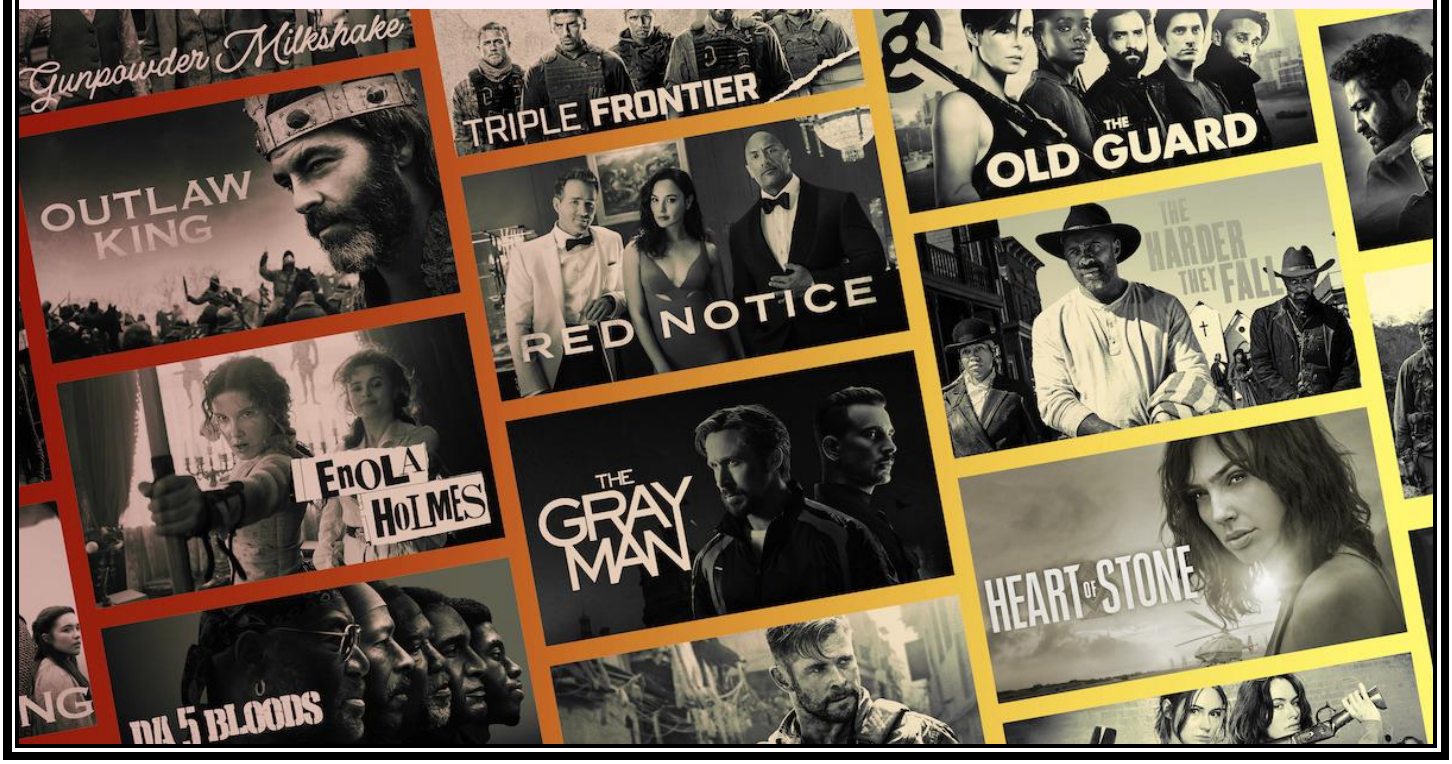
IMDB

MOVIE ANALYSIS



PROJECT MADE BY SAMIYA ALAM

PROJECT DESCRIPTION

The dataset provided is related to IMDB Movies, encompassing various attributes such as director names, movie duration, gross earnings, genres, actors, number of user and critic reviews, language, country, budget, IMDB scores, and Facebook likes. A potential problem to investigate is: "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 aim to understand what makes a movie successful to make informed decisions in their future projects.

IMDB rating, which describes the popularity of a movie in the public eye. This project seeks to analyze various factors that may influence these IMDB ratings.

TECH-STACK USED

Microsoft Excel for:

- **Data Cleaning and Preparation**:
  - Removing duplicates, handling missing values, and formatting data.
- **Data Analysis**:
  - Functions: COUNTIF, AVERAGE, MEDIAN, MODE, MAX, MIN, VAR, STDEV, PERCENTILE, CORREL etc.
  - Descriptive statistics calculations and summary.
  - Pivot tables for summarizing data.
- **Data Visualization**:
  - Scatter plots, bar charts, and trendlines.
  - Visualizing relationships and distributions.

Microsoft word:

- For making the submission presentation.

APPROACH

When conducting the IMDB Movies analysis project in Excel, I followed these steps:

1. **Define Objectives**:
    - Clearly articulated the goals to understand factors influencing movie success on IMDB.
2. **Data Collection**:
    - Downloaded the provided dataset, ensuring it included all relevant data points.
3. **Data Cleaning and Organization**:
    - Removed duplicates, handled missing values, and corrected formatting issues to prepare the data for analysis.
4. **Data Analysis Using Excel**:
    - Employed Excel formulas like COUNTIF, AVERAGE, MEDIAN, MODE, MAX, MIN, VAR, STDEV, PERCENTILE, and CORREL etc to analyze the data and answer specific questions.
5. **Visualization and Reporting**:
    - Created graphs and charts to visualize the findings.

DATA CLEANING

First, I analyzed the dataset in MS Excel to ascertain its structure, finding it comprised 5044 rows and 28 columns.

I decided to remove several columns that did not contribute to the insights we aimed to derive. These columns included: color, num_critic_for_reviews, director_facebook_likes, actor_3_facebook_likes, actor_2_name, actor_1_facebook_likes, num_voted_users, cast_total_facebook_likes, actor_3_name, facenumber_in_poster, plot_keywords, movie_imdb_link, num_user_for_reviews, country, content_rating, actor_2_facebook_likes, aspect_ratio, and movie_facebook_likes.

After streamlining the dataset by removing these unnecessary columns, I addressed missing and duplicate values. I removed all rows containing any null values, ensuring that our analysis would be based on complete data.

Additionally, I eliminated duplicate entries to avoid skewing the results. Following these steps, the dataset was reduced to 3877 rows, now cleaned and ready for detailed analysis

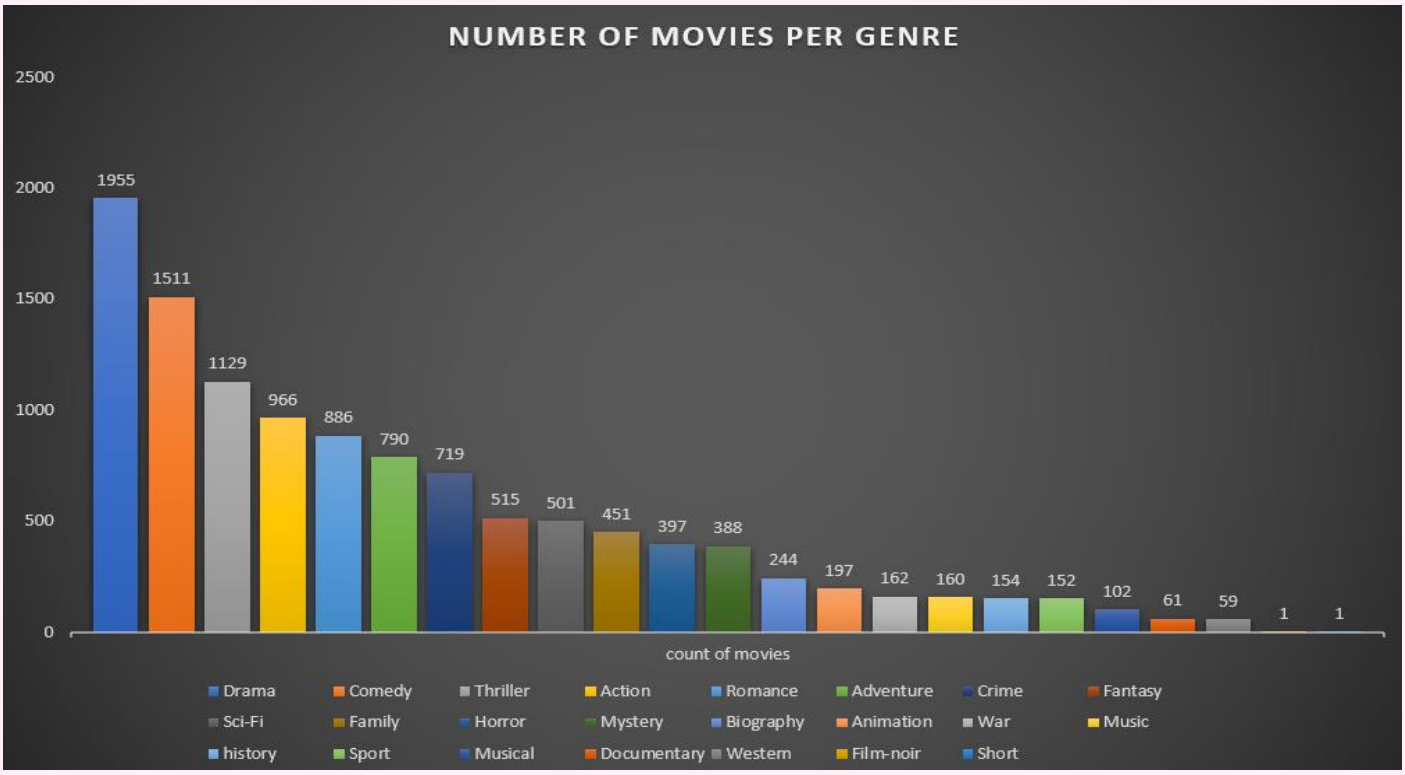
TASK A

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

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.

RESULTS:

					RANGE			
Genres	count of movies	mean	median	mode	max	min	var	stdev
Drama	1955	6.79	6.9	6.7	9.3	2.1	0.794	0.891
Comedy	1511	6.18	6.3	6.3	8.8	1.9	1.081	1.04
Thriller	1129	6.37	6.4	6.5	9	2.7	0.938	0.969
Action	966	6.29	6.3	6.6	9	2.1	1.077	1.038
Romance	886	6.43	6.5	6.5	8.5	2.1	0.938	0.968
Adventure	790	6.45	6.6	6.6	8.9	2.3	1.246	1.116
Crime	719	6.55	6.6	6.6	9.3	2.4	0.967	0.983
Fantasy	515	6.29	6.4	6.7	8.9	2.2	1.298	1.139
Sci-Fi	501	6.33	6.4	7	8.8	1.9	1.361	1.167
Family	451	6.2	6.3	5.4	8.6	1.9	1.365	1.168
Horror	397	5.9	5.9	6.2	8.6	2.3	0.979	0.989
Mystery	388	6.47	6.5	6.6	8.6	3.1	1.012	1.006
Biography	244	7.14	7.2	7	8.9	4.5	0.502	0.709
Animation	197	6.7	6.8	7.3	8.6	2.8	0.982	0.991
War	162	7.05	7.1	7.1	8.6	4.3	0.648	0.805
Music	160	6.37	6.5	6.5	8.5	1.6	1.465	1.21
history	154	7.13	7.2	7.7	8.9	5.5	0.449	0.67
Sport	152	6.6	6.8	7.2	8.4	2	1.091	1.045
Musical	102	6.55	6.7	7.1	8.5	2.1	1.295	1.138
Documentary	61	7.01	7.2	6.6	8.5	1.6	1.418	1.191
Western	59	6.77	6.8	6.8	8.9	4.1	0.98	0.99
Film-noir	1	7.7	7.7		7.7	7.7	0	0
Short	1	6.8	6.8		6.8	6.8	0	0



Initially, I utilized the "**Text to Columns**" function in the Data tab to separate the multiple genres listed in a single column into different columns. This step was essential for accurately analysing the genre information.

Next, I created a separate table that listed the names of the genres in the first row, with the corresponding IMDB ratings for each genre. This table served as the foundation for calculating descriptive statistics.

To derive insights, I employed the following Excel functions:

Count: =COUNTIF().

Mean: =AVERAGE()

Median: =MEDIAN()

Mode: =MODE.SNGL()

Maximum: =MAX()

Minimum: =MIN()

Variance: =VAR.P()

Standard Deviation: =STDEV.P()

After getting the results, I visualized and analysed it. I found that found that he maximum number of movies belong to **Drama genre.**

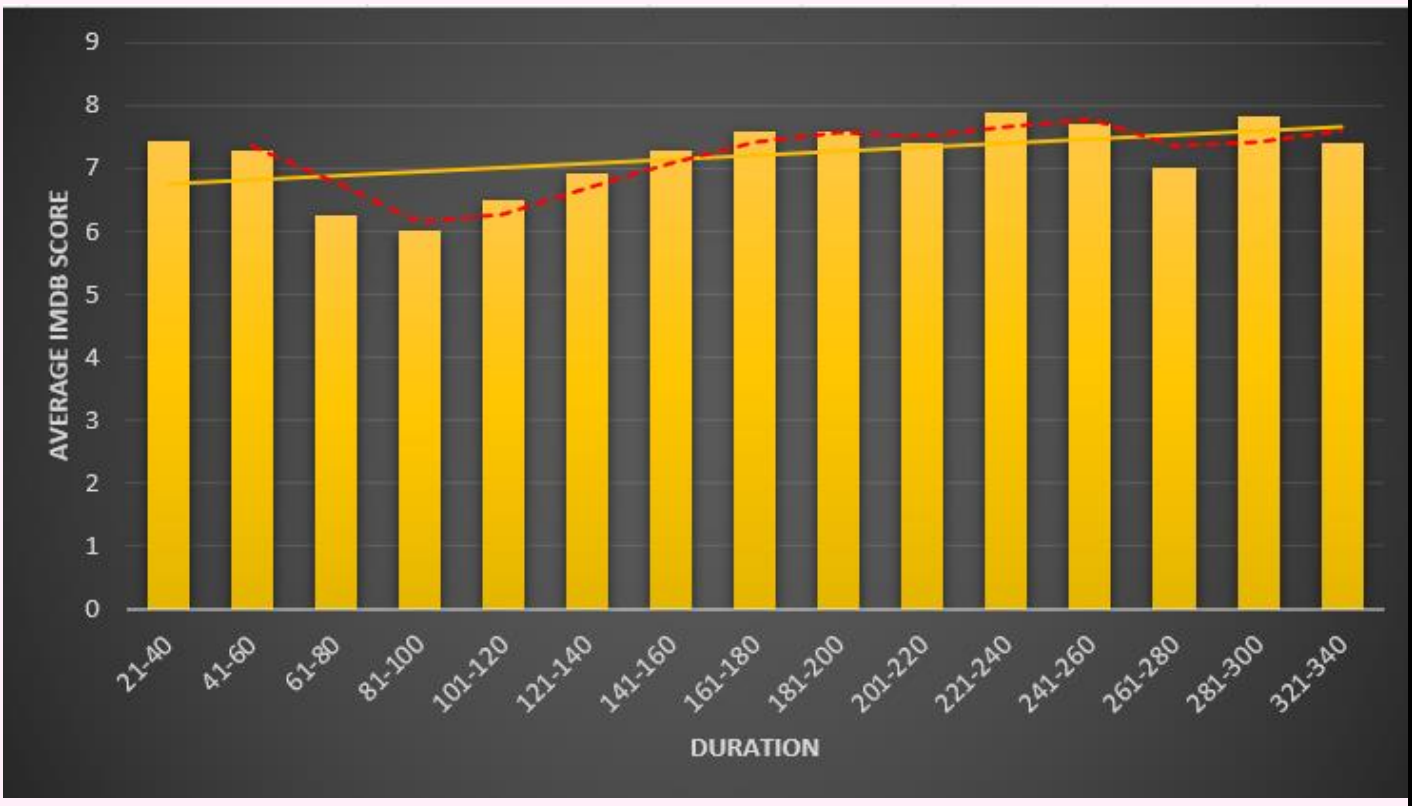
TASK B

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

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

RESULTS:

Duration range	Average of imdb_score	StdDev of imdb_score	Median
21-40	7.45	0.494974747	7.4
41-60	7.3	0.4	7.35
61-80	6.258333333	1.265105908	7.4
81-100	6.02453505	1.11203749	7.409375
101-120	6.51142664	0.895552418	7.41875
121-140	6.910815047	0.836955851	7.45
141-160	7.286705202	0.821935542	7.45
161-180	7.575438596	0.864638798	7.45
181-200	7.5875	0.729688736	7.41875
201-220	7.41875	0.913030668	7.4
221-240	7.88	0.729383301	7.3
241-260	7.7	0.989949494	7.286705202
261-280	7	0.989949494	7.286705202
281-300	7.833333333	1.069267662	7.3
321-340	7.4	0.848528137	7.3

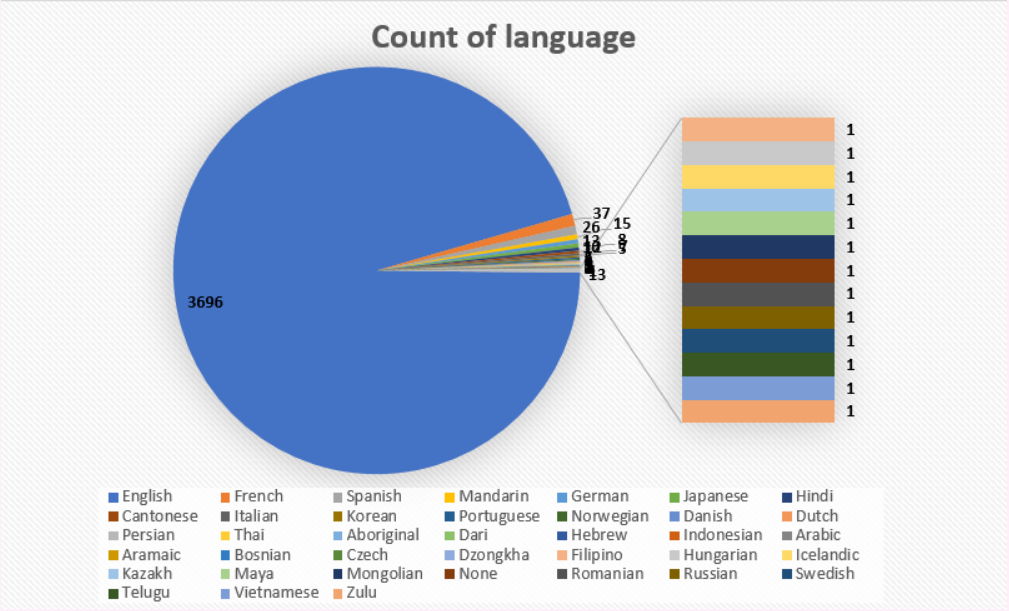


I created a separate table that displays class intervals in minutes in rows and the corresponding average IMDB scores, along with the median and standard deviation for each class interval using formulas. Then I visualized the data using a chart and added a trendline also. The average imdb is highest for 221-240 minutes

TASK C

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

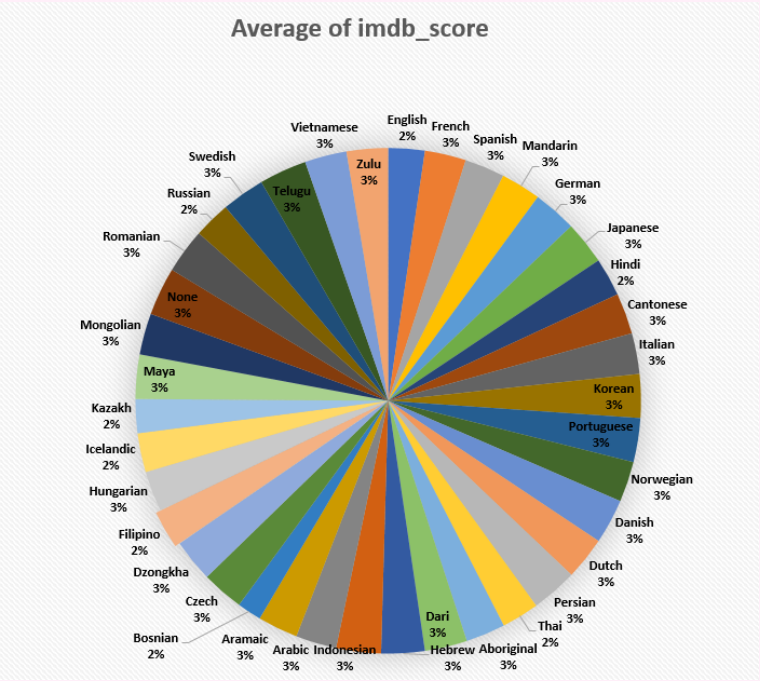
Task: Determine the most common languages used in movies and analyze their impact on the IMDB score using descriptive statistics.



RESULTS:

ENGLISH is found to be the most common language for the movies in this dataset.

LANGUAGE	Count of language	Average of imdb_score	StdDev of imdb_score	Median
English	3696	6.422510823	1.05128404	7.3
French	37	7.286486486	0.561328861	7.3
Spanish	26	7.05	0.826196103	7.35
Mandarin	15	7.08	0.772010363	7.4
German	13	7.692307692	0.640912811	7.4
Japanese	12	7.625	0.899621132	7.4
Hindi	10	6.76	1.111755369	7.35
Cantonese	8	7.2375	0.440575922	7.4
Italian	7	7.185714286	1.155318962	7.4
Korean	5	7.7	0.570087713	7.4
Portuguese	5	7.76	0.978774744	7.4
Norwegian	4	7.15	0.574456265	7.4
Danish	3	7.9	0.529150262	7.4
Dutch	3	7.566666667	0.404145188	7.4
Persian	3	8.133333333	0.550757055	7.35
Thai	3	6.633333333	0.450924975	7.3
Aboriginal	2	6.95	0.777817459	7.35
Dari	2	7.5	0.141421356	7.4
Hebrew	2	7.65	0.494974747	7.35
Indonesian	2	7.9	0.424264069	7.3
Arabic	1	7.2	N.A.	N.A.
Aramaic	1	7.1	N.A.	N.A.
Bosnian	1	4.3	N.A.	N.A.
Czech	1	7.4	N.A.	N.A.
Dzongkha	1	7.5	N.A.	N.A.
Filipino	1	6.7	N.A.	N.A.
Hungarian	1	7.1	N.A.	N.A.
Icelandic	1	6.9	N.A.	N.A.
Kazakh	1	6	N.A.	N.A.
Maya	1	7.8	N.A.	N.A.
Mongolian	1	7.3	N.A.	N.A.
None	1	8.5	N.A.	N.A.
Romanian	1	7.9	N.A.	N.A.
Russian	1	6.5	N.A.	N.A.
Swedish	1	7.6	N.A.	N.A.
Telugu	1	8.4	N.A.	N.A.
Vietnamese	1	7.4	N.A.	N.A.
Zulu	1	7.3	N.A.	N.A.



I used excel formulas for these tasks

For count: =COUNTIF() For mean: =AVERAGEIF() For median: =MEDIAN() For std deviation: =STDEV.P().

TASK D

Director Analysis: Influence of directors on movie ratings.

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

RESULTS:

Top directors according to **movie count:**

director_name	Movie count	Average IMDB	Percentile rank
Steven Spielberg	25	7.544	91
Clint Eastwood	19	7.205263158	84.1
Woody Allen	19	7	72.6
Ridley Scott	17	7.070588235	77.1
Martin Scorsese	16	7.675	93.8
Tim Burton	16	6.93125	71.6
Steven Soderbergh	16	6.70625	62.2
Spike Lee	15	6.733333333	62.5
Renny Harlin	15	5.746666667	24.6
Robert Zemeckis	13	7.307692308	86.9
Ron Howard	13	6.930769231	71.6
John Carpenter	13	6.915384615	71.1
Oliver Stone	13	6.907692308	71.1
Michael Bay	13	6.638461538	58.8
Barry Levinson	13	6.576923077	55.2
Robert Rodriguez	13	5.692307692	22.8
Peter Jackson	12	7.675	93.9
Sam Raimi	12	6.85	67.9
Tony Scott	12	6.791666667	63.9
Joel Schumacher	12	6.341666667	45.5
Shawn Levy	12	6.033333333	33.6
Wes Craven	12	6	31
Richard Linklater	11	7.327272727	87.2
Rob Reiner	11	7.018181818	76.4

Top directors according to **IMDB Score**:

director_name	Movie count	Average IMDB	Percentile rank
Charles Chaplin	1	8.6	99.9
Tony Kaye	1	8.6	99.9
Alfred Hitchcock	1	8.5	99.7
Damien Chazelle	1	8.5	99.7
Majid Majidi	1	8.5	99.7
Ron Fricke	1	8.5	99.7
Sergio Leone	3	8.433333333	99.6
Christopher Nolan	8	8.425	99.5
Asghar Farhadi	1	8.4	99.3
Marius A. Markevicius	1	8.4	99.3
Richard Marquand	1	8.4	99.3
S.S. Rajamouli	1	8.4	99.3
Billy Wilder	1	8.3	99.1
Fritz Lang	1	8.3	99.1
Lee Unkrich	1	8.3	99.1
Lenny Abrahamson	1	8.3	99.1
Pete Docter	3	8.233333333	99
Hayao Miyazaki	4	8.225	99
Quentin Tarantino	8	8.2	98.9
George Roy Hill	2	8.2	98.7
Elia Kazan	1	8.2	98.7
Joshua Oppenheimer	1	8.2	98.7
Juan JosÃ© Campanella	1	8.2	98.7
Victor Fleming	2	8.15	98.6

To complete the above task, I used following formulas in excel:

For Number of movies of each director: =COUNTIF()

For average imbd director wise: =AVERAGEIF()

For calculating percentile rank: =PERCENTRANK.INC([average imdb],[@[average imdb]])*100

Then I pressed ctrl+shift+l to add filter and did sorting accordingly in the table.

The director with the most movies in the dataset is Steven Spielberg, having directed 25 films. However, the highest average IMDB score belongs to Charles Chaplin. But Charles Chaplin directed only one movie in the dataset, using his average score as a basis for analysis may not be correct.

TASK E

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

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

RESULT:

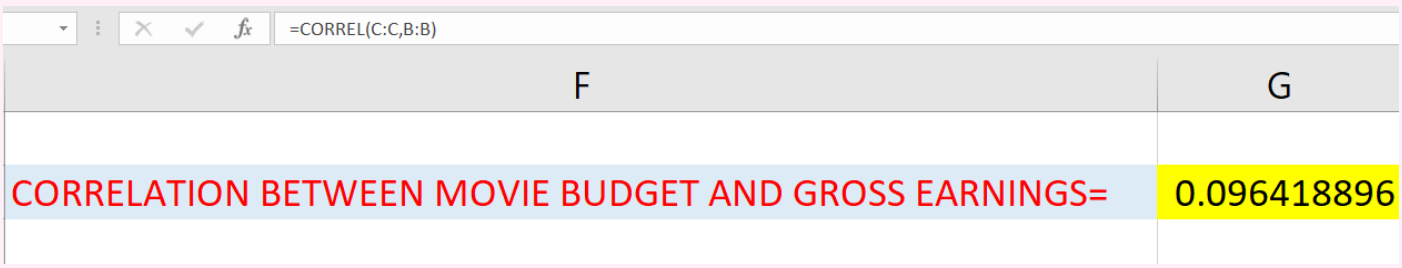
movie_title	gross	budget	profit
AvatarÂ	760505847	237000000	52350584
Jurassic WorldÂ	652177271	150000000	50217727
TitanicÂ	658672302	200000000	45867230
Star Wars: Episode IV - A New HopeÂ	460935665	11000000	44993566
E.T. the Extra-TerrestrialÂ	434949459	10500000	42444945
The AvengersÂ	623279547	220000000	40327954
The Lion KingÂ	422783777	45000000	37778377
Star Wars: Episode I - The Phantom MenaceÂ	474544677	115000000	35954467
The Dark KnightÂ	533316061	185000000	34831606
The Hunger GamesÂ	407999255	78000000	32999925
DeadpoolÂ	363024263	58000000	30502426
The Hunger Games: Catching FireÂ	424645577	130000000	29464557
Jurassic ParkÂ	356784000	63000000	29378400
Despicable Me 2Â	368049635	76000000	29204963
American SniperÂ	350123553	58800000	29132355
Finding NemoÂ	380838870	94000000	28683887
Shrek 2Â	436471036	150000000	28647103
The Lord of the Rings: The Return of the KingÂ	377019252	94000000	28301925
Star Wars: Episode VI - Return of the JediÂ	309125409	32500000	27662540
Forrest GumpÂ	329691196	55000000	27469119
Star Wars: Episode V - The Empire Strikes BackÂ	290158751	18000000	27215875
Home AloneÂ	285761243	18000000	26776124
Star Wars: Episode III - Revenge of the SithÂ	380262555	113000000	26726255
Spider-ManÂ	403706375	139000000	26470637
MinionsÂ	336029560	74000000	26202956

Maximum profit is of “avatar” movie which is Rs. 52,35,05,847.

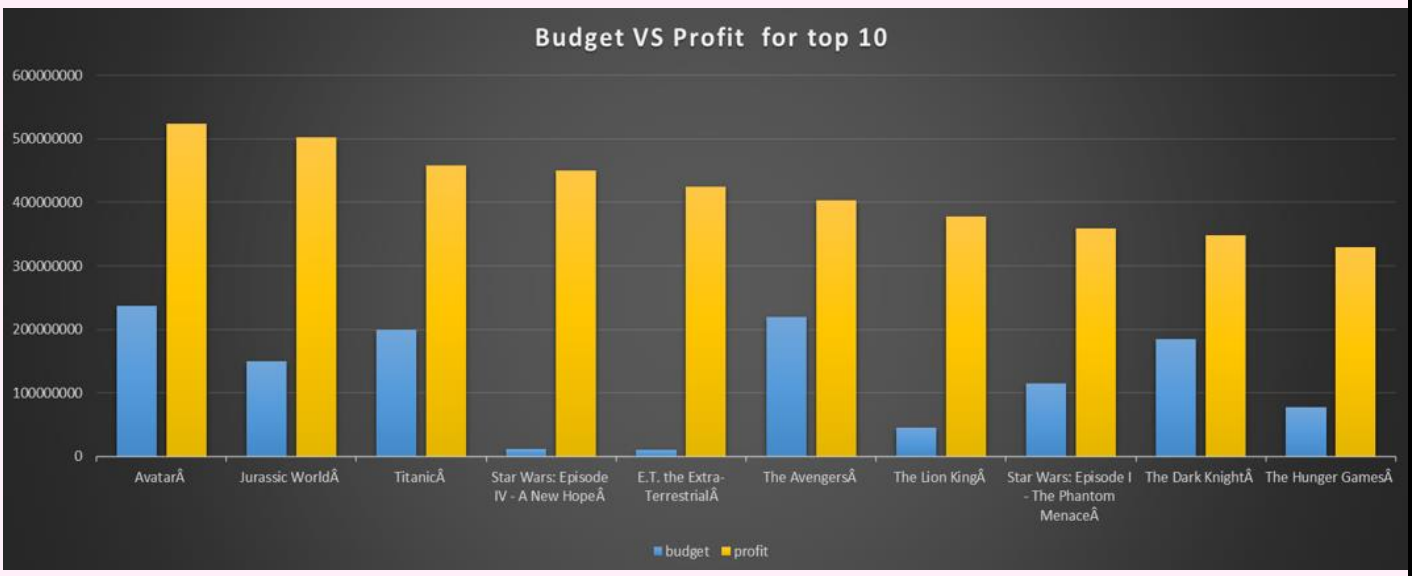
For profit I subtracted budget from gross earnings

For correlation coefficient used formula: =CORREL()

Correlation:



Visualisation:



CONCLUSION

Through this project, I achieved several significant milestones. I ensured the accuracy of my analysis by meticulously cleaning and preparing the dataset, removing unnecessary columns, and handling missing values. I was able to identify the impact of different genres on IMDB scores and explore the relationship between movie durations and their ratings. Additionally, I highlighted the most successful languages and directors based on average IMDB scores and analysed the correlation between movie budgets and gross earnings, pinpointing high-profit films.

This project greatly contributed to my understanding of IMDB movie analysis. I recognized the critical importance of thorough data cleaning to ensure accurate results. My statistical analysis skills were enhanced as I became proficient in using Excel for detailed data analysis. I gained valuable insights into how different genres and movie durations affect ratings, and I learned about the significant impact of directors and languages on a movie's success. Furthermore, I developed a deeper understanding of how movie budgets correlate with financial performance, which can inform better financial planning and investment decisions in the film industry.

For excel sheet:
https://drive.google.com/file/d/1RbqBmKUUuVxcpvQVy3EJJ_2lwLT3k02G/view?usp=sharing

PROJECT MADE BY:

SAMIYA ALAM

Samiyaalam1710@gmail.com