

A dark blue background featuring a complex network graph. The graph consists of numerous nodes, represented by small dots in shades of blue, teal, and magenta, connected by thin, light-colored lines. The nodes are distributed across the frame, with a higher density in the center where the text is located. The overall effect is a sense of interconnectedness and data structure.

IMDb Movie Analysis



INTRODUCTION

The “IMDb Movie Analysis” project demands a complete and real-life analysis of the dataset from the view of a data analyst working for a movie firm. It aims at finding out various trends and anomalies in the dataset that can prove to be helpful in future while releasing new movies. For eg, what genre to target, which actors to cast, how much budget is a good to go budget for a particular genre, which movies are more liked by audience, how much profit are expensive movies making on an average, etc. Our story of the data analysis revolves around the performance of Indian Movies in the IMDb ratings. We will look and dive deep into the world of data to find answers to our questions, following the 5 why’s approach on our backstory.





Tech-Stack Used

Microsoft Excel 2019:

It is a popular and powerful data handling and analytics tool used worldwide as a benchmark.

Its key features include:

- **Enhanced Data Analysis:** Excel 2019 offers improved data analysis tools with new functions and features, making it easier for users to perform complex calculations and gain insights from their data.
- **Advanced Charting:** This version introduces new chart types and enhancements to help users create more informative and visually appealing charts and graphs.





The 5 why's approach

Why Indian Movies make very less profit?

Ans: Because the average gross is 22x less than average gross by movies worldwide. And average budget is 3x more than average budget on movies worldwide.

Why average gross is less in Indian Movies?

Ans: Because less people are watching Indian Movies, also less people are voting for Indian Movies.

Why less people are watching Indian Movies?

Ans: Because the bollywood being in itself a film industry, is not able to impress the audience.

Why are they not able to impress the audience?

Ans: Because of wrong selection of movie duration, directors, and actors.

Why wrong selection is done?

Ans: Because the film-makers do not follow worldwide trends and audience preferences based on data-driven analysis.





IMDb Rating Statistics:

Initial Analysis

Overall		Indian	
Average IMDb score	6.440990178	Average IMDb score	6.5
Median	6.6	Median	6.9
Mode	6.7	Mode	8.4
Range	7.9	Range	5.7
Variance	1.263024155	Variance	2.092727273
Standard Deviation	1.123843474	Standard Deviation	1.446626169

The comparatively high Variance and Standard Deviation signify that data is spread more, and there is more variability and less predictability in case of Indian Movies. However, the average, median, and mode are better than overall stats.

Voter Count Analysis:

Overall		Indian	
Average Votes	83614.02445	Average Votes	15695.21212
Median	34383	Median	7295
Mode	57	Mode	#N/A
Range	1689759	Range	70176
Variance	19086829251	Variance	400600209.5
Standard Deviation	138155.0913	Standard Deviation	20014.99961

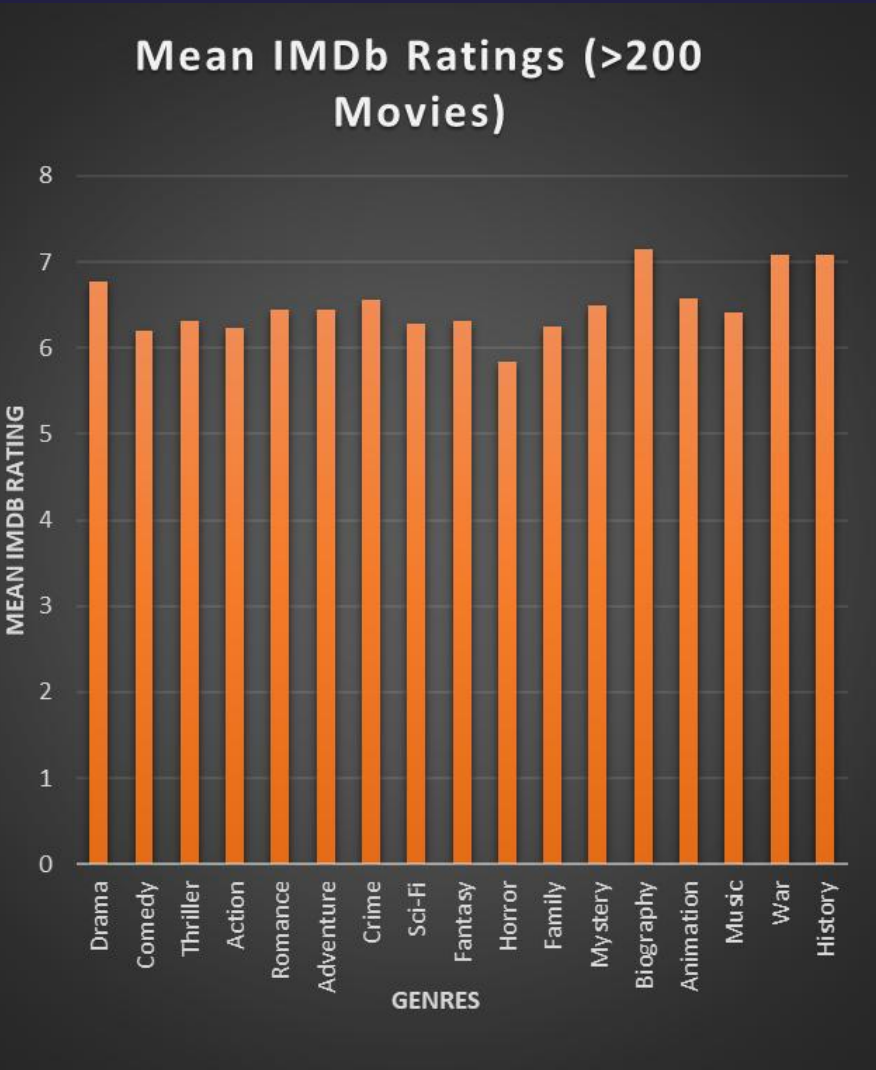
The voter statistics clearly show that the indian movies are not getting sufficient amount of votes despite being the second largest populated country in the world.



Task 1: **Movie Genre Analysis:** Analyze the distribution of movie genres and their impact on the IMDB score.

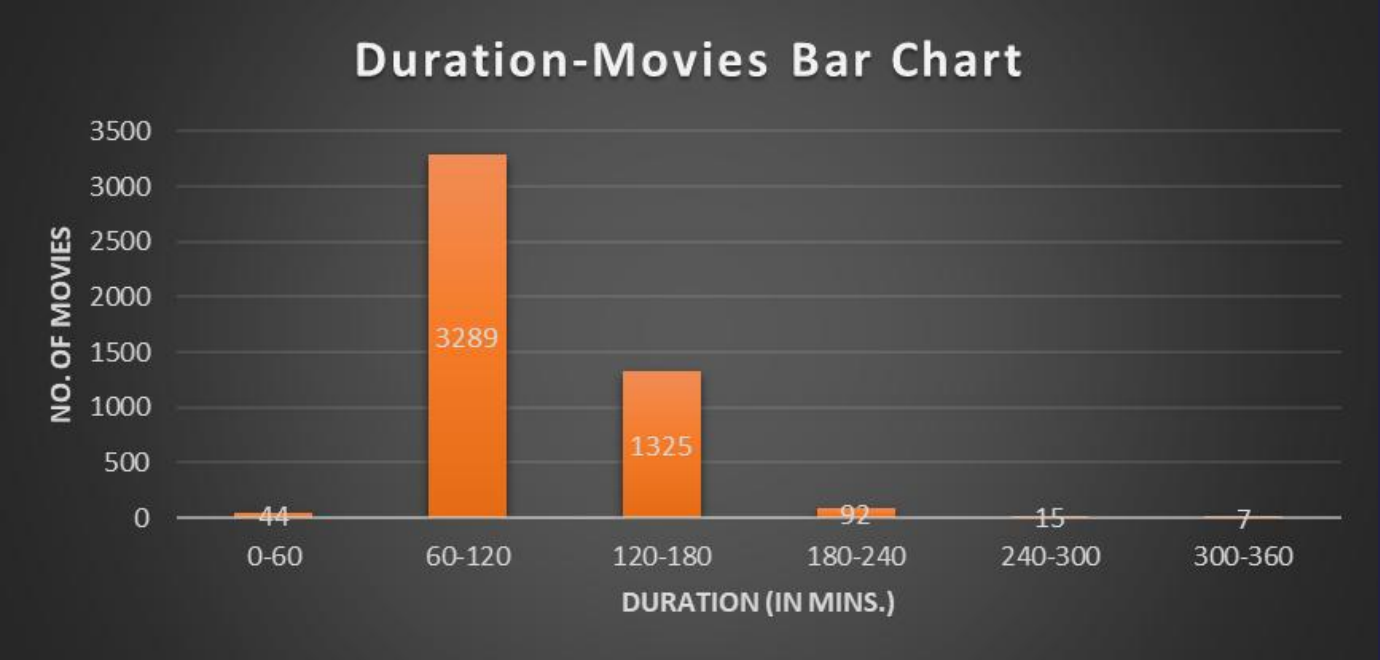
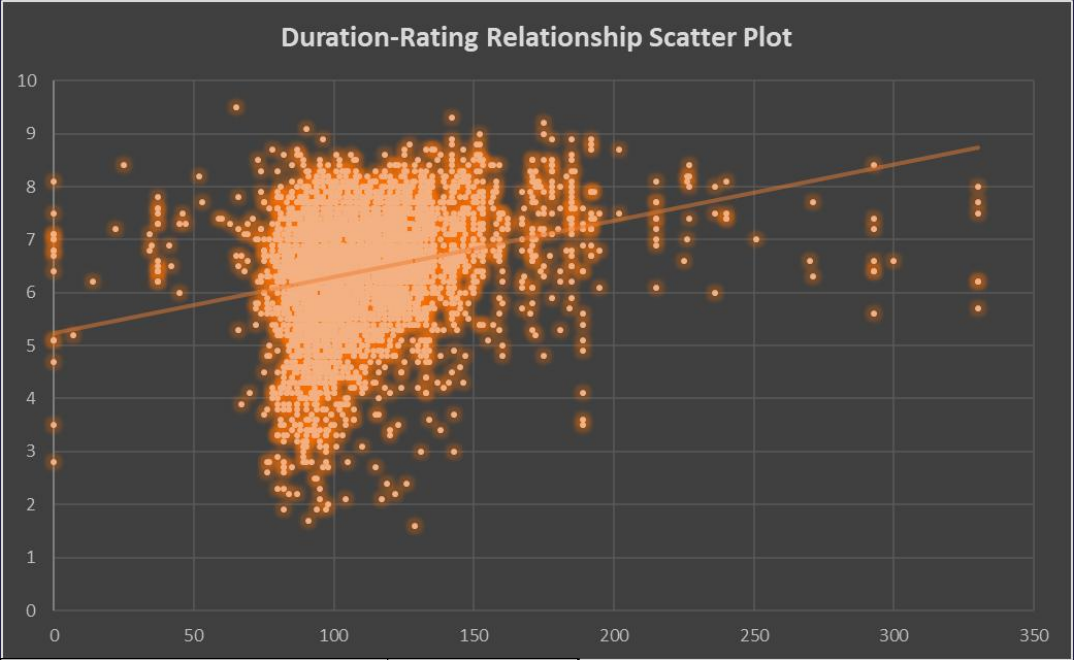
Main Task: Determine the most common genres of movies in the dataset. Then, for each genre, calculate descriptive statistics (mean, median, mode, range, variance, standard deviation) of the IMDB scores.

Genres	No. of Movies	Mean	Median	Mode	Max	Min	Range	Variance	Standard Deviation
Drama	2570	6.76572	6.9	7.2	9.3	2	7.3	0.908944	0.953385424
Comedy	1861	6.194734	6.3	6.7	9.5	1.7	7.8	1.181113	1.086790255
Thriller	1396	6.312894	6.4	6.1	9	2.2	6.8	1.109884	1.053510248
Action	1142	6.238091	6.3	6.1	8.8	1.7	7.1	1.235493	1.108911757
Romance	1096	6.45	6.5	6.5	8.6	2.1	6.5	1.001565	1.000782368
Adventure	913	6.443483	6.6	6.7	8.9	1.9	7	1.266029	1.125179761
Crime	883	6.560929	6.6	6.6	9.2	2.4	6.8	1.046427	1.022949982
Sci-Fi	610	6.280328	6.4	6.7	8.8	1.9	6.9	1.46019	1.208383341
Fantasy	604	6.311093	6.4	6.7	8.9	1.7	7.2	1.33539	1.155590843
Horror	556	5.832734	5.9	6.2	8.7	2.2	6.5	1.273308	1.128409684
Family	543	6.243831	6.4	6.7	8.7	1.7	7	1.434115	1.197545245
Mystery	493	6.488641	6.6	6.6	8.6	2.2	6.4	1.161261	1.077618274
Biography	292	7.15	7.2	7	9.5	4.5	5	0.547975	0.740253094
Animation	242	6.576033	6.7	6.7	8.9	1.7	7.2	1.352198	1.162840461
Music	212	6.406132	6.6	6.5	8.5	1.6	6.9	1.393217	1.180346184
War	210	7.077143	7.1	7.1	8.6	2.7	5.9	0.81102	0.900566715
History	205	7.084878	7.2	7.5	8.9	2	6.9	0.801688	0.895370338
Sport	181	6.602762	6.8	7.2	8.7	2	6.7	1.179812	1.086191499
Musical	132	6.507576	6.75	7	8.5	2.1	6.4	1.541967	1.2417598
Documentary	114	7.202632	7.4	7.5	9.1	1.6	7.5	1.199831	1.095367846
Western	94	6.703191	6.8	6.5	8.9	3.8	5.1	1.109498	1.053326877
Film-Noir	6	7.633333	8	8.2	8.2	7.2	1	0.1425	0.377491722
Short	5	6.38	6.2	#N/A	7.1	2.1	5	3.1336	1.770197729
News	3	7.533333	6.2	#N/A	6.6	5.2	1.4	0.346667	0.588784058
Reality-TV	2	4.75	5.1	5.1	5.1	5.1	0	0	0
Game-Show	1	2.9	7.9	#N/A	7.9	7.9	0	0	0



Task 2: **Movie Duration Analysis:** Analyze the distribution of movie durations and its impact on the IMDB score.

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



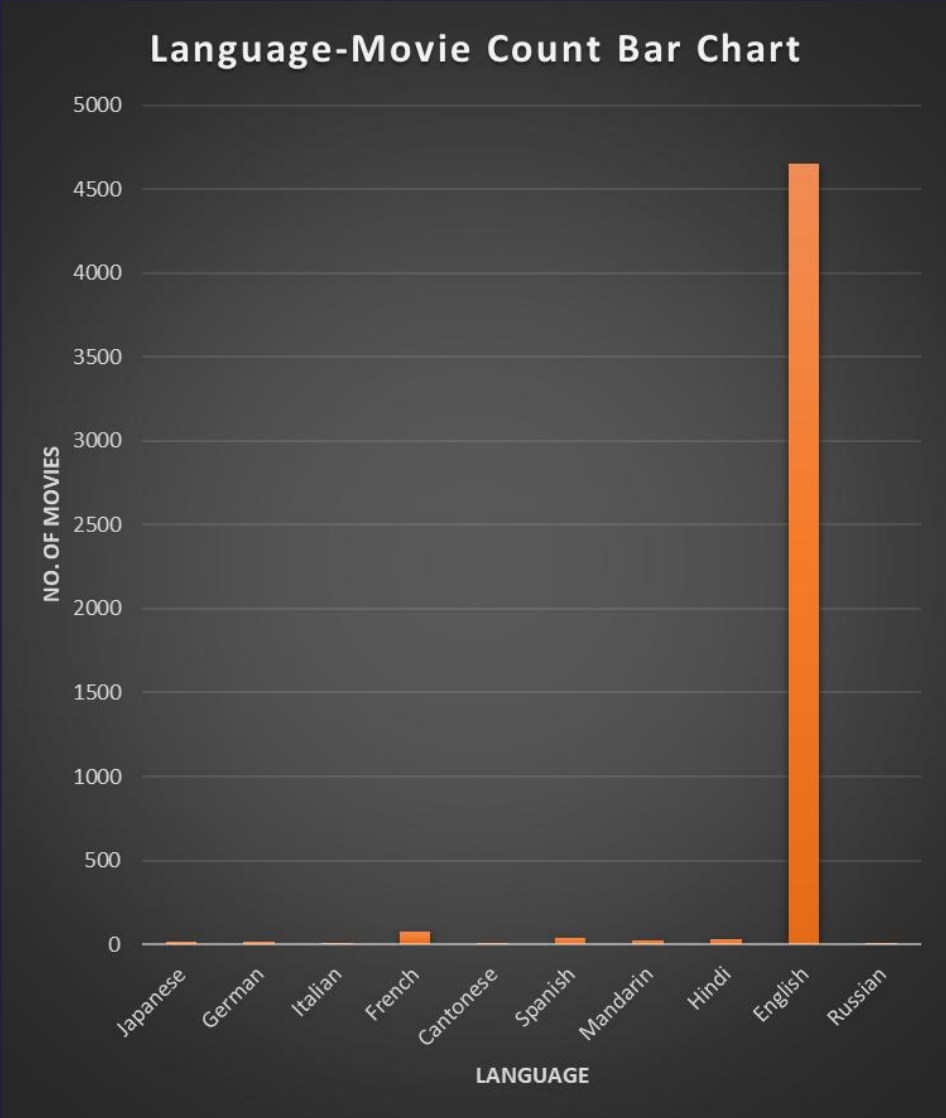
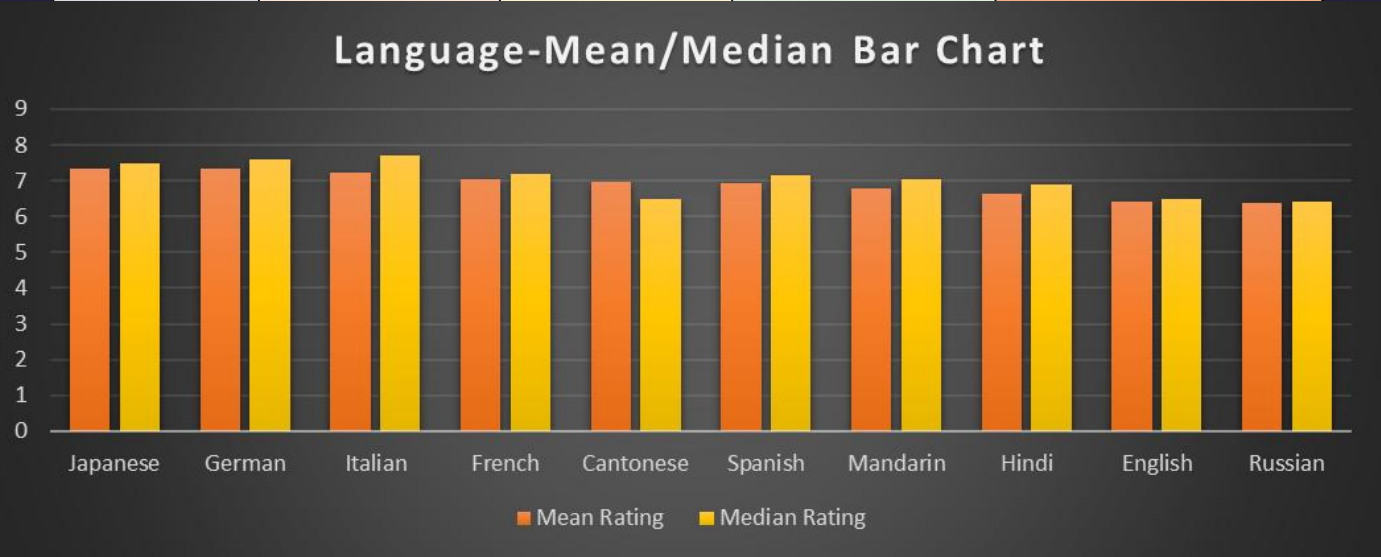
Coefficient of Correlation 0.26078895

Duration Range	Lower Limit	Upper Limit	No. of Movies	Mean Rating	Median Rating	Standard Deviation
0-60	0	60	44	6.684090909	6.9	1.11088482
60-120	60	120	3289	6.224961995	6.4	1.1219988
120-180	120	180	1325	6.806264151	6.9	0.956328675
180-240	180	240	92	7.183695652	7.4	1.065494086
240-300	240	300	15	7.066666667	7.2	0.723571389
300-360	300	360	7	6.842857143	6.6	0.819158354

Task 3: **Language Analysis:** Situation: Examine the distribution of movies based on their language.

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

Language	No. of Movies	Mean Rating	Median Rating	Standard Deviation
Japanese	17	7.347058824	7.5	1.009950494
German	19	7.342105263	7.6	1.014875509
Italian	11	7.227272727	7.7	0.663823137
French	73	7.038356164	7.2	0.724661419
Cantonese	11	6.954545455	6.5	0.747226276
Spanish	40	6.9375	7.15	0.858745451
Mandarin	24	6.7875	7.05	0.729678392
Hindi	28	6.632142857	6.9	1.284761552
English	4654	6.397636442	6.5	1.117734057
Russian	11	6.363636364	6.4	1.201101423



Task 4: **Director Analysis:** Influence of directors on movie ratings.

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

Firstly, we carry out the data of directors that have at least 3 movies on their name to have a fair differentiator on the basis of rating. Then we calculated the 95th percentile of these directors which came out to be 7.6565. Now, we filtered out all the directors that have average rating higher than this 95th percentile. The now obtained directors are the top 5% directors.

Top Directors (95 Percentile)	Rating	More than Mean?	More than mode?	More than median?
Sergio Leone	8.475	Yes	Yes	Yes
Christopher Nolan	8.425	Yes	Yes	Yes
Hayao Miyazaki	8.225	Yes	Yes	Yes
Quentin Tarantino	8.2	Yes	Yes	Yes
Frank Capra	8.06	Yes	Yes	Yes
Stanley Kubrick	8	Yes	Yes	Yes
David Lean	8	Yes	Yes	Yes
Billy Wilder	7.975	Yes	Yes	Yes
Frank Darabont	7.975	Yes	Yes	Yes
James Cameron	7.914286	Yes	Yes	Yes
Richard Brooks	7.8	Yes	Yes	Yes
Alfonso Cuara ³ N	7.8	Yes	Yes	Yes
Alejandro G. Ia±AĵRritu	7.783333	Yes	Yes	Yes
Fred Zinnemann	7.76	Yes	Yes	Yes
David Fincher	7.75	Yes	Yes	Yes
Peter Weir	7.725	Yes	Yes	Yes
Peter Jackson	7.675	Yes	Yes	Yes
Martin Scorsese	7.66	Yes	Yes	Yes

95th Percentile

7.6565

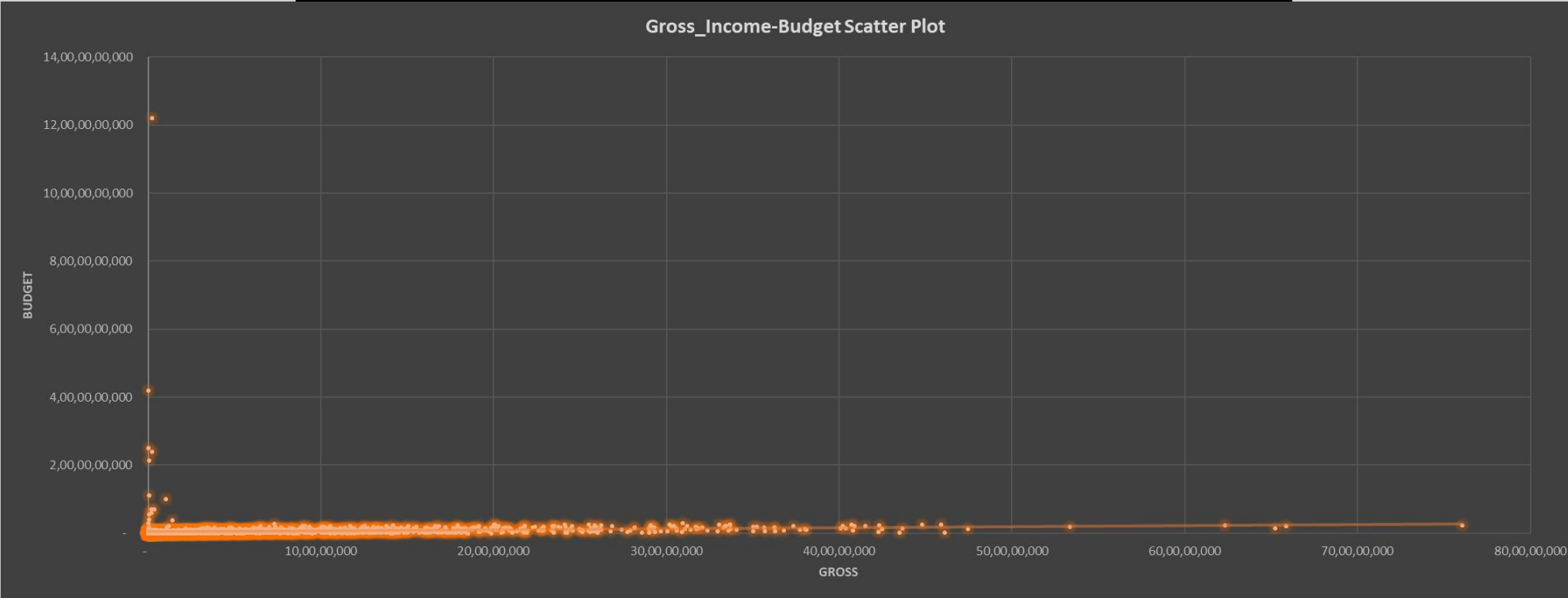


Task 5: **Budget Analysis:** Explore the relationship between movie budgets and their financial success.

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

STATISTICS	
Average Budget	39804015.68
Average Gross	48381280.61
Average Profit Margin	5720005.203
Average Profit %	525.4081404

The scatter plot and even the correlation coefficient does not show any strong relationship between budget and the gross earnings. However, there is an observation that the outliers in budget (exceptionally high) are making very less earning.



There are outliers with respect to the budget and gross that make the data skewed.

Coefficient of Correlation	0.100932					

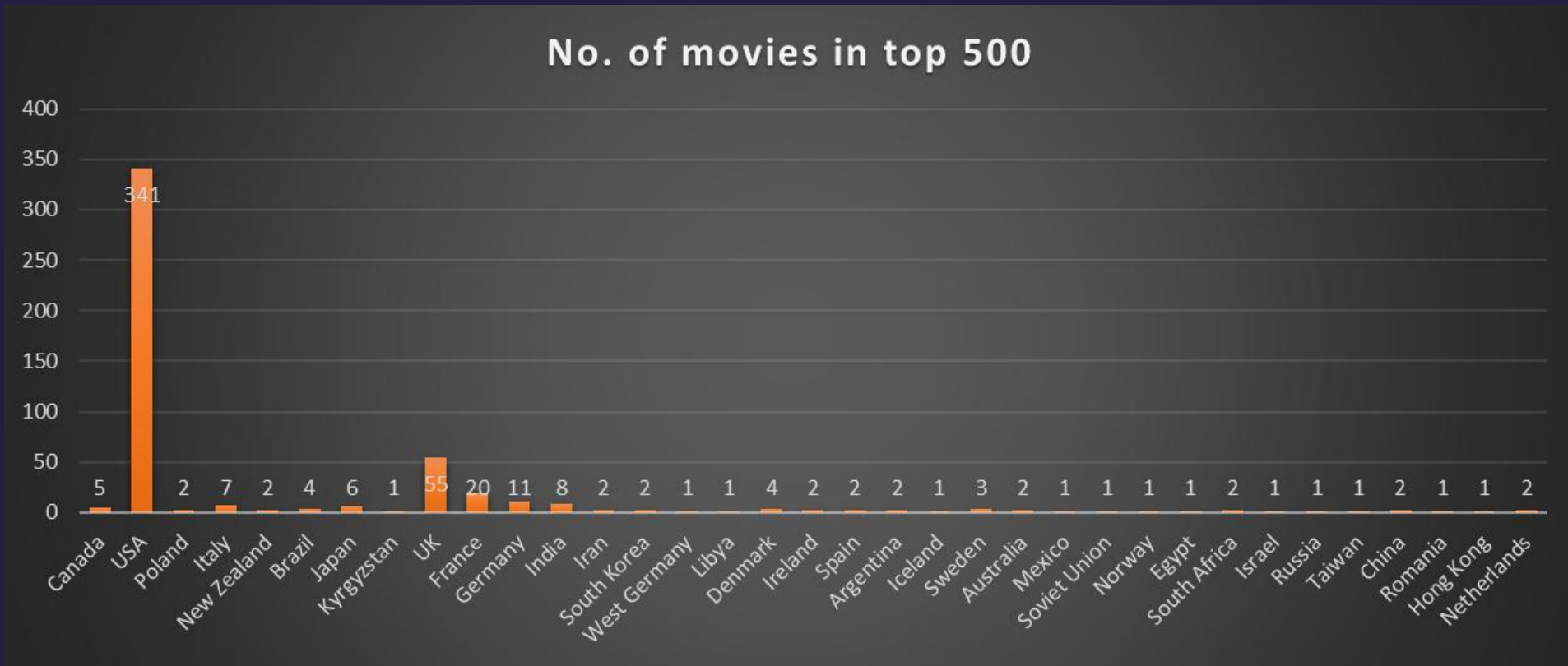
Budget Analysis Results

Highest Profit Making Movies (Top 20)		
movie_title	profit/loss (gross - budget)	profit/loss %
Avatar	52,35,05,847	221
Jurassic World	50,21,77,271	335
Titanic	45,86,72,302	229
Star Wars: Episode IV - A New Hope	44,99,35,665	4,090
E.T. the Extra-Terrestrial	42,44,49,459	4,042
The Avengers	40,32,79,547	183
The Lion King	37,77,83,777	840
Star Wars: Episode I - The Phantom Menace	35,95,44,677	313
The Dark Knight	34,83,16,061	188
The Hunger Games	32,99,99,255	423
Deadpool	30,50,24,263	526
The Hunger Games: Catching Fire	29,46,45,577	227
Jurassic Park	29,37,84,000	466
Despicable Me 2	29,20,49,635	384
American Sniper	29,13,23,553	495
Finding Nemo	28,68,38,870	305
Shrek 2	28,64,71,036	191
The Lord of the Rings: The Return of the King	28,30,19,252	301
Star Wars: Episode VI - Return of the Jedi	27,66,25,409	851
Forrest Gump	27,46,91,196	499

Highest Profit % Making Movies (Top 20)		
movie_title	profit/loss (gross - budget)	profit/loss %
Paranormal Activity	10,79,02,283	7,19,349
Tarnation	5,91,796	2,71,466
The Blair Witch Project	14,04,70,114	2,34,117
The Brothers McMullen	1,02,21,600	40,886
The Texas Chain Saw Massacre	3,07,75,468	36,843
El Mariachi	20,33,920	29,056
The Gallows	2,26,57,819	22,658
Super Size Me	1,14,64,368	17,637
Halloween	4,67,00,000	15,567
American Graffiti	11,42,23,000	14,701
Rocky	11,62,75,247	12,112
In the Company of Men	28,31,622	11,326
Napoleon Dynamite	4,41,40,956	11,035
Facing the Giants	1,00,74,663	10,075
Snow White and the Seven Dwarfs	18,29,25,485	9,146
Benji	3,90,52,600	7,811
My Date with Drew	84,122	7,647
The Circle	6,63,780	6,638
Fireproof	3,29,51,479	6,590
Open Water	3,00,00,882	6,000

Top 500 movies

- We carry out the top 500 movies by sorting using the IMDb rating column and then carried out the country wise share in top 500 movies.
- In this analysis, it was found out that USA being the owner of biggest film industry “Hollywood”, is clearly dominant in top 500 movies while India who also has its own film industry has only 8 movies which is less than 2% of the top 500.
- We further extend this analysis using the solutions to the tasks given. The solutions and process of analysis is provided in further slides.



Indian Movies

- We carry out the genre count, average budget, gross and profit, and average duration of Indian movies in this step.

Genres	No. of movies
Drama	23
Romance	18
Comedy	14
Thriller	7
Action	7
History	4
Musical	4
Adventure	3
Biography	3
War	3
Crime	1
Sci-Fi	1
Fantasy	1
Family	1
Animation	1
Documentary	1
Horror	0
Mystery	0
Music	0
Sport	0
Western	0
Film-Noir	0
Short	0
News	0
Reality-TV	0
Game-Show	0

Targeting trending but less rated genres.

Budget, Gross and Profit

Average Budget	10,33,99,653
Average Gross	22,08,451
Average Profit Margin	-11,81,29,423
Average Profit %	-35


Most Indian Movies are in loss.

Average Duration	138.8333333
------------------	-------------

Not the most preferred duration by audience worldwide.



Insights

- The project “**IMDb Movie Analysis**” has paved a way for a data analytics enthusiast like me to understand and implement industry level data analytics concepts using MS Excel and even what the field of data analytics demands from an individual aspiring to get into the domain.
 - This project has given me a vision of how to deal with real-time data and carry out important demographical trends and distributions over a dataset which is large in size and take out valuable information. I also found out how my analytical skills, understanding and visualization can benefit the film-makers and production houses.
 - During the analysis, there are many insights including that there is **no strong correlation between budget and the profit on movies, duration and IMDb ratings**. Also, an important finding is that **no. of movies should be considered while picking out top directors, languages, and genres**. While analyzing the reason of loss in Indian movies, it is found out that **Indian film-makers are targeting genres based on the trend only, without making a data-driven decision to choose highly rated genres**.
 - This project is a complete test of analytical skills of an individual and has capabilities to evaluate an individual trying to be a data analyst.
- 



Results

- The top trending genres are: **Drama, Comedy, Thriller, Action, Romance.**
 - The top rated genres are: **Biography, History, War, Drama, Animation.**
 - The preferred duration of movies is **1-2 hours.**
 - **Japanese, German, Italian, French** movies are the highest rated, while **English** being the standalone dominator in movies count.
 - **18 Directors** lie in the **95 percentile** category based on average ratings, while **Sergio Leone, Christopher Nolan and Hayao Miyazaki** being the **TOP 3** amongst them.
 - **Avatar, Jurassic World and Titanic** are the **TOP 3** profit making movies, while **Paranormal Activity, Tarnation and The Blair Witch Project** have the **highest profit %.**
 - I was able to solve all the tasks including the extra tasks that I assigned to myself with regards to analyzing the performance of Indian movies, using the concepts of statistics and excel functions and graphs.
- 

An abstract graphic featuring a network of interconnected nodes and lines. The nodes are small circles in various colors (blue, purple, red) and are connected by thin lines, forming a complex, web-like structure. The background is a dark blue gradient. The text "Thank You" is centered in the middle of the image, with a soft red glow behind it.

**Thank
You**