Project 5- IMDB Movie Analysis

By Visharad Singh Chauhan

Project 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.

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.

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.

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.

Remember, these tasks are designed to progressively explore different aspects of the dataset and uncover meaningful insights. Each task builds upon the previous one to provide a comprehensive analysis of the IMDB movie data.

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.

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.

To find Mean, Median and Standard Deviation we use following formulas:

=AVERAGE(C:C)

=MEDIAN(C:C)

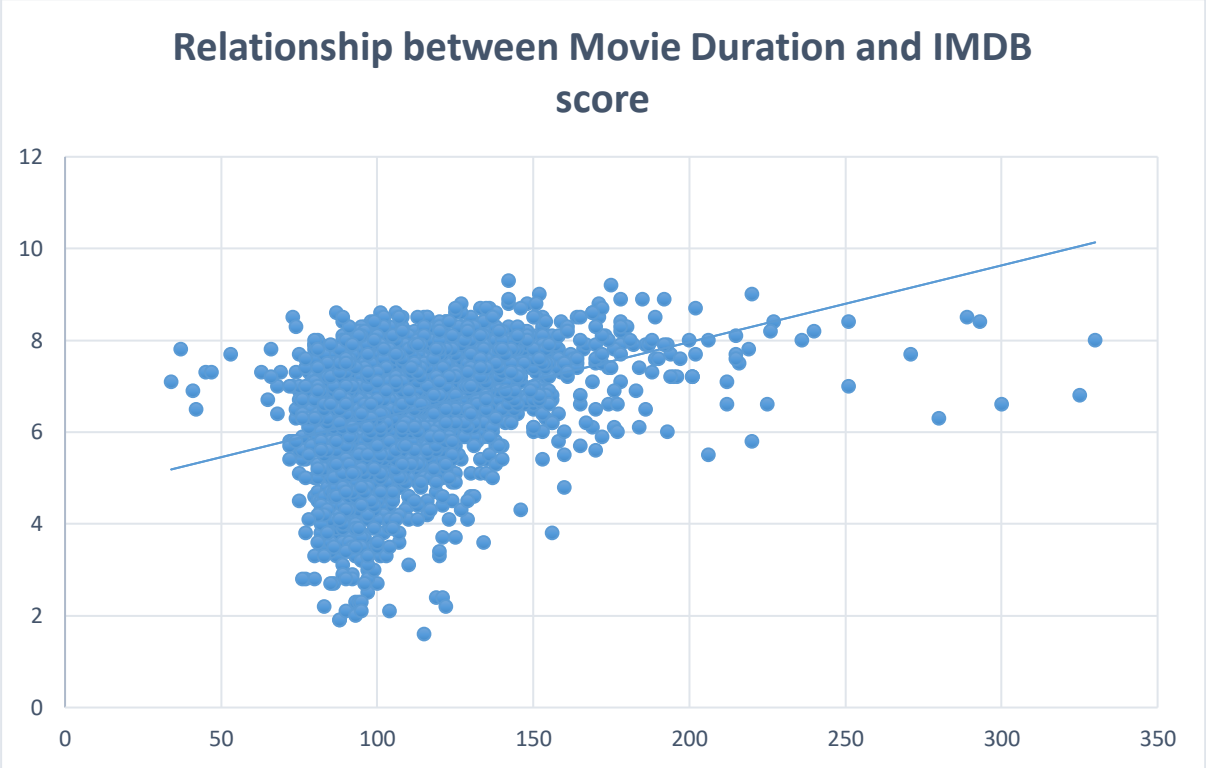
=STDEV(C:C)

We will get the following results after entering these formulas:

Mean	Median	StDev
109.7018	105	22.74435

For visualizing the relationship between movie duration and IMDB score using scatter plot we will select both movie durations and IMDB score

Click on the Insert Chart Option and select Scatter Plot, enable the option show trend line to show the trendline of the chart.



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.

1. First, we will create a pivot table for our table IMDB_movies,

Note: Check the option **Add this to data model**

2. Add Language to the rows section, add movie_title to values section as COUNT, add imdb_score to values section and set value settings to AVERAGE and STDEVP.
3. For Median, add a formula for median in the range section find the median.

This will give us the resultant table:

Row Labels	Count of movie_title	Average of imdb_score	StdDevp of imdb_score	Median IMDB_Score
(blank)	1	5.3	0	5.30000
Aboriginal	2	6.95	0.55	6.95000
Arabic	1	7.2	0	7.20000
Aramaic	1	7.1	0	7.10000
Bosnian	1	4.3	0	4.30000
Cantonese	8	7.2375	0.412121038	7.30000
Czech	1	7.4	0	7.40000
Danish	3	7.9	0.43204938	8.10000
Dari	2	7.5	0.1	7.50000
Dutch	3	7.566666667	0.329983165	7.80000
Dzongkha	1	7.5	0	7.50000
English	3716	6.412540366	1.061062948	6.50000
Filipino	1	6.7	0	6.70000
French	37	7.286486486	0.553691378	7.20000
German	13	7.692307692	0.615769111	7.70000
Hebrew	3	7.5	0.355902608	7.30000
Hindi	10	6.76	1.05470375	7.05000
Hungarian	1	7.1	0	7.10000
Icelandic	1	6.9	0	6.90000
Indonesian	2	7.9	0.3	7.90000
Italian	7	7.185714286	1.069617517	7.00000
Japanese	12	7.625	0.861321659	7.80000
Kazakh	1	6	0	6.00000
Korean	5	7.7	0.509901951	7.70000
Mandarin	14	7.021428571	0.737930089	7.25000
Maya	1	7.8	0	7.80000
Mongolian	1	7.3	0	7.30000
None	1	8.5	0	8.50000
Norwegian	4	7.15	0.497493719	7.30000
Persian	3	8.133333333	0.449691252	8.40000
Portuguese	5	7.76	0.875442745	8.00000
Romanian	1	7.9	0	7.90000
Russian	1	6.5	0	6.50000
Spanish	26	7.05	0.810151933	7.15000
Swedish	1	7.6	0	7.60000
Telugu	1	8.4	0	8.40000
Thai	3	6.633333333	0.368178701	6.60000
Vietnamese	1	7.4	0	7.40000
Zulu	1	7.3	0	7.30000
Grand Total	**3897**	**6.452527585**	**1.065862573**	**6.60000**

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.

To calculate top directors

1. We will create a pivot table of the table imdb_movies.
2. We will put column director_name in rows section and imdb_score in the values section and change it's value settings to AVERAGE.
3. Now we will filter only top 10 from the directors list and sort the data from largest to smallest on the basis of average_of_imdb_scores.

We will get the following table as a result

Row Labels	Average of imdb_score
Tony Kaye	8.6
Charles Chaplin	8.6
Alfred Hitchcock	8.5
Ron Fricke	8.5
Damien Chazelle	8.5
Majid Majidi	8.5
Sergio Leone	8.433333333
Christopher Nolan	8.425
S.S. Rajamouli	8.4
Richard Marquand	8.4
Marius A. Markevicius	8.4
Asghar Farhadi	8.4
Grand Total	**8.452380952**

In this table, we can see that **Tony Kaye**, **Charles Chaplin** and **Alfred Hitchcock** has contributed to the success of movies in the film industry

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.

To find Profit margin,

1. First, we will create a new column "profit" in the cleaned dataset and insert the formula to find the profit ie; **=D2-L2** and extend it to the last row to find profit margin for all the movies.
2. For the Correlation Coefficient, we will use the formula, **=CORREL()** to find it.

It will give us the following result:

Correl Coeff
0.100850218

To find movies with the highest profit margin, we will,

1. Create a pivot table and put movie_title in the rows section and Profit column in the Values section.
2. To filter only top movies, we will change the sorting to Top 10 and we will change the value settings to MAX to show only maximum ones.

It will give us the following result:

Row Labels	Max of Profit margin
AvatarÂ	523505847
E.T. the Extra-TerrestrialÂ	424449459
Jurassic WorldÂ	502177271
Star Wars: Episode I - The Phantom MenaceÂ	359544677
Star Wars: Episode IV - A New HopeÂ	449935665
The AvengersÂ	403279547
The Dark KnightÂ	348316061
The Hunger GamesÂ	329999255
The Lion KingÂ	377783777
TitanicÂ	458672302
Grand Total	**523505847**

The resultant table shows that **Avatar**, **E.T. the Extra-Terrestrial** and **Jurassic World** are the top 3 movies with highest profit margin in the industry.

Here is the link for the complete dataset including the old questions as well

Click here for the datasets

Tech Stack Used- MS Excel 2016