Netflix Movies and TV Shows Exploratory Data Analysis

VINAY SAI RANGUMURI

BATCH NO. 8

INTERNSHIP BATCH – 3

rangumudrivinaysai2001@gmail.com

Netflix Movies and TV Shows EDA □

Netflix is the world’s leading premium media streaming platform, hosting thousands of films and TV series in nearly 200 countries and territories. Initially, a mail-order DVD rental service launched in 1997, the company quickly dominated the streaming sphere when it launched its subscription video-on-demand service a decade later in 2007, the same year that Hulu launched. The field has become quite a bit more crowded since then, and Netflix now competes with the likes of Amazon Prime Video, HBO Max, Disney Plus, Apple TV Plus, and many more, including niche streamers like The Criterion Channel and Shudder.

Project Overview:

This project aims to analyze the content available on netflix streaming platform. The study will analyze the different content available across different countries and its content creators.

Objective:

The Objective of the project is to identify the content types available on netflix, across different countries, types of content available with similarity, top actors appearing in most contents, top directors creating most content and what netflix is focusing on in recent years w.r.t Movies or Tv shows.

Methodology:

Involves the following steps:

- Business Understanding
- Data Understanding
- Data Preparation
- Data Exploration

Business Understanding

The main questions that we have to answer in this notebook

- ❖ Understanding what content is available in different countries
- ❖ Identifying similar content by matching text-based features
- ❖ Network analysis of Actors / Directors and find interesting insights
- ❖ Does Netflix has more focus on TV Shows than movies in recent years

Data Understanding

1. Explored the data using pandas DataFrame
2. Identified the patterns and relationship between columns type, director, cast, country e.t.c

Data Preparation

1. Performed data cleaning by removing missing values.
2. Transformed the data by separating multiple values in the columns by splitting the data column wise.
3. Used text pre-processing technique on column containing data in details or in sentences to transform the data to find similar words or most common words for example column “description” in this data.

Data Exploration

1. Explored the data from previous Data preparation stage using identified patterns and relationships
2. Python libraries used Matplotlib, WordCloud.
3. Tableau used for visualization from data prepared using python.

ANALYSIS

Exploratory Data Analysis

```
# Data analysis
import numpy as np
import pandas as pd
# Visualization
import seaborn as sns
import matplotlib.pyplot as plt
import plotly.express as px
from wordcloud import WordCloud,STOPWORDS
```

Exploratory Data Analysis

```
# Checking Shape of the data
netflix_overall=pd.read_excel("netflix_titles1.xlsx")
netflix_overall.head()
```

About the 12 columns of this interesting dataset:

- **show_id**: A unique ID for each show
- **type:** The category of a show — it can be Movie or TV Show
- **title**: Name of the show
- **director:** Name of the director(s) of the show
- **cast:** Actors involved in the show
- **country:** Country where the show was produced
- **date_added:** Date when the show was added on Netflix
- **release_year**: Release year of the show
- **rating:** TV rating — a content rating system
- **duration:** Time duration — in minutes or number of seasons
- **listed_in:** Genre(s)
- **description:** A summary of the show

Checking Shape of the data → (8807, 12)

Exploratory Data Analysis

- ❖ Checking Shape of the data → (8807, 12)
- ❖ No of Columns present in the Dataset →

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

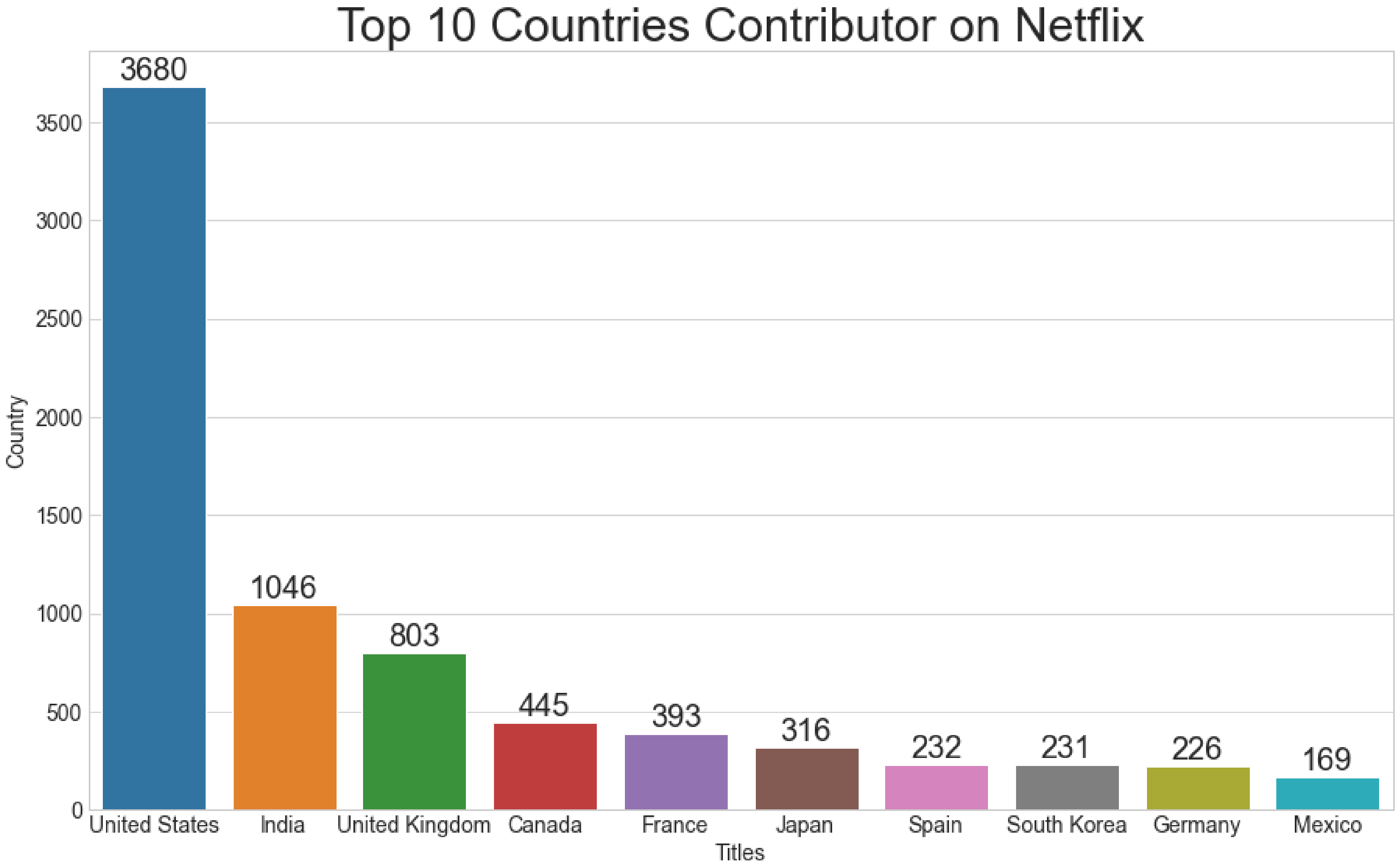
- ❖ Missing data
- ➢ director null rate: 29.91%
- ➢ cast null rate: 9.37%
- ➢ country null rate: 9.44%
- ➢ date_added null rate: 0.11%
- ➢ rating null rate: 0.05%
- ➢ duration null rate: 0.03%

5 columns have missing values, with Director missing 1/3 of the time

Understanding what content is available in different countries

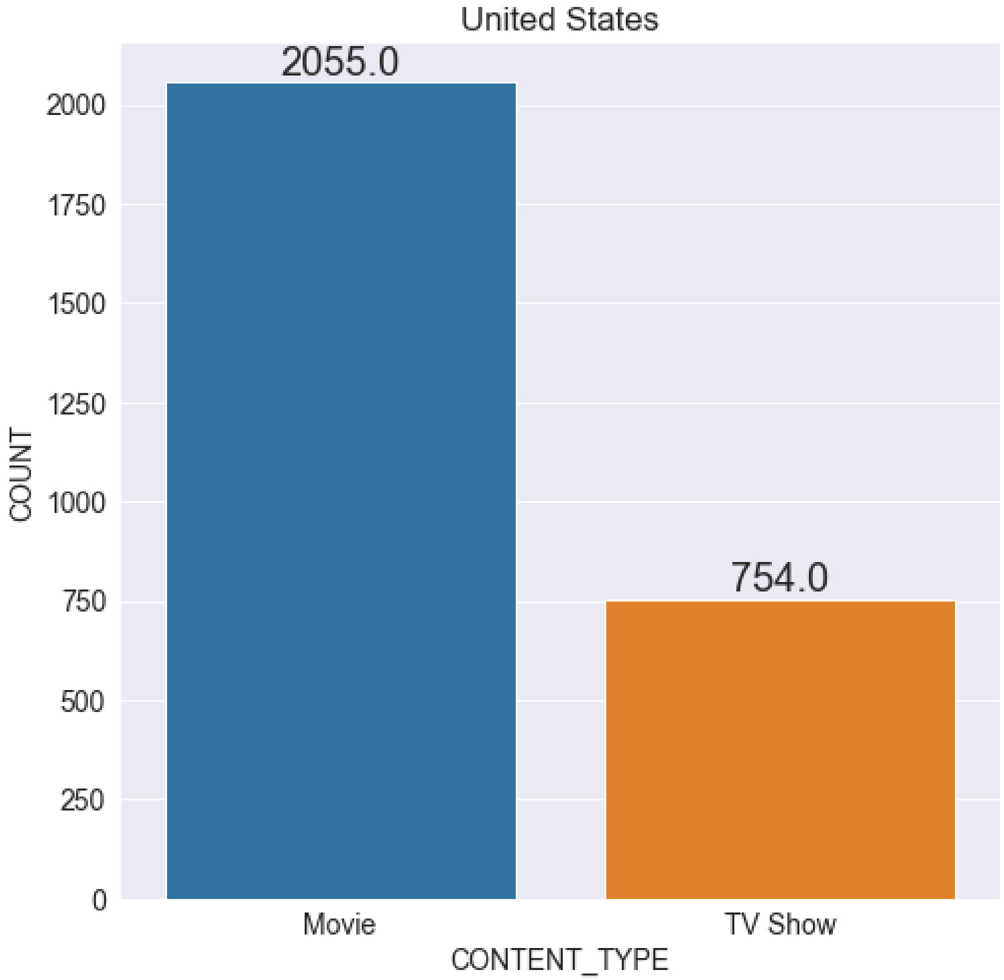
Country	Count
United States	3680
India	1046
United Kingdom	829
Canada	418
France	243
Japan	199
Spain	181
South Korea	145
Germany	124
Mexico	110

Name: country, dtype: int64

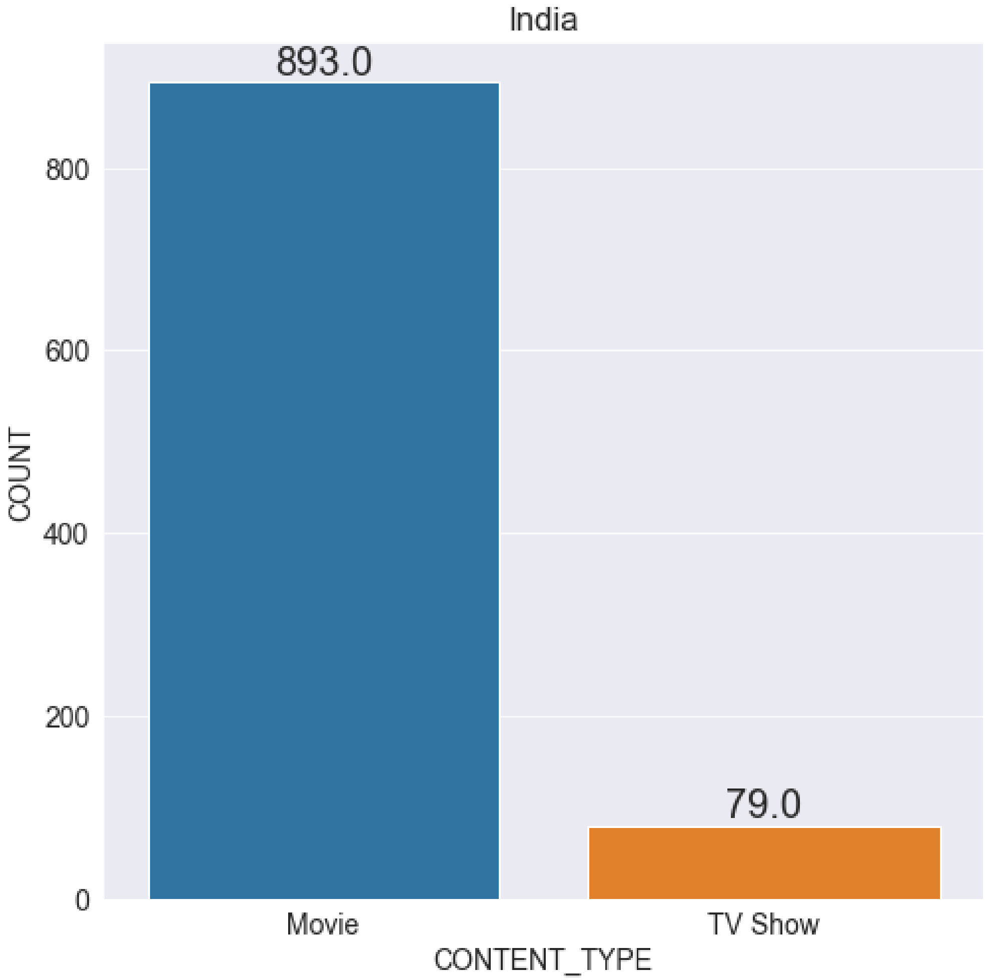


Understanding what content is available in different countries

United States

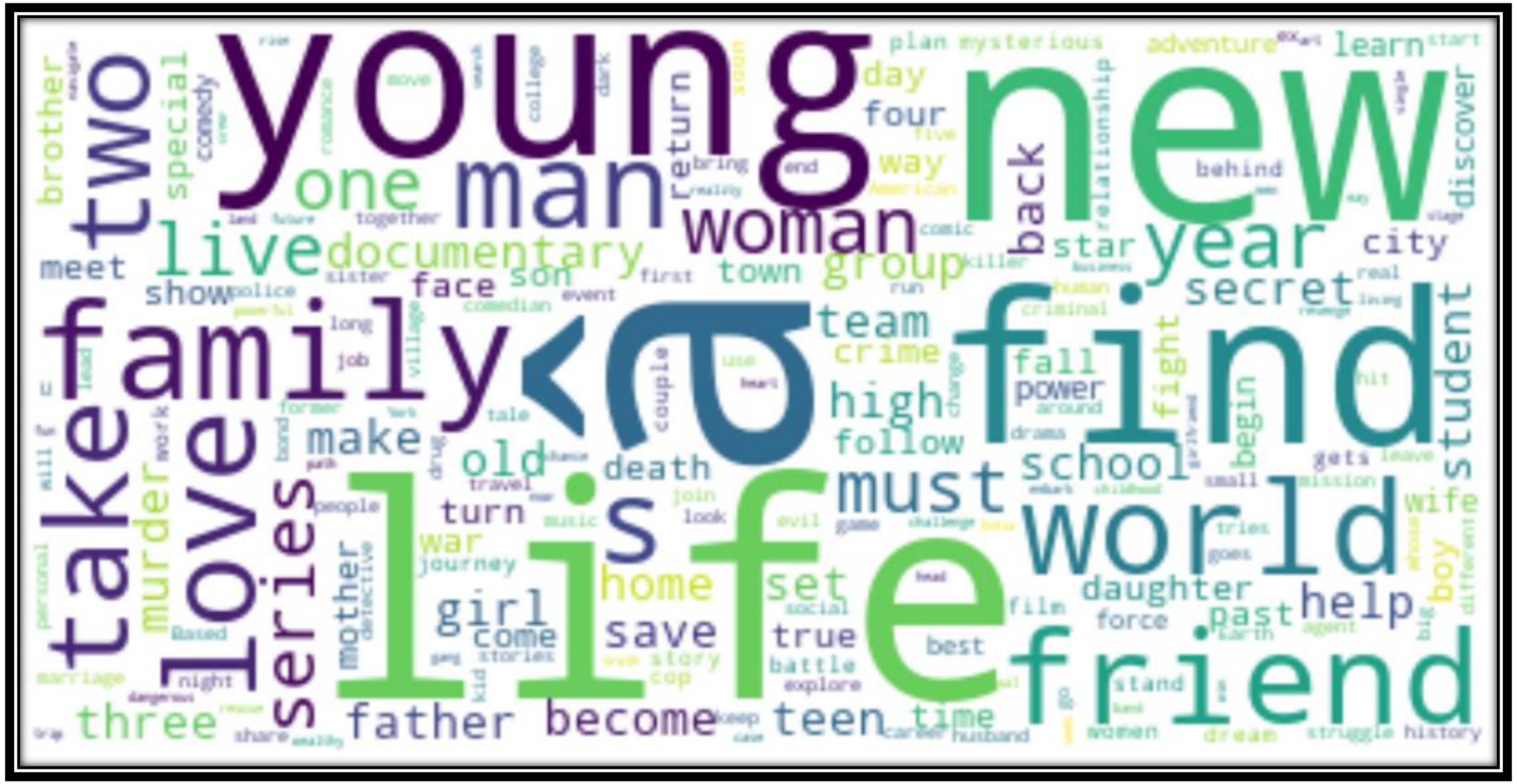


India



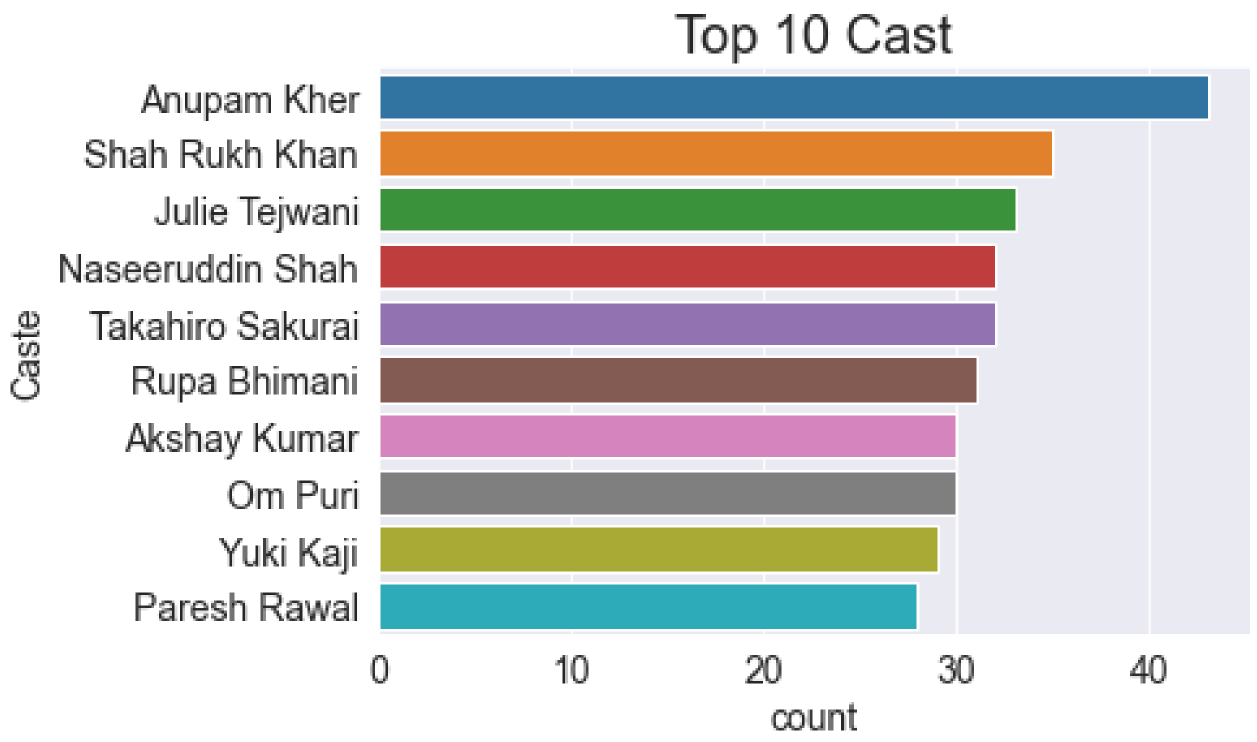
Identifying similar content by matching text-based features

```
text = " ".join(description for description in netflix_overall.description)
word_cloud = WordCloud(collocations = False, background_color = 'white').generate(text)
plt.figure(figsize = (20, 10))
plt.imshow(word_cloud, interpolation = 'bilinear')
plt.axis("off")
plt.show()
```

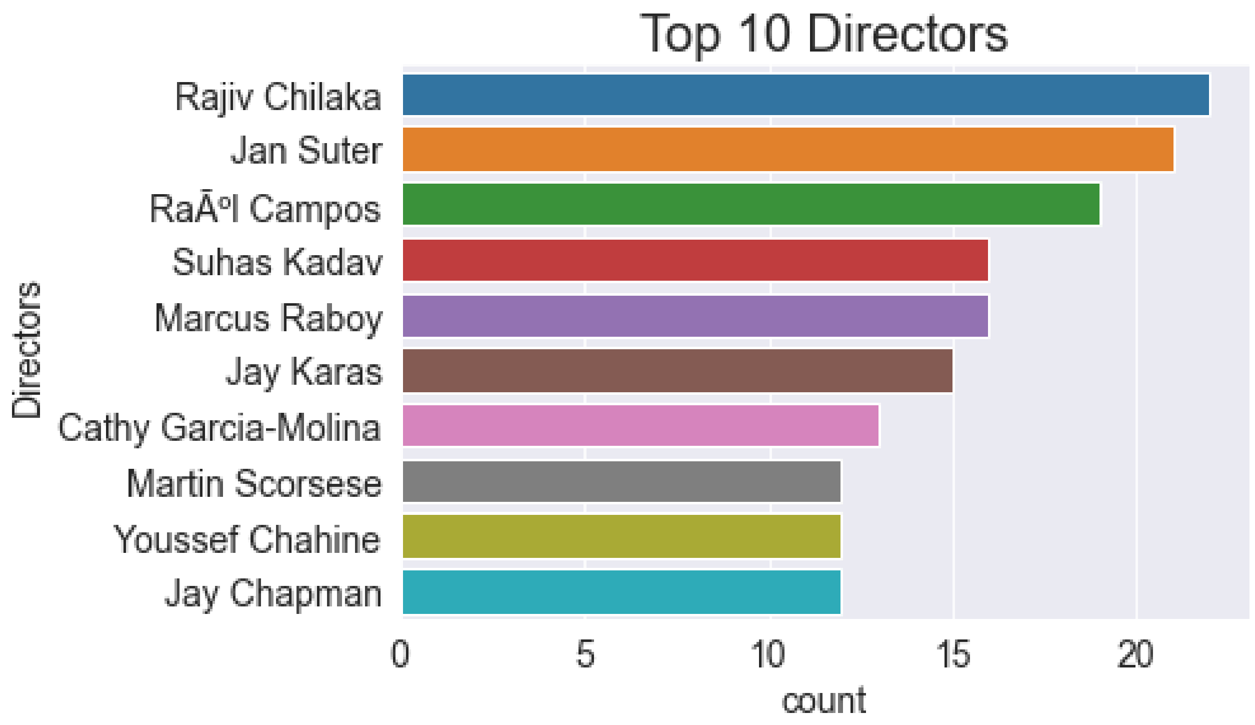


Network analysis of Actors / Directors and find interesting insights

TOP 10 ACTORS (CAST)

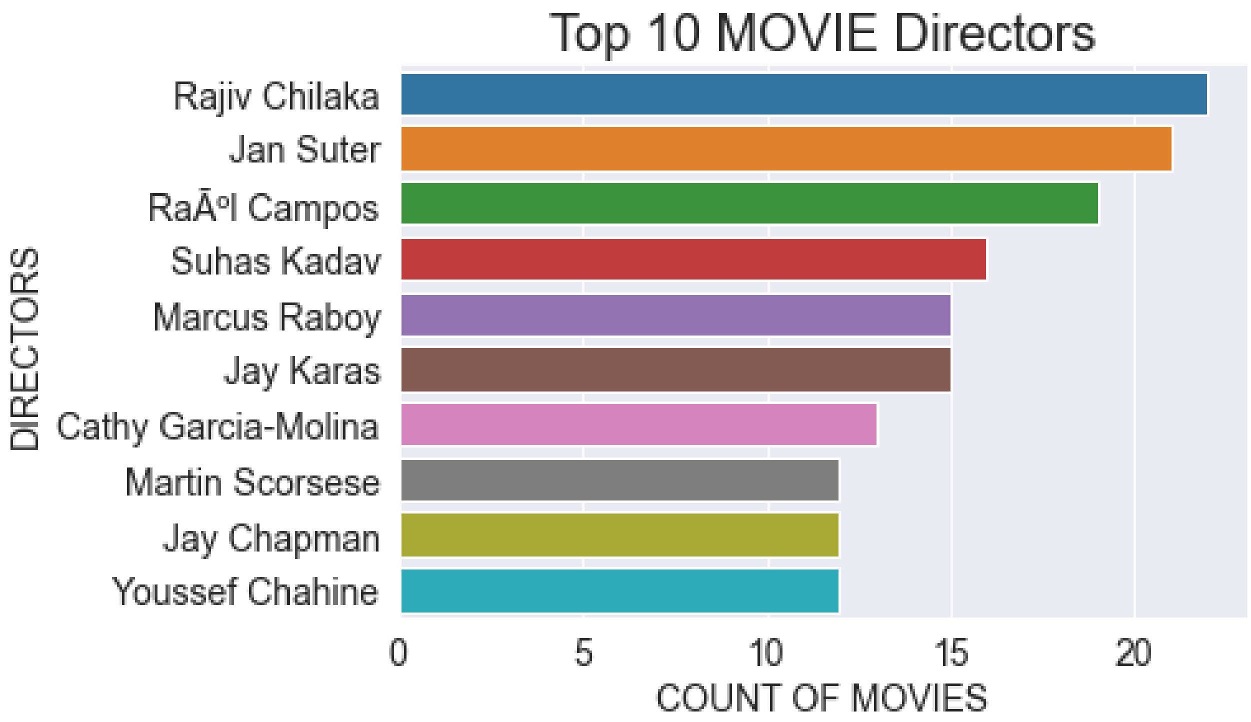


TOP 10 DIRECTORS

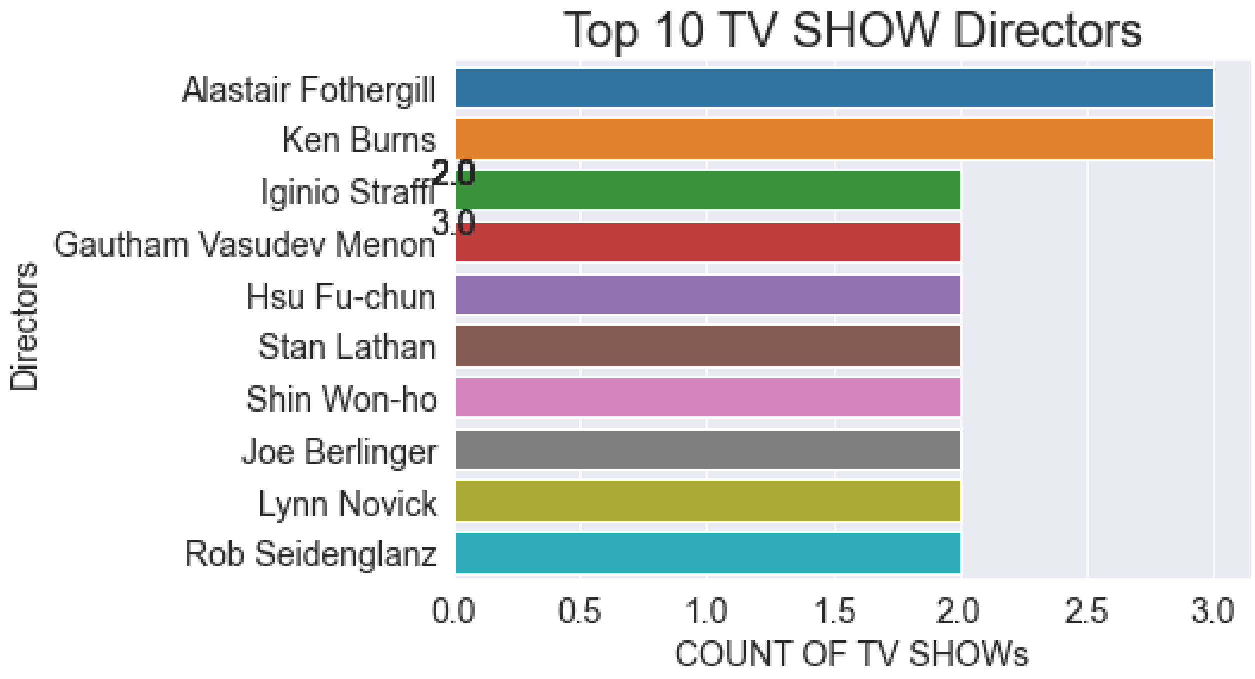


Network analysis of Directors

TOP 10 DIRECTORS IN MOVIES

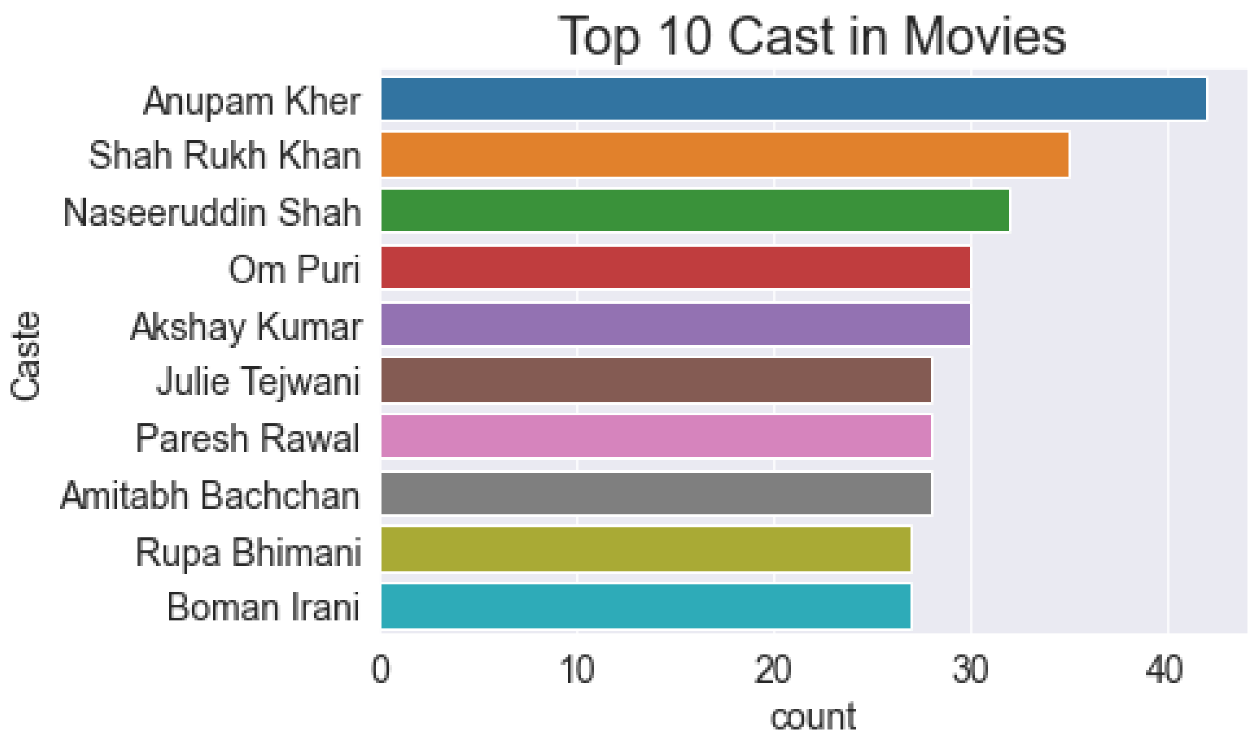


TOP 10 DIRECTORS IN SHOWS

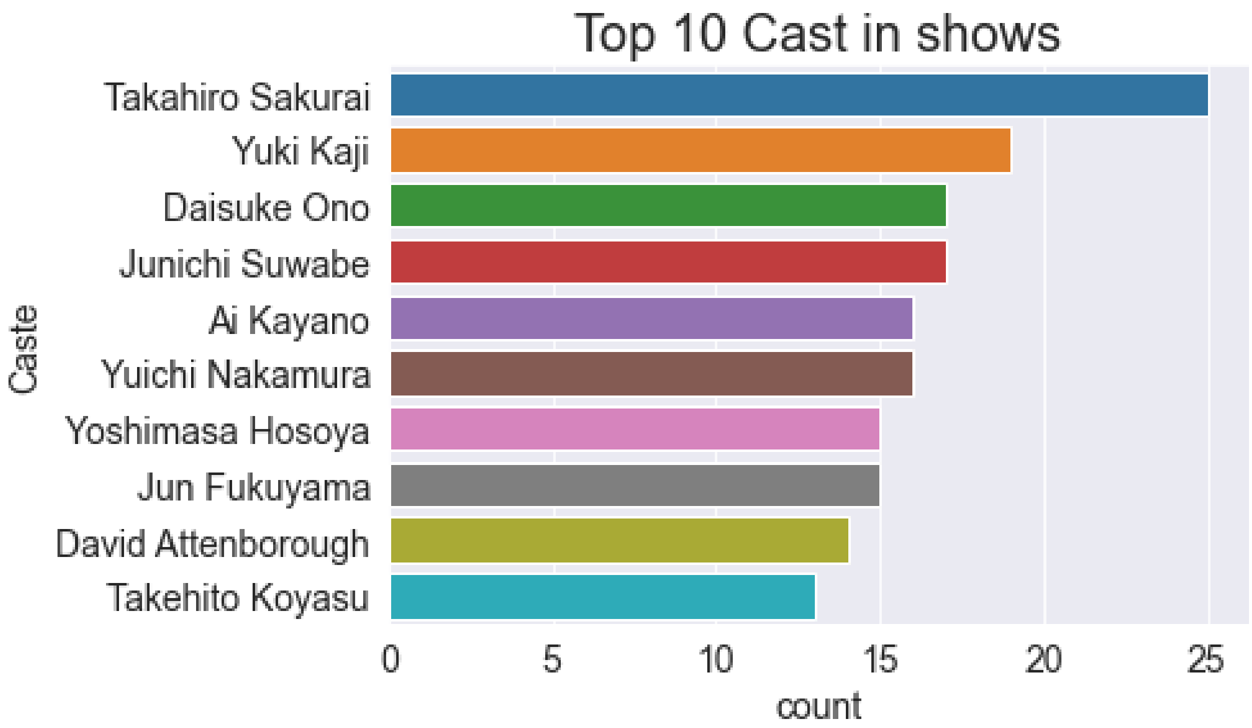


Network analysis of Actors

TOP 10 CAST IN MOVIES

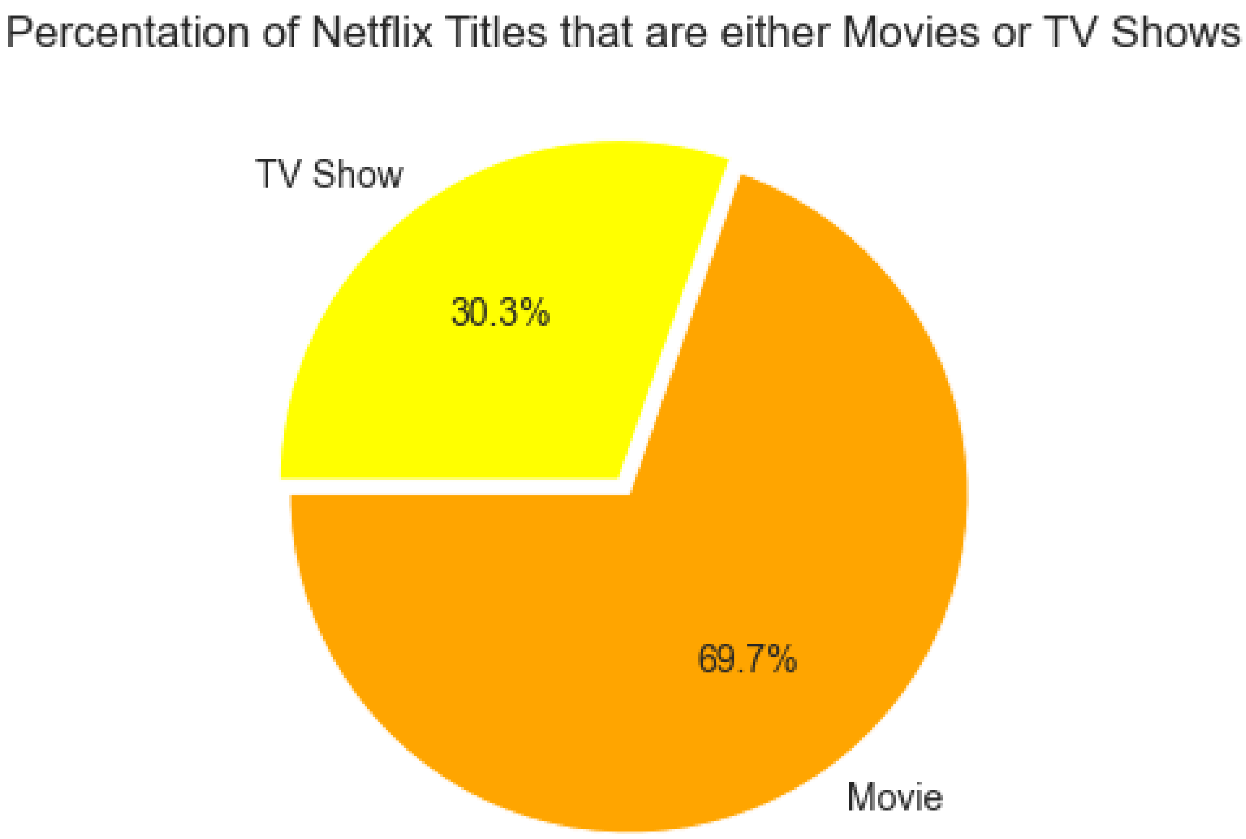


TOP 10 CAST IN SHOWS

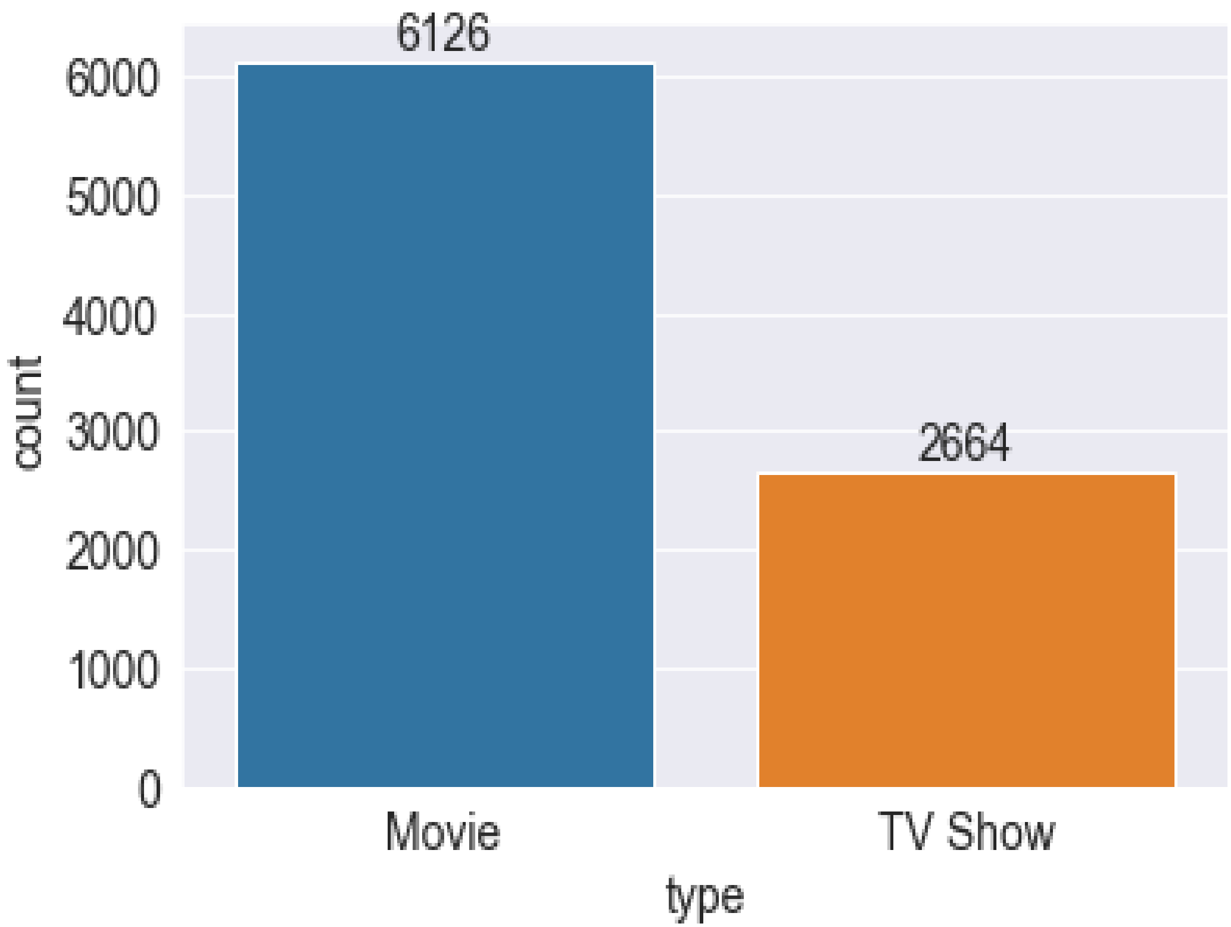


Does Netflix has more focus on TV Shows than movies in recent years

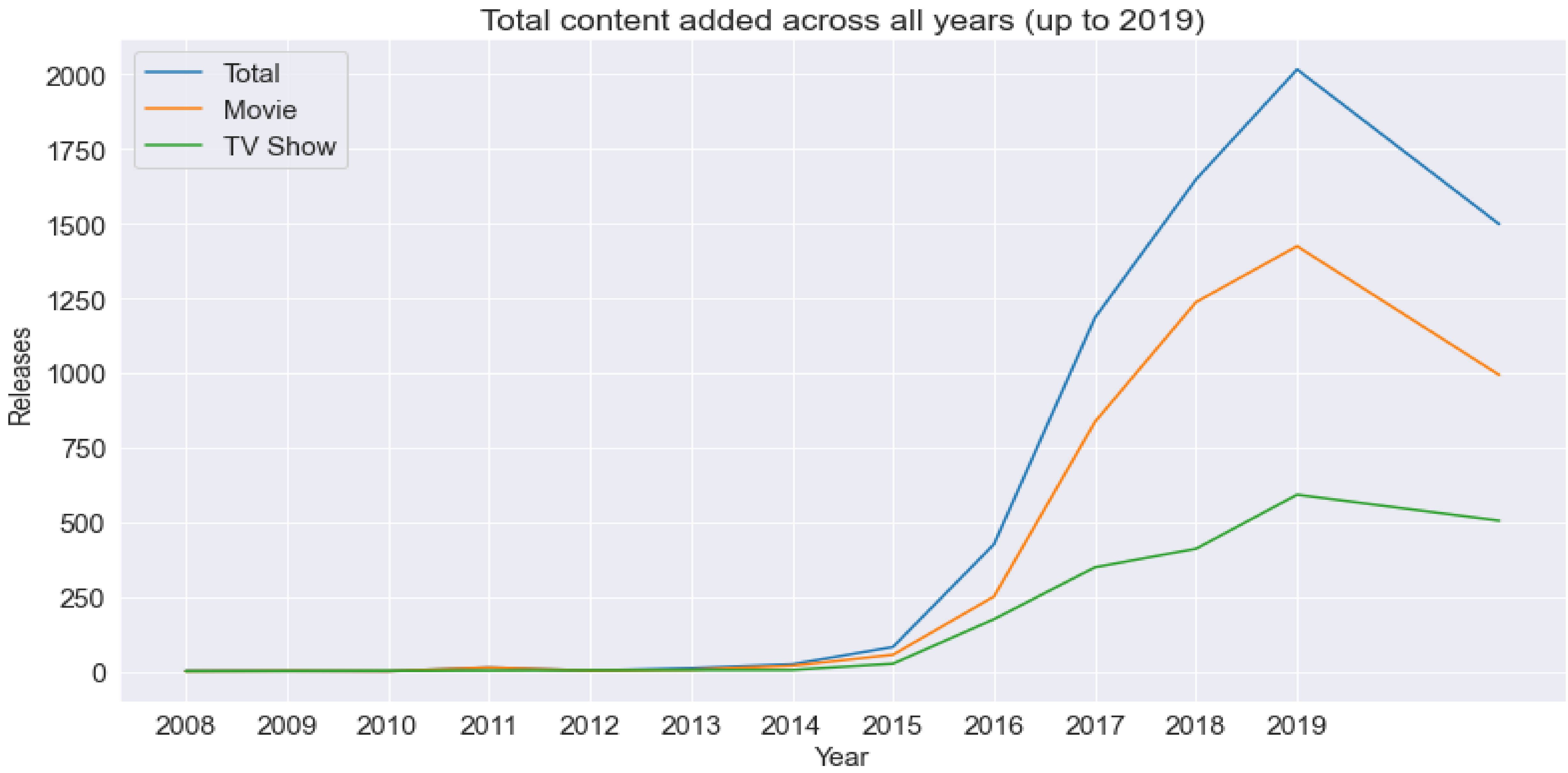
PIE CHART PERCENTAGE INFORMATION



BAR CHART COUNT INFORMATION

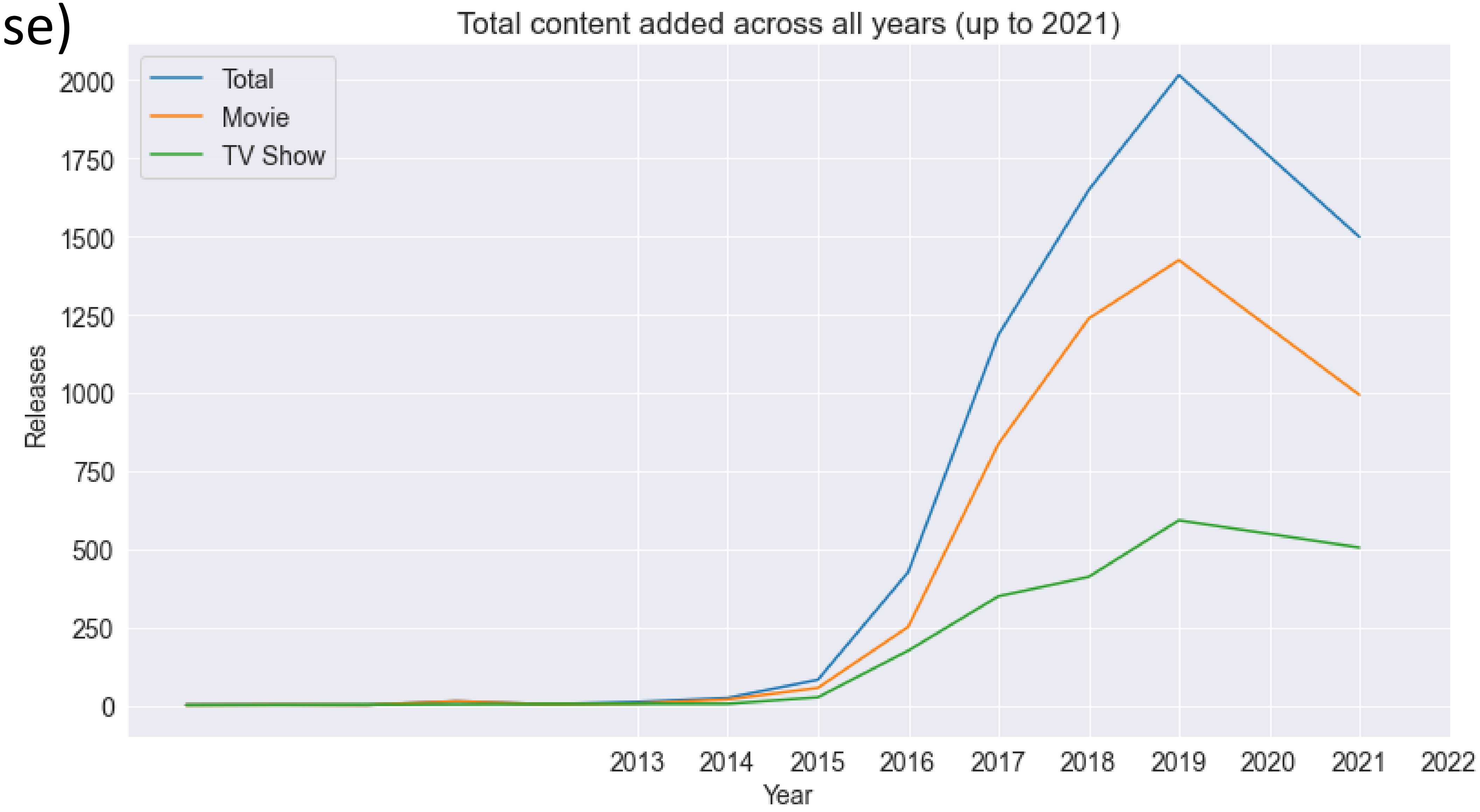


Does Netflix has more focus on TV Shows than movies in recent years



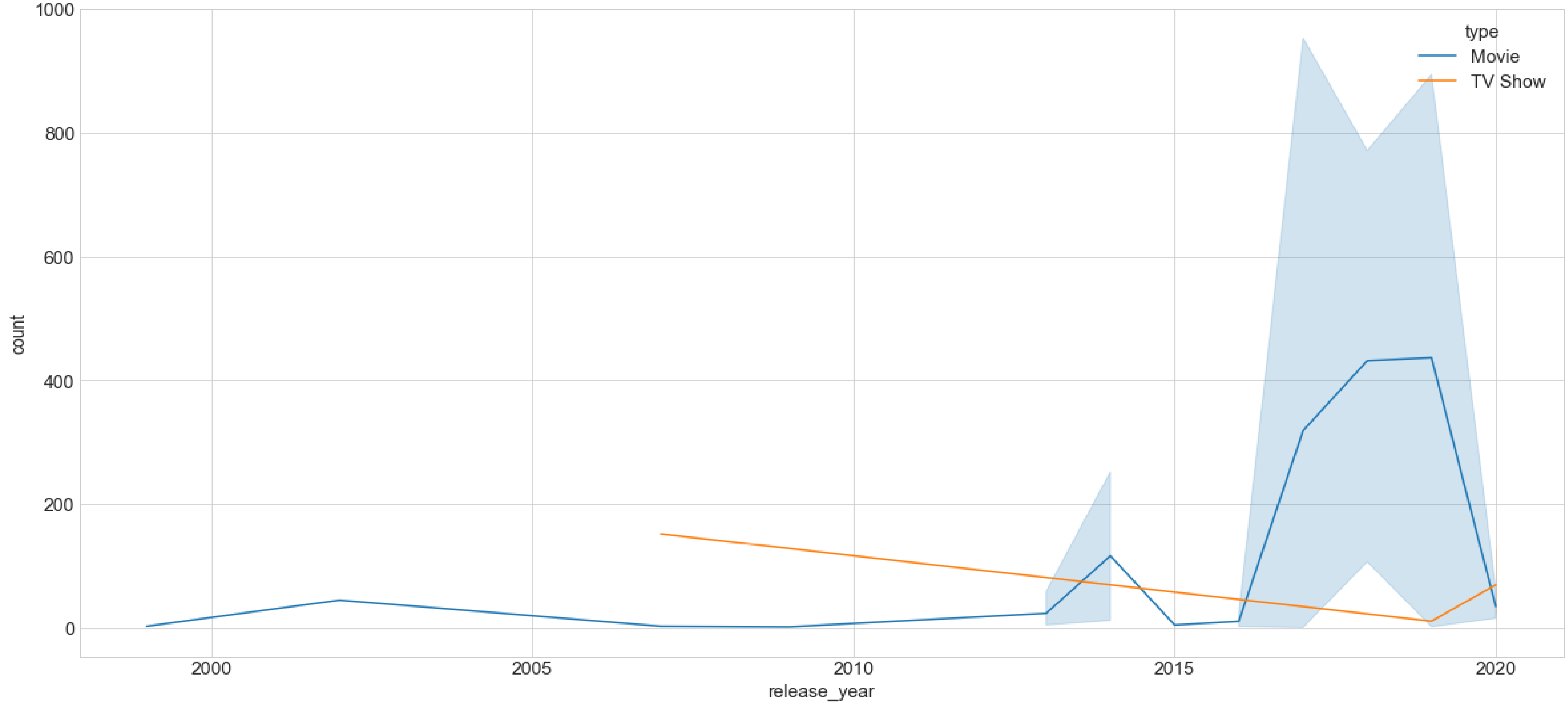
Does Netflix has more focus on TV Shows than movies in recent years

- FROM 2013 TO 2021 Observe (movies completely down, TV shows slant decrease)



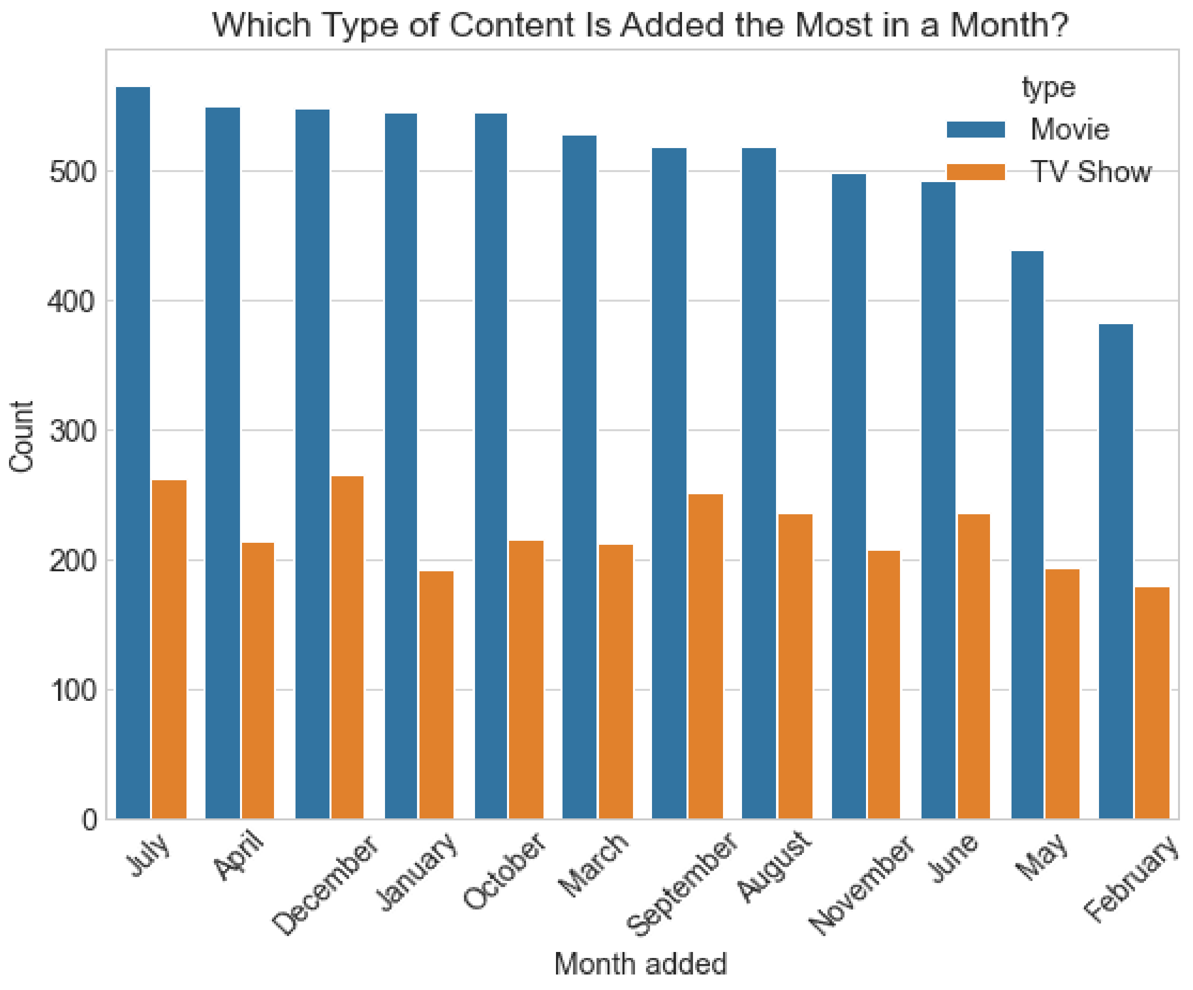
Does Netflix has more focus on TV Shows than movies in recent years

Slant increment of TV shows will be observed in the graph

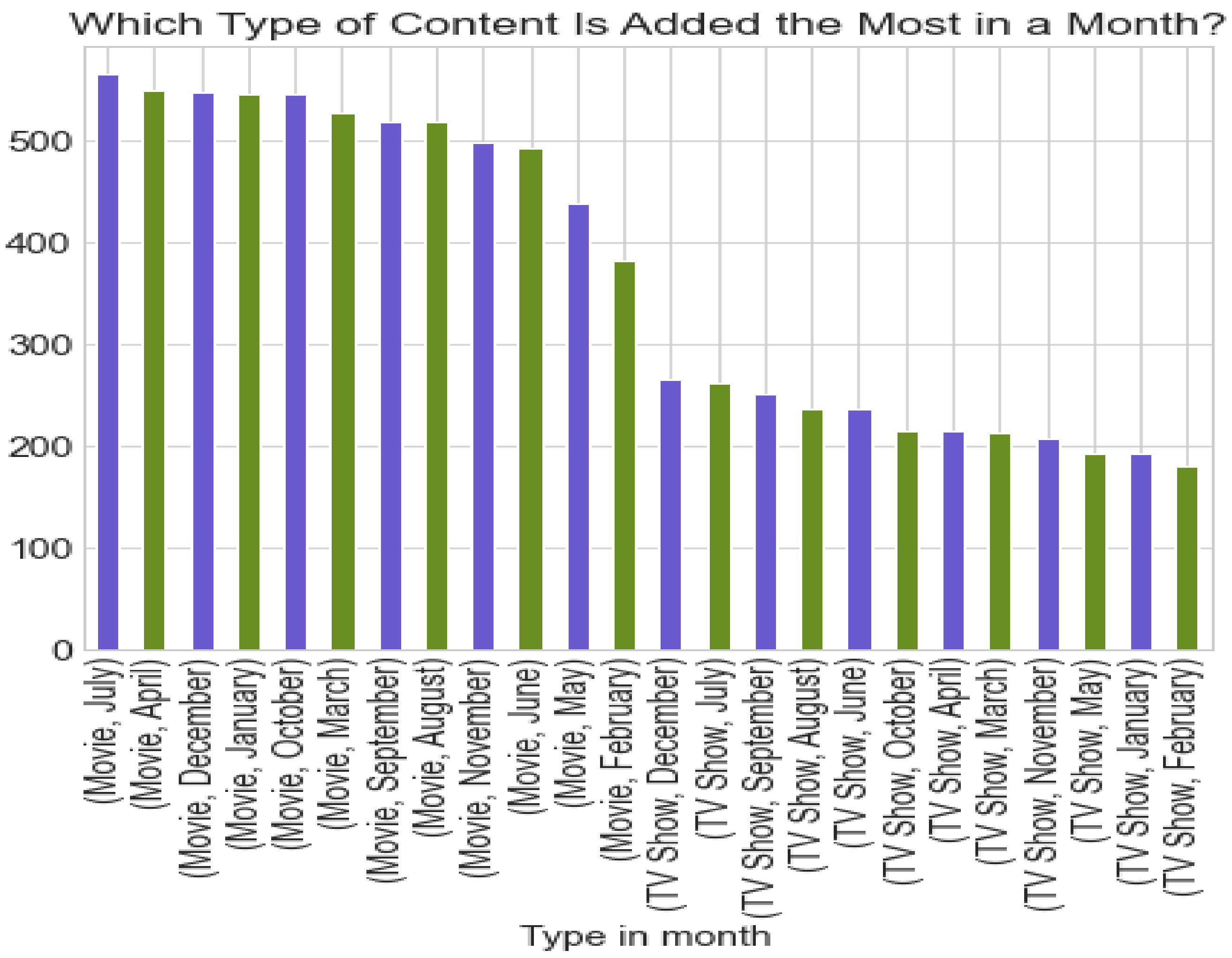


Does Netflix has more focus on TV Shows than movies in recent months

CONTENT ADDED THE MOST IN A MONTH



CONTENT ADDED THE MOST IN A MONTH



THE END