Learning R for Data Analysis: Project One

Tshepo Ralehoko

24 June 2018

Contents

List of Figures

List of Tables

Acknowledgements

I want to acknowledge the useful of the book Elements of Statistical Learning for the theory that is related to the results herein, analysis and interpretation thereof. The book has been a great resource for further developing my statistical computing skills in R and data analysis in general. Furthermore, many thanks to various platforms which are easily accessible, and with great provision and support in R related content for data analysis.

List of Notations

σ^2 : Variance

σ : Standard Deviation

μ : Mean

List of Keywords

Bias-Variance Trade-off

Correlation Matrix

Goodness of fit

Kurtosis

Probability Density Curve

Random Variable

Skewness

Training set

Test set

Configurations

Working directory

Below is the directory that was created for the project as it pertains to the laptop that was used. This can be changed accordingly depending on where the user wants to save their work. We also go ahead and load the dataset.

Introduction

This is a personal project. The project deals with the well-known *Credit* dataset. A brief description of the dataset shall follow. The aim of the project is to build the best model for predicting the output variable using, all or a subset of the features. On that note, we wish to indicate that the dataset we shall be dealing with falls into the *supervised learning* paradigm. For assessing the accuracy of the models in predicting the corresponding *target* variable, we will generate and utilize the necessary *goodness of fit* statistics. We will also keep an eye of the *bias-variance trade-off* during the model building process. This concept is explained in detail under the **Data Modelling** section. Furthermore, we want to underscore that the project is for learning purposes, and as a result, any constructive input is appreciate.

For achieving the aim of the project, our dataset will be randomly split into a *training* and *test* set. We will sometimes refer to the latter set as the *validation* set. This method is widely used for validating the accuracy and performance of the model in predicting observations that were not used in building or training the model (out-of-sample observations).

The next section will take a look at **Data Description and Data Summary**. It will use various functions to study the structure of the dataset; the variables that make up the dataset and summary statistics. The **Data Preparation** section is dedicated to addressing any issues that we might have discovered in the preceding section, and taking the corrective steps to prepare the dataset for model building and analysis.

Data Description and Data Summary

The dataset has *11* predictor variables, and each of the variables contains *400* observations. The names of the features are: Income (in thousands of dollars), Limit (credit limit), Rating (credit rating), Cards (number of credit cards), Age, Education (years of education), Gender, Student (student status), Married (marital status) and Ethnicity (Caucasian, African American or Asian). The class of the variables is split among *integer*, *factor* and *numeric* variables. The dataset has complete cases. For convenience, we will sometimes use the variable names (which provide less description about the variables) instead of the relevant description of the variables. For instance, we might use Education to refer to "the number of years of education".

The results from this section are very insightful, and allows the user to pursue other data mining techniques on the dataset that are beyond the scope covered by the project. This is done to accommodate any further analysis that may be of interest on the dataset in the future. I want to reiterate that, with this project, I desire to take a pragmatic approach and learn new skills beyond what I have gathered from the classroom environment during the course of my studies in Data Science. We now focus our attention to the plots, figures and tables from the *R* output.

In search of discovering interesting insights into the dataset, we decided to plot the column means by both gender and marital status. Below in figure 1 and figure 2 we have the plots of the *column means* by gender and marital status respectively. We have only used numerical variables for the below stacked barplots.

From the plots we can also see that the average of most of the variables is very small across gender and marital status. Further insights from table 2 shows that this is in fact the case for the *column means* for a few variables when ignoring the groupings by gender and marital status. The Limit variable dominates both stacked barcharts with its large mean. Taking a closer look, the average credit limit of females is greater than that of males.

Figure 1: A baplot - Column Means by Gender

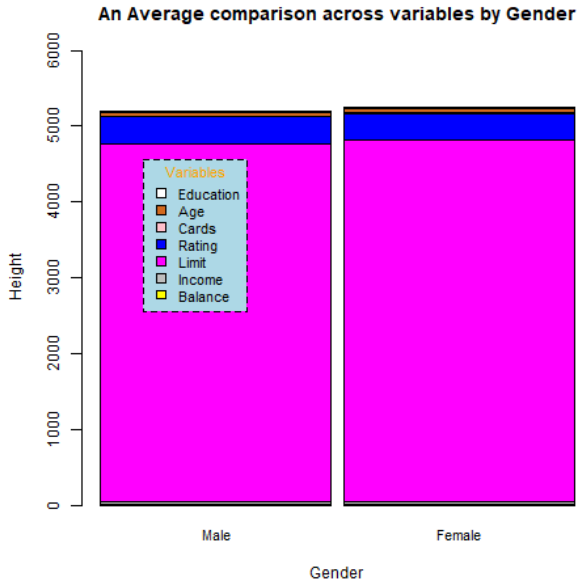
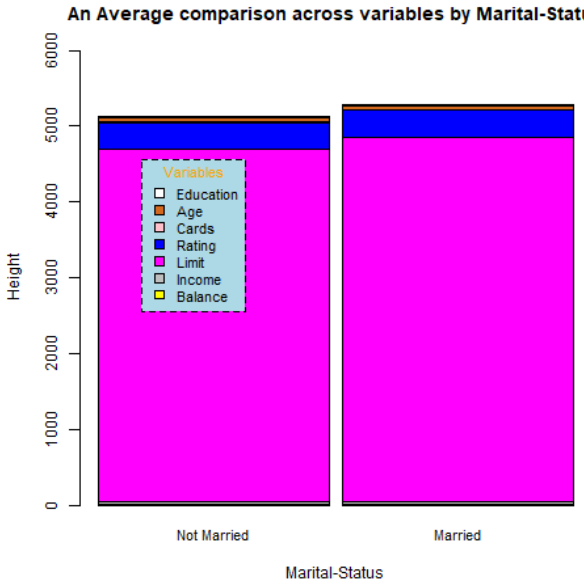


Figure 2: A barplot - Column Means by Marital Status



Below is a box-and-whisker diagram of the numeric variables. We have also marked the outliers (in asterisk-like characters) using a magenta colour. Certainly, these variables can be thought of as *random variables*. In this light, the plot also plays an important role in aiding us to get a rough idea of the distribution of our random variables. It is clear from the figure that the *Limit* predictor variable has a distribution whose underlying statistics can be uniquely identified in this case. It is characterized by a large variance and mean and several outliers on the *upper fence*.

The lower fence and upper fence are situated below the whisker at the bottom of the box and above the whisker at the top side of the box respectively. The respective values for these fences are computed as follows:

$$\text{Lower fence} = Q_1 - 1.5 \times IQR$$

$$\text{Upper fence} = Q_3 + 1.5 \times IQR$$

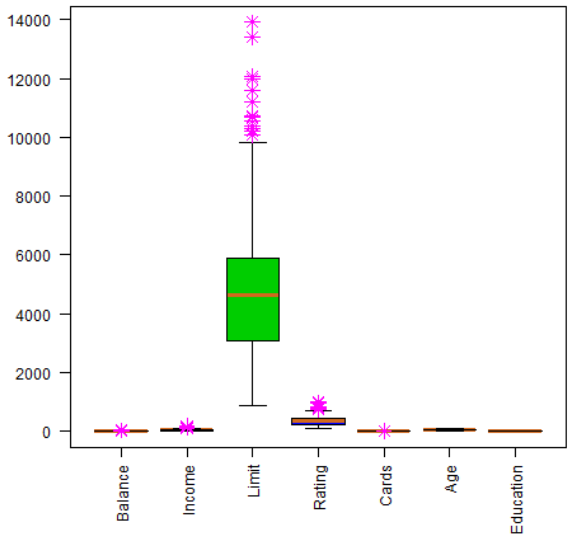
where,

- Q_1 is the first or lower quartile

- Q_3 is the third or upper quartile
- IQR is the interquartile range which is obtained by subtracting Q_1 from Q_3

It is important to keep the outliers in mind when building models. Outliers are generally undesirable and need to be scrutinized in the model building process. Bearing in mind that these are observations that do not fit the general pattern observed in the dataset, they can cause misleading interpretation. For instance, a case could arise where a model is rejected due to a violation of model assumptions caused by outliers, when in actual fact, the correct model is chosen for the dominant pattern of observations in the dataset. This discussion pertains to figure 3 below.

Figure 3: Box and whisker plot - Numeric Variables



The next figure looks at the column variances of our numeric variables. The results below in figure 4 are not surprising. Similar information can be seen from the box-and-whisker plot in figure 3. Therefore, for some analysis, it might be a good idea to standardize the variables so that no one variable is dominant over the others. In any case, we will not consider scaling or standardizing the dataset.

Figure 4: A barplot Depicting Column Variances

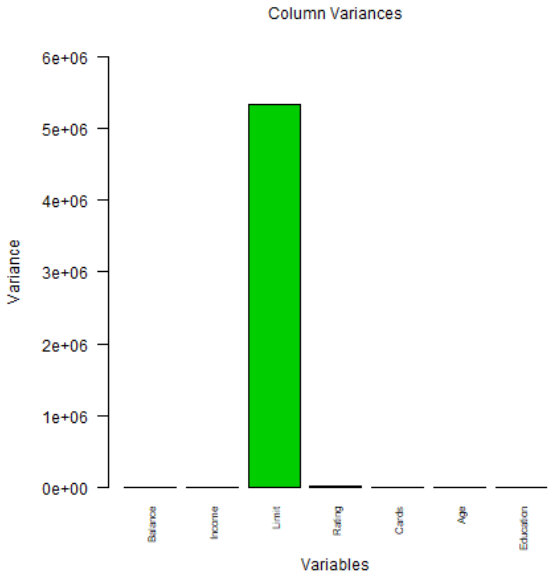


Figure 5 is a plot that represents pairwise scatter plots of the variables. Included in the scatter plots is a "smooth" curve that is fitted to the data points. There is clearly a *linear* association between *Limit* and *Rating*. This is because the data points from the corresponding scatter plot of the two variables can be determined using a liner model that is approximately deterministic . Additionally, similar associated is observed between *Income* and *Limit* and between *Income* and *Rating*, but the strength of the relationships is not as strong.

The strength of the associations between various pairs of variables are found in the correlation matrix in table 1. In reality, we can expect credit card limit to be proportional to income. However, a domain expert would not more about these intricacies. This phenomenon of association between features is known as *collinearity*. From the plot, we also see that quite a number of features seem to be linearly related to the target variable.

When the data points from the pairwise plots are plotted by gender or marital status, it would seem that it is a challenge to pick up any apparent pattern between the variables across both the levels of the gender and marital status factor variable. For this information we refer to figure 6 and figure 7

Figure 5: Pairwise Scatter Plots of Numeric Variables

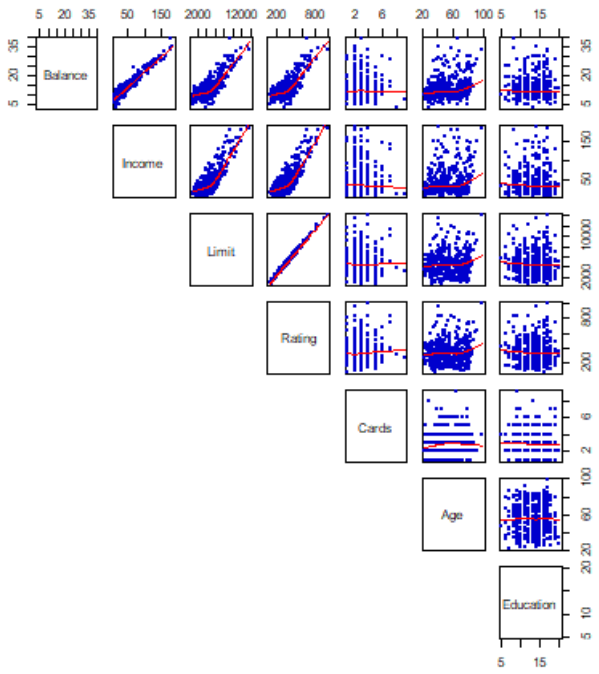


Figure 6: Pairwise Plots - Points Plotted by Marital-Status

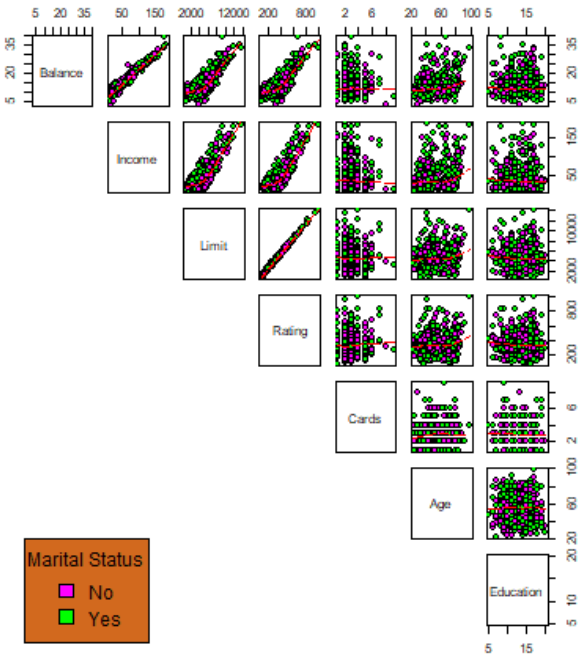


Figure 7: Pairwise Plots - Points Plotted by Gender

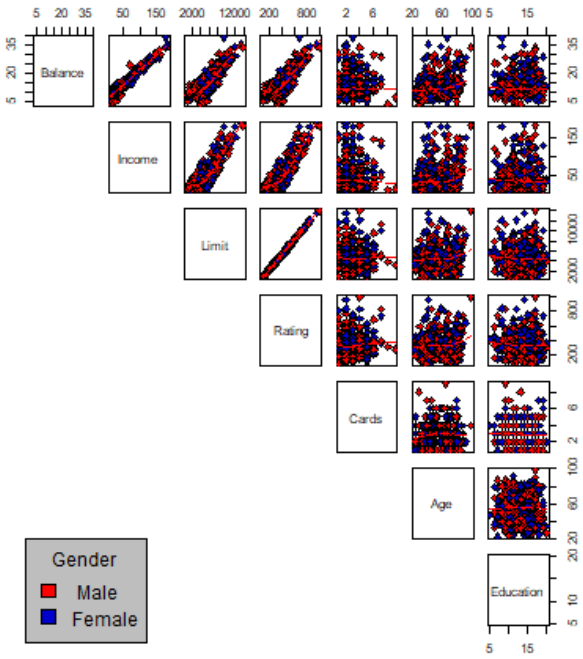


Table 1 below shows the output from the correlation matrix of the numeric variables. It is a measure of the strength of the linear association between the combination of pairwise scatter plots of the numeric variables in the dataset.

Table 1: Correlation Matrix - Tabular representation

Variable	Balance	Income	Limit	Rating	Cards	Age	Education
Balance	1.00	0.97	0.76	0.76	-0.01	0.23	0.01
Income	0.97	1.00	0.79	0.79	-0.02	0.18	-0.03
Limit	0.76	0.79	1.00	1.00	0.01	0.10	-0.02
Rating	0.76	0.79	1.00	1.00	0.05	0.10	-0.03
Cards	-0.01	-0.02	0.01	0.05	1.00	0.04	-0.05
Age	0.23	0.18	0.10	0.10	0.04	1.00	0.00
Education	0.01	-0.03	-0.02	-0.03	-0.05	0.00	1.00

Table 2: Distribution Statistics

	σ^2	σ	μ	minimum	maximum	range	Q_1	Q_2	Q_3	IQR	kurtosis	skewness
Balance	32.14	5.67	13.43	3.75	38.79	35.04	9.89	11.78	15.24	5.35	2.58	1.54
Income	1242.16	35.24	45.22	10.35	186.63	176.28	21.01	33.12	57.47	36.46	2.87	1.73
Limit	5327781.92	2308.20	4735.60	855.00	13913.00	13058.00	3088.00	4622.50	5872.75	2784.75	0.96	0.83
Rating	23939.56	154.72	354.94	93.00	982.00	889.00	247.25	344.00	437.25	190.00	1.01	0.86
Cards	1.88	1.37	2.96	1.00	9.00	8.00	2.00	3.00	4.00	2.00	0.90	0.79
Age	297.56	17.25	55.67	23.00	98.00	75.00	41.75	56.00	70.00	28.25	-1.08	0.01
Education	9.77	3.13	13.45	5.00	20.00	15.00	11.00	14.00	16.00	5.00	-0.60	-0.33

Figure 8: Probability Density Curve of Balance

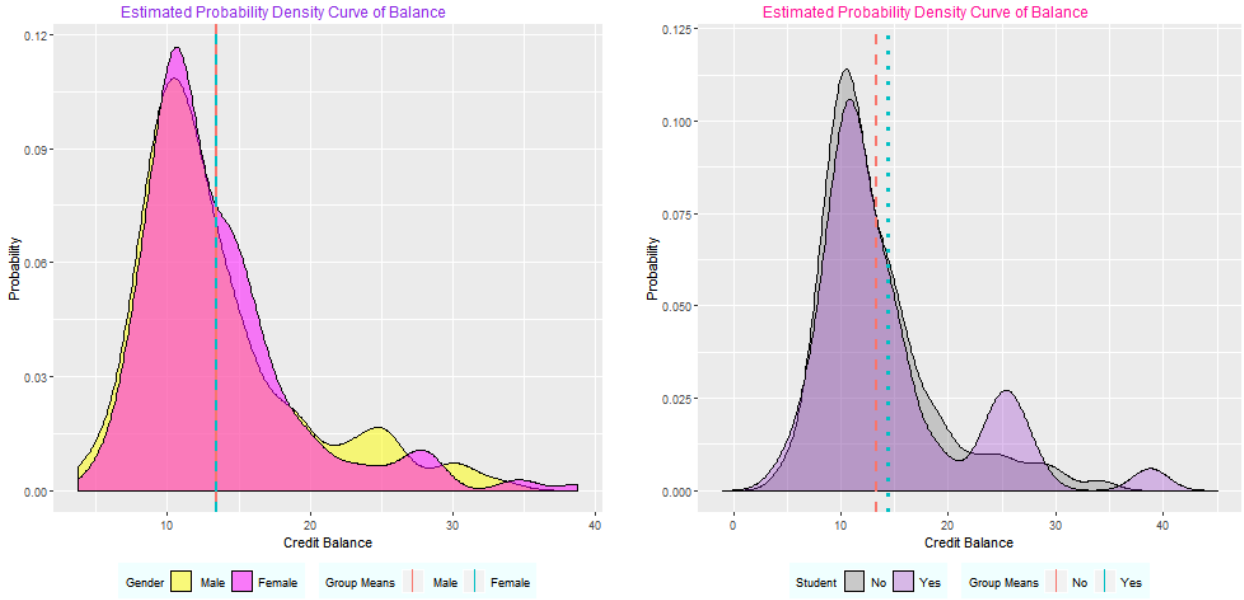


Figure 9: Probability Density Curve of Income

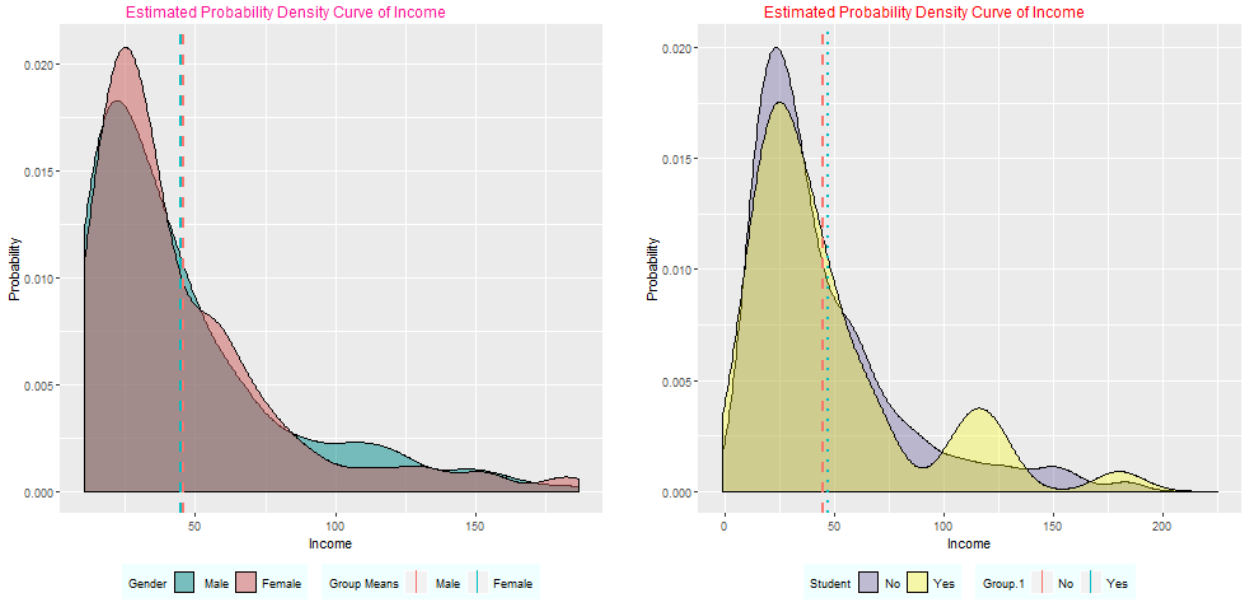


Figure 10: Probability Density Curve of Limit

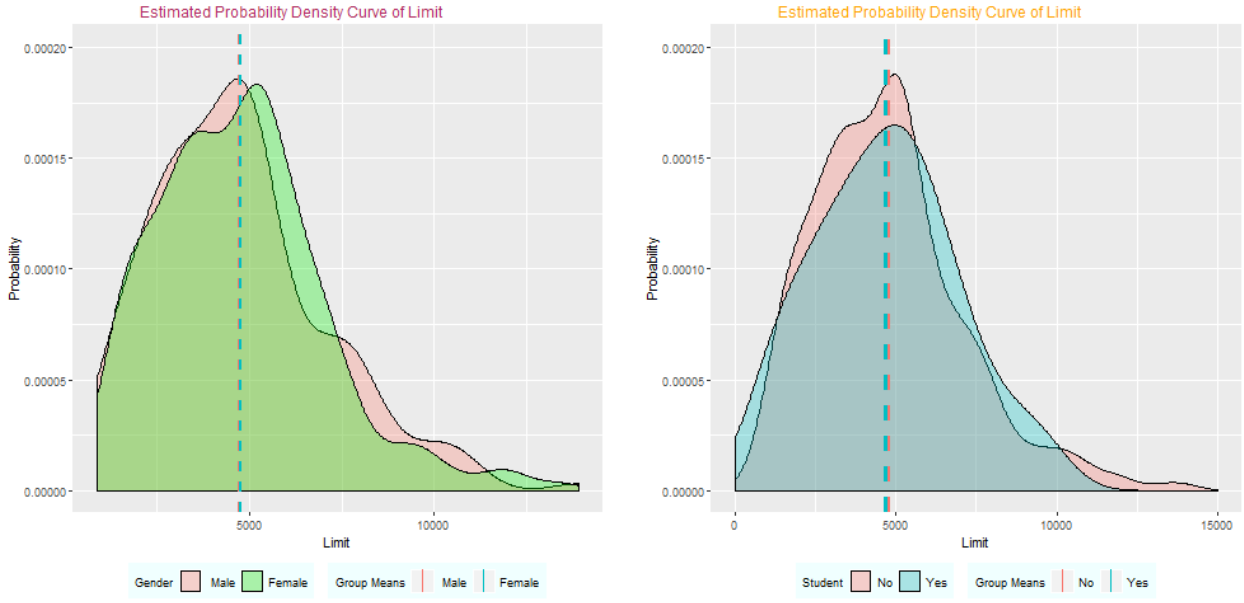


Figure 11: Probability Density Curve of Rating

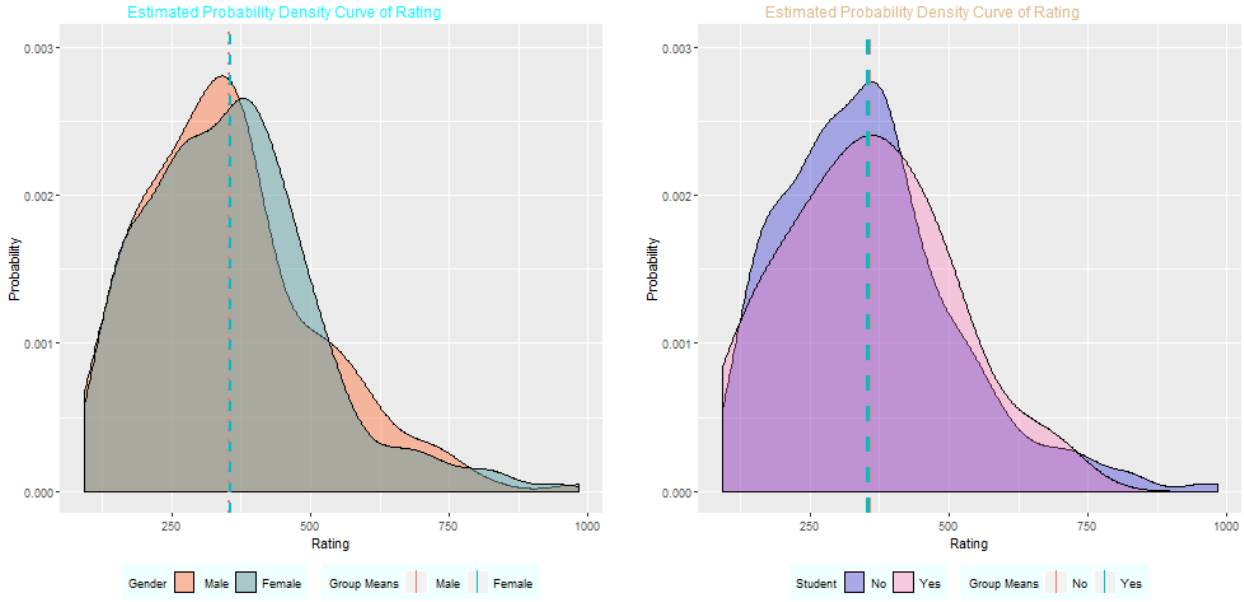


Figure 12: Probability Density Curve of Cards

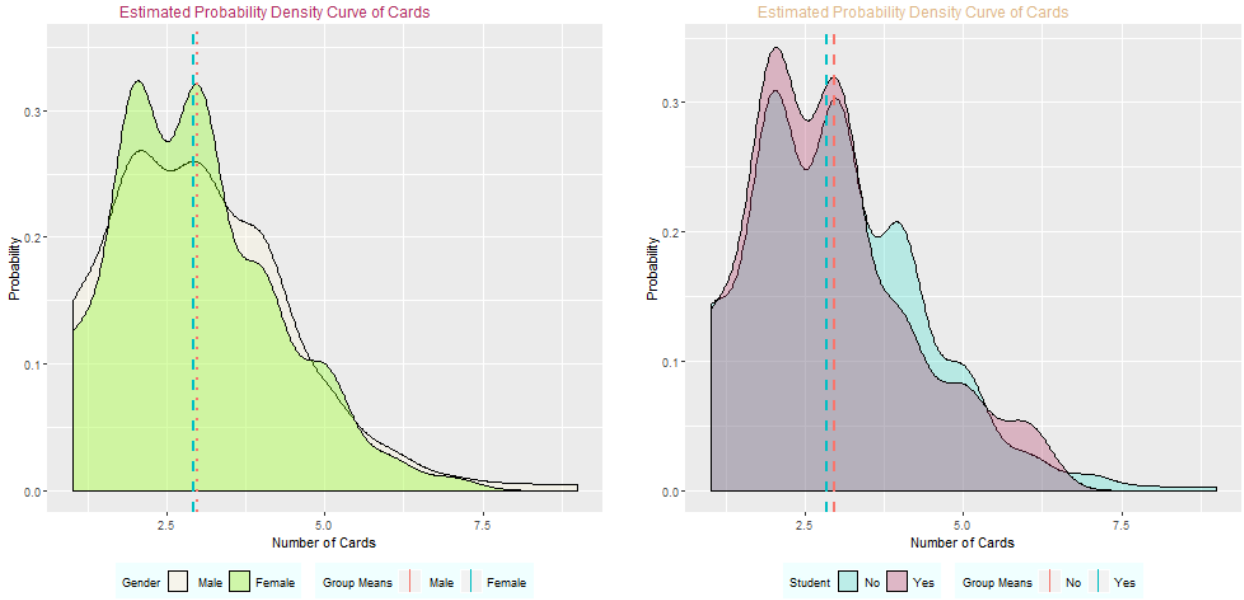


Figure 13: Probability Density Curve of Age

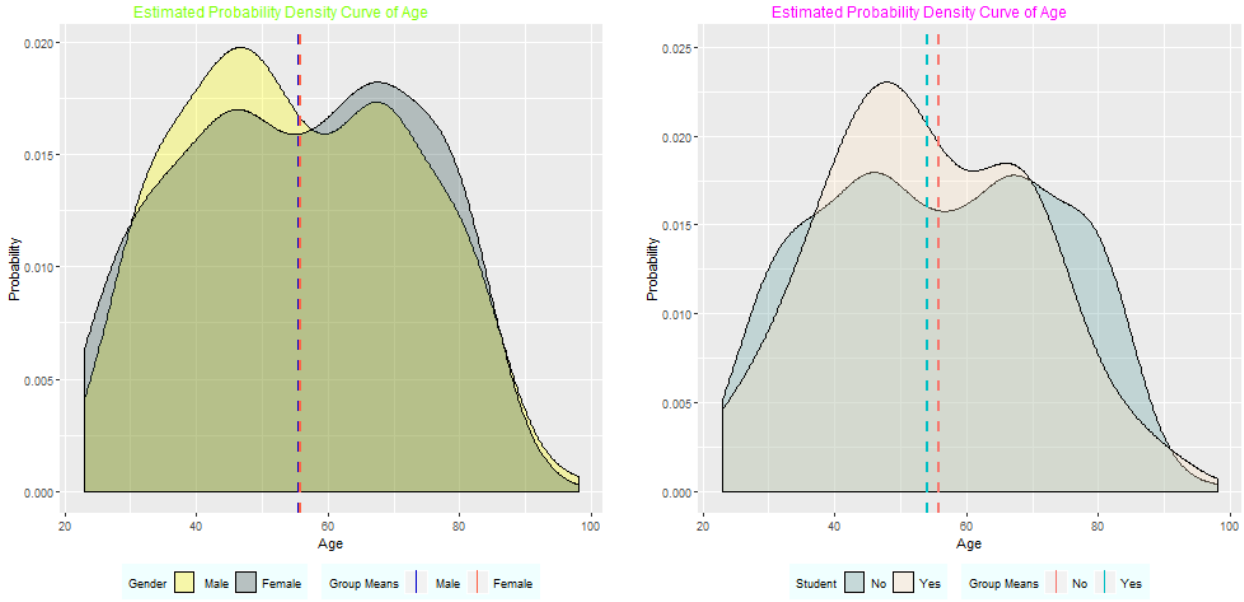


Figure 14: Probability Density Curve of Education

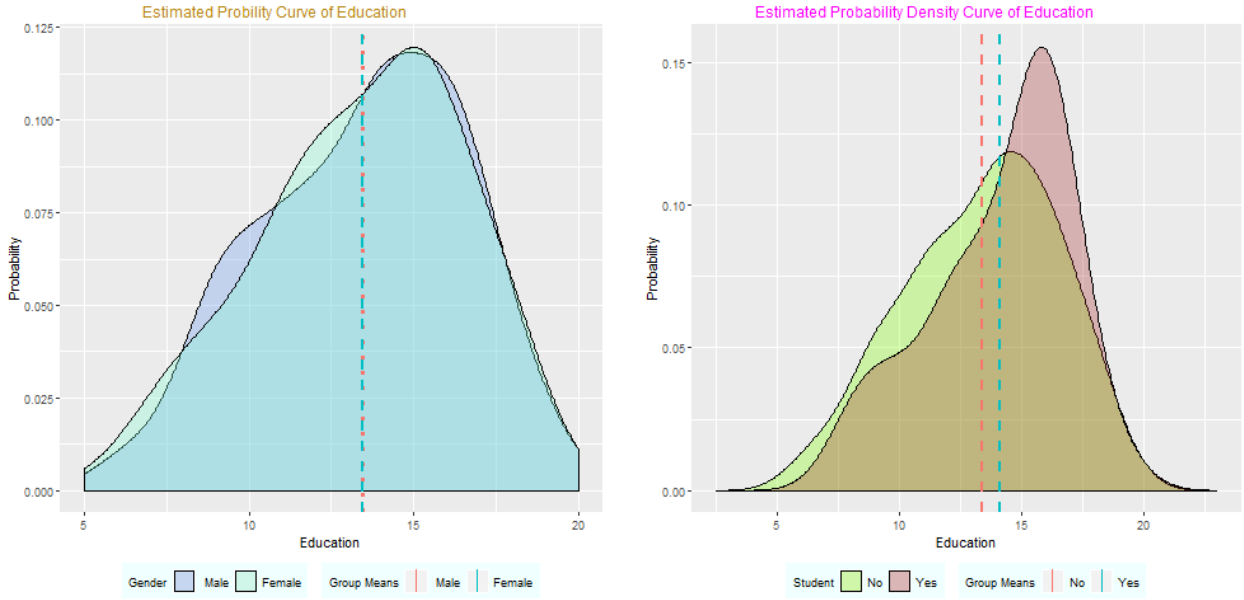


Figure 15: Boxplot of Balance Random Variable

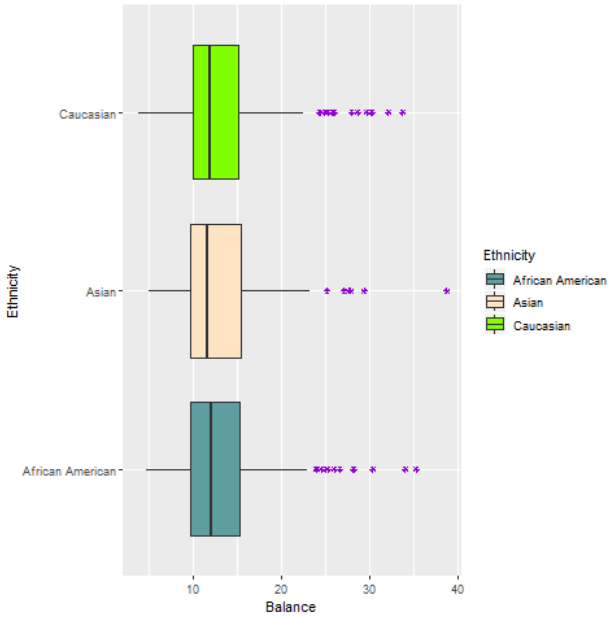


Figure 16: Boxplot of Income Random Variable

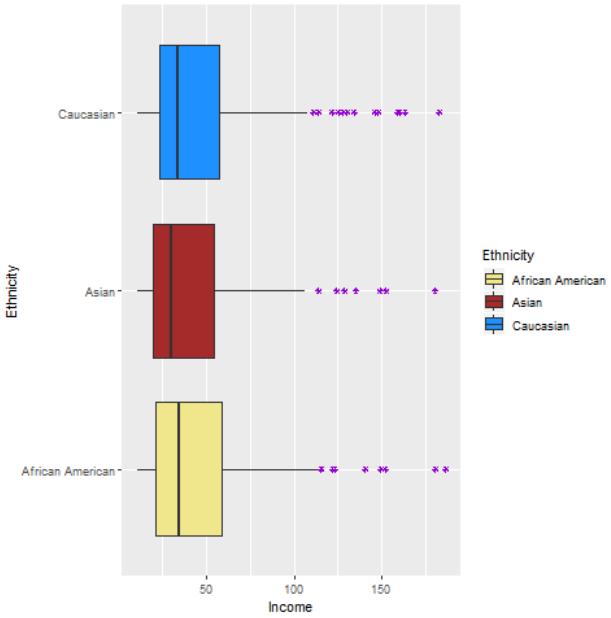


Figure 17: Boxplot of Limit Random Variable

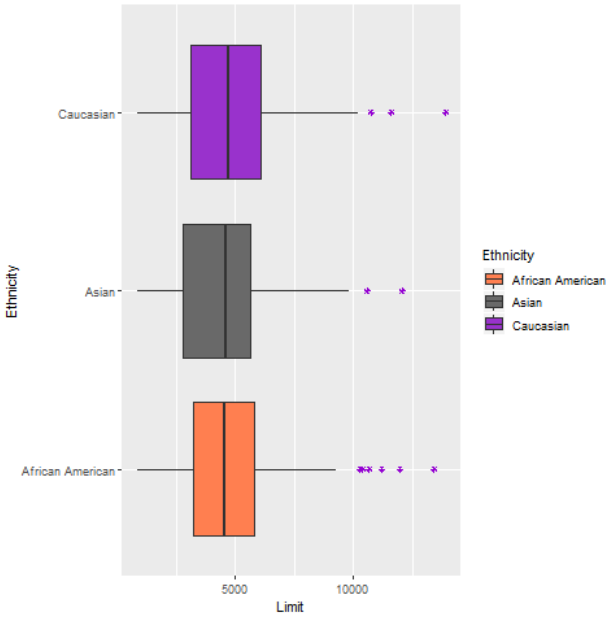


Figure 18: Boxplot of Rating Random Variable

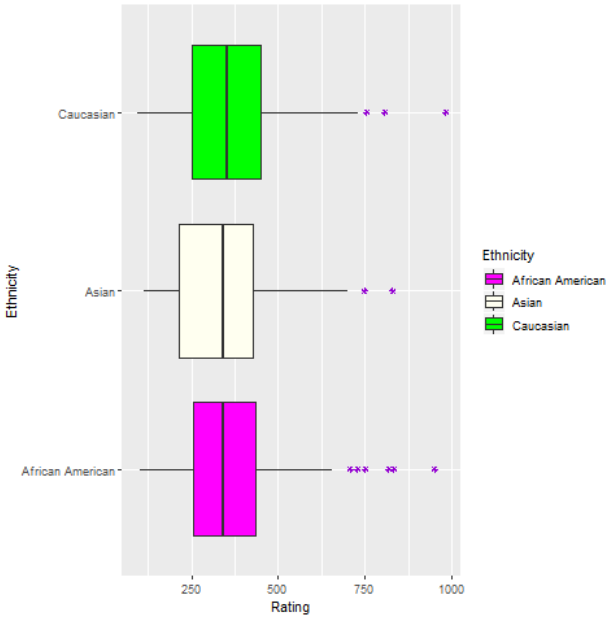


Figure 19: Boxplot of Cards Random Variable

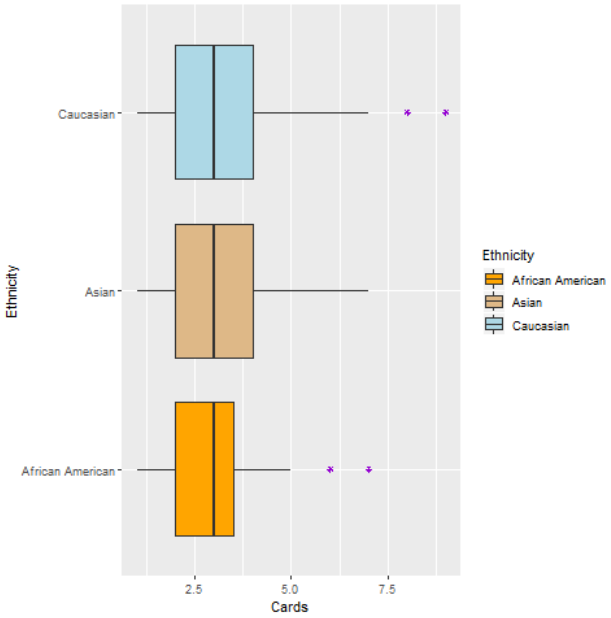


Figure 20: Boxplot of Age Random Variable

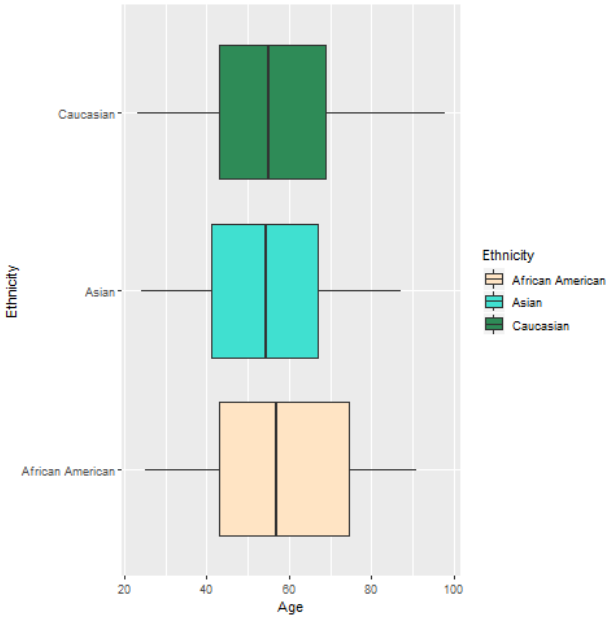
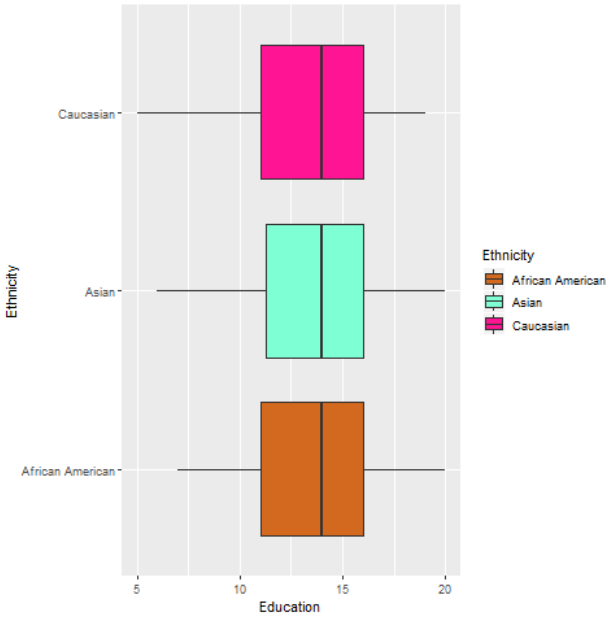
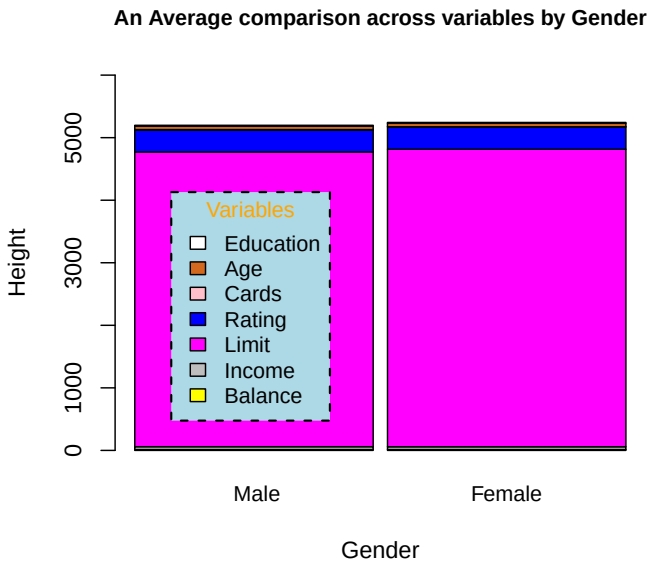
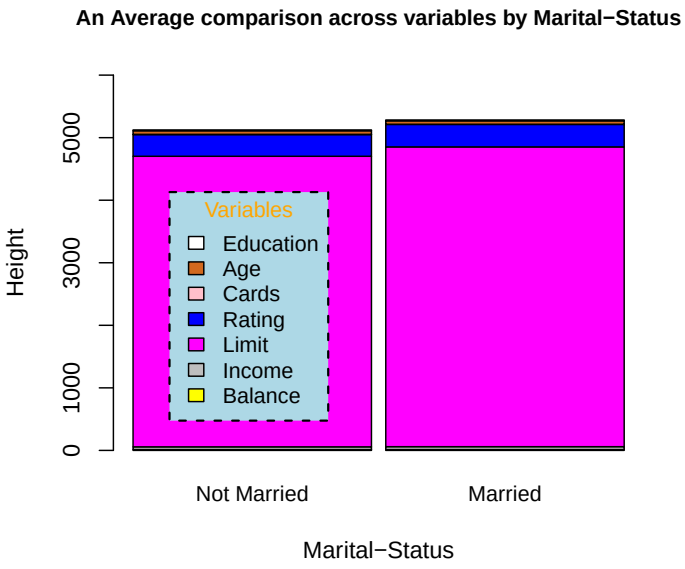
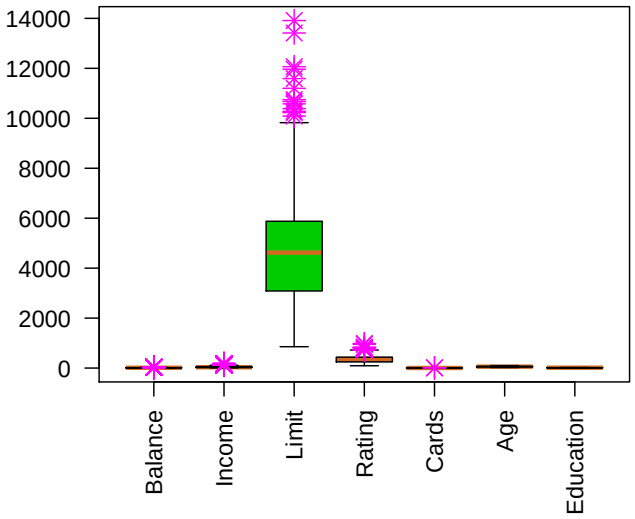


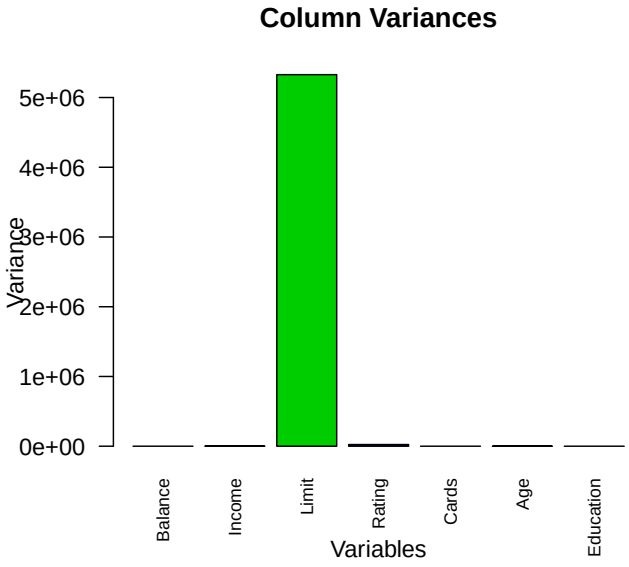
Figure 21: Boxplot of Education Random Variable

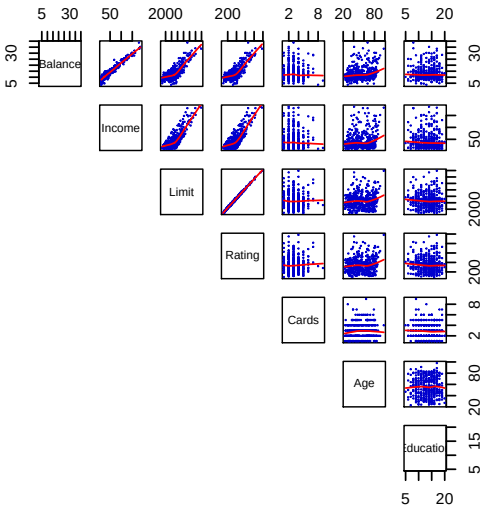


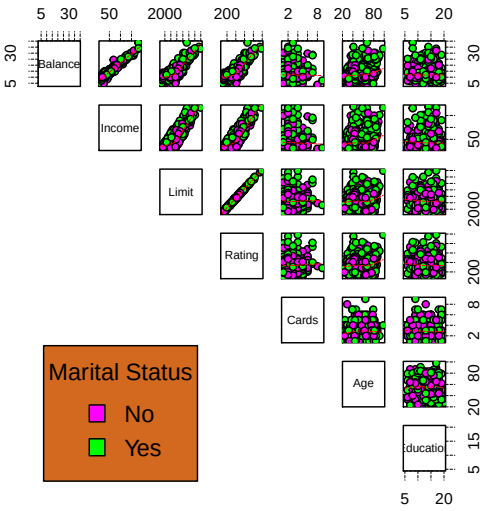

An Average comparison across variables by Gender
Height
0
1000
3000
5000
Variables
Education
Age
Cards
Rating
Limit
Income
Balance
Male
Female
Gender

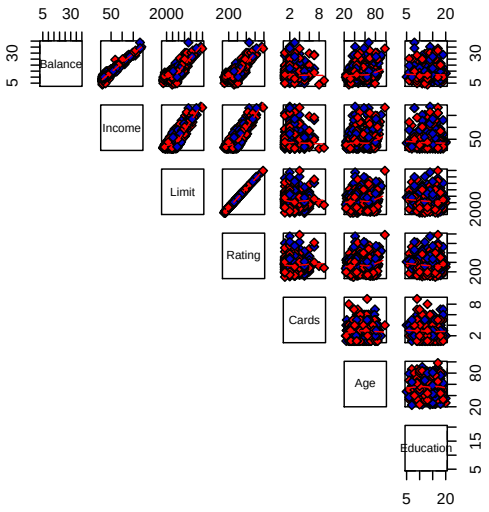

An Average comparison across variables by Marital−Status
Height
0
1000
3000
5000
Variables
Education
Age
Cards
Rating
Limit
Income
Balance
Not Married
Married
Marital−Status

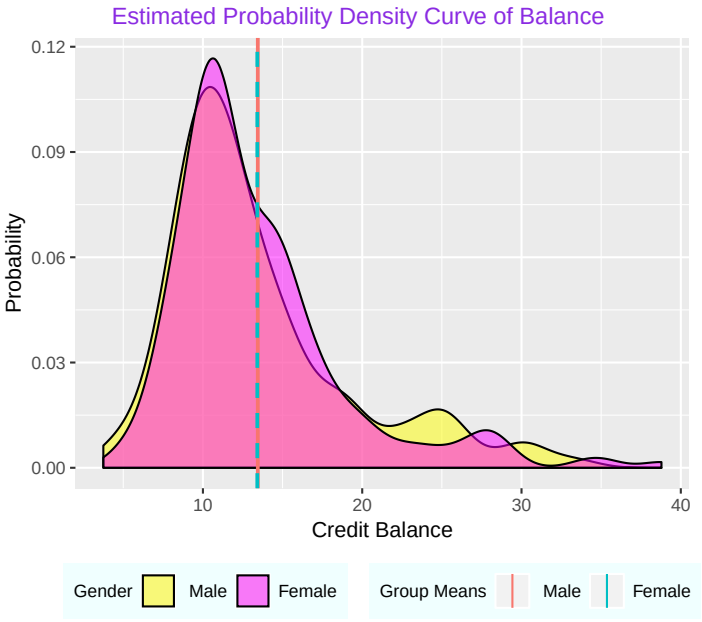

14000
12000
10000
8000
6000
4000
2000
0
Balance
Income
Limit
Rating
Cards
Age
Education

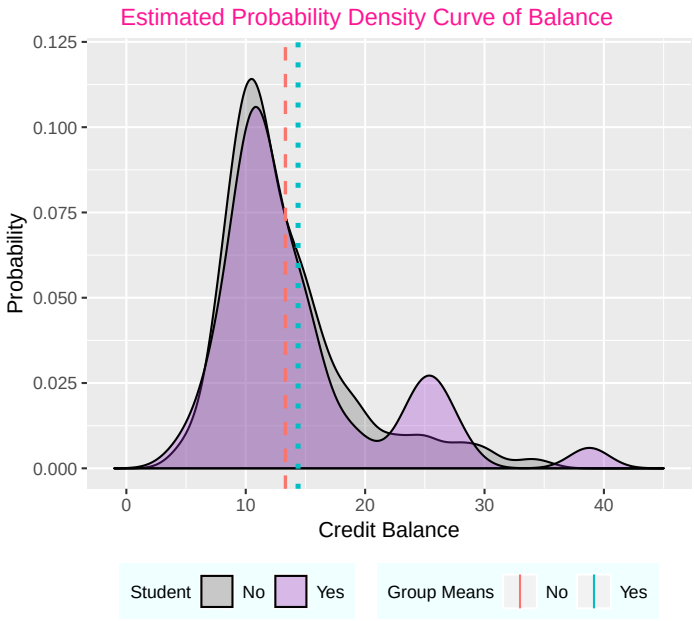

Column Variances
Variance
5e+06
4e+06
3e+06
2e+06
1e+06
0e+00
Balance
Income
Limit
Rating
Cards
Age
Education
Variables

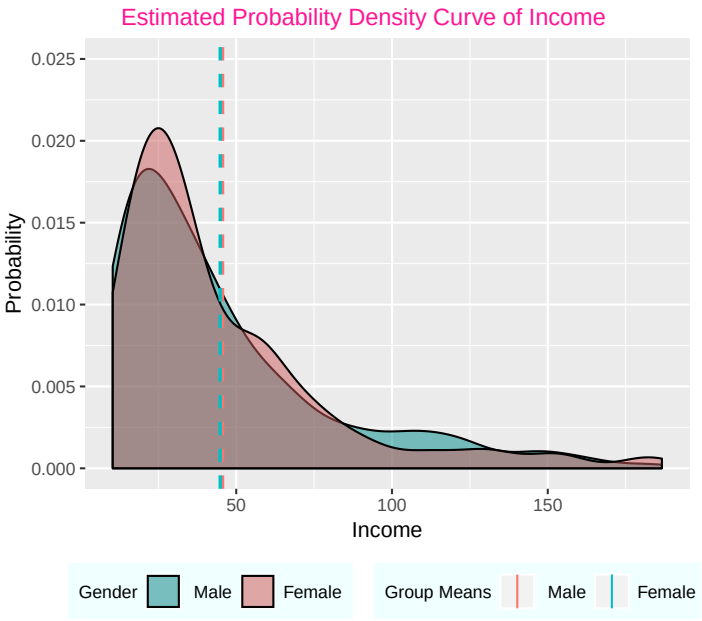

Balance
Income
Limit
Rating
Cards
Age
Education

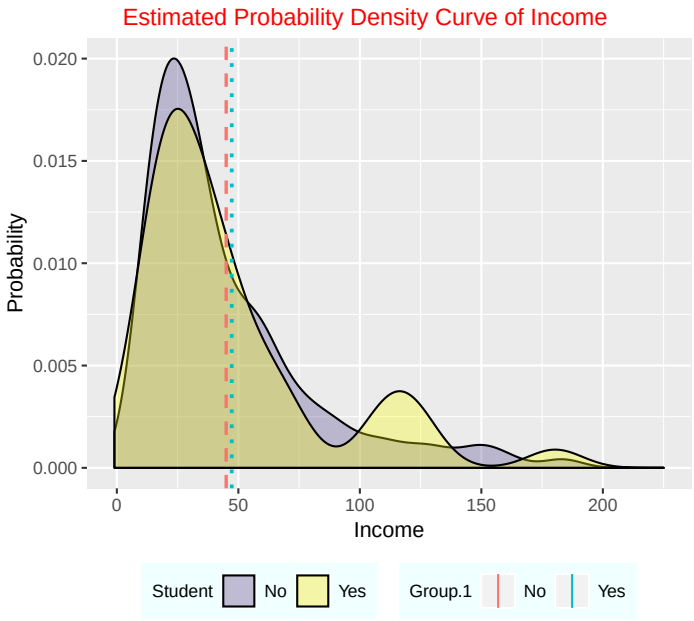

Balance
Income
Limit
Rating
Cards
Age
Marital Status
No
Yes

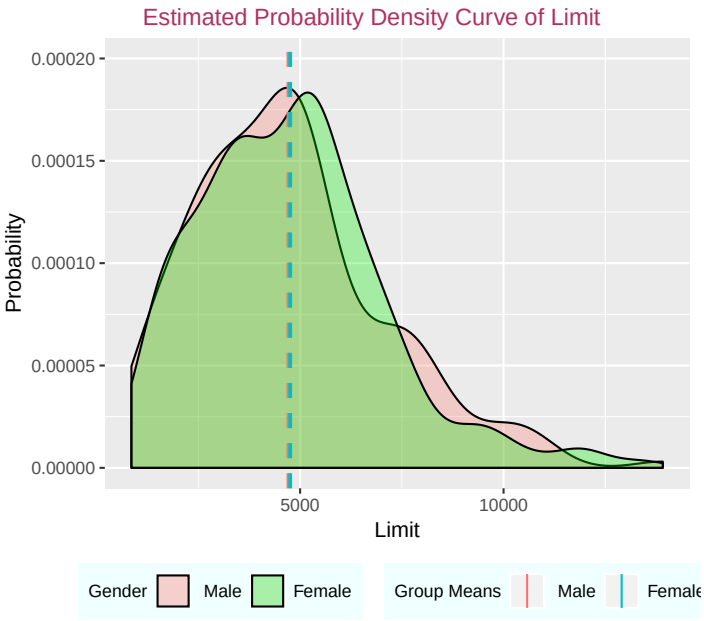

Balance
Income
Limit
Rating
Cards
Age
Education

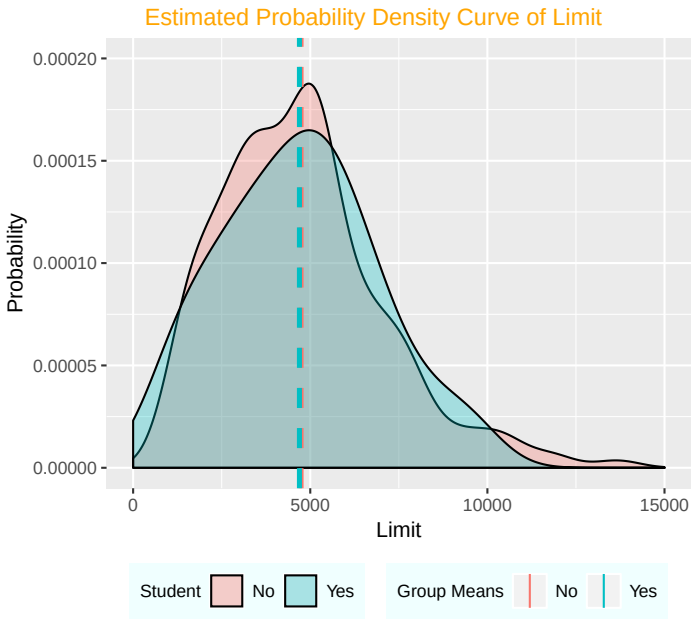

Estimated Probability Density Curve of Balance
Probability
Credit Balance
Gender
Male
Female
Group Means
Male
Female

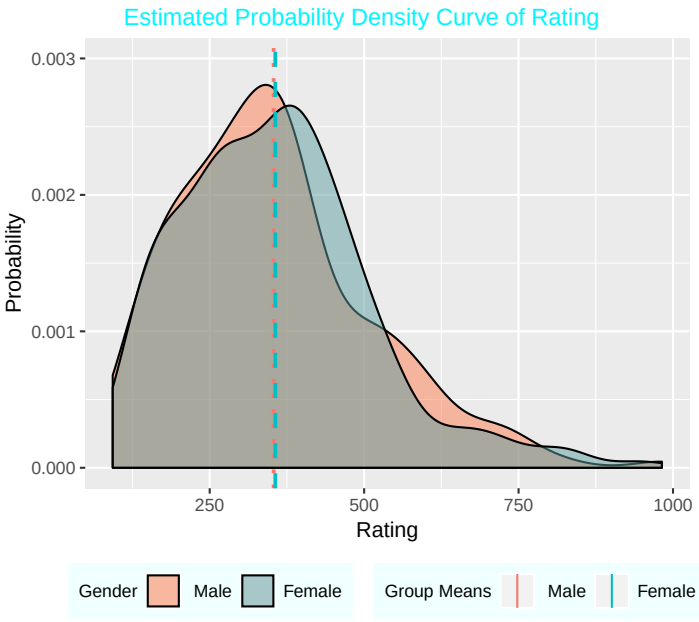

Estimated Probability Density Curve of Balance
Probability
Credit Balance
Student
No
Yes
Group Means
No
Yes

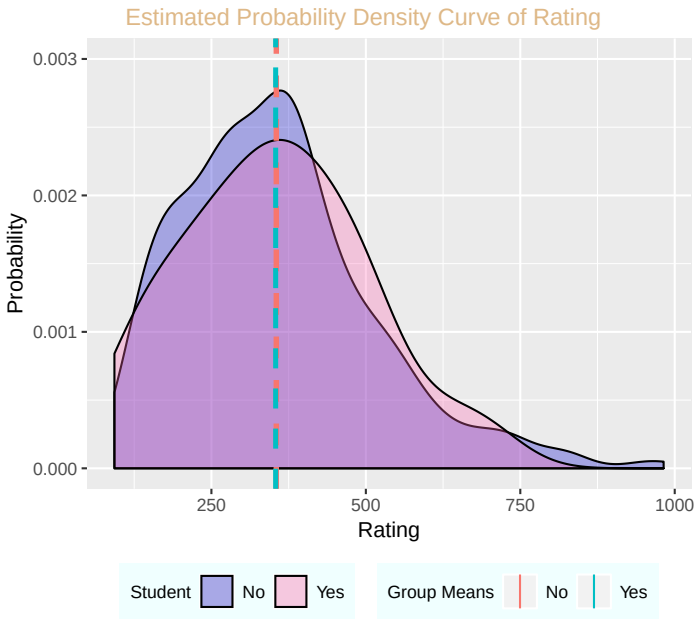

Estimated Probability Density Curve of Income
Probability
Income
Gender
Male
Female
Group Means
Male
Female

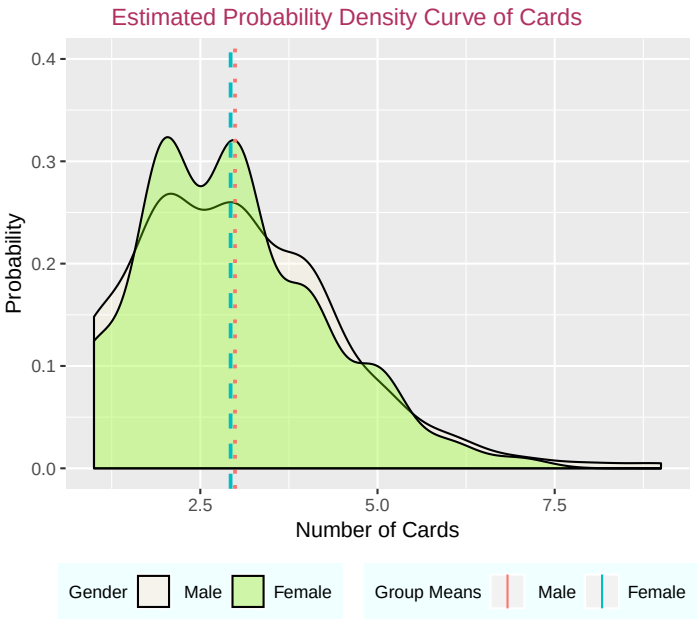

Estimated Probability Density Curve of Income
Probability
Income
Student
No
Yes
Group.1
No
Yes

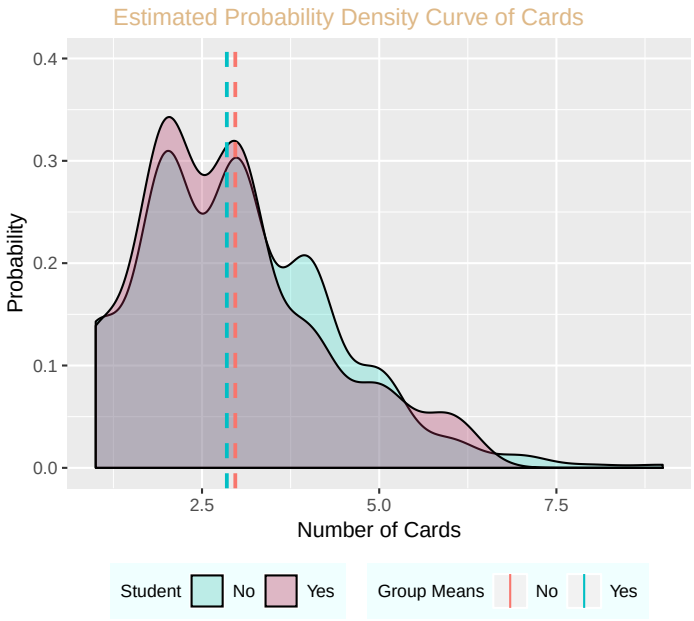

Estimated Probability Density Curve of Limit
Probability
Limit
Gender
Male
Female
Group Means
Male
Female

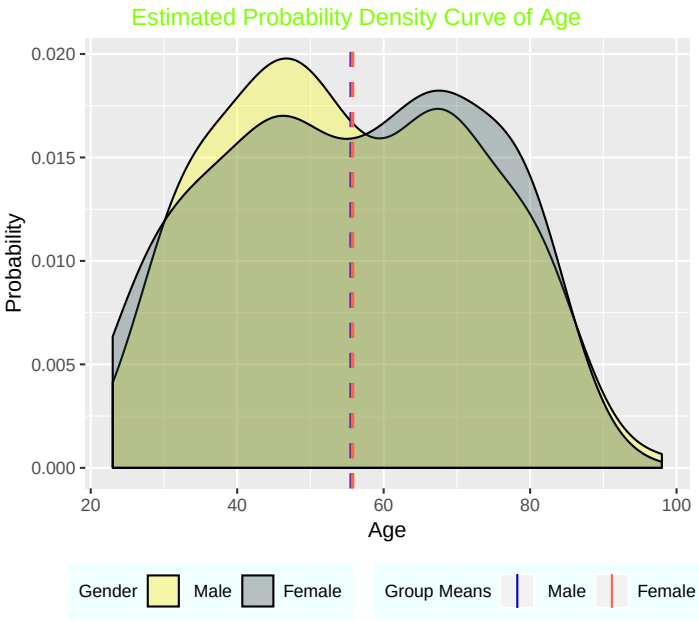

Estimated Probability Density Curve of Limit
Probability
0.00020
0.00015
0.00010
0.00005
0.00000
0
5000
10000
15000
Limit
Student
No
Yes
Group Means
No
Yes

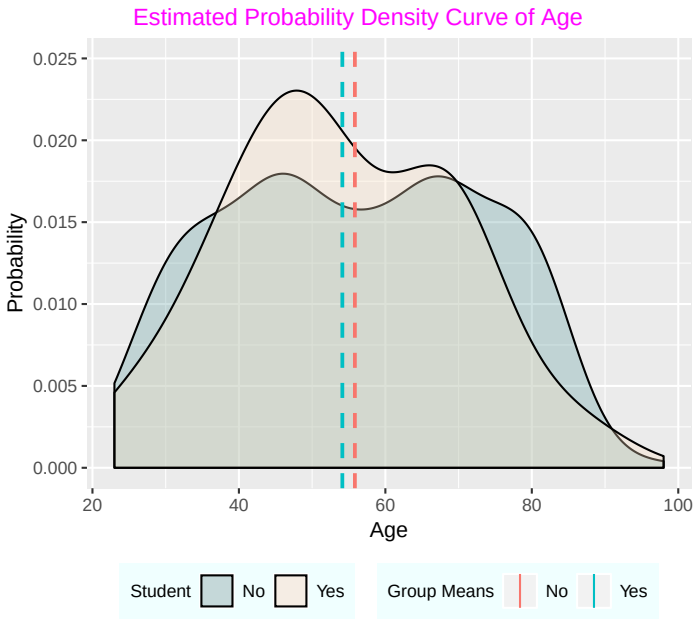

Estimated Probability Density Curve of Rating
Probability
0.003
0.002
0.001
0.000
250
500
750
1000
Rating
Gender
Male
Female
Group Means
Male
Female

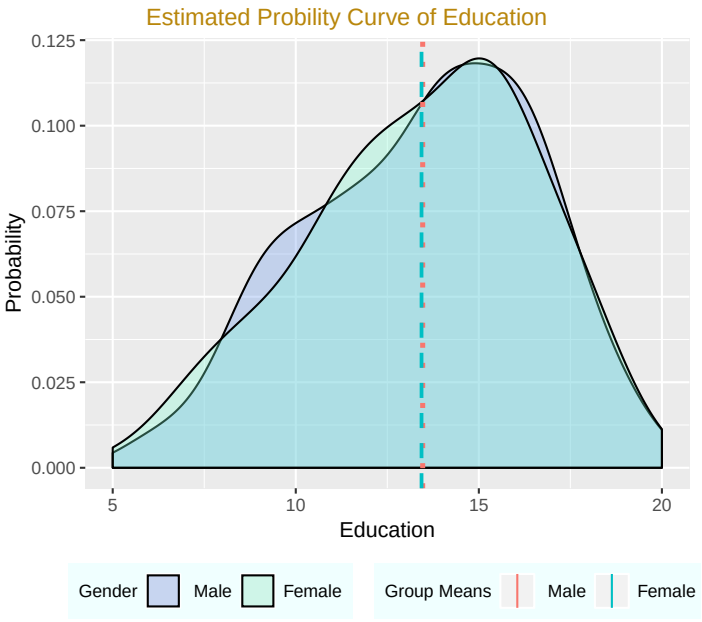

Estimated Probability Density Curve of Rating
Probability
0.003
0.002
0.001
0.000
250
500
750
1000
Rating
Student
No
Yes
Group Means
No
Yes

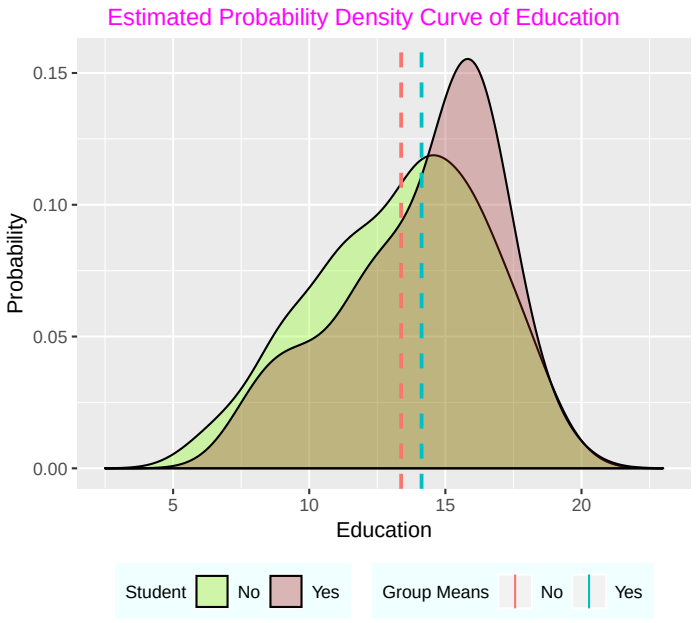

Estimated Probability Density Curve of Cards
Probability
0.4
0.3
0.2
0.1
0.0
2.5
5.0
7.5
Number of Cards
Gender
Male
Female
Group Means
Male
Female

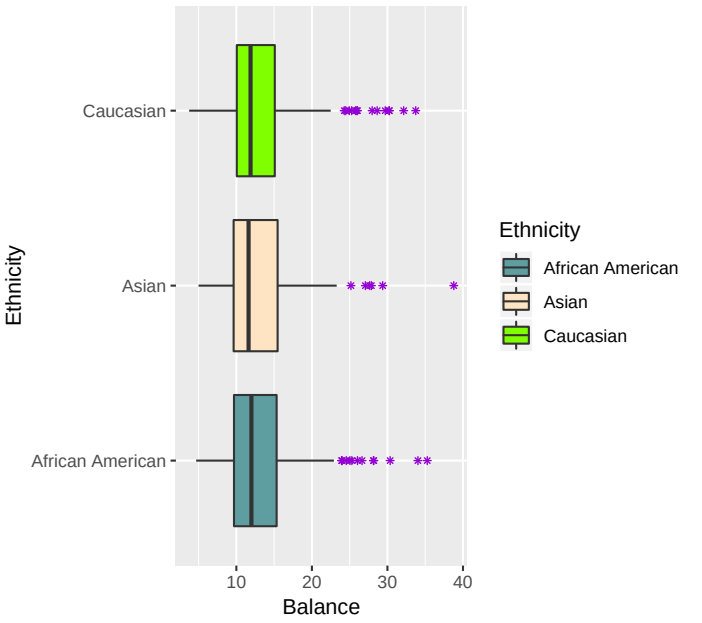

Estimated Probability Density Curve of Cards
Probability
0.4
0.3
0.2
0.1
0.0
2.5
5.0
7.5
Number of Cards
Student
No
Yes
Group Means
No
Yes

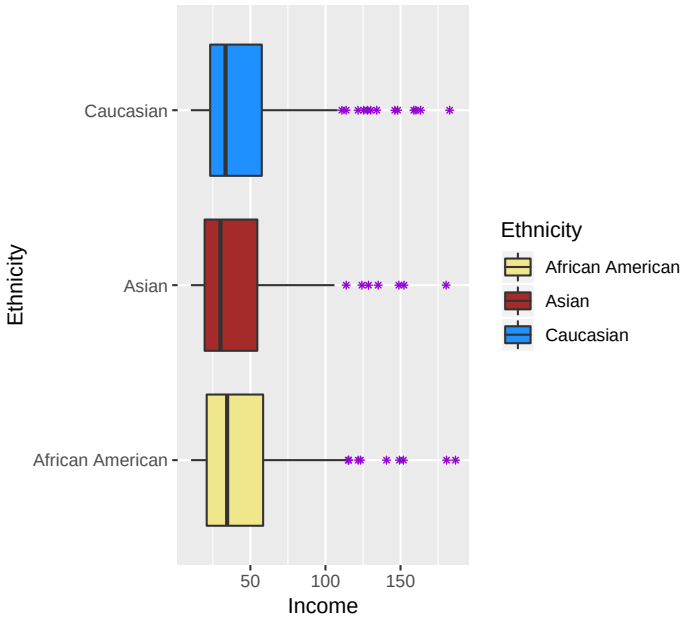

Estimated Probability Density Curve of Age
Probability
0.020
0.015
0.010
0.005
0.000
20
40
60
80
100
Age
Gender
Male
Female
Group Means
Male
Female

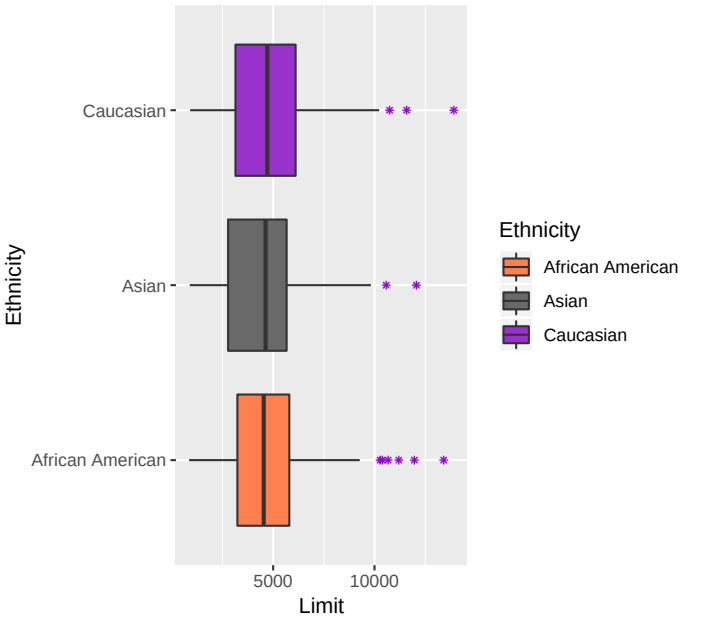

Estimated Probability Density Curve of Age
Probability
0.025
0.020
0.015
0.010
0.005
0.000
20
40
60
80
100
Age
Student
No
Yes
Group Means
No
Yes

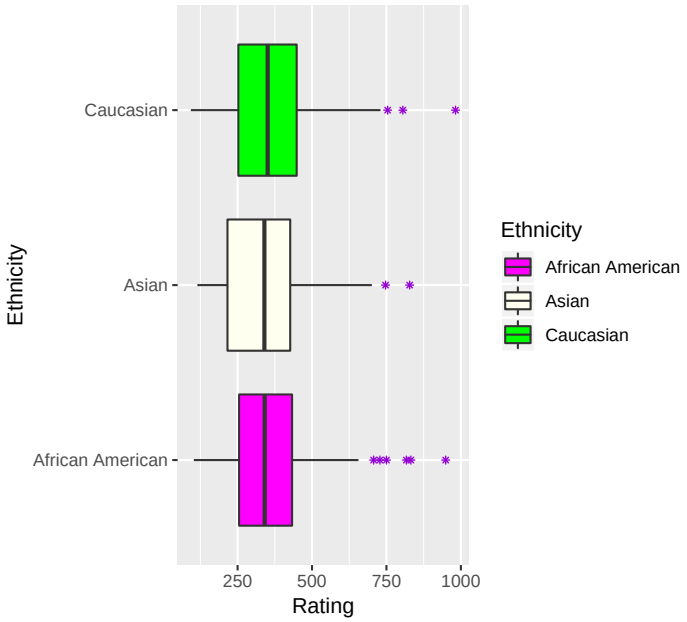

Estimated Probility Curve of Education
Probability
0.125
0.100
0.075
0.050
0.025
0.000
5
10
15
20
Education
Gender
Male
Female
Group Means
Male
Female

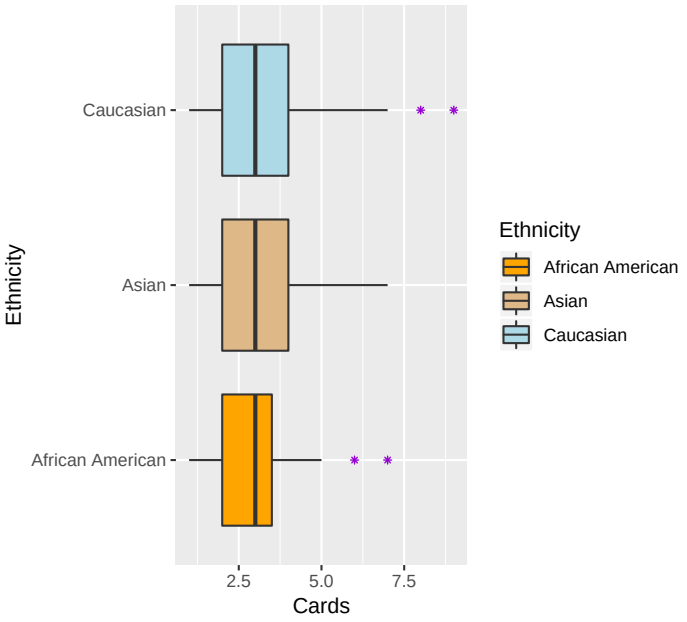

Estimated Probability Density Curve of Education
Probability
0.15
0.10
0.05
0.00
5
10
15
20
Education
Student
No
Yes
Group Means
No
Yes

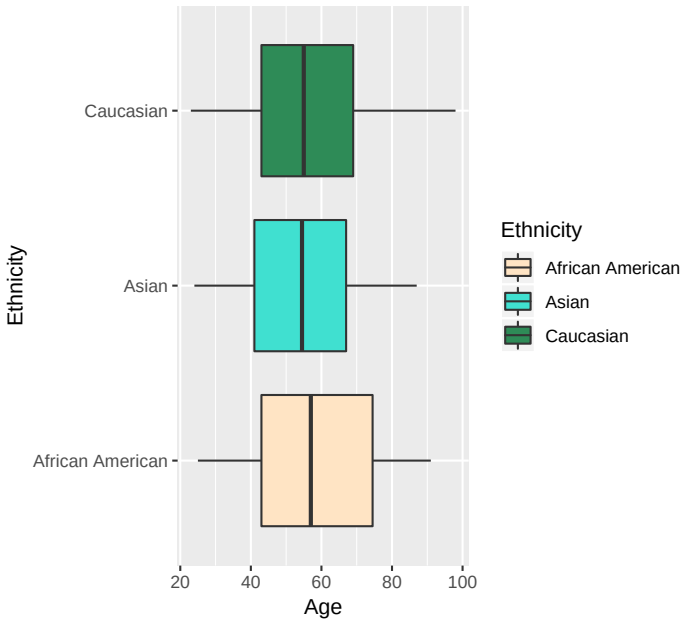

Caucasian
Asian
African American
Ethnicity
10
20
30
40
Balance
Ethnicity
African American
Asian
Caucasian

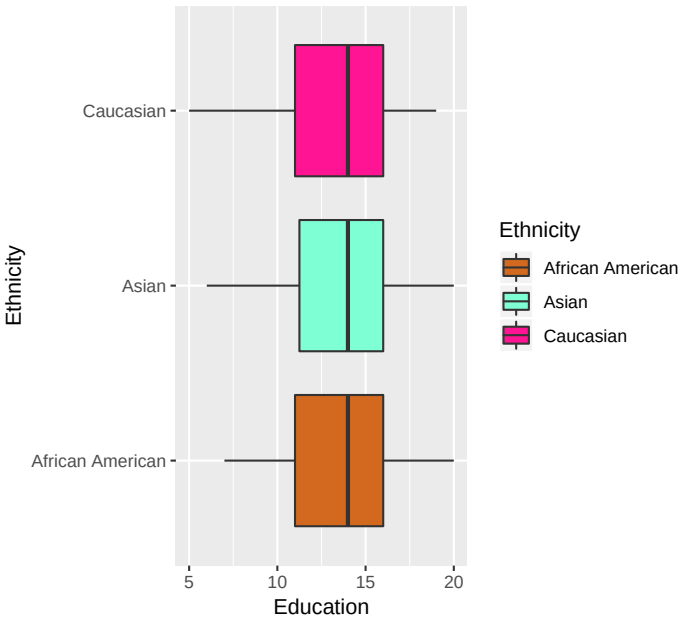

Caucasian
Asian
African American
Ethnicity
50
100
150
Income
Ethnicity
African American
Asian
Caucasian

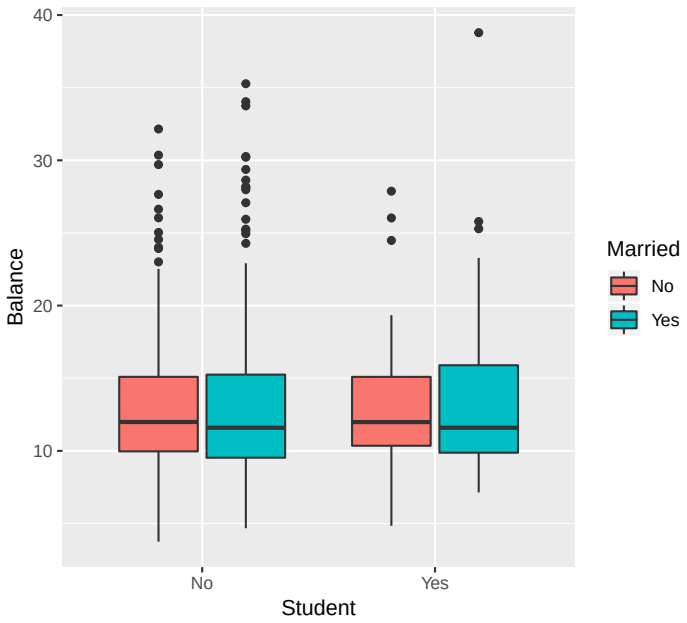

Caucasian
Asian
African American
Ethnicity
5000
10000
Limit
Ethnicity
African American
Asian
Caucasian

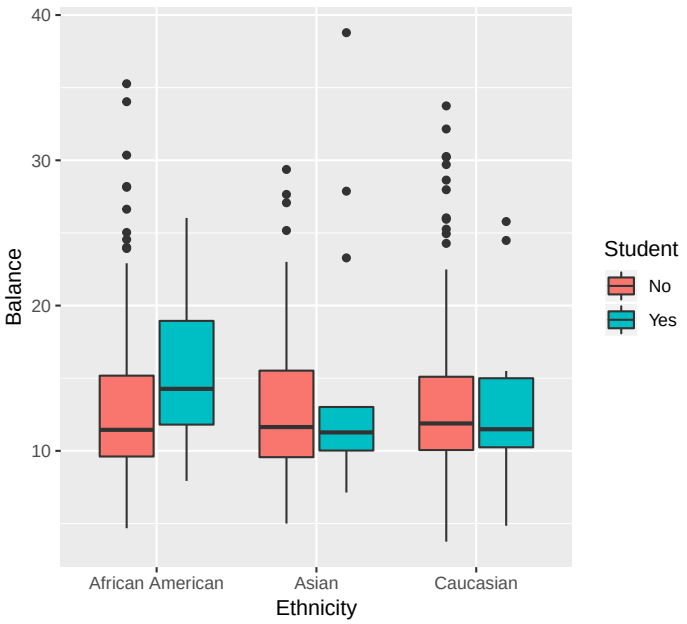

Ethnicity
Caucasian
Asian
African American
250
500
750
1000
Rating
Ethnicity
African American
Asian
Caucasian

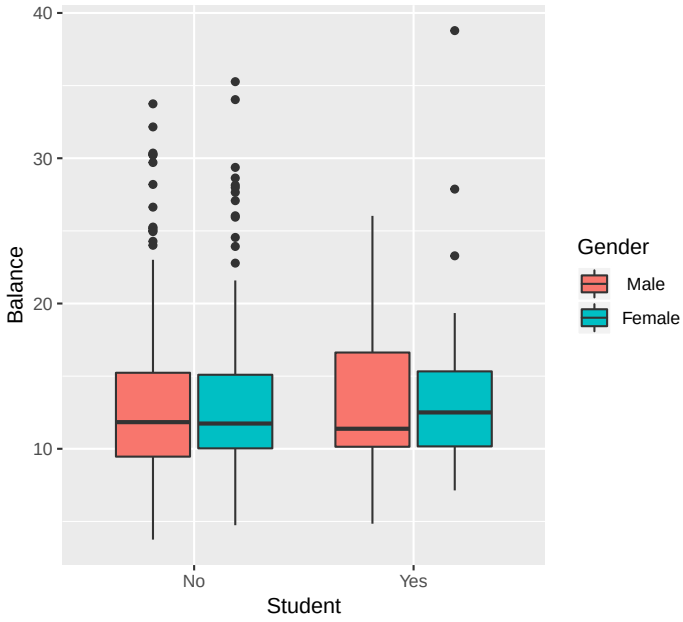

Ethnicity
Caucasian
Asian
African American
2.5
5.0
7.5
Cards
Ethnicity
African American
Asian
Caucasian

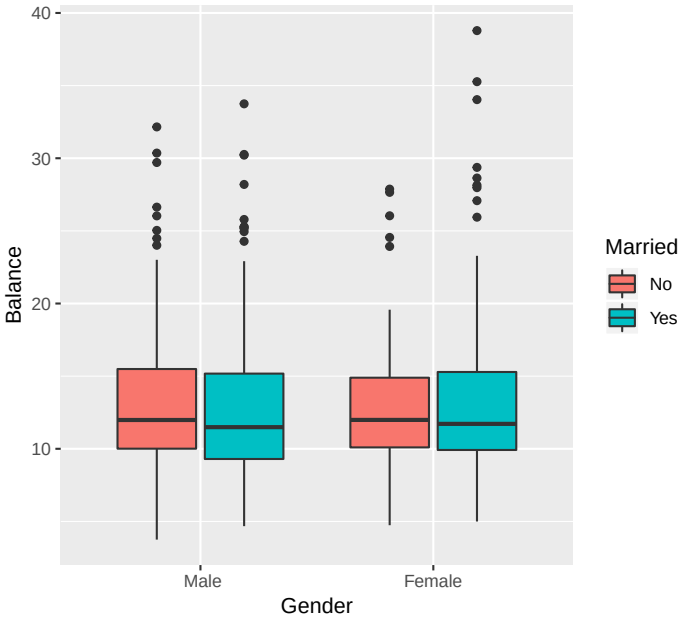

Ethnicity
Caucasian
Asian
African American
20
40
60
80
100
Age
Ethnicity
African American
Asian
Caucasian

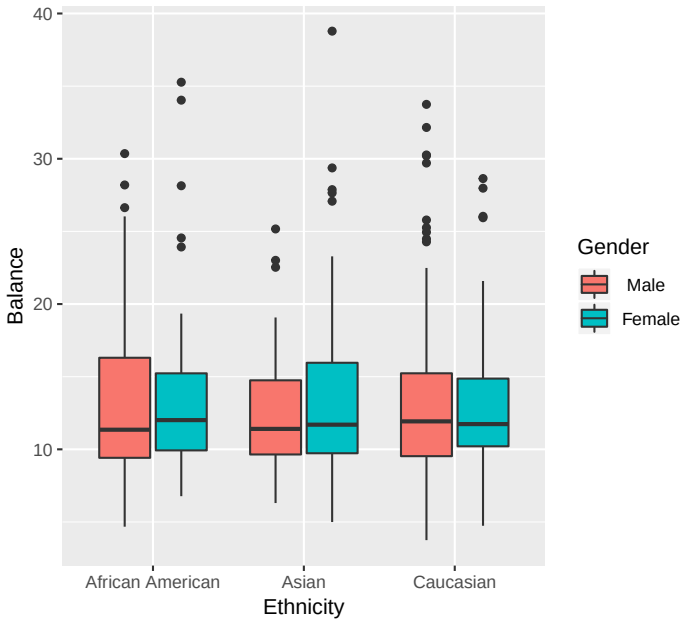

Ethnicity
Caucasian
Asian
African American
5
10
15
20
Education
Ethnicity
African American
Asian
Caucasian

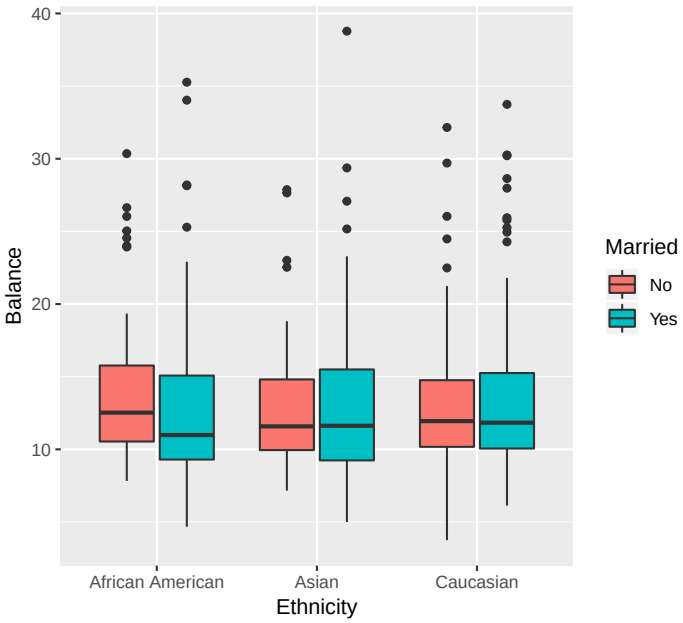

Balance
40
30
20
10
No
Yes
Student
Married
No
Yes


Balance
40
30
20
10
African American
Asian
Caucasian
Ethnicity
Student
No
Yes


40
30
20
10
Balance
No
Yes
Student
Gender
Male
Female


40
30
20
10
Balance
Male
Female
Gender
Married
No
Yes


40
30
20
10
Balance
African American
Asian
Caucasian
Ethnicity
Gender
Male
Female


40
30
20
10
Balance
African American
Asian
Caucasian
Ethnicity
Married
No
Yes

Data Preparation

Data Modelling

Models

Linear Regression

Logistic Regression

Linear Discriminant Analysis

K-Nearest Neighbors

Regression Trees

Bagging

Random Forests

Boosting

Support Vector Machines

Neural Networks

Results

Discussion

Conclusions

Recommendations

Appendix