Stanley College Of Engineering

and Technology For Women



Amazon Prime Dataset

Name: Sree Nidhi

Roll no: 160623747082

Section : AI&DS-B

Problem Statement

Amazon Prime Video hosts thousands of movies, TV shows, and original content. However, with such a large and diverse users may face difficulties discovering relevant content that matches their preferences. This challenge can lead to reduced user satisfaction and decreased viewing time.

**Problem:**

* Users struggle to find engaging content on Amazon Prime Video due to the overwhelming volume of options and limited personalized recommendations and Amazon lacks actionable insights into content distribution by genre, rating, country, and year that could inform better content acquisition and marketing strategies.
* Proposed Solution

**Step 1:**

* Remove duplicates and handle missing values, Normalize textual fields (e.g., genre, cast) and Convert date added and duration into usable formats.

**Step 2: Exploratory Data Analysis (EDA)**

* Visualize the distribution of content by:
* Type (Movie vs TV Show)
* Genre, Country, Year of release
* Ratings (e.g., TV-MA, PG-13)
* Duration (e.g., average runtime of top genres)

**Step 3: Recommendation System:**

* **Content-Based Filtering** using:

Optional: **Collaborative Filtering** if user watch history is available

**Step 4: Predictive Modeling:**

* Predict which type of content is most likely to trend or be watched more.
* Train a classification model using features like genre, duration, country, and rating.

Step Involved:

1.Importing Libraries

• Essential libraries like pandas, numpy, matplotlib, seaborn, wordcloud , confusion matrix and sklearn are imported.

• These are used for data handling, visualization, and machine learning.

2.Data Loading :

• The dataset is loaded using the pandas library from a .csv file.

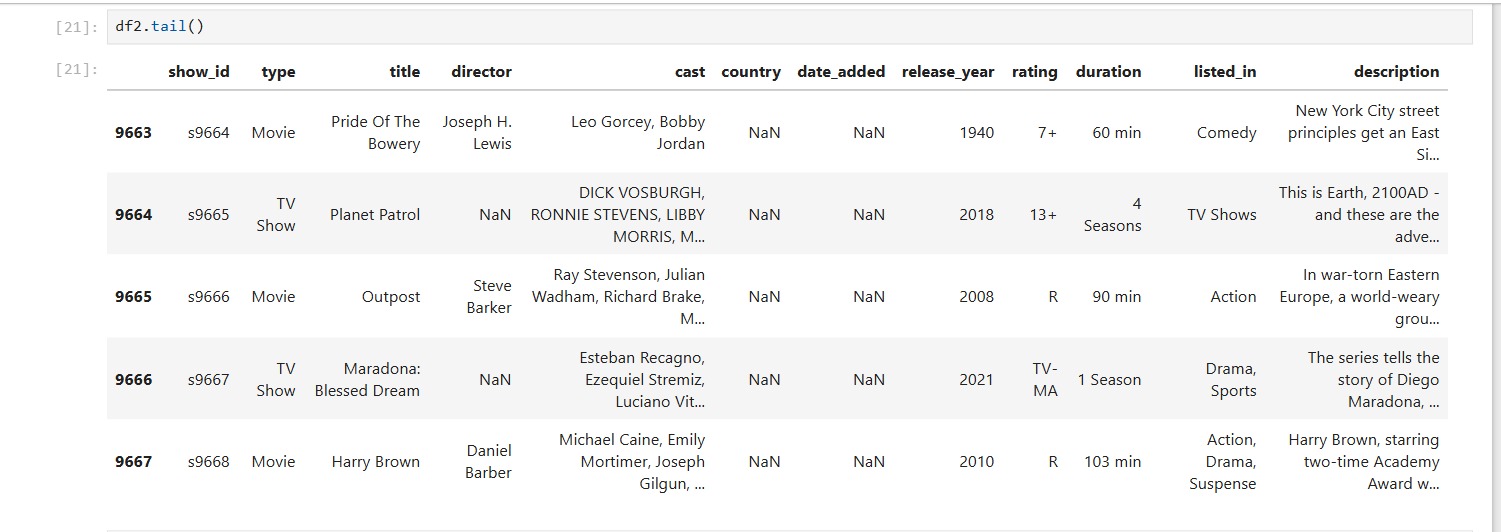
• This is the first step to make the data available for inspection and manipulation.

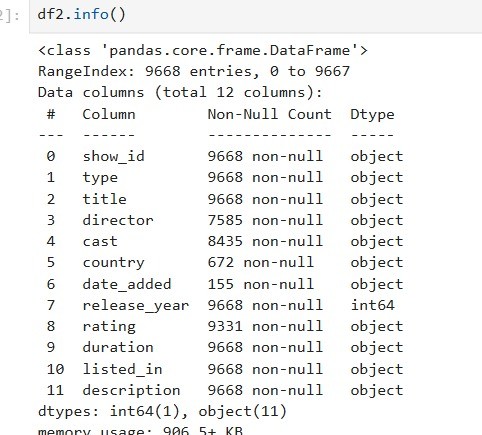
3.Data Inspection

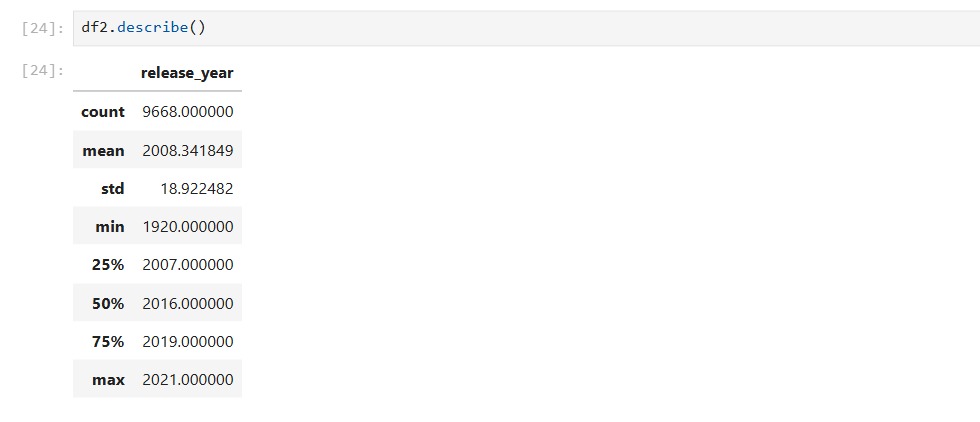
• Functions like .head(), .tail(), .info(), .describe(), and .columns are used.

• These help in understanding the structure, data types, missing values, and summary statistics of the dataset.



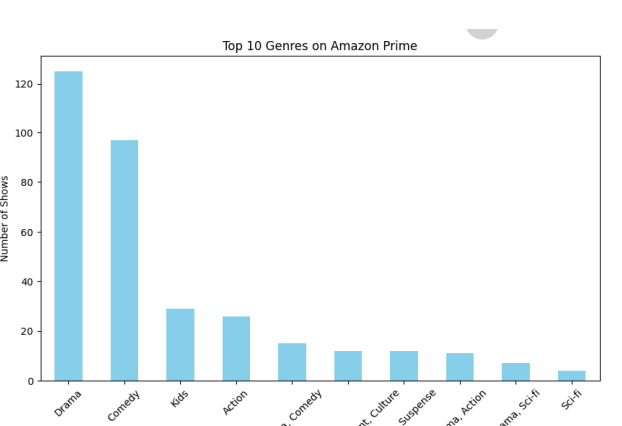






Bar chart





Histogram:

