```
# uploading the libraries
library(dplyr)
library(tidyverse)
library(ggplot2)
library(lubridate)
library(tibble)
library(plotly)
library(gridExtra)
library(scales)
library(crayon)
webshot::install_phantomjs()
```

Introduction

Netflix has grown to be one of the top source of entertainment. The platform has recorded increase of 23 percent in paid memberships during the final quarters of 2019. The projects goal's was to see growth of the Netflix streaming library since 2008 up to early of 2021.

Data

The data set was given by kaggle and can be retrieved from: https://www.kaggle.com/shivamb/netflix-shows

```
# loads that data set
data<- read.csv("Eideted Netlfix sheet.csv")
```

Summary of the Data Set

The overall data set has 11 variables/columns and 7787 objects/rows.

```
# shows first 5 rows of the data set
head(data)
```

```
##   show_id    type title          director
## 1      s1 TV Show    3%
## 2      s2   Movie  7:19 Jorge Michel Grau
## 3      s3   Movie 23:59      Gilbert Chan
## 4      s4   Movie     9       Shane Acker
## 5      s5   Movie    21    Robert Luketic
## 6      s6 TV Show    46       Serdar Akar
##
## 1 JoÃ£o Miguel, Bianca Comparato, Michel Gomes, Rodolfo Valente, Vaneza Oliveira, Rafael Lozano, Vivi
## 2                                                                                    DemiÃ¡n Bichir, H
## 3                                                                          Tedd Chan, Stella Chung, Henley Hi
## 4                                Elijah Wood, John C. Reilly, Jennifer Connelly, Christopher Plummer, C
## 5              Jim Sturgess, Kevin Spacey, Kate Bosworth, Aaron Yoo, Liza Lapira, Jacob Pitts, Laurer
## 6                   Erdal BeÅŸikÃ§ioÄŸlu, Yasemin Allen, Melis Birkan, SaygÄ±n Soysal, Berkan Åžal, H
##         country        date_added release_year rating  duration
## 1        Brazil   August 14, 2020         2020  TV-MA 4 Seasons
## 2        Mexico December 23, 2016         2016  TV-MA    93 min
## 3     Singapore December 20, 2018         2011      R    78 min
## 4 United States November 16, 2017         2009  PG-13    80 min
```

```
## 5 United States   January 1, 2020         2008  PG-13   123 min
## 6        Turkey      July 1, 2017         2016  TV-MA  1 Season
##                                                  listed_in
## 1   International TV Shows, TV Dramas, TV Sci-Fi & Fantasy
## 2                             Dramas, International Movies
## 3                          Horror Movies, International Movies
## 4 Action & Adventure, Independent Movies, Sci-Fi & Fantasy
## 5                                                   Dramas
## 6          International TV Shows, TV Dramas, TV Mysteries
```

```
# outputs the structure of thee data set
str(data)
```

```
## 'data.frame':    7787 obs. of  11 variables:
##  $ show_id     : chr  "s1" "s2" "s3" "s4" ...
##  $ type        : chr  "TV Show" "Movie" "Movie" "Movie" ...
##  $ title       : chr  "3%" "7:19" "23:59" "9" ...
##  $ director    : chr  "" "Jorge Michel Grau" "Gilbert Chan" "Shane Acker" ...
##  $ cast        : chr  "JoÃ£o Miguel, Bianca Comparato, Michel Gomes, Rodolfo Valente, Vaneza Oliveira
##  $ country     : chr  "Brazil" "Mexico" "Singapore" "United States" ...
##  $ date_added  : chr  "August 14, 2020" "December 23, 2016" "December 20, 2018" "November 16, 2017"
##  $ release_year: int  2020 2016 2011 2009 2008 2016 2019 1997 2019 2008 ...
##  $ rating      : chr  "TV-MA" "TV-MA" "R" "PG-13" ...
##  $ duration    : chr  "4 Seasons" "93 min" "78 min" "80 min" ...
##  $ listed_in   : chr  "International TV Shows, TV Dramas, TV Sci-Fi & Fantasy" "Dramas, International
```

```
# outputs the summary of the data set
summary(data)
```

```
##    show_id              type              title             director
##  Length:7787        Length:7787        Length:7787        Length:7787
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##      cast             country           date_added         release_year
##  Length:7787        Length:7787        Length:7787        Min.   :1925
##  Class :character   Class :character   Class :character   1st Qu.:2013
##  Mode  :character   Mode  :character   Mode  :character   Median :2017
##                                                           Mean   :2014
##                                                           3rd Qu.:2018
##                                                           Max.   :2021
##     rating            duration          listed_in
##  Length:7787        Length:7787        Length:7787
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
```

Data Clean-up

Overall Data Clean-up for the Data set: 1) Changed the format of the date_added column 2) Turned all the blank values into NA's 3) Used a fuction to get the mode of country, rating and date added. Used the mode values to replace the values that were NA's 4)Dropped columns director and cast 5)Categorized the rating to an age group

```
# filled all blank values with NA
data_new2 <- data
data_new2[data_new2 == "" | data_new2 == " "] <- NA
head(data_new2)
```

```
##   show_id    type title          director
## 1      s1 TV Show    3%              <NA>
## 2      s2   Movie  7:19 Jorge Michel Grau
## 3      s3   Movie 23:59      Gilbert Chan
## 4      s4   Movie     9       Shane Acker
## 5      s5   Movie    21    Robert Luketic
## 6      s6 TV Show    46       Serdar Akar
##
## 1 JoÃ£o Miguel, Bianca Comparato, Michel Gomes, Rodolfo Valente, Vaneza Oliveira, Rafael Lozano, Vivi
## 2                                                                                 DemiÃ¡n Bichir, H
## 3                                                                  Tedd Chan, Stella Chung, Henley Hi
## 4                                Elijah Wood, John C. Reilly, Jennifer Connelly, Christopher Plummer, C
## 5              Jim Sturgess, Kevin Spacey, Kate Bosworth, Aaron Yoo, Liza Lapira, Jacob Pitts, Laurer
## 6                   Erdal BeÅŸikÃ§ioÄŸlu, Yasemin Allen, Melis Birkan, SaygÄ±n Soysal, Berkan Åžal, H
##         country        date_added release_year rating  duration
## 1        Brazil   August 14, 2020         2020  TV-MA 4 Seasons
## 2        Mexico December 23, 2016         2016  TV-MA    93 min
## 3     Singapore December 20, 2018         2011      R    78 min
## 4 United States November 16, 2017         2009  PG-13    80 min
## 5 United States   January 1, 2020         2008  PG-13   123 min
## 6        Turkey      July 1, 2017         2016  TV-MA  1 Season
##                                                listed_in
## 1   International TV Shows, TV Dramas, TV Sci-Fi & Fantasy
## 2                          Dramas, International Movies
## 3                       Horror Movies, International Movies
## 4 Action & Adventure, Independent Movies, Sci-Fi & Fantasy
## 5                                                 Dramas
## 6          International TV Shows, TV Dramas, TV Mysteries
```

```
# changes format of the date_added
data_new2$date_added <- mdy(data_new2$date_added)
```

```
# shows the number of NA's values in each columns
colSums(is.na(data_new2))
```

```
##      show_id         type        title     director         cast      country
##            0            0            0         2389          718          507
##   date_added release_year       rating     duration    listed_in
##           10            0            7            0            0
```

```r
# gets mode for rating. data_added, and country and replaces NA's with the mode
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v,uniqv)))]
}
rating_mode<-getmode(data_new2$rating)
data_added_mode <- getmode(data_new2$date_added)
country_mode <-getmode(data_new2$country)

data_new2$rating[which(is.na(data_new2$rating))] <- rating_mode

data_new2$date_added[which(is.na(data_new2$date_added))]<- data_added_mode

data_new2$country[which(is.na(data_new2$country))] <- country_mode
```

```r
# drops rows director and cast columns
data_new2 <- data_new2 %>% subset(select=-c(director,cast))
```

```r
# shows total of NA's in each columns
colSums(is.na(data_new2))
```

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

```r
# categorizes the rating to an age group
data_new2<-data_new2 %>% mutate(target_age=case_when(rating == 'TV-PG' ~ 'Older Kids', rating=='TV-MA' 
```

Percentage of TV and Movie in Netflix

Here the next following code chunks are to make the visualization for "Percentage of TV and Movie in Netflix". We start with creating a two new data frame one for Movie and TV.

```r
# creates new data frame for Movie and TV
movie_df<- data_new2 %>% filter(type=='Movie')
head(movie_df)
```

```
##   show_id  type title       country date_added release_year rating duration
## 1      s2 Movie  7:19        Mexico 2016-12-23         2016  TV-MA   93 min
## 2      s3 Movie 23:59     Singapore 2018-12-20         2011      R   78 min
## 3      s4 Movie     9 United States 2017-11-16         2009  PG-13   80 min
## 4      s5 Movie    21 United States 2020-01-01         2008  PG-13  123 min
## 5      s7 Movie   122         Egypt 2020-06-01         2019  TV-MA   95 min
## 6      s8 Movie   187 United States 2019-11-01         1997      R  119 min
##                                              listed_in target_age
## 1                           Dramas, International Movies     Adults
## 2                     Horror Movies, International Movies     Adults
## 3 Action & Adventure, Independent Movies, Sci-Fi & Fantasy      Teens
## 4                                                Dramas      Teens
## 5                     Horror Movies, International Movies     Adults
## 6                                                Dramas     Adults
```

```
tv_df<- data_new2 %>% filter(type=='TV Show')
head(tv_df)
```

```
##   show_id    type                                    title               country
## 1      s1 TV Show                                       3%                Brazil
## 2      s6 TV Show                                       46                Turkey
## 3     s12 TV Show                                     1983 Poland, United States
## 4     s13 TV Show                                     1994                Mexico
## 5     s17 TV Show                                    9-Feb         United States
## 6     s25 TV Show â\200<SAINT SEIYA: Knights of the Zodiac                 Japan
##   date_added release_year rating  duration
## 1 2020-08-14         2020  TV-MA 4 Seasons
## 2 2017-07-01         2016  TV-MA  1 Season
## 3 2018-11-30         2018  TV-MA  1 Season
## 4 2019-05-17         2019  TV-MA  1 Season
## 5 2019-03-20         2018  TV-14  1 Season
## 6 2020-01-23         2020  TV-14 2 Seasons
##                                              listed_in target_age
## 1 International TV Shows, TV Dramas, TV Sci-Fi & Fantasy     Adults
## 2        International TV Shows, TV Dramas, TV Mysteries     Adults
## 3      Crime TV Shows, International TV Shows, TV Dramas     Adults
## 4     Crime TV Shows, Docuseries, International TV Shows     Adults
## 5                       International TV Shows, TV Dramas      Teens
## 6                    Anime Series, International TV Shows      Teens
```

Using the new table I then added the total count of Movie and TV. Along with the total percentage of Movie and TV

```
# gets the total percentage and total count of Movies and Tv's
type_df <- data_new2 %>% group_by(type) %>% summarize(counts=n(),percentage=n()/nrow(data_new2))
type_df
```

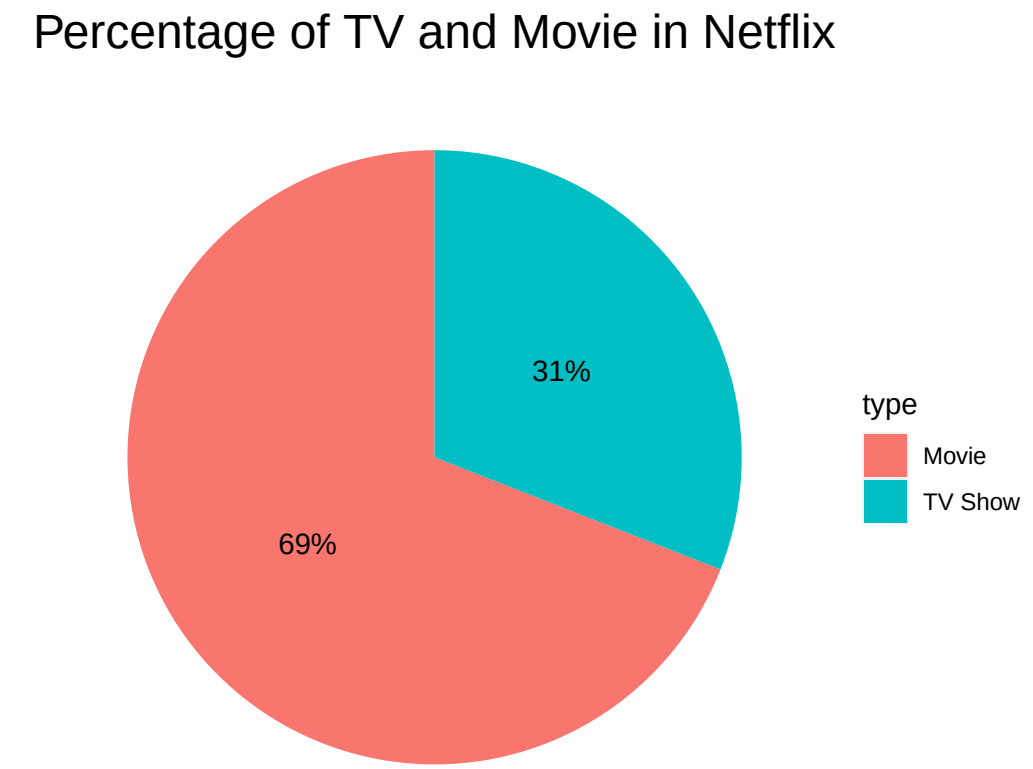
```
## # A tibble: 2 x 3
##   type    counts percentage
##   <chr>    <int>      <dbl>
## 1 Movie     5377      0.691
## 2 TV Show   2410      0.309
```

With type_df created I then added information to a pie chart called type_df_pie. Here we can see Movie is 69% of Netflix's streaming library while TV is 31%

```
# outputs a pie charts of the the total percentage and total count of Movies and Tv's
type_df_pie <- type_df %>% ggplot(aes(x="", y=percentage,fill=type))+
  geom_col(type="black")+
  coord_polar("y",start=0)+
  geom_text(aes(label=paste0(round(percentage*100),"%")),position = position_stack(vjust = 0.5))+ theme
```

```
## Warning: Ignoring unknown parameters: type
```

```
type_df_pie
```



Amount of Netflix Content By Top 10 Countries

This visualization goes over the number Netflix content for the Top 10 Countries. We see from the visualization that many of the other countries have around the same amount of content, while the United States is the outlier with the most amount of content.

```
k<-strsplit(data_new2$country,split = ",") # splits values country columns

netds_countries<- data.frame(type = rep(data_new2$type, sapply(k, length)), country = unlist(k))
head(netds_countries) # creates a data frame
```

```
##      type       country
## 1 TV Show        Brazil
## 2   Movie        Mexico
## 3   Movie     Singapore
## 4   Movie United States
## 5   Movie United States
## 6 TV Show        Turkey
```

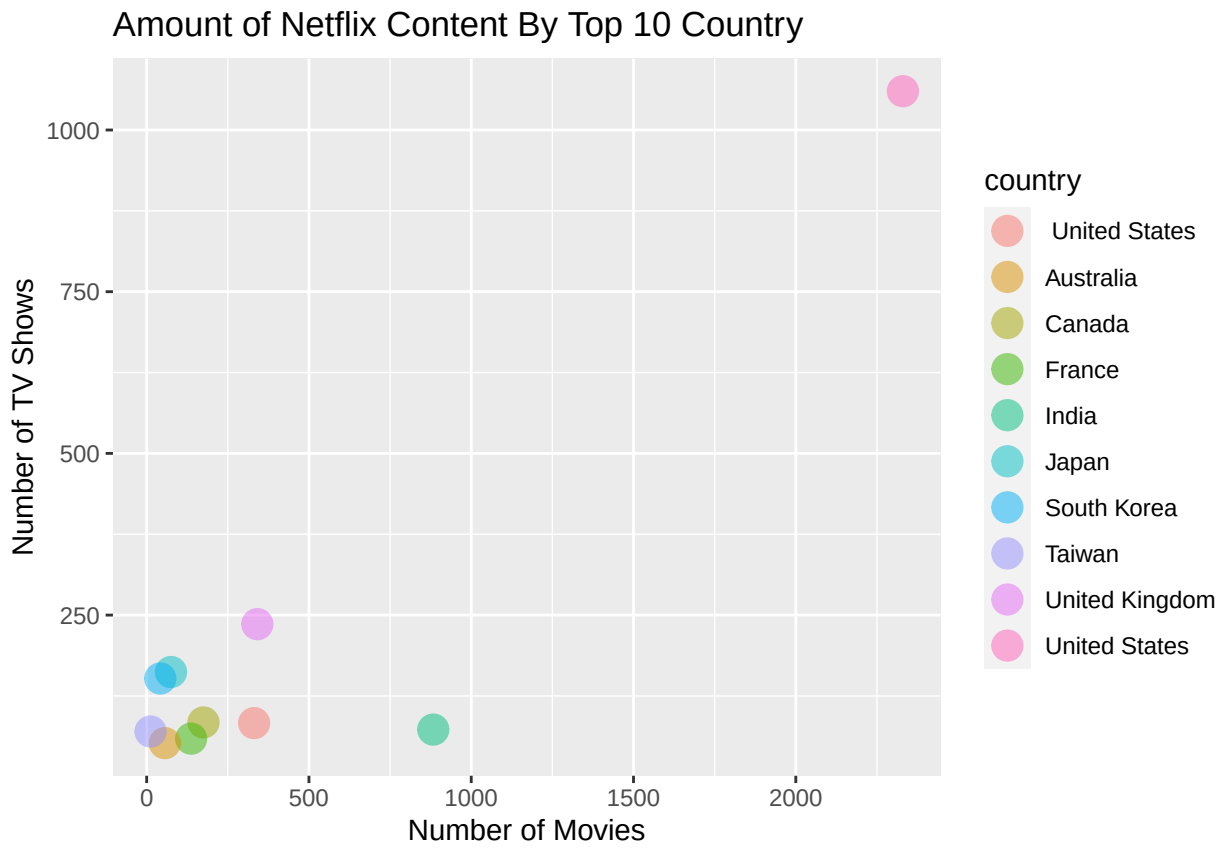
```
netds_countries$country<- as.character(netds_countries$country) # converts country as character

amount_by_country <- netds_countries %>% group_by(country,type)%>% summarise(count=n()) # gets the tota
```

```
## `summarise()` has grouped output by 'country'. You can override using the `.groups` argument.
```

```
u <- reshape(data=data.frame(amount_by_country),idvar="country",  # creates a data frame
                           v.names = "count",
                           timevar = "type",
                           direction="wide") %>% arrange(desc(count.Movie)) %>%
                           top_n(10)
```

```
## Selecting by count.TV Show
```

```
names(u)[2] <- "Number_of_Movies" # changes column names
names(u)[3] <- "Number_of_TV_Shows"

u <- u[order(desc(u$Number_of_Movies +u$Number_of_TV_Shows)),]# orders the values in descending order

figure000 <- ggplot(u, aes(Number_of_Movies, Number_of_TV_Shows, colour=country))+  # outputs the geom_
  geom_point(size=5,alpha=0.5)+
  xlab("Number of Movies") + ylab("Number of TV Shows")+
  ggtitle("Amount of Netflix Content By Top 10 Country")
figure000 # outputs visualization
```



###**Amount of Netflix Content By Time** The visualization goes over the amount of Netflix content by time. We can see there was peak in both TV and Movie around 2019 but there was decrease during 2020.

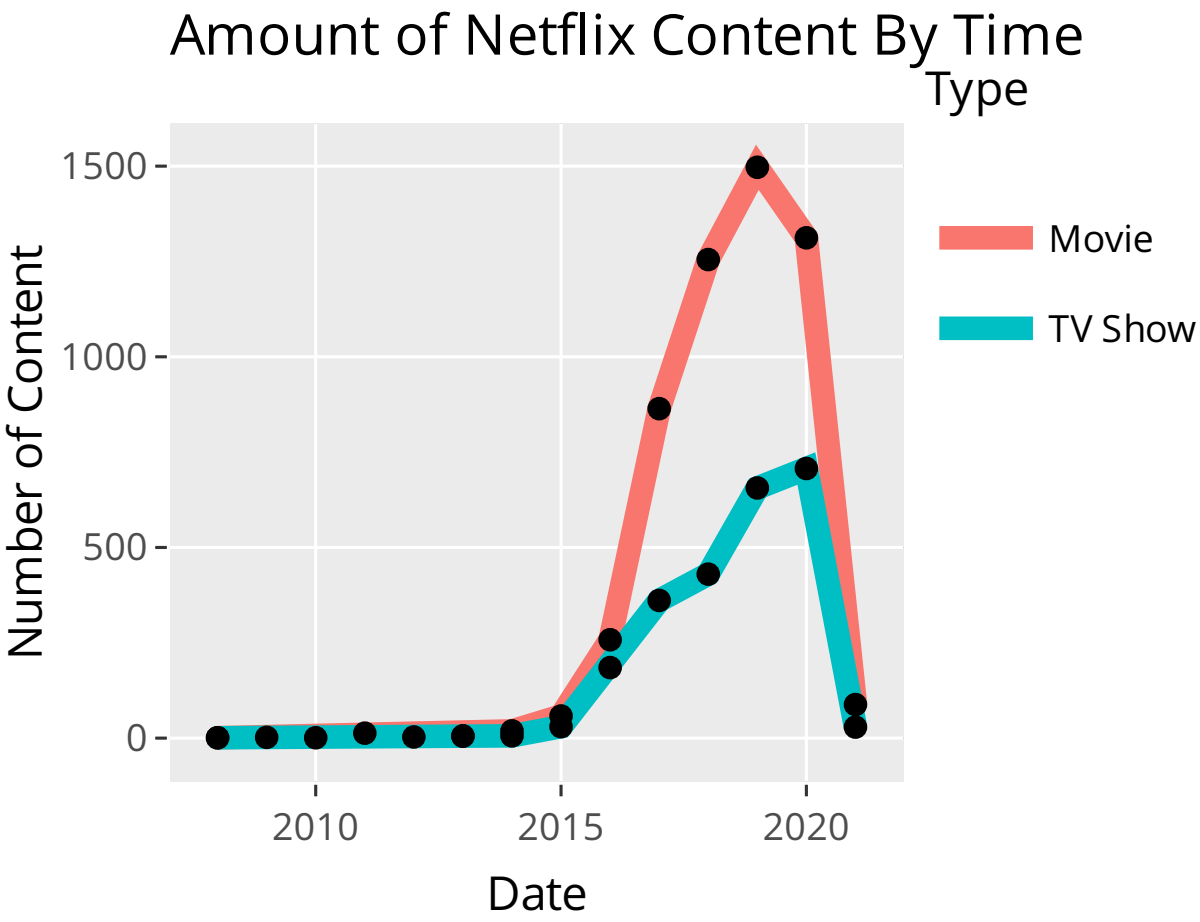
```
f <- data_new2$title # converts the title columns into tibble format
f <-tibble(f)
data_new2$title <- f

# extracts the year from data_added column
data_new2$new_date <- year(data_new2$date_added)

# creates a new data frame  with type and new_date
df_by_date <- data_new2$title %>%
  group_by(data_new2$new_date, data_new2$type) %>%
  na.omit(data_new2$new_date) %>%
  summarise(added_content_num = n())
```

```
## `summarise()` has grouped output by 'data_new2$new_date'. You can override using the `.groups` argume
```

```
head(df_by_date)
```

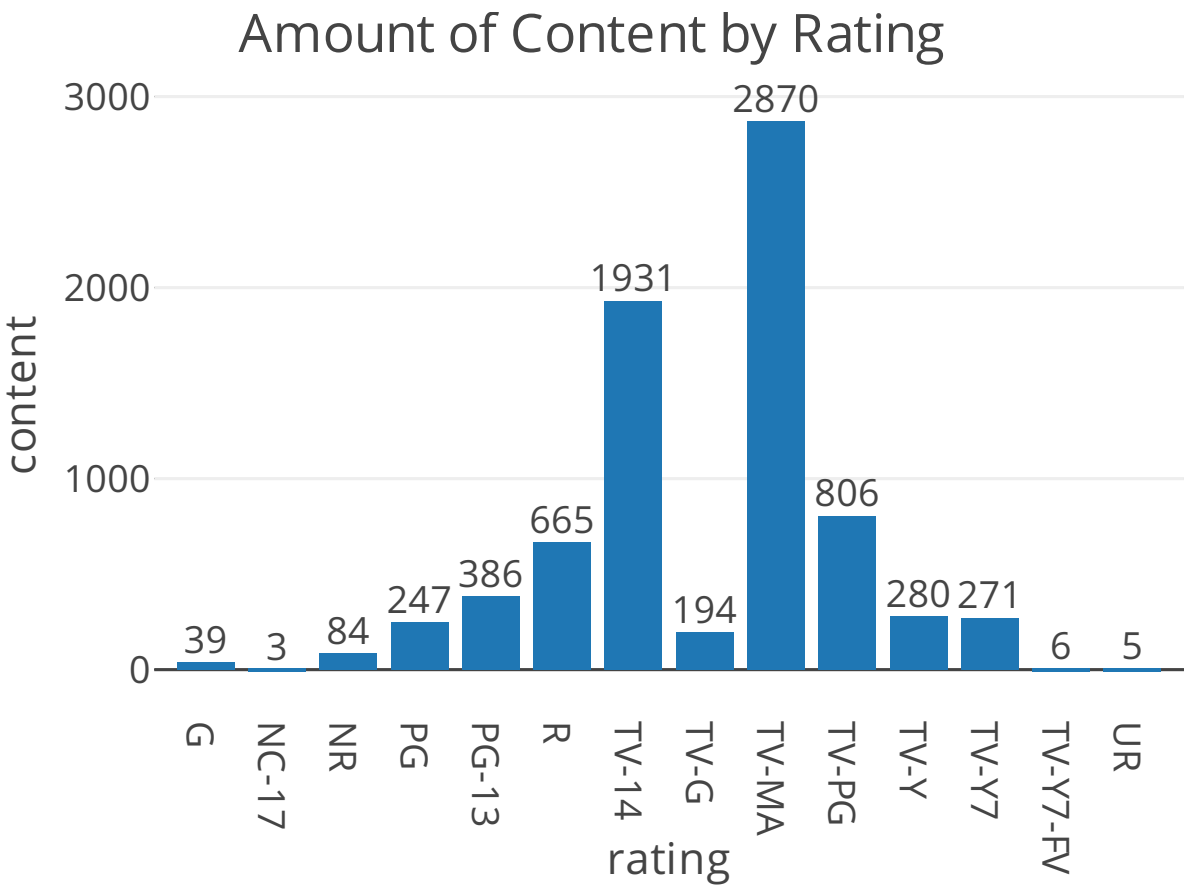
```
## # A tibble: 6 x 3
## # Groups:   data_new2$new_date [5]
##   `data_new2$new_date` `data_new2$type` added_content_num
##                  <dbl> <chr>                        <int>
## 1                 2008 Movie                            1
## 2                 2008 TV Show                          1
## 3                 2009 Movie                            2
## 4                 2010 Movie                            1
## 5                 2011 Movie                           13
## 6                 2012 Movie                            3
```

```
# type,new_date and added_content_num into variables
Type<- df_by_date$`data_new2$type`
Date <- df_by_date$`data_new2$new_date`
Content_Number <- df_by_date$added_content_num
par(mfrow= c(1,2))
# visualization of "Amount of Netflix Content By Time"
g1<- ggplot(df_by_date, aes(Date, Content_Number))+
  geom_line(aes(colour = Type),size=2)+
  geom_point() +
  xlab("Date") +
  ylab("Number of Content")+
  ggtitle("Amount of Netflix Content By Time")

ggplotly(g1, dynamicTicks = T)
```


Amount of Netflix Content By Time
Type
Movie
TV Show
Number of Content
1500
1000
500
0
2010
2015
2020
Date

##**Amount of Content by Rating** The visualization shows the amount of content by rating. When we see the bar char we see the rating of TV-MA has the most with the amount of 2870.

```
# group_by's rating the total count of each
data_new3 <- data_new2$title %>% group_by(data_new2$rating) %>% summarise(content_num=n())

# renames columns to rating and content
names(data_new3) [1] <- "rating"
names(data_new3) [2] <- "content"

# visualization of 'Amount of Content by Rating'
figure_2 <- plot_ly(data_new3, x = ~rating ,y = ~ content, type = 'bar')
figure_2 <- figure_2 %>% layout(title = 'Amount of Content by Rating') %>% add_text(
    text = ~content,
    textposition = "top middle",
    cliponaxis = FALSE,showlegend = FALSE)


figure_2
```


Amount of Content by Rating
content
3000
2000
1000
0
39
3
84
247
386
665
1931
194
2870
806
280
271
6
5
G
NC-17
NR
PG
PG-13
R
TV-14
TV-G
TV-MA
TV-PG
TV-Y
TV-Y7
TV-Y7-FV
UR
rating

##**Amount of Content By Rating (Movie vs. TV Show)** The next visualization goes more in depth with the a stacked bar chart to separate the amount of content by rating for Movie vs. TV Show.

```
# group_by's rating and type with the total count
data2<- data_new2$title %>% group_by(data_new2$rating,data_new2$type) %>% summarise(content_num=n())
```

```
## `summarise()` has grouped output by 'data_new2$rating'. You can override using the `.groups` argumen
```

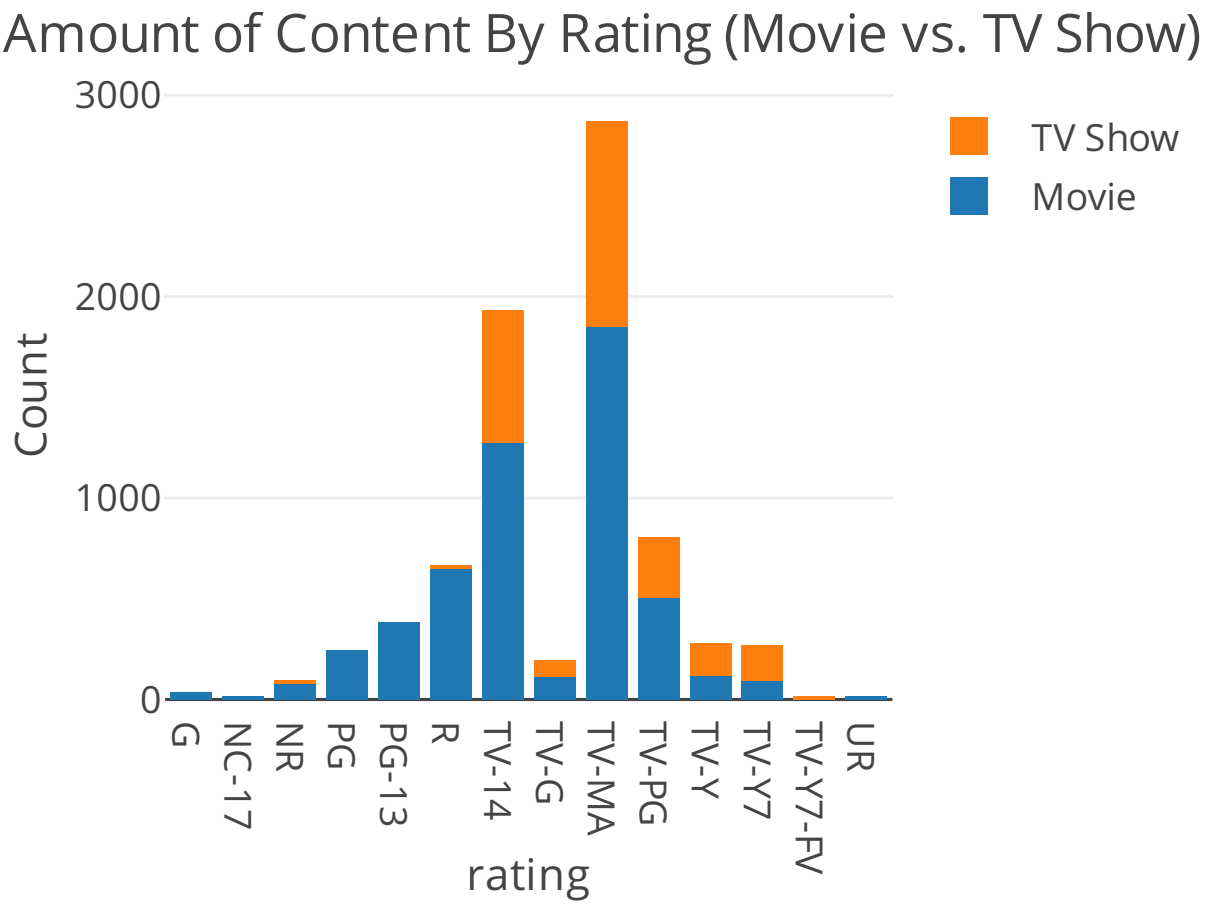
```
#renames columns to rating. type, and content
names(data2) [1] <- "rating"
names(data2) [2] <- "type"
names(data2) [3] <- "content"

head(data2)
```

```
## # A tibble: 6 x 3
## # Groups:   rating [5]
##   rating type    content
##   <chr>  <chr>     <int>
## 1 G      Movie        39
## 2 NC-17  Movie         3
## 3 NR     Movie        79
## 4 NR     TV Show       5
## 5 PG     Movie       247
## 6 PG-13  Movie       386
```

```
# create new data frame
newdata2 <- reshape(data=data.frame(data2),idvar="rating",
                          v.names = "content",
                          timevar = "type",
                          direction="wide")

# changes columns names and and if NA prints 0
names(newdata2)[2] <- "Movie"
names(newdata2)[3] <- "TV Show"
newdata2$`TV Show`[is.na(newdata2$`TV Show`)] <- print(0)
```

```
## [1] 0
```

```
# puts rating, Movie, and TV show into variables
rating <- newdata2$rating
Movie <- newdata2$Movie
Tv_Show <- newdata2$`TV Show`

# visualization for Amount of Content By Rating (Movie vs. TV Show)
figure3 <- plot_ly(newdata2, x = ~rating, y = ~Movie, type = 'bar', name = 'Movie')
figure3 <- figure3 %>% add_trace(y = ~Tv_Show, name = 'TV Show')
figure3 <- figure3 %>% layout(yaxis = list(title = 'Count'),
                         barmode = 'stack',
                         title=("Amount of Content By Rating (Movie vs. TV Show)"))


figure3
```


Amount of Content By Rating (Movie vs. TV Show)
3000
2000
1000
0
Count
TV Show
Movie
G
NC-17
NR
PG
PG-13
R
TV-14
TV-G
TV-MA
TV-PG
TV-Y
TV-Y7
TV-Y7-FV
UR
rating

##**Top 20 Genres On Netflix** The visualization show the top 20 genres on Netflix. When see the genres on the bar graph we can see the international movie as the most amount of content of Netflix.

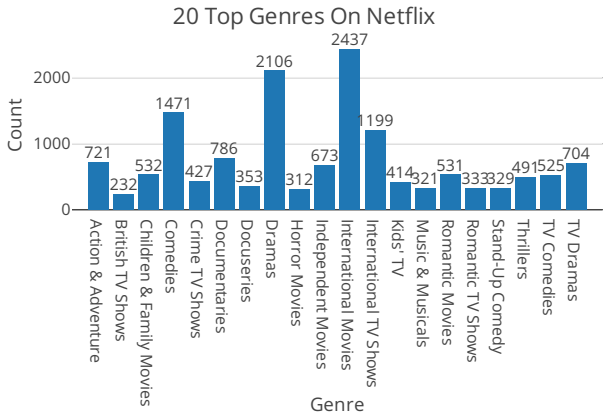
```
# converts listed_in to character
data_new2$listed_in<- as.character(data_new2$listed_in)

# splits listed_in values
t20 <- strsplit(data_new2$listed_in, split = ", ")

# creates data frame and converts to character
count_list_in <- data.frame(type=rep(data_new2$type,sapply(t20,length)),listed_in=unlist(t20))
count_list_in$listed_in <- as.character(gsub(",","",count_list_in$listed_in))

# gets total count of each listed_in values
df_count_listed_in <- count_list_in %>%
                            group_by(listed_in) %>%
                            summarise(count = n()) %>%
                            top_n(20)
```

```
## Selecting by count
```

```
head(df_count_listed_in)
```

```
## # A tibble: 6 x 2
##   listed_in                count
##   <chr>                    <int>
## 1 Action & Adventure         721
## 2 British TV Shows           232
## 3 Children & Family Movies   532
## 4 Comedies                  1471
## 5 Crime TV Shows             427
## 6 Documentaries              786
```

```
# Visualization of "20 Top Genres On Netflix"
figure4 <- plot_ly(df_count_listed_in, x= ~listed_in, y= ~df_count_listed_in$count, type = "bar" )
figure4 <- figure4 %>% layout(xaxis=list(categoryorder = "array",categoryarray =df_count_listed_in$liste
    text = ~count,
    textposition = "top middle",
    cliponaxis = FALSE,showlegend = FALSE)
figure4
```


20 Top Genres On Netflix
Count
2000
1000
0
721
232
532
1471
427
786
353
2106
312
673
2437
1199
414
321
531
333
329
491
525
704
Action & Adventure
British TV Shows
Children & Family Movies
Comedies
Crime TV Shows
Documentaries
Docuseries
Dramas
Horror Movies
Independent Movies
International Movies
International TV Shows
Kids' TV
Music & Musicals
Romantic Movies
Romantic TV Shows
Stand-Up Comedy
Thrillers
TV Comedies
TV Dramas
Genre