Netflix

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 4.1.2
```

```
library(stringr)
library(ggplot2)
library(modeest)
```

```
## Warning: package 'modeest' was built under R version 4.1.2
```

```
data<- read.csv("C:/Users/kkonar2/Downloads/archive/netflix_titles.csv",na.strings = c("","NA"))
str(data)
```

```
## 'data.frame':    8807 obs. of  12 variables:
##  $ show_id     : chr  "s1" "s2" "s3" "s4" ...
##  $ type        : chr  "Movie" "TV Show" "TV Show" "TV Show" ...
##  $ title       : chr  "Dick Johnson Is Dead" "Blood & Water" "Ganglands" "Jailbirds New Orleans" ...
##  $ director    : chr  "Kirsten Johnson" NA "Julien Leclercq" NA ...
##  $ cast        : chr  NA "Ama Qamata, Khosi Ngema, Gail Mabalane, Thabang Molaba, Dillon Windvogel, I
##  $ country     : chr  "United States" "South Africa" NA NA ...
##  $ date_added  : chr  "September 25, 2021" "September 24, 2021" "September 24, 2021" "September 24, 1
##  $ release_year: int  2020 2021 2021 2021 2021 2021 2021 1993 2021 2021 ...
##  $ rating      : chr  "PG-13" "TV-MA" "TV-MA" "TV-MA" ...
##  $ duration    : chr  "90 min" "2 Seasons" "1 Season" "1 Season" ...
##  $ listed_in   : chr  "Documentaries" "International TV Shows, TV Dramas, TV Mysteries" "Crime TV Sho
##  $ description : chr  "As her father nears the end of his life, filmmaker Kirsten Johnson stages his
```

```r
#Checking the number of NA values
colSums(is.na(data))
```

```
##      show_id         type        title     director         cast      country
##            0            0            0         2634          825          831
##   date_added release_year       rating     duration    listed_in  description
##           10            0            4            3            0            0
```

```r
#Removing NA values
df<-na.omit(data)
colSums(is.na(df))
```

```
##      show_id         type        title     director         cast      country
##            0            0            0            0            0            0
##   date_added release_year       rating     duration    listed_in  description
##            0            0            0            0            0            0
```

```r
#Creating a seperate column called month from the date_added column and performing data cleaning
df1<- df %>%
  separate(date_added,c("Month"))
```

```
## Warning: Expected 1 pieces. Additional pieces discarded in 5332 rows [1, 2, 3,
## 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].
```

```r
df1<-df1[!is.na(df1$Month),]
df1$Month[df1$Month==""]<-mfv(df1$Month)
df1$Month<-factor(df1$Month,levels = month.name)
table(df1$type,df1$Month)
```

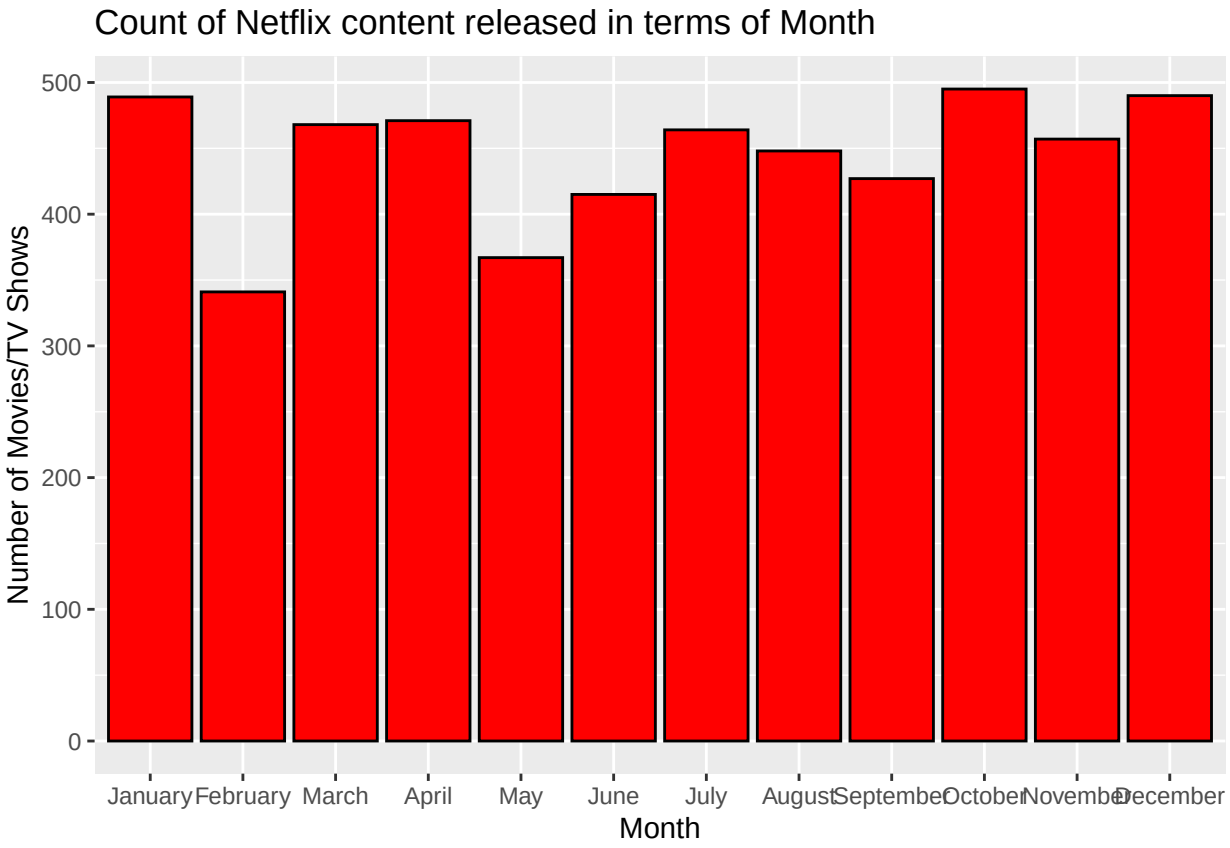
```
##
##           January February March April May June July August September October
##   Movie       478      327   454   460 357  403  451    434       416     480
##   TV Show      11       14    14    11  10   12   13     14        11      15
##
##           November December
##   Movie        452      473
##   TV Show        5       17
```

```r
#Seperating the column country to distinct values and selecting the top countries streaming the most Ne
df2<-separate_rows(df1,country,show_id , convert = TRUE, sep = ', ')
country_count<-sort(table(df2$country),decreasing=TRUE)[1:10]
country_count<-data.frame(country_count)
print(country_count)
```

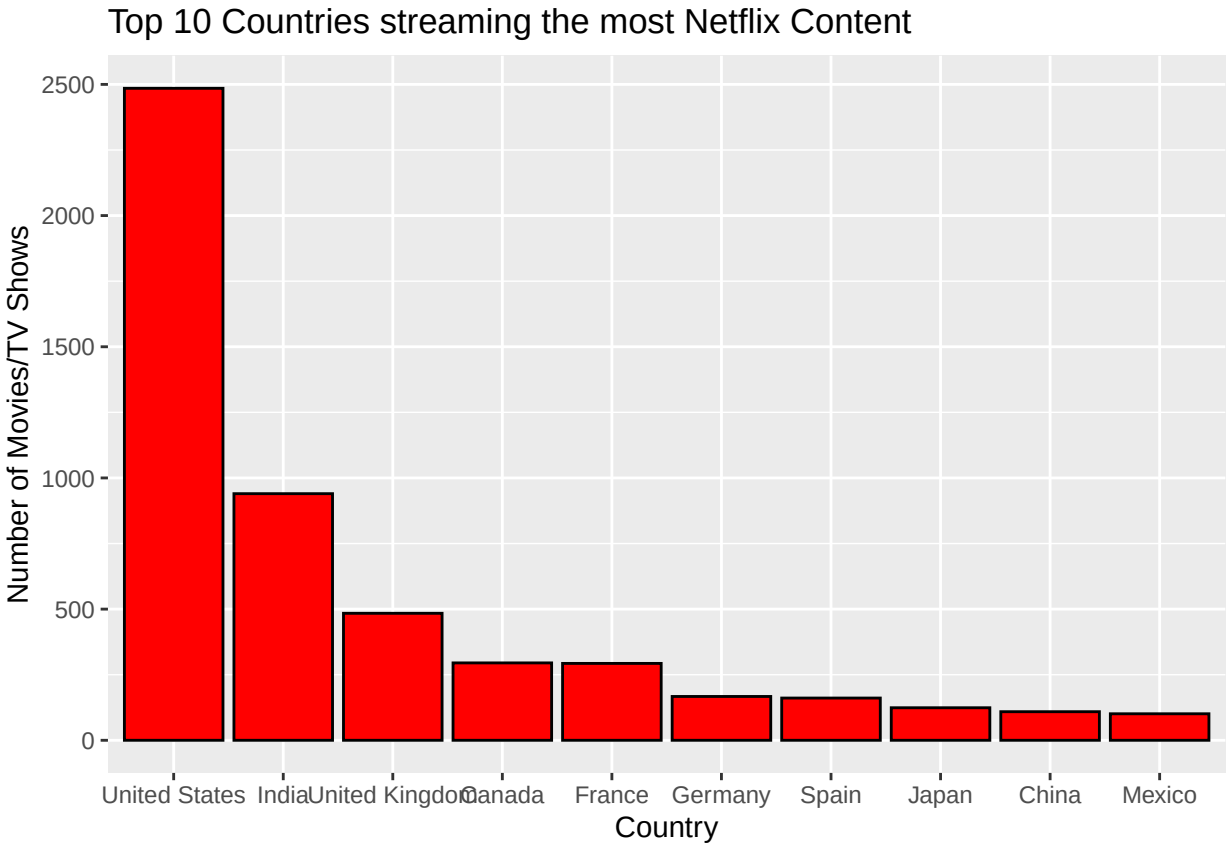
```
##             Var1 Freq
## 1  United States 2485
## 2          India  940
## 3 United Kingdom  484
## 4         Canada  295
## 5         France  293
```

```
## 6         Germany  167
## 7           Spain  161
## 8           Japan  124
## 9           China  109
## 10         Mexico  101
```

```
#Visualizing the association between netflix content type and the release dates
g= ggplot(data=df1, aes(x = Month,fill=type))
g = g + geom_bar(fill = "Red", color = "Black")
g=g+ylab("Number of Movies/TV Shows")+ggtitle("Count of Netflix content released in terms of Month")
g
```




```
#Visualization depicting the countries with the most netflix content
g= ggplot(data=country_count, aes(x = Var1, y=Freq))
g = g + geom_bar(stat = 'Identity',fill = "Red", color = "Black")
g = g + xlab("Country")+ylab("Number of Movies/TV Shows")+ggtitle("Top 10 Countries streaming the most
g
```


Top 10 Countries streaming the most Netflix Content
Number of Movies/TV Shows
2500
2000
1500
1000
500
0
United States
India
United Kingdom
Canada
France
Germany
Spain
Japan
China
Mexico
Country