Date:- November 01, 2023. [Linked A blue square with white letters and a white text

Description automatically generated](https://www.linkedin.com/in/abdulrahman8/)

A multifaceted analysis of top movies from 1980 to 2021 according to IMDb.

Introduction:

Since I have become a movie enthusiast, I decided to do some valuable analysis to unveil informative insights about movies. This project endeavors to undertake a comprehensive analysis of the film industry, with a particular focus to understand the evolution of the movies performance over the decades. By employing data analytics tools and techniques, we aim to uncover underlying trends, patterns and facts of that have shaped the film industry.

This analysis serves an exploration of movie ratings, success at the box office, correlation between budget and gross income of movies and many aspects of the film industry. Through this study I like to provide a data-driven perspective on the ever-evolving world of cinema.

About the dataset:

This project relies on a broad-based dataset compiled from IMDb. Initially, the dataset of movies from 1980 to 2019 was obtained from [Kaggle](https://www.kaggle.com/datasets/danielgrijalvas/movies).com, the owner of the dataset scraped the information about movies from IMDb. Additionally, the movie information for the years 2020-2021 was collected by me using Microsoft power automate from IMDb. Then these datasets were merged using Excel’s power query and cleaned, transformed in excel to suit the requirements.

This dataset encompasses a wide range of information, with 15 fields and it has 8000 of movies totally and each movie in the dataset is characterized by the following attributes,

* Name: The title of the movie.
* Year: Year of release.
* Rating: Content classification of the movie. (advisories for movies regarding a movie’s content)
* Genre: Main genre of the movie.
* Released date: The release date of the movie. (e.g.- ‘January 8, 2021’)
* Score: IMDb user rating.
* Votes: Number of total votes for the movie.
* Director: The director of the movie.
* Writer: Writer of the movie.
* Star: The main actor/star of the movie.
* Country: The country of origin.
* Budget (USD): The budget of a movie (some movies don’t have the information, so it appears as NULL).
* Gross Income (USD): The box office income of a movie, some movies don’t have the values, it appears as NULL.
* Company: The production company.
* Runtime: Duration of the movie.

During the data cleaning process in Excel, I first validated and corrected errors in string and numeric values. Additionally, I removed non-printing characters from certain cells, which contained values. I also had to manually populate missing values in, released date, writer, company, country, gross income and budget fields, as I could only scrape the data for certain attributes for the years 2021 and 2020.

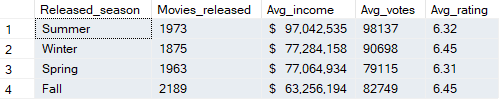
Movies data analysis questions to extract valuable insights.

Top movies released from 1980 – 2021. [8000 Movies.]

Microsoft SQL Server is used to generate insights from the movie dataset through specific questions. Allowing us to understand trends, patterns, and factors of movies.

1). Released season analysis:

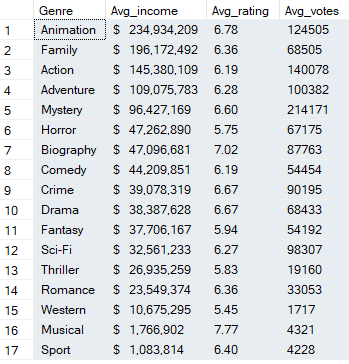
Are there specific released months or seasons when movies tend to perform better at the box office and receive higher ratings/scores?

 Result:-

This analysis is based on the assumption that movies performance is influenced by its released season, which might not capture all factors affecting a movie's success, but it helps to understand whether specific released seasons are associated with better box office performance and higher ratings/scores.

2). Genre performance:

Which movie genres have the highest average gross income and ratings/scores?

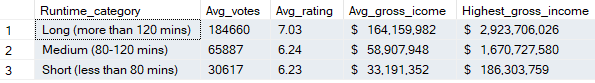


Result :-

This analysis provides insights into the performance of main movie genres focusing on their average gross income and it helps to identify the most profitable genres and their associated scores, votes over the years.

3). Runtime impact:

How does the length (runtime) of a movie affect its audience votes and box office performance ?

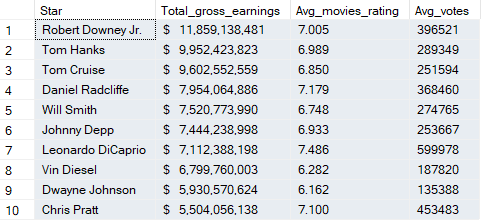


Result:-

This analysis helps understand how runtime relates to a movie’s performance in terms of audience ratings, votes, gross income.

4). Star power at the box office:

Do movies with a specific actor/star tend to earn more at the box office and receive higher ratings/scores?

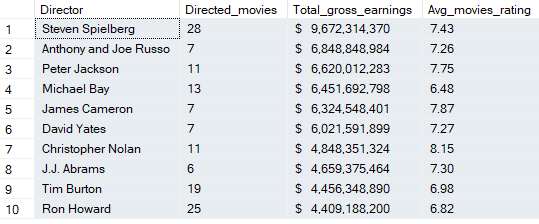


Result:-

This provides information about top ten actors with their significant impact on movies’ box office earnings and the average ratings of the movies in which they have starred as the main actor.

5). Director influence:

Do certain directors have a high impact on the box office earnings and average ratings?



Result:-

This provides information about the top ten directors with their impact at the box office earnings and average ratings of movies that they directed.

6). Company success:

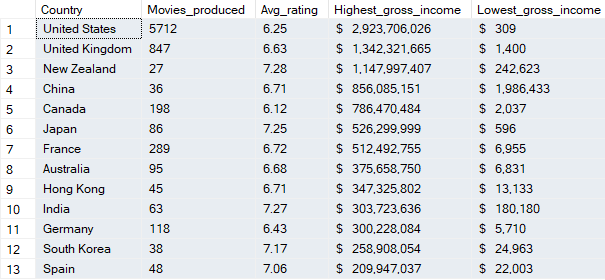
Are certain production companies associated with more success movies in terms of rating and gross income?

 Result:-

This result provides insights about top production companies with the number of best movies produced, highest gross income , total gross earnings, average movies rating. This helps to identify and understand the success and performance of companies in the movie industry.

7). Country based insight:

Are there differences in the box office performance of movies based on the country of production?



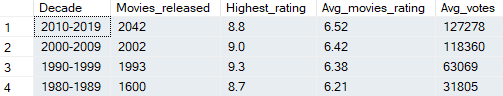
Result:-

This information helps identify countries with a strong presence in the movie industry, highlighting their highest grossing productions and overall movie quality with the best movies produced. These countries were selected based on having more than 1 billion of total gross income over the years.

8). Rating trends over decade:

How have the movie popularity evolved over the decade?

Result:-



This analysis offers the performance of best movies across different decades. It helps to identify which decade has the highest rating and on average, the most highly rated movies and average votes based on the data.

The SQL scripts for these analyses are on [Github.com](https://github.com/AbdulRa8man/Data-Analysis-Portfolio-Projects/blob/e6e229533b6a2cecafb197c51b6534d98782fa76/Movies%20Data%20Project/Movies%20Data%20Analysis%20-%20SQL%20Scripts.sql)

* Correlation analysis

Before the analyses in SQL, I performed a correlation analysis on numerical features in Excel and visualized the correlation matrix and it revealed that some of the correlations exhibited slightly well positive relationships.

The Cleaned dataset and correlation matrix file:- [Excel.xlsm](https://github.com/AbdulRa8man/Data-Analysis-Portfolio-Projects/blob/e6e229533b6a2cecafb197c51b6534d98782fa76/Movies%20Data%20Project/Movies%20Dataset%20%26%20Correlation%20Matrix.xlsm)

* Data visualization:- The visual insights for these analyses are on [Tableau.com](https://public.tableau.com/app/profile/abdul.rahman08/viz/Moviesdatavizunfinished/Dashboard1#1)
* Acknowledgments Data sources are from IMDb.com and [Kaggle](https://www.kaggle.com/datasets/danielgrijalvas/movies).com
* Contacts:-

Email: [abrahman0805@gmail.com](mailto:abrahman0805@gmail.com)

[LinkedIn](https://www.linkedin.com/in/abdulrahman8/)

* Project Portfolio website:- [datascienceportfol.io/AbdulRa8man](https://www.datascienceportfol.io/AbdulRa8man)

Created by: Abdul Rahman.