Project 2

2024-10-01

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

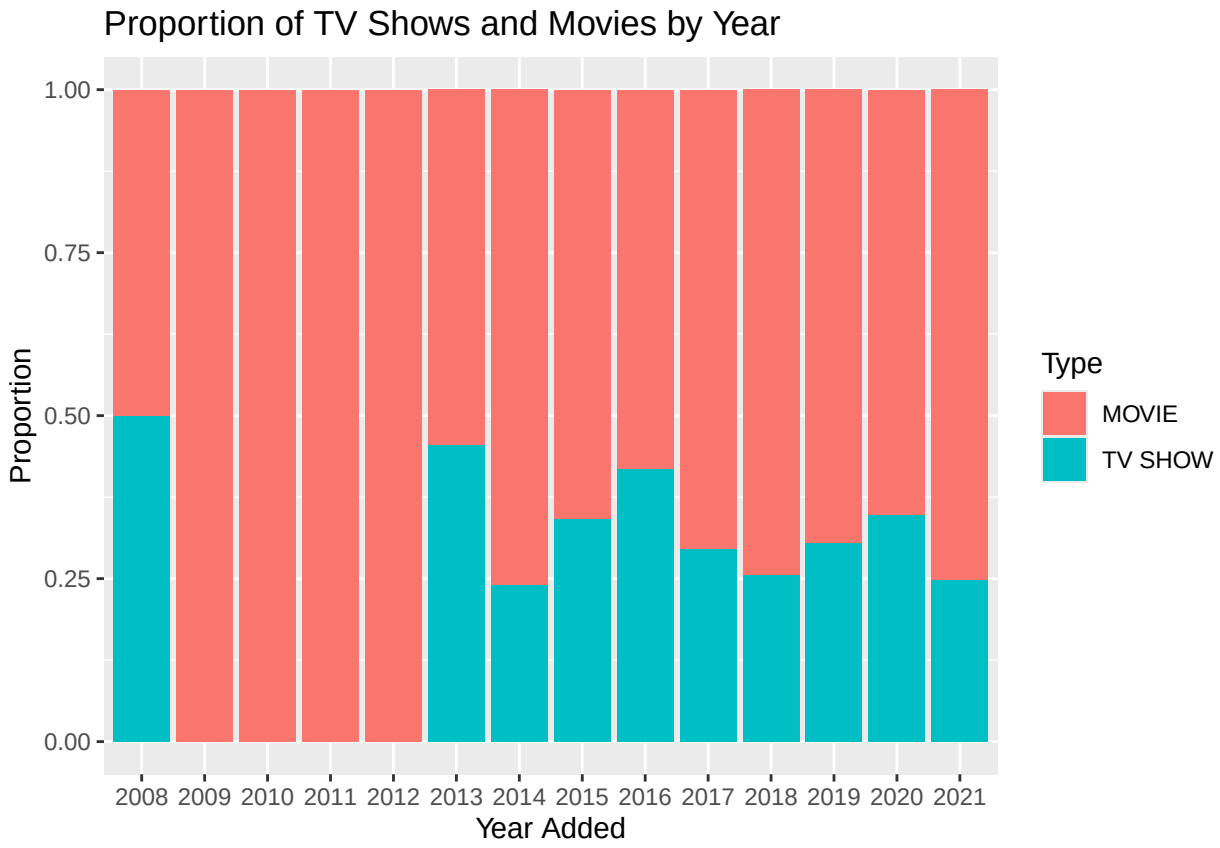
```
library(ggplot2)
netflix_titles <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/
```

```
## Rows: 7787 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.
```

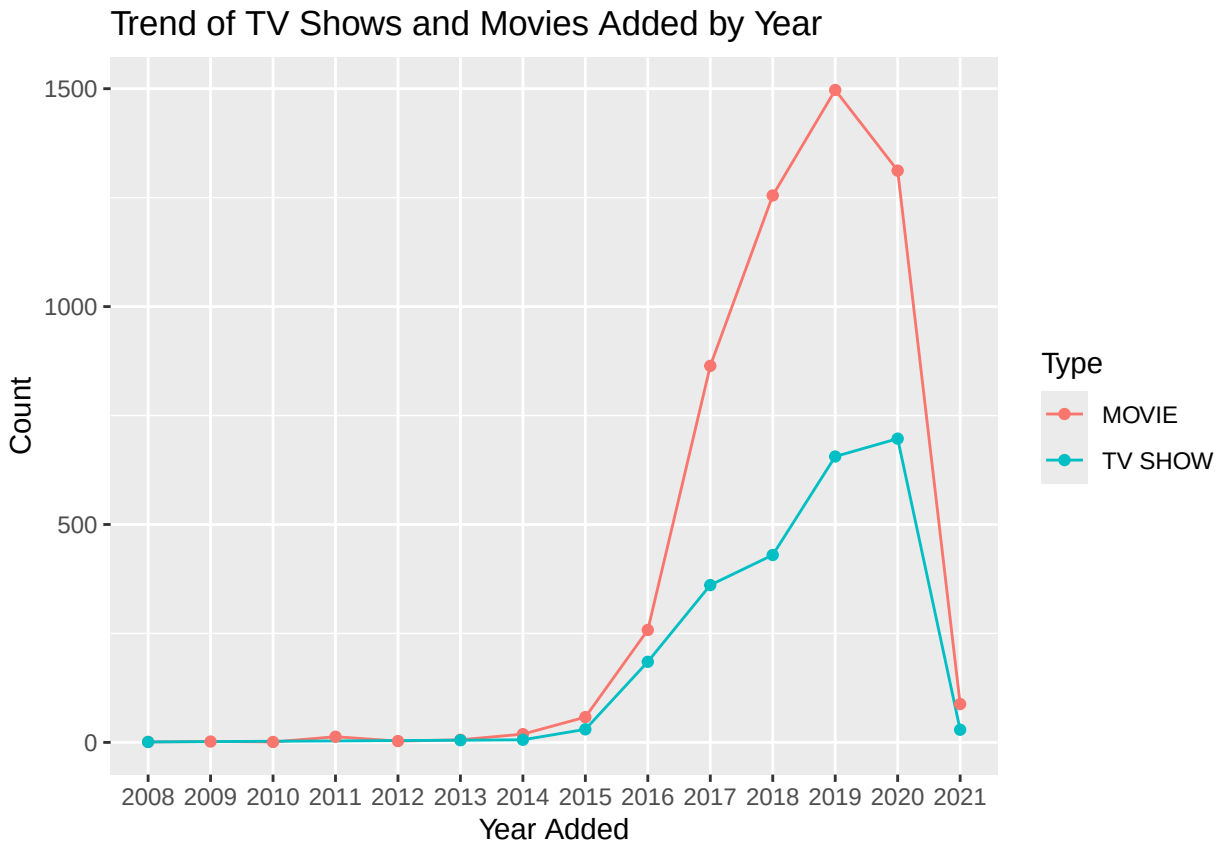
```
netflix_titles_updated <- netflix_titles |>
  mutate(year_added = str_extract(date_added, "\\d{4}"),
         year_added = str_replace_na(year_added, "Unknown"))|>
  mutate(type = str_to_upper(type))|>

  filter(year_added != "Unknown")


#view(netflix_titles_updated)


proportion_data <- netflix_titles_updated |>
  group_by(year_added, type) |>
  summarise(count = n()) |>
  ungroup() |>
  group_by(year_added) |>
  mutate(proportion = count / sum(count))
```

```
## `summarise()` has grouped output by 'year_added'. You can override using the
## `.groups` argument.
```

```
#proportion_data



proportion_plot_1 <- ggplot(proportion_data, aes(x = year_added, y = proportion, fill = type)) +
  geom_bar(stat = "identity", position = "stack") +
  labs(title = "Proportion of TV Shows and Movies by Year",
       x = "Year Added",
       y = "Proportion",
       fill = "Type")

proportion_plot_1
```



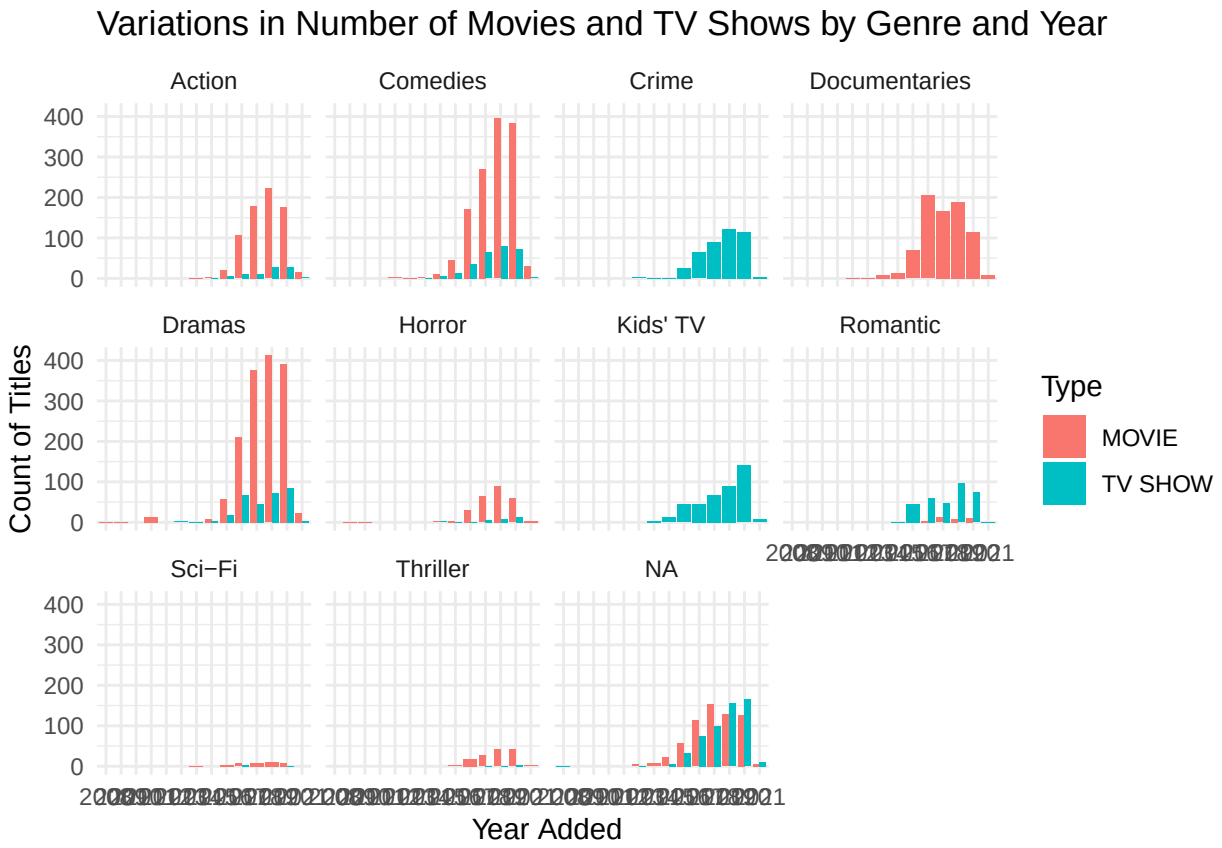

```
trend_plot <- netflix_titles_updated |>
  count(year_added, type) |>
  ggplot(aes(x = year_added, y = n, color = type, group = type)) +
  geom_line() +
  geom_point() +
  labs(title = "Trend of TV Shows and Movies Added by Year",
       x = "Year Added",
       y = "Count",
       color = "Type")
trend_plot
```




```
genres <- c("Thriller", "Action", "Dramas", "Romantic", "Horror", "Crime", "Sci-Fi", "Fantasy", "Comedi

genre<- netflix_titles_updated|>
  mutate(is_genre = str_detect(listed_in, str_c(genres, collapse = "|")))|>
  mutate(genre_label = str_extract(listed_in, str_c(genres, collapse = "|")))|>
  group_by(year_added, type, genre_label) |>
  summarise(count = n())
```

```
## `summarise()` has grouped output by 'year_added', 'type'. You can override
## using the `.groups` argument.
```

```
#genre


ggplot(genre, aes(x = year_added, y = count, fill = type)) +
  geom_bar(stat = "identity", position = "dodge") +
  facet_wrap(~ genre_label) +
  labs(title = "Variations in Number of Movies and TV Shows by Genre and Year",
       x = "Year Added",
       y = "Count of Titles",
       fill = "Type")+
       theme_minimal()
```



ANALYSIS

The first plot, "Proportion of TV Shows and Movies by Year", shows the proportion of TV shows and movies added to Netflix each year. This is a stacked bar chart, where each year is represented by a bar, with segments showing the proportion of TV shows and movies.In the graph, movies tended to dominate the viewing options expecially from 2009 to 2012. Although TV Shows in 2008 had the same as movies. In recent years however,the proportion of TV shows have increased with about 25 percent of viewing options being TV shows.

The second plot, "Trend of TV Shows and Movies Added by Year", shows the actual counts of TV shows and movies added per year. This is a line plot with different colors representing movies and TV shows. Between 2015 and 2020, they both experienced steep growths however movies added far outweighed TV Shows added

The third plot, "Variations in Number of Movies and TV Shows by Genre and Year", shows how different genres have varied over the years with respect to both TV Shows and Movies. In recent years, there's been growth in all the genres, however Comedies, Dramas, Documentary and Action movies rose more highly. Although Crimes, Kid's TV and Romantic movies have all risen in the TV Shows category, they are still below that of movies.

Reference:

https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2021/2021-04-20/netflix_titles.csv

https://github.com/rfordatascience/tidytuesday/blob/master/data/2021/2021-04-20/readme.md