Project Overview

Background of the Business:

Microsoft, a global technology giant, has decided to venture into the entertainment industry by establishing a new movie studio. This decision is driven by the growing trend of large companies creating original video content. However, Microsoft lacks experience in the film industry, making it essential to gather insights and data to make informed decisions.

Domain of the Business:

The domain of this business project is the film and entertainment industry. Microsoft aims to create and produce movies that resonate with audiences and achieve success at the box office. This industry is highly competitive and influenced by various factors, including film genres, release strategies, marketing, and audience preferences.

Business Case:

Microsoft's business case involves creating a new movie studio, but the company lacks expertise in filmmaking. To address this challenge, the project's goal is to explore and analyze existing data related to the film industry. By conducting exploratory data analysis (EDA), the project seeks to identify trends, patterns, and factors associated with successful movies in terms of box office performance.

BUSINESS UNDERSTANDING

The primary business goal is to determine what types of films perform well at the box office and translate these findings into actionable recommendations for Microsoft's new movie studio.

Key agendas:

Data Sources: Microsoft has access to various movie datasets from different sources, including Box Office Mojo, IMDB, Rotten Tomatoes, TheMovieDB, and The Numbers. The data is diverse and comes in different formats, including compressed CSV, TSV, and SQLite databases. Microsoft has provided flexibility in choosing which data to use for analysis.

Data Relationships: The datasets contain information about movies, including their titles, genres, release years, ratings, box office earnings, and more. The IMDB data is central to the analysis and is related to other provided data files. The challenge is to understand how these datasets are interconnected and how to leverage them effectively.

Objectives:

The key business objectives are as follows:

1. Understand Box Office Success: Gain insights into what types of films are currently performing well at the box office. This includes identifying the most successful genres, release years, and other relevant factors.

2. Translate Insights into Action: Convert the findings from data analysis into actionable recommendations. These recommendations will guide Microsoft in deciding the types of films to produce, thereby increasing their chances of success in the movie industry.

3. Effective Communication: Communicate the analysis results and recommendations effectively to the head of Microsoft's new movie studio. The project aims to provide clear and compelling insights through storytelling and visualizations.

This project focuses on helping Microsoft make informed decisions about the types of films to produce, leveraging data-driven insights from the film industry. The ultimate goal is to ensure that Microsoft's new movie studio enters the entertainment industry with a competitive edge and a higher likelihood of producing successful movies at the box office.

DATA UNDERSTANDING

The data was collected from various locations, the different files had different formats in it. My data had 3387 rows and 13 columns in for my analysis and it had various variables which included:

. 'movie_id' which was the primary key and it was essential for merging different dataframes

. runtime in minutes of each movie that is in our data, this help to know the average time for a movie

. domestic and foreign gross which helped us to know how different studios have achieved high number of domestic gross. It also helps in identifying different cost of movie production

. Genres which helped to know which film genres Microsoft studio should focus on in order to achieve its target financially

Before doing my analysis I started by cleaning my data i.e removing null values, checking whether there are duplicate values and any inconsistency in my data. I also found out that some columns with integer datatype had unnecessary data so I removed some and corrected others.

DATA ANALYSIS

The exploratory data analysis (EDA) of the movie dataset ‘movies_review_df’ has provided valuable insights into various aspects of the dataset. Here are the key findings and conclusions from the analysis:

Univariate Analysis:

Categorical Data: The dataset contains categorical variables such as 'movie_id,' 'primary_title,' 'original_title,' 'genres,' 'title,' and 'studio.' These variables have been described, including counts of unique values and the most frequent entries.

Numerical Data: Descriptive statistics were computed for numerical variables, including 'runtime_minutes,' 'average rating,' 'num votes,' 'domestic_gross,' 'start_year,' and 'year.' The measures of central tendency (mean, median, and mode) and measures of dispersion (standard deviation) were analysed for these variables.

Runtime Distribution: The distribution of movie runtimes shows a peak around the average runtime of approximately 90 minutes, indicating a roughly symmetrical distribution.

Movie Release Year Trends: The analysis of movie release years revealed interesting trends. There was a steady decrease in movie releases from 2010 to 2016, followed by an increase from 2017 onwards. 2019 had the highest number of movie releases. The reasons for these trends could be explored further.

Genre Popularity: The most popular movie genre based on the dataset is 'Drama,' followed by 'Documentary.' These findings suggest that viewers and filmmakers have a preference for drama and documentaries.

Bivariate Analysis:

Movie Performance by Studio: The analysis of total domestic gross earnings by studio identified BV studios as the top-performing studio in terms of earnings. This information is valuable for understanding the financial success of studios and making industry comparisons. We can conclude that BV studio is the most preferred studio for film makers and we must try and emulate what they are doing in order to make huge profits

Multivariate Analysis:

Correlation Analysis: A correlation heatmap was generated to explore relationships between multiple numerical variables in the dataset. The analysis found that most variables have low or near-zero correlations, indicating a lack of strong linear relationships. The closest positive correlation observed was between 'start_year' and the 'year' of movie production, but it was relatively low.

RECOMMENDATIONS

Microsft Studio may consider focusing on drama and documentary genres to cater to viewer preferences based on the formulated analysis

Further analysis should be conducted to understand the factors influencing movie release trends but we can conclude that in 2019 more movies we produced because of covid 19. Most guys were at home because of the lockdown.

Microsft Studio can use insights from top-performing studios like BV to inform their strategies.

Explore non-linear relationships and external factors that may affect movie performance.

This comprehensive analysis provides a foundation for deeper investigations and data-driven decision-making in the movie industry.

The next steps:

1. Demographic analysis

Conduct a demographic analysis of the audience to better undersatand who is watching this movies. Analyse factors such as age groups ,gender and geographical locations of viewers. This can help studios tailor their content to specific demographics.

2. genre analysis

Explore how the popularity of movie genres has evolved over the years. Are there specific genres that have gained or lost popularity? Investigate whether certain genres have cyclic trends.

3.Sentiment Analysis

Perform sentiment analysis on viewer reviews or comments if available. This can provide insights into viewer preferences, opinions, and sentiments towards different genres, studios or specific movies

4. Impact of external events
investigate the impact of major natural calamities like a pandemic or an economic recession on movie release trends and performance. Explore how viewer behavior and preferences changed during such events.