Data Wrangling Final Report

Jingyi Gu

2018/4/19

Introduction

Consumer comments and reviews by movie critics may dramatically affect success of movies, and even decision making about whether consumers will see a movie. The most popular movie rating websites are IMDb and Rotten Tomatoes. For my study, I will deeply mine data from these two websites to analyze the factors affecting the success of a movie on each website. This is important as it may provide information that will allow producers to learn which factors contribute to higher profits.

I will scrape the data from IMDb and Rotten Tomatoes websites. Both dataset contains approximately 5,000 samples and variables including budget, genres, keywords, original language, popularity, release date, and ratings from critics and consumers.

I will clean the dataset, then use packages in R such as tidyverse, ggplot2, to find more valuable information and analyze the factors affecting success.

Data cleaning

First we need install some neccessary packages.

```
library(tidyr)
library(tidyverse)
library(ggplot2)
library(jsonlite)
library(stringr)
library(wordcloud)
library(ggrepel)
```

We got data from 2 resources. Firstly we download the information about IMDb from kaggle. Then we got data about Rotten Tomatoes from github. We need to unzip the data. Then we merge them together by the title of movie.

```
imdb<-read.csv("tmdb_5000_movies.csv",stringsAsFactors = F)

rotten<-"http://files.grouplens.org/datasets/hetrec2011/hetrec2011-movielens-2k-v2
.zip"
file <- tempfile(tmpdir=tempdir(), fileext=".zip")
download.file(rotten,file)
rt<-unzip(file,list = TRUE)$Name[c(1,2,3,4,5,6,7,8,9, 10, 11, 12, 13)]
unzip(file,files=rt,exdir=tempdir(), overwrite=TRUE)
imdb_rotten <-read.delim(file.path(tempdir(), rt)[8], header=TRUE,sep="\t")
imdb_rotten<- imdb_rotten[!duplicated(imdb_rotten$title),]

movie <- merge(imdb, imdb_rotten, by.x = "original_title", by.y = "title")
```

We find that there are 40 variables. There are some redundant information and characters in the dataset. Now let's clean the data and find out the information we need.

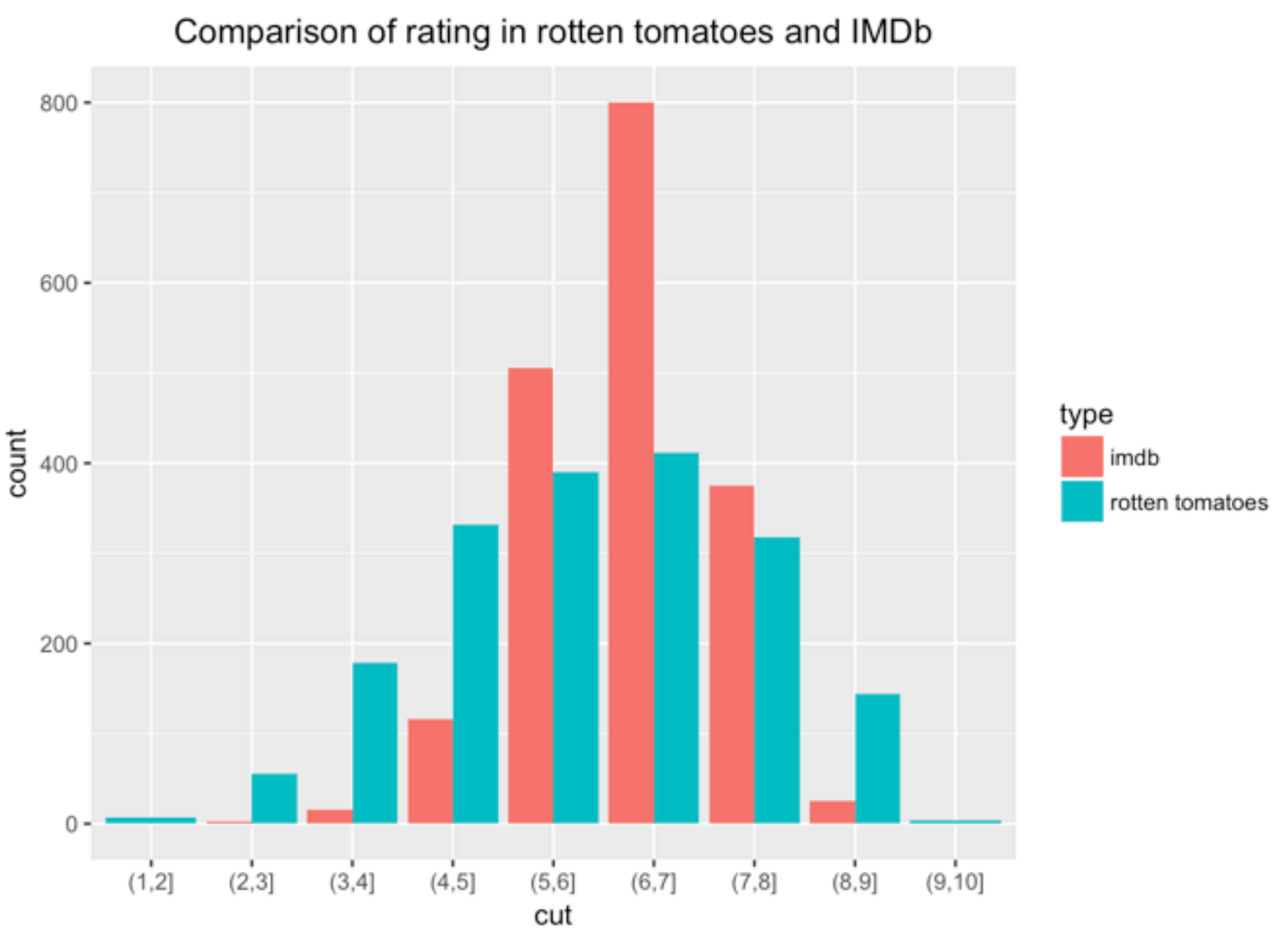
```
movie <- movie %>%
  filter(budget != 0) %>%
  filter(revenue != 0) %>%
  filter(!str_detect(as.character(rtAllCriticsRating),"\\\\N")) %>%
  filter(!is.na(as.numeric(as.character(rtAllCriticsRating)))) %>%
  filter(as.numeric(as.character(rtAllCriticsRating)) != 0) %>%
  select(-c(homepage,id.x,original_title,id.y,imdbID,imdbPictureURL,spanishTitle,r
tPictureURL))

movie$rtAllCriticsRating<-as.numeric(as.character(movie$rtAllCriticsRating))

keyword<-movie %>%
  filter(nchar(keywords)>2) %>%
  unnest(lapply(keywords,fromJSON)) %>%
  select(title,keyword=name)

companies<-movie %>%
  filter(nchar(production_companies)>2) %>%
  unnest(lapply(production_companies,fromJSON)) %>%
  select(budget,revenue,company=name)

genre<-movie %>%
  filter(length(genres)>2) %>%
  unnest(lapply(genres,fromJSON)) %>%
  select(title,genre=name,vote_average,rtAllCriticsRating)

country<-movie %>%
  filter(nchar(production_countries)>4) %>%
  unnest(lapply(production_countries,fromJSON)) %>%
  select(budget,revenue,popularity,country=name,vote_average,rtAllCriticsRating)

write_csv(movie,path = "movie.csv")
write_csv(keyword,path = "keyword.csv")
write_csv(companies,path = "companies.csv")
write_csv(genre,path = "genre.csv")
write_csv(country,path = "country.csv")
```

We filtered the data and deleted some unuseful variables and values which are equal to 0. Now we have 1840 movie samples and 32 variables. Also we transformed type of variables from factor to numeric, which is convenient for us to do more analysis. Then we extracted some variables which we had interest to find more information and made a single data frame for each variable. We saved all tidy version of the data as csv files.

Analysis

First let's see how people rate these movies. The rating ranges from 0 to 10. We made a bar plot to see how rating distributed and compare difference between IMDb and Rotten Tomatoes.

```
m<-seq(0,10,by=1)
imdb = as.character(cut(movie$vote_average,m))
rt = as.character(cut(as.numeric(as.character(movie$rtAllCriticsRating)),m))

rate<-matrix(NA,2*dim(movie)[1],2)
rate[1:dim(movie)[1],1]<-imdb
rate[(1+dim(movie)[1]):(2*dim(movie)[1]),1]<-rt
rate[1:dim(movie)[1],2]<-"imdb"
rate[(1+dim(movie)[1]):(2*dim(movie)[1]),2]<-"rotten tomatoes"
colnames(rate)<-c("cut","type")

ggplot(as.data.frame(rate))+
  geom_bar(aes(x=cut,fill=type), position = "dodge") +
  ggtitle("Comparison of rating in rotten tomatoes and IMDb") +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_colour_discrete(name = "website") +
  scale_colour_discrete(labels = c("Rotten tomatoes", "IMDb"))
```

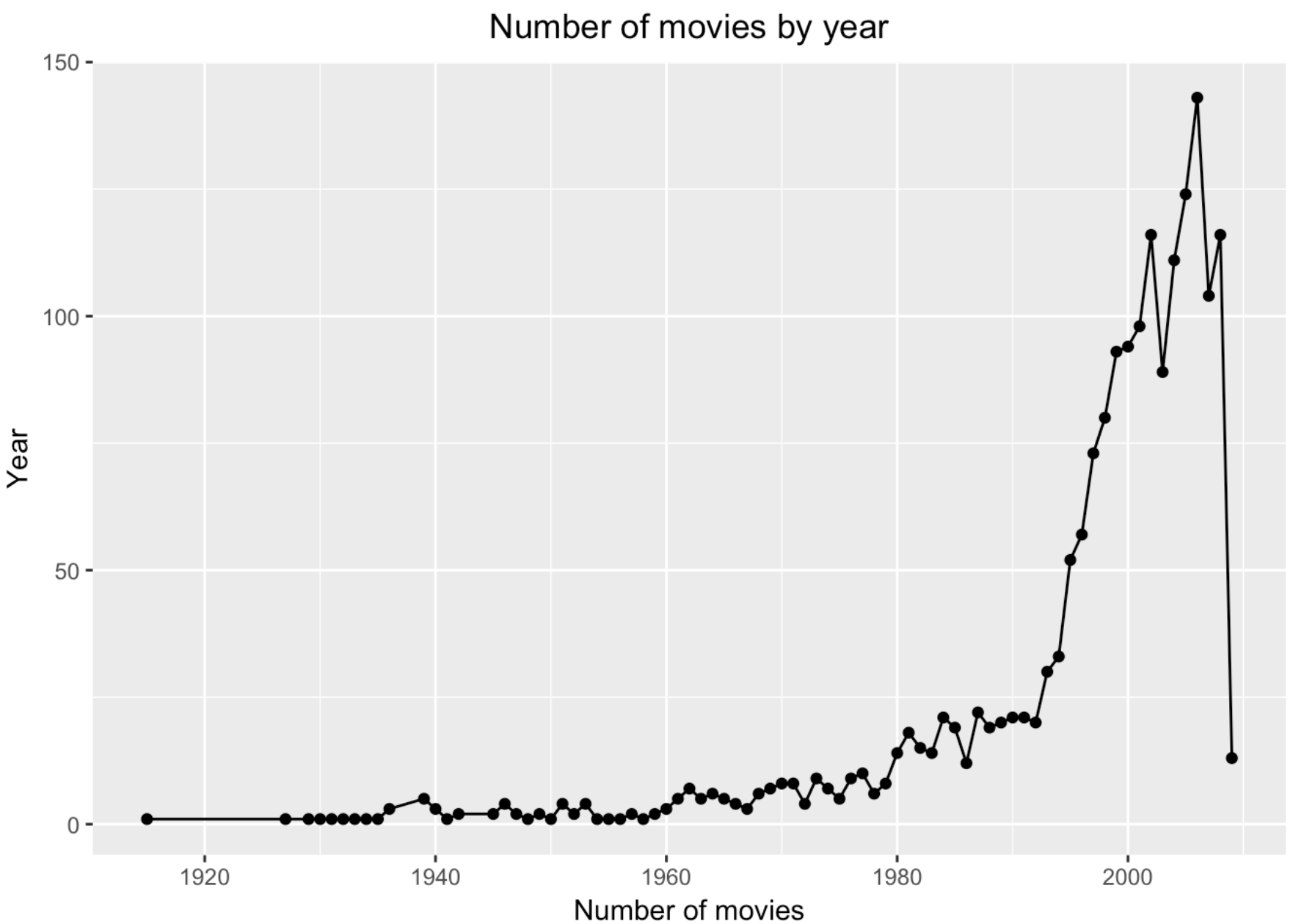


We cut the rating into 9 parts. From the plot we find that the majority of movie in IMDb website are rated between 5 and 7, which is almost twice as those in Rotten Tomatoes. For those rated too low or too high, Rotten Tomatoes has more movies than IMDb.

```
movie_num<-movie %>%
  select(title,year) %>%
  group_by(year) %>%
  summarise(count = n()) %>%
  arrange(desc(count))
head(movie_num)
```

```
## # A tibble: 6 x 2
##    year count
##   <int> <int>
## 1  2006   143
## 2  2005   124
## 3  2002   116
## 4  2008   116
## 5  2004   111
## 6  2007   104
```

```
ggplot(movie_num, aes(year,count)) +  geom_line() + geom_point() +
  ggtitle("Number of movies by year")+
  xlab("Number of movies") + ylab("Year")  +
  theme(plot.title = element_text(hjust = 0.5))
```



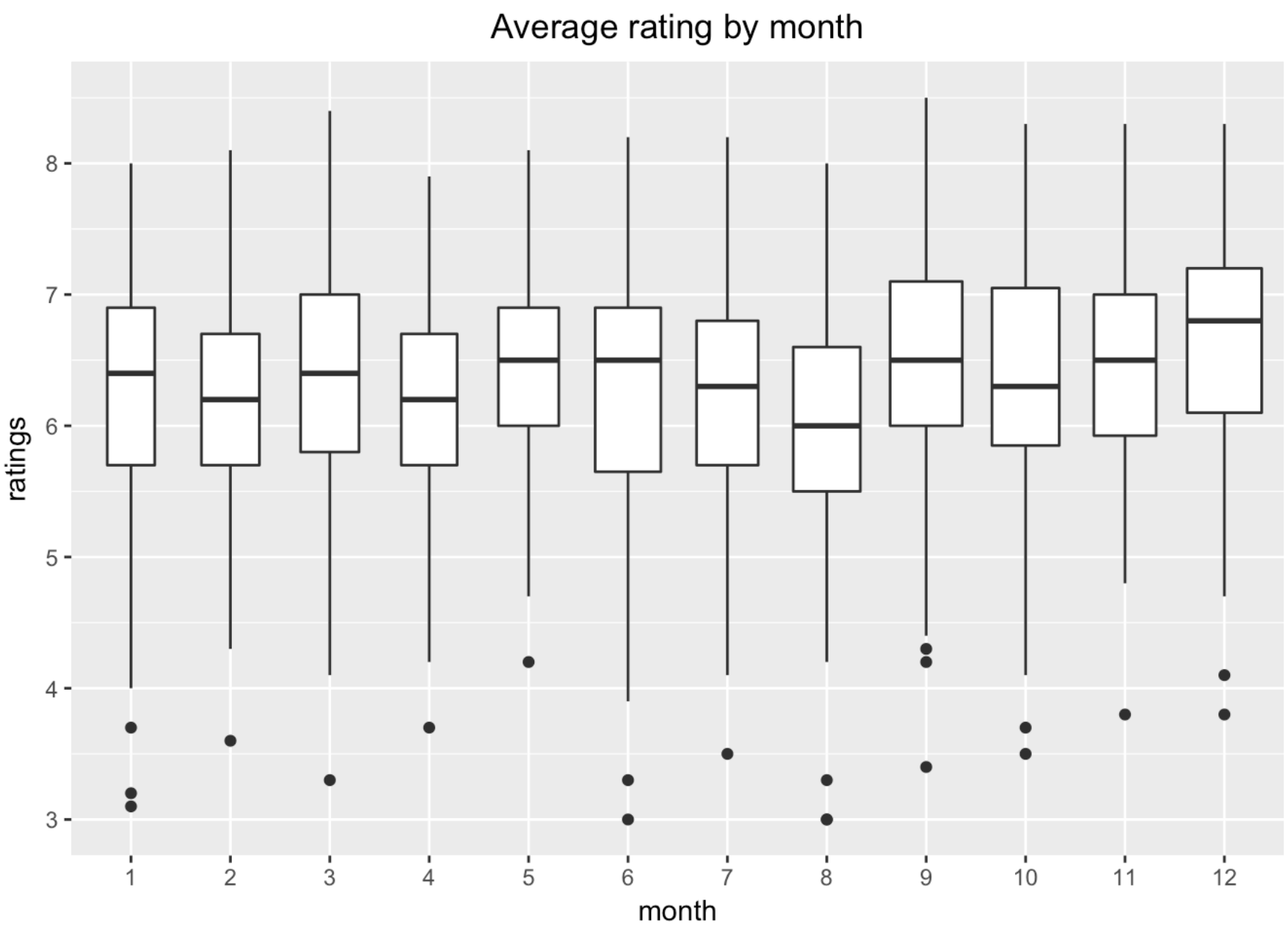
The plot shows that the number of movie was stable before 1980, and began to increse since 1980, and after 1990 the upward trends was obvious. It reached peak at 2006, then it fell sharply. The reason is that the demand of entertainment was increasing with the development of economics after 1990, which leads to explosion of movies.

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

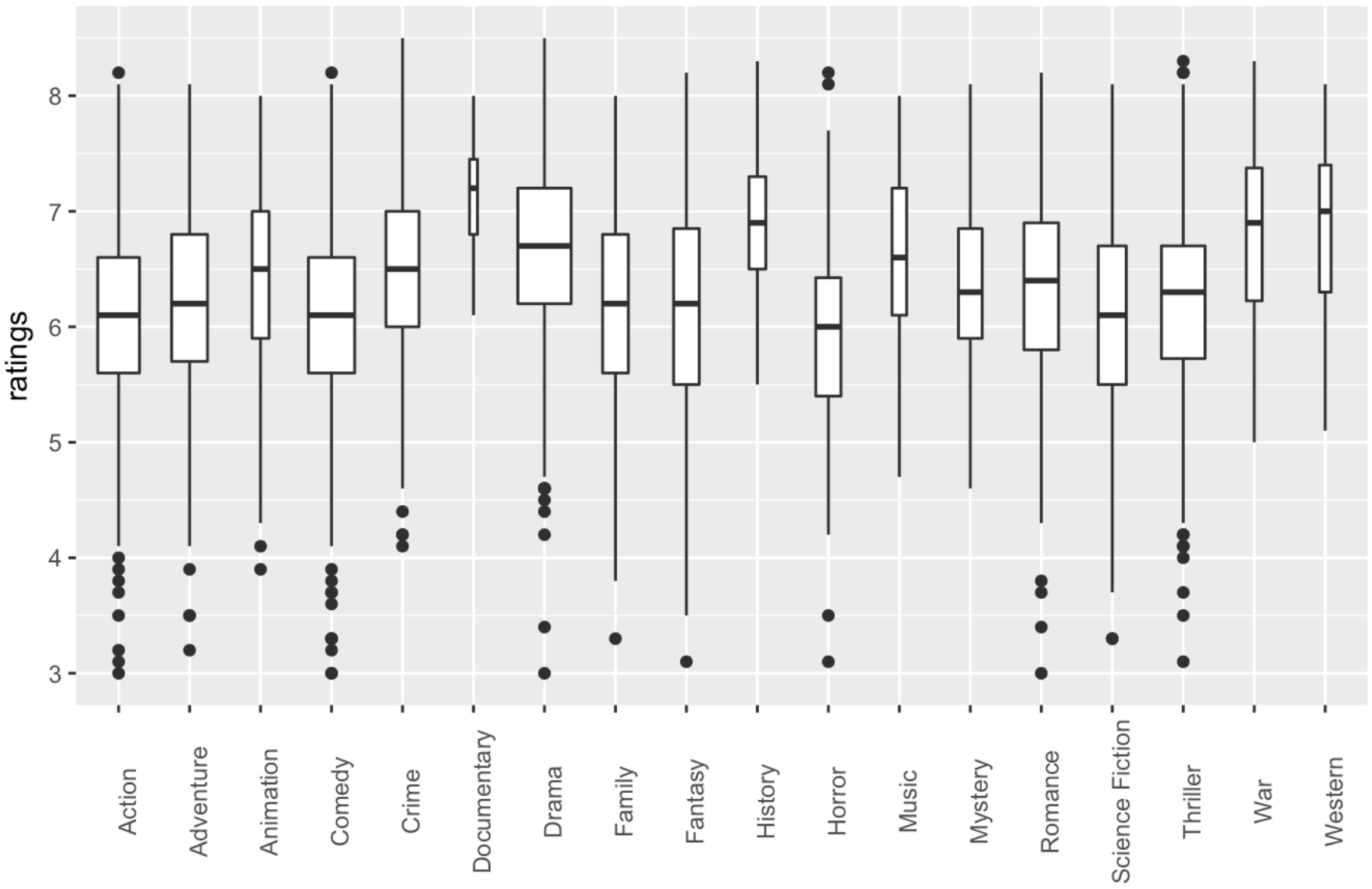
```
month<-data.frame(tibble(m = as.factor(month(ymd(movie$release_date))),
                         rate = movie$vote_average))

month %>%
  ggplot(aes(m,rate))+
  geom_boxplot(varwidth = T) +
  xlab("month") + ylab("ratings") +
  ggtitle("Average rating by month") +
  theme(plot.title = element_text(hjust = 0.5),
        legend.position="none")
```



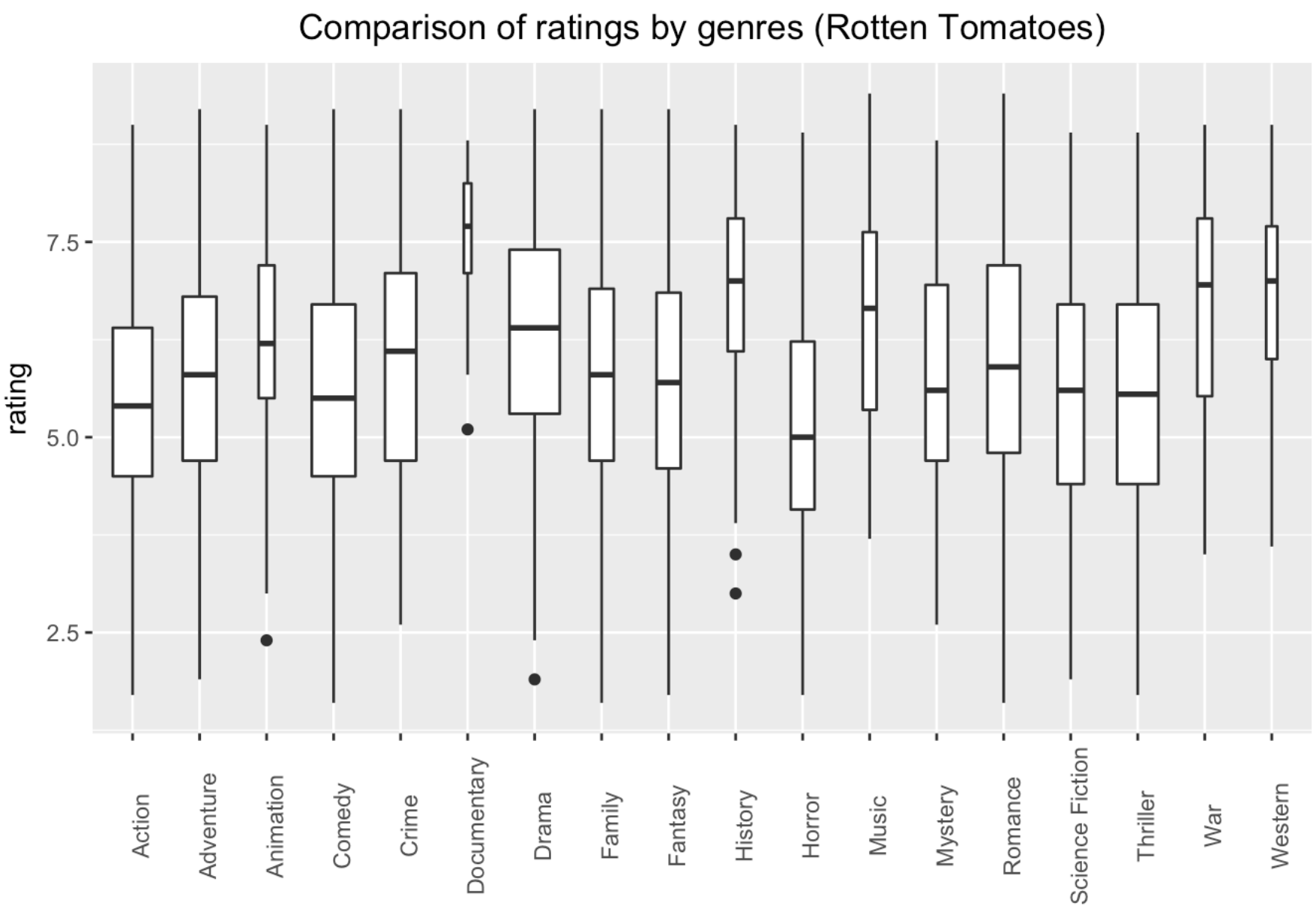
The boxplot shows the average rate and the number of movie by month. The width of box means the number of movie. The wider the box is, the larger the total number of movie released in that month is. It is obvious that September has maximum releases, which is followed by October.

The box plot also shows the range, mid-range, midhinge and trimean of data. Median of average rating in December is the highest one, in opposite average vote rating in August releases is the lowest. The rating in January releases has the largest range.

```
genre %>%
  ggplot(aes(genre,vote_average)) +
  geom_boxplot(varwidth = T) +
  xlab("") + ylab("ratings") +
  ggtitle("Comparison of ratings by genres (IMDb)") +
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle=90),legend.position="none")
```




```
genre %>%
  ggplot(aes(genre,rtAllCriticsRating)) +
  geom_boxplot(varwidth = T) +
  xlab("") + ylab("rating") +
  ggtitle("Comparison of ratings by genres (Rotten Tomatoes)") +
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle=90),legend.position="none")
```



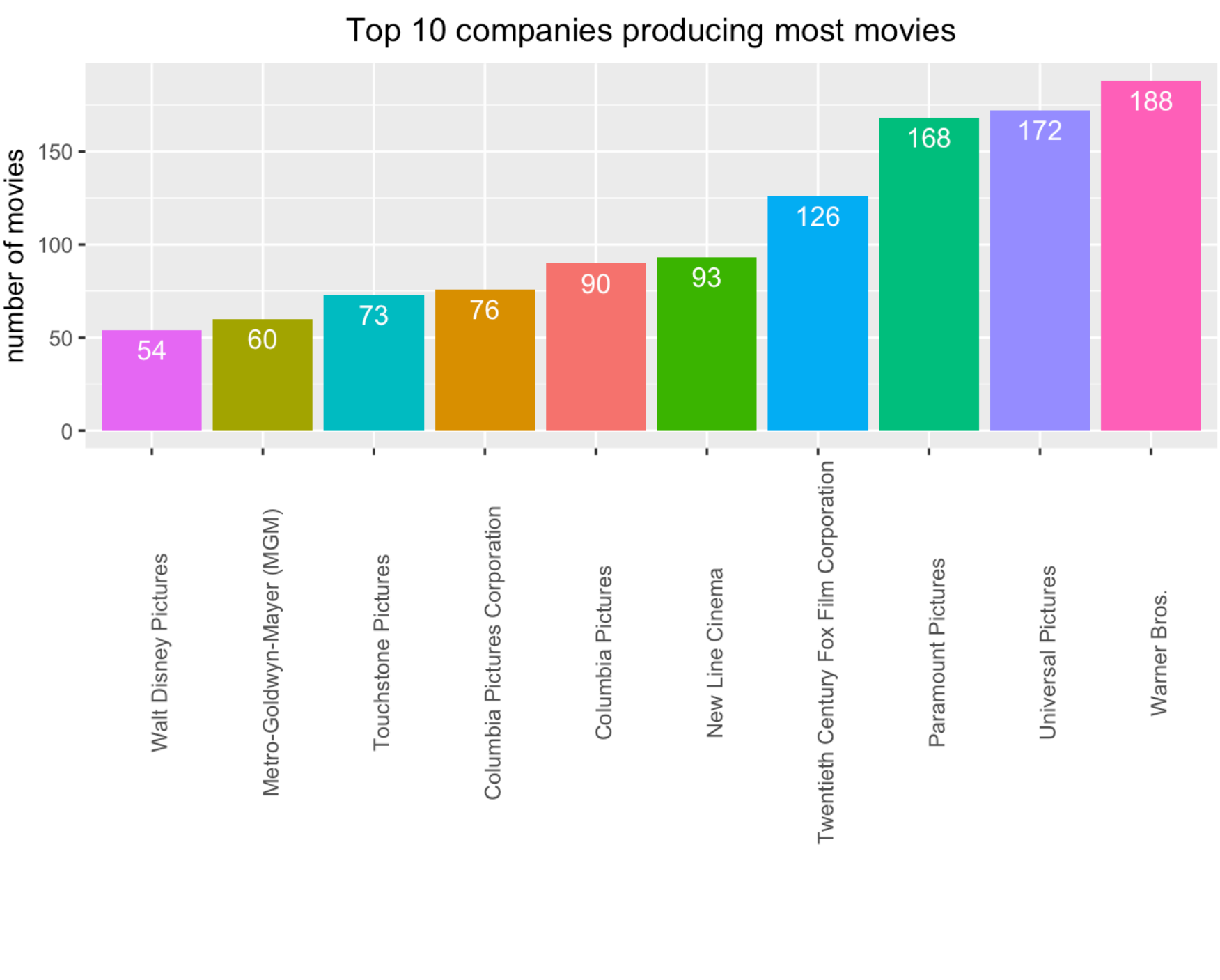
We plot the ratings by genres in Roteen Tomatoes and IMDb seperately. In IMDb, we find that drama, comedy, thriller and action have the largest number of movies. There are less movies in documentary, western, war and history. It is interesting that movies in these types are rated higher than others. Conversely audience do not like horror and family movies.

In Rotten Tomatoes, the result is similar in the number distribution in each type. Average ratings are around 6, and audience like drama and comedy movies most considering rating and populatiry. For those movies with less popularity and high ratings, we guess it is because that all audience rating these movies prefer these types so that they are unlikely to give low ratings.

```
com_movie<-companies %>%
  group_by(company) %>%
  summarise(count = n(),total = sum(revenue)) %>%
  arrange(desc(count)) %>%
  head(10)
com_movie
```

```
## # A tibble: 10 x 3
##                                    company count       total
##                                      <chr> <int>       <dbl>
##  1                           Warner Bros.   188 28898456327
##  2                     Universal Pictures   172 23025593727
##  3                     Paramount Pictures   168 22375257501
##  4 Twentieth Century Fox Film Corporation   126 21255692634
##  5                        New Line Cinema    93 10061196842
##  6                      Columbia Pictures    90 10399918995
##  7          Columbia Pictures Corporation    76  8592903093
##  8                    Touchstone Pictures    73  8227356570
##  9             Metro-Goldwyn-Mayer (MGM)    60  4117011267
## 10                  Walt Disney Pictures    54 13620042534
```

```
com_movie %>%
  ggplot(aes(reorder(company,count),count,fill = company)) +
  geom_bar(stat="identity") +
  xlab("") + ylab("number of movies") +
  ggtitle("Top 10 companies producing most movies") +
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle=90),legend.position="none") +
  geom_text(aes(label=count),vjust=1.5,colour="white")
```



We find top 10 companies with most movies. Obviously Warner Bros produced the largest number of movie in total which is 188, followed by Universe Pictures with 172 movies and Paramont Pictures with 168 movies. The table also shows the 10 companies and their total revenue.

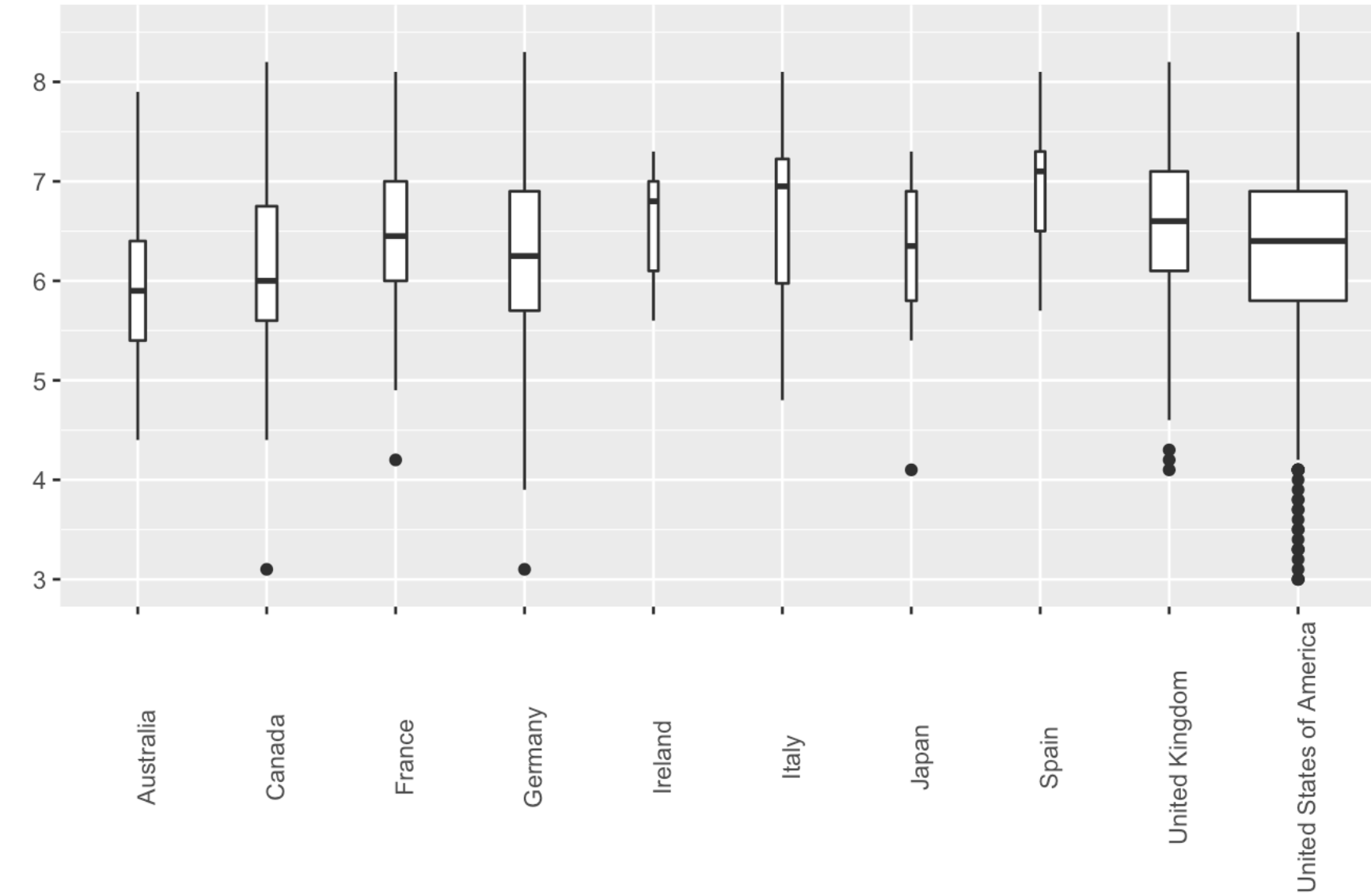
```
word<-keyword %>%
  group_by(keyword) %>%
  summarise(count=length(keyword)) %>%
  arrange(desc(count))
word %>% head(15)
```

```
## # A tibble: 15 x 2
##                 keyword count
##                   <chr> <int>
##  1               murder    88
##  2     independent film    87
##  3       woman director    87
##  4 duringcreditsstinger    78
##  5       based on novel    77
##  6             violence    73
##  7             dystopia    62
##  8  aftercreditsstinger    60
##  9                sport    59
## 10          los angeles    53
## 11              revenge    50
## 12             new york    49
## 13      dying and death    47
## 14               police    46
## 15           friendship    44
```

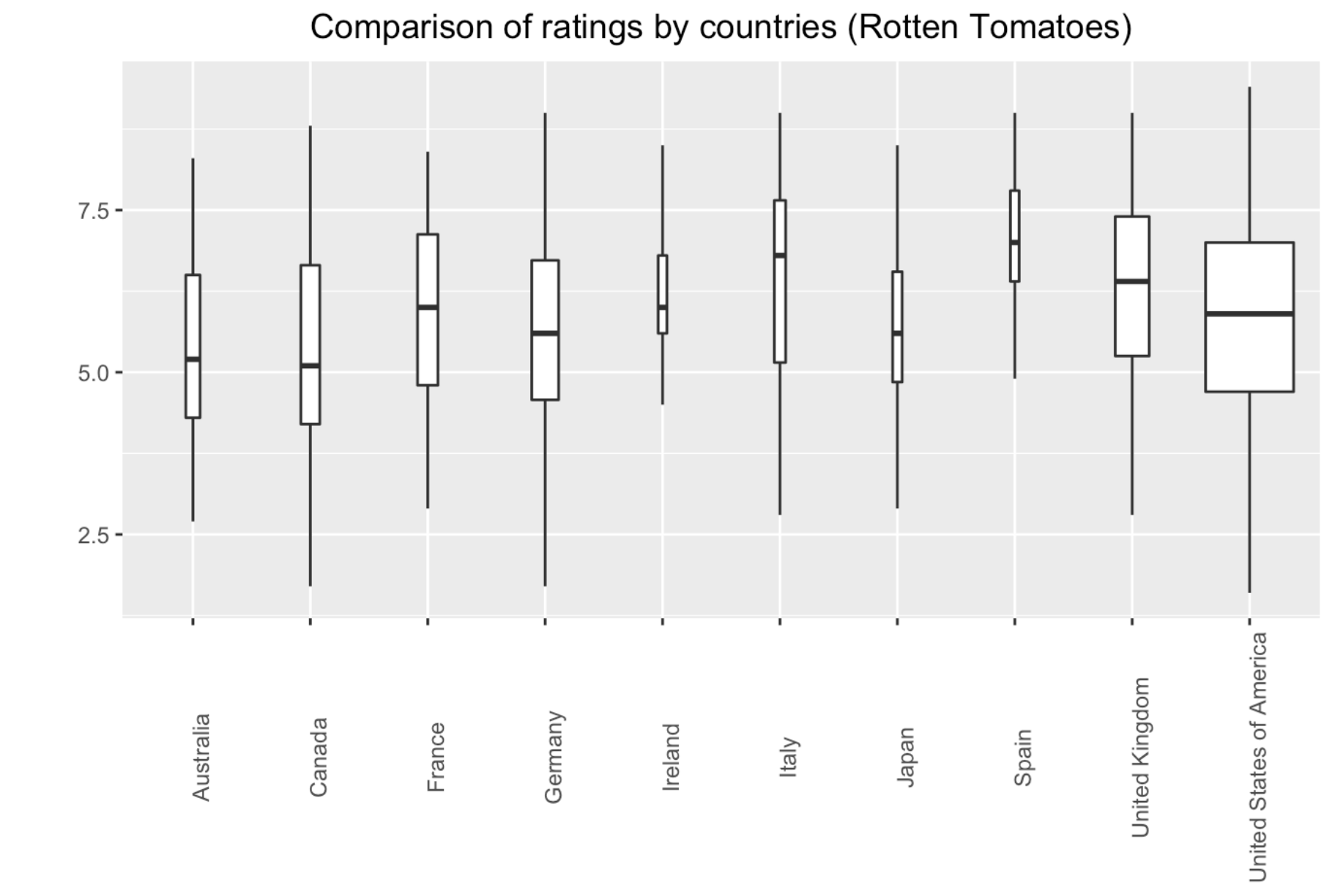
```
wordcloud(words = word$keyword,freq = word$count,
          random.order=FALSE,max.words = 150,rot.per=0.60,
          colors=brewer.pal(8, "Dark2"))
```

We are interested in keywords for each movie, so we had a word cloud plot showing the most frequent words on these 2 websites. Murder, violence and los angeles appears at most times.

```
country10<-country %>%
  group_by(country) %>%
  summarise(count = n()) %>%
  arrange(desc(count)) %>%
  head(10)

country %>%
  filter(country %in% country10$country) %>%
  ggplot(aes(country,vote_average)) +
  geom_boxplot(varwidth = T) +
  xlab("") + ylab("") +
  ggtitle("Comparison of ratings by countries (IMDb)") +
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle=90),legend.position="none")
```




```
country %>%
  filter(country %in% country10$country) %>%
  ggplot(aes(country,rtAllCriticsRating)) +
  geom_boxplot(varwidth = T) +
  xlab("") + ylab("") +
  ggtitle("Comparison of ratings by countries (Rotten Tomatoes)") +
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle=90),legend.position="none")
```



Let's compare the ratings by different countries in 2 websites. In IMDb website, USA is the country with most movies, however, the average rating for American movies is not the highest one. The second country is United Kindom. It is worth mentioning that the median of average vote rating in Spainish movies is the highest one even though the number of Spainish movie is smaller than others. The rating and number of italian movie is similar to Spanish movies. In Rotten Tomatoes, the range of rating in each country is larger than those in IMDb.

```
popular_movie<-movie %>%
  select(title,popularity) %>%
  arrange(desc(popularity)) %>%
  head(10)
popular_movie
```

```
##                                                      title popularity
## 1  Pirates of the Caribbean: The Curse of the Black Pearl   271.9729
## 2                                          The Dark Knight   187.3229
## 3                                               Fight Club   146.7574
## 4              Pirates of the Caribbean: Dead Man's Chest   145.8474
## 5                                            The Godfather   143.6597
## 6                              Teenage Mutant Ninja Turtles   143.3504
## 7                Pirates of the Caribbean: At World's End   139.0826
## 8                                            Forrest Gump   138.1333
## 9       The Lord of the Rings: The Fellowship of the Ring   138.0496
## 10                                The Shawshank Redemption   136.7477
```

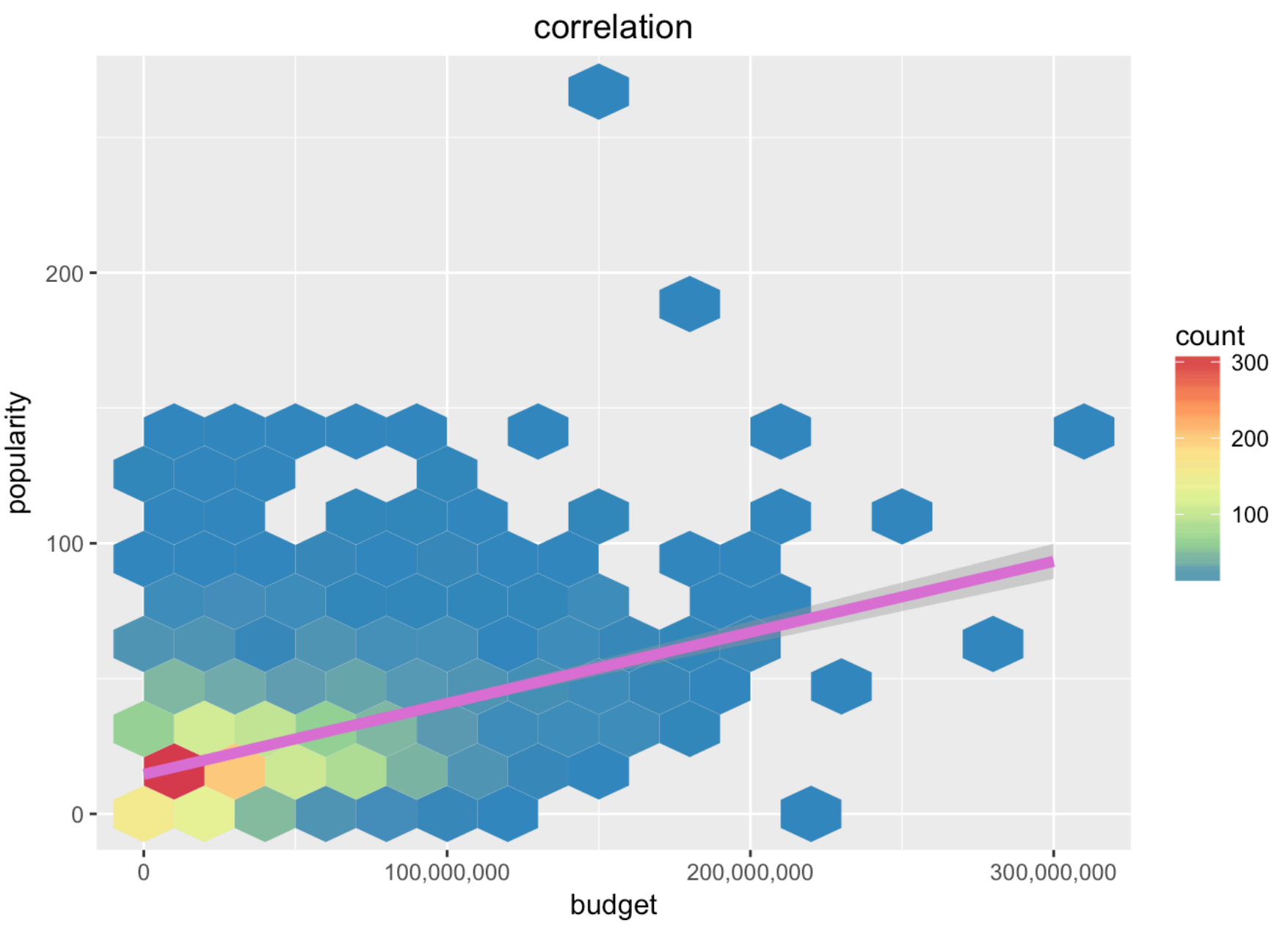
The table shows top 10 most popular movies. There are 3 movies in Pirates of carribbean series，followed by The Dark Knight and Fight club.

```
profit_movie<-movie %>%
  select(title,revenue,budget) %>%
  mutate(profit = revenue - budget) %>%
  arrange(desc(profit)) %>%
  head(10)
profit_movie
```

```
##                                                 title    revenue    budget
## 1                                             Titanic 1845034188 200000000
## 2  The Lord of the Rings: The Return of the King 1118888979  94000000
## 3     Pirates of the Caribbean: Dead Man's Chest 1065659812 200000000
## 4                                    Jurassic Park  920100000  63000000
## 5          The Lord of the Rings: The Two Towers  926287400  79000000
## 6                                     Finding Nemo  940335536  94000000
## 7                            Alice in Wonderland 1025491110 200000000
## 8                                The Dark Knight 1004558444 185000000
## 9                                The Jungle Book  966550600 175000000
## 10     Harry Potter and the Order of the Phoenix  938212738 150000000
##        profit
## 1  1645034188
## 2  1024888979
## 3   865659812
## 4   857100000
## 5   847287400
## 6   846335536
## 7   825491110
## 8   819558444
## 9   791550600
## 10  788212738
```

Also we find top 10 movies with highest profit. Titanic has the highest profit which is over 1.6 billion, followed by The Lord of the Rings: The Return of the King with over 1 billion. Compared with movies with highest popularity, The Dark Knight is the only one in top 10 popularity and profit movie.

```
movie %>%
  select(budget,popularity) %>%
  distinct() %>%
  ggplot(aes(budget,popularity)) +
  stat_bin_hex(bins=15) +
  scale_fill_distiller(palette="Spectral") +
  stat_smooth(method="lm",color="orchid",size=2) +
  scale_x_continuous(labels=scales::comma) +
  ggtitle("correlation") +
  theme(plot.title=element_text(hjust=0.5))
```



We consider popularity and profit before, now we are curious about the relation between popularity and budget. Does high budget lead to high popularity? We got the correlation plot which showing that the movies in higher budget are more likely to be more popular.

Conclusion

In summary, we combined dataset from IMDb and Rotten Tomatoes, cleaned redundant information and extracted neccessary variables we are interested in. We compared how rating distributes in 2 websites, and analyzed the number of movies by year and average rating by month. We find that from 1990 the number of movie had an upward trend then fell after 2006. December releases have highest ratings and September has maximum releases.

We find more valuable things in some variables. For genres, a large proportion of movies is belong to drama, comedy and thriller. Although documentary, western and history have less movies, their ratings are the highest of all. For producing companies, Warner Bros is the first one, followed by Universe Pictures and Paramont Pictures. In keywords, murder, violence and los angeles appear in highest frequency. Of all countries, USA is the country which produced most movies, however the rating is not the highest one. The spanish and italian movies have higher votes.

We also interested in relationship between popularity, profit and budget. We have tables showing the top 10 most popular movies and top 10 movies with highest profit, where The Dark Knight is the only one appearing in both tables. We plot the correlation between budget and popularity and find there is a positive correlation in these 2 variables. Higher budget are more likely to lead to high popularity.