IMDB Movie Analysis

Project Description:

The purpose of this project is to investigate factors which influences the success of movies on IMDB, which is defined by high IMDB ratings. The analysis will help in deriving actionable insights for movie producers, directors, and investors to make informed decisions for future projects.
Through detailed analysis and visualisation and utilizing Excel tools on the provided dataset, the following factors can help in providing comprehensive understanding of movie success on IMDB:

1. **Movie Genre Analysis:** Analyse the distribution of movie genres and their impact on the IMDB score.
2. **Movie Duration Analysis:** Analyse the distribution of movie durations and its impact on the IMDB score.
3. **Language Analysis:** Examine the distribution of movies based on their language.
4. **Director Analysis:** Influence of directors on movie ratings.
5. **Budget Analysis:** Explore the relationship between movie budgets and their financial success.

Approach:

To accomplish the necessary tasks and to finalize the project, the approach to this project involves a structured methodology to analyse the factors influencing the success of movies on IMDB. The following steps were executed:

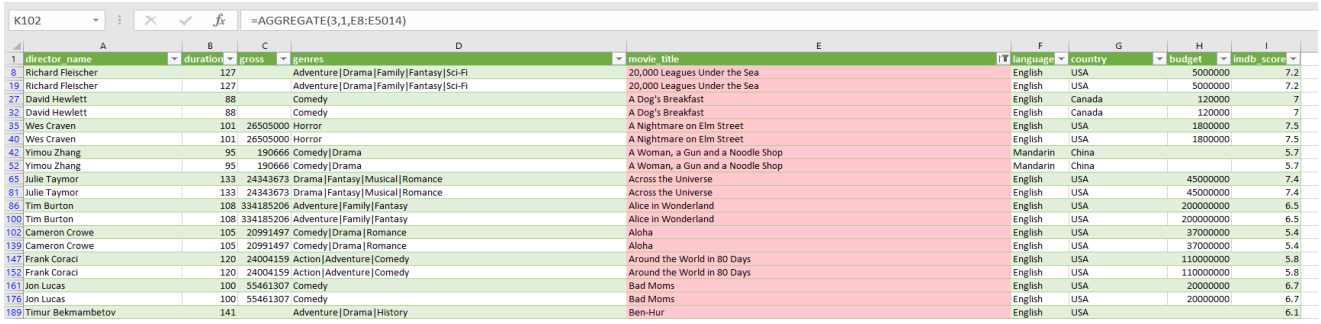
1. **Understanding the requirements of data:** Firstly, the provided dataset was imported into Excel. The raw data was converted into tabular form to handle the data efficiently for analysis.
   Then the structure of the data was observed. The provided dataset is related to IMDB movies and contains records of movies from previous years and different demographic locations. The dataset has the following details:
   - Number of data points: 5043
   - Number of columns: 28

   Each column contains information like director name, actor names, genres, plot keywords, budget, gross collection, IMDB score, etc.

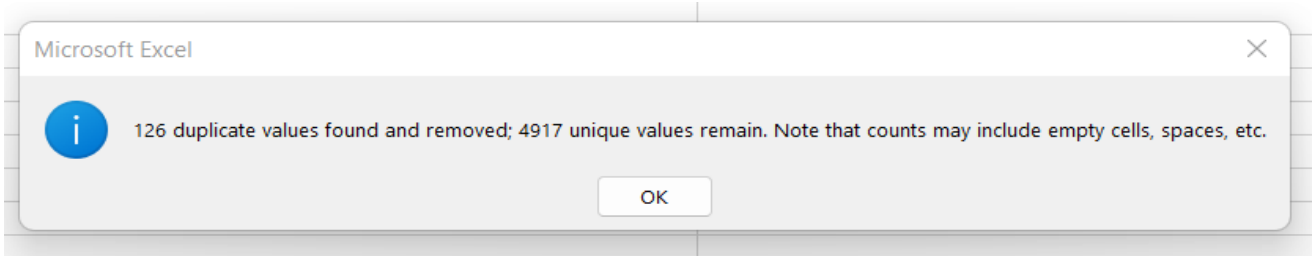
2. **Data Cleaning and Preparation:** To summarize the findings and derive actionable insights, it is crucial to handle them appropriately.
   - **Deleting irrelevant columns:** Based on the objective of the project, certain irrelevant columns were removed which will not provide any valuable insights for our analysis. After removing irrelevant columns there are total 9 columns which will be used for analysis.
     1. director_name: Name of director of the movie.
     2. duration: Duration of the movie.

3. gross: Gross collection of the movie.
4. genres: Genres of the movie.
5. movie_title: Name of the movie.
6. language: Language of the movie.
7. country: Country in which movie was made.
8. budget: Budget of the movie.
9. imdb_score: IMDB score of the movie.

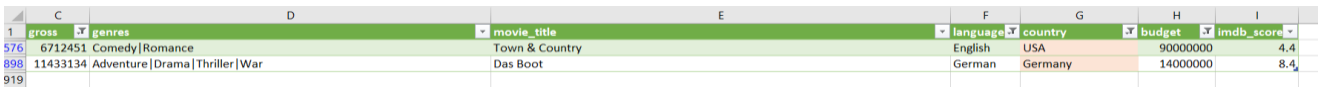
- **Handling missing data, duplicate data and errors:** After deleting irrelevant columns, the datasets will be checked for missing values, duplicate values and errors and the best strategy to handle and rectify them will be determined.
  In the dataset, column name movie_title has 245 duplicates values hence, it needs to be removed otherwise it will create discrepancy in the statistical calculations. Here are few rows highlighting the duplicate values.



	director_name	duration	gross	genres	movie_title	language	country	budget	imdb_score
8	Richard Fleischer	127		Adventure|Drama|Family|Fantasy|Sci-Fi	20,000 Leagues Under the Sea	English	USA	5000000	7.2
19	Richard Fleischer	127		Adventure|Drama|Family|Fantasy|Sci-Fi	20,000 Leagues Under the Sea	English	USA	5000000	7.2
27	David Hewlett	88		Comedy	A Dog's Breakfast	English	Canada	120000	7
32	David Hewlett	88		Comedy	A Dog's Breakfast	English	Canada	120000	7
35	Wes Craven	101	26505000	Horror	A Nightmare on Elm Street	English	USA	1800000	7.5
40	Wes Craven	101	26505000	Horror	A Nightmare on Elm Street	English	USA	1800000	7.5
42	Yimou Zhang	95	190666	Comedy|Drama	A Woman, a Gun and a Noodle Shop	Mandarin	China		5.7
52	Yimou Zhang	95	190666	Comedy|Drama	A Woman, a Gun and a Noodle Shop	Mandarin	China		5.7
65	Julie Taymor	133	24343673	Drama|Fantasy|Musical|Romance	Across the Universe	English	USA	45000000	7.4
81	Julie Taymor	133	24343673	Drama|Fantasy|Musical|Romance	Across the Universe	English	USA	45000000	7.4
86	Tim Burton	108	334185206	Adventure|Family|Fantasy	Alice in Wonderland	English	USA	200000000	6.5
100	Tim Burton	108	334185206	Adventure|Family|Fantasy	Alice in Wonderland	English	USA	200000000	6.5
102	Cameron Crowe	105	20991497	Comedy|Drama|Romance	Aloha	English	USA	37000000	5.4
139	Cameron Crowe	105	20991497	Comedy|Drama|Romance	Aloha	English	USA	37000000	5.4
147	Frank Coraci	120	24004159	Action|Adventure|Comedy	Around the World in 80 Days	English	USA	110000000	5.8
152	Frank Coraci	120	24004159	Action|Adventure|Comedy	Around the World in 80 Days	English	USA	110000000	5.8
161	Jon Lucas	100	55461307	Comedy	Bad Moms	English	USA	20000000	6.7
176	Jon Lucas	100	55461307	Comedy	Bad Moms	English	USA	20000000	6.7
189	Timur Bekmambetov	141		Adventure|Drama|History	Ben-Hur	English	USA		6.1



In the column country, New Line was replaced with USA and West Germany was replaced with Germany.



	gross	genres	movie_title	language	country	budget	imdb_score
576	6712451	Comedy|Romance	Town & Country	English	USA	90000000	4.4
898	11433134	Adventure|Drama|Thriller|War	Das Boot	German	Germany	14000000	8.4
919							

Since there are multiple blank cells in multiple columns, to deal with them all the blank cells have been removed by filtering them out.

- **Clubbing and Separating columns:** Columns with multiple categories that can be combined will be grouped together whereas the columns with

multiple entries will be separated. Our dataset doesn't have columns that needs to be clubbed.
In the column genres there are multiple entries for genres for a single movie, hence separated them into individual genres by using Excel's text to column feature based on delimiter.
There is total 8 genres which are separated into new columns.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	director_name	duration	gross	genre_1	genre_2	genre_3	genre_4	genre_5	genre_6	genre_7	genre_8	movie_title	language
36	Shane Acker	79	31743332	Action	Adventure	Animation	Drama	Mystery	Sci-Fi	Thriller		9	English
38	Mark L. Lester	79	2275557	Action	Comedy	Crime	Thriller					Showdown in Little Tokyo	English
40	Yimou Zhang	80	84961	Action	Adventure	History						Hero	Mandarin
43	Cory Edwards	80	51053787	Action	Animation	Comedy	Crime	Family				Hoodwinked!	English
63	Richard Williams	80	669276	Action	Adventure	Animation	Comedy	Fantasy				The Princess and the Cobbler	English
76	Robert Rodriguez	81	2040920	Action	Crime	Drama	Romance	Thriller				El Mariachi	Spanish
110	Jimmy Hayward	81	10539414	Action	Drama	Fantasy	Thriller	Western				Jonah Hex	English
113	Steve Oedekerk	81	16033556	Action	Comedy							Kung Pow: Enter the Fist	English
117	Louis C.K.	81	3293258	Action	Adventure	Comedy	Musical					Pootie Tang	English
121	Rick Friedberg	81	26906039	Action	Comedy							Spy Hard	English
123	Catherine Gund	82	21199	Action	Biography	Documentary	Sport					Born to Fly: Elizabeth Streb vs. Gravity	English
124	Brad Peyton	82	43575716	Action	Comedy	Family	Fantasy					Cats & Dogs: The Revenge of Kitty Galore	English
128	Mamoru Hosoda	82	9628751	Action	Adventure	Animation	Family	Sci-Fi				Digimon: The Movie	English
129	John A. Davis	82	80920948	Action	Adventure	Animation	Comedy	Family	Sci-Fi			Jimmy Neutron: Boy Genius	English

- **Creating a new column:** The analysis of this project requires creating a column named Profit Margin (gross earning- budget). Hence this column was created.

O	P
budget	Profit margin
30000000	1743332
8000000	-5724443
31000000	-30915039
17500000	33553787
28000000	-27330724
7000	2033920
47000000	-36460586
10000000	6033556
3000000	293258
18000000	8906039
500000	-478801
85000000	-41424284
5000000	4628751
30000000	50920948
27000000	4768374
35000000	-9128166

3. **Data Summary:** After cleaning and preparing the data, statistical measures will be calculated like mean, median, mode, standard deviation, etc. and visualization will be done to derive key insights.

▪ Tech-Stack used

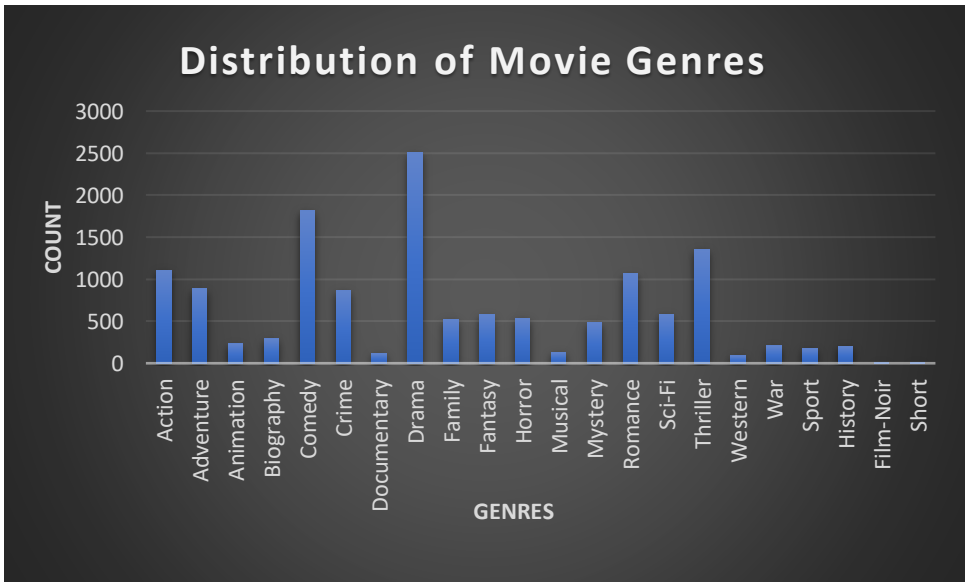
Microsoft Excel 2019 has been used to clean and prepare the dataset for analysis. It was also used for basic statistical calculations. Pivot tables were created, Excel charts were utilised for visualisation.

The reason behind using excel was that is widely used and easy to understand. It is great tool for analysing data for large datasets and works well with other tools like Word and PowerPoint, so it is easy to create reports and presentation based on the analysis. It is cost effective as it is often installed in our computers.

- **Insights:**

1. **Movie Genre Analysis:** The task is to determine the most common genres of movies in the dataset and for each genre calculate descriptive statistics (mean, median, mode, range, variance, standard deviation) of the IMDB scores.

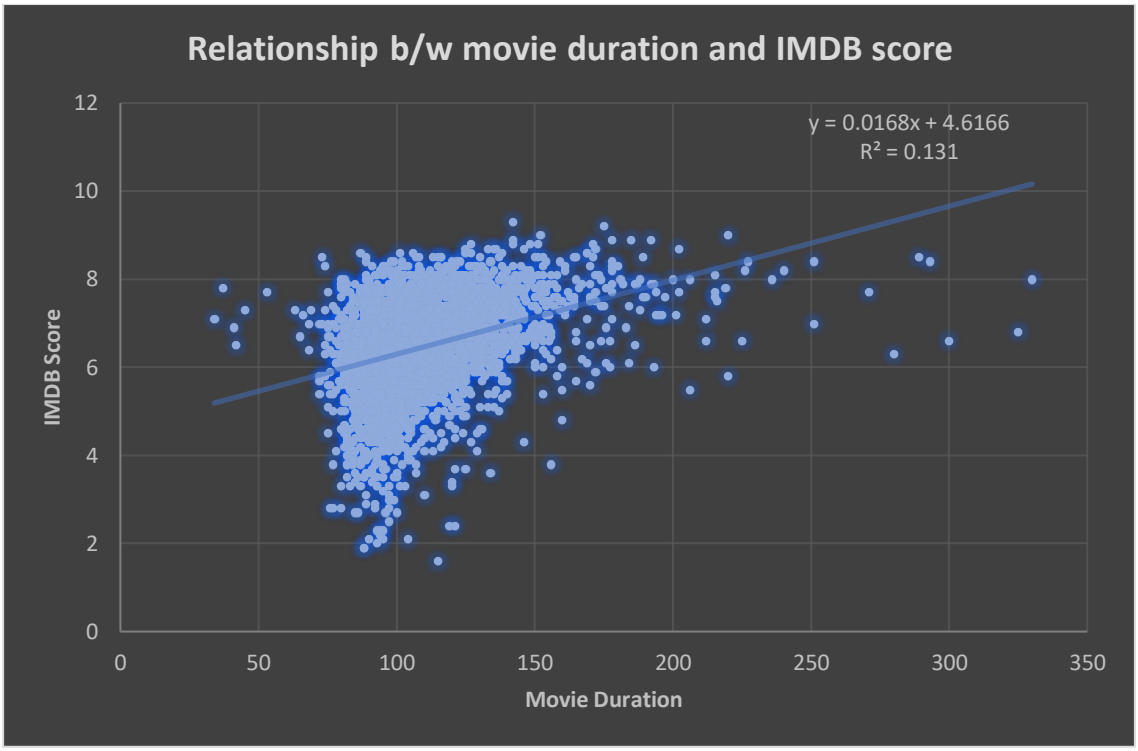
Genres	Count	Average(Mean)	Median	Mode	Max	Min	Range	Variance	Standard Deviation
Action	1107	6.23	6.3	6.1	9.1	1.7	7.4	1.24	1.11
Adventure	884	6.53	6.6	6.7	8.9	1.9	7	1.29	1.14
Animation	231	6.51	6.7	6.7	8.6	1.7	6.9	1.29	1.14
Biography	290	7.16	7.2	7	8.9	4.5	4.4	0.52	0.72
Comedy	1813	6.17	6.3	6.3	9.5	1.7	7.8	1.16	1.08
Crime	862	6.91	6.6	6.3	9.3	2.4	6.9	1.05	1.02
Documentary	119	7.17	7.4	7.5	8.7	1.6	7.1	1.12	1.06
Drama	2513	6.77	6.9	6.7	9.3	2	7.3	0.9	0.95
Family	522	5.71	6.3	6.7	8.6	1.7	6.9	1.45	1.21
Fantasy	576	6.38	6.4	6.7	8.9	1.7	7.2	1.36	1.17
Horror	537	5.63	5.9	6.2	8.7	2.2	6.5	1.25	1.12
Musical	131	6	6.7	7	8.5	2.1	6.4	1.5	1.22
Mystery	481	6.53	6.6	6.6	8.6	2.2	6.4	1.17	1.08
Romance	1073	6.02	6.5	6.5	8.6	2.1	6.5	0.97	0.99
Sci-Fi	586	6	6.35	6.7	8.8	1.9	6.9	1.48	1.22
Thriller	1351	5.58	6.4	6.4	9	2.2	6.8	1.1	1.05
Western	94	6.58	6.8	6.5	8.9	3.8	5.1	1.1	1.05
War	207	7.13	7.1	7.1	8.6	2.7	5.9	0.76	0.87
Sport	177	6.7	6.8	7.2	8.7	2	6.7	1.22	1.11
History	200	7.5	7.2	7.5	8.9	2	6.9	0.78	0.88
Film-Noir	6	7.6	7.65	0	8.2	7.1	1.1	0.16	0.39
Short	2	6.5	6.8	0	7.1	6.5	0.6	0.09	0.3



Insights: From the above data it can be observed that top 5 genres of movies are Drama, Comedy, Thriller, Action and Romance. Since Drama is the most common genre, after analysing its descriptive statistics it can be observed that:
Average IMDB scores of movies falling under Drama genre is 6.7. This means those movies are decent movies with some strength and also some weakness. Viewers enjoy watching these movies but overall, they are not exceptional.
Median IMDB score of 6.9 indicates that half scores are above 6.9 while other half are below 6.9.
Mode of an IMDB score of 6.7 indicates that frequency of 6.7 is occurring most in drama genre. This indicates most viewers have given 6.7 rating to movies in drama genre making it the most common IMDB score.
Other statistic calculation indicates most of the score are clustered around average 6.7 with few variations. Overall, these statistics indicate that movies under drama are moderate quality of movies with some variation.
Now analysing the five 'whys', it is found that drama is the most common genre because drama movies tell story about people emotions and their lives, which viewers find relatable, which in a way is comforting and interesting for the audience. Now because so many people enjoy watching these movies, the stakeholders can create movies that resonates with the audience and help in the success of movie industry.

2. **Movie Duration Analysis:** The task is to analyse the distribution of movie durations and identify the relationship between movie duration and IMDB score.

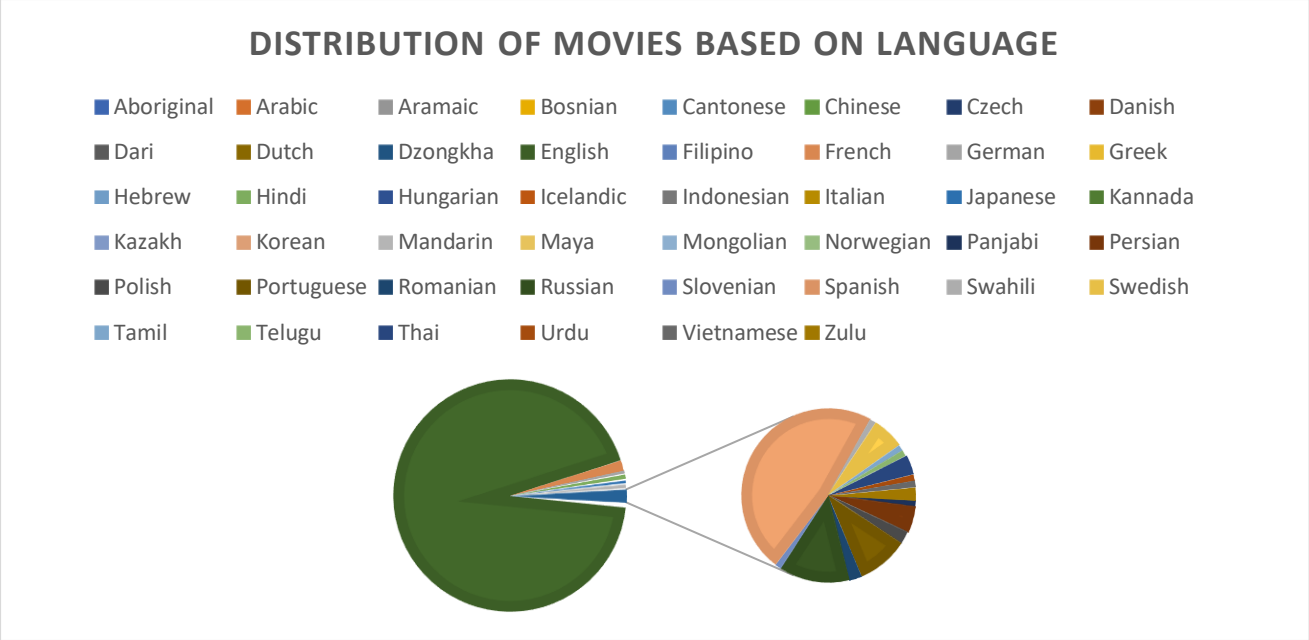
	A	B
1	**Statiscal Calculation**	**Movie Duration**
2	Mean	109.81
3	Median	105
4	Standatd Deviation	22.76



Insights: From the above data it can be observed that average of movie duration is 109.81 which suggests that typically the movie is 1hr 50 minutes long approximately. Median of 105 suggests that half of the movies are shorter than 105 minutes whereas other half are longer than 105 minutes.
The standard deviation of 22.76 suggest that there is moderate amount of variation in movie duration. This indicates that most of the movies cluster around the median of 105, but there are some significant deviations.
The above statistical calculation suggests that most of the movies are not that long and are standard films.
Now analysing the relationship between movie duration and IMDB score from the above scatter plot, it can be observed that the trendline shows positive slope and the R^2 value is 0.1307. It suggests that longer movies have higher IMDB score.
Further analysing the five 'whys', it is found that longer movies have higher IMDB score because this kind of movies allows time for character development and solve complex plots. This in turn engage viewers for long duration and make their experience memorable and enriching leaving strong impression on them and hence the viewers will appreciate the movie which will lead to high IMDB rating. This analysis will further help filmmakers focus on creating movies which will engage audience more.

3. **Language Analysis:** The task is to determine the most common languages used in movies and analyse their impact on the IMDB score using descriptive statistics.

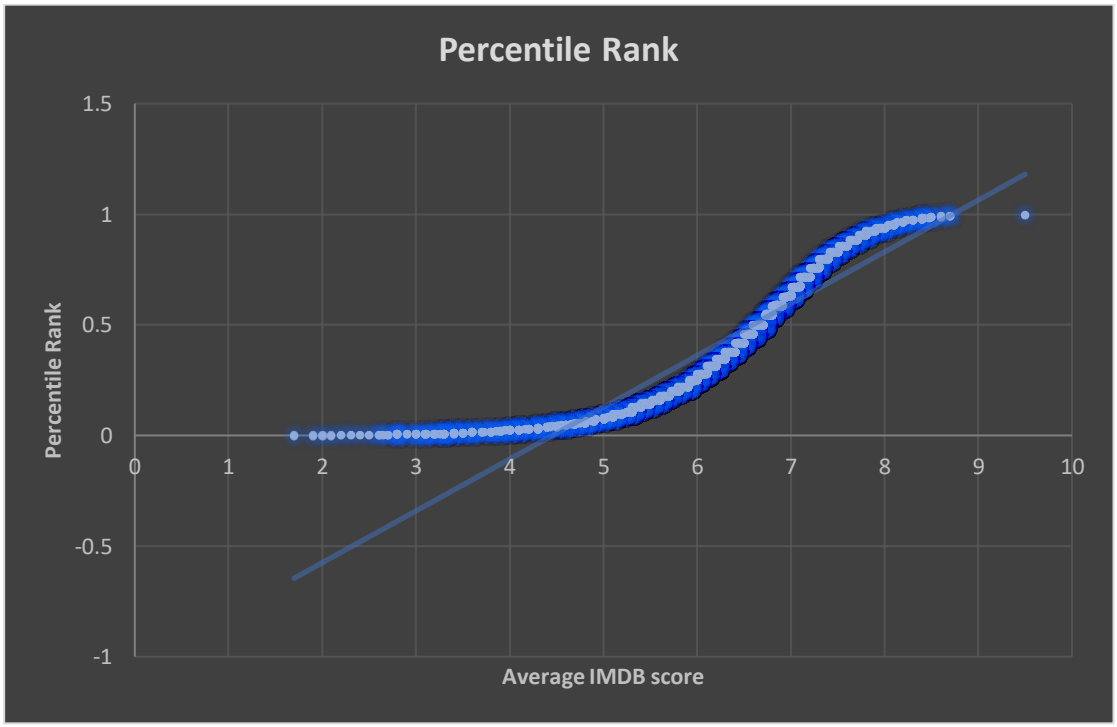
	A	B	C	D	E
1	Language	Count	Mean	Median	Standard Deviation
2	Aboriginal	2	6.95	6.95	0.55
3	Arabic	5	7.38	7.4	0.79
4	Aramaic	1	7.1	7.1	0
5	Bosnian	1	4.3	4.3	0
6	Cantonese	11	6.95	7.2	0.67
7	Chinese	3	5.67	5.7	0.45
8	Czech	1	7.4	7.4	0
9	Danish	5	7.5	8.1	0.96
10	Dari	2	7.5	7.5	0.1
11	Dutch	4	7.43	7.45	0.38
12	Dzongkha	1	7.5	7.5	0
13	English	4537	6.39	6.5	1.12
14	Filipino	1	6.7	6.7	0
15	French	73	7.04	7.2	0.72
16	German	19	7.34	7.6	0.93
17	Greek	1	7.3	7.3	0
18	Hebrew	5	7.58	7.6	0.3
19	Hindi	28	6.63	6.95	1.37
20	Hungarian	1	7.1	7.1	0
21	Icelandic	2	7.55	7.55	0.65
22	Indonesian	2	7.9	7.9	0.3
23	Italian	11	7.23	7.3	1.19
24	Japanese	16	7.37	7.6	1
25	Kannada	1	7.1	7.1	0
26	Kazakh	1	6	6	0
27	Korean	7	7.44	7.7	0.81
28	Mandarin	23	6.74	7	1.01
29	Maya	1	7.8	7.8	0
30	Mongolian	1	7.3	7.3	0
31	Norwegian	4	7.15	7.3	0.5
32	Panjabi	1	6.6	6.6	0
33	Persian	4	7.58	7.95	1.04
34	Polish	2	8.25	8.25	0.85
35	Portuguese	8	7.49	7.7	0.83
36	Romanian	2	7.2	7.2	0.7
37	Russian	11	6.36	6.5	1.32
38	Slovenian	1	6.4	6.4	0
39	Spanish	40	6.94	7.15	0.84
40	Swahili	1	7.4	7.4	0
41	Swedish	5	7.44	7.6	0.68
42	Tamil	1	5.1	5.1	0
43	Telugu	1	8.4	8.4	0
44	Thai	3	6.63	6.6	0.37
45	Urdu	1	7	7	0
46	Vietnamese	1	7.4	7.4	0
47	Zulu	2	7.1	7.1	0.2



Insights: From the above data it can be observed that English is the most common language used in movies.
From the statistical calculation is can be observed that English movies have an average IMDB score of 6.39 suggesting most movies received a score between 6 and 7.
Median of 6.5 implies that half scores are above 6.5 and other half is below 6.5 suggesting a balanced distribution of scores.
Standard Deviation of 1.12 indicates most of the IMDB score of movies are clustered around average score of 6.39 and median score of 6.5 with moderate variations.
Overall, by analysing the above statistical description it can be concluded that the IMDB score of movies in English category is a neutral score reflecting that movies in English language are decent movies liked by the audience.
Now analysing the five 'whys', it is found that English is the most common language in movies because English is a global language understood by wide number of audiences which makes it easier for these movies to reach broad range of viewers and hence maximizes their revenues.

4. **Director Analysis:** The task is to identify the top directors based on their average IMDB score and analyse their contribution to the success of movies using percentile calculations.

A	B	D	E
			Calculating 95th percentile
			8
Row Labels	Average of imdb_score	Percentile Rank	
John Blanchard	9.50	1	
Sadyk Sher-Niyaz	8.70	0.993	
Mitchell Altieri	8.70	0.993	
Cary Bell	8.70	0.993	
Mike Mayhall	8.60	0.99	
Charles Chaplin	8.60	0.99	
Ron Fricke	8.50	0.986	
Raja Menon	8.50	0.986	
Majid Majidi	8.50	0.986	
Damien Chazelle	8.50	0.986	
Sergio Leone	8.48	0.985	
Christopher Nolan	8.43	0.985	
S.S. Rajamouli	8.40	0.98	
Rakeysh Omprakash Mehra	8.40	0.98	
Robert Mulligan	8.40	0.98	
Richard Marquand	8.40	0.98	
Moustapha Akkad	8.40	0.98	
Marius A. Markevicius	8.40	0.98	
Jay Oliva	8.40	0.98	
Catherine Owens	8.40	0.98	
Asghar Farhadi	8.40	0.98	
Bill Melendez	8.40	0.98	



Insights: The following analysis can be done from the above data:
First the overall percentile distribution was calculated to set a benchmark against which director's score can be compared. 95th percentile was calculated which gives

the score below which 95% of the IMDB score fall. So, the 95th percentile score from overall distribution comes out to be 8.

Now the highlighted data above shows the top 10 directors based on their average IMDB score. John Blanchard average score (9.5) is higher than the 8, which means his movies receive higher ratings compared to other movies. Similarly, all top 10 directors have their average IMDB score higher than overall percentile distribution. This indicates the top 10 movie director's movie perform well and receive constantly high ratings.

Now if by analysing the percentile rank, it can be observed that John Blanchard has a percentile rank of 1 which means his average score is higher than 100% of the scores in the dataset. Rest of the directors in top 10 directors has percentile rank between 0.98 and 0.99 approximately which indicates their average score is higher than 98% to 99% of scores in the dataset.

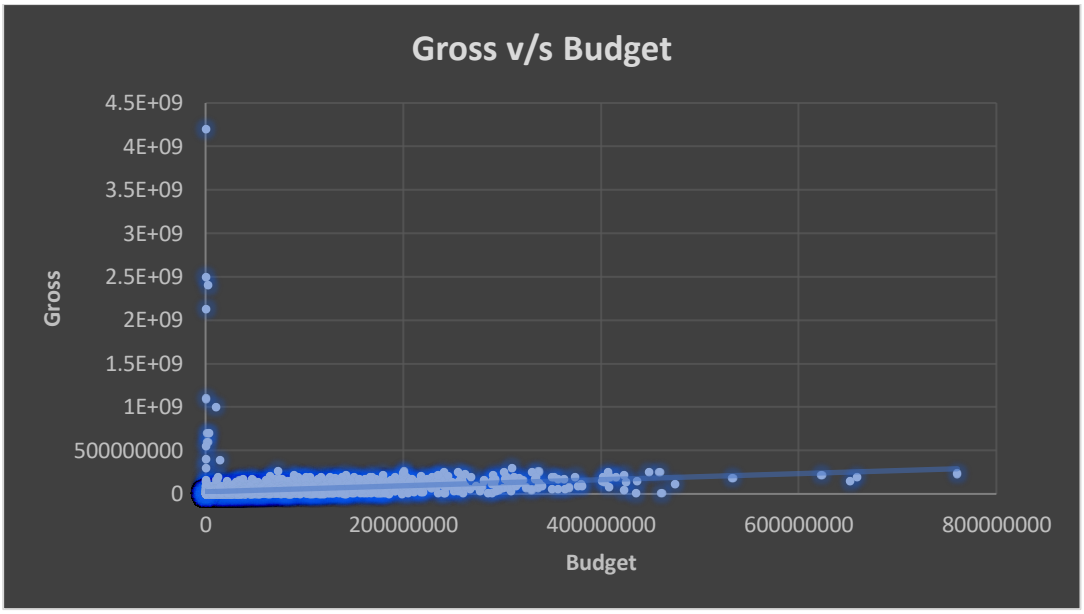
By analysing the scatter plot of the percentile rank the trendline shows positive slope which means as the director's average IMDB score increases then their percentile rank also increases.

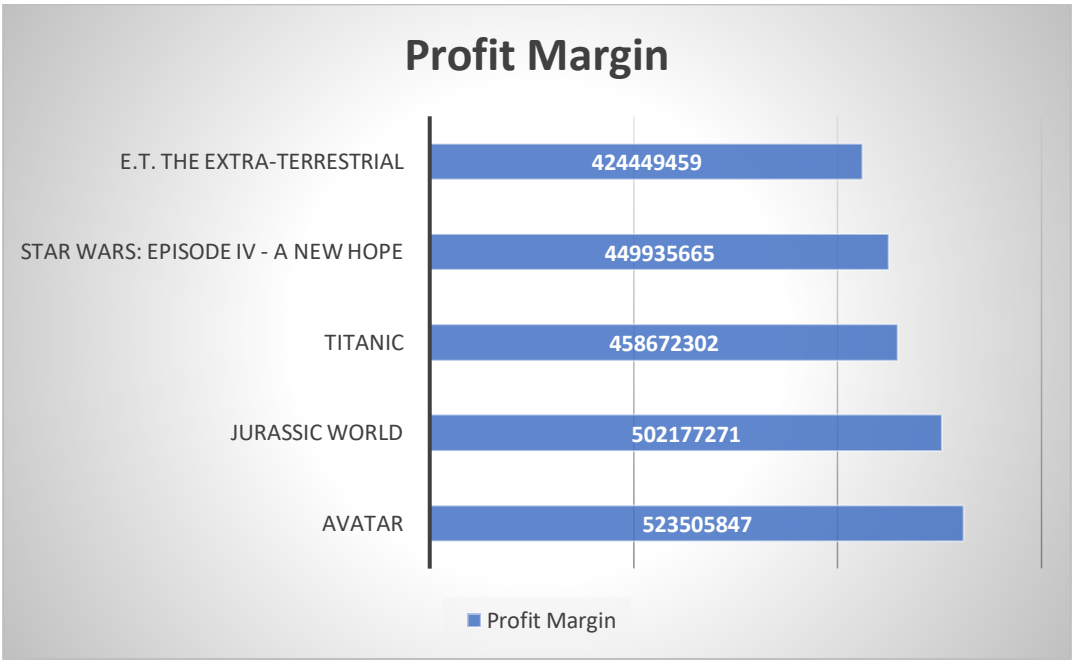
Now analysing the five 'whys', it was observed that as the director's average IMDB score increase their percentile rank also increases, this is because these directors make or produce movies that receive high ratings from audience. This indicates that these directors have deep understanding of what their audience will like and how to execute it, which further suggests that they might have previous successful experience in filmmaking.

5. **Budget Analysis:** The task is to analyse the correlation between movie budgets and gross earnings and identify the movies with the highest profit margin.

Correlation coefficient
0.223252814

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





Insights: From the above data the following interpretation can be made:
First analysing the correlation coefficient, it can be observed that it is 0.223 approximately. It is positive but relatively low, which indicates that the relationship between budget and gross earning is weak. Although it indicates to the tendency that higher movie budgets have high gross earning.
Now analysing the scatter plot of gross v/s budget, it can be observed that trendline is upwards which means slope is positive, which suggests that movies with higher budgets have high gross earning but the trendline is quite close to x-axis which means the relationship is not quite strong. It does suggest that higher budgets have high gross earning but this result is not guaranteed for every movie.

So, the above analysis suggests that budget is not the only factor contributing to movie's financial success, there are other factors responsible like movie plot, star cast, etc.
Now, coming to movies with highest profit margin, the data above shows top 5 movies with high profit margin. Avatar is the movie which has the highest profit margin of 523505847 in the dataset. The reason behind its financial success can be that the film resonated with audience and became a classic, effective marketing and promotions and many other factors.
Now analysing the five 'whys', it is observed overall that movies with high budget do have the tendency of earning high gross which leads to high profit. Understanding this relationship along with other factors helps the stakeholder of films to make informed business decision.

- **Results:**

Through this project as a data analyst, I gained significant insights that proved to be valuable. It deepened my knowledge of data analysis in organisations and process of making data driven decisions. This experience helped in improving my analytical skills and MS-Excel skills. I became comfortable in performing various functions of MS-Excel and how to execute them effectively. I gained experience in data pre-processing like data cleaning. I extracted meaningful insights from datasets and learnt how different factors of movie like genre, duration, language, directors and budgets can impact the movies success.

I analysed the genres having higher IMDB ratings, movie duration affecting the IMDB score, languages popular among audiences and the directors whose movies are highly rated. I also explored the relationship between movie budget and their gross earning and found movies with highest profit margin. I gave insights on how these factors impact the movies which will help the stakeholders of the movie make better strategies and decision.

In conclusion this project was not just about analysing the data but also using insights to improve movies chances of performing well and earning high profits. This experience will further help in boosting my career as data analyst in uncovering valuable insights from data and making informed decisions.

LINK TO EXCEL SHEET