rating part

Mengyan Cui

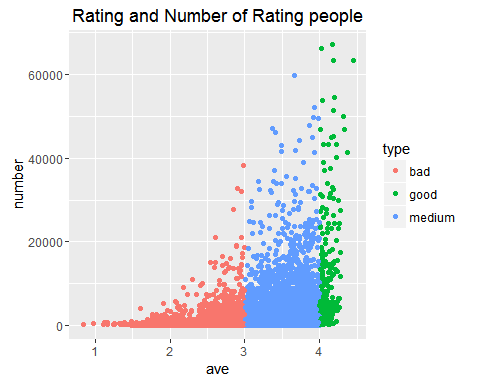
2016年11月30日

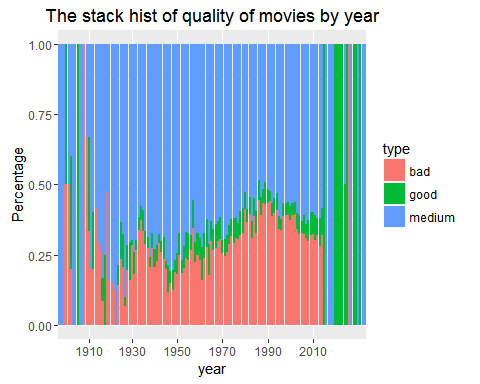
### Statistic analysis about rating dataset

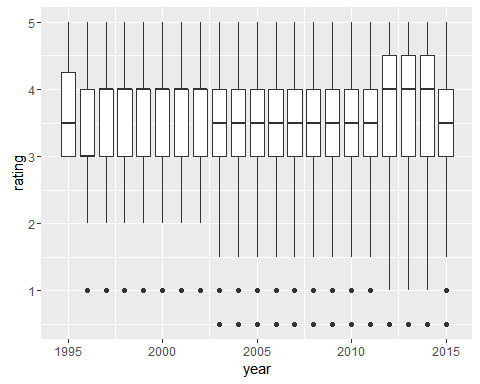
All ratings are contained in the rating dataset. Each line of this file after the header row represents one rating of one movie by one user. Here is a overview about rating dataset.

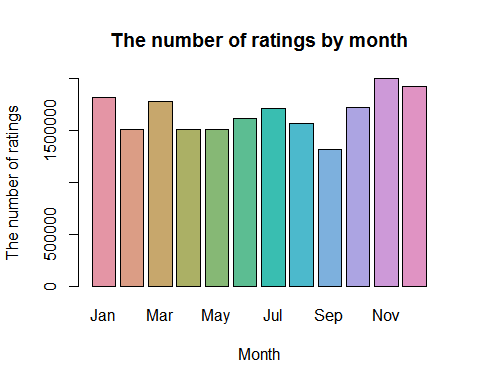
summary(rating)

## userId movieId rating timestamp   
## Min. : 1 Min. : 1 Min. :0.500 Min. :7.897e+08   
## 1st Qu.: 34395 1st Qu.: 902 1st Qu.:3.000 1st Qu.:9.668e+08   
## Median : 69141 Median : 2167 Median :3.500 Median :1.104e+09   
## Mean : 69046 Mean : 9042 Mean :3.526 Mean :1.101e+09   
## 3rd Qu.:103637 3rd Qu.: 4770 3rd Qu.:4.000 3rd Qu.:1.226e+09   
## Max. :138493 Max. :131262 Max. :5.000 Max. :1.428e+09

Since a movie has been rated by plenty of users, the relationship between reveiwers and the average rating is an interesting topic to figure out. The following image reveals that there is correlation between the two variables. That is, the average rating could reflect the real quality of this movie, which is not change with the change of number of reviewers. Besides, movies can be divided into three group based on the average rating: if the average rating is small than 3, the movie is bad, and if is lager than 4, it must be good, and the others are medium ones. 

According the distinguish rules mentioned before, the stack histgram of the quality of moives by year can reveal whether the distribution of good movie or bad movie is concentrated. In fact, the rating data is incompleted since there will still be reviewers rating newest movie. Basically, the tendency of movies' quality is inconspicuous and the fluctuation of good movie is smaller than bad movies and medium movies. 

Here is another chart to figure out the average quality of movies in every year. the barplot is more intuitive. 

Considering that the timestamp is a variable recording comment time, we draw a histgram about the number of ratings by month to see if the rating was effected by month. The result seems no effect of month. 

Similarly, to see whether the hour factor will influence rating, we plot a chart about the rush hourof users' comments. And there is no big fluctuation hour by hour. 