Exploring Loan Data from Prosper

By Kevin Vo (http://kevinvo.xyz)

Based on Prosper website: *Prosper is America's first marketplace lending platform, with over $7 billion in funded loans. Prosper allows people to invest in each other in a way that is financially and socially rewarding. On Prosper, borrowers list loan requests between $2,000 and $35,000 and individual investors invest as little as $25 in each loan listing they select. Prosper handles the servicing of the loan on behalf of the matched borrowers and investors.*

Data Set contains information about Prosper loans:

```
loan %>% dim
```

```
## [1] 113937     81
```

- There are 113937 different loans which are categorized with 82 different variables.

```
loan %>% str
```

```
## 'data.frame':    113937 obs. of  81 variables:
##  $ ListingKey                          : Factor w/ 113066 levels "00003546482094282EF90E5",..: 7180 7193 6647 6
669 6686 6689 6699 6706 6687 6687 ...
##  $ ListingNumber                       : int  193129 1209647 81716 658116 909464 1074836 750899 768193 1023355
1023355 ...
##  $ ListingCreationDate                 : Factor w/ 113064 levels "2005-11-09 20:44:28.847000000",..: 14184 1118
94 6429 64760 85967 100310 72556 74019 97834 97834 ...
##  $ CreditGrade                         : Factor w/ 9 levels "","A","AA","B",..: 5 1 8 1 1 1 1 1 1 1 ...
##  $ Term                                : int  36 36 36 36 36 60 36 36 36 36 ...
##  $ LoanStatus                          : Factor w/ 12 levels "Cancelled","Chargedoff",..: 3 4 3 4 4 4 4 4 4 4 .
..
##  $ ClosedDate                          : Factor w/ 2803 levels "","2005-11-25 00:00:00",..: 1138 1 1263 1 1 1 1
1 1 1 ...
##  $ BorrowerAPR                         : num  0.165 0.12 0.283 0.125 0.246 ...
##  $ BorrowerRate                        : num  0.158 0.092 0.275 0.0974 0.2085 ...
##  $ LenderYield                         : num  0.138 0.082 0.24 0.0874 0.1985 ...
##  $ EstimatedEffectiveYield             : num  NA 0.0796 NA 0.0849 0.1832 ...
##  $ EstimatedLoss                       : num  NA 0.0249 NA 0.0249 0.0925 ...
##  $ EstimatedReturn                     : num  NA 0.0547 NA 0.06 0.0907 ...
##  $ ProsperRating..numeric.             : int  NA 6 NA 6 3 5 2 4 7 7 ...
##  $ ProsperRating..Alpha.               : Factor w/ 8 levels "","A","AA","B",..: 1 2 1 2 6 4 7 5 3 3 ...
##  $ ProsperScore                        : num  NA 7 NA 9 4 10 2 4 9 11 ...
##  $ ListingCategory..numeric.           : int  0 2 0 16 2 1 1 2 7 7 ...
##  $ BorrowerState                       : Factor w/ 52 levels "","AK","AL","AR",..: 7 7 12 12 25 34 18 6 16 16 .
..
##  $ Occupation                          : Factor w/ 68 levels "","Accountant/CPA",..: 37 43 37 52 21 43 50 29 24
24 ...
##  $ EmploymentStatus                    : Factor w/ 9 levels "","Employed",..: 9 2 4 2 2 2 2 2 2 2 ...
##  $ EmploymentStatusDuration            : int  2 44 NA 113 44 82 172 103 269 269 ...
##  $ IsBorrowerHomeowner                 : Factor w/ 2 levels "False","True": 2 1 1 2 2 2 1 1 2 2 ...
##  $ CurrentlyInGroup                    : Factor w/ 2 levels "False","True": 2 1 2 1 1 1 1 1 1 1 ...
##  $ GroupKey                            : Factor w/ 707 levels "","0034337690131242316873l",..: 1 1 335 1 1 1 1
1 1 1 ...
##  $ DateCreditPulled                    : Factor w/ 112992 levels "2005-11-09 00:30:04.487000000",..: 14347 1118
83 6446 64724 85857 100382 72500 73937 97888 97888 ...
##  $ CreditScoreRangeLower               : int  640 680 480 800 680 740 680 700 820 820 ...
##  $ CreditScoreRangeUpper               : int  659 699 499 819 699 759 699 719 839 839 ...
##  $ FirstRecordedCreditLine             : Factor w/ 11586 levels "","1947-08-24 00:00:00",..: 8639 6617 8927 224
7 9498 497 8265 7685 5543 5543 ...
##  $ CurrentCreditLines                  : int  5 14 NA 5 19 21 10 6 17 17 ...
##  $ OpenCreditLines                     : int  4 14 NA 5 19 17 7 6 16 16 ...
##  $ TotalCreditLinespast7years          : int  12 29 3 29 49 49 20 10 32 32 ...
##  $ OpenRevolvingAccounts               : int  1 13 0 7 6 13 6 5 12 12 ...
##  $ OpenRevolvingMonthlyPayment         : num  24 389 0 115 220 1410 214 101 219 219 ...
##  $ InquiriesLast6Months                : int  3 3 0 0 1 0 0 3 1 1 ...
##  $ TotalInquiries                      : num  3 5 1 1 9 2 0 16 6 6 ...
##  $ CurrentDelinquencies                : int  2 0 1 4 0 0 0 0 0 0 ...
##  $ AmountDelinquent                    : num  472 0 NA 10056 0 ...
##  $ DelinquenciesLast7Years             : int  4 0 0 14 0 0 0 0 0 0 ...
##  $ PublicRecordsLast10Years            : int  0 1 0 0 0 0 0 1 0 0 ...
##  $ PublicRecordsLast12Months           : int  0 0 NA 0 0 0 0 0 0 0 ...
##  $ RevolvingCreditBalance              : num  0 3989 NA 1444 6193 ...
##  $ BankcardUtilization                 : num  0 0.21 NA 0.04 0.81 0.39 0.72 0.13 0.11 0.11 ...
##  $ AvailableBankcardCredit             : num  1500 10266 NA 30754 695 ...
##  $ TotalTrades                         : num  11 29 NA 26 39 47 16 10 29 29 ...
##  $ TradesNeverDelinquent..percentage.  : num  0.81 1 NA 0.76 0.95 1 0.68 0.8 1 1 ...
```

```
##  $ TradesOpenedLast6Months            : num  0 2 NA 0 2 0 0 0 1 1 ...
##  $ DebtToIncomeRatio                  : num  0.17 0.18 0.06 0.15 0.26 0.36 0.27 0.24 0.25 0.25 ...
##  $ 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 ...
##  $ StatedMonthlyIncome                : num  3083 6125 2083 2875 9583 ...
##  $ LoanKey                            : Factor w/ 113066 levels "00003683605746079487FF7",..: 100337 69837 463
03 70776 71387 86505 91250 5425 908 908 ...
##  $ TotalProsperLoans                  : int  NA NA NA NA 1 NA NA NA NA NA ...
##  $ TotalProsperPaymentsBilled         : int  NA NA NA NA 11 NA NA NA NA NA ...
##  $ OnTimeProsperPayments              : int  NA NA NA NA 11 NA NA NA NA NA ...
##  $ ProsperPaymentsLessThanOneMonthLate: int  NA NA NA NA 0 NA NA NA NA NA ...
##  $ ProsperPaymentsOneMonthPlusLate    : int  NA NA NA NA 0 NA NA NA NA NA ...
##  $ ProsperPrincipalBorrowed           : num  NA NA NA NA 11000 NA NA NA NA NA ...
##  $ ProsperPrincipalOutstanding        : num  NA NA NA NA 9948 ...
##  $ ScorexChangeAtTimeOfListing        : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ LoanCurrentDaysDelinquent          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ LoanFirstDefaultedCycleNumber      : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ LoanMonthsSinceOrigination         : int  78 0 86 16 6 3 11 10 3 3 ...
##  $ LoanNumber                         : int  19141 134815 6466 77296 102670 123257 88353 90051 121268 121268 .
..
##  $ 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
1821 1649 1666 1813 1813 ...
##  $ LoanOriginationQuarter             : Factor w/ 33 levels "Q1 2006","Q1 2007",..: 18 8 2 32 24 33 16 16 33 3
3 ...
##  $ MemberKey                          : Factor w/ 90831 levels "00003397697413387CAF966",..: 11071 10302 33781
54939 19465 48037 60448 40951 26129 26129 ...
##  $ MonthlyLoanPayment                 : num  330 319 123 321 564 ...
##  $ LP_CustomerPayments                : num  11396 0 4187 5143 2820 ...
##  $ LP_CustomerPrincipalPayments       : num  9425 0 3001 4091 1563 ...
##  $ LP_InterestandFees                 : num  1971 0 1186 1052 1257 ...
##  $ LP_ServiceFees                     : num  -133.2 0 -24.2 -108 -60.3 ...
##  $ LP_CollectionFees                  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ LP_GrossPrincipalLoss              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ LP_NetPrincipalLoss                : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ LP_NonPrincipalRecoverypayments    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ PercentFunded                      : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ Recommendations                    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ InvestmentFromFriendsCount         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ InvestmentFromFriendsAmount        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Investors                          : int  258 1 41 158 20 1 1 1 1 1 ...
```

Before analyzing the data, I have noticed there are two very important variables: CreditGrade and ProsperRating. CreditGrade is used to rate the performance of a loan before 2009. And Prosperating is used for the same purpose since 2009. However, we don't have any variable to mention about year. Let's create the new variable as called as `ListingYear` :

```
loan$ListingYear <- loan$ListingCreationDate %>% year
```

```
table(loan$CreditGrade, loan$ListingYear)
```

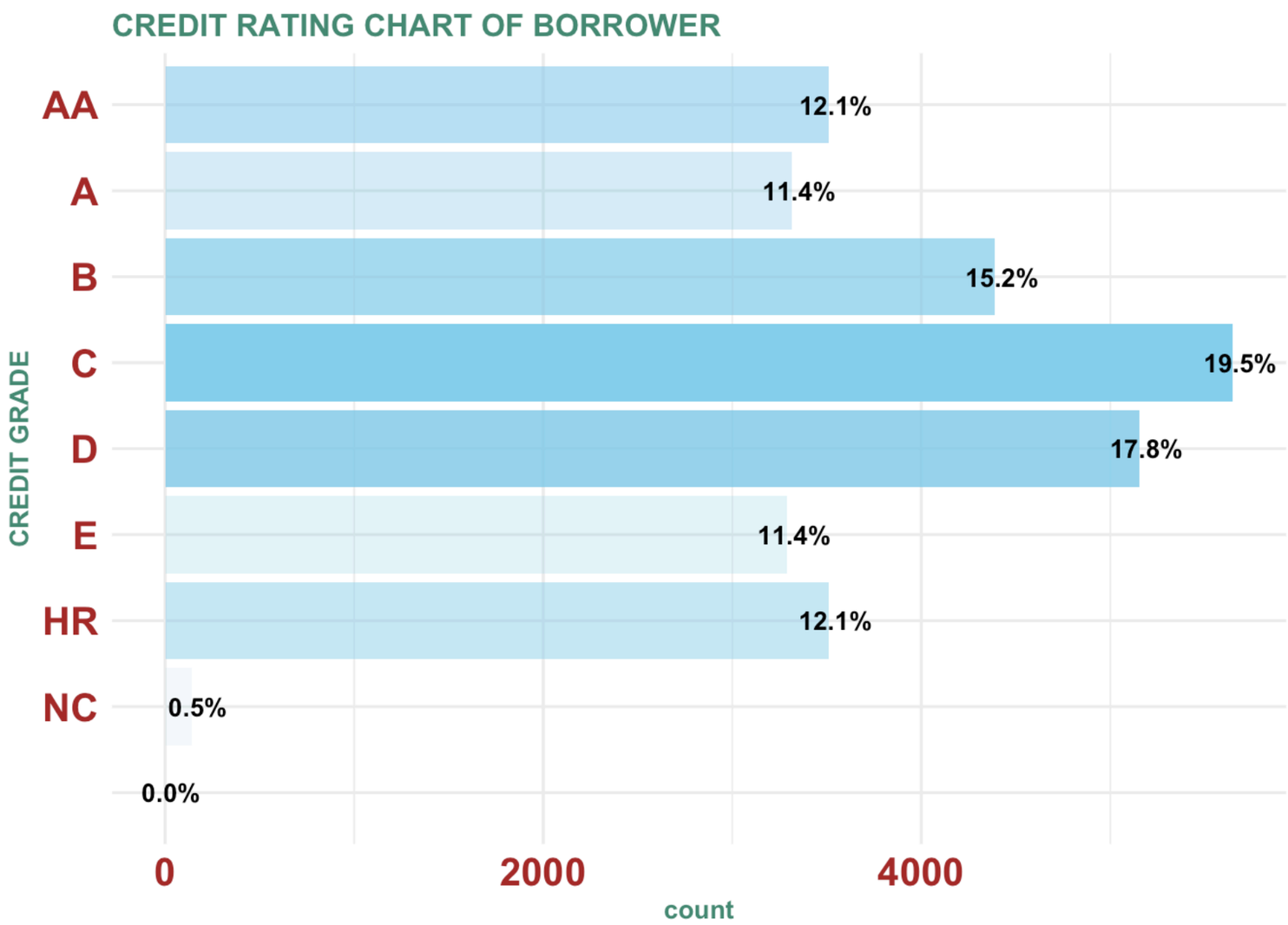
```
##
##        2005  2006  2007  2008  2009  2010  2011  2012  2013  2014
##           0     0    13   103  2193  5530 11442 19556 35413 10734
##   A       1   486  1344  1482     2     0     0     0     0     0
##   AA     12   536  1348  1612     1     0     0     0     0     0
##   B       2   633  1744  2007     3     0     0     0     0     0
##   C       3   934  2325  2382     5     0     0     0     0     0
##   D       1  1002  2174  1974     2     0     0     0     0     0
##   E       1  1207  1196   885     0     0     0     0     0     0
##   HR      2  1306  1382   818     0     0     0     0     0     0
##   NC      1   109    31     0     0     0     0     0     0     0
```

As it is claimed, CreditGrade is used before 2009.

```
table(loan$ProsperRating..numeric.,loan$ListingYear)
```

```
##
##      2005 2006 2007 2008 2009 2010 2011 2012 2013 2014
##   1     0    0    0    0  231  886 1180 3597  955   86
##   2     0    0    0    0  211  635 2321 1903 3969  756
##   3     0    0    0    0  372 1325 3301 3217 5089  970
##   4     0    0    0    0  425  569  912 3945 9522 2972
##   5     0    0    0    0  110  548 1669 3041 7837 2376
##   6     0    0    0    0  509  936 1507 2766 6210 2623
##   7     0    0    0    0  320  631  552 1087 1831  951
```

And ProspeRating is used since 2009. But why does Prosper have to change from CreditCrade to ProsperRating?



If I were one of the investors in Prosper lending platform, I would complain about why there are so many loans are rated as A and AA. I think they were not doing well to assess the risk of all the loans. And investors would not satisfy if they have to rely on this risk-assessing system. It may be the reason why Prosper has changed from using CreditGrade system to ProsperRating.

Since Prosper has changed to new system, therefore we should subset our data with the timeline from 2009 onward.

```
loan <- loan %>% subset(ListingYear >2008)
loan$ListingYear <- loan$ListingYear %>% ordered(levels = 2009:2014)
```

Univariate Plots Section

In this section, we would like to analyze the characteristics of loan and credit risk.

```
loan %>% dim
```

```
## [1] 84881    82
```

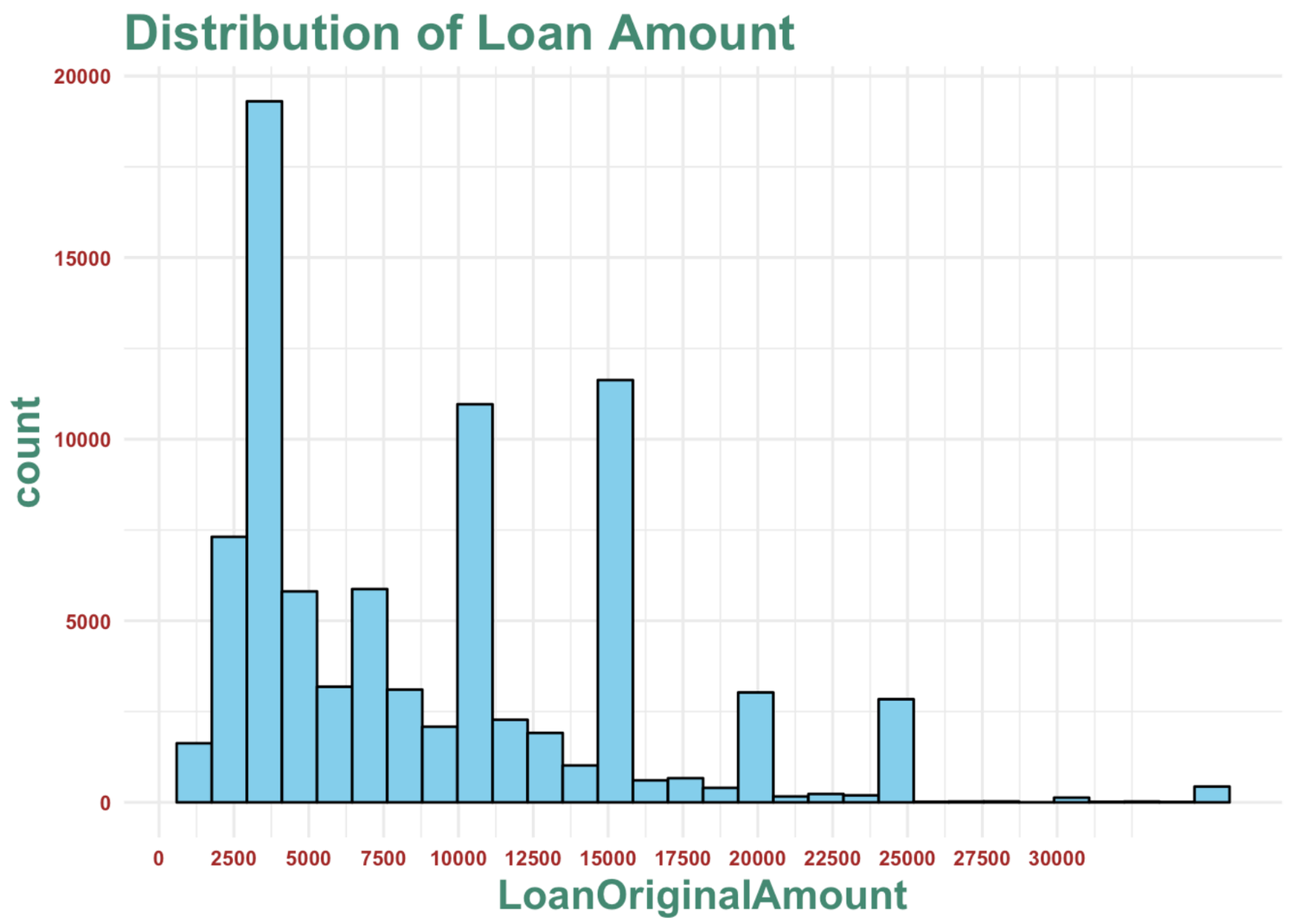
There are 84881 loans listed since July 2009.

1. Characteristics Of Loan (Amount, Payment, Term, Borrow Rate):

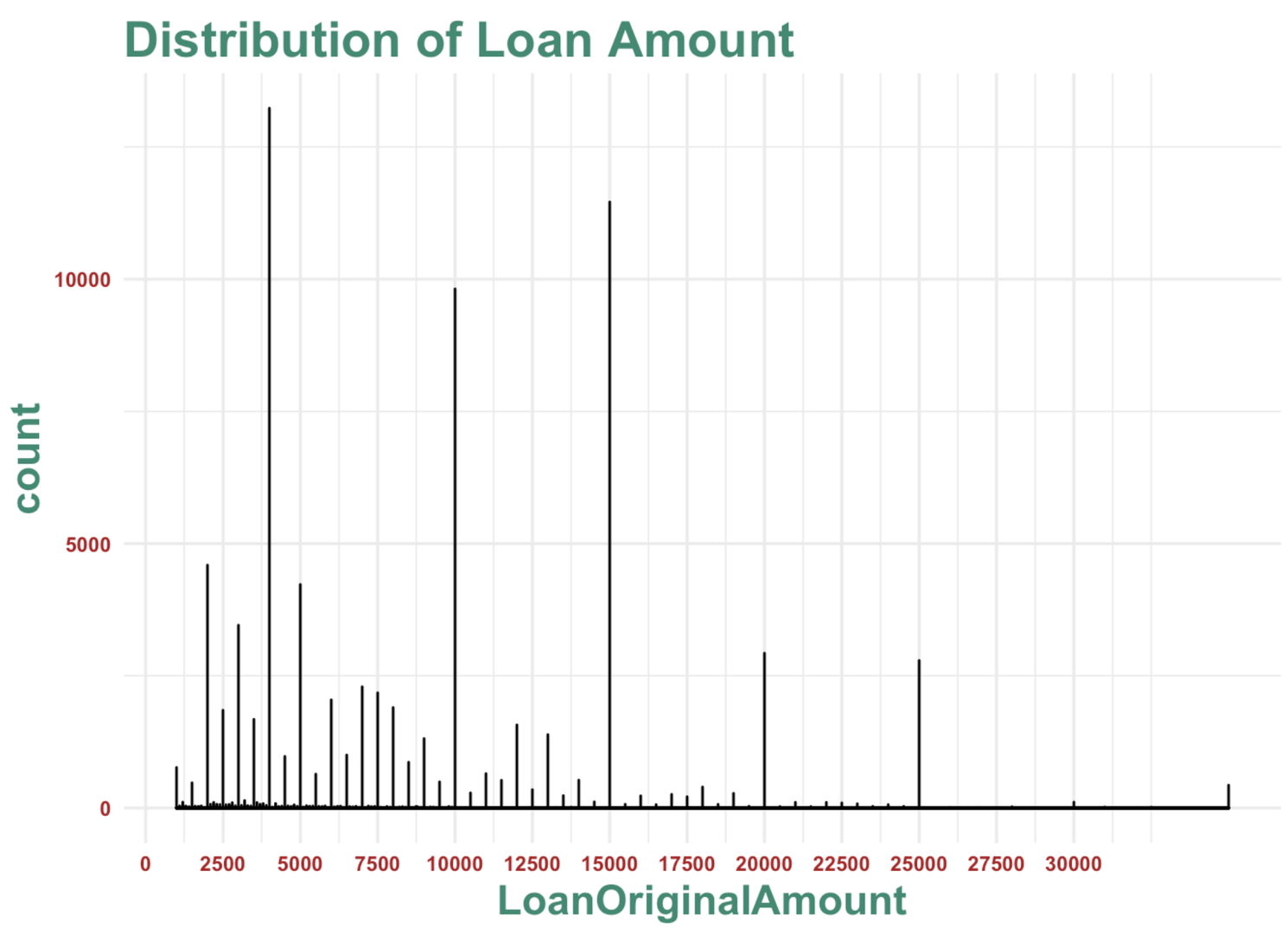
- **Loan Original Amount**:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1000    4000    7500    9082   13500   35000
```

The average loan amount is $9082 and the median loan amount is $7500. Since the median < mean, the distribution of loan amount is skewed to the right. The smallest loan is $1000, the largest loan is $35000.



Observing the distribtuion of loan amount, we notice that there are peaks at $2500, $7500, $10000, $15000, $20000, $25000 (or the peaks appeared at multiple of $2500 or multiple of $5000). But the highest peak is between $2500 and $5000. I guess it is because people try to borrow as much as they could but Prosper make it harder to borrow if the loan is above one specific amount which is in between $2500 and $5000. There's a way to figure out the specific amount by setting the binwidth to 1(each step is $1).

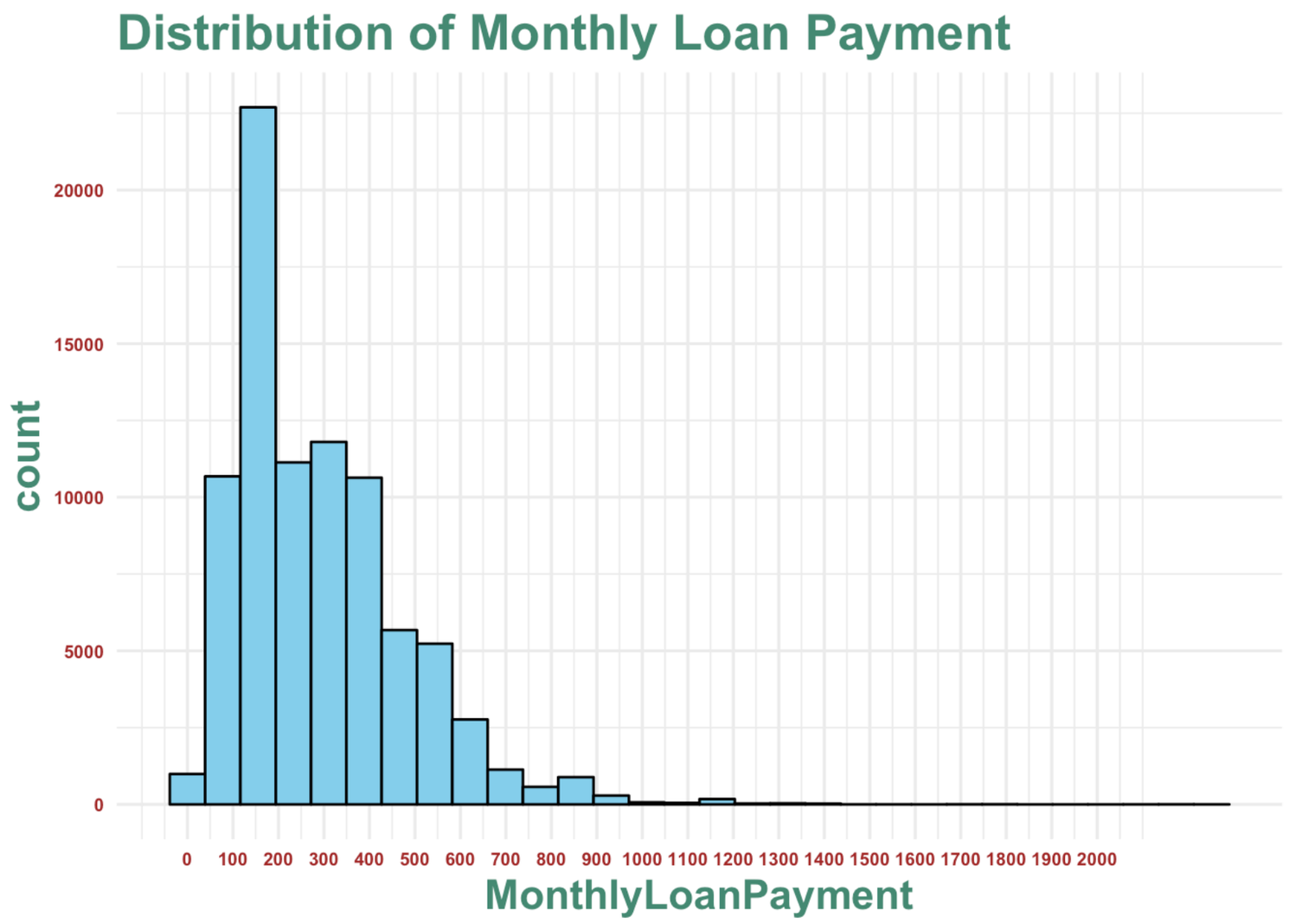


From the above histogram plot, we find that amount is probably $4000. More interestingly, we see that also applies to $5000, $10000, $15000, $20000, $25000.

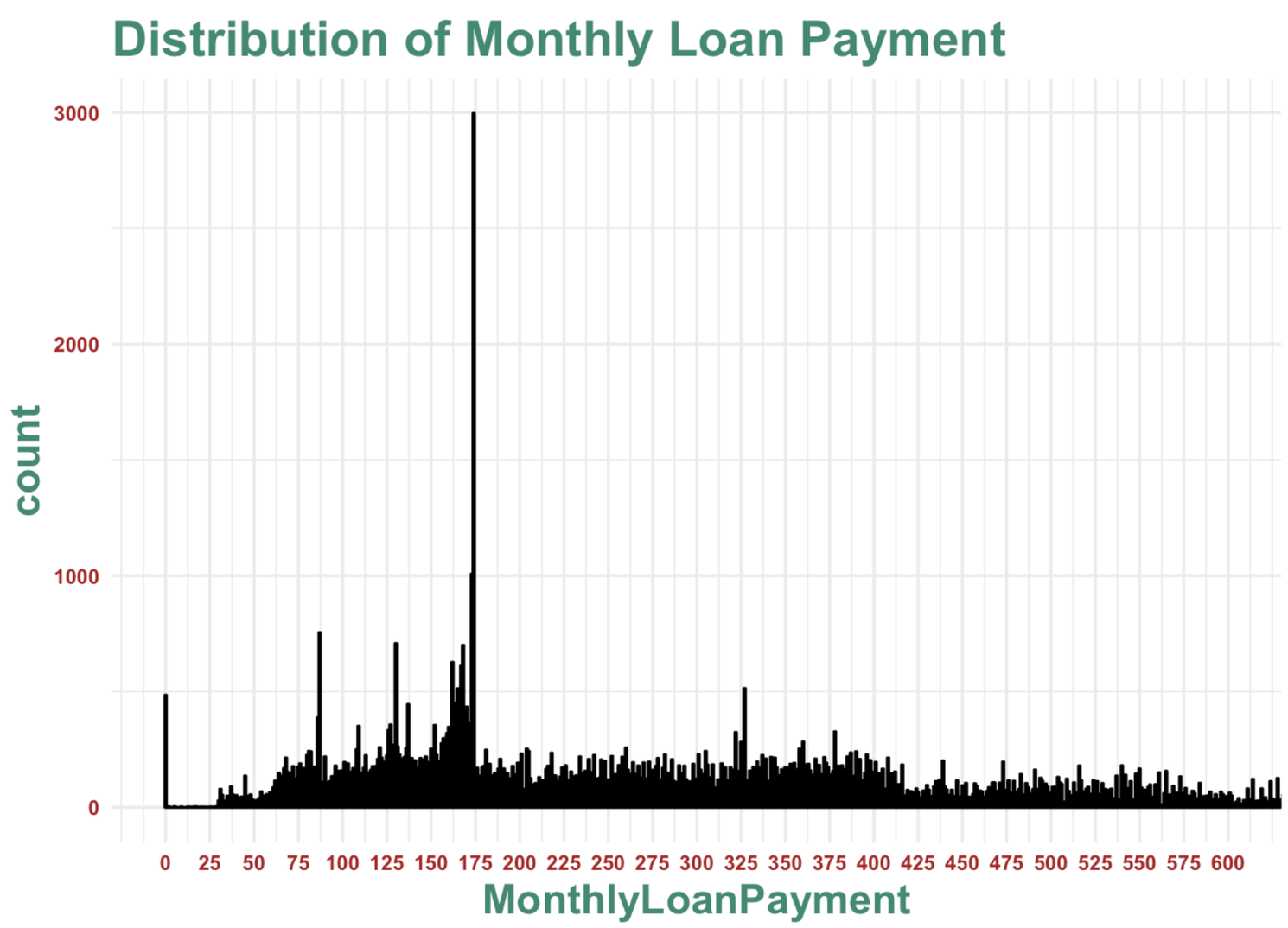
- **Monthly Loan Payment**:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   157.3   251.8   291.9   388.3  2252.0
```

Even though the range between minimum monthly payment and maximum monthly payment is quite large about $2252, half of the number of monthly payment is from $157.3 to $388.3. This makes sense because distribution of original loan amount shows that most of loan amount is from $1000 to $15000. Therefore most of monthly payment must be small amount from $100 to $400.



Clearly, most of the loan from $0 to $600. Let's check some abnomaly by setting the binwidth = 1 (each step is $1) and changing the observing frame from 0 to 600.

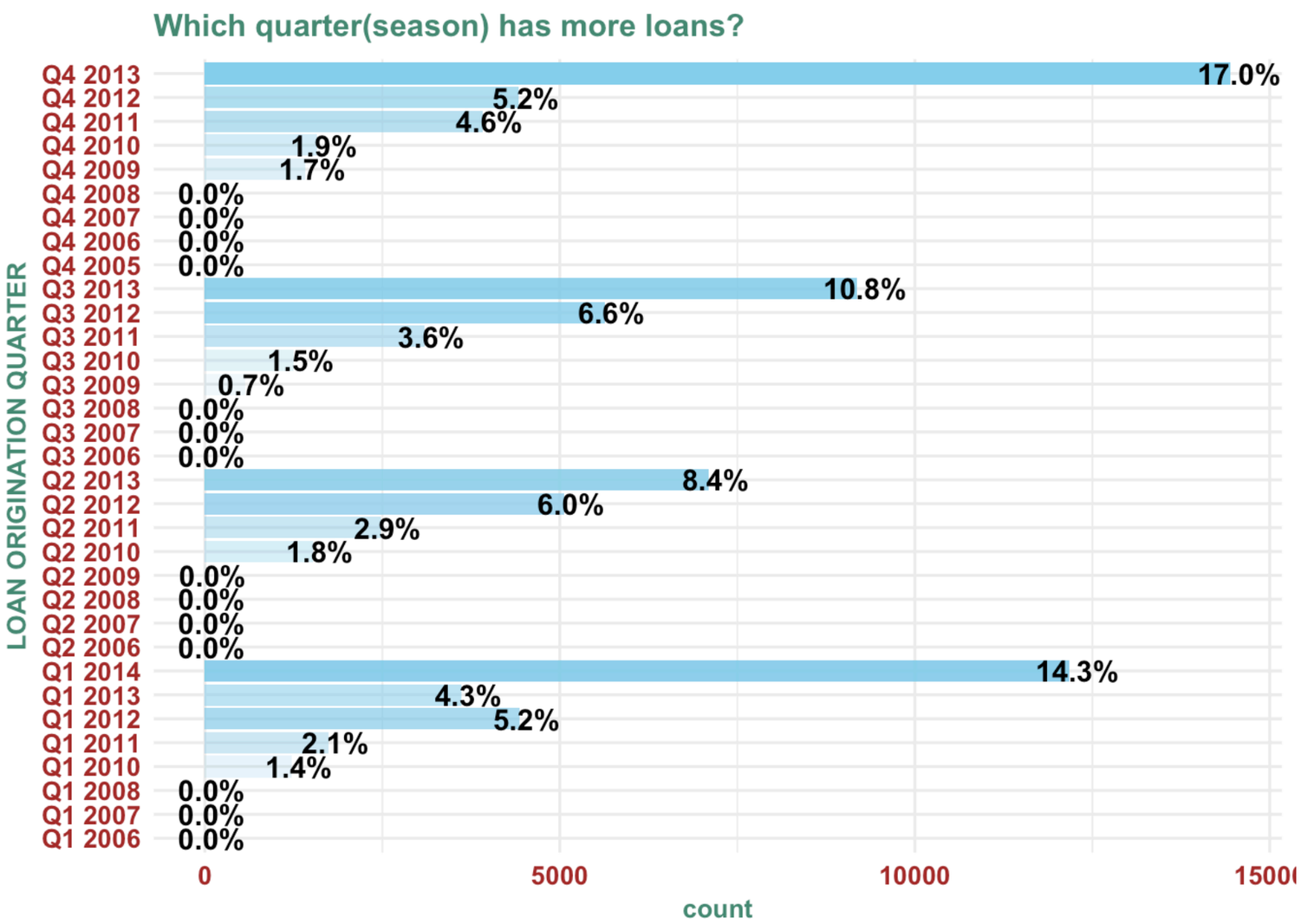


Last time the distribution of loan amount shows some abnormal values corresponding with multiple peaks. However, there is only one noticable peak in the distribution of monthly loan payment. The peak is valued at $175. I believe this $175 monthly payment corresponds to the loan of $4000 since they are two highest peak in those two distribution.

- **Loan Origination Quarter**:

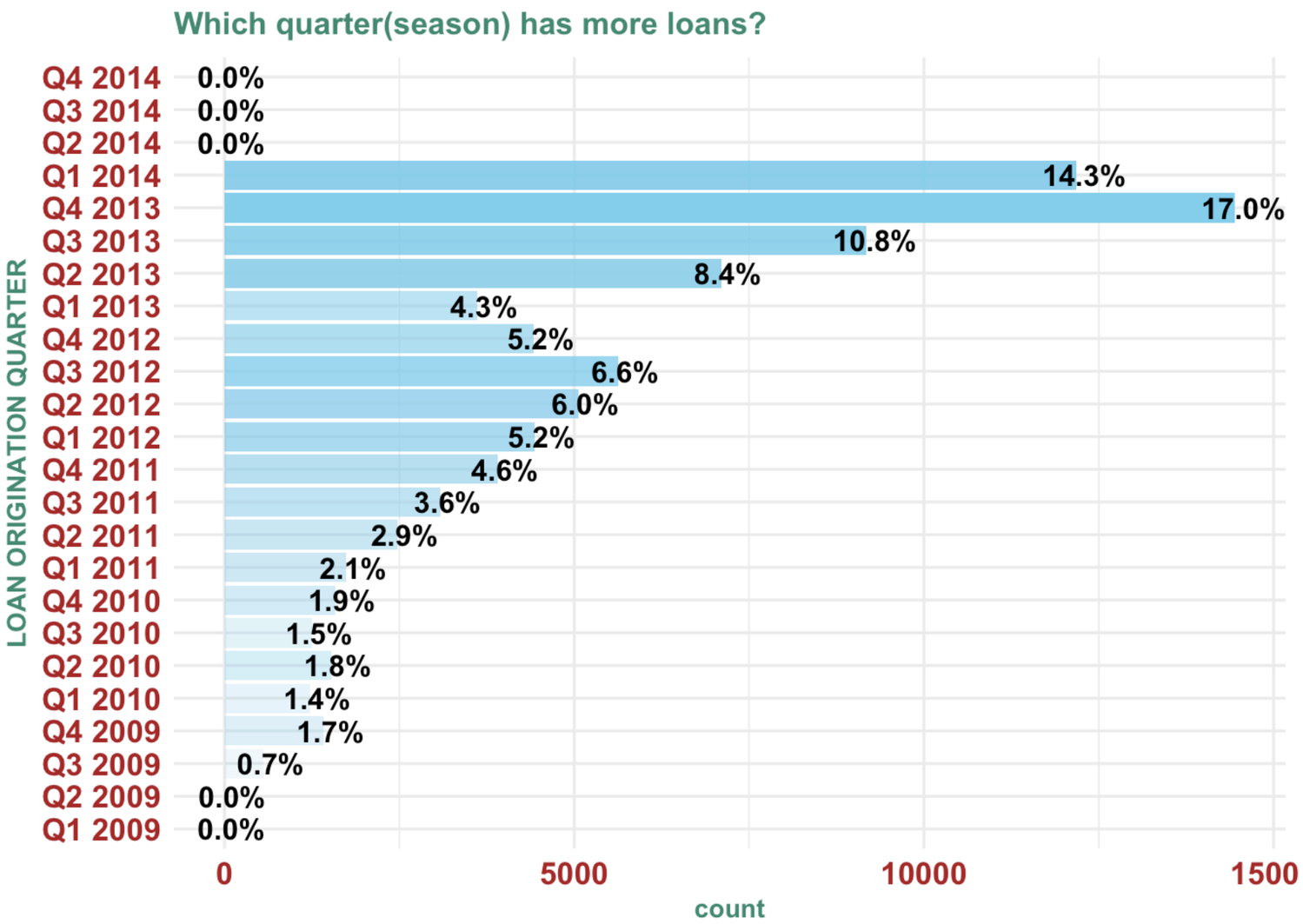
```
loan$LoanOriginationQuarter %>% unique
```

```
##  [1] Q1 2014 Q4 2012 Q3 2013 Q4 2013 Q2 2013 Q2 2012 Q1 2013 Q2 2010
##  [9] Q1 2012 Q3 2012 Q4 2010 Q4 2011 Q2 2011 Q1 2011 Q3 2009 Q3 2011
## [17] Q1 2010 Q4 2009 Q3 2010 Q2 2009
## 33 Levels: Q1 2006 Q1 2007 Q1 2008 Q1 2010 Q1 2011 Q1 2012 ... Q4 2013
```



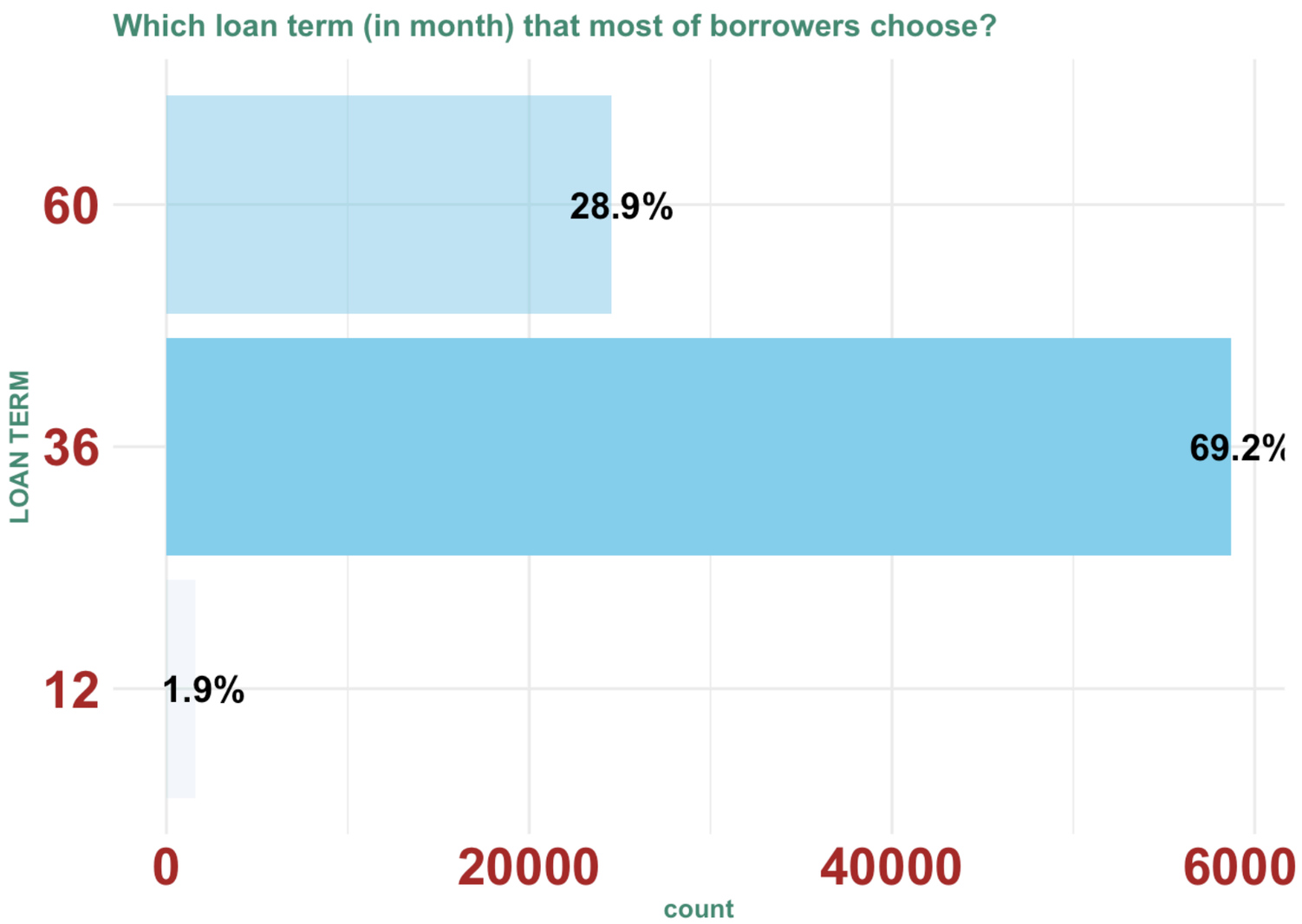
We can see that `LoanOriginationQuarter` has 33 levels but many of them are unused. This causes the bar chart appearing messy and unable to read. Values in `LoanOriginationQuarter` is formatted as 'Quater + Year'. But the levels of this variable is not in neat order. So I add or change the levels by:

```
quarter <- sapply(1:4, function(i) paste0('Q',i))
year <- 2009:2014
loan$LoanOriginationQuarter <- loan$LoanOriginationQuarter %>% ordered(levels =
        sapply(year, function(i) sapply(quarter,
                                function(j) paste(j,i))) %>% as.vector )
```



The number of loans increases every year (we could make the same conclusion with quarter even though there is a decrease in Q4 2012 and Q1 2013).

- **Loan Term**:



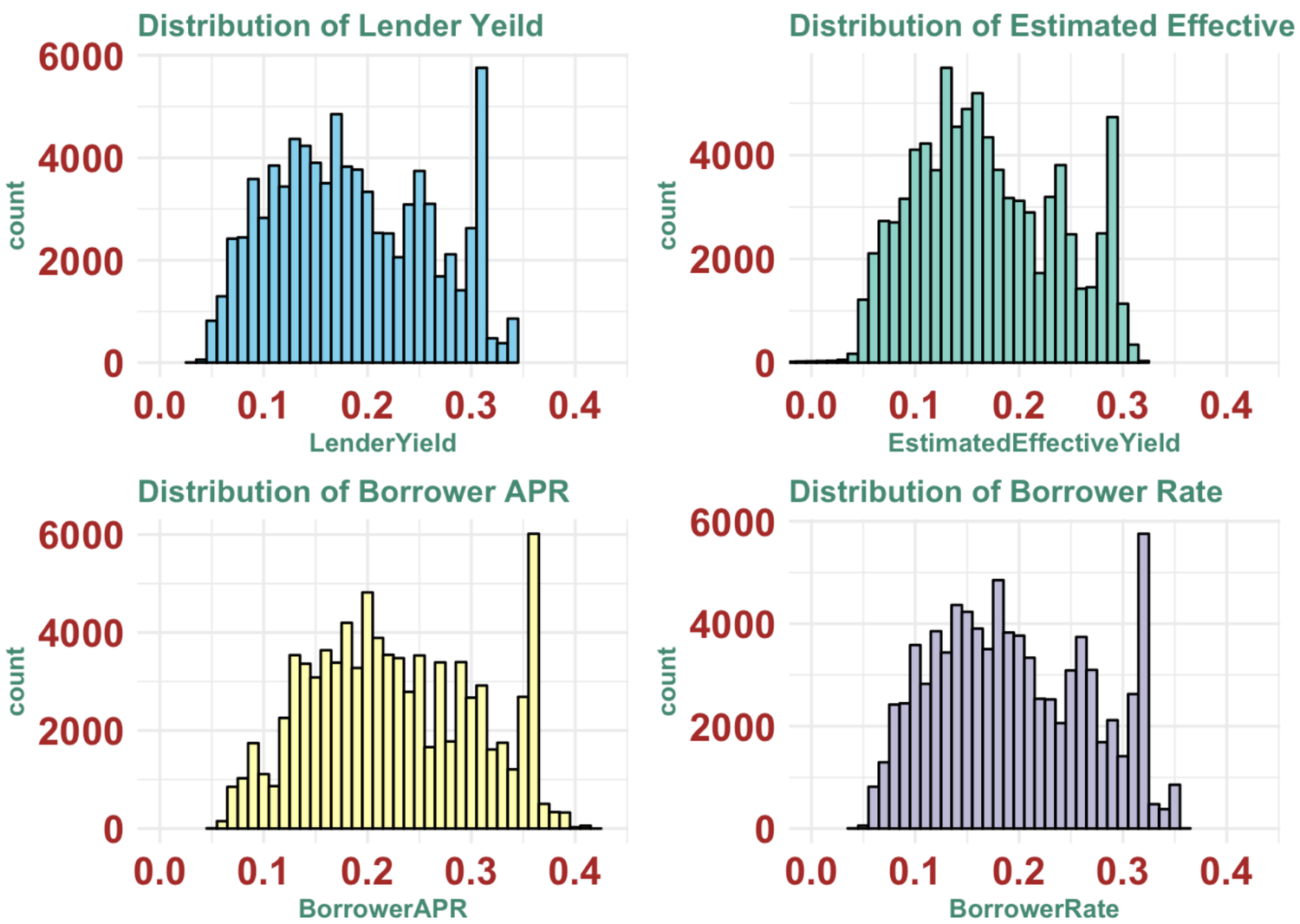
The length of the loan expressed in months. There are three loan terms: 12 months, 36 months and 60 months. Prosper provide mostly 36-months and 60-months loans.

- **Interest Rate**:

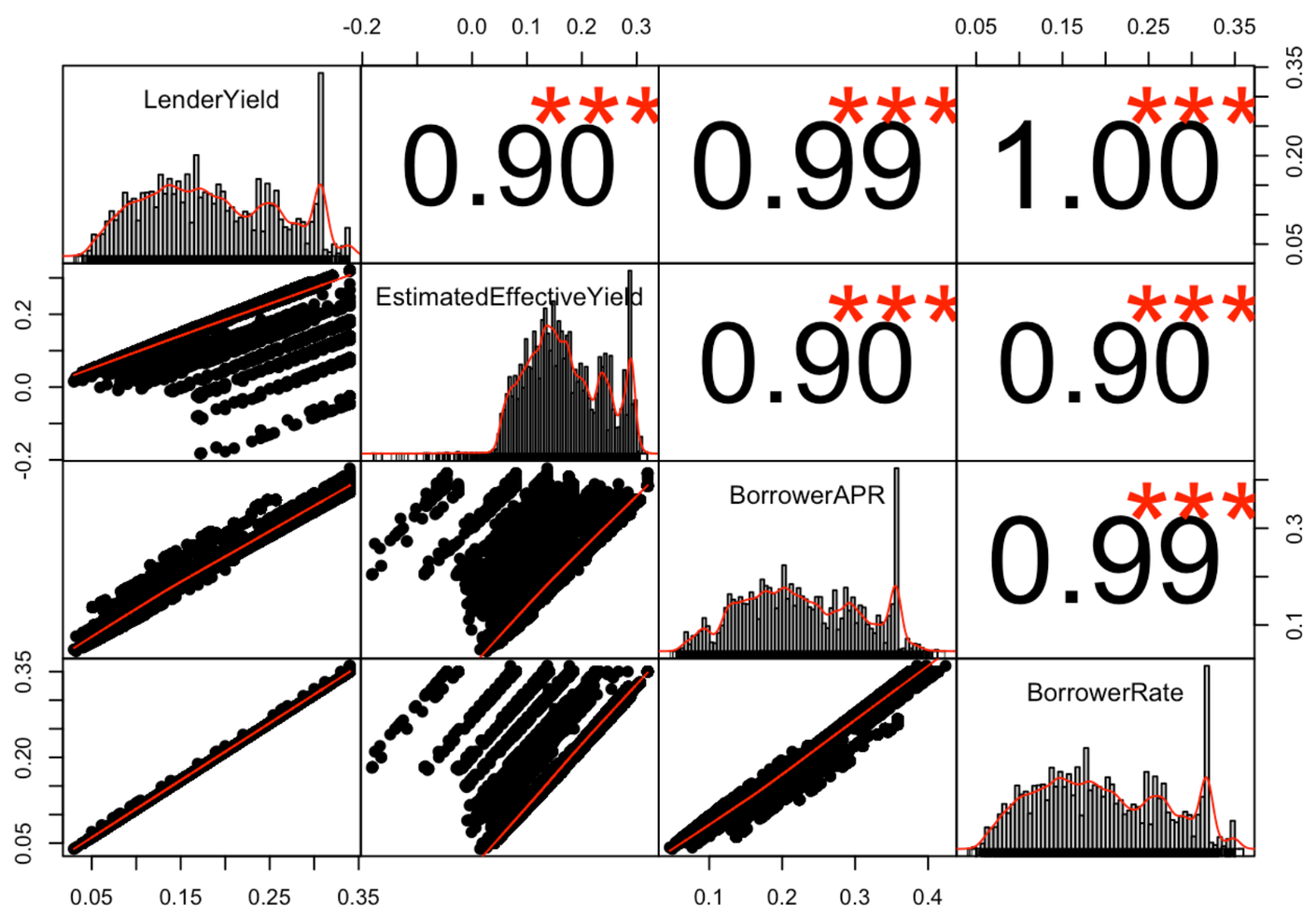
In this dataset, it is confusing to mention about interest rate because there are 4 interest rates: LenderYield, EstimatedEffectiveYield, BorrowerAPR, BorrowerRate.

```
loan[,c("LenderYield",
  "EstimatedEffectiveYield", "BorrowerAPR","BorrowerRate")] %>% summary
```

```
##   LenderYield     EstimatedEffectiveYield  BorrowerAPR
##  Min.   :0.0300   Min.   :-0.1827         Min.   :0.04583
##  1st Qu.:0.1259   1st Qu.: 0.1157         1st Qu.:0.16328
##  Median :0.1775   Median : 0.1615         Median :0.21945
##  Mean   :0.1860   Mean   : 0.1687         Mean   :0.22665
##  3rd Qu.:0.2474   3rd Qu.: 0.2243         3rd Qu.:0.29254
##  Max.   :0.3400   Max.   : 0.3199         Max.   :0.42395
##                   NA's   :28
##   BorrowerRate
##  Min.   :0.0400
##  1st Qu.:0.1359
##  Median :0.1875
##  Mean   :0.1960
##  3rd Qu.:0.2574
##  Max.   :0.3600
##
```



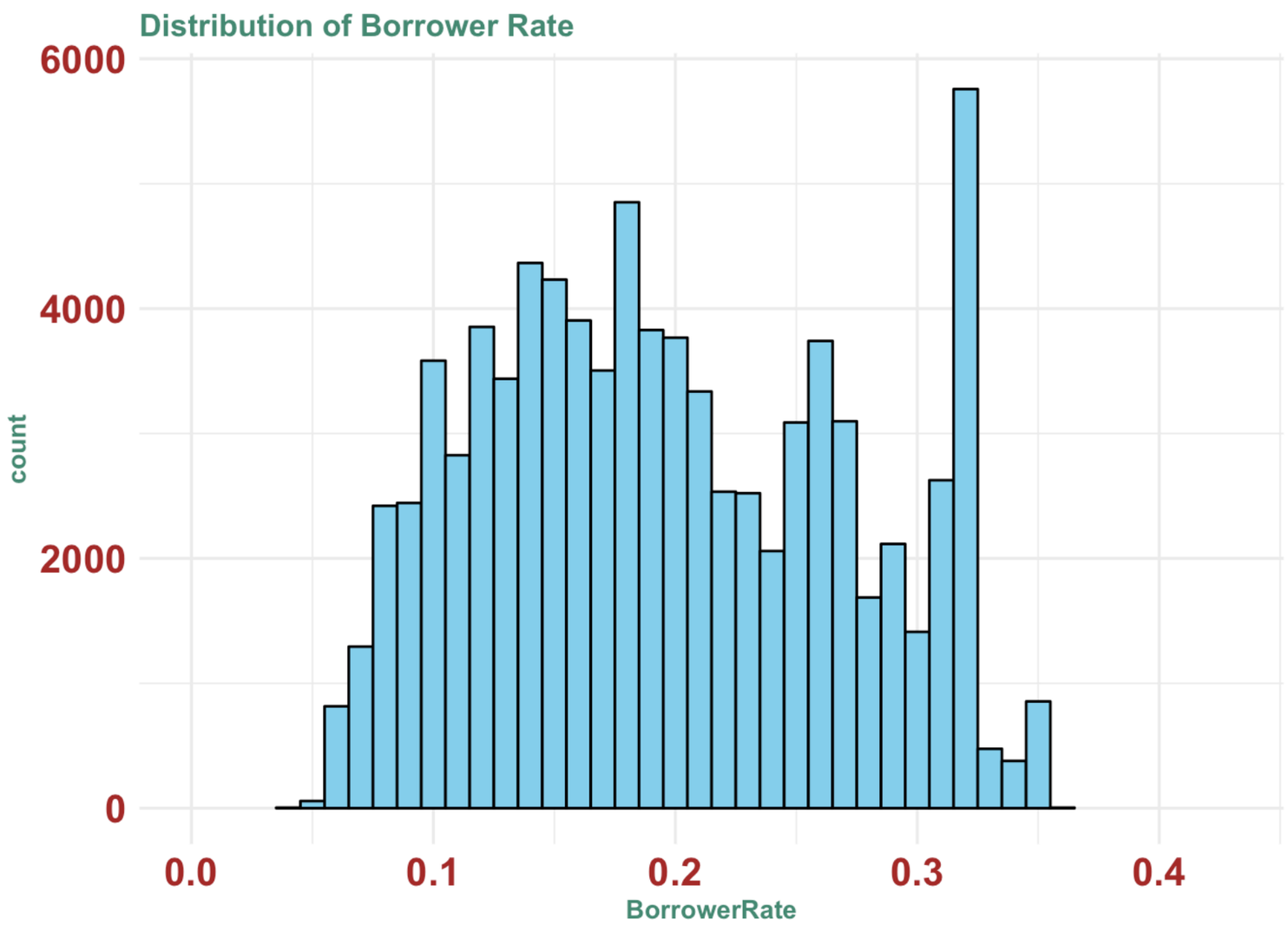
Statistics and distribution of these four interest rate are very similar. They could be slightly changed to solve for different problem. But the change is not so large as we could worry if we choose one of them as the representative for the whole group as an interest rate. But let's check whether they are really related to each other.



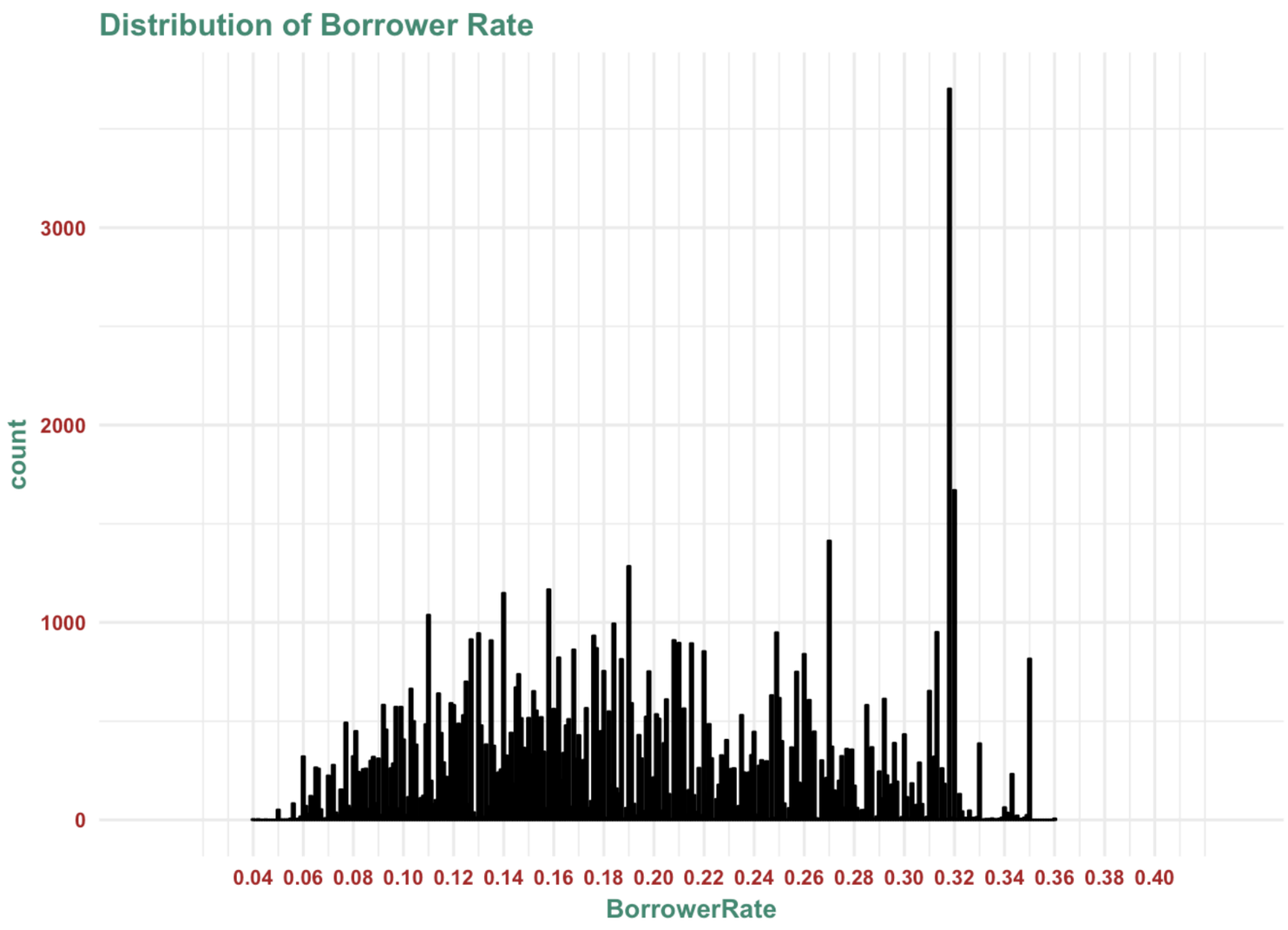

```
loan[,c("LenderYield", "EstimatedEffectiveYield", "BorrowerAPR","BorrowerRate")] %>%
subset(!is.na(LenderYield) & !is.na(EstimatedEffectiveYield) &
       !is.na(BorrowerAPR) & !is.na(BorrowerRate)) %>% cor
```

```
##                          LenderYield EstimatedEffectiveYield BorrowerAPR
## LenderYield                1.0000000               0.8953425   0.9933345
## EstimatedEffectiveYield    0.8953425               1.0000000   0.8956348
## BorrowerAPR                0.9933345               0.8956348   1.0000000
## BorrowerRate               0.9999958               0.8952825   0.9933333
##                          BorrowerRate
## LenderYield                 0.9999958
## EstimatedEffectiveYield     0.8952825
## BorrowerAPR                 0.9933333
## BorrowerRate                1.0000000
```

Clearly, Lender Yield, EstimatedEffectiveYield, BorrowerAPR, BorrowerRate are strongly positively correlated. Therefore I would choose BorrowerRate to represent as interest rate. We might check the distribution of BorrowerRate again.



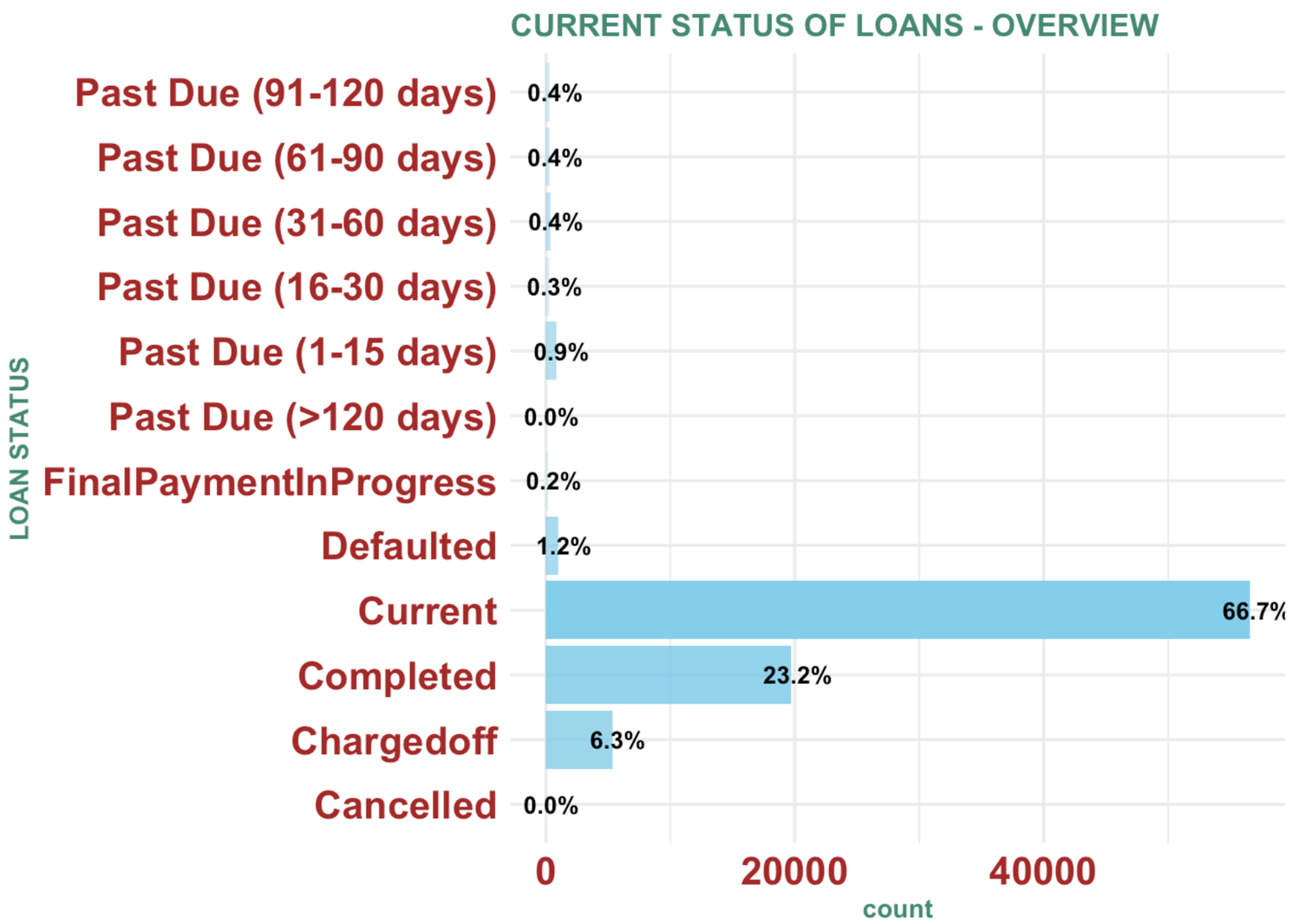
We can set the binwidth to 0.001 instead of 0.01 to see some abnomaly



In this case, I notice that the most common borrow rate is 0.32. I don't think it corresponds to $4000 loan with $175 monthly payment because 0.32 is a very high borrowing interest rate; it is nearly close to the higher end of the distribution.

- **Loan Status**:

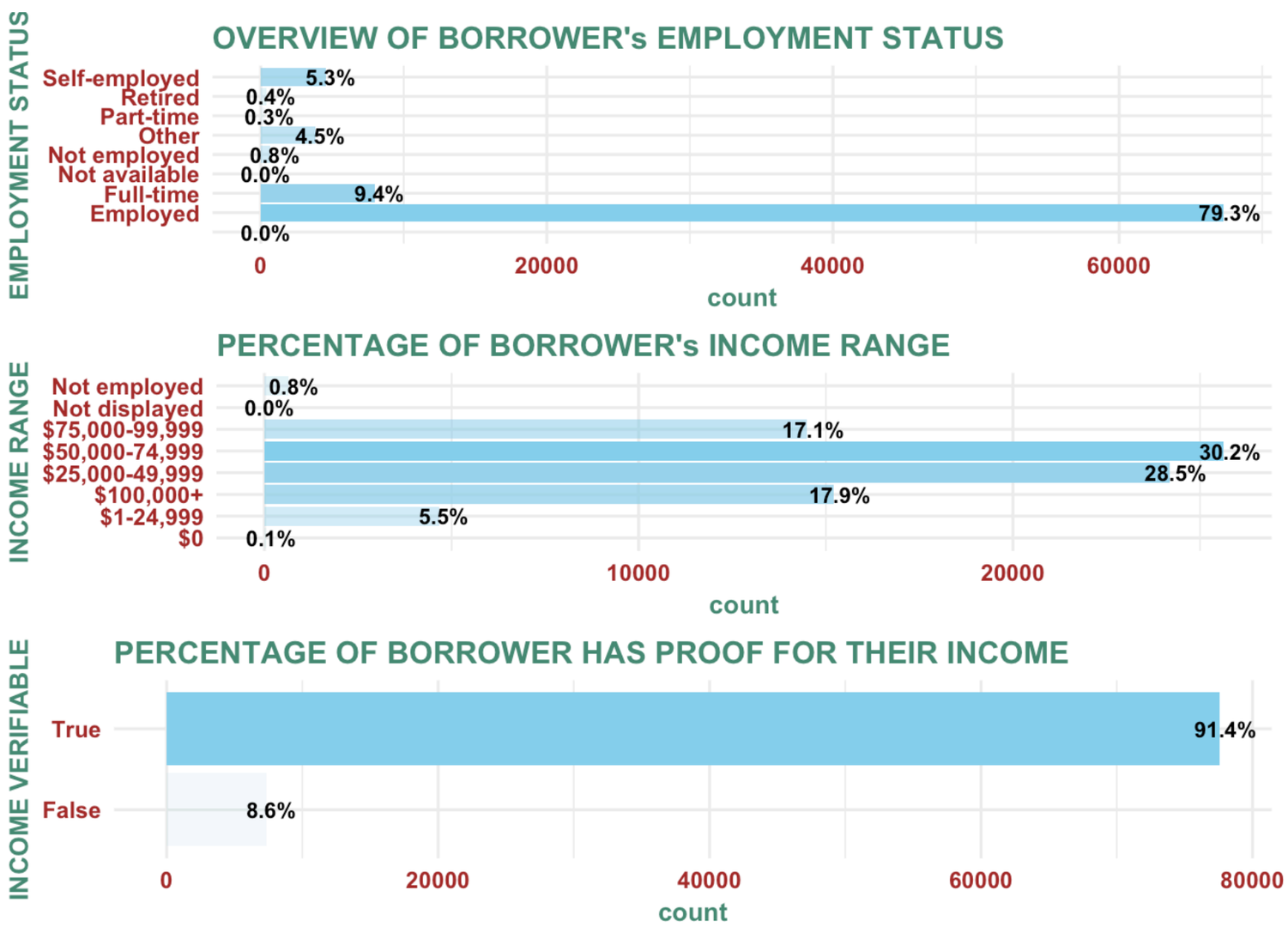
```
The current status of the loan: Cancelled,  Chargedoff, Completed, Current, Defaulted, FinalPaymentInProgress, PastDue. The PastDue
```



Most people will not know what the difference between defaulted loan and charged off loan is. After 130+ days past due, loans enters Default status. After 150 days past due, Default loans enters Charged-Off status. 66.7% of the loans are still active. 23.2% loans are completed. 6.3% loans are charged off. Small fraction of loans have late payments.

2. Credit Risk (or Loan Assessment):

- **Income Factor**:

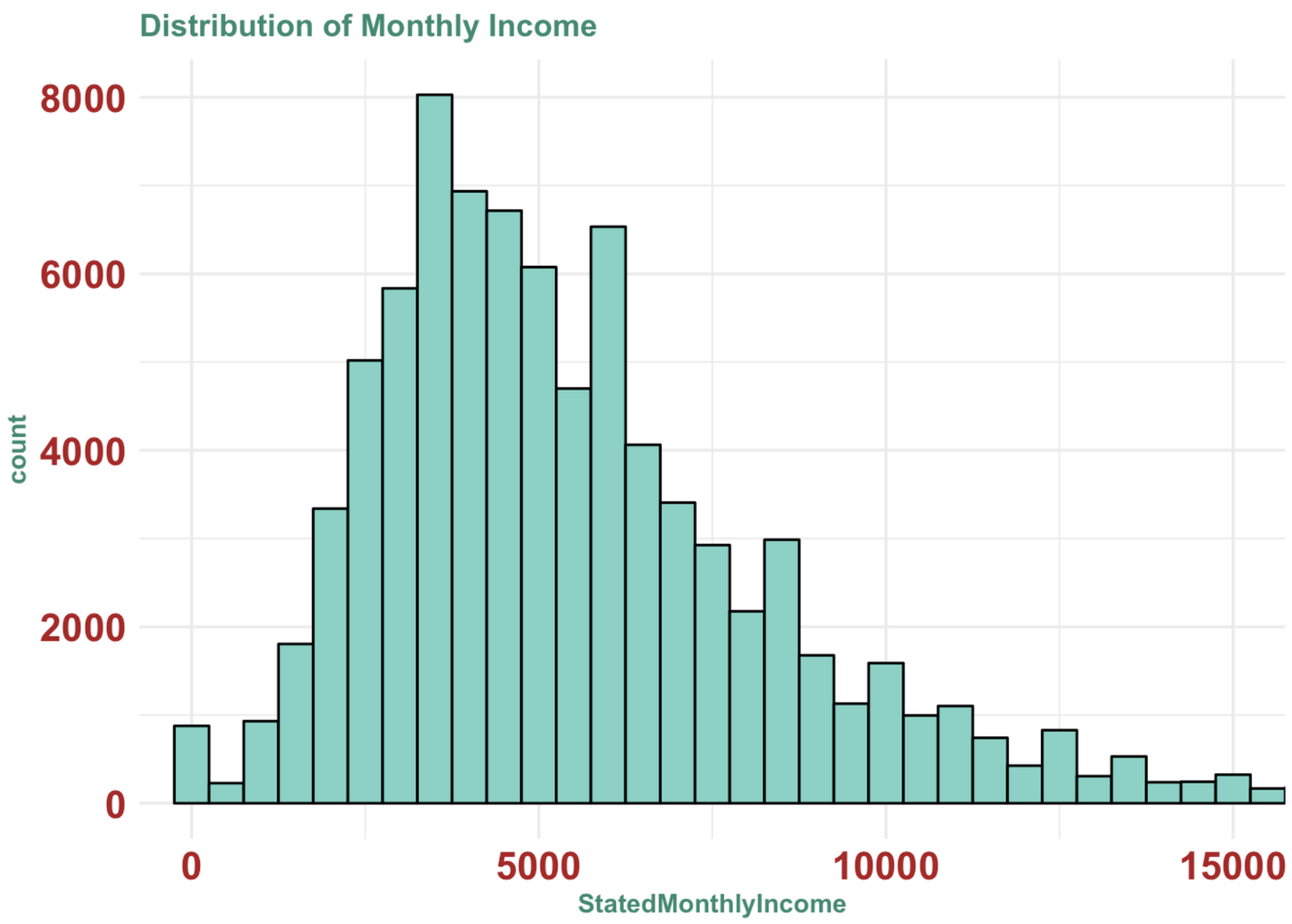


Employed and full-time tend to borrow more than the rest. Among those employed people who have income ranging from $25000 to $75000 take out more loans. And most of them have proof for their income

```
loan$StatedMonthlyIncome %>% summary
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    3433    5000    5931    7083 1750000
```

The average (stated) monthly income of borrowers is $5931. What surprises us is the maximum monthly income is $1,750,000 because we have learnt that maximum loan amount is $35000. It does not make sense that high earner borrows tiny loan amount. Let’s see the distribution of (stated) monthly income:



The distribution is slightly skewed to right (as expected median is not much smaller than mean) with long tail.

```
library(DT)
loan %>%
subset(StatedMonthlyIncome > 15000 & IncomeVerifiable == 'False' &
         ProsperRating..numeric. < 4) %>%
select(StatedMonthlyIncome, IncomeRange, ProsperRating..Alpha.) %>% datatable
```

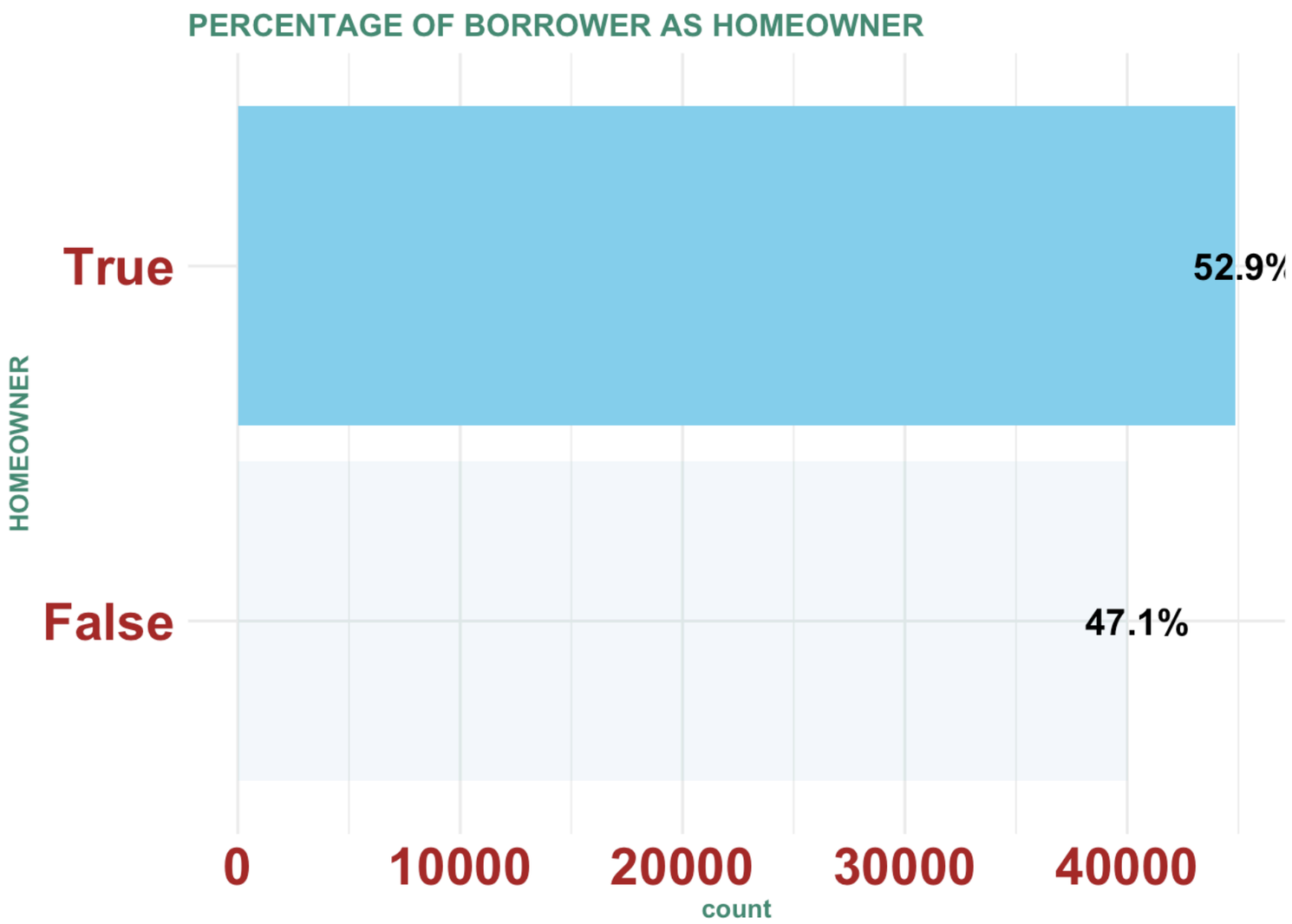
Show 10 entries Search:

	StatedMonthlyIncome	IncomeRange	ProsperRating..Alpha.
1012	16666.666667	$100,000+	D
1610	25000	$100,000+	D
1653	19000	$100,000+	E
1867	41666.666667	$100,000+	D
2902	16666.666667	$100,000+	D
3723	19416.666667	$100,000+	E
5487	20000	$100,000+	HR
7404	15416.666667	$100,000+	HR
8125	20833.333333	$100,000+	HR
8170	16666.666667	$100,000+	D

Showing 1 to 10 of 121 entries Previous 1 2 3 4 5 … 13 Next

Even though some people have stated that they are high earners (person makes more than $15000 a month), it shows that many of them cannot verify their income and they are classified as high risk with D, E or HR rating.

- **Homeowner Factor**:



Half of the loans are made by homeowners.

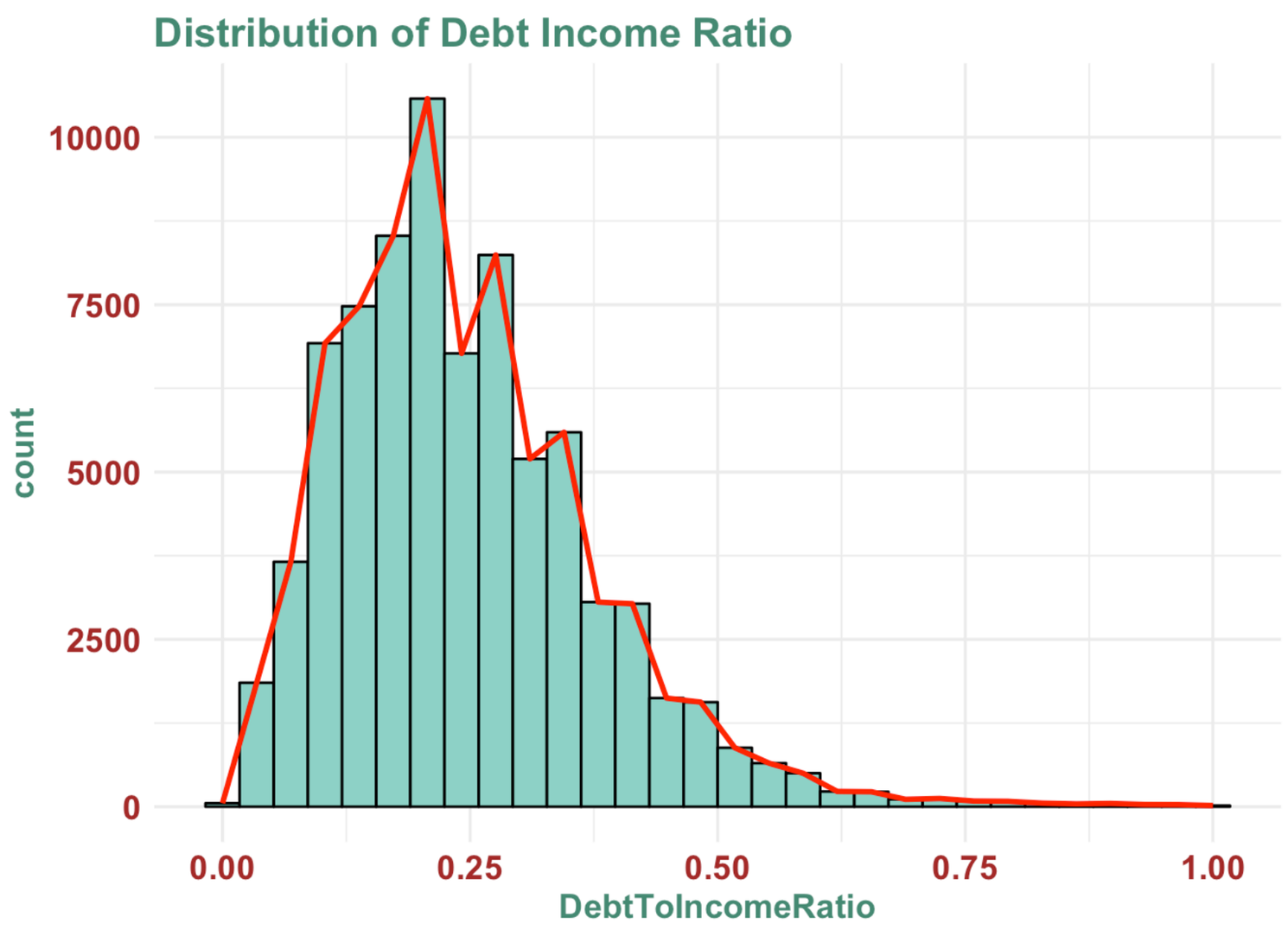
- **Debt To Income Ratio**:

```
Debt To Income Ratio
```

```
loan$DebtToIncomeRatio %>% summary
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   0.150   0.220   0.259   0.320  10.010    7297
```

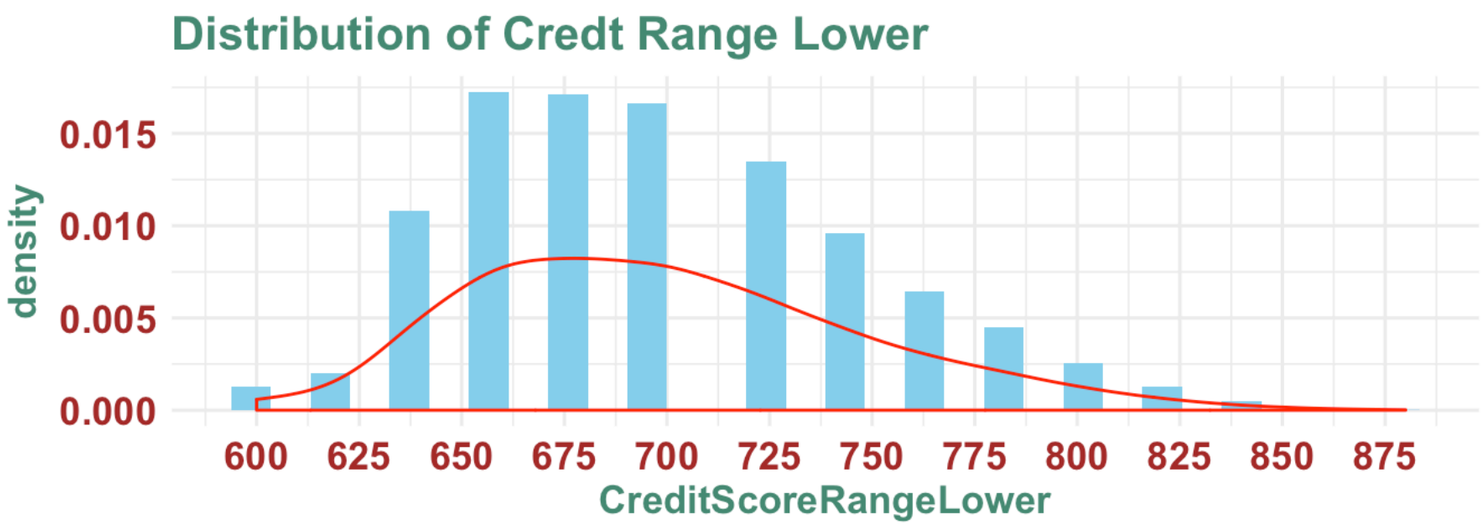
The median debt to income ratio is 0.22, which means that the loan weigh 22% of their annual income.

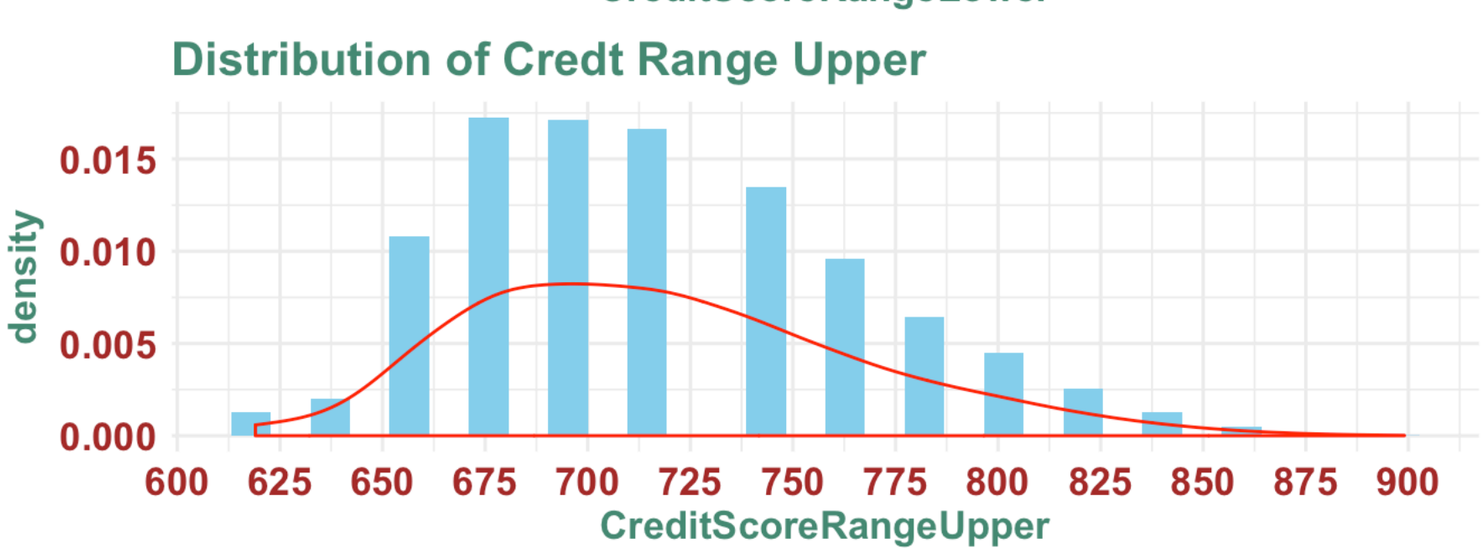


Distribution of Debt Income Ratio is centered around 0.25 and it is skewed to the right.

- **Credit Score**:

Range of credit score is from lower to upper.





The distribution of credit score is slightly skewed to the right, but is centered around 700.

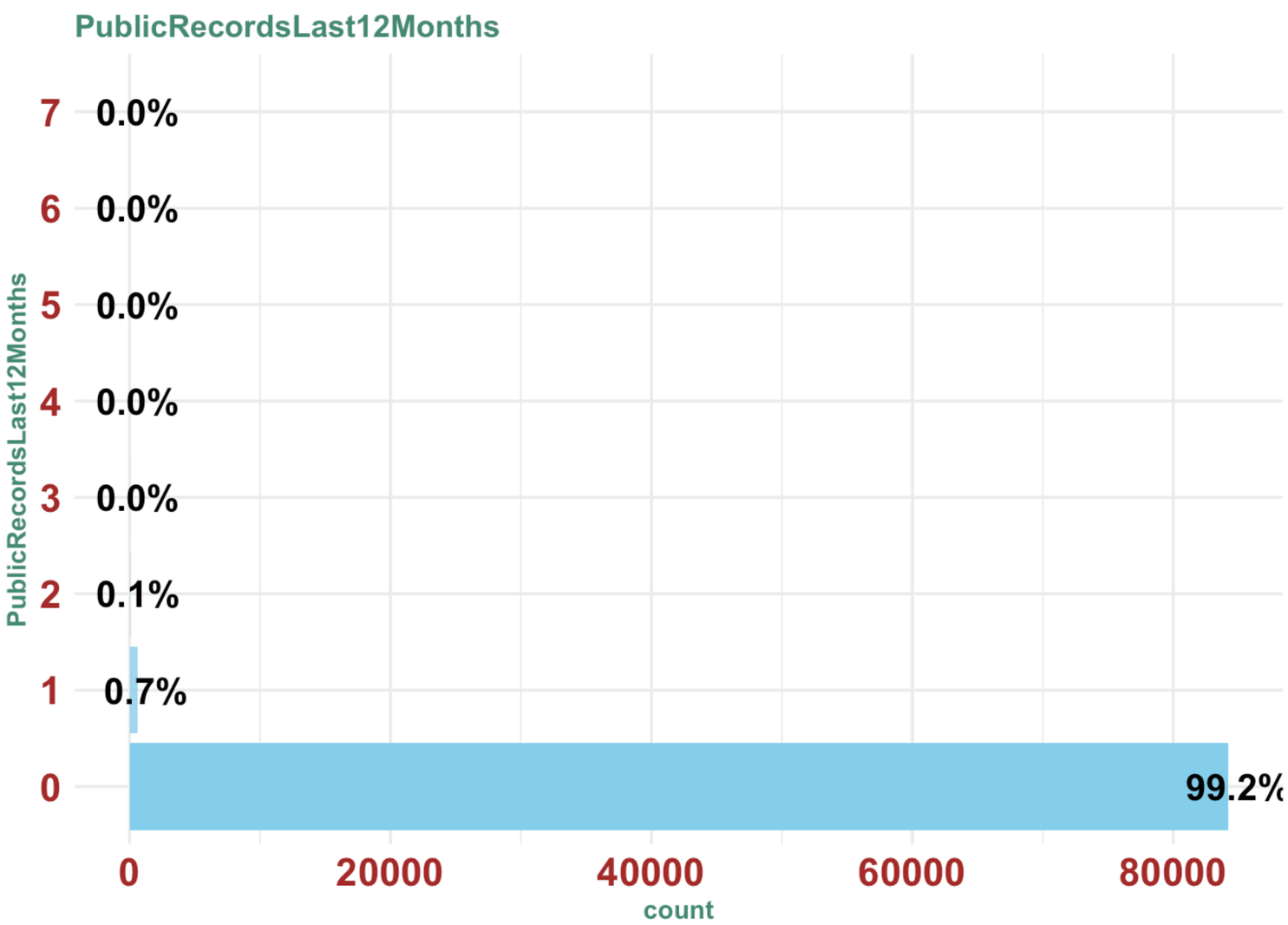
- **Public Records Last 12 Months**:

There are three types of Public Records appear in a credit report including: * status of bankcruptcy (Are you currently bankcrupt?) * tax lien (Do you owe any tax?) * civil judgement (Do you have any unpaid fee or ongoing paying process to the court as a result of your past lawsuit?).

I think the higher number of records you receive, the higher interest rate and the lower amount of loan you can get.

```
loan$PublicRecordsLast12Months %>% table
```

```
## .
##     0     1     2     3     4    20
## 84199   621    47     9     4     1
```



It is very interesting to know that most of the borrowers have no public records in the last 12 months. (In other words, we cannot use PublicRecordLast12Months as a variable to assess the risk of lending money)

- **Inquires Last 6 Months**:

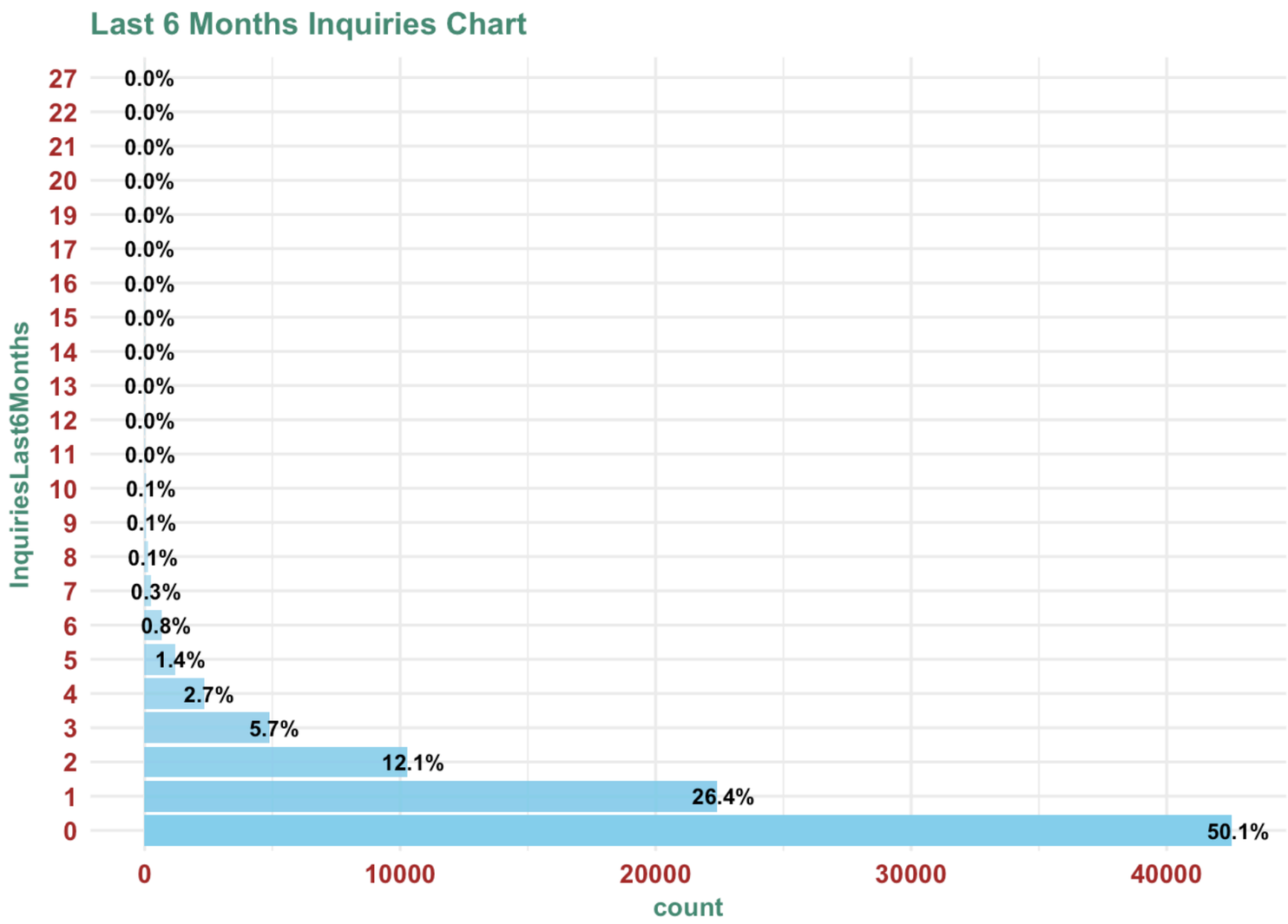
Credit inquiries are the requests to check your credit score when you apply for new credit card or new loan.

The number of inquiries also affect to amount of loan and interest rate,

```
loan$InquiriesLast6Months %>% table
```

```
## .
##     0     1     2     3     4     5     6     7     8     9    10    11
## 42558 22410 10270  4875  2328  1203   648   252   127    73    51    26
##    12    13    14    15    16    17    19    20    21    22    27
##    21     9     9     8     6     2     1     1     1     1     1
```

Very few people have more than 7 inquires in the last 6 months.

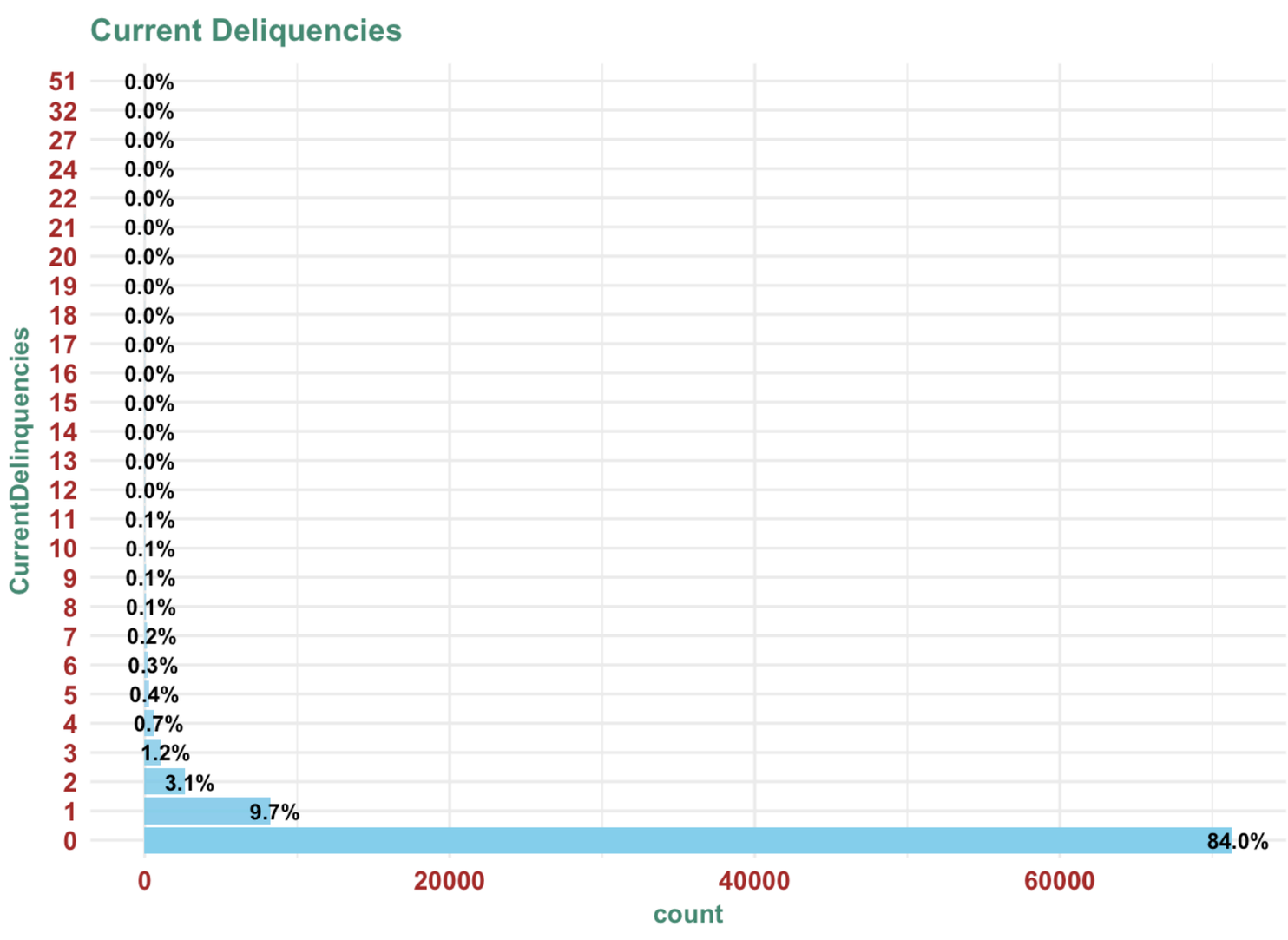


In the last 6 months, half of borrowers have zero inquiries; some have 1 or 2 inquires; very few borrowers have more than 3 inquires.

- **Current Deliquencies**:

```
loan$CurrentDelinquencies %>% summary
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.3223  0.0000 51.0000
```

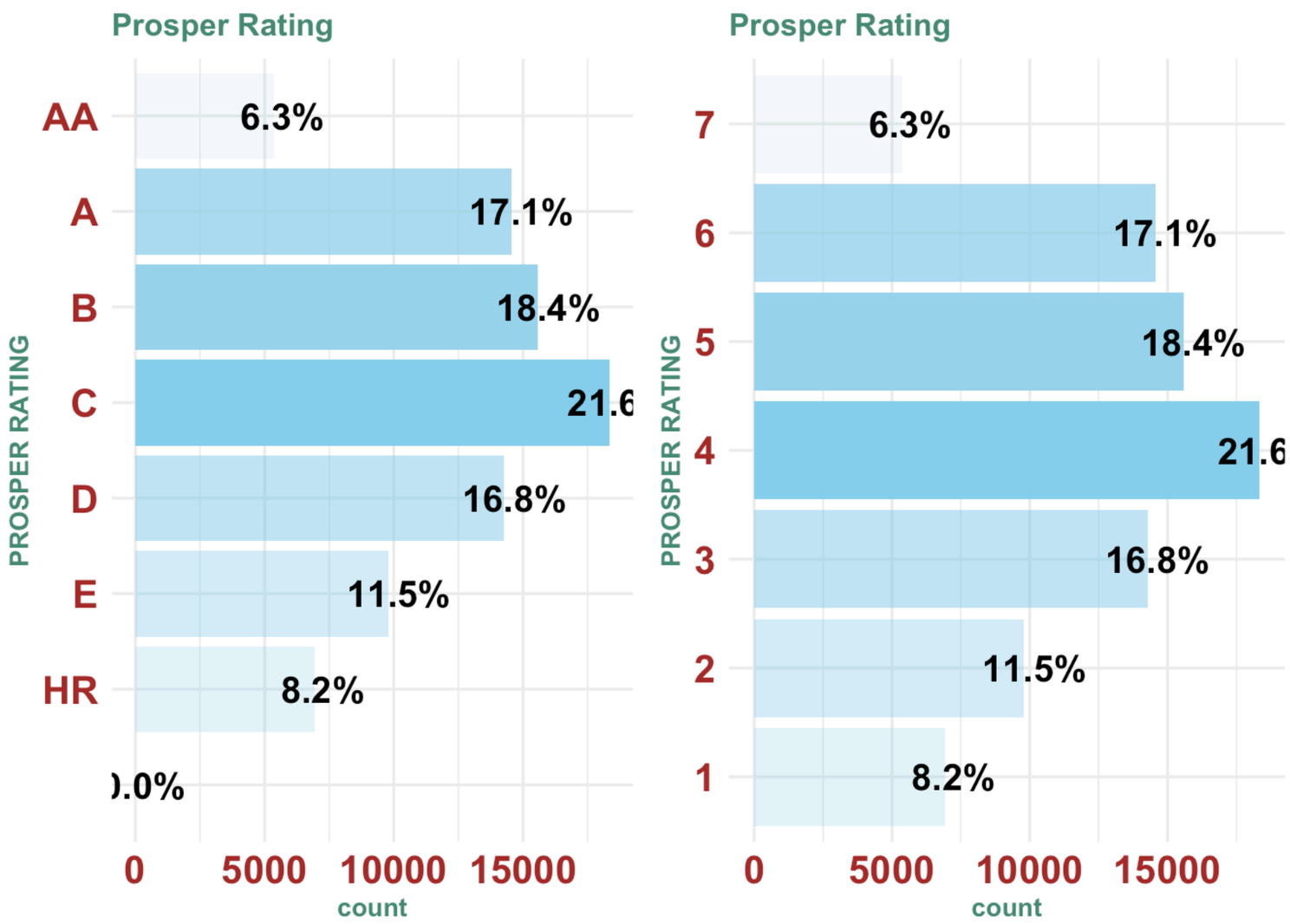


- **Prosper Rating**:

The Prosper Rating assigned at the time the listing was created: 0 - N/A, 1 - HR, 2 - E, 3 - D, 4 - C, 5 - B, 6 - A, 7 - AA. Appli

```
loan$ProsperRating..Alpha. %>% table
```

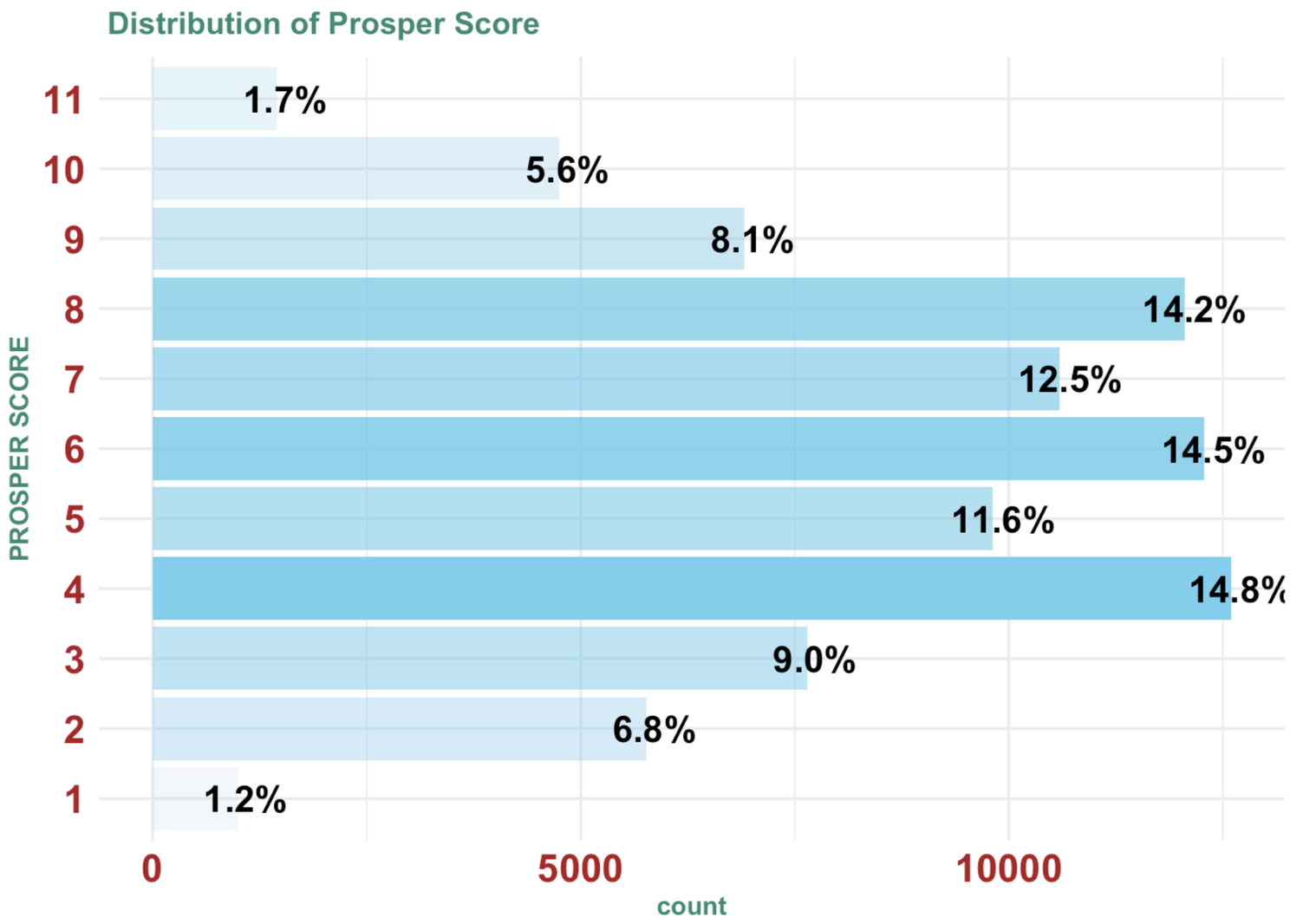
```
## .
##            A     AA      B      C      D      E     HR
##    28 14551   5372 15581 18345 14274   9795   6935
```



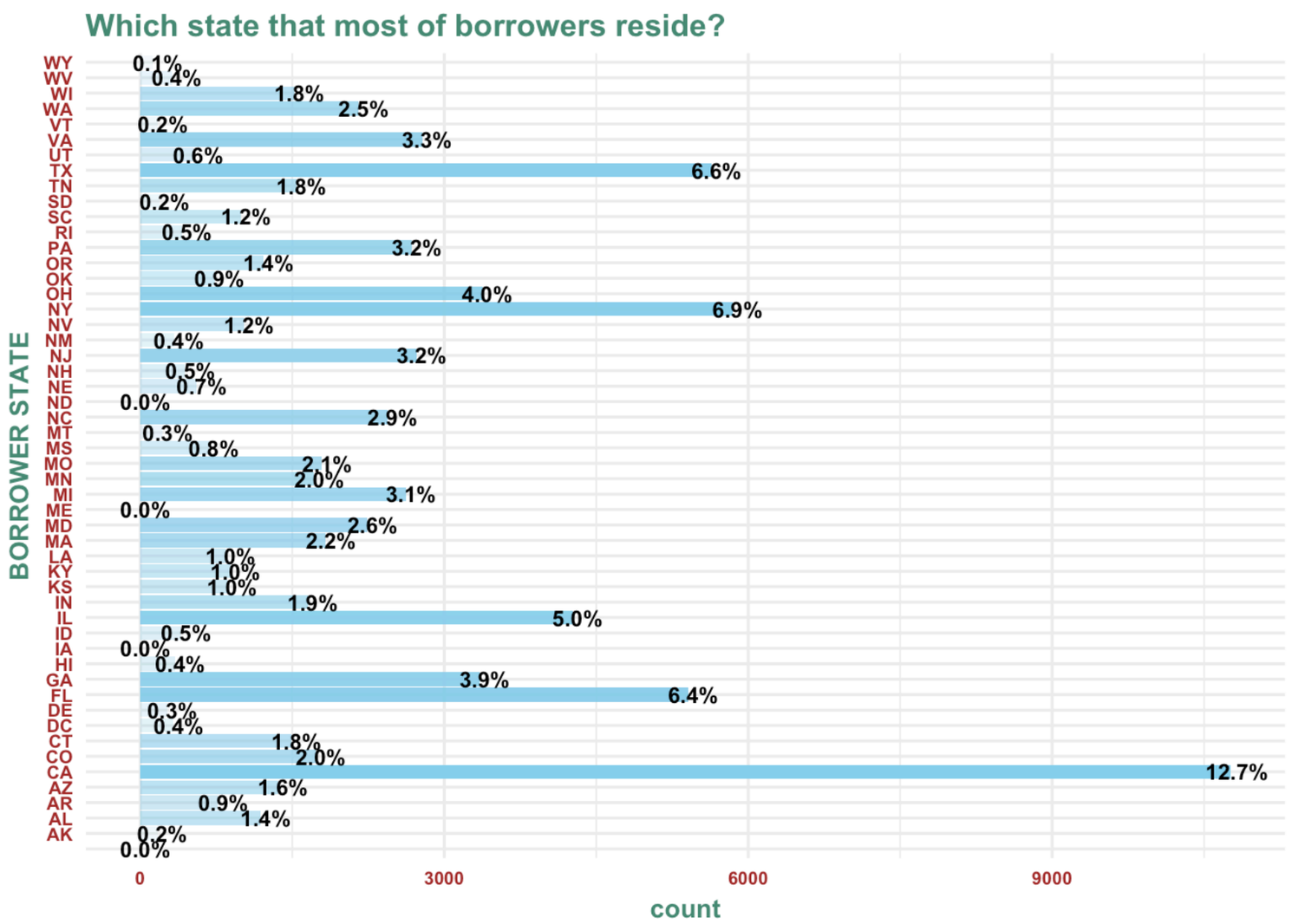
We can see that Prosper Rating Numeric and Prosper Rating Alpha are similar. And investors would feel the distribution of Prosper Rating is better at risk assessment than the distribution of Credit Grade. We also see that three most common loans are A, B and C.

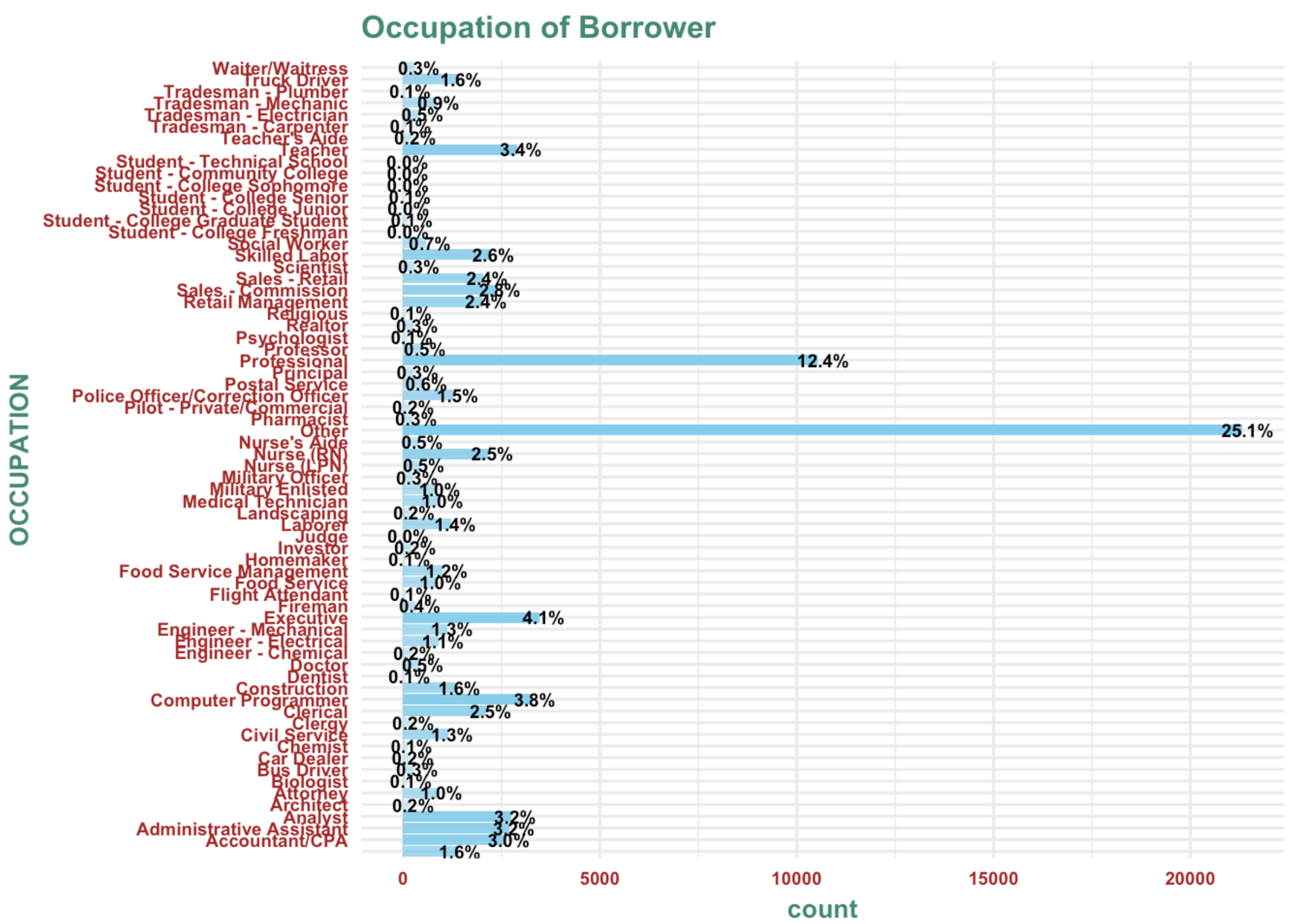
- **Prosper Score**:

```
title_lab <- " Distribution of Prosper Score"
x_lab <- "PROSPER SCORE"
lab_text_size <- 15
loan$ProsperScore %>% draw_bar_plot(title_lab, x_lab, lab_text_size)
```



3. Other Information related to Loan:





Notice that:

- Most loans are made from people in big states.
- Prosper need to improve about the quality of the Occupation data because two highgest percentage of loan makers are listed as Other and Professional. And we cannot know what they exactly do for a living.
- It seems that most of the borrowers have problem with their debts since the purpose of 51.2% of borrowers is debt consolidation.

Univariate Analysis

The dataset contains 113937 observations. Each observation is described by 81 variables. Prosper has changed to use the new rating system since July 2009. Therefore I have create a new variable ListingYear and subset data from 2009 onwards. The new dataset contains 84881 observations.

I choose some features out of 81, and classify them into 3 parts:

- **Characteristics Of Loan**: LoanOriginalAmount, MonthlyLoanPayment, LoanOriginationQuarter, Term, BorrowerRate, LoanStatus
- **Credit Risk (or Loan Assessment)**: EmploymentStatus, IncomeRange, IncomeVerifiable, IsBorrowerHomeowner, DebtToIncomeRatio, CreditScoreRangeLower, CreditScoreRangeUpper, PublicRecordsLast12Months, InquiresLast6Months, CurrentDeliquencies, ProspeRating, ProsperScore.
- **Other Information related to Loan**: BorrowerState, Occupation.

There are some interesting information that I have found:

- Distribution of Loan Original Amount is multi-modal distrubtion. It seems that loans are made in the multiple of $2500 or in the multiple of $5000. I also notice that there are so many $4000 loans. The reason is that it is more difficult to borrow more than $4000 on Prosper.
- The $4000 loan corresponds to the monthly payment of $175.

- The number of loans is increasing sharply. It is a good sign for investors.
- Prosper provide mostly 36-months and 60-months loans.
- We can use borrower rate as an interest rate.
- The number of past due loans and charged-off are small. We can assume if investors diverse their funds into multiple loans, the risk of being charged-off or defaulted is small. (We will check this assumption in bivariate analysis or in multivariate analysis)
- Majority of borrowers are employed and have income ranging from $25000 to $10000+. And more than 90% of them have proof for their income.
- Half of the loans are made by homeowners.
- Debt To Income Ratio is centered around 0.25. And very few people have debt to ratio above 0.6, which means very few people are at high risk of debt that they could not pay back their loans.
- More than 99% of borrowers have zero public records in the last 12 months. It means public records in the last 12 months is not a good indicator to assess the risk of default loans.
- Whereas to public records in the last 12 months, inquires in the last 6 months and current deliquencies are better indicators. Most of borrowers have 0 to 6 inquires in the last 6 months. Most of them currently have 0 to 4 deliquencies.

Since I have used alot of bar plots in univariate plot session, I have created a function as called as draw_bar_plot. In this function, it has already clean the data such as removing NA and tidy them by generating factors for categorical variables.

My main interest in this dataset are ProsperRating and ProsperScore. These are metrics that Prosper have used to evaluate potential loan opportunities. Some questions should be asked for bivariate and multivariate analysis:

- Which are factors that influence or have relationship with ProsperRating ?
- Which are factors that influence or have relationship with ProsperScore ?
- Is there any overlap among these two metrics?
- Which risk-assessment metric perform better? ProsperRating or ProsperScore?

In term of characteristic of Loan, I will try to find the relationship among these following features such as LoanOriginalAmount, MonthlyLoanPayment, LoanOriginationQuarter, Term, BorrowerRate, LoanStatus.

In term of risk assessment, some supporting features that I have used to analyze ProsperRating and ProsperScore are: EmploymentStatus, IncomeRange, IncomeVerifiable, IsBorrowerHomeowner, DebtToIncomeRatio, CreditScoreRangeLower (lower credit score expose more risks than upper level), InquiresLast6Months, CurrentDeliquencies.

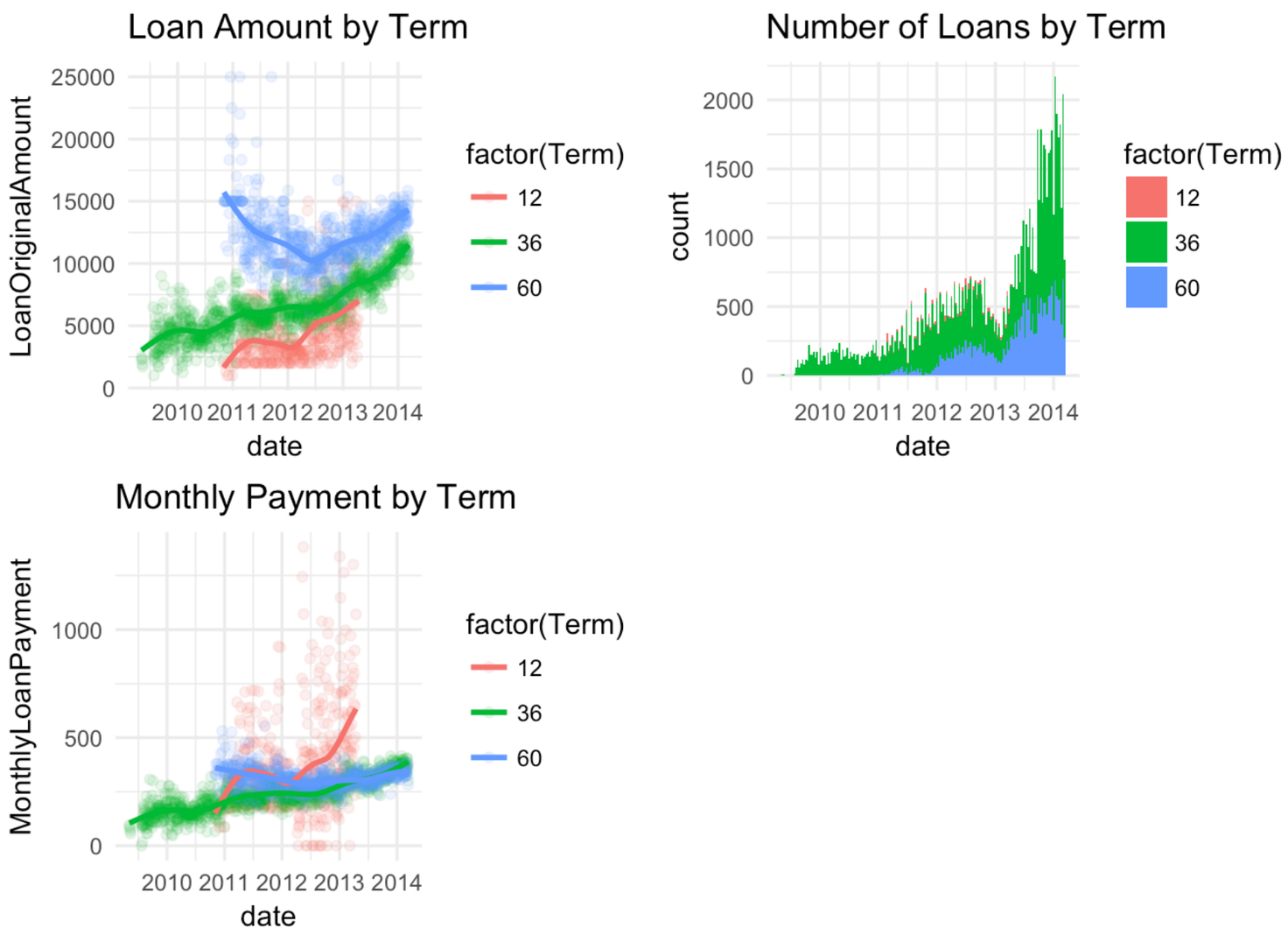
Bivariate Plots Section

1. Characteristics of Loan:

- **Loan Original Amount, Monthly Loan Payment and Loan Term**:

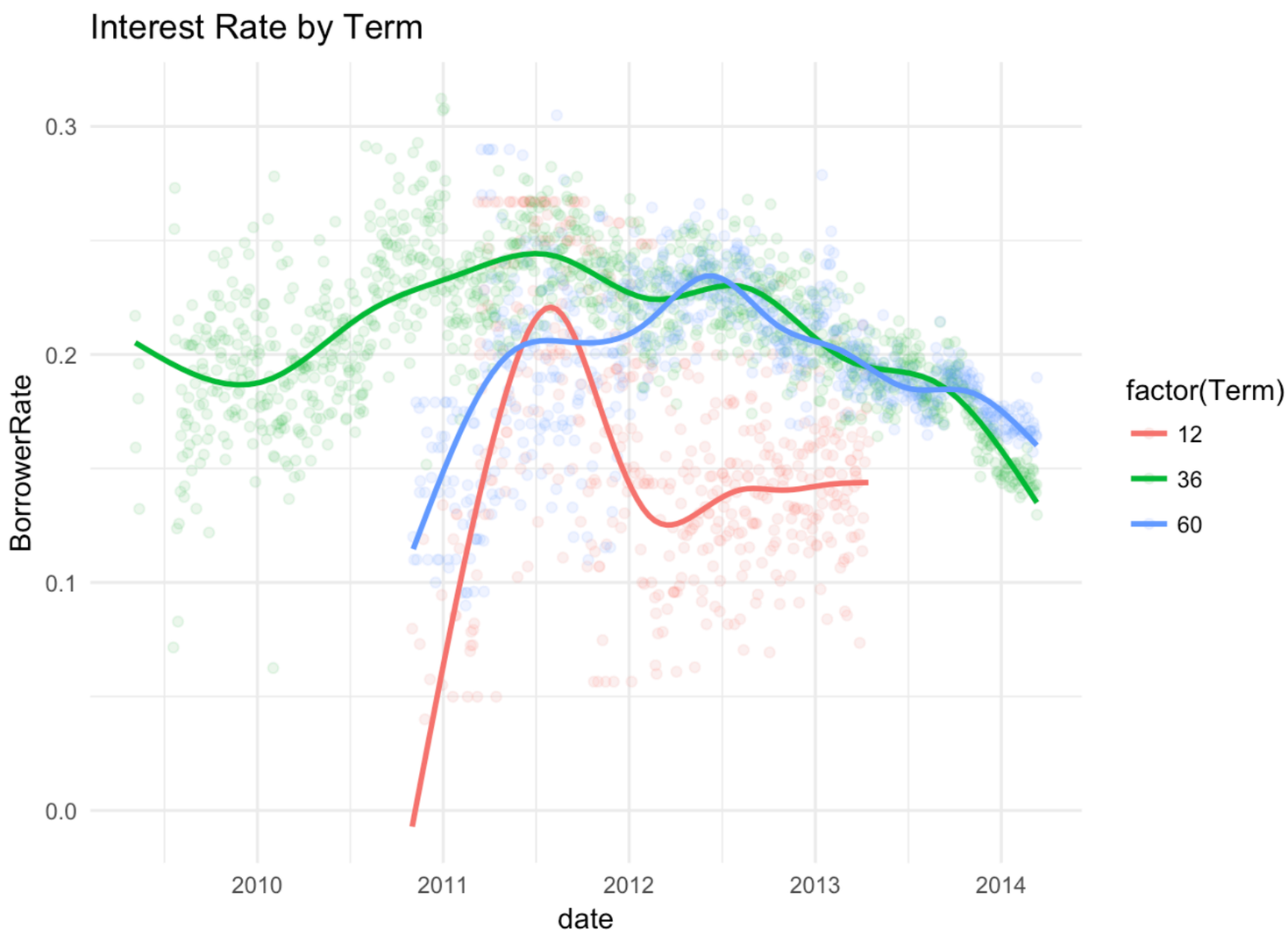
In order to know the relationship between loan original amount and loan term, we need to create another vairable as called as date. Because these are time series data, we need to have date (fortmat: ymd) for each observation. It can be extracted from the variable LoanOriginationDate (format : ymd_hms).

```
loan$date <- loan$LoanOriginationDate %>% as.Date
```



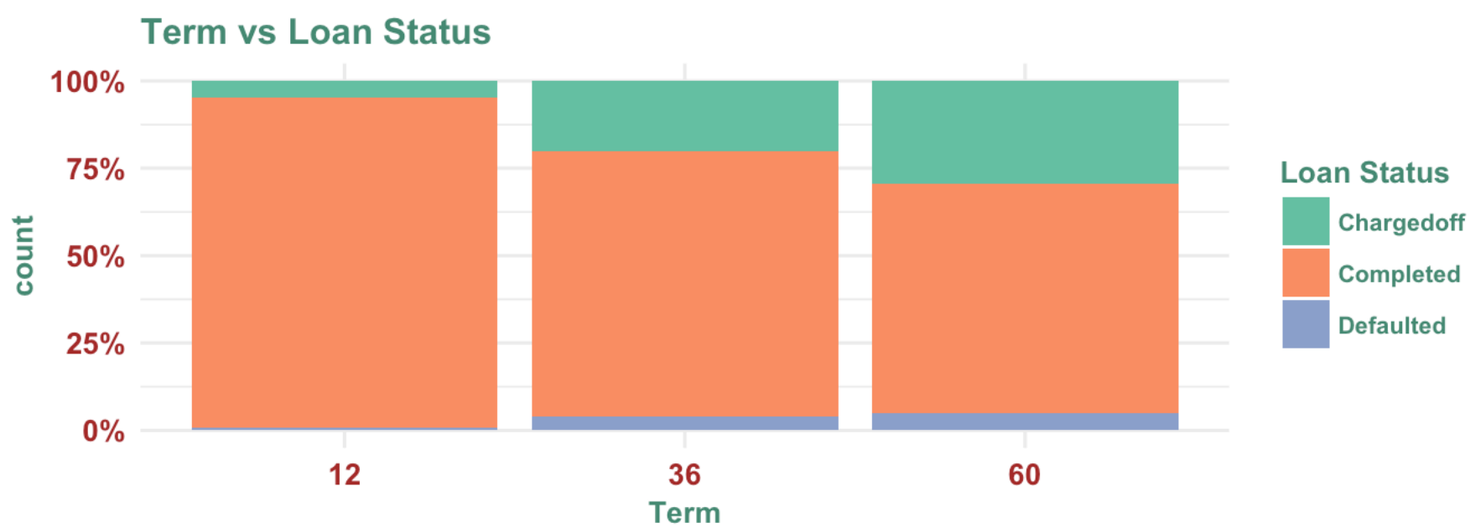
Both loan amount and loan payment has increased every year. Larger loan amount has longer term. And the number of loan has been increasing, especially for 36-months loans and 60-months loans. We notice that the average loan amount has increased for 12-month loans and 36-months loans. However, on average, 60-months loan amount had dropped from 2011 to mid 2012 then has gone up again since mid 2012.

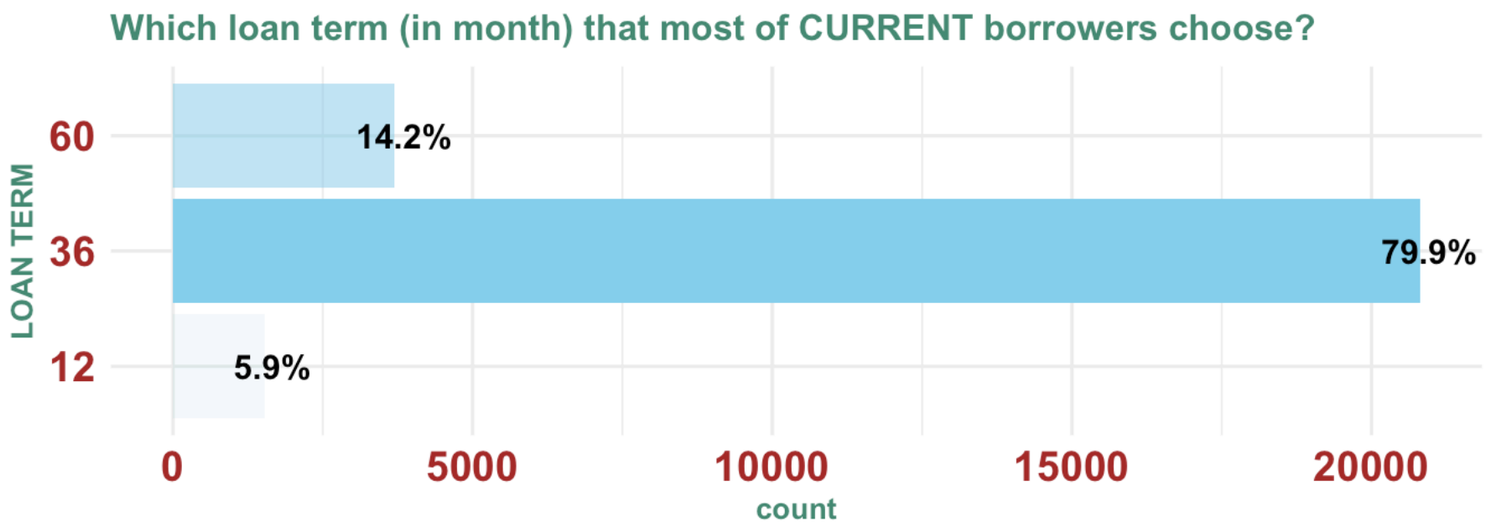
- **Interest Rate and Loan Term**:



Interest rate went up from 2009 to 2011 and it has dropped since 2012.

- **Loan Term and Loan Status**:





From Univariate Plot of Loan Status, there are 4 loan status that we need to focus: Defaulted, Completed, Chargedoff and Current.

It is surprising that the percentage of charged-off loan and defailted loan is higher for longer loan term. Most of the current loans are 36-months loans. Only 14.2% of borrowers makes 60-months loans and 5.9% of borrowers makes 12-months loans.

2. Credit Risk & Loan Assessment:

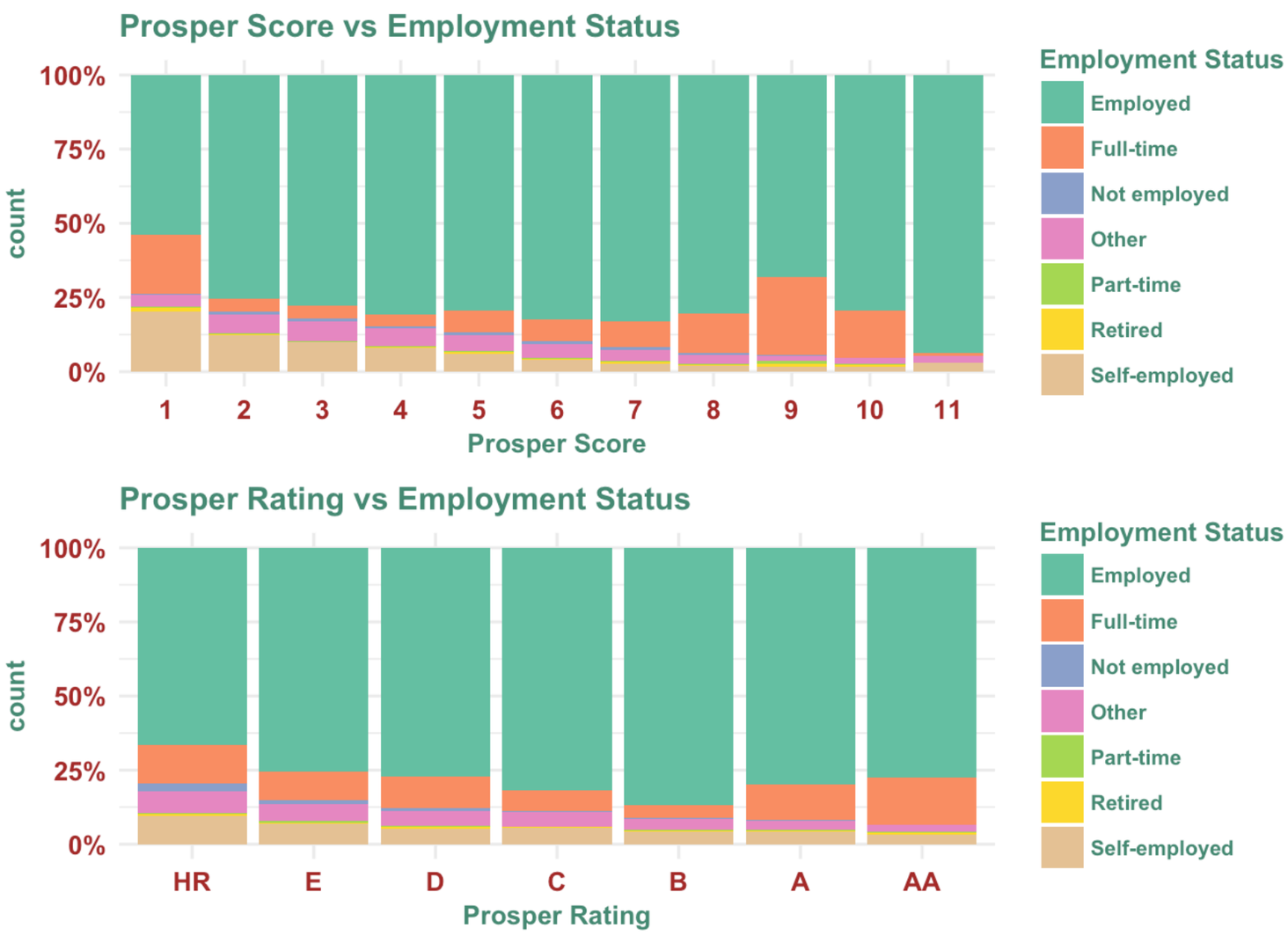
We conduct analyses of credit related variables that investors might use to evaluate potential loan opportunities. However, what is less commonly known is the fact that Prosper has more than one credit rating. They are Prosper Score and Prosper Rating (numeric or alpha are basically the same thing).

With Prosper Score, 1 is highest risk and 11 is lowest risk.

With Prosper Rating, the risk increases in the following order: HR, E, D, C, B, A, AA.

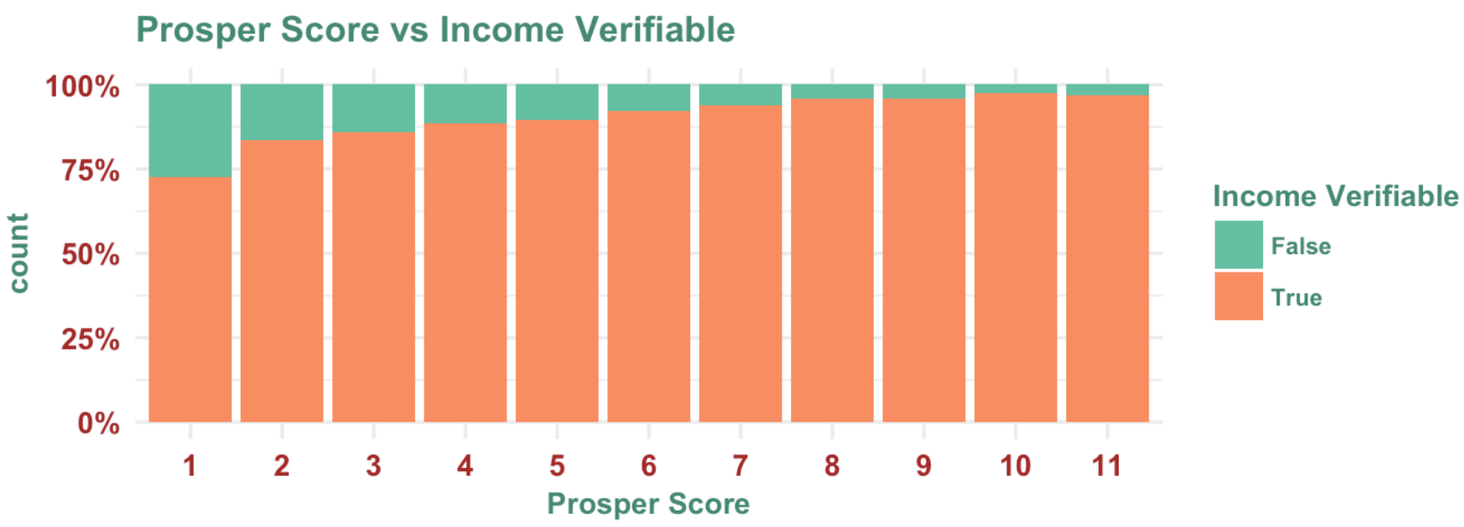
So which one is the better metric? Let's find out.

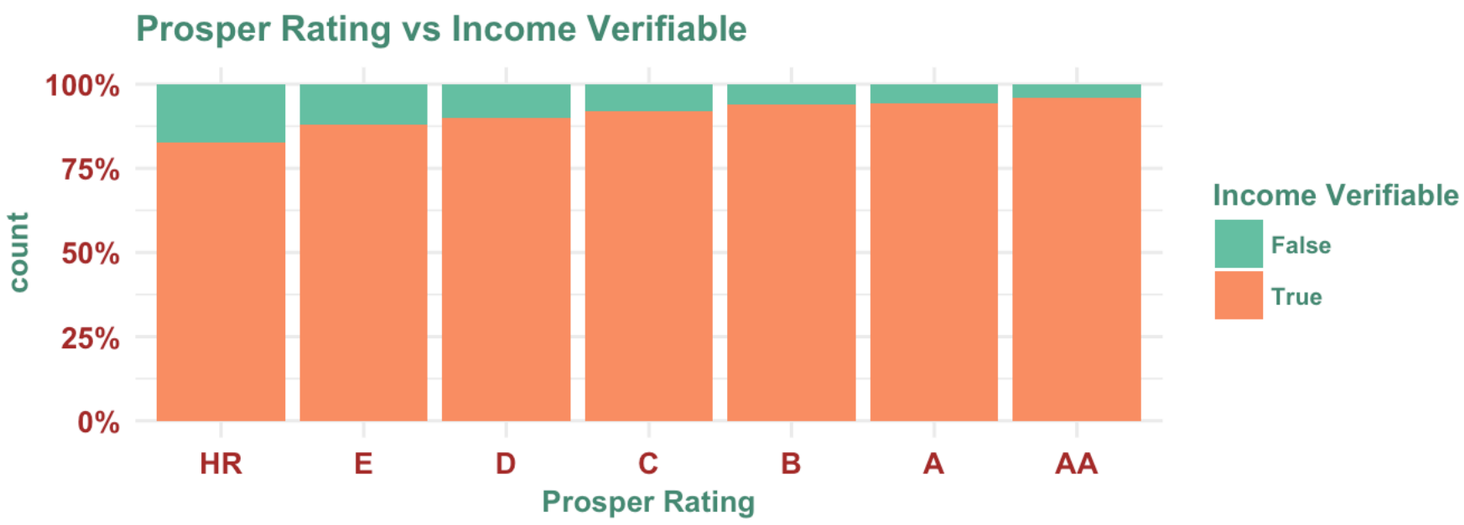
- **EmploymentStatus**:



We will drop Employment Status as a variable that we are interested in. Because the above two charts do not show any meaningful information. For example, 'Employed' has already included 'Full Time' and 'Part Time'; but the person lists his/her employment status as 'Full Time' or 'Part Time'. Also 'Other' means 'Employed' or 'Not employed', or maybe 'Retired'.

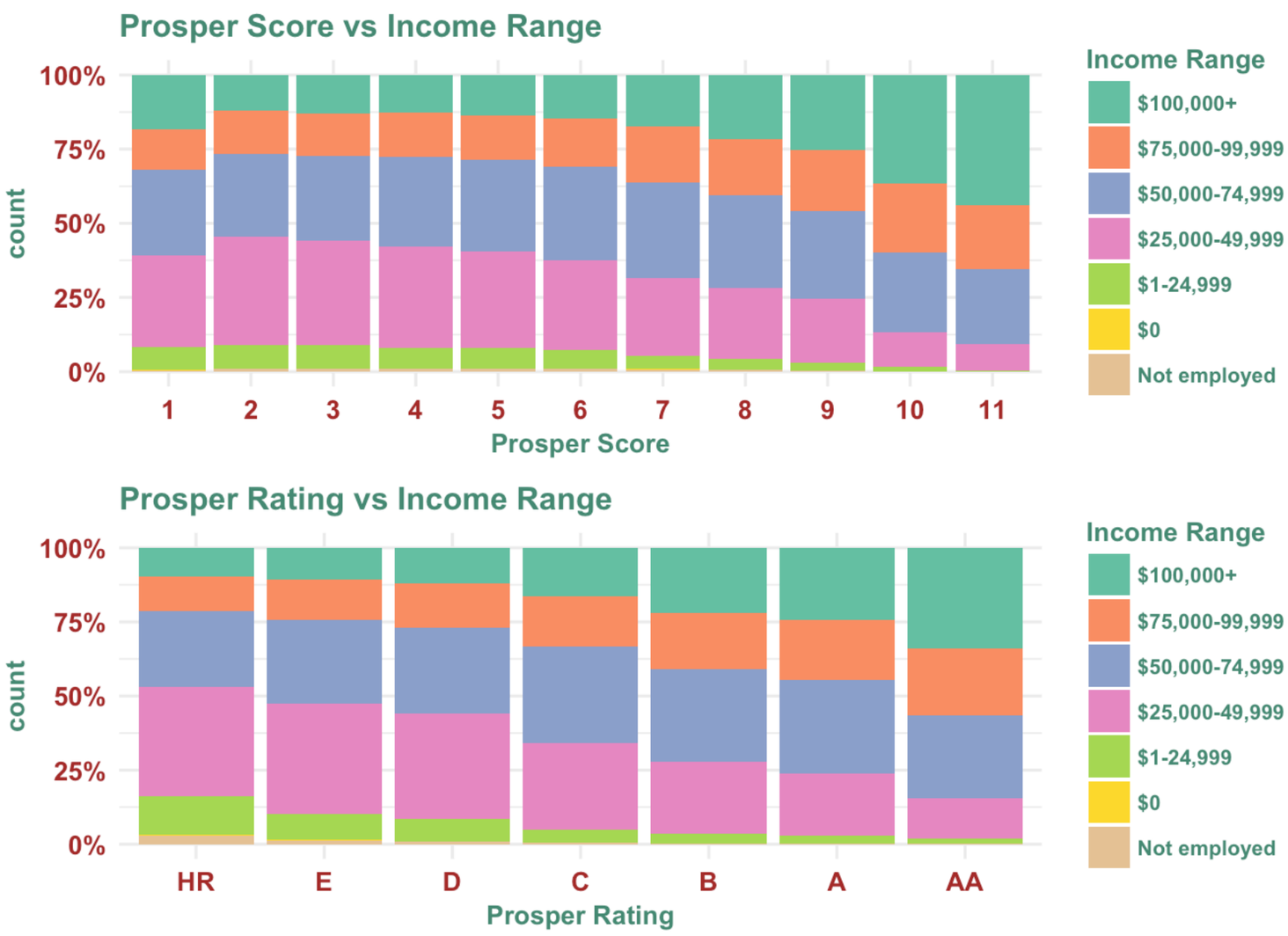
- **IncomeVerifiable:**





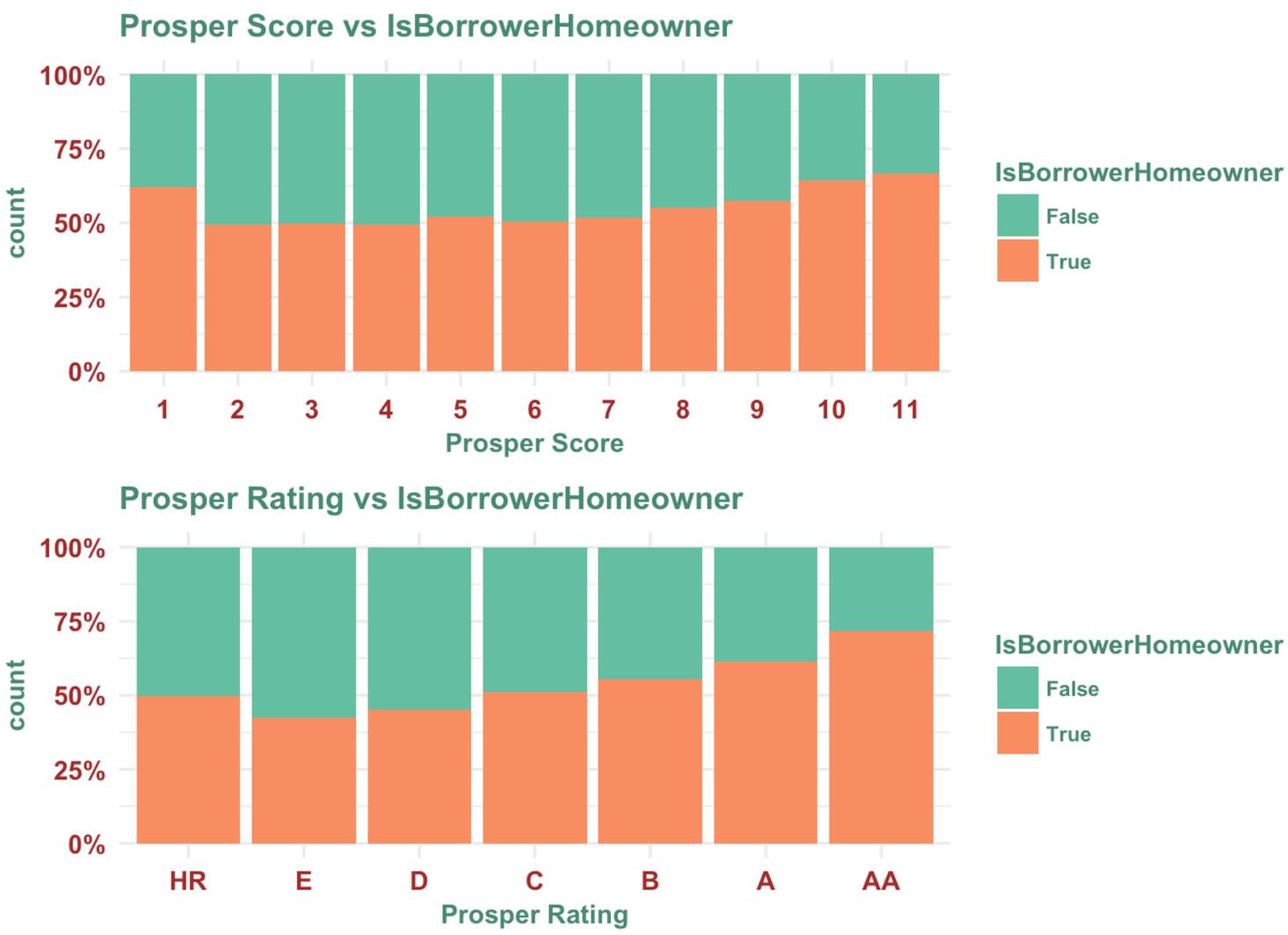
Both chart shows very few people cannot show proof of their income can be rated as low risk. We can see that Income Verfiable are very consistent with both metrics.

- **IncomeRange**:



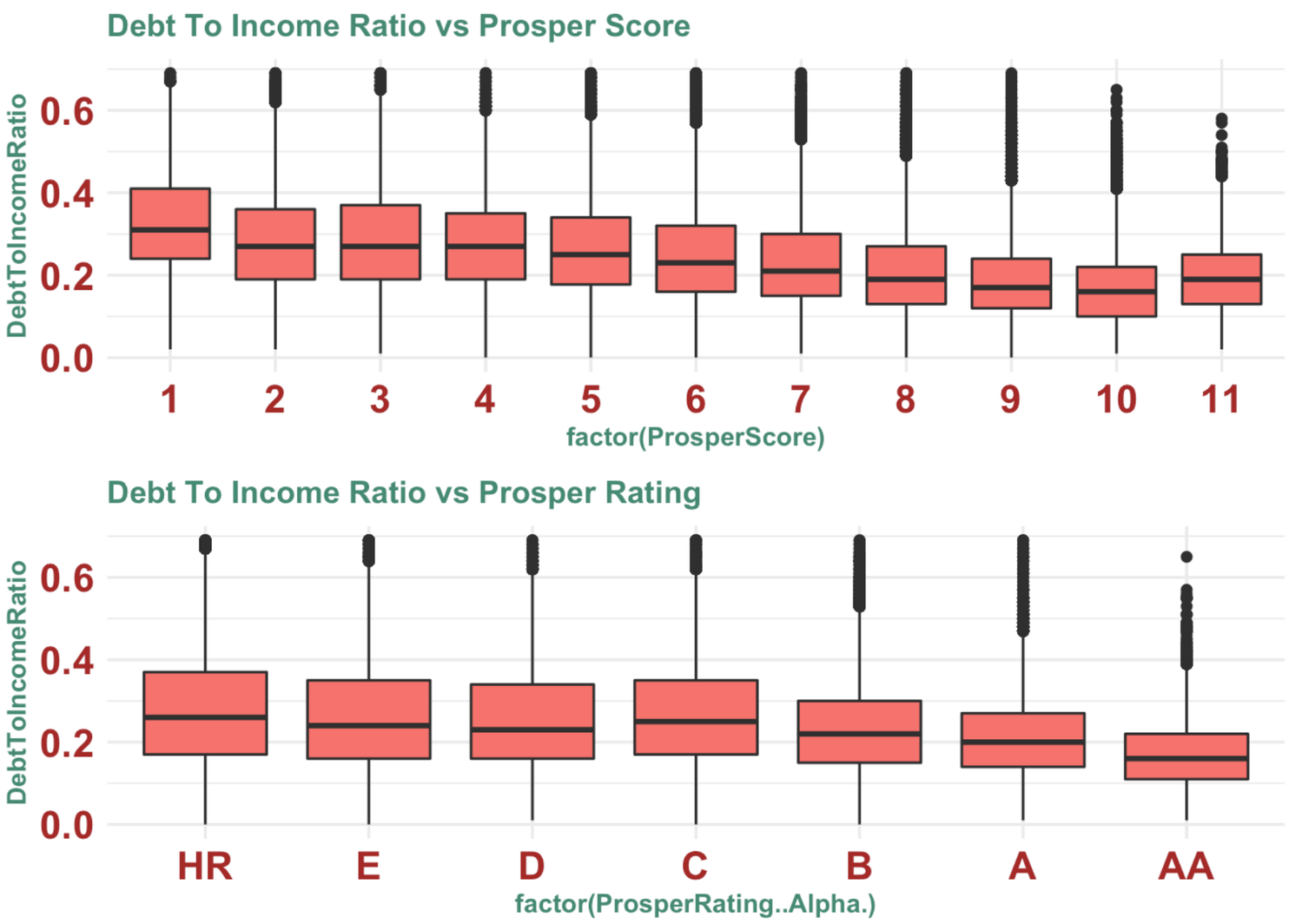
It is the income range that it seems to be not associated with Prosper Score in some cases. The portion of people who have income in the range of $75000-$99999 and $10000+ are nearly identical in prosper score 2, 3,4 and 5. Whereas, income range is very consistent with Prosper Rating. There are more higher income people in lower risk loan.

- **IsBorrowerHomeowner**:



There are many homeowners are in high risk group. The Prosper Rating is also more consistent in this case.

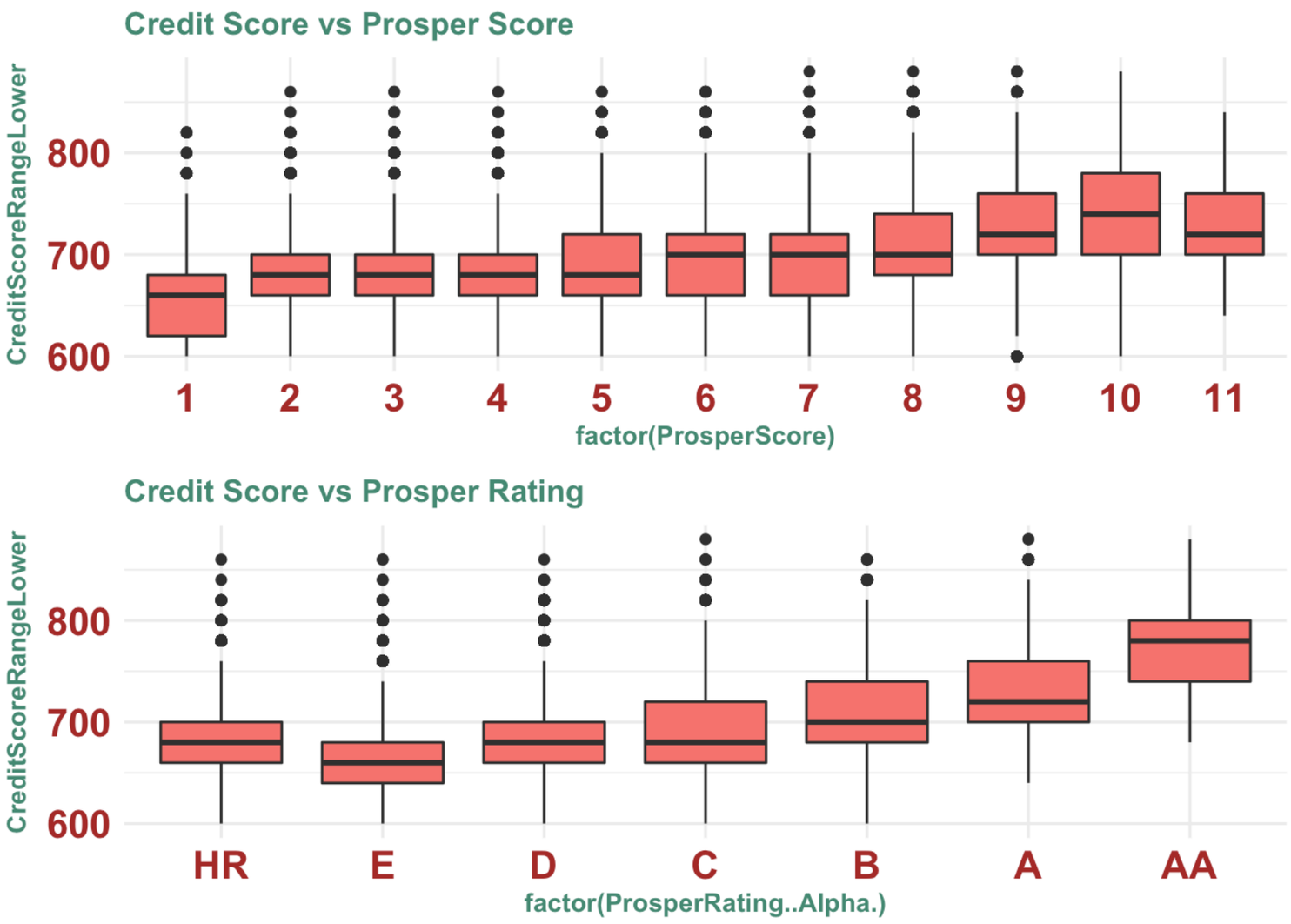
- **DebtToIncomeRatio**:



It is hard to identify which one is better in this case. Except the last score (which is 11), Prosper Score is showing good downward trend with debt to income ratio for lower risk loan.

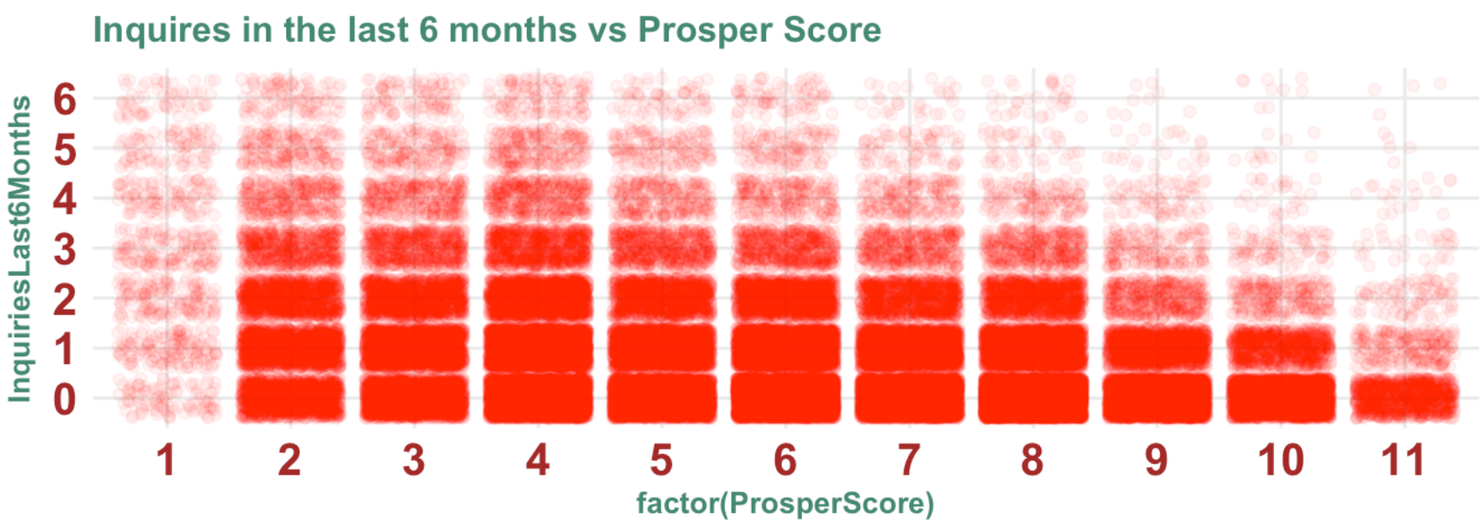
Prosper Rating does not perform well with Debt To Income Ratio. When rating C is lower risk than rating D, but corresponding debt to ratio with rating C is higher than debt to ratio with rating D.

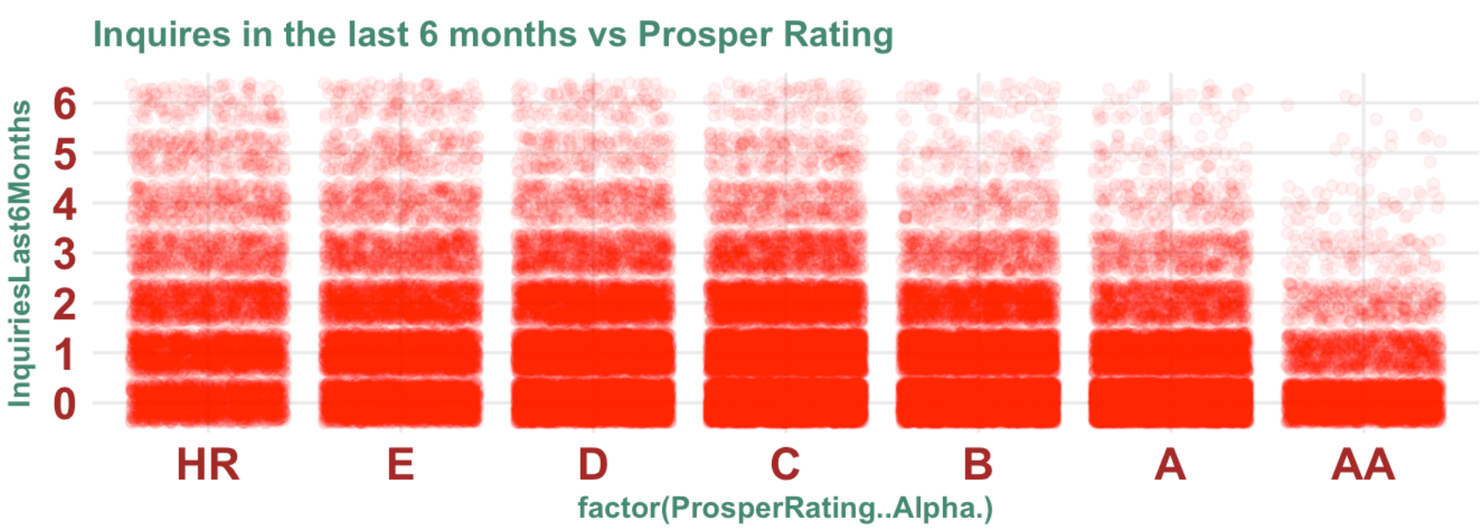
- **CreditScoreRangeLower**:



High credit score means the borrowers have very good credit history. In other words, high credit score borrowers should be placed as lower risk. In this case, Prosper Rating performs well to identify higher credit score borrowers, whereas Prosper Score perform poorly in the same task.

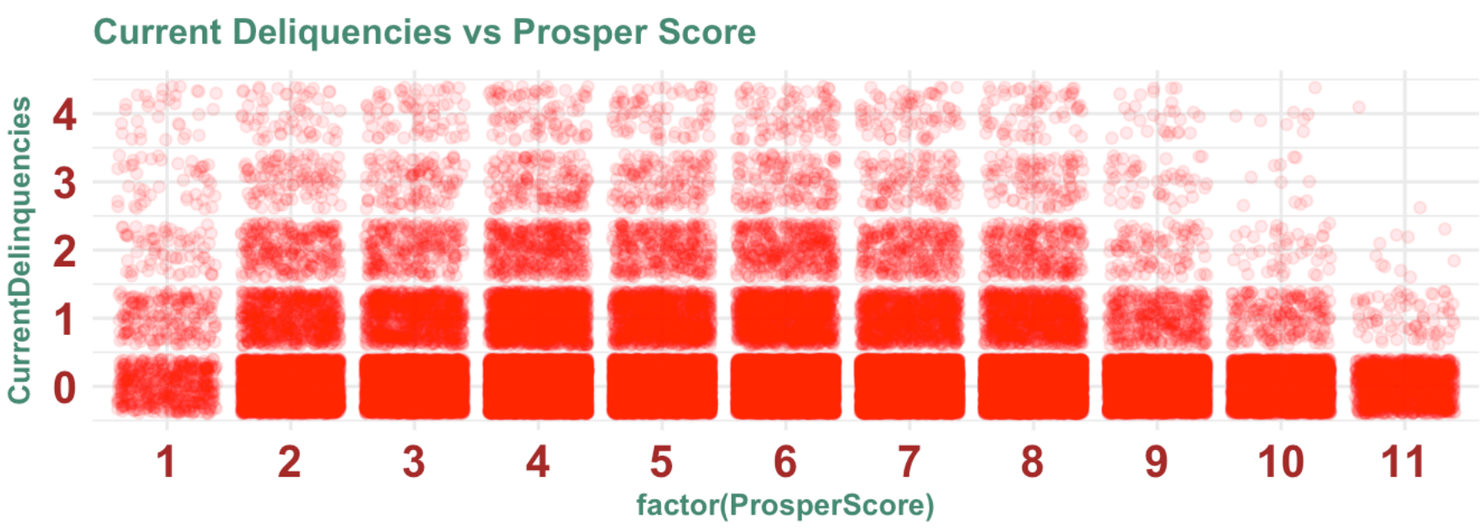
- **InquiresLast6Months**:





It seems that Prosper Score and Prosper Rating don't weigh in the factor 'Number of Inquires in the last 6 months'

- **CurrentDeliquencie**:



As similar to 'number of Inquires in the last 6 months', Prosper Score and Prosper Rating does not weigh in the factor 'number of current deliquencies' in their risk assessment.

Bivariate Analysis

Some interesting relationships that I have found in the characteristics of loan is:

- Both loan amount and loan payment has increased every year. Larger loan amount has longer term
- The percentage of charged-off loan and defailted loan is higher for longer loan term
- Interest rate has dropped since 2012.

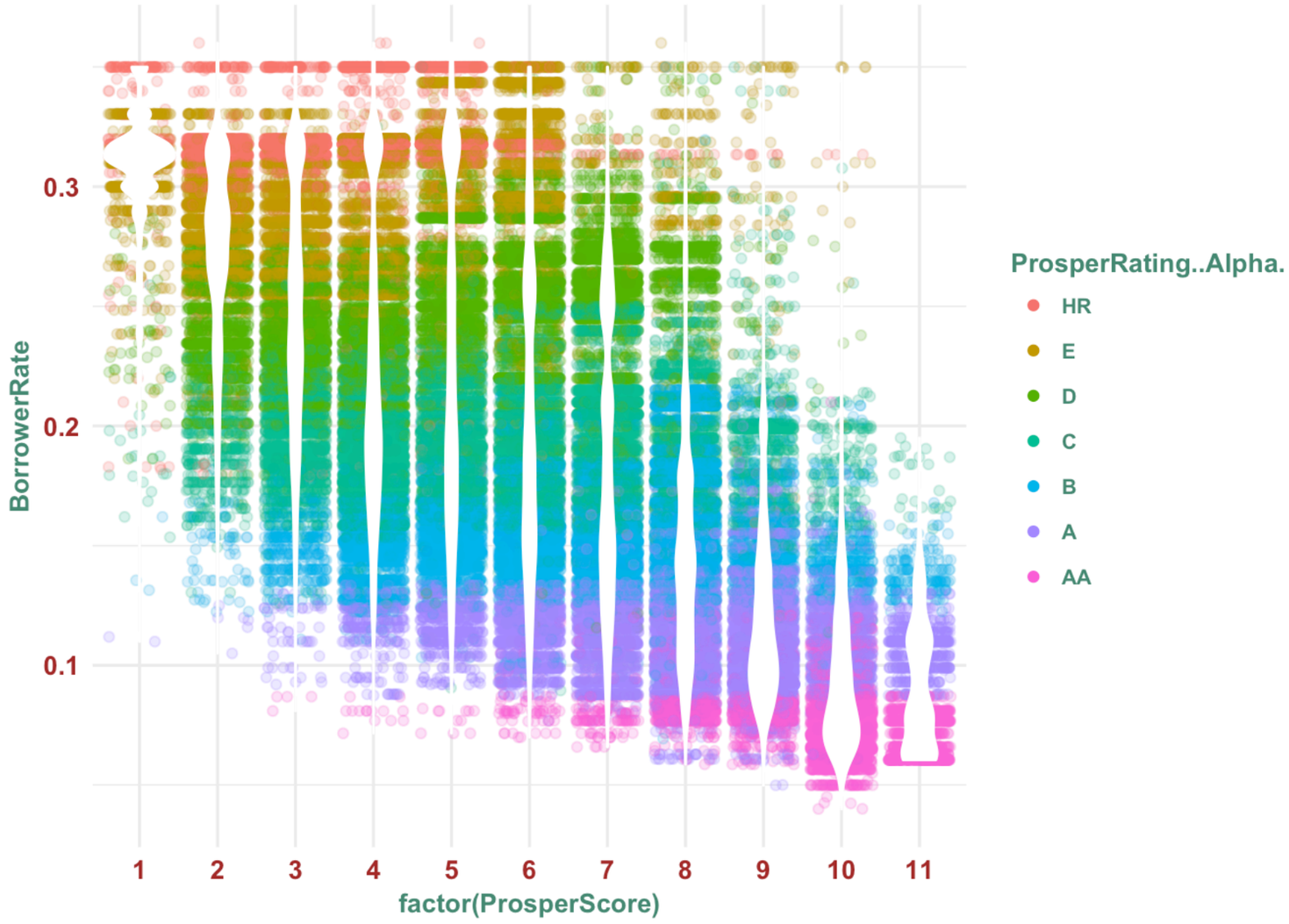
About Loan assessment, It seems that Prosper Rating is a better metrics than Prosper Score:

- Income Verfiable are very consistent with both metrics.
- It is the income range that it seems to be not associated with Prosper Score in some cases. The portion of people who have income in the range of $75000-$99999 and $10000+ are nearly identical in prosper score 2, 3,4 and 5. Whereas, income range is very consistent with Prosper Rating. There are more higher income people in lower risk loan.
- The proportion of homeowner in lower risk loan is higher.
- Prosper Rating performs well to identify higher credit score borrowers, whereas Prosper Score perform poorly in the same task

For the multivariate analysis, we need to combine between the characteristics of loan and credit risk assessment by examine the relationship between the variables of two groups:

- Characteristics of Loan: Loan Original Amount, Loan Term, Monthly Loan Payment, Loan Status.
- Credit Risk: IncomeVerifiable, IncomeRange, IsBorrowerHomeOwner, CreditScoreRangeLower.

Multivariate Plots Section



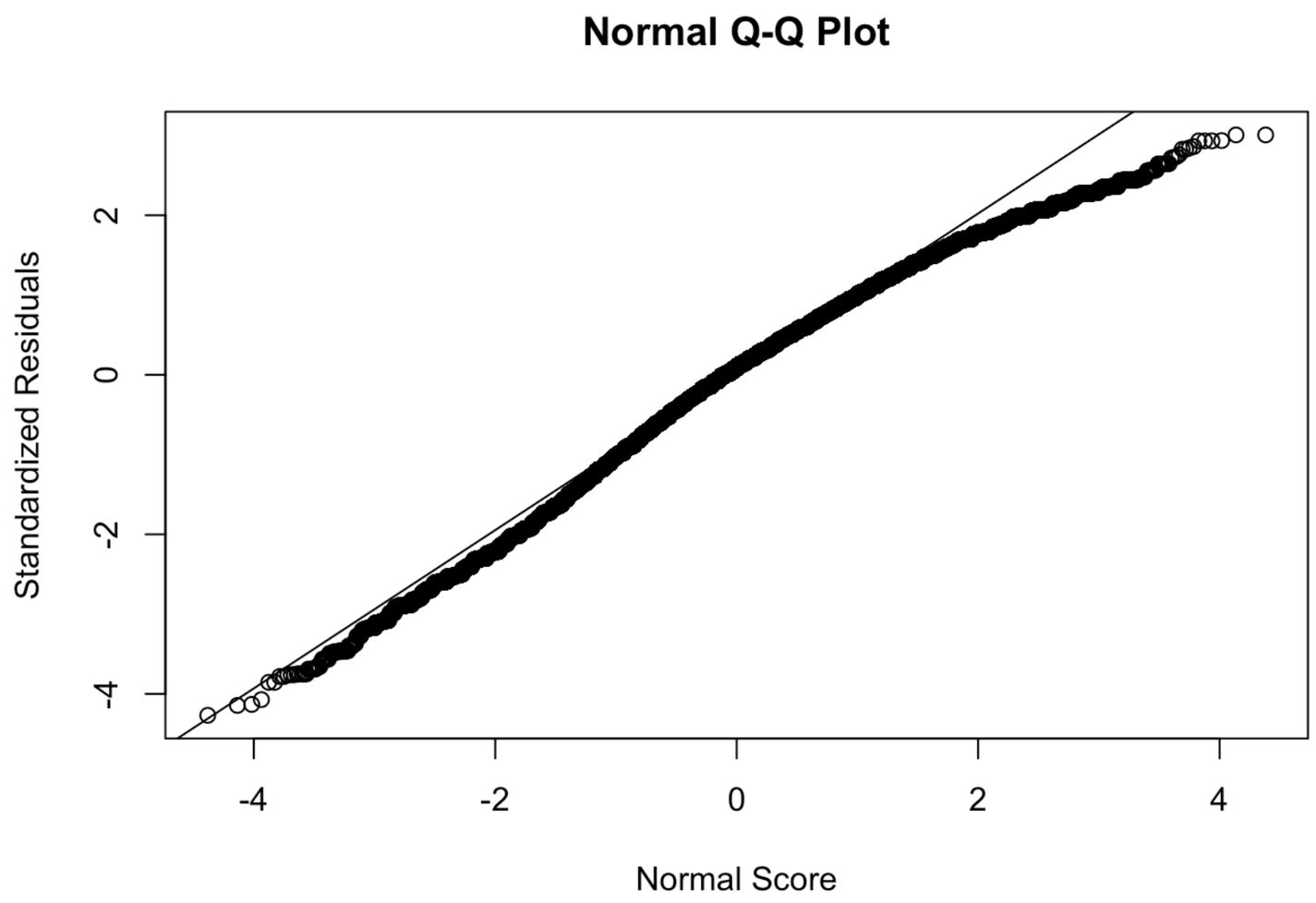
The chart above shows the average interest rates for each combination of rating and score on loans originated since 2009. What we can see is that while interest rates are very consistent within alphabetic credit grades, they can vary widely among loans of the same numeric score.

Therefore we can conclude that Prosper Rating perform better at assessing risk and are more consistent with interest rate than Prosper Score. So we will use Prosper Rating for the further analysis.

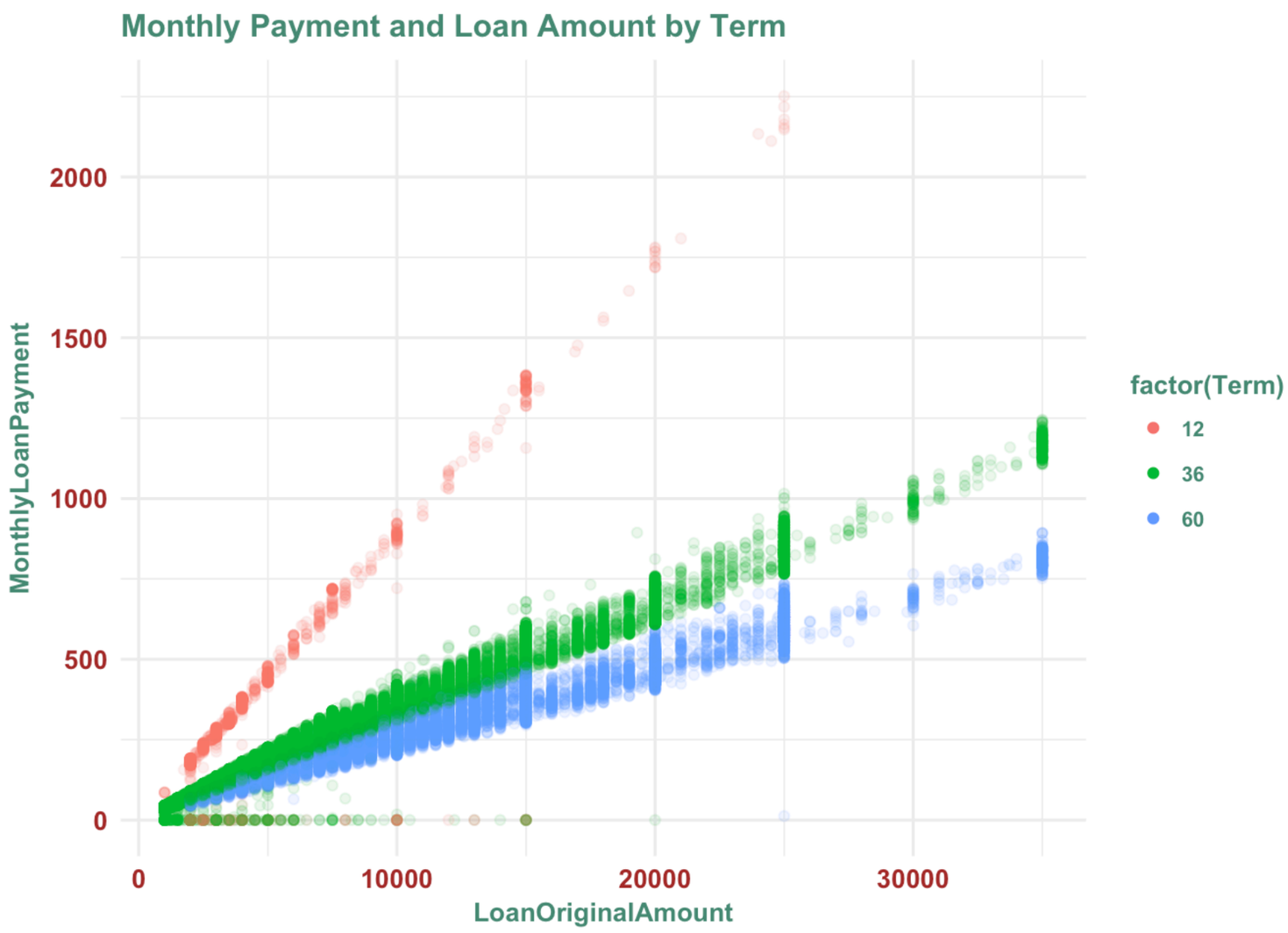
Based on the bivariate analysis, I want to check the linear relationship between ProsperRatingNumeric and predictors such as IncomeVerifiable, IncomeRange, IsBorrowerHomeOwner, CreditScoreRangeLower; and I want check the linear relationship between ProsperRatingScore and predictors such as IncomeVerifiable, IncomeRange, IsBorrowerHomeOwner, CreditScoreRangeLower

```
ProsperRating.lm %>% summary
```

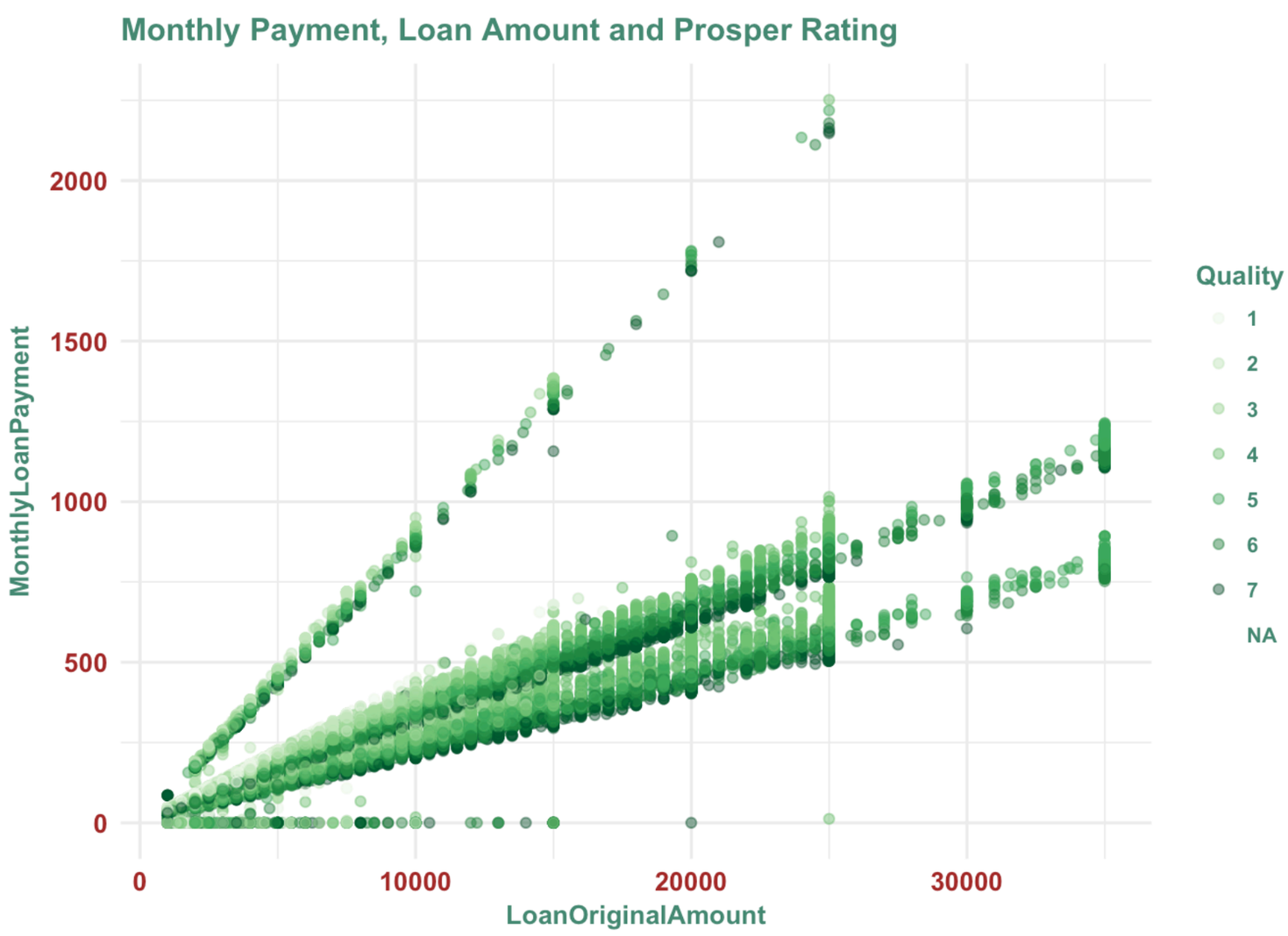
```
##
## Call:
## lm(formula = ProsperRating..numeric. ~ CreditScoreRangeLower +
##     factor(IncomeVerifiable) + factor(IncomeRange) + factor(IsBorrowerHomeowner) +
##     -1, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7461 -0.8490  0.1231  0.9532  4.0476
##
## Coefficients:
##                                   Estimate Std. Error  t value Pr(>|t|)
## CreditScoreRangeLower            1.950e-02  1.027e-04  189.852  < 2e-16
## factor(IncomeVerifiable)False   -1.034e+01  7.813e-02 -132.372  < 2e-16
## factor(IncomeVerifiable)True    -9.687e+00  7.674e-02 -126.234  < 2e-16
## factor(IncomeRange).L           -1.353e+00  8.333e-02  -16.234  < 2e-16
## factor(IncomeRange).Q            7.435e-02  3.161e-02    2.352 0.018650
## factor(IncomeRange).C            3.038e-01  8.541e-02    3.557 0.000376
## factor(IncomeRange)^4            8.308e-02  1.143e-01    0.727 0.467173
## factor(IncomeRange)^5           -1.097e-01  8.876e-02   -1.236 0.216347
## factor(IncomeRange)^6           -3.391e-02  4.152e-02   -0.817 0.414176
## factor(IsBorrowerHomeowner)True -2.703e-01  1.001e-02  -26.996  < 2e-16
##
## CreditScoreRangeLower           ***
## factor(IncomeVerifiable)False   ***
## factor(IncomeVerifiable)True    ***
## factor(IncomeRange).L           ***
## factor(IncomeRange).Q           *
## factor(IncomeRange).C           ***
## factor(IncomeRange)^4
## factor(IncomeRange)^5
## factor(IncomeRange)^6
## factor(IsBorrowerHomeowner)True ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.346 on 84843 degrees of freedom
##   (28 observations deleted due to missingness)
## Multiple R-squared:  0.9065, Adjusted R-squared:  0.9065
## F-statistic: 8.227e+04 on 10 and 84843 DF,  p-value: < 2.2e-16
```



Based on the coefficient table and QQ-Plot Figure19 , it shows that there is a strong linear relationship and goodness of fit of the model between Prosper Rating and following predictors: CreditScoreRangeLower, IncomeVerifiable, IncomeRange, IsBorrowerHomeowner.



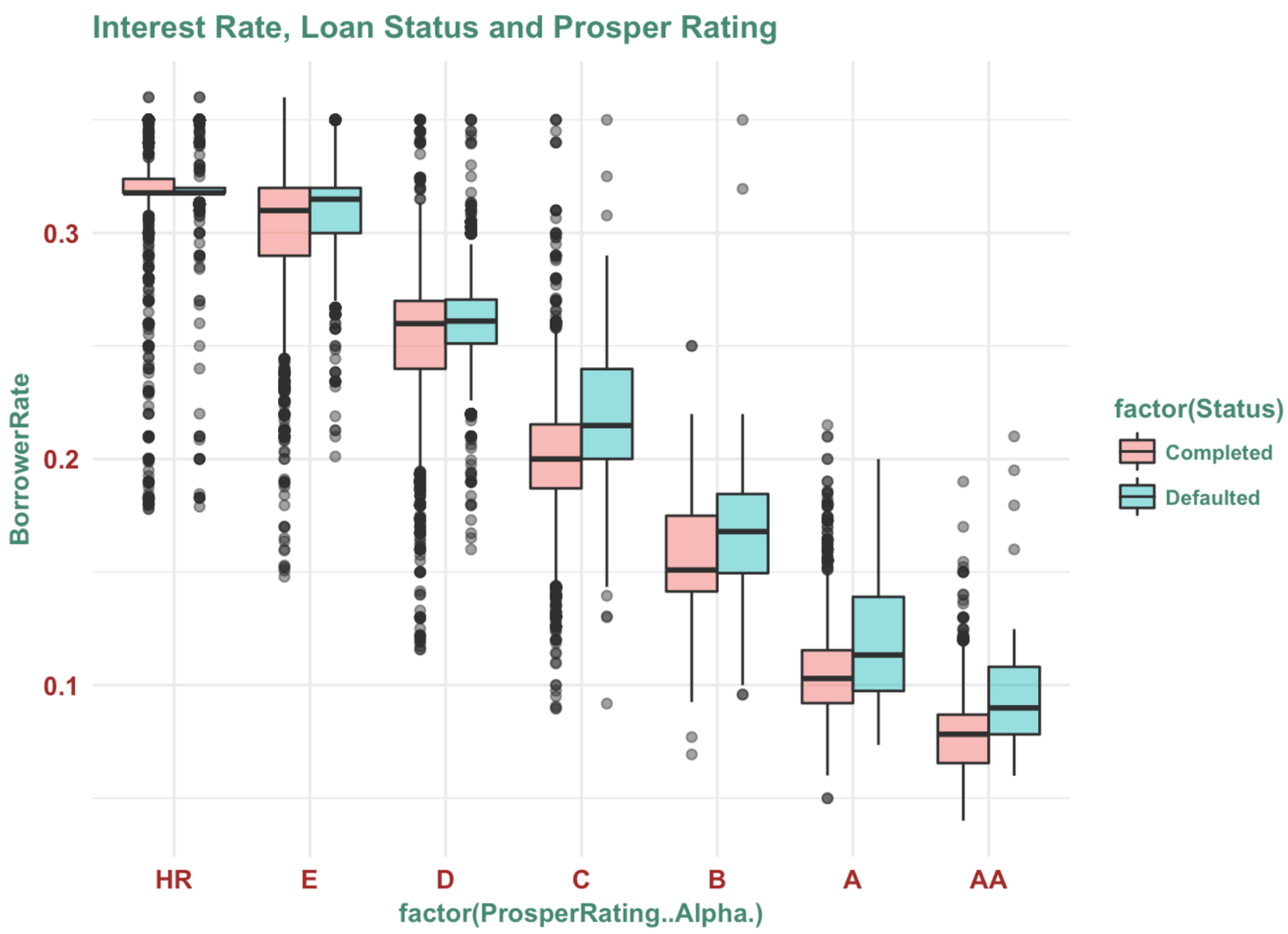
We can see three cluster in the plot between Monthly Payment and Loan Amount by Term. The first cluster is 12-month loan, the second is 36-month loan, the last is 60-month loan.

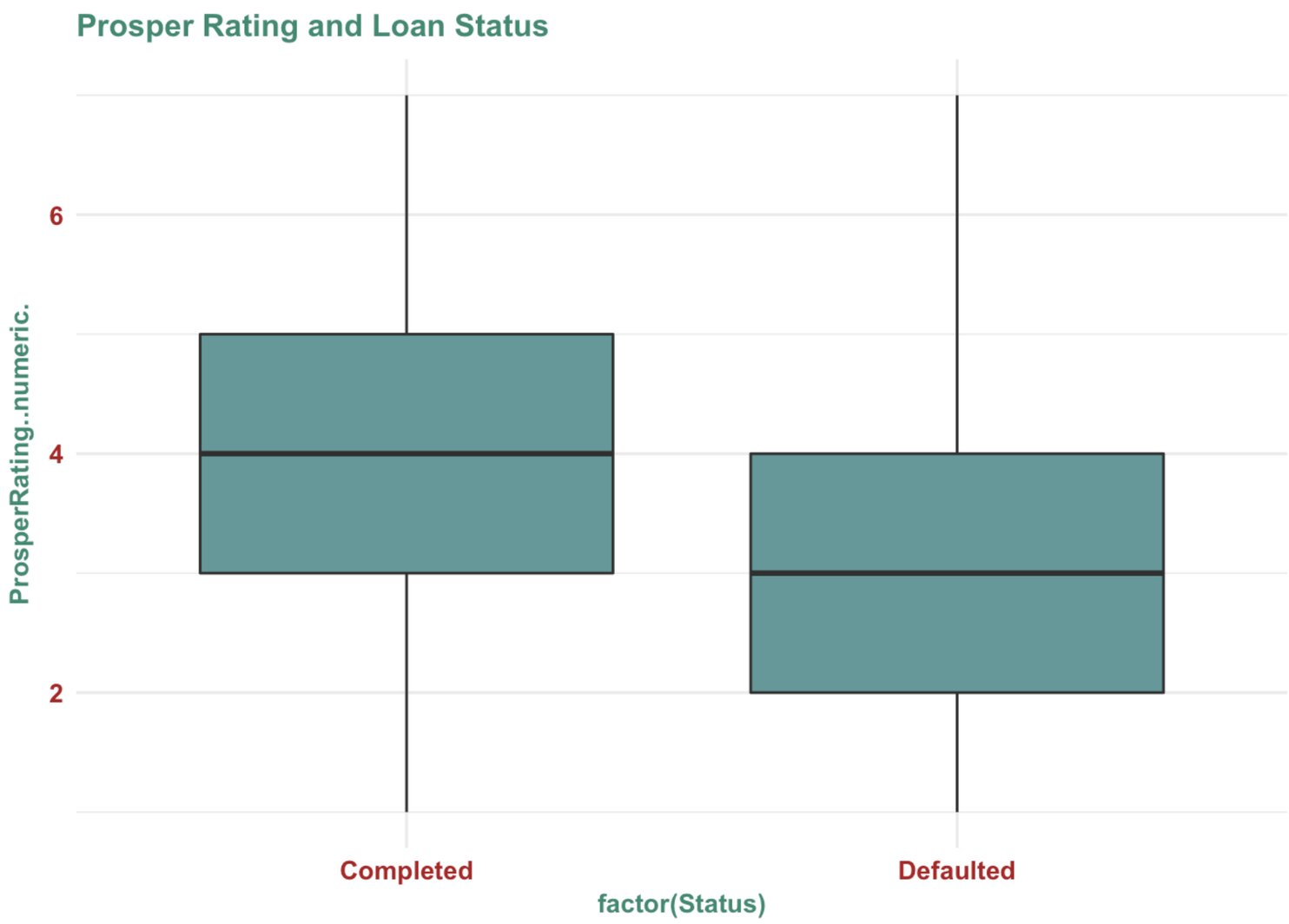


We can see that the top of the scatter plot is dominated by loans with Prosper Rating equal to 1, which represents high risks. Th bottom of the scatter plot is dominated by loans with Prosper Rating equal to 7, which represents low risk. Investors should invest their money to loans which is bigger than $25000 because the above chart shows that the right side of $25000 is dominated of green points, which represent for low risk.

We can see the linear relationship between Interest Rate and Prosper Rating. It seems the risk of bad loan is low with higher Prosper Rating. However it is not clear with this chart. Let's make another that combine 'Defaulted' and 'ChargedOff' into the same name as 'Defaulted'. So all the past loans are splitted into two categories: bad loans and good loans. Bad loan is named as 'Defaulted', and good loan is named as 'Completed'.

```
loan$Status <- loan$LoanStatus
loan$Status[loan$Status == 'Chargedoff'] = 'Defaulted'
```





It is clear to notice that when we move from high risk rating to low risk rating, the density of default loan reduces. Moreover, the boxplot tells us that Defaulted loan and non-defaulted loans(Completed) have a very different Prosper Rating (AA:7, HR: 1).

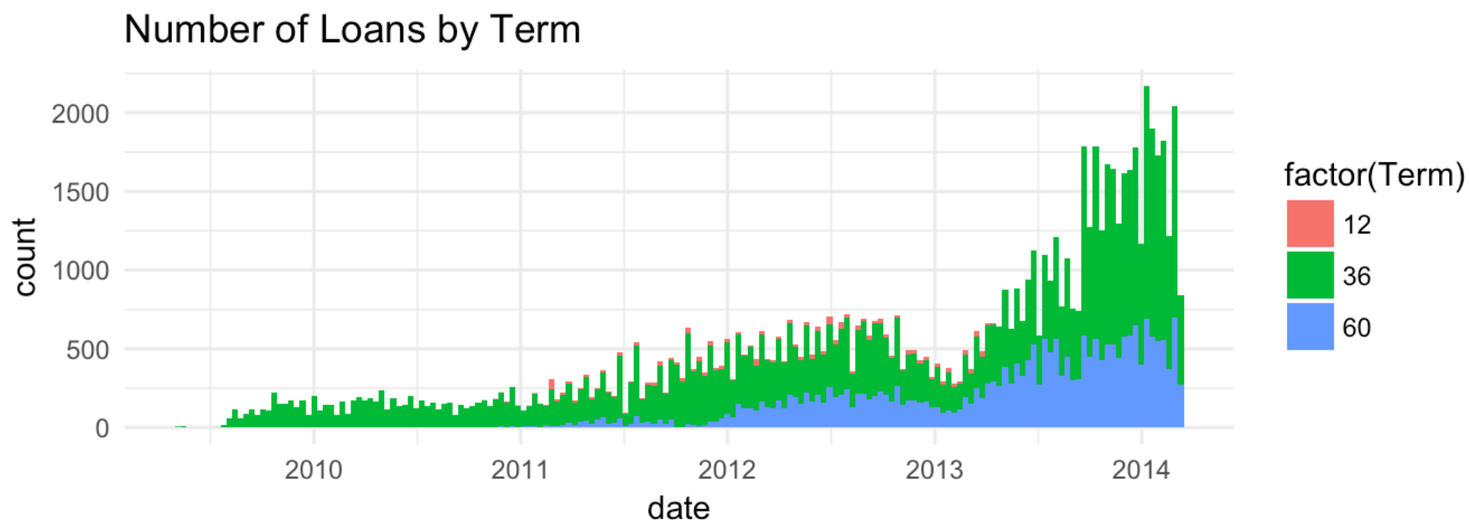
Multivariate Analysis

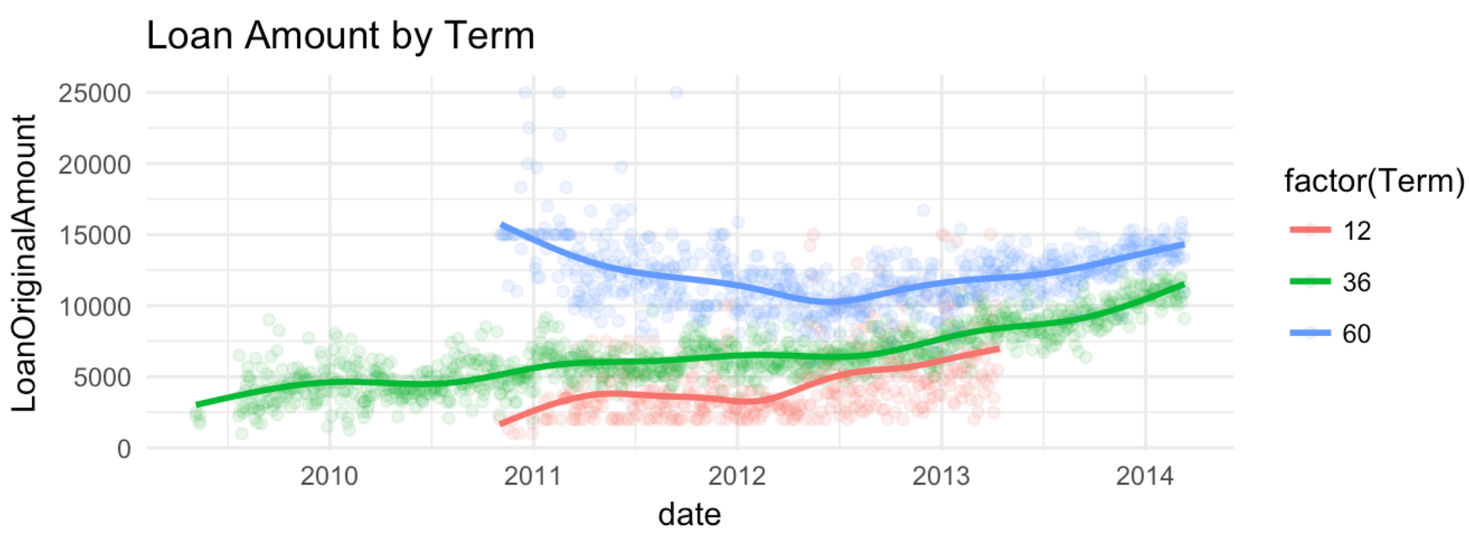
In this part of the investigation, we have observed:

- Prosper Score is not a good indicator for investors. And Prosper Rating is a better metric.
- We have observed that Prosper Rating has influence on Monthly Payment for the same loan original amount.
- When the risk is low, the proportion of bad loan is low. And when the risk is high, the proportion of bad loan is high.
- There is a linear relationship between Interest Rate and Prosper Rating: When the rating is low risk, the interest rate is low. And when the rating is high risk, the interest rate is high. What it means to investor is high risk loans offer high reward, low risk loans offer low yield.

Final Plots and Summary

Plot One



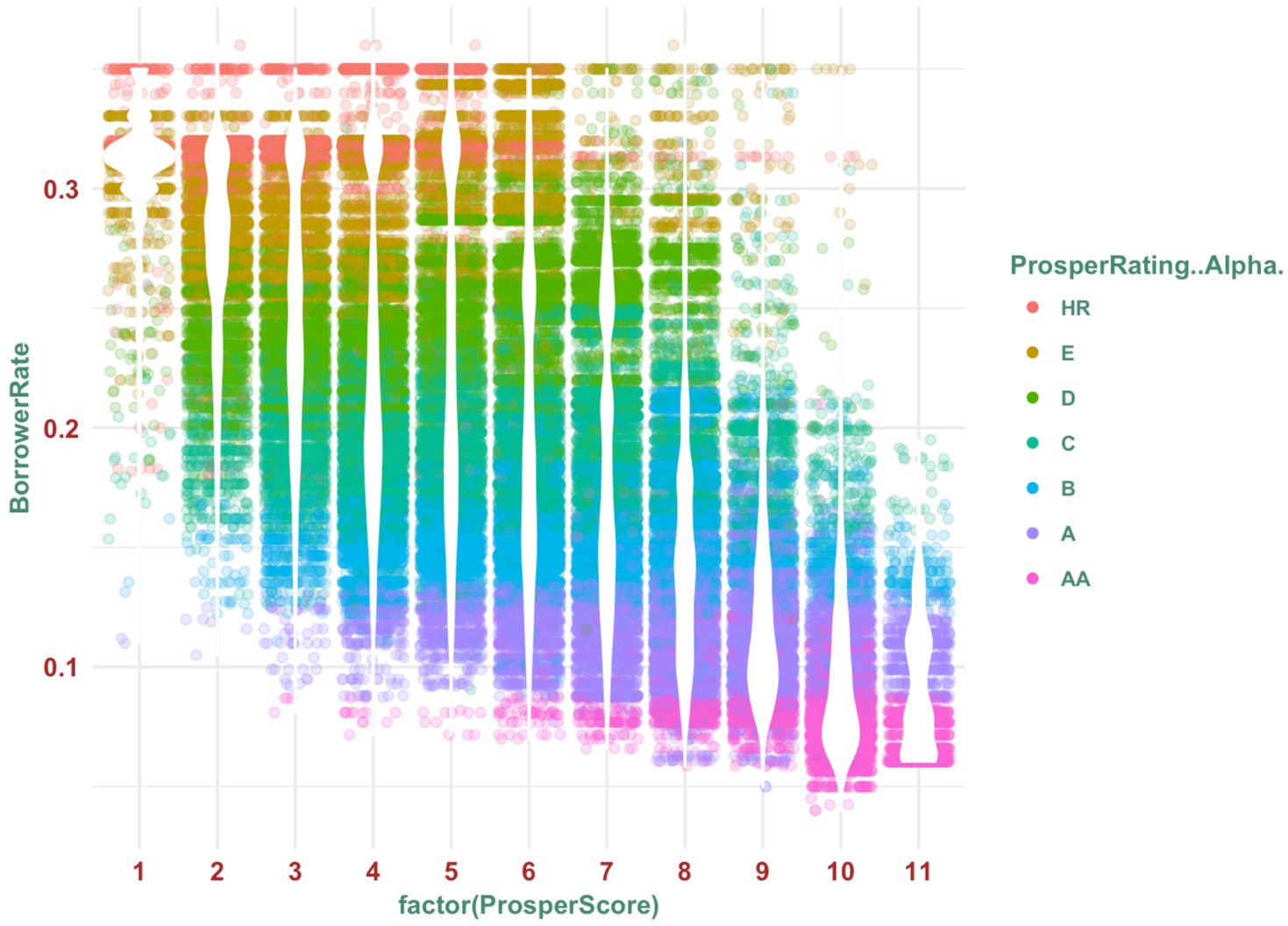


Description One

The plot above describe how the Prosper Loan platform has evolved between 2009 and 2014. The significant increase in number of loans indicates the maturity of the loan platform. In 2009, Prosper issue only 36-months loan. In 2011, Prosper introduce two more loan terms which are 60-months and 12-months. However, the number of 12-months loans are very limited. And the loan amount is increased. In other words, 12-months loans have low volume and high loan amount in a very short term. Investors are exposed to high risk with 12 months loans, therefore Prosper stopped issuing 12 months loan since 2013.

Since 2009, most of the loans are 36-months, and the average loan amount of 36-months has gradually increased from 2009 to 2014. Borrows who borrow 60-month loans can borrow higher amount of money. On average, this amount dropped in 2012 and went up again in 2013.

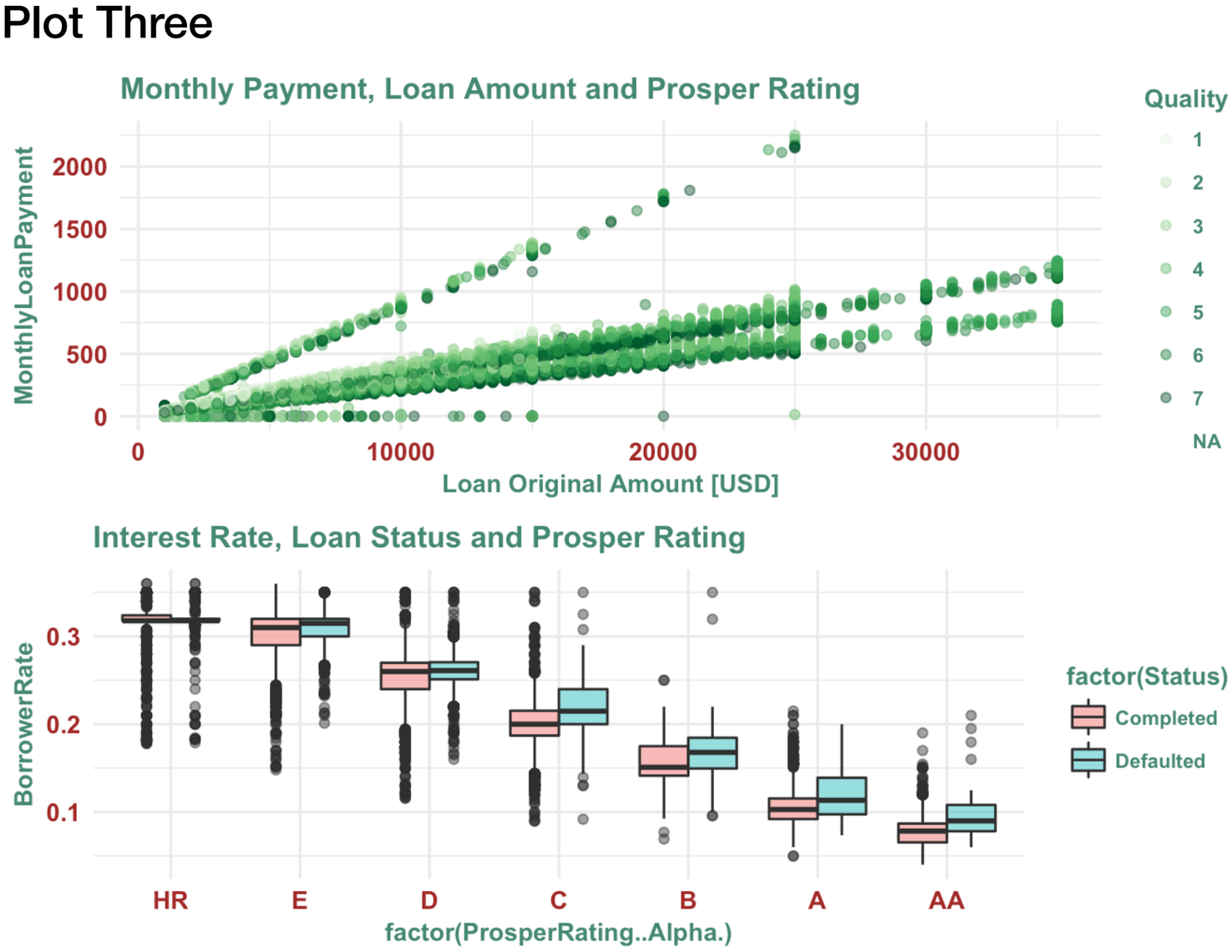
Plot Two



Description Two

- There are many plots in the univariate and bivariate section to make a comparison between Prosper Score and Prosper rating. The objective is to identify which one is more rigorous in risk assessment.
- The idea of this chart is very similar to two-way contigency table with one way is Prosper Score, the other way is Prosper Rating. And the table is based on the third variable as Interes Rate (Borrow Rate). A look through this chart give us an impression that the color factor of Prosper Rating is very consistent with the interest Rate. In other words, when the interest rate is low, the rating is high and vice versa. However, with Prosper score, the interest rate is widely spread. For example, with the average interest rate is 0.25, we can find its correspond prosper scores are in many value as 2,3,4,5,6,7,8.
- The reason why I choose this plot is that I can combine both Prosper Score and Prosper Rating on the same chart. As well as the other plots, this plot also shows Prosper Rating is a better metric in assessing risk.

Plot Three



Description Three

- If we look at the plot between Monthly Payment and Loan Original Amount with Prosper Rating, we find that there is a linear relationship between Monthly Payment and Loan Original Amount. Also, for the same loan amount, the average monthly payment is influenced by Prosper Rating. In other words, Prosper Rating is explained for the variation in monthly payment at a certain level of loan original amount. It means, with the same loan amount, if Prosper rate the borrowers as low risk, their monthly payment is lower than the borrowers who is rated as high risk.
- If we look at the plot between Interest Rate and Prosper Rating with Loan Status(Completed/Defaulted), we find that there is a linear relationship between Interest Rate and Prosper Rating. When the rating is low risk, the interest rate is low. And when the rating is high risk, the interest rate is high.

In conclusion, with low risk rating at a certain loan original amount, the average monthly payment is low and the average interest rate is. With high risk rating at a certain loan original amount, the average monthly payment is high and the average interest rate is high.

Reflection

In the beginning, I stated with exploring the definition of individual variables. It is a very good approach to know what Prosper Loan Platform is and to understand how borrowers and investors are using this platform.So I can classify what I want to analyze into 2 groups: characteristics of loans and credit risk assessment. In addition, I realize that throughout the timeline of Prosper, there are two different credit rating systems. It leads to my decision of subsetting the data because I have to choose the new credit rating, which is used since 2009. Also I have created another variable representing the origination year of each loan.

As I moved to univariate analysis, I have found frequency table, histogram, bar plot very useful to explore trends, patterns and distribution of each variable. I realize that I have used many bar charts. So I have a function as callled as `draw_bar_plot` to reduce the number of code lines that I have to write. And I prefer to put that function into a seperate file named as `pfunction.R` . Based on the finding in characteristics of loans ad credit risk assessment, I compose a list questions that I have to answer in the bivariate analysis and multivariate analysis. Moreover, I found that there is not only one new credit rating. They are Prosper Rating (0 - N/A, 1 - HR, 2 - E, 3 - D, 4 - C, 5 - B, 6 - A, 7 - AA) and Prosper Score (1-11, 1 is highest risk and 11 is lowest risk).

At first, I thought Prosper Score is better because it has wider range. But after moving to bivariate analysis, I realize that Prosper Rating is better at assessing risk. And I have used more stacked bars than the other type of plot in bivariate analysis. Therefore I create `draw_group_bar_plot` to reduce the repeating codes in my report.

One of the problem I got is that the categorical variables or factor variables have made my chart look very messy. So I have to reorganize levels of those variables.

After having analyzed Prosper loan platform, we have observed:

- The number of loans has drastically increase between 2009 and 2014.
- It is harder to borrow more than $4000 on Prosper. And the corresponding average monthly payment for $4000 loan is $175.
- Distribution of loan original amount is multi-modal distribtuion with peaks as multiple of $2500 or $5000. It means Borrowers usually borrow $2500, $5000, $10000 or $15000 etc.
- Prosper has stopped issuing 12-month loans since 2013. Most common loans are 36-months loan.
- In term of risk assessment, there are two metrics such as Prosper Score and Prosper Rating. Prosper Rating is a better metrics.
- There is a linear relationship between Prosper Rating and predictors such as Income Verified, Income Range, IsBorrowerHomeOwner, CreditScoreRangeLower.
- There is a linear relationship between Interest Rate and Prosper Rating.
- When we move from high risk rating to low risk rating, the density of default loan reduces.

Even though we have said that Prosper Rating is a better tool to analyze loan risk for investor than Prosper Score, we must be careful to some situation where potential risk exists in the small sample size. Therefore it is helpful if we can use Prosper Score to evaluate the risk again. Dual score matrix offers further risk segmentation.