Exploring Netflix Movies Dataset with TidyVerse

Sabina Baraili

November 22, 2025

Contents

Introduction

The purpose of this vignette is to demonstrate how to use **TidyVerse** packages - mainly `dplyr`, `ggplot2`, and `tidyr` - for data cleaning, manipulation, and visualization using a real-world dataset.

We will explore the **Netflix Movies and TV Shows dataset** from Kaggle, which includes details such as title, type, country, release year, and genre.

Dataset link: Netflix Titles (Kaggle)

Load and Inspect the Dataset

```
# Load dataset (make sure netflix_titles.csv is saved in your working directory)
netflix <- read_csv("netflix_titles.csv")
```

```
## Rows: 8807 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (11): show_id, type, title, director, cast, country, date_added, rating,...
## 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.
```

```
# View structure and sample data
glimpse(netflix)
```

```
## Rows: 8,807
## Columns: 12
## $ show_id      <chr> "s1", "s2", "s3", "s4", "s5", "s6", "s7", "s8", "s9", "s1~
## $ type         <chr> "Movie", "TV Show", "TV Show", "TV Show", "TV Show", "TV ~
## $ title        <chr> "Dick Johnson Is Dead", "Blood & Water", "Ganglands", "Ja~
## $ director     <chr> "Kirsten Johnson", NA, "Julien Leclercq", NA, NA, "Mike F~
## $ cast         <chr> NA, "Ama Qamata, Khosi Ngema, Gail Mabalane, Thabang Mola~
## $ country      <chr> "United States", "South Africa", NA, NA, "India", NA, NA,~
## $ date_added   <chr> "September 25, 2021", "September 24, 2021", "September 24~
## $ release_year <dbl> 2020, 2021, 2021, 2021, 2021, 2021, 2021, 1993, 2021, 202~
## $ rating       <chr> "PG-13", "TV-MA", "TV-MA", "TV-MA", "TV-MA", "TV-MA", "PG~
## $ duration     <chr> "90 min", "2 Seasons", "1 Season", "1 Season", "2 Seasons~
## $ listed_in    <chr> "Documentaries", "International TV Shows, TV Dramas, TV M~
## $ description  <chr> "As her father nears the end of his life, filmmaker Kirst~
```

```
head(netflix)
```

```
## # A tibble: 6 x 12
##   show_id type    title    director cast  country date_added release_year rating
##   <chr>   <chr>   <chr>    <chr>    <chr> <chr>   <chr>             <dbl> <chr>
## 1 s1      Movie   Dick Jo~ Kirsten~ <NA>  United~ September~         2020 PG-13
## 2 s2      TV Show Blood &~ <NA>     Ama ~ South ~ September~         2021 TV-MA
## 3 s3      TV Show Ganglan~ Julien ~ Sami~ <NA>    September~         2021 TV-MA
## 4 s4      TV Show Jailbir~ <NA>     <NA>  <NA>    September~         2021 TV-MA
## 5 s5      TV Show Kota Fa~ <NA>     Mayu~ India   September~         2021 TV-MA
## 6 s6      TV Show Midnigh~ Mike Fl~ Kate~ <NA>    September~         2021 TV-MA
## # i 3 more variables: duration <chr>, listed_in <chr>, description <chr>
```

Clean and Prepare the Data

We'll filter only **Movies** and remove records with missing values in important columns like `release_year` and `country`.

```r
netflix_clean <- netflix %>%
  filter(type == "Movie") %>%
  drop_na(release_year, country)

# Quick check of the cleaned data
summary(netflix_clean)
```

```
##    show_id              type              title              director
##  Length:5691        Length:5691        Length:5691        Length:5691
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##      cast              country           date_added         release_year
##  Length:5691        Length:5691        Length:5691        Min.   :1942
##  Class :character   Class :character   Class :character   1st Qu.:2012
##  Mode  :character   Mode  :character   Mode  :character   Median :2016
##                                                           Mean   :2013
##                                                           3rd Qu.:2018
##                                                           Max.   :2021
##     rating             duration          listed_in          description
##  Length:5691        Length:5691        Length:5691        Length:5691
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
```

Top Movie-Producing Countries

Let's find the **top 10 countries** that have produced the most Netflix movies using `dplyr::count()`.

```r
top_countries <- netflix_clean %>%
  count(country, sort = TRUE) %>%
  head(10)

top_countries
```

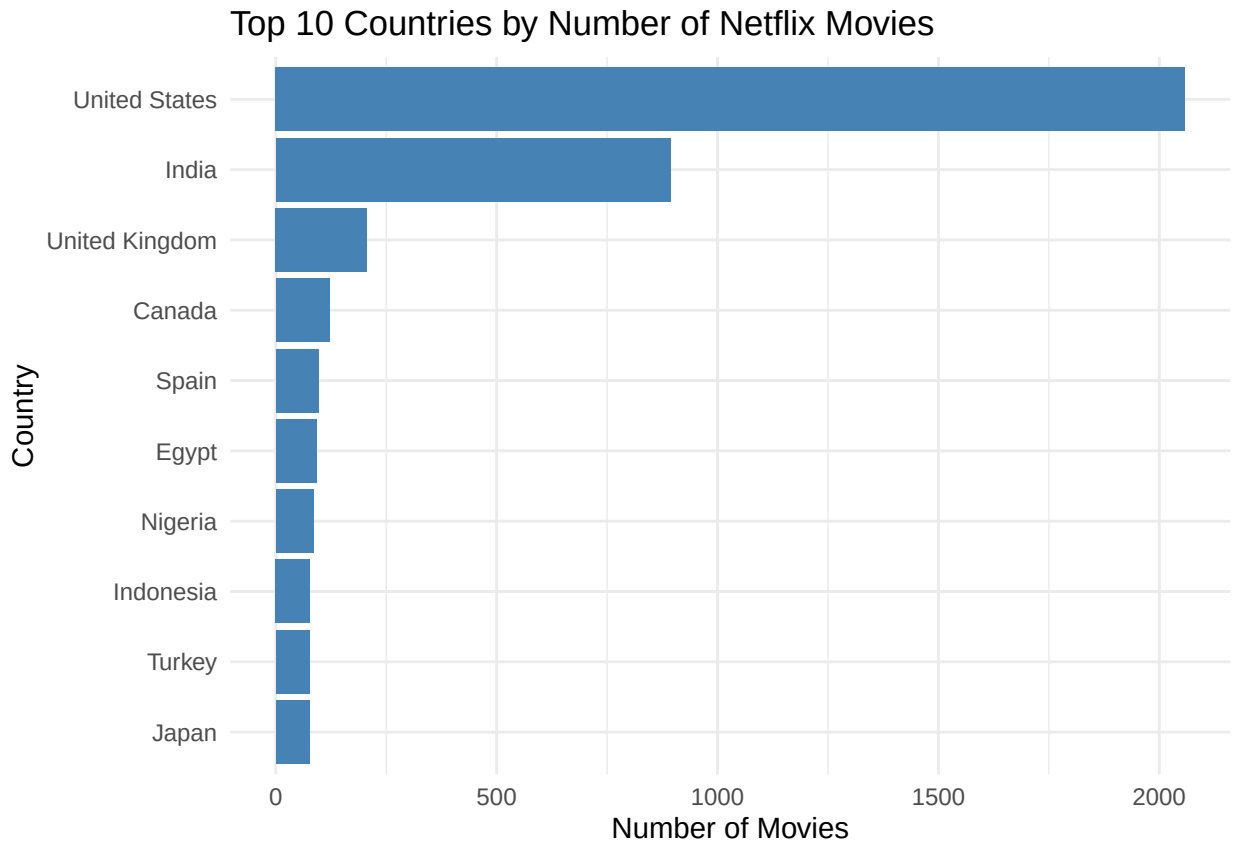
```
## # A tibble: 10 x 2
##    country            n
##    <chr>          <int>
##  1 United States   2058
##  2 India            893
##  3 United Kingdom   206
##  4 Canada           122
##  5 Spain             97
##  6 Egypt             92
##  7 Nigeria           86
```

```
##  8 Indonesia          77
##  9 Japan              76
## 10 Turkey             76
```

Visualization - Top 10 Countries by Movie Count

We'll visualize the results using a horizontal bar chart with `ggplot2`.

```
ggplot(top_countries, aes(x = reorder(country, n), y = n)) +
  geom_col(fill = "steelblue") +
  coord_flip() +
  labs(
    title = "Top 10 Countries by Number of Netflix Movies",
    x = "Country",
    y = "Number of Movies"
  ) +
  theme_minimal()
```

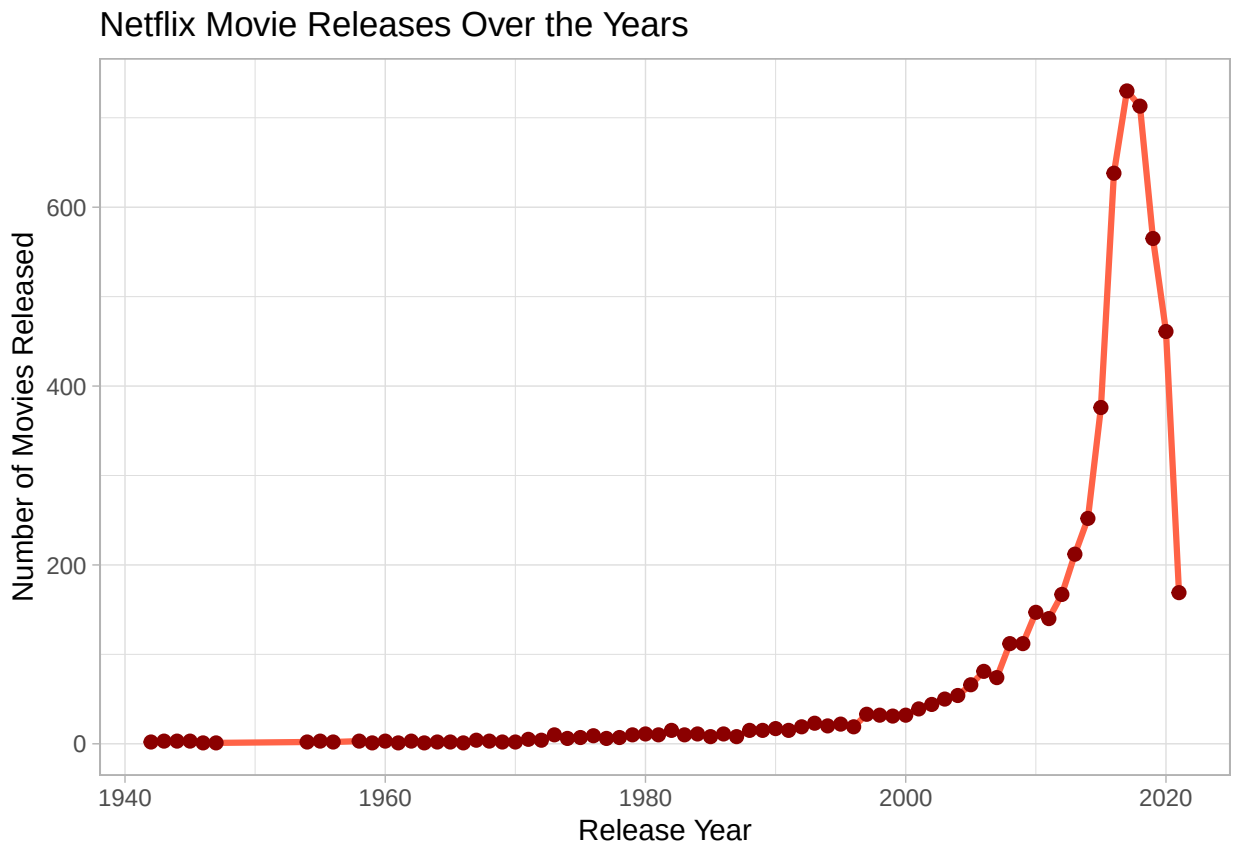


Movie Release Trend Over the Years

Here, we'll analyze how Netflix movie releases have changed over time.

```
movies_by_year <- netflix_clean %>%
  group_by(release_year) %>%
  summarise(total_movies = n())

ggplot(movies_by_year, aes(release_year, total_movies)) +
  geom_line(color = "tomato", linewidth = 1.1) +
  geom_point(color = "darkred", size = 2) +
  labs(
    title = "Netflix Movie Releases Over the Years",
    x = "Release Year",
    y = "Number of Movies Released"
  ) +
  theme_light()
```



Most Common Genres on Netflix

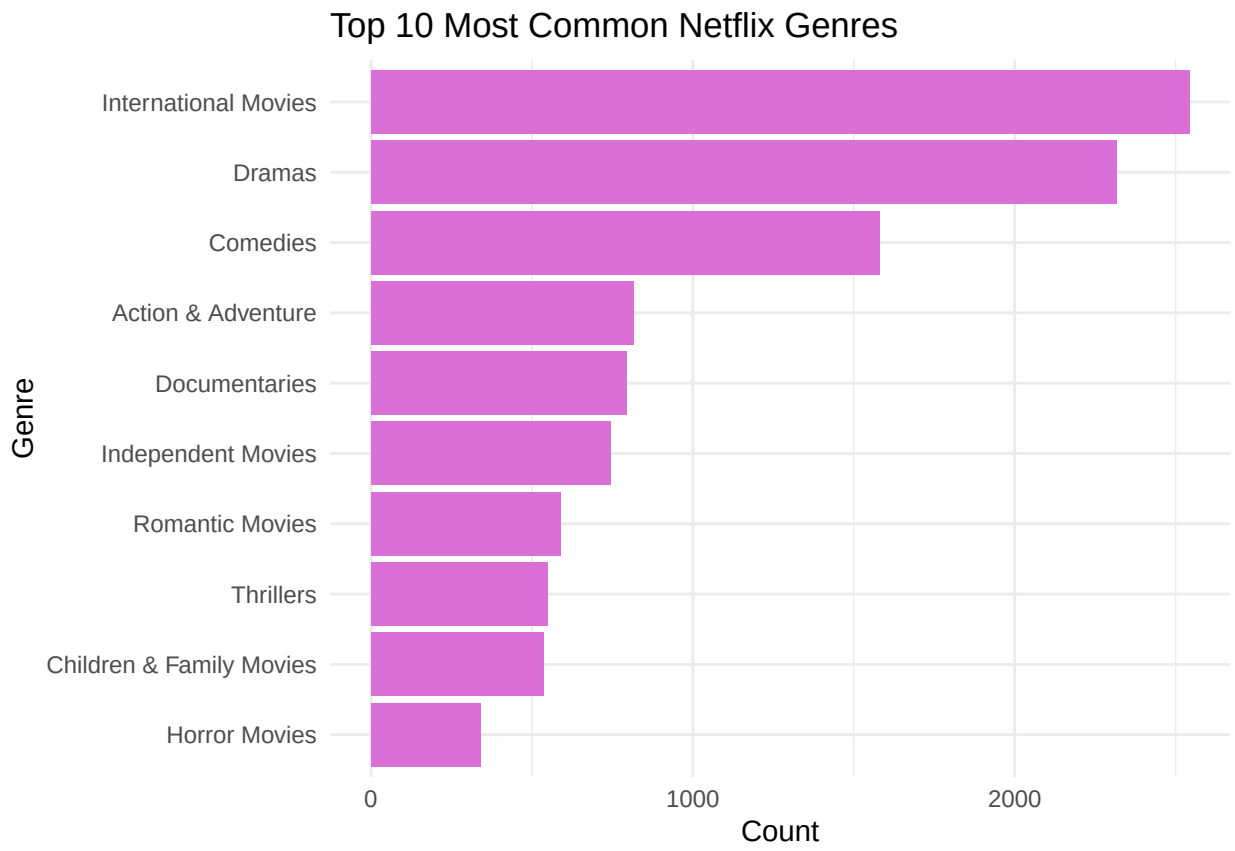
Each title can have multiple genres separated by commas.
We'll use `tidyr::separate_rows()` to split them and count the most frequent genres.

```r
genre_data <- netflix_clean %>%
  separate_rows(listed_in, sep = ", ") %>%
  count(listed_in, sort = TRUE) %>%
  head(10)

genre_data
```

```
## # A tibble: 10 x 2
##    listed_in                    n
##    <chr>                    <int>
##  1 International Movies      2543
##  2 Dramas                    2317
##  3 Comedies                  1580
##  4 Action & Adventure         817
##  5 Documentaries              794
##  6 Independent Movies         745
##  7 Romantic Movies            588
##  8 Thrillers                  549
##  9 Children & Family Movies   535
## 10 Horror Movies              340
```

```r
ggplot(genre_data, aes(x = reorder(listed_in, n), y = n)) +
  geom_col(fill = "orchid") +
  coord_flip() +
  labs(
    title = "Top 10 Most Common Netflix Genres",
    x = "Genre",
    y = "Count"
  ) +
  theme_minimal()
```



Jacob Shapiro Continuation

Movie Ratings with dplyr

The dplyr package has the select function, which isolates certain columns from a data frame. For example:

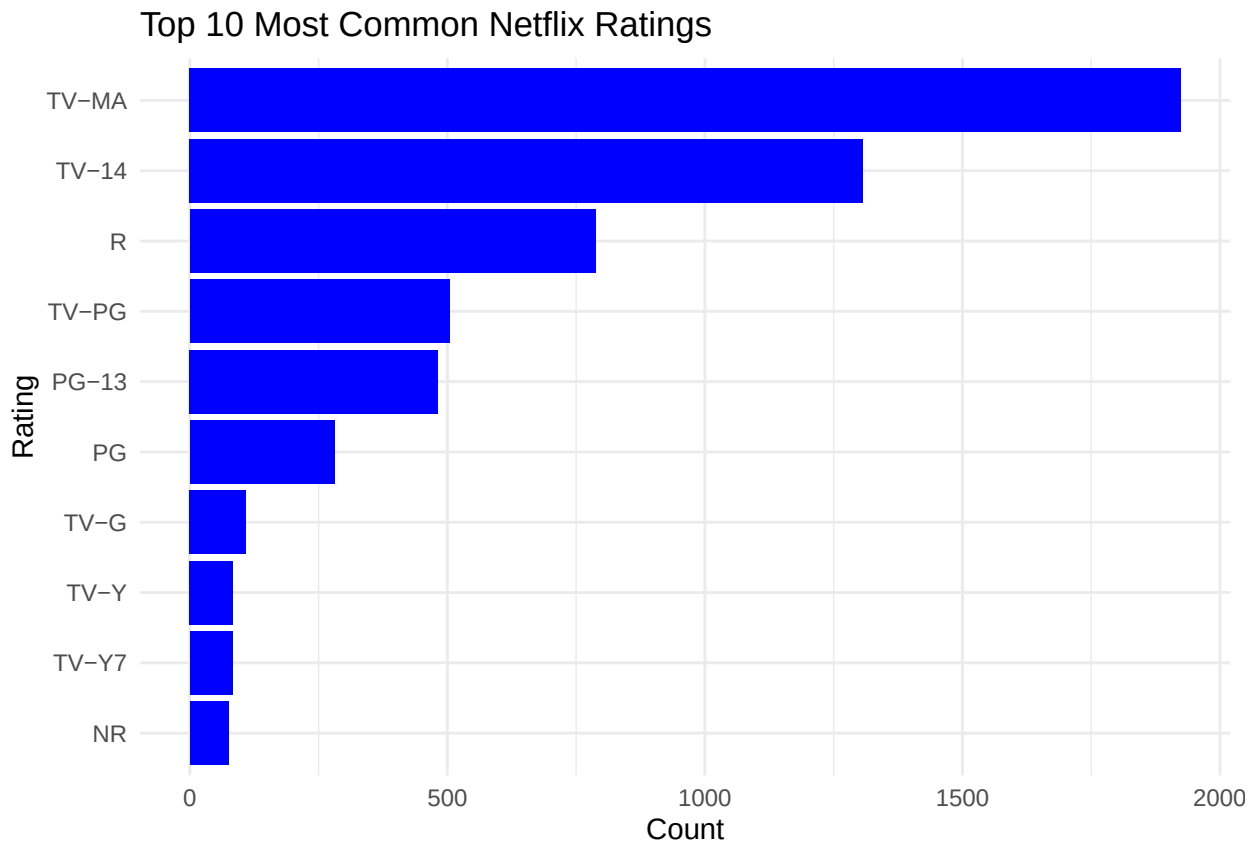
```
head(netflix_clean %>% select(rating))
```

```
## # A tibble: 6 x 1
##   rating
##   <chr>
## 1 PG-13
## 2 TV-MA
## 3 PG-13
## 4 TV-MA
## 5 TV-14
## 6 PG-13
```

We could look at the top ratings via ggplot

```r
rating_data <- netflix_clean %>%
  separate_rows(rating, sep = ", ") %>%
  count(rating, sort = TRUE) %>%
  head(10)

ggplot(rating_data, aes(x = reorder(rating, n), y = n)) +
  geom_col(fill = "blue") +
  coord_flip() +
  labs(
    title = "Top 10 Most Common Netflix Ratings",
    x = "Rating",
    y = "Count"
  ) +
  theme_minimal()
```



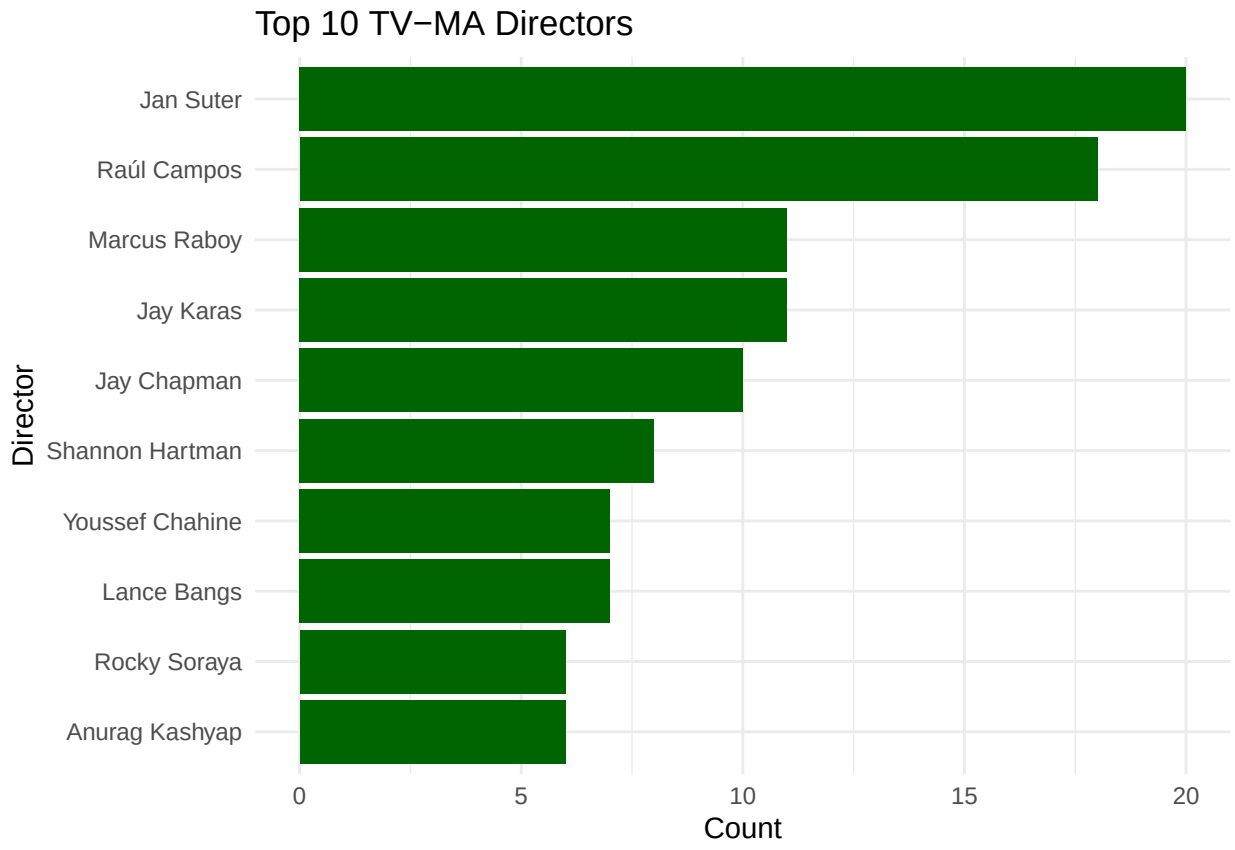
All items in netflix_clean are movies. A movie can have a "TV" rating if it's shown on TV or streaming service.

Top Directors of TV-MA

As TV-MA is the most used rating, we can use dplyr to find the director who's made the most TV-MA movies.

```r
tvma <- netflix_clean %>%
  filter(rating == "TV-MA") %>%
  drop_na(director) %>%
  separate_rows(director, sep = ", ") %>%
  select(director, rating) %>%
  count(director, sort = TRUE) %>%
  head(10)

ggplot(tvma, aes(x = reorder(director, n), y = n)) +
  geom_col(fill = "darkgreen") +
  coord_flip() +
  labs(
    title = "Top 10 TV-MA Directors",
    x = "Director",
    y = "Count"
  ) +
  theme_minimal()
```



Summary

This vignette demonstrates how to: - Use **dplyr** for data manipulation (`filter`, `count`, `group_by`, `summarise`, and `select`). - Use **tidyr** to reshape data with `separate_rows`, `drop_na`. - Use **ggplot2** to

visualize results in bar and line charts.

References

- Kaggle Netflix Dataset: https://www.kaggle.com/datasets/shivamb/netflix-shows
- TidyVerse Documentation: https://www.tidyverse.org/packages/