**IMDB Movie Analysis Project**

**Project Description:**

IMDb (Internet Movie Database) is an [online database](https://en.wikipedia.org/wiki/Online_database) of information related to films, television series, podcasts, home videos, video games, and streaming content online including cast, production crew and personal biographies, plot summaries, trivia, ratings, and fan and critical reviews. IMDb began as a fan-operated movie database in 1990, and moved to the Web in 1993. Since 1998, it has been owned and operated by IMDb.com, Inc., a subsidiary of [Amazon](https://en.wikipedia.org/wiki/Amazon_(company)).

**Approach:**

I went through the Excel data provided by the Trainity IMDB Movie Analysis project and understood that there were columns related to the movie in the dataset. Further, I understood the columns and their respective constraints to do the analysis. I was given a set of questions to solve as part of the analysis. By using the Microsoft Excel, I did solve the queries and provided the result as expected.

**Tech-Stack Used:**

Microsoft Excel 2021 – To answer the queries with the help of Excel formulas in the tool.

**Insights:**

Did the data cleaning like:

* Removing null values.
* Removed the columns which we don’t use for the analysis.
* Removing the Duplicate rows.

Have used the in-built formulas in excel for the descriptive analysis such as:  
Mean – average()

Median – median()

Mode – mode()  
Max – max()  
Min – min()

Variance – VAR.P()

Standard Deviation - STDEV.P()

With the help of the Excel formulas, I found out many insights which include –

**Task A - Genre Analysis:**

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.

Used the columns Genres and found out the count of each Genres using the countif formula in Excel.

And sorted the Top 10 count to find out the Top 10 in pivot table and created the Column chart for it.

The Descriptive Analysis for the above is:

|  |  |
| --- | --- |
| Mean | 226 |
| Median | 39 |
| Mode | 3 |
| Max | 1028 |
| Min | 2 |
| Var | 109876.8 |
| Std Dev. | 331.5 |

We could see that the most popular genre is **Comedy** and followed by the other genres.

**Task B – Movie Duration Analysis:**

Analyse the distribution of movie durations and identify the relationship between movie duration and IMDB score. Created the Scatter plot as instructed in the task.

The Descriptive Analysis is as below:

|  |  |
| --- | --- |
| Mean | 109.9 |
| Median | 106 |
| Mode | 101 |
| Max | 330 |
| Min | 34 |
| Var | 517.6844 |
| Std Dev. | 22.75268 |

We could see that the trendline is increasing with the duration increase.

But mostly the IMDB score is more when the duration is between 80 mins to 150mins.

**Task C – Language Analysis:**

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

From the above Column chart, it is evident that the language “**English**” is the most common language with **23556** times present in the dataset. It was found by counting the number of times its present using the countif formula in excel.

The descriptive analysis is as below:

|  |  |
| --- | --- |
| Mean | 6.46 |
| Median | 0 |
| Mode | 0 |
| Max | 23556 |
| Min | 0 |
| Var | 144273.8 |
| Std Dev. | 379.8339 |

**Task D – Director Analysis:**

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

Have plotted the Bar graphs for the top directors with more IMDB scores as well as with the percent rank based on the IMDB scores.

**Task E – Budget Analysis:**

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

Found the Profit of each movie by the difference of the gross with budget used for the movie.

These are the top 5 Profitable movies in the database:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Director Name** | **Gross** | **Budget** | **Year** | **IMDB Score** | **Movie Title** | **Profit** |
| James Cameron | 760505847 | 237000000 | 2009 | 7.9 | Avatar | 523505847 |
| Colin Trevorrow | 652177271 | 150000000 | 2015 | 7 | Jurassic World | 502177271 |
| James Cameron | 658672302 | 200000000 | 1997 | 7.7 | Titanic | 458672302 |
| George Lucas | 460935665 | 11000000 | 1977 | 8.7 | Star Wars: Episode IV - A New Hope | 449935665 |
| Steven Spielberg | 434949459 | 10500000 | 1982 | 7.9 | E.T. the Extra-Terrestrial | 424449459 |

The most profitable movie is **Avatar** with a profit of **523,505,847 Dollars (5.2 billion Approx.)**.

**Result:**

Through this project I was able to understand the formulas being used in the Excel which can be used to find the Statistical measures such as Mean, Median, Mode, Max, Min, Variance and Standard Deviation. I got used to the Excel formulas and how to convert the Raw Data into meaningful insights. And the steps which I used are – cleansing the data and using the formulas to find the desired outcome and also learnt how to convert the data into a visualized chart so that the insights can be drawn within seconds by seeing the graphs instead of searching the whole data.

I have achieved the end result and I think I have contributed my full support into the Analysis. I hope this project helps the Analysis and it achieves what it was tend to achieve.

**Hyperlink for the Excel sheet:**

[IMDB-Movie-Analysis-Excel-file](https://github.com/Arunmaran21/IMDB-Movie-Analysis-Excel-file)