Process_notebook

2022-06-25

Abstract

- In today's society, streaming platforms have become a worldwide trend in mainstream media, where a good majority of people have owned at least one subscription in their day-to-day lives. Whether it is from Netflix, Hulu, Amazon Prime, or Disney+, there are a boatload of streaming platforms catered to every audience member's demographic and personal preferences. This project explores the thousands of movies and TV shows available across said platforms and how age demographics, specific actors, genres, and the actual amount of content comes into play using the particular coding techniques we have discussed throughout the course. Some techniques covered include, but are not limited to, hypothesis testing, tackling the dplyr package, and data visualization.

Overview and Motivation

- The goal of our project is to discover which platforms provide content geared towards specific age groups or are content heavy in one specific genre, discover which platforms would be best fit for a person that prefers access to more movies or TV shows, and if the actors in movies leads to a better IMDb score. We felt that streaming platforms and movies play such a huge role in the lives of so many people in today's society and is very relatable and well understood by many. The questions that we answered in this project were ones that we specifically were interested in which was also a huge motivation for us since it was an enjoyable topic to look into, and we felt that it could help many others when trying to answer these same questions when deciding which streaming platform to purchase a subscription to.

Related Work

- We were inspired to dedicate our project to this topic because like mentioned before, we have all used a subscription for a respective streaming platform and wanted to find a topic that we all can relate to. We were also inspired by this article, "https://www.theguardian.com/tv-and-radio/2022/jun/21/netflix-and-bills-which-streaming-services-are-really-worth-shelling-out-for" which talks about the breakdown of each streaming service and what each platform has to offer in terms of content and cost.

Initial Questions

- Initially we were interested in answering the following questions:
  1. Which streaming platform is best for each age demographic?
  2. What are the common features of top Rated shows/Movies?
  3. Are movies and TV shows always improving its quality through the time? (See average rate from two platforms by year)
  4. Are there certain types of movies on different platforms (action, drama, comedy, etc)?
  5. Do certain streaming platforms have more movies vs TV shows or vice versa? Along these lines do certain platforms have more original content than others?
- These questions evolved into the following:
  1.Do the streaming platforms gear their content towards a certain age group more than others?
  2.Does an "A list" actor always guarantee a higher rating on IMDb? Do some platforms have more content with "A list" actors?
  3. Are movies and TV shows always improving its quality through the time? (See average rate from two platforms by year)
  4. Do certain platforms contain more movies in specific genres than other genres?
  5. Do certain streaming platforms have more movies vs TV shows or vice versa?
- When first exploring the possibility of questions to answer we tried to be extremely broad in order to give ourselves the chance to narrow down the scope of our research and not limit ourselves from the start.
- Our first question we decided to change the wording to allow us to properly use the data we had. We realized that we would not be able to conclude anything having to do with the popularity of platforms with specific age groups, but rather we had the information about which age groups specific platforms carried more or less content for.
- Our second question, "What are the common features of top Rated shows/Movies?" we realized was extremely broad and we were not going to be able to complete a factor analysis based on the scope of this course, so we decided to look specifically at how "A List" actors affect the IMDb rating of movies and how many movies/shows containing "A List" actors each platform carries.
- All we did was reword our question regarding the genre of movies to make it more clear what we were trying to answer as at first it was a bit difficult to know whether we were comparing one platform to another or comparing the different genres within each platform.
- Our last question we realized when it came to original content it was extremely difficult to find a dataset that differentiated between what is original content on specific platforms as there are so many production companies, and when original movies or shows are created for a platform any number of production companies could have been involved, and a platform such as Disney+ most of the content is technically considered original.

Data

- The following data was found on Kaggle and we were able to download the two .csv files directly to our computer and upload onto R studio.
  https://www.kaggle.com/code/ruchi798/movies-and-tv-shows-eda/data <- Movie and TV Shows Data
  https://www.kaggle.com/datasets/wrandrall/imdb-new-dataset?resource=download <- IMDb Database
- We found the data with information regarding actors from IMDb.
  https://www.imdb.com/list/ls058011111/ <- A-list actors

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
tv_show <- read.csv(file = "tv_shows.csv")
head(tv_show)
```

```
##   X ID                      Title Year Age   IMDb Rotten.Tomatoes Netflix Hulu
## 1 0  1              Breaking Bad 2008 18+ 9.4/10         100/100       1    0
## 2 1  2          Stranger Things 2016 16+ 8.7/10          96/100       1    0
## 3 2  3          Attack on Titan 2013 18+ 9.0/10          95/100       1    1
## 4 3  4        Better Call Saul 2015 18+ 8.8/10          94/100       1    0
## 5 4  5                    Dark 2017 16+ 8.8/10          93/100       1    0
## 6 5  6 Avatar: The Last Airbender 2005  7+ 9.3/10          93/100       1    0
##   Prime.Video Disney. Type
## 1           0       0    1
## 2           0       0    1
## 3           0       0    1
## 4           0       0    1
## 5           0       0    1
## 6           1       0    1
```

```
movie <- read.csv(file = "MoviesOnStreamingPlatforms.csv")
head(movie)
```

```
##   X ID                                  Title Year Age Rotten.Tomatoes
## 1 0  1                            The Irishman 2019 18+          98/100
## 2 1  2                                 Dangal 2016  7+          97/100
## 3 2  3 David Attenborough: A Life on Our Planet 2020  7+          95/100
## 4 3  4        Lagaan: Once Upon a Time in India 2001  7+          94/100
## 5 4  5                                   Roma 2018 18+          94/100
## 6 5  6        To All the Boys I've Loved Before 2018 13+          94/100
##   Netflix Hulu Prime.Video Disney. Type
## 1       1    0           0       0    0
## 2       1    0           0       0    0
## 3       1    0           0       0    0
## 4       1    0           0       0    0
```

```
## 5       1    0           0       0    0
## 6       1    0           0       0    0
```

```
imdb <- read.csv(file = "imdb_database.csv") %>%
  rename(Title=Movie.Name) #only movie type column is used

Alist <- read.csv(file = "Top 1000 Actors and Actresses.csv") %>%
  select(Name, Known.For) %>% rename(Title=Known.For)
head(Alist)
```

```
##                Name              Title
## 1    Robert De Niro        Raging Bull
## 2    Jack Nicholson          Chinatown
## 3     Marlon Brando     Apocalypse Now
## 4 Denzel Washington             Fences
## 5 Katharine Hepburn The Lion in Winter
## 6   Humphrey Bogart         Casablanca
```

Below we tidy our data. (We specify what each step is within the code)

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v stringr 1.4.0
## v tidyr   1.2.0     v forcats 0.5.1
## v readr   2.1.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
#Change IMDb and Rotten.Tomatoes column to be numeric
TV_shows <- tv_show %>%
  separate(IMDb, c("IMDb", "I-fullRate"), sep = "/") %>%
  separate(Rotten.Tomatoes, c("Rotten.Tomatoes", "R-fullRate"), sep = "/") %>%
  mutate(IMDb = as.numeric(IMDb)/10,
         Rotten.Tomatoes = as.numeric(Rotten.Tomatoes)/100) %>%
  select(c(-1,-2,-7,-9)) %>% mutate(MT = "TV")

Movie <- movie %>%
  separate(Rotten.Tomatoes, c("Rotten.Tomatoes", "R-fullRate"), sep = "/") %>%
  mutate(Rotten.Tomatoes = as.numeric(Rotten.Tomatoes)/100) %>%
  select(-1,-2,-7) %>% mutate(MT = "Movie")

data <-
  bind_rows(TV_shows, Movie) %>%              #Combine two data sets
  rename(Disney=Disney.) %>%                  #Change the name of "Disney." to "Disney"
  mutate(Age = ifelse(Age=="", NA, Age)) %>% #Add NA to missing values in Age column
  pivot_longer(6:9, names_to = "Platform", values_to = "OnPlatform") %>% #Tidy data
  filter(OnPlatform == 1) %>%
  select(-9)

head(data)
```

```
## # A tibble: 6 x 8
##   Title             Year Age    IMDb Rotten.Tomatoes  Type MT    Platform
```

```
##   <chr>            <int> <chr> <dbl>           <dbl> <int> <chr> <chr>
## 1 Breaking Bad      2008 18+    0.94             1        1 TV    Netflix
## 2 Stranger Things   2016 16+    0.87             0.96     1 TV    Netflix
## 3 Attack on Titan   2013 18+    0.9              0.95     1 TV    Netflix
## 4 Attack on Titan   2013 18+    0.9              0.95     1 TV    Hulu
## 5 Better Call Saul  2015 18+    0.88             0.94     1 TV    Netflix
## 6 Dark              2017 16+    0.88             0.93     1 TV    Netflix
```

Exploratory Data Analysis

1. Do the streaming platforms gear their content towards a certain age group more than others?

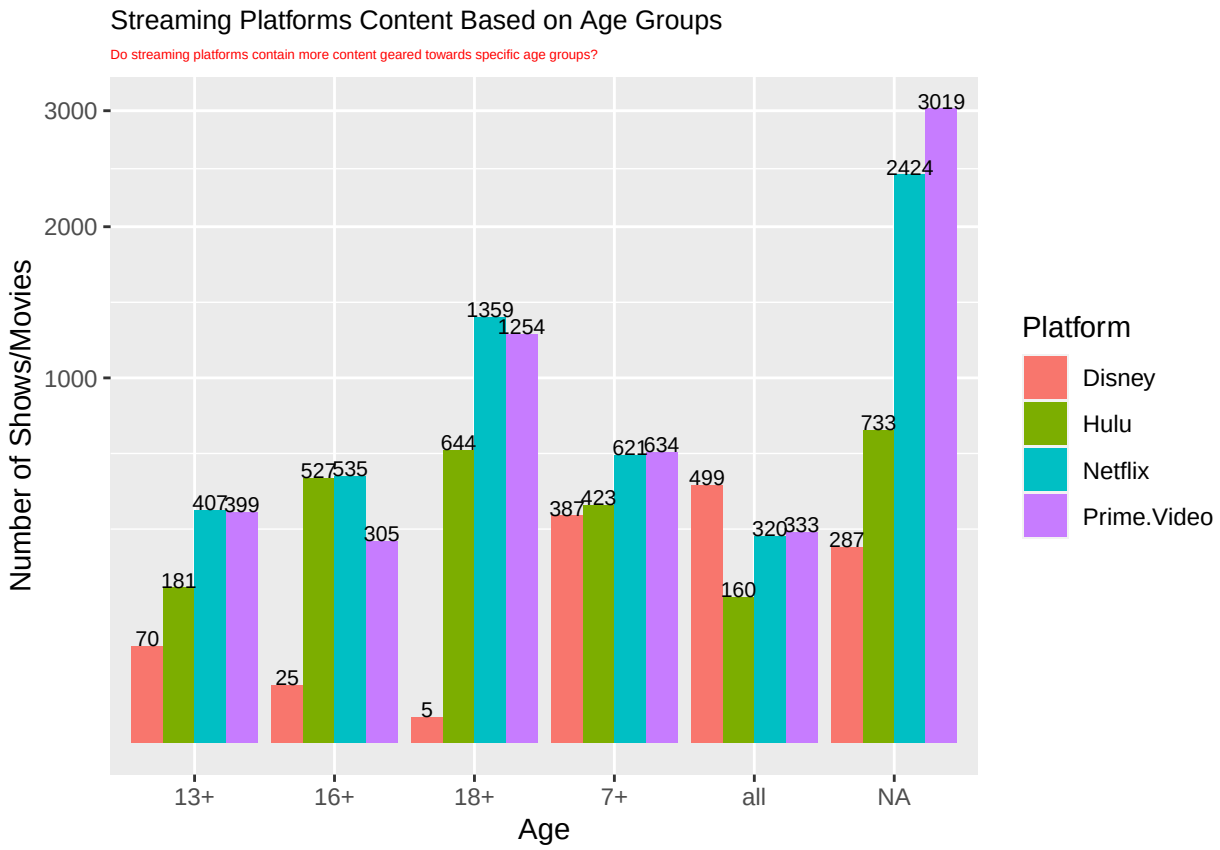
    - To explore our first question we knew we wanted to plot the data in some sort of graphic as it would be best to actually visualize the information. We found our best option was to create a bargraph. We were able to group our data by both Age and Platform to use summarise to count how many different TV shows and Movies each platform has for each age group. The Age variable is a categorical variable with categories: "13+", "16+", "18+", "all ages". There are also some NA values which we chose to omit here since it did not add any value to our analysis as we needed to know the specific age categories.
    - Taxonomy for Data Graphics:

Visual Cues: we used color to represent each platform, position to show which platforms had more or less content geared towards a specific age group, and length to show how much content each platform had in each age group category.
Coordinate system: Cartesian
Scale:x-axis: categorical (age groups), y-axis: linear (amount of shows/movies in each age category), color: categorical (each platform)
Context: We have a title, subtitle, labeled axes, a labeled legend, and we have the exact number for each bar labeled at the top.

```
# Question 1
agedemo = data %>% group_by(Age, Platform) %>% summarise(n = n()) %>% mutate(prop = n/sum(n))
```

```
## `summarise()` has grouped output by 'Age'. You can override using the `.groups`
## argument.
```

```
agedemo %>% ggplot(aes(x = Age, y=n, fill = Platform)) +
  geom_bar(position = "dodge", stat = 'identity') +
  scale_y_sqrt() +
  geom_text(
    aes(label = n, y = n + 0.05),
    position = position_dodge(0.9),
    vjust = 0,
    size = 7.5 / .pt
  ) + labs(title = "Streaming Platforms Content Based on Age Groups ", subtitle = "Do streaming platfor
  theme(plot.title = element_text(size = 10),
        plot.subtitle = element_text(size = 5, color = "red")
        )
```



- Disney has the most "family friendly" content with most of their content being for all ages or 13+, whereas there are very few content geared towards ages 16+ or 18+.
- Netflix has the most content geared towards individuals 18+, so it might not make as much sense for families with very young children to have a Netflix subscription vs a Disney subscription, however Netflix does have more content for younger ages than platforms Hulu and Prime Video.
- Prime Video has content mostly 18+ similar to Netflix, so for individuals that are older, or families with older children it would make more sense to own a Prime Video subscription, but there isn't as much content in this category as Netflix.
- Hulu has content mostly geared towards ages 16+ and 18+, so once again it would not make as much sense for families with young children to have a Hulu subscritption as it would for these families to have a Disney subscription. Hulu has the least content for 13+ and all ages than all the other platforms.

2. Does an "A list" actor always guarantee a higher rating on IMDb? Do some platforms have more content with "A list" actors?

```
library(stringr)
# Tidying up the data to get rid of the brackets and quotations on the Actor's names
imdbact <- imdb %>% select(Title, Actors, Score) %>% mutate(Actors = str_replace_all(imdb$Actors, c("\\

# combining the imdb, Alist actors, and TV shows/Movies databases into one table based off their Movie
imdb_alist <- left_join(imdbact, Alist, by = "Title") %>% left_join(., data, by = "Title") %>% group_by
head(imdb_alist)
```

```
## # A tibble: 6 x 7
## # Groups:   Title [5]
##   Title            Actors                    Name  Score MT    Platform is_Alist
##   <chr>            <chr>                     <chr> <dbl> <chr> <chr>    <lgl>
## 1 Fight Club       Brad Pitt, Edward Norton~ Edwa~   8.8 Movie Prime.V~ TRUE
## 2 Seven            Morgan Freeman, Brad Pit~ <NA>    8.6 Movie Netflix  FALSE
## 3 Breaking Bad     Bryan Cranston, Aaron Pa~ Brya~   9.5 TV    Netflix  TRUE
## 4 Breaking Bad     Bryan Cranston, Aaron Pa~ Aaro~   9.5 TV    Netflix  TRUE
## 5 Django Unchained Jamie Foxx, Christoph Wa~ Chri~   8.4 Movie Netflix  TRUE
## 6 Batman Begins    Christian Bale, Michael ~ Kati~   8.2 Movie Hulu     TRUE
```

```
# Linear Regression
 # For just movies
Alist_movie <- imdb_alist %>% filter(MT == "Movie")
mod2 <- lm(Score ~ is_Alist, data = Alist_movie)
summary(mod2)$coefficients
```

```
##              Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)   6.22805 0.01836593 339.10899  0.000000e+00
## is_AlistTRUE  1.74375 0.04755858  36.66531 1.772742e-267
```

```
summary(mod2)
```

```
##
## Call:
## lm(formula = Score ~ is_Alist, data = Alist_movie)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9718 -0.8281  0.1719  0.8282  3.7719
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.22805    0.01837  339.11   <2e-16 ***
## is_AlistTRUE  1.74375    0.04756   36.66   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.36 on 6442 degrees of freedom
## Multiple R-squared:  0.1727, Adjusted R-squared:  0.1725
## F-statistic:  1344 on 1 and 6442 DF,  p-value: < 2.2e-16
```

```
qt(0.05/2, 6442, lower.tail=FALSE)
```

```
## [1] 1.960332
```

```r
t.test(Alist_movie$Score, Alist_movie$is_Alist, alternative = "greater", conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  Alist_movie$Score and Alist_movie$is_Alist
## t = 331.1, df = 7172.3, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  6.307471      Inf
## sample estimates:
## mean of x mean of y
##  6.488097  0.149131
```

```r
 # For just TV shows
Alist_TV <- imdb_alist %>% filter(MT == "TV")
Alist_TV
```

```
## # A tibble: 4,082 x 7
## # Groups:   Title [1,890]
##    Title        Actors                        Name  Score MT    Platform is_Alist
##    <chr>        <chr>                         <chr> <dbl> <chr> <chr>    <lgl>
##  1 Breaking Bad Bryan Cranston, Aaron Paul,~ Brya~   9.5 TV    Netflix  TRUE
##  2 Breaking Bad Bryan Cranston, Aaron Paul,~ Aaro~   9.5 TV    Netflix  TRUE
##  3 Titanic      Leonardo DiCaprio, Kate Win~ Kate~   7.8 TV    Prime.V~ TRUE
##  4 Titanic      Leonardo DiCaprio, Kate Win~ Glor~   7.8 TV    Prime.V~ TRUE
##  5 Titanic      Leonardo DiCaprio, Kate Win~ Bill~   7.8 TV    Prime.V~ TRUE
##  6 Breaking Bad Bryan Cranston, Aaron Paul,~ Brya~   9.5 TV    Netflix  TRUE
##  7 Breaking Bad Bryan Cranston, Aaron Paul,~ Aaro~   9.5 TV    Netflix  TRUE
##  8 Titanic      Leonardo DiCaprio, Kate Win~ Kate~   7.8 TV    Prime.V~ TRUE
##  9 Titanic      Leonardo DiCaprio, Kate Win~ Glor~   7.8 TV    Prime.V~ TRUE
## 10 Titanic      Leonardo DiCaprio, Kate Win~ Bill~   7.8 TV    Prime.V~ TRUE
## # ... with 4,072 more rows
```

```r
mod3 <- lm(Score ~ is_Alist, data = Alist_TV)
summary(mod3)$coefficients
```

```
##              Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)  6.592843 0.02287993 288.14958  0.000000e+00
## is_AlistTRUE 1.678229 0.04367996  38.42103 6.136821e-276
```

```r
summary(mod3)
```

```
##
## Call:
## lm(formula = Score ~ is_Alist, data = Alist_TV)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4711 -0.4711  0.0072  1.1072  3.1072
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.59284    0.02288  288.15   <2e-16 ***
## is_AlistTRUE  1.67823    0.04368   38.42   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.245 on 4080 degrees of freedom
## Multiple R-squared:  0.2657, Adjusted R-squared:  0.2655
## F-statistic:  1476 on 1 and 4080 DF,  p-value: < 2.2e-16
```

```
qt(0.05/2, 4080, lower.tail = FALSE)
```

```
## [1] 1.960546
```

```
t.test(Alist_TV$Score, Alist_TV$is_Alist, alternative = "greater", conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  Alist_TV$Score and Alist_TV$is_Alist
## t = 284.95, df = 4844.1, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  6.739794      Inf
## sample estimates:
## mean of x mean of y
## 7.0533072 0.2743753
```

```
# Null Hypothesis: A movie with an A list actor rates the same as a movie/TV show without an A list act
# Alt Hypothesis: A movie with an A list actor rates higher than a movie/TV show without an A list acto


# table with the counts for each platform
with(imdb_alist, table(Platform))
```

```
## Platform
##      Disney        Hulu     Netflix Prime.Video
##         793        1774        3985        3974
```

- Based on the linear regression and comparing the t-score to the critical t-score, we can conclude that there is evidence that there is a statistical significance when a movie/TV show that contains an A-list actor generates a higher score than a movie/TV show without one. We can also see that this same conclusion can be made when looking at summary of our linear regression models. The p-values, when looking at both movies and TV shows, are the same and are much smaller than an alpha level of 0.001 allowing us to reject the null hypothesis.
- As shown in the table, we are able to see how many movies/TV shows contain an A-list actor on each platform, both Netflix and Prime Video have close to 4,000 movies and shows with an A-list actor, followed be Hulu around 1700 and Disney around 800.

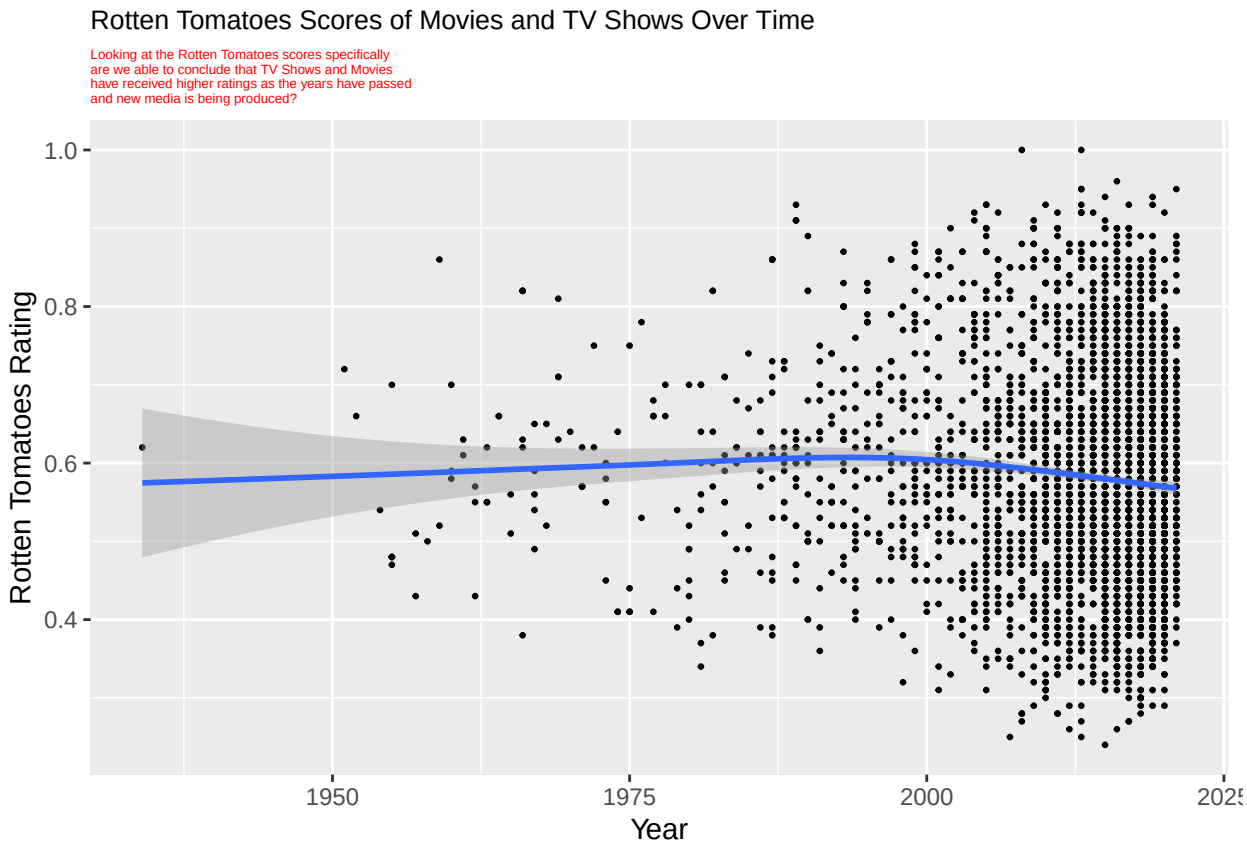
3. Are movies and TV shows always improving its quality through the time? (See average rate from two platforms by year)

    - Our third question we knew off the bat we would need to definitely visualize any trends in the ratings of movies and TV shows. The first thing we realized when looking into this question further and our dataset is that our movies data only had Rotten Tomato scores, not IMDb scores, yet our TV show data had both IMDb and Rotten Tomato scores which gave us an inclining that we might have to do multiple graphics. We knew a scatterplot here was definitely our best bet since we had two quantitative variables to be visualizing. Once the scatterplot was created we realized we had such a large sample of movies and TV shows that we could not see any specific trends without using a smoother, so we added one in and were able to make general conclusions. We then tried to narrow down our sample size by splitting up movies and TV shows, that way we could also look at both the IMDb and Rotten Tomato scores for the TV shows. After doing this we still ran into an issue with the datasets being too large. To fix this issue, we decided to look at the average ratings for each year for movies and TV shows individually. This helped narrow down the sample size to be able to draw more specific conclusions about the trends in scores (or in our case the lack thereof).
    - Taxonomy for All Data Graphics in This Section:

Visual Cues: We used position to show based on the year a movie or show came out where it fell on the rating system and we used a smoother to show direction of the trend of how the ratings of movies and TV shows have changed over time.
Coordinate system: Cartesian
Scale: x-axis: linear (years), y-axis: linear (possible ratings)
Context: We have a title, subtitle, and labeled axes.

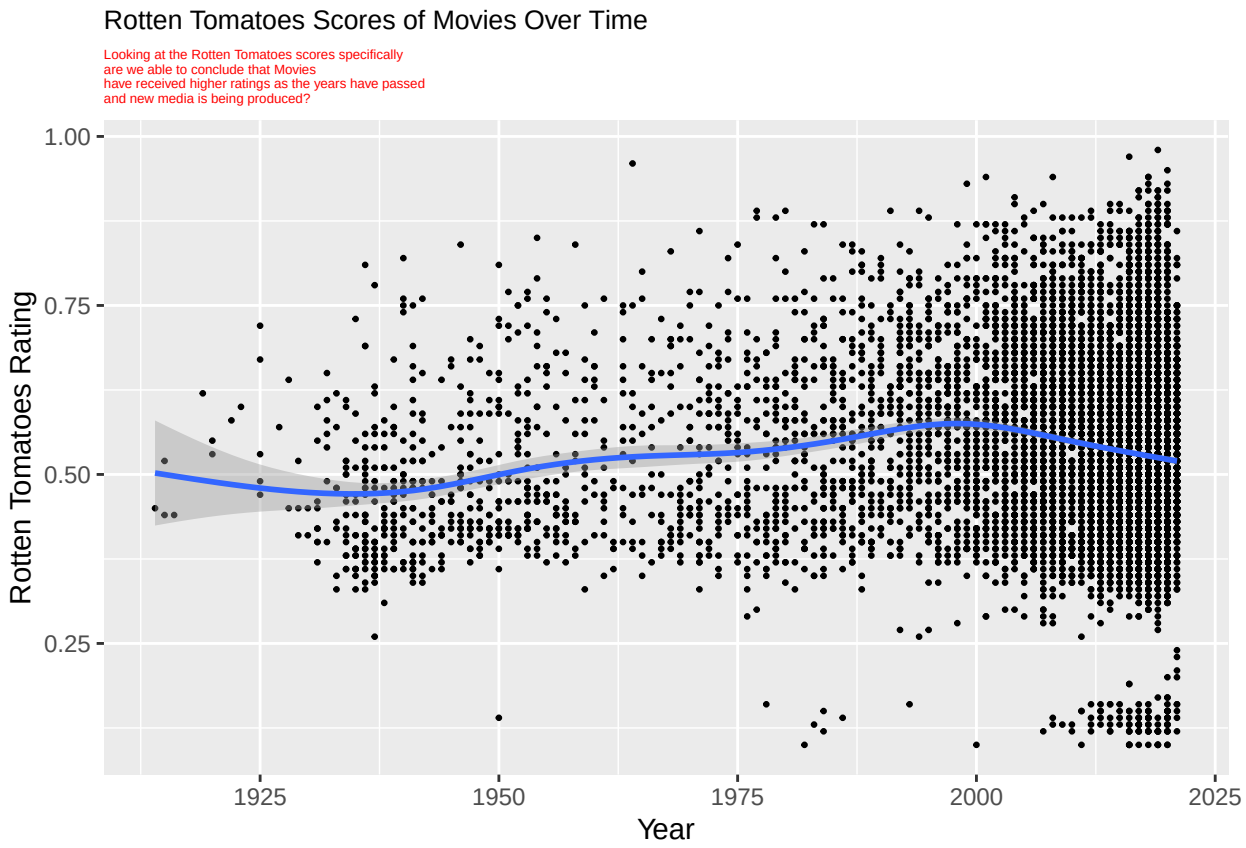
```
# Question 3
library(ggplot2)
##Rotten Tomato Scores for both Movies and TV Shows
ratings = data %>%
  group_by(Year, Rotten.Tomatoes) %>%
  arrange(Year) %>%
  na.omit()
p1 = ratings %>% ggplot(aes(x = Year, y = Rotten.Tomatoes)) +
  geom_point(size=0.5) +
  geom_smooth() + labs(title = "Rotten Tomatoes Scores of Movies and TV Shows Over Time", subtitle = "L
are we able to conclude that TV Shows and Movies
have received higher ratings as the years have passed
and new media is being produced?") + ylab("Rotten Tomatoes Rating")+
  theme(plot.title = element_text(size = 10),
        plot.subtitle = element_text(size = 5, color = "red")
        )
p1
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



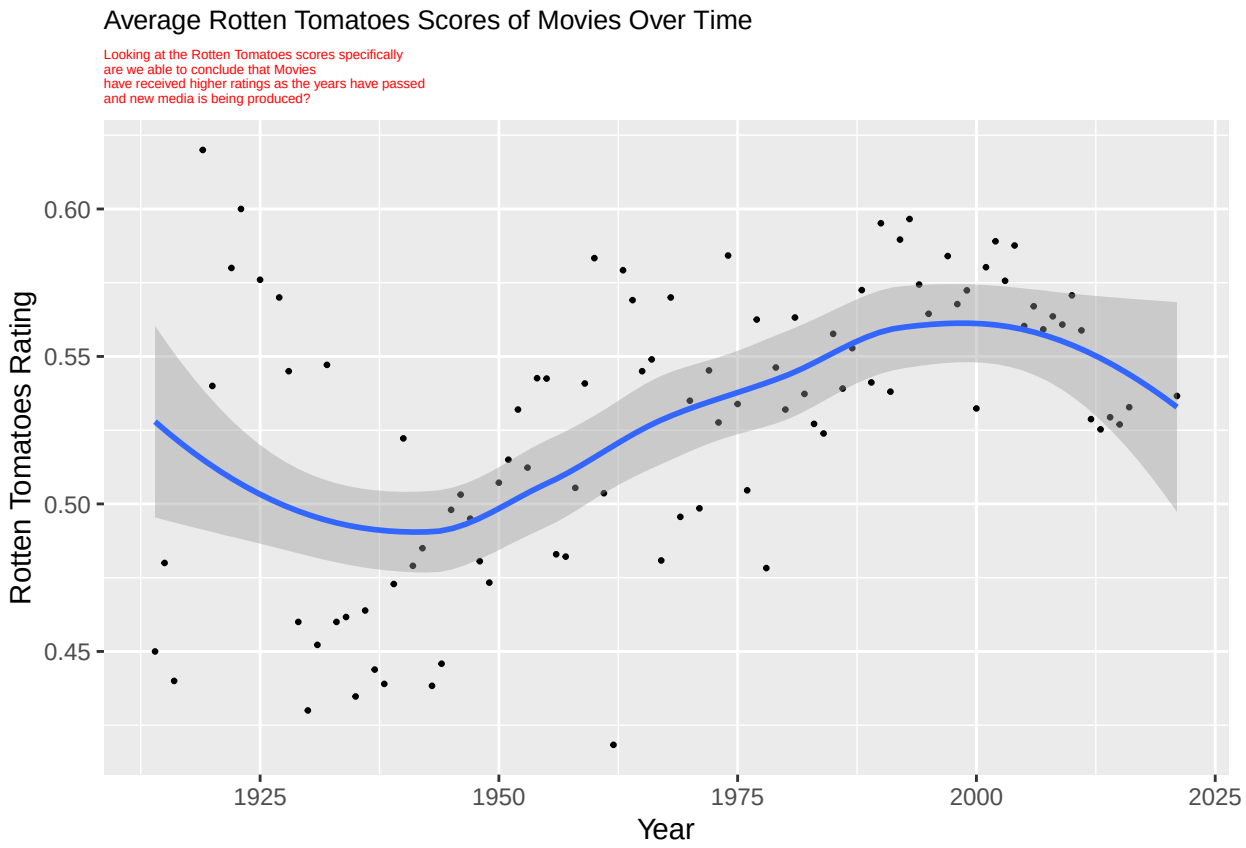

```
###Rotten Tomato Scores for Movies Only
ratingsmovie = Movie %>%
  group_by(Year, Rotten.Tomatoes) %>%
  arrange(Year) %>% na.omit()
p2 = ratingsmovie %>% ggplot(aes(x = Year, y = Rotten.Tomatoes)) +
  geom_point(size=0.5) +
  geom_smooth() + labs(title = "Rotten Tomatoes Scores of Movies Over Time", subtitle = "Looking at the
are we able to conclude that Movies
have received higher ratings as the years have passed
and new media is being produced?") + ylab("Rotten Tomatoes Rating")+
  theme(plot.title = element_text(size = 10),
        plot.subtitle = element_text(size = 5, color = "red")
        )
p2
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
###Average Rotten Tomato Scores for Movies
ratingsmovie2 = Movie %>%
  group_by(Year) %>% summarise(Mean_Score = mean(Rotten.Tomatoes)) %>% arrange(Year) %>% na.omit()
p3 = ratingsmovie2 %>% ggplot(aes(x = Year, y = Mean_Score)) +
  geom_point(size=0.5) +
  geom_smooth() + labs(title = "Average Rotten Tomatoes Scores of Movies Over Time", subtitle = "Lookin
are we able to conclude that Movies
have received higher ratings as the years have passed
and new media is being produced?") + ylab("Rotten Tomatoes Rating")+
  theme(plot.title = element_text(size = 10),
        plot.subtitle = element_text(size = 5, color = "red")
        )
p3
```

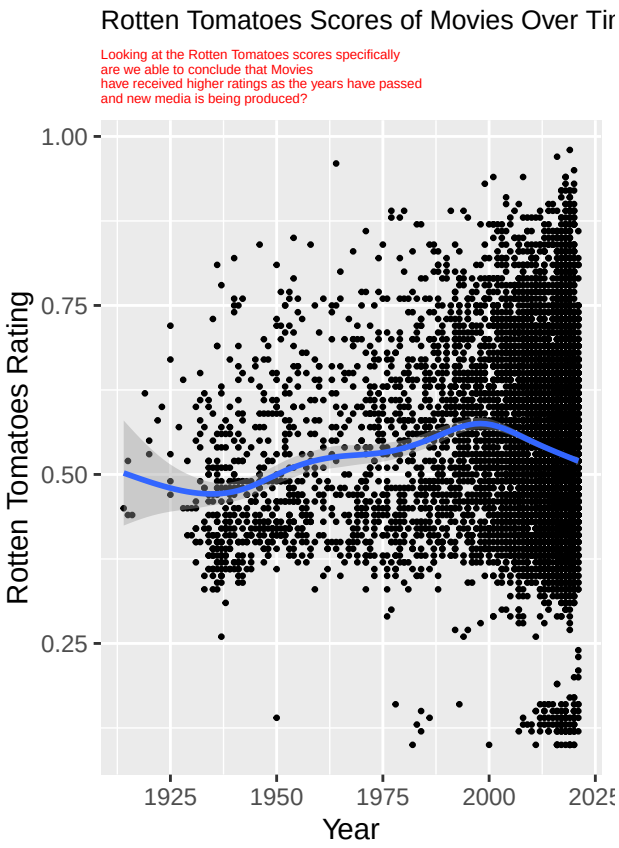
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

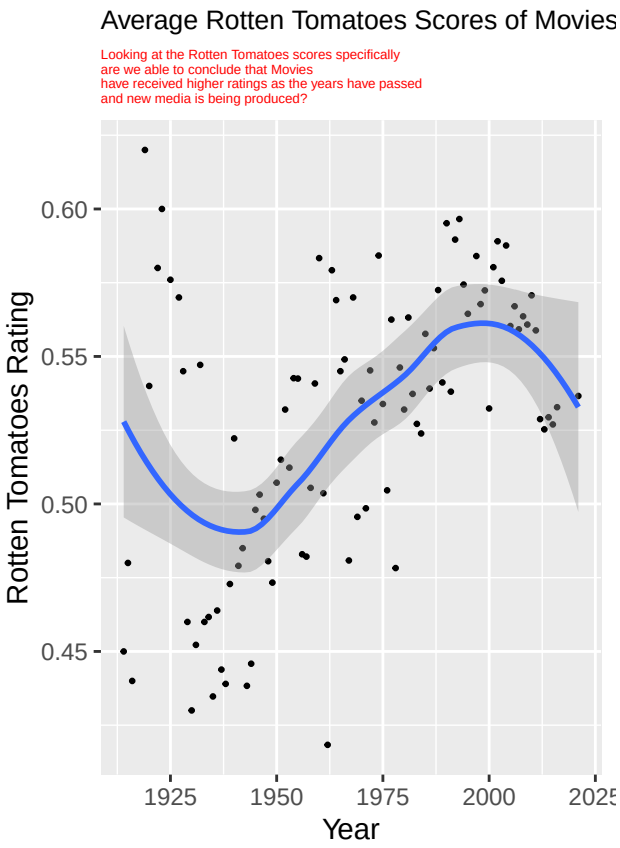



```
library(ggpubr)
ggarrange(p2, p3)
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

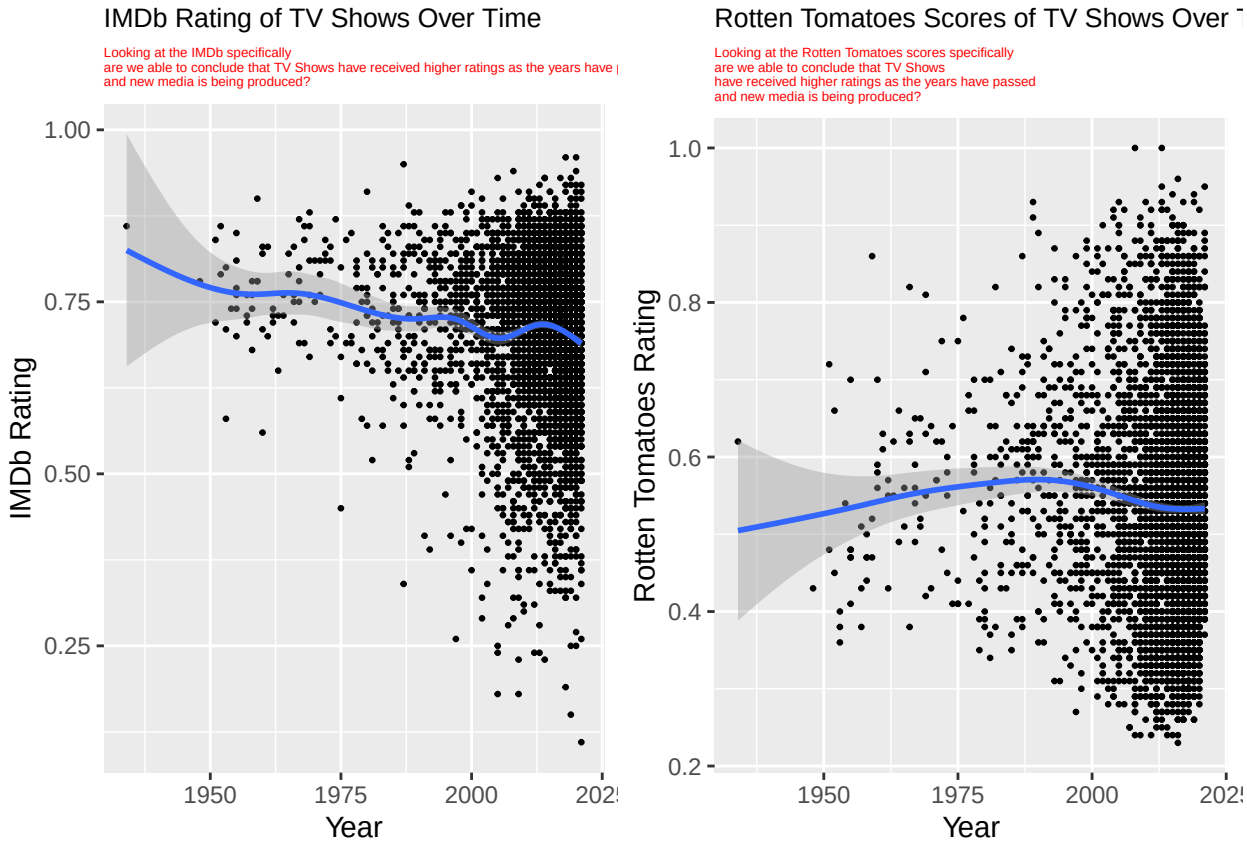





```
###Ratings for TV Shows Only
ratingstv = TV_shows %>%
  group_by(Year, IMDb, Rotten.Tomatoes) %>%
  arrange(Year) %>%
  na.omit()
p4 = ratingstv %>% ggplot(aes(x = Year, y = IMDb)) +
  geom_point(size=0.5) +
  geom_smooth() + labs(title = "IMDb Rating of TV Shows Over Time", subtitle = "Looking at the IMDb spe
are we able to conclude that TV Shows have received higher ratings as the years have passed
and new media is being produced?") + ylab("IMDb Rating")+
  theme(plot.title = element_text(size = 10),
        plot.subtitle = element_text(size = 5, color = "red")
        )
p5 = ratingstv %>% ggplot(aes(x = Year, y = Rotten.Tomatoes)) +
  geom_point(size=0.5) +
  geom_smooth() + labs(title = "Rotten Tomatoes Scores of TV Shows Over Time", subtitle = "Looking at t
are we able to conclude that TV Shows
have received higher ratings as the years have passed
and new media is being produced?") + ylab("Rotten Tomatoes Rating")+
  theme(plot.title = element_text(size = 10),
        plot.subtitle = element_text(size = 5, color = "red")
        )
library(ggpubr)
ggarrange(p4, p5)
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

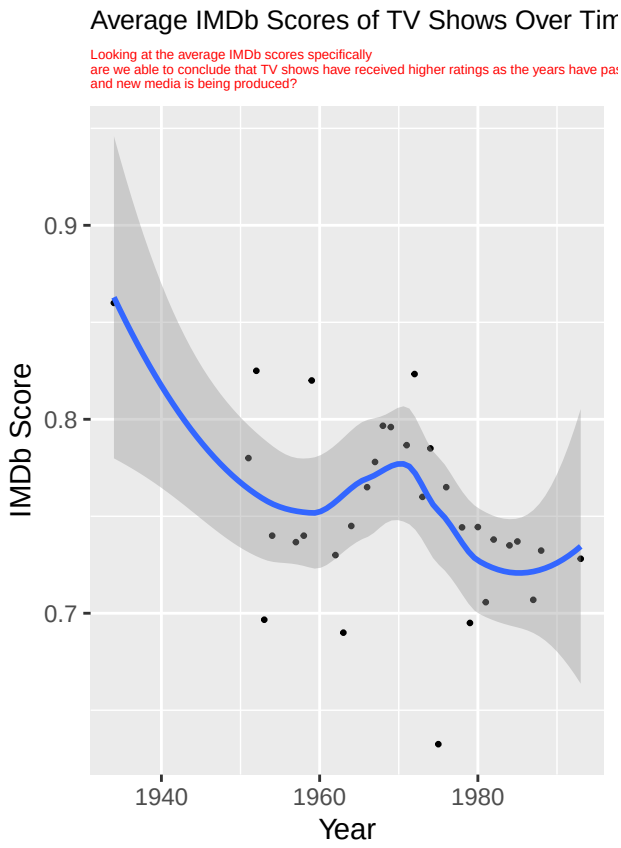
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

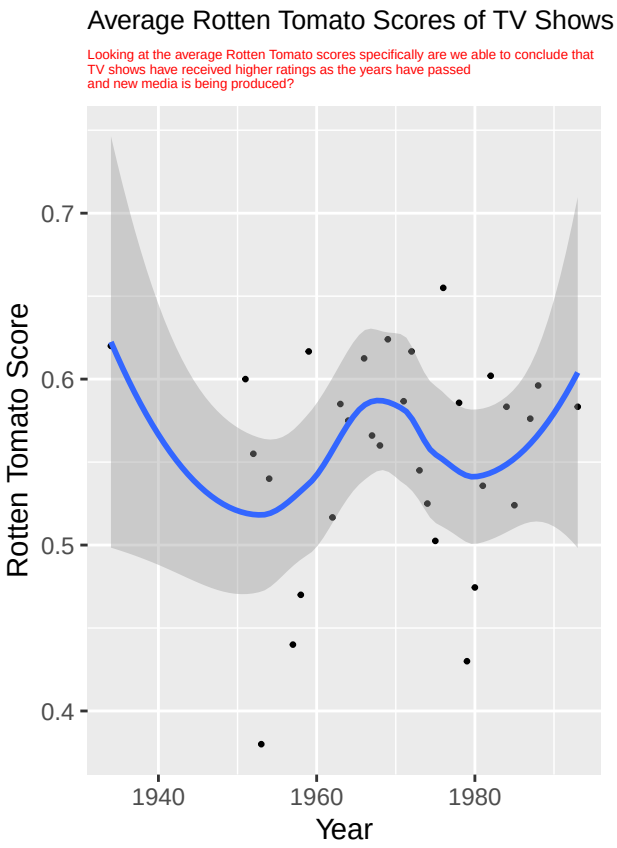



```
###Average Ratings for TV Shows Only
ratingstv2 = TV_shows %>%
  group_by(Year) %>% summarise(Mean_IMDb = mean(IMDb), Mean_Rotten.Tomato = mean(Rotten.Tomatoes)) %>%
  arrange(Year) %>%
  na.omit()
p6 = ratingstv2 %>% ggplot(aes(x = Year, y = Mean_IMDb)) +
  geom_point(size=0.5) +
  geom_smooth() + labs(title = "Average IMDb Scores of TV Shows Over Time", subtitle = "Looking at the a
are we able to conclude that TV shows have received higher ratings as the years have passed
and new media is being produced?") + ylab("IMDb Score") +
  theme(plot.title = element_text(size = 10),
        plot.subtitle = element_text(size = 5, color = "red")
        )
p7 = ratingstv2 %>% ggplot(aes(x = Year, y = Mean_Rotten.Tomato)) +
  geom_point(size=0.5) +
  geom_smooth() + labs(title = "Average Rotten Tomato Scores of TV Shows Over Time", subtitle = "Lookin
TV shows have received higher ratings as the years have passed
and new media is being produced?") + ylab("Rotten Tomato Score") +
  theme(plot.title = element_text(size = 10),
        plot.subtitle = element_text(size = 5, color = "red")
        )
ggarrange(p6, p7)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



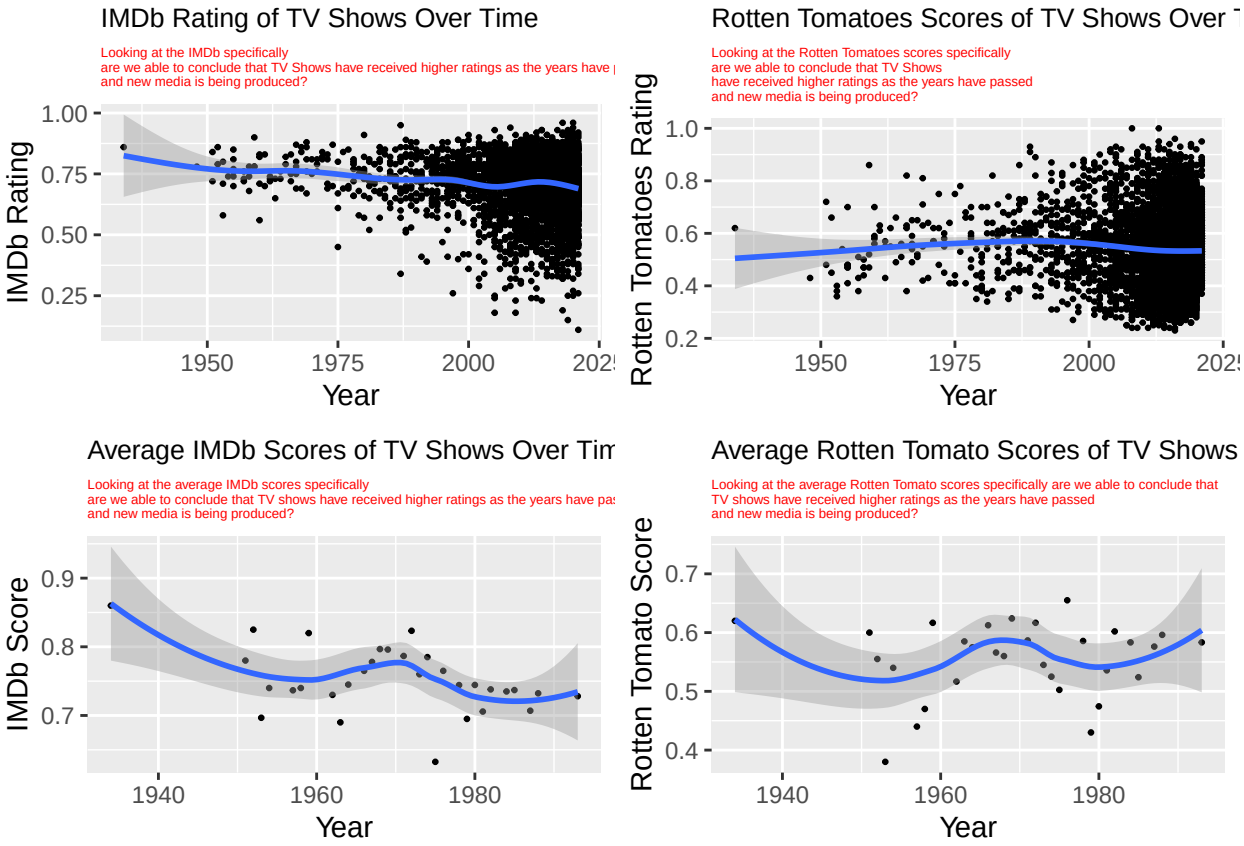



```
ggarrange(p4, p5, p6, p7)
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



- The dataset provides only the Rotten Tomato scores for movies, but both the IMDb scores and Rotten Tomato scores for TV Shows.
- When looking at the graphic comparing the years the movies and TV shows came out vs the Rotten Tomatoes rating we can see between a bit before 1950 through right before 2000 we see a slight increase in Rotten Tomato scores, but a decrease through the present. However, this is just a general trend as there are way too many movies and shows to be able to draw an absolute conclusion about ratings in this way.
- The scores, especially through the 2000s, seem to be very spread out with some movies and shows receiving extremely high scores and some receiving extremely low scores. This definitely affects the way that the smoother gets placed in the graphic and makes it difficult to actually say that there was a definitive decrease in Rotten Tomato scoring over time.
- The graphic showing just the Rotten Tomato scores of movies still seems to be way too large of a sample to draw any definitive conclusions, but the graphic showing the average Rotten Tomato scores of movies in each year shows a bit more. We can see that over time the average scores of movies that came out fluctuated a bit in their scores. It does not seem that the average scores of movies has consistently increased or decreased over time. It looks as though from 1914 until right before 1950 the average Rotten Tomato score decreased, then increased until around 2000, and has since been in a decline.
- The dataset provided both the IMDb and Rotten Tomato scores for TV shows. Looking at the graphics showing the IMDb scores and Rotten Tomato scores of all TV shows we still cannot draw a direct conclusion about the trends of scores over time as the sample size is just too large. Looking at the graphics showing the average IMDb scores and average Rotten Tomato scores of TV shows there is no consistent trend of scores increasing or decreasing consistently. Both graphics look to increase and decrease within the same years which is interesting since the scoring methods used for IMDb and Rotten Tomato scores are different.

- IMDb scores are based on users of IMDb submitting a score and a review of the movie, and Rotten Tomato scores are based on critics that are approved by the Rotten Tomato creators. Overall, it would seem IMDb scores are a bit more accurate as anyone can submit a review not jut a critic that has been approved.
- Overall, it seems as though the average rating of scores for both movies and TV shows do not follow a specific pattern of increasing or decreasing, but rather there are some years where scores increase and some where they decrease.

4. Do certain platforms contain more movies in specific genres than other genres?

    - Our fourth question again provides a great way to showcase our skills in ggplot. We knew we wanted to easily visualize the amount of movies in each specific genre on the different platforms. The genre and platform variables are both categorical, so we went with a bargraph. To be able to best visualize each platform individually we decided to facet_wrap using the platform variable.
    - Taxonomy for Data Graphics:

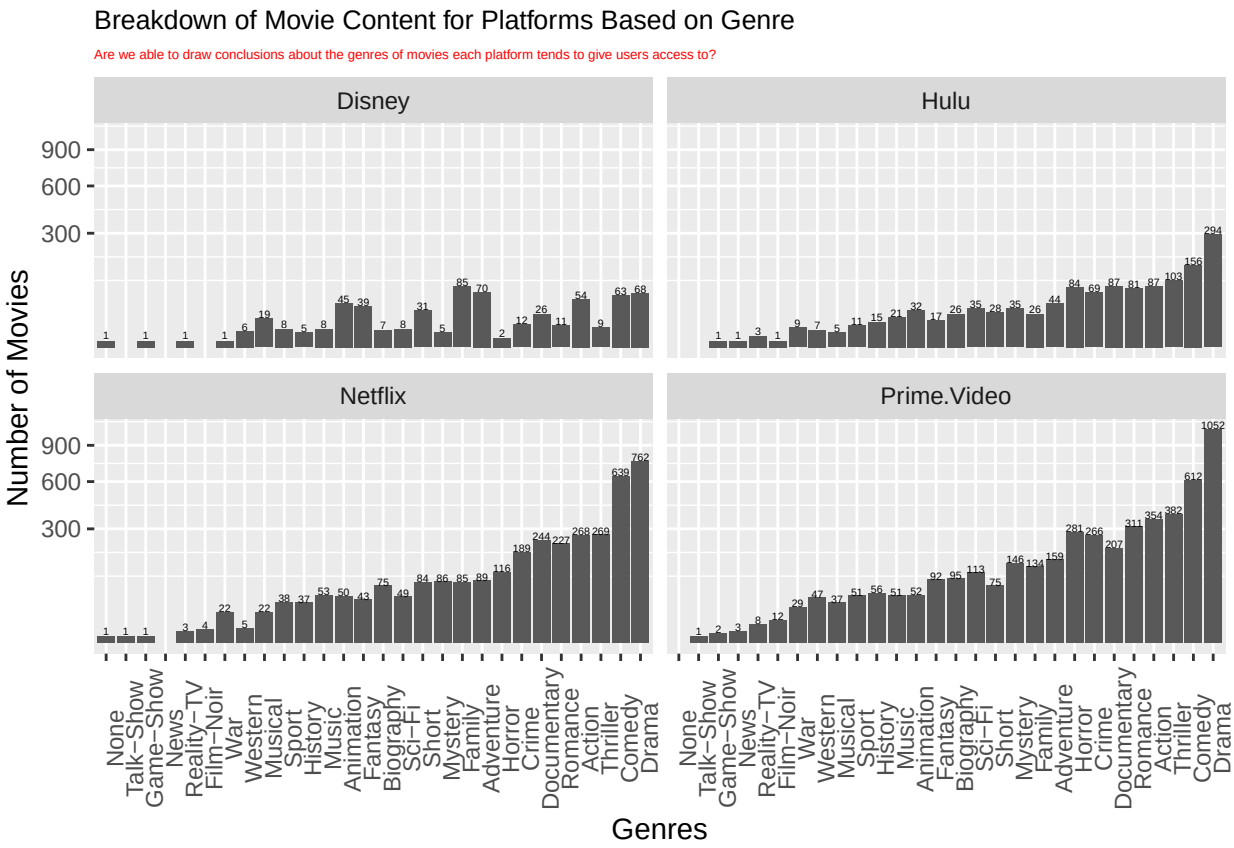
Visual Cues: We used length to show how many movies in each genre was available on each platform and position to show which genre a specific amount of movies were in.
Coordinate system: Cartesian
Scale: x-axis: categorical(the genres), y-axis: linear (number of movies)
Context: We have a title, subtitle, labeled axes, and we have the exact number labeled at the top of each bar.

```
#Question 4
genres <-
  data %>%
  filter(Type==0) %>%
  left_join(imdb, by = "Title") %>%
  select(Title, Movie.Type, Platform) %>%
  arrange(Movie.Type) %>%
  unique() %>%
  separate(Movie.Type, c("genres1", "genres2", "genres3"), sep = ",") %>%
  mutate(genres2 = str_sub(genres2, 2, -1),
         genres3 = str_sub(genres3, 2, -1)) %>%
  pivot_longer(2:4, names_to = "Genres", values_to = "genres") %>%
  filter(!is.na(genres)) %>%
  select(-Genres) %>%
  group_by(Platform, genres) %>%
  summarise(Number = n())
```

```
## Warning: Expected 3 pieces. Missing pieces filled with `NA` in 2884 rows [1, 2,
## 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].
```

```
## `summarise()` has grouped output by 'Platform'. You can override using the
## `.groups` argument.
```

```
genres %>%
  ggplot(aes(x = reorder(genres, Number), y=Number)) +
  geom_bar(position = "dodge", stat = 'identity') +
  scale_y_sqrt() +
  geom_text(
    aes(label = Number, y = Number + 0.5),
    position = position_dodge(0.9),
    vjust = 0,
    size = 4 / .pt) +
  facet_wrap(~Platform) +
  labs(x = "Genres", y = "Number of Movies", title = "Breakdown of Movie Content for Platforms Based on
  theme(axis.text.x = ggplot2::element_text(angle = 90)) +
  theme(plot.title = element_text(size = 10),
        plot.subtitle = element_text(size = 5, color = "red")
        )
```



- Yes.
- Disney has more family, adventure, animation movies. It is obvious that the kinds of movie on Disney's platform has a more evenly distributed than other platforms.
- Netflix, Prime Video and Hulu all have more drama, comedy movies than any other kinds of movies.
- As we can see different platforms have different genres represented more or less heavily through the movies that are available. However, we must take this with a grain of salt, since certain genres of movies may be more popular than others and some movies can fall into more than one genre. For example, we see comedy and drama are among the most highly shown genres on all platforms, but these two genres have more movies in them than genres like war or western overall, so it would make sense to see these genres more represented among the platforms since more movies are available in these two genres than all others.

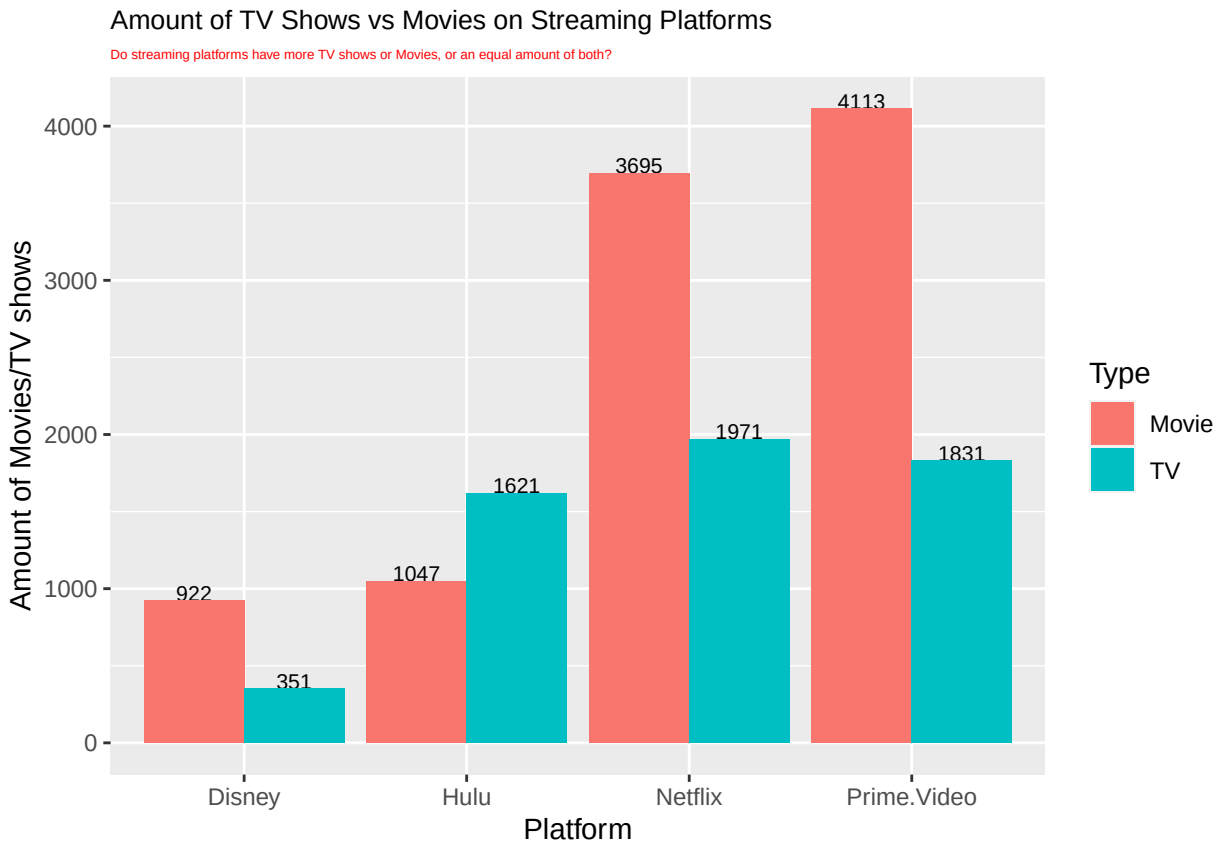
5. Do certain streaming platforms have more movies vs TV shows or vice versa?

    - To visualize the information needed to answer this question our immediate instinct was to create a bargraph since our platform variable and type variable are both categorical. We grouped_by MT (which was our column we created to distinguish which titles were Movies or TV Shows in order to more easily visualize the information when creating the graph so that we would have two distinct colors to represent the information rather than a gradient since the "Type" variable only has two values of 0 and 1) and platform. We were then able to summarise and calculate the amount of movies and shows on each platform.
    - Taxonomy for Data Graphics:

Visual Cues: We used color to represent whether the title belonged to the Movie or TV show categories, length to show how many movies and TV shows are available and position to show how many movies and TV shows are available on each platform.
Coordinate system: Cartesian
Scale: x-axis: Categorical (platforms), y-axis: linear (amount of movies/shows), color: categorical (media type)
Context: We have a title, subtitle, labeled axes, and we have the exact number labeled at the top of each bar.

```
# Question 5
data %>% group_by(MT, Platform) %>%
  summarise(n = n()) %>%
  ggplot(aes(x = Platform, y = n, fill = MT)) +
  geom_text(
    aes(label = n, y = n + 0.05),
    position = position_dodge(0.9),
    vjust = 0,
    size = 7.5 / .pt
  ) +
   geom_bar(position = "dodge", stat = 'identity') + labs(title = "Amount of TV Shows vs Movies on Strea
  theme(plot.title = element_text(size = 10),
        plot.subtitle = element_text(size = 5, color = "red")
        )
```

```
## `summarise()` has grouped output by 'MT'. You can override using the `.groups`
## argument.
```



- Yes. Disney, Netflix, and Prime Video have more movies than TV shows, Hulu has more TV shows than movies.

Final Analysis

- We learned many different things about our data and how different streaming platforms actually are. When we first began our project we had no idea the extent to which platforms offered their content and how this could possibly change the streaming platforms we might choose to subscribe to. We were able to draw conclusions about which platforms might have more content for a younger audience vs an older audience, or which platforms contain more movies in specific genres that a user might want access to. We were able to answer questions 1, 3, 4 and 5 after creating different data graphics, which allowed us to draw very direct conclusions that are visible to the eye and can be easily interpreted. To answer question 2 we completed a hypothesis test to be able to draw our conclusion, which can be repeated by other statisticians and the p-values can allow us to make concrete conclusions in regards to whether an "A" list actor influences the IMDb rating of a show.
- There is definitely much further analysis that can be completed in regards to this topic such as looking into the cost of each platform and how much content you are receiving or which platforms contain more highly rated movies and shows, there are also so many other streaming platforms like HBOmax and Paramount+ that could be included in the discussion of this topic if a similar dataset containing information about these platforms could be compiled. When it comes to selecting which streaming platform to purchase a subscription to we feel as though this assignment and all the analysis we were able to complete gives a great overview of each of the four streaming services we looked at and could help someone come to a decision if choosing a certain platform is the goal.