Casestudy 2

Shaowei Gong, Houjue Wang, William Grudzinski

Problem 1:

1. Download the 1 million ratings data set.

We download the 1 million ratings data set as requirement.

1. Merge all of the data into a single Pandas DataFrame

We use pd.read\_table to read these 3 dat format file into python as dataframe, and then use pd.merge to merge these three dataframe into one dataframe for the convenience of our analysis.

1. Store the data into an HDF5 file.

Use store = pd.HDFStore('C:\Users\Shaowei\Desktop\DS501\mydata.h5'), store['obj1'] = data to store dataframe *data* into a hdf format file.

1. How many movies have an average rating over 4.5 overall?

First we group our data by movie\_id, then we select out the movies that have an average rating over 4.5 overall, finally, we count the number of movies that meet with requirement.

We find that only 29 movies meet this requirement.

5. How many movies have an average rating over 4.5 among men? How about women?

First we select out the rating rated by female and male separately and then do the same the just like above.

We find that 70 movies have an average rating over 4.5 among women and 29 movies have an average rating over 4.5 among men.

6.How many movies have an median rating over 4.5 among men over age 30? How about women over age 30?

First we select out the rating rated by female and male whose age is above 30 separately, then we group our data by movie\_id, then we select out the movies that have an median rating over 4.5 overall, finally, we count the number of movies that meet with requirement.

We find that 161 movies have an median rating over 4.5 among women and 93 movies have an median rating over 4.5 among men.

7.What are the ten most popular movies?

Choose what you consider to be a reasonable definition of "popular".

Be prepared to defend this choice.

We think that the most popular movies will have the most amount of ratings. According to this standard, we group movies by movie\_id, then then count them, finally we sort our movies according to the amount of ratings of per movie.

We find that movies with movie\_id 2858 260 1196 1210 480 2028 589 2571 1270 593 are the most popular 10 movies.

8. Make some conjectures about how easy various groups are to please? Support your answers with data!

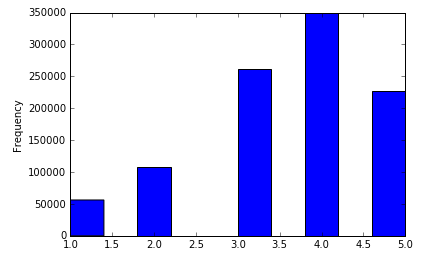
For example, one might conjecture that people between the ages of 1 and 10 are the easiest to please since they are all young children. This conjecture may or may not be true, but how would you support or disprove either conclusion with with data?

Be sure to come up with your own conjectures and support them with data!

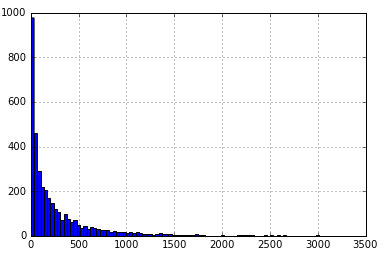
We think that women are more easy to please than men. We find that in this dataset, women have average 3.62 rating for movies, however, men only have 3.57 rating for movies which just support our conjecture.

Problem 2: Expand our investigation to histograms

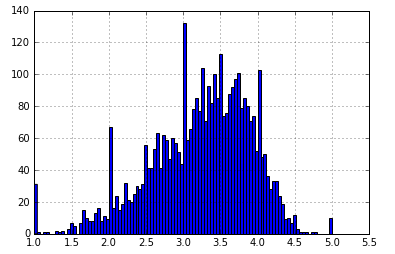
1. Plot a histogram of the ratings of all movies.



1. Plot a histogram of the number of ratings each movie received.



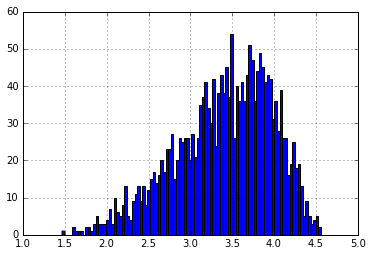
3. Plot a histogram of the average rating for each movie.



4. Plot a histogram of the average rating for movies which are rated more than 100 times.

What do you observe about the tails of the histogram where you use all the movies versus the one where you only use movies rated more than 100 times?

Which highly rated movies would you trust are actually good? Those rated more than 100 times or those rated less than 100 times?



Histogram of movies rated more than 100 times has much more thin tail compared with histogram of all the movies.

For me, I would like to trust movies rated more than 100 times. Because bigger sample can represent the population better.

5.Make some conjectures about the distribution of ratings? Support your answers with data!

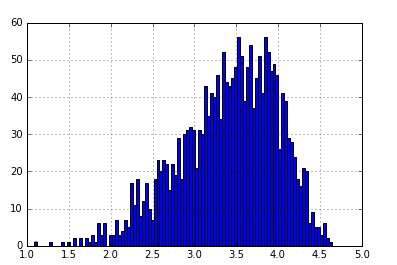
For example, what age range do you think has more extreme ratings? Do you think children are more or less likely to rate a movie 1 or 5?

Be sure to come up with your own conjectures and support them with data!

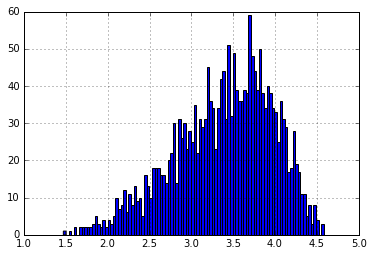
I conject that the rate of Female has more extreme rates than Male.

In order to eliminate invalid rate, we plot rate histograms of Female and Male with more than 100 rates separately .

We find that histogram of female has thicker tails than histogram of male and lower peak. This indicates that Female has more extreme rates.



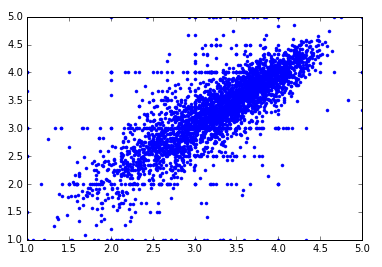
Distribution of Female ratings



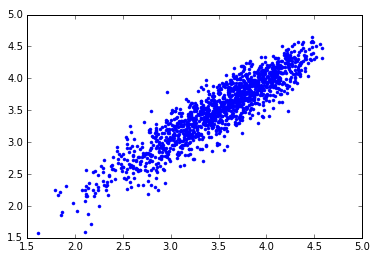
Distribution of Male ratings

Problem 3:

1. Make a scatter plot of men versus women and their mean rating for every movie.



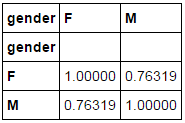
1. Make a scatter plot of men versus women and their mean rating for movies rated more than 200 times.



3.Compute the correlation coefficient between the ratings of men and women.

What do you observe?

Are the ratings similiar or not? Support your answer with data!



The rating between men and women is very similar. According to the correlation coefficient, it is as high as 0.763, so they are highly correlated.

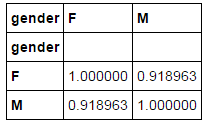
4. Conjecture under what circumstances the rating given by one gender can be used to predict the rating given by the other gender.

For example, are men and women more similar when they are younger or older?

Be sure to come up with your own conjectures and support them with data!

For the movies rated more than 200 times, given the rating by one gender, we can predict the rating given by another gender.

According to the correlation coefficient of rating between male and female for movies rated more than 200 times, their correlation coefficient is as high as 0.92, which indicates that male and female almost have the same behavior in matter of movies rated more than 200 times.



Problem 4:

1. What data you collected?

We collect data from http://grouplens.org/datasets/movielens/ and download the MovieLens 1M Dataset.

This dateset include 3 tables: users(user\_id, gender, age, occupation, zip), ratings(user\_id, movie\_id, rating, timestamp), movies(movie\_id, title, genres). For the convenience of our analysis, we merge these three datasets, and get a new dataset called *data*.

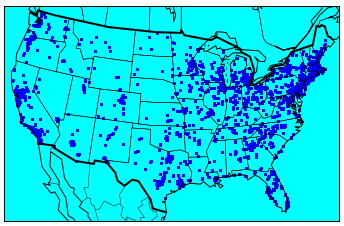
1. Why this topic is interesting or important to you?

We want improve the impact of our advertisements, so we can save cost of our ads and avoid useless ads. In order to do this, we need to understand where is the right place to put our ads based on our data.

3.How did you analyze the data?

We plan to visualize every rating into American map, so we can know where people are more passionate about movie intuitively. Luckily, in our dataset, we have zipcode of every rating, so we can get the corresponded coordinate of every rating. Then we mark every pair of corresponded coordinate as a dot into US map. Place where there are more dense dot is the place that there are more movie watcher.

1. What did you find in the data? (please include figures or tables in the report, but no source code)



We find that most of ratings come from east and west coasts especially like big city like Seattle and Los Angles in west coast and New York in east coast and New England Area . In the middle states, there are much less ratings than other places.

If the sampling of this dataset including 1 million observations can represent the whole US appropriately, we highly suggest that movies companies put their ads in east coast and west coast rather than middle area of US.