Final Report

Analysis Report on Netflix Streaming Services

TV SHOWS

MKT 445 – 001

Submitted by:

Jagannath Kharel
Illinois State University

Abstract

Netflix is distinctly more popular with young viewers in the United States and the rest of the world than with the older generation. The industry speeds up and simplifies access to digital content, to keep it in a leading position among other competitors for gaining success and controlling the downward trend of subscriptions of media-streaming and video-rental companies. This study aims to understand the usage and awareness of Netflix streaming services among different customers, especially young adults aged above 18, to gain a better knowledge about the industry. By using competitors' data such as Hulu, Prime Video and Disney Plus, this study focuses on where Netflix can maximize growth in the targeted age groups they are most popular with. The study's primary research was mix methods, both qualitative and quantitative, followed by descriptive and analytical using. The secondary dataset from a public source, called Kaggle.com, was used after merging both datasets to analyze the data on Netflix using R programming. To find the results of every research question, different methods for each question were used to analyze the data set and to reach the findings. After cleaning the final dataset, we used altogether 1062 responses with 8 variables for more than a decade period 2012 to 2022.

The results indicated that TV shows targeted the age group of 18 and Rotten Tomato scores scores that were greater or equal to 80/100. It is further found that Netflix usage and awareness is focused on young adults more than the other 3 streaming services. Netflix also has more of the newer and popular shows based on the "start year" and "Rotten Tomato" scores. In addition, it is found the Age variable is significant at 0.05 level with the TV shows and also found negative association among all variables. The fitness of multiple regression explained the Neflix watching variation with help of R-Squred values about 51% of the change of independent variables, such as Age and other predictors i.e.Netflix, Hulu, Prime Video, Disney plus. Surprisingly, there are multiple popular genres on each streaming service. It is interesting that the drama genre is the winner with having three out of the four popular genres on the four streaming services. Moreover, the result shows what emotions are most popular on Netflix and it is found that most of Netflix shows have a positive sentiment. The conclusion implies that if you are the average young adult, you should purchase Netflix over other streaming services.

Introduction

Online Streaming is becoming more and more popular every day. It makes access to digital content, whether it be videos or audio, much faster and easier. While streaming content, the user is not required to download that content but can watch it online. What the user requires is a good Internet connection and a good service provider. This is where Netflix comes in. Netflix is among various service providers that provide access to unlimited content for a fixed monthly fee. Video on demand (VOD) is a programming system which allows users to select and watch/listen to video or audio content such as movies and TV shows whenever they choose, rather than at a scheduled broadcast time, the method that prevailed with over-the-air programming during the 20th century.

Television VOD systems can stream content through either a set-top box, a computer or other device, allowing viewing in real time, or download it to a device such as a computer, digital video recorder or portable media player for viewing at any time. The arrival of video streaming services revolutionized the entertainment industry and with it, many of our habits also changed. Whether you spend hours or minutes watching a movie or TV series, Netflix has become a favorite on the scene. Netflix's primary focus shifted to video on demand via the internet in 2007. Even though DVD sales dropped drastically, Netflix started growing exponentially. Netflix was launched on April 14, 1998, as the world's first online DVD rental store, with only 30 employees and 925 titles available, which was almost the entire catalog of DVDs in print at the time, through the pay-per-rent model with rates and due dates that were similar to its bricks-and-mortar rival, Blockbuster International expansion. For this analysis, the data set consists of TV shows and movies available on Netflix, Hulu, Prime Video and Disney Plus as of 2010 and part of 2022

Main objective: To understand the usage and awareness of Netflix streaming services among young adults aged above 18, to gain a better knowledge about the industry.

The specific objectives of this research project are:

a) Develop a model to analyze and predict rating score and trend among ages 18 and over on Netflix
b) Determine whether or not the Netflix target age group is justifiable
c) What is the most popular genre on Netflix?

Research Questions:

Research Question 1: What types of shows on different streaming services have a score of at least 80/100 , within the last 12 years,for young adults aged between 18 and over?

Research Question 2: What age group is Netflix mainly trying to target? To what extent and in what manner,Age and other predictors could explain the Netflix watching variation?

Research Question 3: What is the most popular genre throughout the different streaming services using results from Research question one?

Research Question 4: What sentimental factors are the most popular on Netflix?

Our approach is taking the bigger picture and narrowing down the data within Netflix to see what the data gives us as a whole. This will allow us to target each age demographic as a whole overall and give us the most hits with each genre of a tv show. We will visualize the data with graphs to forecast the age group and the most popular genre on multiple streaming services. This will compare why people choose Netflix over other streaming platforms that are available to the public in looking at the show ratings. These ratings will provide insight to customers on what streaming service has the higher ratings overall.

With these results, we hope to benefit young adults, aged 18 and over, with reliable streaming service that will provide them with an enjoyable show to watch. We will conclude our study by providing a brief summary of our results. The limitation to this study is getting data from the internet. We will discuss further why this will hinder our research at the end.

Data:

We found our databases in Kaggle. The first dataset we are using is called TV shows on Netflix, Prime video, Hulu and Disney plus. This dataset is authored by Ruchi Bhatia. There are 11 different variables within this dataset which are: ID, Title, Year, Age, IMDb, Rotten Tomatoes, Netflix, Hulu, Prime Video, Disney Plus. The second dataset is called Netflix Movie and TV shows (June2021) by Snehaan Bhawal. There are 19 variables included in this dataset are Imdb_id, Title, popular Rank, certificate, Start Year, End year, Episodes, runtime, type, origin country,language, plot, summary, rating, popular_rank, num votes, genres, is adult cast and image_url. For our first research question we will use Start Year, title, Rotten Tomatoes , age, Netflix, Hulu, Disney +, and Prime Video variables from the two data sets. Start Year will help us see what show started within the last 12 years. The title will provide us with the show names. Rotten Tomatoes will give us a score for what show is currently good to watch. Age will provide us with what age group the show is attempting to target. Netflix, Hulu, Disney +, and Prime Video will provide us with which show is on what stream. For our second research question, we will use the title to provide us with the show names. Age to determine what age group a show is trying to target. With this we will be able to find the mode of the age group. Netflix will tell us what show is on that specific stream. For our third research question, we will use the results from question one in order to have the targeted shows needed. The genre variable will give us further insight as to the most popular among tv shows on streaming services. For our fourth research question, we will use the Summary to determine the show's sentimental score. This will provide us with what sentiments are the most popular on Netflix.

The First step is to read the two datasets into R. The second step is to merge the two datasets into one dataset using the merge function in R. We are using two different datasets for this project.We merged the two files together by title since that was one common variable between the two datasets. The first dataset is called netflix list, which has shows all around the country that are currently streaming on Netflix or was on Netflix. The variables consist of the title, popular rank, certificate, start year, end year, episodes, runtime, type, origin country, language, plot, summary, rating, numvotes, genres, is adult, cast, and image url. The second data is called tv shows and contains information about the year it started, age appropriate is the show, the IMDB score, the Rotten Tomato's score and the different streaming platforms the shows are on. This will allow us to get more of a complete picture with the different TV shows. We removed the following variables from the dataset to focus in on the our research questions, which are the numbers column for each movie listed, IMDB, IMDB id, Popular rank, certificate, End

Year, episodes, runtime, type, origin country, Language, plot, Rating Num votes, adult, case, x, id, year, Image Url and the count of type variable for the show. The variable Age has an ALL category which we replaced with the value 0, this is because anyone can watch the show. The last step is to create a text document for all the tv show summaries on Netflix. This will allow us to identify the sentimental scores for the summaries overall.

After cleaning the data we had 1062 different tv shows to work with. The columns left that we will be using are title, start year, summary, genres, age, rotten tomato, Netflix, Hulu, Prime Video, and Disney. Using these variables will help us lead to answers for our four research questions. Eventually, leading us to understand the usage and awareness of Netflix streaming services among young adults aged above 18, to gain a better knowledge about the industry.

Methods and process

The key research design of the study will be descriptive, analytic and prescriptive to understand the different aspects of streaming tv shows such as pattern of traction and rating and popular genre among college aged groups in about more than a decade. Here, a predictive model will be used to analyze the dataset to understand patterns, relations between ages and popular shows on different streaming services and trends of the trending shows among the target groups (youths). This model will be able to help us identify our objective of analyzing and spotting trends among young adults aged above 18 using streaming services. Including correlation, regression model fitness is run to explore the relationship between the age and the shows and targeted audiences and their attributes.

Firstly, study variables will be selected and filtered from the CSV file, then data cleaning will be done along with checking the relevancy, authenticity and validity of data.The linear model will help to explain the impact of attributes on popularity of the shows. The dependent variable is Rotten Tomato, whereas the independent variables will be features of shows such as genres of the shows on streaming services. Variables will be checked for their significance and kept or removed from the regression model. This will assist the audience if they are interested at the current time. This model will be drawn from the csv data file of all the data on the streaming services.

This ultimately will help to bring the broad understanding how the tv shows are popular in the specific age group and what other factors are associated with, which will increase the tv show ratings.

The following steps will be taken to analyze the data:

A short data set will be taken to remove variables that are not relevant to the study. We will mainly use variables such as the age of the individuals watching, tv show ratings, what shows are on what streaming service, and the genre of the tv show. Also, the dataset will be checked for missing values and the data will be imputed based on central tendency.

Research Question 1 Method:

For our first research question, we filtered out the shows that were older than 12 years. This will provide us with which TV shows are newer. We used the start year and filtered out the shows that were greater than or equal to 2010. The next step we did is to filter out the TV shows that didn't target our age range for our main objective of 18 . We used the age variable in our data set. The next filter we applied was the Rotten Tomato Variable to focus on what shows will be a good show for young adults to watch on different streaming services. The Rotten Tomato value for the filter was greater or equal to 80 out of 100 rating. We used the sum code to figure out how many shows on each streaming service they had total.

Research Question 2 Method:

For our second research question, we want to see what age group Netflix mainly targets and every TV show that is on Netflix. . To figure this out, the multiple targeted age groups given were blank, all, 7, 16, and 18. We subset the data to remove the blank age group because this was missing data and doesn't tell us what age appropriate group a specific TV show is targeting. The "all" value is considered to be all ages for the age variable. We replaced the value "all" with the value of 0 due to the fact that anyone could watch the show. We selected the values of 0, 7 ,16 and 18 to be able to find the average age group for Netflix's target age group. We took out the blank value because there was a missing value. R SQL used to filter the data. .Apart from this, we analyzed the correlation and fitness of the regression model with Age and other variables on Netflix. Applying chart correlation test and regression fitness test, we analyzed to what extent and in what manner Age and other predictors could explain the Netflix show watching variation.

Research Question 3 Method:

To find the popular genre amongst the different streaming services, we will analyze the results from research question one to start answering this third research question. The data is going to be subset into each streaming service to identify the data into four parts. We will conduct text analysis in producing a word cloud image of each one to gain insight of what genre is the most popular a streaming service has to offer.

Research Question 4 Method:

The sentiment analysis is also called opinion (positive, negative, and neutral) mining involves extracting opinions, sentiments, feedback and emotion on the dataset. The sentimental analysis is widely used to track attitudes and feelings on text data for measuring performance. For this, we filtered out the age group using the Age variable to include the values of 0,7,16,18. We extracted only Netflix's data to provide the show's summary on Netflix only. This will assist us in identifying the sentimental factors overall. The next step is to focus on the summary variable to identify the wording for each tv show. The analysis shows us the most frequently

used eight emotions in the tv shows summaries on Netflix. Leading us to better understanding why Netflix chooses shows on emotions on their platform.

Results:

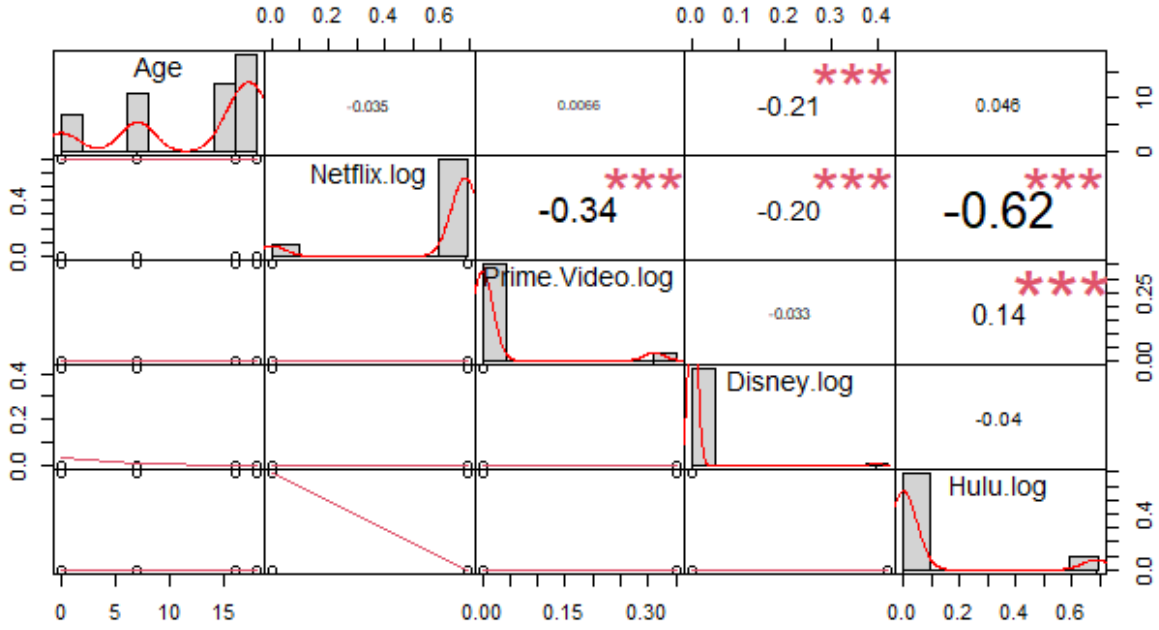
Research Question 1 Results:

When filtering out TV shows that are older than 12 years, this provided us with a smaller list of TV shows. Giving us a count of 986. Using our next step to figure out what TV shows targeted the age groups of 18 and older, it provided us with a smaller list of TV shows. Now giving us a total count of 290. The Rotten Tomato value for the filter was greater or equal to 80 out of 100 rating. The results showed us that after this step, it only included 48 shows on the four main streaming services. Some of the shows were on multiple streaming services. This gave us the answer to our first research question because Netflix had way more shows that fit our research purposes. We summed up how many TV shows were on each streaming services. Netflix has 40 shows, Hulu has 8 shows, Prime Video has 10 shows. Disney + has 0 shows. This helped us gain a better knowledge of industry. We found that Netflix usage and awareness is focused on young adults more than the other 3 streaming services. This would make sense because everyone in our group is over the age of 18, and we all use Netflix the most to watch TV shows. These results also tell us that Netflix is up to date with adding more of the popular and newer shows. It also makes sense that Disney + has 0 because you always hear about the cartoons that they have. Anyone is able to watch their shows because it is made for kids.

Research Question 2 Results:

We filtered out the Netflix variable to only include the value of 1 for all the shows. This allowed us to see every TV show on Netflix. This left us with 940 shows as the result of this filter. The Second filter applied was the Age variable to only include the 0, 7, 16, 18, values. We took out the blank value because there was a missing value. After this filter was applied, this gave us 718 shows as a result. Using (R SQL) we averaged out the total and came out to 12.50 Age group overall for Netflix . Out of total data, about 277 were missing about the age variable of TV viewers and among the remaining information, the result showed that Netflix is emphasizing only 18 years or older aged groups instead of other age group.Correlation of Age and some other variables are also analysed with the help of chart correlation analysis. Besides this, Age and other variables for fitness test is analysed on the regression model. The detailed result are in below;

The chart correlation function of the Performance analytics package is a shortcut to create a correlation plot in R with histograms, density functions, smoothed regression lines and correlation coefficients with the corresponding significance levels (if no stars, the variable is not statistically significant, while one, two and three stars mean that the corresponding variable is significant at 10%, 5% and 1% levels, respectively) with a single line of code: The results show negative relationship at 1% level among Age, Netflix and Prime video.



Regression fitness model relating overall satisfaction with all four features, such as Disney, Prime video and Hulu and Age. According to regression model analysis, it shows some level of significant on Age at 0.05 level and other predictors at 0 level. The base model includes intercept (0.70) and the adjusted R-Squared is 51. 54%. This is a better measure, as this accounts for multiple Independent variables. It implies that the model can explain, with help of R-Squared value, 51.1% of the change independent variable.

```
> m2<-lm(Netflix.log~ Hulu + Age + Prime.Video + Disney, data=Streaming)
> summary(m2)

Call:
lm(formula = Netflix.log ~ Hulu + Age + Prime.Video + Disney,
    data = Streaming)

Residuals:
     Min       1Q   Median       3Q      Max
-0.51995 -0.00149  0.01753  0.02176  0.62830

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.709439   0.012939  54.828   <2e-16 ***
Hulu        -0.417055   0.016648 -25.051   <2e-16 ***
Age         -0.002114   0.000893  -2.367   0.0182 *
Prime.Video -0.447928   0.048767  -9.185   <2e-16 ***
Disney      -0.935736   0.089932 -10.405   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1654 on 824 degrees of freedom
  (232 observations deleted due to missingness)
Multiple R-squared:  0.5154,    Adjusted R-squared:  0.513
F-statistic: 219.1 on 4 and 824 DF,  p-value: < 2.2e-16
```

Research Question 3 Results:

For our third research question, we will use the results from research question 1 and apply the genre filter to see what Genres are popular on each streaming service. This will give us insight as to the top genre on which each streaming service is focused on. This will determine what the industry focuses on to get the most ratings for each streaming platform. The results of the analysis are the following popular genres for the different platforms: Netflix are Comedy and Drama, Hulu and Prime Video have the same results which are Crime, Drama and Mystery and Disney Plus is no genre since the shows are geared toward children. Surprisingly, there are multiple popular genres on each streaming service. It seems the drama genre is the winner with having three out of the four popular genres on the four streaming services that the data is provided. The code to generate the top 5 genres for Prime Video was not able to be generated in R.

NETFLIX WORD CLOUDS

HULU WORD CLOUDS

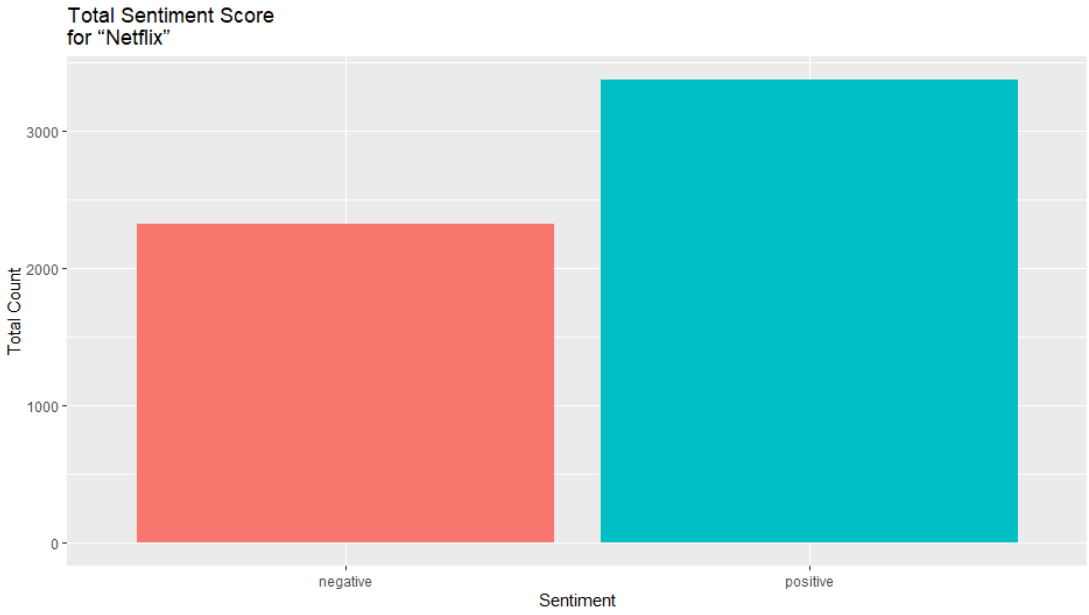
PRIME VIDEO WORD CLOUD

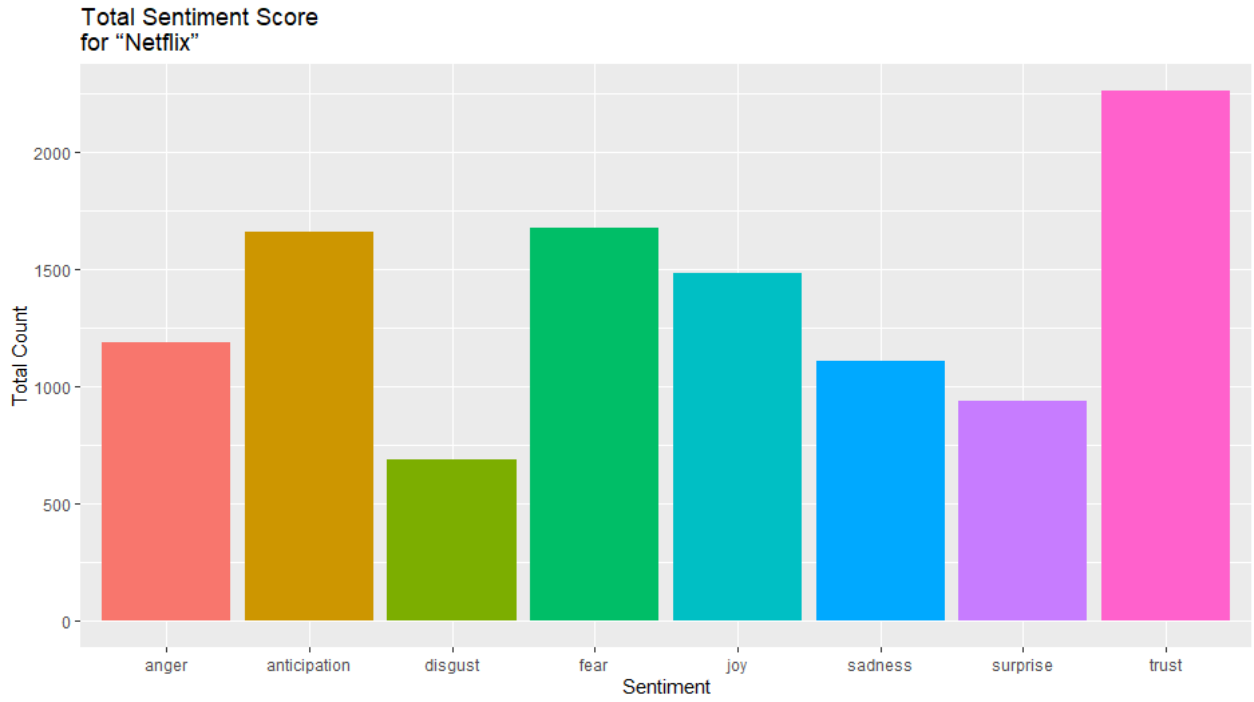
Research Question 4 Results:

For getting our results for research question 4, we converted our data from Excel to a text file, then we ran the sentimental analysis to receive the emotions of the TV shows and determine if they were positive or not.

Anger had a sentimental score of about 1200, anticipation had an score of about 1700, disgust had a score of about 700 being the lowest score, fear had one of about 1700, joy had a sentimental score of 1500, sadness is about 1100, surprise is about 900 and lastly we got trust with the highest score with about 2500. Based on this data we found what emotions are most popular on Netflix. Giving us a better understanding about Netflix. We also ran the Negative and Positive test and we found that most of Netflix shows have a positive sentiment.

Let's put a key graph in the results and interpret briefly.





Managerial Implications

Knowing that Netflix has been the pioneer in the world of streaming platforms and today maintains the leadership that gives it the advantage it has gained over the rest of the services. Netflix streams shows through any internet-connected device that offers the Netflix app, including smart TVs, game consoles, streaming media players, set-top boxes, smartphones, and tablets. You can also watch Netflix on your computer using an internet browser.

The audiovisual industry is in the midst of a revolution, in which it has gone from a distribution paradigm to a consumption paradigm, where decisions are in the hands of the consumer, once the content is digitized, controlling its distribution becomes an extremely complex task where the power and logistics seem to be in the hands of the user. Therefore, some benefits that are expected from our study are knowing what Netflix's competitive advantage is and how they have tried to differentiate themselves by offering more original series and implementing the internationalization of series, since in our database we were able to observe that Netflix has series that were recorded in Asia, Europe and Latin America. Some other benefits that will come are knowing which streaming service to purchase based on an individual's needs. Our data can provide them with what newer TV shows currently have a good Rotten Tomato score with what specific streaming service. Leading them to a TV show that came out within the last 12 years, that is worth the watch.

Those who benefit from this study are consumers as our results will provide us with information on the best Netflix series that young adults can watch during their free time. The evolution of audiences/consumers requires an evolution in the measurement, approximation and study of them. To the extent that traditional screens no longer have a predominant role in people's lives. The flexibility of the processes in which the industry generates content impacts the way in which the relationship with the public that consumes this content is carried out. In addition, in a multimedia environment, the public is a consumer of content, a user of technology. This polysemic condition of audiences calls for the need to find different ways of understanding the consumption of a permanently connected public that obtains satisfactions other than the communication products it seeks, which is why our study also benefits Netflix so that they can see what the consumer trends are. The other 3 streaming services can also benefit because they will see what is working for Netflix, and how they can improve for that target audience.

Conclusion:

Streaming shows such as Netflix, Hulu, Prime Video, Disney + , in general, change in pattern or popularity of streaming and specific targeted age groups will be analyzed. The study aims to understand the usage and awareness of different streaming services among young adults aged above 18 and analyze various factors associated with popularity of the TV streaming among the certain age group and its association with genre or review in overall from the available dataset. According to our results, we found that Netflix usage and awareness is focused on young adults more than the other 3 streaming services. Netflix also has more of the newer and popular shows based on the “start year” and “Rotten Tomato” scores. It concludes that the association among variables are negative at 1% level among Age, Netflix and Prime video. Regression fitness shows some level of significant on Age at 0.05 level and other predictors significant at 0.000 level, the lowest level of significant. It tells us that we would reject the Null hypothesis. The fitness of regression indicated that the model can explain 51% (R-Squre) of the change for multiple Independent predictors . Overall, the result showed that Netflix is emphasizing mainly on the targeted age group of 18. We also analyzed that Netflix likes to focus on TV shows that are in the comedy and drama genre category. When looking at the emotions, trust, fear, and anticipation is the most used. The total sentiment score was mainly positive. The limitation of the study was the availability of ‘Age’ variable in the dataset was only age group responses like 7,

12 and 18 years and older, whereas we did not find specific years of adults or older age group more than 18 on the available data. Because of the limitation, we could not compare the consumers of different age groups above 18. Another limitation is that we don't have data on more streaming services such as, Apple TV, HBO Max, Sling, YouTube TV. We would suggest future research to have more data on the different targeted age groups. This would allow us to get a better understanding of who Netflix is mainly trying to target. An individual that is 60, compared to an 18 year old will most of the time want to watch different things. In future, researchers need to look at what the difference is in TV shows among adults.

References:

Bhatia, R. (2021, August 2). *TV shows on Netflix, Prime Video, hulu and disney+*. Kaggle. Retrieved December 5, 2022, from https://www.kaggle.com/datasets/ruchi798/tv-shows-on-netflix-prime-video-hulu-and-disney

Bhawal, S. (2021, June 12). *Netflix movie and TV shows (June 2021)*. Kaggle. Retrieved December 5, 2022, from https://www.kaggle.com/datasets/snehaanbhawal/netflix-tv-shows-and-movie-list

Appendix:

i. Summary of Data Set

Codes:

```
setwd("C:/Users/jnkha/OneDrive/Desktop/445 Final/group project 445")
Streaming=read.csv("Streaming shows.csv")
nf<-read.csv("Streaming shows.csv")
```

Final summary of the data after cleaning:

```
str(nf)

summary(nf)
```

```
summary(nf)
   Title             startYear      summary            genres
Length:1061        Min.   :1966   Length:1061        Length:1061
Class :character   1st Qu.:2016   Class :character   Class :character
Mode  :character   Median :2018   Mode  :character   Mode  :character
                   Mean   :2016
                   3rd Qu.:2019
                   Max.   :2021

      Age         Rotten.Tomatoes       Netflix           Hulu         Prime.Video
Min.   : 0.00   Length:1061        Min.   :0.000   Min.   :0.000   Min.   :0.000
1st Qu.: 7.00   Class :character   1st Qu.:1.000   1st Qu.:0.000   1st Qu.:0.000
Median :16.00   Mode  :character   Median :1.000   Median :0.000   Median :0.000
Mean   :12.48                      Mean   :0.885   Mean   :0.114   Mean   :0.082
3rd Qu.:18.00                      3rd Qu.:1.000   3rd Qu.:0.000   3rd Qu.:0.000
Max.   :18.00                      Max.   :1.000   Max.   :1.000   Max.   :1.000
NA's   :232
   Disney.
Min.   :0.00000
1st Qu.:0.00000
Median :0.00000
Mean   :0.01225
3rd Qu.:0.00000
Max.   :1.00000
```

Summary before cleaning data.

```
> summary(Netflix.dta)
      X.1            Title             imdb_id          popular_rank       certificate
 Min.   :   1.0   Length:705         Length:705         Length:705         Length:705
 1st Qu.: 288.0   Class :character   Class :character   Class :character   Class :character
 Median : 583.0   Mode  :character   Mode  :character   Mode  :character   Mode  :character
 Mean   : 599.7
 3rd Qu.: 904.0
 Max.   :1244.0

   startYear       endYear        episodes        runtime             type
 Min.   :1966   Min.   :1969   Min.   :  1.00   Length:705         Length:705
 1st Qu.:2015   1st Qu.:2017   1st Qu.: 10.00   Class :character   Class :character
 Median :2018   Median :2018   Median : 20.00   Mode  :character   Mode  :character
 Mean   :2016   Mean   :2017   Mean   : 41.99
 3rd Qu.:2019   3rd Qu.:2020   3rd Qu.: 45.25
 Max.   :2021   Max.   :2022   Max.   :381.00
 NA's   :1      NA's   :355    NA's   :1
 orign_country       language            plot              rating         numVotes
 Length:705         Length:705         Length:705         Min.   :2.500   Min.   :     39
 Class :character   Class :character   Class :character   1st Qu.:6.700   1st Qu.:   1476
 Mode  :character   Mode  :character   Mode  :character   Median :7.400   Median :   5505
                                                          Mean   :7.233   Mean   :  40761
                                                          3rd Qu.:8.000   3rd Qu.:  25782
                                                          Max.   :9.400   Max.   :1523446
                                                          NA's   :6       NA's   :6
    genres             isAdult      cast                X                ID
 Length:705         Min.   :0   Length:705         Min.   :   0.0   Min.   :   1
 Class :character   1st Qu.:0   Class :character   1st Qu.: 288.0   1st Qu.: 290
 Mode  :character   Median :0   Mode  :character   Median : 757.0   Median : 760
                    Mean   :0                      Mean   : 947.5   Mean   : 961
                    3rd Qu.:0                      3rd Qu.:1383.0   3rd Qu.:1387
                    Max.   :0                      Max.   :5124.0   Max.   :5456

      Age             IMDb           Rotten.Tomatoes       Netflix           Hulu
 Min.   : 7.00   Length:705         Length:705         Min.   :0.000   Min.   :0.000
 1st Qu.: 7.00   Class :character   Class :character   1st Qu.:1.000   1st Qu.:0.000
 Median :16.00   Mode  :character   Mode  :character   Median :1.000   Median :0.000
 Mean   :14.55                                         Mean   :0.861   Mean   :0.156
```

Research Question 1

```
Netflix<-read.csv("Streamingshows.csv")
head(Streaming.Shows)
Netflix=Netflix[Netflix$Rotten.Tomatoes>="80/100",]
Netflix=Netflix[Netflix$Age=="18",]
Netflix=Netflix[Netflix$startYear>="2010",]
head(Netflix)
sum(Netflix$Netflix >= 1)
sum(Netflix$Hulu >= 1)
sum(Netflix$Prime.Video >= 1)
sum(Netflix$Disney >= 1)
write.csv(Netflix, "Filteredobjective.csv")
```

Netflix = 40
Hulu = 8
Prime Video = 10
Disney + = 0

Research Question 2

```
setwd("C:/Users/jnkha/OneDrive/Desktop/445 Final/group project 445")
Streaming=read.csv("Streaming shows.csv")
str(Streaming)
```

For mode:

```
Netflixtv=Streaming[Streaming$Netflix=="1",]
Netflixtv=Netflix[Netflix$Age>="0",]
na.omit(Netflixtv$Age)
summary(Netflixtv$Age)
mode=function(){
  return(sort(-table(Netflixtv$Age))[1])
}
mode
```

Scatterplot:

Scatter plot (Age and Netflic)

```
scatterplotMatrix(~Age+Netflix, data=Streaming, diagonal="histogram")
```

str(Netflixtv)

install.packages("corrgram")

library(corrgram)

corrgram(Streaming)

cor.test(Streaming $Age,Streaming$Hulu, conf.level=0.95)

cor.test(Streaming$Age,Streaming$Netflix,conf.level=0.95)

cor.test(Streaming $Age,Streaming$Prime.Video, conf.level=0.95)

cor.test(Streaming $Age,Streaming$Disney, conf.level=0.95)

Regression model:

hist(Streaming$Netflix)

m2<-lm(Netflix.log~ Hulu + Age + Prime.Video + Disney, data=Streaming)

summary(m2)

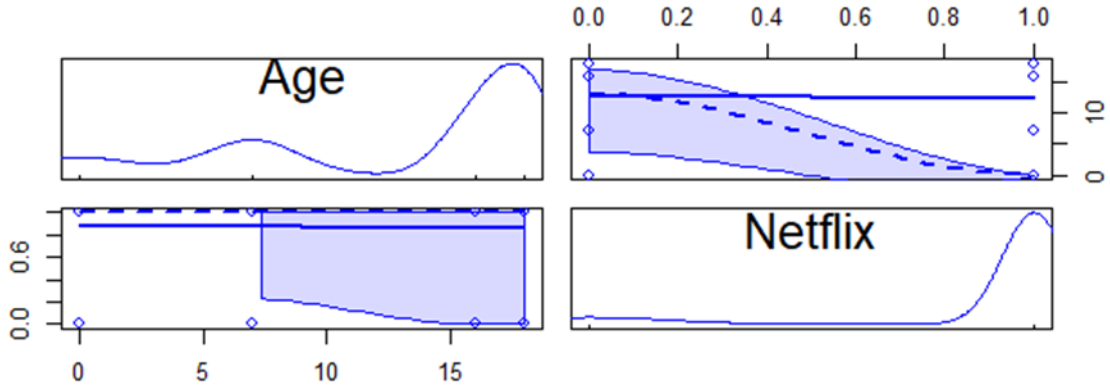
```
Streaming=read.csv("Streaming shows.csv")
Netflix=Streaming[Streaming$Netflix=="1",]
Netflix=Netflix[Netflix$Age>="0",]
na.omit(Netflix$Age)
mode(na.omit(Netflix$Age))
summary(Netflix$Age)
mode=function(){
  return(sort(-table(Netflix$Age))[1])
}
mode()
```

```
> cor.test(Streaming $Age,Streaming$Disney, conf.level=0.95)

        Pearson's product-moment correlation

data:  Streaming$Age and Streaming$Disney
t = -6.1267, df = 827, p-value = 1.387e-09
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2725913 -0.1422970
sample estimates:
       cor
-0.2083684
```

```
> cor.test(Streaming $Age,Streaming$Prime.Video, conf.level=0.95)

        Pearson's product-moment correlation

data:  Streaming$Age and Streaming$Prime.Video
t = 0.19033, df = 827, p-value = 0.8491
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.06149962  0.07467514
sample estimates:
        cor
0.006618443
```

```
> cor.test(Streaming $Age,Streaming$Hulu, conf.level=0.95)

        Pearson's product-moment correlation

data:  Streaming$Age and Streaming$Hulu
t = 1.3273, df = 827, p-value = 0.1848
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.02205308  0.11383946
sample estimates:
      cor
0.0461065
```

```
> cor.test(Streaming$Age,Streaming$Netflix,conf.level=0.95)

        Pearson's product-moment correlation

data:  Streaming$Age and Streaming$Netflix
t = -0.99982, df = 827, p-value = 0.3177
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1025939  0.0334231
sample estimates:
        cor
-0.03474633
```



```
> m2<-lm(Netflix.log~ Hulu + Age + Prime.Video + Disney, data=Streaming)
> summary(m2)

Call:
lm(formula = Netflix.log ~ Hulu + Age + Prime.Video + Disney,
    data = Streaming)

Residuals:
     Min       1Q   Median       3Q      Max
-0.51995 -0.00149  0.01753  0.02176  0.62830

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.709439   0.012939  54.828   <2e-16 ***
Hulu        -0.417055   0.016648 -25.051   <2e-16 ***
Age         -0.002114   0.000893  -2.367   0.0182 *
Prime.Video -0.447928   0.048767  -9.185   <2e-16 ***
Disney      -0.935736   0.089932 -10.405   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1654 on 824 degrees of freedom
  (232 observations deleted due to missingness)
Multiple R-squared:  0.5154,    Adjusted R-squared:  0.513
F-statistic: 219.1 on 4 and 824 DF,  p-value: < 2.2e-16
```

```
setwd("C:/Users/jnkha/OneDrive/Desktop/445 Final/group project 445")
Streaming=read.csv("Streaming shows.csv")
str(Streaming)

Netflixtv=Streaming[Streaming$Netflix=="1",]
Netflixtv=Netflix[Netflix$Age>="0",]
na.omit(Netflixtv$Age)

summary(Netflixtv$Age)
mode=function(){
  return(sort(-table(Netflixtv$Age))[1])
}
mode
```

Correlation and Regression

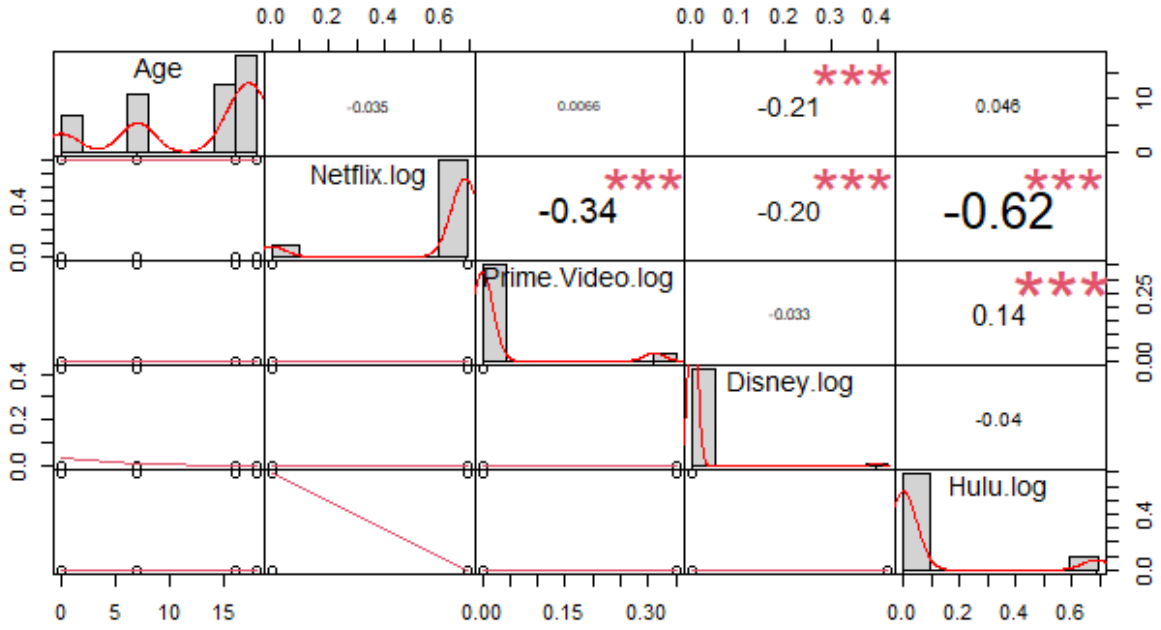
```
str(Netflixtv)
install.packages("corrgram")
library(corrgram)
corrgram(Streaming)

cor.test(Netflixtv$Age,Netflixtv$Hulu)
```

```
cor.test(Netflixtv$Age,Netflixtv$Hulu, conf.level=0.90)

cor.test(Netflixtv$Age,Netflixtv$Prime.Video,conf.level=0.90)

cor.test(Streaming$Age,Streaming$Disney,conf.level=0.90)

cor.test(Netflixtv$Age,Netflixtv$Disney,conf.level=0.90)
##worked
cor.test(Streaming$Age,Streaming$Netflix,conf.level=0.90)


##
library(car)
scatterplotMatrix(~Age+Netflix, data=Streaming, diagonal="histogram")



Streaming$Netflix.log<-log(Streaming$Netflix+1)
Streaming$Hulu.log<-log(Streaming$Hulu+1)
Streaming$Prime.Video.log<-log(Streaming$Prime.Video+1)
Streaming$Disney.log<-log(Streaming$Disney+1)


#AGE
cor.test(Streaming$Age, Streaming$Netflix)
cor.test(Streaming$Age, Streaming$Hulu)
cor.test(Streaming$Age, Streaming$Prime.Video)
cor.test(Streaming$Age, Streaming$Disney)


model<-lm(Netflix.log~Prime.Video.log, data=Streaming)
summary(model)

model<-lm(Netflix.log~Hulu + Age + Prime.Video + Disney, data=Streaming)
summary(model)


cor.test(Streaming$Netflix.log, Streaming$Hulu.log,Streaming$Prime.Video.log,
Streaming$Disney.log)

hist(Streaming$Netflix)
m2<-lm(overall~Netflix+ Hulu + Age + Prime.Video + Disney, data=Streaming)
summary(m2)

hist(Streaming$Netflix)
```

```
m2<-lm(Netflix.log~ Hulu + Age + Prime.Video + Disney, data=Streaming)
summary(m2)
```

Correlation:

```
chart.Correlation(Streaming[,c("Age","Netflix.log","Prime.Video.log",
                               "Disney.log","Hulu.log")]
                  , histogram=TRUE, method="pearson")
```



Research Question 3

```
Netflix=Streaming[Streaming$Netflix=="1",]
Netflix=Netflix[Netflix$Rotten.Tomatoes>="80/100",]
Netflix=Netflix[Netflix$Age=="18",]
Netflix$genres
Netflix=data.frame(Netflix)
netflix.corpus=Corpus(VectorSource(Netflix$genres))
netflix.clean=tm_map(netflix.corpus, PlainTextDocument)
netflix.clean<-tm_map(netflix.corpus,tolower)
netflix.clean<-tm_map(netflix.corpus,removeNumbers)
netflix.clean<-tm_map(netflix.corpus,removeWords,stopwords("english"))
netflix.clean<-tm_map(netflix.corpus,removePunctuation)
netflix.clean<-tm_map(netflix.corpus,stripWhitespace)
netflix.clean<-tm_map(netflix.corpus,stemDocument)
wordcloud(netflix.clean,max.words = 100,random.color = TRUE,random.order=FALSE)
wordcloud(words = netflix.clean, min.freq = 1,
          max.words=100, random.order=FALSE, rot.per=0.1,
```

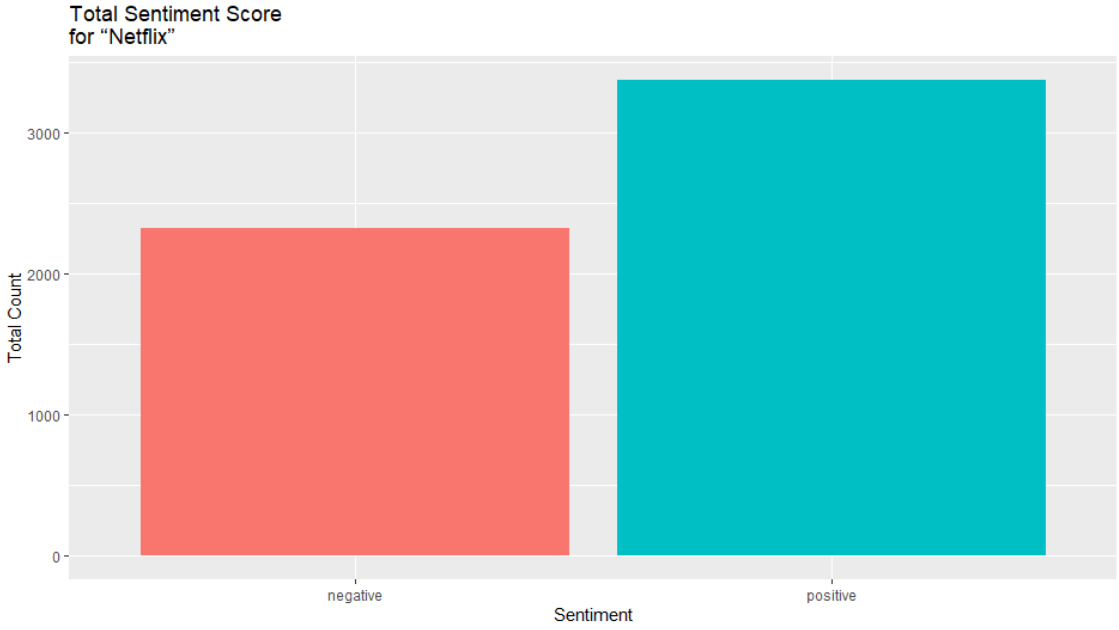
```
        colors=brewer.pal(8, "Dark2"))
mode=function(){
  return(sort(-table(Netflix$genres))[1])
}
mode()

Hulu=Streaming[Streaming$Hulu=="1",]
Hulu=Hulu[Hulu$Rotten.Tomatoes>="80/100",]
Hulu=Hulu[Hulu$Age=="18",]
Hulu$genres
Hulu=data.frame(Hulu)
Hulu.corpus=Corpus(VectorSourceHulu$genres))
Hulu.clean=tm_map(Hulu.corpus, PlainTextDocument)
Hulu.clean<-tm_map(Hulu.corpus,tolower)
Hulu.clean<-tm_map(Hulu.corpus,removeNumbers)
Hulu.clean<-tm_map(Hulu.corpus,removeWords,stopwords("english"))
Hulu.clean<-tm_map(Hulu.corpus,removePunctuation)
Hulu.clean<-tm_map(Hulu.corpus,stripWhitespace)
Hulu.clean<-tm_map(Hulu.corpus,stemDocument)
wordcloud(Hulu.clean,max.words = 100,random.color = TRUE,random.order=FALSE)
wordcloud(words = Hulu.clean, min.freq = 1,
        max.words=100, random.order=FALSE, rot.per=0.1,
        colors=brewer.pal(8, "Dark2"))
mode=function(){
  return(sort(-table(Hulu$genres))[1])
}
mode()

PrimeVideo=Streaming[Streaming$Prime.Video=="1",]
PrimeVideo=PrimeVideo[PrimeVideo$Rotten.Tomatoes>="80/100",]
PrimeVideo=PrimeVideo[PrimeVideo$Age=="18",]
PrimeVideo$genres
PrimeVideo=data.frame(PrimeVideo)
PrimeVideo.corpus=Corpus(VectorSource(PrimeVideo$genres))
PrimeVideo.clean=tm_map(PrimeVideo.corpus, PlainTextDocument)
PrimeVideo.clean<-tm_map(PrimeVideo.corpus,tolower)
PrimeVideo.clean<-tm_map(PrimeVideo.corpus,removeNumbers)
PrimeVideo.clean<-tm_map(PrimeVideo.corpus,removeWords,stopwords("english"))
PrimeVideo.clean<-tm_map(PrimeVideo.corpus,removePunctuation)
PrimeVideo.clean<-tm_map(PrimeVideo.corpus,stripWhitespace)
PrimeVideo.clean<-tm_map(PrimeVideo.corpus,stemDocument)
wordcloud(PrimeVideo.clean,max.words = 3,random.color = TRUE,random.order=FALSE)
        wordcloud(words = PrimeVideo.clean, min.freq = 1,
        max.words=100, random.order=FALSE, rot.per=0.1,
        colors=brewer.pal(8, "Dark2"))
mode=function(){
```

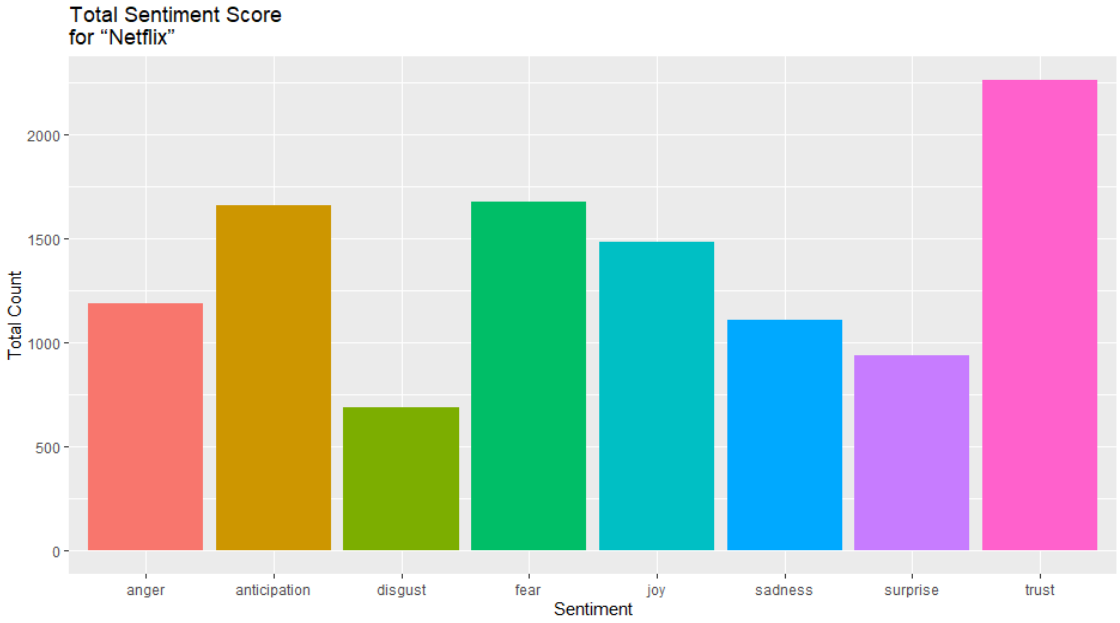
```
  return(sort(-table(PrimeVideo$genres))[1])
}
mode()

Disney=Streaming[Streaming$Disney=="1",]
Disney=Disney[Disney$Rotten.Tomatoes>="80/100",]
Disney=Disney[Disney$Age=="18",]
Disney$genres
Disney=data.frame(Disney)
Disney.corpus=Corpus(VectorSource(Disney$genres))
Disney.clean=tm_map(Disney.corpus, PlainTextDocument)
Disney.clean<-tm_map(Disney.corpus,tolower)
Disney.clean<-tm_map(Disney.corpus,removeNumbers)
Disney.clean<-tm_map(Disney.corpus,removeWords,stopwords("english"))
Disney.clean<-tm_map(Disney.corpus,removePunctuation)
Disney.clean<-tm_map(Disney.corpus,stripWhitespace)
Disney.clean<-tm_map(Disney.corpus,stemDocument)
wordcloud(Disney.clean,max.words = 1,random.color = TRUE,random.order=FALSE)
wordcloud(words = Disney.clean, min.freq = 1,
        max.words=5, random.order=FALSE, rot.per=0.1,
        colors=brewer.pal(8, "Dark2"))
mode=function(){
  return(sort(-table(Disney$genres))[1])
}
mode()
```

Research Question 4

```
text = readLines("Netflix.txt")


library(syuzhet)
library(lubridate)
library(ggplot2)
library(scales)
library(reshape2)
library(dplyr)

sent.text <- get_nrc_sentiment(text)
head(sent.text)
sent.text

full.sent.text <- cbind(text, sent.text)

sentimentTotals <- data.frame(colSums(full.sent.text[,c(2:9)]))
```

```
names(sentimentTotals) <- "count"
sentimentTotals <- cbind("sentiment" = rownames(sentimentTotals), sentimentTotals)
rownames(sentimentTotals) <- NULL
ggplot(data = sentimentTotals, aes(x = sentiment, y = count)) +
  geom_bar(aes(fill = sentiment), stat = "identity") +
  theme(legend.position = "none") +
  xlab("Sentiment") + ylab("Total Count") + ggtitle("Total Sentiment Score
for “Netflix”")

sentimentNegPos <- data.frame(colSums(full.sent.text[,c(10:11)]))
names(sentimentNegPos) <- "count"
sentimentNegPos <- cbind("sentiment" = rownames(sentimentNegPos), sentimentNegPos)
rownames(sentimentNegPos) <- NULL
ggplot(data = sentimentNegPos, aes(x = sentiment, y = count)) +
  geom_bar(aes(fill = sentiment), stat = "identity") +
  theme(legend.position = "none") +
  xlab("Sentiment") + ylab("Total Count") + ggtitle("Total Sentiment Score
for “Netflix”")
```





Scatterpot:
scatterplotMatrix(~Age+Netflix+Prime.Video+Disney+Hulu, data=Streaming, diagonal="histogram")