netflix-project

February 5, 2024

##**Business Case: Netflix - Data Exploration and Visualisation**

Mindset

Evaluation will be kept lenient, so make sure you attempt this case study. Read the question carefully and try to understand what exactly is being asked. Brainstorm a little. If you're getting an error, remember that Google is your best friend. You can watch the lecture recordings or go through your lecture notes once again if you feel like you're getting confused over some specific topics. Discuss your problems with your peers. Make use of the Slack channel and WhatsApp group. Only if you think that there's a major issue, you can reach out to your Instructor via Slack or Email. There is no right or wrong answer. We have to get used to dealing with uncertainty in business. This is exactly the skill we want to develop. 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.

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

Show_id: Unique ID for every Movie / Tv Show Type: Identifier - A Movie or TV Show Title: Title of the Movie / Tv Show Director: Director of the Movie Cast: Actors involved in the movie/show Country: Country where the movie/show was produced Date_added: Date it was added on Netflix Release_year: Actual Release year of the movie/show Rating: TV Rating of the movie/show Duration: Total Duration - in minutes or number of seasons Listed_in: Genre Description: The summary description

#Importing Libraries

[1]:

```
#Importing Libraries
import warnings
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import io
warnings.filterwarnings('ignore')
```

#Loading Data into Colab

[2]:

```
#For Import the file
from google.colab import files
uploaded = files.upload()
```

```
<IPython.core.display.HTML object>
```

```
Saving Scalar_DA_Project.csv to Scalar_DA_Project.csv
```

[3]:

```
#Import the file and save as netflix
netflix = pd.read_csv(io.StringIO(uploaded['Scalar_DA_Project.csv'].
 ↪decode('utf-8')))
```

Inspecting first few rows of dataset

1 Q1. Defining Problem Statement and Analysing basic metrics (10 Points)

[4]:

```
#To check the head values of the dataset
netflix.head()
```

[4]:

```
  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…
```

Basic Dataset checkings:

```
[5]: #Check the size of the Dataset
     netflix.size
```

```
[5]: 105684
```

```
[6]: #Check the shape of the dataset
     netflix.shape
```

```
[6]: (8807, 12)
```

```
[7]: #Check the columns of the dataset
     netflix.columns
```

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

2 Q2. Observations on the shape of data, data types of all the attributes, conversion of categorical attributes to 'category' (If required), missing value detection, statistical summary (10 Points)

```
[8]: #Check the data types of the dataset
     netflix.dtypes
```

```
[8]: 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
```

[9]:

```
#Check the info of the dataset
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
```

3 Q3. Non-Graphical Analysis: Value counts and unique attributes

Checking the null values:

[10]:

```
mode_values = netflix.mode().iloc[0]
mode_values
```

[10]:

```
show_id                                                        s1
type                                                        Movie
title                                                      #Alive
director                                            Rajiv Chilaka
cast                                           David Attenborough
country                                             United States
date_added                                        January 1, 2020
release_year                                               2018.0
rating                                                      TV-MA
duration                                                 1 Season
listed_in                             Dramas, International Movies
description     Paranormal activity at a lush, abandoned prope…
```

```
Name: 0, dtype: object
```

[11]:

```
#Check the null values
netflix.isnull()
```

[11]:

```
      show_id   type  title  director   cast  country  date_added  \
0       False  False  False     False   True    False       False
1       False  False  False      True  False    False       False
2       False  False  False     False  False     True       False
3       False  False  False      True   True     True       False
4       False  False  False      True  False    False       False
…         …      …      …         …      …        …
8802    False  False  False     False  False    False       False
8803    False  False  False      True   True     True       False
8804    False  False  False     False  False    False       False
8805    False  False  False     False  False    False       False
8806    False  False  False     False  False    False       False

      release_year  rating  duration  listed_in  description
0            False   False     False      False        False
1            False   False     False      False        False
2            False   False     False      False        False
3            False   False     False      False        False
4            False   False     False      False        False
…              …       …        …          …          …
8802         False   False     False      False        False
8803         False   False     False      False        False
8804         False   False     False      False        False
8805         False   False     False      False        False
8806         False   False     False      False        False

[8807 rows x 12 columns]
```

[12]:

```
#Checking for unique values
netflix.nunique()
```

[12]:

```
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
```

[13]:
```
#Check the columns has null or not
null_content_check = netflix.isnull().all()
null_content_check
```

[13]:
```
show_id         False
type            False
title           False
director        False
cast            False
country         False
date_added      False
release_year    False
rating          False
duration        False
listed_in       False
description     False
dtype: bool
```

[14]:
```
#Total null value count in each column
null_count = netflix.isnull().sum()
null_count
```

[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]:
```
#Total available datas in each row
total_available_data = netflix.count()
total_available_data
```

[15]:
```
show_id         8807
type            8807
title           8807
director        6173
```

```
cast            7982
country         7976
date_added      8797
release_year    8807
rating          8803
duration        8804
listed_in       8807
description     8807
dtype: int64
```

[16]:

```
#Total Null values
total_null_values = netflix.isnull().values.sum()
total_null_values
```

[16]:

```
4307
```

##Data Preprocessing

Unique Values of each column

[17]:

```
#Checking the unique values in each row
for i in netflix.columns:
  print(f'{i} has {netflix[i].nunique()} unique values')
```

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

[18]:

```
# Inspecting Null values in the date_added, rating and duration columns

netflix[(netflix.rating.isnull()) | (netflix.duration.isnull())]
```

[18]:

```
     show_id     type                                             title  \
5541   s5542    Movie                                   Louis C.K. 2017
5794   s5795    Movie                             Louis C.K.: Hilarious
5813   s5814    Movie              Louis C.K.: Live at the Comedy Store
5989   s5990    Movie  13TH: A Conversation with Oprah Winfrey & Ava …
6827   s6828  TV Show                  Gargantia on the Verdurous Planet
7312   s7313  TV Show                                      Little Lunch
```

```
7537   s7538    Movie                               My Honor Was Loyalty

             director                                               cast  \
5541       Louis C.K.                                         Louis C.K.
5794       Louis C.K.                                         Louis C.K.
5813       Louis C.K.                                         Louis C.K.
5989              NaN                       Oprah Winfrey, Ava DuVernay
6827              NaN  Kaito Ishikawa, Hisako Kanemoto, Ai Kayano, Ka…
7312              NaN  Flynn Curry, Olivia Deeble, Madison Lu, Oisín …
7537  Alessandro Pepe  Leone Frisa, Paolo Vaccarino, Francesco Miglio…

            country          date_added  release_year  rating  duration  \
5541  United States       April 4, 2017          2017  74 min       NaN
5794  United States  September 16, 2016          2010  84 min       NaN
5813  United States     August 15, 2016          2015  66 min       NaN
5989            NaN    January 26, 2017          2017     NaN    37 min
6827          Japan    December 1, 2016          2013     NaN  1 Season
7312      Australia    February 1, 2018          2015     NaN  1 Season
7537          Italy       March 1, 2017          2015     NaN   115 min

                                 listed_in  \
5541                                Movies
5794                                Movies
5813                                Movies
5989                                Movies
6827  Anime Series, International TV Shows
7312                 Kids' TV, TV Comedies
7537                                Dramas

                                            description
5541  Louis C.K. muses on religion, eternal love, gi…
5794  Emmy-winning comedy writer Louis C.K. brings h…
5813  The comic puts his trademark hilarious/thought…
5989  Oprah Winfrey sits down with director Ava DuVe…
6827  After falling through a wormhole, a space-dwel…
7312  Adopting a child's perspective, this show take…
7537  Amid the chaos and horror of World War II, a c…
```

Seems like 'duration' for show_id - 5542, 5543 and 5544 have been wrongly entered into 'rating' column. So we need to move it back to 'duration' column, making rating column empty/null for show ids - 5542, 5543 and 5544

[19]:
```
#Replcaing the Duration and Rating columns
netflix.loc[netflix["show_id"] == "s5542", "duration"] = '74 min'
netflix.loc[netflix["show_id"] == "s5795", "duration"] = '84 min'
netflix.loc[netflix["show_id"] == "s5814", "duration"] = '66 min'
netflix.loc[netflix["show_id"] == "s5542", "rating"] = np.nan
netflix.loc[netflix["show_id"] == "s5795", "rating"] = np.nan
netflix.loc[netflix["show_id"] == "s5814", "rating"] = np.nan
```

[20]:
```
# drop show id column, as that is not needed for the analysis
netflix.drop('show_id',axis=1, inplace = True)
```

[21]:
```
#Dataset after removing the show id
netflix.head()
```

[21]:
```
      type                  title         director  \
0    Movie   Dick Johnson Is Dead  Kirsten Johnson
1  TV Show          Blood & Water              NaN
2  TV Show              Ganglands  Julien Leclercq
3  TV Show  Jailbirds New Orleans              NaN
4  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…
```

[22]:
```
# Replace nan values in data_added with January 1,{release_year}
netflix['date_added']=netflix['date_added'].fillna('January 1, {}'.
 ↪format(str(netflix['release_year'].mode()[0])))
```

[23]:

```
#Dataset after adding the nan values of date added with the Jan 1 and release␣
 ↪year
netflix.head()
```

[23]:

```
      type                 title         director  \
0    Movie  Dick Johnson Is Dead  Kirsten Johnson
1  TV Show         Blood & Water              NaN
2  TV Show             Ganglands  Julien Leclercq
3  TV Show  Jailbirds New Orleans             NaN
4  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…
```

[24]:

```
# to check there are any nll values in the date added column
netflix['date_added'].isnull().sum()
```

[24]:

```
0
```

[25]:

```
#To create a new column release month from the date added column
netflix["release_month"] = netflix['date_added'].apply(lambda x: x.lstrip().
 ↪split(" ")[0])
```

[26]:
```
#The dataset after creating a new column release month
netflix.head()
```

[26]:
```
      type                  title         director  \
0    Movie   Dick Johnson Is Dead  Kirsten Johnson
1  TV Show          Blood & Water              NaN
2  TV Show              Ganglands  Julien Leclercq
3  TV Show  Jailbirds New Orleans              NaN
4  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 release_month
0  As her father nears the end of his life, filmm…     September
1  After crossing paths at a party, a Cape Town t…     September
2  To protect his family from a powerful drug lor…     September
3  Feuds, flirtations and toilet talk go down amo…     September
4  In a city of coaching centers known to train I…     September
```

[27]:
```
#To check the value counts of the rating column in dataset
netflix['rating'].value_counts()
```

[27]:
```
TV-MA       3207
TV-14       2160
TV-PG        863
R            799
PG-13        490
TV-Y7        334
```

```
TV-Y         307
PG           287
TV-G         220
NR            80
G             41
TV-Y7-FV       6
NC-17          3
UR             3
Name: rating, dtype: int64
```

```
[28]: #  Replace nan values in rating column with TV-MA as the TV-MA is more value in␣
       ↪the rating column
      netflix['rating'].replace(np.nan, 'TV-MA',inplace  = True)
```

```
[29]: #To check that there is no null values in the rating column
      netflix['rating'].isnull().sum()
```

```
[29]: 0
```

```
[30]: # to check there are any nll values in the date added column
      netflix['duration'].isnull().sum()
```

```
[30]: 0
```

```
[31]: #Total country counts in the dataset
      netflix['country'].value_counts()
```

```
[31]: United States                               2818
      India                                        972
      United Kingdom                               419
      Japan                                        245
      South Korea                                  199
                                                  …
      Romania, Bulgaria, Hungary                     1
      Uruguay, Guatemala                             1
      France, Senegal, Belgium                       1
      Mexico, United States, Spain, Colombia         1
      United Arab Emirates, Jordan                   1
      Name: country, Length: 748, dtype: int64
```

```
[32]: # Replace nan values in country with United States as united states has mostly␣
       ↪found in the dataset
      netflix['country'].replace(np.nan, 'United States',inplace  = True)
```

```
[33]: #Total country counts in the dataset after updating the United States for nan␣
       ↪in dataset
      netflix['country'].value_counts()
```

[33]:
```
United States                             3649
India                                      972
United Kingdom                             419
Japan                                      245
South Korea                                199
                                          …
Romania, Bulgaria, Hungary                   1
Uruguay, Guatemala                           1
France, Senegal, Belgium                     1
Mexico, United States, Spain, Colombia       1
United Arab Emirates, Jordan                 1
Name: country, Length: 748, dtype: int64
```

[34]:
```
#Only Director and Cast has the null values apart from that no null values in␣
 ↪other columns
netflix.isnull().sum()
```

[34]:
```
type                0
title               0
director         2634
cast              825
country             0
date_added          0
release_year        0
rating              0
duration            0
listed_in           0
description         0
release_month       0
dtype: int64
```

As we could see that the null values found in the Director and Cast and we can't able to fill that with the other details. Also for our analysis we are going to use the other details and not the director and cast. For this we are going to drop down the Director and cast columns form this.

[35]:
```
# Drop the director and cast columns completely.
netflix.drop(['director','cast'],axis=1, inplace = True)
```

[36]:
```
#The new dataset after removing all the null values and not needed columns
netflix.head()
```

[36]:
```
      type                  title        country          date_added  \
0    Movie   Dick Johnson Is Dead  United States  September 25, 2021
1  TV Show          Blood & Water   South Africa  September 24, 2021
2  TV Show              Ganglands  United States  September 24, 2021
3  TV Show  Jailbirds New Orleans  United States  September 24, 2021
```

```
4  TV Show           Kota Factory         India  September 24, 2021

   release_year rating   duration  \
0          2020  PG-13     90 min
1          2021  TV-MA  2 Seasons
2          2021  TV-MA   1 Season
3          2021  TV-MA   1 Season
4          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 release_month
0  As her father nears the end of his life, filmm…     September
1  After crossing paths at a party, a Cape Town t…     September
2  To protect his family from a powerful drug lor…     September
3  Feuds, flirtations and toilet talk go down amo…     September
4  In a city of coaching centers known to train I…     September
```

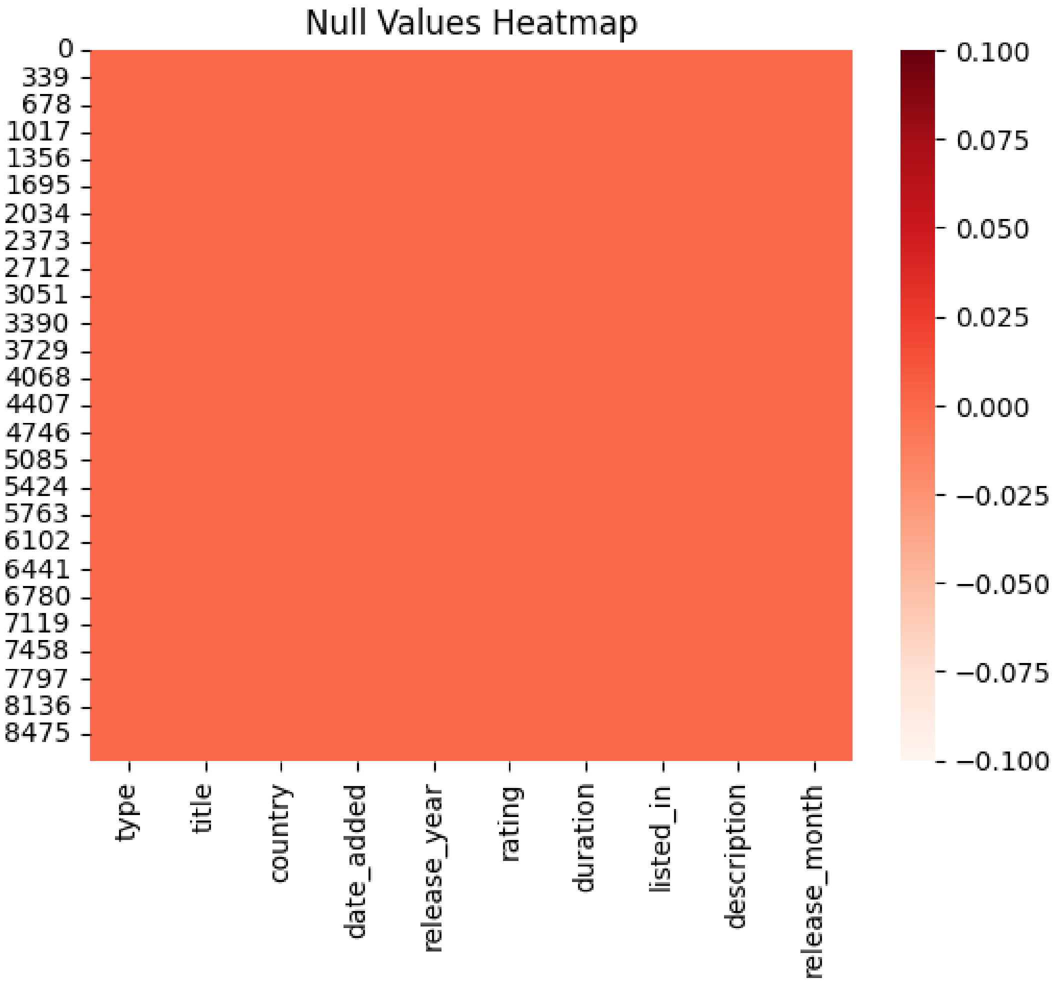
[37]:
```
#Clean dataset with unique and non null values
netflix.isnull().sum()
```

[37]:
```
type             0
title            0
country          0
date_added       0
release_year     0
rating           0
duration         0
listed_in        0
description      0
release_month    0
dtype: int64
```

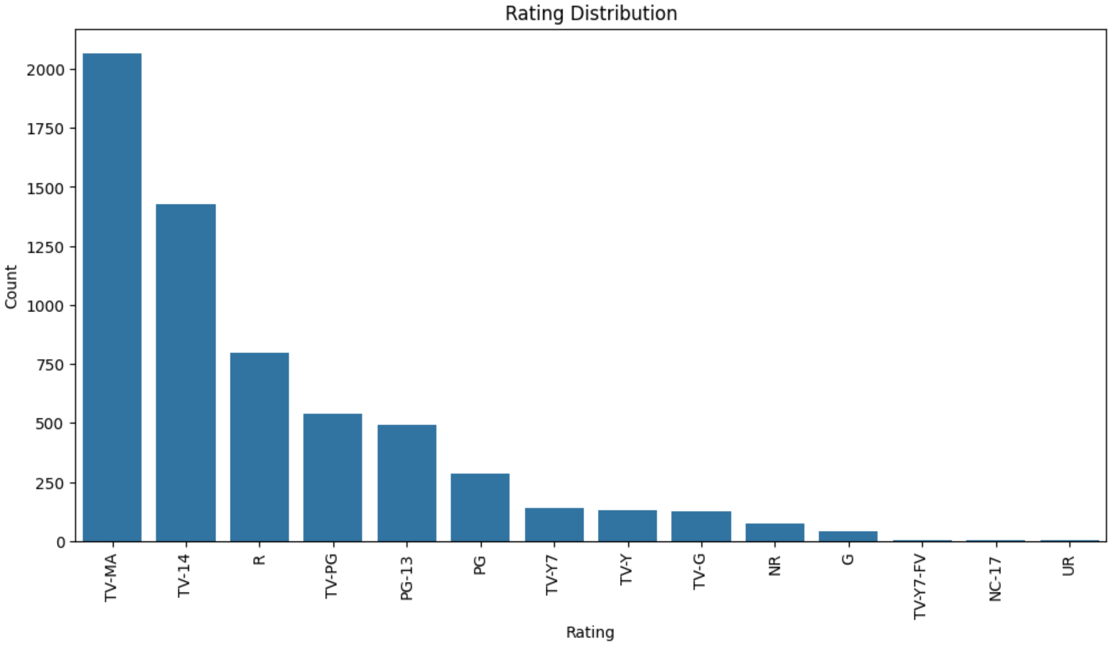
##It seems that there are no null values, duplicate values and missing values in the dataset. Now we can start using this for the analysis.

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

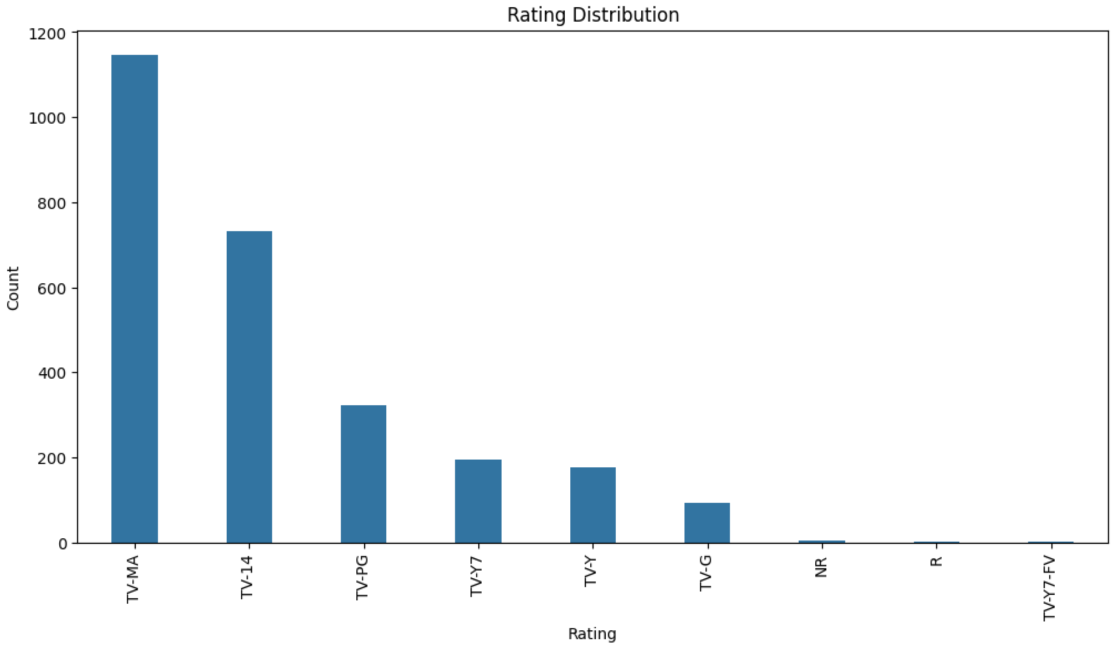
[38]:
```
#To check the null values in the heat map
sns.heatmap(netflix.isnull(),cmap = 'Reds')
plt.title('Null Values Heatmap')
plt.show()
```



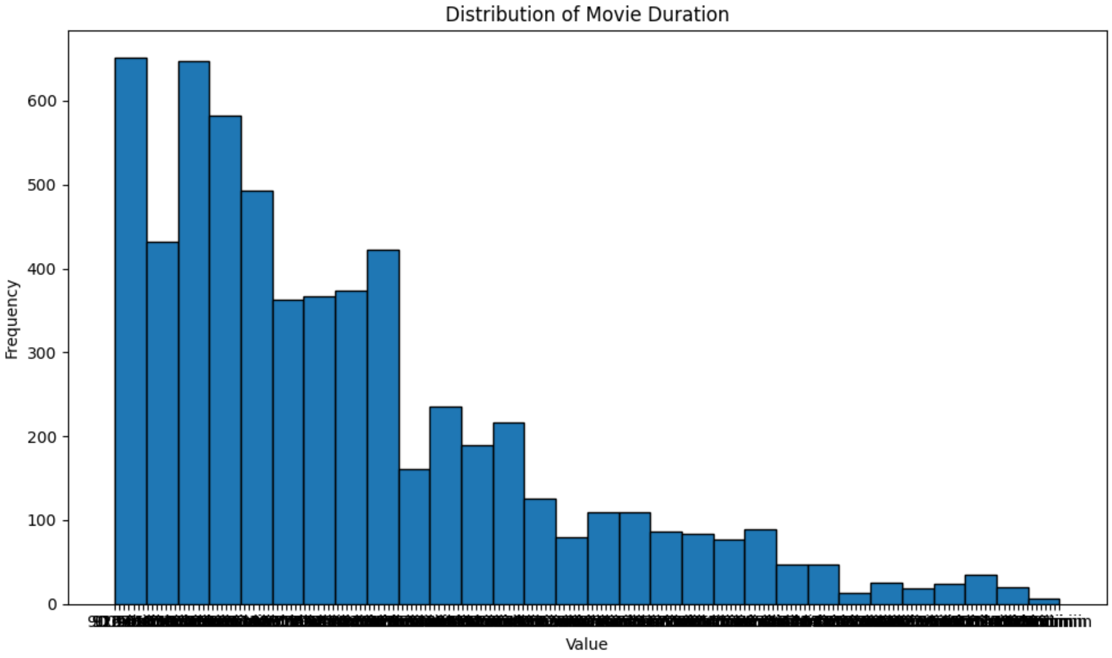
[39]:
```
genere=netflix[["title","rating", "type"]]
genere=genere.drop_duplicates()
movies_df = genere[genere['type'] == 'Movie']
tv_shows_df = genere[genere['type'] == 'TV Show']

movie_genre_counts = movies_df['rating'].value_counts()
# Plot the bar graph for movie genres
plt.figure(figsize=(12, 6))
sns.barplot(x=movie_genre_counts.index, y=movie_genre_counts.values)
plt.xticks(rotation=90)
plt.xlabel('Rating')
plt.ylabel('Count')
plt.title('Rating Distribution')
plt.show()
```



[40]:
```
tv_show_counts = tv_shows_df['rating'].value_counts()
# Plot the bar graph for movie genres
plt.figure(figsize=(12, 6))
sns.barplot(x=tv_show_counts.index, y=tv_show_counts.values, width=0.4)
plt.xticks(rotation=90)
plt.xlabel('Rating')
plt.ylabel('Count')
plt.title('Rating Distribution')
plt.show()
```

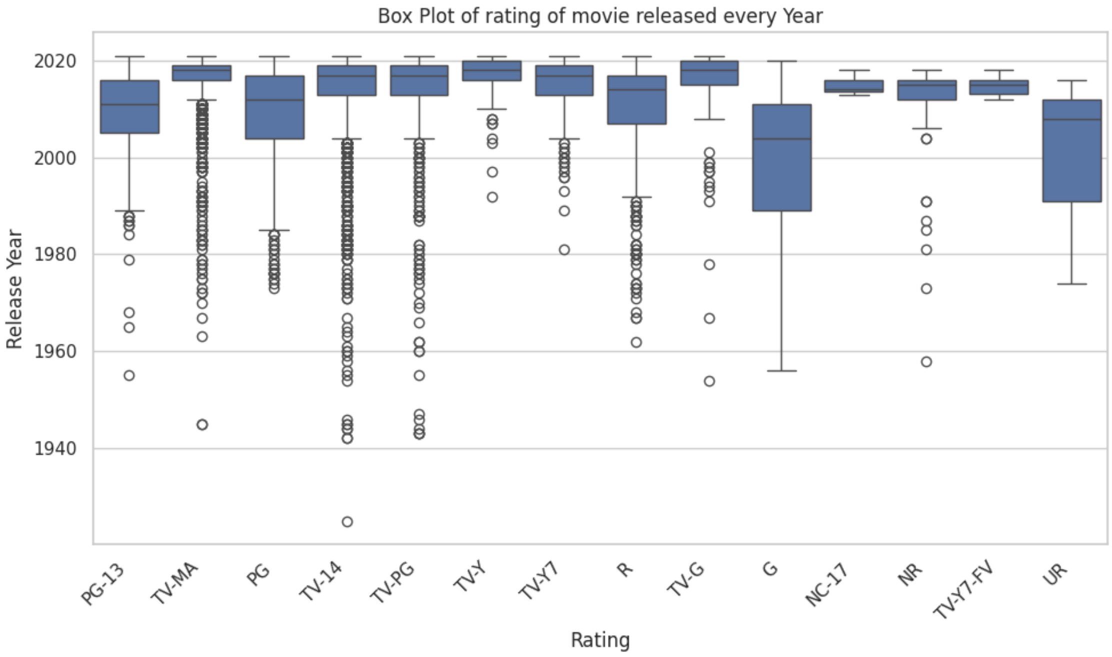


```
[41]: # Ploting the histogram of Duration of Movies
      genere=netflix[["title","duration", "type"]]
      genere=genere.drop_duplicates()
      movies_df = genere[genere['type'] == 'Movie']

      plt.figure(figsize=(10, 6))
      plt.hist(movies_df['duration'], bins=30, edgecolor='k')  # You can adjust the␣
       ↪number of bins for better granularity
      plt.xlabel('Value')
      plt.ylabel('Frequency')
      plt.title('Distribution of Movie Duration')
      plt.tight_layout()
      plt.show()
```

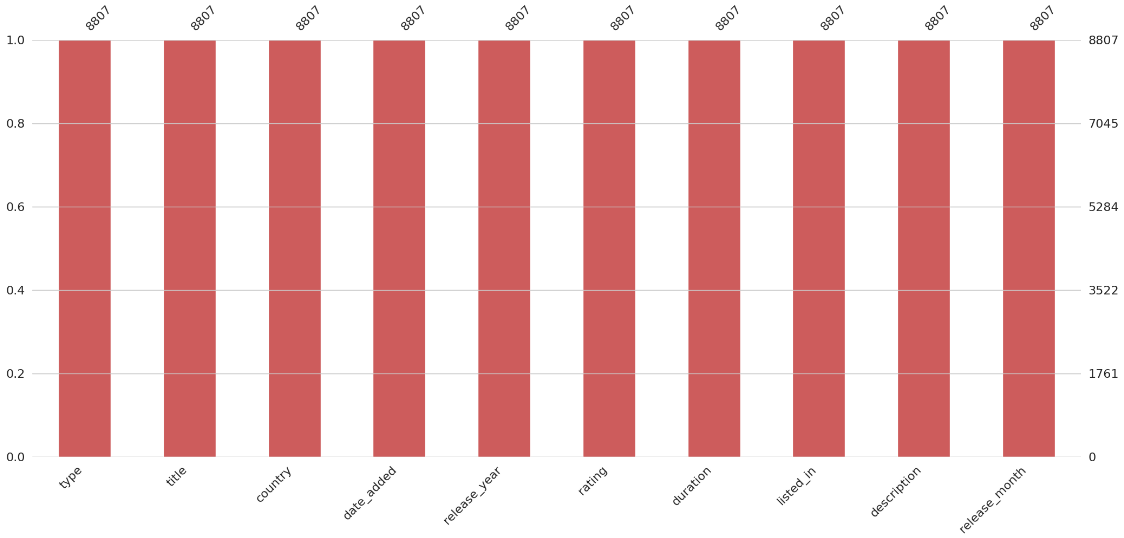


[42]:

```
sns.set(style="whitegrid")
plt.figure(figsize=(10, 6))
sns.boxplot(x='rating', y='release_year', data=netflix)
plt.xticks(rotation=45, ha='right')  # Rotating the x-axis labels for better visibility
plt.xlabel('Rating')
plt.ylabel('Release Year')
plt.title('Box Plot of rating of movie released every Year')
plt.tight_layout()
plt.show()
```



[43]:
```
#To check the missing numbers in the bar chart
import missingno as msno
msno.bar(netflix,color = 'indianred')
plt.show()
```



So the following columns have null values in the dataset:

director - 2634 null values

cast - 825 null values

country - 931 null values

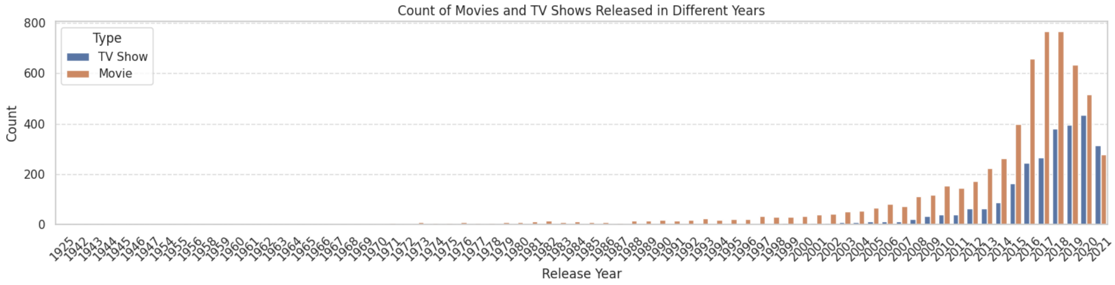
date_added - 10 null values

rating - 4 null values

duration - 3 null values

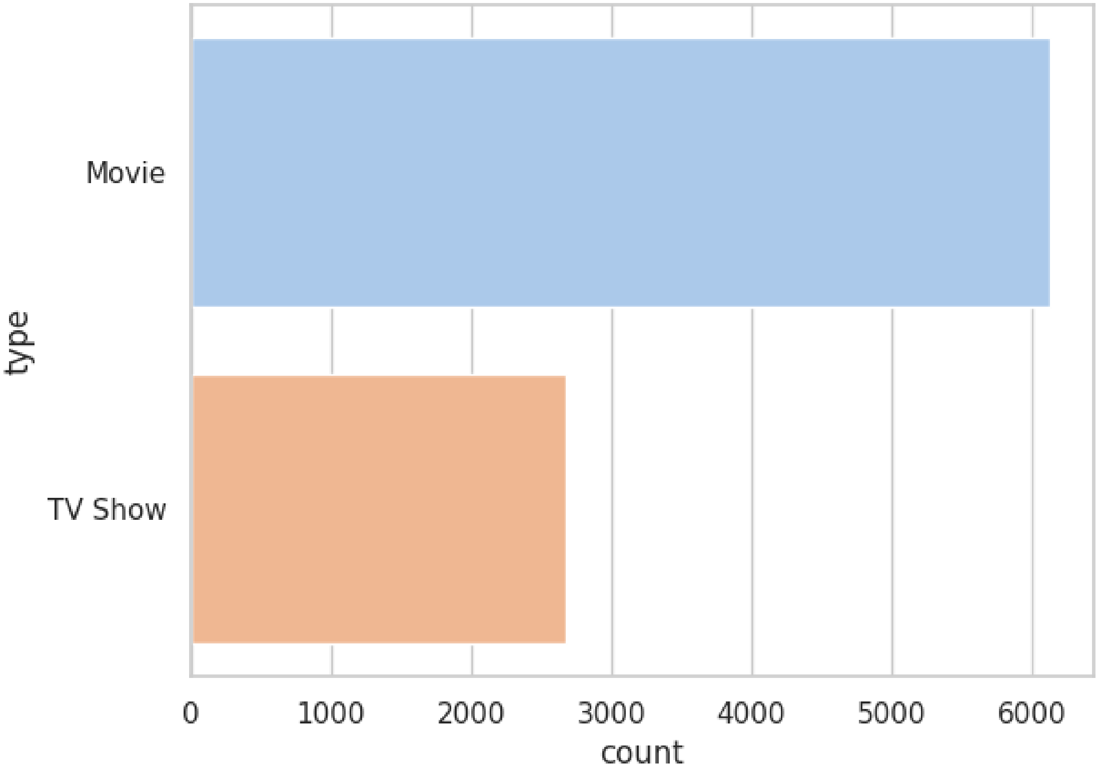
Analysis of Movies vs TV Shows

[44]:
```
# Create the Count Plot of Release Year of Movies & Tv series
plt.figure(figsize=(15, 4))
sns.countplot(data=netflix, x='release_year', hue='type')
plt.xlabel('Release Year')
plt.ylabel('Count')
plt.title('Count of Movies and TV Shows Released in Different Years')


plt.legend(title='Type', loc='upper left')
plt.xticks(rotation=45)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()

# Show the plot
plt.show()
```



[45]:
```
print(netflix.type.value_counts())
sns.countplot(netflix.type,palette="pastel")
plt.show()
```

```
Movie      6131
TV Show    2676
Name: type, dtype: int64
```

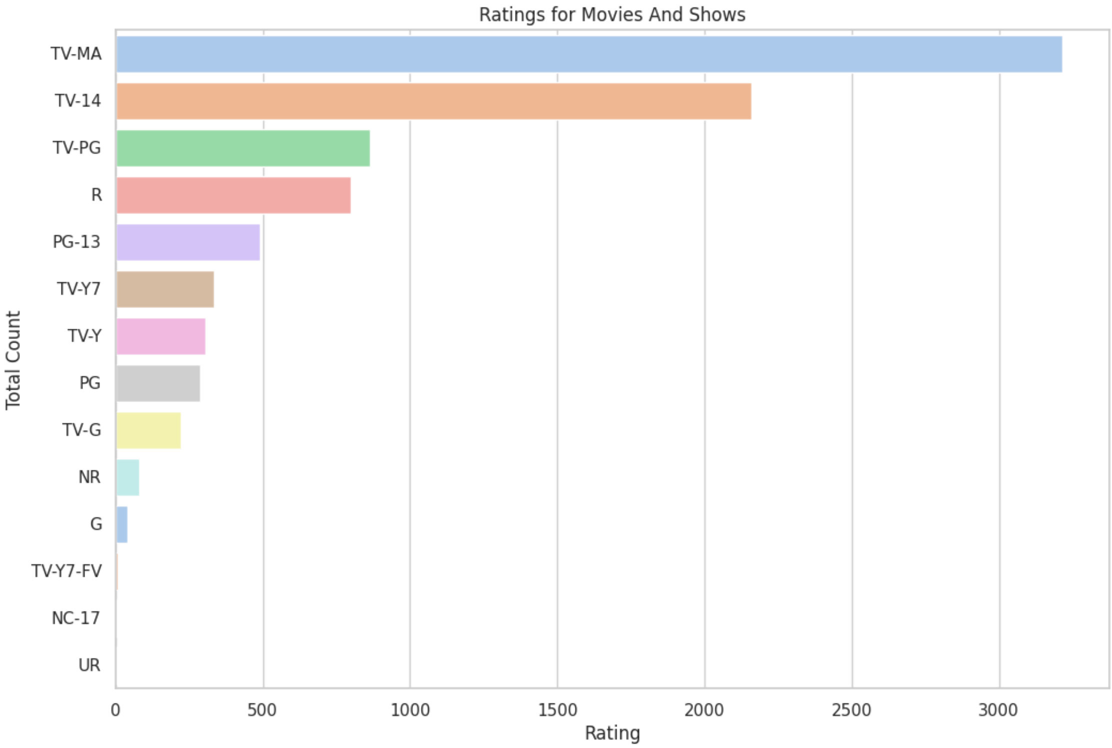


Analysis of Ratings

[46]:
```
rating_counts = netflix.rating.value_counts()
print(rating_counts)
plt.figure(figsize = (12,8))
sns.countplot(netflix.rating, order = rating_counts.index[0:
  ↪15],palette="pastel")
plt.title("Ratings for Movies And Shows")
plt.xlabel("Rating")
plt.ylabel("Total Count")
plt.show()
```

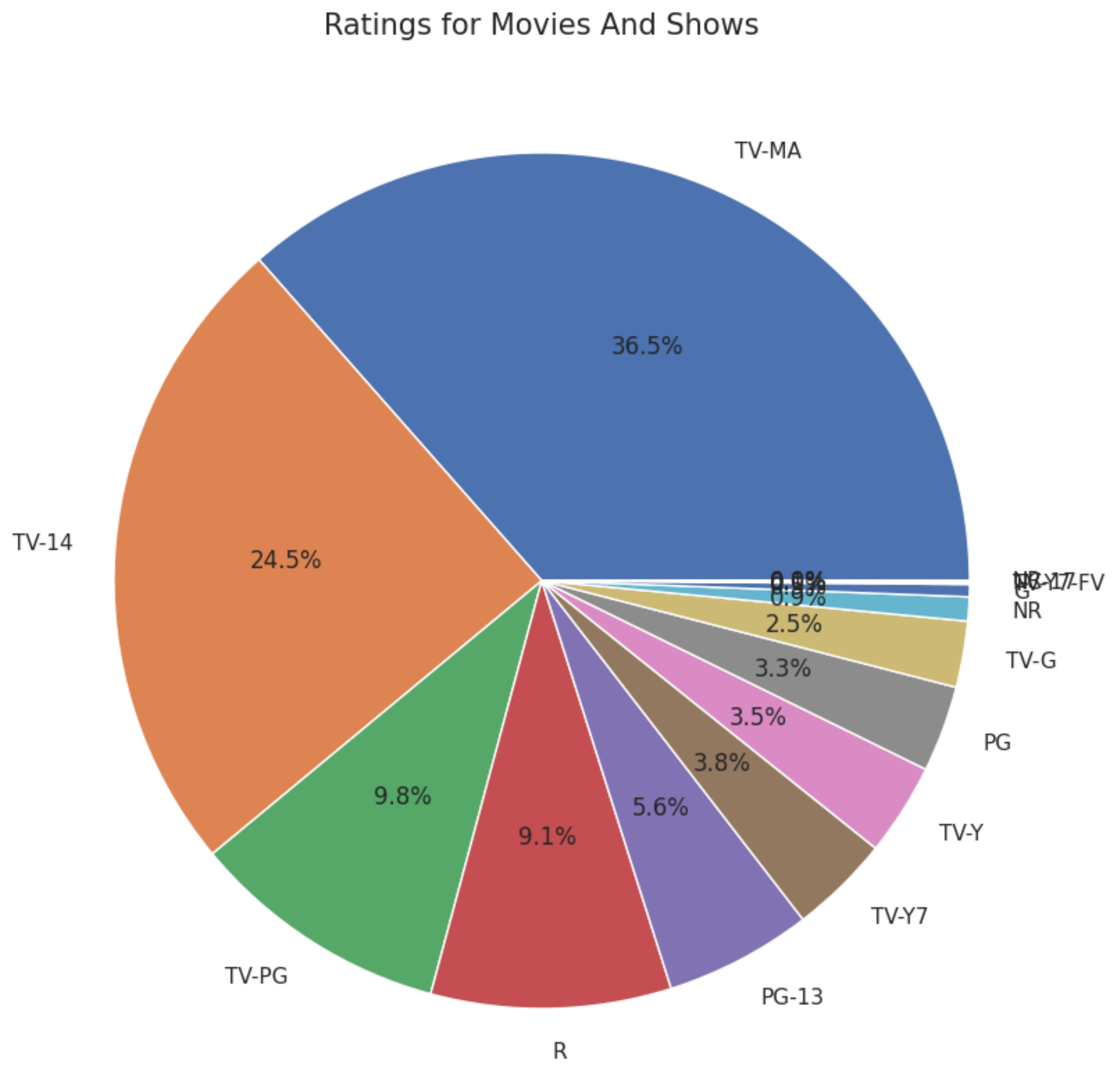
```
TV-MA       3214
TV-14       2160
TV-PG        863
R            799
PG-13        490
TV-Y7        334
TV-Y         307
PG           287
TV-G         220
NR            80
```

```
G              41
TV-Y7-FV        6
NC-17           3
UR              3
Name: rating, dtype: int64
```

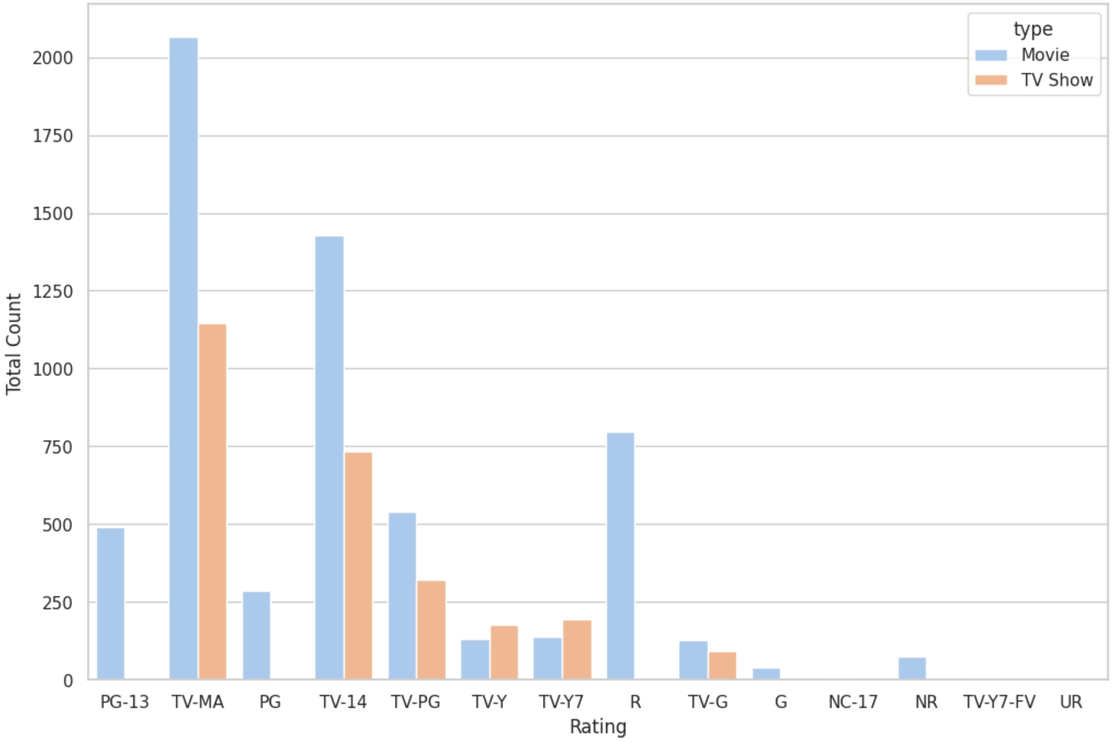


[47]:

```
explode = [0,0,0,0,0,0,0,0,0,0,0,0,0,0]
sizes = rating_counts.values

# visual
plt.figure(figsize = (10,10))
plt.pie(sizes, explode=explode, labels=rating_counts.index, autopct='%1.1f%%')
plt.title('Ratings for Movies And Shows',fontsize = 15)
plt.show()
```



```
[48]: #Type - rating
      plt.figure(figsize = (12,8))
      sns.countplot(x='rating',data = netflix,hue='type',palette="pastel")
      plt.xlabel("Rating")
      plt.ylabel("Total Count")
      plt.show()
```

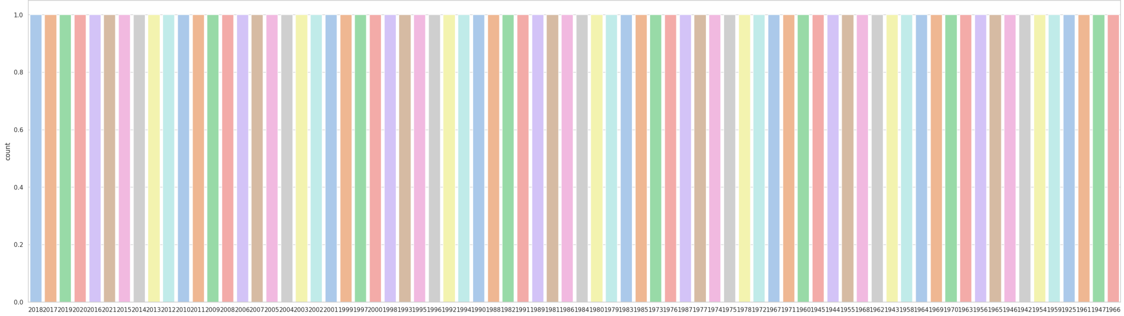


Year wise analysis

[49]:
```
release_year_counts = netflix.release_year.value_counts()
print(release_year_counts)
```

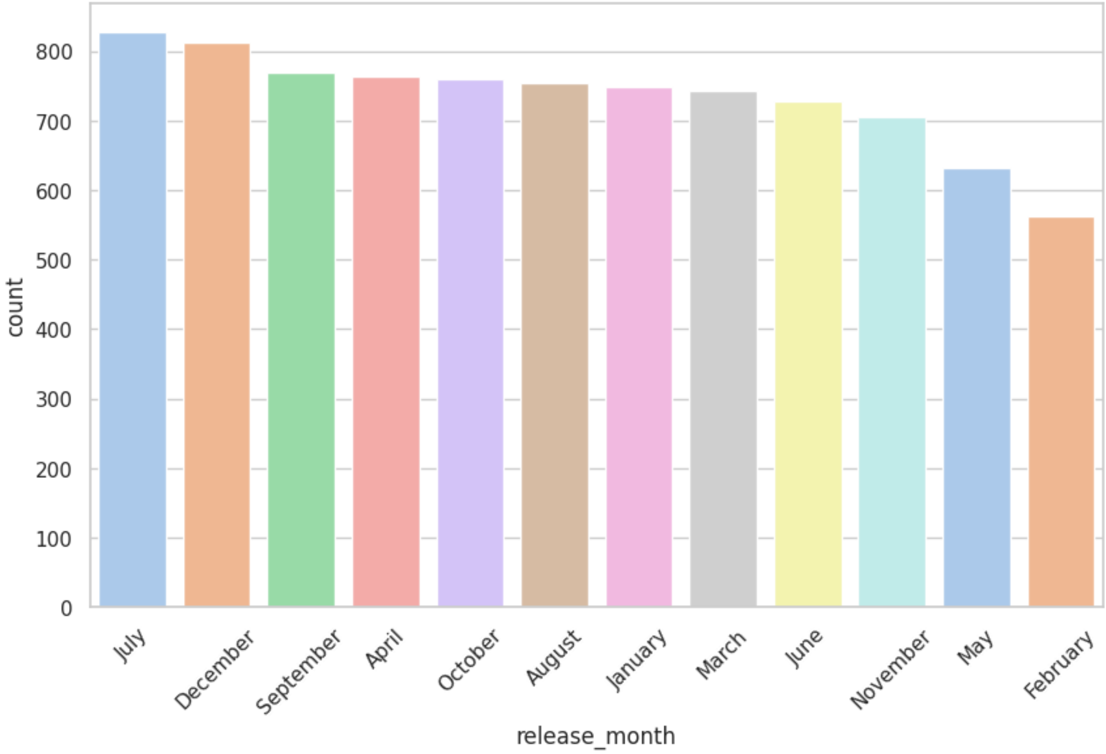
```
2018    1147
2017    1032
2019    1030
2020     953
2016     902
        …
1959       1
1925       1
1961       1
1947       1
1966       1
Name: release_year, Length: 74, dtype: int64
```

[50]:
```
plt.figure(figsize = (36,10))
sns.countplot(netflix.release_year, order = release_year_counts.index[0:
 ↪200],palette="pastel")
plt.show()
```



As we can see most of the movies and Tv shows on Netflix are released in 2018.Let's see which month directors prefer most to release their Movies & Tv Shows.

```
[51]: plt.figure(figsize=(10,6))
sns.countplot(x="release_month",data= netflix,order = netflix['release_month'].
 ↪value_counts().index[0:12],palette="pastel")
plt.xticks(rotation=45)
plt.show()
```

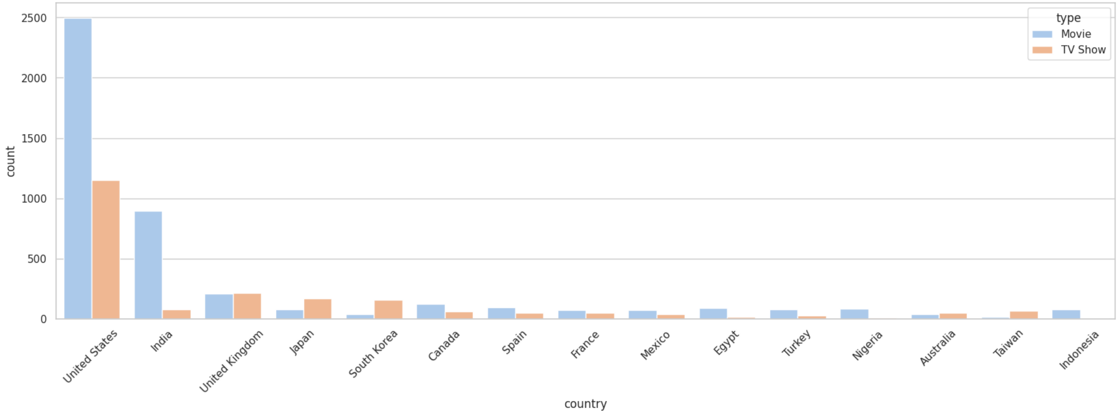


Countries with the most content available

[52]:

```
print(netflix["country"].value_counts().head())
plt.figure(figsize=(20,6))
sns.countplot(x="country",data= netflix,hue= "type",order = netflix['country'].
  ↪value_counts().index[0:15],palette="pastel")
plt.xticks(rotation=45)
plt.show()
```

```
United States     3649
India              972
United Kingdom     419
Japan              245
South Korea        199
Name: country, dtype: int64
```



Unsurprisingly, the United States stands out because Netflix is an American company. India surprisingly ranks second in the film, followed by the UK.

#Top 10 Genres of Movies

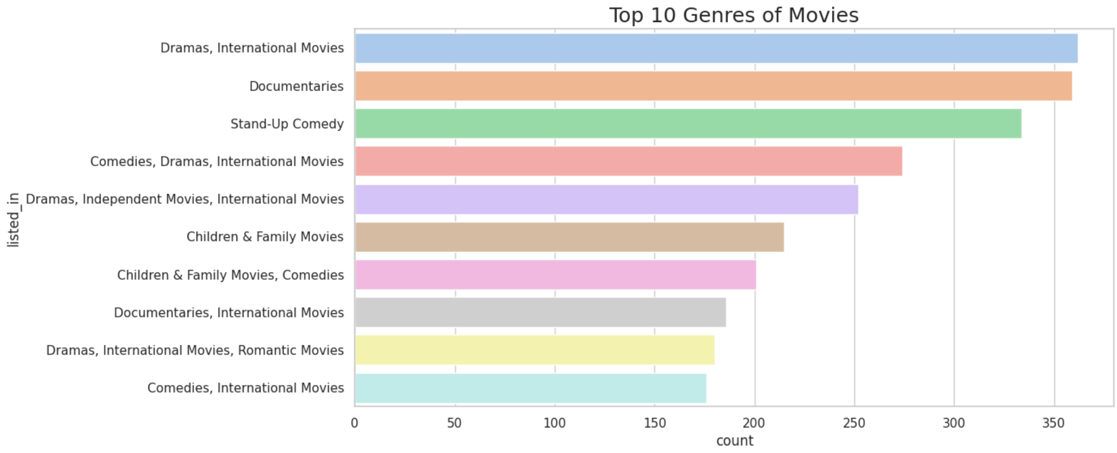
[54]:

```
netflix_movies = netflix[netflix['type']=='Movie']
```

[55]:

```
print(netflix_movies["listed_in"].value_counts()[:10])
plt.figure(figsize=(12,6))
sns.countplot(y='listed_in',data = netflix_movies,order␣
  ↪=netflix_movies["listed_in"].value_counts().index[0:10],palette="pastel")
plt.title("Top 10 Genres of Movies",size=18)
plt.show()
```

```
Dramas, International Movies                          362
Documentaries                                         359
Stand-Up Comedy                                       334
Comedies, Dramas, International Movies                274
Dramas, Independent Movies, International Movies      252
Children & Family Movies                              215
Children & Family Movies, Comedies                 201
Documentaries, International Movies                186
Dramas, International Movies, Romantic Movies      180
Comedies, International Movies                     176
Name: listed_in, dtype: int64
```

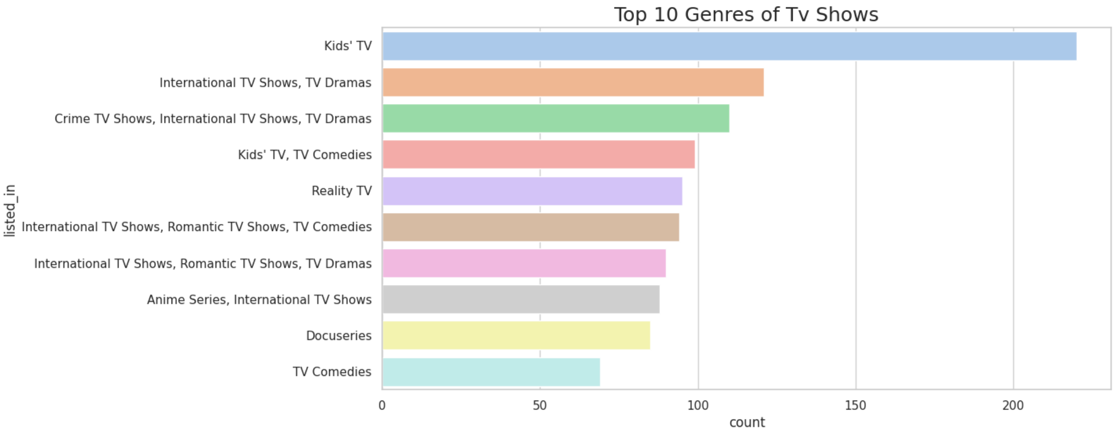


#Top 10 Genres of Tv Shows

```
[56]: netflix_shows = netflix[netflix['type']=='TV Show']
```

```
[57]: print(netflix_shows["listed_in"].value_counts()[:10])
plt.figure(figsize=(12,6))
sns.countplot(y='listed_in',data = netflix_shows,order␣
 ↪=netflix_shows["listed_in"].value_counts().index[0:10],palette="pastel")
plt.title("Top 10 Genres of Tv Shows",size=18)
plt.show()
```

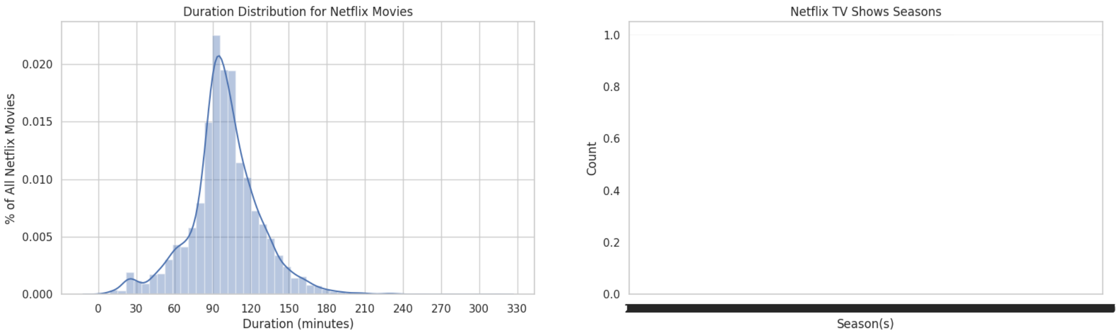
```
Kids' TV                                                   220
International TV Shows, TV Dramas                          121
Crime TV Shows, International TV Shows, TV Dramas          110
Kids' TV, TV Comedies                                       99
Reality TV                                                  95
International TV Shows, Romantic TV Shows, TV Comedies      94
International TV Shows, Romantic TV Shows, TV Dramas        90
Anime Series, International TV Shows                        88
Docuseries                                                  85
TV Comedies                                                 69
Name: listed_in, dtype: int64
```



#Analysis of duration of movies and Tv Shows

[59]:
```
netflix_movies.duration = netflix_movies.duration.str.replace(' min','').
  ↪astype(int)
netflix_shows.rename(columns={'duration':'seasons'}, inplace=True)
netflix_shows.replace({'seasons':{'1 Season':'1 Seasons'}}, inplace=True)
netflix_shows.seasons = netflix_shows.seasons.str.replace(' Seasons','').
  ↪astype(int)
```

[61]:
```
fig, ax = plt.subplots(1,2, figsize=(19, 5))
g1 = sns.distplot(netflix_movies.duration,ax=ax[0]);
g1.set_xticks(np.arange(0,360,30))
g1.set_title("Duration Distribution for Netflix Movies")
g1.set_ylabel("% of All Netflix Movies")
g1.set_xlabel("Duration (minutes)")
g2 = sns.countplot(netflix_shows.seasons,ax=ax[1],palette="pastel");
g2.set_title("Netflix TV Shows Seasons")
g2.set_ylabel("Count")
g2.set_xlabel("Season(s)")
fig.show()
```



As you can see, movies are usually between 75-120 minutes and TV shows are usually 1 season.

```
[62]: oldest = netflix.sort_values("release_year", ascending = True)
      oldest[['title', "release_year"]][:10]
```

```
[62]:
                                                 title  release_year
4250                 Pioneers: First Women Filmmakers*          1925
7790                                    Prelude to War          1942
8205                              The Battle of Midway          1942
8660     Undercover: How to Operate Behind Enemy Lines          1943
8739                Why We Fight: The Battle of Russia          1943
8763                   WWII: Report from the Aleutians          1943
8640                                  Tunisian Victory          1944
8436                                 The Negro Soldier          1944
8419  The Memphis Belle: A Story of a\nFlying Fortress          1944
7930                                        San Pietro          1945
```

Standup shows on Netflix

```
[63]: standup=netflix[netflix["listed_in"] == "Stand-Up Comedy"]
      standup[["title","country","release_year"]].head(10)
```

```
[63]:
                                             title        country  \
278                    Lokillo: Nothing's the Same       Colombia
359                   The Original Kings of Comedy  United States
475                                   The Stand-Up  United States
484                   Lee Su-geun: The Sense Coach  United States
766                        Alan Saldaña: Locked Up         Mexico
826                             Bo Burnham: Inside  United States
838                          Soy Rada: Serendipity      Argentina
1172                       Loyiso Gola: Unlearning   South Africa
1189  Nate Bargatze: The Greatest Average American  United States
1191                              The Fluffy Movie  United States

      release_year
278           2021
359           2000
475           2019
484           2021
766           2021
826           2021
838           2021
1172          2021
1189          2021
1191          2014
```

Kids TV shows on Netflix

[64]:

```
kids=netflix[netflix["listed_in"] == "Kids' TV"]
kids[["title","country","release_year"]].head(10)
```

[64]:

```
                       title                   country  release_year
34   Tayo and Little Wizards             United States          2020
39              Chhota Bheem                     India          2021
65             Numberblocks            United Kingdom          2021
89               Mighty Raju             United States          2017
100  Tobot Galaxy Detectives             United States          2019
111                 Sharkdog  United States, Singapore          2021
123                Luv Kushh             United States          2012
153               Kid-E-Cats                    Russia          2016
254        Go! Go! Cory Carson             United States          2021
263        Mother Goose Club             United States          2016
```

3.0.1 Business Insight from the plot - TV Shows released per Month

From the graph we can see that most of the TV shows are released in the months of ‘June’,‘July’,‘August’,‘September’ and ‘December’

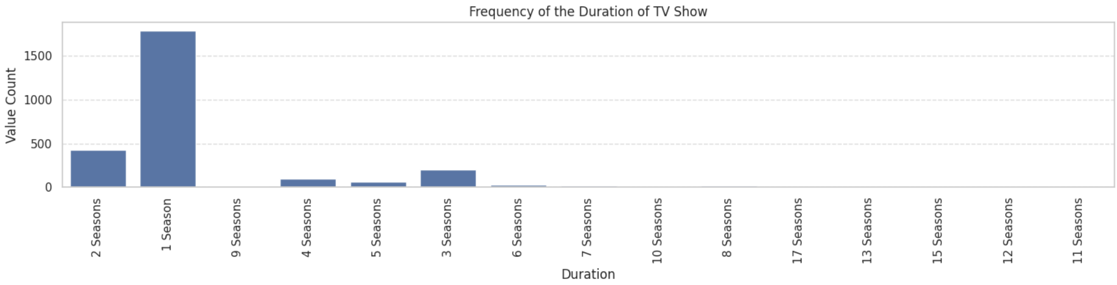
Relatively less number of shows are release in the months of ‘January’,‘February’ and ”May’

Average amount of TV Shows are released in the months of ‘March’, ‘April’, ‘October’ and ‘November’

4 Duration of Tv Shows and Movies.

[65]:

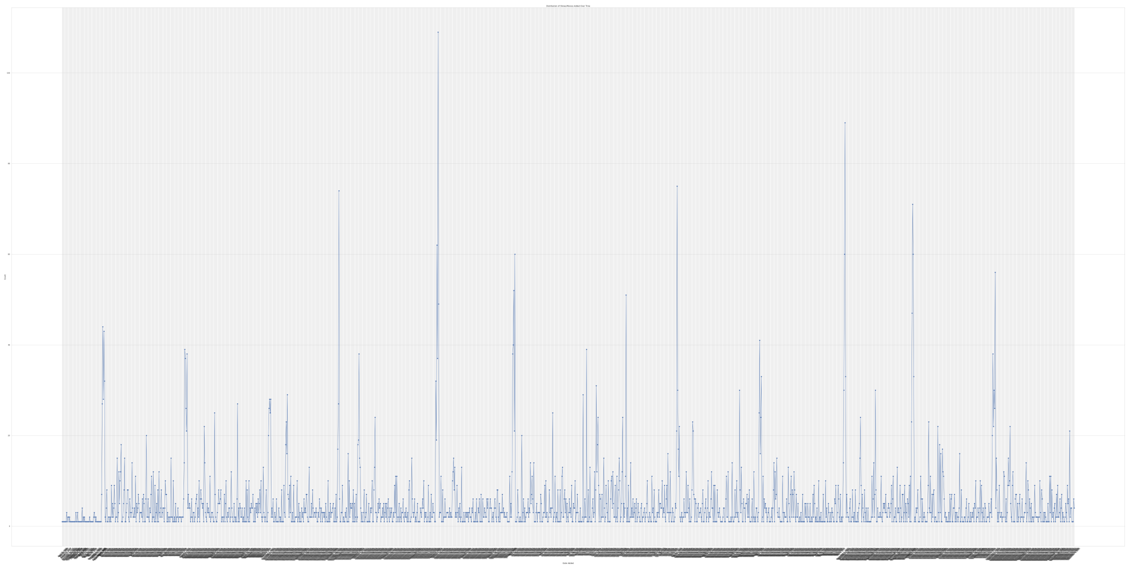
```
#dist plot of duration
 # Separate the data for movies and TV shows
movies_data = netflix[netflix['type'] == 'Movie']
tv_shows_data = netflix[netflix['type'] == 'TV Show']

plt.figure(figsize=(15, 4))
sns.countplot(data=tv_shows_data, x='duration')
plt.xlabel('Duration')
plt.ylabel('Value Count')
plt.title('Frequency of the Duration of TV Show')
plt.xticks(rotation=90)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()

# Show the plot
plt.show()
```



5 Time series plot of date_added column

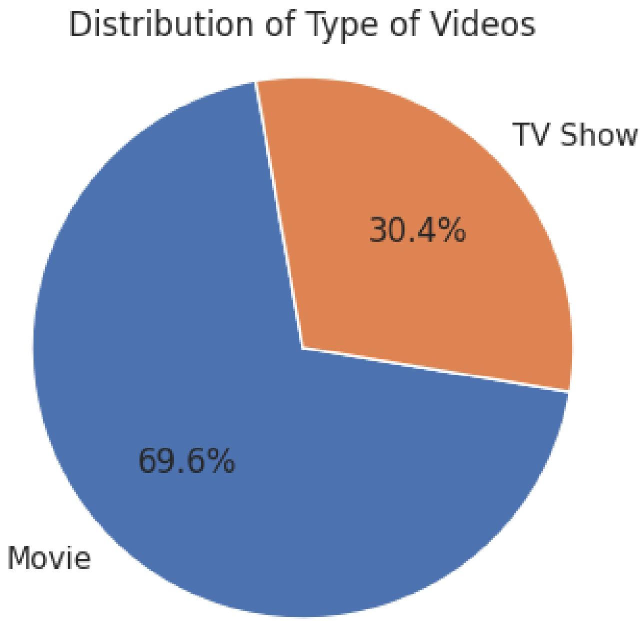
[69]:

```
# Time series plot of date_added column
date_counts = netflix.groupby('date_added').size()
plt.figure(figsize=(100, 50))
plt.plot(date_counts.index, date_counts.values, marker='o', linestyle='-',␣
 ↪color='b')
plt.xlabel('Date Added')
plt.ylabel('Count')
plt.title('Distribution of Shows/Movies Added Over Time')
plt.xticks(rotation=45)
plt.grid(True)
plt.tight_layout()
plt.show()
```



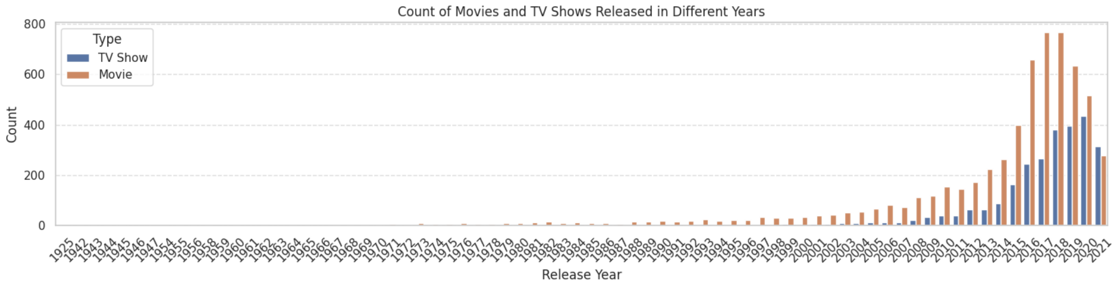
[70]:

```python
# Create the Pie Chart for type
type_count = netflix['type'].value_counts()
plt.figure(figsize=(4,4))
plt.pie(type_count, labels=type_count.index,  autopct='%1.1f%%', startangle=100)
plt.axis('equal')
plt.title('Distribution of Type of Videos')
plt.show()
```



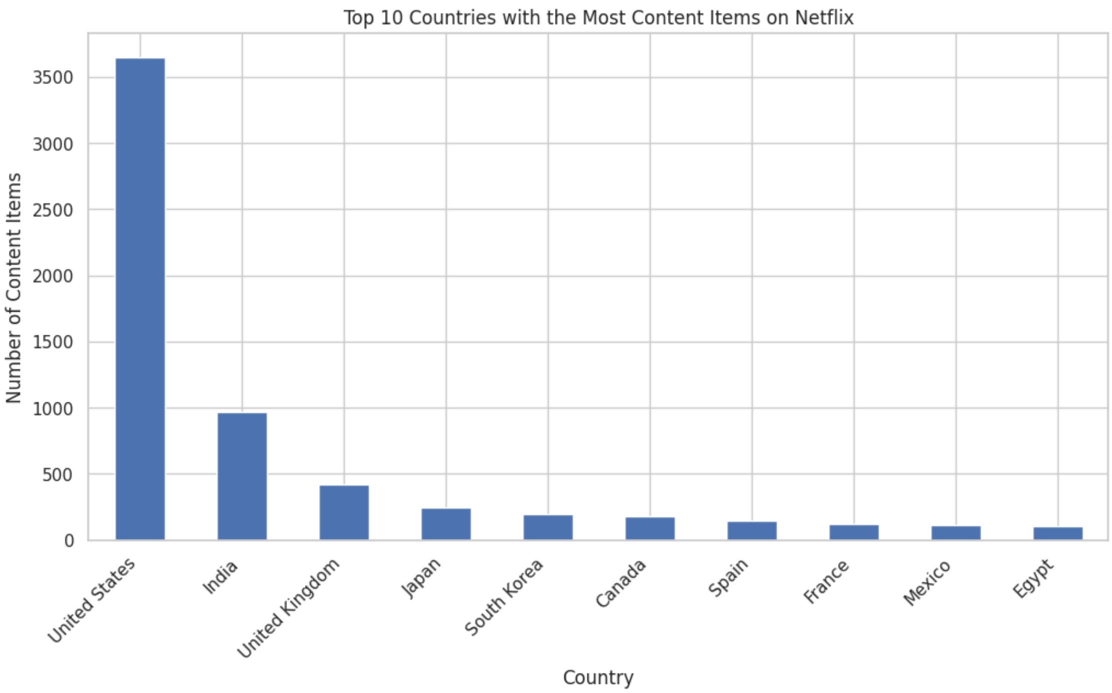
[71]:

```python
# Create the Count Plot
plt.figure(figsize=(15, 4))
sns.countplot(x='release_year', hue='type', data=netflix)
plt.xlabel('Release Year')
plt.ylabel('Count')
plt.title('Count of Movies and TV Shows Released in Different Years')
plt.xticks(rotation=45)
plt.legend(title='Type', loc='upper left')
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()
```



[72]:

```
# Group the content (movies/TV shows) by country and count the number of items␣
 ↪in each country
content_by_country = netflix.groupby('country').size().
 ↪sort_values(ascending=False)

# Select the top 'n' countries to plot (you can adjust the value of 'n' as per␣
 ↪your requirement)
n = 10
top_countries = content_by_country.head(n)

# Plotting the number of content items by country using a bar chart
plt.figure(figsize=(12, 6))
top_countries.plot(kind='bar')
plt.xlabel('Country')
plt.ylabel('Number of Content Items')
plt.title(f'Top {n} Countries with the Most Content Items on Netflix')
plt.xticks(rotation=45, ha='right')
plt.show()

Highest_content_country = content_by_country.idxmax()
lowest_content_country= content_by_country.idxmin()
print(f"The Highest_content_country is: {Highest_content_country }")
print(f"The lowest_content_country is: {lowest_content_country}")
```



```
The Highest_content_country is: United States
The lowest_content_country is: United Kingdom, China
```

6 Q6. Insights based on Non-Graphical and Visual Analysis

[73]:
```
genere=netflix[["title","listed_in"]]
genere=genere.drop_duplicates()

most_common_directors = genere['listed_in'].mode()
most_common_directors
```

[73]:
```
0    Dramas, International Movies
Name: listed_in, dtype: object
```

We could see that there are two types 1. Movies 2. TV Shows

We saw the below details,

- There are 8807 Unique titles
- There are total of 39296 Cast members
- There are total of 197 countries for these movies and TV Shows
- The Whole data of the movies is between the 1925 to 2021
- We found there are 17 types of ratings
- The Duration for the TV shows is from season 1 to 10
- Also for the movies iot ranges form 22 to 230 minutes

Types- There are 6131 Movies & 2676 TV Show listed in Netfilx according to given dataframe

Titles- Not a valuable insight

Directors- Rajiv Chilak

Cast- Anupam Kher

Country - United State

Date Added- Not a valuable insight

Release_Year - 2018

Rating - TV-MA

Duration - For TV Shows season 1 is the highest and for movies highest is in range of 90-110

Listed_in - International Movies

##Business Insights

- The high demand in the watching movies and the TV shows in the OTT platforms are increased in the last 20 years
- The Best time to release the movies is from June to December
- the movies should not be released in the January to May month
- Creating Movies and TV Shows and releasing only in OTT is a good option of expanding the business.
- Adding More and More Movies, Series and TV Shows would increase the chance of generating more revenue
- Our most of the audiences are adult we can sat that by highest content available is of genere TV-MA, we can target those audience and encourage then to take subscription by giving sone good offer.
- Highest movies listed in OTT release in United States. US is the highest content creator and India is second highest content creator. Listing More and More content fron these two contries help us to give most benifits.