NETFLIX

Netflix Data: Data Cleaning, Analysis, Visualization Using Python & PowerBI

NETFLIX

Introduction

Netflix is a global streaming platform offering diverse content. It uses data analysis to understand user behavior, personalize recommendations, improve content decisions, and enhance user experience. This data-driven approach helps Netflix stay competitive, boost engagement, and deliver content that matches viewer preferences efficiently and effectively.

Objective

This project involves loading, cleaning, analyzing, and visualizing data from a Netflix dataset. We'll use Python libraries like Pandas, Matplotlib, and Seaborn to work through the project. The goal is to explore the dataset, derive insights, and prepare good visualization dashboard.

NETFLIX

DataSet Over View

Dataset Name:	netflix1	
Source:	Unified Mentor (Internship Project)	
Total Entries:	8790	
Total Columns:	10 columns	show_id, type, title, director, country, date_added, release_year, rating, duration, listed_in
Purpose:	Study and analyze content trends, distribution, and user preferences using python and powerBI.	

NETFLIX

Step 1: Data Preparation

- Importing libraries
- loading datase
- checking datasets

Import Required Libraries

```
[ ]  import pandas as pd
     import numpy as np
     import matplotlib.pyplot as plt
     import seaborn as sns
     from wordcloud import WordCloud
```

Step 2: Load the Dataset

- Load the dataset

```
[ ]  # Load the dataset
     data = pd.read_csv('/content/netflix1.csv', encoding='latin-1')
```

- Display the first few rows of the dataset

```
print(data.head())
```

```
  show_id     type                                  title        director  \
0      s1    Movie                   Dick Johnson Is Dead  Kirsten Johnson
1      s3  TV Show                              Ganglands  Julien Leclercq
2      s6  TV Show                          Midnight Mass    Mike Flanagan
3     s14    Movie  Confessions of an Invisible Girl    Bruno Garotti
4      s8    Movie                                Sankofa     Haile Gerima

         country date_added  release_year rating  duration  \
0  United States  9/25/2021          2020  PG-13    90 min
1         France  9/24/2021          2021  TV-MA  1 Season
2  United States  9/24/2021          2021  TV-MA  1 Season
3         Brazil  9/22/2021          2021  TV-PG    91 min
4  United States  9/24/2021          1993  TV-MA   125 min

                                           listed_in
0                                        Documentaries
1  Crime TV Shows, International TV Shows, TV Act...
2                TV Dramas, TV Horror, TV Mysteries
3                Children & Family Movies, Comedies
4   Dramas, Independent Movies, International Movies
```

NETFLIX

Step 3: Data Cleaning

- Check for missing values

```
# Check for missing values
print(data.isnull().sum())
```

```
show_id         0
type            0
title           0
director        0
country         0
date_added      0
release_year    0
rating          0
duration        0
listed_in       0
dtype: int64
```

NETFLIX

Step 3: Data Cleaning

- Drop duplicates if any

```
# Drop duplicates if any

data.drop_duplicates(inplace=True)
```

- Drop rows with missing critical information

```
data.dropna(subset=['director' , 'type', 'title', 'date_added',
                    'release_year', 'rating', 'duration',
                    'listed_in',  'country'], inplace=True)
print(data.isnull().sum())
```

```
show_id         0
type            0
title           0
director        0
country         0
date_added      0
release_year    0
rating          0
duration        0
listed_in       0
dtype: int64
```

- Converting data type: 'date_added' to datetime

```
# Convert 'date_added' to datetime
data['date_added'] = pd.to_datetime(data['date_added'])
print(data.head())
```

- Date_added data before conversion

	country	date_added	release_year	rating	duration	\
0	United States	9/25/2021	2020	PG-13	90 min	
1	France	9/24/2021	2021	TV-MA	1 Season	
2	United States	9/24/2021	2021	TV-MA	1 Season	
3	Brazil	9/22/2021	2021	TV-PG	91 min	
4	United States	9/24/2021	1993	TV-MA	125 min	

	country	date_added	release_year	rating	duration	\
0	United States	2021-09-25	2020	PG-13	90 min	
1	France	2021-09-24	2021	TV-MA	1 Season	
2	United States	2021-09-24	2021	TV-MA	1 Season	
3	Brazil	2021-09-22	2021	TV-PG	91 min	
4	United States	2021-09-24	1993	TV-MA	125 min	

NETFLIX

- Show data types to confirm changes

- Previous Datatype

```
# previous data type
print(data.dtypes)
```

```
show_id          object
type             object
title            object
director         object
country          object
date_added       object
release_year      int64
rating           object
duration         object
listed_in        object
dtype: object
```

- Datatype after changes

```
# Show data types to confirm changes
print(data.dtypes)
```

```
show_id                object
type                   object
title                  object
director               object
country                object
date_added     datetime64[ns]
release_year            int64
rating                 object
duration               object
listed_in              object
dtype: object
```

NETFLIX

Step 4: Exploratory Data Analysis (EDA)

- Content Type Distribution (Movies vs. TV Shows)

```
# Count the number of Movies and TV Shows

type_counts = data['type'].value_counts()
print(type_counts)
```

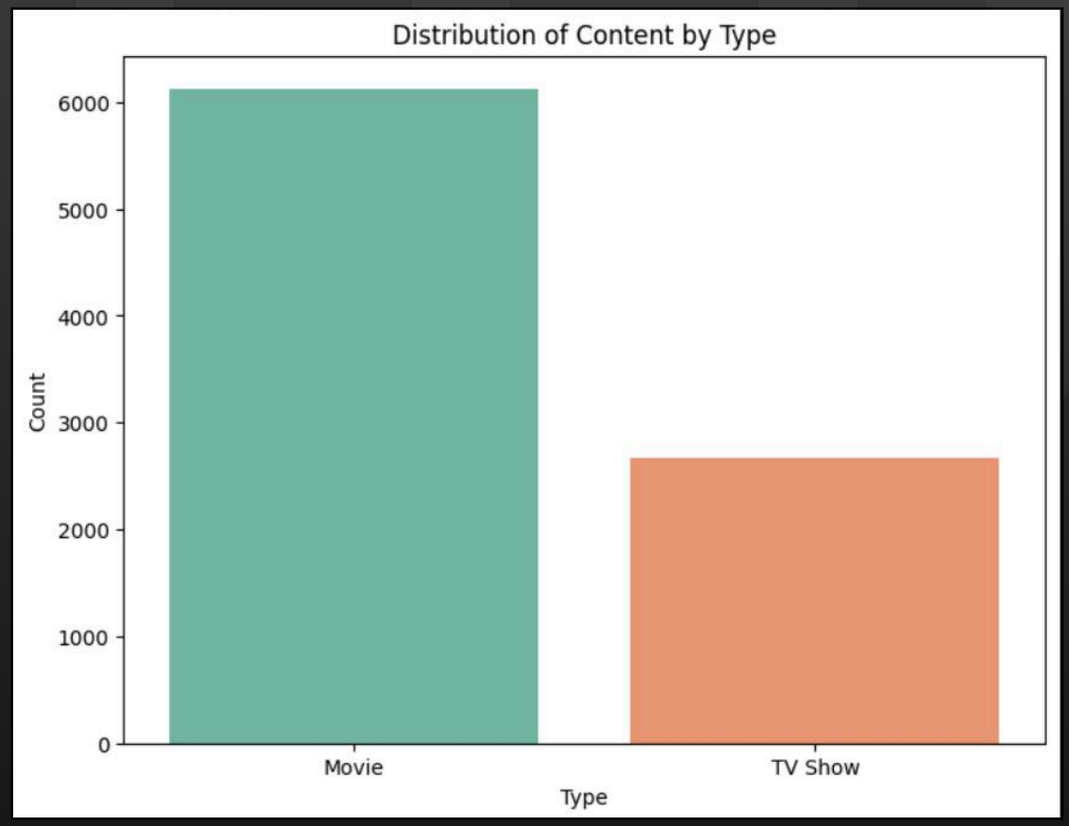
```
type
Movie      6126
TV Show    2664
Name: count, dtype: int64
```

- The code and its output shows the count of Movies and TV Shows available on Netflix.

NETFLIX

```
[ ]  # Plot the distribution
     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()
```

- This chart shows the count of Movies and TV Shows available on Netflix.
- Netflix has a higher number of movies compared to TV shows.



- Most Common Genres
  - Split the 'listed_in' column and count genres

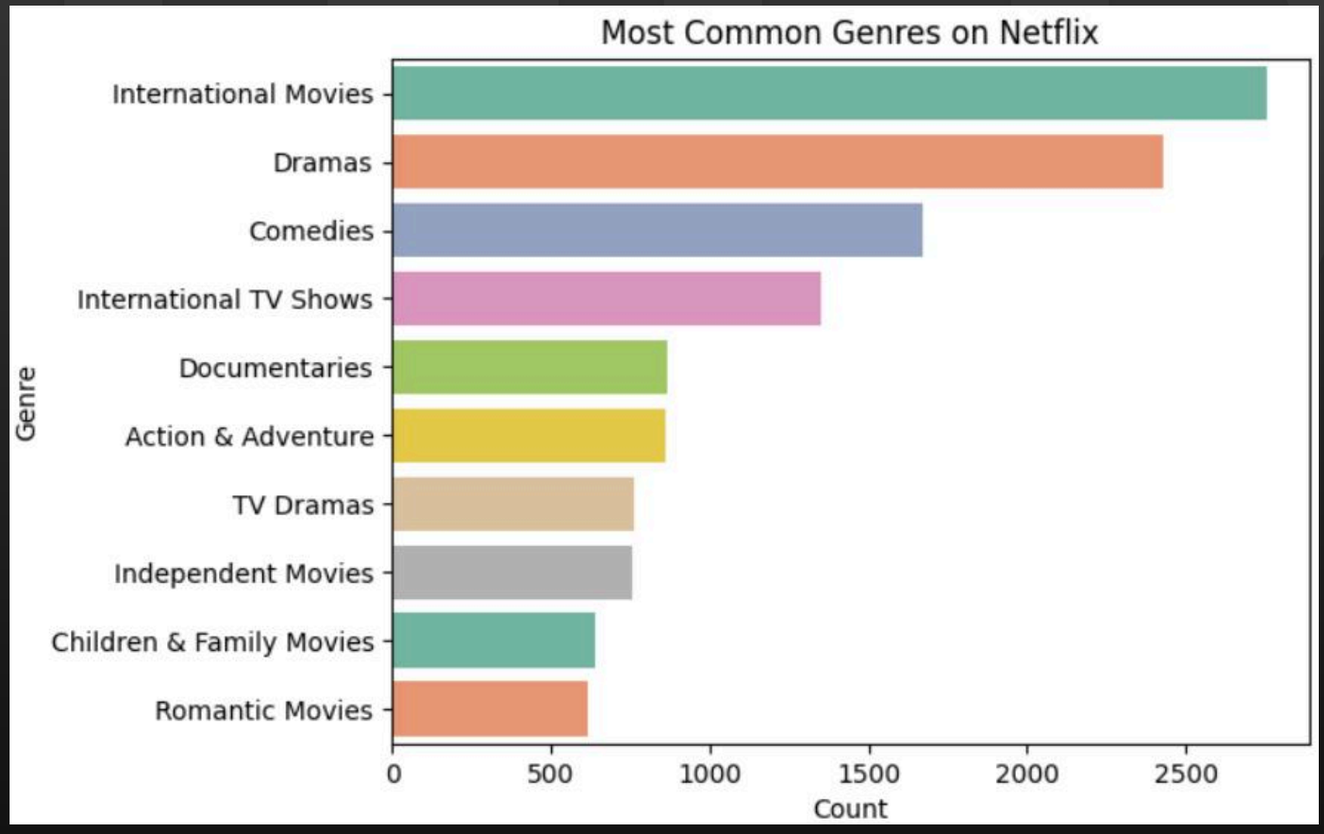
```python
# Split the 'listed_in' column and count genres
data['genres'] = data['listed_in'].apply(lambda x: x.split(', '))
all_genres = sum(data['genres'], [])
genre_counts = pd.Series(all_genres).value_counts().head(10)
print(genre_counts)
```

```
International Movies        2752
Dramas                      2426
Comedies                    1674
International TV Shows      1349
Documentaries                869
Action & Adventure           859
TV Dramas                    762
Independent Movies           756
Children & Family Movies     641
Romantic Movies              616
Name: count, dtype: int64
```

- The most common genres

```
# Plot the most common genres
# plt.figure(figsize=(10, 6))
sns.barplot(x=genre_counts.values, y=genre_counts.index,palette='Set2')
plt.title('Most Common Genres on Netflix')
plt.xlabel('Count')
plt.ylabel('Genre')
```

- The chart displays the top 10 most common genres of Netflix's contents.



NETFLIX

- Content Added Over Time
  - Extract year and month from 'date_added'

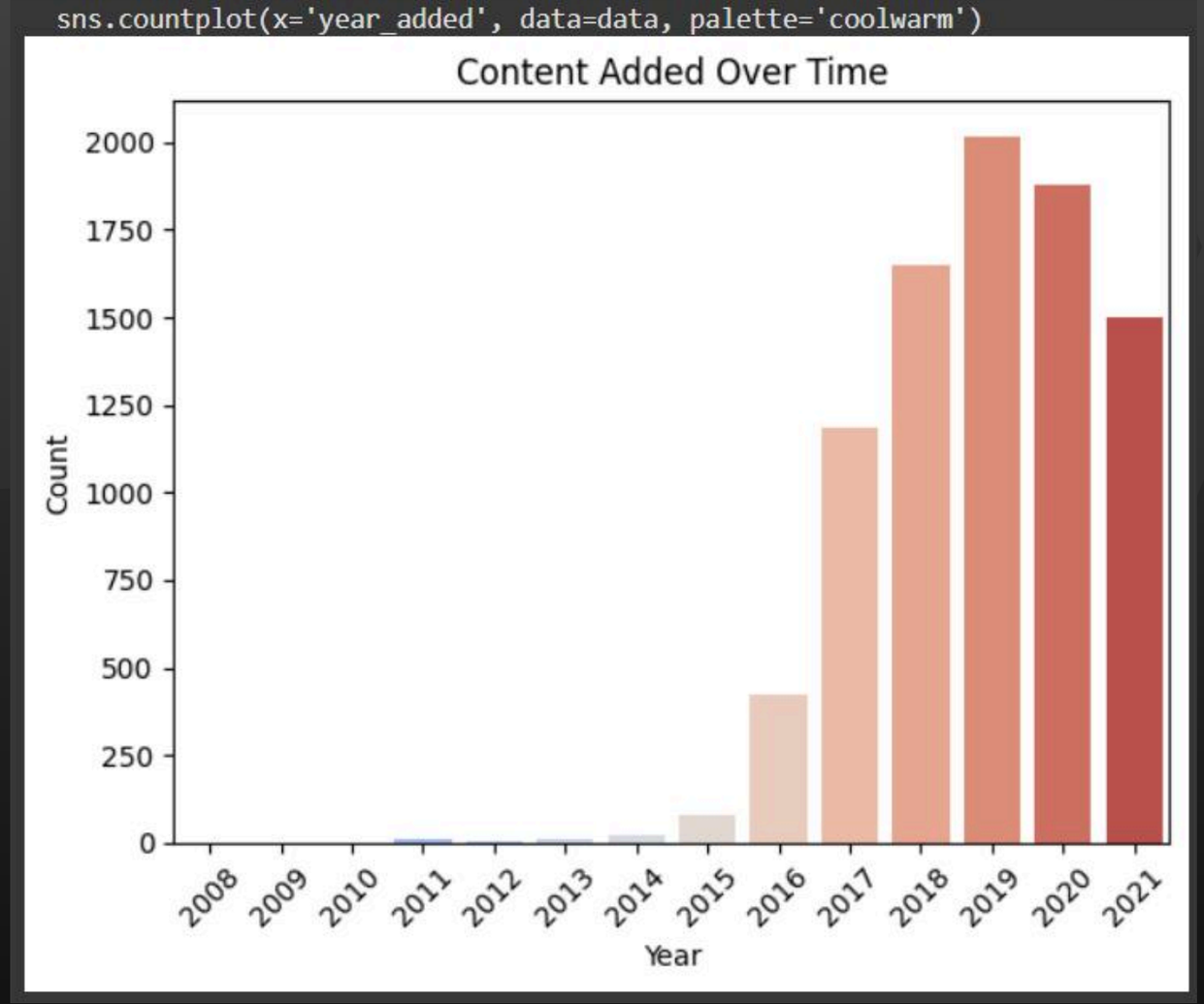
```
# Extract year and month from 'date_added'
data['year_added'] = data['date_added'].dt.year
data['month_added'] = data['date_added'].dt.month
print(data.head())
```

```
  show_id     type                          title        director  \
0      s1    Movie           Dick Johnson Is Dead  Kirsten Johnson
1      s3  TV Show                      Ganglands  Julien Leclercq
2      s6  TV Show                  Midnight Mass    Mike Flanagan
3     s14    Movie  Confessions of an Invisible Girl    Bruno Garotti
4      s8    Movie                        Sankofa     Haile Gerima

         country date_added  release_year rating  duration  \
0  United States 2021-09-25          2020  PG-13    90 min
1         France 2021-09-24          2021  TV-MA  1 Season
2  United States 2021-09-24          2021  TV-MA  1 Season
3         Brazil 2021-09-22          2021  TV-PG    91 min
4  United States 2021-09-24          1993  TV-MA   125 min

                                      listed_in  \
0                                  Documentaries
1  Crime TV Shows, International TV Shows, TV Act...
2                TV Dramas, TV Horror, TV Mysteries
3                Children & Family Movies, Comedies
4   Dramas, Independent Movies, International Movies

                                      genres  year_added  month_added
0                                [Documentaries]        2021            9
1  [Crime TV Shows, International TV Shows, TV Ac...        2021            9
2              [TV Dramas, TV Horror, TV Mysteries]        2021            9
3              [Children & Family Movies, Comedies]        2021            9
4  [Dramas, Independent Movies, International Mov...        2021            9
```

- Content added over the years

```
# Plot content added over the years
# plt.figure(figsize=(12, 6))
sns.countplot(x='year_added', data=data, palette='coolwarm')
plt.title('Content Added Over Time')
plt.xlabel('Year')
plt.ylabel('Count')
plt.xticks(rotation=45)
plt.show()
```

- This chart shows content added over years, year 2019 has most content added in a year.

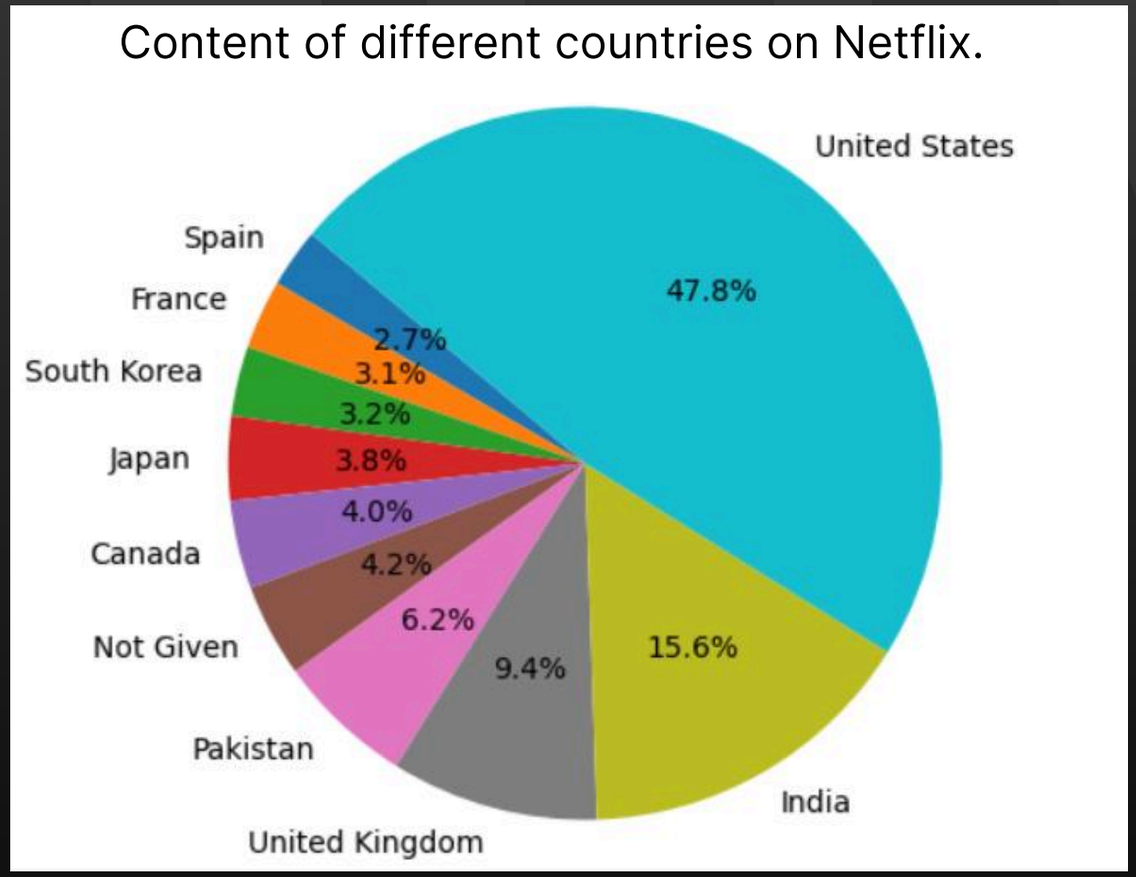


NETFLIX

- Netflix Content Distribution by Country

```
# content distribution by countries

country_counts = data['country'].value_counts().head(10)
country_counts.sort_values().plot(kind='pie', autopct='%1.1f%%', startangle=140)
plt.title('Most Common Countries on Netflix')
plt.ylabel('')
plt.axis('equal')
plt.show()
```

- The pie chart illustrates the proportion of content available of different countries on Netflix.



NETFLIX

Step 5: Creating PowerBI Dashboard

After performing initial data checks in Python, we used Power BI for advanced data transformation and visualization tailored to Netflix data. Power BI allowed us to:

- Clean and Transform Data – Handled issues like removing empty entries and standardizing formats.
- Build Interactive Dashboards – To explore Netflix content trends in a dynamic and user-friendly way.
- Visualize Key Insights – Utilized bar charts, pie charts, and line graphs to present data clearly and effectively.

Through Power BI, we transformed raw Netflix data into a visually rich, interactive dashboard that highlights insights on content distribution, genre popularity, and yearly content additions.

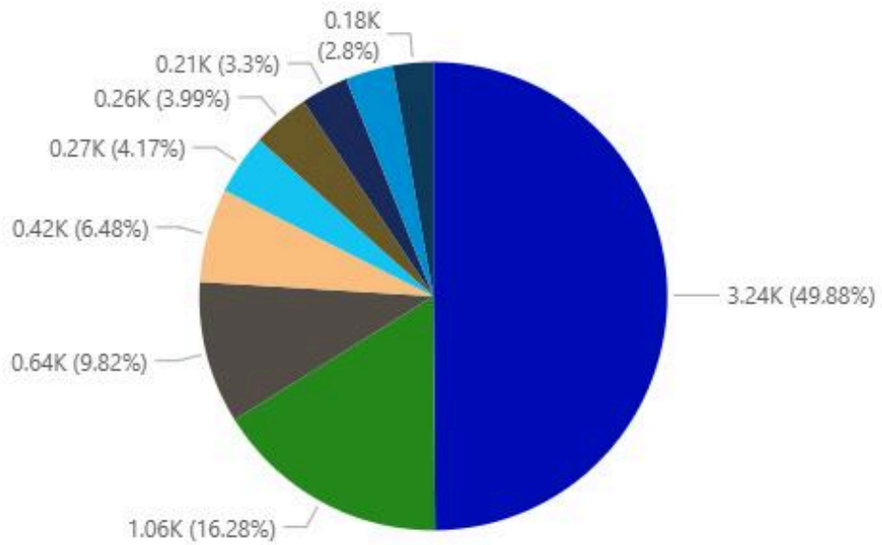
NETFLIX

PowerBI Dashboard

NETFLIX

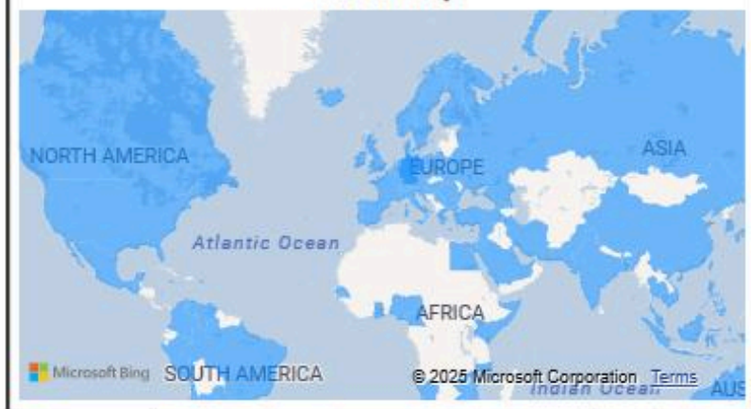
Netflix Data Analysis: Dashboard

Content by Countries



First country

World Map



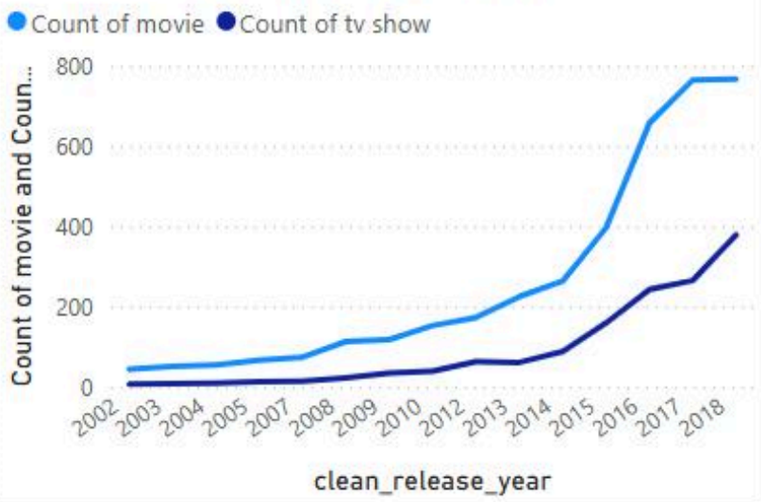
TV Shows

2.664K

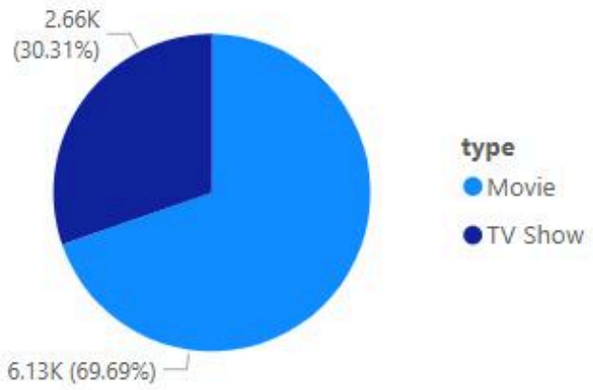
MOVIES

6.126K

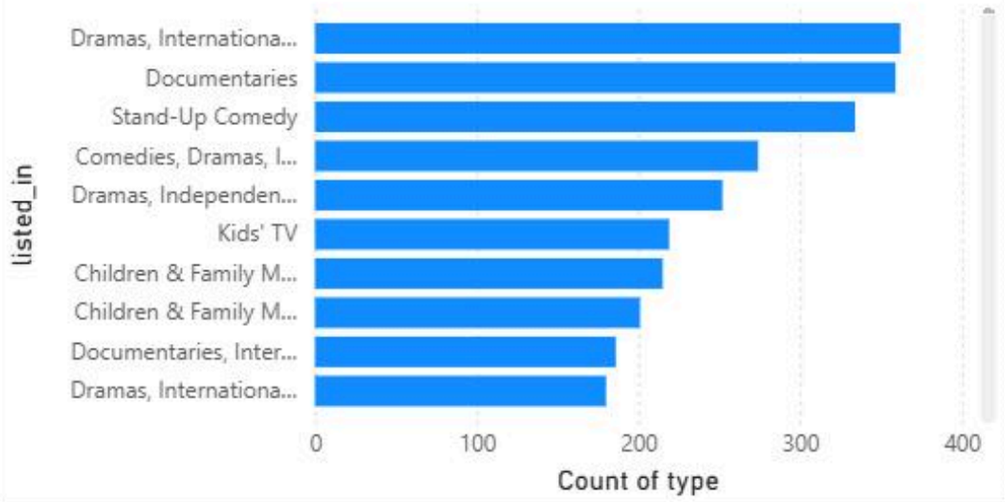
Content added over years



Movies and TV Shows



Top Genre



NETFLIX

Conclusion & Key Takeaways

- Clean Data: No missing values or duplicates found; fixed date format errors in python.
- Movies Dominate: Netflix has more movies than TV shows.
- Most content from the U.S followed by India, U.K
- Popular Genres: Drama and international.
- Power of Visuals: Python & Power BI helped uncover key trends.
- Clean data: Well-structured data ensures meaningful and accurate insights.

Thank You