Data_Analysis_With_R_Final_Project

Erwin

18 March 2017

First look at Prosper Loan Data Set

```
##load the loan data to dataframe
loan_full <- read.csv(file='data/prosperLoanData.csv')

##take a look at the list of variables. 81 variables in total
names(loan_full)
```

```
##  [1] "ListingKey"
##  [2] "ListingNumber"
##  [3] "ListingCreationDate"
##  [4] "CreditGrade"
##  [5] "Term"
##  [6] "LoanStatus"
##  [7] "ClosedDate"
##  [8] "BorrowerAPR"
##  [9] "BorrowerRate"
## [10] "LenderYield"
## [11] "EstimatedEffectiveYield"
## [12] "EstimatedLoss"
## [13] "EstimatedReturn"
## [14] "ProsperRating..numeric."
## [15] "ProsperRating..Alpha."
## [16] "ProsperScore"
## [17] "ListingCategory..numeric."
## [18] "BorrowerState"
## [19] "Occupation"
## [20] "EmploymentStatus"
## [21] "EmploymentStatusDuration"
## [22] "IsBorrowerHomeowner"
## [23] "CurrentlyInGroup"
## [24] "GroupKey"
## [25] "DateCreditPulled"
## [26] "CreditScoreRangeLower"
## [27] "CreditScoreRangeUpper"
## [28] "FirstRecordedCreditLine"
## [29] "CurrentCreditLines"
## [30] "OpenCreditLines"
## [31] "TotalCreditLinespast7years"
## [32] "OpenRevolvingAccounts"
## [33] "OpenRevolvingMonthlyPayment"
## [34] "InquiriesLast6Months"
## [35] "TotalInquiries"
## [36] "CurrentDelinquencies"
## [37] "AmountDelinquent"
## [38] "DelinquenciesLast7Years"
## [39] "PublicRecordsLast10Years"
```

```
## [40] "PublicRecordsLast12Months"
## [41] "RevolvingCreditBalance"
## [42] "BankcardUtilization"
## [43] "AvailableBankcardCredit"
## [44] "TotalTrades"
## [45] "TradesNeverDelinquent..percentage."
## [46] "TradesOpenedLast6Months"
## [47] "DebtToIncomeRatio"
## [48] "IncomeRange"
## [49] "IncomeVerifiable"
## [50] "StatedMonthlyIncome"
## [51] "LoanKey"
## [52] "TotalProsperLoans"
## [53] "TotalProsperPaymentsBilled"
## [54] "OnTimeProsperPayments"
## [55] "ProsperPaymentsLessThanOneMonthLate"
## [56] "ProsperPaymentsOneMonthPlusLate"
## [57] "ProsperPrincipalBorrowed"
## [58] "ProsperPrincipalOutstanding"
## [59] "ScorexChangeAtTimeOfListing"
## [60] "LoanCurrentDaysDelinquent"
## [61] "LoanFirstDefaultedCycleNumber"
## [62] "LoanMonthsSinceOrigination"
## [63] "LoanNumber"
## [64] "LoanOriginalAmount"
## [65] "LoanOriginationDate"
## [66] "LoanOriginationQuarter"
## [67] "MemberKey"
## [68] "MonthlyLoanPayment"
## [69] "LP_CustomerPayments"
## [70] "LP_CustomerPrincipalPayments"
## [71] "LP_InterestandFees"
## [72] "LP_ServiceFees"
## [73] "LP_CollectionFees"
## [74] "LP_GrossPrincipalLoss"
## [75] "LP_NetPrincipalLoss"
## [76] "LP_NonPrincipalRecoverypayments"
## [77] "PercentFunded"
## [78] "Recommendations"
## [79] "InvestmentFromFriendsCount"
## [80] "InvestmentFromFriendsAmount"
## [81] "Investors"
```

With so many variables in the datasets, I need to first understand the meaning of the variables based on the given variable definitions.

Because it is a loan data, some of the initial thoughts I have is to analyze on variables related to income, profession / employment, loan amount, credit info of the borrower.

As such, I have shortlisted 19 variables for the analysis in this project: EmploymentStatusDuration, CreditScoreRangeUpper, ListingCategory, CreditScoreRangeLower, CurrentCreditLines, DebtToIncomeRatio, StatedMonthlyIncome, Occupation, BorrowerState, EmploymentStatus, IsBorrowerHomeowner, IncomeRange, IncomeVerifiable, BorrowerAPR, BorrowerRate, Term, LoanStatus, LoanOriginalAmount, LoanOriginationDate, MonthlyLoanPayment

```
## Create a subset of the full dataset with preselected variables that will be analyzed
loan <- select(loan_full,
            EmploymentStatusDuration, ListingCategory..numeric., CreditScoreRangeLower,
            CurrentCreditLines, DebtToIncomeRatio,
            StatedMonthlyIncome, Occupation,
            BorrowerState, EmploymentStatus,
            IsBorrowerHomeowner, IncomeRange,
            IncomeVerifiable, BorrowerAPR,
            BorrowerRate, Term,
            LoanStatus, LoanOriginalAmount,
            LoanOriginationDate, MonthlyLoanPayment,
            ProsperRating..Alpha.,ProsperRating..numeric.)

loan$Term <- factor(loan$Term)


summary(loan)
```

```
##  EmploymentStatusDuration ListingCategory..numeric. CreditScoreRangeLower
##  Min.   :  0.00           Min.   : 0.000            Min.   :  0.0
##  1st Qu.: 26.00           1st Qu.: 1.000            1st Qu.:660.0
##  Median : 67.00           Median : 1.000            Median :680.0
##  Mean   : 96.07           Mean   : 2.774            Mean   :685.6
##  3rd Qu.:137.00           3rd Qu.: 3.000            3rd Qu.:720.0
##  Max.   :755.00           Max.   :20.000            Max.   :880.0
##  NA's   :7625                                       NA's   :591
##  CurrentCreditLines DebtToIncomeRatio StatedMonthlyIncome
##  Min.   : 0.00      Min.   : 0.000    Min.   :      0
##  1st Qu.: 7.00      1st Qu.: 0.140    1st Qu.:   3200
##  Median :10.00      Median : 0.220    Median :   4667
##  Mean   :10.32      Mean   : 0.276    Mean   :   5608
##  3rd Qu.:13.00      3rd Qu.: 0.320    3rd Qu.:   6825
##  Max.   :59.00      Max.   :10.010    Max.   :1750003
##  NA's   :7604       NA's   :8554
##                     Occupation    BorrowerState        EmploymentStatus
##  Other                   :28617   CA     :14717   Employed     :67322
##  Professional            :13628   TX     : 6842   Full-time    :26355
##  Computer Programmer     : 4478   NY     : 6729   Self-employed: 6134
##  Executive               : 4311   FL     : 6720   Not available: 5347
##  Teacher                 : 3759   IL     : 5921   Other        : 3806
##  Administrative Assistant: 3688          : 5515                : 2255
##  (Other)                 :55456   (Other):67493   (Other)      : 2718
##  IsBorrowerHomeowner         IncomeRange    IncomeVerifiable
##  False:56459         $25,000-49,999:32192   False:  8669
##  True :57478         $50,000-74,999:31050   True :105268
##                      $100,000+     :17337
##                      $75,000-99,999:16916
##                      Not displayed : 7741
##                      $1-24,999     : 7274
##                      (Other)       : 1427
##   BorrowerAPR       BorrowerRate    Term
##  Min.   :0.00653   Min.   :0.0000   12: 1614
##  1st Qu.:0.15629   1st Qu.:0.1340   36:87778
##  Median :0.20976   Median :0.1840   60:24545
```

```
##  Mean   :0.21883   Mean   :0.1928
##  3rd Qu.:0.28381   3rd Qu.:0.2500
##  Max.   :0.51229   Max.   :0.4975
##  NA's   :25
##                  LoanStatus    LoanOriginalAmount
##  Current              :56576   Min.   : 1000
##  Completed            :38074   1st Qu.: 4000
##  Chargedoff           :11992   Median : 6500
##  Defaulted            : 5018   Mean   : 8337
##  Past Due (1-15 days) :  806   3rd Qu.:12000
##  Past Due (31-60 days):  363   Max.   :35000
##  (Other)              : 1108
##           LoanOriginationDate MonthlyLoanPayment ProsperRating..Alpha.
##  2014-01-22 00:00:00:   491   Min.   :   0.0            :29084
##  2013-11-13 00:00:00:   490   1st Qu.: 131.6     C      :18345
##  2014-02-19 00:00:00:   439   Median : 217.7     B      :15581
##  2013-10-16 00:00:00:   434   Mean   : 272.5     A      :14551
##  2014-01-28 00:00:00:   339   3rd Qu.: 371.6     D      :14274
##  2013-09-24 00:00:00:   316   Max.   :2251.5     E      : 9795
##  (Other)            :111428                      (Other):12307
##  ProsperRating..numeric.
##  Min.   :1.000
##  1st Qu.:3.000
##  Median :4.000
##  Mean   :4.072
##  3rd Qu.:5.000
##  Max.   :7.000
##  NA's   :29084
```

```
str(loan)
```

```
## 'data.frame':    113937 obs. of  21 variables:
##  $ EmploymentStatusDuration : int  2 44 NA 113 44 82 172 103 269 269 ...
##  $ ListingCategory..numeric.: int  0 2 0 16 2 1 1 2 7 7 ...
##  $ CreditScoreRangeLower    : int  640 680 480 800 680 740 680 700 820 820 ...
##  $ CurrentCreditLines       : int  5 14 NA 5 19 21 10 6 17 17 ...
##  $ DebtToIncomeRatio        : num  0.17 0.18 0.06 0.15 0.26 0.36 0.27 0.24 0.25 0.25 ...
##  $ StatedMonthlyIncome      : num  3083 6125 2083 2875 9583 ...
##  $ Occupation               : Factor w/ 68 levels "","Accountant/CPA",..: 37 43 37 52 21 43 50 29 24
##  $ BorrowerState            : Factor w/ 52 levels "","AK","AL","AR",..: 7 7 12 12 25 34 18 6 16 16 .
##  $ EmploymentStatus         : Factor w/ 9 levels "","Employed",..: 9 2 4 2 2 2 2 2 2 2 ...
##  $ IsBorrowerHomeowner      : Factor w/ 2 levels "False","True": 2 1 1 2 2 2 1 1 2 2 ...
##  $ IncomeRange              : Factor w/ 8 levels "$0","$1-24,999",..: 4 5 7 4 3 3 4 4 4 4 ...
##  $ IncomeVerifiable         : Factor w/ 2 levels "False","True": 2 2 2 2 2 2 2 2 2 2 ...
##  $ BorrowerAPR              : num  0.165 0.12 0.283 0.125 0.246 ...
##  $ BorrowerRate             : num  0.158 0.092 0.275 0.0974 0.2085 ...
##  $ Term                     : Factor w/ 3 levels "12","36","60": 2 2 2 2 2 3 2 2 2 2 ...
##  $ LoanStatus               : Factor w/ 12 levels "Cancelled","Chargedoff",..: 3 4 3 4 4 4 4 4 4 4 .
##  $ LoanOriginalAmount       : int  9425 10000 3001 10000 15000 15000 3000 10000 10000 10000 ...
##  $ LoanOriginationDate      : Factor w/ 1873 levels "2005-11-15 00:00:00",..: 426 1866 260 1535 1757
##  $ MonthlyLoanPayment       : num  330 319 123 321 564 ...
##  $ ProsperRating..Alpha.    : Factor w/ 8 levels "","A","AA","B",..: 1 2 1 2 6 4 7 5 3 3 ...
##  $ ProsperRating..numeric.  : int  NA 6 NA 6 3 5 2 4 7 7 ...
```

```
names(loan)
```

```
##  [1] "EmploymentStatusDuration"  "ListingCategory..numeric."
##  [3] "CreditScoreRangeLower"     "CurrentCreditLines"
##  [5] "DebtToIncomeRatio"         "StatedMonthlyIncome"
##  [7] "Occupation"                "BorrowerState"
##  [9] "EmploymentStatus"          "IsBorrowerHomeowner"
## [11] "IncomeRange"               "IncomeVerifiable"
## [13] "BorrowerAPR"               "BorrowerRate"
## [15] "Term"                      "LoanStatus"
## [17] "LoanOriginalAmount"        "LoanOriginationDate"
## [19] "MonthlyLoanPayment"        "ProsperRating..Alpha."
## [21] "ProsperRating..numeric."
```

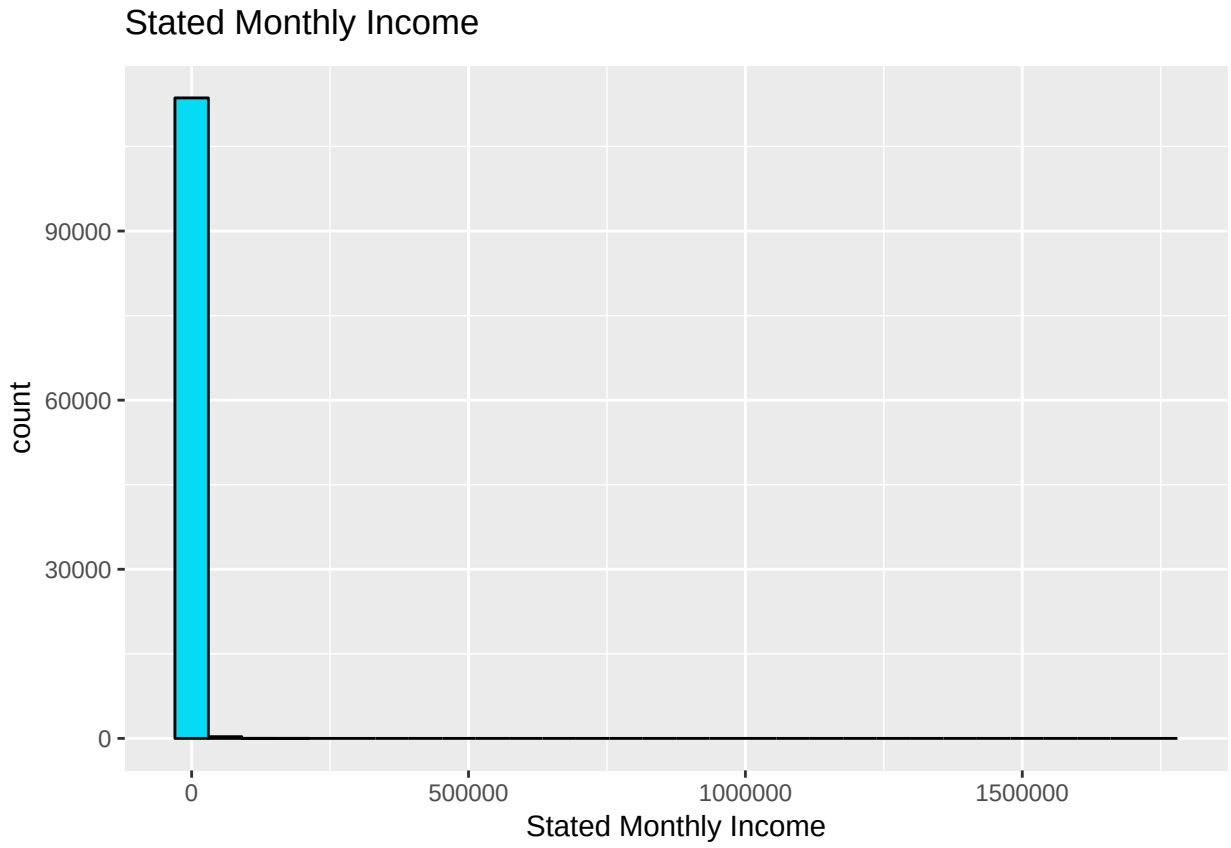
There is a good mix of both discrete and continuous variables, which can be further explored in the univariate plot section below.

Univariate Plots

This section analyzes the distribution and characteristics of single variables that were selected above.

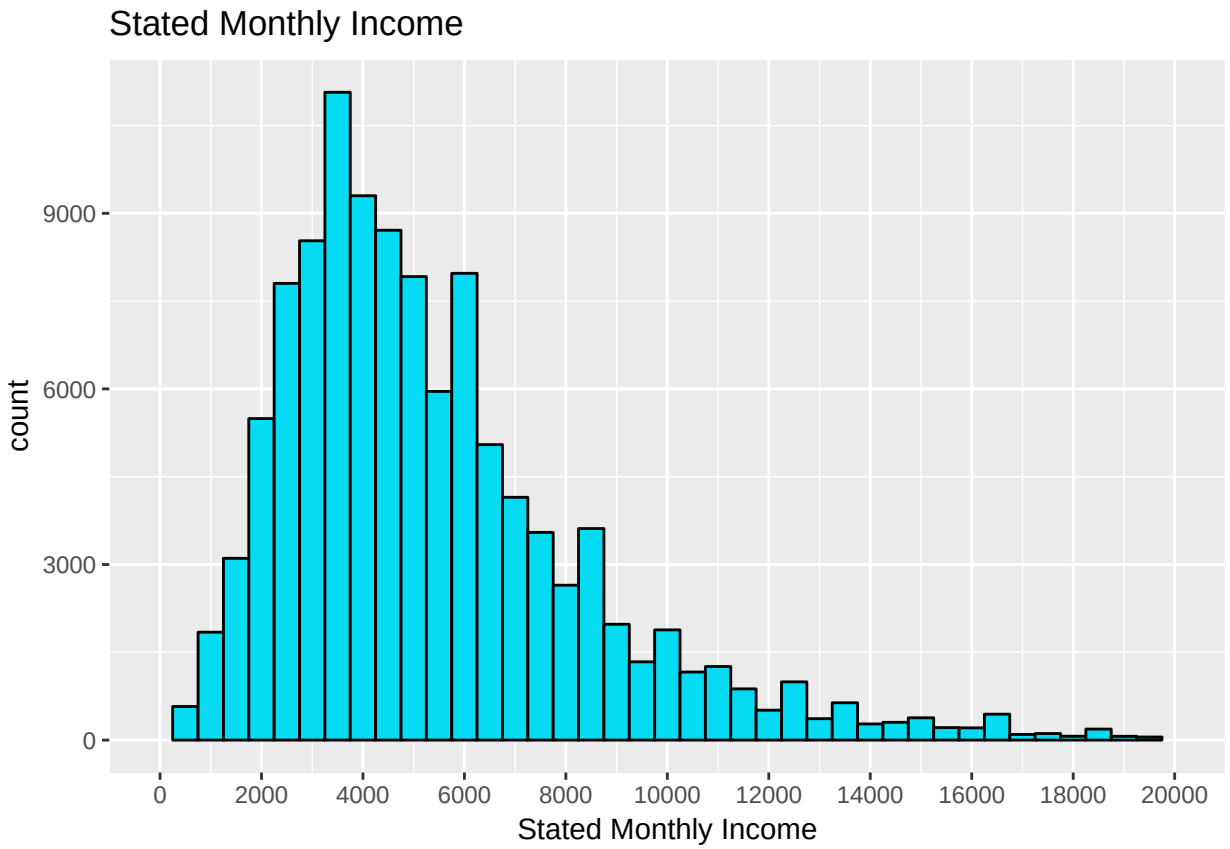
Here are some custom functions to help summarize the information of variables that will be used later.

First thing that comes to my mind when we talk about loan is income. So, let's first look at the stated monthly



income of users.

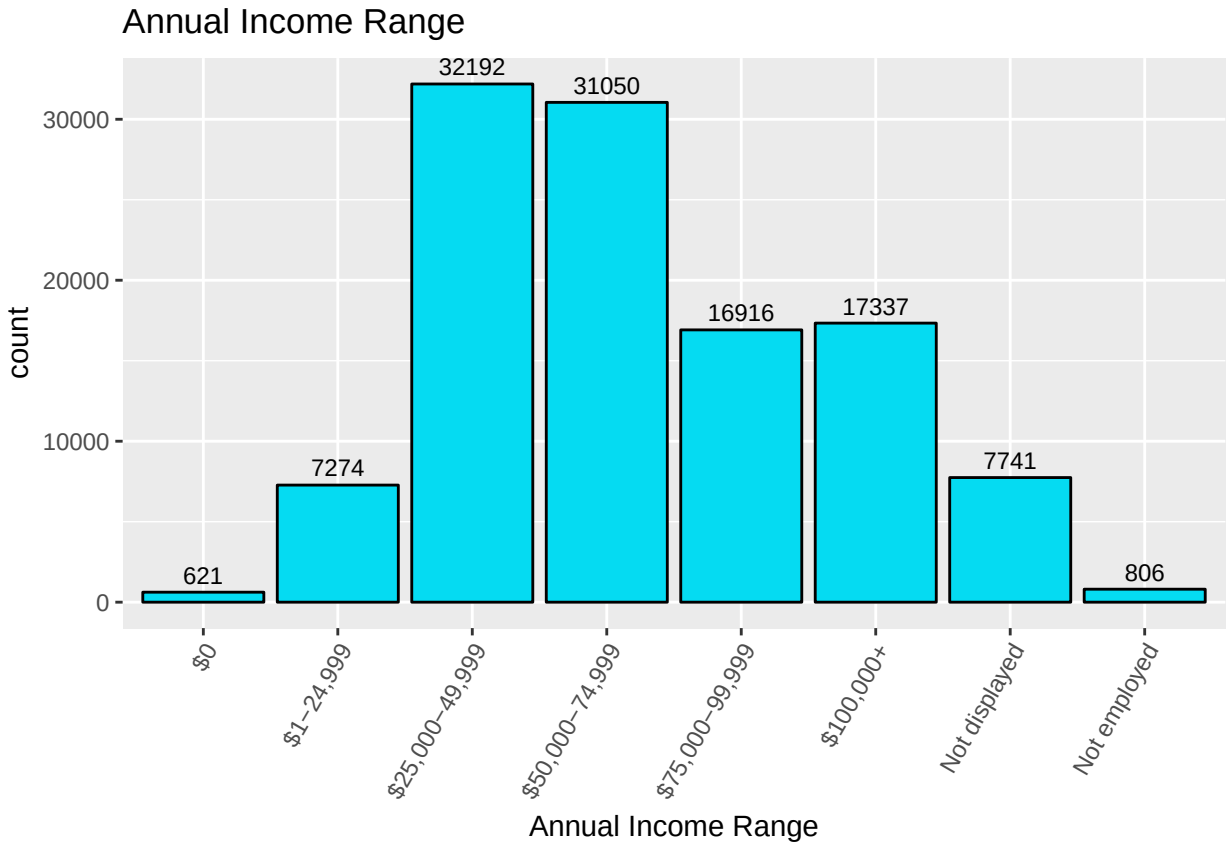
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    3200    4667    5608    6825 1750000
```



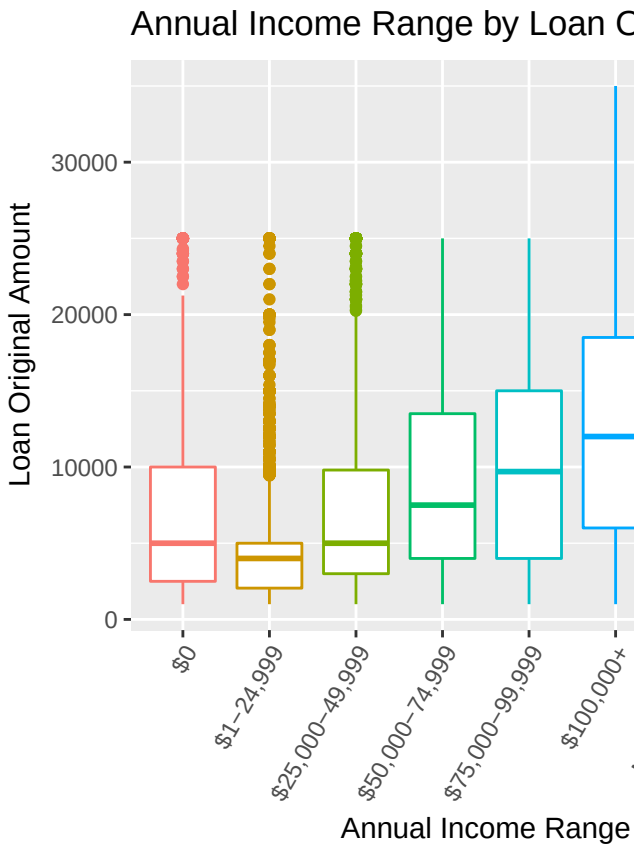
Because the stated monthly income is extremely skewed in the positive side (max income = 1750000), we need to limit the x-axis to display a more observable histogram distribution. After limiting the x-axis to 20000, We can see that the common range of stated monthly income falls between 3000 to 6000, which is in line with the IQR distribution shown.

Next, we will analyze the annual income range of users.

```
##                  freq percentage
## $0                621  0.5450380
## $1-24,999        7274  6.3842299
## $100,000+       17337 15.2163037
## $25,000-49,999  32192 28.2542107
## $50,000-74,999  31050 27.2519024
## $75,000-99,999  16916 14.8468013
## Not displayed    7741  6.7941055
## Not employed      806  0.7074085
```



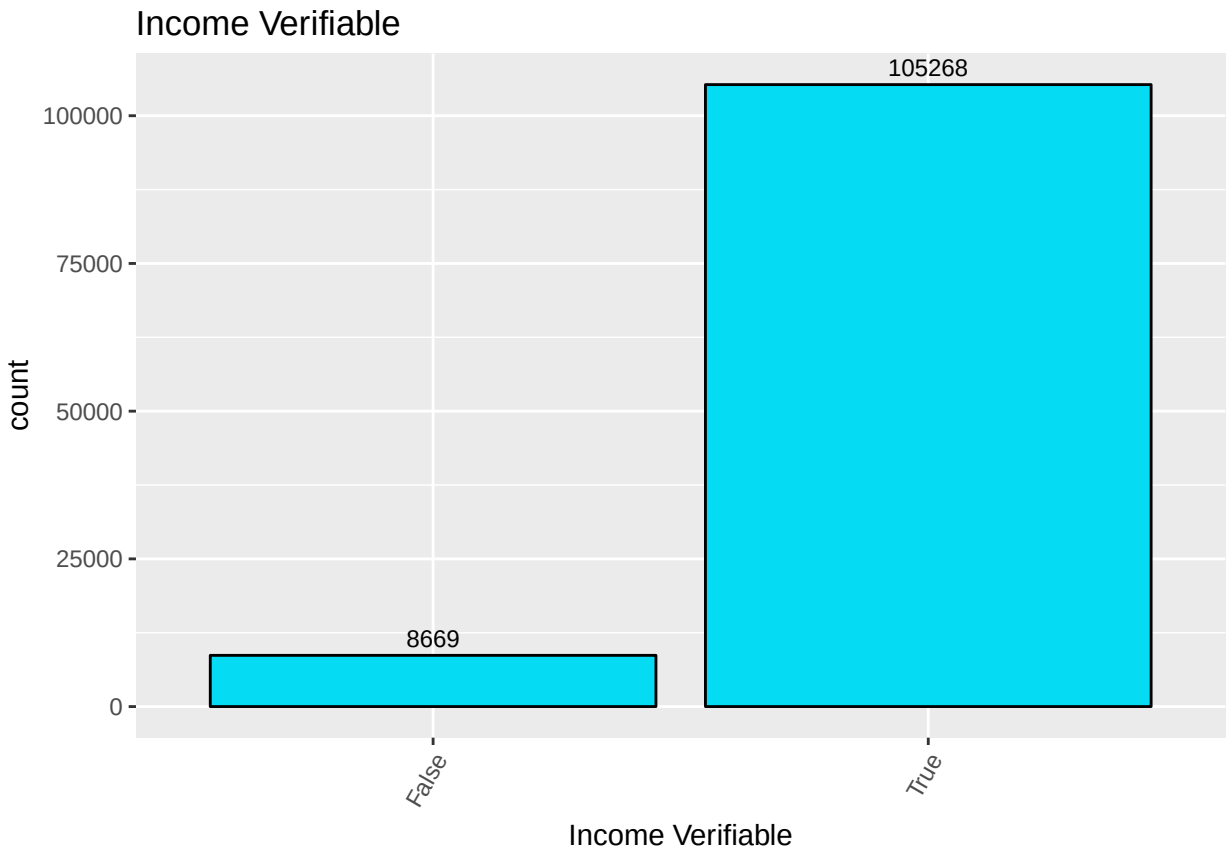
The annual income range tallies with the stated monthly income. In this case, 55% of the users are in the income range of \$25K to \$75K. Most of the users (>90%) have stated their income range, except for the category of \$0, "Not displayed", "Not employed".



Next, we are going to analyze the loan amount by annual income range.
From the boxplot above, we can observe that users with higher income tend to take a higher loan amount.

After getting some ideas about the income range of users, we may be interested to know whether the stated income range is already verified or not.

```
##         freq percentage
## False   8669   7.608591
## True  105268  92.391409
```



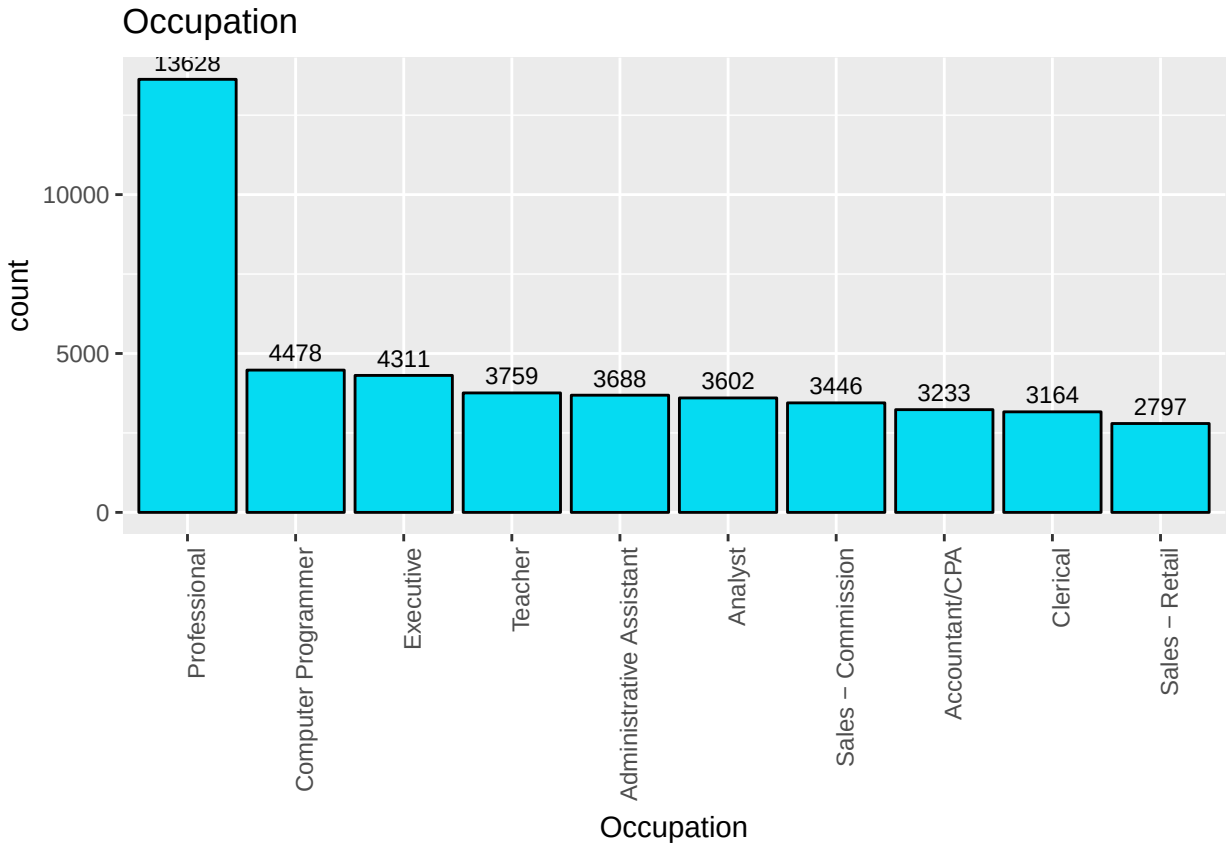
Fortunately, most of the users (>90%) have their income verified in the Prosper platform.

After understanding the distribution of income-related variables, we will move on to analyze the occupation of users.

```
##  Factor w/ 68 levels "","Accountant/CPA",..: 37 43 37 52 21 43 50 29 24 24 ...
```

```
##                                    freq  percentage
## Other                             28617 25.11651176
## Professional                      13628 11.96099599
## Computer Programmer                4478  3.93024215
## Executive                          4311  3.78366992
## Teacher                            3759  3.29919166
## Administrative Assistant           3688  3.23687652
## Analyst                            3602  3.16139621
##                                    3588  3.14910872
## Sales - Commission                 3446  3.02447844
## Accountant/CPA                     3233  2.83753302
## Clerical                           3164  2.77697324
## Sales - Retail                     2797  2.45486541
## Skilled Labor                      2746  2.41010383
## Retail Management                  2602  2.28371820
## Nurse (RN)                         2489  2.18454058
## Construction                       1790  1.57104365
## Truck Driver                       1675  1.47011068
## Laborer                            1595  1.39989643
## Police Officer/Correction Officer  1578  1.38497591
## Civil Service                      1457  1.27877687
```

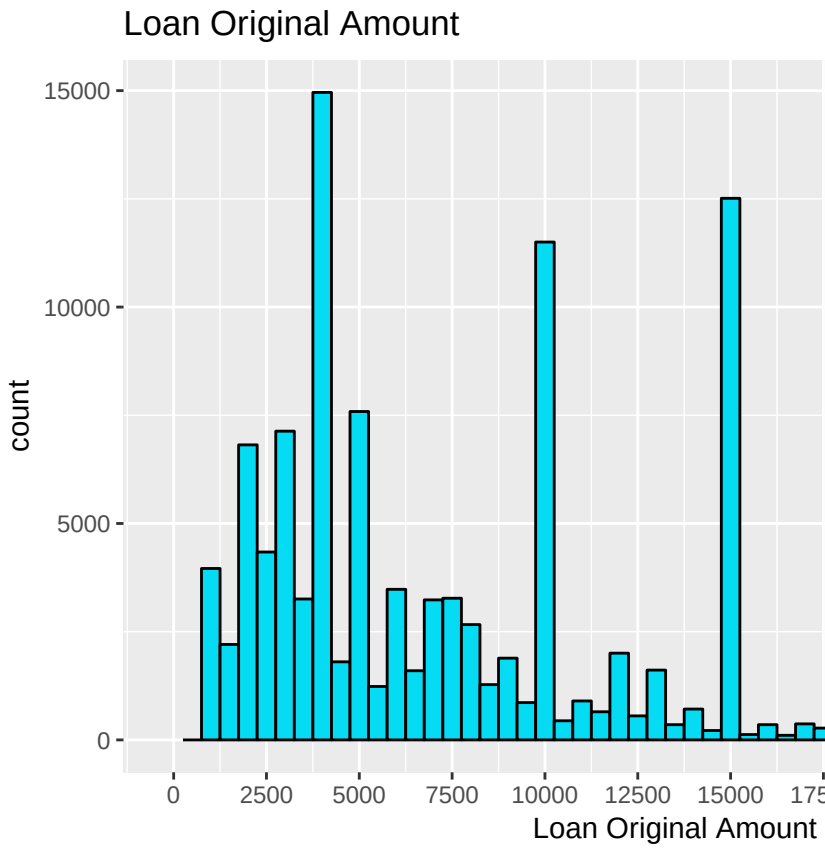
```
## Engineer - Mechanical               1406  1.23401529
## Military Enlisted                   1272  1.11640644
## Food Service Management             1239  1.08744306
## Engineer - Electrical               1125  0.98738777
## Food Service                        1123  0.98563241
## Medical Technician                  1117  0.98036634
## Attorney                            1046  0.91805120
## Tradesman - Mechanic                 951  0.83467179
## Social Worker                        741  0.65035941
## Postal Service                       627  0.55030412
## Professor                            557  0.48886665
## Realtor                              543  0.47657916
## Doctor                               494  0.43357294
## Nurse (LPN)                          492  0.43181758
## Nurse's Aide                         491  0.43093991
## Tradesman - Electrician              477  0.41865241
## Waiter/Waitress                      436  0.38266761
## Fireman                              422  0.37038012
## Scientist                            372  0.32649622
## Military Officer                     346  0.30367659
## Bus Driver                           316  0.27734625
## Principal                            312  0.27383554
## Teacher's Aide                       276  0.24223913
## Pharmacist                           257  0.22556325
## Student - College Graduate Student   245  0.21503111
## Landscaping                          236  0.20713201
## Engineer - Chemical                  225  0.19747755
## Investor                             214  0.18782310
## Architect                            213  0.18694542
## Pilot - Private/Commercial           199  0.17465792
## Clergy                               196  0.17202489
## Student - College Senior             188  0.16500347
## Car Dealer                           180  0.15798204
## Chemist                              145  0.12726331
## Psychologist                         145  0.12726331
## Biologist                            125  0.10970975
## Religious                            124  0.10883207
## Flight Attendant                     123  0.10795440
## Homemaker                            120  0.10532136
## Tradesman - Carpenter                120  0.10532136
## Student - College Junior             112  0.09829994
## Tradesman - Plumber                  102  0.08952316
## Student - College Sophomore           69  0.06055978
## Dentist                               68  0.05968211
## Student - College Freshman            41  0.03598480
## Student - Community College           28  0.02457498
## Judge                                 22  0.01930892
## Student - Technical School            16  0.01404285
```



There are many categories (68) in occupation. As such, to visualize it in a bar chart, we will need to: - First, subset the occupation data to exclude 'Other' and '' - Get the top 10 occupation from the subset of occupation for a more accurate reflection of top occupations.

Afterwards, we are able to visually comprehend the distribution of occupations among users.

From the summary table and the bar chart, we can observe that: - At least a quarter of the users do not wish to indicate their occupation. This is shown by occupation of 'Other' (>25%) and '' (>3%) - The next top occupation (>11%) is 'Professional'. This is a very broad categorization of occupation, which may suggest that some users do not want to specifically indicate their occupation. - Combining both of the observations above, we can understand that around 40% of users have not stated their occupation specifically in the Prosper platform. - Interestingly, 'Computer Programmer' comes in 3rd. This occurence is probably because Prosper is an advanced fintech platform that many programmers are interested to try.

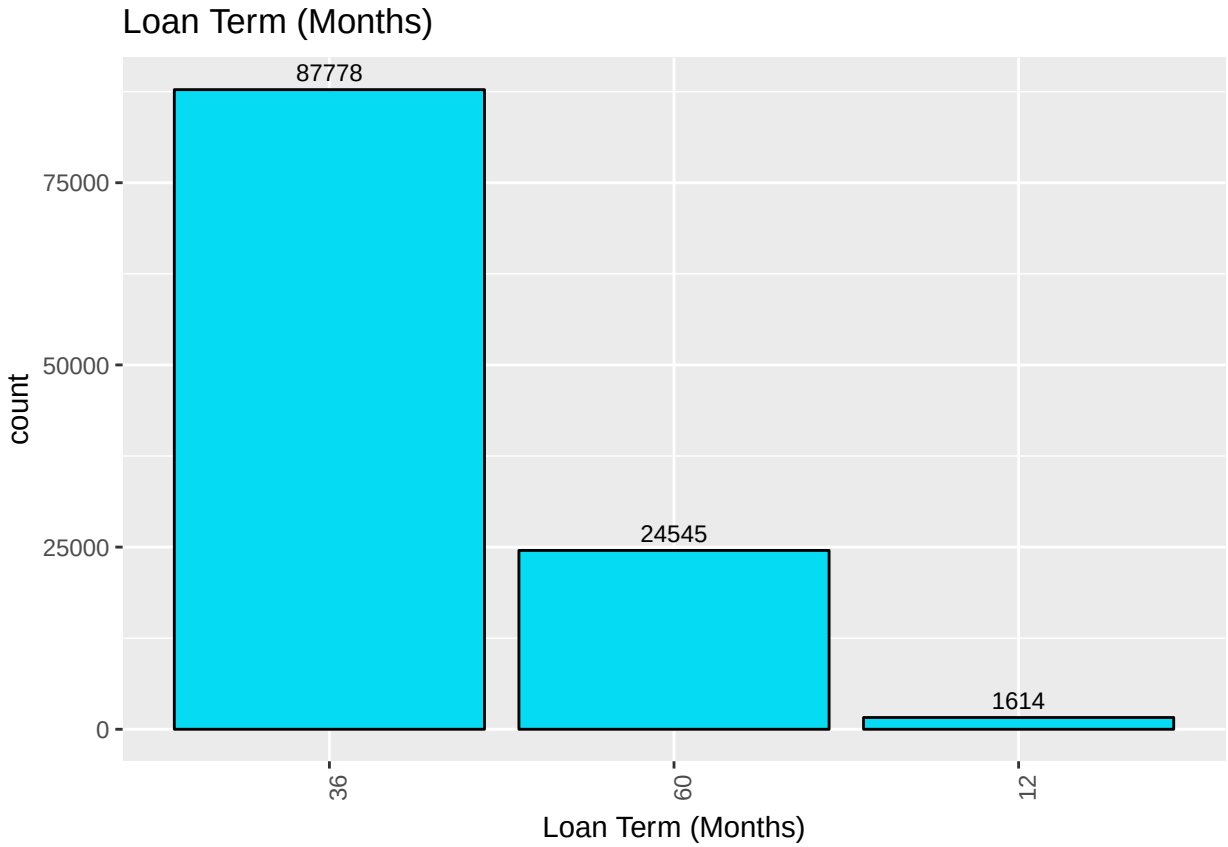


Next, we will analyze Loan Original Amount variable.

From the histogram above, We can see that there are some prominent peaks for original amount of loan at 4000, 10000, 15000. Some minor peaks include 2000, 3000, 5000. Interestingly, for higher amount of loan, 20000 and 25000 are the most common ones.

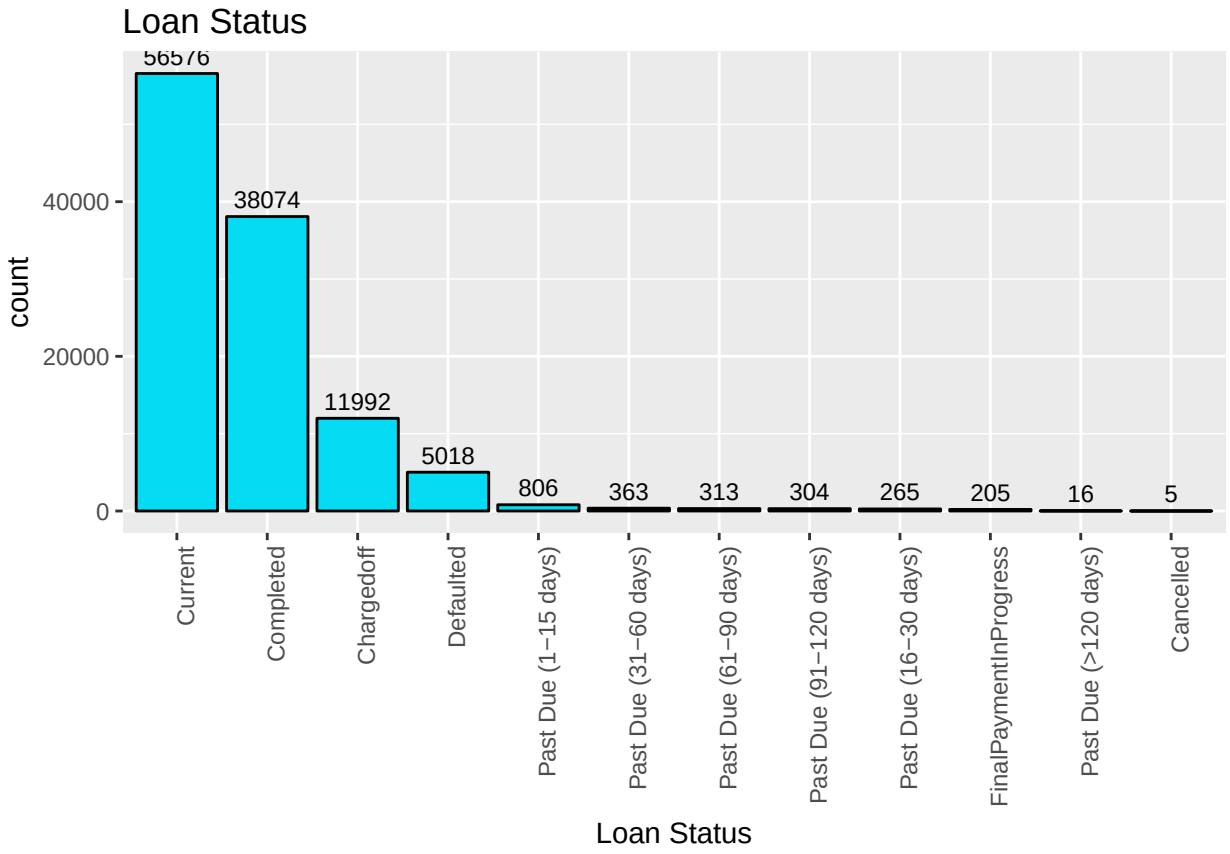
Next, we will look at loan term variable.

```
##     freq percentage
## 12  1614   1.416572
## 36 87778  77.040821
## 60 24545  21.542607
```



77% of the users have a loan term of 36 months (3 years), followed by 60 months and 12 months.

```
##                           freq  percentage
## Cancelled                    5  0.00438839
## Chargedoff               11992 10.52511476
## Completed                38074 33.41671274
## Current                  56576 49.65551138
## Defaulted                 5018  4.40418828
## FinalPaymentInProgress     205  0.17992399
## Past Due (>120 days)        16  0.01404285
## Past Due (1-15 days)       806  0.70740848
## Past Due (16-30 days)      265  0.23258467
## Past Due (31-60 days)      363  0.31859712
## Past Due (61-90 days)      313  0.27471322
## Past Due (91-120 days)     304  0.26681412
```



Fortunately, around 82% of the users have "Current" or "Completed" loan status. These two statuses can be considered as positive categories of loan status because they mean that the users have either completed their loan payment or on track in making their loan payment.

On the flip side, around 15% of the users have "Chargedoff" or "Defaulted" status. These two statuses can be considered as negative statuses because they mean that the users are not able to make the loan payment and defaulted their loan.

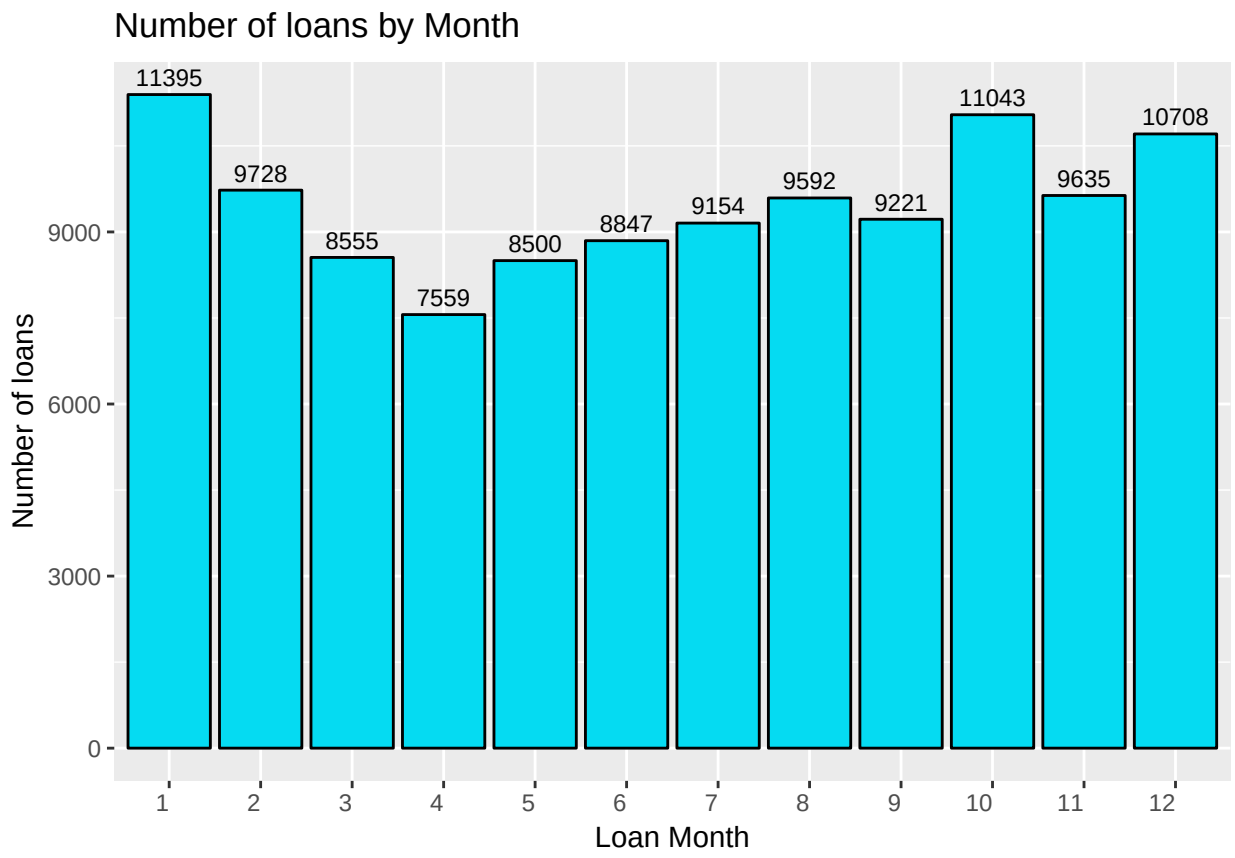
```
##       freq  percentage
## 2005    22  0.01930892
## 2006  5906  5.18356636
## 2007 11460 10.05819005
## 2008 11552 10.13893643
## 2009  2047  1.79660690
## 2010  5652  4.96063614
## 2011 11228  9.85456875
## 2012 19553 17.16123823
## 2013 34345 30.14385143
## 2014 12172 10.68309680
```

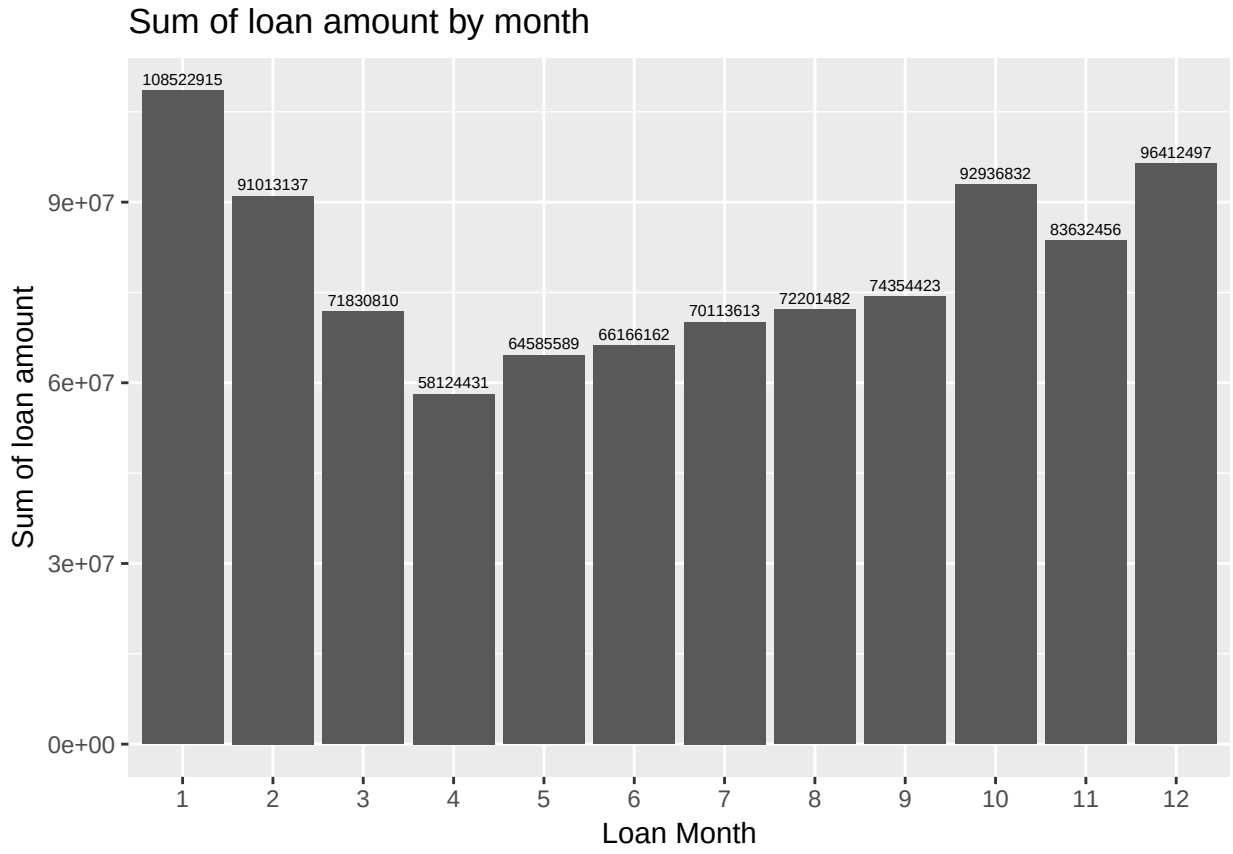
```
##     freq percentage
## 1  11395  10.001141
## 2   9728   8.538052
## 3   8555   7.508535
## 4   7559   6.634368
## 5   8500   7.460263
## 6   8847   7.764817
## 7   9154   8.034265
```

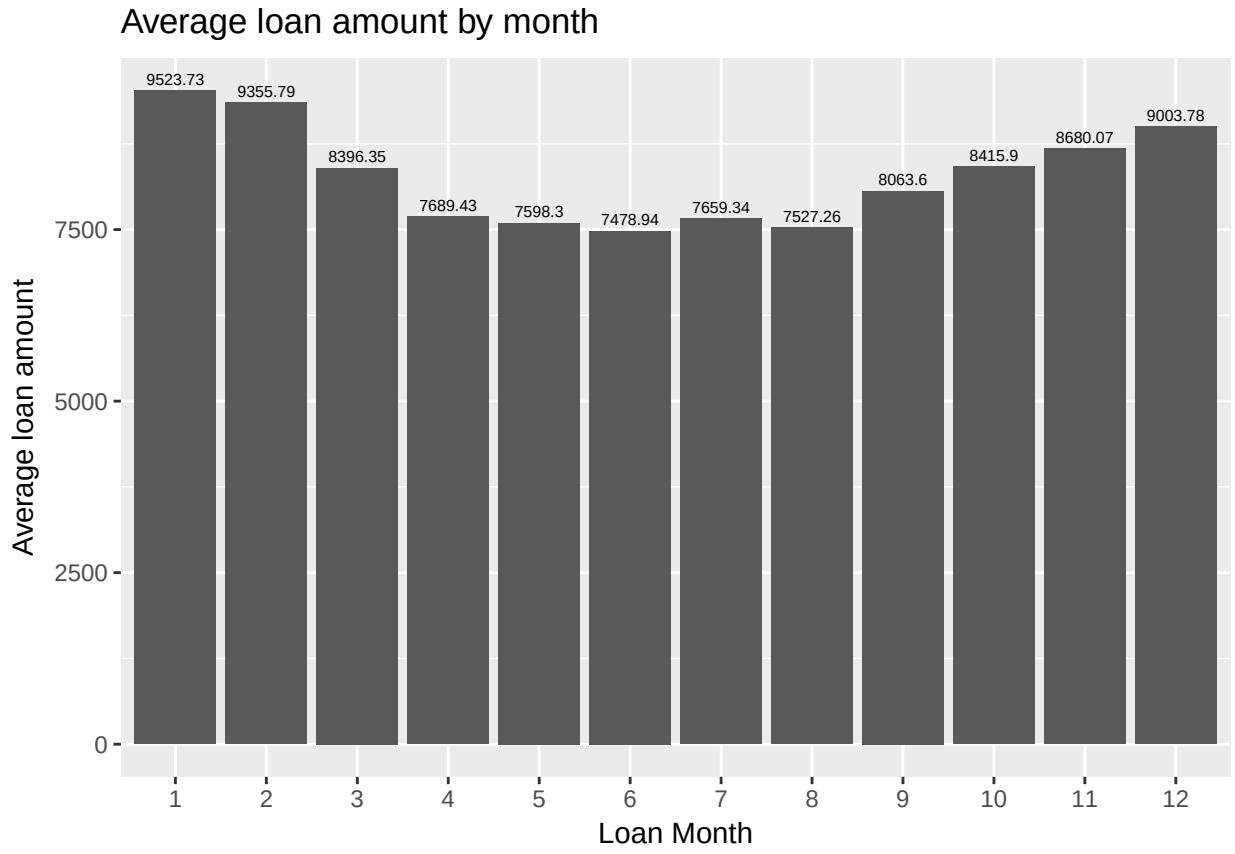
```
## 8   9592   8.418688
## 9   9221   8.093069
## 10 11043   9.692198
## 11  9635   8.456428
## 12 10708   9.398176
```

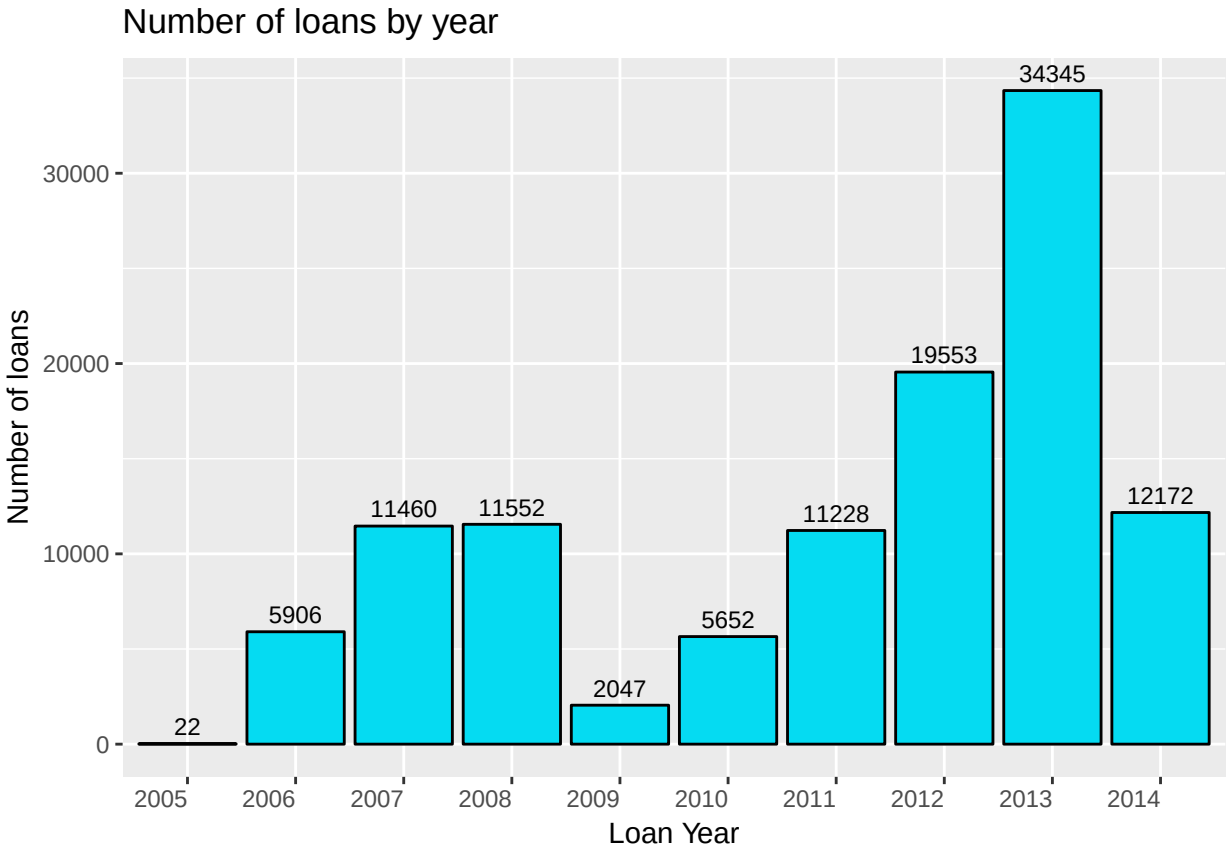
```
## # A tibble: 10 × 4
##    LoanYear LoanTotalSum LoanTotalAvg LoanCount
##      <fctr>        <int>        <dbl>     <int>
## 1      2014    144995536    11912.220     12172
## 2      2013    362170278    10545.066     34345
## 3      2012    153175116     7833.842     19553
## 4      2011     75138013     6692.021     11228
## 5      2010     26940486     4766.540      5652
## 6      2009      8914396     4354.859      2047
## 7      2008     69561850     6021.628     11552
## 8      2007     80787786     7049.545     11460
## 9      2006     28132199     4763.325      5906
## 10     2005        78687     3576.682        22
```

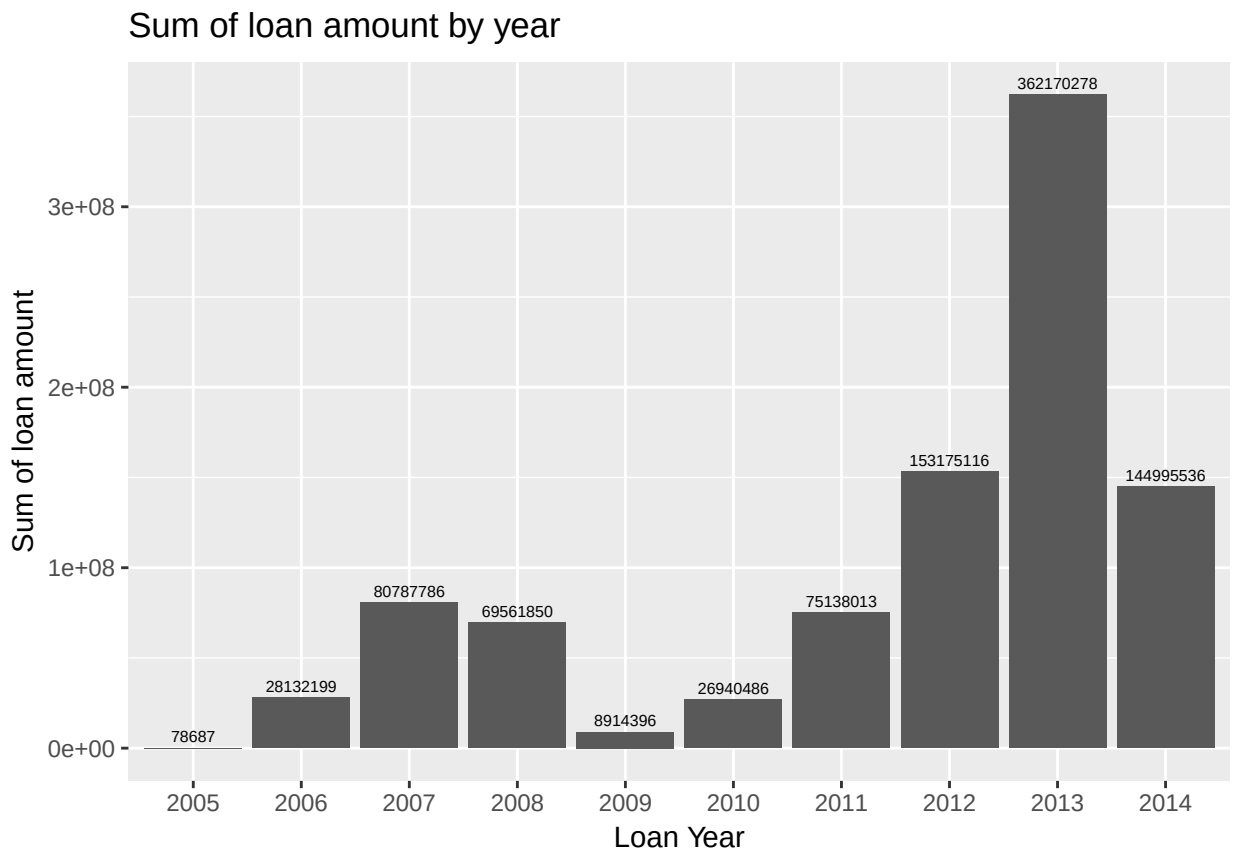
```
## # A tibble: 12 × 4
##    LoanMonth LoanTotalSum LoanTotalAvg LoanCount
##       <fctr>        <int>        <dbl>     <int>
## 1         12     96412497     9003.782     10708
## 2         11     83632456     8680.068      9635
## 3         10     92936832     8415.904     11043
## 4          9     74354423     8063.596      9221
## 5          8     72201482     7527.260      9592
## 6          7     70113613     7659.342      9154
## 7          6     66166162     7478.938      8847
## 8          5     64585589     7598.305      8500
## 9          4     58124431     7689.434      7559
## 10         3     71830810     8396.354      8555
## 11         2     91013137     9355.791      9728
## 12         1    108522915     9523.731     11395
```

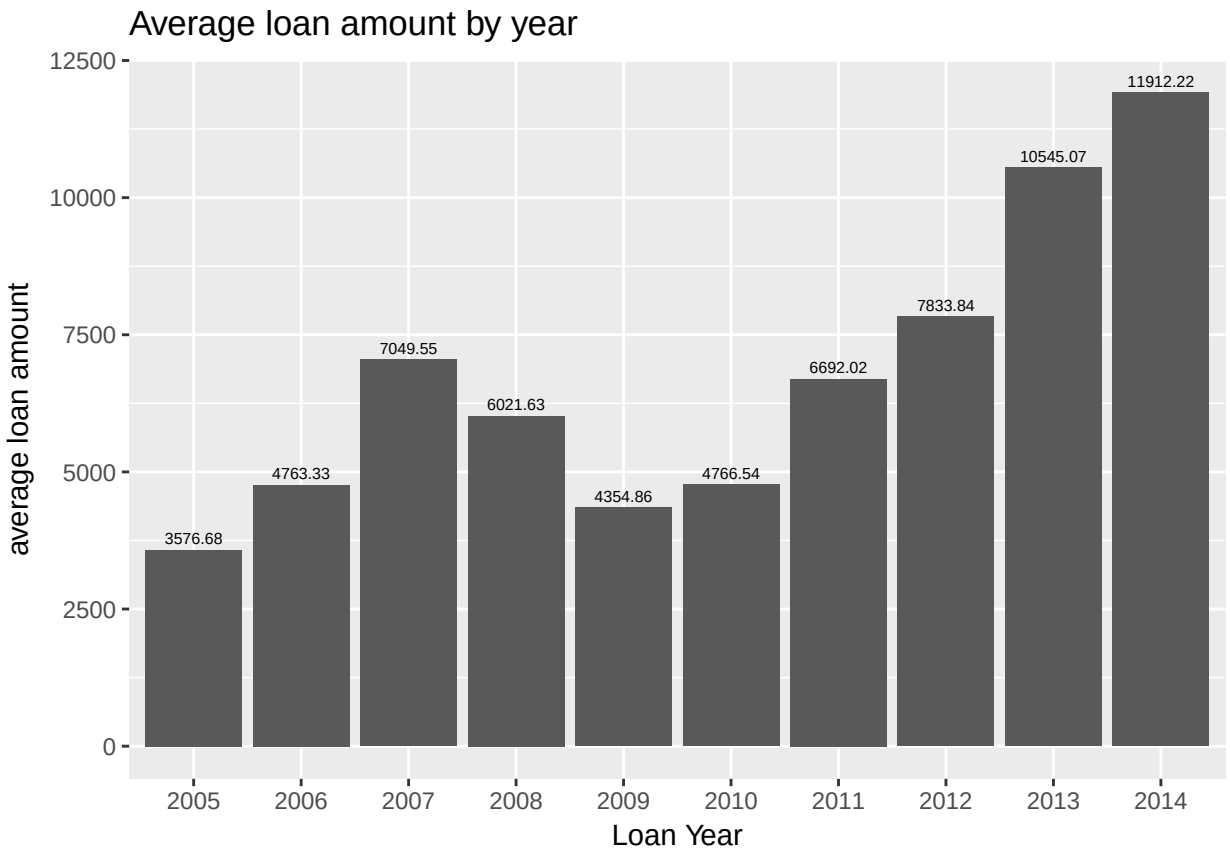

Number of loans by Month
11395
9728
8555
7559
8500
8847
9154
9592
9221
11043
9635
10708
9000
6000
3000
0
Number of loans
1
2
3
4
5
6
7
8
9
10
11
12
Loan Month


Sum of loan amount by month
108522915
91013137
71830810
58124431
64585589
66166162
70113613
72201482
74354423
92936832
83632456
96412497
9e+07
6e+07
3e+07
0e+00
Sum of loan amount
1
2
3
4
5
6
7
8
9
10
11
12
Loan Month


Average loan amount by month
9523.73
9355.79
8396.35
7689.43
7598.3
7478.94
7659.34
7527.26
8063.6
8415.9
8680.07
9003.78
7500
5000
2500
0
Average loan amount
1
2
3
4
5
6
7
8
9
10
11
12
Loan Month


Number of loans by year
Number of loans
30000
20000
10000
0
22
5906
11460
11552
2047
5652
11228
19553
34345
12172
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
Loan Year

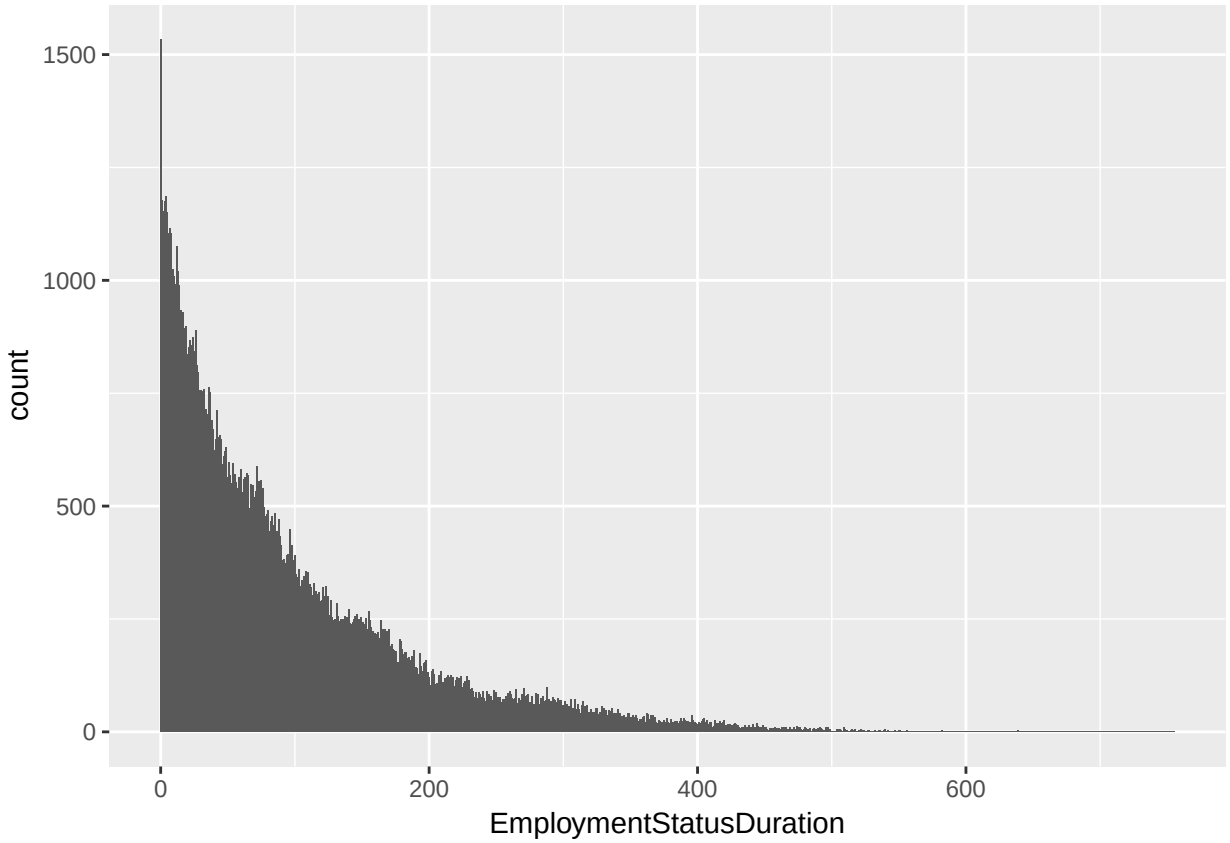

Sum of loan amount by year
Sum of loan amount
3e+08
2e+08
1e+08
0e+00
78687
28132199
80787786
69561850
8914396
26940486
75138013
153175116
362170278
144995536
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
Loan Year



From the graphs and tables above, we can see a very different trend between comparison of average and frequency or sum of loan amount by month or year.

For loan amount by month: - Both sum and average of loan amount show similar months in which there are some peaks. However, if we rank it by the measures (either average or sum or freq), the top months will be different: 1) For average loan amount: Jan, Feb, Oct, Nov, Dec have the most average loan amount (ordered from top to bottom). The average loan is generally lower in the month of April to August. 2) For sum or frequency of loan amount: Jan, Oct, Dec, Feb, Nov have the most average loan amount (ordered by top to bottom). There is an obvious dip in sum and frequency of loan amount in April.

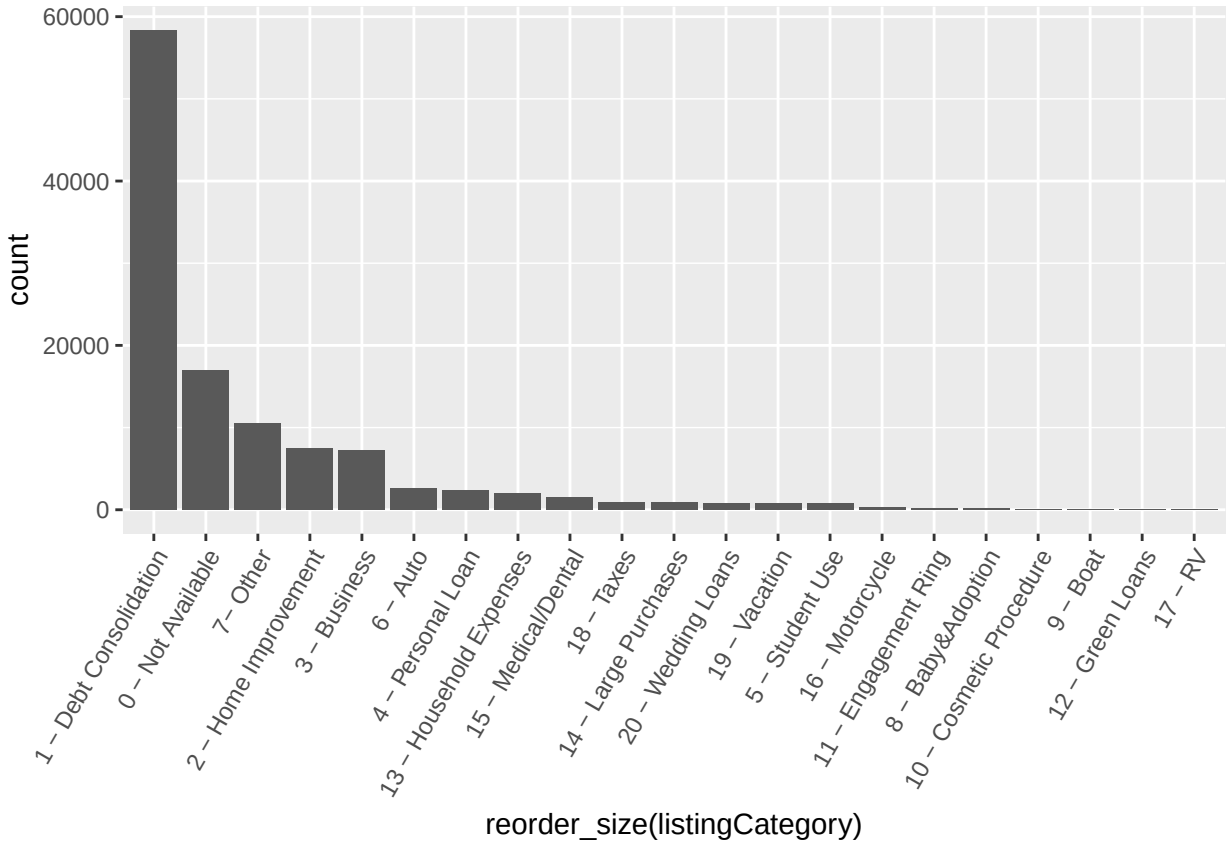
For loan amount by year: 1) Generally, the average loan amount is increasing throughout the years, only with some dips in 2008 to 2010. 2) The sum of loan amount peaks in 2013, with similar dips observed in 2008 to 2010. Perhaps, those dips are caused by financial crisis?

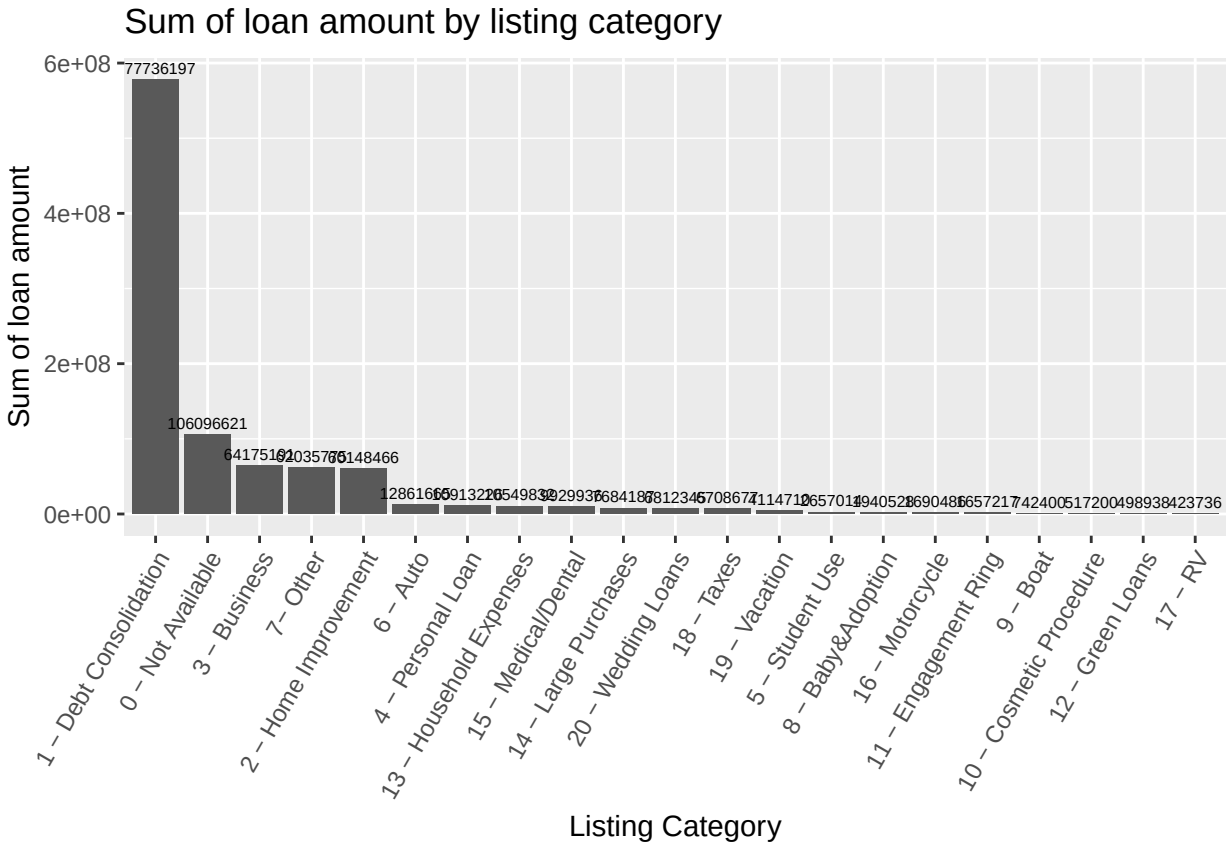


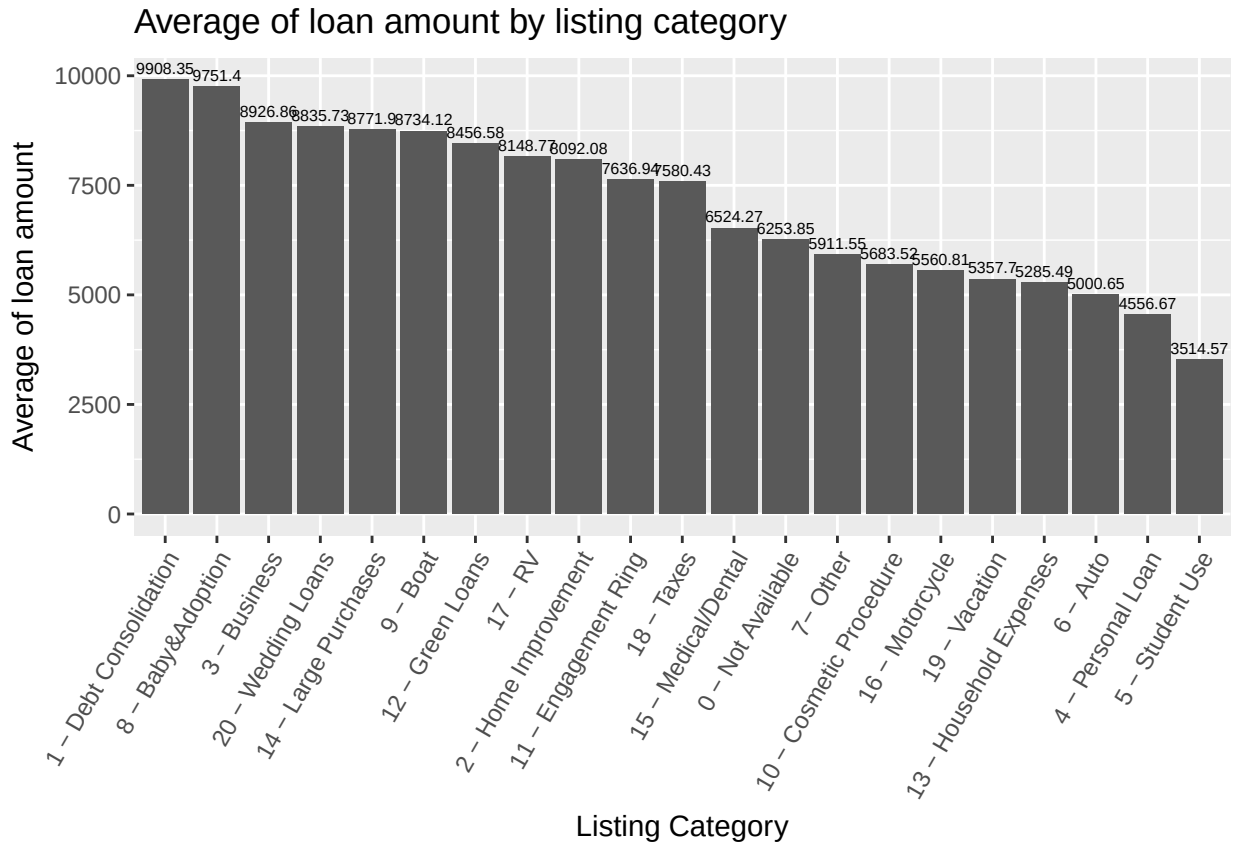

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   26.00   67.00   96.07  137.00  755.00    7625
```

EmploymentStatusDuration shows a right-skewed distribution with mean (96 months) > median (67 months). This shows that the employment status of users are more likely to stay unchanged in a shorter duration.

```
## [1] 113937
```


60000
40000
20000
0
count
1 – Debt Consolidation
0 – Not Available
7 – Other
2 – Home Improvement
3 – Business
6 – Auto
4 – Personal Loan
13 – Household Expenses
15 – Medical/Dental
18 – Taxes
14 – Large Purchases
20 – Wedding Loans
19 – Vacation
5 – Student Use
16 – Motorcycle
11 – Engagement Ring
8 – Baby&Adoption
10 – Cosmetic Procedure
9 – Boat
12 – Green Loans
17 – RV
reorder_size(listingCategory)


Sum of loan amount by listing category
Sum of loan amount
6e+08
4e+08
2e+08
0e+00
77736197
106096621
64175162
62035776
61148466
1 – Debt Consolidation
0 – Not Available
3 – Business
7 – Other
2 – Home Improvement
6 – Auto
4 – Personal Loan
13 – Household Expenses
15 – Medical/Dental
14 – Large Purchases
20 – Wedding Loans
18 – Taxes
19 – Vacation
5 – Student Use
8 – Baby&Adoption
16 – Motorcycle
11 – Engagement Ring
9 – Boat
10 – Cosmetic Procedure
12 – Green Loans
17 – RV
Listing Category




```
## Warning: 'plyr' namespace cannot be unloaded:
##   namespace 'plyr' is imported by 'scales', 'ggplot2' so cannot be unloaded
```

```
## # A tibble: 21 × 3
##            listingCategory SumDollarTotal CountTotal
##                      <fctr>          <int>      <int>
## 1   1 - Debt Consolidation      577736197      58308
## 2        0 - Not Available      106096621      16965
## 3             3 - Business       64175191       7189
## 4                7- Other       62035775      10494
## 5     2 - Home Improvement       60148466       7433
## 6                 6 - Auto       12861665       2572
## 7        4 - Personal Loan       10913226       2395
## 8  13 - Household Expenses       10549832       1996
## 9      15 - Medical/Dental        9929936       1522
## 10    14 - Large Purchases        7684187        876
## # ... with 11 more rows
```

```
## # A tibble: 21 × 3
##           listingCategory AvgDollarTotal CountTotal
##                     <fctr>          <dbl>      <int>
## 1  1 - Debt Consolidation       9908.352      58308
## 2       8 - Baby&Adoption       9751.397        199
## 3            3 - Business       8926.859       7189
## 4      20 - Wedding Loans       8835.726        771
## 5    14 - Large Purchases       8771.903        876
## 6                9 - Boat       8734.118         85
```

```
## 7        12 - Green Loans       8456.576         59
## 8                17 - RV       8148.769         52
## 9    2 - Home Improvement       8092.085       7433
## 10   11 - Engagement Ring       7636.945        217
## # ... with 11 more rows
```

```
##                           freq  percentage
## 1 - Debt Consolidation  58308 51.17564970
## 0 - Not Available       16965 14.88980753
## 7- Other                10494  9.21035309
## 2 - Home Improvement     7433  6.52378069
## 3 - Business             7189  6.30962725
## 6 - Auto                 2572  2.25738785
## 4 - Personal Loan        2395  2.10203885
## 13 - Household Expenses  1996  1.75184532
## 15 - Medical/Dental      1522  1.33582594
## 18 - Taxes                885  0.77674504
## 14 - Large Purchases      876  0.76884594
## 20 - Wedding Loans        771  0.67668975
## 19 - Vacation             768  0.67405672
## 5 - Student Use           756  0.66352458
## 16 - Motorcycle           304  0.26681412
## 11 - Engagement Ring      217  0.19045613
## 8 - Baby&Adoption         199  0.17465792
## 10 - Cosmetic Procedure    91  0.07986870
## 9 - Boat                   85  0.07460263
## 12 - Green Loans           59  0.05178300
## 17 - RV                    52  0.04563926
```

Based on the results shown above, if we look at the bar chart showing the frequency of listing categories, we can observe that around 51% of the listing categories are of Debt Consolidation. Therefore, it does not really tell us much insight as debt consolidation is a very generic term for loan repayment.

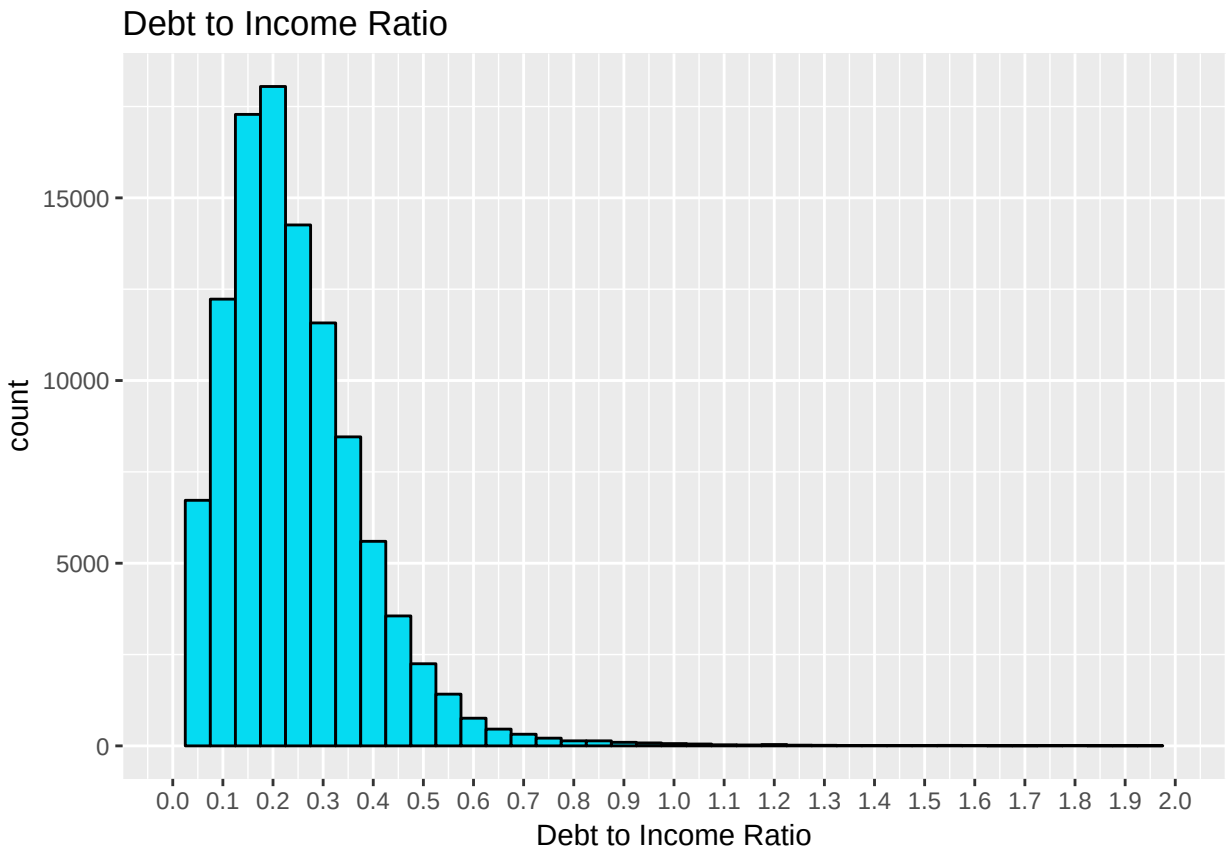
As such, by observing the listing category by average loan amount, we can observe some interesting insights. The top 10 listing categories are: Debt Consolidation, Baby&Adoption, Business, Wedding Loans, Large Purchases, Boat, Green Loans, RV, Home Improvement, Engagement Ring.

Some interesting observations: - Marriage seems to play a huge role in loan: Baby&Adoption, Wedding Loans, Engagement Ring - Home improvement has a higher frequency than Business. - Beyond the top 10, Household Expenses, Auto, Vacation, Medical/Dental have significant frequency number as well.

Combining all the three points above, we can see that most of the users are willing to take loan in category such as family, wedding, housing, health, business, and lifestyle.

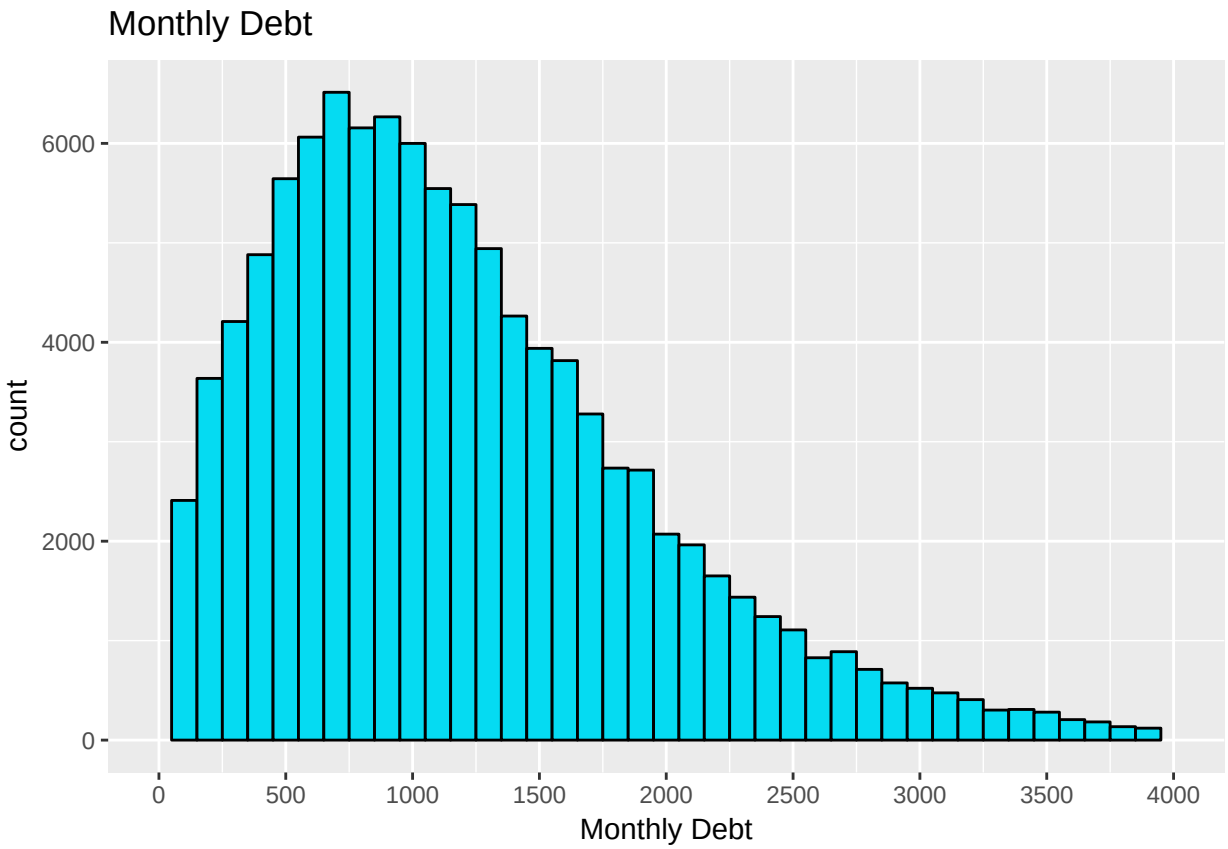
```
ggplot(aes(x=DebtToIncomeRatio), data=loan) +
    geom_histogram(fill='#05DBF2',
                   color='black',
                   binwidth=0.05) +
    scale_x_continuous(breaks=seq(0,2,0.1),limits=c(0,2)) +
    labs(title="Debt to Income Ratio",
         x='Debt to Income Ratio')
```

```
## Warning: Removed 9033 rows containing non-finite values (stat_bin).
```



From the histogram above, we can see that most of the users have debt that is around 10% to 30% of their income. It makes me wonder who will take loan which is 50% or more above their income? Maybe, this is worth investigating in the multivariate analysis in later part of the project.

the next step is to estimate the monthly debt by multplying the debt to income ratio with the stated monthly in-



come.

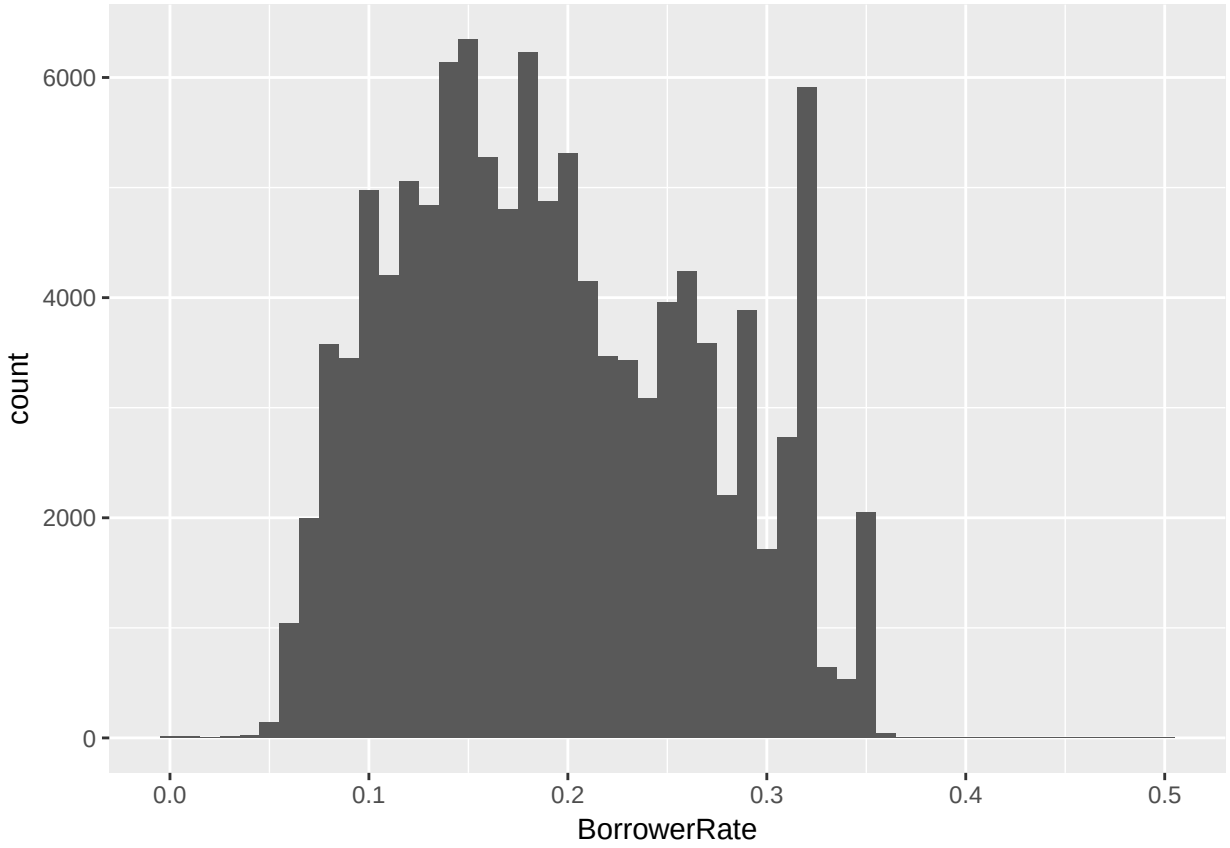
```
##      Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##       0.0    637.5   1062.0   1225.0   1622.0 171000.0     8554
```

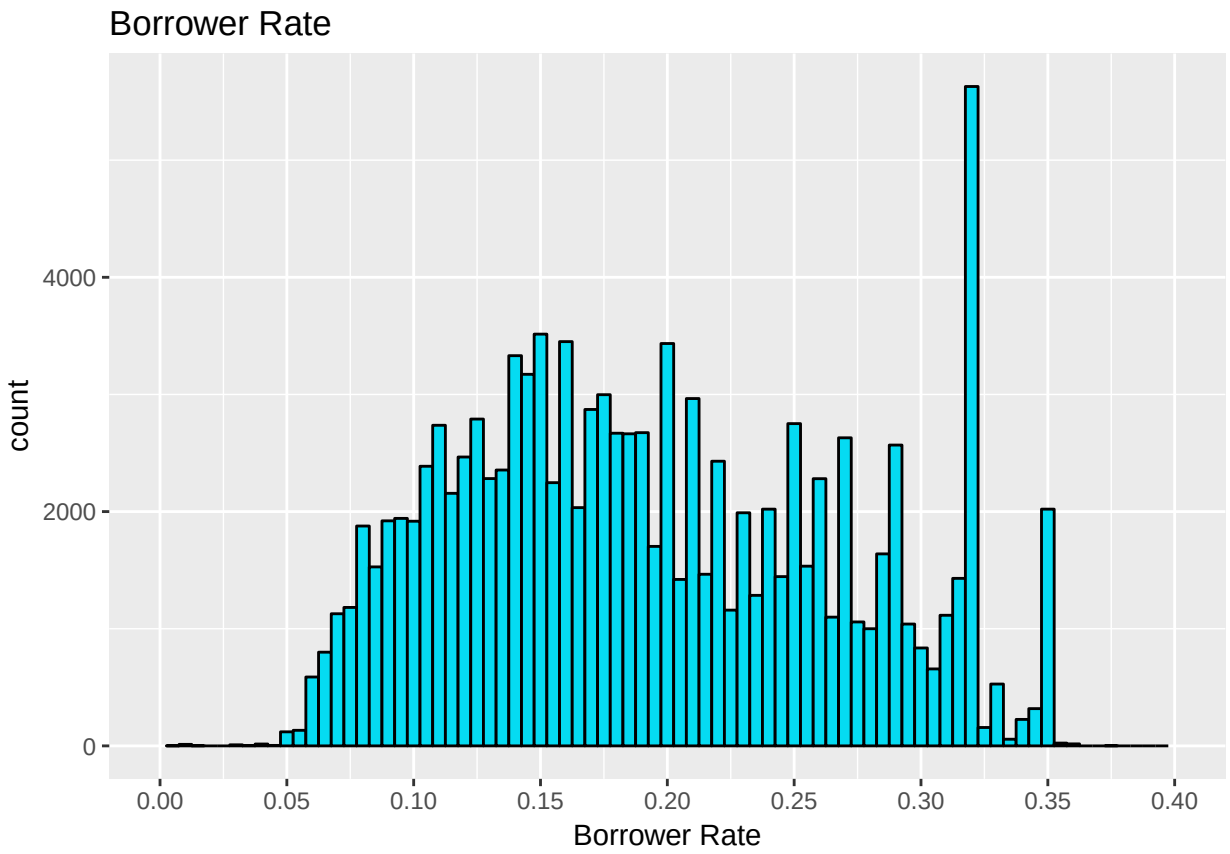
From the distribution of monthly debt above, it can be seen that most of the users have monthly debt around 300 to 1500. This is in line with the debt-to-income ratio, which is around 10% to 30% of the monthly income (around $3000 to $6000) for most of the users.

```
##  num [1:113937] 0.192 0.145 0.0755 0.0925 0.1355 ...
```

```
## [1] "double"
```

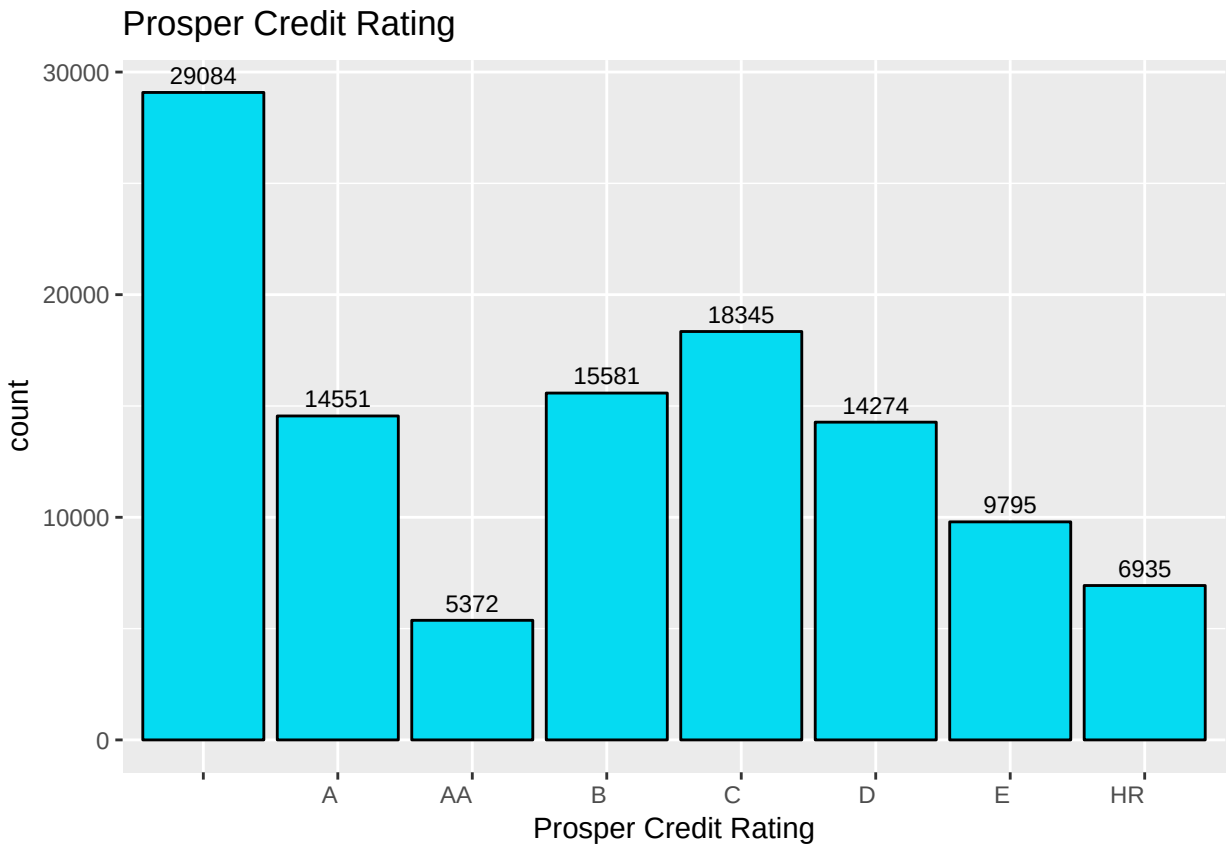
```
## [1] "numeric"
```


6000
4000
2000
0
count
0.0
0.1
0.2
0.3
0.4
0.5
BorrowerRate

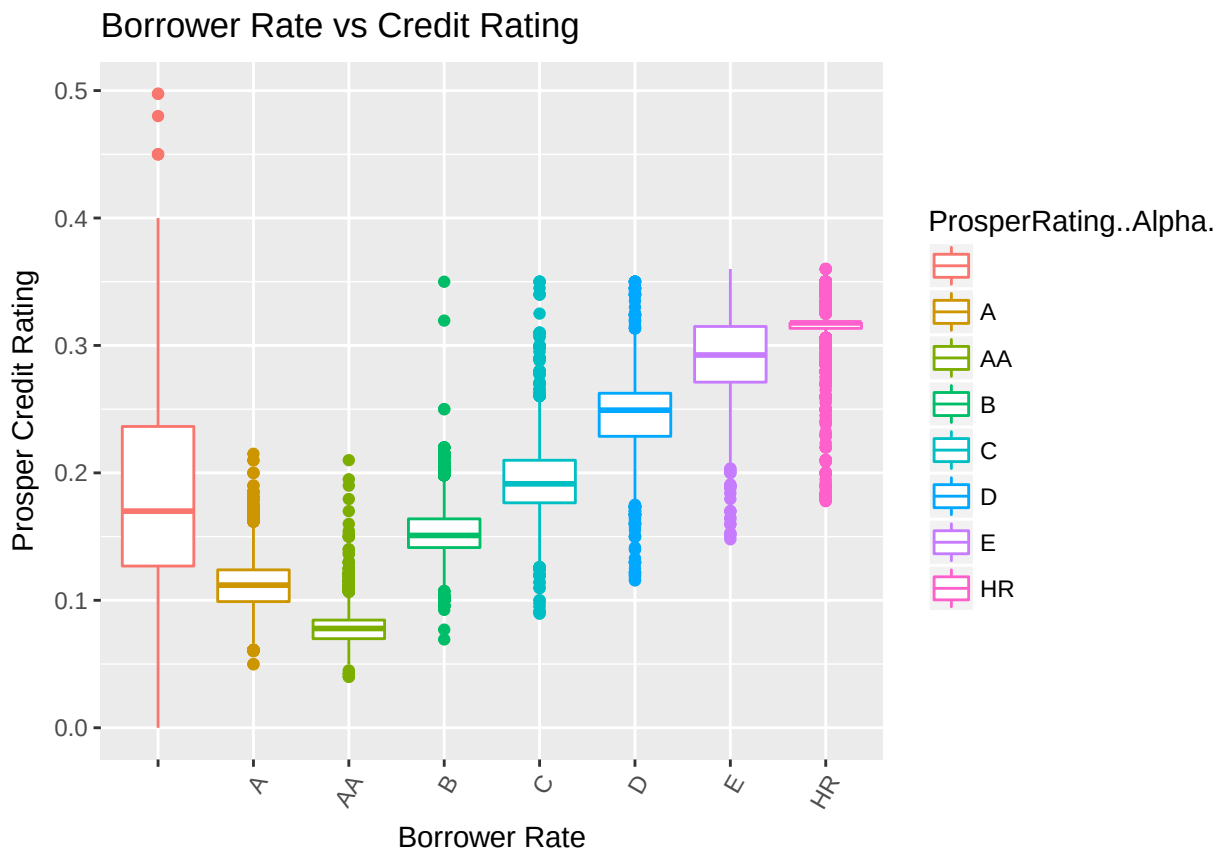


The distribution of borrower rate is somewhat bell-curve like with slightly right-skewed trend and some intermittent spikes throughout the trend. There is a very obvious mega spike at around 31%. After some online research (http://www.lendingmemo.com/rates-fees-lending-club-prosper/),we can understand that users with 31% borrower rate usually falls within Prosper credit rating of E or HR.

```
##     freq percentage
##    29084  25.526387
## C  18345  16.101003
## B  15581  13.675101
## A  14551  12.771093
## D  14274  12.527976
## E   9795   8.596856
## HR  6935   6.086697
## AA  5372   4.714886
```



Around 25% of the loans are not rated, while the rest of the ratings are almost similar at around 15%, except for the smaller proportion of AA, E, and HR rating.



From the boxplot above, we cannot really see any relationship between borrower rate and Prosper Credit Rating. Perhaps, the source (http://www.lendingmemo.com/rates-fees-lending-club-prosper/) mentioned are not really credible? The Borrower Rate seems to be an interesting area to explore in multivariate analysis.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0   660.0   680.0   685.6   720.0   880.0     591
```

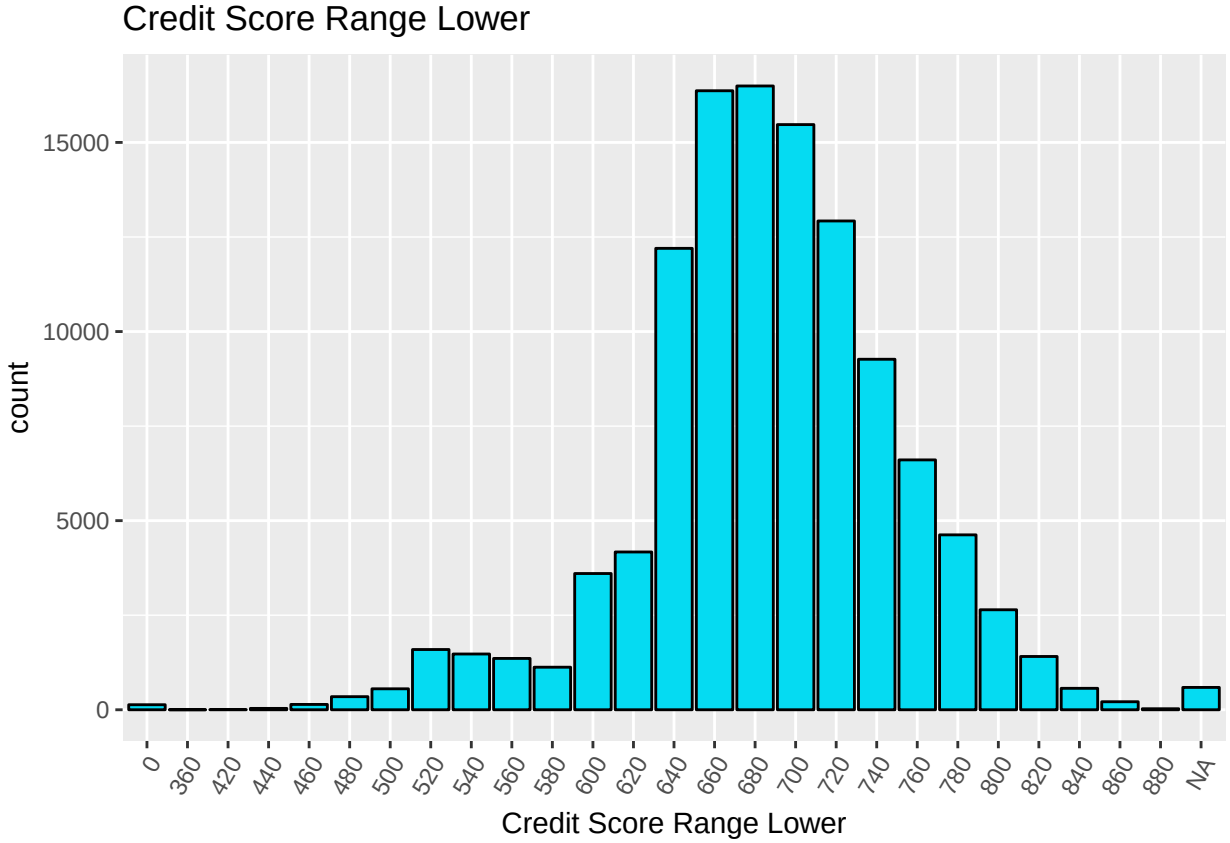
```
## [1] "integer"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0   660.0   680.0   685.6   720.0   880.0     591
```

```
##  [1] 640 680 820 740 680 760 540  NA 700 700 540 720 600 680 520 760 580
## [18]  NA 580 700
```

```
##      freq   percentage
## 0     133 1.173398e-01
## 360     1 8.822543e-04
## 420     5 4.411272e-03
## 440    36 3.176116e-02
## 460   141 1.243979e-01
## 480   346 3.052600e-01
## 500   554 4.887689e-01
## 520  1593 1.405431e+00
## 540  1474 1.300443e+00
## 560  1357 1.197219e+00
## 580  1125 9.925361e-01
## 600  3602 3.177880e+00
## 620  4172 3.680765e+00
```

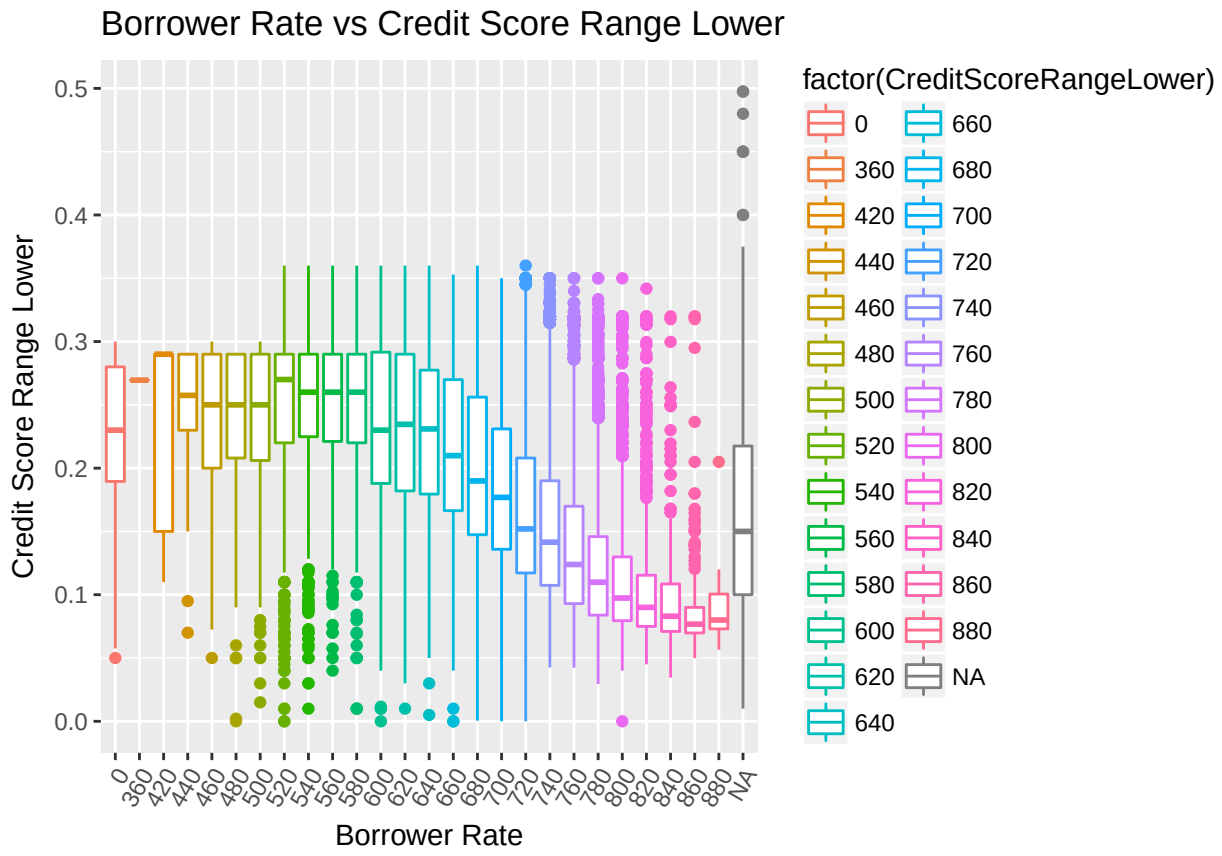
```
## 640 12199 1.076262e+01
## 660 16366 1.443897e+01
## 680 16492 1.455014e+01
## 700 15471 1.364936e+01
## 720 12923 1.140137e+01
## 740  9267 8.175851e+00
## 760  6606 5.828172e+00
## 780  4624 4.079544e+00
## 800  2644 2.332680e+00
## 820  1409 1.243096e+00
## 840   567 5.002382e-01
## 860   212 1.870379e-01
## 880    27 2.382087e-02
```



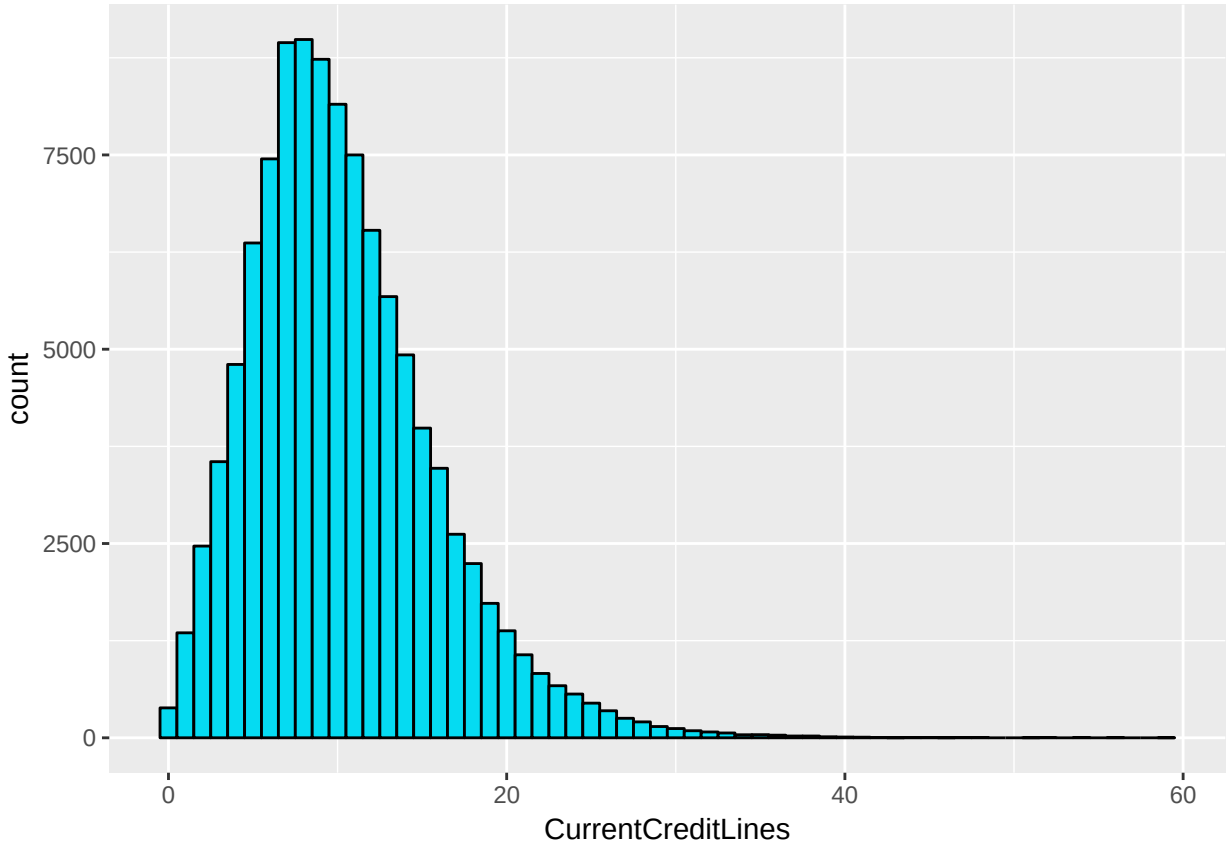
After summarizing CreditScoreRangeLower in frequency table, we can see that the variable is more appropriate to be converted as a discrete variable. As such, a bar chart is chosen over a histogram.

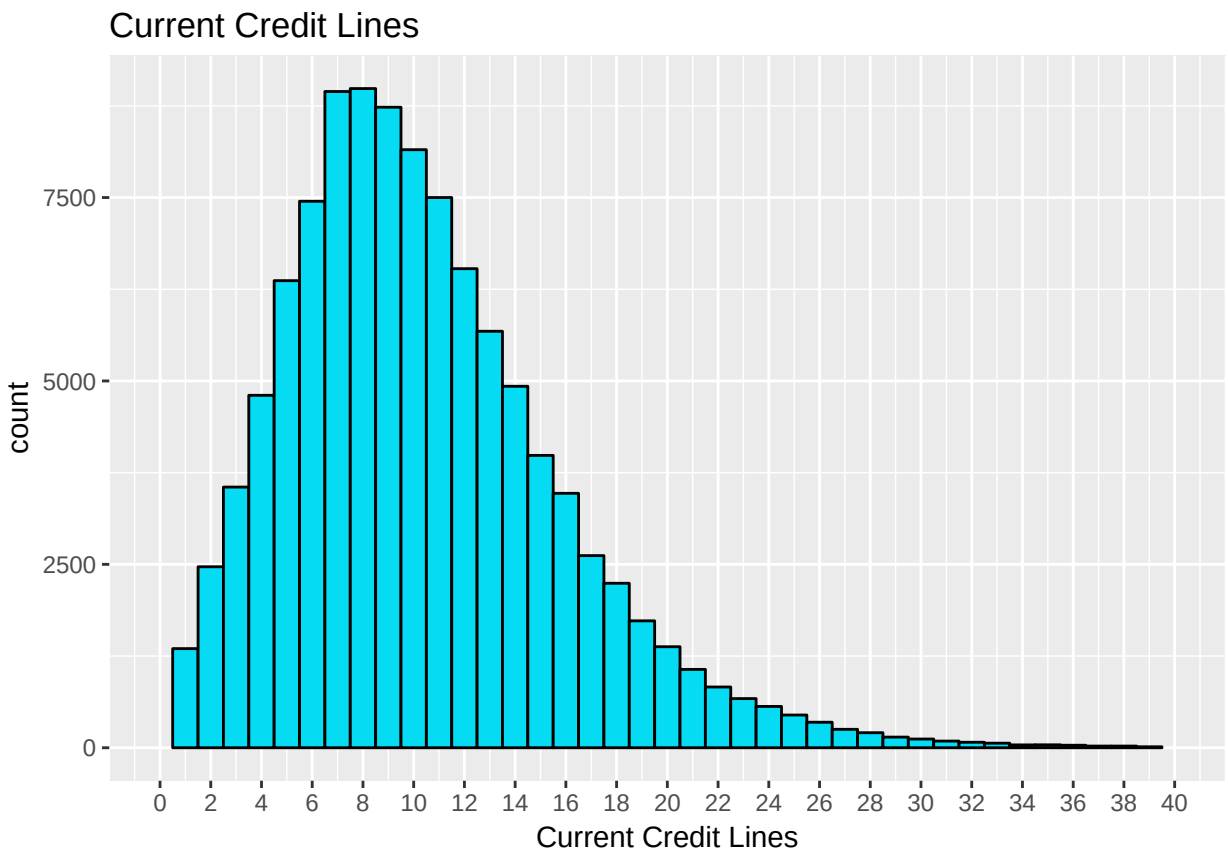
From the bar chart of credit score range lower, we can see that the most common credit score is between 640 to 740.

What if we analyze both borrower rate and credit score range lower together? Is there any relationship between them?

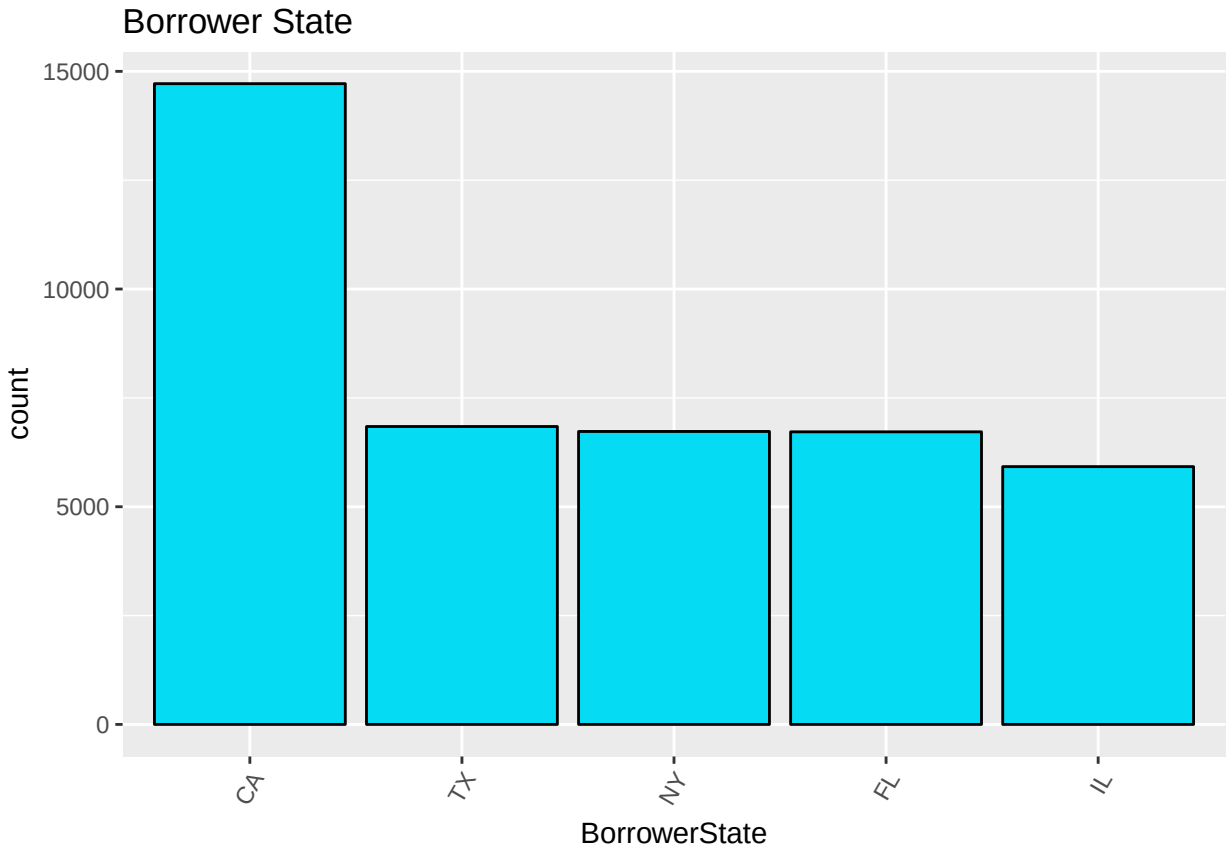


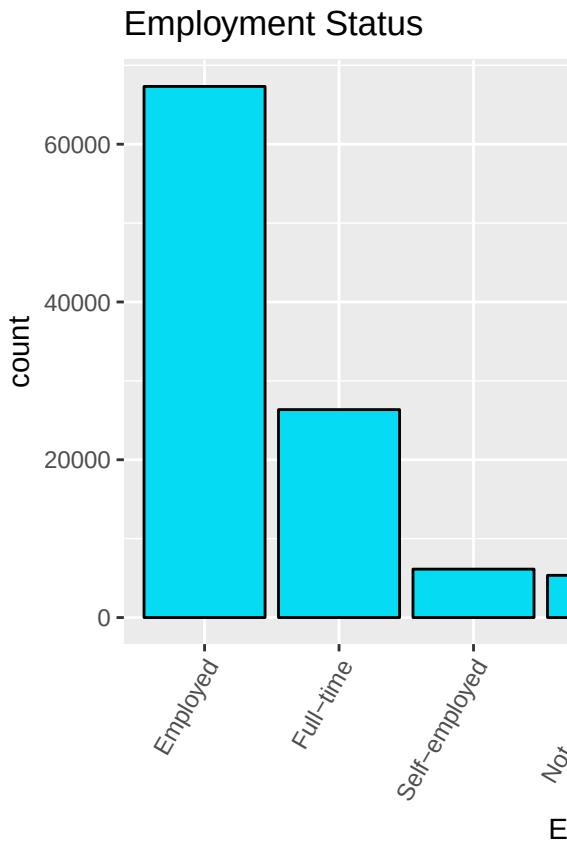
Now, we can see that borrower rate goes lower as the credit score goes higher. This means that users with higher credit score are usually perceived as more credible, and as such, may be more likely to be given a lower borrower rate.


7500
5000
2500
0
count
0
20
40
60
CurrentCreditLines



Most of the users have around 3 to 15 current credit lines with the median around 6 to 8 credit lines.


Borrower State
15000
10000
5000
0
count
CA
TX
NY
FL
IL
BorrowerState



The top borrower states are California, Texas, New York, Florida, and Illinois