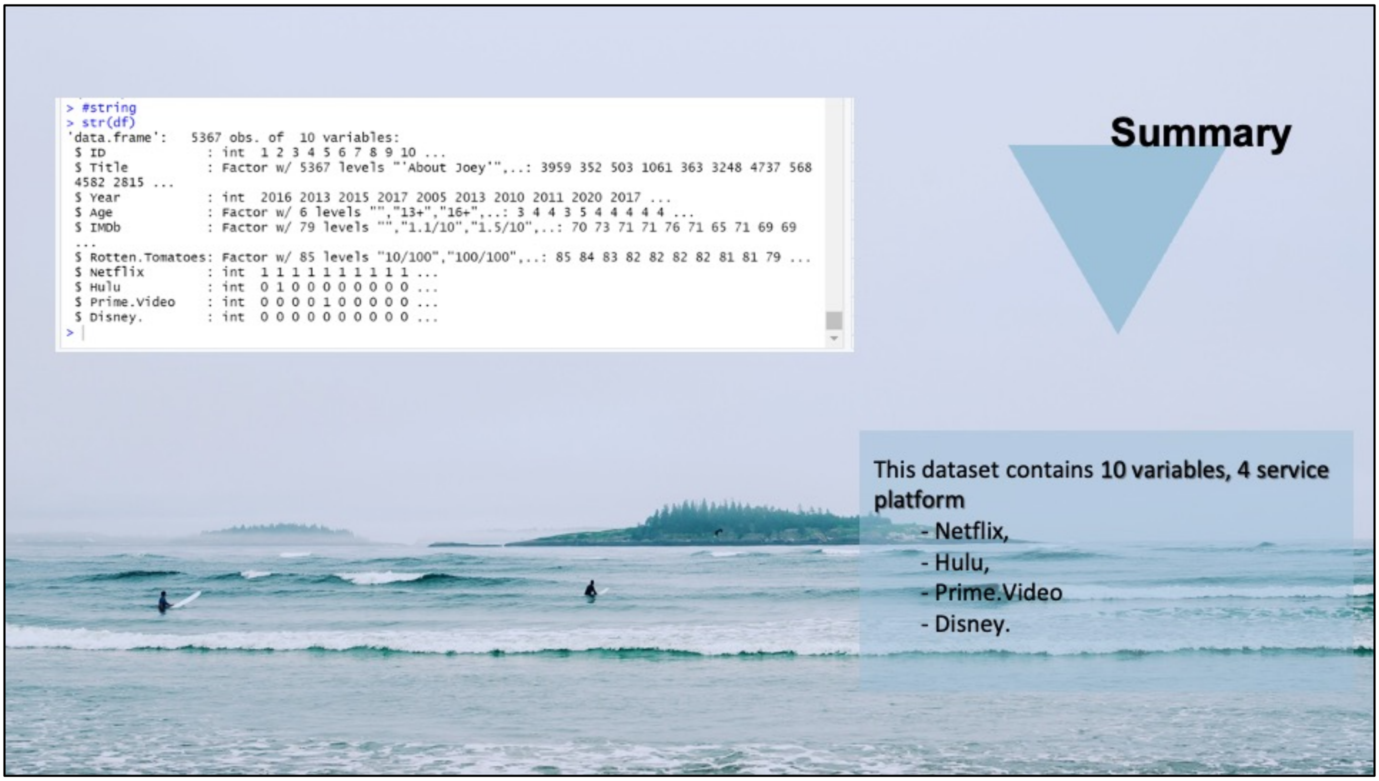
# AnalysIs on Movie

By : Yijun Wang



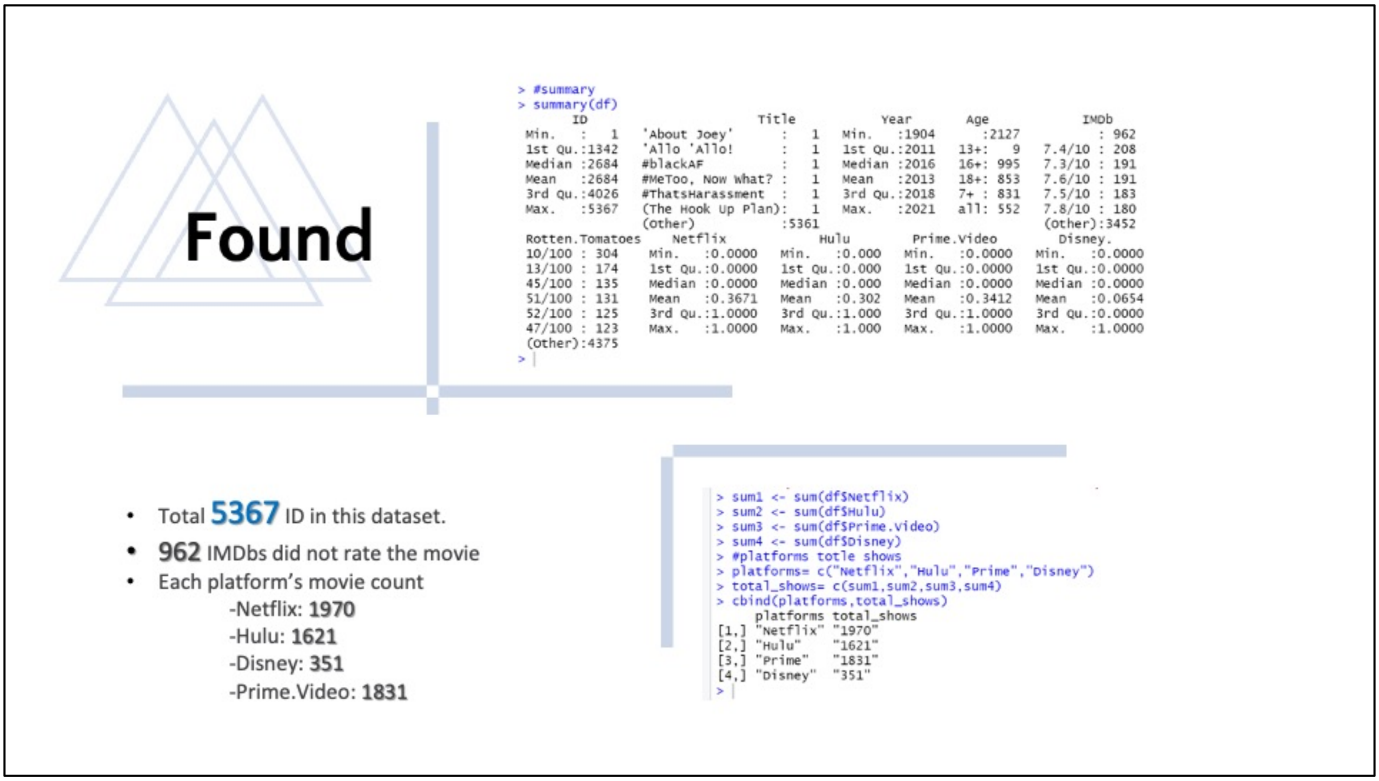
## Summary

```
> #string
> str(df)
'data.frame':	5367 obs. of  10 variables:
 $ ID             : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Title          : Factor w/ 5367 levels "'About Joey'",..: 3959 352 503 1061 363 3248 4737 568
 4582 2815 ...
 $ Year           : int  2016 2013 2015 2017 2005 2013 2010 2011 2020 2017 ...
 $ Age            : Factor w/ 6 levels "","13+","16+",..: 3 4 4 3 5 4 4 4 4 4 ...
 $ IMDb           : Factor w/ 79 levels "","1.1/10","1.5/10",..: 70 73 71 71 76 71 65 71 69 69
 ...
 $ Rotten.Tomatoes: Factor w/ 85 levels "10/100","100/100",..: 85 84 83 82 82 82 82 81 81 79 ...
 $ Netflix        : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Hulu           : int  0 1 0 0 0 0 0 0 0 0 ...
 $ Prime.Video    : int  0 0 0 0 1 0 0 0 0 0 ...
 $ Disney.        : int  0 0 0 0 0 0 0 0 0 0 ...
> 
```

This dataset contains **10 variables, 4 service platform**

- Netflix,
- Hulu,
- Prime.Video
- Disney.

#string
str(df)



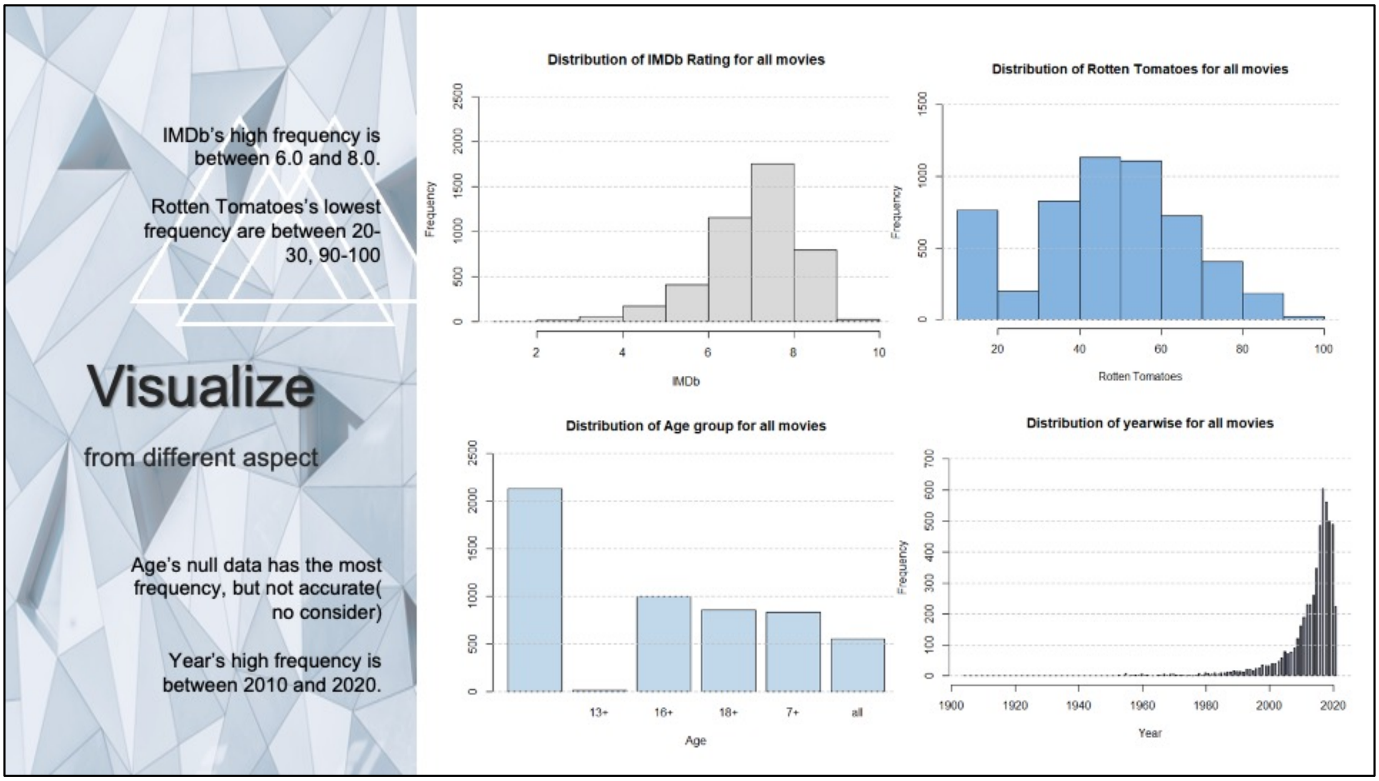
## Found

```
> #summary
> summary(df)
       ID                           Title            Year          Age              IMDb     
 Min.   :   1   'About Joey'        :   1   Min.   :1904       :2127            : 962  
 1st Qu.:1342   'Allo 'Allo!        :   1   1st Qu.:2011   13+:    9   7.4/10 : 208  
 Median :2684   #blackAF            :   1   Median :2016   16+: 995   7.3/10 : 191  
 Mean   :2684   #MeToo, Now What? :   1   Mean   :2013   18+: 853   7.6/10 : 191  
 3rd Qu.:4026   #ThatsHarassment  :   1   3rd Qu.:2018   7+ : 831   7.5/10 : 183  
 Max.   :5367   (The Hook Up Plan):   1   Max.   :2021   all: 552   7.8/10 : 180  
                (Other)             :5361                            (Other):3452  
 Rotten.Tomatoes     Netflix             Hulu         Prime.Video         Disney.      
 10/100 : 304   Min.   :0.0000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000  
 13/100 : 174   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000  
 45/100 : 135   Median :0.0000   Median :0.000   Median :0.0000   Median :0.0000  
 51/100 : 131   Mean   :0.3671   Mean   :0.302   Mean   :0.3412   Mean   :0.0654  
 52/100 : 125   3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:0.0000  
 47/100 : 123   Max.   :1.0000   Max.   :1.000   Max.   :1.0000   Max.   :1.0000  
 (Other):4375                                                                     
> |
```

- Total **5367** ID in this dataset.
- **962** IMDbs did not rate the movie
- Each platform's movie count
  - -Netflix: **1970**
  - -Hulu: **1621**
  - -Disney: **351**
  - -Prime.Video: **1831**

```
> sum1 <- sum(df$Netflix)
> sum2 <- sum(df$Hulu)
> sum3 <- sum(df$Prime.Video)
> sum4 <- sum(df$Disney)
> #platforms totle shows
> platforms= c("Netflix","Hulu","Prime","Disney")
> total_shows= c(sum1,sum2,sum3,sum4)
> cbind(platforms,total_shows)
     platforms total_shows
[1,] "Netflix" "1970"     
[2,] "Hulu"    "1621"     
[3,] "Prime"   "1831"     
[4,] "Disney"  "351"      
> |
```

```
#summary
summary(df)


#sum
sum1 <- sum(df$Netflix)
sum2 <- sum(df$Hulu)
sum3 <- sum(df$Prime.Video)
sum4 <- sum(df$Disney)

#platforms totle shows
platforms= c("Netflix","Hulu","Prime","Disney")
total_shows= c(sum1,sum2,sum3,sum4)
cbind(platforms,total_shows)
```



IMDb's high frequency is between 6.0 and 8.0.

Rotten Tomatoes's lowest frequency are between 20-30, 90-100

Visualize

from different aspect

Age's null data has the most frequency, but not accurate( no consider)

Year's high frequency is between 2010 and 2020.

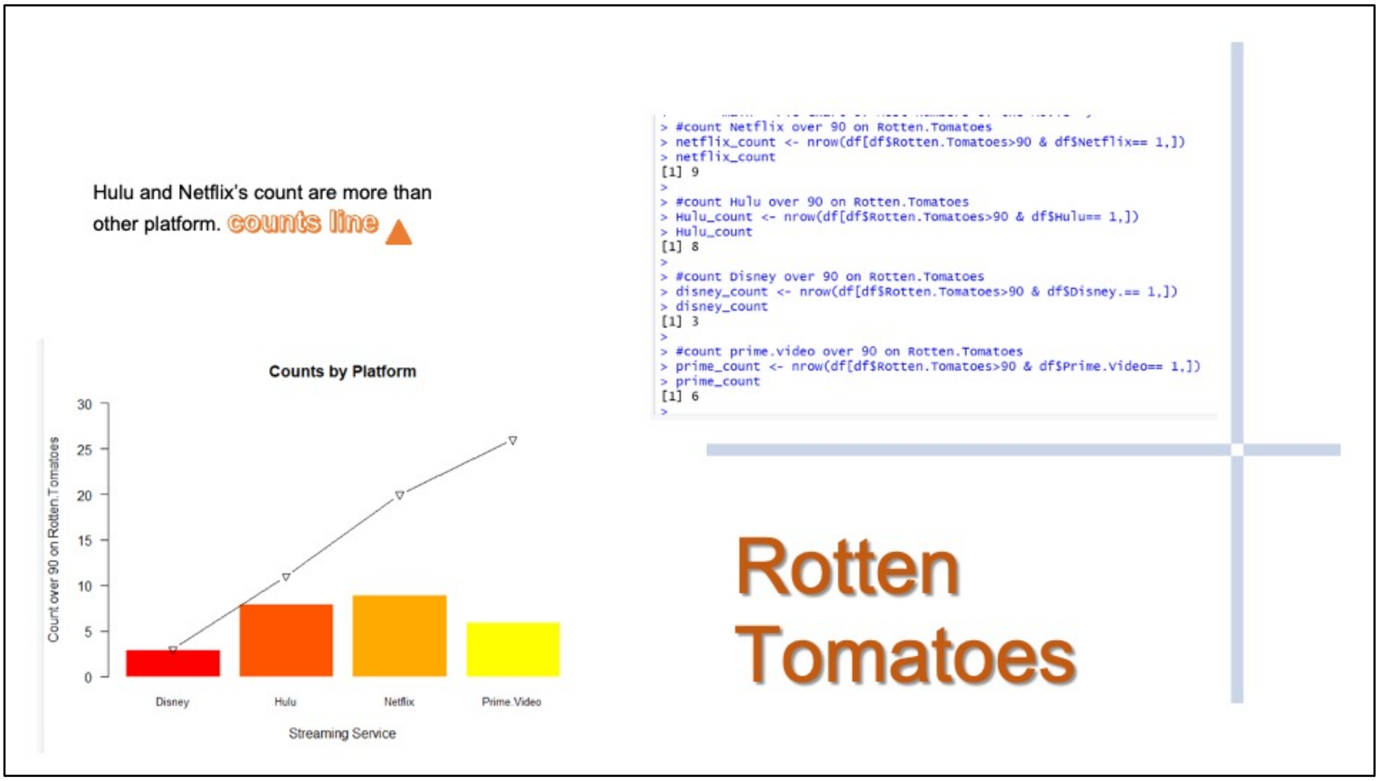
Distribution of IMDb Rating for all movies

Distribution of Rotten Tomatoes for all movies

Distribution of Age group for all movies

Distribution of yearwise for all movies

```
#Histgram of IMDb
hist(df$IMDb,col= "#D9D9D9",
    main= "Distribution of IMDb Rating for all movies",xlab= "IMDb",breaks= 7,
    ylim= c(0,2500))
grid(nx= NA, ny= NULL, lty= 2, col= "gray", lwd= 1)

#Histgram of Rotten Tomatoes
hist(df$Rotten.Tomatoes,col= "#85B4E0",
    main= "Distribution of Rotten Tomatoes for all movies",xlab= "Rotten Tomatoes",
    breaks= 7, ylim= c(0,1500))
grid(nx= NA, ny= NULL, lty= 2, col= "gray", lwd= 1)

#Plot of Age
plot(df$Age,col= "#C1D8EA",
    main= "Distribution of Age group for all movies",xlab= "Age",breaks= 7,
    ylim= c(0,2500))
grid(nx= NA, ny= NULL, lty= 2, col= "gray", lwd= 1)

#Histgram of Year
hist(df$Year,col= "#727a93",
    main= "Distribution of yearwise for all movies",xlab= "Year",breaks= 200,
    ylim= c(0,5000))
grid(nx= NA, ny= NULL, lty= 2, col= "gray", lwd= 1)
```



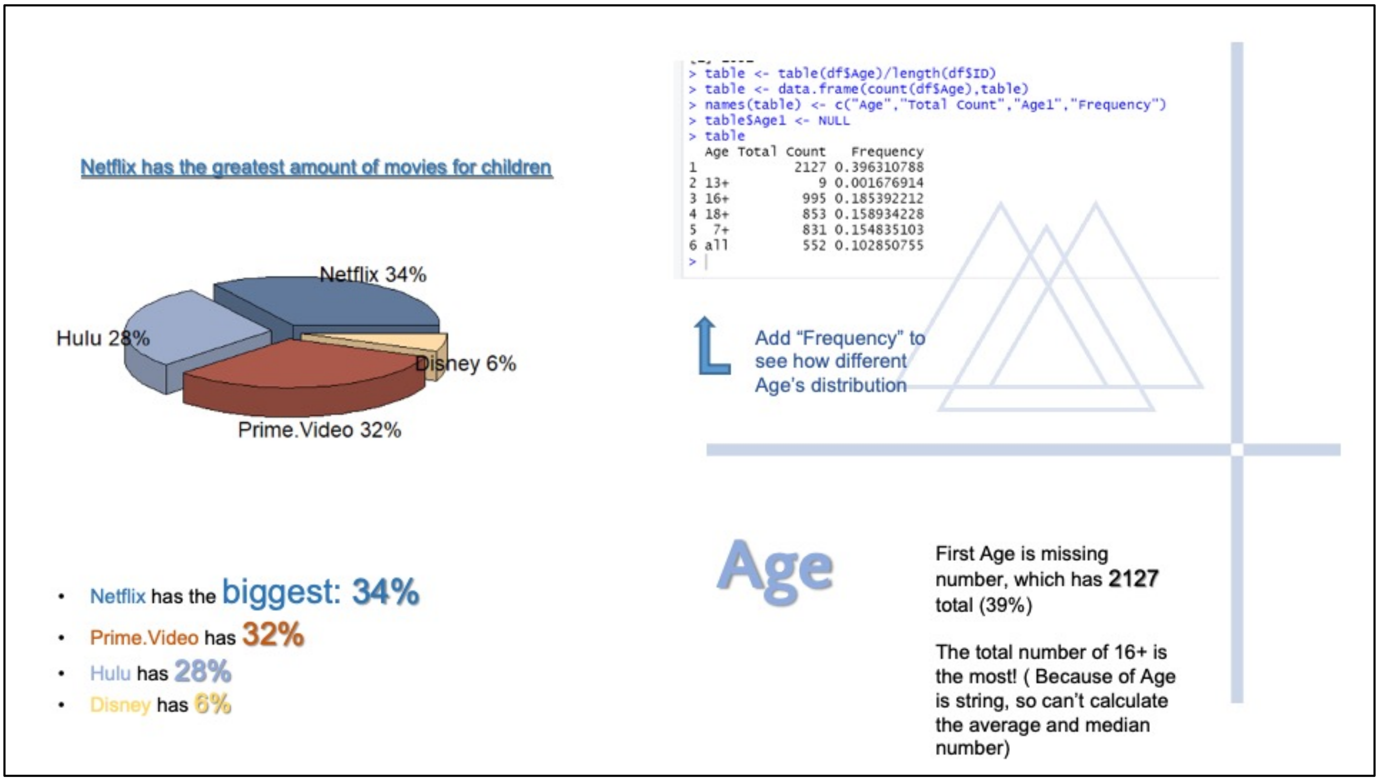
Hulu and Netflix's count are more than other platform. counts line ▲

```
> #count Netflix over 90 on Rotten.Tomatoes
> netflix_count <- nrow(df[df$Rotten.Tomatoes>90 & df$Netflix== 1,])
> netflix_count
[1] 9
>
> #count Hulu over 90 on Rotten.Tomatoes
> Hulu_count <- nrow(df[df$Rotten.Tomatoes>90 & df$Hulu== 1,])
> Hulu_count
[1] 8
>
> #count Disney over 90 on Rotten.Tomatoes
> disney_count <- nrow(df[df$Rotten.Tomatoes>90 & df$Disney.== 1,])
> disney_count
[1] 3
>
> #count prime.video over 90 on Rotten.Tomatoes
> prime_count <- nrow(df[df$Rotten.Tomatoes>90 & df$Prime.Video== 1,])
> prime_count
[1] 6
>
```

## Rotten Tomatoes

```
#count Netflix over 90 on Rotten.Tomatoes
netflix_count <- nrow(df[df$Rotten.Tomatoes>90 & df$Netflix== 1,])
netflix_count

#count Hulu over 90 on Rotten.Tomatoes
Hulu_count <- nrow(df[df$Rotten.Tomatoes>90 & df$Hulu== 1,])
Hulu_count

#count Disney over 90 on Rotten.Tomatoes
disney_count <- nrow(df[df$Rotten.Tomatoes>90 & df$Disney.== 1,])
disney_count

#count prime.video over 90 on Rotten.Tomatoes
prime_count <- nrow(df[df$Rotten.Tomatoes>90 & df$Prime.Video== 1,])
prime_count

#Barplot of over 90 on Rotten.Tomatoes
names<-c("Disney","Hulu","Netflix","Prime.Video")
counts<-c(disney_count,Hulu_count, netflix_count,prime_count)
p <-barplot(counts,names.arg= names,las= 1,cex.names= 0.8, ylim= c(0,30),
        main= "Counts by Platform",col= heat.colors(5),border= "white",
        ylab= "Count over 90 on Rotten.Tomatoes",
        xlab= "Streaming Service")
```

```
cum_sums<-cumsum(counts)
lines(p, cum_sums, type= 'b', pch= 6, col= 'black')
```



Netflix has the greatest amount of movies for children

- Netflix has the **biggest: 34%**
- Prime.Video has **32%**
- Hulu has **28%**
- Disney has **6%**

```
> table <- table(df$Age)/length(df$ID)
> table <- data.frame(count(df$Age),table)
> names(table) <- c("Age","Total Count","Age1","Frequency")
> table$Age1 <- NULL
> table
  Age Total Count   Frequency
1            2127 0.396310788
2 13+           9 0.001676914
3 16+         995 0.185392212
4 18+         853 0.158934228
5  7+         831 0.154835103
6 all         552 0.102850755
> |
```

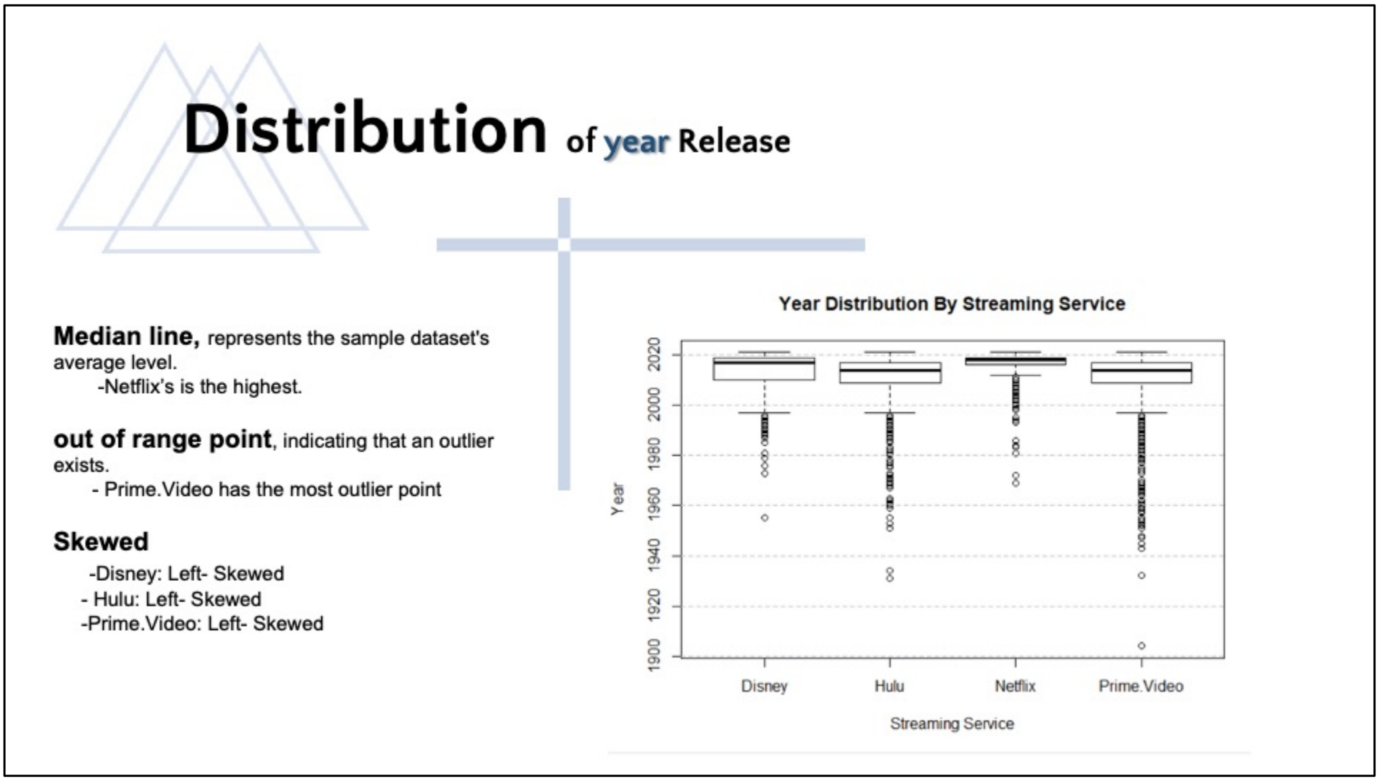
Add “Frequency” to see how different Age’s distribution

**Age**

First Age is missing number, which has **2127** total (39%)

The total number of 16+ is the most! ( Because of Age is string, so can’t calculate the average and median number)

#3D PIE CHART: the Service has the highest number of the movie
slices <- c(sum1,sum2,sum3,sum4)
lbls <- c("Netflix", "Hulu", "Prime.Video", "Disney")
pct <- round(slices/sum(slices)*100)
lbls <- paste(lbls, pct)
lbls <- paste(lbls,"%",sep= "")
colors <- c("#61799b", "#9daccb", "#ab594b", "#ffdba7")
pie3D(slices,labels= lbls,explode= 0.1,col= colors,
      main= "Pie Chart of Most Numbers of the Movie ")



## Distribution of year Release

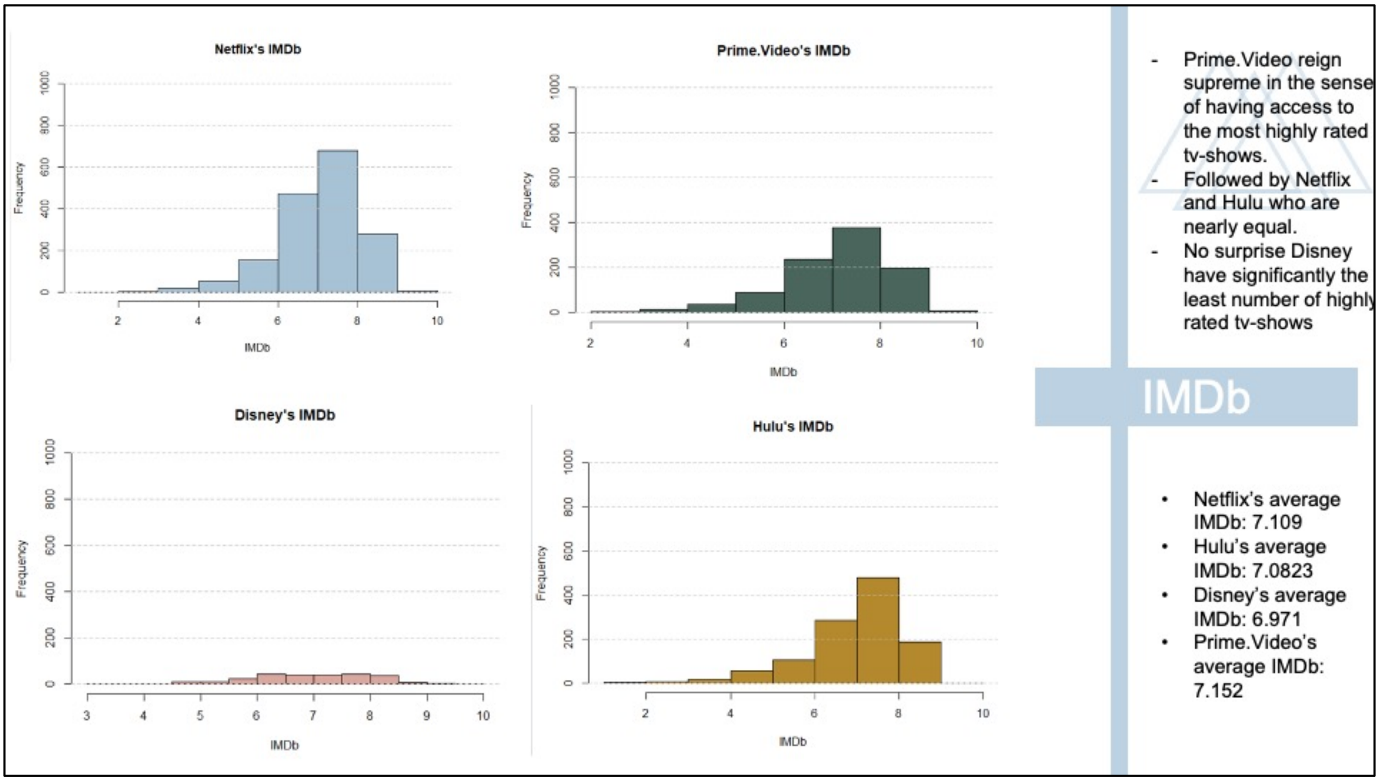
**Median line,** represents the sample dataset's average level.
-Netflix's is the highest.

**out of range point**, indicating that an outlier exists.
- Prime.Video has the most outlier point

**Skewed**

-Disney: Left- Skewed
- Hulu: Left- Skewed
-Prime.Video: Left- Skewed

```
#Boxplot of 4 Platform's year
list <-list(D$Year,H$Year,N$Year,P$Year)
boxplot(list, main= "Year Distribution By Streaming Service",
        xlab= "Streaming Service",ylab= "Year",names= names)
grid(nx= NA, ny= NULL, lty= 2, col= "gray", lwd= 1)
```



Netflix's IMDb

Prime.Video's IMDb

Disney's IMDb

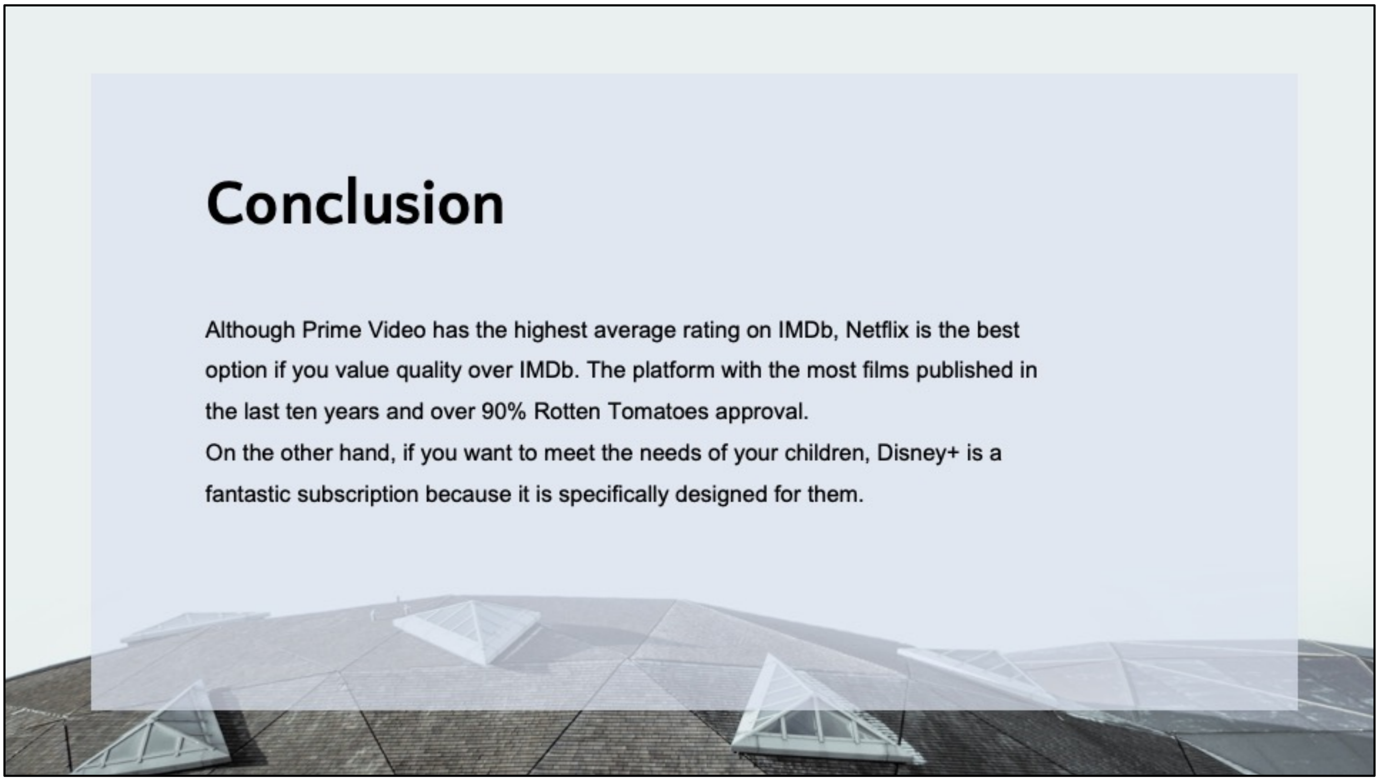
Hulu's IMDb

IMDb

- Prime.Video reign supreme in the sense of having access to the most highly rated tv-shows.
- Followed by Netflix and Hulu who are nearly equal.
- No surprise Disney have significantly the least number of highly rated tv-shows

- Netflix's average IMDb: 7.109
- Hulu's average IMDb: 7.0823
- Disney's average IMDb: 6.971
- Prime.Video's average IMDb: 7.152

```
#Distribution of Netflix
hist(N$IMDb,col= "#a7c2d5",
    main= "Netflix's IMDb",
    xlab= "IMDb", ylim= c(0,1000),breaks= 10)
grid(nx= NA, ny= NULL, lty= 2, col= "gray", lwd= 1)

#Distribution of prime.Video
hist(P$IMDb,col= "#49655c",
    main= "Prime.Video's IMDb",
    xlab= "IMDb", ylim= c(0,1000),breaks= 10)
grid(nx= NA, ny= NULL, lty= 2, col= "gray", lwd= 1)

#Distribution of Disney
hist(D$IMDb,col= "#dba89f",
    main= "Disney's IMDb",
    xlab= "IMDb", ylim= c(0,1000),breaks= 10)
grid(nx= NA, ny= NULL, lty= 2, col= "gray", lwd= 1)

#Distribution of Hulu
hist(H$IMDb,col= "#b4892e",
    main= "Hulu's IMDb",
    xlab= "IMDb", ylim= c(0,1000),breaks= 10)
grid(nx= NA, ny= NULL, lty= 2, col= "gray", lwd= 1)
```

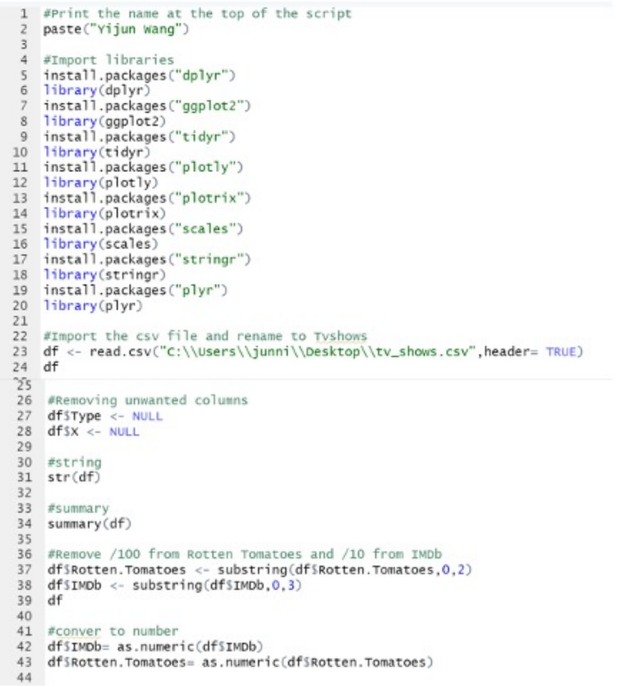


# Conclusion

Although Prime Video has the highest average rating on IMDb, 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% Rotten Tomatoes approval.

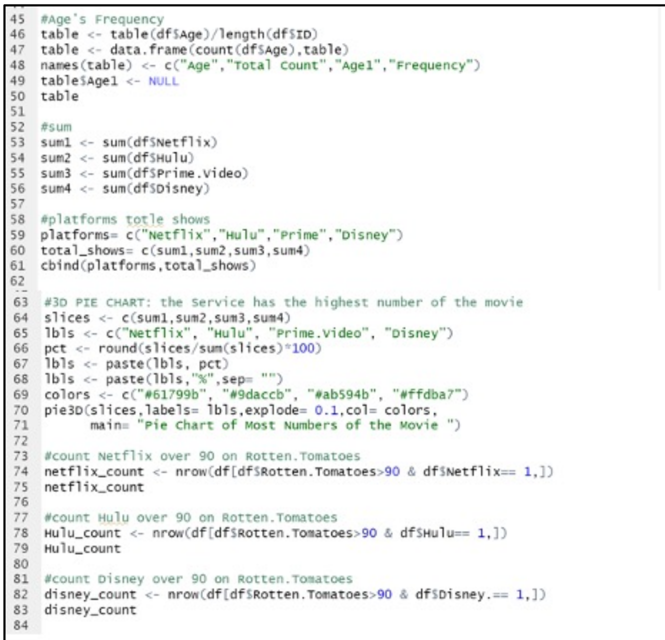
On the other hand, if you want to meet the needs of your children, Disney+ is a fantastic subscription because it is specifically designed for them.

# Code



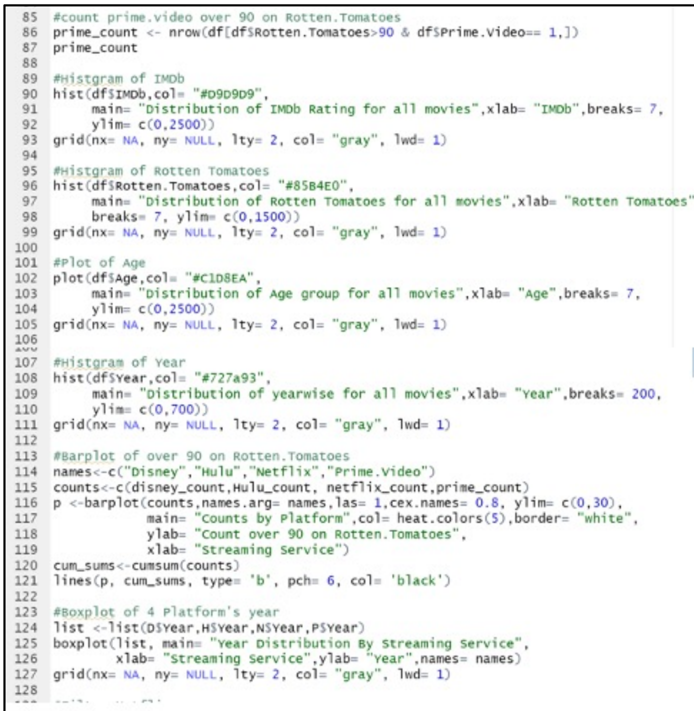
```
#Print the name at the top of the script
paste("Yijun Wang")

#Import libraries
install.packages("dplyr")
library(dplyr)
install.packages("ggplot2")
library(ggplot2)
install.packages("tidyr")
library(tidyr)
install.packages("plotly")
library(plotly)
install.packages("plotrix")
library(plotrix)
install.packages("scales")
library(scales)
install.packages("stringr")
library(stringr)
install.packages("plyr")
library(plyr)

#Import the csv file and rename to Tvshows
df <- read.csv("C:\\Users\\junni\\Desktop\\tv_shows.csv",header= TRUE)
df

#Removing unwanted columns
df$Type <- NULL
df$X <- NULL

#string
str(df)

#summary
summary(df)

#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)
df

#conver to number
df$IMDb= as.numeric(df$IMDb)
df$Rotten.Tomatoes= as.numeric(df$Rotten.Tomatoes)

```

# Code



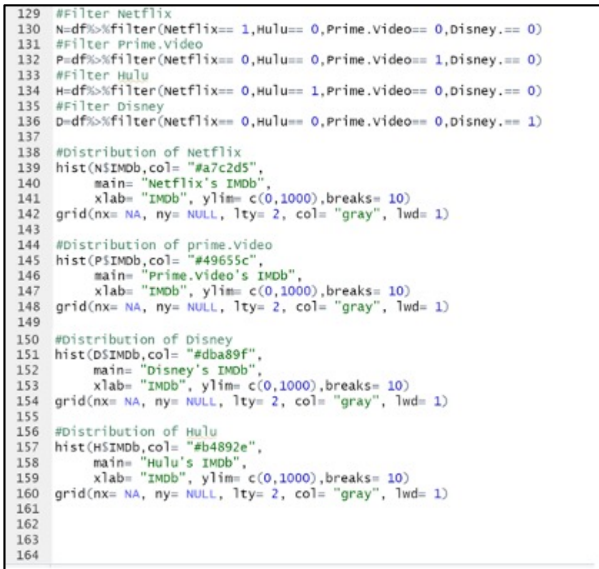
```r
#Age's Frequency
table <- table(df$Age)/length(df$ID)
table <- data.frame(count(df$Age),table)
names(table) <- c("Age","Total Count","Age1","Frequency")
table$Age1 <- NULL
table

#sum
sum1 <- sum(df$Netflix)
sum2 <- sum(df$Hulu)
sum3 <- sum(df$Prime.Video)
sum4 <- sum(df$Disney)

#platforms totle shows
platforms= c("Netflix","Hulu","Prime","Disney")
total_shows= c(sum1,sum2,sum3,sum4)
cbind(platforms,total_shows)

#3D PIE CHART: the Service has the highest number of the movie
slices <- c(sum1,sum2,sum3,sum4)
lbls <- c("Netflix", "Hulu", "Prime.Video", "Disney")
pct <- round(slices/sum(slices)*100)
lbls <- paste(lbls, pct)
lbls <- paste(lbls,"%",sep= "")
colors <- c("#61799b", "#9daccb", "#ab594b", "#ffdba7")
pie3D(slices,labels= lbls,explode= 0.1,col= colors,
      main= "Pie Chart of Most Numbers of the Movie ")

#count Netflix over 90 on Rotten.Tomatoes
netflix_count <- nrow(df[df$Rotten.Tomatoes>90 & df$Netflix== 1,])
netflix_count

#count Hulu over 90 on Rotten.Tomatoes
Hulu_count <- nrow(df[df$Rotten.Tomatoes>90 & df$Hulu== 1,])
Hulu_count

#count Disney over 90 on Rotten.Tomatoes
disney_count <- nrow(df[df$Rotten.Tomatoes>90 & df$Disney.== 1,])
disney_count
```

# Code



```r
#count prime.video over 90 on Rotten.Tomatoes
prime_count <- nrow(df[df$Rotten.Tomatoes>90 & df$Prime.Video== 1,])
prime_count

#Histgram of IMDb
hist(df$IMDb,col= "#D9D9D9",
     main= "Distribution of IMDb Rating for all movies",xlab= "IMDb",breaks= 7,
     ylim= c(0,2500))
grid(nx= NA, ny= NULL, lty= 2, col= "gray", lwd= 1)

#Histgram of Rotten Tomatoes
hist(df$Rotten.Tomatoes,col= "#85B4E0",
     main= "Distribution of Rotten Tomatoes for all movies",xlab= "Rotten Tomatoes",
     breaks= 7, ylim= c(0,1500))
grid(nx= NA, ny= NULL, lty= 2, col= "gray", lwd= 1)

#Plot of Age
plot(df$Age,col= "#C1D8EA",
     main= "Distribution of Age group for all movies",xlab= "Age",breaks= 7,
     ylim= c(0,2500))
grid(nx= NA, ny= NULL, lty= 2, col= "gray", lwd= 1)

#Histgram of Year
hist(df$Year,col= "#727a93",
     main= "Distribution of yearwise for all movies",xlab= "Year",breaks= 200,
     ylim= c(0,700))
grid(nx= NA, ny= NULL, lty= 2, col= "gray", lwd= 1)

#Barplot of over 90 on Rotten.Tomatoes
names<-c("Disney","Hulu","Netflix","Prime.Video")
counts<-c(disney_count,Hulu_count, netflix_count,prime_count)
p <-barplot(counts,names.arg= names,las= 1,cex.names= 0.8, ylim= c(0,30),
            main= "Counts by Platform",col= heat.colors(5),border= "white",
            ylab= "Count over 90 on Rotten.Tomatoes",
            xlab= "Streaming Service")
cum_sums<-cumsum(counts)
lines(p, cum_sums, type= 'b', pch= 6, col= 'black')

#Boxplot of 4 Platform's year
list <-list(D$Year,H$Year,N$Year,P$Year)
boxplot(list, main= "Year Distribution By Streaming Service",
        xlab= "Streaming Service",ylab= "Year",names= names)
grid(nx= NA, ny= NULL, lty= 2, col= "gray", lwd= 1)
```

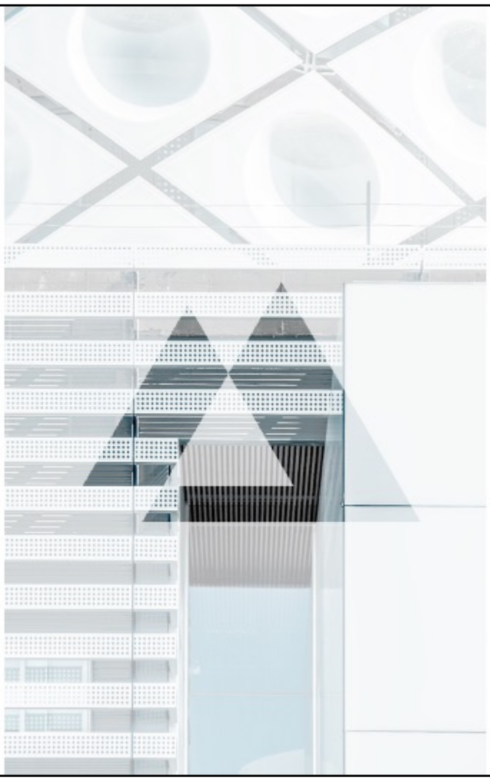
# Code



```r
#Filter Netflix
N=df%>%filter(Netflix== 1,Hulu== 0,Prime.Video== 0,Disney.== 0)
#Filter Prime.Video
P=df%>%filter(Netflix== 0,Hulu== 0,Prime.Video== 1,Disney.== 0)
#Filter Hulu
H=df%>%filter(Netflix== 0,Hulu== 1,Prime.Video== 0,Disney.== 0)
#Filter Disney
D=df%>%filter(Netflix== 0,Hulu== 0,Prime.Video== 0,Disney.== 1)

#Distribution of Netflix
hist(N$IMDb,col= "#a7c2d5",
     main= "Netflix's IMDb",
     xlab= "IMDb", ylim= c(0,1000),breaks= 10)
grid(nx= NA, ny= NULL, lty= 2, col= "gray", lwd= 1)

#Distribution of prime.Video
hist(P$IMDb,col= "#49655c",
     main= "Prime.Video's IMDb",
     xlab= "IMDb", ylim= c(0,1000),breaks= 10)
grid(nx= NA, ny= NULL, lty= 2, col= "gray", lwd= 1)

#Distribution of Disney
hist(D$IMDb,col= "#dba89f",
     main= "Disney's IMDb",
     xlab= "IMDb", ylim= c(0,1000),breaks= 10)
grid(nx= NA, ny= NULL, lty= 2, col= "gray", lwd= 1)

#Distribution of Hulu
hist(H$IMDb,col= "#b4892e",
     main= "Hulu's IMDb",
     xlab= "IMDb", ylim= c(0,1000),breaks= 10)
grid(nx= NA, ny= NULL, lty= 2, col= "gray", lwd= 1)
```

# THANKS

Yijun Wang



## 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/