datamining

Mengyan Cui

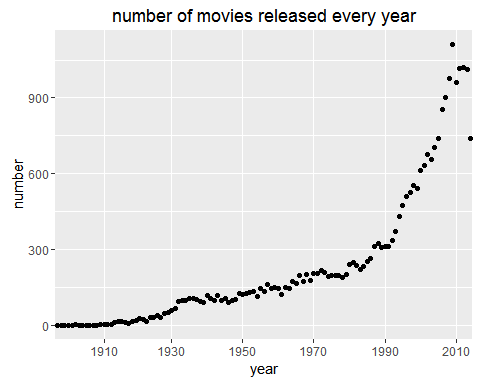
2016年11月28日

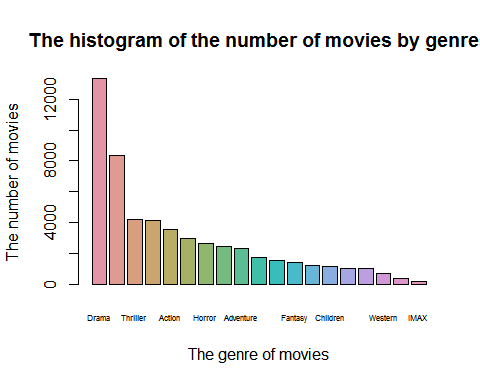
### Statistical analysis about movie data set

The movie dataset collect the films that are released from 1981 to 2015 and has received at least one rating, including 27278 movies. Here is a simple summary of the movie dataset.

## # A tibble: 6 × 4  
## movieId title  
## <int> <chr>  
## 1 1 Toy Story (1995)  
## 2 2 Jumanji (1995)  
## 3 3 Grumpier Old Men (1995)  
## 4 4 Waiting to Exhale (1995)  
## 5 5 Father of the Bride Part II (1995)  
## 6 6 Heat (1995)  
## # ... with 2 more variables: genres <chr>, year <chr>

The first chart is a scatter plot about the number of movies released every year. This figure shows that the number of movies increased every year without no exception. Especially, during the period from 1990 to 2010, the number increase from 300/year to 900+/year, by almost three times, which is significant development.



The following figure count movies by genres. A important note about a movie is that one movie can have one or more genres. Therefore, the total number of movies by genres is much larger than 27278-the total number of movies. It is obvious that the drama is the most common genre of movie while IMAX is the most rare genre. 

To specify whether the genres of released movies is different by year, a stack hist of different genres movies by year is very useful. The following chart shows that there is no big difference among years after 1910 and before that Comedy、Drama、MUsical and Romance are most popular. 