Double-click (or enter) to edit

Importing required libraries

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

Loading Netflix Dataset

```
netflix_data= pd.read_csv('netflix.csv')
netflix_data.head()
```

	show_id	type	title	director	cast	country	date_added	release_ye
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States	September 25, 2021	2(
1	s2	TV Show	Blood & Water	NaN	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2(
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy	NaN	September 24, 2021	2(

Next steps: Generate code with `netflix_data` View recommended plots

Understanding the Data

```
#Finding the shape, size, data types , missing values etc

netflix_data.shape
```

```
(8807, 12)
```

Observation: Dataset has 8807 records/rows and 12 columns.

```
netflix_data.size # size gives total number of elements in the dataset/dataframe
```

```
105684
```

```
# info method gives the details of the dataframe like index range ,column names,
# non-null value count and data type

netflix_data.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
```

```
#Getting statistics of numerical columns in the dataframe

netflix_data.describe()
```

	release_year
count	8807.000000
mean	2014.180198
std	8.819312
min	1925.000000
25%	2013.000000
50%	2017.000000
75%	2019.000000
max	2021.000000

Observation: Dataset has data from the year 1925 to the year 2021 , i.e 96 years of data

```
#Finding the unique values in the dataset

netflix_data.nunique()
```

```
show_id         8807
type               2
title           8807
director        4528
cast            7692
country          748
date_added      1767
release_year      74
rating            17
duration         220
listed_in        514
description     8775
dtype: int64
```

Data Preperation

1. Unnesting the columns that have multiple values
2. Finding null/missing values and filling them
3. Data type conversions
4. Renaming columns

```
# Checking for duplicate records
netflix_data.duplicated().sum()
```

```
0
```

Observation: There are no duplicate records

```
#Finding null/missing values

missing_values=pd.Series(netflix_data.isna().sum())
missing_values
```

```
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
```

```
missing_values_percentage= pd.Series((netflix_data.isna().sum()/len(netflix_data))*100)
missing_values_percentage
```

```
show_id          0.000000
type             0.000000
title            0.000000
director        29.908028
cast             9.367549
country          9.435676
date_added       0.113546
release_year     0.000000
rating           0.045418
duration         0.034064
listed_in        0.000000
description      0.000000
dtype: float64
```

```
# Missing values dataframe
missing_df=pd.concat([missing_values,missing_values_percentage],axis=1,keys=['Total','Percentage'])
missing_df
```

	Total	Percentage
show_id	0	0.000000
type	0	0.000000
title	0	0.000000
director	2634	29.908028
cast	825	9.367549
country	831	9.435676
date_added	10	0.113546
release_year	0	0.000000
rating	4	0.045418
duration	3	0.034064
listed_in	0	0.000000
description	0	0.000000

Next steps: Generate code with missing_df View recommended plots

Observation: Netflix dataset has almost 30% of director column has null values followed by cast and country

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

```
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
Name: count, dtype: int64
```

Observation: Rating has values in min , that should be filled in duration

```
# forward filling rating values to duration where valid observation is given

ind=netflix_data[netflix_data['duration'].isna()].index
ind
```

```
Index([5541, 5794, 5813], dtype='int64')
```

```
netflix_data.loc[ind]= netflix_data.loc[ind].fillna(method='ffill',axis=1)
netflix_data.loc[ind]
```

	show_id	type	title	director	cast	country	date_added	release_year
5541	s5542	Movie	Louis C.K. 2017	Louis C.K.	Louis C.K.	United States	April 4, 2017	2017
5794	s5795	Movie	Louis C.K.: Hilarious	Louis C.K.	Louis C.K.	United States	September 16, 2016	2010

```
netflix_data.loc[ind,'rating']='Not Avaliable'
netflix_data.loc[ind]
```

	show_id	type	title	director	cast	country	date_added	release_year
5541	s5542	Movie	Louis C.K. 2017	Louis C.K.	Louis C.K.	United States	April 4, 2017	2017
5794	s5795	Movie	Louis C.K.: Hilarious	Louis C.K.	Louis C.K.	United States	September 16, 2016	2010

**Filling missing values **

```
netflix_data['director'] = netflix_data['director'].fillna('Unspecified')
netflix_data['cast'] = netflix_data['cast'].fillna('Unknown')
netflix_data['country'] = netflix_data['country'].fillna(netflix_data['country'].mode()[0])
netflix_data['date_added'] = netflix_data['date_added'].fillna(netflix_data['date_added'].mode()[0])
netflix_data['duration']= netflix_data['duration'].fillna(netflix_data['duration'].mode()[0])
netflix_data['rating'] = netflix_data['rating'].fillna('Not Available')
```

```
netflix_data.isna().sum()
```

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

```
netflix_data.head()
```

	show_id	type	title	director	cast	country	date_added	release_y
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Unknown	United States	September 25, 2021	2
1	s2	TV Show	Blood & Water	Unspecified	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi	United States	September 24, 2021	2

Next steps: Generate code with `netflix_data` View recommended plots

```
# Copying the dataset before cleaning
netflix_original=netflix_data.copy()
netflix_original.head()
```

	show_id	type	title	director	cast	country	date_added	release_y
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Unknown	United States	September 25, 2021	2
1	s2	TV Show	Blood & Water	Unspecified	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi	United States	September 24, 2021	2

Next steps: Generate code with netflix_original View recommended plots

```
#Columns director, cast, listend_in have multiple values seperated by comma

#unnesting the cast column and considering only the main actor for the analasys

netflix_data['cast']=netflix_data['cast'].str.split(",").str[0]

#Renaming the cast to "main_actor"

netflix_data.rename(columns={'cast':'main_actor'},inplace=True)
netflix_data.head()
```

	show_id	type	title	director	main_actor	country	date_added	release
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Unknown	United States	September 25, 2021	
1	s2	TV Show	Blood & Water	Unspecified	Ama Qamata	South Africa	September 24, 2021	
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila	United States	September 24, 2021	

Next steps: Generate code with netflix_data View recommended plots

```
#unnesting director column

netflix_data['director']=netflix_data['director'].str.split(",")
netflix_data=netflix_data.explode('director')
netflix_data.head(20)
```

	show_id	type	title	director	main_actor	country	date_added	rele
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Unknown	United States	September 25, 2021	
1	s2	TV Show	Blood & Water	Unspecified	Ama Qamata	South Africa	September 24, 2021	
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila	United States	September 24, 2021	
3	s4	TV Show	Jailbirds New Orleans	Unspecified	Unknown	United States	September 24, 2021	
4	s5	TV Show	Kota Factory	Unspecified	Mayur More	India	September 24, 2021	
5	s6	TV Show	Midnight Mass	Mike Flanagan	Kate Siegel	United States	September 24, 2021	
6	s7	Movie	My Little Pony: A New Generation	Robert Cullen	Vanessa Hudgens	United States	September 24, 2021	
6	s7	Movie	My Little Pony: A New Generation	José Luis Ucha	Vanessa Hudgens	United States	September 24, 2021	
7	s8	Movie	Sankofa	Haile Gerima	Kofi Ghanaba	United States, Ghana, Burkina Faso, United Kin...	September 24, 2021	
8	s9	TV Show	The Great British Baking Show	Andy Devonshire	Mel Giedroyc	United Kingdom	September 24, 2021	
9	s10	Movie	The Starling	Theodore Melfi	Melissa McCarthy	United States	September 24, 2021	
10	s11	TV Show	Vendetta: Truth, Lies and The Mafia	Unspecified	Unknown	United States	September 24, 2021	
11	s12	TV Show	Bangkok Breaking	Kongkiat Komesiri	Sukollawat Kanarot	United States	September 23, 2021	
12	s13	Movie	Je Suis Karl	Christian Schwochow	Luna Wedler	Germany, Czech Republic	September 23, 2021	
13	s14	Movie	Confessions of an Invisible Girl	Bruno Garotti	Klara Castanho	United States	September 22, 2021	

Next steps: Generate code with `netflix_data` View recommended plots

```
#unnesting listed_in column

netflix_data['listed_in']=netflix_data['listed_in'].str.split(",")
netflix_data=netflix_data.explode('listed_in')
netflix_data.head(20)
```

	show_id	type	title	director	main_actor	country	date_added	release
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Unknown	United States	September 25, 2021	
1	s2	TV Show	Blood & Water	Unspecified	Ama Qamata	South Africa	September 24, 2021	
1	s2	TV Show	Blood & Water	Unspecified	Ama Qamata	South Africa	September 24, 2021	
1	s2	TV Show	Blood & Water	Unspecified	Ama Qamata	South Africa	September 24, 2021	
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila	United States	September 24, 2021	
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila	United States	September 24, 2021	
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila	United States	September 24, 2021	
3	s4	TV Show	Jailbirds New Orleans	Unspecified	Unknown	United States	September 24, 2021	
3	s4	TV Show	Jailbirds New Orleans	Unspecified	Unknown	United States	September 24, 2021	
4	s5	TV Show	Kota Factory	Unspecified	Mayur More	India	September 24, 2021	
4	s5	TV Show	Kota Factory	Unspecified	Mayur More	India	September 24, 2021	
4	s5	TV Show	Kota Factory	Unspecified	Mayur More	India	September 24, 2021	
5	s6	TV Show	Midnight Mass	Mike Flanagan	Kate Siegel	United States	September 24, 2021	
5	s6	TV Show	Midnight Mass	Mike Flanagan	Kate Siegel	United States	September 24, 2021	
5	s6	TV Show	Midnight Mass	Mike Flanagan	Kate Siegel	United States	September 24, 2021	
6	s7	Movie	My Little Pony: A New	Robert Cullen	Vanessa Hudgens	United States	September 24, 2021	

Next steps: Generate code with `netflix_data` View recommended plots

```
#unnesting country column

netflix_data['country']=netflix_data['country'].str.split(",")
new=netflix_data.explode('country')


netflix_data=new
netflix_data.head()
```

	show_id	type	title	director	main_actor	country	date_added	release
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Unknown	United States	September 25, 2021	
1	s2	TV Show	Blood & Water	Unspecified	Ama Qamata	South Africa	September 24, 2021	
1	s2	TV Show	Blood & Water	Unspecified	Ama Qamata	South Africa	September 24, 2021	

Next steps: Generate code with netflix_data View recommended plots

```
#Replacing show_id values with int values instead of s1 .. etc

netflix_data['show_id']=netflix_data['show_id'].str.replace('s','')
netflix_data.head()
```

	show_id	type	title	director	main_actor	country	date_added	release
0	1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Unknown	United States	September 25, 2021	
1	2	TV Show	Blood & Water	Unspecified	Ama Qamata	South Africa	September 24, 2021	
1	2	TV Show	Blood & Water	Unspecified	Ama Qamata	South Africa	September 24, 2021	

Next steps: Generate code with netflix_data View recommended plots

```
# reset index after unnesting different columns
netflix_data.reset_index()
```

	index	show_id	type	title	director	main_actor	country	date_add
0	0	1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Unknown	United States	Septemb 25, 20
1	1	2	TV Show	Blood & Water	Unspecified	Ama Qamata	South Africa	Septemb 24, 20
2	1	2	TV Show	Blood & Water	Unspecified	Ama Qamata	South Africa	Septemb 24, 20
3	1	2	TV Show	Blood & Water	Unspecified	Ama Qamata	South Africa	Septemb 24, 20
4	2	3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila	United States	Septemb 24, 20
...	...	...	...	...	...	...	...	
25895	8805	8806	Movie	Zoom	Peter Hewitt	Tim Allen	United States	January 20
25896	8805	8806	Movie	Zoom	Peter Hewitt	Tim Allen	United States	January 20
25897	8806	8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal	India	March 20

```
# Renaming listed_in column to genre

netflix_data.rename(columns={'listed_in':'genre'},inplace=True)
netflix_data.head()
```

	show_id	type	title	director	main_actor	country	date_added	release
0	1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Unknown	United States	September 25, 2021	
1	2	TV Show	Blood & Water	Unspecified	Ama Qamata	South Africa	September 24, 2021	
1	2	TV Show	Blood & Water	Unspecified	Ama Qamata	South Africa	September 24, 2021	

Next steps: Generate code with `netflix_data` View recommended plots

```
netflix_data.dtypes
```

```
show_id         object
type            object
title           object
director        object
main_actor      object
country         object
date_added      object
release_year    object
rating          object
duration        object
genre           object
description     object
dtype: object
```

```
# stripping the spaces for the date_added column values

netflix_data['date_added']=netflix_data['date_added'].str.strip()


# Converting object to datetime for the column date_added
date_format = "%B %d, %Y"

netflix_data['date_added']=pd.to_datetime(netflix_data['date_added'],format=date_format)

netflix_data.head()
```

	show_id	type	title	director	main_actor	country	date_added	release
0	1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Unknown	United States	2021-09-25	
1	2	TV Show	Blood & Water	Unspecified	Ama Qamata	South Africa	2021-09-24	
1	2	TV Show	Blood & Water	Unspecified	Ama Qamata	South Africa	2021-09-24	

Next steps: Generate code with netflix_data View recommended plots

```
# Adding columns Year_added and month_added based on the date_added column

netflix_data['year_added']=netflix_data['date_added'].dt.year
netflix_data['month_added']=netflix_data['date_added'].dt.month
netflix_data.head()
```

	show_id	type	title	director	main_actor	country	date_added	release
0	1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Unknown	United States	2021-09-25	
1	2	TV Show	Blood & Water	Unspecified	Ama Qamata	South Africa	2021-09-24	
1	2	TV Show	Blood & Water	Unspecified	Ama Qamata	South Africa	2021-09-24	
1	2	TV Show	Blood & Water	Unspecified	Ama Qamata	South Africa	2021-09-24	
2	3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila	United States	2021-09-24	

Next steps: Generate code with netflix_data View recommended plots

Exploratory Analysis and Visualization

** Total no of titles uploaded on Netflix**

```
netflix_original['show_id'].count()
```

```
8807
```

Observation: Out Dataset has a total of 8807 shows uploaded

Start coding or generate with AI.

Percentage of show types uploaded on Netflix

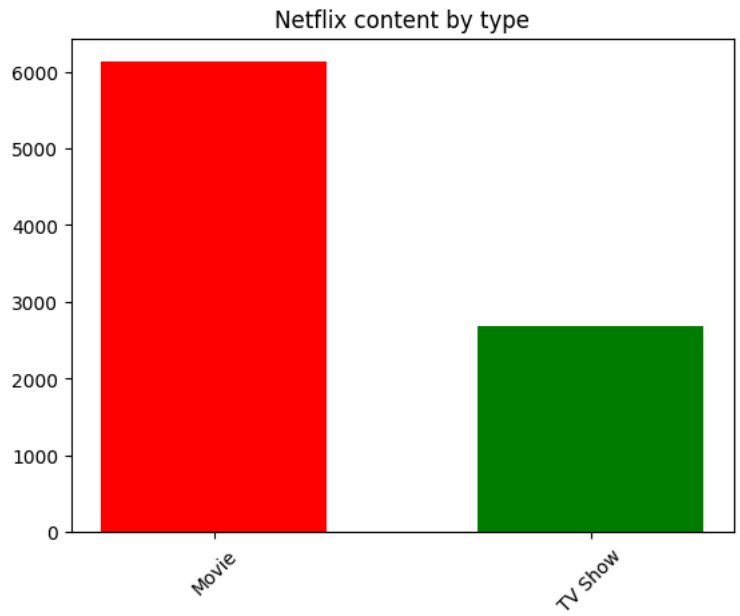
```
type_count=netflix_original['type'].value_counts().reset_index()
```

```
type_count.columns=['Type','Number of Titles']
type_count
```

	Type	Number of Titles
0	Movie	6131
1	TV Show	2676

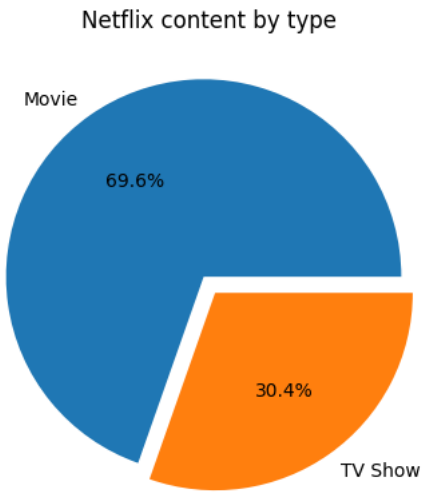
Next steps: Generate code with type_count View recommended plots

```
x_bar=type_count['Type']
y_bar=type_count['Number of Titles']

plt.bar(x_bar,y_bar,color=['r','g'],width=0.6)
plt.xticks(rotation=45, fontsize=10)
plt.title("Netflix content by type")
plt.ylabel='Number of Titles'
plt.show()
```



Start coding or generate with AI.

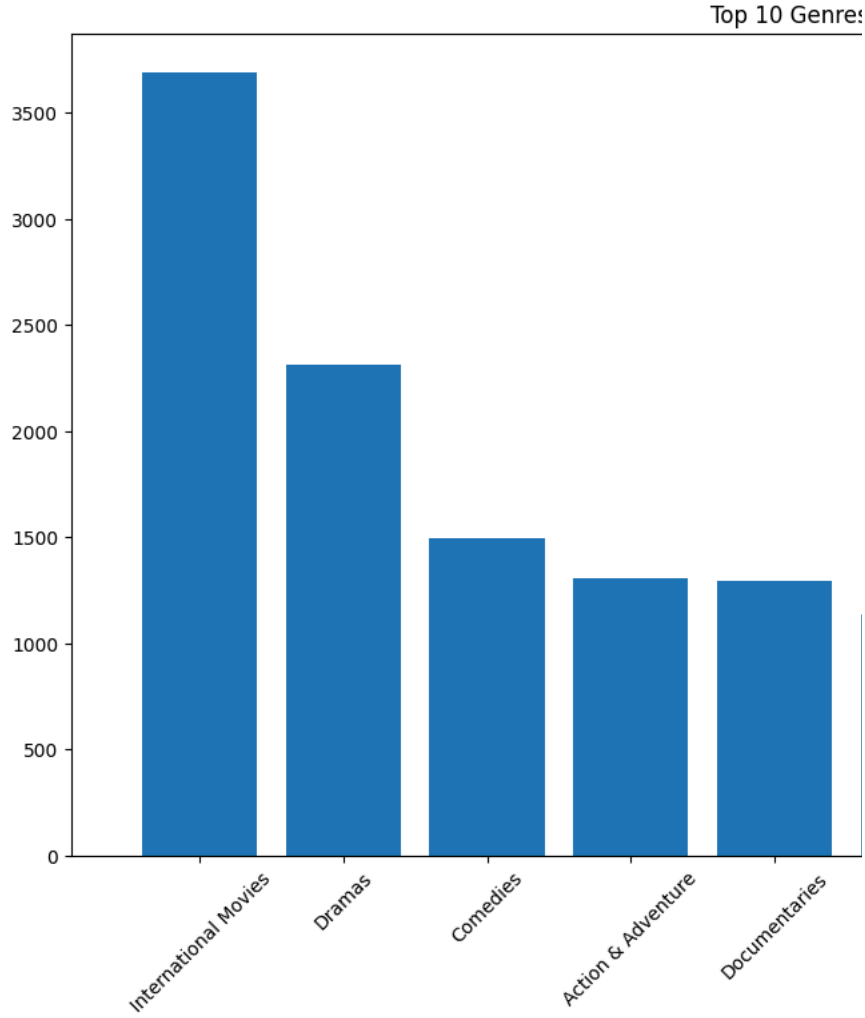
```
plt.title('Netflix content by type')
plt.pie(netflix_original['type'].value_counts(),labels=netflix_original['type'].value_counts().index,explode=(0.05,0.05),au
plt.show()
```




```
genre_counts=netflix_data['genre'].value_counts()
genre_counts
```

```
genre
 International Movies        3689
Dramas                       2313
Comedies                     1494
Action & Adventure           1305
Documentaries                1292
                             ...
Spanish-Language TV Shows       3
Romantic Movies                 3
LGBTQ Movies                    1
TV Sci-Fi & Fantasy             1
Sports Movies                   1
Name: count, Length: 73, dtype: int64
```

```
plt.figure(figsize=(15,8))
x_bar=genre_counts.index[:10]
y_bar=genre_counts[:10]
plt.bar(x_bar,y_bar)
plt.xticks(rotation=45, fontsize=10)
plt.title("Top 10 Genres on Netflix")
plt.show()
```




```
netflix_original.head()
```

	show_id	type	title	director	cast	country	date_added	release_y
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Unknown	United States	September 25, 2021	2
1	s2	TV Show	Blood & Water	Unspecified	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021	2
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi	United States	September 24, 2021	2

Next steps: Generate code with `netflix_original` View recommended plots

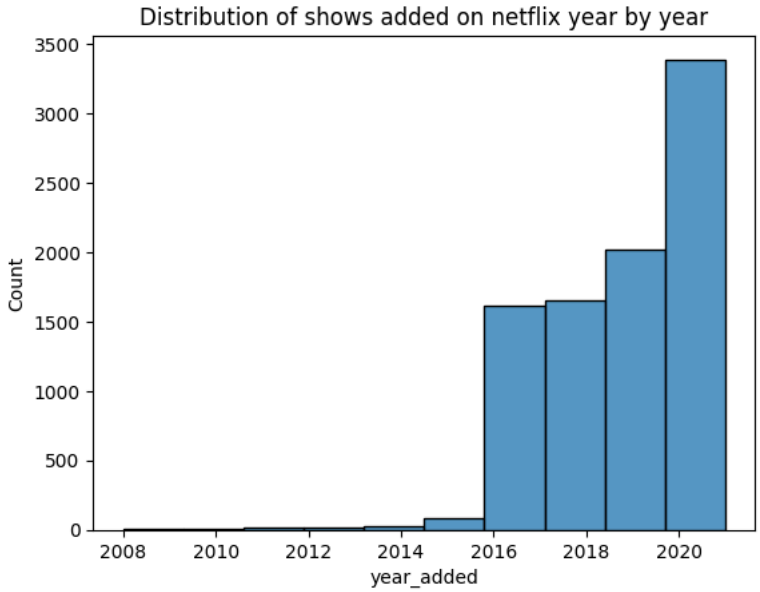
```
netflix_original['date_added']=netflix_original['date_added'].str.strip()
netflix_original['date_added']=pd.to_datetime(netflix_original['date_added'],format=date_format)
netflix_original['year_added']=netflix_original['date_added'].dt.year
netflix_original['month_added']=netflix_original['date_added'].dt.month
```

```
netflix_original['year_added'].value_counts()
```

```
year_added
2019    2016
2020    1889
2018    1649
2021    1498
```

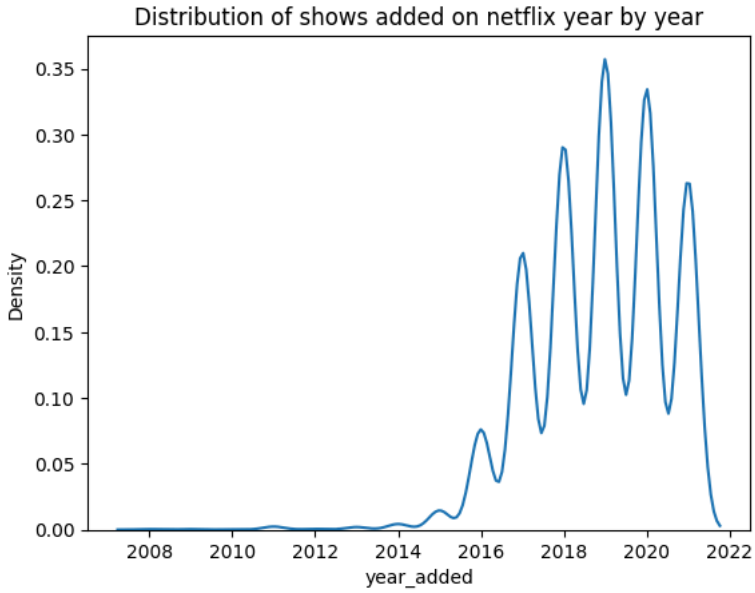
```
2017    1188
2016     429
2015      82
2014      24
2011      13
2013      11
2012       3
2009       2
2008       2
2010       1
Name: count, dtype: int64
```

```
sns.histplot(netflix_original['year_added'], bins=10)
plt.title("Distribution of shows added on netflix year by year")
plt.show()
```



Observation: The curve seems to be left skewed , indicating shows were added more from the year 2016 due to popularity of OTT platforms on internet.

```
# KDE plot to show more interpretable data
sns.kdeplot(netflix_original['year_added'])
plt.title("Distribution of shows added on netflix year by year")
plt.show()
```




```
netflix_data.head()
netflix_data['country']=netflix_data['country'].str.strip() #stipping off spaces in country column
```

```
country_data= netflix_data.set_index('country')
country_data.head()
```

country	show_id	type	title	director	main_actor	date_added	release_y
United States	1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Unknown	2021-09-25	2
South Africa	2	TV Show	Blood & Water	Unspecified	Ama Qamata	2021-09-24	2
South Africa	2	TV Show	Blood & Water	Unspecified	Ama Qamata	2021-09-24	2
South Africa	2	TV Show	Blood & Water	Unspecified	Ama Qamata	2021-09-24	2
United States	3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila	2021-09-24	2

Next steps: Generate code with country_data View recommended plots

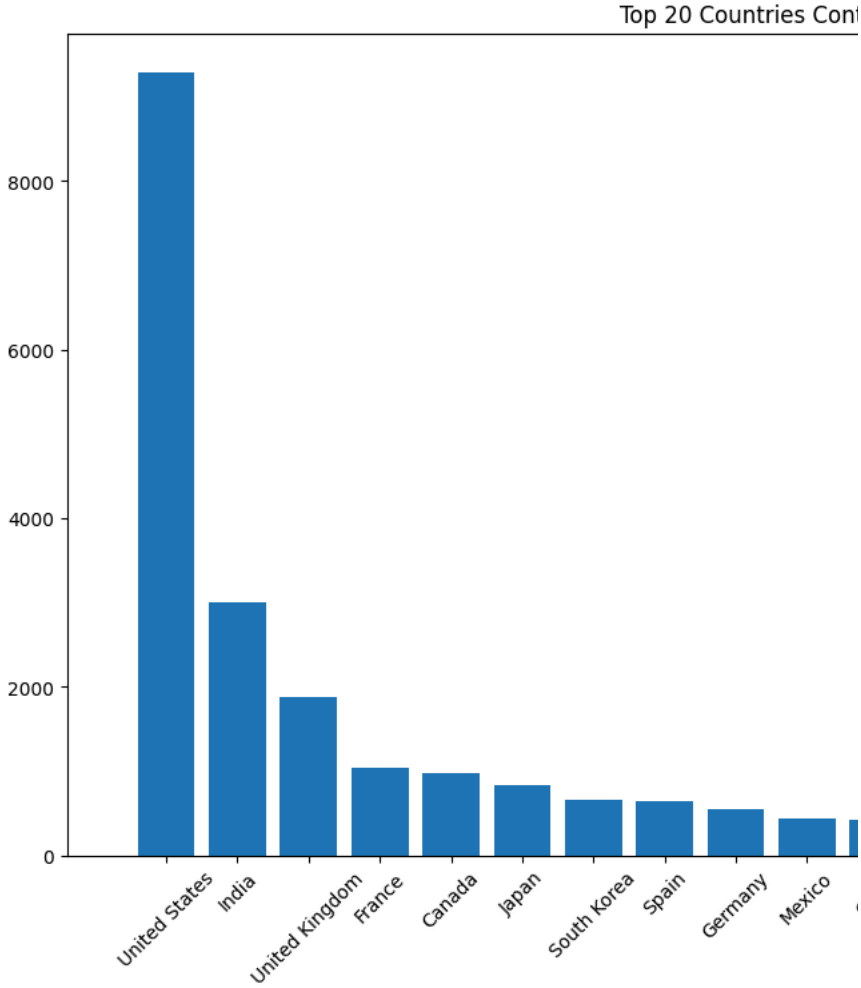
```
country_data=netflix_data.groupby('country').aggregate({'title':'count'}).sort_values(by='title',ascending=False)

country_data.reset_index(inplace=True)
country_data
```

	country	title
0	United States	9289
1	India	2999
2	United Kingdom	1880
3	France	1042
4	Canada	980
...	...	...
118	Bahamas	1
119	Uganda	1
120	Somalia	1
121	Nicaragua	1
122	Sudan	1

123 rows × 2 columns

Next steps: Generate code with country_data View recommended plots

```
plt.figure(figsize=(15,8))
x_bar=country_data['country'][:20]
y_bar=country_data['title'][:20]
plt.bar(x_bar,y_bar)
plt.xticks(rotation=45, fontsize=10)
plt.title("Top 20 Countries Contributors on Netflix")
plt.show()
```




```
year_count=netflix_original.groupby('year_added').aggregate({'title':'count'}).sort_values(by='title',ascending=False)
year_count.reset_index(inplace=True)
year_count
```

	year_added	title
0	2019	2016
1	2020	1889
2	2018	1649
3	2021	1498
4	2017	1188
5	2016	429
6	2015	82
7	2014	24
8	2011	13
9	2013	11
10	2012	3
11	2008	2
12	2009	2
13	2010	1

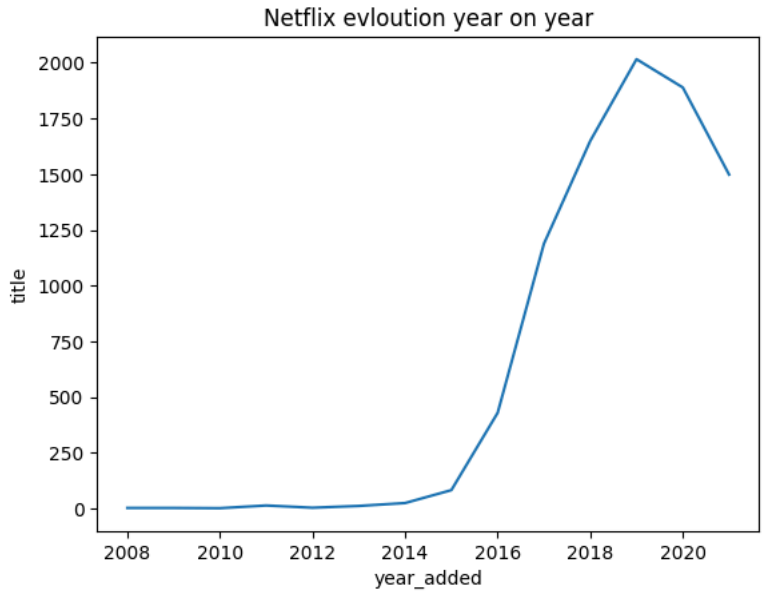
Next steps: Generate code with year_count View recommended plots

```
type_count=netflix_original.groupby(['year_added','type']).aggregate({'title':'count'})
type_count.reset_index(inplace=True)
type_count
```

	year_added	type	title
0	2008	Movie	1
1	2008	TV Show	1
2	2009	Movie	2
3	2010	Movie	1
4	2011	Movie	13
5	2012	Movie	3
6	2013	Movie	6
7	2013	TV Show	5
8	2014	Movie	19
9	2014	TV Show	5
10	2015	Movie	56
11	2015	TV Show	26
12	2016	Movie	253
13	2016	TV Show	176
14	2017	Movie	839
15	2017	TV Show	349
16	2018	Movie	1237
17	2018	TV Show	412
18	2019	Movie	1424
19	2019	TV Show	592
20	2020	Movie	1284
21	2020	TV Show	605
22	2021	Movie	993
23	2021	TV Show	505

Next steps: Generate code with type_count View recommended plots

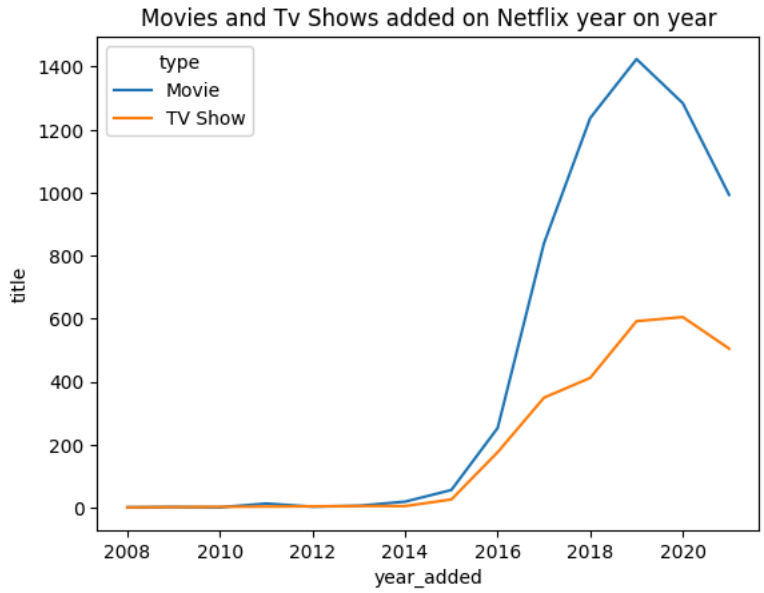
```
sns.lineplot(data=year_count,x=year_count['year_added'],y=year_count['title'])
plt.title("Netflix evloution year on year")
plt.show()
```



Average Movies vs TV Shows added per year

```
# @title Average Movies vs TV Shows added per year

sns.lineplot(x = 'year_added', y = 'title', hue = 'type', data = type_count)
plt.title('Movies and Tv Shows added on Netflix year on year')
```

```
Text(0.5, 1.0, 'Movies and Tv Shows added on Netflix year on year')
```

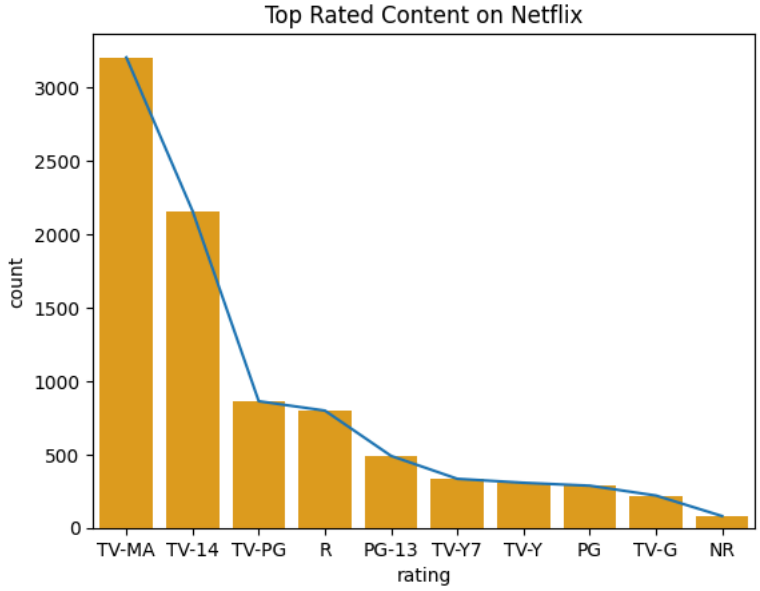



```
rating_count=netflix_original['rating'].value_counts()
rating_count=pd.DataFrame(rating_count)
rating_count.reset_index()
top_rated=rating_count.iloc[:10]
top_rated.reset_index(inplace=True)
top_rated
```

	rating	count
0	TV-MA	3207
1	TV-14	2160
2	TV-PG	863
3	R	799
4	PG-13	490
5	TV-Y7	334
6	TV-Y	307
7	PG	287
8	TV-G	220
9	NR	80

Next steps: Generate code with top_rated View recommended plots

```
sns.lineplot(x = 'rating', y = 'count',  data = top_rated)
sns.barplot(data=top_rated, x="rating", y="count",color='orange')
plt.title('Top Rated Content on Netflix')
plt.show()
```

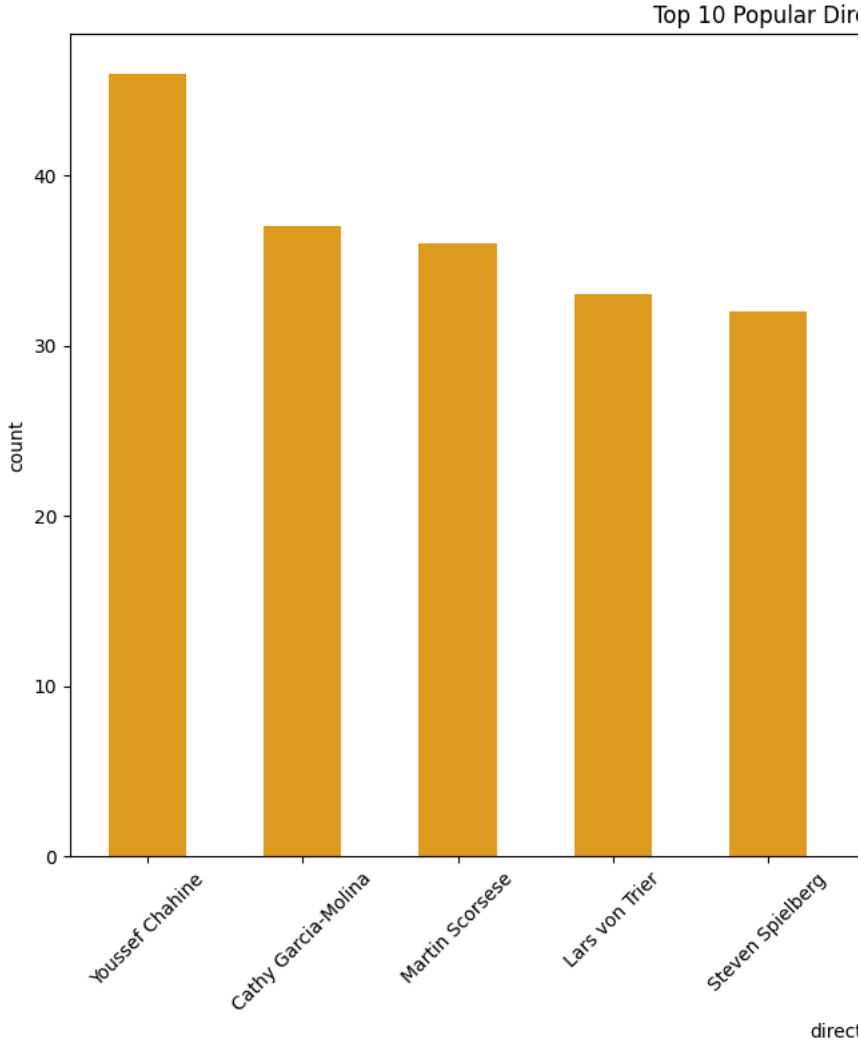



```
#Show top 10 director, who gave the highest number of TV shows & Movies to Netflix?
top_directors=netflix_data[netflix_data['director']!="Unspecified"]['director'].value_counts().reset_index()[:10]
top_directors
```

	director	count
0	Youssef Chahine	46
1	Cathy Garcia-Molina	37
2	Martin Scorsese	36
3	Lars von Trier	33
4	Steven Spielberg	32
5	Olivier Assayas	30
6	Tom Hooper	30
7	Suhas Kadav	29
8	Don Michael Paul	29
9	Johnnie To	28

Next steps: Generate code with `top_directors` View recommended plots

```
plt.figure(figsize=(15,8))
sns.barplot(data=top_directors, x="director", y="count",color='orange',width=0.5)
plt.width=0.5
plt.xlabel='Director'
plt.ylabel='Number of Shows'
plt.xticks(rotation=45, fontsize=10)
plt.title('Top 10 Popular Directors on Netflix')
plt.show()
```

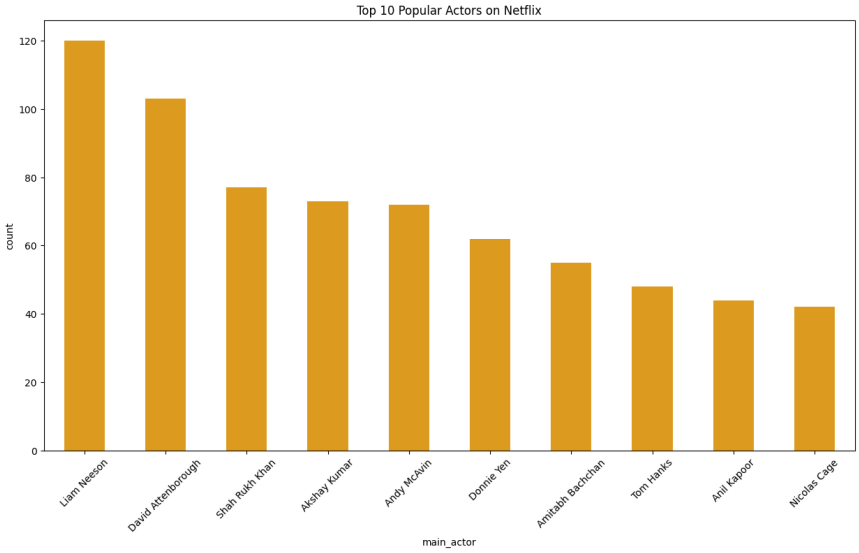



```
top_actors=netflix_data[netflix_data['main_actor']!="Unknown"]['main_actor'].value_counts().reset_index()[:10]
top_actors
```

	main_actor	count
0	Liam Neeson	120
1	David Attenborough	103
2	Shah Rukh Khan	77
3	Akshay Kumar	73
4	Andy McAvin	72
5	Donnie Yen	62
6	Amitabh Bachchan	55
7	Tom Hanks	48
8	Anil Kapoor	44
9	Nicolas Cage	42

Next steps: Generate code with top_actors View recommended plots

```
plt.figure(figsize=(15,8))
sns.barplot(data=top_actors, x="main_actor", y="count",color='orange',width=0.5)
plt.width=0.5
plt.xlabel='Actor'
plt.ylabel='Number of Shows'
plt.title('Top 10 Popular Actors on Netflix')
plt.xticks(rotation=45, fontsize=10)
plt.show()
```




```
netflix_data.dtypes
```

```
show_id                 object
type                    object
title                   object
director                object
main_actor              object
country                 object
date_added      datetime64[ns]
release_year            object
rating                  object
duration                object
genre                   object
description             object
year_added               int32
month_added              int32
dtype: object
```

```
netflix_data['year_added']=netflix_data['year_added'].astype('int64')
netflix_data['month_added']=netflix_data['month_added'].astype('int64')
```

```
corr_data=netflix_data[['date_added','year_added','month_added']]
corr_data
```

	date_added	year_added	month_added
0	2021-09-25	2021	9
1	2021-09-24	2021	9
1	2021-09-24	2021	9
1	2021-09-24	2021	9
2	2021-09-24	2021	9
...	...	...	...
8805	2020-01-11	2020	1
8805	2020-01-11	2020	1
8806	2019-03-02	2019	3
8806	2019-03-02	2019	3
8806	2019-03-02	2019	3

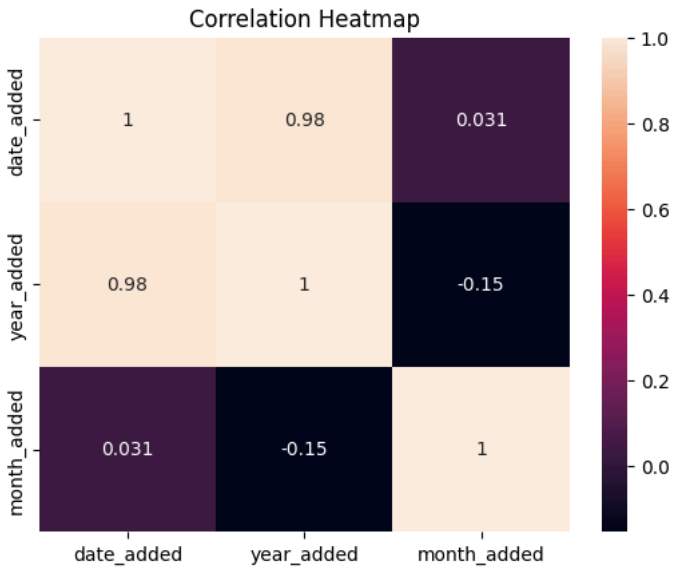
25900 rows × 3 columns

Next steps: Generate code with corr_data View recommended plots

```
corr_data.corr()
```

	date_added	year_added	month_added
date_added	1.000000	0.982798	0.030835
year_added	0.982798	1.000000	-0.153480
month_added	0.030835	-0.153480	1.000000

```
#plt.subplots(figsize=(5,5))
sns.heatmap(corr_data.corr(),annot=True)
plt.title("Correlation Heatmap")
plt.show()
```



Obeservation : Above Heatmap shows correlation between release_year,year_added & month_added.

Finding the best Month for releasing content?

```
# converting month number to month name
netflix_data['month_name'] = netflix_data['month_added'].replace({1:'Jan', 2:'Feb', 3:'Mar', 4:'Apr', 5:'May', 6:'June', 7:
netflix_data.head(2)
```

	show_id	type	title	director	main_actor	country	date_added	release_y
0	1	Movie	Dick Johnson Is Dead	Kirsten Johnson	Unknown	United States	2021-09-25	
1	2	TV Show	Blood & Water	Unspecified	Ama Qamata	South Africa	2021-09-24	

Next steps: Generate code with netflix_data View recommended plots

Generate Using dataframe: netflix_data suggest a plot Close

Generate is available for a limited time for unsubscribed users. Upgrade to Colab Pro

Start coding or generate with AI.

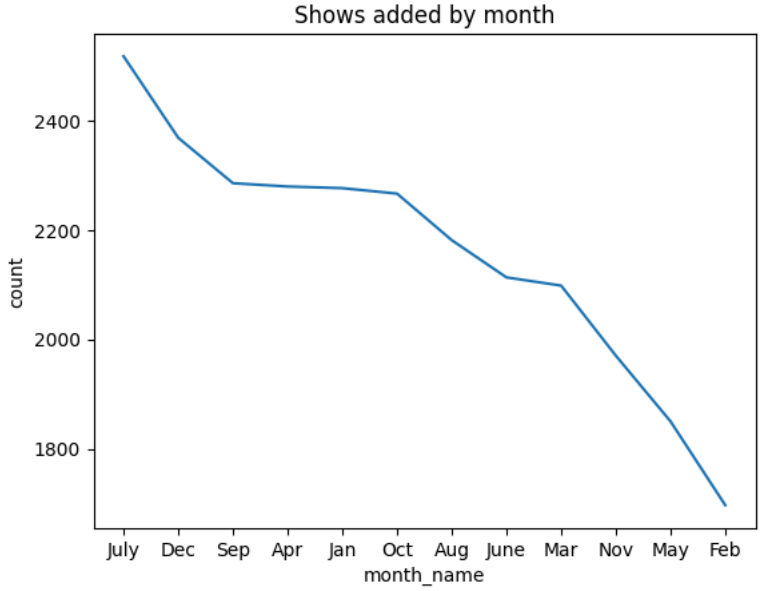
```
month_data=netflix_data['month_name'].value_counts().reset_index()
month_data
```

	month_name	count
0	July	2517
1	Dec	2368
2	Sep	2285
3	Apr	2279
4	Jan	2276
5	Oct	2266
6	Aug	2181
7	June	2113
8	Mar	2098
9	Nov	1970
10	May	1850
11	Feb	1697

Next steps: Generate code with month_data View recommended plots

Count per month

```
# @title Count per month

sns.lineplot(data=month_data, x='month_name', y='count')
plt.title("Shows added by month")
plt.show()
```



Observation:

1. Maximum shows were added in the month of July followed by December,October,January, April
2. Content was added keeping holidays and festive seaons

Double-click (or enter) to edit

```
#Number of movies added on netflix after 2015
movie_data=netflix_original[(netflix_original['type'] == 'Movie' ) & (netflix_original['year_added'] >= 2015)].reset_index(
movie_data.head()
movie_data['type'].value_counts()
```

```
type
Movie    6086
Name: count, dtype: int64
```

Observation: Maximum number of movies were added on Netflix from the year 2015.

```
netflix_data['country'].str.strip().value_counts()[:10]
```

```
country
United States     9289
India             2999
United Kingdom    1880
France            1042
Canada             980
Japan              828
South Korea        652
Spain              642
Germany            553
Mexico             437
Name: count, dtype: int64
```

```
#country_wise movies
usa_data=netflix_data[netflix_data['country']=='United States'].reset_index()
usa_data=usa_data.groupby(['year_added']).aggregate({'title':'count'}).sort_values(by=['title'],ascending=False).reset_inde
usa_data
```

	year_added	title
0	2021	2114
1	2019	2113
2	2020	1907
3	2018	1464
4	2017	1066
5	2016	399
6	2015	129

Next steps: Generate code with usa_data View recommended plots

```
india_data=netflix_data[netflix_data['country']=='India'].reset_index()
india_data=india_data.groupby(['year_added']).aggregate({'title':'count'}).sort_values(by=['title'],ascending=False).reset_
india_data
```

	year_added	title
0	2018	999
1	2019	618
2	2020	560
3	2017	449
4	2021	309
5	2016	64

Next steps: Generate code with india_data View recommended plots

```
uk_data=netflix_data[netflix_data['country']=='United Kingdom'].reset_index()
uk_data=uk_data.groupby(['year_added']).aggregate({'title':'count'}).sort_values(by=['title'],ascending=False).reset_index(
uk_data
```

	year_added	title
0	2019	439
1	2020	356
2	2018	338
3	2017	326
4	2021	259
5	2016	125
6	2015	22
7	2011	6
8	2014	6
9	2013	3

Next steps: Generate code with uk_data View recommended plots

```
plt.subplot(1,1,1)
sns.lineplot(x = 'year_added', y = 'title',  data = usa_data,label='United States')
sns.lineplot(x = 'year_added', y = 'title',  data = india_data,label='India')
sns.lineplot(x = 'year_added', y = 'title',  data = uk_data,label='United Kingdom')
plt.title("Year on Year content for top 3 countries")
plt.legend()
plt.show()
```