Final Project-1

PROJECT 5

IMDB Movie Analysis

RAKSHA NAYAK

Project Description: This Project "IMDB Movie Analysis" is designed to progressively explore different aspects of the dataset and uncover meaningful insights and provide a comprehensive analysis. Further, this project helps to dig deeper into the problems using "Why" approaches to uncover the root cause.

A potential problem to investigate could be: "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 want to understand what makes a movie successful to make informed decisions in their future projects.

Approach: The dataset provided is,

https://drive.google.com/file/d/1XpGThHzLnXxL_7aQo2sCpYL3SeB18MMB/view?usp=sharing

The dataset after analysis with answers, insights and visualization is,

https://docs.google.com/spreadsheets/d/1ddSoHI_4oFwEMMunYbHAfee9dUo627it/edit?usp=sharing&ouid=108154584635151678812&rtpof=true&sd=true

The above dataset contains records of IMDB movies.
The dataset contains,

Total data points: 5043

Attributes/Columns: 28

Column names:

1. color – Movie colour could be colored or black and white
2. director_name – Movie director name
3. num_critic_for_reviews – Number of reviews by film critics
4. duration – Movie duration
5. director_facebook_likes – Facebook likes for the director
6. actor_3_facebook_likes – Actor 3 facebook likes
7. actor_2_name – Actor 2 name
8. actor_1_facebook_likes – Actor 1 facebook likes
9. gross – Gross collection of the movie
10. genres – Genre of the movie
11. actor_1_name – Actor 1 name
12. movie_title – Movie Name
13. num_voted_users – Number of users who voted for the movie
14. cast_total_facebook_likes – Total movie cast's facebook likes
15. actor_3_name – Actor 3 name
16. facenumber_in_poster – Number of faces in the movie's poster
17. plot_keywords – Some keywords from plot of the movie
18. movie_imdb_link – IMDB link of the movie
19. num_user_for_reviews – Number of users who reviewed the movie

20. language – Original language of the movie
21. country – Country of origin of the movie
22. content_rating – Content rating of the movie
23. budget – Budget of the movie
24. title_year – Year in which the movie was released
25. actor_2_facebook_likes – Actor 2 facebook likes
26. imdb_score – IMDB score of the movie
27. aspect_ratio – Aspect ratio in which the movie was made
28. movie_facebook_likes – Facebook likes for the movie

After downloading the data set, we need to pre-process and clean (Data Clean) it, this is one of the most important steps to perform before performing analysis.

1. **Omission/Dropping Columns/Removing unwanted columns:** To make analysis easy, we need to first understand what questions we are supposed to answer and which columns are needed to support it. Once that is clear, we can proceed and drop columns which are not useful for that particular analysis in order to reduce the size of the dataset and avoid confusions.

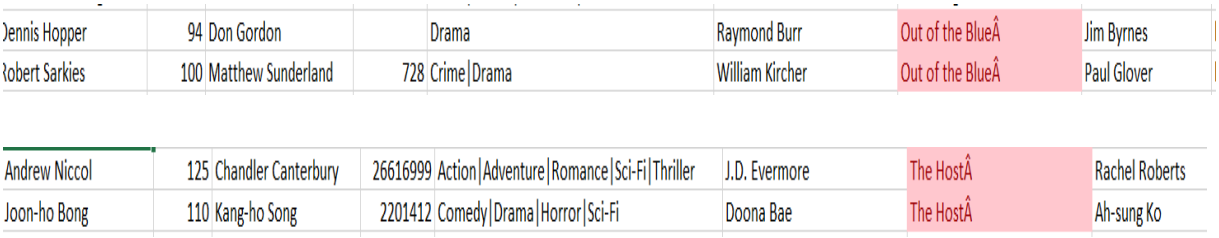
   We can drop below columns,

   1. color
   2. num_critic_for_reviews
   3. director_facebook_likes
   4. actor_3_facebook_likes
   5. actor_1_facebook_likes
   6. num_voted_users
   7. cast_total_facebook_likes
   8. facenumber_in_poster
   9. plot_keywords
   10. num_user_for_reviews
   11. content_rating
   12. actor_2_facebook_likes
   13. aspect_ratio
   14. movie_facebook_likes

   Total number of columns after cleaning: 14

2. **Handling Duplicate Values:** I have used **conditional formatting** on movie_title to look for a duplicate record as there would be less chances of 2 or more movies with the same name. Later, **sorted** the

column to spot the duplicates. I am deleting them by keeping the first occurred duplicates.



)ennis Hopper	94	Don Gordon		Drama	Raymond Burr	Out of the BlueÂ	Jim Byrnes
Robert Sarkies	100	Matthew Sunderland	728	Crime|Drama	William Kircher	Out of the BlueÂ	Paul Glover

Andrew Niccol	125	Chandler Canterbury	26616999	Action|Adventure|Romance|Sci-Fi|Thriller	J.D. Evermore	The HostÂ	Rachel Roberts
Joon-ho Bong	110	Kang-ho Song	2201412	Comedy|Drama|Horror|Sci-Fi	Doona Bae	The HostÂ	Ah-sung Ko

Except these 2 above records with same Movie name but different directors and language, rest all are deleted.

The records are now reduced from 5043 to 4931

3. **Handling Missing Data**: We need to check if there are any missing values in the dataset. We can click on movie_imdb_link columns link and try to fill as many missing values as possible.
   NOTE: This was the main reason for not dropping column, movie_imdb_link

   If there are many missing values in 1 single record then we can drop that record if any of the many details are not available on internet. Example: 12 Monkeys and 10,000 B.C that they are referring in the dataset is the series and not a movie for which there is no budget nor gross available on the internet.

   Total blanks/missing values – 1155
   Total records retained by filling missing values from the internet - 5
   Total Records after deleting blanks – 3781

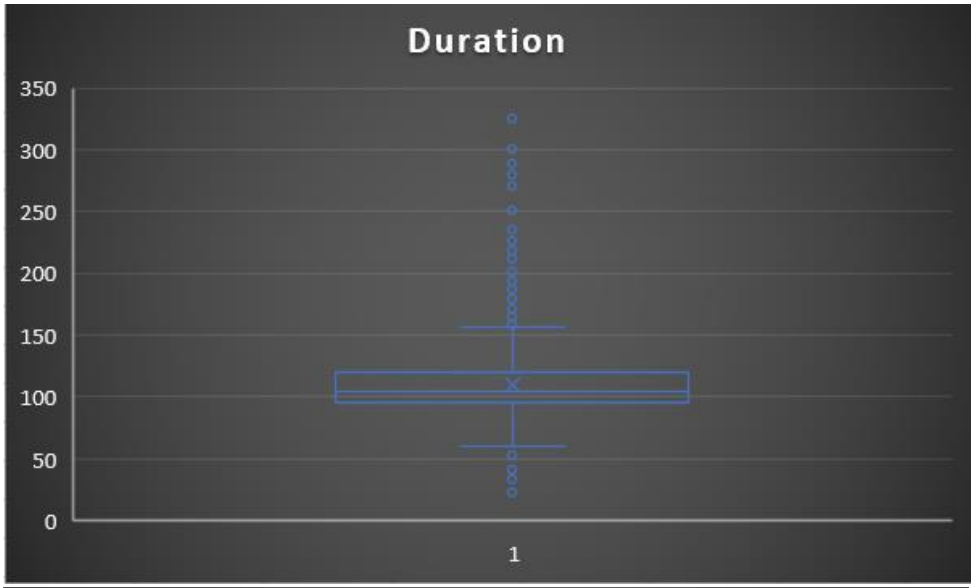
4. **Error Correction/Rectification**:
   - In the column, “movie_title”, we can observe that all the movie names end with Â. So we would be finding and replacing Â with “ “.

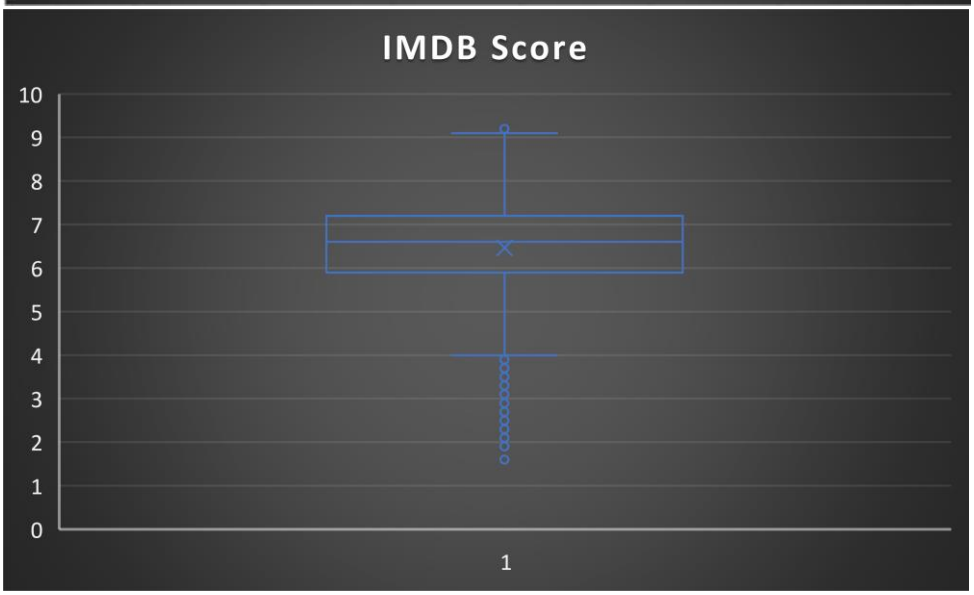
1	movie_title	1	movie_title
2	OsamaÂ	2	Osama
3	Nine QueensÂ	3	Nine Queens
4	The Holy GirlÂ	4	The Holy Girl
5	The Secret in Their EyesÂ	5	The Secret in Their Eyes
6	Knock OffÂ	6	Knock Off
7	AdoreÂ	7	Adore
8	AustraliaÂ	8	Australia
9	Babe: Pig in the CityÂ	9	Babe: Pig in the City
10	BabeÂ	10	Babe
11	Bran Nue DaeÂ	11	Bran Nue Dae
12	Critical CareÂ	12	Critical Care
13	Crocodile Dundee IIÂ	13	Crocodile Dundee II
14	Crocodile Dundee in Los AngelesÂ	14	Crocodile Dundee in Los Angeles
15	Crocodile DundeeÂ	15	Crocodile Dundee
16	Dark CityÂ	16	Dark City

- In column “country”, we have Official site and New Line which can be replaced with USA as they belong to USA on further research. Similarly, West Germany can be replaced as Germany for uniformity.

431	Country StrongÂ		2010	117	English	Official site	15000000	20218921	6.3 Drama|Music
417	Town & CountryÂ			2001	104	English	New Line	90000000	6712451
3781	Das BootÂ		1981	293	German	West Germany	14000000	11433134	8.4 Adventure|Drama|Thriller|War

5. **Outlier Detection**: From the below box plots, we get to know that there are outliers for duration, budget, gross and imdb_score and that may skew our analysis.





6. **Removing/Replacing Outliers**: During the process of looking for an outlier values, we understood that few values in gross and budget was wrongly mentioned considering only specific region collections. We changed it to the worldwide collection. Few examples as below.

Skin Trade (2014)

Budget. **$9,000,000 (estimated)** · Gross US & Canada. $1,242 · Opening weekend US & Canada. $162; May 10, 2015 · Gross worldwide. $595,268.

1	Skin Trade	2014	96	English	Thailand	9000000	162

The Jimmy Show (2001)

Budget. $1,500,000 (estimated) · **Gross** US & Canada. $1,000 · Opening weekend US & Canada. $703; Dec 15, 2002 **Gross** worldwide. $1,000.

5.3/10 ★★½☆☆ (632)

Missing: ~~Line~~ ~~Fire~~ ~~2006~~

The Jimmy Show	2001	96	English	USA	1500000	703

7. **Data Summary**: We have cleaned the data by managing missing values, errors and replacing the outliers with actual values. Summary of the cleaned data is as shown below,

Provided Dataset Details:						
Total Records in the dataset - 5043						
Total Clumns - 28						
After Data Cleaning:						
Total Records after deleting duplicates - 4931						
Total columns after dropping unwanted columns - 14						
Total blank/missing values - 1155						
Total records retained by filling missing values from the internet - 5						
Total Records after deleting blanks - 3781						

Budget		*Gross*		*IMDB Score*	
Mean	44778569.38	Mean	50360708.82	Mean	6.461164021
Standard Error	3665487.69	Standard Error	1118798.986	Standard Error	0.017239144
Median	24000000	Median	27347061.5	Median	6.6
Mode	20000000	Mode	8000000	Mode	6.7
Standard Deviation	225360431.3	Standard Deviation	68785668.75	Standard Deviation	1.059891959
Sample Variance	5.07873E+16	Sample Variance	4.73147E+15	Sample Variance	1.123370965
Kurtosis	2291.326957	Kurtosis	14.19808824	Kurtosis	1.129020774
Skewness	44.29636758	Skewness	3.048928729	Skewness	-0.722296188
Range	12215499782	Range	760504963	Range	7.7
Minimum	218	Minimum	884	Minimum	1.6
Maximum	12215500000	Maximum	760505847	Maximum	9.3
Sum	1.69263E+11	Sum	1.90363E+11	Sum	24423.2
Count	3780	Count	3780	Count	3780

Tech stack used: Microsoft Excel Version 2407, 2019 – Excel is a spreadsheet editor developed by Microsoft. It features calculation or computation capabilities, graphing tools, pivot tables etc.

Data Analytics Tasks:

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

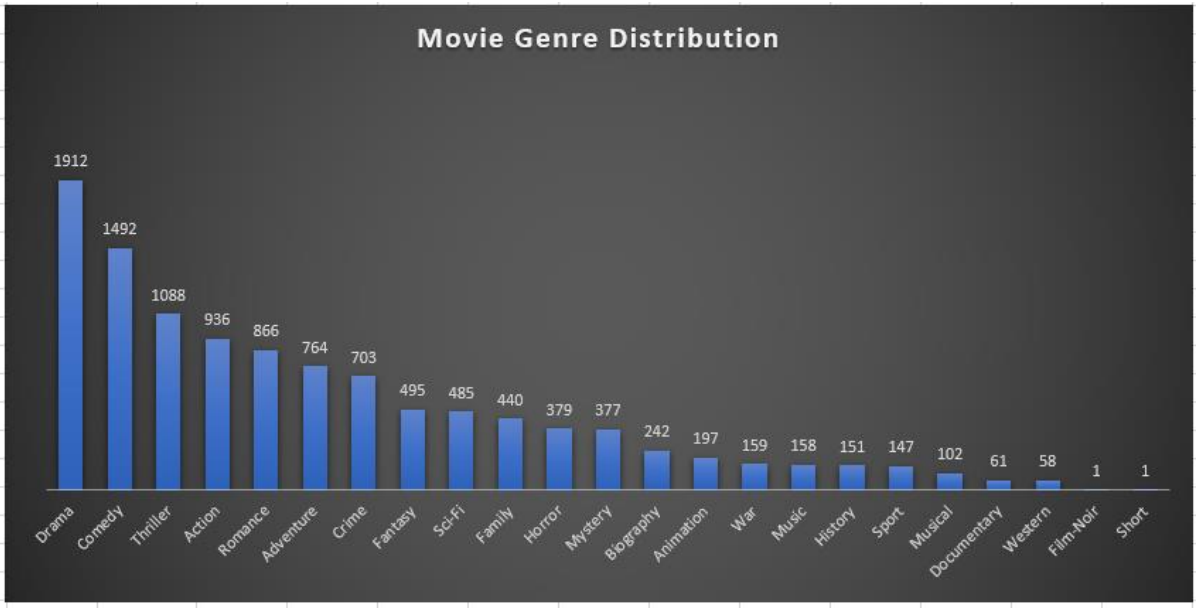
- **Task:** Determine the most common genres of movies in the dataset. Then, for each genre, calculate descriptive statistics (mean, median, mode, range, variance, standard deviation) of the IMDB scores.
- **Hint:** Use Excel's COUNTIF function to count the number of movies for each genre. You might need to manipulate the 'genres' column to separate multiple genres for a single movie. Use Excel's functions like AVERAGE, MEDIAN, MODE, MAX, MIN, VAR, and STDEV to calculate descriptive statistics. Compare the statistics to understand the impact of genre on movie ratings.

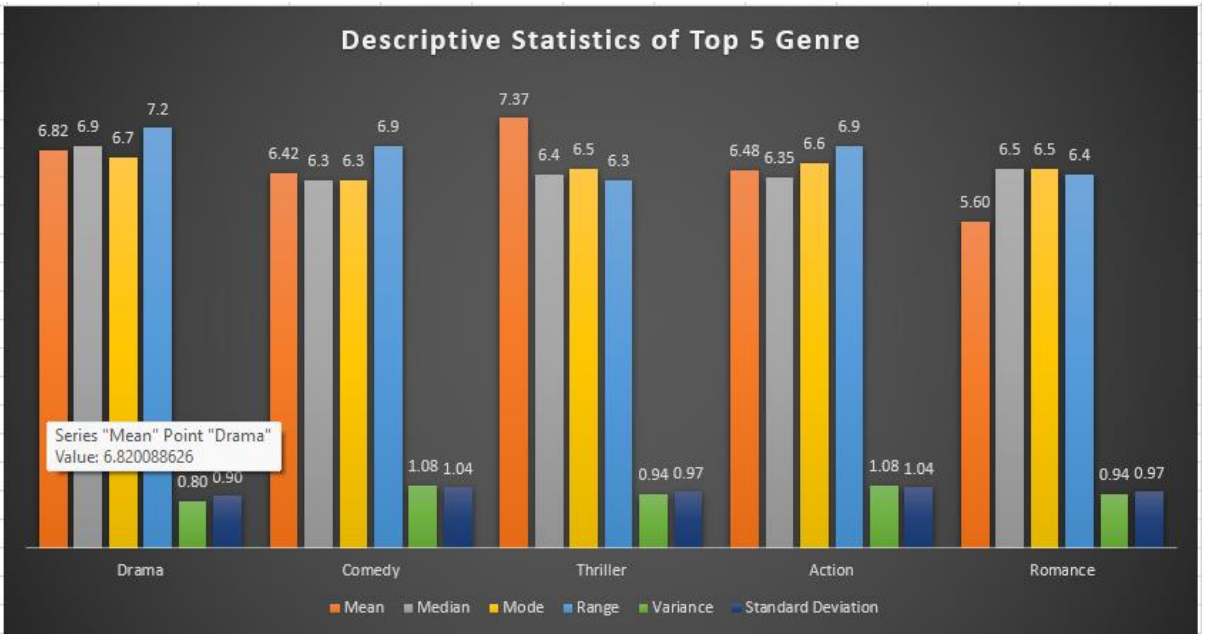
Output:

	N	O	P	Q	R	S	T	U	V	W	X
1	genres	genre1	genre2	genre3	genre4	genre5	genre6	genre7	genre8	Each Genre	Genre Count
2	Adventure|Comedy|Drama|Family|Fantasy	Adventure	Comedy	Drama	Family	Fantasy				Drama	1912
3	Action|Comedy|Thriller	Action	Comedy	Thriller						Comedy	1492
4	Action|Fantasy|Sci-Fi|Thriller	Action	Fantasy	Sci-Fi	Thriller					Thriller	1088
5	Drama	Drama								Action	936
6	Horror|Mystery|Sci-Fi|Thriller	Horror	Mystery	Sci-Fi	Thriller					Romance	866
7	Crime|Drama|Thriller	Crime	Drama	Thriller						Adventure	764
8	Drama|Romance	Drama	Romance							Crime	703
9	Comedy|Drama|Musical	Comedy	Drama	Musical						Fantasy	495
10	Action|Crime|Sci-Fi|Thriller	Action	Crime	Sci-Fi	Thriller					Sci-Fi	485
11	Comedy|Drama|Family	Comedy	Drama	Family						Family	440
12	Adventure|Comedy|Crime	Adventure	Comedy	Crime						Horror	379
13	Action|Horror|Sci-Fi|Thriller	Action	Horror	Sci-Fi	Thriller					Mystery	377
14	Drama|Mystery|Thriller	Drama	Mystery	Thriller						Biography	242
15	Drama	Drama								Animation	197
16	Mystery|Thriller	Mystery	Thriller							War	159
17	Adventure|Comedy	Adventure	Comedy							Music	158
18	Action|Adventure|Comedy	Action	Adventure	Comedy						History	151
19	Horror	Horror								Sport	147
20	Action|Drama|Fantasy|Mystery|Sci-Fi|Thriller	Action	Drama	Fantasy	Mystery	Sci-Fi	Thriller			Musical	102
21	Adventure|Drama|Romance|War	Adventure	Drama	Romance	War					Documentary	61
22	Crime|Drama|Mystery|Thriller	Crime	Drama	Mystery	Thriller					Western	58
23	Comedy|Drama	Comedy	Drama							Film-Noir	1
24	Animation|Comedy|Family|Musical	Animation	Comedy	Family	Musical					Short	1
25	Comedy|Romance	Comedy	Romance								

Each Genre	Genre Count	Mean	Median	Mode	Range	Variance	Standard Deviation
Drama	1912	6.820088626	6.9	6.7	7.2	0.803043	0.896126917
Comedy	1492	6.421539961	6.3	6.3	6.9	1.080707	1.039570455
Thriller	1088	7.366666667	6.4	6.5	6.3	0.938554	0.968790095
Action	936	6.481303419	6.35	6.6	6.9	1.079832	1.039149814
Romance	866	5.6	6.5	6.5	6.4	0.937869	0.968436621
Adventure	764	6.530769231	6.6	6.6	6.6	1.251541	1.118722905
Crime	703	6.418253968	6.6	6.6	6.9	0.966995	0.983358924
Fantasy	495	6.325714286	6.4	6.7	6.7	1.300829	1.14053879
Sci-Fi	485	6.7625	6.4	7	6.9	1.36179	1.166957778
Family	440	6.066666667	6.3	5.4	6.7	1.362078	1.16708084
Horror	379	6.354487179	5.9	6.2	6.3	0.979536	0.989715003
Mystery	377	6.517391304	6.5	6.6	5.5	1.014759	1.007352324
Biography	242	6.584466019	7.2	7	4.4	0.502154	0.708628049
Animation	197	6.593333333	6.8	7	5.8	0.983451	0.991690852
War	159	7.048427673	7.1	7.1	4.3	0.648284	0.805160662
Music	158	6.467721519	6.5	6.5	6.9	1.460947	1.208696556
History	151	6.594701987	7.2	7.7	3.4	0.451548	0.67197302
Sport	147	6.471428571	6.8	7.2	6.4	1.091291	1.044648585
Musical	102	6.3	6.7	7.1	6.4	1.294852	1.13791563
Documentary	61	6.271052632	7.2	6.6	6.9	1.52738	1.235872055
Western	58	5.9	6.8	6.8	4.8	0.979845	0.989871417
Film-Noir	1	6.2	7.7	#N/A	0	0	0
Short	1	7.1	7.1	#N/A	0	0	0

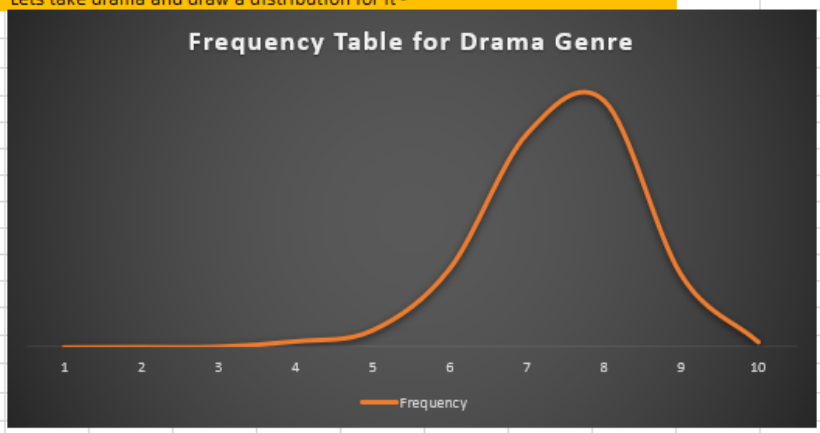
Visual Representation:





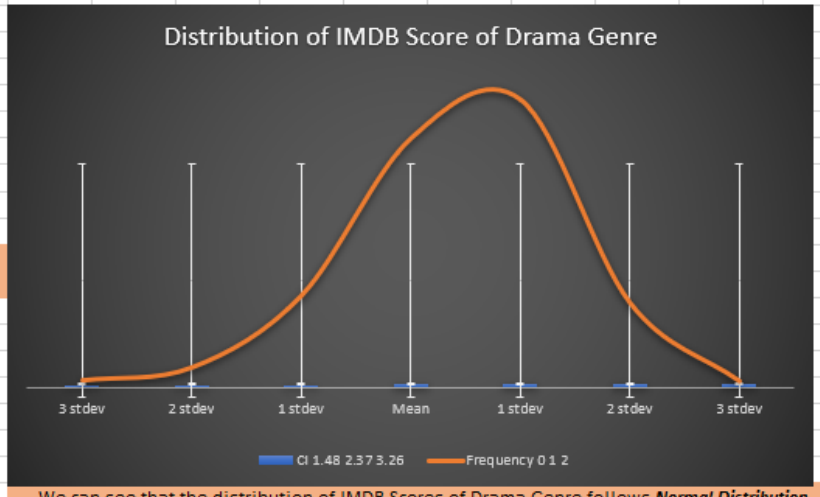
Lets take drama and draw a distribution for it -

	CI	Frequency
	1.48	0
	2.37	1
	3.26	2
3 stdev	4.15	18
2 stdev	5.04	51
1 stdev	5.93	235
Mean	6.82	637
1 stdev	7.71	736
2 stdev	8.6	216
3 stdev	9.49	16



	CI	frequency
	1.48-2.37	1
	2.37-3.26	2
	3.26-4.15	18
3 stdev	4.15-5.04	51
2 stdev	5.04-5.93	235
1 stdev	5.93-6.82	637
1 stdev	6.82-7.71	736
2 stdev	7.71-8.6	216
3 stdev	8.6-9.49	16
		1912

3 values falls out	0.1569038	percent
Values falling inside	99.843096	percent



We can see that the distribution of IMDB Scores of Drama Genre follows ***Normal Distribution***. Since it follows Normal Distribution, we can say that 99.8% of its data lies above IMDB Score of 4.15. This is the reason that it's Standard Deviation is small.

Therefore, empirical rule of 68%,95% and 99.7% holds good as 99.8% vales are falling within the 3 standard deviation

Insight: The top 5 most common genres are Drama, Comedy, Thriller, Action and Romance.

Observation: All top 5 genres descriptive statistics (mean, median, mode, range, variance, standard deviation) is almost same. Also, the drama genre has a normal distribution, where 99.8% of IMDB score values fall within 3 standard deviation and thus satisfies **Empirical Rule**.

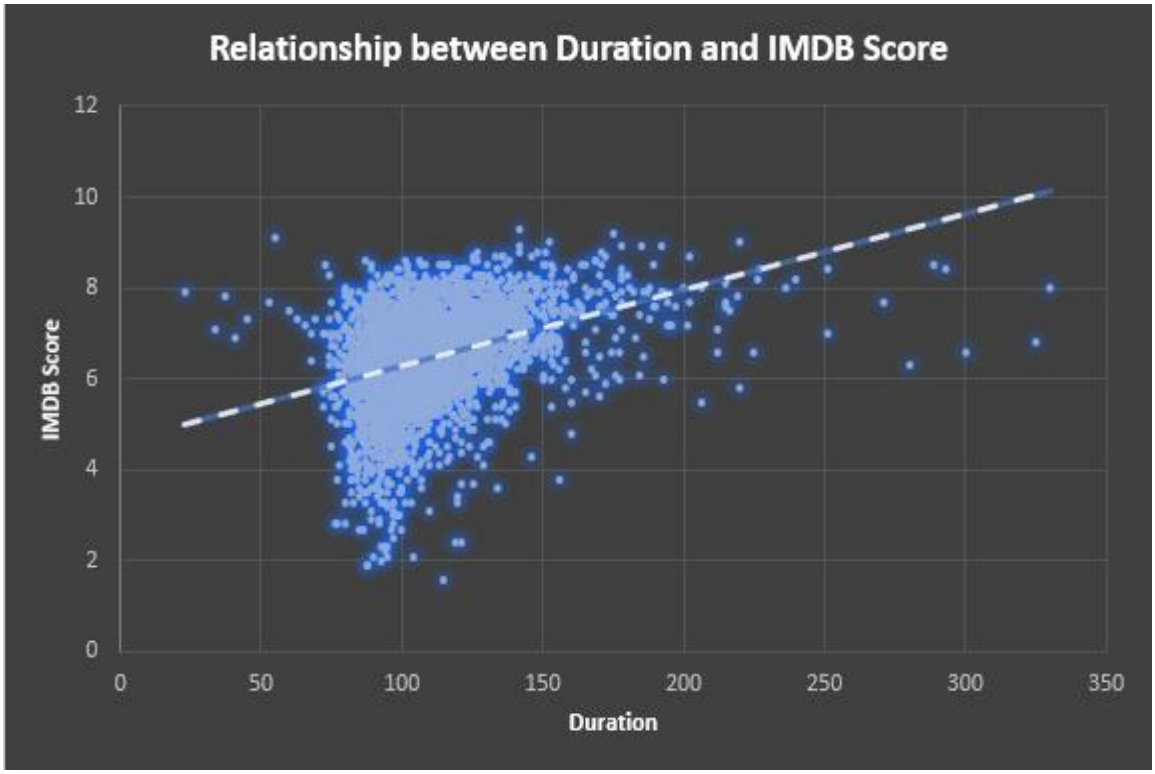
B. Movie Duration Analysis: Analyze the distribution of movie durations and its impact on the IMDB score.

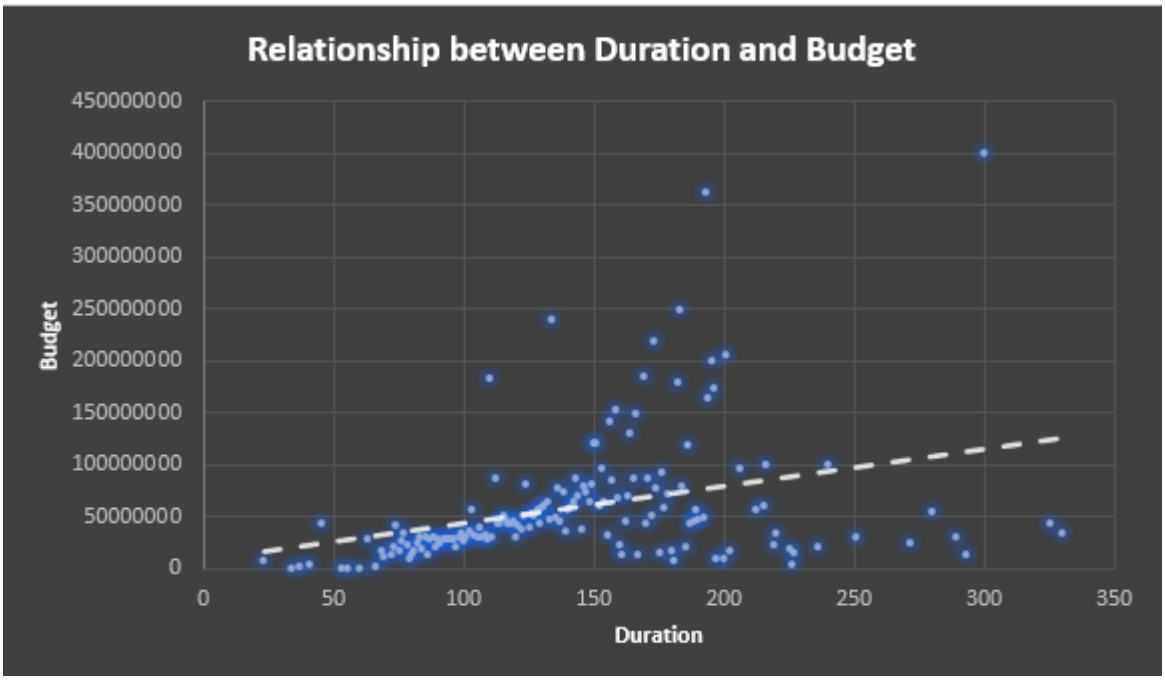
- Task: Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.
- Hint: Calculate descriptive statistics such as mean, median, and standard deviation for movie durations. Use Excel's functions like AVERAGE, MEDIAN, and STDEV. Create a scatter plot to visualize the relationship between movie duration and IMDB score. Add a trendline to assess the direction and strength of the relationship.

Output:

E	F	G
Mean	Median	Standard Deviation
109.8114	105	22.80083141

Visual Representation:





Insight: The first plot shows a positive trendline, where more movies lie in between the duration of 75 to 200 minutes with good IMDB scores.

The one good reason behind the high duration movies having good ratings can be understood from second scatter plot, where high duration movies are having high budget. High budget movies have sufficient funds for,

- Big stars: Actors and actresses with greater fan followings, which in turn will increase the popularity of the movie and might result in more people watching the movie and liking it. Thus, resulting in increase in IMDB Score.
- Good Plots, Special-Effects, Graphics and Music: This attracts the viewers and leaves behind a good view/experience about the movie along with sense of satisfaction and good feeling.
- Broader Marketing/Publicity and Distribution

These above factors play an important role in the success of a movie and in increasing the IMDB Score.

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

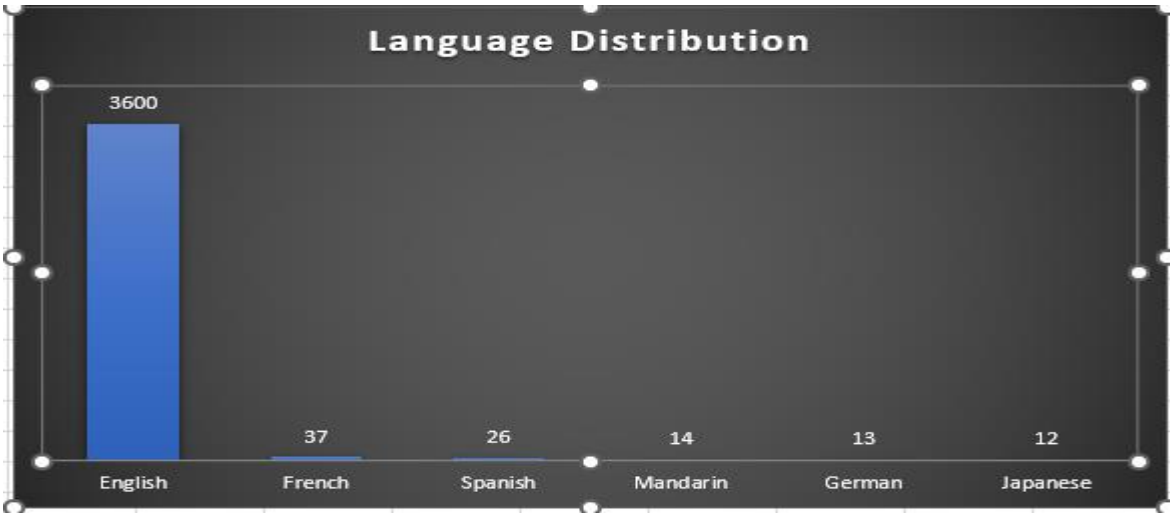
- **Task:** Determine the most common languages used in movies and analyze their impact on the IMDB score using descriptive statistics.
- **Hint:** Use Excel's COUNTIF function to count the number of movies for each language. Calculate the mean, median, and standard deviation of the IMDB scores for each language. Compare the statistics to understand the impact of language on movie ratings.

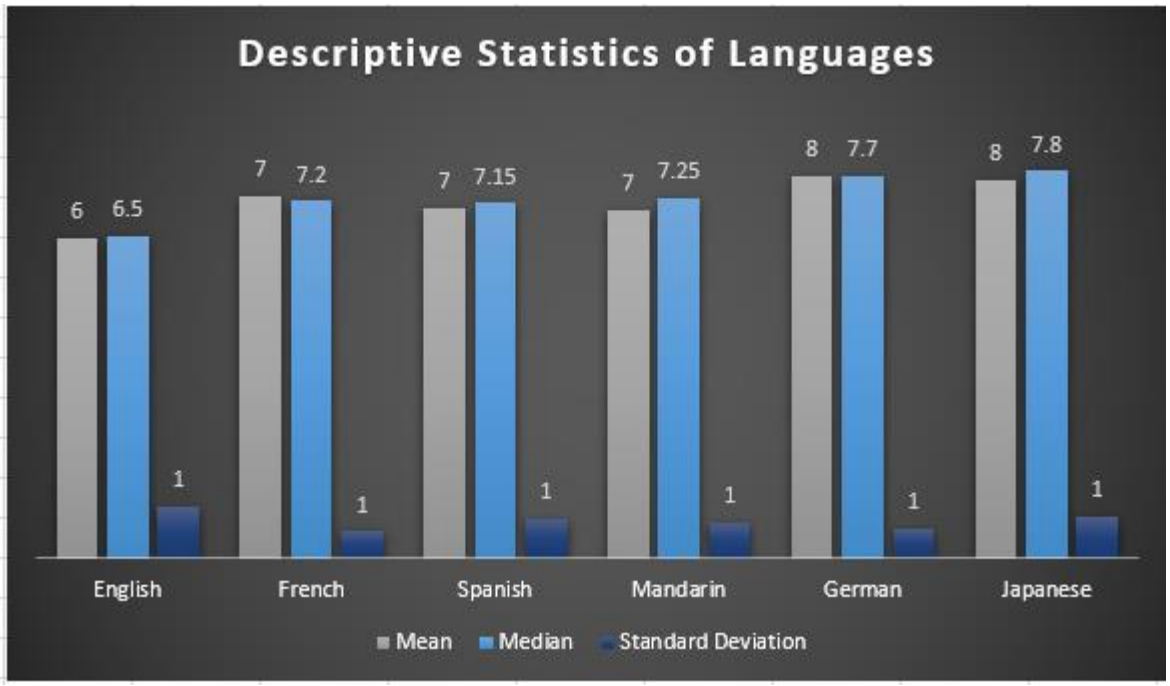
Output:

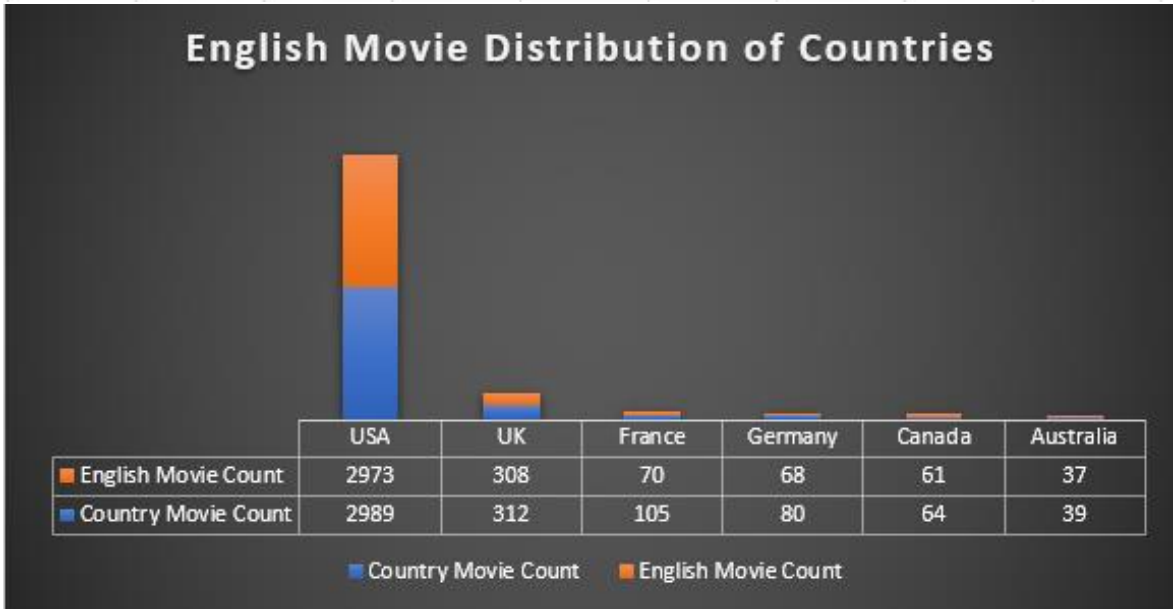
Language	Language Count	Mean	Median	Standard Deviation
English	3600	6.461164	6.5	1.053937722
French	37	7.286486	7.2	0.553691378
Spanish	26	7.05	7.15	0.810151933
Mandarin	14	7.021429	7.25	0.737930089
German	13	7.692308	7.7	0.615769111
Japanese	12	7.625	7.8	0.861321659
Hindi	10	6.76	7.05	1.05470375
Cantonese	8	7.2375	7.3	0.412121038
Italian	7	7.185714	7	1.069617517
Korean	5	7.7	7.7	0.509901951
Portuguese	5	7.76	8	0.875442745
Norwegian	4	7.15	7.3	0.497493719
Danish	3	7.9	8.1	0.43204938
Dutch	3	7.566667	7.8	0.329983165
Persian	3	8.133333	8.4	0.449691252
Thai	3	6.633333	6.6	0.368178701
Aboriginal	2	6.95	6.95	0.55
Dari	2	7.5	7.5	0.1
Hebrew	2	7.65	7.65	0.35
Indonesian	2	7.9	7.9	0.3
Arabic	1	7.2	7.2	0
Aramaic	1	7.1	7.1	0
Bosnian	1	4.3	4.3	0
Czech	1	7.4	7.4	0
Dzongkha	1	7.5	7.5	0
Filipino	1	6.7	6.7	0
Hungarian	1	7.1	7.1	0
Icelandic	1	6.9	6.9	0
Kazakh	1	6	6	0
Maya	1	7.8	7.8	0
Mongolian	1	7.3	7.3	0
None	1	8.5	8.5	0
Polish	1	9.1	9.1	0
Romanian	1	7.9	7.9	0
Russian	1	6.5	6.5	0
Swedish	1	7.6	7.6	0
Telugu	1	8.4	8.4	0
Vietnamese	1	7.4	7.4	0
Zulu	1	7.3	7.3	0
	3780			

Country	Country Movie	English Movie Co
USA	2989	2973
UK	312	308
France	105	70
Germany	80	68
Canada	64	61
Australia	39	37
Spain	22	15
New Zealand	9	9
Ireland	7	7
Denmark	9	6
Japan	17	6
Italy	11	5
Hong Kong	13	4
South Korea	9	4
China	14	3
South Africa	3	3
Belgium	2	2
Czech Republic	3	2
India	12	2
Mexico	11	2
Aruba	1	1
Chile	1	1
Georgia	1	1
Greece	1	1
Hungary	2	1
Iceland	2	1
Iran	4	1
Peru	1	1
Philippines	1	1
Poland	2	1
Romania	2	1
Russia	3	1
Thailand	4	1
Afghanistan	1	0
Argentina	3	0
Brazil	5	0
Colombia	1	0
Finland	1	0
Indonesia	1	0
Israel	2	0
Netherlands	3	0
Norway	4	0
Sweden	1	0
Taiwan	2	0
	3780	3600

Visual Representation:







Insight: From the above first visualization, we get to know that most popular language movies are English, French, Spanish, Mandarin, German and Japanese.

From second plot, we get to learn that Japanese has comparatively higher mean and median with standard deviation of 1, implying that most of the Japanese movies have higher IMDB scores.

Third plot is created to study why is English the most common language. So according to the output and plot, we can understand that USA has the highest movies released with the total of 2989 movies, out of which 2973 movies are in English as it is the most spoken language in the country.

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.
- Hint: Calculate the average IMDB score for each director. Use Excel's PERCENTILE function to identify the directors with the highest scores. Compare the scores of these directors to the overall distribution of scores.

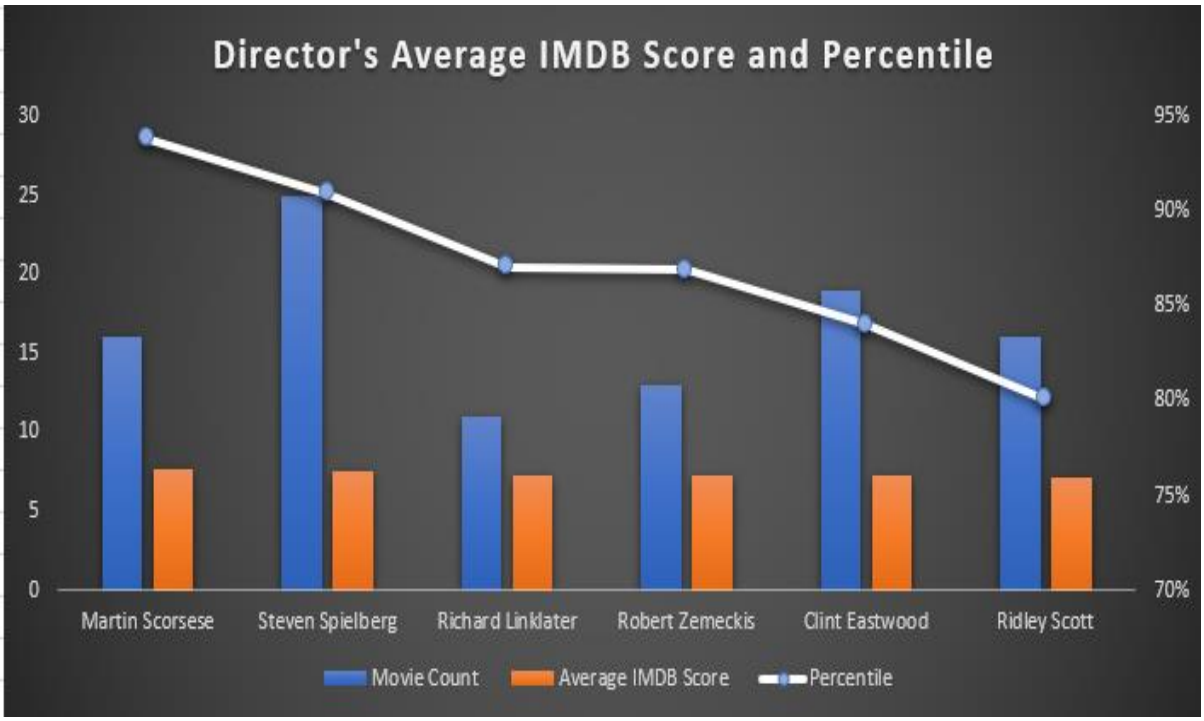
Output:

Director Names	Movie Count	Average IMDB Scor	Percentile
Krzysztof Kieślowski	1	9.1	100%
Charles Chaplin	1	8.6	100%
Tony Kaye	1	8.6	100%
Alfred Hitchcock	1	8.5	100%
Damien Chazelle	1	8.5	100%
Majid Majidi	1	8.5	100%
Ron Fricke	1	8.5	100%
Sergio Leone	3	8.433333333	100%
Christopher Nolan	8	8.425	100%
Asghar Farhadi	1	8.4	99%
Marius A. Markevicius	1	8.4	99%
Richard Marquand	1	8.4	99%
S.S. Rajamouli	1	8.4	99%
Billy Wilder	1	8.3	99%
Fritz Lang	1	8.3	99%
Lee Unkrich	1	8.3	99%
Lenny Abrahamson	1	8.3	99%
Pete Docter	3	8.233333333	99%
Hayao Miyazaki	4	8.225	99%
Quentin Tarantino	8	8.2	99%
George Roy Hill	2	8.2	99%
Elia Kazan	1	8.2	99%
Joshua Oppenheimer	1	8.2	99%
Juan JosÃ© Campanella	1	8.2	99%
Victor Fleming	2	8.15	99%
Milos Forman	3	8.133333333	99%
Akira Kurosawa	2	8.1	98%
David Sington	1	8.1	98%
Je-kyu Kang	1	8.1	98%
Michael Roemer	1	8.1	98%
Michael Wadleigh	1	8.1	98%
Terry George	1	8.1	98%
Tim Miller	1	8.1	98%
William Wyler	1	8.1	98%
David Lean	4	8	98%

Director Names	Movie Count	Average II	Percentile
Martin Scorsese	16	7.675	94%
Steven Spielberg	25	7.544	91%
Richard Linklater	11	7.327273	87%
Robert Zemeckis	13	7.307692	87%
Clint Eastwood	19	7.205263	84%
Ridley Scott	16	7.13125	80%

Filtered the movie count greater than 10 and sorted with largest average IMDB Score and Percentile.

Visual Representation:



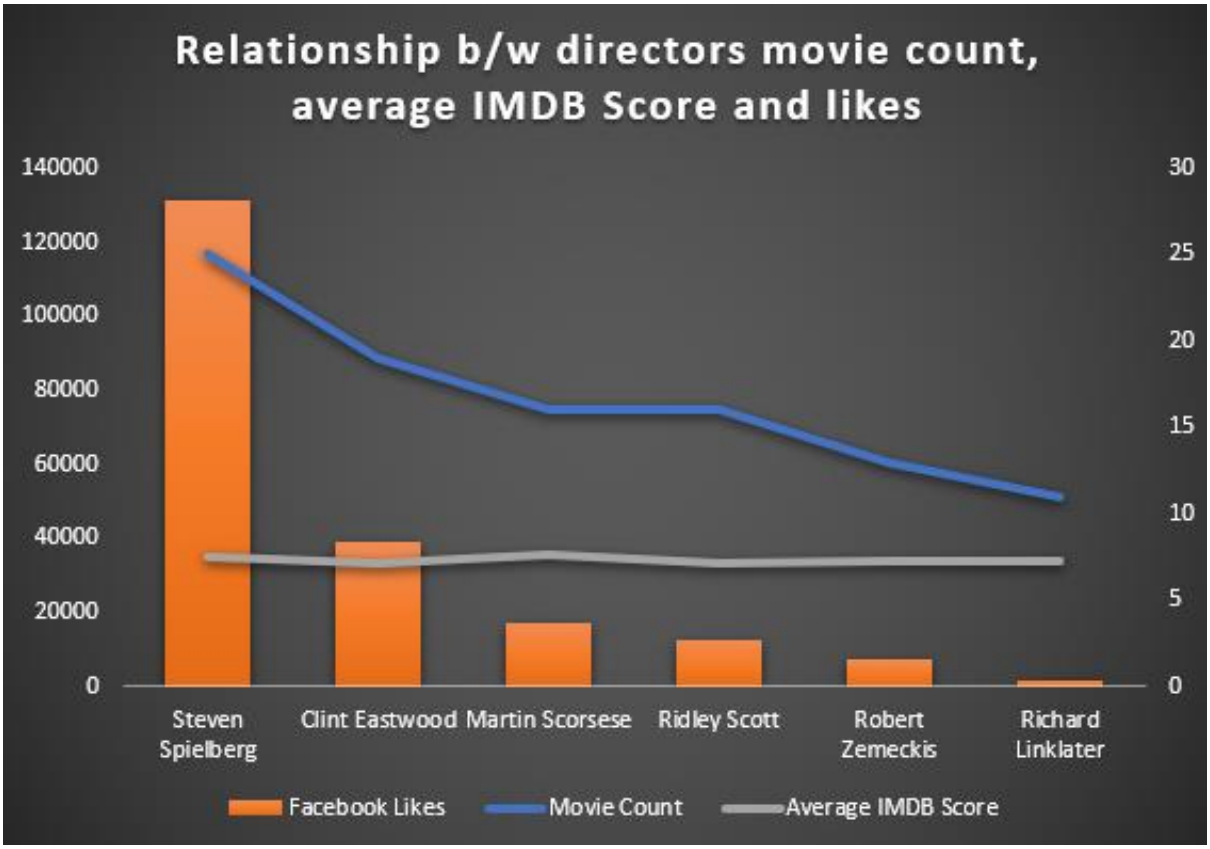
Insight: Top 6 directors are chosen based on condition that they must have **more than 10 movies** and should have IMDB Score greater than 7 and percentile greater or equal to 80%. These conditions help to check for consistency and picks the best ones.

The top 6 directors have high average IMDB score and percentile with more than 10 movies thus reflecting their work and influence on the audience.

To support the above conclusions, I have taken the Facebook likes of each director from internet and created a table as below.

Director Names	Movie Count	Facebook Likes	Average IMDB Score	Percentile
Steven Spielberg	25	131404	7.544	91%
Clint Eastwood	19	39243	7.205263158	84%
Martin Scorsese	16	17000	7.675	94%
Ridley Scott	16	12872	7.13125	80%
Robert Zemeckis	13	7520	7.307692308	87%
Richard Linklater	11	1785	7.327272727	87%

To visually analyse the relationship between the movie counts, average IMDB Score and Facebook likes of a specific director, below chart is plotted.



Observation: More the number of movies, more people recognition for good work, thus more Facebook likes, followers and popularity. Which will eventually result in more viewers and probability of increasing IMDB Score.

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.

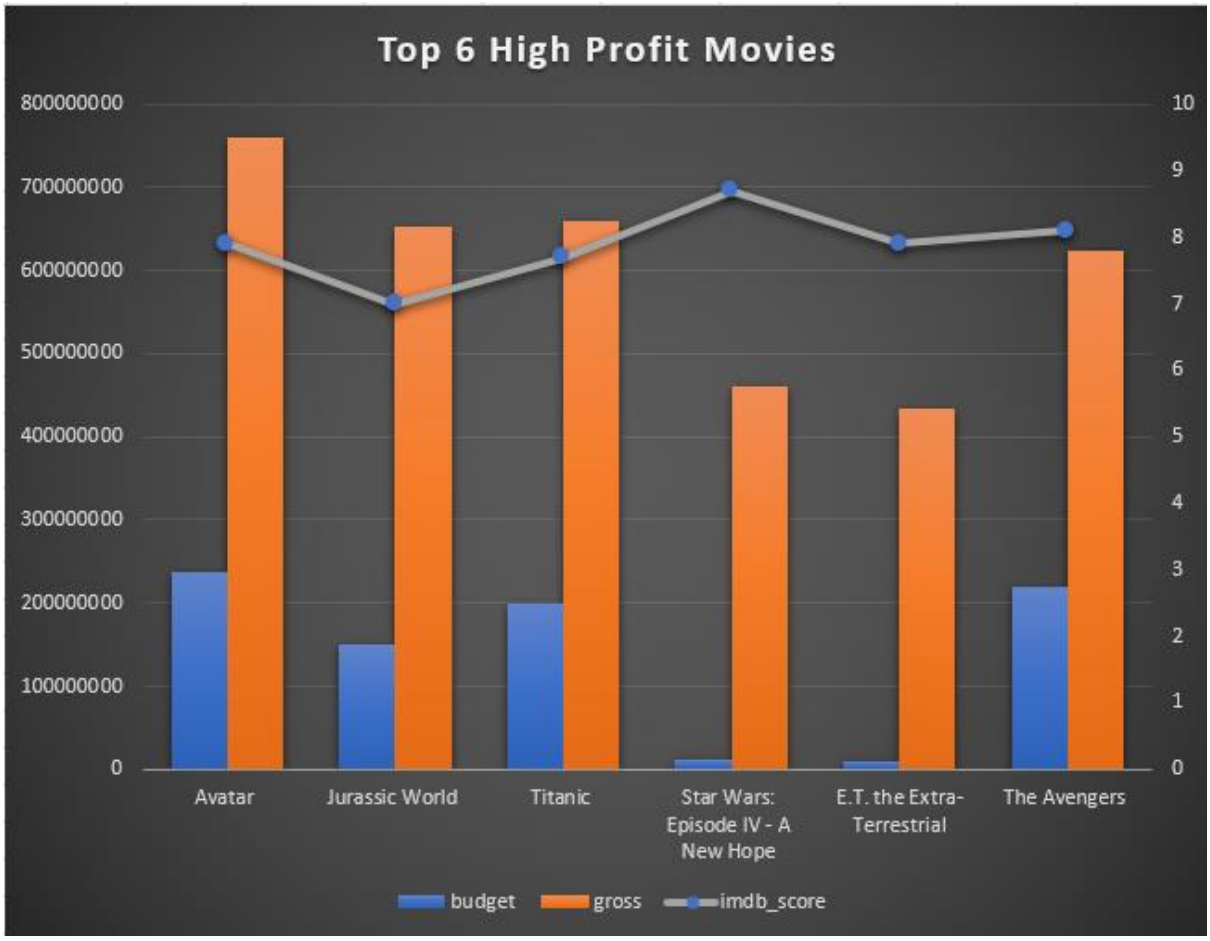
- Hint: Calculate the correlation coefficient between movie budgets and gross earnings using Excel's CORREL function. Calculate the profit margin (gross earnings - budget) for each movie and identify the movies with the highest profit margin using Excel's MAX function.

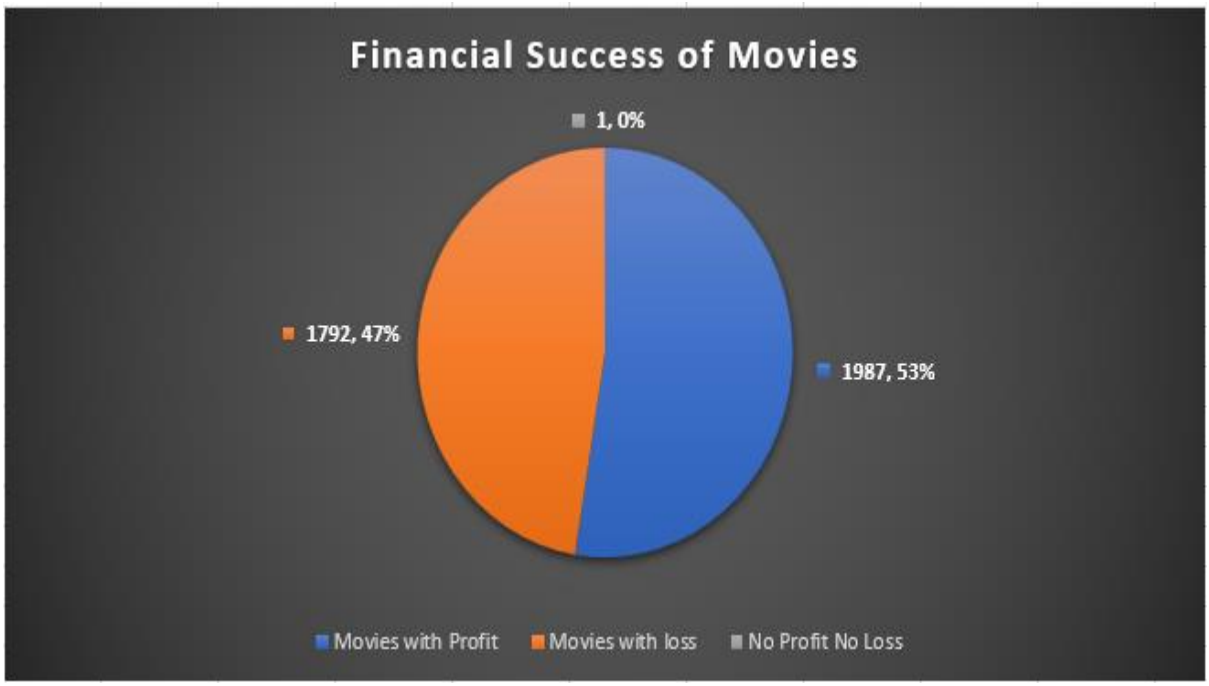
Output:

movie_title	budget	gross	imdb_score	Profit Margin	Max Profit M	Movie Name	Correlation Coefficient
Avatar	237000000	760505847	7.9	523505847	523505847	Avatar	0.096363836
Jurassic World	150000000	652177271	7	502177271			
Titanic	200000000	658672302	7.7	458672302			
Star Wars: Episode IV - A New Hope	11000000	460935665	8.7	449935665			
E.T. the Extra-Terrestrial	10500000	434949459	7.9	424449459			
The Avengers	220000000	623279547	8.1	403279547			

Movies with Profit	Movies with loss	No Profit No Loss	Total Movies
1987	1792	1	3780

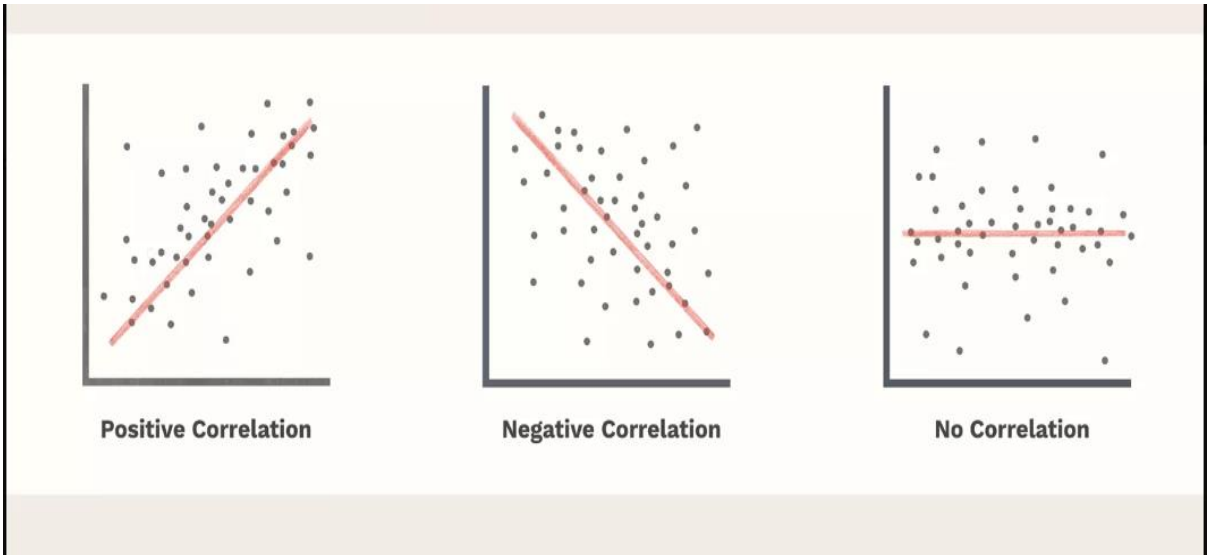
Visual Representation:





Note: and A correlation coefficient is a number between -1 and 1 that tells us the strength and direction of a relationship between variables, tells us how similar the measurements of two or more variables are across a dataset.

- 1 value indicates a **perfect positive correlation** – all data points align in a straight line
- -1 value indicates a **perfect negative/inverse correlation** - all data points align in a straight line
- 0 value indicates **no linear relationship** or a weak correlation.
- Closer to 0 – weaker correlation
- Closer to 1 or -1 – stronger correlation



Insight: The highest profit margin movie is Avatar with a profit of 523505847 USD.

The top 6 movies with highest profits are plotted above and along with their relationship with budget, gross and IMDB Score. From the plot, we understand that **higher the IMDB Score, higher the profit irrespective of budget.**

Therefore, **the correlation between budget and gross is weak with 0.09 value**.

As per the second output and plot, there are 1987 movies with profit and 1792 movies with loss and 1 movie with neither profit nor loss. Therefore, there are more movies with profit than loss.

Result: This project helped me to understand

- The process of data cleaning. First start with deleting duplicate values using "Remove Duplicate" option of Excel. Then drop unwanted columns followed by deleting rows with many missing values. Also check if missing values can be fetched and retained with the help of internet. Finally, look for errors and correct them.
- Helped me understand how to find the outlier and use best strategy to remove or replace or keep them depending on the situation.
- Has made me think and apply logic and use the best out of them with proper explanation so that the audience can relate with it.
- This project helped me to think and dive deep by keeping in mind "**the why approach**" and to keep questioning myself how is the obtained output related or dependant on other variables to **uncover the root cause**.
- How to retrieve/extract the answers and what technique or formulas to use to achieve it.
- Has made me think like an analyst, considering what makes the entire thing perfect with relevant and valuable insights.
- Along with approach and output, visual representation plays a very important role for faster understanding of analysis and this project has made me more confident and aware about various options that are available. I have used combo of various charts this time and have gained a knowledge of when to use which chart that perfectly suits the situation and makes it easy to understand.

Links: My excel worksheet link with different sheets for each task,

https://docs.google.com/spreadsheets/d/1ddSoHI_4oFwEMMunYbHAfee9dUo627it/edit?usp=sharing&ouid=108154584635151678812&rtpof=true&sd=true

You can connect with me on LinkedIn account,

https://www.linkedin.com/in/raksha-nayak-41578738/

Loom Video link,

https://www.loom.com/share/88927cfe821d421691bc0b06a7adc665?sid=5d3937c0-f03b-4907-af50-67cae2da018f