Netflix Titles

```
library(tidyverse)
library(tidytuesdayR)
library(scales)
theme_set(theme_light())
```

Reading the data

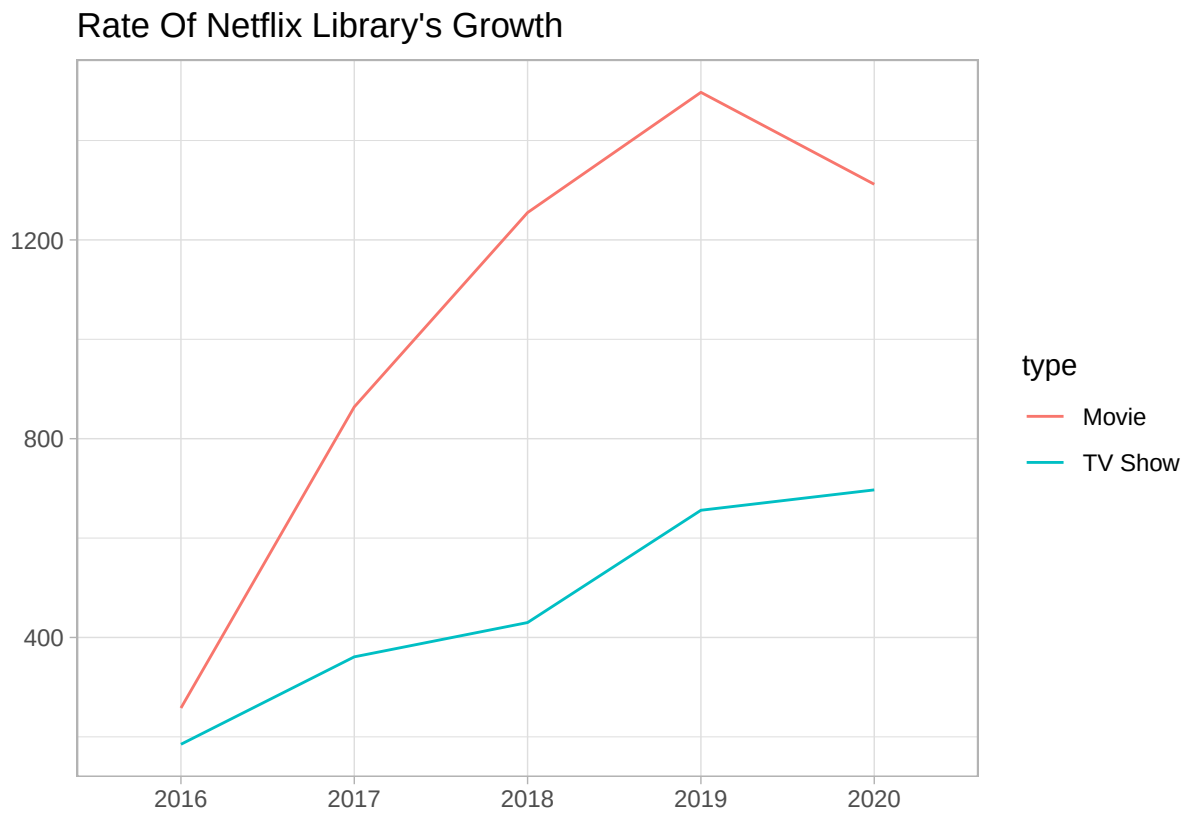
```
netflix_titles <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/
```

Processing The Data

```
# Extract exact date from date
titles_processed <- netflix_titles  %>%
  mutate(date_added_year = format(as.POSIXct(date_added, format="%b %d, %Y"), format="%Y"))  %>%
  separate(duration, c("duration", "unit"), sep = " ", convert = TRUE)
```

What is the growth of Netflix library through time?

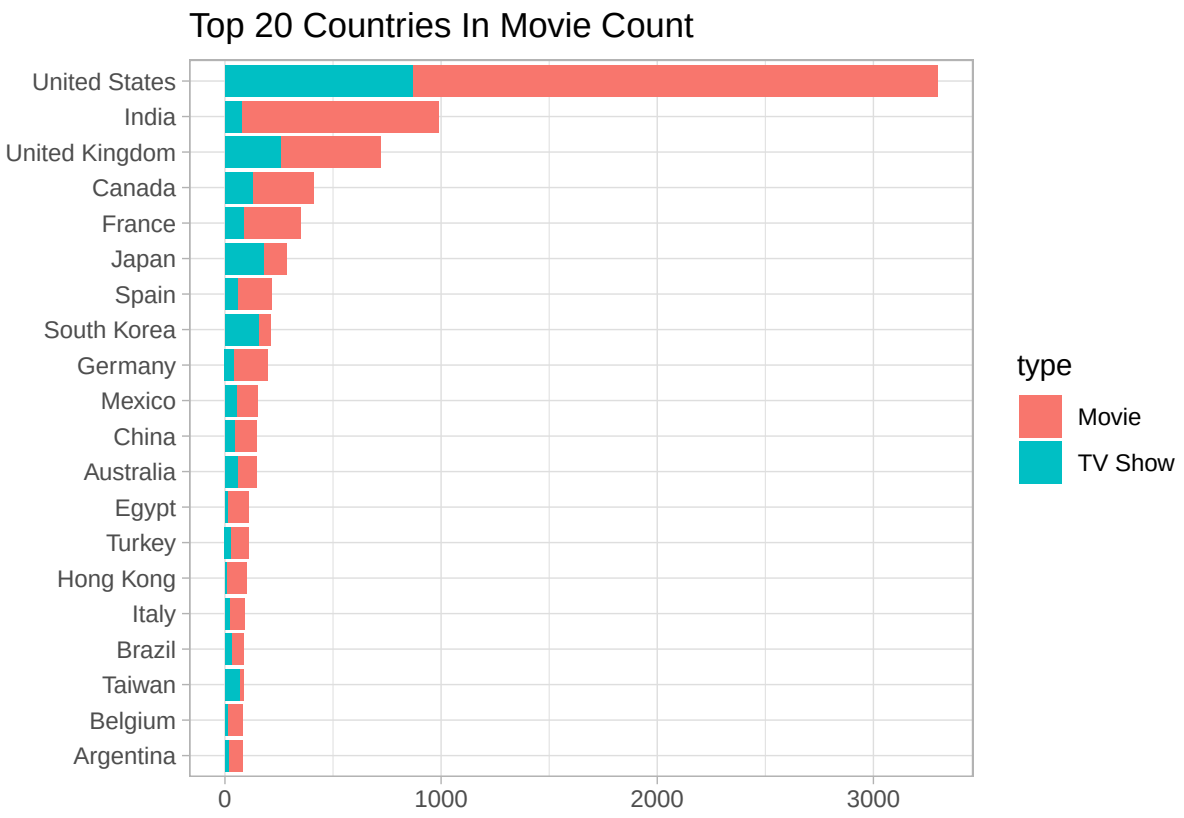
```
# We exclude 2021 as its incomplete
# And Years < 2015 as it's before Netflix actual kick off (insignificant)
titles_processed %>%
  count(date_added_year, type)  %>%
  filter(!is.na(date_added_year),
         date_added_year != 2021,
         date_added_year > 2015)  %>%
  ggplot(aes(x=date_added_year, y=n, color = type, group=type)) +
  geom_line() +
  labs(title=str_to_title("Rate of Netflix Library's Growth"),
       x = "",
       y = "")
```



So rate of adding both movies and TV shows declined at 2020, after some investigation and according to data from https://www.mediaplaynews.com/report-netflix-lost-30-percent-us-market-share-in-2020/ the number of Netflix subscribers fell from number of subscriptions to SVOD services in the US fell from 29% to 20%, hence, that explains why would they buy less shows

What countries contribute to most movies in Netflix library?

```
titles_processed %>%
  separate_rows(country, sep = ',') %>%
  mutate(
    country = str_trim(country),
    country = fct_lump(country, 20))  %>%
  filter(!is.na(country), country != "Other") %>%
  count(country, type, sort = TRUE)   %>%
   mutate(
    country = fct_reorder(country, n)) %>%
  ggplot(aes(x=country, y=n, fill = type, group=type)) +
  geom_col() +
  coord_flip() +
  labs(title=str_to_title("top 20 countries in movie count"),
       x = "",
       y = "")
```

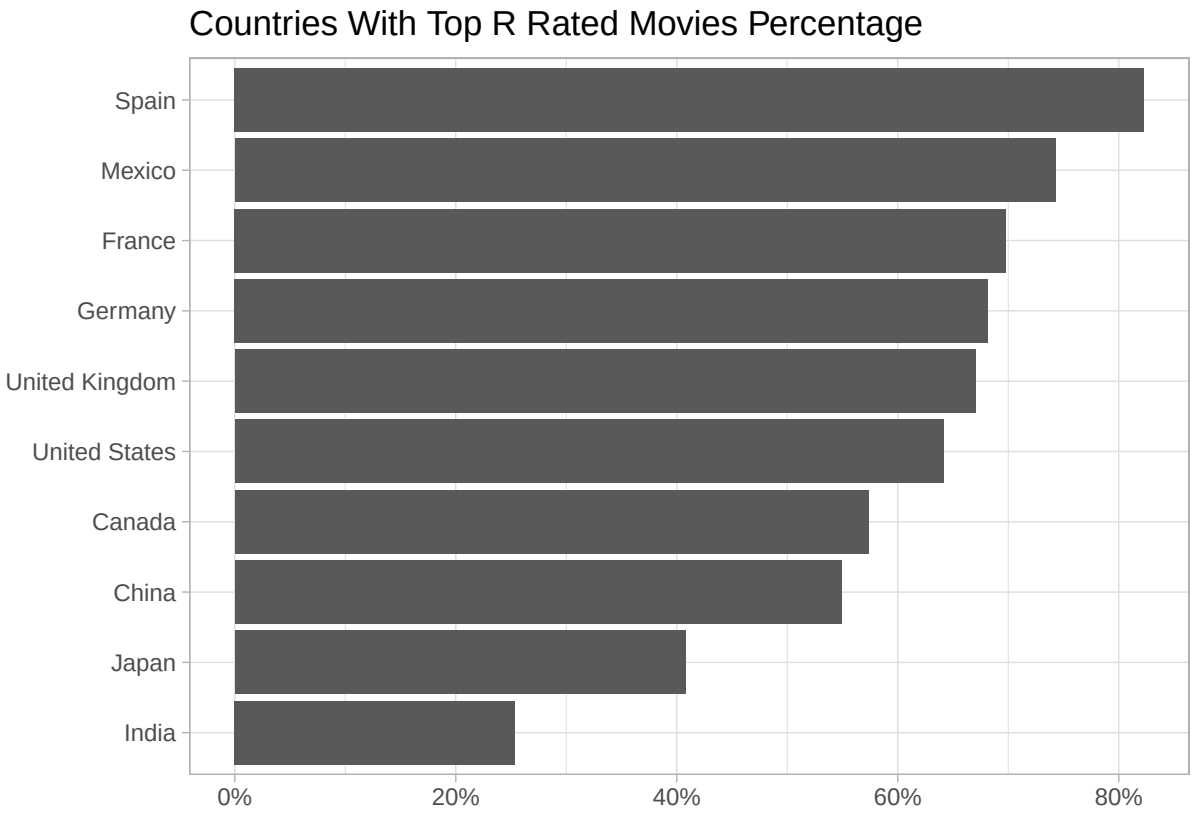


* Countries contribute mostly with movies such as USA and India, while other mostly TV shows such Japan and South Korea

Countries with highest R rated percentage?

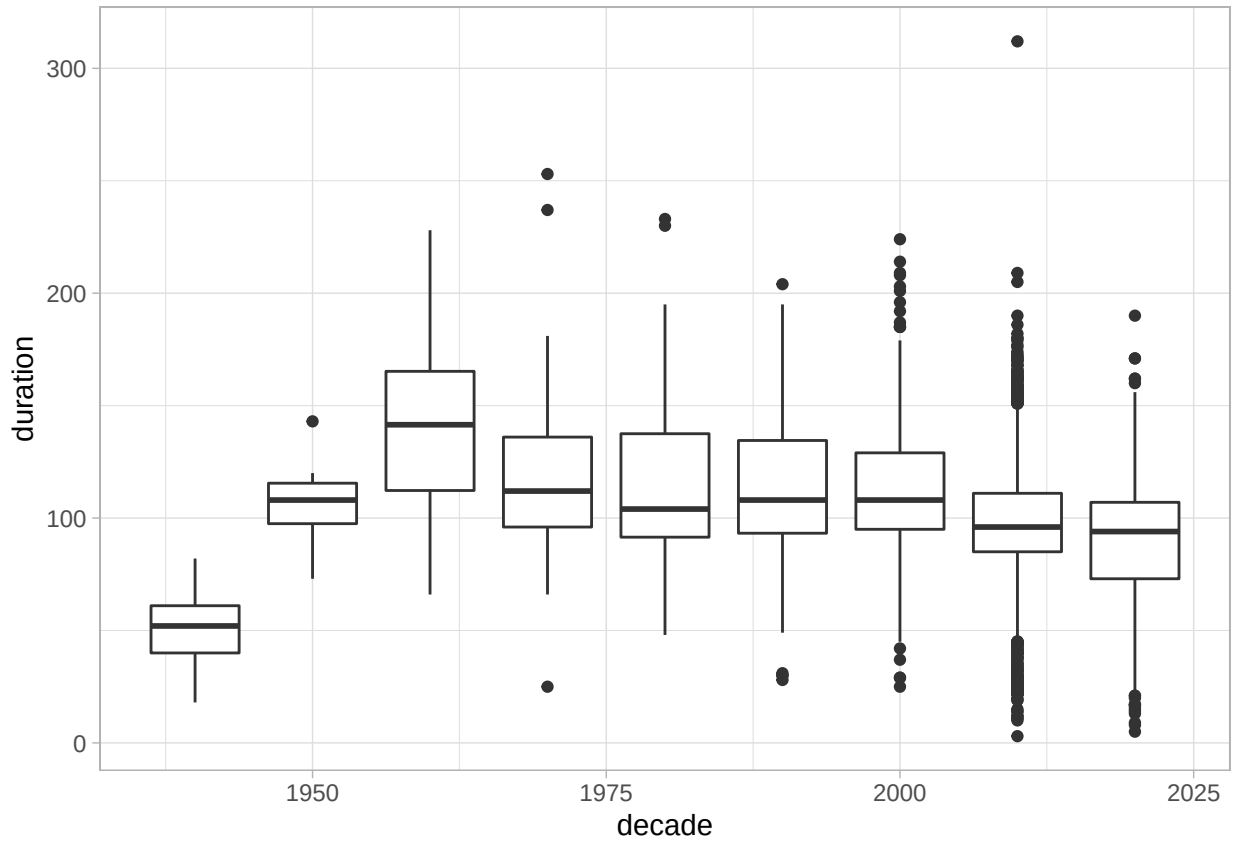
Since India and USA represents most of Netflix library (not a fair comparison) we'd try get countries with highest R rated percentage

```
titles_processed %>%
    filter(type == "Movie", !is.na(country))  %>%
    select(country, rating) %>%
    separate_rows(country, sep = ',') %>%
    mutate(country = str_trim(country)) %>%
    group_by(country) %>%
  # All adult ratings
    summarise(r_percentage = (sum(rating %in% c('PG-13','R','TV-MA', 'NC-17'))/n()), movies = n())  %>%
  mutate(
    country = fct_reorder(country, r_percentage)) %>%
  # We use minimum a 100 movies to get reliable results
    filter(movies > 100)  %>%
    ggplot(aes(x=r_percentage, y=country)) +
    geom_col() +
    scale_x_continuous(labels=label_percent()) +
  labs(title=str_to_title("Countries with top R rated movies percentage "),
       x = "",
       y = "")
```



Are older movies have longer duration, or vice versa?

```
titles_processed %>%
    filter(type == "Movie")  %>%
    mutate(decade = 10 * (release_year %/% 10))  %>%
    ggplot(aes(x=decade, y=duration, group = decade)) +
    geom_boxplot()
```



60s has the most longer movie duration, I did some investigation and found that maybe because scenes back then where a bit longer, as audience needed more elaboration to understand what's going on, unlike fast paced 2 3 minute scenes nowadays, hence, shorter movies towads todate.

How words in description relates to movies titles

Peering words with titles

```
library(tidytext)
library(widyr)
library(ggraph)

titles_processed %>%
    # Seperate words from description
    unnest_tokens(word, description)  %>%
    # Remove stopping words (a, the, to, and.. etc.)
    anti_join(stop_words, by = "word") %>%
    # Distinct for display
    distinct(type, title, word)  %>%
    add_count(word, name = "word_total") %>%
    # Min 20 words
    filter(word_total >= 20) %>%
    pairwise_cor(word, title, sort = TRUE) %>%
    filter(correlation >= 0.2)  %>%
    igraph::graph_from_data_frame()  %>%
    ggraph(layout = "fr") +
```

```
  geom_edge_link(aes(alpha =correlation)) +
  geom_node_point() +
  geom_node_text(aes(label=name),repel = TRUE)  +
  theme(legend.position = "none")
```