Project Title: Netflix Data Analysis with Python

Netflix is one of the largest providers of online streaming services, boasting a massive subscriber base that generates vast amounts of data. In this project, I'm going to walk you through a data science project focused on analyzing Netflix data using Python.

Introduction to Netflix Data Analysis

Netflix has continually adapted its business model to meet evolving market demands, transitioning from on-demand DVD rentals to becoming a major producer of original content. This shift has generated a wealth of data that can be analyzed to glean insights into Netflix's content strategy and user preferences.

In this project, I will explore several key aspects of Netflix's data to understand what drives their business. Key areas of analysis include:

Content availability: Understanding what content is available on Netflix.

Content similarity: Analyzing the similarities between different content.

Network analysis: Examining the relationships between actors and directors.

Business focus: Identifying the trends Netflix is focusing on.

Sentiment analysis: Evaluating the sentiment of the content available on Netflix. Dataset Overview

The dataset I'm using for this Netflix data analysis contains information on TV shows and movies streamed on Netflix as of 2019. This dataset is provided by Flixable, a third-party research engine for Netflix.

In [ ]:

In [1]:
```
import numpy as np
import pandas as pd
import plotly.express as px
from textblob import TextBlob
```

In [2]:
```
dff = pd.read_csv(r'C:\Users\User\Desktop\NDA\netflix_titles.csv')
```

In [3]:
```
dff.columns
```

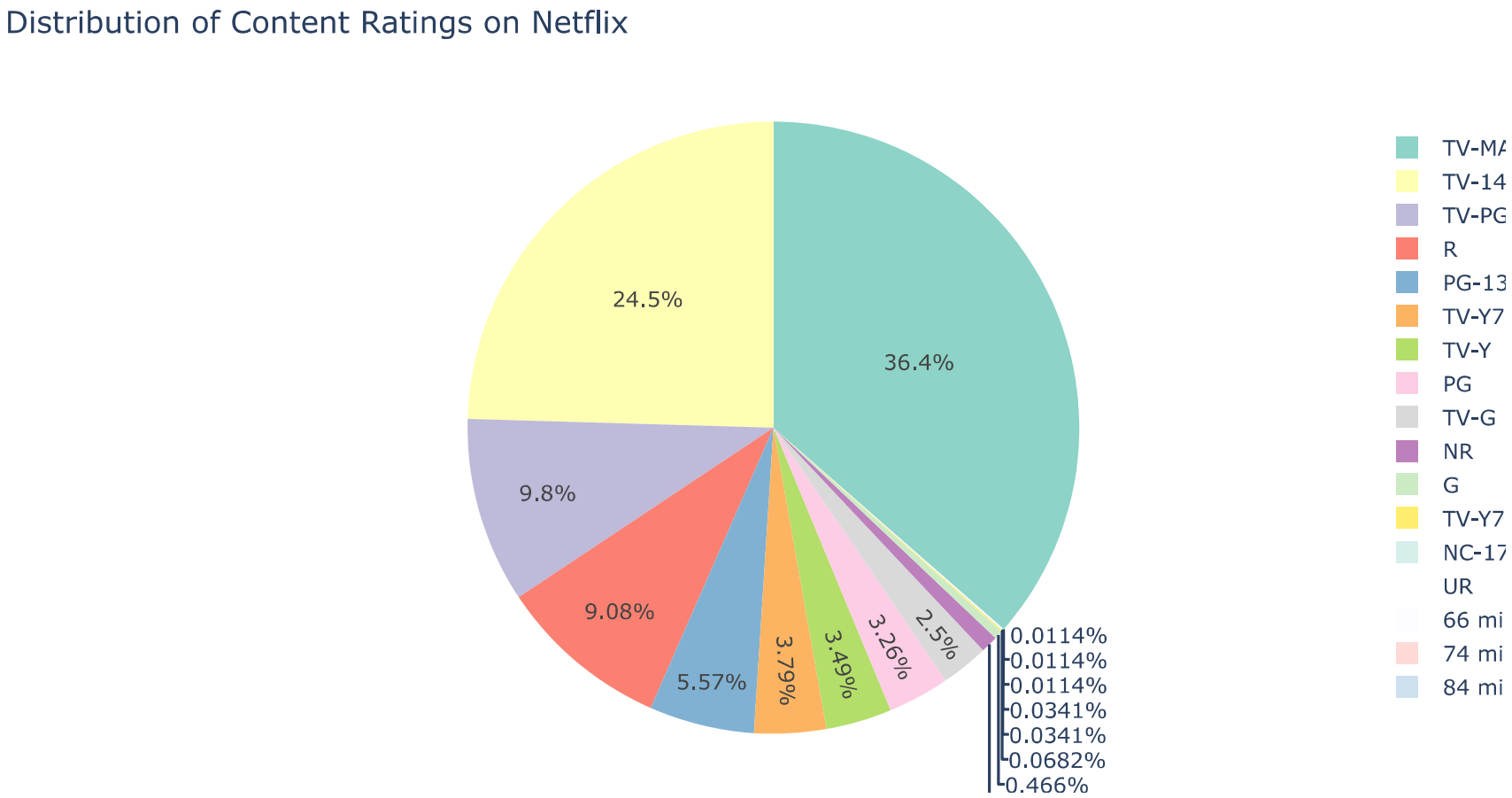
Out[3]:
```
Index(['show_id', 'type', 'title', 'director', 'cast', 'country', 'date_added',
       'release_year', 'rating', 'duration', 'listed_in', 'description'],
      dtype='object')
```

Distribution of Content:

To begin the task of analyzing Netflix data, I'll start by looking at the distribution of content ratings on Netflix:

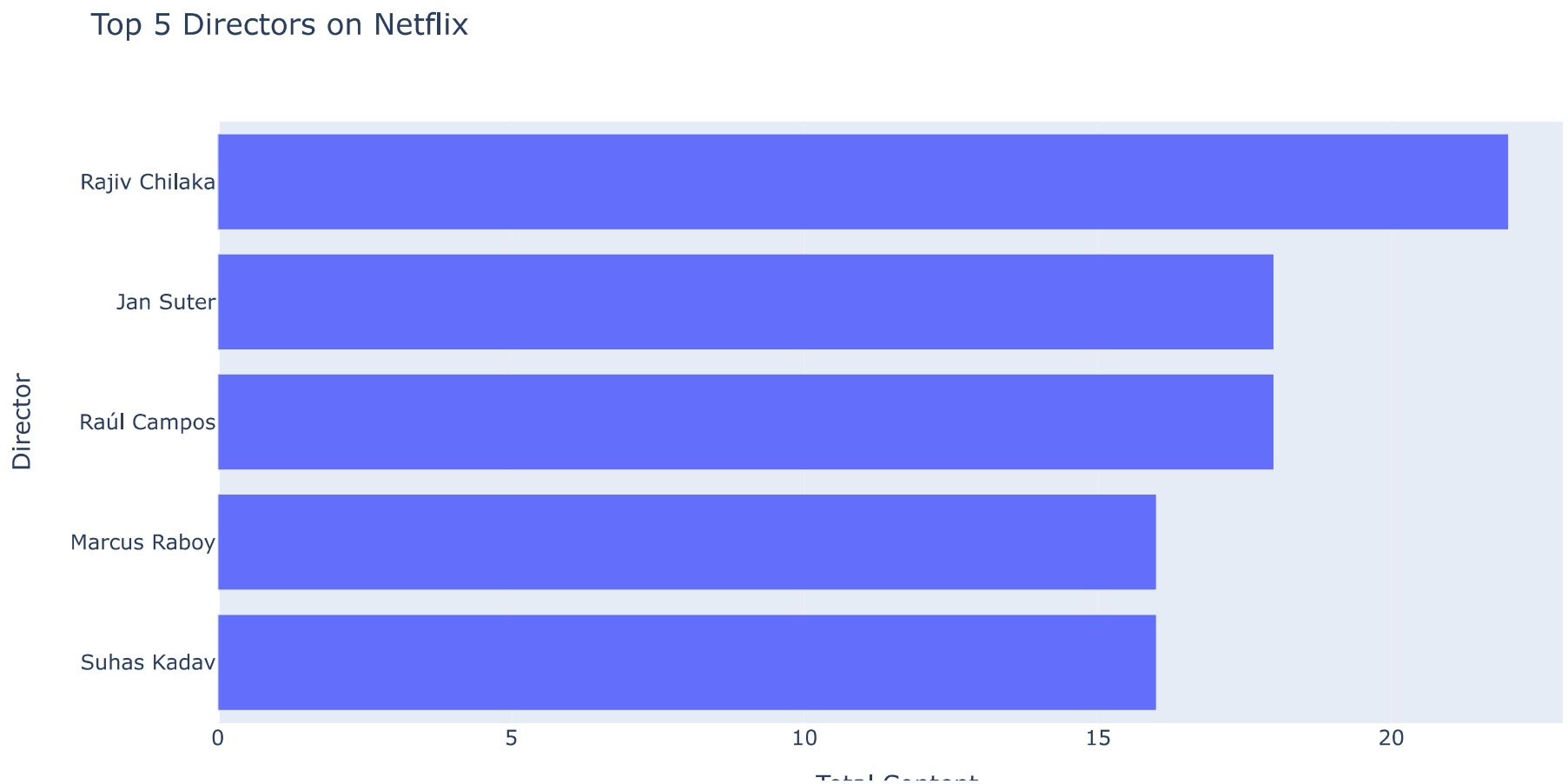
In [4]:
```
z = dff.groupby(['rating']).size().reset_index(name='counts')
```

In [5]:
```
pieChart = px.pie(z, values='counts', names='rating',
                  title='Distribution of Content Ratings on Netflix',
                  color_discrete_sequence=px.colors.qualitative.Set3)
pieChart.show()
```



Top 5 Actors and Directors: Now let’s see the top 5 successful directors on this platform:

In [6]:
```
dff['director']=dff['director'].fillna('No Director Specified')
filtered_directors=pd.DataFrame()
filtered_directors=dff['director'].str.split(',',expand=True).stack()
filtered_directors=filtered_directors.to_frame()
filtered_directors.columns=['Director']
directors=filtered_directors.groupby(['Director']).size().reset_index(name='Total Content')
directors=directors[directors.Director !='No Director Specified']
directors=directors.sort_values(by=['Total Content'],ascending=False)
directorsTop5=directors.head()
directorsTop5=directorsTop5.sort_values(by=['Total Content'])
fig1=px.bar(directorsTop5,x='Total Content',y='Director',title='Top 5 Directors on Netflix')
fig1.show()
```



From the above graph it is derived that the top 5 directors on this platform are:

- Rajiv chilaka
- Jan Suter

raul campos

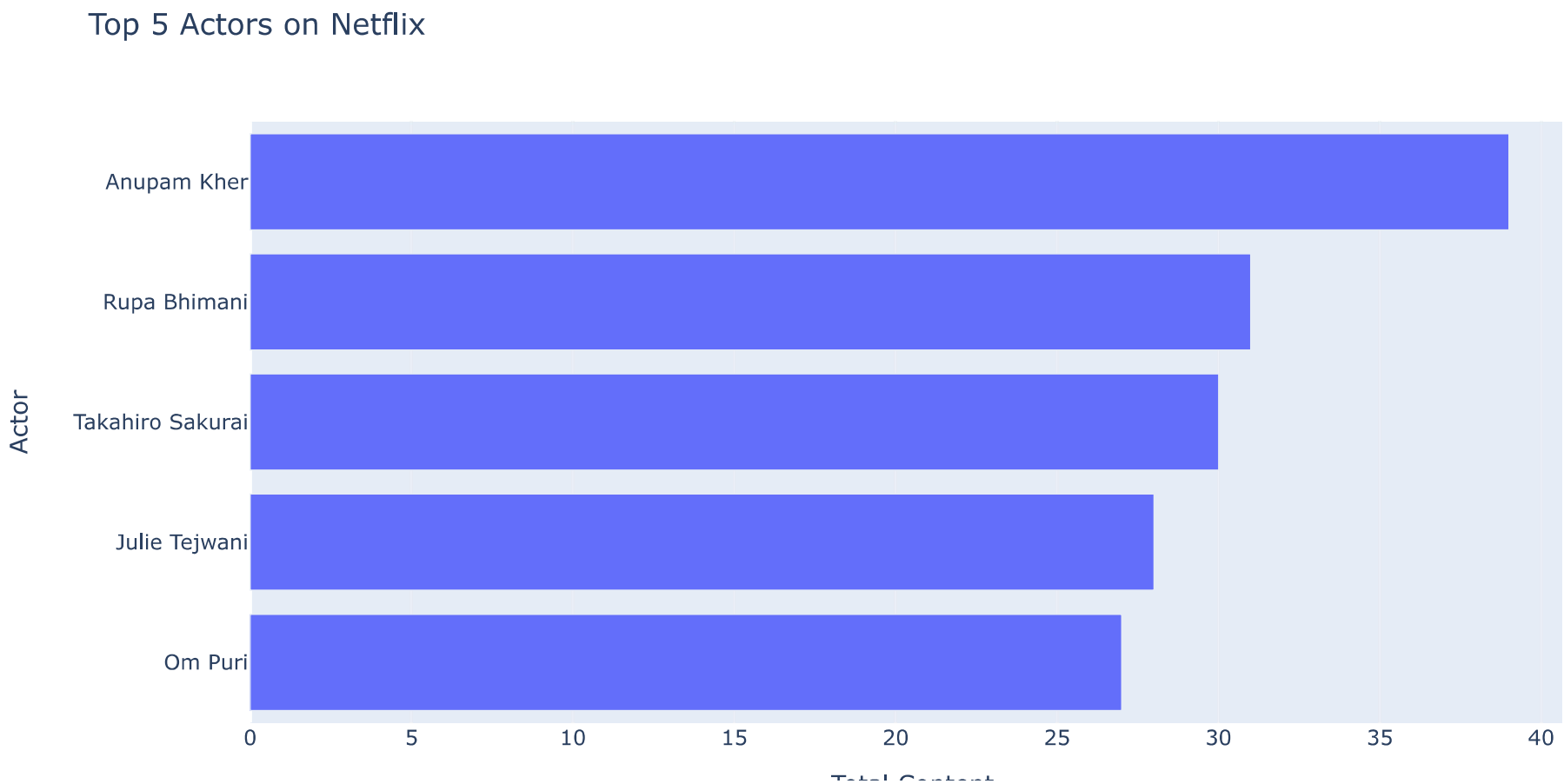
Marcus Raboy

suhas kadav

Now let’s have a look at the top 5 successful actors on this platform:

In [ ]:

In [7]:
```
dff['cast']=dff['cast'].fillna('No Cast Specified')
filtered_cast=pd.DataFrame()
filtered_cast=dff['cast'].str.split(',',expand=True).stack()
filtered_cast=filtered_cast.to_frame()
filtered_cast.columns=['Actor']
actors=filtered_cast.groupby(['Actor']).size().reset_index(name='Total Content')
actors=actors[actors.Actor !='No Cast Specified']
actors=actors.sort_values(by=['Total Content'],ascending=False)
actorsTop5=actors.head()
actorsTop5=actorsTop5.sort_values(by=['Total Content'])
fig2=px.bar(actorsTop5,x='Total Content',y='Actor', title='Top 5 Actors on Netflix')
fig2.show()
```



The above line graph shows that there has been a decline in the production of the content for both movies and other shows since From the above plot, it is derived that the top 5 actors on Netflix are:

- Anupam Kher
- rupa bhimani

Takahiro sakurai

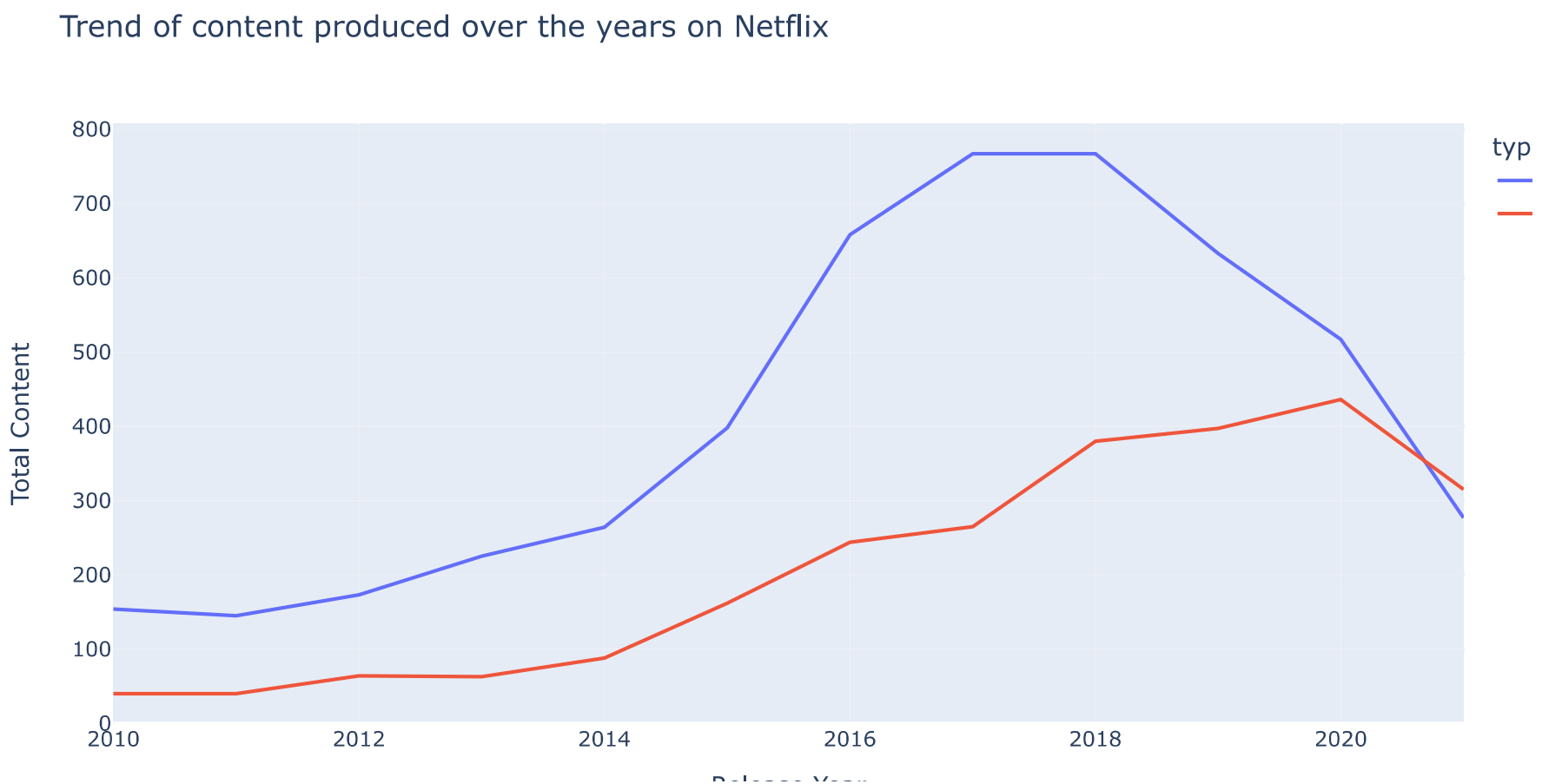
julie tejwani

om puri

Analyzing Content on Netflix:

The next thing to analyze from this data is the trend of production over the years on Netflix:

In [8]:
```
df1=dff[['type','release_year']]
df1=df1.rename(columns={"release_year": "Release Year"})
df2=df1.groupby(['Release Year','type']).size().reset_index(name='Total Content')
df2=df2[df2['Release Year']>=2010]
fig3 = px.line(df2, x="Release Year", y="Total Content", color='type',title='Trend of content produced over the years
fig3.show()
```



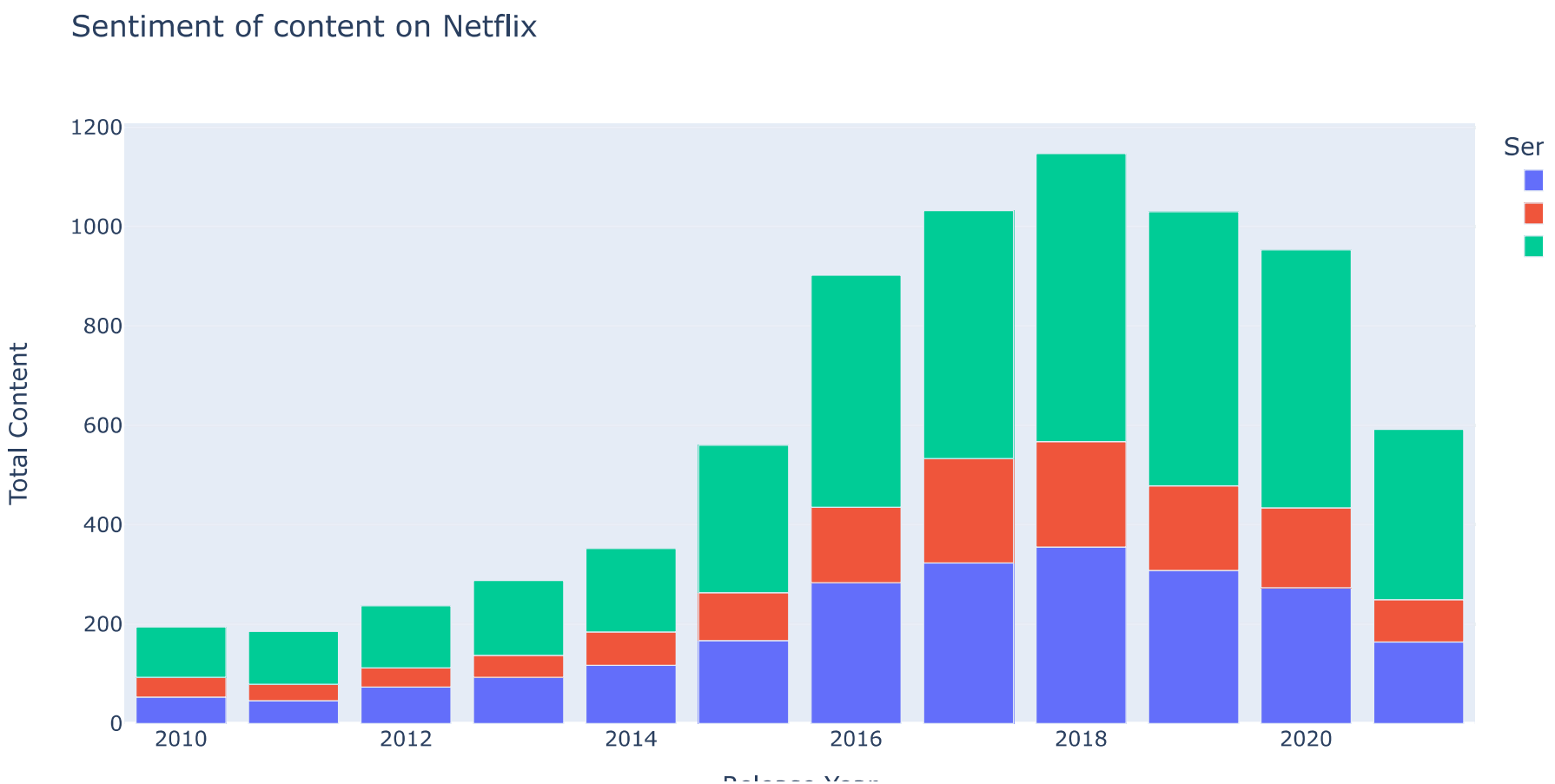
The above line graph shows that there has been a decline in the production of the content for both movies and other shows since 2018.

At last, to conclude our analysis, I will analyze the sentiment of content on Netflix:

In [9]:
```
dfx=dff[['release_year','description']]
dfx=dfx.rename(columns={'release_year':'Release Year'})
for index,row in dfx.iterrows():
    z=row['description']
    testimonial=TextBlob(z)
    p=testimonial.sentiment.polarity
    if p==0:
        sent='Neutral'
    elif p>0:
        sent='Positive'
    else:
        sent='Negative'
    dfx.loc[[index,2],'Sentiment']=sent
```

In [10]:
```
dfx=dfx.groupby(['Release Year','Sentiment']).size().reset_index(name='Total Content')
```

In [11]:
```
dfx=dfx[dfx['Release Year']>=2010]
fig4 = px.bar(dfx, x="Release Year", y="Total Content", color="Sentiment", title="Sentiment of content on Netflix")
fig4.show()
```



So the above graph shows that the overall positive content is always greater than the neutral and negative content combined.

In [ ]: