

NETFLIX MOVIES AND  
TV SHOWS

# ANALYSIS

PRESENTED BY:

**MUHAMMAD  
TAHIR**

# netflix-movies-and-tv-shows

September 9, 2024

## 0.1 Import Libraries

```
[1]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
```

## 0.2 Load Data and get Some Information

```
[2]: data = pd.read_csv("netflix_titles.csv")
```

```
[3]: data.head()
```

```
[3]: show_id      type      title      director \
0      s1      Movie  Dick Johnson Is Dead  Kirsten Johnson
1      s2  TV Show      Blood & Water      NaN
2      s3  TV Show      Ganglands  Julien Leclercq
3      s4  TV Show  Jailbirds New Orleans      NaN
4      s5  TV Show      Kota Factory      NaN

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

      date_added  release_year  rating  duration \
0  September 25, 2021      2020  PG-13      90 min
1  September 24, 2021      2021  TV-MA  2 Seasons
2  September 24, 2021      2021  TV-MA  1 Season
3  September 24, 2021      2021  TV-MA  1 Season
4  September 24, 2021      2021  TV-MA  2 Seasons

      listed_in \
0      Documentaries
```

```

1 International TV Shows, TV Dramas, TV Mysteries
2 Crime TV Shows, International TV Shows, TV Act...
3 Docuseries, Reality TV
4 International TV Shows, Romantic TV Shows, TV ...

```

```

description
0 As her father nears the end of his life, filmm...
1 After crossing paths at a party, a Cape Town t...
2 To protect his family from a powerful drug lor...
3 Feuds, flirtations and toilet talk go down amo...
4 In a city of coaching centers known to train I...

```

```
[4]: data.tail()
```

```

[4]:      show_id      type      title      director \
8802   s8803    Movie    Zodiac    David Fincher
8803   s8804  TV Show  Zombie Dumb          NaN
8804   s8805    Movie  Zombieland  Ruben Fleischer
8805   s8806    Movie    Zoom    Peter Hewitt
8806   s8807    Movie    Zubaan    Mozez Singh

```

```

cast      country \
8802 Mark Ruffalo, Jake Gyllenhaal, Robert Downey J... United States
8803                                     NaN          NaN
8804 Jesse Eisenberg, Woody Harrelson, Emma Stone, ... United States
8805 Tim Allen, Courteney Cox, Chevy Chase, Kate Ma... United States
8806 Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan... India

```

```

date_added  release_year  rating  duration \
8802 November 20, 2019      2007     R    158 min
8803      July 1, 2019      2018  TV-Y7  2 Seasons
8804 November 1, 2019      2009     R     88 min
8805 January 11, 2020      2006    PG     88 min
8806 March 2, 2019      2015  TV-14    111 min

```

```

listed_in \
8802      Cult Movies, Dramas, Thrillers
8803      Kids' TV, Korean TV Shows, TV Comedies
8804      Comedies, Horror Movies
8805      Children & Family Movies, Comedies
8806 Dramas, International Movies, Music & Musicals

```

```

description
8802 A political cartoonist, a crime reporter and a...
8803 While living alone in a spooky town, a young g...
8804 Looking to survive in a world taken over by zo...
8805 Dragged from civilian life, a former superhero...

```

8806 A scrappy but poor boy worms his way into a ty...

```
[5]: data.shape
```

```
[5]: (8807, 12)
```

```
[6]: data.columns
```

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

```
[7]: data.dtypes
```

```
[7]: 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
```

```
[8]: 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
```

```
[9]: data.isnull().sum()
```

```
[9]: 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
```

### 0.3 Data Cleaning and Preprocessing

```
[10]: data.duplicated().sum()
```

```
[10]: 0
```

```
[11]: data = data.set_index("show_id")
```

```
[12]: data.head(2)
```

```
[12]:
```

	type	title	director	\
show_id				
s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	
s2	TV Show	Blood & Water	NaN	

	cast	country	\
show_id			
s1	NaN	United States	
s2	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	

	date_added	release_year	rating	duration	\
show_id					
s1	September 25, 2021	2020	PG-13	90 min	
s2	September 24, 2021	2021	TV-MA	2 Seasons	

	listed_in	\
show_id		
s1	Documentaries	

	International TV Shows, TV Dramas, TV Mysteries		
			description
show_id			
s1	As her father nears the end of his life, filmm...		
s2	After crossing paths at a party, a Cape Town t...		

```
[13]: data.isnull().sum()
```

```
[13]: 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
```

```
[14]: columns_to_replace = ["director", "cast", "country"]
      data[columns_to_replace] = data[columns_to_replace].replace(np.nan, "Unknown")
```

```
[15]: columns_to_drop = ["date_added", "rating", "duration"]
      data.dropna(subset=columns_to_drop, inplace=True)
```

```
[16]: data.isna().sum()
```

```
[16]: 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
```

```
[17]: # changing date type
      data["date_added"] = data["date_added"].astype("datetime64[ms]")
```

```
[18]: data.dtypes
```

```
[18]: type          object
      title         object
      director      object
      cast          object
      country       object
      date_added    datetime64[ms]
      release_year  int64
      rating        object
      duration      object
      listed_in     object
      description   object
      dtype: object
```

```
[19]: data.head(3)
```

```
[19]:
```

	type	title	director \
show_id			
s1	Movie	Dick Johnson Is Dead	Kirsten Johnson
s2	TV Show	Blood & Water	Unknown
s3	TV Show	Ganglands	Julien Leclercq

	cast	country \
show_id		
s1	Unknown	United States
s2	Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa
s3	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	Unknown

	date_added	release_year	rating	duration \
show_id				
s1	2021-09-25	2020	PG-13	90 min
s2	2021-09-24	2021	TV-MA	2 Seasons
s3	2021-09-24	2021	TV-MA	1 Season

	listed_in \
show_id	
s1	Documentaries
s2	International TV Shows, TV Dramas, TV Mysteries
s3	Crime TV Shows, International TV Shows, TV Act...

	description
show_id	
s1	As her father nears the end of his life, filmm...
s2	After crossing paths at a party, a Cape Town t...
s3	To protect his family from a powerful drug lor...

```
[20]: data['year_added'] = data['date_added'].dt.year
      print(data['year_added'])
```

```

show_id
s1      2021
s2      2021
s3      2021
s4      2021
s5      2021
...
s8803   2019
s8804   2019
s8805   2019
s8806   2020
s8807   2019
Name: year_added, Length: 8790, dtype: int32

```

```
[21]: data['month_added'] = data['date_added'].dt.month_name()
      print(data['month_added'])
```

```

show_id
s1      September
s2      September
s3      September
s4      September
s5      September
...
s8803   November
s8804       July
s8805   November
s8806    January
s8807     March
Name: month_added, Length: 8790, dtype: object

```

## 0.4 Data Analysis and Visualiation

```
[22]: data.head(3)
```

```
[22]:
```

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



	date_added	release_year	rating	duration	\
show_id					
s1	2021-09-25	2020	PG-13	90 min	
s2	2021-09-24	2021	TV-MA	2 Seasons	
s3	2021-09-24	2021	TV-MA	1 Season	

	listed_in	\
show_id		
s1	Documentaries	
s2	International TV Shows, TV Dramas, TV Mysteries	
s3	Crime TV Shows, International TV Shows, TV Act...	

	description	year_added	\
show_id			
s1	As her father nears the end of his life, filmm...	2021	
s2	After crossing paths at a party, a Cape Town t...	2021	
s3	To protect his family from a powerful drug lor...	2021	

	month_added
show_id	
s1	September
s2	September
s3	September

```
[23]: data.describe()
```

```
[23]:
```

	date_added	release_year	year_added
count	8790	8790.000000	8790.000000
mean	2019-05-17 21:44:01.638000	2014.183163	2018.873606
min	2008-01-01 00:00:00	1925.000000	2008.000000
25%	2018-04-06 00:00:00	2013.000000	2018.000000
50%	2019-07-03 00:00:00	2017.000000	2019.000000
75%	2020-08-19 18:00:00	2019.000000	2020.000000
max	2021-09-25 00:00:00	2021.000000	2021.000000
std	NaN	8.825466	1.573568

1. How has the distribution of different types of Netflix content (Movies vs. TV Shows) evolved over the years?

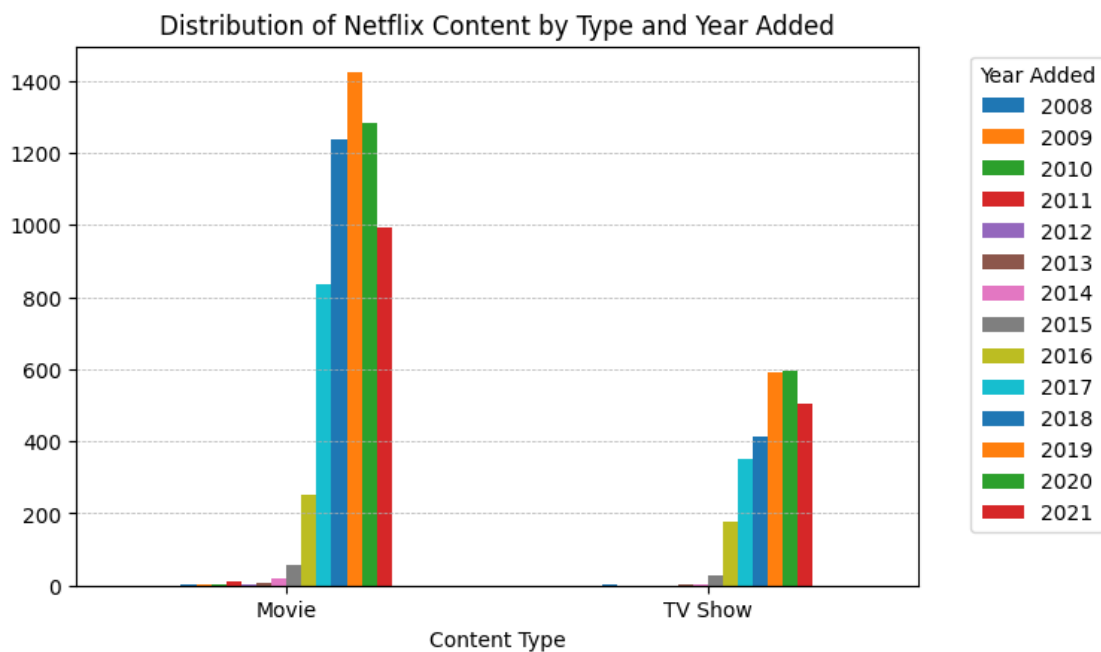
```
[24]: pd.crosstab(data.type, data.year_added)
```

```
[24]:
```

	year_added	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	\
type													
Movie		1	2	1	13	3	6	19	56	251	836	1237	
TV Show		1	0	0	0	0	5	5	26	175	349	411	
year_added		2019	2020	2021									

type			
Movie	1424	1284	993
TV Show	592	595	505

```
[25]: pd.crosstab(data.type, data.year_added).plot(kind="bar")
plt.tight_layout()
plt.title("Distribution of Netflix Content by Type and Year Added")
plt.xlabel("Content Type")
plt.legend(title='Year Added', bbox_to_anchor=(1.05, 1), loc='upper left')
plt.grid(True, which='both', axis='y', linestyle='--', linewidth=0.5)
plt.xticks(rotation=0)
plt.show()
```



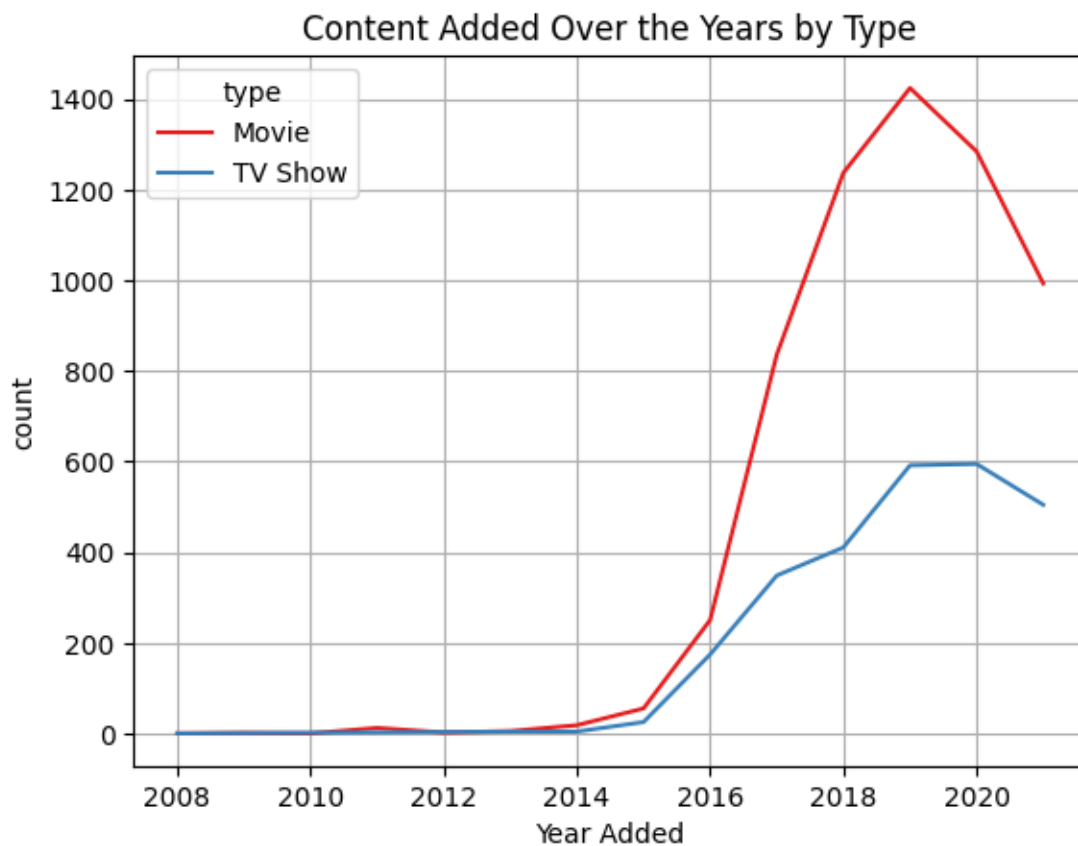
```
[26]: data.groupby(['year_added', 'type']).size().reset_index(name='count').
      ↪sort_values('count', ascending=False).head(10)
```

```
[26]:
```

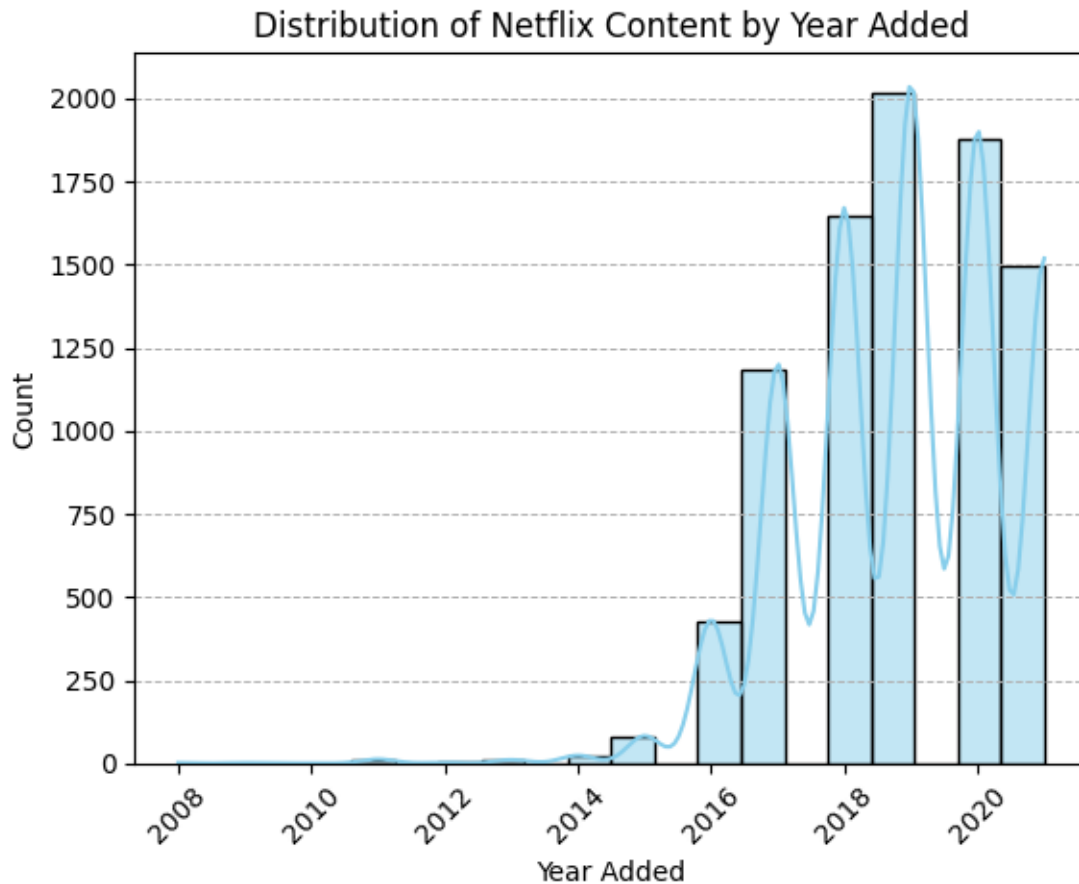
	year_added	type	count
18	2019	Movie	1424
20	2020	Movie	1284
16	2018	Movie	1237
22	2021	Movie	993
14	2017	Movie	836
21	2020	TV Show	595
19	2019	TV Show	592
23	2021	TV Show	505

17	2018	TV Show	411
15	2017	TV Show	349

```
[27]: year_type_count = data.groupby(['year_added', 'type']).size().
      ↪reset_index(name='count')
sns.lineplot(data=year_type_count, x='year_added', y='count', hue='type',
      ↪palette='Set1')
plt.title("Content Added Over the Years by Type")
plt.xlabel("Year Added")
plt.grid(True)
plt.show()
```



```
[28]: plt.tight_layout()
sns.histplot(data=data, x="year_added", bins=20, kde=True, color="skyblue")
plt.title("Distribution of Netflix Content by Year Added")
plt.xlabel("Year Added")
plt.xticks(rotation=45)
plt.grid(True, which='both', axis='y', linestyle='--', linewidth=0.7)
plt.show()
```



#### 0.4.1 2. What is the distribution of different content types on Netflix?

```
[29]: Movies = data['type'].value_counts()['Movie']
      tv_shows = data['type'].value_counts()['TV Show']
      print(f"Total Movies: {Movies}")
      print(f"Total TV Shows: {tv_shows}")
```

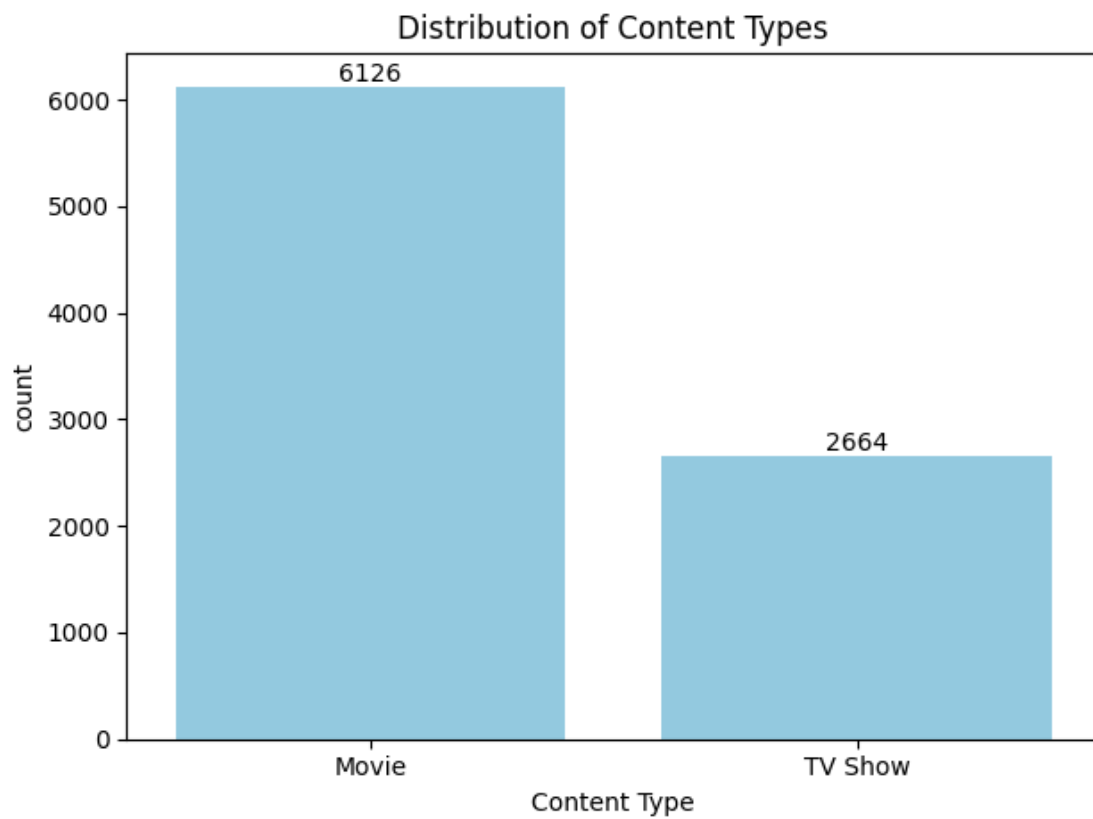
Total Movies: 6126  
Total TV Shows: 2664

```
[30]: sns.countplot(data=data, x="type", color='skyblue')
      plt.title("Distribution of Content Types")
      plt.xlabel("Content Type")

      for index, value in enumerate(data['type'].value_counts()):
          plt.text(index, value, str(value), ha='center', va='bottom')

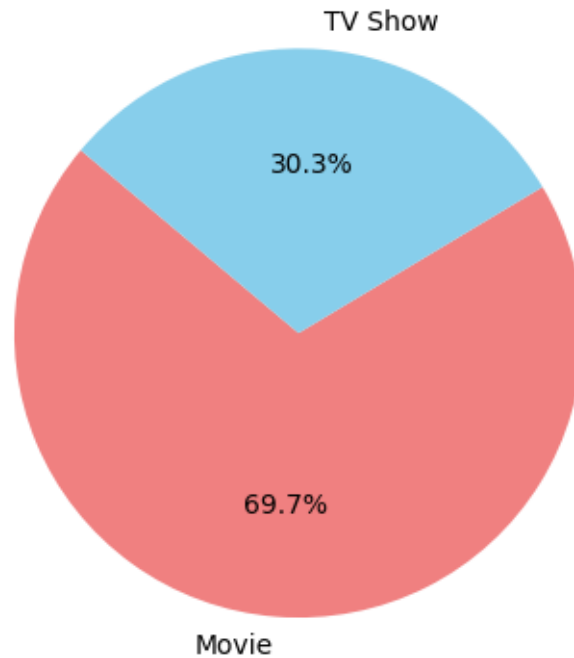
      plt.tight_layout()
```

```
plt.show()
```



```
[31]: plt.tight_layout()
data['type'].value_counts().plot(kind='pie', autopct='%1.1f%%', startangle=140,
    colors=['lightcoral', 'skyblue'])
plt.title('Percentage Distribution of Content Types')
plt.ylabel('')
plt.show()
```

### Percentage Distribution of Content Types



#### 0.4.2 3. Which countries have the most content available on Netflix?

```
[32]: top_10_countries = data['country'].replace("Unknown", pd.NA).dropna().  
      ↪ value_counts().head(10)  
      top_10_countries
```

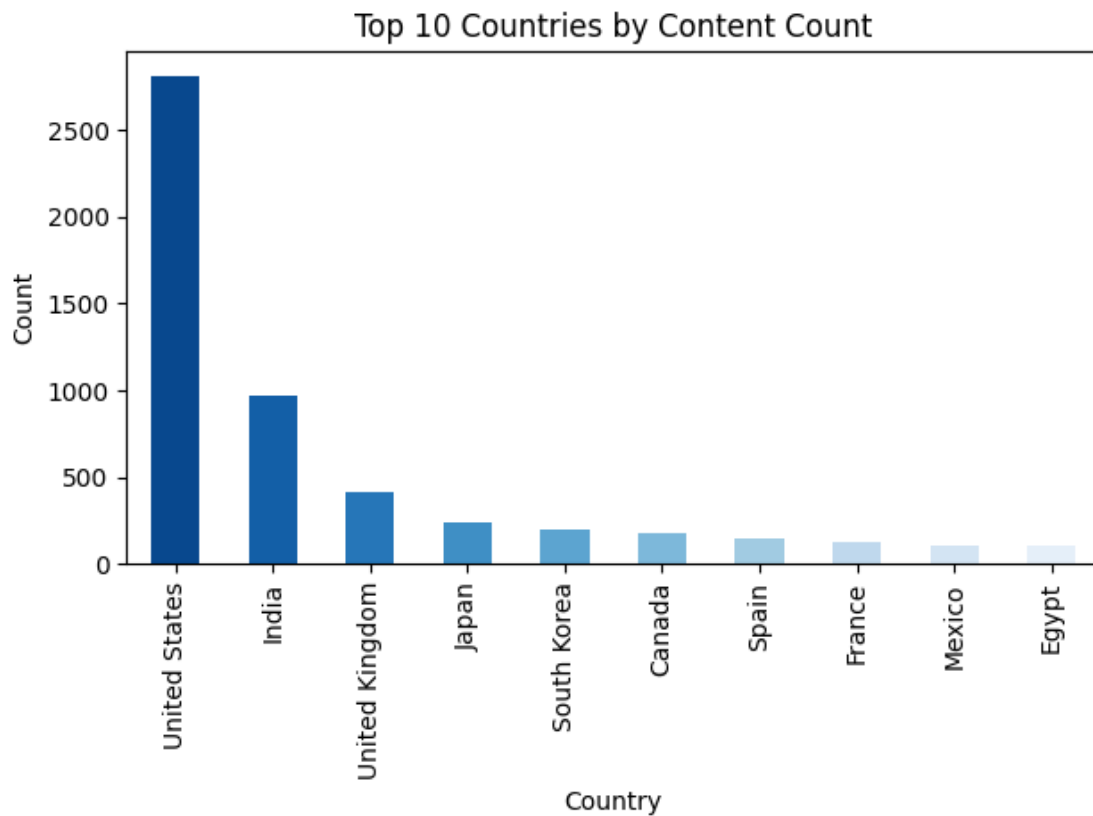
```
[32]: country  
      United States      2809  
      India              972  
      United Kingdom    418  
      Japan             243  
      South Korea       199  
      Canada            181  
      Spain             145  
      France            124  
      Mexico            110  
      Egypt             106  
      Name: count, dtype: int64
```

```
[33]: # Top 10 countries with most content  
      top_countries = data['country'].replace("Unknown", pd.NA).dropna().  
      ↪ value_counts().head(10)
```

```

colors = sns.color_palette("Blues", len(top_countries))[:-1]
top_countries.plot(kind='bar', color=colors)
plt.title("Top 10 Countries by Content Count")
plt.xlabel("Country")
plt.ylabel("Count")
plt.tight_layout()
plt.show()

```



```

[34]: temp = data.groupby(['year_added', 'country']).size().reset_index(name='count').
      ↪sort_values("count",ascending=False)
temp

```

```

[34]:
   year_added  country  count
717      2019  United States    677
952      2020  United States    624
491      2018  United States    451
1163     2021  United States    448
278      2017  United States    361
...         ...      ...      ...
625      2019  Italy, United States    1
222      2017  Netherlands, Belgium    1

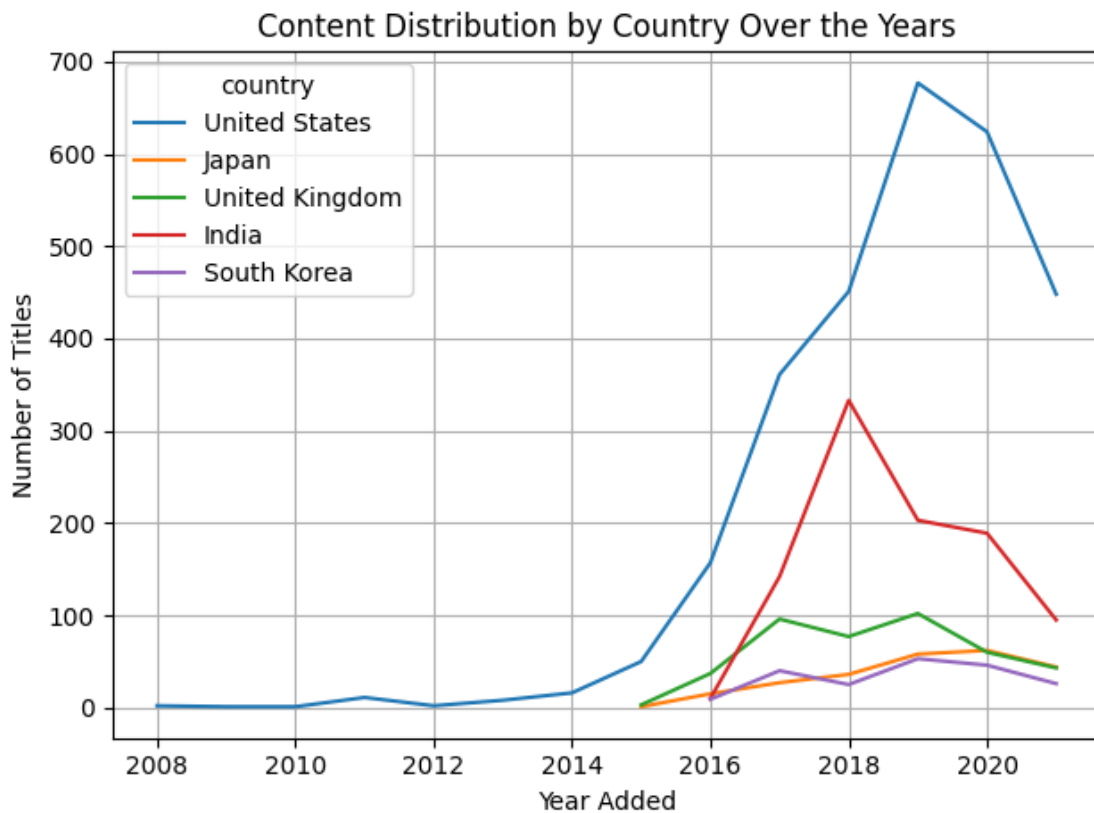
```

627	2019	Japan, United States	1
628	2019	Jordan	1
169	2017	Colombia, United States	1

[1206 rows x 3 columns]

```
[35]: country_year_trends = data.groupby(["year_added", "country"]).size().
      ↪reset_index(name='count')
top_countries = country_year_trends[country_year_trends['country'].
      ↪isin(['United States', 'India', 'United Kingdom', 'Japan', 'South Korea'])]

sns.lineplot(data=top_countries, x='year_added', y='count', hue='country')
plt.title("Content Distribution by Country Over the Years")
plt.xlabel("Year Added")
plt.ylabel("Number of Titles")
plt.grid(True)
plt.tight_layout()
plt.show()
```



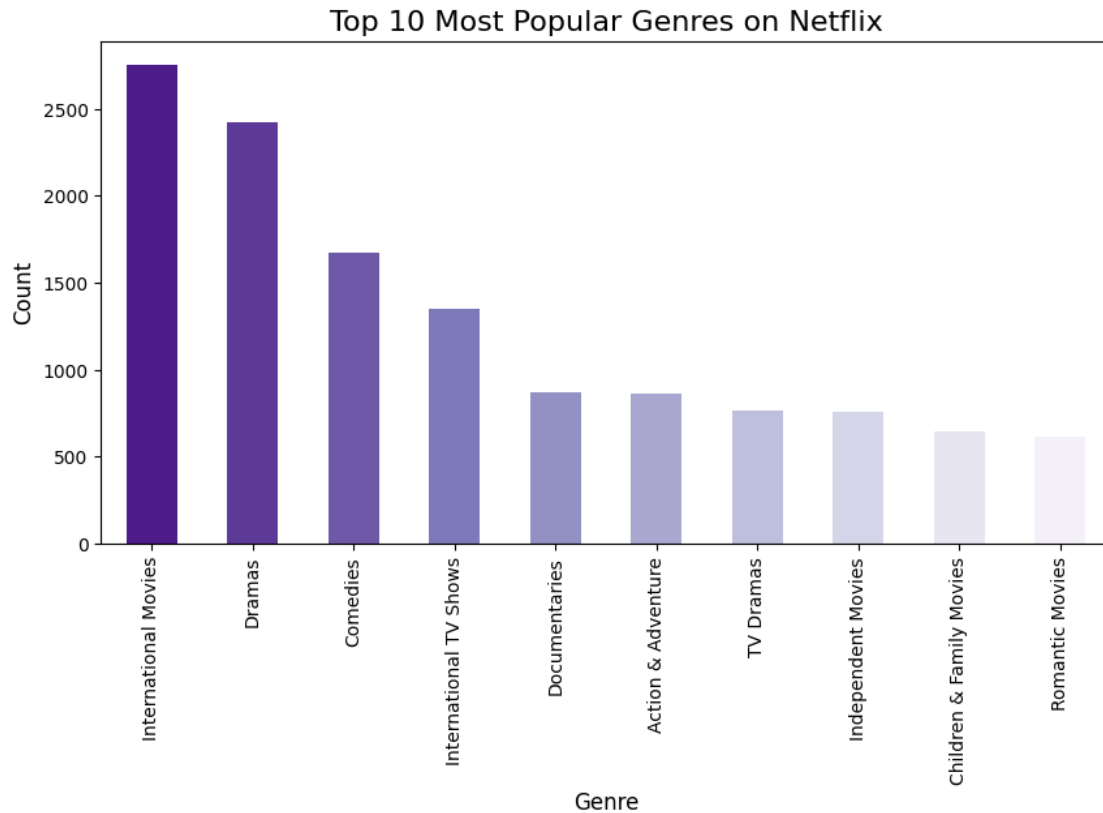


#### 0.4.3 4. What are the most popular genres available on Netflix?

```
[36]: data['listed_in'].str.split(', ').explode().value_counts().head(10)
```

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

```
[37]: # Top 10 most common genres
top_10_genres_count = data['listed_in'].str.split(', ').explode().
    ↪value_counts().head(10)
colors = sns.color_palette("Purples", len(top_10_genres_count))[::-1]
plt.figure(figsize=(10,5))
top_10_genres_count.plot(kind='bar', color=colors)
plt.title("Top 10 Most Popular Genres on Netflix", fontsize=16)
plt.xlabel("Genre", fontsize=12)
plt.ylabel("Count", fontsize=12)
plt.show()
```



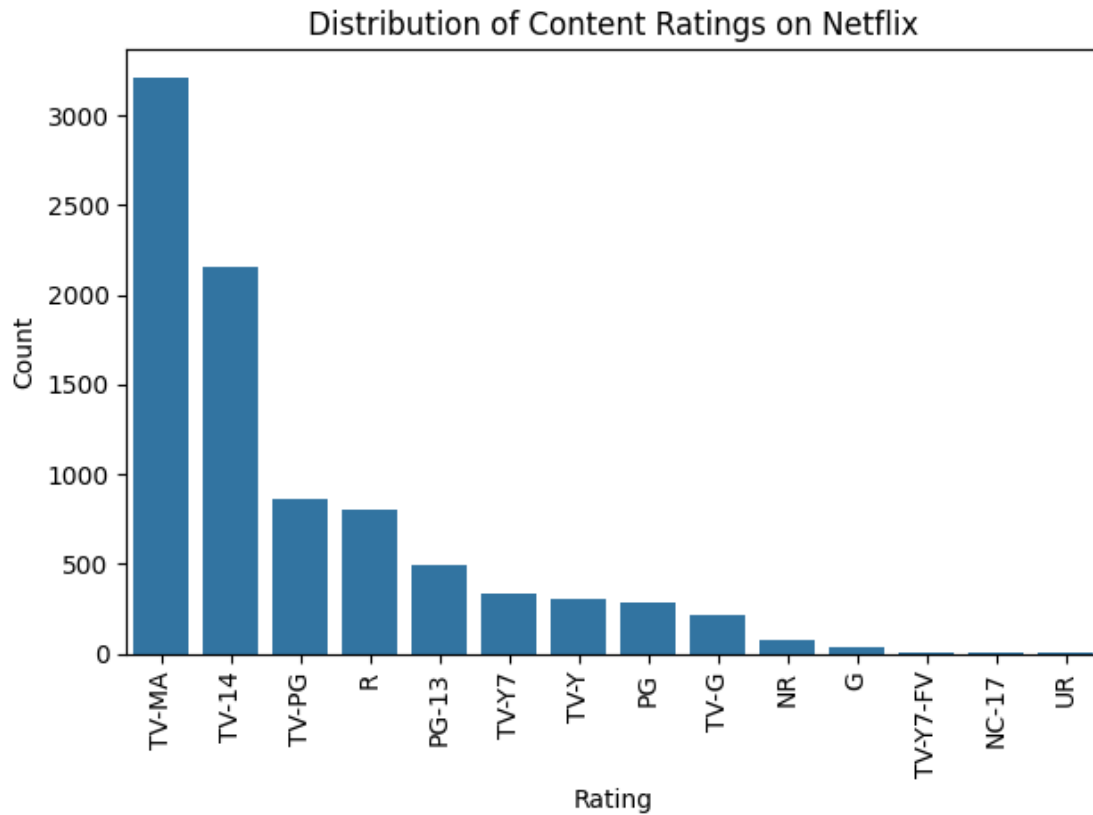
#### 0.4.4 5. How are content ratings distributed on Netflix?

```
[38]: # content ratings distributions  
data['rating'].value_counts().head(10)
```

```
[38]: rating  
TV-MA    3205  
TV-14    2157  
TV-PG    861  
R         799  
PG-13    490  
TV-Y7     333  
TV-Y      306  
PG        287  
TV-G      220  
NR         79  
Name: count, dtype: int64
```

```
[39]: ratings = data['rating'].value_counts().index  
sns.countplot(data=data, x="rating", order=ratings)  
plt.title("Distribution of Content Ratings on Netflix")
```

```
plt.xlabel("Rating")
plt.ylabel('Count')
plt.xticks(rotation=90)
plt.tight_layout()
plt.show()
```



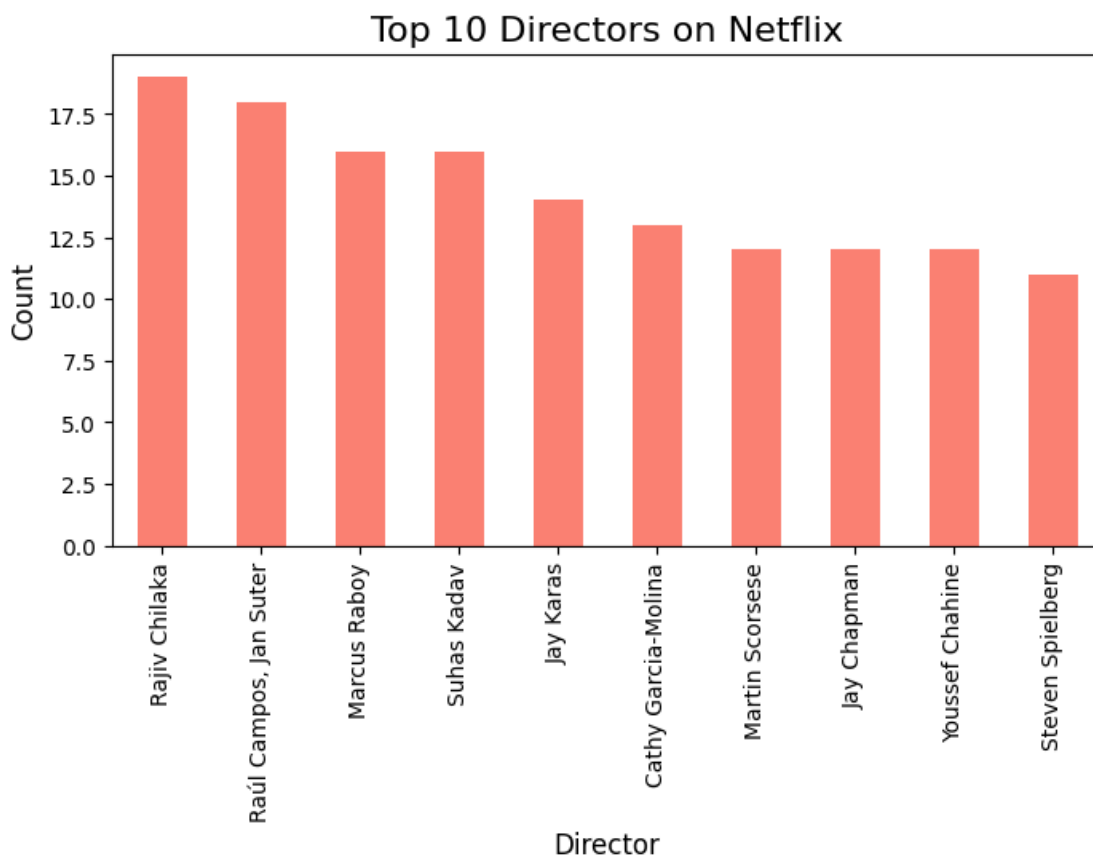
#### 0.4.5 6. Who are the top 10 directors with the most content on Netflix?

```
[40]: data['director'].replace("Unknown", pd.NA).dropna().value_counts().head(10)
```

```
[40]: director
Rajiv Chilaka          19
Raúl Campos, Jan Suter  18
Marcus Raboy           16
Suhas Kadav            16
Jay Karas              14
Cathy Garcia-Molina    13
Martin Scorsese        12
Jay Chapman            12
Youssef Chahine        12
```

Steven Spielberg 11  
Name: count, dtype: int64

```
[41]: # top 10 directors
plt.figure(figsize=(8,4))
top_directors = data['director'].replace("Unknown", pd.NA).dropna().
    ↪value_counts().head(10)
top_directors.plot(kind='bar', color='salmon')
plt.title("Top 10 Directors on Netflix", fontsize=16)
plt.xlabel("Director", fontsize=12)
plt.ylabel("Count", fontsize=12)
plt.show()
```



#### 0.4.6 7. Find average duration of movies?

```
[42]: movie_data = data[data['type'] == "Movie"]
avg_movie_duration = movie_data['duration'].str.replace(' min', '').astype(int).
    ↪mean()
print(f"Average Movie Duration (Minutes): {avg_movie_duration:.2f}")
```

Average Movie Duration (Minutes): 99.58

```
[43]: import plotly.graph_objects as go

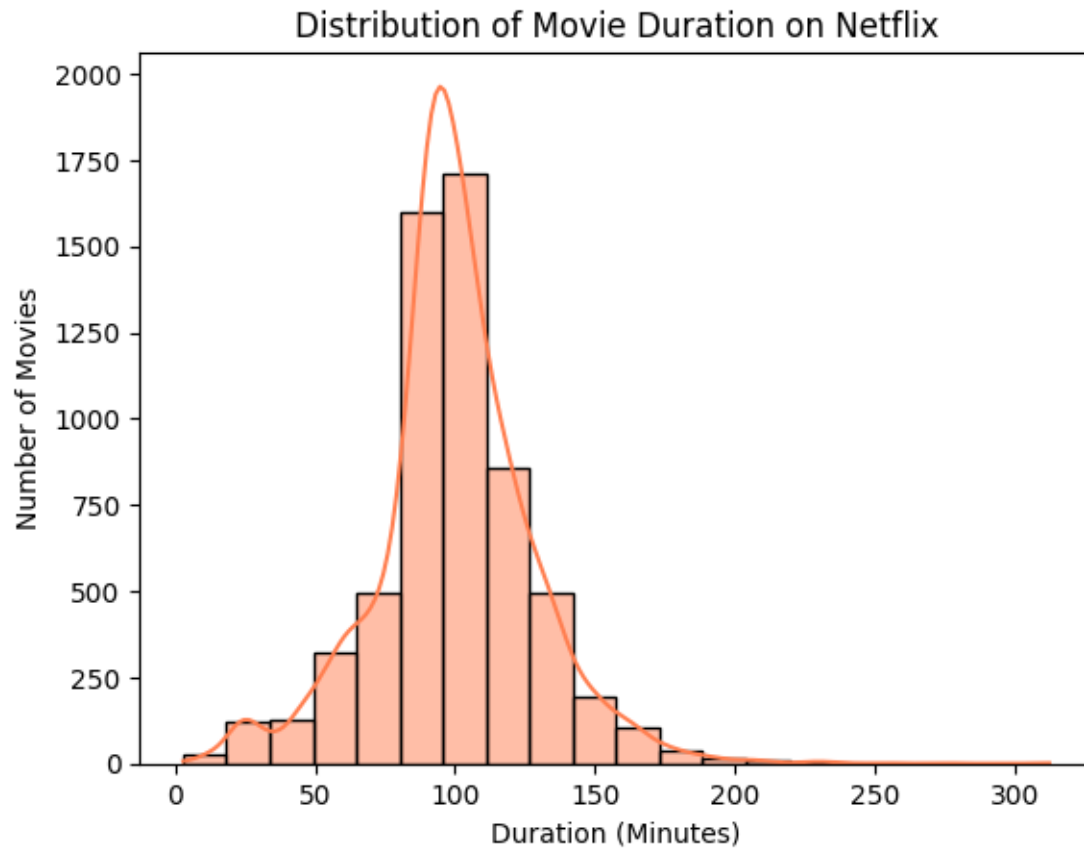
fig = go.Figure(go.Indicator(mode='number',
                             value=avg_movie_duration,
                             title={'text': "Average Movie Duration (Minutes)",
                                     "font":{"size":16, "color":"darkblue"}},
                             number={"font":{"size":40, "color":"green"}},
                             domain={'x':[0,1], 'y':[0,1]}))

fig.update_layout(
    paper_bgcolor='skyblue',
    height=200,
    width=300,
)
fig.show()
```

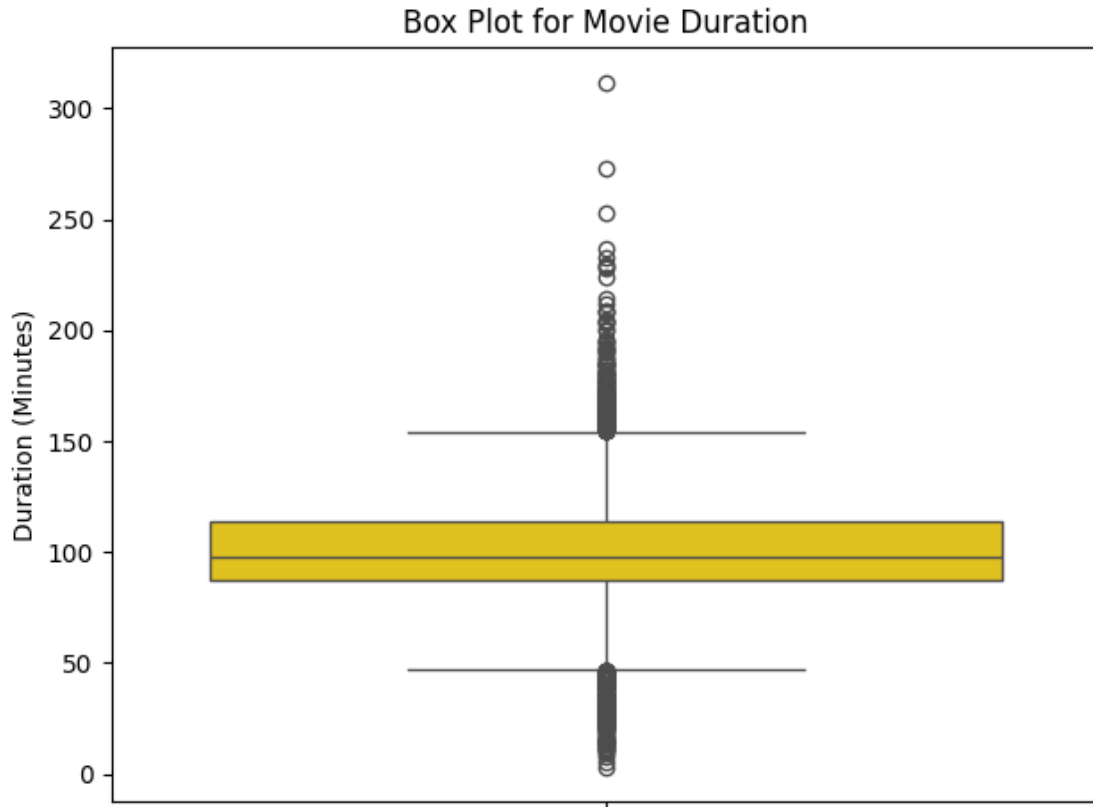


#### 0.4.7 8. How is the duration of movies distributed on Netflix?

```
[44]: # Histogram for movie duration
movie_duration = data[data['type']=='Movie']['duration'].str.replace(' min',
    ' ').astype(int)
sns.histplot(movie_duration, bins=20, kde=True, color='coral')
plt.title("Distribution of Movie Duration on Netflix")
plt.xlabel("Duration (Minutes)")
plt.ylabel("Number of Movies")
plt.show()
```



```
[45]: # Box plot for movie duration to identify outliers
sns.boxplot(y=movie_duration, color='Gold')
plt.title("Box Plot for Movie Duration")
plt.ylabel("Duration (Minutes)")
plt.tight_layout()
plt.show()
```

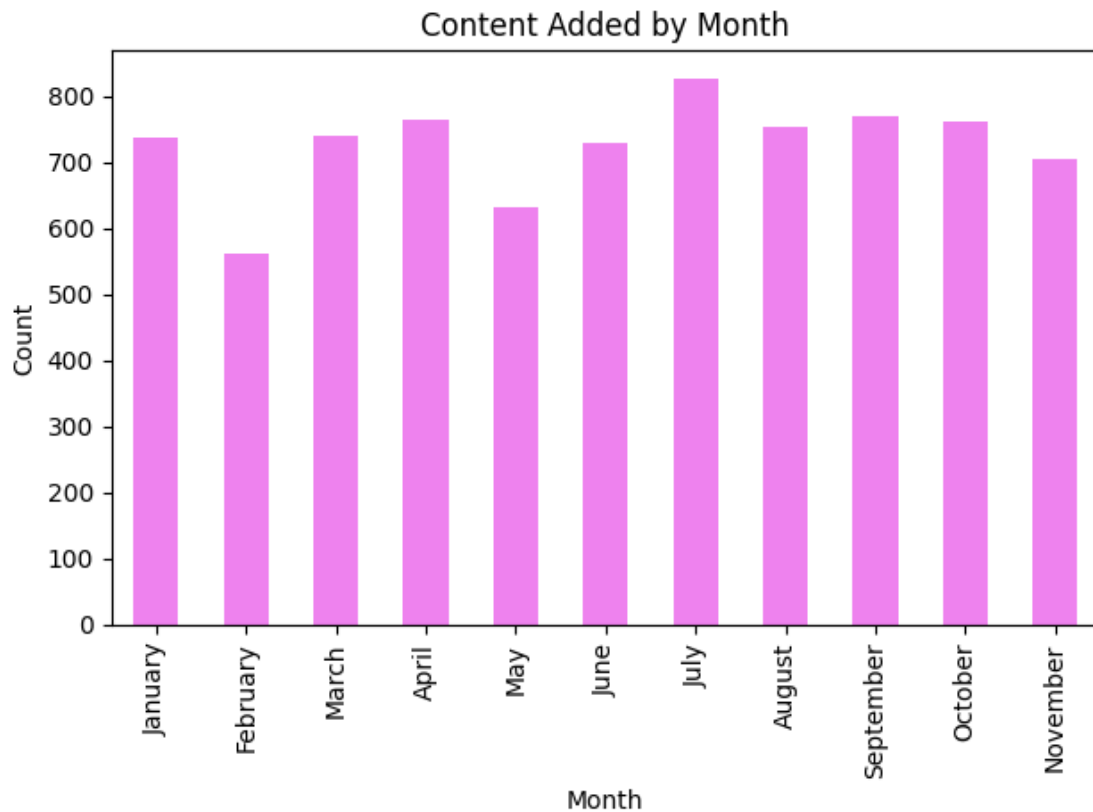


#### 0.4.8 9. How does the number of content additions vary by month throughout the year 2021?

```
[46]: data['month_added'].value_counts().reindex(pd.date_range('2021-01', '2021-12',  
freq='ME').strftime('%B'), fill_value=0)
```

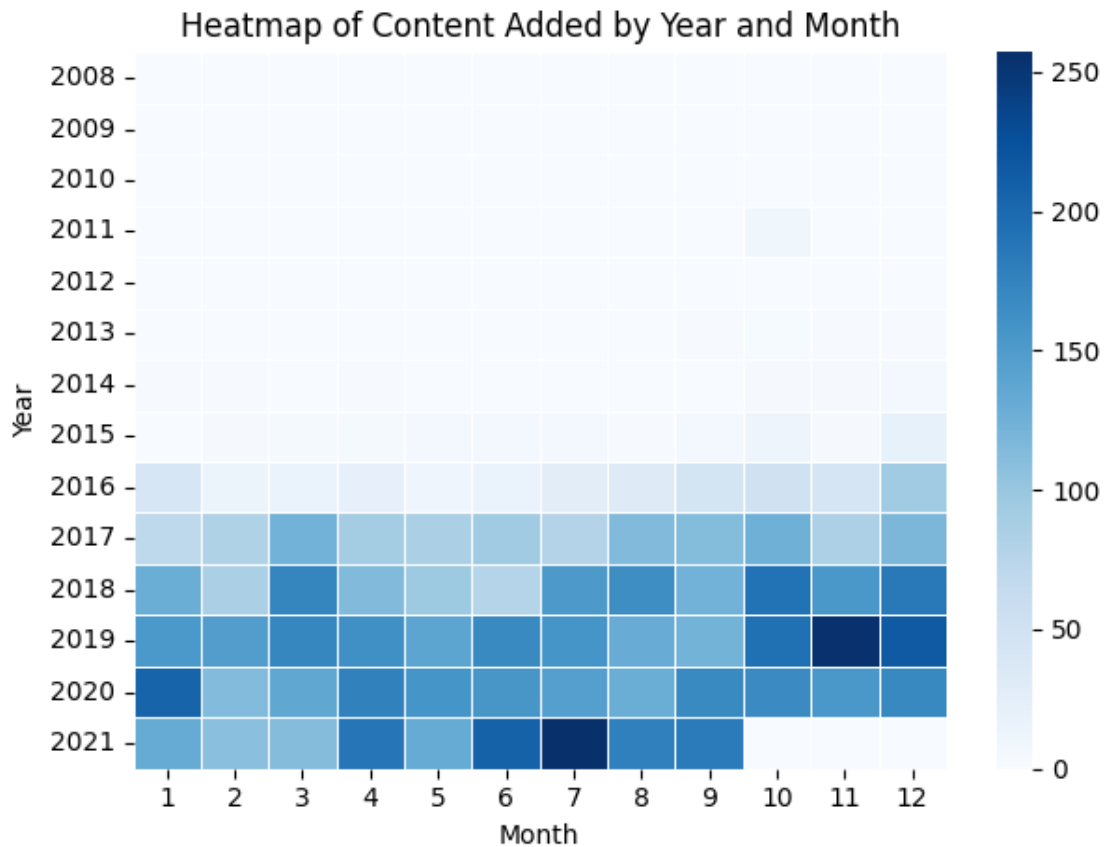
```
[46]: January      737
      February    562
      March       741
      April       763
      May        632
      June       728
      July       827
      August     754
      September  769
      October    760
      November   705
      Name: count, dtype: int64
```

```
[47]: # Content Added per month
month_count = data['month_added'].value_counts().reindex(pd.
    ↳date_range('2021-01', '2021-12', freq='ME').strftime('%B'), fill_value=0)
month_count.plot(kind='bar', color='violet')
plt.title("Content Added by Month")
plt.xlabel("Month")
plt.ylabel('Count')
plt.tight_layout()
plt.show()
```



```
[48]: # heatmap for title added by year and month
month_no = data['date_added'].dt.month
heatmap_data = pd.crosstab(data['year_added'], month_no)
sns.heatmap(data=heatmap_data, cmap='Blues', linewidths=.5)
plt.title("Heatmap of Content Added by Year and Month")
plt.xlabel("Month")
plt.ylabel("Year")
plt.tight_layout()
plt.show()
```



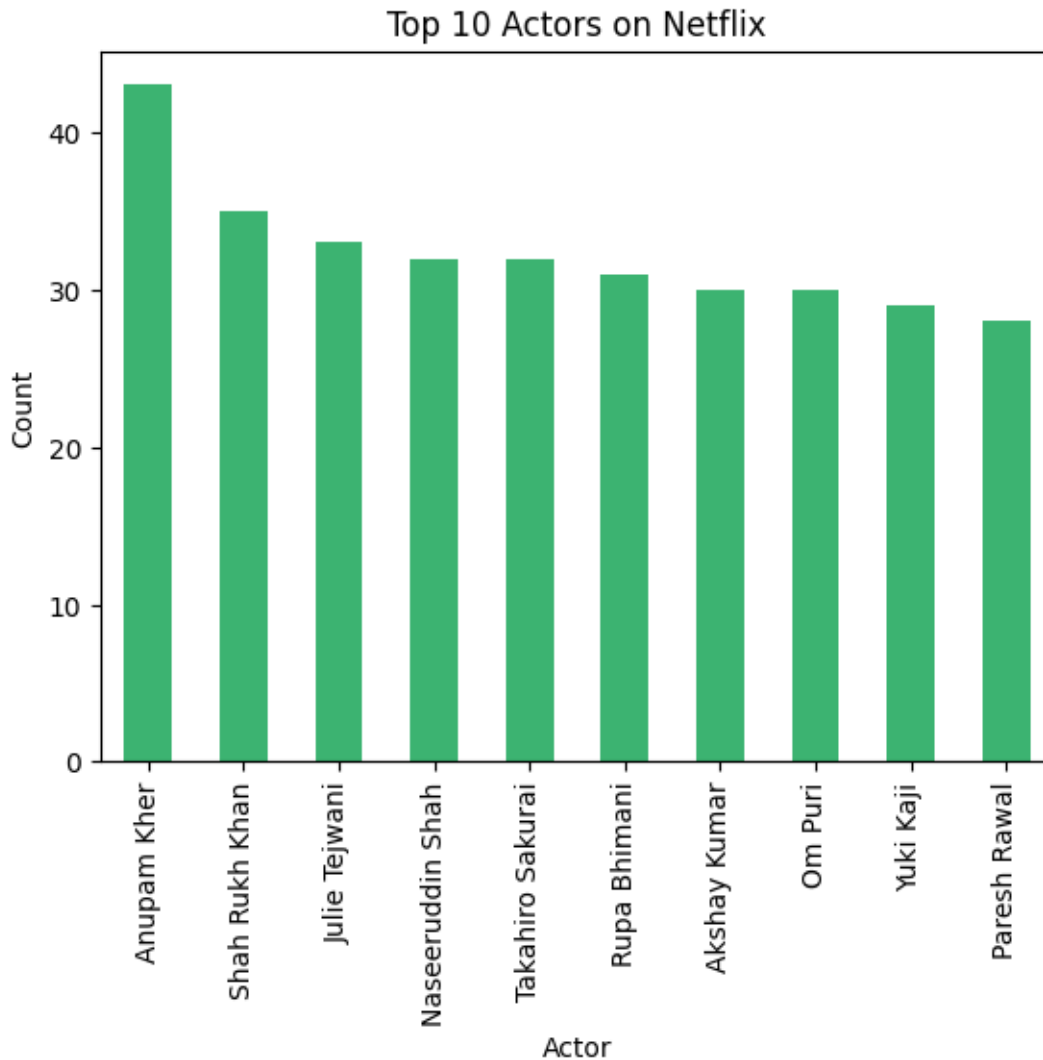


#### 0.4.9 10. Who are the top 10 most frequently appearing actors in the Netflix dataset?

```
[49]: top_10_actors = data['cast'].replace('Unknown', pd.NA).dropna().str.split(', ').
      ↪explode().value_counts().head(10)
top_10_actors
```

```
[49]: cast
Anupam Kher          43
Shah Rukh Khan       35
Julie Tejewani       33
Naseeruddin Shah     32
Takahiro Sakurai     32
Rupa Bhimani         31
Akshay Kumar         30
Om Puri              30
Yuki Kaji            29
Paresh Rawal         28
Name: count, dtype: int64
```

```
[50]: top_10_actors.plot(kind='bar', color="MediumSeaGreen")
plt.title("Top 10 Actors on Netflix")
plt.xlabel("Actor")
plt.ylabel("Count")
plt.show()
```



0.4.10 11. How has the number of Netflix titles released each year evolved between 1990 and 2021?

```
[51]: pd.crosstab(data.type, data.release_year)
```

```
[51]: release_year  1925  1942  1943  1944  1945  1946  1947  1954  1955  1956  ...  \
type
Movie              0     2     3     3     3     1     1     2     3     2  ...
```

TV Show	1	0	0	0	1	1	0	0	0	0	...
---------	---	---	---	---	---	---	---	---	---	---	-----

release_year	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
type										
Movie	173	225	264	396	658	765	767	633	517	277
TV Show	63	61	88	159	243	265	379	397	436	315

[2 rows x 74 columns]

```
[52]: filtered_data = data[(data['release_year'] >= 1990) & (data['release_year'] <= 2021)]
sns.countplot(data=filtered_data, x="release_year", hue="release_year", palette='viridis', legend=False)
plt.title("Content Count by Release Year (1990-2021)")
plt.xticks(rotation=90)
plt.tight_layout()
plt.show()
```

