1 Explore the data

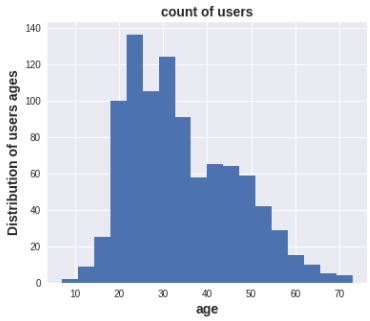
1.1 Data explenarsion

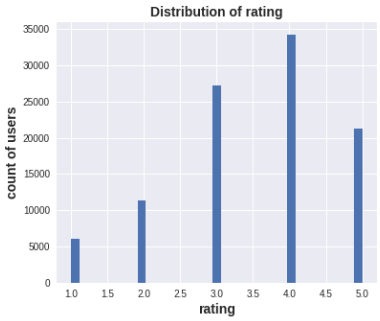
This dataset deals with the rating (1-5) of 1682 different films and classifier this films in several groups like (Children, Adventure,...).

The first file (u.user) contain information from the film tester like their ages, sex or occupation. The data in this file are nominal(sex, occupation) and the other columns are numeric values.

For the nominal columns there could be 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" $\rightarrow 1$ and "M" $\rightarrow 0$. 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 best and 5 is the worst.
The lastimportant file is (u.item). It has binary types for the classification 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.





3.52986 It is not a well distributed ages column. I would be better if its better distributed. On the other side the graph shows that there are roughly twice men as woman. This could be crucial for the rating for example more women orientated films. For the rating distribution there is a tendency for bad rated films.

The target for a regression and a classification task could be the "rating" column.

For the classification there are multiple classes from (1-5) and the numbers itself for the regression. The input for the training of this model could be the columns to categorize the move. The problem is, that they are no clear correlation between thees columns $|C| = 0.11$ where C is the correlation.