# EXPERIMENT 6

| | |
|---|---|
| **NAME** | **Shreya Shetty** |
| **UID** | **2019140059** |
| **CLASS** | **TE IT** |
| **BATCH** | **A** |
| **SUBJECT** | **Big Data Analytics Lab** |

**AIM:** Demonstrate use of modern tools like R/Matlab for EDA

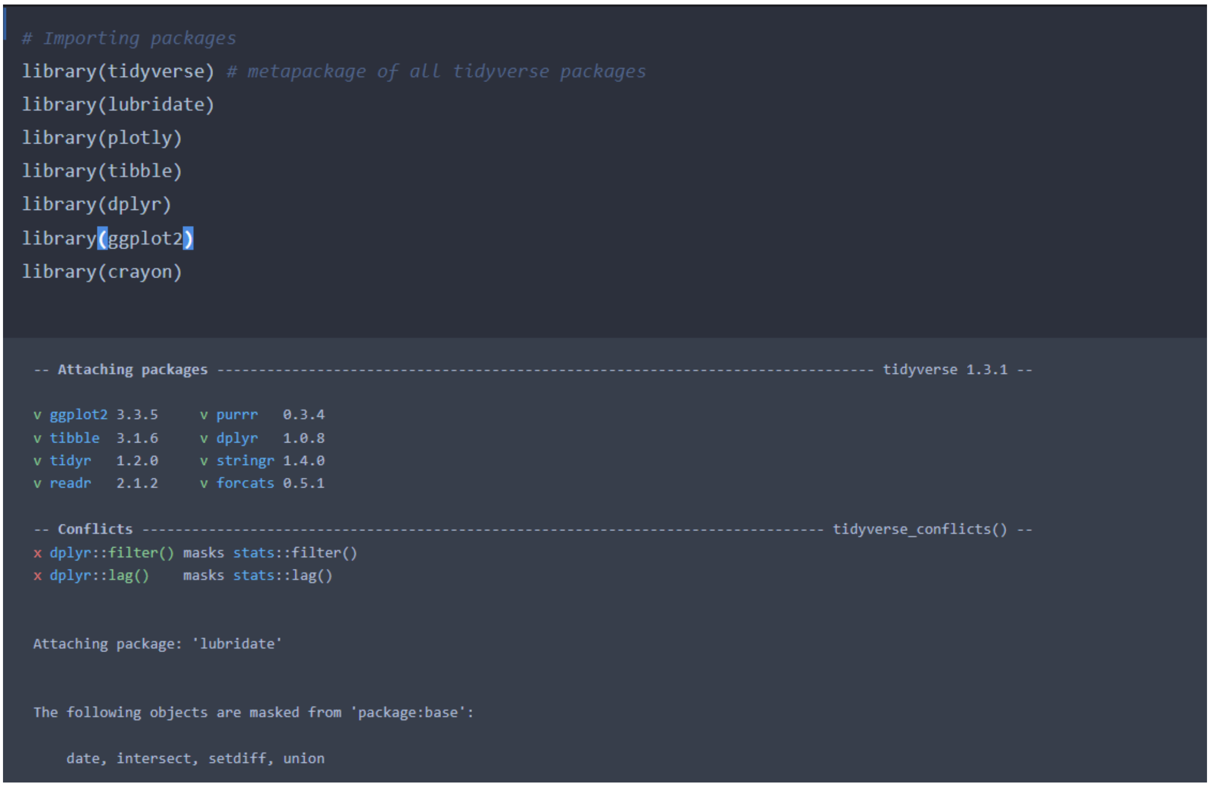
**DATASET:** Netflix Movies and TV Shows

**KAGGLE LINK:** https://www.kaggle.com/shivamb/netflix-shows/

There are 8807 unique rows in the dataset chosen. Following are the 12 attributes in this dataset-

| Sr. No. | Attribute Name | Attribute Description | Datatype |
|---|---|---|---|
| 1 | Show Id | Unique ID for every Movie / Tv Show | String |
| 2 | Type | Identifier - A Movie or TV Show | String |
| 3 | Title | Title of the Movie / Tv Show | String |
| 4 | Director | Director of the Movie | String |
| 5 | Cast | Actors involved in the movie / show | String |
| 6 | Country | Country where the movie / show was produced | String |
| 7 | Date Added | Date it was added on Netflix | Date |
| 8 | Release Year | Actual Release year of the move / show | Integer |
| 9 | Rating | TV Rating of the movie / show | String |
| 10 | Duration | Total Duration - in minutes or number of seasons | String |
| 11 | Listed In | Genre of Movie / TV Show | String |
| 12 | Description | The summary description | String |

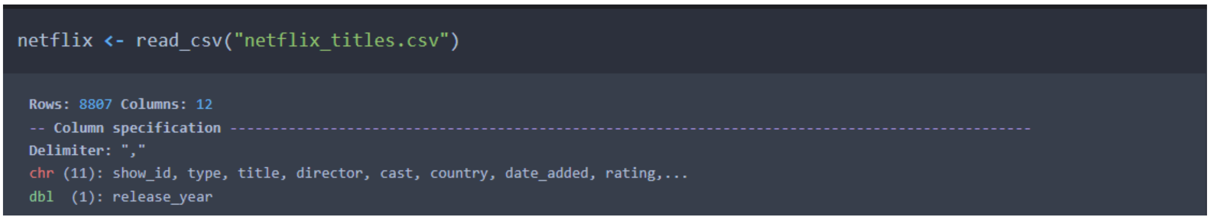
## Importing Required Packages:



```
# Importing packages
library(tidyverse) # metapackage of all tidyverse packages
library(lubridate)
library(plotly)
library(tibble)
library(dplyr)
library(ggplot2)
library(crayon)
```

```
-- Attaching packages --------------------------------------------------------------------------- tidyverse 1.3.1 --

v ggplot2 3.3.5     v purrr   0.3.4
v tibble  3.1.6     v dplyr   1.0.8
v tidyr   1.2.0     v stringr 1.4.0
v readr   2.1.2     v forcats 0.5.1

-- Conflicts ------------------------------------------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()


Attaching package: 'lubridate'


The following objects are masked from 'package:base':

    date, intersect, setdiff, union
```

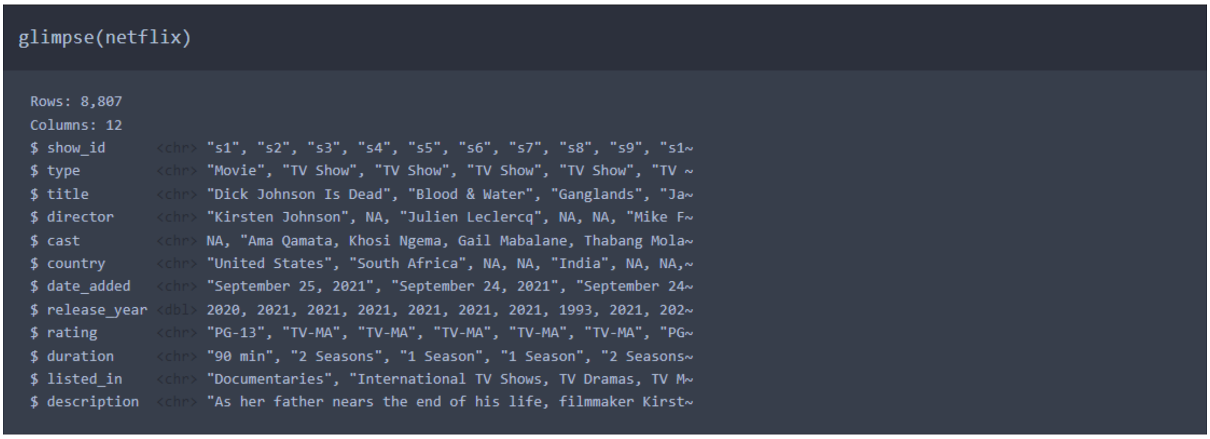
## Reading the Dataset:



```
netflix <- read_csv("netflix_titles.csv")
```

```
Rows: 8807 Columns: 12
-- Column specification ------------------------------------------------------------------------------------------
Delimiter: ","
chr (11): show_id, type, title, director, cast, country, date_added, rating,...
dbl  (1): release_year
```

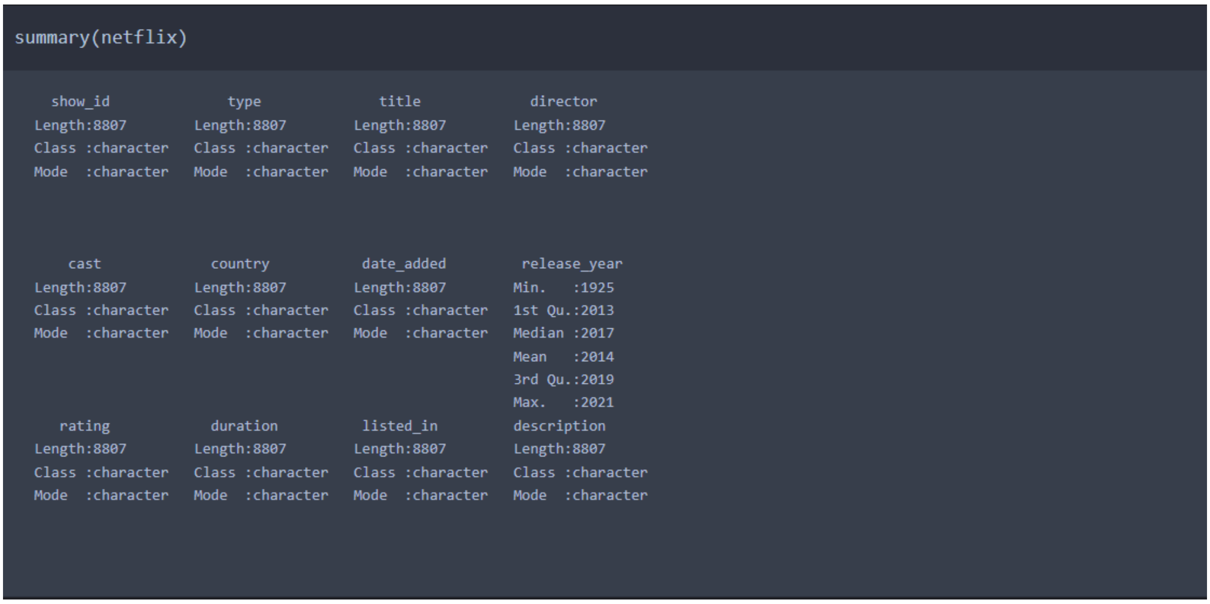
## Glimpse() gives details of all the columns:



```
glimpse(netflix)
```

```
Rows: 8,807
Columns: 12
$ show_id      <chr> "s1", "s2", "s3", "s4", "s5", "s6", "s7", "s8", "s9", "s1~
$ type         <chr> "Movie", "TV Show", "TV Show", "TV Show", "TV Show", "TV ~
$ title        <chr> "Dick Johnson Is Dead", "Blood & Water", "Ganglands", "Ja~
$ director     <chr> "Kirsten Johnson", NA, "Julien Leclercq", NA, NA, "Mike F~
$ cast         <chr> NA, "Ama Qamata, Khosi Ngema, Gail Mabalane, Thabang Mola~
$ country      <chr> "United States", "South Africa", NA, NA, "India", NA, NA,~
$ date_added   <chr> "September 25, 2021", "September 24, 2021", "September 24~
$ release_year <dbl> 2020, 2021, 2021, 2021, 2021, 2021, 2021, 1993, 2021, 202~
$ rating       <chr> "PG-13", "TV-MA", "TV-MA", "TV-MA", "TV-MA", "TV-MA", "PG~
$ duration     <chr> "90 min", "2 Seasons", "1 Season", "1 Season", "2 Seasons~
$ listed_in    <chr> "Documentaries", "International TV Shows, TV Dramas, TV M~
$ description  <chr> "As her father nears the end of his life, filmmaker Kirst~
```

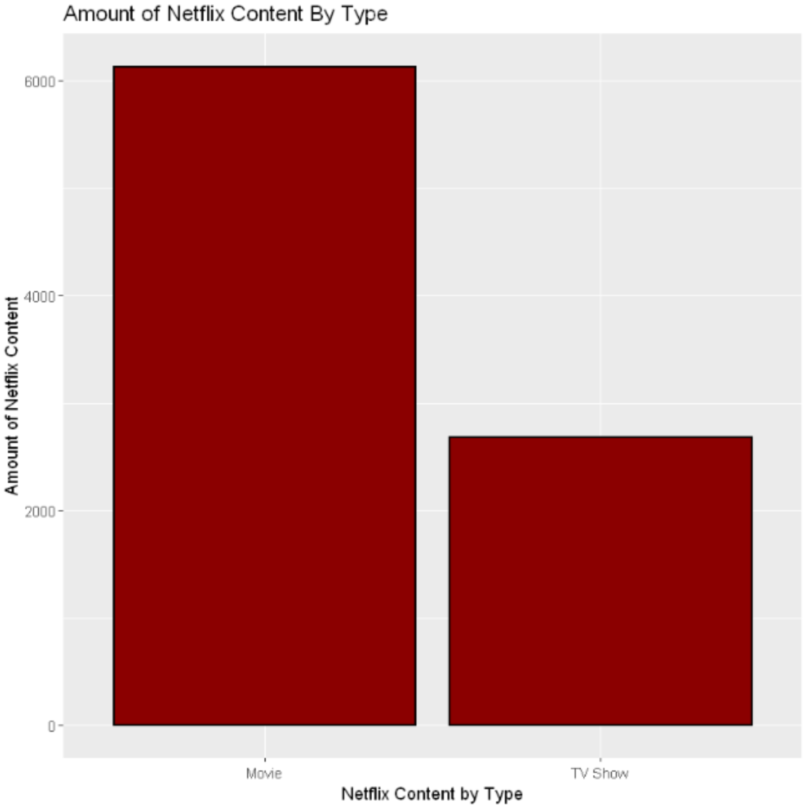
## Summary of all attributes of the Dataset:



```
summary(netflix)
```

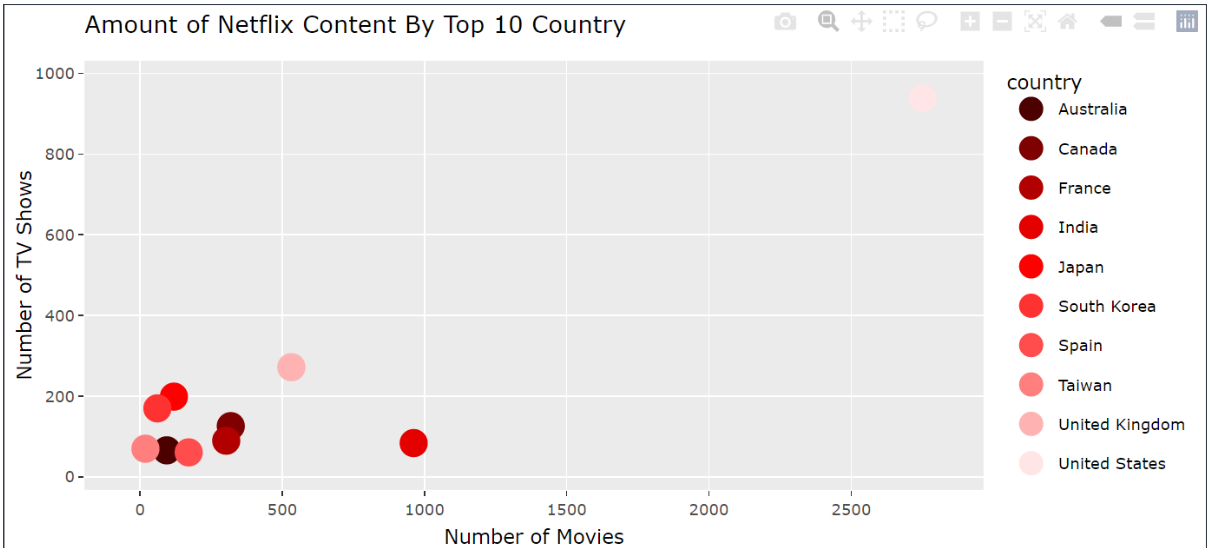
```
   show_id              type               title              director        
 Length:8807        Length:8807        Length:8807        Length:8807       
 Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character  



     cast              country            date_added         release_year 
 Length:8807        Length:8807        Length:8807        Min.   :1925  
 Class :character   Class :character   Class :character   1st Qu.:2013  
 Mode  :character   Mode  :character   Mode  :character   Median :2017  
                                                          Mean   :2014  
                                                          3rd Qu.:2019  
                                                          Max.   :2021  
    rating            duration           listed_in          description       
 Length:8807        Length:8807        Length:8807        Length:8807       
 Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character  
```

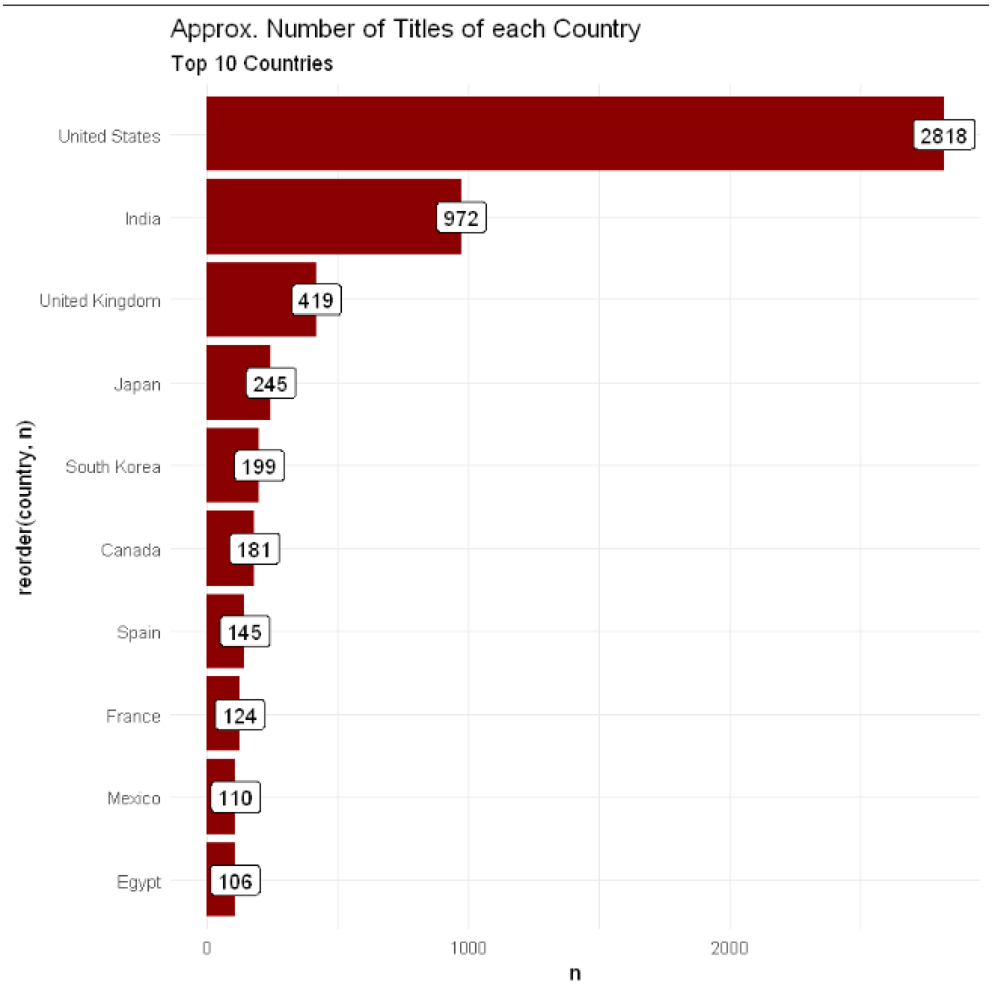
## PLOT 1: Amount of Netflix Content by Type



In the above chart, a bar graph is plotted between count of  attribute (Movie & TV Show) and the Content Type.

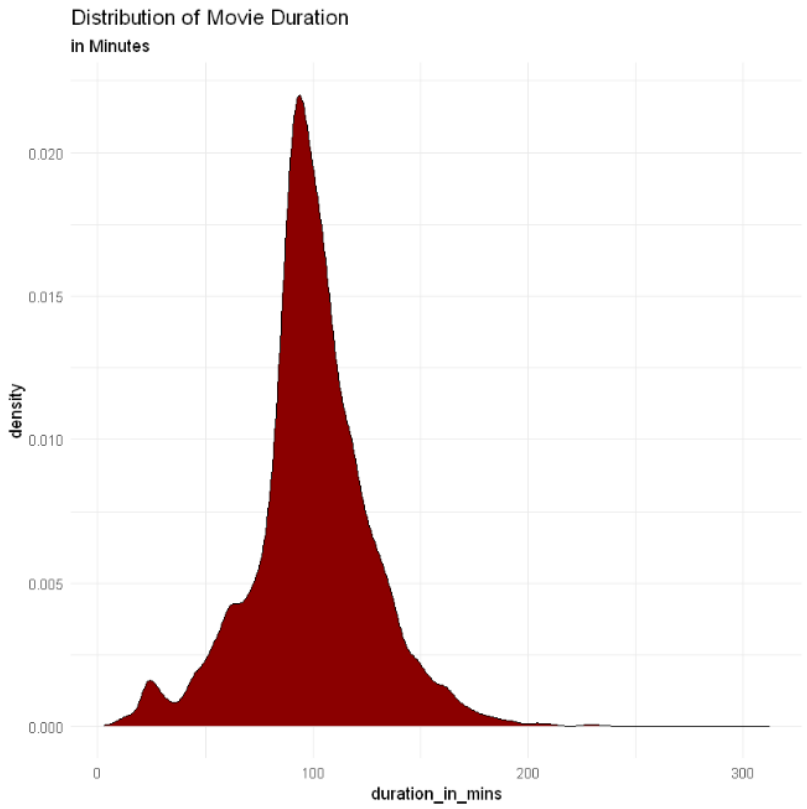
## PLOT 2: Amount of Netflix Content By Top 10 Country





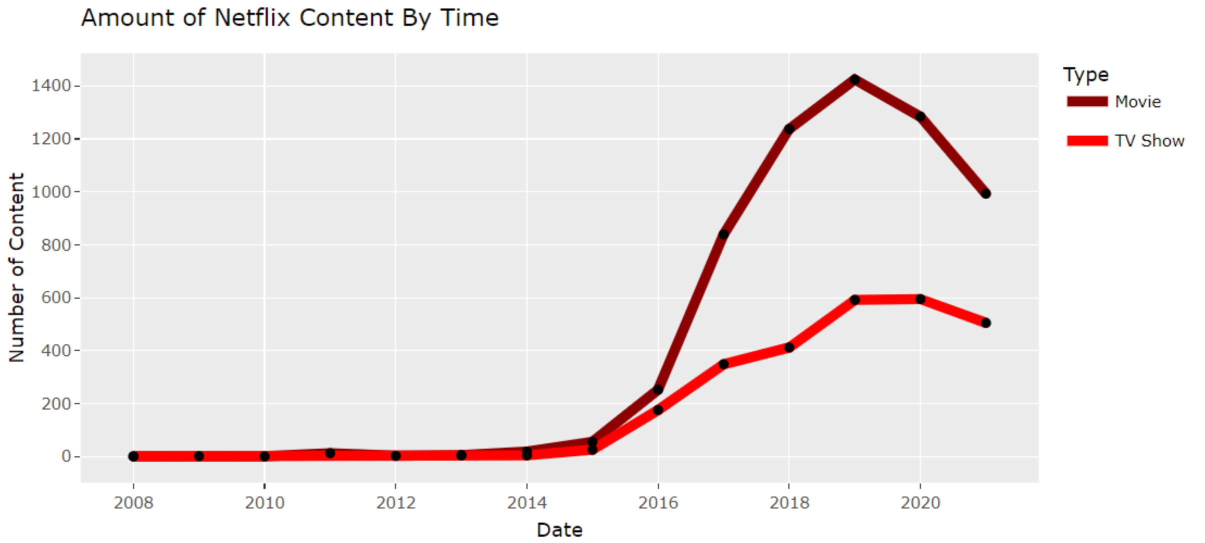
In the above 2 charts, a graph is plotted between top 10 countries and number of titles. We see that the United States is a clear leader in the amount of content on Netflix.

## PLOT 3: Mean Duration of Movies



In the above chart, a density plot of duration of Content Type = ‘Movie’ is plotted. It helps to identify that maximum movies have duration of around 100 minutes.

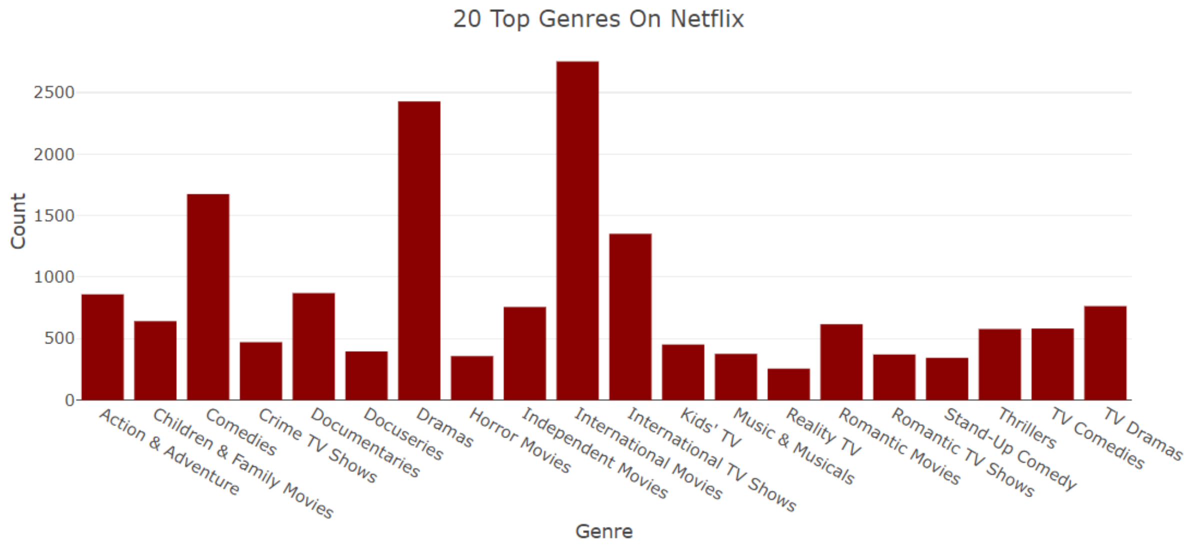
## PLOT 4: Movies & TV Shows Added Over The Years



In the above chart, a line graph is plotted between count of Type attribute (Movie & TV Show) and Date Added. It helps to visualize the number of TV Shows and Movies added to

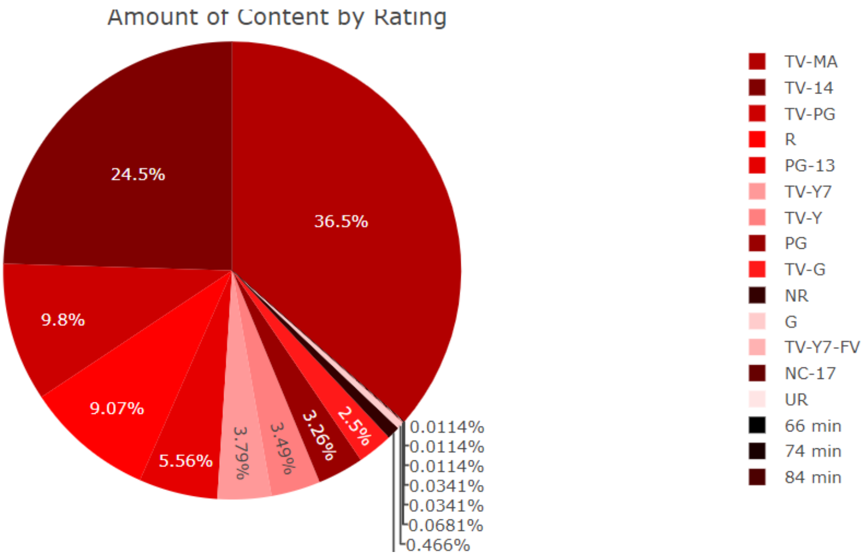
Netflix over the years. Maximum TV Shows and Movies were added in the year 2019. Also, there is an increase in the amount of content added on Netflix from 2016 onwards. This is because of the launch of the Netflix app in 2016 and expansion in Asia.

## PLOT 5: Top 20 Genres (By Number Of Titles)



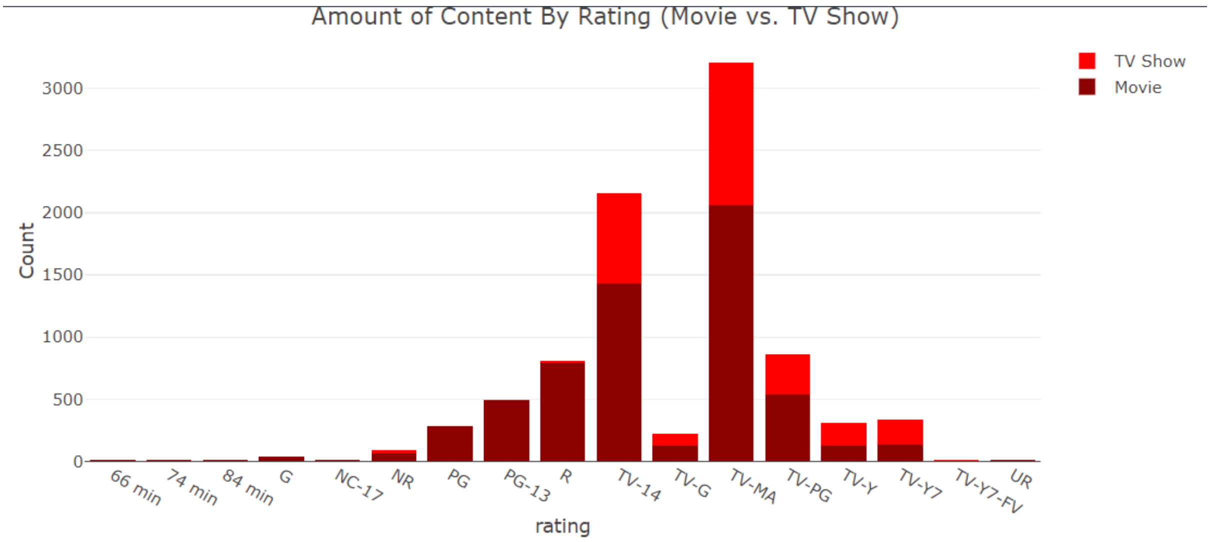
The above chart was plotted between the Genres i.e. Listed In Attribute and number of titles. It helps us to identify the genre having maximum titles. Here, international movies has the maximum number of titles.

## PLOT 6: Amount Of Content By Rating



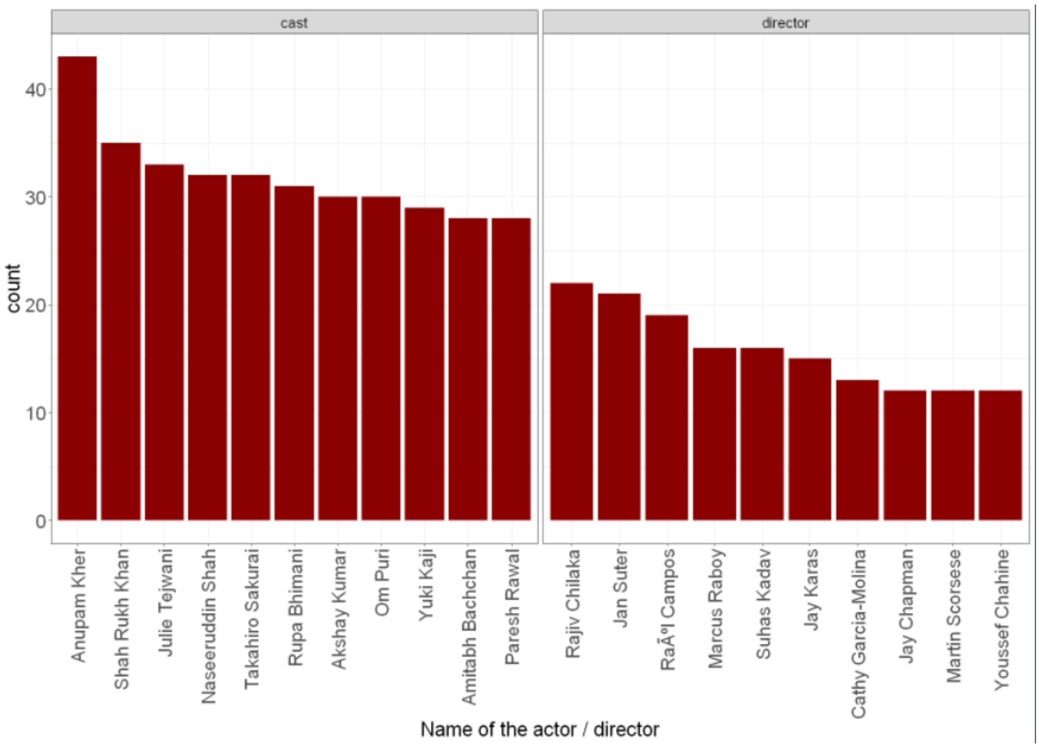
The above chart was plotted between the Rating and Content. It helps us to identify which content has maximum Rating. In this dataset, the percentage of TV-MA is maximum.

## PLOT 7: Amount Of Content By Rating Movies V/S TV Shows



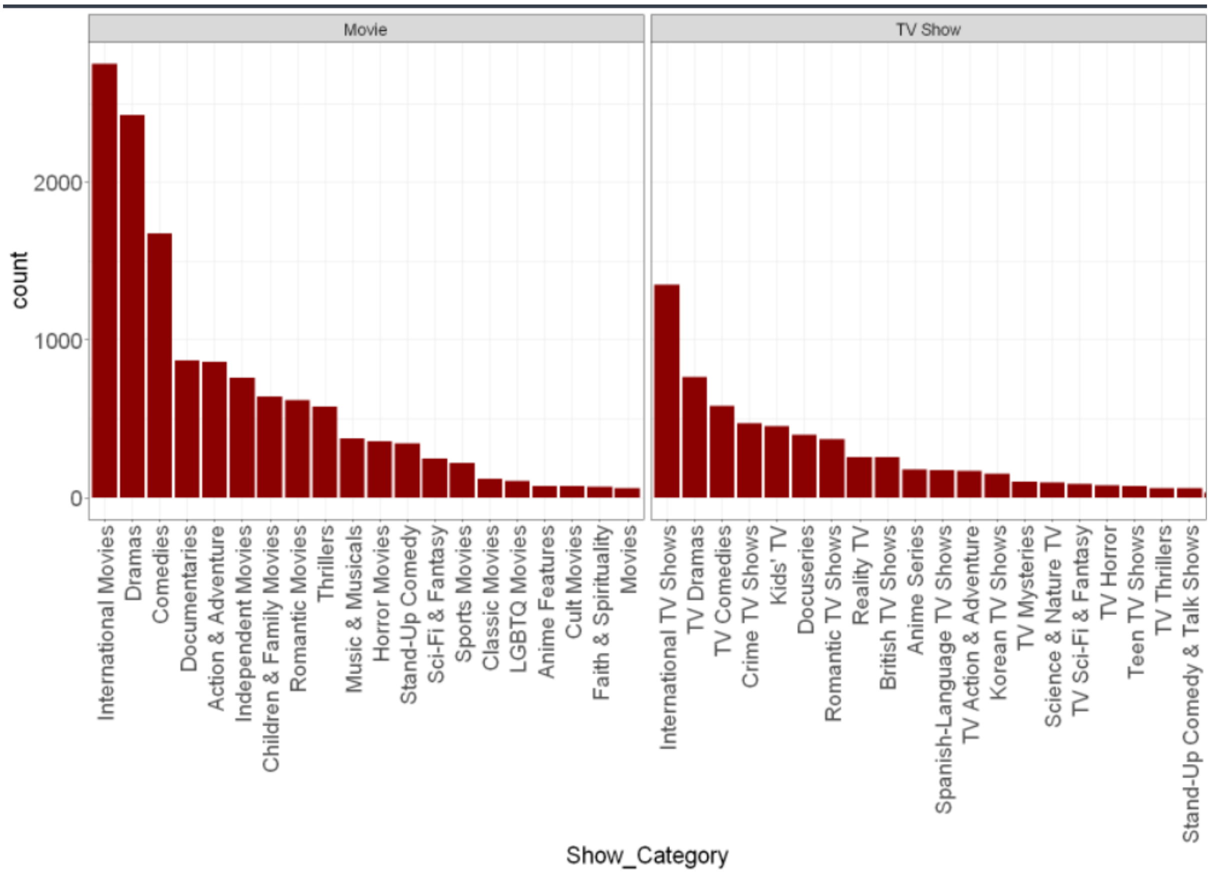
The above chart was plotted between the Rating and Content along with Content Type. It helps us to identify which content has maximum Rating.

## PLOT 8: Top 10 Directors & Actors By Titles



The above chart was plotted between Directors v/s title count and Cast Member v/s title count. It helps us to identify the most prolific directors, and actors who are associated with most movies/shows. It helps us to identify the director/actor having maximum titles. Here, the director with maximum Titles is Rajiv Chilaka and actor is Anupam Kher.

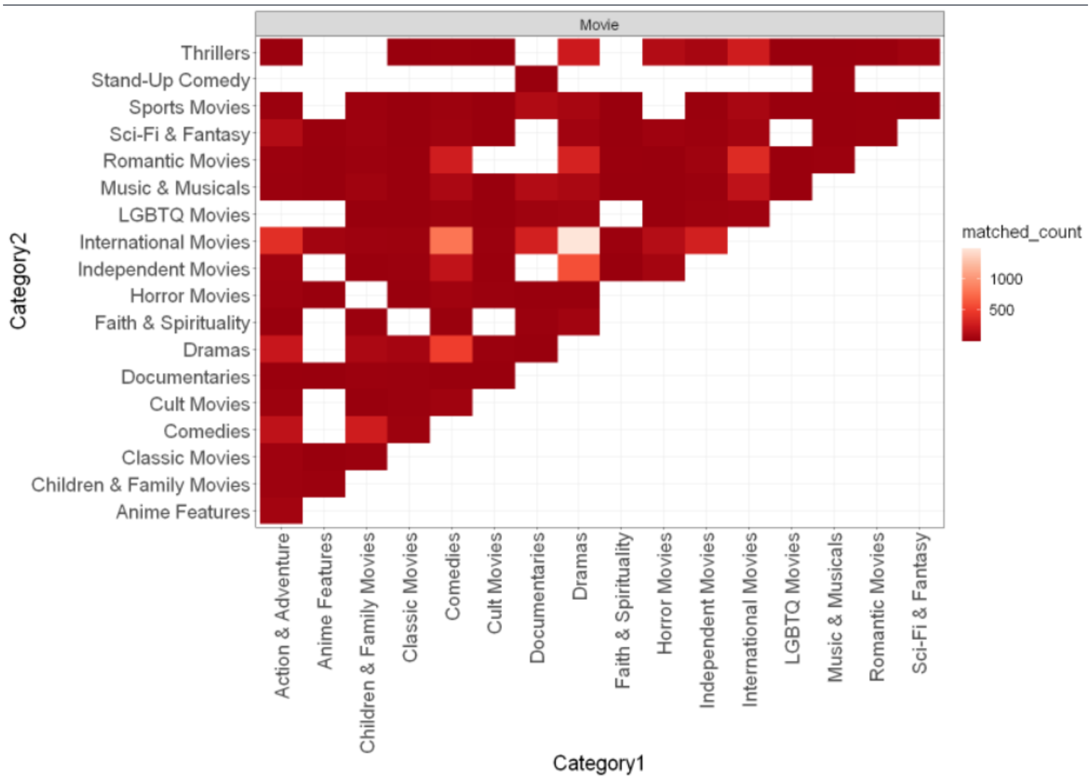
## PLOT 9: Top TV Show Categories & Top Movie Categories

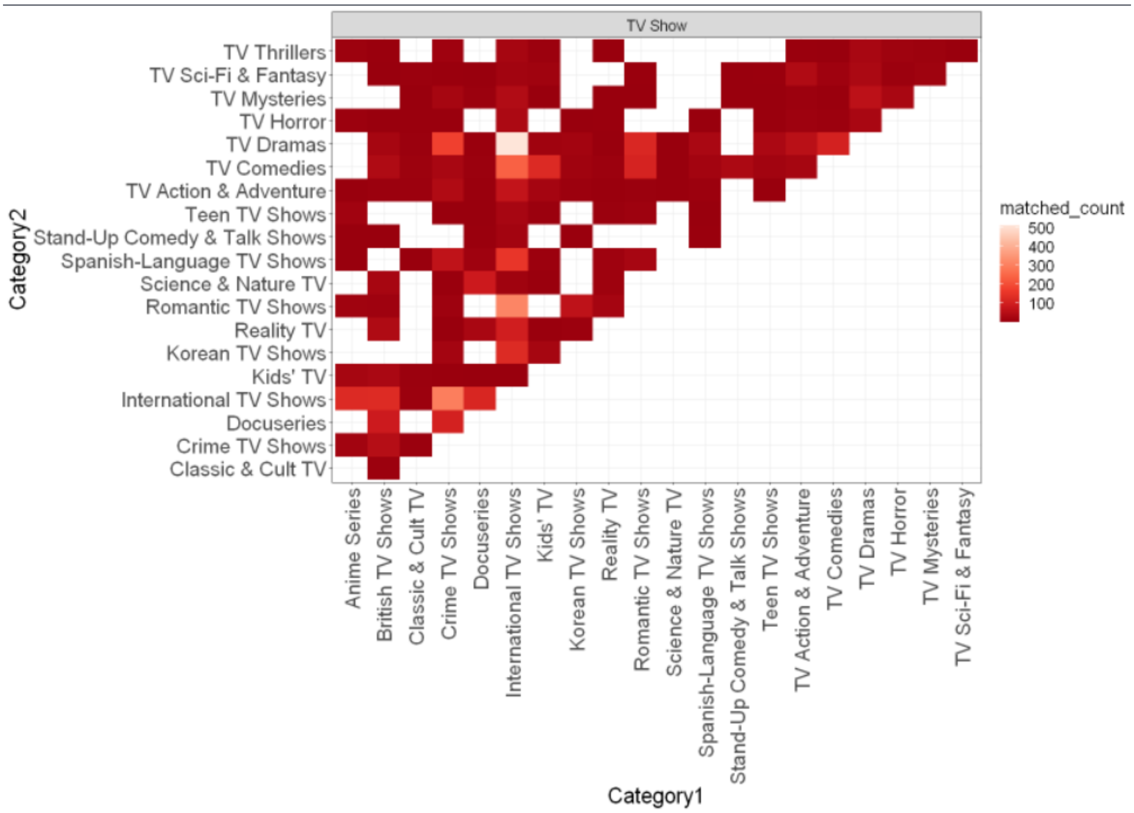


The above chart consists of 2 subplots. A bar graph was plotted between Genres (Listed In) and Count of TV Shows (From Type: TV Shows). It helps us to identify the genres with highest number of TV Shows. The 2nd above bar graph was plotted between Genres (Listed In) and Count of Movies (From Type: Movies). It helps us to identify the genres with highest number of Movies. It can be seen that International Movies / TV Shows are showing up as the dominant category in both Movies and TV shows, followed by Drama and Comedies

## PLOT 10: Correlation Between Categories (Genre) In Movies & TV Shows

Many Movies/ TV Shows are listed in multiple categories, so the below 2 correlation plots (heat map) helps to identify what categories more correlated with each other. From the below 2 Correlation Plots (1st one is genres correlation plot for content type Movies and the 2nd one is for Type TV Shows), it can be observed that many of the international movies are listed "Dramas", also there are some interesting missing overlaps in categories, such as "Faith and Spirituality" doesn't seem to overlap with "LGBTQ" movies while there is a significant overlap between "Drama" and "Comedy" as expected.





## CONCLUSION:

In this experiment, I learnt to use R for Exploratory Data Analysis and plotted multiple colour coordinated plots to draw insights from the Netflix Movies and TV Shows Dataset.