An analysis of movies by ratings

Thomas Boulenger

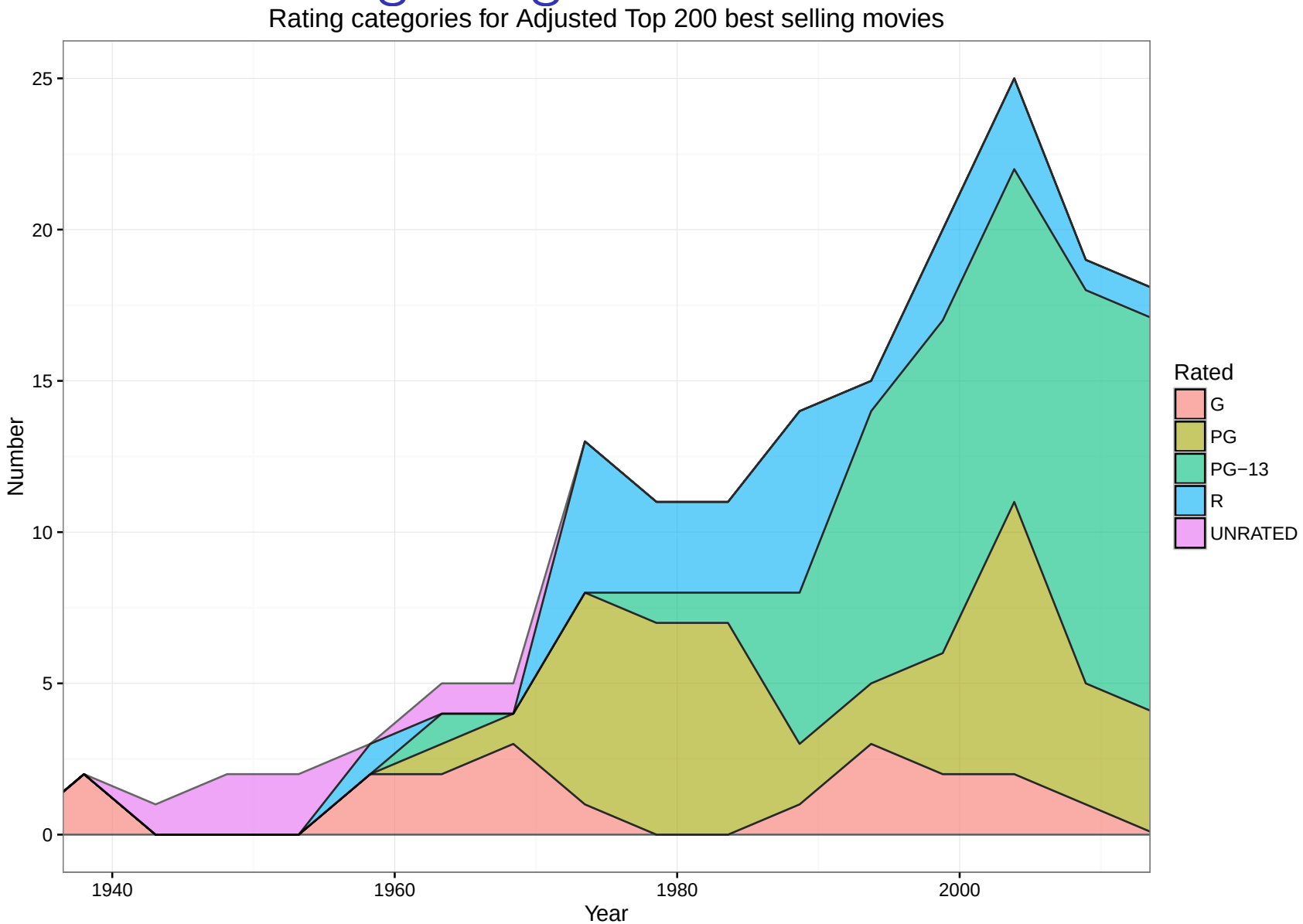
Jan 28th, 2016

Datasets

- Scrapping a table from boxofficemojo.com for a Top 200 best selling movies in the US (two 'profits' variables: unadjusted and adjusted to the inflation).
- Scrapping a tables from the-numbers.com : $\sim$ 5,000 movies with both the US and Worldwide profits, and a Top 100 best selling in unadjusted money (recent years).
- Using the references from these tables to sift through the IMDB databse via the omdbapi library, thus adding 22 more variables (Actors, Directors, Genres, Runtime, Ratings etc..)

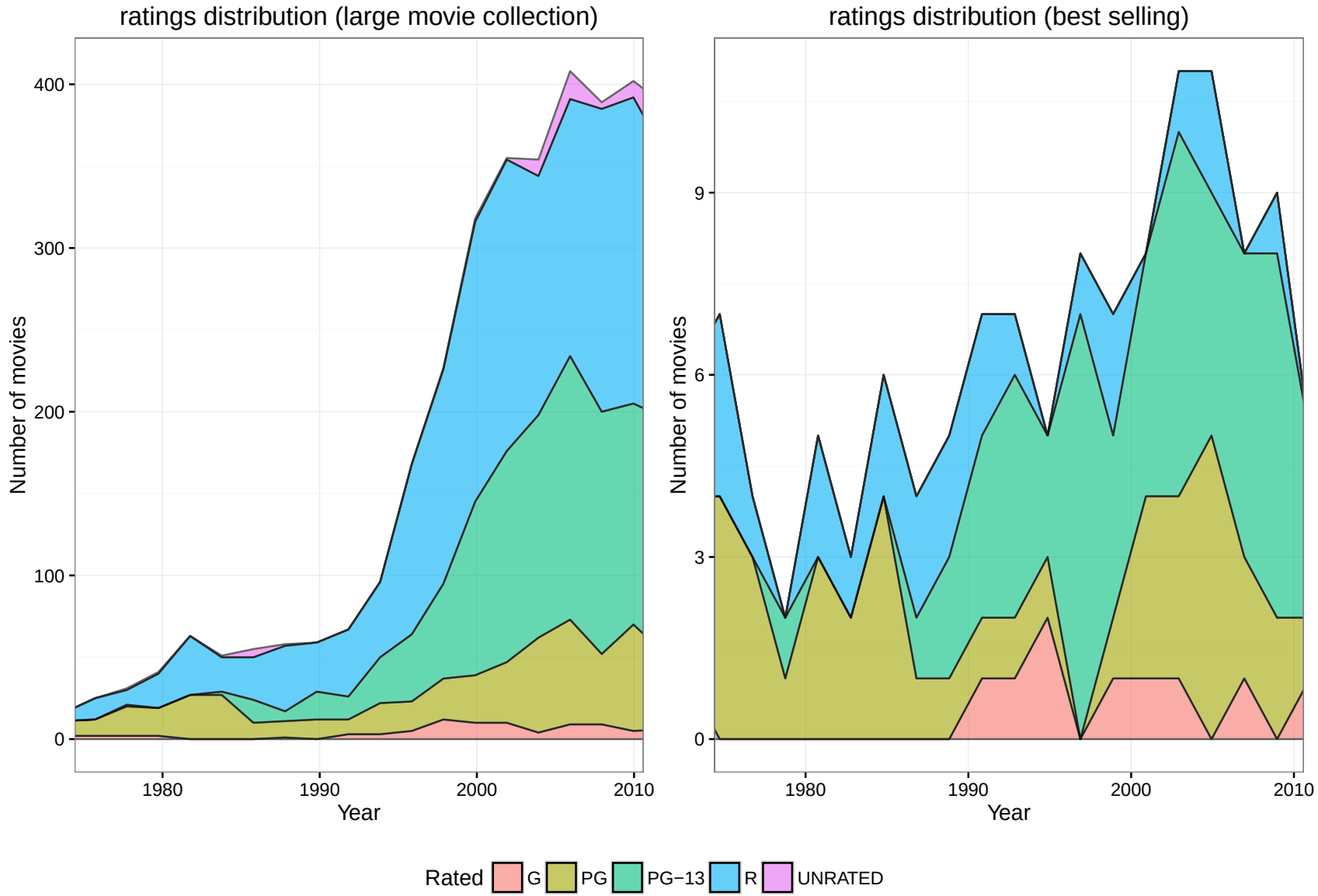
Required a lot of cleaning work after joining datasets, many mismatches (* $, etc..)

Evolution of Rating categories



Obviously, the number of movies went up, but more so among certain categories

Comparison Best selling/Large sample : 1975-2010



Big increase in R rated and PG-13, though R rated appear less financially successful

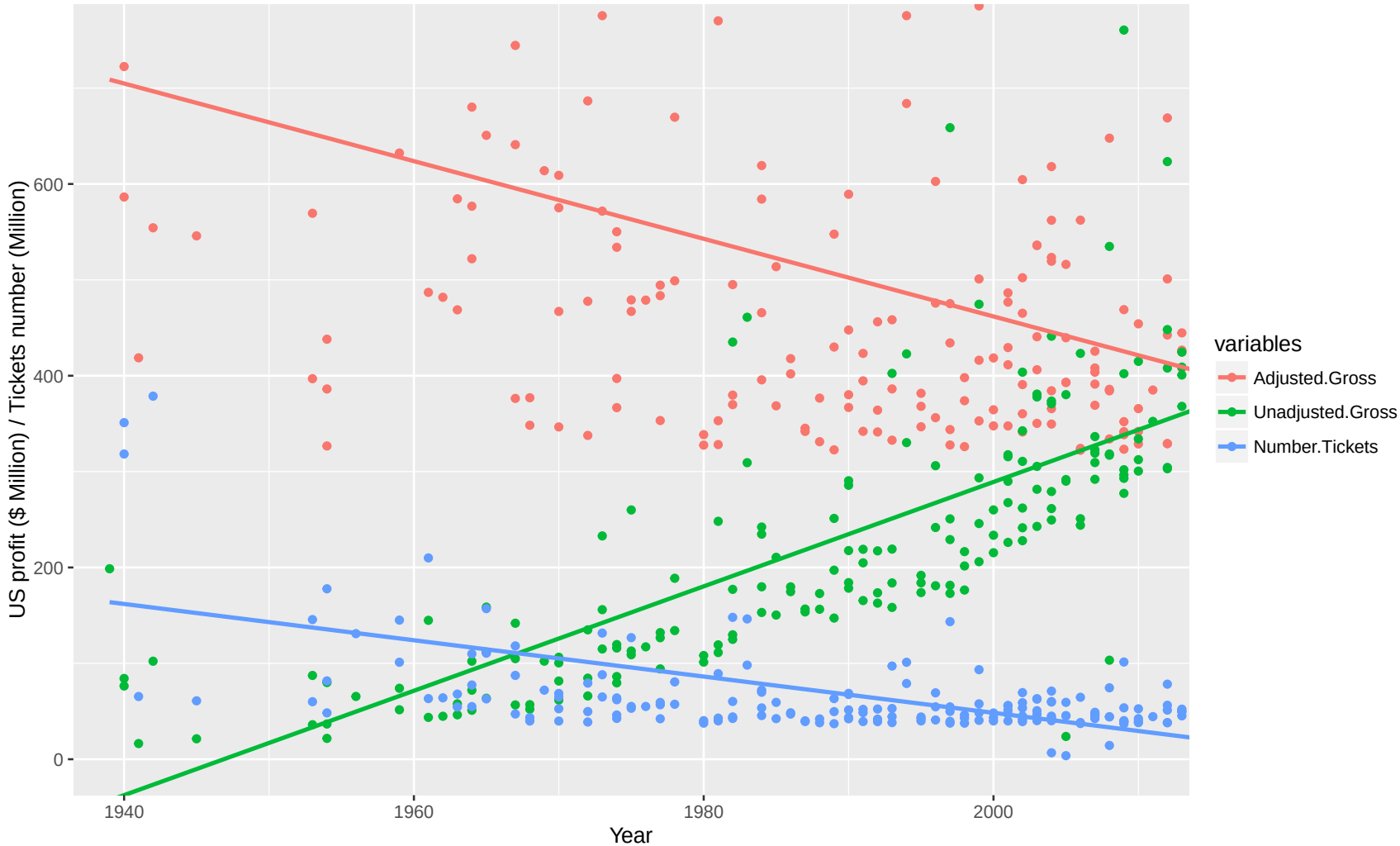
Adjusted vs Unadjusted US profits

The unadjusted variable looks almost quadratic, when the adjusted one appear 'close' to linear.

Questions raised

- might the faster increase in unadjusted profits vs adjusted ones be a sign of an accelerated inflation?
- decrease in adjusted profits $\Rightarrow$ decrease in movies attendance in the US? Based on a dataset of tickets prices per year and the unadjusted US profits, we draw the number of tickets alongside the profits

Closer look and number of tickets

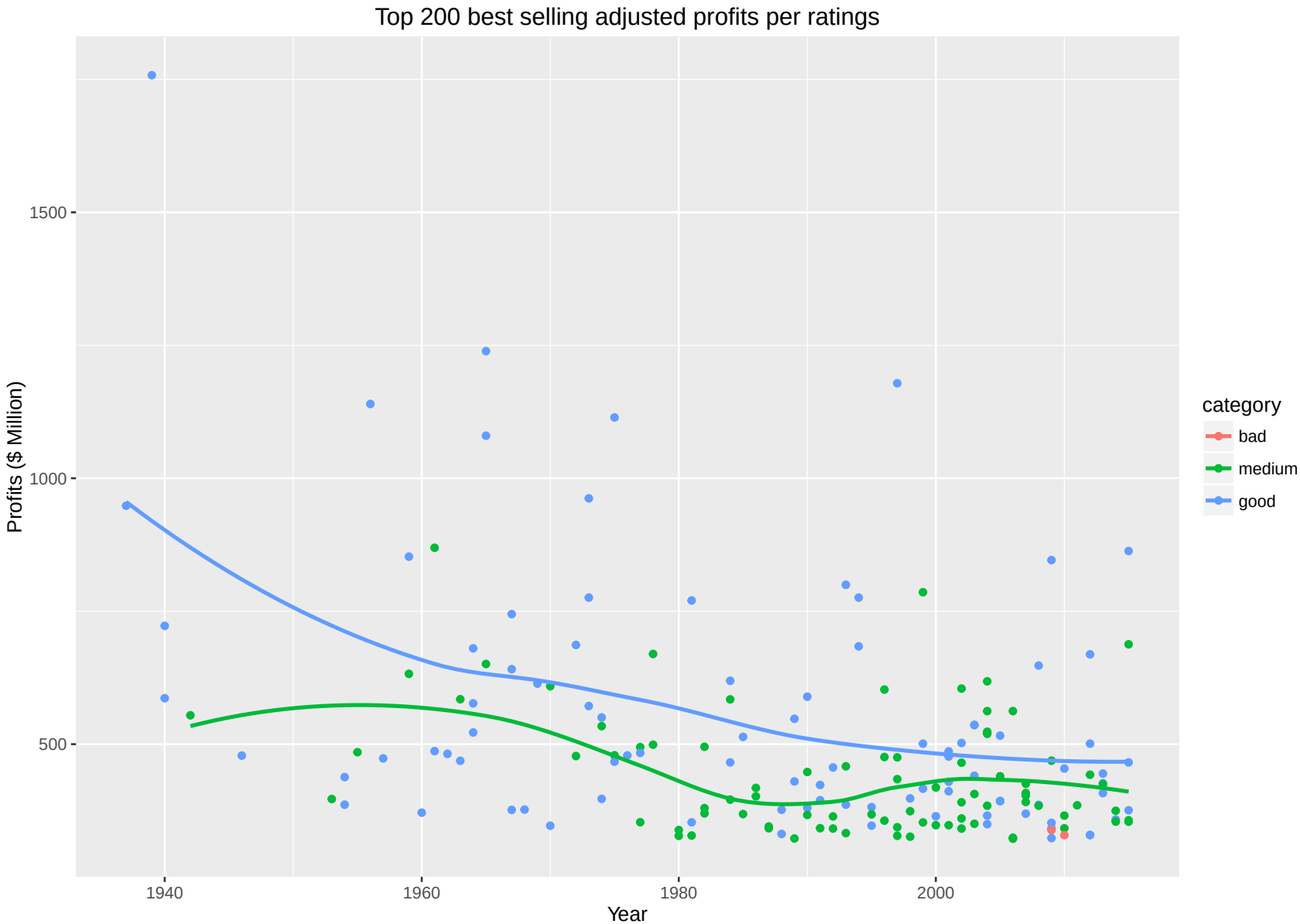


Towards present days, the 2 different inflation-free sources Tickets and Unadjusted tend to parallel downward trends: inflation grows too fast for Unadjusted to be affected as it wildly goes up

US vs Worldwide profits

Do Unadjusted money evolve the same in all areas? From another Top 100 best selling dataset

Movies quality perceived through IMDB votes



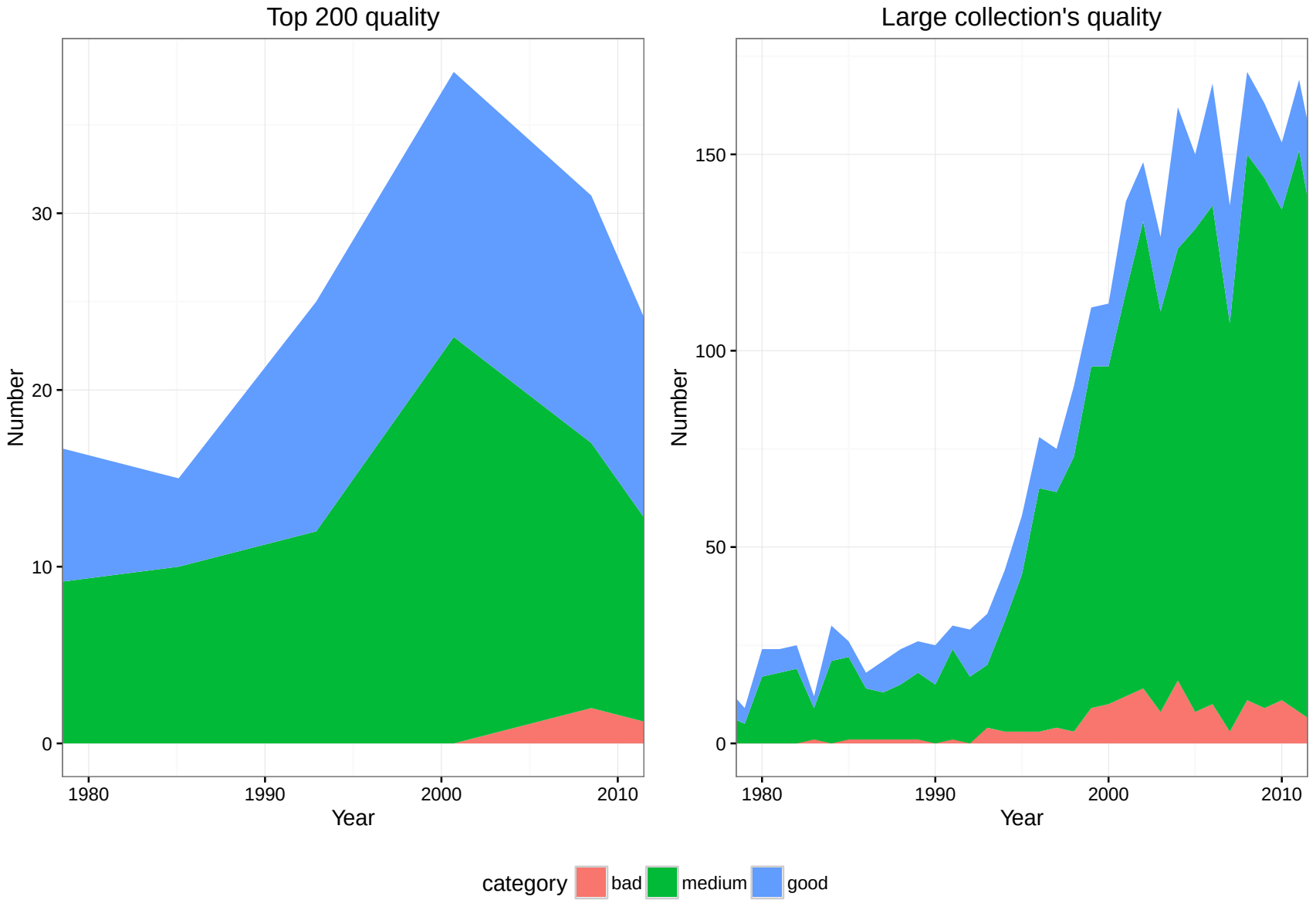
Categories: $2.5 \leq bad < 5$; $5 \leq medium < 7.5$; $7.5 \leq good < 10$

Another downward trend

- It appears that the adjusted profits for 'good' movies tend to decrease faster than the profits for 'medium' ones.
- Wussy vampires are 'bad' for business

```
## Source: local data frame [2 x 1]
##
##                            Title
##                            (chr)
## 1 The Twilight Saga: New Moon
## 2  The Twilight Saga: Eclipse
```

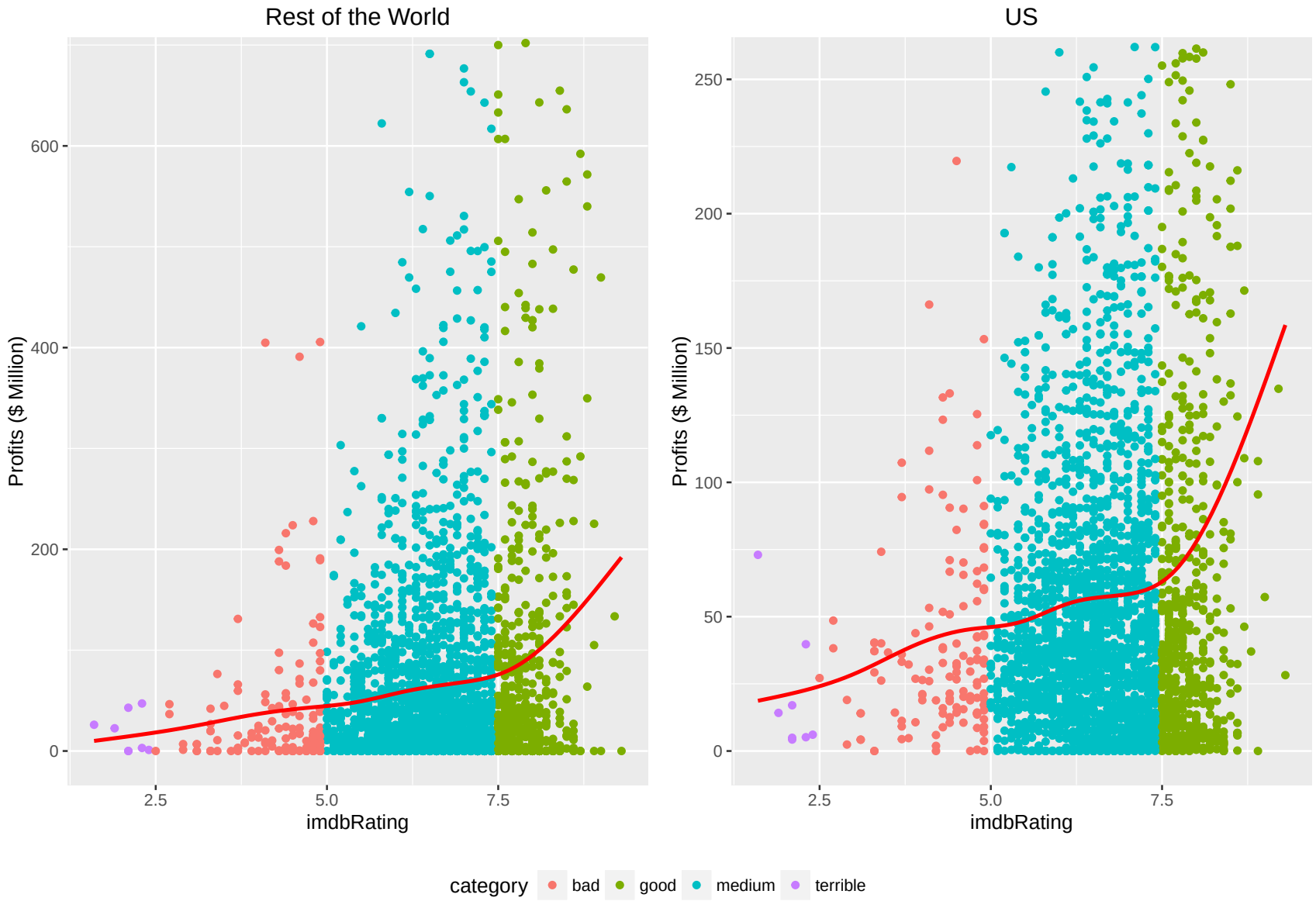
IMDB Ratings distributions



IMDB Ratings distributions: 1980-2010 trend

- Among the 'Top 200 best selling' subclass, we tend to see a rather similar explosion in the number of both medium and good
- Among a larger *a priori* random collection, the situation is very different with mainly an explosion of medium quality movies

Worldwide and US profits vs IMDB votes



How does quality affect sales

- After rescaling, it seems the better the movie the higher the profits... but even more so in the US
- χ^2 test to check that those variables are not independent

```
##
##  Pearson's Chi-squared test
##
## data:  movie.data
## X-squared = 6876100, df = 6537300, p-value < 2.2e-16
```