Netflix Data Analysis and Visualization

Introduction

Netflix, Inc. (Netflix) is a retailer of movie rental services. The company offers an internet subscriptionservice for enjoying TV shows and movies, where subscribers can instantly watch unlimited TV showsand movies streamed over the Internet to their TVs, computers and mobile devices. It offers DVDs and Bluray titles to its subscribers. Netflix also operates a separate library of movies that can be watched instantly on subscriber's TV through a Netflix ready device or on their computers.

Table of Contents

- · Import Necessary Libraries
- Data Understanding
- · Data Preparation
- · Exploratory Data Analysis
- Conclusions

Import Necessary Libraries

Data Understanding

- · head and tail
- sample
- Dataframe shape
- size
- columns
- dtypes
- info
- describe

```
In [3]: 1 # preview the first 5 rows of dataset
2 df.head()
```

Out[3]:

	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
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi	NaN	September 24, 2021	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	September 24, 2021	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	September 24, 2021	2021	TV- MA	2 Seasons	International TV Shows, Romantic TV Shows, TV	In a city of coaching centers known to train I

	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
8804	s8805	Movie	Zombieland	Ruben Fleischer	Jesse Eisenberg, Woody Harrelson, Emma Stone,	United States	November 1, 2019	2009	R	88 min	Comedies, Horror Movies	Looking to survive in a world taken over by zo
8805	s8806	Movie	Zoom	Peter Hewitt	Tim Allen, Courteney Cox, Chevy Chase, Kate Ma	United States	January 11, 2020	2006	PG	88 min	Children & Family Movies, Comedies	Dragged from civilian life, a former superhero
8806	s8807	Movie	Zubaan	Mozez Singh	Vicky Kaushal, Sarah-Jane Dias, Raaghav Chanan	India	March 2, 2019	2015	TV-14	111 min	Dramas, International Movies, Music & Musicals	A scrappy but poor boy worms his way into a ty

In [5]: 1 # to see any random rows to ensure about the data quality

3 df.sample(7, random_state=12345)

Out[5]:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description
4632	s4633	TV Show	The Good Cop	NaN	Tony Danza, Josh Groban, Monica Barbaro, Isiah	United States	September 21, 2018	2018	TV-14	1 Season	Crime TV Shows, TV Comedies, TV Dramas	When he's not solving murders, a pathologicall
15	s16	TV Show	Dear White People	NaN	Logan Browning, Brandon P. Bell, DeRon Horton,	United States	September 22, 2021	2021	TV- MA	4 Seasons	TV Comedies, TV Dramas	Students of color navigate the daily slights a
3102	s3103	Movie	Sweetheart	J.D. Dillard	Kiersey Clemons, Emory Cohen, Hanna Mangan Law	United States	December 25, 2019	2019	PG- 13	83 min	Horror Movies, Sci-Fi & Fantasy, Thrillers	A woman shipwrecked on a remote island discove
8781	s8782	Movie	Yo-Kai Watch: The Movie	Mark Risley	Johnny Yong Bosch, Joey D'Auria, Alicyn Packar	Japan, United States	December 1, 2016	2016	TV-Y7	96 min	Anime Features, Children & Family Movies	Nate's special Yo- Kai Watch is missing. He tra
7829	s7830	Movie	Ramen Shop	Eric Khoo	Takumi Saitoh, Seiko Matsuda, Mark Lee, Jeanet	Singapore, Japan, France	July 31, 2019	2018	TV- PG	90 min	Dramas, International Movies	Ramen chef Masato travels to Singapore to retr
1770	s1771	Movie	Wheels of Fortune	Shaun Paul Piccinino	Matt Jones, Noureen DeWulf, John Ducey, Matty	United States	November 1, 2020	2020	R	107 min	Comedies, Sports Movies	To claim a big inheritance, a down-on-his-luck
8296	s8297	Movie	The Fighter	David O. Russell	Mark Wahlberg, Christian Bale, Amy Adams, Meli	United States	January 13, 2019	2010	R	116 min	Dramas, Sports Movies	After a string of defeats, a boxer rediscovers

Out[6]: (8807, 12)

```
In [8]:
        1 # the summary of the dataset
         3 df.info(memory_usage="deep")
       <class 'pandas.core.frame.DataFrame'>
       RangeIndex: 8807 entries, 0 to 8806
       Data columns (total 12 columns):
        # Column
                       Non-Null Count Dtype
        0
                     8807 non-null
                                       object
            show_id
            type 8807 non-null
        1
                                        object
        2
            title
                        8807 non-null
                                        object
            director 6173 non-null
                                        object
        4
                      7982 rion ...
7976 non-null
                         7982 non-null
            cast
                                        object
            country
        5
                                        object
            date_added 8797 non-null
        6
                                        object
        7
            release_year 8807 non-null
        8 rating
                      8803 non-null
                                        object
            duration
                        8804 non-null
                                        object
                        8807 non-null
        10 listed_in
                                        object
        11 description 8807 non-null
                                        object
       dtypes: int64(1), object(11)
       memory usage: 8.5 MB
In [9]: 1 # the descriptive statistics summary
         3 df.describe().round(2)
Out[9]:
```

	release_year
count	8807.00
mean	2014.18
std	8.82
min	1925.00
25%	2013.00
50%	2017.00
75%	2019.00
max	2021.00

This dataset does not contain any numerical data for the the descriptive statistics.

Data Preparation

- Dropping irrelevant columns and rows
- Identifying missing values
- · Identifying duplicated columns
- · Renaming columns
- Feature Creation

```
In [10]:
          1 # the number of missing values
           3 df.isna().sum().sort_values(ascending=False)
Out[10]: director
                         2634
         country
                          831
                          825
         cast
         date_added
                           10
         rating
         duration
                            0
         show_id
         type
         title
                            0
                            0
         release year
         listed_in
                            0
         description
                            0
         dtype: int64
```

There are high amounts of missing values in 3 columns and it is better to keep these rows for analysing by replacing null values with the string for these specific columns that are director, country, cast.

```
In [11]: 1 # filling all missing values in the director column with unknown director
           3 df["director"].fillna("unknown director", inplace=True)
In [65]: 1 df["director"].isna().sum()
Out[65]: 0
In [13]: 1 # filling all missing values in the country column with unknown country
           3 df["country"].fillna("unknown country", inplace=True)
In [66]: 1 df["country"].isna().sum()
Out[66]: 0
In [15]: | 1 # filling all missing values in the cast column with unknown cast
           3 df["cast"].fillna("unknown cast", inplace=True)
In [48]: 1 df["cast"].isna().sum()
Out[48]: 0
         The missing values were filled witht the string per specific column for the analysis.
In [17]: 1 # dropping the missing values on specific columns
           3 df.dropna(subset=["date_added", "rating", "duration"], inplace=True)
In [18]: 1 df.isna().sum()
Out[18]: show_id
         type
         title
         director
                         0
         cast
                         0
         country
         date_added
                         0
         release_year
         rating
         duration
                         0
         listed_in
                         a
         description
                         0
         dtype: int64
         After the necessary action such as replacing the null values and dropping null values, there is no missing value on dataset.
In [19]: 1 # the number of duplicate rows of data
           3 df.duplicated().sum()
Out[19]: 0
         There is no duplicate rows of data.
In [23]: 1 # the data type of the date_added column is converted to "datetime64"
           3 # date_added_year and date_added_month are created.
           5 df = df.assign(
           6
                   date_added = df["date_added"].astype("datetime64"),
           7
                   date_added_year = lambda x: x["date_added"].dt.year,
                   date_added_month = lambda x: x["date_added"].dt.month
```

9)

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

2021-09-24

2021-09-24

2021

2021

MA

TV-

MA Seasons

Season

TV Shows, TV Act...

Docuseries,

International

TV Shows, Romantic TV

Shows, TV ...

Reality TV

powerful drug lor...

Feuds, flirtations

talk go

down amo... 2021

2021

and toilet

In a city of

coaching

centers

known to

train I...

Exploratory Data Analysis

Jailbirds

Orleans

New

Kota

Factory

unknown

director

unknown

director

TV

TV

Show

Show

s4

3

1. How is the number of content added to Netflix changed over time?

Samuel

Jouy,

unknown unknown

country

India

cast

Mayur More,

Jitendra

Kumar,

Ranjan

Raj, Alam

```
In [68]:  # type_pivot is created to plot the trend over time

type_pivot = df.pivot_table(
    index="date_added_year",
    columns="type",
    values="title",
    aggfunc="count",
    # margins=True
    ).fillna(0).astype("int16")

type_pivot

type_pivot is created to plot the trend over time

type_pivot is created to plot the trend over time

type_pivot is created to plot the trend over time

type_pivot is created to plot the trend over time

type_pivot is created to plot the trend over time

type_pivot is created to plot the trend over time

type_pivot = type_pivot is created to plot the trend over time

type_pivot = type_pivot is created to plot the trend over time

type_pivot = type_pivot is created to plot the trend over time

type_pivot = type_pivot is created to plot the trend over time

type_pivot = type_pivot is created to plot the trend over time

type_pivot = type_pivot is created to plot the trend over time

type_pivot = type_pivot is created to plot the trend over time

type_pivot = type_pivot is created to plot the trend over time

type_pivot = type_pivot is created to plot the trend over time

type_pivot = type_pivot is created to plot the trend over time

type_pivot = type_pivot is created to plot the trend over time

type_pivot = type_pivot is created to plot the trend over time

type_pivot = type_pivot is created to plot the trend over time

type_pivot = type_pivot is created to plot the trend over time

type_pivot = type_pivot is created to plot the trend over time

type_pivot = type_pivot is created to plot the trend over time

type_pivot = type_pivot is created to plot the trend over time

type_pivot = type_pivot is created to plot the trend over time

type_pivot = type_pivot is created to plot the trend over time

type_pivot = type_pivot is created to plot the trend over time

type_pivot = type_pivot is created to plot the trend over time

type_pivot = type_pivot is created to plot the trend over time

type_pivot = type_pivot is created to plot the trend over time

type_pivot
```

Out[68]:

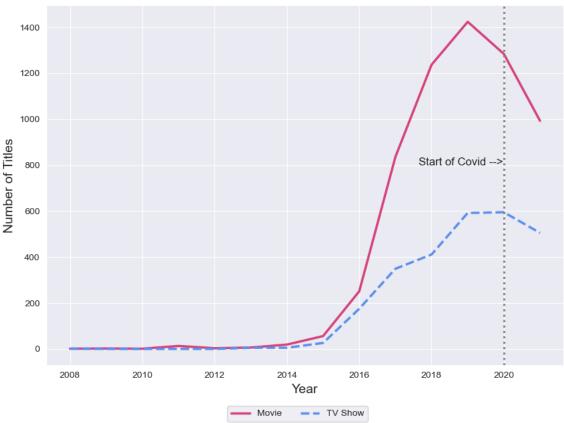
In [24]:

1 df.head()

type	Movie	TV Show
date_added_year		
2008	1	1
2009	2	0
2010	1	0
2011	13	0
2012	3	0
2013	6	5
2014	19	5
2015	56	26
2016	251	175
2017	836	349
2018	1237	411
2019	1424	592
2020	1284	595
2021	993	505

```
In [40]:
          fig, ax = plt.subplots(figsize=(10, 7))
           3
              ax.plot(
           4
                  type_pivot.index,
                  type_pivot["Movie"],
           5
                  color="#d6427b",
           6
                  label="Movie",
           7
                  ls="-"
           8
           9
                  linewidth="2.5"
          10 )
          11
          12
             ax.plot(
          13
                  type_pivot.index,
                  type_pivot["TV Show"],
          14
                  color="#5a90ed",
          15
                  label="TV Show",
          16
                  ls="--"
          17
          18
                  linewidth="2.5"
          19 )
          20
          21 ax.set_title("Movie versus TV Show in 2008-2021", fontsize=16)
          22 ax.set_xlabel("Year", fontsize=14)
          23
             ax.set_ylabel("Number of Titles", fontsize=14)
          25 ax.axvline(2020, c="grey", 1s=":", linewidth=2.2)
          26
          27
             ax.text(2017.65, 800,
          28
                       "Start of Covid -->",
          29
                       fontsize=12)
          30
          31
             ax.legend(bbox_to_anchor=(.63,-.1), ncol=2);
```





The number of movies and TV shows added to Netflix has generally increased over time, with a significant spike in 2017, 2018, and 2019...

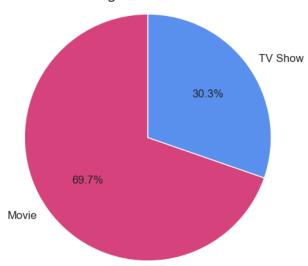
In 2019 and 2020, there were a similar number of TV shows added, with over 500 new shows added in both years. In 2021, the number of TV shows decreased for the first time after the upward trend.

Additionally, in 2020 and 2021, the number of movies added to the platform decreased slightly compared to the previous year.

This could be due to the impact of the COVID-19 pandemic on the entertainment industry, causing delays in movie releases and production.

```
In [34]:
          1 fig, ax = plt.subplots()
           3 type_color = ["#d6427b","#5a90ed" ]
              df.groupby("type")["title"].count() \
              .plot.pie(
                  ylabel="",
           8
                  colors= type_color,
           9
                  startangle=90,
                  autopct="%1.1f%%",
figsize=(6,6),
          10
          11
          12
                  fontsize=12
          13 )
          15
              plt.title("The Percentage of TV Show and Movie",
                        y=0.93,
          16
          17
                        fontsize=16);
```

The Percentage of TV Show and Movie



The percentage of movies is 69.7% of the total types and this is more than double the percentage of TV shows in the total types. The percentage of TV shows is 30.3% of the total types.

3.In which months is more content being added to Netflix?

In [36]:

1 df.head()

Out[36]:

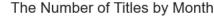
	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	listed_in	description	date_added_year	date_add
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	unknown cast	United States	2021-09-25	2020	PG- 13	90 min	Documentaries	As her father nears the end of his life, filmm	2021	
1	s2	TV Show	Blood & Water	unknown director	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	2021	
2	s3	TV Show	Ganglands	Julien Leclercq	Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi	unknown country	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	2021	
3	s4	TV Show	Jailbirds New Orleans	unknown director	unknown cast	unknown country	2021-09-24	2021	TV- MA	1 Season	Docuseries, Reality TV	Feuds, flirtations and toilet talk go down amo	2021	
4	s5	TV Show	Kota Factory	unknown director	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	2021	

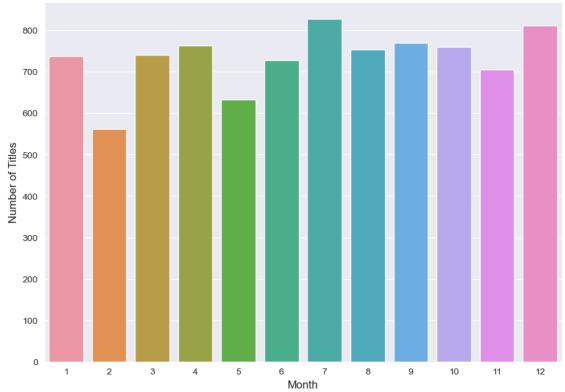
```
In [37]: | 1 # the creation of month_df to present the number of content by month
          month_df = df.groupby("date_added_month").agg(number_of_content=("title","count")).reset_index()
          4
5 month_df
```

Out[37]:

	date_added_month	number_of_content
0	1	737
1	2	562
2	3	741
3	4	763
4	5	632
5	6	728
6	7	827
7	8	754
8	9	769
9	10	760
10	11	705
11	12	812

```
In [39]:
            fig, ax = plt.subplots(figsize=(10, 7))
               sns.barplot(
                   x="date_added_month",
y="number_of_content",
data=month_df
            4
            5
            6
            7
               )
            9
               ax.set_title(
           10
                    "The Number of Titles by Month",
           11
12
                    fontsize=16)
           13
               ax.set_xlabel(
           14
                    "Month",
           15
                    fontsize=12)
           16
           17
               ax.set_ylabel(
                    "Number of Titles",
           18
                    fontsize=12);
```





The months with the highest number of content additions are July and December, with 827 and 812 additions, respectively.

The month with the lowest number of content additions is February, with 562 additions.

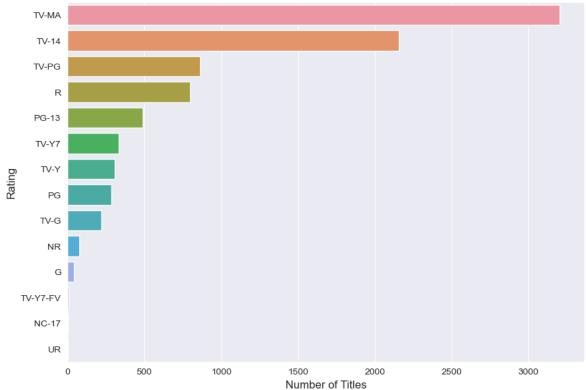
4. Which rating is the highest on Netflix ?

Out[42]:

rating	number_of_titles
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
G	41
TV-Y7-FV	6
NC-17	3
UR	3
	TV-MA TV-14 TV-PG R PG-13 TV-Y7 TV-Y PG TV-G NR G TV-Y7-FV NC-17

```
In [45]:
          fig, ax = plt.subplots(figsize=(10, 7))
             sns.barplot(
                 x="number_of_titles",
           4
                 y="rating",
                 data=rating_df
           7
             )
           8
           9
             ax.set_title(
          10
                  "The Number of Titles by Rating",
          11
                 fontsize=16)
          12
          13
             ax.set_xlabel(
                  "Number of Titles",
          15
                 fontsize=12)
          16
          17
             ax.set_ylabel(
          18
                 fontsize=12);
```





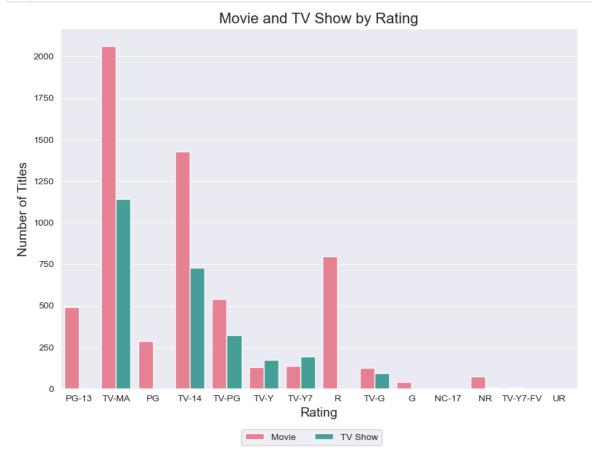
The rating with the largest number of content titles is TV-MA, which is content intended for adult viewers and may be unsuitable for children under 17.

Meanwhile, the rating with the smallest number of content titles are NC-17 (no children under 17 admitted) and UR(unrated).

5. What is the most common rating for Netflix movies and TV shows?

```
In [46]: | 1 | # the creation of rating_pivot to present the number of titles by rating
              rating_pivot = df.pivot_table(index="rating",
                                                        columns="type",
values="title",
aggfunc="count",
margins=True) \
              5
6
7
8
9
                                 .fillna(0).astype("int16") \
.reset_index() \
.sort_values("Movie")
             10
             11
             12
             13 rating_pivot
```

Out[46]:					
	type	rating	Movie	TV Show	All
	1	NC-17	3	0	3
	13	UR	3	0	3
	12	TV-Y7-FV	5	1	6
	0	G	41	0	41
	2	NR	75	4	79
	7	TV-G	126	94	220
	10	TV-Y	131	175	306
	11	TV-Y7	139	194	333
	3	PG	287	0	287
	4	PG-13	490	0	490
	9	TV-PG	540	321	861
	5	R	797	2	799
	6	TV-14	1427	730	2157
	8	TV-MA	2062	1143	3205
	14	All	6126	2664	8790



The rating with the highest number of movies is TV-MA with 6126 titles.

The rating with the highest number of TV shows is TV-MA with 2664 titles.

6. Which are the top 10 content producing countries?

Out[55]:

number_of_titles

country	
United States	2809
India	972
United Kingdom	418
Japan	243
South Korea	199
Canada	181
Spain	145
France	124
Mexico	110
Egypt	106

The highest number of titles on Netflix is the United States, with 2809 titles. India has the second-highest number of titles with 972, while the United Kingdom has 418 titles.

7. Which directors or actors/actresses appear most frequently on Netflix?

Out[70]:

number_of_titles

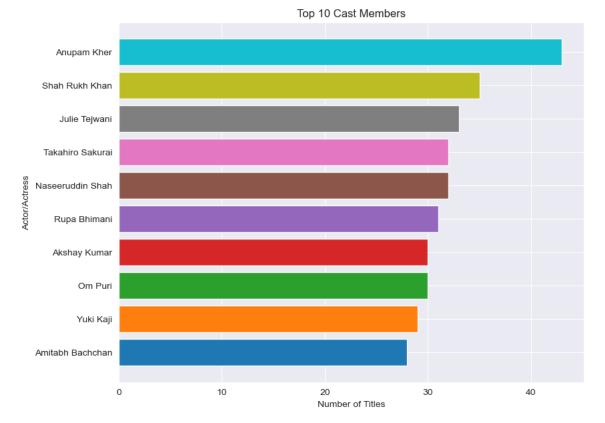
director	
Rajiv Chilaka	19
Raúl Campos, Jan Suter	18
Marcus Raboy	16
Suhas Kadav	16
Jay Karas	14
Cathy Garcia-Molina	13
Jay Chapman	12
Martin Scorsese	12
Youssef Chahine	12
Steven Spielberg	11

The table shows the top 10 directors with the most titles on Netflix.

The first position belongs to Rajiv Chilaka, an Indian director who has managed 19 content titles at Netflix.

Out[62]:

```
number_of_films
             cast
                               43
    Anupam Kher
  Shah Rukh Khan
                               35
     Julie Tejwani
                               33
  Takahiro Sakurai
                               32
Naseeruddin Shah
                               32
    Rupa Bhimani
                               31
   Akshay Kumar
                               30
         Om Puri
                               30
         Yuki Kaji
                               29
Amitabh Bachchan
                               28
```



The bar chart shows the top 10 cast members with the most titles on Netflix.

The first position belongs to Shah Rukh Khan, an Indian actor who has appeared in 43 content titles on Netflix.

Conclusions

1. How is the number of content added to Netflix changed over time?

The number of movies and TV shows added to Netflix has generally increased over time, with a significant spike in 2017, 2018, and 2019...

In 2019 and 2020, there were a similar number of TV shows added, with over 500 new shows added in both years. In 2021, the number of TV shows decreased for the first time after the upward trend.

Additionally, in 2020 and 2021, the number of movies added to the platform decreased slightly compared to the previous year.

This could be due to the impact of the COVID-19 pandemic on the entertainment industry, causing delays in movie releases and production.

2. What is the percentage of TV Show and Movie on Netflix?

The percentage of movies is 69.7% of the total types and this is more than double the percentage of TV shows in the total types. The percentage of TV shows is 30.3% of the total types.

3. In which months is more content being added to Netflix?

The months with the highest number of content additions are July and December, with 827 and 812 additions, respectively.

The month with the lowest number of content additions is February, with 562 additions.

4. Which rating is the highest on Netflix?

The rating with the largest number of content titles is TV-MA, which is content intended for adult viewers and may be unsuitable for children under 17.

Meanwhile, the rating with the smallest number of content titles are NC-17 (no children under 17 admitted) and UR(unrated).

5. What is the most common rating for Netflix movies and TV shows?

The rating with the highest number of movies is TV-MA with 6126 titles.

The rating with the highest number of TV shows is TV-MA with 2664 titles

6. Which are the top 10 content producing countries?

The highest number of titles available on Netlifx is the United States, with 2809 titles. India has the second-highest number of titles with 972, while the United Kingdom has 418 titles. Japan, South Korea, Canada, Spain, France, Mexico, and Egypt round out the remaining countries on the top 10 list.

7. Which directors or actors/actresses appear most frequently on Netflix?

The first position belongs to Rajiv Chilaka, an Indian director who has managed 19 content titles at Netflix.

The first position belongs to Shah Rukh Khan, an Indian actor who has appeared in 43 content titles on Netflix.