Investment strategy for development of movie industry

1) Project Overview

1.1 Background

- The global film industry generates billions of dollars annually and produces thousands of movies across various genres and countries.
- Understanding the factors that drive box office success such as budget, genre, country and audience ratings is crucial for production companies and investors.
- This project aims to apply data analytics and visualization techniques to explore the relationships between movie features, financial performance, and audience perception, helping identify key factors that contribute to a movie's success.

1.2 Objectives

- Analyze the relationship between budget, revenue, profit, and ratings.
- Identify top-performing genres and countries in terms of box office success.
- Explore correlations among quantitative factors (e.g., Budget, IMDb Rating, Profit).
- Apply What-if Analysis to simulate the effect of budget changes on profit.
- Provide actionable insights for movie producers, marketers, and investors.

2) Data Sources

*** Dataset:** Movie dataset sourced from:
https://www.kaggle.com/code/emanfatima2025/exploratory-data-analysis-of-movies-dataset/input

*** Data Description:**

- The dataset is a collection of nearly 1 million films worldwide, capturing detailed information about production, revenue, and reviews, used to analyze trends, discover insights and predict film success.
- It has totally 999.999 rows and 17 columns including:
  - ❖ MovieID: unique ID for the movie
  - ❖ Title: movie name

- ReleaseYear: The year when the movie was released
- ReleaseDate: The datetime when the movie was released
- Coutry: The name of the nation which released the movie
- BudgetUSD: The amount of money that was spent to budget the movie
- US_BoxOfficeUSD: The movie's domestic box office revenue
- Global_BoxOfficeUSD: The movie's global box office revenue
- Opening_Day_SalesUSD: The movie's opening day revenue
- One_Week_SalesUSD: Revenue earned in the first week of the movie's release
- IMDbRating: Average IMDb score rating (1–10 scale)
- RottenTomatoesRating: Rotten Tomatoes critic or audience rating (0–100%)
- NumVotesIMDb: Number of audience votes on IMDB
- NumVotesRT: Number of audience votes on RottenTomatoe
- Director: The director of the movie
- LeadActor: The lead actor of the movie

3) Project Workflow and Methodology

- Step 1: Data Collection:
  - Download the marketing dataset from the provided source.
  - Store the dataset in an accessible format (CSV, Excel) for analysis.
- Step 2: Data Cleaning & Preprocessing:
  - Load dataset into Python Notebook for preprocessing.
  - Check for missing values, outliers and inconsistencies.
  - Convert a column with numerical data to object
  - Convert a column with object data to datetime
  - Add two new columns in order to analyze.

- Step 3: Exploratory Data Analysis (EDA) – Analyze key statistics and visualize financial and rating distributions:
    - Check important qualitative attributes: Count the number of different values in each attribute
    - Check the distribution of quantitative variables with each other
    - Using the Outlier method to detect outliers for the most important attribute
    - Check the distribution of quantitative variables
    - Analyze the correlation (relationship) between quantitative variables in the dataset
- Step 4: Data Visualization (Power BI & Python) : Generate visual results that are different charts that represent quantitative variables against qualitative variables through Power BI and Python.
- Step 5: What-if Analysis: Apply LinearRegression model in machine learning to predict revenue based on budget and predict profit based on number of votes
- Step 6: Model Evaluation: To evaluate the performance of the linear regression model, the following metrics will be used:
    - Mean Absolute Error (MAE): Measures the average absolute errors
    - between predicted and actual purchases.
    - Mean Squared Error (MSE): Measures the average squared errors,
    - penalizing larger errors more.
    - Root Mean Squared Error (RMSE): Provides a more interpretable
    - measure of model accuracy.
    - R-squared (R2 Score): Determines how well the model explains the
    - variance in the data.
- Step 7: Reporting & Storytelling : Summarize and present results in Power BI dashboards and charts made by Python.

4) Expected Outcomes

- Identification of key success drivers (budget, genre, rating, or region).
- Understanding of how audience perception relates to revenue.
- Data-driven recommendations for film investment and marketing focus.
- Interactive Power BI dashboard for film performance tracking.
- Documented analytical workflow for entertainment data analysis.

5) Tools and Technologies

- Data Collection and Reprocessing: Python (Pandas, NumPy, Matplotlib, Seaborn)
- Visualization, EDA: Power BI, Python (Matplotlib)
- Machine Learning: Scikit-learn (LinearRegression)
- Data Storage: CSV file format.

6) Conclusion

This project integrates data analytics, visualization, and AI reasoning to analyze nearly one million movies worldwide. Through exploratory and predictive techniques, it uncovers patterns explaining how production budgets, genres, and ratings contribute to box office success. The findings will help filmmakers, producers, and analysts make informed decisions regarding investment allocation, marketing focus, and production planning and fostering a data-driven approach to the film industry's future growth.