## **EDA-Netflix Titles EDA**

### **Objective**

Analyze Netflix content by:

* Country distribution
* Genre types
* Yearly trends
* Duration of content

### **Dataset**

* **Source**: [Netflix Movies and TV Shows – Kaggle](https://www.kaggle.com/datasets/shivamb/netflix-shows)
* Filename: netflix\_titles.csv

### **Tools**

* **Python Libraries**: pandas, matplotlib, seaborn, plotly
* **Optional**: Tableau for dynamic dashboards

### **Key EDA Tasks**

* Handling missing data
* Count of titles by **genre**, **country**, and **release year**
* Analysis of **TV Show duration** and **Movie runtime**

### **Step-by-Step Code**

import pandas as pd  
import seaborn as sns  
import matplotlib.pyplot as plt  
import plotly.express as px  
  
**# Load dataset**df = pd.read\_csv("netflix\_titles.csv")  
  
# Preview data  
print(df.head())  
print(df.info())  
  
**# --- Handling Missing Values ---**df.fillna({'country': 'Unknown', 'director': 'Unknown', 'cast': 'Unknown'}, inplace=True)  
  
# Convert dates  
df['date\_added'] = pd.to\_datetime(df['date\_added'])  
df['year\_added'] = df['date\_added'].dt.year  
df['month\_added'] = df['date\_added'].dt.month  
  
**# --- Genre Count ---**# Split multiple genres  
df['genre'] = df['listed\_in'].str.split(', ')  
genre\_explode = df.explode('genre')  
genre\_count = genre\_explode['genre'].value\_counts().reset\_index()  
genre\_count.columns = ['Genre', 'Count']  
  
plt.figure(figsize=(10, 6))  
sns.barplot(data=genre\_count.head(10), x='Count', y='Genre', palette='Blues\_d')  
plt.title('Top 10 Genres on Netflix')  
plt.tight\_layout()  
plt.show()  
  
**# --- Year-wise Trend of Content Added ---**yearly = df['release\_year'].value\_counts().sort\_index()  
  
plt.figure(figsize=(12, 5))  
sns.lineplot(x=yearly.index, y=yearly.values)  
plt.title("Trend of Netflix Content Releases Over Years")  
plt.xlabel("Release Year")  
plt.ylabel("Number of Titles")  
plt.grid(True)  
plt.tight\_layout()  
plt.show()  
  
**# --- Country-wise Distribution ---**top\_countries = df['country'].value\_counts().head(10).reset\_index()  
top\_countries.columns = ['Country', 'Count']  
  
fig = px.bar(top\_countries, x='Country', y='Count', title='Top 10 Countries with Most Netflix Titles')  
fig.show()  
  
**# --- Duration Analysis ---**# Split TV Shows and Movies  
tv\_shows = df[df['type'] == 'TV Show']  
movies = df[df['type'] == 'Movie']  
  
**# Movie duration (in minutes)**movies['duration'] = movies['duration'].str.replace(' min', '').astype(float)  
plt.figure(figsize=(10, 4))  
sns.histplot(movies['duration'].dropna(), bins=30, kde=True)  
plt.title('Movie Duration Distribution')  
plt.xlabel('Duration (minutes)')  
plt.tight\_layout()  
plt.show()  
  
**# TV Show seasons**tv\_shows['duration'] = tv\_shows['duration'].str.replace(' Season', '').str.replace('s', '').astype(float)  
plt.figure(figsize=(10, 4))  
sns.countplot(data=tv\_shows, x='duration', order=tv\_shows['duration'].value\_counts().index[:10])  
plt.title('TV Show Season Count Distribution')  
plt.xlabel('Number of Seasons')  
plt.tight\_layout()  
plt.show()

### **Optional –Dashboard Ideas**

Export cleaned CSV using:

df.to\_csv("netflix\_cleaned.csv", index=False)

**And upload:**

* Genre-wise donut chart
* Country-wise choropleth map
* Timeline of release years
* Heatmap: year vs genre

### **Key Insights You Can Present**

* Most common genres on Netflix (Drama, Comedy, etc.)
* Growth pattern: more content added post-2015
* Country dominance: US, India, UK
* Movie lengths mostly between 80–120 mins
* Most TV shows are limited to 1–2 seasons