IMDB Movie Analysis

BY-
RAHUL M
RAMCHANDANI

Project Description

We are given a dataset which is related to IMDB Movies. The problem to investigate is that: "What factors influence the success of a movie on IMDB?"

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. The project is about finding out valuable insights that can help to make informed decisions.

Tasks

SQL Tasks to be performed –

A. Movie Genre Analysis

B. Movie Duration Analysis

C. Language Analysis

D. Director Analysis

E. Budget Analysis

Software used :-

Microsoft Excel 2021

Approach

- To achieve our project goals, we'll start by downloading and familiarizing ourselves with the dataset. Once we have the data, we'll clean it up by removing any null values and deleting unnecessary columns.

- Then after data cleaning, we'll dive into the data analysis using Excel. This will involve manipulating the data to extract key information, calculating various statistics, and visualizing relationships between different variables. We'll use Excel functions and charts to find the answers we need.

- Finally, based on our analysis, we will create comprehensive reports to present our findings and insights.

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.

- First step involves cleaning the data.
- Columns like color, director_facebook_likes, actor_3_facebook_likes, actor_2_name, actor_1_facebook_likes, cast_total_facebook_likes, actor_3_name, facenumber_in_poster, plot_keywords, movie_imdb_link, content_rating, actor_2_facebook_likes, aspect_ratio, movie_facebook_likes are irrelevant data. It needs to be dropped.
- Now we need to remove the rows which contains null values. Then we need to remove duplicates from dataset.
- We will then we will separate multiple genres and use COUNTIF function to count the number of movies for each genre.
- Then we will use Excel's functions like AVERAGE, MEDIAN, MODE, MAX, MIN, VAR, and STDEV to calculate descriptive statistics.

Movie Genre Analysis

Formulas used-

1. Count : =COUNTIF($A$2:$A$3849, A2)
2. Mean : =AVERAGEIF(IMDB_Movies_Cleaned!$E$2:$E$3849, A2, IMDB_Movies_Cleaned!$N$2:$N$3849)
3. Median: =MEDIAN(IF(IMDB_Movies_Cleaned!$E$2:$E$3849=A2, IMDB_Movies_Cleaned!$N$2:$N$3849))
4. Mode: =MODE.SNGL(IF(IMDB_Movies_Cleaned!$E$2:$E$3849=A2, IMDB_Movies_Cleaned!$N$2:$N$3849))
5. Max: =MAX(IF(IMDB_Movies_Cleaned!$E$2:$E$3849=A2, IMDB_Movies_Cleaned!$N$2:$N$3849))
6. Min: =MIN(IF(IMDB_Movies_Cleaned!$E$2:$E$3849=A2, IMDB_Movies_Cleaned!$N$2:$N$3849))
7. Variance: =VAR.S(IF(IMDB_Movies_Cleaned!$E$2:$E$3849=A2, IMDB_Movies_Cleaned!$N$2:$N$3849))
8. Standard Deviation: =STDEV.S(IF(IMDB_Movies_Cleaned!$E$2:$E$3849=A2, IMDB_Movies_Cleaned!$N$2:$N$3849))

Movie Genre Analysis

Result -

The most common genres are -								
Genres	**Count**	**Average**	**Median**	**Mode**	**Max**	**Min**	**Variance**	**Standard Deviation**
Comedy|Drama|Romance	151	6.494702	6.5	6.5	8	4.3	0.562772	0.750181141
Comedy|Drama	147	6.583673	6.7	6.7	8.8	3.3	0.7348	0.857204825
Comedy	145	5.84069	6	6.5	8	1.9	1.481875	1.217322686
Comedy|Romance	135	5.896296	6	6.1	8.4	2.7	0.76827	0.87650999
Drama	153	7.04183	7.2	7.3	8.8	3.4	0.687055	0.828887522

Movie Duration Analysis

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

- First we will select columns- *duration* and *imdb_score*.
- Then we will use Excel's functions like AVERAGE, MEDIAN, and STDEV to calculate descriptive statistics.

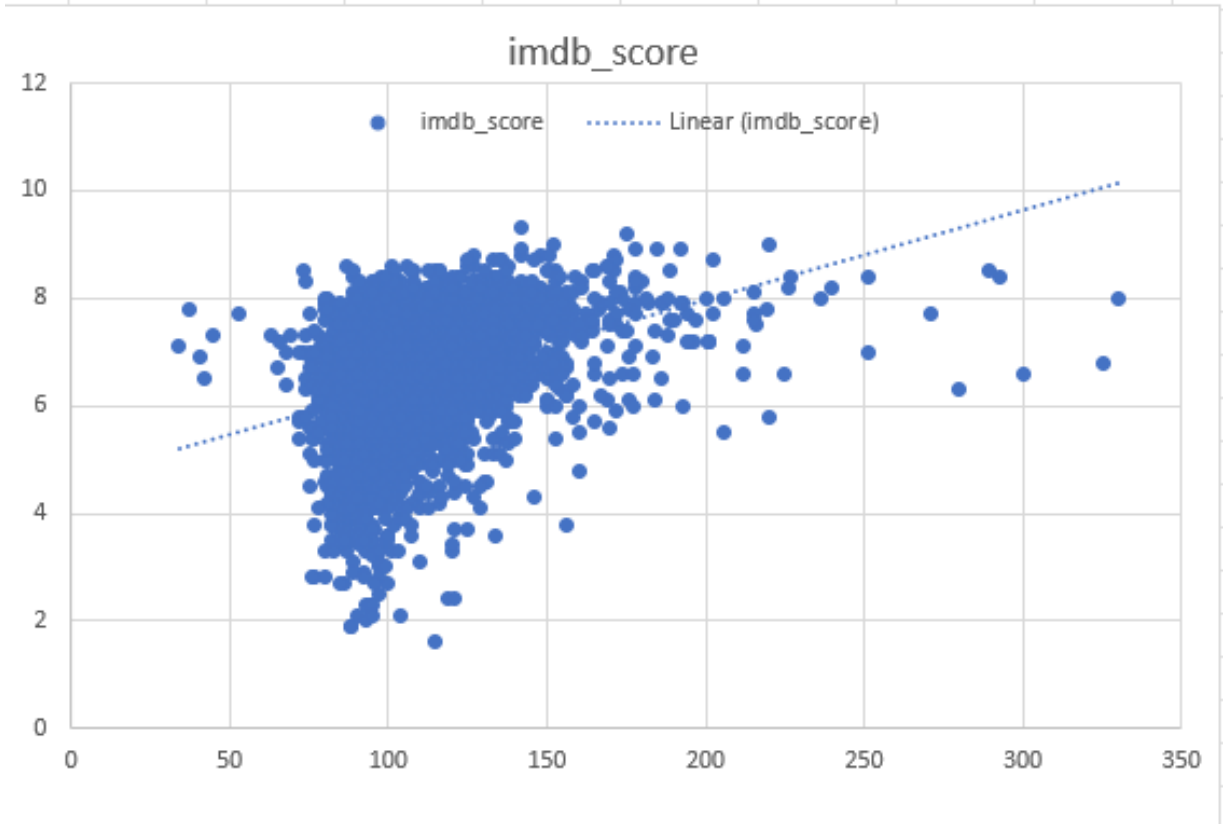
Formulas to be used :-

- Mean: =AVERAGE(A:A)
- Median:=MEDIAN(A:A)
- Standard deviation: =STDEV.S(A:A)

Movie Duration Analysis

Result -

Average	109.9241164
Median	106
Standard Deviation	22.75364979



Language Analysis

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

- First we will select these columns- *language* and *imdb_score*.
- We will then use COUNTIF function to count the number of movies for each language.
- Then we will calculate the Average, Median and standard deviation using the formulas below.

Formulas to be used :-

- Count: =COUNTIFS(IMDB_Movies_Cleaned!$J$2:$J$3849,A2)
- Mean: =AVERAGEIFS(IMDB_Movies_Cleaned!$N$2:$N$3849, IMDB_Movies_Cleaned!$J$2:$J$3849, A2)
- Median: =MEDIAN(IF(IMDB_Movies_Cleaned!$J$2:$J$3849=A2, IMDB_Movies_Cleaned!$N$2:$N$3849))
- Standard Deviation: =STDEV.S(IF(IMDB_Movies_Cleaned!$J$2:$J$3849=A2, IMDB_Movies_Cleaned!$N$2:$N$3849))

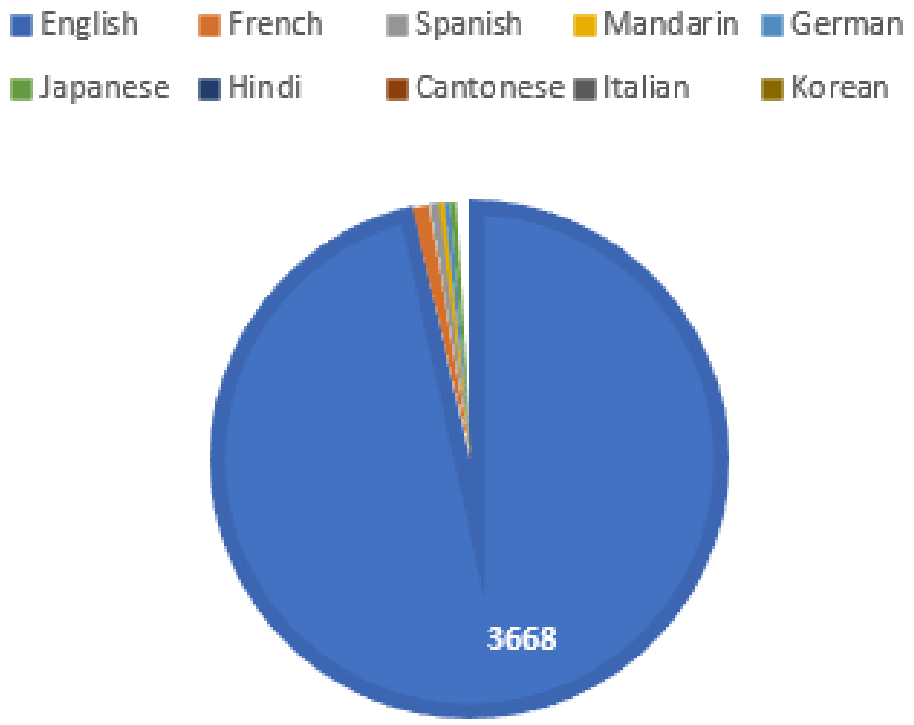
Language Analysis

Result -

	Most common Languages are:-			
Language	Count	Mean	Median	Standard Deviation
English	3668	6.423909	6.5	1.048750752
French	37	7.286486	7.2	0.561328861
Spanish	26	7.05	7.15	0.826196103
Mandarin	14	7.021429	7.25	0.765786244
German	13	7.692308	7.7	0.640912811
Japanese	12	7.625	7.8	0.899621132
Hindi	10	6.76	7.05	1.111755369
Cantonese	8	7.2375	7.3	0.440575922
Italian	7	7.185714	7	1.155318962
Korean	5	7.7	7.7	0.570087713

Language Analysis

Result -



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.

- We will select column- director_name and imdb_score.
- Then we will use AVERAGE function to Calculate the average IMDB score for each director.
- Then we will calculate percentrank and use PERCENTILE function to identify the directors with the highest scores.
- Formulas : -
- Average : =AVERAGE(IF(IMDB_Movies_Cleaned!$A$2:$A$3849=A2,IMDB_Movies_Cleaned!$N$2:$N$3849))
- Percentile : =PERCENTILE(J3:J11, J18)
- PercentRank : =PERCENTRANK(J3:J12,8.6)

Director Analysis

Result -

Top 10 Directors	Average
Charles Chaplin	8.60
Tony Kaye	8.60
Alfred Hitchcock	8.50
Damien Chazelle	8.50
Majid Majidi	8.50
Ron Fricke	8.50
Sergio Leone	8.43
Christopher Nolan	8.43
Asghar Farhadi	8.40
Marius A. Markevicius	8.40

LARGE	8.6
percentrank	0.888
percentile	8.6

Budget Analysis

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

- First we will have to calculate profit margin for each movie by subtracting movie budget value from gross value.
- Then, we will use CORREL function to calculate correlation coefficients between movie budgets and gross earnings.
- Using MAX function we will get highest profit margin.
- Formulas -
- Correlation : =CORREL('[imdb movies cleaned]IMDB_Movies_Cleaned'!D:D, '[imdb movies cleaned]IMDB_Movies_Cleaned'!L:L)
- Max: =Max(C:C)

Budget Analysis

Result -

CORRELATION
0.100850218

Highest Profit Margin	movie_title
523505847	Avatar

Insights/Conclusion

1. Most Common Genre is Drama
2. Most Common Language is English
3. Top Directors are **Charles Chaplin** and **Tony Kaye**
4. Movies with Highest Profit Margin is **Avatar**

Google drive link for Excel sheet –

https://docs.google.com/spreadsheets/d/19EF9tsH8ExaAtYUOeem00qrjdQC5kvBd/edit?usp=drive_link&ouid=109013092337571372406&rtpof=true&sd=true