IMDB MOVIE ANALYSIS

By Himanshu Pal

IMDB Movie Analysis

Submitted by Himanshu Pal

PROJECT DESCRIPTION

The IMDb Movie Analysis project focuses on examining and analyzing a detailed dataset of movies from the IMDb platform. This dataset includes key information such as movie titles, director names, duration, genres, budgets, gross earnings, IMDb ratings, and more. By applying data analysis methods using Excel, as well as data visualization and statistical techniques, the project aims to uncover meaningful insights and trends that influence a movie's success.

Problem Statement: 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.

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, you'll explore the data to understand the relationships between different variables. You might look at the correlation between movie ratings and other factors like genre, director, budget, etc. You might also want to consider the year of release, the actors involved, and other relevant factors.

Five 'Whys' Approach: This technique will help you dig deeper into the problem. For instance, if you find that movies with higher budgets tend to have higher ratings, you can ask "Why?" repeatedly to uncover the root cause. Here's an example:

Q: "Why do movies with higher budgets tend to have higher ratings?"
A: They can afford better production quality.
Q: "Why does better production quality lead to higher ratings?"
A: It enhances the viewer's experience.
Q: "Why does an enhanced viewer experience lead to higher ratings?"
A: Viewers are more likely to rate a movie highly if they enjoyed watching it.
Q: "Why are viewers more likely to rate a movie highly if they enjoyed watching it?"

A: Positive experiences lead to positive reviews.
Q: "Why do positive reviews matter?"
A: They influence other viewers' decisions to watch the movie, increasing its popularity and success.
Report and Data Story: After your analysis, you'll create a report that tells a story with your data. This should include your initial problem, your findings, and the insights you've gained. Use visualizations to help tell your story and make your findings more understandable.

Remember, as a data analyst, your goal is not just to answer questions but to provide insights that can drive decision-making. Your analysis should aim to provide actionable insights that can help stakeholders make informed decisions.

In this project, I will provide a detailed report for the below data record mentioning the answers of the questions that follows:

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

The Task is 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.

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

The Task is to Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.

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

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

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

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

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

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

APPROACH

I began by examining the dataset to understand all the available columns. The dataset originally contained 28 columns and 5,043 rows, but I noticed it included unnecessary columns, null values, and blank rows. To address these issues, I decided to clean the dataset thoroughly.

First, I removed columns that were irrelevant to the project and did not contribute valuable insights, reducing the dataset to nine key columns: director's name, duration, movie title, genre, budget, gross, IMDb rating, language, and country.

Next, I identified and removed blank rows by using the "Find & Select" feature, choosing "Go to Special," and selecting the "Blank" option to highlight them. I then used the shortcut "CTRL + -" and selected "Entire rows" to delete them.

Finally, I removed duplicate rows from the dataset, resulting in a cleaned dataset with 9 columns and 3,786 rows. The cleaned version of the dataset is provided below.

https://docs.google.com/spreadsheets/d/1QZcrT5BZhKOTA9_pnpaorlPPRI7wW4BCzT_FyVd0YQY/edit?usp=sharing

TECH STACK

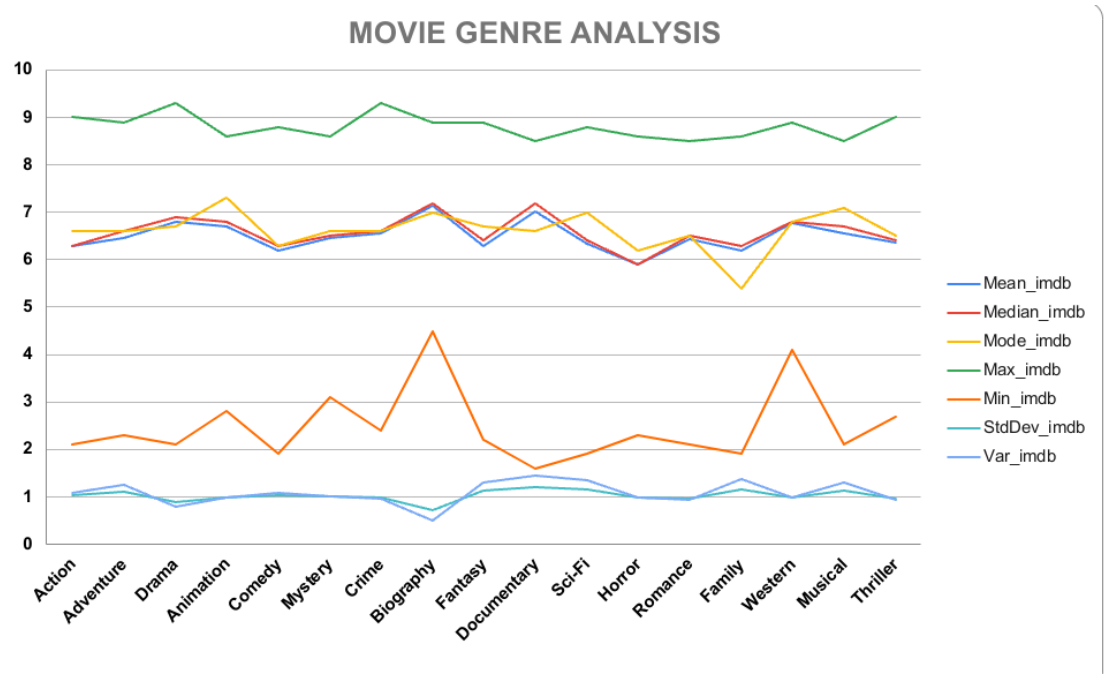
I have used Microsoft Excel 365 to run the functions and to find the answers for the given questions and to uncover valuable insights.

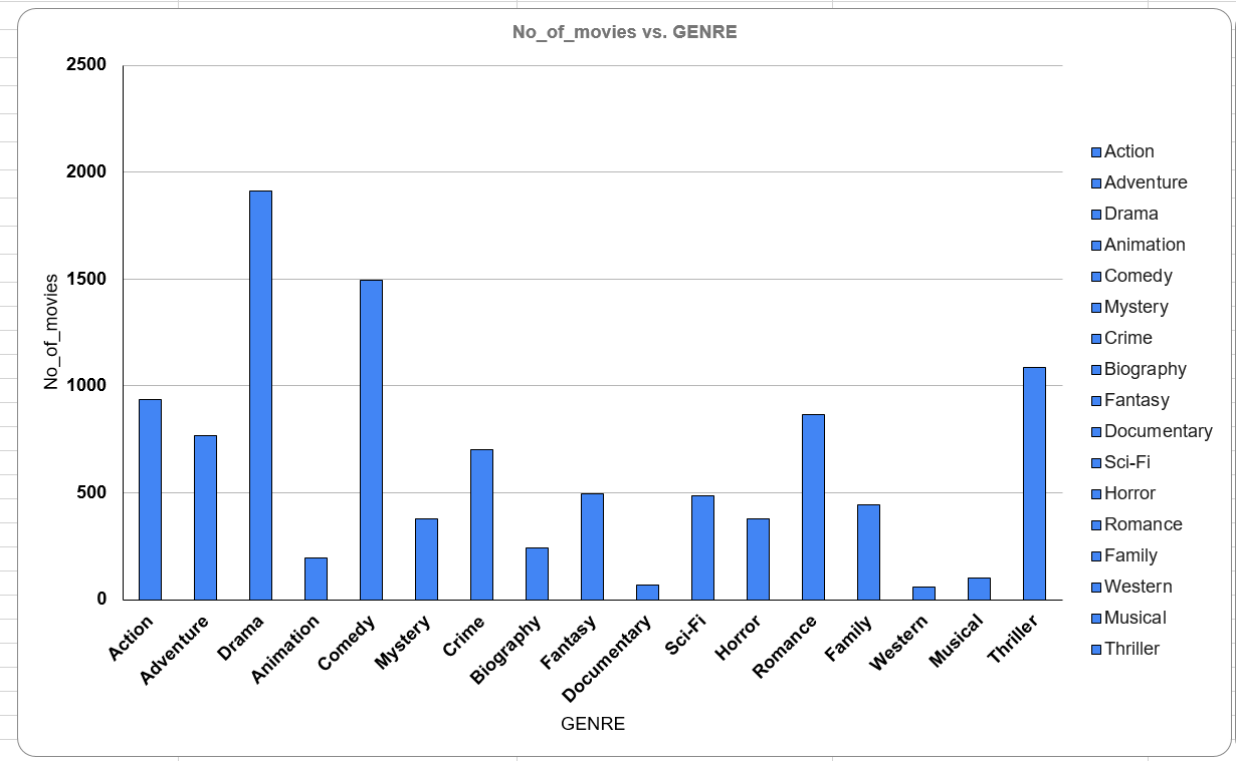
UNCOVERING INSIGHTS

1) **Movie Genre Analysis:**

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	No_of_movies	Mean_imdb	Median_imdb	Mode_imdb	Max_imdb	Min_imdb	StdDev_imdb	Var_imdb
Action	935	6.285989305	6.3	6.6	9	2.1	1.038357736	1.078186788
Adventure	766	6.454960836	6.6	6.6	8.9	2.3	1.116926308	1.247524378
Drama	1911	6.789115646	6.9	6.7	9.3	2.1	0.891064898	0.793996652
Animation	197	6.700507614	6.8	7.3	8.6	2.8	0.993627526	0.987295659
Comedy	1492	6.183310992	6.3	6.3	8.8	1.9	1.039919012	1.081431552
Mystery	377	6.469496021	6.5	6.6	8.6	3.1	1.007391835	1.014838309
Crime	702	6.548148148	6.6	6.6	9.3	2.4	0.984105199	0.968463042
Biography	242	7.140082645	7.2	7	8.9	4.5	0.71009671	0.504237338
Fantasy	496	6.285080645	6.4	6.7	8.9	2.2	1.140414241	1.30054464
Documentary	67	7.011940299	7.2	6.6	8.5	1.6	1.199939694	1.439855269
Sci-Fi	484	6.327272727	6.4	7	8.8	1.9	1.16718415	1.362318841
Horror	379	5.903957784	5.9	6.2	8.6	2.3	0.991023285	0.982127152
Romance	866	6.426212471	6.5	6.5	8.5	2.1	0.968996249	0.938953731
Family	441	6.2	6.3	5.4	8.6	1.9	1.169576458	1.367909091
Western	58	6.765517241	6.8	6.8	8.9	4.1	0.998516746	0.997035693
Musical	102	6.550980392	6.7	7.1	8.5	2.1	1.143535	1.307672297
Thriller	1087	6.372309108	6.4	6.5	9	2.7	0.969078327	0.939112803

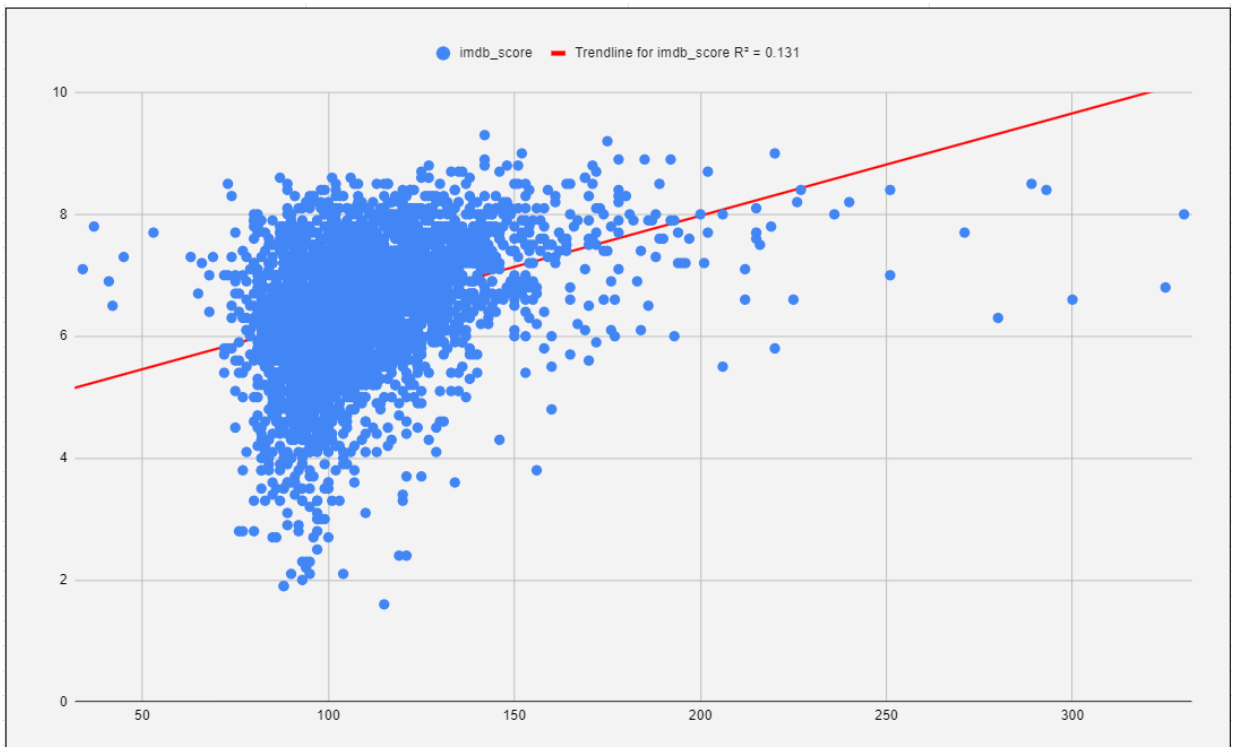

MOVIE GENRE ANALYSIS
10
9
8
7
6
5
4
3
2
1
0
Action
Adventure
Drama
Animation
Comedy
Mystery
Crime
Biography
Fantasy
Documentary
Sci-Fi
Horror
Romance
Family
Western
Musical
Thriller
Mean_imdb
Median_imdb
Mode_imdb
Max_imdb
Min_imdb
StdDev_imdb
Var_imdb


No_of_movies vs. GENRE
2500
2000
1500
1000
500
0
No_of_movies
Action
Adventure
Drama
Animation
Comedy
Mystery
Crime
Biography
Fantasy
Documentary
Sci-Fi
Horror
Romance
Family
Western
Musical
Thriller
GENRE

2. Movie Duration Analysis:

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

Operations	Values
Mean	109.808505
Median	105
Mode	101
Standard Devation	22.763201
Variance	518.16332



3. Language Analysis:

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

Language	No_of_movies	Average_imdb	Median_imdb	Var_imdb	StdDev_imdb
English	3606	6.421436495	6.5	1.107753941	1.052498903
French	37	7.286486486	7.2	0.3150900901	0.5613288609
Spanish	26	7.05	7.15	0.6826	0.8261961026
Mandarin	14	7.021428571	7.25	0.5864285714	0.765786244
German	13	7.692307692	7.7	0.4107692308	0.6409128106
Japanese	12	7.625	7.8	0.8093181818	0.8996211324
Hindi	10	6.76	7.05	1.236	1.111755369
Cantonese	8	7.2375	7.3	0.1941071429	0.4405759218
Italian	7	7.185714286	7	1.334761905	1.155318962
Korean	5	7.7	7.7	0.325	0.5700877125
Portuguese	5	7.76	8	0.958	0.9787747443
Norwegian	4	7.15	7.3	0.33	0.5744562647
Dutch	3	7.566666667	7.8	0.1633333333	0.4041451884
Thai	3	6.633333333	6.6	0.2033333333	0.4509249753
Danish	3	7.9	8.1	0.28	0.5291502622
Hebrew	3	7.5	7.3	0.19	0.4358898944

4. Movie Director Analysis:

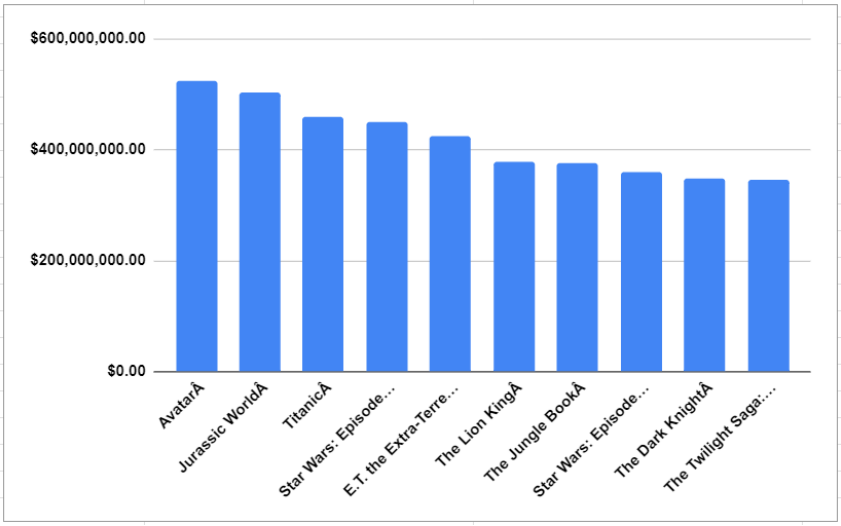
Task: Identify the top directors based on their average IMDB score and analyze their contribution to the success of movies using percentile calculations.

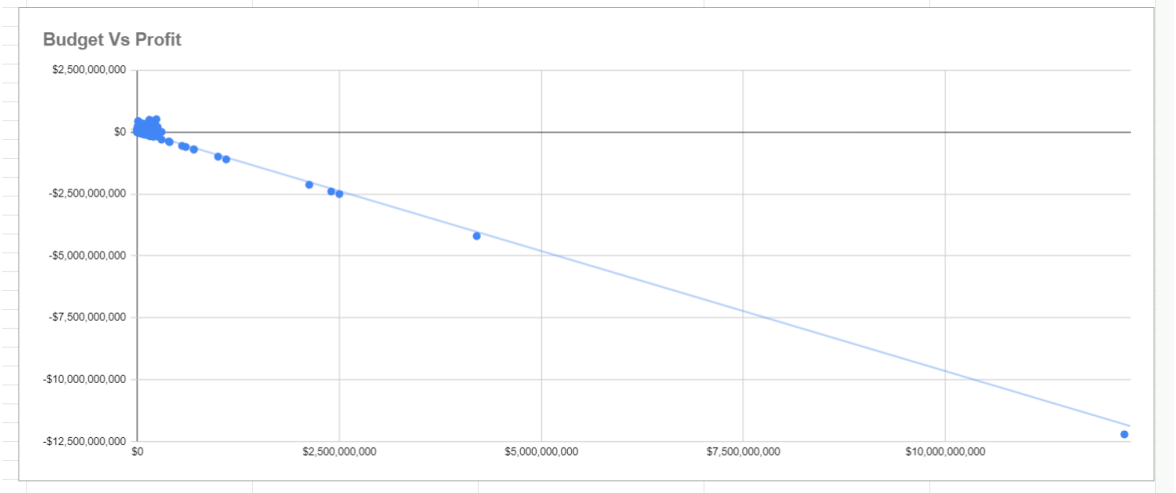
Director	Average_imdb	percentile	Count_movies
Tony Kaye	8.6	0.999	1
Charles Chaplin	8.6	0.999	1
Alfred Hitchcock	8.5	0.997	1
Ron Fricke	8.5	0.997	1
Damien Chazelle	8.5	0.997	1
Majid Majidi	8.5	0.997	1
Sergio Leone	8.433333333	0.996	3
Christopher Nolan	8.425	0.995	8
S.S. Rajamouli	8.4	0.993	1
Richard Marquand	8.4	0.993	1
Asghar Farhadi	8.4	0.993	1
Marius A. Markevicius	8.4	0.993	1
Lee Unkrich	8.3	0.991	1
Fritz Lang	8.3	0.991	1
Lenny Abrahamson	8.3	0.991	1
Billy Wilder	8.3	0.991	1

5. Movie Budget Analysis:

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

Movies	Profits in Millions
AvatarÂ	523505847
Jurassic WorldÂ	502177271
TitanicÂ	458672302
Star Wars: Episode IV - A New HopeÂ	449935665
E.T. the Extra-TerrestrialÂ	424449459
The Lion KingÂ	377783777
The Jungle BookÂ	375290282
Star Wars: Episode I - The Phantom Mena	359544677
The Dark KnightÂ	348316061
The Twilight Saga: Breaking Dawn - Part 2	344597846





Dataset link

https://1drv.ms/x/s!Am1F7NaSAdHZmUKQ1iLM6zjRIdJ_

I observed several key insights from the dataset:

- The most common movie genres are Drama, Comedy, Thriller, and Action.
- The average movie duration is 109 minutes, and the trendline for duration versus IMDb score shows an upward slope with an R^2 value of 0.131.
- The most frequently used languages in movies are English, French, Spanish, Mandarin, and German. Interestingly, Telugu and Persian have the highest average IMDb scores.
- The top 10 directors with an average IMDb score of 8.4 or higher include Tony Kaye, Charles Chaplin, Alfred Hitchcock, Ron Fricke, Damien Chazelle, Majid Majidi, Sergio Leone, Christopher Nolan, SS Rajamouli, and Richard Marquand.
- The top 5 most profitable movies are Avatar, Jurassic World, Titanic, Star Wars: Episode IV - A New Hope, and E.T. The Extra-Terrestrial. Additionally, there is a positive correlation between budget and gross earnings.

RESULTS

This project has provided me with valuable experience in data analysis, combining statistical knowledge with Excel's data visualization tools. Through this process, I have learned to effectively apply my data analysis skills to solve real-world problems.