netflix-business-case

September 10, 2024

1 Netflix_Business_Case

About Netflix

Netflix is one of the most popular media and video streaming platforms. They have over 10000 movies or tv shows available on their platform, as of mid-2021, they have over 222M Subscribers globally. This tabular dataset consists of listings of all the movies and tv shows available on Netflix, along with details such as - cast, directors, ratings, release year, duration, etc.

2 Business Problem

Analyze the data and generate insights that could help Netflix ijn deciding which type of shows/movies to produce and how they can grow the business in different countries

1.(Analysing Basic Metrics)

[1]:
```
import numpy as np
import pandas as pd
```

Loading Dataset

[2]:
```
netflix = pd.read_csv('Downloads/netflix.csv')
netflix.head()
```

[2]:
```
  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…
```

Observing null values along columns

```
[3]: print('columns','      ', 'count')
     netflix.isnull().count()
```

```
columns        count

[3]: show_id         8807
     type            8807
     title           8807
     director        8807
     cast            8807
     country         8807
     date_added      8807
     release_year    8807
     rating          8807
     duration        8807
     listed_in       8807
     description     8807
     dtype: int64
```

Observing names of columns present in the dataset

```
[4]: netflix.columns
```

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

observing dimention, size and shape of dataset

[5]:

```
print("dimention of dataset", netflix.ndim)
```

```
dimention of dataset 2
```

[6]:

```
print("size of dataset", netflix.size)
```

```
size of dataset 105684
```

[7]:

```
print("shape of dataset", netflix.shape)
```

```
shape of dataset (8807, 12)
```

2. Observing informatiion of Dataset, includes column name, Not-Null Count and Datatype of each Column(Basic Data Matrix)

[8]:

```
netflix.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
```

Checking for columns with missing values

[9]:

```
print("columns with missing values")
netflix.isnull().any()
```

```
columns with missing values
```

```
[9]: show_id         False
     type            False
     title           False
     director         True
     cast             True
     country          True
     date_added       True
     release_year    False
     rating           True
     duration         True
     listed_in       False
     description     False
     dtype: bool
```

Number of null values in the Dataset

```
[10]: netflix.isna().sum().sum()
```

```
[10]: 4307
```

Describing dataset

```
[11]: netflix.describe()
```

```
[11]:        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
```

Columns with number of missing value/Null values

```
[12]: print('columns with missung values')
      netflix.isnull().sum()
```

```
columns with missung values
```

```
[12]: 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
```

checking unique values

```
[13]: netflix.nunique()
```

```
[13]: 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
```

Finding and Handling Null values

Process of finding , cleaning, analyzing the missing values of data and providing the correctly formated data for further analysis is the major part of the Data cleaning process.

```
[14]: netflix.isna().sum()
```

```
[14]: 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
```

```
[15]: netflix_df = netflix
```

[16]:
```
netflix['director']=netflix['director'].str.split(",")
netflix["country"]= netflix["country"].str.split(",")
netflix["cast"]=netflix["cast"].str.split(",")
netflix["listed_in"] =netflix["listed_in"].str.split(",")
netflix.head(5)
```

[16]:
```
  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,  Th…   [South Africa]
2  [Sami Bouajila,  Tracy Gotoas,  Samuel Jouy,  …              NaN
3                                                NaN              NaN
4  [Mayur More,  Jitendra Kumar,  Ranjan Raj,  Al…          [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 Myste…
2  [Crime TV Shows,  International TV Shows,  TV …
3                          [Docuseries,  Reality TV]
4  [International TV Shows,  Romantic TV Shows,  …

                                         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…
```

exploding the data to individual rows based on the multiple supported entries for further data imputation

[17]:
```
# exloding data into new columns
def explode_columns(row):
    row['director'] = pd.Series(row['director']).explode()
    row['cast'] = pd.Series(row['cast']).explode()
    row['country'] = pd.Series(row['country']).explode()
    row['listed_in'] = pd.Series(row['listed_in']).explode()
    return row

netflix_df_explode = netflix.apply(explode_columns, axis= 1).
 ↪explode('director').explode('cast').explode('country').explode('listed_in').
 ↪reset_index(drop = True)
netflix_df_explode.head()
```

[17]:

```
  show_id     type                 title         director          cast  \
0      s1    Movie  Dick Johnson Is Dead  Kirsten Johnson           NaN
1      s2  TV Show         Blood & Water              NaN    Ama Qamata
2      s2  TV Show         Blood & Water              NaN    Ama Qamata
3      s2  TV Show         Blood & Water              NaN    Ama Qamata
4      s2  TV Show         Blood & Water              NaN   Khosi Ngema

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

                listed_in                                        description
0           Documentaries  As her father nears the end of his life, filmm…
1  International TV Shows  After crossing paths at a party, a Cape Town t…
2               TV Dramas  After crossing paths at a party, a Cape Town t…
3            TV Mysteries  After crossing paths at a party, a Cape Town t…
4  International TV Shows  After crossing paths at a party, a Cape Town t…
```

Replacing NaN value with imputation method

[18]:

```
# replacing Nan value in 'directors' with Unkown_Directors, 'cast' with␣
 ↪'Unknown Actors' and 'listed_in' with 'Not listed'

netflix_df_explode['director'] = netflix_df_explode.director.fillna('Unknown␣
 ↪Directors')
netflix_df_explode['cast'] = netflix_df_explode.cast.fillna('Unknown Actors')
netflix_df_explode['listed_in'] = netflix_df_explode.listed_in.fillna('Not␣
 ↪listed')
netflix_df_explode.head()
```

[18]:

```
  show_id     type                 title           director            cast  \
0      s1    Movie  Dick Johnson Is Dead    Kirsten Johnson  Unknown Actors
1      s2  TV Show         Blood & Water  Unknown Directors      Ama Qamata
2      s2  TV Show         Blood & Water  Unknown Directors      Ama Qamata
```

```
3      s2  TV Show         Blood & Water  Unknown Directors      Ama Qamata
4      s2  TV Show         Blood & Water  Unknown Directors     Khosi Ngema

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

                listed_in                                        description
0           Documentaries  As her father nears the end of his life, filmm…
1  International TV Shows  After crossing paths at a party, a Cape Town t…
2               TV Dramas  After crossing paths at a party, a Cape Town t…
3            TV Mysteries  After crossing paths at a party, a Cape Town t…
4  International TV Shows  After crossing paths at a party, a Cape Town t…
```

[19]:

```
# dataframe after handling NaN values
netflix_df_explode.isna().sum()
```

[19]:

```
show_id             0
type                0
title               0
director            0
cast                0
country         11897
date_added        158
release_year        0
rating             67
duration            3
listed_in           0
description         0
dtype: int64
```

Now using mode imputation method to fill NaN values in 'country', 'date_added' For 'rating' column rating cannot be predicted so filling with 'Not-rated'(NR)

[20]:

```
netflix_df_explode['rating'] = netflix_df_explode.rating.fillna('NR')
netflix_df_explode.head()
```

[20]:

```
  show_id     type                 title           director            cast  \
0      s1    Movie  Dick Johnson Is Dead    Kirsten Johnson  Unknown Actors
1      s2  TV Show         Blood & Water  Unknown Directors      Ama Qamata
2      s2  TV Show         Blood & Water  Unknown Directors      Ama Qamata
3      s2  TV Show         Blood & Water  Unknown Directors      Ama Qamata
4      s2  TV Show         Blood & Water  Unknown Directors     Khosi Ngema
```

```
         country          date_added  release_year rating   duration  \
0  United States  September 25, 2021          2020  PG-13     90 min
1   South Africa  September 24, 2021          2021  TV-MA  2 Seasons
2   South Africa  September 24, 2021          2021  TV-MA  2 Seasons
3   South Africa  September 24, 2021          2021  TV-MA  2 Seasons
4   South Africa  September 24, 2021          2021  TV-MA  2 Seasons

                listed_in                                        description
0           Documentaries  As her father nears the end of his life, filmm…
1  International TV Shows  After crossing paths at a party, a Cape Town t…
2               TV Dramas  After crossing paths at a party, a Cape Town t…
3            TV Mysteries  After crossing paths at a party, a Cape Town t…
4  International TV Shows  After crossing paths at a party, a Cape Town t…
```

[21]:

```
netflix_df_explode.isna().sum()
```

[21]:

```
show_id             0
type                0
title               0
director            0
cast                0
country         11897
date_added        158
release_year        0
rating              0
duration            3
listed_in           0
description         0
dtype: int64
```

[22]:

```
# date_added cannot be imputed manually so imputing date_added with release year

for i in netflix_df_explode[netflix_df_explode['date_added'].
 ↪isnull()]['release_year'].unique():
    impu = netflix_df_explode[netflix_df_explode['release_year'] ==␣
 ↪i]['date_added'].mode().values[0]
    netflix_df_explode.loc[netflix_df_explode['release_year'] ==␣
 ↪i,'date_added'] = netflix_df_explode.loc[netflix_df_explode['release_year']␣
 ↪== i, 'date_added'].fillna(impu)

netflix_df_explode.head()
```

[22]:

```
  show_id     type               title          director            cast  \
0      s1    Movie  Dick Johnson Is Dead   Kirsten Johnson  Unknown Actors
1      s2  TV Show        Blood & Water  Unknown Directors      Ama Qamata
2      s2  TV Show        Blood & Water  Unknown Directors      Ama Qamata
3      s2  TV Show        Blood & Water  Unknown Directors      Ama Qamata
```

```
4      s2  TV Show         Blood & Water  Unknown Directors     Khosi Ngema

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

                listed_in                                       description
0           Documentaries  As her father nears the end of his life, filmm…
1  International TV Shows  After crossing paths at a party, a Cape Town t…
2               TV Dramas  After crossing paths at a party, a Cape Town t…
3            TV Mysteries  After crossing paths at a party, a Cape Town t…
4  International TV Shows  After crossing paths at a party, a Cape Town t…
```

[23]:

```
netflix_df_explode.isna().sum()
```

[23]:

```
show_id             0
type                0
title               0
director            0
cast                0
country         11897
date_added          0
release_year        0
rating              0
duration            3
listed_in           0
description         0
dtype: int64
```

[24]:

```
#Now 'country' column will be imputed with origin of director

for i in netflix_df_explode[netflix_df_explode['country'].isnull()]['director'].
 ↪unique():
    if i in netflix_df_explode[~netflix_df_explode['country'].
 ↪isnull()]['director'].unique():
        impu = netflix_df_explode[netflix_df_explode['director']==␣
 ↪i]['country'].mode().values[0]
        netflix_df_explode.loc[netflix_df_explode['director'] ==i,'country'] =␣
 ↪netflix_df_explode.loc[netflix_df_explode['director']== i,'country'].
 ↪fillna(impu)
```

[25]:

```
netflix_df_explode.isna().sum()
```

[25]:

```
show_id           0
type              0
title             0
director          0
cast              0
country        4673
date_added        0
release_year      0
rating            0
duration          3
listed_in         0
description       0
dtype: int64
```

[26]:

```
# there are still Nan values so replacing those with 'country unavailable'
netflix_df_explode['country'] = netflix_df_explode.country.fillna('country␣
 ↪unavailable')
```

[27]:

```
netflix_df_explode.isna().sum()
```

[27]:

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

[28]:

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

[28]:

```
rating
TV-MA       73915
TV-14       43957
R           25860
PG-13       16246
TV-PG       14926
PG          10919
TV-Y7        6304
TV-Y         3665
TV-G         2779
NR           1640
```

```
G            1530
NC-17         149
TV-Y7-FV       86
UR             86
74 min          1
84 min          1
66 min          1
Name: count, dtype: int64
```

As we can see that 3 of the nan values of 'dureation' column might captured in 'rating' column, so imputing those Nan values with rating values

[29]:

```
netflix_df_explode['duration'].value_counts()
```

[29]:

```
duration
1 Season      35035
2 Seasons      9559
3 Seasons      5084
94 min         4343
106 min        4040
              …
3 min             4
5 min             3
11 min            2
8 min             2
9 min             2
Name: count, Length: 220, dtype: int64
```

[ ]:

```

```

[ ]:

```

```

[30]:

```
netflix_df_explode.loc[netflix_df_explode['duration'].
  ↪isnull(),'duration']=netflix_df_explode.loc[netflix_df_explode['duration'].
  ↪isnull(),'duration'].fillna(netflix_df_explode['rating'])
netflix_df_explode.isnull().sum()
```

[30]:

```
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
```

3. Non-Graphical Analysis: Value counts and unique attributes

```
[31]: netflix_df_explode['duration'].value_counts()
      netflix_df_explode['duration']=netflix_df_explode['duration'].str.replace("␣
        ↪min","")
      netflix_df_explode['duration']=netflix_df_explode['duration'].str.replace("␣
        ↪Seasons","")
      netflix_df_explode['duration'].unique()
```

```
[31]: array(['90', '2', '1 Season', '91', '125', '9', '104', '127', '4', '67',
             '94', '5', '161', '61', '166', '147', '103', '97', '106', '111',
             '3', '110', '105', '96', '124', '116', '98', '23', '115', '122',
             '99', '88', '100', '6', '102', '93', '95', '85', '83', '113', '13',
             '182', '48', '145', '87', '92', '80', '117', '128', '119', '143',
             '114', '118', '108', '63', '121', '142', '154', '120', '82', '109',
             '101', '86', '229', '76', '89', '156', '112', '107', '129', '135',
             '136', '165', '150', '133', '70', '84', '140', '78', '7', '64',
             '59', '139', '69', '148', '189', '141', '130', '138', '81', '132',
             '10', '123', '65', '68', '66', '62', '74', '131', '39', '46', '38',
             '8', '17', '126', '155', '159', '137', '12', '273', '36', '34',
             '77', '60', '49', '58', '72', '204', '212', '25', '73', '29', '47',
             '32', '35', '71', '149', '33', '15', '54', '224', '162', '37',
             '75', '79', '55', '158', '164', '173', '181', '185', '21', '24',
             '51', '151', '42', '22', '134', '177', '52', '14', '53', '57',
             '28', '50', '26', '45', '171', '27', '44', '146', '20', '157',
             '203', '41', '30', '194', '233', '237', '230', '195', '253', '152',
             '190', '160', '208', '180', '144', '174', '170', '192', '209',
             '187', '172', '16', '186', '11', '193', '176', '56', '169', '40',
             '168', '312', '153', '214', '31', '163', '19', '179', '43', '200',
             '196', '167', '178', '228', '18', '205', '201', '191'],
            dtype=object)
```

```
[32]: netflix_df_explode1= netflix_df_explode.copy()
      netflix_df_explode1.head()
```

```
[32]:   show_id     type                 title          director           cast  \
      0      s1    Movie  Dick Johnson Is Dead   Kirsten Johnson  Unknown Actors
      1      s2  TV Show         Blood & Water  Unknown Directors      Ama Qamata
      2      s2  TV Show         Blood & Water  Unknown Directors      Ama Qamata
      3      s2  TV Show         Blood & Water  Unknown Directors      Ama Qamata
      4      s2  TV Show         Blood & Water  Unknown Directors     Khosi Ngema

               country          date_added  release_year rating duration  \
      0  United States  September 25, 2021          2020  PG-13       90
      1   South Africa  September 24, 2021          2021  TV-MA        2
      2   South Africa  September 24, 2021          2021  TV-MA        2
      3   South Africa  September 24, 2021          2021  TV-MA        2
      4   South Africa  September 24, 2021          2021  TV-MA        2

                      listed_in                                       description
      0          Documentaries  As her father nears the end of his life, filmm…
      1  International TV Shows  After crossing paths at a party, a Cape Town t…
      2              TV Dramas  After crossing paths at a party, a Cape Town t…
      3            TV Mysteries  After crossing paths at a party, a Cape Town t…
      4  International TV Shows  After crossing paths at a party, a Cape Town t…
```

[33]:
```
netflix_df_explode['duration'].describe()
```

[33]:
```
count        202065
unique          210
top        1 Season
freq          35035
Name: duration, dtype: object
```

[34]:
```
netflix_df_explode['duration'].value_counts()
netflix_df_explode['duration']=netflix_df_explode['duration'].str.replace("␣
  ↪Season","")
netflix_df_explode['duration'].unique()
```

[34]:
```
array(['90', '2', '1', '91', '125', '9', '104', '127', '4', '67', '94',
       '5', '161', '61', '166', '147', '103', '97', '106', '111', '3',
       '110', '105', '96', '124', '116', '98', '23', '115', '122', '99',
       '88', '100', '6', '102', '93', '95', '85', '83', '113', '13',
       '182', '48', '145', '87', '92', '80', '117', '128', '119', '143',
       '114', '118', '108', '63', '121', '142', '154', '120', '82', '109',
       '101', '86', '229', '76', '89', '156', '112', '107', '129', '135',
       '136', '165', '150', '133', '70', '84', '140', '78', '7', '64',
       '59', '139', '69', '148', '189', '141', '130', '138', '81', '132',
       '10', '123', '65', '68', '66', '62', '74', '131', '39', '46', '38',
       '8', '17', '126', '155', '159', '137', '12', '273', '36', '34',
       '77', '60', '49', '58', '72', '204', '212', '25', '73', '29', '47',
       '32', '35', '71', '149', '33', '15', '54', '224', '162', '37',
       '75', '79', '55', '158', '164', '173', '181', '185', '21', '24',
       '51', '151', '42', '22', '134', '177', '52', '14', '53', '57',
       '28', '50', '26', '45', '171', '27', '44', '146', '20', '157',
       '203', '41', '30', '194', '233', '237', '230', '195', '253', '152',
       '190', '160', '208', '180', '144', '174', '170', '192', '209',
       '187', '172', '16', '186', '11', '193', '176', '56', '169', '40',
       '168', '312', '153', '214', '31', '163', '19', '179', '43', '200',
       '196', '167', '178', '228', '18', '205', '201', '191'],
      dtype=object)
```

[35]:

```
netflix_df_explode['duration'].describe()
```

[35]:

```
count      202065
unique        210
top             1
freq        35035
Name: duration, dtype: object
```

As we had seen that one of value was mentiones as 'Season' that has now been replaced with numerical value 1.

[36]:

```
netflix_df_explode1['duration without season'] = netflix_df_explode['duration'].
  ↪copy()
netflix_df_explode1.loc[netflix_df_explode1['duration without season'].str.
  ↪contains('Season'),'duration without season'] = 0
netflix_df_explode1['duration without season'] = netflix_df_explode1['duration␣
  ↪without season'].astype(int)
netflix_df_explode1['duration without season'].describe()
```

[36]:

```
count    202065.000000
mean         77.687828
std          51.481723
min           1.000000
25%           4.000000
50%          95.000000
75%         112.000000
max         312.000000
Name: duration without season, dtype: float64
```

4. Visual Analysis - Univariate, Bivariate after pre-processing of the data

[37]:

```
netflix_df_explode1.head()
```

[37]:

```
  show_id     type                 title           director           cast  \
0      s1    Movie  Dick Johnson Is Dead    Kirsten Johnson  Unknown Actors
1      s2  TV Show         Blood & Water  Unknown Directors      Ama Qamata
2      s2  TV Show         Blood & Water  Unknown Directors      Ama Qamata
3      s2  TV Show         Blood & Water  Unknown Directors      Ama Qamata
4      s2  TV Show         Blood & Water  Unknown Directors     Khosi Ngema

         country          date_added  release_year rating duration  \
0  United States  September 25, 2021          2020  PG-13       90
1   South Africa  September 24, 2021          2021  TV-MA        2
2   South Africa  September 24, 2021          2021  TV-MA        2
3   South Africa  September 24, 2021          2021  TV-MA        2
4   South Africa  September 24, 2021          2021  TV-MA        2
```

```
                listed_in                                        description  \
0           Documentaries  As her father nears the end of his life, filmm…
1  International TV Shows  After crossing paths at a party, a Cape Town t…
2               TV Dramas  After crossing paths at a party, a Cape Town t…
3            TV Mysteries  After crossing paths at a party, a Cape Town t…
4  International TV Shows  After crossing paths at a party, a Cape Town t…

   duration without season
0                       90
1                        2
2                        2
3                        2
4                        2
```

[38]:

```
from datetime import datetime
from dateutil.parser import parse
import pandas as pd

arr = []
for i in netflix_df_explode1['date_added'].values:
    dt1 = parse(i)
    arr.append(dt1.strftime('%Y-%m-%d'))

# Convert 'Modified_Added_date' to datetime
netflix_df_explode1['Modified_Added_date'] = pd.to_datetime(arr)

# Extract month, week, and year
netflix_df_explode1['month_added'] = netflix_df_explode1['Modified_Added_date'].
 ↪dt.month
netflix_df_explode1['week_Added'] = netflix_df_explode1['Modified_Added_date'].
 ↪dt.isocalendar().week
netflix_df_explode1['year'] = netflix_df_explode1['Modified_Added_date'].dt.year

# Display the first 5 rows
netflix_df_explode1.head(5)
```

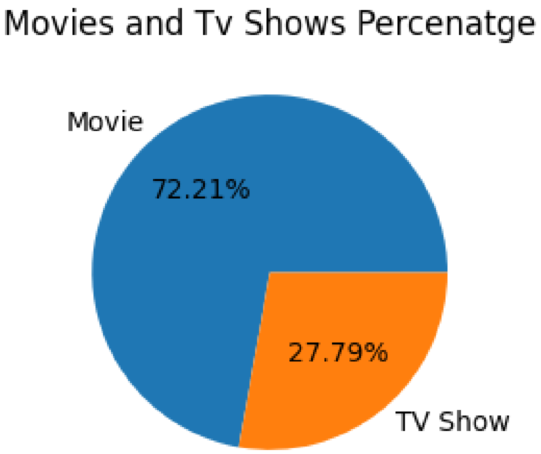
[38]:

```
  show_id     type                 title           director            cast  \
0      s1    Movie  Dick Johnson Is Dead    Kirsten Johnson  Unknown Actors
1      s2  TV Show         Blood & Water  Unknown Directors      Ama Qamata
2      s2  TV Show         Blood & Water  Unknown Directors      Ama Qamata
3      s2  TV Show         Blood & Water  Unknown Directors      Ama Qamata
4      s2  TV Show         Blood & Water  Unknown Directors     Khosi Ngema

         country          date_added  release_year rating duration  \
0  United States  September 25, 2021          2020  PG-13       90
1   South Africa  September 24, 2021          2021  TV-MA        2
```

```
2   South Africa  September 24, 2021         2021  TV-MA        2
3   South Africa  September 24, 2021         2021  TV-MA        2
4   South Africa  September 24, 2021         2021  TV-MA        2

                listed_in                                        description  \
0           Documentaries  As her father nears the end of his life, filmm…
1  International TV Shows  After crossing paths at a party, a Cape Town t…
2               TV Dramas  After crossing paths at a party, a Cape Town t…
3            TV Mysteries  After crossing paths at a party, a Cape Town t…
4  International TV Shows  After crossing paths at a party, a Cape Town t…

   duration without season Modified_Added_date  month_added  week_Added  year
0                       90          2021-09-25            9          38  2021
1                        2          2021-09-24            9          38  2021
2                        2          2021-09-24            9          38  2021
3                        2          2021-09-24            9          38  2021
4                        2          2021-09-24            9          38  2021
```

Univariate Analysis

[39]:
```python
import matplotlib.pyplot as plt
# Now we will proceed with comparison of TVshows vs. movies.
plt.figure(figsize=(5,3))
plt.title("Movies and Tv Shows Percenatge")
plt.pie(netflix_df_explode1.type.value_counts(),
        labels=netflix_df_explode1.type.value_counts().index,
        autopct='%.2f%%')
plt.show()
```



from overall data of netflix we can observe that Movies contribute 72.21% and TV shows contribute

27.79%

[40]:
```
# Convert 'date_added' to datetime, handling errors by coercing invalid formats
netflix_df_explode1["Modified_Added_date"] = pd.
 ↪to_datetime(netflix_df_explode1["date_added"], errors='coerce')

# Extract the year from the 'Modified_Added_date' column
netflix_df_explode1["year_added"] = netflix_df_explode1["Modified_Added_date"].
 ↪dt.year

# Display the year_added column
netflix_df_explode1["year_added"]
```

[40]:
```
0         2021.0
1         2021.0
2         2021.0
3         2021.0
4         2021.0
           …
202060    2019.0
202061    2019.0
202062    2019.0
202063    2019.0
202064    2019.0
Name: year_added, Length: 202065, dtype: float64
```

[41]:
```
#now checking for the count of movies every year
# Filter the movies and count the number of movies added by year
netflix_df_explode_movies = netflix_df_explode1[netflix_df_explode1.type ==␣
 ↪'Movie']
df_movies = netflix_df_explode_movies.year_added.value_counts().reset_index()

# Rename the columns correctly
df_movies = df_movies.rename(columns={"year_added": "year"})

df_movies['year'] = df_movies['year'].astype(int)

# Display the result
df_movies
```

[41]:
```
   year  count
0  2019  34473
1  2020  32488
2  2018  28050
3  2021  25709
4  2017  18252
5  2016   4858
```

```
6   2015   1125
7   2011    438
8   2014    345
9   2013     75
10  2012     36
11  2009     30
12  2010     20
13  2008     18
```

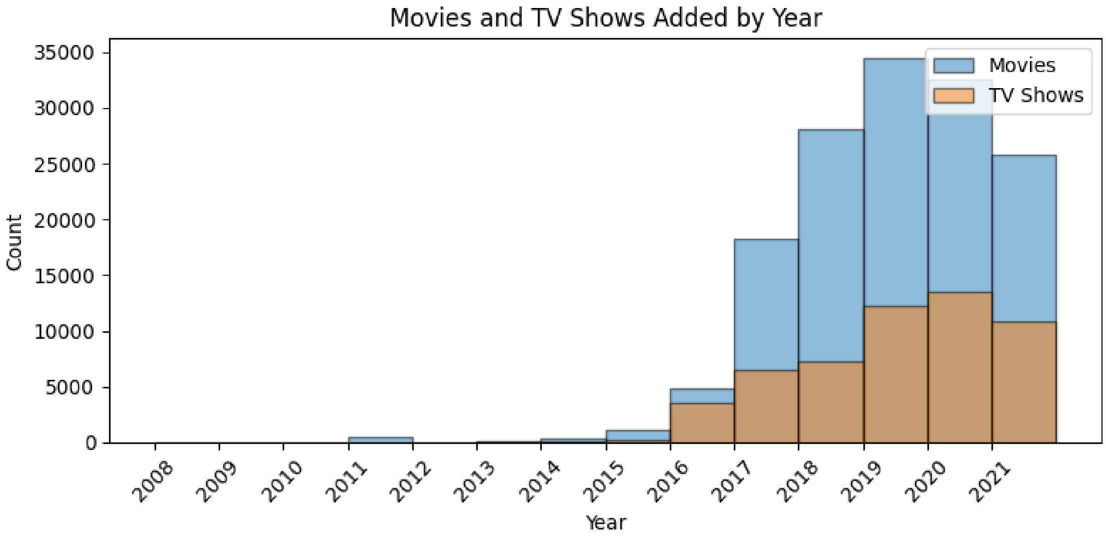
[42]:
```python
# Filter the TV Show and count the number of TV show added by year
netflix_df_explode_tvshow = netflix_df_explode1[netflix_df_explode1.type == 'TV␣
  ↪Show']
df_tvshow = netflix_df_explode_tvshow.year_added.value_counts().reset_index()

# Rename the columns correctly
df_tvshow = df_tvshow.rename(columns={"year_added": "year"})
df_tvshow['year'] = df_tvshow['year'].astype(int)
# Display the result
df_tvshow
```

[42]:
```
   year  count
0  2020  13545
1  2019  12272
2  2021  10850
3  2018   7317
4  2017   6555
5  2016   3574
6  2015    232
7  2013    110
8  2014    104
9  2008      1
```

4.1 Histogram For continuous variable(s)

[43]:
```python
import matplotlib.pyplot as plt



# Plot histogram for Movies and TV Shows on the same plot
plt.figure(figsize=(8, 4))

# Histogram for Movies
plt.hist(df_movies['year'], weights=df_movies['count'],␣
  ↪bins=range(df_movies['year'].min(), df_movies['year'].max() + 2), alpha=0.5,␣
  ↪label='Movies', edgecolor='black')

# Histogram for TV Shows
```

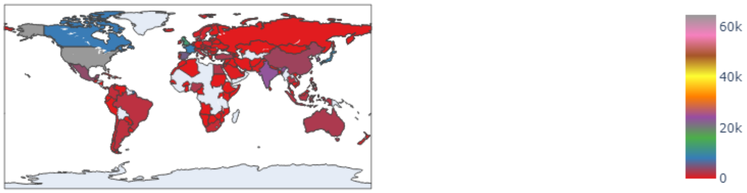
```
plt.hist(df_tvshow['year'], weights=df_tvshow['count'],␣
  ↪bins=range(df_tvshow['year'].min(), df_tvshow['year'].max() + 2), alpha=0.5,␣
  ↪label='TV Shows', edgecolor='black')

plt.title('Movies and TV Shows Added by Year')
plt.xlabel('Year')
plt.ylabel('Count')
plt.legend(loc='upper right')
plt.xticks(range(min(df_movies['year'].min(), df_tvshow['year'].min()),␣
  ↪max(df_movies['year'].max(), df_tvshow['year'].max()) + 1), rotation=45)
plt.tight_layout()
plt.show()
```



4.2 Boxplot For categorical variable(s)

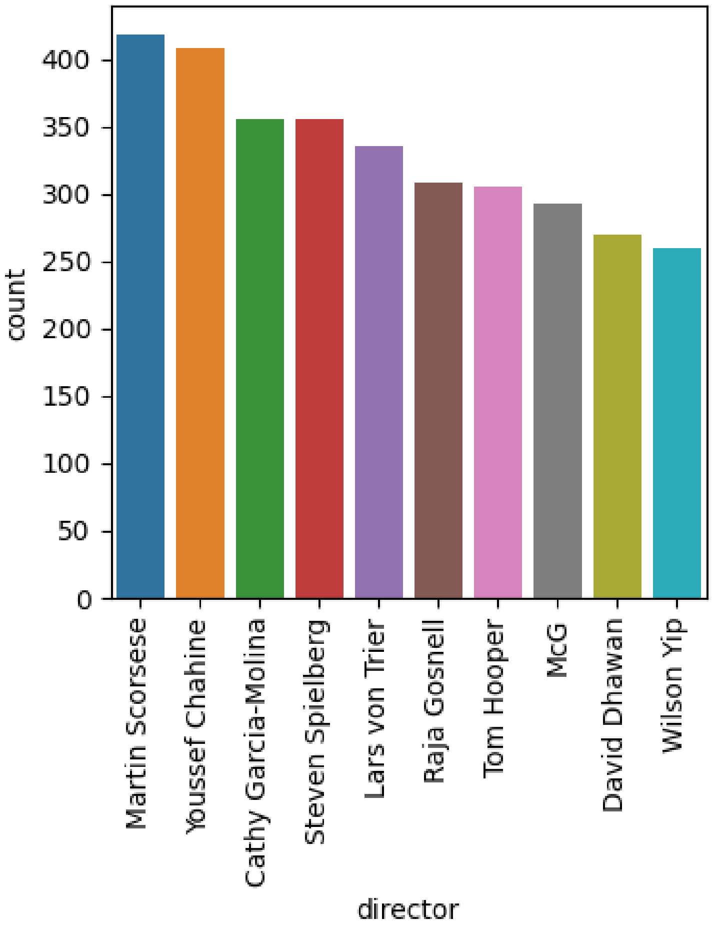
[44]:
```
import plotly.graph_objects as go
from plotly.offline import init_notebook_mode, iplot
import plotly.express as px

country_counts = netflix_df_explode1['country'].str.strip().value_counts()
country_counts=country_counts[country_counts!='country unavailable']
# Get the top 5 countries
top_5_countries = country_counts
top_5_countries
# # Create a custom color scale
colors = px.colors.qualitative.Set1

# Create a Choropleth plot with distinct colors
fig = go.Figure()
```

```
fig.add_trace(go.Choropleth(
    locationmode='country names',
    locations=top_5_countries.index,
    z=top_5_countries.values,
    colorscale=colors,
    showscale=True
))

# Customize the layout
fig.update_geos(showcoastlines=True)  # Show country boundaries
fig.update_coloraxes(colorbar_title="Count")
```



[46]:
```
import seaborn as sns
df_netflix_country = netflix_df_explode1[netflix_df_explode1.director!='Unknown␣
 ↪Directors']
# df_netflix_country.director.value_counts()

plt.figure(figsize=(4,4))
sns.countplot(x=df_netflix_country.director,order=df_netflix_country.director.
 ↪value_counts().index[:10])
plt.xticks(rotation=90)
```

[46]:
```
(array([0, 1, 2, 3, 4, 5, 6, 7, 8, 9]),
 [Text(0, 0, 'Martin Scorsese'),
  Text(1, 0, 'Youssef Chahine'),
  Text(2, 0, 'Cathy Garcia-Molina'),
  Text(3, 0, 'Steven Spielberg'),
  Text(4, 0, 'Lars von Trier'),
  Text(5, 0, 'Raja Gosnell'),
  Text(6, 0, 'Tom Hooper'),
  Text(7, 0, 'McG'),
  Text(8, 0, 'David Dhawan'),
  Text(9, 0, 'Wilson Yip')])
```

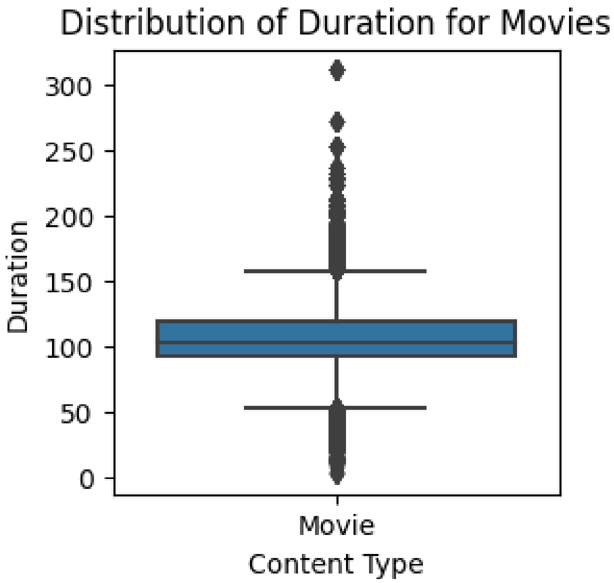


From the above we can see top ten 'directors' with max number of movies they did and ranges from 250+ to 400+

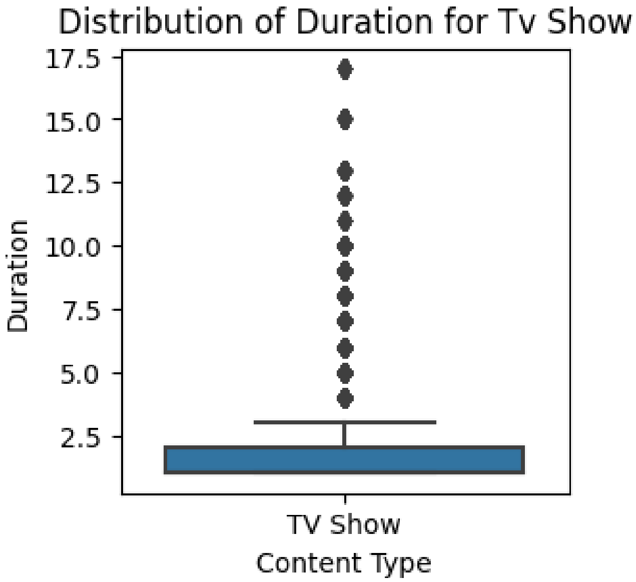
Bivariate Analysis- to understand relationship between two variables we willl proceed with bivariate analysis

For understanding this we will plot Box-plot for duration between TV show and movies and can watch outliers.

[47]:
```
#Using Box plot for undderstanding outliers for duration in movies
import seaborn as sns
plt.figure(figsize=(3,3))
sns.boxplot(data=netflix_df_explode_movies, x='type', y='duration without␣
 ↪season')
plt.xlabel('Content Type')
plt.ylabel('Duration')
plt.title('Distribution of Duration for Movies')
plt.show()
```



[48]:
```
#Using Box plot for undderstanding outliers for duration in TV Show
plt.figure(figsize=(3,3))
sns.boxplot(data=netflix_df_explode_tvshow, x='type', y='duration without␣
 ↪season')
plt.xlabel('Content Type')
plt.ylabel('Duration')
plt.title('Distribution of Duration for Tv Show')
plt.show()
```

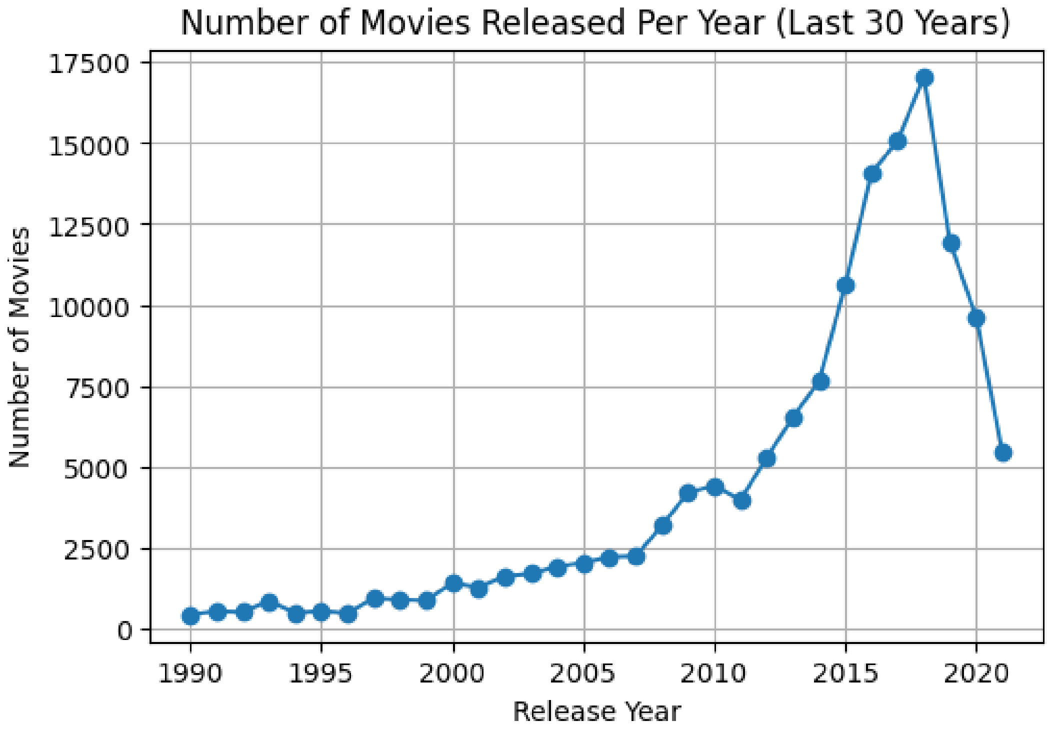


[49]:
```
netflix_df_explode1.head(2)
```

[49]:
```
  show_id     type                 title           director           cast  \
0      s1    Movie  Dick Johnson Is Dead    Kirsten Johnson  Unknown Actors
1      s2  TV Show         Blood & Water  Unknown Directors      Ama Qamata

        country          date_added  release_year rating duration  \
0  United States  September 25, 2021          2020  PG-13       90
1   South Africa  September 24, 2021          2021  TV-MA        2

               listed_in                                        description  \
0          Documentaries  As her father nears the end of his life, filmm…
1  International TV Shows  After crossing paths at a party, a Cape Town t…

  duration without season Modified_Added_date  month_added  week_Added  year  \
0                      90          2021-09-25            9          38  2021
1                       2          2021-09-24            9          38  2021

  year_added
0     2021.0
1     2021.0
```

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

[50]:
```
import matplotlib.pyplot as plt

# Filter data for movies
df_movies = netflix_df_explode1[netflix_df_explode1['type'] == 'Movie']

# Filter for movies released in the last 30 years (1990 onwards)
df_movies_last_30_years = df_movies[df_movies['release_year'] >= 1990]

# Group by release_year and count the number of movies released each year
movies_per_year = df_movies_last_30_years.groupby('release_year').size().
  ↪reset_index(name='count')

# Plot the trend of movies released per year
plt.figure(figsize=(6, 4))
plt.plot(movies_per_year['release_year'], movies_per_year['count'], marker='o')
plt.title('Number of Movies Released Per Year (Last 30 Years)')
plt.xlabel('Release Year')
plt.ylabel('Number of Movies')
plt.grid(True)
plt.show()
```



[ ]:

Comparison of tv shows vs. movies in last 20 years.

[51]:

```
# Find the most recent year in the dataset
most_recent_year = netflix_df_explode1['release_year'].max()

# Calculate the threshold year for the last 20 years
threshold_year = most_recent_year - 20

# Filter the dataset for movies and TV shows released in the last 20 years
df_last_20_years = netflix_df_explode1[netflix_df_explode1['release_year'] >=␣
 ↪threshold_year]

# Display the filtered data
df_last_20_years.head()
```

[51]:

```
  show_id     type                 title           director           cast  \
0      s1    Movie  Dick Johnson Is Dead   Kirsten Johnson  Unknown Actors
1      s2  TV Show         Blood & Water  Unknown Directors     Ama Qamata
2      s2  TV Show         Blood & Water  Unknown Directors     Ama Qamata
3      s2  TV Show         Blood & Water  Unknown Directors     Ama Qamata
4      s2  TV Show         Blood & Water  Unknown Directors    Khosi Ngema
```

```
         country          date_added  release_year rating duration  \
0  United States  September 25, 2021          2020  PG-13       90
1   South Africa  September 24, 2021          2021  TV-MA        2
2   South Africa  September 24, 2021          2021  TV-MA        2
3   South Africa  September 24, 2021          2021  TV-MA        2
4   South Africa  September 24, 2021          2021  TV-MA        2

                listed_in                                        description  \
0           Documentaries  As her father nears the end of his life, filmm…
1  International TV Shows  After crossing paths at a party, a Cape Town t…
2               TV Dramas  After crossing paths at a party, a Cape Town t…
3             TV Mysteries  After crossing paths at a party, a Cape Town t…
4  International TV Shows  After crossing paths at a party, a Cape Town t…

  duration without season Modified_Added_date  month_added  week_Added  year  \
0                      90          2021-09-25            9          38  2021
1                       2          2021-09-24            9          38  2021
2                       2          2021-09-24            9          38  2021
3                       2          2021-09-24            9          38  2021
4                       2          2021-09-24            9          38  2021

  year_added
0     2021.0
1     2021.0
2     2021.0
3     2021.0
4     2021.0
```

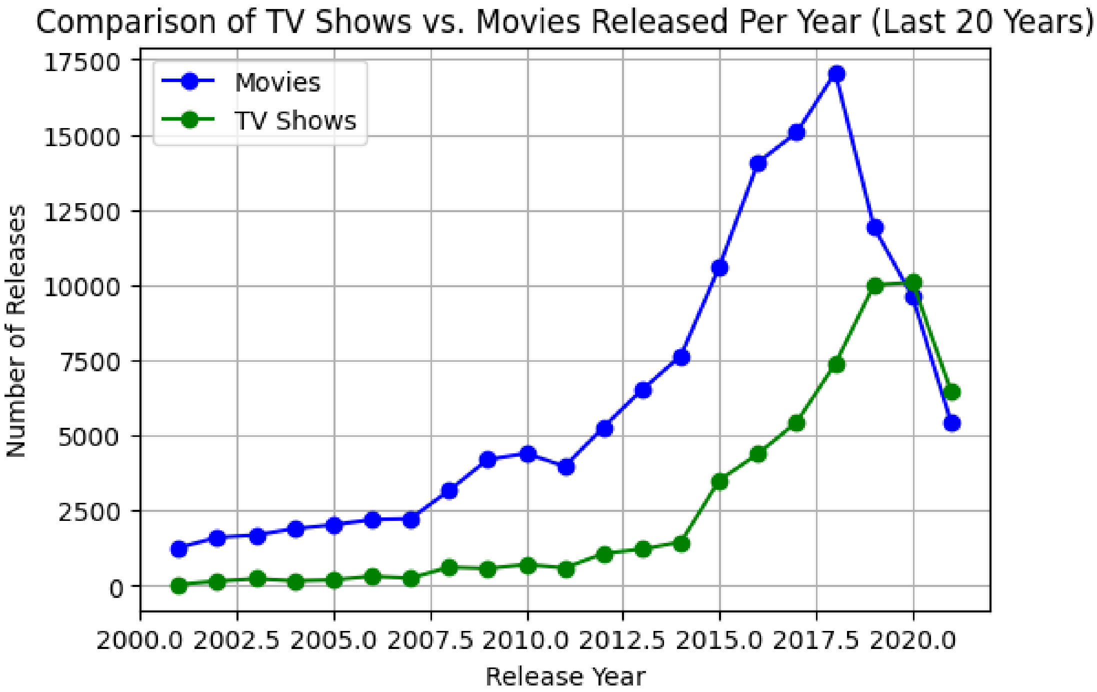
[52]:

```
import matplotlib.pyplot as plt



# Filter the data for movies and TV shows released in the last 20 years
df_last_20_years = netflix_df_explode1[netflix_df_explode1['release_year'] >=␣
 ↪threshold_year]
df_last_20_years['release_year'] = df_last_20_years['release_year'].round().
 ↪astype(int)
# Separate movies and TV shows
df_movies_last_20_years = df_last_20_years[df_last_20_years['type'] == 'Movie']
df_tvshows_last_20_years = df_last_20_years[df_last_20_years['type'] == 'TV␣
 ↪Show']

# Group by release_year and count the number of releases for both
movies_per_year = df_movies_last_20_years.groupby('release_year').size().
 ↪reset_index(name='count_movies')
tvshows_per_year = df_tvshows_last_20_years.groupby('release_year').size().
 ↪reset_index(name='count_tvshows')
```

```python
# Merge the two datasets on release_year
comparison_df = pd.merge(movies_per_year, tvshows_per_year, on='release_year',␣
 ↪how='outer').fillna(0)

# Plot the comparison
plt.figure(figsize=(6,4))
plt.plot(comparison_df['release_year'], comparison_df['count_movies'],␣
 ↪marker='o', label='Movies', color='blue')
plt.plot(comparison_df['release_year'], comparison_df['count_tvshows'],␣
 ↪marker='o', label='TV Shows', color='green')
plt.title('Comparison of TV Shows vs. Movies Released Per Year (Last 20 Years)')
plt.xlabel('Release Year')
plt.ylabel('Number of Releases')
plt.grid(True)
plt.legend()
plt.show
```

```
C:\Users\admin\AppData\Local\Temp\ipykernel_6444\1092556704.py:6:
SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

```
[52]: <function matplotlib.pyplot.show(close=None, block=None)>
```



```
[53]: netflix_df_explode1.head(2)
```

```
[53]:   show_id     type                  title            director            cast  \
      0      s1    Movie   Dick Johnson Is Dead     Kirsten Johnson  Unknown Actors
      1      s2  TV Show          Blood & Water  Unknown Directors       Ama Qamata

               country          date_added  release_year rating duration  \
      0  United States  September 25, 2021          2020  PG-13       90
      1   South Africa  September 24, 2021          2021  TV-MA        2

                     listed_in                                        description  \
      0          Documentaries  As her father nears the end of his life, filmm…
      1  International TV Shows  After crossing paths at a party, a Cape Town t…

         duration without season Modified_Added_date  month_added  week_Added  year  \
      0                       90          2021-09-25            9          38  2021
      1                        2          2021-09-24            9          38  2021

         year_added
      0      2021.0
      1      2021.0
```

What is the best time to launch a TV show?

[54]:

```
netflix_data = netflix_df_explode1.copy()
netflix_data.head(2)
```

[54]:

```
  show_id     type                 title           director            cast  \
0      s1    Movie  Dick Johnson Is Dead    Kirsten Johnson  Unknown Actors
1      s2  TV Show         Blood & Water  Unknown Directors      Ama Qamata

         country          date_added  release_year rating duration  \
0  United States  September 25, 2021          2020  PG-13       90
1   South Africa  September 24, 2021          2021  TV-MA        2

                listed_in                                        description  \
0           Documentaries  As her father nears the end of his life, filmm…
1  International TV Shows  After crossing paths at a party, a Cape Town t…

   duration without season Modified_Added_date  month_added  week_Added  year  \
0                       90          2021-09-25            9          38  2021
1                        2          2021-09-24            9          38  2021

   year_added
0      2021.0
1      2021.0
```

[55]:

```
# Convert 'date_added' column to datetime format
netflix_data['date_added'] = pd.to_datetime(netflix_data['date_added'],␣
 ↪errors='coerce')

# Filter for TV Shows
tv_shows = netflix_data[netflix_data['type'] == 'TV Show']

# Extract the month and day of the week from the 'date_added' column
tv_shows['month_added'] = tv_shows['date_added'].dt.month
tv_shows['day_of_week_added'] = tv_shows['date_added'].dt.day_name()

# Count the number of shows added per month
month_counts = tv_shows['month_added'].value_counts().sort_index()

# Count the number of shows added per day of the week
day_counts = tv_shows['day_of_week_added'].value_counts()

# Print the results
print("TV Shows Released by Month:")
print(month_counts)

print("\nTV Shows Released by Day of the Week:")
print(day_counts)
```

```
TV Shows Released by Month:
month_added
1.0     3941
2.0     3786
3.0     4201
4.0     4487
5.0     4111
6.0     4959
7.0     5211
8.0     5053
9.0     4842
10.0    4199
11.0    4429
12.0    5341
Name: count, dtype: int64

TV Shows Released by Day of the Week:
day_of_week_added
Friday       20890
Thursday      8431
Tuesday       6385
Wednesday     6279
Saturday      5312
Monday        4138
Sunday        3125
Name: count, dtype: int64
```

```
C:\Users\admin\AppData\Local\Temp\ipykernel_6444\1560419489.py:8:
SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

C:\Users\admin\AppData\Local\Temp\ipykernel_6444\1560419489.py:9:
SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

From the above analysis we can see tha in the month of December there were highest release and

Friday most of the TV shows had been released

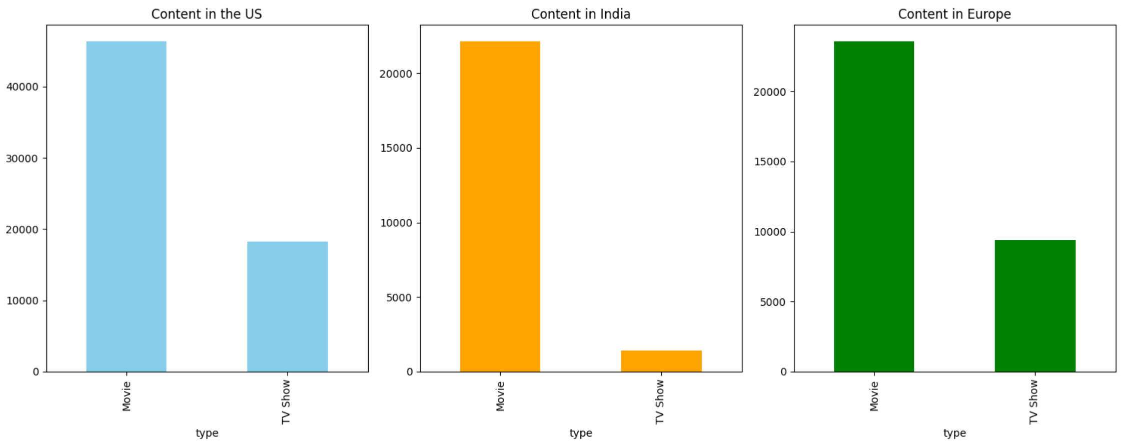
```
[56]: netflix_data1 = netflix_df_explode1.copy()
      netflix_data1.head(2)
```

```
[56]:   show_id     type                 title             director            cast  \
      0      s1    Movie  Dick Johnson Is Dead     Kirsten Johnson  Unknown Actors
      1      s2  TV Show         Blood & Water  Unknown Directors       Ama Qamata

               country          date_added  release_year rating duration  \
      0  United States  September 25, 2021          2020  PG-13       90
      1   South Africa  September 24, 2021          2021  TV-MA        2

                    listed_in                                        description  \
      0           Documentaries  As her father nears the end of his life, filmm…
      1  International TV Shows  After crossing paths at a party, a Cape Town t…

         duration without season Modified_Added_date  month_added  week_Added  year  \
      0                       90          2021-09-25            9          38  2021
      1                        2          2021-09-24            9          38  2021

         year_added
      0      2021.0
      1      2021.0
```

Understanding what content is available in different countries

```
[57]: # Filter content by country
      # US, India, and selected European countries
      us_content = netflix_data1[netflix_data1['country'].str.contains('United␣
       ↪States', na=False)]
      india_content = netflix_data1[netflix_data1['country'].str.contains('India',␣
       ↪na=False)]
      europe_content = netflix_data1[netflix_data1['country'].str.contains('United␣
       ↪Kingdom|France|Germany|Spain|Italy', na=False)]

      # Compare the number of TV shows and movies in each region
      us_count = us_content['type'].value_counts()
      india_count = india_content['type'].value_counts()
      europe_count = europe_content['type'].value_counts()

      # Visualize the content distribution in each region using bar plots
      fig, axes = plt.subplots(1, 3, figsize=(15, 6))

      us_count.plot(kind='bar', ax=axes[0], title='Content in the US',␣
       ↪color='skyblue')
```

```
india_count.plot(kind='bar', ax=axes[1], title='Content in India',␣
  ↪color='orange')
europe_count.plot(kind='bar', ax=axes[2], title='Content in Europe',␣
  ↪color='green')

plt.tight_layout()
plt.show()
```



From the above we can see that in India TV show are less as compared to US and European countries

Business Insights for Netflix

The US typically has the largest variety of content, including both TV shows and movies. This indicates that Netflix can prioritize the US market with diverse content to cater to different audiences.

Across all regions, there is a clear demand for both TV shows and movies, but the US and Europe show a slightly stronger preference for TV shows, while India leans more toward movies. To grow, Netflix can increase its catalog of TV shows in India and boost original movie production in the US and Europe.

Holiday seasons like December month and weekends are ideal for launching new TV shows, especially in the US and Europe, where viewers have more free time to engage with new content.

[ ]:

[ ]:

[ ]: