NETFLIX DATA- ANALYSIS ,CLEANING AND VISUALIZATION PROJECT

NETFLIX

BY-SANJAY SINGH BISHT

Content table

INTRODUCTION

Title- NETFLIX DATA –CLEANING,ANALYSIS AND VISUALIZATION

Netflix is a popular streaming service that offers a vast catalog of movies, TV shows, and original contents. This dataset is a cleaned version of the original version which can be found here. The data consist of contents added to Netflix from 2008 to 2021.

The oldest content is as old as 1925 and the newest as 2021. This dataset will be cleaned with PostgreSQL and visualized with Tableau. The purpose of this dataset is to test my data cleaning and visualization skills.

PROBLEM STATEMENT

- **Understanding Netflix Content Trends**

-Understand the growth and distribution of Netflix content (movies and TV shows) by release year to identify production trends and inform content strategy.

-To explore and analyze Netflix's content dataset to understand content distribution, trends, and insights by performing data cleaning and visualization.

Objective and overview

-Objective: To explore and visualize the trends in Netflix content production over the years and analyze also Netflix's content to identify trends in content types ,geographical contributions and release patterns.
.It’s main work to analyze Netflix content trends using data cleaning and visualization techniques
•**Challenges**:
•Handling missing or inconsistent data in the Netflix dataset (e.g., director, country, or date_added fields).
•Identifying patterns in content type (Movies vs. TV Shows) and release years.
•Visualizing the growth of content to understand production trends.
•**Dataset**: Contains 8,790 entries with 10 attributes (show_id, type, title, director, country, date_added, release_year, rating, duration, listed_in).

🛠 Tools and technologies

Technologies Used for Analysis and Visualization

- **Python**: Core programming language for data processing and analysis.
- **Pandas**: For data manipulation and cleaning (e.g., handling missing values, filtering data).
- **NumPy**: For numerical operations and data handling.
- **Matplotlib & Seaborn**: For creating visualizations such as bar plots to display trends.
- **WordCloud**: For visualizing frequent words in titles or genres (potential use, not shown in provided code).
- **Dataset Source**: 'netflix_titles.csv' loaded into a Pandas DataFrame for analysis.
- **Environment**: jupyter notebook(likely run on google colab).
- **Purpose**: These tools enable efficient data cleaning, analysis, and visualization to derive meaningful insights.

Cleaning ,analysis and visualization process

- 🧹 Data cleaning
- Removed duplicates and dropped rows with null values in director ,country and title.
- Converted date_added to datetime format.
- Extracted year from date_added for trend analysis
- Split duration into numerical and categorical columns(min/seasons).

- 📊 Data analysis and visualization:
- Trend analysis:
- **Content Type Distribution**: More *Movies* than *TV Shows*
- **Top Countries** producing content: *USA*, *India*, *UK*, etc.
- **Rating Distribution**: Majority of content falls under *TV-MA* and *TV-14*
- **Top genres/keywords** visualized via WordCloud.
- Created a derived dataset (df2) with 'Release Year' and 'Total Count' for trend visualization.
- Bar plot created using Seaborn to visualize the count of content produced per release year.

- <u>VISUALIZATION :</u>
  - Used 'Set2' palette for aesthetic and clear differentiation of bars.
  - Plot size set to (10,6) for better readability.
  - Bar plot using Seaborn to display the trend of content produced on Netflix by release year.
  - X-axis: Release Year, Y-axis: Total Count of content (movies and TV shows).
  - Customizations: Rotated x-axis labels (70°) for readability, used 'Set2' color palette.
  - Top countries: US(3240 entries),INDIA(1057),UK(638).
  - Content distribution :

    -Bar plot: movies(70%) dominate over tv shows (30%).

RESULTS

ANALYZING DATASET AND DATA CLEANING

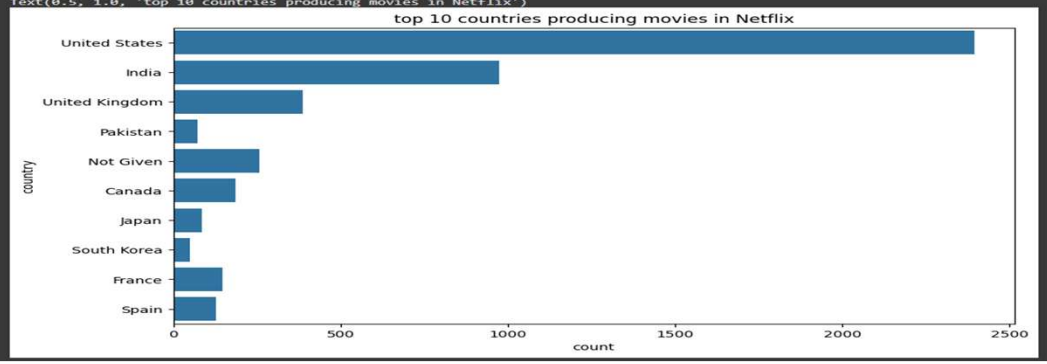
```
[10] df.head()
```

	show_id	type	title	director	country	date_added	release_year	rating	duration	listed_in
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	United States	2021	2020	PG-13	90 min	Documentaries
1	s3	TV Show	Ganglands	Julien Leclercq	France	2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...
2	s6	TV Show	Midnight Mass	Mike Flanagan	United States	2021	2021	TV-MA	1 Season	TV Dramas, TV Horror, TV Mysteries
3	s14	Movie	Confessions of an Invisible Girl	Bruno Garotti	Brazil	2021	2021	TV-PG	91 min	Children & Family Movies, Comedies
4	s8	Movie	Sankofa	Haile Gerima	United States	2021	1993	TV-MA	125 min	Dramas, Independent Movies, International Movies

```
movies_countries=df[df['type']=='Movie']
tv_show_countries=df[df['type']=='TV Show']
```

```
[ ] plt.figure(figsize=(10, 6))
    sns.countplot(y='country',order =df['country'].value_counts().index[0:10],data=movies_countries)
    plt.title('top 10 countries producing movies in Netflix')
```

```
Text(0.5, 1.0, 'top 10 countries producing movies in Netflix')
```



DATA VISUALIZATION

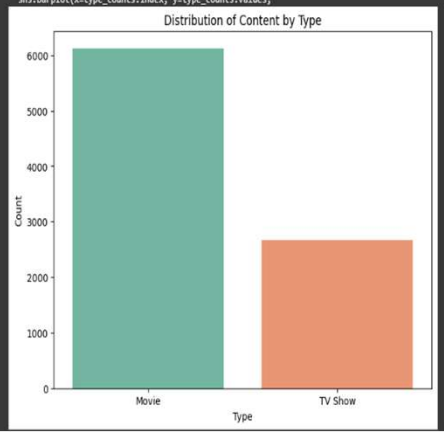
```
df['type'].value_counts()
```

```
count
type
Movie    6126
TV Show  2664
dtype: int64
```

```
[ ] type_counts = df['type'].value_counts()
```

```
plt.figure(figsize=(8, 6))
sns.barplot(x=type_counts.index, y=type_counts.values,
palette='Set2')
plt.title('Distribution of Content by Type')
plt.xlabel('Type')
plt.ylabel('Count')
plt.show()
```

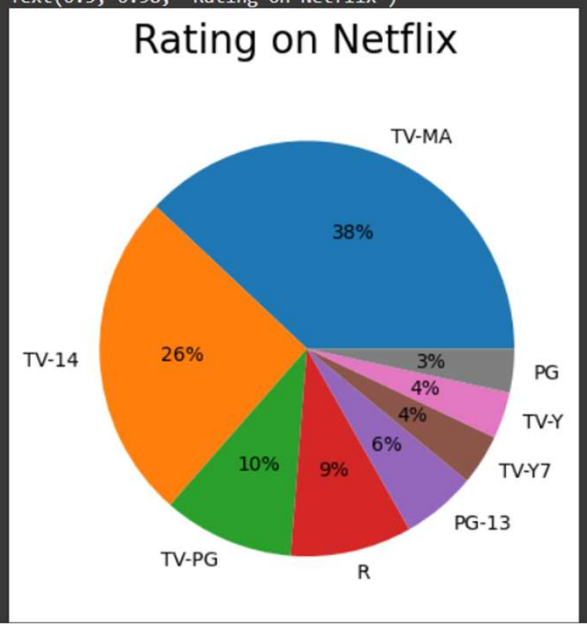
```
<ipython-input-148-d75cf4bc67e5>:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.barplot(x=type_counts.index, y=type_counts.values,
```



RATING AND GENRES OUTCOMES

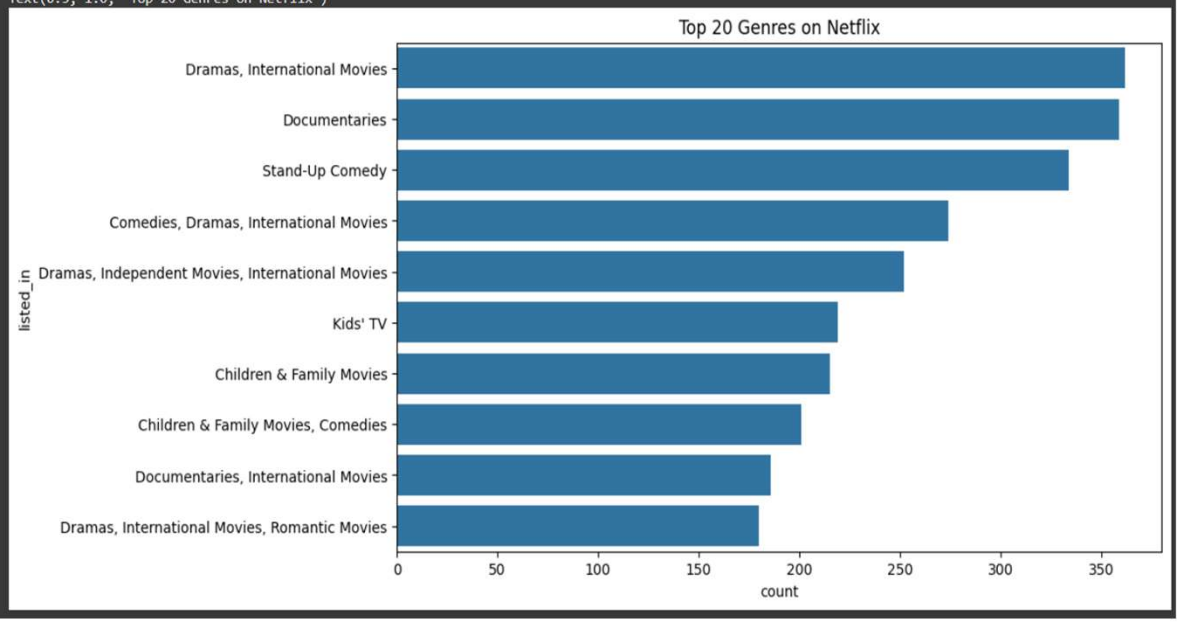
```
plt.pie(ratings['count'][:8], labels=ratings['rating'][:8],
autopct='%.0f%%')
plt.suptitle('Rating on Netflix', fontsize=20)
```

```
Text(0.5, 0.98, 'Rating on Netflix')
```




```
plt.figure(figsize=(10, 6))
sns.countplot(y='listed_in', order=df['listed_in'].value_counts().index[0:10], data=df)
plt.title('Top 20 Genres on Netflix')
```

```
Text(0.5, 1.0, 'Top 20 Genres on Netflix')
```



CONCLUSION

•**Summary**:
•Successfully analyzed Netflix content trends using Python and visualization libraries.
•Data cleaning ensured reliable analysis of 8,790 records.
•Visualizations revealed a clear upward trend in content production, especially post-2010.
•**Key Insights**:
•Netflix has significantly expanded its content library in recent decades.
•Understanding these trends can guide future content investments.
•**Future Work**:
•Analyze content by genre, country, or rating for deeper insights.
•Incorporate predictive modeling to forecast content demand.

THANK YOU