Box Office Film Analysis

Overview

This project takes in data about movies from different sources. Box Office movie data, The Movie Database data, The Numbers movie data, Rotten Tomatoes movie reviews data and Internet Movie Database data.

This data is analysed to give business recommendations on starting up a film industry. Microsoft is a large Tech company and is looking to expand into film production. This project will provide actionable information to help determine the type of films to create

Business Understanding

The film industry is a large industry worth over 40 billion USD. Our main objective in this project will be to provide information to help Microsoft predict the best outcomes in this industry. Our specific objectives are to find out:

- Most popular genres of film at the Box Office
- Highest grossing films.
- Films with the highest ratings.
- Relationship between budget and revenue

Data Understanding

The data provided is from five different sources and as such it comes in different formats. These formats are: comma separated values, tab separated values, and sql .db files. Some files also had unique encoding. As such each data source had to be treated uniquely.

The sources provided different information arranged in tables and columns. Some of this information includes: movie titles, popularity, rating, release date, reviews, genre, production budget, local gross revenue, worldwide gross revenue, synopsis and even directors.

For our objectives not all the information was required and this led us to our next step.

Data Analysis

Our Data Analysis was divided into sections.

Data Extraction

We extracted the data from the different sources into formats we could use in our analysis. We used available methods to describe the data and find out information about the data.

Data Cleaning

Here we selected the data we required from each source to help us achieve our objective. We removed unnecessary columns and information. After careful consideration we removed null values and duplicated entries.

We also did data transformation in areas where it was necessary e.g converting revenue fields stored as string to usable float or integer data.

Data Merging and Aggregation

Some data sources had fields that coud relate to fields in other data sources and we joined this sources to help us get even further insights from the sources given

We also merged relevant data and used aggregation methods to extract aggregate data from different groups of data.

Stats and Measures

We then derived the necessary statistical information we required for our business recommendations.

Data Visualization

A combination of bar graphs, line graphs and scatter plots were used to visualize the analysed data better.

Recommendations

From this analysis we can conclude that the movie industry is a high profit industry. We found out that the mean average expenditure was 31,587,757 USD and average gross income was 91,487,460 USD leaving a tidy figure of 59,899,703 USD as profit. With some fields grossing 2,776,345,279 USD in revenue.

From the relationship between revenue and profit, it was seen that they are directly proportional, showing that a larger budget meant a higher revenue.

We also determined the most popular genres in film and from this we can draw recommendations on what type of films to produce. These would be:

- Drama
- Documentary
- Comedy
- Horror
- Comedy, Drama

Although, while doing an analysis of the highest rated films, we realised that they were a mix of genres. the highest rated mix of genres were:

- Comedy, Documentary, Fantasy
- Documentary, Family, Musical
- History, Sport
- Music, Mystery
- Drama, Fantasy, war

Digging deeper into films, we analysed the highest grossing and most popular films at the Box Office, these were:

Most popular:

- Avengers Infinity War
- John Wick
- Spider-Man Into the Spiderverse
- The Hobbit - Battle of Five Armies
- The Avengers

Highest Grossing:

- Avatar
- Titanic
- Star Wars VII - The Force Awakens
- Avengers Infinity War
- Jurrasic World

English was by far the most used language in film. Followed by French, Spanish, Russian and Japanese.

Overall. The film industry is a high expenditure, high gains industry and the above information allows us to understand the industry further.

Next Steps

More data is required to give a deeper understanding of the industry. This could be acquired through Web Scrapping and APIs. We could draw further insights for example, the relationship between movie popularity and revenue.

We could also further analyse the relationship between budget and revenue and understand why some films made more revenue despite a smaller budget and some made less revenue with a larger budget.

Further analysis in film release dates could allow us to know if they is any relationship between film success and release dates.