MICROSOFT MOVIE INSIGHT PROJECT

BUSINESS UNDERSTANDING

Link to Google Colab Notebook: [Python Notebook]

Link to GitHub Repository: [GitHub Repo]

Understanding the Problem

A good businessperson is one who is constantly looking for a market niche to fill or a growing company to invest in. Whatever the reason, investing in a new company involves risk and expense, so for an investor, the true question is always: Is the risk worth taking? Microsoft wants to start a movie studio, but they know very little about the movie business. Therefore, this project does an exploratory data analysis on information from three datasets to assist Microsoft in learning more about the movie industry and determining whether the venture is worthwhile.

Problem Statement

The major issue is deciding whether or not to invest in the new business idea, as well as offering guidance on the aspects to consider before making the investment. What are the variables that are causing other film-related businesses to succeed so admirably?

Business Objectives

To evaluate the expected production budget

To identify which Genre to invest in

To approximate the runtime for the movie in minutes

DATA UNDERSTANDING

Data Collection

The information was gathered from the Internet Movie Database and Box Office Mojo (IMDb). An American website called Box Office Mojo uses an algorithm to systematically track box office receipts. Box Office Mojo was and continues to be considered as one of the most dependable box office trackers available, despite the fact that there are other websites that offer comprehensive box office tracking and a small number that offer a resource for final totals. An online database called IMDB contains data about movies, TV shows, home videos, video games, and internet streaming entertainment. IMDb ratings are considered "accurate" because they are produced using a reliable, unbiased formula.

DATA DESCRIPTION

Data Frame	Number rows	Number of columns	Missing values	Duplicates	Columns used
Df1	3387	5	1383	0	Domestic gross, Foreign gross, title

Df2	5782	6	0	0	Productionbudget, domesticgross, worldwide gross, title
Df3	73856	8	8424	0	Average rating, genres, runtime, original title, primary title

DATA PREPARATION

Loading Data

In each dataset all columns will be loaded but only those relevant to the business questions at hand.

Data Cleaning

Data cleaning involves removing outliers, missing and duplicate values to improve the quality and accuracy of the analysis. It also promotes consistency and uniformity of the results. In this analysis, two datasets had missing values. For the first dataset the missing data in the numerical datatype was replaced with the median of the individual columns as they had skewed distribution. For the categorical datatype missing values were replaced with the mode of the respective column. In the second dataset, the missing values were very minimal (10% in one column and less than 5% of the total values) prompting drop as a way of handling them. Below are displays for the first and third dataset's missing value percentages, respectively.

All datasets had too many outliers which if dropped would result in an inaccurate analysis of the data which will lead to inaccurate insights and recommendations. Fortunately, our data did not have any duplicates.

DATA ANALYSIS

Exploratory Data Analysis

UNIVARIATE DATA ANALYSIS

a.) **Distribution of data**

The foreign, production budget, worldwide and domestic gross are positively skewed; for this kind of distribution, most points are clustered towards the left side of the tail. The best measure for this kind of distribution is the median. This also confirms our previous assertion of the data being highly skewed towards the right.The runtime data had a normal distribution. Insinuating that its best measure of central tendency is the mean.

b.) **Summary Statistics**

Columns	Mean	Median	Standard Dev.	Max.	Min.

Product budget	46994109.88211708	25000000.0	55942082.485717684	410600000.0	5000
Profit	165730412.65677625	67249045.0	275149073.38101226	2426949682.0	-89057484.0
runtime minutes	94.7322732805843	91.0	209.37701687634805	51420.0	3.0

BIVARIATE DATA ANALYSIS

Correlation Between Variables of Interest

There was a strong positive correlation between production budget and the profit earned (0.677062).There was a weak positive correlation of 0.1278 between movie runtime and total gross earned.There exists a weak positive correlation of 0.1222 between runtime and average rating. This implies that a change in either of the variables will have very little or no effect on the other.

The adventure genre had the highest average rating (9.2) .The adventure drama sport genre had the highest total gross of 1,276,400,000 dollars.Adventure drama Sci-Fi had the longest average runtime of 156.5 minutes.P/DW is the highest earning studio with a profit of 542,694,200 dollars.P/DW studio has the highest production budgets of 133,400,000.

Bivariate Analysis Recommendation

To choose what genre to invest in look at three factors; the total gross, the number of votes and the average rating. In the analysis above, there is one consistent genre in the top five of each category mentioned that is Adventure, Drama and Sci-Fi. Therefore, I would recommend Microsoft to invest in Adventure Drama Sci-Fi.

From the analysis above, P/DW studio is the highest earning studio with the highest production budget. The studio is earning 542,694,200 dollars as profit from as little as 133,400,000. Given that our data is highly skewed, the best measure for central tendency would be the median. Hence to start the movie studio I recommend that Microsoft sets aside 25,000,000 dollars as their production budget for the investment. The production budget might be higher than this but it is good to remember that the production budget and the profit have a strong (0.6) positive correlation.

I would recommend a movie with a duration of 107minutes or longer. The data had a normal distribution prompting the use of the mean as or central measure of tendency.