IMDB MOVIE ANALYSIS

A project about descriptive analysis of movies

Project Description

a brief overview of the project

- The dataset is related to IMDB Movies the objective of the project is to investigating the main factors influencing the IMBD movies success .
- The project leads to understanding the dataset and to gain insights we need to find out the correlation in various factors such as IMBD ratings , budget , directors and many other factors which ultimately lead to aim of these project.
- This includes the various step to followed such as data cleaning, finding correlation , visualization of data which helps in better understanding of data.
- Overall the project describe how such insights empower the end the users in this case the stakeholders to take further decision in future as it will lead to strategic move informed by data.

Approach

- Data preparation :The first approach towards the project is understanding the data and cleaning it which includes various techniques such as removing duplicate values, handing the missing values by replacing it with mean(average) or ignoring it if the values is not affecting the analysis.
- Exploring correlations: Analyze links between movie ratings and attributes like genre, director, budget, language, movie duration.
- The "five whys" : Which helps to dig deeper and to find out the keen insights .
- Visualization of data : It further helps in better understanding of correlations between like histogram ,bar charts scatter plots will helps to identify the top numbers in data .
- Insights : By analyzing the correlations we will be able to find out the insights .
- Result : The end result is to give answers to problematic question by ultimately analyzing the data to stakeholders which helps them to make strategic decision making.

Tech-Stack Used

- Microsoft excel 2022 : it is used for the analyzing of dataset mainly for the descriptive statistics and further visualization of data .
- PowerPoint 2021 : It is used for making the presentation.

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 and then compare the statistics to understand the impact of genre on movie ratings.

- The dataset included various genre the first step is data cleaning removed duplicate values which gives us total of 914 unique values then using COUNTIF function to count the number of movies against each genre which gives MAX value that is 236 and MIN value that is 1 and then filter out top 10 genre using number filter which gives the most common genres. Then, for each genre calculated mean, median , mode, standard deviation which helps in the following comparison. And make bar graph which show the top 10 movies count and mean .
- Functions used are
- Averageif for mean   max(if)   stdev.p(if) for standard deviation
- Median(if)   min(if)   var.p(if) for variance
- Mode(if)   range(max-min)

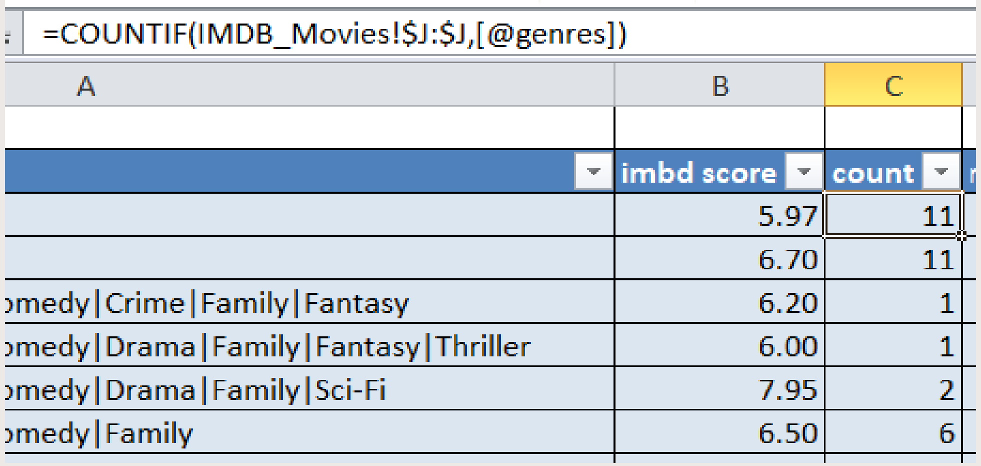
•By comparison we find out genres with highest and lowest average rating .

1.What are the top 3 most common genres in the dataset?

•According to the count top 3 genres are comedy , comedy/drama, drama with the maximum count whereas all the movies with the most highest IMBD score range 8-8.60 have only count of 1 this draw the conclusion despite of minimum number of count with these particular genres they are tend to have higher IMBD ratings.

- **B. How Consistent Are Audience Opinions within Genres?**
- What is the variance of IMDB scores within each genre?

A variance is something that measures how far each number in the set is from the mean (average), and thus from every other number in the set. In the dataset the maximum variances is 0and 1 which shows and draw conclusion that there are standard and normal distribution .



```
=COUNTIF(IMDB_Movies!$J:$J,[@genres])
```

A	B	C
	imbd score	count
	5.97	11
	6.70	11
omedy|Crime|Family|Fantasy	6.20	1
omedy|Drama|Family|Fantasy|Thriller	6.00	1
omedy|Drama|Family|Sci-Fi	7.95	2
omedy|Family	6.50	6

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 analyze this the first step is data preparation there are some missing values in movie duration it will affect the overall analysis of data so the missing value is replaced by average of movie duration that is 107.
- Then calculated descriptive statistics for both duration and imbd score using various function and inbuild function of data analysis which include mean, median, standard deviation , count, variance range.
- These statical measures give insights about data
- Made scatter plot to visualize relation between imbd score and duration and adding trendline show how the positive relation between them.

By calculating descriptive statistic we are able to find out some trend also the scatter plot defines how the data is overall distributed. The insights are

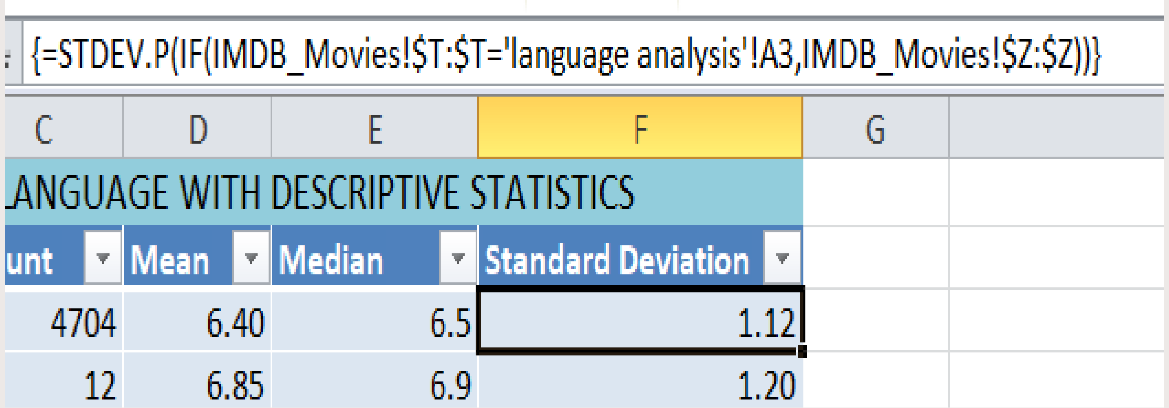
- The one trend we can clearly see is the average IMBD score is 6.44 and highest is 9.5 and all the movies that falls between these range are tend to have duration between 50 to 200 as the average of movie duration is around 107 and maximum it goes to 511. So it is clearly understandable from the data that audience is more likely to watch the movie with normal duration as per the trends the time span of audience is less so the movies with very long duration does not have a very high IMBD ratings .
- The second point is mean and median both of movie duration and IMBD ratings are close which indicates symmetrical distribution, and are both centrally located close to the high point of the distribution.

- **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.

- By examine the language clearly the English language is the most common language used in movie with the maximum no. of count and it also have IMBD score that is closer to average IMBD score.
- The movies with minimum no. Of count such as Telugu and Polish have the highest IMBD score through which we can draw a conclusion that in spite of low variability these movies have some specific preferences and cultural factors that impact ratings.
- We can also say that while the data doesn't show much variation, it's important to note that factors beyond language could still contribute to IMDb scores, such as plot, acting, and production quality.

- The language analysis also followed by the same steps just as genre analysis first the count of each language is find out by using count if which allow to count on specific criterion in this case it is particular language like English, Telugu etc. And then to calculate mean, median ,standard deviation of the IMBD score for each language used the average, median, standard deviation function all along with if to create nested function .
- Now for the comparison three different types of charts are made to show all the relation between various components and to gain insights.
- Nested function to calculate standard deviation.



```
{=STDEV.P(IF(IMDB_Movies!$T:$T='language analysis'!A3,IMDB_Movies!$Z:$Z))}
```

LANGUAGE WITH DESCRIPTIVE STATISTICS

unt	Mean	Median	Standard Deviation
4704	6.40	6.5	1.12
12	6.85	6.9	1.20

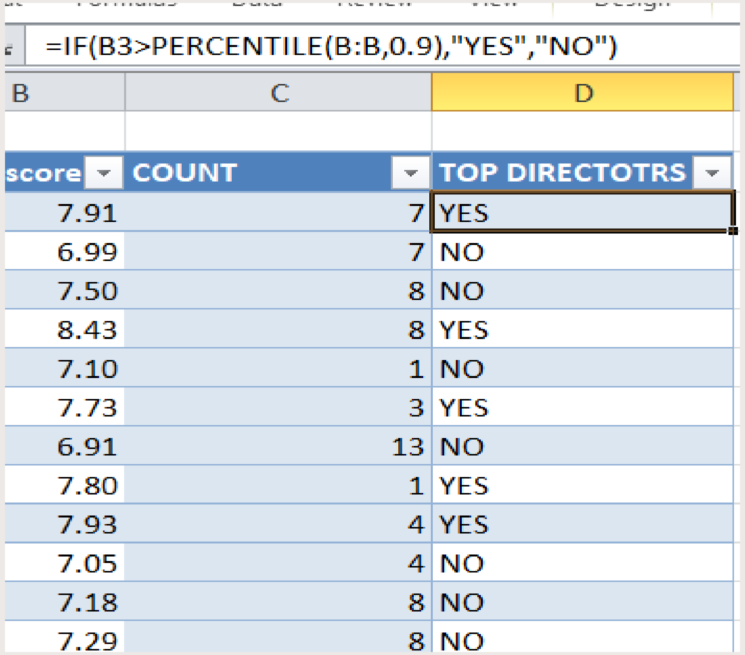
D. Director Analysis: Influence of directors on movie ratings.

- Task: Identify the top directors based on their average IMDB score and analyze their contribution to the success of movies using percentile calculations.

➢ To identify the top directors the percentile function is used first set the threshold percentage which is 90% in this dataset so percentile function filter out all the values from IMBD score which is above from these threshold.

➢ Again the count for each director is calculated using count if function which display the number of movies directed by each director.

➢ Histogram display the count and imbd score of each director.



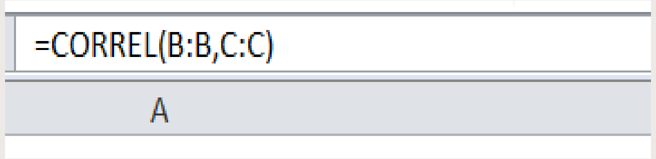
```
=IF(B3>PERCENTILE(B:B,0.9),"YES","NO")
```

score	COUNT	TOP DIRECTOTRS
7.91	7	YES
6.99	7	NO
7.50	8	NO
8.43	8	YES
7.10	1	NO
7.73	3	YES
6.91	13	NO
7.80	1	YES
7.93	4	YES
7.05	4	NO
7.18	8	NO
7.29	8	NO

- **Range of IMDb Scores:** The dataset spans a range of IMDb scores, from the lowest of 7.50 to the highest of 9.50. This showcases that the film industry has covers a wide range of different levels of quality and even though all directors on the list are considered "top," there is still variation in the excellence of their films.
- The directors' IMDb scores vary. This variation suggests that the quality of their works is not uniform, indicating that some directors have consistently produced films with higher audience ratings. The list includes directors with IMDb scores above 7.5, which is considered relatively good. Notably, several directors like Christopher Nolan, Lee Unkrich, Pete Docter, Hideaki Anno, and Quentin Tarantino have consistently received high scores for their works. This suggests that they have a strong track record of producing movies that resonate well with audiences and critics alike.

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 analyze the correlation first I prepare the data by replacing missing values with median as it is the most appropriate way according to nature of data as it is less sensitive to outliers.
  - The budget and gross are given so by subtracting gross from budget we get the profit for every movie label as profit margin.
  - After finding out profit margin we can filter out top movies using number filter and Avtar and Happiness A are highest earning movies.
  - The correlation between budget and gross is find out by using correlation function .
  - Out of total 5042 we can identify the nearly half of the movies had suffer losses this indicates the budget is not one of the critical factor that influence ratings .



```
=CORREL(B:B,C:C)
```

The excel sheet which is used in this project which contains all the IMBD_movies data and all the descriptive statistic is attached below.

https://docs.google.com/spreadsheets/d/1ZSTT3yOUQKCdmXQ6GNz2IAq1NUqm3fSg/edit?usp=drive_link&ouid=108529123799510054473&rtpof=true&sd=true

Insights

- To find out insights I took the why approach which helps in dig deeper
- Q: "Why do movies with higher budgets tend to have higher ratings?
- The movies with higher budgets can afford better production and picture quality with better VFX affect which increase the viewer experience and thus lead to positive reviews which can clearly anticipate in the quantitative form of higher ratings.
- Q. why do movies with longer duration does not have higher ratings?
- The viewers have a attention span for a short period thus the movies with average running time tends to have higher ratings as audience are likely to watch these short duration movies.
- Many other insights are mention in the previous slides.

Result

- Through this project it is evident that how we can identify the factors which affects the IMBD ratings of movies.
- It will help in better understanding of audience preferences through data analysis which further leads to take more of data driven decisions.
- Overall the project contributed towards the more data analytical thinking which focus on asking more of relevant question and not only to provide insights but also stating answers to problematic questions for end users.