Yijun Wang

ALY 6010

Introduction

I got this dataset from the Kaggle website, and I picked the data set that included movies from Netflix, Prime Video, Hulu, and Disney+. This topic was chosen because of covid-19. The long-term decline of movie theaters has created an opportunity for enormous expansion for different streaming media providers. The reason I'd like to study this data is because I'm curious about which streaming service individuals should use.

```
> dim(df)
[1] 6664   17
>
```

```
> #string
> str(df)
'data.frame':   6664 obs. of  17 variables:
 $ X              : int  0 1 2 3 4 5 6 7 8 9 ...
 $ ID             : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Title          : Factor w/ 6664 levels "'71","#Alive",..: 5492 1403 1432 2985 4377 6123 5872 3825 5131 5942 ...
 $ Year           : int  2019 2016 2020 2001 2018 2018 2020 2017 2018 2020 ...
 $ Age            : Factor w/ 6 levels "","13+","16+",..: 4 5 5 5 4 2 2 2 3 4 ...
 $ IMDb           : Factor w/ 75 levels "","1.5/10","1.7/10",..: 62 68 74 65 61 55 60 57 57 62 ...
 $ Rotten.Tomatoes: Factor w/ 83 levels "","10/100","12/100",..: 83 82 81 80 80 80 79 78 78 78 ...
 $ Netflix        : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Hulu           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Prime.Video    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Disney.        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Type           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Directors      : Factor w/ 4793 levels "","A-mer Faruk Sorak",..: 2844 3240 135 432 1 4313 1970 599 1331 30 ...
 $ Genres         : Factor w/ 1078 levels "","Action","Action,Adventure",..: 548 118 809 965 207 666 845 71 660 939 ...
 $ Country        : Factor w/ 781 levels "","Afghanistan,United States",..: 626 281 516 276 611 626 626 449 626 759 ...
 $ Language       : Factor w/ 727 levels "","Afrikaans",..: 247 477 65 477 65 65 65 264 136 267 ...
 $ Runtime        : int  209 161 83 224 52 99 94 120 133 129 ...
>
```

this dataset has 6664 observations and 17 columns. All of dataset are talking around four different platforms: Netflix, Hulu, Disney, and Prime. Video. The goal of this study is to examine the dataset statistically and provide relevant, in-depth findings about the dataset, including data levels, descriptive statistics, and data distribution. Rest of dataset are Movies' title, publish year, Age's level, Rotten.Tomatoes's scores, IMDb's score, Directors, Genres, Country, Language and run time. These data are composed of int and factors.

From the column of “IMDb”, the number is followed by the symbol "/10." Although we recognize that this is a scoring system, it is quite cumbersome when undertaking data processing. As a result, I eliminated the "/10" and changed these as number.

```
> #Remove /100 from Rotten Tomatoes and /10 from IMDb
> df$Rotten.Tomatoes <- substring(df$Rotten.Tomatoes,0,2)
> df$IMDb <- substring(df$IMDb,0,3)
> head(df)
  X ID                                Title Year Age IMDb Rotten.Tomatoes Netflix Hulu Prime.Video Disney. Type
1 0  1                          The Irishman 2019 18+  7.8              98       1    0           0       0    0
2 1  2                               Dangal 2016  7+  8.4              97       1    0           0       0    0
3 2  3 David Attenborough: A Life on Our Planet 2020  7+    9              95       1    0           0       0    0
4 3  4        Lagaan: Once Upon a Time in India 2001  7+  8.1              94       1    0           0       0    0
5 4  5                                    Roma 2018 18+  7.7              94       1    0           0       0    0
6 5  6        To All the Boys I've Loved Before 2018 13+  7.1              94       1    0           0       0    0
                                    Directors                          Genres
1                                Martin Scorsese            Biography,Crime,Drama
2                                  Nitesh Tiwari     Action,Biography,Drama,Sport
3 Alastair Fothergill,Jonathan Hughes,Keith Scholey            Documentary,Biography
4                              Ashutosh Gowariker              Drama,Musical,Sport
5                                                Action,Drama,History,Romance,War
6                                  Susan Johnson             Comedy,Drama,Romance
                                        Country                                 Language Runtime
1                                 United States English,Italian,Latin,Spanish,German     209
2 India,United States,United Kingdom,Australia,Kenya,Namibia                   Hindi,English     161
3                                 United Kingdom                                 English      83
```

And I did same thing to “Rotten.Tomatoes”.

```
6                                                                    United States
> #conver to number
> df$IMDb= as.numeric(df$IMDb)
> df$Rotten.Tomatoes= as.numeric(df$Rotten.Tomatoes)
> #conver to number
> df$IMDb= as.numeric(df$IMDb)
> df$Rotten.Tomatoes= as.numeric(df$Rotten.Tomatoes)
> class(df$IMDb)
[1] "numeric"
> class(df$Rotten.Tomatoes )
[1] "numeric"
>
```

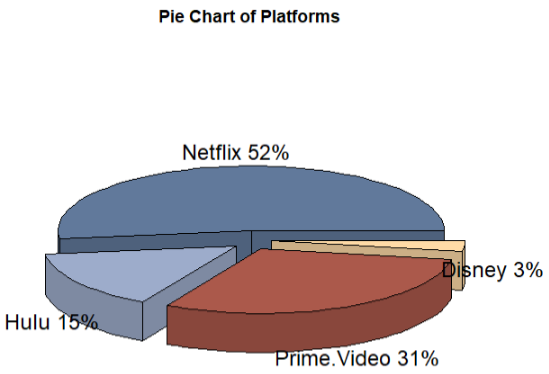
Besides that, I remove blank value from “Age”.

```
+   scale_fill_manual(values=c( #999999 ,  #E69F00 ,
> new_age <- sub(" ",NA,df$Age, fixed =TRUE)
> ggplot(data = df, aes(x = Age)) +
```

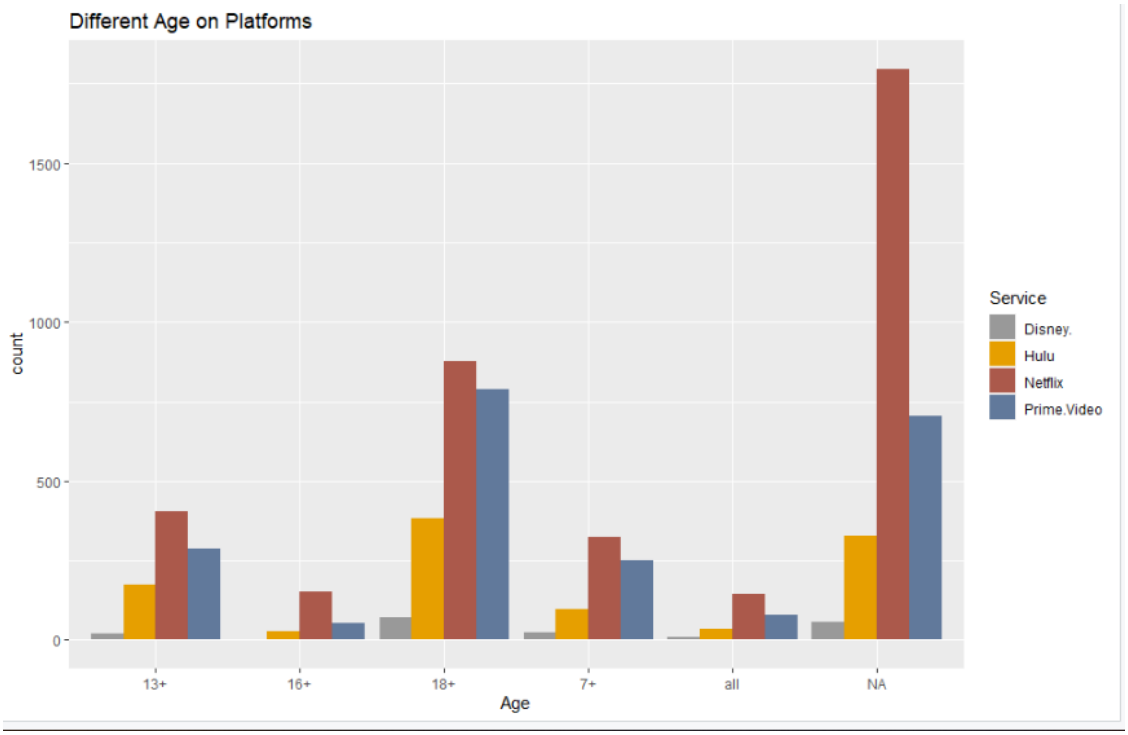
Here is the summary of this dataset.

```
> #summary
> summary(df)
       X                ID                                Title          Year          Age            IMDb      Rotten.Tomatoes
 Min.   :   0   Min.   :   1   '71                              :   1   Min.   :1915      :2747   Min.   :1.50   Min.   :10.00
 1st Qu.:1666   1st Qu.:1667   #Alive                           :   1   1st Qu.:2010   13+: 822   1st Qu.:5.60   1st Qu.:49.00
 Median :3332   Median :3332   #AnneFrank. Parallel Stories:   1   Median :2016   16+: 226   Median :6.40   Median :56.00
 Mean   :3205   Mean   :3206   #cats_the_mewvie                 :   1   Mean   :2011   18+:1972   Mean   :6.29   Mean   :56.62
 3rd Qu.:4997   3rd Qu.:4998   #FriendButMarried                :   1   3rd Qu.:2019   7+ : 646   3rd Qu.:7.10   3rd Qu.:64.00
 Max.   :5366   Max.   :5367   #FriendButMarried 2              :   1   Max.   :2021   all: 251   Max.   :9.20   Max.   :98.00
                               (Other)                          :6658                             NA's   :113    NA's   :7
    Netflix            Hulu          Prime.Video       Disney.             Type                    Directors
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0                          : 246
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0   RaÃ³l Campos,Jan Suter:  22
 Median :1.0000   Median :0.0000   Median :0.0000   Median :0.00000   Median :0   Jay Chapman           :  19
 Mean   :0.5545   Mean   :0.1571   Mean   :0.3244   Mean   :0.02746   Mean   :0   Jay Karas              :  17
 3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0   Marcus Raboy          :  17
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :0   Cathy Garcia-Molina   :  11
                                                                                  (Other)               :6332
                 Genres              Country              Language       Runtime
 Comedy              : 575   United States :2779   English        :3489   Min.   :  3.0
 Drama               : 444   India         : 710   Hindi          : 344   1st Qu.: 89.0
 Documentary         : 344   United Kingdom: 253                   : 186   Median : 98.0
 Comedy,Drama        : 275                 : 159   Spanish        : 185   Mean   :100.1
 Drama,Romance       : 224   Canada        : 142   English,Spanish: 154   3rd Qu.:112.0
 Comedy,Drama,Romance: 218   Japan         : 105   Japanese       :  98   Max.   :359.0
 (Other)             :4584   (Other)       :2516   (Other)        :2208   NA's   :195
>
```

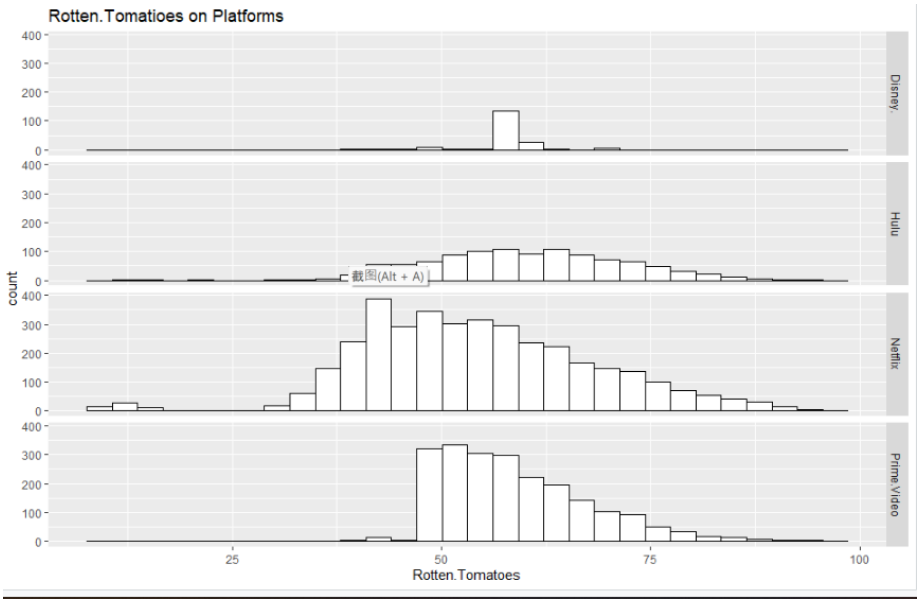
Data Analysis



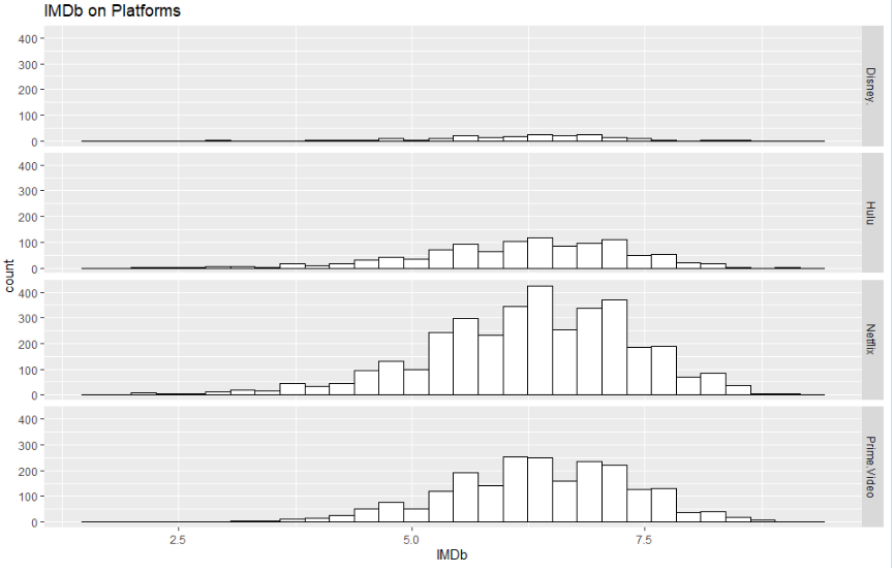
According to this pie chart, Netflix has the most total number of movies played, followed by Prime. The last but not least is Disney.



This bar graph depicts the total number of movies watched by each age group on various platforms. According to the graph, Netflix has the most broadcasts, while Disney has the fewest.



As shown on the following graph, the average score of Netflix platform on Rotten. Tomatoes is around 50. Prime. Video has a substantially higher average score than the other three formats.



This is the number of times count has received a rating on IMDb. In contrast to Rotten. Tomatoes, Netflix have considerably better IMDb ratings, especially when compared to Disney.

Summary

Although Prime Video has the highest average rating on Rotten. Tomatoes, Netflix is the best option if you value quality over IMDb. The platform with the most films published in the last ten years and over 90% IMDb approval. On the other hand, if you want to meet the needs of your children, Disney+ is a fantastic subscription because it is specifically designed for them.

Kaggle's movie data contains information about movies that have been played on four different platforms, as well as other relevant information. This data set comprises 17 attributes associated with 6664 observations, as well as the title, year, and director of the film. The process of data cleansing is critical. We can remove extraneous information, such as blank values in Age. Furthermore, the use of vertical charts, pie charts, or box plots helps us to examine the direction of the data and the relevant correlations in a more natural manner.

Question: Based on the years in the data, can it be concluded about which year the movies are more popular?

Bibliography

Schork, J. (2020, 11 20). *Change Colors of Axis Labels & Values of Base R Plot (2 Examples)*. Retrieved from statisticsglobe: https://statisticsglobe.com/r-change-colors-axis-labels-values-of-plot

Zach. (2021, 02 04). *Format Numbers as Percentages in R (With Examples)*. Retrieved from statology: https://www.statology.org/percentage-in-r/

ZACH. (2021, 04 21). *How to Create Relative Frequency Tables in R*. Retrieved from statology: https://www.statology.org/relative-frequency-table-in-r/