue with a higher budget.
- Avatar(2009) made the highest gross profit.

Based on the analysis of imdb movie data, several key findings that provide insights into the film industry. These findings collectively reveal interesting trends in various aspects, including genre preferences, the significance of movie duration, comfort of language, and notable influences from directors.

These findings provide valuable information for stakeholders in the film industry, enabling them to make data-driven decisions for their future projects.