Visualization project

ashtom

22 Jan 2016

Through Google Docs and from 'BoxOfficeMojo.com', I have first obtained a table 'adj_gross.csv' for the top 200 selling tickets movies, adjusted to the inflation in the US. The dataset needs some cleaning work to get rid of some special characters, which for instance make budget/profits appear as factors rather than numeric value.

```
adj_gross = read.csv('/Users/boulenge/Desktop/Projects/Project 1/AdjustedGross.csv')
library(dplyr); adj_gross = tbl_df(adj_gross)
adj_gross = dplyr::rename(adj_gross, Title = Title..click.to.view.,
                          Adjusted.Gross = X.Adjusted.Gross.,
                        Year = Year.)
adj_gross$Title = gsub("^[\\*]", '', adj_gross$Title)
adj_gross$Title = gsub("[\\*]$", '', adj_gross$Title)
adj_gross$Adjusted.Gross = gsub("^[\\*]", '', adj_gross$Adjusted.Gross)
adj_gross$Adjusted.Gross = gsub("[\\*]$", '', adj_gross$Adjusted.Gross)
adj_gross$Adjusted.Gross = gsub('[\\$,]', '', adj_gross$Adjusted.Gross)
adj_gross$Unadjusted.Gross = gsub('[\\$,]', '', adj_gross$Unadjusted.Gross)
adj_gross$Year = gsub('\\^$', '', adj_gross$Year)
adj_gross$Title[adj_gross$Title == 'A Star Is Born (1976)'] = 'A Star Is Born'
adj_gross$Title[adj_gross$Title == 'Alice in Wonderland (2010)'] = 'Alice in Wonderland'
adj_gross$Title[adj_gross$Title == 'Kramer Vs. Kramer'] = 'Kramer vs Kramer'
adj_gross$Title[adj_gross$Title == 'Cleopatra (1963)'] = 'Cleopatra'
adj_gross$Title[adj_gross$Title == "Marvel's The Avengers"] = 'The Avengers'
```

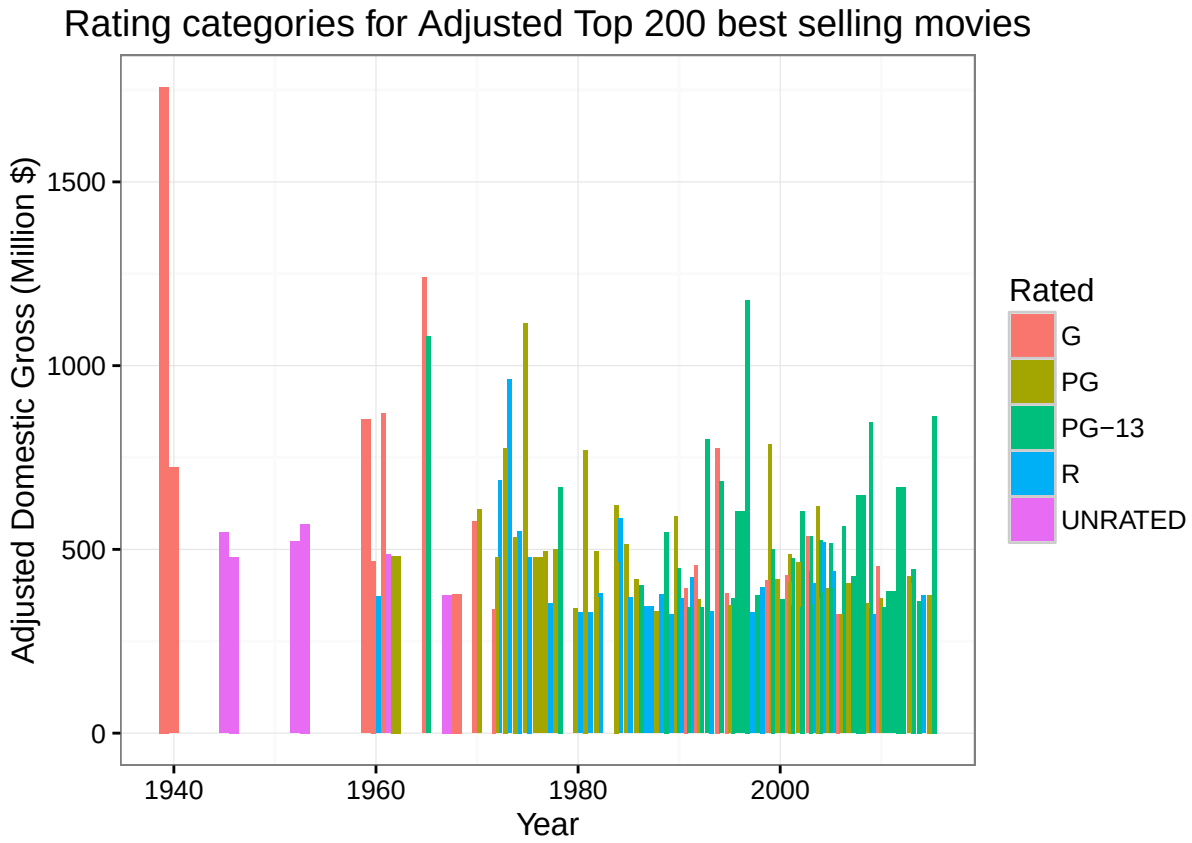
In order to complete the table, I then sift through the IMDB database with the OMDB API, using the list of movie titles from adj_gross. For that purpose I need to use the library 'omdbapi' in R.

```
library(omdbapi)
movie_adj = data.frame()
for (str in adj_gross$Title) {
  movie_adj = rbind(movie_adj, find_by_title(as.character(str)))
}
movie_adj = inner_join(movie_adj, adj_gross[, -6], by = 'Title')
movie_adj$Year = as.numeric(movie_adj$Year)
movie_adj$Adjusted.Gross = as.numeric(movie_adj$Adjusted.Gross)
movie_adj$Unadjusted.Gross = as.numeric(movie_adj$Unadjusted.Gross)
movie_adj$Year[movie_adj$Title == "M.A.S.H."] = 1970
movie_adj$Adjusted.Gross = movie_adj$Adjusted.Gross/10^6
movie_adj$Unadjusted.Gross = movie_adj$Unadjusted.Gross/10^6
write.csv(movie_adj, file = '/Users/boulenge/movie_adj.csv')

rate_dist = tbl_df(movie_adj)
rate_dist$Rated[rate_dist$Rated == 'NOT RATED'] = 'UNRATED'
rate_dist = dplyr::filter(rate_dist, Rated == 'G' | Rated == 'PG' |
                                Rated == 'PG-13' | Rated == 'R' |
                                Rated == 'UNRATED')
```

Evolution of Ratings

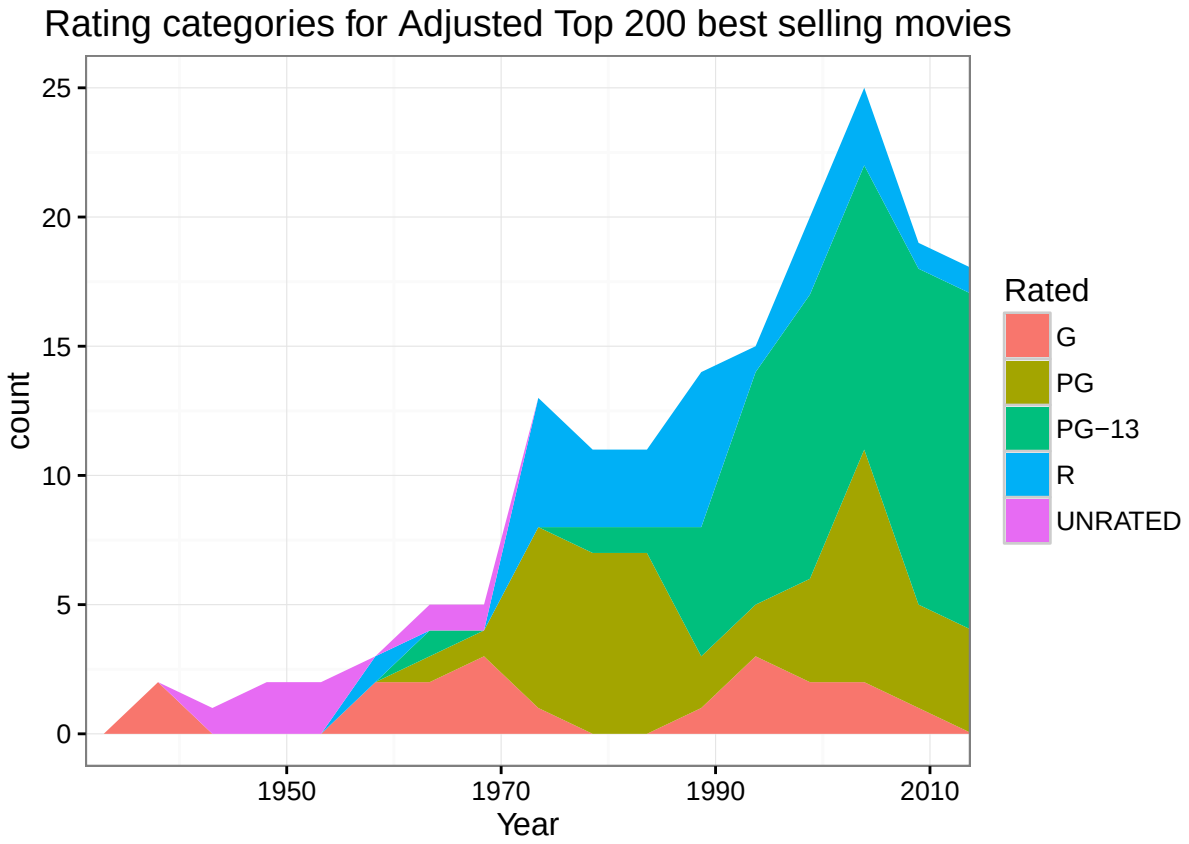
First of all, let us look at how the Top 200 best selling movies (adjusted to inflation) where rated over the years

```
library(ggplot2)
ggplot(rate_dist, aes(x = Year, y = Adjusted.Gross, fill = Rated)) +
  theme_bw() +
  geom_bar(stat = 'identity', position = 'dodge') +
  ggtitle('Rating categories for Adjusted Top 200 best selling movies') +
  ylab('Adjusted Domestic Gross (Million $)')
```



When adjusted to inflation, the profits of best selling movies seem to follow a downward trend, but more importantly for now, a change in the rating distribution appears to be at hand. Forgetting about profits, we see it more clearly by counting movies by rating types.

```
ggplot(rate_dist, aes(Year)) +
  geom_area(aes(fill = Rated, group = Rated), stat = 'bin', bins = 15) +
  theme_bw() +
  ggtitle('Rating categories for Adjusted Top 200 best selling movies') +
  coord_cartesian(xlim = c(1935, 2010))
```

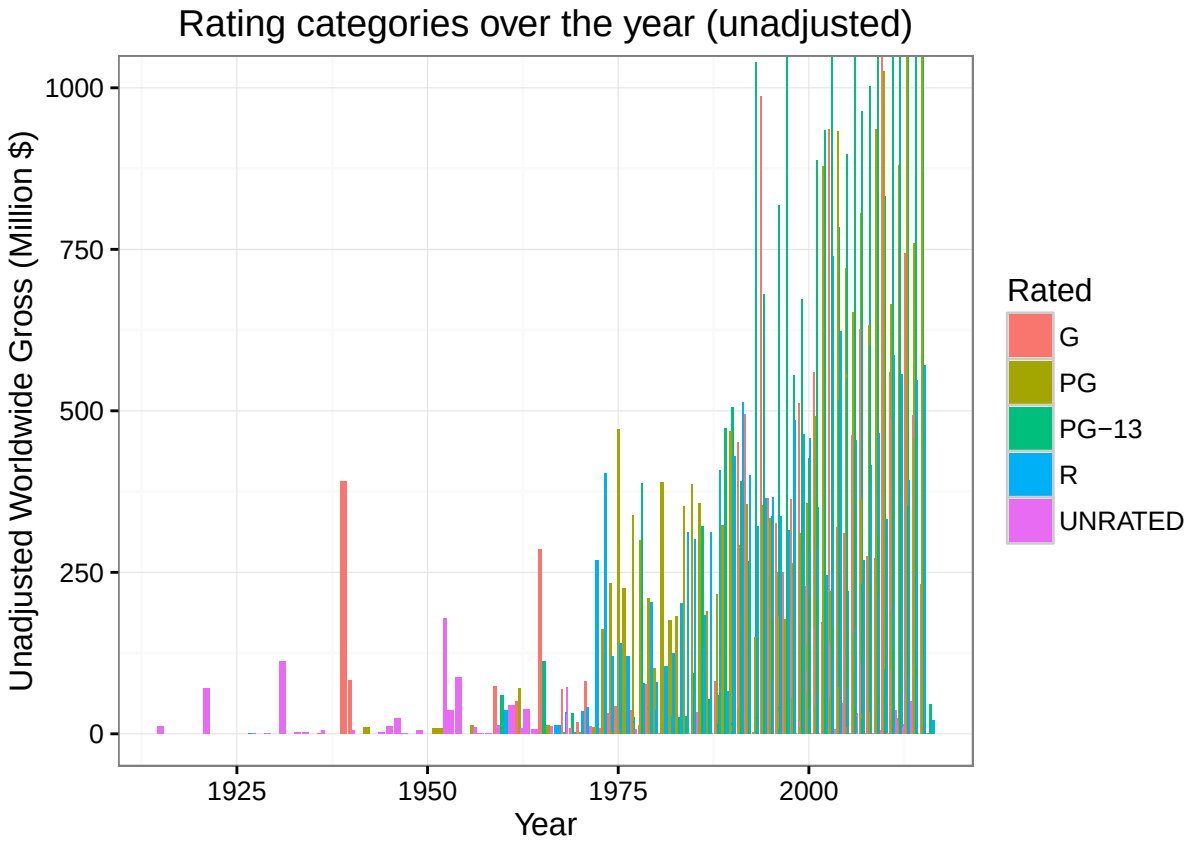


Notice here we look only at 5 different types of Rating that seem most meaningful for the US film industry. On the long run, G rated movies appear periodically over short periods of times among the best selling list. At the beginning of the 1970s, as the number of movies increase, we see an explosion in the number of PG-13 and R rated movies, with a peak of PG rated movies in the 2000s. This suggests a big change in movie goers habits, with more familities and teenagers enjoying going to the movies. The width of the R rated movie also seem to be going through some change, as it appears thicker in the 1980s, is reduced drastically through the 1990s, only to remain somewhat constant until now.

Following the same method over a much bigger list of movies, we construct a large dataset of 4769 movies, based on a large dataset of financial figures extracted from 'the-numbers.com'. Although this time, the profits are not adjusted to inflation, and therefore less relevant when comparing movies as their distance in time grows.

```
library(dplyr)
movie_finance = read.csv('/Users/boulenge/Desktop/Projects/Project 1/movie_finance.csv')
movie_finance = na.omit(movie_finance)
movie_finance = rename(movie_finance, Title = Movie)
##movie_finance = movie_finance[-is.na(movie_finance$X)]
movie_finance$Title = gsub("^[\\*]", '', movie_finance$Title)
movie_finance$Title = gsub("[\\*]$", '', movie_finance$Title)

## use the movie_finance$Title to sift through IMDB data via OMDB API, a very long process
##movie = data.frame()
##for (str in movie_finance$Title) {
##  movie = rbind(movie, find_by_title(as.character(str)))
##}
##movie = inner_join(movie, movie_finance, by = 'Title')
##write.csv(movie, file = '/Users/boulenge/Desktop/Projects/Project 1/movie.csv')
```

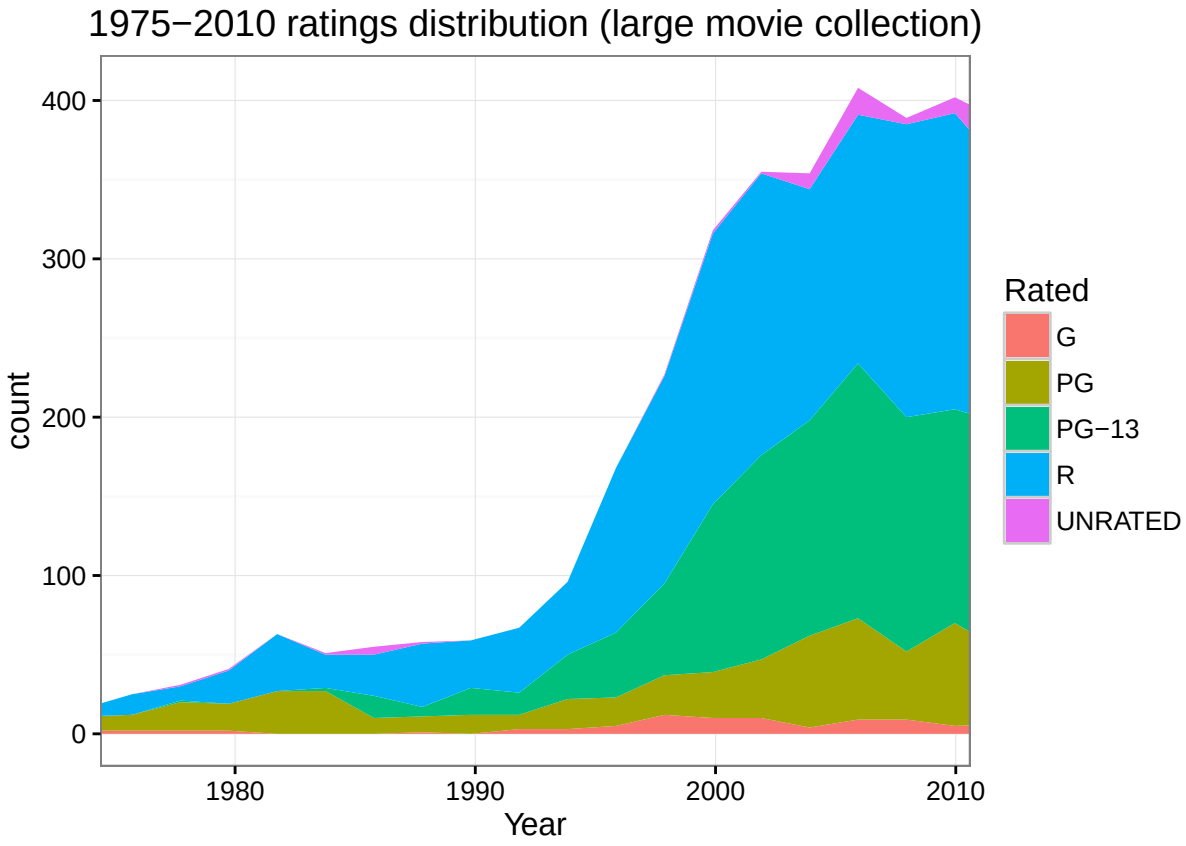
```
movie = read.csv('/Users/boulenge/Desktop/Projects/Project 1/movie.csv')

movie$Worldwide.Gross = gsub('[\\$,]', '', movie$Worldwide.Gross)
movie$Production.Budget = gsub('[\\$,]', '', movie$Production.Budget)
movie$Title[movie$Title == 'A Star Is Born (1976)'] = 'A Star Is Born'
movie$Title[movie$Title == 'Alice in Wonderland (2010)'] = 'Alice in Wonderland'
movie$Title[movie$Title == 'Kramer Vs. Kramer'] = 'Kramer vs Kramer'
movie$Title[movie$Title == 'Cleopatra (1963)'] = 'Cleopatra'
movie$Title[movie$Title == "Marvel's The Avengers"] = 'The Avengers'

movie$Worldwide.Gross = as.numeric(movie$Worldwide.Gross)
movie$Production.Budget = as.numeric(movie$Production.Budget)
movie$Year[movie$Title == "M.A.S.H."] = 1970
movie$Worldwide.Gross = movie$Worldwide.Gross/10^6
movie$Production.Budget = movie$Production.Budget/10^6
write.csv(movie, file = '/Users/boulenge/Desktop/Projects/Project 1/movie_adj.csv')

rate_world = tbl_df(movie)
rate_world$Rated[rate_world$Rated == 'NOT RATED'] = 'UNRATED'
rate_world = dplyr::filter(rate_world, Rated == 'G' | Rated == 'PG' |
                                       Rated == 'PG-13' | Rated == 'R' |
                                       Rated == 'UNRATED')
rate_world$Year = as.numeric(as.character(rate_world$Year))

ggplot(rate_world, aes(x = Year, y = Worldwide.Gross, fill = Rated)) +
  geom_bar(stat = 'identity', position = 'dodge') +
  theme_bw() +
  ggtitle('Rating categories over the year (unadjusted)') +
  ylab('Unadjusted Worldwide Gross (Million $)') +
  coord_cartesian(ylim = c(0, 1000))
```



Not very surprinsingly, the unadjusted worlwide profits follow an upward trend, as the G, PG and PG-13 rated movies dominate the market in terms of worlwide sales over the last three decades.

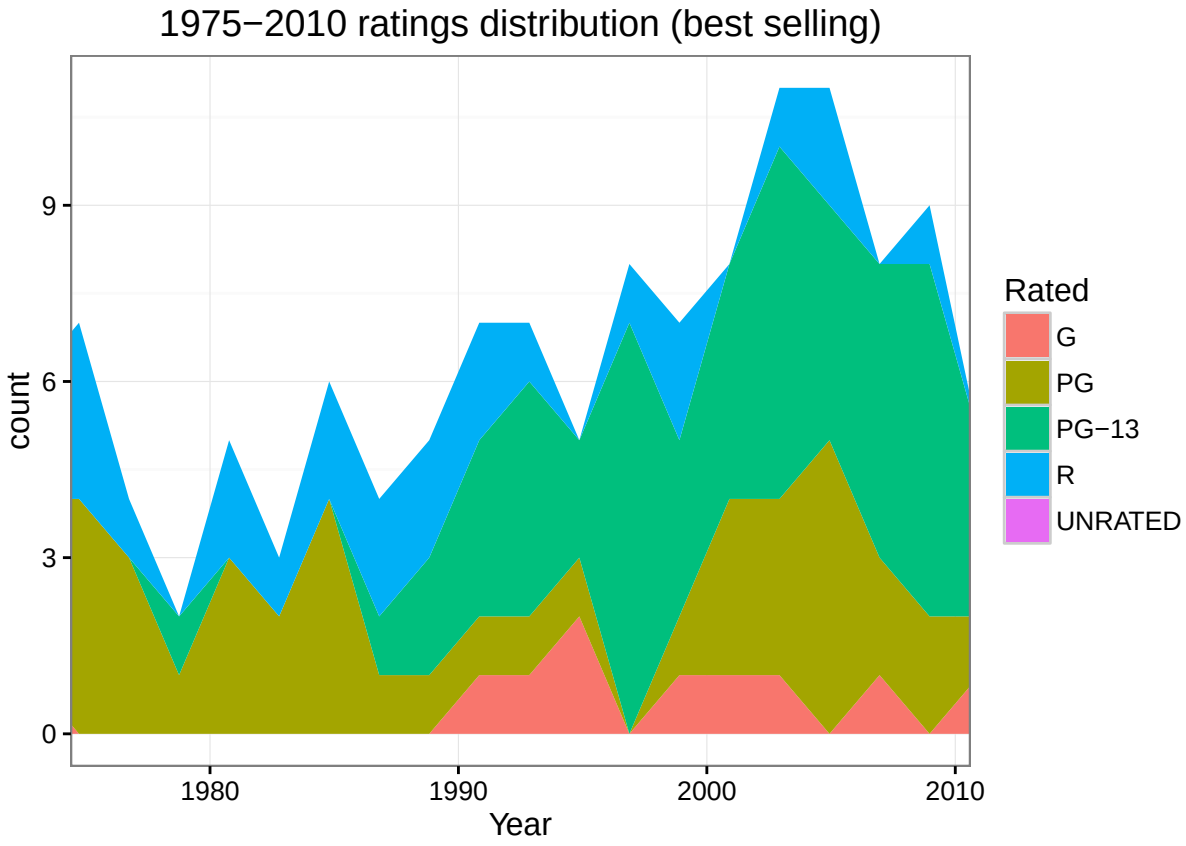
Focusing now on counting movies by Rating, for our large movie collection we see at the beginning of the 1990s a huge increase in numbers of both PG-13 and R rated movies.

```
ggplot(rate_world, aes(Year)) +
  geom_area(aes(fill = Rated, group = Rated), stat = 'bin', bins = 1000) +
  theme_bw() +
  scale_x_discrete(breaks = c(1980,1990,2000, 2010)) +
  ggtitle('1975-2010 ratings distribution (large movie collection)') +
  coord_cartesian(xlim = c(1975, 2010))
```



However, counting only among the top 200 best selling movies, we see indeed an explosion in the number of PG-13 rated movies at the beginning of the 1990s, but the number of R rated movies now seem to follow a steady pace, even decreasing at times. This suggests that while R rated movies are becoming very common, their financial success is far less assured than PG-13 movies, which again appear to dominate the market in terms of sales.

```
ggplot(rate_dist, aes(Year)) +
  geom_area(aes(fill = Rated, group = Rated), stat = 'bin', bins = 1000) +
  theme_bw() +
  scale_x_discrete(breaks = c(1980,1990,2000, 2010)) +
  ggtitle('1975-2010 ratings distribution (best selling)') +
  coord_cartesian(xlim = c(1975, 2010))
```



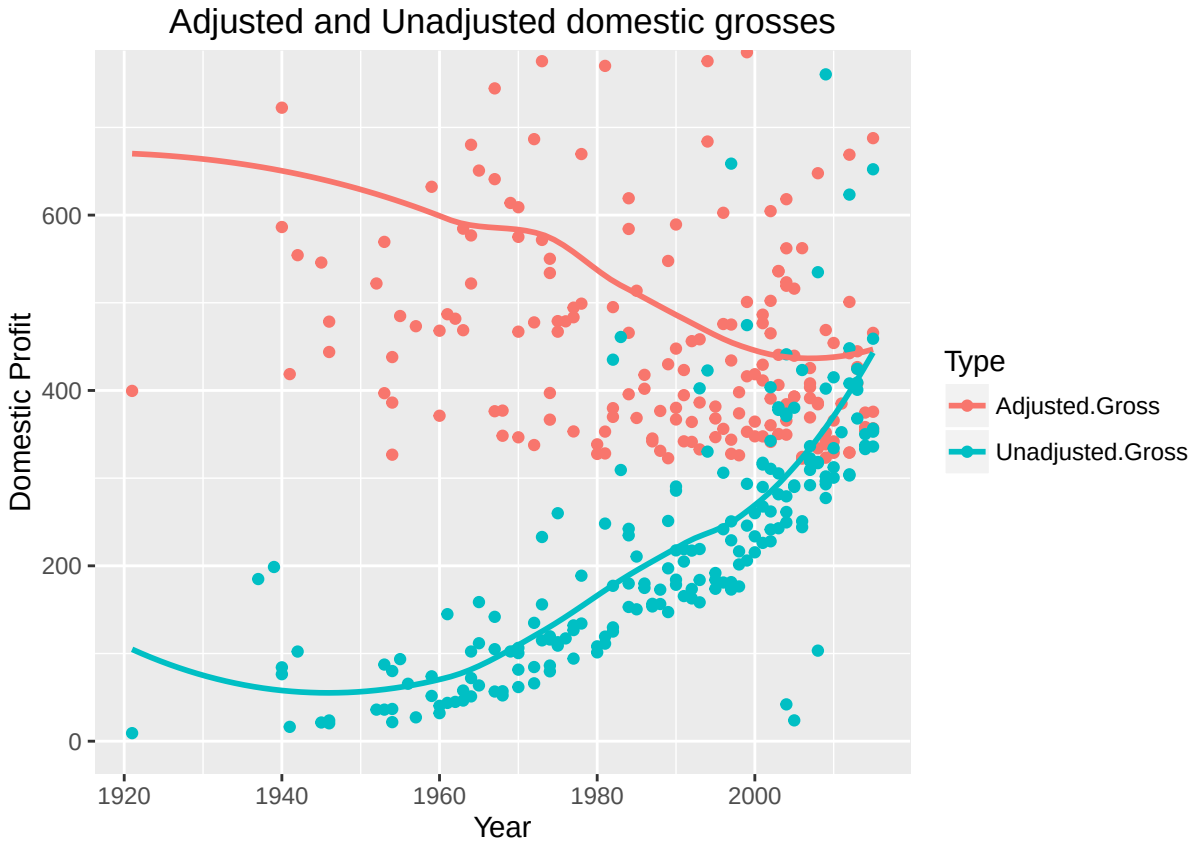
All this suggests a shift in mentalities, with

When looking at best selling movies, over the past three decades, the Rating categories seem to shift from a large public (G and PG), to mostly family and PG-13 movies. We also notice what could be seen as a change in mentalities, as in the 1980s, most movies are rated either R or PG, so either large or restricted audiences. In the 1990s however, we see an explosion in the number of PG-13 movies. R rated movies however increase in numbers but seem less financially successful. This suggests what we see is a new niche of gore movies.

Evolution of Domestic profits for the Top 200 best selling movies

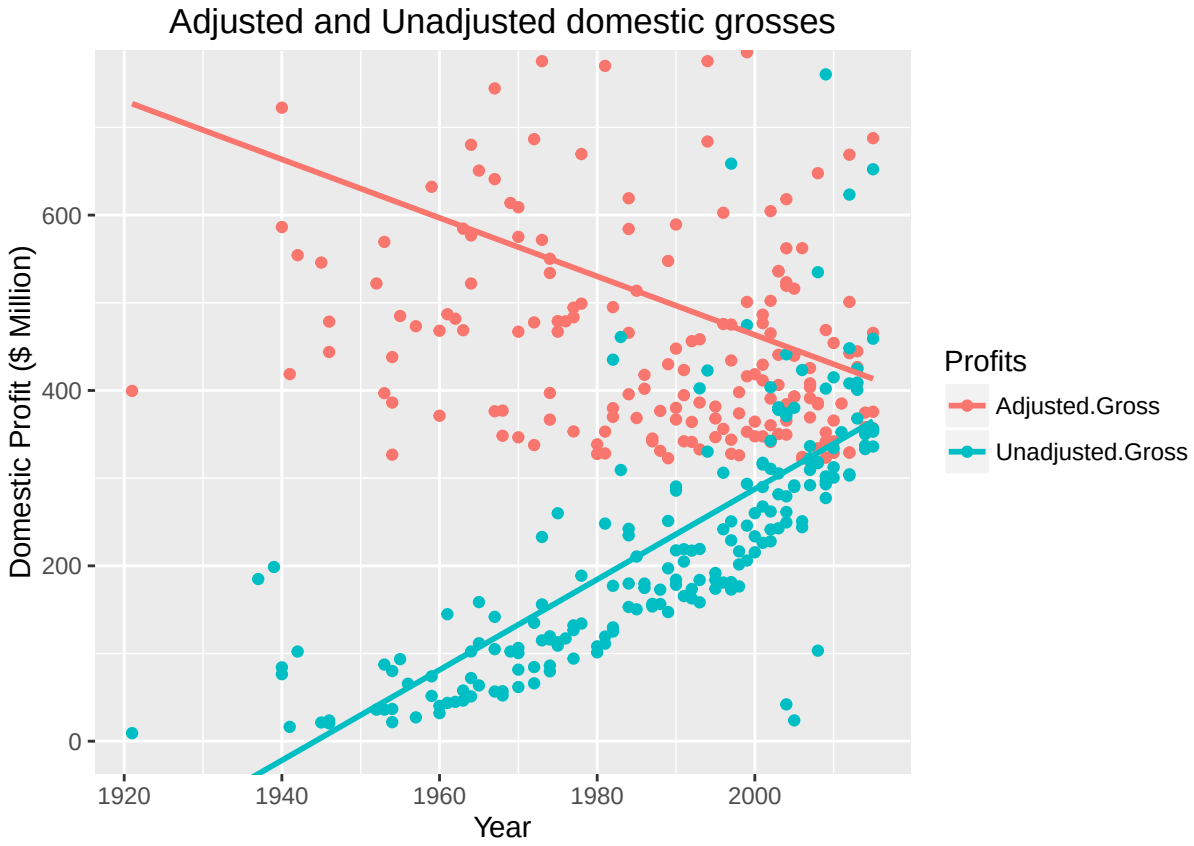
Based on a table from 'BoxOfficeMojo.com', we now look again at the Top 200 best selling movies, and compare their adjusted and unadjusted US ticket sales.

```
library(reshape)
temp = tbl_df(movie_adj) %>% select(Year, Rank, Adjusted.Gross, Unadjusted.Gross)
temp = as.data.frame(temp)
temp = melt(temp, id = c('Year', 'Rank'))
ggplot(temp, aes(x = Year, y = value, colour = variable)) + geom_point() +
  geom_smooth(se = FALSE) +
  ggtitle('Adjusted and Unadjusted domestic grosses') +
  ylab('Domestic Profit') +
  labs(colour = 'Type') +
  coord_cartesian(ylim = c(0, 750))
```



Although the unadjusted gross seems to be increasing, the adjusted gross clearly seems to follow a downward trend over the years. The upward slope of the unadjusted gross also seems to increase almost 'quadratically', definitely faster than the downward adjusted gross goes down. Could that be a hint for an acceleration of the inflation in the US over the last 50 years? Other variables such as the population size would be needed to confirm. Using a linear regression we may compare the slopes

```
ggplot(temp, aes(x = Year, y = value, colour = variable)) + geom_point() +
  geom_smooth(se = FALSE, method = 'lm') +
  ggtitle('Adjusted and Unadjusted domestic grosses') +
  ylab('Domestic Profit ($ Million)') +
  coord_cartesian(ylim = c(0, 750)) +
  labs(colour = 'Profits')
```




```
library(data.table)
```

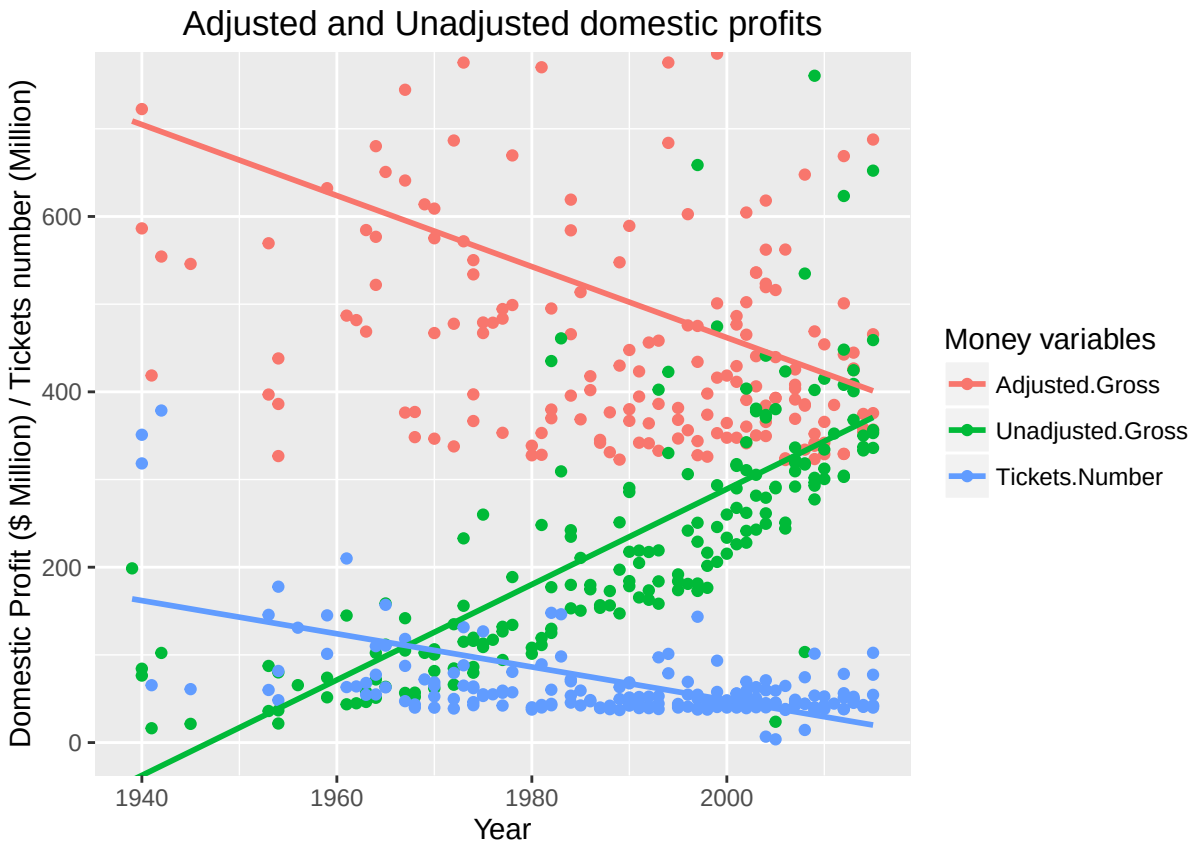
```
##
## Attaching package: 'data.table'

## The following object is masked from 'package:reshape':
##
##     melt

## The following objects are masked from 'package:dplyr':
##
##     between, last
```

```
dat = data.table(x = temp$Year, y = temp$value, grp = temp$variable)
dat[, list(intercept = coef(lm(y ~ x))[1], slope = coef(lm(y ~ x))[2]), by = grp]
```

```
##                 grp  intercept     slope
## 1:   Adjusted.Gross   7138.722 -3.337679
## 2: Unadjusted.Gross -10027.348  5.157490
```

Indeed, the slope for the adjusted gross is roughly -3.3 $ Million per year, while the unadjusted slope goes up by roughly 5.5 $ Million per year. This suggests an acceleration of inflation. Adding the rescaled average of tickets price, we find a hint that the inflation was indeed much faster than the decrease of the attendance for US audience, hence the appearance that tickets sales increase, when they actually decrease when tickets prices are adjusted to inflation.

```
tickets = read.csv('/Users/boulenge/Desktop/Projects/Project 1/tickets_price.csv')
tickets = tbl_df(tickets)
tickets = dplyr::rename(tickets, Year = X.Year., price = X.Avg..Price.)
tickets$price = gsub("^[\\$]", '', tickets$price)
tickets$price[1] = 8.70
tickets$price = as.numeric(tickets$price)

temp2 = tbl_df(movie_adj) %>% dplyr::select(., Year, Adjusted.Gross, Unadjusted.Gross)
temp2 = as.data.frame(temp2)

temp3 = inner_join(temp2, tickets, by = 'Year') %>%
  dplyr::mutate(., Tickets.Number = Unadjusted.Gross/price) %>%
  select(., -price) %>%
  melt(., id = 'Year')

ggplot(temp3, aes(x = Year, y = value, colour = variable)) + geom_point() +
  geom_smooth(se = FALSE, method = 'lm') +
  ggtitle('Adjusted and Unadjusted domestic profits') +
  ylab('Domestic Profit ($ Million) / Tickets number (Million)') +
  coord_cartesian(ylim = c(0, 750)) +
  labs(colour = 'Money variables')
```




```
dat3 = data.table(x = temp3$Year, y = temp3$value, grp = temp3$variable)
dat3[, list(intercept = coef(lm(y ~ x))[1], slope = coef(lm(y ~ x))[2]), by = grp]
```

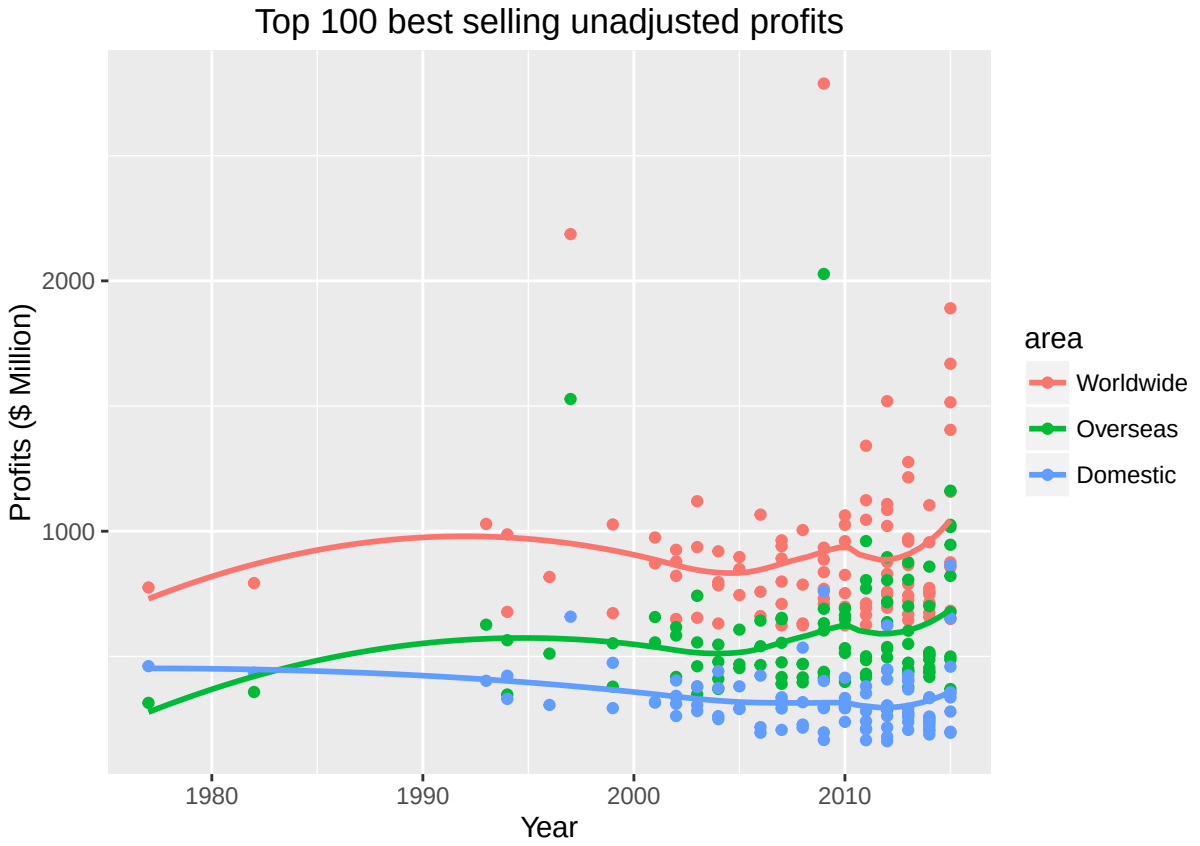
```
FALSE                 grp  intercept     slope
FALSE 1:   Adjusted.Gross   8560.433 -4.049299
FALSE 2: Unadjusted.Gross -10599.086  5.444139
FALSE 3:   Tickets.Number   3833.738 -1.892670
```

Using a linear regression, we do see that when adjusted to inflation, Tickets sales do decrease by roughly 1.9 Million a year.

Comparison of profits for the Top 100 best selling movies

Looking now at worlwide data for the Top 100 best selling movies from 'BoxOfficeMojo', this time unadjusted to inflation, we see the Overseas profits seem to be responsible for the steady growth of profits in the movies industry. For this particular dataset, even not adjusted, the trend for the Domestic gross is a downward slope.

```
world_gross = read.csv('/Users/boulenge/Desktop/Projects/Project 1/worldwide.csv')
library(dplyr); world_gross = tbl_df(world_gross)
world_gross = dplyr::rename(world_gross, Year = Year., Domestic = Domestic....,
                                         Overseas = Overseas....)
world_gross$Title = gsub("^[\\*]", '', world_gross$Title)
world_gross$Title = gsub("[\\*]$", '', world_gross$Title)
world_gross$Domestic = gsub("[\\$,]", '', world_gross$Domestic)
world_gross$Domestic = gsub("[\\$,]", '', world_gross$Domestic)
world_gross$Overseas = gsub('[\\$,]', '', world_gross$Overseas)
world_gross$Overseas = gsub('[\\$,]', '', world_gross$Overseas)
world_gross$Worldwide = gsub('[\\$,]', '', world_gross$Worldwide)
world_gross$Worldwide = gsub('[\\$,]', '', world_gross$Worldwide)
world_gross$Year = gsub('\\^$', '', world_gross$Year)
world_gross$Overseas = as.numeric(world_gross$Overseas)
world_gross$Worldwide = as.numeric(world_gross$Worldwide)
world_gross$Domestic = as.numeric(world_gross$Domestic)
world_gross$Year = as.numeric(world_gross$Year)


library(reshape)
temp_w = tbl_df(world_gross) %>% select(Year, Worldwide, Overseas, Domestic)
temp_w = as.data.frame(temp_w)
temp_w = melt(temp_w, id = c('Year'))

ggplot(temp_w, aes(x = Year, y = value, colour = variable)) + geom_point() +
geom_smooth(se = FALSE) +
  ggtitle('Top 100 best selling unadjusted profits') +
  ylab('Profits ($ Million)') +
  labs(colour = 'area')
```



That may explain why the industry in Hollywood is increasingly trying to seduce foreign audiences by shooting in European and now increasingly often in Asian locations. The Chinese market is evidently very promising for the future. A strong incentisive for producers to make part of their stories depending on our Chinese friends (Transformers (Chinese officials were sooo effective..), The Martian (thank god their have boosters..)).

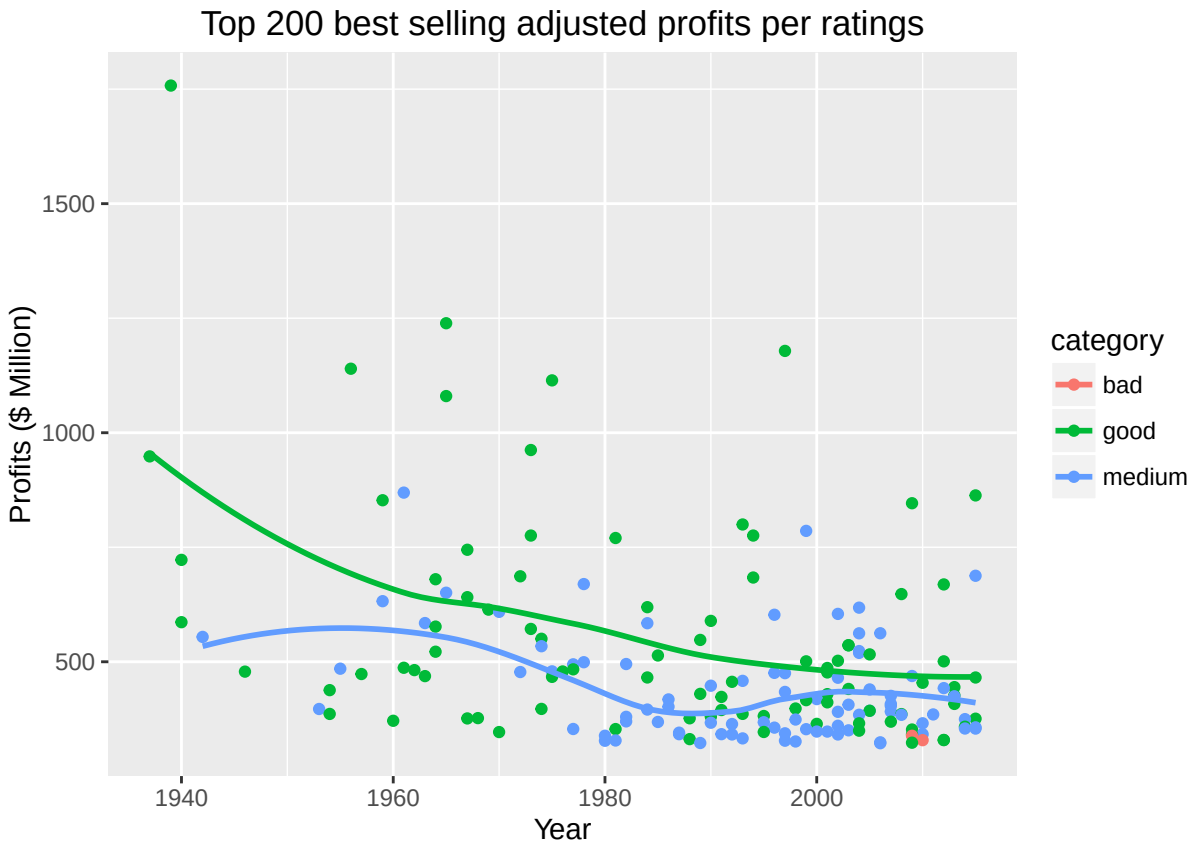
```
world_adj = read.csv('/Users/boulenge/Desktop/Projects/Project 1/world_adj.csv')
library(dplyr); world_adj = tbl_df(world_adj)
world_adj = dplyr::rename(world_adj, Year = YEAR, Title = FILM,
                                     World.Adj = ADJUSTED....2011.) %>%
  dplyr::select(., Year, Title, World.Adj)
world_adj$World.Adj = gsub("[,]", '', world_adj$World.Adj)
world_adj$World.Adj = gsub("[,]", '', world_adj$World.Adj)
world_adj$World.Adj = as.numeric(world_adj$World.Adj)/10^6
world_adj$Year = as.numeric(world_adj$Year)
```

IMDB Ratings, an insight in movies quality over the years

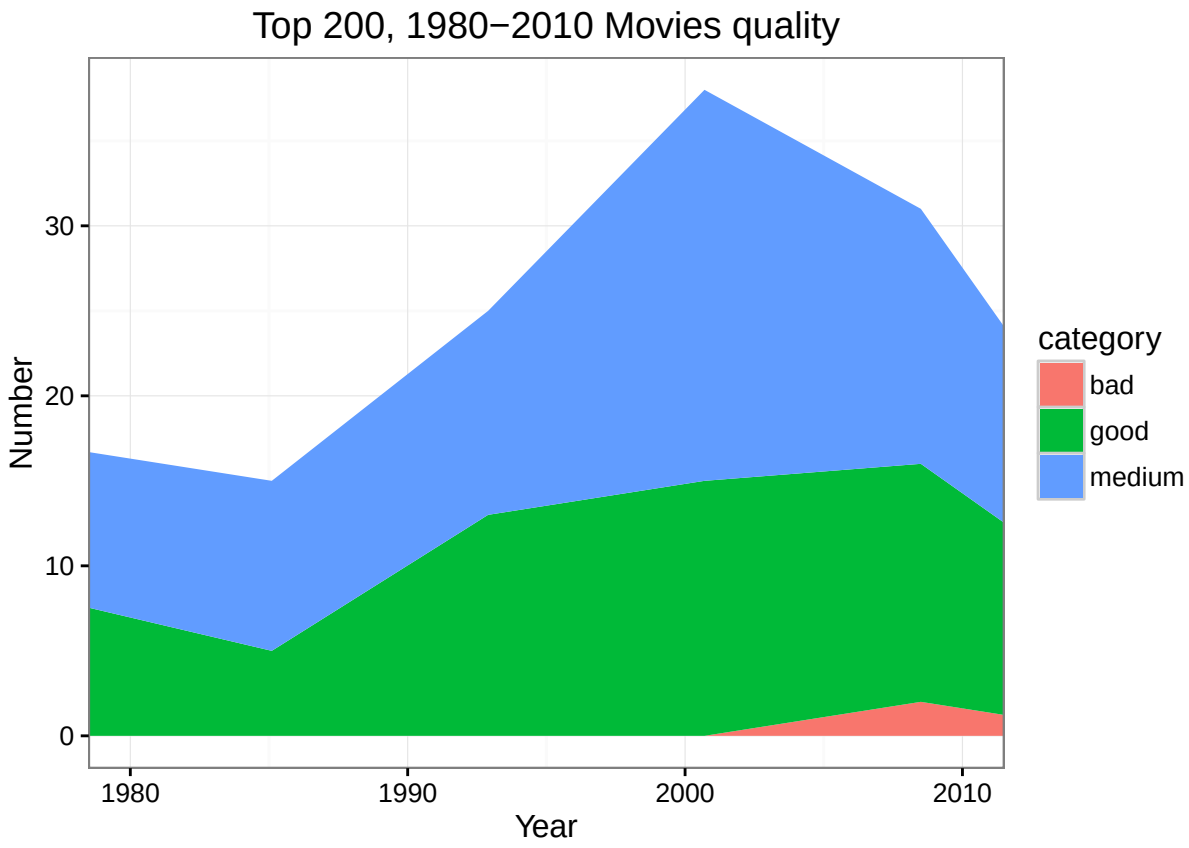
Interestingly, plotting the adjusted profits over the years and grouping by categories, the downward slope in profits for 'good movies' (grades > 7.5) seem to be faster than the one for 'medium' movies ($5 <$ grades < 7.5). As for bad movies, there aren't enough in this dataset for the top 200 best selling. Is it a hint that the only to appear in the dataset were released in the 2000s?

```
movie_adj = tbl_df(movie_adj)
movie_adj = movie_adj[!is.na(movie_adj$imdbVotes), ]
```

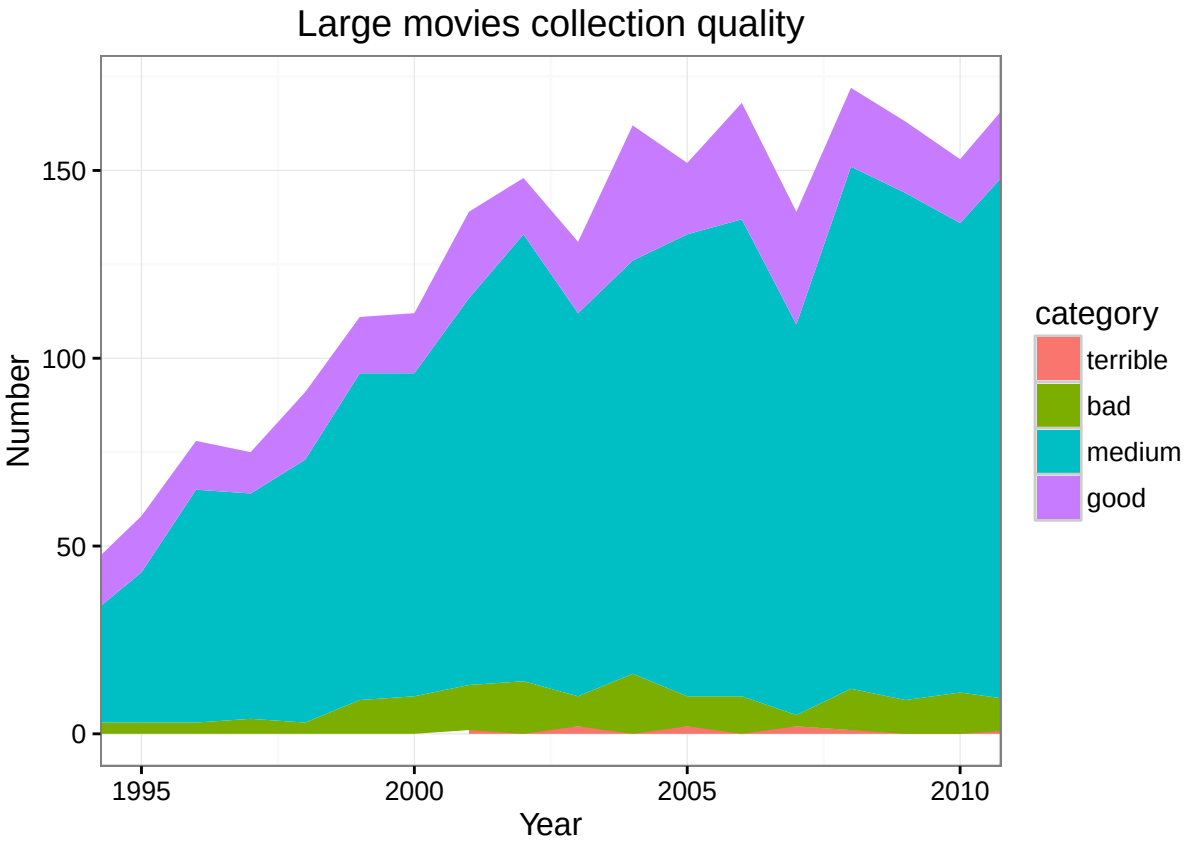
```r
fun_cat = function(x) {
  stopifnot(is.numeric(x))
  y = character(length(x))
  for (i in 1:length(x)) {
  if (x[i] < 2.5) {
    y[i] = 'terrible'
  }
  else if (x[i] < 5) {
    y[i] = 'bad'
  }
  else if (x[i] < 7.5) {
    y[i] ='medium'
  }
  else {
    y[i] = 'good'
  }
  }
  return(y)
}

movie_votes = dplyr::filter(movie_adj, imdbVotes > 0.01*max(imdbVotes))
movie_votes = mutate(movie_votes, cat = fun_cat(imdbRating))

ggplot(movie_votes, aes(x = Year, y = Adjusted.Gross, colour = cat)) + geom_point() +
  geom_smooth(data = filter(movie_votes, cat != 'bad'), se = FALSE) +
  ggtitle('Top 200 best selling adjusted profits per ratings') +
  ylab('Profits ($ Million)') +
  labs(colour = 'category')
```




```
ggplot(movie_votes, aes(Year)) +
  geom_area(aes(fill = cat, group = cat), stat = 'bin', bins = 10) +
  theme_bw() +
  ggtitle('Top 200, 1980-2010 Movies quality') +
  coord_cartesian(xlim = c(1980, 2010)) +
  ylab('Number') +
  labs(fill = 'category')
```




```
movie_votes2 = tbl_df(movie) %>%
  dplyr::select(., Title, Year, imdbVotes, imdbRating)
movie_votes2 = na.omit(movie_votes2)
movie_votes2 = movie_votes2[movie_votes2$imdbVotes > 10000, ]
movie_votes2 = mutate(movie_votes2, cat = fun_cat(imdbRating))

movie_votes2$Year = as.numeric(as.character(movie_votes2$Year))
movie_votes2$cat = factor(movie_votes2$cat,
                          levels = c('terrible', 'bad', 'medium', 'good'))

ggplot(movie_votes2, aes(Year)) +
  geom_area(aes(fill = cat, group = cat), stat = 'count') +
  theme_bw() +
  ggtitle('Large movies collection quality') +
  coord_cartesian(xlim = c(1995, 2010)) +
  ylab('Number') +
  labs(fill = 'category')
```



While the number of movies has hugely increased, the number of good ones has seen little change, while most of the increase is taken by 'medium' movies. As the proportion of 'medium' movies increase, at least as perceived by IMDB users, the public may get the impression that quality itself decreases. This might explain why people go less often to the movies (beside the tickets price increase).