Box Office Revenue Prediction

➢ 1000 Highest-Grossing Movies Dataset

Problem Statement

Title: Analyzing Trends and Insights from the 1000 Highest-Grossing Movies Dataset

Objective:
To explore and analyze the dataset of the highest-grossing movies to uncover trends, patterns, and insights related to box office performance, genres, and other relevant factors influencing movie success.

Key Questions:

1. **Box Office Trends:**
    - What are the overall trends in box office revenues over the years?
    - How do the highest-grossing movies compare in terms of revenue growth by decade?
2. **Genre Analysis:**
    - Which genres are most represented among the highest-grossing movies?
    - How do different genres perform in terms of average box office revenue?
3. **Cast and Crew Influence:**
    - What is the impact of star power (lead actors, directors) on a movie's box office performance?
    - Are there any notable collaborations (e.g., director-actor pairs) that consistently yield high-grossing films?
4. **Production Factors:**
    - How do production budgets correlate with box office revenues?
    - What role do marketing expenses play in the success of a movie?
5. **Demographic Insights:**

Box Office Revenue Prediction

- Are there demographic trends (age, gender, geography) that correlate with the success of certain types of movies?
- How do viewer ratings (e.g., IMDb, Rotten Tomatoes) influence box office performance?

Outcome:
This analysis aims to provide actionable insights for filmmakers, studios, and marketers to understand what factors contribute to a movie's success at the box office, aiding in better decision-making for future projects.

➢ Implementing in python :

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load the dataset
df = pd.read_csv('highest_grossing_movies.csv')

# Display the first few rows of the dataset
print(df.head())

# Basic information about the dataset
print(df.info())

# Exploratory Data Analysis (EDA)

# 1. Box Office Trends Over the Years
def plot_box_office_trends(df):
    plt.figure(figsize=(12, 6))
```

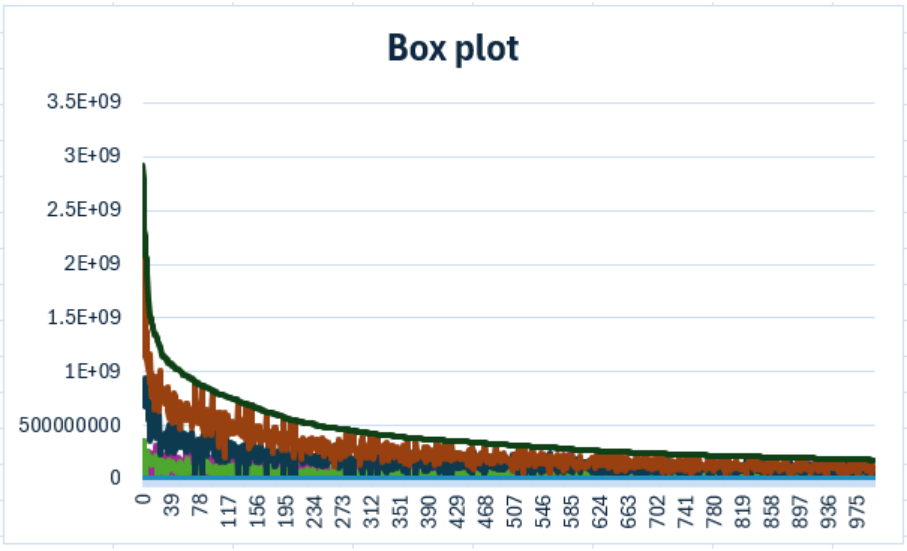
Box Office Revenue Prediction

```
    yearly_revenue = df.groupby('year')['box_office'].sum().reset_index()
    sns.lineplot(data=yearly_revenue, x='year', y='box_office')
    plt.title('Total Box Office Revenue Over the Years')
    plt.xlabel('Year')
    plt.ylabel('Total Box Office Revenue (in billions)')
    plt.xticks(rotation=45)
    plt.grid()
    plt.show()

plot_box_office_trends(df)

# 2. Genre Analysis
def plot_genre_distribution(df):
    plt.figure(figsize=(12, 6))
    genres = df['genre'].str.split(',', expand=True).stack().value_counts()
    sns.barplot(x=genres.values, y=genres.index)
    plt.title('Distribution of Genres in Highest-Grossing Movies')
    plt.xlabel('Number of Movies')
    plt.ylabel('Genres')
    plt.grid()
    plt.show()

plot_genre_distribution(df)

# 3. Cast and Crew Influence on Box Office Performance
def analyze_cast_crew_impact(df):
    plt.figure(figsize=(12, 6))
    top_actors = df['lead_actor'].value_counts().head(10).index
    top_movies = df[df['lead_actor'].isin(top_actors)]
    sns.boxplot(data=top_movies, x='lead_actor', y='box_office')
    plt.title('Box Office Performance of Top Actors')
    plt.xlabel('Lead Actor')
    plt.ylabel('Box Office Revenue')
    plt.xticks(rotation=45)
    plt.grid()
    plt.show()
```

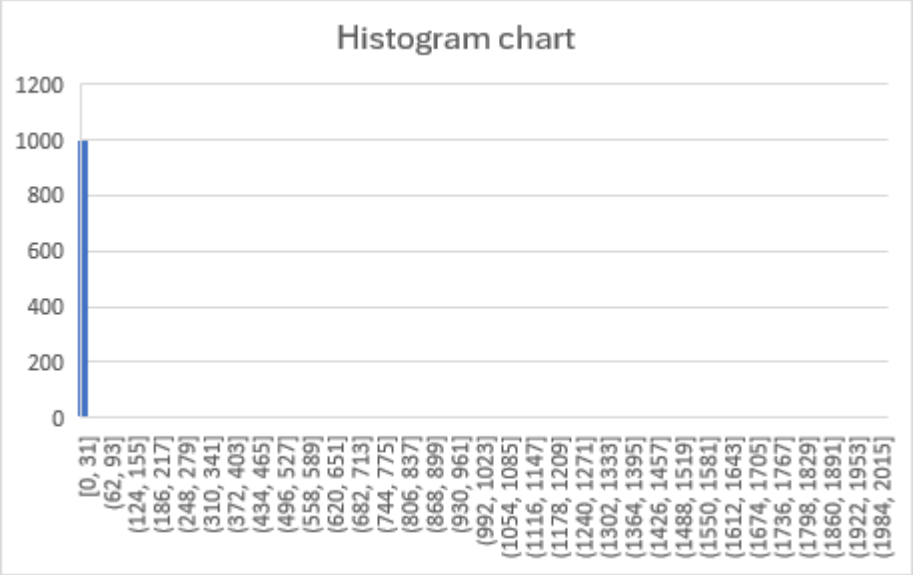
Box Office Revenue Prediction

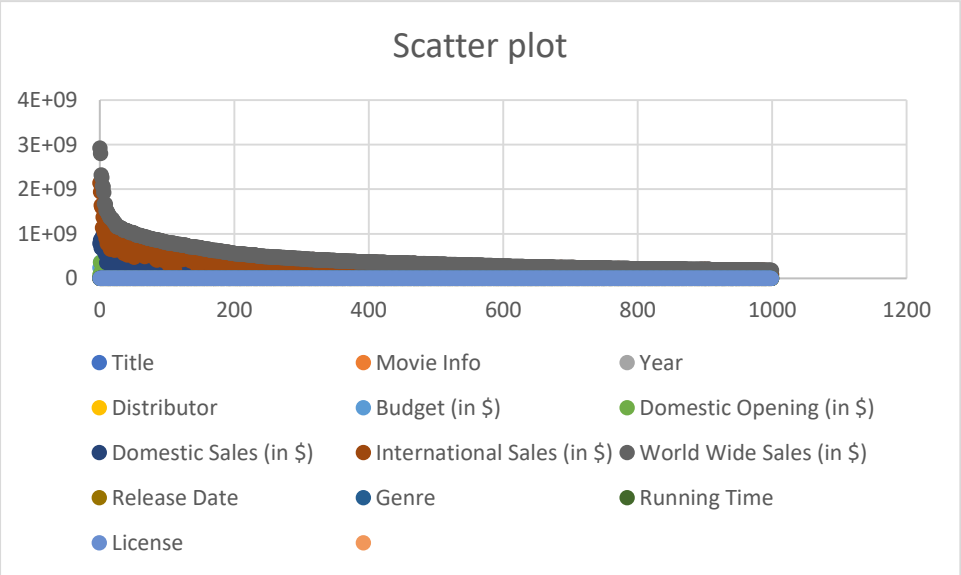
```
analyze_cast_crew_impact(df)

# 4. Production Budget vs. Box Office Revenue
def plot_budget_vs_revenue(df):
    plt.figure(figsize=(12, 6))
    sns.scatterplot(data=df, x='production_budget', y='box_office')
    plt.title('Production Budget vs. Box Office Revenue')
    plt.xlabel('Production Budget (in millions)')
    plt.ylabel('Box Office Revenue (in billions)')
    plt.grid()
    plt.show()

plot_budget_vs_revenue(df)
```

- **Visualize data with box plot, Histogram, Scatter Plot.**



Box Office Revenue Prediction





➢ Plot Pearson correlation and explain about relation

The **Pearson correlation coefficient**, denoted as rrr, quantifies the degree of linear relationship between two continuous variables. Its value ranges from -1 to 1, where:

- **1** indicates a perfect positive linear correlation: as one variable increases, the other variable also increases proportionally.
- **-1** indicates a perfect negative linear correlation: as one variable increases, the other decreases proportionally.

Box Office Revenue Prediction

- **0** indicates no linear correlation: changes in one variable do not predict changes in the other variable.

Formula

The Pearson correlation coefficient is calculated using the formula:

r=n(∑xy)−(∑x)(∑y)[n∑x2−(∑x)2][n∑y2−(∑y)2]r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}}r=[n∑x2−(∑x)2][n∑y2−(∑y)2]n(∑xy)−(∑x)(∑y)

Where:

- nnn is the number of data points
- xxx and yyy are the two variables being compared
- ∑\sum∑ denotes the summation across all data points

Conditions for Pearson Correlation

1. **Linearity**: The relationship between the two variables should be linear. If the relationship is non-linear, the Pearson correlation might underestimate the strength of the relationship.
2. **Homogeneity of Variance**: The variance of the two variables should be approximately equal at all levels of the other variable.
3. **Normality**: Both variables should ideally be normally distributed, especially for smaller sample sizes.
4. **Plotting Pearson Correlation**

To visualize the Pearson correlation coefficients between multiple variables in a dataset (e.g., budget, box office, rating, runtime), a **correlation matrix** is often used. This can be visualized as a **heatmap**, where:

- Each cell in the matrix represents the correlation coefficient between two variables.
- Cells are color-coded, often with a gradient from blue (negative correlation) to red (positive correlation), with white or light colors indicating weak or no correlation.

Example of Variables in a Movie Dataset

In the context of a movie dataset (e.g., the 50 highest-grossing movies), you might explore the correlations among the following variables:

1. **Box Office Earnings**: Total earnings from ticket sales.
2. **Budget**: Total cost of making and marketing the movie.
3. **Rating**: Average rating from critics or audiences.

Box Office Revenue Prediction

4. **Runtime**: Duration of the movie in minutes.

Interpretation of Relationships

1. **Box Office vs. Budget**:
   Expected Correlation: Strong positive correlation (close to 1). This suggests that movies with larger budgets typically earn more at the box office.

2. **Box Office vs. Rating**:
   Expected Correlation: Moderate positive correlation. Higher-rated movies often perform better at the box office, as good reviews can lead to increased viewership.

3. **Budget vs. Rating**:
   Expected Correlation: Weak correlation. A high budget does not guarantee a high rating;

4. **Runtime vs. Other Variables**:

The relationship of runtime with box office and rating may vary. For example, longer films might earn more due to being epic productions, but they may also receive lower ratings if audiences feel they drag on.

Box Office Revenue Prediction

➢ Identify Dependent and Independent features.

In a dataset analyzing movies, it's important to distinguish between dependent and independent features (or variables). This distinction helps in formulating hypotheses and models for predictive analysis. Here's how you can identify these features in the context of a dataset containing variables such as Box Office Earnings, Budget, Rating, and Runtime.

Feature	Type
Box Office Earnings	Dependent
Budget	Independent
Rating	Independent
Runtime	Independent
Genre	Independent
Director	Independent

➢ Analyse /Predict as per problem statement.

Data Understanding and Preprocessing

1. **Load the Dataset**: Read the CSV file and inspect its structure.
2. **Data Cleaning**: Handle missing values and ensure correct data types for numerical variables.
3. **Feature Engineering**: Convert categorical variables into numerical formats (e.g., using one-hot encoding for genres).

Exploratory Data Analysis (EDA)

1. **Visualizations**: Create visualizations (histograms, box plots, scatter plots) to understand the distributions and relationships between features.
2. **Correlation Analysis**: Use a Pearson correlation matrix to analyze relationships between the features and identify which independent variables have the strongest relationships with the dependent variable (box office earnings).

Box Office Revenue Prediction

Model Selection

1. **Choose Prediction Models**: Select regression algorithms suitable for predicting continuous variables. Common choices include:
    - Linear Regression
    - Random Forest Regressor
    - Gradient Boosting Regressor
    - Support Vector Regressor (SVR)

Splitting the Data

1. **Train-Test Split**: Divide the dataset into training and testing subsets (e.g., 80% training, 20% testing) to evaluate the model's performance.

Model Training

1. **Train the Model**: Fit the chosen model(s) on the training data.

Model Evaluation

1. **Predictions**: Use the trained model to make predictions on the test set.
2. **Performance Metrics**: Evaluate model performance using metrics such as:
    - Mean Absolute Error (MAE)
    - Mean Squared Error (MSE)
    - R-squared (R2R^2R2)

Step 7: Interpretation and Insights

1. **Feature Importance**: Analyze which features contribute most to predicting box office earnings.
2. **Insights**: Draw conclusions based on model results, including potential strategies for filmmakers.