Prosper loans

5 October 2017

This report explores a dataset containing 113,937 loans with 81 variables on each loan, including loan amount, borrower rate (or interest rate), current loan status, borrower income, etc. The source of the data is Prosper, or Prosper Marketplace, an online peer-to-peer lending company.

The detailed description of the variables can be found on this page: https://docs.google.com/spreadsheets/d/1gDyi_L4UvIrLTEC6Wri5nbaMmkGmLQBk-Yx3z0XDEtI/edit#gid=0

Number of records and variables

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

After checking the dataset for duplicated keys and missing values, I noticed that:

- ListingKey has duplicated values. Some ids have all the fields duplicated with the same values,the only exception being the Prosper Score. I removed the duplicated records and kept only the first row for each id;

- the missing values are occuring especially before July 2009;

- 'NC' score category appears only before July 2009. We know that the scoring system was changed in July 2009.The Prosper Score will be an important feature in my analysis. So, I'm going to filter out records before July 2009;

In the end we are left with 83,982 records in the dataset.

```
## [1] 83982    83
```

```
##                    ListingKey    ListingNumber
##  00003546482094282EF90E5:    1   Min.    : 416275
##  00013542762124763F20254:    1   1st Qu.: 557061
##  00013592180058085D3AC05:    1   Median : 734179
##  00033533702971 4661FD4AD:    1   Mean    : 771240
##  00053536716875 50573289D:    1   3rd Qu.: 975678
##  00053596416593761653B0E:    1   Max.    :1255725
##  (Other)                :83976
##                     ListingCreationDate  CreditGrade         Term
##  2012-10-20 12:21:46.333000000:    2           :83982   Min.   :12.00
##  2013-06-03 17:27:50.540000000:    2    A      :    0   1st Qu.:36.00
##  2009-07-13 18:01:24.347000000:    1    AA     :    0   Median :36.00
##  2009-07-13 18:04:40.220000000:    1    B      :    0   Mean   :42.46
##  2009-07-13 19:08:45.763000000:    1    C      :    0   3rd Qu.:60.00
##  2009-07-13 19:09:37.827000000:    1    D      :    0   Max.   :60.00
##  (Other)                       :83974    (Other):    0
##                  LoanStatus                  ClosedDate
##  Current              :55730                      :57990
##  Completed            :19651   2014-03-04 00:00:00:  105
##  Chargedoff           : 5336   2014-02-19 00:00:00:  100
##  Defaulted            : 1005   2014-02-11 00:00:00:   92
##  Past Due (1-15 days) :  800   2012-10-30 00:00:00:   81
##  Past Due (31-60 days):  361   2013-02-26 00:00:00:   78
```

```
##  (Other)              : 1099   (Other)            :25536
##   BorrowerAPR       BorrowerRate     LenderYield
##  Min.   :0.04583   Min.   :0.0400   Min.   :0.0300
##  1st Qu.:0.16361   1st Qu.:0.1359   1st Qu.:0.1259
##  Median :0.21945   Median :0.1875   Median :0.1775
##  Mean   :0.22694   Mean   :0.1963   Mean   :0.1863
##  3rd Qu.:0.29254   3rd Qu.:0.2574   3rd Qu.:0.2474
##  Max.   :0.42395   Max.   :0.3600   Max.   :0.3400
##
##  EstimatedEffectiveYield EstimatedLoss     EstimatedReturn
##  Min.   :-0.1827         Min.   :0.00490   Min.   :-0.18270
##  1st Qu.: 0.1160         1st Qu.:0.04240   1st Qu.: 0.07463
##  Median : 0.1616         Median :0.07240   Median : 0.09211
##  Mean   : 0.1689         Mean   :0.08042   Mean   : 0.09625
##  3rd Qu.: 0.2253         3rd Qu.:0.11200   3rd Qu.: 0.11710
##  Max.   : 0.3199         Max.   :0.36600   Max.   : 0.28370
##
##  ProsperRating..numeric. ProsperRating..Alpha.  ProsperScore
##  Min.   :1.000           C      :18096         Min.   : 1.000
##  1st Qu.:3.000           B      :15368         1st Qu.: 4.000
##  Median :4.000           A      :14390         Median : 6.000
##  Mean   :4.069           D      :14170         Mean   : 5.953
##  3rd Qu.:5.000           E      : 9716         3rd Qu.: 8.000
##  Max.   :7.000           HR     : 6917         Max.   :11.000
##                          (Other): 5325
##  ListingCategory..numeric. BorrowerState                 Occupation
##  Min.   : 0.000            CA     :10638   Other              :21122
##  1st Qu.: 1.000            NY     : 5775   Professional       :10445
##  Median : 1.000            TX     : 5578   Executive          : 3437
##  Mean   : 3.322            FL     : 5353   Computer Programmer: 3200
##  3rd Qu.: 3.000            IL     : 4215   Teacher            : 2858
##  Max.   :20.000            OH     : 3340   Analyst            : 2711
##                            (Other):49083   (Other)            :40209
##       EmploymentStatus EmploymentStatusDuration IsBorrowerHomeowner
##  Employed     :66586   Min.   :  0.0            False:39560
##  Full-time    : 7926   1st Qu.: 30.0            True :44422
##  Self-employed: 4456   Median : 74.0
##  Other        : 3742   Mean   :103.1
##  Not employed :  649   3rd Qu.:148.0
##  Retired      :  367   Max.   :755.0
##  (Other)      :  256   NA's   :19
##  CurrentlyInGroup                    GroupKey
##  False:81743                            :81958
##  True : 2239      3D4D3366260257624AB272D:  310
##                   783C3371218786870A73D20:  208
##                   52EA3425051368132B80C96:  150
##                   B047336437692012837OB13:   83
##                   FEF83377364176536637E50:   82
##                   (Other)                : 1191
##             DateCreditPulled CreditScoreRangeLower CreditScoreRangeUpper
##  2010-04-27 09:47:43:    2   Min.   :600.0         Min.   :619.0
```

```
##  2010-07-18 14:45:02:    2   1st Qu.:660.0         1st Qu.:679.0
##  2010-07-29 14:29:08:    2   Median :700.0         Median :719.0
##  2011-06-27 03:36:39:    2   Mean   :699.5         Mean   :718.5
##  2011-09-20 13:24:11:    2   3rd Qu.:720.0         3rd Qu.:739.0
##  2011-12-05 07:17:32:    2   Max.   :880.0         Max.   :899.0
##  (Other)            :83970
##         FirstRecordedCreditLine CurrentCreditLines OpenCreditLines
##  1993-12-01 00:00:00:  148      Min.   : 0.0       Min.   : 0.000
##  1994-11-01 00:00:00:  136      1st Qu.: 7.0       1st Qu.: 6.000
##  1995-11-01 00:00:00:  124      Median :10.0       Median : 9.000
##  1996-03-01 00:00:00:  122      Mean   :10.5       Mean   : 9.517
##  1995-03-01 00:00:00:  120      3rd Qu.:13.0       3rd Qu.:12.000
##  1989-11-01 00:00:00:  117      Max.   :59.0       Max.   :54.000
##  (Other)            :83215
##  TotalCreditLinespast7years OpenRevolvingAccounts
##  Min.   :  2.00             Min.   : 0.000
##  1st Qu.: 18.00             1st Qu.: 4.000
##  Median : 26.00             Median : 7.000
##  Mean   : 27.66             Mean   : 7.377
##  3rd Qu.: 35.00             3rd Qu.:10.000
##  Max.   :125.00             Max.   :50.000
##
##  OpenRevolvingMonthlyPayment InquiriesLast6Months TotalInquiries
##  Min.   :    0.0             Min.   : 0.0000      Min.   : 0.000
##  1st Qu.:  155.0             1st Qu.: 0.0000      1st Qu.: 2.000
##  Median :  310.0             Median : 0.0000      Median : 3.000
##  Mean   :  430.4             Mean   : 0.9644      Mean   : 4.286
##  3rd Qu.:  563.0             3rd Qu.: 1.0000      3rd Qu.: 6.000
##  Max.   :13765.0             Max.   :27.0000      Max.   :78.000
##
##  CurrentDelinquencies AmountDelinquent   DelinquenciesLast7Years
##  Min.   : 0.0000      Min.   :     0.0   Min.   : 0.00
##  1st Qu.: 0.0000      1st Qu.:     0.0   1st Qu.: 0.00
##  Median : 0.0000      Median :     0.0   Median : 0.00
##  Mean   : 0.3238      Mean   :   953.3   Mean   : 3.66
##  3rd Qu.: 0.0000      3rd Qu.:     0.0   3rd Qu.: 2.00
##  Max.   :51.0000      Max.   :463881.0   Max.   :99.00
##
##  PublicRecordsLast10Years PublicRecordsLast12Months RevolvingCreditBalance
##  Min.   : 0.000           Min.   : 0.000000         Min.   :     0
##  1st Qu.: 0.000           1st Qu.: 0.000000         1st Qu.:  3804
##  Median : 0.000           Median : 0.000000         Median :  9310
##  Mean   : 0.285           Mean   : 0.009252         Mean   : 17936
##  3rd Qu.: 0.000           3rd Qu.: 0.000000         3rd Qu.: 20335
##  Max.   :38.000           Max.   :20.000000         Max.   :999165
##
##  BankcardUtilization AvailableBankcardCredit  TotalTrades
##  Min.   :0.000       Min.   :     0          Min.   :  1.00
##  1st Qu.:0.330       1st Qu.:  1141          1st Qu.: 15.00
##  Median :0.600       Median :  4568          Median : 23.00
##  Mean   :0.564       Mean   : 11400          Mean   : 23.93
##  3rd Qu.:0.830       3rd Qu.: 13002          3rd Qu.: 31.00
```

```
##  3rd Qu.:0.830       3rd Qu.: 13902          3rd Qu.: 31.00
##  Max.   :2.500       Max.   :498374          Max.   :122.00
##
##  TradesNeverDelinquent..percentage. TradesOpenedLast6Months
##  Min.   :0.0800                     Min.   : 0.0000
##  1st Qu.:0.8500                     1st Qu.: 0.0000
##  Median :0.9500                     Median : 0.0000
##  Mean   :0.9057                     Mean   : 0.7288
##  3rd Qu.:1.0000                     3rd Qu.: 1.0000
##  Max.   :1.0000                     Max.   :20.0000
##
##  DebtToIncomeRatio         IncomeRange    IncomeVerifiable
##  Min.   : 0.000    $50,000-74,999:25326   False: 7251
##  1st Qu.: 0.150    $25,000-49,999:23923   True :76731
##  Median : 0.220    $100,000+     :15056
##  Mean   : 0.259    $75,000-99,999:14362
##  3rd Qu.: 0.320    $1-24,999     : 4621
##  Max.   :10.010    Not employed  :  649
##  NA's   :7214      (Other)       :   45
##  StatedMonthlyIncome                   LoanKey      TotalProsperLoans
##  Min.   :      0     00003683605746079487FF7:    1   Min.   :0.00
##  1st Qu.:   3427     0002365050369681OC531F7:    1   1st Qu.:1.00
##  Median :   5000     0003370344524056436F5AC:    1   Median :1.00
##  Mean   :   5931     0004363753221955965B646:    1   Mean   :1.46
##  3rd Qu.:   7083     0005370013632204511EA011:    1   3rd Qu.:2.00
##  Max.   :1750003     0006369562505202 9E10FDB:    1   Max.   :8.00
##                      (Other)               :83976   NA's   :64347
##  TotalProsperPaymentsBilled OnTimeProsperPayments
##  Min.   :  0.00             Min.   :  0.0
##  1st Qu.: 10.00             1st Qu.:  9.0
##  Median : 18.00             Median : 17.0
##  Mean   : 24.31             Mean   : 23.6
##  3rd Qu.: 35.00             3rd Qu.: 34.0
##  Max.   :141.00             Max.   :141.0
##  NA's   :64347              NA's   :64347
##  ProsperPaymentsLessThanOneMonthLate ProsperPaymentsOneMonthPlusLate
##  Min.   : 0.00                       Min.   : 0.00
##  1st Qu.: 0.00                       1st Qu.: 0.00
##  Median : 0.00                       Median : 0.00
##  Mean   : 0.66                       Mean   : 0.05
##  3rd Qu.: 0.00                       3rd Qu.: 0.00
##  Max.   :42.00                       Max.   :21.00
##  NA's   :64347                       NA's   :64347
##  ProsperPrincipalBorrowed ProsperPrincipalOutstanding
##  Min.   :    0            Min.   :    0
##  1st Qu.: 3700            1st Qu.:    0
##  Median : 6300            Median : 1591
##  Mean   : 8737            Mean   : 2918
##  3rd Qu.:11600            3rd Qu.: 4112
##  Max.   :72499            Max.   :23451
##  NA's   :64347            NA's   :64347
##  ScorexChangeAtTimeOfListing LoanCurrentDaysDelinquent
```

```
##  ScoreExchangeAtTimeOfListing LoanCurrentDaysDelinquent
##  Min.   :-209.00              Min.   :   0.00
##  1st Qu.: -35.00              1st Qu.:   0.00
##  Median :  -6.00              Median :   0.00
##  Mean   :  -4.69              Mean   :  37.01
##  3rd Qu.:  23.00              3rd Qu.:   0.00
##  Max.   : 286.00              Max.   :1593.00
##  NA's   :67356
##  LoanFirstDefaultedCycleNumber LoanMonthsSinceOrigination   LoanNumber
##  Min.   : 1.00                 Min.   : 0.00              Min.   : 38045
##  1st Qu.: 9.00                 1st Qu.: 4.00              1st Qu.: 60728
##  Median :13.00                 Median :11.00              Median : 87354
##  Mean   :14.47                 Mean   :16.16              Mean   : 86231
##  3rd Qu.:19.00                 3rd Qu.:25.00              3rd Qu.:108609
##  Max.   :41.00                 Max.   :56.00              Max.   :136486
##  NA's   :77738
##  LoanOriginalAmount          LoanOriginationDate LoanOriginationQuarter
##  Min.   : 1000      2013-11-13 00:00:00:  469    Q4 2013:14054
##  1st Qu.: 4000      2014-01-22 00:00:00:  468    Q1 2014:11734
##  Median : 7500      2013-10-16 00:00:00:  424    Q3 2013: 9143
##  Mean   : 9061      2014-02-19 00:00:00:  424    Q2 2013: 7099
##  3rd Qu.:13500      2014-01-28 00:00:00:  329    Q3 2012: 5632
##  Max.   :35000      2013-09-24 00:00:00:  309    Q2 2012: 5061
##                     (Other)            :81559    (Other):31259
##                    MemberKey     MonthlyLoanPayment LP_CustomerPayments
##  720D3508651090808DC328F:    7   Min.   :   0.0     Min.   :   -2.35
##  C7093420657523078260C7:    7   1st Qu.: 157.1     1st Qu.:  818.35
##  E4AF3422677498955FFA00E:    7   Median : 251.3     Median : 2247.31
##  18F6337949289842881D0A8:    6   Mean   : 291.4     Mean   : 3698.95
##  3D6B34225353312993B9700:    6   3rd Qu.: 387.6     3rd Qu.: 4910.83
##  43DB3366978035224D7D9E3:    6   Max.   :2251.5     Max.   :37369.16
##  (Other)                :83943
##  LP_CustomerPrincipalPayments LP_InterestandFees LP_ServiceFees
##  Min.   :    0.0              Min.   :   -2.35   Min.   :-589.95
##  1st Qu.:  405.2              1st Qu.:  259.80   1st Qu.: -72.91
##  Median : 1274.4              Median :  683.13   Median : -35.39
##  Mean   : 2648.8              Mean   : 1050.19   Mean   : -55.04
##  3rd Qu.: 3471.2              3rd Qu.: 1447.97   3rd Qu.: -14.62
##  Max.   :35000.0              Max.   :10572.78   Max.   :   3.01
##
##  LP_CollectionFees   LP_GrossPrincipalLoss LP_NetPrincipalLoss
##  Min.   :-4865.080   Min.   :  -94.2       Min.   : -504.4
##  1st Qu.:    0.000   1st Qu.:    0.0       1st Qu.:    0.0
##  Median :    0.000   Median :    0.0       Median :    0.0
##  Mean   :   -8.289   Mean   :  380.0       Mean   :  371.5
##  3rd Qu.:    0.000   3rd Qu.:    0.0       3rd Qu.:    0.0
##  Max.   :    0.000   Max.   :25000.0       Max.   :25000.0
##
##  LP_NonPrincipalRecoverypayments PercentFunded    Recommendations
##  Min.   :   0.000                Min.   :0.7000   Min.   : 0.00000
##  1st Qu.:   0.000                1st Qu.:1.0000   1st Qu.: 0.00000
##  Median :   0.000                Median :1.0000   Median : 0.00000
```

```
##  Median :   0.000                  Median :1.0000   Median : 0.00000
##  Mean   :   7.726                  Mean   :0.9981   Mean   : 0.01772
##  3rd Qu.:   0.000                  3rd Qu.:1.0000   3rd Qu.: 0.00000
##  Max.   :7780.030                  Max.   :1.0125   Max.   :19.00000
##
##  InvestmentFromFriendsCount InvestmentFromFriendsAmount   Investors
##  Min.   :0.000000           Min.   :    0.00            Min.   :   1.00
##  1st Qu.:0.000000           1st Qu.:    0.00            1st Qu.:   1.00
##  Median :0.000000           Median :    0.00            Median :  32.00
##  Mean   :0.008133           Mean   :    4.36            Mean   :  68.68
##  3rd Qu.:0.000000           3rd Qu.:    0.00            3rd Qu.:  98.00
##  Max.   :9.000000           Max.   :11000.00            Max.   :1189.00
##
##  ListingCreationDate2    Score
##  Min.   :2009-07-13   Length:83982
##  1st Qu.:2012-02-07   Class :character
##  Median :2013-03-20   Mode  :character
##  Mean   :2012-10-30
##  3rd Qu.:2013-10-23
##  Max.   :2014-03-10
##
```

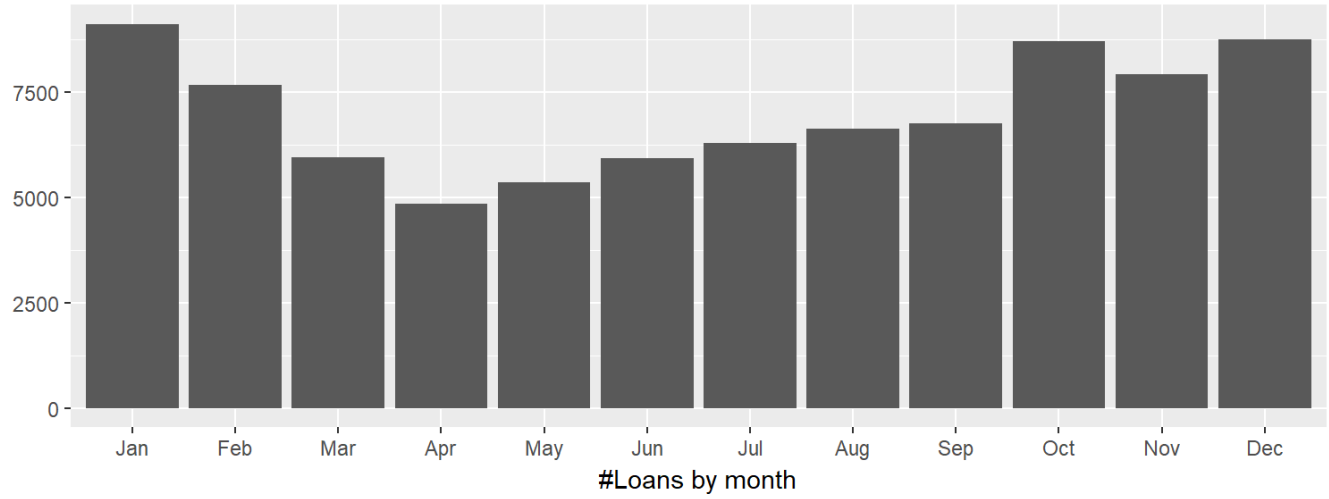
The cleaned dataset shows better summaries: the missings are gone, minimum BorrowerRate is no longer 0, but StatedMonthlyIncome has value zero.

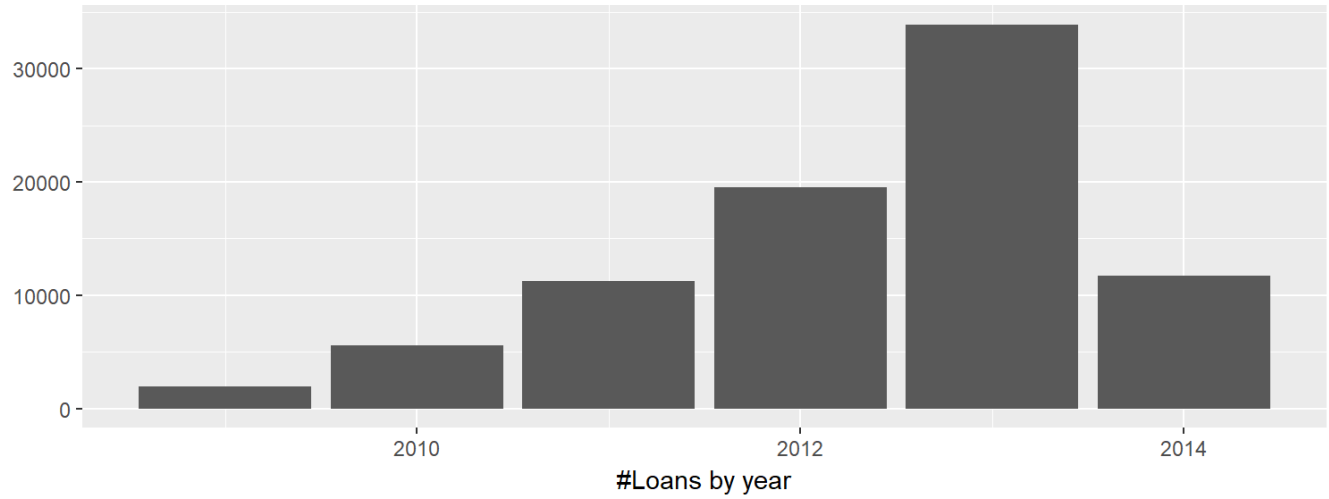
UNIVARIATE PLOTS SECTION

1. NUMBER OF LOANS

First, let's see the evolution in time of the number of loans. For this I need to create new variables for month, year and month&year.I will extract this information from LoanOriginationDate variable.

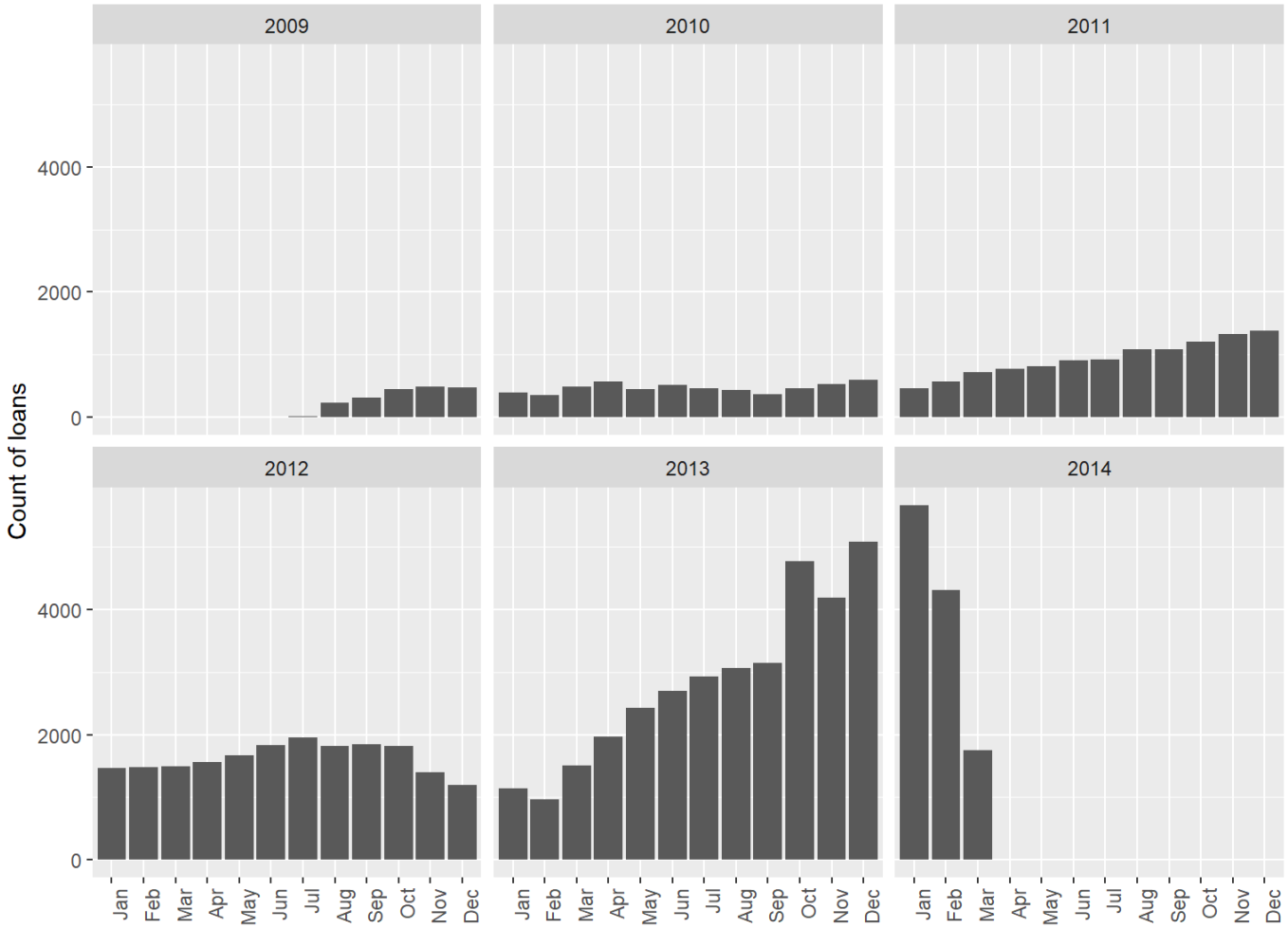
1.1 Number of loans by month and year.





Number of loans are increasing by year. Years 2009 and 2014 don't have data for all 12 months: we have data from July 2009 until March 2014.

We see more loans in the first and last months of the year. Is there a sesonality for loans? Are people borrowing more in January and December? Let's plot the same thing, but facet wraped by year.



From this plot we see that loans are increasing over time, with high peaks in Oct-Dec 2013 and Jan 2014. There is actually no seasonality in data. The peaks in 2013 and 2014 are affecting the total loans distribution by month.

2. VALUES OF LOANS

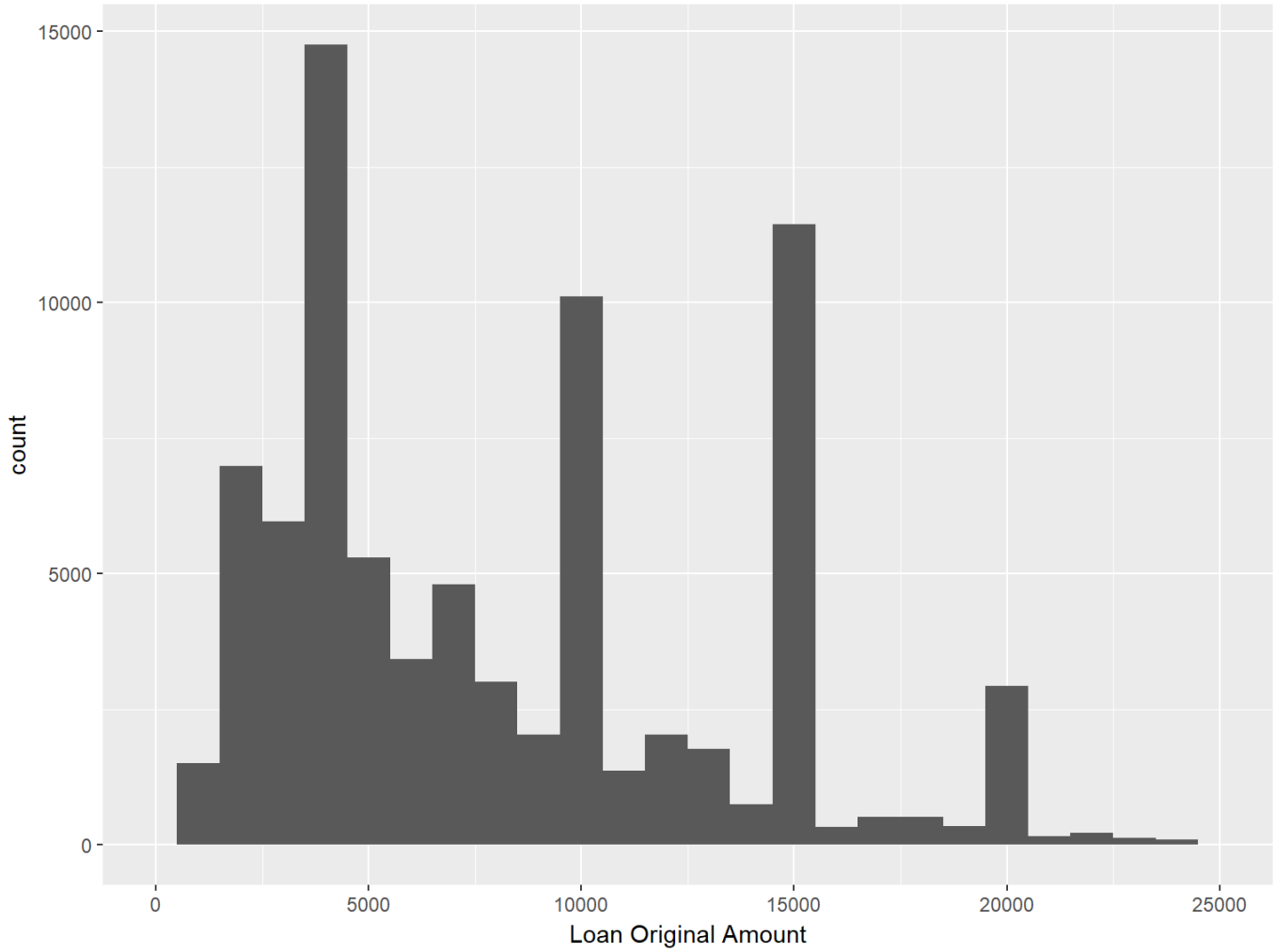
Total amount of loans by year:

```
## # A tibble: 6 × 3
##   yearCreditDate value_loans no_loans
##            <dbl>       <int>    <int>
## 1           2009     8657561     1976
## 2           2010    26620674     5579
## 3           2011    75138013    11228
## 4           2012   153175116    19553
## 5           2013   357437822    33912
## 6           2014   139950560    11734
```

Summary:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1000    4000    7500    9061   13500   35000
```

Histogram:



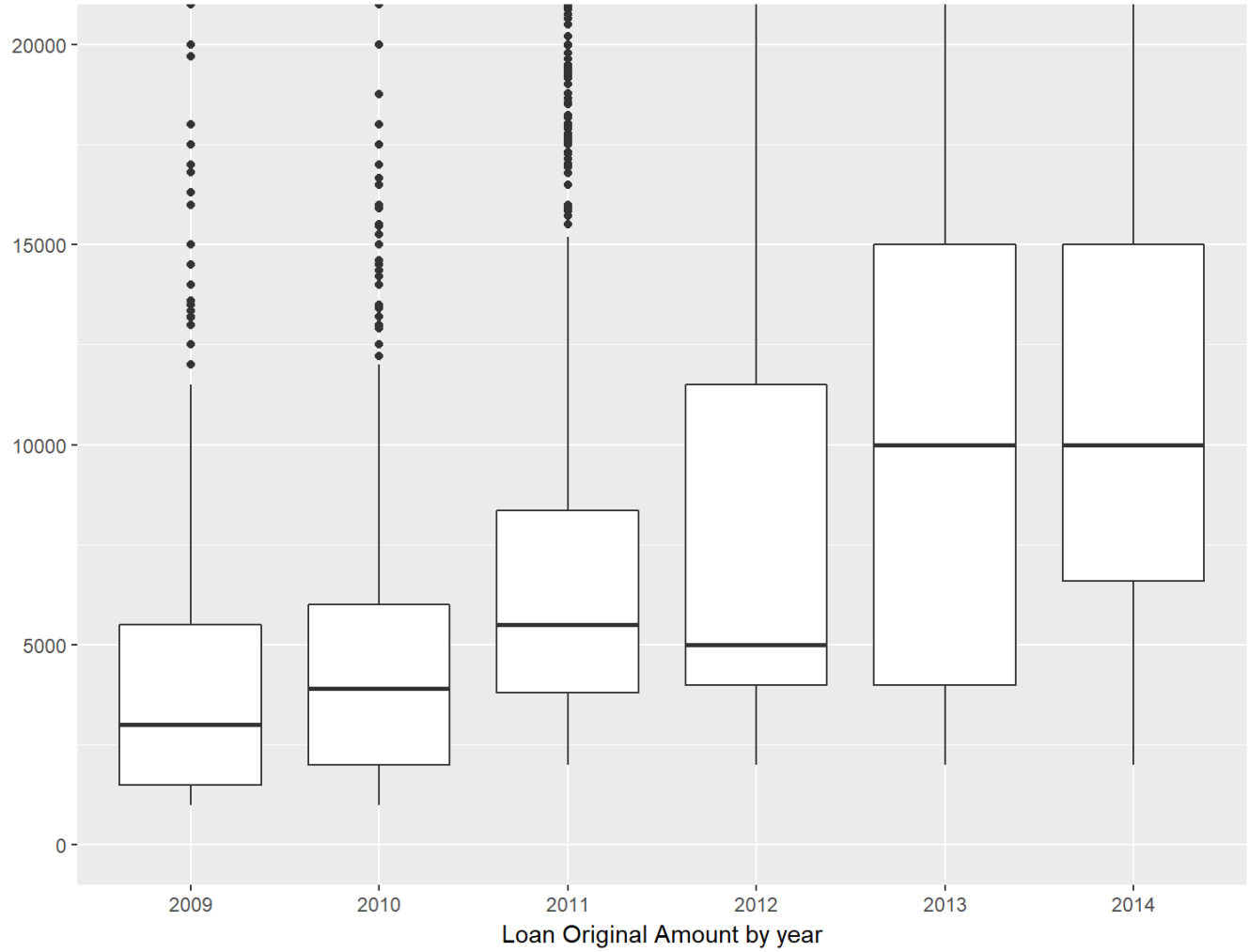
People preffer to borrow round values, like $4000,$10000,$15000,$20000 (in the histogram above we have high peaks for these values).

Summary loans by year:

```
## df_loans$yearCreditDate: 2009
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1000    1500    3000    4381    5500   25000
## ------------------------------------------------------------
## df_loans$yearCreditDate: 2010
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1000    2000    3900    4772    6000   25000
## ------------------------------------------------------------
## df_loans$yearCreditDate: 2011
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2000    3800    5500    6692    8361   25000
## ------------------------------------------------------------
## df_loans$yearCreditDate: 2012
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2000    4000    5000    7834   11500   25000
## ------------------------------------------------------------
## df_loans$yearCreditDate: 2013
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2000    4000   10000   10540   15000   35000
## ------------------------------------------------------------
## df_loans$yearCreditDate: 2014
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2000    6600   10000   11930   15000   35000
```

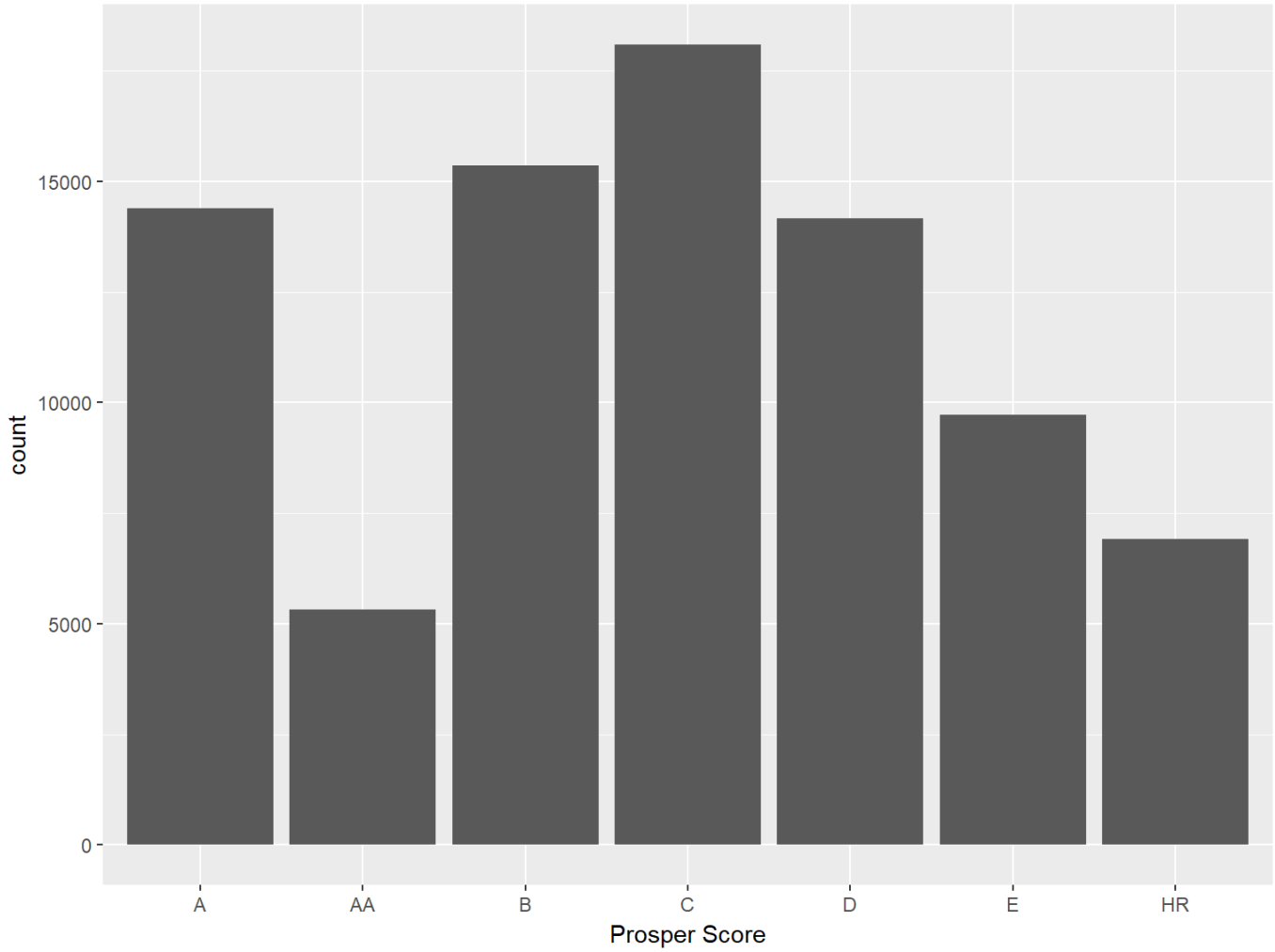
Loan amount is increasing by year. In the last 2 years, 25% of the people borrowed more than $15000, compared to $6000 in 2010.

Boxplot loan amount per year:

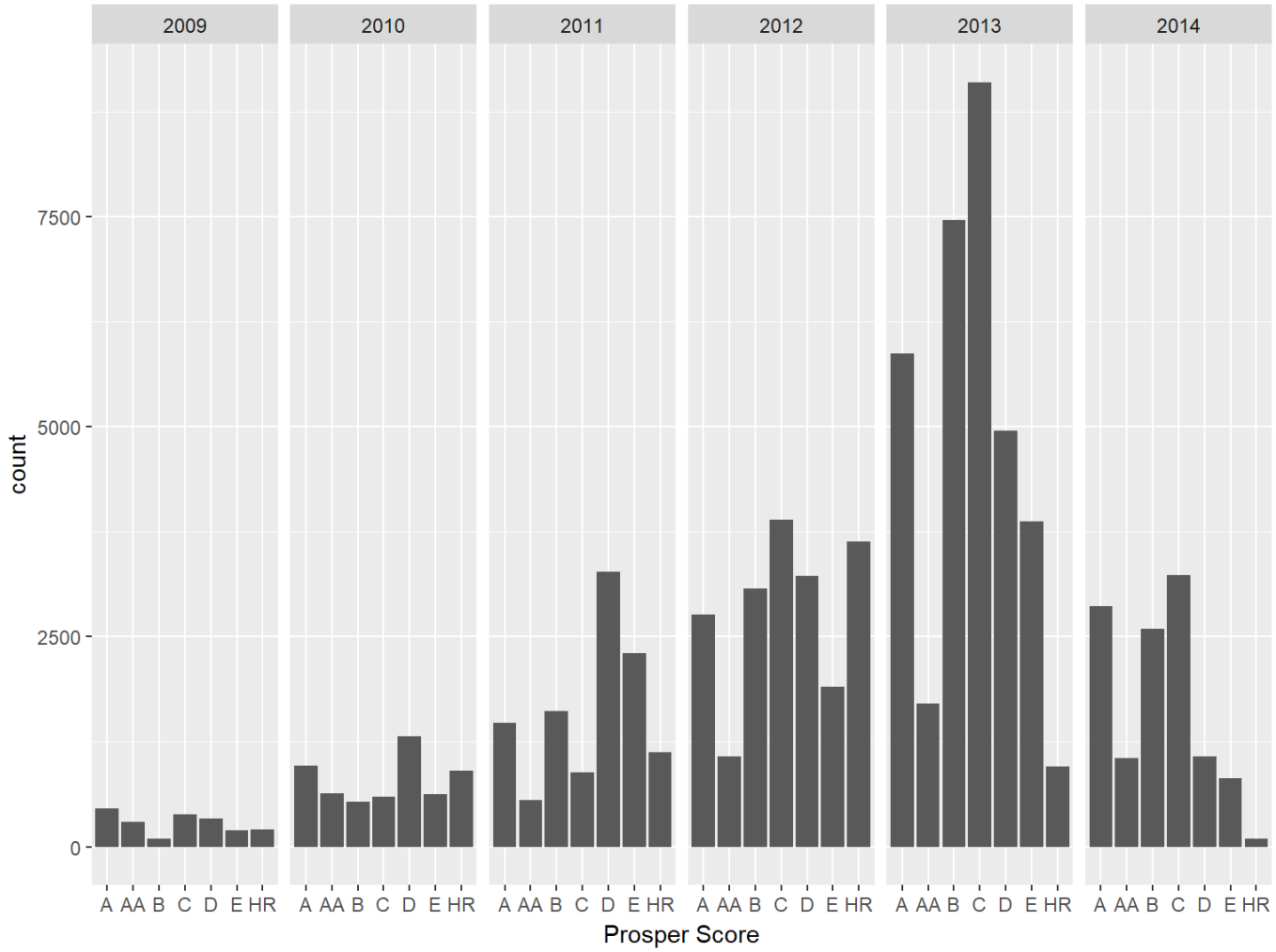


The total values and the median of the loans are increasing by year. In 2012, the median is very low, which means there were many loans of low values; 50% of the people borrowed less than $5000.

3. CREDIT SCORE

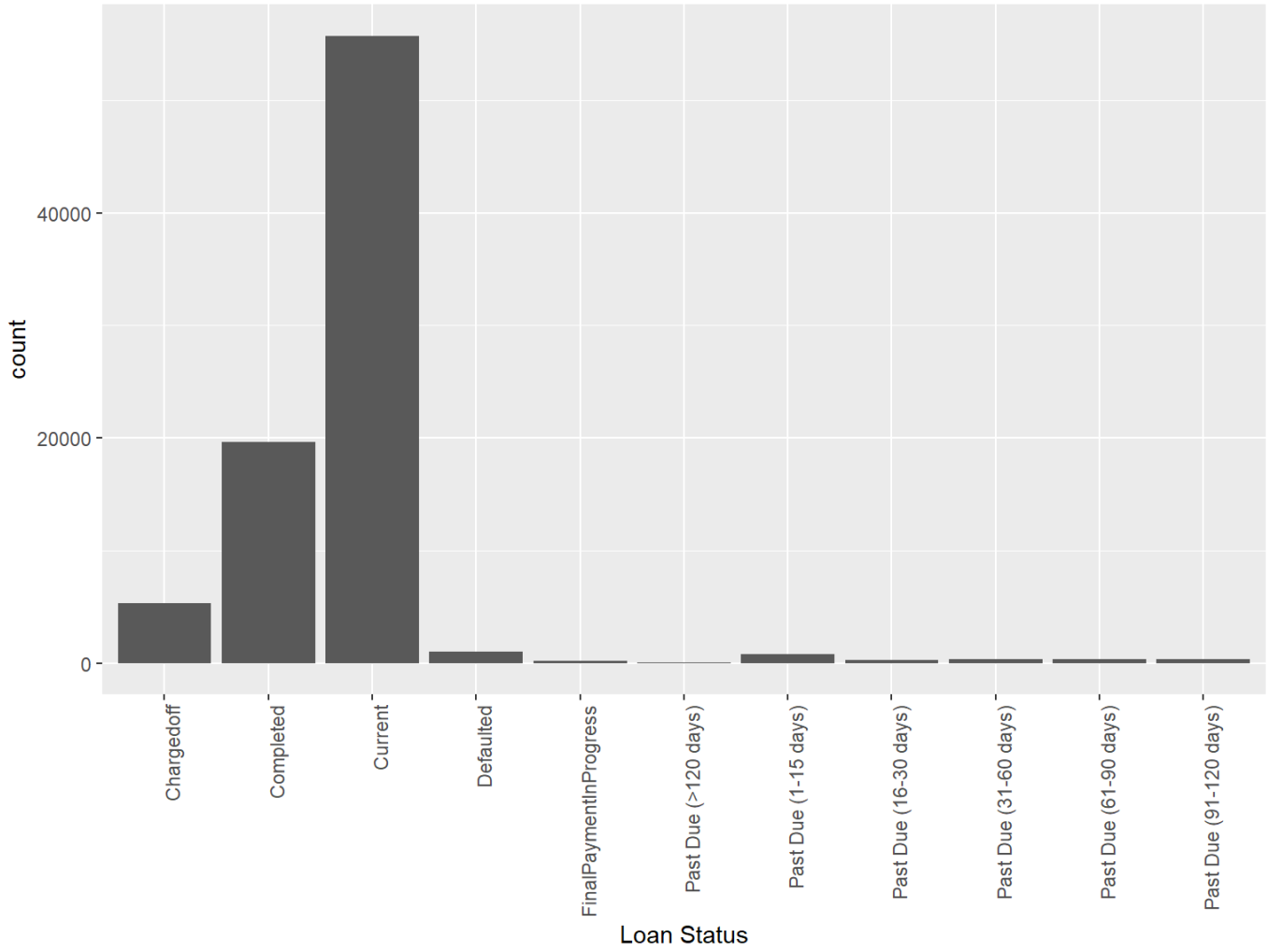


Prosper has many high value borrowers. A,B and C are the most frequent scores. Lets see if the scores changed over time.



We see a descrease in E and D and increase in B and C scores. In 2012, there was a great increase of HR score. This could explain why the meadian of loans was so low in 2012, as we saw in a previous boxplot.

4. LOAN STATUS

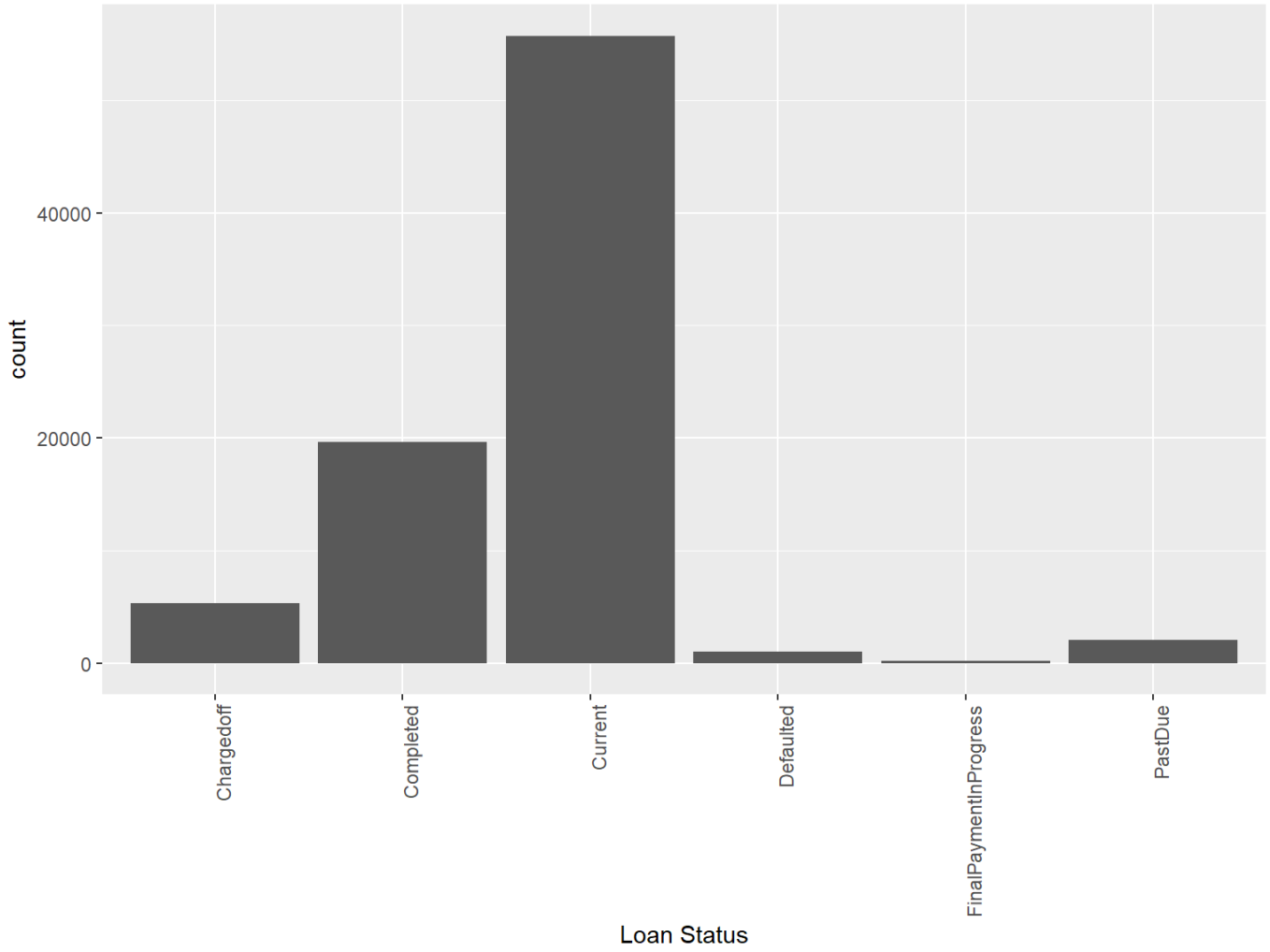


Most of the loans have 'Current' status. I need to create a new variable with the Past due categories grouped.

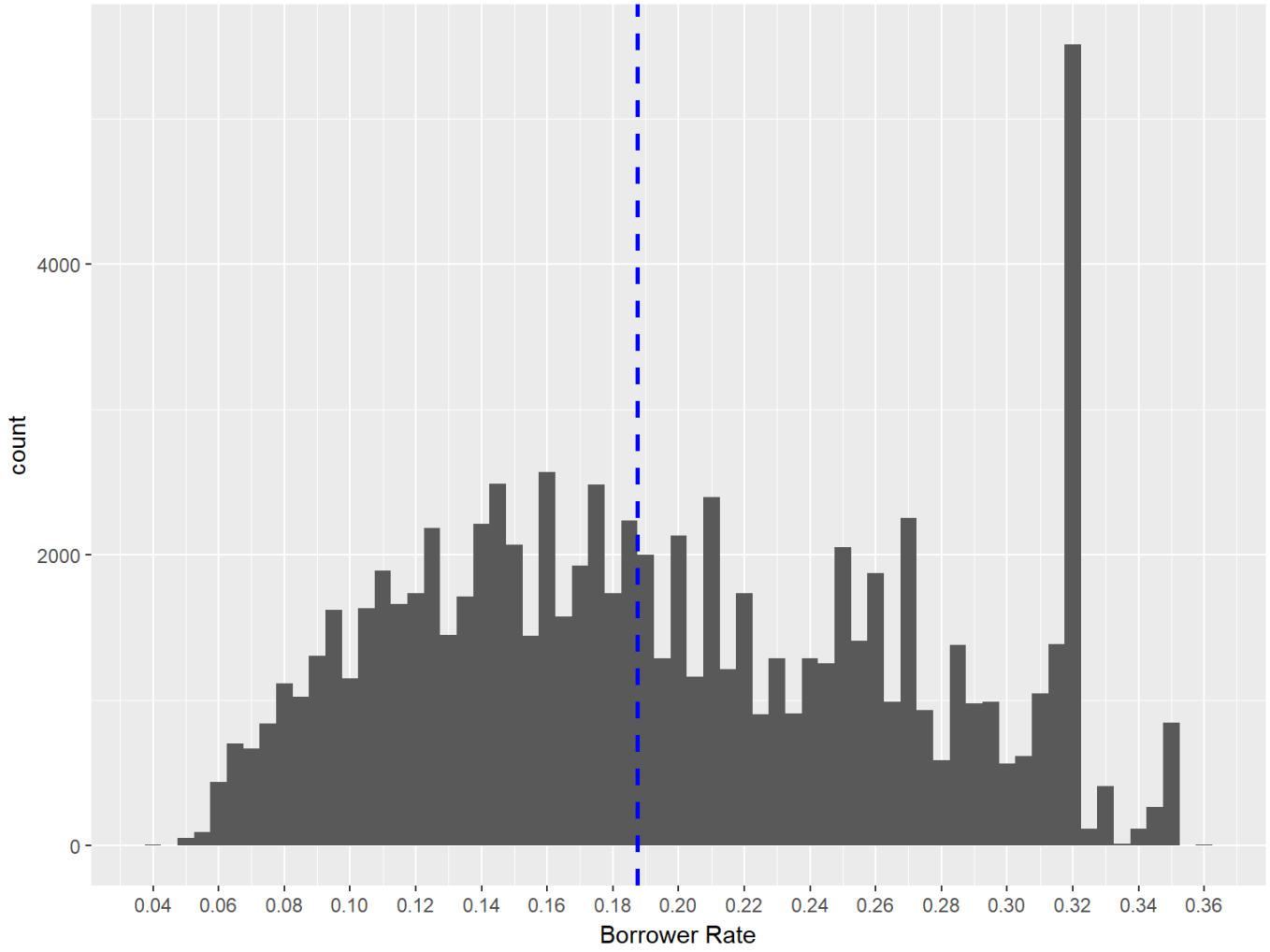
```
## # A tibble: 6 × 3
##           LoanStatus_rec     N        freq
##                    <chr> <int>       <dbl>
## 1             Chargedoff  5336 0.063537425
## 2              Completed 19651 0.233990617
## 3                Current 55730 0.663594580
## 4              Defaulted  1005 0.011966850
## 5 FinalPaymentInProgress   203 0.002417185
## 6                PastDue  2057 0.024493344
```

66% of the loans are Current, 23% are Completed, 0.12% are Defaulted and 0.63% are Chargedoff.

Plot the recoded loan status variable:


count
40000
20000
0
Chargedoff
Completed
Current
Defaulted
FinalPaymentInProgress
PastDue
Loan Status

5. RATES




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0400  0.1359  0.1875  0.1963  0.2574  0.3600
```

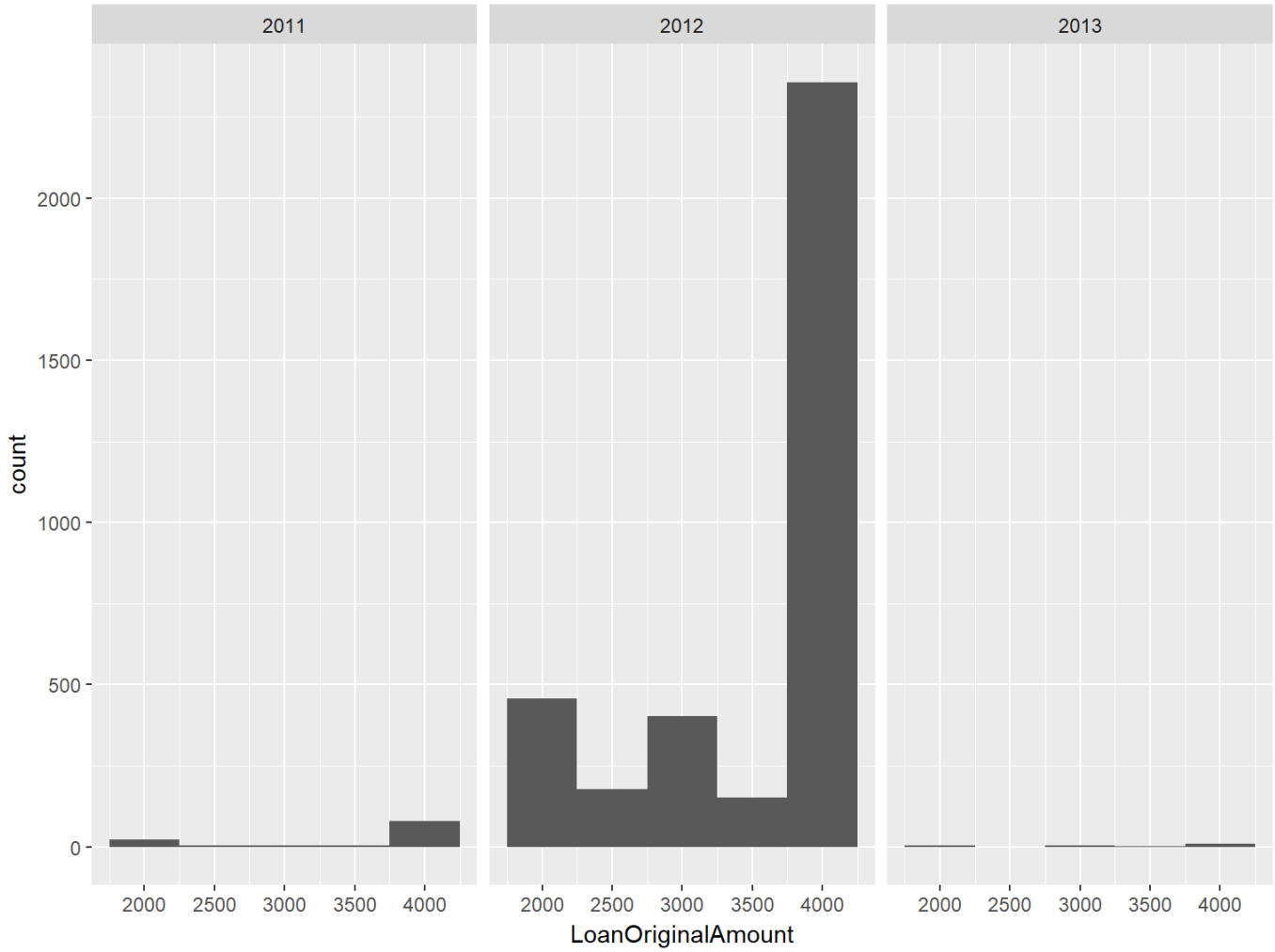
Borrowerrate has a peak around 0.32, value close to the maximun rate.Is this a normal value for the borrower rate? My guess is,this value has something to do with year 2012, when Prosper had many low loans in HR category.

Let’s see if this is true. First, let’s find the exact value where this peak happens.

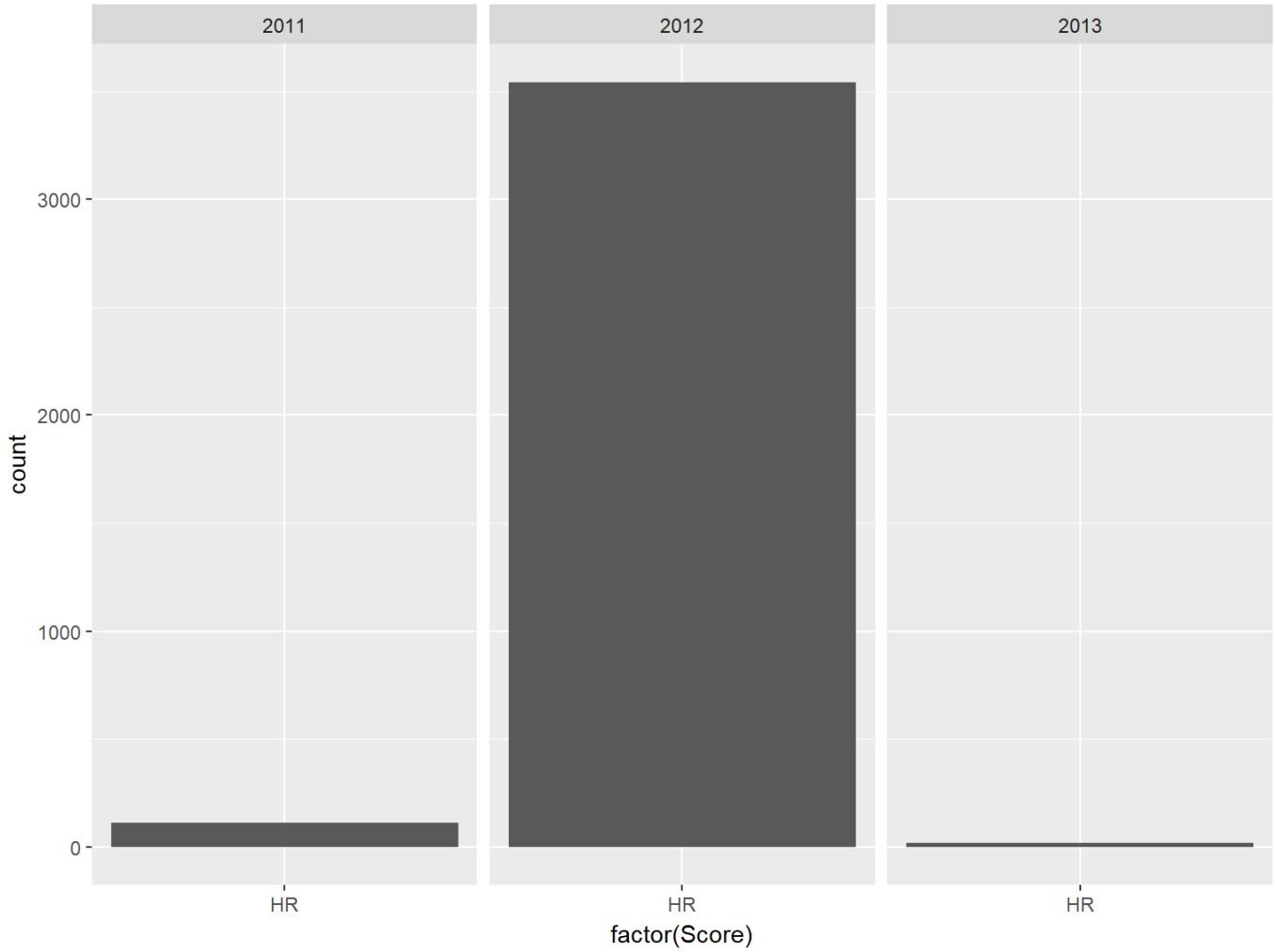
```
##     mode
## 1 0.3177
```

This value is 0.3177.

Next let’s plot the 0.3177 BorrowerRate, facet warped by year.

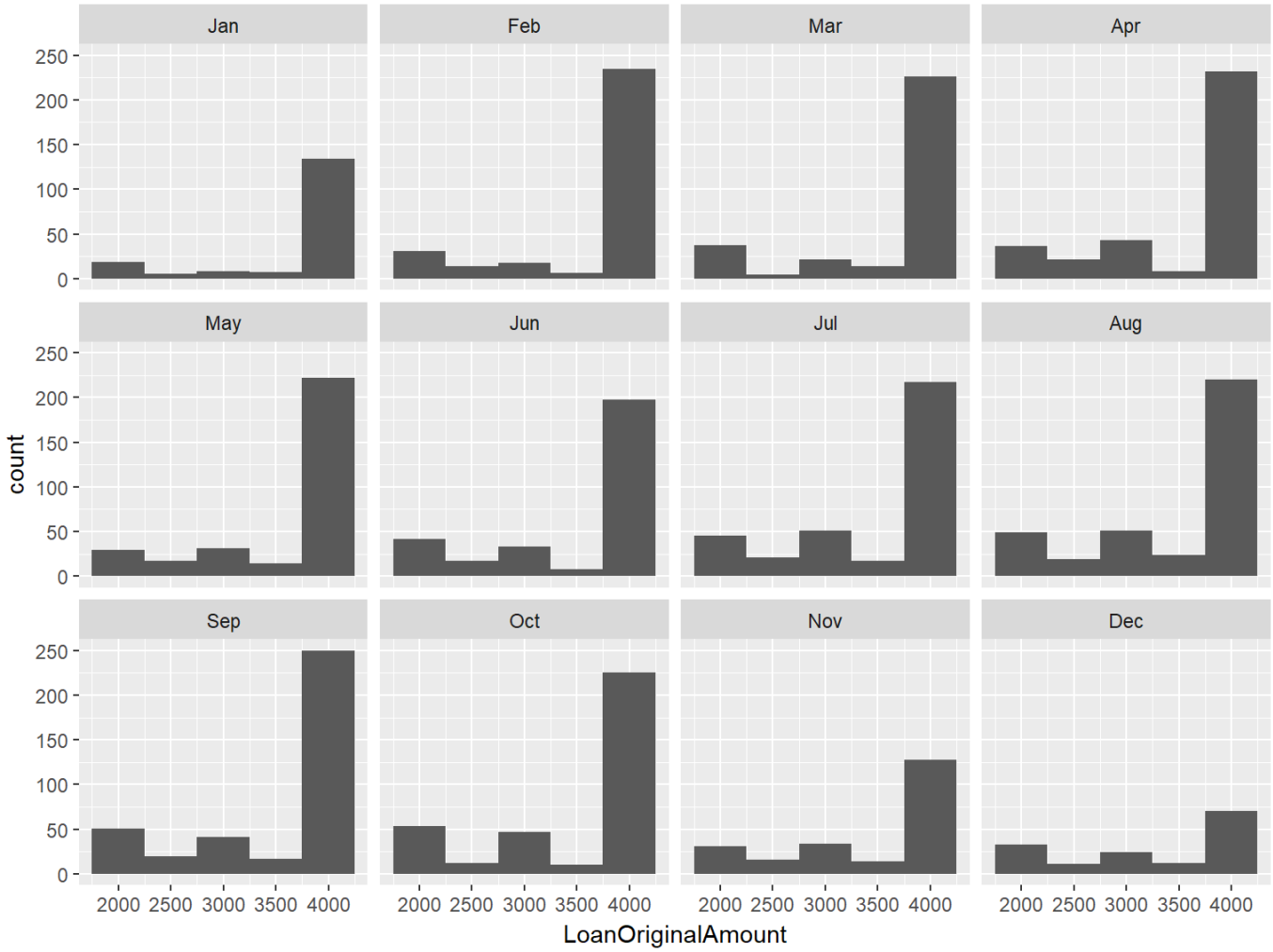


In 2012, there was a high amount of loans of $4000 with a borrower rate of 0.3177. Where they also in HR category ?



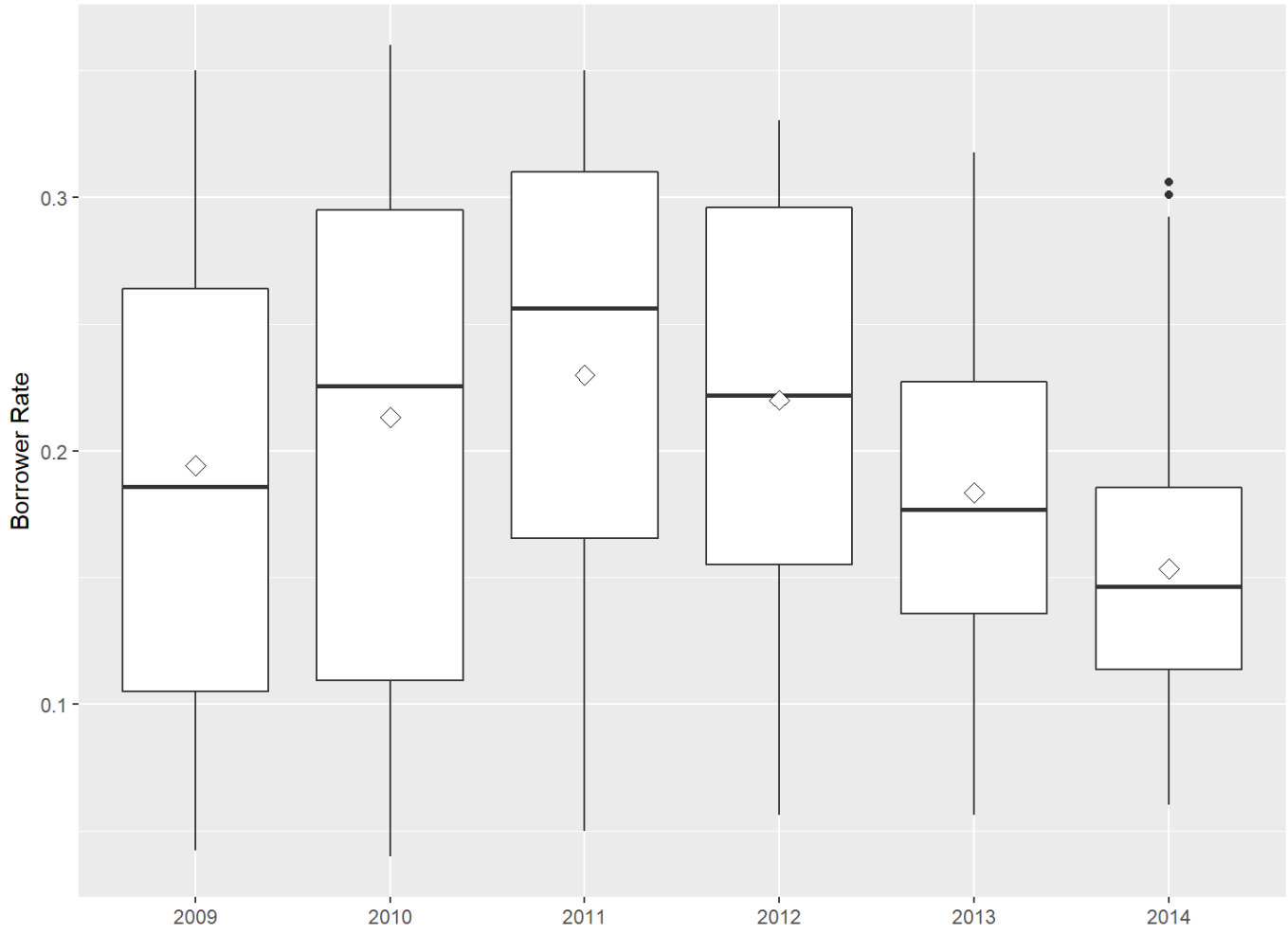
Yes, my intuition was correct. Borrower rate 0.3177 was applied to HR category.

One more curiosity: was this rate applied only in certain months of 2012?



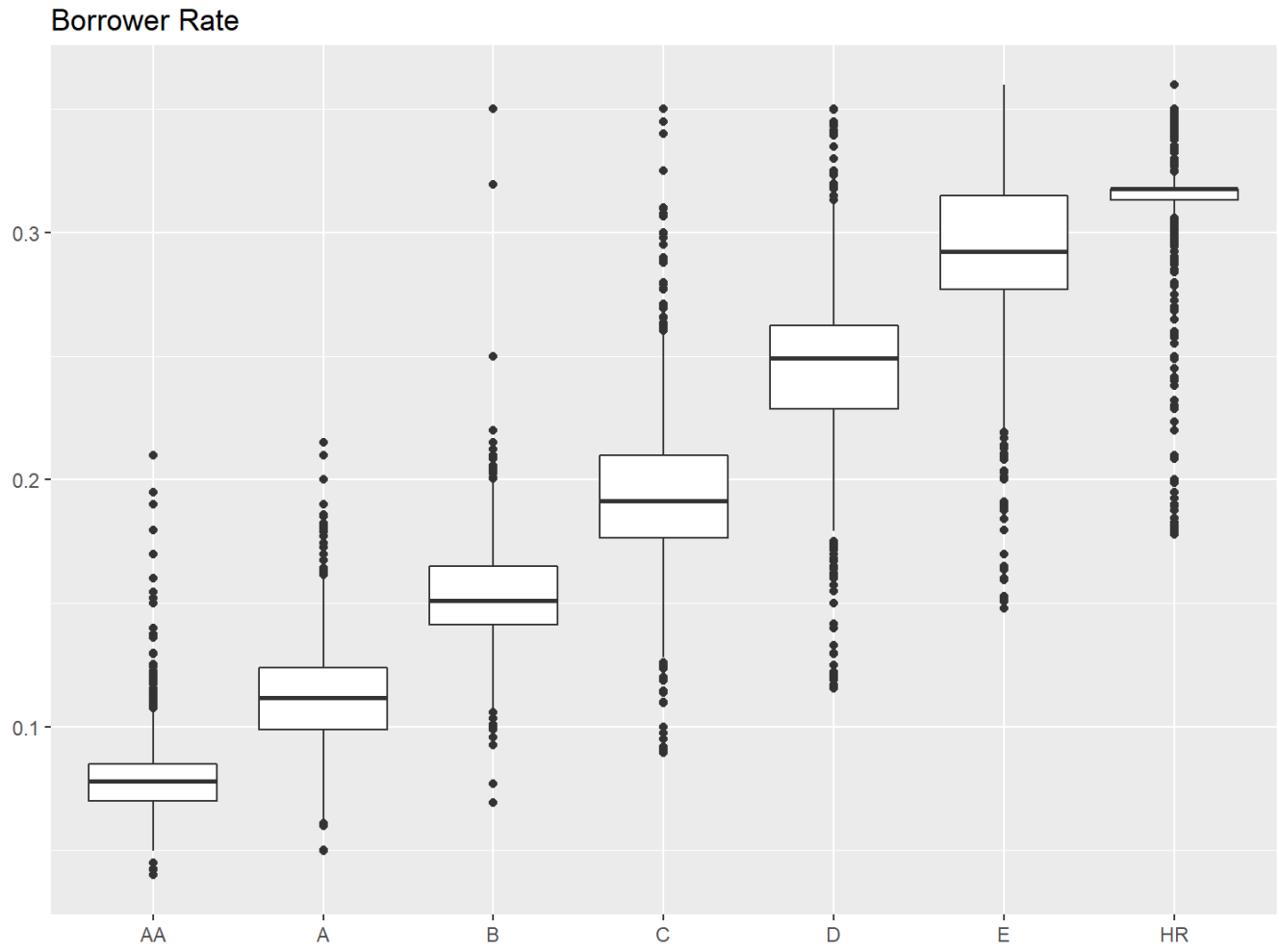
This rate is present in all the months of 2012, with a high decrease at the end of the year.

RATES - BOXPLOTS:



Borrower rates have an ascendent trend from 2009 to 2011 and start decreasing in 2012.

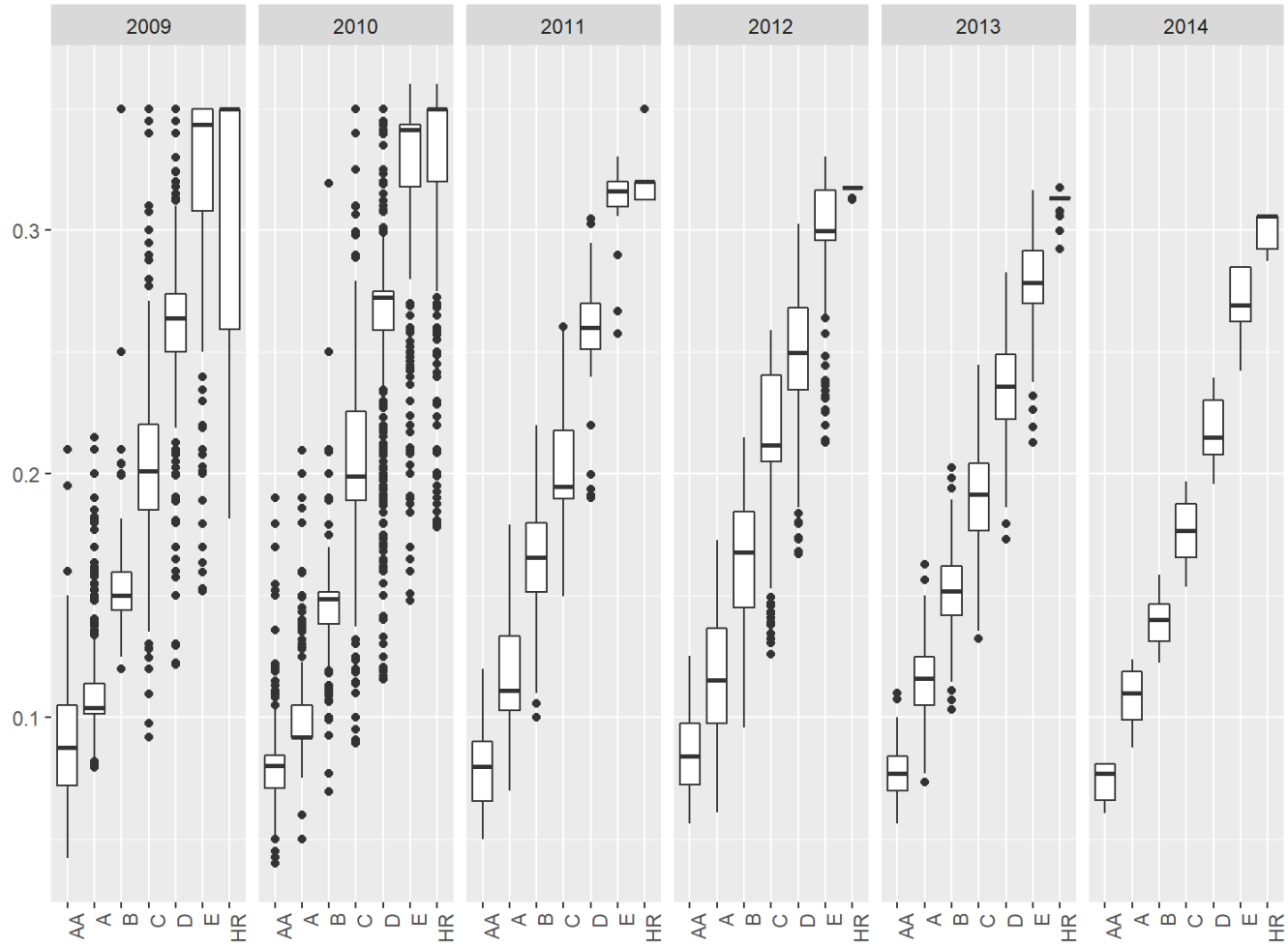
Let’s see if the borrower rate is influenced by Prosper Score?



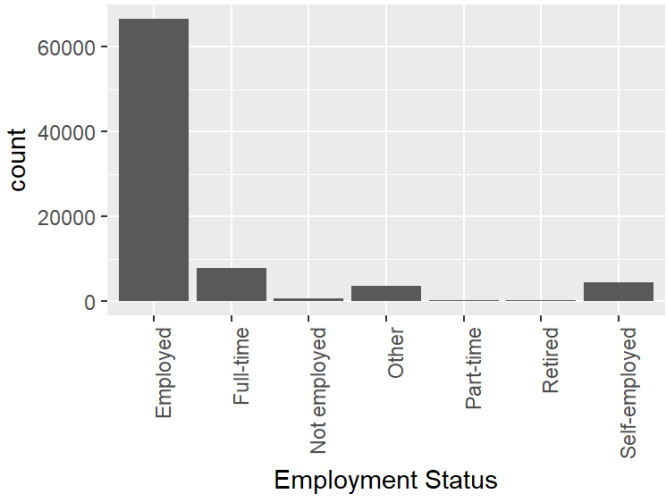
Clearly, the most valuable borrowers (AA category) have the lowest rates, while the risky ones have higher rates.

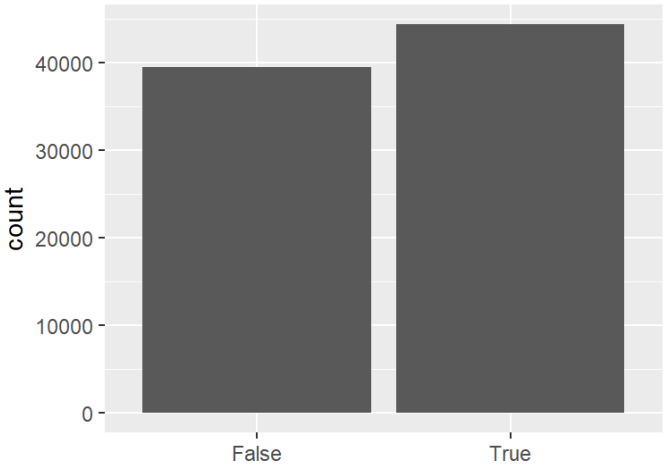
Rate increases by prosper score. There's a high negative correlation of -0.95 between borrower rate and prosper score.

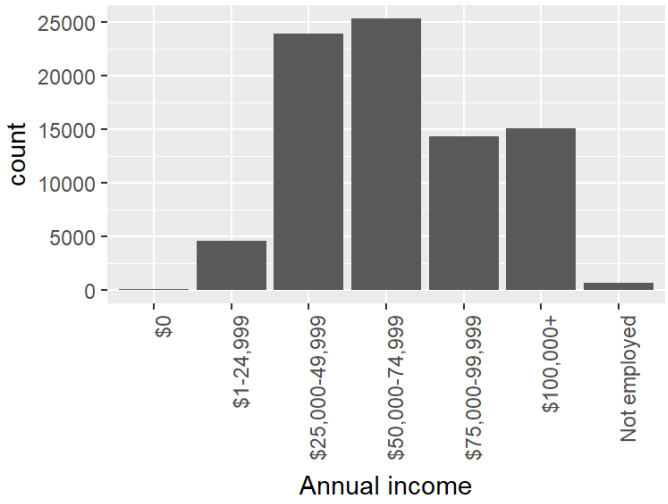
```
##
##  Pearson's product-moment correlation
##
## data:  ProsperRating..numeric. and BorrowerRate
## t = -912.61, df = 83980, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9537173 -0.9524783
## sample estimates:
##       cor
## -0.9531018
```



6. BORROWERS CHARACTERISTICS: EMPLOYMENT STATUS, Is Borrower Homeowner?, IncomeRange







7. INCOME AND DEBT TO INCOME

Summary StatedMonthlyIncome:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    3427    5000    5931    7083 1750000
```

1.750.003 maximum monthly income value doesn't look like a realistic number.

Summary DebtToIncomeRatio:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   0.150   0.220   0.259   0.320  10.010    7214
```

Income=0 is an issue.But why DebtToIncomeRatio has NA's? This should be calculated as ratio of debt to income. We don't have any missings StatedMonthlyIncome, so DebtToIncomeRatio should be fully populated. Is this variable left NA in case the income was not verifiable by Prosper?

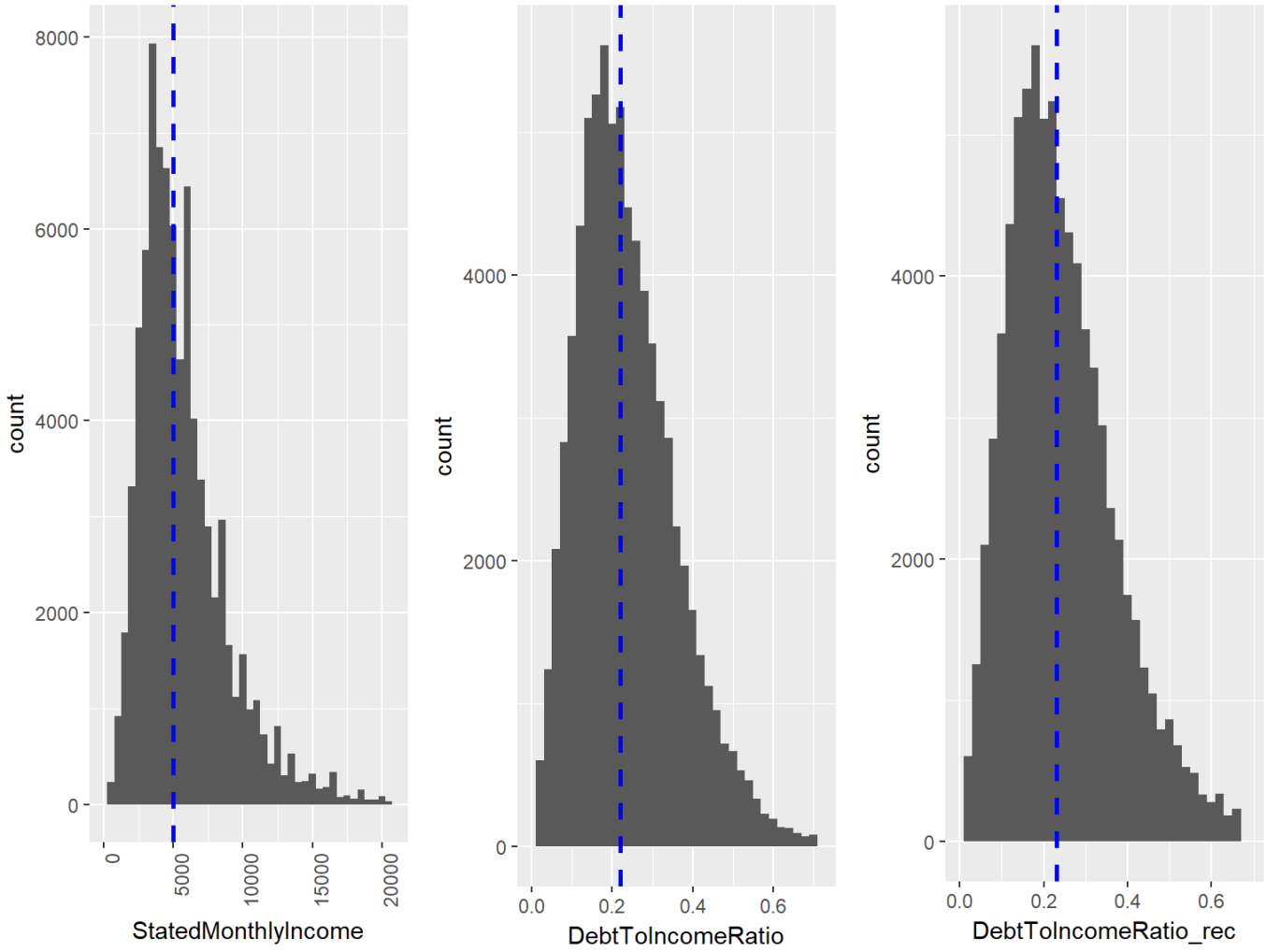
```
## # A tibble: 2 × 2
##   IncomeVerifiable     n
##              <fctr> <int>
## 1            False  7203
## 2             True    11
```

This seems to be true. Prosper don't calculate Debt to Income without the required documents to support borrowers income.

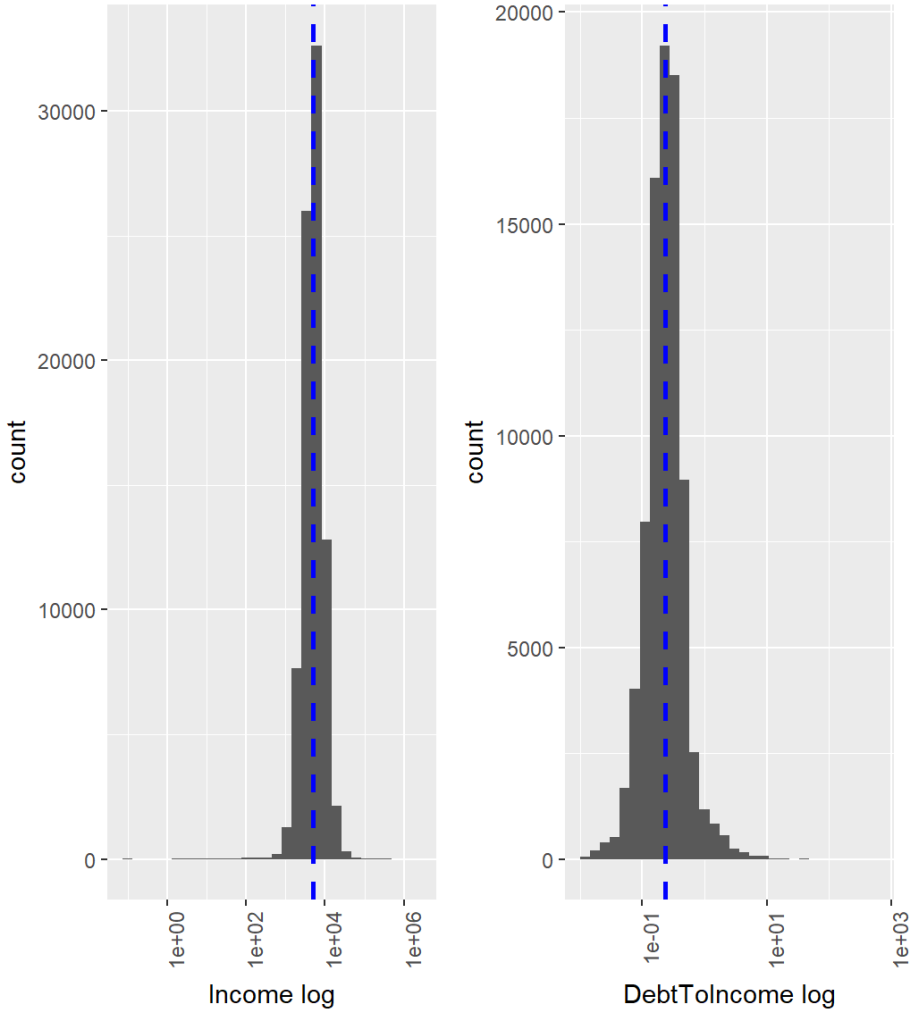
Could consider the income as being the correct one and add the missing debt to income for those with null values.

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##    0.0000    0.1600    0.2300    0.3287    0.3400 437.5000
```

The summaries are changing a lot after recoding missing values in Debt to income variable. The maximum becomes 437… was 10 before.This means we have one or more very high income values.

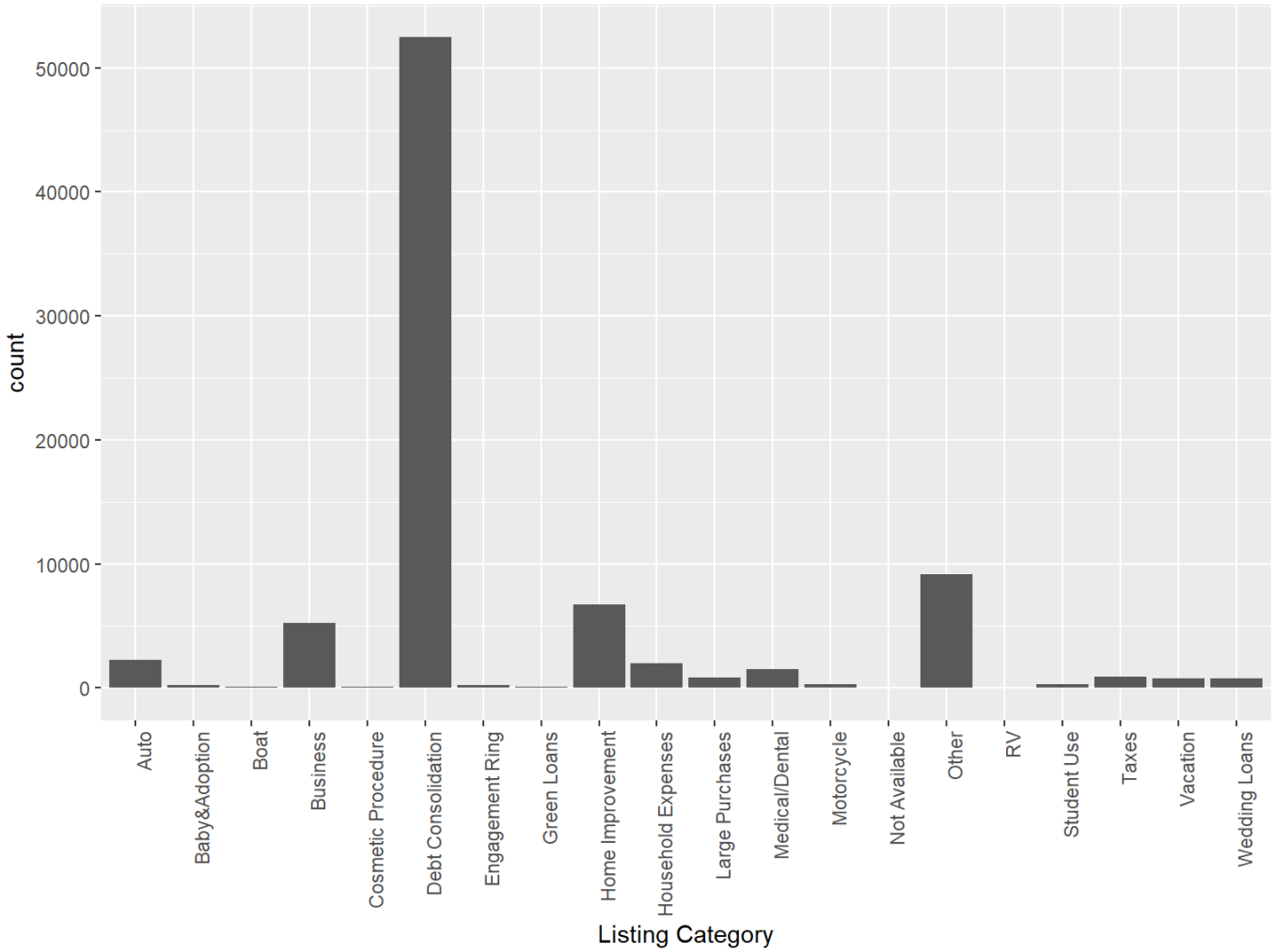


I will log transform the StatedMonthlyIncome and DebtToIncomeRatio_rec to get more normal distributions and to better understand their distributions.



8. LISTING CATEGORY

The purpose of the loanS was mainly debt consolidation.



Univariate Analysis

What is the structure of your dataset?

The initial dataset has 113937 obervations and 81 variables.

After cleaning the table as explained in the dedicated section above, I reduced dimension of the table to 83982 rows.

What is/are the main feature(s) of interest in your dataset?

I'm especially interested in the borrowers' characteristics like: Income, borrower rate, debt to income etc., how are these variables influencing the prosper score and also, what is the evolution in time of the number/amount of loans, or of the financial indicators.

What other features in the dataset do you think will help support your investigation into your feature(s) of interest?

I will also take into consideration:

-Term,

-BankcardUtilization,

-Available Bankcard Credit,

-Current Credit Lines,

-Open Revolving Accounts

Did you create any new variables from existing variables in the dataset?

I created a new variable named Score. This variable takes values from ProsperRating..Alpha. or CreditGrade, so in the end we have a valid value for each borrower.

Also, I created variables for month, year, month&year from the loan origination date.

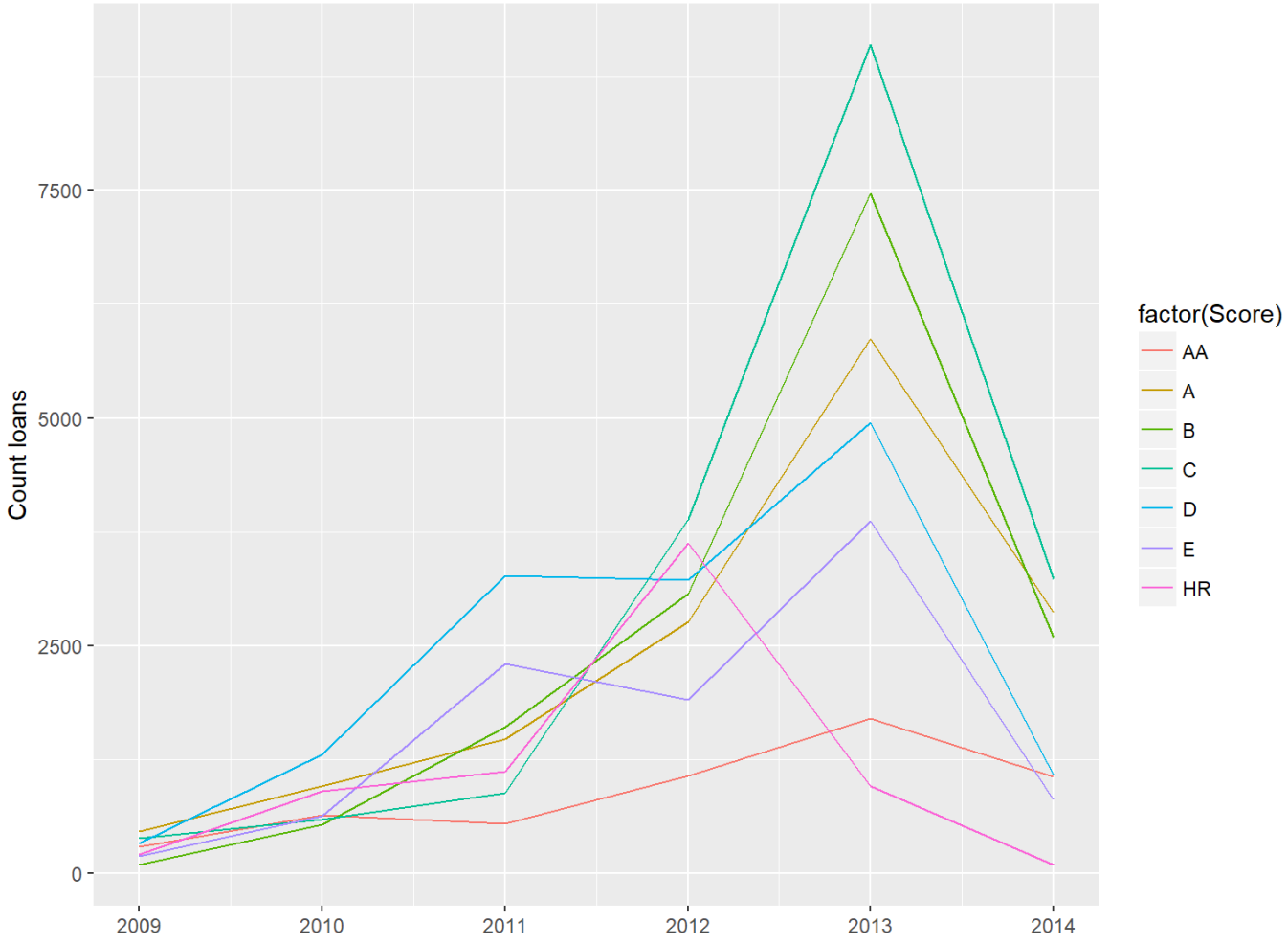
Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

I log transformed the StatedMonthlyIncome and DebtToIncomeRatio_rec. These features are highly positively skewed and I transformed them to show a more normal distribution.

Bivariate Plots Section

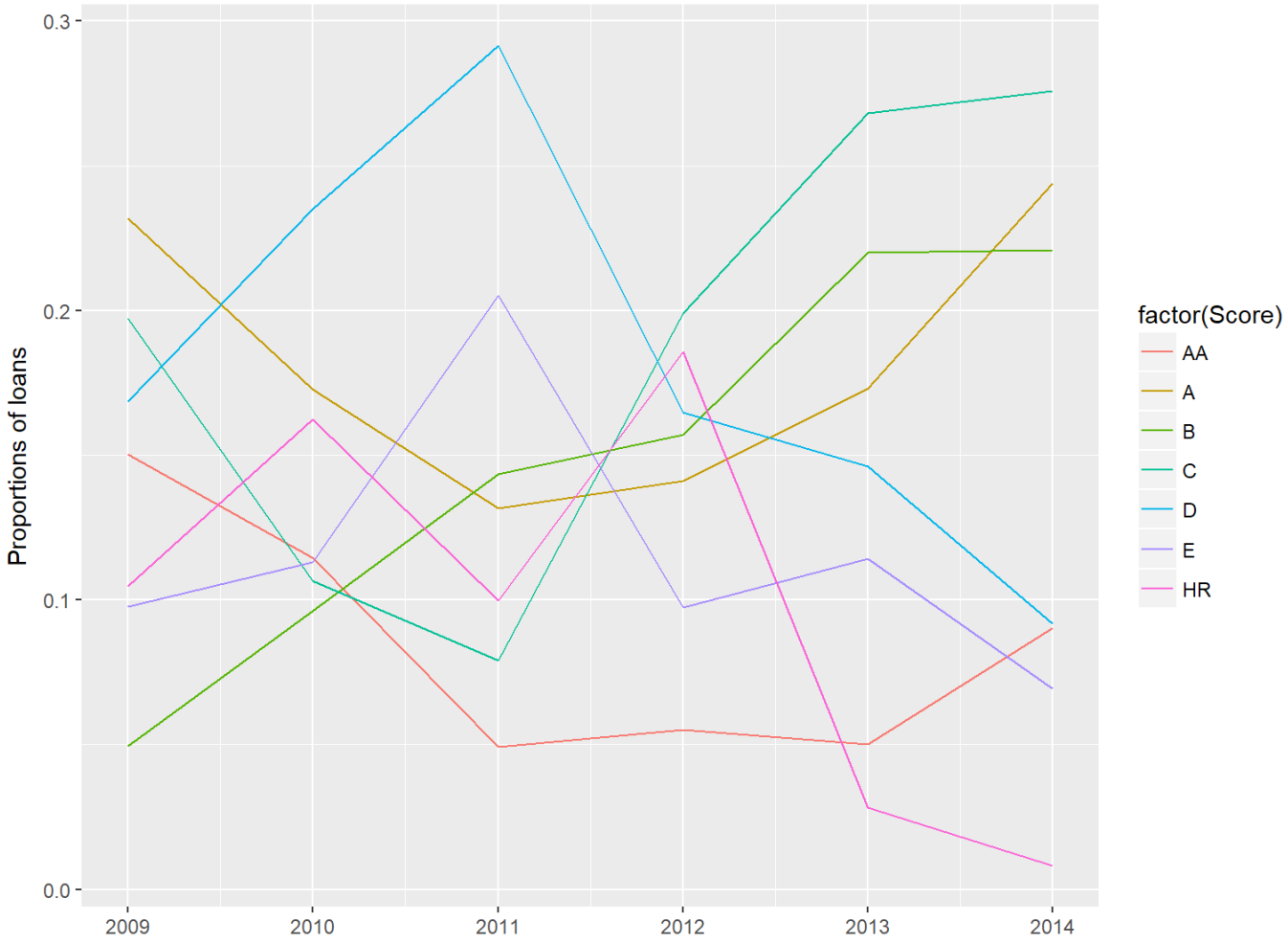
1. HOW MANY LOANS BY PROSPER SCORE AND YEAR?

Aggregate data by year and score, calculate absolute counts and percents, then plot the data.



2014 is deceiving as it contains data for only 3 months, so the plot shows a decreasing trend in 2014.

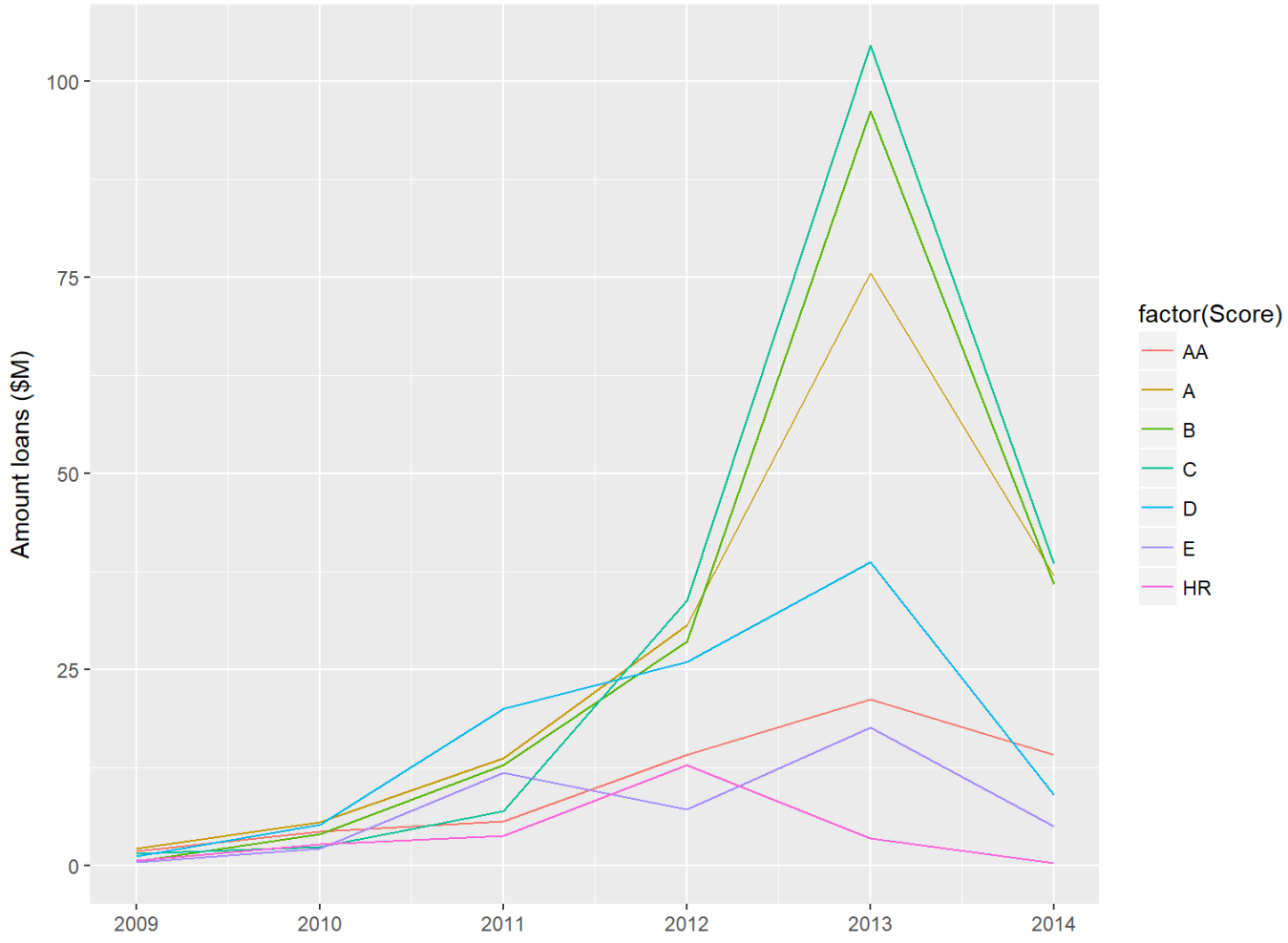
A solution would be to plot loans as % from total of each year.

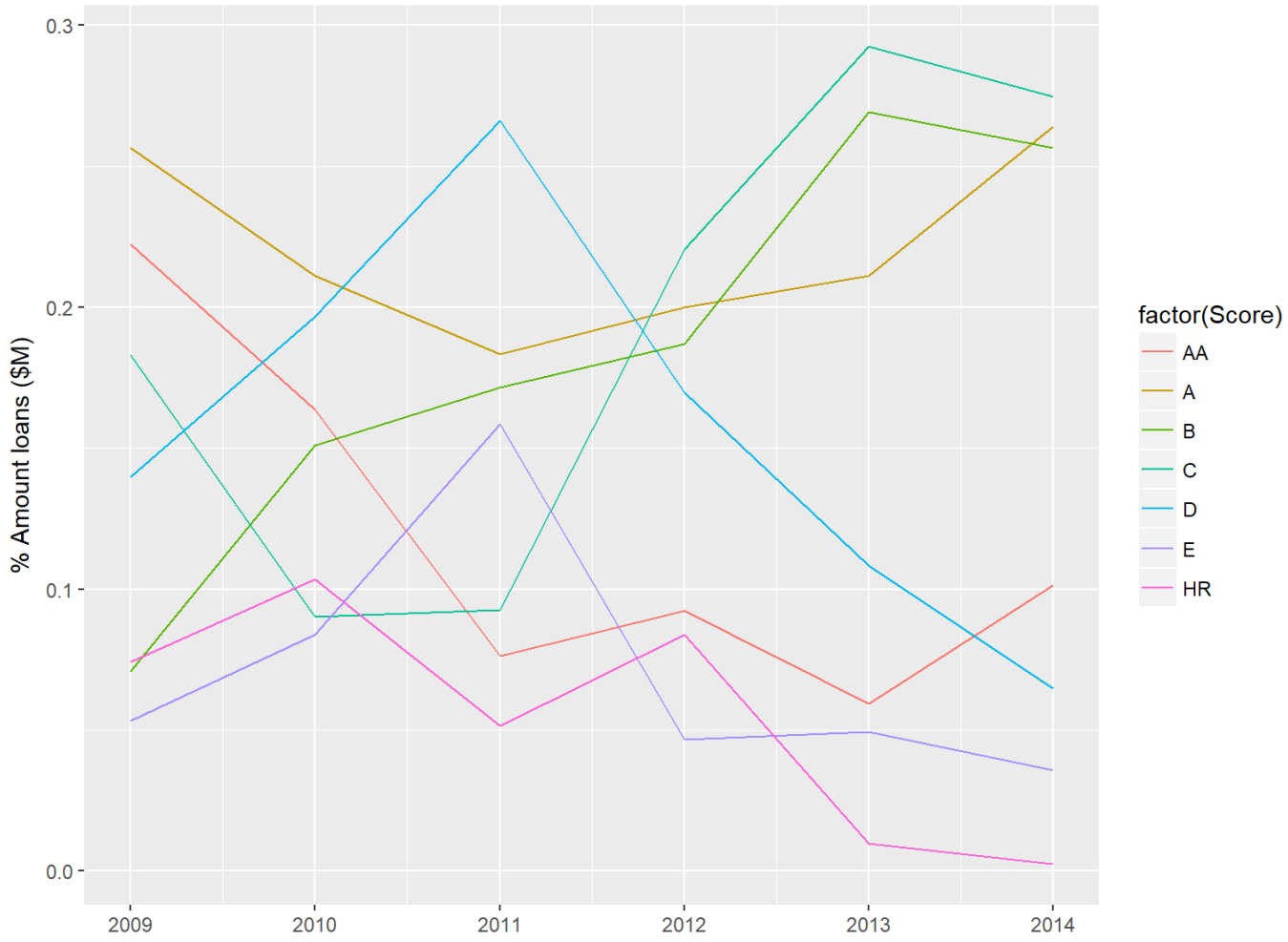


It looks like the number of loans in A,B,C categories have an ascending trend, while D,E,HR tend to descrease over time.

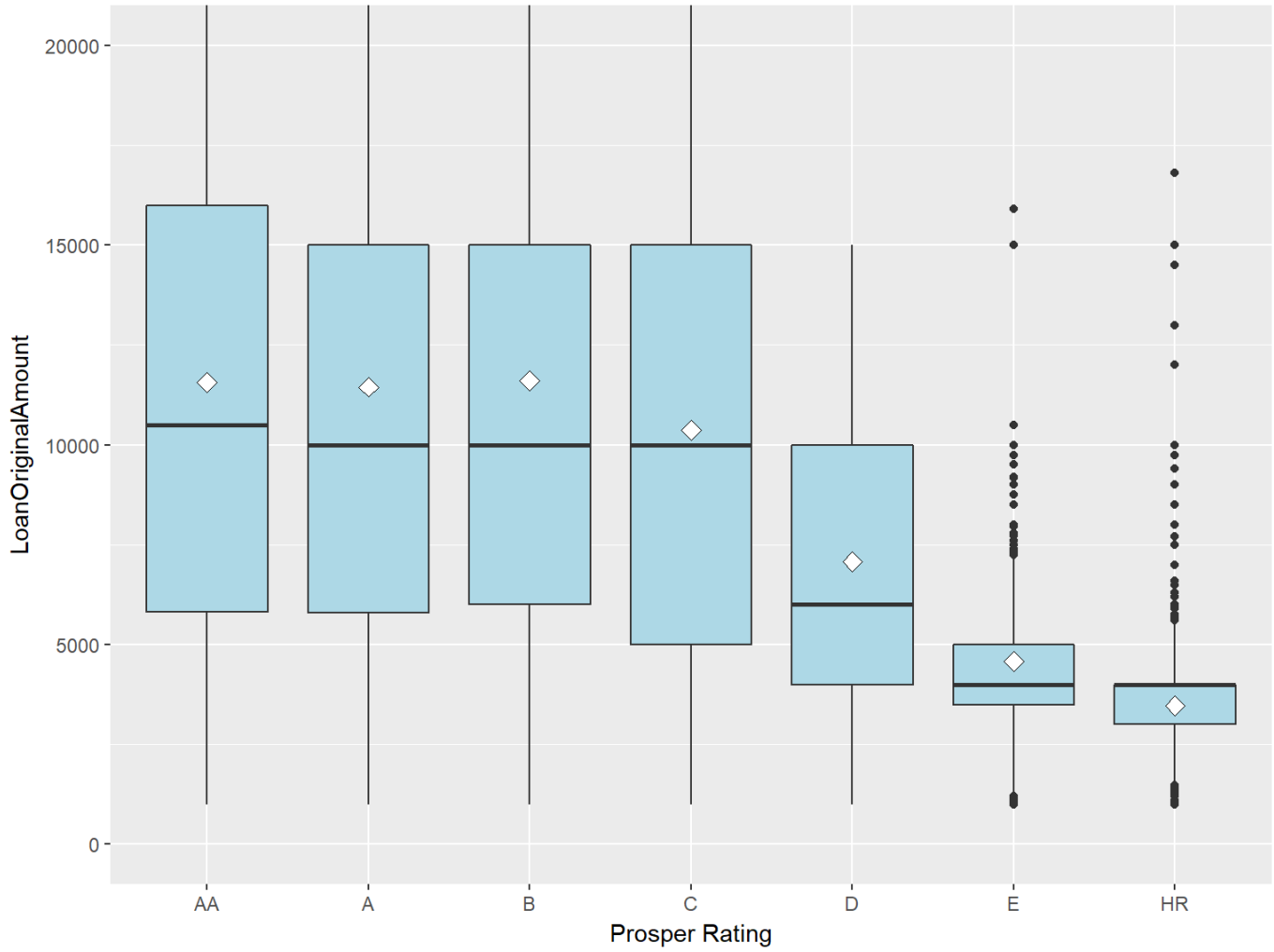
2. LOAN AMOUNT BY PROSPER SCORE AND YEAR

I will apply the same transformation as above and plot the proportions.


Amount loans ($M)
100
75
50
25
0
2009
2010
2011
2012
2013
2014
factor(Score)
AA
A
B
C
D
E
HR


% Amount loans ($M)
0.3
0.2
0.1
0.0
2009
2010
2011
2012
2013
2014
factor(Score)
AA
A
B
C
D
E
HR

3.BOXPLOTS - LOAN AMOUNT BY PROSPER RATING



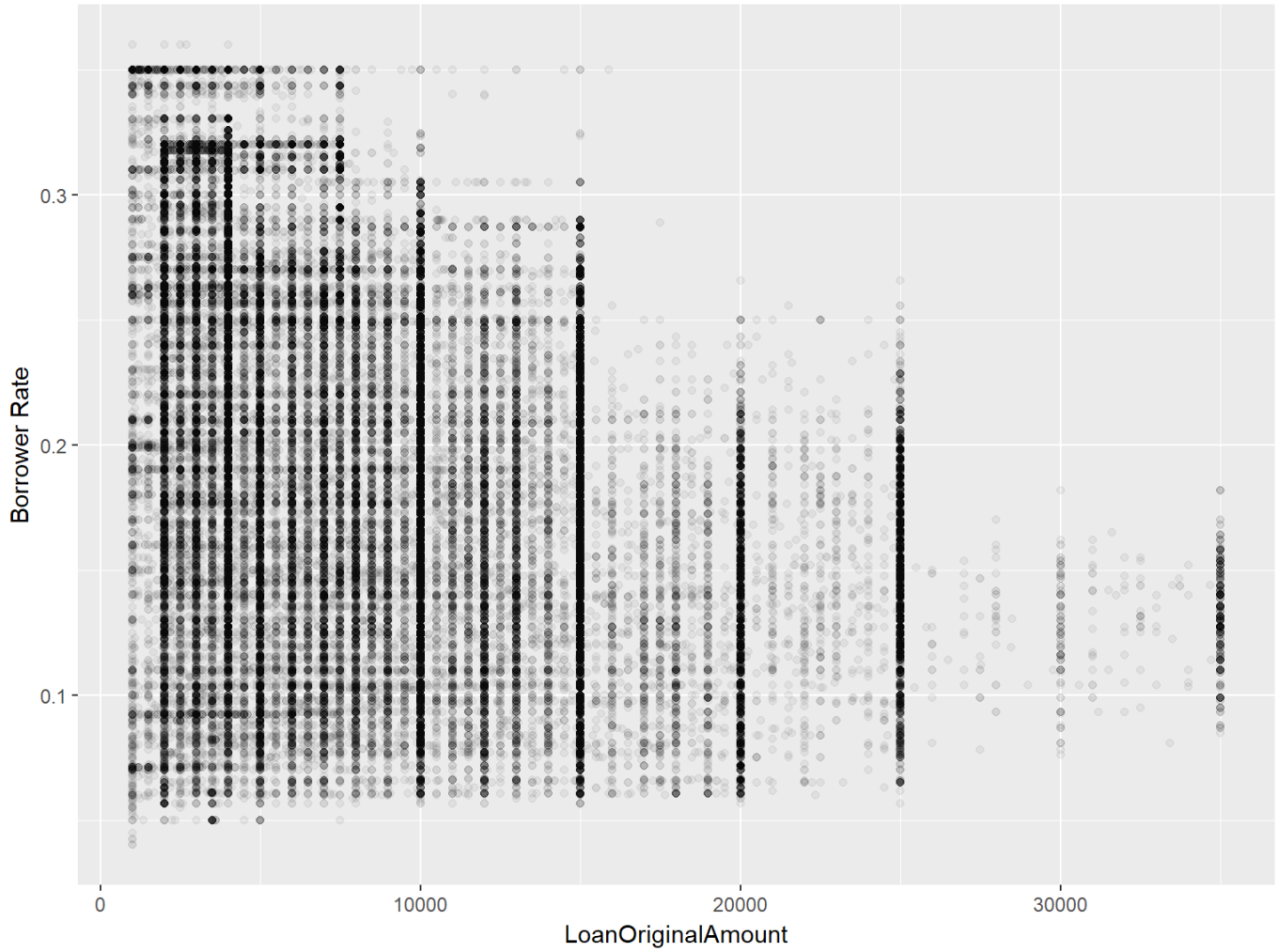
We notice lower loan values for high risk clients E and HR.

HR has negative skewed distribution.

```
##
##  Pearson's product-moment correlation
##
## data:  ProsperRating..numeric. and LoanOriginalAmount
## t = 137.71, df = 83980, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4236691 0.4347039
## sample estimates:
##       cor
## 0.4292025
```

There is a week positive association of 0.43 between loan amount and prosper score.

4. RELATIONSHIP BETWEEN LOAN AMOUNT ANF BORROWER RATE



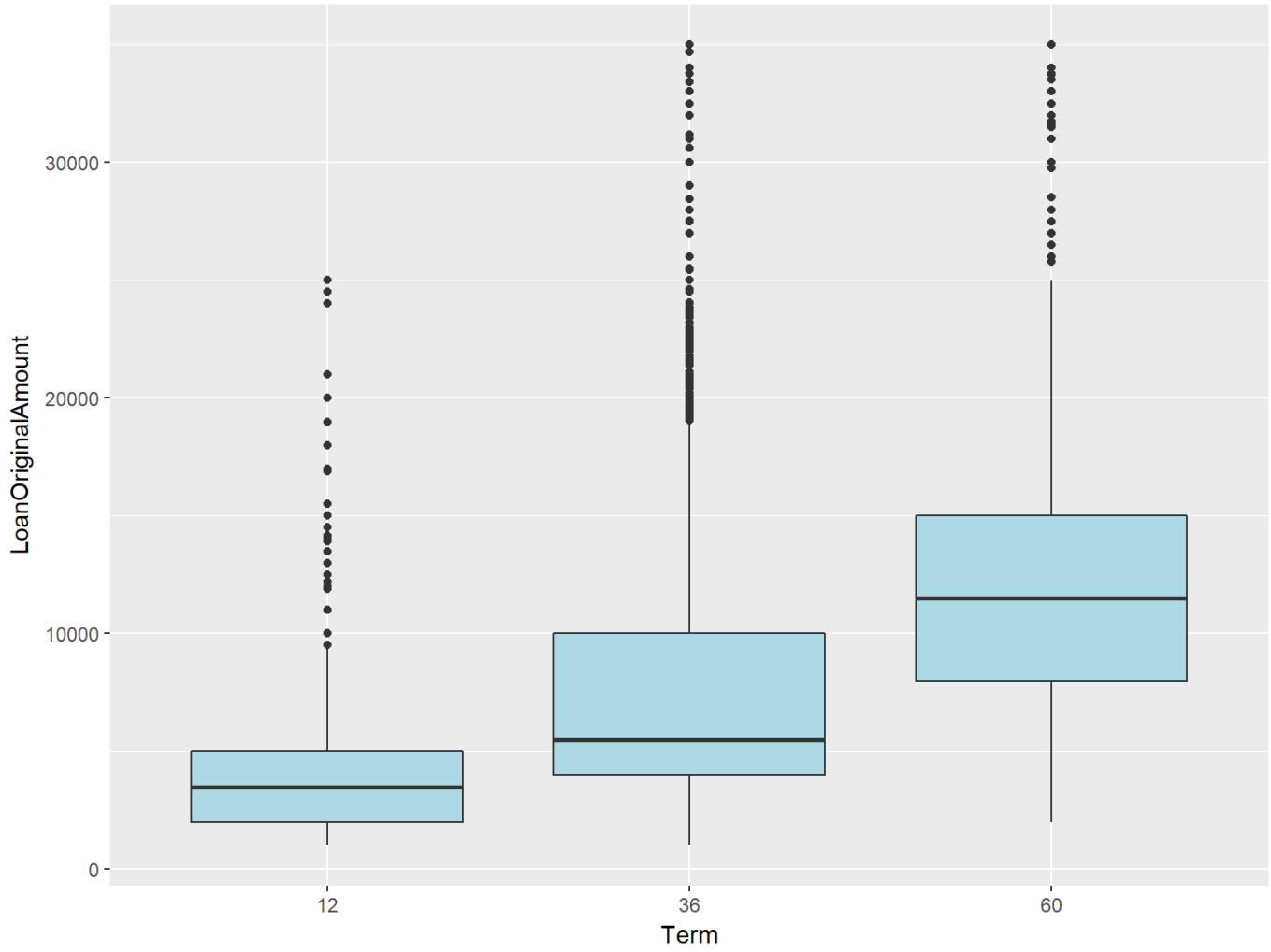
Borower rate is higher for small loans and descreases as loans increase.

Looks like there is a correlation between these two variable. Let's see the Pearson's correlation.

```
##
##  Pearson's product-moment correlation
##
## data:  BorrowerRate and LoanOriginalAmount
## t = -131.64, df = 83980, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4191876 -0.4079748
## sample estimates:
##       cor
## -0.4135969
```

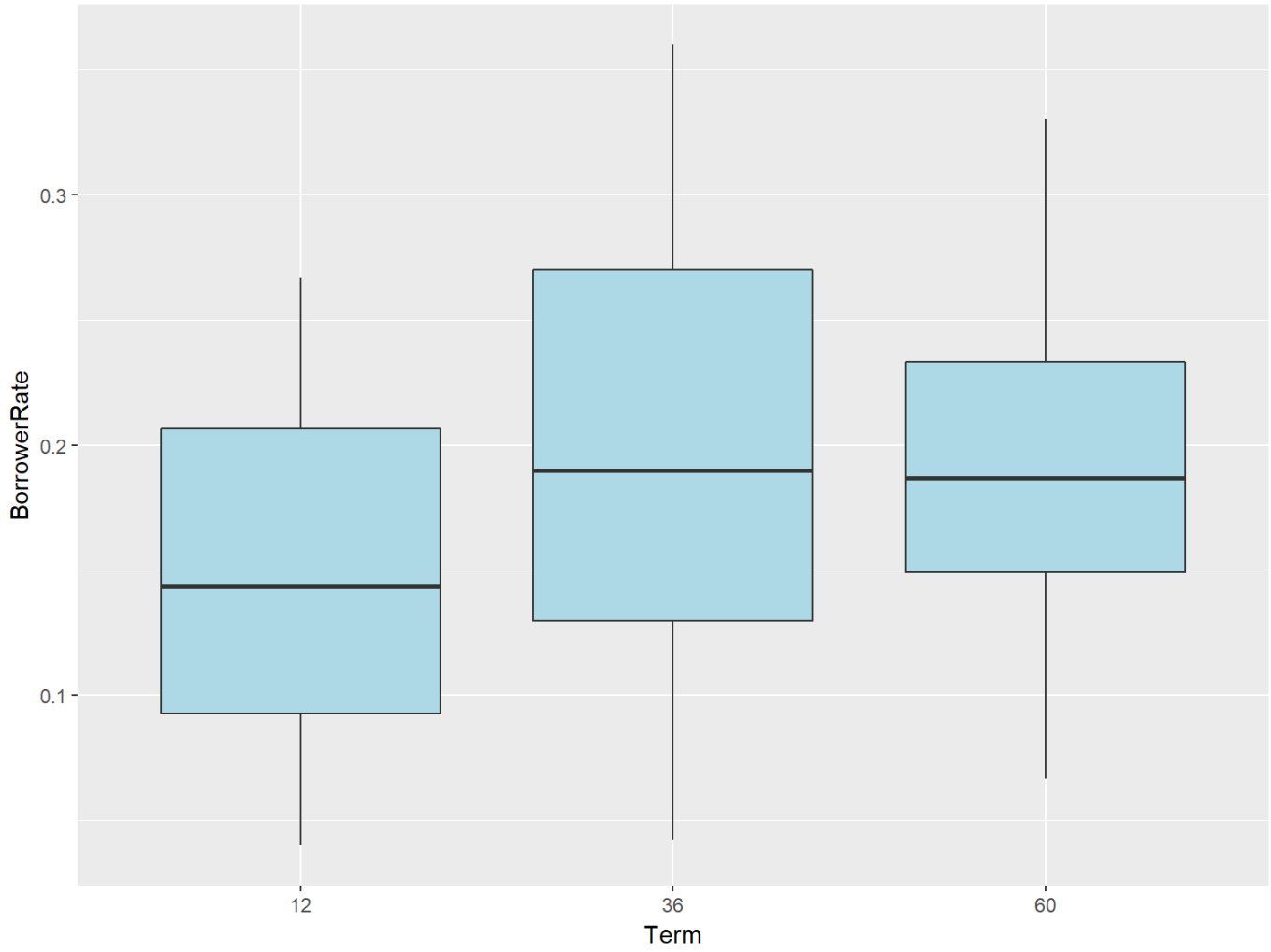
There is a week negative association of -0.41 between loan amount and borrower rate.

5. HOW ARE LOAN AMOUNT AND TERM RELATED? LARGER AMOUNTS ARE MADE ON LONGER TERMS?



Loan terms are 12/36/60 months.For small amounts of money, the loans are made on 12 months term. As the loans amount increase, the term is also increasing.

6. WHAT IS THE RELATIONSHIP BETWEEN TERM AND BORROWER RATE?



The highest borrower rates are for 36 months loans. Is it because most of loans are made on 36 months term?

Table of term:

```
##
##    12    36    60
##  1613 58141 24228
```

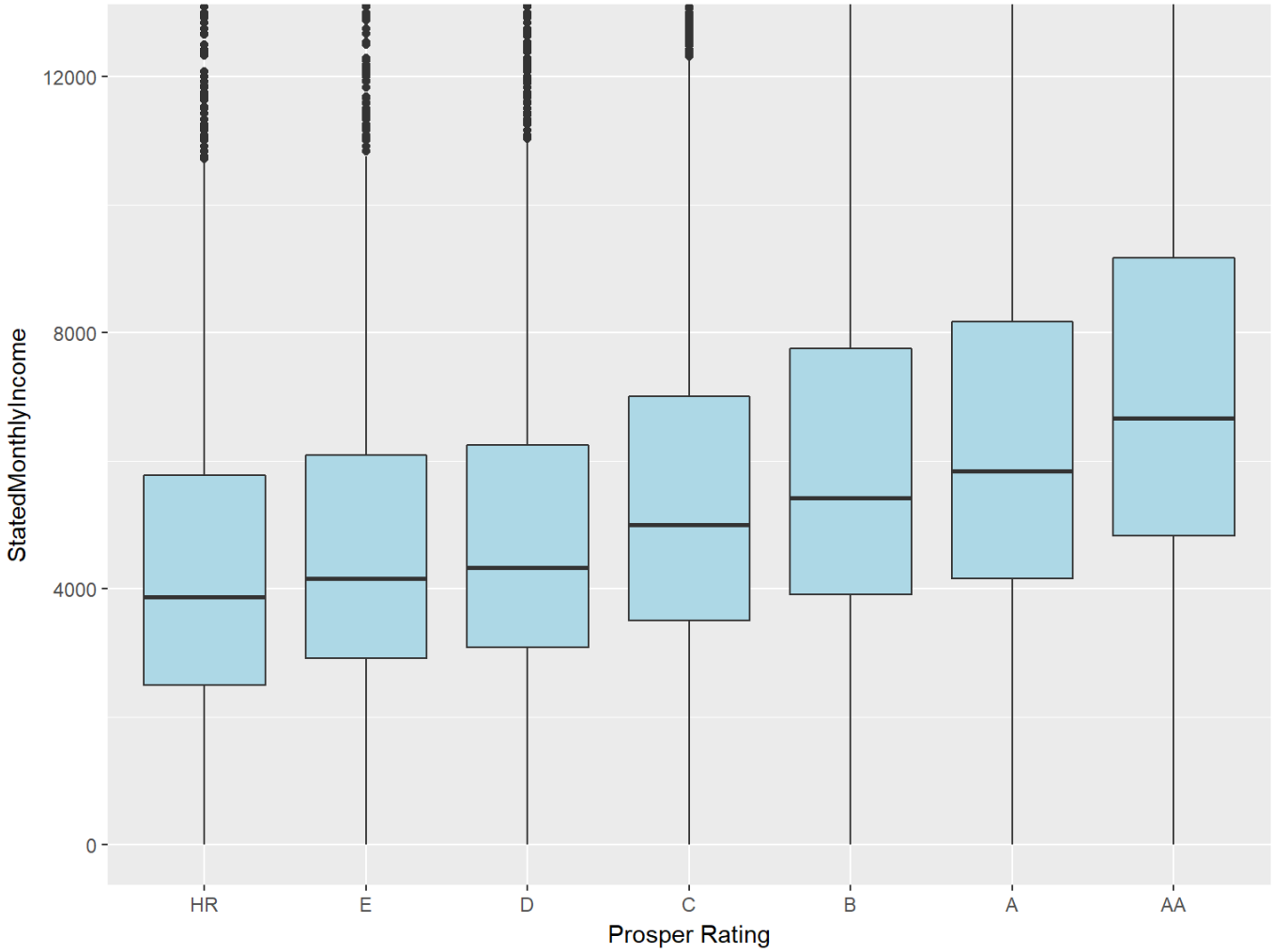
Yes, most of the loans are on 36 month term.

Is there a correlation between borrower rate and term?

```
##
##  Pearson's product-moment correlation
##
## data:  Term and BorrowerRate
## t = -0.097424, df = 83980, p-value = 0.9224
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.007099431  0.006427089
## sample estimates:
##           cor
## -0.0003361864
```

No correlation.

7. MONTHLY INCOME AND PROSPER RATING. DOES A HIGH INCOME ALSO MEAN A BETTER PROSPER SCORE?



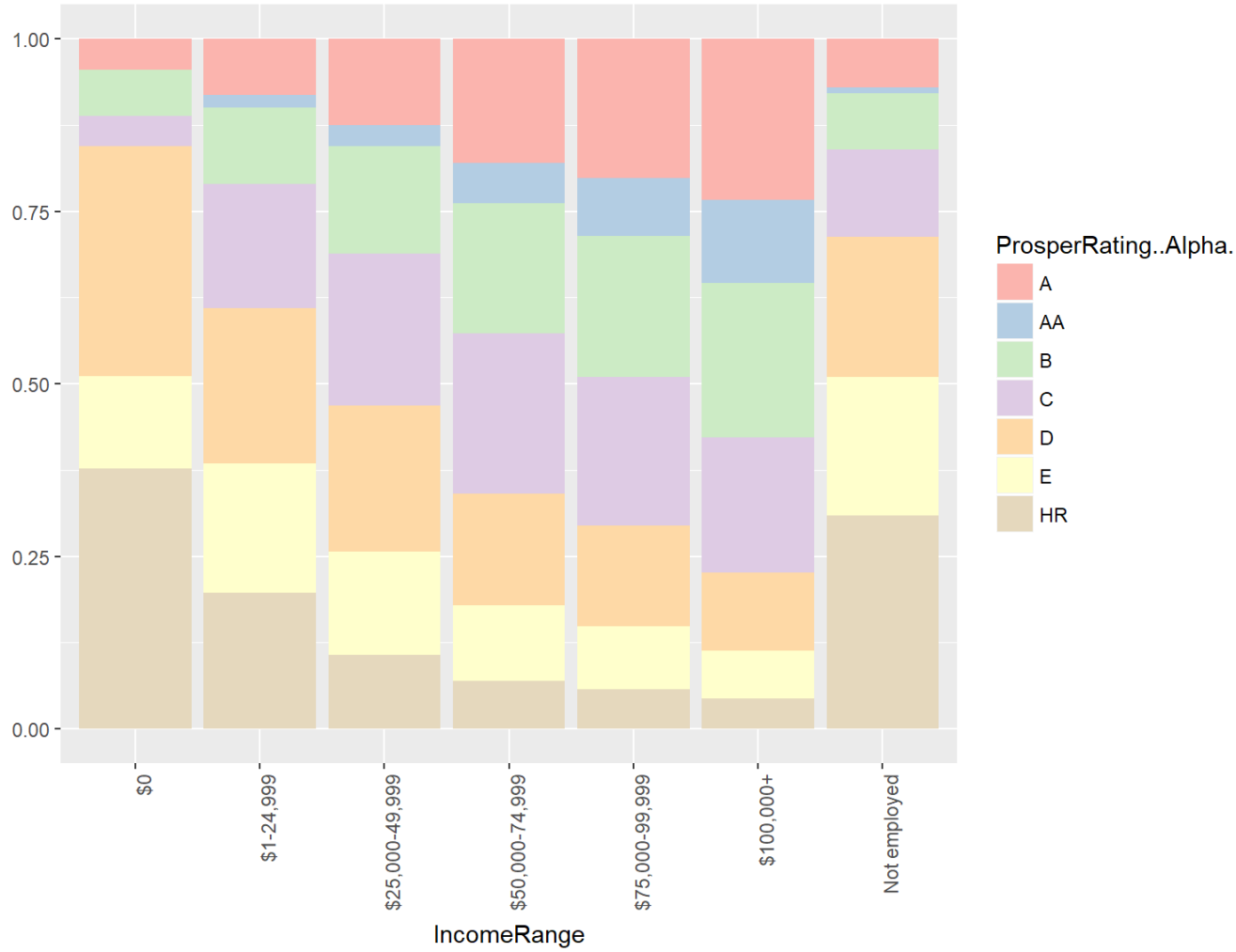
There are better Prosper scores for higher incomes.The income medians are increasing as the score gets better.

Is there any correlation between income and score?

```
##
##  Pearson's product-moment correlation
##
## data:  ProsperRating..numeric. and StatedMonthlyIncome
## t = 27.396, df = 83980, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.08740874 0.10081545
## sample estimates:
##       cor
## 0.09411636
```

No correlation.

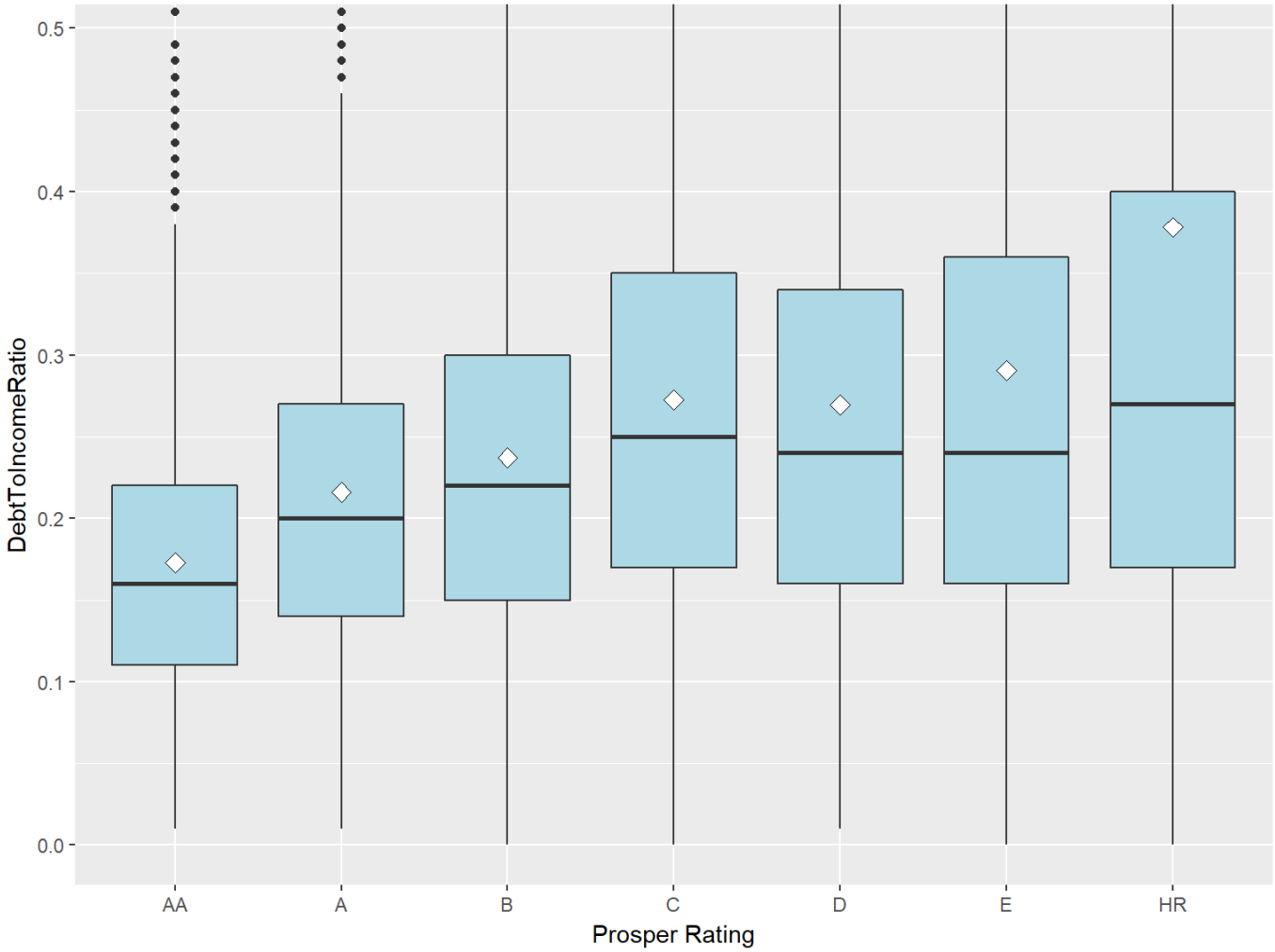
8. INCOME RANGE AND PROSPER RATING



Those with undeclared income ($0 or NA) are scored worse, mostly in HR, E and D category.

As the income increses, the A,AA,B proportions also increase. D,E,HR proportions increase as the income decrease.

9. DEBT TO INCOME RATIO AND PROSPER SCORE



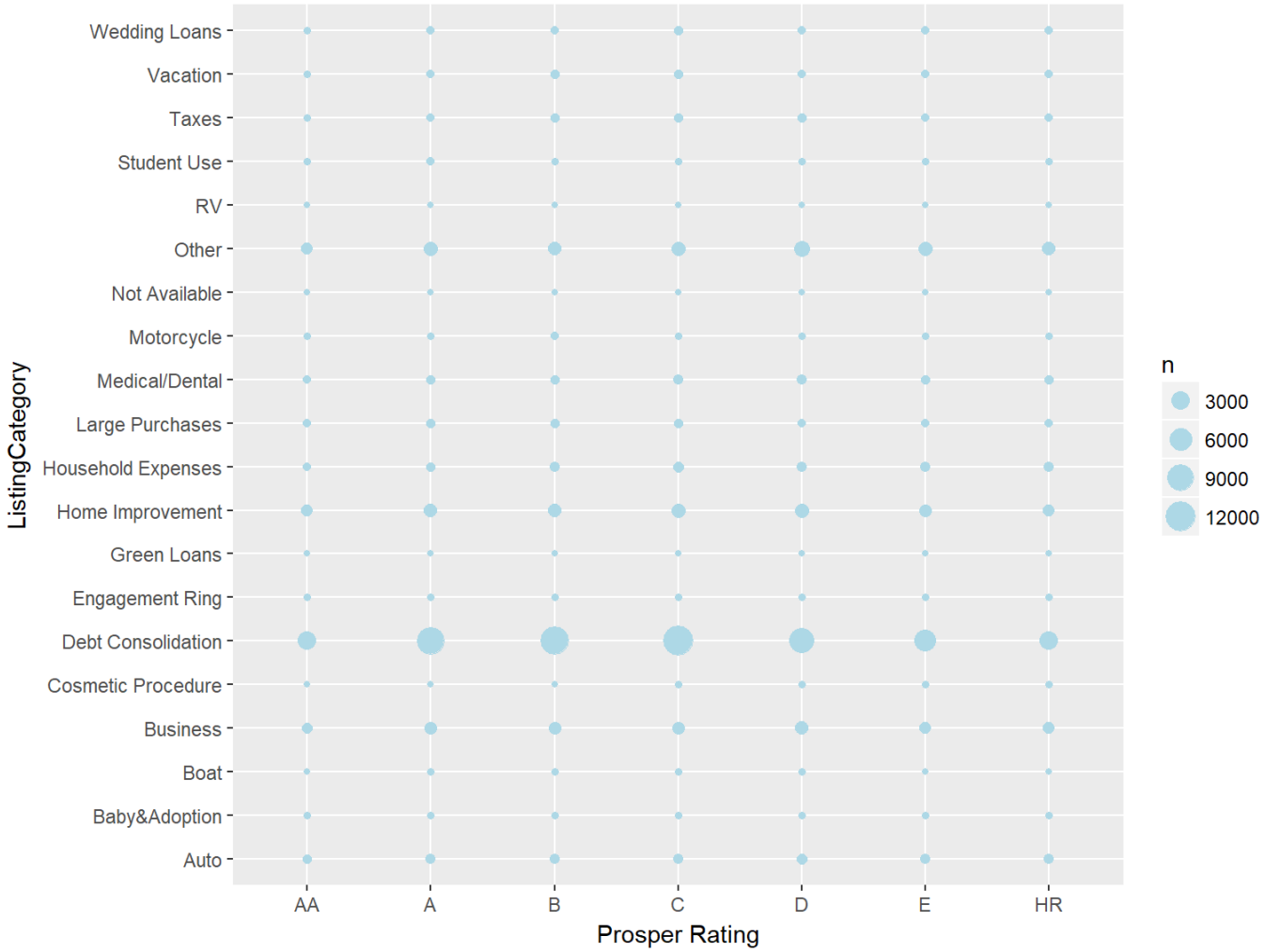
Worse scores for those with high Debt To Income Ratio.

Is there any correlations between score and debt to income ratio?

```
##
##  Pearson's product-moment correlation
##
## data:  ProsperRating..numeric. and DebtToIncomeRatio
## t = -37.845, df = 76766, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1422738 -0.1283851
## sample estimates:
##        cor
## -0.1353361
```

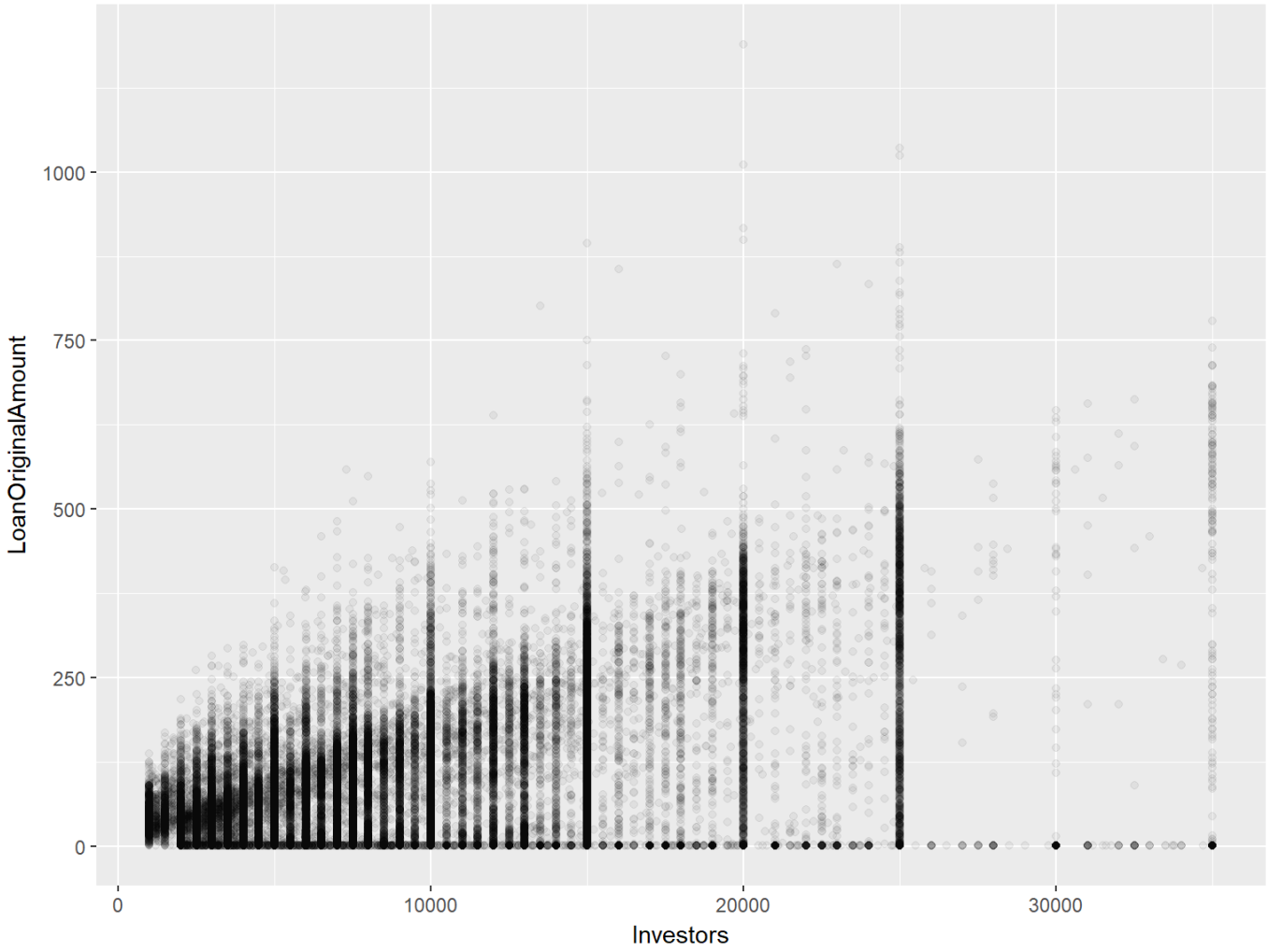
No correlation.

10. LISTING CATEGORY BY SCORE



Most loans are for debit consolidation.

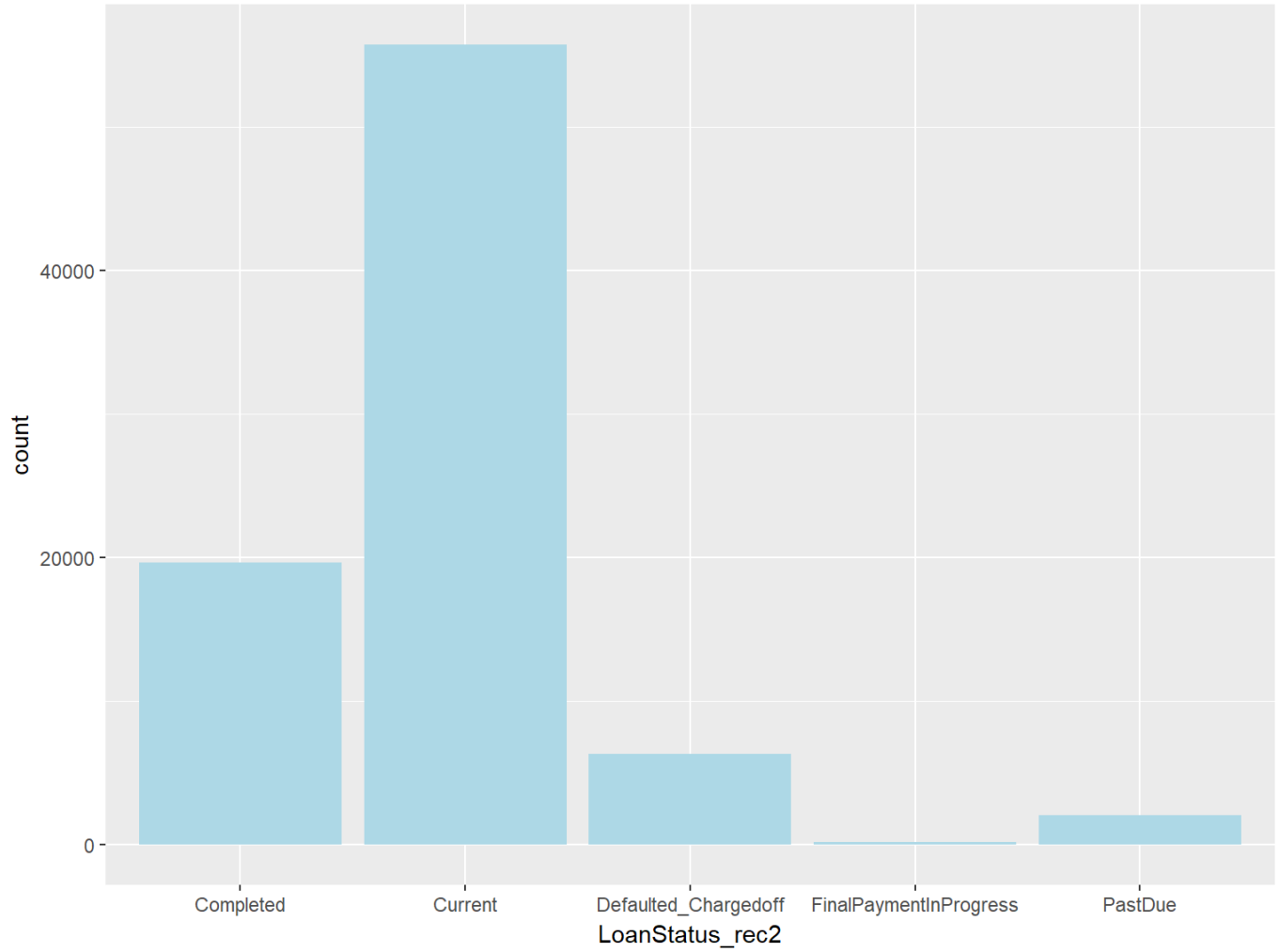
11. LOAN AMOUNT AND NUMBER OF INVESTORS. A high loan amount needs more investors?



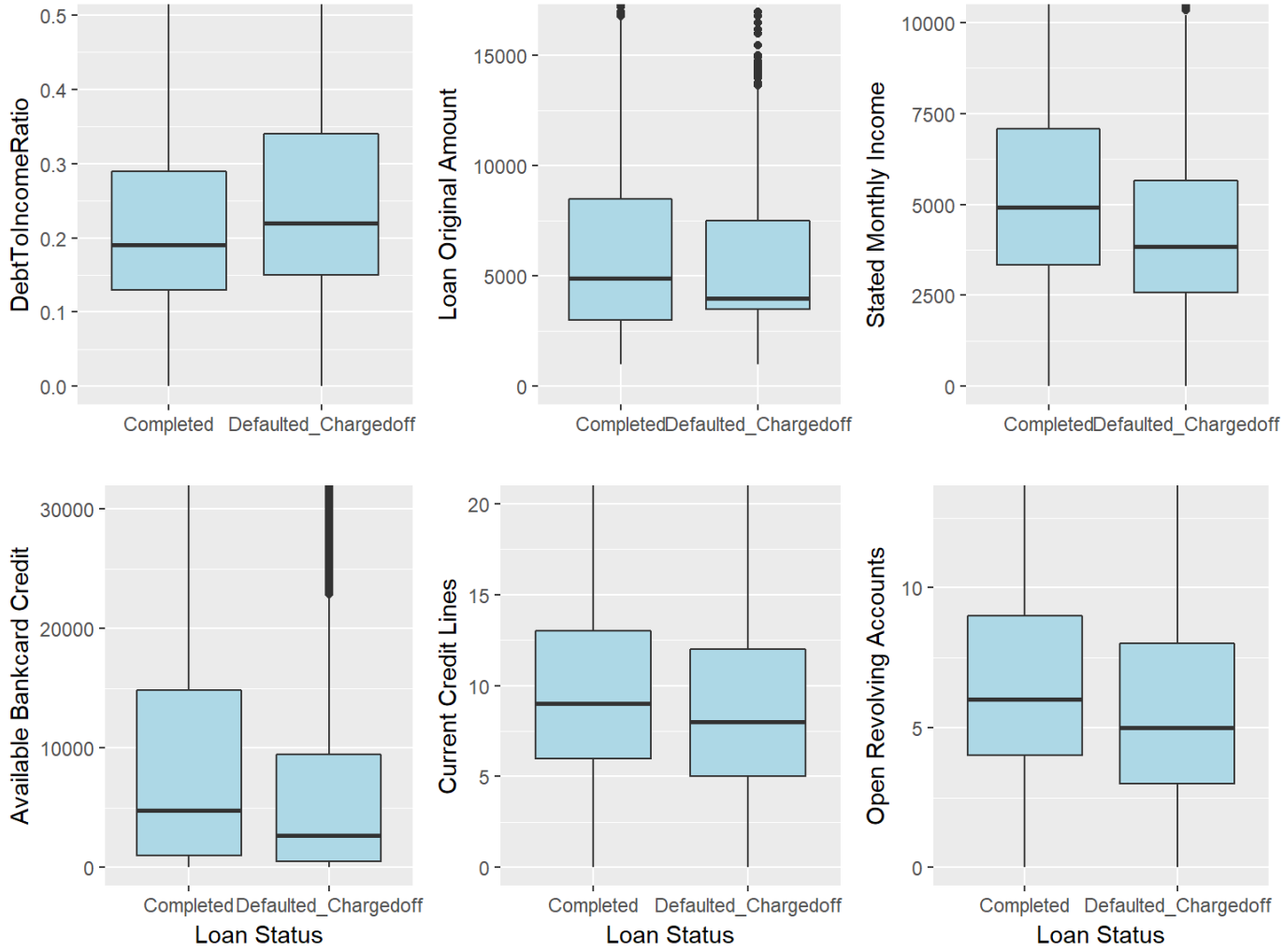
As loan amounts increase, more investors are needed to finance it.

Recode LoanStatus to create a new category Defaulted_Chargedoff which combines Chargedoff and Defaulted borrowers into a single category:

```
## # A tibble: 5 × 2
##          LoanStatus_rec2     N
##                    <chr> <int>
## 1              Completed 19651
## 2                Current 55730
## 3   Defaulted_Chargedoff  6341
## 4 FinalPaymentInProgress   203
## 5                PastDue  2057
```



12. RELATIONSHIPS BETWEEN DEFAULTED STATUS AND FINANCIAL INFORMATIONS ABOUT BORROWERS



Compared to the borrowers who completed the loan payment, the borrower who defaulted have higer debt to income ratio, lower income, lower bankcard credit. These are people who borrow small amount of money.

Bivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

- The number of loans in A,B,C categories have an ascending trend over years, while D,E,HR tend to descrease.

- Week positive association of 0.43 between loan amount and prosper score.

- High risk clients E and HR are borrowing less money.

- Borower rate is higher for small loans and descreases as loans increase.There is a week negative association of -0.41 between loan amount and borrower rate.

- Usually, for small amounts of money, the loans are made on 12 months term. As the loan amounts increase, the term is also increasing.

- Compared to the borrowers who completed the loan payment, the borrower who defaulted have higer debt to income ratio, lower income, lower bankcard credit.

- There are better Prosper scores for higher incomes.

- Those with undeclared income ($0 or NA) are scored worse, mostly in HR, E and D category.

- Worse Prosper scores for those with high Debt To Income Ratio.

Did you observe any interesting relationships between the other features (not the main

feature(s) of interest)?

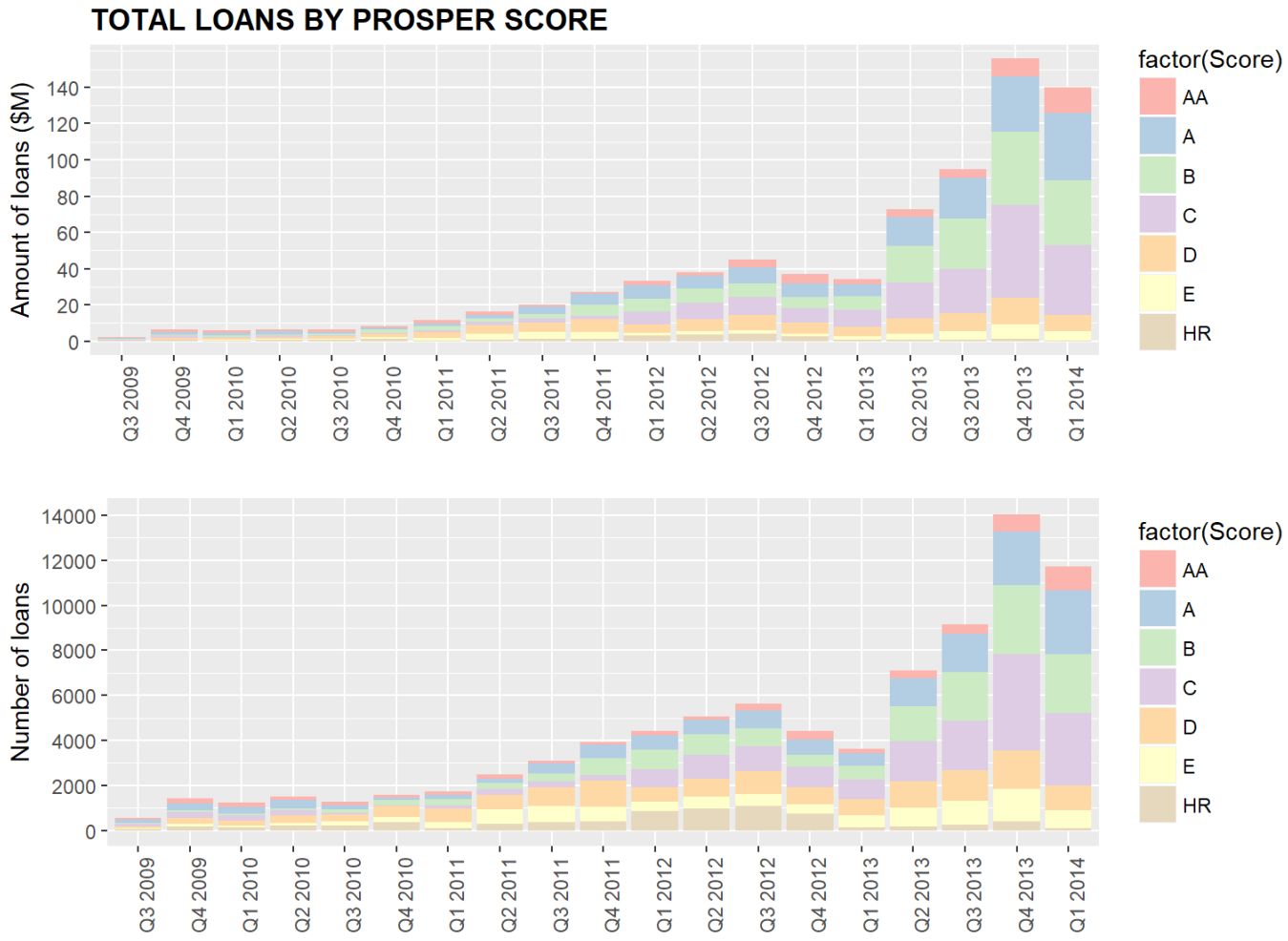
- The number of investors is growing when the large loans amount.
- Most loans are for debit consolidation.

What was the strongest relationship you found?

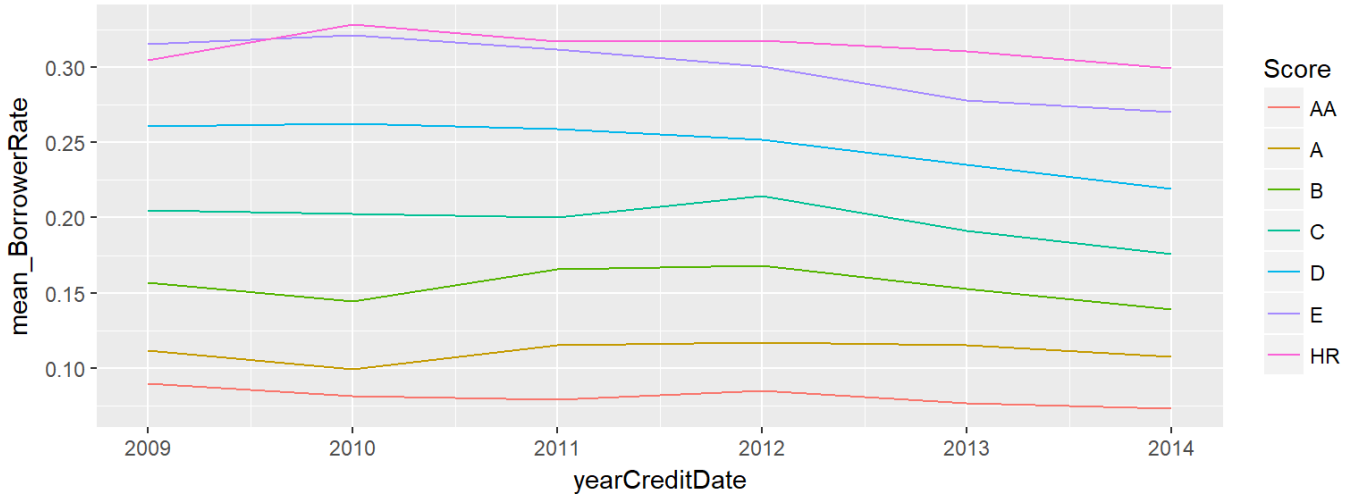
There's a high negative correlation of -0.95 between borrower rate and prosper score.

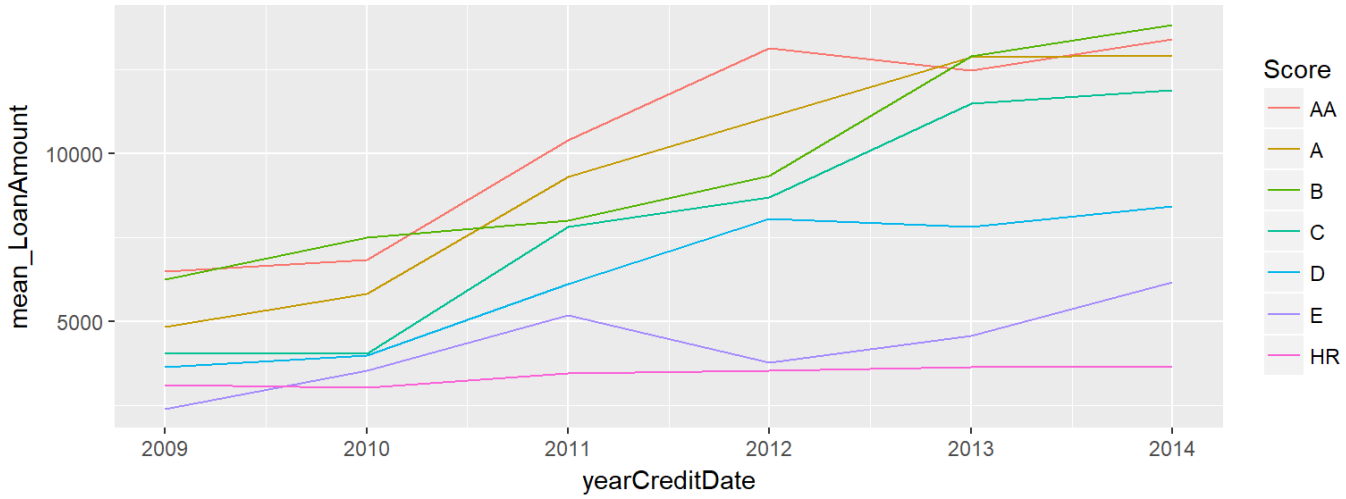
MULTIVARIATE PLOTS SECTION

1. QUARTERLY NUMBER AND AMOUNT OF LOANS BY PROSPER SCORE



2.BORROWER RATE/LENDER YIELD/LOAN AMOUNT MEANS BY PROSPER SCORE YEARLY EVOLUTION

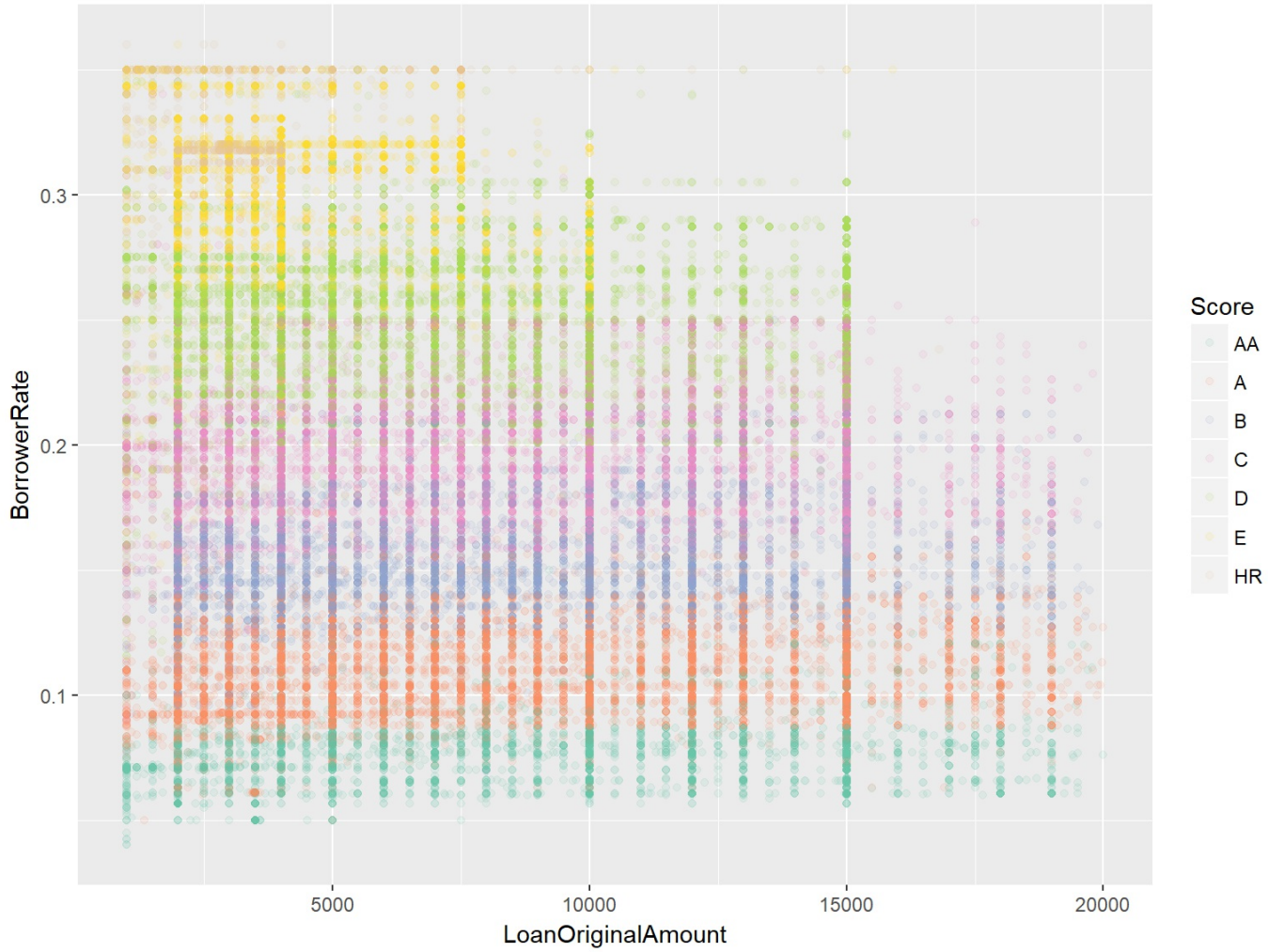




Borrower rate seems slightly decrease over time for all Prosper score category.

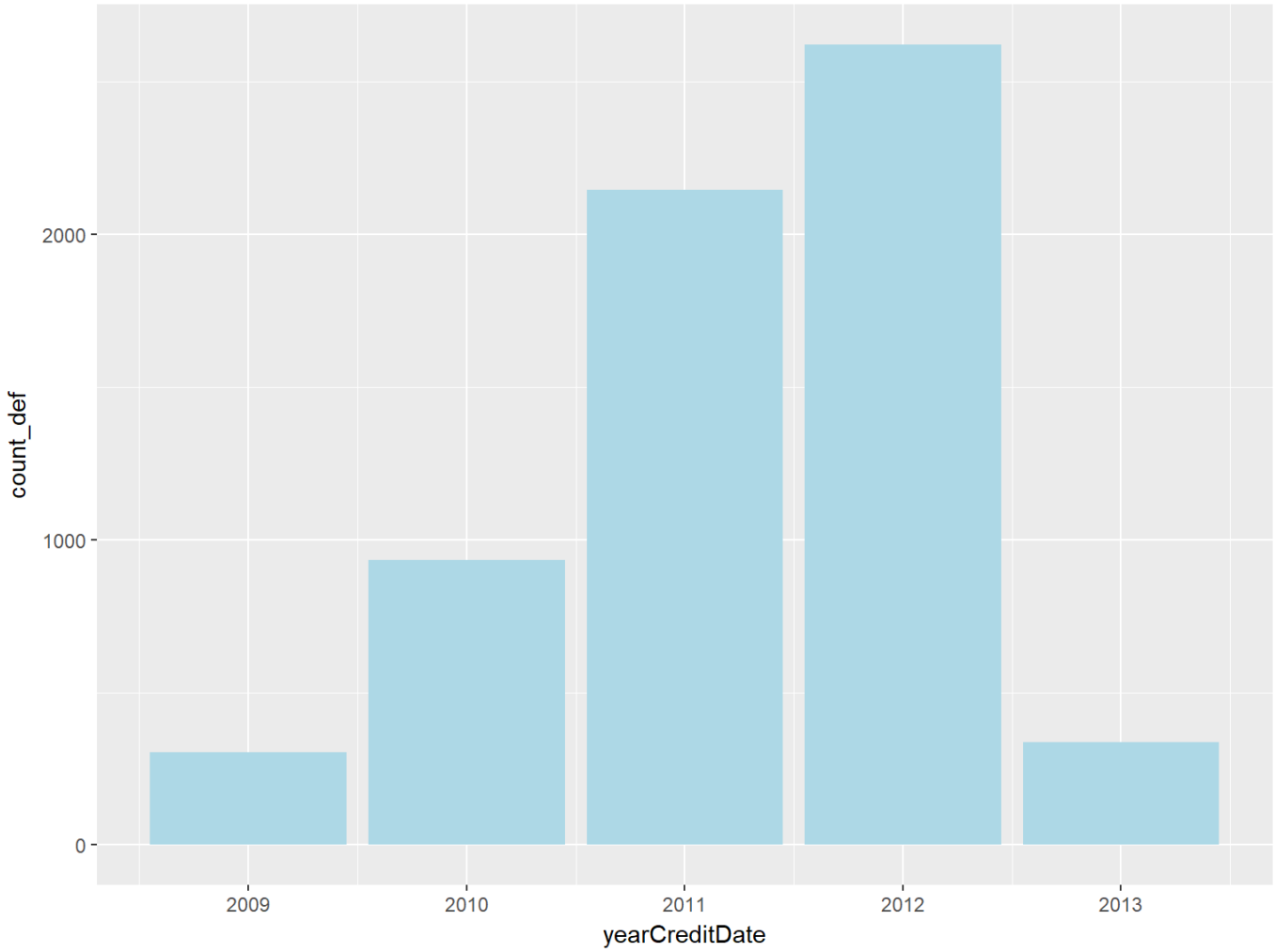
Loan amount is increasing over time. AA, A,B and D categories have significant increases, D has a moderate increase, while HR and E very small or no increases.

3.BORROWER RATE, LOAN AMOUNT AND PROSPER SCORE



The borrowers with higher scores have lower rates. E and HR categories borrow less money, but they pay highest rates.

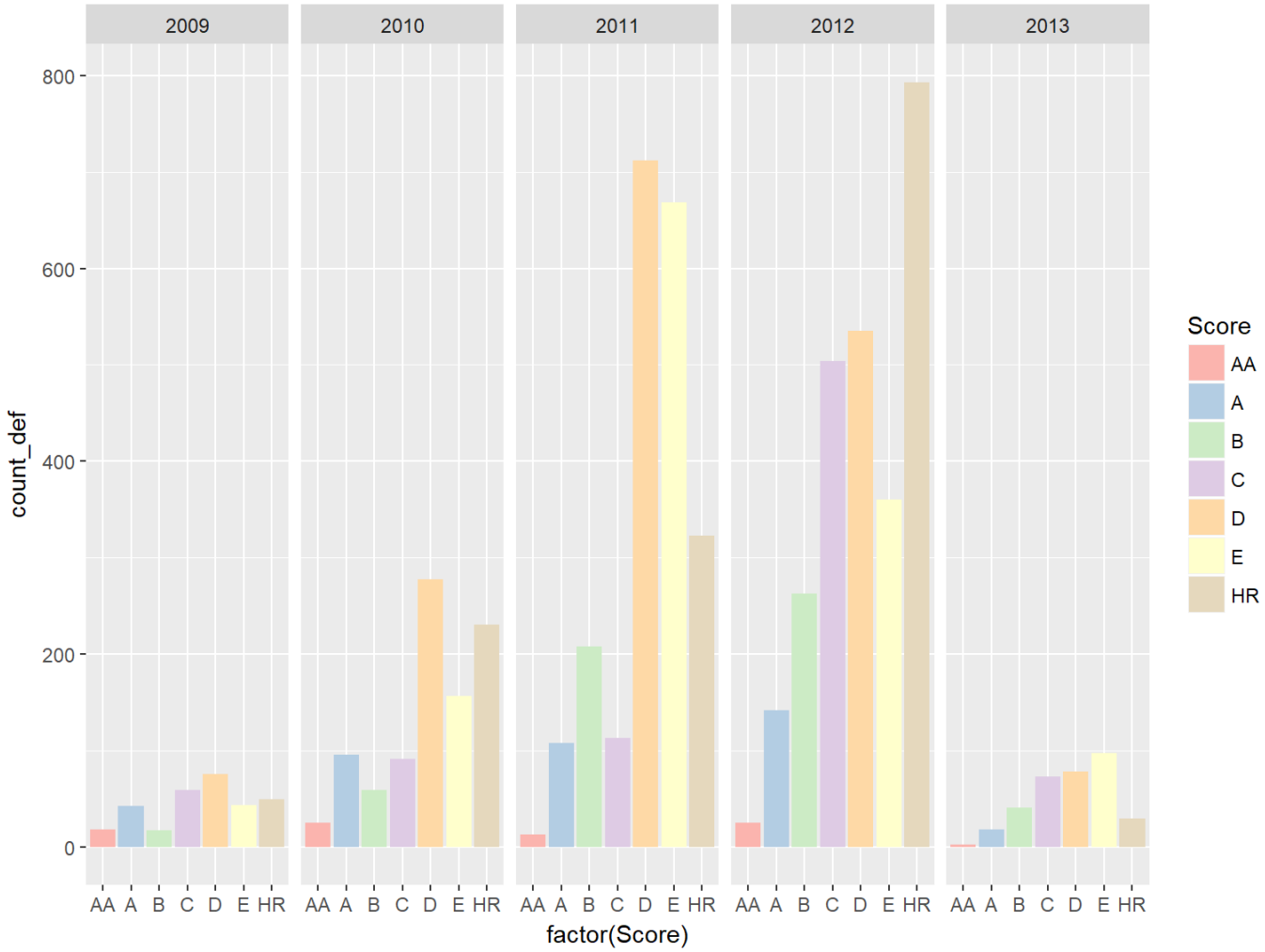
4. YEAR WHEN THE CREDIT WAS TAKEN AND THE DEFAULTED STATUS



We see an increase in 2011 and 2012, but that doesn't necessarily mean that loans originated in these periods defaulted more.

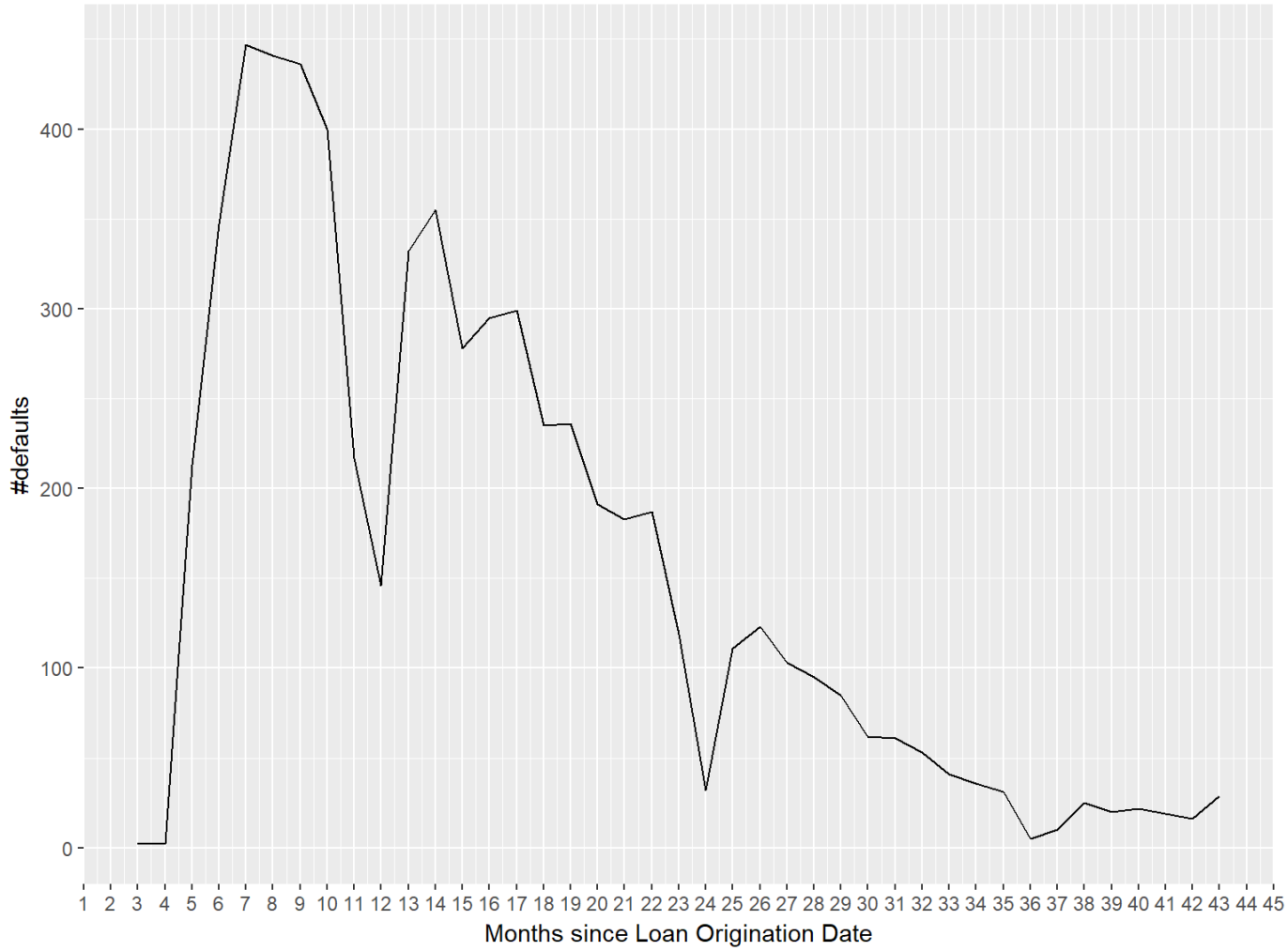
The number of loans started to significant increase in 2011, so more loans result in more chances to default. 2013 has low counts most probably because the short time since credit was taken.A different approach is needed (see the final section of this project)

5. DEFAULTS PER YEAR AND SCORE



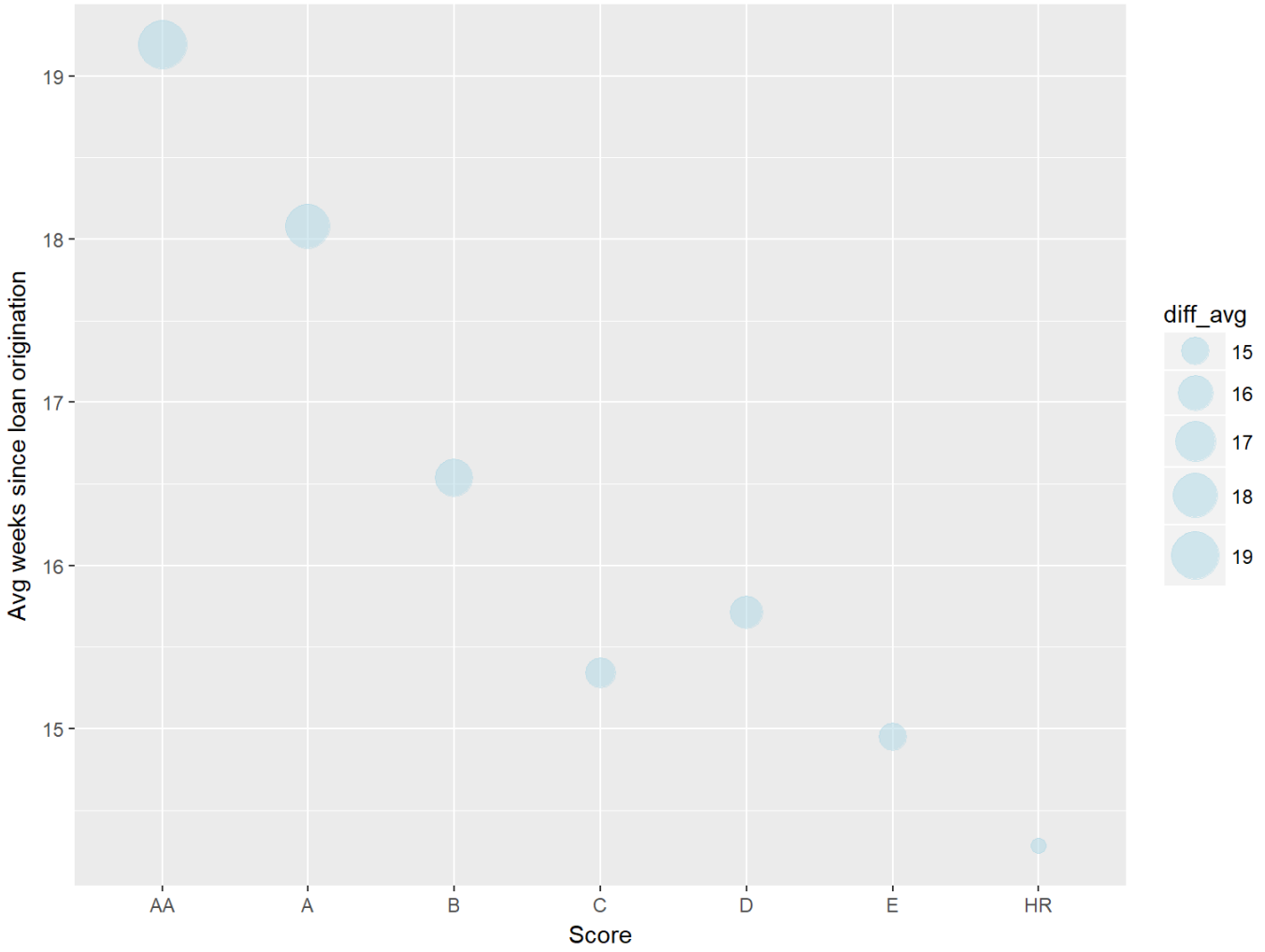
Most of defaults happen from D,E,HR categories. In 2012, there was a a very high increase of defaults in C and HR categories.

6. After how many months since credit was initiated do borrowers default?



There is a huge peak around the 6th month after the credit was originated, then the trend is descending, with a second peak around the 14th month.

7. AFTER HOW MANY AVERAGE WEEKS SINCE LOAN ORIGINATION DO BORROWERS OF EACH PROSPER SCORE CATEGORY DEFALUT ?



As the Prosper score is better, the default is happening later, after more weeks since loan origination.

Multivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. Were there features that strengthened each other in terms of looking at your feature(s) of interest?

- Borrower rate seems to slightly decrease over time for each Prosper score category.

- The borrowers with higher scores have lower rates.

- Most of defaults happen from D,E,HR categories. In 2012, there was a a very high increase of defaluts in C and HR categories.

- The defaults happens mostly in the first 6 months after the credit was originated.

- For higher Prosper score the default happens later.

Were there any interesting or surprising interactions between features?

I didn’t expect to find such nice delimitations in the scatterplot of borrower rate, loan amount and prosper score.

I knew that the rate was influenced by the score, but I was expecting a higher variation among loan amount.

OPTIONAL: Did you create any models with your dataset? Discuss the strengths and limitations of your model.

I created a logistic regression model to predict the borrowers with default risk.

From the variables I initialy selected, I removed the ones which were not significant for the model and rebuilt the model with only the significant features.

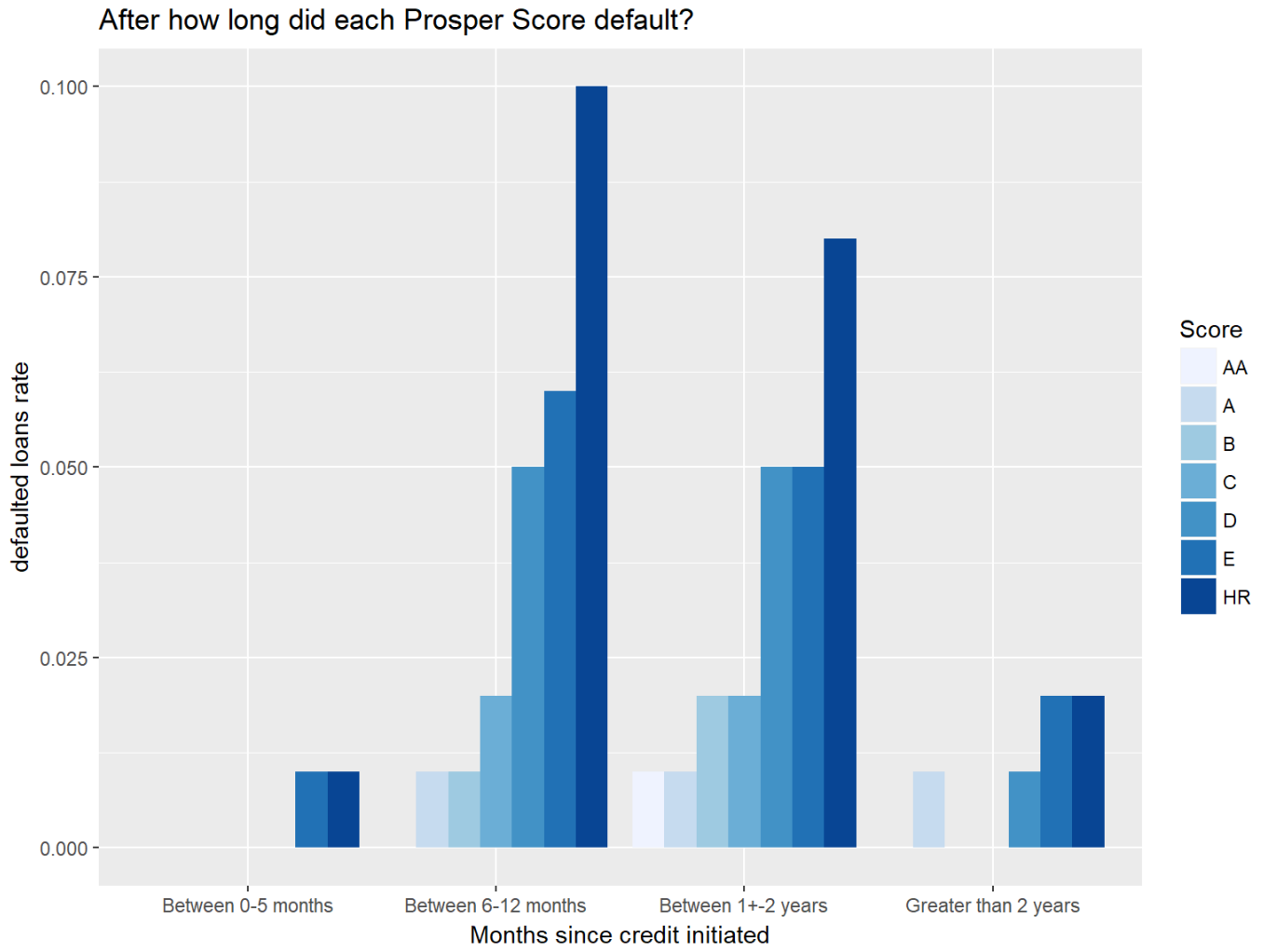
Measuring the accuracy of the model on the test set, at a trashhold cutoff equal to 0.5, I got 75.9% accuracy, 10% sensitivity and 97% specificity. The baseline model accuracy was 75.6%.Ploting the ROC curve and calculating the AUC (area under the curve) which is typical performance measurement for a binary classifier, I got AUC=0.72.

The sensitivity score I obtained is very low, only 10%. This means that the model identify only 10% of the true positives (borrowers who are likely to default). But, a credit company is likely to be more concerned with sensititivy because this way they can reduce the risk. Therefore, they may be more concerned with tuning a model so that their sensititivy is improved.I played with the trashhold cutoff to see how the accuracy of the model is changing. If I decrease the cutoff, the sensitivity is increasing but at the cost of a lower overall accuracy and specificity. Playing around with the trashhold cutoff, the features included in model or even using a more complex model, the accuracy may be improved.

Final Plots and Summary

Plot One

% OF DEFAULTS BY PROSPER SCORE AND CREDIT ORIGINATED DATE



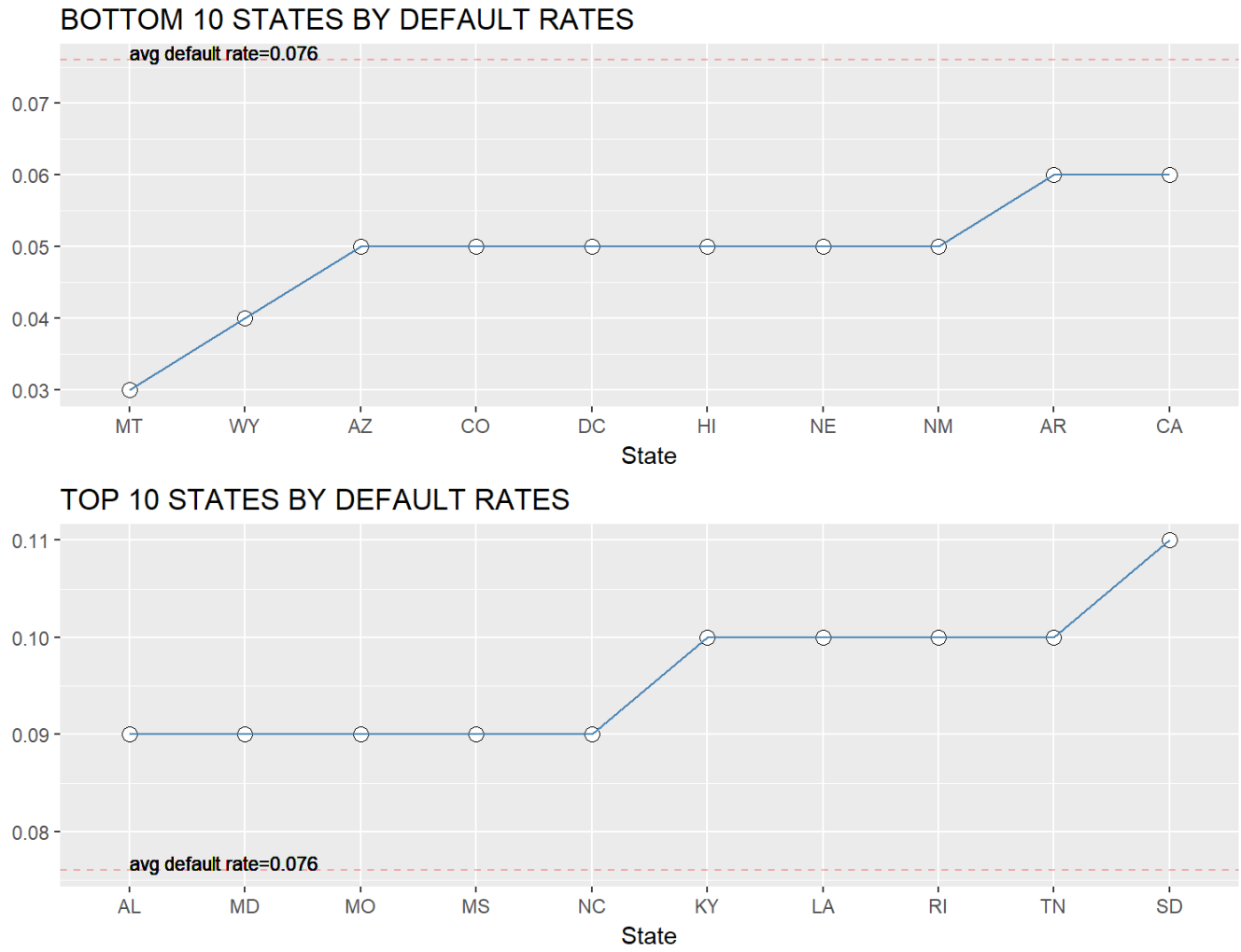
Description One

The chart shows the percentages from total loans of the defaulted loans, considering also the number of months since credit initiated until defaulted date and the Prosper score.The higher the score, the lower are the chances of a default. Only E and HR defaulted in the first 5 months since credit initiated. E and HR have also the highest default rates between 6 and 12 months. AA and B tend to default more

between 1 and 2 years ,while A,C and D have a constant default rate between 1 and 3 years. AA has a very low default rate of 0.01% and is present only in the group ‘between 1-2 years’.

Plot Two

TOP/BOTTOM 10 COUNTIES BY DEFAULT RATES



Description Two

The above charts try to find out which are the top 10 / bottom 10 countries with the highest/lowest default rates. The red line shows the average default rate across all states. The top country, with 0.11% defaulted loans was South Dakota, while the country with the lowest rate was Montana. This rate was calculated by taking the number of defaults by state and dividing it to the total loans. It would be interesting to also calculate the rate by divinding the defaults to the toatl loans for each state.

Plot Three

DEFAULTED VS COMPLETED BORROWERS COMPARISON: Avg borrower rate / Avg Stated Monthly Income / Avg Debt to Income ratio FOR EACH PROSPER SCORE

Description Three

Borrowers who defaulted had an average borrower rate greater than those who completed the payment. The avg rate is increasing with each score, putting more pressure on the borrowers who were classified in a low score category and making it harder to pay the loan.

Those who defaulted had an avg monthly income lower than those who completed the payment across all score categories. This two facors make the avg debt to income ratio to increase by score and to have higher values for borrowers who defaulted compared with those who didn't. So a smaller income and a high borrower rate have an important impact on the success of completing the payment or not.

Reflection

The difficulties I encounter with the dataset were mainly related to the financial terms. I didn't know anything about peer-to-peer lending company, in Romania the legislation doesn't allow such companies to exist.

After an initial quick summary run on the dataset, I noticed many missings, inexplicable values of 0 and an inconsistency in the Prosper Score categories. Exploring the data, I removed the duplicates and decided to keep only the loans from July 2009 to March 2014 which solved many of the issues.

I started the EDA by running some simple plots to make sense of the data and to find a possible scenario to present.

But every new plot rised new questions and each question could be answered in many ways. I think I spent most of my time searching for a story to tell. After I did few plots to reveal the loans trend as values and numbers, I become curious about one thing: which variables influence the prosper score? Seeing the behaviour of each category in prosper score, the next curiosity was: which of these borrowers will not complete their payment and why? Answering this questions helped me to develop a logistic

regression to predict risky borrowers. The model can be improved,but I learned a lot about a simple machine learning algorithm and it was a great experience.