

✓ Welcome to Colab!

Explore the Gemini API

The Gemini API gives you access to Gemini models created by Google DeepMind. Gemini models are built from the ground up to be multimodal, so you can reason seamlessly across text, images, code, and audio.

How to get started?

- Go to [Google AI Studio](#) and log in with your Google account.
- [Create an API key](#).
- Use a quickstart for [Python](#), or call the REST API using [curl](#).

Discover Gemini's advanced capabilities

- Play with Gemini [multimodal outputs](#), mixing text and images in an iterative way.
- Discover the [multimodal Live API](#) (demo [here](#)).
- Learn how to [analyze images and detect items in your pictures](#) using Gemini (bonus, there's a [3D version](#) as well!).
- Unlock the power of [Gemini thinking model](#), capable of solving complex task with its inner thoughts.

Explore complex use cases

- Use [Gemini grounding capabilities](#) to create a report on a company based on what the model can find on internet.
- Extract [invoices and form data from PDF](#) in a structured way.
- Create [illustrations based on a whole book](#) using Gemini large context window and Imagen.

To learn more, check out the [Gemini cookbook](#) or visit the [Gemini API documentation](#).

Colab now has AI features powered by [Gemini](#). The video below provides information on how to use these features, whether you're new to Python, or a seasoned veteran.

ValueError

KeyError

...



Please explain [this code](#).



Explain Code

The code snippet you provided is processing the `date_added` column of the DataFrame `df`. It aims to split the dates into individual components.



```
z=df[ 'date_added' ].apply
```



[Use code with caution](#)

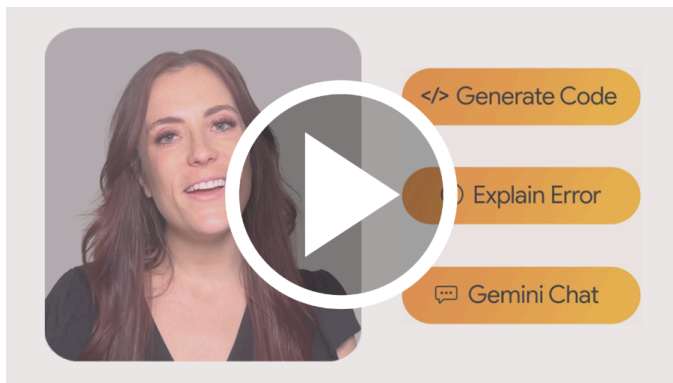
This line of code selects the `date_added` column from the `df` DataFrame. It then applies a function to each value in this column using the `.apply()` method. The function being applied is a `lambda` function, which is a small anonymous function. This `lambda` function takes an input `x` (which represents each value in the `date_added` column), converts it to a string using `str(x)`, and then splits the string by spaces using `.split(' ')`. This split operation will likely separate the month, day, and year components of the date. Finally, `.to_list()` converts the result of applying this function to the entire column into a Python list, which is then assigned to the variable `z`.



```
z=pd.DataFrame(z)
```

[Use code with caution](#)

This line takes the list `z` (which contains lists of the date components) and converts it into a



What is Colab?

Colab, or "Colaboratory", allows you to write and execute Python in your browser, with

- Zero configuration required
- Access to GPUs free of charge
- Easy sharing

Whether you're a **student**, a **data scientist** or an **AI researcher**, Colab can make your work easier. Watch [Introduction to Colab](#) or [Colab Features You May Have Missed](#) to learn more, or just get started below!

✓ Getting started

The document you are reading is not a static web page, but an interactive environment called a **Colab notebook** that lets you write and execute code.

For example, here is a **code cell** with a short Python script that computes a value, stores it in a variable, and prints the result:

```
seconds_in_a_day = 24 * 60 * 60
seconds_in_a_day
```

↩ 86400

To execute the code in the above cell, select it with a click and then either press the play button to the left of the code, or use the keyboard shortcut "Command/Ctrl+Enter". To edit the code, just click the cell and start editing.

pandas DataFrame. Each inner list from the original list becomes a row in the new DataFrame.

```
z=z.stack().reset_index()
```

◀ ————— ▶
[Use code with caution](#)

This line performs two operations. First, `z.stack()` "stacks" or "melts" the DataFrame. It transforms the DataFrame from a wide format (multiple columns for date components) into a long format, where all the date components are in a single column. The original column names become part of a new index level. Second, `.reset_index()` converts the index of the stacked DataFrame back into regular columns. This typically results in a DataFrame with columns for the original index (which was the row index from the previous step), the original column name, and the unstacked value (the date component).

z

[Use code with caution](#)

This line simply displays the contents of the `z` DataFrame in the output of the notebook cell. This allows you to see the result of the unstacking operation, where each individual date component from the `date_added` column is now on its own row, along with information about its origin.

Rate this answer 👍 👎



WHAT BUSINESS INSIGHTS CAN I SAY FROM MY DATA

Variables that you define in one cell can later be used in other cells:

```
seconds_in_a_week = 7 * seconds_in_a_day  
seconds_in_a_week
```

→ 604800

Colab notebooks allow you to combine **executable code** and **rich text** in a single document, along with **images**, **HTML**, **LaTeX** and more. When you create your own Colab notebooks, they are stored in your Google Drive account. You can easily share your Colab notebooks with co-workers or friends, allowing them to comment on your notebooks or even edit them. To learn more, see [Overview of Colab](#). To create a new Colab notebook you can use the File menu above, or use the following link: [create a new Colab notebook](#).

Colab notebooks are Jupyter notebooks that are hosted by Colab. To learn more about the Jupyter project, see jupyter.org.

✓ Data science

With Colab you can harness the full power of popular Python libraries to analyze and visualize data. The code cell below uses **numpy** to generate some random data, and uses **matplotlib** to visualize it. To edit the code, just click the cell and start editing.

You can import your own data into Colab notebooks from your Google Drive account, including from spreadsheets, as well as from Github and many other sources. To learn more about importing data, and how Colab can be used for data science, see the links below under [Working with Data](#).

```
import numpy as np  
import IPython.display as display  
from matplotlib import pyplot as plt  
import io
```

◆ Business Insights from the Netflix Dataset

Based on the analysis performed in the notebook, here are some potential business insights that Netflix could gain from this dataset:

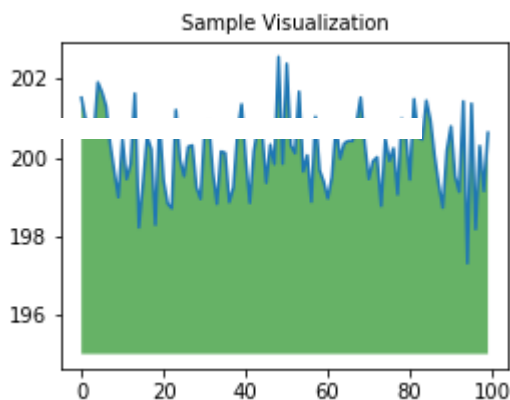
- **Content Mix:** The analysis shows that the number of movies is more than double the number of TV shows. This indicates that movies are a significantly larger part of Netflix's content library. Netflix should consider if this aligns with their overall strategy and subscriber preferences. While movies are popular, investing more in TV shows, particularly those that have multiple seasons, could potentially increase subscriber engagement and retention. The analysis also highlighted that "one season" shows are the most popular type of TV show. This could suggest that investing in limited series or shorter-run shows is a viable strategy.
- **Geographic Focus:** The top 5 countries where Netflix is most popular are the United States, India, the United Kingdom, Japan, and South Korea. Focusing marketing efforts, content acquisition, and even local content production in these countries could be a key driver for subscriber growth

```
import base64

ys = 200 + np.random.randn(100)
x = [x for x in range(len(ys))]

fig = plt.figure(figsize=(4, 3), facecolor='w')
plt.plot(x, ys, '-')
plt.fill_between(x, ys, 195, where=(ys > 195),
plt.title("Sample Visualization", fontsize=10)

data = io.BytesIO()
plt.savefig(data)
image = F"data:image/png;base64,{base64.b64encod
alt = "Sample Visualization"
display.display(display.Markdown(F"""![{alt}]({:
plt.close(fig)
```



Colab notebooks execute code on Google's cloud servers, meaning you can leverage the power of Google hardware, including [GPUs and TPUs](#), regardless of the power of your machine. All you need is a browser.

For example, if you find yourself waiting for **pandas** code to finish running and want to go faster, you can switch to a GPU Runtime and use libraries like [RAPIDS cuDF](#) that provide zero-code-change acceleration.

To learn more about accelerating pandas on Colab, see the [10 minute guide](#) or [US stock market data analysis demo](#).

✓ Machine learning

and engagement. Since Netflix is already popular in these regions, it might be easier to acquire new customers compared to less penetrated markets.

- **Popular Genres:** Identifying the top 10 popular genres is crucial for content strategy. Netflix should prioritize acquiring or producing content within these popular genres to cater to existing subscriber preferences and attract new ones.

Understanding which genres perform best can help in making informed decisions about future content investments.

- **Release Year Trends:** The histogram of release years shows a sharp increase in the number of movies released after 2000. This suggests a trend towards more recent content being added to the platform. While this is expected, analyzing the performance of older content versus newer content could provide insights into subscriber viewing habits and content longevity.

- **Content Addition Timing:** The analysis of the month of content addition shows that October has the highest number of movies added. While this might be tied to specific release schedules or holidays, it raises the question of whether content additions are optimized

With Colab you can import an image dataset, train an image classifier on it, and evaluate the model, all in just [a few lines of code](#).

Colab is used extensively in the machine learning community with applications including:

- Getting started with TensorFlow
- Developing and training neural networks
- Experimenting with TPUs
- Disseminating AI research
- Creating tutorials

To see sample Colab notebooks that demonstrate machine learning applications, see the [machine learning examples](#) below.

✓ More Resources

Working with Notebooks in Colab

- [Overview of Colab](#)
- [Guide to Markdown](#)
- [Importing libraries and installing dependencies](#)
- [Saving and loading notebooks in GitHub](#)
- [Interactive forms](#)
- [Interactive widgets](#)

Working with Data

- [Loading data: Drive, Sheets, and Google Cloud Storage](#)
- [Charts: visualizing data](#)
- [Getting started with BigQuery](#)

Machine Learning

These are a few of the notebooks related to Machine Learning, including Google's online Machine Learning course. See the [full course website](#) for more.

- [Intro to Pandas DataFrame](#)
- [Intro to RAPIDS cuDF to accelerate pandas](#)
- [Getting Started with cuML's accelerator mode](#)
- [Linear regression with tf.keras using synthetic data](#)

throughout the year. Adding content uniformly throughout the year, or strategically aligning additions with festivals and cultural events in different countries, could potentially lead to more consistent engagement.

- **Director and Cast**

Combinations: The analysis of popular director and cast combinations highlights successful creative partnerships. Identifying these popular combinations can inform decisions about which directors and actors to work with on future projects. Leveraging the established popularity of these individuals could increase the likelihood of a show or movie's success. Specifically, focusing on the director-actor combinations that appear together frequently (more than 10 times in the analyzed data) could be a targeted approach.

- **Data Quality:** The high number of unique entries in the `country` column (748 compared to around 200 countries globally) suggests potential data quality issues. Addressing this inaccuracy is important for accurate analysis of geographic trends and targeting. This could involve standardizing country names, identifying combined entries (country and state), or investigating potential data entry errors.

Using Accelerated Hardware

- [TensorFlow with GPUs](#)
- [TPUs in Colab](#)

✓ Featured examples

- [Retraining an Image Classifier](#): Build a Keras model on top of a pre-trained image classifier to distinguish flowers.
- [Text Classification](#): Classify IMDB movie reviews as either *positive* or *negative*.
- [Style Transfer](#): Use deep learning to transfer style between images.
- [Multilingual Universal Sentence Encoder Q&A](#): Use a machine learning model to answer questions from the SQuAD dataset.
- [Video Interpolation](#): Predict what happened in a video between the first and the last frame.

Netflix is one of the most popular media and video streaming platforms. They have over 10000 movies or tv shows available on their platform, as of mid-2021, they have over 222M Subscribers globally. This tabular dataset consists of listings of all the movies and tv shows available on Netflix, along with details such as - cast, directors, ratings, release year, duration, etc.

Business Problem:


Analyze the data and generate insights that could help Netflix in deciding which type of shows/movies to produce and how they can grow the business in different countries

Below is the top 10 rows of the data.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
df=pd.read_csv('/content/netflix.csv')
```

These insights can be used to inform Netflix's content production, acquisition, marketing, and overall business strategy to potentially increase subscriber numbers and revenue. Remember, these are insights derived from the provided dataset and further analysis or external data could provide a more complete picture.

► Sources

Rate this answer  



WHAT RECOMMENDATIONS SHOULD I GIVE TO NETFLIX



Recommendations for Netflix

Based on the business insights derived from the analysis of the dataset, here are some recommendations Netflix could consider:

- **Diversify TV Show Investment:** While movies are currently the dominant content type, consider strategically increasing investment in TV shows, particularly those with multiple seasons. This could lead to sustained subscriber engagement and reduce churn, as viewers tend to stay subscribed to follow ongoing series. The popularity of "one season" shows also suggests exploring more limited series or anthology formats.
- **Geo-Targeted Content and Marketing:** Capitalize on the strong presence in the United States, India, the

<pre>df.head(10)</pre>	#top :	<p>United Kingdom, Japan, and South Korea. Develop and market content specifically tailored to the cultural preferences and languages of these countries. This could include acquiring more local content, producing original shows and movies with local talent, and running targeted marketing campaigns. Expanding efforts in these already successful markets might yield better returns than trying to penetrate entirely new territories immediately.</p> <ul style="list-style-type: none">• Prioritize Popular Genres: Use the identified top 10 genres to guide content acquisition and production decisions. Ensure a consistent flow of high-quality content within these popular categories to satisfy existing subscriber demand. Explore sub-genres within these popular categories that might have untapped potential.• Analyze Performance of Older Content: While the focus is on newer releases, analyze the viewing data for older content. If certain older titles still attract significant viewership, consider promoting them more actively or even investing in sequels, prequels, or spin-offs. This could leverage existing intellectual property and cater to nostalgic audiences.
------------------------	--------	--



	show_id	type	title	director	country
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	USA
1	s2	TV Show	Blood & Water	NaN	South Africa
2	s3	TV Show	Ganglands	Julien Leclercq	France
3	s4	TV Show	Jailbirds New Orleans	NaN	USA
4	s5	TV Show	Kota Factory	NaN	India
5	s6	TV Show	Midnight Mass	Mike Flanagan	USA
6	s7	Movie	My Little Pony: A New Generation	Robert Cullen, José Luis Ucha	USA
7	s8	Movie	Sankofa	Haile Gerima	Nigeria
8	s9	TV	The Great British	Andy	UK

- **Optimize Content Addition Schedule:** Review the content addition schedule. While October shows a peak, aim for a more uniform distribution of new content throughout the year. Consider aligning new releases with significant holidays, festivals, or cultural events in target countries to maximize visibility and potential viewership. Analyze subscriber activity patterns to identify optimal times for adding different types of content.
- **Foster Successful Creative Partnerships:** Identify and actively seek collaborations with the director and cast combinations that have proven popular on the platform. These established partnerships have a track record of attracting audiences and could be valuable assets for future projects. Negotiate deals that secure their talent for exclusive Netflix content.
- **Improve Data Quality for Geographic Analysis:** Invest in cleaning and standardizing the country data to ensure accurate geographical analysis. This will provide a clearer picture of where content is being viewed and inform more precise geo-targeting strategies. Implement data validation processes to prevent future inaccuracies.

The total number of rows: 8807 The total number of columns: 9

```
df.shape
```

```
(8807, 12)
```

The director column has high number of null values or missing value

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8807 entries, 0 to 8806
Data columns (total 12 columns):
#   Column                Non-Null Count  Dtype
---  -
0   show_id                8807 non-null   object
1   type                   8807 non-null   object
2   title                  8807 non-null   object
3   director               6173 non-null   object
4   cast                   7982 non-null   object
5   country                7976 non-null   object
6   date_added             8797 non-null   object
7   release_year           8807 non-null   int64
8   rating                 8803 non-null   object
9   duration               8804 non-null   object
10  listed_in              8807 non-null   object
11  description             8807 non-null   object
dtypes: int64(1), object(11)
memory usage: 825.8+ KB
```

So we having data for 96 years

```
df['release_year'].max()-df['release_year'].min()
```

```
96
```

Below table shows that director column has 2634 null values. Removing the rows containing null values will not be much beneficial. In the next step lets remove unnecessary columns like show_id,description for better and cleaner view of data

```
df.isna().sum()
```

- **Explore User Behavior Beyond Content Type:** While the dataset provides valuable information on content attributes, consider integrating user behavior data (e.g., viewing time, time of day watched, device used) to gain deeper insights into subscriber preferences and viewing habits. [1] [2] This can lead to more personalized recommendations and content strategy adjustments.

By implementing these recommendations, Netflix can leverage the insights from the dataset to refine its content strategy, improve subscriber engagement, and potentially drive further business growth.

► Sources

Rate this answer  



	0
show_id	0
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

In the below code the unnecessary columns have been removed

```
a=df.drop(['show_id','description'],axis=1)
a.head()
```



	director	cast	country	count
1	Kirsten Johnson	NaN	United States	1
2	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	1
3	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	1
4	NaN	NaN	NaN	1
5	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	1

We can say the number of movies more than double of number of tv shows

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



	count
type	
Movie	6131
TV Show	2676

dtype: int64

It is important to note that the number of unique directors are 4528.

The shows and movies of netflix are broadcasted in 748 countries. But there are around 200 countries in the world which means there could be few possibilities:

1. There might be different combination of country and state in the country column
2. There could be inaccurate data entry

So we will fix this issue of country column in upcoming steps

```
a.nunique()
```

```
↔
```

	0
type	2
title	8807
director	4528
cast	7692
country	748
date_added	1767
release_year	74
rating	17
duration	220
listed_in	514

dtype: int64

We can see most of the rating is TV-MA. apart from these, there is duration also inserted in rating. so we will remove those rows.

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



	count
rating	
TV-MA	3207
TV-14	2160
TV-PG	863
R	799
PG-13	490
TV-Y7	334
TV-Y	307
PG	287
TV-G	220
NR	80
G	41
TV-Y7-FV	6
NC-17	3
UR	3
74 min	1
84 min	1
66 min	1

dtype: int64

```
#removing the rows in which rating is given in r
df=df[~df['rating'].str.contains('min',na=False)]
df['rating'].value_counts()
```



	count
rating	
TV-MA	3207
TV-14	2160
TV-PG	863
R	799
PG-13	490
TV-Y7	334
TV-Y	307
PG	287
TV-G	220
NR	80
G	41
TV-Y7-FV	6
NC-17	3
UR	3

dtype: int64

Below are the countries where the Netflix is popular. The top 5 countries are United States, India, United Kingdom, Japan and South Korea so we should focus on these countries as because Netflix is already popular in this countries so it will be easier to Grab more customers or more registered users in this countries

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



	count
country	
United States	2818
India	972
United Kingdom	419
Japan	245
South Korea	199
...	...
Mexico, United States, Spain, Colombia	1
Canada, Norway	1
Finland, Germany, Belgium	1
Argentina, United States, Mexico	1
United Kingdom, United States, Germany, Denmark, Belgium, Japan	1

748 rows × 1 columns

dtype: int64

So one season shows are most popular in Netflix apart from that the two season and season are also popular

```
#unnesting the director column
c1=df['director'].apply(lambda x:str(x).split('.')
df_new1=pd.DataFrame(c1,index=df['title'])
df_new1=df_new1.stack().reset_index()
df_new1.drop(columns=['level_1'],inplace=True)
df_new1.rename(columns={0:'director'},inplace=True)
```

```
#unnesting the cast column
c2=df['cast'].apply(lambda x:str(x).split(',')
df_new2=pd.DataFrame(c2,index=df['title'])
df_new2=df_new2.stack().reset_index()
df_new2.drop(columns=['level_1'],inplace=True)
df_new2.rename(columns={0:'cast'},inplace=True)
```

```
#unnesting the genre
c3=df['listed_in'].apply(lambda x:str(x).split(
df_new3=pd.DataFrame(c3,index=df['title'])
df_new3=df_new3.stack().reset_index()
df_new3.drop(columns=['level_1'],inplace=True)
df_new3.rename(columns={0:'genre'},inplace=True)
```



```
#merging the unnested director with unnested cast
df_new4=df_new1.merge(df_new2,on='title',how='inner')
df_new5=df_new4.merge(df_new3,on='title',how='inner')
```

These are the top cast director combination popular in Netflix

```
a=df_new5.groupby(['director','cast']).size()
a=pd.DataFrame(a)
a.rename(columns={0:'count'},inplace=True)
a.sort_values(by='count',ascending=False).head(20)
```



		count
director	cast	
Toshiya Shinohara	Koji Tsujitani	21
	Houko Kuwashima	21
	Satsuki Yukino	21
	Kumiko Watanabe	21
	Kappei Yamaguchi	21
David Dhawan	Anupam Kher	18
Cathy Garcia-Molina	Joross Gamboa	17
Yılmaz Erdoğan	Yılmaz Erdoğan	16
Sooraj R. Barjatya	Alok Nath	15
Toshiya Shinohara	Ken Narita	15
Thierry Donard	Wille Lindberg	15
Toshiya Shinohara	Noriko Hidaka	15
Cathy Garcia-Molina	John Lloyd Cruz	13
Hakan Algül	Salih Kalyon	13

So below are the actor director combination which is more popular in Netflix. So Netflix should look for these actor

director combination

```
a[a['count']>10]
```



count



director

cast



Cathy Garcia-Molina	Al Tantay	11
	Joross Gamboa	17
	John Lloyd Cruz	13
David Dhawan	Anupam Kher	18
Hakan Algül	Salih Kalyon	13
Khaled Marei	Ahmed Helmy	12
Masahiko Murata	Chie Nakamura	12
	Junko Takeuchi	12
Omoni Oboli	Omoni Oboli	12
Priyadarshan	Rajpal Yadav	11
Sooraj R. Barjatya	Alok Nath	15
	Mohnish Bahl	12
	Salman Khan	12
Steven Spielberg	Harrison Ford	12
Thierry Donard	Jesse Richman	12
	Matt Annetts	12
	Wille Lindberg	15
Toshiya Shinohara	Houko Kuwashima	21
	Ken Narita	15
	Koji Tsujitani	21
	Kumiko Watanabe	21
	Noriko Hidaka	15
	Satsuki Yukino	21
	Kappei Yamaguchi	21
Wilson Yip	Donnie Yen	13
Youssef Chahine	Mahmoud El Meleigy	11
Yılmaz Erdoğan	Yılmaz Erdoğan	16

These are the top 10 genres popular in Netflix. So Netflix should always look for these kind of genres

```
df_new3['genre'].value_counts().head(10)
```




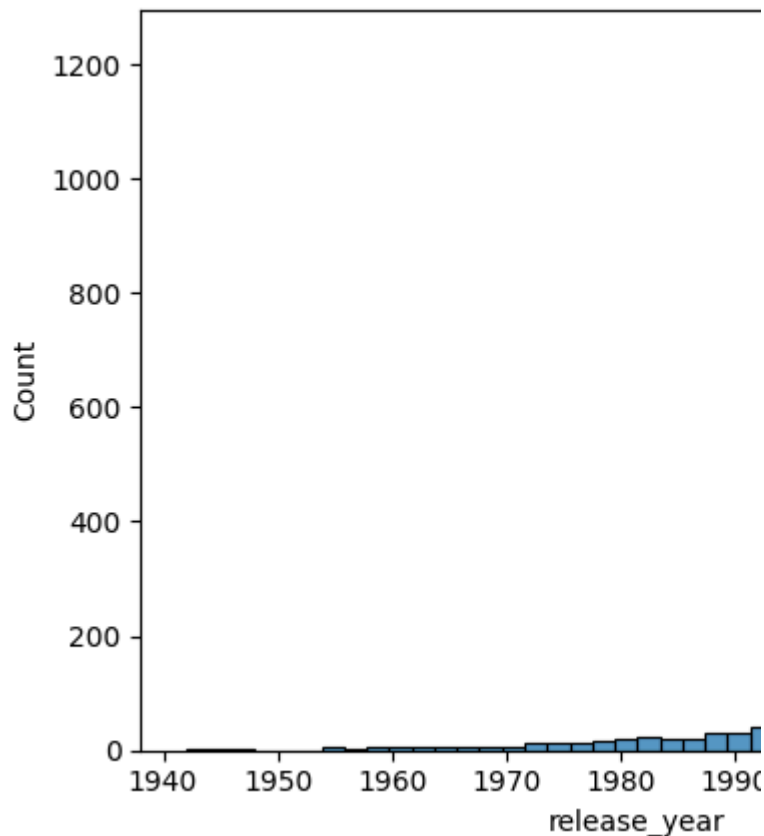
	count
genre	
International Movies	2369
Dramas	2294
Comedies	1553
Action & Adventure	806
Independent Movies	740
Romantic Movies	579
Thrillers	547
Children & Family Movies	503
Documentaries	391
Horror Movies	336

dtype: int64

By looking at the histplot of release year we can say the the number of movies released has been increase sharply after 2000

```
sns.histplot(df['release_year'],bins=40)
```

 <Axes: xlabel='release_year', ylabel='Count'>



So the month of october has the highest number of movies added. Movies should be added uniformly every month. also more movies sholud be added according to the festivals of different countries

```
z=df['date_added'].apply(lambda x:str(x).split(
z=pd.DataFrame(z)
z.drop(columns=[1,2,3],inplace=True)
z.groupby(0).size().sort_values(ascending=False)
```



0

0

October	491
December	490
January	489
April	471
March	469
July	464
November	457
August	448
September	427
June	415
May	367
February	341
	4

dtype: int64

Business Insights from the Netflix Dataset Based on the analysis performed in the notebook, here are some potential business insights that Netflix could gain from this dataset:

Content Mix: The analysis shows that the number of movies is more than double the number of TV shows. This indicates that movies are a significantly larger part of Netflix's content library. Netflix should consider if this aligns with their overall strategy and subscriber preferences. While movies are popular, investing more in TV shows, particularly those that have multiple seasons, could potentially increase subscriber engagement and retention. The analysis also highlighted that "one season" shows are the most popular type of TV show. This could suggest that investing in limited series or shorter-run shows is a viable strategy.

Geographic Focus: The top 5 countries where Netflix is most popular are the United States, India, the United Kingdom, Japan, and South Korea. Focusing marketing

efforts, content acquisition, and even local content production in these countries could be a key driver for subscriber growth and engagement. Since Netflix is already popular in these regions, it might be easier to acquire new customers compared to less penetrated markets.

Popular Genres: Identifying the top 10 popular genres is crucial for content strategy. Netflix should prioritize acquiring or producing content within these popular genres to cater to existing subscriber preferences and attract new ones. Understanding which genres perform best can help in making informed decisions about future content investments.

Release Year Trends: The histogram of release years shows a sharp increase in the number of movies released after 2000. This suggests a trend towards more recent content being added to the platform. While this is expected, analyzing the performance of older content versus newer content could provide insights into subscriber viewing habits and content longevity.

Content Addition Timing: The analysis of the month of content addition shows that October has the highest number of movies added. While this might be tied to specific release schedules or holidays, it raises the

Enter a prompt here



0 / 2000

Gemini can make mistakes so double-check responses and use code with caution. [Learn more](#)