

IMDB Movie Analysis

Project – 5

Charmy Raj

Project Description

As a professional data analyst, I was given a large dataset comprising specific information about past all the details and performance of IMDB movies. To help the filmmakers, directors, and financiers who want to know what makes a movie successful so they may make wise choices for their upcoming ventures, it was my critical responsibility to investigate this abundance of data and draw forth key insights.

Tech-Stack

This analytical endeavour is carried out with the expert use of Microsoft Excel, utilising its comprehensive array of statistical functions such as COUNTIF, Pivot Charts, and other statistical tools. My knowledge of decoding this data will aid in the identification of actionable patterns and trends, ultimately guiding the organisation to more informed and effective decision-making processes.

Excel file



<https://docs.google.com/spreadsheets/d/1mBW5oZ0MeDr4a1QkwTciBIP8vNUUjqxE/edit?usp=sharing&ouid=115236894826816815181&rtpof=true&sd=true>

Table of Contents

Project Description.....	1
Tech-Stack	1
Excel file	1
Removed Duplicate Values.....	3
A. Movie Genre Analysis:	4
STATMENT:	4
APPROACH:	4
INTERPRETATION:.....	5
B. Movie Duration Analysis:	6
STATMENT:	6
APPROACH:	6
INTERPRETATION:.....	6
C. language Analysis:.....	7
STATMENT:	7
APPROACH:	7
INTERPRETATION:.....	8
D. Director Analysis:	8
STATMENT:	8
APPROACH:	8
INTERPRETATION:.....	10
E. Budget Analysis:	10
STATMENT:	10
APPROACH:	10
INTERPRETATION:.....	11
Data Story Report.....	11
Initial Problem	11
Finding.....	11
Insights	12

Removed Duplicate Values

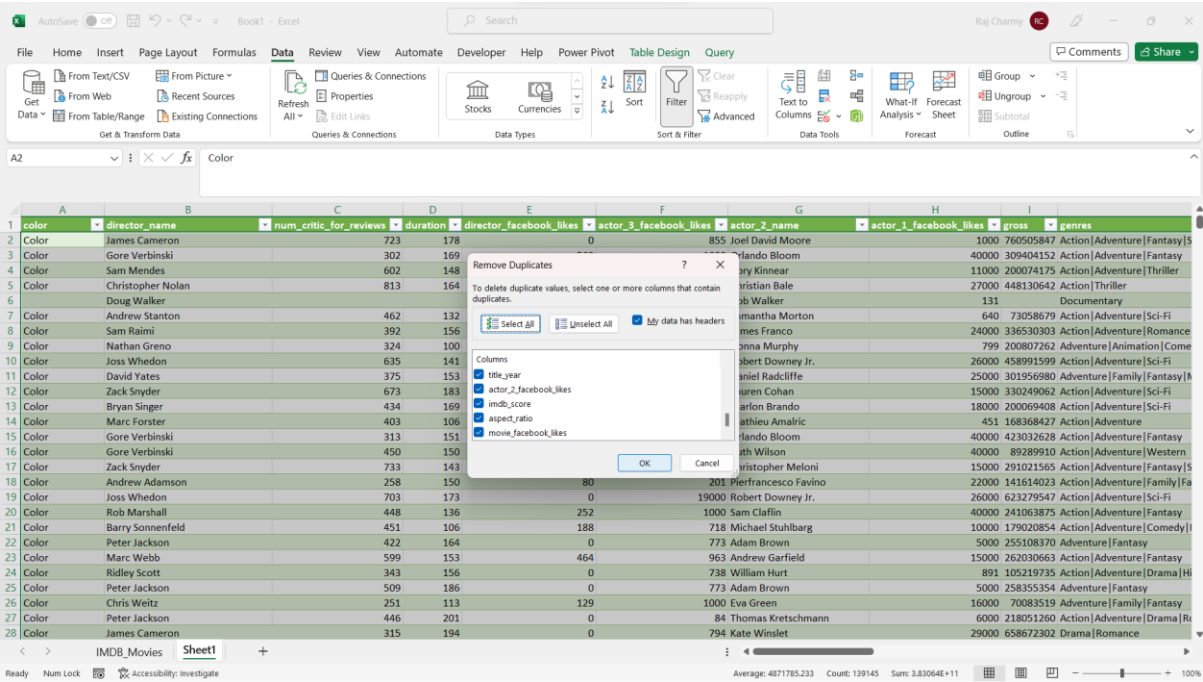


FIG 1 Cleaning of Data

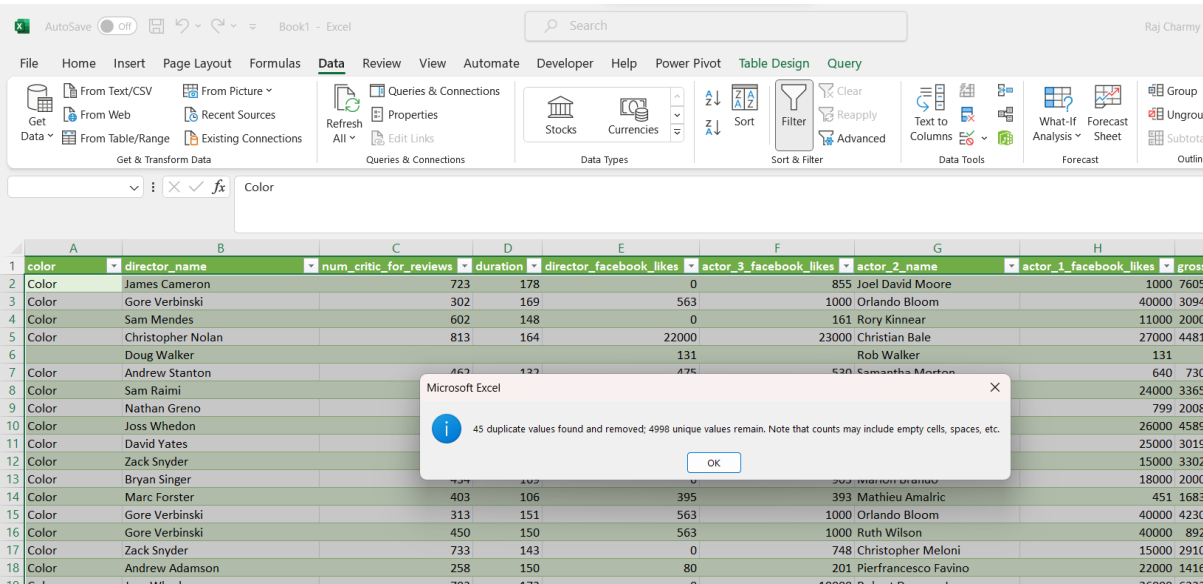


FIG 2 Cleaning of Data

A. Movie Genre Analysis:

STATEMENT:

Analyse how different movie genres are distributed and how it affects the IMDB score. Find the dataset's most common movie genres. Next, determine the descriptive statistics (mean, median, mode, range, variance, and standard deviation) for each genre based on the IMDB ratings.

APPROACH:

To do the movie analysis, following functions were used:

- Countif function: to calculate number of movies for each genre.
- Max function: to calculate maximum number of movies for a particular genre.
- Min function: to calculate minimum number of movies for a particular genre.
- Average function: to calculate mean of the total number of movies.
- Median Function: to calculate middle value of the total number of movies.
- Mode function: to calculate repetition of number of movies for a particular genre.
- VAR function: to calculate how much the values in the dataset deviate from the mean.
- Stdev function: to calculate Standard deviation of the dataset.

	Total Genres	Total Number of movies										
	Film-Noir	1										
	Documentary	45										
	Western	57										
	Musical	96										
	History	147										
	Sport	147										
mily Fantasy Musical Romance	War	150										
	Animation	196										
y	Music	231										
	Biography	238										
	Mystery	378										
	Horror	386										
	Family	440										
	Sci-Fi	492										
	Fantasy	504										
	Crime	704										
	Adventure	773										
	Romance	851										
Fantasy Sci-Fi	Action	951										
	Thriller	1105										
	Comedy	1455										
	Drama	1876										

FIG 3 Movie Genres Calculation

INTERPRETATION:

- The dataset contains a total of 22 distinct movie genres.
- 11,223 total films have been produced. The overall number of movies in the dataset, across all genres, is 11,223 according to the "Total Number of Movies" column's sum.

Let's now compute some statistical conclusions using the supplied data:

- Mean: 510.1364 The average number of films across all genres is shown by the mean. For each genre, there are typically 510.1364 films.
- Average: 382 When the quantity of films in each genre is listed in ascending order, the median is the midway value. According to the median of 382, approximately half of the genres have fewer than 382 films, and the other half have more than 382 films.
- Mode: 147 The value that appears the most frequently in the dataset is the mode. In this instance, 147 refers to the quantity of films in the most prevalent genres—"History" and "Sport"—together.
- Difference: 241147.7 The variance quantifies the degree to which the dataset's values depart from the mean. The large variance value shows that there are more films in each genre than the mean, suggesting distributional diversity.
- 491.068 Standard Deviation: The average number of films in each genre deviates from the mean of 510.1364 by around 491 films, according to the standard deviation of 491.068.

CONCLUSION:

- The dataset includes a wide range of film genres, from "Film-Noir" with just 1 film to "Drama" with the highest number of films (1876).
- The properties of the dataset, such as the relatively high mean and standard deviation, show that there is a lack of balance in the distribution of movies among genres.
- As an alternative, there are both genres with few films and those with numerous.
- The fact that there is a mode at 147 indicates that there are frequently this many movies (representing the "History" and "Sport" categories).
- The high variance and standard deviation indicate a significant range in the quantity of films across genres, which reflects the diversity of the film industry's production priorities.

B. Movie Duration Analysis:

STATEMENT:

Examine the distribution of movie lengths and how it affects the IMDB score. To determine the connection between duration and IMDB score.

APPROACH:

To do the movie duration analysis, following functions were used:

- Average function: to calculate mean of the duration of the movie.
- Median Function: to calculate middle value of the duration of the movie.
- Stdev function: to calculate Standard deviation of the dataset.
- Correl function: to find out the relationship between the two variables duration and IMDB_SCORE.
- Scatter graph: To show relationship between duration and IMDB_SCORE.

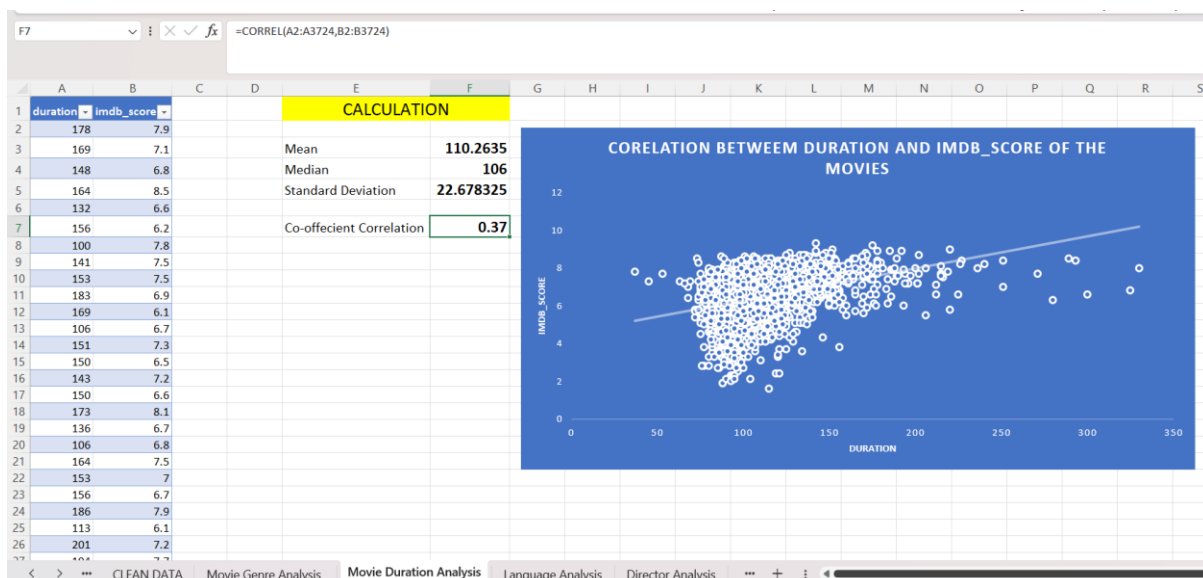


FIG 4 Partially Co-RELATED

INTERPRETATION:

- The average duration of a movie is 110.
- As the value of co-efficient correlation is far from positive 1, it is to have partially positive correlation between 2 variables duration and IMDB_SCORE.

C. language Analysis:

STATMENT:

Examine how movies are distributed according to their language. Analyse the impact of the most frequently used film languages on the IMDB score using descriptive statistics.

APPROACH:

To do the language analysis, following functions were used:

- Countif function: to calculate number of languages used by the IMDB.
- Max function: to calculate maximum number of Language used for the movies.
- Min function: to calculate minimum number of Language used for the movies.
- Average function: to calculate mean of the IMDB_SCORE.
- Median Function: to calculate middle value of the IMDB_SCORE.
- Mode function: to calculate repetition of number of the IMDB_SCORE.
- VAR function: to calculate how much the values in the dataset deviate from the mean.
- Stdev function: to calculate Standard deviation of the dataset.
- Correl function: to find out the relationship between the two variables Language and IMDB_SCORE.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	language	imdb_sc		All the Lang	Total NO. of Lang					Row Lab	Count of lang	Sum of imdb_score	
2	English	7.9		Aboriginal	2					Aboriginal	2	13.9	
3	English	7.1		Aramaic	1					Arabic	1	7.2	
4	English	6.8		Bosnian	1					Aramaic	1	7.1	
5	English	8.5		Cantonese	7					Bosnian	1	4.3	
6	English	6.6		Czech	1					Cantonese	7	51.4	
7	English	6.2		Danish	3					Czech	1	7.4	
8	English	7.8		Dari	2		Maximum Value	3558		Danish	3	23.7	
9	English	7.5		Dutch	3		Minimum Value	1		Dari	2	15	
10	English	7.5		English	3558					Dutch	3	22.7	
11	English	6.9		Filipino	1		Mean	6.465672844		English	3566	22920.5	
12	English	6.1		French	34		Median	6.6		Filipino	1	6.7	
13	English	6.7		German	10		Mode	6.7		French	34	250.1	
14	English	7.3		Hebrew	1					German	10	77.7	
15	English	6.5		Hindi	5		Variance	1.110164694		Hebrew	1	8	
16	English	7.2		Hungarian	1		Standard Deviation	1.053643533		Hindi	5	36.1	
17	English	6.6		Indonesian	2		Co-efficient correlation	0.999998735		Hungarian	1	7.1	
18	English	8.1		Italian	7					Indonesian	2	15.8	
19	English	6.7		Japanese	10					Italian	7	50.3	
20	English	6.8		Kazakh	1					Japanese	10	75.6	
21	English	7.5		Korean	5					Kazakh	1	6	
22	English	7		Mandarin	14					Korean	5	38.5	
23	English	6.7		Maya	1					Mandarin	14	98.3	
24	English	7.9		Mongolian	1					Maya	1	7.8	
25	English	6.1		None	1					Mongolian	1	7.3	
26	English	7.2		Norwegian	4					None	1	8.5	
27	English	7.7		Persian	3					Norwegian	4	28.6	
28	English	8.2		Portuguese	5					Persian	3	24.4	
29	English	5.9		Romanian	1					Portuguese	5	38.8	
30	English	7		Russian	1					Romanian	1	7.9	
31	English	7.8		Spanish	23					Russian	1	6.5	
32	English	7.3		Thai	3					Spanish	23	162.9	
33	English	7.2		Vietnamese	1					Thai	3	19.9	
34	English	6.5		Zulu	1					Vietnamese	1	7.4	
35	English	6.8								Zulu	1	7.3	
36	English	7.3								Grand Total	3723	24071.7	
37	English	6											

FIG 5 Language Analysis

INTERPRETATION:

- From the above images we can see that, English is the language which is used in majority of the movies.
- Language and IMDB_SCORE is perfectly positively correlated with each other the co-efficient correlation value is close to 1.

D. Director Analysis:

STATEMENT:

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

APPROACH:

To do director analysis, I have used Pivot Table. Following are the steps that I have taken:

- I selected the columns which I needed to analyse the same.
- From the Fig6 and Fig 7, we can clearly see the output of the steps.
- Index function was used to match the columns.
- Percentile function was used to find out the output.

INTERPRETATION:

- Steven Spielberg stands 1st in terms of contributing to the success of movies with 25 movies and on an average 7.5 IMDB score.

E. Budget Analysis:

STATEMENT:

Investigate the relationship between movie budgets and financial performance. Determine which films have the largest profit margins by examining the relationship between movie budgets and gross earnings.

APPROACH:

To do the budget analysis, following functions were used:

- Max function: to calculate maximum profit margin of a movie.
- Index function was used to match the columns.
- Correl function: to find out the relationship between the two variables budget and gross.

	A	B	C	D	E	F	G	H	I	J	K	L	M
	movie_title	budget	gross	Profit Margin									
1	Avatar	237000000	760505847	523505847									
2	Jurassic World	150000000	652177271	502177271		Maximum Profit	523505847						
3	Titanic	200000000	658672302	458672302		Name of the Movie	Avatar						
4	Star Wars: Episode IV - A New Hope	11000000	460935665	449935665									
5	E.T. the Extra-Terrestrial	10500000	434949459	424449459									
6	The Avengers	220000000	623279547	403279547		Co-officient Correlation	0.10						
7	The Lion King	45000000	422783777	377783777									
8	Star Wars: Episode I - The Phantom Menace	115000000	474544677	359544677									
9	The Dark Knight	185000000	53316061	348316061									
10	The Hunger Games	78000000	407999255	329999255									
11	Deadpool	58000000	363024263	305024263									
12	The Hunger Games: Catching Fire	130000000	424645577	294645577									
13	Jurassic Park	63000000	356784000	293784000									
14	Despicable Me 2	76000000	368049635	292049635									
15	American Sniper	58800000	350123553	291323553									
16	Finding Nemo	94000000	380838870	286838870									
17	Shrek 2	150000000	436471036	286471036									
18	The Lord of the Rings: The Return of the King	94000000	377019252	283019252									
19	Star Wars: Episode VI - Return of the Jedi	32500000	309125409	276625409									
20	Forrest Gump	55000000	329691196	274691196									
21	Star Wars: Episode V - The Empire Strikes Back	18000000	290158751	272158751									
22	Home Alone	18000000	285761243	267761243									
23	Star Wars: Episode III - Revenge of the Sith	113000000	380262555	267262555									
24	Spider-Man	139000000	403706375	264706375									
25	Minions	74000000	336029560	262029560									

FIG 9 Position Tier Analysis

INTERPRETATION:

- From the above images we can figure out that, budget and gross are partially positively correlated with each other as the value is far from 1.
- The maximum profit margin was of Avatar movie.

Data Story Report

Initial Problem

Consider that we are exploring a movie informational treasure box. In a world of genres, ratings, languages, directors, budgets, and revenues, it is our goal to find undiscovered gems. Searching for patterns and clues that reveal stories about the inner workings of movies is like being detectives.

Finding

- The most common movie genre is DRAMA.
- The high variance and standard deviation indicate a significant range in the quantity of films across genres, which reflects the diversity of the film industry's production priorities.
- Due to which we can say that the impact of genre on movie ratings is high.
- The movie duration and IMDB score is partially co related.
- The average distribution of movie duration 110.26.
- English is the most common language used in the IMDB movies.
- The impact of the language on IMDB score is perfectly positively correlated with each other.
- Steven Spielberg stands 1st in terms of contributing to the success of movies with 25 movies and on an average 7.5 IMDB score.
- The relationship between movie budget and gross earning is partially positively correlated.

- The movie with the highest profit margin is Avatar.

Insights

You can use the following information from the movie data to make wise decisions:

1. Leverage Drama's Popularity: To make the most of the most popular genre, concentrate on making distinctive and interesting drama films while considering genre-specific marketing techniques.
2. Customise by Genre: To increase audience engagement, adjust your production and marketing techniques in accordance with the various priorities of each genre.
3. Create Genre-Aligned Stories: Match your genre selections to the target audience perceptions to produce films that connect and get good reviews.
4. Put quality first within Maintain a balance between the length of the film and the level of narrative, putting an emphasis on interesting stories and well-rounded characters.
5. Consider creating films in English to reach a wider worldwide audience and take advantage of the language's familiarity.
6. Enhance Language Appeal: Create engrossing dialogue and stories that complement the beneficial effects of the selected language on IMDb ratings.
7. Work with Successful Directors: To improve the calibre and reception of your films, collaborate with accomplished directors like Steven Spielberg.
8. Script quality, casting, and marketing should be prioritised in addition to the money to create financial success that goes beyond the initial investment.
9. Learn from Successful Case Studies: Examine successful budget-to-earnings ratios used in films like Avatar to achieve a high profit margin as an example of how to succeed.
10. Execute with Innovation: Put these insights into practise to produce engrossing films that connect with viewers and are profitable both artistically and commercially.