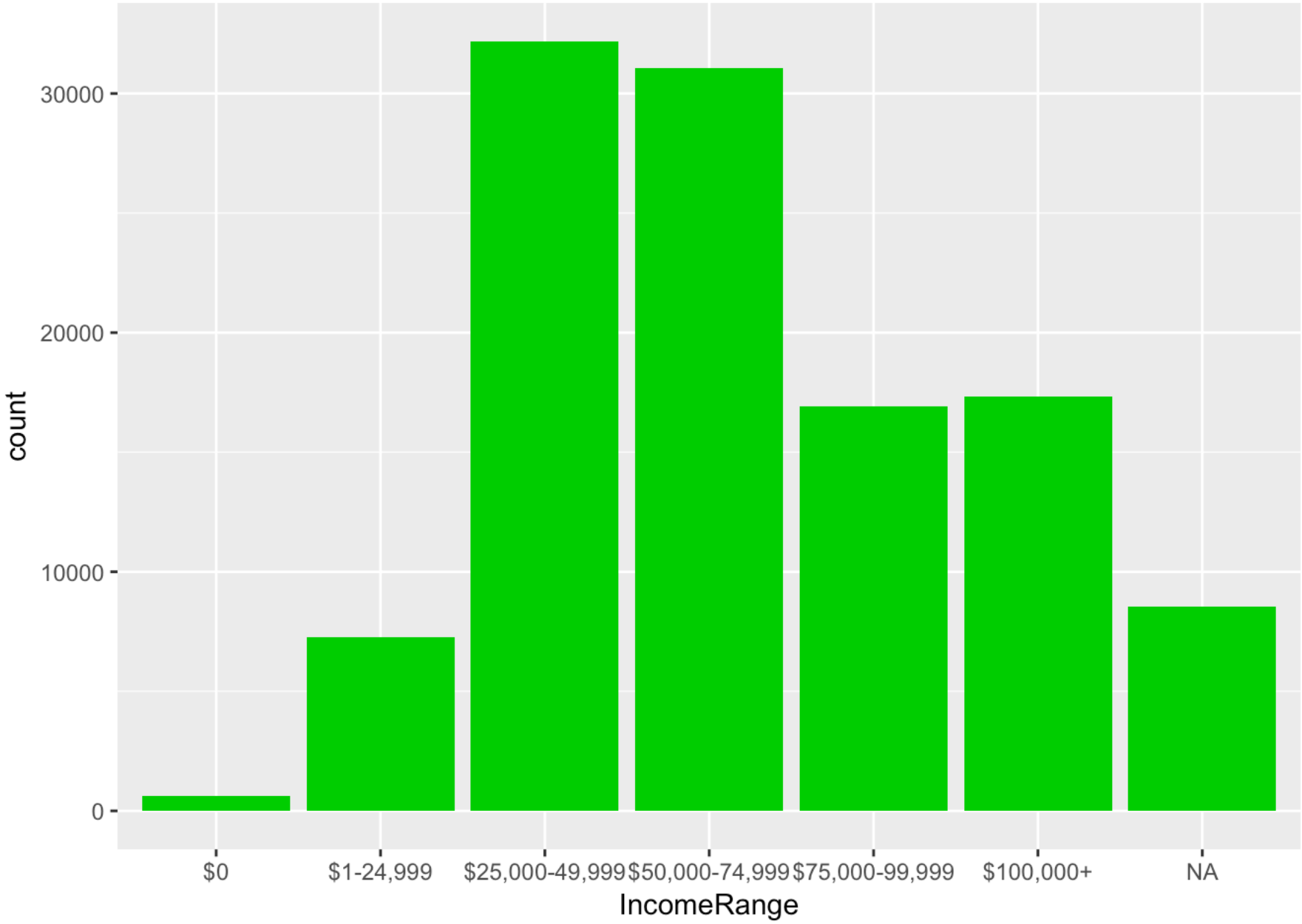
Analysis of Prosper Loan Data

Overview

Prosper Marketplace is America's first peer-to-peer lending marketplace, with over $7 billion in funded loans. Borrowers request personal loans on Prosper and investors (individual or institutional) can fund anywhere from $2,000 to $35,000 per loan request. In addition to credit scores, ratings, and histories, investors can consider borrowers' personal loan descriptions, endorsements from friends, and community affiliations. Prosper handles the servicing of the loan and collects and distributes borrower payments and interest back to the loan investors.[https://en.wikipedia.org/wiki/Prosper_Marketplace (https://en.wikipedia.org/wiki/Prosper_Marketplace)].Here I used the data available to the public from the institution and analyse the borrower market (including the demographic segmentation and beyond) and try to let data tell some 'behind scene' stories about the borrowers.

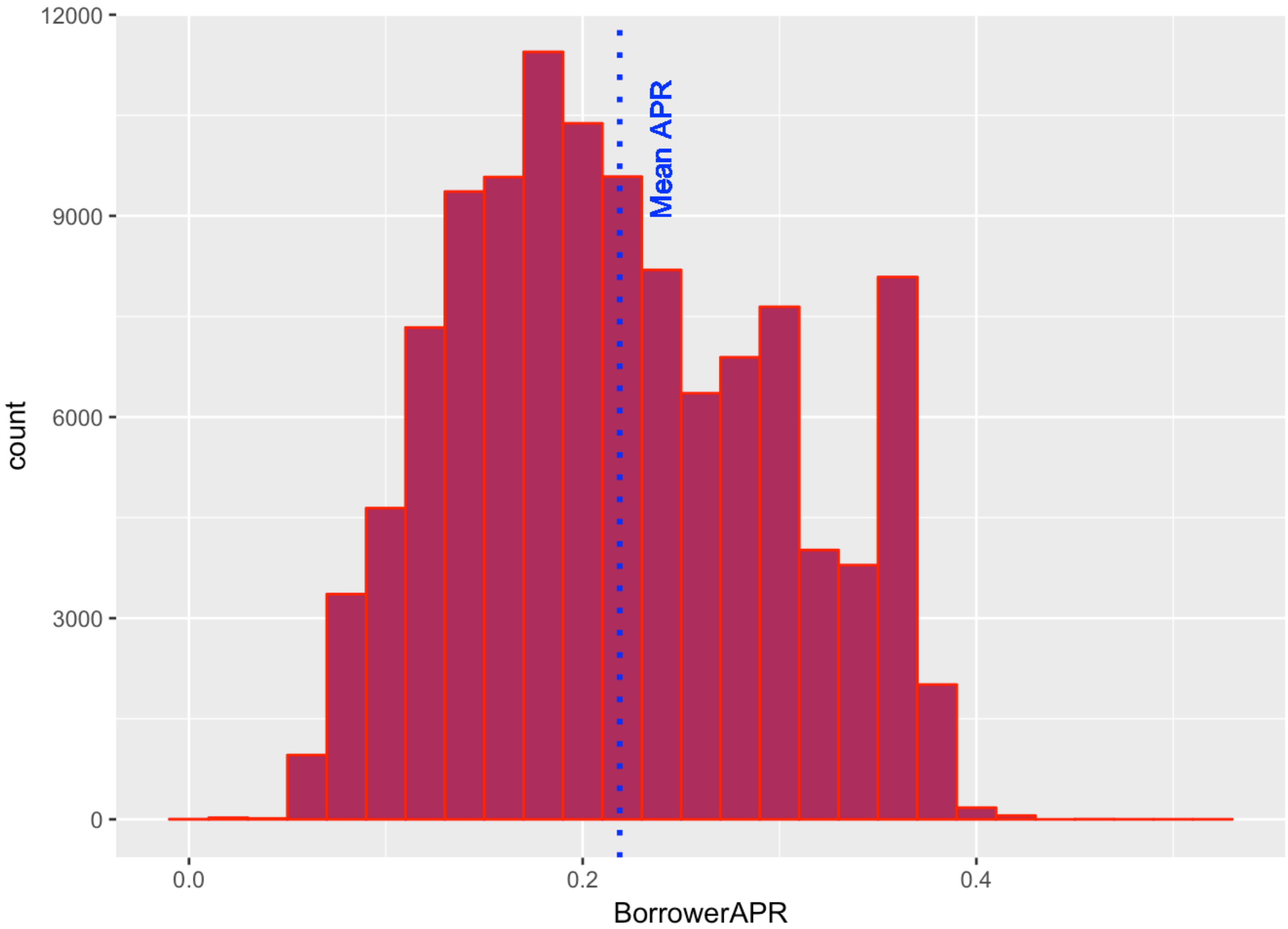
UNIVARIAT ANALYSIS

Income distribution of borrowers



The top borrowers are customers that have income ranges from $25,000-$49,999 followed by those with income range of $50,000 - $74,999 and there are fewer zero income borrowers as well.

APR distribution

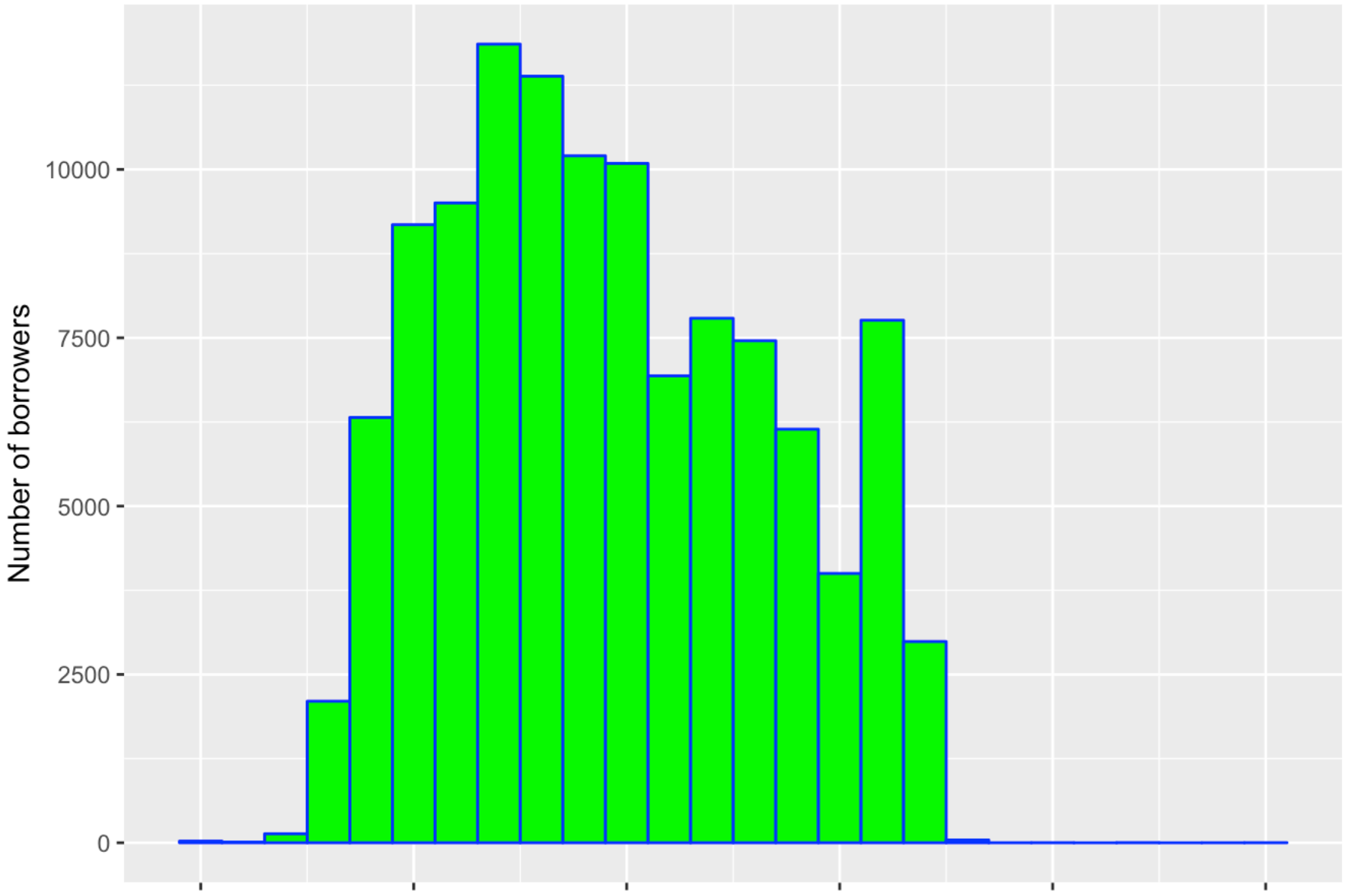



```
## [1] "The APR summary statistics is :"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## 0.00653 0.15630 0.20980 0.21880 0.28380 0.51230      25
```

The loan APR ranges from less than 5 % up to about 51 % though there are only very few borrowers with such high and lower APR values. Most borrowers seems to have a APR from 10 % upto 30 % with somehow a substantial amount of borrowers found to have APR value of ~ 36 %. The median APR value is ~21 %.

Borrower Interest rate distribution

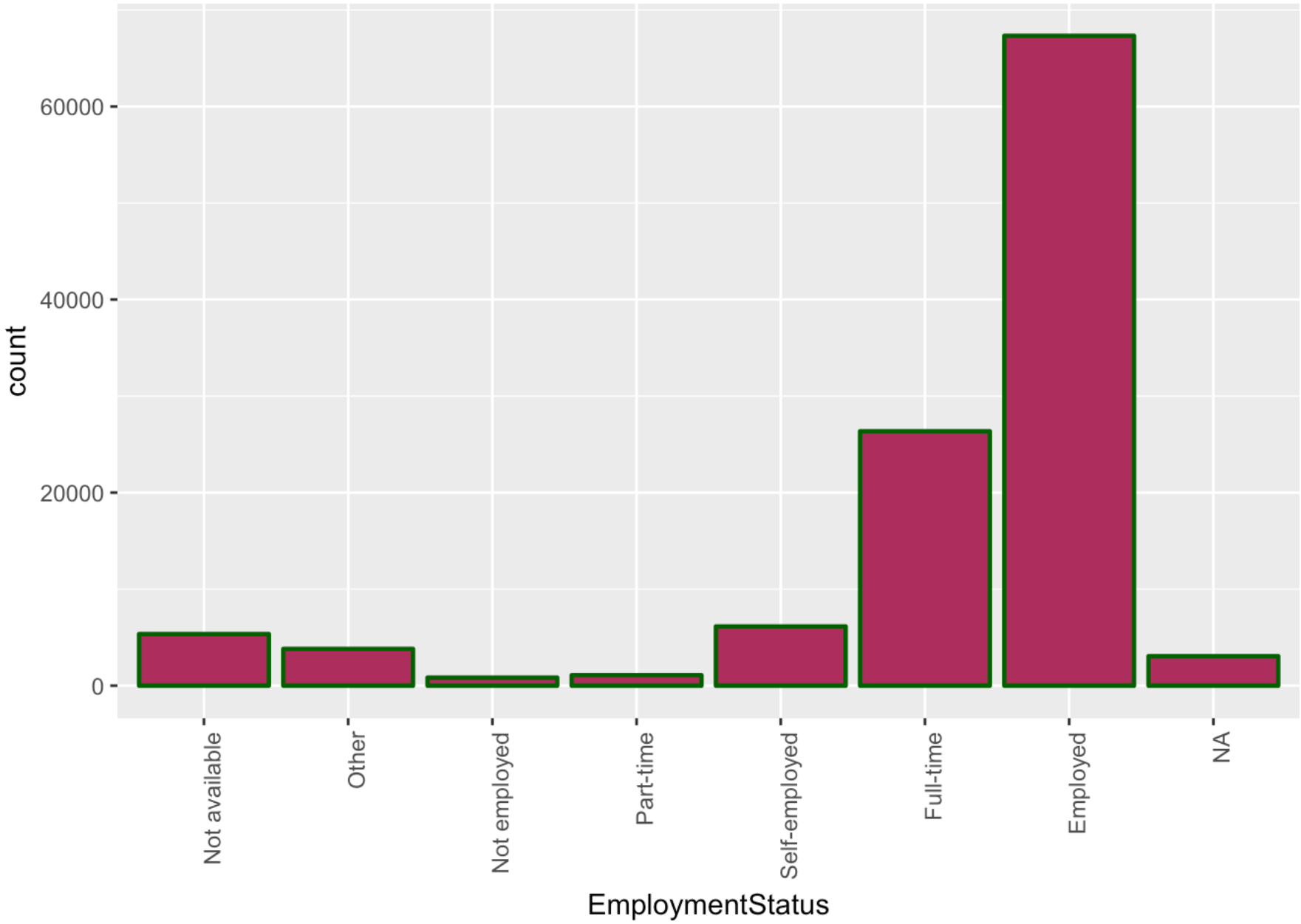



```
## [1] "The interest rate summary statistics is :"
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.1340  0.1840  0.1928  0.2500  0.4975
```

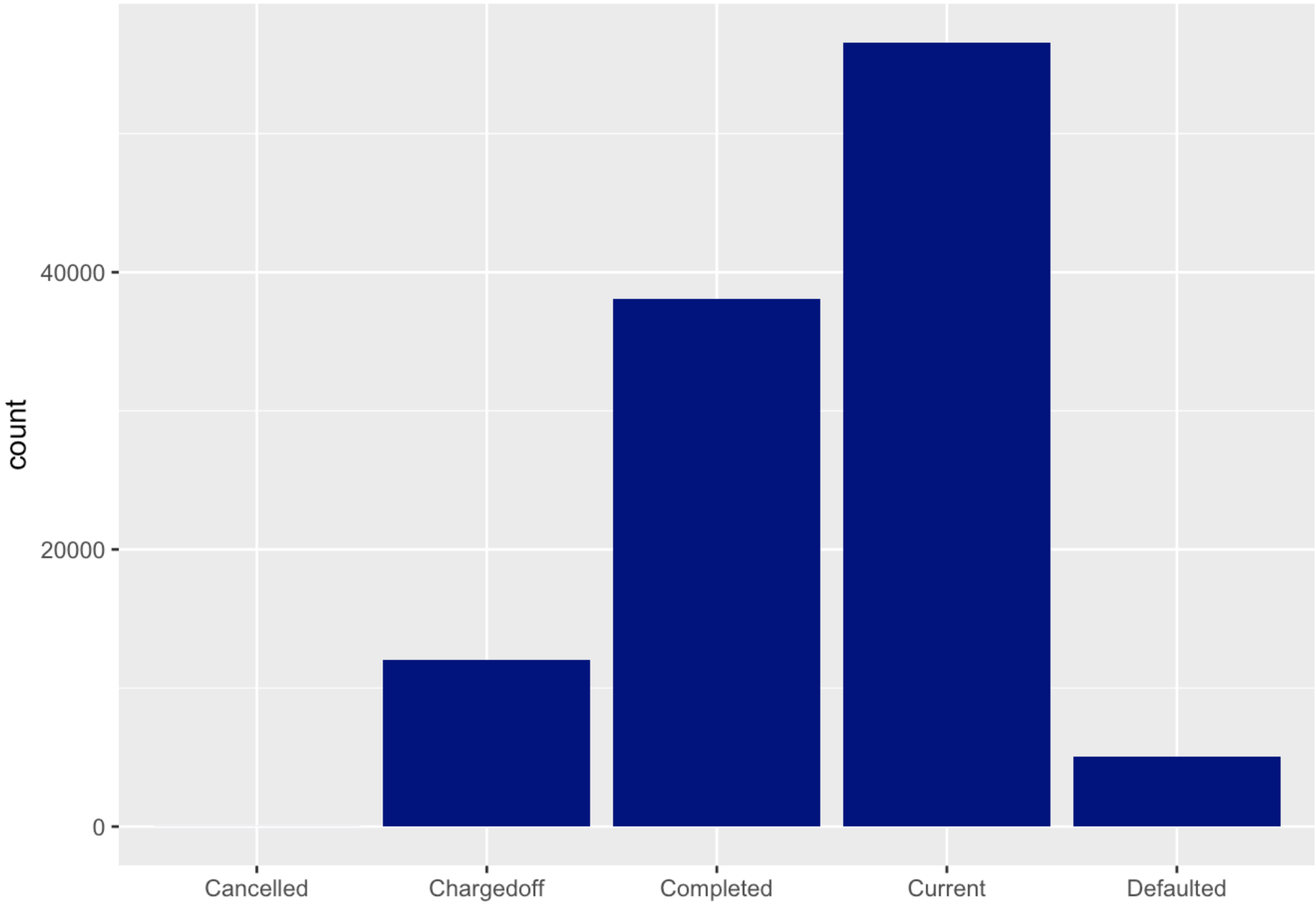
The borrower rate has roughly the same distribution as the borrower APR. However, the maximum rate is less than 50 %.

Employment status



Most borrowers are employed and also there are few unemployed borrowers. It also shows full time employees have either higher chance of getting prosper loan or most of the loan applicants are under this category. On the other hand, there are few unemployed and Part time borrowers.This could be, due to their lower chance of getting approved to the loan because of their income or or there are fewer number of loan applicants under these categories.

Loan status

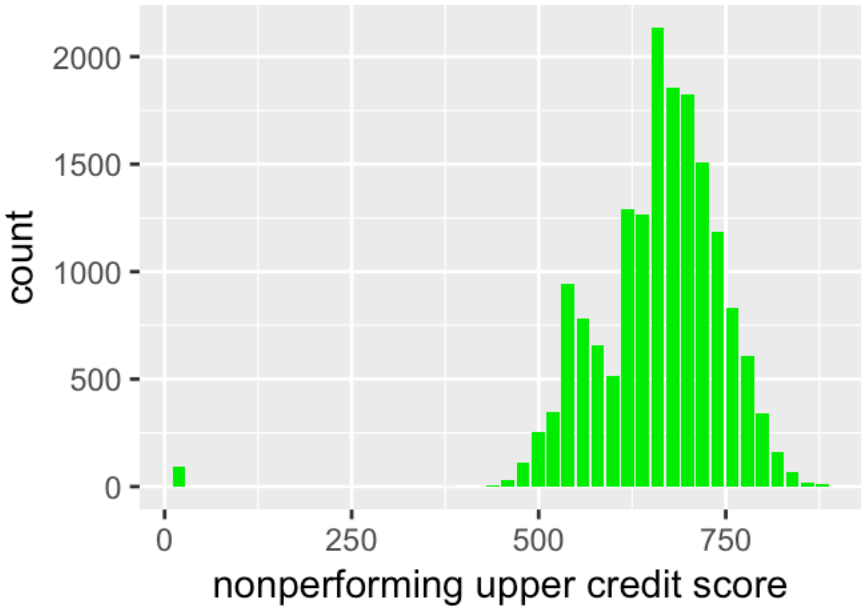


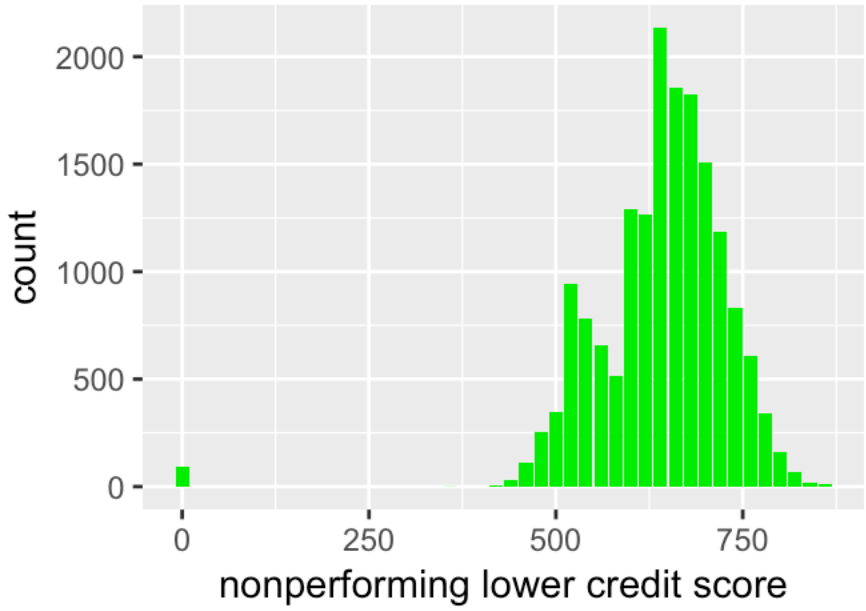

```
##   LoanStatus      Percent
## 1  Cancelled  0.004477679
## 2 Chargedoff 10.739264765
## 3  Completed 34.096628308
## 4    Current 50.665830833
## 5  Defaulted  4.493798415
```

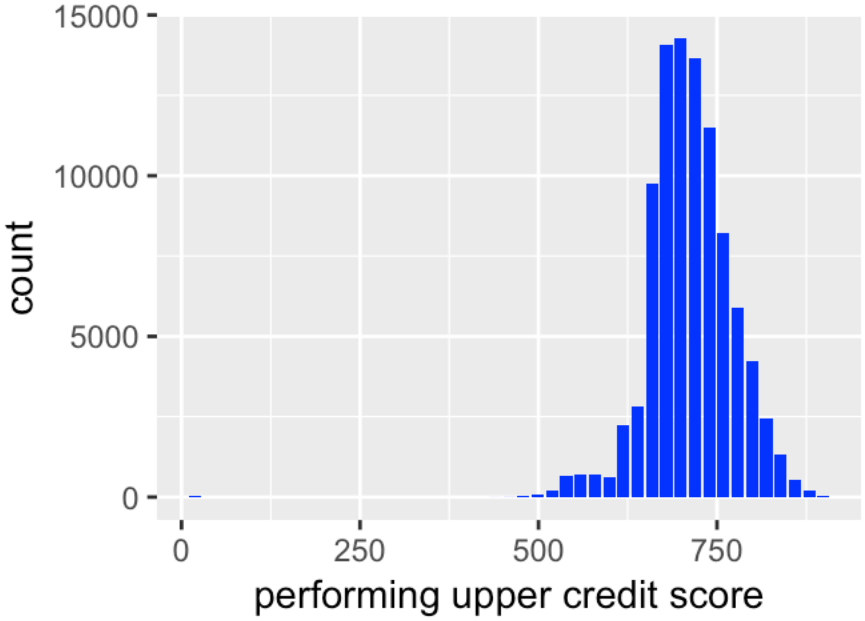
Even though most of the loan is active and completed, about 10.739265 % of the total loan is Chargedoff and 4.493798 % is defaulted.

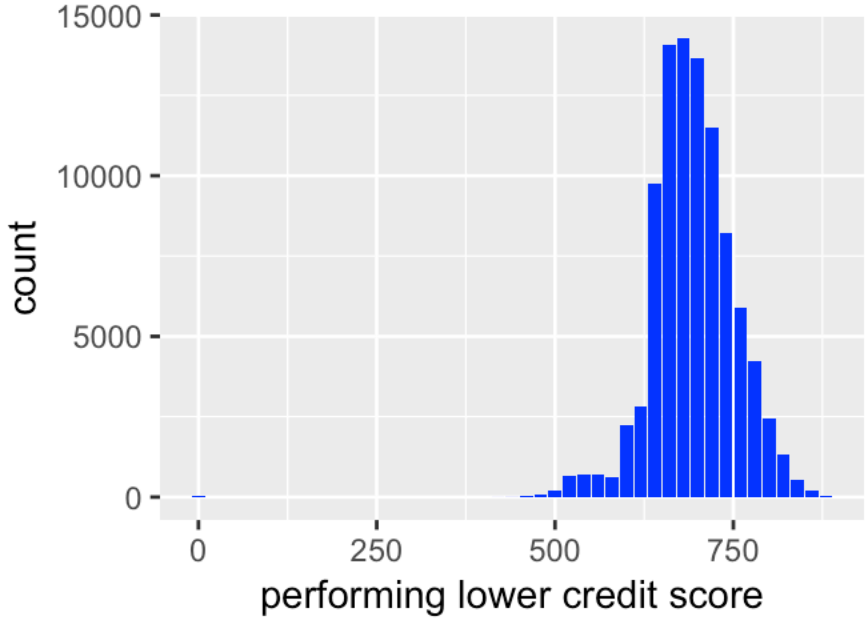
Loan Performance by credit score

I categorized Chargedoff and Defaulted loans as non performing loans for simplicity of analysis.



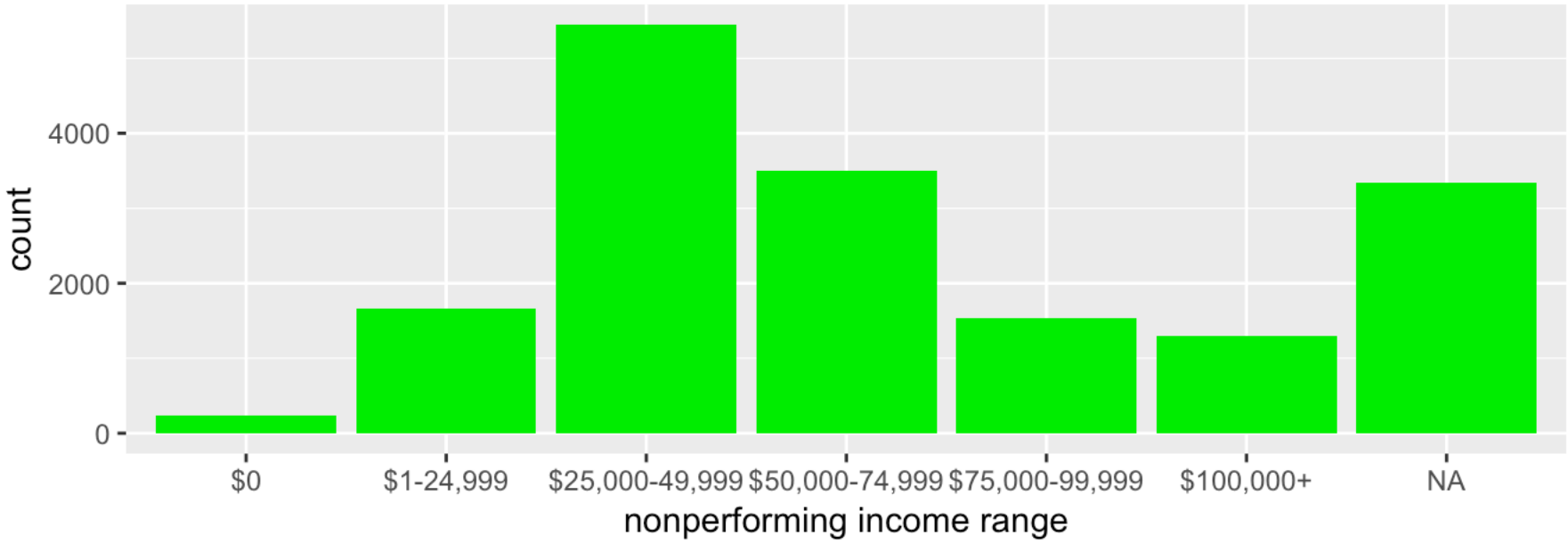


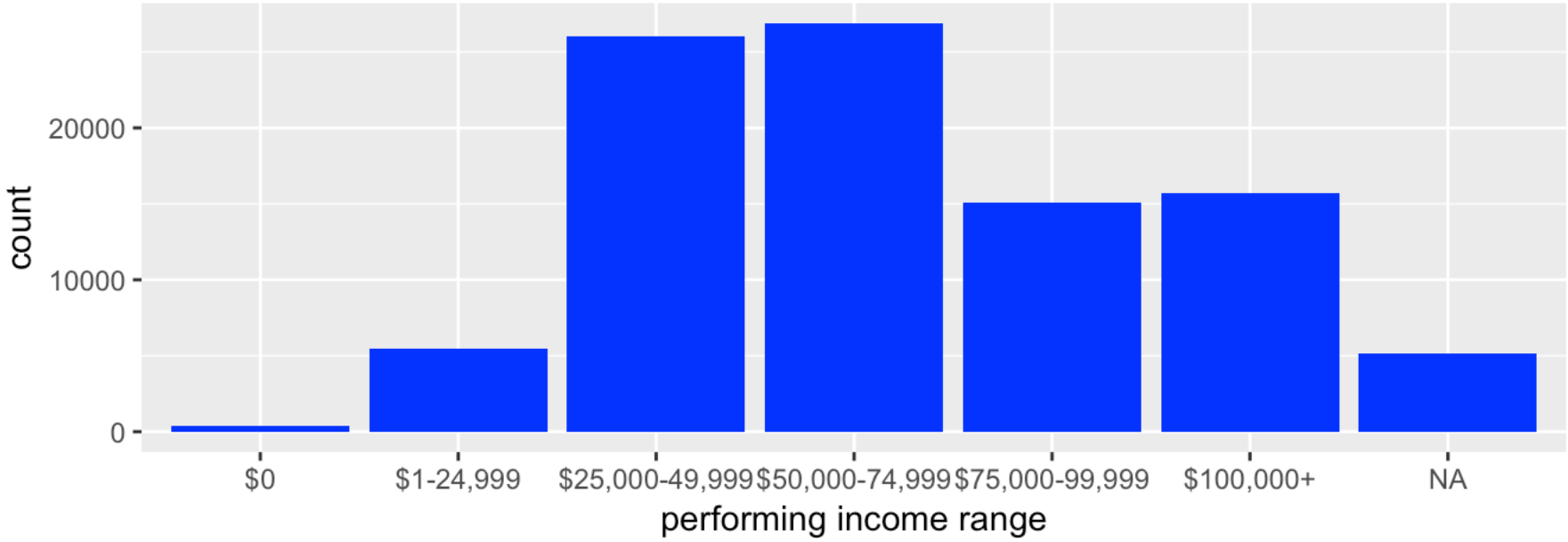




From the distributions of credit score we can see that, there are non performing borrowers across all credit scores from high to low. However, there are many borrowers with lower credit scores (< ~ 600) fall under the credit score distribution of non performing borrowers and relatively fewer borrowers with lower credit scores (< ~ 600) fall under the performing distribution.

Loan Performance by Income range



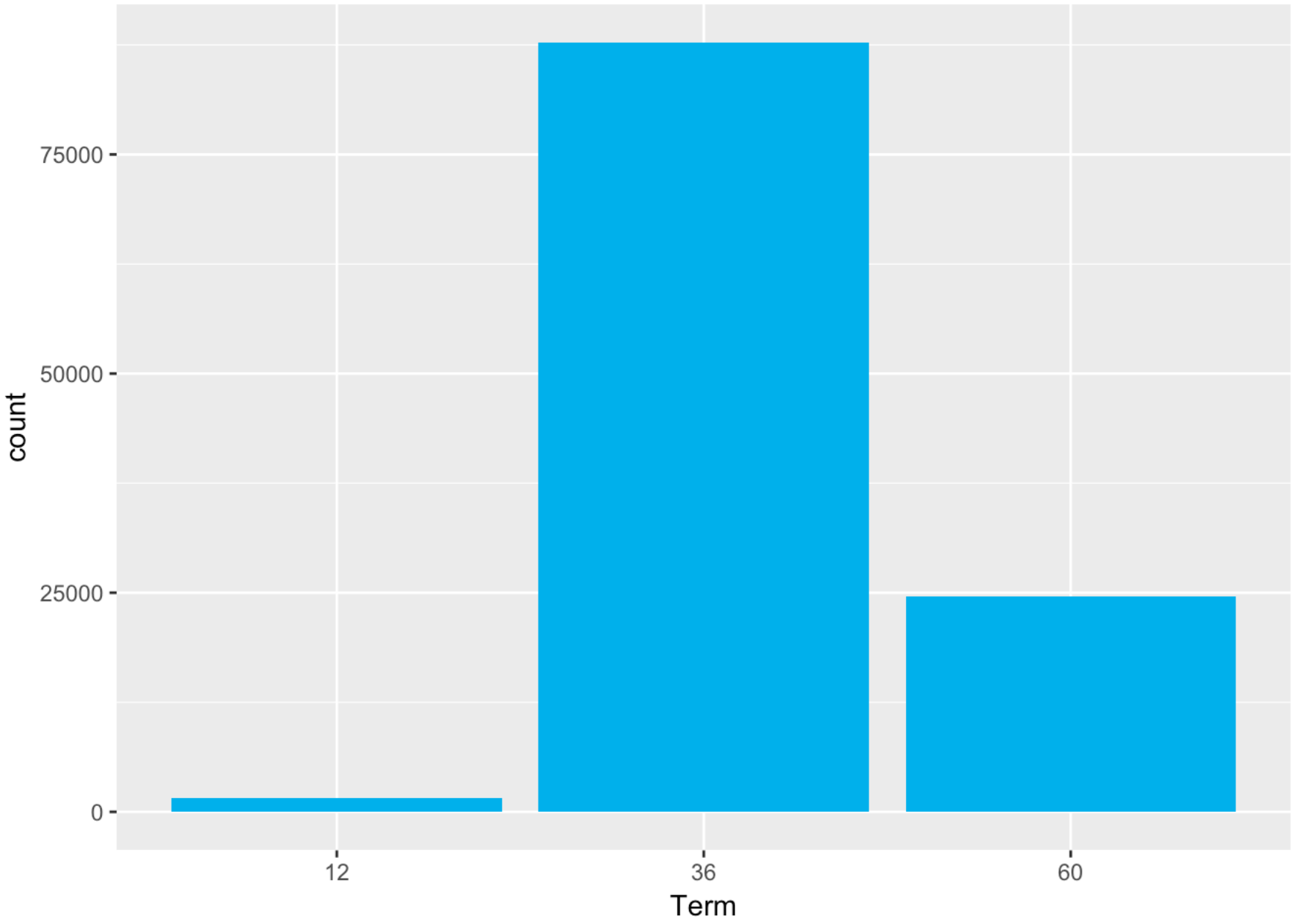



```
## NULL
```

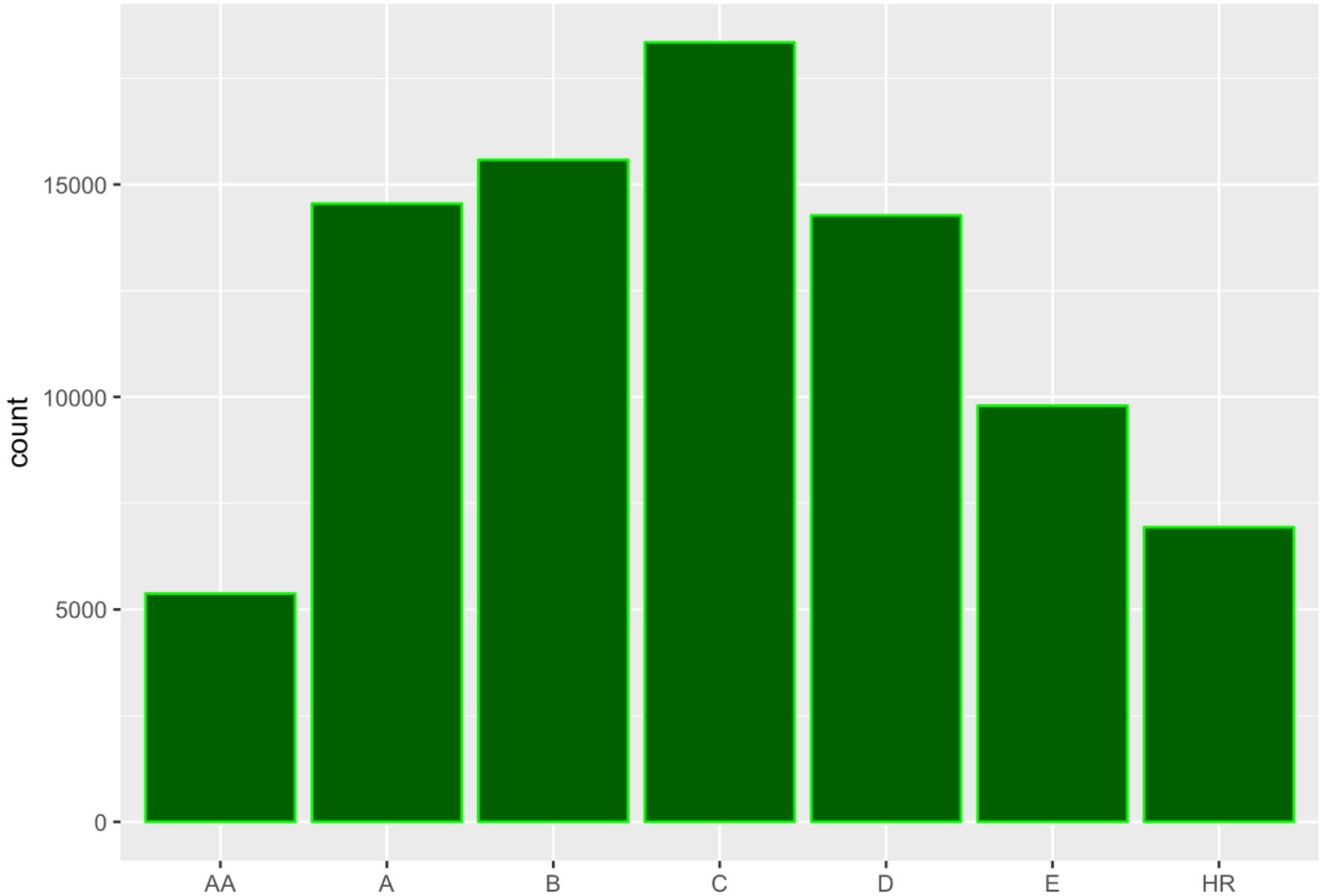
```
##      IncomeRange PerformingCount NonPerformingCount
## 1             $0              381                 236
## 2       $1-24,999             5444                1663
## 3 $25,000-49,999            26003                5452
## 4 $50,000-74,999            26897                3507
## 5 $75,000-99,999            15053                1528
## 6       $100,000+            15690                1290
##   NonPerformingPercentage
## 1                   38.25
## 2                   23.40
## 3                   17.33
## 4                   11.53
## 5                    9.22
## 6                    7.60
```

From this chart we can see an interesting trend of increasing performance with increasing income range. This could be explained by, borrowers with good incomes are more likely to pay their loans off. It is interesting to find that borrowers with no income are the most non performing borrowers (38.25 %) followed by borrowers with income ranges from $1-$24,999 (23.4 %) compared to all borrowers with registered income ranges.

Loan duration

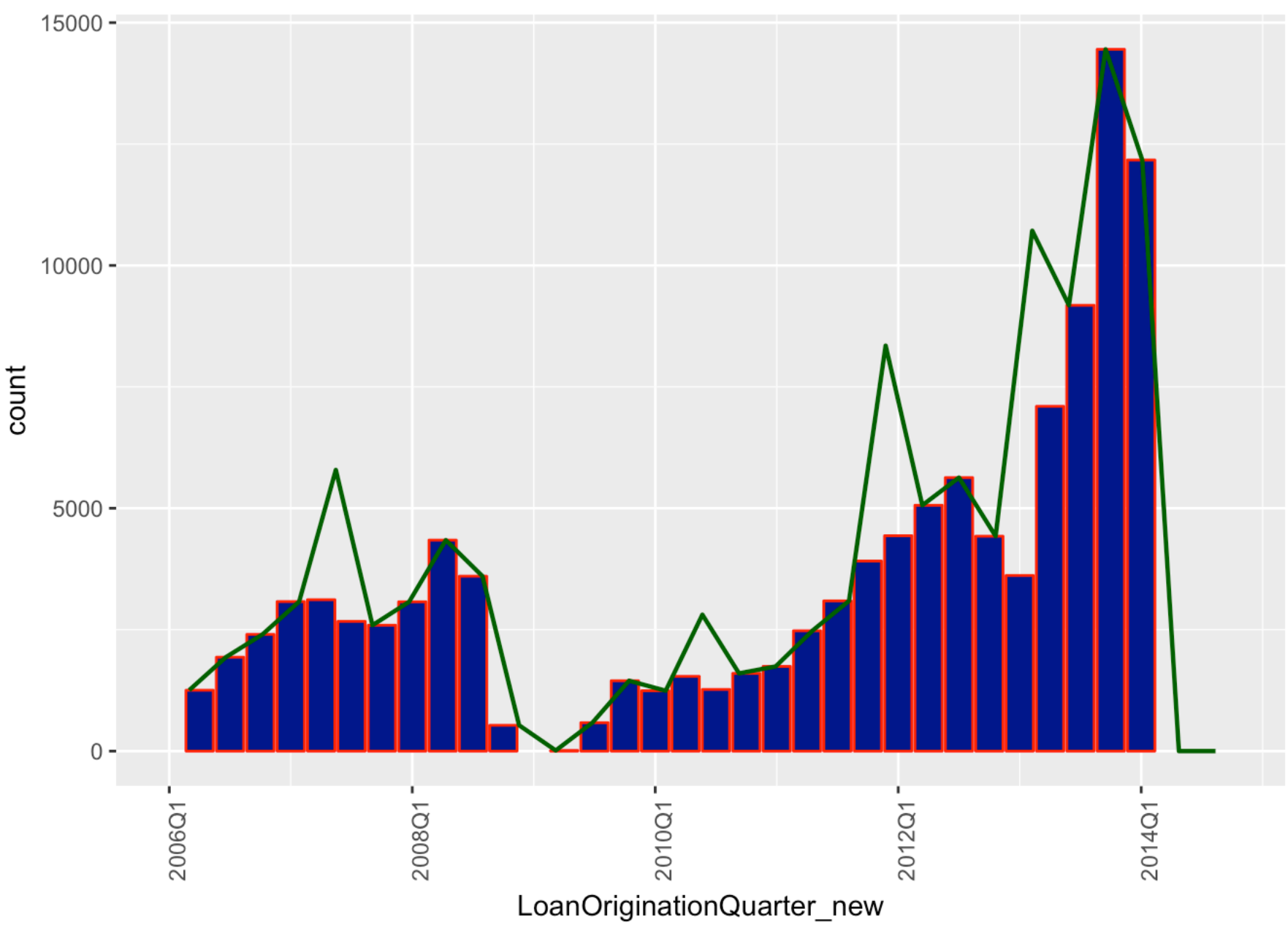


It is clear that Most of the borrowers are under the 3 year loan period and relatively small number of borrowers borrowed for the 12 month period.



The Prosper rating distribution is almost normal with C the most freqent rating and AA the least frequent rating.

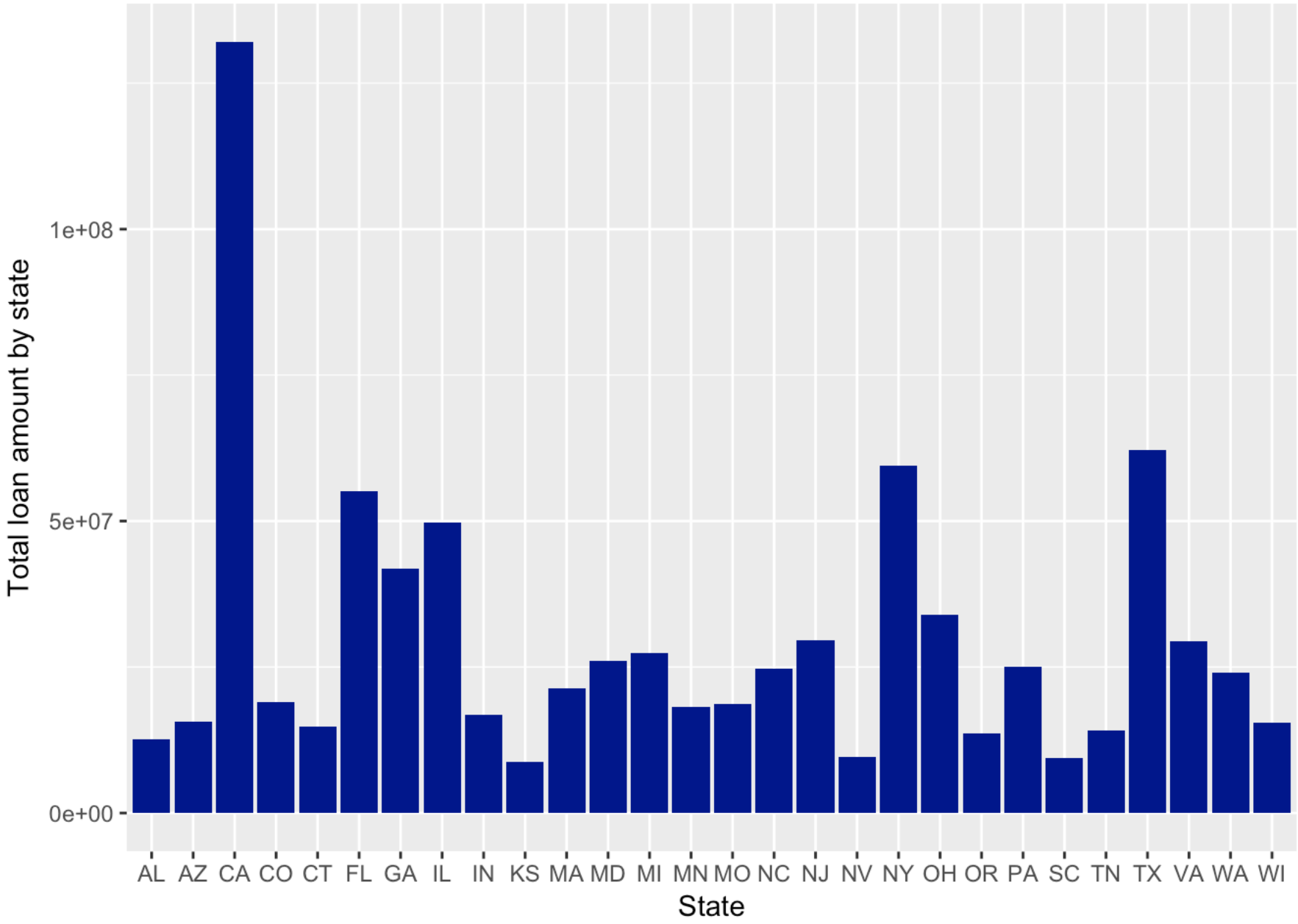
Loan origination by quarter year

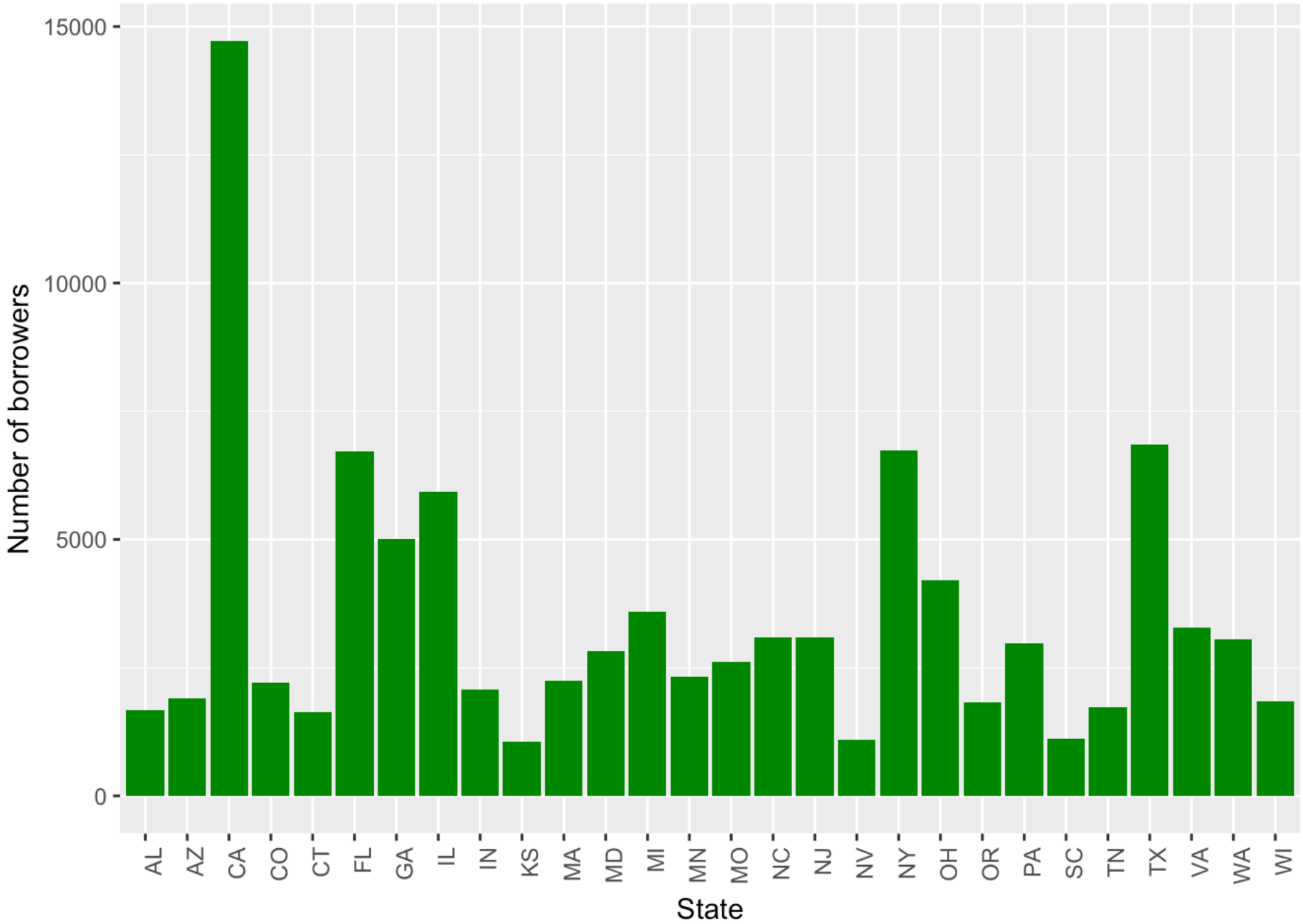


There is abig dip in listings from 2008 Q4 into 2009. This time period coincides with the collapse of Lehman brothers and the fallout of the global financial system.

BIVARIAT ANALYSIS

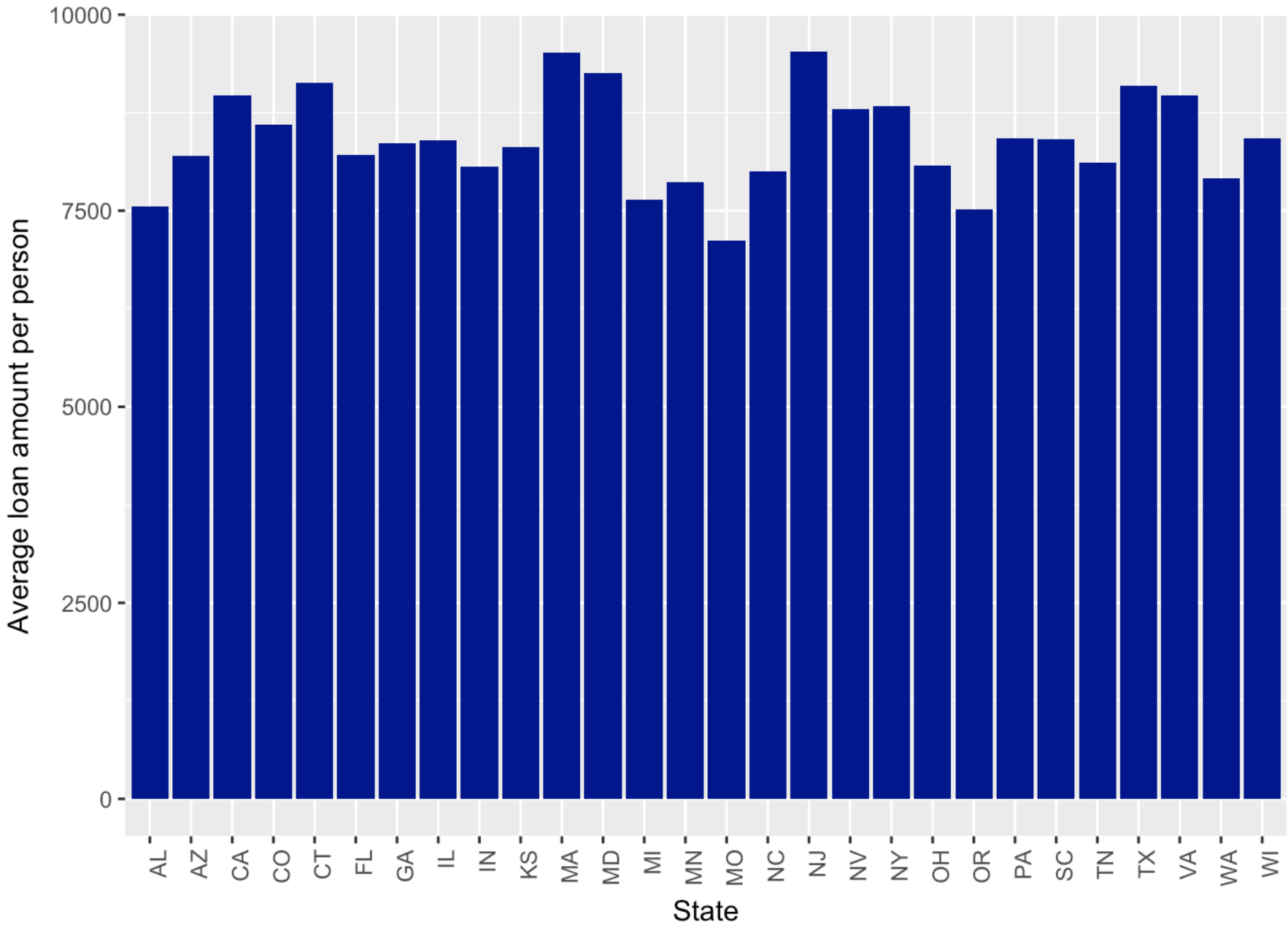
Total loan amount by state where number of borrowers >1000


Total loan amount by state
1e+08
5e+07
0e+00
AL AZ CA CO CT FL GA IL IN KS MA MD MI MN MO NC NJ NV NY OH OR PA SC TN TX VA WA WI
State



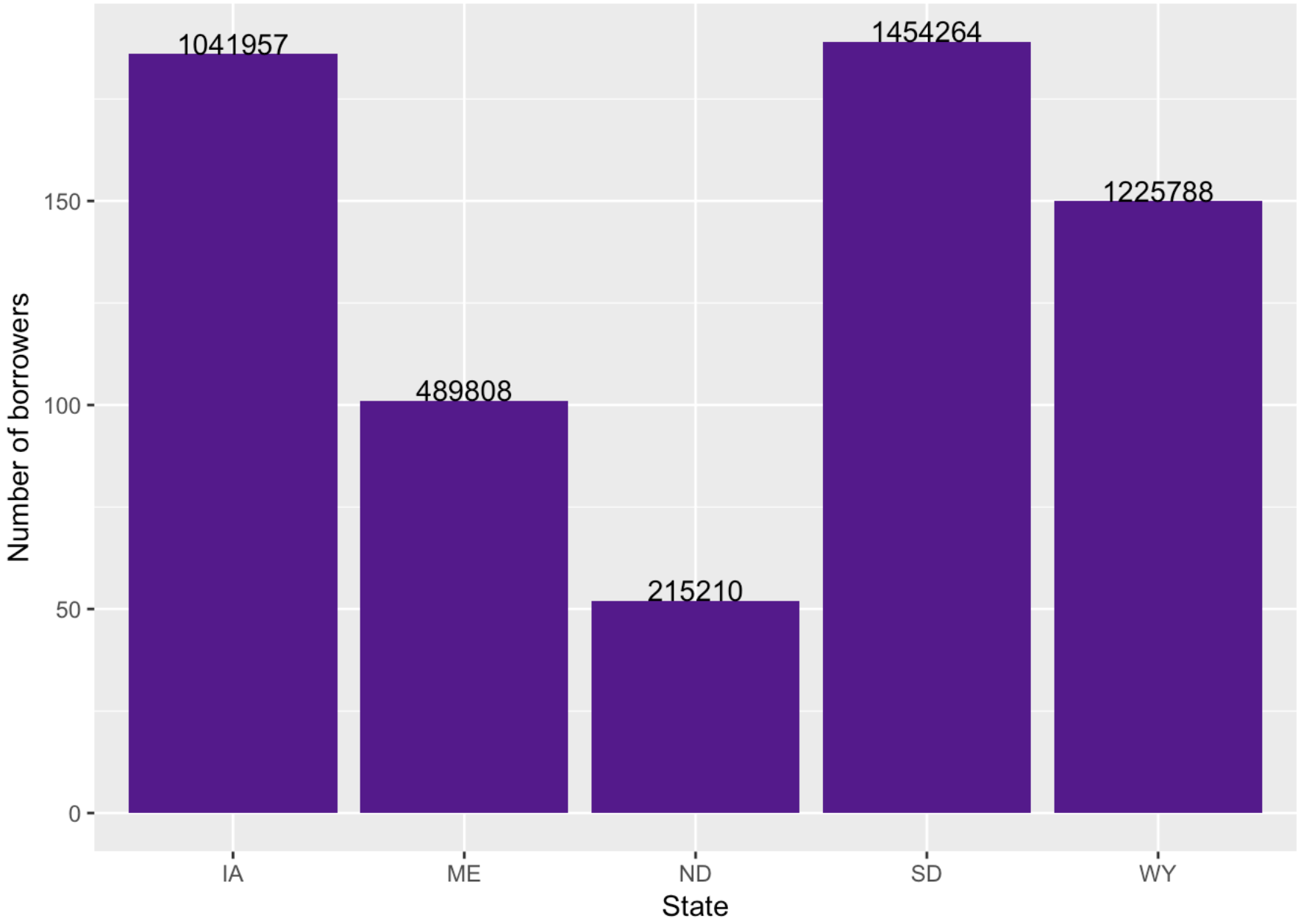
I categorized the states as high loan and low loan states by number of borrowers (number of borrowers in thestate >1000 and number of borrowers in the state < 200). The above two distributions show the total loan amount and number of borrowers of the high loan states. As we can see from the histograms, CA,TX,NY and FL are the top states by total loan amount. This is consistent with the number of borrowers and I found it is related to the population size of the states. In order to check whether this high total loan amount is related to the population of the states I normalized the total loan amount by the total number of borrowers as the average loan amount per person as shown below.

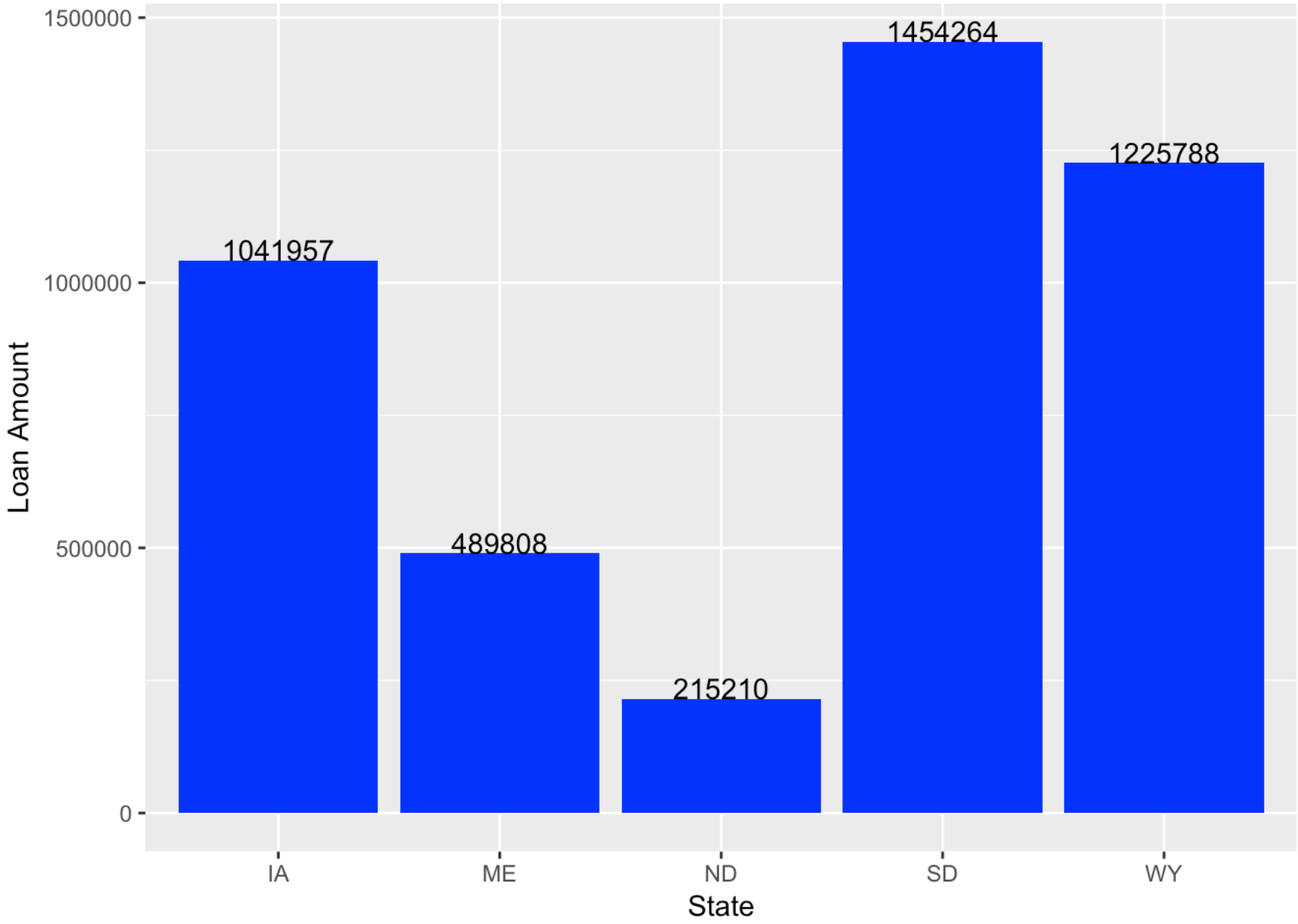
Average loan amount per person

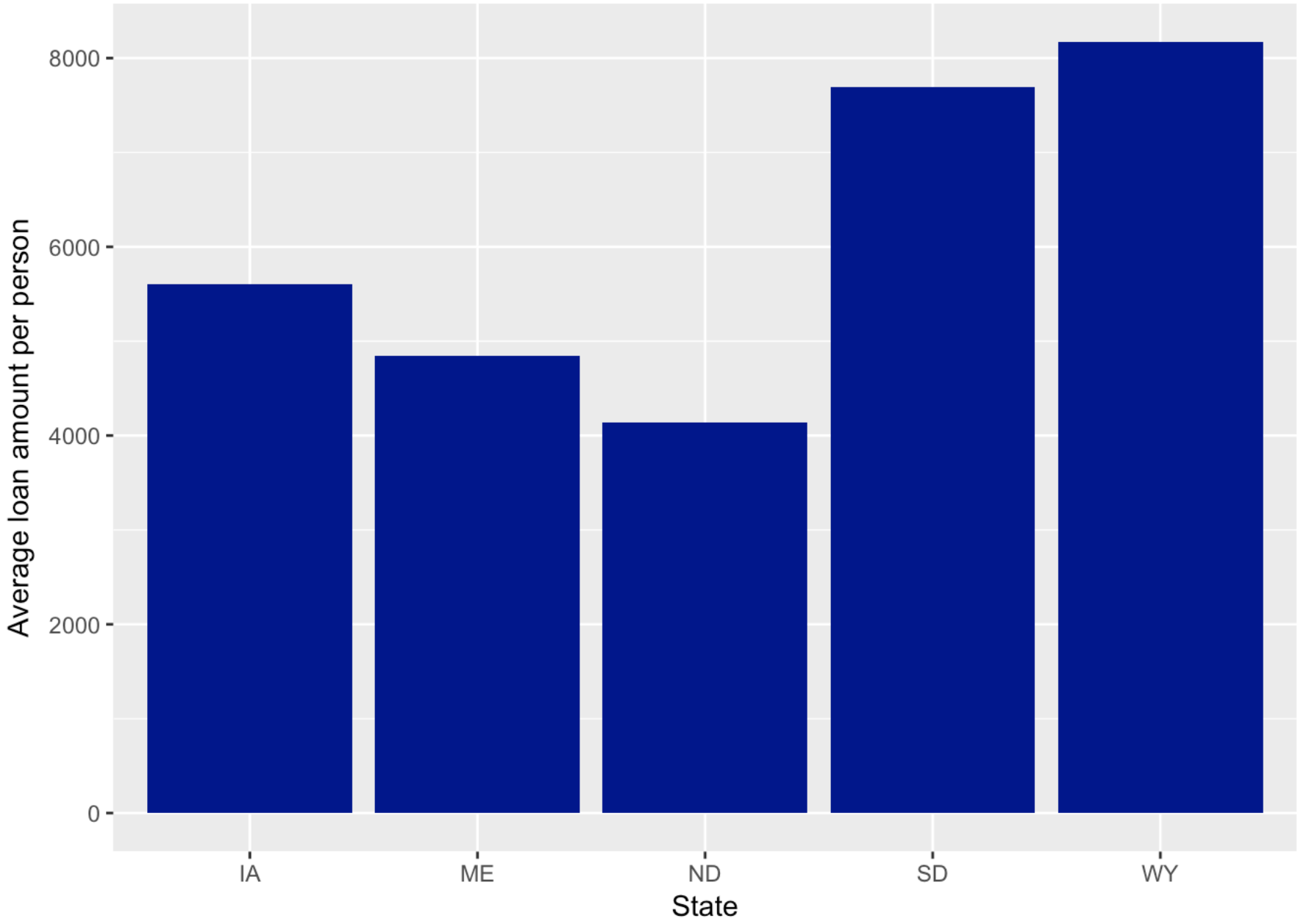


From the above histogram, we can see that the average loan per person distribution is almost kind of flat and interestingly the normalized loan amount of CA, TX, NY, and FL is less than the the normalized loan amount of some states like NJ and MA. Therefore the higher total loan amount in larger states is related to the size of their populations.

States with least number of borroers (< 200)

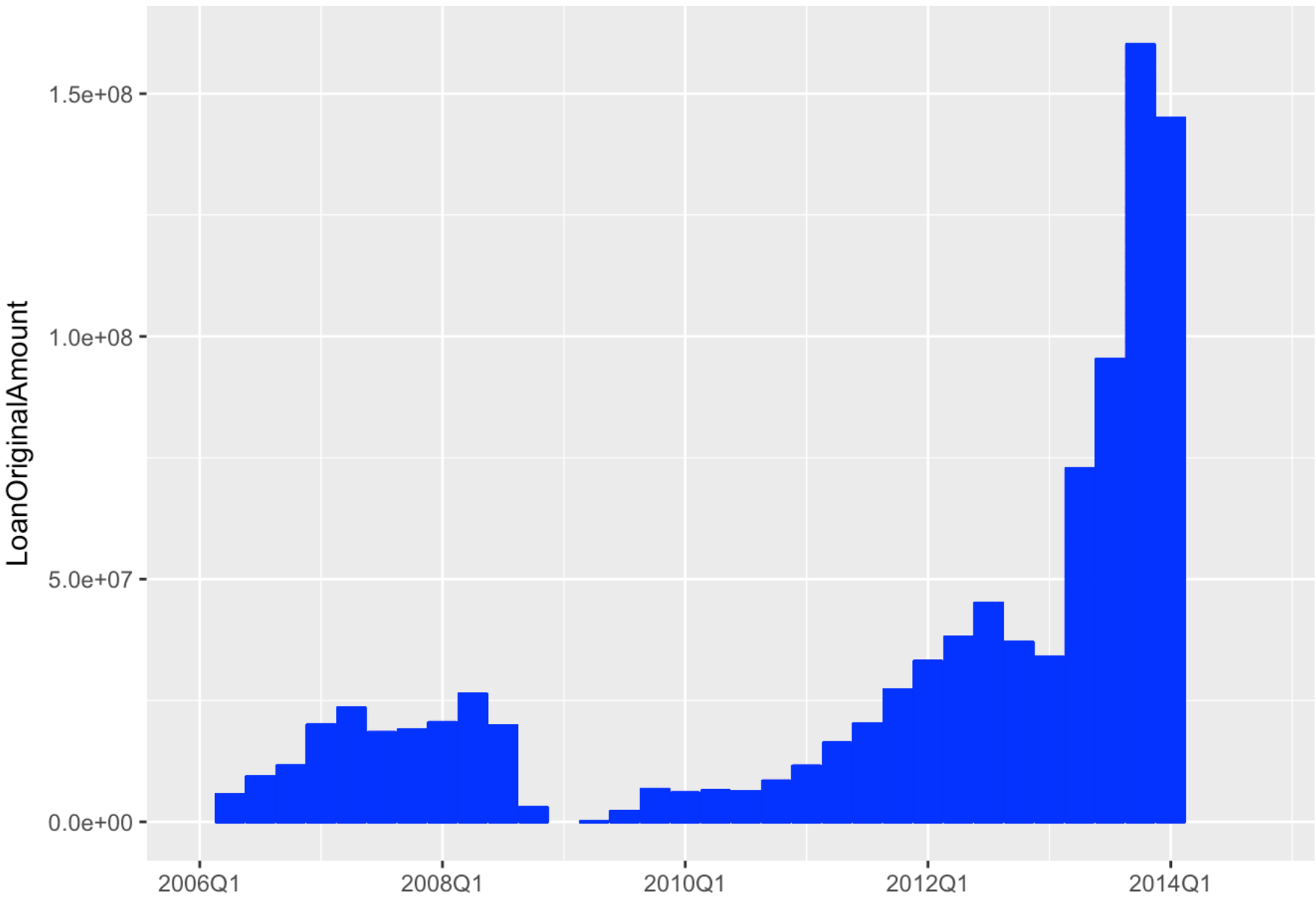

1041957
489808
215210
1454264
1225788
150
100
50
0
Number of borrowers
IA
ME
ND
SD
WY
State


1500000
1000000
500000
0
Loan Amount
1041957
489808
215210
1454264
1225788
IA
ME
ND
SD
WY
State



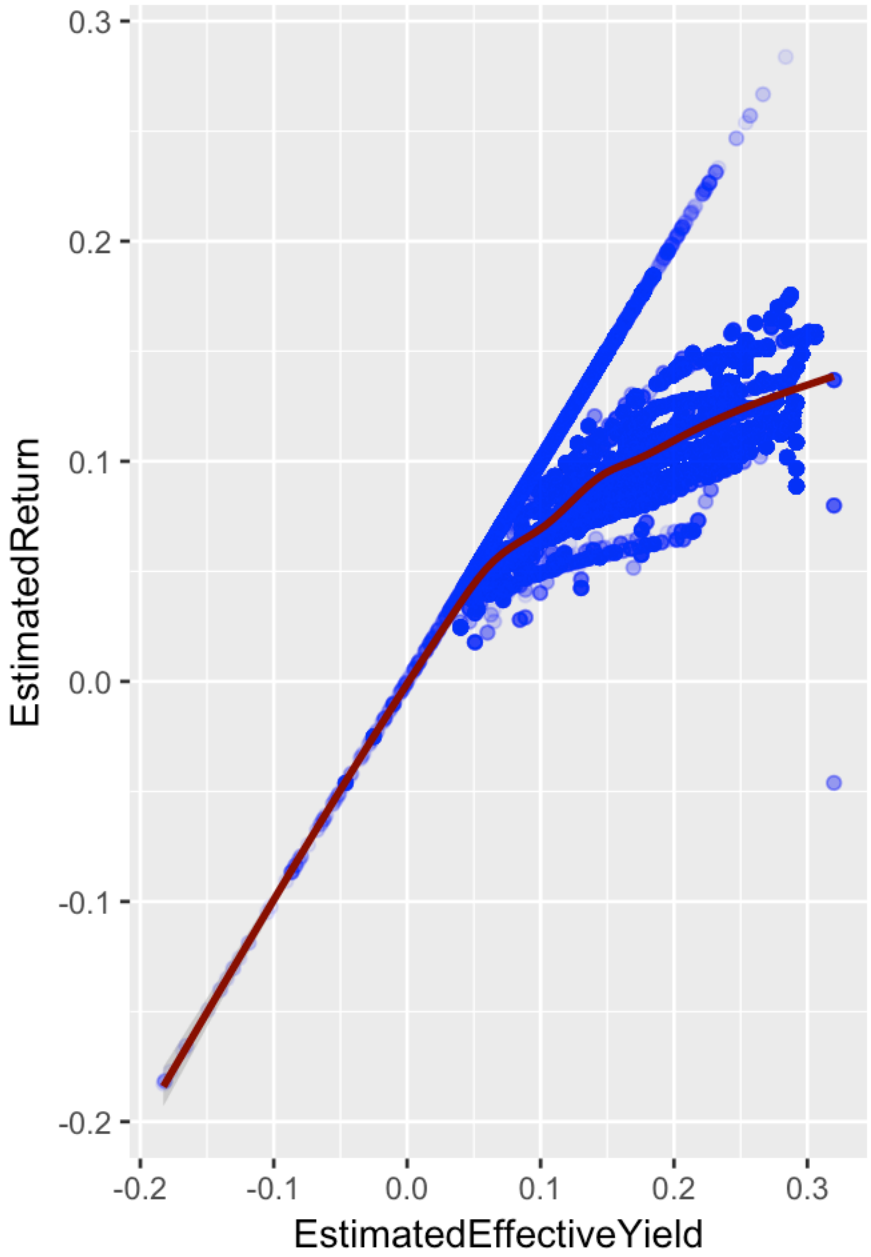
It is interesting that 5 states (IA, ME,ND, SD, WY) have the least number of borrowers.

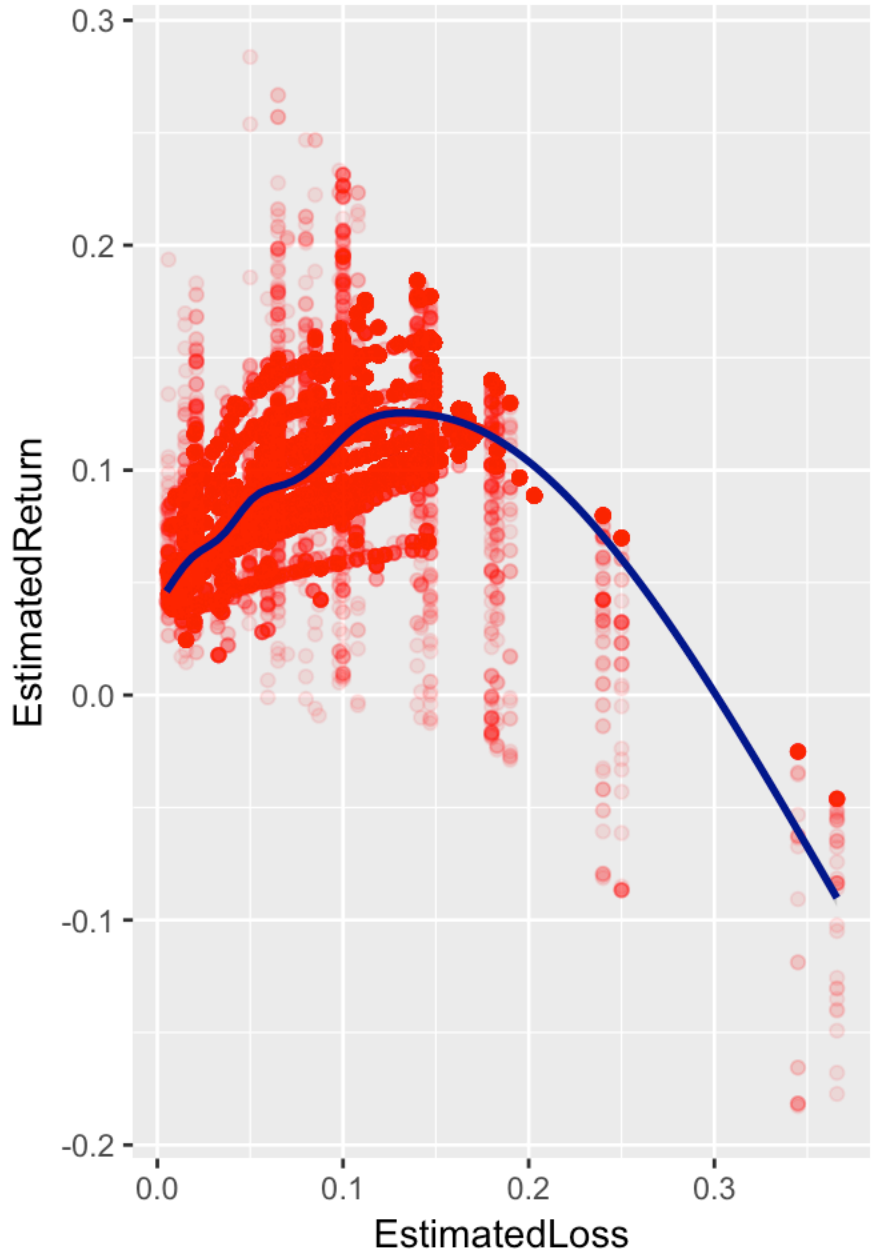
Loan Amount by quarter year



The loan amount shows generally an increasing pattern especially since the fourth quarter of 2009.

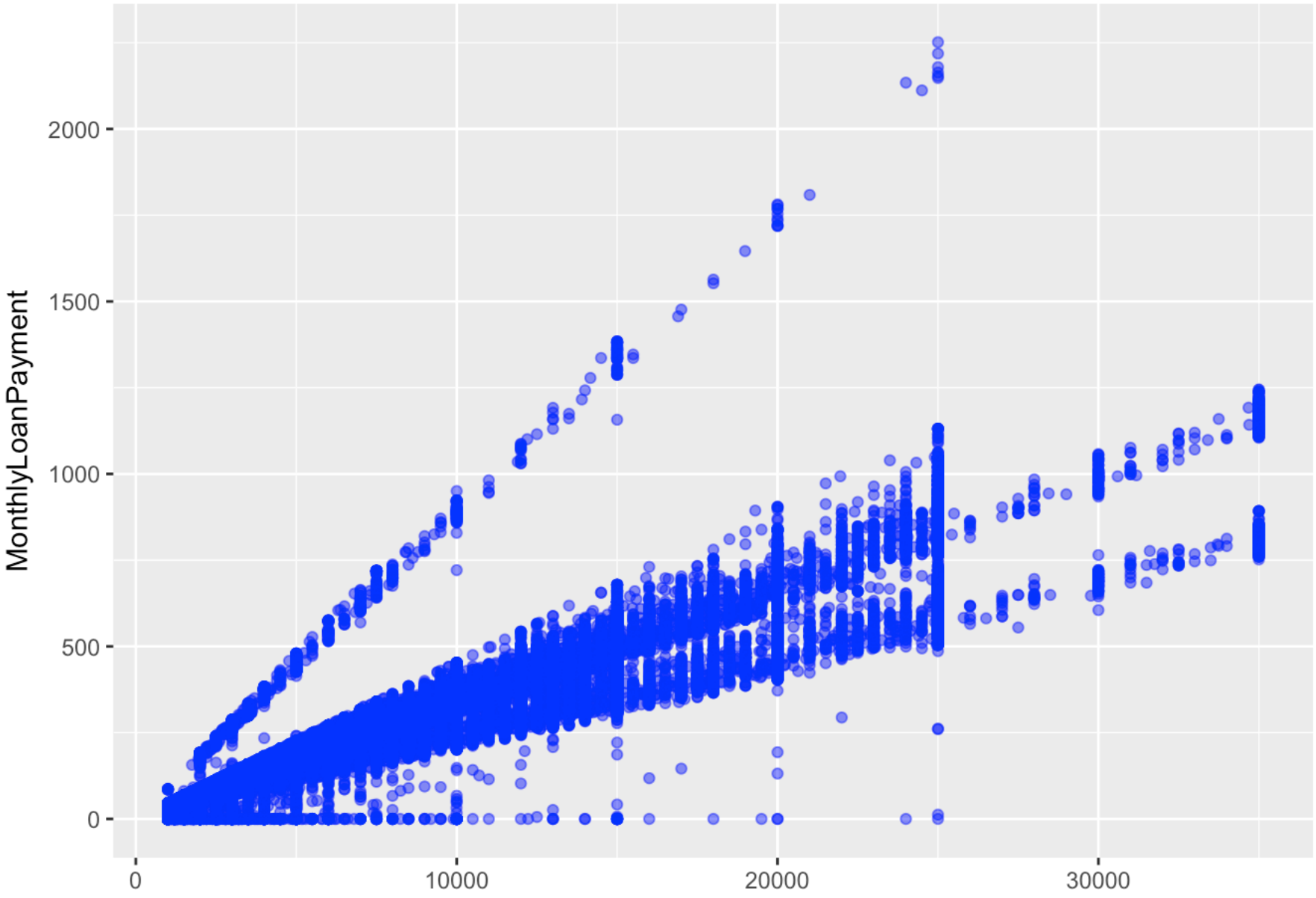
Estimated Effective Yield, EstimatedReturn and Estimated Loss





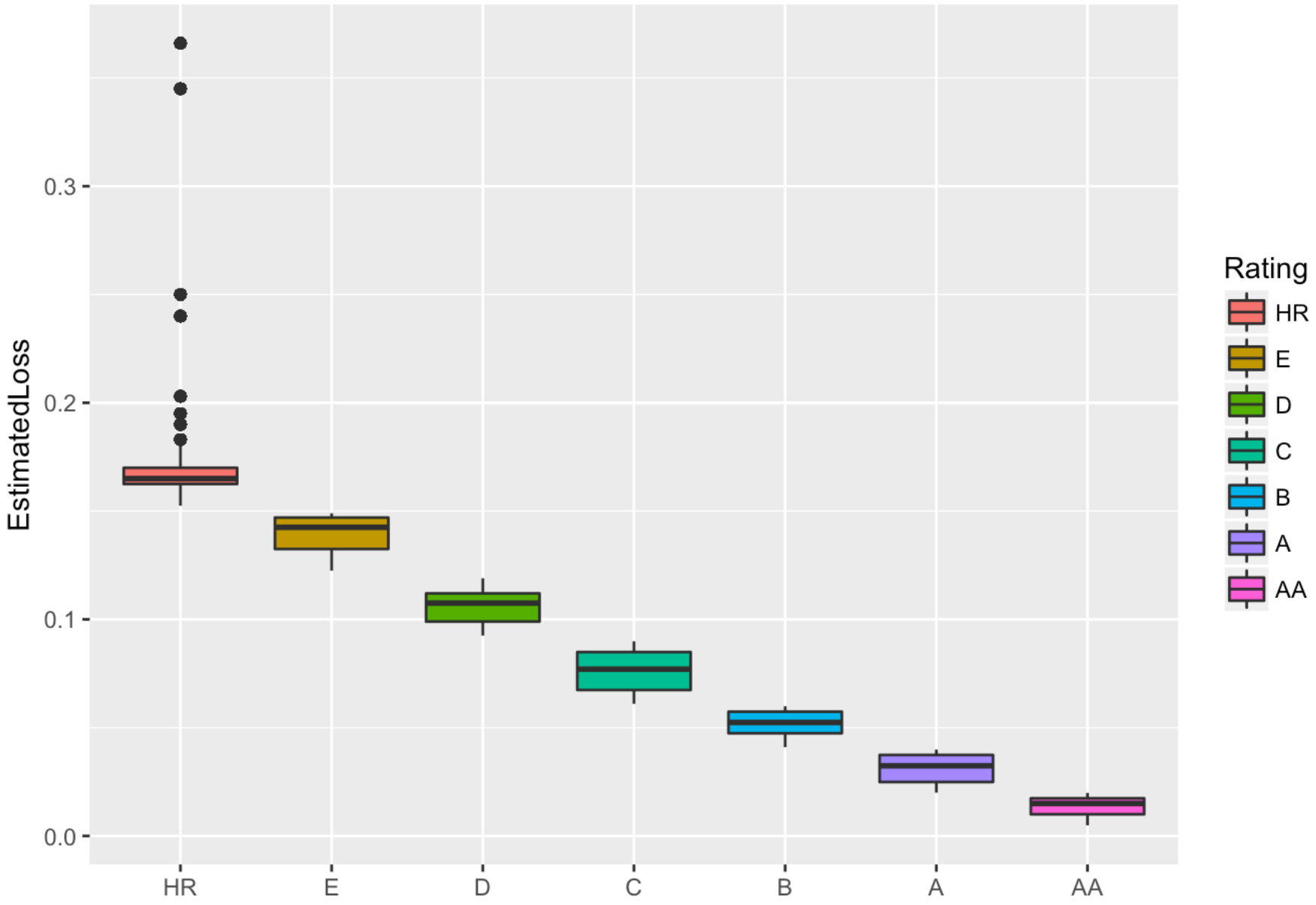
The estimated return increses with increasing estimiated effective yield and decreases with increasing estimated loss.

Monthly payments with loan amount



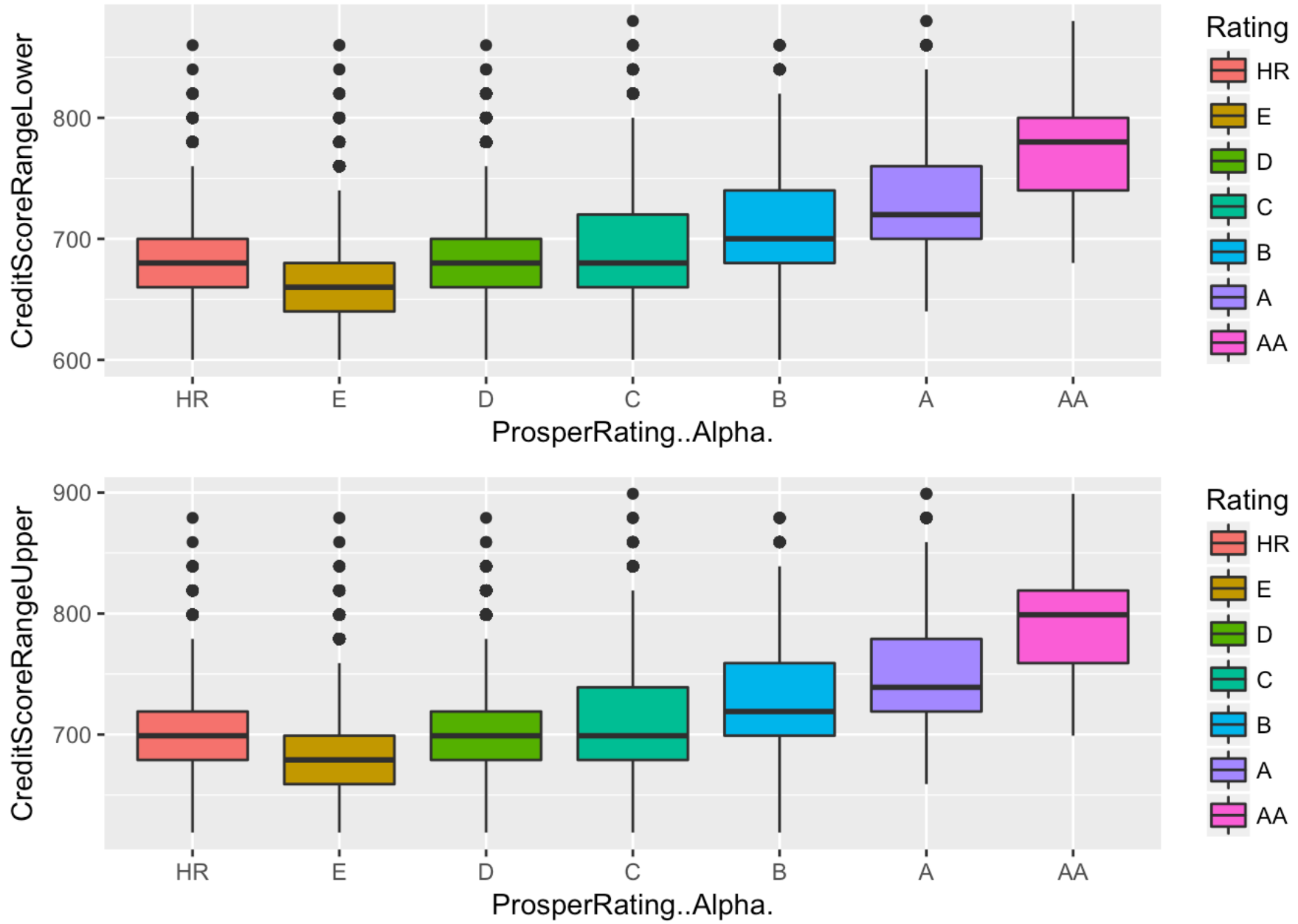
The monthly payment increases with the amount borrowed. This is expected since borrowers with higher loans need to pay more to pay off their debt in the specified loan period.

Estimated loss and Prosper rating



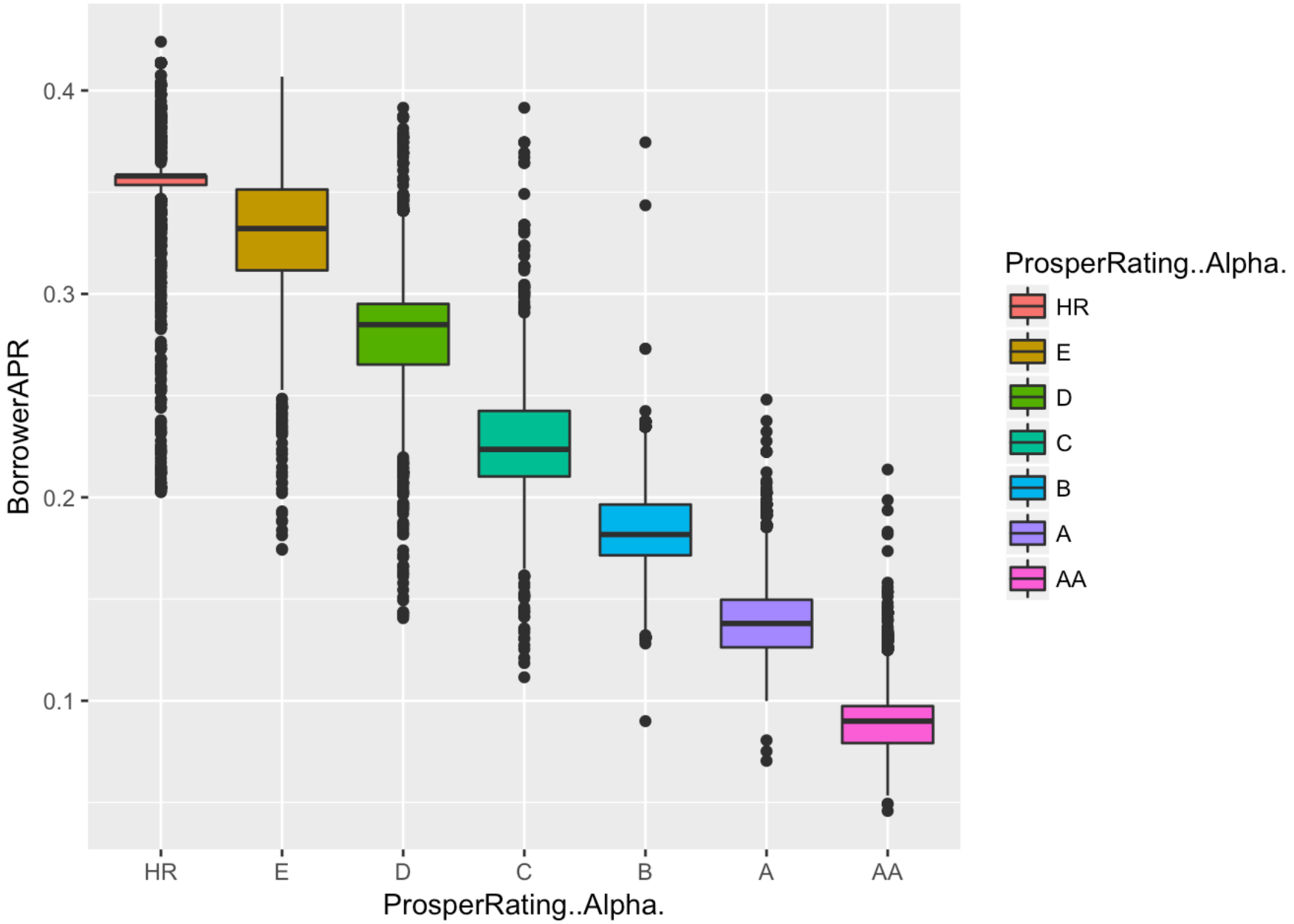
The estimated loss is low for borrowers with good prosper rating and high for those with bad prosper rating.

Prosper rating and credit score



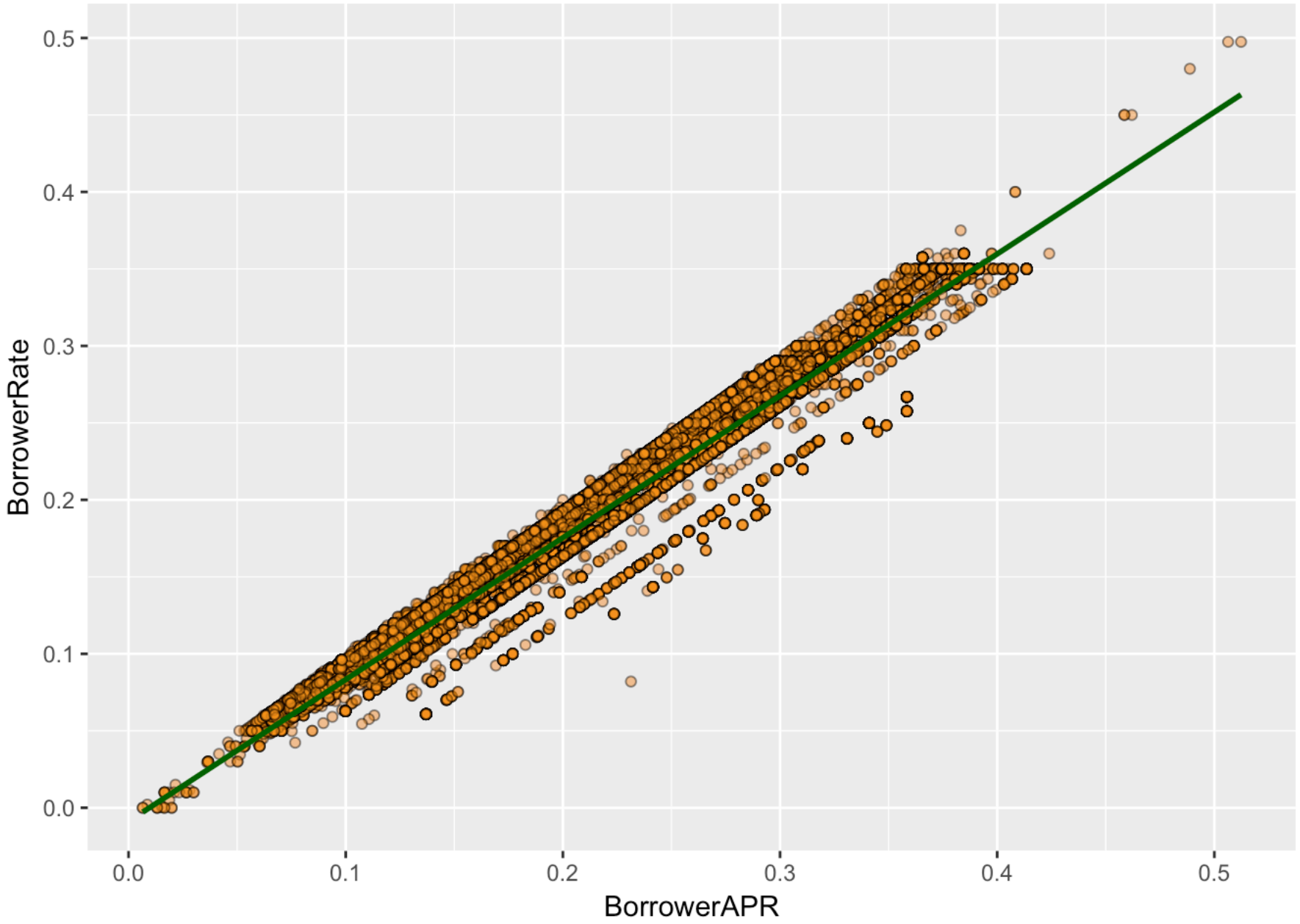
The plots show that, borrowers with good prosper ratings have higher credit scores and those with bad ratings have realtively lower credit scores.

APR and Prosper rating



Similarly, borrowers with good prosper rating have lower APRs and those with bad ratings have high APRs.

APR and interest rate



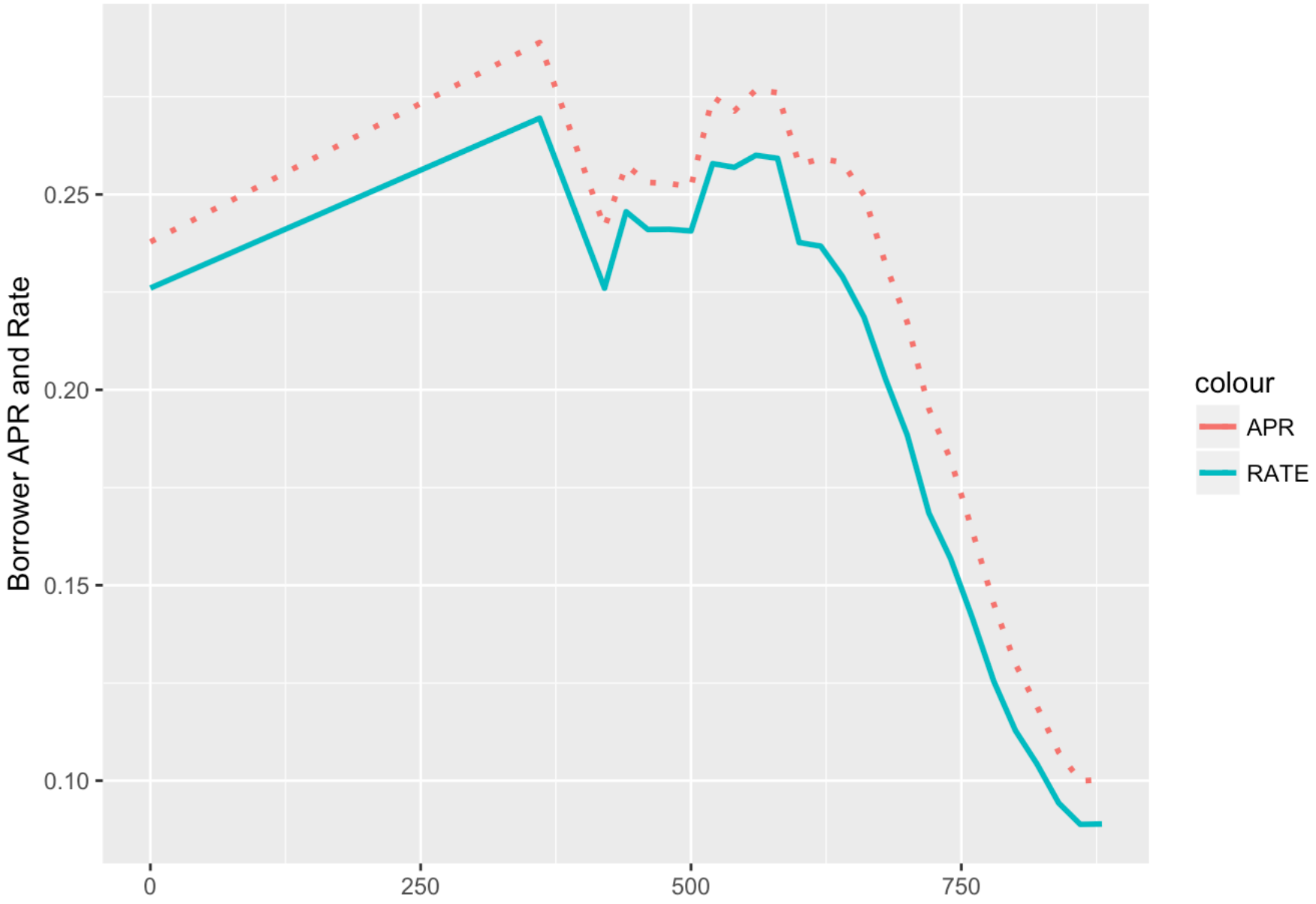

```
## [1] "The summary statistics of borrower APR is "
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## 0.00653 0.15630 0.20980 0.21880 0.28380 0.51230      25
```

```
## [1] "The summary statistics of borrower rate is "
```

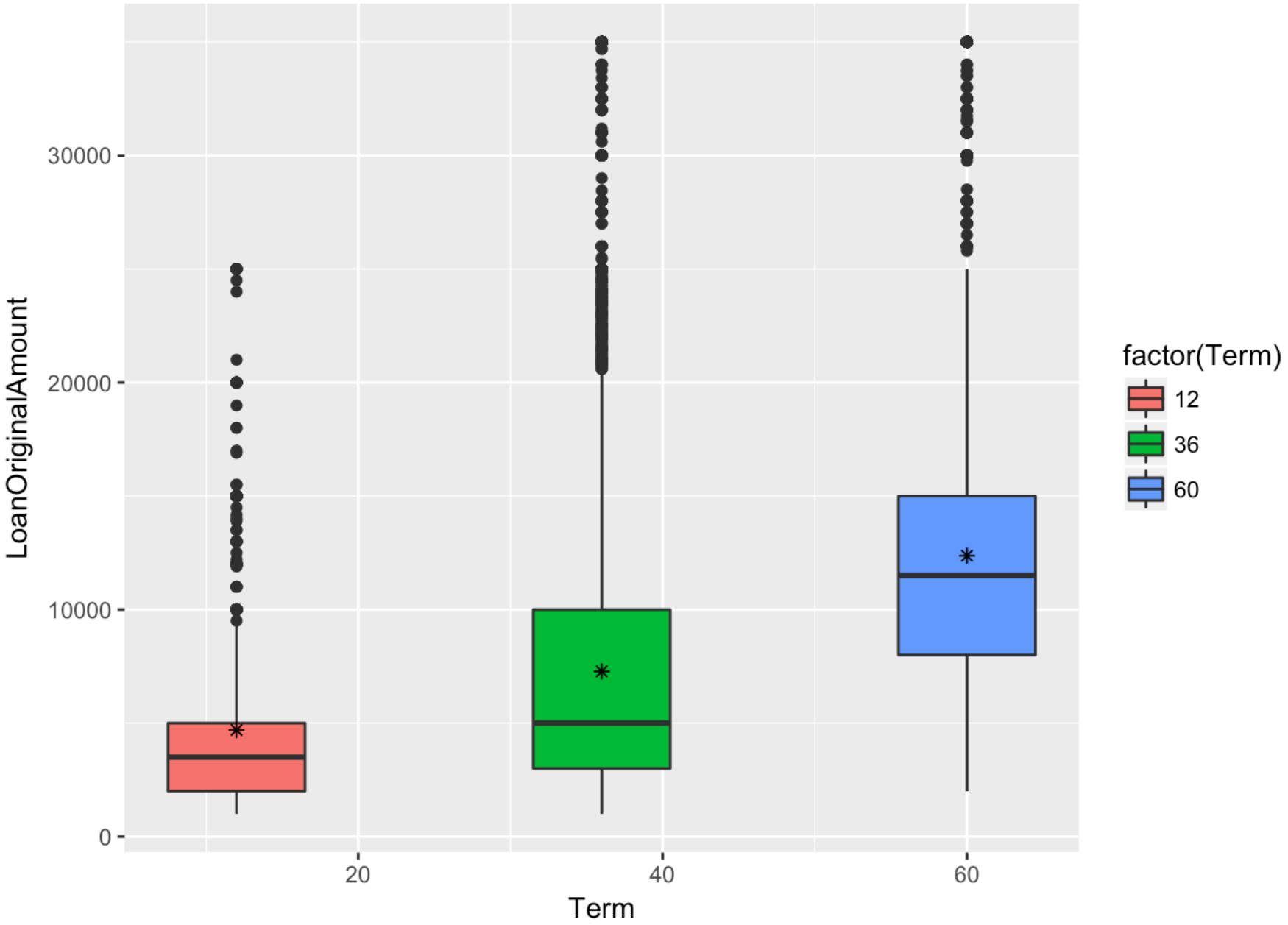
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.1340  0.1840  0.1928  0.2500  0.4975
```

The borrower APR and rate are linearly related. The average APR and rate are ~ 22% and 19 % respectively. The minimum 0 % interest rate is
strange and needs to be investigated further.



Over all the Borrower APR and interest rate have inverse relation with credit score.

Loan Amount vs loan term




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1000    4000    6500    8337   12000   35000
```

It is clear that borrowers who owe larger loan amounts borrow for longer durations. This is expected since they need longer time to pay off their loans or otherwise need to pay higher monthly paymemnts with shorter loan terms. The stars in the boxes show the average loan amount. From the summary statistics we can see the average loan amount is $8337 and the loan amount ranges from $1000 to $35000.

MULTIVARIATE ANALYSIS

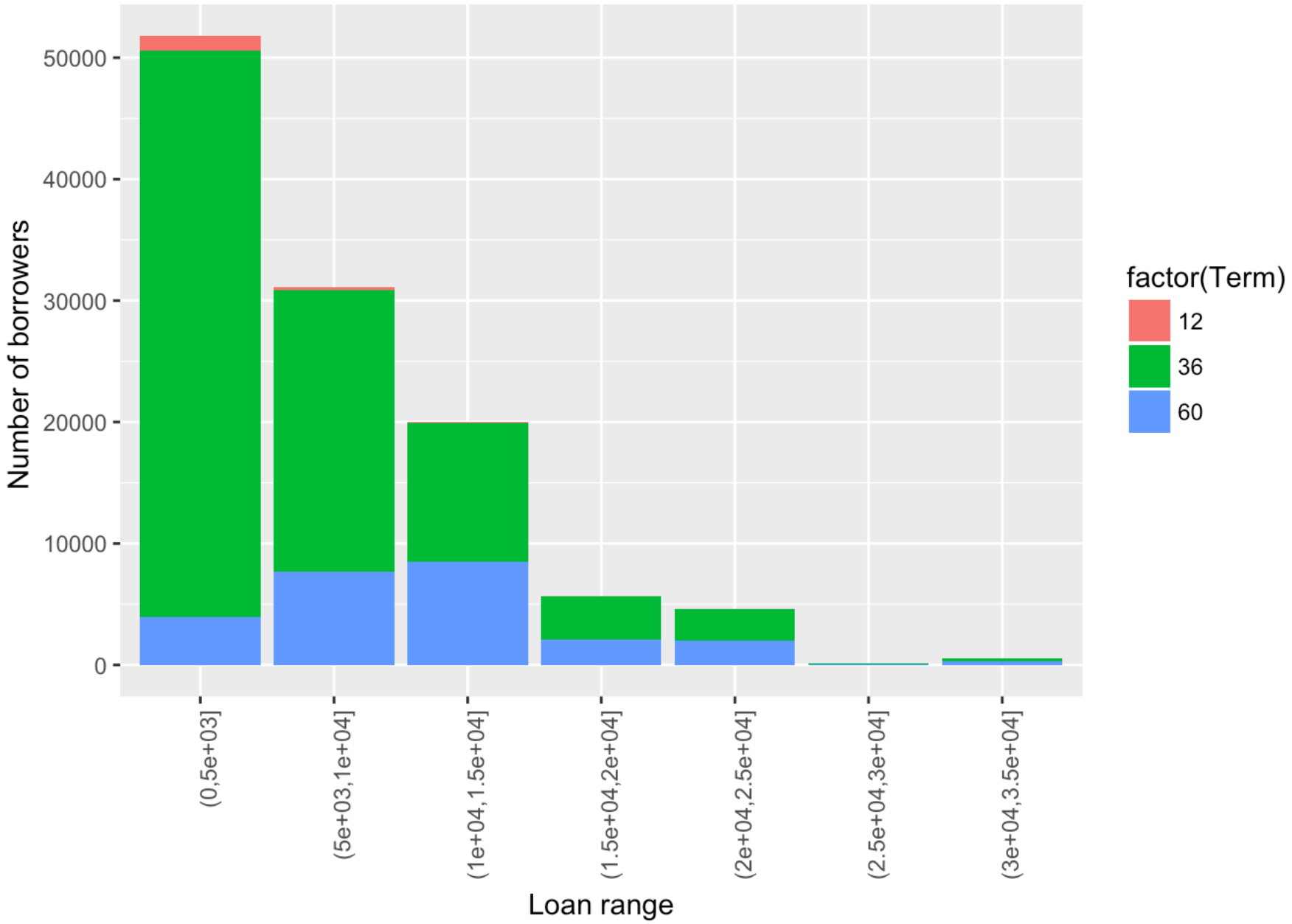
Monthly payement by total loan and term


```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   131.6   217.7   272.5   371.6  2252.0
```

The monthly payment increases with loan amount. For the same loan amount the monthly payment amount is higher for shorter terms. This is expected since borrowers need to pay higher monthly payments in order to payoff the balance with in shorter terms. From the summary statistics we can see that the average monthly payment is ~ $272, and the maximum monthly payment is $2252. The minimum monthly payment which is 0.0 is clearly an outlier associated with cancelled loans.

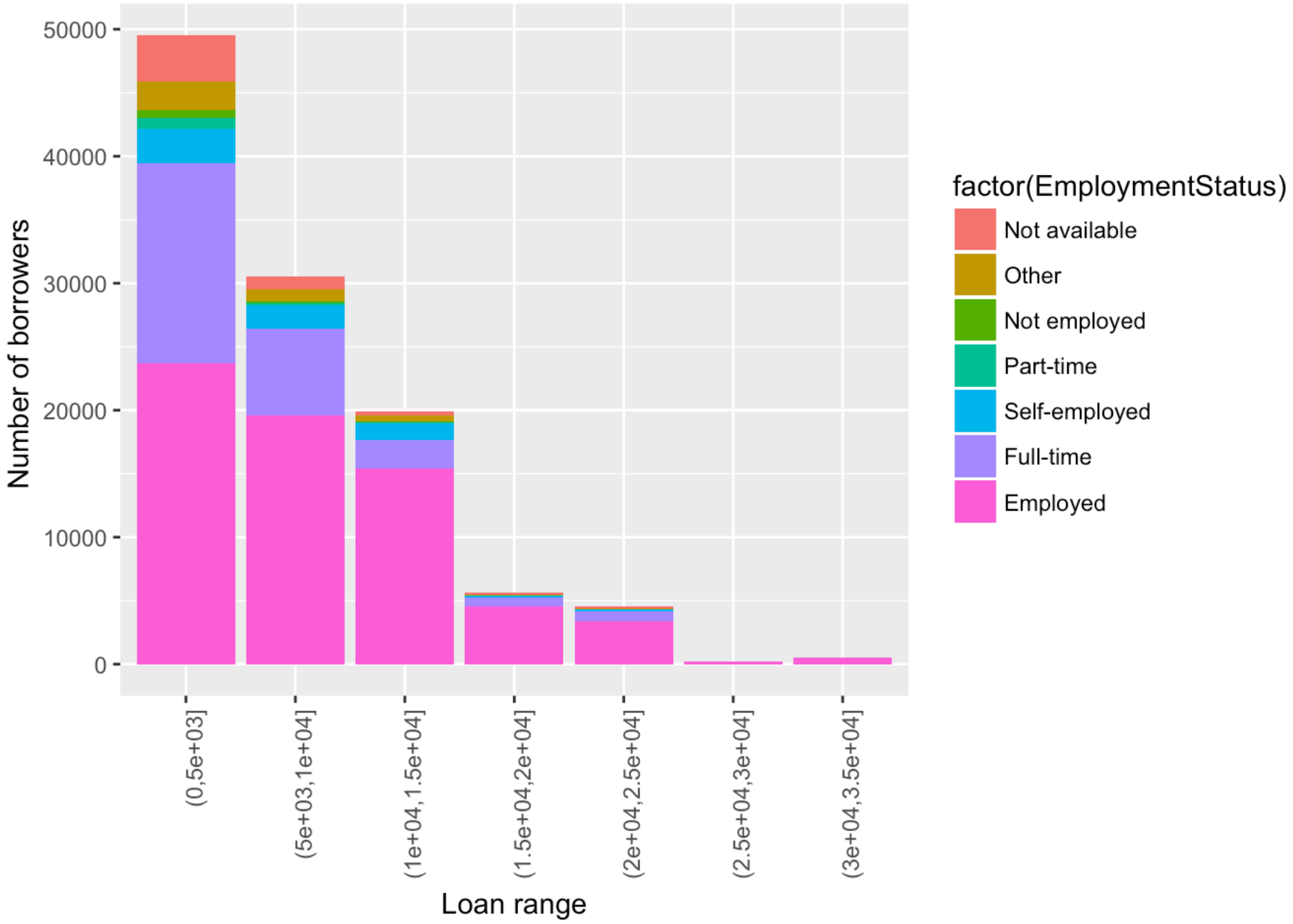
Borrower classification by loan range and loan term?

```
##        Loan_range Number_of_borrowers
## 1       (0,5e+03]                51824
## 2   (5e+03,1e+04]                31136
## 3 (1e+04,1.5e+04]                20033
## 4 (1.5e+04,2e+04]                 5659
## 5 (2e+04,2.5e+04]                 4605
## 6 (2.5e+04,3e+04]                  177
## 7 (3e+04,3.5e+04]                  503
```



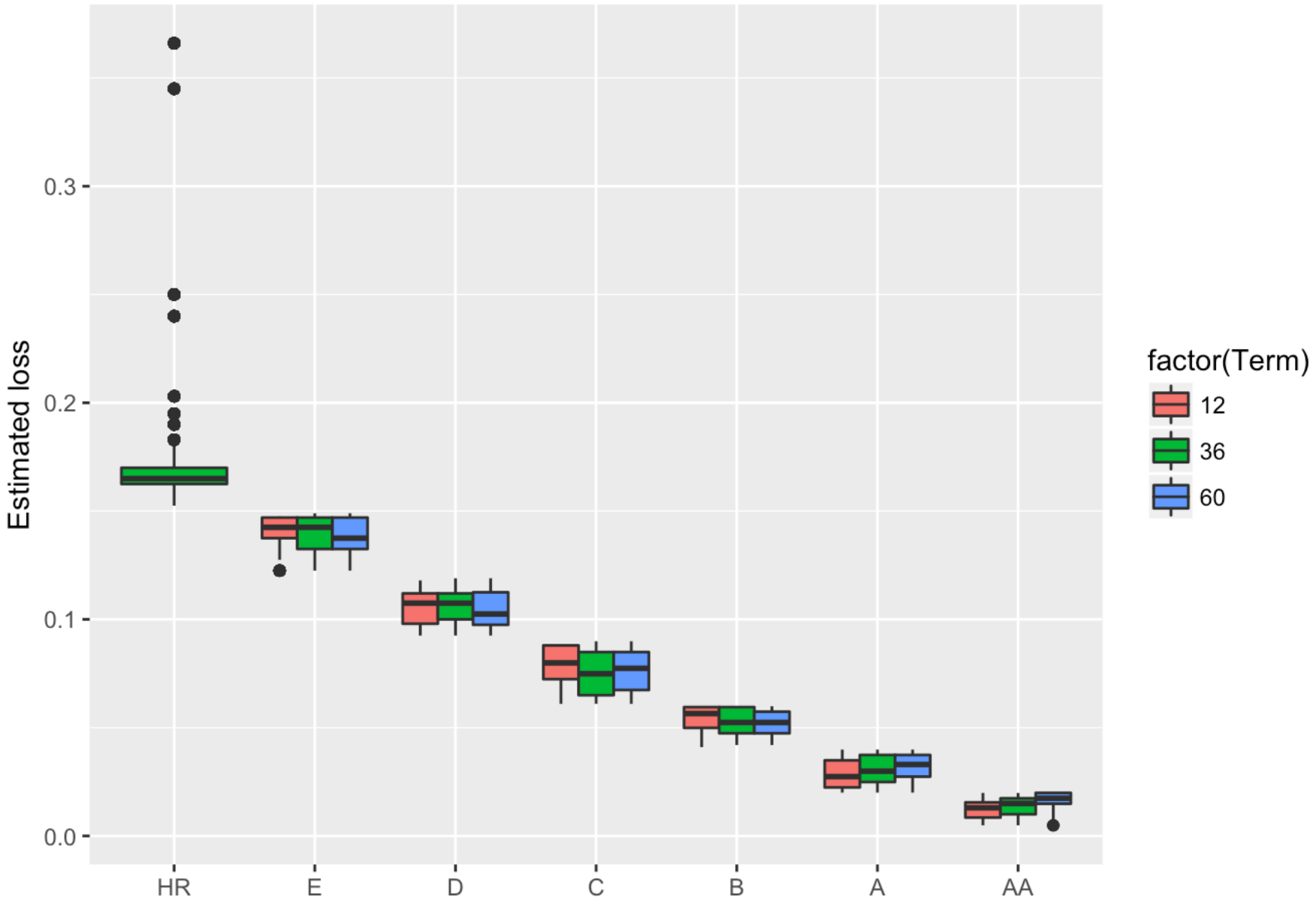
It is clear to see that most borrowers have loans in the range from 1-5000 for aperiod of 36 months. Generallly most borrowers borrowed for 36 and 60 months and there are no borrowers borrowed more than 15000 for a period of 12 months.

Distribution of borrowers by loan range and employment status?



It is interesting to see that most prosper loan amounts are in the range of 0,5000 followed by 5000 - 10,000 and most borrowers are employed.

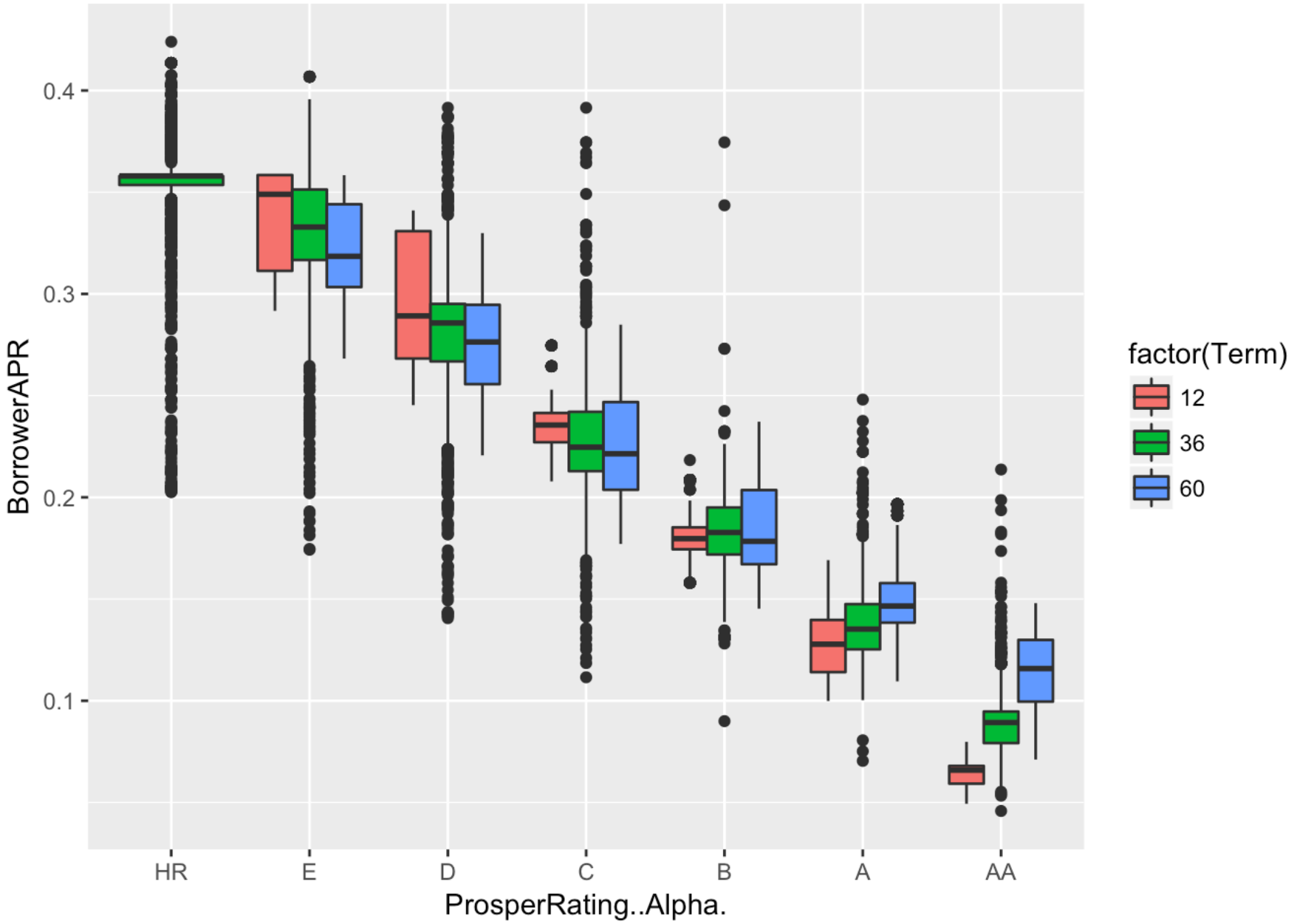
Effective yield with rating and term



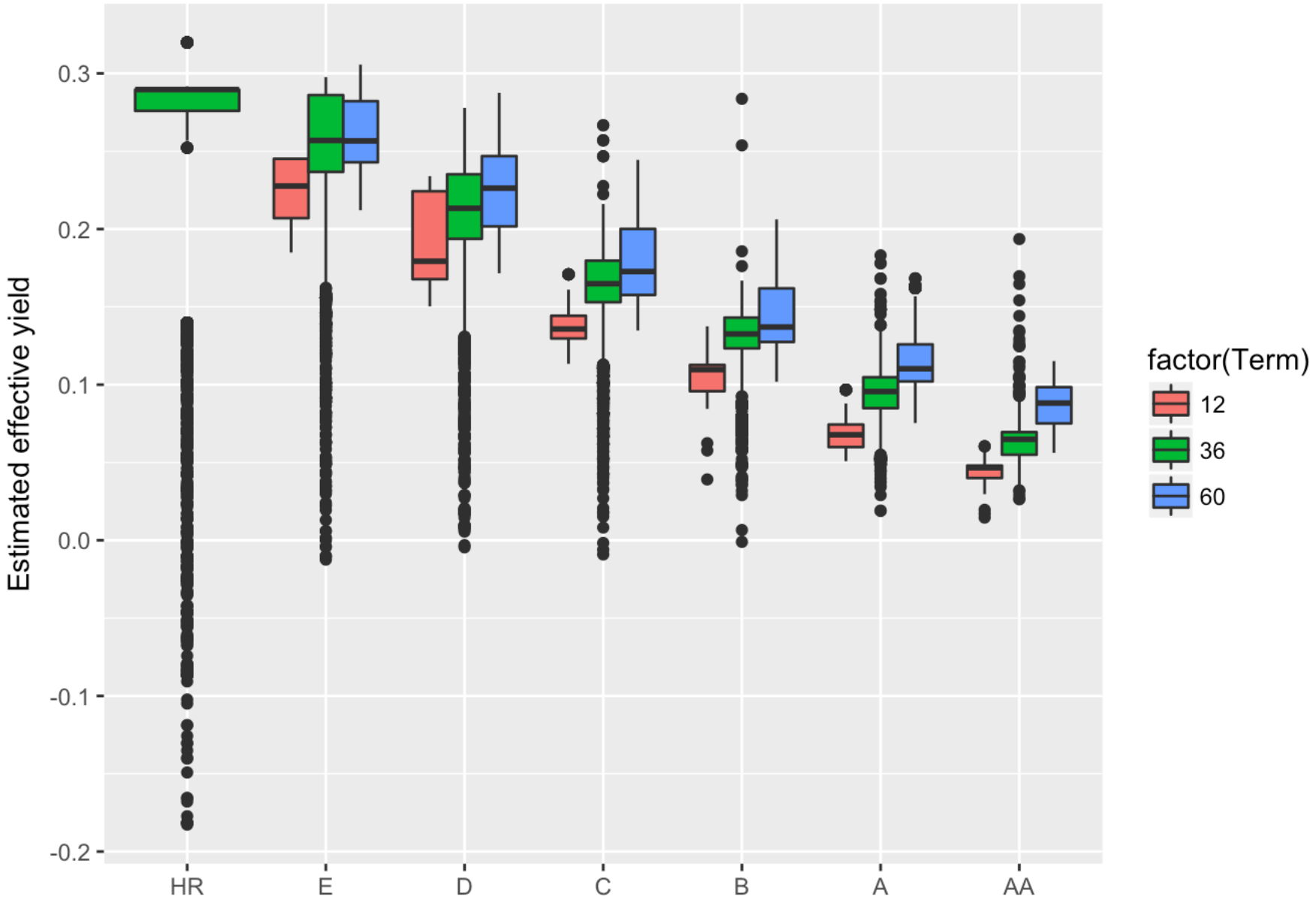

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.005   0.042   0.072   0.080   0.112   0.366   29084
```

The estimated loss increases with decreasing prosper rating. For same proper rating of (E,D,C,B), the median estimated loss is a little bit higher with shorter loan periods and vice versa for A and AA ratings. The average estimated loss is 8 % and the maximum estimated loss is 36.6 %.

APR with term and rating



The borrower APR decreases with increasing prosper rating grade. At the same rating, for boorrowers with lower rating (ratings E,D,C and B), the median borrower APR decreases with loan period and increases with loan period for borrowers with rating A and AA.

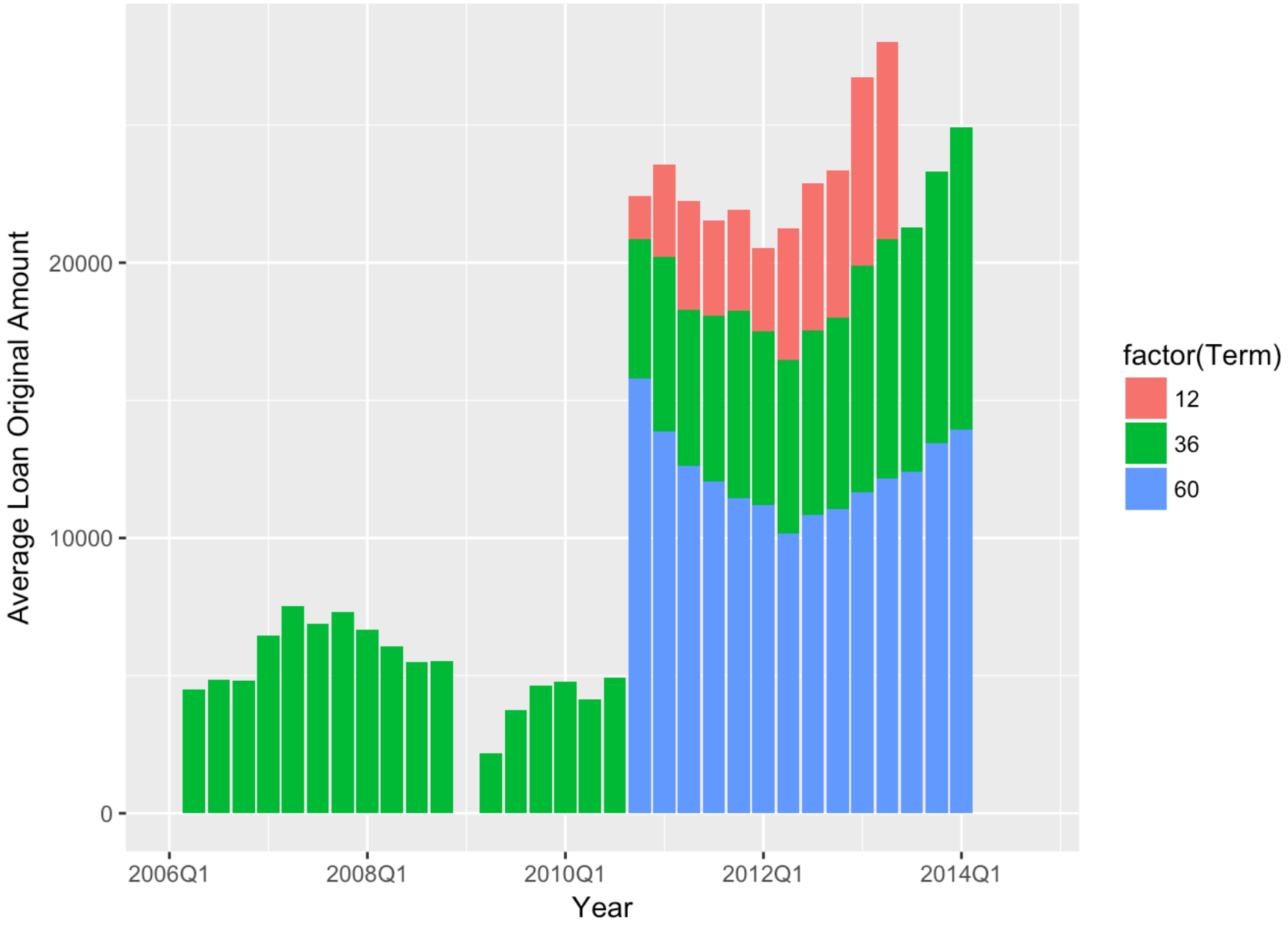


Prosper rating

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  -0.183   0.116   0.162   0.169   0.224   0.320   29084
```

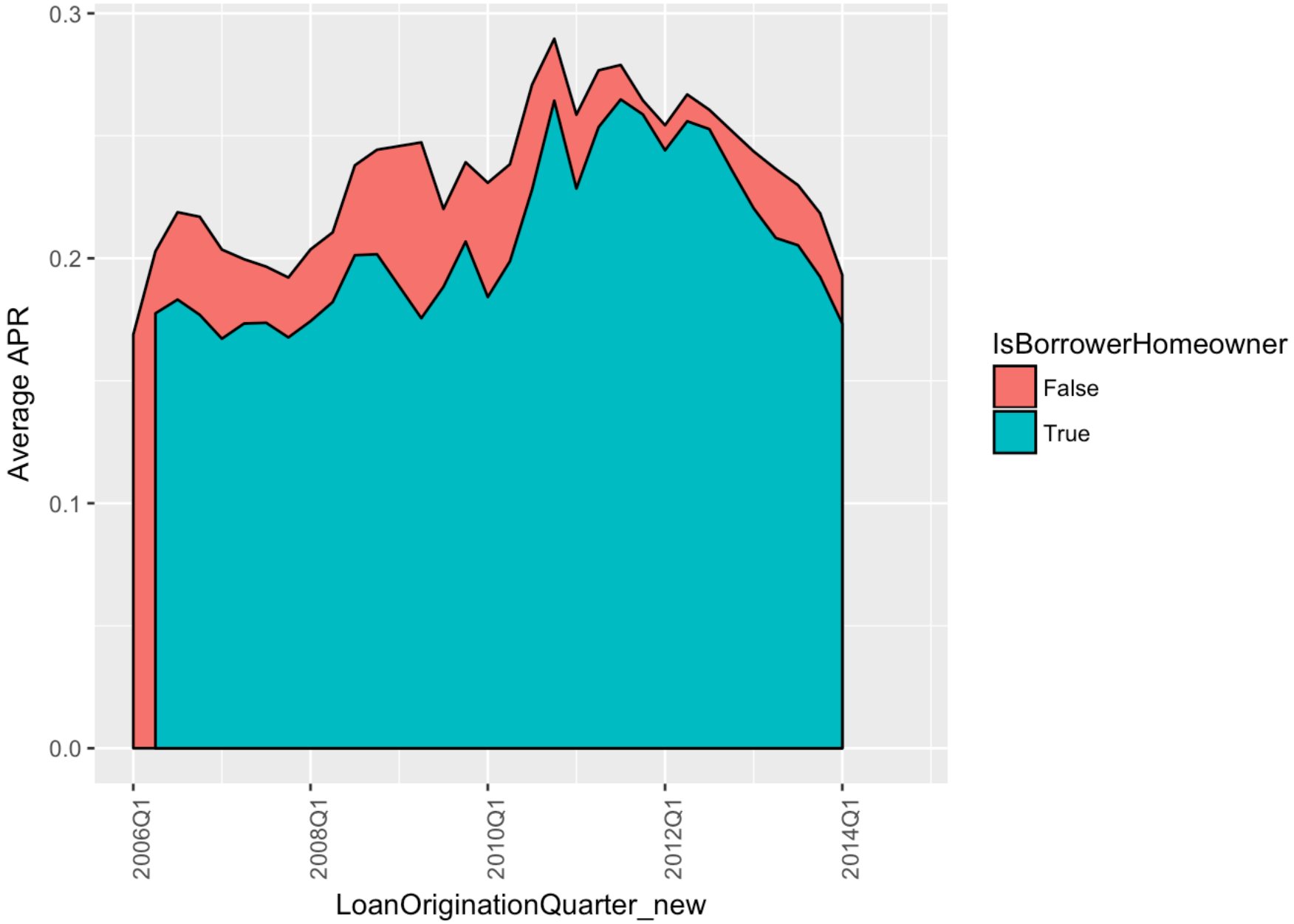
The estimated effective yield obtained from borrowers with lower prosper ratings is generally higher. This might be due to the fact that there are fewer customers with good ratings and also the APR and rate is higher for the borrowers with lower prosper ratings as we have already seen.
Insterestingly, for borrowers with the same rating, the estimated yield increases with increasing loan period. This is also explained by the fact that the more borrowers stayed in the loan the more interest they would pay and hence increase the estimated effective yield. The all over average estimated effective yield is ~ 17 % which is close to the difference between average estimated return and average estimated loss.

Average loan amount per quarter year



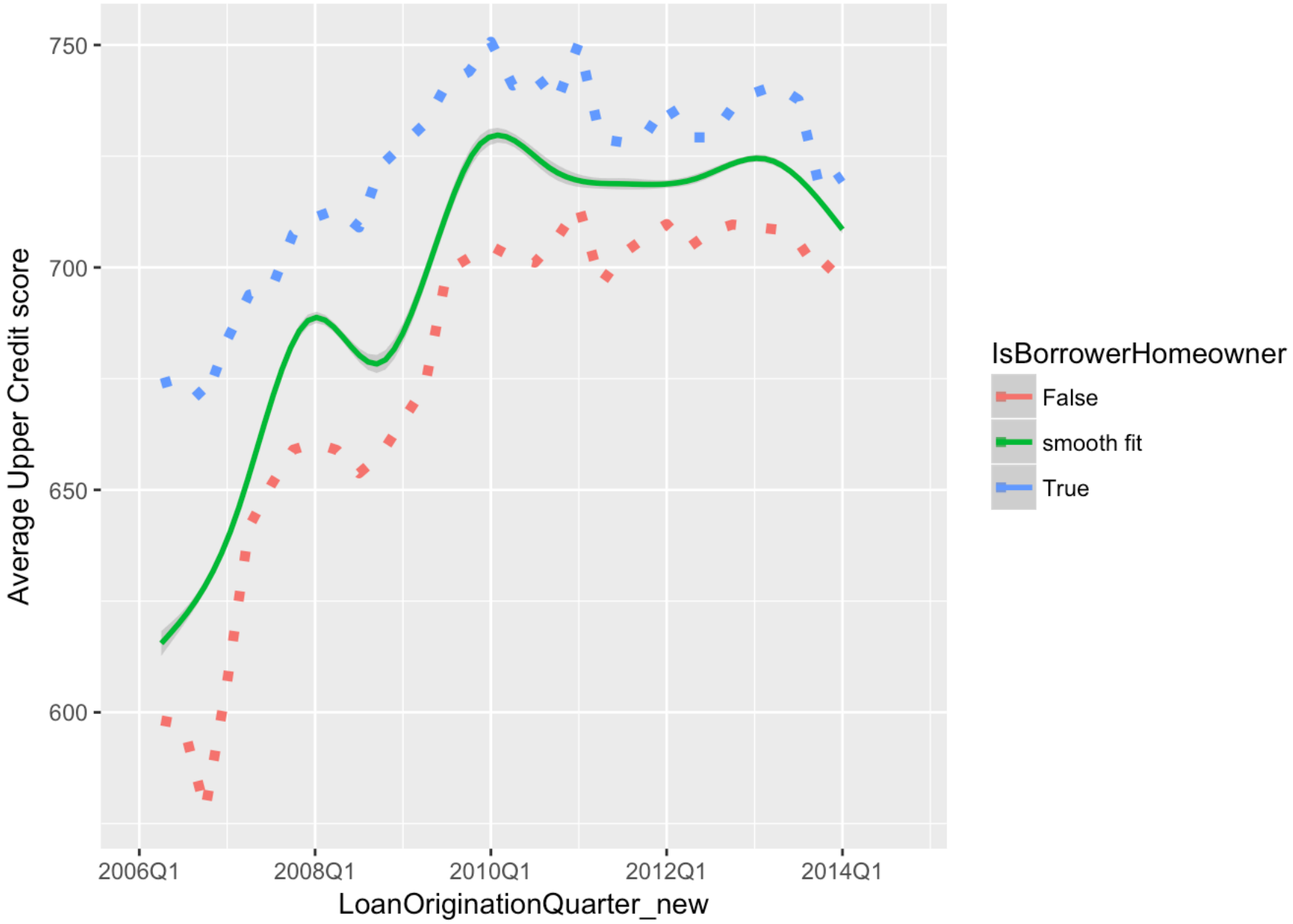
The loan amount borrowed until the third quarter of 2010 was less than $10000 and was only for a period of 36 months. It is interesting to see that both higher loan amount of more than 10000 and flexible payment plans were introduced after the fourth quarter of 2010.

Average APR by quarter year and home owner



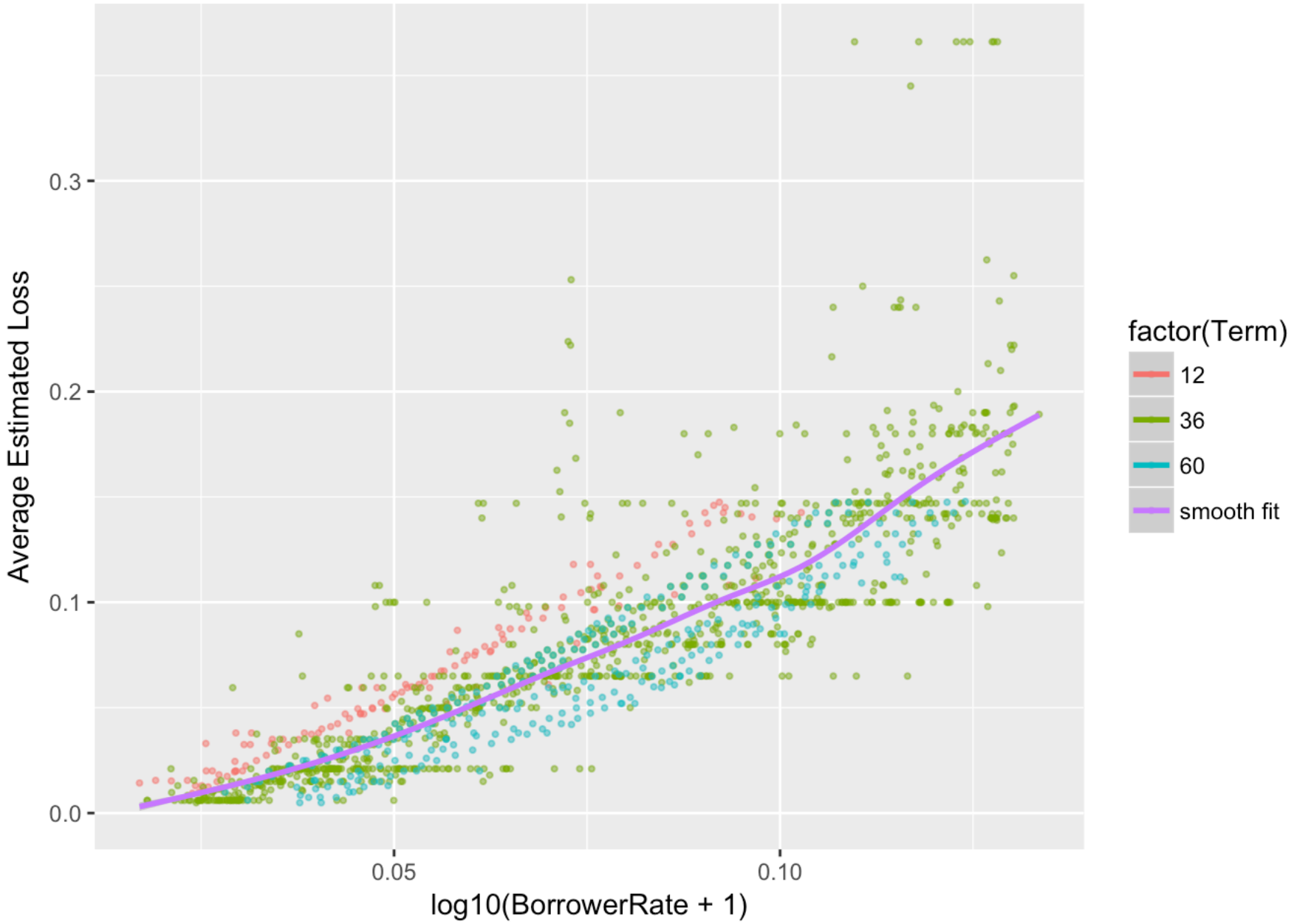
The average APR of home owners is less than the others over the whole year. This might be associated with the credit history of the borrowers.

Credit score



The average credit score of borrowers with home owners is above 650 over the whole year.

Average estimated loss and borrower rate



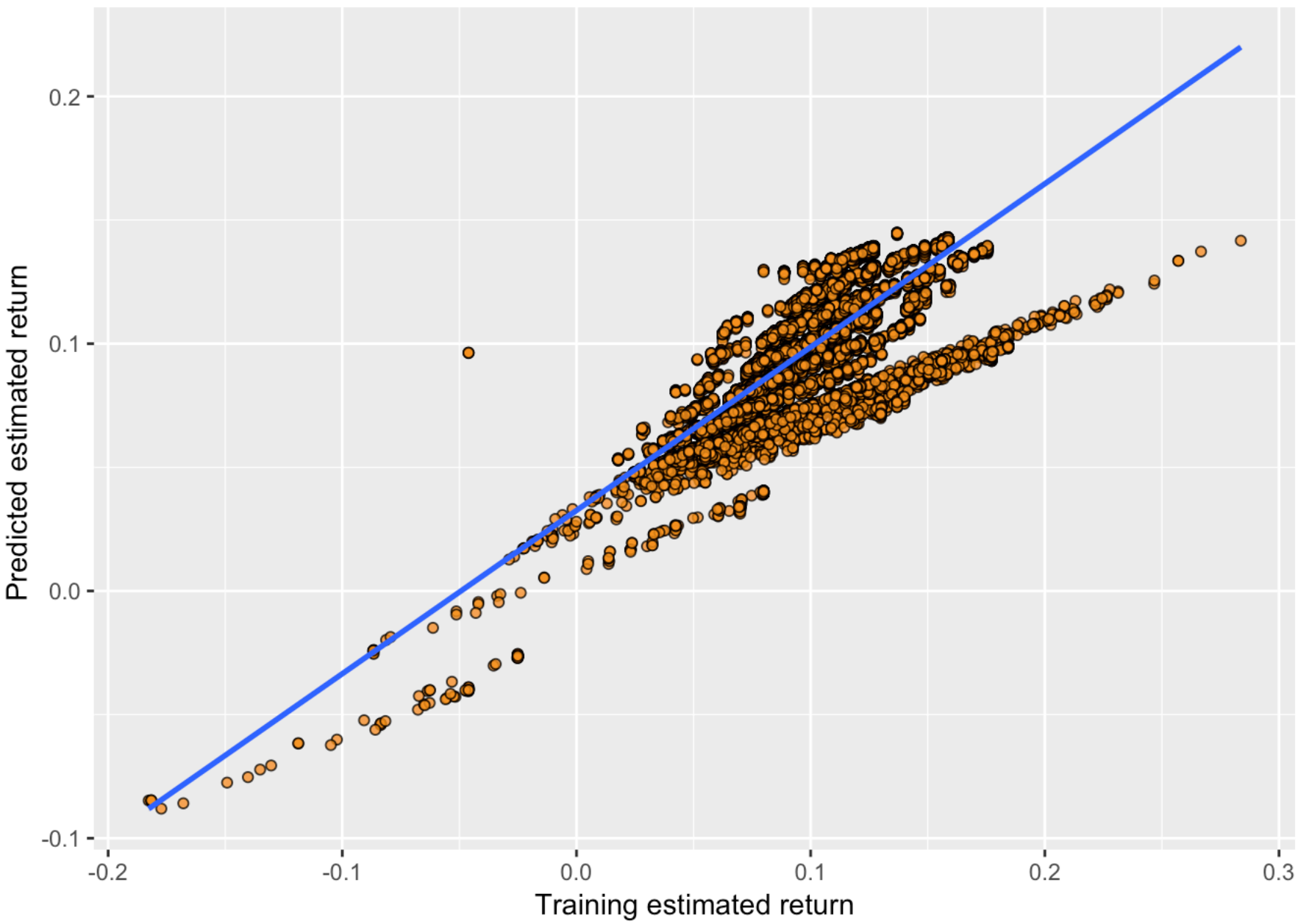
It is interesting to see that the average estimated loss is higher for borrowers with higher interest rates.

Correlation between different variables

From the above correlation table we can see, there is a strong negative relatioship between prosper rating and estimated loss.

Predicting estimated return using multivariat regression

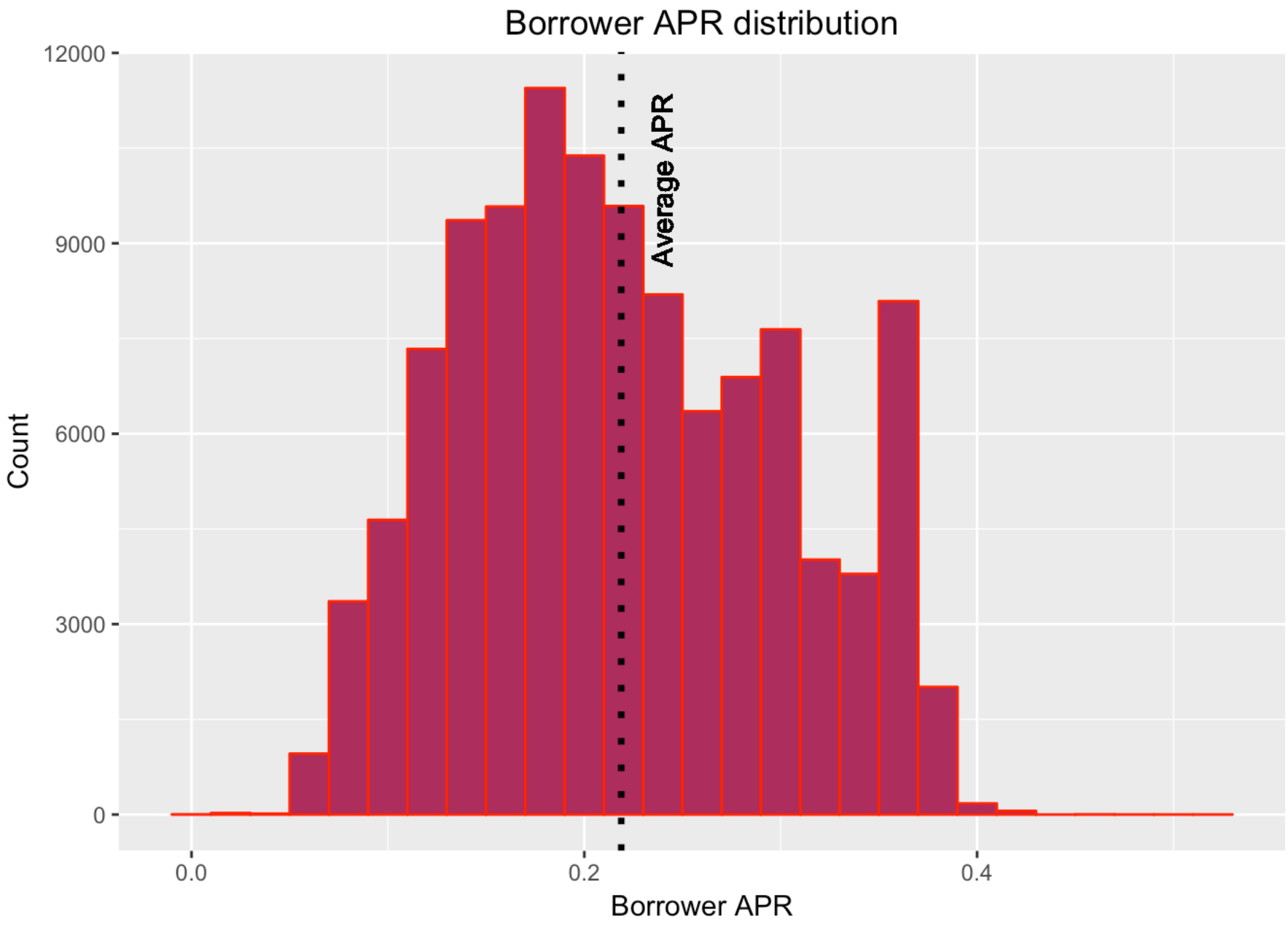
```
##
## Call:
## lm(formula = y ~ ., data = ddf2)
##
## Residuals:
##       Min         1Q    Median         3Q       Max
## -0.142468 -0.010746 -0.003912  0.006014  0.142047
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                7.716e-02  1.719e-03  44.895   <2e-16 ***
## EstimatedEffectiveYield    3.722e-01  2.024e-03 183.914   <2e-16 ***
## EstimatedLoss             -2.612e-01  5.948e-03 -43.908   <2e-16 ***
## ProsperRating..numeric.   -6.171e-03  1.938e-04 -31.836   <2e-16 ***
## EmploymentStatusDuration -6.490e-06  7.429e-07  -8.736   <2e-16 ***
## CreditScoreRangeLower      4.035e-06  1.843e-06   2.189   0.0286 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0177 on 59994 degrees of freedom
## Multiple R-squared:  0.6604, Adjusted R-squared:  0.6604
## F-statistic: 2.334e+04 on 5 and 59994 DF,  p-value: < 2.2e-16
```



I used multivariat linear regression to see how the estimated return related to other variables describing the borrowers. The detail results of the model is described in the summary. Though the model is not well tunned, it explains up to 69% of the variation in the estimated returns.

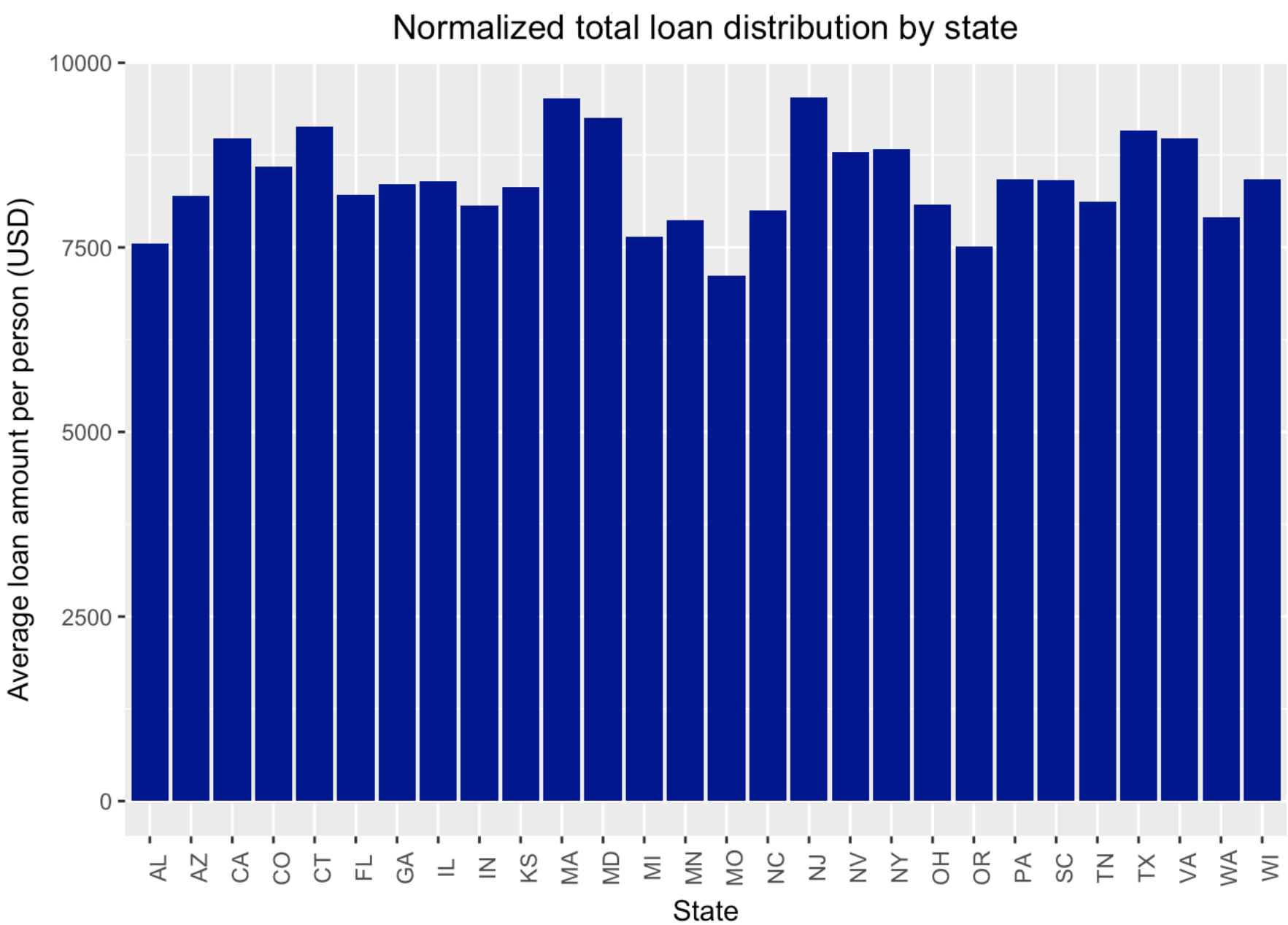
FINAL PLOTS

Plot 1



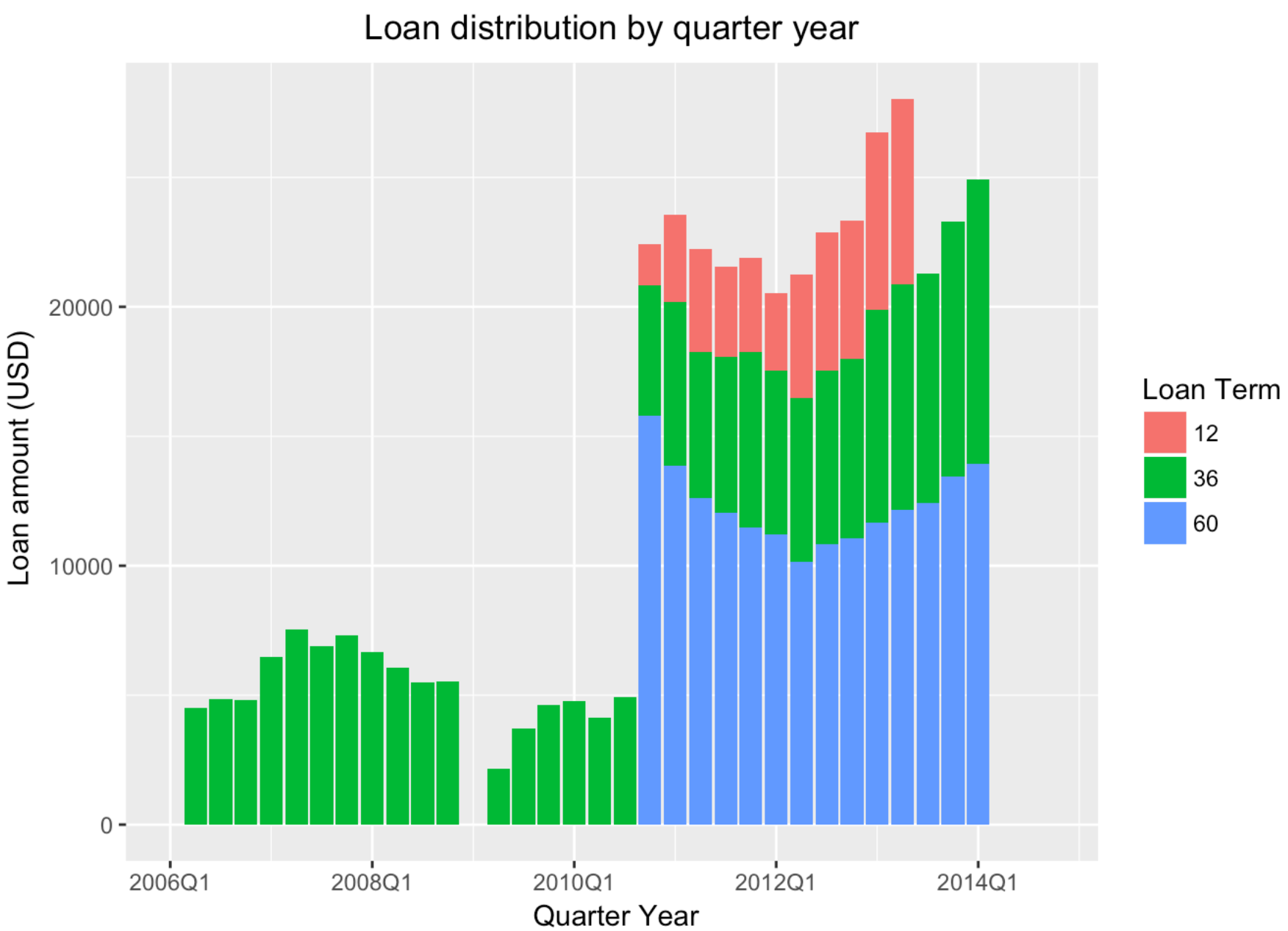
The loan APR ranges from less than 5 % up to about 51 %. However, there are only very few borrowers with such high and lower APR values. Most borrowers seems to have a APR from 10 % to 30 %. The median APR value is ~21 %.

Plot 2



When I investigated the demography of borrowers by state, I found a higher value of total loan amount and larger number of borrowers especially for the states CA, TX, NY and FL. This led me to ask is there any special reason for this? are people from these states tend to borrow more likely compared to people from other states? In order to answer this question and compare borrowers by state, I normalized the total loan amount by the number of borrowers to see the average loan amount per person. As can be seen above the distribution of the normalized loan amount is almost nearly uniform and the states even found to have normalized loan amount less than some other states. This shows that the higher number of borrowers in these states might be due to higher number population in the states so that the number of applicants is also higher.

Plot 3



The loan amount borrowed until the third quarter of 2010 was less than 10000 and was only for a period of 36 months. It is interesting to see that both higher loan amount of more than 10000 and flexible payment plans were introduced after the fourth quarter of 2010.

REFLECTION

The Prosper loan data set has 113,937 rows and 81 features. I explored some of the 81 variables and tried to find interesting relationships among the
variables. My focus of investigation was mostly on the profile of the borrowers such as income range, employment, credit score etc.. and the features of the loan such as APR, estimated loss, loan amount with time etc..,. I used ggplot to check how these variable distributed and related. On the process of exploration, I found many missing values in the data and I handeled the problem using R data cleaning tools such as from the deplyer package. Another challenge I encountered was the datset contains a lot of outliers and noise that makes hardly easy to extract the information I want using scatter or bar plots directly from the variables of interest. I handeled this problem by using transformations such as scaling the variables from linear to logarithimic, by averaging and choosing more approprioate ploting styles for example using box plot instead of histograms and scatter plots. In this way I was able find informative trends that explains the relationships between the variables of interest.

From the exploration of the data, I found the APR varies from less than 1 % upto a little higher than 50 %. The APR is related to the credit score of the borrowers. It is foud that borrowers with high credit scores have good prosper rating and lower APR. It is also found that most borrowers have loans in the range from 1-5000 for aperiod of 36 months and the loan amount grown over time since the second quarter of 2009. I estimated the percentage of loan status and found that about 10.74 % of the loan is Chargedoff and 4.5 % is defaulted. Finally,I used multivariat linear regression to predict the estimated return. and the model explains up to 69% of the variation in the estimated returns. Even though this rough estimation gives a clue about the possibility of using machine learning to investigate the data, I recommend it to be done more carefully with proper feature selection, careful algorithm choices and proper parametric optimization. Since the data has many features, using combination of principal component analysis (PCA) to reduce the dimensionality and univariate feature selection could be helpful to develop a well tuned model of the data.

REFERENCES

http://napitupulu-jon.appspot.com/posts/eda-prosper-loan.html (http://napitupulu-jon.appspot.com/posts/eda-prosper-loan.html) https://rpubs.com/grace-pehl/prosper (https://rpubs.com/grace-pehl/prosper) https://s3.amazonaws.com/content.udacity-data.com/courses/ud651/ (https://s3.amazonaws.com/content.udacity-data.com/courses/ud651/) diamondsExample_2016-05.html http://stackoverflow.com/ (http://stackoverflow.com/)