Project Title	**Netflix Data: Cleaning, Analysis, and Visualization**
Tools	**Python, ML**
Domain	**Data Analyst & Data Scientist**
Project Difficulties Level	**intermediate**

Dataset: The dataset is available at the given link.

Click here to download data set

About Dataset: Netflix is a leading streaming platform offering a wide range of movies, TV shows, and original content. This cleaned dataset includes titles added to Netflix from 2008 to 2021, with some content dating back to 1925. The original raw data was cleaned using PostgreSQL to ensure accuracy and consistency. Visualizations were created using Tableau to showcase insights and trends. This project highlights my data cleaning and visualization skills.

1. Importing Required Libraries: Import essential Python libraries for data manipulation and visualization.

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns
```

2. Loading the Dataset: Read the Netflix dataset from a CSV file and display the first few rows.

```
data = pd.read_csv("netflix1.csv")

data.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	9/25/2021	2020	PG-13	90 min	Documentaries
1	s3	TV Show	Ganglands	Julien Leclercq	France	9/24/2021	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...
2	s6	TV Show	Midnight Mass	Mike Flanagan	United States	9/24/2021	2021	TV-MA	1 Season	TV Dramas, TV Horror, TV Mysteries
3	s14	Movie	Confessions of an Invisible Girl	Bruno Garotti	Brazil	9/22/2021	2021	TV-PG	91 min	Children & Family Movies, Comedies
4	s8	Movie	Sankofa	Haile Gerima	United States	9/24/2021	1993	TV-MA	125 min	Dramas, Independent Movies, International Movies

3. Exploring Dataset Information: Check the structure and dimensions of the dataset, including column types and null values.

data.info()

```
<class 'pandas.core.frame.DataFrame'>

RangeIndex: 8790 entries, 0 to 8789

Data columns (total 10 columns):

 #  Column       Non-Null Count  Dtype

---  ------       --------------  -----

 0  show_id      8790 non-null   object

 1  type         8790 non-null   object

 2  title        8790 non-null   object

 3  director     8790 non-null   object

 4  country      8790 non-null   object

 5  date_added   8790 non-null   object

 6  release_year  8790 non-null   int64

 7  rating       8790 non-null   object

 8  duration     8790 non-null   object

 9  listed_in    8790 non-null   object

dtypes: int64(1), object(9)

memory usage: 686.8+ KB
```

```
data.shape
```

```
(8790, 10)
```

4. Removing Duplicate Records: Drop any duplicate rows from the dataset to maintain data quality.

```
data=data.drop_duplicates()
```

5. Content Type Count: Get the frequency count of different content types (Movies and TV Shows).

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

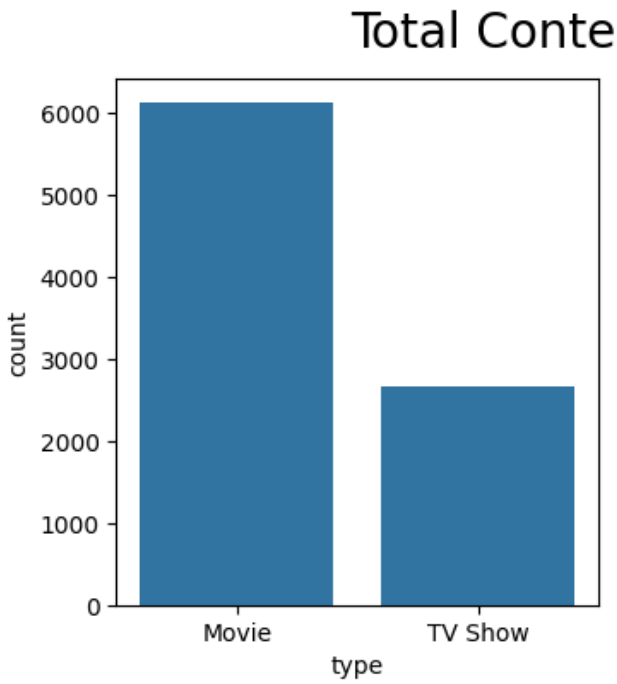
	count
type	
Movie	6126
TV Show	2664

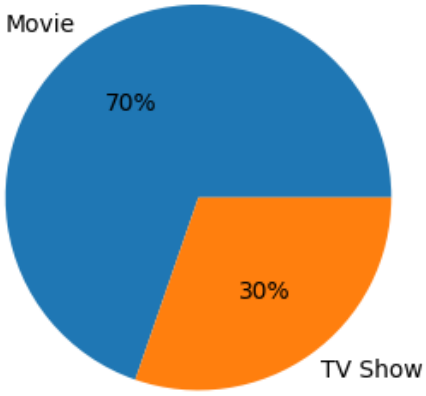
dtype: int64

6. Visualizing Content Type Distribution: Use a bar chart and a pie chart to visualize the proportion of Movies and TV Shows on Netflix.

```
freq=data['type'].value_counts()
fig, axes=plt.subplots(1,2, figsize=(8, 4))
sns.countplot(data, x=data['type'], ax=axes[0])
plt.pie(freq, labels=['Movie', 'TV Show'], autopct='%.0f%%')
plt.suptitle('Total Content on Netflix', fontsize=20)
```

```
Text(0.5, 0.98, 'Total Content on Netflix')
```





7. Dataset Info After Cleaning: Re-check the dataset info to confirm any changes after cleaning.

data.info()

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 8790 entries, 0 to 8789

Data columns (total 10 columns):

#	Column	Non-Null Count	Dtype
0	show_id	8790 non-null	object
1	type	8790 non-null	object
2	title	8790 non-null	object
3	director	8790 non-null	object
4	country	8790 non-null	object
5	date_added	8790 non-null	object
6	release_year	8790 non-null	int64
7	rating	8790 non-null	object
8	duration	8790 non-null	object
9	listed_in	8790 non-null	object

dtypes: int64(1), object(9)

memory usage: 686.8+ KB

8. Rating Value Counts: View the distribution of different content ratings on Netflix.

data['rating'].value_counts()

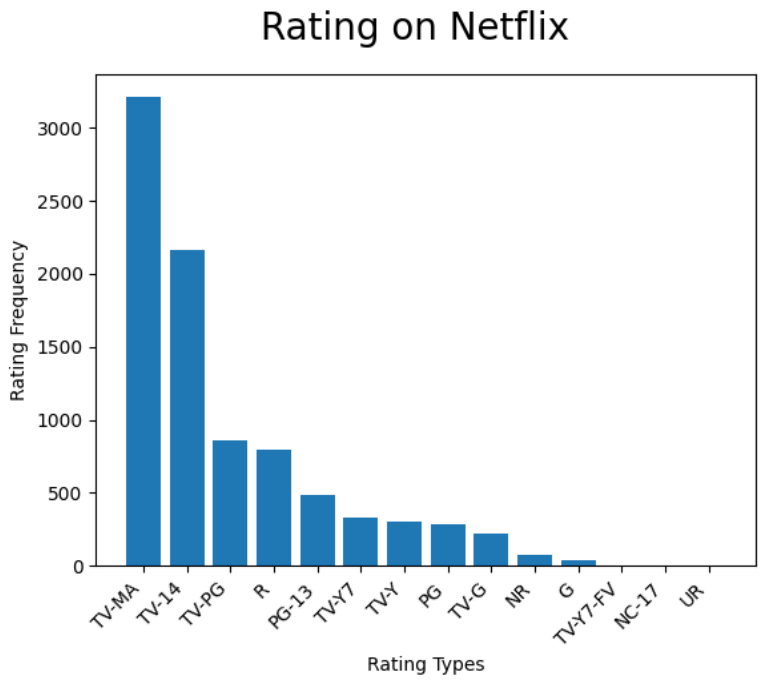
rating	count
TV-MA	3205
TV-14	2157
TV-PG	861
R	799
PG-13	490
TV-Y7	333
TV-Y	306
PG	287
TV-G	220
NR	79
G	41
TV-Y7-FV	6
NC-17	3
UR	3

dtype: int64

9. Visualizing Ratings Distribution (Bar Chart): Plot a bar chart showing how frequently each rating appears on Netflix.

```
ratings=data['rating'].value_counts().reset_index().sort_values(by='count', ascending=False)

plt.bar(ratings['rating'], ratings['count'])

plt.xticks(rotation=45, ha='right')

plt.xlabel("Rating Types")

plt.ylabel("Rating Frequency")

plt.suptitle('Rating on Netflix', fontsize=20)
```

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

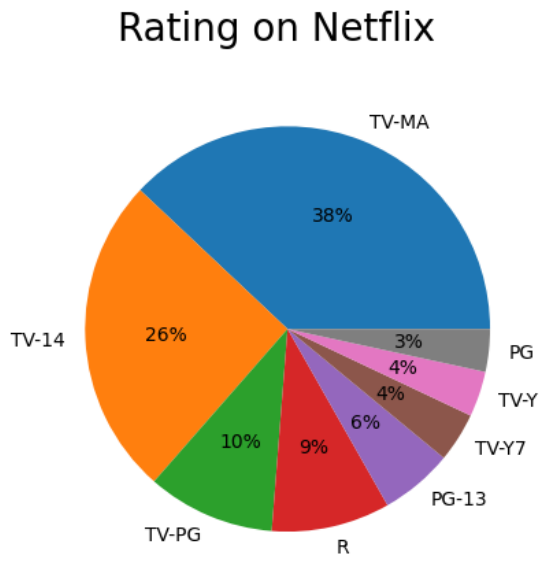


10. Visualizing Top Ratings (Pie Chart): Create a pie chart for the top 8 most common content ratings on Netflix.

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')



11. Converting Date Column to DateTime: Convert the date_added column to datetime format for time-based analysis.

```
data['date_added']=pd.to_datetime(data['date_added'])
```

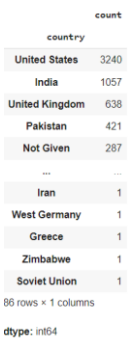
12. Summary Statistics: Get descriptive statistics for numerical columns in the dataset.

```
data.describe()
```

	date_added	release_year
count	8790	8790.000000
mean	2019-05-17 21:44:01.638225408	2014.183163
min	2008-01-01 00:00:00	1925.000000
25%	2018-04-06 00:00:00	2013.000000
50%	2019-07-03 00:00:00	2017.000000
75%	2020-08-19 18:00:00	2019.000000
max	2021-09-25 00:00:00	2021.000000
std	NaN	8.825466

13. Country-wise Content Count: Count the number of contents produced by each country.

```
data['country'].value_counts()
```



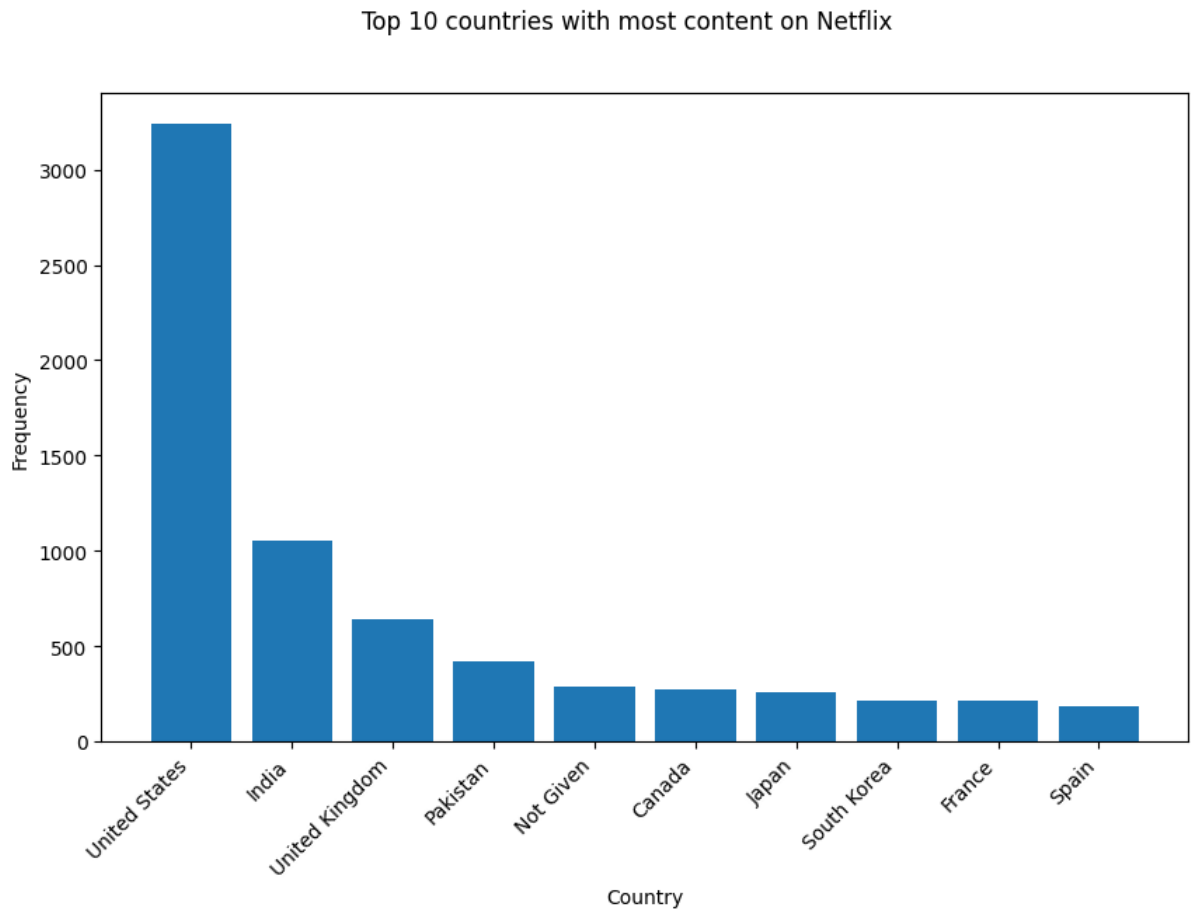
country	count
United States	3240
India	1057
United Kingdom	638
Pakistan	421
Not Given	287
...	...
Iran	1
West Germany	1
Greece	1
Zimbabwe	1
Soviet Union	1

86 rows × 1 columns

dtype: int64

14. Top 10 Countries with Most Content: Visualize the top 10 countries that have the highest number of titles on Netflix using a bar chart.

```
top_ten_countries=data['country'].value_counts().reset_index().sort_values(by='count', ascending=False)[:10]

plt.figure(figsize=(10, 6))

plt.bar(top_ten_countries['country'],

top_ten_countries['count'])

plt.xticks(rotation=45, ha='right')

plt.xlabel("Country")

plt.ylabel("Frequency")

plt.suptitle("Top 10 countries with most content on Netflix")

plt.show()
```



Conclusion: This analysis of Netflix content from 2008 to 2021 was successfully carried out using Python in Google Colab. It revealed that movies make up a larger portion of the content compared to TV shows. The United States leads in terms of content production on Netflix, and the majority of content is rated for general or teen audiences. Visualizations using Matplotlib and Seaborn provided clear insights into content type, rating distribution, and country-wise contributions. This project helped enhance my data analysis and visualization skills using Python-based tools within the Google Colab environment.