**About the data:**

Most of the data are categorical variables where they only take a few discrete values. R does a pretty good job of turning the words into factors and saving us some work.

There are a couple of things we may want to do before any analysis though. To prepare the data we may want to:

1. Change Dependents, Cred\_here, installment rate, and at address since to categorical (factor) variables.
2. Scale the two continuous time variables age and duration so the mean is 0 and standard deviation is 1.

These would just be proper things to do but may not make the algorithm perform better.

Another consideration is the “real” size of the data set. One way to interpret categorical variables where there are more than 2 options is via “one hot encoding.” One hot encoding creates columns for each type of factor and is labeled 1 where it is true and a 0 otherwise. I believe these are essentially factors in R. If we assume each of the options in each of the categorical variables is a separate variable, there are about 45 independent variables. This leaves ~18 observations/ variable which is not that many. This observations/ variable ratio would shrink if we use a train/ test set.

By looking at summary of the data, we can see many variables are not well balanced (more of 1 class than another).

Variable selection should be useful later on.

**Algorithms:**

A quick look at the prediction variable cred\_approved shows that about 70% of the data is “yes”. Basically without any algorithm, if we just guessed “yes” we would be right 70% of the time. We can think of this as a baseline or “number to beat” for our algorithm.

I ran a quick logistic regression on all of the data and got a cross validation score of 83%. When I used a test train set, I got an accuracy of 90% on the test set.

Next we should use some variable selection techniques on the logistic regression and possibly try using a random forest and playing with those parameters.