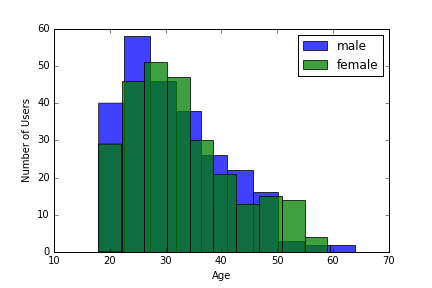
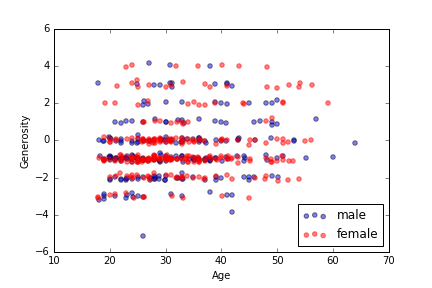
Abhilasha Deka and Hoa Ly

Introduction to Data Science

 The TSV file of [Odnoklassniki.ru](http://odnoklassniki.ru/) is made of 3500 users. There were 2,977 null values in the response grade (OGrade), leaving us with 523 usable records. After dropping the null values, there are 253 male and 270 female users.

As the histogram dictates, the website is most popular for males between the ages 22-25 and females between the ages 27-30. The generosity we calculated for each user is pretty much uniform as shown in the following scatter plot. Using normally distributed random displacement to both X and Y coordinates of each point helped displace the points a little bit, making it easier to see more points:



Before splitting the data set for the regression models, we replaced the Ms and Fs in the gender column into 0s and 1s respectively. We split the data randomly: 70% going into the training set and 30% going into the testing or validation set. The score of the linear model is 42.7%. In the second regression model, we used indicators for the response grades, meaning the predictions returned six arrays 0s and 1s. We kept a predicted score only if all the six models agreed on it. If they did not, the predicted OGrade was left 0. The scores on the six logistic models are as follows: 76.9%, 92%,95%, 98.7%, 83.8% and 97.9%.

We recommend using the logistic models to make the predictions, for they are more accurate than the linear model.