Data 698 Final Project

Bryan Persaud

5/18/2021

Abstract

Watching TV used to be you watch shows on TV channels provided by some TV provider. This has seen some change in the modern world with the introduction of streaming services. This project will look to monitor this change by looking to answer these two questions, 1. Are streaming services being used and how much growth have they seen? 2. Can we see if streaming services are having an impact on the TV industry? The methodology will include using different datasets containing information on streaming service usage. The parts of the project that will show this are data exploration, data preparation, data analysis, and model building. The models will be built to show if streaming services rely on highly rated shows, which the random forest model created shows that streaming services do not rely on highly rated shows.

Keywords: Streaming services, TV, Ratings, Subscriptions, Netflix

Contents

Introduction

Streaming services are impacting the way we watch TV. With the start of Netflix viewership of TV shows have started to change and people started to move away from the norm of watching cable TV. Over the years there have been a boom of streaming services that changed the way TV is looked at. My project looks to see the trends in streaming services and how these trends have caused change in the TV industry.

This project will look to answer these two questions:

1. Are streaming services being used and how much growth have they seen?
2. Can we see if streaming services are having an impact on the TV industry?

The rest of the project will include a literature review, an insight to the methodology, the details of the data analysis and model building, the findings of these details and the possible future works to look for from these findings, and finally the appendix and references.

Literature Review

Streaming services have grown over the years and has changed the TV industry. The way we watch TV has changed over the years as we no longer have just one option of watching TV. Before it would be you watch TV through cable or satellite provided by a TV company. But now with there are ways to watch TV over the Internet either through websites or streaming services provided by different companies. This literature review will focus on seeing how much of an impact streaming services has had on the TV industry by looking at the changes and trends caused by streaming service usage.

Johannes H Snyman, & Debora J Gilliard (2019) discuss the history of the TV industry and how the start and rise of streaming services began. TV has evolved over the years from broadcasting to cable television to satellite television and now streaming television using the Internet. In the early 2000s the way TV was being viewed changed by certain services such as iTunes and Amazon Video offering people the ability to purchase TV shows and companies such as YouTube and Netflix that offered streaming services. Someone who wanted to watch their favorite TV show had to wait to watch it live, record the show, or rent/buy a DVD/VHS tape. Even with streaming services people at first preferred this way of watching TV, but over the years people would shift to using streaming services.

In 2007 Netflix came out with a different option to watch TV by shifting their business model from sending DVDs through mail to offering a monthly paid service that offered unlimited viewing on streaming shows. This would pay off as more and more people would switch to streaming services and as a result Netflix would become the largest streaming service to this day. This switch from cable or satellite television to streaming television would be known as chord cutting and this has been on the rise since 2012 [1]. People would see the benefit of streaming shows as it gave them a way to watch shows when they wanted to and be able to watch these shows with little to no commercials interrupting.

The ability to watch your favorite shows with no commercials or advertisements interrupting has caused changed in the TV industry. This change has made it so TV providers have to rethink how they handle advertisements so that people do not resort to cord cutting. One such thing TV providers have done is provide people with video on-demand, where a collection of aired episodes for certain shows can be viewed. This allows people to watch episodes for shows they may have missed or want to watch again, and people have the option to fast forward through the episode. People using this have some way to avoid advertisements, but people using streaming services can do this and still be able to binge-watch shows.

Binge-watching is a term used to describe when someone watches multiple episodes to a show in one sitting, usually it is multiple episodes for a single show. With the rise of streaming service usage there has been an increase in binge-watching, as over 70% of Americans binge-watch [2]. Streaming services make it so you have access to almost all of the episodes for a show and access to all these episodes promotes binge-watching as you

have the option to continue watching. Also adding in having no commercials interrupting an episode makes majority of people pick binge-watching on streaming services rather than TV. Binge-watching is changing the way people view television as you went from having to wait a week between viewing episodes to having multiple episodes ready to view at once.

Methodology

As for the research to be done I will look at different datasets containing information on streaming services. The information that will be looked at include the number of subscriptions, number of TV shows offered, and TV ratings. I will also look to monitor trends in streaming services and what changes are done. I will do some data analysis to show the growth of streaming services and how this growth has caused an increased use of streaming services. I will look to model if streaming services rely on highly rated shows to gain more subscriptions.

Two datasets were taken from Kaggle to do this project, one showing the different shows offered by streaming services and the other one showing the number of subscriptions over the years for Netflix. The project consists of the following parts: data exploration, data preparation, data analysis, and model building. The data exploration helps to show the data we are working with and helps to see the characteristics of each variable in the dataset. Data preparation will be done for any changes to these characteristics and handling any missing data. The data analysis done includes splitting the dataset to see the different shows offered for each streaming service and visualizing these shows based on their ratings. Also a shiny app was done to help show the dataset with the number of subscriptions for Netflix. Two models were built, one a multiple linear regression model and the one model created using random forest. These models will be compared to see which one better shows if a streaming service relies on shows with high ratings.

Experimentation and Results

Data Exploration

```
##     X                           Title Year Age IMDb Rotten.Tomatoes Netflix
## 1   0                     Breaking Bad 2008 18+  9.5             96%       1
## 2   1                   Stranger Things 2016 16+  8.8             93%       1
## 3   2                       Money Heist 2017 18+  8.4             91%       1
## 4   3                         Sherlock 2010 16+  9.1             78%       1
## 5   4                 Better Call Saul 2015 18+  8.7             97%       1
## 6   5                       The Office 2005 16+  8.9             81%       1
## 7   6                     Black Mirror 2011 18+  8.8             83%       1
## 8   7                     Supernatural 2005 16+  8.4             93%       1
## 9   8                     Peaky Blinders 2013 18+  8.8             92%       1
## 10  9       Avatar: The Last Airbender 2005  7+  9.2            100%       1
## 11 10                 The Walking Dead 2010 18+  8.2             81%       1
## 12 11                             Dark 2017 16+  8.7             94%       1
## 13 12                            Ozark 2017 18+  8.4             81%       1
## 14 13                 Attack on Titan 2013 16+  8.8             94%       1
## 15 14                           Narcos 2015 18+  8.8             89%       1
## 16 15 Fullmetal Alchemist: Brotherhood 2009 18+  9.1            100%       1
## 17 16                       Community 2009  7+  8.5             88%       1
## 18 17                      Mindhunter 2017 18+  8.6             96%       1
## 19 18            Parks and Recreation 2009 16+  8.6             93%       1
## 20 19                          Dexter 2006 18+  8.6             72%       1
## 21 20              Marvel's Daredevil 2015 18+  8.6             92%       1
```

```
## 22 21                    The Witcher 2019 18+  8.3            67%       1
## 23 22                     Twin Peaks 1990 18+  8.8            89%       1
## 24 23                  One-Punch Man 2015 16+  8.8           100%       1
## 25 24                      Outlander 2014 18+  8.4            91%       1
##    Hulu Prime.Video Disney. type
## 1     0           0       0    1
## 2     0           0       0    1
## 3     0           0       0    1
## 4     0           0       0    1
## 5     0           0       0    1
## 6     0           0       0    1
## 7     0           0       0    1
## 8     0           0       0    1
## 9     0           0       0    1
## 10    0           0       0    1
## 11    0           0       0    1
## 12    0           0       0    1
## 13    0           0       0    1
## 14    1           0       0    1
## 15    0           0       0    1
## 16    1           0       0    1
## 17    1           0       0    1
## 18    0           0       0    1
## 19    1           1       0    1
## 20    0           0       0    1
## 21    0           0       0    1
## 22    0           0       0    1
## 23    1           0       0    1
## 24    1           0       0    1
## 25    0           0       0    1
```

```
##        X           Title                Year          Age
##  Min.   :   0   Length:5611        Min.   :1901   Length:5611
##  1st Qu.:1402   Class :character   1st Qu.:2010   Class :character
##  Median :2805   Mode  :character   Median :2015   Mode  :character
##  Mean   :2805                      Mean   :2011
##  3rd Qu.:4208                      3rd Qu.:2017
##  Max.   :5610                      Max.   :2020
##
##       IMDb       Rotten.Tomatoes       Netflix            Hulu
##  Min.   :1.000   Length:5611        Min.   :0.0000   Min.   :0.0000
##  1st Qu.:6.600   Class :character   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :7.300   Mode  :character   Median :0.0000   Median :0.0000
##  Mean   :7.113                      Mean   :0.3441   Mean   :0.3126
##  3rd Qu.:7.900                      3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :9.600                      Max.   :1.0000   Max.   :1.0000
##  NA's   :1161
##   Prime.Video        Disney.            type
##  Min.   :0.0000   Min.   :0.00000   Min.   :1
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:1
##  Median :0.0000   Median :0.00000   Median :1
##  Mean   :0.3821   Mean   :0.03208   Mean   :1
##  3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:1
##  Max.   :1.0000   Max.   :1.00000   Max.   :1
```

```
##
```

```
##                                Area     Years Subscribers
## 1         United States and Canada Q1 - 2018    60909000
## 2  Europe,  Middle East and Africa Q1 - 2018    29339000
## 3                    Latin America Q1 - 2018    21260000
## 4                     Asia-Pacific Q1 - 2018     7394000
## 5         United States and Canada Q2 - 2018    61870000
## 6  Europe,  Middle East and Africa Q2 - 2018    31317000
## 7                    Latin America Q2 - 2018    22795000
## 8                     Asia-Pacific Q2 - 2018     8372000
## 9         United States and Canada Q3 - 2018    63010000
## 10 Europe,  Middle East and Africa Q3 - 2018    33836000
## 11                   Latin America Q3 - 2018    24115000
## 12                    Asia-Pacific Q3 - 2018     9461000
## 13        United States and Canada Q4 - 2018    64757000
## 14 Europe,  Middle East and Africa Q4 - 2018    37818000
## 15                   Latin America Q4 - 2018    26077000
## 16                    Asia-Pacific Q4 - 2018    10607000
## 17        United States and Canada Q1 - 2019    66633000
## 18 Europe,  Middle East and Africa Q1 - 2019    42542000
## 19                   Latin America Q1 - 2019    27547000
## 20                    Asia-Pacific Q1 - 2019    12141000
## 21        United States and Canada Q2 - 2019    66501000
## 22 Europe,  Middle East and Africa Q2 - 2019    44229000
## 23                   Latin America Q2 - 2019    27890000
## 24                    Asia-Pacific Q2 - 2019    12942000
## 25        United States and Canada Q3 - 2019    67114000
## 26 Europe,  Middle East and Africa Q3 - 2019    47355000
## 27                   Latin America Q3 - 2019    29380000
## 28                    Asia-Pacific Q3 - 2019    14485000
## 29        United States and Canada Q4 - 2019    67662000
## 30 Europe,  Middle East and Africa Q4 - 2019    51778000
## 31                   Latin America Q4 - 2019    31417000
## 32                    Asia-Pacific Q4 - 2019    16233000
## 33        United States and Canada Q1 - 2020    69969000
## 34 Europe,  Middle East and Africa Q1 - 2020    58734000
## 35                   Latin America Q1 - 2020    34318000
## 36                    Asia-Pacific Q1 - 2020    19835000
## 37        United States and Canada Q2 - 2020    72904000
## 38 Europe,  Middle East and Africa Q2 - 2020    61483000
## 39                   Latin America Q2 - 2020    36068000
## 40                    Asia-Pacific Q2 - 2020    22492000
```

```
##      Area              Years            Subscribers
##  Length:40          Length:40          Min.   : 7394000
##  Class :character   Class :character   1st Qu.:22184000
##  Mode  :character   Mode  :character   Median :32626500
##                                        Mean   :37864725
##                                        3rd Qu.:61052500
##                                        Max.   :72904000
```

The first dataset has each observation as a different TV show, there are 5611 different observation. For each of them you can see the year the show came out, the IMDb and Rotten Tomatoes rating, and whether the

show is in the specified streaming service. A value of 1 means that the show is included in the streaming service and a value of 0 means that the show is not in the streaming service.

The second dataset has each observation as a certain area for each quarter from the year 2018 to the first half of 2020. There are 40 observations in total with 4 different areas, United States and Canada, Europe, Middle East, and Africa, Latin America, and Asia-Pacific. The number of subscribers are shown for each area by the quarter of that year.

Data Preperation

We can see above that some columns for the first dataset are not accurately as each streaming service should be a categorical factor. Let us transform the data to change the type of Netflix, Hulu, Prime.Video, and Disney. Also for the IMDb column there are NAs, these won't be removed just yet so we can get an accurate number of the number of shows offered by each streaming service.

```
##        X           Title                Year          Age
##  Min.   :   0   Length:5611        Min.   :1901   Length:5611
##  1st Qu.:1402   Class :character   1st Qu.:2010   Class :character
##  Median :2805   Mode  :character   Median :2015   Mode  :character
##  Mean   :2805                      Mean   :2011
##  3rd Qu.:4208                      3rd Qu.:2017
##  Max.   :5610                      Max.   :2020
##
##       IMDb       Rotten.Tomatoes    Netflix  Hulu     Prime.Video Disney.
##  Min.   :1.000   Length:5611        0:3680   0:3857   0:3467      0:5431
##  1st Qu.:6.600   Class :character   1:1931   1:1754   1:2144      1: 180
##  Median :7.300   Mode  :character
##  Mean   :7.113
##  3rd Qu.:7.900
##  Max.   :9.600
##  NA's   :1161
##       type
##  Min.   :1
##  1st Qu.:1
##  Median :1
##  Mean   :1
##  3rd Qu.:1
##  Max.   :1
##
```

Now we can see how many shows each streaming service has and how many shows they do not have. Amazon Prime Video has the most with 2144 and Disney+ has the least with 180. However, for each streaming service you can see that there are more shows not included then there are included.

Data Analysis

Some data analysis will be done to see the shows that are offered for each streaming service and the ratings for each show. There are two different ratings, IMDb and Rotten Tomatoes. The IMDb rating for a show is a score from 1 to 10 and the Rotten Tomatoes rating is a percentage from 0 to 100. I will be using the IMDb rating as the data for this rating is more reliable to use.

```
##                              Title Year Age IMDb Rotten.Tomatoes Netflix
```

```
## 1                      Breaking Bad 2008 18+  9.5             96%       1
## 2                        Our Planet 2019  7+  9.3             93%       1
## 3                           Ramayan 1987 all  9.3                       1
## 4        Avatar: The Last Airbender 2005  7+  9.2            100%       1
## 5                    Yeh Meri Family 2018      9.2                       1
## 6                          Sherlock 2010 16+  9.1             78%       1
## 7  Fullmetal Alchemist: Brotherhood 2009 18+  9.1            100%       1
## 8                    The Vietnam War 2017 18+  9.1             98%       1
## 9                 The Twilight Zone 1959  7+  9.0             82%       1
## 10                       Death Note 2006 18+  9.0                       1
```

```
##     Title                Year          Age                IMDb
##  Length:1931        Min.   :1914   Length:1931        Min.   :1.000
##  Class :character   1st Qu.:2013   Class :character   1st Qu.:6.600
##  Mode  :character   Median :2016   Mode  :character   Median :7.400
##                     Mean   :2014                      Mean   :7.163
##                     3rd Qu.:2018                      3rd Qu.:8.000
##                     Max.   :2020                      Max.   :9.500
##                                                       NA's   :120
##  Rotten.Tomatoes       Netflix
##  Length:1931        Min.   :1
##  Class :character   1st Qu.:1
##  Mode  :character   Median :1
##                     Mean   :1
##                     3rd Qu.:1
##                     Max.   :1
##
```

```
##                                       Title Year Age IMDb Rotten.Tomatoes
## 1                                   Destiny 2014      9.6
## 2                                Hungry Henry 2014      9.5
## 3                          The Joy of Painting 1983 all  9.4
## 4                               Rick and Morty 2013 18+  9.2             94%
## 5              Fullmetal Alchemist: Brotherhood 2009 18+  9.1            100%
## 6  Leah Remini: Scientology and the Aftermath 2016 16+  9.1
## 7                             The Twilight Zone 1959  7+  9.0             82%
## 8                                   Death Note 2006 18+  9.0
## 9                                      Firefly 2002 16+  9.0             85%
## 10                       How the Universe Works 2010  7+  9.0
##    Hulu
## 1     1
## 2     1
## 3     1
## 4     1
## 5     1
## 6     1
## 7     1
## 8     1
## 9     1
## 10    1
```

```
##     Title                Year          Age                IMDb
##  Length:1754        Min.   :1931   Length:1754        Min.   :1.700
```

```
##  Class :character   1st Qu.:2007   Class :character   1st Qu.:6.600
##  Mode  :character   Median :2013   Mode  :character   Median :7.300
##                     Mean   :2010                      Mean   :7.061
##                     3rd Qu.:2016                      3rd Qu.:7.900
##                     Max.   :2020                      Max.   :9.600
##                                                       NA's   :237
##  Rotten.Tomatoes         Hulu
##  Length:1754        Min.   :1
##  Class :character   1st Qu.:1
##  Mode  :character   Median :1
##                     Mean   :1
##                     3rd Qu.:1
##                     Max.   :1
##
```

```
##                      Title Year Age IMDb Rotten.Tomatoes Prime.Video
## 1              Malgudi Days 1987 all  9.5                          1
## 2      The Joy of Painting 1983 all  9.4                          1
## 3         Band of Brothers 2001 18+  9.4             94%          1
## 4                 The Wire 2002 18+  9.3             94%          1
## 5           Green Paradise 2011 all  9.3                          1
## 6             The Sopranos 1999 18+  9.2             92%          1
## 7                 Baseball 1994 16+  9.2                          1
## 8                  The Bay 2010      9.2                          1
## 9  Harmony with A R Rahman 2018      9.2                          1
## 10         Everyday Driver 2017      9.2                          1
```

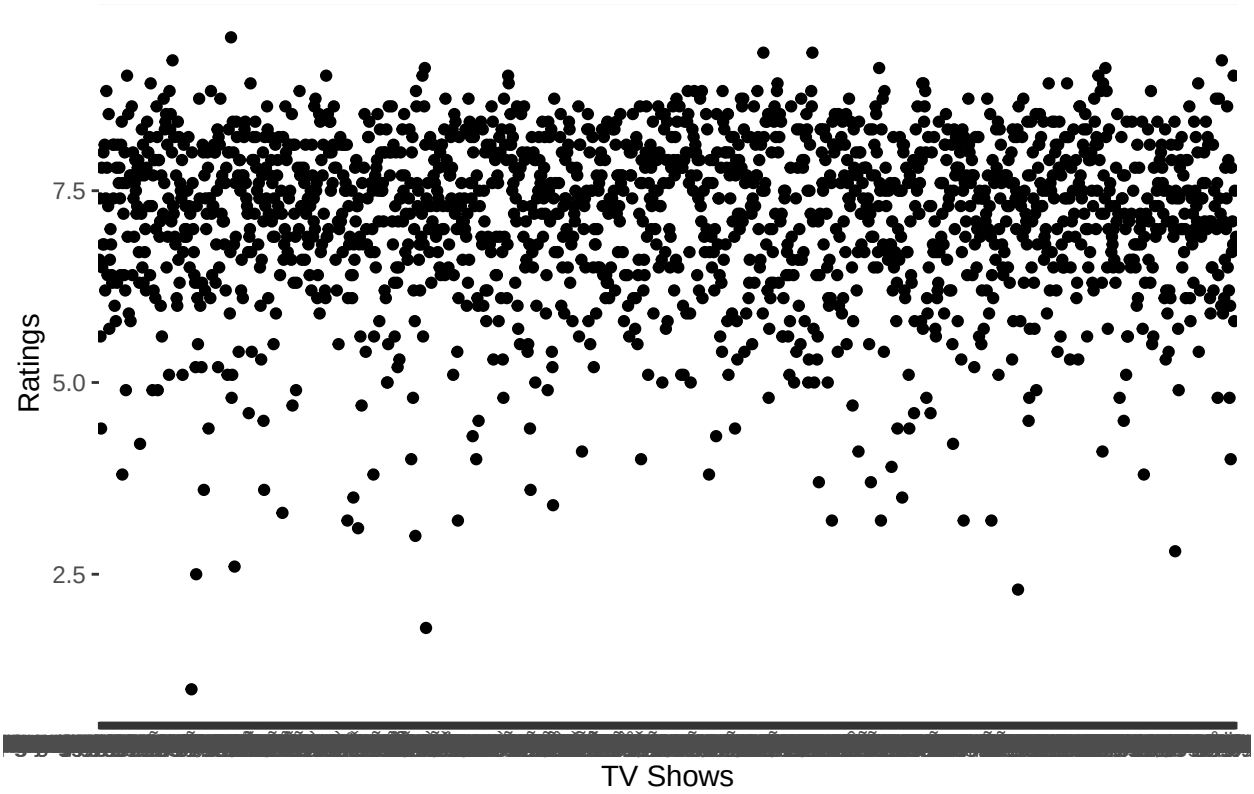
```
##     Title                Year          Age                 IMDb
##  Length:2144        Min.   :1901   Length:2144        Min.   :1.80
##  Class :character   1st Qu.:2007   Class :character   1st Qu.:6.60
##  Mode  :character   Median :2013   Mode  :character   Median :7.40
##                     Mean   :2009                      Mean   :7.18
##                     3rd Qu.:2016                      3rd Qu.:8.00
##                     Max.   :2020                      Max.   :9.50
##                                                       NA's   :837
##  Rotten.Tomatoes     Prime.Video
##  Length:2144        Min.   :1
##  Class :character   1st Qu.:1
##  Mode  :character   Median :1
##                     Mean   :1
##                     3rd Qu.:1
##                     Max.   :1
##
```

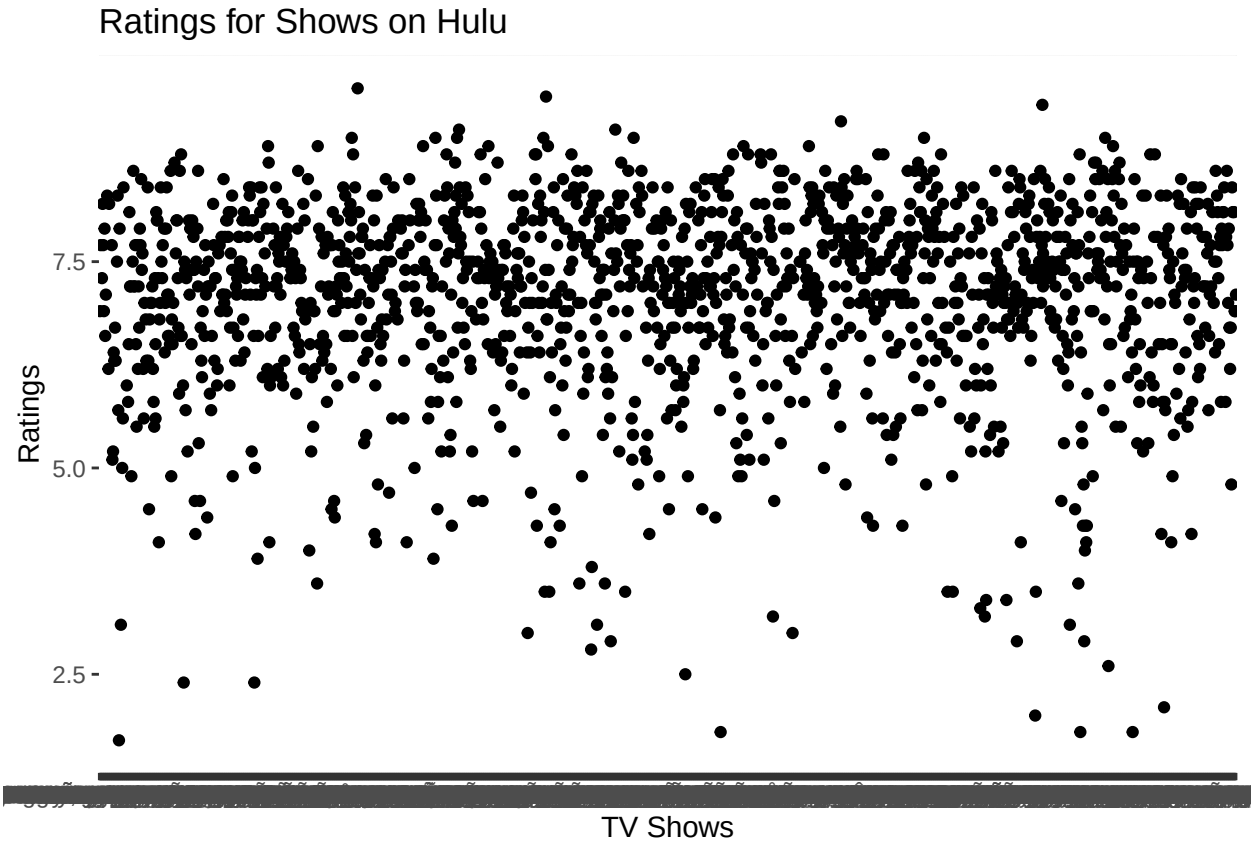
```
##                                    Title Year Age IMDb Rotten.Tomatoes
## 1                  The Imagineering Story 2019  7+  9.1            100%
## 2                           Gravity Falls 2012  7+  8.9            100%
## 3                       One Strange Rock 2018 all  8.8             83%
## 4                           The Simpsons 1989  7+  8.7             85%
## 5                         The Mandalorian 2019  7+  8.7             93%
## 6                  The Incredible Dr. Pol 2011  7+  8.6
## 7                           Prop Culture 2020  7+  8.6
## 8                      Prairie Dog Manor 2019      8.6
```

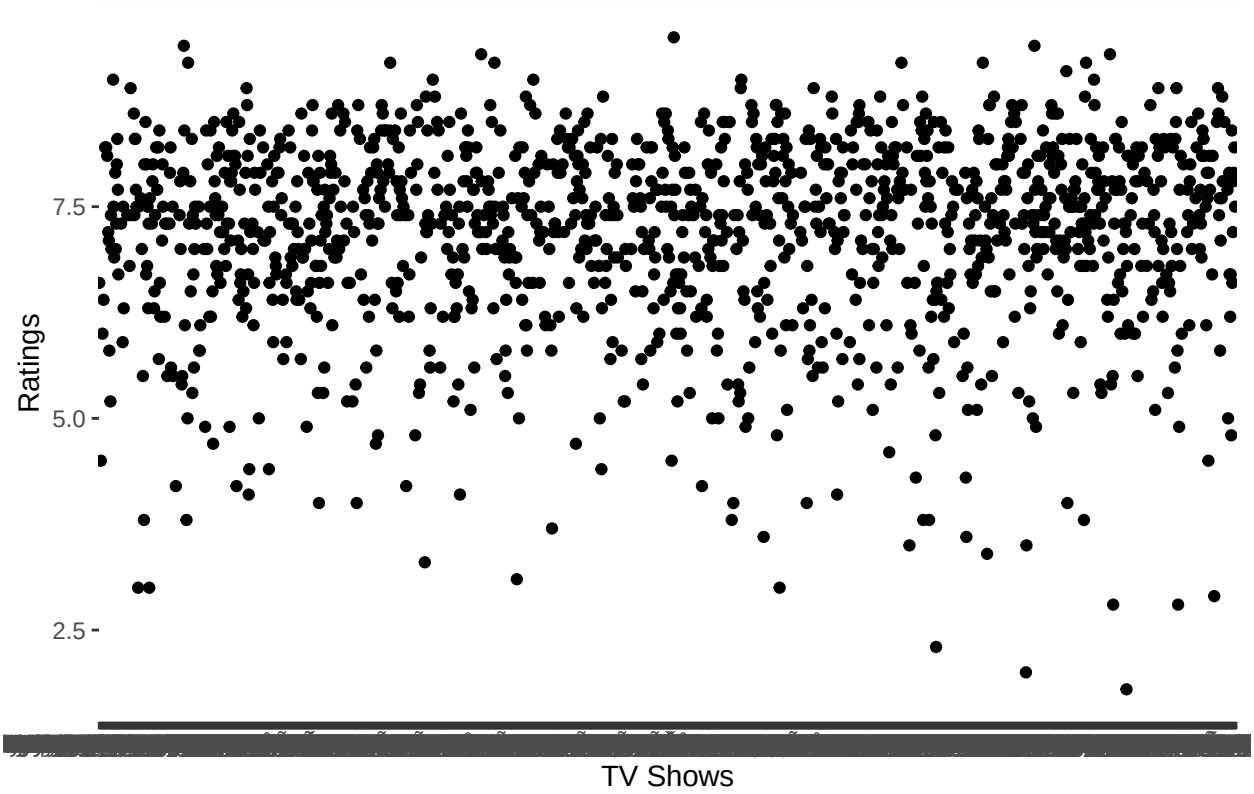
```
## 9  Disney Gallery / Star Wars: The Mandalorian 2020  7+  8.5
## 10                                    So Weird 1999  7+  8.5
##    Disney.
## 1        1
## 2        1
## 3        1
## 4        1
## 5        1
## 6        1
## 7        1
## 8        1
## 9        1
## 10       1
```

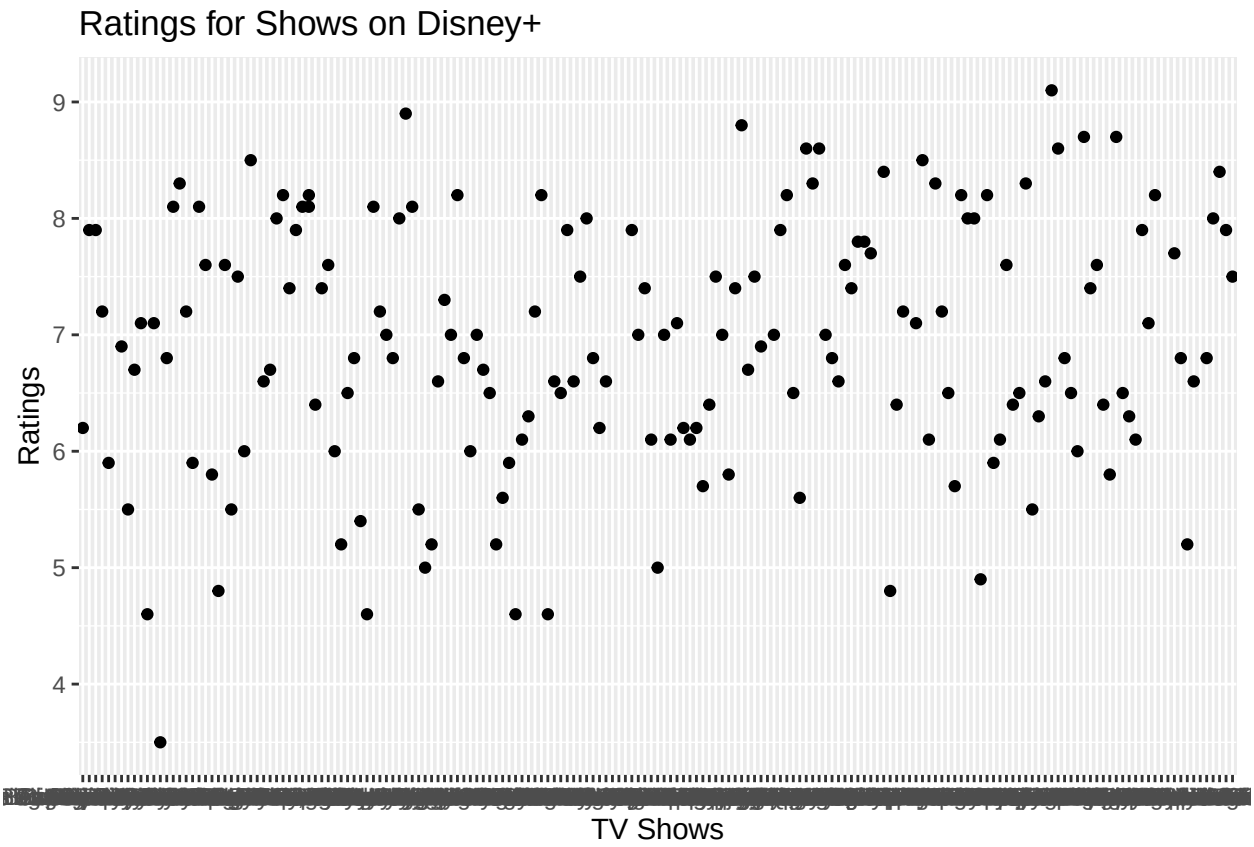
```
##     Title                Year          Age                IMDb
##  Length:180         Min.   :1955   Length:180         Min.   :3.500
##  Class :character   1st Qu.:2006   Class :character   1st Qu.:6.200
##  Mode  :character   Median :2013   Mode  :character   Median :7.000
##                     Mean   :2010                      Mean   :6.924
##                     3rd Qu.:2017                      3rd Qu.:7.900
##                     Max.   :2020                      Max.   :9.100
##                                                       NA's   :11
##  Rotten.Tomatoes       Disney.
##  Length:180         Min.   :1
##  Class :character   1st Qu.:1
##  Mode  :character   Median :1
##                     Mean   :1
##                     3rd Qu.:1
##                     Max.   :1
##
```

From sorting the data we see what shows are available for each streaming services, Netflix, Hulu, Prime video, and Disney+. In addition to seeing which shows are available for each streaming service, the data is also sorted by their IMDb rating. Let us visualize this data to get a better look at these ratings.


Ratings for Shows on Netflix
Ratings
7.5
5.0
2.5
TV Shows


Ratings for Shows on Hulu
Ratings
7.5
5.0
2.5
TV Shows


Ratings for Shows on Amazon Prime Video
Ratings
7.5
5.0
2.5
TV Shows



From the data we can see that Netflix and Amazon Prime Video contain more shows that have higher ratings. Netflix is seen as the biggest streaming service and we can see that by the number of shows it offers and the high ratings for some of these shows. Let us look into the number of people who have subscribed to Netflix over the years. This will give us an insight on how much streaming services have grown.

For the second dataset since the data is split by different regions and shows the number of subscribers from the year 2018 to the first half of 2020 a shiny app was created to help visualize this data. It can be viewed here: https://bpersaud104.shinyapps.io/Netflix_Analysis/

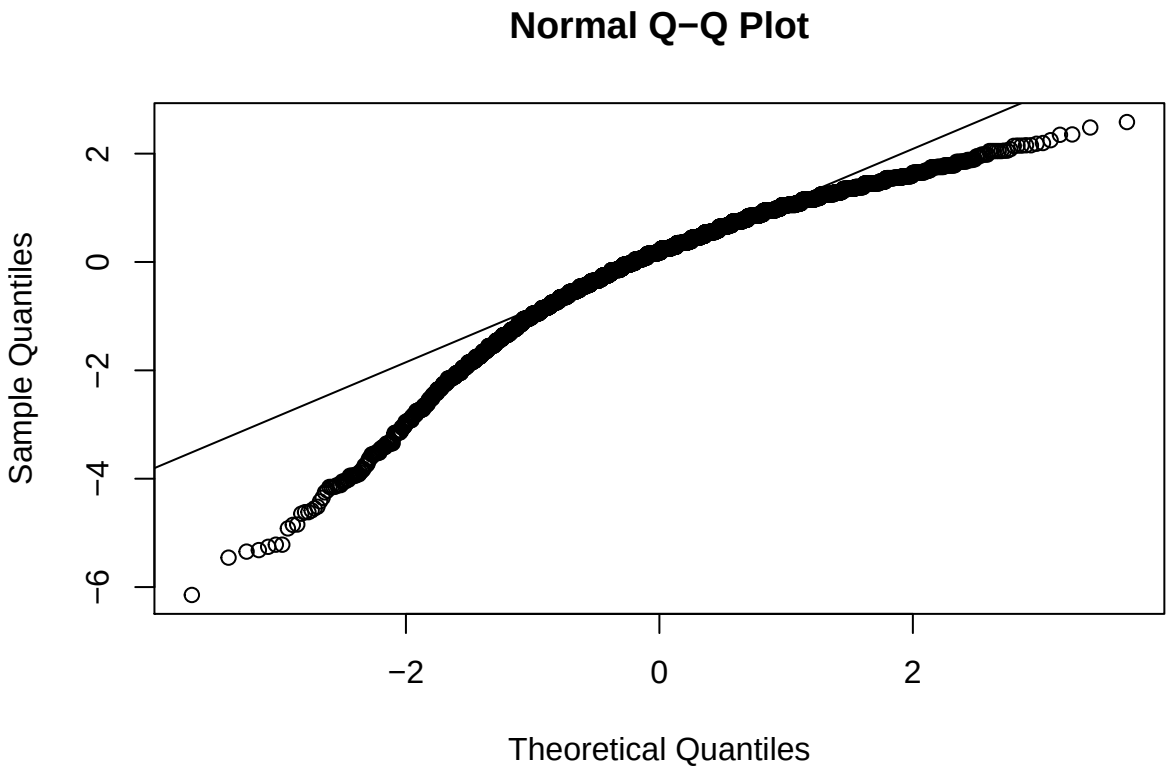
From the shiny app we can see the number of subscriptions for Netflix divided by four areas, United States and Canada, Europe, Middle East and Africa, Latin America, and Asia-Pacific. For all four areas the number of subscribers have increased since 2018. North America and Canada has the most subscribers going from around sixty million in 2018 to around 72 million in the first half of 2020. Europe, Middle East, and Africa has seen the most rise in subscribers going from around 29 million in 2018 to around 61 million in the first half of 2020.

Model Building

The first model to be built is a multiple linear regression model. The model will consist of the IMDb rating and each streaming service.

```
##
## Call:
## lm(formula = IMDb ~ Netflix + Hulu + Prime.Video + Disney., data = shows)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -6.1456 -0.5456  0.1825  0.7825  2.5825
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.912859   0.064577 107.049  < 2e-16 ***
## Netflix      0.232791   0.064223   3.625 0.000292 ***
## Hulu         0.104682   0.061036   1.715 0.086398 .
## Prime.Video  0.239362   0.062738   3.815 0.000138 ***
## Disney.     -0.008604   0.103832  -0.083 0.933961
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.129 on 4445 degrees of freedom
##   (1161 observations deleted due to missingness)
## Multiple R-squared:  0.005946,   Adjusted R-squared:  0.005051
## F-statistic: 6.647 on 4 and 4445 DF,  p-value: 2.493e-05
```



The first model shows a p-value that is very high. The Q-Q-plot shown is skewed and does not follow the line. This means that this model does not do a good job of showing whether a streaming service relies on shows with high ratings.

For the second model it will be built using random forest. Since there are NAs in the IMDB column they will be removed for the random forest model. A train/test split will be also be set up.

```
##
```

```
## Call:
##  randomForest(formula = IMDb ~ Netflix + Hulu + Prime.Video +      Disney., data = train)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 1
##
##           Mean of squared residuals: 1.283645
##                     % Var explained: 0.39
```

A confusion matrix will be made from the random forest model.

```
##         predicted
## observed 6.95951062706041 7.03644469923655 7.05683446482275 7.12964541163825
##        1               33              255                2                 9
##         predicted
## observed 7.14934971767413 7.15013755835843 7.15520181183265 7.18787459701853
##        1              346               27              194                16
##         predicted
## observed 7.2232114043871
##        1               8
```

Here we can see a confusion matrix set up using the model created. It shows that for the train dataset most of the data lies in ratings between 6.9 and 7.2. I would say that this model shows that a streaming service does not rely on shows with high ratings, since the confusion matrix is showing us scores around a 7.

Findings and Conclusion

From the start there were two questions to answer:

1. Are streaming services being used and how much growth have they seen?
2. Can we see if streaming services are having an impact on the TV industry?

From our data exploration, preparation and analysis we can see the growth of streaming services over the years. Going through a dataset containing 5611 different shows offered by different streaming services helps us to see the amount of shows offered by each of them. You can see that more shows are not offered than offered for each streaming service. This is probably because of licensing restrictions and each streaming service offering their own original shows, but more data is needed to confirm this. Also from the shows that are offered there are some with high ratings. From this we can answer the first question, but not so much the second question.

For the first model, the multiple linear regression model, it does not do a good job of showing evidence of a streaming service relying oh highly rated shows. The second model, the random forest model, is the better of the two models because it helps us to see that streaming services do not rely on highly rated shows. Streaming services might not be getting more subscriptions through the use of highly rated shows but from the shiny app we can see that the number of subscriptions are increasing, at least for Netflix anyways. Further investigation will be needed to help show this increase in subscriptions and if this is happening for the other streaming services as well.

This project is a baseline to help show the growth of streaming services and the impact they have had on the TV industry. Streaming services has helped influence movements such as cord cutting and binge-watching through the use of more and more people seeing the value in using streaming services over cable TV. Future

works in this area of study can include comparing this data to TV providers and top TV channels to see if this growth in streaming services has caused a diminish in TV usage. Also there are some TV channels, such as HBO coming out with HBO Max and NBC coming out with Peacock, that can also be looked at to monitor the trends in streaming services and how these streaming services will affect the viewership of their counterpart TV channel. There are also streaming services such as Hulu that offer packages that include live TV, so this is another area that can be used in future works.

Appendix

```
library(dplyr)
library(ggplot2)
library(caTools)
library(randomForest)
```

```
shows <- read.csv("https://raw.githubusercontent.com/bpersaud104/Data698/main/Data%20Collection%20and%2
Netflix_subs <- read.csv("https://raw.githubusercontent.com/bpersaud104/Data698/main/Data%20Collection%
```

```
head(shows, 25)
summary(shows)
head(Netflix_subs, 40)
summary(Netflix_subs)
```

```
# Change variables to factors
streaming_shows <- transform(shows, X = as.integer(X), Title = as.character(Title), Year = as.integer(Y
summary(streaming_shows)
```

```
# Get shows that are on Netflix
Netflix_data <- shows %>%
  select(Title, Year, Age, IMDb, Rotten.Tomatoes, Netflix) %>%
  filter(Netflix == 1) %>%
  arrange(desc(IMDb))
head(Netflix_data, 10)
summary(Netflix_data)

# Get shows that are on Hulu
Hulu_data <- shows %>%
  select(Title, Year, Age, IMDb, Rotten.Tomatoes, Hulu) %>%
  filter(Hulu == 1) %>%
  arrange(desc(IMDb))
head(Hulu_data, 10)
summary(Hulu_data)

# Get shows on Amazon Prime Video
Prime_video_data <- shows %>%
  select(Title, Year, Age, IMDb, Rotten.Tomatoes, Prime.Video) %>%
  filter(Prime.Video == 1) %>%
  arrange(desc(IMDb))
head(Prime_video_data, 10)
summary(Prime_video_data)
```

```r
# Get shows on Disney+
Disney_plus_data <- shows %>%
  select(Title, Year, Age, IMDb, Rotten.Tomatoes, Disney.) %>%
  filter(Disney. == 1) %>%
  arrange(desc(IMDb))
head(Disney_plus_data, 10)
summary(Disney_plus_data)
```

```r
#Plot Netflix shows by ratings
ggplot(Netflix_data, aes(x = Title, y = IMDb)) + geom_point() + xlab("TV Shows") + ylab("Ratings") + gg

# Plot Hulu shows by ratings
ggplot(Hulu_data, aes(x = Title, y = IMDb)) + geom_point() + xlab("TV Shows") + ylab("Ratings") +  ggti

# Plot Amazon Prime Video shows by ratings
ggplot(Prime_video_data, aes(x = Title, y = IMDb)) + geom_point() + xlab("TV Shows") + ylab("Ratings")

# Plot Disney+ shows by ratings
ggplot(Disney_plus_data, aes(x = Title, y = IMDb)) + geom_point() + xlab("TV Shows") + ylab("Ratings")
```

```r
# Multiple linear regression model using IMDb
model1 <- lm(IMDb ~ Netflix + Hulu + Prime.Video + Disney., data = shows)
summary(model1)

# Plot for model
qqnorm(model1$residuals)
qqline(model1$residuals)
```

```r
# Build model using random forest
model2 <- randomForest(IMDb ~ Netflix + Hulu + Prime.Video + Disney., data = train)
model2
```

```r
# Set seed
set.seed(120)
# Get rid of NAs
ratings <- na.omit(streaming_shows)
# Split the data into train and test datasets
sample <- sample.split(ratings$IMDb, SplitRatio = 0.80)
train <- subset(ratings, sample == TRUE)
test <- subset(ratings, sample == FALSE)
```

```r
# Build model using random forest
model2 <- randomForest(IMDb ~ Netflix + Hulu + Prime.Video + Disney., data = train)
model2
```

Code for Shiny App located here:

https://github.com/bpersaud104/Data698/blob/main/Data%20Collection%20and%20Analysis/Netflix_Analysis/app.R

References

[1] Johannes H Snyman, & Debora J Gilliard. (2019). The Streaming Television Industry: Mature or

Still Growing? Journal of Marketing Development and Competitiveness, 13(4), 94–105. https://doi.org/10.33423/jmdc.v13i4.2355

[2] Steiner, E., & Xu, K. (2020). Binge-watching motivates change: Uses and gratifications of streaming video viewers challenge traditional TV research. Convergence (London, England), 26(1), 82–101. https://doi.org/10.1177/1354856517750365

[3] David A. Schweidel, & Wendy W. Moe. (2016). Binge Watching and Advertising. Journal of Marketing, 80(5), 1–19. https://doi.org/10.1509/jm.15.0258

[4] Wayne, M. (2018). Netflix, Amazon, and branded television content in subscription video on-demand portals. Media, Culture & Society, 40(5), 725–741. https://doi.org/10.1177/0163443717736118

[5] Bikfalvi, A., García-Reinoso, J., Vidal, I., Valera, F., & Azcorra, A. (2011). P2P vs. IP multicast: Comparing approaches to IPTV streaming based on TV channel popularity. Computer Networks (Amsterdam, Netherlands : 1999), 55(6), 1310–1325. https://doi.org/10.1016/j.comnet.2010.12.020

[6] Kelly, J. (2019). Television by the numbers: The challenges of audience measurement in the age of Big Data. Convergence (London, England), 25(1), 113–132. https://doi.org/10.1177/1354856517700854

[7] Christian, A. (2020). Beyond Branding: The Value of Intersectionality on Streaming TV Channels. Television & New Media, 21(5), 457–474. https://doi.org/10.1177/1527476419852241

[8] Schauerte, R., Feiereisen, S., & Malter, A. (2020). What does it take to survive in a digital world? Resource-based theory and strategic change in the TV industry. Journal of Cultural Economics. https://doi.org/10.1007/s10824-020-09389-x