PROJECT 5

Project Description:

Movie Database(IMBD) Also Known as Internet movie Database, Provides us the details related to movie. It helps to Acknowledged about the movie reviews and most importantly the actual ratings based on the users. It gives the details about the characters in the movies and their roles.

The Datasets are given to me, and we need to Analyze the questions given below. We need to Analyze and make beautiful visualization through this data. The Dataset was given name as IMBD MOVIES which consists of 5044 rows and 28 columns. I need to make same good visualization through this data.
As this is IMBD movies analysis, here we are going to use Descriptive analysis To analyze the data.

Descriptive Analysis is the type of analysis of data that helps describe, show or summarize data points in a constructive way such that patterns might emerge that fulfil every condition of the data. It is one of the most important steps for conducting statistical data analysis. It gives you a conclusion of the distribution of your data, helps you detect typos and outliers, and enables you to identify similarities among variables, thus making you ready for conducting further statistical analyses.

Techniques for Descriptive Analysis:

Data aggregation and data mining are two techniques used in descriptive analysis to churn out historical data. In Data aggregation, data is first collected and then sorted to make the datasets more manageable.

1. Descriptive techniques often include constructing tables of quantiles and means, methods of dispersion such as variance or standard deviation, and cross-tabulations or "crosstabs" that can be used to carry out many disparate hypotheses. These hypotheses often highlight differences among subgroups.

2. Measures like segregation, discrimination, and inequality are studied using specialised descriptive techniques. Discrimination is measured with the help of audit studies or decomposition methods. More segregation on the basis of type or inequality of outcomes need not be wholly good or bad in itself, but it is often considered a marker of unjust social processes; accurate measurement of the different steps across space and time is a prerequisite to understanding these processes.

3. A table of means by subgroup is used to show important differences across subgroups, which mostly results in inference and conclusions being made. When we notice a gap in earnings, for example, we naturally tend to extrapolate reasons for those patterns complying.

4. But this also enters the province of measuring impacts which requires the use of different techniques. Often, random variation causes difference in means, and statistical inference is required to determine whether observed differences could happen merely due to chance.

5. A crosstab or two-way tabulation is supposed to show the proportions of components with unique values for each of two variables available, or cell proportions. For example, we might tabulate the proportion of the population that has a high school degree and to receives food or cash assistance, meaning a crosstab of education versus receipt of is supposed to be made. Then we might also want to examine row proportions, or the fractions in each education group who receive food or cash assistance, perhaps seeing assistance levels dip extraordinarily at higher education levels.

6. Column proportions can also be examined, for the fraction of population with different levels of education, but this is the opposite from any causal effects. We might come across a surprisingly high number or proportion of recipients with a college education, but this might be a result of larger numbers of people being college graduates than people who have less than a high school degree.

Advantages of Descriptive Analysis:

1. High degree of objectivity and neutrality of the researchers are one of the main advantages of Descriptive Analysis. The reason why researchers need to be extra vigilant is because descriptive analysis shows different characteristics of the data extracted and if the data doesn't match with the trends then it will lead to major dumping of data.

2. Descriptive analysis is considered to be more vast than other quantitative methods and provide a broader picture of an event or phenomenon. It can use any number of variables or even a single number of variables to conduct descriptive research.

3.This type of analysis is considered as a better method for collecting information that describes relationships as natural and exhibits the world as it exists. This reason makes this analysis very real and close to humanity as all the trends are made after research about the real-life behaviour of the data.

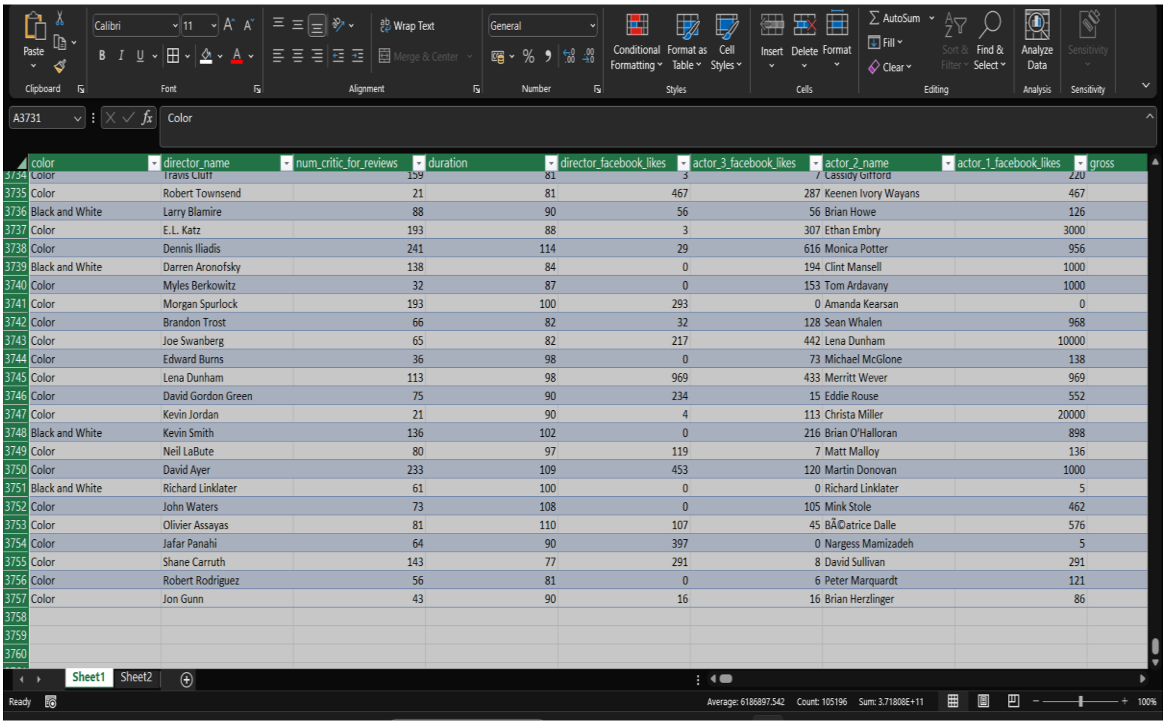
4.It is considered useful for identifying variables and new hypotheses which can be further analyzed through experimental and inferential studies. It is considered useful because the margin for error is very less as we are taking the trends straight from the data properties.

5.This type of study gives the researcher the flexibility to use both quantitative and qualitative data to discover the properties of the population.

Approach:

- ➢ I downloaded the database of IMBD movies.
- ➢ We had been given movies database with 5044 rows and 28 columns.
- ➢ Then I cleaned the data , I checked for null values, duplicates values and then I deleted it and cleaned the data, after cleaning there is only 3757 rows left with me .

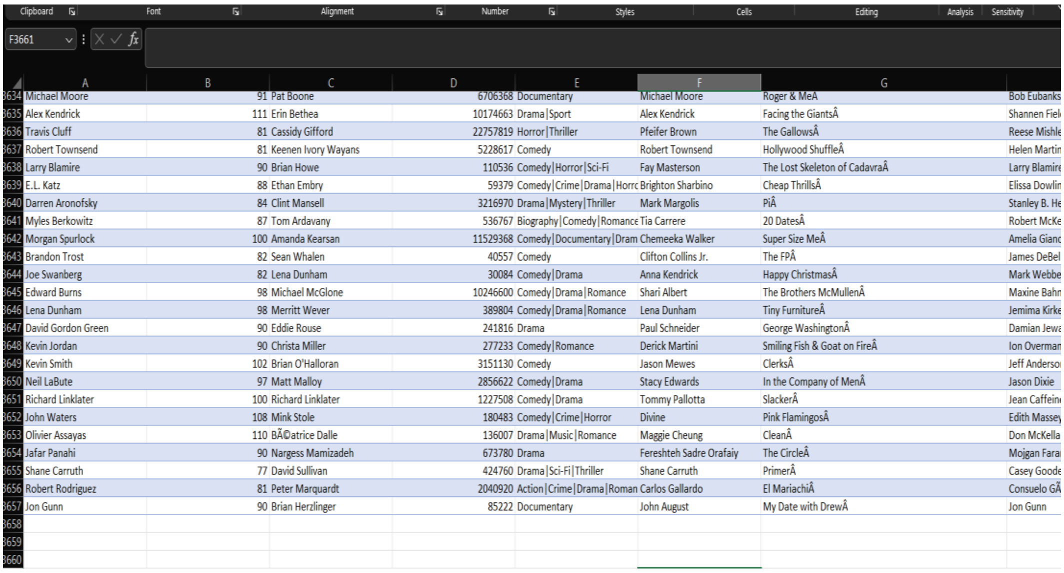
CTRL+A ⟹ CTRL+G ⟹ Specials(blanks) ⟹ DELETE(Delete sheet rows).



	color	director_name	num_critic_for_reviews	duration	director_facebook_likes	actor_3_facebook_likes	actor_2_name	actor_1_facebook_likes	gross
3734	Color	Travis Cluff	159	81	3	7	Cassidy Gifford	220	
3735	Color	Robert Townsend	21	81	467	287	Keenen Ivory Wayans	467	
3736	Black and White	Larry Blamire	88	90	56	56	Brian Howe	126	
3737	Color	E.L. Katz	193	88	3	307	Ethan Embry	3000	
3738	Color	Dennis Iliadis	241	114	29	616	Monica Potter	956	
3739	Black and White	Darren Aronofsky	138	84	0	194	Clint Mansell	1000	
3740	Color	Myles Berkowitz	32	87	0	153	Tom Ardavany	1000	
3741	Color	Morgan Spurlock	193	100	293	0	Amanda Kearsan	0	
3742	Color	Brandon Trost	66	82	32	128	Sean Whalen	968	
3743	Color	Joe Swanberg	65	82	217	442	Lena Dunham	10000	
3744	Color	Edward Burns	36	98	0	73	Michael McGlone	138	
3745	Color	Lena Dunham	113	98	969	433	Merritt Wever	969	
3746	Color	David Gordon Green	75	90	234	15	Eddie Rouse	552	
3747	Color	Kevin Jordan	21	90	4	113	Christa Miller	20000	
3748	Black and White	Kevin Smith	136	102	0	216	Brian O'Halloran	898	
3749	Color	Neil LaBute	80	97	119	7	Matt Malloy	136	
3750	Color	David Ayer	233	109	453	120	Martin Donovan	1000	
3751	Black and White	Richard Linklater	61	100	0	0	Richard Linklater	5	
3752	Color	John Waters	73	108	0	105	Mink Stole	462	
3753	Color	Olivier Assayas	81	110	107	45	BÃ©atrice Dalle	576	
3754	Color	Jafar Panahi	64	90	397	0	Nargess Mamizadeh	5	
3755	Color	Shane Carruth	143	77	291	8	David Sullivan	291	
3756	Color	Robert Rodriguez	56	81	0	6	Peter Marquardt	121	
3757	Color	Jon Gunn	43	90	16	16	Brian Herzlinger	86	
3758									
3759									
3760									

- ➢ Then I cleaned the data , I checked for null values, duplicates values and then I deleted it and cleaned the data, I similarly applied on the columns and then deleted the duplicate data .

- After removing the duplicate movie name and its features . I removed the duplicates using conditional formatting , and now we have 3657 rows with us.



	A	B	C	D	E	F	G	
3634	Michael Moore	91	Pat Boone	6706368	Documentary	Michael Moore	Roger & MeA	Bob Eubanks
3635	Alex Kendrick	111	Erin Bethea	10174663	Drama|Sport	Alex Kendrick	Facing the GiantsÂ	Shannen Fiel
3636	Travis Cluff	81	Cassidy Gifford	22757819	Horror|Thriller	Pfeifer Brown	The GallowsÂ	Reese Mishle
3637	Robert Townsend	81	Keenen Ivory Wayans	5228617	Comedy	Robert Townsend	Hollywood ShuffleÂ	Helen Martin
3638	Larry Blamire	90	Brian Howe	110536	Comedy|Horror|Sci-Fi	Fay Masterson	The Lost Skeleton of CadavraÂ	Larry Blamire
3639	E.L. Katz	88	Ethan Embry	59379	Comedy|Crime|Drama|Horrc	Brighton Sharbino	Cheap ThrillsÂ	Elissa Dowlin
3640	Darren Aronofsky	84	Clint Mansell	3216970	Drama|Mystery|Thriller	Mark Margolis	PiÂ	Stanley B. He
3641	Myles Berkowitz	87	Tom Ardavany	536767	Biography|Comedy|Romance	Tia Carrere	20 DatesÂ	Robert McKe
3642	Morgan Spurlock	100	Amanda Kearsan	11529368	Comedy|Documentary|Dram	Chemeeka Walker	Super Size MeÂ	Amelia Gianc
3643	Brandon Trost	82	Sean Whalen	40557	Comedy	Clifton Collins Jr.	The FPÂ	James DeBel
3644	Joe Swanberg	82	Lena Dunham	30084	Comedy|Drama	Anna Kendrick	Happy ChristmasÂ	Mark Webbe
3645	Edward Burns	98	Michael McGlone	10246600	Comedy|Drama|Romance	Shari Albert	The Brothers McMullenÂ	Maxine Bahn
3646	Lena Dunham	98	Merritt Wever	389804	Comedy|Drama|Romance	Lena Dunham	Tiny FurnitureÂ	Jemima Kirke
3647	David Gordon Green	90	Eddie Rouse	241816	Drama	Paul Schneider	George WashingtonÂ	Damian Jewa
3648	Kevin Jordan	90	Christa Miller	277233	Comedy|Romance	Derick Martini	Smiling Fish & Goat on FireÂ	Ion Overman
3649	Kevin Smith	102	Brian O'Halloran	3151130	Comedy	Jason Mewes	ClerksÂ	Jeff Anderso
3650	Neil LaBute	97	Matt Malloy	2856622	Comedy|Drama	Stacy Edwards	In the Company of MenÂ	Jason Dixie
3651	Richard Linklater	100	Richard Linklater	1227508	Comedy|Drama	Tommy Pallotta	SlackerÂ	Jean Caffein
3652	John Waters	108	Mink Stole	180483	Comedy|Crime|Horror	Divine	Pink FlamingosÂ	Edith Massey
3653	Olivier Assayas	110	BÃ©atrice Dalle	136007	Drama|Music|Romance	Maggie Cheung	CleanÂ	Don McKella
3654	Jafar Panahi	90	Nargess Mamizadeh	673780	Drama	Fereshteh Sadre Orafaiy	The CircleÂ	Mojgan Farai
3655	Shane Carruth	77	David Sullivan	424760	Drama|Sci-Fi|Thriller	Shane Carruth	PrimerÂ	Casey Goode
3656	Robert Rodriguez	81	Peter Marquardt	2040920	Action|Crime|Drama|Roman	Carlos Gallardo	El MariachiÂ	Consuelo GÃ
3657	Jon Gunn	90	Brian Herzlinger	85222	Documentary	John August	My Date with DrewÂ	Jon Gunn
3658								
3659								
3660								

- **I dropped down the some columns which are not useful in analysis.**
  - Color
  - num_critic_for_reviews
  - director_facebook_like
  - actor_3_facebook_likes
  - actor_1_facebook_likes
  - num_voted_users
  - cast_total_facebook_likes
  - actor_2_facebook_likes
  - movie_facebook_likes

Tech-Stack Used:

- Microsoft Excel 365(2023).

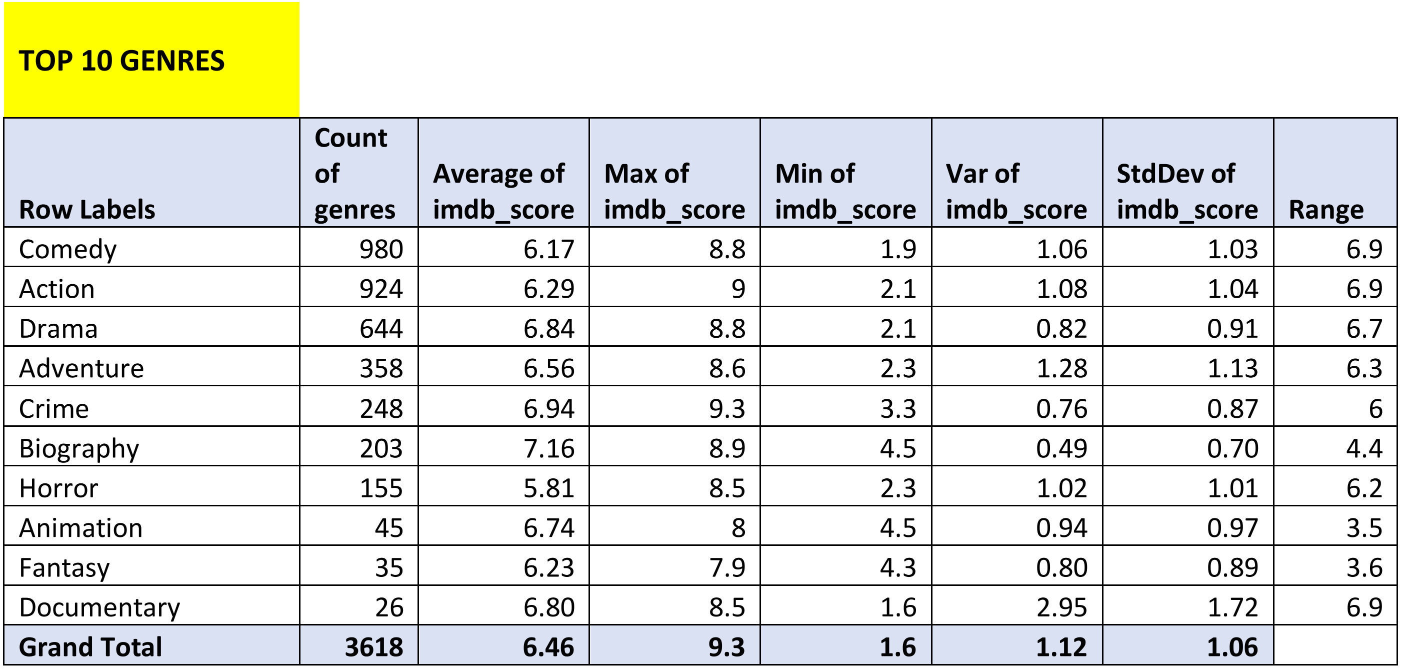
Insights:

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.



TOP 10 GENRES

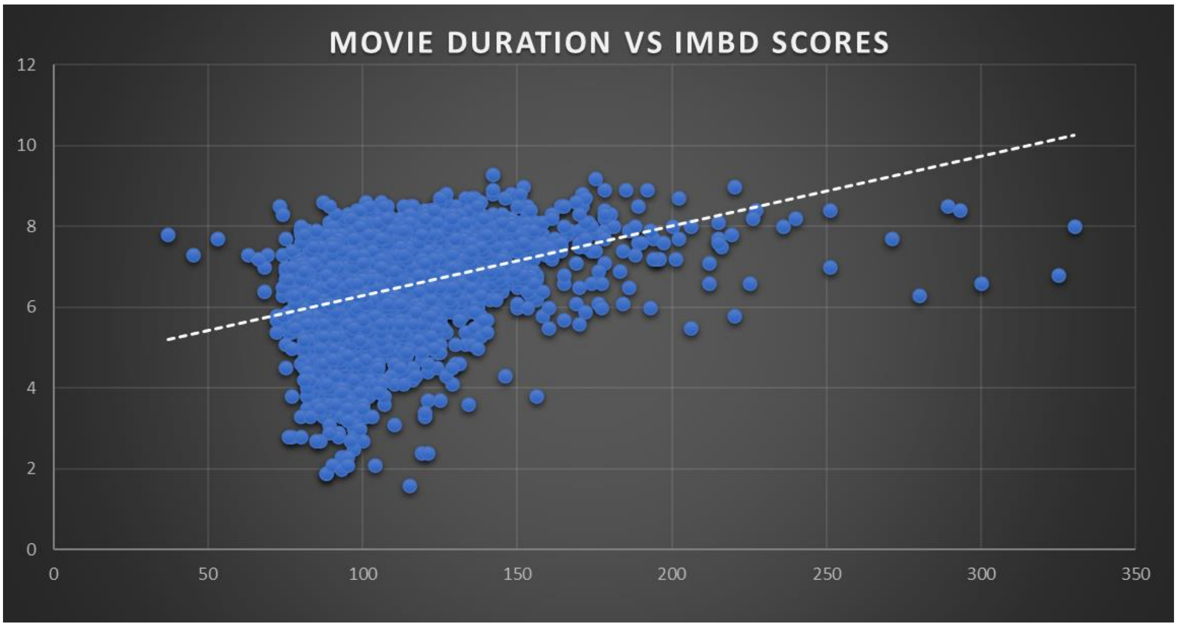
Row Labels	Count of genres	Average of imdb_score	Max of imdb_score	Min of imdb_score	Var of imdb_score	StdDev of imdb_score	Range
Comedy	980	6.17	8.8	1.9	1.06	1.03	6.9
Action	924	6.29	9	2.1	1.08	1.04	6.9
Drama	644	6.84	8.8	2.1	0.82	0.91	6.7
Adventure	358	6.56	8.6	2.3	1.28	1.13	6.3
Crime	248	6.94	9.3	3.3	0.76	0.87	6
Biography	203	7.16	8.9	4.5	0.49	0.70	4.4
Horror	155	5.81	8.5	2.3	1.02	1.01	6.2
Animation	45	6.74	8	4.5	0.94	0.97	3.5
Fantasy	35	6.23	7.9	4.3	0.80	0.89	3.6
Documentary	26	6.80	8.5	1.6	2.95	1.72	6.9
Grand Total	**3618**	**6.46**	**9.3**	**1.6**	**1.12**	**1.06**	

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.

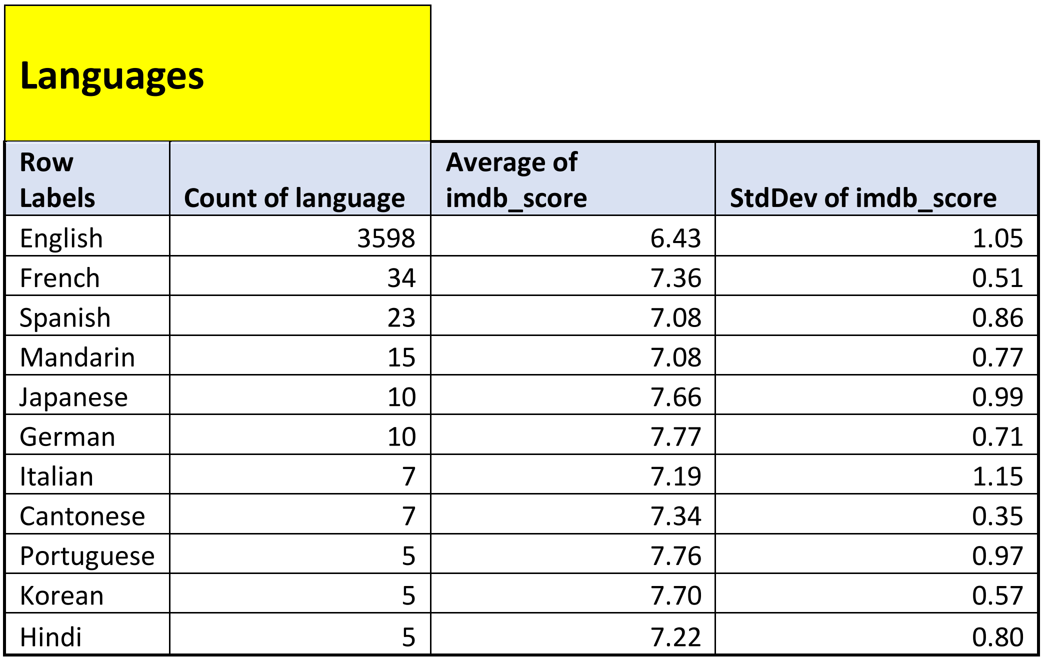


C. **Language Analysis**:

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.



Languages

Row Labels	Count of language	Average of imdb_score	StdDev of imdb_score
English	3598	6.43	1.05
French	34	7.36	0.51
Spanish	23	7.08	0.86
Mandarin	15	7.08	0.77
Japanese	10	7.66	0.99
German	10	7.77	0.71
Italian	7	7.19	1.15
Cantonese	7	7.34	0.35
Portuguese	5	7.76	0.97
Korean	5	7.70	0.57
Hindi	5	7.22	0.80

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.

TOP 10 DIRECTORS

Directors	movies	IMBD RATINGS
Akira Kurosawa	Seven SamuraiÂ	8.70
Tony Kaye	American History XÂ	8.60
Charles Chaplin	Modern TimesÂ	8.60
Alfred Hitchcock	PsychoÂ	8.50
Ron Fricke	SamsaraÂ	8.50
Majid Majidi	Children of HeavenÂ	8.50
Damien Chazelle	WhiplashÂ	8.50
Sergio Leone	The Good, the Bad and the UglyÂ	8.43
Christopher Nolan	The Dark KnightÂ	8.43
Richard Marquand	Star Wars: Episode VI - Return of the JediÂ	8.40
Asghar Farhadi	A SeparationÂ	8.40

D. Budget Analysis:

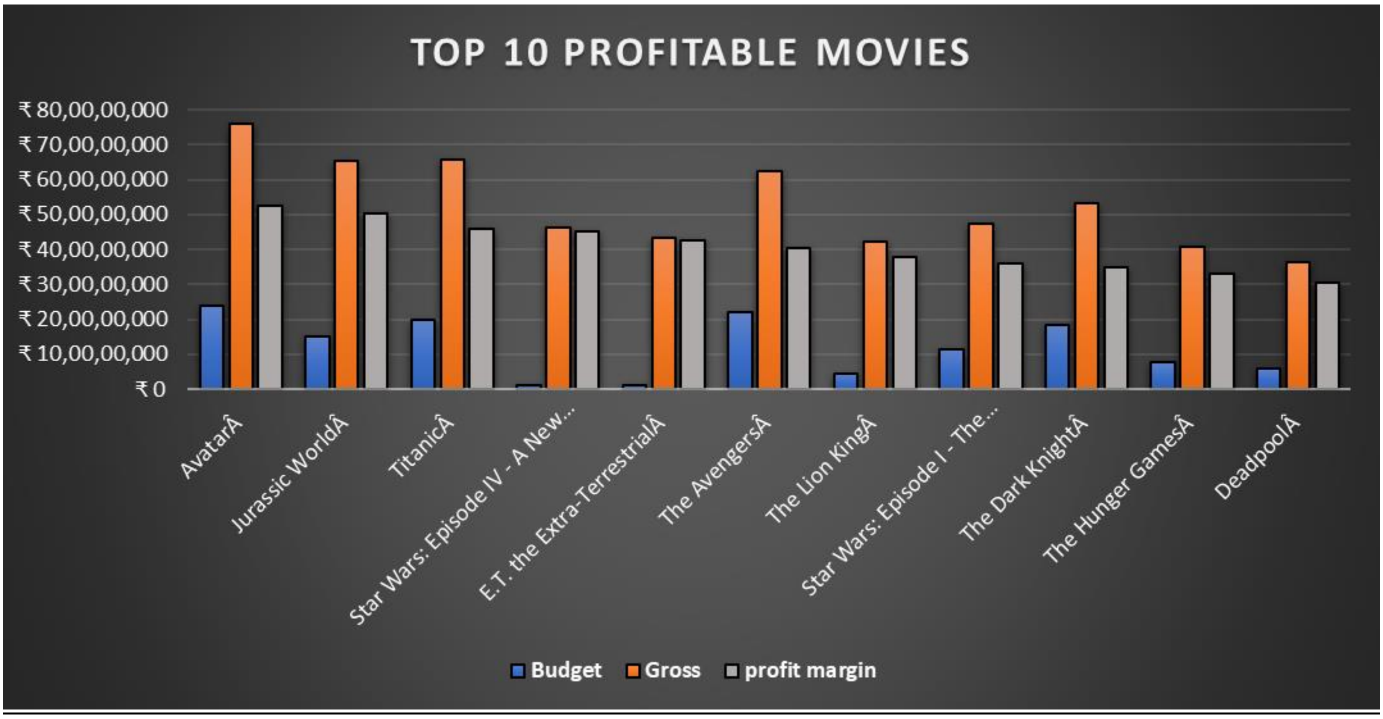
Explore the relationship between movie budgets and their financial success.

- 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.

Top 10 movies with higher profit rate

Movies	Budget	Gross	profit margin
AvatarÂ	₹ 23,70,00,000	₹ 76,05,05,847	₹ 52,35,05,847
Jurassic WorldÂ	₹ 15,00,00,000	₹ 65,21,77,271	₹ 50,21,77,271
TitanicÂ	₹ 20,00,00,000	₹ 65,86,72,302	₹ 45,86,72,302
Star Wars: Episode IV - A New HopeÂ	₹ 1,10,00,000	₹ 46,09,35,665	₹ 44,99,35,665
E.T. the Extra-TerrestrialÂ	₹ 1,05,00,000	₹ 43,49,49,459	₹ 42,44,49,459
The AvengersÂ	₹ 22,00,00,000	₹ 62,32,79,547	₹ 40,32,79,547
The Lion KingÂ	₹ 4,50,00,000	₹ 42,27,83,777	₹ 37,77,83,777
Star Wars: Episode I - The Phantom MenaceÂ	₹ 11,50,00,000	₹ 47,45,44,677	₹ 35,95,44,677
The Dark KnightÂ	₹ 18,50,00,000	₹ 53,33,16,061	₹ 34,83,16,061
The Hunger GamesÂ	₹ 7,80,00,000	₹ 40,79,99,255	₹ 32,99,99,255
DeadpoolÂ	₹ 5,80,00,000	₹ 36,30,24,263	₹ 30,50,24,263



RESULT:

I have gained knowledge about descriptive statistics . I solved problems with Average ,mean , median ,mode ,variance ,standard deviation. I worked on outliers and other Statistical implementations.
I used my Technical, visualization and statistical knowledge to complete this project. After using my statistical I came with lot of challenges to deliver that data into visualization terms. I worked with different charts and different data insights modules. Here, I learned to mange them all and add some meaningful data with charts and other correlations of profit margins.
I have used Microsoft Excel 365 to solve the given problems. I have came up with solutions finding in details and solving each of the problem with my technical and visualization knowledge.

VIDEO LINK : https://www.loom.com/share/0232a95810c04ed7a80a67a2e6d2c23b?sid=8b953297-6636-4e3b-ad42-d9c3c604f9a1

EXCEL SHEET LINK:
https://docs.google.com/spreadsheets/d/114kTlnzgewbTBucRUvXFhVukFLXAhBbs/edit?usp=drive_link&ouid=113818516476537685883&rtpof=true&sd=true

**(Download the file in .xlsx format, otherwise features will not be displayed.)

THANK YOU.