A Analysis of Netflix's Content Distribution

ManavSharma-400451088

2023-04-17

Introduction

Netflix is a popular worldwide streaming platform that offers a wide variety of movies, TV shows, and documentaries to its subscribers. Founded in 1997, the platform allows its subscribers to have access to millions of titles through a monthly subscription. Over the years the platform has reached over 190 countries, it has dominated its market and become a billion dollar company, consistently earning billions in revenue annually.

In this analysis, we will seek to answer three key questions regarding Netflix's content distribution strategy:

1. What types of media are most commonly released on the platform?
2. Which countries release the most content?
3. Which rating categories are most frequently released?

The Netflix dataset from Kaggle.com is the dataset that I will be using to conduct this analysis. It is a valuable resource for analyzing the companies content distribution. The dataset contains information on thousands of movies and TV shows available on Netflix, including title, country of release, rating, genre, and more. What makes this data set interesting is that it will be useful for answering my questions since It has all the columns that I need and it will give us a insight on other things such as Netflix's distribution strategy.

Data Wrangling Plan

Iteration 1:

Phase 1:

1. Read the csv file into R
2. Make columns lowercase
3. Determine if data is tidy
4. Reorder so uid's are first and drop unnecessary columns

Phase 2:

```r
#1
data <- read_csv("/Users/manavsharma/Desktop/Data Science/Final project/netflix1.csv")
```

```
## Rows: 8790 Columns: 10
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (9): show_id, type, title, director, country, date_added, rating, durati...
## dbl (1): release_year
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
#2
data %<>%
  rename_with(tolower)
#3
data %>%
  glimpse()#The data is tidy
```

```
## Rows: 8,790
## Columns: 10
## $ show_id      <chr> "s1", "s3", "s6", "s14", "s8", "s9", "s10", "s939", "s13"~
## $ type         <chr> "Movie", "TV Show", "TV Show", "Movie", "Movie", "TV Show~
## $ title        <chr> "Dick Johnson Is Dead", "Ganglands", "Midnight Mass", "Co~
## $ director     <chr> "Kirsten Johnson", "Julien Leclercq", "Mike Flanagan", "B~
## $ country      <chr> "United States", "France", "United States", "Brazil", "Un~
## $ date_added   <chr> "9/25/2021", "9/24/2021", "9/24/2021", "9/22/2021", "9/24~
## $ release_year <dbl> 2020, 2021, 2021, 2021, 1993, 2021, 2021, 2019, 2021, 201~
## $ rating       <chr> "PG-13", "TV-MA", "TV-MA", "TV-PG", "TV-MA", "TV-14", "PG~
## $ duration     <chr> "90 min", "1 Season", "1 Season", "91 min", "125 min", "9~
## $ listed_in    <chr> "Documentaries", "Crime TV Shows, International TV Shows,~
```

```r
#4
data %>%
  count(title)%>%
  filter(n>1) #title is a unique uid after removing the duplicate movies
```

```
## # A tibble: 3 x 2
##   title      n
##   <chr>  <int>
## 1 15-Aug     2
## 2 22-Jul     2
## 3 9-Feb      2
```

```r
data <- data %>%
  select(title,type,country,rating)

data %>%
  glimpse()
```

```
## Rows: 8,790
## Columns: 4
## $ title   <chr> "Dick Johnson Is Dead", "Ganglands", "Midnight Mass", "Confess~
## $ type    <chr> "Movie", "TV Show", "TV Show", "Movie", "Movie", "TV Show", "M~
## $ country <chr> "United States", "France", "United States", "Brazil", "United ~
## $ rating  <chr> "PG-13", "TV-MA", "TV-MA", "TV-PG", "TV-MA", "TV-14", "PG-13",~
```

Iteration 2

Phase 1:

1. Check for NA values
2. Convert any misleading values such as “Not Given” into NA values
3. Remove any duplicate values
4. Drop all NA values
5. Convert columns to factor columns
6. Sort by UID

Phase 2:

```
#1
data %>%
  summary() #No NA values
```

```
##     title               type             country             rating
##  Length:8790        Length:8790        Length:8790        Length:8790
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

```
"Not Given" %in% data$country #There are "Not Given" values present
```

```
## [1] TRUE
```

```
#2
data <- data %>%
  mutate(country = na_if(country, "Not Given"))

"Not Given" %in% data$country #"Not Given" values have been removed
```

```
## [1] FALSE
```

```
#3
data %>%
  count(title) %>%
  filter(n>1) #There are duplicates
```

```
## # A tibble: 3 x 2
##   title      n
##   <chr>  <int>
## 1 15-Aug     2
## 2 22-Jul     2
## 3 9-Feb      2
```

```
data <- data %>%
  distinct(title, .keep_all = TRUE)

data %>%
  count(title) %>%
  filter(n>1) #Duplicates have been removed and title is now a unique UID
```

```
## # A tibble: 0 x 2
## # ... with 2 variables: title <chr>, n <int>
```

```
#4
data %>%
  dim_desc()
```

```
## [1] "[8,787 x 4]"
```

```
data %<>%
  drop_na()

data %>%
  dim_desc() #287 rows containing NA values have been dropped
```

```
## [1] "[8,500 x 4]"
```

```
#5
data %<>%
  mutate_all(as.factor)
```

```
#6
data %<>%
  arrange(title)

data %>%
  glimpse()
```

```
## Rows: 8,500
## Columns: 4
## $ title   <fct> "¡Ay, mi madre!", "'76", "'89", "(T)ERROR", "(Un)Well", "#Aliv~
## $ type    <fct> Movie, Movie, Movie, Movie, TV Show, Movie, Movie, TV Show, Mo~
## $ country <fct> Spain, Nigeria, United Kingdom, United States, United States, ~
## $ rating  <fct> TV-MA, TV-PG, TV-PG, NR, TV-MA, TV-MA, TV-14, TV-MA, TV-14, TV~
```

Data Wrangling Plan is finished

Results/Discussion:

Question 1:

What types of media are most commonly released on the platform?

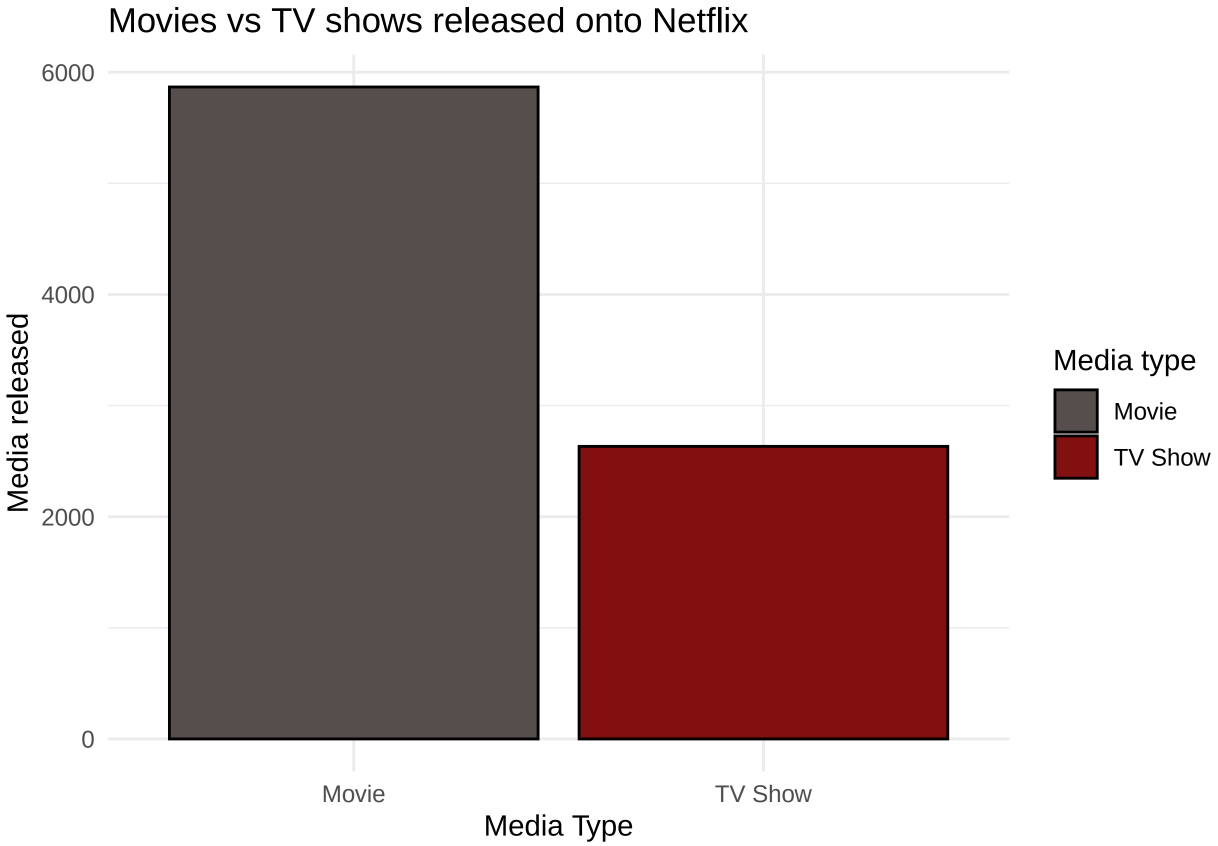

Figure 1: Depicts that the platform Netflix exhibits a greater tendency to release movies.

There could be several reasons why movies are released more frequently on Netflix than TV shows. One major reason is that movies have a wider appeal and are marketed much more heavily compared to the average TV show causing movies to attract a bigger audience. Additionally, producing a TV show requires ongoing production and investment, while movies are often one-time investments. Finally, Netflix may have a larger selection of movies available for licensing compared to TV shows.

Question 2:

What countries release the most content?

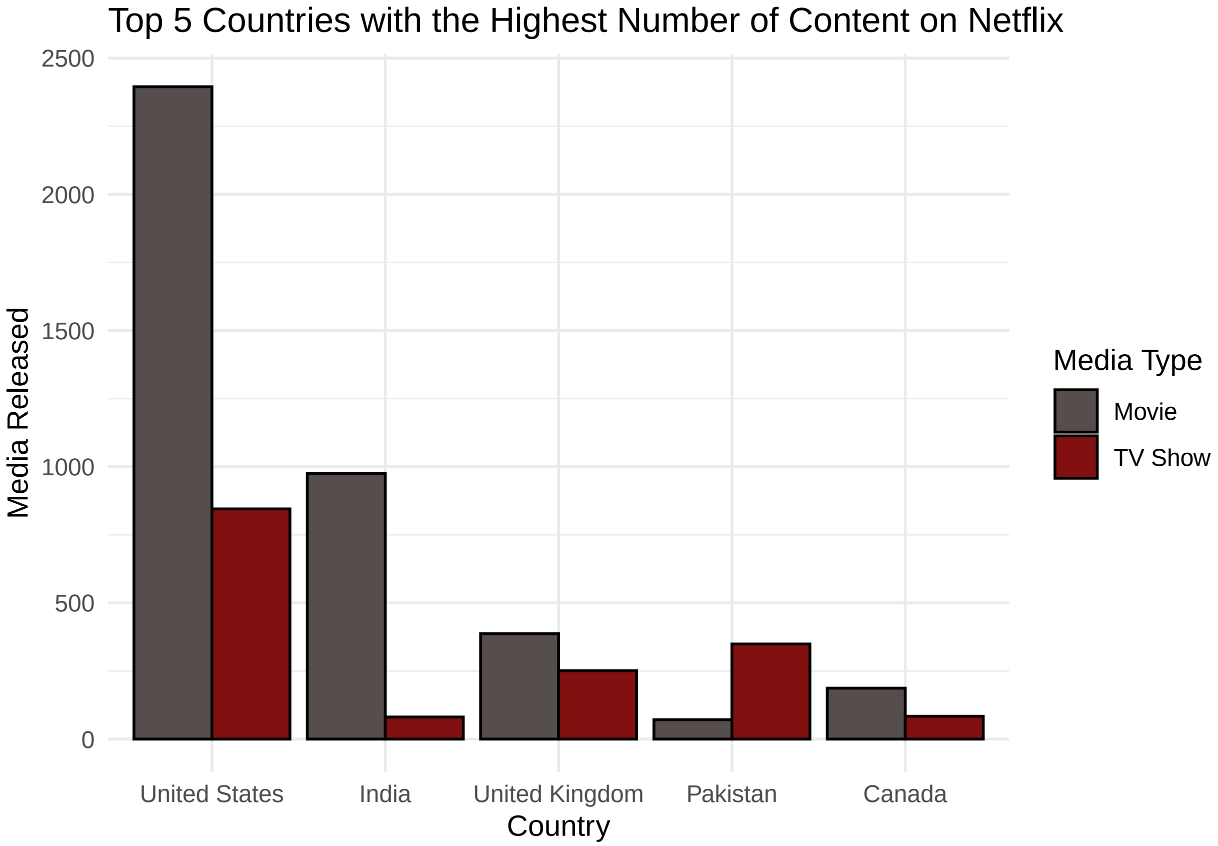

Figure 2: Displays the countries with the most media content released on Netflix, ranked in descending order.

From our result we can conclude that the United States by far releases the most media onto Netflix. My hypothesis for this question was incorrect, India falls significantly short in terms of the number of releases compared to the United States. While Bollywood surpasses Hollywood in terms of production, Netflix acquires more films for its larger demographic audience, such as Canada, the United States, and Europe.

Question 3:

Which rating categories are most frequently released?

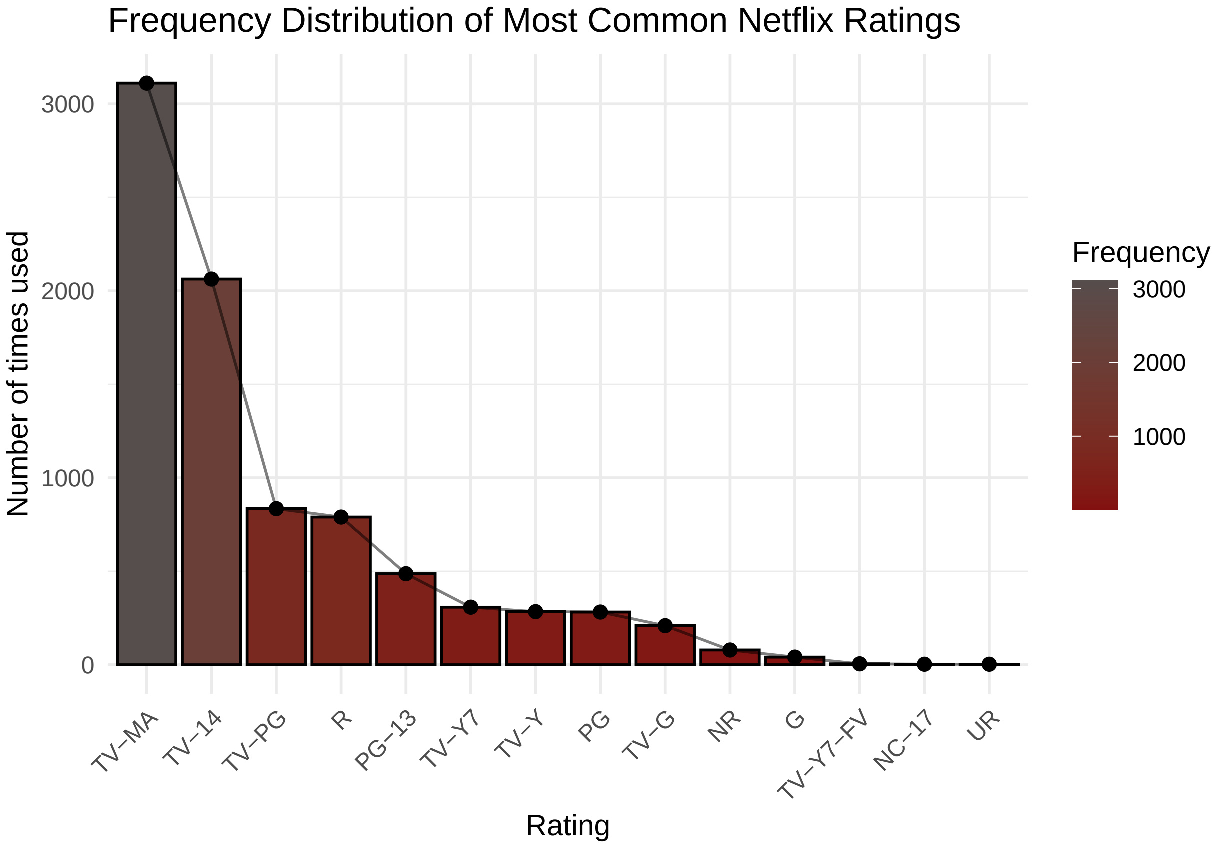

Figure 3: Displays all of Netflix's ratings in order from most frequent to least.

After analyzing the data, it turns out that my initial prediction of PG-13 being the most commonly used rating on Netflix is incorrect. Surprisingly, the most frequently used rating is TV-MA. This finding is particularly interesting as I expected Netflix to produce more PG-13 content considering the large number of children who watch TV for extended periods of time. However, the prevalence of TV-MA rating is significantly higher than that of PG-13.

Conclusion

My analysis provides insights into three different hypotheses related to the content available on Netflix.

1. My first hypothesis was that movies are released more frequently than TV shows on the platform. My analysis supports this hypothesis, and we have identified several reasons why this may be the case, including wider appeal, heavy marketing, and easier licensing.
2. My second hypothesis was that India, being the largest producer of films, releases the most media onto Netflix. However, my analysis found that the United States is the clear leader in terms of the number of releases, followed by India. This could be because Netflix acquires more films for its larger demographic audience in North America and Europe.
3. My third hypothesis was that PG-13 would be the most common rating for content on Netflix. However, my analysis found that TV-MA is the most commonly used rating, which was unexpected, given the large number of children who watch TV on the platform.

Overall, while some of our hypotheses were proven incorrect, the data analysis provides us with valuable insights into the trends of media released on Netflix.