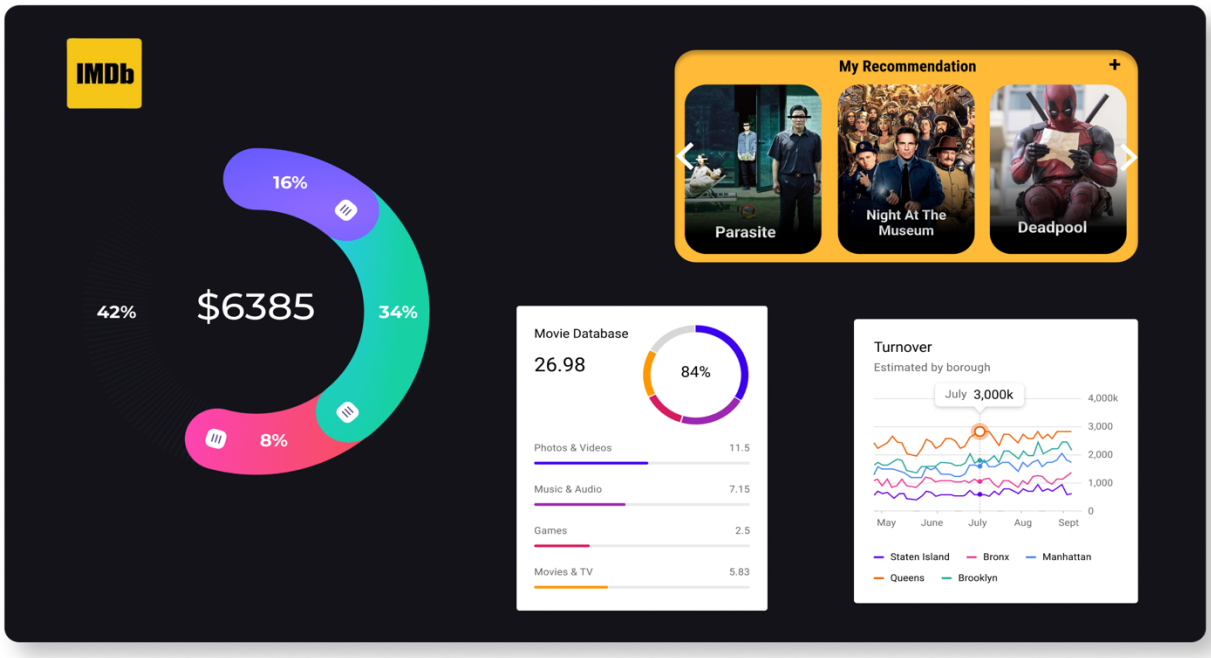
IMDB Movie Analysis



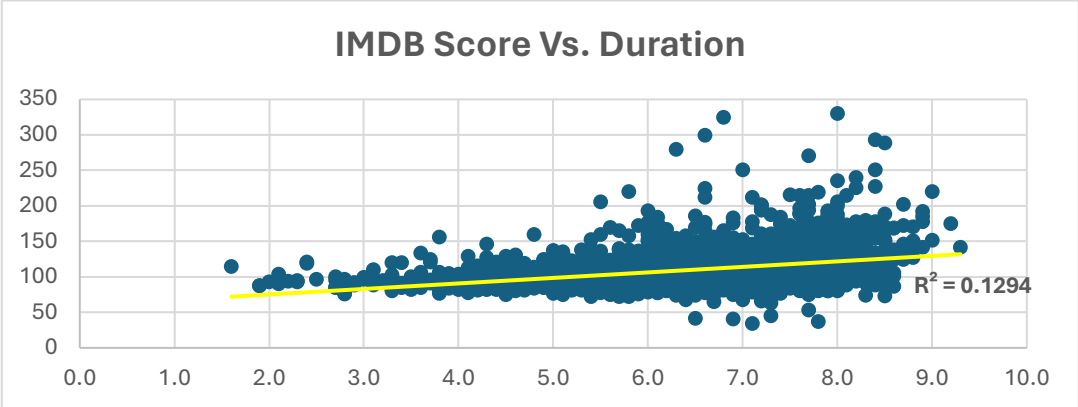
Project Description:

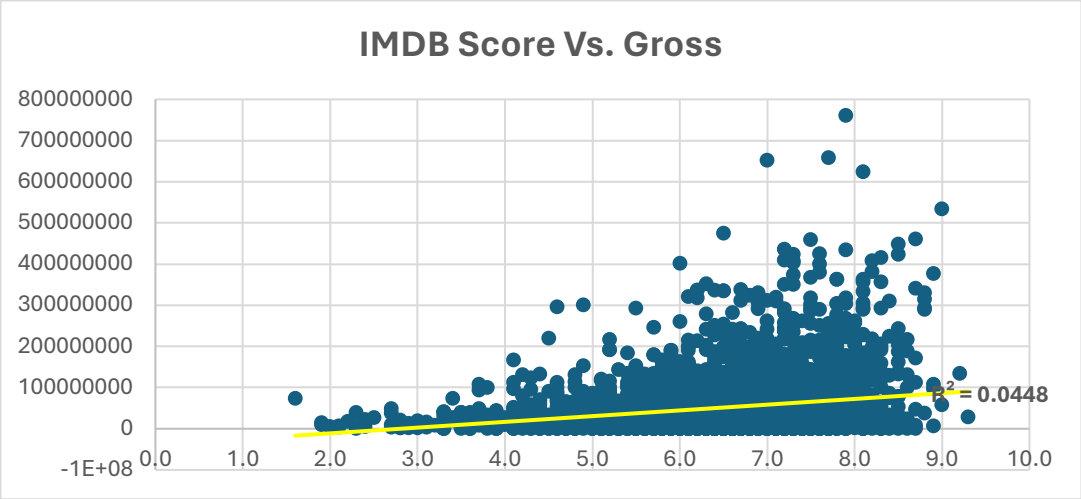
As a data analyst, I analyse the movie data on IMDB to find potential reasons or factors responsible for a film's high ratings, which contribute to its success and What impact this might have on filmmakers, financiers, and producers of motion pictures who want to know what makes a film successful so they can make wise choices for their next endeavours.

Approach:

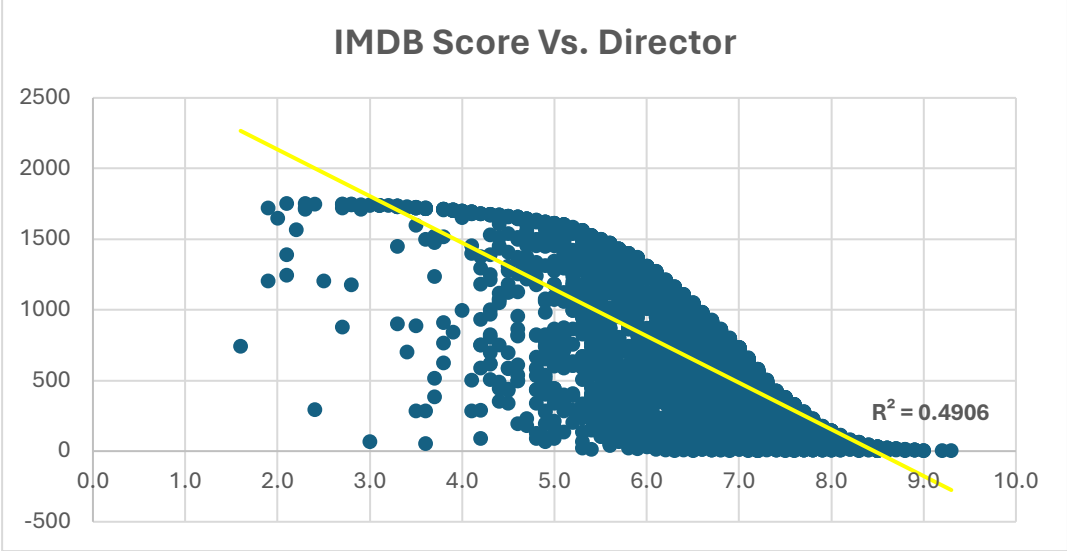
To extract useful insights from the data and enhance the understanding of what makes movies successful, first understand the data that has been provided. Next, Clean the data to make it feasible for analysis by removing duplicates, handling missing values and removing some of the unwanted columns that are present in the data sets.

To explore the data to understand the relation between different variables, done some data analysis using Excel and found out that IMDB rating is more correlated to durations and gross earnings than the other variables.





On the other hand, the correlation of the IMDB score with the director is strongly negative which means that directors who have made one or two movies tend to have high IMDB scores.



Tech-Stack Used:

Microsoft® Excel – Version 16.83 is used in this project as it is free and open-source. It provides a user-friendly interface. It offers a wide range of built-in mathematical and statistical functions. It has various types of visualisation to explore the data. It also has a Pivot Table which allows to summary and analysis of datasets by creating dynamic tables.

Insights:

Summarise the insights gained from analysing the data:

- From an understanding of the relationship between different variables, it is observed that the IMDB Movies Score is more correlated with duration and gross earnings than other variables, suggesting that longer-duration and higher-grossing movies are likely to receive higher ratings.
- It was noted from the genre distribution of movies that the majority of movies fall under the drama, comedy, and thriller genres, it was found that:

    - Drama, Crime, Biography, and Documentary genres tend to have higher average ratings and less variability, showing there are consistent positive responses from the audience.
    - Genres like Thriller, Action, Fantasy, Sci-Fi, and Family have moderate ratings and variability which suggests engaging a wider variety of audiences.
    - Musical, Romance, and Mystery genres show moderate to slightly lower ratings with varied audience perceptions.
    - Genres with higher mean ratings and stable reception may indicate areas of strength to focus on or genres to explore further for success.
    - Genres with lower mean ratings or higher variability may require deeper audience analysis, content adjustments, or targeted marketing approaches.
- From the analysis of movie duration distribution, it was observed that there is a wide range of movie durations, from relatively short movies to longer ones, with an average duration of around 109.91 minutes. By creating the scatter plot between the movie duration and IMDB Score, it indicates a positive correlation between them and the trend line is slightly going upward which means that there is a chance of getting a high score if the duration of the movie is longer.
- From Language Analysis, it was observed that the most common language used in movies is English followed by French and Spanish. By comparing the statistics of each language with the IMDB Score, it was found that:
    - English movies have a mean IMDB score of 6.42, with a standard deviation of 1.05, which shows a moderate difference in IMBD ratings.
    - The mean scores and standard deviations of other languages vary, indicating the diversity of audience opinions and ratings in various languages.
    - While English dominates in terms of the number of movies, it's interesting to note that even though languages like French, Spanish, Mandarin, and German with smaller amounts of movies have relatively higher mean IMDB scores, suggesting potentially positive responses or critical acclaim for movies in these languages.
    - Movies in languages like Persian, Portuguese, and Korean show relatively higher mean scores, indicating a positive response among audiences with these languages and cultures.
- From the Director Analysis, it was observed that Directors like Charles Chaplin, Tony Kaye, Ron Fricke, Majid Majidi, Damien Chazelle, Alfred Hitchcock, Sergio Leone, and Christopher Nolan have achieved a 100% percentile score which indicates that they are among the top directors in terms of their impact or success on IMDB ratings.
    - Directors like Charles Chaplin, Tony Kaye, Alfred Hitchcock, and Damien Chazelle with a small amount of movies as compared to other directors have an average IMDB score of 8.5 or higher which indicates their quality across the works.
    - Directors like Christopher Nolan, Sergio Leone, and Hayao Miyazaki have directed multiple movies with relatively high average IMDB scores which indicates a track record of successful and well-received movies by the audience. They consistently produce movies that are rated highly on IMDB, reflecting their reputation for quality filmmaking.

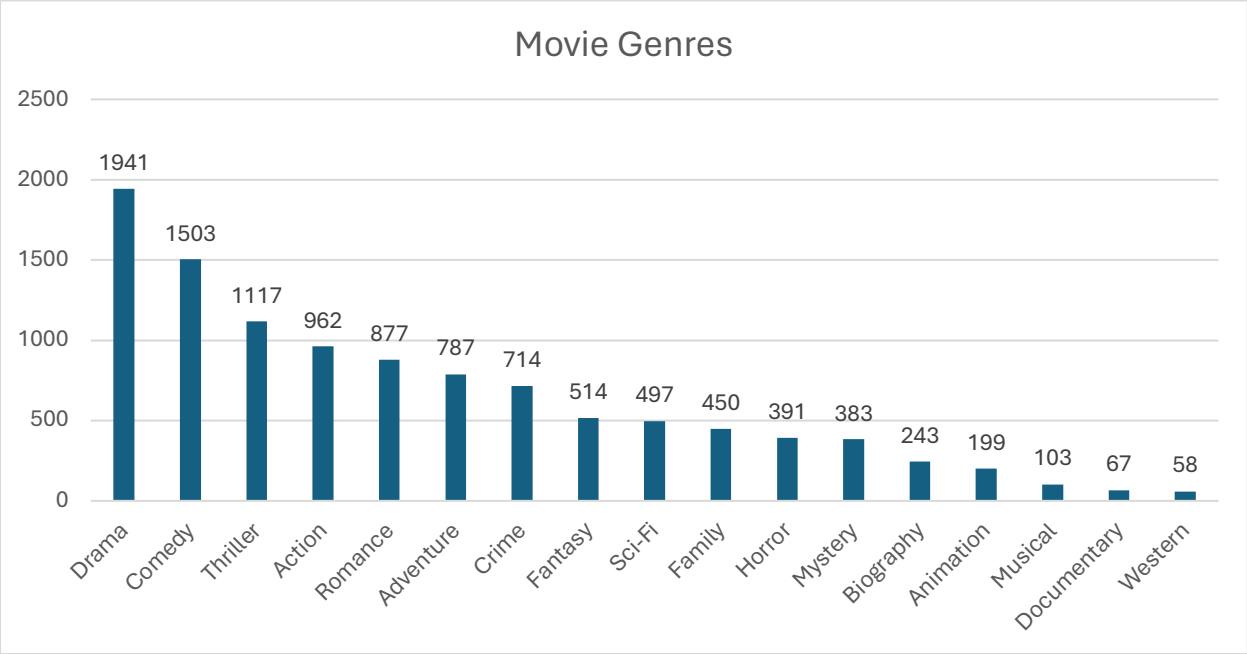
    - Directors with a percentile of 99% and above may have a strong potential for future success and continued high ratings for their upcoming projects.
- While exploring the relationship between movie budgets and their financial success, it was observed that the correlation coefficient between movie budgets and gross earnings is 0.10094171 which is a very weak positive correlation between movie budgets and gross earnings indicating that there is a tendency for higher movie budgets to be associated with higher gross earnings. However, due to almost no linear relationship between movie budgets and gross earnings, knowing the budget of a movie does not provide much information about its expected gross earnings, and vice versa.
    - This suggests that we cannot rely on the movie budget alone to predict or explain variations in gross earnings. Other factors likely play a more significant role in determining a movie's financial success.

The following are some insights from the data by answering the questions using Excel:

A. Movie Genre Analysis: Analyse 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.

⇒ Movies Count =COUNTIF('A. Movie Genre'!$C$2:$J$3852,L6)

Genre	Count of Movies
Drama	1941
Comedy	1503
Thriller	1117
Action	962
Romance	877
Adventure	787
Crime	714
Fantasy	514
Sci-Fi	497
Family	450
Horror	391
Mystery	383
Biography	243
Animation	199
Musical	103
Documentary	67
Western	58



➢ From the above distribution, Drama, Comedy and Thriller are the genres where most of the movies are made.

- **Descriptive statistics of the IMDB scores for each Genre.**

⇒ Minimum Value =MINIFS(Table1[imdb_score], Table1[genres1], L6)
⇒ Maximum Value =MAXIFS(Table1[imdb_score], Table1[genres1], L6)
⇒ Mean =AVERAGEIF(Table1[[genres1]:[genres8]], L6, Table1[imdb_score])
⇒ Median =MEDIAN(IF(Table1[[genres1]:[genres8]]=L6, Table1[imdb_score]))
⇒ Mode=MODE.SNGL(IF(Table1[[genres1]:[genres8]]=L6, Table1[imdb_score]))
⇒ Range =MAX(IF(Table1[[genres1]:[genres8]]=L6, Table1[imdb_score])) - MIN(IF(Table1[[genres1]:[genres8]]=L6, Table1[imdb_score]))
⇒ Standard Deviation = STDEV(IF(Table1[[genres1]:[genres8]]=L6, Table1[imdb_score]))

Genre	Min	Max	Range	Mean	Median	Mode	Variance	SD
Drama	2.1	8.8	7.2	6.8	6.9	6.7	0.8	0.9
Comedy	1.9	8.8	6.9	6.2	6.3	6.7	1.1	1.0
Thriller	4.8	6.3	6.3	5.3	6.4	6.5	0.9	1.0
Action	2.1	9.0	6.9	6.3	6.3	6.1	1.1	1.0
Romance	6.2	7.1	6.4	6.7	6.5	6.5	0.9	1.0
Adventure	2.3	8.6	6.6	6.6	6.6	6.7	1.2	1.1
Crime	3.3	9.3	6.9	6.9	6.6	6.6	1.0	1.0
Fantasy	4.3	7.9	6.7	6.3	6.4	6.7	1.3	1.1
Sci-Fi	5.0	8.2	6.9	6.6	6.4	6.7	1.3	1.2
Family	5.7	7.9	6.7	6.5	6.3	6.7	1.3	1.2
Horror	2.3	8.5	6.3	5.8	6.0	5.9	1.0	1.0
Mystery	3.3	8.5	5.5	6.7	6.5	6.6	1.0	1.0
Biography	4.5	8.9	4.4	7.2	7.2	7.0	0.5	0.7
Animation	4.5	8.0	5.8	6.8	6.8	6.7	1.0	1.0

Musical	6.3	7.2	6.4	6.8	6.7	7.1	1.3	1.1
Documentary	1.6	8.5	6.9	7.0	7.2	7.6	1.4	1.2
Western	4.1	8.9	4.8	6.8	6.8	6.8	1.0	1.0

- From the descriptive statistics, the following impact of the genres on the movie ratings are
  - From a range perspective, Drama and Comedy have wider ranges of IMDB Ratings as compared to Mystery or Western which shows diversity in audience perceptions within each genre.
  - Genres like Drama, Crime, Biography, Animation, Musical, Documentary and Western have relatively high average ratings which indicate that there is a positive response from the audience regarding these films.
  - Western genre stands out with an equal mean, median, and mode rating of 6.8, which indicates a stable response from the audience for this genre.
  - Genres like Thriller and Horror have slightly lower average IMDB Ratings than the other genres.
  - For Drama, Comedy, and Adventure, the mode is close to the median and mean ratings, indicating a concentration of ratings around these values.
  - Some genres, like Documentary and Musical, have modes that are slightly higher than the mean, suggesting a peak in higher ratings.
  - Documentary, Sci-Fi and Family are the genres where there is more variability in audience ratings than the genres like Mystery, Horror, and Western show.
  - Overall, Drama, Crime, Biography, and Documentary genres tend to have higher average ratings and lower variability, indicating consistent positive responses from the audience.

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

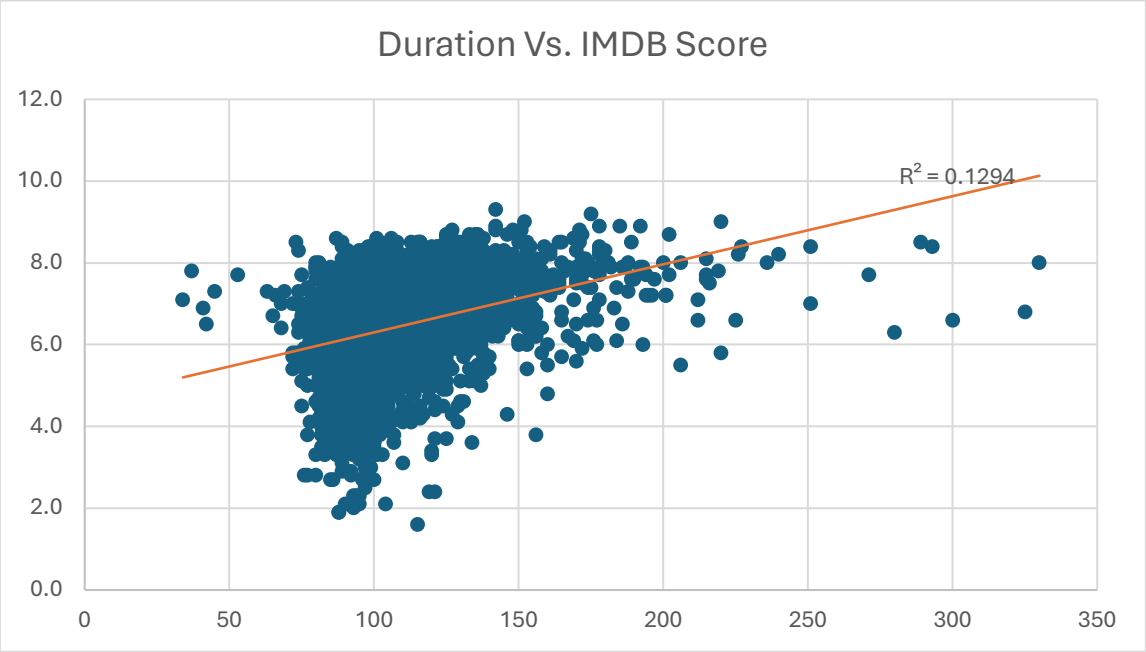
Task: Analyse the distribution of movie durations and identify the relationship between movie duration and IMDB score.

⇒ Descriptive Statistics of Duration:

Mean	109.91	=AVERAGE(B2:B3852)
Median	106.00	=MEDIAN(B2:B3852)
Mode	101.00	=MODE(B2:B3852)
Standard Deviation	22.75	=STDEV(B2:B3852)

- From Descriptive Statistics of Duration, it is depicted that the average duration of the movies is 109.91 minutes. Most of the movies have a duration of 101 minutes.
- The dispersion of movie durations around the mean is more due to a higher standard deviation of 22.75 which indicates greater variability.

- **Scatter plot to visualize the relationship between movie duration and IMDB score.**



- Most of the IMDB Scores lie in the duration between 80 – 150 minutes.
- As per the graph, the trendlines go slightly upward which means movie duration increases, with a tendency to increase in IMDB scores which implies that longer movies tend to receive higher ratings from viewers.

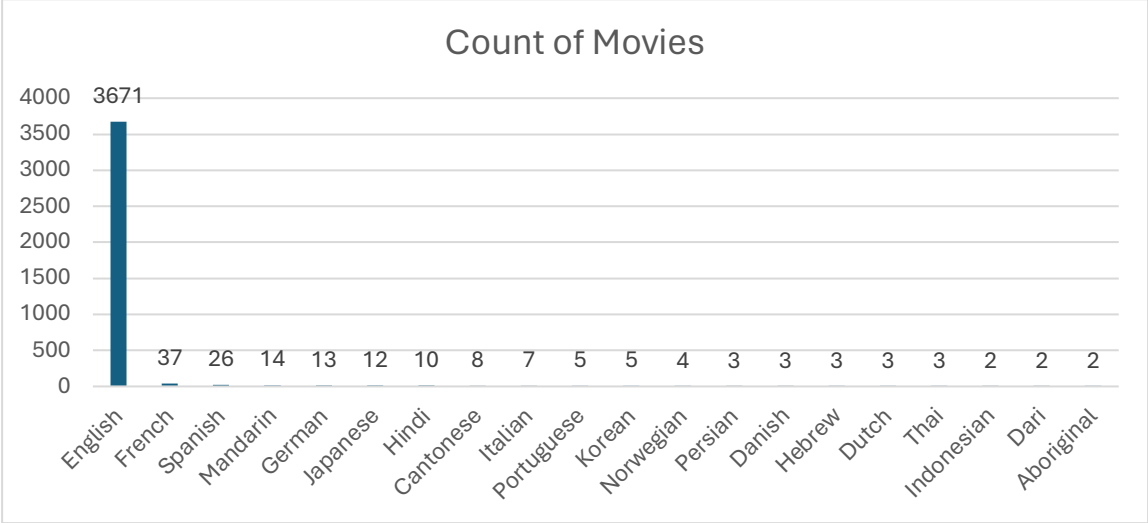
C. Language Analysis: Situation: Examine the distribution of movies based on their language.

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

⇒ =COUNTIF('C. Language'!$A$2:$B$3852, E3)

Language	Count of Movies
English	3671
French	37
Spanish	26
Mandarin	14
German	13
Japanese	12
Hindi	10
Cantonese	8
Italian	7
Portuguese	5
Korean	5
Norwegian	4
Persian	3
Danish	3
Hebrew	3
Dutch	3
Thai	3
Indonesian	2

Dari	2
Aboriginal	2



- The most commonly used language in the movies is English.
- English dominates the dataset with 3671 movies in total. This suggests that English-language movies have a significant presence in the global film industry and likely contribute significantly to the overall dataset's statistics.

- **Descriptive statistics of the IMDB scores for each Genre.**

Language	Mean	Median	Std. Dev.
English	6.42	6.50	1.05
French	7.29	7.20	0.56
Spanish	7.05	7.15	0.83
Mandarin	7.02	7.25	0.77
German	7.69	7.70	0.64
Japanese	7.63	7.80	0.90
Hindi	6.76	7.05	1.11
Cantonese	7.24	7.30	0.44
Italian	7.19	7.00	1.16
Portuguese	7.76	8.00	0.98
Korean	7.70	7.70	0.57
Norwegian	7.15	7.30	0.57
Persian	8.13	8.40	0.55
Danish	7.90	8.10	0.53
Hebrew	7.50	7.30	0.44
Dutch	7.57	7.80	0.40
Thai	6.63	6.60	0.45
Indonesian	7.90	7.90	0.42
Dari	7.50	7.50	0.14
Aboriginal	6.95	6.95	0.78

- Languages with higher mean and median scores, such as Persian, Portuguese, and Japanese, indicate that movies in these languages tend to receive positive ratings on average.

- On the other hand, languages like Aboriginal and Thai have a lower mean and median scores, indicating a mixed reception among viewers or a wider range of quality in movies from those languages.
- As per the variability of the IMDB Score, Languages with lower standard deviations, such as Dutch and Cantonese, show more consistent ratings among their movies.
- On the other hand, languages with higher standard deviations, such as Hindi and Italian, have a wider spread of ratings which indicates there are diverse opinions of the audience.

D. Director Analysis: Influence of directors on movie ratings.
Task: Identify the top directors based on their average IMDB score and analyse their contribution to the success of movies using percentile calculations.

- **Top 20 Directors as per Average IMDB Scores**

⇒ =AVERAGEIF($B$2:$B$3852, E3, $C$2:$C$3852)

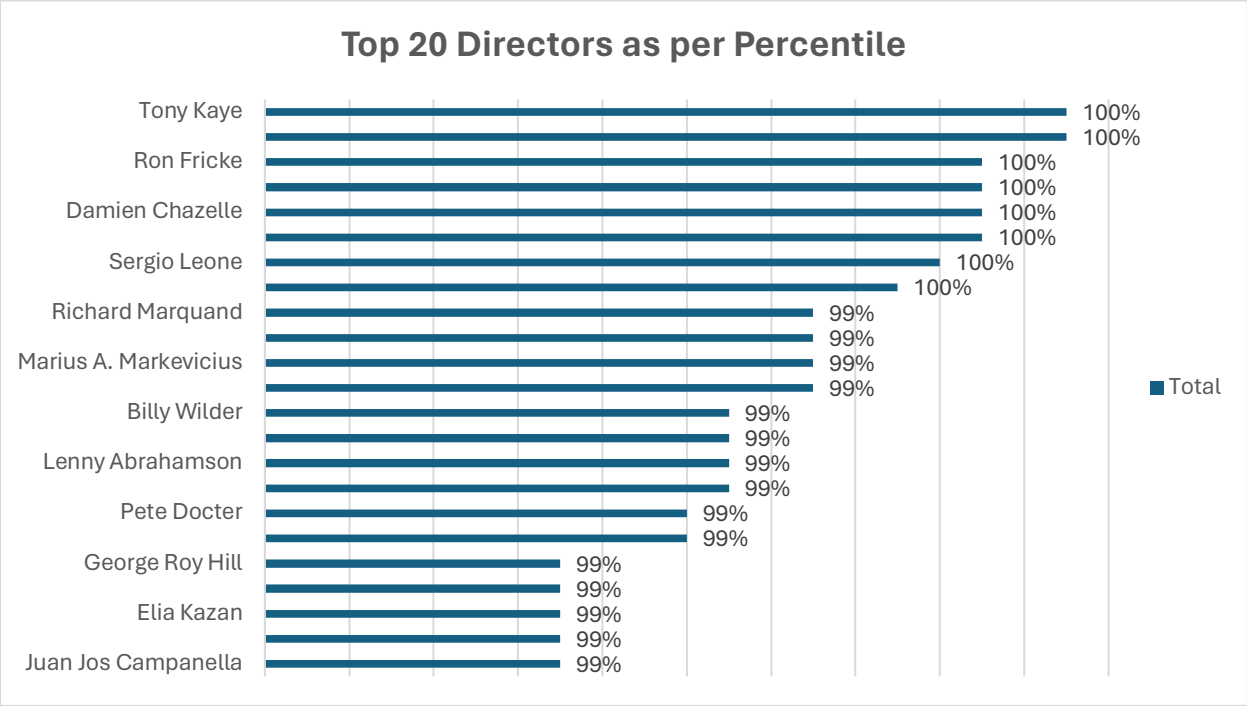
Directors	Avg. IMDB Score
Charles Chaplin	8.60
Tony Kaye	8.60
Ron Fricke	8.50
Majid Majidi	8.50
Damien Chazelle	8.50
Alfred Hitchcock	8.50
Sergio Leone	8.43
Christopher Nolan	8.43
S.S. Rajamouli	8.40
Marius A. Markevicius	8.40
Richard Marquand	8.40
Asghar Farhadi	8.40
Fritz Lang	8.30
Billy Wilder	8.30
Lee Unkrich	8.30
Lenny Abrahamson	8.30
Pete Docter	8.23
Hayao Miyazaki	8.23
George Roy Hill	8.20
Quentin Tarantino	8.20
Elia Kazan	8.20
Joshua Oppenheimer	8.20
Juan Jos Campanella	8.20

- Directors like Charles Chaplin and Tony Kaye have an average IMDB score of 8.60 even though they have made fewer movies than the other directors which indicates the quality of their work liked by the audience.

- On the other hand, Directors like Christopher Nolan, Sergio Leone, and Hayao Miyazaki have directed multiple movies with relatively high average IMDB scores, which indicates a track record of successful and well-received films by the audience.
- Directors like Christopher Nolan and Quentin Tarantino, with a larger count of movies and consistently high IMDB scores as compared to other directors, have established a strong reputation and popularity among viewers.
- Directors with fewer movies but high average scores, such as Hayao Miyazaki, have a dedicated fan base and are known for producing quality films.

- **Top 20 Directors as per Percentile**

Directors	Percentile of Directors
Charles Chaplin	100%
Tony Kaye	100%
Ron Fricke	100%
Majid Majidi	100%
Damien Chazelle	100%
Alfred Hitchcock	100%
Sergio Leone	100%
Christopher Nolan	100%
S.S. Rajamouli	99%
Marius A. Markevicius	99%
Richard Marquand	99%
Asghar Farhadi	99%
Fritz Lang	99%
Billy Wilder	99%
Lee Unkrich	99%
Lenny Abrahamson	99%
Hayao Miyazaki	99%
Pete Docter	99%
George Roy Hill	99%
Quentin Tarantino	99%
Elia Kazan	99%
Joshua Oppenheimer	99%
Juan Jos√© Campanella	99%



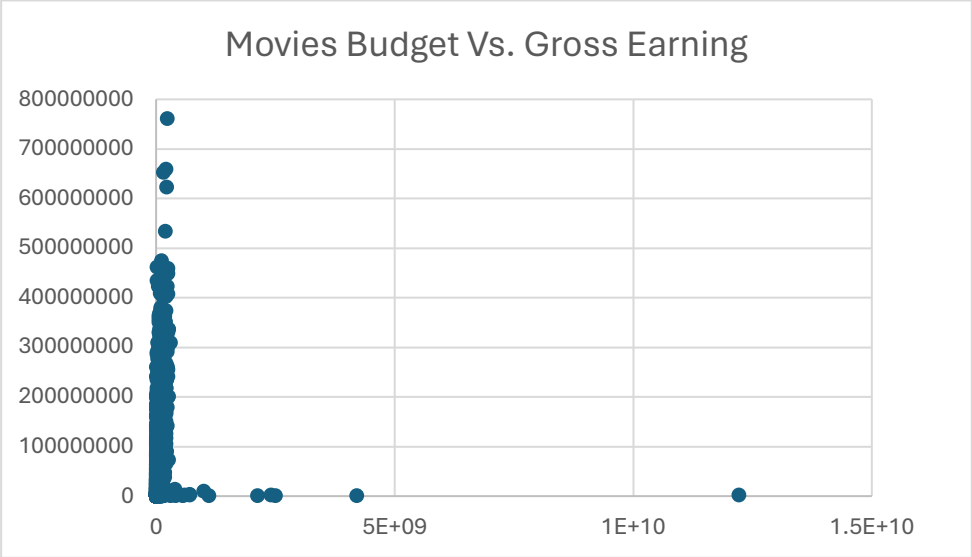
- Directors like Charles Chaplin, Tony Kaye, Ron Fricke, Majid Majidi, Damien Chazelle, and Alfred Hitchcock have achieved a 100% percentile, indicating consistently high performance in terms of their movies' IMDB scores. This suggests a track record of producing well-received films.
- Directors with a percentile of 99%, including Christopher Nolan, Asghar Farhadi, Hayao Miyazaki, Quentin Tarantino, and others, are highly acclaimed in the film industry. They consistently produce movies that are rated highly on IMDB, reflecting their reputation for quality filmmaking.

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

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

⇒ =CORREL(B2:B3852, C2:C3852)

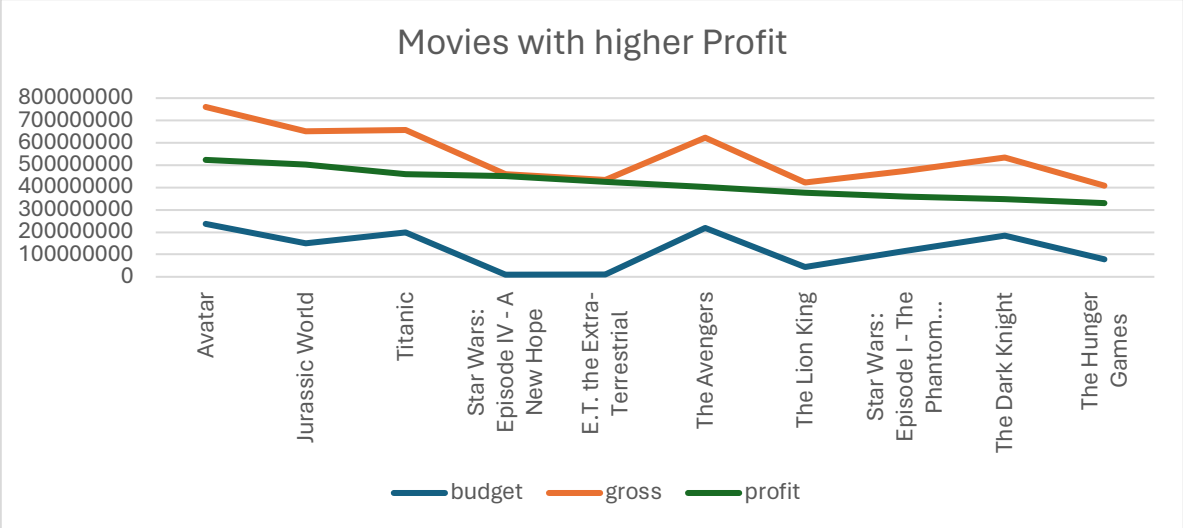
- The correlation coefficient between movie budgets and gross earnings is 0.1009417 (10.09%) which indicates a very weak positive correlation between movie budgets and gross earnings, it means that it's not necessary that higher-budget movies will have higher gross earnings.
- A weak correlation means that there is no linear relationship between movie budgets and gross earnings which indicates that the budget of a movie alone does not provide much information about its expected gross earnings, there are other factors that can also affect the gross earnings.



- **The top 10 movies with the highest profit margin:**

  - The Avatar movie has the highest profit margin of $523 million.

Movie Title	Budget	Gross	Profit
Avatar	237000000	760505847	523505847
Jurassic World	150000000	652177271	502177271
Titanic	200000000	658672302	458672302
Star Wars: Episode IV - A New Hope	11000000	460935665	449935665
E.T. the Extra-Terrestrial	10500000	434949459	424449459
The Avengers	220000000	623279547	403279547
The Lion King	45000000	422783777	377783777
Star Wars: Episode I - The Phantom Menace	115000000	474544677	359544677
The Dark Knight	185000000	533316061	348316061
The Hunger Games	78000000	407999255	329999255



  - Movies like Avatar, Titanic, and The Avengers had significantly high budgets, ranging from $200 million to $237 million.

- Despite their high budgets, movies like Avatar, Jurassic World, Titanic, and Star Wars: Episode IV - A New Hope made substantial profits, ranging from hundreds of millions to over half a billion dollars.
- Movies such as Avatar, Jurassic World, and Titanic achieved blockbuster success, generating massive gross earnings exceeding $650 million.

Results:

While doing this project, I have learned about Excel formulas and functions, Descriptive statistics, and Measure of variability using Excel, Pivot Tables and Data visualization using Excel to extract valuable insights from the data.

While analysing IMDB Movie Data, here are some of the insights derived from the data are as follows:

- By analysing the relation between different variables, I got to know which variable has more impact on making movies successful. Duration and Gross Earning do impact the IMDB Scores.
- By analysing the movie genres, I got to know that the majority of the movies are made in drama, comedy, and thriller genres. Understanding the ratings distribution across genres can inform marketing strategies, content development, and audience targeting for future movie projects.
- I learned that there is a significant impact of Movie Duration on IMDB Scores as it is positively correlated.
- By analysing the data, I got to know that most of the common language used in the movie is English. Other languages like French, Spanish, Mandarin, and German have relatively higher mean IMDB scores, suggesting potentially positive reception or critical acclaim for movies in these languages.
- By analysing Director's contributions to the success of movies, it is found that Directors like Charles Chaplin, Tony Kaye, Ron Fricke, Majid Majidi, Damien Chazelle, Alfred Hitchcock, Sergio Leone, and Christopher Nolan are the top directors that impacted the success of IMDB ratings with 100% percentile score.
- It is observed that there is weak positive relationship of Budget with gross earning which means that they can not necessarily affect each other. And the movies like Avatar, Jurassic World, Titanic, and Avengers made substantial profits despite of having high budget.

Overall, this project helps me to derive some valuable insights and meaningful conclusions that will help the movie producers, directors, and investors to make informed decisions in their future projects to make a successful movies.

The link for the Excel sheet is below:
https://docs.google.com/spreadsheets/d/1pTy68W72htnq098-KIgCwLRfS1llPn2y/edit?usp=sharing&ouid=111731928816147649891&rtpof=true&sd=true