IMBD Movie Analysis

- Rukmini Annadata
rukmini2k2@gmail.com

Project Description:

IMBD is a globally recognized platform for movie and series ratings, catering to both users and critics. It offers comprehensive information about movies and series, including their ratings, director and actor profiles, as well as financial data. In this project, we have access to IMBD's dataset spanning the years 1920 to 2010, which contains a wealth of information about movies, their cast and crew, budgets, box office collections, and more.

Our objective in this project is to clean and preprocess the IMBD dataset and extract valuable insights by applying the Five Whys method for analytics. We will be utilizing Office 365 Excel to perform data analysis, identify underlying issues or trends, and answer specific questions related to the dataset. Through this process, we aim to uncover meaningful patterns and gain a deeper understanding of the movie industry during the specified time period.

Approach:

In this project, our initial step involves comprehending the provided dataset. Subsequently, we will meticulously clean the data by eliminating null values, discarding superfluous columns, and ensuring it meets our specific requirements. Following data cleaning, we will employ pivot tables, diverse functions, and charts to extract the desired insights and answers to our questions. Our analytical approach will employ the Five Whys method, delving deeper into the data to uncover the root causes and underlying trends. To present our findings effectively, we will format our results in tables and graphical representations, providing a comprehensive and visually appealing analysis of the dataset.

Tech-Stack Used:

In the IMBD Movie Analytics project, Microsoft Excel from the Office 365 suite played a pivotal role. The Office 365 suite, developed by Microsoft Corporation, offers a wide range of products designed to enhance productivity across various tasks and data management needs. Microsoft Excel, a prominent component of this suite since its inception in 1985, has been widely adopted in both business and personal contexts for tasks such as data organization, analysis, and visualization. Additionally, Excel supports Visual Basic for Applications (VBA), enabling the creation and execution of macros for extended functionality and automation in data processing and analysis.

Insights:

Cleaning the data: This is one of the most important steps to perform before moving forward with the analysis. Use your knowledge learned till now to do this. (Dropping columns, removing null values, etc.)

Your task: Clean the data.

Dropping unnecessary columns.

(Color, director_facebook_likes, actor_3_facebook_likes,

actor_2_name, actor_1_facebook_likes, cast_total_facebook_likes, actor_3_name, facenumber_in_posts, plot_keywords, movie_imdb_link, content_rating, actor_2_facebook_likes, aspect_ratio, movie_facebook_likes)

Remove Blank Cell / Null Value.

Removing Duplicate.

Outliers:

-12213298588

-4199788333

-2499804112

-2397701809

-2127109510

Task 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.**

AVERAGE (Mean):

To calculate the average of a range of data (e.g., column A), use this formula:

scssCopy code

=AVERAGE(A2:Axxxx)

Replace "xxxx" with the last row number where you have data.

MEDIAN:

To calculate the median of a range of data (e.g., column A), use this formula:

scssCopy code

=MEDIAN(A2:Axxxx)

MODE:

To calculate the mode (most frequently occurring value) of a range of data (e.g., column A), use this formula:

lessCopy code

=MODE.SNGL(A2:Axxxx)

Note: Use MODE.SNGL for a single mode, or MODE.MULT for multiple modes if applicable.

MAX (Maximum):

To find the maximum value in a range of data (e.g., column A), use this formula:

scssCopy code

=MAX(A2:Axxxx)

MIN (Minimum):

To find the minimum value in a range of data (e.g., column A), use this formula:

scssCopy code

=MIN(A2:Axxxx)

VAR (Variance):

To calculate the sample variance of a range of data (e.g., column A), use this formula:

lessCopy code

=VAR.S(A2:Axxxx)

Use VAR.P for population variance.

STDEV (Standard Deviation):

To calculate the sample standard deviation of a range of data (e.g., column A), use this formula:

lessCopy code

=STDEV.S(A2:Axxxx)

Use STDEV.P for population standard deviation.

AVERAGE	497.0869565
MEDIAN	383
MODE	Drama
MAX	1941
MIN	1
VAR	245389.5577
SD	495.3681032

Action	Descriptive Statistics
Mean	6.464144566
Median	6.6
Mode	6.7
range	7.7
Variance	1.11076587
Standard deviation	1.053928778

Adventure	Stats
Mean	6.462785208
Median	6.6
Mode	6.7
Range	7.7
Var	1.113445901
S.D	1.05519946

Animation	Descriptive Statistic
Mean	6.46375455
Median	6.6
Mode	6.7
range	7.7
Variance	1.110230729
Standard deviation	1.053674869

Biography	stats
mean	6.448317631
median	6.6
mode	6.7
range	7.7
var	1.116591382
sd	1.056688877

comedy	stats
mean	6.4630044
median	6.6
mode	6.7
range	7.7
variance	1.1104381
sd	1.0537733

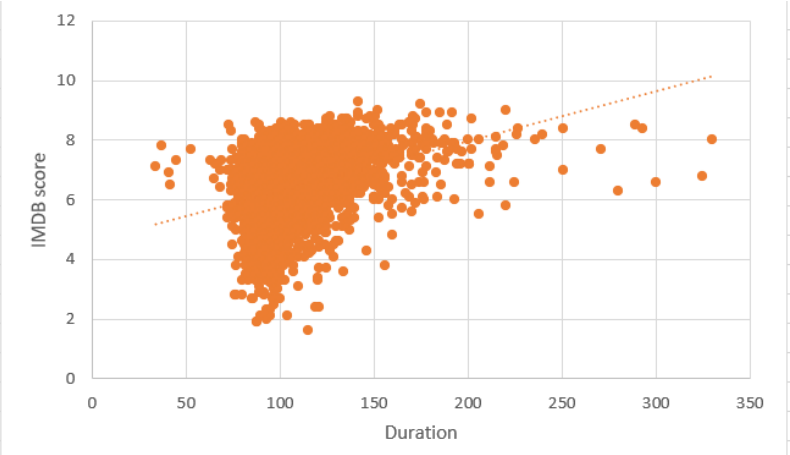
drama	stats
mean	6.4605855
median	6.6
mode	6.7
range	7.7
var	1.1122259
sd	1.0546212

Task 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.**

Column1	Duration	imdb score
mean	109.9241164	6.464163202
median	106	6.6
standard dev	22.75364979	1.053794573



Distribution of Movie Durations:

The mean (average) movie duration is approximately 109.9 minutes, suggesting that, on average, movies in the dataset have a duration of about 109.9 minutes.

The median movie duration is 106 minutes, indicating that half of the movies have a duration of 106 minutes or less, while the other half have a duration of more than 106 minutes.

The standard deviation of approximately 22.75 minutes reflects the variability or dispersion in movie durations. Movies in the dataset can have durations that deviate from the mean by about 22.75 minutes on average.

Impact on IMDB Scores:

To understand the impact of movie duration on IMDB scores, it's important to note that the mean IMDB score is approximately 6.46, and the median IMDB score is 6.6. This suggests that, on average, movies in the dataset have IMDB scores around 6.46, with a median score slightly higher at 6.6. These statistics indicate that the majority of movies in the dataset have IMDB scores clustered around the 6.5 to 6.6 range.

Relationship Between Movie Duration and IMDB Score:

To further analyze the relationship between movie duration and IMDB scores, I created a scatter plot.

A scatter plot can help visualize if there's any discernible pattern or trend between these two variables. In this case, it can show whether longer or shorter movie durations tend to correlate with higher or lower IMDB scores.

Adding a trendline to the scatter plot can help assess the direction and strength of the relationship. If the trendline slopes upward, it suggests a positive correlation (longer movies tend to have higher IMDB scores), while a downward slope indicates a negative correlation (longer movies tend to have lower IMDB scores).

Interpretation:

Based on the provided statistics, it appears that movie duration alone may not strongly determine IMDB scores. The mean and median IMDB scores are relatively consistent, regardless of movie duration.

However, the scatter plot and trendline analysis can provide a more detailed understanding of the relationship. If the trendline has a noticeable slope, it may indicate a weak to moderate relationship between the two variables.

Further Analysis:

To gain deeper insights, you can conduct additional statistical tests or subgroup analyses to explore whether specific genres or other factors influence the relationship between movie duration and IMDB scores.

Task 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.**

unique_language	Count	Mean	Median	SD
English	3602	6.42085	6.6	1.057599
Mandarin	14	7.021429	6.6	0
Aboriginal	2	6.95	6.6	0
Spanish	26	7.05	6.6	0
French	37	7.286486	6.6	0
Filipino	1	6.7	6.6	0
Maya	1	7.8	6.6	0
Kazakh	1	6	6.6	0
Telugu	1	8.4	6.6	0
Cantonese	8	7.2375	6.6	0
Japanese	12	7.625	6.6	0
Aramaic	1	7.1	6.6	0
Italian	7	7.185714	6.6	0
Dutch	3	7.566667	6.6	0
Dari	2	7.5	6.6	0
German	13	7.692308	6.6	0
Mongolian	1	7.3	6.6	0
Thai	3	6.633333	6.6	0
Bosnian	1	4.3	6.6	0
Korean	4	7.875	6.6	0
Hungarian	1	7.1	6.6	0
Hindi	10	6.76	6.6	0
Icelandic	1	6.9	6.6	0
Danish	3	7.9	6.6	0
Portuguese	5	7.76	6.6	0
Norwegian	4	7.15	6.6	0
Czech	1	7.4	6.6	0
Russian	1	6.5	6.6	0
Zulu	1	7.3	6.6	0
Hebrew	3	7.5	6.6	0
Dzongkha	1	7.5	6.6	0
Arabic	1	7.2	6.6	0
Vietnamese	1	7.4	6.6	0
Indonesian	2	7.9	6.6	0
Romanian	1	7.9	6.6	0
Persian	3	8.133333	6.6	0
Swedish	1	7.6	6.6	0
GRAND TOTAL	3781			

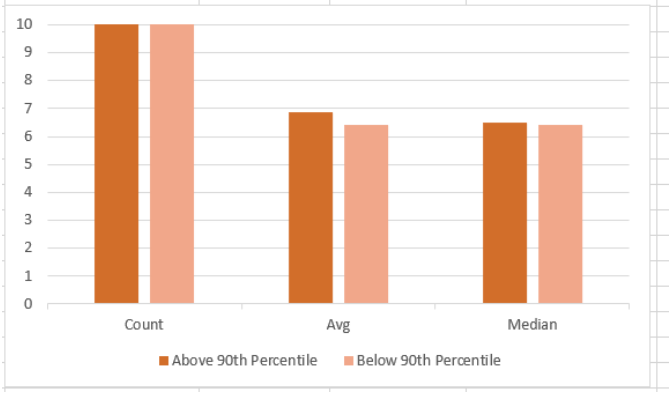
Most common languages used in movies- English

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.**

unique directors	SORTED avg imdb
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

Category	Count	Avg	Median
Above 90th Percentile	160	6.8725	6.5
Below 90th Percentile	1587	6.40063012	6.4



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.

Top 5 Profitable Movies:

director_name	actor_1_name	movie_title	title_year	imdb_score	Profit
James Cameron	CCH Pounder	Avatar	2009	7.9	523505847
Colin Trevorrow	Bryce Dallas Howard	Jurassic World	2015	7	502177271
James Cameron	Leonardo DiCaprio	Titanic	1997	7.7	458672302

George Lucas	Harrison Ford	Star Wars: Episode IV – A New Hope	1977	8.7	449935665
Steven Spielberg	Henry Thomas	E.T. the Extra-Terrestrial	1982	7.9	424449459

CORRELATION				
0.100850218	weak positive relationship			
Highest Profit Margin	**Movie - Country**	**title-year**	**imdb score**	**director**
523505847	Avatar - USA	2009	7.9	James Cameron

Result:

Throughout the IMBD Movie Analysis project, I honed a range of critical skills, encompassing logical reasoning, statistical acumen, and technical proficiency to extract meaningful insights from the dataset. Concepts like calculating averages, creating frequency tables, and identifying outliers proved invaluable in comprehending the data thoroughly and laying the foundation for analytical exploration.

My proficiency in applying statistical techniques, combined with the technical capabilities of Microsoft Excel, significantly expedited data analysis tasks, making complex calculations more manageable. The power of data visualization within Excel also became evident, simplifying data comprehension through charts and graphs. I acquired the knowledge to judiciously select the right visualization method based on the data and desired outcomes, enhancing my ability to convey insights effectively. Overall, this project equipped me with a robust skill set for data-driven decision-making and analysis.