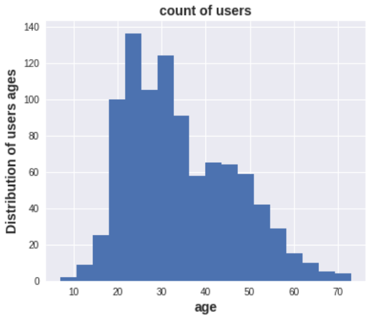
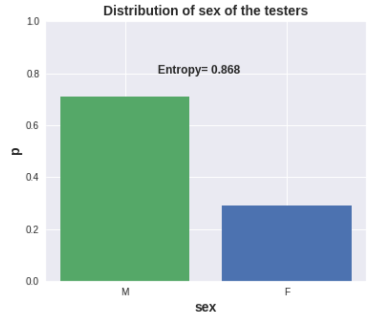
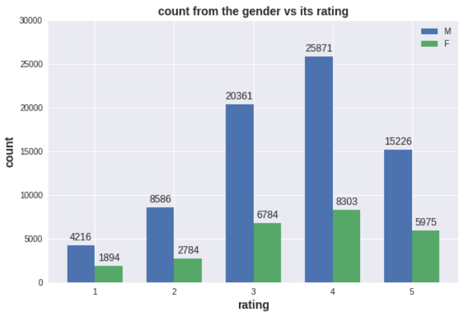
1 Explore the data

1.1 Data explenarsion

This dataset deals with the rating (1-5) of 1682 different films and classifier this films in several groups like (Children, Adventure,...).
The first file (u.user) contain information from the film tester like their ages, gender or occupation. The data in this file are nominal(gender, occupation) and the other columns are numeric values. There are now ethnical problems because the connection between the user's age, gender or occupation is a idNumber. There is no way that you can find out the real name to that test user otherwise there are ethnical problems. A solution for the age column could be that they can divide it into 4 parts to secure the information. For the sex column it is possible to decode it for example "F" → 1 and "M" → 0. So the reader could not know the gender and for the algorithm it does not change anything. For the third column it is also possible to divide this into other sections like "official", "private sector".
The second file u.data contained only numerical values. The rating column is divided into (1-5) where 1 is the worst and 5 is the best evaluation.
The last important file is (u.item). It has binary types for the classification of the film, a column with only *nan* values, nominal values (film names, Url sources) and dates.
The column with the *nan* values could not predict because there is no way to do this only to look on the internet and add this information.







It is not a well distributed ages column. It would be better if it's better distributed. For example to increase the number of 38 age-old test users. On the other side the graph shows that there are roughly twice men as woman. This could be crucial for the rating for example for more women orientated films. Then the men would rate these movies worse than the women or in the other direction. On the right side we could definitely see the difference in the numbers of the rating from the Woman "F" and the men "M". The mean value of the rating from the men is 3.53 and for the women 3.52. That means that they choose wisely movies for women and also for men.

Children	Drama	Sci Fi	Thriller	Romance
-0.033961	0.034519	0.044447	0.030092	-0.049035

This table shows the correlation between the sex or gender column with the different binomial classifiers form the movie. There is definetly no correlation between these values and the gender.

Children	Drama	Sci Fi	Thriller	Romance
-0.043644	0.114006	0.010471	-0.009802	0.040107

Similar to the upper table there is also no correlation between in's binomial classifiers and the rating column.