IMDB Movie Analysis

Final Project-1

- **Project Description:** The provided dataset pertains to IMDb movies, offering an opportunity to investigate a key question: **"What factors influence the success of a movie on IMDb?"** Success, in this context, is defined by high IMDb ratings. Understanding these factors is crucial for movie producers, directors, and investors who aim to make informed decisions in their future projects. Analysing these elements can provide valuable insights into what contributes to a movie's success, ultimately aiding in the creation of more critically acclaimed and commercially successful films.

- **Approach:** Performing IMDB movie analysis using Microsoft Excel involves several steps, including data collection, cleaning, analysis, visualization, and reporting. Here is a detailed approach:

1. Data Collection and Preparation

- **Collect Data:**
  - Downloaded the IMDb dataset provided which contained 28 number of features and 5043 data points. The names of features were color, movie titles, ratings, genres, director name, duration, actor name, budgets, gross, etc.
- **Clean the Data:**
  - Handle Duplicate values: Found 45 rows where all column values were duplicate. Keeping the first occurrence of each duplicate, dropped rest of the duplicates.
  - Handle missing values: Checked for null values and dropped all the rows which contained any cells which were missing.

2. Exploratory Data Analysis (EDA)

- **Understand the Data:**
  - Use descriptive statistics to summarize the data.
  - Visualize distributions of key variables (e.g., IMDB ratings, budgets).
  - Identify and visualize relationships between variables (e.g., scatter plots, correlation matrices).
- **Identify Key Features:**
  - Determine which features (e.g., genres, budget, actors, directors) might influence IMDb ratings.

3. Feature Engineering

- **Create New Features:** Separated the genres in **genres** using **pipe** (**|**) as separator and then deleted the **genres** column. It created 7 genres named as genres.1, generes.2 etc.

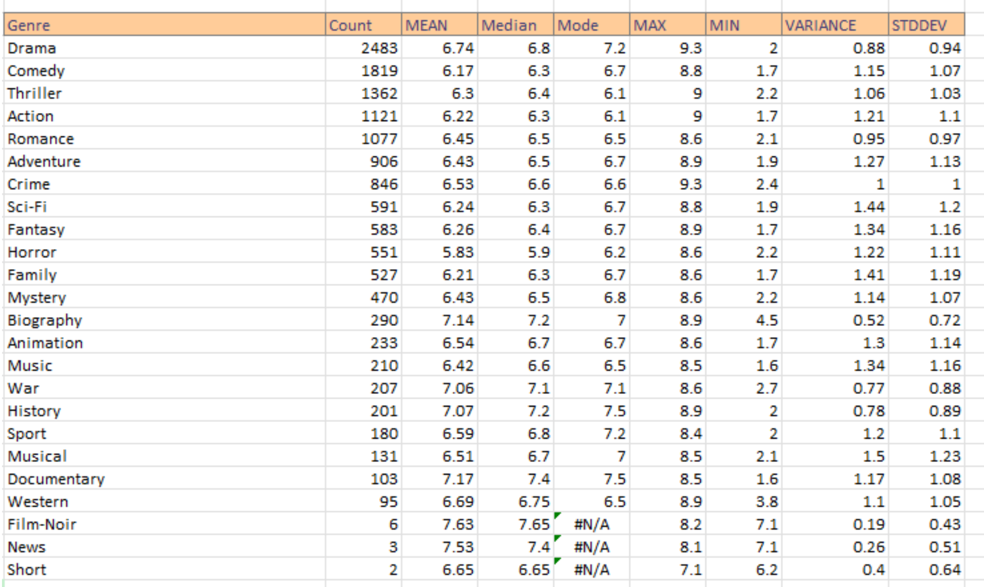
Tech Stack Used

1. **Microsoft Excel 2021** — A spreadsheet editor software used mainly by professionals to enter data in table format, perform computations, plot graphs etc. Here Microsoft Excel is used to filter data and plot graphs to get insights about the movies.
2. **Microsoft Word:** A word processing application for preparing report

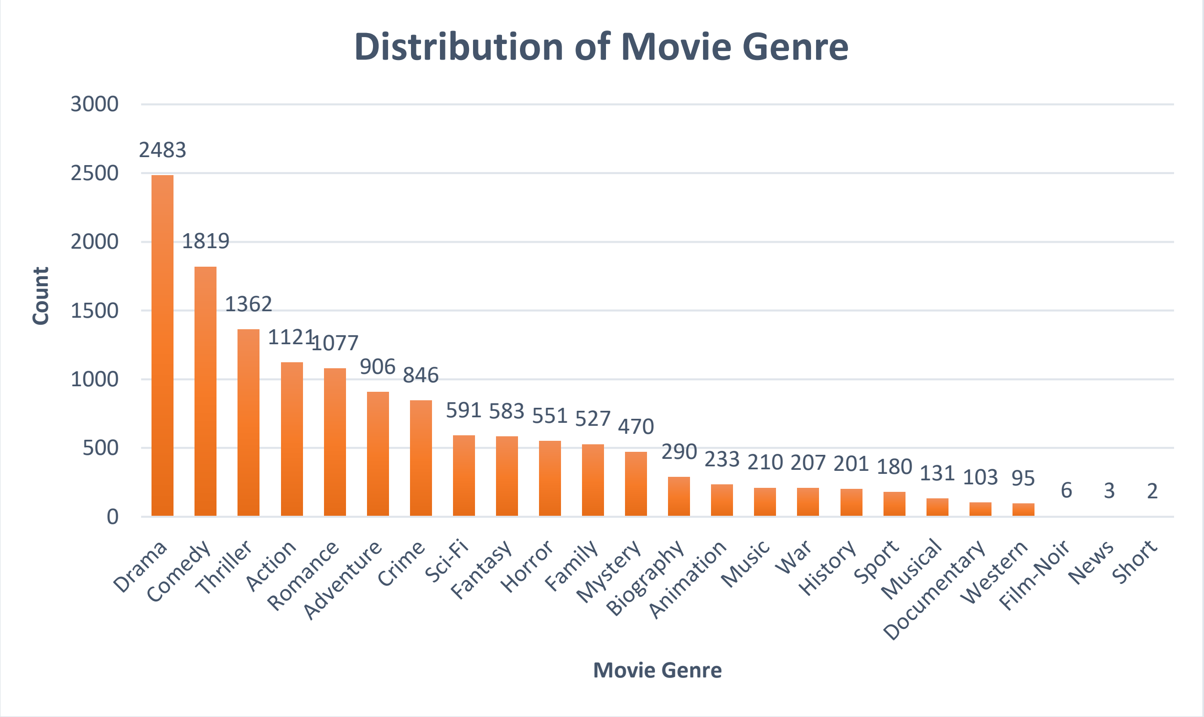
DATA ANALYTICS TASKS:

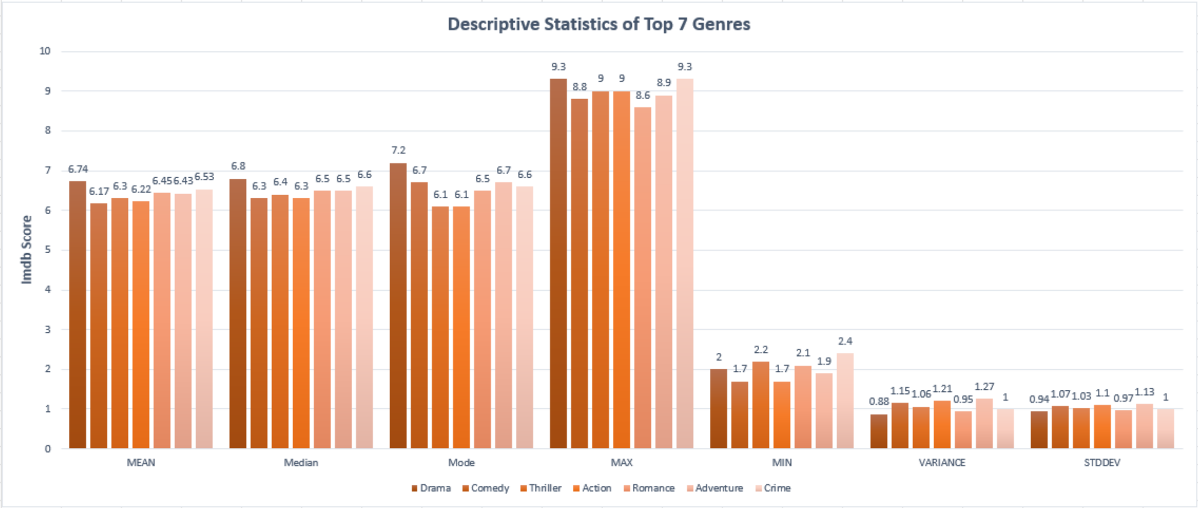
A. **Movie Genre Analysis:** Analyse the distribution of movie genres and their impact on the IMDB score.

- **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.
- **Result:** The top 7 most common genres are Drama, Comedy, Thriller, Action, Romance, Adventure and Crime. Also, all the top 7 genres descriptive statistics (mean, median, mode, std. dev. Variance) are almost at same level.



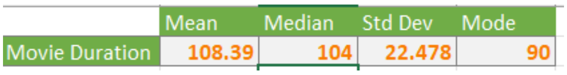
Genre	Count	MEAN	Median	Mode	MAX	MIN	VARIANCE	STDDEV
Drama	2483	6.74	6.8	7.2	9.3	2	0.88	0.94
Comedy	1819	6.17	6.3	6.7	8.8	1.7	1.15	1.07
Thriller	1362	6.3	6.4	6.1	9	2.2	1.06	1.03
Action	1121	6.22	6.3	6.1	9	1.7	1.21	1.1
Romance	1077	6.45	6.5	6.5	8.6	2.1	0.95	0.97
Adventure	906	6.43	6.5	6.7	8.9	1.9	1.27	1.13
Crime	846	6.53	6.6	6.6	9.3	2.4	1	1
Sci-Fi	591	6.24	6.3	6.7	8.8	1.9	1.44	1.2
Fantasy	583	6.26	6.4	6.7	8.9	1.7	1.34	1.16
Horror	551	5.83	5.9	6.2	8.6	2.2	1.22	1.11
Family	527	6.21	6.3	6.7	8.6	1.7	1.41	1.19
Mystery	470	6.43	6.5	6.8	8.6	2.2	1.14	1.07
Biography	290	7.14	7.2	7	8.9	4.5	0.52	0.72
Animation	233	6.54	6.7	6.7	8.6	1.7	1.3	1.14
Music	210	6.42	6.6	6.5	8.5	1.6	1.34	1.16
War	207	7.06	7.1	7.1	8.6	2.7	0.77	0.88
History	201	7.07	7.2	7.5	8.9	2	0.78	0.89
Sport	180	6.59	6.8	7.2	8.4	2	1.2	1.1
Musical	131	6.51	6.7	7	8.5	2.1	1.5	1.23
Documentary	103	7.17	7.4	7.5	8.5	1.6	1.17	1.08
Western	95	6.69	6.75	6.5	8.9	3.8	1.1	1.05
Film-Noir	6	7.63	7.65	#N/A	8.2	7.1	0.19	0.43
News	3	7.53	7.4	#N/A	8.1	7.1	0.26	0.51
Short	2	6.65	6.65	#N/A	7.1	6.2	0.4	0.64



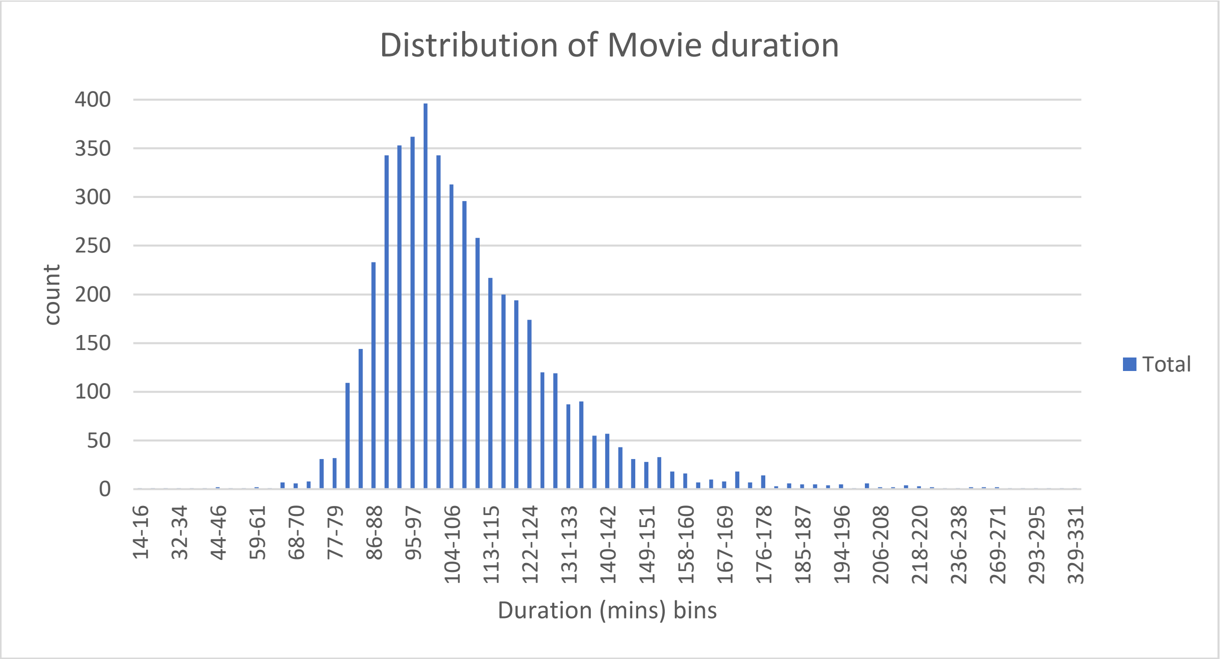


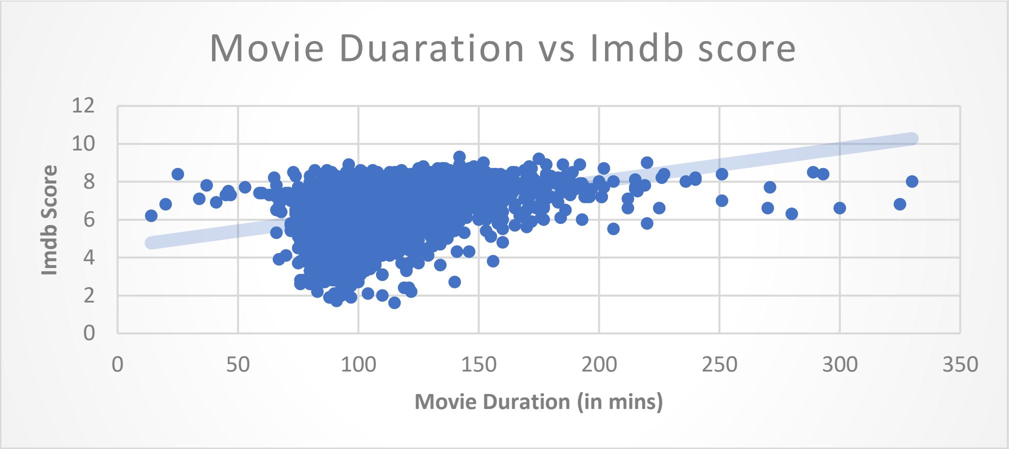
B. Movie Duration Analysis: Analyse the distribution of movie durations and its impact on the IMDB score.

- **Task:** Analyse the distribution of movie durations and identify the relationship between movie duration and IMDB score.
- **Result:** The distribution of Movie Durations shows that it closely follows a Normal Distribution. Also, the scatter plot shows that duration and imdb_scores have a positive relationship.



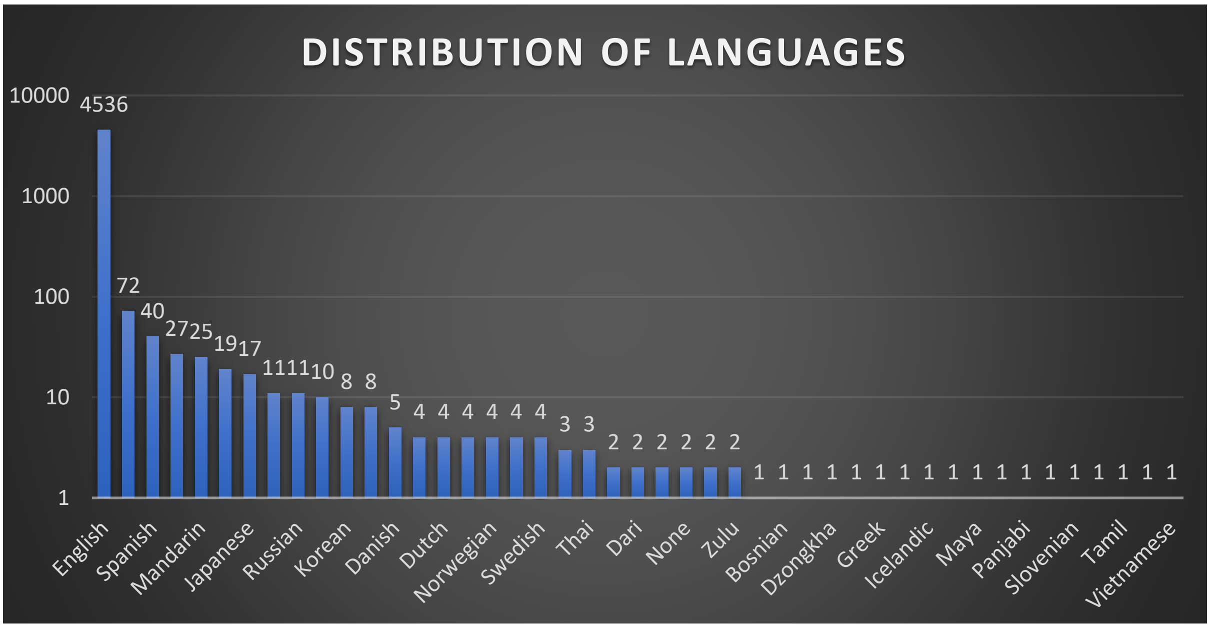
	Mean	Median	Std Dev	Mode
Movie Duration	108.39	104	22.478	90

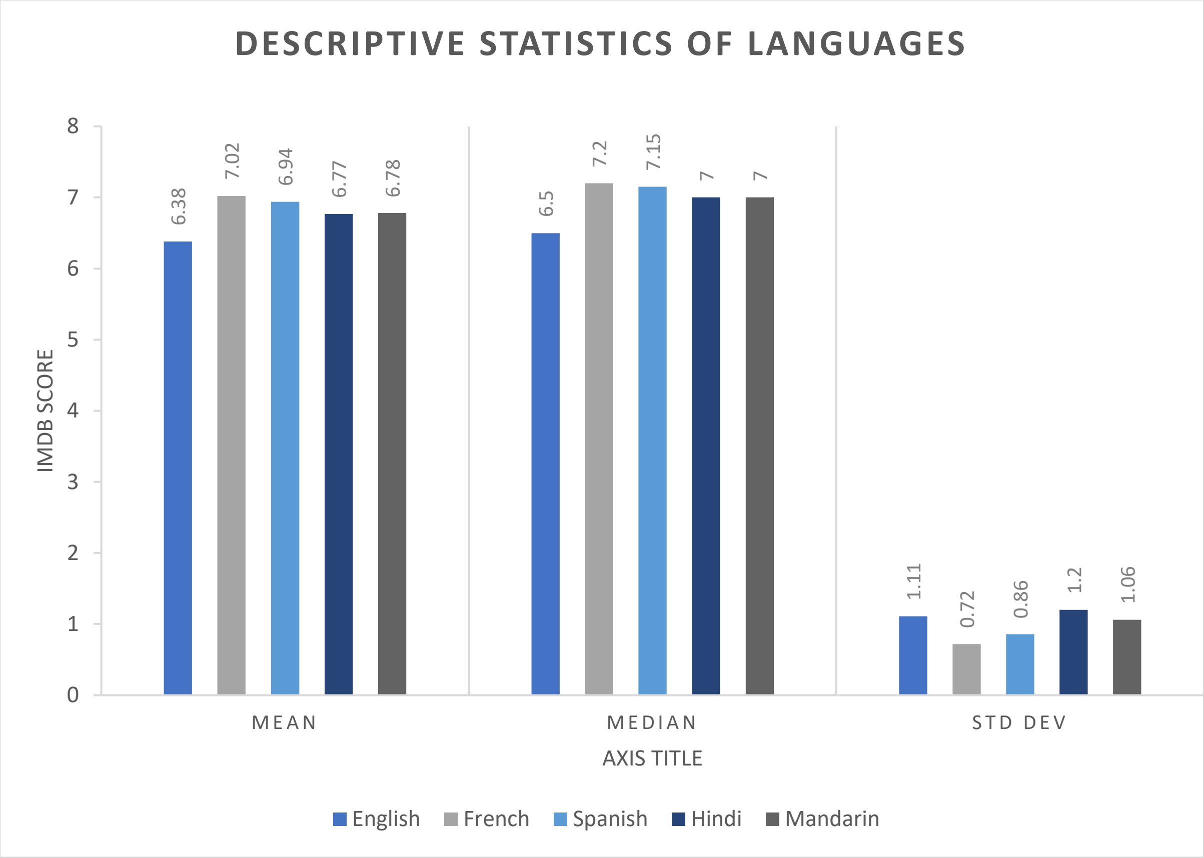




C. Language Analysis: Situation: Examine the distribution of movies based on their language.

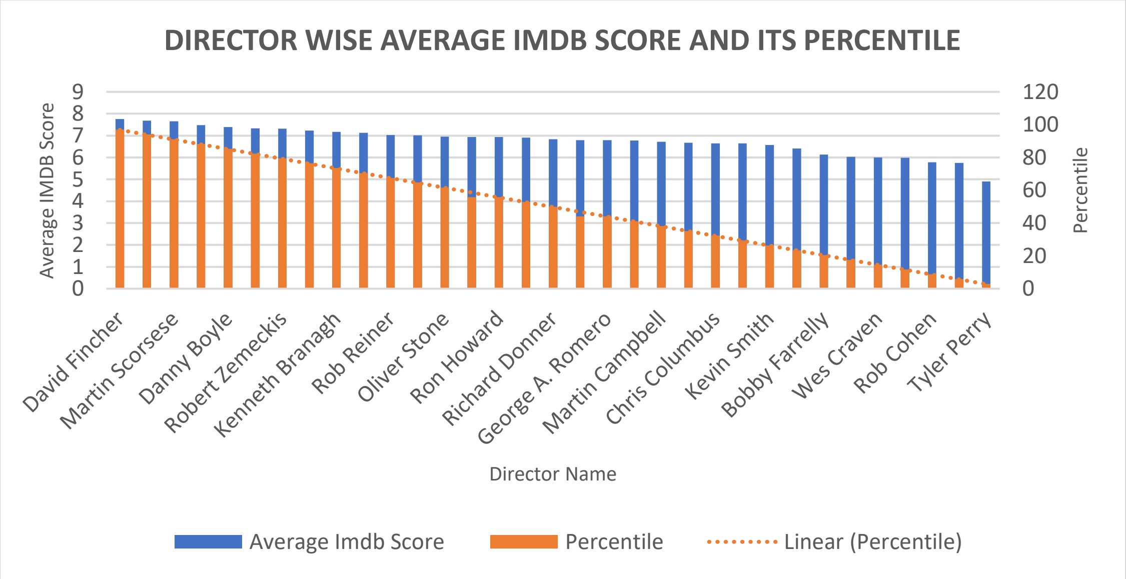
- **Task**: Determine the most common languages used in movies and analyse their impact on the IMDB score using descriptive statistics.
- **Result:** The first plot below shows that **English** is the most common language used in movies followed by **French, Spanish, Hindi and Mandarin**. The second plot below shows that French language has comparatively higher mean and median but lower standard deviation implying that most of the French language movies have their imdb score on the higher side.



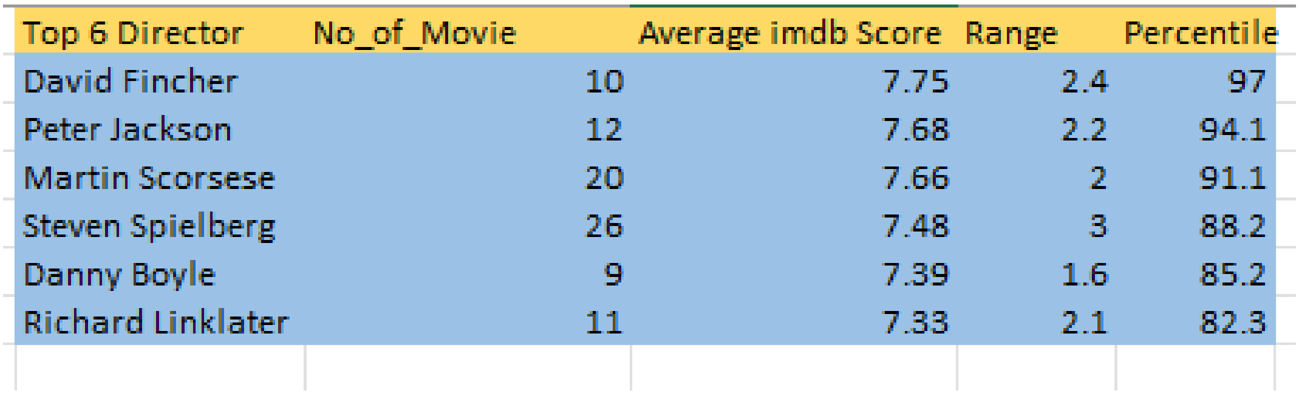


D. Director Analysis: Influence of directors on movie ratings.

- **Task:** Identify the top directors based on their average IMDB score and analyse their contribution to the success of movies using percentile calculations.
- **Result:** The plot considers only those directors whose movie counts are more than or equal to 9 and the range of IMDB Scores is less than equal to 3 as otherwise it would be unfair for those who has maintained consistently high scores for large number of movies to be compared for top directors to those who has performed well in few movies**.**



- **Result:** For top directors, only the top 6 directors are considered as there is a drop in percentile after the first 6 directors. The average **IMDB** Scores are between **7** and **8** for the top directors with the above condition. Also, their percentile score is above **80%.**



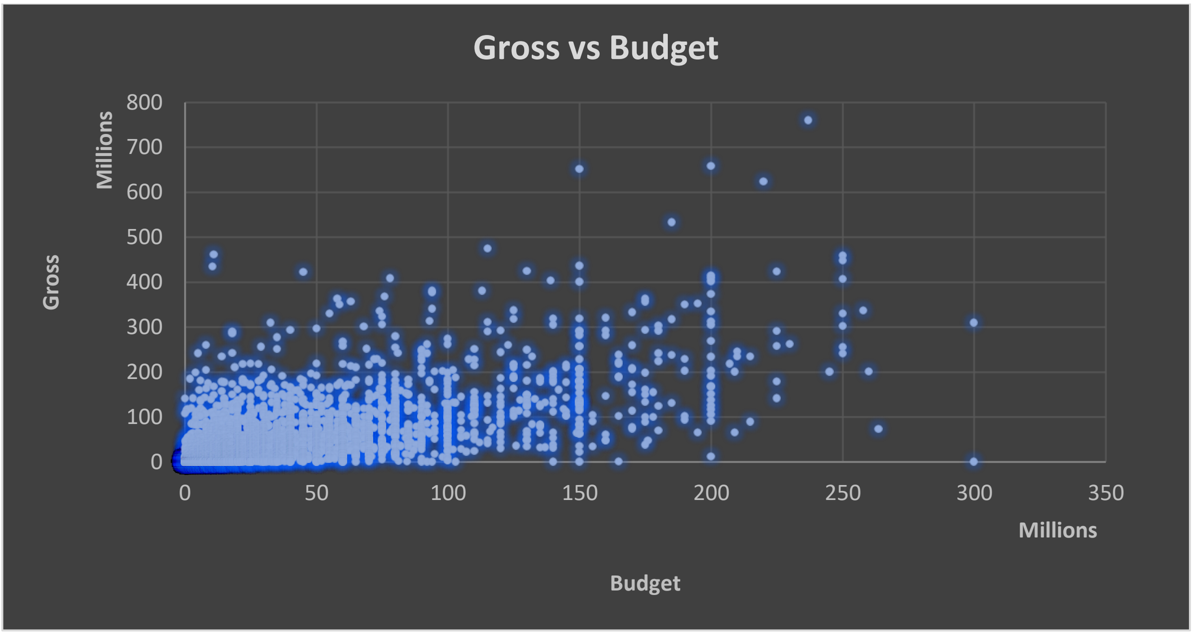
Top 6 Director	No_of_Movie	Average imdb Score	Range	Percentile
David Fincher	10	7.75	2.4	97
Peter Jackson	12	7.68	2.2	94.1
Martin Scorsese	20	7.66	2	91.1
Steven Spielberg	26	7.48	3	88.2
Danny Boyle	9	7.39	1.6	85.2
Richard Linklater	11	7.33	2.1	82.3

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**.**
- **Result:** The table shows that the correlation between **Gross** and **Budget** is **positive** and more than **0.5**. That is, the relationship shows that as budget of movies increase, there is a very high probability that the gross collection of the movie will also increase. The plot shows the relationship between **Gross** and **Budget**. The overall trendline has a slope close to **1**.

Correlation between gross and budget
0.774133278

Movies with Highest Profit Margin	
Movie Title	Margin
Avatar	523505847
Jurassic World	502177271
Titanic	458672302
Star Wars: Episode IV - A New Hope	449935665
E.T. the Extra-Terrestrial	424449459
The Avengers	403279547



- **Conclusion:**

This project highlighted the significance of data analytics in movie analysis, revealing crucial insights such as the correlation between directors and IMDb scores, the impact of genres on IMDb ratings, and the relationship between budgets and IMDb scores. These insights are invaluable for making informed, data-driven decisions in the film industry.

- LINK To MS Excel: https://docs.google.com/spreadsheets/d/1HpF5Rg-eXAjrSu1qEHX84f6QEvWNqJQV/edit?usp=sharing&ouid=104742351045324653369&rtpof=true&sd=true