

!gdown [https://d2beiqkhq929f0.cloudfront.net/public\\_assets/assets/000/000/940/original/netflix.csv](https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/000/940/original/netflix.csv)

Downloading...  
 From: [https://d2beiqkhq929f0.cloudfront.net/public\\_assets/assets/000/000/940/original/netflix.csv](https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/000/940/original/netflix.csv)  
 To: /content/netflix.csv  
 100% 3.40M/3.40M [00:00<00:00, 49.4MB/s]

Start coding or [generate](#) with AI.

## ✓ UNDERSTANDING THE BUSINESS PROBLEM

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 deciding which type of shows/movies to produce and how they can grow the business in different countries

Start coding or [generate](#) with AI.

## ✓ OBJECTIVES

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

### OBJECTIVE OR THE AREAS TO FIGURE OUT

How has the number of movies released per year changed over the last 20-30 years

Comparison of tv shows vs. movies.

What is the best time to launch a TV show

Analysis of actors/directors of different types of shows/movies.

Does Netflix has more focus on TV Shows than movies in recent years

Understanding what content is available in different countries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

```
df=pd.read_csv("netflix.csv")
```

## ✓ DATA INSPECTION

```
df.head()
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...
		TV		Julian	Sami Bouajila, Tracy		September				Crime TV Shows	To protect his family from a

```
df.tail()
```

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	November 20, 2019	2007	R	158 min	Cult Movies, Dramas, Thrillers	A political cartoonist, a crime reporter and a...
8803	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	July 1, 2019	2018	TV-Y7	2 Seasons	Kids' TV, Korean TV Shows, TV Comedies	While living alone in a spooky town, a young g...
				Ruben	Jesse Eisenberg, Woody	United	November				Comedies,	Looking to survive in a

```
df.columns
```

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

```
df.dtypes
```

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

CONVERTING THE DATE\_ADDED INTO DATETIME FORMAT

```
df['date_added'] = pd.to_datetime(df['date_added'], errors='coerce')
```


```
df['date_added']
```

```
0    2021-09-25
1    2021-09-24
2    2021-09-24
3    2021-09-24
4    2021-09-24
...
```



```
8802    2019-11-20
8803    2019-07-01
8804    2019-11-01
8805    2020-01-11
8806    2019-03-02
Name: date_added, Length: 8807, dtype: datetime64[ns]
```

```
df["year_added"]=df['date_added'].dt.year
df["month_added"]=df['date_added'].dt.month
```


```
df.head()
```



	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description	year_ac
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	2021-09-25	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...	20
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...	20
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel	NaN	2021-09-24	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV	To protect his family from a powerful drug lor...	20



```
df["release_year"]
```



```
0      2020
1      2021
2      2021
3      2021
4      2021
...
8802   2007
8803   2018
8804   2009
8805   2006
8806   2015
Name: release_year, Length: 8807, dtype: int64
```

```
df['Movie_Mins'] = df[df['type']=='m']['duration'].apply(lambda x: int(x.split(" ")[0]))
df['Number of Seasons'] = df[df['type']=='tv']['duration'].apply(lambda x: int(x.split(" ")[0]))
```

```
df
```



	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	2021-09-25	2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	2021-09-24	2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	2021-09-24	2021	TV-MA	1 Season	Docuseries, Reality TV	Feuds, flirtations and toilet talk go down amo...
4	s5	TV Show	Kota Factory	NaN	Mayur More, Jitendra Kumar, Ranjan Raj, Alam K...	India	2021-09-24	2021	TV-MA	2 Seasons	International TV Shows, Romantic TV Shows, TV ...	In a city of coaching centers known to train I...
...	...	...	...	...	...	...	...	...	...	...	...	...
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	2019-11-20	2007	R	158 min	Cult Movies, Dramas, Thrillers	A political cartoonist, a crime reporter and a...
8803	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	2019-07-01	2018	TV-Y7	2 Seasons	Kids' TV, Korean TV Shows, TV Comedies	While living alone in a spooky town, a young g...
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	2019-11-01	2009	R	88 min	Comedies, Horror Movies	Looking to survive in a world taken over by zo...

df.tail()



	show_id	type	title	director	cast	country	date_added	release_year
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	2019-11-20	2007
8803	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	2019-07-01	2018
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone, ...	United States	2019-11-01	2009
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma...	United States	2020-01-11	2006
				Mozez	Vicky Kaushal, Sarah-			

```
df['duration']
```



```
0      90 min
1      2 Seasons
2      1 Season
3      1 Season
4      2 Seasons
...
8802    158 min
8803    2 Seasons
8804     88 min
8805     88 min
8806    111 min
Name: duration, Length: 8807, dtype: object
```

✓ DURATION COLUMN CONSISTS OF BOTH THE MINUTES AND THE SEASONS DETAILS ,THEREFORE THAT NEEDES TO BE SEPARATED INTO TWO COLUMNS

```
df['Movie_Mins'] = df[df['type'] == 'Movie']['duration'].apply(
    lambda x: int(x.split(" ")[0]) if pd.notna(x) else None
)
```

```
df['Number of Seasons'] = df[df['type'] == 'TV Show']['duration'].apply(
    lambda x: int(x.split(" ")[0]) if pd.notna(x) else None
)
```

```
df.head()
```



	show_id	type	title	director	cast	country	date_added	release_year	rat
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	2021-09-25	2020	PG
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV-
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	2021-09-24	2021	TV-
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	2021-09-24	2021	TV-
					Mayur More, Jitendra				

```
df.tail(10)
```



	show_id	type	title	director	cast	country	date_added	release_y
8797	s8798	TV Show	Zak Storm	NaN	Michael Johnston, Jessica Gee-George, Christin...	United States, France, South Korea, Indonesia	2018-09-13	2
8798	s8799	Movie	Zed Plus	Chandra Prakash Dwivedi	Adil Hussain, Mona Singh, K.K. Raina, Sanjay M...	India	2019-12-31	2
8799	s8800	Movie	Zenda	Avadhoot Gupte	Santosh Juvekar, Siddharth Chandekar, Sachit P...	India	2018-02-15	2
8800	s8801	TV Show	Zindagi Gulzar Hai	NaN	Sanam Saeed, Fawad Khan, Ayesha Omer, Mehreen ...	Pakistan	2016-12-15	2
8801	s8802	Movie	Zinzana	Majid Al Ansari	Ali Suliman, Saleh Bakri, Yasa, Ali Al-Jabri, ...	United Arab Emirates, Jordan	2016-03-09	2
8802	s8803	Movie	Zodiac	David Fincher	Mark Ruffalo, Jake Gyllenhaal, Robert Downey J...	United States	2019-11-20	2
8803	s8804	TV Show	Zombie Dumb	NaN	NaN	NaN	2019-07-01	2
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma	United States	2019-11-01	2

```
df.head()
```



	show_id	type	title	director	cast	country	date_added	release_year	rat
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	2021-09-25	2020	PG
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	2021-09-24	2021	TV-
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	2021-09-24	2021	TV-
3	s4	TV Show	Jailbirds New Orleans	NaN	NaN	NaN	2021-09-24	2021	TV-
					Mayur More, Jitendra				

Start coding or [generate](#) with AI.

WORKING ON THE OBJECTIVES

How has the number of movies released per year changed over the last 20-30 years?

```
df.columns
Index(['show_id', 'type', 'title', 'director', 'cast', 'country', 'date_added',
      'release_year', 'rating', 'duration', 'listed_in', 'description',
      'year_added', 'month_added', 'Movie_Mins', 'Number of Seasons'],
      dtype='object')
```

Double-click (or enter) to edit

```
df['date_added'].dt.year
0      2021.0
1      2021.0
2      2021.0
3      2021.0
4      2021.0
...
8802   2019.0
8803   2019.0
8804   2019.0
8805   2020.0
8806   2019.0
Name: date_added, Length: 8807, dtype: float64
```

```
movies_per_release_year=df.groupby("release_year").size()
```

```
movies_per_release_year
```

```
release_year
1925      1
1942      2
```



```

1943      3
1944      3
1945      4
...
2017    1032
2018    1147
2019    1030
2020     953
2021     592
Length: 74, dtype: int64

```

Start coding or [generate](#) with AI.

```
movies_per_addede_year=df.groupby("year_added").size()
```

```
movies_per_addede_year
```

```

↩ year_added
2008.0      2
2009.0      2
2010.0      1
2011.0     13
2012.0      3
2013.0     10
2014.0     23
2015.0     73
2016.0    418
2017.0   1164
2018.0   1625
2019.0   1999
2020.0   1878
2021.0   1498
dtype: int64

```

✓ CONSIDERING THE CURRENT YEAR AS 2024 APPLYING FOR THE PROBLEM STATEMENT TO CALCULATE FOR THE MOVIES RELEASE PER YEAR AND HAS CHANGED OVER 20-30 YEARS

```
current_year =2024
```

```
years_range=range(current_year -30,current_year +1)
```

```

movies_per_release_year = movies_per_release_year[movies_per_release_year.index.isin(years_range)]
movies_per_addede_year = movies_per_addede_year[movies_per_addede_year.index.isin(years_range)]

```

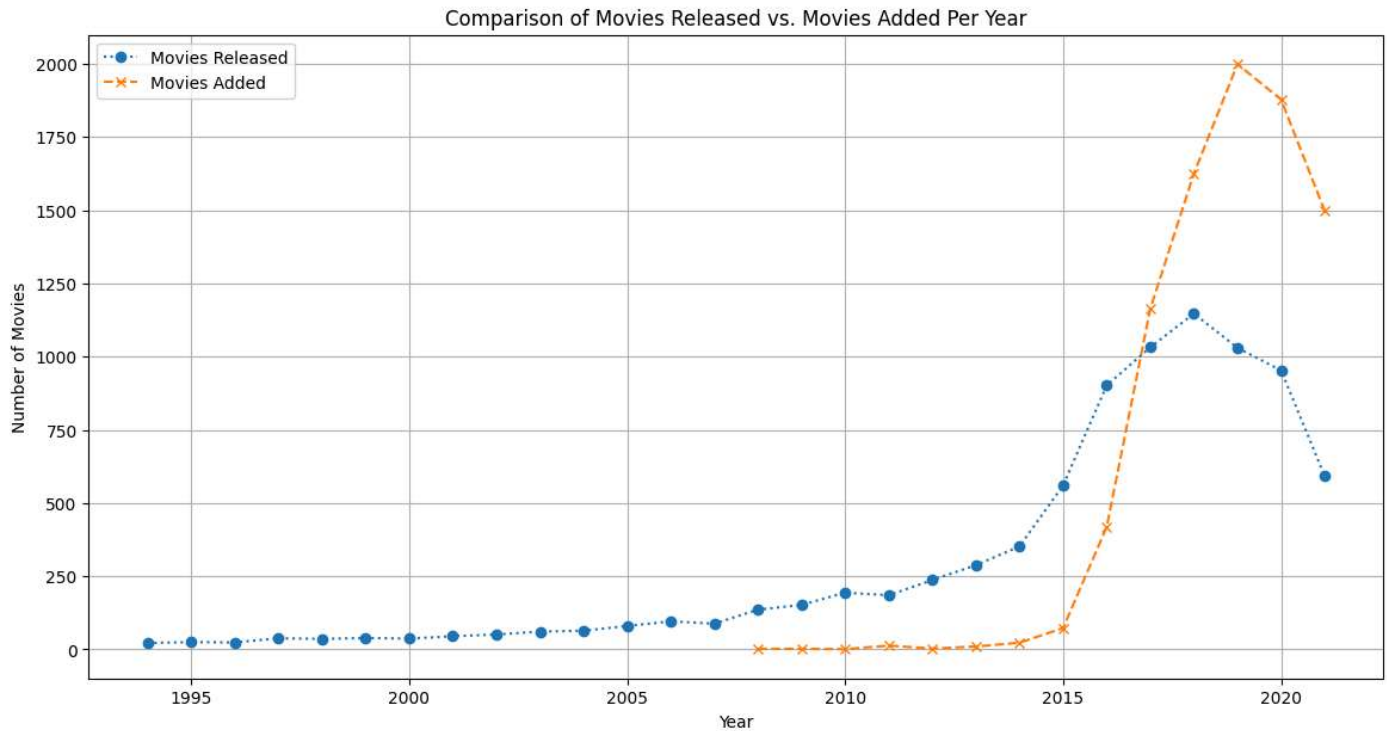
Start coding or [generate](#) with AI.

```

plt.figure(figsize=(14, 7))
plt.plot(movies_per_release_year.index, movies_per_release_year, label='Movies Released', marker='o', linestyle=':')
plt.plot(movies_per_addede_year.index, movies_per_addede_year, label='Movies Added', marker='x', linestyle='--')

plt.title('Comparison of Movies Released vs. Movies Added Per Year')
plt.xlabel('Year')
plt.ylabel('Number of Movies')
plt.legend()
plt.grid(True)
plt.show()

```



Start coding or [generate](#) with AI.

#### ✓ THE INSIGHT DRAFTED STATES THAT THE RELAESED MOVIES

THE ADDED MOVIES HAS FOUND THE MASSIVE INCREASE FROM THE YEAR 2015 ONWARDS

NUMBER OF MOVIES RELEASED IS >2000 IN THE YEAR 2019

Start coding or [generate](#) with AI.

#### ✓ Comparison of tv shows vs. movies.

```
df["Number of Seasons"]
```



```
0      NaN
1      2.0
2      1.0
3      1.0
4      2.0
...
8802   NaN
8803   2.0
8804   NaN
8805   NaN
8806   NaN
Name: Number of Seasons, Length: 8807, dtype: float64
```

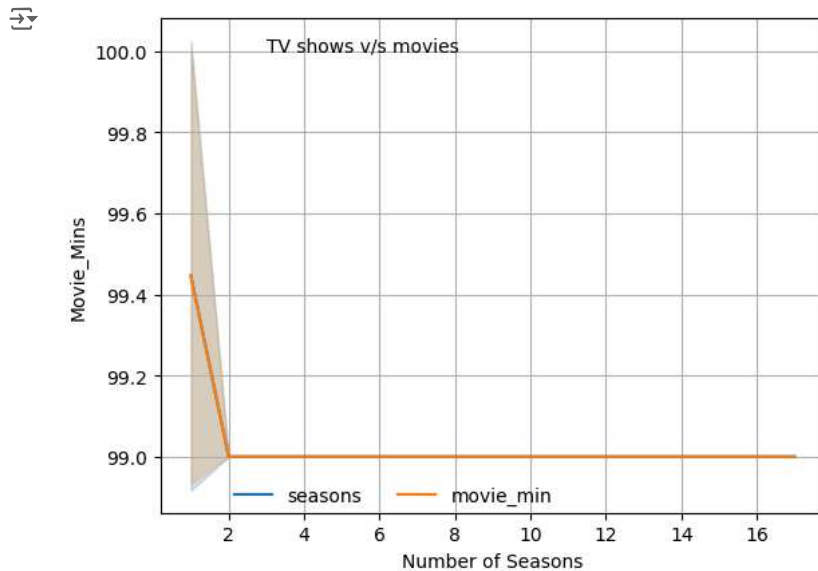
```
df.columns
```



```
Index(['show_id', 'type', 'title', 'director', 'cast', 'country', 'date_added',
       'release_year', 'rating', 'duration', 'listed_in', 'description',
       'year_added', 'month_added', 'Movie_Mins', 'Number of Seasons'],
      dtype='object')
```

```
df['Number of Seasons'].fillna(int(df['Number of Seasons'].mean()), inplace=True)
df['Movie_Mins'].fillna(int(df['Movie_Mins'].mean()), inplace=True)
```

```
sns.lineplot(x='Number of Seasons', y='Movie_Mins', data=df, label='seasons')
sns.lineplot(x='Number of Seasons', y='Movie_Mins', data=df, label='movie_min')
plt.legend(loc=(0.1, 0), ncol=2, frameon=False)
plt.text(x=3, y=100, s="TV shows v/s movies")
plt.grid()
plt.show()
```



## INSIGHT

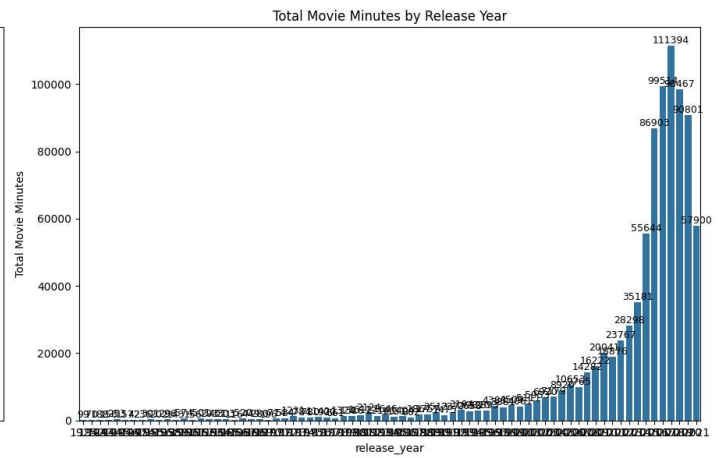
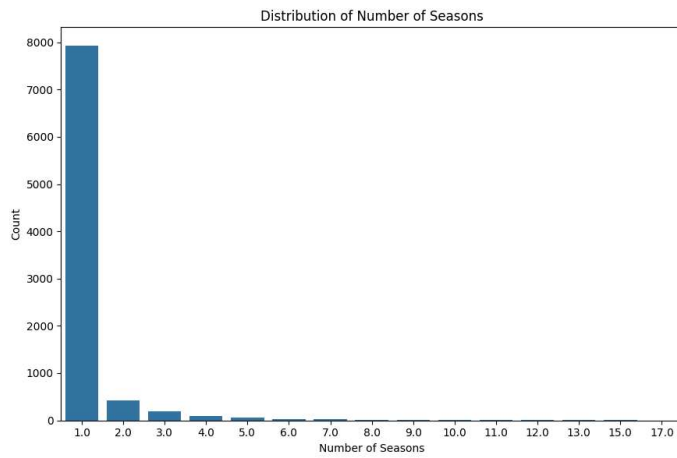
COMPASION MADE BETWEEN THE NUMBER OF SEASONS AND MOVIES MADE SHOWS THAT PEOPLE ARE SHOWING MORE INTREST ON MOVIES WITH MORE MINS OR THE TIME AND TV SHOWS WITH 1 OR 2 SEASONS

```
plt.figure(figsize=(18, 6)) # Set the overall figure size

# First subplot for the countplot
plt.subplot(1, 2, 1) # 1 row, 2 columns, first plot
ax1 = sns.countplot(x="Number of Seasons", data=df)
ax1.set_title('Distribution of Number of Seasons')
ax1.set_ylabel('Count')

# Second subplot for the barplot
plt.subplot(1, 2, 2) # 1 row, 2 columns, second plot
ax2 = sns.barplot(data=df, x="release_year", y="Movie_Mins", estimator=sum, errorbar=None)
ax2.set_title('Total Movie Minutes by Release Year')
ax2.bar_label(ax2.containers[0], fontsize=9)
ax2.set_ylabel('Total Movie Minutes')

# Show the plots
plt.tight_layout() # Adjusts subplots to give some padding and prevent overlap
plt.show()
```



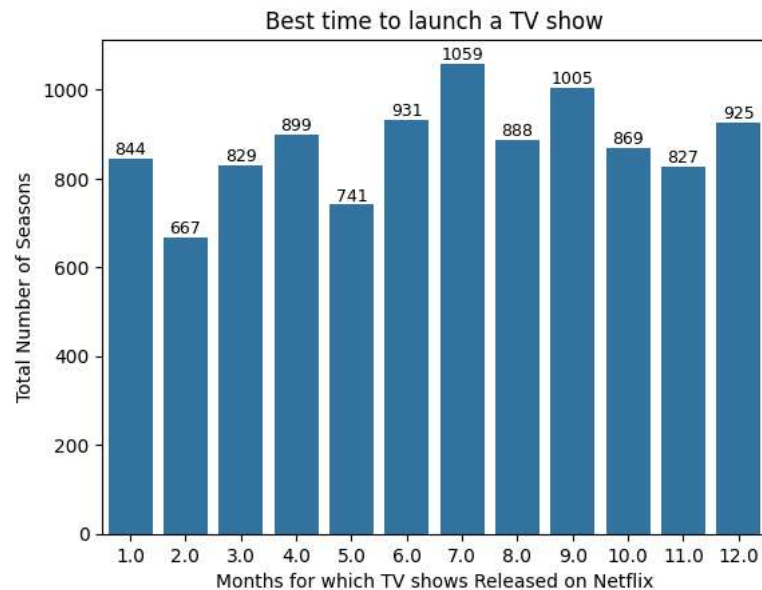
Start coding or [generate](#) with AI.

✓ What is the best time to launch a TV show?

```
ax2 = sns.barplot(data=df, x="month_added", y="Number of Seasons", estimator=sum, errorbar=None)
ax2.set_title('Best time to launch a TV show')
ax2.bar_label(ax2.containers[0], fontsize=9)
ax2.set_ylabel('Total Number of Seasons')
ax2.set_xlabel('Months for which TV shows Released on Netflix')
```



Text(0.5, 0, 'Months for which TV shows Released on Netflix')



Start coding or [generate](#) with AI.

###INSIGHTS

✓ FROM THE ABOVE GRAPH WE CAN DRAFT THE INSIGHT THAT MORE NUMBER OF TV SHOWS CAN BE LAUNCHED IN THE MONTH OF THE JULY AND THE SEPTEMBER MONTH ALSO HAS GOT MORE

```
### Analysis of actors/directors of different types of shows/movies.
```

```
df["director"]
```

```
0      Kirsten Johnson
1      NaN
2      Julien Leclercq
3      NaN
4      NaN
...
8802    David Fincher
8803    NaN
8804    Ruben Fleischer
8805    Peter Hewitt
8806    Mozez Singh
Name: director, Length: 8807, dtype: object
```

```
df.head()
```

```
0      0      s1      Movie      Dick Johnson Is Dead      Kirsten Johnson      NaN      United States      2021-09-25      2020      PG

1      1      s2      TV Show      Blood & Water      NaN      Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...      South Africa      2021-09-24      2021      TV-

2      2      s3      TV Show      Ganglands      Julien Leclercq      Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...      NaN      2021-09-24      2021      TV-

3      3      s4      TV Show      Jailbirds New Orleans      NaN      NaN      NaN      2021-09-24      2021      TV-

Mayur More, Jitendra
```

```
df.loc[df["type"] == "Movie", "director"]
```

```
0      Kirsten Johnson
6      Robert Cullen, José Luis Ucha
7      Haile Gerima
9      Theodore Melfi
12     Christian Schwochow
...
8801    Majid Al Ansari
8802    David Fincher
8804    Ruben Fleischer
8805    Peter Hewitt
8806    Mozez Singh
Name: director, Length: 6131, dtype: object
```

```
df.loc[df["type"] == "TV Show", "director"]
```

```
1      NaN
2      Julien Leclercq
3      NaN
4      NaN
5      Mike Flanagan
...
8795    NaN
```

```

8796      NaN
8797      NaN
8800      NaN
8803      NaN
Name: director, Length: 2676, dtype: object

```

```
df.columns
```

```

Index(['show_id', 'type', 'title', 'director', 'cast', 'country', 'date_added',
      'release_year', 'rating', 'duration', 'listed_in', 'description',
      'year_added', 'month_added', 'Movie_Mins', 'Number of Seasons'],
      dtype='object')

```

```
df['director'].fillna('Unknown Director', inplace=True)
```

```
df["Number of Seasons"]
```

```

0      1.0
1      2.0
2      1.0
3      1.0
4      2.0
...
8802    1.0
8803    2.0
8804    1.0
8805    1.0
8806    1.0
Name: Number of Seasons, Length: 8807, dtype: float64

```

```
df['director']
```

```

0      Kirsten Johnson
1      Unknown Director
2      Julien Leclercq
3      Unknown Director
4      Unknown Director
...
8802    David Fincher
8803    Unknown Director
8804    Ruben Fleischer
8805    Peter Hewitt
8806    Mozez Singh
Name: director, Length: 8807, dtype: object

```

```
# Group by 'director' and sum 'Number of Seasons', assuming 'Number of Seasons' is a column in df
```

```
director_seasons = df.groupby('director')['Number of Seasons'].sum().reset_index()
```

```
# Sort the results for better readability and focus on the most prolific directors
```

```
director_seasons = director_seasons.sort_values(by='Number of Seasons', ascending=False)
```

```
# Print the summarized data
```

```
print(director_seasons)
```

```


director  Number of Seasons
4303      Unknown Director    4536.0
3392      Rajiv Chilaka       19.0
3443      Raúl Campos, Jan Suter  18.0
2598      Marcus Raboy        16.0
4046      Suhas Kadav         16.0
...      ...
2070      Joshua Oppenheimer     1.0
2071      Joshua Rofé           1.0
2072      Joshua Tickell, Rebecca Harrell Tickell  1.0
2073      Joshua Zeman          1.0
2264      Khaled Youssef        1.0

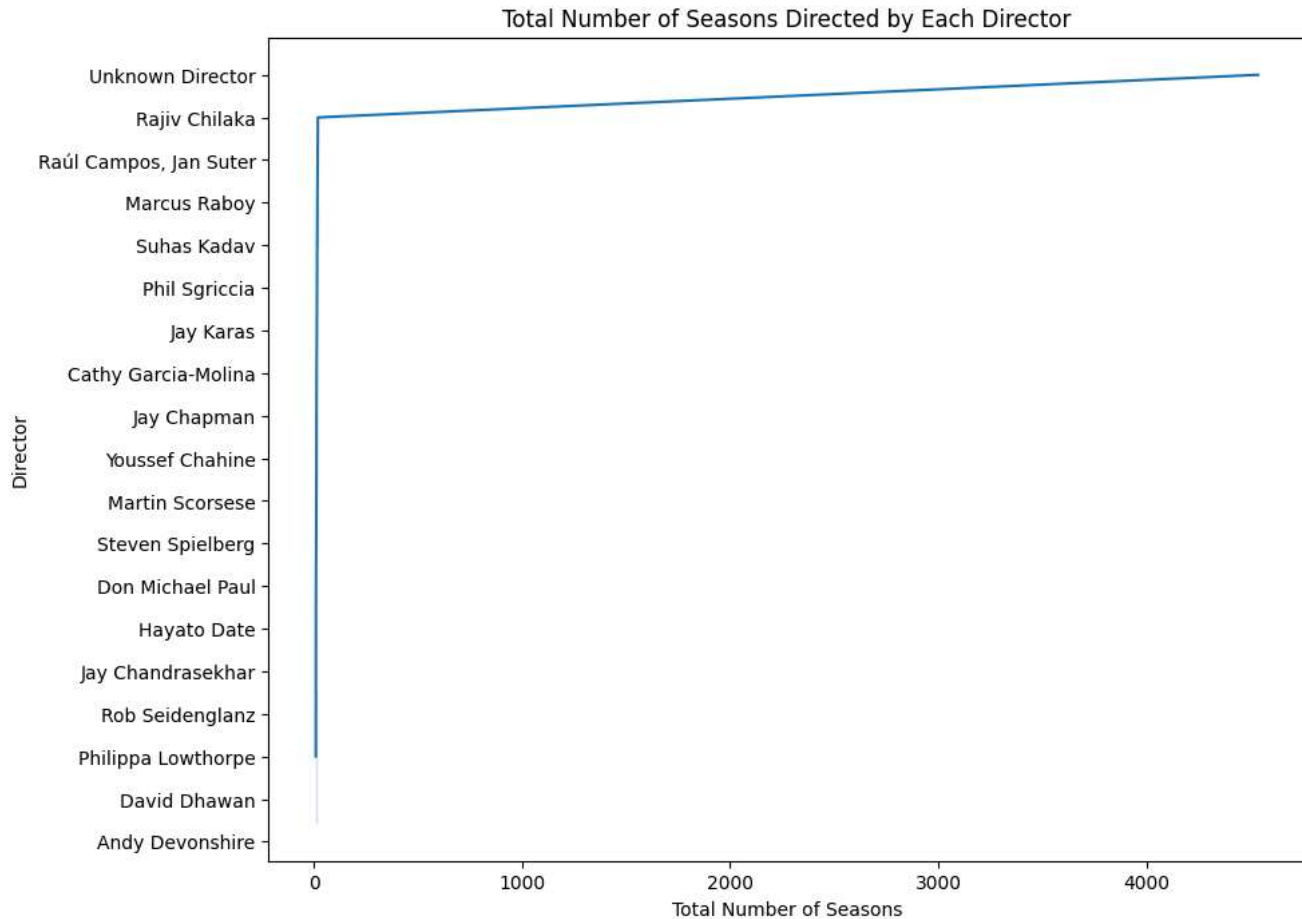
```

```
[4529 rows x 2 columns]
```

```
# Plotting using a horizontal bar chart for better readability
```

```
plt.figure(figsize=(10, 8))
sns.lineplot(x='Number of Seasons', y='director', data=director_seasons.head(20), palette='viridis') # Show top 20 for brevity
plt.title('Total Number of Seasons Directed by Each Director')
plt.xlabel('Total Number of Seasons')
plt.ylabel('Director')
plt.show()
```


 <ipython-input-219-4e78cd8e873f>:4: UserWarning: Ignoring `palette` because no `hue` variable has been assigned.  
sns.lineplot(x='Number of Seasons', y='director', data=director\_seasons.head(20), palette='viridis') # Show top 20 for brevity




## INSIGHTS

MAXIMUM NUMBER OF TV SEASONS ARE MADE BY DIRECTOR RAJIV CHILAKA

```
df.columns
```

 Index(['show\_id', 'type', 'title', 'director', 'cast', 'country', 'date\_added', 'release\_year', 'rating', 'duration', 'listed\_in', 'description', 'year\_added', 'month\_added', 'Movie\_Mins', 'Number of Seasons'], dtype='object')

```
df["Movie_Mins"]
```

 0 90.0  
1 99.0  
2 99.0  
3 99.0  
4 99.0  
...  
8802 158.0  
8803 99.0  
8804 88.0  
8805 88.0

```
8806      111.0
Name: Movie Mins, Length: 8807, dtype: float64
```

```
# Step 1: Find the maximum number of movie minutes
```

```
max_duration = df['Movie_Mins'].max()
```

```
# Step 2: Filter the DataFrame to find the movie(s) with this maximum duration
```

```
max_duration_movies = df[df['Movie_Mins'] == max_duration]
```

```
# Display the directors of these movies
```

```
print(max_duration_movies[['director', 'Movie_Mins']])
```

```
↔
director  Movie_Mins
4253  Unknown Director    312.0
```

```
plt.figure(figsize=(10, 8))
```

```
sns.barplot(x='Movie_Mins', y='director', data=df.head(20), palette='viridis') # Show top 20 for brevity
```

```
plt.title('Total Number of Seasons Directed by Each Director')
```

```
plt.xlabel('Total Number of Seasons')
```

```
plt.ylabel('Director')
```

```
plt.show()
```