Can we reduce the cost of survey about movies?

Group 3 Yao Chun Hsieh, Yuchen Shen, Yang Tao, Ying Fang, Nathan Hsu,

Motivation:

In the perspective of film industry, the box is the most important thing. As a theater, he would like to know whether the audience like the movie or not, so he can decide how long the movie would be on play. To DVDs sellers they would also like to know the rating of audience, so he can decide how many DVD to purchase as inventory, or how much premium he would like to pay. In short, survey is important. However, survey is difficult since it takes long time, human resource and it's not easy to avoid each kind of bias. As a result, it would be a great project if we can use a small group to predict the whole population's tendency. That way, we only need to get the predict model the first time, then we can use them for the predictions thereafter, and reduce the cost of survey.

Data:

We use the data from Movielens, our dataset is 1M. The period is from 2000 to 2003 year, and there are about 4000 movies, from like 1944 to 21th century. From each survey, we have user ID, movie ID, movie Name, user occupation, location, timestamp, and ratings.

Data resample:

Since the users in the dataset cannot present the real demographic, we can't use the result from the dataset to imply that the real world would behave the same way. To solve this problem, we tried to simulate a condition that was the same as the real demographic distribution. First, we collect the real population statistic from government website. We calculate that the proportion of different age intervals. Below is the population distribution of real world and from our dataset.

Demographic Data		Movielens Data	
15-19	7.20%	under 18	3.67%
20-24	6.70%	18-24	18.24%
25-34	14.20%	25-34	34.67%
35-44	16.00%	35-44	19.73%
45-54	13.40%	45-55	17.30%
55+	21%	56+	6.38%

Notice that we drop the data below 15 years old, since we don't think there's too many web users in this range. Then we fit the distribution into our dataset. Because there's slightly difference, so we show the match table here to prevent ambiguous. As you can see the distribution is quite difference between the real world and the dataset we have, therefore it's meaningful to resample the data. Our method is using bootstrap. We first decide the population number we want, say 5000 people. Then from each age category, we randomly chose the instances, and the number is just 5,000*p, where p stands the proportion of each age category. Our method is with replacement. As a result, we have a sample of 5,000, and the distribution is just like the real world. Now we have confidence to use our results to present the real situation.

Method:

We will discuss some basic problem first, trying to get some features about our data. Then we can decide which is the best way to reduce cost. We will define the popular movie, the rating distributions of different groups, the relationship between men and women. Finally, we will show the best way to predict the rating of the whole population from a small group.

Problem 1: The popularity of movie

To determine which attribute has important influence on the rating, we check several of them and find some interesting results.

There are 25 movies having the top average rating over 4.5, but when the data is divided by gender, there is huge movie amount change from 25 to 69(female) and 25(male) which makes us focus on the gender attribute. Considering age attribute based on gender, many movie rating points gathered at the high rating, since the amount of movies' median of rating over age 35 is 166 for female and 109 for male, which is actually much bigger than total age range. As a result, people from different ages have different views of a same movie.

The next part of report will focus on choosing some popular movies. As we think what is a popular movie, it must be watched by many people and have a high rating score. But how to define a high rating score can cause difference. A popular movie need to be favored by all people whatever age they are, so we recalculate the rating of each movie following the below rules:

1. the amount of rating of each movie is bigger than the mean amount;
2. mean rating of each movie separately in different age should be high;
3. when the rating applies to the real world, it should be high among all age;
4. sort the top ten rating movies;

In order to use our result in the real world to reduce cost, we use real world population apply to our rating rule as we show previous.

Movie Title (Top 1 ~ 5)	Movie Title (Top 6 ~ 10)
Shawshank Redemption, The (1994)	Close Shave, A (1995)
Schindler's List (1993)	Wrong Trousers, The (1993)
Godfather, The (1972)	Raiders of the Lost Ark (1981)
Seven Samurai (1954)	Rear Window (1954)
Usual Suspects, The (1995)	Sunset Blvd. (a.k.a. Sunset Boulevard) (1950)

Using our rule for calculation, the result shows the top rating is 4.543. Among the top 10 movies, several of them are a series movies, and they actually are old movies. As a conclusion, old movies are more likely to be favored and when the problem comes to a series of movies, the survey can just apply to one of them, since these kinds of movie always have the similar evaluation. After applying this method to get the popular movies, the movie company or DVD company will know which should be put on more concentration.

Which group is the easiest to be pleased:

To decide which age category is the easiest to please, we first need to define "please". We conclude that those who are easily to give good credit to movies, are easy to please. In another word, the higher the mean rating a person gave, the easier he can be pleased. To implement the concept, we first calculate the average rating for each person. If a person always gives 5, then he got an average rating of 5, which means that he is super-easy-to-be-pleased. The second step is to group by occupations and calculate the mean of

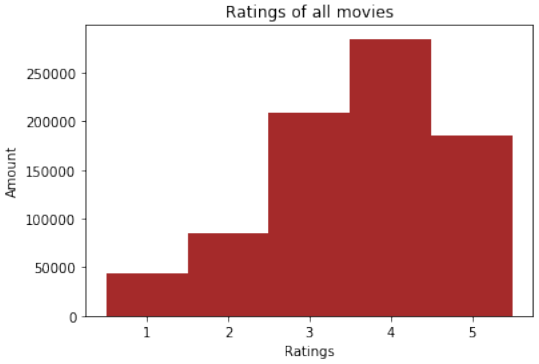
each occupation groups. We also calculate the standard deviation of each group. We conjecture that the occupation that requires more knowledge, or we say the smart guy, may be more critical about the movies. So, they will be the hardest to please. On the opposite, we think students are the easiest to please since they are young, and may not be that critical from those who are working. The result is as follows:

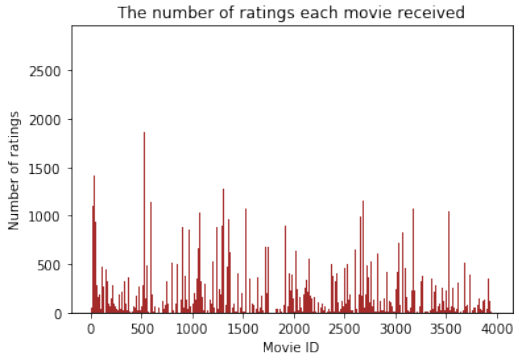
occupation	Occupation	MeanOfEachOcc	StdOfEachOcc
8	Farmer	3.405448	0.698204
5	Customer/service	3.517509	0.438053
19	Unemployed	3.578627	0.592160
10	K-12 Student	3.594185	0.514304
…	…	…	…
17	Technician/engineer	3.754289	0.369672
6	Doctor/health care	3.775062	0.398196
13	Retired	3.840425	0.422330
11	Lawyer	3.858736	0.401172

The result shows a totally opposite conclusion. Lawyer gave us the highest rating, meaning they are the easiest to please. And the farmer is the hardest one to please. Notice the standard deviation of farmer is very high, so we say they are hard to please, but in a non-significant way. And also, the retired people are not that critical as our conjecture, maybe the reason is that they have gone through a long life, and became careless about the rating.

Problem 2: The distributions of ratings

- Overall of movies ratings

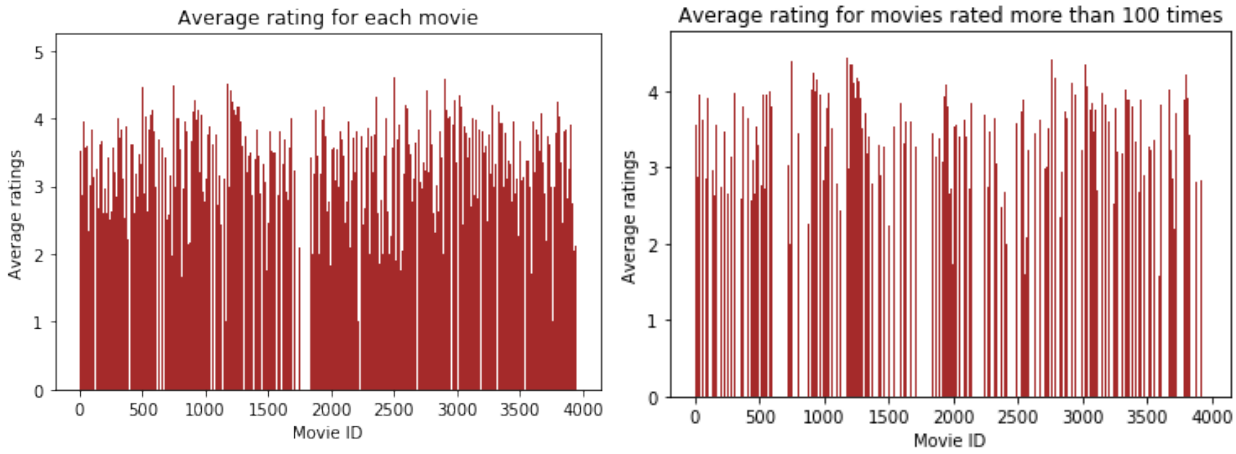




-

We plot the rating of all movies received as show above. The rating that movies received most is 4, then is 3, and very few of movies got rating 1. As shown in the graph right hand above, we can easily see that most of movies were rated less than 500 times, but there are still some very popular movies which were rated more than 1000 times.

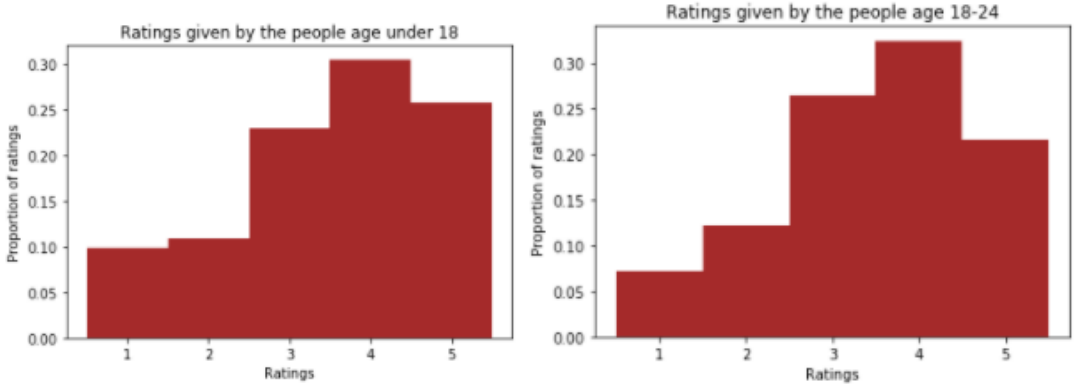
- What kinds of ratings do you trust?

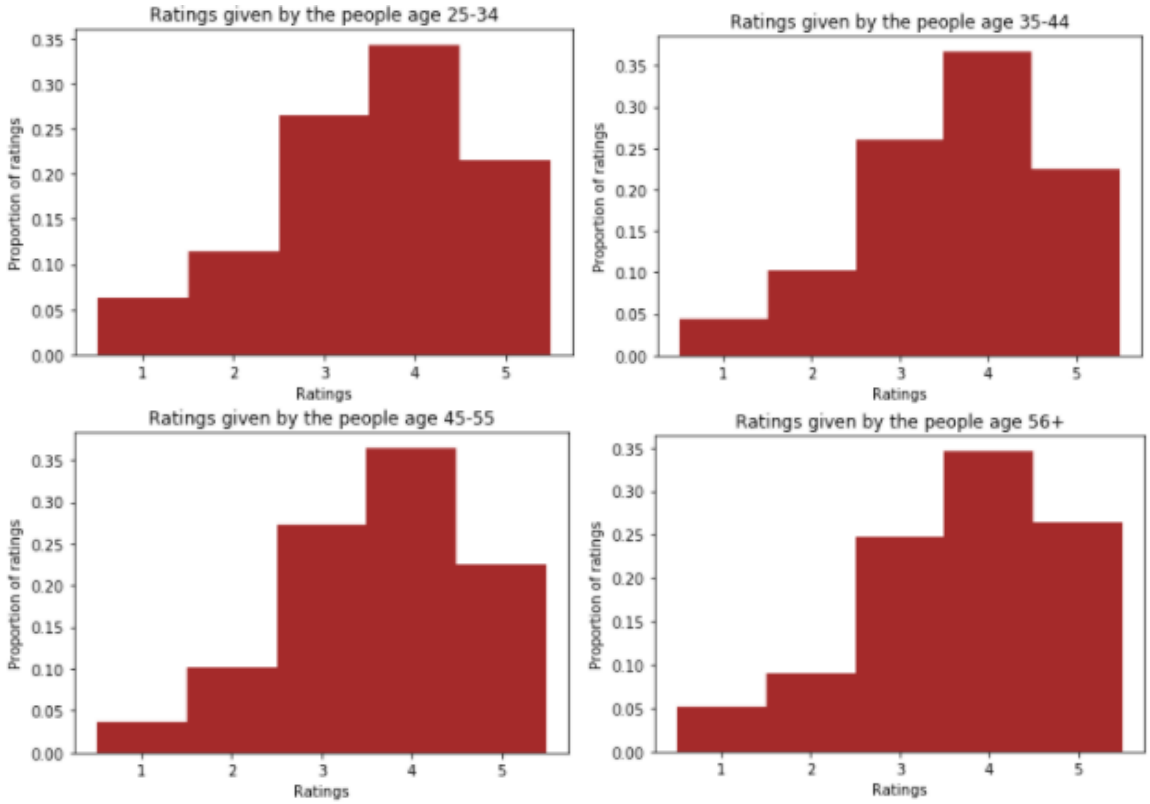


Compare with the average rating for each movie, average rating for movies which are rated more than 100 times generally has high ratings. In other words, imposing restrictions on the number of ratings will be helpful to filter some movies which are not so much popular. In general, people tend to trust that the highly rated movies are actually good since they have been rated many times. In this case, the high ratings movies which were rated more than 100 times are highly trusted they are actually good.

- Conjuncture: older people tend to make higher ratings for movies.

To find out more about movies ratings, we decide to figure out the relationship between rater's age and ratings. We suggest that older people tend to make higher ratings for movies. We plot the average ratings given by each age range (under 18, 18-24, 25-34, 35-44, 45-55, 56+) to see the features of rating habits at different ages. Because the number of raters at each age range is different so we use the proportion of raters at each rating to show the distributions.



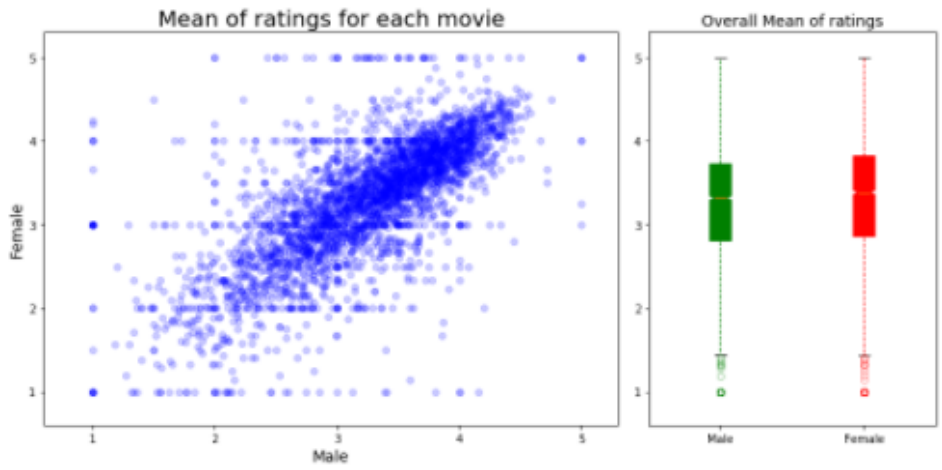


As shown in graphs above, most people would like to rate 4 at each age range. Although all these six charts have similar shape of bars, we can easily find out that people with higher age tend to rate higher. There is a certain possibility of people at young ages to make low ratings. As the age rising, less proportion of people rated 1 and 2, while higher proportion of people would like to give rating 5. It is pretty obvious in the group of the people older than 56: It is more possible for people in this range to rate 5 than to rate 3, but there is no phenomenon like that in other age ranges. To summarize, we can conclude that people at higher age tend to make higher ratings for these movies. This is the same result with problem 1. And since the distributions are similar to each other, we can use this finding to solve our business question.

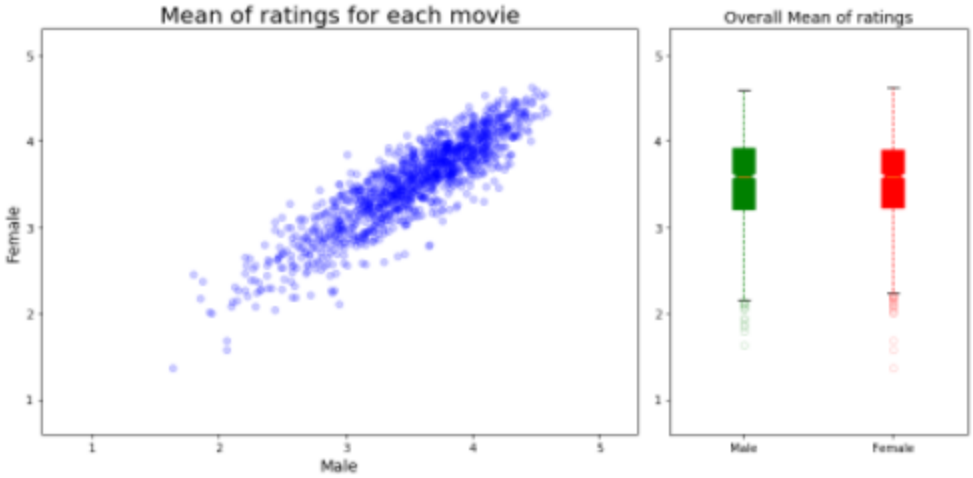
Problem 3: The correlation between men and women

For the ratings of all the movies in our database, what about considering problems with the aspect of gender? Is there any correlation between men and women? Can the rating given by one gender be used to predict the rating given by the other gender? Let's delve into the data to find more information.

Firstly, we make a scatter plot of men versus women and their mean rating for every movie to see if there exists some interesting insight. The result is below:



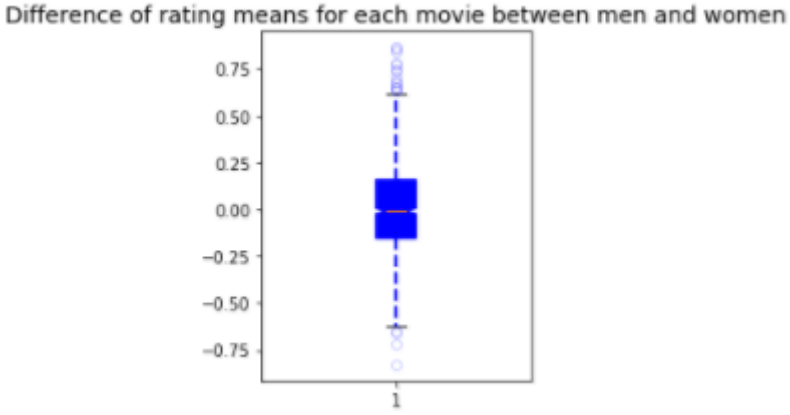
From the plot, an obvious main trend for the ratings between men and women is observed. For most of the cases, the ratings given by these two groups are quite similar. However, there are also a part of data not following the trend; a movie can be rated very high by women while very low by men, and vice versa. Maybe this is caused by the number of rating is too small. When the rating count of a movie is small, it gets biased easily. Also, if there are too many noises in the data, the performance of a prediction model will be poor. Thus, in order to eliminate this factor, we make a scatter plot of men versus women and their mean rating for movies rated more than 200 times. Moreover, we use a boxplot to illustrate the distribution of rating values of men and women. This time, the trend is still obvious while less point data running out of the trend. The result helps us to come up with a conclusion that for movies having more than 200 ratings, there is a linear correlation between men and women when they give rates to the same movie.



Let's go further with the concept of correlation between men and women. How well do the rating values given by men correlated to women? How to demonstrate that concept with statistical numbers instead of just plots and conjecture? By computing the correlation coefficient between the rating values given by men and women and the p-value, we can get the answer easily. For all the movies, the correlation coefficient is 0.699, and the p-value is 0.0. For movies having ratings more than 200, the correlation coefficient is 0.890 and the p-value is 0.0. To interpret the correlation coefficients and the p-value we get, let's denote the dataset of all the movies as DataSet_A, and the dataset of movies having rating more than 200 as DataSet_B. According to the results, we declare that correlations between men and women are

more significant under the environment setting of DataSet_B; favors of men or women are more predictable under this condition.

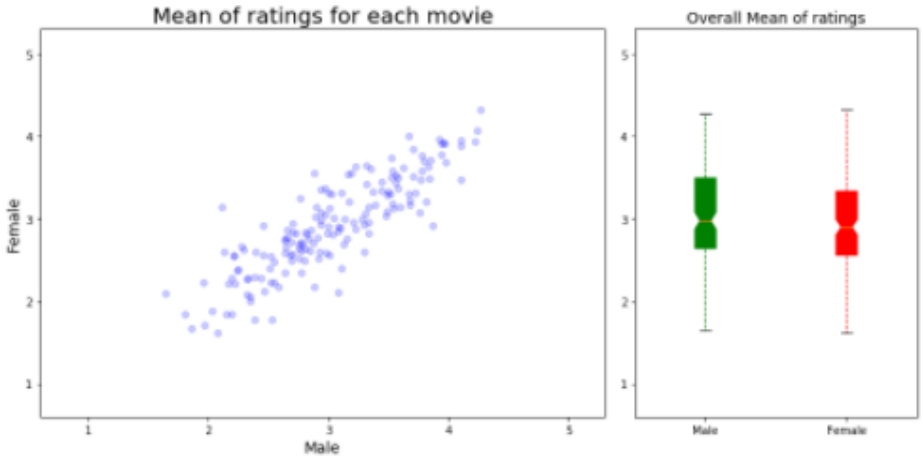
Now, we got the trend, and we realize the correlated coefficient is significant and we know while a gender gives a relatively high rate, the other gender would also behave the same. One more interesting question is if the rating values between men and women are similar. Here is an example for explaining the difference between the high correlation and the similarity of rating values. For 3 movies, men give a mean rating value with 1, 2, 3 and women give 3, 4, 5 respectively. We say the trend is similar, the data between men and women has high correlation, but the rating values are different. Therefore, to find the answer of this question, a box plot is used for showing the overall mean of ratings for men and women.



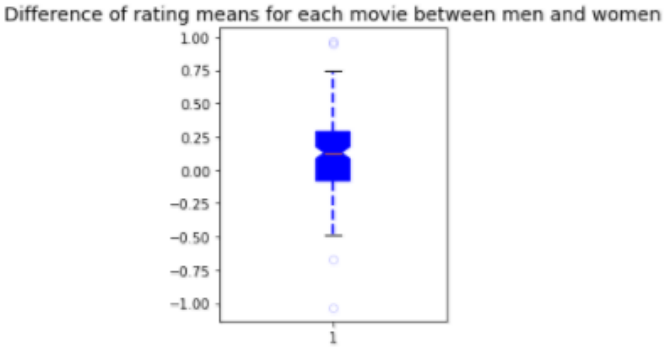
According to this boxplot, the IQR(Q1-Q3) is smaller than 0.5, with median close to 0. Thus, it is confident to have the conclusion that not only the trend is similar, but the values of rating are also similar between men and women for movies having more than 200 ratings.

For predicting ratings between different gender, we conjecture that the rating values are more predictable for comedy movies than for horror movies. The reason is because we believe that when it comes to scary things, men have generally been trained to be more fearless than women, and they also become calm about this. The consequence is that men might give similar rating values for all the scary movies, while women might give a wide range of values since they express their emotion in the different way from men. Therefore, the rating trend between men and women for horror movies may not be significant. On the other hand, we think people's reaction are much similar for funny things regardless the gender. The rating trend for comedy movies may be highly correlated between men and women. If there exists sufficient trend, the rating values given by one gender might be used to predict the rating given by the other gender.

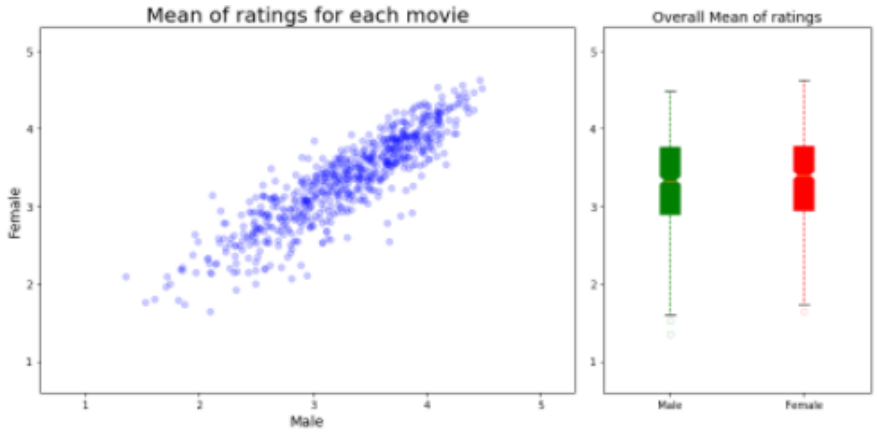
To support our conjecture with data, we first plot the mean of ratings for horror movies between men and women:



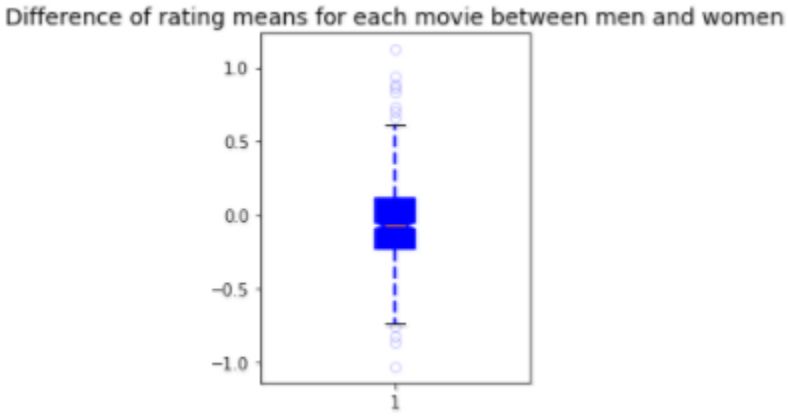
According to these two plots, it seems like there is a trend between men and women for horror movies, and the ranges of rating values are also looked alike. We then calculate the correlated coefficient and the p-value for the data. The correlated coefficient is 0.862 and the p-value is less than 0.01. Also, the box plot shown below is the different rating means for each horror movie between men and women, the IQR(Q1-Q3) is small. From the first glimpse, ratings for men and women might not be that different for horror movies as we expected.



Now we do exactly the same analyzing for comedy movies. The mean of ratings for comedy movies between men and women is as below. A trend can be observed, and the range of mean ratings are similar.



The correlation coefficient is 0.890, which is better than horror movies, and the p-value is also less than 0.01. The IQR(Q1-Q3) shown in box plot below is also small. The last step is to build a linear regression model and check if the model can predict well.



From the result above, although the correlation between men and women for horror movies is high, the correlation for comedy movies is higher. Therefore, our conjecture that the rating values are more predictable for comedy movies than for horror movies is correct and well supported.

Find the group that can be used to predict other's rating:
In order to find a group that can be used to predict other groups, we need regression. Before regression, we need to deal the data first. We first divide the data into 6 groups by age, as shown in the previous table. The reason we pick age as our dividing attribute, is from the problem 2 that between each age range, the distributions are very similar to each other. And we want to find the group which is the most representative. Then for each group of age, we calculate the mean ratings of each movie. And also, we calculate the mean ratings of each movie of "other groups". For example, we calculate the average ratings of each movie for people UNDER 18, and average ratings of each movie for people that is ABOVE 18. As a result, we will have 6 datasets. The independent variable is the rating of a specific age group, and the dependent variable is of the people from all the other groups. The regression model is as follows:

$$y_i = \beta_0 + \beta_1 * x_i$$

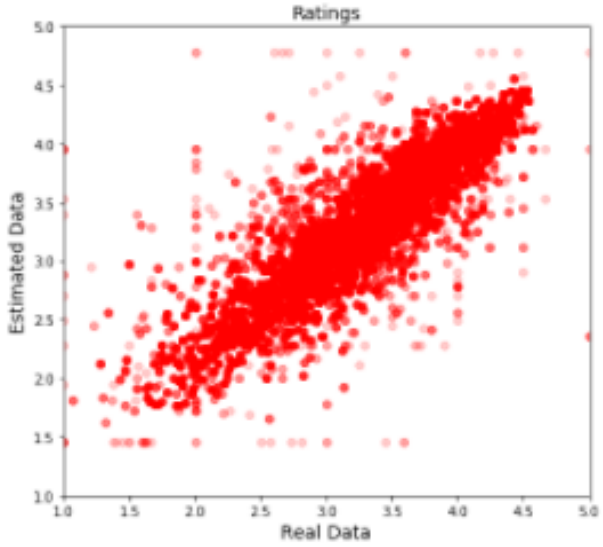
	Yi (avg rating of group)	Xi (avg rating of group)
Regression 1	Above 18	Under 18
Regression 2	Not in 18-24	18-24
Regression 3	Not in 25-34	25-34
Regression 4	Not in 35-44	35-44
Regression 5	Not in 45-55	45-55
Regression 6	Under 56	56+

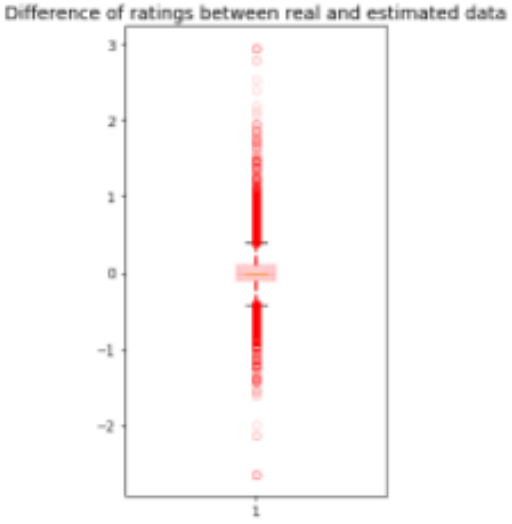
For each regression, we separate them into 70% as training dataset, 30% as testing dataset. Before the regression, we shuffle the dataset. We will show the best and worst regression results, and the test error result (predicted value minus real value).

OLS Regression Results

Dep. Variable:	y	R-squared:	0.869
Model:	OLS	Adj. R-squared:	0.869
Method:	Least Squares	F-statistic:	1.476e+06
Date:	Wed, 25 Oct 2017	Prob (F-statistic):	0.00
Time:	23:56:37	Log-Likelihood:	50188.
No. Observations:	223037	AIC:	-1.004e+05
Df Residuals:	223035	BIC:	-1.004e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.6176	0.002	250.962	0.000	0.613	0.622
x1	0.8334	0.001	1215.086	0.000	0.832	0.835





Summary:

The best result is from regression 3, which is using people between 25-34 to predict the other groups. The R squared is 87%, and the P value is zero, shows that the model is significant and the predictability is quite well.

$$y_i = 0.6176 + 0.8334 * x_i$$

For the testing data, regression 3 showed a linear relationship, and 99% of the test error are within a range less than 0.5 with our regression model. As a result, we can declare that this model works pretty well, and we can use this group to predict other groups. From the real demographic, this age group takes about 14% of the whole population, and the distribution of our dataset is the same since we use bootstrap. In other words, we can do survey on only 14% of the whole population, and use the result to predict the rest 86% people. For the MovieLens case, we can keep only the rating feedbacks of people in range of 25-34, therefore we can decrease the storage and maintenance cost. If we want to do the survey on the street, we can choose the time after work, and some fancy places like movie theater, department stores that attracts these kinds of people. Then, we have a solution toward our problem.