College of Professional Studies

Northeastern University San Jose

MPS Analytics

Course: ALY6000 - Introduction to Data Analytics

Assignment:

MODULE PROJECT - 6

EXECUTIVE SUMMARY REPORT – 6

Submitted on:

October 28, 2022

Submitted to:

Professor: BEHZAD AHMADI

Submitted by:

ARCHIT BARUA

NIKSHITA RANGANATHAN

ABSTRACT

The importance of data analytics is growing across all industries, generating vast amounts of knowledge acumen that can offer insightful information. To gain decision-making insights, it is very much significant that data compilation can be accompanied by varied analysis. Data analytics assist organizations in understanding the massive amount of knowledge required for continued success and growth. The significance of data analytics in organizations is a component of strategic growth, enabling them to foresee consumer trends, take actions and decisions based on best-available evidence and increase competitiveness.

In this project, we have considered Netflix which is a No#1 OTT (Over the Top) streaming content platform since its launch in 2007. Netflix reached 203.67 million subscribers worldwide as of the last quarter of 2020 and was predicted and reached an additional 6 million at the beginning of the quarter of 2021. Approximately 36% of the global subscriptions are from The United States with around 73 million new subscribers in 2020 facing additional competition, and the growth of the business is promoted by its international subscribers. And the secret to Netflix's success is Data and analytics.

INTRODUCTION

Netflix has always understood the value of data analytics and has built its business model around it. Netflix started its video streaming services in 2007 but now it has transformed much more than what it offered to the customer. They are massive now that they use analytics to know their customers: What their customer watches? What do they watch? From which geographic location do they watch? How long do they spend watching? When do they watch the content? How many times do they pause the content? Which content do they pause? Which content do they fast-forward and watch? Which content do they repeatedly watch? And many more to add.

The purpose of this project is to provide an exploratory analysis of the movies and tv shows that are available on the Netflix platform. The data that is used within the project is sourced from *Kaggle* by the data owner, Shivam Bansal. The project focuses on the overview of the insight including the growth of Netflix content by year and percentage of the content compared to TV Shows and Movies which aims to provide movie fans to discover the Netflix contents which are presented in a variety of data visualizations.

About this Dataset: Netflix is one of the most popular media and video streaming platforms. They have movies or tv shows available on their platform, as of mid-2021, it has around 200M subscribers worldwide. This tabular dataset comprises of records of all the content available on Netflix, along with details such as - titles, directors, ratings, time duration, etc.(source-*Kaggle*)

The data set consists of numerical as well as categorical values. It has 6164 observations with 10 features.

Below are the data descriptions of each variable that briefly describe the contents of the data set. The feature of the dataset is as follows:

No	Feature	Dictionary
1	show id	The unique id represents the content/show (TV Shows/Movies)
2	type	The type of content/show (TV Shows/Movies)
3	title	The title of the content/show (TV Shows/Movies)
4	director	Name of the director(s) of the content/show
5	country	Country in which contents were produced
6	date added	Date of the content/show added to the Netflix platform
7	year added	Actual year in which the content/show got added to Netflix
8	rating	Ratings of the content/show (viewer ratings)
9	duration	Length of the duration of the content/show (number of series for TV Shows and number of minutes for Movies)
10	listed_in	List of genres in which the contents were listed in

Table 1: Features of the Netflix data set with their dictionary

DATA CLEANING AND MANIPULATION

- Importing the Netflix csv file
  <netflix> vector contains the details of the content in Netflix.

```
> getwd()
[1] "C:/Users/14086/OneDrive - Northeastern University/Desktop/ALY 6000/Assignments/Module -6"
> netflix=read.csv(file = 'netflix_titles.csv')
```

Figure 1-read.csv()

- Cleaning the dataset
  Data cleaning is a method used to remove inaccurate information to simplify analysis.
  We cleaned the data by checking if there are any missing values or duplicate values.

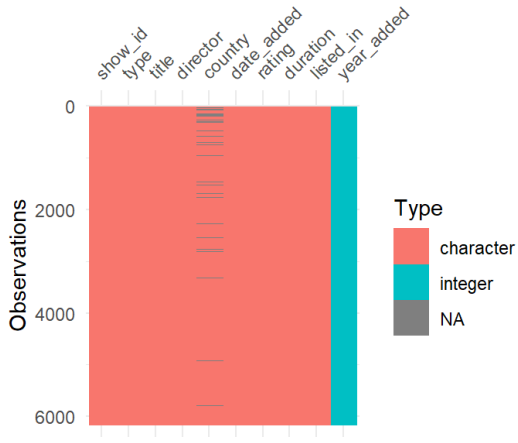
  - Replacing missing values with <NA> values

```
> netflix1<-netflix
> netflix1[netflix1 == "" | netflix1 == " "] <- NA
```

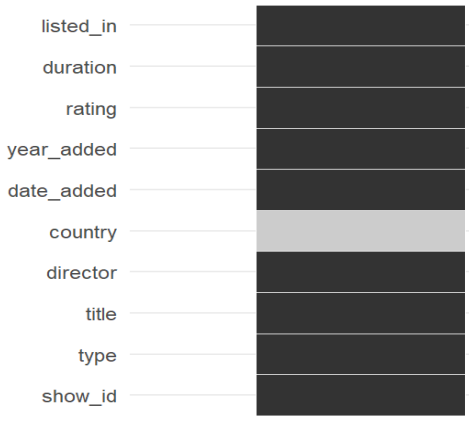
Figure 2 – Adding <NA>s to blanks

  - Visualization of <NA> values
    vis_dat provides a glance of missing data.
    gg_miss_which is a function under the naniar package and is used to check the column that has the blanks.
    miss_var_summary() summarizes the missingness under each variable. The country column has 422 missing values

```
> vis_dat(netflix1)
```




```
> library(naniar)
> gg_miss_which(netflix1)
```




```
> miss_var_summary(netflix1)
# A tibble: 10 × 3
   variable   n_miss pct_miss
   <chr>       <int>    <dbl>
 1 country       422     6.85
 2 show_id         0        0
 3 type            0        0
 4 title           0        0
 5 director        0        0
 6 date_added      0        0
 7 year_added      0        0
 8 rating          0        0
 9 duration        0        0
10 listed_in       0        0
```

Figure 3 - <NA> value representation

- Eliminating the <NA>s
  The number of rows reduced from 6164 to 5742 (i.e 422 rows corresponds to Figure 3).

```
> complete.cases(netflix1)
> which(!complete.cases(netflix1))
> na<-which(!complete.cases(netflix1))
> cleanset<- netflix1[-na,]
> view(cleanset)
```

Figure 4 – Code for obtaining clean Netflix dataset without <NA> values

- Manipulating cleanset
  Using only the first content of the columns country, listed_in and directors and eliminating others from the list. This is to visualize better.

```
> cleanset$country<- sub(",.*","",cleanset$country)
> view(cleanset)
> cleanset$listed_in<- sub(",.*","",cleanset$listed_in)
> view(cleanset)
> cleanset$director<- sub(",.*","",cleanset$director)
> view(cleanset)
```

Figure 5 – Extracting first contents

The column name listed_in is replaced by genre. We also removed columns 1 and 5 (show id and date_added).

```
> colnames(cleanset)[colnames(cleanset) == "listed_in"] <- "genre"
> view(cleanset)
> which(!complete.cases(cleanset) )
integer(0)
> cleanset<-cleanset[ , -c(1)]
> cleanset<-cleanset[ , -c(5)]
```

Figure 6 – Changing column names and removing columns

The rating column is replaced by 3 categories Adults, Teens, and Kids with the help of recode() and mutate(). The column name now is the audience category.
We used as.factor() to change the data type of the audience category to factor.

```
> rating<-cleanset$rating
> cleanset=cleanset %>% mutate(rating=recode(rating,"TV-PG"="Kids","TV-MA"="Adults","TV-Y7-
FV"="Kids","TV-Y7"="Kids","TV-14"="Teens","R"="Adults","TV-Y"="Kids","NR"="Adults","TV-
G"="Kids","PG-13"="Teens","PG"="Kids","G"="Kids","UR"="Adults","NC-17"="Adults"))
> colnames(cleanset)[colnames(cleanset) == "rating"] <- "audience_category"
> view(cleanset)
> cleanset$audience_category <- as.factor(cleanset$audience_category)
> class(cleanset$audience_category)
[1] "factor"
```

Figure 7 – Modifying rating as audience category and changing datatypes

DATA ANALYSIS

- Structure
  The str() function displays the various datatypes for the variables.
  Cleanset has 5762 observations and 8 variables. It can be observed that the audience category is a factor with 3 levels and the year added is an integer. All other variables are characters.

```
> str(cleanset)
'data.frame':   5742 obs. of  8 variables:
 $ type            : chr  "Movie" "Movie" "TV Show" "Movie" ...
 $ title           : chr  "Dick Johnson Is Dead" "Sankofa" "The Great British Baking Show" "The Starling"
 ...
 $ director        : chr  "Kirsten Johnson" "Haile Gerima" "Andy Devonshire" "Theodore Melfi" ...
 $ country         : chr  "United States" "United States" "United Kingdom" "United States" ...
 $ year_added      : int  2021 2021 2021 2021 2021 2021 2021 2021 2021 2021 ...
 $ audience_category: Factor w/ 3 levels "Adults","Kids",..: 3 1 3 3 1 3 3 3 3 3 ...
 $ duration        : chr  "90 min" "125 min" "9 Seasons" "104 min" ...
 $ genre           : chr  "Documentaries" "Dramas" "British TV Shows" "Comedies" ...
```

Figure 8 – str()

- Summary
  There are no NA values in the cleanset. There are 2843, 1089 and 1810 observations in the Adults, Kids, and Teens categories respectively.

```
> summary(cleanset)
     type               title              director
 Length:5742        Length:5742        Length:5742
 Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character


   country            year_added   audience_category
 Length:5742        Min.   :2008   Adults:2843
 Class :character   1st Qu.:2018   Kids  :1089
 Mode  :character   Median :2019   Teens :1810
                    Mean   :2019
                    3rd Qu.:2020
                    Max.   :2021
   duration             genre
 Length:5742        Length:5742
 Class :character   Class :character
 Mode  :character   Mode  :character
```

Figure 9 – summary()

- Glimpse
  glimpse() is part of the dplyr package and we can see the preview of columns of the dataset with the help of this function.

```
> glimpse(cleanset)
Rows: 5,742
Columns: 8
$ type             <chr> "Movie", "Movie", "TV Show", "Movie", …
$ title            <chr> "Dick Johnson Is Dead", "Sankofa", "Th…
$ director         <chr> "Kirsten Johnson", "Haile Gerima", "An…
$ country          <chr> "United States", "United States", "Uni…
$ year_added       <int> 2021, 2021, 2021, 2021, 2021, 2021, 20…
$ audience_category <fct> Teens, Adults, Teens, Teens, Adults, T…
$ duration         <chr> "90 min", "125 min", "9 Seasons", "104…
$ genre            <chr> "Documentaries", "Dramas", "British TV…
```

Figure 10 – glimpse()

- Dimension
  dim() views the number of rows and columns of a matrix, data frame, or array. In this case, we have 5742 rows and 8 columns.

```
> dim(cleanset)
[1] 5742    8
```

Figure 11 – dim()

- Headtail
  This function is a part of the FSA package. We can see the first 3 and last 3 records of the cleanset below.

```
> headtail(cleanset)
        type                          title        director
1      Movie          Dick Johnson Is Dead Kirsten Johnson
5      Movie                       Sankofa    Haile Gerima
6    TV Show The Great British Baking Show Andy Devonshire
6162   Movie                    Zombieland Ruben Fleischer
6163   Movie                          Zoom    Peter Hewitt
6164   Movie                        Zubaan     Mozez Singh
            country year_added audience_category  duration
1     United States       2021             Teens    90 min
5     United States       2021            Adults   125 min
6    United Kingdom       2021             Teens 9 Seasons
6162  United States       2019            Adults    88 min
6163  United States       2020              Kids    88 min
6164          India       2019             Teens   111 min
```

Figure 12 – headtail()

DATA VISUALIZATION

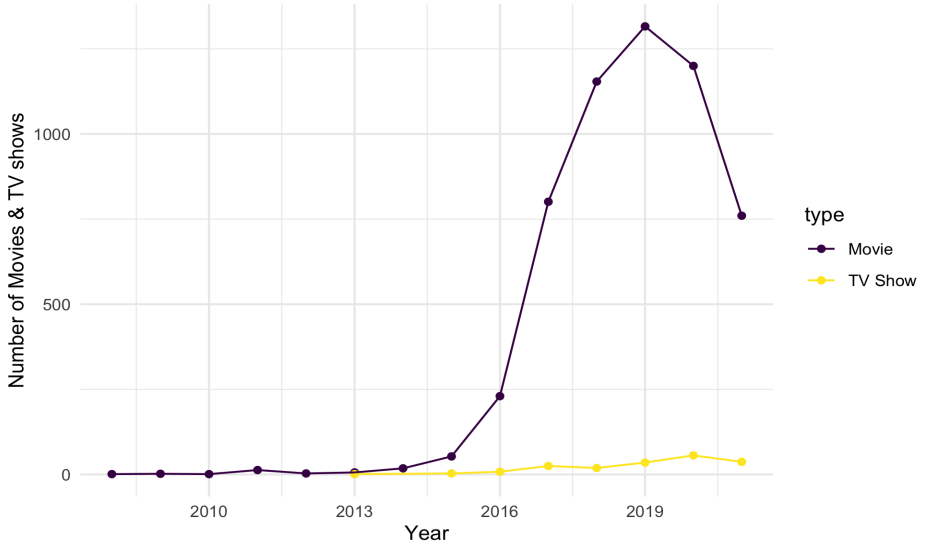


Figure 13 – Line chart

The use of a line chart in this case helps us visualize the increase in the number of contents over a certain period of time. It shows us that the trend of watching content on online platforms sharply rose post-2015, especially for movies. The rise of big data in the last 5-6 years can also be justified by looking at this trend.

Personalization is one of the key aspects of big data and not surprisingly Netflix gets 70% of its total views because of the recommendation page put forward to its consumers. The dramatic difference between movies and tv shows post-2014 gives us an idea that the viewers preferred watching short content over 1 to 2 hours rather than content having multiple seasons such as tv shows.

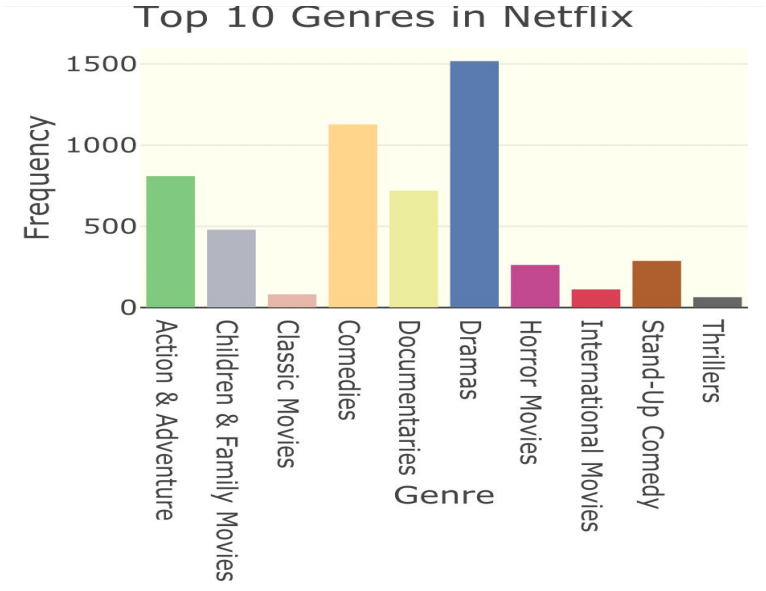


Figure 14 – Bar graph

The bar plot above shows us the top 10 genres on Netflix. There is no surprise drama tops it,

As we humans love drama, be it on or off the screen! The comedy genre takes the second place and it tells us a story that most viewers after a hard day of work just want to ease down to something simple yet entertaining without having to focus much. Comedy is also the number 1 killer of boredom.

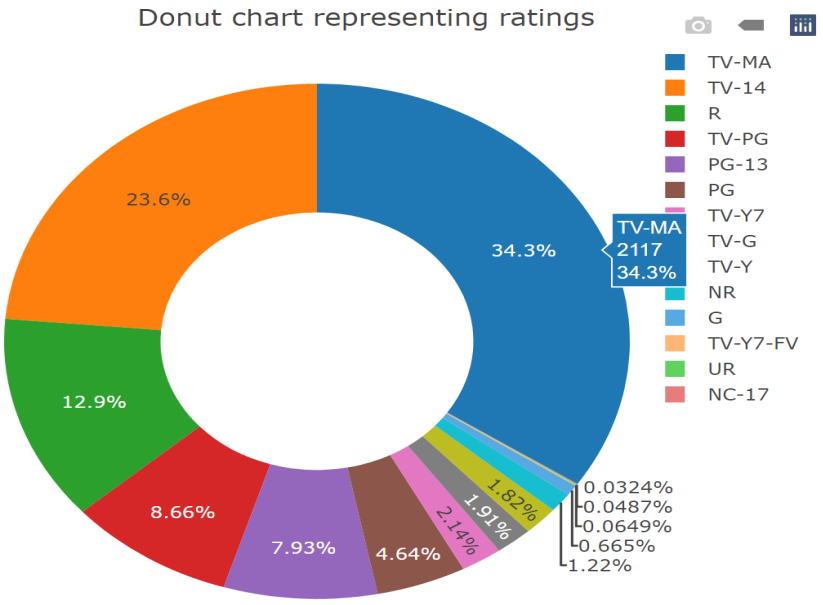


Figure 15 – Donut Chart

The following chart gives us an in-depth explanation of the targeted audience of Netflix. 34.3% of the shows are for mature audiences (18+) followed by 23.6% for PG-13 which are teens between 14-17. It gives us information that Netflix is mostly targeting the age group which is highly active on social media. That is a smart business move as engaging viewers who are active on social media platforms will give Netflix free marketing as most of these users and viewers will take to social media to post about a tv show or movie they just watched on Netflix hence attracting more viewers.

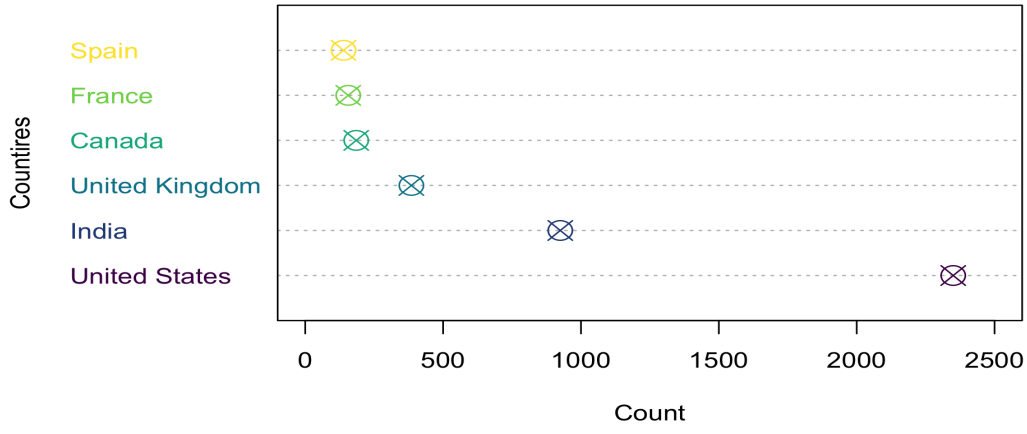


Figure 16 – Dot chart

The above map gives us the top 6 countries with the number of content respectively on Netflix. Large numbers of content are driven for and by the USA market which is not a surprise as Netflix is a US-based firm. India and UK follow behind respectively. A key takeaway in terms of a business opportunity for Netflix is to tap the Indian subcontinent market, it has a huge potential in terms of growing its overall viewers as India is the 2nd most populated country in the world with a population of 1.4 billion compared to the 332 million of the United States.

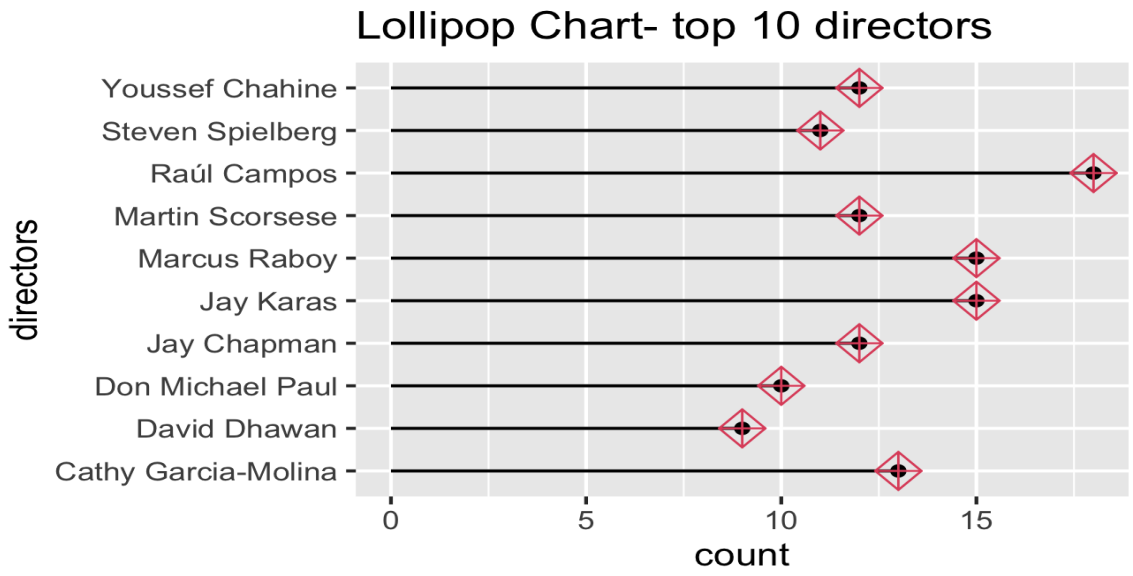


Figure 17 – Lollipop chart

The chart above gives us a detailed visualization of the top 10 directors in terms of frequency on

Netflix. Raul Campos who is a director, writer, and producer with a Spanish background tops the chat.

There are some big star names also on this list such as Steven Spielberg and David Dhavan.

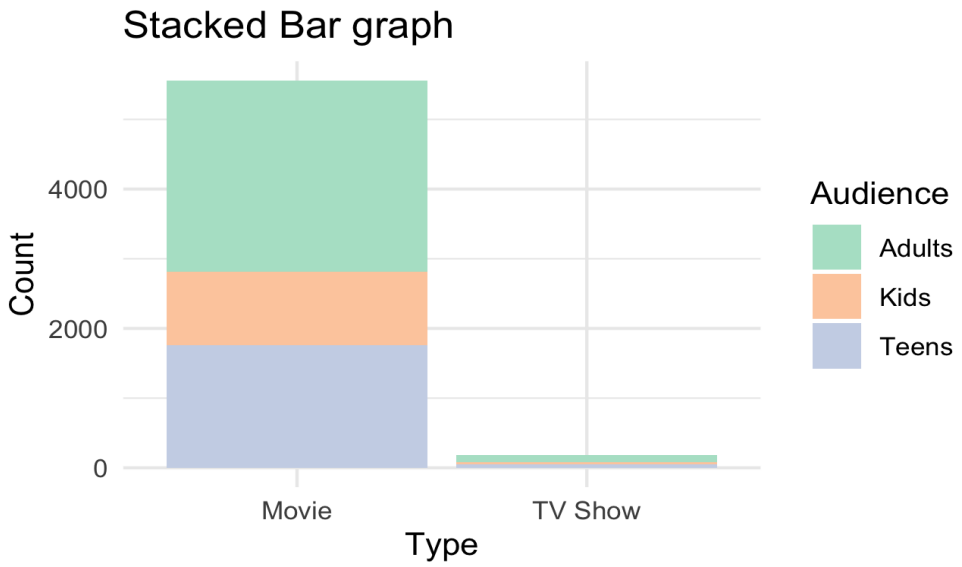


Figure 18 – Stacked bar graph

The stacked bar graph helps us visualize the density of the 3 audience category factors with the two main variables such as the movie and tv show. As we can clearly see, movies outweigh the number of tv shows by a huge margin. While just taking the movie variable into account, content for adults dominates the movie type followed by teens and kids respectively.

CONCLUSION

This particular Netflix data set has some informative data and a lot of variables that could be compared and contrasted against each other to explore future deep analysis and produce data-driven answers and questions. The graphs presented in this file gives us a visualized picture of what this data set is all about and the key is to make the data set easy to understand and analyze for the business managers with the help of data visualization to make vital data-driven business decisions.

Key takeaway points-

- The effects of big data in the last 5-6 years can be seen by the unprecedented growth of content on online platforms such as Netflix.
- Drama and comedy are the go-to genres for the majority of the masses.
- The targeted audience for Netflix based on the age demography is mostly adults and teens.
- USA dominates the number of content on Netflix but there is an interesting question to be asked which is, will tapping into the Indian subcontinent turn out to be a profitable decision for Netflix? - as it is a country with almost 1.4 billion people.
- The directors chart shows us that though there are some big names present on this platform, there is plenty of room for new directors to showcase their talents.

REFERENCES

Bansal, S. (2021, September 27). *Netflix movies and TV shows*. Kaggle. Retrieved October 28, 2022, from https://www.kaggle.com/datasets/shivamb/netflix-shows

Bluman, A. G. (2018). Elementary Statistics, 10th ed. McGraw Hill.

Kabacoff, R. I. (2011). R in action: Data analysis and graphics with R. Manning Publications Co.

Low-code data app development. Plotly. (n.d.). Retrieved October 28, 2022, from https://plotly.com/

APPENDIX

```
---
title: "Module Project-6"
author: “Group”
output:  word_document: default
date: "2022-10-28"
---

# INSTALLING LIBRARIES
library(tidyverse)
library(scales)
library(lubridate)
library(igraph)
library(networkD3)
library(ggplot2)
install.packages("viridis")
library(viridis)
library(plotly)
install.packages("plotly")
library(RColorBrewer)
install.packages("visdat")
install.packages("naniar")
library(visdat)
library(naniar)
library(car)
install.packages("dplyr")
library(dplyr)
library(FSA)
library(magrittr)

# UPLOADED NETFLIX DATA SET CONTAINING 6164 OBS WITH 10 VARIABLES
getwd()
netflix=read.csv(file = 'netflix_titles.csv')

# CLEANING AND MANIPULALTING THE DATA SET FOR VISUALIZATION
# REDUCING THE OBSERVATIONS TO 5742
netflix1<-netflix
netflix1[netflix1 == "" | netflix1 == " "] <- NA
vis_dat(netflix1)
gg_miss_which(netflix1)
miss_var_summary(netflix1)
complete.cases(netflix1)
which(!complete.cases(netflix1))
na<-which(!complete.cases(netflix1))
cleanset<- netflix1[-na,]
view(cleanset)
```

```
# TAKING THE FIRST CONTENT OF THE FOLLOWING VARIABLES AS THEY ORIGINALLY HAVE A LONG LIST WHICH WOULD BE DIFFICULT TO VISUALIZE
cleanset$country<- sub(",.*","",cleanset$country)
view(cleanset)
cleanset$listed_in<- sub(",.*","",cleanset$listed_in)
view(cleanset)
cleanset$director<- sub(",.*","",cleanset$director)
view(cleanset)

# CHANGING THE COLUMN NAME AND DELETING COLUMNS THAT ARE NOT REQUIRED IN ORDER TO HAVE 8 VARIABLES
colnames(cleanset)[colnames(cleanset) == "listed_in"] <- "genre"
view(cleanset)
which(!complete.cases(cleanset) )
cleanset<-cleanset[ , -c(1)]
cleanset<-cleanset[ , -c(5)]


#REPLACING OR MODIFYING THE RATING COLUMN TO AUDIENCE CATEGORY WITH KIDS, TEENS, AND ADULTS.

rating<-cleanset$rating
cleanset=cleanset %>% mutate (rating= recode (rating, "TV-PG" = "Kids" , "TV-MA" = "Adults" , "TV-Y7-FV" = "Kids" , "TV-Y7" = "Kids" , "TV-14" = "Teens" , "R" = "Adults" , "TV-Y" = "Kids" , "NR" = "Adults" , "TV-G" = "Kids" , "PG-13" = "Teens" , "PG" = "Kids" , "G" = "Kids" , "UR" = "Adults" , "NC-17" = "Adults "))
colnames(cleanset)[colnames(cleanset) == "rating"] <- "audience_category"
view(cleanset)

# CHANGING DATA TYPES AND VIEWING THE STR
cleanset$audience_category <- as.factor(cleanset$audience_category)
class(cleanset$audience_category)
str(cleanset)
glimpse(cleanset)
headtail(cleanset)
dim(cleanset)
summary(cleanset)
view(cleanset)

# GRAPH 1
cleanset %>% group_by(type, year_added) %>% summarize(count=n()) %>%
ggplot(aes(year_added, count)) + geom_point(aes(color=type)) + geom_line(aes(color=type)) + scale_color_viridis(discrete=TRUE) + scale_x_log10() + theme_minimal() + labs(y='Number of Movies & TV shows',x='Year')

# GRAPH 2
df2<-cleanset %>% group_by(genre) %>% summarise(count = n()) %>% top_n(10)
RColorBrewer::display.brewer.all()
fig2<-plot_ly(df2,x= ~genre,y= ~count,type = 'bar',color= ~ genre,colors = "Accent")
fig2<-fig2 %>% layout(title="Top 10 Genre in Netflix",
plot_bgcolor = "Ivory",showlegend = FALSE,xaxis = list(title = 'Genre'),
```

```
yaxis=list(title='Frequency'))
fig2

# GRAPH 3
df1 <- netflix1 %>% group_by(rating) %>% summarise(count = n())
View(df1)
fig <- df1 %>% plot_ly(labels = ~rating, values = ~count)
fig <- fig %>% add_pie(hole = 0.5)
fig <- fig %>% layout(title = "Donut chart representing ratings",xaxis = list
(zeroline = F, showticklabels = F,showgrid = T),yaxis = list
(zeroline = F, showticklabels = F,showgrid = T))
fig


# GRAPH 4
type<-cleanset$type
ggplot(cleanset,aes(x=type,fill=audience_category))+geom_bar(position="stack")+
scale_fill_brewer(palette = "Pastel2")+
labs(y = "Count", fill = "Audience",x = "Type",title = "Stacked Bar graph") + theme_minimal()

# GRAPH 5
x<-table(cleanset$country)
x
y<-data.frame(x)
y
aw<-y[order(y$Freq, decreasing = TRUE),]
b<-head(aw)
class(b)
b
data_bar<-b$Freq
names(data_bar) <-b$Var1
data_bar
dotchart(data_bar,pch= 13,col = hcl.colors(6),pt.cex = 2, xlim = c(0,2500),ylab = "Countries",
xlab = "Count")

# GRAPH 6
m<-table(cleanset$director)
m
n<-data.frame(m)
n
mw<-n[order(n$Freq, decreasing = TRUE),]
mw
v<-head(mw,n=10)
v
b_bar<-v$Freq
names(b_bar)<-v$Var1
b_bar

ggplot(v, aes(x = v$Var1, y = v$Freq)) +geom_segment(aes(x = v$Var1, xend = v$Var1, y =
0, yend = v$Freq)) + geom_point() +geom_point(size = 4, pch = 9, bg = 4, col = 18) +
coord_flip()+ labs(y="count", x= "directors",title= "Lollipop Chart- top 10 directors")
```