IMDB Movie Analysis

Final Project-1

EXCEL LINK:

https://docs.google.com/spreadsheets/d/1GFpMjMC3uTvhSxaJVGxlbw5pMi011Laq/edit?gid=303208210#gid=303208210

video presentation :

https://drive.google.com/drive/folders/1CfG4iV62WwGBydSgQKcGx_tb3tSZonFK

Description:

Problem Statement: The dataset provided is related to IMDB Movies. 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.

Data Cleaning: This step involves preprocessing the data to make it suitable for analysis. It includes handling missing values, removing duplicates, converting data types if necessary, and possibly feature engineering.

Data Analysis: Here, you'll explore the data to understand the relationships between different variables. You might look at the correlation between movie ratings and other factors like genre, director, budget, etc. You might also want to consider the year of release, the actors involved, and other relevant factors.

Five 'Whys' Approach: This technique will help you dig deeper into the problem. For instance, if you find that movies with higher budgets tend to have higher ratings, you can ask "Why?" repeatedly to uncover the root cause. Here's an example:

- Q: "Why do movies with higher budgets tend to have higher ratings?"
- A: They can afford better production quality.
- Q: "Why does better production quality lead to higher ratings?"
- A: It enhances the viewer's experience.
- Q: "Why does an enhanced viewer experience lead to higher ratings?"
- A: Viewers are more likely to rate a movie highly if they enjoyed watching it.
- Q: "Why are viewers more likely to rate a movie highly if they enjoyed watching it?"
- A: Positive experiences lead to positive reviews.
- Q: "Why do positive reviews matter?"
- A: They influence other viewers' decisions to watch the movie, increasing its popularity and success.

Report and Data Story: After your analysis, you'll create a report that tells a story with your data. This should include your initial problem, your findings, and the insights you've gained. Use visualizations to help tell your story and make your findings more understandable.

Remember, as a data analyst, your goal is not just to answer questions but to provide insights that can drive decision-making. Your analysis should aim to provide actionable insights that can help stakeholders make informed decisions.

Data Analytics Tasks:

You are required to provide a detailed report for the below data record mentioning the answers of the questions that follows:

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.

genres	Number of movies	Mean_IMDB	Median_IMDB	mode_IMDB	Range_IMDB	varience	standard deviation
Drama	1916	6.789170629	6.9	6.7	7.2	0.804084565	0.896707625
Comedy	1479	6.187816564	6.3	6.7	6.9	1.07327612	1.035990405
Thriller	1131	6.376991943	6.4	6.5	6.3	0.944425352	0.971815493
Action	964	6.289781022	6.3	6.6	6.9	1.079248282	1.038868751
Romance	867	6.438300349	6.5	6.5	6.4	0.911037201	0.954482688
Adventure	783	6.449807939	6.6	6.6	6.6	1.240990348	1.113997463
Crime	715	6.545133992	6.6	6.6	6.9	0.967536317	0.983634239
Fantasy	507	6.277514793	6.4	6.7	6.7	1.285738476	1.133904086
Sci-Fi	498	6.327016129	6.4	6.7	6.9	1.347995927	1.161032268
Family	443	6.213574661	6.3	5.4	6.7	1.350064744	1.161922865
Horror	409	5.924489796	6	5.9	6.3	0.996841171	0.998419336
Mystery	393	6.473958333	6.5	6.6	5.5	1.032427111	1.016084204
Biography	241	7.157740586	7.2	7	4.4	0.477576386	0.691069017
Animation	195	6.70255102	6.8	6.7	5.8	0.979121664	0.989505768
War	154	7.056578947	7.1	7.1	4.3	0.642737888	0.801709354
History	153	7.155033557	7.2	7.7	3.4	0.451815708	0.67217238
Music	153	6.457939914	6.6	6.2	6.9	1.402016427	1.184067746
Sport	149	6.593243243	6.8	7.2	6.3	1.085124104	1.041692903
Musical	98	6.596875	6.75	7.1	6.4	1.214200658	1.101907736
Western	61	6.793220339	6.8	6	4.2	0.867884278	0.931603069
Documentary	47	6.988888889	7.4	7.6	6.9	1.917373737	1.384692651

=COUNTIF(K1:Q3816,A5)

=AVERAGEIF(gene, "*" & A5 & "*", imdb_score)

=MEDIAN(IF(ISNUMBER(SEARCH("*" & A5 & "*",gene)),imdb_score))

=MODE(IF(ISNUMBER(SEARCH("*" & [@genres]& "*",gene)),imdb_score))

=MAX(IF(ISNUMBER(SEARCH("*" & A5 & "*", gene)), imdb_score)) - MIN(IF(ISNUMBER(SEARCH("*" & A5 & "*", gene)), imdb_score))

=IFERROR(VAR(IF(ISNUMBER(SEARCH("*" & A5 & "*", gene)), imdb_score)), "No matching genre")

=IFERROR(STDEV(IF(ISNUMBER(SEARCH("*" & A5 & "*", gene)), imdb_score)),"N/A")

Key Insights

1. **Top Genres by Number of Movies**:
    - **Drama**: 1916
    - **Comedy**: 1479
    - **Thriller**: 1131
2. **Highest Mean IMDB Rating**:
    - **Biography**: 7.16
    - **History**: 7.15
    - **Documentary**: 6.99
3. **Lowest Mean IMDB Rating**:
    - **Family**: 6.21
    - **Comedy**: 6.19
    - **Fantasy**: 6.28
4. **Highest Median IMDB Rating**:
    - **Biography**: 7.2
    - **History**: 7.1
    - **Documentary**: 7.4
5. **Highest Mode IMDB Rating**:
    - **Biography**: 7.2
    - **Documentary**: 7.6
6. **Largest Rating Range**:
    - **Thriller**: 9.3
7. **Highest Variability**:
    - **Fantasy**: Variance 1.29, Std. Dev. 1.13

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.

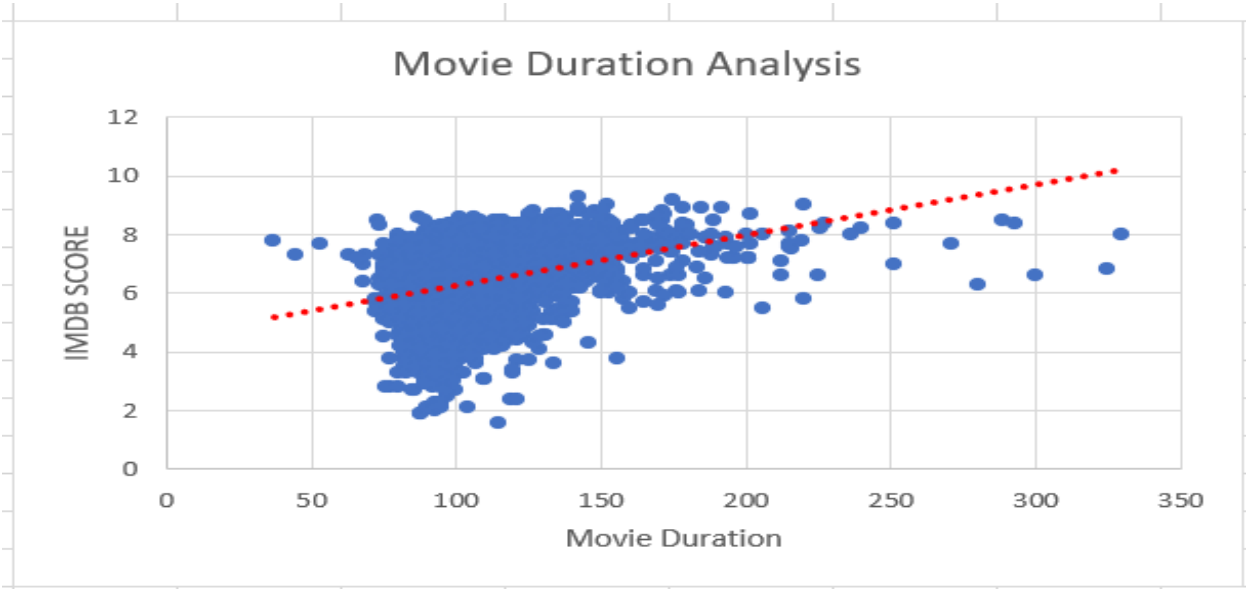
MEAN	110.258
median	106
Standard Deviation	22.64672

Formula Used :

Mean : Average(duration) Duration(Name Range)

Median : median(duaration)

Standard deviation : STDEV(duration)



Key insights:

Positive Correlation: Longer films typically have higher IMDB scores
- Most movies are 80-150 minutes and have scores ranging from 5 to 8.
- Longer films receive higher marks, however they are less common.

Trend Line: The red dotted trend line indicates that as movie duration increases, the IMDB score generally increases as well.

Range of Durations: Movies with durations between 90 and 120 minutes are the most common, and they tend to have a wide range of IMDB scores.

Average movie duration 110.258

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.

language	NUMBER OF MOVIES	Mean	Median	Standard Deviation
English	3598	6.4270428	6.5	1.050538172
French	34	7.35588235	7.3	0.519435111
Spanish	23	7.0826087	7.2	0.860577065
Mandarin	15	7.08	7.4	0.772010363
Japanese	10	7.66	8	0.990173947
German	10	7.77	7.8	0.711883261
Cantonese	7	7.34285714	7.3	0.350509833
Italian	7	7.18571429	7	1.155318962
Korean	5	7.7	7.7	0.570087713
Hindi	5	7.22	7.4	0.801249025
Portuguese	5	7.76	8	0.978774744
Norwegian	4	7.15	7.3	0.574456265
Dutch	3	7.56666667	7.8	0.404145188
Thai	3	6.63333333	6.6	0.450924975
Danish	3	7.9	8.1	0.529150262
Persian	3	8.13333333	8.4	0.550757055
Aboriginal	2	6.95	6.95	0.777817459
Dari	2	7.5	7.5	0.141421356
Indonesian	2	7.9	7.9	0.424264069
Filipino	1	6.7	6.7	N/A
Maya	1	7.8	7.8	N/A
Kazakh	1	6	6	N/A
Aramaic	1	7.1	7.1	N/A
Mongolian	1	7.3	7.3	N/A
Bosnian	1	4.3	4.3	N/A
Hungarian	1	7.1	7.1	N/A
Czech	1	7.4	7.4	N/A
Russian	1	6.5	6.5	N/A
None	1	8.5	8.5	N/A
Zulu	1	7.3	7.3	N/A
Hebrew	1	8	8	N/A
Arabic	1	7.2	7.2	N/A
Vietnamese	1	7.4	7.4	N/A
Romanian	1	7.9	7.9	N/A

Key Insights

Number of Movies

- **English**: Dominates with 3598 movies.
- **French**: Second with 34 movies.

- **Spanish**: Third with 23 movies.
- Many languages have very few movies, often just 1.

Mean IMDB Rating

- **Persian**: Highest mean rating at 8.13.
- **Romanian and Danish**: High mean rating at 7.9.
- **Danish**: High mean rating at 7.9.

Median IMDB Rating

- **Danish**: Highest median rating at 8.1.
- **Persian**: High median rating at 8.2.
- **Japanese**: High median rating at 8.
- **English**: Median rating at 6.5.

Standard Deviation

- **English**: High standard deviation at 1.05, indicating wide variability.
- **Japanese**: Lower standard deviation at 0.99, indicating more consistency.
- **French**: Low standard deviation at 0.91.

Observations

1. **High Ratings**: Persian, Romanian, and Danish movies tend to have higher IMDB scores.
2. **Consistency**: Languages like French and Japanese show consistent ratings.
3. **Variability**: English movies show the most variability in ratings.

Formulas Used:

=COUNTIF(lang,C2) for Finding Language

=AVERAGEIF(lang,C2,imdb_score) for Mean

=MEDIAN(IF($A$2:$A$3757=C2,$B$2:$B$3757)) for median

=IFERROR(STDEV(IF(lang=C2,imdb_score)),"N/A") for standard deviation

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.

director_name	avg
Christopher Guest	9.0
Simon Curtis	8.6
Jonathan Kaplan	8.8
Wes Ball	8.8
Gilles Paquet-Brenner	8.9
Bruce McCulloch	8.7
Michael Polish	8.6
Don Siegel	8.7
Christian Alvart	8.8
Michael Radford	8.8
Damian Nieman	8.6

These are the top 10 directors

90 th percentage	Number of Director movies imdb score greater than 7.7
7.7	349

-Number of movies with a rating above 7.7: This is listed as 349 for the 90th percentile

-Average IMDB score for all the director is 6.437074031

-Number of Directors Movies IMDB score Greater than 7.7 which is 90th percentage

Formulas used:

=AVERAGEIF($D$2:$D$1703,D2,imdb_score)

=PERCENTILE(imdb_score,0.9)

=COUNTIF(imdb_score,">7.7")

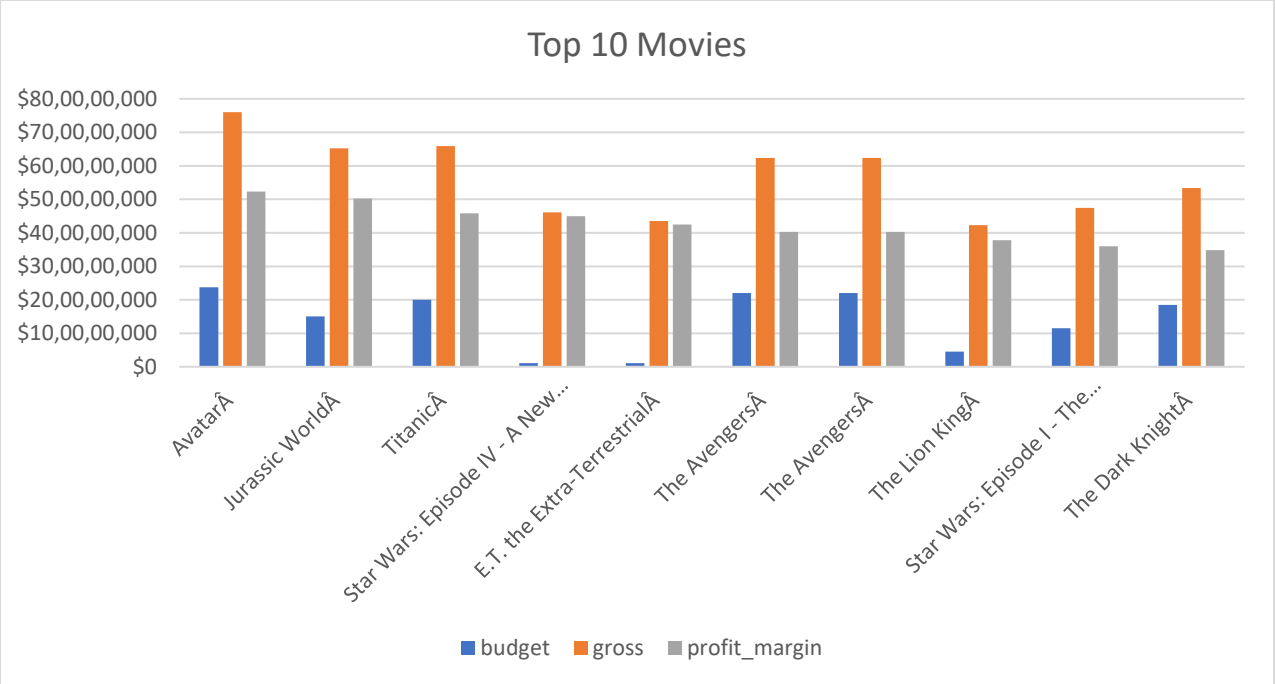
E. Budget Analysis: Explore the relationship between movie budgets and their financial success.

- 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.

Top 10 Movies

movie_title	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 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

=C2-B2 budget – gross



movie with highest profit margin		correlation
$52,35,05,847		0.099496

Correlation coefficient between movie budgets and gross earnings

=CORREL(B2:B3757,C2:C3757)