Problem statement: - *MOVIE RATING PREDICTION*

In the context of predicting movie ratings, the goal is to estimate how well a specific film will score based on different factors like its genre, length, cast, and release year. The dataset you've shared can be used as input for this prediction task, where we seek to create a model that forecasts the movie rating using these characteristics.

The dataset include the following key features:-

1. **Movie Title**: The name of the movie.

2. **Release Year**: The year the movie is released.

3. **Duration**: The length of the movie in minutes.

4. **Cast**: The list of main actors in the movie.

5. **Genres**: The genre(s) of the movie (e.g., Sci-Fi, Drama, Action).

**Libraries Used:**

In a typical movie rating prediction task, several Python libraries are used to process data, train models, and make predictions. Below is an explanation of the key libraries:

1. **Pandas**:-

**Purpose**: It is used for data manipulation and analysis. In the case of this dataset, pandas helps in reading, cleaning, and structuring the data, such as extracting the relevant features (e.g., duration, genre) and handling missing or malformed data.

1. **NumPy**:-

**Purpose**: NumPy is a core library for numerical and matrix operations. It's especially helpful in performing calculations related to features such as movie duration, year, or any transformations applied to the data.

**3. Matplotlib/Seaborn**:

**Purpose**: These are visualization libraries. Matplotlib is used for creating static, interactive, and animated plots, while Seaborn provides an easier interface to create attractive statistical graphics. These libraries are used to visualize relationships in the data (e.g., correlation between movie duration and rating).

**4. Scikit-learn**:

**Purpose**: Scikit-learn is a powerful library for machine learning. It offers a variety of models that can be used for prediction (e.g., regression, classification), as well as tools for preprocessing data (e.g., scaling or encoding categorical variables), splitting datasets into training and testing sets, and evaluating model performance.

5. scipy:-

**SciPy** is a Python library used for scientific and technical computing. It builds on top of **NumPy** and provides a large number of functions and algorithms for mathematical, scientific, and engineering applications. It's commonly used in fields like data analysis, optimization, signal processing, statistics, and more.

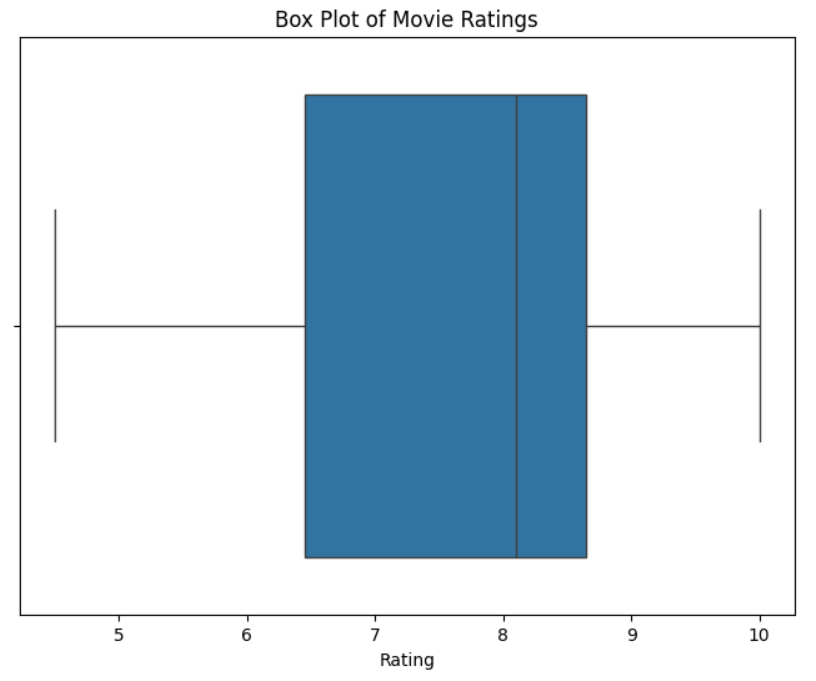
*Process for Rating Prediction*:

**Data Preprocessing**:

* Convert categorical data (like movie genres or actors) into numerical representations, such as using one-hot encoding or label encoding.
* Normalize or standardize numerical features (e.g., duration, release year) to ensure the model treats them equally.
* Handle missing values, if any, using methods like imputation or removal.

Data Visualization

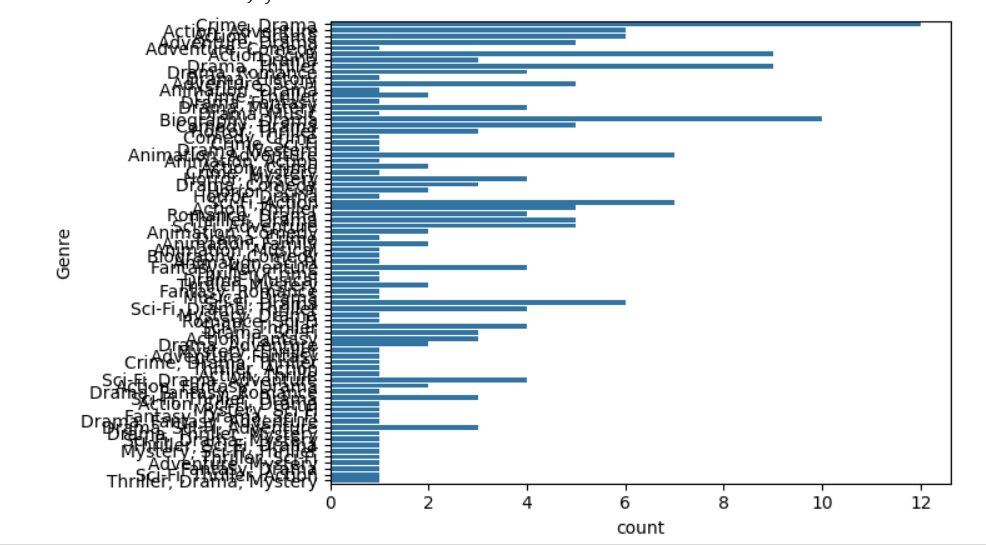
**Data Visualization** is the graphical representation of data and information. It uses visual elements like charts, graphs, and maps to make complex data more accessible and understandable. Data visualization helps to identify trends, patterns, and outliers in data, making it a crucial part of data analysis and decision-making.

1.Box plotting:-

This code creates a **box plot** to visually represent the distribution of movie ratings. A box plot helps show how the ratings are spread out and whether there are any unusual values (outliers).

* The **box** shows the middle 50% of the ratings, with the line inside the box representing the median (the middle rating).
* The **whiskers** (lines extending from the box) show the range of the ratings, excluding any outliers.
* **Outliers** (ratings that are much higher or lower than most of the others) are marked separately.

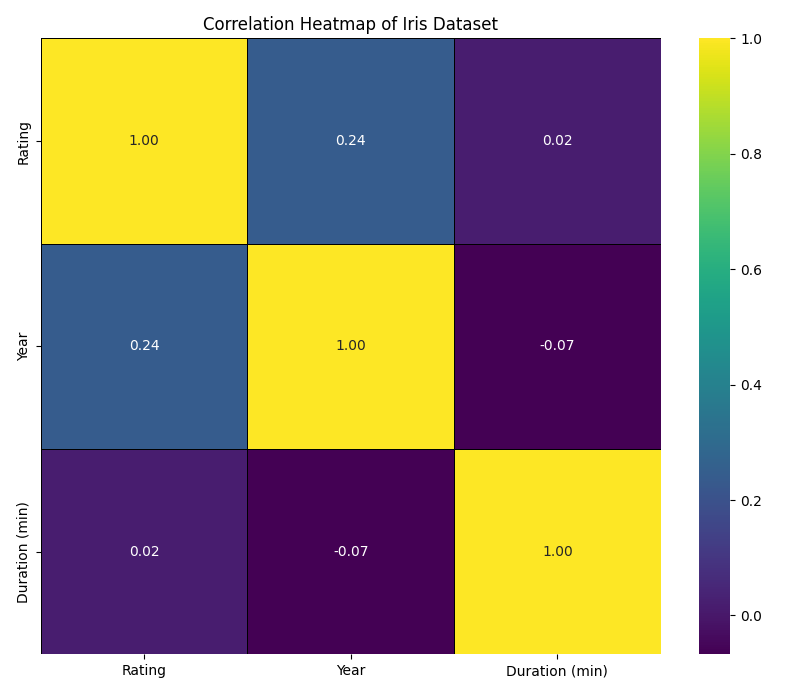
The plot helps us understand the general range of ratings, the typical rating value, and if there are any extreme ratings that stand out.

2. 

**Displays the Count of Each Genre**: It first calculates and prints how many movies there are in each genre (for example, how many Sci-Fi movies, how many Drama movies, etc.). This helps us see the distribution of genres in the dataset.

**Creates a Count Plot**: It then creates a **count plot**, which is a type of bar chart. On the y-axis (vertical axis), it shows the different genres, and on the x-axis (horizontal axis), it shows how many movies belong to each genre. The longer the bar, the more movies there are in that genre.

3.



**Selecting Numeric Data**: It first selects only the numeric columns from the dataset. This is because the correlation matrix is calculated based on numerical data (e.g., ratings, duration, etc.), not categorical data (like genre names or cast members).

**Calculating the Correlation Matrix**: It calculates how strongly the numeric columns in the dataset are related to each other. Correlation values range from -1 to 1:

* A value closer to 1 means a strong positive relationship (as one value increases, the other also increases).
* A value closer to -1 means a strong negative relationship (as one value increases, the other decreases).
* A value close to 0 means no significant relationship.

**Creating a Heatmap**: The correlation matrix is then displayed as a **heatmap**, where each number in the matrix is represented by a color:

* **Darker colors** represent stronger correlations.
* The numbers are also shown on the heatmap to provide precise values for each correlation.
* The title "Correlation Heatmap of Iris Dataset" will be displayed (though this seems to be a placeholder and should refer to your dataset).

Data Visualization (Screen shot) : -  


explain in short

This image shows various visualizations of IMDb movie ratings and genres. Key points include:

1. **Ratings by Movie**: "The Godfather" has the highest rating.
2. **Ratings by Year**: A funnel chart showing how ratings vary by year.
3. **Genres by Rating**: A line chart indicating a peak at rating 10.
4. **Genres by Movie**: "Avatar" has the highest genre count.
5. **Genre-Movie Table**: Lists genres and corresponding movie names.
6. **Actors Count**: There are 203 actors listed.

Final Output:-



Team :-

Data commanders’