Predictive Modeling Process

Timothy Park, Gary Nguyen

November 4th, 2016

Abstract

Incorporating all the resources learned in Stat 159, we will dive deeper into data analytics by applying more advanced statistical learning methods to model a data set listing credit information of 400 randomly selected individuals. This project will reproduce the methods performed in Chapter 5 and 6 from the textbook, *An Introduction to Statistical Learning" by James et al.* These sections cover cross-validation techniques and linear model selection/regularization methods, both of which are necessary in locating the optimal model for a given data set. The credit data set is sourced from the following url: `http://www-bcf.usc.edu/~gareth/ISL/Credit.csv`. Collaborating with a partner was mandatory, which allowed us to become more comfortable with Github and Git processes.

Introduction

The motivation for this project is to perform a predictive modeling process to the Credit data set. The first step in data analysis is to understand and search for key characteristics in the working data set. Gaining insight into the relationships among the multiple variables is essential in subset selection and prior knowledge into locating which variables are insignificant. After completing the section covering exploratory data analysis, the Credit data set will be processed to be prepared for model fitting, which includes centering/standardizing each variable column and dealing with categorical variables. The processed Credit data set is then split into training and test sets and thereafter, five different regression models are fitted to select the best model in accurately predicting the response variable Balance.

Data

The Credit data set presents information on 11 different characteristics of 400 bank accounts. Of these 11 variables, 7 of them are quantitative and the remaining 4 are qualitative. Below contains a brief description of each variable.

Quantitative

- `Income`: customer's income
- `Limit`: customer's credit limit
- `Rating`: customer's credit rating
- `Cards`: number of credit cards
- `Age`: customer's age
- `Education`: number of years in education
- `Balance`: current balance in the customer's bank

Qualitative

- `Gender`: customer's gender (factor with two levels - Male/Female)
- `Student`: customer's current student status (factor with 2 levels - Yes/No)
- `Married`: customer's current marital status (factor with 2 levels - Yes/No)
- `Ethnicity`: customer's ethnicity (factor with 3 levels - Asian/Caucasian/AfricanAmerican)

For this project, we will see which model we fit will most closely predict with the quantitative variable `Balance`. Thus, we can consider the Balance variable as a response and the other variables as predictors. Since there are 11 variables and 400 customers, the Credit data set has a 400 x 11 dimension. In the next section, in order to assess the accuracy of each regression fit, the Credit data set must be split into two, a training set and a test set.

Methods

Exploratory Data Analysis

Descriptive statistics of `Credit` were explored by displaying their summary statistics. All quantitative variables in `Credit` were plotted with the following: histograms, boxplots, and frequency tables. All qualitative variables in `Credit` were visually displayed by conditional boxplots. With both quantitative and qualitative variables, a scatterplot matrix, matrix of correlations, and anova outputs were produced. All of these images and descriptions are stored in the data directory of this project.

Training and Test Sets

To split the Credit data set, the `test_split_script.R` was used. In the source code, 300 rows were randomly selected and stored into a csv file called `train_credit.csv`. The remaining 100 rows were written into a csv file called `test_credit.csv`. The regression models will be trained onto the 300-row data set and their fitted predictions will be measured by how much they deviate from the test set predictions.

Pre-modeling Data Processing

Since the splitted data sets were still raw, pre-processing techniques were done to properly prepare for modeling. The four categorical variables cannot be interpreted by the model functions in R, so it is better to convert them to numerical values. The categorical variables with 2 levels can be converted to a dummy variable while the variable with 3 levels such as Ethnicity needs two binary columns or two dummy variables. The function `model.matrix()` does this automatically, converting the data set into a 400 x 12 data frame.

In addition, the coefficients after fitting each model may be unexpectedly altered by the scaling of each variable. Thus, it is proper to standardize and center each column so that comparisons among each variable are done fairly and that the beta coefficients do not blow up.

Regression Models

The following lists the 5 regressions that will be applied to the `Credit` data set:

- Ordinary Least Squares Regression
- Ridge Regression
- Lasso Regression
- Principal Components Regression
- Partial Least Squares Regression

We start with the ordinary least squares regression model to fit onto the Credit data set. After computing the test mean standard error, this statistic will serve as a basis to compare the accuracies of the other models. The OLS regression will be performed by the `ols_script.R`, using the `lm()` function to model Balance over the 11 predictors (with the 2 dummy variables). From this model object, we then use the `predict()` function on the test predictors and calculate its mean squared error by comparing the test's values of `Balance`

Another set of more complex models considers shrinkage methods. These models, ridge regression and lasso regression, shrink the coeffcients to 0 by setting a constraint to the beta coefficients, having lambda as the tuning parameter. The difference in ridge and lasso is the formula inside of the sum of the constraint. The following provides a step-by-step instruction on how these methods work:

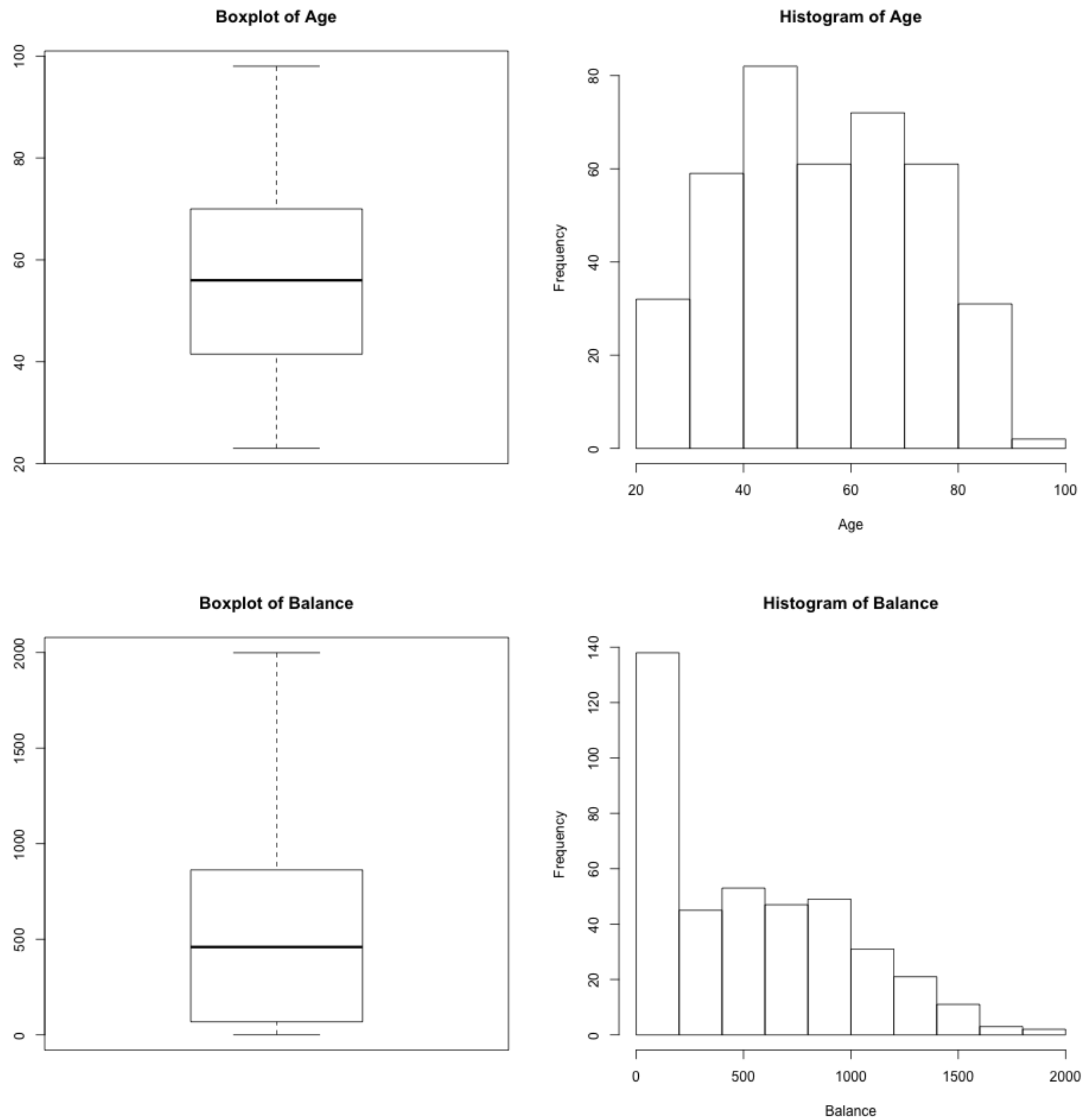
- Load the library called `glmnet` and read in the training and test sets.
- Split the the training set by subsetting every column except Balance into x and Balance into y.
- Perform the regression using the function `cv.glmnet()`. A 10-fold cross validation technique is applied by setting `nfold = 10`. Since we already standardized and centered the variables, the argument `intercept` and `standardize` are set to `FALSE`. By default, `cv.glmnet()` incorporates all lambdas, but in this project, we will assume that `lambda = 10^seq(10, -2, length = 100)`. The alpha argument differentiates between the two models: ridge regression is performed when alpha = 0 and lasso regression is performed when alpha = 1.
- The model object after fitting one of these shrinkage regression models displays a unique model for each lambda gived from `grid`. We use this object and retrieve the lowest lambda by `model_object$min.lambda`.
- A plot of the cross-validation errors can be done easily by using the `plot()` function. The lowest lambda is saved onto the text file and the plot is saved as a png image.
- Use the model object and the `predict()` function to make predictions on the predictor variables of the test set. The mean squared error is then calculated by comparing the predictions and the actual response values. The MSE is stored into the text file as well.
- Using the lowest lambda, the optimal model is refitted onto the whole data set `scaled_credit.csv` and the model's coefficients are then saved into the same text file.

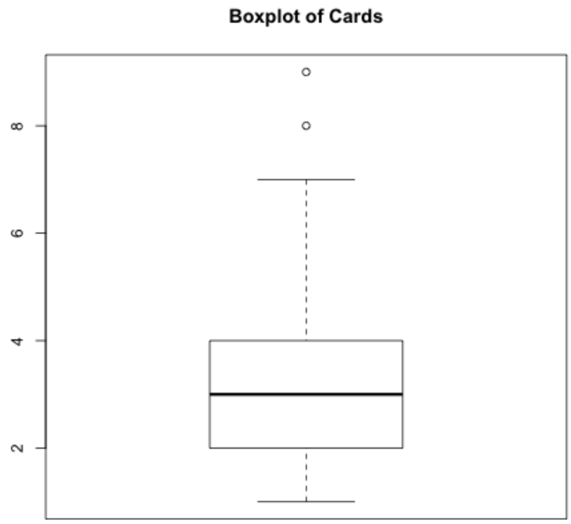
The last two regression models deal with dimension reduction by locating which predictors explain the most variance of Balance. Knowing how many predictors or components to use depends on which set of components has the lowest MSE. Principal Components Regression and Partial Least Squares Regression are run by the `pcr_script.R` and `plsr_script.R`.

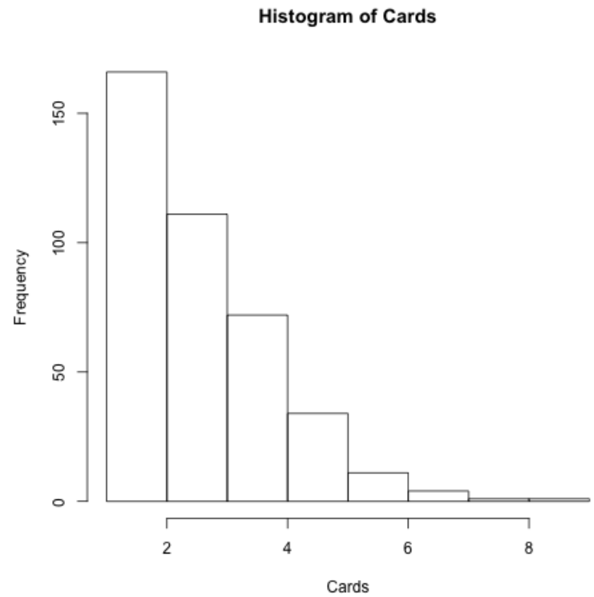
- Load the library called `pls` and read in the training and test sets.
- Use the model fitting functions `pcr()` and `plsr()` to run the regressions, setting `formula = Balance ~.` and `validation = 'CV'`. The data should only cover the training set.
- The best model in the cross-validations is found using `which.min(model_object$validation$PRESS)`. This value is stored into a text file.
- Plot the model using the `plot()` function and save it as a png image.
- Use the test set to predict the response Balance and compare it to the real Balance observations, thereafter calculating the test mean squared error of the model. The MSE is stored in a text file.
- Fit the model onto the `scaled_credit.csv` with `ncomp` equal to the minimum principal components found before in cross-validation since this model will lead to the lowest MSE. The coefficients of this official model is then saved into a text file.

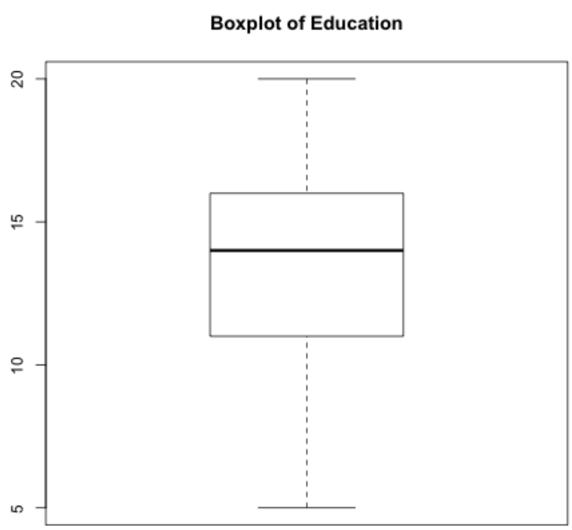
These five regression methods are then compared by their MSEs, and the model with the lowest MSE is considered the best. Other model statistics can be used for comparisons, but for this project, the MSE is the most convenient.

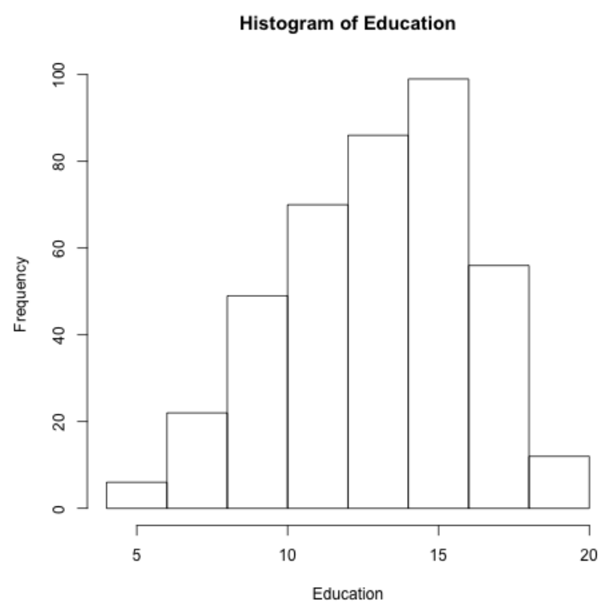
Analysis# Results

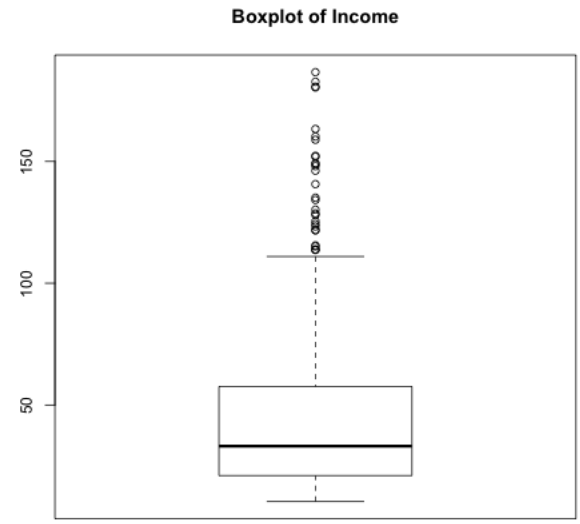


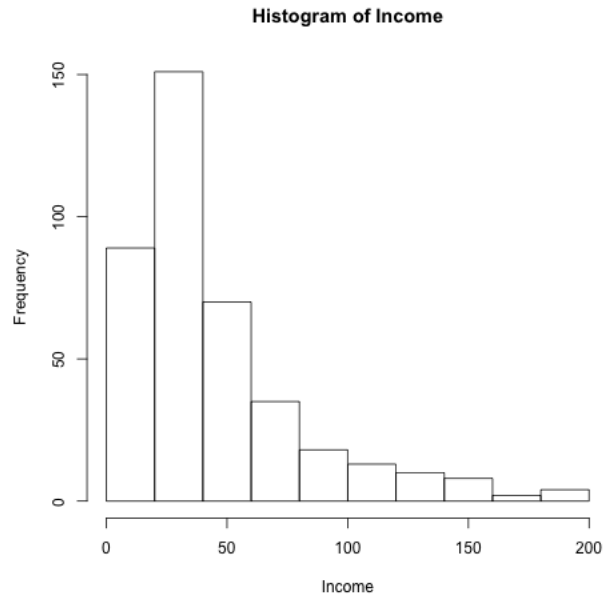

Boxplot of Cards
2
4
6
8

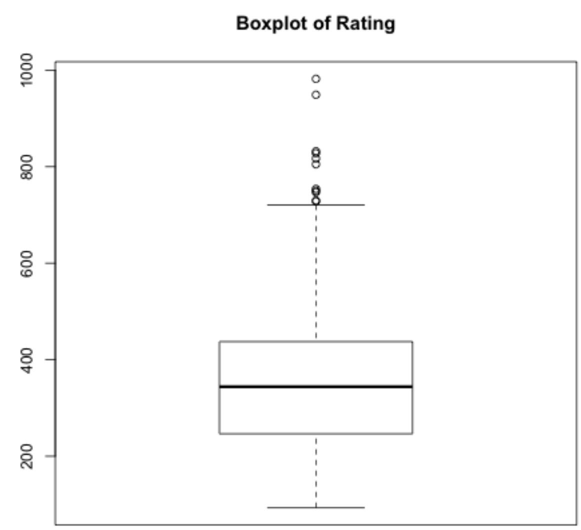

Histogram of Cards
Frequency
0
50
100
150
2
4
6
8
Cards

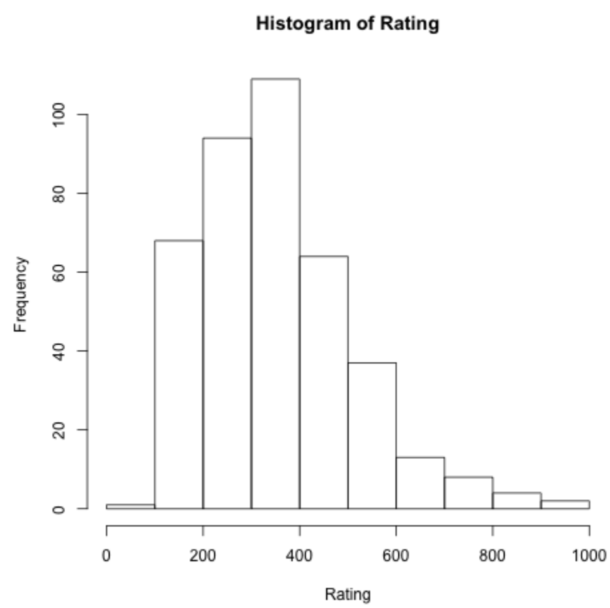

Boxplot of Education
5
10
15
20

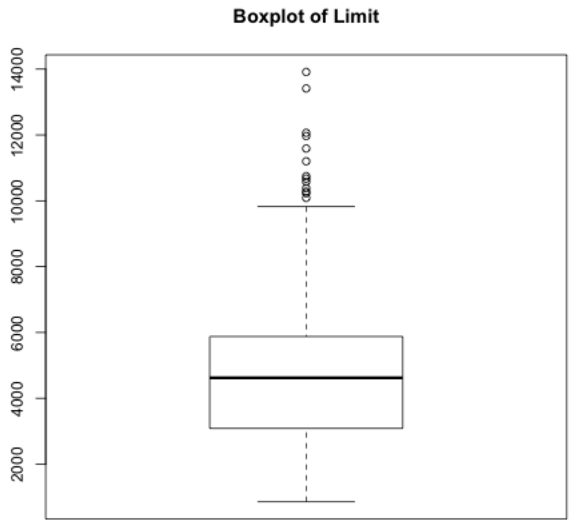

Histogram of Education
Frequency
0
20
40
60
80
100
5
10
15
20
Education

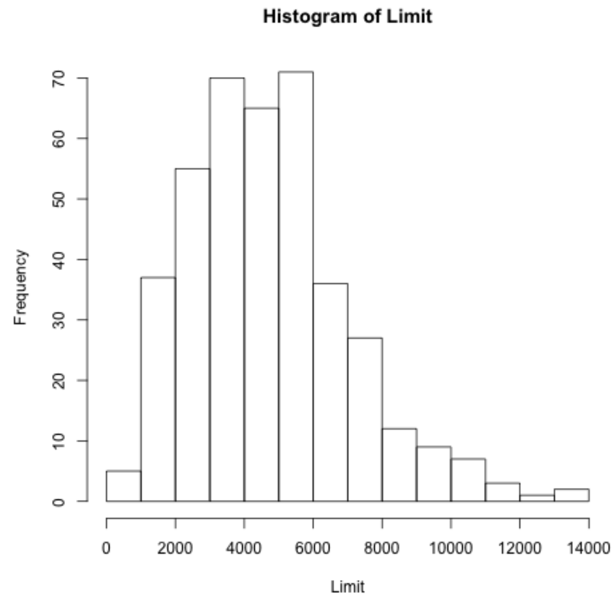

Boxplot of Income
50
100
150

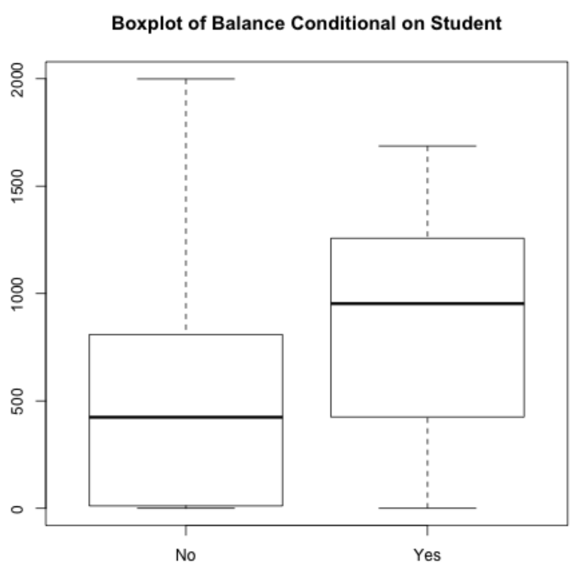

Histogram of Income
Frequency
0
50
100
150
Income
0
50
100
150
200

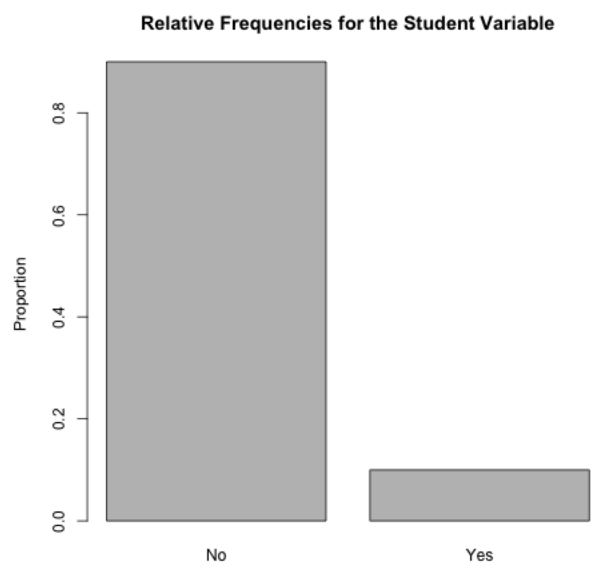

Boxplot of Rating
200
400
600
800
1000

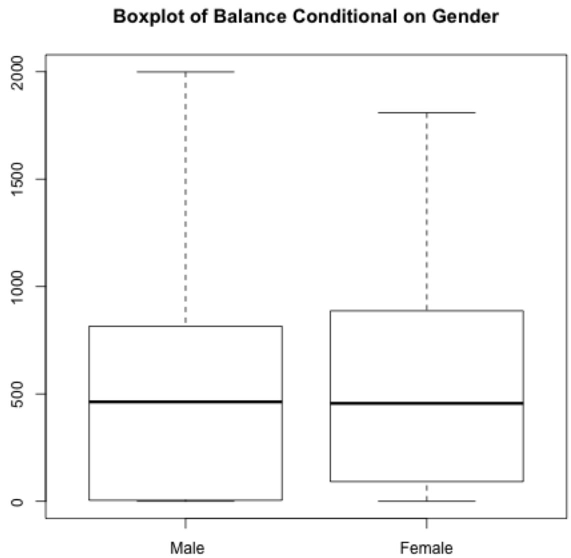

Histogram of Rating
Frequency
0
20
40
60
80
100
Rating
0
200
400
600
800
1000

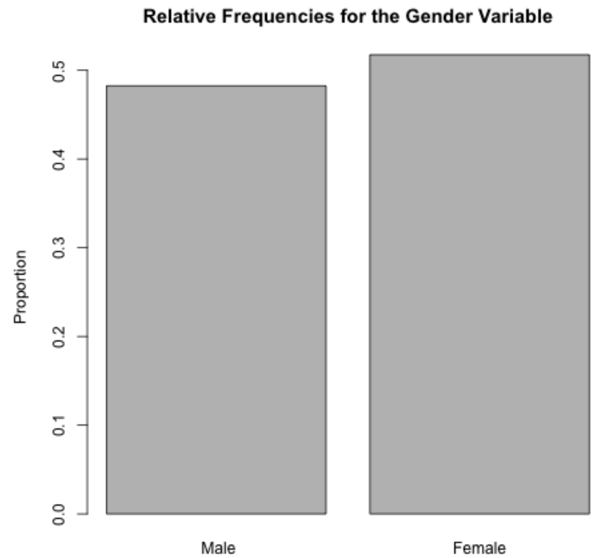

Boxplot of Limit
2000
4000
6000
8000
10000
12000
14000

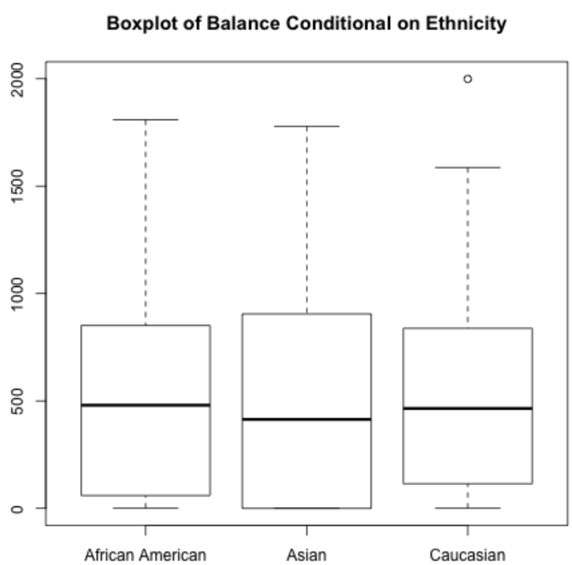

Histogram of Limit
Frequency
0
10
20
30
40
50
60
70
0
2000
4000
6000
8000
10000
12000
14000
Limit

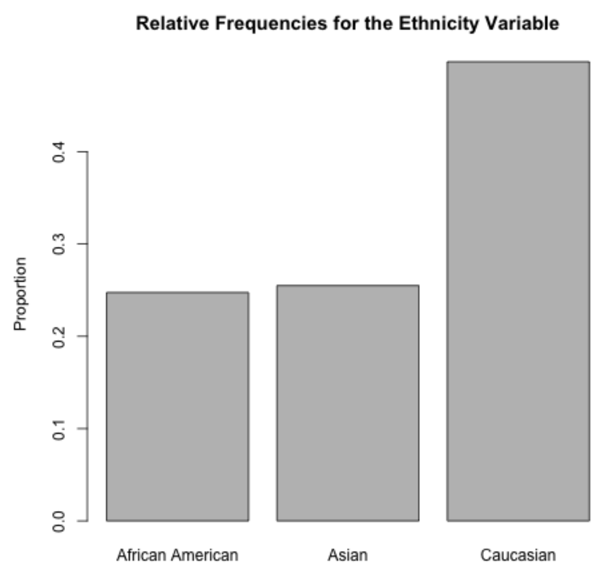

Boxplot of Balance Conditional on Student
0
500
1000
1500
2000
No
Yes

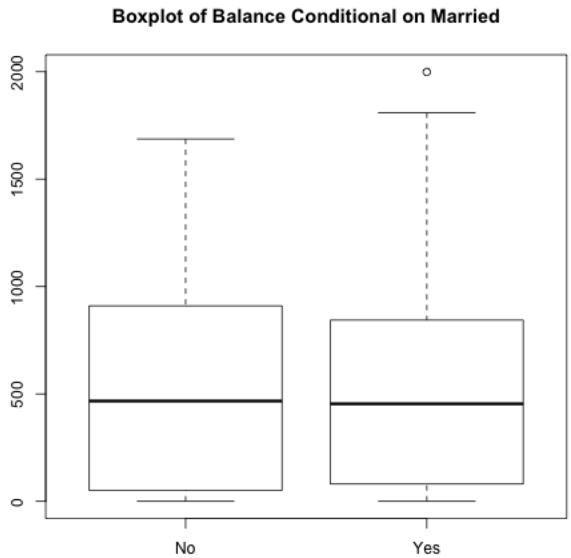

Relative Frequencies for the Student Variable
Proportion
0.0
0.2
0.4
0.6
0.8
No
Yes

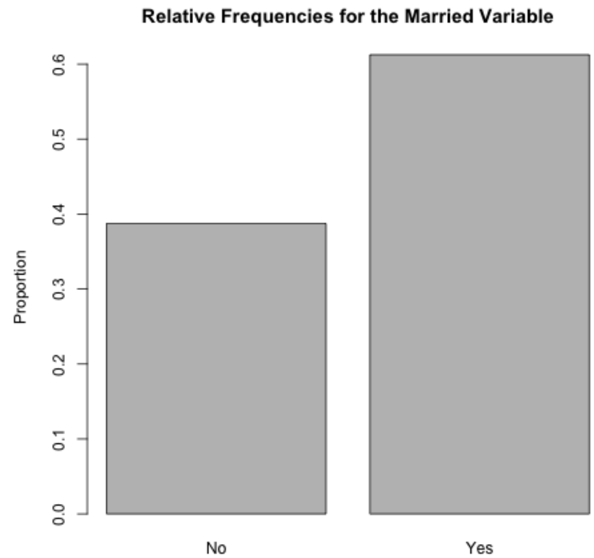

Boxplot of Balance Conditional on Gender
0
500
1000
1500
2000
Male
Female

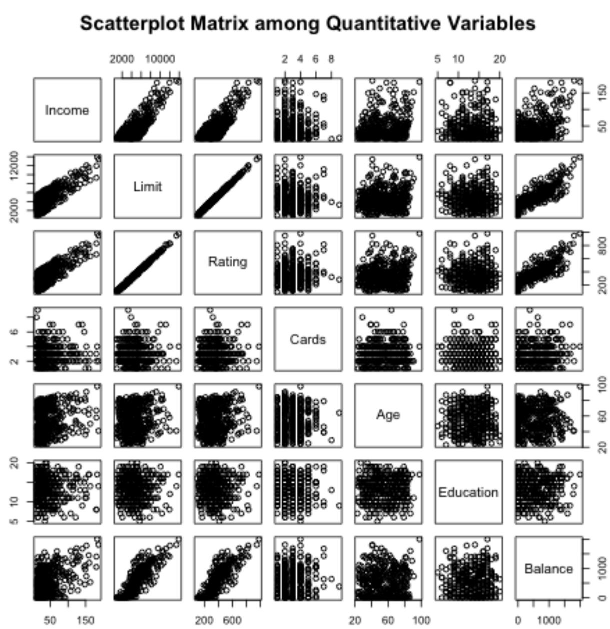

Relative Frequencies for the Gender Variable
Proportion
0.0
0.1
0.2
0.3
0.4
0.5
Male
Female

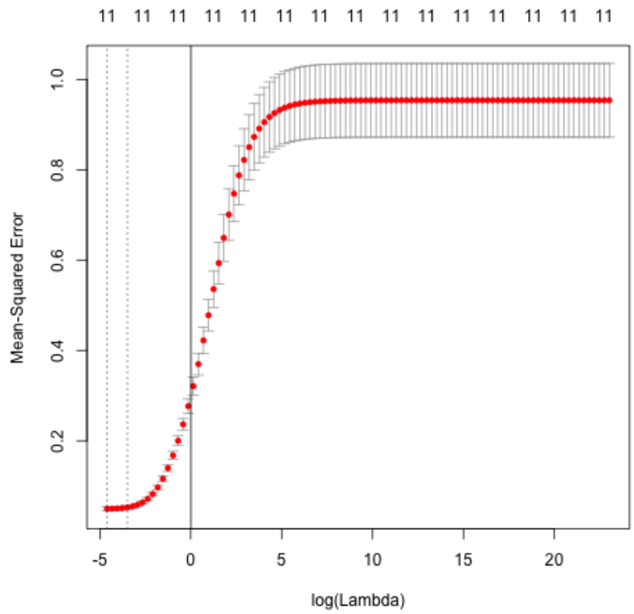

Boxplot of Balance Conditional on Ethnicity
0
500
1000
1500
2000
African American
Asian
Caucasian

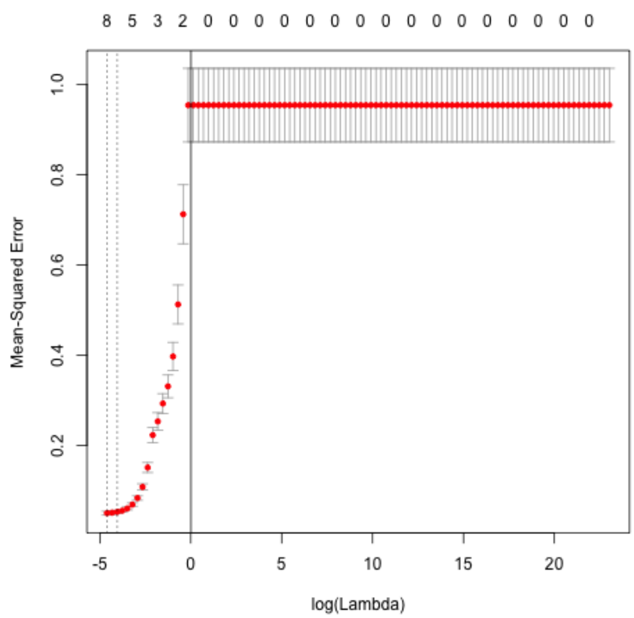

Relative Frequencies for the Ethnicity Variable
Proportion
0.0
0.1
0.2
0.3
0.4
African American
Asian
Caucasian

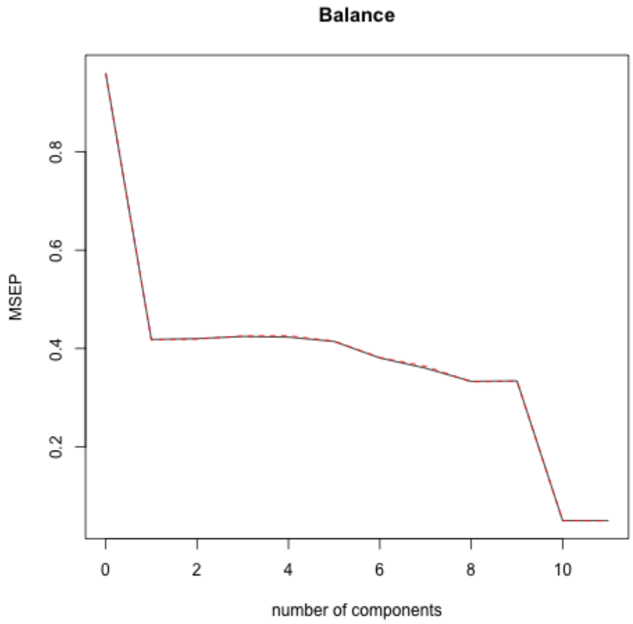

Boxplot of Balance Conditional on Married
0
500
1000
1500
2000
No
Yes

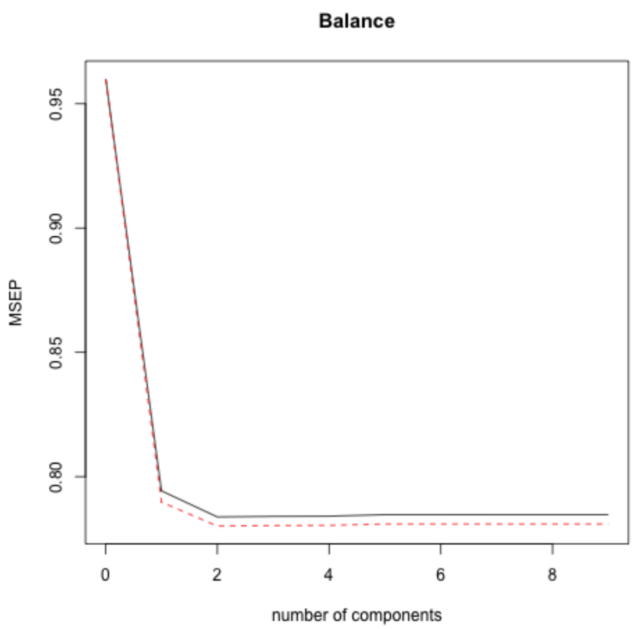

Relative Frequencies for the Married Variable
Proportion
0.0
0.1
0.2
0.3
0.4
0.5
0.6
No
Yes


Scatterplot Matrix among Quantitative Variables
Income
Limit
Rating
Cards
Age
Education
Balance


11 11 11 11 11 11 11 11 11 11 11 11 11 11 11
Mean-Squared Error
0.2
0.4
0.6
0.8
1.0
-5 0 5 10 15 20
log(Lambda)


8 5 3 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Mean-Squared Error
0.2
0.4
0.6
0.8
1.0
-5 0 5 10 15 20
log(Lambda)


Balance
MSEP
0.2
0.4
0.6
0.8
0
2
4
6
8
10
number of components


Balance
MSEP
0.80
0.85
0.90
0.95
0
2
4
6
8
number of components

Data