Online Streaming Services: Netflix & Amazon Prime

STAT 3355.001

Group 24 - Vishnu Nambiar, Victoria Puente, Nrityya Sivakumar Annu

April 24th 2023

Table of Contents

1 Introduction

This project's purpose is to understand how to use R in order to analyze and make use of information within a massive dataset along with a secondary dataset to supplement data. The idea is to clean the data in order to be able to combine different datasets, as well as create plots that help readers understand and visualize any patterns that may occur within the data. The goal is to have these plots result in some conclusion that could be reasoned and expanded on.

There were two datasets we had decided to use. The first being a dataset that covered Netflix movies and tv shows, and the second covering Amazon Prime's selection of movies and tv shows.

- The main dataset is titled "Netflix Movies and TV Shows" (*netflix_titles.csv*), posted by Shivam Bansal on kaggle.com.
- The secondary dataset is titled "Amazon Prime Movies and TV Shows" (*amazon_prime_titles.csv*), posted by Shivam Bansal on kaggle.com.

Both datasets displayed whether the media was a tv show or a movie (*type*), their name (*title*), the director (*director*), the cast (*cast*), where the media originated (*country*), the date of when the media was added to the online platform (*data_added*), the date of when the media was released to the public (*release_year*), the age rating (*rating*), the length of the show in seasons or the length of the movie in minutes (*duration*), in what genre/category it was listed in within the platform (*listed_in*), and the description under each piece of media on the platform (*description*). A primary advantage of picking Amazon Prime was the identical categories of the dataset. The reason was due to the creator of the dataset being the same individual, and they had utilized similar code in order to receive/organize the data. This also means that the countries would be named the same and there wouldn't be an issue with capitalization or choice of word when attempting to combine datasets.

Nevertheless, we cover the relationships with the various variables within the dataset, mainly concerning the relationship between age ratings and countries of origin, release year and age rating, the date added across months, and the number of movies and tv shows across release years. Each relationship will be analyzed to see if there is any possible conclusion or correlation that can be seen within the patterns of the data.

2 Data Cleaning

Data Cleaning for Figure 3.1

The dataset was subsetted based on month_added, title and release year. A new column was created called month_added to easily access the months from the date_added attribute.

Data Cleaning for Figure 3.2

This dataset had been edited slightly in order to allow for an easier understanding if a movie or tv show was before or after the year 2000 by adding a column that describes it as such (*beforeAfter2000*). Although this wasn't necessarily utilized in its entirety, it allowed for the exploration of the topic.

Data Cleaning for Figure 3.3

The country names of the dataset had to be changed for the sake of a similar naming structure across countries between *map_data* from *tidyr*, and *netflix_titles.csv* and *amazon_prime_titles.csv*. The spelling differences had to be manually corrected. To add onto this, the *country* section of the data had to be changed from comma-delimited to entirely separated, so that certain media repeat more than once for each country that it had listed. This then allows us to match the country names between datasets.

Data Cleaning for Figure 3.4

For this graph we created two subsets, *netflix_year* and *amazon_year*. These datasets were obtained by subsetting the netflix and amazon datasets by a new vector, *year_vars*. *year_vars* is a concatenated vector which contains the character strings of the names of our *release_year* and *type* variables. This allowed us to have a data frame with solely the variables we needed to work on this graph. Additionally, the *type* variable was factored. It became factored with two levels: "Movie" and "TV Show".

3 Questions and Discoveries

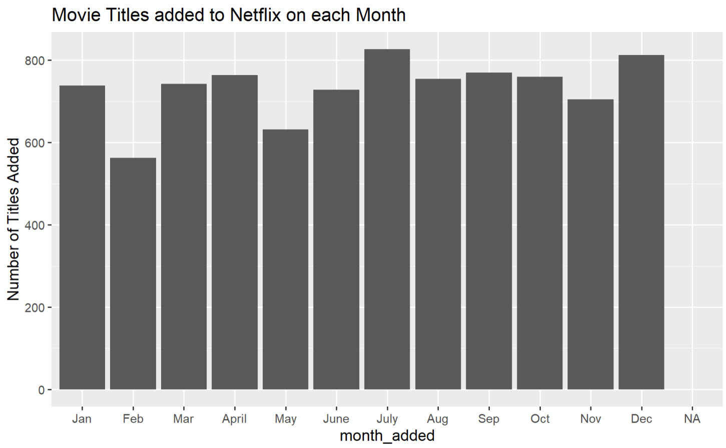
3.1 Number of titles added each month

The purpose of this analysis was to determine the number of titles added each month to Netflix and Amazon Prime, and to identify any trends in their release patterns.

According to the bar charts, Netflix added more titles throughout the month of July. This finding correlates to the summer season, when many people have more free time to watch movies and television series. Furthermore, July is often a month when many networks and studios release new content, so Netflix may add fresh programs to capitalize on this trend. In contrast, Netflix added a relatively modest number of titles in February compared to the rest of the year.

Our analysis revealed that the Amazon Prime dataset did not have sufficient "date_added" data. The dataset only contained the "date_added" information for a small proportion of the entries. Specifically, only 20% of the entries in the dataset contained the "date_added" information, while the remaining 80% of the entries did not have any "date_added" data. In conclusion, this analysis revealed interesting insights into the release patterns of Netflix.

Netflix:



Amazon:

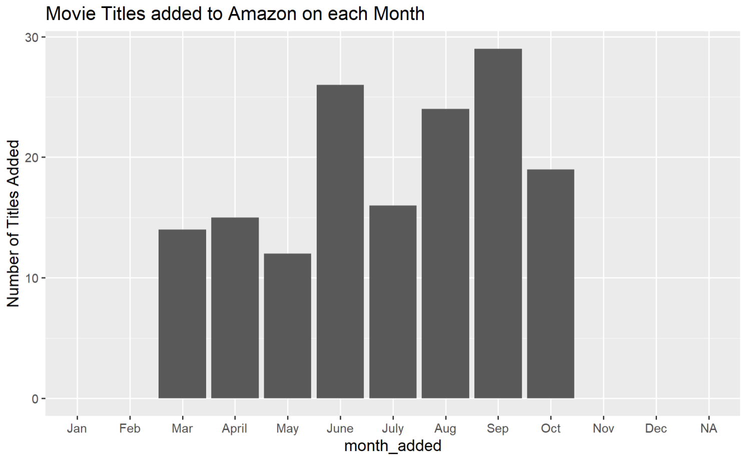
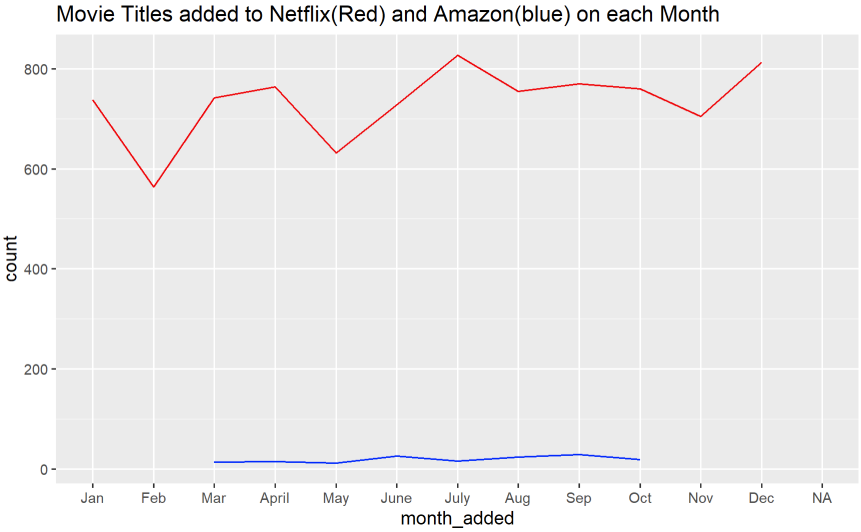


Figure 3.1 includes the two graphs above and the one below.

In addition to the bar plots, I also used a one-line plot to visualize the monthly counts of titles added to Netflix and Amazon. The line plot provides a clear trend of the number of titles added each month, making it easier to identify patterns and trends over time. Overall, the line plot seems to be a more effective visualization tool than the bar plot for this particular dataset.



3.2 Proportion of Ratings of Movies Before v.s. After 2000

We also had decided to analyze the relationship of movie ratings across different time periods. Before the 2000s, movies were entirely different in themes and storytelling. Moreover, Netflix had only started their streaming services in 2007, so there is likely to be a difference in choice when older movies were selected versus newer movies. However, a clear difference between movies from then and now is the age rating. Older movies tended to be rated for younger audiences as certain aspects of movies, such as violence, would be rated less harshly and critically. Also, rarely do movies get re-rated. Movies that were rated PG-13 at one point may be rated R today, and as a result, we wanted to see if any similar pattern existed between movies in the previous millenia versus now within the selection of Netflix.

We had split up the data into two sections for both Amazon and Netflix. First we would have to separate the movies from the tv shows. Once that was done, we split them even further by creating a dataset with movies that had release years from 2000 and greater, and another that had release years less than 2000. Once that was done, we used the library “patchwork” to place all four graphs together.

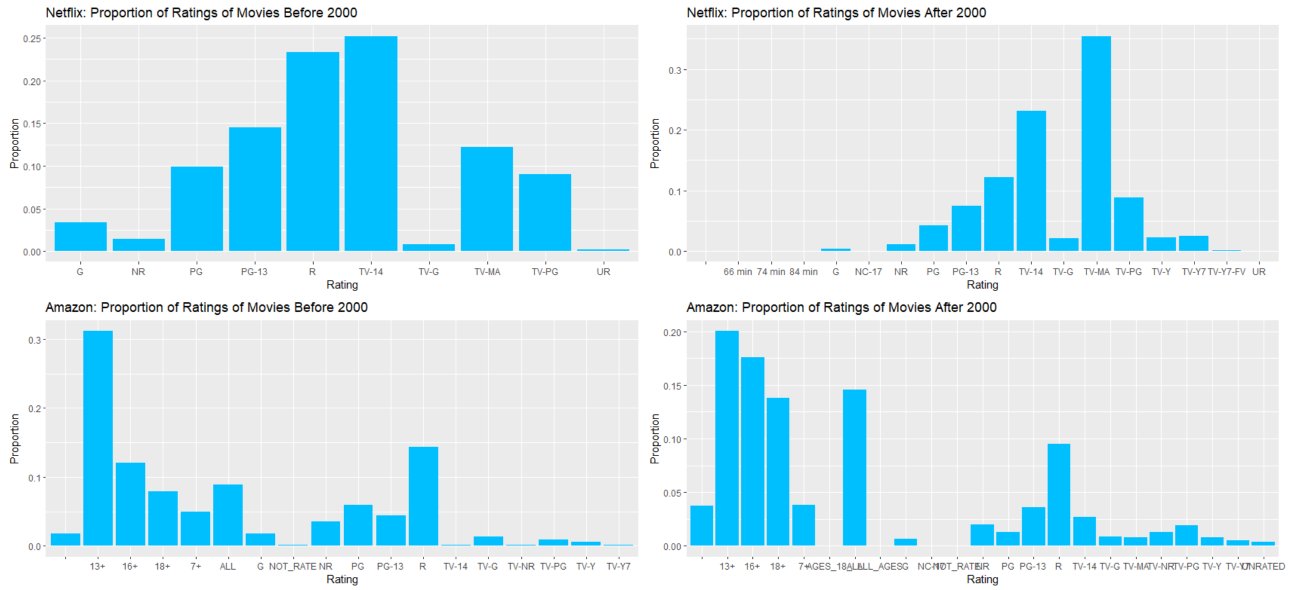


Figure 3.2

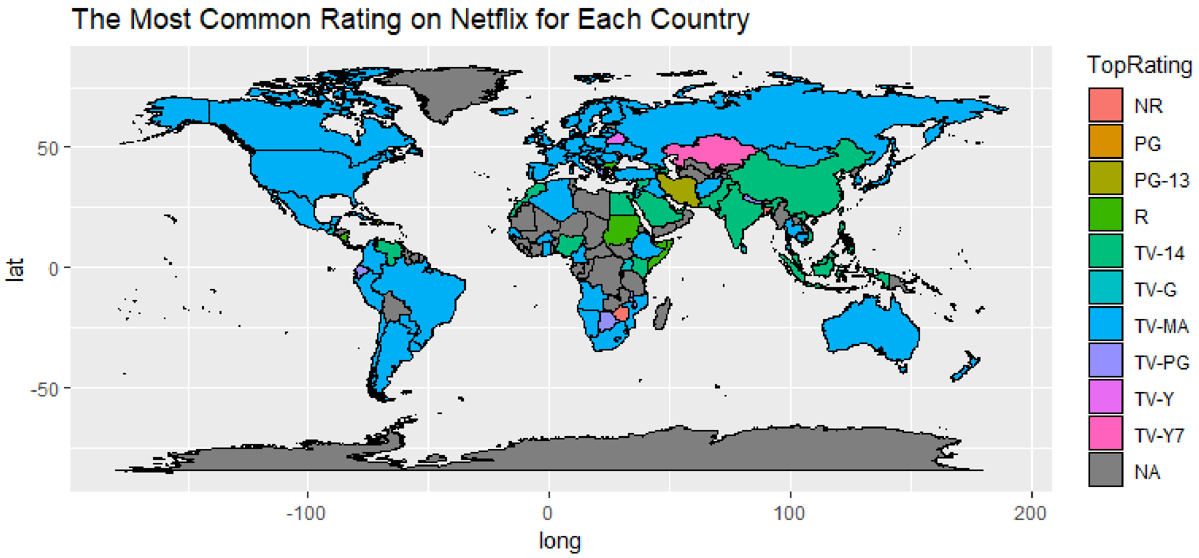
A possible conclusion is that the proportions tend to be drastically different between Netflix and Amazon Prime. A reason may be that their choice of rating tends to differ, since Amazon chooses to display entirely different ratings that Neflix never uses. However, for Netflix solely, there tends to be a clearer pattern between the years before and after 2000. For example, TV-MA is a little over 25% of all movies before the 2000s, whereas after 2000 TV-MA is over 35% of all movies. Moreover, R movies are above 22% of all movies before 2000, but are only a little more than 12% of all movies after 2000. The most significant difference, however, is that there is less variety in the Netflix movies before 2000, as the spread of the age ratings tend to be a lot more even than in the movies after 2000. This is likely because of Netflix having less of a priority of satisfying their larger audience when it comes to older movies, as they tend to target a much more niche group of individuals looking for classic movies. On the contrary, Netflix’s audience is largely adults, and so the latest R or TV-MA movies seem to be popular choices for Netflix’s selection.

3.3 Most Common Ratings of Countries

Another question we wanted to ask is if there are certain countries where Netflix prefers to involve different age ratings of content? This was especially interesting since Netflix offers a different selection based on the country in which Netflix is being viewed, and Neftlix is originally an American company, which expanded outwards in recent years. Netflix can only show movies and shows that they are licensed to in the United States within the United States, and this goes for other countries. As a result, we were curious to see if certain age ratings were more common in some countries rather than others.

We chose to display the data with a map using “tidyr”. Using “map_data(“world”)”, we are able to insert a map into a plot through R Studio. The best method in which we could implement this map is by creating a new dataset that merges the map data with the Netflix/Amazon data. There were some issues with the spelling of countries however. For example, within the Netflix data, the “United States” is the same as “USA” within the map data. Otherwise there was no other problem in compatibility.

Netflix:



Amazon:

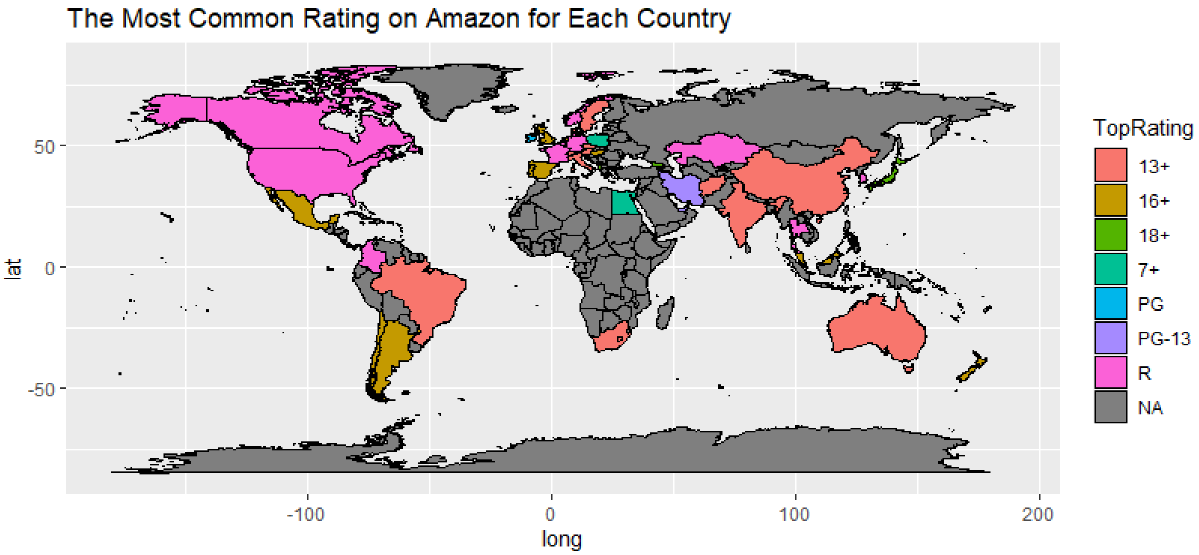


Figure 3.3 includes the two graphs above.

For both graphs, a lower maturity rating is more common across countries such as China and India, whereas a higher maturity rating is more common across countries such as the United States, the UK, and Canada. More specifically to Netflix, countries such as India and China have more TV-14 ratings in proportion to all Indian and Chinese Netflix shows in comparison to countries such as the United States, the United Kingdom, and Australia. Most countries around the world have more movies and shows that are rated to be TV-MA, which is Netflix's most popular age rating. In Amazon Prime, this is not as clear cut, where R and 13+ tend to be the most popular age ratings. It seems that countries with western cultures are more likely to have more TV-MA/R ratings than other cultures based on both datasets.

3.4 Available Listings by Release Year

One of the main questions we wanted to find answers for is: Are more TV shows or movies released in comparison per year? During our initial brainstorm, we wanted to approach this question by date_added. However, for the Amazon data, only 155 out of 9,668 observations contained values for date_added and the rest were missing values. We thus proceeded with this question from the perspective of the distribution of movies and TV shows on Netflix and Amazon based on the titles' release year, since missing values were not an issue with the release_year variable in either dataset.

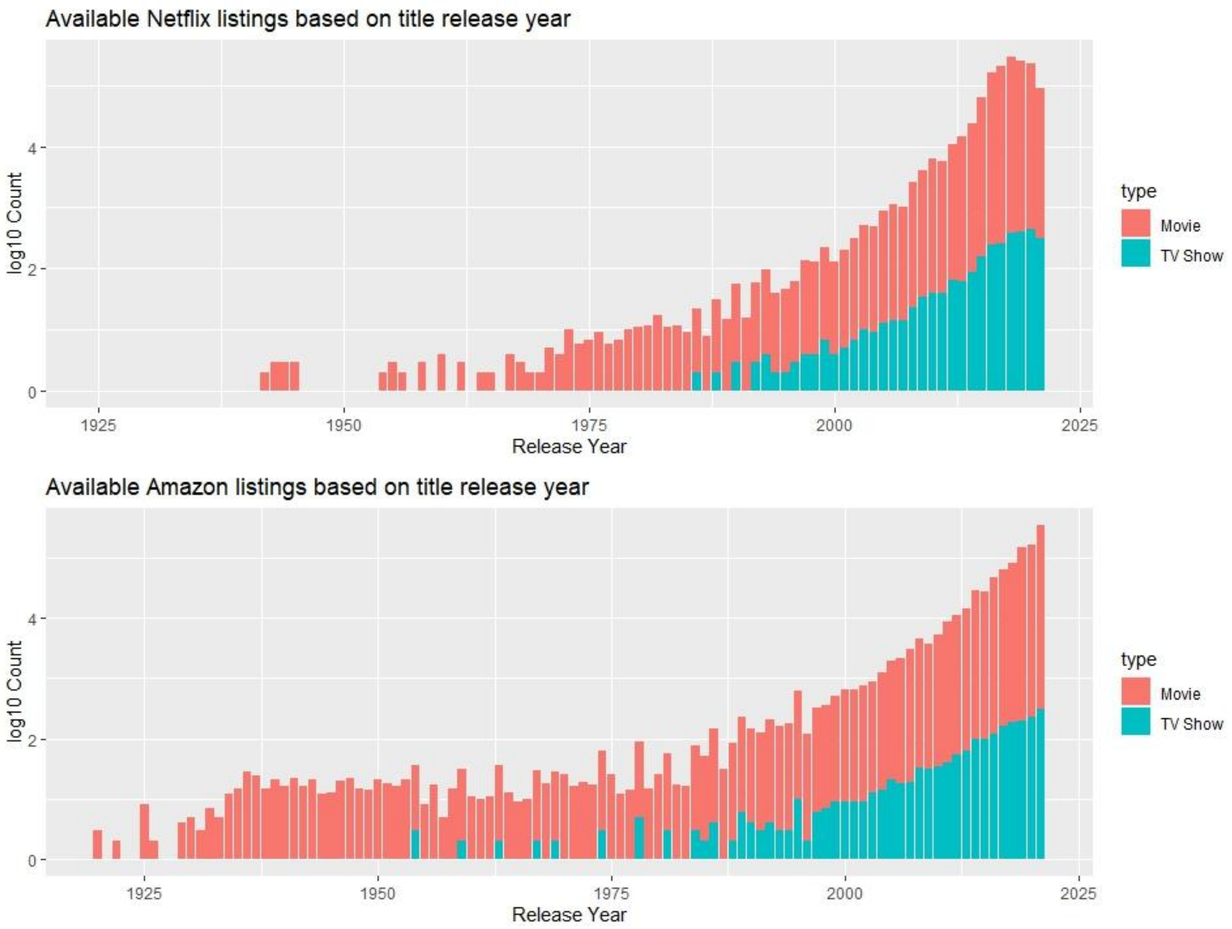


Figure 3.4

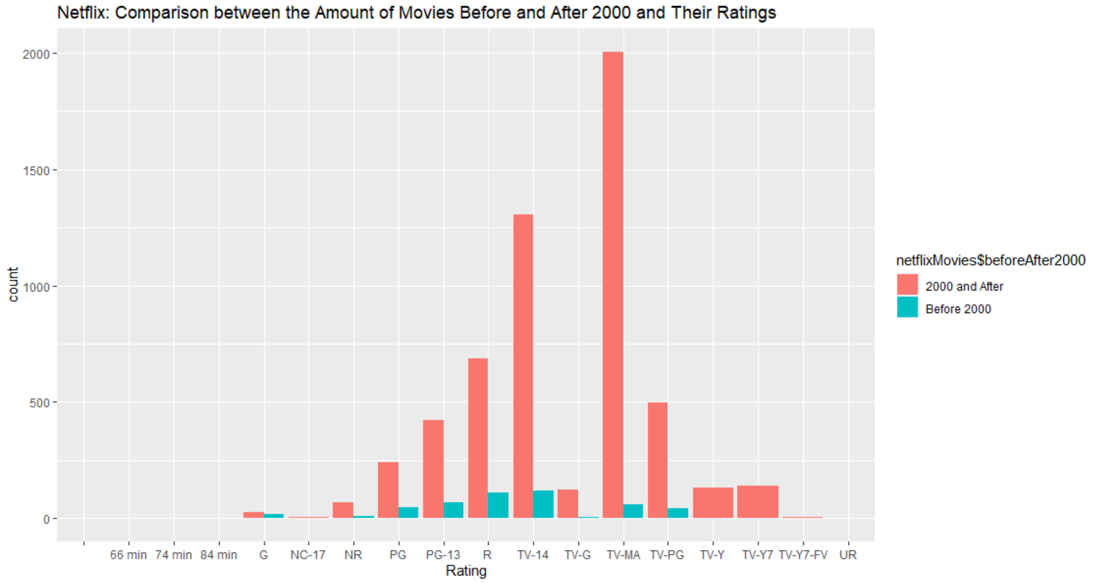
To arrange our graphs in a way that would allow for seamless and direct comparison, we used the “gridExtra” package. The first iteration graph used a count for each release year, and the resulting graphs were left-skewed. We could not draw many conclusions about the data due to the skew and large difference in values. This graph was revised to the figure 3.4 graph, which uses the log10 count per release year.

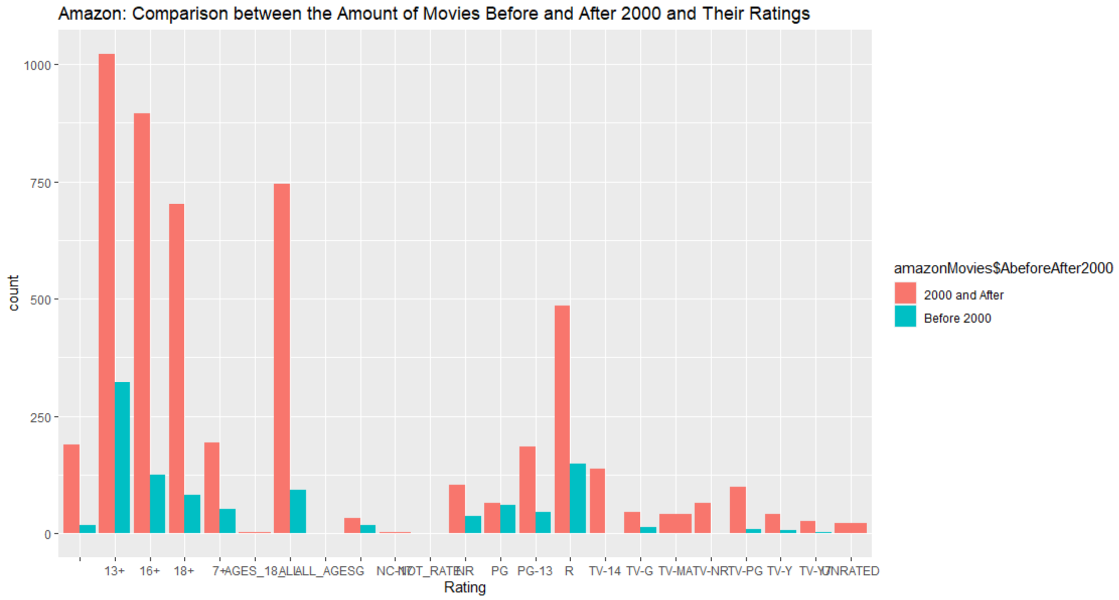
Switching from count to the log10 count made it easier to see patterns in the graph. Figure 3.4 highlights that of the years where there are both TV shows and movies on Netflix and Amazon, they each respectively make up close to half of the catalog for that release year. We concluded that for Netflix the year with the largest amount of titles is 2018 and for Amazon the release year with the largest amount of titles is 2021. Before release years in the 1960s, there were not many titles available on Netflix, and those that are available are of the type “Movie”. Amazon titles before the 1950s are of the type “Movie”. The most obvious pattern we saw from our graphs shows that there has been an increase in the amount of titles on these platforms as the years progress, with the difference between release years being smaller for Amazon than on Netflix.

4 Attempted Questions

Number of Ratings of Movies Before v.s. After 2000

The issue with trying to come to a conclusion between the number of ratings of movies before and after the year 2000 is that inherently there was going to be less movies before the year 2000 as Netflix only became a streaming service during 2007. This meant that Netflix was less likely to include a movie before this year than after this year. This was clear in the graphs below.

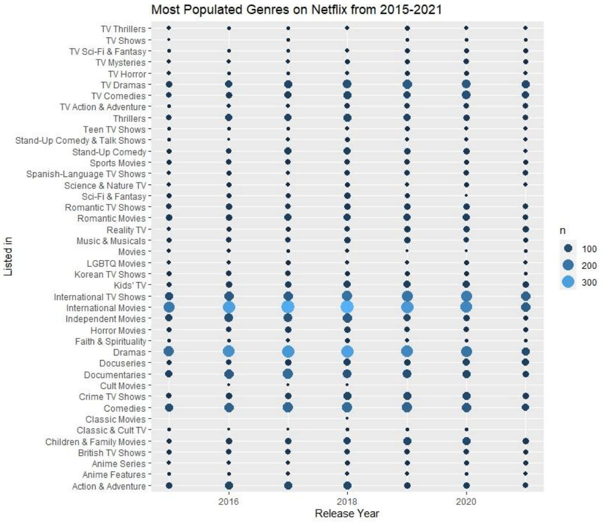


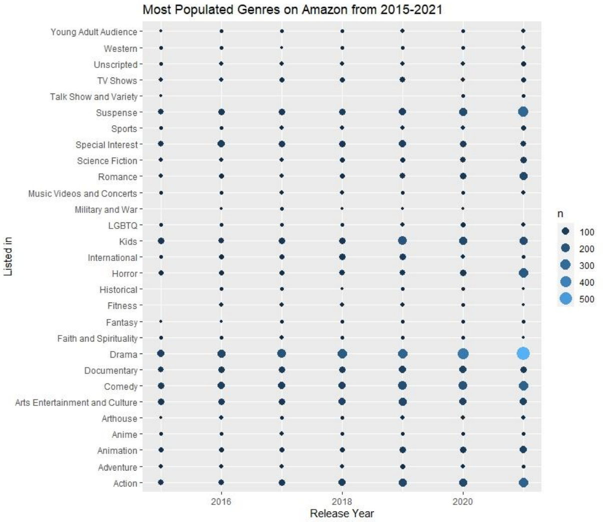


Throughout the years, has Netflix shifted from adding one particular genre to another?

I. Is there a genre that is more largely populated compared to other genres?

The issue with choosing to find patterns with the listed_in variable (even when it was factored) is that the Netflix dataset had 42 different genres, and Amazon Prime had 29 genres. Each title could also be listed in multiple genres so many of the observations were concatenated character vectors. We subsetted the data to only release years 2015-2021, but the resulting graph had weak correlations and interfered with our ability to come up with a strong conclusion. It made it both visually difficult to analyze and difficult to make a great conclusion, especially for the data between Amazon Prime and Netflix, as genres weren't always the same. However it did have some results, we could see that Dramas were both popular between Amazon and Netflix.

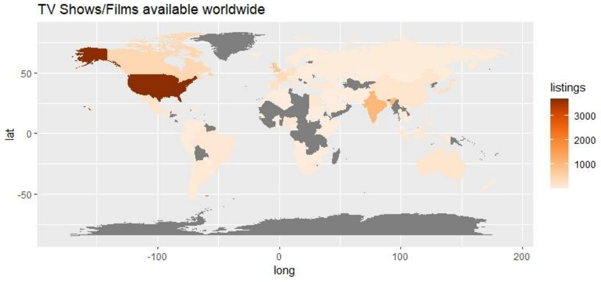




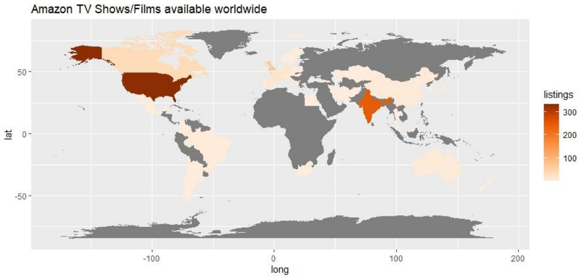
What is the distribution of listings available worldwide?

The question was simply not an interesting question, as the dataset would definitely be biased to wherever Netflix/Amazon originated. There wasn't enough interest in the question as one could guess that the United States would dominate in the number of movies within Netflix. This is likely due to the sheer size and popularity of the Hollywood industry and how Netflix was originally an only American company, before they started going international in 2011. Amazon is similar, as they are mainly an American corporation.

Netflix:



Amazon:



How do categories gain or lose popularity as years pass?

Unfortunately, the Netflix data we have only includes information about the kind of movies and TV shows that are added each month. It does not include any ratings or other popularity indices. We would need to use more data sources to assess how the popularity of different categories varies over time. One possible source of data could be IMDb, a popular movie and TV show database that provides ratings and reviews for millions of titles. By cross-referencing the categories of movies and shows added to Netflix with their IMDb ratings, we could gain insights into how the popularity of different categories has changed over time.

Are certain categories more or less likely to have a large runtime for movies and TV shows separately?

We decided to not pursue this question due to the fact that there were different sample sizes for different categories, and there may have been skewed and disproportionate data when comparing runtimes, and that would make it difficult to come to a conclusion. What would be needed is a larger sample size, or a categorization of genres in order to help come up with a more general conclusion.

What kind of rating are certain categories more likely to be?

A question that emerges when studying Netflix data is what sort of rating particular categories are more likely to be connected with. However, because the data comprises several categories for each title, identifying a clear pattern in the association between category and rating is challenging. Additional data with more precise information on each title may be necessary to study this further. Nonetheless, it's worth investigating the distribution of ratings across the Netflix collection and comparing it to the distribution of ratings across other media platforms. This might possibly shed light on Netflix consumers' tastes and give insights for content developers and distributors.

5 Conclusion

The goal was to discover patterns within the selection of movies and tv shows within Netflix and Amazon Prime. Multiple methods were used to help depict relationships, from simple bar graphs to entire maps.

When analyzing when movies and tv shows were added to Netflix and Amazon, we discovered that the most popular times of the year to do so were during summertime (July, August) and wintertime (January, December). When looking at the age ratings of movies before and after the year 2000, we found that there was a more even distribution of ratings for movies before 2000 than after. Moreover, for the years after 2000, more adult rated content existed in proportion. This was likely because the selection of older movies were probably rated more child friendly at the time, and the fact that older movies are not necessarily prioritized for an adult audience, which is true for current day movies as adults compromise most of the streaming service's audience. When we analyzed the relationship between country and the most common age rating within Netflix and Amazon Prime listings, we found out that countries with more western cultures were more likely to have more mature ratings as their most common age rating than any other country that has a lack of that culture. We also analyze the number of listings based on release year for both streaming services. A clear pattern that occurred was that the streaming services were adding more movies from the current year as time passed, likely due to the platforms' success.

Overall, there were significant discoveries, however, there is potential for improvement. We could've found and used data that would help us analyze the relationship of movies and their actors, and the impact a singular actor can have on a movie. If we found supplementing data referring to the popularity ratings of the movies, we could find which directors were more popular than others. We also could have tried to correlate the movie or tv show length with their success in popularity ratings, as it would give us insight if certain lengths indicated success, as some shows supposedly were aired for longer if they were popular, and some shows were cut short especially if it never gained traction. Instead of using information from another streaming service, supplementing data that gave more insight on each movie could've allowed us to ask entirely separate questions that could concern different aspects of movies and how those factors could relate to Netflix. Nevertheless, our plots gave us insight on multiple aspects of Netflix and Amazon Prime and provided us with many theories as a result.

6 Code

6.1 Code for Figure 3.1

6.1.1 Data Cleaning for Figure 3.1

```
# For netflix data
netflix <- read.csv("netflix.csv")

 library(ggplot2)
 library(tidyverse)

 # Convert the date_added column to a Date object
 netflix$date_added <- lubridate::mdy(netflix$date_added)

# Create a new column with the month and year of each movie's addition to
Netflix

 netflix <- netflix %>%
   mutate(month_added = format(date_added, "%m"))

 # Count the frequency of each movie category added to Netflix by month
   title_counts1 <- netflix %>%
   group_by(month_added, title, release_year) %>%
   summarize(count = n()) %>%
   ungroup()

# For Amazon data,
amazon <- read.csv("amazon.csv")

library(ggplot2)
library(tidyverse)

#Convert the date_added column to a Date object
 amazon$date_added <- lubridate::mdy(amazon$date_added)
```

```
#Create a new column with the month and year of each movie's addition to
Netflix

amazon <- amazon %>%
   mutate(month_added = format(date_added, "%m"))

#Count the frequency of each movie category added to Netflix by month
   title_counts2 <- amazon %>%
   group_by(month_added, title, release_year) %>%
   summarize(count = n()) %>%
   ungroup()
```

6.1.2 Construction of Graph for Figure 3.1

```
# For netflix data,
ggplot(data = title_counts1, mapping = aes(x = month_added)) +
     geom_bar() +
     scale_y_continuous(name = "Number of Titles Added") +
     scale_x_discrete(limit = c("01", "02", "03", "04", "05", "06", "07", "08",
"09", "10","11", "12", "NA"),
                     labels = c("Jan","Feb","Mar", "April", "May", "June",
"July", "Aug", "Sep", "Oct", "Nov", "Dec", "NA")) + labs(title = "Movie Titles
added to Netflix on each Month")

# For Amazon data,
ggplot(title_counts2, mapping = aes(x = month_added)) +
   geom_bar() +
     scale_y_continuous(name = "Number of Titles Added") +
     scale_x_discrete(limit = c("01", "02", "03", "04", "05", "06", "07", "08",
"09", "10","11", "12", "NA"),
                     labels = c("Jan","Feb","Mar", "April", "May", "June",
"July", "Aug", "Sep", "Oct", "Nov", "Dec", "NA"))+ labs(title = "Movie Titles
added to Amazon on each Month")

# Line plot for both the streaming platforms,
ggplot(mapping = aes(x= month_added, group = 1)) +
         geom_line(data  = title_counts1, stat = "count", col = "red") +
         geom_line(data = title_counts2, stat = "count", col = "blue")+
     scale_x_discrete(limit = c("01", "02", "03", "04", "05", "06", "07", "08", 
"09", "10","11", "12", "NA"),
                     labels = c("Jan","Feb","Mar", "April", "May", "June", 
"July", "Aug", "Sep", "Oct", "Nov", "Dec", "NA"))+ labs(title = "Movie Titles 
added to Netflix(Red) and Amazon(blue) on each Month")
```

6.2 Code for Figure 3.2

6.2.1 Data Cleaning for Figure 3.2

```
netflix <- read.csv("netflix.csv")
amazon <- read.csv("amazon.csv")

#seperates Movies from TV Shows
netflixMoviesIndex <- which(netflix$type == "Movie")
netflixTvShowsIndex <- which(netflix$type == "TV Show")
amazonMoviesIndex <- which(netflix$type == "Movie")
amazonTvShowsIndex <- which(netflix$type == "TV Show")

#found the indices where movies are released either before or after the year 
2000
before2000Index <- which(netflix$release_year < 2000)
after2000Index <- which(netflix$release_year >= 2000)
Abefore2000Index <- which(amazon$release_year < 2000)
Aafter2000Index <- which(amazon$release_year >= 2000)

#implements the new column
beforeAfter2000 <- rep("Before 2000", nrow(netflix))
netflix <- cbind(netflix, beforeAfter2000)
netflix$beforeAfter2000[after2000Index] <- "2000 and After"
AbeforeAfter2000 <- rep("Before 2000", nrow(amazon))
amazon <- cbind(amazon, AbeforeAfter2000)
amazon$AbeforeAfter2000[Aafter2000Index] <- "2000 and After"

#creates a subset for graph use
netflixMovies <- netflix[netflixMoviesIndex, ]
amazonMovies <- amazon[amazonMoviesIndex, ]
```

6.2.2 Construction of Graph for Figure 3.2

```
netflix <- read.csv("netflix.csv")
amazon <- read.csv("amazon.csv")

#seperates Movies from TV Shows
netflixMoviesIndex <- which(netflix$type == "Movie")
netflixTvShowsIndex <- which(netflix$type == "TV Show")
amazonMoviesIndex <- which(netflix$type == "Movie")
amazonTvShowsIndex <- which(netflix$type == "TV Show")

#creating separate subset for movies and tv shows before and after 2000
Moviesbefore2000Index <- which(netflixMovies$beforeAfter2000 == "Before 2000")
Moviesafter2000Index <- which(netflixMovies$beforeAfter2000 == "2000 and
After")
netflixMoviesBefore2000 <- netflixMovies[Moviesbefore2000Index, ]
netflixMoviesAfter2000 <- netflixMovies[Moviesafter2000Index, ]

netflixTvShows <- netflix[netflixTvShowsIndex, ]
Showsbefore2000Index <- which(netflixTvShows$beforeAfter2000 == "Before 2000")
Showsafter2000Index <- which(netflixTvShows$beforeAfter2000 == "2000 and
After")
netflixShowsBefore2000 <- netflixTvShows[Showsbefore2000Index, ]
netflixShowsAfter2000 <- netflixTvShows[Showsafter2000Index, ]

AMoviesbefore2000Index <- which(amazonMovies$AbeforeAfter2000 == "Before 2000")
AMoviesafter2000Index <- which(amazonMovies$AbeforeAfter2000 == "2000 and
After")
amazonMoviesBefore2000 <- amazonMovies[AMoviesbefore2000Index, ]
amazonMoviesAfter2000 <- amazonMovies[AMoviesafter2000Index, ]

amazonTvShows <- amazon[amazonTvShowsIndex, ]
AShowsbefore2000Index <- which(amazonTvShows$AbeforeAfter2000 == "Before 2000")
AShowsafter2000Index <- which(amazonTvShows$AbeforeAfter2000 == "2000 and
After")
amazonShowsBefore2000 <- amazonTvShows[AShowsbefore2000Index, ]
amazonShowsAfter2000 <- amazonTvShows[AShowsafter2000Index, ]
```

```
#library allows for combining graphs into one
library(patchwork)

plot1 <- ggplot(netflixMoviesBefore2000, aes(x = rating, y =
after_stat(count/sum(count)))) +
  geom_bar(fill = "deepskyblue") + ylab("Proportion") + xlab("Rating") +
ggtitle("Netflix: Proportion of Ratings of Movies Before 2000")

plot2 <- ggplot(netflixMoviesAfter2000, aes(x = rating, y =
after_stat(count/sum(count)))) +
  geom_bar(fill = "deepskyblue") + ylab("Proportion") + xlab("Rating") +
ggtitle("Netflix: Proportion of Ratings of Movies After 2000")

plot3 <- ggplot(amazonMoviesBefore2000, aes(x = amazonMoviesBefore2000$rating,
y = after_stat(count/sum(count)))) +
  geom_bar(fill = "deepskyblue") + ylab("Proportion") + xlab("Rating") +
ggtitle("Amazon: Proportion of Ratings of Movies Before 2000")

plot4 <- ggplot(amazonMoviesAfter2000, aes(x = amazonMoviesAfter2000$rating, y
= after_stat(count/sum(count)))) +
  geom_bar(fill = "deepskyblue") + ylab("Proportion") + xlab("Rating") +
ggtitle("Amazon: Proportion of Ratings of Movies After 2000")

#creates a singular graph out of the four
layout <- wrap_plots(plot1, plot2, plot3, plot4, ncol = 2)

#displays the graph
layout
```

6.3 Code for Figure 3.3

6.3.1 Data Cleaning for Figure 3.3

```
#Removing all data points with empty values within “country”
netflixCountryIndex <- which(netflixData$country != "")
netflixCountryData <- netflixData[netflixCountryIndex, ]

#function that finds the mode of a dataset
find_mode <- function(x) {
  u <- unique(x)
  tab <- tabulate(match(x, u))
  u[tab == max(tab)]
}

library("tidyr")

#remove the commas between countries listed and separate unique countries
netflixCountryData <- netflixCountryData %>% separate_rows(country, sep=", ")

#fixing spelling differences
unitedStatesIndex <- which(netflixCountryData$country == "United States")
UKIndex <- which(netflixCountryData$country == "United Kingdom")
netflixCountryData[unitedStatesIndex, ]$country <- "USA"
netflixCountryData[UKIndex, ]$country <- "UK"

#create dataset called world
world <- data.frame()
world <- map_data("world")

TopRating = rep(NA, nrow(world))
world = cbind(world, TopRating)

# appends the most common rating for each country to “world”
for (i in unique(netflixCountryData$country)) {
  index <- which(netflixCountryData$country == i)
  x <- names(which.max(table(netflixCountryData[index, ]$rating)))
  if (length(which(world$region == i)) > 0) {
    index2 <- which(world$region == i)
    world[index2, ]$TopRating <- x
  }
 }

#same is repeated for Amazon Prime dataset
unitedStatesIndex <- which(amazonCountryData$country == "United States")
UKIndex <- which(amazonCountryData$country == "United Kingdom")

world2 <- data.frame()
```

```
world2 <- map_data("world")

amazonCountryIndex <- which(amazonData$country != "")
amazonCountryData <- amazonData[amazonCountryIndex, ]
amazonCountryData <- amazonCountryData %>% separate_rows(country, sep=", ")

amazonCountryData[unitedStatesIndex, ]$country <- "USA"
amazonCountryData[UKIndex, ]$country <- "UK"

TopRating <- rep(NA, nrow(world2))
world2 <- cbind(world2, TopRating)
allCountries <- unique(world2$region)
print(allCountries)
for (i in unique(amazonCountryData$country)) {
  index <- which(amazonCountryData$country == i)
  x <- names(which.max(table(amazonCountryData[index, ]$rating)))

  if (length(which(world2$region == i)) > 0) {
    index2 <- which(world2$region == i)
    world2[index2, ]$TopRating <- x
  }
}
```

6.3.2 Construction of Graph for Figure 3.3

```
#graph for Netflix dataset
ggplot(world) + geom_polygon(mapping = aes(x = long , y = lat , group = group ,
fill = TopRating), color = "black") + coord_quickmap() +
  ggtitle("The Most Common Rating on Netflix for Each Country")

#graph for Amazon Prime dataset
ggplot(world2) + geom_polygon(mapping = aes(x = long , y = lat , group = group
, fill = TopRating), color = "black") + coord_quickmap() +
  ggtitle("The Most Common Rating on Amazon for Each Country")
```

6.4 Code for Figure 3.4

6.4.1 Data Cleaning for Figure 3.4

```
# Question 1:  Are tv shows or movies released in comparison per year?
#  Read data
netflixData <- read.csv("netflix.csv", header = TRUE, sep = ",",
                        na.strings = "NA", stringsAsFactors = FALSE)
amazonData <- read.csv("amazon.csv", header = TRUE, sep = ",",
                    na.strings = "NA", stringsAsFactors = FALSE)

# Libraries needed: ggplot2 is used to construct graphs
# gridExtra is for grid.arrange capabilities once graphs are constructed
library(ggplot2)
library(gridExtra)

# Create new dataframes for use with this question and graphs
netflix_year <- netflixData
amazon_year <- amazonData

# Create character vector with names of variables needed to make graphs
year_vars <- c("release_year", "type")

# Dataframes with only variables that are in the year_vars vector
netflix_year <- netflixData[year_vars]
amazon_year <-  amazonData[year_vars]

# Factor type variable in each dataframe with two levels: Movie and TV Show
netflix_year$type <- factor(netflix_year$type, levels = c("Movie", "TV Show"),
                            labels = c("Movie", "TV Show"))
amazon_year$type <- factor(amazon_year$type, levels = c("Movie", "TV Show"),
                            labels = c("Movie", "TV Show"))
```

6.4.2 Construction of Graph for Figure 3.4

```
# Create bar graph for netflix_year
# Counts for bar graph displayed using log10 transformation and after_stat in
aesthetics
netflix_bar <- ggplot(data = netflix_year) +
  geom_bar(mapping = aes(x = release_year, y = after_stat(log10(count)), fill =
type)) +
  labs(x = "Release Year", y = "log10 Count", title = "Available Netflix
listings based on title release year")

# Create bar graph for amazon_year
# Counts for bar graph displayed using log10 transformation and after_stat in
aesthetics
amazon_bar <- ggplot(data = amazon_year) +
  geom_bar(mapping = aes(x = release_year, y = after_stat(log10(count)), fill =
type)) +
  labs(x = "Release Year", y = "log10 Count", title = "Available Amazon
listings based on title release year")

# Arrange graph to show each on one row
grid.arrange(netflix_bar, amazon_bar, nrow = 2)
```

7 References

[1] Bansal, S. (2021, September 27). *Netflix Movies and TV Shows*. Kaggle.
https://www.kaggle.com/datasets/shivamb/netflix-shows

[2] Bansal, S. (2021b, October 12). *Amazon Prime Movies and TV Shows*. Kaggle.
https://www.kaggle.com/datasets/shivamb/amazon-prime-movies-and-tv-shows