Final Project 1:
IMDB Movie Analysis

by-Syed Ali Ashraf

Project Description

The project focuses on understanding relationships between different variables given in data set such movie ratings ,genre, director, budget , language etc and its impact on movies for movie producers, directors, and investors who want to understand what makes a movie successful to make informed decisions in their future projects

First Step : Data Cleaning

First process involves cleaning the data.

As per the information required, the columns which had no relevant importance were dropped , Column names: 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 .

Total Count of column reduced from 28 in raw dataset to 14 .

Rows having null values were removed using the helper column and countblank() .

Further , duplicates were removed and data was cleaned to be used for analysis.

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

Sheet : Common Genre

1)Each Genre was counted using countif() function in count column

=COUNTIF(IMDB_Movies!$E$2:$E$3849,'Common Genre'!A2)

2)Statistical measures such as Mean , Median, Mode , Max , Min , Variance and Standard Deviation were calculated in separate columns for each genre.

Mean=AVERAGEIF(IMDB_Movies!$E$2:$E$3849,'Common Genre'!A2,IMDB_Movies!$N$2:$N$3849)

=MEDIAN(IF(IMDB_Movies!$E$2:$E$3849='Common Genre'!A2,IMDB_Movies!$N$2:$N$3849))

=MODE(IF(IMDB_Movies!$E$2:$E$3849='Common Genre'!A2,IMDB_Movies!$N$2:$N$3849))

=MAX(IF(IMDB_Movies!$E$2:$E$3849='Common Genre'!A2,IMDB_Movies!$N$2:$N$3849))

=MIN(IF(IMDB_Movies!$E$2:$E$3849='Common Genre'!A2,IMDB_Movies!$N$2:$N$3849))

=VAR(IF(IMDB_Movies!$E$2:$E$3849='Common Genre'!A2,IMDB_Movies!$N$2:$N$3849))

=STDEV.S(IF(IMDB_Movies!$E$2:$E$3849='Common Genre'!A2,IMDB_Movies!$N$2:$N$3849))

Array Functions

Values for all genre were calculated and sorting was done on count of genre from largest to smallest.

Data table:

Top 5 Genre								
Genre	Count	Mean	Median	Mode	Max	Min	Variance	Standard Deviation
Drama	153	7.041830065	7.2	7.3	8.8	3.4	0.687054524	0.828887522
Comedy|Drama|Romance	151	6.494701987	6.5	6.5	8	4.3	0.562771744	0.750181141
Comedy|Drama	147	6.583673469	6.7	6.7	8.8	3.3	0.734800112	0.857204825
Comedy	145	5.840689655	6	6.5	8	1.9	1.481874521	1.217322686
Comedy|Romance	135	5.896296296	6	6.1	8.4	2.7	0.768269762	0.87650999

Observations:

- Drama has the highest average rating (Mean = 7.04), showing that audiences generally rate drama movies higher than other genres in this list.

- Drama has the lowest variance (0.68) and standard deviation(0.83) indicating consistency in ratings indicating consistency in ratings while Comedy genre has the highest variance (1.48) and the standard deviation (1.21) showing the ratings very widely , some movies perform well and some not well .

- From the above data of Genre and Movie Ratings ,Genre Drama has overall better statistical outputs and resonates with audiences preferences.

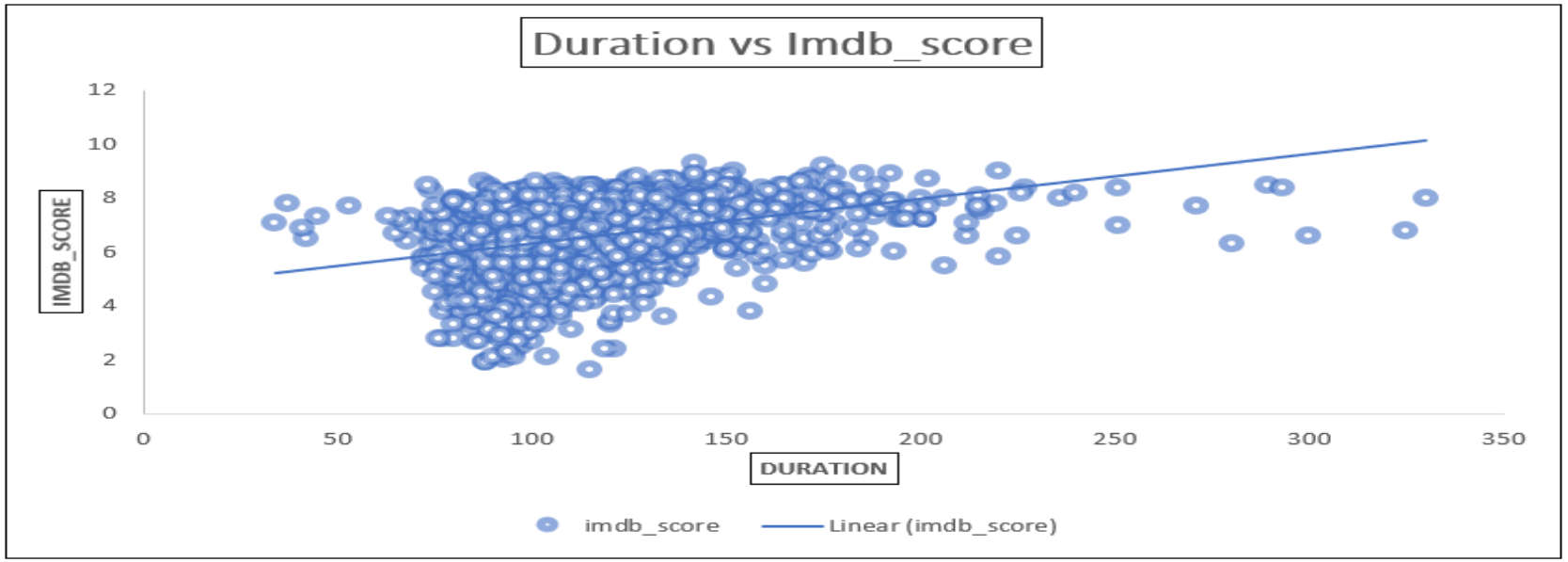
Task B. Movie Duration Analysis: Analyze the distribution of movie durations and its impact on the IMDB score.

Sheet : Movie Distribution

Descriptive Statistical values such as Average , Median and Standard Deviation were calculated from Duration Table.

Scatter Plot was plotted to visualize the relationship between Duration and IMDB score.

Duration's	
Average	109.924
Mean	106
Standard Deviation	22.7536



Observations from Scatter Plot

- Most movies fall between 50 to 200 minutes in duration.
- IMDb scores mostly range between 4 to 9, with a few outliers above 9 or below 4.
- A weak positive trend line suggests that increasing movie duration may result in higher scores,
- but not always.
- Some long-duration movies exist, but they have the highest ratings.

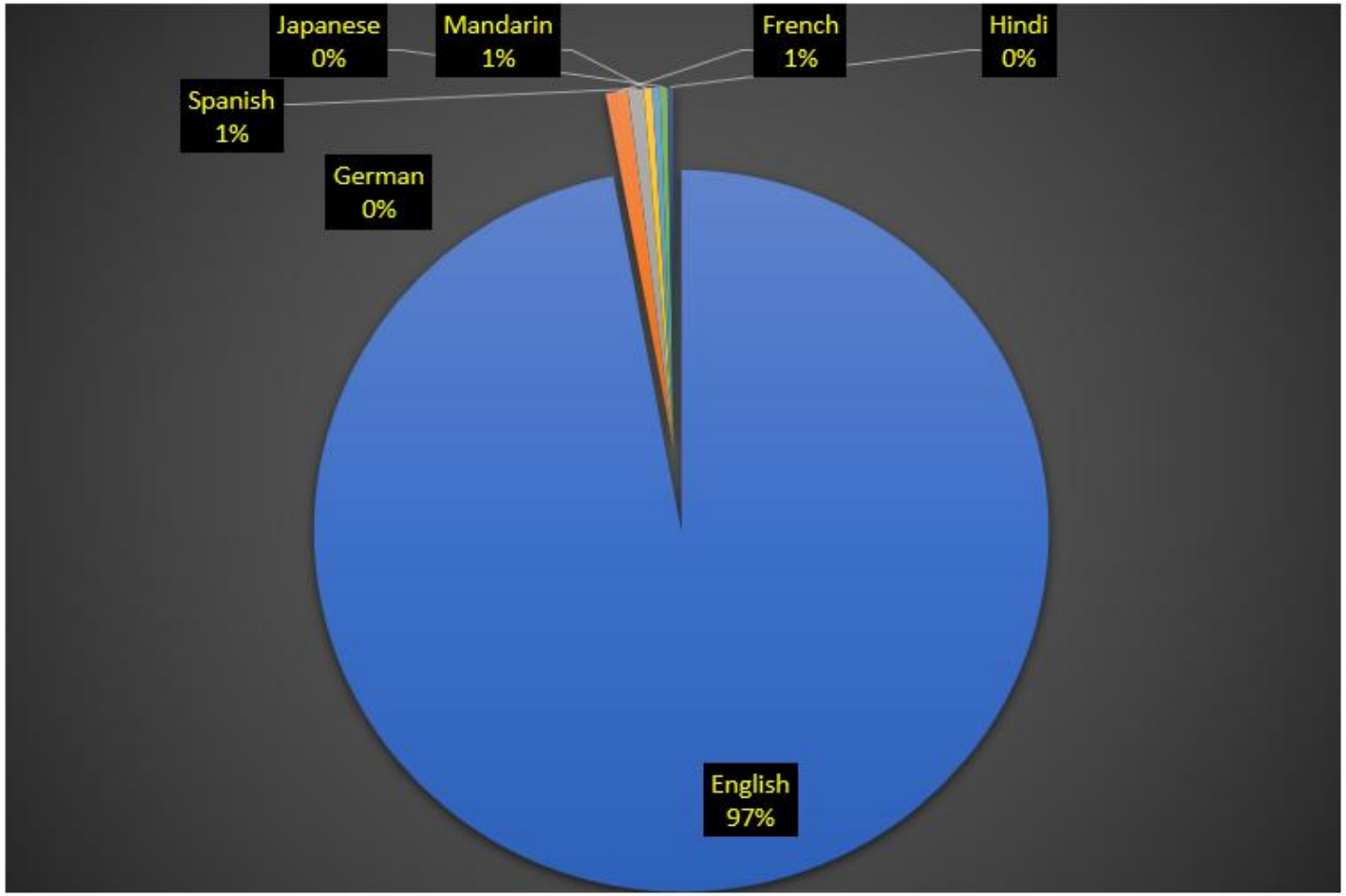
Task C. Language Analysis: Examine the distribution of movies based on their language

Sheet : Common Language

Determining the most common languages used in movies and analyze their impact on the IMDB score using descriptive statistics.
Used COUNTIF() function to count the number of movies for each language.
Using AVERAGE, MEDIAN, and Standard deviation function we will calculate Mean, Median and Standard Deviation of IMDB Scores for each language.

Top 7 Preferred Language

Language	Count	Mean	Median	Standard Deviation
English	3668	6.42391	6.5	1.048750752
French	37	7.28649	7.2	0.561328861
Spanish	26	7.05	7.15	0.826196103
Mandarin	14	7.02143	7.25	0.765786244
German	13	7.69231	7.7	0.640912811
Japanese	12	7.625	7.8	0.899621132
Hindi	10	6.76	7.05	1.111755369



Pie Chart distribution of languages.

Observations:

- English language movies dominate the dataset.
- German language(7.69) though in low count gets the highest mean rating followed by Japanese(7.62) , French(7.2) showing non English films receive higher ratings may be because of niche audiences.

Task D. Director Analysis: Influence of directors on movie ratings.

Sheet : Director's Analysis

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

Calculated Average IMDB rating for each Director followed by
Mean , Median and Percentile for top 10% IMDB ratings

Percentile =PERCENTILE(B2:B3849,0.9) , here 0.9 (or the 90th percentile) means that 90% of the data falls below this value, while the remaining 10% is above it.

Observations:

Director_name	Average IMDB
Charles Chaplin	8.6
Tony Kaye	8.6
Alfred Hitchcock	8.5
Damien Chazelle	8.5
Majid Majidi	8.5
Ron Fricke	8.5
Sergio Leone	8.433333333
Christopher Nolan	8.425
Asghar Farhadi	8.4
Marius A. Markevicius	8.4
Richard Marquand	8.4
S.S. Rajamouli	8.4

Percentile for top 10% IMDB ratings
7.7

Average of IMDB score
6.464163202

Median of IMDB score
6.6

- The top 10% of movies have IMDb ratings of 7.7 or higher, which means the directors listed in the table predominantly fall within this elite group.
- Directors like Charles Chaplin (8.6), Tony Kaye (8.6), Hitchcock (8.5), Nolan (8.42), and Sergio Leone (8.43) consistently create highly rated films

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

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

Profit was calculated by = **Gross-Budget**.

Correlation between Budgets and earnings =**CORREL(B2:B3849,C2:C3849)**

Correlation Coefficient
0.100850218

Calculated value of Correlation coefficient indicates a very weak positive correlation between budget and earnings and there is a slight tendency for higher budgets to result in higher earnings , but the relationship is not strong or consistent.

Profit was sorted from largest to lowest and highlighted using conditional formatting ,`

Movie with maximum profit :

Max Profit	Movie name
523505847	AvatarÂ

Conclusion:

Most Common Genre is **Drama**

Most Common Language is **English**

Top Directors are Charles Chaplin and Tony Kaye..

Movie with Highest Profit Margin is **AvatarÂ**

Tool used : MS Excel

Google drive link for excel sheet:
https://docs.google.com/spreadsheets/d/1jre39dwWyu2hOVxozdUn6dGrPRt_3PiE/edit?usp=drive_link&ouid=105115788278784376851&rtpof=true&sd=true