



# Streaming TV Shows

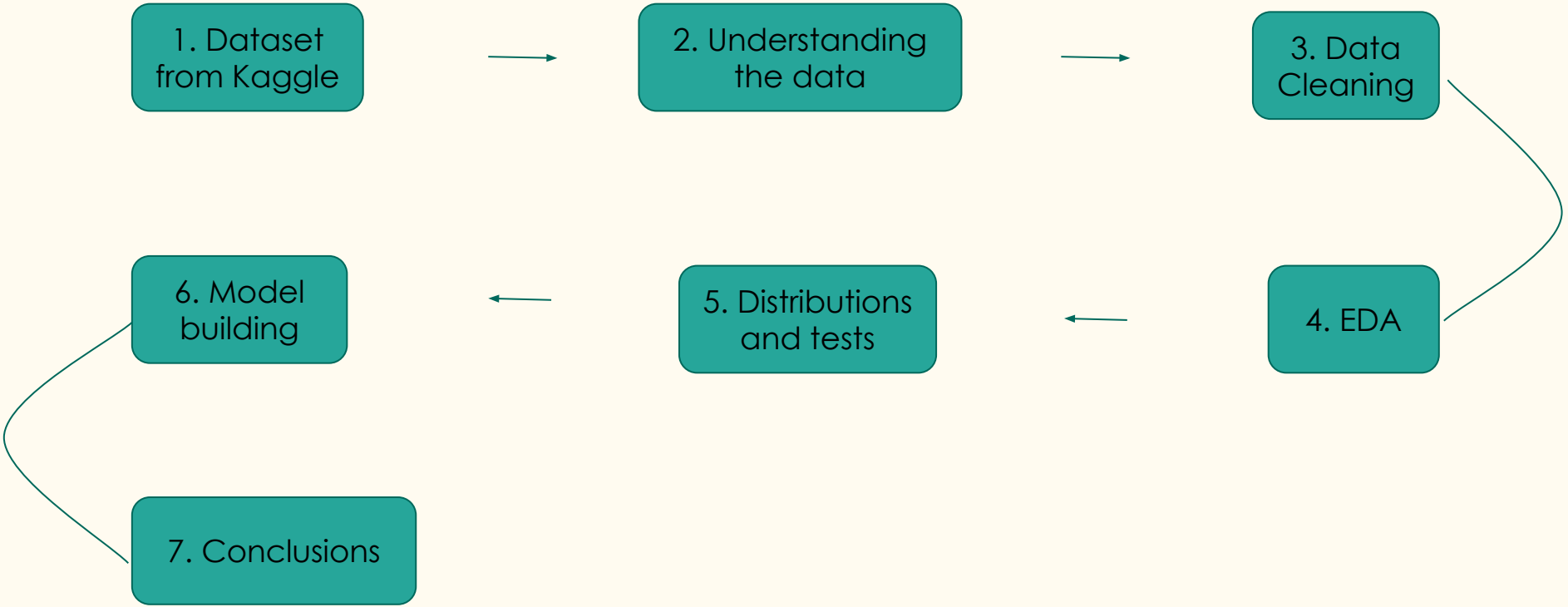
**Data With Density:** Evelyn Campo, Nusrat Prithee, Xiao Qi, Roman Kosarzycki

# Introduction

SMART Questions:

- What are the most targeted age groups for the TV shows by Netflix, Hulu, PrimeVideo, Disney+?
- Which year published the highest number of TV shows?
- Which streaming platform has the highest average rating (according to Rotten Tomatoes and IMDb)?
- What is the relationship between IMDb and Rotten Tomatoes?

# Steps Conducted



# Understanding the data

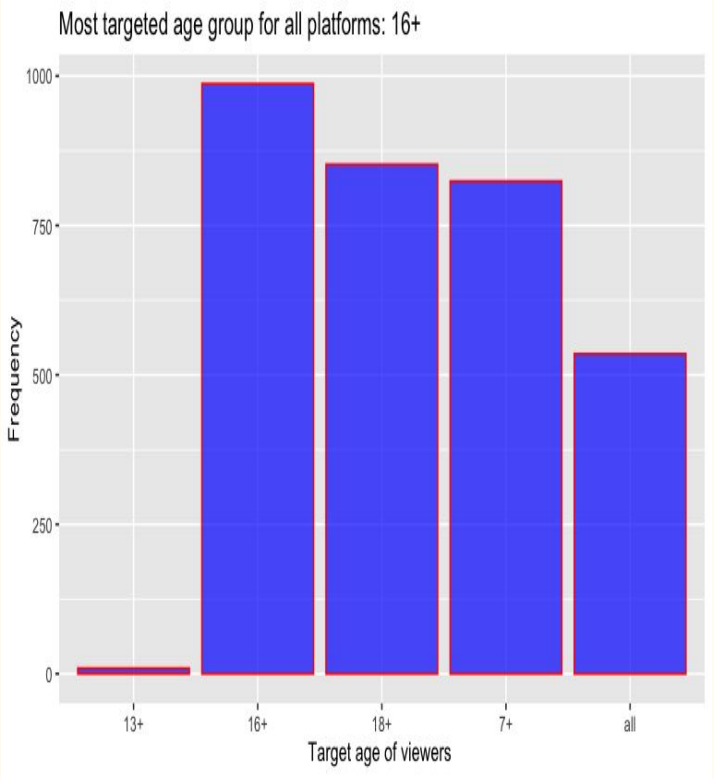
## Dataset Summary

| **Observations** | 5368 | **NA** | **Description** |
|---|---|---|---|
| **Variables:** 9 | **Class** | 3089 | |
| Title | Character | - | Name of the TV show |
| Year | Number | - | The year in which the tv show was produced |
| Age | Character | 2127 | Target age group |
| IMDb | Character | 962 | Rating/10 of Internet Movie Database: help fans explore the world of movies and shows and decide |
| Rotten Tomatoes | Character | - | percentage of professional critic reviews that are positive for a given film or television show. |
| Netflix | Factor | - | Streaming platform: Whether the tv show is found |
| Hulu | Factor | - | Streaming platform: Whether the tv show is found |
| Prime Video | Factor | - | Streaming platform: Whether the tv show is found |
| Disney+ | Factor | - | Streaming platform: Whether the tv show is found |

# Data Cleaning

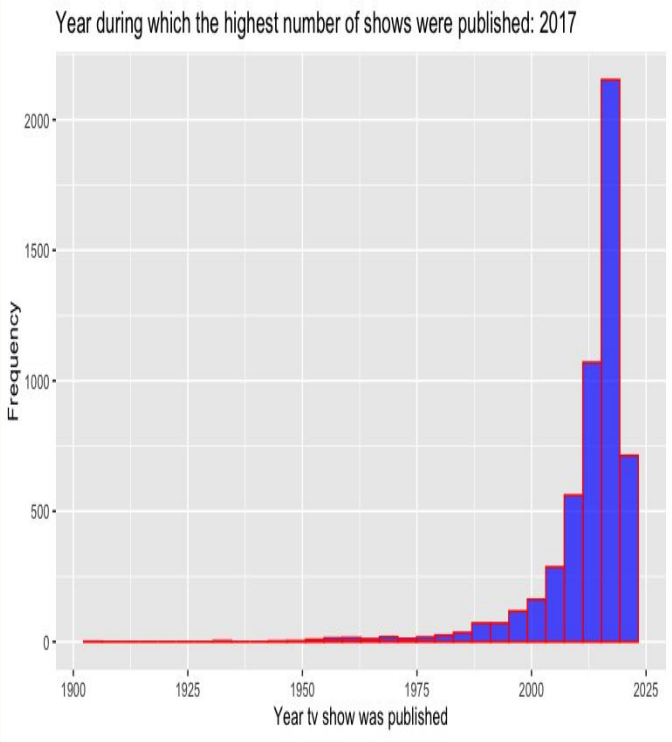
- Dropped variables X1 and ID
- Assigned "NA" to all blank cells (specifically NA for Age)
- Replacement of substrings with  gsub(old, new, string) function for variables IDMb and Rotten Tomatoes
- Turned the variables for the streaming platforms into as.factor()
- Counted missing values for each variable

# Exploratory Data Analysis
## Age and Year




Positively skewed

| Frequency Table | | | | |
|---|---|---|---|---|
| 13+ | 16+ | 18+ | 7+ | all |
| 9 | 995 | 854 | 831 | 552 |
| 0.0020 | 0.30 | 0.3 | 0.25 | 0.25 |




Variance: 102.8

Mean: 2012

SD: 10.14

Negatively skewed

# Exploratory Data Analysis

## Normality Test

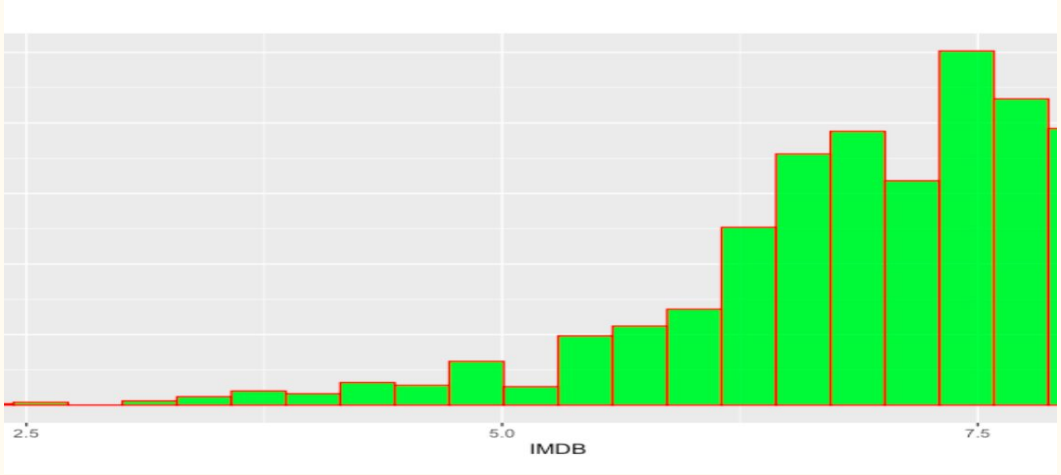
Normality check for the variables IMDb and Rotten.Tomatoes for Netflix

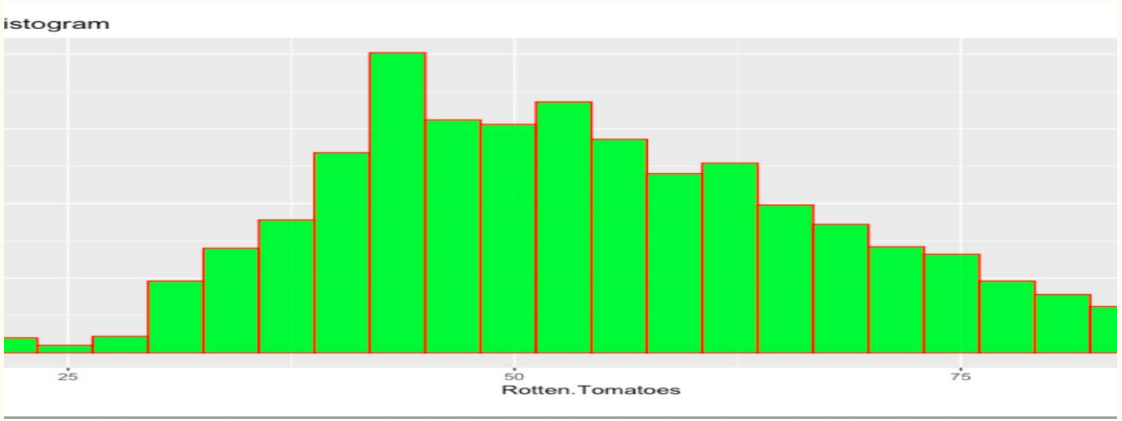
```
        Shapiro-Wilk normality test

data:  netflixtv$IMDb
W = 0.94818, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  netflixtv$Rotten.Tomatoes
W = 0.99215, p-value = 8.532e-09
```



# Exploratory Data Analysis

## Normality Test

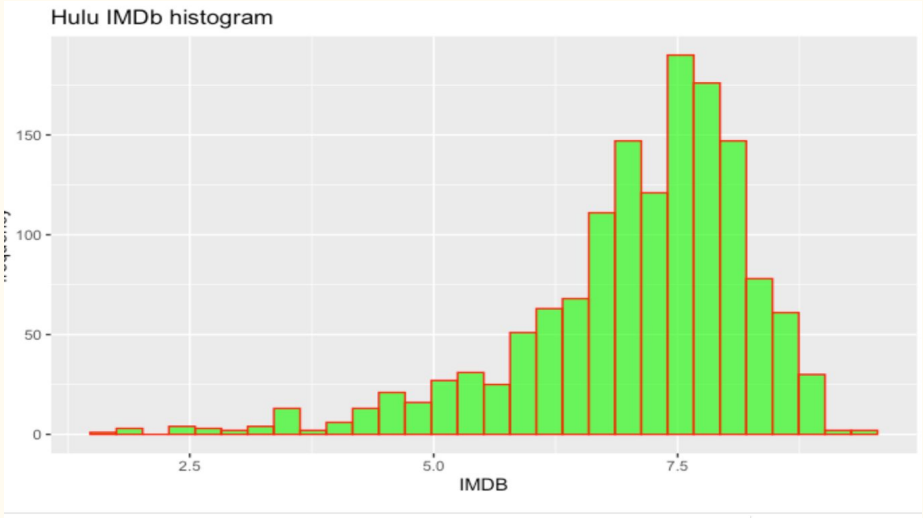
Normality check of the variable IMDb and Rotten.Tomatoes for Hulu

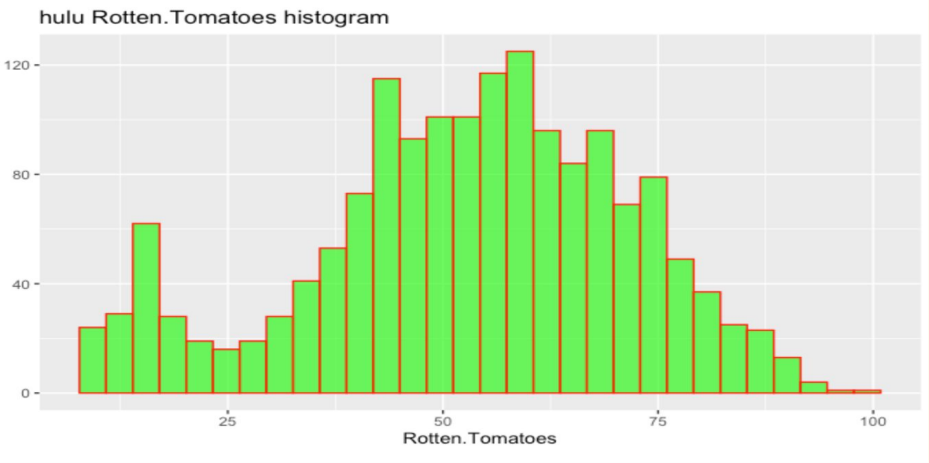
```
        Shapiro-Wilk normality test

data:  hulutv$IMDb
W = 0.91477, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  hulutv$Rotten.Tomatoes
W = 0.97572, p-value = 6.306e-16
```



# Exploratory Data Analysis

## Normality Test

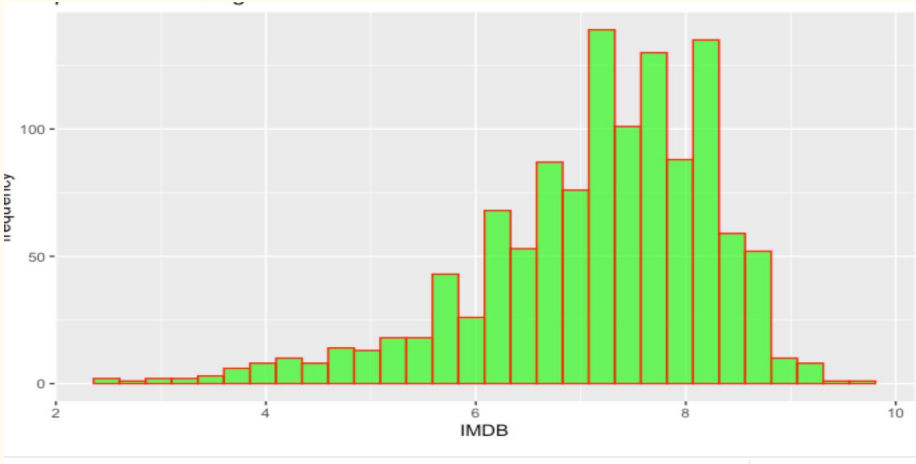
Normality check of the variable IMDb and Rotten.Tomatoes for Prime Videos

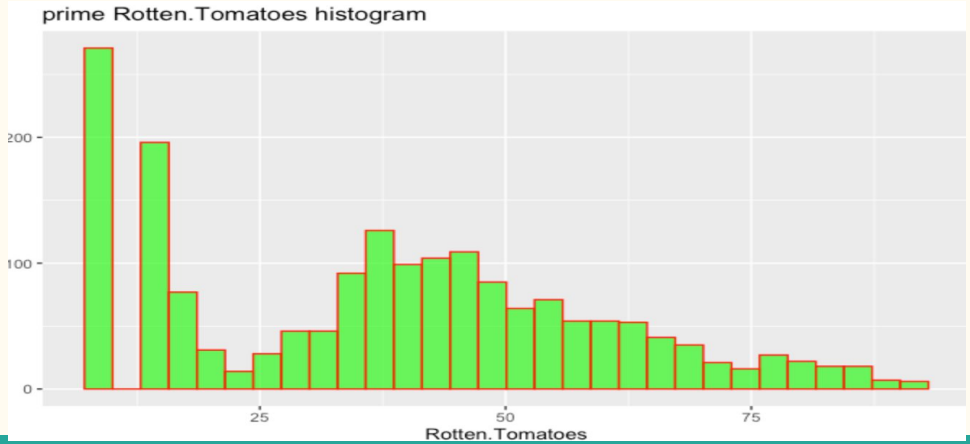
```
        Shapiro-Wilk normality test

data:  primetv$IMDb
W = 0.94083, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  primetv$Rotten.Tomatoes
W = 0.93943, p-value < 2.2e-16
```



# Exploratory Data Analysis

## Normality Test

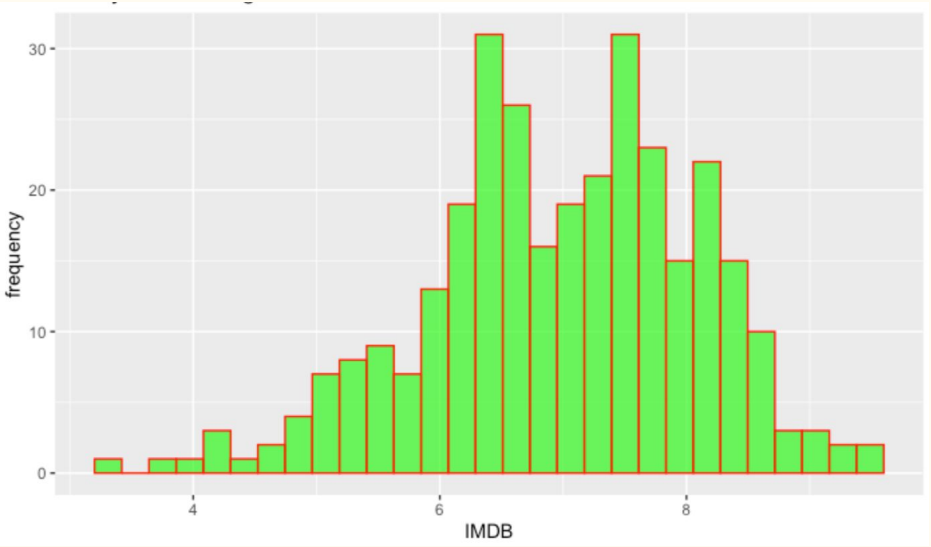
Normality check of the variable IMDb and Rotten.Tomatoes for Disney+

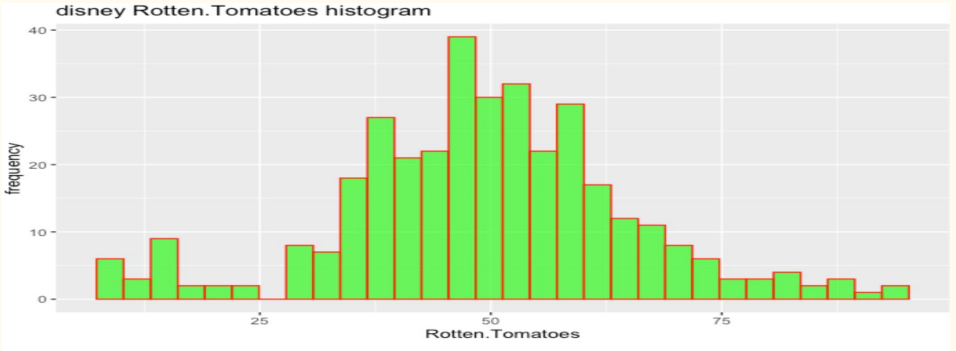
```
        Shapiro-Wilk normality test

data:  disneytv$IMDb
W = 0.98668, p-value = 0.005289
```

```
        Shapiro-Wilk normality test

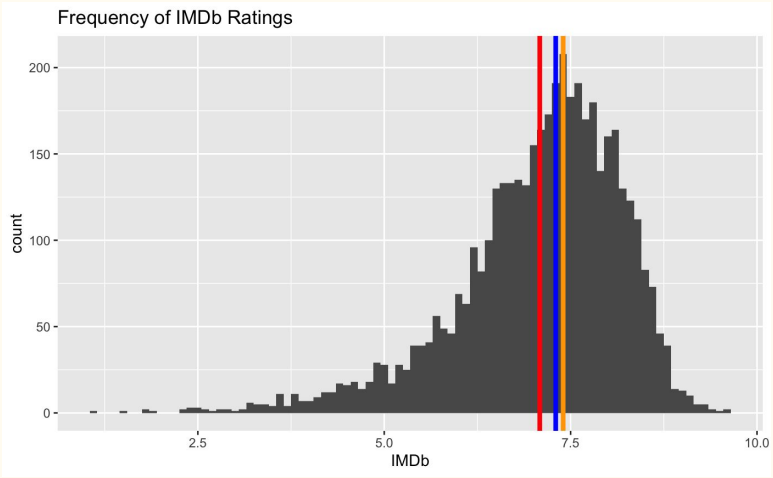
data:  disneytv$Rotten.Tomatoes
W = 0.98002, p-value = 8.655e-05
```



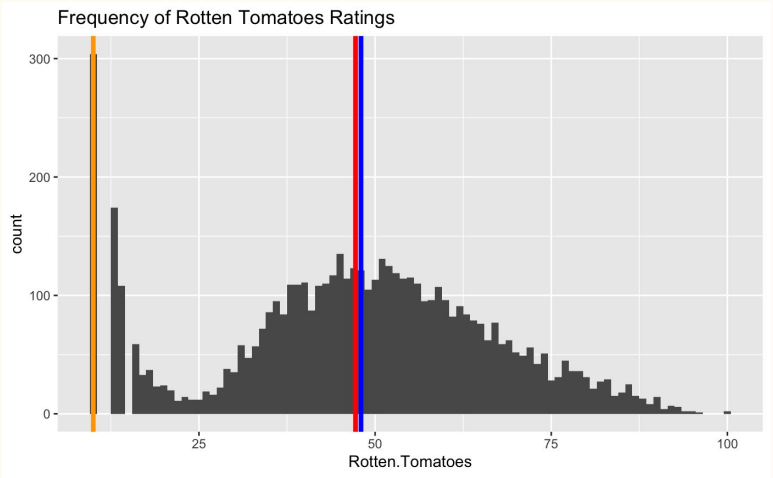
# Exploratory Data Analysis
## Rating Methods

IMDb has higher average rating than Rotten Tomatoes




Mean: 7.09 Median: 7.3 Mode: 7.4

Left Skewed

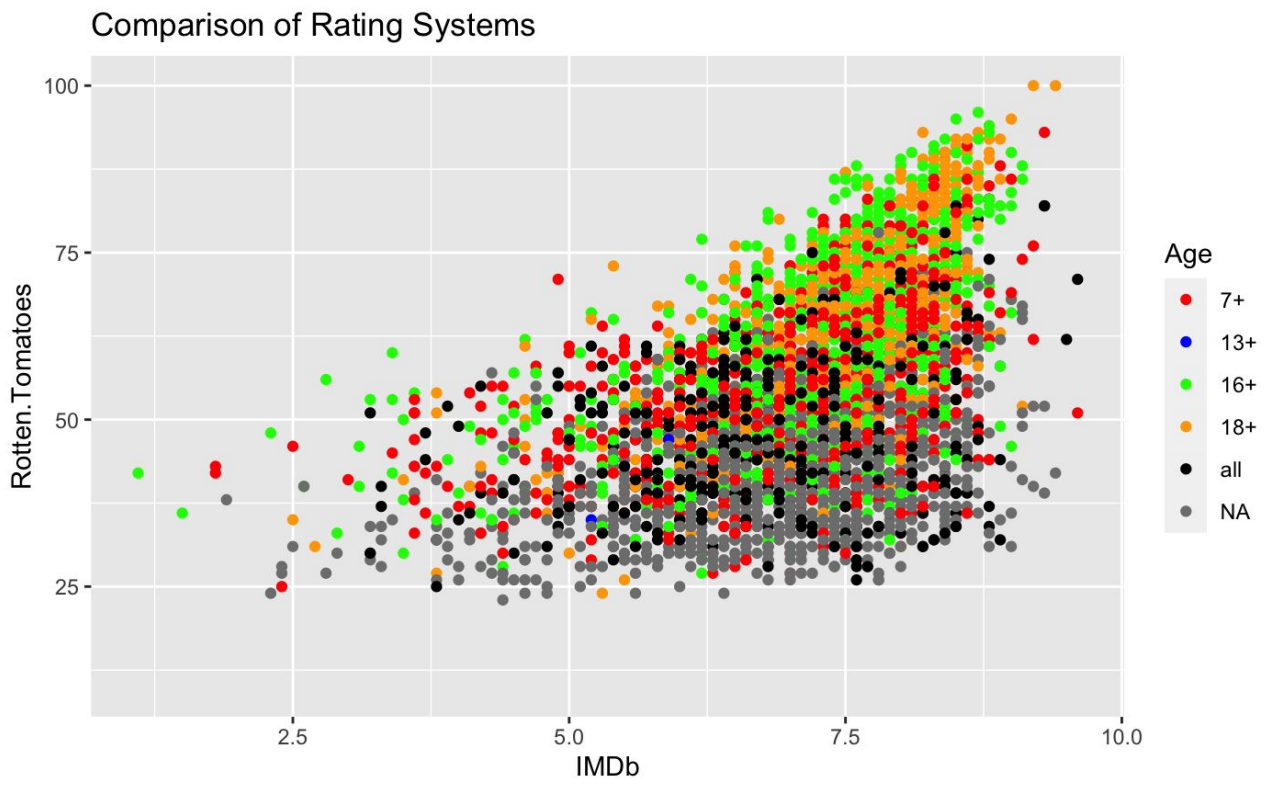



Mode: 10 Mean: 47.2 Median: 48

Outliers at lower ratings

# Exploratory Data Analysis

## Rating Methods and Age




7+: IMDb - 7.01
RT - 55.0

13+: IMDb - 6.83
RT - 54.2

16+: IMDb - 7.25
RT - 60.3

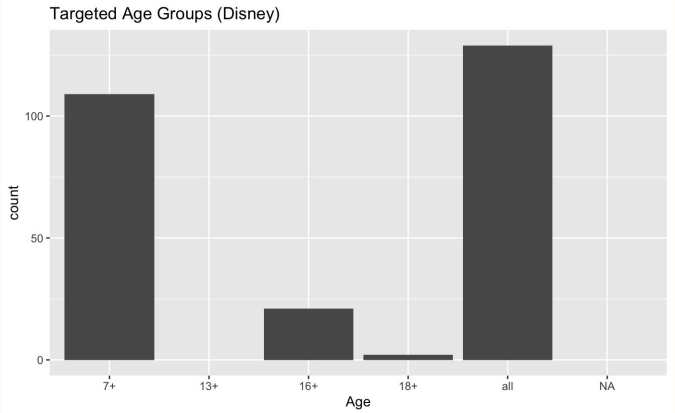
18+: IMDb - 7.30
RT - 62.7

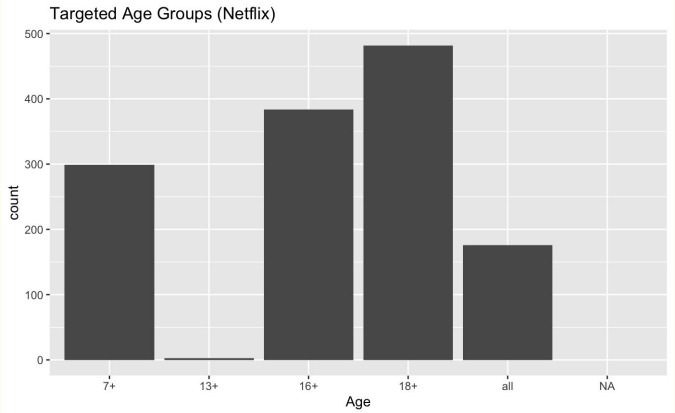
all: IMDb - 6.85
RT - 47.7

NA: IMDb - 6.96
RT - 31.7

# Exploratory Data Analysis
## Platforms and Age

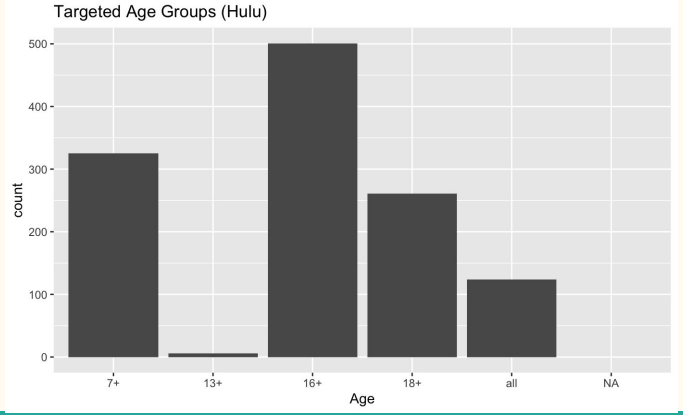
all: 36.8%

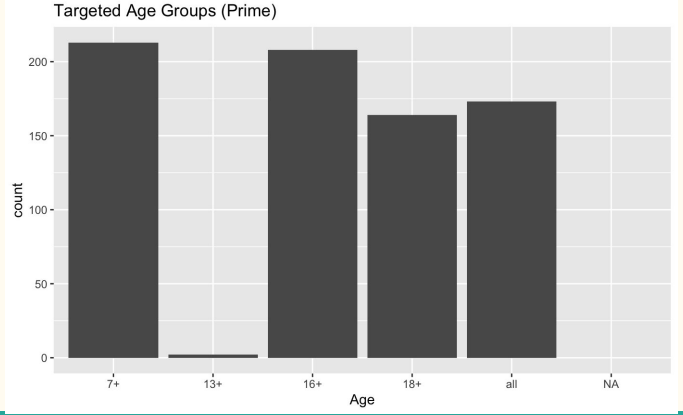



18+: 24.5%

18+: 30.9%

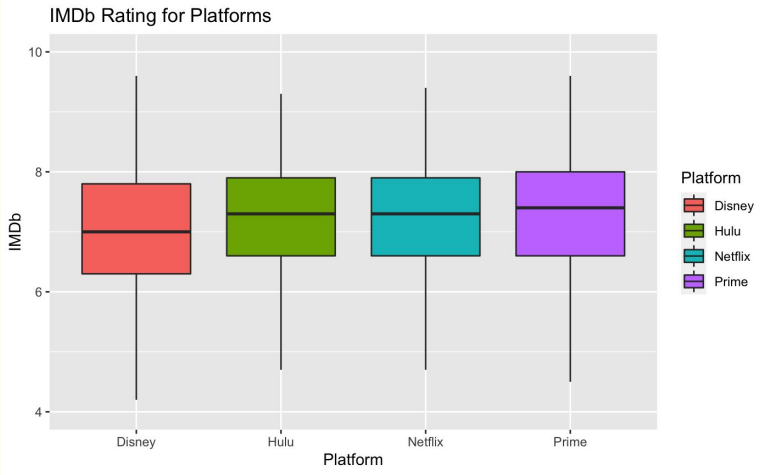


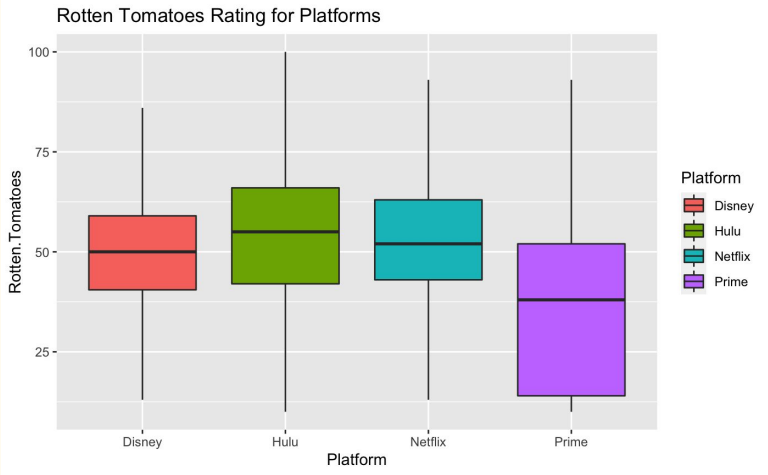




7+: 11.6%

# Exploratory Data Analysis

## Ratings for Platforms




6.97 7.08 7.11 7.15 ⟵ Mean ⟶ 49.4 52.8 53.6 37.8

# Distributions and tests

## T-test

Which streaming platform has the highest average rating (according to Rotten Tomatoes and IMDb)?

Prime Videos has the highest average Imdb rating  which is 7.152538. Also, Netflix has the highest average RTT rating  which is 53.559107

```
        Welch Two Sample t-test

data:  primetv$IMDb and primetv$Rotten.Tomatoes
t = -61.595, df = 1846.2, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -31.58340 -29.63418
sample estimates:
mean of x mean of y
 7.152538 37.761333


        Welch Two Sample t-test

data:  disneytv$IMDb and disneytv$Rotten.Tomatoes
t = -50.711, df = 353.85, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -44.09983 -40.80695
sample estimates:
mean of x mean of y
 6.971111 49.424501
```

```
        Welch Two Sample t-test

data:  netflixtv$IMDb and netflixtv$Rotten.Tomatoes
t = -134.4, df = 1989.8, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -47.12593 -45.77042
sample estimates:
mean of x mean of y
 7.110933 53.559107


        Welch Two Sample t-test

data:  hulutv$IMDb and hulutv$Rotten.Tomatoes
t = -98.133, df = 1634.6, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -46.66991 -44.84086
sample estimates:
mean of x mean of y
  7.08237  52.83775
```

# Distributions and tests

## ChiSquare Test

Test of Independence. To check whether IMDb and RRT is independent.

Pearson's Chi-squared test

data: contable1 (netflixtv$Age, netflixtv$IMDb)

X-squared = 377.75, df = 325, p-value = 0.02316

data: contable2(netflixtv$Age, netflixtv$Rotten.Tomatoes)

X-squared = 984.28, df = 415, p-value < 2.2e-16

data: contable3(hulutv$Age, hulutv$IMDb)

X-squared = 354.48, df = 355, p-value = 0.4978

data: contable4 (hulutv$Age, hulutv$Rotten.Tomatoes)

X-squared = 1216.2, df = 410, p-value < 2.2e-16

Pearson's Chi-squared test

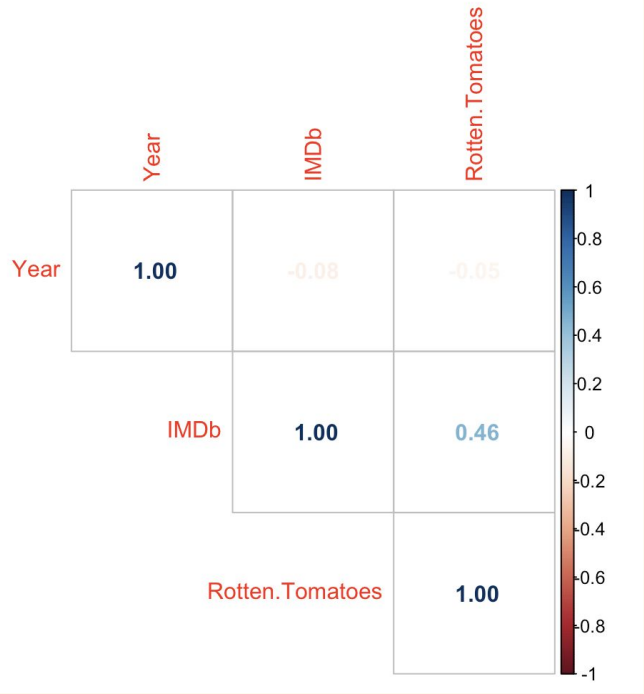
data: contable5(primetv$Age, primetv$IMDb)
X-squared = 405.74, df = 325, p-value = 0.001524

data: contable6(primetv$Age, primetv$Rotten.Tomatoes)
X-squared = 1632.5, df = 395, p-value < 2.2e-16

data: contable7(disneytv$Age, disneytv$IMDb)
X-squared = 325.04, df = 220, p-value = 5.189e-06

data: contable8(disneytv$Age, disneytv$Rotten.Tomatoes)
X-squared = 460.17, df = 272, p-value = 7.556e-12

# Model Building




```
Call:
lm(formula = IMDb ~ Rotten.Tomatoes, data = tvdata_rank)

Residuals:
    Min      1Q  Median      3Q     Max
-5.562 -0.495   0.087   0.628   2.746

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       5.15419    0.05785    89.1   <2e-16 ***
Rotten.Tomatoes   0.03590    0.00104    34.6   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.989 on 4404 degrees of freedom
Multiple R-squared:  0.213,     Adjusted R-squared:  0.213
F-statistic: 1.19e+03 on 1 and 4404 DF,  p-value: <2e-16
```

R2 equals 0.213, shows IMDb and Rotten.Tomatoes only have a weak correlation

# Model Building

The variable netflix, prime.video, disney and Age13+ are insignificant

There are only 9 shows in Age13+

```
```{r}
table(tvdata_rank$Age)
```
```

```
      13+  16+  18+   7+  all
1199    9  987  852  824  535
```

```
Call:
lm(formula = IMDb ~ ., data = tvdata_rank)

Residuals:
   Min     1Q Median     3Q    Max
-5.332 -0.485  0.082  0.624  2.750

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     16.42346    3.20303    5.13  3.1e-07 ***
Year            -0.00562    0.00159   -3.54   0.0004 ***
Age13+          -0.52233    0.32370   -1.61   0.1067
Age16+          -0.37963    0.04721   -8.04  1.1e-15 ***
Age18+          -0.44175    0.04963   -8.90  < 2e-16 ***
Age7+           -0.42070    0.04678   -8.99  < 2e-16 ***
Ageall          -0.32024    0.05253   -6.10  1.2e-09 ***
Rotten.Tomatoes  0.04376    0.00123   35.45  < 2e-16 ***
Netflix1        -0.10072    0.05389   -1.87   0.0617 .
Hulu1           -0.23752    0.05294   -4.49  7.4e-06 ***
Prime.Video1     0.09660    0.05321    1.82   0.0695 .
Disney.1        -0.12233    0.07458   -1.64   0.1010
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.965 on 4394 degrees of freedom
Multiple R-squared:  0.252,     Adjusted R-squared:  0.25
F-statistic:  135 on 11 and 4394 DF,  p-value: <2e-16
```

# Model Building

Now, the model looks better, and value of vif is not too high so there is no problem of multicollinearity.

VIFs of the model

| Age16+ | Age18+ | Age7+ | Ageall |
|---|---|---|---|
| 1.11 | 1.83 | 1.8 | 1.55 |

| Hulu1 | Rotten.Tomatoes | Year |
|---|---|---|
| 1.32 | | 1.43 | 1.12 |

```
Call:
lm(formula = IMDb ~ ., data = tvdata_rank_no13)

Residuals:
   Min     1Q Median     3Q    Max
-5.397 -0.495  0.084  0.614  2.770

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     21.68127    3.03570    7.14  1.1e-12 ***
Year            -0.00821    0.00150   -5.46  5.0e-08 ***
Age16+          -0.37745    0.04734   -7.97  2.0e-15 ***
Age18+          -0.44184    0.04959   -8.91  < 2e-16 ***
Age7+           -0.43639    0.04660   -9.36  < 2e-16 ***
Ageall          -0.34673    0.05143   -6.74  1.8e-11 ***
Rotten.Tomatoes  0.04269    0.00122   35.11  < 2e-16 ***
Hulu1           -0.20879    0.03306   -6.32  3.0e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.969 on 4389 degrees of freedom
Multiple R-squared:  0.247,     Adjusted R-squared:  0.246
F-statistic:  205 on 7 and 4389 DF,  p-value: <2e-16
```

# Model Building

Now I add the interaction between Rotten.Tomatoes and Age into the model.

The overall model is significant as well and the adjusted R-squared is 0.249, a little higher than before.

```
Call:
lm(formula = IMDb ~ . + Rotten.Tomatoes:Age, data = tvdata_rank_no13)

Residuals:
   Min     1Q Median     3Q    Max
-5.453 -0.493  0.072  0.607  2.799

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              23.464467   3.068632    7.65  2.5e-14 ***
Year                     -0.009186   0.001525   -6.02  1.9e-09 ***
Age16+                    0.005813   0.189196    0.03   0.9755
Age18+                   -0.453837   0.199101   -2.28   0.0227 *
Age7+                    -0.516497   0.198813   -2.60   0.0094 **
Ageall                    0.609364   0.235339    2.59   0.0096 **
Rotten.Tomatoes           0.046692   0.002810   16.62  < 2e-16 ***
Hulu1                    -0.209429   0.033037   -6.34  2.5e-10 ***
Age16+:Rotten.Tomatoes   -0.007472   0.003620   -2.06   0.0391 *
Age18+:Rotten.Tomatoes   -0.001010   0.003700   -0.27   0.7849
Age7+:Rotten.Tomatoes     0.000526   0.003932    0.13   0.8936
Ageall:Rotten.Tomatoes   -0.020359   0.004954   -4.11  4.0e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.966 on 4385 degrees of freedom
Multiple R-squared:  0.251,     Adjusted R-squared:  0.249
F-statistic:  134 on 11 and 4385 DF,  p-value: <2e-16
```

# Model Building

Compared with the 3 models, the one with interaction is the best (although it's still a weak correlation)

```
Analysis of Variance Table

Model 1: IMDb ~ Rotten.Tomatoes
Model 2: IMDb ~ Year + Age + Rotten.Tomatoes + Hulu
Model 3: IMDb ~ Year + Age + Rotten.Tomatoes + Hulu + Rotten.Tomatoes:Age
  Res.Df  RSS Df Sum of Sq     F  Pr(>F)
1   4395 4304
2   4389 4117  6     186.9 33.36 < 2e-16 ***
3   4385 4093  4      23.4  6.26 5.1e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

IMDb = 23.46 + -0.0092 * Year + 0.047 * Rotten.Tomatoes + -0.21 * Hulu

+ (-0.52 + 0.00053 * Rotten.Tomatoes) * (Age7+)

+ (0.0058 + -0.0075 * Rotten.Tomatoes) * (Age16+)

+ (-0.45 + -0.0010* Rotten.Tomatoes) * (Age18+)

+ (0.61 + -0.020 * Rotten.Tomatoes) * Ageall

# Conclusions

- Targeted age groups highly dependent on streaming platform
  - Specific correlation shown between Hulu and Age
- Very few shows produced in the 20th century, most produced in 2017
  - Age can be directly related to the RT rating using a linear model
- IMDb and Rotten Tomatoes disagree about the highest rated platforms
  - Prime is highest for IMDb, Hulu (median) and Netflix (mean) for RT
- IMDb and Rotten Tomatoes only have a weak correlation
  - IMDb tends to give higher ratings than RT
  - Both show normality, IMDb is more left skewed and RT has low-value outliers



# Questions?