Netflix Movies and TV Shows: Exploratory Data Analysis

Introduction This project explores a dataset of movies and TV shows available on Netflix. The dataset contains information such as title, director, cast, release year, country of origin, rating, duration, and the date the content was added to Netflix. The goal is to uncover insights about the type of content Netflix hosts, how it has evolved over time, and what patterns emerge in terms of genre, country, and release trends.

By cleaning and analyzing this data using Python libraries such as pandas, matplotlib, and seaborn, we aim to answer questions like:

How much content is added each year?

What's the balance between Movies and TV Shows?

Which countries and genres are most represented?

Dataset Source Kaggle.com

Libraries used | Library | Purpose | | ---------- | ---------- | | **pandas** | Data loading, cleaning, transformation, and analysis | | **numpy** | Numeric operations (used indirectly via pandas) | | **matplotlib** | Basic plotting and chart customization | | **seaborn** | High-level data visualization (e.g., bar plots, line plots) | | **wordcloud** | Generating word cloud visualizations from text data | | **datetime** *(optional)* | Working with dates (if not fully handled by pandas) |

DATASET LOADING, CLEANING AND ANALYSIS

In [175…
```
import pandas as pd
import datetime as dt
```

In [177…
```
netflix = pd.read_csv('netflix_titles.csv', index_col = 'show_id')
netflix.head()
```

Out[177…

show_id	type	title	director	cast	country	date_added	release_year	rating
s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13
s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA
s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021	2021	TV-MA
s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	September 24, 2021	2021	TV-MA
s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	September 24, 2021	2021	TV-MA

Understanding the Shape of the Dataset

In [179…

```
netflix.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 8807 entries, s1 to s8807
Data columns (total 11 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   type          8807 non-null   object
 1   title         8807 non-null   object
 2   director      6173 non-null   object
 3   cast          7982 non-null   object
 4   country       7976 non-null   object
 5   date_added    8797 non-null   object
 6   release_year  8807 non-null   int64 
 7   rating        8803 non-null   object
 8   duration      8804 non-null   object
 9   listed_in     8807 non-null   object
 10  description   8807 non-null   object
dtypes: int64(1), object(10)
memory usage: 825.7+ KB
```

In [181…
```
print('Shape of the dataset:',netflix.shape)
```

```
Shape of the dataset: (8807, 11)
```

In [183…
```
print('\nColumn names\n', netflix.columns)
```

```
Column names
 Index(['type', 'title', 'director', 'cast', 'country', 'date_added',
       'release_year', 'rating', 'duration', 'listed_in', 'description'],
      dtype='object')
```

In [185…
```
netflix.isnull().sum()
```

Out[185…
```
type              0
title             0
director       2634
cast            825
country         831
date_added       10
release_year      0
rating            4
duration          3
listed_in         0
description       0
dtype: int64
```

Cleaning Missing Values

Filling missing Values in 'Director' and 'Country' columns with 'Unknown'

In [342…
```
netflix['director'] = netflix['director'].fillna('Unknown')
```

In [344…
```
netflix['country'] = netflix['country'].fillna('Unknown')
```

Filling missing values in 'date_added' column with the values in the 'release year'

In [206…

```
#checking null values in date_added column
netflix[netflix['date_added'].isnull()]
```

Out[206...

show_id	type	title	director	cast	country	date_added	release_year	rating
s6080	TV Show	Abnormal Summit	Jung-ah Im, Seung-uk Jo	Hyun-moo Jun, Si-kyung Sung, Se-yoon Yoo	South Korea	NaT	2017	TV-PG
s6178	TV Show	忍者ハットリくん	Unknown	NaN	Japan	NaT	2012	TV-Y7
s6214	TV Show	Bad Education	Unknown	Jack Whitehall, Mathew Horne, Sarah Solemani, ...	United Kingdom	NaT	2014	TV-MA
s6280	TV Show	Being Mary Jane: The Series	Unknown	Gabrielle Union, Lisa Vidal, Margaret Avery, O...	United States	NaT	2016	TV-14
s6305	TV Show	Big Dreams, Small Spaces	Unknown	Monty Don	United Kingdom	NaT	2017	TV-G
...	...	...	...	...	...	...	...	...
s8540	TV Show	The Tudors	Unknown	Jonathan Rhys Meyers, Henry Cavill, James Frai...	Ireland, Canada, United States, United Kingdom	NaT	2010	TV-MA
s8558	TV Show	The West Wing	Unknown	Martin Sheen, Rob Lowe, Allison Janney, John S...	United States	NaT	2005	TV-14
s8685	TV Show	Vroomiz	Unknown	Joon-seok Song, Jeong-	South Korea	NaT	2016	TV-Y

show_id	type	title	director	cast	country	date_added	release_year	rating
				hwa Yang, Sang-hyun Um, ...				
s8713	TV Show	Weird Wonders of the World	Unknown	Chris Packham	United Kingdom	NaT	2016	TV-PG
s8756	TV Show	Women Behind Bars	Unknown	NaN	United States	NaT	2010	TV-14

88 rows × 11 columns

In [347…
```
netflix['date_added'] = netflix['date_added'].fillna(pd.to_datetime(netflix['releas
```

In [349…
```
netflix['date_added'] = pd.to_datetime(netflix['date_added'], errors = 'coerce')
```

In [351…
```
netflix.dtypes
```

Out[351…
```
type                    object
title                   object
director                object
cast                    object
country                 object
date_added      datetime64[ns]
release_year             int64
rating                  object
listed_in               object
description             object
duration_int           float64
durtion_type            object
year_added               int32
dtype: object
```

Cleaning the 'duration' column

In [222…
```
#splitting 'duration' into two new columns
netflix[['duration_int','durtion_type']] = netflix['duration'].str.split(' ', expan
```

In [238…
```
netflix['duration_int'] = pd.to_numeric(netflix['duration_int'])
```

In [240…
```
netflix.dtypes
```

Out[240…

```
type                    object
title                   object
director                object
cast                    object
country                 object
date_added      datetime64[ns]
release_year             int64
rating                  object
duration               float64
listed_in               object
description             object
duration_int           float64
durtion_type            object
dtype: object
```

In [250…

```
netflix.drop(columns='duration', axis=1, inplace=True)
netflix
```

Out[250…

show_id	type	title	director	cast	country	date_added	release_year	rati
s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	2021-09-25	2020	PG
s2	TV Show	Blood & Water	Unknown	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	
s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	Unknown	2021-09-24	2021	
s4	TV Show	Jailbirds New Orleans	Unknown	NaN	Unknown	2021-09-24	2021	
s5	TV Show	Kota Factory	Unknown	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	2021-09-24	2021	
...	...	...	...	...	...	...	...	
s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	2019-11-20	2007	
s8804	TV Show	Zombie Dumb	Unknown	NaN	Unknown	2019-07-01	2018	TV

show_id	type	title	director	cast	country	date_added	release_year	rati
s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	2019-11-01	2009	
s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	2020-01-11	2006	
s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	2019-03-02	2015	TV

8807 rows × 12 columns

VISUALIZATION

Number of Movies vs TV shows

In [253…
```
import matplotlib.pyplot as plt
import seaborn as sns
```

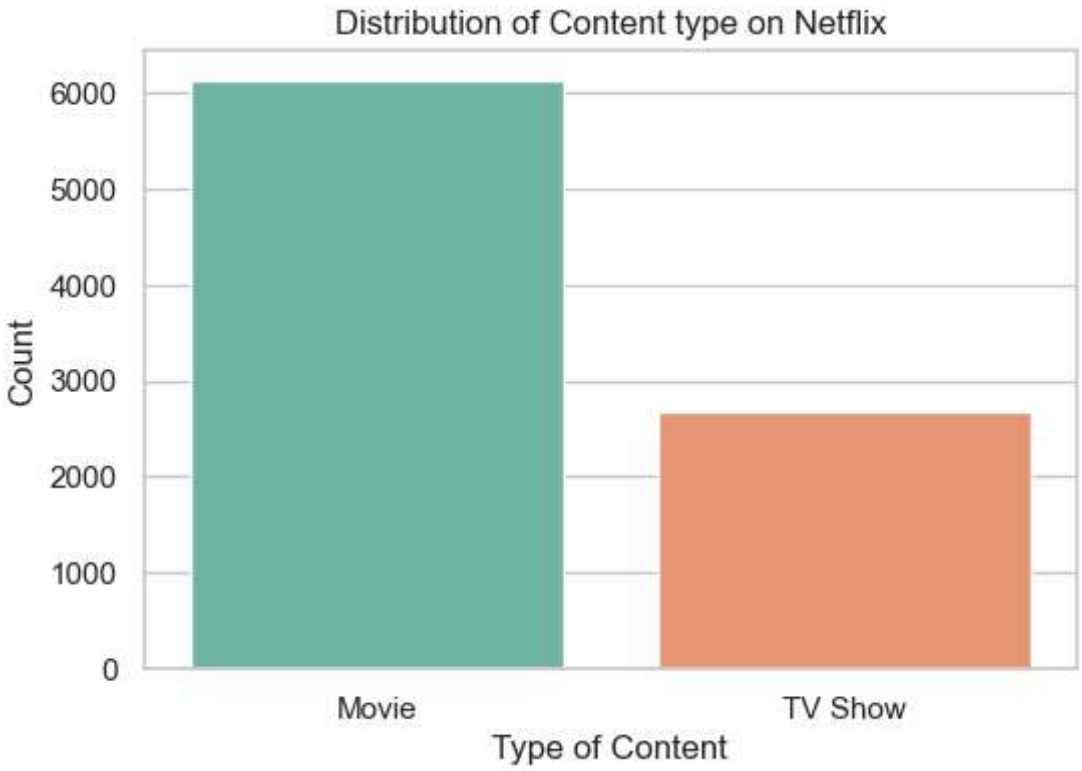
In [255…
```
sns.set(style = 'whitegrid')
```

In [259…
```
plt.figure(figsize = (6, 4))
sns.countplot(data=netflix, x='type', palette = 'Set2')
plt.title('Distribution of Content type on Netflix')
plt.xlabel('Type of Content')
plt.ylabel('Count')
plt.show()
```

```
C:\Users\saras\AppData\Local\Temp\ipykernel_2480\3304694227.py:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.1
4.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.countplot(data=netflix, x='type', palette = 'Set2')
```

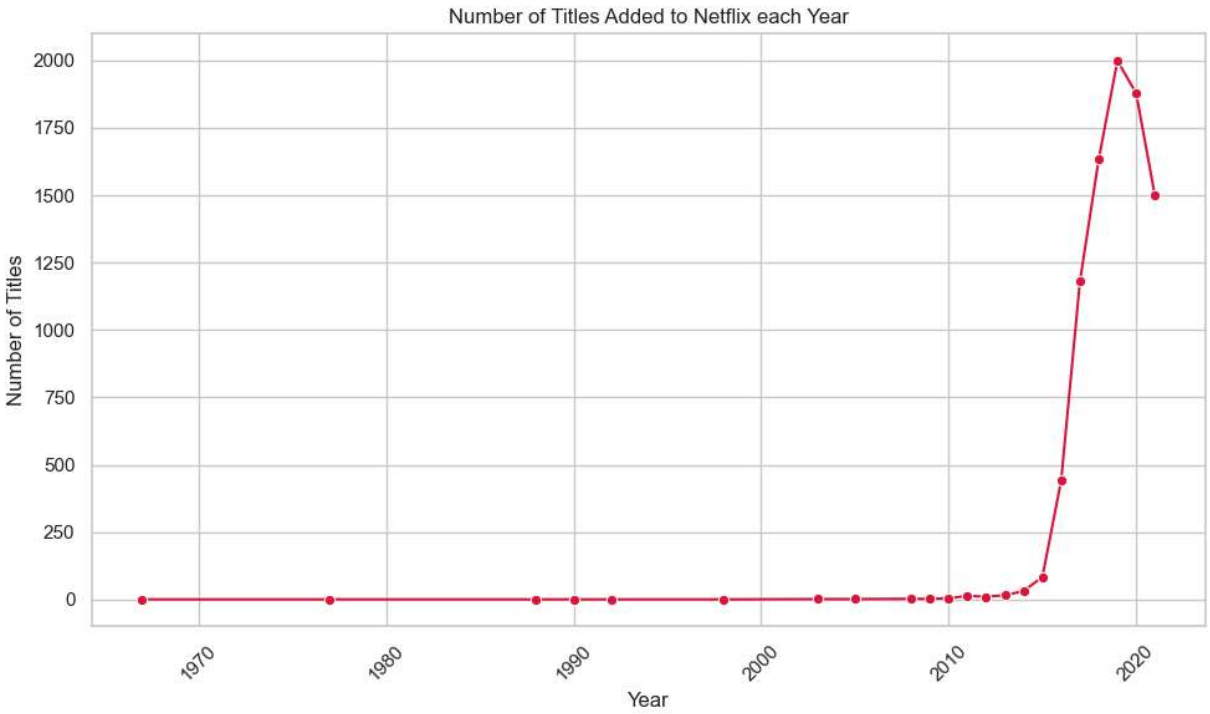


Titles added per year

In [262…
```
#Extracting the year from the 'date_added' column
netflix['year_added'] = netflix['date_added'].dt.year
```

In [268…
```
#counting titles per year
titles_per_year = netflix['year_added'].value_counts().sort_index()
```

In [270…
```
plt.figure(figsize = (10,6))
sns.lineplot(x=titles_per_year.index, y=titles_per_year.values, marker='o', color='
plt.title('Number of Titles Added to Netflix each Year')
plt.xlabel('Year')
plt.ylabel('Number of Titles')
plt.xticks(rotation = 45)
plt.grid(True)
plt.tight_layout()
plt.show()
```

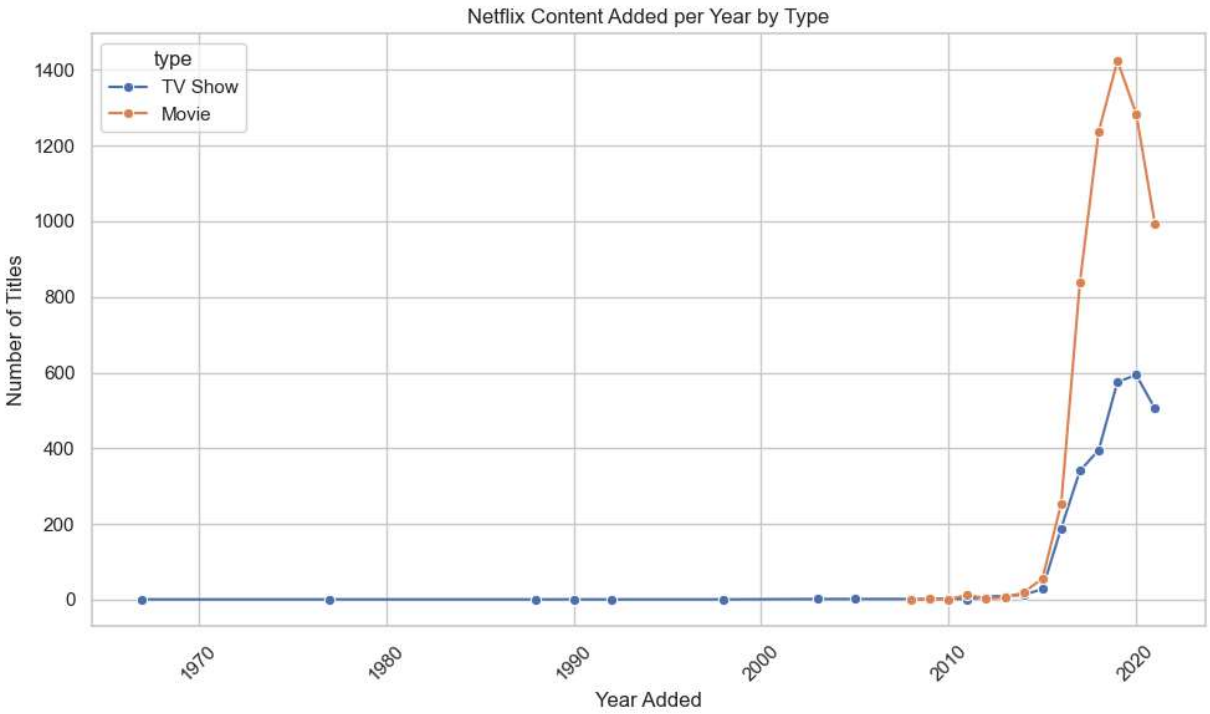


The chart shows a sudden spike during 2020 most likely due to COVID-19 pandemic and the curfew

Content Type added per Year

In [280…
```
year_type_counts = netflix.groupby(['year_added','type']).size().reset_index(name='
```

In [282…
```
plt.figure(figsize=(10,6))
sns.lineplot(data=year_type_counts, x='year_added', y='count', hue='type', marker='
plt.title('Netflix Content Added per Year by Type')
plt.xlabel('Year Added')
plt.ylabel('Number of Titles')
plt.xticks(rotation=45)
plt.grid(True)
plt.tight_layout()
plt.show()
```



During the pandemic Netflix focus was mainly on adding movies rather than TV shows

Top Countries on Netflix

In [287…
```
#dropping 'Unknown' values
country_data = netflix[netflix['country'] != 'Unknown'].copy()
```

In [289…
```
#splitting values where two or more countries are in the cell
country_data ['country'] = country_data['country'].str.split(', ')
country_exploded = country_data.explode('country')
```

In [299…
```
top_countries = country_exploded['country'].value_counts().head(10)
top_countries
```

Out[299…
```
country
United States     3689
India             1046
United Kingdom     804
Canada             445
France             393
Japan              318
Spain              232
South Korea        231
Germany            226
Mexico             169
Name: count, dtype: int64
```

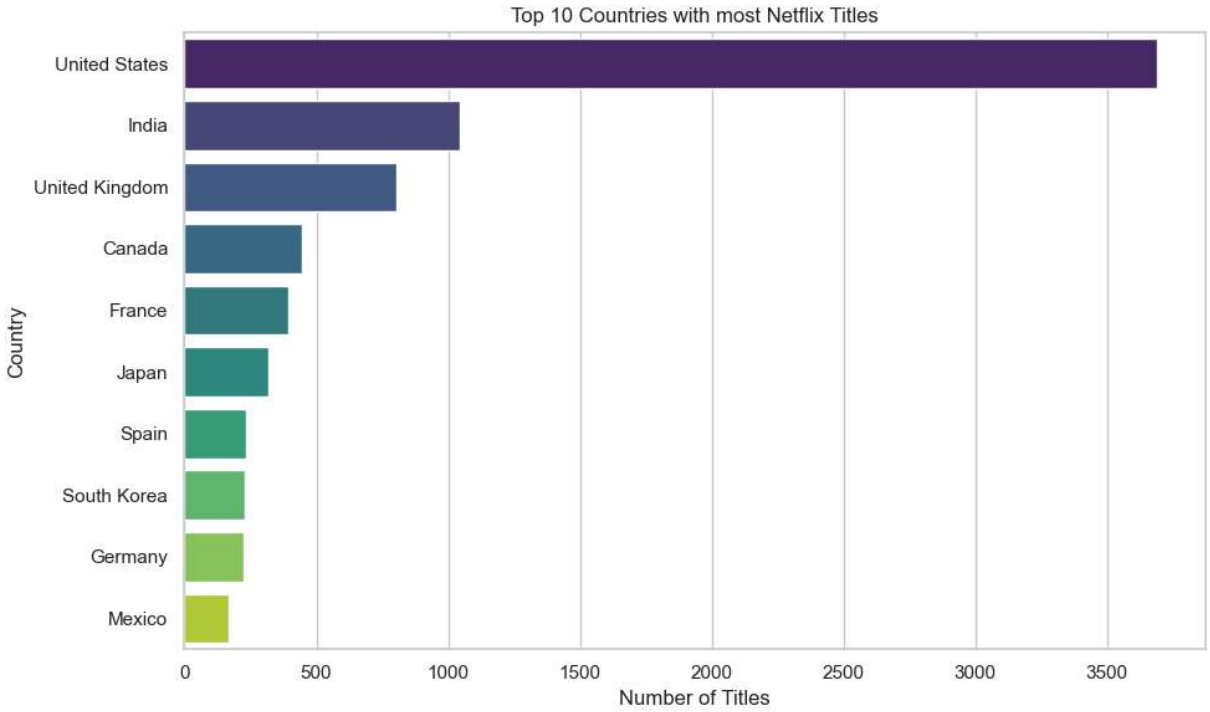
In [301…
```
plt.figure(figsize=(10,6))
sns.barplot(x=top_countries.values, y=top_countries.index, palette = 'viridis')
plt.title('Top 10 Countries with most Netflix Titles')
plt.xlabel('Number of Titles')
plt.ylabel('Country')
```

```
plt.tight_layout()
plt.show()
```

```
C:\Users\saras\AppData\Local\Temp\ipykernel_2480\3939814251.py:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.1
4.0. Assign the `y` variable to `hue` and set `legend=False` for the same effect.

  sns.barplot(x=top_countries.values, y=top_countries.index, palette = 'viridis')
```



Most Common Genres on Netflix

In [306…
```
#splitting genres
genre_data = netflix.copy()
genre_data['genre'] = genre_data['listed_in'].str.split(', ')
genre_exploded = genre_data.explode('genre')
```

In [310…
```
top_genres = genre_exploded['genre'].value_counts().head(10)
top_genres
```

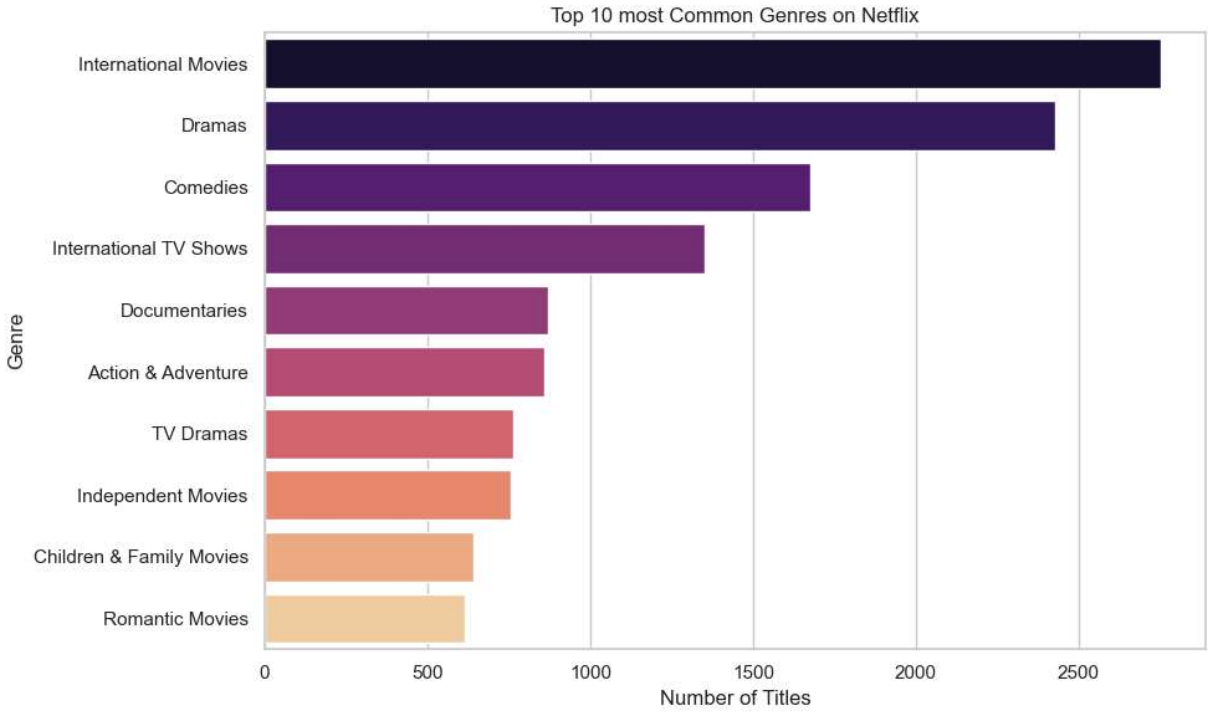
Out[310…
```
genre
International Movies        2752
Dramas                      2427
Comedies                    1674
International TV Shows      1351
Documentaries                869
Action & Adventure           859
TV Dramas                    763
Independent Movies           756
Children & Family Movies     641
Romantic Movies              616
Name: count, dtype: int64
```

In [314…
```
plt.figure(figsize = (10,6))
sns.barplot(x=top_genres.values, y=top_genres.index, palette = 'magma')
plt.title('Top 10 most Common Genres on Netflix')
plt.xlabel('Number of Titles')
plt.ylabel('Genre')
plt.tight_layout()
plt.show()
```

```
C:\Users\saras\AppData\Local\Temp\ipykernel_2480\1261391661.py:2: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.1
4.0. Assign the `y` variable to `hue` and set `legend=False` for the same effect.

  sns.barplot(x=top_genres.values, y=top_genres.index, palette = 'magma')
```



This shows that Netflix prioritizes more 'International Movies', 'Dramas' and 'Comedies' while the least of generes presented on Netflix is 'Romantic Movies'

Text Analysis on Titles and Descriptions

In [318…
```
pip install wordcloud
```

```
Collecting wordcloud
  Downloading wordcloud-1.9.4-cp312-cp312-win_amd64.whl.metadata (3.5 kB)
Requirement already satisfied: numpy>=1.6.1 in c:\users\saras\anaconda3\lib\site-pac
kages (from wordcloud) (1.26.4)
Requirement already satisfied: pillow in c:\users\saras\anaconda3\lib\site-packages
(from wordcloud) (10.4.0)
Requirement already satisfied: matplotlib in c:\users\saras\anaconda3\lib\site-packa
ges (from wordcloud) (3.9.2)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\saras\anaconda3\lib\site
-packages (from matplotlib->wordcloud) (1.2.0)
Requirement already satisfied: cycler>=0.10 in c:\users\saras\anaconda3\lib\site-pac
kages (from matplotlib->wordcloud) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\saras\anaconda3\lib\sit
e-packages (from matplotlib->wordcloud) (4.51.0)
Requirement already satisfied: kiwisolver>=1.3.1 in c:\users\saras\anaconda3\lib\sit
e-packages (from matplotlib->wordcloud) (1.4.4)
Requirement already satisfied: packaging>=20.0 in c:\users\saras\anaconda3\lib\site-
packages (from matplotlib->wordcloud) (24.1)
Requirement already satisfied: pyparsing>=2.3.1 in c:\users\saras\anaconda3\lib\site
-packages (from matplotlib->wordcloud) (3.1.2)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\saras\anaconda3\lib
\site-packages (from matplotlib->wordcloud) (2.9.0.post0)
Requirement already satisfied: six>=1.5 in c:\users\saras\anaconda3\lib\site-package
s (from python-dateutil>=2.7->matplotlib->wordcloud) (1.16.0)
Downloading wordcloud-1.9.4-cp312-cp312-win_amd64.whl (301 kB)
Installing collected packages: wordcloud
Successfully installed wordcloud-1.9.4
Note: you may need to restart the kernel to use updated packages.
```

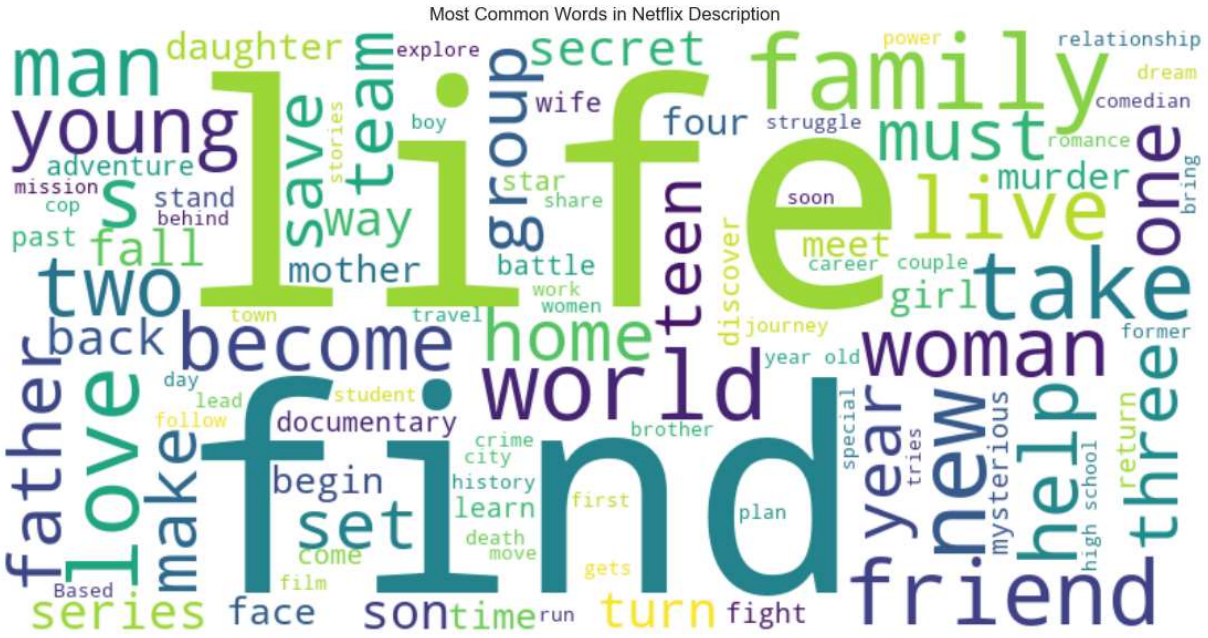
In [320…
```
from wordcloud import WordCloud, STOPWORDS
import matplotlib.pyplot as plt
```

In [322…
```
text = ' '.join(netflix['description'].dropna().tolist())
```

In [324…
```
stopwords = set(STOPWORDS)
```

In [330…
```
wordcloud = WordCloud(stopwords = stopwords,
                      background_color='white',
                      max_words=100,
                      width=800,
                      height=400).generate(text)
```

In [331…
```
plt.figure(figsize=(12,6))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.title('Most Common Words in Netflix Description')
plt.tight_layout()
plt.show()
```



Final Look of the Dataset

In [366…

```
netflix.index = netflix.index.str.title()
netflix.columns = netflix.columns.str.title()
netflix
```

Out[366...

show_id	Type	Title	Director	Cast	Country	Date_Added	Release_Year	Ra
S1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	2021-09-25	2020	P
S2	TV Show	Blood & Water	Unknown	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	
S3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	Unknown	2021-09-24	2021	
S4	TV Show	Jailbirds New Orleans	Unknown	NaN	Unknown	2021-09-24	2021	
S5	TV Show	Kota Factory	Unknown	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	2021-09-24	2021	
...	...	...	...	...	...	...	...	
S8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	2019-11-20	2007	
S8804	TV Show	Zombie Dumb	Unknown	NaN	Unknown	2019-07-01	2018	T
S8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg,	United States	2019-11-01	2009	

show_id	Type	Title	Director	Cast	Country	Date_Added	Release_Year	Ra
				Woody Harrelson, Emma Stone, ...				
S8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	2020-01-11	2006	
S8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan...	India	2019-03-02	2015	T

8807 rows × 13 columns

Insights

Content Growth: Netflix has steadily increased the number of titles added each year, showing consistent content expansion.

Movies vs TV Shows: Movies dominate the platform, but TV shows have seen a steady rise, reflecting the shift toward binge-worthy content.

Top Countries: The United States leads in content production, followed by India, the United Kingdom, and other global contributors.

Popular Genres: Dramas and Comedies are the most common genres, followed by Documentaries and International content.

Frequent Themes: From the text descriptions, words like life, love, family, and murder frequently appear, revealing common narrative elements.

Diverse Content: Netflix features content from a wide range of countries and genres, appealing to a global audience.

Conclusion

This project showcases how Python can be used for real-world data analysis tasks in media and entertainment. Using pandas, seaborn, and wordcloud, we extracted meaningful insights from Netflix's catalog, such as content trends, popular genres, and regional diversity.

Such analysis is not only useful for understanding platform strategy but can also support personalized recommendations, market research, and content planning.

In a professional context, this kind of exploration demonstrates a data analyst's ability to clean, visualize, and interpret large datasets.

Credits and Extras

Dataset: Provided by Kaggle – Netflix Titles Dataset

Tools Used: Python, Jupyter Notebook, pandas, matplotlib, seaborn, wordcloud

Author: (Your Name Here – feel free to personalize)

💼 This notebook is part of my data analysis portfolio. Feel free to explore more of my work on [GitHub/Portfolio link].

In [ ]: