Problem Statement EDA Analysis

Riyas KP

6/24/2019

Problem Statement

You work for a **consumer finance company** which specializes in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. **Two types** of risks are associated with the bank's decision:

- If the applicant is likely **to repay the loan**, then not approving the loan results in a **loss of business to the company**
- If the applicant is **not likely to repay the loan**, i.e. he/she is likely to default, then approving the loan may lead to a **financial loss** for the company

The data given below contains the information about past loan applicants and whether they 'defaulted' or not. The aim is to identify patterns which indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

In this case study, you will use EDA to understand how **consumer attributes** and **loan attributes** influence the tendency of default.

When a person applies for a loan, there are two types of decisions that could be taken by the company:

1. **Loan accepted**: If the company approves the loan, there are 3 possible scenarios described below:

- **Fully paid**: Applicant has fully paid the loan (the principal and the interest rate)
- **Current**: Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.
- **Charged-off**: Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has **defaulted** on the loan

2. **Loan rejected**: The company had rejected the loan (because the candidate does not meet their requirements etc.). Since the loan was rejected, there is no transactional history of those applicants with the company and so this data is not available with the company (and thus in this dataset)

Business Objectives

This company is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast-online interface.

Like most other lending companies, lending loans to 'risky' applicants is the largest source of financial loss (called credit loss). The credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed. In other words, borrowers who **default** cause the largest amount of loss to the lenders. In this case, the customers labelled as 'charged-off' are the 'defaulters'.

If one is able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants using EDA is the aim of this case study.

In other words, the company wants to understand the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

Analysis

Overview of the Dataset

Browsing through the Data:

This data set contains 39,717 loans with 111 variables on each loan.

```
## [1] 39717   111
```

Here are the names of the 111 variables:

```
##   [1] "id"                          "member_id"
##   [3] "loan_amnt"                   "funded_amnt"
##   [5] "funded_amnt_inv"             "term"
##   [7] "int_rate"                    "installment"
##   [9] "grade"                       "sub_grade"
##  [11] "emp_title"                   "emp_length"
##  [13] "home_ownership"              "annual_inc"
##  [15] "verification_status"         "issue_d"
##  [17] "loan_status"                 "pymnt_plan"
##  [19] "url"                         "desc"
##  [21] "purpose"                     "title"
##  [23] "zip_code"                    "addr_state"
##  [25] "dti"                         "delinq_2yrs"
##  [27] "earliest_cr_line"            "inq_last_6mths"
##  [29] "mths_since_last_delinq"      "mths_since_last_record"
##  [31] "open_acc"                    "pub_rec"
##  [33] "revol_bal"                   "revol_util"
##  [35] "total_acc"                   "initial_list_status"
```

```
##  [37] "out_prncp"                        "out_prncp_inv"
##  [39] "total_pymnt"                      "total_pymnt_inv"
##  [41] "total_rec_prncp"                  "total_rec_int"
##  [43] "total_rec_late_fee"               "recoveries"
##  [45] "collection_recovery_fee"          "last_pymnt_d"
##  [47] "last_pymnt_amnt"                  "next_pymnt_d"
##  [49] "last_credit_pull_d"               "collections_12_mths_ex_med"
##  [51] "mths_since_last_major_derog"      "policy_code"
##  [53] "application_type"                 "annual_inc_joint"
##  [55] "dti_joint"                        "verification_status_joint"
##  [57] "acc_now_delinq"                   "tot_coll_amt"
##  [59] "tot_cur_bal"                      "open_acc_6m"
##  [61] "open_il_6m"                       "open_il_12m"
##  [63] "open_il_24m"                      "mths_since_rcnt_il"
##  [65] "total_bal_il"                     "il_util"
##  [67] "open_rv_12m"                      "open_rv_24m"
##  [69] "max_bal_bc"                       "all_util"
##  [71] "total_rev_hi_lim"                 "inq_fi"
##  [73] "total_cu_tl"                      "inq_last_12m"
##  [75] "acc_open_past_24mths"             "avg_cur_bal"
##  [77] "bc_open_to_buy"                   "bc_util"
##  [79] "chargeoff_within_12_mths"         "delinq_amnt"
##  [81] "mo_sin_old_il_acct"               "mo_sin_old_rev_tl_op"
##  [83] "mo_sin_rcnt_rev_tl_op"            "mo_sin_rcnt_tl"
##  [85] "mort_acc"                         "mths_since_recent_bc"
##  [87] "mths_since_recent_bc_dlq"         "mths_since_recent_inq"
##  [89] "mths_since_recent_revol_delinq"   "num_accts_ever_120_pd"
##  [91] "num_actv_bc_tl"                   "num_actv_rev_tl"
##  [93] "num_bc_sats"                      "num_bc_tl"
##  [95] "num_il_tl"                        "num_op_rev_tl"
##  [97] "num_rev_accts"                    "num_rev_tl_bal_gt_0"
##  [99] "num_sats"                         "num_tl_120dpd_2m"
## [101] "num_tl_30dpd"                     "num_tl_90g_dpd_24m"
## [103] "num_tl_op_past_12m"               "pct_tl_nvr_dlq"
## [105] "percent_bc_gt_75"                 "pub_rec_bankruptcies"
## [107] "tax_liens"                        "tot_hi_cred_lim"
## [109] "total_bal_ex_mort"                "total_bc_limit"
## [111] "total_il_high_credit_limit"
```

Lets have a look at the structure of the dataset:

We can see that there are both quatitative and categorical variables, quantitaive variales are numerical numbers and Factor is a categorical varible, right next to the Factor you can identify how many categorical levels are there in the specified vatiable.

```
## 'data.frame':    39717 obs. of  111 variables:
##  $ id                            : int  1077501 1077430 1077175 1076863
1075358 1075269 1069639 1072053 1071795 1071570 ...
##  $ member_id                     : int  1296599 1314167 1313524 1277178
1311748 1311441 1304742 1288686 1306957 1306721 ...
```

```
##  $ loan_amnt                   : int  5000 2500 2400 10000 3000 5000
7000 3000 5600 5375 ...
##  $ funded_amnt                 : int  5000 2500 2400 10000 3000 5000
7000 3000 5600 5375 ...
##  $ funded_amnt_inv             : num  4975 2500 2400 10000 3000 ...
##  $ term                        : Factor w/ 2 levels " 36 months"," 60
months": 1 2 1 1 2 1 2 1 2 2 ...
##  $ int_rate                    : Factor w/ 371 levels
"10.00%","10.01%",..: 18 159 176 100 75 339 176 240 287 75 ...
##  $ installment                 : num  162.9 59.8 84.3 339.3 67.8 ...
##  $ grade                       : Factor w/ 7 levels "A","B","C","D",..:
2 3 3 3 2 1 3 5 6 2 ...
##  $ sub_grade                   : Factor w/ 35 levels "A1","A2","A3",..:
7 14 15 11 10 4 15 21 27 10 ...
##  $ emp_title                   : Factor w/ 28823 levels "","  old palm
inc",..: 1 21526 1 736 26536 27227 23139 16692 1 23624 ...
##  $ emp_length                  : Factor w/ 12 levels "< 1 year","1
year",..: 3 1 3 3 2 5 10 11 6 1 ...
##  $ home_ownership              : Factor w/ 5 levels
"MORTGAGE","NONE",..: 5 5 5 5 5 5 5 5 4 5 ...
##  $ annual_inc                  : num  24000 30000 12252 49200 80000 ...
##  $ verification_status         : Factor w/ 3 levels "Not Verified",..: 3
2 1 2 2 2 1 2 2 3 ...
##  $ issue_d                     : Factor w/ 55 levels "Apr-08","Apr-
09",..: 14 14 14 14 14 14 14 14 14 14 ...
##  $ loan_status                 : Factor w/ 3 levels "Charged Off",..: 3
1 3 3 2 3 3 3 1 1 ...
##  $ pymnt_plan                  : Factor w/ 1 level "n": 1 1 1 1 1 1 1 1
1 1 ...
##  $ url                         : Factor w/ 39717 levels
"https://lendingclub.com/browse/loanDetail.action?loan_id=1000007",..: 4364
4363 4362 4361 4360 4359 4341 4358 4357 4356 ...
##  $ desc                        : Factor w/ 26529 levels "","- Pay off
Dell Financial: $ 1300.00 - Pay off IRS for 2005: $ 1400.00 - Pay off Mac
Comp    : $ 1700.00 - Pay o"| __truncated__,..: 19926 19927 1 19882 19879 1
19763 19672 19878 19673 ...
##  $ purpose                     : Factor w/ 14 levels
"car","credit_card",..: 2 1 12 10 10 14 3 1 12 10 ...
##  $ title                       : Factor w/ 19617 levels "","'08 & '09
Roth IRA Investments",..: 3449 1746 15905 15302 15067 12963 11132 2531 7040
2142 ...
##  $ zip_code                    : Factor w/ 823 levels
"007xx","010xx",..: 714 278 503 750 799 708 249 735 788 640 ...
##  $ addr_state                  : Factor w/ 50 levels "AK","AL","AR",..:
4 11 15 5 37 4 28 5 5 43 ...
##  $ dti                         : num  27.65 1 8.72 20 17.94 ...
##  $ delinq_2yrs                 : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ earliest_cr_line            : Factor w/ 526 levels "Apr-00","Apr-
01",..: 201 43 388 170 212 391 221 181 5 484 ...
##  $ inq_last_6mths              : int  1 5 2 1 0 3 1 2 2 0 ...
```

```
##  $ mths_since_last_delinq        : int  NA NA NA 35 38 NA NA NA NA NA ...
##  $ mths_since_last_record        : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ open_acc                      : int  3 3 2 10 15 9 7 4 11 2 ...
##  $ pub_rec                       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ revol_bal                     : int  13648 1687 2956 5598 27783 7963
17726 8221 5210 9279 ...
##  $ revol_util                    : Factor w/ 1090 levels
"","0%","0.01%",..: 915 984 1076 168 567 254 934 953 312 357 ...
##  $ total_acc                     : int  9 4 10 37 38 12 11 4 13 3 ...
##  $ initial_list_status           : Factor w/ 1 level "f": 1 1 1 1 1 1 1 1
1 1 ...
##  $ out_prncp                     : num  0 0 0 0 524 ...
##  $ out_prncp_inv                 : num  0 0 0 0 524 ...
##  $ total_pymnt                   : num  5863 1009 3006 12232 3513 ...
##  $ total_pymnt_inv               : num  5834 1009 3006 12232 3513 ...
##  $ total_rec_prncp               : num  5000 456 2400 10000 2476 ...
##  $ total_rec_int                 : num  863 435 606 2215 1037 ...
##  $ total_rec_late_fee            : num  0 0 0 17 0 ...
##  $ recoveries                    : num  0 117 0 0 0 ...
##  $ collection_recovery_fee       : num  0 1.11 0 0 0 0 0 0 2.09 2.52 ...
##  $ last_pymnt_d                  : Factor w/ 102 levels "","Apr-08","Apr-
09",..: 43 7 59 43 78 43 78 43 6 83 ...
##  $ last_pymnt_amnt               : num  171.6 119.7 649.9 357.5 67.8 ...
##  $ next_pymnt_d                  : Factor w/ 3 levels "","Jul-16","Jun-
16": 1 1 1 1 3 1 1 1 1 1 ...
##  $ last_credit_pull_d            : Factor w/ 107 levels "","Apr-09","Apr-
10",..: 82 105 82 9 82 45 82 26 15 69 ...
##  $ collections_12_mths_ex_med    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ mths_since_last_major_derog   : logi  NA NA NA NA NA NA ...
##  $ policy_code                   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ application_type              : Factor w/ 1 level "INDIVIDUAL": 1 1 1 1
1 1 1 1 1 1 ...
##  $ annual_inc_joint              : logi  NA NA NA NA NA NA ...
##  $ dti_joint                     : logi  NA NA NA NA NA NA ...
##  $ verification_status_joint     : logi  NA NA NA NA NA NA ...
##  $ acc_now_delinq                : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ tot_coll_amt                  : logi  NA NA NA NA NA NA ...
##  $ tot_cur_bal                   : logi  NA NA NA NA NA NA ...
##  $ open_acc_6m                   : logi  NA NA NA NA NA NA ...
##  $ open_il_6m                    : logi  NA NA NA NA NA NA ...
##  $ open_il_12m                   : logi  NA NA NA NA NA NA ...
##  $ open_il_24m                   : logi  NA NA NA NA NA NA ...
##  $ mths_since_rcnt_il            : logi  NA NA NA NA NA NA ...
##  $ total_bal_il                  : logi  NA NA NA NA NA NA ...
##  $ il_util                       : logi  NA NA NA NA NA NA ...
##  $ open_rv_12m                   : logi  NA NA NA NA NA NA ...
##  $ open_rv_24m                   : logi  NA NA NA NA NA NA ...
##  $ max_bal_bc                    : logi  NA NA NA NA NA NA ...
##  $ all_util                      : logi  NA NA NA NA NA NA ...
##  $ total_rev_hi_lim              : logi  NA NA NA NA NA NA ...
```

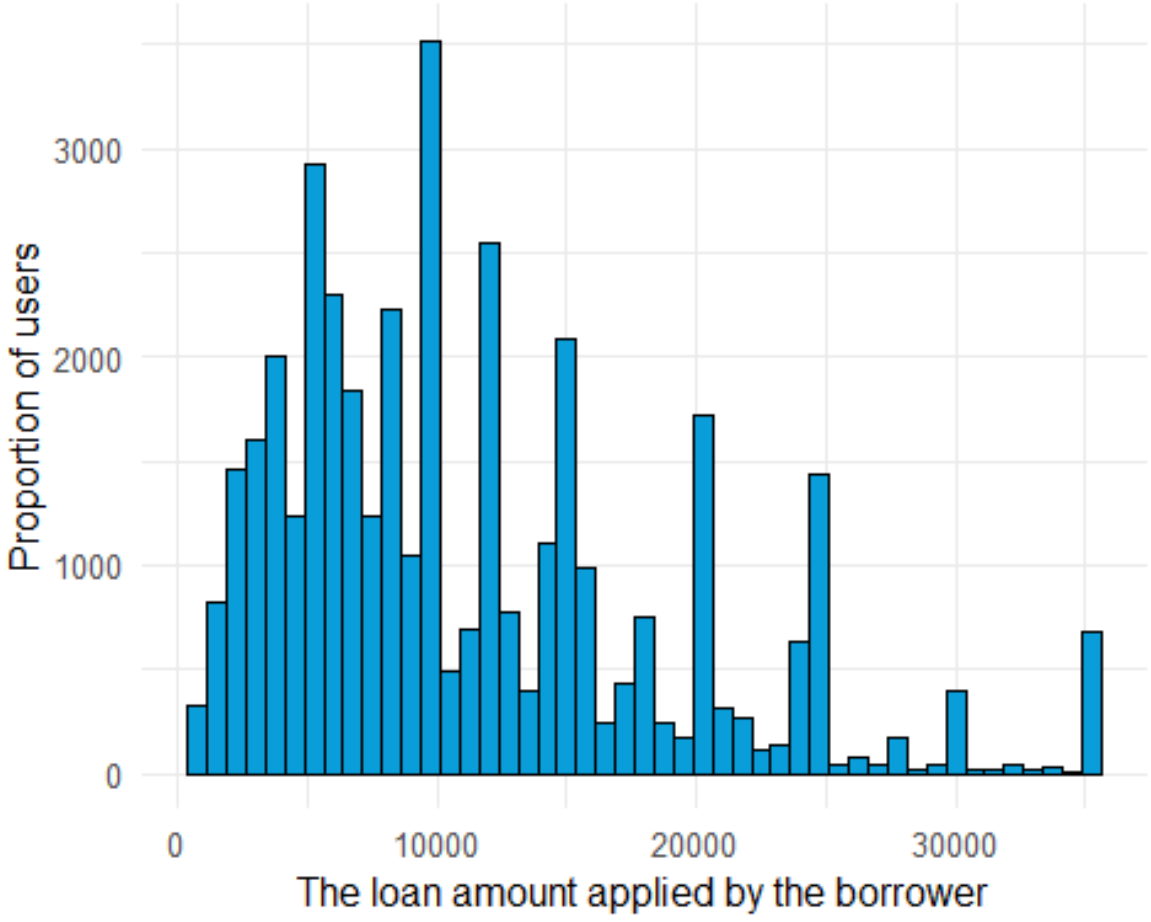
```
##  $ inq_fi                          : logi  NA NA NA NA NA NA ...
##  $ total_cu_tl                     : logi  NA NA NA NA NA NA ...
##  $ inq_last_12m                    : logi  NA NA NA NA NA NA ...
##  $ acc_open_past_24mths            : logi  NA NA NA NA NA NA ...
##  $ avg_cur_bal                     : logi  NA NA NA NA NA NA ...
##  $ bc_open_to_buy                  : logi  NA NA NA NA NA NA ...
##  $ bc_util                         : logi  NA NA NA NA NA NA ...
##  $ chargeoff_within_12_mths        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ delinq_amnt                     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ mo_sin_old_il_acct              : logi  NA NA NA NA NA NA ...
##  $ mo_sin_old_rev_tl_op            : logi  NA NA NA NA NA NA ...
##  $ mo_sin_rcnt_rev_tl_op           : logi  NA NA NA NA NA NA ...
##  $ mo_sin_rcnt_tl                  : logi  NA NA NA NA NA NA ...
##  $ mort_acc                        : logi  NA NA NA NA NA NA ...
##  $ mths_since_recent_bc            : logi  NA NA NA NA NA NA ...
##  $ mths_since_recent_bc_dlq        : logi  NA NA NA NA NA NA ...
##  $ mths_since_recent_inq           : logi  NA NA NA NA NA NA ...
##  $ mths_since_recent_revol_delinq: logi  NA NA NA NA NA NA ...
##  $ num_accts_ever_120_pd           : logi  NA NA NA NA NA NA ...
##  $ num_actv_bc_tl                  : logi  NA NA NA NA NA NA ...
##  $ num_actv_rev_tl                 : logi  NA NA NA NA NA NA ...
##  $ num_bc_sats                     : logi  NA NA NA NA NA NA ...
##  $ num_bc_tl                       : logi  NA NA NA NA NA NA ...
##  $ num_il_tl                       : logi  NA NA NA NA NA NA ...
##  $ num_op_rev_tl                   : logi  NA NA NA NA NA NA ...
##  $ num_rev_accts                   : logi  NA NA NA NA NA NA ...
##  $ num_rev_tl_bal_gt_0             : logi  NA NA NA NA NA NA ...
##  $ num_sats                        : logi  NA NA NA NA NA NA ...
##   [list output truncated]
```

Considering the below table of application_type varibale, We can identify that there are 39717 inidvidual applications.

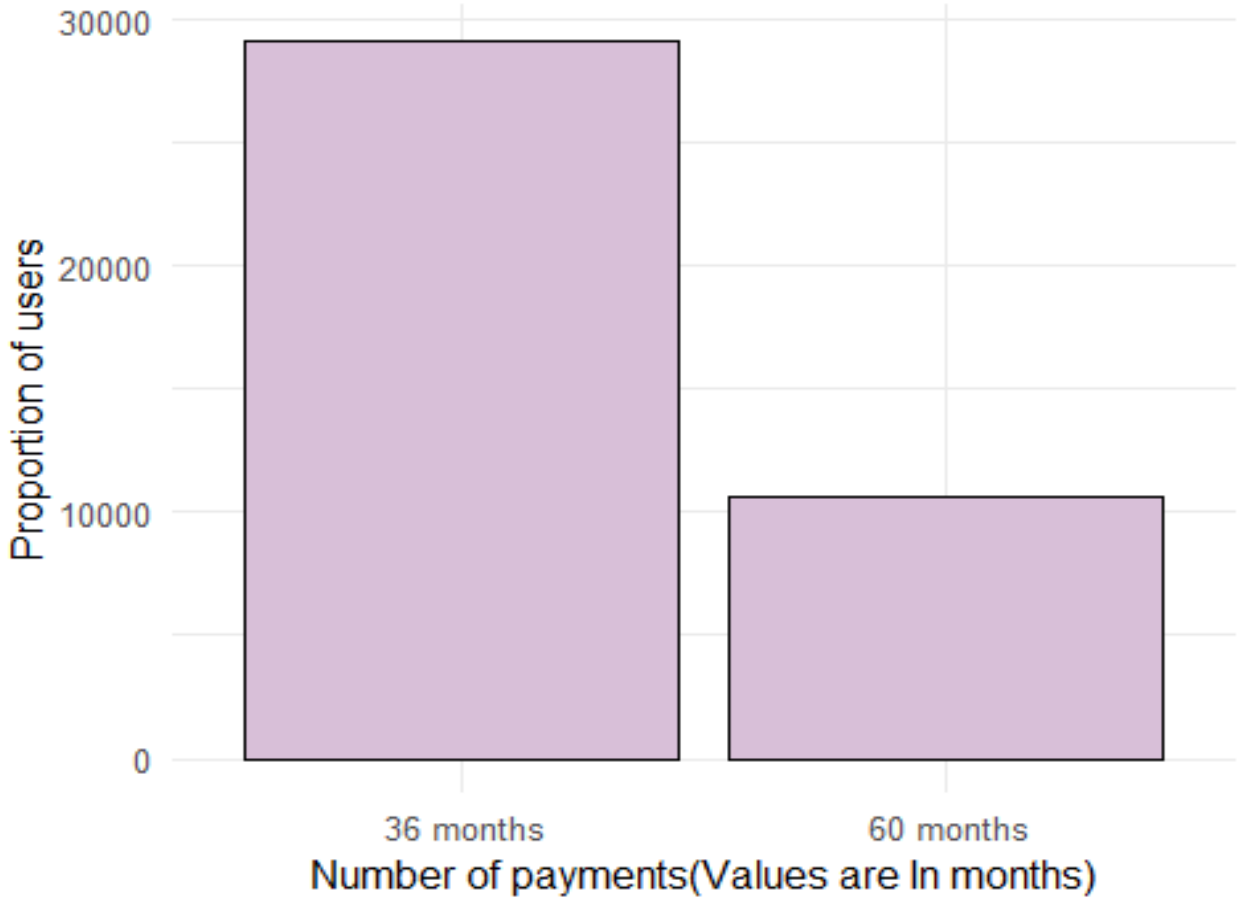
```
##
## INDIVIDUAL
##      39717
```

Univariate Analysis

Lets have look into the amount of the loan applied for by the borrower.

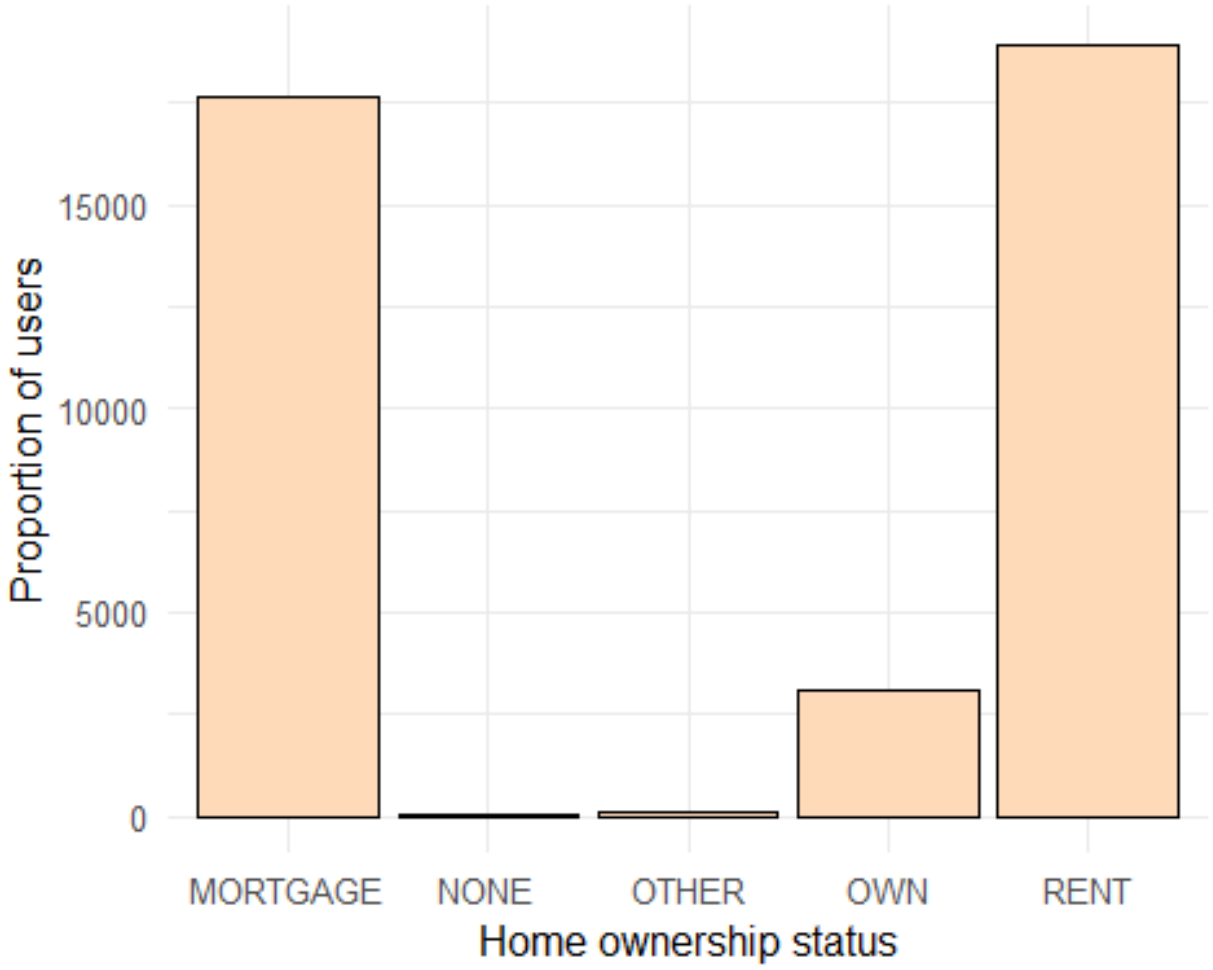


The number of payments on the loan in months.



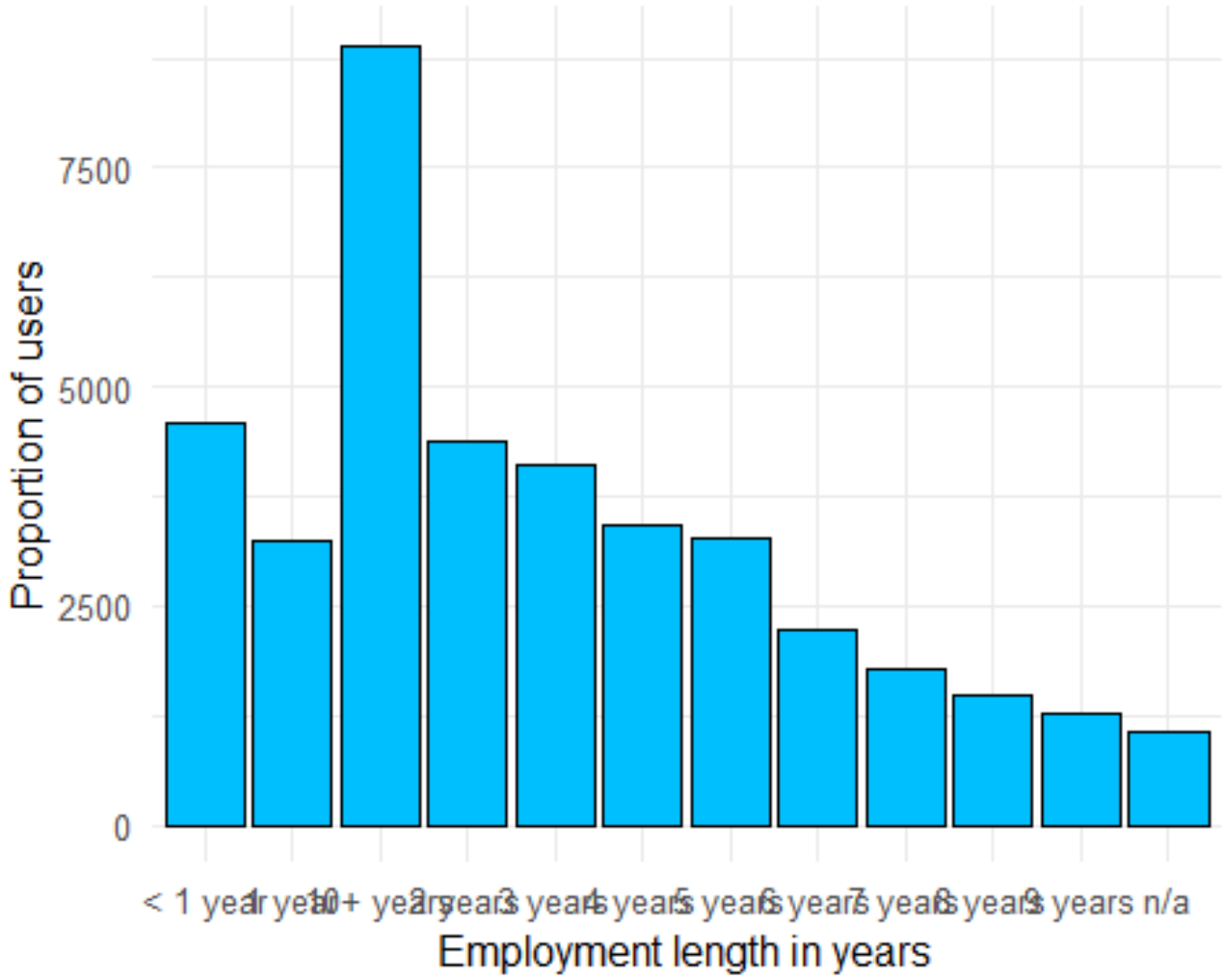
From the above observations we can clearly understand that the most of the borrowed amount is near to $10,000. And most chosen number of payment is 36 months which is three years.

Let's have a look into the status of home ownership:



From the above plot we understood that most of the people are renting the their property and ther is also much similar number of people who are on a mortgage.

Below plot will show when does a person get a loan in terms of the length of their working status.

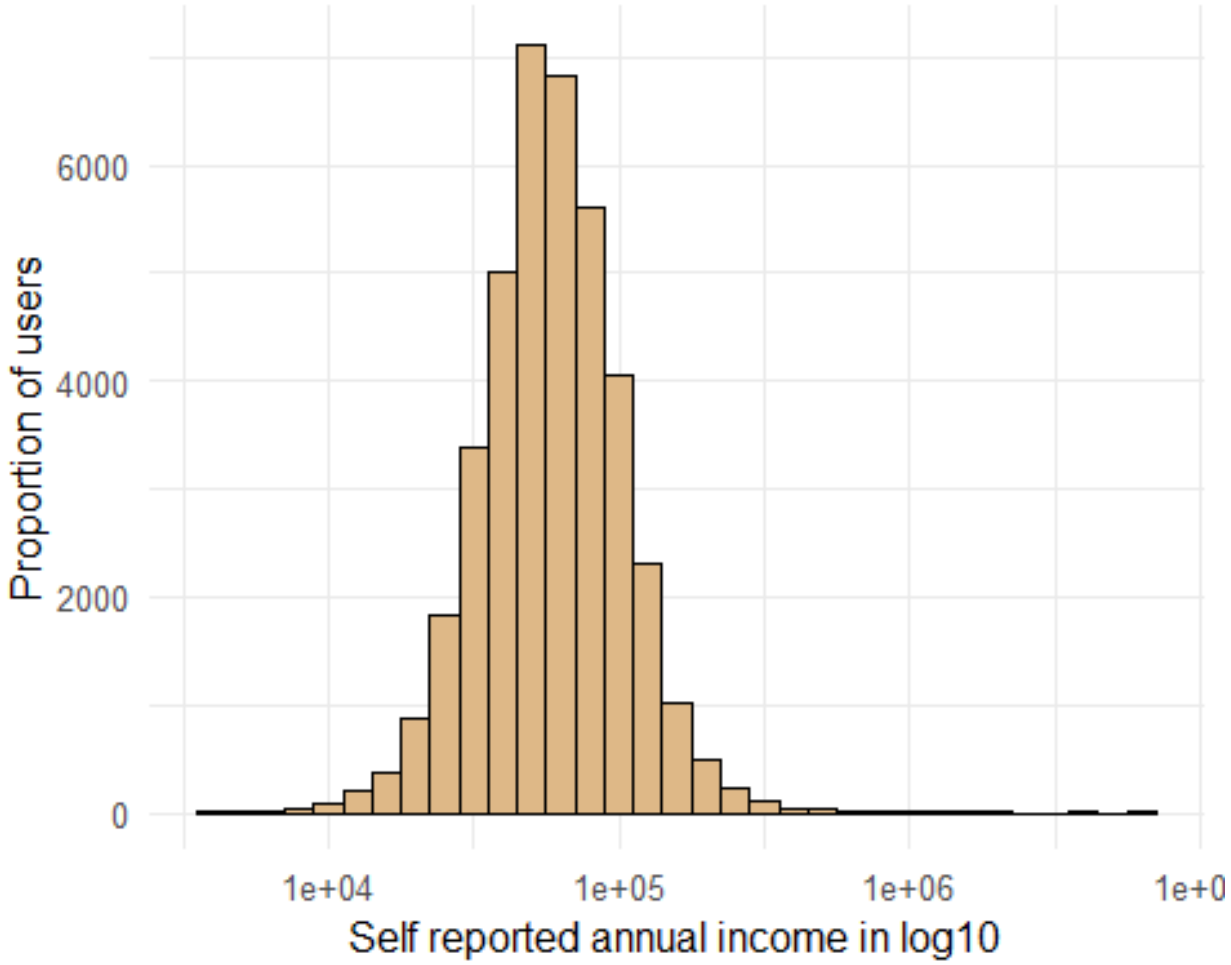


Tabel below is showing the grouped summary of the employment length.

```
##
##  < 1 year    1 year 10+ years   2 years   3 years   4 years   5 years
##      4583      3240      8879      4388      4095      3436      3282
##   6 years   7 years   8 years   9 years       n/a
##      2229      1773      1479      1258      1075
```

In the above observations we can make out that there are 1075 loans with no recorded employment status, these number shows us that either 1075 number of people are unemployed or we have missed the information of theire status of employment. This plot also shows that a good number of people with more than ten years of employment are taking loans, which is almost a 23%. And 12% of people who took loans are less than a year employed. During First year and fifth year we see almost similar number of loans. Second year we have almost similar number of loans compare to the less than a year employed people. from the third year onwards we see a decreasing trend till nineth year.

Let's have a look at the annual income of the borrowers:

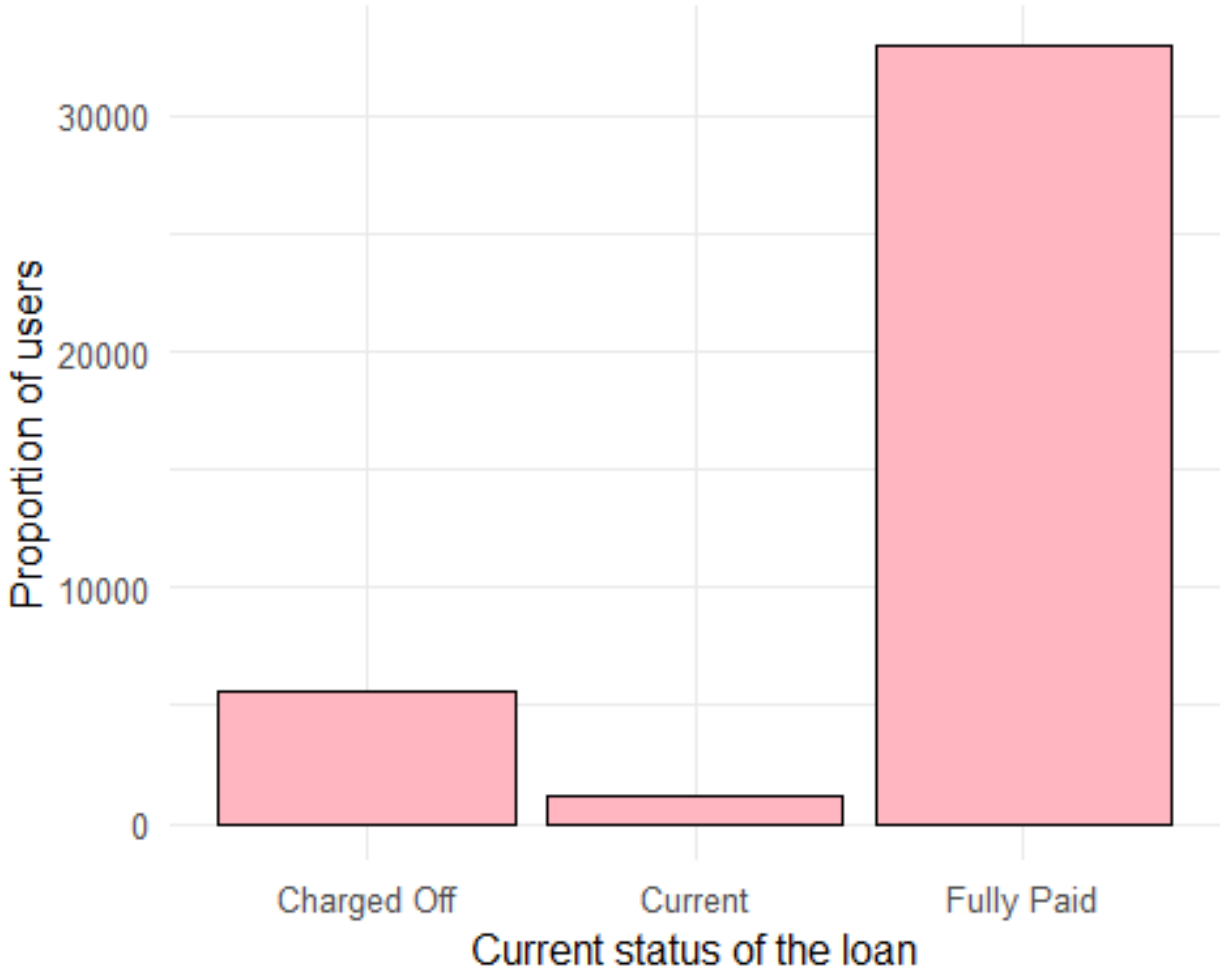


Below is the summary of the annual income provided by the borrower.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4000   40404   59000   68969   82300 6000000
```

The income of most borrowers range from $40,000 to $82,000, the highest recorded income is $6,000,000.

The below plot will give us a picture of the current staus of the loan:

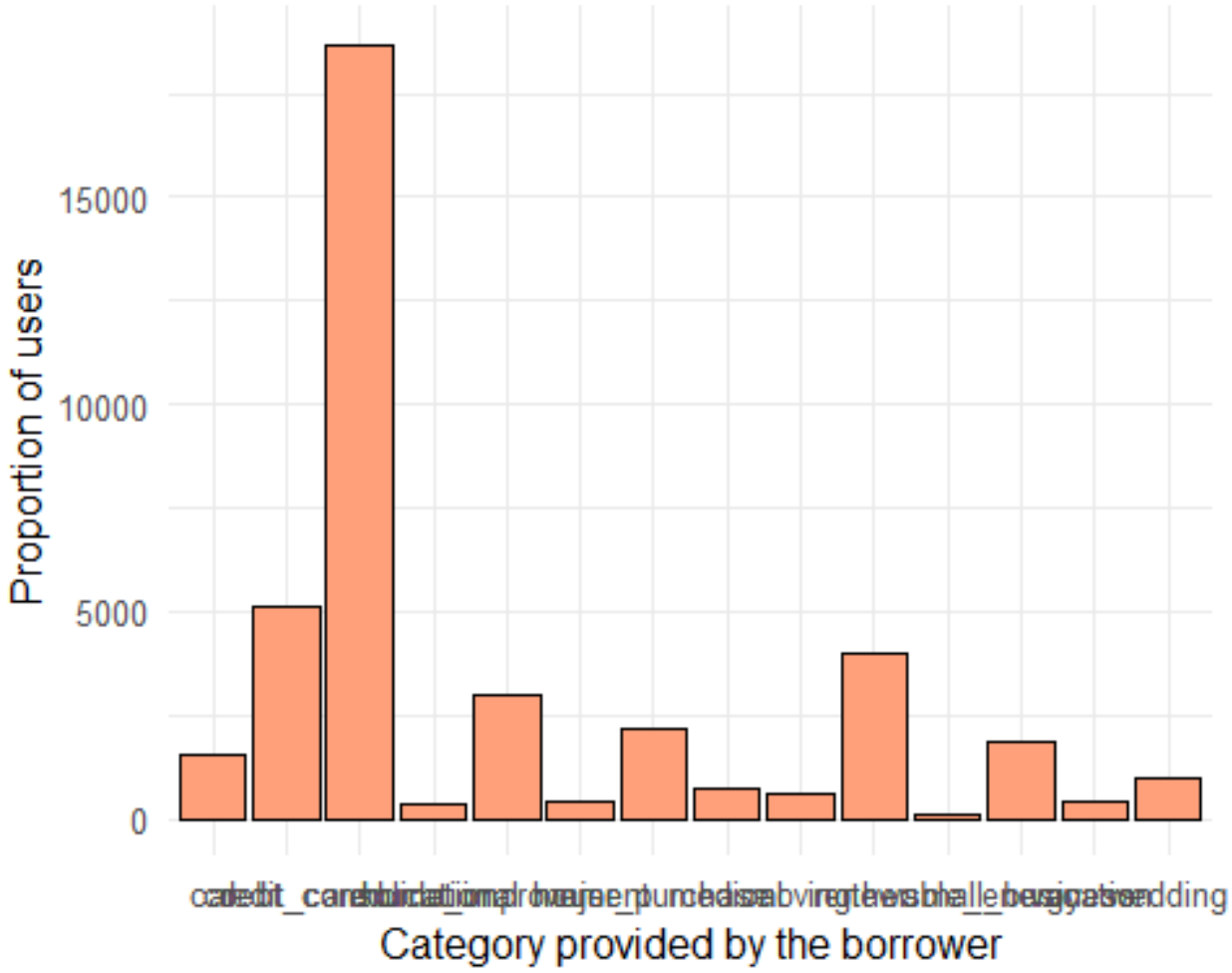


Grouped loan status table.

```
##
## Charged Off     Current  Fully Paid
##        5627        1140       32950
```

From the above observation we can make out that there are 5627 loans which had been defaulted. Which is a 14% of total number of loan given.

Let's see for what purpose people borrow money for:



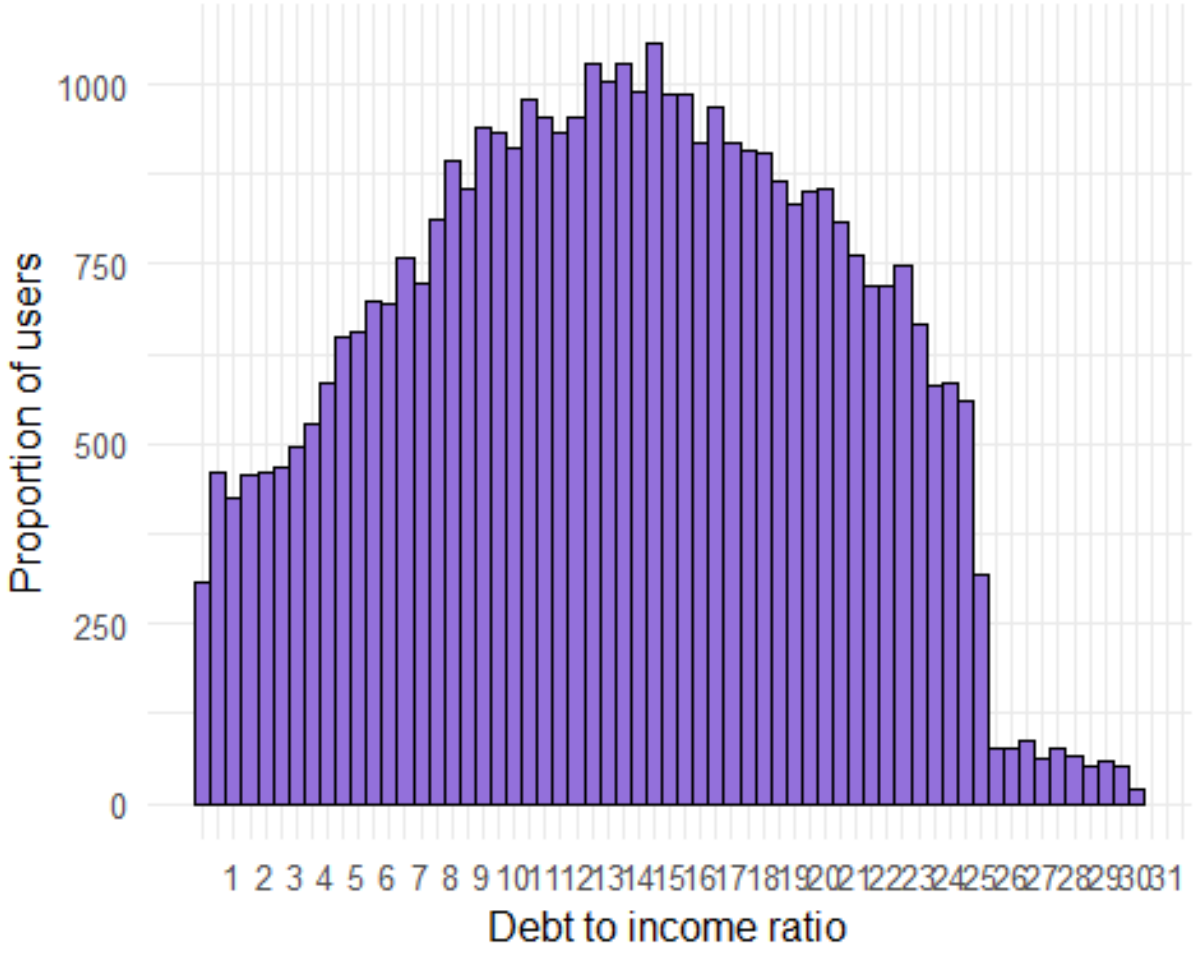
Below table give us the exact number of different loan purpose.

```
##                 car         credit_card debt_consolidation 
##                1549                5130               18641 
##         educational    home_improvement               house 
##                 325                2976                 381 
##      major_purchase             medical              moving 
##                2187                 693                 583 
##               other    renewable_energy      small_business 
##                3993                 103                1828 
##            vacation             wedding                     
##                 381                 947
```

Majority of the people got the loan for debt consolidation. Credit card is also an important purpose of the loan. Another thing I found is about somewhat similar amount of people got loans for home improvement and major purchase. One interesting thing i found that few number of people are taking loan for renewable energy. This is a good sign that people are thinking about the climate change and they are acting. We need to have a deeper

understanding of this and need to think about how can we encourage more people apply for taking loan for renewable energy. May be we need to consider some special interest rates for this in the future. And a good portion of the people didnt mentioned the specific purpose of the loan.

Lets have a look into the Debt to income ratio of borrowers:



Inorder to find out how many people are having dti <1, we will use below quantile function.

```
##   99%
## 26.68
```

After limiting the ratio to <1, we can understand that only 29% has debt to income ratio less than 1. And if we call the function quantile(loan$dti, 0.99, na.rm=T), it tells us that almost 75% of the borrowers has their dti > 1. Which is dnagerously high. Evidence from studies of loans suggest that borrowers with a higher debt to income ratio are more likely to run into trouble making monthly payments.

Income Verification:


```
##
##    Not Verified Source Verified        Verified
##           16921             9987           12809
```

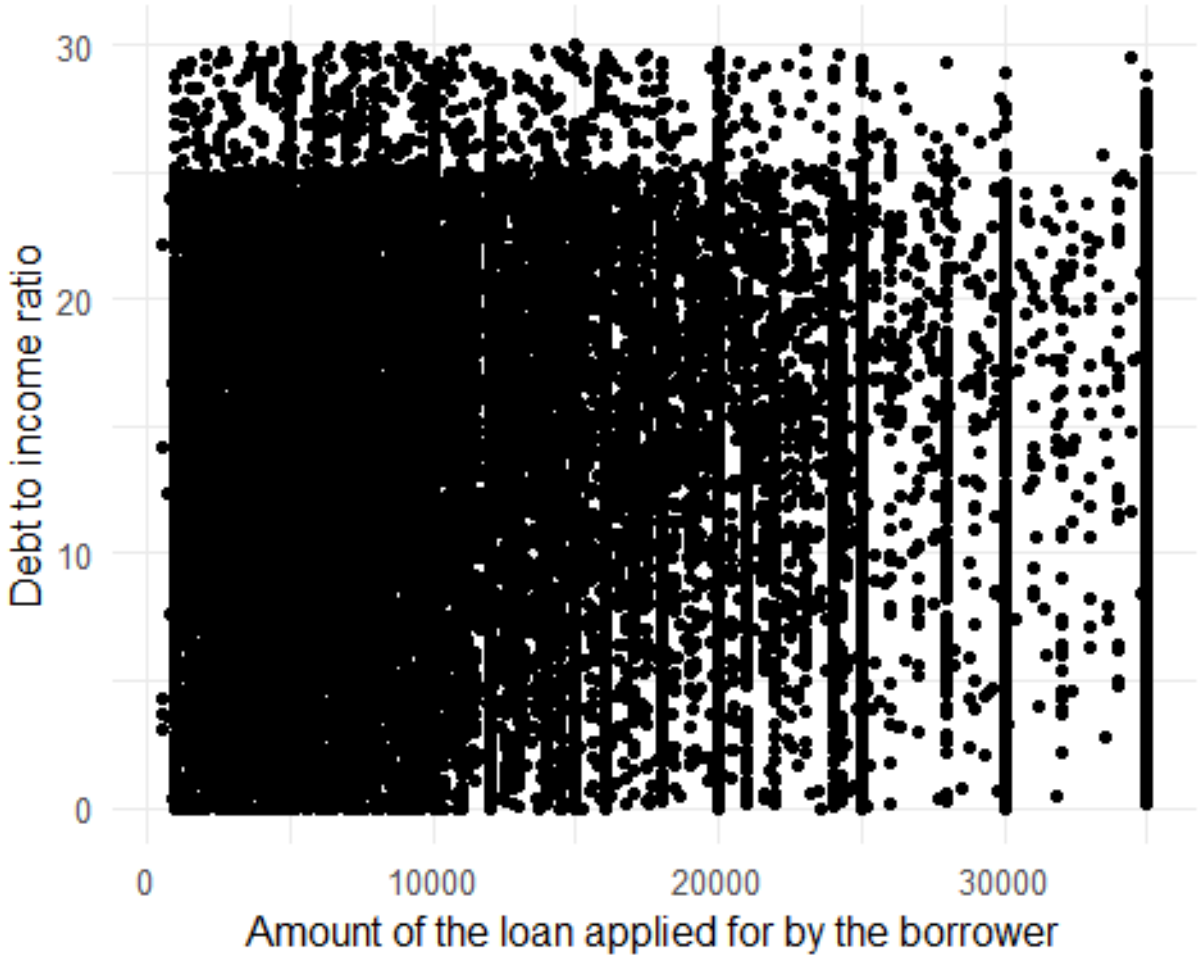
From the above plot we can understand that half of the loan was allowed wihout any verification. Approving a loan without verifing documents and backgroudnd is also a cause to default.

Bi-Variate Analysis

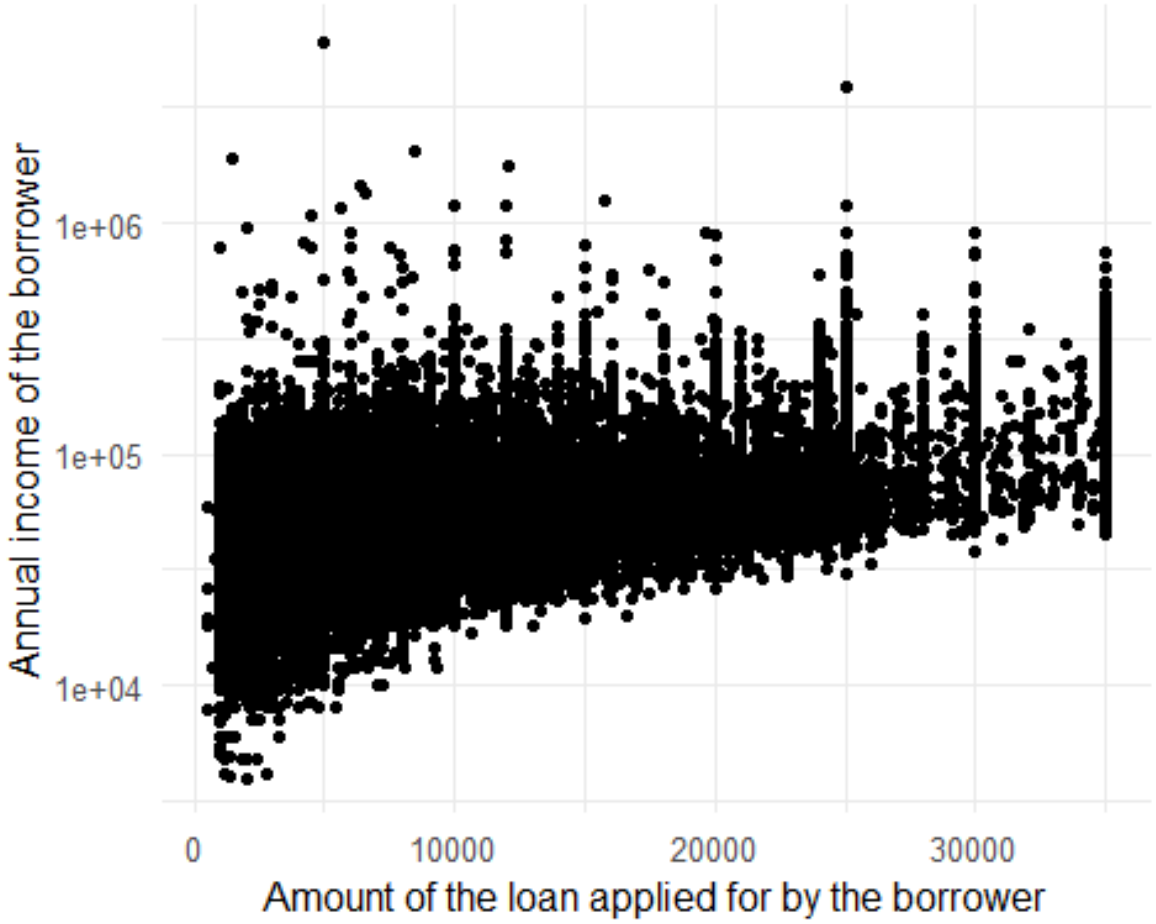
Let's have a look into the annual income and monthly installment of the borrowers.

From the above plot we can identify a positive moderate upward curve from the obervations. It shows us the streangth and direction of the varibles. If we verify annual income of all the borrowers, we dont have to do anything new with the monthly payment calculation for each loan.

Loan amount and debt to income ratio plot:



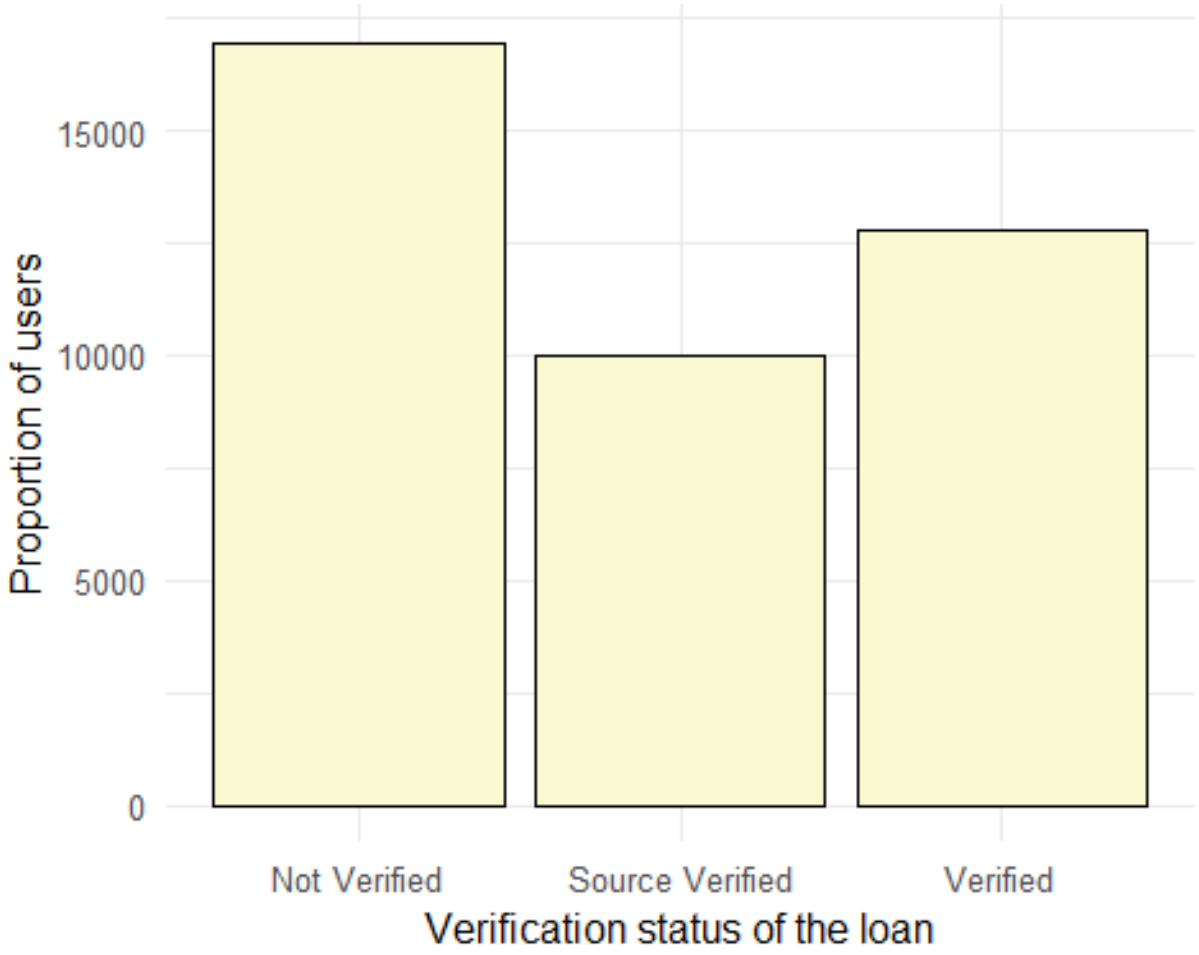
The above plot is giving us a weak spread out picture of observations, it shows us how inappropiate and high debt to income ratio was when we comapare it to the loan amount had taken by the borrower.



The above plot is reflecting that the points are spreaded in a slight moderate positive upward direction. which showing us that the people having higher annual income takes highest loan amount. The higher number of loans are from people who are having an annual income of $10,0000 and less, which indicates the wide spread dark concentrated points in the plot. Vertical bar shows us that most people takes a rounded figure like, $10k, $15k, $20k and so on.

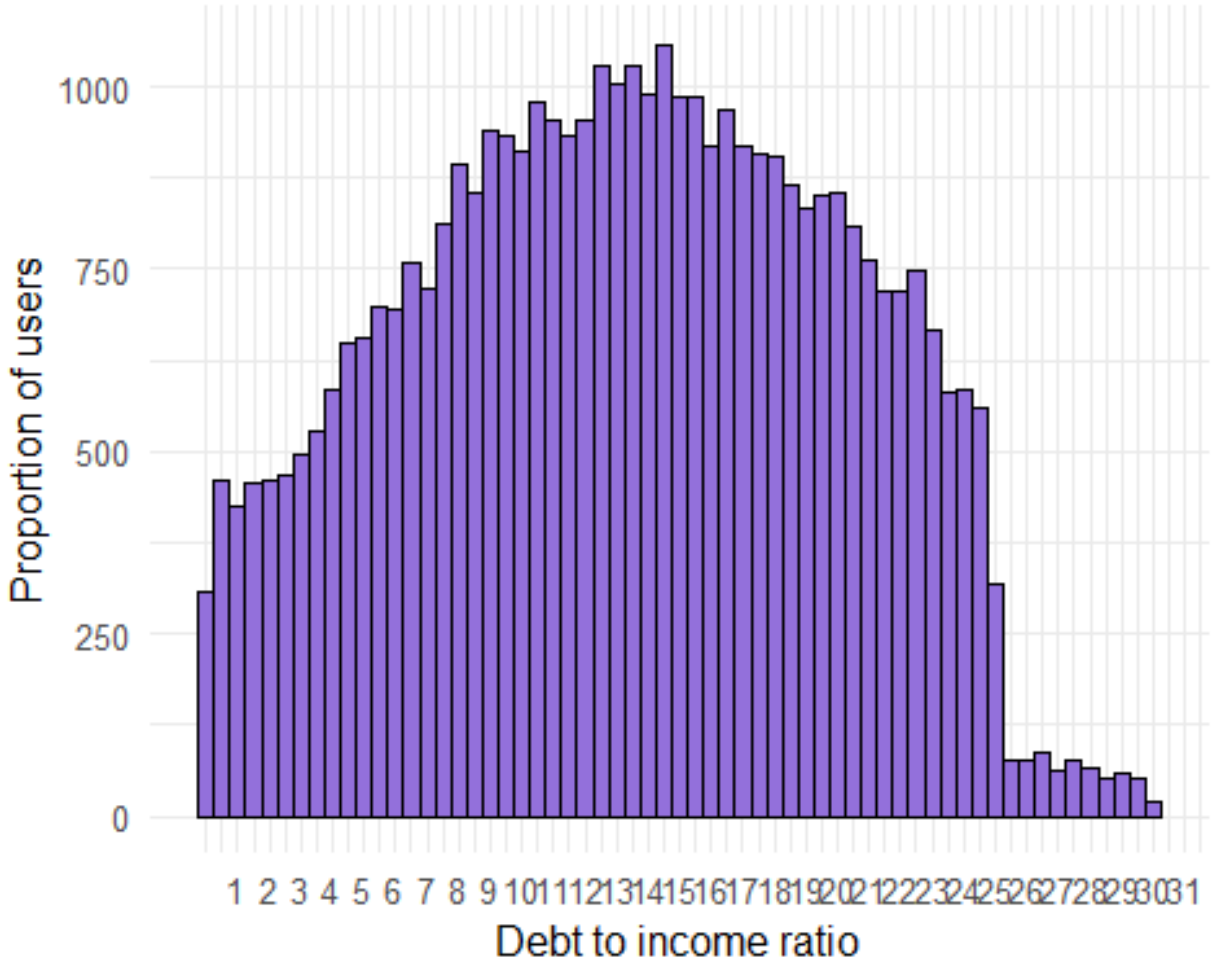
Final Plots and Summary

1. **Income Verification:**



Verfication is a keyfactor, and income verification has to be done in a way that there is no forgery involved. If the annual incomes of all the borrowers are verified then only we will be able to get a proper debt to income ratio.

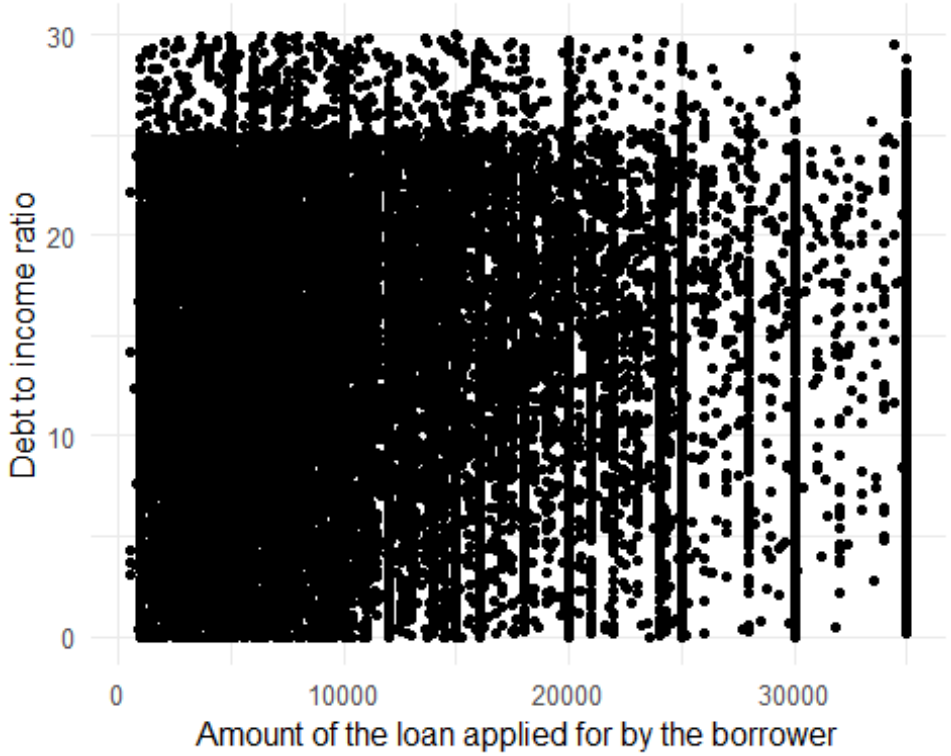
2. **Debt to income ratio:**



Debt to income ratio is the monthly debt payments divided by gross monthly income. This number is one way lenders measure borrowers ability to manage the payments they make every month to repay the money borrowed.

We understood that only 26% has debt to income ratio less than 1. It shows us that almost 75% of the borrowers has their dti > 1. Which is dnagerously high. Evidence from studies of loans suggest that borrowers with a higher debt to income ratio are more likely to run into trouble in making monthly payments.Higher risk is usually associated with higher monthly payment. And borrowers must keep their debt to income ratio low. Paying a huge amount of money from monthly salary in every month for a period of 3 years is not possible, of course you have other things to do in daily life, you need money for that also.

3. **Loan amount and debt to income ratio plot:**



The above plot is giving us a weak spread out picture of observations, it is a reflection of how inappropiate and high debt to income ration was when we comapare it to the loan amount had taken by the borrower.

End

Reflection

I have created this markdown in R language inside RStudio. I have used ggplot to create different types of plots and visualization, i found it simple and efficient using ggplot rather than the default R syntax.

PDF document is for just a reference, please view the visualization in Rstudio for better rendering.

Since i have been learning R for some time, this is very first interaction with a live project. Thank you for showing interest and sharing the information.

Please find the below session info.

```
## R version 3.6.0 (2019-04-26)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 18362)
##
## Matrix products: default
##
```

```
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] tinytex_0.14  ggplot2_3.2.0
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.1       knitr_1.23       magrittr_1.5     munsell_0.5.0
##  [5] colorspace_1.4-1 rlang_0.3.4      stringr_1.4.0    tools_3.6.0
##  [9] grid_3.6.0       gtable_0.3.0     xfun_0.7         withr_2.1.2
## [13] htmltools_0.3.6  yaml_2.2.0       lazyeval_0.2.2   digest_0.6.19
## [17] tibble_2.1.3     crayon_1.3.4     evaluate_0.14    rmarkdown_1.13
## [21] labeling_0.3     stringi_1.4.3    compiler_3.6.0   pillar_1.4.1
## [25] scales_1.0.0     pkgconfig_2.0.2
```