Final Report

Siwei Hu

December 12, 2018

Abstract

Movie is one of the most important entertainment ways in our daily life. I find a movie dataset, including 7 thousand+ movies, that scraped from Internet Movie Database (aka IMDB). I'm interested in several topics includes genres, movie budget & gross, geographic movie distribution, rating distribution and IMDB scores & votes. I check the relationship between them and get several findings. First, movie industry is becoming more and more important. Second, high budget movie does not mean high earning. Third, USA can be defined as movie center in this world. Fourth, R rating movies were pictured more than others. Fifth, Score distribution is not follow the normal distribution. Finally, votes data follows the Benfold distribution. This helped us figure out that the scores in imdb deserve our believe.

Introduction:

I. Background

Internet Movie database is one of the most useful movie website to help us rate which movie deserve to watch. It includes films, television programs, home videos and video games, and internet streams, including cast, production crew and personnel biographies, plot summaries, trivia, and fan reviews and ratings.As a very famous website, imdb can be a very good project to make me do some anaylsis for movies on it.

II. data source

I get a dataset includes nearly 7000 movies from a kaggle data challenge. Data in this data set comes from imdb.com. The author scraped data from imdb. In the data set,it has budget,company,country,director,genre,gross,name,rating,released,runtime,score,star,votes,writer,year.

```
## Parsed with column specification:
```

Figure 1: “imdb”

```
## cols(
##   budget = col_double(),
##   company = col_character(),
##   country = col_character(),
##   director = col_character(),
##   genre = col_character(),
##   gross = col_double(),
##   name = col_character(),
##   rating = col_character(),
##   released = col_character(),
##   runtime = col_integer(),
##   score = col_double(),
##   star = col_character(),
##   votes = col_integer(),
##   writer = col_character(),
##   year = col_integer()
## )
```

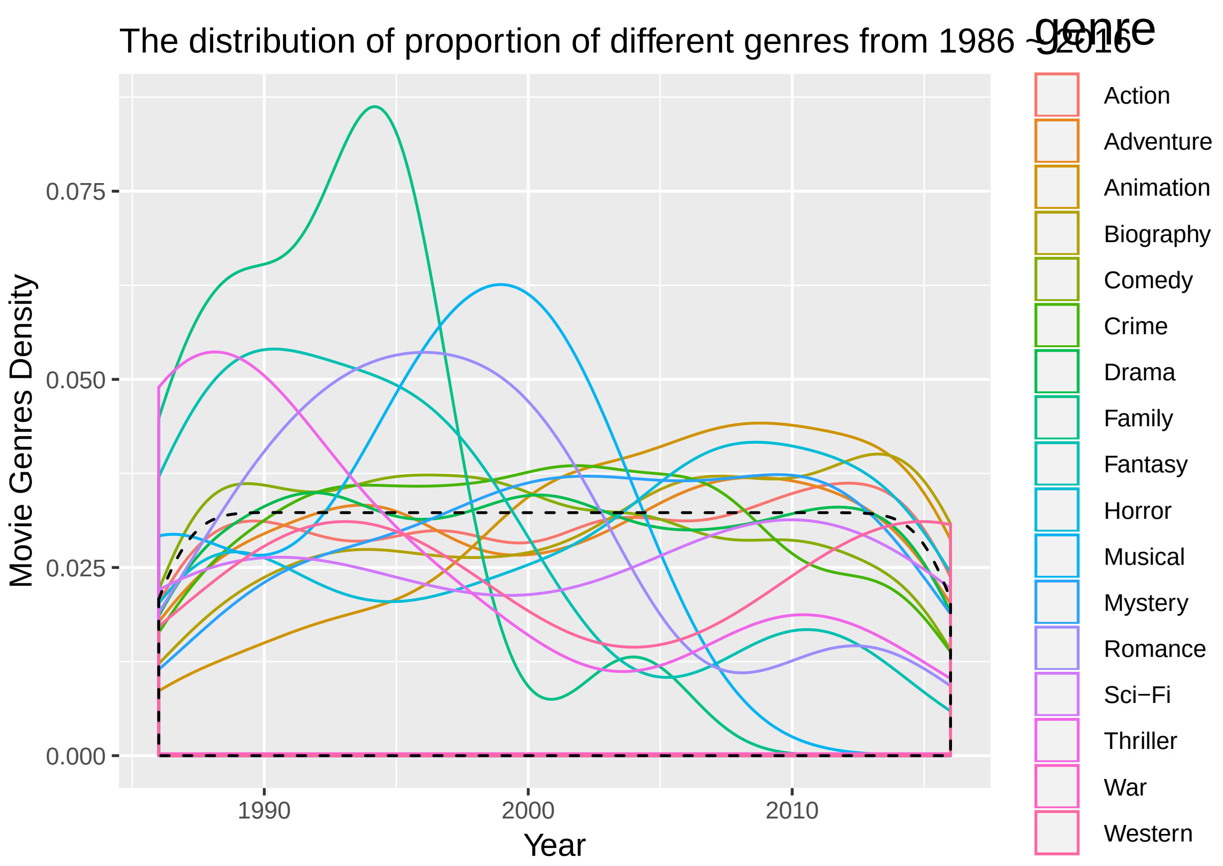
i.EDA:

a.Genre

From plot, we know how is the distribution of Years about different genres movies in this density plot.

We can know from plot:

1. Before 2000, the drama ,musical,Romance and Thriller are more popular than after 2000.
2. Scentific Movie become more popular from 2000 to 2010.



ii.Budget

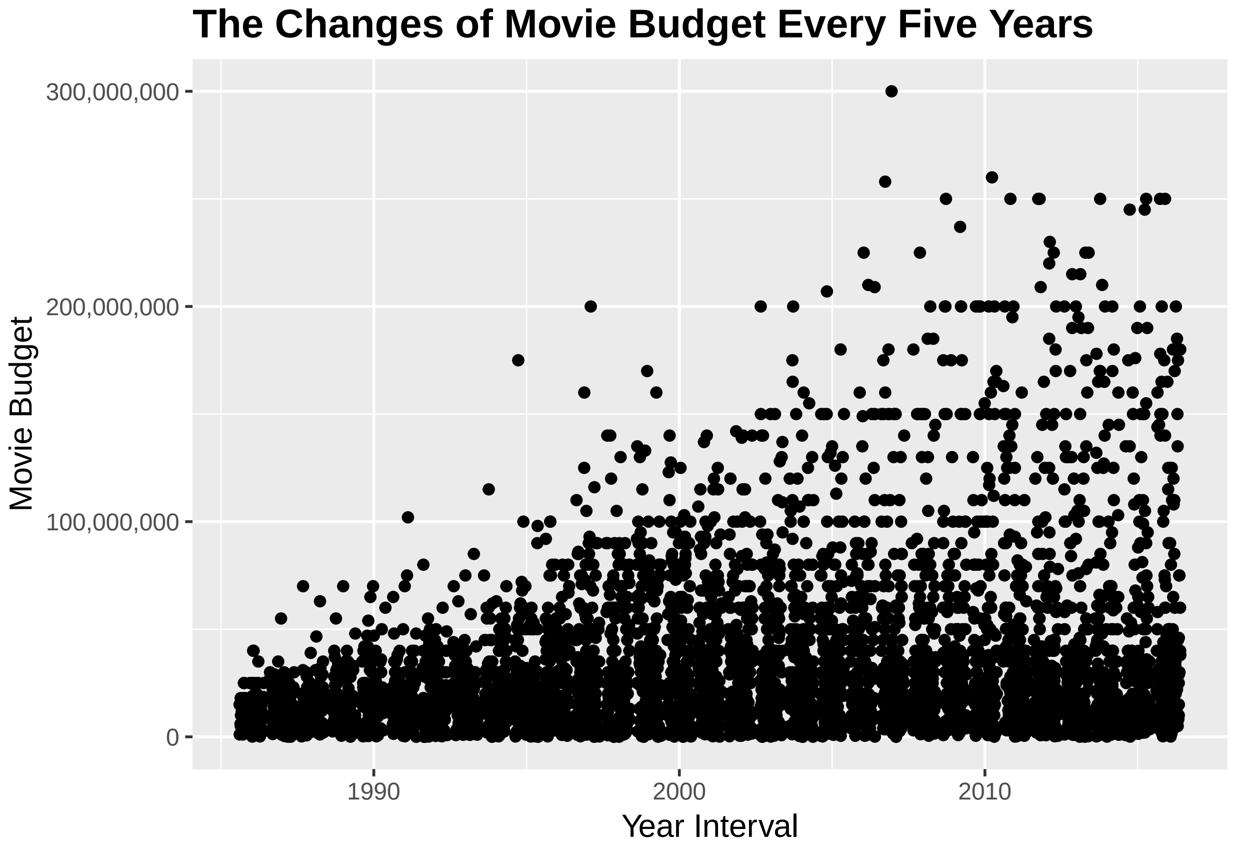
a.plots of budget of every five years' movies.

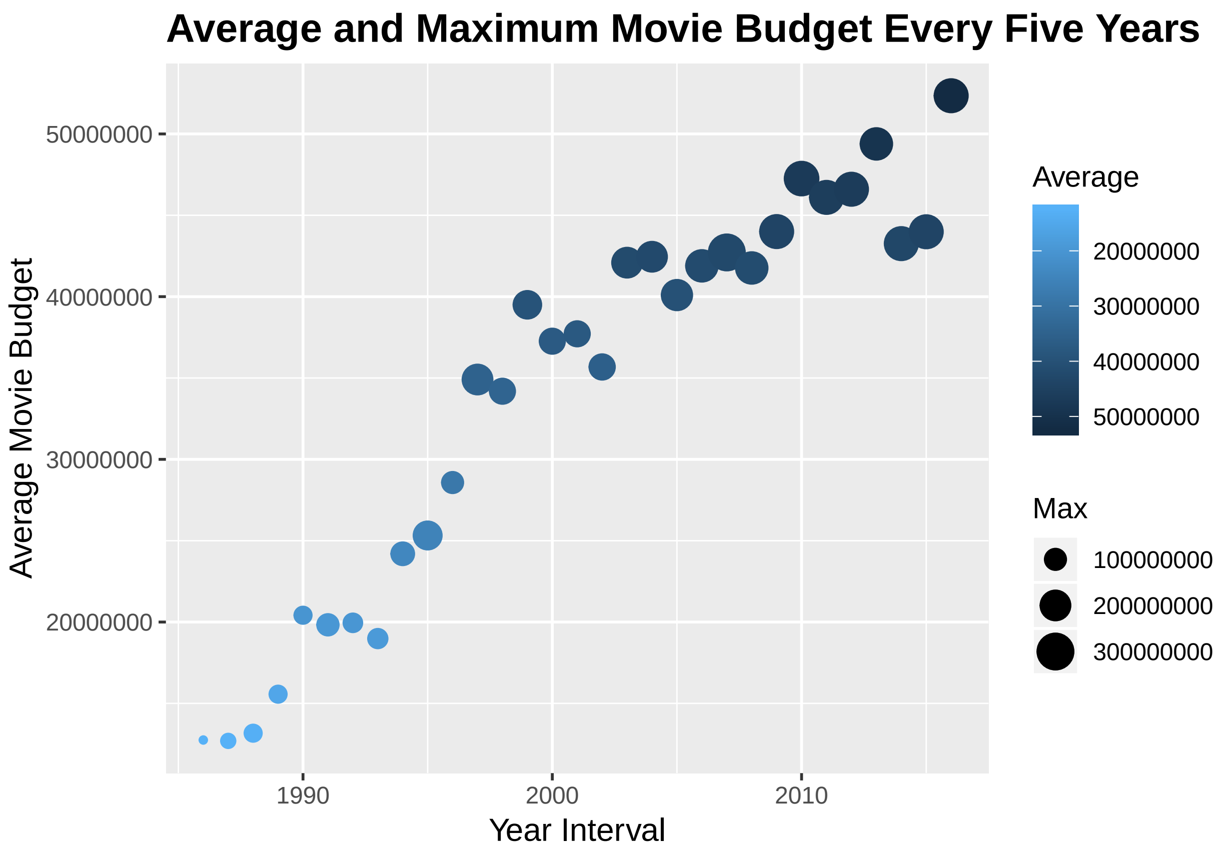
A increasing level of budget are showed in plot from 1989 to 2016. The budget becomes higher and higher along with time.

I calculated the maximum budget and average budget in the second plot. We can understand that:

1.There exists linear relationship between year and budget.

2.Average level of movie budget increases continously.


The Changes of Movie Budget Every Five Years
Movie Budget
300,000,000
200,000,000
100,000,000
0
1990
2000
2010
Year Interval



iii.Company

a. Company earning

I choose 100 higest budget movies and check their earning (gross - budget). Since i want to do facet, so i only keep company have more than one movie. only 67 left.

I found for high budget movies, the companies did not always earn money on it. Some of them made a loss. We can see this from Universal Pictures and WaltDisney Pictures.

Also, we can know that there exist several company earn a lot from high budget movies, like 21 Century Fox.

In conclusion, the ratio of positive earning is more than negative earning from this Top 100 highest budget.

```
comp <- imdb %>% arrange(desc(budget))
comp <- comp[1:100,]
comp1 <- comp %>%  group_by(company,year) %>% summarise(budget = sum(budget)) %>% ungro
```

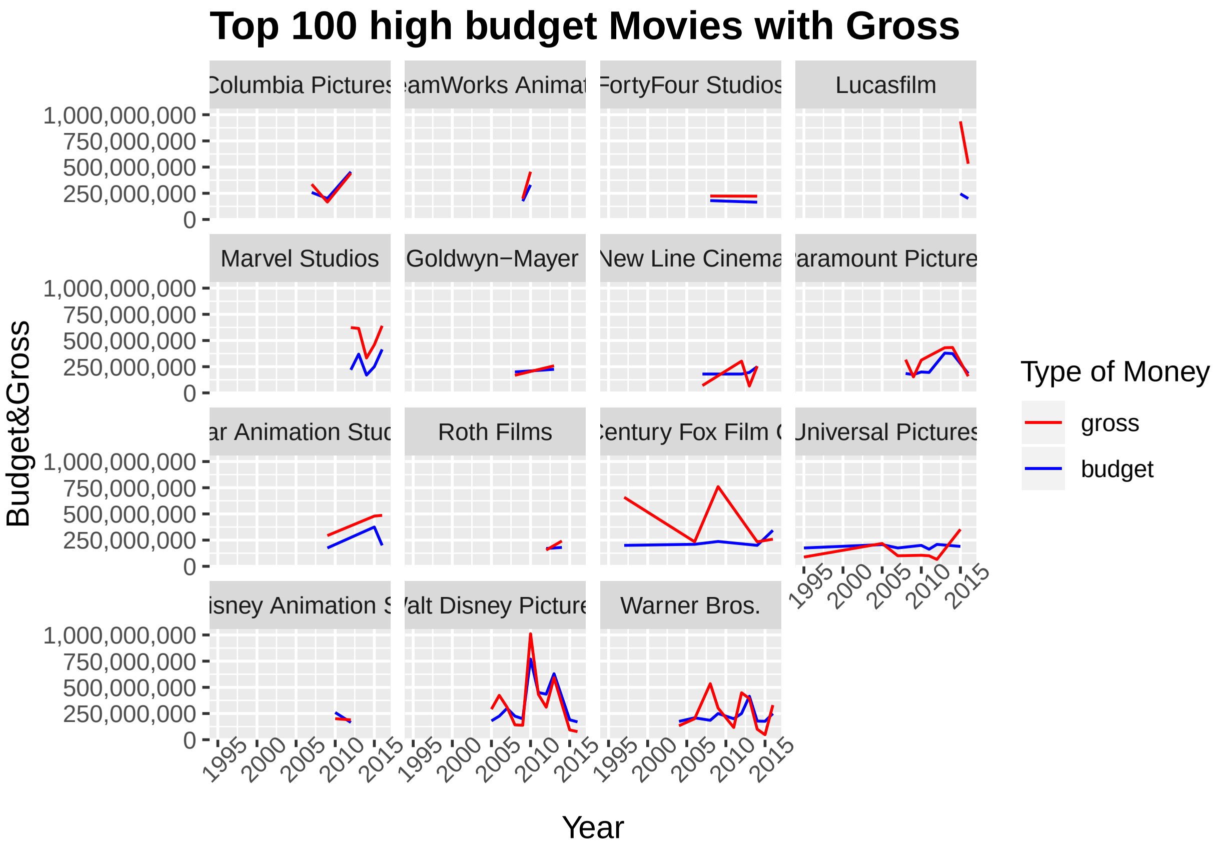
```
comp2 <- comp %>% group_by(company, year) %>% summarise(gross = sum(gross)) %>%  ungrou

a <- ifelse(comp$gross - comp$budget >0, 1 ,-1)
sum(a)
```

```
## [1] 22
```

```
year <- c(95, 97, 04 ,05 ,06, 07, 08, 09 ,10, 11, 12 ,13, 14 ,15 ,16)

ggplot(data = comp1)+ geom_line(aes(x = year,y = budget,color = "red")) + geom_line(aes
```

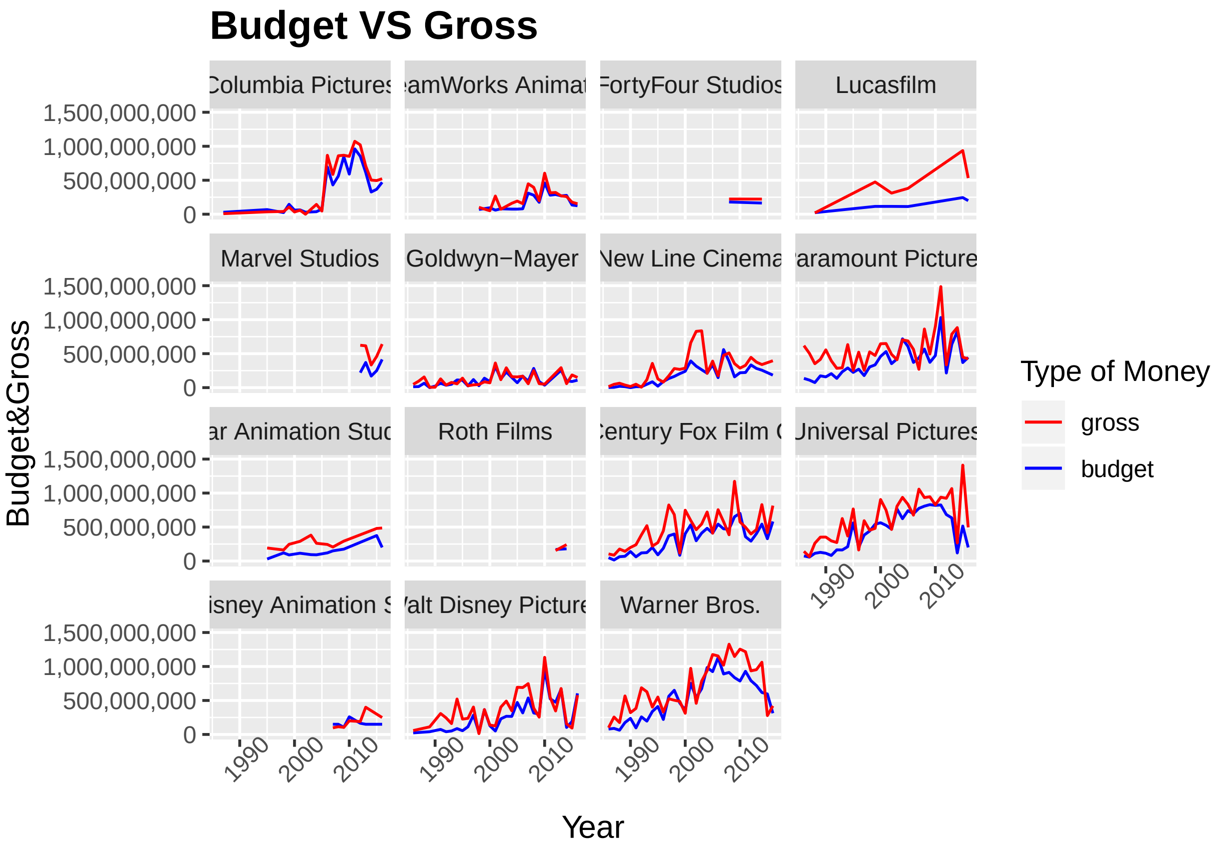


b. Company's suitation

From the previous plot, i am interested in companies earning suituation. So I use the company names from 100 highest budget and analysis their all movies in my data set.

See their earning suitation from the line plot.

From these plot, these Movie company still earn a lot from all movies. High budget Movies have both high compentasion and high risk. However, big movie companies have ability to hedge risk.



iv. Geograph

I draw a geograph plot to see how many movies was made in different countries from 1989 to 2016.

America made the most movie in my dataset. I think this makes sense because

1.hollywood and the biggest movie companies are in the America.

2.The audiences are attracted by the Superhero theme.

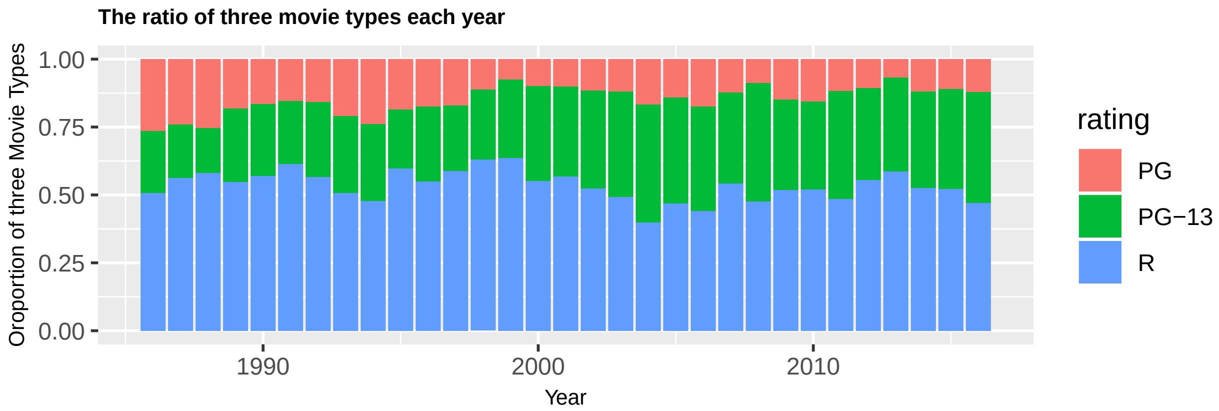
```
{r, echo=FALSE} #  # imdb.map <- read_csv("imdb.map.csv")
# wm <- geojsonio::geojson_read("WorldMap.geo.json",
what = "sp") # wm@data <- left_join(wm@data,imdb.map,
by ="name")  #  # ## Leaflet Map # bins <- c(1,5,10,25,50,10
# pal <- colorBin("YlGnBu", domain = wm@data$count,
bins = bins) #  # leaflet(wm) %>% #   setView(41.5,12.6,2)
%>%  #   addTiles() %>% #   addPolygons(stroke =
FALSE, smoothFactor = 0.3, fillOpacity = 1, #     fillColor
= ~pal(count), #     label = ~paste0("Movies made
by ",name, ":", formatC(count))) %>% #   addLegend(pal
= pal,title = "# of Movies", values = ~count, opacity
= 1.0)  #  #  #
```

v.Rating ratio

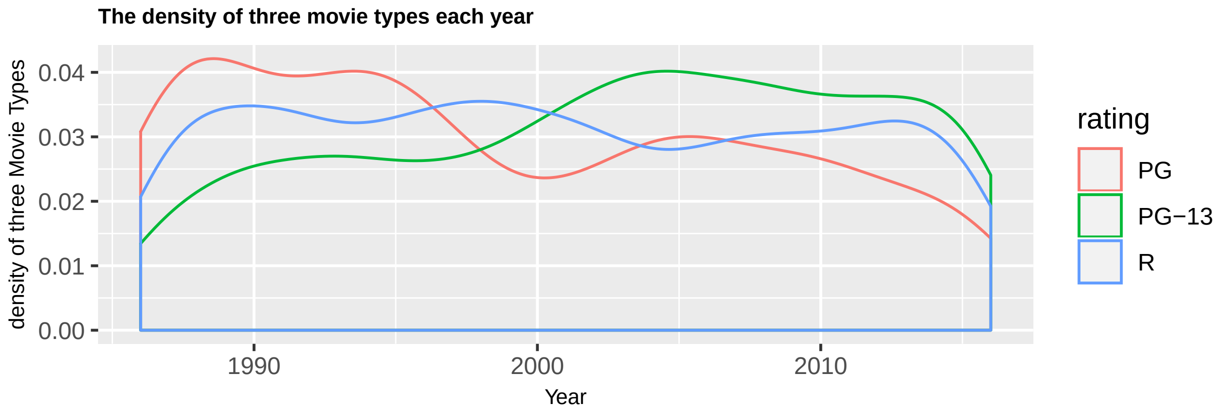
a.Rating distribution

From plot, R rating movies were pictured most each year.This can be checked each year movie voters and score via three ratings.

The density plot displays a decreasing trend of density of PG. From 1986 to 2005, pg-13 shows a increasing trend and then it shows a decreasing trend from 2006 to 2016. R rating keeps same.



The ratio of three movie types each year

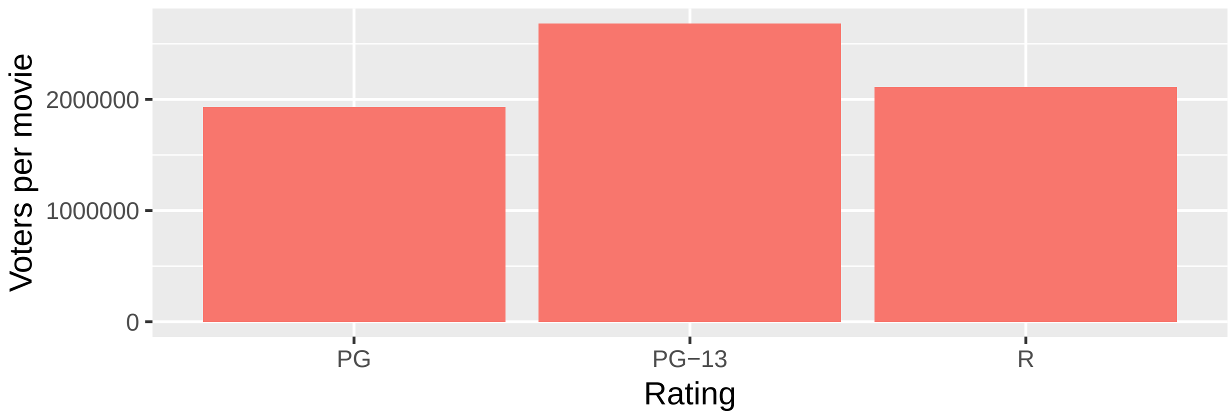


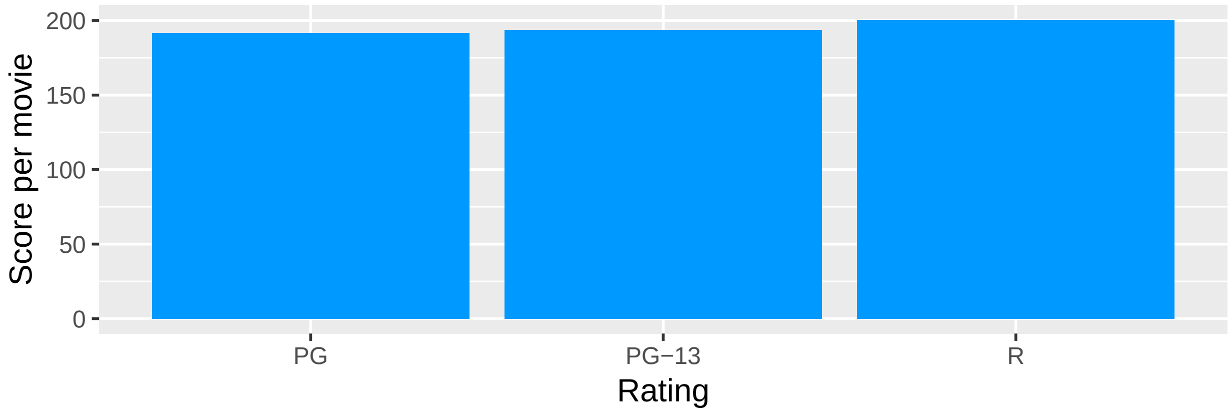
The density of three movie types each year

b. people's attitude to three rating

From plots, it's clear to see from voter distribution for three rating.

More people are willing to vote for PG-13 movies.





vi.runtime with earning

Interested in whether longer runtime movies have higher earning.

See smooth lines, there exists higher se at longer runtime because sample size with long run time is not too much.

I found runtime from 125 to 150 exist a significant higher earning than other runtime.

```
##Year vs runtime
imdb.length <- imdb  %>% dplyr::select(runtime,year,gross,budget,released)%>% filter(bud


ggplot(data = imdb.length) + geom_smooth(aes(x= runtime, y = budget,color = "budget"))+g
```