Project 2

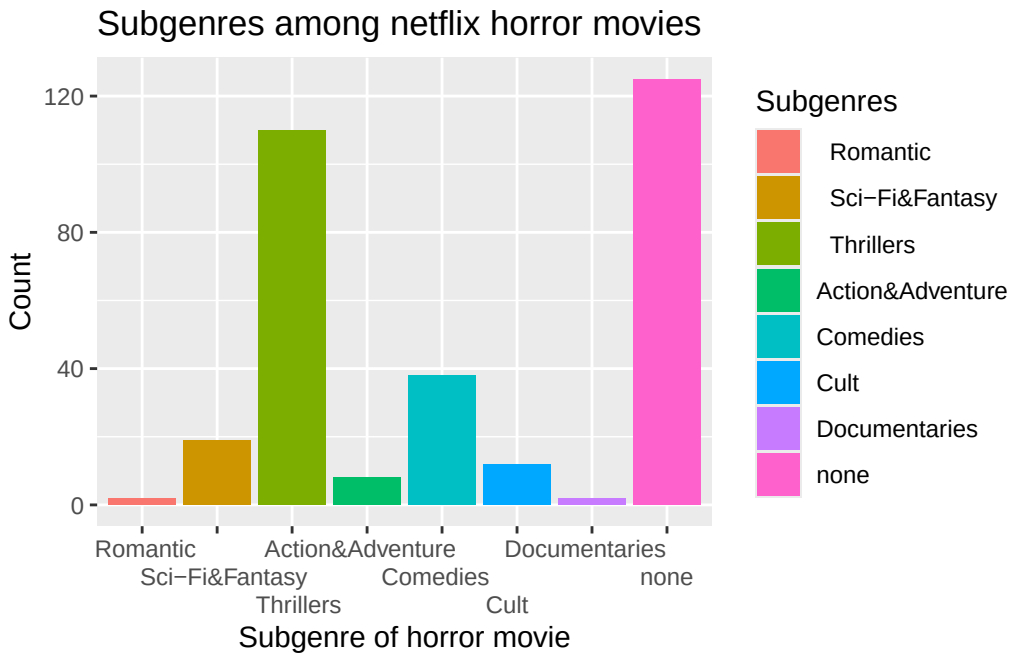
```
library(RTextTools)
library(tidyverse)
library(tidytuesdayR)
```

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

Netflix Horror Movies

```
horror_subgenre <- netflix %>%
  filter(str_detect(listed_in, "Horror"), type == "Movie") %>%
  mutate(listed_in = str_remove(listed_in, ", International Movies")) %>%
  mutate(listed_in = str_remove(listed_in, ", Independent Movies")) %>%
  mutate(listed_in = str_remove_all(listed_in, "[,]")) %>%
  mutate(listed_in = str_remove_all(listed_in, "Movies")) %>%
  mutate(listed_in = str_replace(listed_in, "[:space:]&[:space:]", "&")) %>%
  group_by(listed_in) %>%
  summarize(num = n()) %>%
  arrange(desc(num)) %>%
  mutate(listed_in = str_remove(listed_in, "Horror"))

horror_df <- horror_subgenre[c(1:5,7,12, 13),]
add <- horror_subgenre[-c(1:5,7,12, 13),]

for (i in 1:nrow(add)){
  if(str_detect(add[i,1], "Thrillers") == TRUE){
    horror_df[2,2] = horror_df[2,2] + add[i,2]
  }
  if(str_detect(add[i,1], "Comedies") == TRUE){
    horror_df[3,2] = horror_df[3,2] + add[i,2]
```

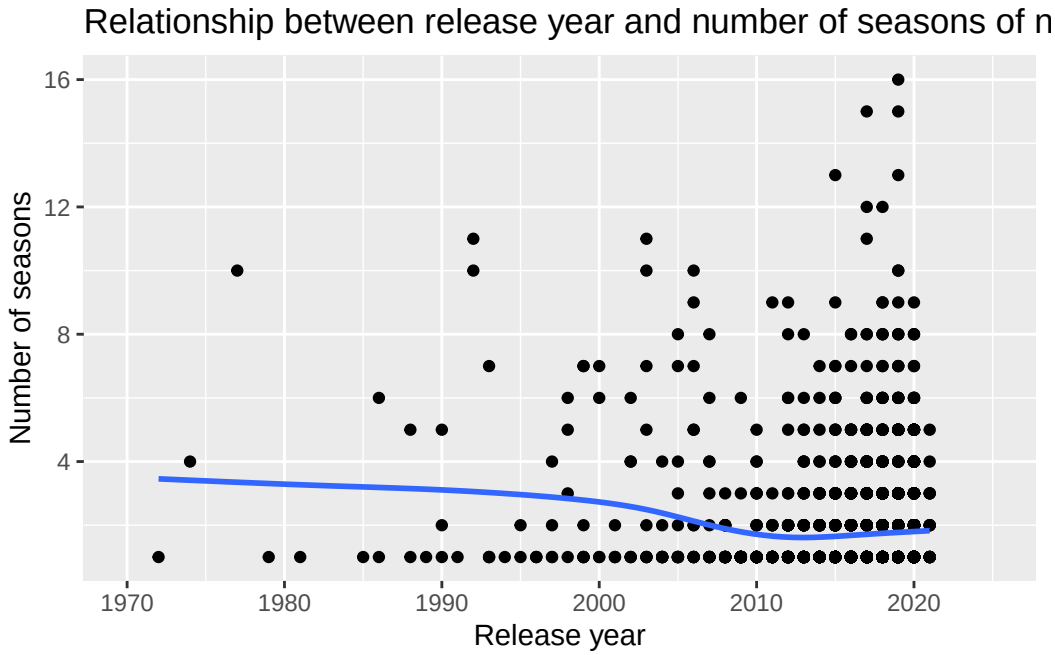
```
  }
  if(str_detect(add[i,1], "Action/Adventure") == TRUE){
    horror_df[4,2] = horror_df[4,2] + add[i,2]
  }
  if(str_detect(add[i,1], "Sci-Fi&Fantasy") == TRUE){
    horror_df[5,2] = horror_df[5,2] + add[i,2]
  }
  if(str_detect(add[i,1], "Cult") == TRUE){
    horror_df[6,2] = horror_df[6,2] + add[i,2]
  }
  if(str_detect(add[i,1], "Documentaries") == TRUE){
    horror_df[7,2] =  horror_df[7,2] + add[i,2]
  }
  if(str_detect(add[i,1], "Romantic") == TRUE){
    horror_df[8,2] = horror_df[8,2] + add[i,2]
  }
}

horror_df[1,1] = "none"

ggplot(horror_df, aes(x = listed_in, y = num)) +
  geom_bar(stat = "identity", aes(fill = listed_in)) + labs(
    title = "Subgenres among netflix horror movies",
    x = "Subgenre of horror movie",
    y = "Count"
  ) +
  scale_x_discrete(guide = guide_axis(n.dodge=3)) +
  guides(fill=guide_legend(title="Subgenres"))
```



Based on this plot of the number of horror movies with specific subgenres, it seems most netflix horror movies either do not have a subgenre or have thriller as their subgenre. Other subgenres appear substantially less present among Netflix horror movies.

Netflix TV Shows

```
data <- netflix %>%
  filter(type == "TV Show") %>%
  mutate(num_seasons = as.numeric(str_extract(duration, "\\d+" )))

ggplot(data, aes(x = release_year, y = num_seasons)) +
  geom_point() +
  xlim(1970,2025)+
  geom_smooth(se = FALSE) +
  labs(
    title = "Relationship between release year and number of seasons of netflix tv show",
    x = "Release year",
    y = "Number of seasons"
  )
```



Based on this plot, the number of seasons that a Netflix show has tends to be lower for shows released between 2005 and 2020 than for other shows. Due to the discreteness of values on the x-axis, multiple points are most likely overlapping. This explains why it appears the right side of the graph has less points below the line of best fit than above it even with the line trending downwards.

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