MICROSOFT MOVIE STUDIO PROJECT

1. BUSINESS UNDERSTANDING

Link to Google Colab Notebook: [Python Notebook]
Link to GitHub Repository: [GitHub Repo]

1.1. UNDERSTANDING THE PROBLEM

Businesses must constantly be on the lookout for the newest and greatest developments to remain relevant in the ever-expanding commercial sector. A new film studio is the next best thing for Microsoft. That is only a concept to them, unrealized as of yet. They need insights into the movie industry before they settle on that investment.

1.2. PROBLEM STATEMENT

Understanding the movie industry in order to assist Microsoft realize their concept is the current issue. Consequently, this project does an exploratory data analysis on data from four datasets in order to help Microsoft understand the film industry.

1.3 BUSINESS OBJECTIVES

1) To identify studio with the highest titles.
2) To identify studio with the highest total gross.
3) To identify most popular titles (in terms of highest vote count)
4) To identify movies most used original language.
5) To identify movies with the highest production budget.
6) To identify genres with the highest number of titles.
7) To identify genres with the highest popularity.
8) To identify genres with the highest profit.

2. DATA UNDERSTANDING

2.1. DATA COLLECTION

Box Office Mojo, tmdb, budgets, and Movie Basics were four trustworthy sources from which the four datasets used for the analysis were gathered. The datasets were in three different formats, including a database, tab separated values (TSV), and comma separated values (CSV). While IMDB deals with movie basics, the box office is mostly concerned with the money movie studios make, tmdb deals with popularity and budget deals with the movie budget. This sources are trustworthy because the calculations were made using unbiased formulas.

2.2. DATA DESCRIPTION

Column	**Description**
movie	The name of the movie
date	date of the row aggregation
genres	Different categories different identical movies belong to
studio	Production firm where individual movie is being produced
production_budget	The total cost of movie production
domestic_gross	Amount of income made from local market
worldwide_gross	Amount of income from market outside the country
Original_title	Same as title of the movie
original_language	Language used in the original production
popularity	How a specific movie is well know
vote_average	Average number of vote voted for an individual movie
vote_count	Total count of votes per movie
runtime_minutes	Total minutes per movies

3. DATA PREPARATION

3.1. SELECTING DATA

We'll use all of the columns relevant to our analysis and drop columns that will not be of great importance.

3.2. DATA CLEANING

This was done to ensure the Validity, Accuracy, Completeness, Consistency and Uniformity of the Data.

The first thing done was to rename the columns to make them uniform and readable. The columns were then checked to see if they were of the appropriate types / dtypes. After this, missing values in the datasets were checked for and were found to be none. The data was also found to be consistent there being no duplicated data.

4. DATA ANALYSIS

4.1. EXPLORATORY DATA ANALYSIS

4.1.1. UNIVARIATE DATA ANALYSIS

a. Numerical Data

Productive budget, domestic gross, worldwide gross, vote count, vote average

b. Categorical Data

Time, genre, original language

c. Summary Statistics

Statistic	***production_budget***	***domestic_gross***	***worldwide_gross***
mode	20000000.0	13	0
range	5782	5782	5782
Standard deviation	41812076.83	68240597.35690415	174719968.78
Variance	1748249768582191.8	4656779127627114.0	3.052706749010146e+16
1st Quartile (Q1)	5000000.0	1429534.5	4125414.75
Median (Q2)	17000000.0	17225945.0	27984448.5
3rd Quartile (Q3)	40000000.0	52348661.5	97645836.5
Skewness	2.7183	3.7589	4.4914
Kurtosis	10.2859	22.4188	31.928
mean	31,587,757.096	41,873,326.867	91487460.90

5. CONCLUSION

IFC studios has the highest number of production (has the highest number of title) 166 in total.

BV studio have the highest total gross of $44,250,280,000. En original language was the most used language with a total of 22380 vote count.

Foreign gross and total gross are highly positively correlated with a value of 0.968957. Vote count and popularity are highly positively correlated with a value of 0.685287.

Inception is the title with the highest vote count of 22186. Avatar is the movie with the highest Production Budget.

Production budget, domestic gross and worldwide gross are positively skewed. Production budget is correlated with domestic gross; worldwide gross is positively skewed with domestic gross.

Adventure, Fantasy, Mystery Genre has the highest popularity.

Adventure, Drama, Sport Genre has the highest profit of $1,373,208,000.

Should also consider set aside production budget of around $425,000,000 (maximum production budget).

6. RECOMMENDATION

Language to use

To use EU as the original language since it’s the most used and voted.

Genre to produce

To choose from the following genres;

1. Drama since it most big companies have also invested greatly on them.

2. Adventure, Fantasy, Mystery Genre since they are the most popular.

3. Adventure, Drama, Sport Genre since they have the highest profit.

Production budget

To set aside production budget of around $425,000,000 (maximum production budget).

To set aside $31,587,757.096 production budget.