DS 501

Case Study 2

Report

Group 10:
Chao xu, Congyang Wang, Zhaoning Su, Jiexuan Sun

A. The Dataset Description:

The data we collected contain 1,000,209 anonymous ratings of approximately 3,900 movies made by 6,040 MovieLens users who joined MovieLens in 2000. And the dataset including three files – rating file, user file and movie file.

- RATINGS FILE DESCRIPTION

All ratings are contained in the file "ratings.dat" and are in the following format:

UserID::MovieID::Rating::Timestamp

- USERS FILE DESCRIPTION

User information is in the file "users.dat" and is in the following format:

UserID::Gender::Age::Occupation::Zip-code

- MOVIES FILE DESCRIPTION

Movie information is in the file "movies.dat" and is in the following format:

MovieID::Title::Genres

B. The Relationship Between the Topic and Business Intelligence:

We assume that we are a movie company. And the data we collected is about the movie ratings among the United State. Now, we want to expand our market. Analyzing this dataset can help us figure out our target customer and target market. Then, we could decide our next movies' filming topic.

C. How We Analyze the Data:

We mainly use the pandas library under python programming language. Also, we use the matplotlib library to help us visualize the result which can show the statistical relationship among data clearly. And, a plotly library is used in our analyzing for showing the geographical distribution of costumers' rating.
For more details, please see the following part.

D. What We Get from It and the conjectures we make:

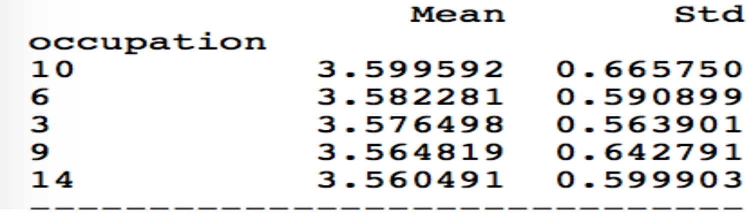
1. Importing the 1 million ratings data set from the MovieLens data set and merging the 3 parts data into a single Pandas Data Frame. Then we store the data into an HDF5 file for further analyzing.

2. First, we analyze some basic details about the data. We found that:
    - There're totally about **3,900** movies are rated by **6,040** MovieLens users.
    - There're **21** movies have an average rating over 4.5 overall.

- There’re **23** movies have an average rating over 4.5 among men, while **51** movies have an average rating over 4.5 among women. (Obviously more movies get a high rating among women than men.)
- There’re **86** movies have a median rating over 4.5 among men over age 30, while **149** movies have an median rating over 4.5 among women over age 30. (We can see that even in people who over age 30, women are still more easily to than men.)
- We define ‘popular’ is the movie with large number of ratings. Then, we find out the movies with top ten rating numbers, they are:

```
American Beauty (1999)                                      3428
Star Wars: Episode IV - A New Hope (1977)                   2991
Star Wars: Episode V - The Empire Strikes Back (1980)       2990
Star Wars: Episode VI - Return of the Jedi (1983)           2883
Jurassic Park (1993)                                        2672
Saving Private Ryan (1998)                                  2653
Terminator 2: Judgment Day (1991)                           2649
Matrix, The (1999)                                          2590
Back to the Future (1985)                                   2583
Silence of the Lambs, The (1991)                            2578
```

- We want to find people with which kind of occupation are the easiest to please. We first calculate the mean rating of each movie of different occupations. Then, we calculate the mean rating and its standard deviation of all movies of different occupations. After computing the ratings mean and standard deviation of whole movies grouped by occupations, we find that the top 5 occupation that are the easiest to please are: K-12 students, doctors and health care, clericals and admins, homemakers, sales and marketing.
-



```
                   Mean        Std
occupation
10             3.599592   0.665750
6              3.582281   0.590899
3              3.576498   0.563901
9              3.564819   0.642791
14             3.560491   0.599903
------------------------------------
```

Table 1 occupation with the top 5 high average ratings

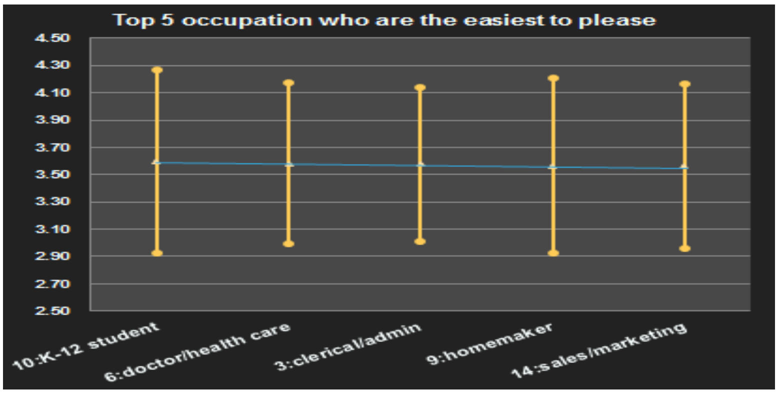
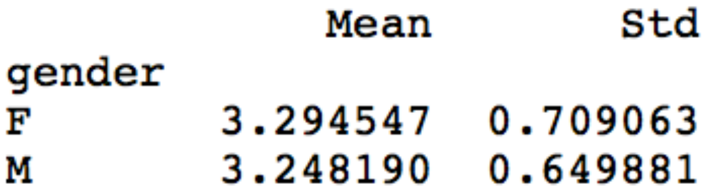


Figure 1 The error bar figure of top 5 occupation that are the easiest to please

- We have conjectured that female are easier to please than male. Now we use the same method to confirm it. Our result show that Female have higher average movie rating among different movie. So it proves our conjecture. But, the female has bigger standard deviation, which means their rating have more variance.



```
                Mean        Std
gender
F           3.294547   0.709063
M           3.248190   0.649881
```

Table 2 The mean and std of ratings among women and men

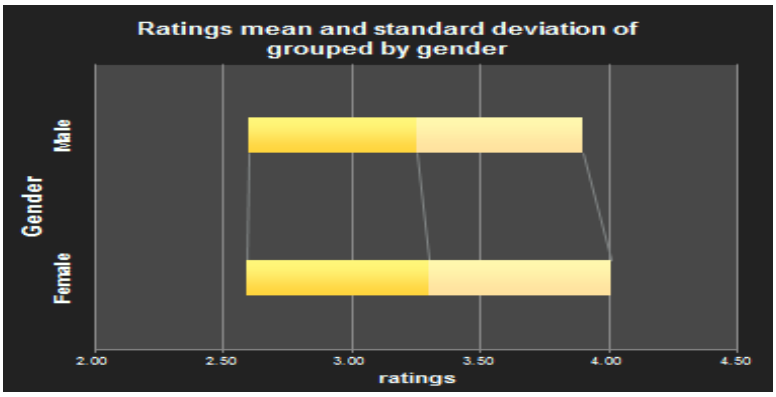
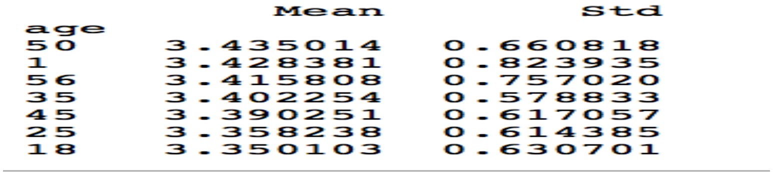


Figure 2 The error bar figure by gender

- We also find that people in 50-55 are the easiest to please, followed by people under 18 and then. So, it seems that people aged between 50 and 56 are the easiest group of people to please. And people whose age between 1 and 12 is at the second place. But, they have quite large variance, which means their rating have more variance.



```
            Mean         Std
age
50      3.435014    0.660818
1       3.428381    0.823935
56      3.415808    0.757020
35      3.402254    0.578833
45      3.390251    0.617057
25      3.358238    0.614385
18      3.350103    0.630701
```

Table 3 The mean and std of ratings according to age

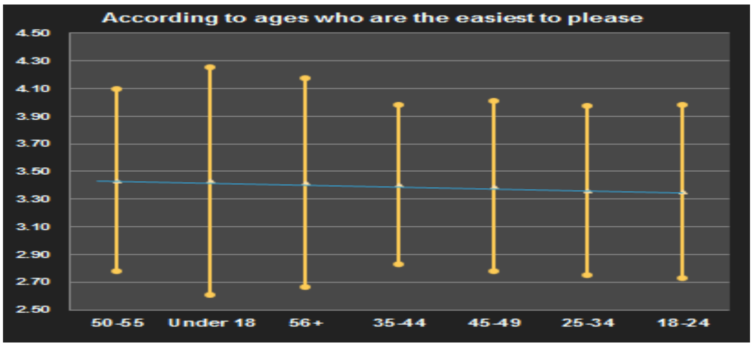
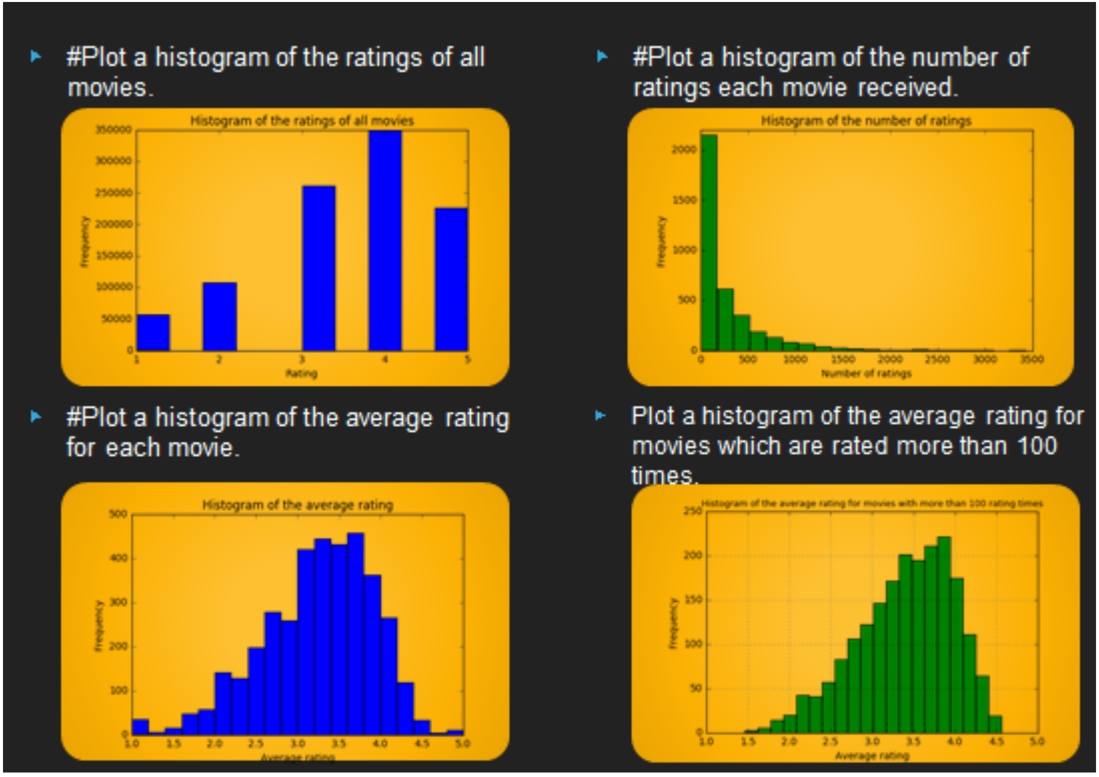


Figure 3 The error bar figure of ratings according to age

3. Then, we expand our investigation to histograms. We perform the figure as following:

   - The first four figure:

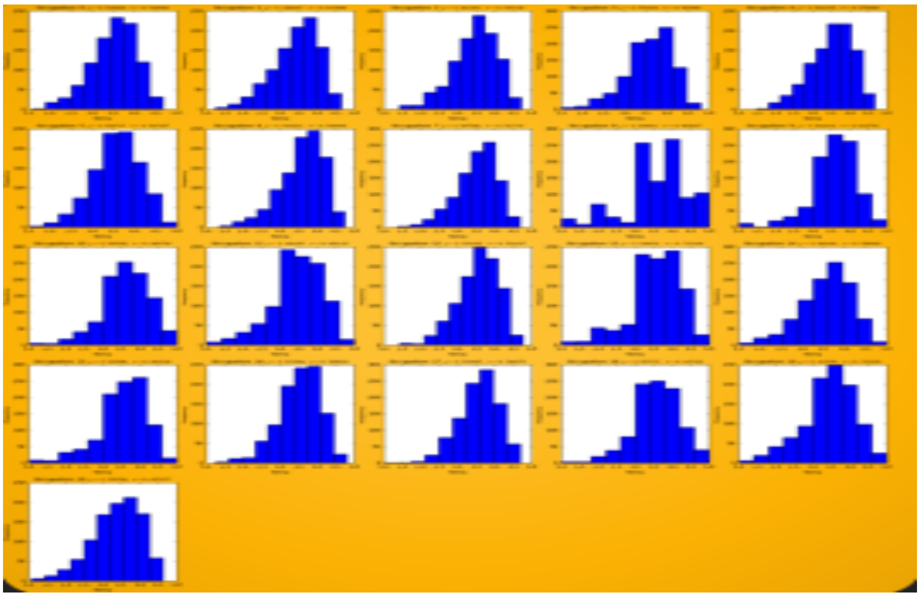


- #Plot a histogram of the ratings of all movies.
- #Plot a histogram of the number of ratings each movie received.
- #Plot a histogram of the average rating for each movie.
- Plot a histogram of the average rating for movies which are rated more than 100 times.

Figure 4 Histograms

   - From the second histogram, we find a long tail here, which can be explained that most customers just rate a few specific movies, and the many other movies just have a little ratings. This implies the pareto principle" 80% of ratings come from 20% of movies." Here we assume that a movie with ratings over 100 can be a popular one and perform the histograms of ratings for these movies. Of course those rated more than 100 times movies we trust are more convincible actually good than those rated less than 100 times since a larger size of sample means a less sample error.

- Then we perform the average rating for movies histograms grouped by different occupations and find farmers (the 8th occupation) have more extreme ratings. Compared with other occupations, Occupation 8 "farmer" looks like a bi-modal distribution. And, obviously, it has the largest variance with more extreme ratings.



-

Figure 5 average rating for movies histograms grouped by different occupations

- Also we compare the average rating for movies histogram between women and men and find that women have more extreme ratings, and its standard deviation is bigger than male.

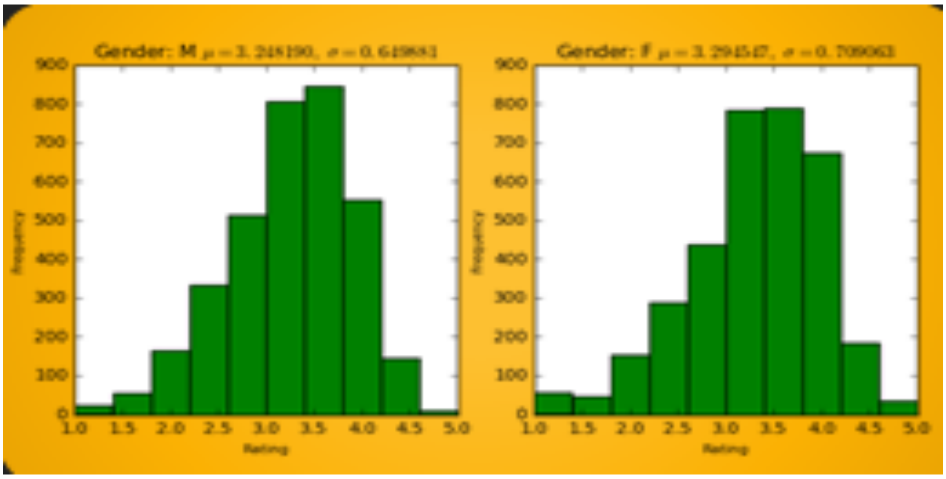


Figure 6 average rating for movies histograms comparing women and men

- Next we plot a histogram of the average rating for movies among different ages and find that, obviously, the people age between 1 and 18, have more extreme ratings. And its standard deviation is bigger than other age groups.

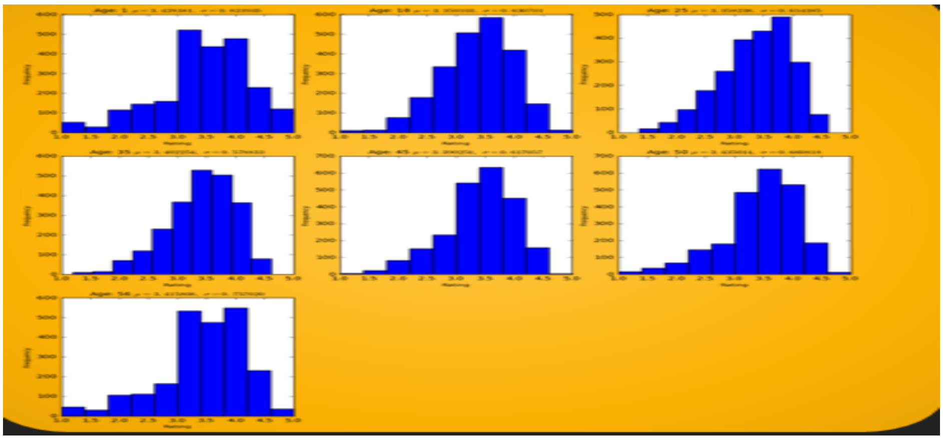


Figure 7 average rating for movies histograms group by age

4. After that, we try to find the correlation between men and women.

- We make two scatter plots of men versus women, one uses the mean rating data for every movie, while the other just uses the mean mean rating for movies rated more than 200 times:

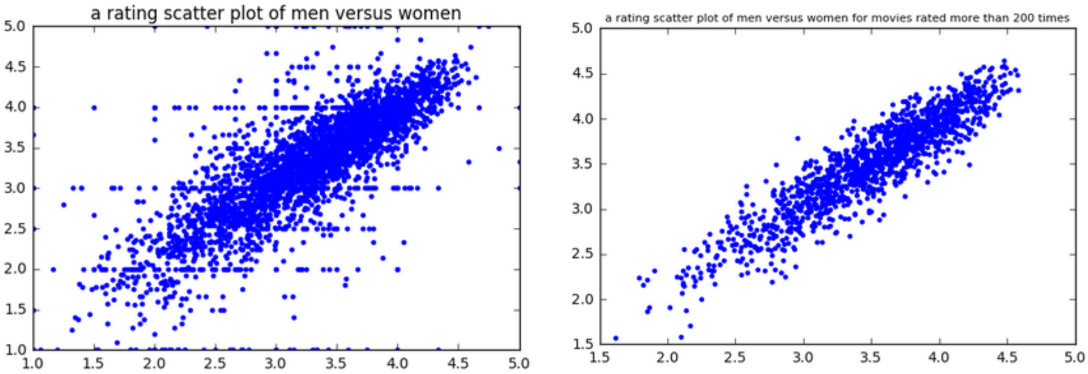
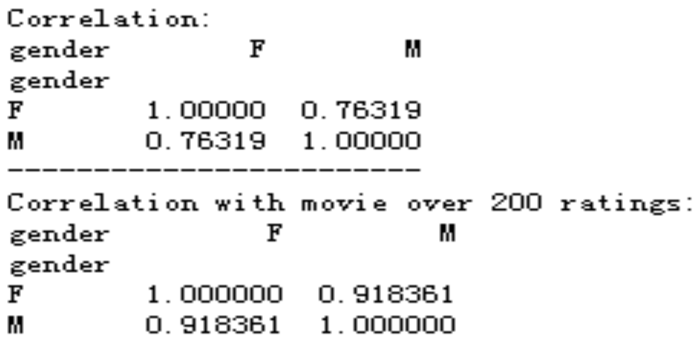


Figure 8 a rating scatter plot of men versus women for every movie and rated more than 200 times movies

- From the figure we can see that men and women have some differences on rating movies. However, for those popular movies which are rated more than 200 times, men and women have strong positive linear relation.

- To convince this, we compute the following correlation coefficient between the ratings of men and women. According to the table above we find that the correlation for every movies between men and women is just 76.319%, while the correlation is 91.8361% between men and women's ratings for movies rated more than 200 times, which just proves our assumption that for those popular movies which are rated more than 200 times, their rating are quite similar.



```
Correlation:
gender          F          M
gender
F         1.00000   0.76319
M         0.76319   1.00000
-------------------------
Correlation with movie over 200 ratings:
gender           F           M
gender
F         1.000000   0.918361
M         0.918361   1.000000
```

Table 4 correlation coefficient between the ratings of men and women for every movies and movies rated over 200 times

- We conjecture that under the circumstance that "age between 25-34", we can relatively accurately predict the ratings by other gender. Compute the correlation coefficients between the ratings of men and women among different age groups. The result show that the correlation between male and female is largest at age group between 25 and 34. So, it proves our conjecture. Also, the scatter plot proves our conjecture also.

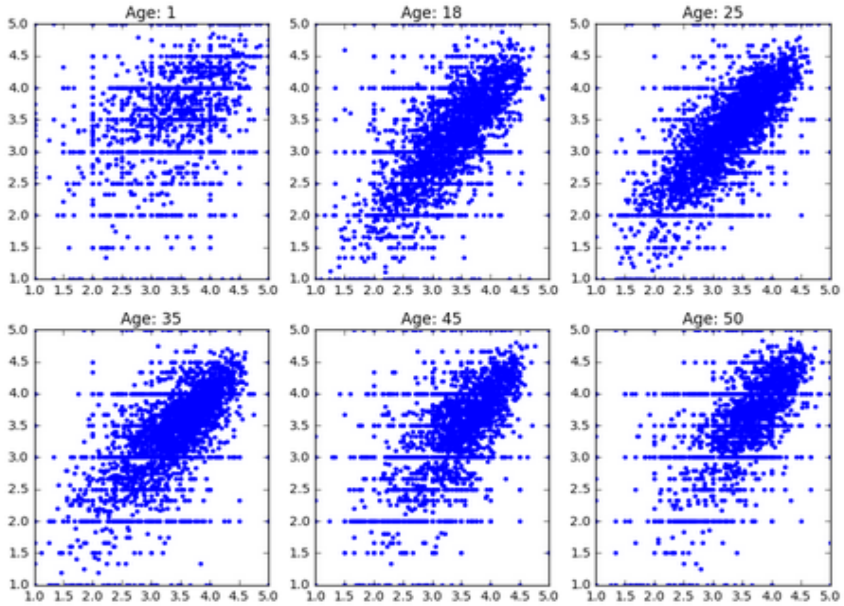


Figure 9 men vesus women on average ratings of movies of different age

- We conjecture that under the circumstance that "the occupation is academic/educator", we can relatively accurately predict the ratings by other gender. Compute the correlation coefficients between the ratings of men and women among different age groups. The result show that the correlation between male and female is largest at occupation group of academic/educator. So, it prove our conjecture. Also, the scatter plot prove our conjecture also.

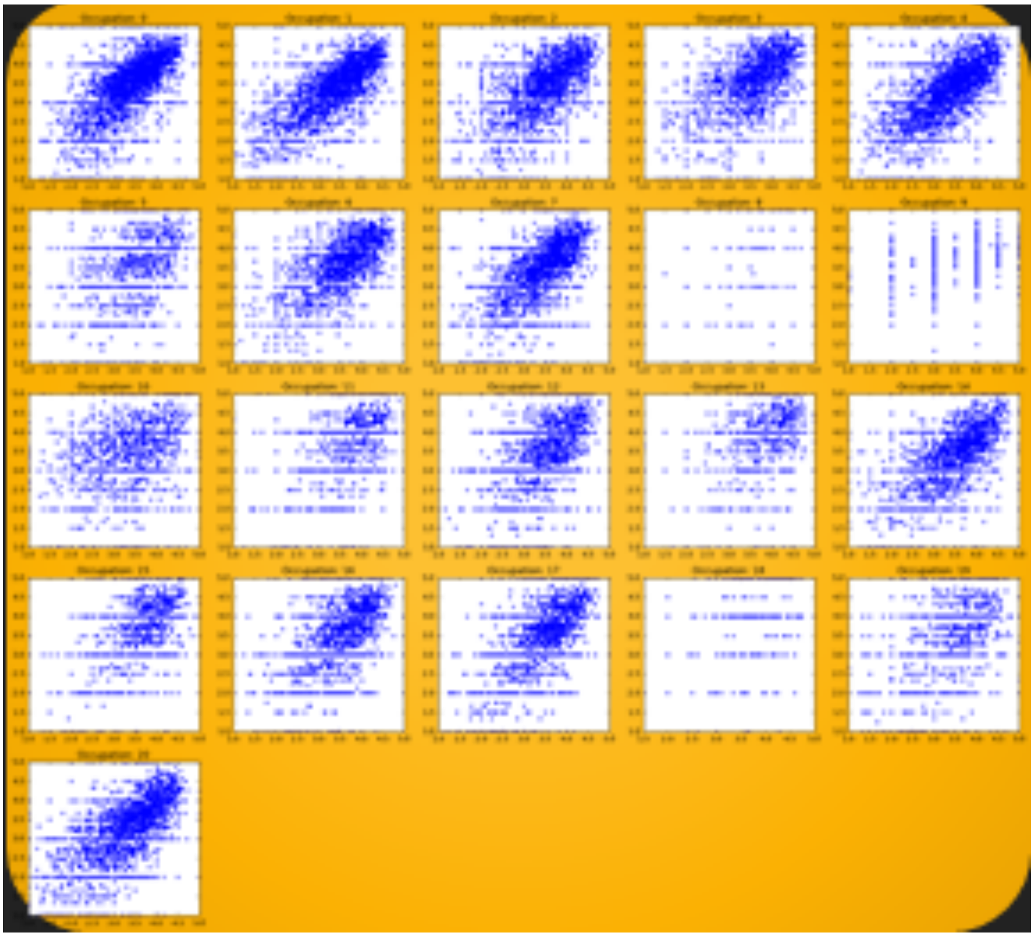
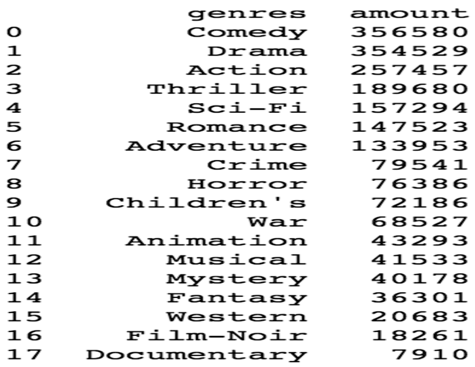


Figure 10 men vesus women on average ratings of movies of different occupation

5. Next, we have a business question about which kind of movie we should film? We calculate the number of the rating among different genres. We found that comedy, drama and action are the top three popular movie genre. Maybe we should consider that our next movie should be in one of these three genre.

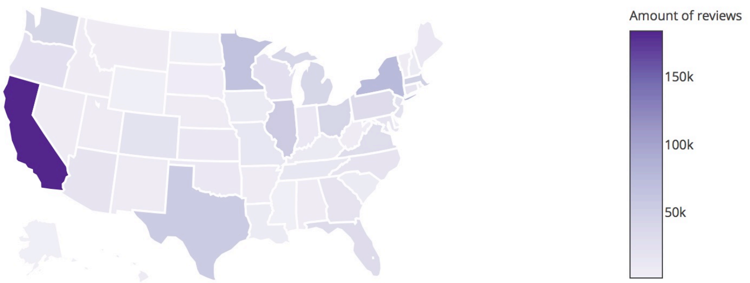


```
          genres   amount
0         Comedy   356580
1          Drama   354529
2         Action   257457
3       Thriller   189680
4         Sci-Fi   157294
5        Romance   147523
6      Adventure   133953
7          Crime    79541
8         Horror    76386
9     Children's    72186
10           War    68527
11     Animation    43293
12       Musical    41533
13       Mystery    40178
14       Fantasy    36301
15       Western    20683
16     Film-Noir    18261
17   Documentary     7910
```

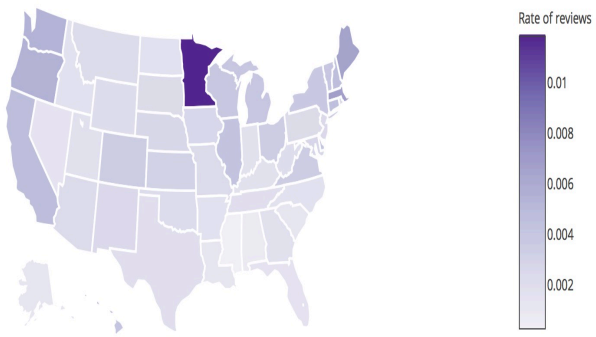
6. At last, we want to know Which state should we advertise for our new movies. We are eager to find the distribution of these rating geographically. We calculate the movie rating number and the number of the rating/population among different states. And we also visualize it as the following graphs. We find that California states residents have relative more movie rating number, which means that they are keen on film. So, we should advertise our new movie at those states. What's more, we

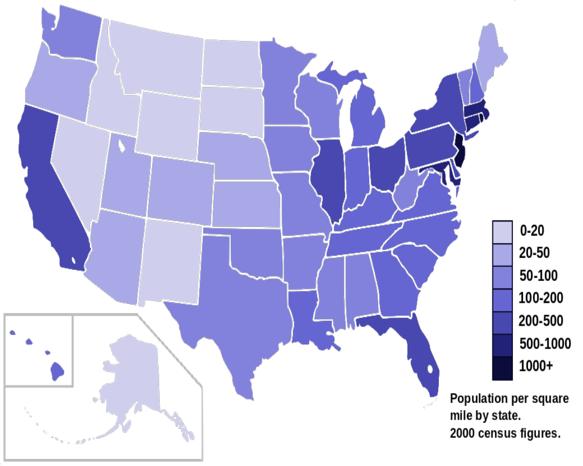
found some state have very low (number of the rating/population) rate, such as Alabama, which means that the rate of people who loved watching movie is very low. But, considering the population of each state, we would like to choose those state having relatively large population and low (number of the rating/population) rate, such as Florida and Texas as our new market areas. We would like to popularize movie culture and do some activities related to films. And interestingly, Minnesota have very large (number of the rating/population) rate. We find that the GroupLens, which is the website we get the data, is owned by University of Minnesota. So, it explained why Minnesota have such a high rate. And, the data is obviously biased geographically.

Moive Reviews Amount among Different States



Movie Review Amount/Population among Different States





E. Business Decision:

Based on the above result, even the K-12 student and people under 18-year-old are easier to please, they are also the group of people who have high variance. We want to choose a group of people who have relative high average rating and relative low rating variance, like occupation 17 - technician/engineer or people age between 25 and 34 as company's target customer. So, our next movie should focus on those group of costumers

we should consider that our next movie should be in as one of the following genre: comedy, drama and action. Also, as a movie company, we would like to advertise our new movie at California and We would like to popularize movie culture and do some activities related to films at Texas/ Florida. And we will treat this area as our new marketing area.