```
import pandas as pd
import matplotlib.pyplot as plt
```

```
Matplotlib is building the font cache; this may take a moment.
```

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

```
df.head()
```

```
  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...
```

```
print(df.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8807 entries, 0 to 8806
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
```

```
 0   show_id       8807 non-null   object
 1   type          8807 non-null   object
 2   title         8807 non-null   object
 3   director      6173 non-null   object
 4   cast          7982 non-null   object
 5   country       7976 non-null   object
 6   date_added    8797 non-null   object
 7   release_year  8807 non-null   int64
 8   rating        8803 non-null   object
 9   duration      8804 non-null   object
 10  listed_in     8807 non-null   object
 11  description   8807 non-null   object
dtypes: int64(1), object(11)
memory usage: 447.3+ KB
None
```

```
df.describe(include='all')
```

```
       show_id   type                  title       director  \
count     8807   8807                   8807           6173
unique    8807      2                   8807           4528
top         s1  Movie  Dick Johnson Is Dead  Rajiv Chilaka
freq         1   6131                      1             19
mean       NaN    NaN                    NaN            NaN
std        NaN    NaN                    NaN            NaN
min        NaN    NaN                    NaN            NaN
25%        NaN    NaN                    NaN            NaN
50%        NaN    NaN                    NaN            NaN
75%        NaN    NaN                    NaN            NaN
max        NaN    NaN                    NaN            NaN

                        cast        country       date_added
release_year  \
count                   7982           7976             8797
8807.000000
unique                  7692            748             1767
NaN
top     David Attenborough  United States  January 1, 2020
NaN
freq                      19           2818              109
NaN
mean                     NaN            NaN              NaN
2014.180198
std                      NaN            NaN              NaN
8.819312
min                      NaN            NaN              NaN
1925.000000
25%                      NaN            NaN              NaN
2013.000000
50%                      NaN            NaN              NaN
```

```
2017.000000
75%                    NaN             NaN                NaN
2019.000000
max                    NaN             NaN                NaN
2021.000000

       rating  duration                      listed_in  \
count    8803      8804                           8807
unique     17       220                            514
top     TV-MA  1 Season  Dramas, International Movies
freq     3207      1793                            362
mean      NaN       NaN                            NaN
std       NaN       NaN                            NaN
min       NaN       NaN                            NaN
25%       NaN       NaN                            NaN
50%       NaN       NaN                            NaN
75%       NaN       NaN                            NaN
max       NaN       NaN                            NaN

                                                 description
count                                                   8807
unique                                                  8775
top     Paranormal activity at a lush, abandoned prope...
freq                                                       4
mean                                                     NaN
std                                                      NaN
min                                                      NaN
25%                                                      NaN
50%                                                      NaN
75%                                                      NaN
max                                                      NaN
```

```
df.isnull().sum()
```

```
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
```

#Filling null values in director column

```
df['director']=df['director'].fillna('N.A')
```

```
df.isnull().sum()
```

```
show_id           0
type              0
title             0
director          0
cast            825
country         831
date_added       10
release_year      0
rating            4
duration          3
listed_in         0
description       0
dtype: int64
```

```
df['cast']=df['cast'].fillna('N.A')
```

```
df['country']=df['country'].fillna('unknown')
```

```
df['date_added'] = df['date_added'].str.strip()

# Convert to datetime invalid dates NaT
df['date_added'] = pd.to_datetime(df['date_added'], errors='coerce')
```

```
df['date_added'] = df['date_added'].fillna(pd.NaT)
```

```
df['rating']=df['rating'].fillna('unknown')
```

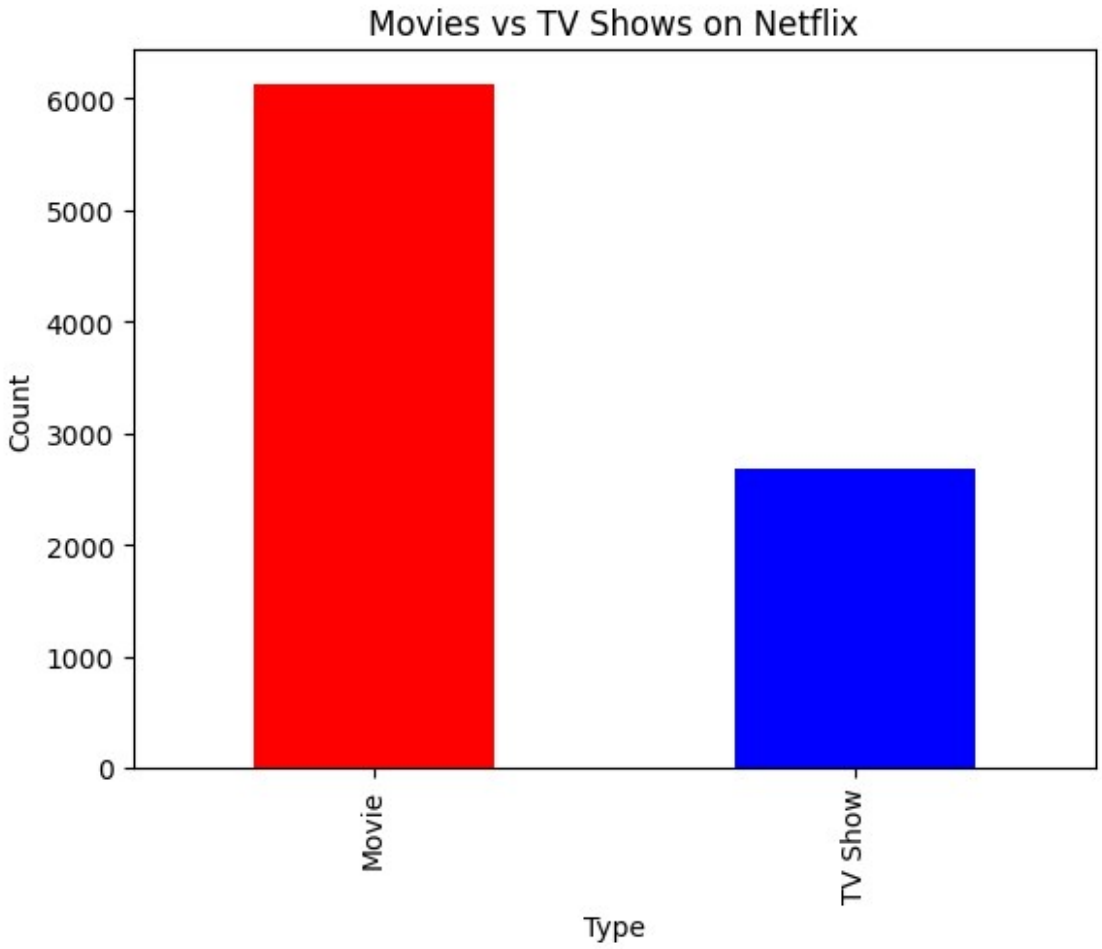
```
df['duration']=df['duration'].fillna("N.A")
```

```
df.isnull().sum()
```

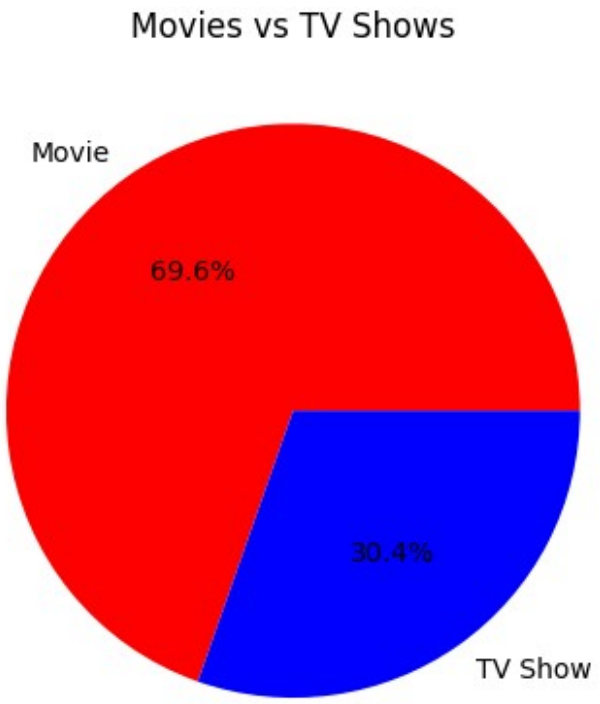
```
show_id           0
type              0
title             0
director          0
cast              0
country           0
date_added       10
release_year      0
rating            0
duration          0
listed_in         0
description       0
dtype: int64
```

#Here Completed the Data Cleaning Part

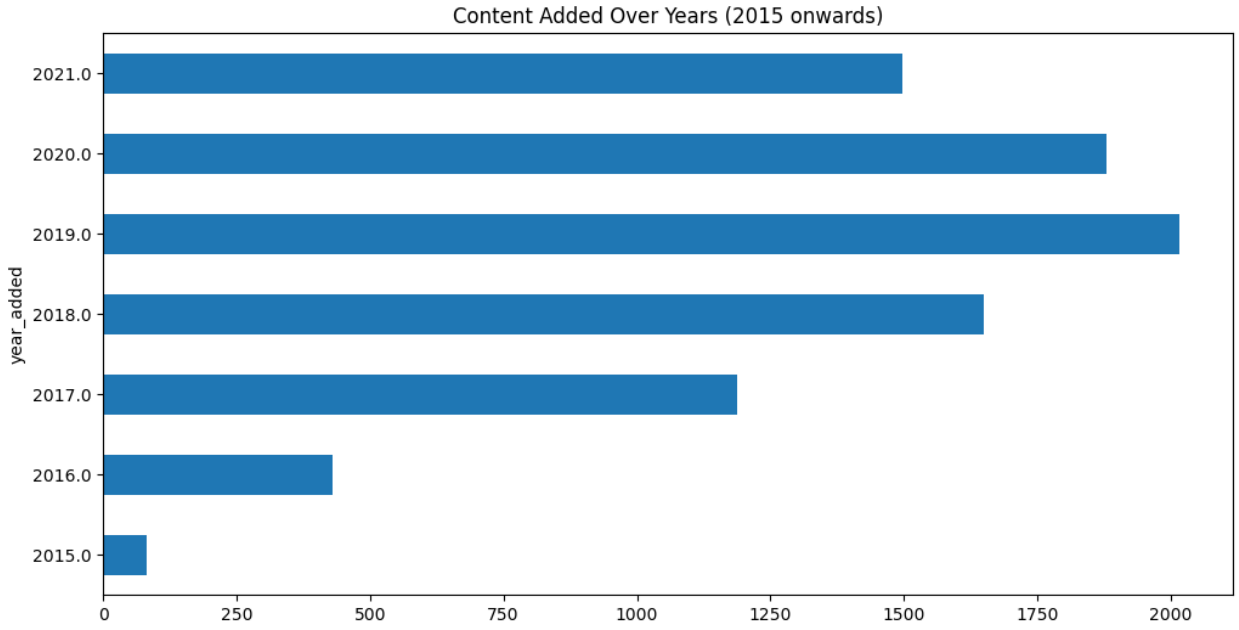
```
# Movies vs TV Shows
df['type'].value_counts().plot(kind='bar', color=['red','blue'])
plt.title("Movies vs TV Shows on Netflix")
plt.xlabel("Type")
plt.ylabel("Count")
plt.show()
```



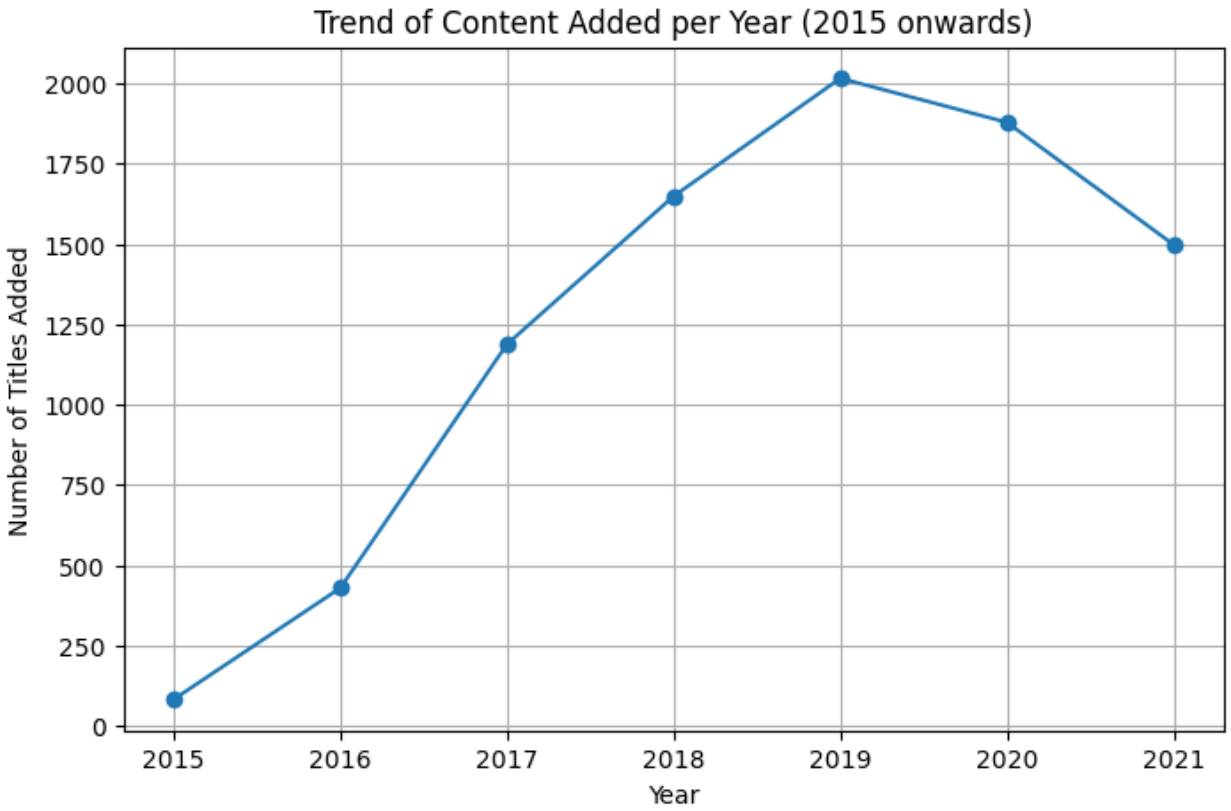

```
df['type'].value_counts().plot(kind='pie', autopct='%1.1f%%',
colors=['red','blue'])
plt.title("Movies vs TV Shows")
plt.ylabel("") # Hide y-label
plt.show()
```



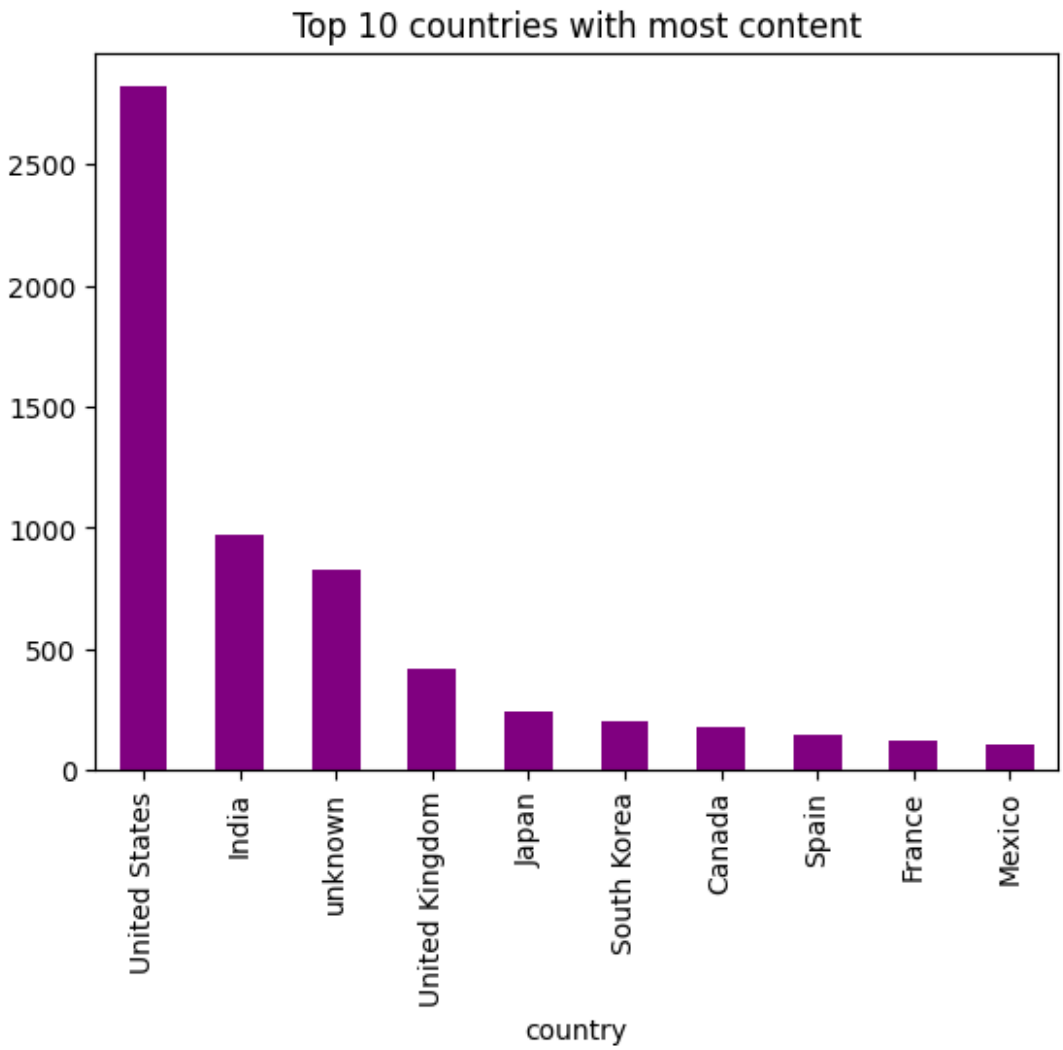

```
df_recent = df[df['year_added'] >= 2015]

df_recent['year_added'].value_counts().sort_index().plot(kind='barh',
figsize=(12,6))
plt.title("Content Added Over Years (2015 onwards)")
plt.show()
```



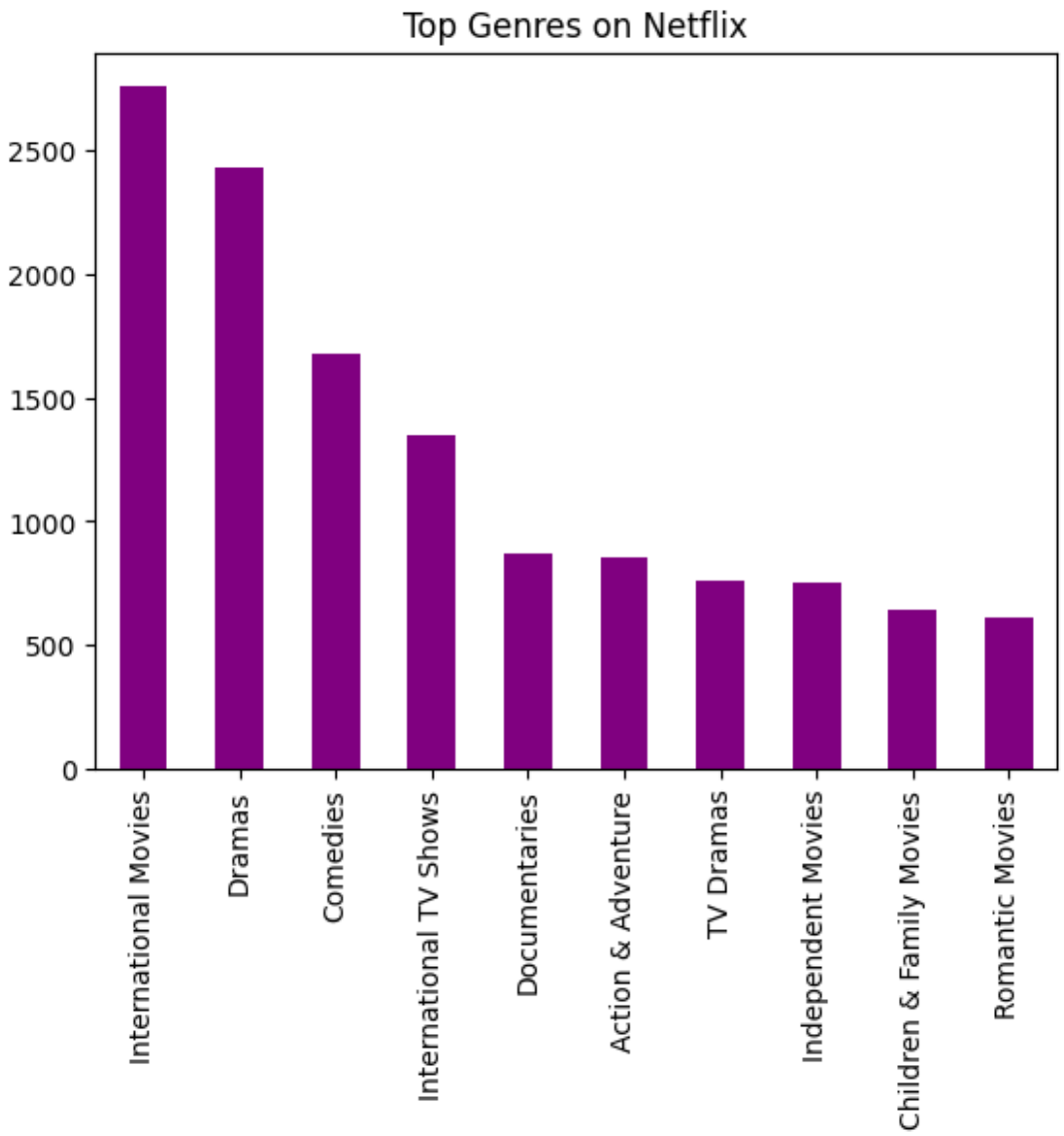
```
year_counts = df_recent['year_added'].value_counts().sort_index()
year_counts.plot(kind='line', marker='o', figsize=(8,5))
plt.title("Trend of Content Added per Year (2015 onwards)")
plt.xlabel("Year")
plt.ylabel("Number of Titles Added")
plt.grid(True)
plt.show()
```



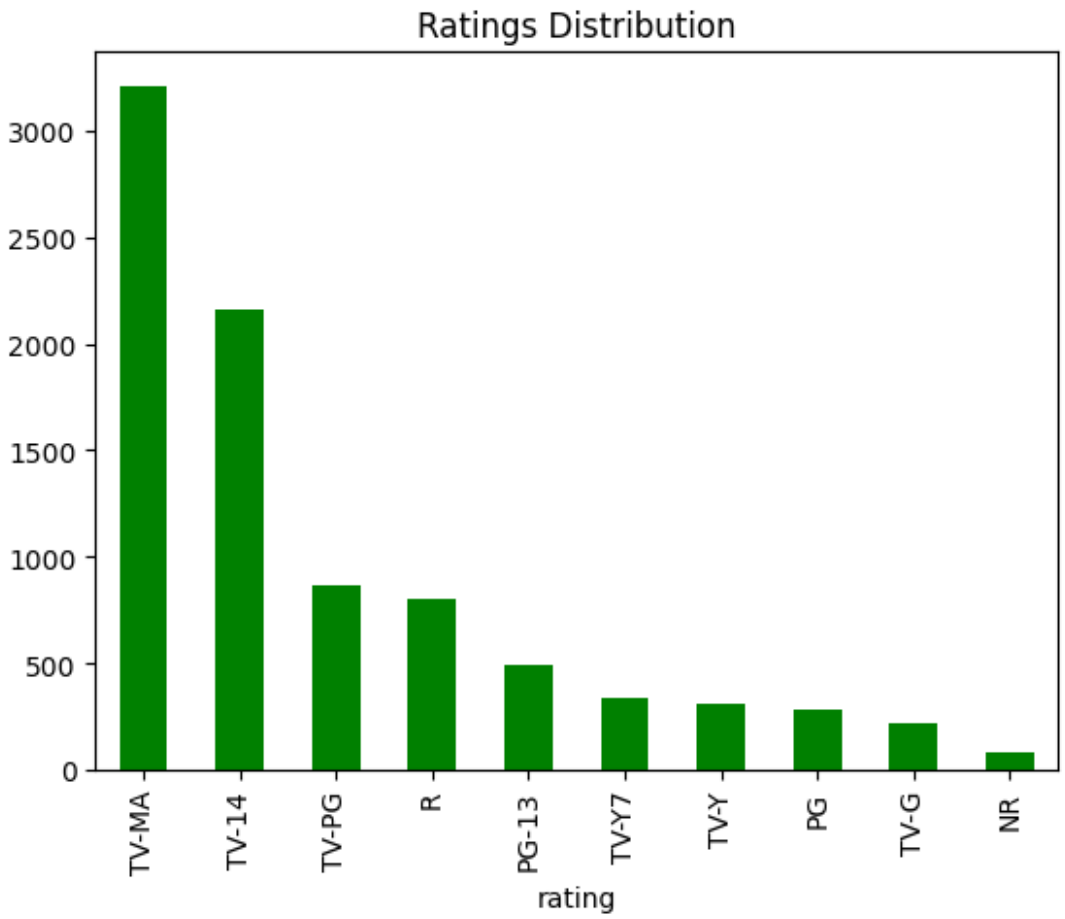

```
df['country'].value_counts().head(10).plot(kind='bar',color='purple')
plt.title("Top 10 countries with most content")
plt.show()
```



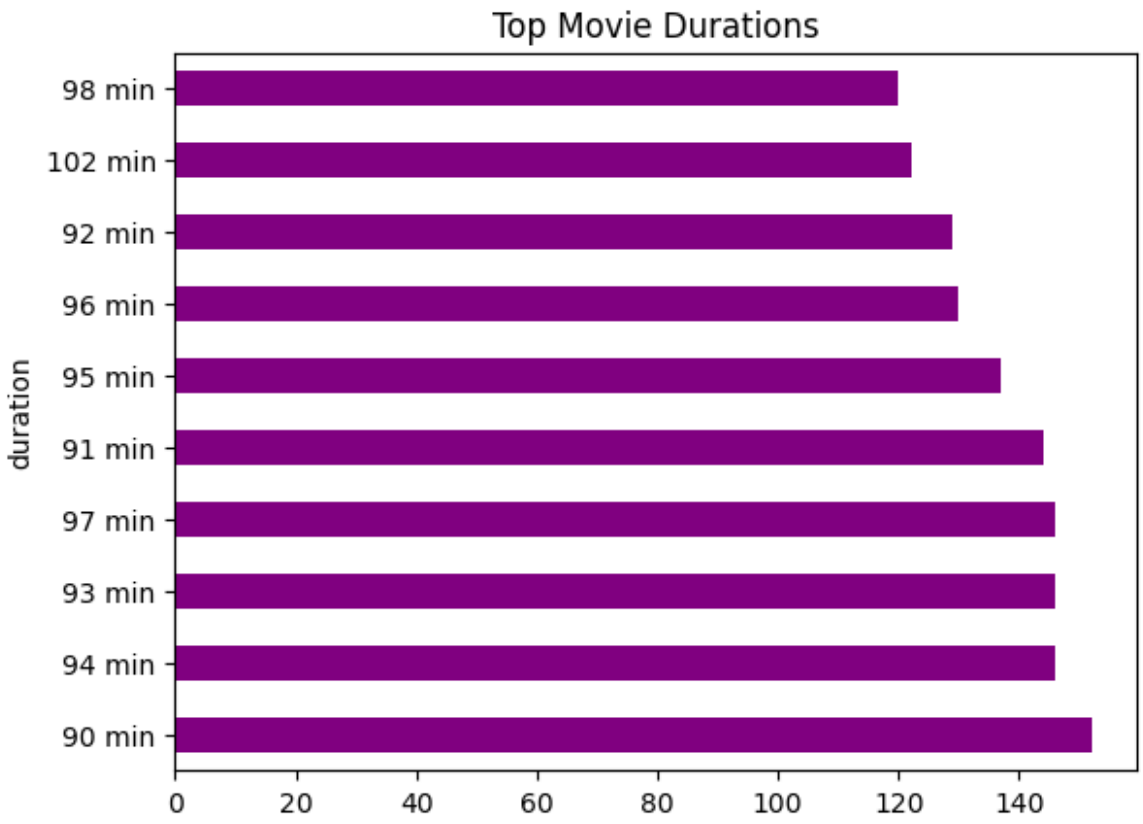

```
genres = df['listed_in'].dropna().str.split(', ')
genre_list = [g for sublist in genres for g in sublist]
pd.Series(genre_list).value_counts().head(10).plot(kind='bar',color='p
urple')
plt.title("Top Genres on Netflix")
plt.show()
```

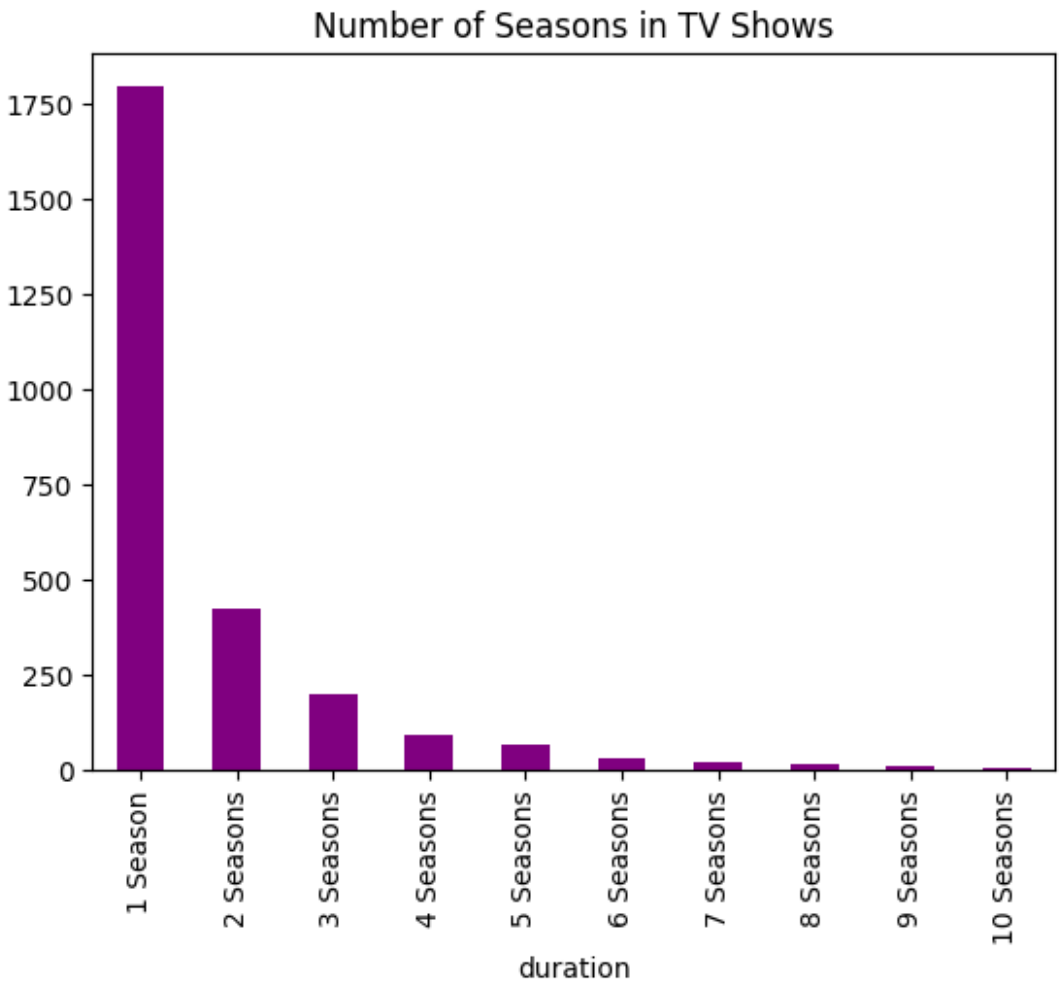



```
df['rating'].value_counts().head(10).plot(kind='bar',color='green')
plt.title("Ratings Distribution")
plt.show()
```




```
movies = df[df['type'] == 'Movie']
tv_shows = df[df['type'] == 'TV Show']

# Movies duration
movies['duration'].value_counts().head(10).plot(kind='barh',color='pur
ple')
plt.title("Top Movie Durations")
plt.show()

# TV Shows seasons
tv_shows['duration'].value_counts().head(10).plot(kind='bar',color='pu
rple')
plt.title("Number of Seasons in TV Shows")
plt.show()
```


Top Movie Durations
duration
98 min
102 min
92 min
96 min
95 min
91 min
97 min
93 min
94 min
90 min
0
20
40
60
80
100
120
140




```
df.to_csv('Netflix_Data_analysis.csv', index=False)
```