WOE/LOGISTIC REGRESSION FINAL PROJECT

By:

Andrew Lundborg

STATS 451/551 Predictive Analytics

Spring 2015

**DATA PREPARATION**

For my data I started with all available observations, including those who did not meet the credit policy listed at the bottom of the given excel file.

My first objective was to create a binary response variable from the given loan\_status variable. I wanted those who are currently paying off their loan and those who have paid off their loan to be in the “good” group (represented by a 0) and put everyone else in the “bad” group (represented by a 1).

Of the available variables, I chose ten of them for consideration, to keep the model relatively simple:

1. Loan amount (loan\_amnt) was an easy choice as a useful variable. I expected that those who have larger loans are less likely to be able to pay them off.
2. Lending Club applies a Grade (grade) to each loan, which represents the calculated likelihood that the person will pay of their loan. In other words, a grade of ‘A’ means that the loan will most likely be paid off.
3. The provided home ownership status by the borrower (home\_ownership) may also help in the model, as there may be some difference between those who were currently renting their property and those who were paying off mortgage loans, etc.
4. Annual Income (annual\_inc) was another easy choice to consider. Those who obtain more cash are easily more able to pay off their loans.
5. Debt-to-Income (dti) is the ratio of the borrower’s monthly debt payments divided by their reported monthly income. In other words, it is the percentage of the borrower’s income that goes towards paying off their debts. Therefore, one would expect that those with a high DtI have more debt for their income.
6. A FICO score (fico\_range\_low) is another metric that represents a person’s credit score. Higher FICO scores means that the person has a good credit record and is more likely to pay off their debt.
7. A credit inquiry (inq\_last\_6mths) is a request by a business to check an individual’s credit. While not always negative, credit checks can occur when an individual applies for things such as loans or credit cards. Therefore, if a person has a lot of credit checks on their record, then it is possible that their credit may not be so great.
8. A delinquency (mths\_since\_last\_delinq) is essentially a mark on the person’s record where they failed to make a payment within the allotted time. Therefore, a person who has many delinquencies is more likely to not pay off a loan.
9. The amount of open credit lines (open\_acc), on a borrower’s account may imply that the person is a more risky investment, but it could potentially mean that the person is good at managing their credit and can afford to have those lines open. Therefore, I chose this variable to see if/how it may explain loan status.
10. While a borrower is trusted to report their own income, there is always the possibility that it may be false. Therefore, Lending Club looks into the income sources and amounts to verify them (is\_inc\_v). I chose this variable, as it may be possible that those borrowers who have unverified incomes may also be untrustworthy and not able to pay off their loans.

As for the variables I did not choose, some of them I did not choose for specific reasons. For instance, the variable addr\_state is the state provided by the borrower, which has 50 levels. It would be possible to group them up, such as by time zones, but I decided against it. Other variables would require text analysis, such as the description and title of the loan, which would take some time. Some of the variables had the same value for each person, such as collections\_12\_mths\_ex\_med, and policy code.

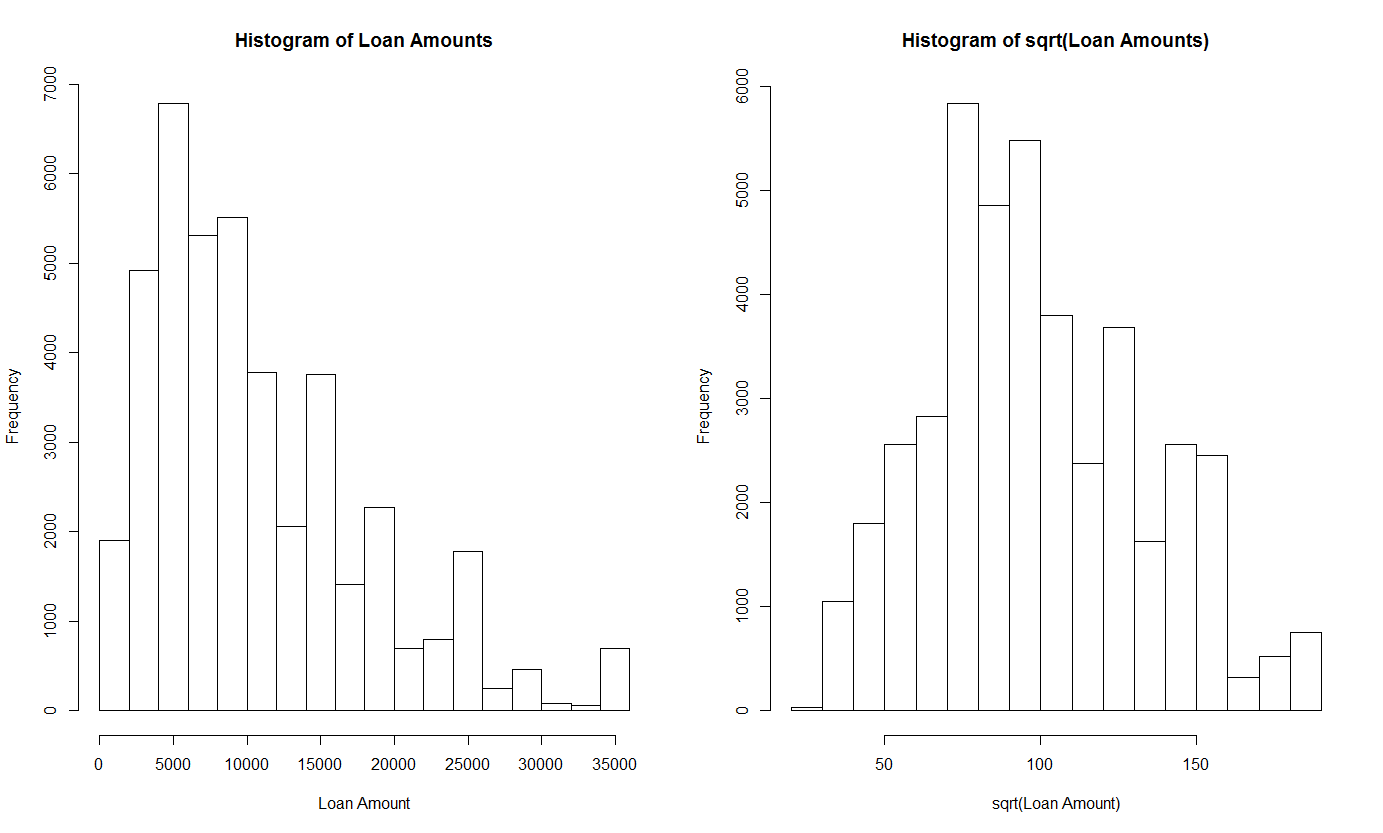
**VARIABLE: LOAN AMOUNT**

As stated prior, one would think that the amount of money being requested would definitely help determine whether or not the borrower would pay off the loan. First, I decided to take a look at the summary statistics for loan amounts, given below in Table 1.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Minimum** | **1st Quantile** | **Median** | **Mean** | **3rd Quantile** | **Maximum** |
| 500 | 5200 | 9700 | 11,090 | 15,000 | 35,000 |

**Table 1.** *Summary statistics for loan amounts.*

After viewing the histogram of loan amounts (Figure 1: Left), I determined that a square root transform moves the skewed data into a more normal distribution (Figure 1: Right). Therefore, I will use the new transformed variable in my analysis.



**Figure 1.** *Left: Distribution of loan amounts; Right: Distribution of loan amounts (square root transformed)*

Now that the new variable has a more normal distribution, I will leave it as is for the future binning step.

**VARIABLE: GRADE**

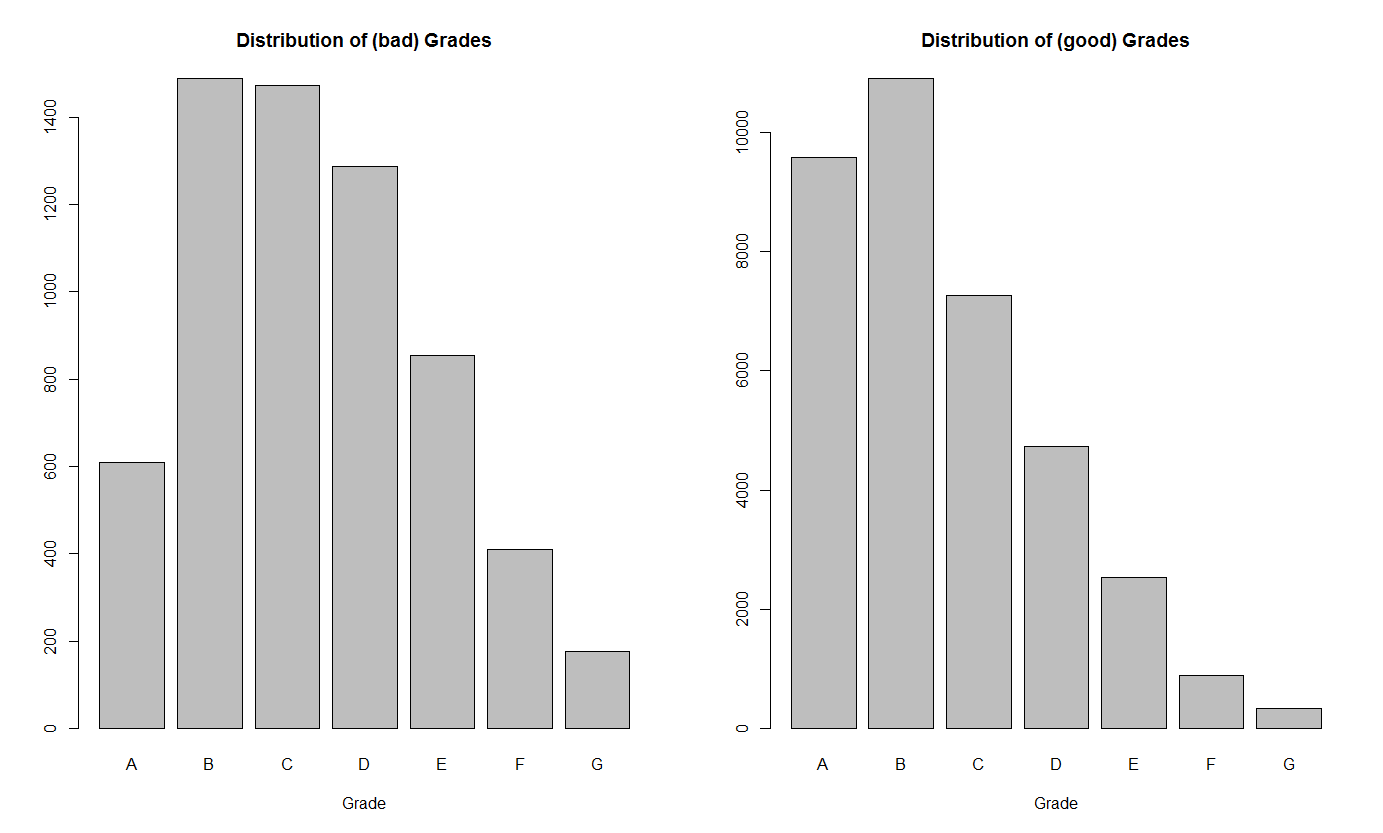
The Grade variable was definitely my first variable of interest, as the variable, taking on values of A through G, is a grade already applied to the customer by Lending Club. As stated, this grade is the customers predetermined level of risk of defaulting on their loan. For instance, a grade of A means that the person has good credit history and is most likely to successfully pay off their loan.

First I wanted to look at the distribution of Grades among all borrowers, which is shown in Table 2.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Grade:** | A | B | C | D | E | F | G |
| **Count:** | 10,183 | 12,389 | 8,740 | 6016 | 3394 | 1301 | 512 |

**Table 2.** *Grade level counts*

So it appears that most borrowers actually have high grades, which is to be expected, as having a lot of borrowers with low grades would be bad for business. Next I decided to look at the distribution of grades by loan status, shown in Figure 2.



**Figure 2.** *Left: Distribution of Grades for people who didn’t pay off their loan; Right: Distribution of Grades for people who are good at paying off their loan.*

These distributions do reflect what we would expect. Those who have been good with their credit have a higher percentage of good grades. Those who have been bad with their credit have a higher percentage of worse grades. Note that since the amount of people who have been good with their loan is significantly higher than those who have been bad with their loan, there are actually higher counts of “good” people with bad grades. Therefore, we have to look at the percentages, as mentioned.

**VARIABLE: HOME OWNERSHIP**

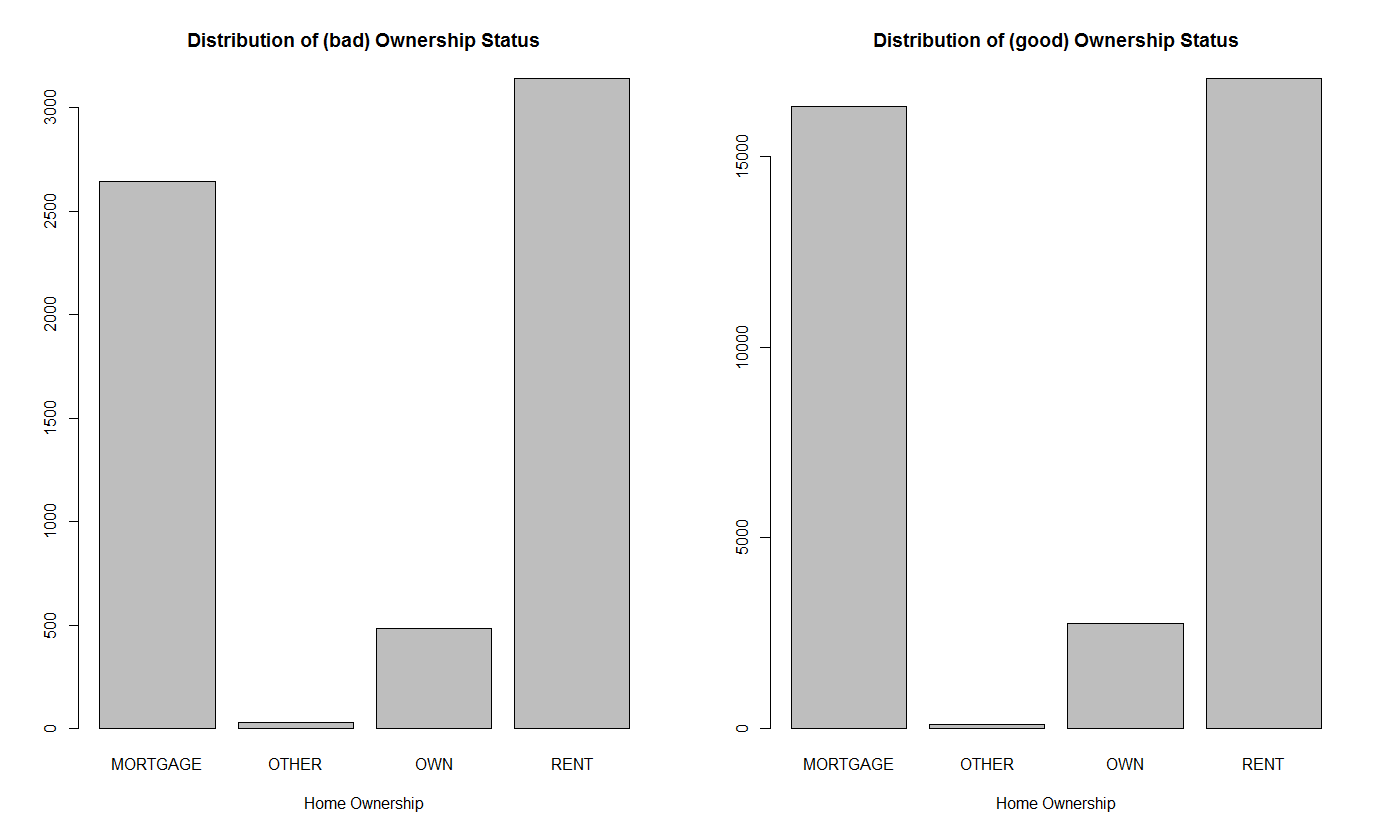
This variable was selected mostly because it is a nice factor variable to work with and investigate. I do not know how home ownership reflects on credit scores, so it was interesting to see whether or not it was significant. First I wanted to look at the distribution of the factors, which is shown in Table 3.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Ownership:** | Mortgage | None | Other | Own | Rent |
| **Counts:** | 18,959 | 8 | 136 | 3251 | 20,181 |

**Table 3.** *Home Ownership counts*

As the level Other is pretty ambiguous, I felt it was okay to place those eight unlisted observations in the Other group. I could be wrong about this decision, but even if so, that is eight out of over forty thousand observations, so the impact would hopefully not be too great.

Next I decided to see if there was any noticeable difference between the home ownership distributions across loan status. This is given in Figure 3 below.



**Figure 3.** *Left: Distribution of Home Ownership for people who didn’t pay off their loan; Right: Distribution of Home Ownership for people who are good at paying off their loan.*

Interestingly, the distribution almost looks exactly the same for both response types, this means that this variable is most likely not going to be helpful, but I will confirm later on whether this is true.

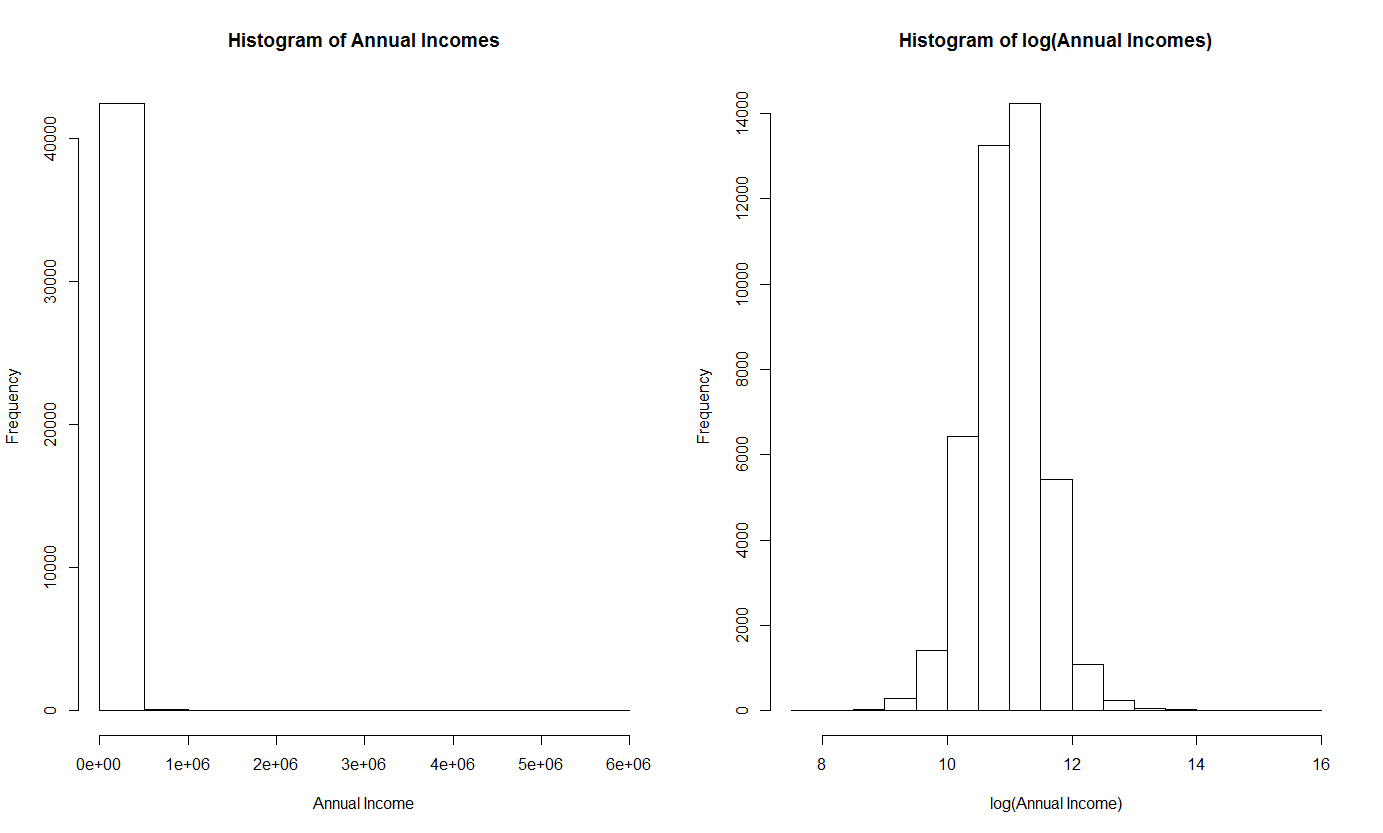
**VARIABLE: ANNUAL INCOME**

As stated previously, this variable seemed like a necessity. Higher incomes should imply higher probabilities of being able to pay off loans. As other variables, I started with the summary statistics, given in Table 4.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Minimum** | **1st Quantile** | **Median** | **Mean** | **3rd Quantile** | **Maximum** | **Missing** |
| 1896 | 40,000 | 59,000 | 69,140 | 82,500 | 6,000,000 | 4 |

**Table 4.** *Summary statistics for annual income.*

First, I opted to simply remove the four missing observations, as they should not affect my analyses too much. Then, after viewing the histogram of annual incomes (Figure 4: Left), I determined that a log transform moves the skewed data into a more normal distribution (Figure 4: Right).This fixes the issue of the large outliers, so this new log variable is the one I will use in my analysis.



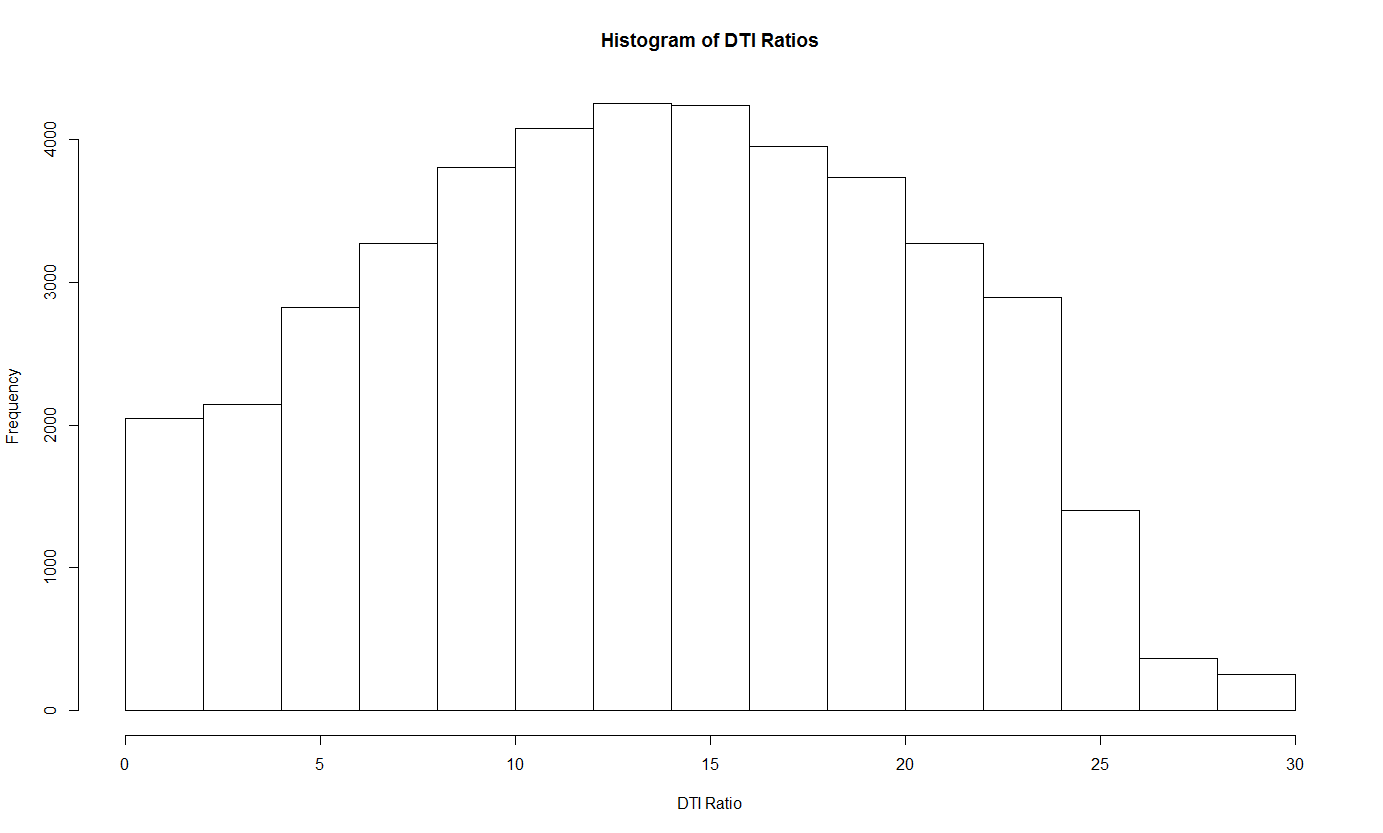
**Figure 4.** *Left: Distribution of annual incomes; Right: Distribution of annual incomes (log transformed)*

**VARIABLE: DEBT-TO-INCOME RATIO**

A borrower’s

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Minimum** | **1st Quantile** | **Median** | **Mean** | **3rd Quantile** | **Maximum** |
| 0.00 | 8.20 | 13.47 | 13.37 | 18.68 | 29.99 |

**Table 5.** *Summary statistics for DTI ratios.*

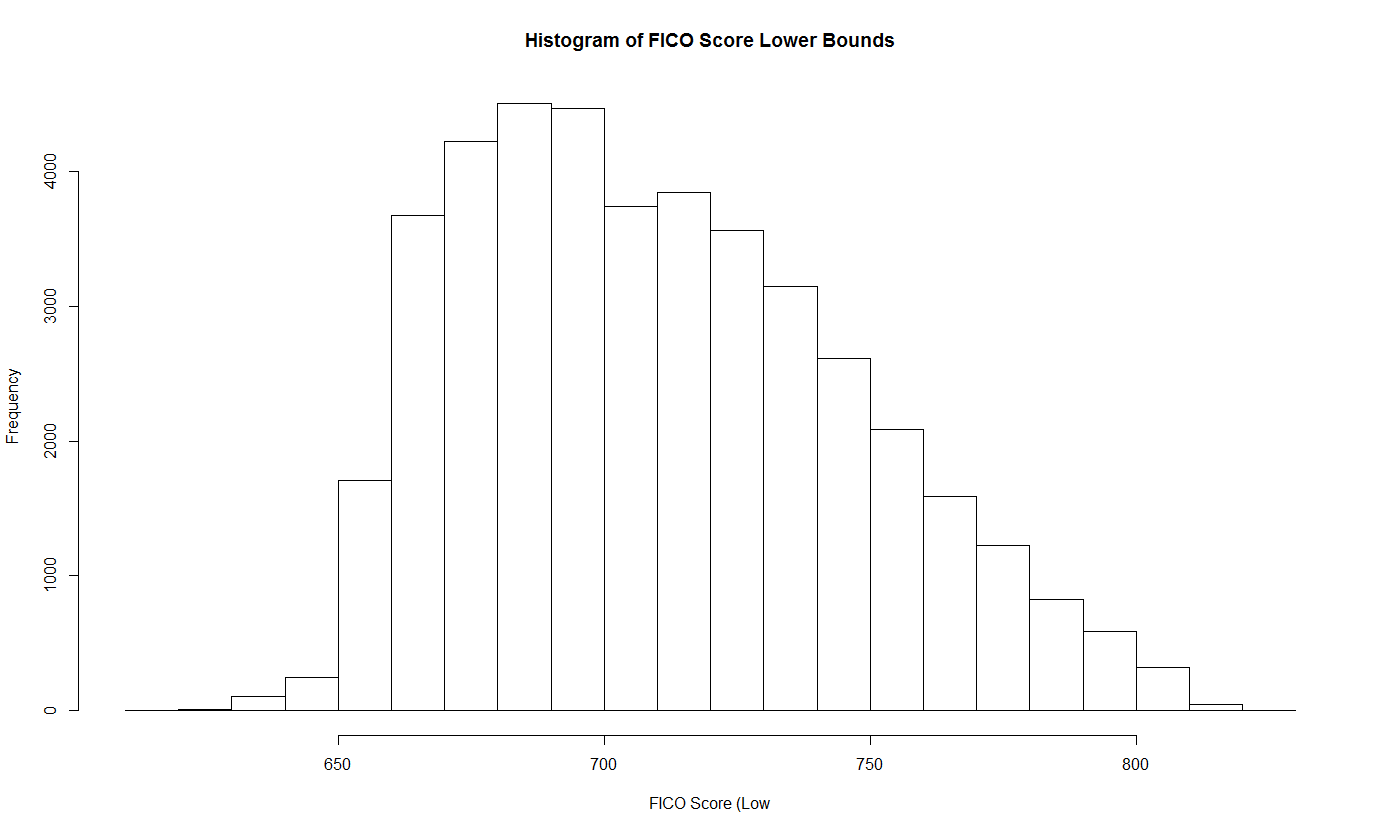


**Figure 5.** *Distribution of DTI ratios*

**VARIABLE: FICO RANGE LOW**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Minimum** | **1st Quantile** | **Median** | **Mean** | **3rd Quantile** | **Maximum** |
| 610 | 685 | 710 | 713 | 740 | 825 |

**Table 6.** *Summary statistics for lower bound of FICO Scores.*

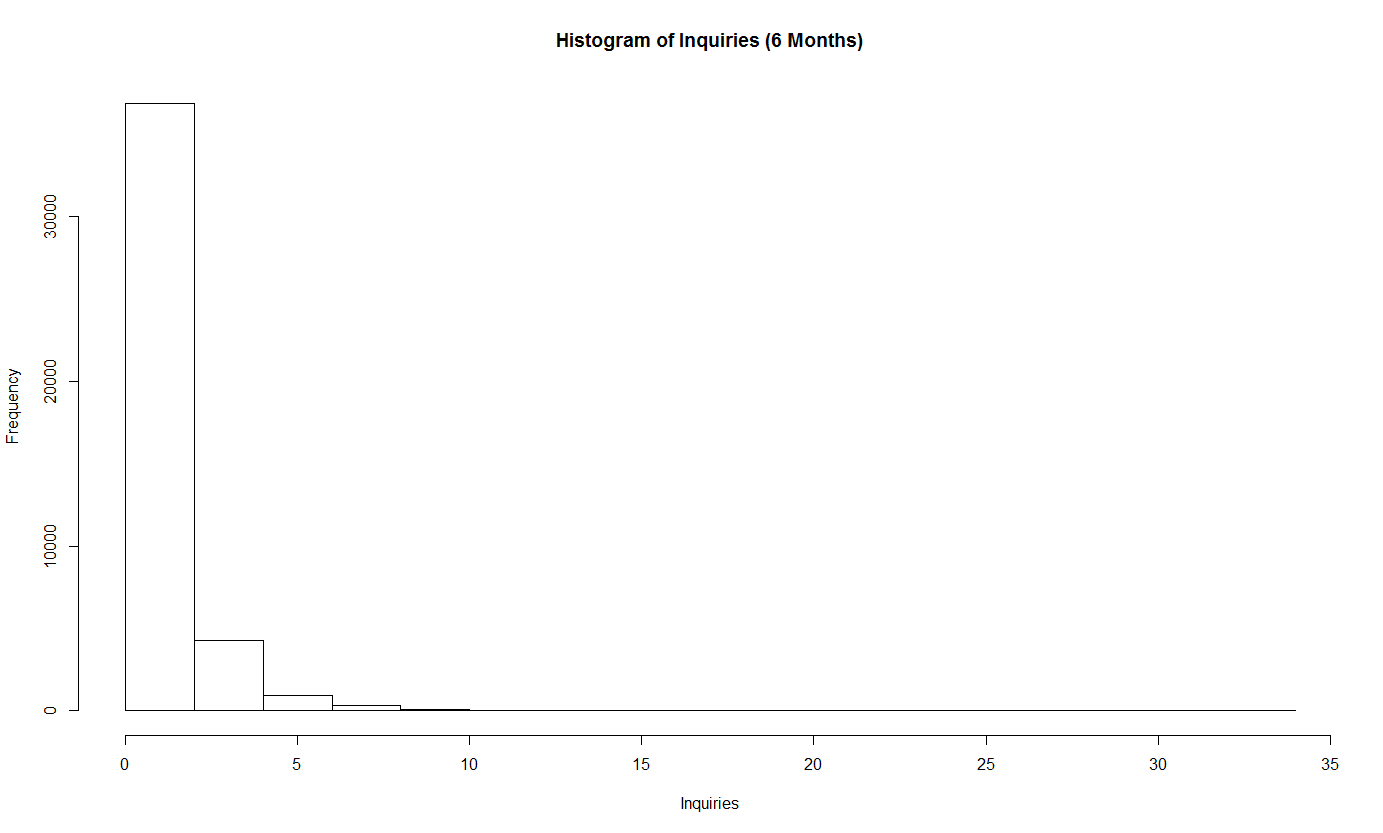


**Figure 6.** *Distribution of FICO score (lower bounds)*

**VARIABLE: INQUIRIES IN LAST 6 MONTHS**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Minimum** | **1st Quantile** | **Median** | **Mean** | **3rd Quantile** | **Maximum** | **Missing** |
| 0.000 | 0.000 | 1.000 | 1.081 | 2.000 | 33.000 | 25 |

**Table 7.** *Summary statistics for inquiries in the last 6 months.*

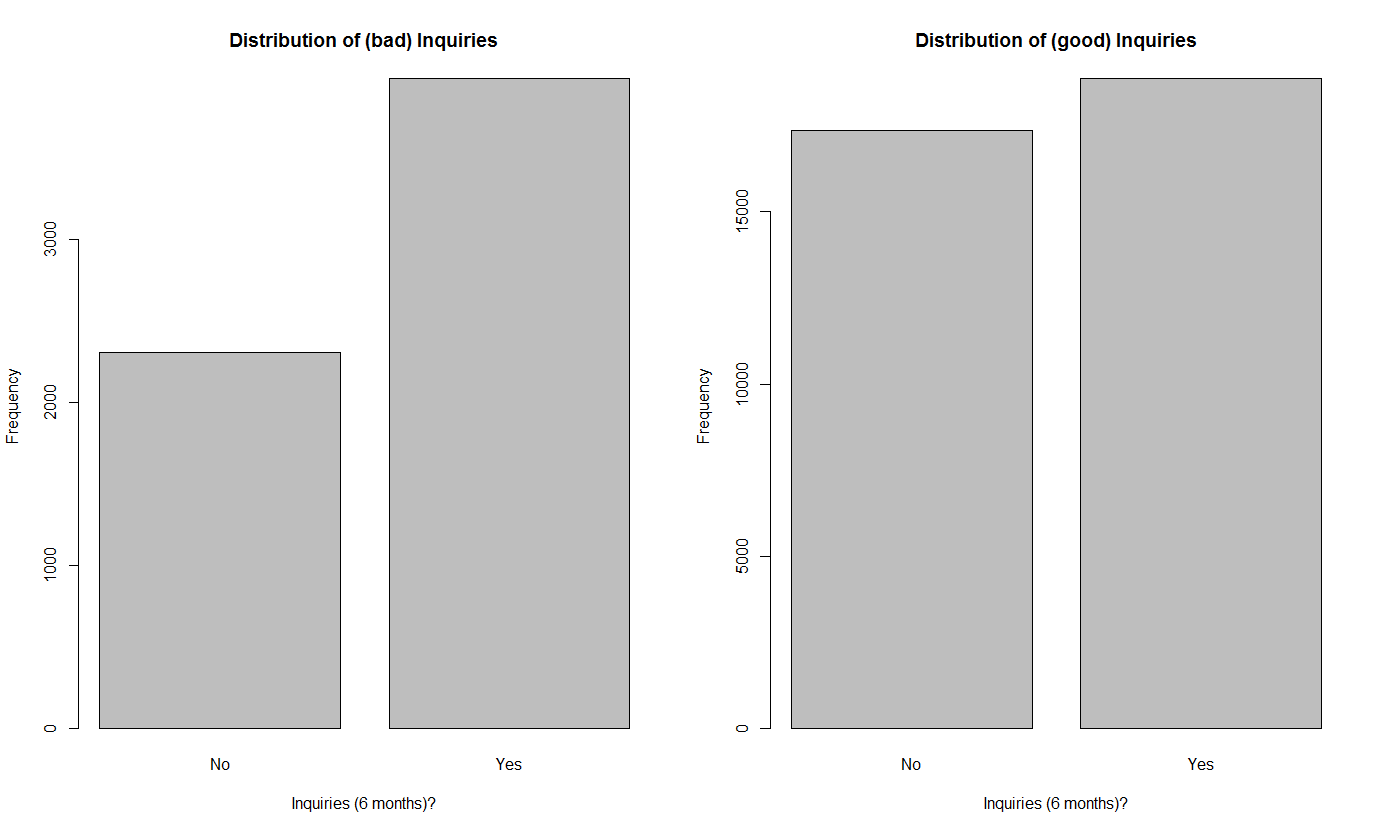


**Figure 7.** *Distribution of Inquiries in the last six months*

**TRANSFORMED VARIABLE: INQUIRIES (FACTOR)**

|  |  |  |
| --- | --- | --- |
| **Inquiries?:** | No | Yes |
| **Counts:** | 19,657 | 22,849 |

**Table 8.** *Counts for whether or not the borrower has had any inquiries in the last 6 months*

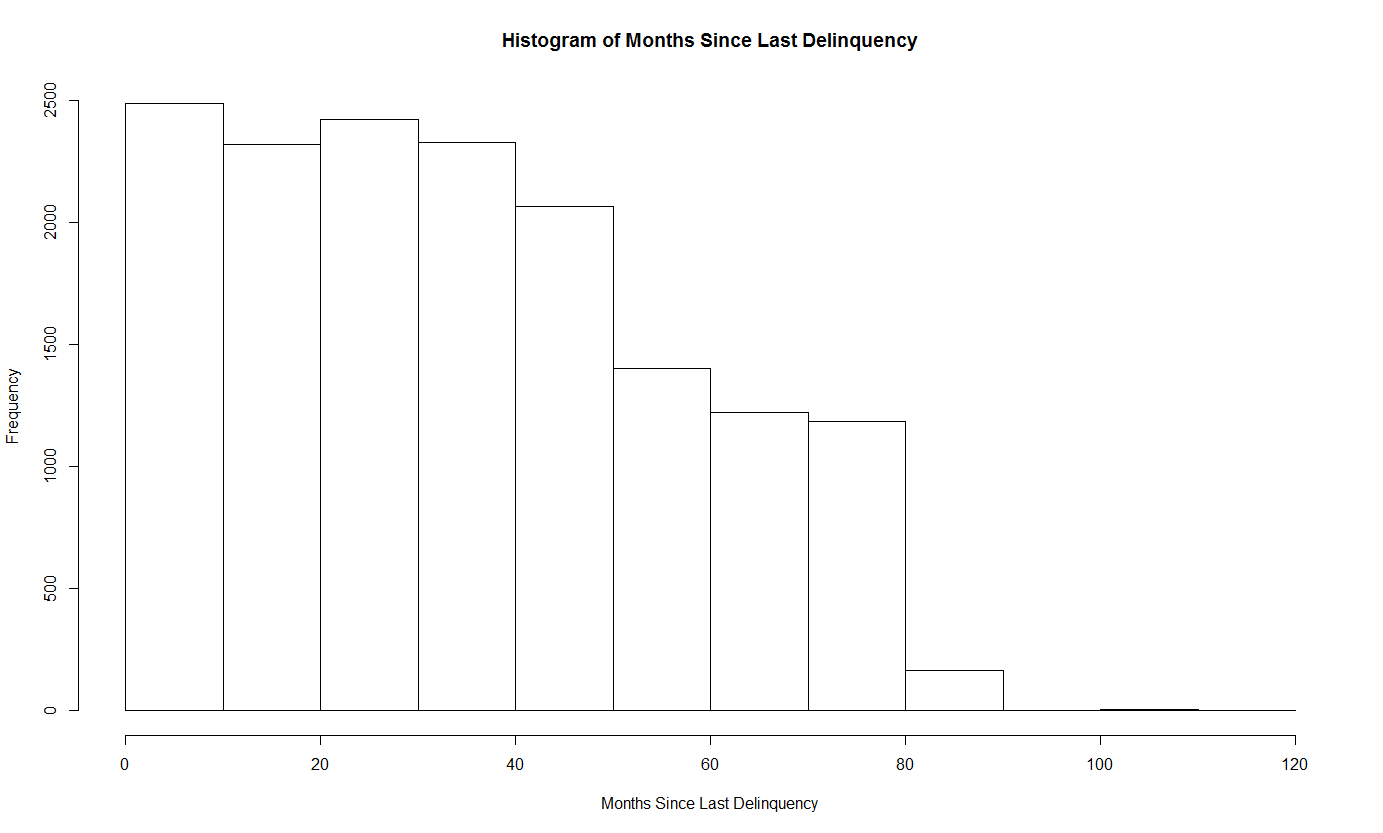


**Figure 8.** *Left: Distribution of Inquiries for people who didn’t pay off their loan; Right: Distribution of Inquiries for people who are good at paying off their loan.*

**VARIABLE: MONTHS SINCE LAST DELINQUENCY**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Minimum** | **1st Quantile** | **Median** | **Mean** | **3rd Quantile** | **Maximum** | **Missing** |
| 0.00 | 17.00 | 33.00 | 35.02 | 51.00 | 120.00 | 26,897 |

**Table 9.** *Summary statistics for months since last delinquency*

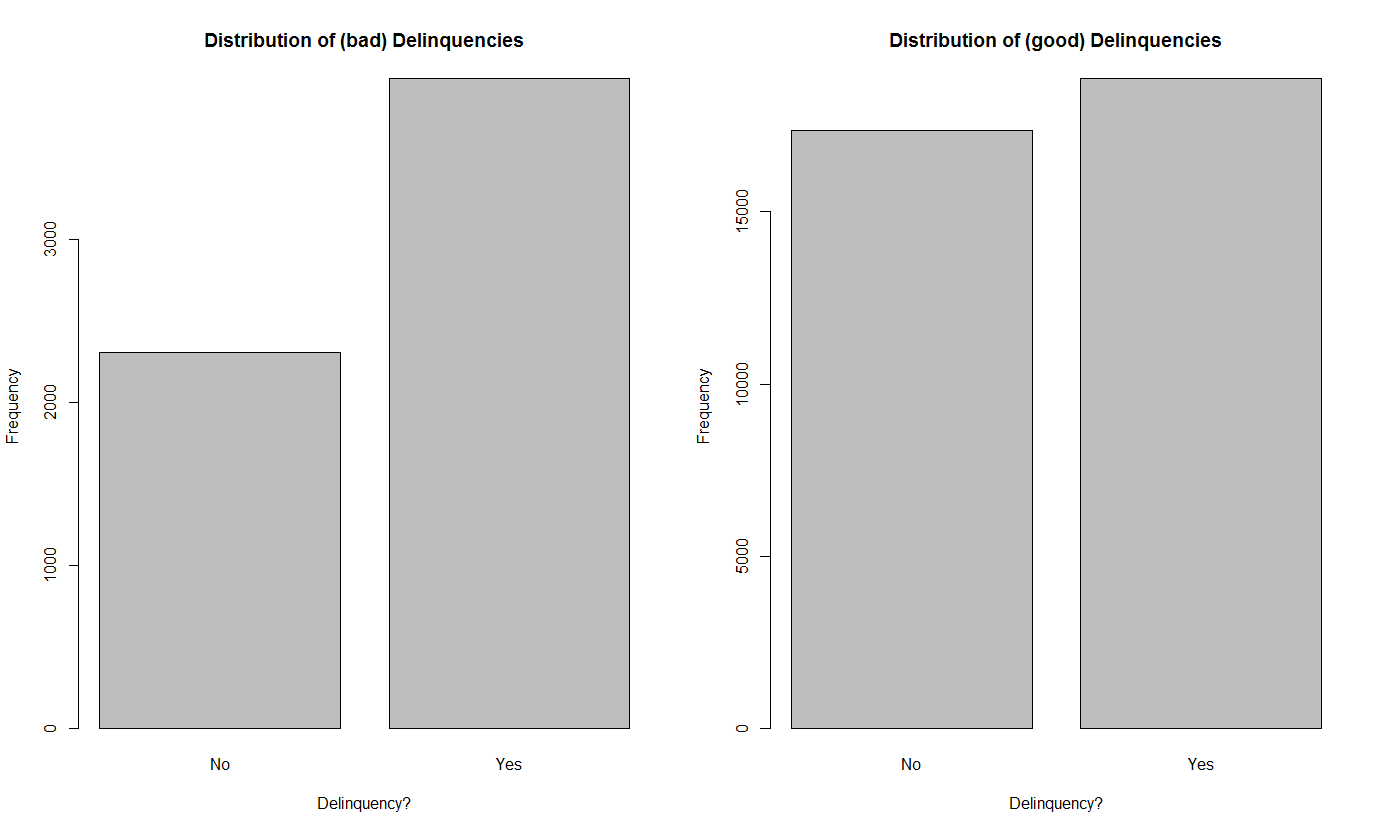


**Figure 9.** *Distribution of Months Since Last Delinquency*

**TRANSFORMED VARIABLE: DELINQUENCIES (FACTOR)**

|  |  |  |
| --- | --- | --- |
| **Delinquency?** | No | Yes |
| **Counts:** | 26,897 | 15,609 |

**Table 10.** *Counts for whether or not a borrower had a delinquency*

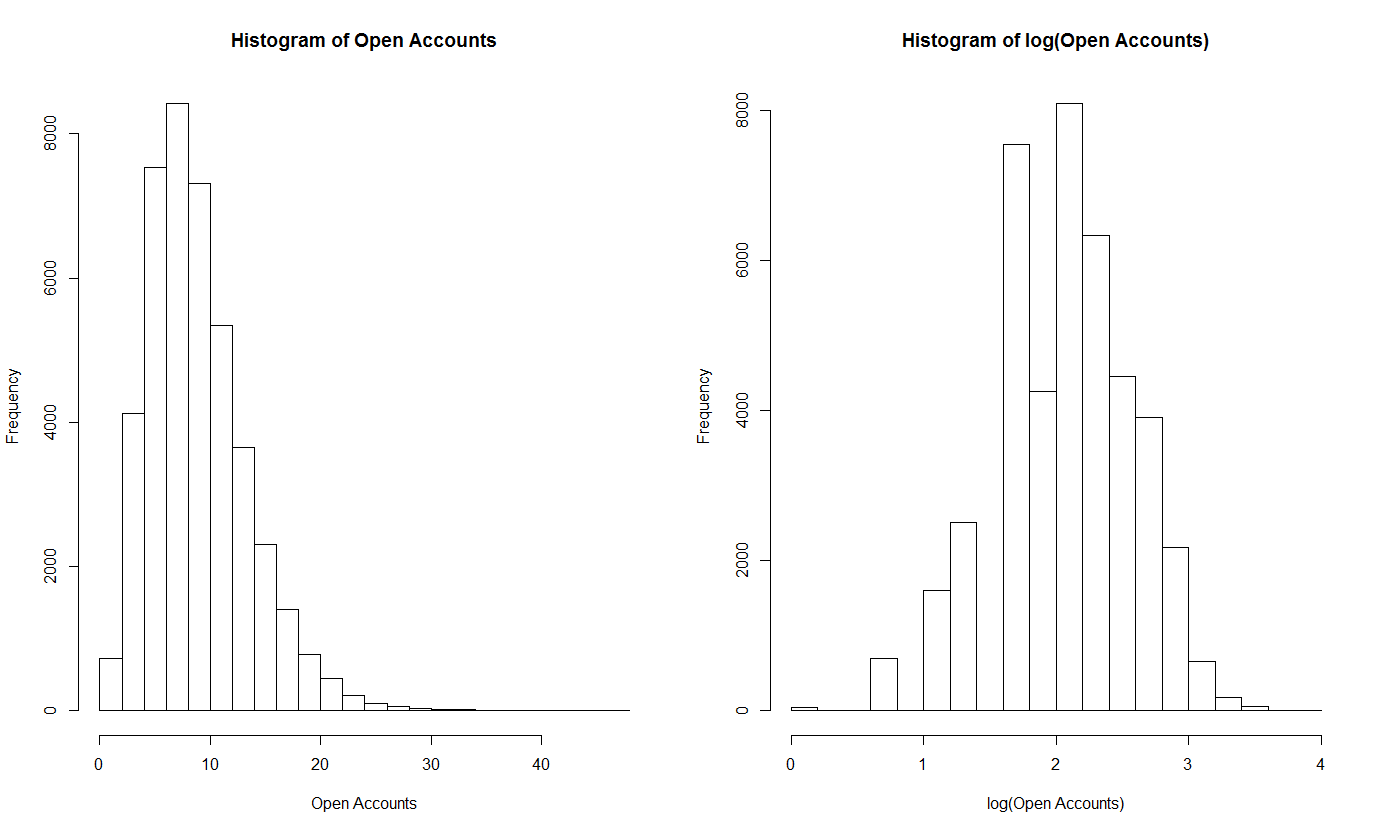


**Figure 10.** *Left: Distribution of Delinquencies for people who didn’t pay off their loan; Right: Distribution of Delinquencies for people who are good at paying off their loan.*

**VARIABLE: OPEN ACCOUNTS**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Minimum** | **1st Quantile** | **Median** | **Mean** | **3rd Quantile** | **Maximum** |
| 1.000 | 6.000 | 9.000 | 9.344 | 12.000 | 47.000 |

**Table 11.** *Summary statistics for Open Accounts*

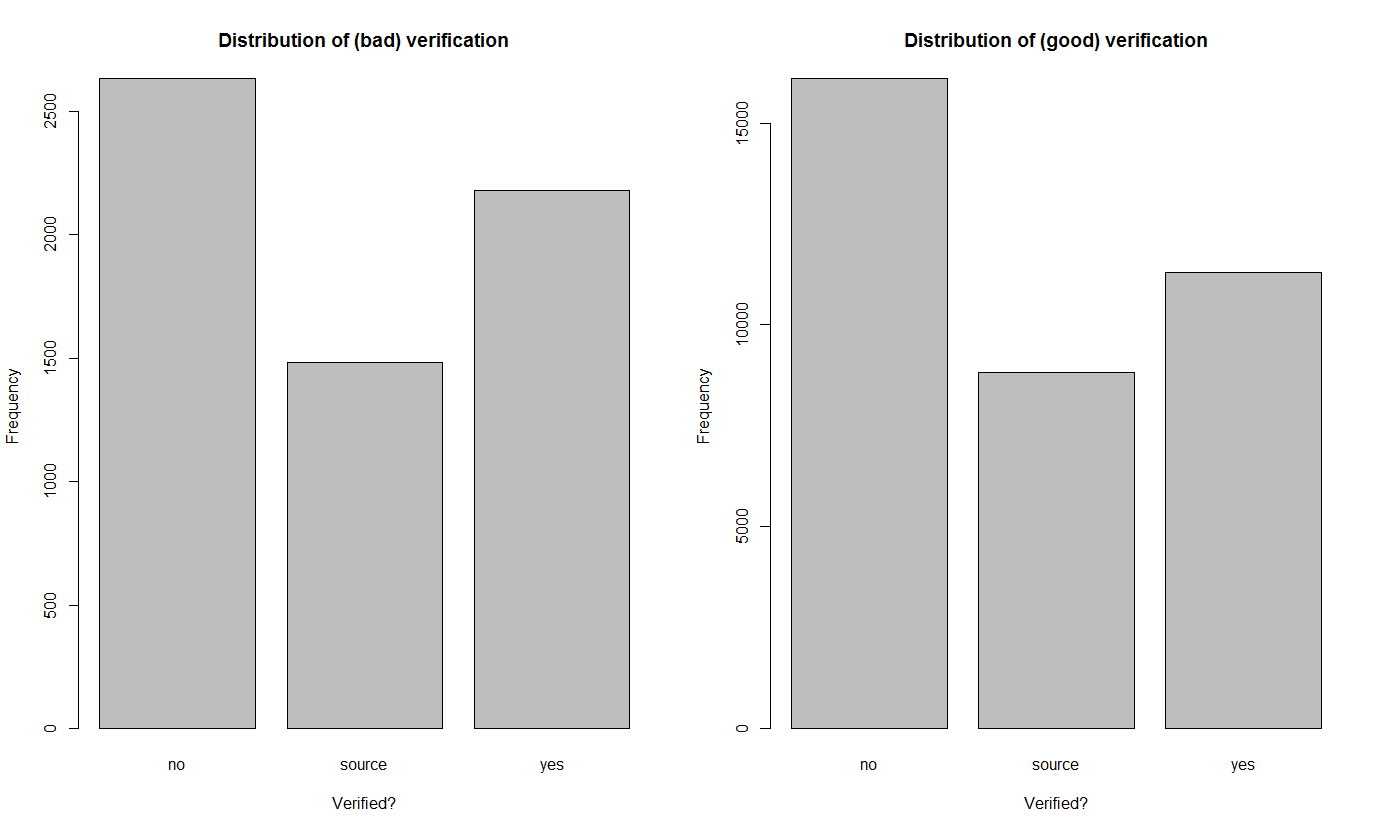


**Figure 11.** *Left: Distribution of Open Accounts; Right: Distribution of Open Accounts (log transformed)*

**VARIABLE: VERIFIED INCOME**

|  |  |  |  |
| --- | --- | --- | --- |
| **Verified?:** | No | Source | Yes |
| **Count:** | 18,729 | 10,306 | 13,471 |

**Table 12.** *Counts for different levels of Verified Income*



**Figure 10.** *Left: Distribution of Verification for people who didn’t pay off their loan; Right: Distribution of Verification for people who are good at paying off their loan.*

Annual\_inc, the annual income reported by the person asking for the loan, had some very extreme outliers, such as one person who reportedly had an annual income of six million. Therefore, I decided to make a factor variable from annual\_inc. I created the variable IncFact and gave it three levels: low, medium, and high. Those in the low bracket had an income less than $40,000, those in medium bracket had an income between $40,000 and $82,500, and those in the high bracket had an income greater than $82,500.

The variable is\_inc\_v was a bit wordy, as values would be “Verified”, “Source Verified” or “Not Verified,” therefore, I recoded that into a new variable VerifInc, which has values of “Yes”, ”Source”, or “No.”

The mths\_since\_last\_delinq has a numerical variable if the person ever had a delinquency, so rather than using this variable as is, I created a factor variable, DelinqFact, that has a value of “Yes” if there was a number value for mths\_since\_last\_delinq, and “No” if not.

The variable home\_ownership was recoded into the variable HomeFact, which is the same, except the few cases where the person put “NONE” as their answer were placed into the “OTHER” group.

After creating the variables above, I dropped the old variables mentioned. The other variables in consideration where left as is, just renamed. Finally, I dropped any incomplete cases. It is not the best option for dealing with missing data values, but in our case there were only 29 incomplete cases of the total 42,535 so it hopefully was not too costly of a decision to make.

So our variables are:

* loan\_amnt, renamed Loan\_Amnt, a numerical variable.
* grade, renamed Grade, a factor variable.
* HomeFact, a factor variable.
* IncFact, a factor variable.
* dti, renamed DTI, a numerical variable.
* fico\_range\_low, renamed FICO\_Low, a numerical variable.
* inq\_last\_6mths, renamed Inq\_6mnths, a numerical variable.
* DelinqFact, a factor variable.
* open\_acc, renamed Open\_Acc, a numerical variable.
* VerifInc, a factor variable.

**VARIABLE: GRADE**

The Grade variable was definitely my first of interest, as the variable, taking on values of A through G, is a grade already applied to the customer by Lending Club. This grade is the customers predetermined level of risk of defaulting on their loan. For instance, a grade of A means that the person has good credit history and is most likely to successfully pay off their loan. First I wanted to look at the distribution of Grades among all customers, which is shown in Table 1.