IMDB-Movie Data Set EDA Project

❖ Problem Statement:

The rookie movie producer has enlisted the expertise of a data scientist to analyze a dataset containing information on 3,000 movies. The objective is to extract meaningful insights that will guide decision-making in the production process. The dataset needs to be thoroughly explored, cleaned, and analyzed to provide actionable recommendations for the producer.

❖ Background:

Hired by a rookie movie producer, I analyze a dataset of 3,000 movies, ensuring accuracy through data cleaning. Uncovering trends and patterns, I identify profitable movies, popular genres, and successful actors. By developing predictive models based on genre, budget, and cast, I provide recommendations, guiding the producer to make informed decisions for a successful venture in the film industry.

❖ Dataset Information:

The imdb_data.csv file contains a dataset is a collection of information about movies. from the imdb platform. The dataset has movie details like ID, budget, genres, release date, revenue, and production info for analysis. The dataset contains 3000 rows of data

❖ Objective and Description of Project:

This project supports a novice movie producer by cleaning and validating a dataset of 3,000 movies, ensuring accuracy. The analysis uncovers trends in profitability, popular genres, and successful actors. Utilizing predictive models based on variables like genre and budget, the recommendations include genre-specific production advice and casting choices backed by historical success. Ultimately, the aim is to empower the producer with actionable insights, aligning data-driven recommendations with strategic decisions for increased success in the competitive film industry.

Further, you have to answer the following questions:

1. Which movie made the highest profit? Who were its producer and director? Identify the actors in that film.
2. This data has information about movies made in different languages. Which language has the highest average ROI (return on investment)?
3. Find out the unique genres of movies in this dataset.
4. Make a table of all the producers and directors of each movie. Find the top 3 producers who have produced movies with the highest average RoI?
5. Which actor has acted in the most number of movies? Deep dive into the movies, genres and profits corresponding to this actor.
6. Top 3 directors prefer which actors the most?

❖ **Data Dictionary:**

"imdb_data.csv": The imdb_data.csv file contains a dataset is a collection of information about movies. from the imdb platform. The dataset has movie details like ID, budget, genres, release date, revenue, and production info for analysis. The dataset contains 3000 rows of data

1. **id:** Unique identifier for each row.
2. **belongs_to_collection**: Information about movie collections and associated posters.
3. **budget:** The budget allocated for each movie.
4. **genres:** Types of movies, such as drama, comedy, horror, etc.
5. **homepage:** URL of the movie's official website.
6. **imdb_id:** IMDb identifier for the movie.
7. **original_language:** The original language in which the movie was produced.
8. **original_title**: The original title of the movie.
9. **overview:** A brief summary or description of the movie.
10. **popularity:** Popularity score of the movie.
11. **poster_path:** Path to the movie's poster image.
12. **production_companies:** Companies involved in the movie's production.
13. **production_countries:** Countries where the movie was produced.
14. **release_date:** The date when the movie was released.
15. **runtime:** Duration of the movie in minutes.
16. **spoken_languages:** Languages spoken in the movie.
17. **status:** Current status of the movie (e.g., released, in production).

18. **tagline:** A memorable phrase associated with the movie.
19. **title:** The title of the movie.
20. **Keywords:** Keywords associated with the movie.
21. **Castcrew:** Information about the cast and crew.
22. **revenue:** The revenue generated by the movie.

❖ Null values of each column:

id	0
belongs_to_collection	2396
budget	0
genres	7
homepage	2054
imdb_id	0
original_language	0
original_title	0
overview	8
popularity	0
poster_path	1
production_companies	156
production_countries	55
release_date	0
runtime	2
spoken_languages	20
status	0
tagline	597
title	0
Keywords	276
cast	13
crew	16
revenue	0

IMDB-Movie Data Set EDA Project

Approach:

1. **Data Exploration and Preprocessing:**

- ✓ Perform exploratory data analysis (EDA) to gain insights into the dataset and identify patterns.
- ✓ Handle missing values by imputing or dropping them based on the analysis.
- ✓ Clean the data by removing irrelevant columns and filtering out invalid records.
- ✓ Encode categorical variables to numerical values for model compatibility.
- ✓ Scale numerical features to ensure they are on a similar scale.

2. **Feature Engineering:**

- ✓ Create new features if necessary, such as the total number of nights stayed by guests.
- ✓ Analyze feature correlations and select relevant features for model training.

IMDB-Movie Data Set EDA Project

Framework:

Importing Required Libraries: Begin by importing the necessary libraries such as pandas, numpy, matplotlib, and seaborn. These libraries will be used for data manipulation, visualization, and building machine learning models.

Loading the Dataset: Read the "Imdb_data" dataset using the pandas library's read_csv() function. The dataset should be stored in a suitable location. Display the first few rows of the dataset using the head () function to understand its structure and contents.

Data Exploration and Preprocessing:

- ✓ Perform exploratory data analysis (EDA) to gain insights into the dataset. Use visualization libraries like matplotlib and seaborn to create meaningful plots and identify patterns in the data.
- ✓ Check for missing values using the isna(). sum() function and calculate the percentage of null values for each feature.
- ✓ Handle missing values by filling them with appropriate values. In the provided code, the fillna() function is used to replace missing values with zeros.
- ✓ Visualize missing values using the msno.bar () function from the missing no library to get a visual representation of the missing data.

Exploratory Data Analysis (EDA):

- ✓ Conduct various analyses to understand the data and extract insights.
- ✓ Explore the countries from which the most guests are coming using visualizations like choropleth maps.
- ✓ Analyze the distribution of room prices per night based on different room types using box plots.
- ✓ Examine how the room prices vary over the months to identify seasonal patterns.