# Data

*Liang Hao, Andrew Shibata*

*10/24/2016*

## Data

### Description

The *Credit* data set contains information about 400 bank customers. It has eleven columns in total, six of which are quantitative, including:

- *Income*, the customer's stated income,
- *Limit*, the customer's credit limit,
- *Rating*, the customer's credit rating,
- *Cards*, the number cards that the customer has,
- *Age*, the age of the customer,
- *Education*, the year of education that the customer has;

four of which are qualitative, including:

- *Gender*, the gender of the customer, with two levels: *Male* or *Female*
- *Student*, whether the customer is currently a student, with two levels: *Yes* or *No*
- *Married*, the marital status of the customer, with two levels: *Single* or *Married*,
- *Ethnicity*, the ethnicity of the customer, with three levels: *Caucasian*, *African American* and *Asian*;

and a dependent quantitative variable *Balance*, describing the current balance of the customer in his or her bank account.

In this report, we focus on choosing the independent variables that helps to predict the dependent variables, and choosing the best models with the optimal parameters.

### Exploratory Data Analysis

We may first run some preliminary analysis on each of these variables.

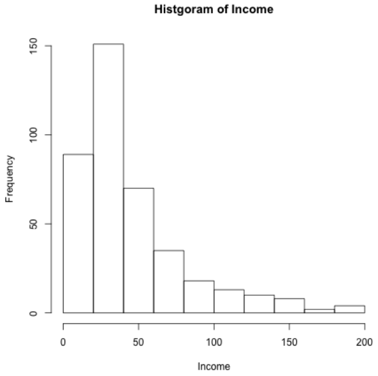
#### Quantitative Variables

| | Min | Max | Range | Median | 25% | 75% | IQR | Mean | Var | Std. Dev. |
|---|---|---|---|---|---|---|---|---|---|---|
| Income | 10.35 | 186.63 | 176.28 | 33.12 | 21.01 | 57.47 | 36.46 | 45.22 | 1242.16 | 35.24 |
| Limit | 855.00 | 13913.00 | 13058.00 | 4622.50 | 3088.00 | 5872.75 | 2784.75 | 4735.60 | 5327781.92 | 2308.20 |
| Rating | 93.00 | 982.00 | 889.00 | 344.00 | 247.25 | 437.25 | 190.00 | 354.94 | 23939.56 | 154.72 |
| Cards | 1.00 | 9.00 | 8.00 | 3.00 | 2.00 | 4.00 | 2.00 | 2.96 | 1.88 | 1.37 |
| Age | 23.00 | 98.00 | 75.00 | 56.00 | 41.75 | 70.00 | 28.25 | 55.67 | 297.56 | 17.25 |
| Education | 5.00 | 20.00 | 15.00 | 14.00 | 11.00 | 16.00 | 5.00 | 13.45 | 9.77 | 3.13 |
| Balance | 0.00 | 1999.00 | 1999.00 | 459.50 | 68.75 | 863.00 | 794.25 | 520.01 | 211378.23 | 459.76 |

Table 1: Summary Statistics for Quantitative Variables

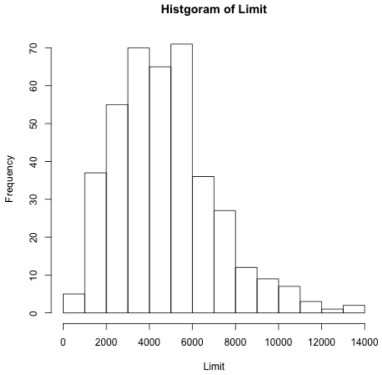
The distribution of the variables might not be necessarily clear. We may also have a look at the histograms:

**Income:**



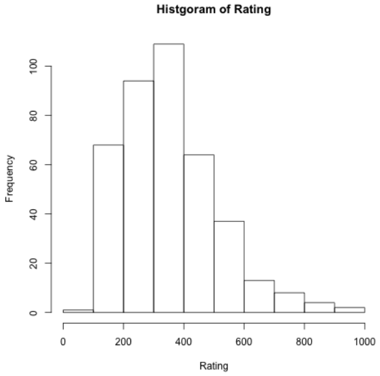
We can see that *Income* has a skewed-to-the-right distribution, with a peak of distribution at around 25 unit, which also explains why its mean of 45.22 is quite larger than its median 33.12.

**Limit:**



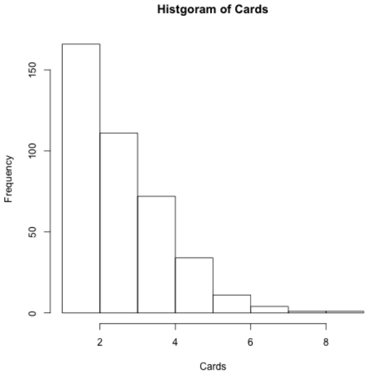
We see that *Rating* basically follows a normal distribution, with the majority of the data in range between 1000 and 8000. We can also see that the median 4622 is relatively close to its mean 4735.

**Rating:**



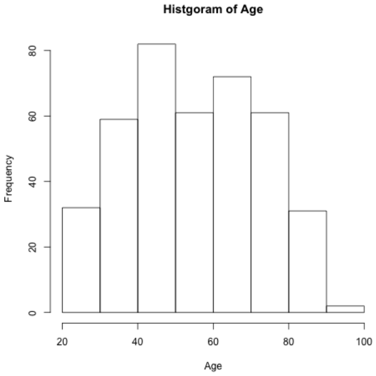
We see that *Rating* basically follows a normal distribution with a long right tail. Most of the data lie in the range between 300 and 600. Also, its median 344 is relatively close to its mean of 355.

**Cards:**



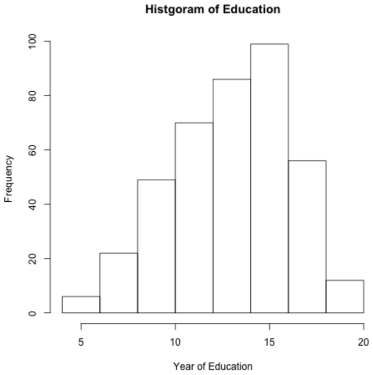
We see that *Cards* follows a skewed to the right distribution, with a peak of distribution between 0 and 2. Because the range of this variable, 8, is relatively small, we may consider scaling for future analysis.

**Age:**



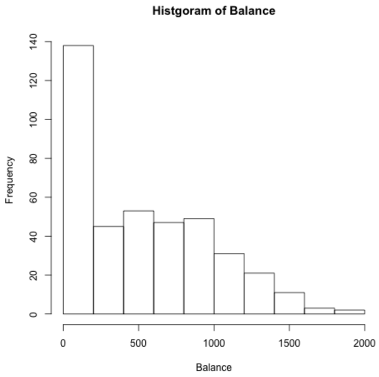
We may see that *Age* has two peaks, at 40-50 and 60-70 years old, which can also be infereced by the fact that both of its median 56 and its mean 55 are not in the peak ranges.

**Education:**



We may see that *Education* has a skewed-to-the-left distribution, with a peak at 15 years, saying that most people have years of education upto undergraduate levels. We may also see this trend from the fact that its median 14 is slightly greater than its mean of 13.5.

**Balance:**



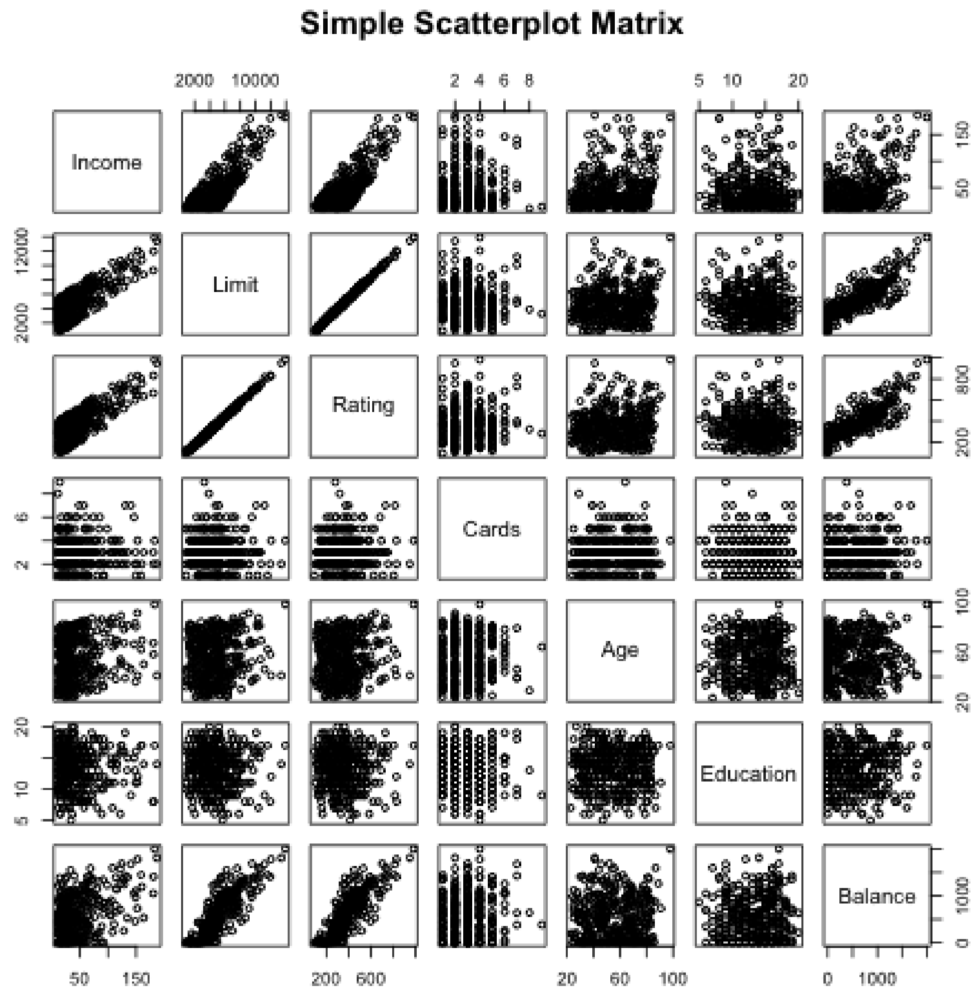
For the dependent variable *Balance*, we may see that the distribution has a long right tail, with a peak between 0 and 250 units. Such a trend can also be inferenced from the fact that its meidan 459.5 is quite smaller than its mean of 520.

Furthermore, we may also have a look at the correlation between the variables:

| | Income | Limit | Rating | Cards | Age | Education | Balance |
|---|---|---|---|---|---|---|---|
| Income | 1.00 | 0.79 | 0.79 | -0.02 | 0.18 | -0.03 | 0.46 |
| Limit | 0.79 | 1.00 | 1.00 | 0.01 | 0.10 | -0.02 | 0.86 |
| Rating | 0.79 | 1.00 | 1.00 | 0.05 | 0.10 | -0.03 | 0.86 |
| Cards | -0.02 | 0.01 | 0.05 | 1.00 | 0.04 | -0.05 | 0.09 |
| Age | 0.18 | 0.10 | 0.10 | 0.04 | 1.00 | 0.00 | 0.00 |
| Education | -0.03 | -0.02 | -0.03 | -0.05 | 0.00 | 1.00 | -0.01 |
| Balance | 0.46 | 0.86 | 0.86 | 0.09 | 0.00 | -0.01 | 1.00 |

Table 2: Correlation matrix for the Quantitative Variables

Additionally the plot of correlations:



**Conclusion:**

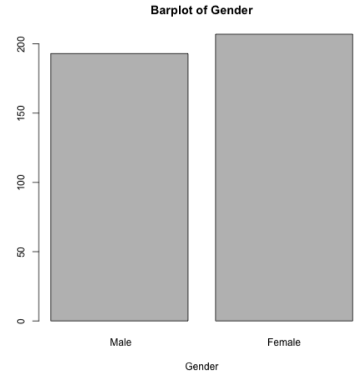
We may see that *Income* has a somewhat strong and positive correlation with *Balance*, and strong correlation with *Limit* and *Rating*; *Limit* has a strong positive correlation with *Balance*, and one-to-one correlation with *Rating*; *Rating* has a strong positive correlation with *Balance*; and *Cards*, *Age* and *Education* have small or even no correlation with *Balance.*

Therefore, for future analysis, we may consider using *Income* and *Limit* as two major predictors, for other independent variables are either dependent on these two, or unlikely to have much predicting power for the dependent variable *Balance.*

Also, we may want to normalize and standardize the data set for the range and scale of the raw data may have an impact on how our model would perform.

**Qualitative Variables:**

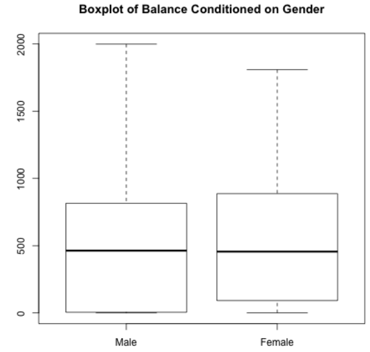
**Gender:**



| | Gender | Frequency | RelativeFrequency |
|---|---|---|---|
| 1 | Male | 193 | 0.48 |
| 2 | Female | 207 | 0.52 |

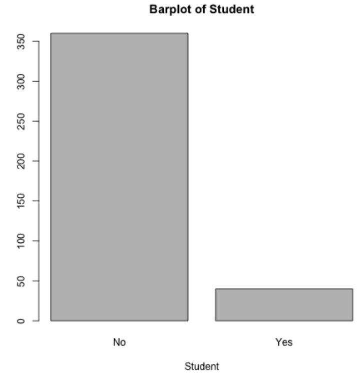
Table 3: Frequency Table for Gender

We may see that the two levels *Male* and *Female* basically have the same amount of input in this data set. Additionally, we may have a look at the boxplot os *Balance* conditioned on the two levels of *Gender*:



We may see that the two boxes seem quite similar to each other, which may indicate that *Gender* might not have a strong correlation with *Balance*.

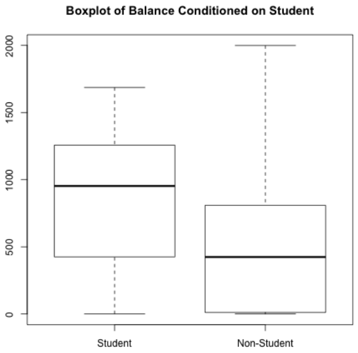
**Student:**



|   | Student | Frequency | RelativeFrequency |
|---|---|---:|---:|
| 1 | No | 360 | 0.90 |
| 2 | Yes | 40 | 0.10 |

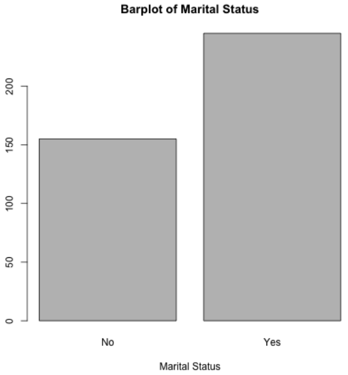
Table 4: Frequency Table for Student

We may see that majority input of the data set are not students. Additionally, we may have a look at the boxplot os *Balance* conditioned on the two levels of *Student*:



We may see that the *Balance* of *Students* look much higher than those of *Non-Students*, which may indicate that *Student* may be a good predictor for the *Balance*. We may keep this in mind for future analysis.

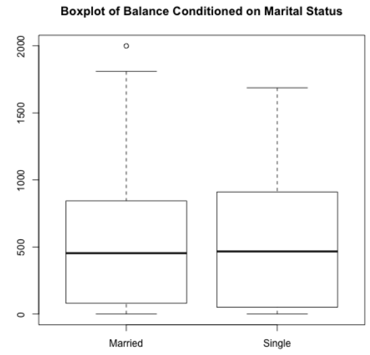
**Married:**



|   | Married | Frequency | RelativeFrequency |
|---|---|---:|---:|
| 1 | No | 155 | 0.39 |
| 2 | Yes | 245 | 0.61 |

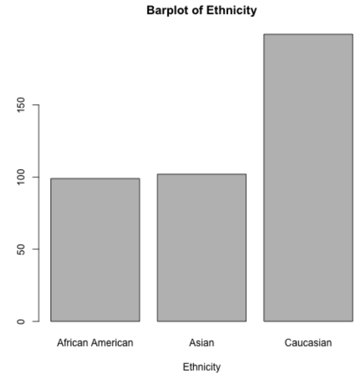
Table 5: Frequency Table for Marital Status

We may see that more married customers are in the data set than single ones. Additionally, we may have a look at the boxplot os *Balance* conditioned on the two levels of *Married*:



We can see that the two boxplots have basically the same distribution, with may indicate that the *Balance* might not depend on *Married*.

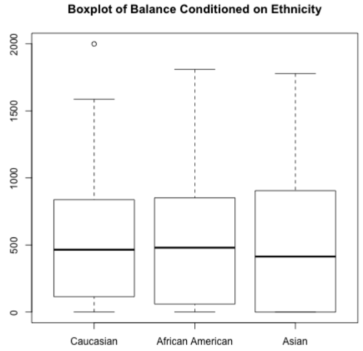
**Ethnicity:**



| | Ethnicity | Frequency | RelativeFrequency |
|---|---|---|---|
| 1 | African American | 99 | 0.25 |
| 2 | Asian | 102 | 0.26 |
| 3 | Caucasian | 199 | 0.50 |

Table 6: Frequency Table for Ethnicity

We may see that *Caucasian* takes about half in the ethnicity of the customers, and *African American* and *Asian* take the other half combined. Additionally, we may have a look at the boxplot os *Balance* conditioned on the three levels of *Ethnicity*:



We may see that the *Balance* distribution of the three ethnicities are basically the same, meaning that *Balance* might not depend on the *Ethnicity* of the customers.

**Conclusion:**

With the analysis above, we may consider using *Student* as a predictor, for *Balance* are quite different for the two levels. For other variables, we may want to figure out with future analysis.