Exploratory Data Analysis on Netflix Dataset

What is EDA?

Exploratory Data Analysis (EDA) is the process of analyzing datasets to summarize their main characteristics, often using visual methods. It helps understand the data, clean it, discover trends, and prepare it for further modeling.

1. Importing Libraries

Libraries like numpy, pandas, matplotlib, and seaborn are imported for numerical operations, data handling, and visualization.

2. Loading the Dataset

The Netflix dataset is loaded using pandas' read_csv function into a DataFrame.

3. Viewing the Dataset

df.head() is used to preview the first 5 rows of the dataset.

4. Dataset Information

df.info() gives an overview of column types, non-null values, and memory usage.

5. Checking for Null Values

df.isnull().sum() checks how many missing values each column has.

6. Removing Duplicates

df.drop_duplicates(inplace=True) removes repeated rows to avoid biased analysis.

7. Filling Missing Values

Missing values in important columns are filled with placeholders like 'Not Available' or 'Not Rated'.

Exploratory Data Analysis on Netflix Dataset

8. Fixing Data Types

The 'date_added' column is converted to datetime type for easier date-based analysis.

9. Creating New Columns for Year and Month

New columns 'year_added' and 'month_added' are extracted from the date.

10. Visualizing Count of Movies vs TV Shows

Bar chart using seaborn to show the number of movies and TV shows.

11. Top 10 Countries with Most Content

Identifies countries with the most shows/movies on Netflix using value_counts.

12. Year-wise Content Added

Shows how content addition has changed over the years with a bar chart.

13. Top 10 Directors

Displays top directors by number of shows/movies on Netflix.

14. WordCloud for Descriptions

Generates a word cloud from the description text to highlight common terms.