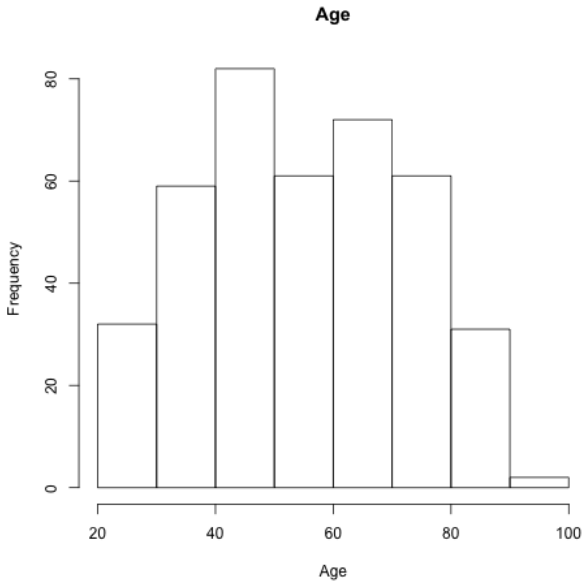
Data

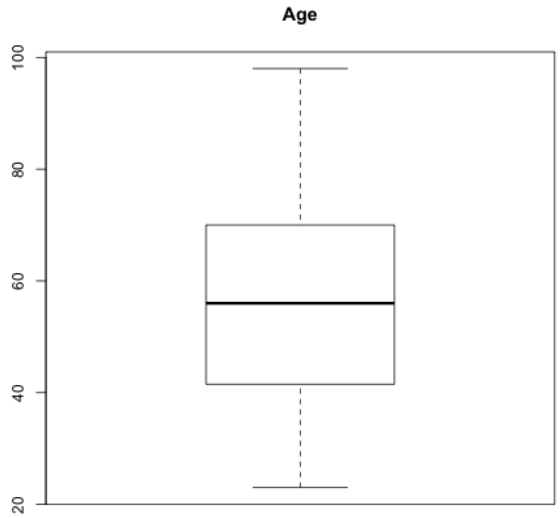
The `Credit.csv` dataset consists of `balance` observations (average credit card debt) for 400 different individuals, as well as observations for a number of quantitative and qualitative variables detailed below:

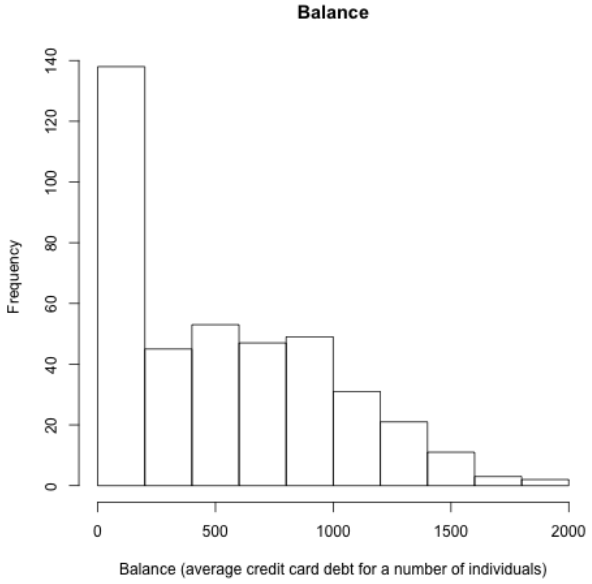
Quantitative Variables

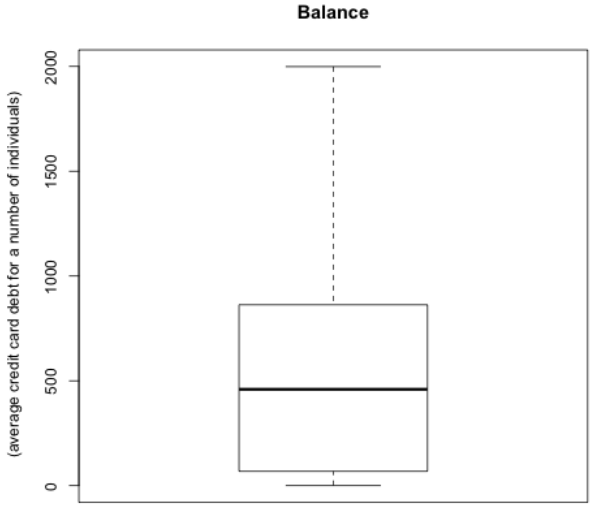
- `age`
- `cards` (number of credit cards)
- `education` (years of education)
- `income`(in thousands of dollars)
- `limit` (credit limit)
- `rating` (credit rating)

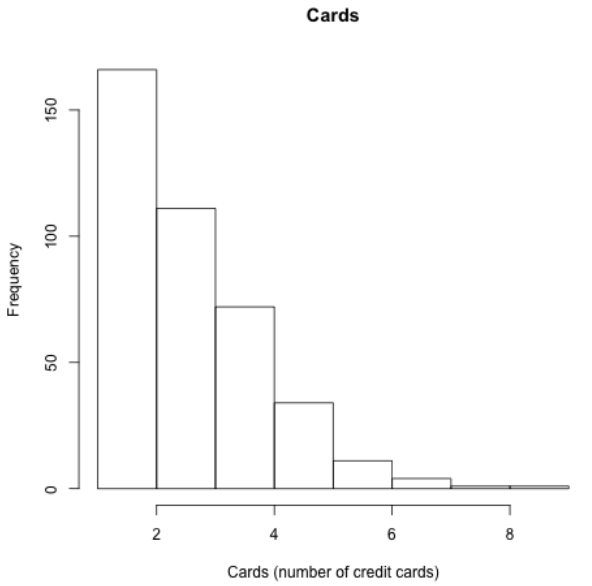
A very general overview of the distribution for each variable is provided in the histograms and boxplots below:

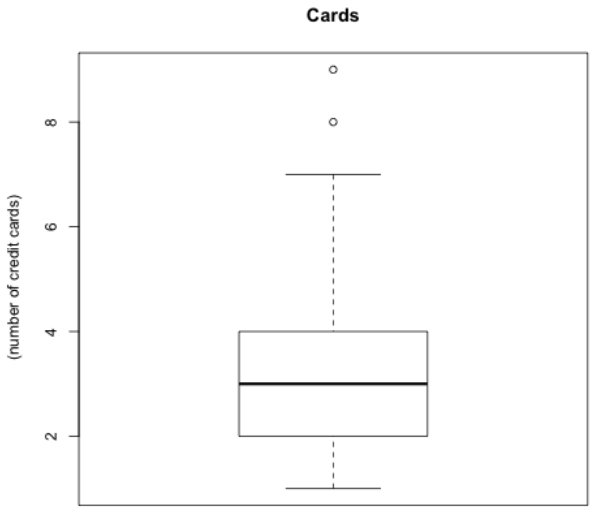


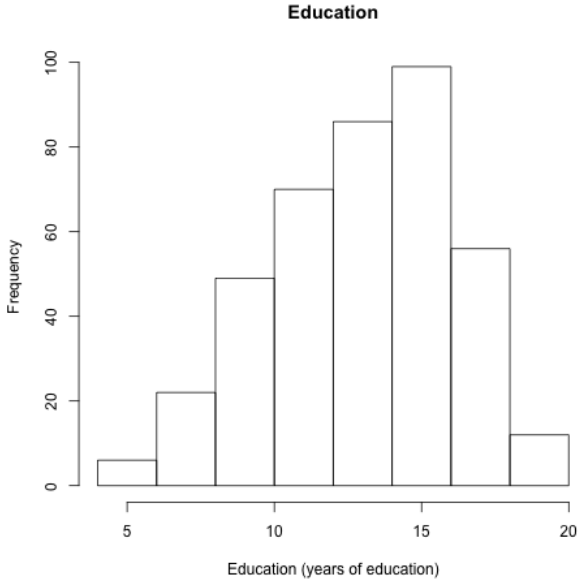


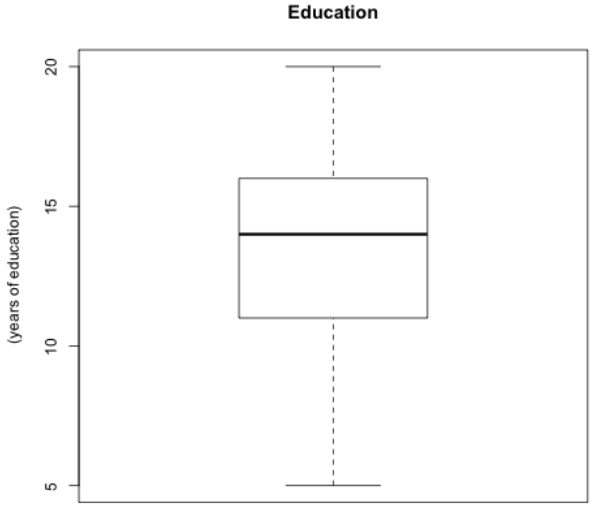

Balance
Frequency
0
20
40
60
80
100
120
140
0
500
1000
1500
2000
Balance (average credit card debt for a number of individuals)

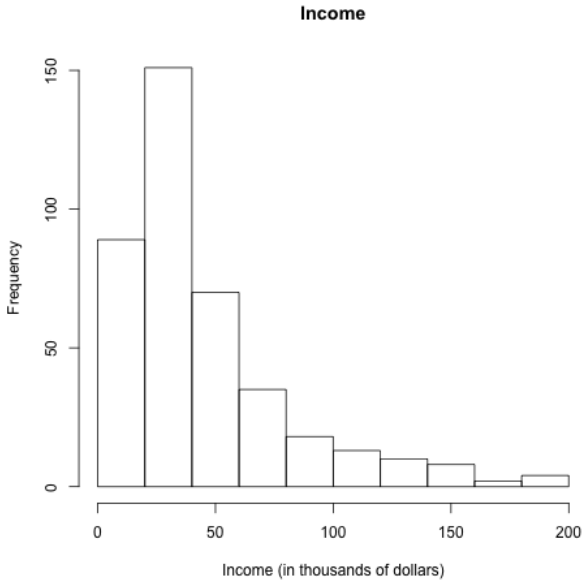

Balance
(average credit card debt for a number of individuals)
0
500
1000
1500
2000

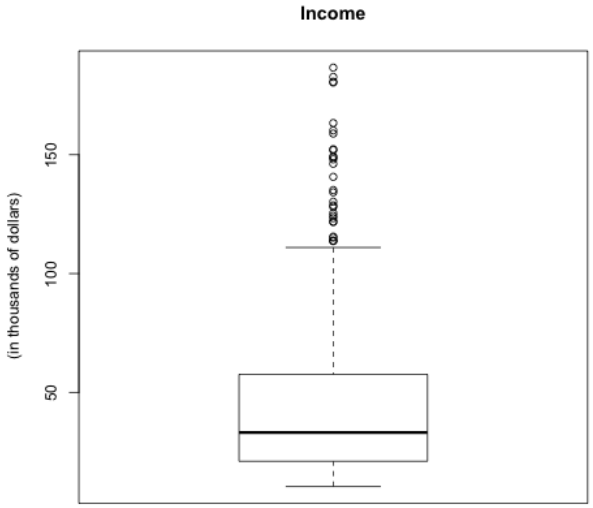

Cards
Frequency
0
50
100
150
2
4
6
8
Cards (number of credit cards)

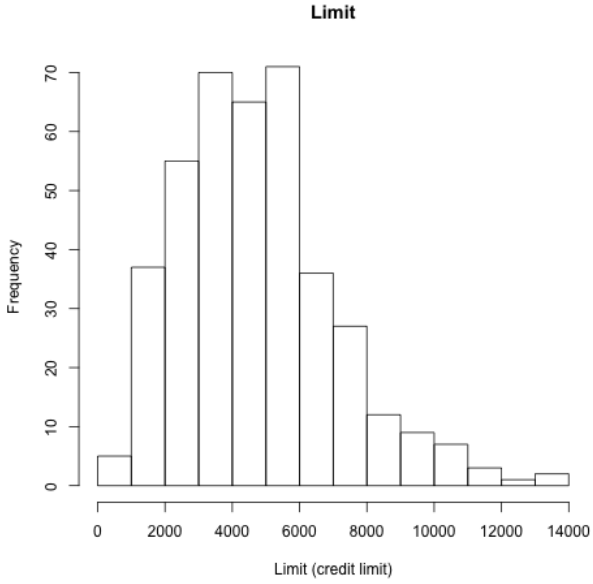

Cards
(number of credit cards)
2
4
6
8

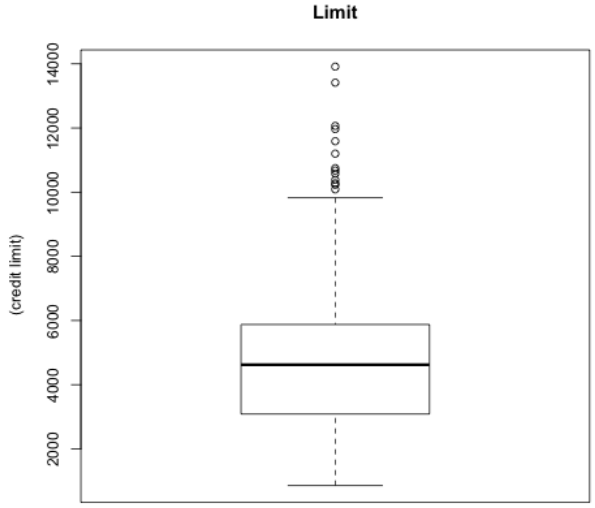

Education
Frequency
0
20
40
60
80
100
5
10
15
20
Education (years of education)

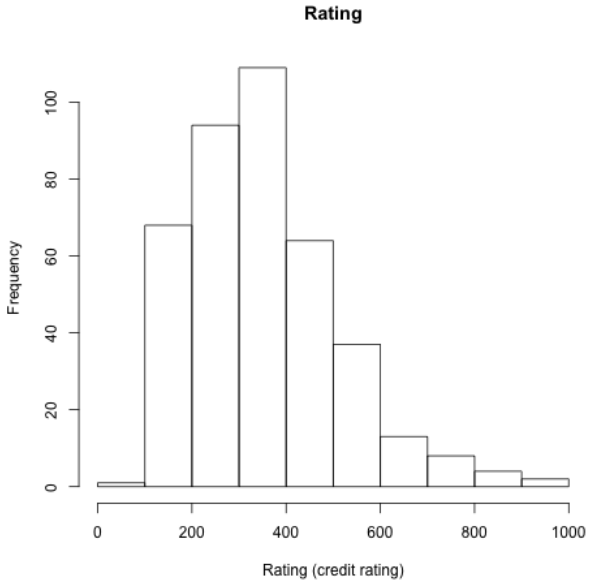

Education
(years of education)
5
10
15
20

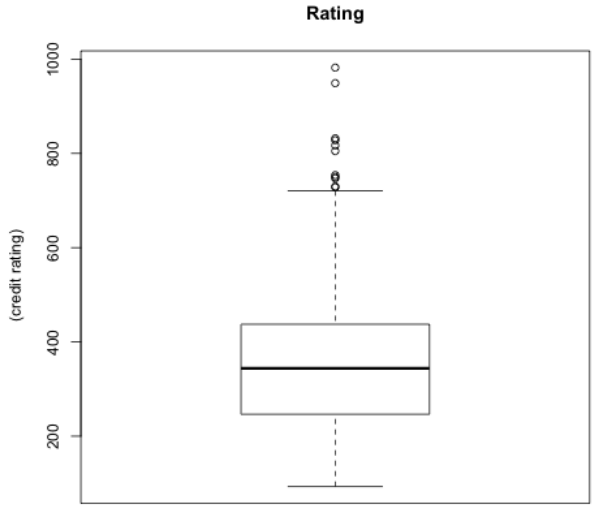

Income
Frequency
0
50
100
150
0
50
100
150
200
Income (in thousands of dollars)


Income
(in thousands of dollars)
50
100
150









A scatterplot matrix for all quantitative variables and a matrix of correlations are displayed below:

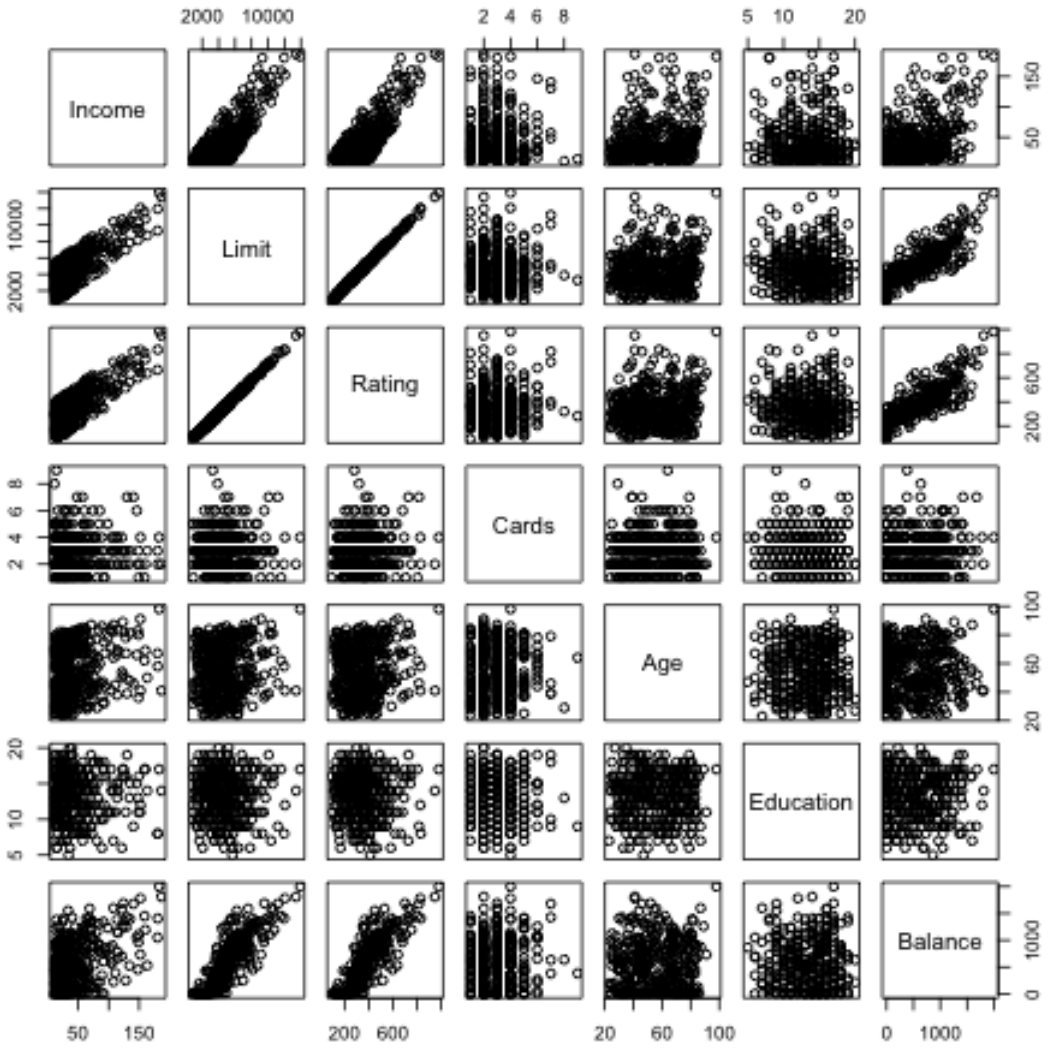


Table 1: Matrix of Correlations for all Quantitative Variables

	Income	Limit	Rating	Cards	Age	Education	Balance
Income	1.0000	0.7921	0.7914	-0.0183	0.1753	-0.0277	0.4637
Limit	0.7921	1.0000	0.9969	0.0102	0.1009	-0.0235	0.8617
Rating	0.7914	0.9969	1.0000	0.0532	0.1032	-0.0301	0.8636
Cards	-0.0183	0.0102	0.0532	1.0000	0.0429	-0.0511	0.0865
Age	0.1753	0.1009	0.1032	0.0429	1.0000	0.0036	0.0018
Education	-0.0277	-0.0235	-0.0301	-0.0511	0.0036	1.0000	-0.0081
Balance	0.4637	0.8617	0.8636	0.0865	0.0018	-0.0081	1.0000

Qualitative Variables

- `gender` (m/f)
- `student` (yes/no)

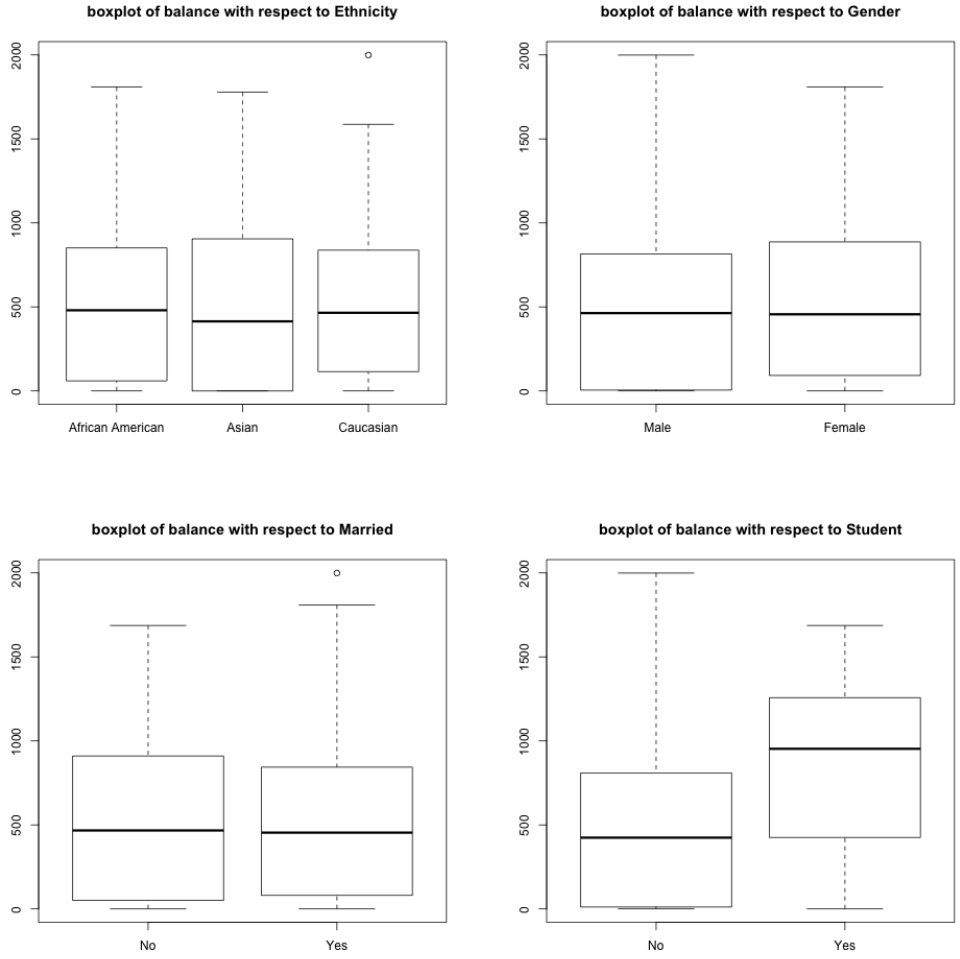
- `married` (yes/no)
- `ethnicity`(caucasian/asian/african american)

All of these variables are factors with two or three levels, so the best way to explore their distribution is through frequency tables and plots that display `Balance` with respect to a qualitative variable.

Below is a proportional frequency table for all qualitative variables:

```
##                         Ethnicity African American  Asian Caucasian
## Gender Student Married
##  Male  No      No                          0.0450 0.0400    0.0825
##                Yes                         0.0650 0.0675    0.1425
##        Yes     No                          0.0050 0.0075    0.0100
##                Yes                         0.0075 0.0025    0.0075
## Female No      No                          0.0625 0.0250    0.0825
##                Yes                         0.0500 0.0900    0.1475
##        Yes     No                          0.0050 0.0075    0.0150
##                Yes                         0.0075 0.0150    0.0100
```

Below are conditional boxplots of `Balance` with respect to each qualitative variable:



Data Pre-Processing

Scaling

Using data with multiple predictors presents the problem of different scalings. For example, a person's age is typically between 0 and 100, while a person's income is at a much larger scale. In order to use both as predictors for a response, it is essential to standardize their ranges. To do this, we use the R function "scale", which subtracts from each vector its mean and divides each vector by its standard deviation.

Training and Testing Sets

Oftentimes fitting a model to data is for the purpose of predicting future observations. Given a full dataset we can simulate "past" and "future" observations by dividing the data into *training* and *testing* sets. A model is fit, or "trained", to the training set. The testing set becomes "new" data that the model tries to predict. Since we have the response values of our testing set, we can calculate the difference between our predictions and our true values to assess model. Here is what we did:

```
# take a random sample whose size is 75% of the number of observations of our data (number_rows):

# compute total number of rows of our data; should be 400
number_rows <- nrow(scaled_credit)

# set seed
set.seed(10)

# take 75% of the rows as training by randomly sampling from number_rows
training_rows <- sample(1:number_rows, 0.75*number_rows,
                        replace = F)

# assign training and testing sets based on these rows

# y is our response, "Balance".
y <- as.matrix(scaled_credit$Balance)

# x is our predictors
x <- as.matrix(scaled_credit[,-ncol(scaled_credit)]) #removing "balance"

# split into training and testing for x and y
y_train <- y[training_rows,]
x_train <- x[training_rows,]
y_test <- y[-training_rows,]
x_test <- x[-training_rows,]
```