Capstone Project-4

Netflix movies and TV shows clustering

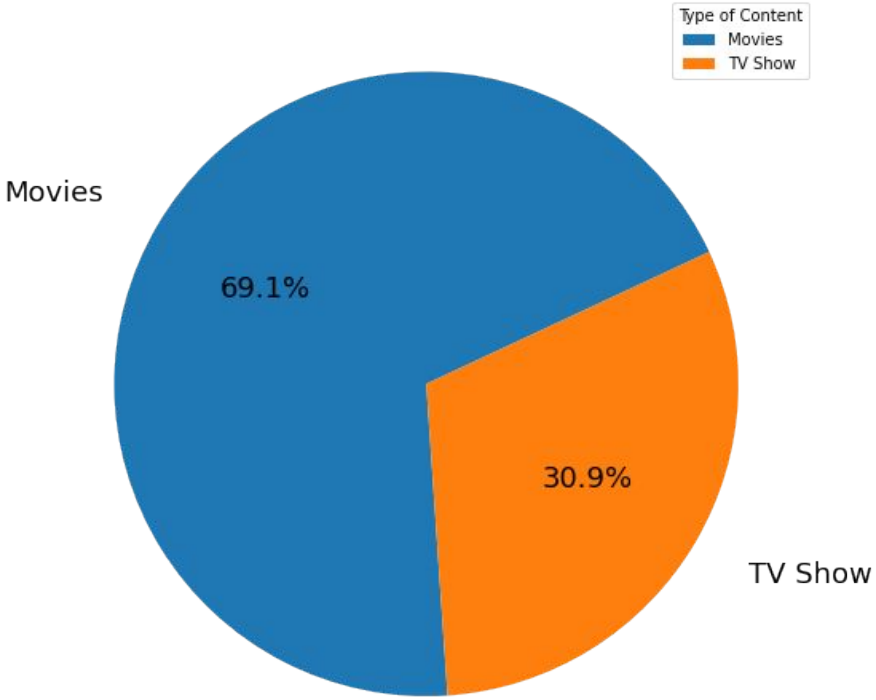
By
Harshavardhan M. Shete

Netflix movies and TV shows clustering

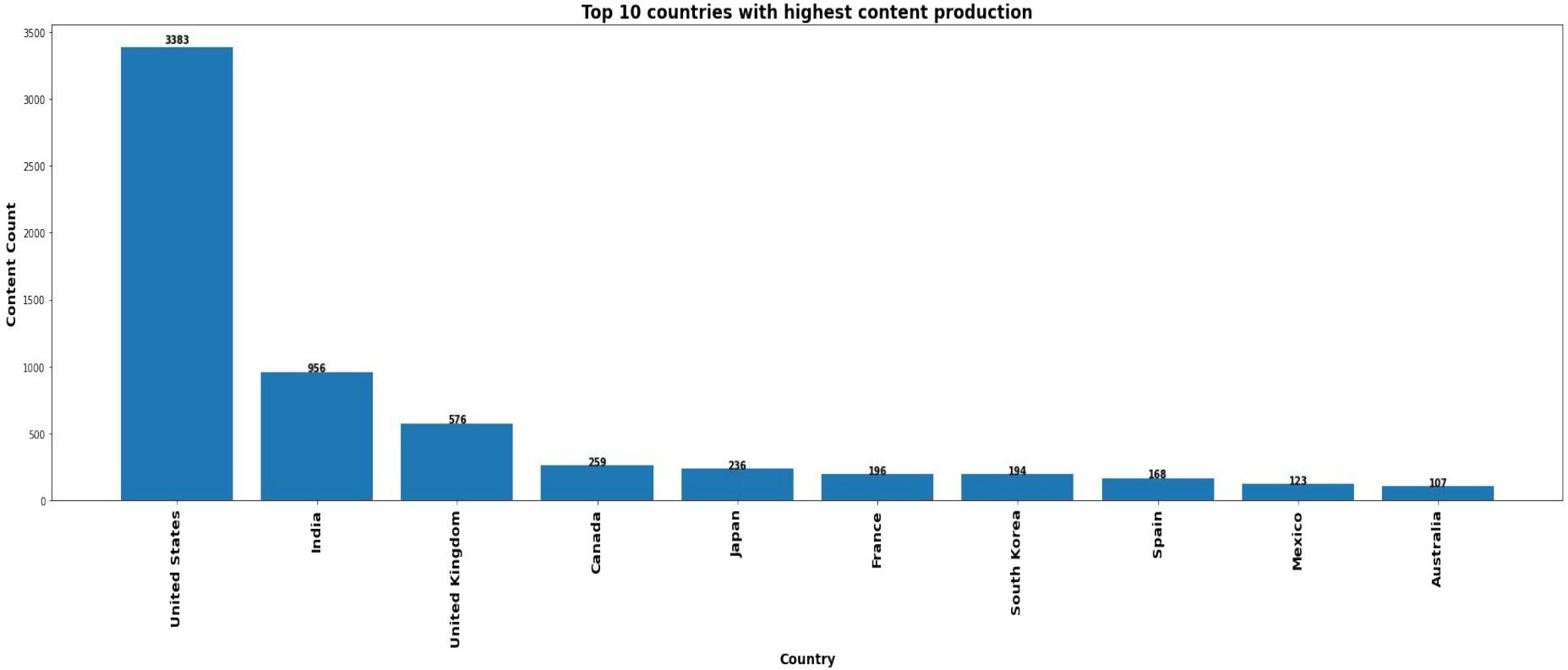
- **What is required to do:**
  1. Exploratory data analysis
  2. Clustering using appropriate clustering algorithm
- **What is the data all about:**
  - TV shows and movies available on Netflix till 2019
  - Shape: 7787 rows, 12 columns
  - Attributes: Show id, Type, Title, Director, Cast, Country, Date added, Release year, Rating, Duration, Listed in and Description

Exploratory Data Analysis

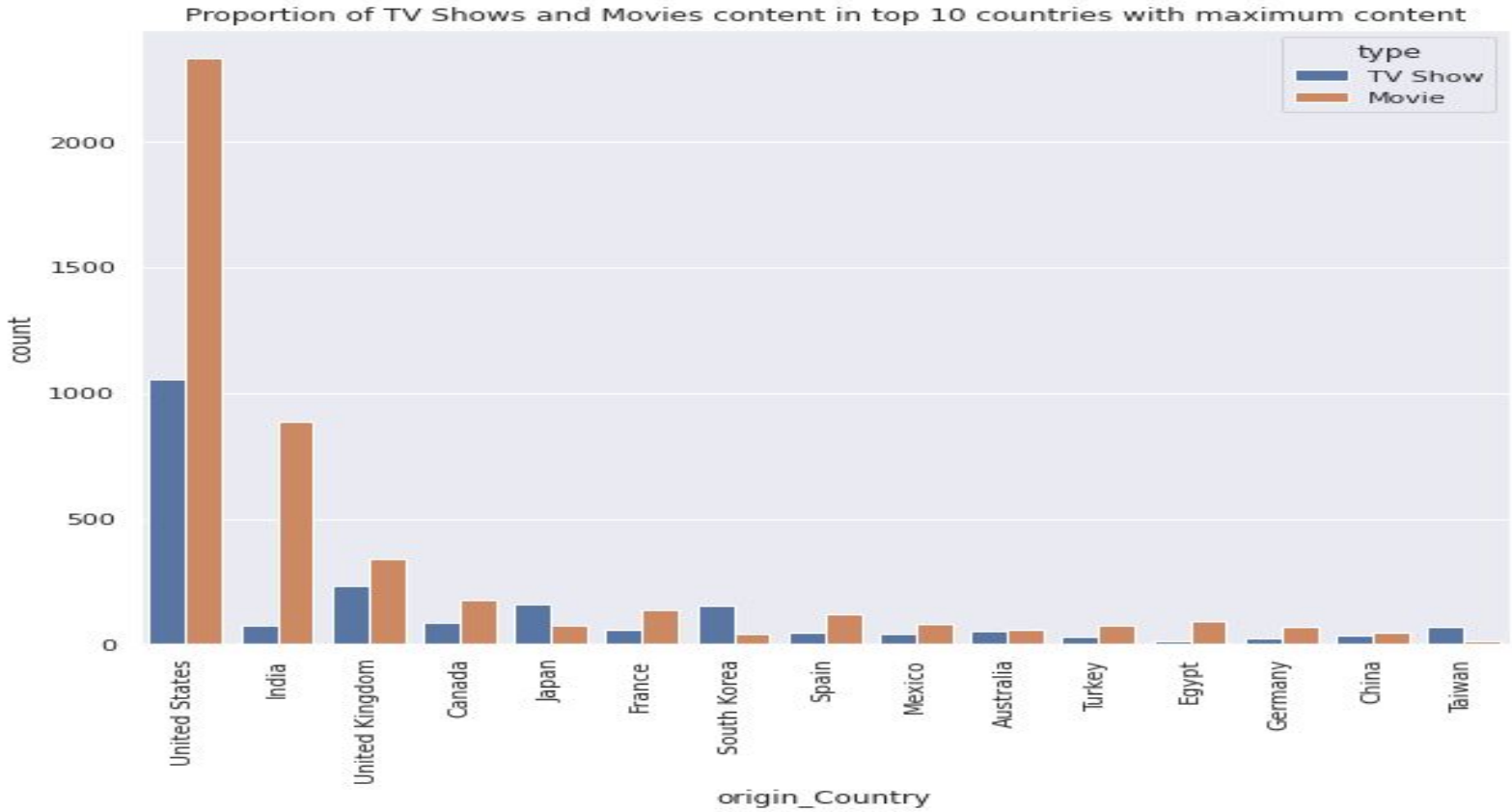
Proportion of movies and TV shows



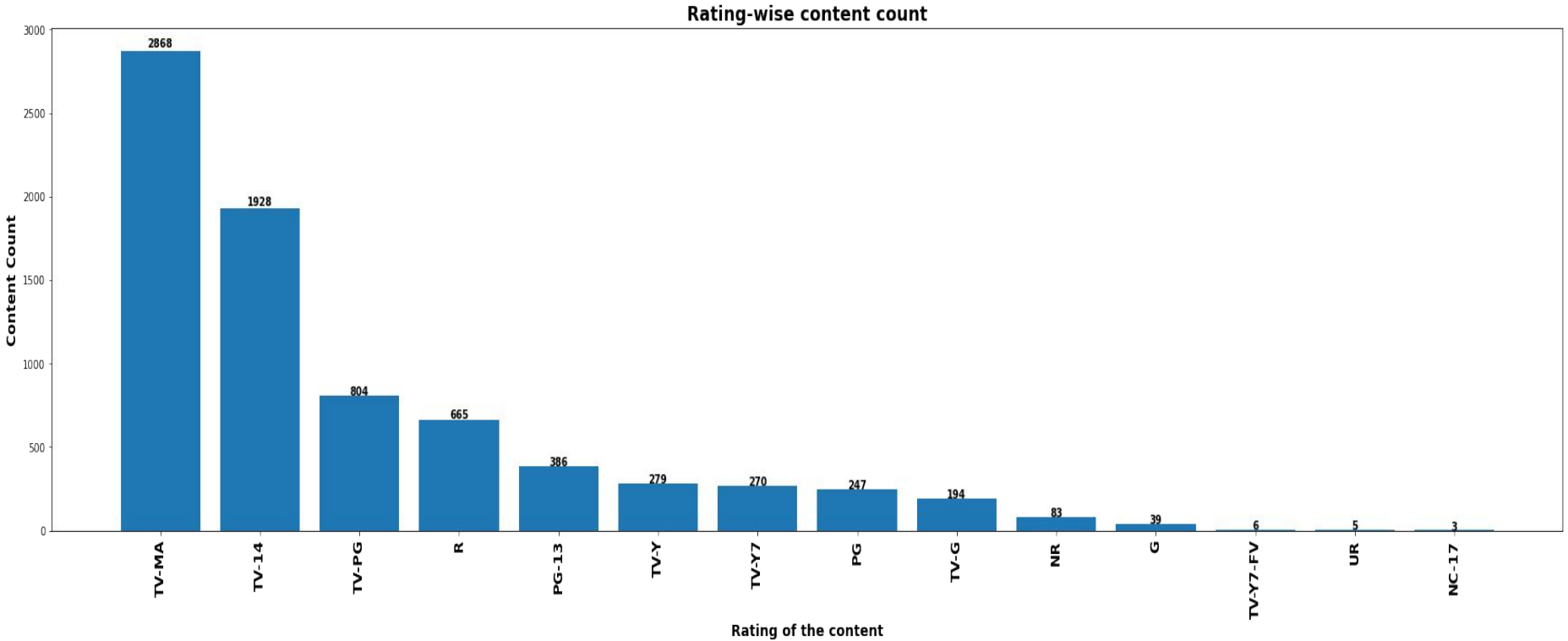
Top 10 countries with highest content production



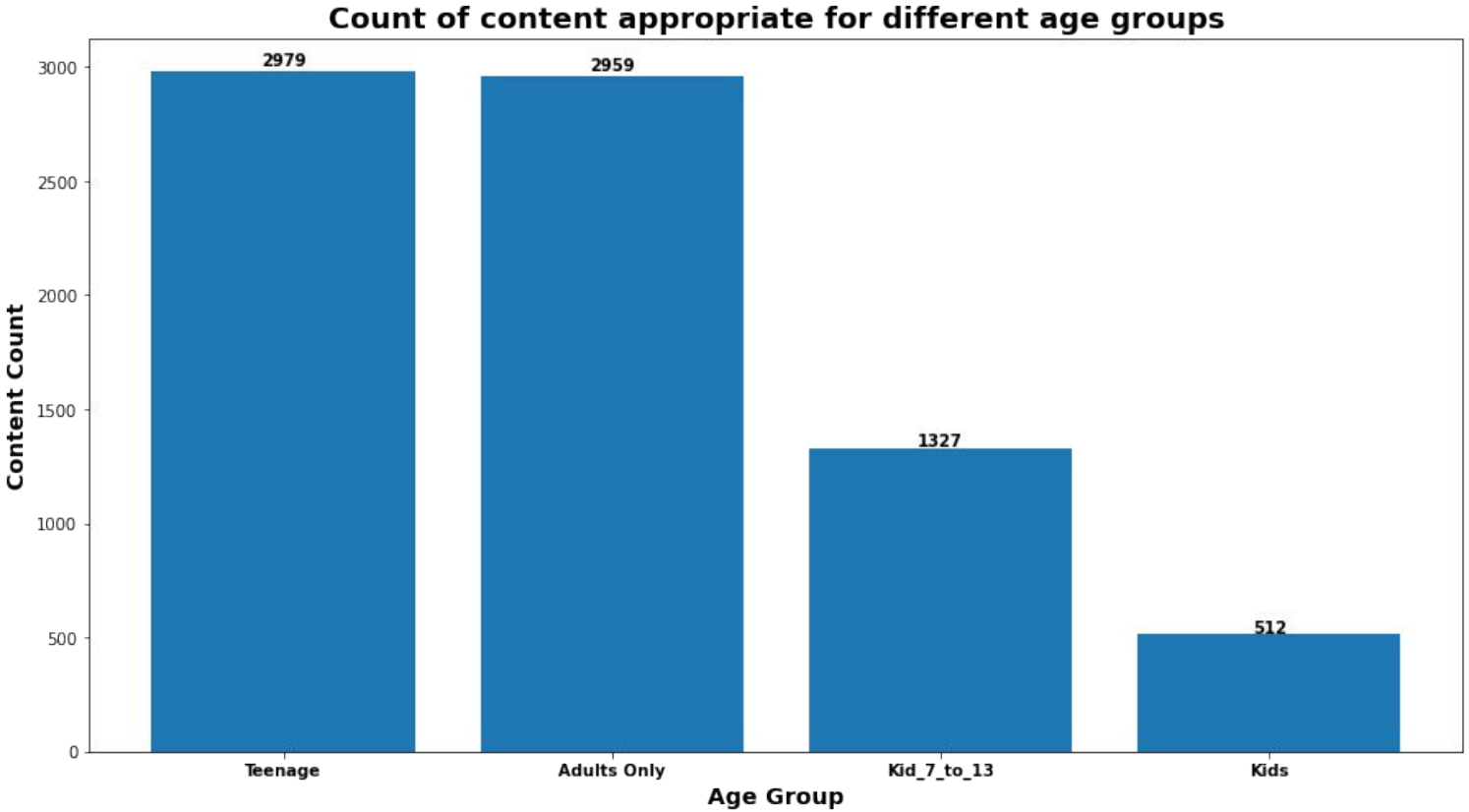
Proportion of TV shows and movies



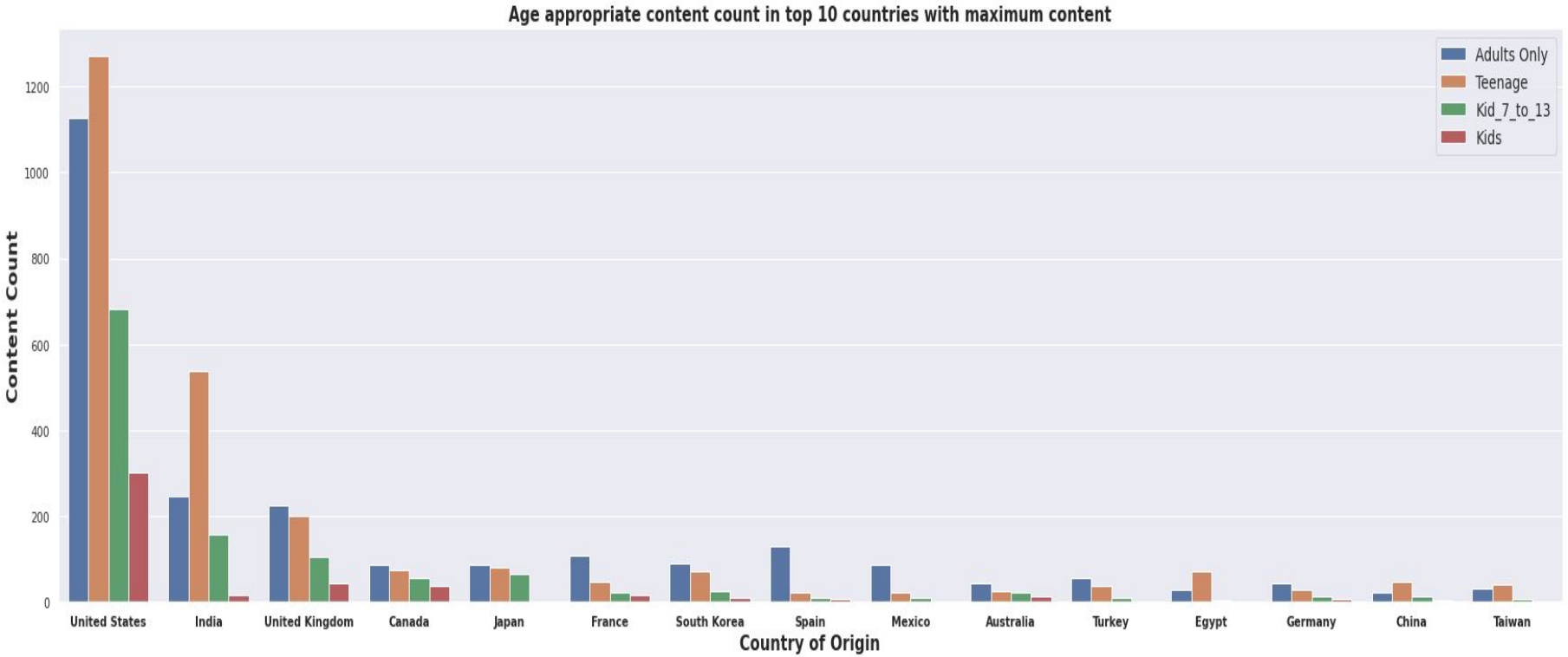
Rating wise content count



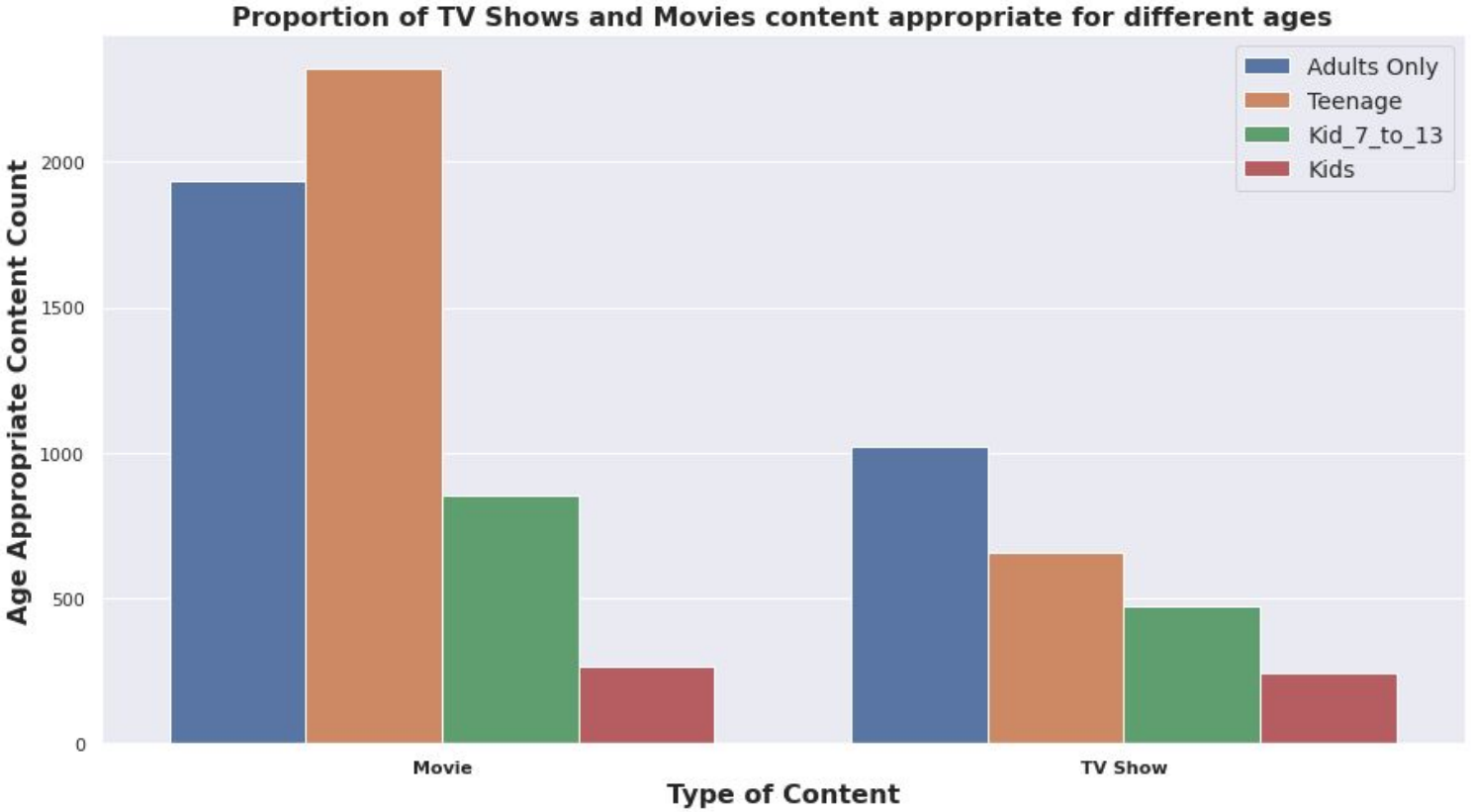
Age appropriate content count



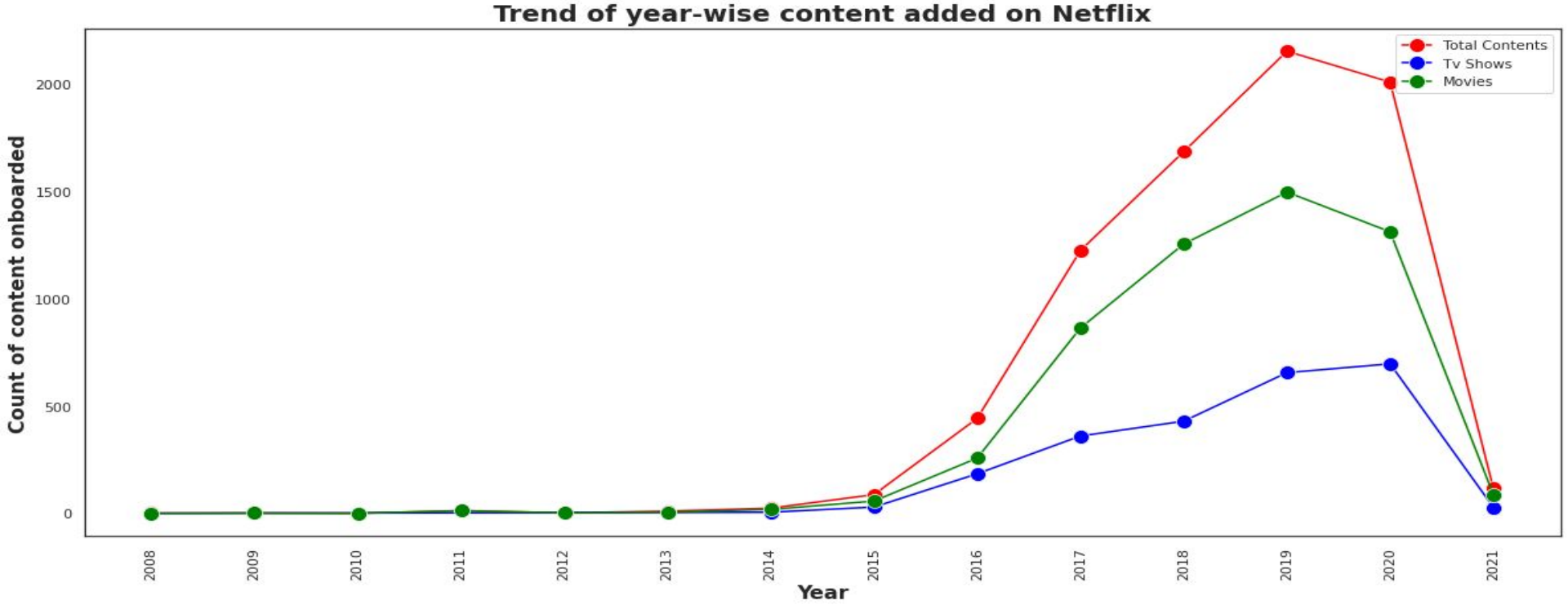
Age appropriate content in top 10 countries



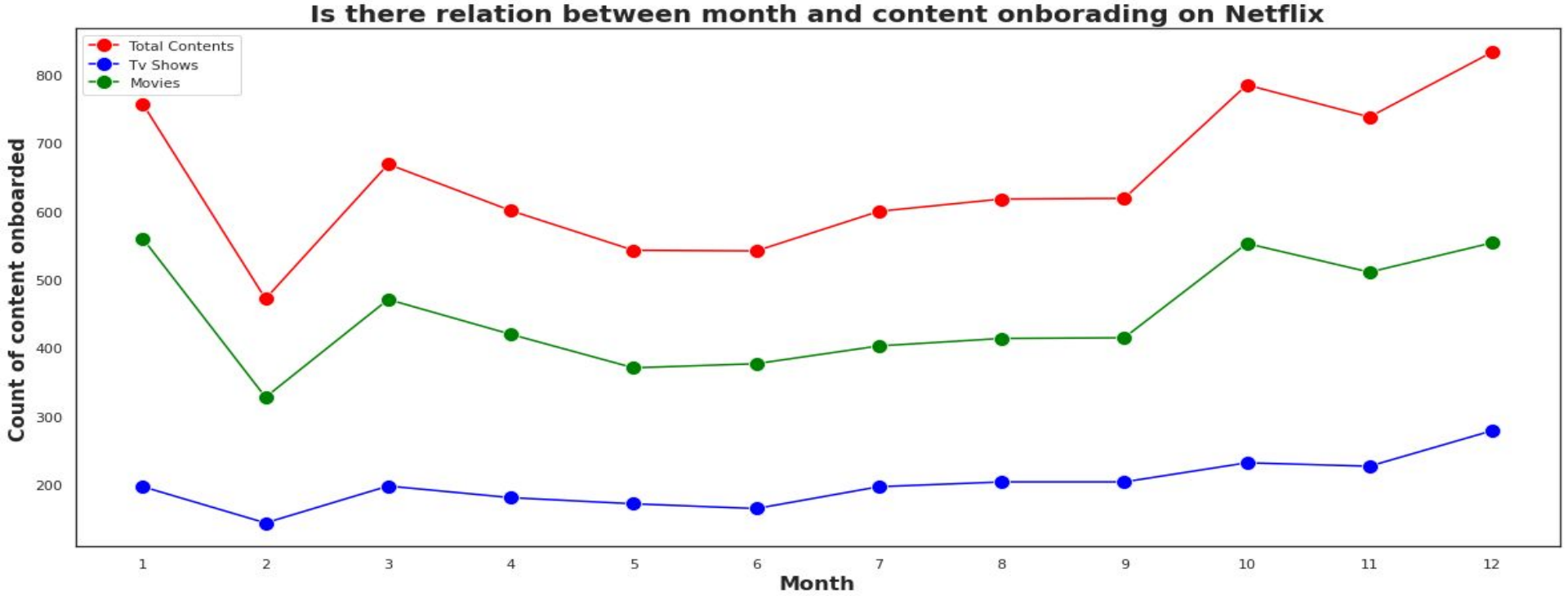
Age appropriate content count in TV shows and movies



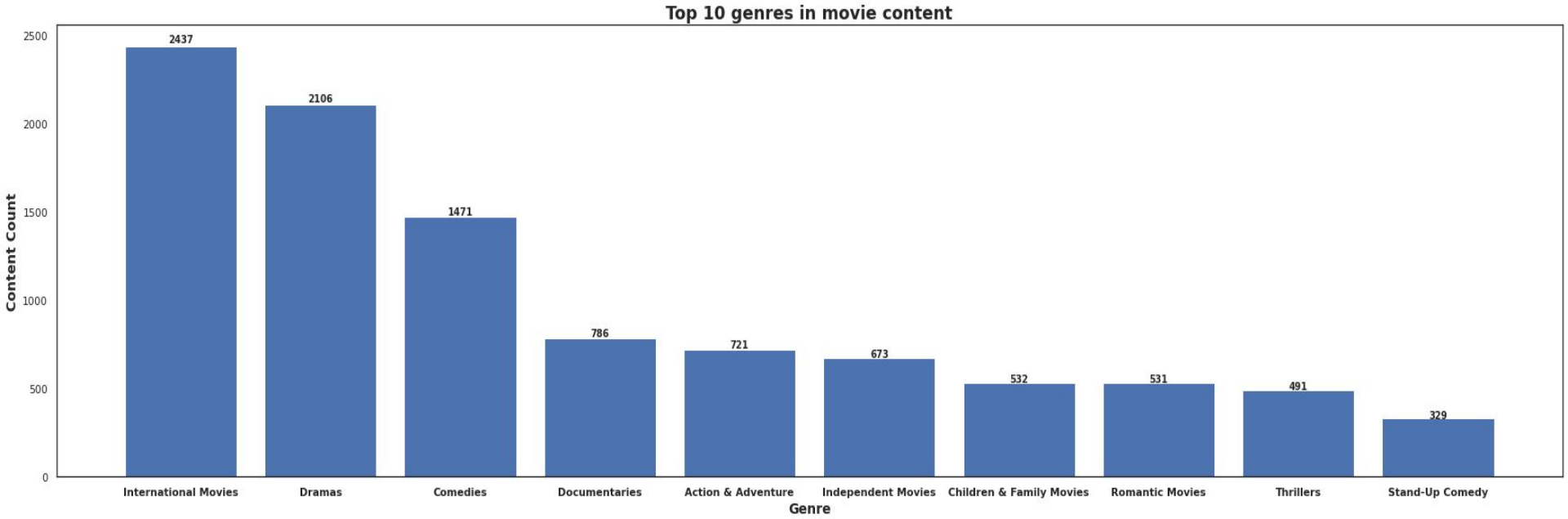
Trend of year-wise content on-boarded on Netflix



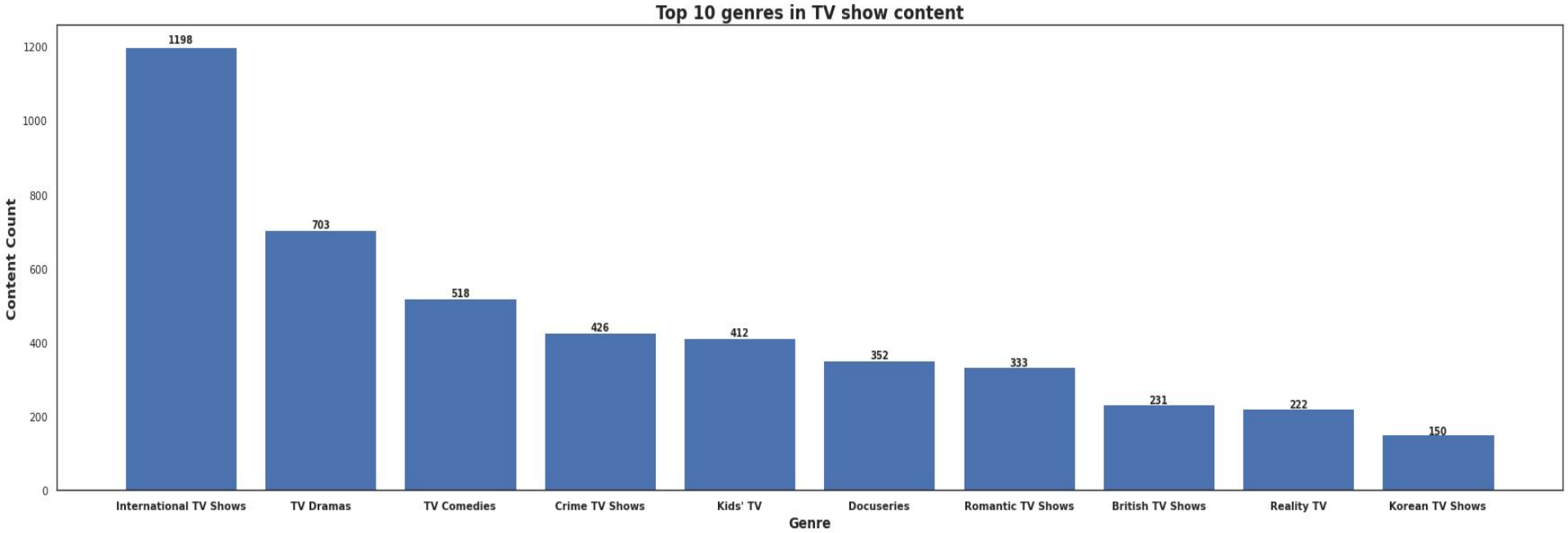
Month vs. Content Onboarding on Netflix



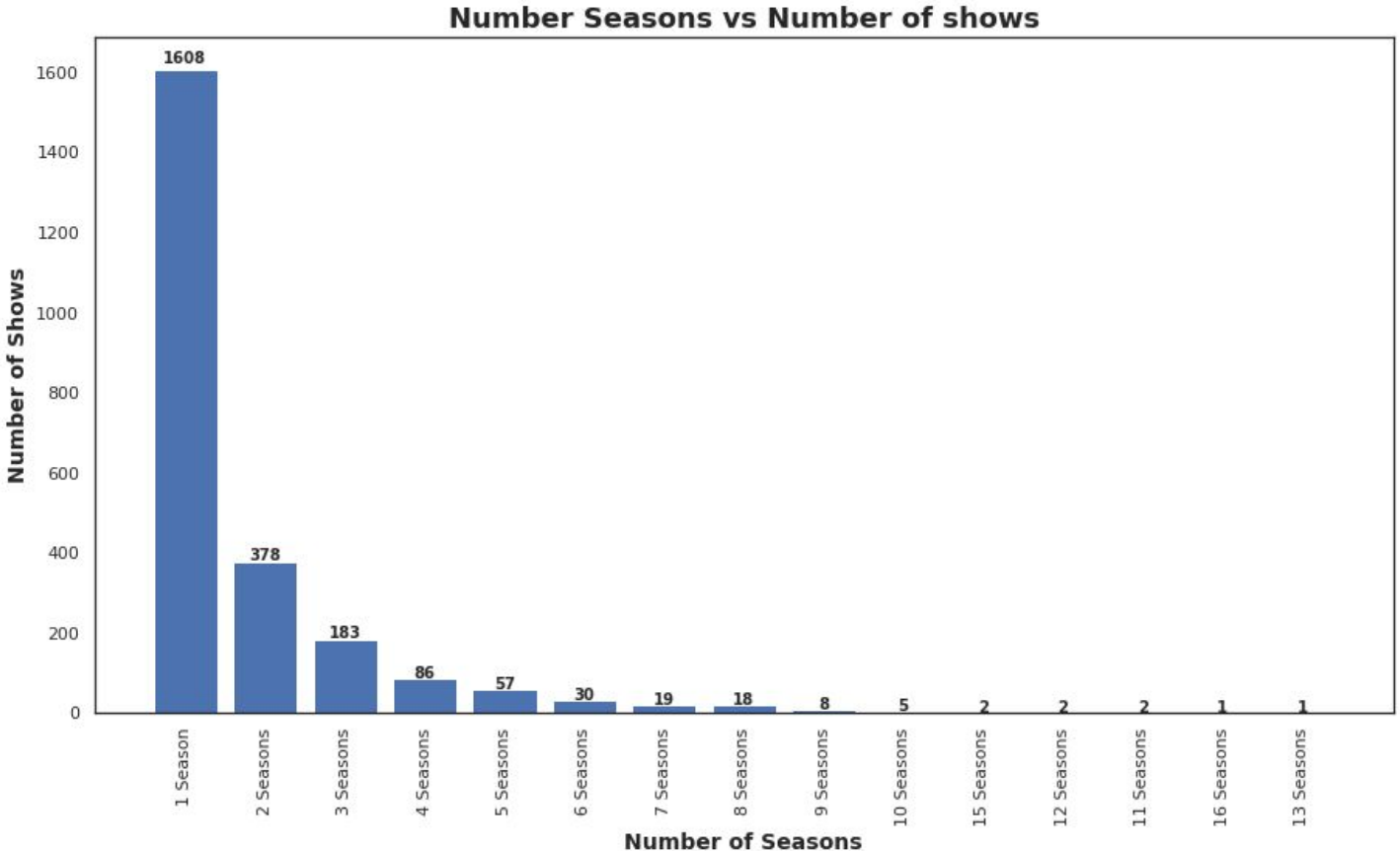
Top 10 genres in movies



Top 10 genres in TV shows



Season vs. Number of TV shows

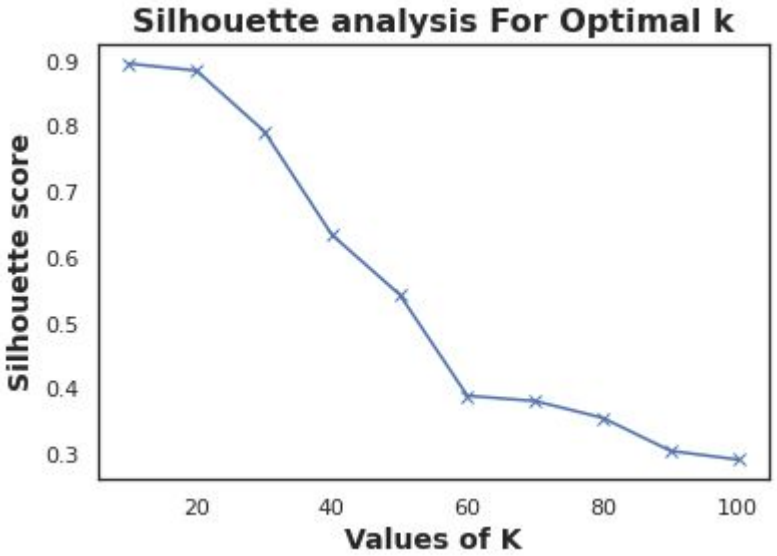


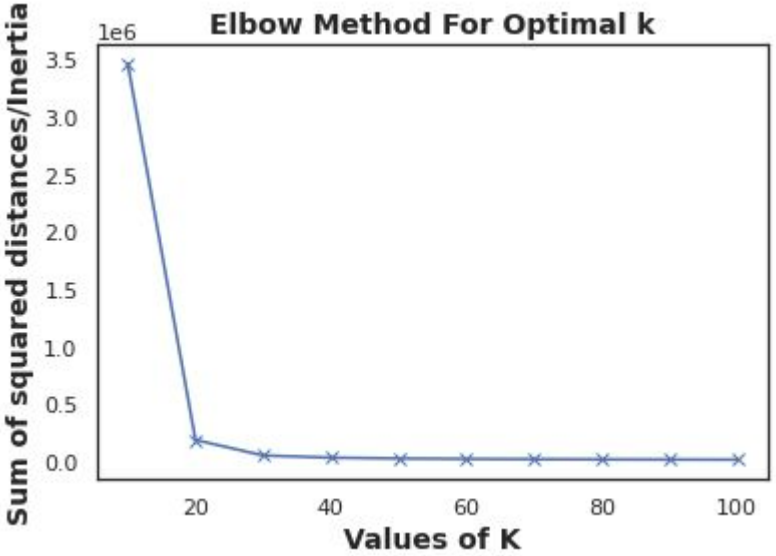
K means Clustering

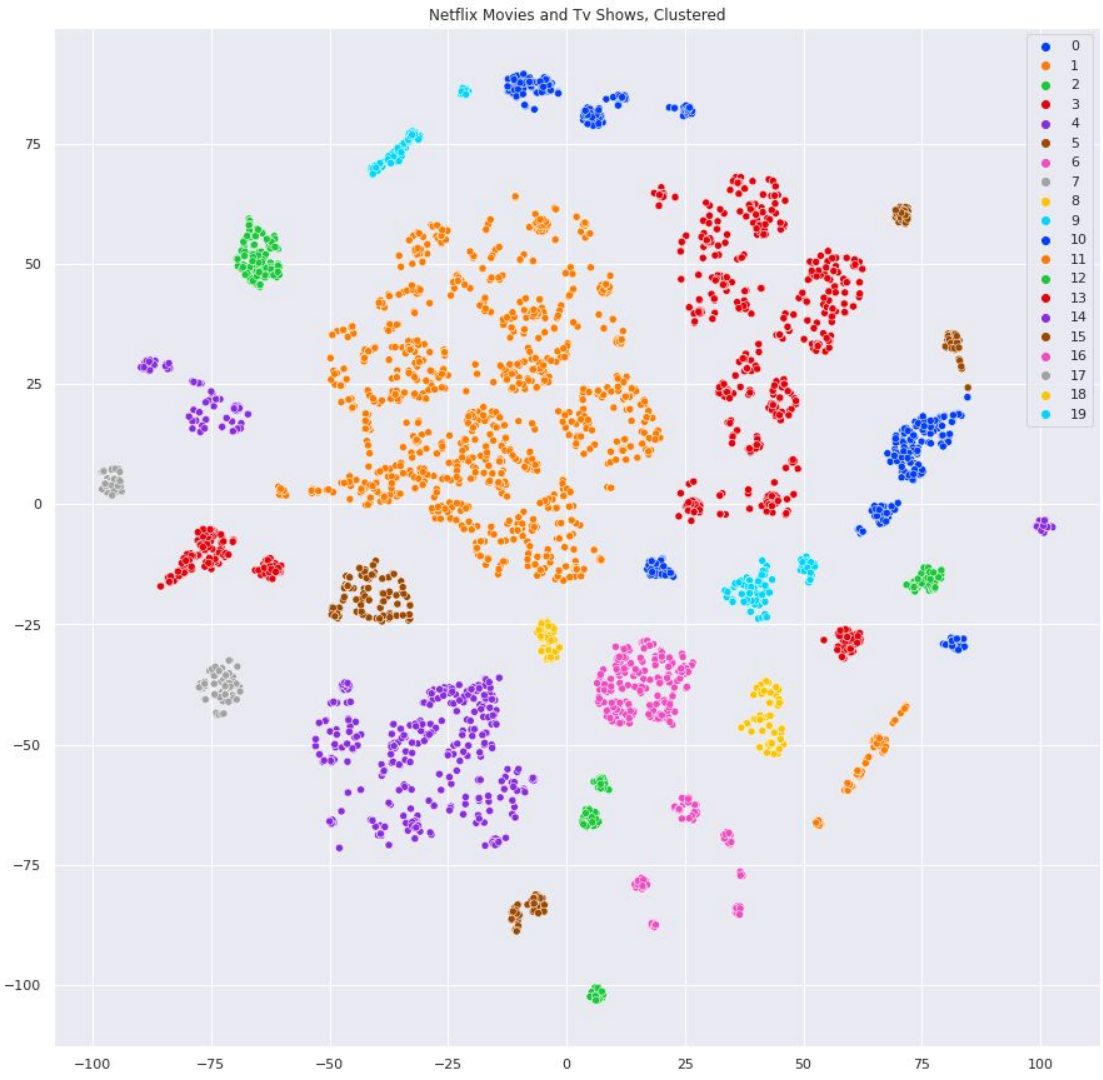
How does it work?

1. To begin, we first select a number of classes/groups to use and randomly initialize their respective center points. To figure out the number of classes to use, it's good to take a quick look at the data and try to identify any distinct groupings.
2. Each data point is classified by computing the distance between that point and each group center, and then classifying the point to be in the group whose center is closest to it.
3. Based on these classified points, we recompute the group center by taking the mean of all the vectors in the group.
4. Repeat these steps for a set number of iterations or until the group centers don't change much between iterations.

Optimal k


Silhouette analysis For Optimal k
Silhouette score
0.9
0.8
0.7
0.6
0.5
0.4
0.3
20
40
60
80
100
Values of K


Elbow Method For Optimal k
1e6
Sum of squared distances/Inertia
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
20
40
60
80
100
Values of K


Netflix Movies and Tv Shows, Clustered
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
75
50
25
0
−25
−50
−75
−100
−100
−75
−50
−25
0
25
50
75
100

Conclusion

- Major findings from EDA
    - Higher amount of movie content than TV shows
    - Exponential growth in content onboarding in 2015-16 and sudden deep in 2021
    - USA has the highest amount of content
    - Overall content appropriate for kids is very less as compared with adult and teen appropriate content
- Optimal value of clusters is “20” for K-mean clustering