IMDB Movie Analysis

Objective:

As a data analyst intern at IMDB, you have been tasked with exploring and analyzing the IMDB Movies dataset. Your goal is to answer specific business questions, gain insights into movie trends, and deliver actionable recommendations. Using Python and libraries such as Pandas, NumPy, Seaborn, and Matplotlib, perform analysis to help IMDB better understand genre popularity, rating trends, and factors influencing movie success.

Dataset: https://drive.google.com/file/d/1lruT50ZWD4PtvDbIn4VnepZvSoeO9BrA/view?usp=sharing

Tasks:

1. Project Setup and Data Loading

Task: Load the dataset and perform initial setup.

Questions:

- What libraries are required for this project, and why are they useful in data analysis?
- Load the dataset. What is the shape of the dataset? What does each row and column represent?

2. Data Overview and Basic Exploration

Task: Explore the structure and composition of the dataset.

Questions:

- Use .info() to understand the data types and missing values. What potential issues can you spot?
- Describe the main characteristics of each column using .describe(). What can you infer from the mean, median, and distribution of numerical columns?

3. Data Cleaning

Task: Address missing values, data types, and outliers.

Questions:

- Which columns contain missing values? How would you handle them?
- Are there any columns where data types need conversion (e.g., date, ratings)? Explain your decision.

4. Univariate Analysis: Explore each column individually.

Task: Perform univariate analysis on numerical and categorical variables.

Questions:

- What is the distribution of movie runtimes? Plot a histogram and describe its shape.
- What are the most common genres in the dataset? Use a bar chart to show their distribution.

5. Bivariate Analysis: Explore relationships between two variables.

Task: Use scatter plots, box plots, and correlation analysis.

Questions:

- Is there a relationship between a movie's runtime and its rating? Plot a scatter plot and describe any observed trend.
- How do ratings vary by genre? Use a boxplot to visualize the differences in ratings across genres.
- Is there a correlation between the number of votes a budget and revenue? Create a scatter plot and calculate the correlation coefficient. What can you conclude?

6. Genre-Specific Analysis

Task: Delve deeper into the genre of movies.

Questions:

- Which genre has the highest average rating? Calculate the average rating for each genre and plot the results.
- How does the popularity of genres vary over time? Plot the number of movies released per genre each year.

7. Year and Trend Analysis

Task: Analyze trends over time.

Questions:

- How has the average movie rating changed over the years? Plot the average rating for each year.
- Which years had the highest and lowest number of movie releases? Plot the number of movies released each year.

8. Multivariate Analysis: Analyze multiple variables together.

Task: Combine insights from multiple columns to explore complex relationships.

Questions:

- Which genres are most popular in each decade? Create a bar plot showing the most frequent genres by decade.
- Plot a heatmap or pairplot to examine relationships between budget, revenue, scores.
- Are there specific genres or release years with higher-rated movies? Group by genre and year, then analyze the average rating.

9. Insights and Summary

Task: Summarize key findings**.**

Questions:

- Based on your analysis, what are three major insights you learned about movie trends, popular genres, or movie ratings?
- What additional questions could be explored with this dataset, or what other data would be helpful to gain a deeper understanding?