IMDB Movie Analysis

Prachi Ranjan

PROJECT DESCRIPTION

The project is about finding out valuable insights that can help stakeholders make informed decisions. We analyze this data on the following points:

A. Movie Genre Analysis
B. Movie Duration Analysis
C. Language Analysis
D. Director Analysis
E. Budget Analysis

Software used:-

- Microsoft Excel 2307

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

Task A: 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 process involves cleaning the data. So dropping the columns which we have no use for analysis.
- 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.
- 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: Analyze the distribution of movie genres and their impact on the IMDB score.

Formulas:-

To count : =COUNTIF('cleaned data'!E$2:$E$3849, K2)

Mean : =AVERAGE(IF('cleaned data'!$E$2:$E$3849=A2, 'cleaned data'!$N$2:$N$3849))

Median: =MEDIAN(IF('cleaned data'!$E$2:$E$3849=A2, 'cleaned data'!$N$2:$N$3849))

Mode: =MODE(IF('cleaned data'!$E$2:$E$3849=A2, 'cleaned data'!$N$2:$N$3849))

Max: =MAX(IF('cleaned data'!$E$2:$E$3849=A2, 'cleaned data'!$N$2:$N$3849))

Min: =MIN(IF('cleaned data'!$E$2:$E$3849=A2, 'cleaned data'!$N$2:$N$3849))

Variance: =VAR(IF('cleaned data'!$E$2:$E$3849=A2, 'cleaned data'!$N$2:$N$3849))

Standard Deviation: =STDEV.S(IF('cleaned data'!$E$2:$E$3849=A2, 'cleaned data'!$N$2:$N$3849))

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

Output/Result:-

Most common genres are:-

genres	Count	Mean	Median	Mode	Max	Min	Variance	Standard Deviation
Drama	153	7.04183	7.2	7.3	8.8	3.4	0.687055	0.828887522
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

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

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

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

Formulas:-

Mean: =AVERAGE(A:A)

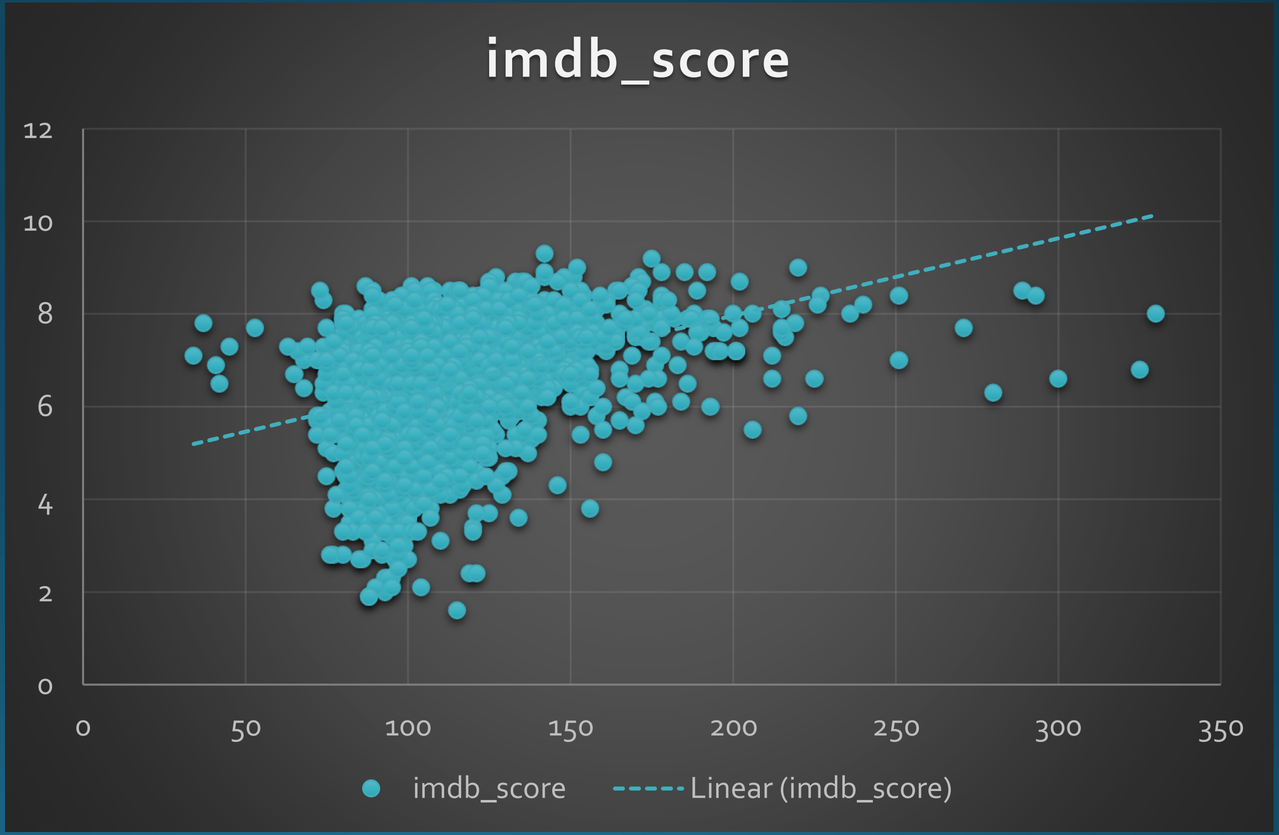
Median: =MEDIAN(A:A)

Standard deviation: =STDEV.S(A:A)

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

Output/Result:-

Average	109.9241164
Median	106
Standard Deviation	22.75364979



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

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

- First we will select Column **language** and **imdb_score.**
- Then we will use COUNTIF function to count the number of movies for each language.
- Using AVERAGE, MEDIAN, and STDEV function we will calculate Mean, Median and Standard Deviation of IMDB Scores for each language.

Formulas:-

Count: =COUNTIFS('cleaned data'!$J$2:$J$3849, J2)

Mean: =AVERAGE(IF('cleaned data'!$J$2:$J$3849=J2, 'cleaned data'!$N$2:$N$3849))

Median: =MEDIAN(IF('cleaned data'!$J$2:$J$3849=J2, 'cleaned data'!$N$2:$N$3849))

Standard Deviation: =STDEV.S(IF('cleaned data'!$J$2:$J$3849=J2, 'cleaned data'!$N$2:$N$3849))

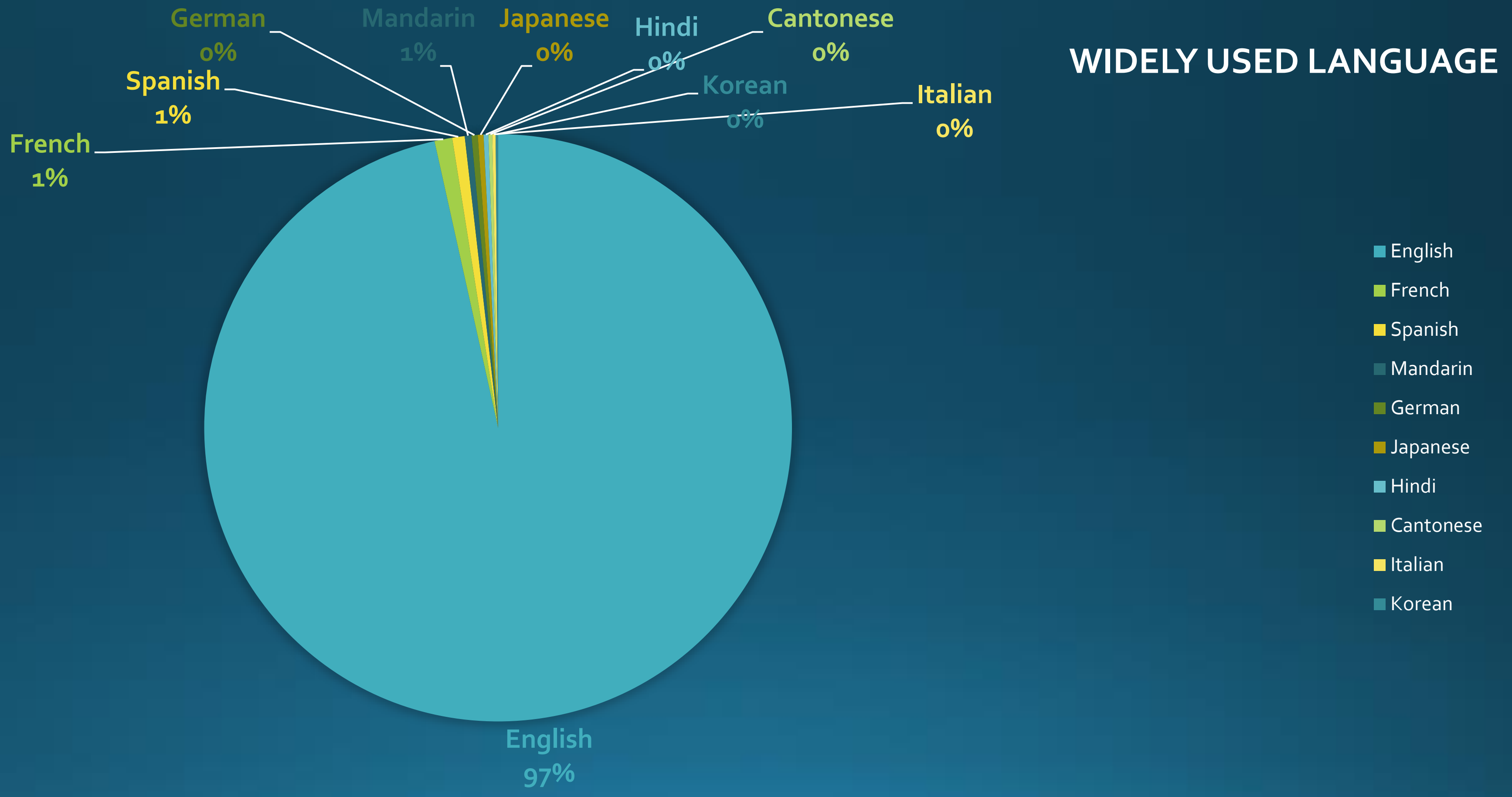
Language Analysis: Examine the distribution of movies based on their language.

Output/Results:-

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: Examine the distribution of movies based on their language.



Director Analysis: Influence of directors on movie ratings.

Task D: 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('cleaned data'!$A$2:$A$3849=A2, 'cleaned data'!$N$2:$N$3849))

Percentile: =PERCENTILE(H2:H11, H15)

Director Analysis: Influence of directors on movie ratings.

Output/Results:-

director_name	**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

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

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

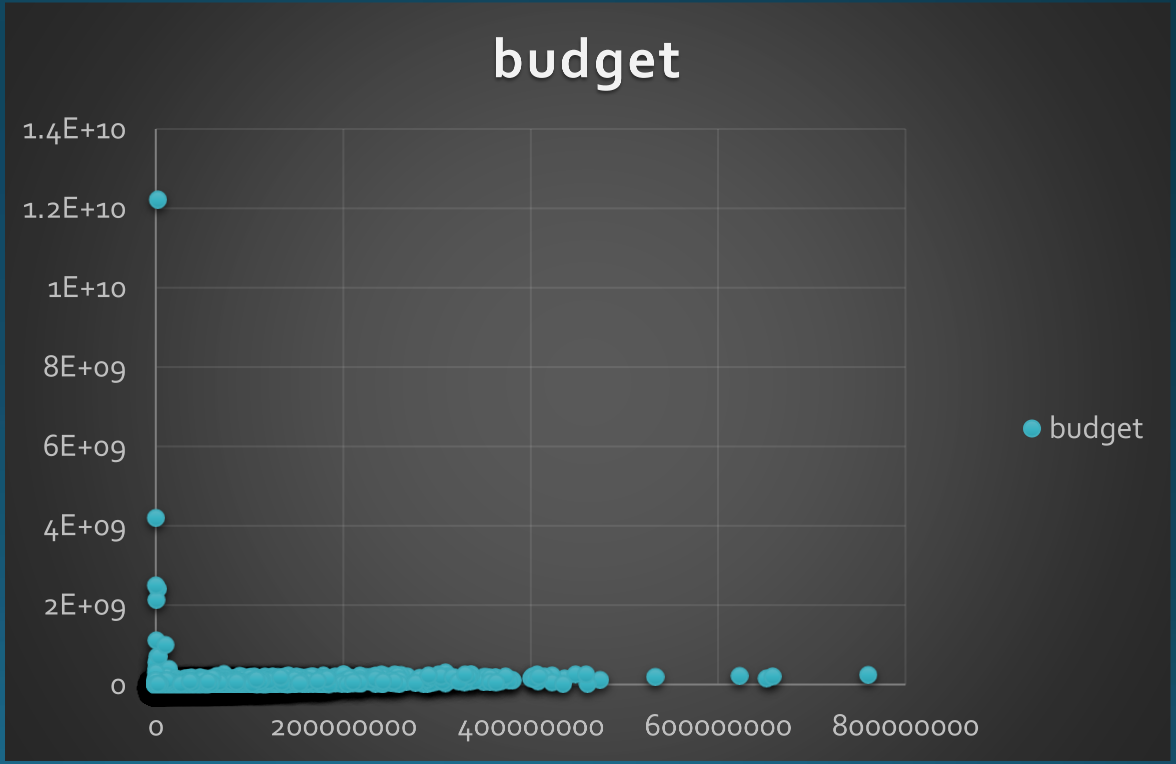
- First we will calculate profit margin for each movie by subtracting budget value from gross value.
- We will use CORREL function to calculate correlation coefficients between movie budgets and gross earnings.
- Using MAX function we will get highest profit margin then we will use =INDEX(B2:B3849, MATCH(1,IF(D2:D3849=G11, 1),0)) to get the title of the movie.

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

Output/Results:

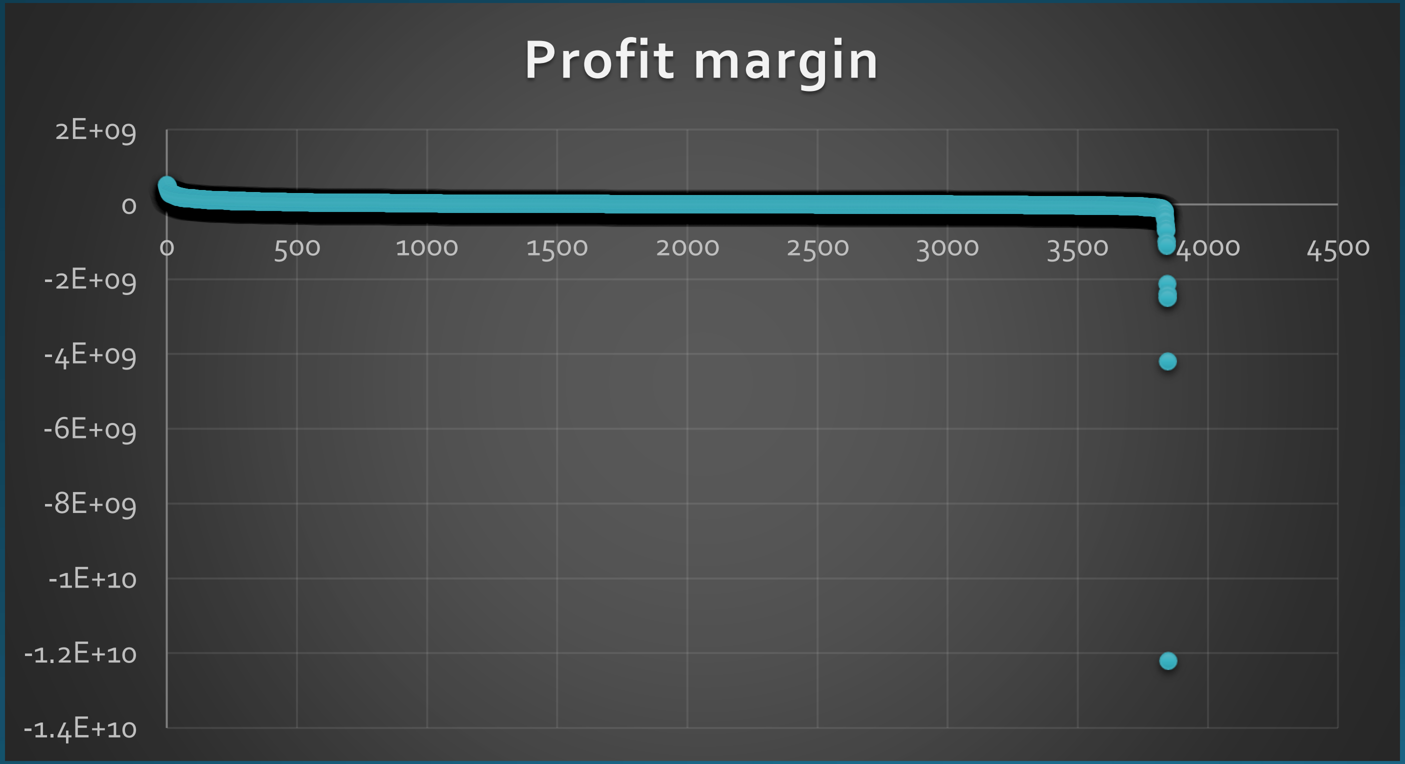
CORRELATION
0.100850218

MAX PROFIT	MOVIE TITLE
523505847	AvatarÂ



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

Output/Results:



CONCLUSION

- Most Common Genre is Drama
- Most Common Language is English
- Top Directors are Charles Chaplin and Tony Kaye
- Movies with Highest Profit Margin is AvatarÂ

Google Drive Link for Excel sheets:-

https://docs.google.com/spreadsheets/d/1B-i3ZebzaOnBpwvGrvuE0mjZasogFDP8pcobDq8ibeY/edit?usp=sharing