Netflix 2019 Dataset

Irene Limcolioc

November 2020

Overview

This a dataset of Netflix as of 2019. It consists of TV Shows and Movies available in Netflix. The dataset can be found in: https://www.kaggle.com/shivamb/netflix-shows.

I will explore and observe the variables and values in the given data by Data Reading and Data Visualizations. I will use 2 types of libraries: `ggplot2` and `tidyverse`.

Data Reading

```
netflix = read.csv("~/Downloads/netflix_titles.csv")
```

In the dataset, there are a total of 6234 observations of 12 variables describing TV Shows and Movies.

I will now make some observations and get familiarize with the data `netflix`.

```
colnames(netflix)
```

```
##  [1] "show_id"      "type"         "title"        "director"     "cast"
##  [6] "country"      "date_added"   "release_year" "rating"       "duratio
n"
## [11] "listed_in"    "description"
```

```
summary(netflix)
```

```
##     show_id             type               title             director
##  Min.   :  247747   Length:6234        Length:6234        Length:6234
##  1st Qu.:80035802   Class :character   Class :character   Class :characte
r
##  Median :80163367   Mode  :character   Mode  :character   Mode  :characte
r
##  Mean   :76703679
##  3rd Qu.:80244889
##  Max.   :81235729
##      cast             country           date_added         release_year
##  Length:6234        Length:6234        Length:6234        Min.   :1925
##  Class :character   Class :character   Class :character   1st Qu.:2013
##  Mode  :character   Mode  :character   Mode  :character   Median :2016
##                                                           Mean   :2013
##                                                           3rd Qu.:2018
##                                                           Max.   :2020
##     rating            duration          listed_in         description
##  Length:6234        Length:6234        Length:6234        Length:6234
##  Class :character   Class :character   Class :character   Class :characte
r
##  Mode  :character   Mode  :character   Mode  :character   Mode  :characte
r
##
##
##
```

```
head(netflix)
```

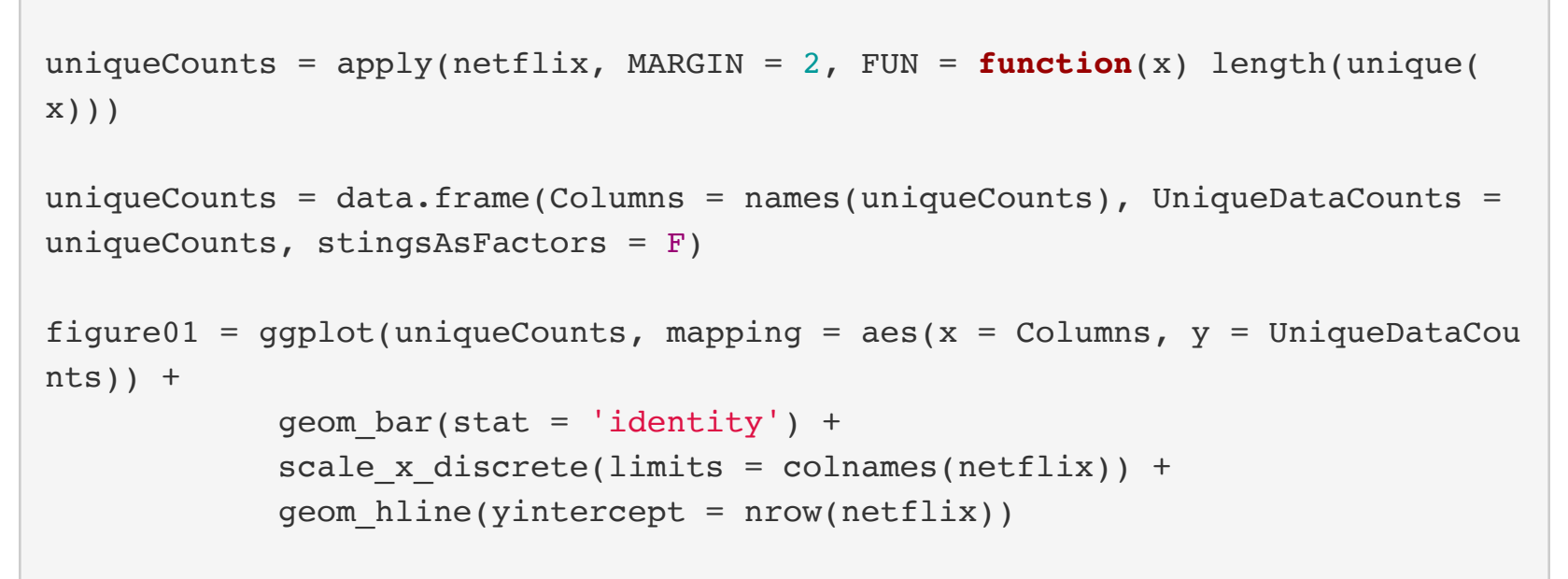
```
##    show_id    type                                    title
## 1 81145628   Movie Norm of the North: King Sized Adventure
## 2 80117401   Movie               Jandino: Whatever it Takes
## 3 70234439 TV Show                       Transformers Prime
## 4 80058654 TV Show         Transformers: Robots in Disguise
## 5 80125979   Movie                             #realityhigh
## 6 80163890 TV Show                                  Apaches
##                   director
## 1 Richard Finn, Tim Maltby
## 2
## 3
## 4
## 5         Fernando Lebrija
## 6
##
cast
## 1                                            Alan Marriott, Andrew Toth, Bria
n Dobson, Cole Howard, Jennifer Cameron, Jonathan Holmes, Lee Tockar, Lisa D
urupt, Maya Kay, Michael Dobson
## 2
Jandino Asporaat
## 3 Peter Cullen, Sumalee Montano, Frank Welker, Jeffrey Combs, Kevin Micha
el Richardson, Tania Gunadi, Josh Keaton, Steve Blum, Andy Pessoa, Ernie Hud
son, Daran Norris, Will Friedle
## 4                                                             Will Friedle,
Darren Criss, Constance Zimmer, Khary Payton, Mitchell Whitfield, Stuart All
an, Ted McGinley, Peter Cullen
## 5           Nesta Cooper, Kate Walsh, John Michael Higgins, Keith Powers,
Alicia Sanz, Jake Borelli, Kid Ink, Yousef Erakat, Rebekah Graf, Anne Winter
s, Peter Gilroy, Patrick Davis
## 6
Alberto Ammann, Eloy Azorín, Verónica Echegui, Lucía Jiménez, Claudia Traisa
c
##                                     country        date_added release_year
## 1 United States, India, South Korea, China September 9, 2019         2019
## 2                           United Kingdom September 9, 2016         2016
## 3                            United States September 8, 2018         2013
## 4                            United States September 8, 2018         2016
## 5                            United States September 8, 2017         2017
## 6                                    Spain September 8, 2017         2016
##     rating duration
## 1    TV-PG   90 min
## 2    TV-MA   94 min
## 3 TV-Y7-FV 1 Season
## 4    TV-Y7 1 Season
## 5    TV-14   99 min
## 6    TV-MA 1 Season
##                                                             listed_in
## 1                                  Children & Family Movies, Comedies
## 2                                                     Stand-Up Comedy
## 3                                                            Kids' TV
## 4                                                            Kids' TV
## 5                                                            Comedies
## 6 Crime TV Shows, International TV Shows, Spanish-Language TV Shows
##
description
## 1         Before planning an awesome wedding for his grandfather, a polar
bear king must take back a stolen artifact from an evil archaeologist first.
## 2    Jandino Asporaat riffs on the challenges of raising kids and serenad
es the audience with a rousing rendition of "Sex on Fire" in his comedy sho
w.
## 3         With the help of three human allies, the Autobots once again pr
otect Earth from the onslaught of the Decepticons and their leader, Megatro
n.
## 4                          When a prison ship crash unleashes hundreds of De
cepticons on Earth, Bumblebee leads a new Autobot force to protect humankin
d.
## 5 When nerdy high schooler Dani finally attracts the interest of her long
time crush, she lands in the cross hairs of his ex, a social media celebrit
y.
## 6             A young journalist is forced into a life of crime to save h
is father and family in this series based on the novel by Miguel Sáez Carra
l.
```

```
tail(netflix)
```

```
##       show_id    type                                       title directo
r
## 6229 80159925 TV Show                                    Kikoriki
## 6230 80000063 TV Show                                 Red vs. Blue
## 6231 70286564 TV Show                                       Maron
## 6232 80116008   Movie      Little Baby Bum: Nursery Rhyme Friends
## 6233 70281022 TV Show A Young Doctor's Notebook and Other Stories
## 6234 70153404 TV Show                                     Friends
##
cast
## 6229
Igor Dmitriev
## 6230 Burnie Burns, Jason Saldaña, Gustavo Sorola, Geoff Lazer Ramsey, Joe
l Heyman, Matt Hullum, Dan Godwin, Kathleen Zuelch, Yomary Cruz, Nathan Zell
ner
## 6231
Marc Maron, Judd Hirsch, Josh Brener, Nora Zehetner, Andy Kindler
## 6232
## 6233                 Daniel Radcliffe, Jon Hamm, Adam Godley, Christopher G
odwin, Rosie Cavaliero, Vicki Pepperdine, Margaret Clunie, Tim Steed, Shaun
Pye
## 6234                                                         Jennifer An
iston, Courteney Cox, Lisa Kudrow, Matt LeBlanc, Matthew Perry, David Schwim
mer
##             country date_added release_year rating   duration
## 6229                                     2010   TV-Y  2 Seasons
## 6230  United States                      2015     NR 13 Seasons
## 6231  United States                      2016  TV-MA  4 Seasons
## 6232                                     2016            60 min
## 6233 United Kingdom                      2013  TV-MA  2 Seasons
## 6234  United States                      2003  TV-14 10 Seasons
##                                                   listed_in
## 6229                                               Kids' TV
## 6230 TV Action & Adventure, TV Comedies, TV Sci-Fi & Fantasy
## 6231                                            TV Comedies
## 6232                                                 Movies
## 6233               British TV Shows, TV Comedies, TV Dramas
## 6234                         Classic & Cult TV, TV Comedies
##
description
## 6229     A wacky rabbit and his gang of animal pals have fun solving prob
lems, sharing stories and exploring their sometimes magical, always special
world.
## 6230   This parody of first-person shooter games, military life and scien
ce-fiction films centers on a civil war fought in the middle of a desolate c
anyon.
## 6231 Marc Maron stars as Marc Maron, who interviews fellow comedians for
his popular podcast, only to reveal more about his own neuroses and relation
ships.
## 6232     Nursery rhymes and original music for children accompanied by br
ight, playful animation engage and educate about numbers, shapes, colors and
more.
## 6233     Set during the Russian Revolution, this comic miniseries is base
d on a doctor's memories of his early career working in an out-of-the-way vi
llage.
## 6234      This hit sitcom follows the merry misadventures of six 20-somet
hing pals as they navigate the pitfalls of work, life and love in 1990s Manh
attan.
```

```
netflix$date_added = as.Date(netflix$date_added, format = "%B %d, %Y")
```

```
library(ggplot2)

# I would like to get a sense of unique values in all of the 12 variables. B
y using the `unique` function, I will be able to see duplicate elements or i
dentical values in the dataset.

uniqueCounts = apply(netflix, MARGIN = 2, FUN = function(x) length(unique(
x)))

uniqueCounts = data.frame(Columns = names(uniqueCounts), UniqueDataCounts =
uniqueCounts, stingsAsFactors = F)

figure01 = ggplot(uniqueCounts, mapping = aes(x = Columns, y = UniqueDataCou
nts)) +
            geom_bar(stat = 'identity') +
            scale_x_discrete(limits = colnames(netflix)) +
            geom_hline(yintercept = nrow(netflix))

print(figure01)
```



By observation, this plot shows `show_id`, `description`, and `title` are particularly unique. This tells us that there is no duplicate input of data for these three variables. We can also clearly see that there are possible missing values in `director`, `cast`, `country`, `data_added`, and `rating` variables.

Data Visualizations

The following Visualizations are: The Total Number of Netflix Content by Type, The Total Number of Genres, and The Total Number of Movies and TV Shows Available Over Time.

The Total Number of Netflix Content by Type:

Which content does Netflix provide more of for viewing?

```
# I will use the library `titdyverse`. The operator `%>%` extremely improves
my ability to build a function that is less difficult to design. Here, I use
the function `%>%` to pass the operator from the `netflix` data to operators
`group_by ()` and then to `summarise ()`.

library(tidyverse)
```

```
## ── Attaching packages ───────────────────────────────────────
──────────────────── tidyverse 1.3.0 ──
```

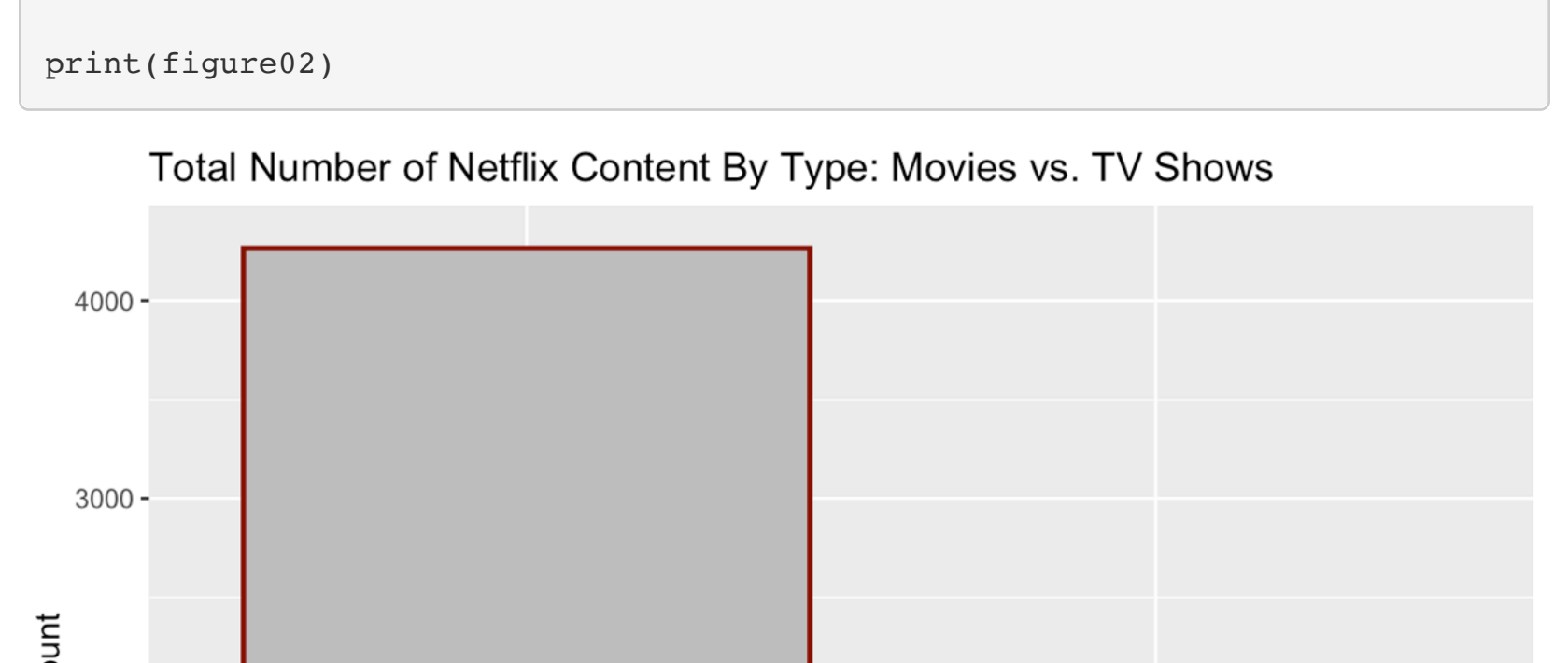
```
## ✓ tibble  3.0.3     ✓ dplyr   1.0.2
## ✓ tidyr   1.1.2     ✓ stringr 1.4.0
## ✓ readr   1.4.0     ✓ forcats 0.5.0
## ✓ purrr   0.3.4
```

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

```
dfAmountByType = netflix %>% group_by(type) %>% summarise(count = n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
figure02 =  ggplot(data = dfAmountByType, aes(x= type, y= count, fill= typ
e)) +
             geom_bar(colour ="dark red", size= 0.8, fill = "gray " ,  stat
= "identity") +
             guides(fill= FALSE) +
             xlab("Content by Type") + ylab("Amount") +
             ggtitle("Total Number of Netflix Content By Type: Movies vs. TV
Shows")

print(figure02)
```



Based on the visualization, there are twice as much of Movies than TV Shows available on Netflix.

The Total Number of Genres:

What are the top genres of Movies and TV Shows available in Netflix?

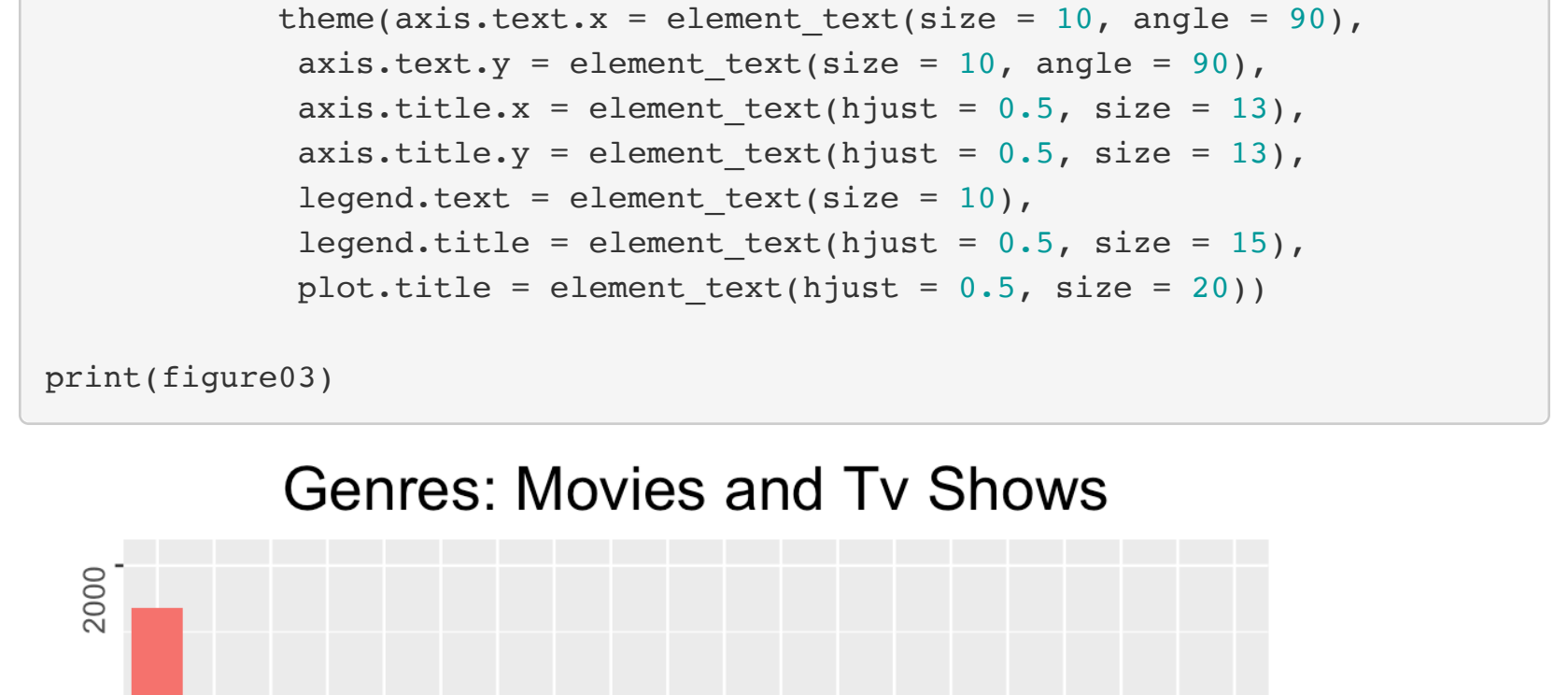
```
# The `trimws ()` removes the whitespace from the left from the character st
rings in the variable `listed_in`.

# And for the sake of argument, I will first show the first 50 listed types
of Genres offered in Netflix.

netflix$listed_in = trimws(netflix$listed_in, which = 'left')

netflix$listed_in[1:50]
```

```
##  [1] "Children & Family Movies, Comedies"
##  [2] "Stand-Up Comedy"
##  [3] "Kids' TV"
##  [4] "Kids' TV"
##  [5] "Comedies"
##  [6] "Crime TV Shows, International TV Shows, Spanish-Language TV Shows"
##  [7] "International Movies, Sci-Fi & Fantasy, Thrillers"
##  [8] "Stand-Up Comedy"
##  [9] "Docuseries, Science & Nature TV"
## [10] "Action & Adventure, Thrillers"
## [11] "Stand-Up Comedy"
## [12] "Action & Adventure, Dramas, International Movies"
## [13] "Children & Family Movies"
## [14] "Children & Family Movies"
## [15] "Children & Family Movies"
## [16] "Children & Family Movies"
## [17] "Children & Family Movies"
## [18] "Children & Family Movies"
## [19] "Children & Family Movies"
## [20] "Cult Movies, Dramas, Independent Movies"
## [21] "Comedies, Independent Movies, Romantic Movies"
## [22] "Action & Adventure, Comedies, International Movies"
## [23] "Documentaries"
## [24] "Horror Movies, Thrillers"
## [25] "Dramas, Independent Movies"
## [26] "Dramas, Independent Movies, Romantic Movies"
## [27] "International TV Shows, Romantic TV Shows, TV Comedies"
## [28] "Documentaries"
## [29] "Docuseries"
## [30] "Horror Movies, International Movies"
## [31] "Children & Family Movies, Comedies, Sci-Fi & Fantasy"
## [32] "Comedies, Romantic Movies"
## [33] "Dramas, International Movies, Thrillers"
## [34] "Dramas, Independent Movies, International Movies"
## [35] "Kids' TV, TV Comedies"
## [36] "Dramas, International Movies, Thrillers"
## [37] "Comedies, Dramas, Independent Movies"
## [38] "Comedies, International Movies"
## [39] "Comedies, International Movies, Romantic Movies"
## [40] "International TV Shows, TV Dramas, TV Thrillers"
## [41] "Action & Adventure, Comedies, Independent Movies"
## [42] "Comedies, Dramas, International Movies"
## [43] "Dramas, Independent Movies"
## [44] "Documentaries"
## [45] "Documentaries"
## [46] "Comedies, International Movies, Romantic Movies"
## [47] "Comedies, International Movies, Romantic Movies"
## [48] "Comedies, International Movies, Romantic Movies"
## [49] "Dramas, International Movies, Romantic Movies"
## [50] "Horror Movies, International Movies"
```

```
# By grouping `type` and `genre` together` and flattening `genre` back out i
nto it's regular column, a new variable with `genre` and a split character v
ector `listed_in` will be created.

dfGenres = netflix %>% mutate(genre = strsplit(listed_in, ',')) %>% unnest(g
enre) %>% group_by(type, genre)                %>% summarise(count = n()) %>%
unique() %>% arrange(desc(count)) %>% top_n(10, count)
```

```
## `summarise()` regrouping output by 'type' (override with `.groups` argume
nt)
```

```
figure03 = ggplot(data = dfGenres, aes(x = fct_reorder(genre,count,.desc =
T), y = count, fill = type)) +
            geom_col() +
            scale_y_continuous(limits = c(0,2000), breaks = seq(0, 2000, 50
0)) +
            labs(title = 'Genres: Movies and Tv Shows', x = 'Genres', y = 'A
mount') +
            theme(axis.text.x = element_text(size = 10, angle = 90),
             axis.text.y = element_text(size = 10, angle = 90),
             axis.title.x = element_text(hjust = 0.5, size = 13),
             axis.title.y = element_text(hjust = 0.5, size = 13),
             legend.text = element_text(size = 10),
             legend.title = element_text(hjust = 0.5, size = 15),
             plot.title = element_text(hjust = 0.5, size = 20))

print(figure03)
```



By observation, consider the following:

1.) It appears that International Movies and International TV Shows are the top genre accessible in Netflix.

2.) Movie genres has the highest amount in comparison to TV genres offered in Netflix.

3.) The topmost 3 Movie genres are: International Movies, Drama, and Comedies. Similarly, the top 3 TV genres are: International TV Shows, Drama, and TV Comedies.

4.) The bottommost 3 Movie genres are: Romantic Movies, Children & Family Movies, and Thrillers. Likewise, the bottommost 3 TV genres are: British TV Shows, Docuseries, and Korean TV Shows.

The Total Number of Movies and TV Shows Available Over Time:

Is there any correlation between Movies and TV Shows offered by Netflix over time?

```
netflix$date_added = as.Date(netflix$date_added, format = "%B %d, %Y")

netflix$date_added[1:30]
```

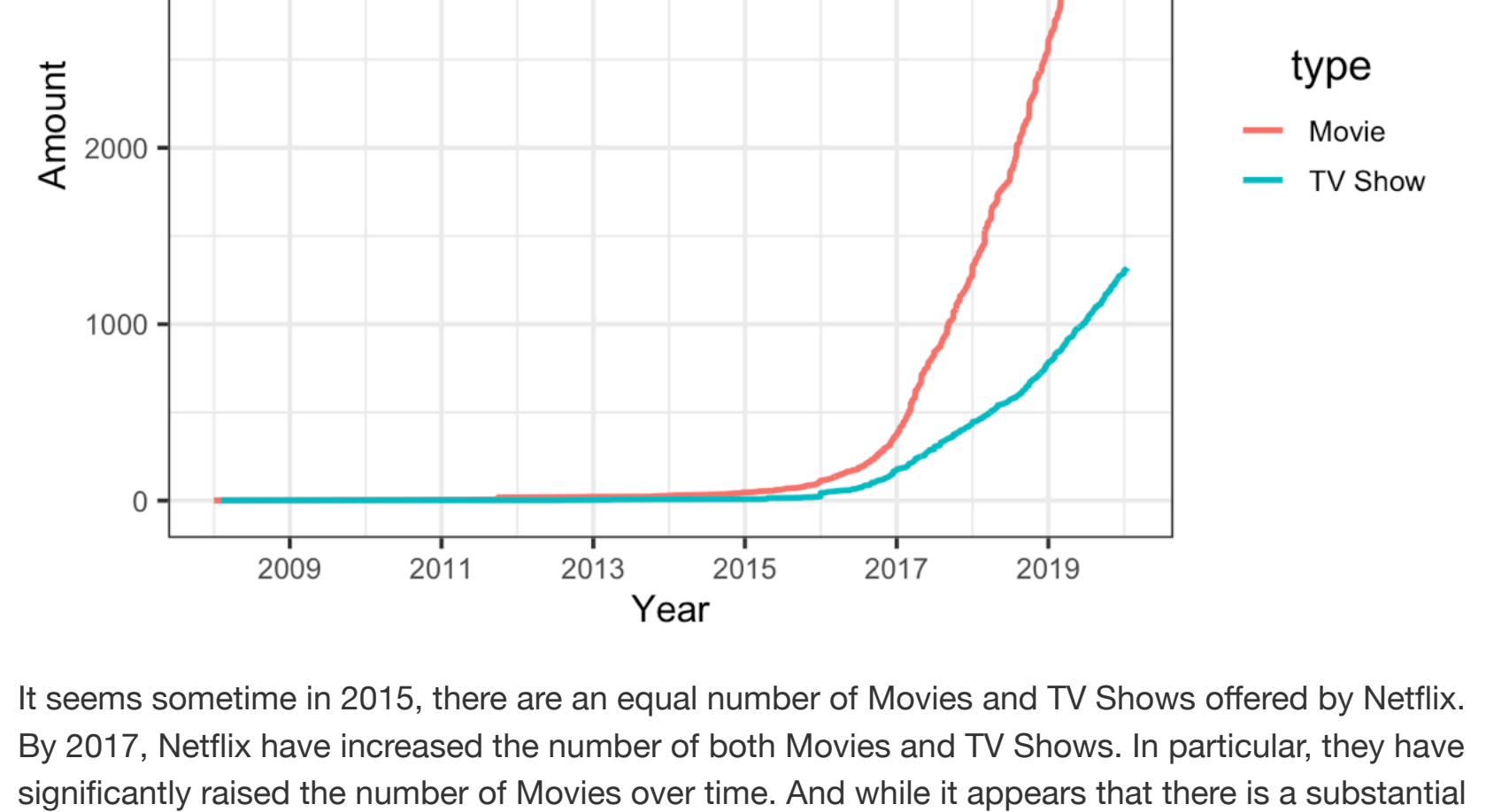
```
##  [1] "2019-09-09" "2016-09-09" "2018-09-08" "2018-09-08" "2017-09-08"
##  [6] "2017-09-08" "2017-09-08" "2017-09-08" "2017-09-08" "2017-09-08"
## [11] "2017-09-08" "2017-09-08" "2017-09-08" "2017-09-08" "2017-09-08"
## [16] "2017-09-08" "2017-09-08" "2017-09-08" "2017-09-08" "2017-09-08"
## [21] "2017-09-08" "2017-09-08" "2017-09-08" "2017-09-08" "2017-09-08"
## [26] "2015-09-08" "2018-09-07" "2018-09-07" "2018-09-07" "2018-09-07"
```

```
# Variables `type` and `date_added` will be combined to create a new data fr
ame.

dfDateTimePeriod = netflix %>% group_by(type, date_added) %>% summarise(coun
tt = n()) %>%
                 mutate(total_shows = cumsum(countt))
```

```
## `summarise()` regrouping output by 'type' (override with `.groups` argume
nt)
```

```
figure04 = ggplot(data = dfDateTimePeriod, aes(x = date_added, y = total_sho
ws, color = type)) +
           geom_line(size = 1) + theme_bw(base_size = 15) +
           scale_x_date(breaks = '2 years', date_labels = '%Y') +
           labs(title = 'Movies and TV Shows Over Years', x = 'Year', y = 'A
mount') +
           theme(axis.text.x = element_text(size = 10),
              axis.text.y = element_text(size = 10),
              axis.title.x = element_text(hjust = 0.5,size = 13),
              axis.title.y = element_text(hjust = 0.5,size = 13),
              legend.text = element_text(size = 10),
              legend.title = element_text(hjust = 0.5,size = 15),
              plot.title = element_text(hjust = 0.5,size =20))

print(figure04)
```

```
## Warning: Removed 2 row(s) containing missing values (geom_path).
```



It seems sometime in 2015, there are an equal number of Movies and TV Shows offered by Netflix. By 2017, Netflix have increased the number of both Movies and TV Shows. In particular, they have significantly raised the number of Movies over time. And while it appears that there is a substantial number of Movies available in comparison to TV Shows, Netflix have boosted both types in a span of 4 to 5 years. One can only suggest that this is due to the possibility of viewer demands. And imagine how this visualization would dramatically shift from 2020 and on.