IMDB MOVIE ANALYSIS

Excel hyperlink

Video hyperlink

PROJECT DESCRIPTION

In this project, I'm working as a Data Analyst and doing project of IMDb movie analysis. The dataset provided is related to IMDb Movies which includes **3757 rows 28 columns** after the **data cleaning**.

In this project we're doing analysis and investigate about **What factors influence the success of a movie on IMDb ?**

Here, success can be defined by high IMDb ratings. So we have to find the valuable insights to better understanding to the directors, movie producers, investors what makes a movie successful and to make informed decisions in their future projects.

The valuable insights will provide the better decision making to the directors, movie producers and investors how to create successful movies that increase their gross earning from the movie and make higher profit.

From this data, we can see that the higher budgets movies tends to higher ratings because they can afford the better production quality and vfx which can interact the people. And we know the it can also enhance our viewing experience. Because we know the viewers are more likely to see and

more likely to rate a movie which they enjoyed, because we know positive experiences lead to positive reviews.

As a Data Analyst, My analysis should aim to provide actionable insights that can help stakeholders make informed decisions.

APPROACH

Firstly, I have downloaded the dataset, and after this I handled the missing values and removed duplicates and make useful data to do feature engineering.

After doing this I solved all the given problems by using the excel function, pivot table, graphs and many more to extract the valuable insights from the data.

TECH-STACK USED

I have used the MS Excel 2019 version.

INSIGHTS

A. 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.

In this question we have to analysis the most common genres of movies in the dataset and after this we have to calculate the descriptive statistics for each genres of the IMDb scores.

Firstly, I copied **genres** column to the new sheet to better analysis and remove **duplicate values** and extract the **unique values** from the genres by using **filters** and **count** the number of movies with respect to genre to better understanding.

After this, we find the most common genres of the movie by using the **filter** function. Then **find the mean, median, mode (descriptive statistics)** of the data.

The **Mean** will help us to to know which value is most commonly used in data and also used to know presence of extreme observations has the least impact on it. Also it will help us to measure the **central tendency** of data.

From the given data **mean is 5.03** which shows that any movie of genres contains approximately **5 IMDb ratings**. If we analyze from the **business perspective** we can conclude that any upcoming movies would have rating approximately **5**. This can be used in business perspective and also to director, movie producers and investors for their future projects.

Median will help us in open ended distribution since the position rather than the value of item that matters in median. **Median** value in this data also used to measure

central tendency for **skewed distributions** also used in distribution of **outliers**.

From the data the **Median is 1**. It means that majority of genres as atleast one movie. We can't say that median is good approach to predict on analysis our data in terms of genres.

But we conclude that majority of genres contain atleast **1 movie.**

Mode will help us easy to identify in a data set in a discrete frequency distribution. It's also useful for qualitative data.

From the analysis we conclude that mode has value of 1 which means that majority of genres having 1 movie.

It shows there's many genres are available for viewers and they can enjoy any movie according to their choice.

From the insights we can say that, **Comedy|Drama|Romance is most common genre**.

Here's our output :

Most Common Genres	
genres	**no of movies**
Action|Adventure|Sci-Fi	48
Drama|Romance	115
Action|Crime|Thriller	56
Action|Crime|Drama|Thriller	50
Comedy|Drama|Romance	147
Comedy|Romance	131
Comedy	138
Crime|Drama|Thriller	82
Comedy|Drama	138
Drama	141

Here's our descriptive statistics :

Descriptive Statistics	
Mean	5.032258065
Standard Error	0.506683407
Median	1
Mode	1
Standard Deviation	13.82048073
Sample Variance	191.0056875
Kurtosis	65.3558989
Skewness	7.505270486
Range	146
Minimum	1
Maximum	147
Sum	3744
Count	744
Largest(1)	147
Smallest(1)	1

B. Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.

In this question we have to analyze the distribution of movie durations and identify the relationship between **movie duration and IMDb score.**

This can help to know the **producers, directors and investors** about the viewers interaction.

Firstly, I copied **duration and imdb score** column into the new sheet for better analysis and after this I make **scatter plot** to better visualizations and also plotted a **trend line** to assess the direction and strength of the relationship by using excel function.

This can help to know about the **relationship between Movie durations and IMDb scores.**

From the data, our **Mean value is 110.24** with respect to their movie duration. We can say that the average movie duration is approximately **110 minutes of durations**.

Also, we can say that the movies having approximately 110 minutes of duration that movie having higher IMDb ratings which shows the popularity and interest of viewers.

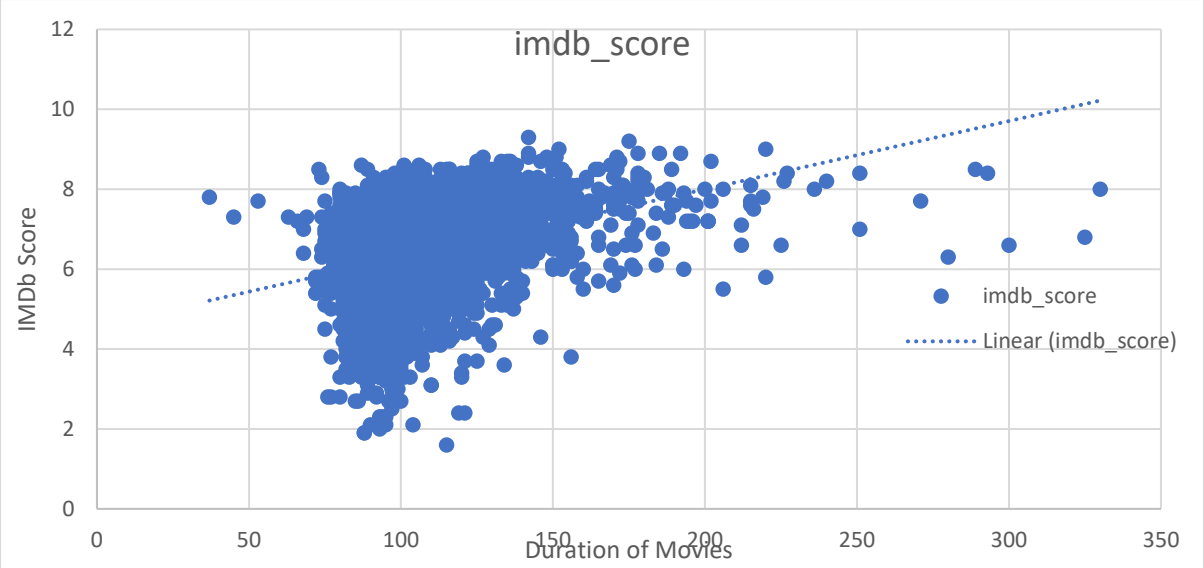
From the data, the **median is 106 with respect to movies** duration which means that many movies having 106 minutes of duration and this kind of movies people are enjoying also movie having more than **6** IMDb ratings.

From the analysis, **Mode comes 101** which means that majority of movies having approximately 101 minutes of duration and people loving it and contains higher amount of IMDb ratings.

The scatter plot shows the **higher IMDb ratings which have more than 100 minutes of duration**.

After this, I make a Descriptive Statistics to know about the data.

Here’s our output :



Descriptive Statistics	
Mean	110.2399
Standard Error	0.369181
Median	106
Mode	101
Standard Deviation	22.62272
Sample Variance	511.7876
Kurtosis	12.68422
Skewness	2.406084
Range	293
Minimum	37
Maximum	330
Sum	413951
Count	3755
Largest(1)	330
Smallest(1)	37

C. Determine the most common languages used in movies and analyze their impact on the IMDB score using descriptive statistics.

In this question, we have to analyze the most common languages used in movies and also analyze their impact on the IMDb score by using descriptive statistics.

Firstly, I copied **language and imdb score** columns into the new sheet to better analysis.

After this, I filter the unique values from the data and make **frequency of number of movies** with respect to the **languages**.

Then, I made a **descriptive statistics table** and **chart** to better visualize the output and to understand the impact of languages on movie ratings.

From analysis, **Mean is approximately 6.46** which means that every language having **6 number of movies**. But if we observe the data English language has majority of movies and other languages are not comparable according to this data. Because there is vast difference between any one of the languages with English.

From the data, **Median is found to be 6.6** which shows that **every languages has approximately 6 or 7 movies**. But already we know there's vast difference English and other languages so we can take approach this data with the business perspective but we can say that English language having highest number of movies.

From the analysis, **mode is 6.7** which shows that majority of languages having 6 or 7 movies which is absolutely wrong. So we can't take this approach to business analysis. But we can say that English movies having higher numbers and viewers

will enjoy English movies more and from business perspective Directors have to more focus on English movie to increase their gross earnings and make more profits.

From the insights we can say that**, English is most common languages used in movies which is 3598.**

Here's our output :

language	No of movies
English	3598
Mandarin	15
Aboriginal	2
Spanish	23
French	34
Filipino	1
Maya	1
Kazakh	1
Cantonese	7
Japanese	10
Aramaic	1
Italian	7
Dutch	3
Dari	2
German	10
Mongolian	1
Thai	3
Bosnian	1
Korean	5
Hungarian	1
Hindi	5
Danish	3
Portuguese	5
Norwegian	4
Czech	1
Russian	1
None	1
Zulu	1
Hebrew	1
Arabic	1
Vietnamese	1
Indonesian	2
Romanian	1
Persian	3
total	**3756**

Mean	6.4649
Standard Error	0.017235
Median	6.6
Mode	6.7
Standard Deviation	1.056128
Sample Variance	1.115406
Kurtosis	1.148004
Skewness	-0.72335
Range	7.7
Minimum	1.6
Maximum	9.3
Sum	24275.7
Count	3755
Largest(1)	9.3
Smallest(1)	1.6

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

In this question, we have to identify the top directors based on their average IMDb score and also analyze their contribution to the success of movies using the percentile calculations.

Firstly, I copied **director name and imdb score** column into the new sheet to better analysis.

After this, I filtered out the unique values of the **director** from the column and make frequency with respect to the director name to get know how many movies are made by directors.

Then, find the **average imdb score** with respect to the **directors** to know to success rate of directors to their movies.

After this, I make **a percentile table** and chart **to visualize** and to know overall distribution of scores.

We can see that **Akira Kurosawa has 8.7 score which is one of the highest IMDb score.**

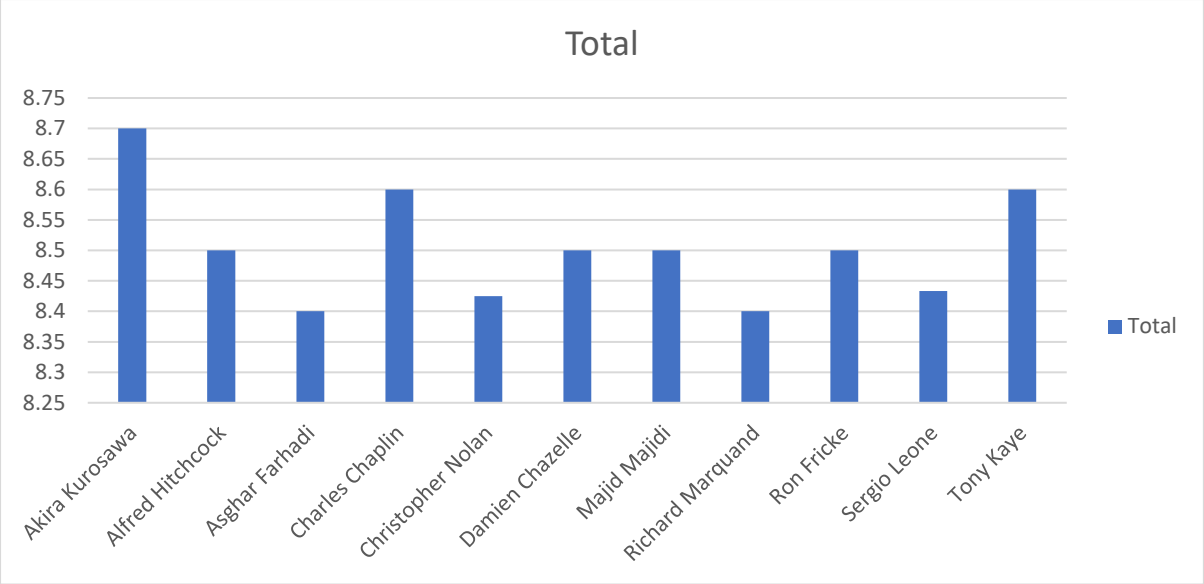
Here's our output :

Row Labels	Average of imdb_score
Akira Kurosawa	8.7
Alfred Hitchcock	8.5
Asghar Farhadi	8.4
Charles Chaplin	8.6
Christopher Nolan	8.425
Damien Chazelle	8.5
Majid Majidi	8.5
Richard Marquand	8.4
Ron Fricke	8.5
Sergio Leone	8.433333333
Tony Kaye	8.6
Grand Total	**8.47**

Percentile Table :

25th percentile	5.9
50th percentile	6.6
75th percentile	7.2

Graph :



E. Analyze the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.

In this question, we have to analyze the correlation between the movie budgets and gross earning and also identify the movies with the highest profit margin.

Firstly, I copied **movie name, gross, budget columns** to the new sheet to better analysis.

After this I make a **profit** column to know the how much movie earned the profit. So for this I find the **difference between gross and budget** by using the excel function.

After this I extract the **correalation between budget and gross** to know their relation and make graph to better visualization.

AvatarA is most profit margin movie which has 523505847 profit.

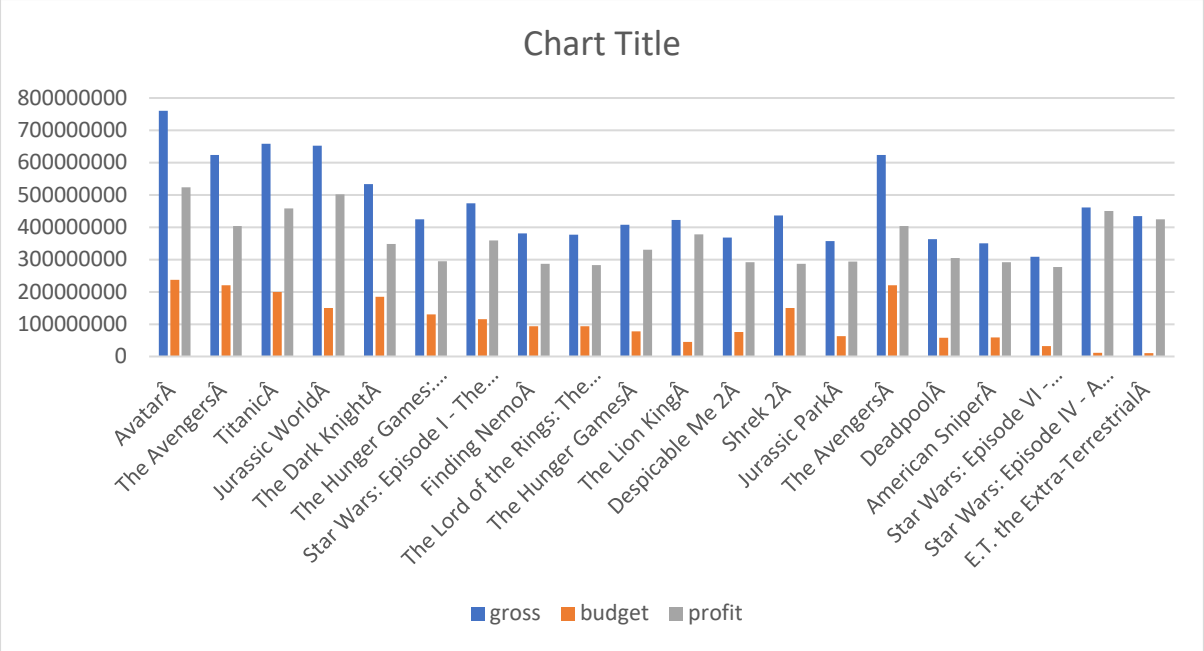
Here's our output :

Correlation between Budget and Gross Earning			
		budget	***gross***
	budget	1	
	gross	0.099496	1

Movies with higher profit margin :

movie_title	gross	budget	**profit**
AvatarÂ	760505847	237000000	523505847
The AvengersÂ	623279547	220000000	403279547
TitanicÂ	658672302	200000000	458672302
Jurassic WorldÂ	652177271	150000000	502177271
The Dark KnightÂ	533316061	185000000	348316061
The Hunger Games: Catching FireÂ	424645577	130000000	294645577
Star Wars: Episode I - The Phantom MenaceÂ	474544677	115000000	359544677
Finding NemoÂ	380838870	94000000	286838870
The Lord of the Rings: The Return of the KingÂ	377019252	94000000	283019252
The Hunger GamesÂ	407999255	78000000	329999255
The Lion KingÂ	422783777	45000000	377783777
Despicable Me 2Â	368049635	76000000	292049635
Shrek 2Â	436471036	150000000	286471036
Jurassic ParkÂ	356784000	63000000	293784000
The AvengersÂ	623279547	220000000	403279547
DeadpoolÂ	363024263	58000000	305024263
American SniperÂ	350123553	58800000	291323553
Star Wars: Episode VI - Return of the JediÂ	309125409	32500000	276625409
Star Wars: Episode IV - A New HopeÂ	460935665	11000000	449935665
E.T. the Extra-TerrestrialÂ	434949459	10500000	424449459

Graph :



CONCLUSION

- From this project, I gained my Advanced Excel skills and also learnt about how Data Analyst extract the valuable insights from the data.
- Also, I gained skills how to use pivot table and how to visualize by using graph, chart, scatter plot to the given output to better understanding the data.
- I learnt how to deal with null or missing values and how to clean the data it can helps to enhance the better feature engineering.
- Also I learnt the trend line analysis it can help to give valuable insights and to better understanding about the data.
- I also learnt about the which factors influence the the success of a movie on IMDb and the directors, movie

producers, investors what makes a movie successful and to make informed decisions in their future projects.

- We can see that the higher budgets movies tends to higher ratings because they can afford the better production quality and vfx which can interact the people. And we know the it can also enhance our viewing experience. Because we know the viewers are more likely to see and more likely to rate a movie which they enjoyed, because we know positive experiences lead to positive reviews.

Linkedin Profile : https://www.linkedin.com/in/pawan-kashyap-832515230

Instagram : https://www.instagram.com/pawankkkashyap