ML_B_ex1 - Netflix Dataset

Gil Diamant 314978412, Yishay Shlezinger 208438119 & Itamar Twersky 311587489

08 4 2022

setup & configure of the plots (each plot present in different figure)

load all relevant libraries for using on EDA:

```
library(ggplot2) # load the ggplot2 library
library(dplyr)
library(tidyverse)
library(scales)
```

import the dataset to a local variable called surveys:

```
surveys <- read_csv("netflix-rotten-tomatoes-metacritic-imdb.csv")
```

Firstly, lets observe the properties of the data for getting ideas for EDA -

```
dim(surveys)
```

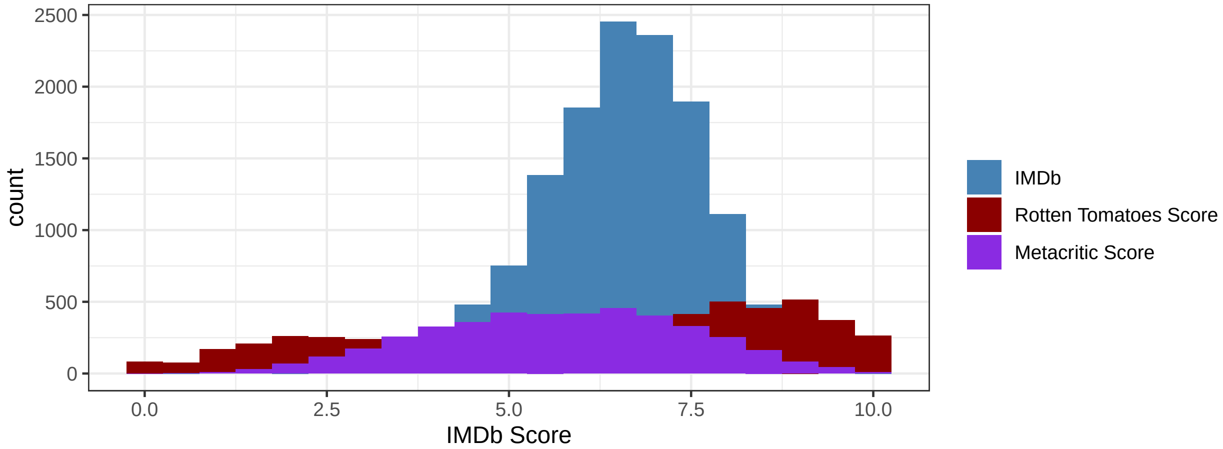
```
## [1] 15480    29
```

```
names(surveys)
```

```
##  [1] "Title"                 "Genre"                 "Tags"
##  [4] "Languages"             "Series or Movie"       "Hidden Gem Score"
##  [7] "Country Availability"  "Runtime"               "Director"
## [10] "Writer"                "Actors"                "View Rating"
## [13] "IMDb Score"            "Rotten Tomatoes Score" "Metacritic Score"
## [16] "Awards Received"       "Awards Nominated For"  "Boxoffice"
## [19] "Release Date"          "Netflix Release Date"  "Production House"
## [22] "Netflix Link"          "IMDb Link"             "Summary"
## [25] "IMDb Votes"            "Image"                 "Poster"
## [28] "TMDb Trailer"          "Trailer Site"
```

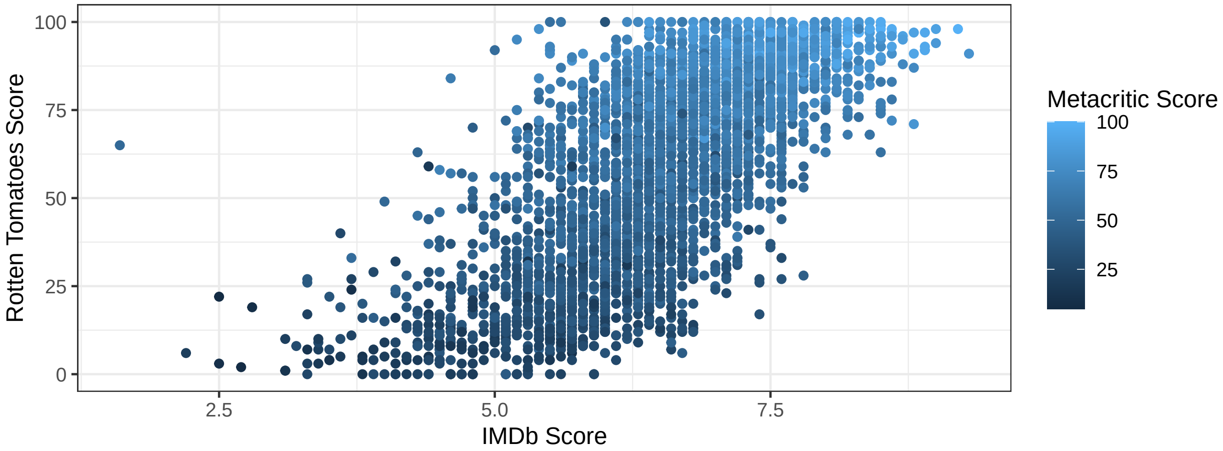
In the data, there are movie scores from three different sites - IMDb, Rotten Tomatoes and Metacritic, we were intersted in the difference of the ditribution of the 3 scores. we moved all the 3 scores to be in scale of 0-10, for the clarity of the comparsion. The results shows a significant different in the distribution of the 3 scores - while IMDb score is much centerd around ~6.5, the Rotten-Tomatoes score is widely distributed along all the score-range and the Metacritic-score have a distribution with propertis in the middle of the previous two.

```
ggplot(data = surveys) +
  geom_histogram(mapping = aes(x = `IMDb Score`,
                               fill="IMDb"),
                 binwidth = 0.5) +
  geom_histogram(mapping = aes(x = surveys$`Rotten Tomatoes Score`/10,
                               fill="Rotten Tomatoes Score"),
binwidth = 0.5) +
  geom_histogram(mapping = aes(x = surveys$`Metacritic Score`/10,

                 binwidth = 0.5) + scale_fill_manual(name = "",
                  values = c("IMDb" = "steelblue",
           "Rotten Tomatoes Score" = "darkred",
           "Metacritic Score" = "blueviolet")) +
  theme_bw()
```



Following the distributions difference results, it was interesting to check if there is correlation in the different scores' values or they are incompatible in the score values which means the scores are totaly different. If it is, its can be bad news for the who that want to choose good movie/series to watch, but wont be able to trust score if everyone says different. But, the results below shows that there is a corraltion between the 3 scores, you can see that we have nice diagonal and at most of the movies the 3 scores are agreed.

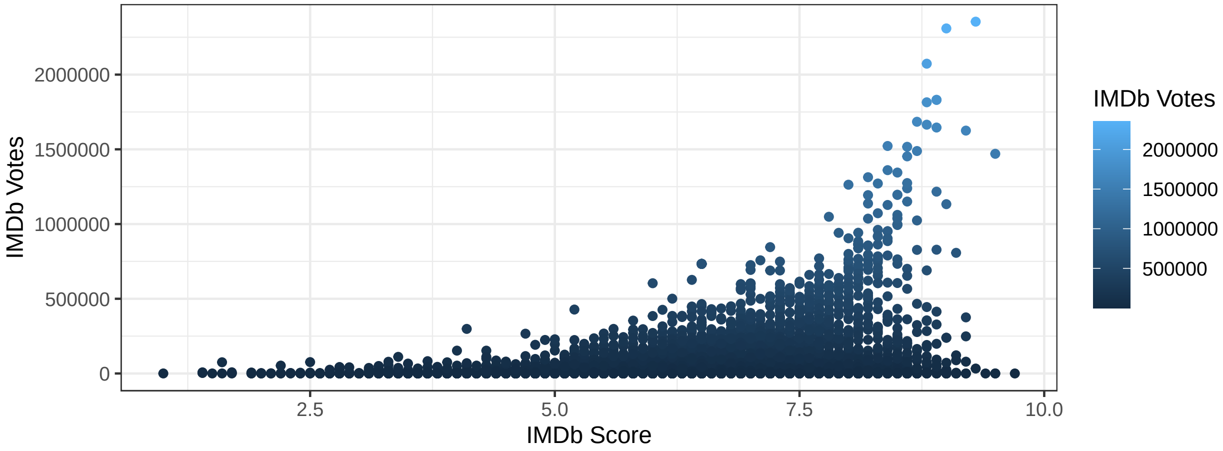
```
surveys_noNA_Meta<-subset(surveys, !is.na(`Metacritic Score`))
ggplot(data = surveys_noNA_Meta) +
  geom_point(mapping = aes(x = `IMDb Score`,
                           y = `Rotten Tomatoes Score`, color=`Metacritic Score`)) +
  theme_bw()
```



for the next parts, we continued with the IMDb score.

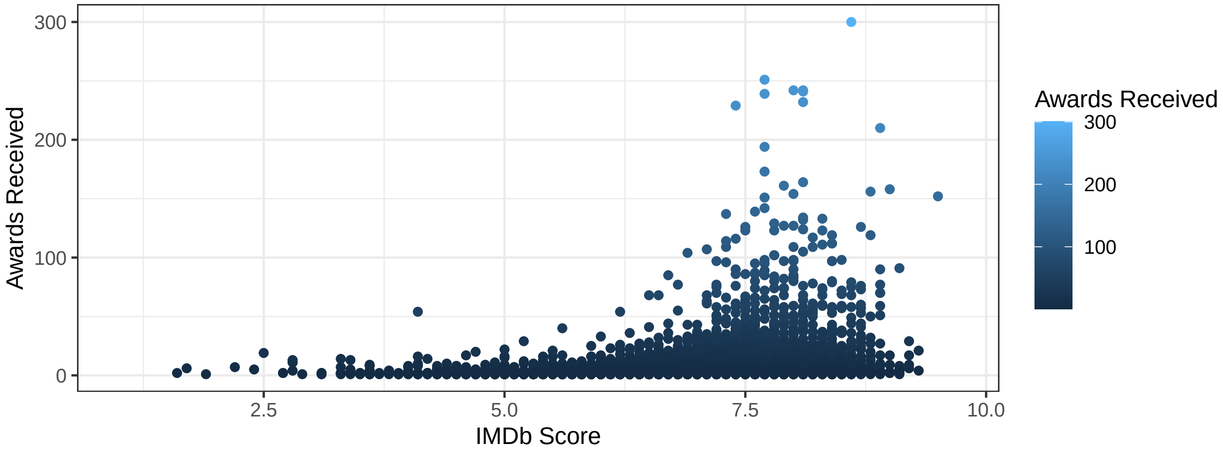
We were also interested whether high score movies are tend to be more popular and therefore will have more votes

```
ggplot(data = surveys) +
  geom_point(mapping = aes(x = `IMDb Score`, y = `IMDb Votes`, color= `IMDb Votes`)) +
  theme_bw()
```



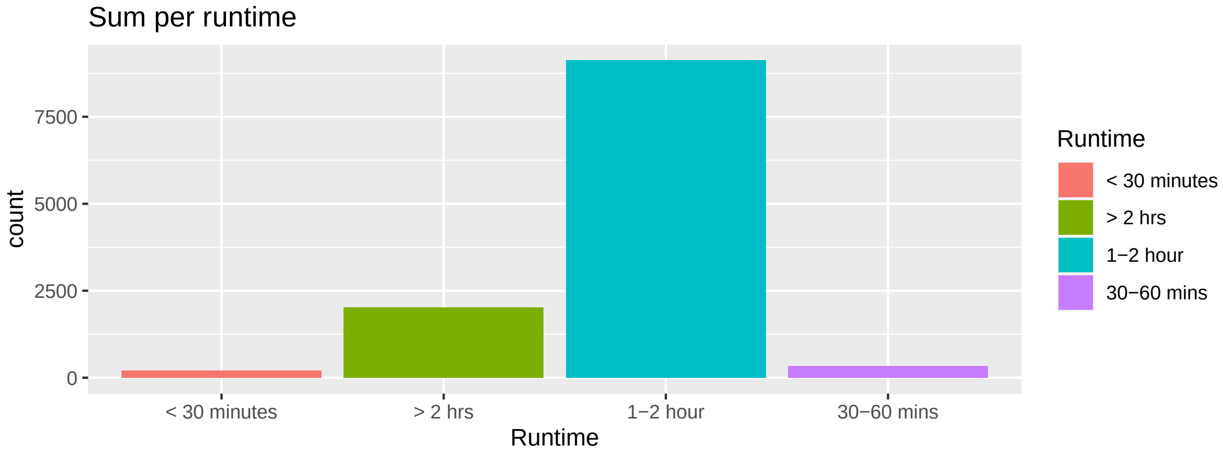
Last thing to look at about the score is the question - if an high score cause more Awards? We saw that the answer is yes, but not completely. Also movies/series which have <7 score almost have no rewards, the movies/series with the high scores have won much more rewards per movies/series

```
ggplot(data = surveys) +
  geom_point(mapping = aes(x = `IMDb Score`, y = `Awards Received`, color= `Awards Received`)) +
  theme_bw()
```



We was interest to check the distribution of length of movies

```
surveys_movies <- surveys %>% filter(`Series or Movie` == 'Movie')
ggplot(data = surveys_movies , color = Runtime) +
  geom_bar(mapping = aes(x = Runtime, fill = Runtime))+
  labs( title = "Sum per runtime")
```

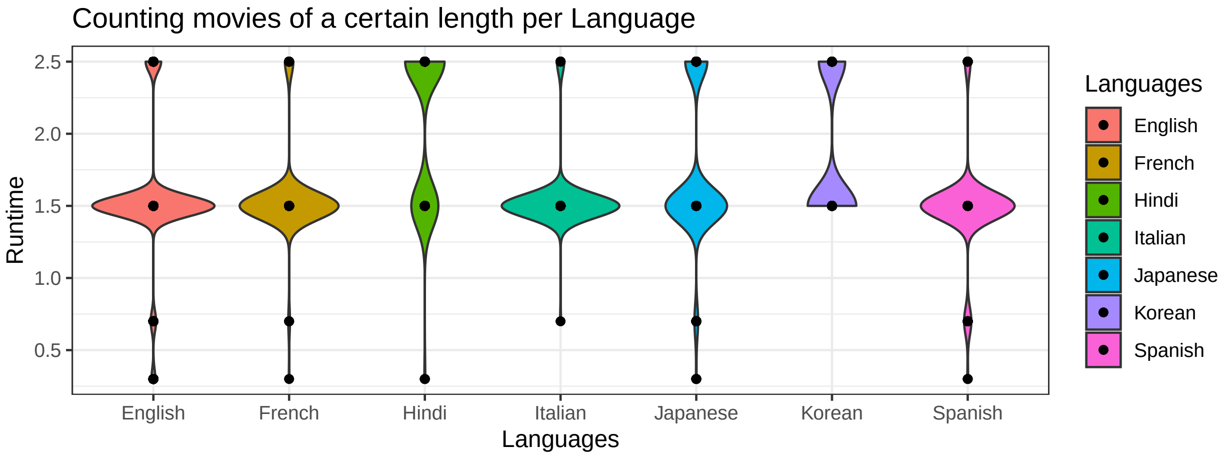


We assume that movies with a certain language are have properties that represent the culture the language belong to. One property that can be related to a culture is the movie length. check the correlation between the movie's language and the movie's time length. we looked for the top 5 popular languages, and check for them. From results we conclude interesting findings: - the movies in languages of countries with more western culture (English,French,Spanish,Italian) tend to be with length of 1-2 hours while movies with other languages are have also significant piece of longer(>2) movies. that may point out that the western culture are tend to be less patient to see a long movie

```
surveys_movies %>% group_by(Languages) %>%  tally(sort = TRUE) %>% head(10)
```

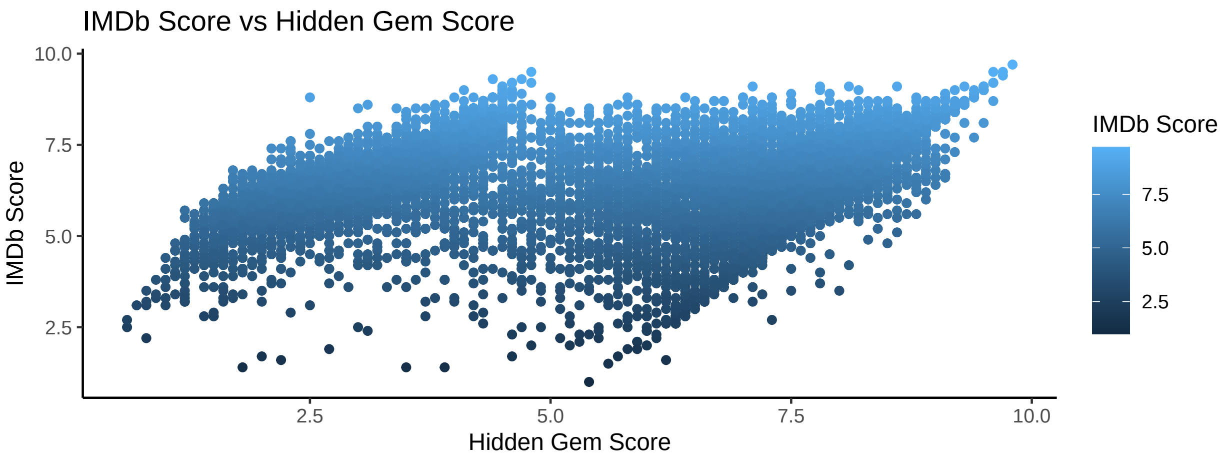
```
## # A tibble: 10 x 2
##    Languages            n
##    <chr>            <int>
##  1 English           3799
##  2 <NA>              1281
##  3 Japanese           768
##  4 Korean             296
##  5 Hindi              294
##  6 Spanish            267
##  7 English, Spanish   246
##  8 French             196
##  9 English, French    141
## 10 Italian            139
```

```
surveys_movies_fixed_runtime <- surveys_movies

surveys_movies_fixed_runtime$Runtime[surveys_movies_fixed_runtime$Runtime == "1-2 hour"] <- "1.5"
surveys_movies_fixed_runtime$Runtime[surveys_movies_fixed_runtime$Runtime == "< 30 minutes"] <- "0.3"
surveys_movies_fixed_runtime$Runtime[surveys_movies_fixed_runtime$Runtime == "> 2 hrs"] <- "2.5"
surveys_movies_fixed_runtime$Runtime[surveys_movies_fixed_runtime$Runtime == "30-60 mins"] <- "0.7"
surveys_movies_fixed_runtime$Runtime <- as.numeric(surveys_movies_fixed_runtime$Runtime)

one_lan <- surveys_movies_fixed_runtime %>% filter(Languages=="English" |
                                                    Languages=="Japanese" | Languages=="Korean" |
                                                    Languages=="Spanish" | Languages=="Hindi" |
Languages=="French" | Languages=="Italian")

ggplot(data = one_lan, aes(x=Languages, y=Runtime, fill=Languages)) + geom_violin()+ geom_point()+
  theme_bw() +
  labs(title = "Counting movies of a certain length per Language")
```



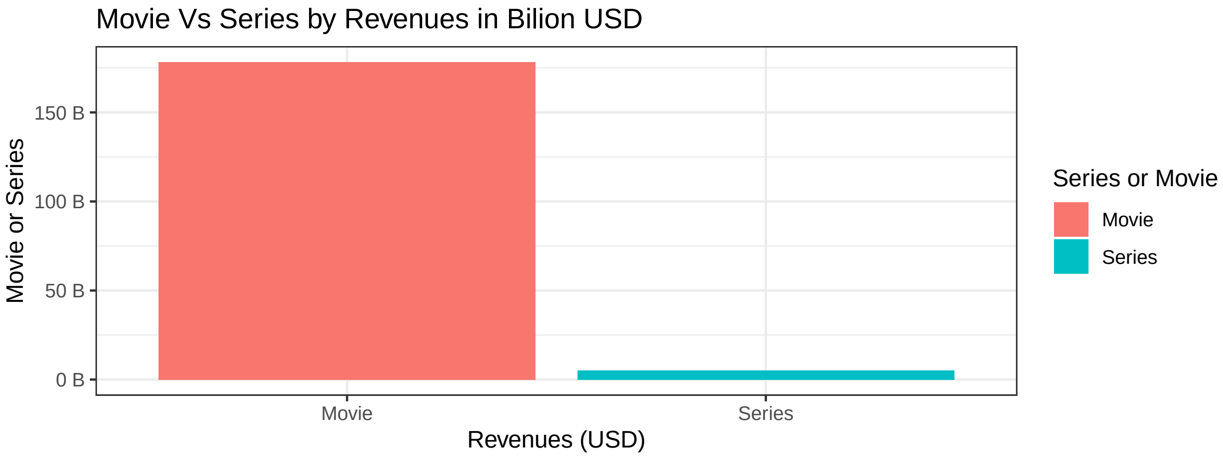
Hidden gem score is a measurement for an putative unknown good movie. We wanted to explore whether the unknown good movies is also counted as good movies in the general IMDB score. The results of those two plots show us that there is movies with mid gem-score but with very low IMDB score which means there are probably bad movies. but the movies with the high gem score are mostly have very good IMDB-score

```
ggplot(data = surveys,  aes(y =`IMDb Score` , x= `Hidden Gem Score`, color=`IMDb Score`)) +
  geom_point() + theme_classic() +
  labs(title = "IMDb Score vs Hidden Gem Score")
```

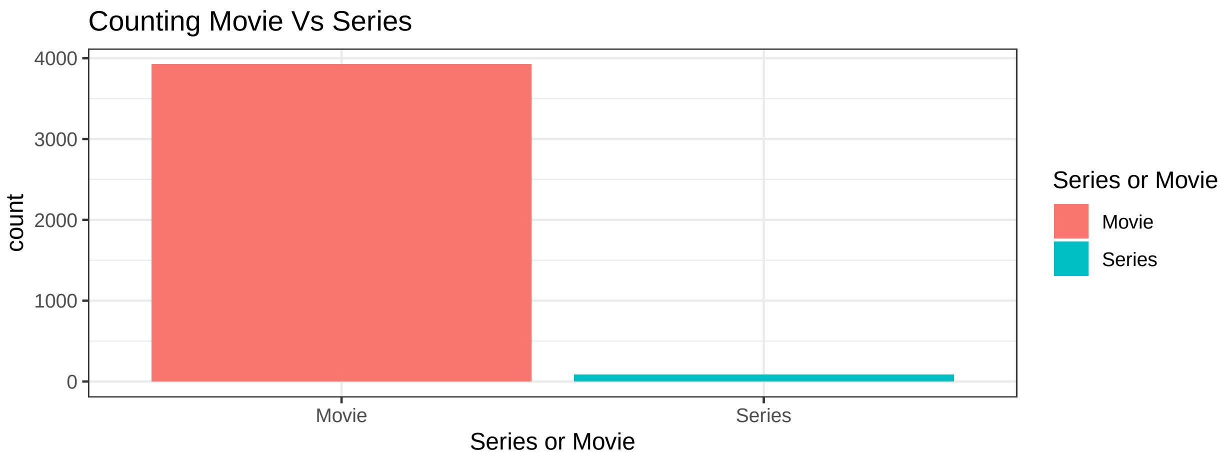


In this plot , We wanted to check the differences in revenues (Boxoffice) between movies & series. we manipulate the Boxoffice-data for having the ability to use it as a number in the graph.
We found out that movies are much more successful and have much more revenues: All movies in Netflix have together more than 150B dollars of revenue, in the other side all series have less then 50B dollars together. We also checked the number of movies and Series with data about their revenues and got that there is much more movies then series, and the difference in the sum of revenues is maybe explained by that.

```
# convert the income to integer
Boxoffice_fixed <- surveys
Boxoffice_fixed$Boxoffice <- gsub("\\$", "", surveys$Boxoffice)
Boxoffice_fixed$Boxoffice <- gsub("\\,", "", Boxoffice_fixed$Boxoffice)
Boxoffice_fixed$Boxoffice <- as.numeric(Boxoffice_fixed$Boxoffice)
# Movie vs Series on boxoffice in Billion dollars
ggplot(Boxoffice_fixed,aes(`Series or Movie`,Boxoffice,fill = `Series or Movie`)) +
   geom_col() +
  scale_y_continuous(labels = label_number(suffix = " B", scale = 1e-9)) +
  labs(title = "Movie Vs Series by Revenues in Bilion USD",
   x = "Revenues (USD)",y = "Movie or Series",) + theme_bw()
```

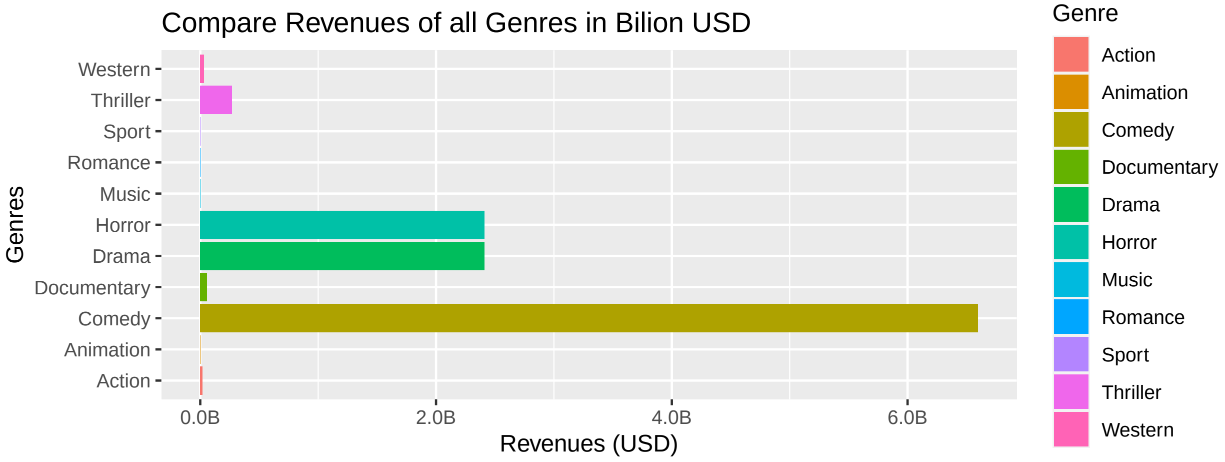


```
Boxoffice_fixed_no_NA <- subset(Boxoffice_fixed, !is.na(Boxoffice))
ggplot(Boxoffice_fixed_no_NA,aes(`Series or Movie`,fill=`Series or Movie`)) +
  stat_count(geom = "bar") +
  theme_bw() + labs(title = "Counting Movie Vs Series")
```



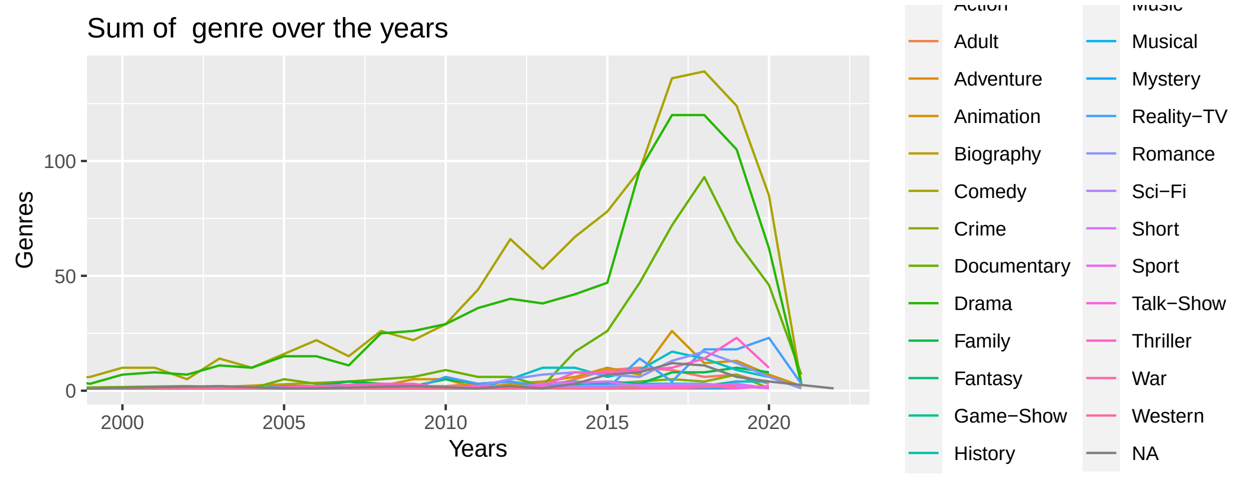
In this plot we want to check the correlation between the genre type to revenues in Billion dollars. We used the same fixed dataset from the previous plot (“Boxoffice_fixed”) with numeric values in Boxoffice column, we also edit the column of Genres - on many rows there were multiple genre in one row, so we keep only the rows with one Genre only. We also remove all rows with N/A values in genres and on Boxoffice. The remain table has ~400 row which we count as a represented sample. We saw in the plot that according to the dataset That Comedy genre has the biggest revenues. We want to emphasize the point that on this dataset there are more than 10,000 movies without any data on their revenues at all (have N/A values)

```
surveys_fixed_one_genre <- Boxoffice_fixed[!grepl(",", Boxoffice_fixed$Genre), ]
surveys_fixed_one_genre_no_NA_Genre <- subset(surveys_fixed_one_genre, !is.na(Genre))
fixed_Genres_Boxoffice<- subset(surveys_fixed_one_genre_no_NA_Genre , !is.na(Boxoffice))
# Revenues by Genres in Billion dollars
ggplot(data = fixed_Genres_Boxoffice ,mapping = aes(x=Boxoffice, y= Genre, fill = Genre)) +
  geom_col() +
  scale_x_continuous(labels = label_number(suffix = "B", scale = 1e-9))+
  labs(title = "Compare Revenues of all Genres in Bilion USD",
  y = "Genres", x = "Revenues (USD)")
```



On this plot we check the correlation between genre type and the real release date (not Netflix release date) of the movies/series. we use the same fixed dataset from the previous plot - “surveys_fixed_no_NA_Genres” and also use year column which we create from `Release date` column We found out that on the last 20 years the biggest increase is in Comedy and after that in Drama, and Family movies.

```r
# Sum of Genres per years
surveys_one_genre_fixed_date <- surveys[!grepl(",", Boxoffice_fixed$Genre), ]
fixed_Genres <- surveys_one_genre_fixed_date %>% separate(`Release Date`,
c("day", "month", "year"), sep = " ", convert = TRUE)
ggplot(fixed_Genres,aes(x = year, color = Genre)) +
     geom_line(stat = 'count') +
  labs( title = "Sum of  genre over the years",
  y = "Genres", x = "Years") + coord_cartesian(xlim = c(2000, 2022))
```


Sum of genre over the years
Genres
Years
Action
Adult
Adventure
Animation
Biography
Comedy
Crime
Documentary
Drama
Family
Fantasy
Game−Show
History
Music
Musical
Mystery
Reality−TV
Romance
Sci−Fi
Short
Sport
Talk−Show
Thriller
War
Western
NA