Stat 159 Project 2

Author: Lydia Maher, Ziao Liu

Abstract

In this project, we are doing a typical data analysis cycle which contains getting raw unstructured datasets, data cleaning and processment, exploratory data analysis, modeling and tuning parameters, visualization of results, report and presentation.

We are replicating work on chapter 6: Linear Model Selection and Regularization (from "An Introduction to Statistical Learning" by James et al) with the dataset "Credit". In detail, we are trying to predict "Balance", which is the response variable, from 11 predictors including Income, Limit, Rating, Cards, Age, Education, Gender, Student, Maaried, Ethinicity.

Since it is a continuous prediction, we use five algorithms: Least Square Regression, Ridge Regression, Lasso Regression, Principal Components Regression, and Partial Least Squares Regression. From our output, we found that Lasso Regression performs best, followed by Ridge Regression. Dimension Reduction Regression, in this case, do not perform as well as Shrinkage Methods.

Introduction

Since we are predicting for a continuous output, we choose Least Square Regression as our base model. Then we choose two Shrinkage Methods (Ridge Regression and Lasso Regression) and two Dimension Reduction Methods (Principal Components Regression and Partial Least Squares Regression) to compare the results with our base model.

For parameter choosing, we use 10-fold Cross Validation to choose minimum lambda for Shrinkage Methods and minimum validation components for Dimension Reduction Methods. Then we compare results by splitting dataset into training set and test set and calculating Mean Square Error on test dataset. Finally we calculate the coefficients in each of five methods used for prediction to see which factor matters most in prediction.

Data

In this project, we are using the "Credit" dataset. The dataset contains 12 columns each representing a variable and 400 rows each representing a unique person.

Predictors: 11 predictors including Income, Limit, Rating, Cards, Age, Education, Gender, Student, Maaried, Ethinicity. Response: 1 response variable namely Balance

Methods

Algorithms For Prediction

Least Squares Regression:

One way to predict sales based on multiple predictors is to fit separate linear regression model for each predictor, and the other way is to extend the linear model so that it can directly accommodate multiple predictors. For a one predictor case, we can just use a simple linear regression with Balance = Beta-0 + Beta-1*Income where Beta-0 is the intercept and Beta-1 is the slope of the regreession line, and replace Income with the other predictors to get different results. For a multiple linear regression, we need more coefficients to represent the data properly. We can formulate a linear regression with all predictors via the least square criterion which is the same criterion in the simple linear regression.

Ridge Regression:

Ridge regression is very similar to least squares, except that the coefficients are estimated by minimizing a slightly different quantity. Ridge regression seeks coefficients that fit the data well by making RSS small which is the same intuition as least square regression. The difference between Ridge Regression and Least Regression is that Ridge Regression adds a l2 penalty to the RSS which is small when Beta is close to zero and shrinks the estimate of Beta to zero. Ridge regression's advantage over least squares is rooted in the bias-variance trade-off.

Lasso Regression:

Lasso Regression shares the same intuition as Ridge Regression which is adding a penalty to RSS to shrink coefficients and make better prediction. The difference between Lasso and Ridge is that Lasso uses a l1 penalty rather than a l2 penalty. In this case, Lasso is able to shrink coefficients to exactly zero and would return only a subset of predictors.

Principal Components Regression:

Principal components analysis (PCA) is a common approach for deriving a low-dimensional set of features from a large set of variables. Principal Components Regression is a dimension reduction technique for regression. The principal components regression (PCR) approach involves constructing the first M principal components, Z1,. . . ,ZM, and then using these components as the predictors in a linear regression model that is fit using least squares. we assume that the directions in which X1,. . . , Xp show the most variation are the directions that are associated with Y.

Partial Least Squares Regression:

PCR has a major disadvantage that there is no guarantee that the directions that best explain the predictors will also be the best directions to use for predicting the response. Similarly, PLS is a dimension reduction method which first identifies a new set of features Z1,. . . ,ZM that are linear combinations of the original features, and then fits a linear model via least squares using these M new features. The difference between PLS and PCR is that PLS learns the new features in a supervised way which means that PLS find directions that help to explain both response and predictors.

Parameter Choosing for Algorithms

Cross Validation:

We use Cross Validation to choose best parameters that give us best prediction accuracy. Specifically we use 10-fold Cross Validation. In this case, The data set is divided into 10 subsets, and the holdout method is repeated 10 times. Each time, one of the 10 subsets is used as the test set and the other 9 subsets form a training set. Then the average error across all 10 trials is computed. The advantage of this method is that it matters less how the data gets divided. Every data point gets to be in a test set exactly once, and gets to be in a training set 9 times.

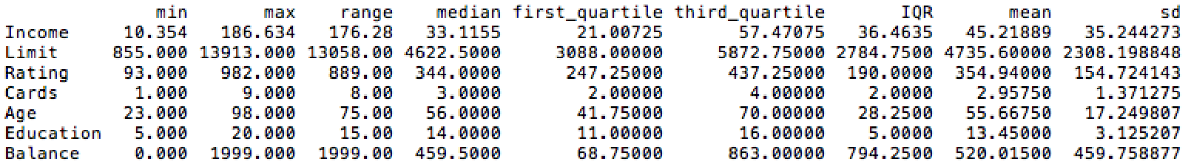
Analysis

Exploratry Data Analysis

Quantitative Variables

There are 7 quantitative variables in the dataset, including Income, Limit, Rating, Cards, Age, Education, Balance. For each of the variable, we computed minimum, maximum, range, median, quantile, IQR, mean, and standard deviation for each variable.

The following table is the summary statistics

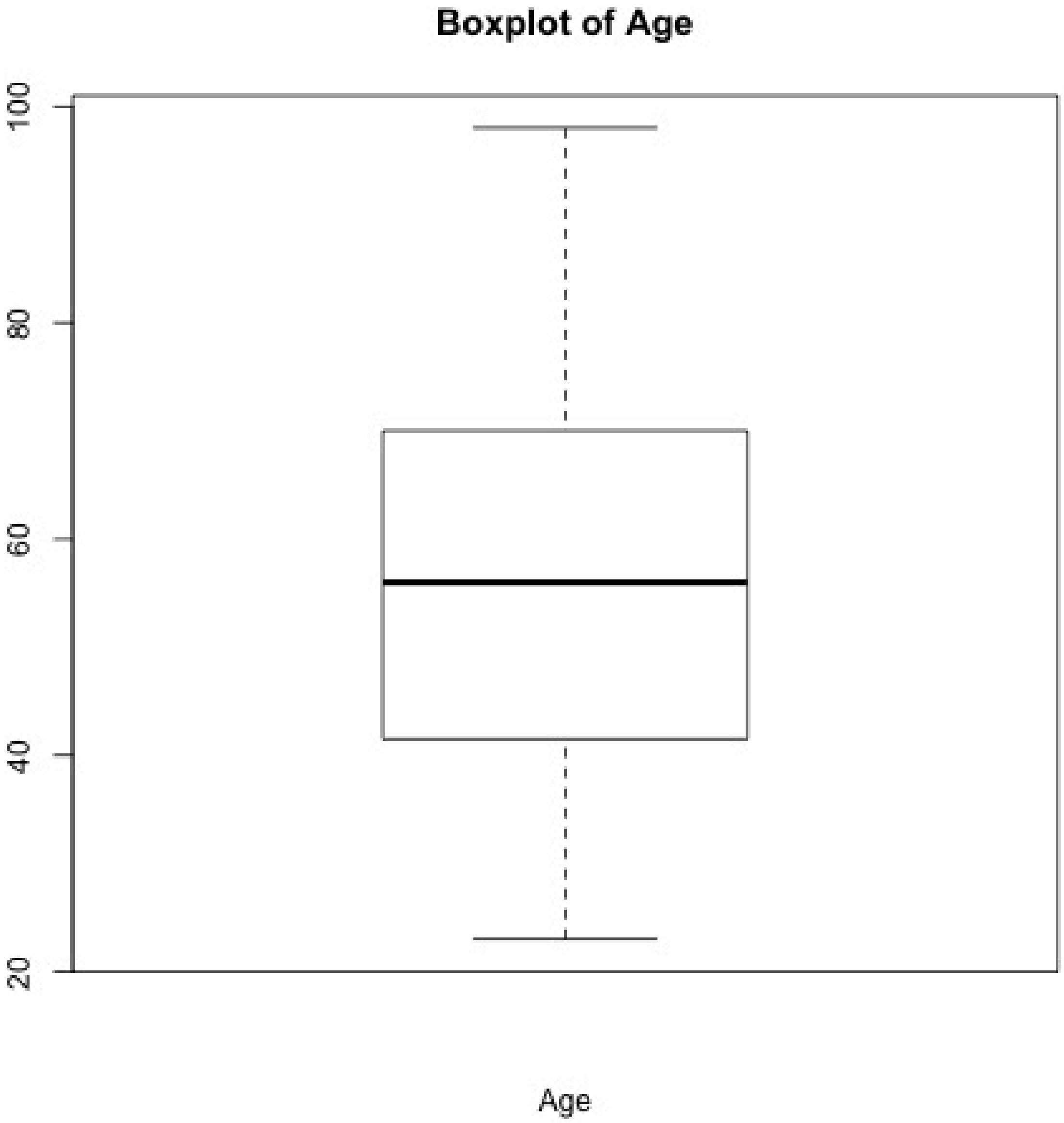


```
              min       max    range    median first_quartile third_quartile       IQR       mean          sd
Income     10.354   186.634   176.28   33.1155       21.00725       57.47075   36.4635   45.21889   35.244273
Limit     855.000 13913.000 13058.00 4622.5000     3088.00000     5872.75000 2784.7500 4735.60000 2308.198848
Rating     93.000   982.000   889.00  344.0000      247.25000      437.25000  190.0000  354.94000  154.724143
Cards       1.000     9.000     8.00    3.0000        2.00000        4.00000    2.0000    2.95750    1.371275
Age        23.000    98.000    75.00   56.0000       41.75000       70.00000   28.2500   55.66750   17.249807
Education   5.000    20.000    15.00   14.0000       11.00000       16.00000    5.0000   13.45000    3.125207
Balance     0.000  1999.000  1999.00  459.5000       68.75000      863.00000  794.2500  520.01500  459.758877
```

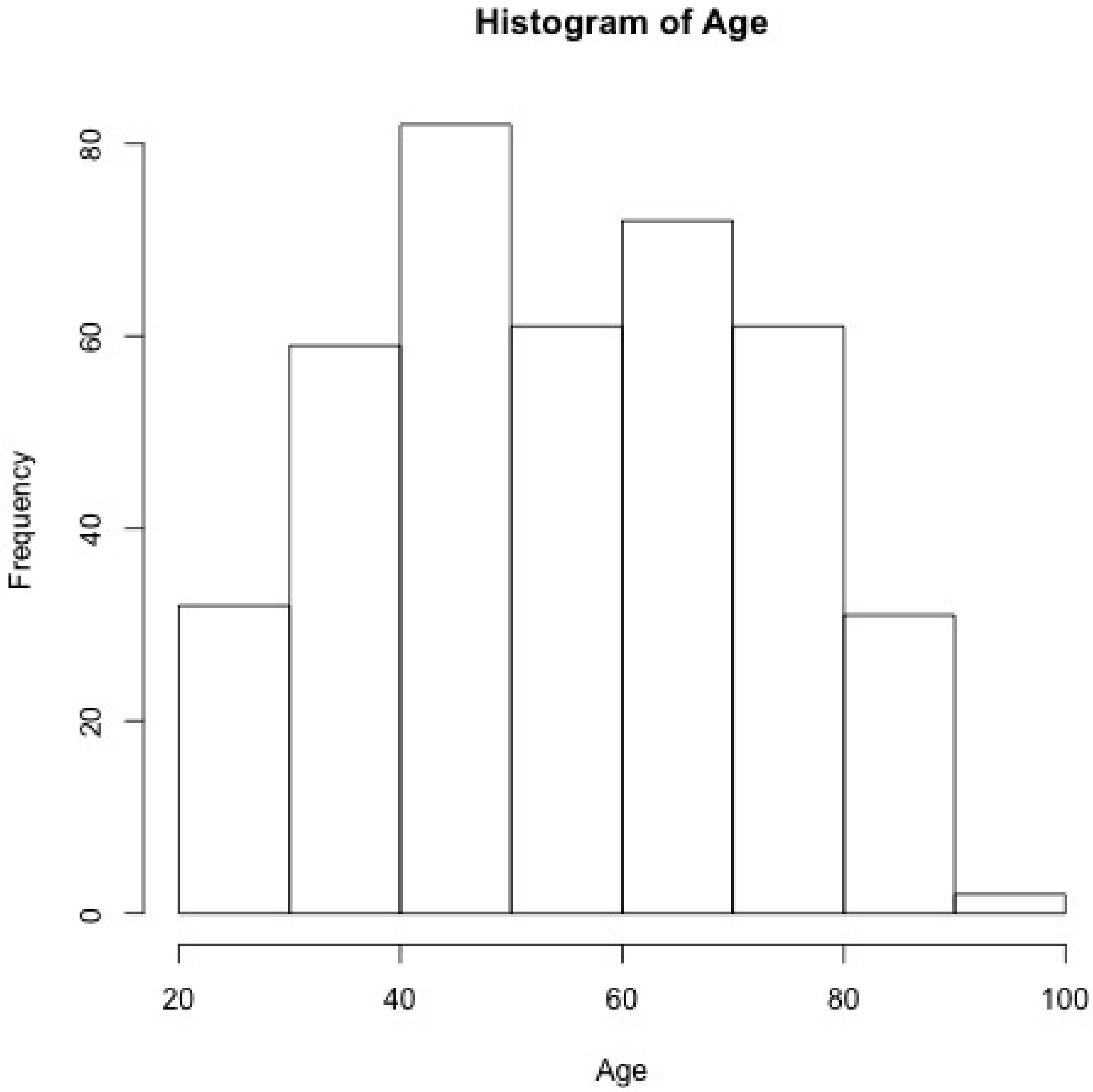
Then we make Histograms and Boxplots for each of the variables

For Predictor Age:

Boxplot

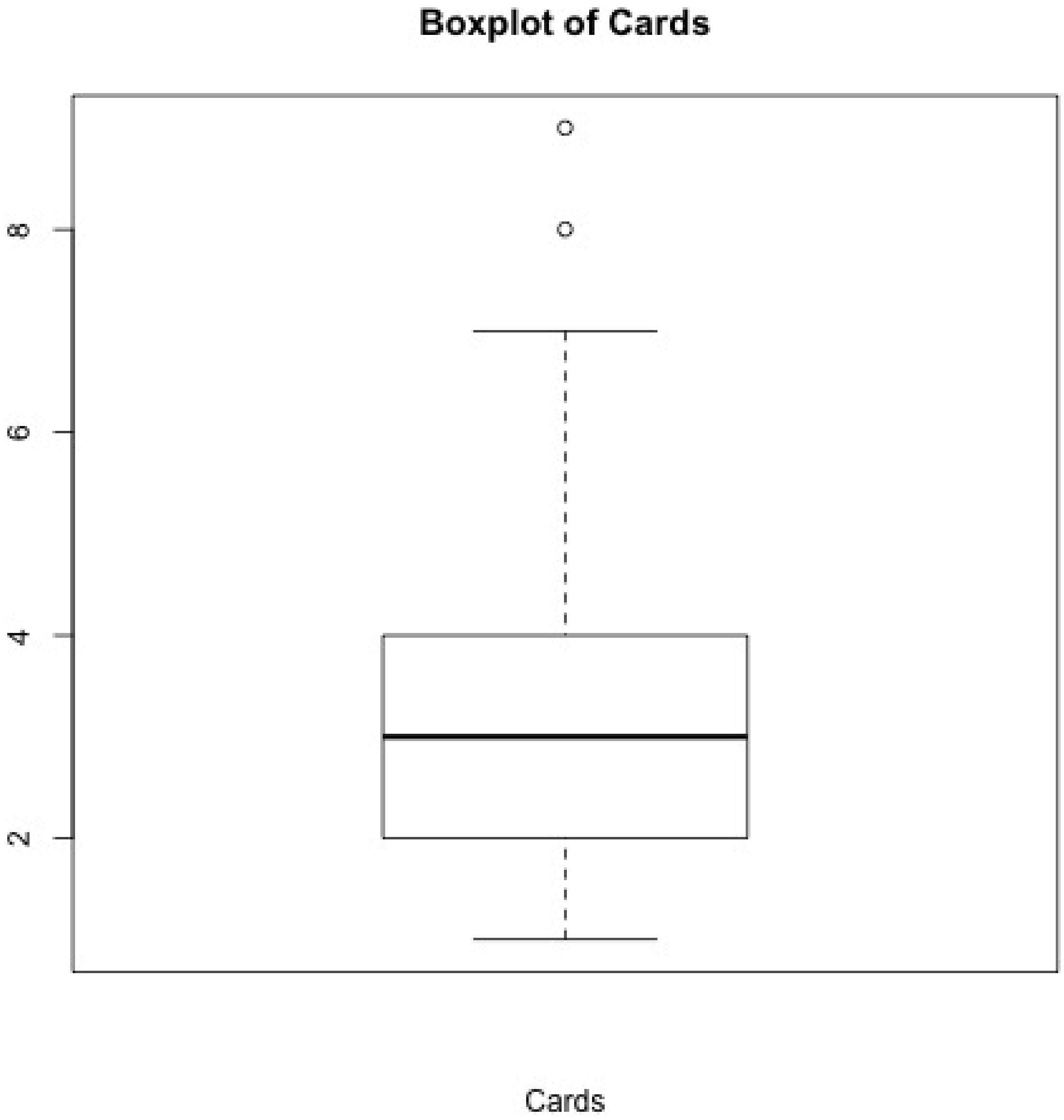


Histogram

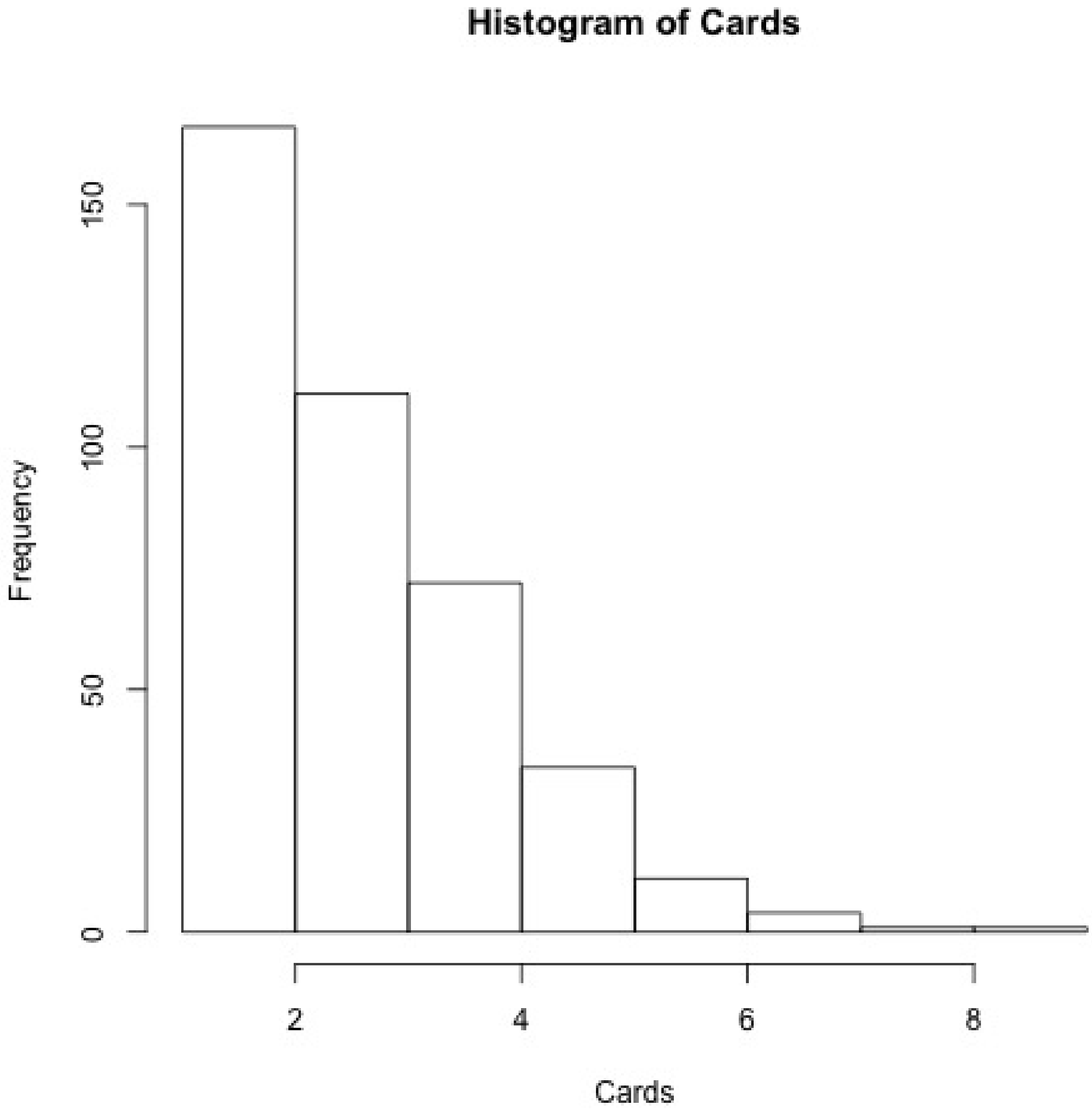


For Predictor Cards

Boxplot

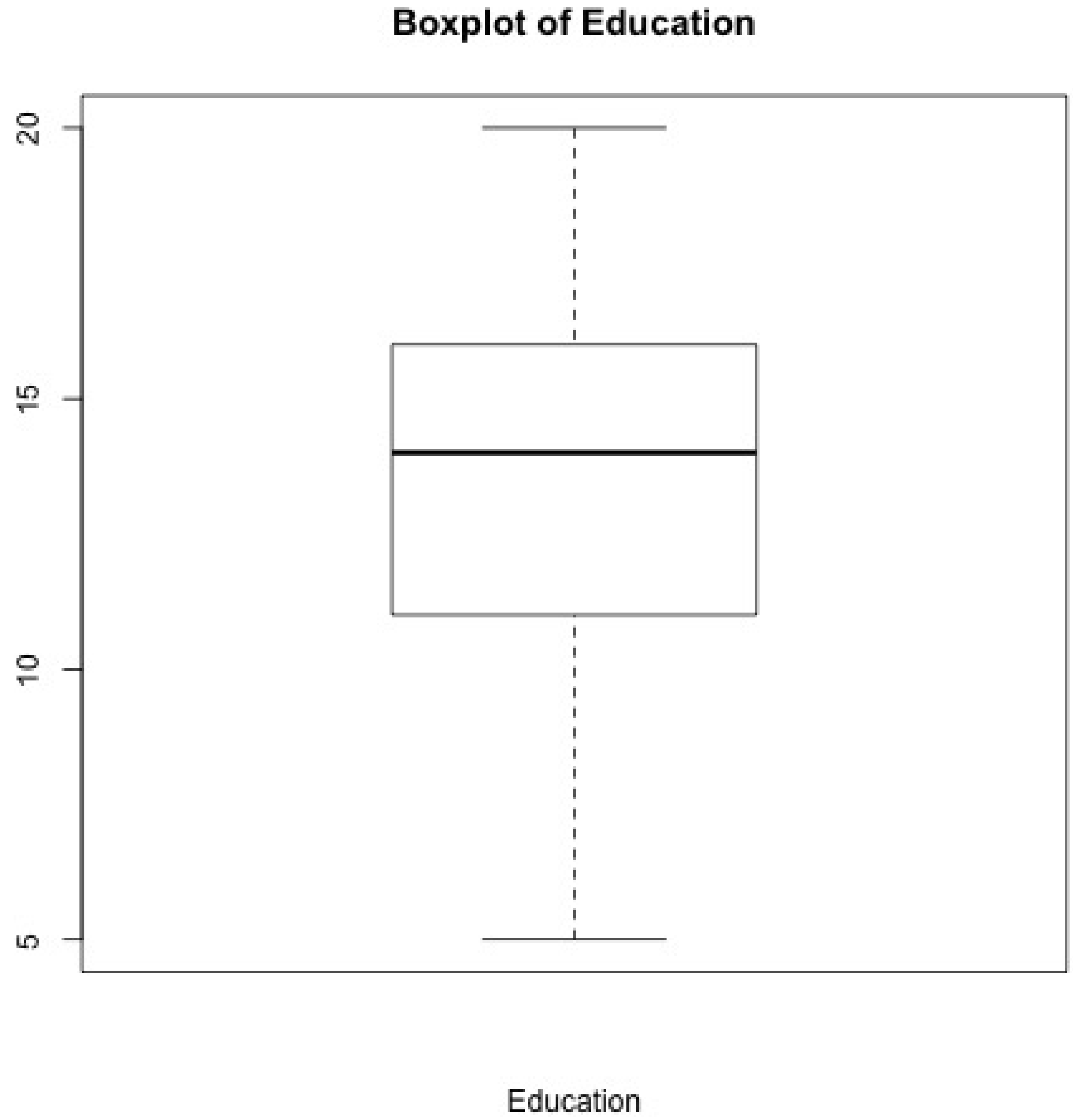


Histogram

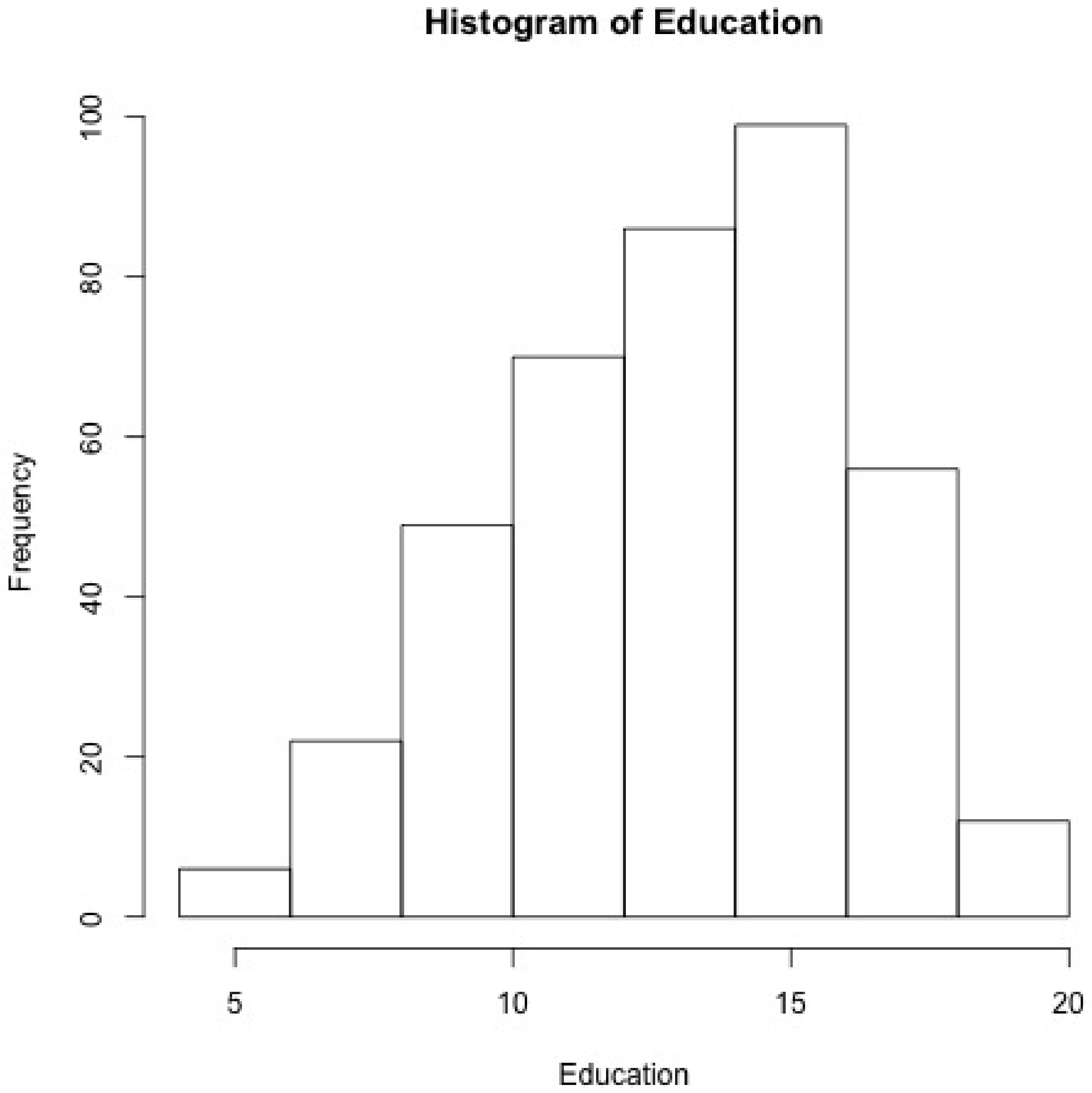


For Predictor Education

Boxplot

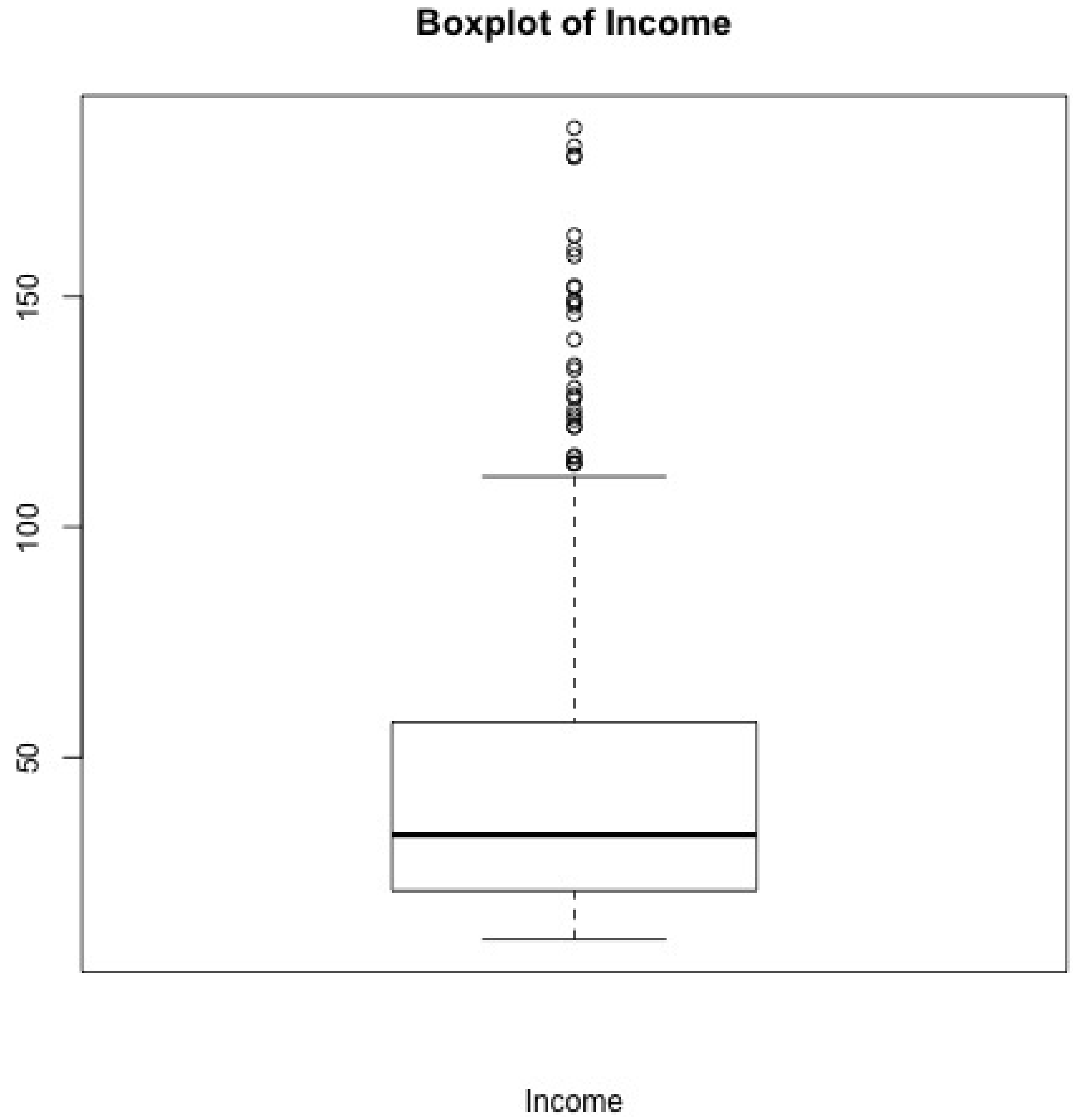


Histogram

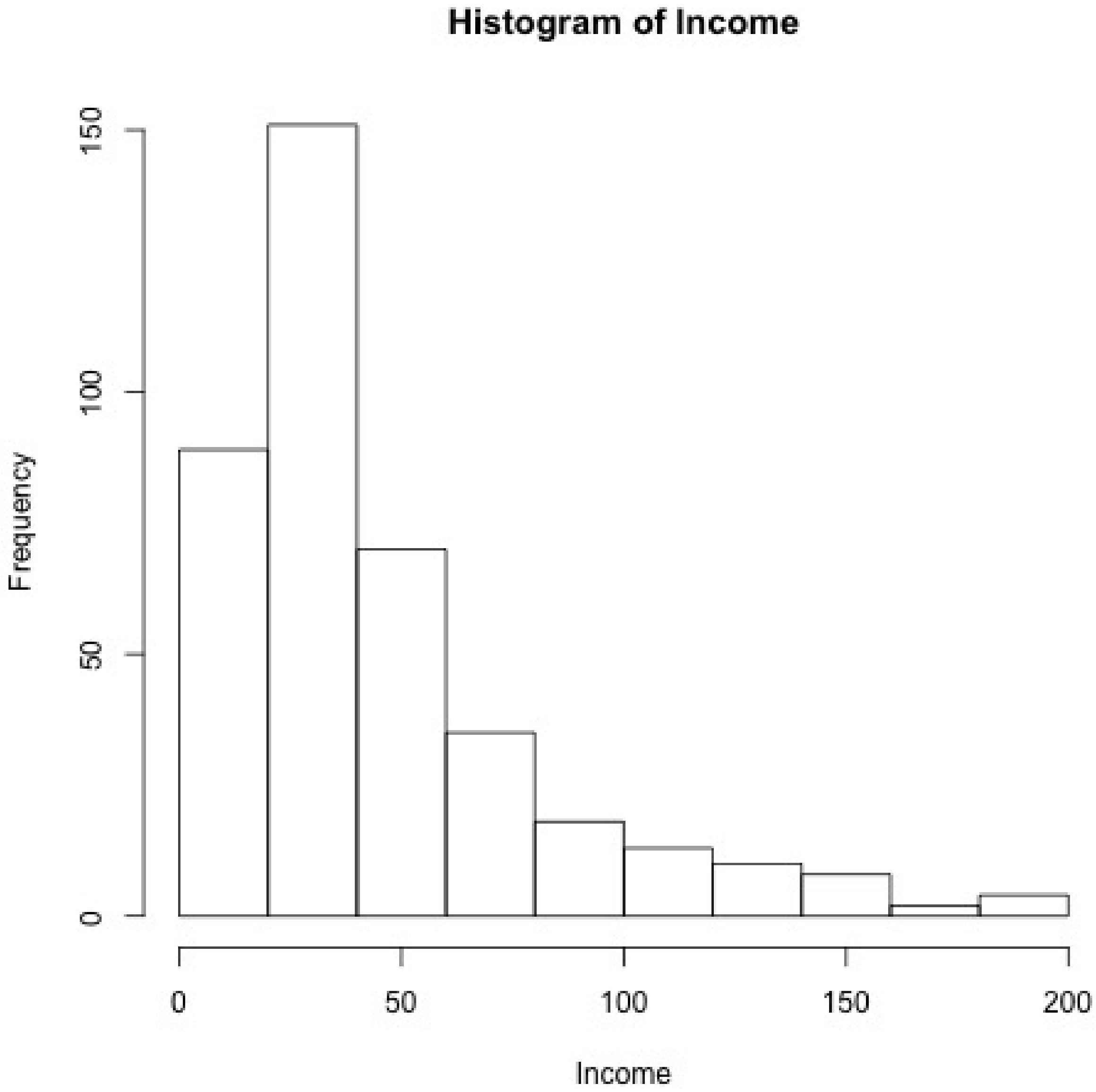


For Predictor Income

Boxplot

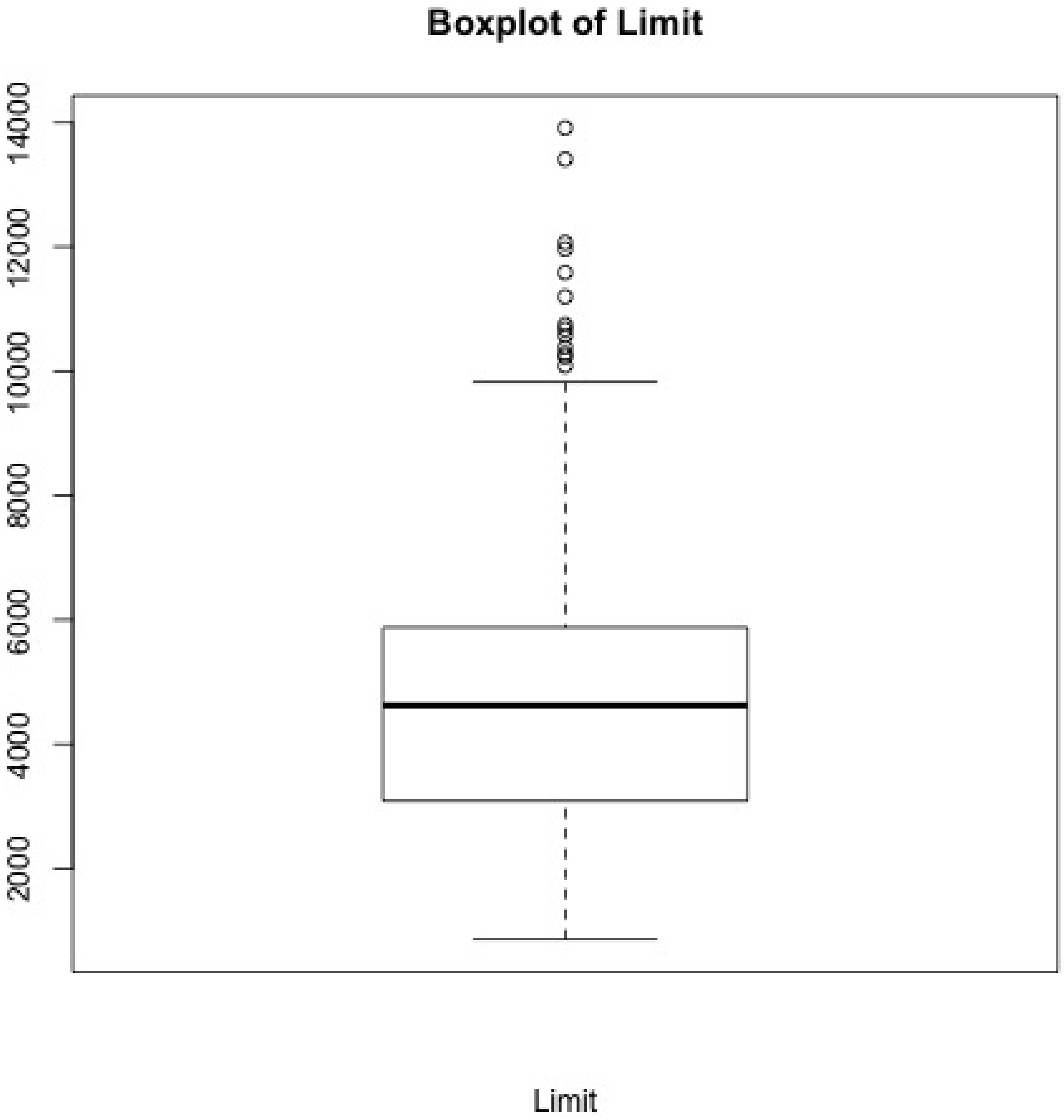


Histogram

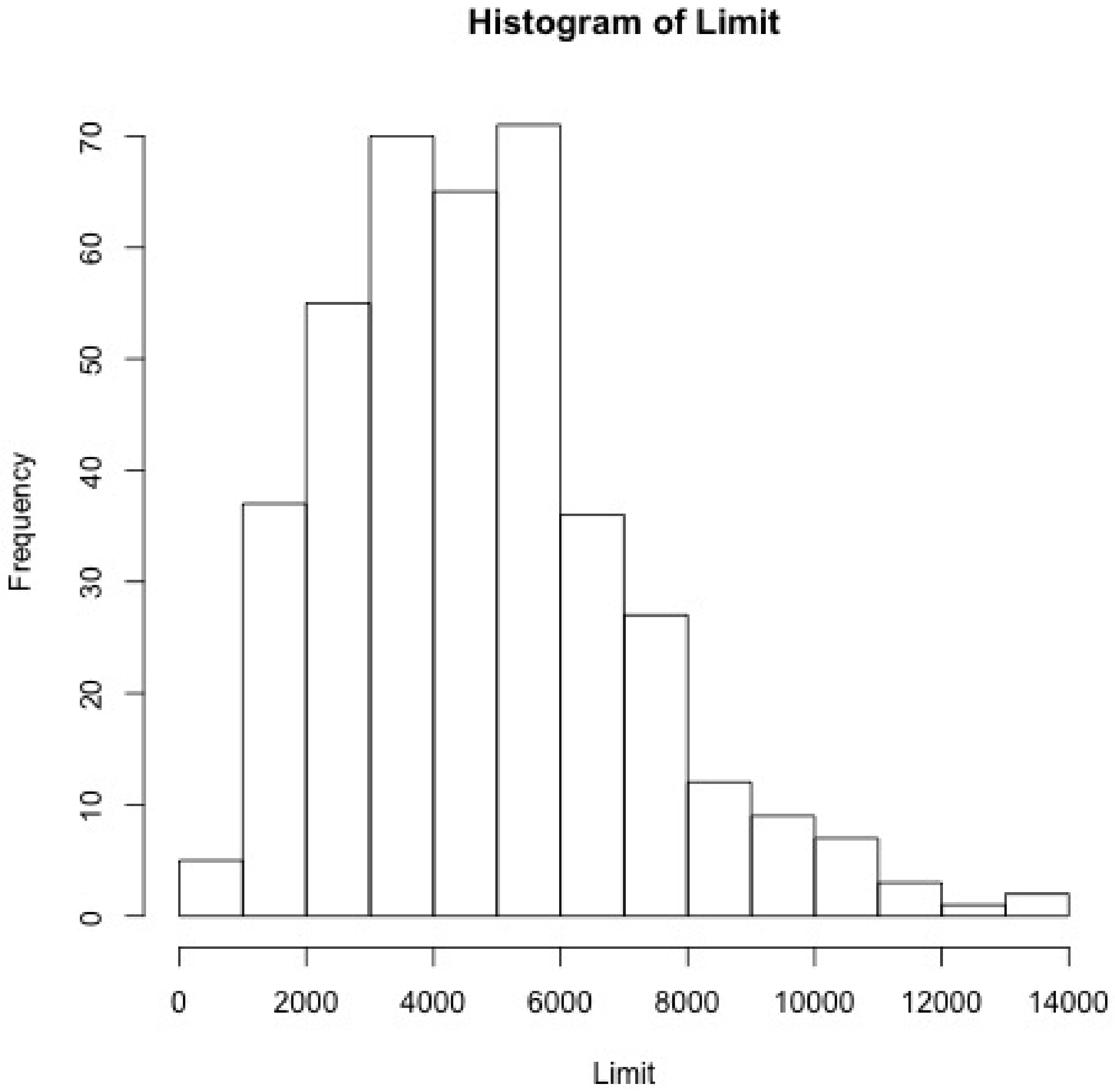


For Predictor Limit

Boxplot

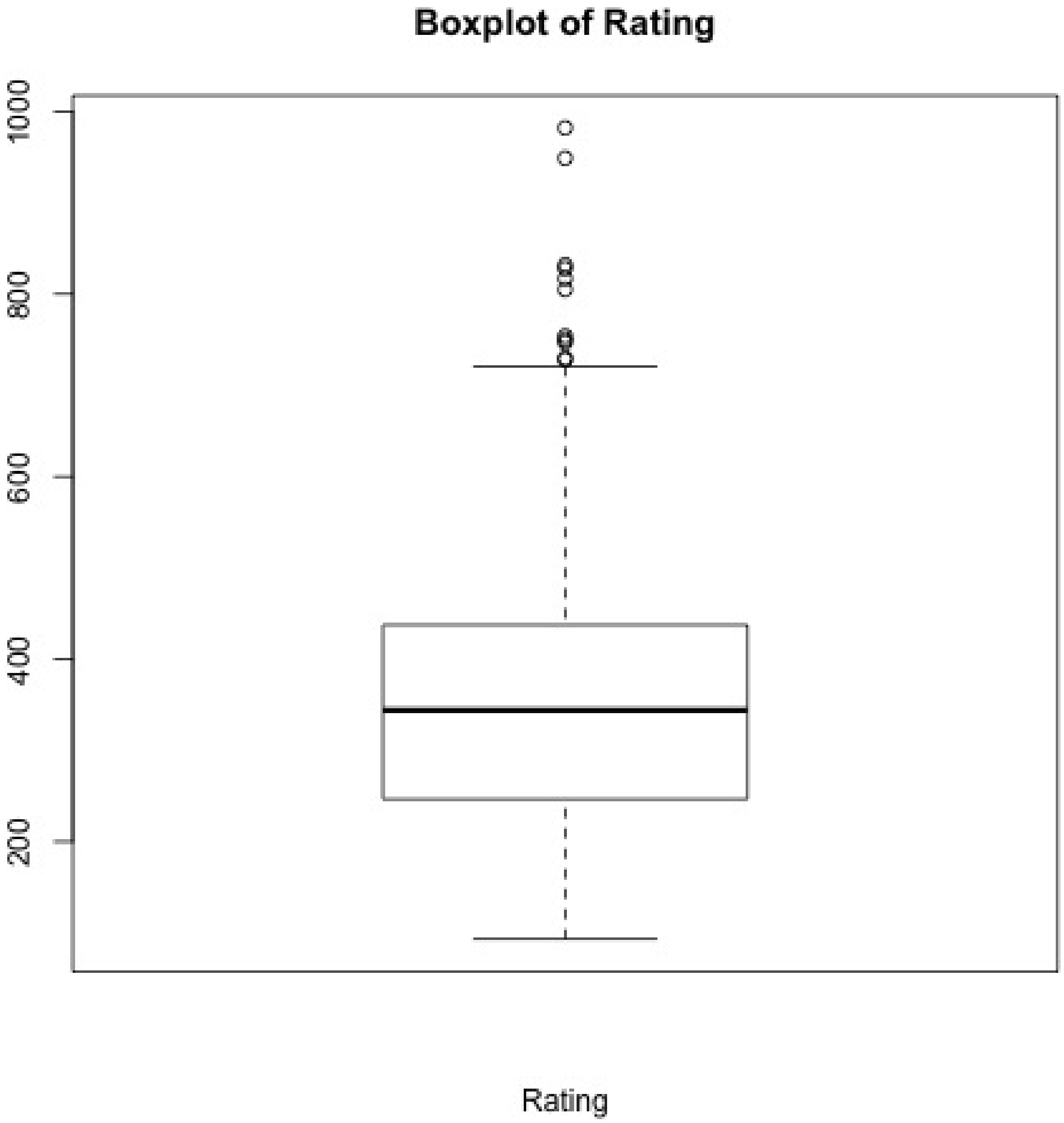

Boxplot of Limit
14000
12000
10000
8000
6000
4000
2000
Limit

Histogram

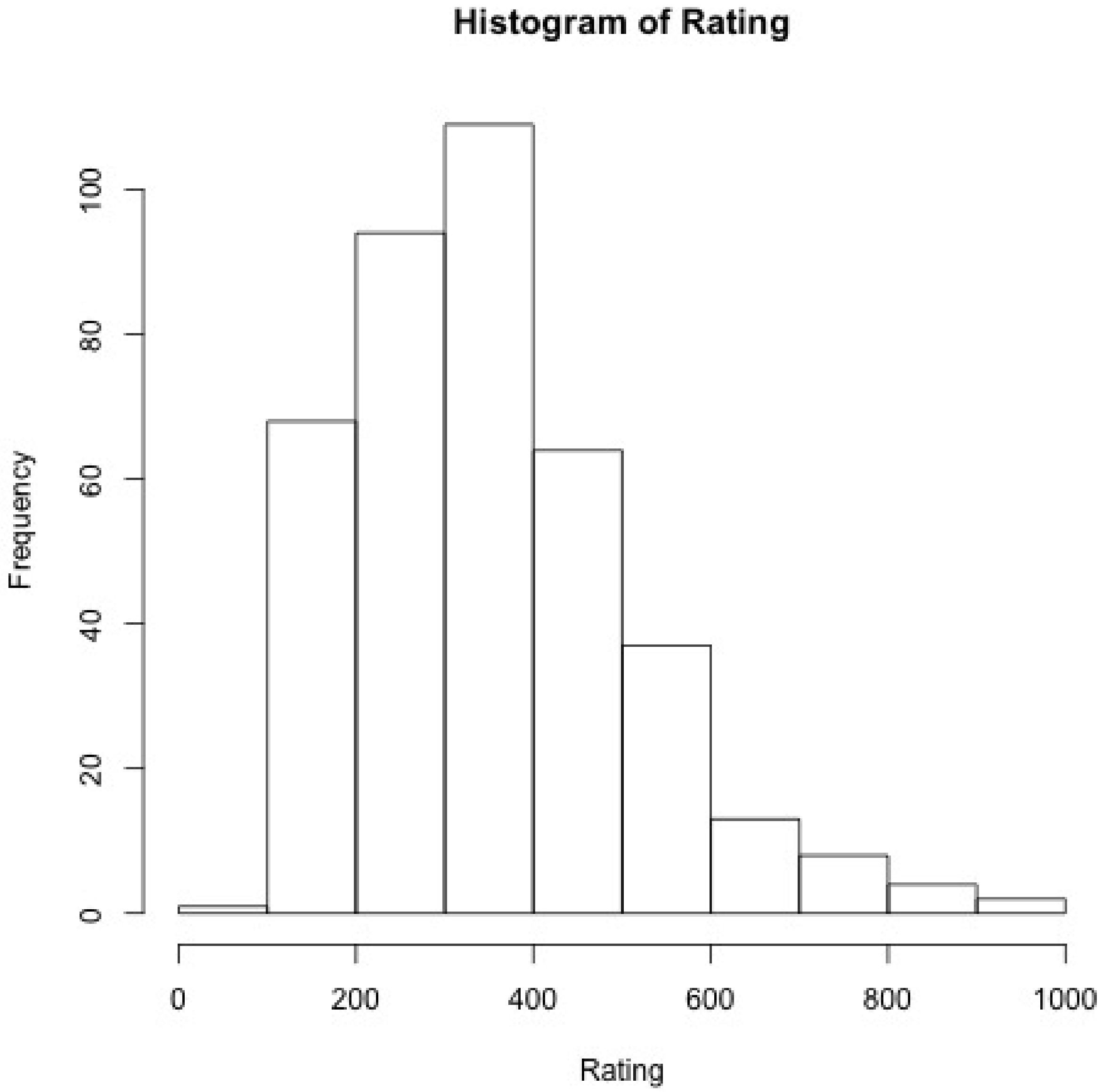


For Predictor Rating

Boxplot



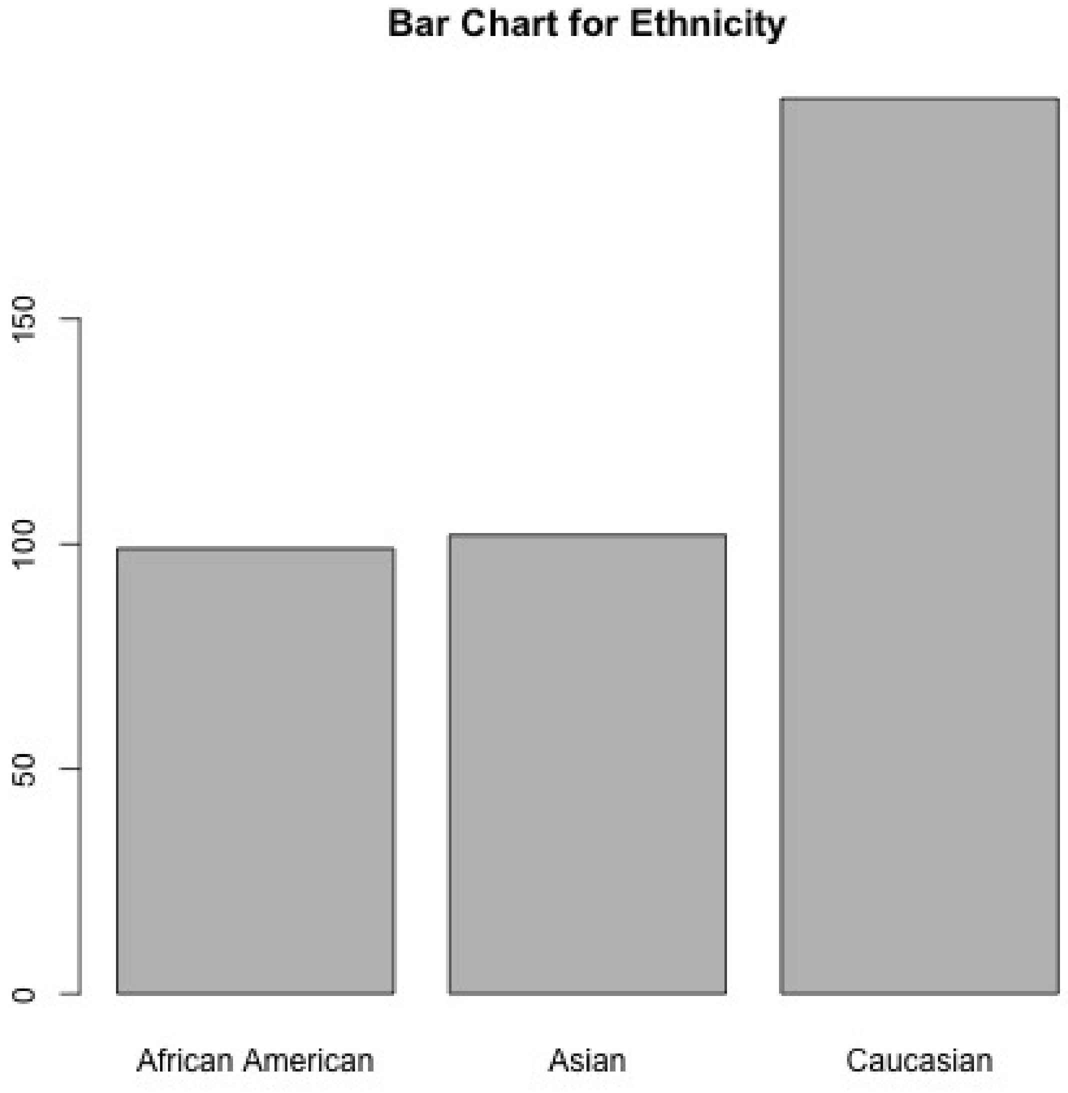
Histogram



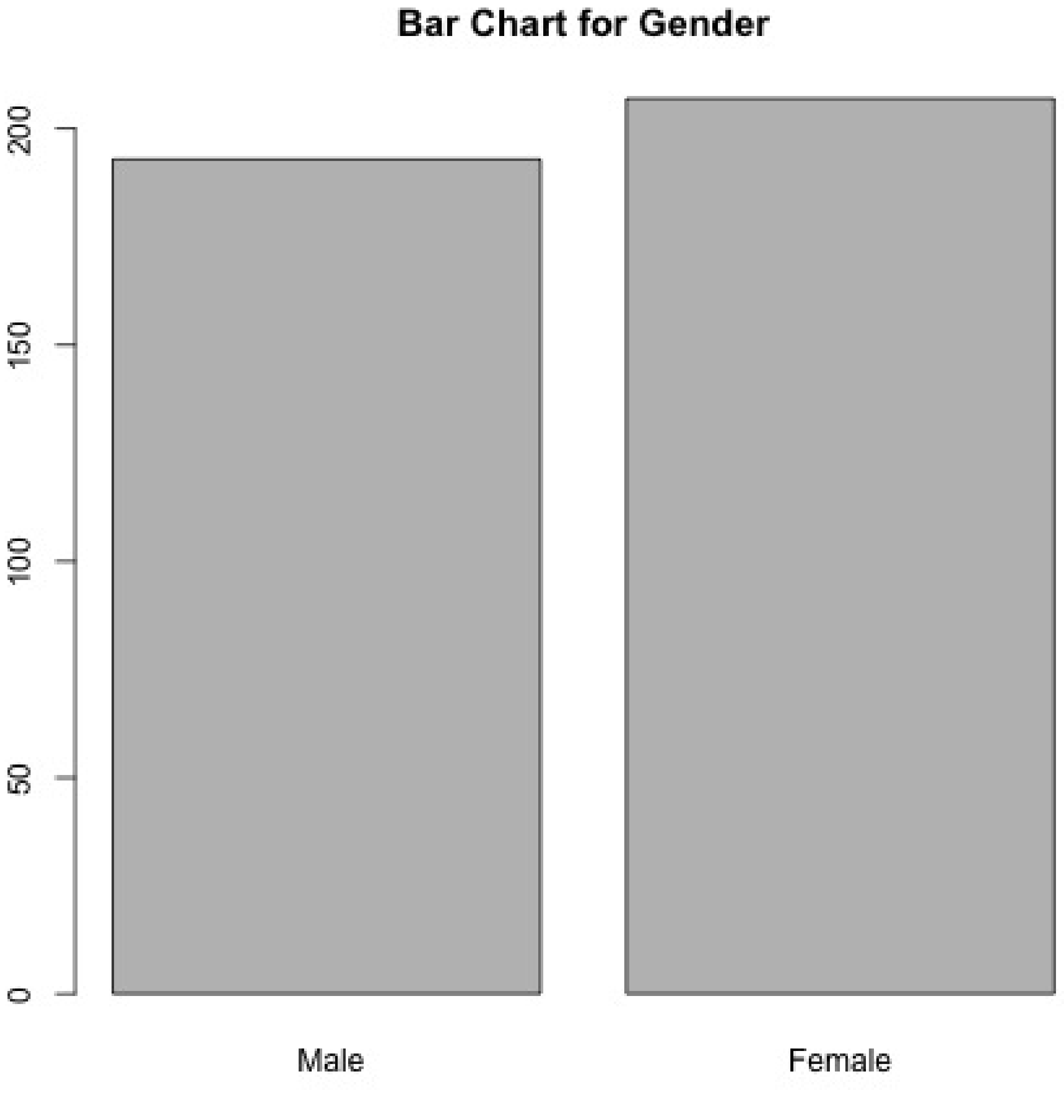
Qualitative Variables

There are 4 qualitative variables in the dataset, including Ethinicity, Gender, Married, Student. For each of the variable, we computed the frequency of each category within the variable and visualize the results by Bar Chart.

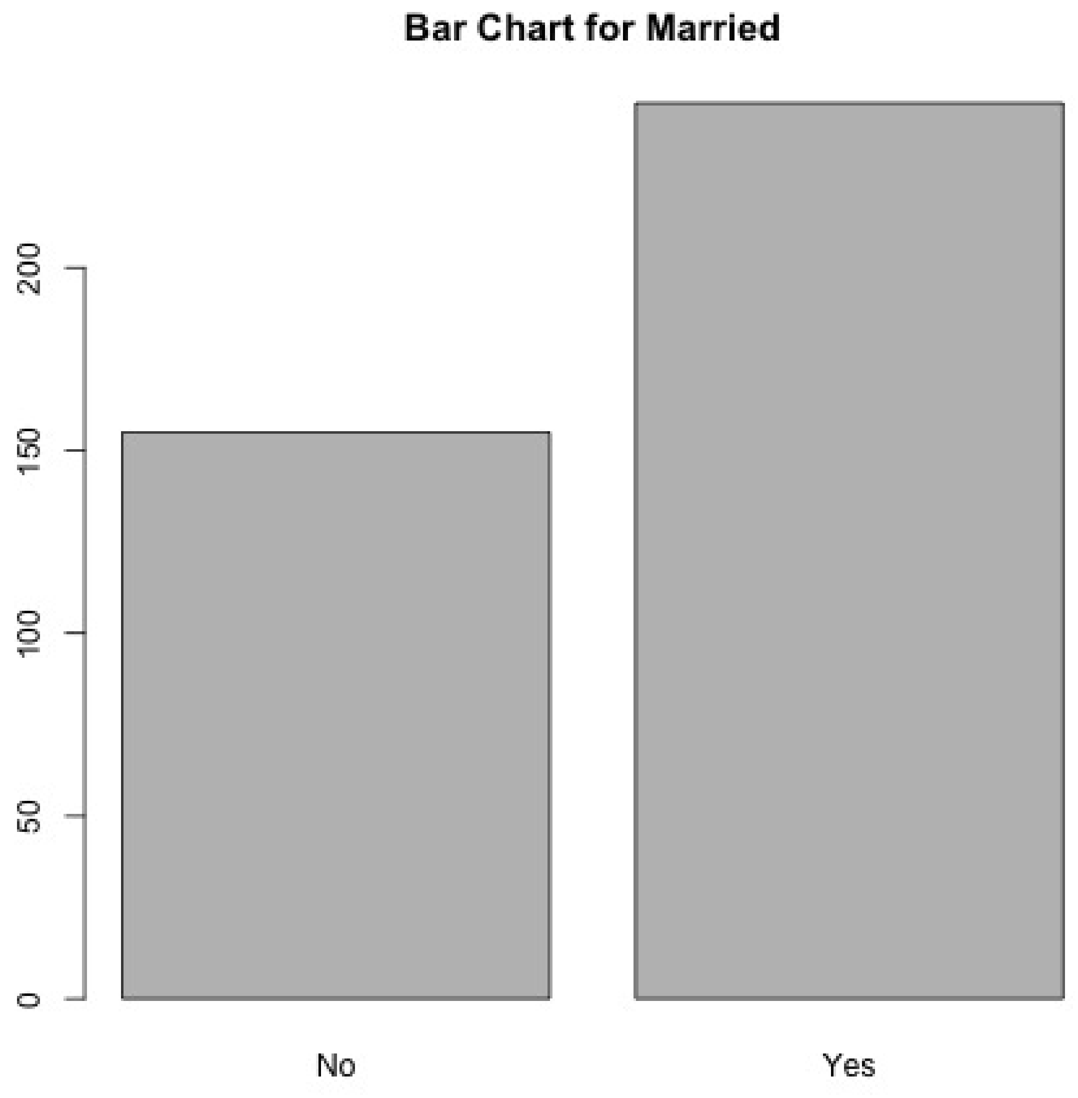
Ethinicity


Bar Chart for Ethnicity
150
100
50
0
African American
Asian
Caucasian

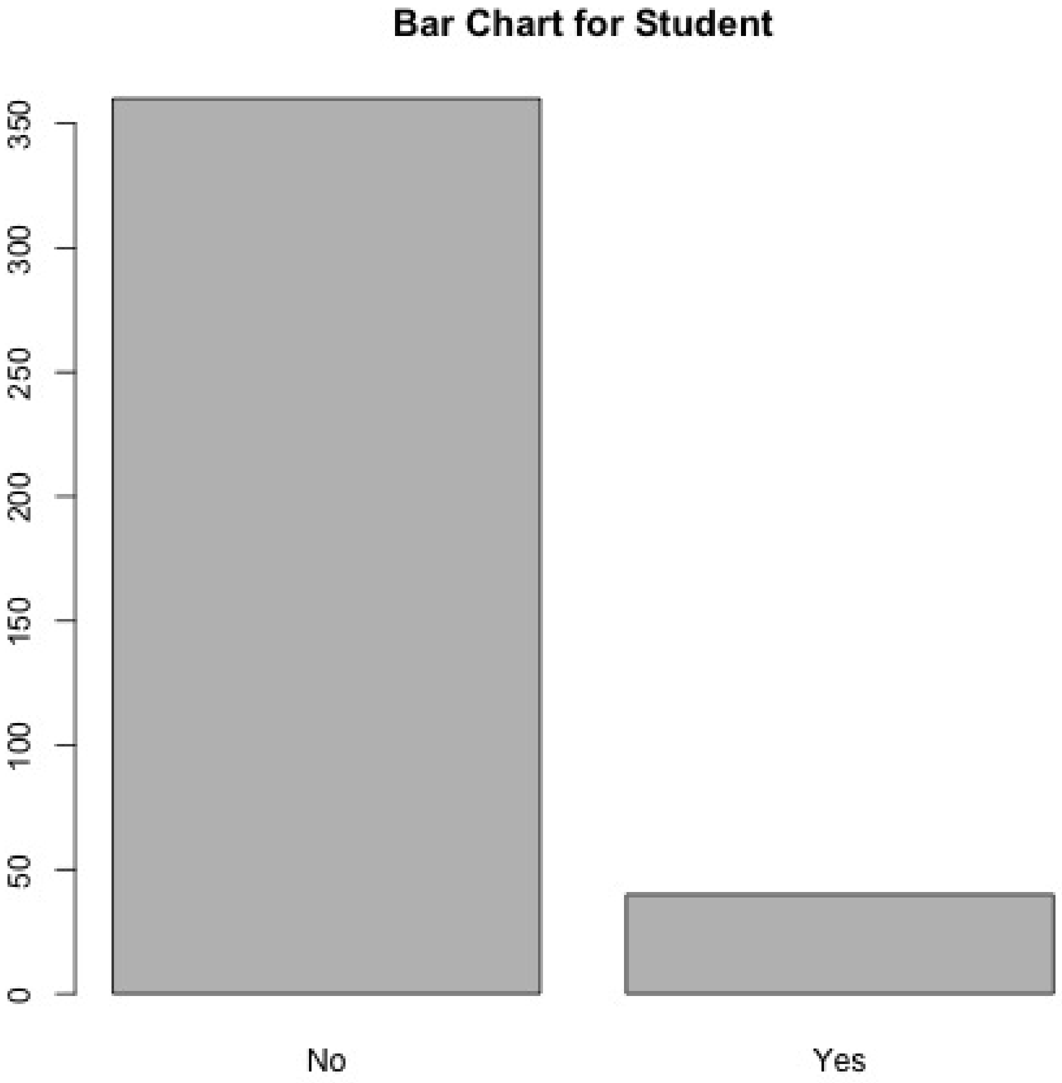
Gender



Married

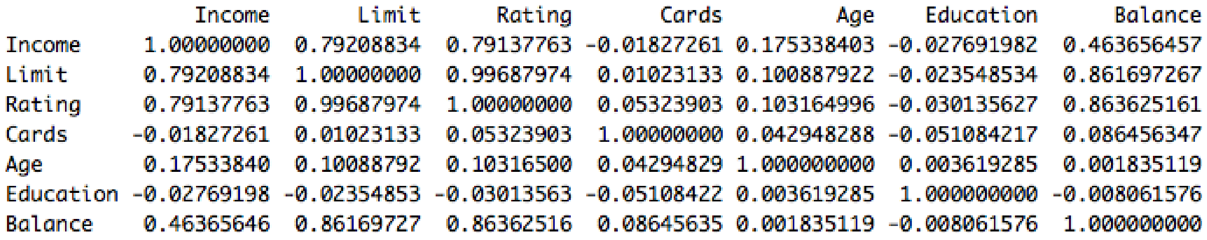


Student



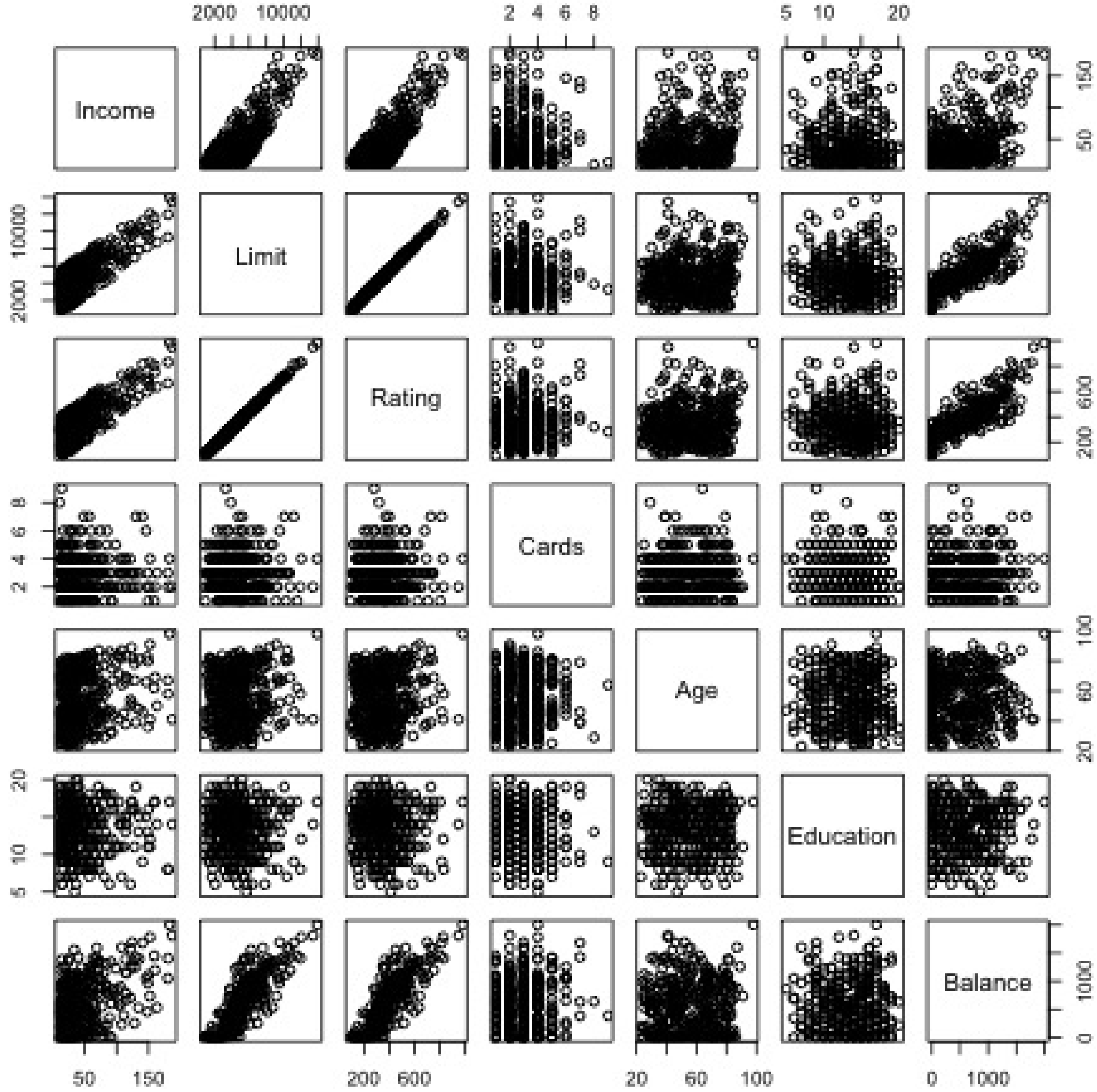
Association

To study the association between predictors and response variable, we first calculate the correlation matrix among all predictors.

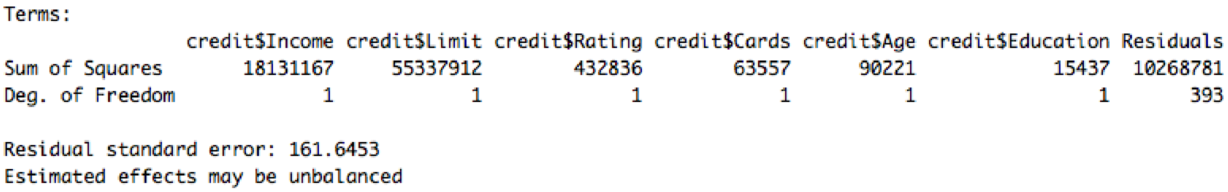


```
              Income       Limit      Rating       Cards         Age    Education      Balance
Income     1.00000000  0.79208834  0.79137763 -0.01827261 0.175338403 -0.027691982  0.463656457
Limit      0.79208834  1.00000000  0.99687974  0.01023133 0.100887922 -0.023548534  0.861697267
Rating     0.79137763  0.99687974  1.00000000  0.05323903 0.103164996 -0.030135627  0.863625161
Cards     -0.01827261  0.01023133  0.05323903  1.00000000 0.042948288 -0.051084217  0.086456347
Age        0.17533840  0.10088792  0.10316500  0.04294829 1.000000000  0.003619285  0.001835119
Education -0.02769198 -0.02354853 -0.03013563 -0.05108422 0.003619285  1.000000000 -0.008061576
Balance    0.46365646  0.86169727  0.86362516  0.08645635 0.001835119 -0.008061576  1.000000000
```

Then we visualize the scatterplot mattrix



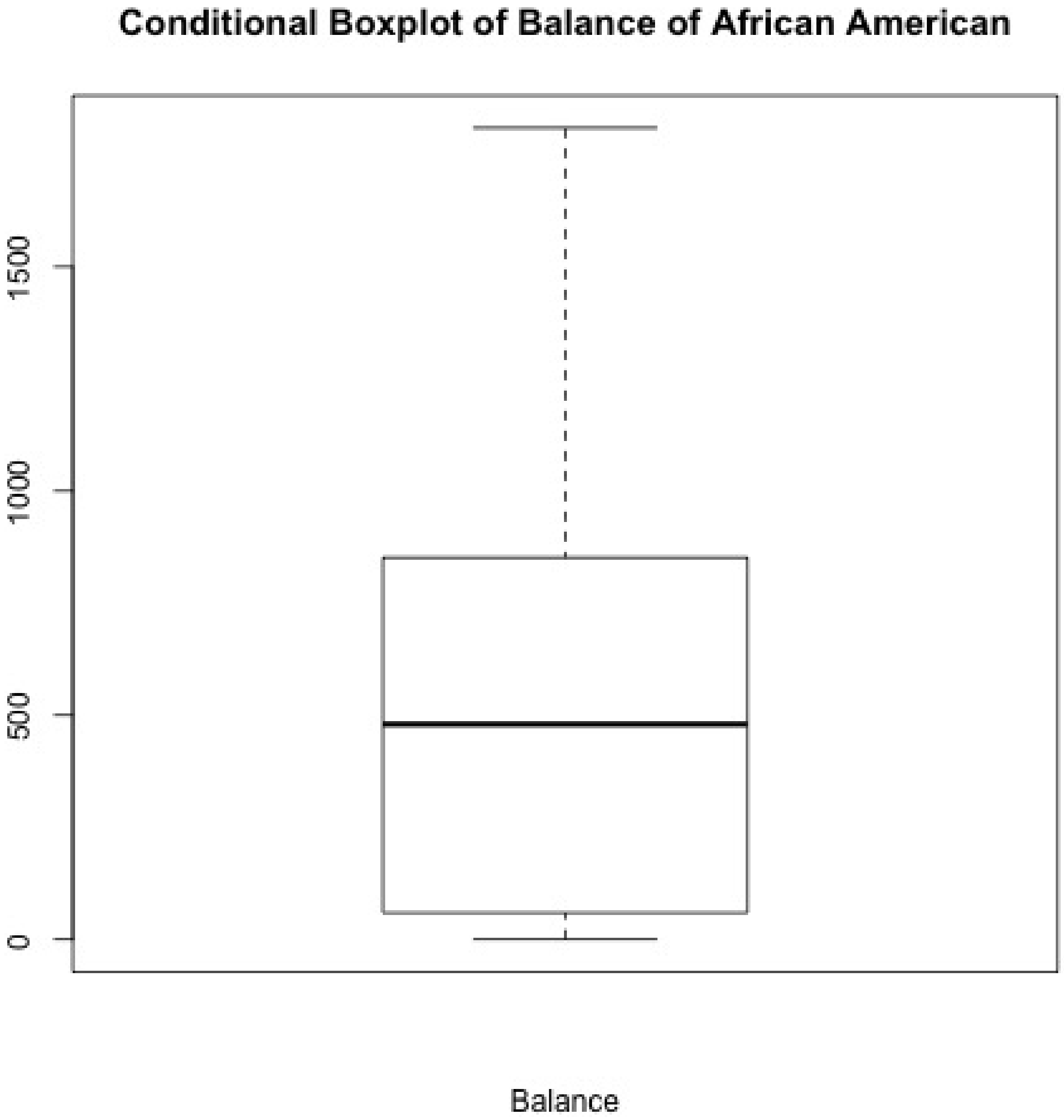
Analysis of variance (ANOVA) is a collection of statistical models used to analyze the differences among group means and their associated procedures (such as “variation” among and between groups)



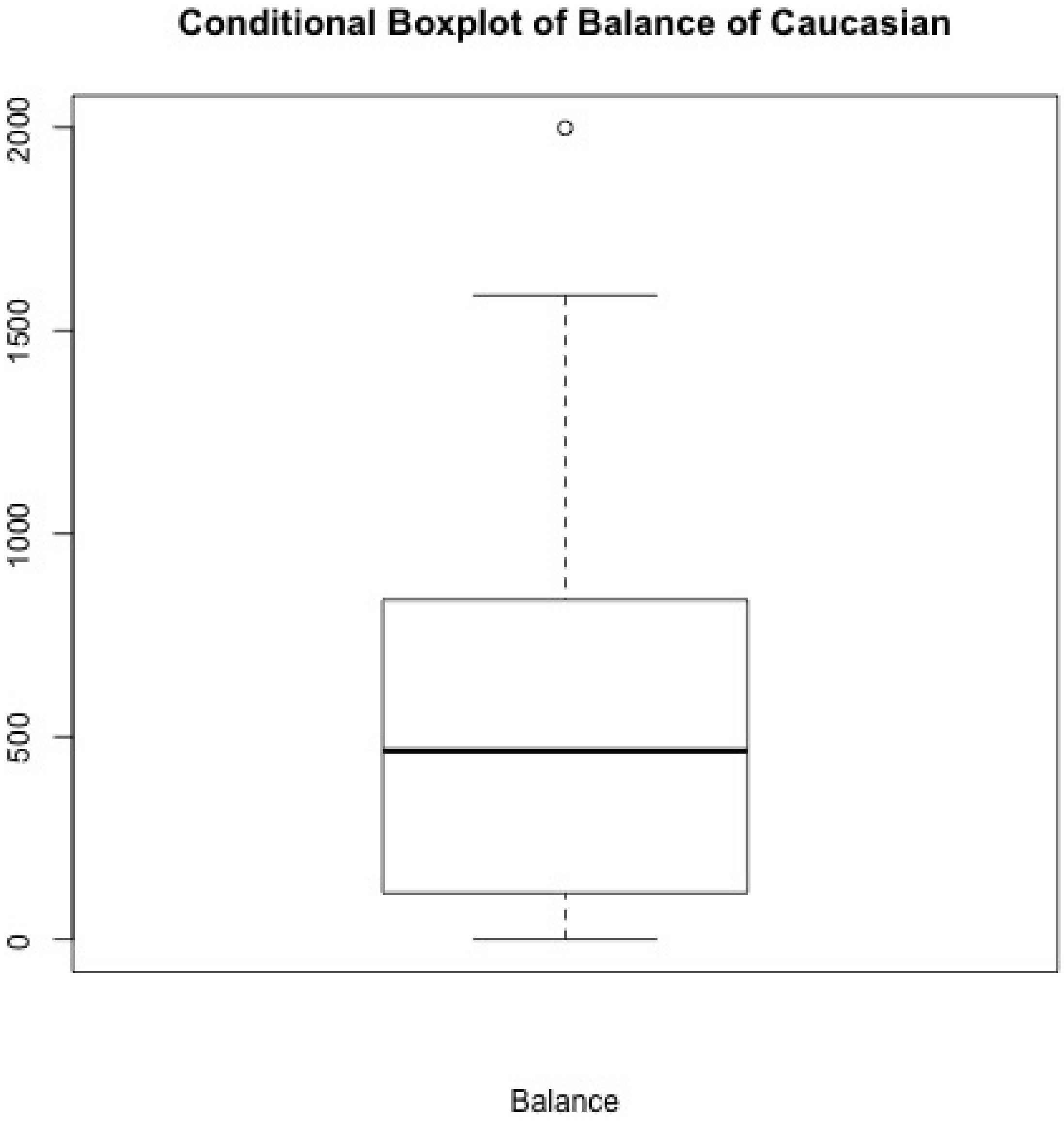
```
Terms:
                credit$Income credit$Limit credit$Rating credit$Cards credit$Age credit$Education Residuals
Sum of Squares       18131167     55337912        432836        63557      90221            15437  10268781
Deg. of Freedom             1            1             1            1          1                1       393

Residual standard error: 161.6453
Estimated effects may be unbalanced
```

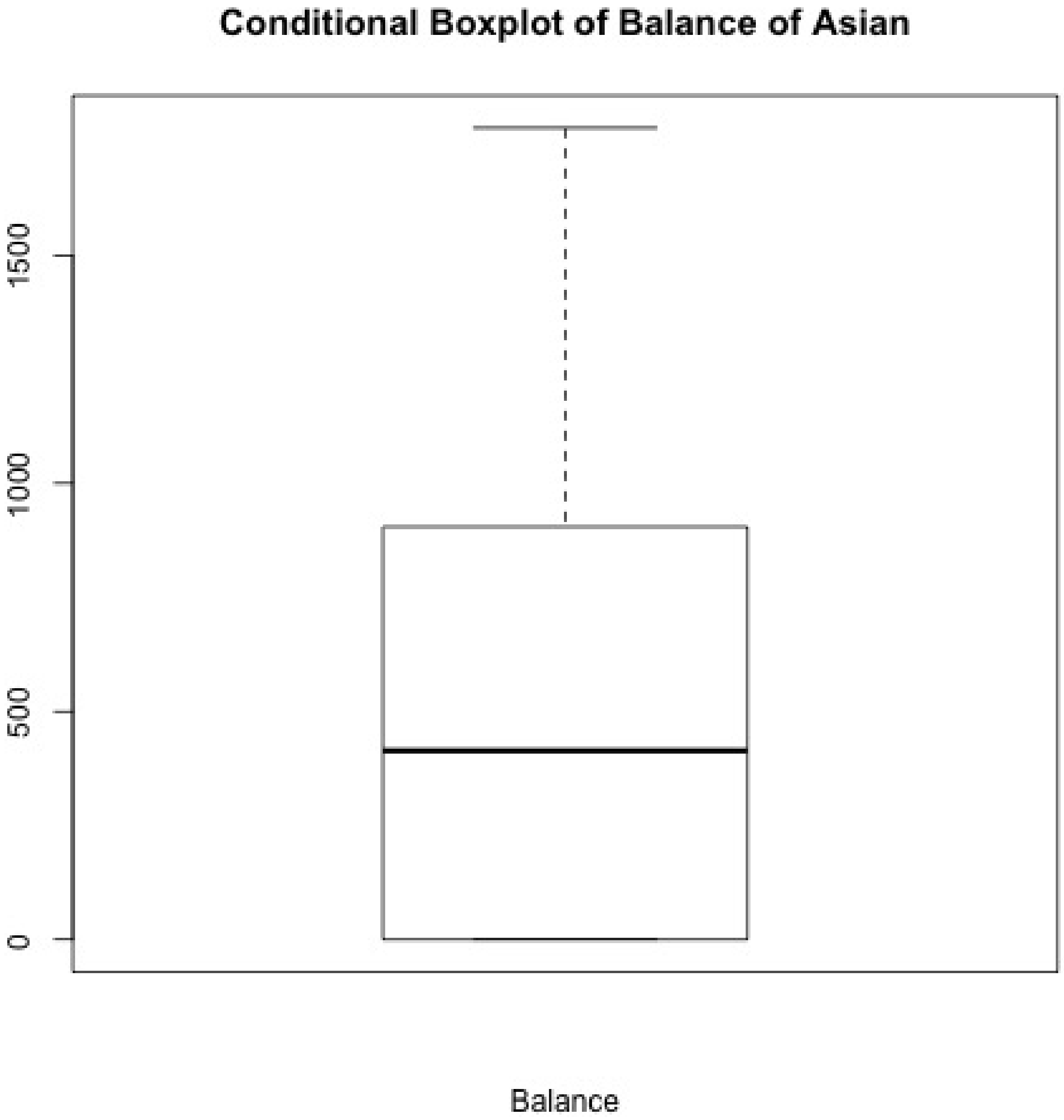
Finally we visualize the conditional boxplots of Balance conditioned to each category of Gender, Ethnicity, Student and Married.

Ethinicity

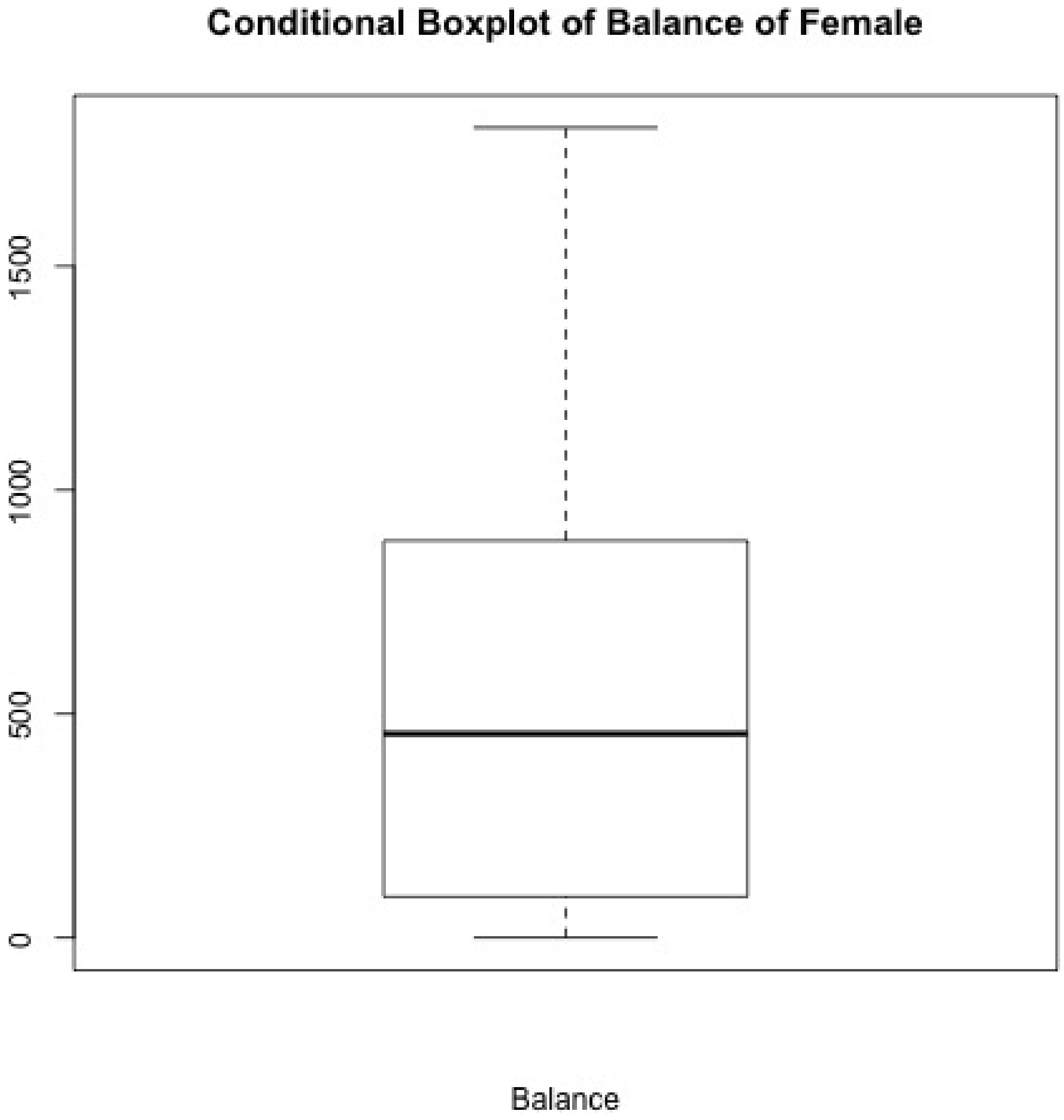


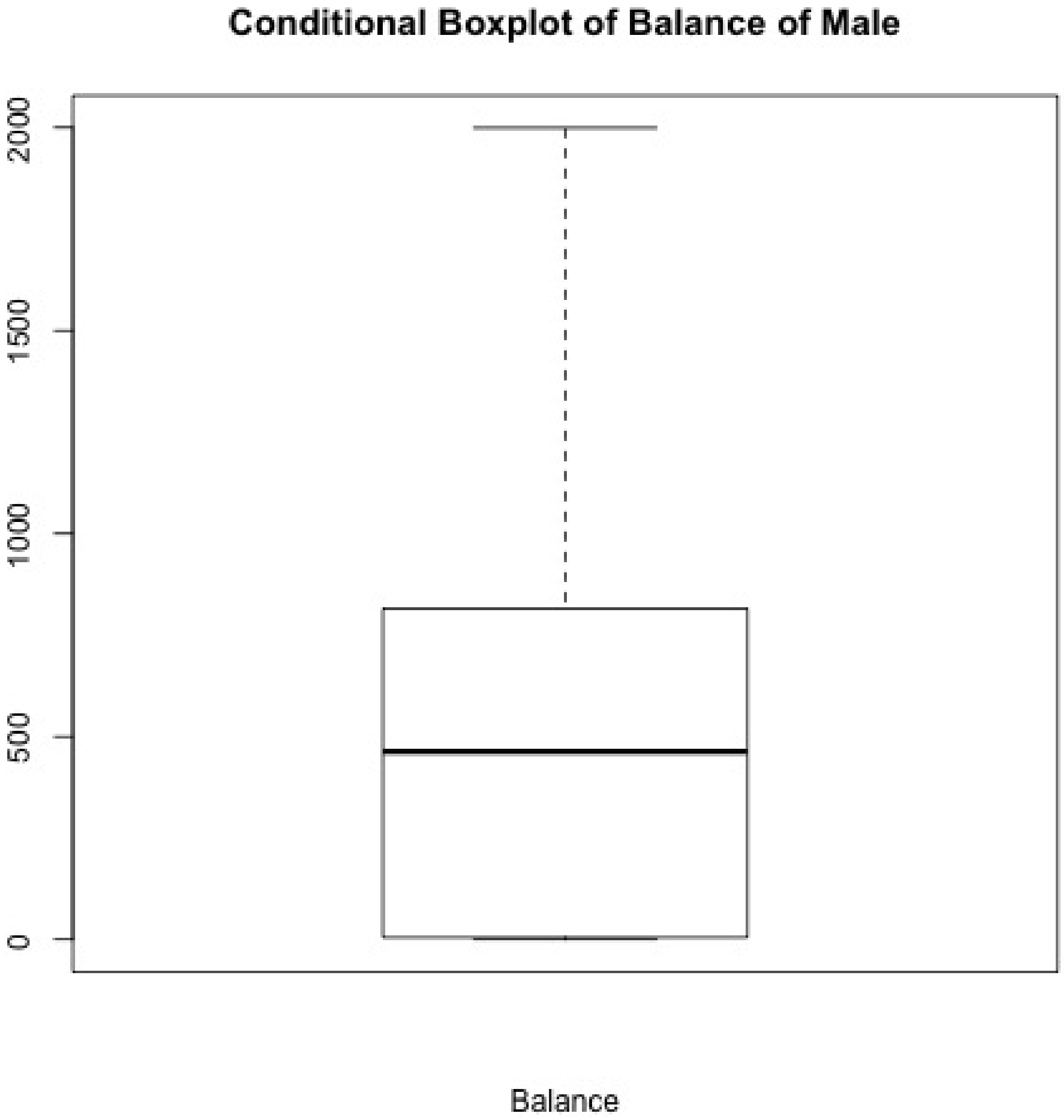
Conditional Boxplot of Balance of African American


Conditional Boxplot of Balance of Caucasian
2000
1500
1000
500
0
Balance

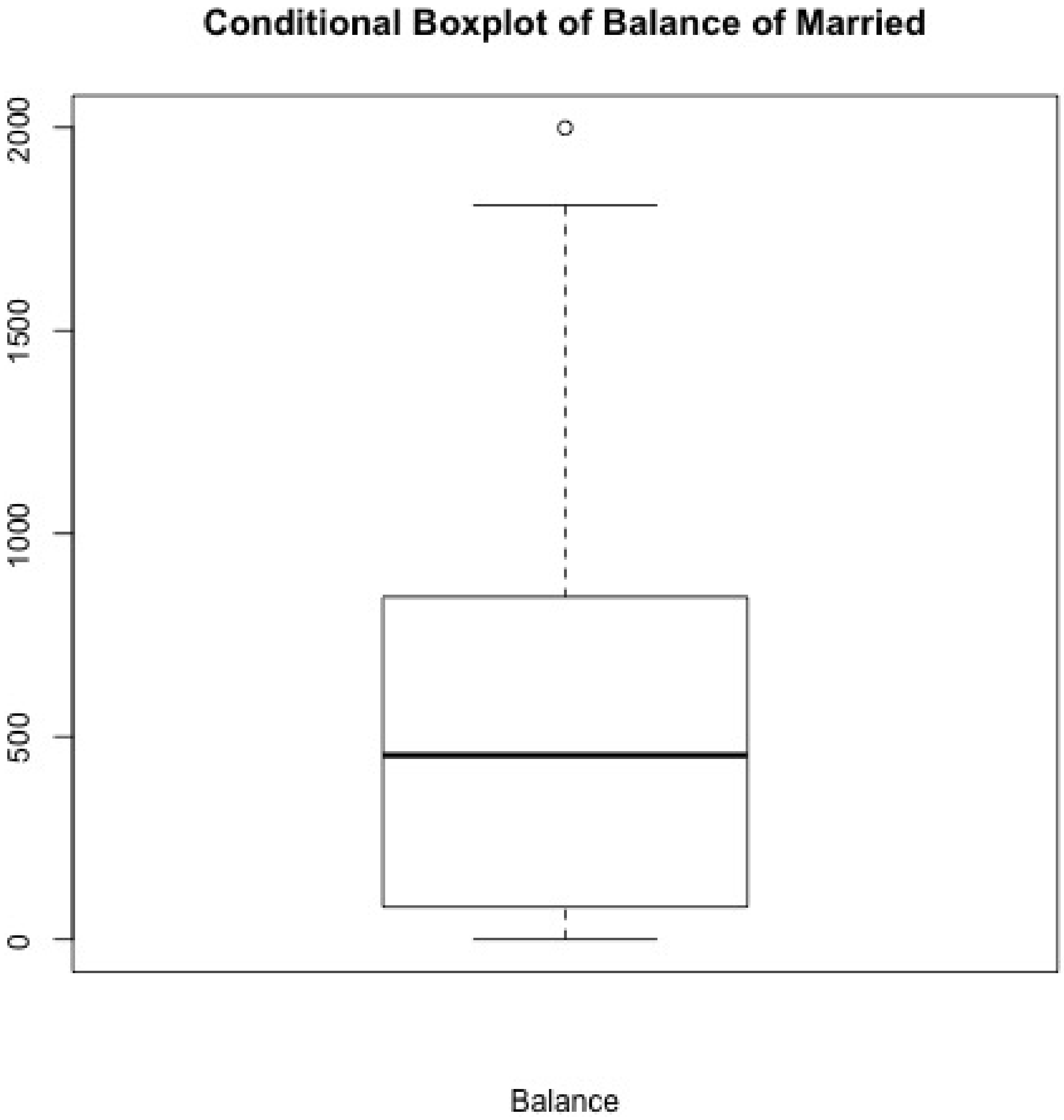

Conditional Boxplot of Balance of Asian
0
500
1000
1500
Balance

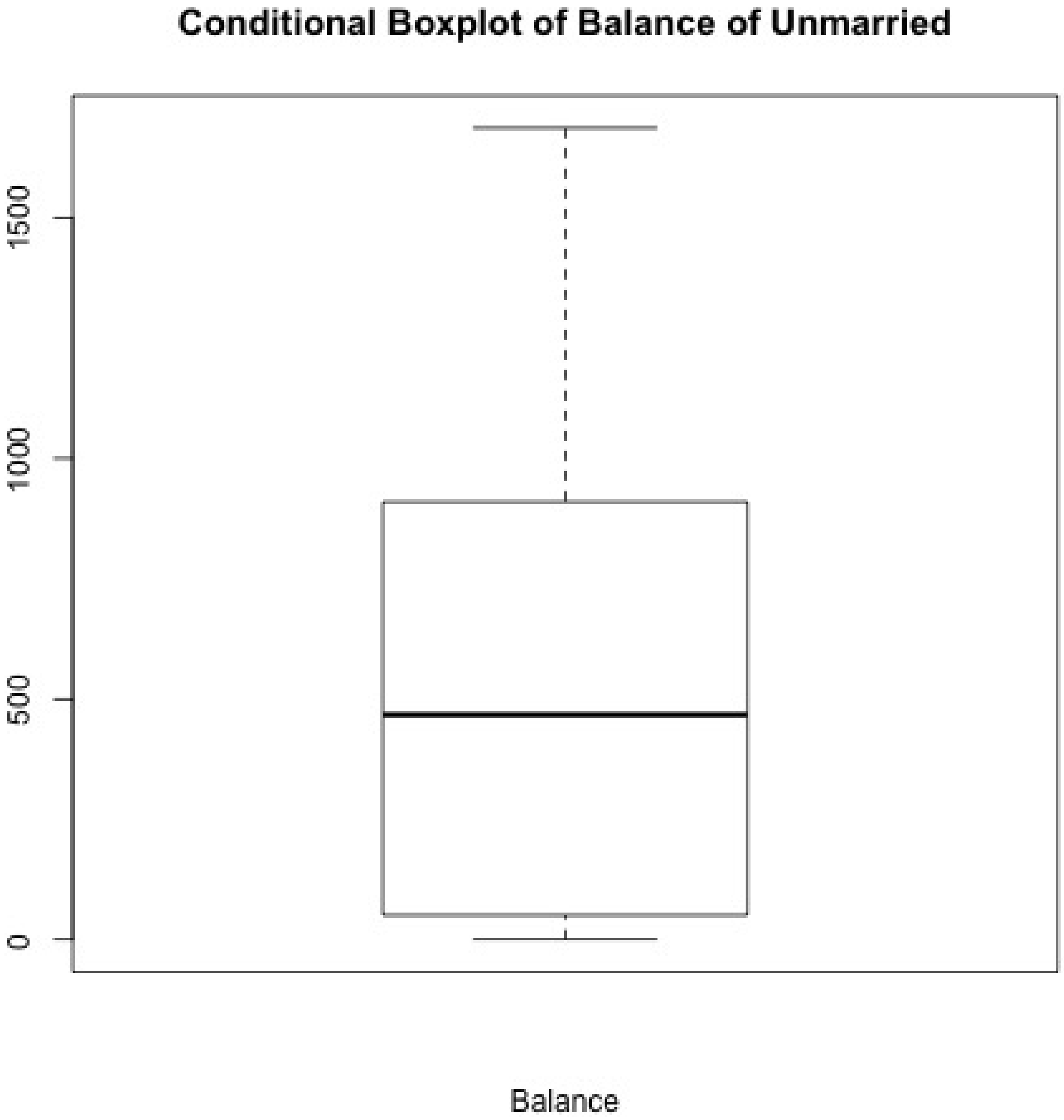
Gender


Conditional Boxplot of Balance of Female
0
500
1000
1500
Balance

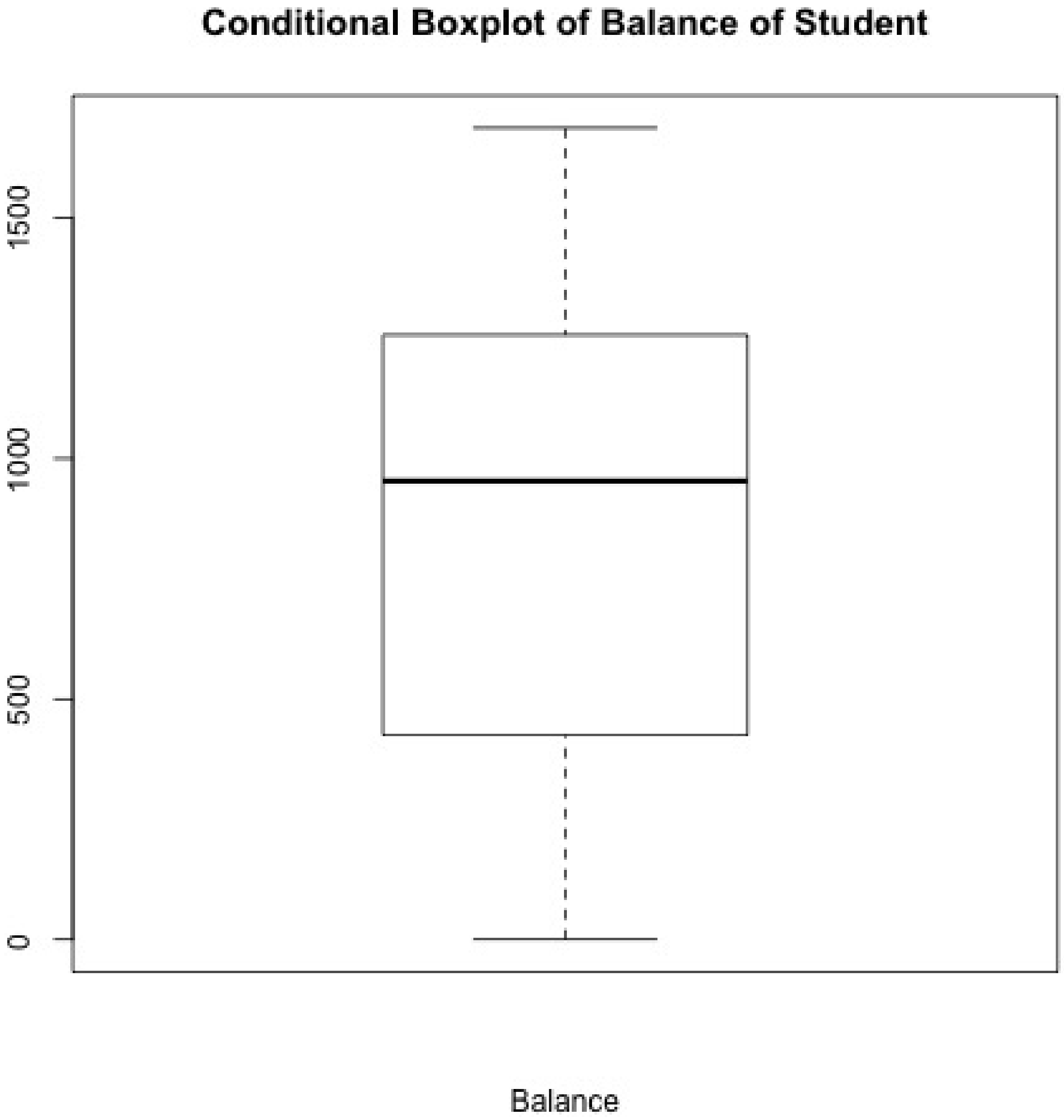

Conditional Boxplot of Balance of Male
2000
1500
1000
500
0
Balance

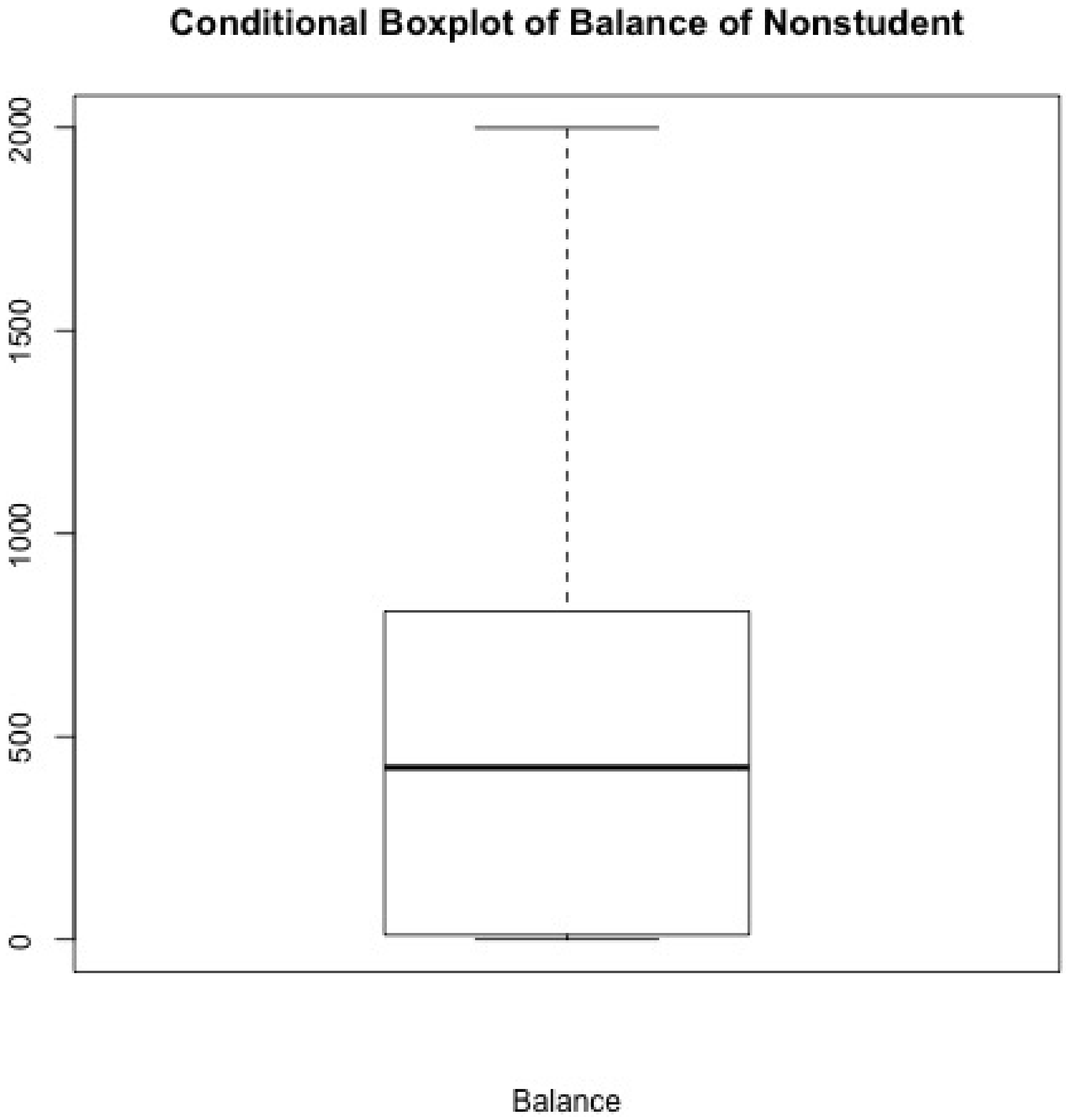
Married


Conditional Boxplot of Balance of Married
2000
1500
1000
500
0
Balance


Conditional Boxplot of Balance of Unmarried
0
500
1000
1500
Balance

Student


Conditional Boxplot of Balance of Student
1500
1000
500
0
Balance

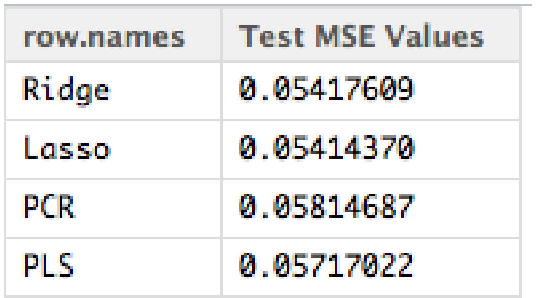


Results

Since it is a continuous prediction problem, we choose Least Squares Regression as our model. In addition, we performed two Shrinkage Methods (Ridge Regression and Lasso Regression) and two Dimension Reduction Methods (Principal Components Regression and Partial Least Squares Regression) to compare the results with our base model.

Then we used 10-fold Cross Validation to choose minimum lambda for Shrinkage Methods and minimum validation components for Dimension Reduction Methods. With the chosen parameter, we compare results by calculating Mean Square Error on test dataset.

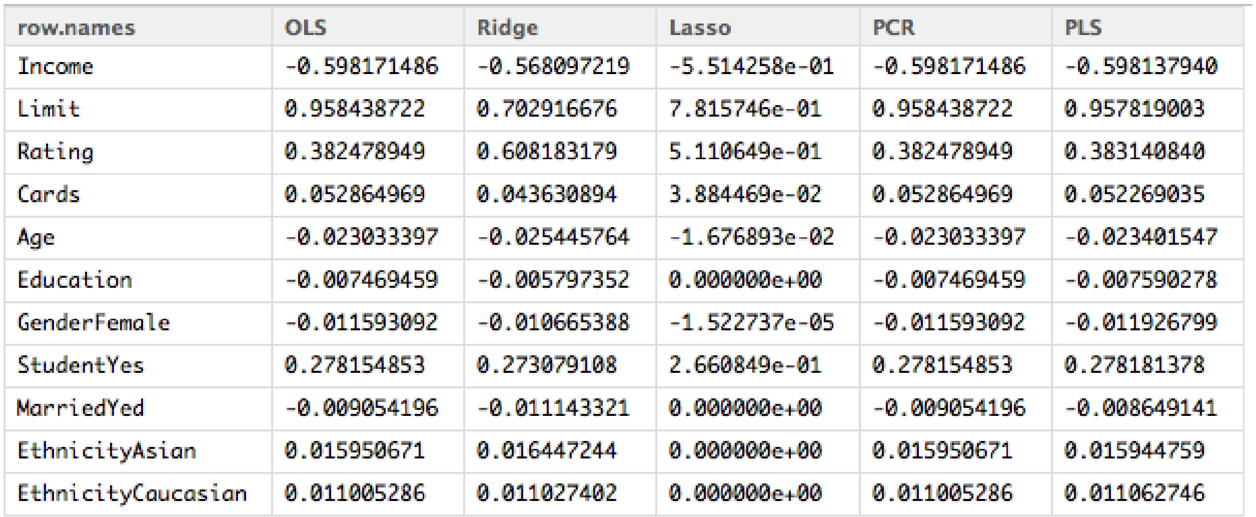
The following table is the test Mean Square Error for the four methods.



row.names	Test MSE Values
Ridge	0.05417609
Lasso	0.05414370
PCR	0.05814687
PLS	0.05717022

From the table, we can see that the Lasso Regression has the minimum test Mean Squared Error in this case, followed by Ridge Regression. The Dimension Reduction Methods in this case do not perform as well as Shrinkage Methods.

Then to see which predictors matter most in prediction, we visualize the coefficients of predictors within each method as a table.

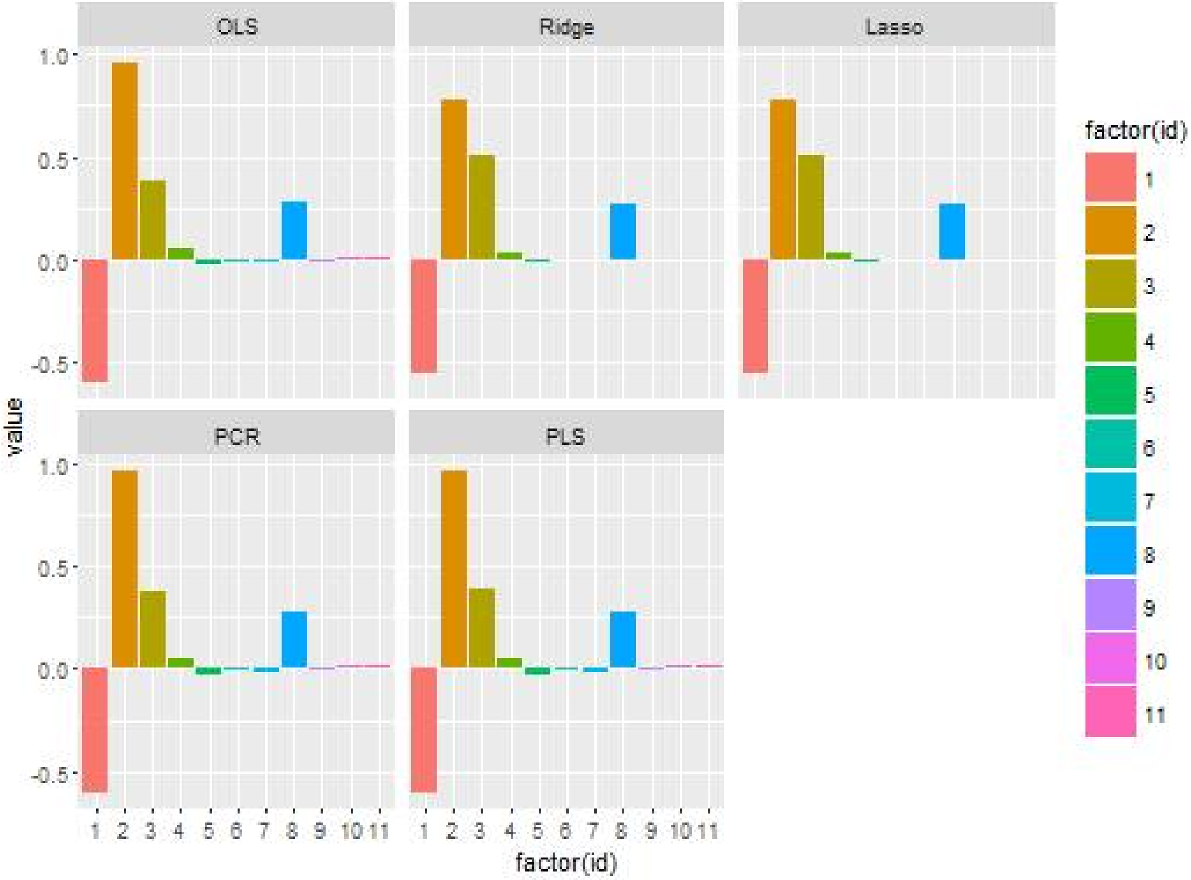


row.names	OLS	Ridge	Lasso	PCR	PLS
Income	-0.598171486	-0.568097219	-5.514258e-01	-0.598171486	-0.598137940
Limit	0.958438722	0.702916676	7.815746e-01	0.958438722	0.957819003
Rating	0.382478949	0.608183179	5.110649e-01	0.382478949	0.383140840
Cards	0.052864969	0.043630894	3.884469e-02	0.052864969	0.052269035
Age	-0.023033397	-0.025445764	-1.676893e-02	-0.023033397	-0.023401547
Education	-0.007469459	-0.005797352	0.000000e+00	-0.007469459	-0.007590278
GenderFemale	-0.011593092	-0.010665388	-1.522737e-05	-0.011593092	-0.011926799
StudentYes	0.278154853	0.273079108	2.660849e-01	0.278154853	0.278181378
MarriedYed	-0.009054196	-0.011143321	0.000000e+00	-0.009054196	-0.008649141
EthnicityAsian	0.015950671	0.016447244	0.000000e+00	0.015950671	0.015944759
EthnicityCaucasian	0.011005286	0.011027402	0.000000e+00	0.011005286	0.011062746

From the table, we can see that in all five methods we used for prediction, Limit is the single most important predictor for Balance which corresponds to our common sense that amount of limit is usually highly correlated to amount of Balance.

For Ridge Regression, the second and third important factors are Rating and Income. For Lasso Regression, the second and third important factors are Income and Rating. For Pricipal Components Regression, the second and third important factors are Income and Rating. For Parse Least Squares Regression, the second and third important factors are Income and Rating. Hence we can conclude that the three leading factors are Limit, Income, and Rating.

To get a better intuition of the importance of each factor, we visualize the coefficients of each predictor in each five methods in bar plots.



Conclusion

In this project, we performed a data analysis cycle which contains getting raw datasets, data cleaning and processment, exploratory data analysis, modeling and tuning parameters, visualization of results, report and presentation.

We replicated major findings on chapter 6: Linear Model Selection and Regularization (from "An Introduction to Statistical Learning" by James et al) with the dataset "Credit". we performed five algorithms: Least Square Regression, Ridge Regression, Lasso Regression, Principal Components Regression, and Partial Least Squares Regression.

From our result, we found that Lasso Regression performs best, followed by Ridge Regression using the measure of test Mean Squared Error. Dimension Reduction Regression, in this case, do not perform as well as Shrinkage Methods. The three leading predictors in prediction are Limit, Income and Rating.

References Lecture slides of Stat 159

"An Introduction to Statistical Learning" by James et al