

Netflix Data Visualisation Project in R

Iona

14/01/2021

This project will use data visualisation techniques on the Netflix Movies and TV Shows using R. We will calculate which year will have the most releases. The most popular genres on netflix for tv shows and movies will be uncovered and the most frequent director in the dataset. These will also be visualised in different types of plots throughout.

Data available from <https://www.kaggle.com/shivamb/netflix-shows>.

Explore the Data

First the data set and the relevant packages will be loaded then the first few lines of data will be viewed.

```
#import relevant packages
library(tidyverse)

## — Attaching packages ————— tidyverse 1.3.0 —

## ✓ ggplot2 3.3.2      ✓ purrr  0.3.4
## ✓ tibble  3.0.3      ✓ dplyr  1.0.2
## ✓ tidyr   1.1.2      ✓ stringr 1.4.0
## ✓ readr   1.4.0      ✓ forcats 0.5.0

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

library(wordcloud)

## Loading required package: RColorBrewer

library(tokenizers)
library(repr)
library(ggplot2)

#read the netflix dataset
netflix <- read.csv("netflix_titles.csv")

#view the first few lines of data
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, Brian
Dobson, Cole Howard, Jennifer Cameron, Jonathan Holmes, Lee Tockar, Lisa Duru
pt, Maya Kay, Michael Dobson
## 2
Jandino Asporaat
## 3 Peter Cullen, Sumalee Montano, Frank Welker, Jeffrey Combs, Kevin Michae
l Richardson, Tania Gunadi, Josh Keaton, Steve Blum, Andy Pessoa, Ernie Hudso
n, Daran Norris, Will Friedle
## 4                                Will Friedle,
Darren Criss, Constance Zimmer, Khary Payton, Mitchell Whitfield, Stuart Alla
n, 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 Winters
, Peter Gilroy, Patrick Davis
## 6
Alberto Ammann, Eloy Azorín, Verónica Echegui, Lucía Jiménez, Claudia Traisac
##
##      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 serenade
s the audience with a rousing rendition of "Sex on Fire" in his comedy show.
## 3 With the help of three human allies, the Autobots once again pro
tect Earth from the onslaught of the Decepticons and their leader, Megatron.
## 4 When a prison ship crash unleashes hundreds of Dec
epticons on Earth, Bumblebee leads a new Autobot force to protect humankind.
## 5 When nerdy high schooler Dani finally attracts the interest of her longt
ime crush, she lands in the cross hairs of his ex, a social media celebrity.
## 6 A young journalist is forced into a life of crime to save hi
s father and family in this series based on the novel by Miguel Sáez Carral.
```

Top Release Year

After grouping the data by release year we can count both types released each year. This is shown below where it is clear that there is spike of numbers after the year 2000.

```
# group the data by release date
group_by_date <- netflix %>% group_by(release_year)

# count the number of shows released per date
count_by_date <- netflix %>% count(release_year)

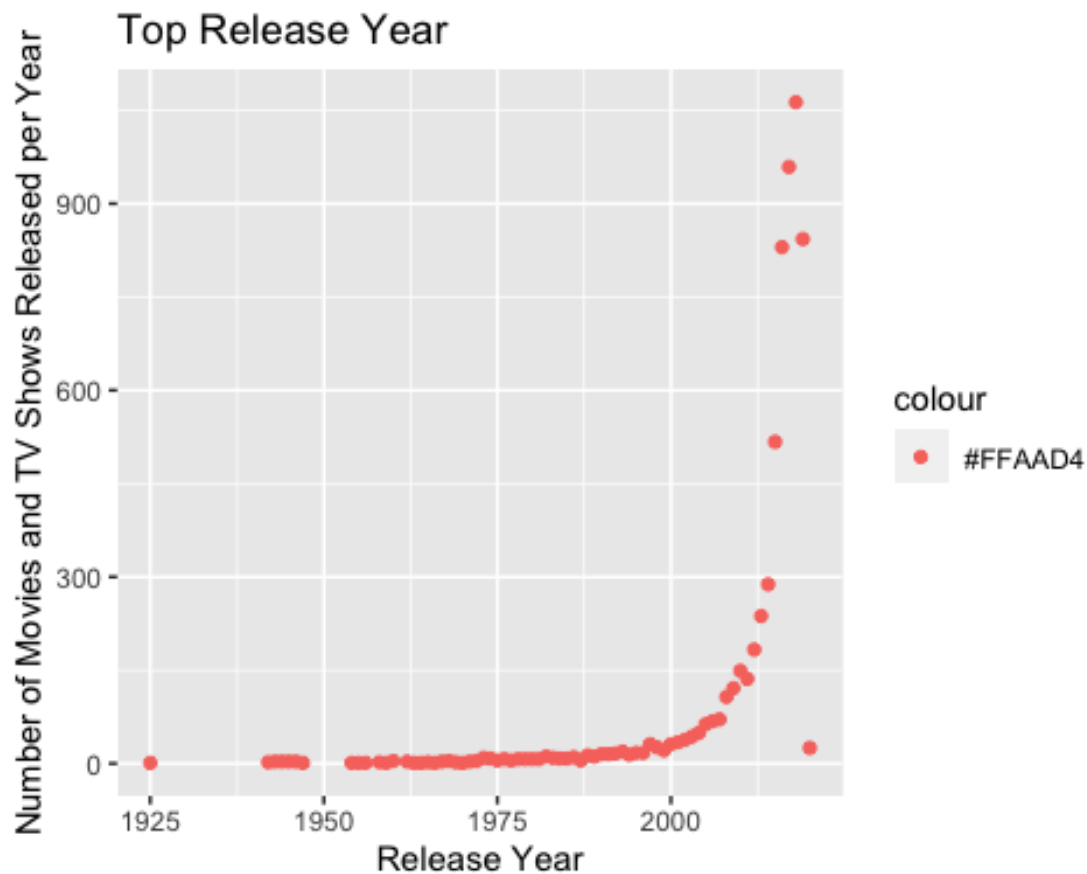
# view the number of shows released per date
head(count_by_date)

##   release_year n
## 1         1925 1
## 2         1942 2
## 3         1943 3
## 4         1944 3
## 5         1945 3
## 6         1946 3

# create the plot
plot2 <- ggplot(count_by_date, aes(x = release_year, y = n, backgroundColor="
white", color='#FFAAD4',
removePanelGrid=TRUE,removePanelBorder=TRUE))+
labs(title='Top Release Year',x='Release Year', y='Number of Movies and TV Sh
ows Released per Year')

# Display the scatterplot
```

```
plot2 +  
  geom_point()
```



Most Popular Genres

We find the most popular genres for both tv shows and movies below. This shows the international movies and international tv shows are the most popular genres found in the data set.

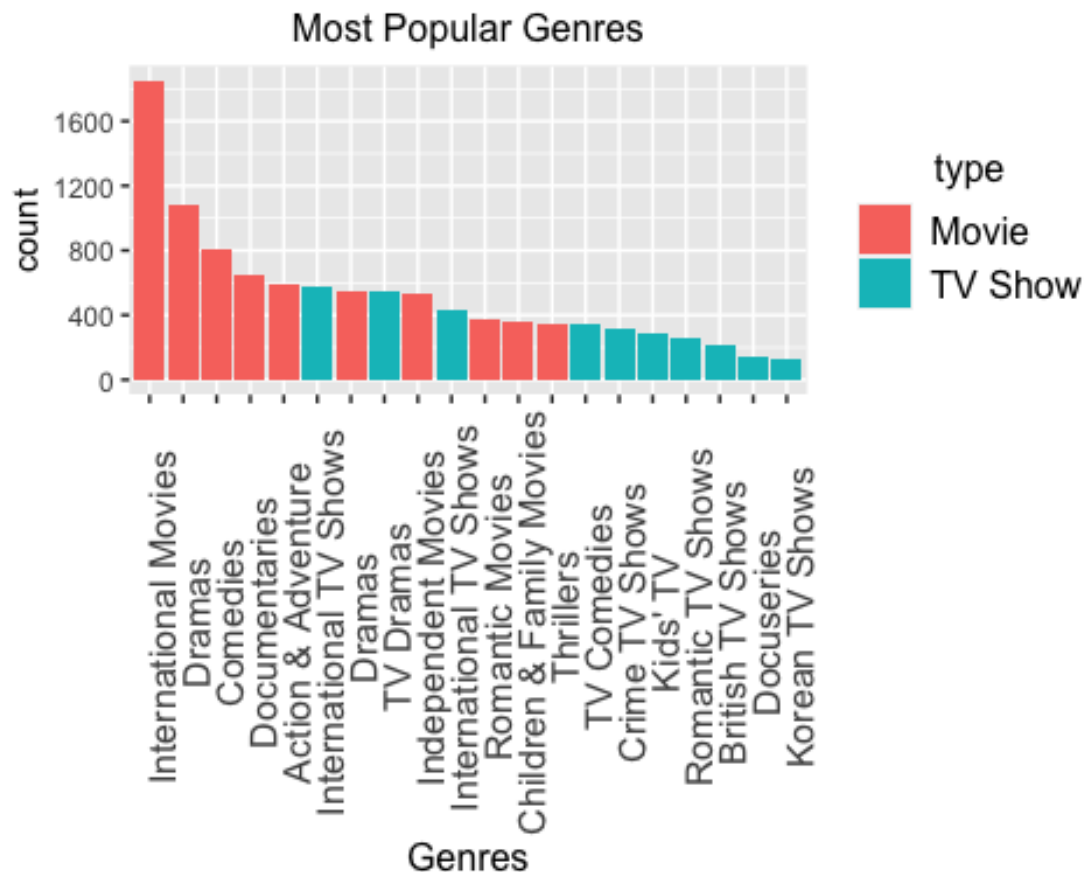
```
size <- function(width, height){ options(repr.plot.width = width, repr.plot.h  
eight = height)}
```

```
size(16,10)
```

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

```
## `summarise()` regrouping output by 'type' (override with `.groups` argumen  
t)
```

```
genres %>%
  ggplot(aes(x=fct_reorder(genre,count,.desc = T), y=count,fill=type))+
  geom_col()+scale_y_continuous(limits =c(0,1850),breaks =seq(0,1850,400))+
  labs(title='Most Popular Genres',x='Genres')+
  theme(axis.text.x = element_text(size = 12,angle =90),
        axis.title.x = element_text(hjust = 0.5,size = 12),
        legend.text = element_text(size = 12),
        legend.title=element_text(hjust = 0.5,size = 12),
        plot.title=element_text(hjust = 0.5,size = 12))
```



Top Directors

The data set can be arranged to show the most frequently found director on netflix by first selecting the show ids with the director. After counting and arranging them in descending order the top director found in the data set can be found as Jan Suter with over 20 counts.

```
#create a variable for the director per show
director <- netflix %>%
  select(c('show_id', 'director')) %>%
  gather(key = 'role', value = 'person', director) %>%
  filter(person != "") %>%
  separate_rows(person, sep = ',')
```

#view the first few lines

```
head(director)
```

```
## # A tibble: 6 x 3
##   show_id role      person
##   <int> <chr>    <chr>
## 1 81145628 director "Richard Finn"
## 2 81145628 director " Tim Maltby"
## 3 80125979 director "Fernando Lebrija"
## 4 70304989 director "Gabe Ibáñez"
## 5 80164077 director "Rodrigo Toro"
## 6 80164077 director " Francisco Schultz"
```

#count directors and arrange in descending order

```
count_director<- director %>%
  group_by(person,role) %>%
  summarise(count = n()) %>%
  arrange(desc(count))
```

```
## `summarise()` regrouping output by 'person' (override with `.groups` argument)
```

#create the plot to display top 10 directors

```
count_director %>%
  group_by(role) %>%
  top_n(10,count) %>%
  ungroup() %>%
```

```
ggplot(aes(x = fct_reorder(person,count,.desc = T), y = count, fill = role))
+
geom_bar(stat = 'identity') +
scale_x_discrete() +
facet_wrap(~role, scales = 'free_x') +
theme(legend.position = 'none') +
labs(x = 'Director', y='Count')
```

