Insights into LendingClub.com Client Loan Portfolio: An Analytical Breakdown

Key words: Lending, Borrowing, Supervised Learning, Unsupervised Learning, Statistics, Clustering

1. Introduction

The project analyses the data provided by LendingClub.com, a financial services company that gives the opportunity to investors to choose potential borrowers.
It is divided into two parts:

1. **Unsupervised Learning**: this involves clustering clients based on their characteristics when they first approach the company. The goal is to identify specific groups of clients who are more likely to receive a loan.
2. **Supervised Learning**: this part aims to understand the variables that have the greatest influence on the investors and on the company's decision-making process regarding the client, such as the interest rate on the loan and whether the client complies with the company's credit policy. Subsequently, predictive models are created to forecast the company's decisions, with the best model(s) being evaluated for effectiveness.

The data was processed using R programming language within the Rsudio environment. Version control was conducted through GitHub.

2. Data Used

2.1 Source and Columns Outline

The dataset, uploaded by Sara Mahdavi on Kaggle in CSV format, contains lending data from 2007 to 2010. Here is an outline of the columns:

Categorical Variables:

- credit.policy: binary variable indicating whether the customer meets the credit underwriting criteria of LendingClub.com (1 for meeting criteria, 0 otherwise);
- purpose: the purpose of the loan, with categories including *"credit_card*", "*debt_consolidation*", "*educational*", "*major_purchase*", "*small_business*", and "*all_other*";
- not.fully.paid: Binary variable indicating whether the loan is not fully paid (1) or is fully paid (0).

Numerical Variables:

- int.rate: the interest rate of the loan;
- log.annual.inc: the natural log of the borrower self-reported annual income;
- installment: the monthly installments owed by the borrower if the loan is funded;
- dti: the debt-to-income ratio of the borrower, calculated as the amount of debt divided by annual income;
- fico: the FICO credit score of the borrower;

- days.with.cr.line: the number of days the borrower has had a credit line;
- revol.bal: the borrower's revolving balance, i.e., the amount unpaid at the end of the credit card billing cycle;
- revol.util: this represents the borrower's revolving line utilization rate, which indicates the proportion of the credit line used relative to the total credit available, expressed as a percentage. Generally, a higher utilization rate suggests that the client relies more heavily on debt, posing greater risk;
- risk.inq.last.6mths: the borrower's number of inquiries by creditors in the last 6 months.
- delinq.2yrs: The number of times the borrower has been 30+ days past due on a payment in the past 2 years;
- pub.rec: the borrower's number of derogatory public records, such as bankruptcy filings, tax liens, or judgments.

The dataset comes with 9488 records, with each line corresponding to a user of the website.

2.2 Data Manipulation

Utilising the column provided by the dataset three other columns have been created:

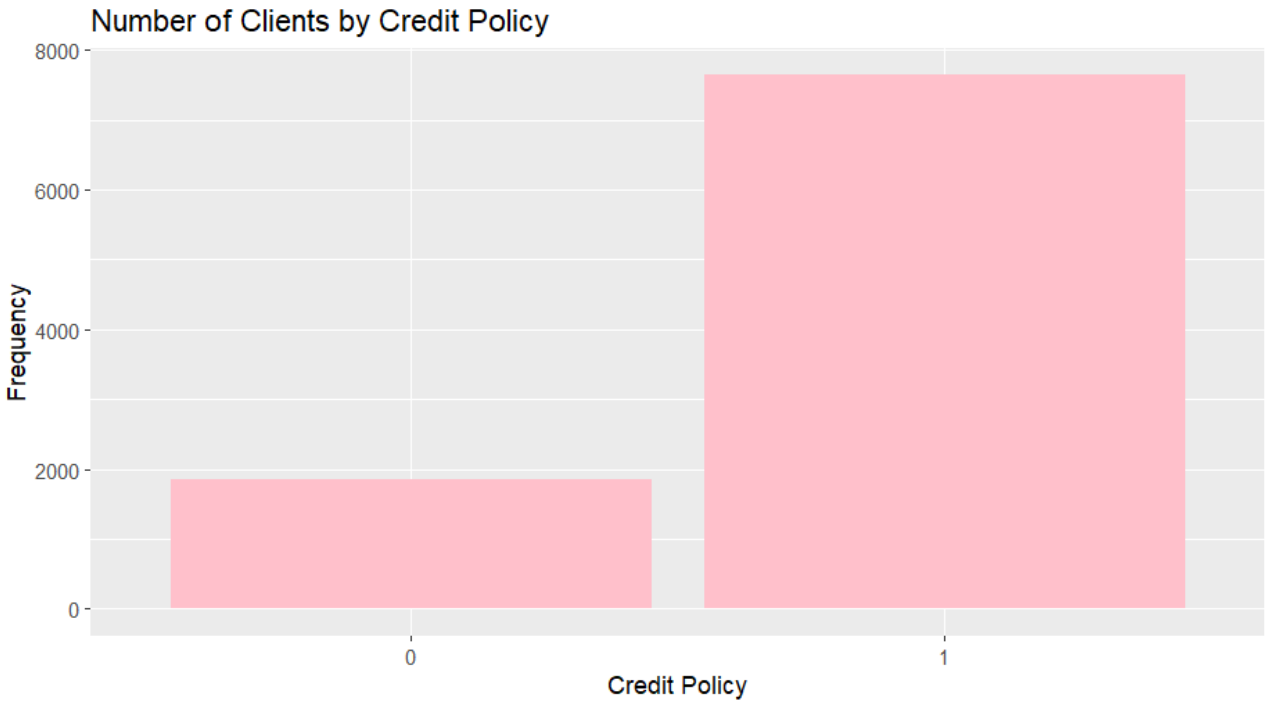
- annual.inc: exponential of the natural logarithm of the annual income (log.annual.inc);
- debt: product of the annual income (annual.inc) and the debt-to-income ratio (dti);
- iti: installment to monthly income ratio.

3. Data Visualization

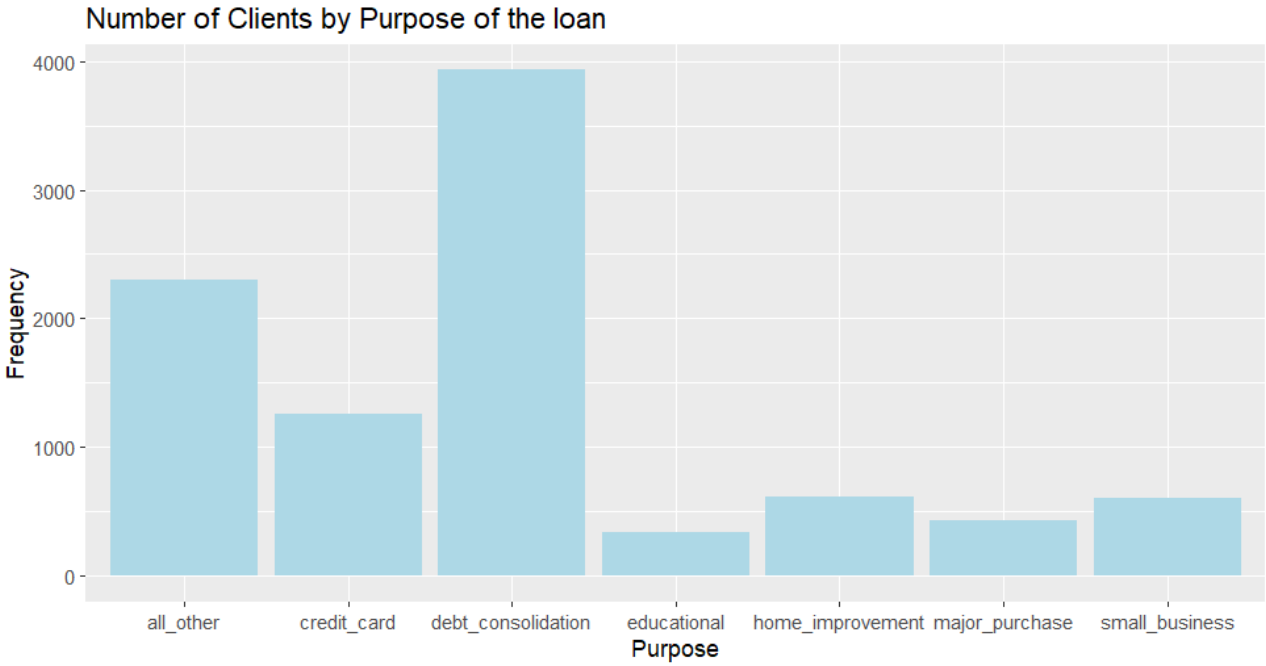
Title: Exploring Data Through Visualization

The objective is to delve into the data, seeking out useful features for answering to our questions.

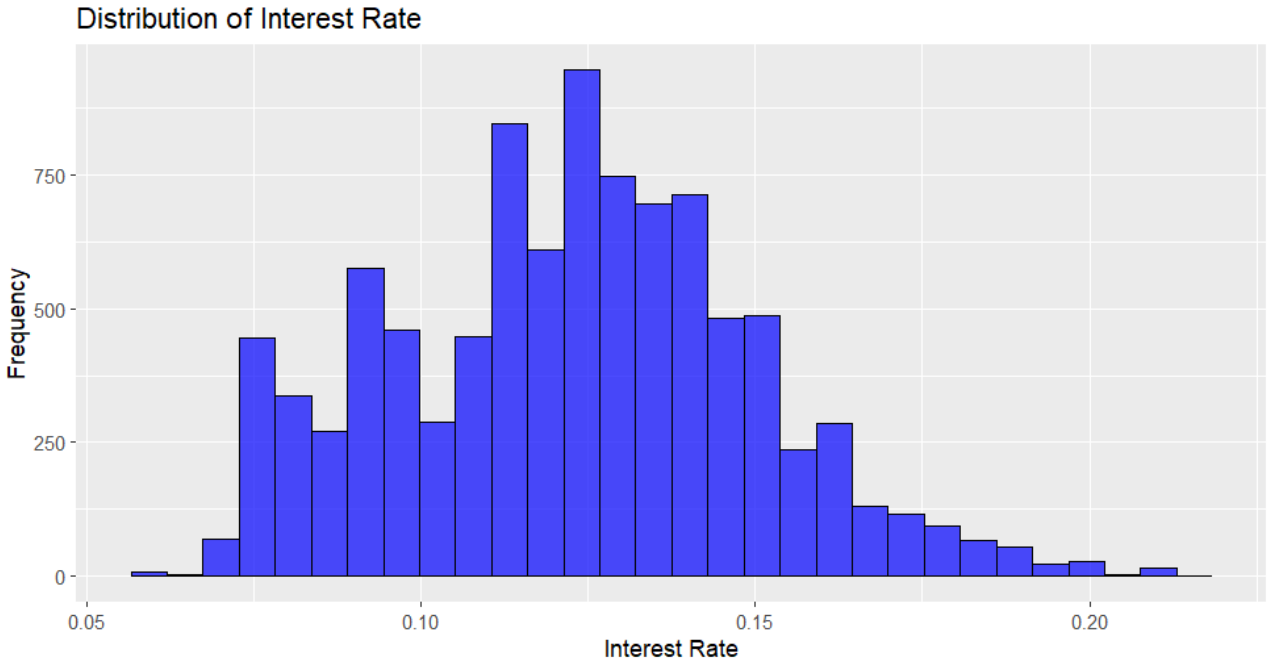
Borrowers will be referred to as “Clients” sometimes.



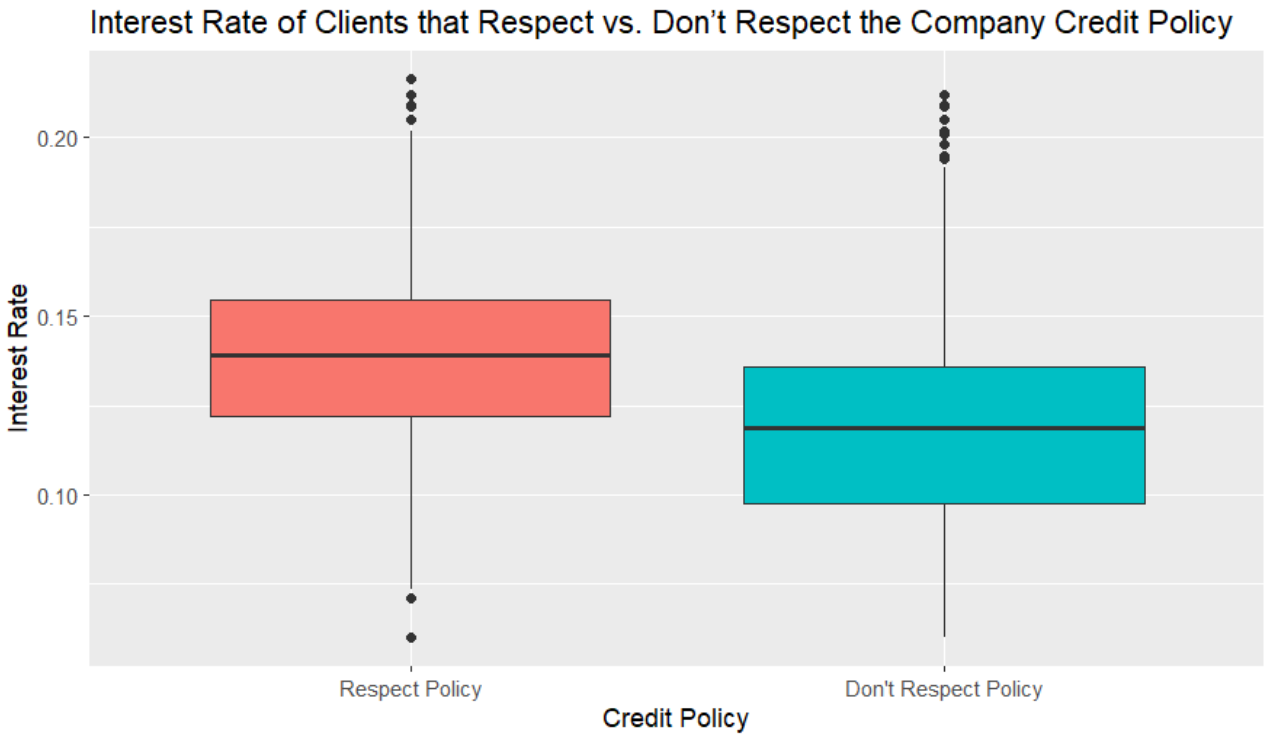
A high majority of the borrowers (80.5%) respect the underwriting criteria of LendingClub.com, while the others (19.5%) don't. This variable will be used to compare different clients, to understand if there's a difference between clients that respect the policy and the others.



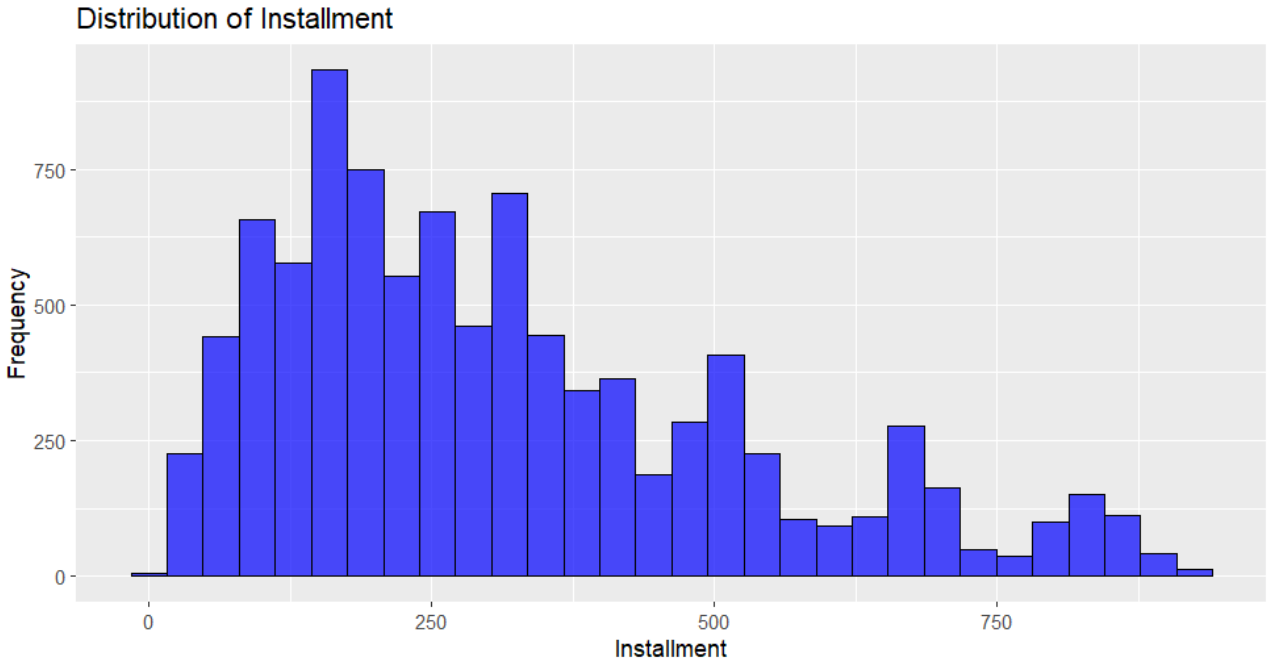
Most of the users ask for a loan for debit consolidation, followed by credit card payments, home renovations, small business funding, major purchases, educational purposes and others.



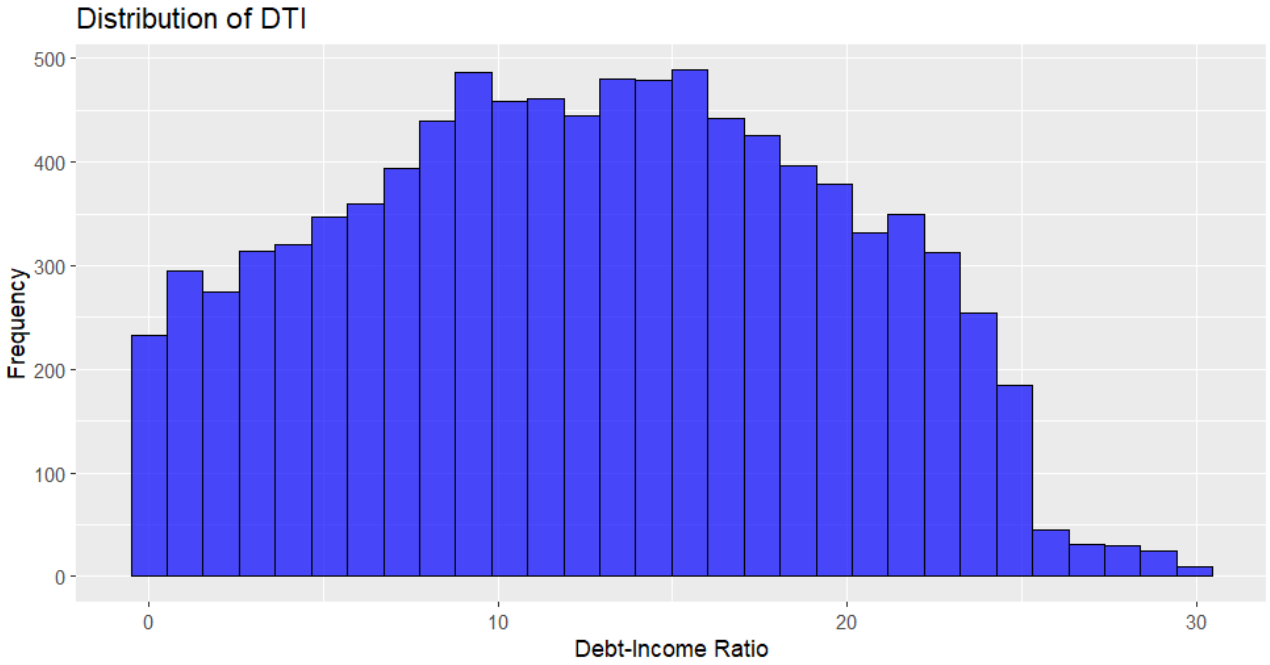
The interest rate data looks like a right-skewed distribution. The median interest rate is 0.1221. The 75% of the records presents an interest rate lower than 0.14.



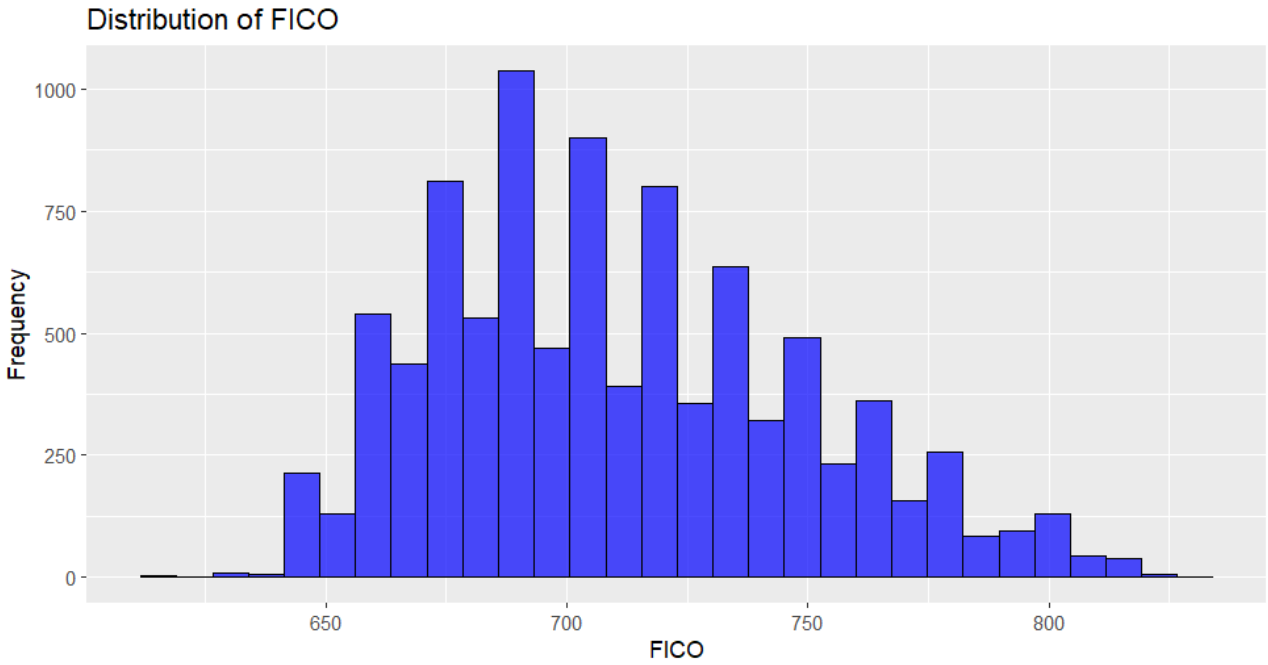
The Interest rate is higher for the clients that respect the company policy.



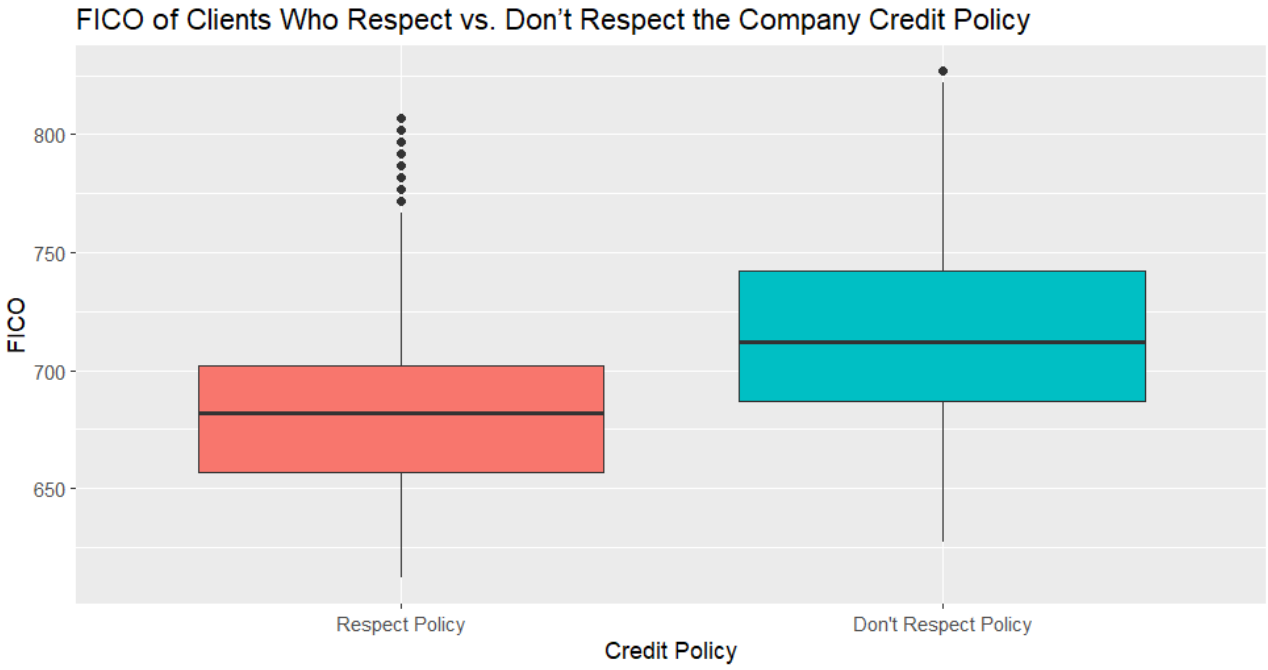
The Installment, the monthtly payment if the lending is funded, has a right-skewed distribution. The median is 268.42 and the 75% of the records have an installment lower than 428.65.



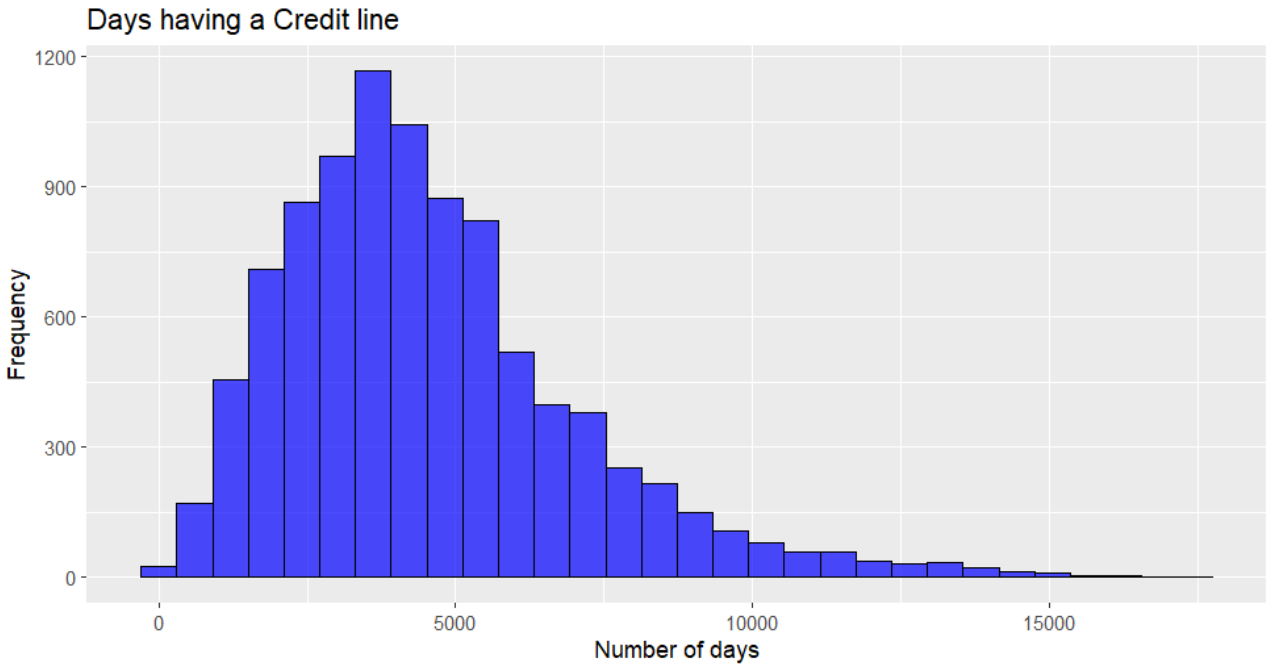
Debt-Income Ratio is a measure that can give an idea of the stability of the potential borrower. The median is 12.72. The frequency sinks when the dti is bigger than 25.



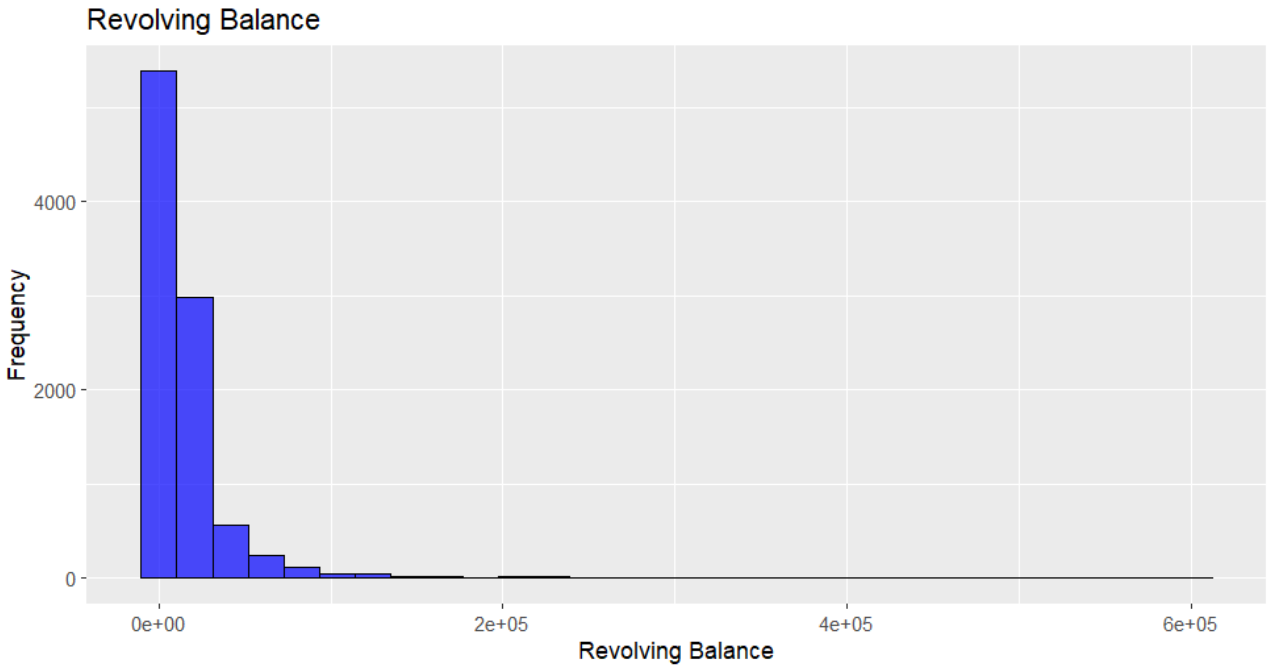
The FICO Credit Score has a range of [612-827], with a median of 707.



Borrowers that respect the credit underwriting criteria have a lower credit score.

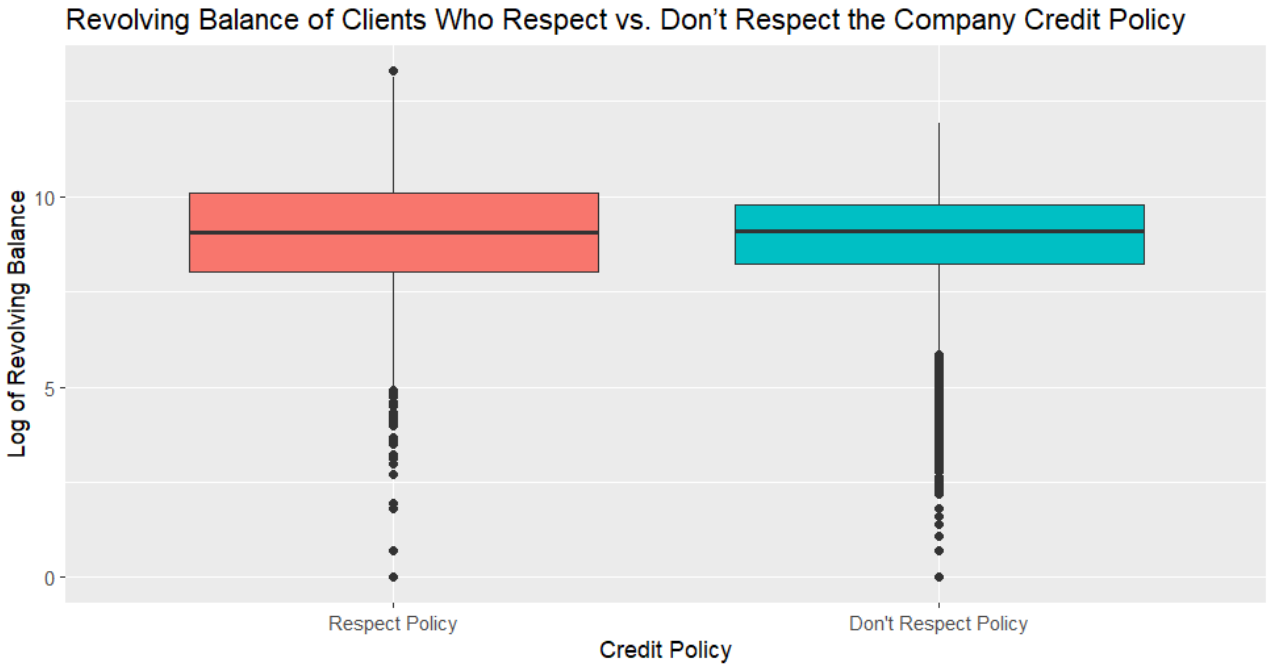


Days of Credit lines follow a right-skewed distribution, with a median of 4110.5.

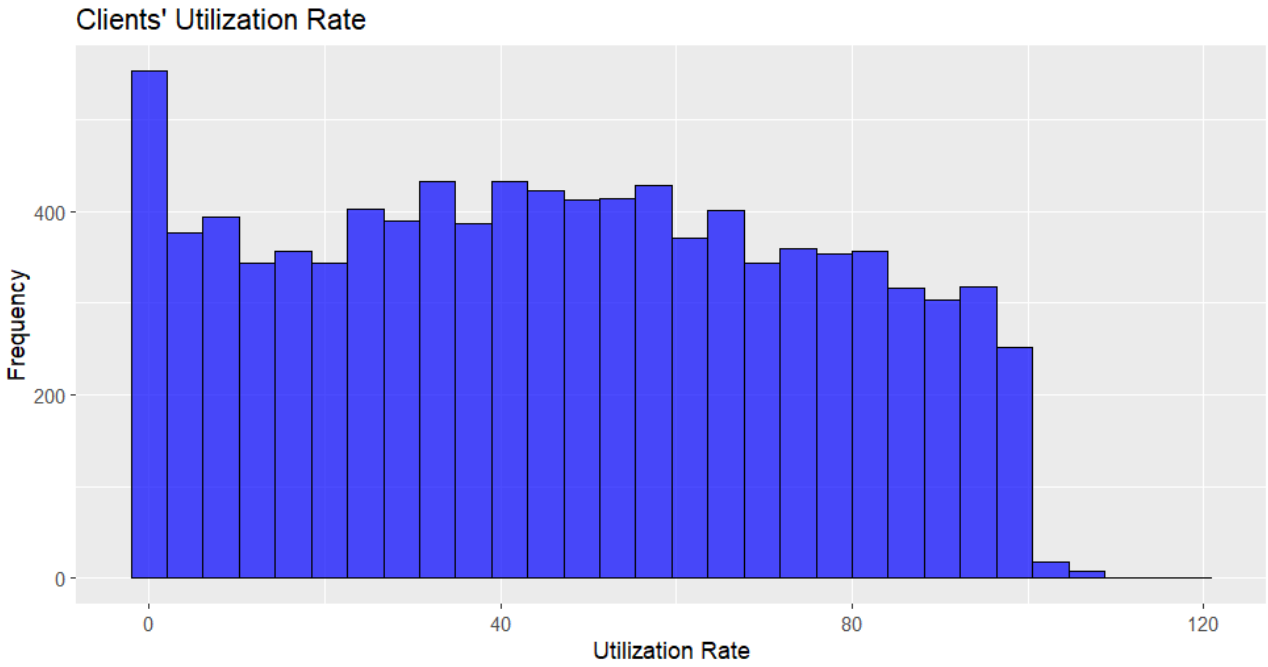


Most of the borrowers do not have a debt due to monlthy credit card payments.
It's iteresting to notice that the percentage of clients having 0 revolving balance within the two categories:

- Only 2.8% of clients that respect the credit criteria have a 0 revolving balance;
- The 5.2% of clients that respect the credit criteria have a 0 revolving balance.

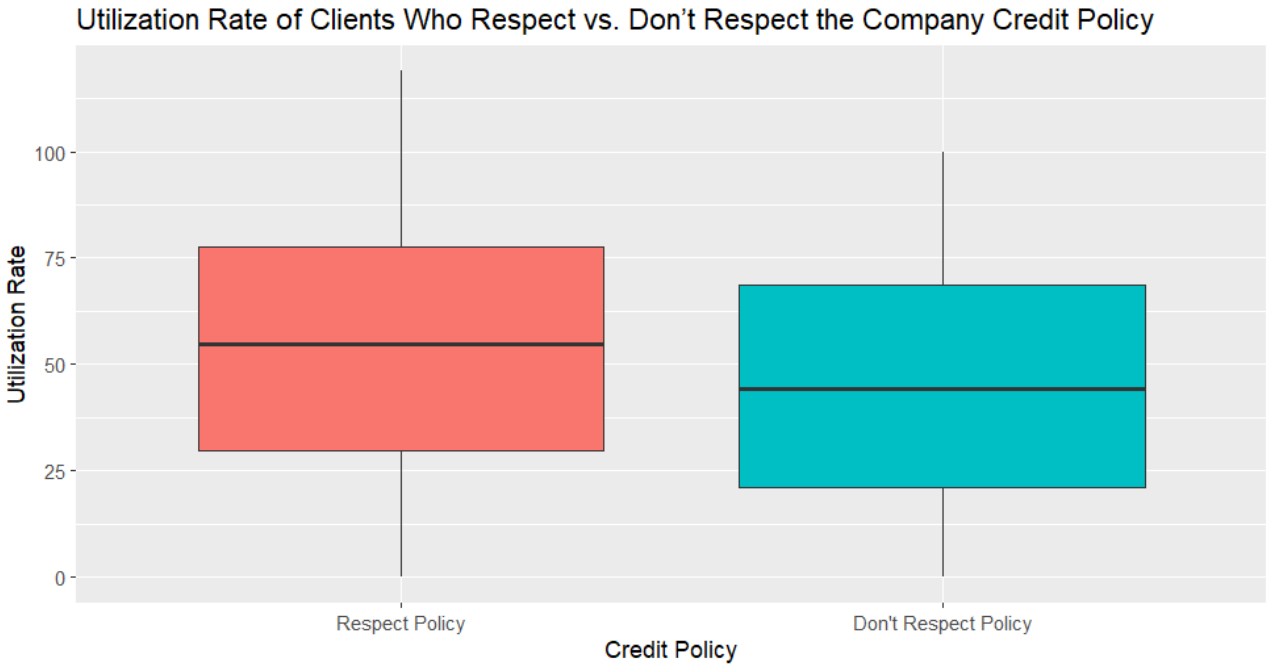


There's not significant difference between the two category, concerning Revolving Balance.

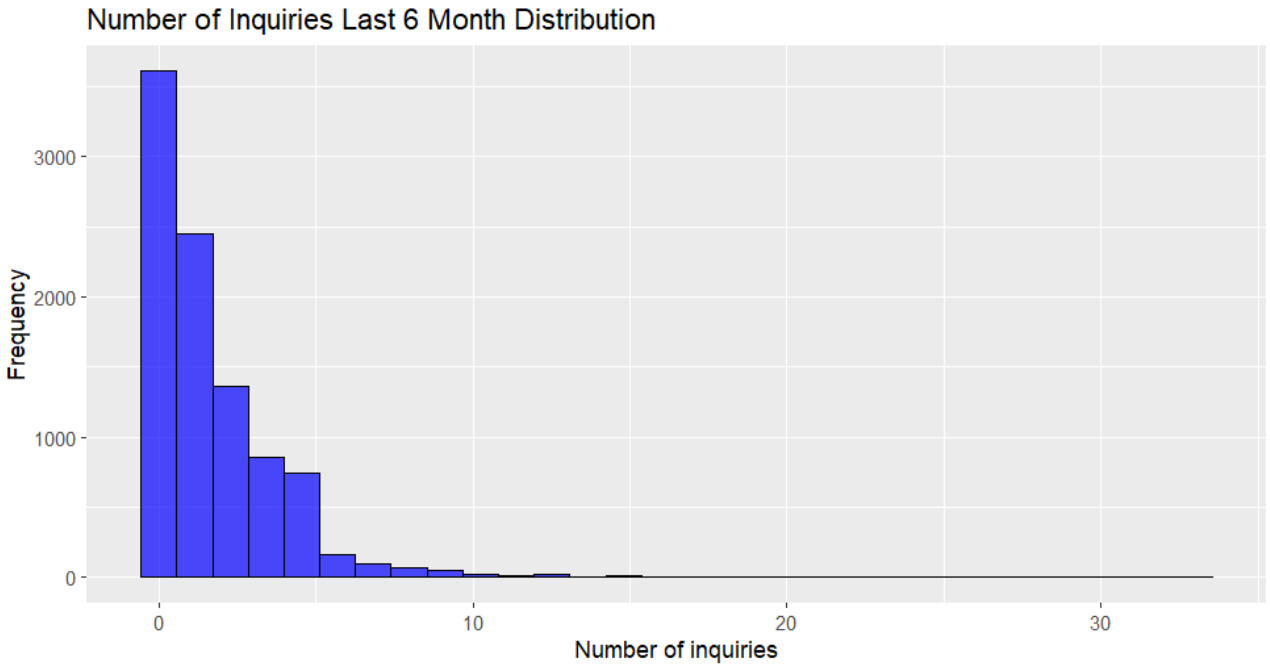


The median utilisation rate is 46.2%.

Few borrowers have an utilisation rate lower than 100%

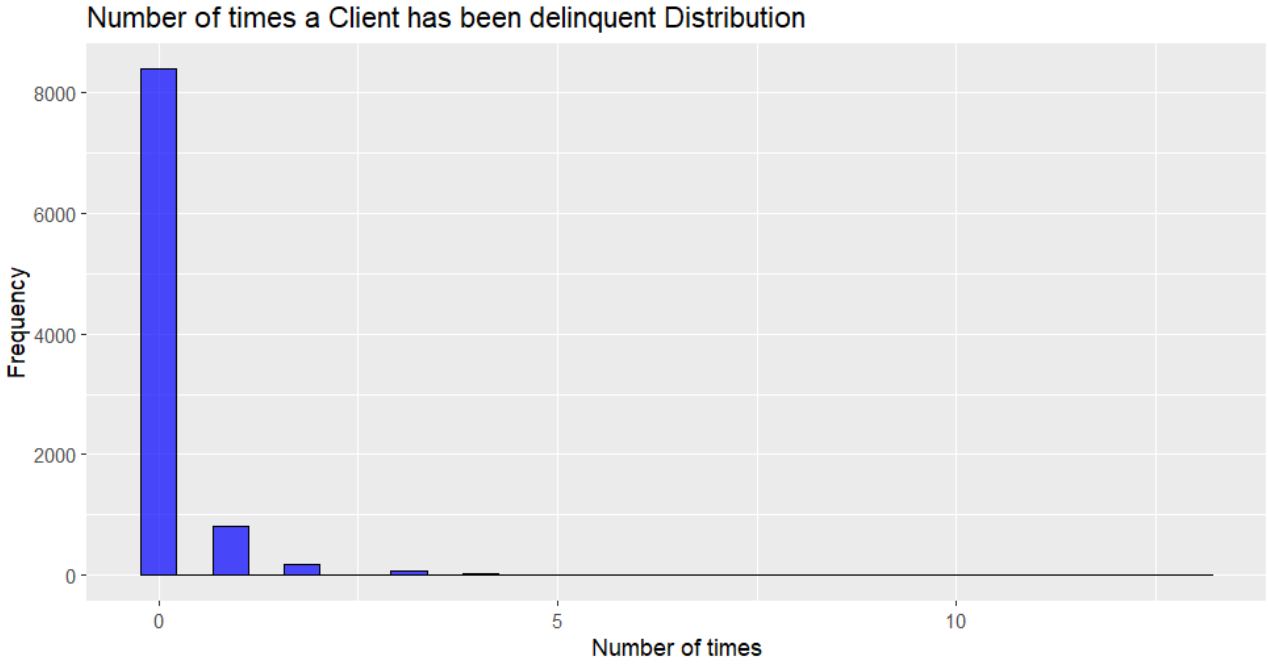


We have an higher utilisation rate for clients that respect the company's credit criteria.

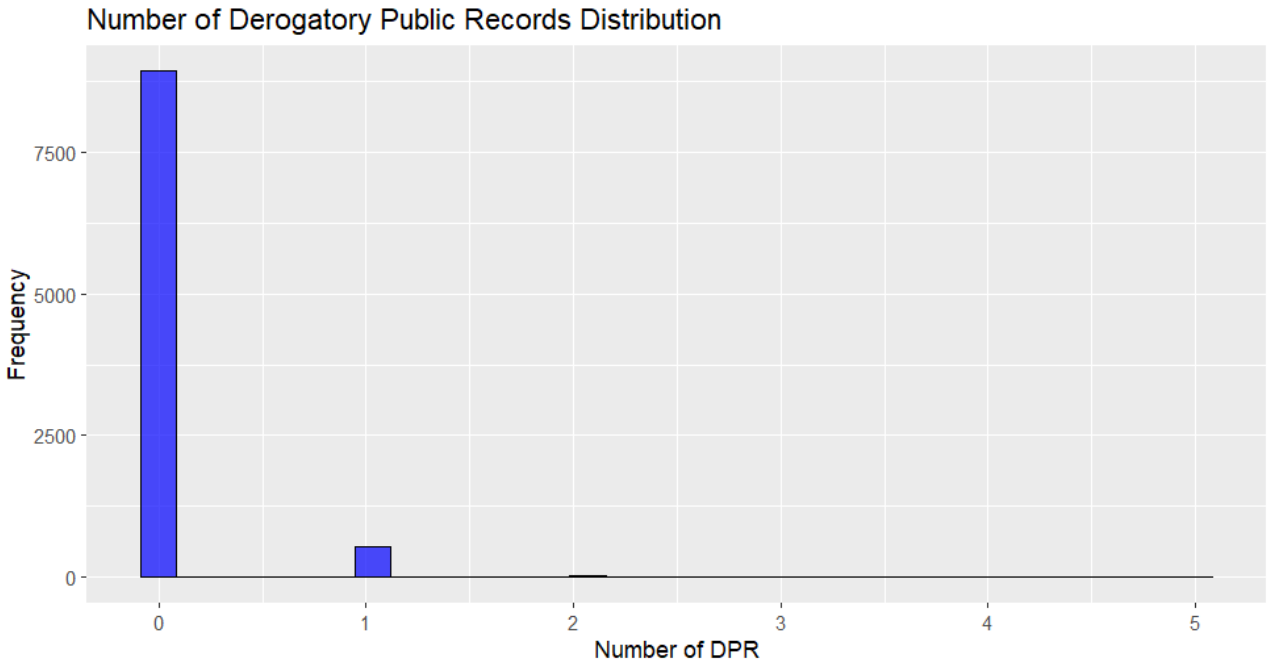


The 75% percent of the borrowers have a number of inquiries lower or equal to 2 in the last 6 months.

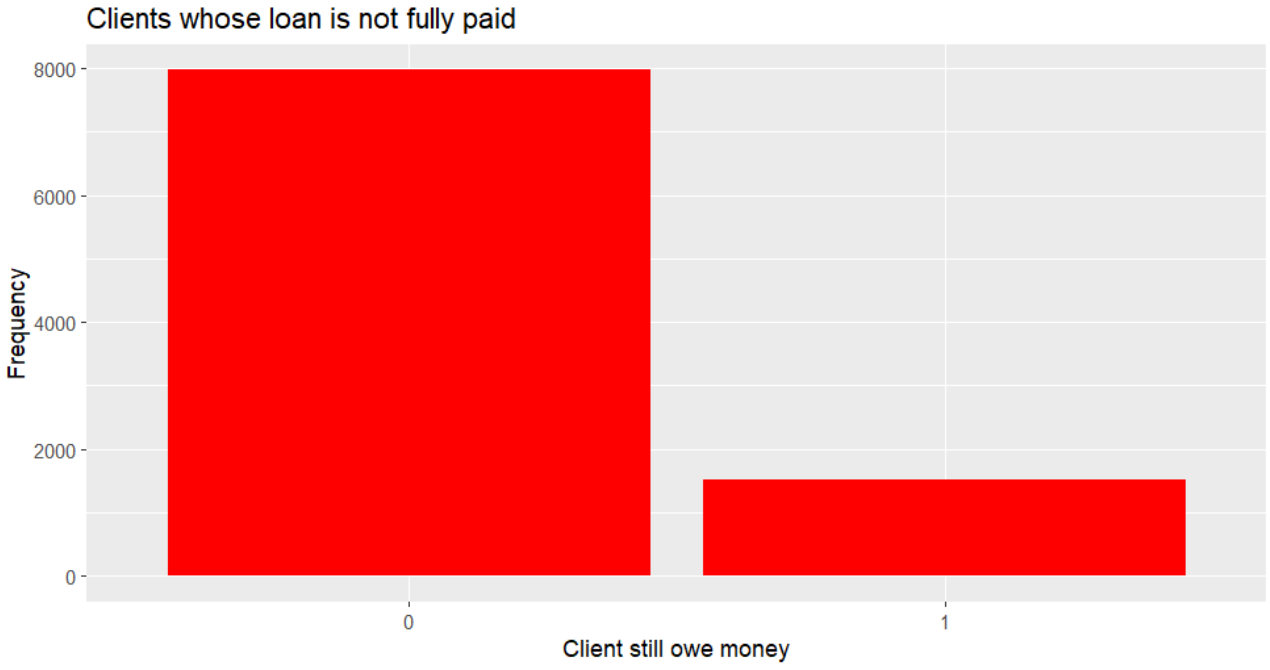
Borrowers that respect the company's credit policy have an higher number of inquiries.



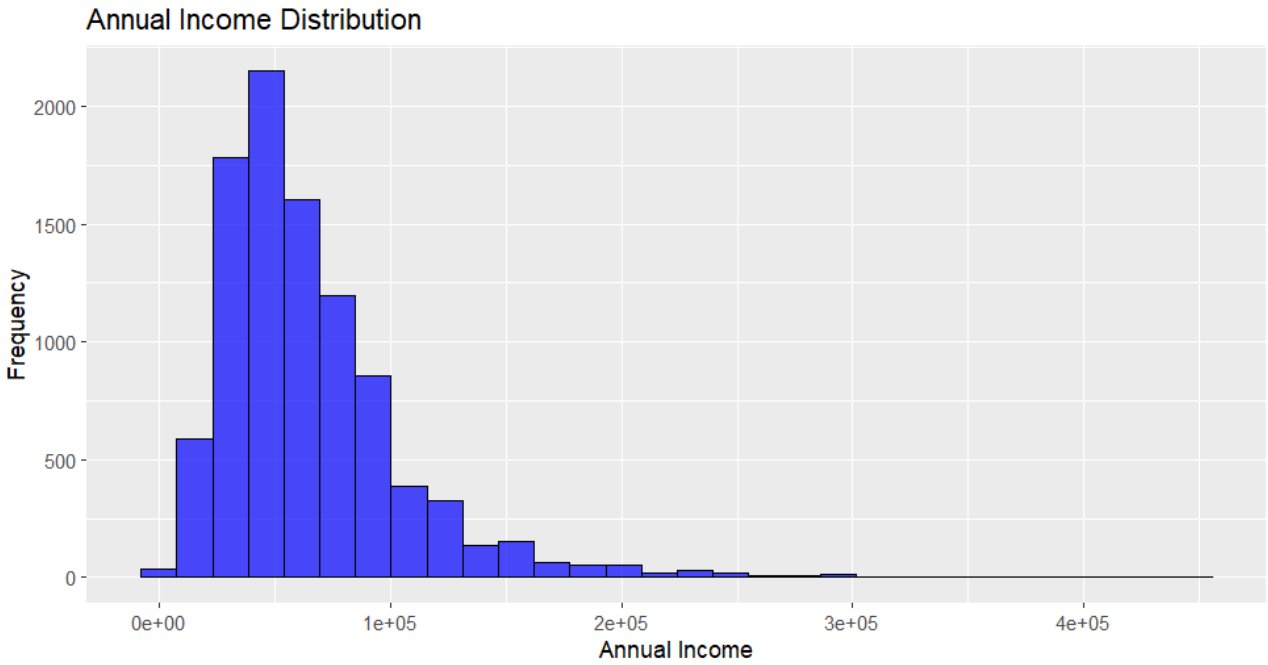
Most of the borrowers have never been delinquent in the last 2 years.



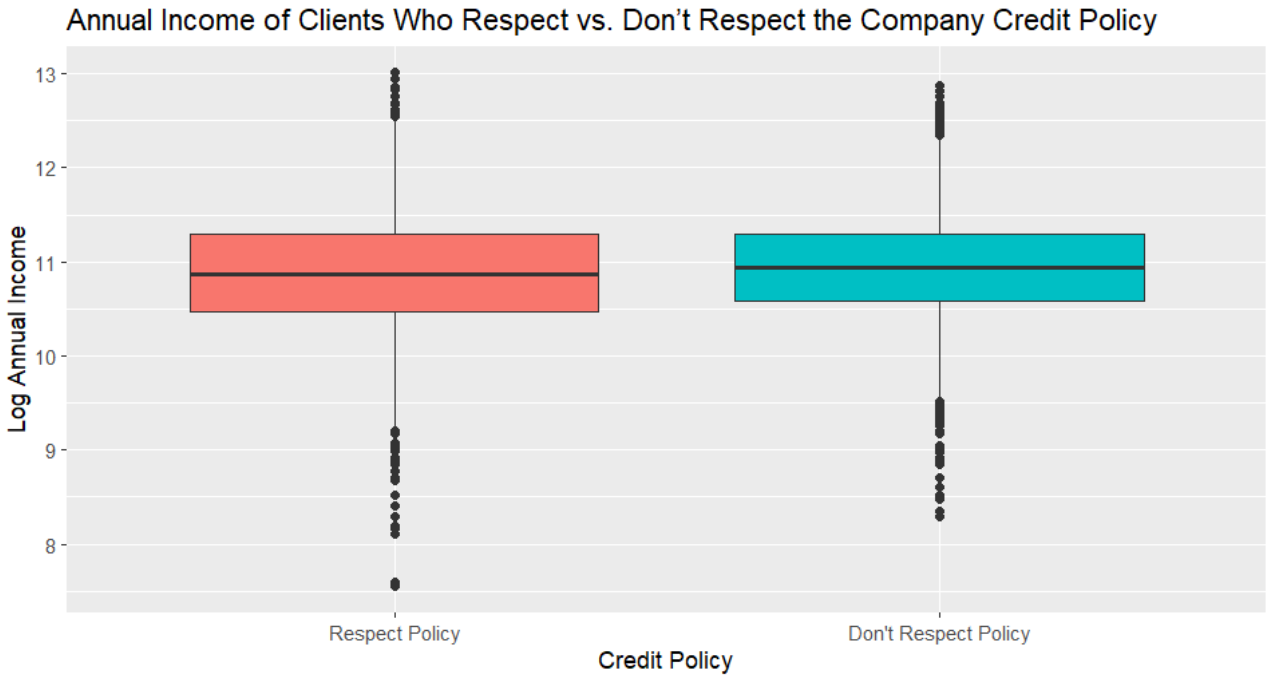
Most of the borrowers have not derogatory public records (bankruptcy filings, tax liens, or judgments).



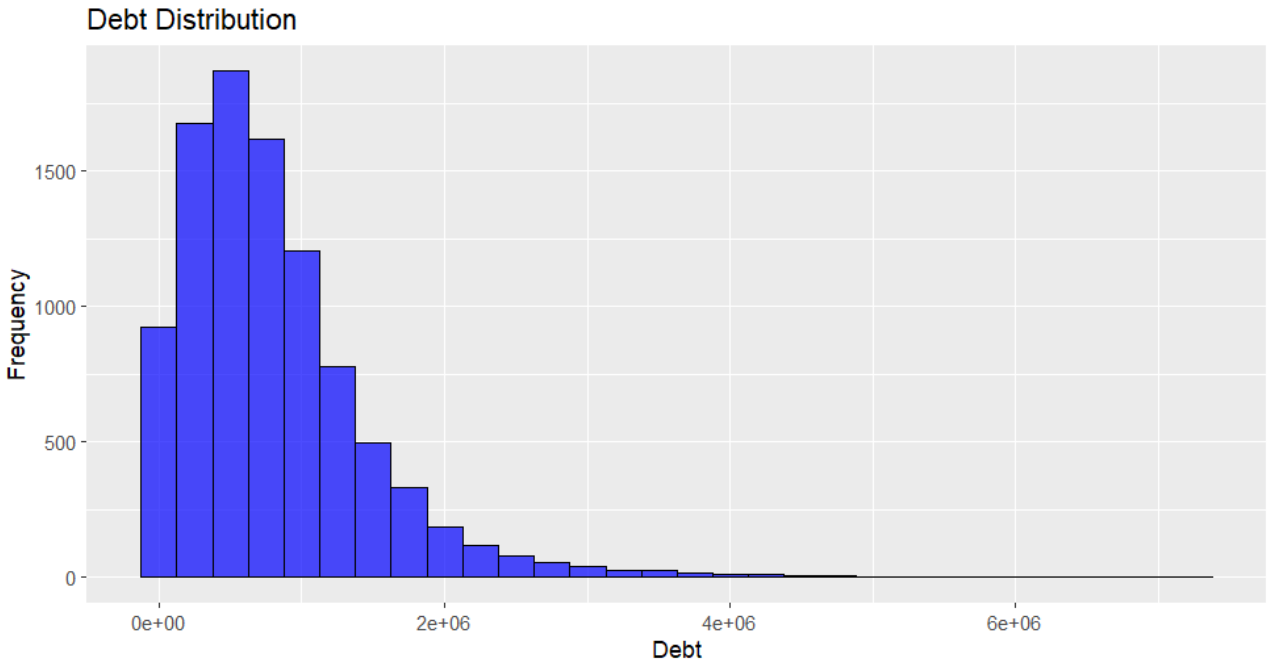
Most of the clients have not fully paid the loan.



Annual income follow a right-skewed distribution with median equal to 55,764.

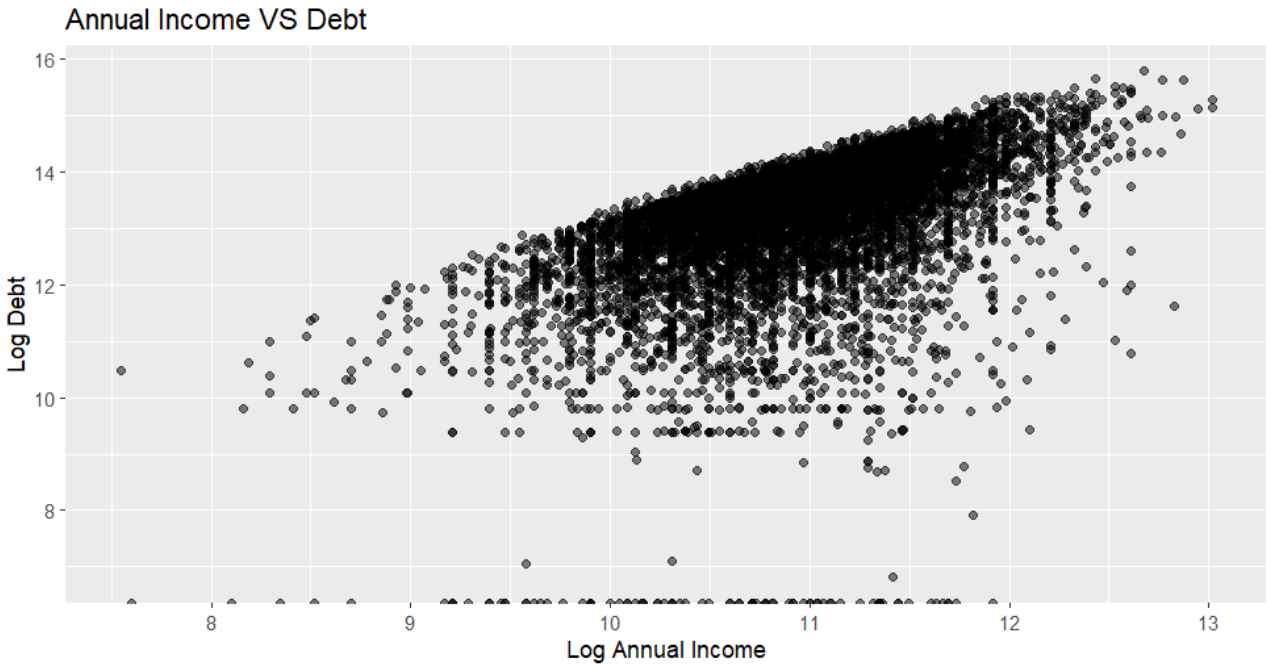


There is not a significant difference between the two groups reguarding Annual Income.



Debt follow a right-skewed distribution with median 666,795.7.

There's not significant difference reguarding debt between the two groups.



It appears that there's a relationship between income and maximum debt: as income increases, the maximum debt also tends to increase. There are instances where borrowers request amounts below this limit, indicating that they don't always seek the maximum loan they could qualify for.

4. Insights from the data

There are multiple variables used to measure the risk of a borrower defaulting on their loan. Some like the FICO credit score, utilization rate, number of inquiries, and the absence of monthly credit card unpaid debt, suggest that those not adhering to the company's credit policy present a lower risk. Furthermore, clients who abide by the policy are offered higher interest rates. This seems counterintuitive, as the credit policy is designed to categorize clients based on their financial stability and maximize the likelihood of repayment. These doubts highlight the need for further data analysis.

5. Ask Fase

The main goal of this analysis is to answer the following questions:

- What variables influence the company's decision on whether clients meet the company's credit policy or not?
- Which variables impact the decision on interest rates?
- Can borrowers be grouped (clustered) based on the available data before investors decide to lend them money? Are these clusters meaningful, and what are the significant differences between them?

6. What variables influence the company's decision on whether clients meet the company's credit policy or not?

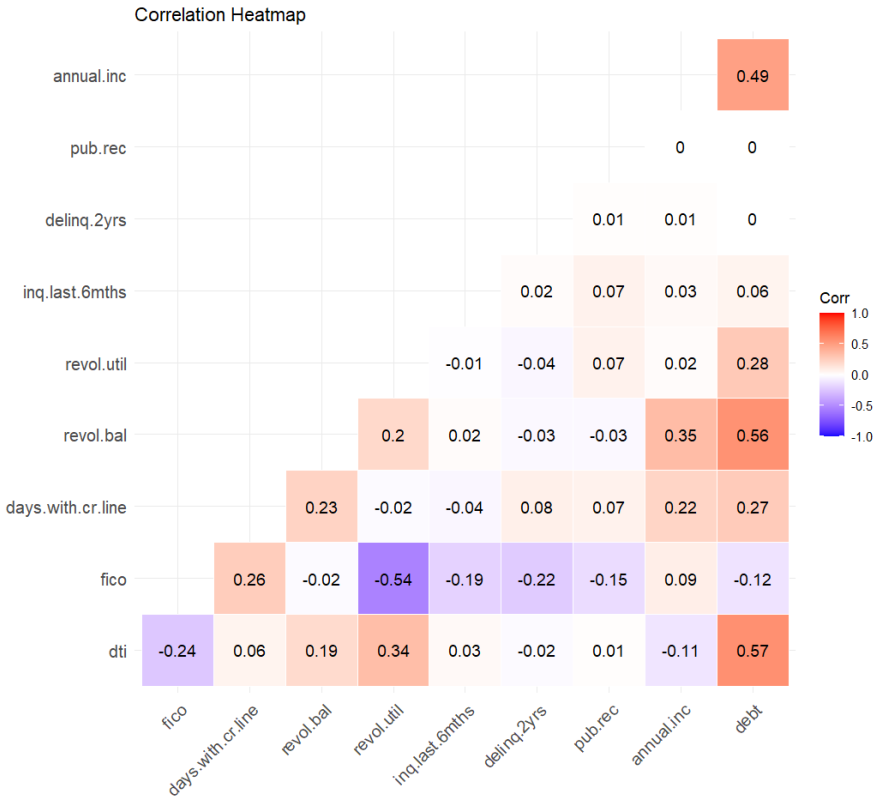
To explore this, supervised learning techniques have been employed.

6.1 Column Selection

Given that the goal is trying to understand which variables make the company put its mark on the potential borrower, the not relevant variables have been excluded by the model: the interest rate, the total interest, the monthly instalment and the fact that a borrower has repaid the loan are information not available when the borrower first approach the website. This because they are decided later, when the investor already selected the potential borrower.

The idea is to create a general linear model specifying the binomial family. This because we are modelling a binary variable that can't be negative.

The following is a correlation heatmap, to check if there is a risk of multicollinearity between the independent variables of the model:



There is a high positive correlation between debt amount and debt to income ratio, debt amount and revolving balance, debt and annual income, and credit score and revolving line utilisation rate. This could cause multicollinearity.

To decide if the high correlation is a problem the Variance Inflation Factor (VIF) has been computed:

$$VIF = \frac{1}{1 + R^2}$$

The VIF results for these variables are the following:

Variable	Variance Inflation Factor
FICO	1.594081
Annual Income	3.727031
Debt	5.865977

The variable *debt* has a moderately high VIF, meaning that could cause multicollinearity.
3.727031
The decision is to keep it for the moment, trying to get as much information as possible from the dataset.

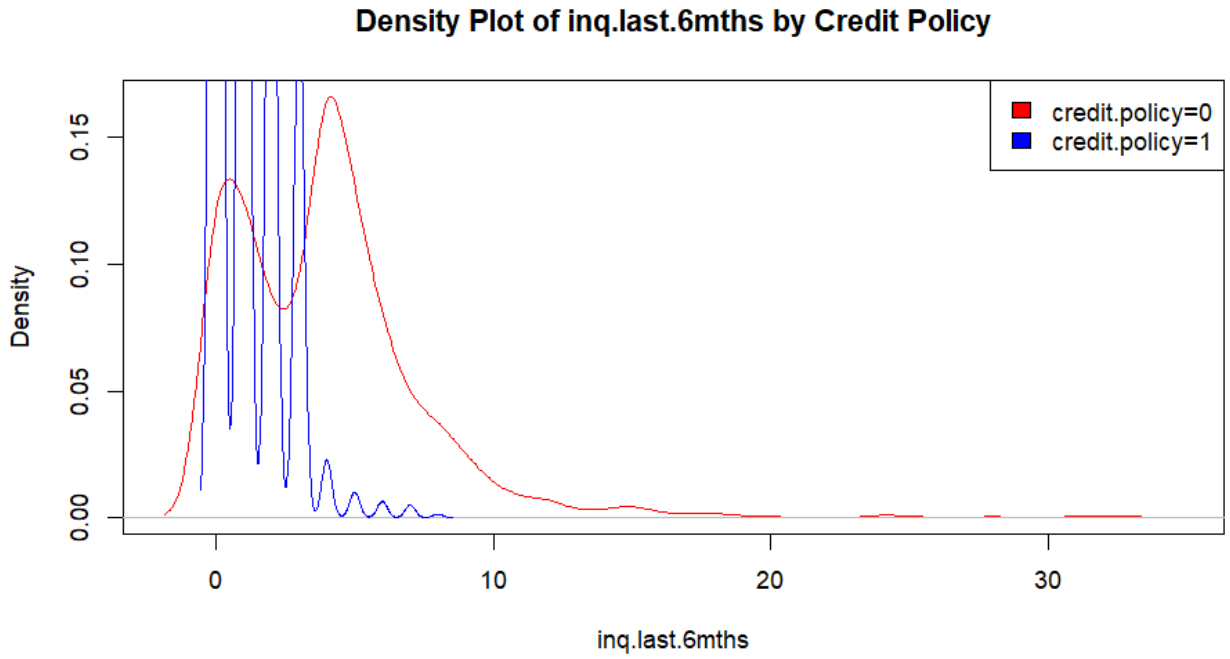
6.2 Simple Generalised Linear Model

The model is the following, and it includes all the variables selected in the *6.1* step:

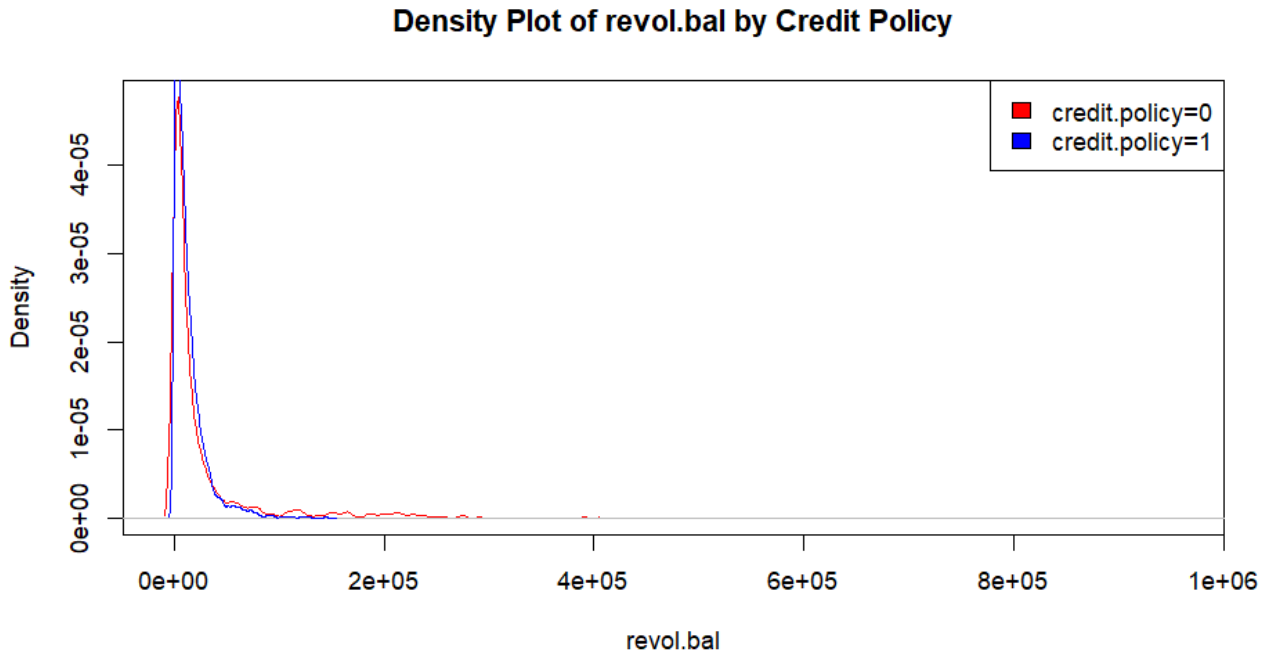
```
Call:
glm(formula = credit.policy ~ ., family = binomial(), data = data.train.1)

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               -3.027e+01  1.359e+00 -22.278  < 2e-16 ***
purposecredit_card         1.916e-01  1.440e-01   1.330  0.18335
purposedebt_consolidation  4.585e-01  1.078e-01   4.253 2.11e-05 ***
purposeeducational         2.823e-01  2.289e-01   1.233  0.21745
purposehome_improvement    4.377e-01  1.997e-01   2.192  0.02838 *
purposemajor_purchase      2.509e-01  2.349e-01   1.068  0.28552
purposesmall_business      5.782e-01  1.971e-01   2.934  0.00334 **
dti                        2.273e-03  1.067e-02   0.213  0.83138
fico                       4.633e-02  1.934e-03  23.953  < 2e-16 ***
days.with.cr.line          1.496e-04  2.174e-05   6.885 5.80e-12 ***
revol.bal                 -4.408e-05  2.283e-06 -19.304  < 2e-16 ***
revol.util                 9.549e-03  1.832e-03   5.213 1.86e-07 ***
inq.last.6mths            -9.417e-01  2.843e-02 -33.128  < 2e-16 ***
delinq.2yrs               -8.232e-02  7.164e-02  -1.149  0.25048
pub.rec                    4.101e-02  1.488e-01   0.276  0.78279
annual.inc                 8.869e-06  1.867e-06   4.750 2.04e-06 ***
debt                       1.606e-07  1.475e-07   1.089  0.27611
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The most statistically significant variables are FICO Credit Score (positive coefficient), days with a credit line (with a positive coefficient), revolving balance (negative coefficient), the revolving line utilisation rate (positive coefficient), number of inquiries in the last 6 months (negative coefficient), purpose equal to debt consolidation (positive coefficient) and annual income (positive coefficient). When running the model, the RStudio compiler returned a warning: some variables may be deterministic, potentially leading to misleading results and higher coefficients. Comparing the distribution of the variables of the model in the two cases (respecting company policy or not), it

has been possible to highlight the two deterministic variables: the number of inquiries in the last 6 months and the revolving balance.



After 9 inquiries, there are not records that respect company's credit policy.



After a certain treshold of revolving balance, there are not records that respect company's credit policy. One solution could be deleting the variables from the model, losing some information, or to use alternative models that shrink the coefficient of deterministic variables.

Before introducing these alternatives model, the current one has been tested, splitting the dataset in training set (80% of the records) and test set (20% of the records).

Then some indices for model evaluation has been computed.

Confusion Matrix:

Prediction\Reference	0	1
0	233	47
1	136	1500

The model obtained 1550 true positives (TP), 233 true negatives (TN), 136 false positives (FP)and 47 false negative (FN).

The AUC (Area under the curve of the Receiver Operating Characteristic) is 0.8045, meaning that the model has a high discrimination power.

The Accuracy $\frac{TP+TN}{TP+TN+FP+FN}$ is equal to 0.9044885.

The Precision $\frac{TP}{TP+FP}$ is equal to 0.9168704.

The Recall $\frac{TP}{TP+FN}$ is equal to 0.9696186.

The F1 score, a harmonic average between of Precision and Recall, is 0.9425071.

The model achieved excellent results, indicating that the variables used by the company to determine "company.policy" align closely with those included in the model.

6.3 Ridge Linear Regression

Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where the independent variables are highly correlated. In the previous model, the variable “debt” is highly correlated with some other variables, and it has an high VIF score.

Creating a new model using Ridge Linear Regression, we obtain the following coefficient:

```
18 x 1 sparse Matrix of class "dgCMatrix"
                                      s1
(Intercept)                -1.556303e+01
purposeall_other           -1.824037e-01
purposecredit_card         -4.478438e-02
purposedebt_consolidation   1.211160e-01
purposeeducational         -9.640654e-02
purposehome_improvement     6.915024e-02
purposemajor_purchase       2.449676e-02
purposesmall_business       1.240550e-01
dti                        -4.730238e-03
fico                        2.547477e-02
days.with.cr.line           1.019167e-04
revol.bal                  -2.139990e-05
revol.util                  2.383396e-03
inq.last.6mths             -5.986062e-01
delinq.2yrs                -1.196706e-01
pub.rec                    -2.032905e-02
annual.inc                  2.968252e-06
debt                        4.849131e-08
```

The coefficient of the revolving balance is now closer to zero, meaning that it affects less the decision. On the contrary, the variable corresponding to the number of inquiries has now a coefficient bigger in absolute terms. This happens because the Ridge Regression does not shrink directly the coefficient, but adds a penalty term to the loss function, meaning is not guaranteed that the deterministic variables will have smaller coefficients.

Some other variables changed considerably their coefficient: “days.with.cr.line” has now o coefficient of $8.859031 * e^{-5}$, while before was $1.350 * e^{-4}$.

Some other coefficients changed sign, such number of derogatory public records and the categorical variable “purpose” when is equal to “credit_card”.

Testing the model dividing the data (80-20) we obtain the following results:

Confusion Matrix

Prediction\Reference	0	1
0	197	18
1	172	1529

The model obtained 1529 true positives (TP), 197 true negatives (TN), 172 false positives (FP) and 18 false negative (FN).

The AUC (Area under the curve of the Receiver Operating Characteristic) is 0.7635, meaning that the model has a high discrimination power. The result is lower (-4.2%) than with the model 1 (*6.2*). The Accuracy $\frac{TP+TN}{TP+TN+FP+FN}$ is equal to 0.9008351. The result is lower but similar than with the model 1.

The Precision $\frac{TP}{TP+FP}$ is equal to 0.8988. The result lower (-1.8%) than with the model 1.

The Recall $\frac{TP}{TP+FN}$ is equal to 0.9883. The result is higher (+1.8%) than with the model 1.

The F1 score is 0.9415025, similar to the one with model 1.

The best model could be chosen based on the final objective: in the context of classifying clients, it's important to define both 0 and 1 instances, because the investor must know if the potential borrowers are reliable or not (from the company point of view). Consequently, model 1 is better than model 2, even if it's simpler.

6.4 Bi-Directional Step-wise Selection

Another way to improve the model could be delete from the model the less relevant variables. The bi-directional Step-wise selection has been employed, which combine both forward and backward selection.

```
Call:
glm(formula = credit.policy ~ purpose + fico + days.with.cr.line +
    revol.bal + revol.util + inq.last.6mths + annual.inc + debt,
    family = binomial(), data = data.train.1)

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               -3.061e+01  1.307e+00 -23.421  < 2e-16 ***
purposecredit_card         1.927e-01  1.439e-01   1.339  0.18042
purposedebt_consolidation  4.601e-01  1.076e-01   4.276 1.90e-05 ***
purposeeducational         2.851e-01  2.287e-01   1.247  0.21248
purposehome_improvement    4.453e-01  1.996e-01   2.231  0.02570 *
purposemajor_purchase      2.455e-01  2.344e-01   1.047  0.29498
purposesmall_business      5.757e-01  1.966e-01   2.928  0.00341 **
fico                       4.682e-02  1.870e-03  25.040  < 2e-16 ***
days.with.cr.line          1.454e-04  2.095e-05   6.942 3.87e-12 ***
revol.bal                 -4.398e-05  2.268e-06 -19.386  < 2e-16 ***
revol.util                 1.007e-02  1.776e-03   5.667 1.45e-08 ***
inq.last.6mths            -9.397e-01  2.835e-02 -33.150  < 2e-16 ***
annual.inc                 8.522e-06  1.392e-06   6.121 9.31e-10 ***
debt                       1.868e-07  9.352e-08   1.997  0.04577 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The variable of the debt to income ratio, delinquent in the last 2 years and number of derogatory public records have been removed from the model. All of them were statistically not significant in model 1 (*6.2*).

The coefficient are almost the same comparing with the previous model, with some small variations.

Testing the model dividing the data (80-20) we obtain the following results:

Confusion Matrix:

Prediction\Reference	0	1
0	232	47
1	137	1500

The model obtained 1550 true positives (TP), 232 true negatives (TN), 137 false positives (FP)and 47 false negative (FN).

The AUC (Area under the curve of the Receiver Operating Characteristic) is 0.7992, meaning that the model has a high discrimination power. It's slightly lower than the one with model 1.
The Accuracy $\frac{TP+TN}{TP+TN+FP+FN}$ is equal to 0.90396, almost the same than with model 1.

The Precision $\frac{TP}{TP+FP}$ is equal to 0.916310, almost the same than with model 1.

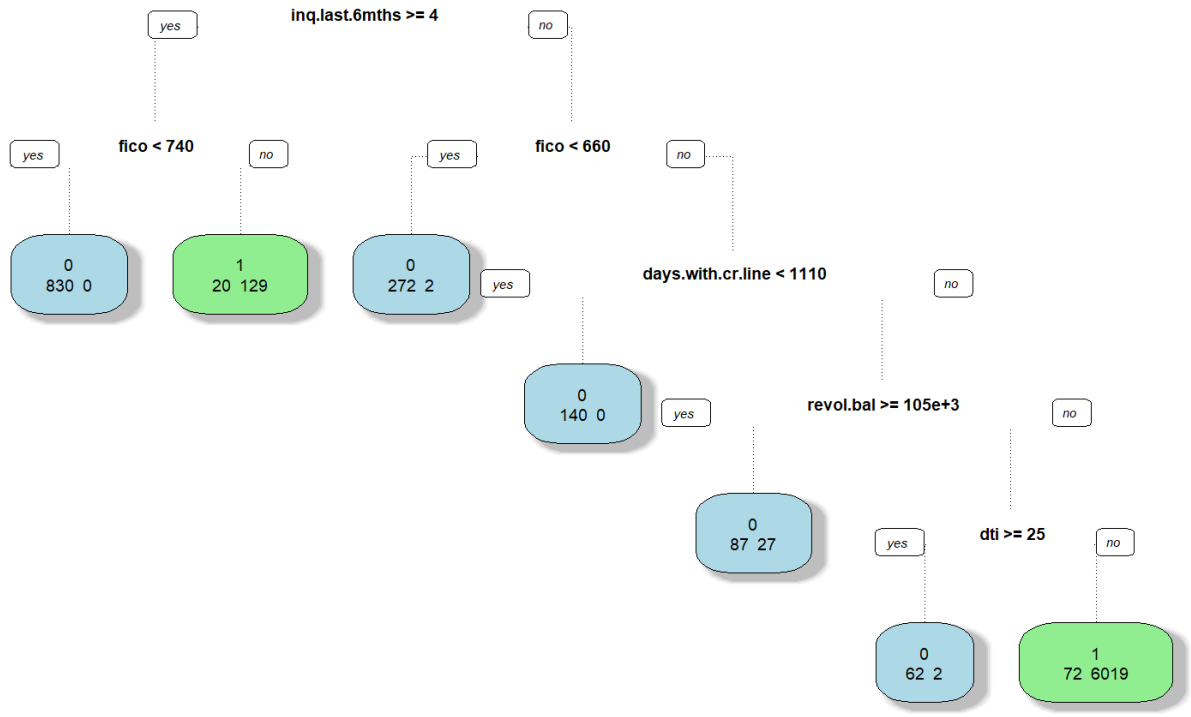
The Recall $\frac{TP}{TP+FN}$ is equal to 0.9696186, the same than with model 1.

The F1 score, a harmonic average between of Precision and Recall, is 0.94221, almost the same that in model 1.

Model 1 and Model 3 yield similar performances, but Model 3 utilizes fewer variables.Then, it could be a better option due to its reduced need for computational and storage resources.

6.5 Decision Tree

Given the presence of some determinist variables (after a certain treshold) in the dataset, a decision tree could be a good model to predict values.



The one above is a graph of the decision tree. As shown in the graph, there are 5 main splitting variables: the number of inquiries in the last 6 months, the fico credit score, the days of utilisation of the credit line, the revolving balance and the debt to income ratio. It's interesting to note that both of the deterministic variables we individuated before are now included in the decision tree.

Testing the model [80-20] we obtain:

Confusion Matrix:

Prediction\Reference	0	1
0	320	28
1	6	1562

The model obtained 1562 true positives (TP), 320 true negatives (TN), 6 false positives (FP) and 28 false negative (FN).

The Accuracy $\frac{TP+TN}{TP+TN+FP+FN}$ is equal to 0.9822547, higher than every other implemented model.

The Precision $\frac{TP}{TP+FP}$ is equal to 0.9961735, higher than every other implemented model.

The Recall $\frac{TP}{TP+FN}$ is equal to 0.9823899, higher than every other implemented model.

The F1 score, a harmonic average between of Precision and Recall, is 0.9892337, higher than every other implemented model.

The decision tree outperforms all the other model.

7. Which variables impact the decision on interest rates?

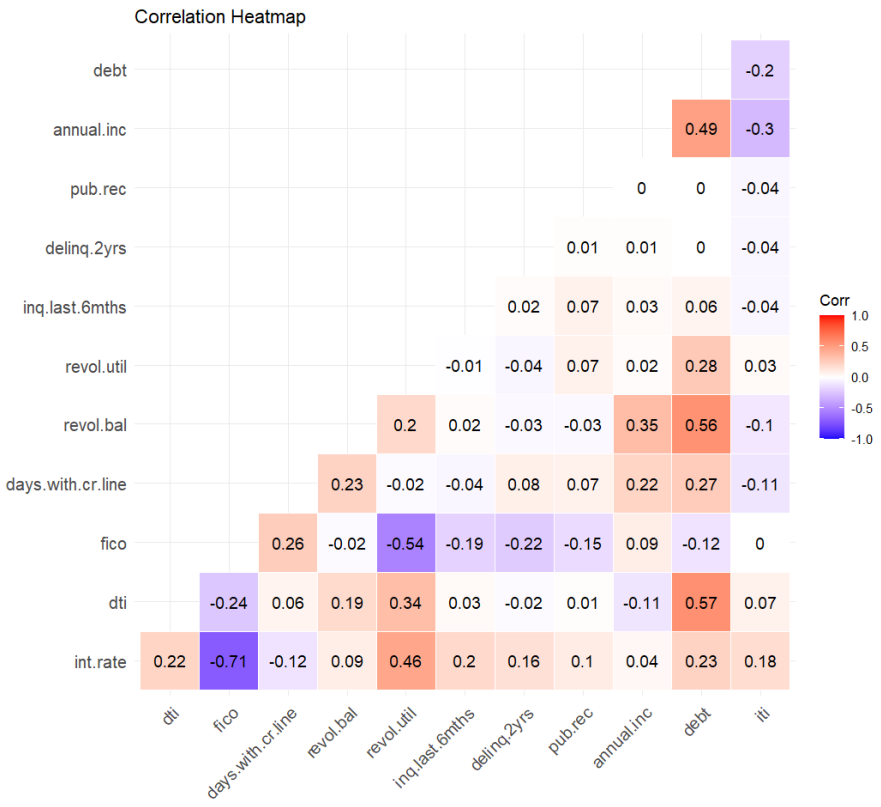
To explore this supervised techniques have been implemented.

7.1 General Linear Model

Given that the goal is trying to understand which variables make the investors choose the interest rate on the loan, the not relevant variables have been excluded by the model: the total interest, and the fact that a borrower has repaid the loan are information not available when the borrower first approach the website. This because they are decided later, when the investor already selected the potential borrower, or they are consequences of the choice of the interest rate. Also instalment has been removed, since there is the variable dti that summarise its information.

The idea is to create a general linear model with a Gamma family, assuming interest rates follow a gamma distribution.

Plotting the correlation heat map:



And computing the VIF for each variable:

Variable	Variance Inflation Factor
credit.policy	1.295343
purpose	1.021245
dti	1.927235
fico	1.450821
days.with.cr.line	1.299572
revol.bal	1.835580

revol.util	1.762344
inq.last.6mths	1.461668
delinq.2yrs	1.139414
pub.rec	1.049414
annual.inc	5.405079
debt	7.080556
iti	3.157088

We notice that annual income and debt could cause multicollinearity problems.

```
Call:
glm(formula = int.rate ~ ., family = Gamma("log"), data = data.train.2)

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                5.519e-01  4.240e-02  13.015  < 2e-16 ***
credit.policy1            -5.928e-03  5.005e-03  -1.184  0.23633
purposecredit_card        -2.266e-02  5.388e-03  -4.205 2.64e-05 ***
purposedebt_consolidation -3.690e-03  4.167e-03  -0.886  0.37582
purposeeducational        -7.892e-03  8.837e-03  -0.893  0.37189
purposehome_improvement    3.006e-02  6.801e-03   4.420 1.00e-05 ***
purposemajor_purchase      1.004e-02  7.872e-03   1.276  0.20216
purposesmall_business      1.610e-01  7.008e-03  22.967  < 2e-16 ***
dti                       -2.228e-03  3.454e-04  -6.450 1.19e-10 ***
fico                      -3.990e-03  5.839e-05 -68.324  < 2e-16 ***
days.with.cr.line          2.000e-06  6.989e-07   2.861  0.00423 **
revol.bal                 -2.942e-07  5.758e-08  -5.109 3.31e-07 ***
revol.util                 7.374e-04  7.055e-05  10.451  < 2e-16 ***
inq.last.6mths             7.192e-03  8.665e-04   8.300  < 2e-16 ***
delinq.2yrs                7.986e-03  2.986e-03   2.675  0.00749 **
pub.rec                   -3.832e-03  5.999e-03  -0.639  0.52301
annual.inc                 2.241e-07  3.487e-08   6.425 1.40e-10 ***
debt                       5.660e-08  4.281e-09  13.222  < 2e-16 ***
iti                        1.212e+00  3.861e-02  31.399  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The most statistically significant coefficient are purpose equal to “credit card” (negative coefficient), purpose equal to “home improvement” (positive coefficient), purpose equal to “small business” (positive coefficient), debt to income ratio (negative coefficient), fico credit score (coefficient), revolving balance (negative coefficient), revolving utilisation rate (positive coefficient), inquiries in the last 6 months (positive coefficient), annual income (positive coefficient), and total debt (positive coefficient).

We can also notice than the variable created utilising some dataset column (income to installment ratio) is highly statistically significant, with a positive coefficient. Same for debt.

Testing the model, we obtain the following results:

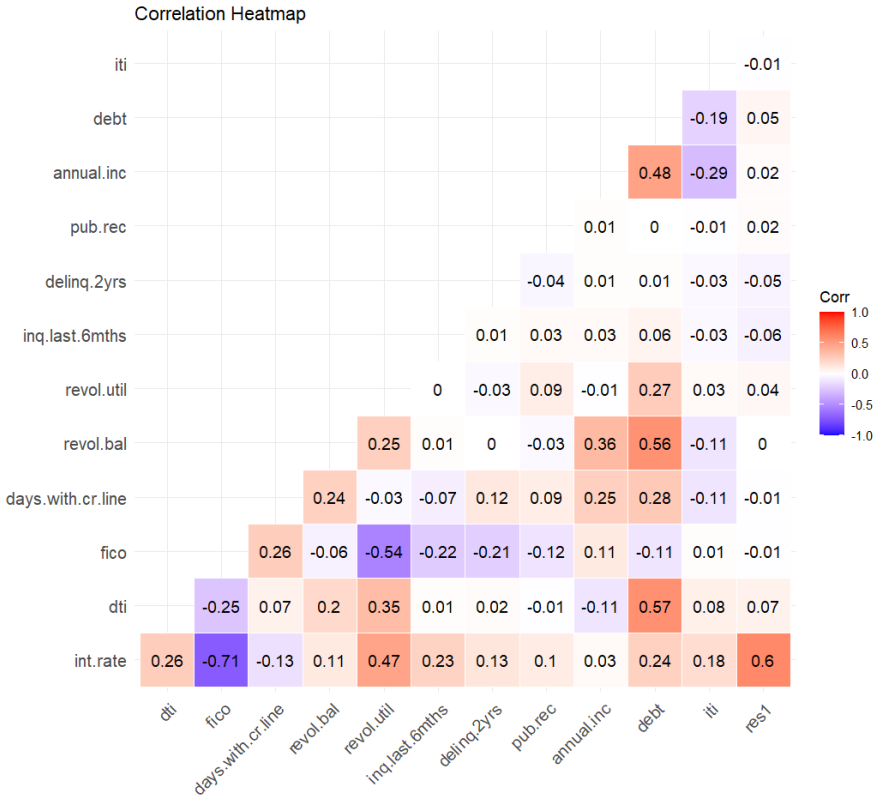
The R^2 score is 0.6093039, meaning that the model explain the 60,93% of the variability of the interest rate.

Given that the model is made with General Linear Regression, we can't use adjusted R^2 to compare the models. To solve this problem, McFadden Pseudo R^2 has been computed:

$$R^2_{McFadden} = 1 - \frac{\log likelihood\ model}{loglikelihood\ null\ model}$$

Which is equal to: log lik -0.2305983 (df=20)

To compute the error, having a numerical variable, the Root Mean Square Error (RMSQ) has been computed and is equal to 0.01695035.



The error (res1) is not correlated or weakly correlated with all the independent variables, meaning that they are exogenous. It's correlated with the inflation rate. This makes sense because the interest rate rate is affected by many other variables, such as the inflation rate and the general economical situation, which are not included in the dataset.

7.2 Step-wise Selection

We apply the step-wise selection to verify if we can remove from the model some not relevant variable:

```
Call:
glm(formula = int.rate ~ purpose + dti + fico + days.with.cr.line +
    revol.bal + revol.util + inq.last.6mths + delinq.2yrs + annual.inc +
    debt + iti, family = Gamma("log"), data = data.train.2)

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                5.544e-01  4.142e-02  13.385  < 2e-16 ***
purposecredit_card        -2.289e-02  5.386e-03  -4.250 2.16e-05 ***
purposedebt_consolidation -4.010e-03  4.160e-03  -0.964  0.33507
purposeeducational        -7.834e-03  8.839e-03  -0.886  0.37548
purposehome_improvement    2.980e-02  6.800e-03   4.383 1.19e-05 ***
purposemajor_purchase      9.781e-03  7.871e-03   1.243  0.21399
purposesmall_business      1.606e-01  7.003e-03  22.927  < 2e-16 ***
dti                       -2.214e-03  3.455e-04  -6.408 1.57e-10 ***
fico                      -4.001e-03  5.577e-05 -71.733  < 2e-16 ***
days.with.cr.line          1.916e-06  6.930e-07   2.764  0.00571 **
revol.bal                 -2.772e-07  5.598e-08  -4.952 7.51e-07 ***
revol.util                 7.326e-04  7.036e-05  10.411  < 2e-16 ***
inq.last.6mths             7.713e-03  7.434e-04  10.375  < 2e-16 ***
delinq.2yrs                8.120e-03  2.983e-03   2.722  0.00651 **
annual.inc                 2.228e-07  3.487e-08   6.389 1.76e-10 ***
debt                       5.649e-08  4.279e-09  13.200  < 2e-16 ***
iti                        1.212e+00  3.856e-02  31.432  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

pub.rec and credit.policy have been removed. None of them where statistically significant. It’s important to note that debt and annual income were variables with an high VIF. They are still in the model but now they have a lower VIF (respectively 3.663582 and 1.929935). All the other variables have a VIT lower than 5.

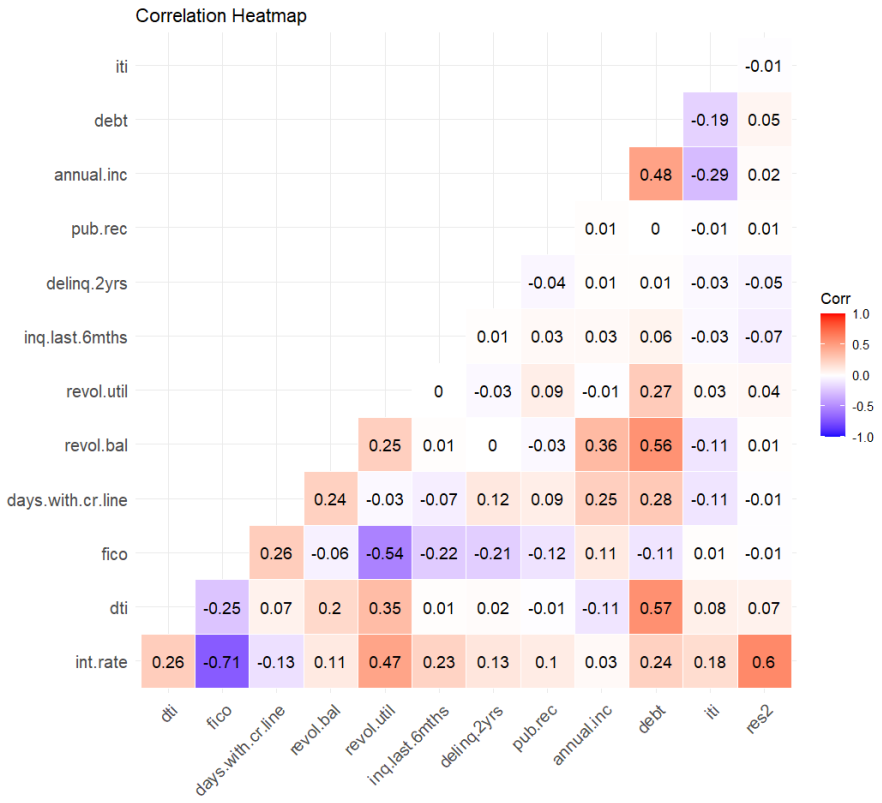
It’s important to notice that both the variables created (debt and iti) are highly statistically significant, with a positive coefficient.

Testing the model we obtain the following results:

The R^2 score is 0.6088126, meaning that the model explain the 60,88% of the variability of the interest rate.

McFadden Pseudo R^2 is equal to: log lik -0.2305437 (df=18), higher and with a smaller number of variables, which makes it better.

The Root Mean Square Error (RMSQ) has been computed and is equal to 0.0169607, which is similar to the previous model.



The results in the heatmap are almost identical to the previous model: the error is highly correlated with the int.rate and independent or weakly correlated with the independent variables.

8. Can borrowers be clustered based on the available data before investors decide to lend them money?

To understand this, clustering methods have been used.

8.1 Data Encoding and Manipulation

To prepare the dataset for clustering, the categorical variable "purpose," which contains seven different values, has been transformed into seven binary columns. Each new column corresponds to one of the original purpose values. If a record's purpose matches the value of the specific column, it is marked as 1; otherwise, it is marked as 0. This binary representation enables computing distances between records for clustering purposes.

The new column are the following:

- all_other;
- credit_card;
- debt_consolidation;
- educational;
- home_improvement;
- major_purchase;
- small_business.

Then only the variables available to the website LendigClub.com when the potential borrower registers has been selected, in order to group the clients before they are selected by the investors.

dti, fico, days.with.cr.line, revol.bal, revol.util, inq.last.6mths, delinq.2yrs, pub.rec, annual.inc, all_other, credit_card, debt_consolidation, educational, home_improvement, major_purchase and small_business have been selected.

credit.policy, int.rate, installment, not.fully.paid and debt have been excluded.

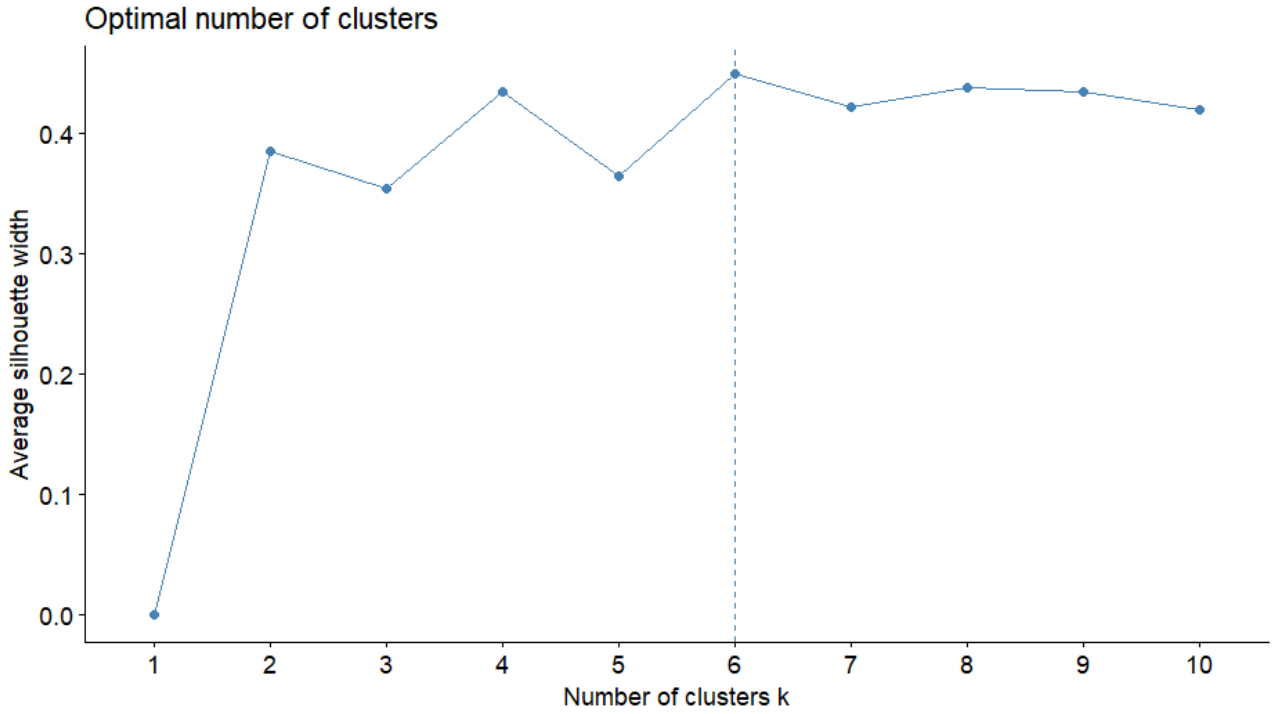
Furthermore, the data has been scaled, given the columns have different scale and unit of measure. The categorical variables has been excluded from the scaling, because otherwise they would lose their interpretability.

Then the data has been reducted, both in 3 and 2 dimensions to compare the results, using T-SNE technique, suitable for both numerical and categorical variables.

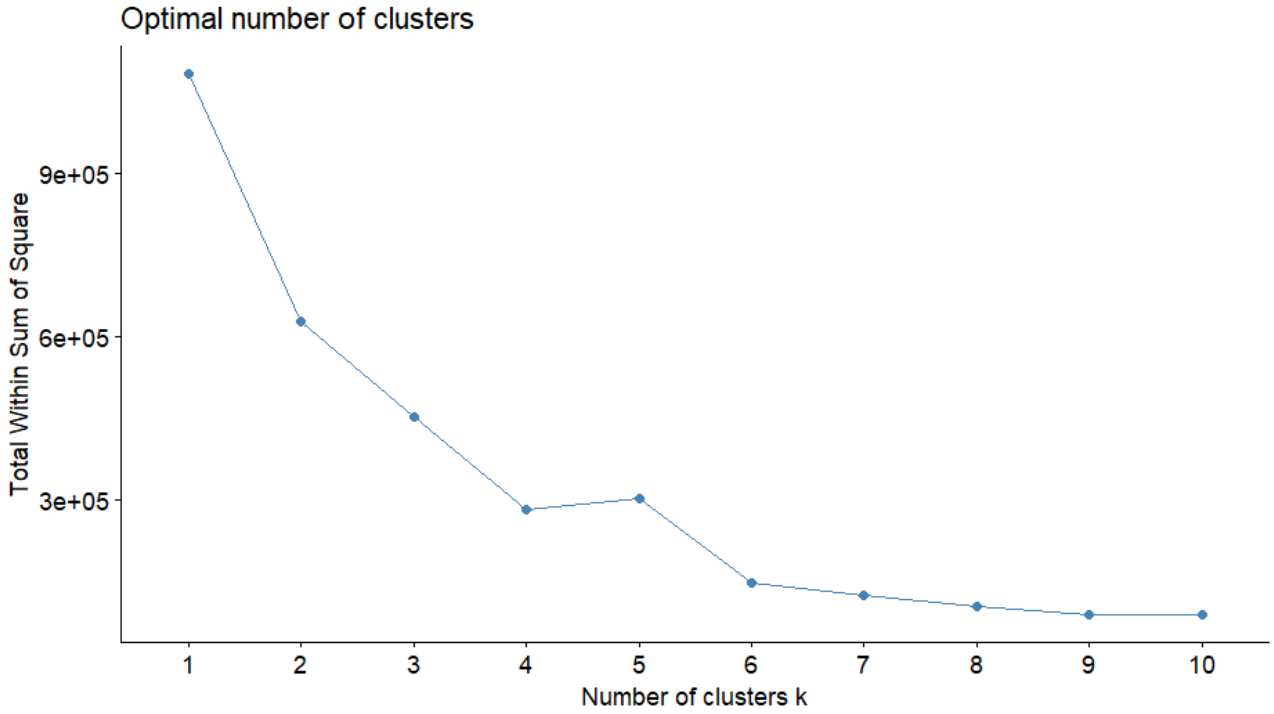
Finally, the distance matrix has been computed using Gower, a method that can be used both for categorical and numerical variables.

8.2 Clustering (2 dimensions)

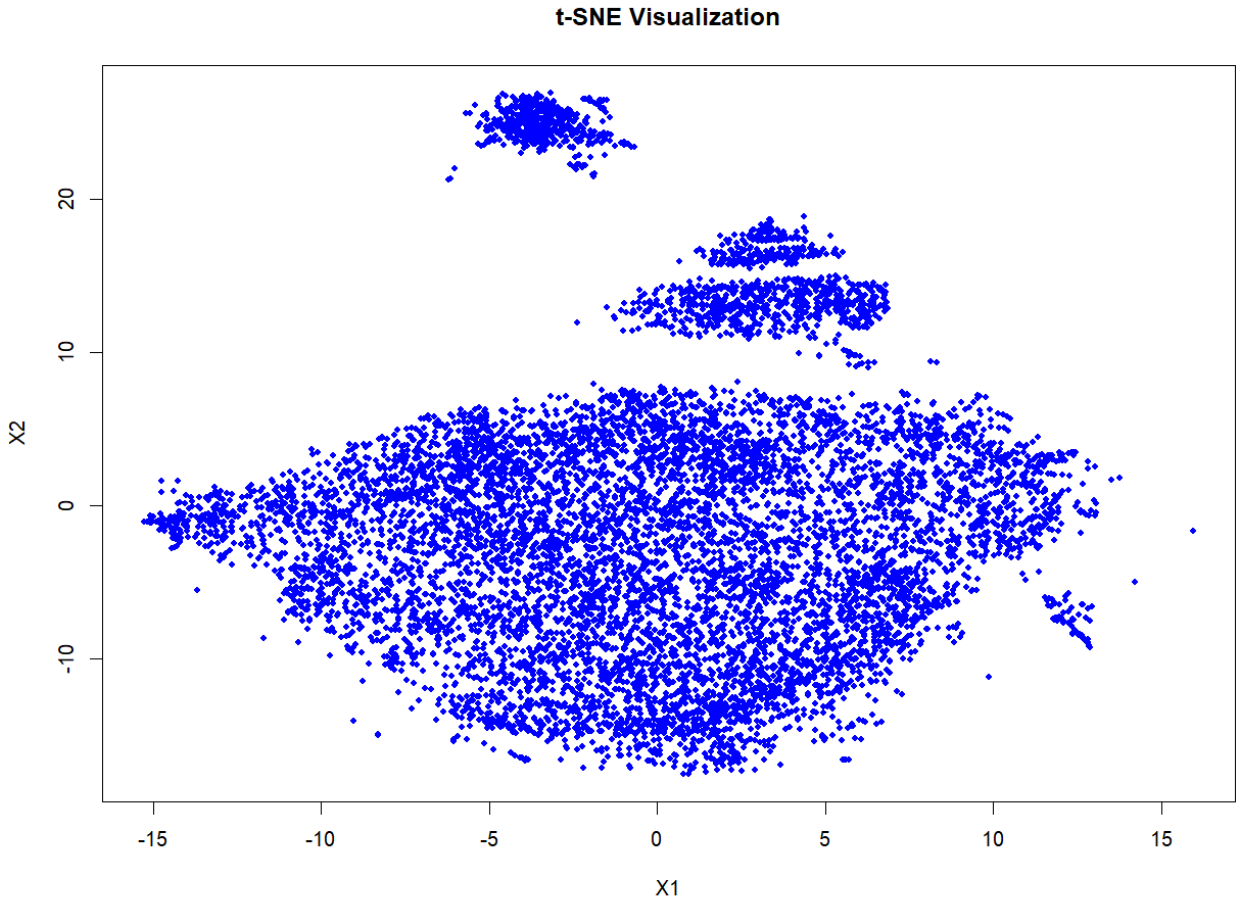
On of the first step is understanding what is the maximum amount of clusters for the goal.



The maximization of the Average Silhouette Width metod suggests that the optimal amount of clusters could be 4 or 6.



The elbow method suggests 6.



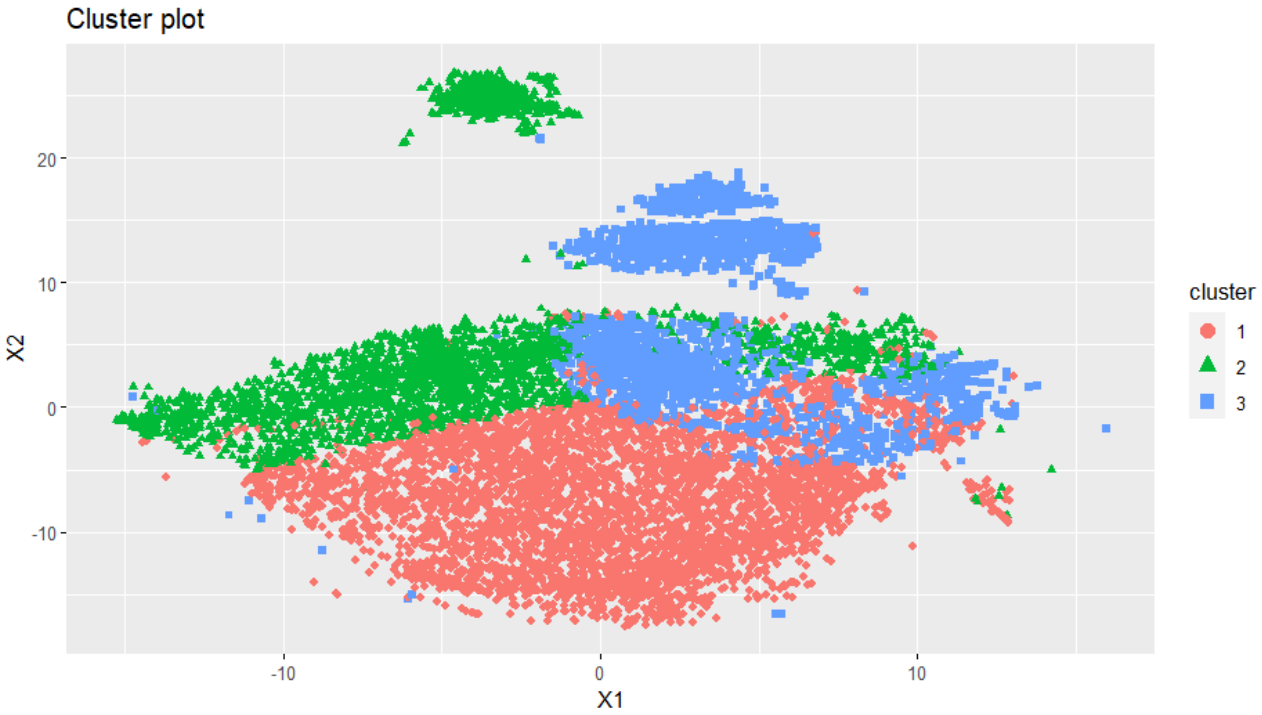
The plot of the reducted records show 3 or 4 block of points. Choosing 4 clusters could give more significant cluster, but the block between 10-20 X2 would be divided into two, and too small clusters not always are useful and they could also be expensive. That's why graphically 3 clusters seems the best choice.

Running ECLUST, an R function that suggests the best number of clusters this was the output:

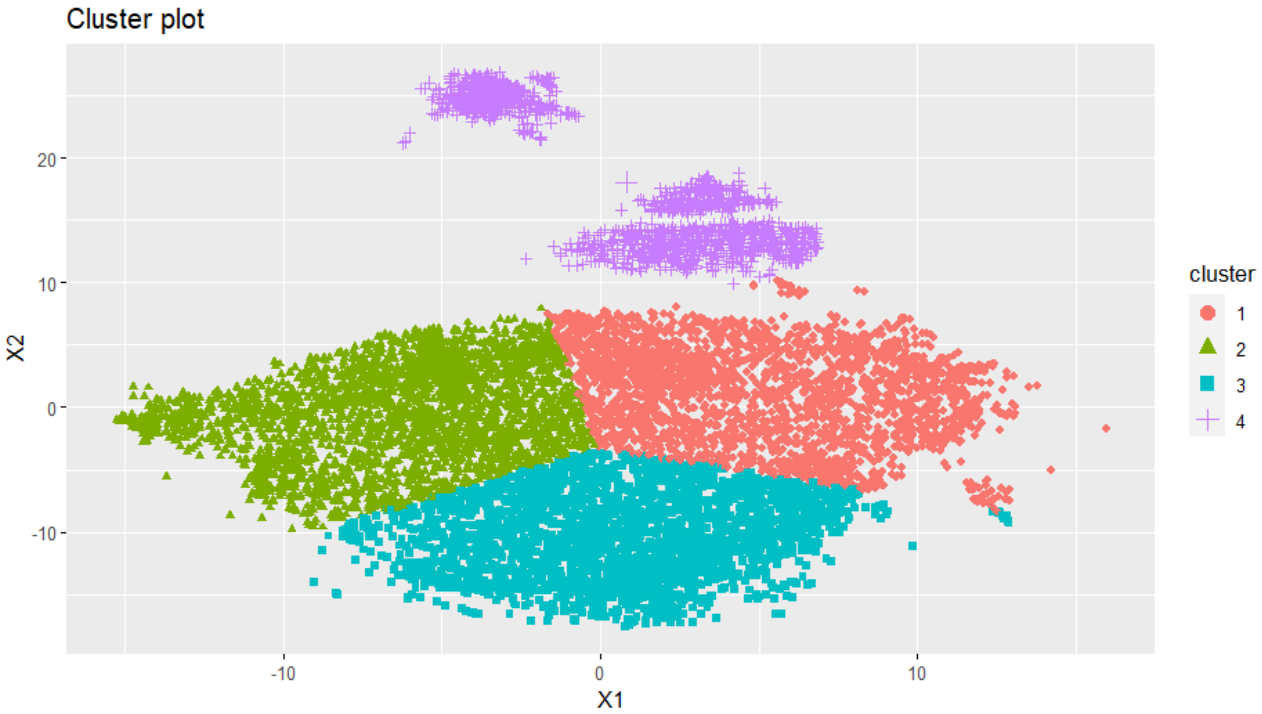
```
# ******************************************************************
#  * Among all indices:
#  * 2 proposed 2 as the best number of clusters
# * 8 proposed 3 as the best number of clusters
# * 6 proposed 4 as the best number of clusters
# * 7 proposed 6 as the best number of clusters
# ***** Conclusion *****
#  * According to the majority rule, the best number of clusters is  3
# ******************************************************************
```

3 cluster is the best number according to majority rule, but 4 and 6 clusters obtained good results as well.

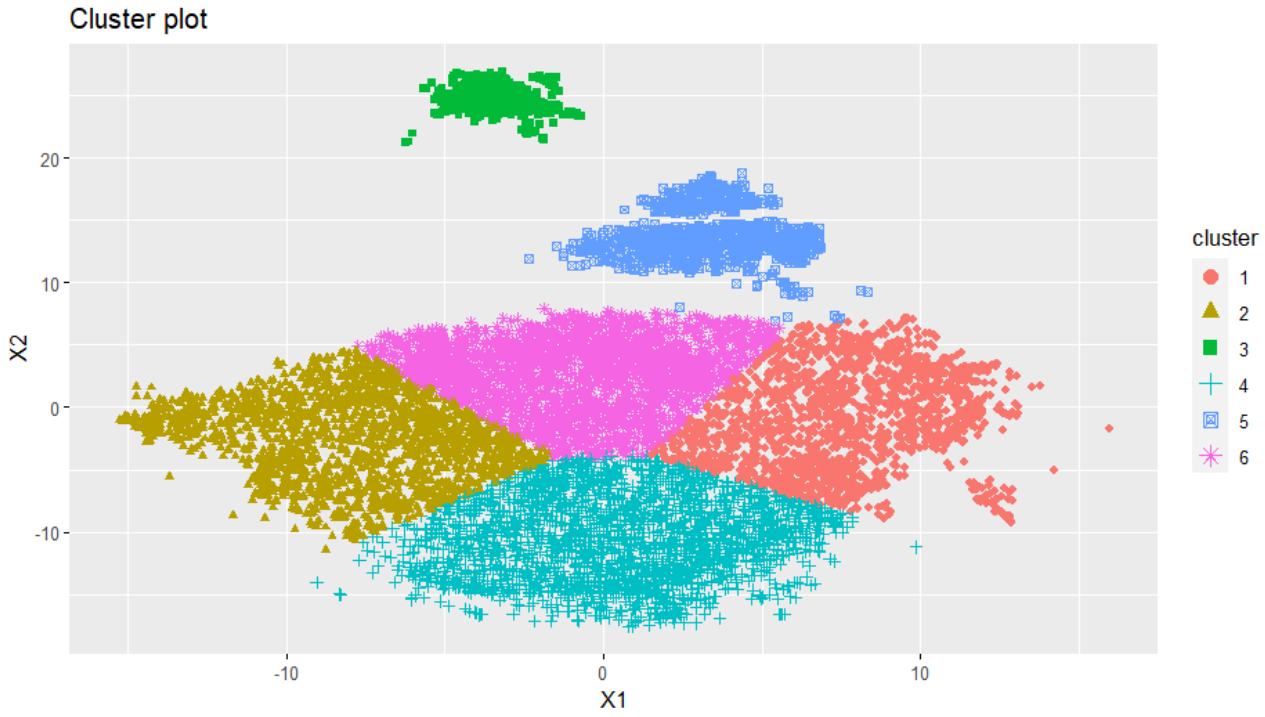
Let’s try all of the clusters using K-Means (given that now we have all numeric values):



With 3 clusters, there is too much overlapping, also the clusters don’t consider the shape of the blocks.

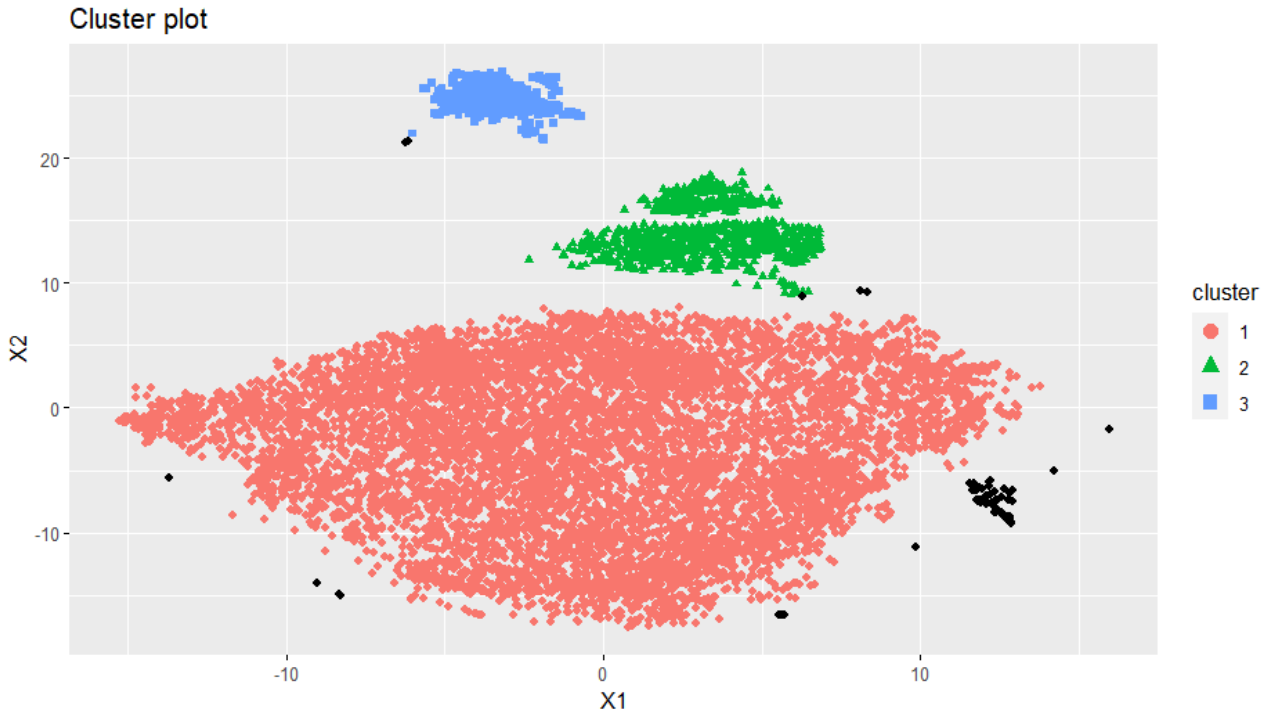


With 4 clusters we don’t have overlapping, but there still is have the shape problem.



Same problem with 6 clusters.

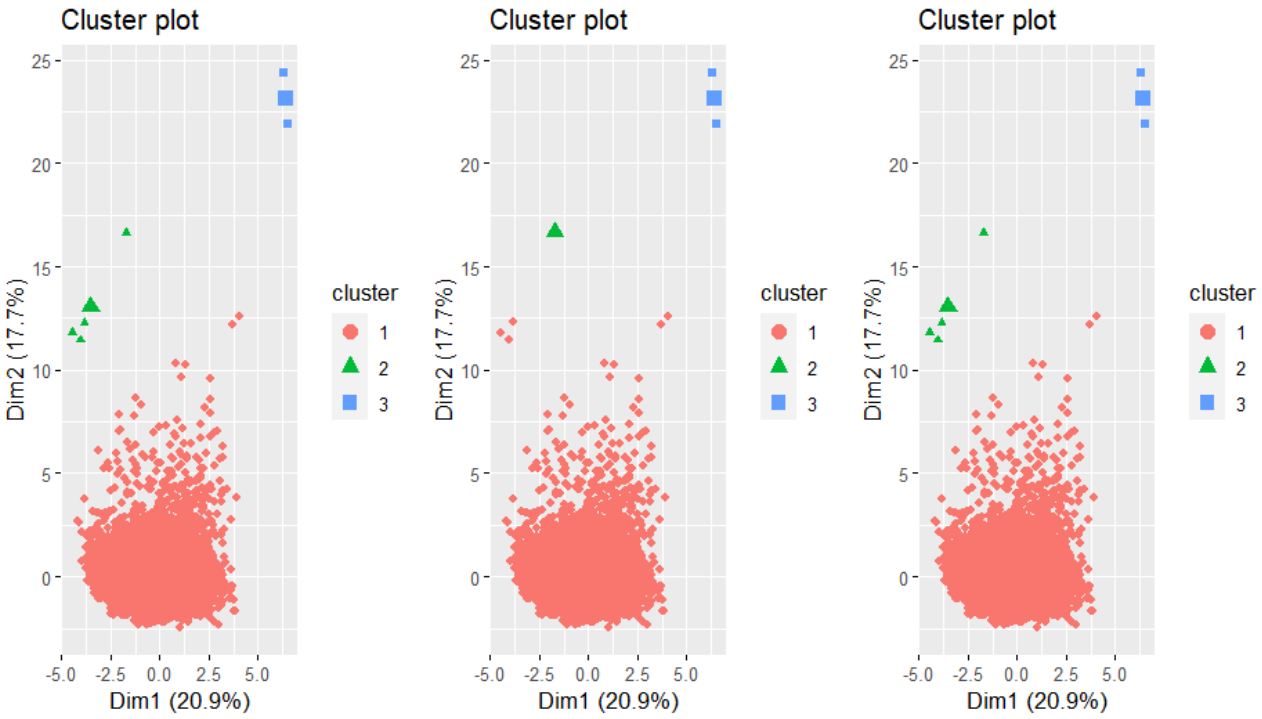
Using DBSCAN, a clustering thechnic that consider the shape of the dots distribution, the results are better:



The black dots are records that were too far from a cluster to be consider part of it and, at the same time, not enough close dots to be considered a separate cluster. This happened because a minimum amount of 84 dots (around 9% of the dataset's records) per cluster has been defined.

The cluster are unbalanced: cluster 1 is way bigger than the others.

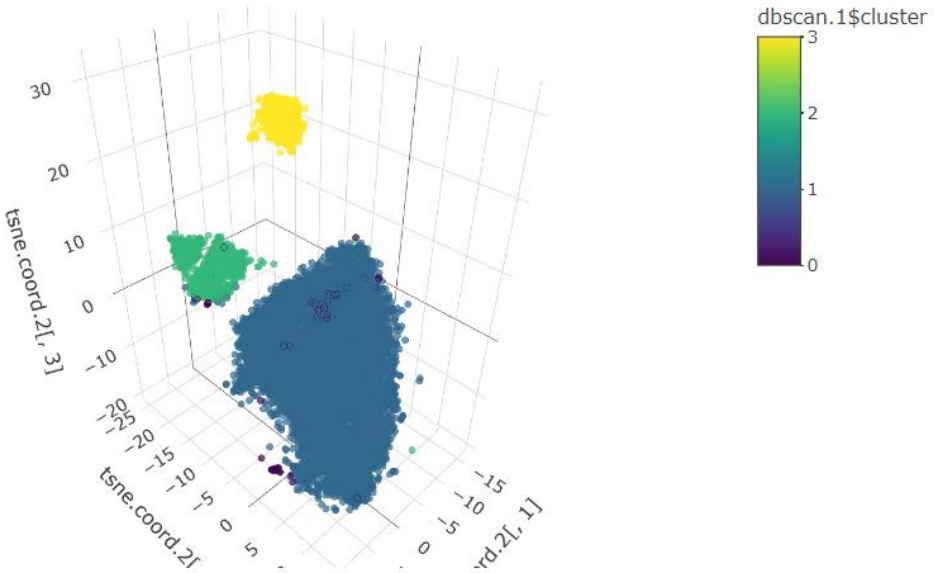
Hierarchical clustering has been tested as well:



From left to right: Complete Linkage Method, Single Linkage Method and Average Linkage Method.

Almost all the records have been assigned to cluster 1. The result is not optimal.

The following are the results with 3 dimensions data reduction combined with DBSCAN:



Click here for the 3d model

Even though we obtained more information with the 3 axes, comparing the clustering classification between two and three dimensions, the results are almost identical.

That's why the best clusters are the ones obtained with 2 dimensions data reduction and DBSCAN, because they can obtain the same result with less data.

9. Are the clusters meaningful, and what are the differences between them?

To answer this question data visualization techniques have been used.

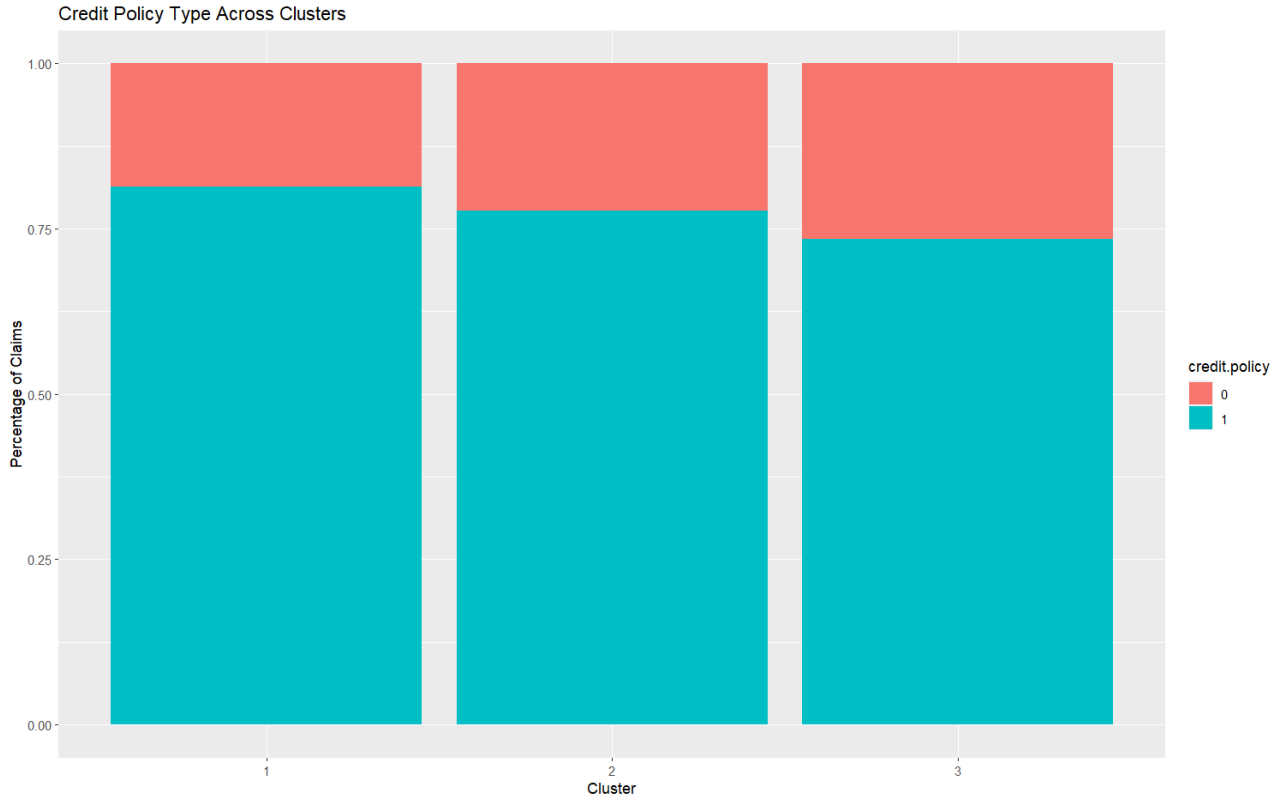
9.1 Clustering filtering

Before starting the analysis, a decision on how to deal with records without clusters had to be made. They are 90 records, representing less than 1% of the whole dataset.

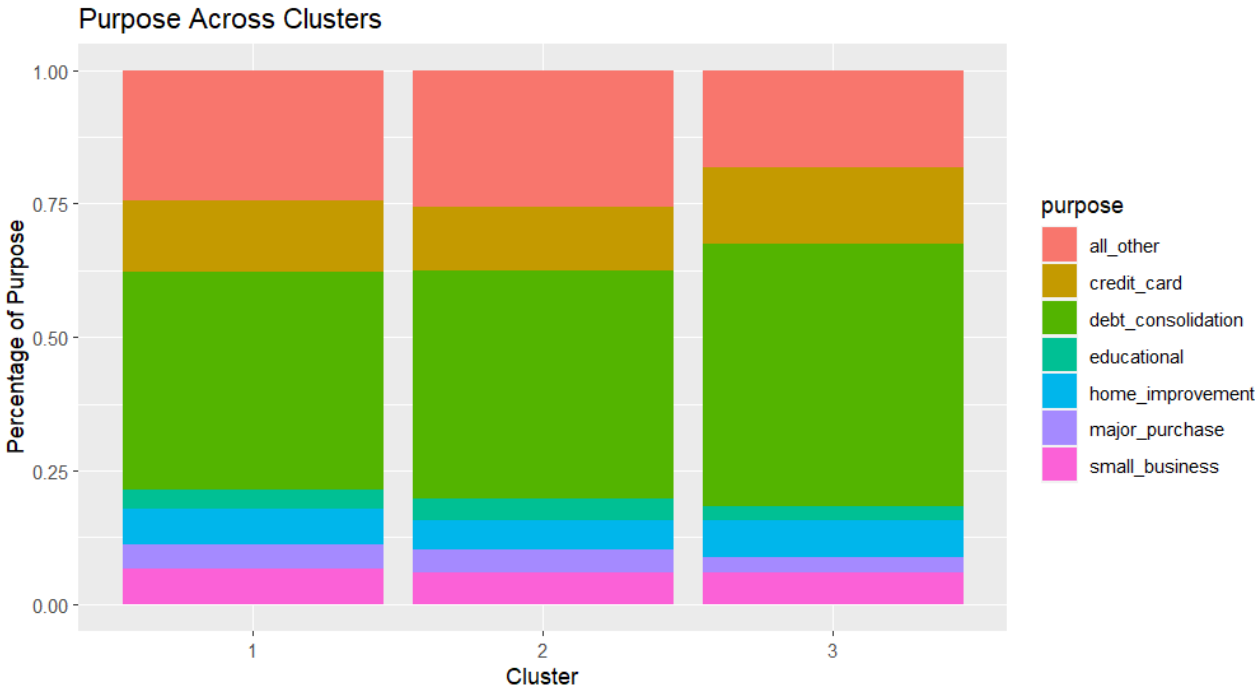
Given that the objective was to form clusters, and that the black dots present many differences between each other, analysing them in group would have been of little significance. Analyse them singulary or in smaller groups could be a choise, but in our case it would be too expensive and time consuming, given our goal. The decision is to exclude them from the analysis, filtering them out.

9.2 Data Visualisation

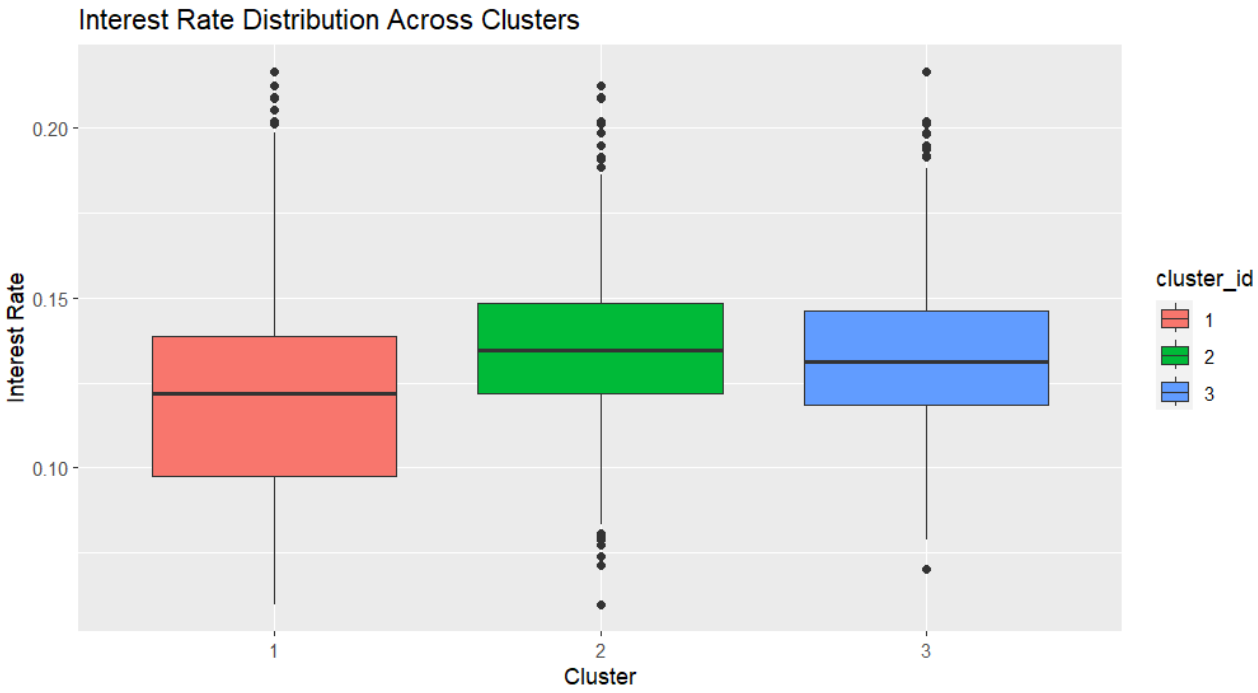
The data for each cluster has been compared to see if there are any significant differences:



There is not much difference between clusters in terms of respecting the company's credit policy. It looks slightly decreasing from cluster 1 to cluster 3.

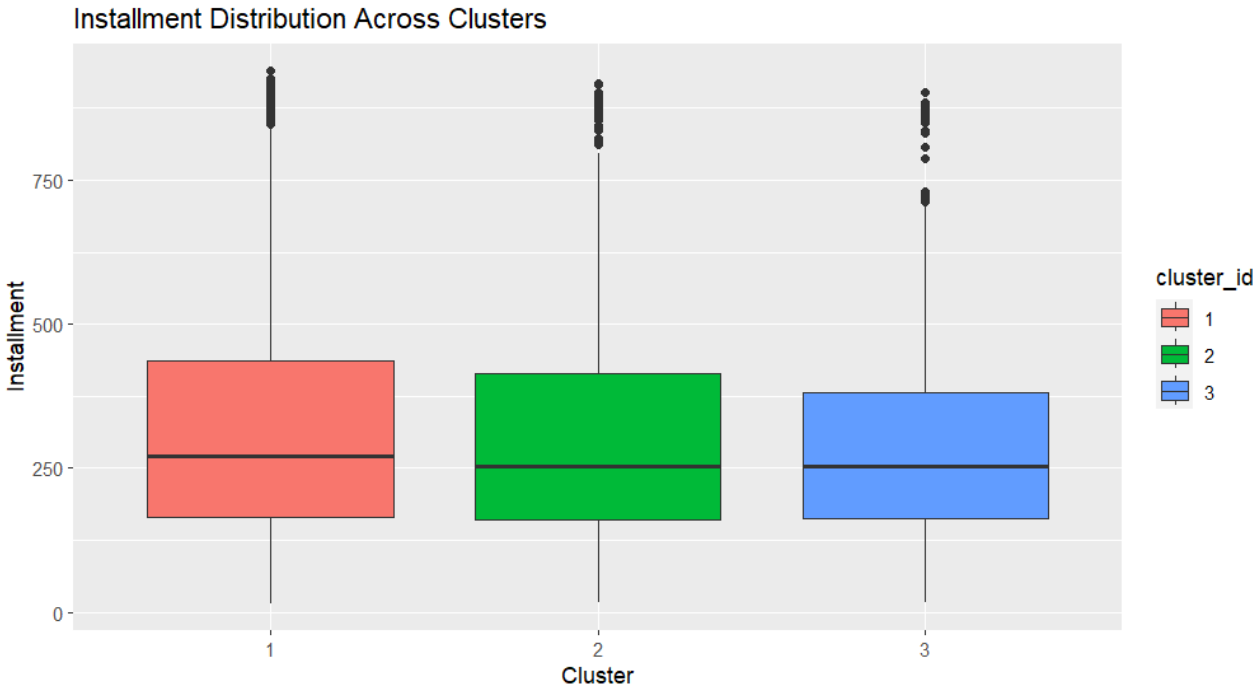


The clusters are not divided based on purpose of the loan.



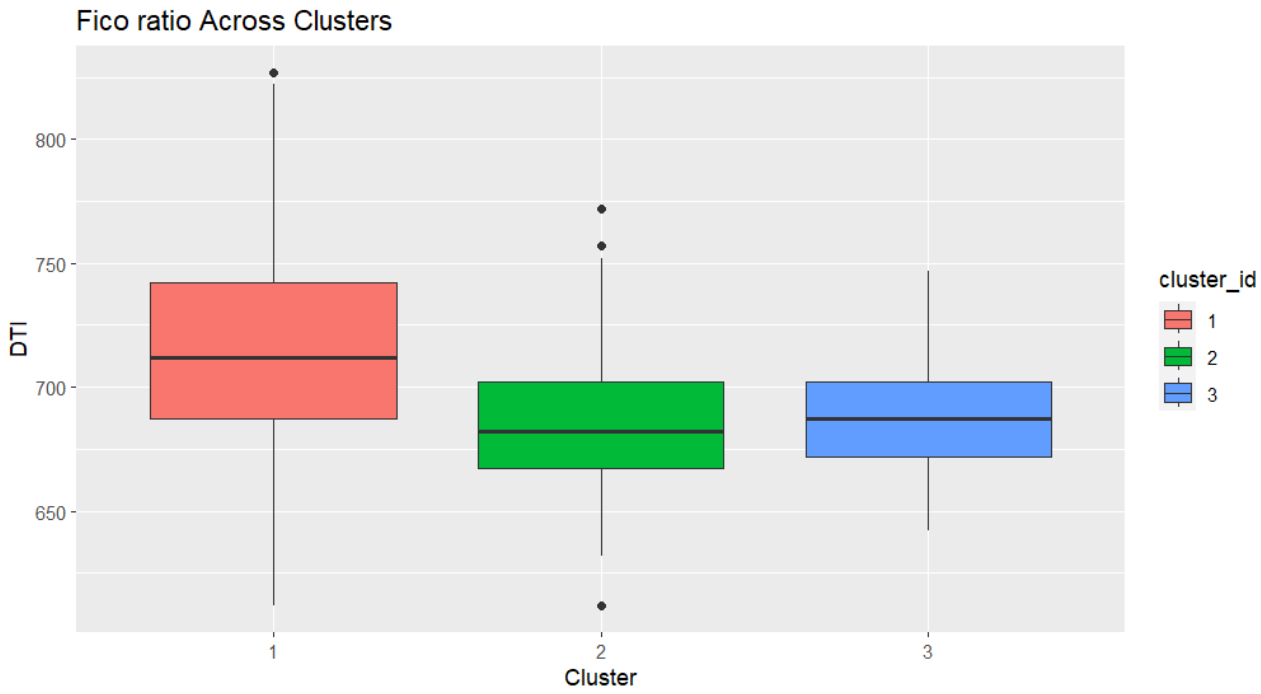
The cluster 1 has the lowest median interest rate, followed by cluster 3 and cluster 2.

We can also notice that cluster 1 has the highest within variance, which makes sense given the graph that shows it as the biggest cluster in terms of record and dimension.

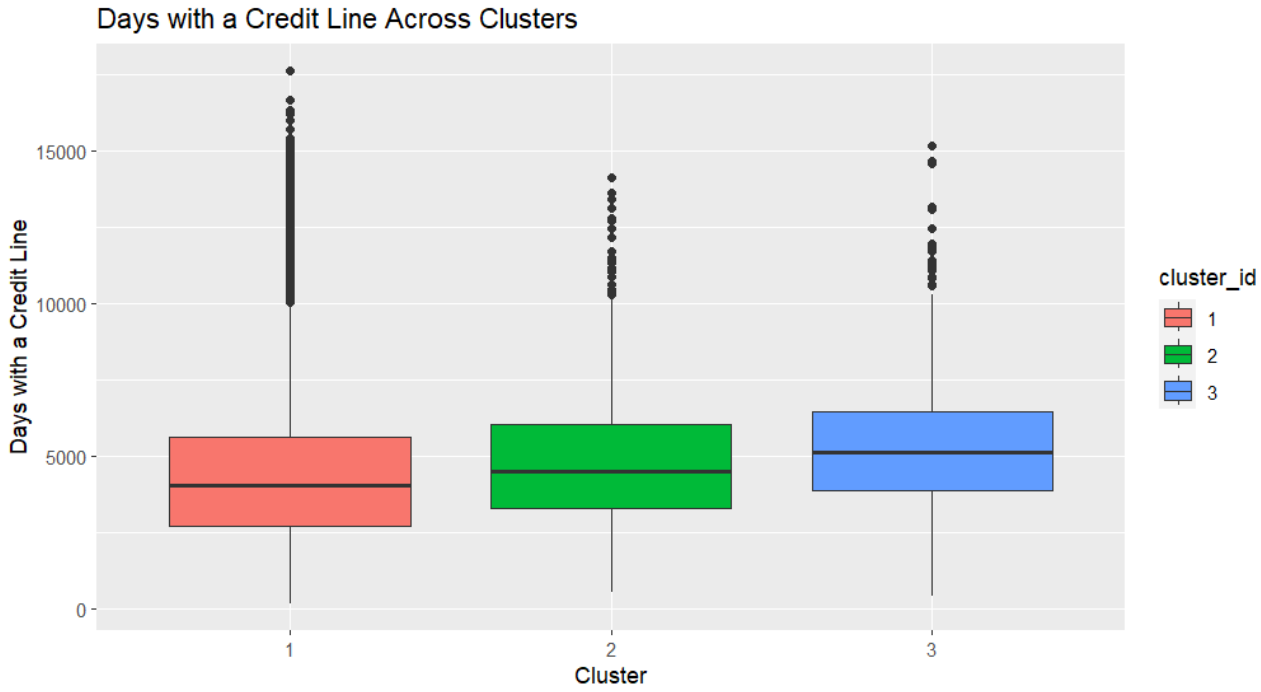


There is not much difference in the median installment across clusters, but cluster 1's is slightly higher.

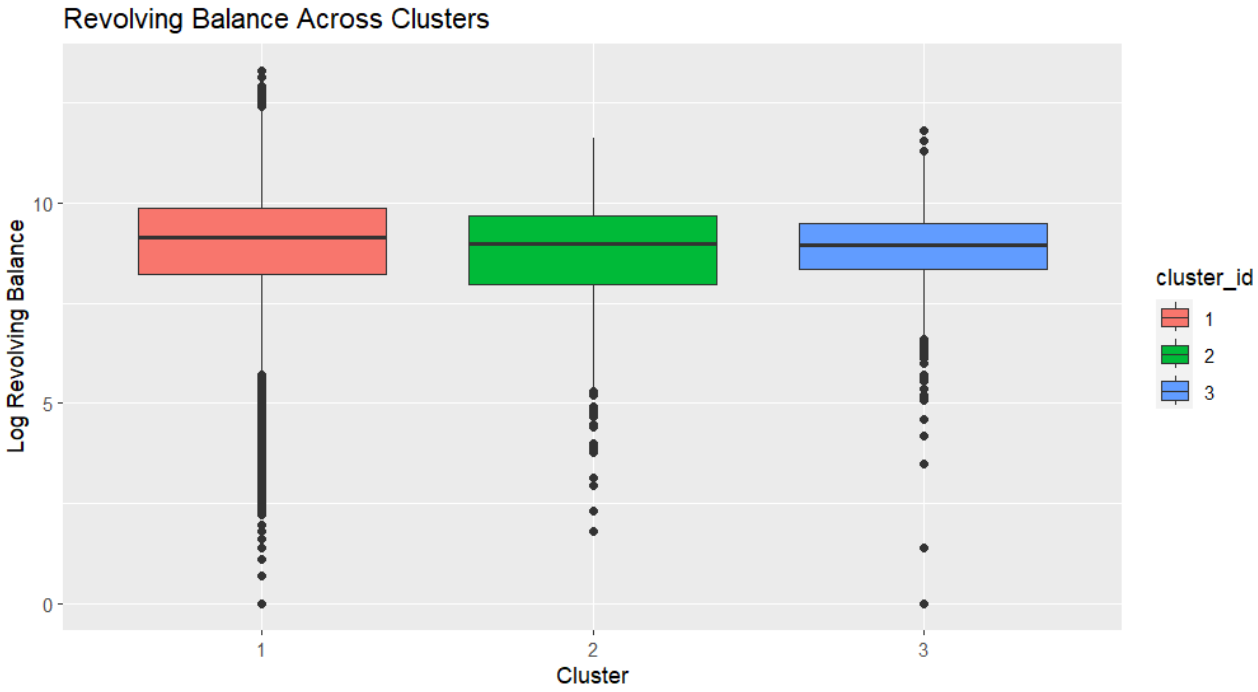
Similar debt to income ratio between clusters.



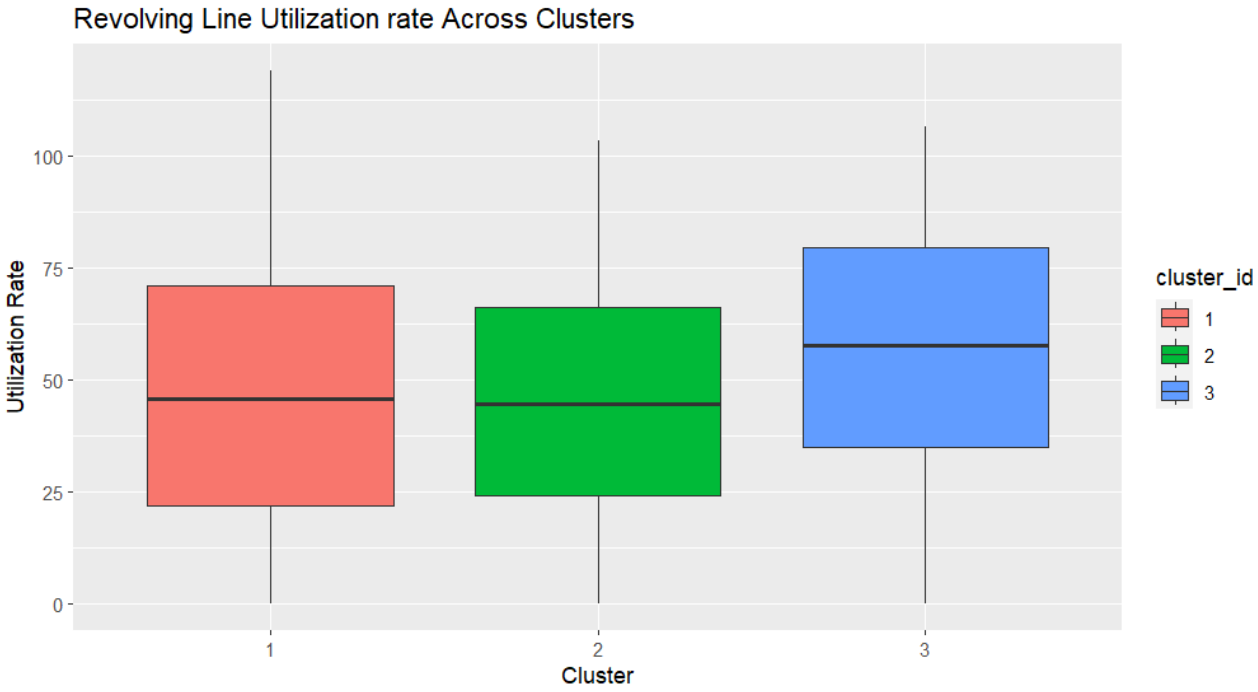
Cluster 1's records have an higher FICO credit score.



The median number of days with a credit card are increasing across clusters, from 1 to 3.

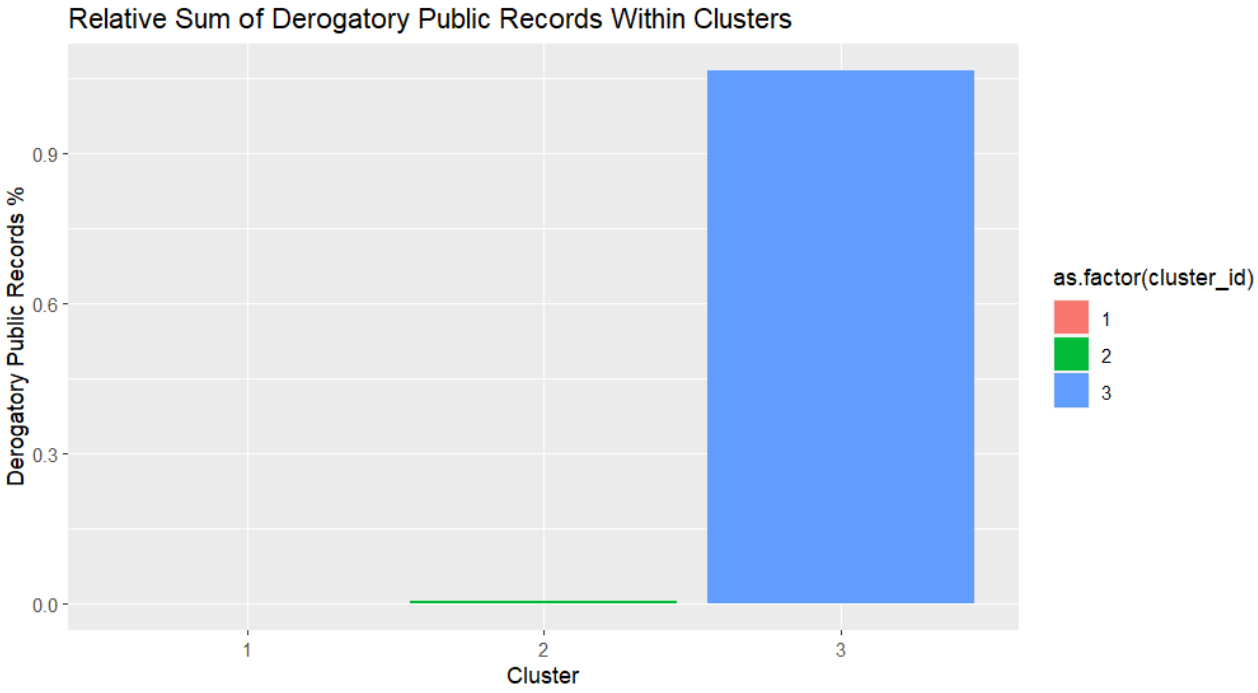


The median revolving balance is similar across clusters, with cluster 3 having a lower within variance.



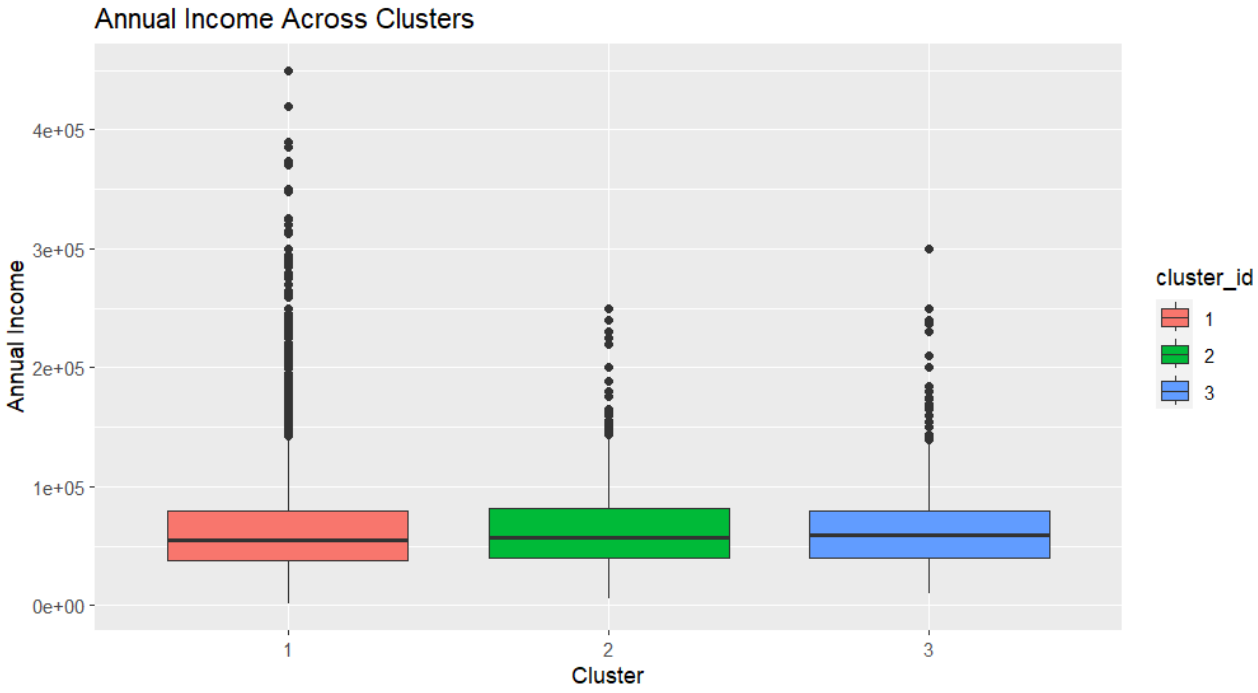
Cluster 3 have a higher median utilisation rate.

The ratio between the number of inquiries in the last 6 months and the number of records within the cluster shows that cluster 3 has the highest relative number of inquiries.

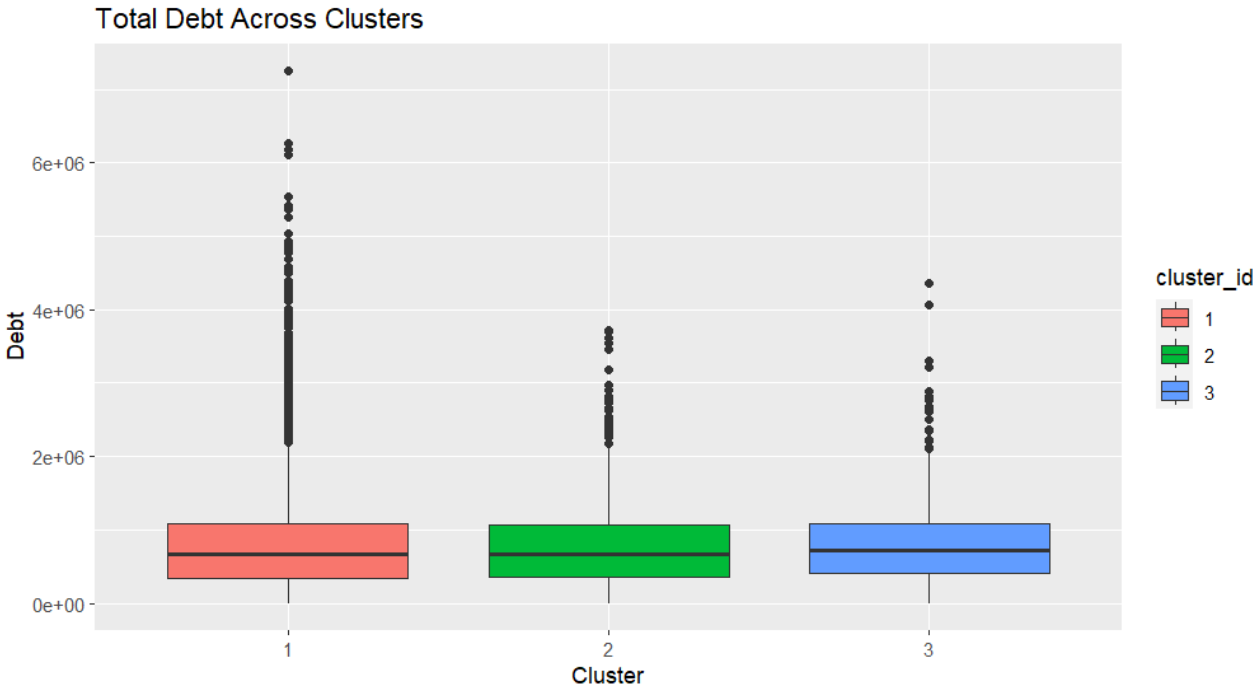


Cluster 3 represents almost all the derogatory public records.

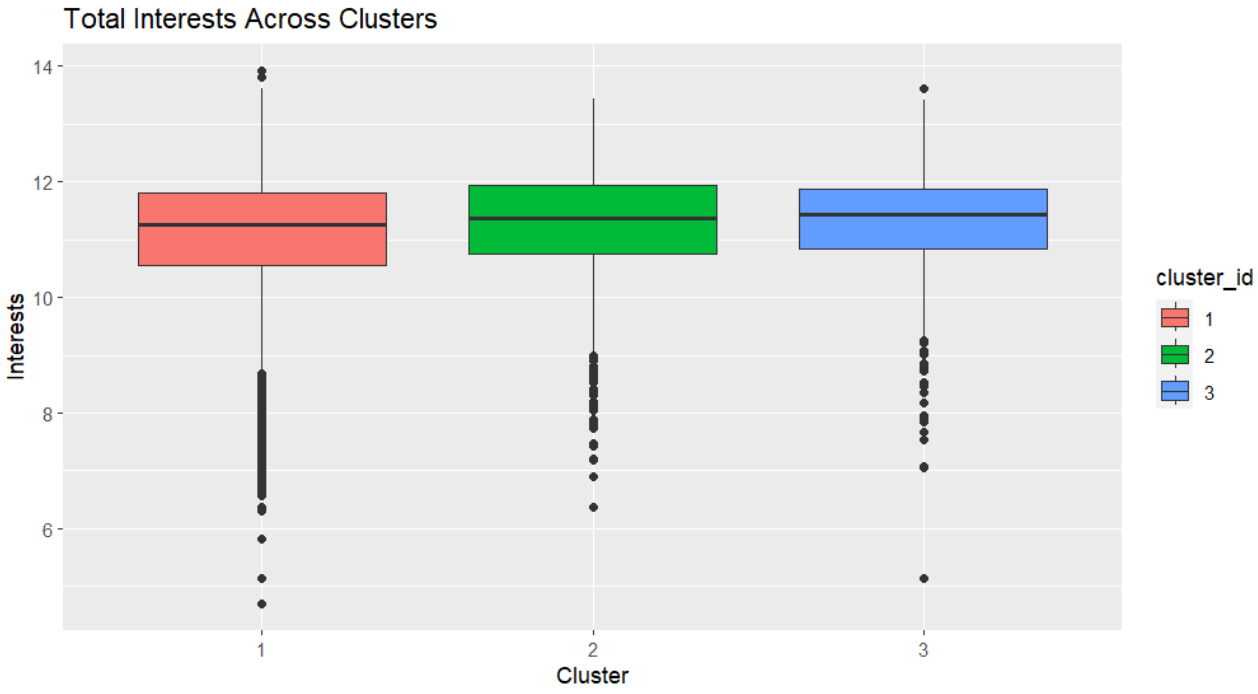
The difference in the percentage of records within cluster that have not fully paid the loan is not much different, but it's slightly increasing from cluster 1 to cluster 3.



Annual income across clusters is similar.



Debt across cluster is similar.



Total interests across clusters are similar.

10. Conclusions

10.1 What variables influence the company's decision on whether clients meet the company's credit policy or not?

According to decision tree (*6.5*), the model with higher prediction power, the variables that affect the most if the company is considering a client respecting the company policy are inquiries in the last 6 months, fico credit score, days with credit line, revolving balance and debt to income ratio.

Analysing the independent variables included in the step-wise general linear regression (*6.4*), we can see that the most statistically relevant variables are the purpose when is equal to debt consolidation (positive coefficient), the fico credit score (positive coefficient), days with credit line (positive coefficient), revolving balance (negative coefficient), credit card utilization rate (positive coefficient), inquiries in the last 6 months (positive coefficient), annual income (positive coefficient) and debt (positive coefficient.)

We can see that the variables that effect the most the odds of being assigned to credit.policy equal to 1 are the purpose of the loan, the credit score and the inquiries in the last 6 months (which is also a deterministic variable after a certain threshold). These mean that the company rely more on these variables when determining if a client is reliable or not.

The prediction results using the model are very good, almost without an error in the case of the decision tree and a low error with the other models. This could mean that the variables utilised by LendingClub.com are the same we have in the dataset, and that’s why the result are so good.

On the other hand, predicting something that is assigned by the website has not a real practical application.

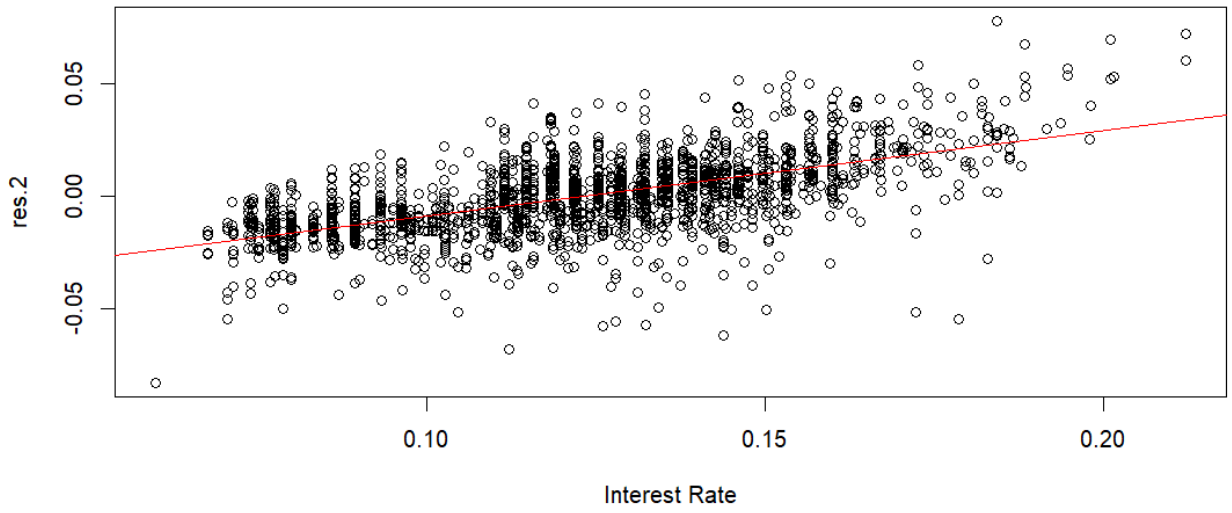
10.2 Which variables impact the decision on interest rates?

According to the generalised linear model obtained with step-wise selection (*7.2*), the most statistically significant variables to assign interest rate are purpose when is equal to credit card payment (negative coefficient), purpose when is equal to house improvement (positive coefficient), purpose when is equal to small business (positive coefficient), debt to income ratio (negative coefficient), fico credit score (negative coefficient), revolving balance (negative coefficient), credit card utilisation rate (positive coefficient), inquiries in the last six months (potive coefficient), annual income (positive coefficient), debt (positive coefficient) and installment to income ratio (positive coefficient).

The most relevant, with higher coefficient, is installment to income ratio, the variable created for the analysis.

It’s very important to notice that credit.policy was excluded from the model, because it wasn’t statistically significant. This could mean that the investors don’t take in consideration the suggestion of the company when they choose the borrower.

Also in both the models created, the interest rate was highly correlated with the error. There are probably some variables that are not included in the dataset that affect the decision on interest rate, such inflation rate, the predicted inflation rate and the economic growth rate, variables that the credit institutes and investors usually consider when they are negotiating the interest rate. This could be also noticed by the R^2 score, that states that the model (*7.2*) explains the 60.88% of the interest rate variance. Another reason why error could be hihgly correlated with the interest rate is that for higher values of interest rate we have less records. Heteroskedasticity has been confirmed by checking the scatterplot between error and interest:



The error of the model, given the RMSE, is 0.01695035. This mean the predicted interest rates differ from the actual interest rates by about 1.695035%.

Given the range of interest rates (0.1564), the error seems reasonable.

Also comparing it with the median (0.1221), the error seems reasonable.

10.3 Can borrowers be grouped (clustered) based on the available data before investors decide to lend them money? Are these clusters meaningful, and what are the significant differences between them?

The optimal number of clusters is 3.

It seems that in cluster 1 there are the more reliable clients: this because the median interest rate is lower than the others, the median fico credit score is higher and they also have slightly higher percentage of borrowers respecting company policy.

Clusters 2 and 3 are similar, but, according to the data, cluster 3 have less reliable clients. This happens because in cluster 3 we almost have all the derogatory public records registered in the dataset, there is the highest percentage of clients not having fully paid the loan, the highest relative number of inquiries in the last 6 months, the highest utilisation rate of the credit card. In these variables, cluster 2 and cluster 1 scored similarly, meaning that the main difference was made by credit score and interest rate difference.