Loan data EDA

Aziz Mamatov

Dataset parameters

The document is not very large so we can analyze it in full without creating a sample. For large document we would need to create a sample, say choosing each 10th or 100th data points. There are 81 variables with 113937 observations. There are several factor variables. Clearly we need to focus on certain variables as there are too much data. Variable_list is a list of all variables with type of variables explained. Some factor variables are clearly not factor as there are too many sets, like ClosedDate has 2803 levels.

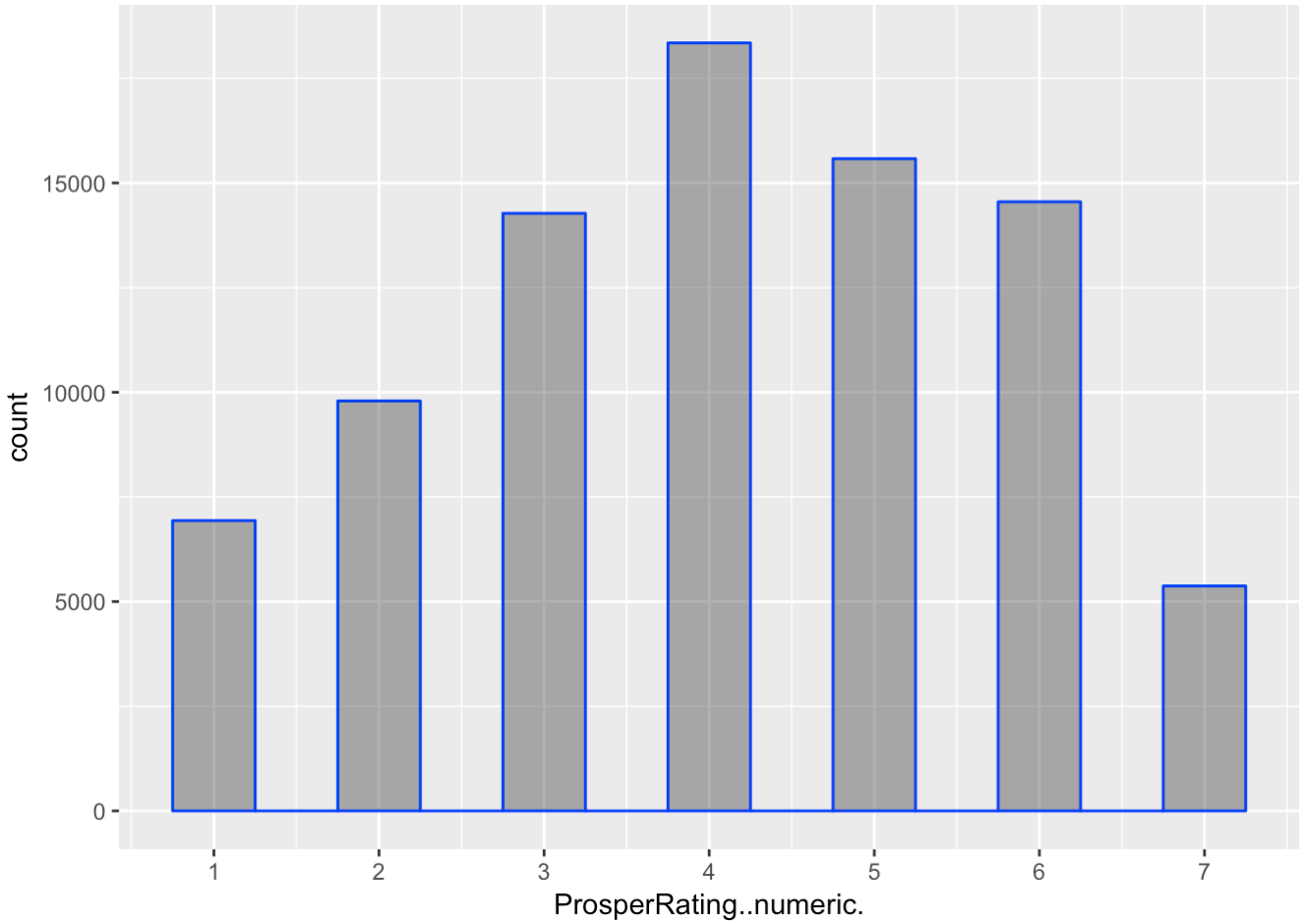
Potential issues

This is a loan performance data from Prosper and from the dictionary. I would like to explore the dependency of certain variables on others like Loan status, EmploymentStatus, Occupation, CreditGrade, Incomerange, LoanOriginalAmount, ProsperRating, CreditGrade, BorrowerState, LoanOriginalAmount, ProsperScore, ProsperPaymentsOneMonthPlusLate

From looking at individual columns we can see from that there are way too many occupations to be a factor variable. But ProsperRating, CreditGrade and ProsperScore look promising.

Initial analysis - Prosper Rating and BorrowerAPR

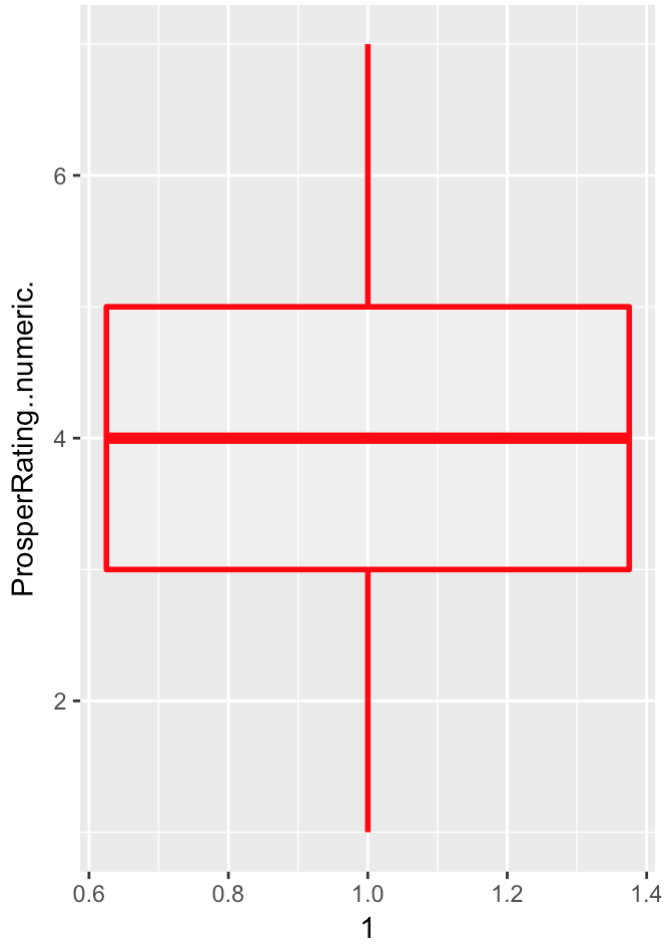
We can start with Rating and APR information as being one of the most imporant outcomes of the loan process. We can see below that for Rating the histogram is fairly symmetrical with defined mean and median value around 4 (see summary below plot)

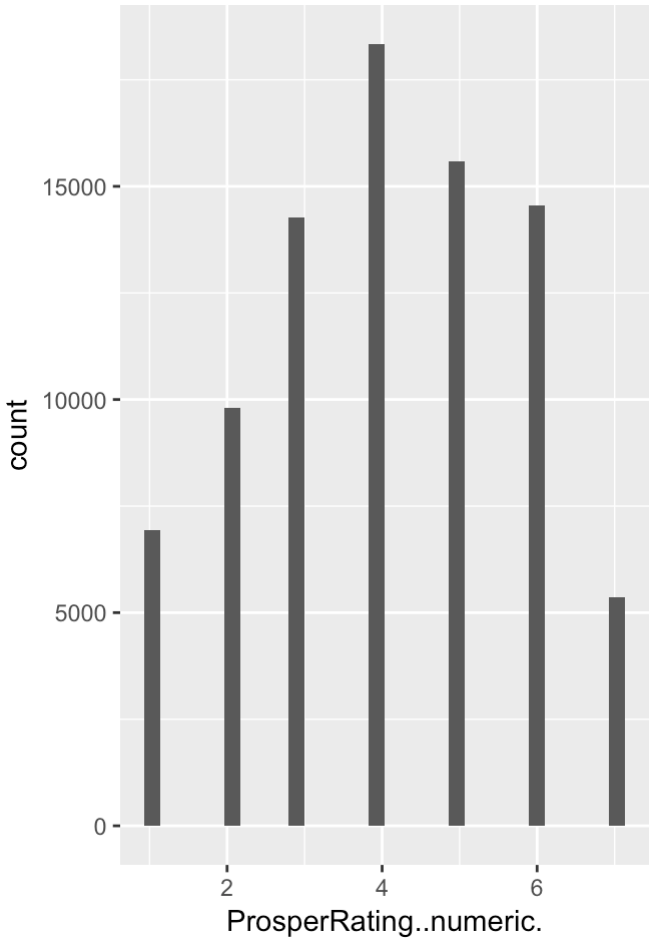



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   3.000   4.000   4.072   5.000   7.000   29084
```

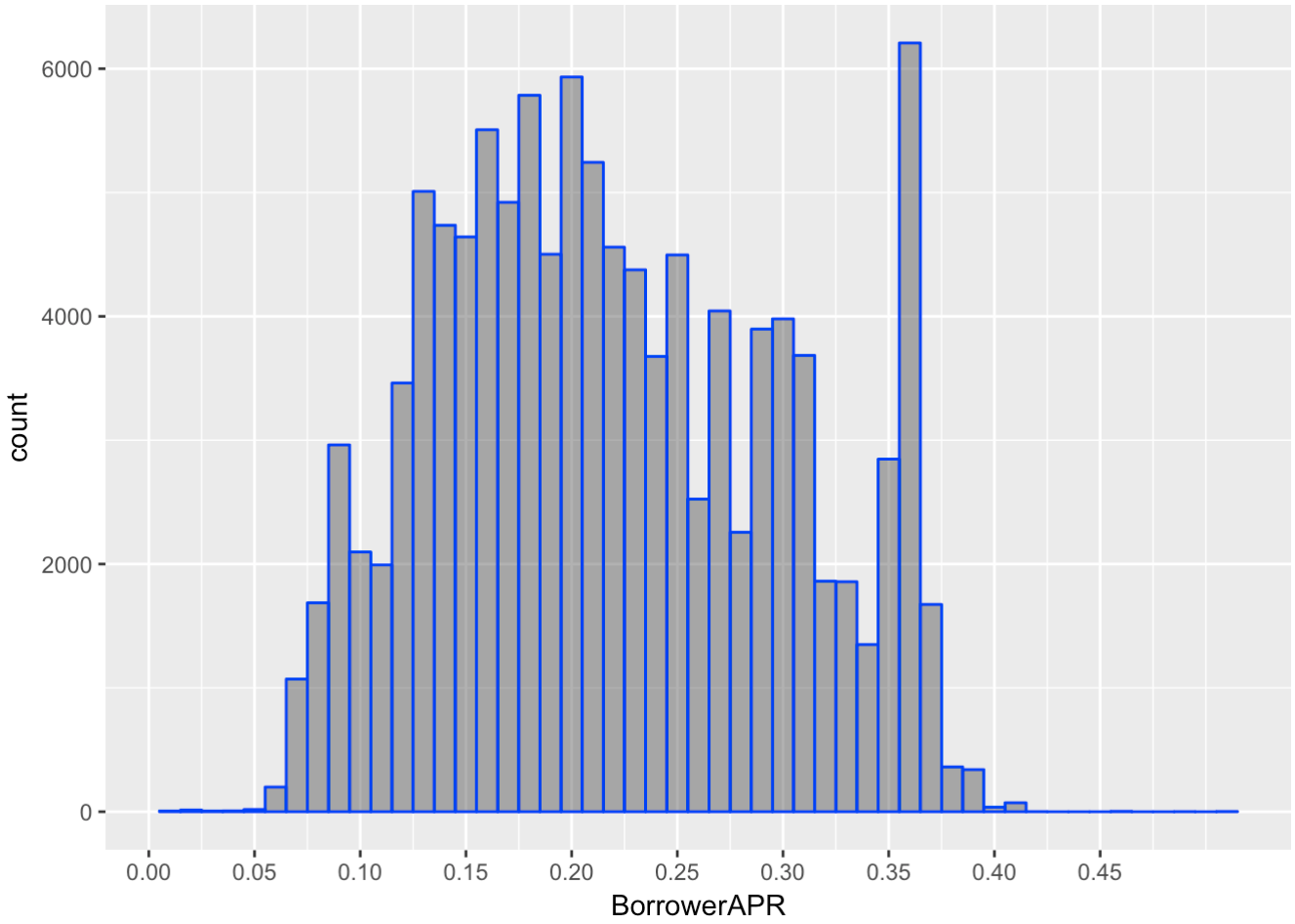
Outliers for Prosper rating

It can be seen the majority of data is properly distributed between ratings especially around median rates.



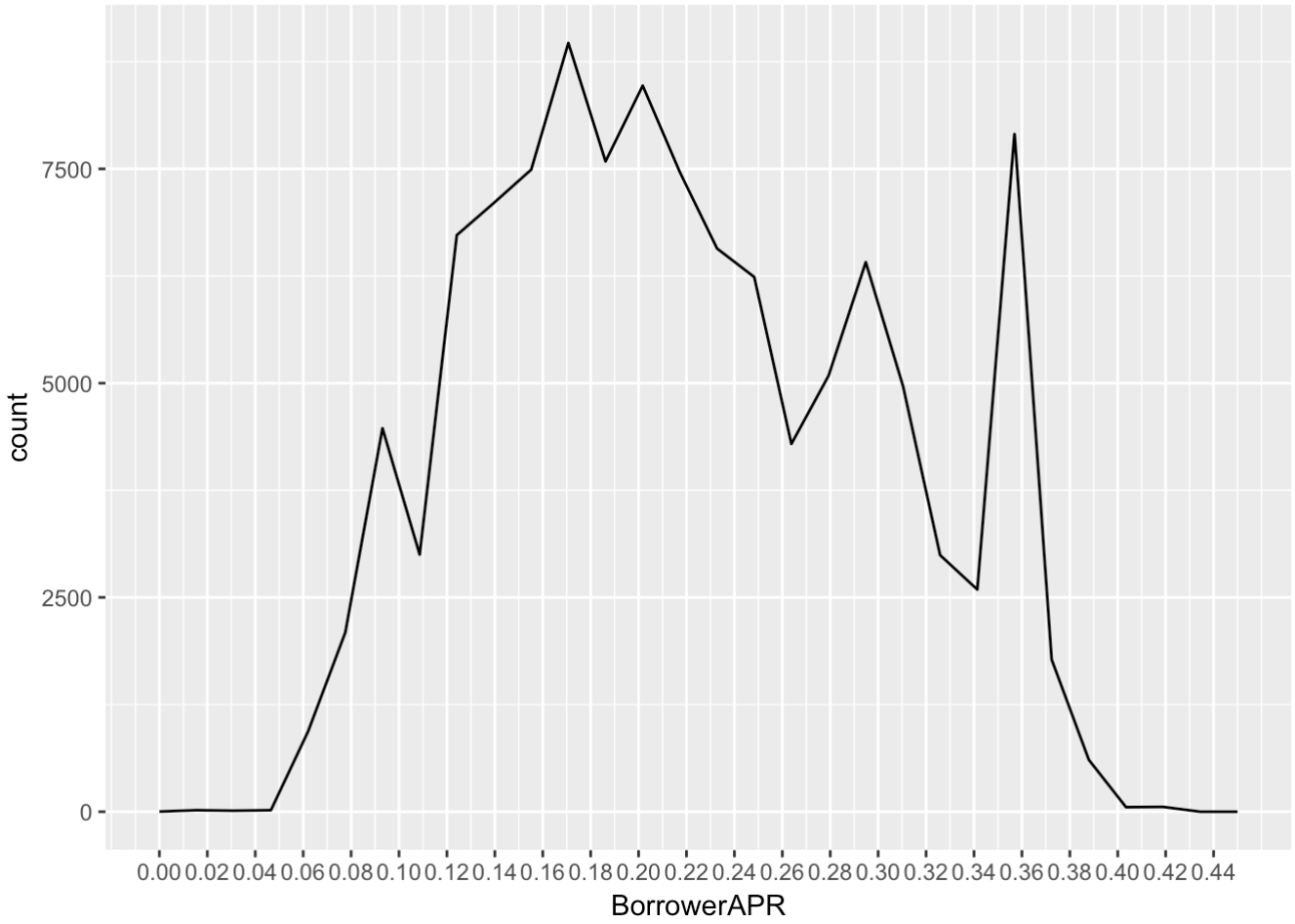


For the Borrower APR there is clear preference of the rate to be around 38% while the rest of the APR is fairly distributed. According to summary stats, the min APR is 6.5% and max is 51.2% and we will see how it is distributed.



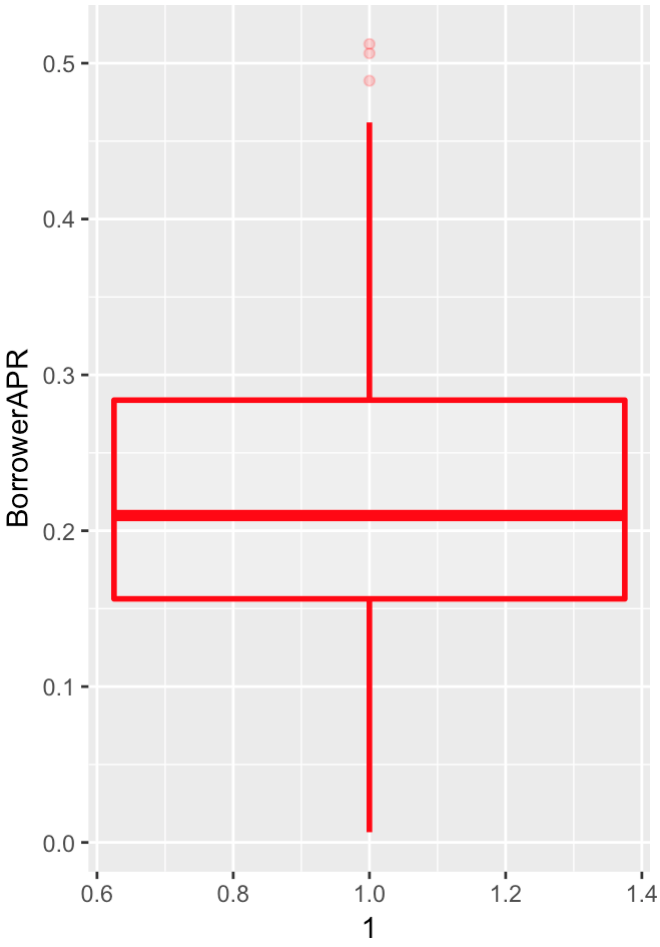

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## 0.00653 0.15630 0.20980 0.21880 0.28380 0.51230      25
```

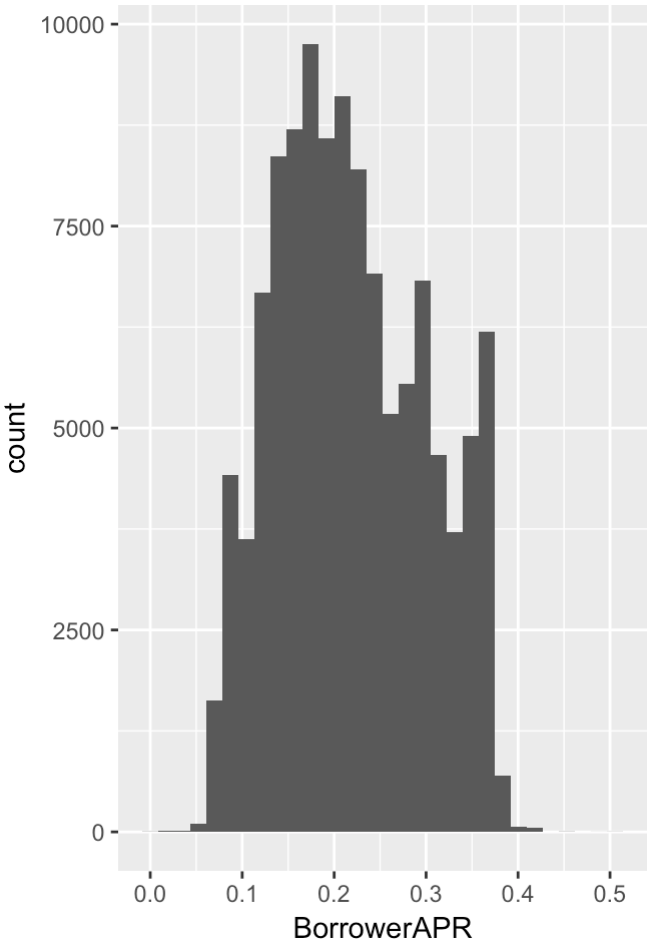
There is another way to look at data and to determine what is the most frequen rate, apparently it is not 36% but rather 17% (see below)



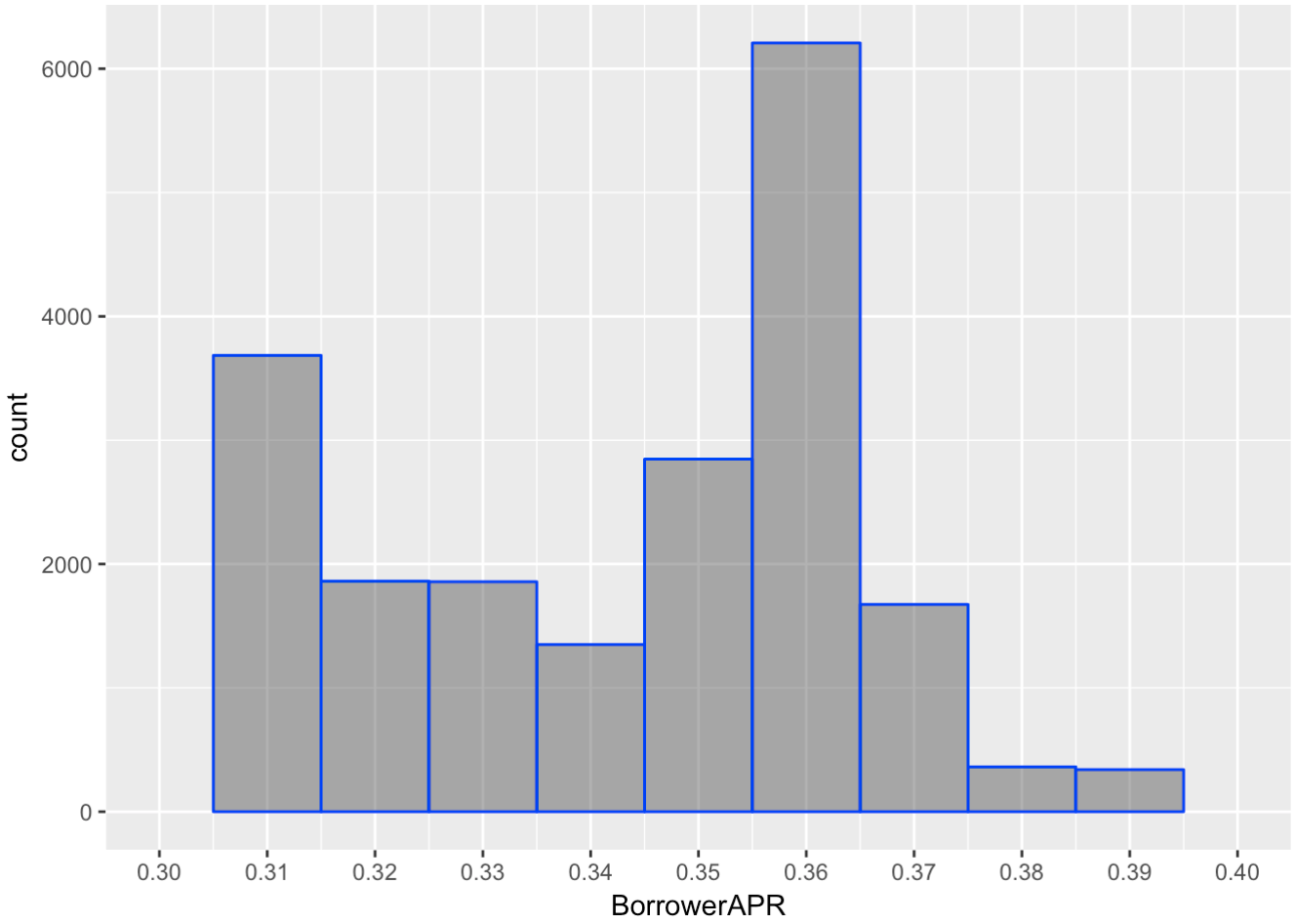
Outliers for Borrower APR data

There are few outliers for Borrower APR at more than 45%. However, the data is more or less evenly distributed around median at 20 - 25% rate.





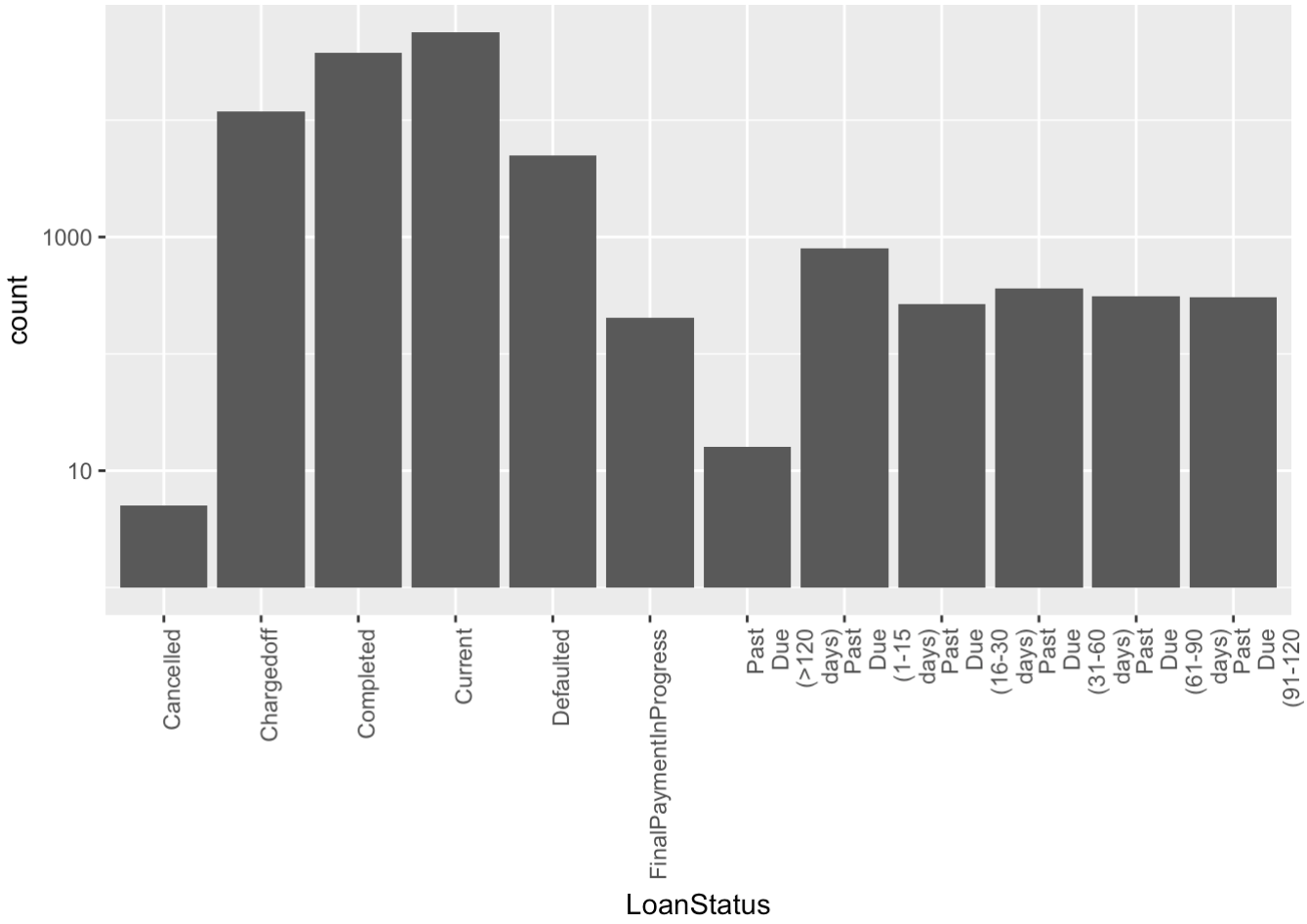
Now we will focus is on the amount of borrowing between 30 and 40%, where it is apparent that the 36% rate is the most prevalent.



Univaraibale analysis of other variables

Loan status data

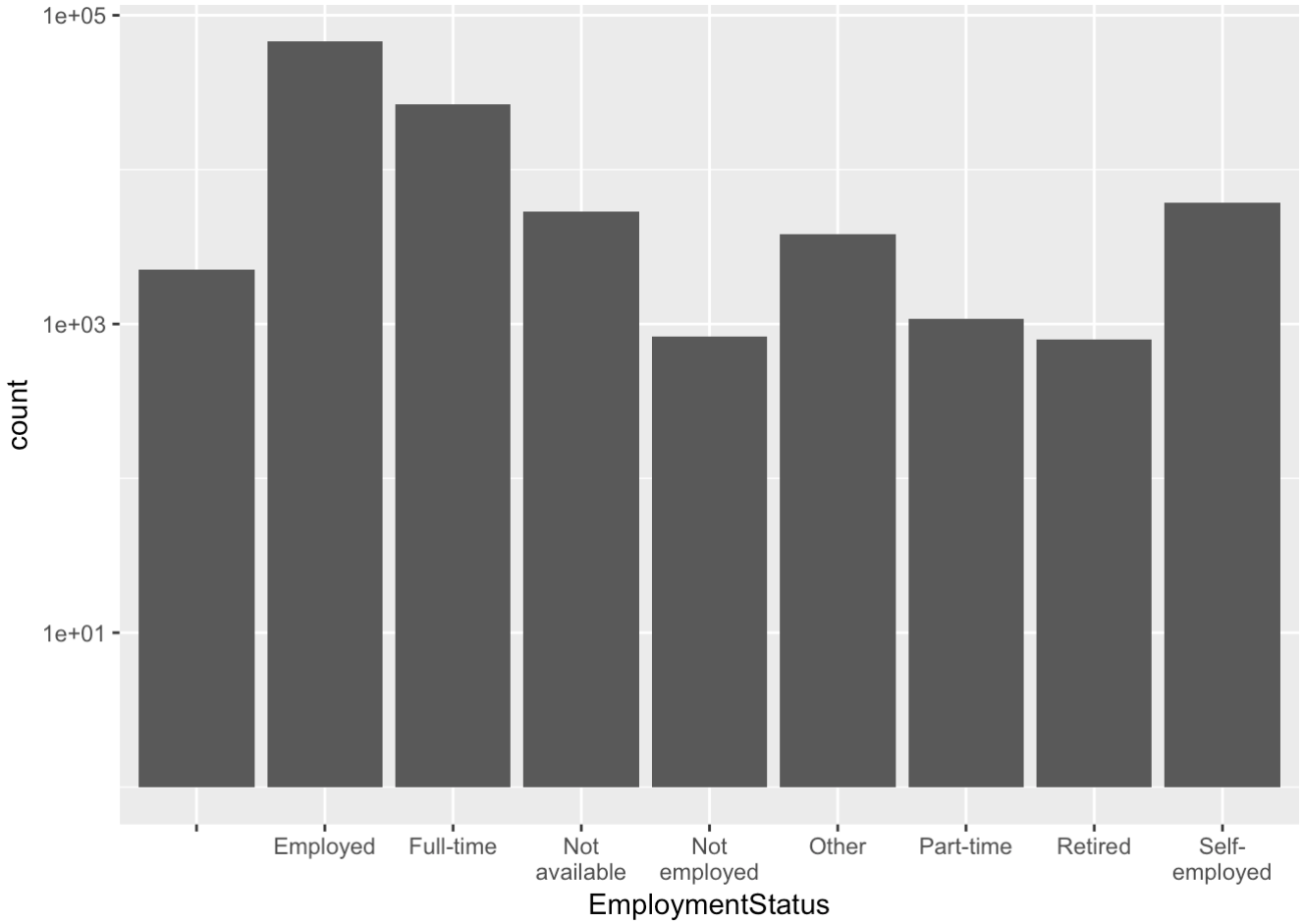
There are only few loans which are cancelled - less than 0.004% and past due less than 2%. Majority of loans are completed, current, charged.




```
##             Cancelled             Chargedoff             Completed
##            0.00438839            10.52511476           33.41671274
##               Current               Defaulted FinalPaymentInProgress
##           49.65551138             4.40418828             0.17992399
##  Past Due (>120 days)  Past Due (1-15 days)  Past Due (16-30 days)
##            0.01404285            0.70740848            0.23258467
##  Past Due (31-60 days)  Past Due (61-90 days) Past Due (91-120 days)
##            0.31859712            0.27471322            0.26681412
```

Employment status

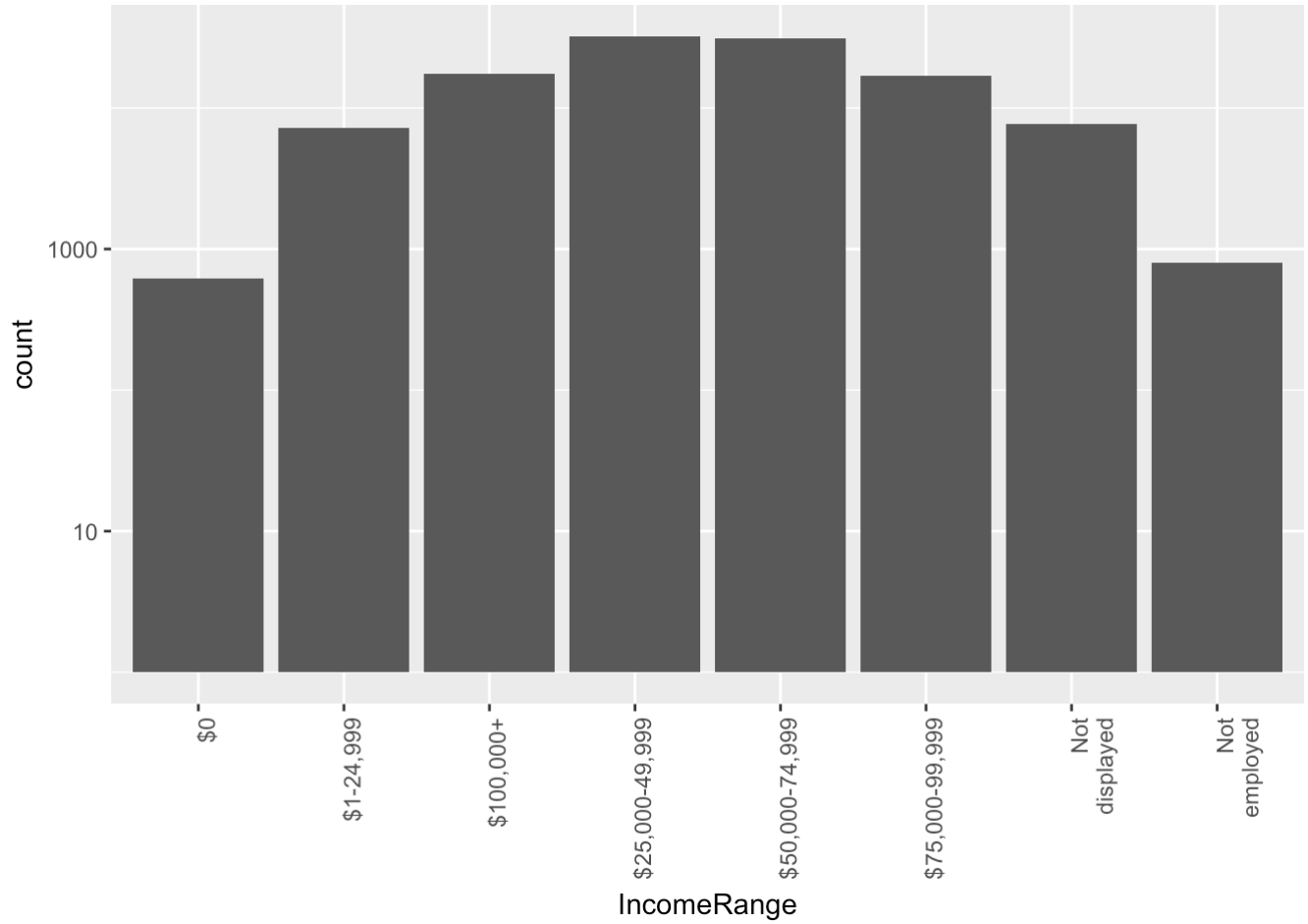
Vast majority of borrowers and applicants are employed. However, there is some labels data missing in Employed category for almost 2% of applicants.




```
##                   Employed     Full-time Not available  Not employed
##     1.9791639    59.0870393    23.1312041     4.6929443     0.7328611
##         Other     Part-time       Retired Self-employed
##     3.3404425     0.9549137     0.6977540     5.3836769
```

Income range

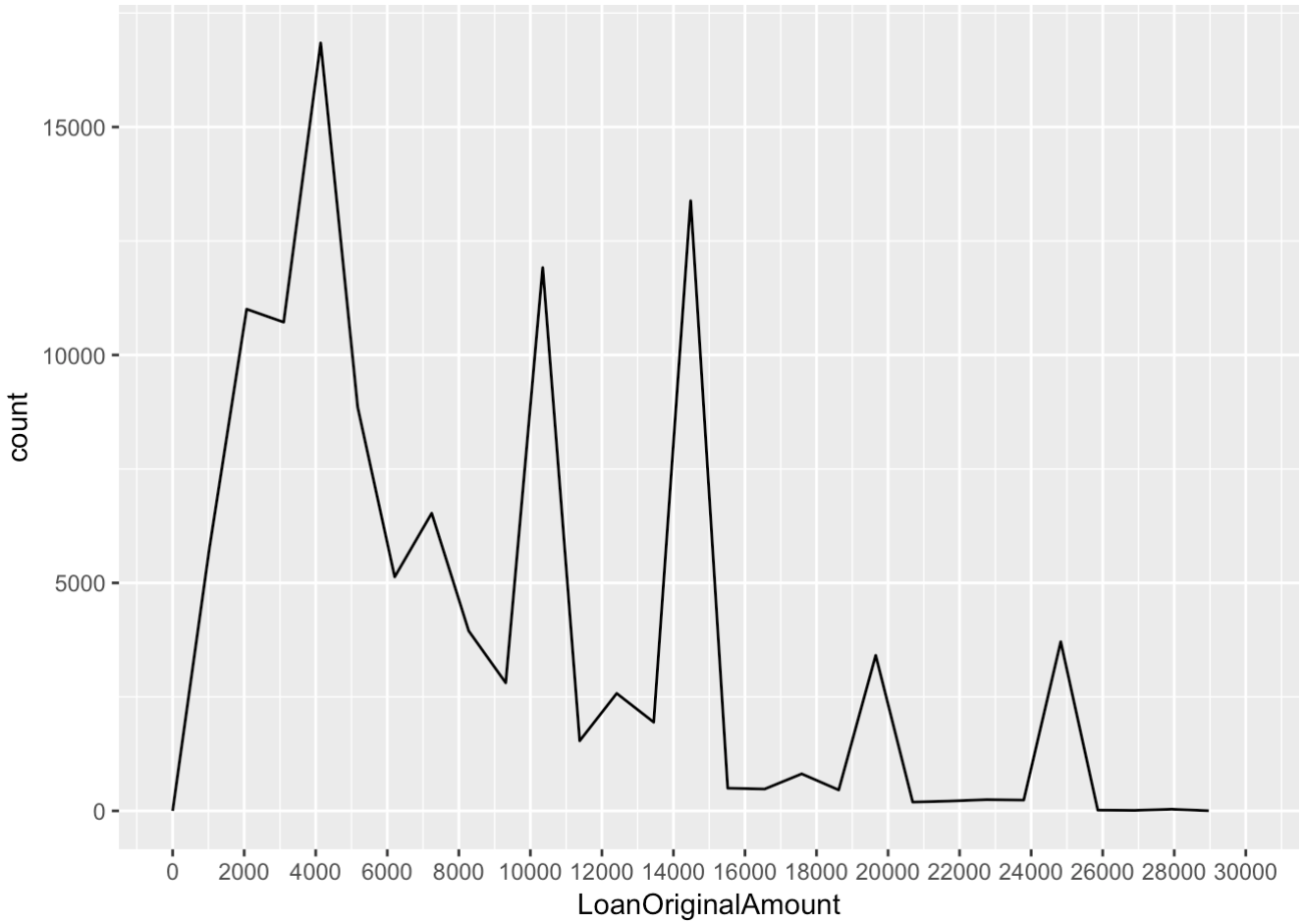
Majority of applicants have income range from 25 to 75K. Seems like a normal distribution of income. Less than 1.5% showed not employed or no income while not displayed is for 6.8% of applicants.




```
##             $0      $1-24,999      $100,000+ $25,000-49,999 $50,000-74,999
##      0.5450380      6.3842299     15.2163037     28.2542107     27.2519024
## $75,000-99,999  Not displayed   Not employed
##     14.8468013      6.7941055      0.7074085
```

Loan original amount

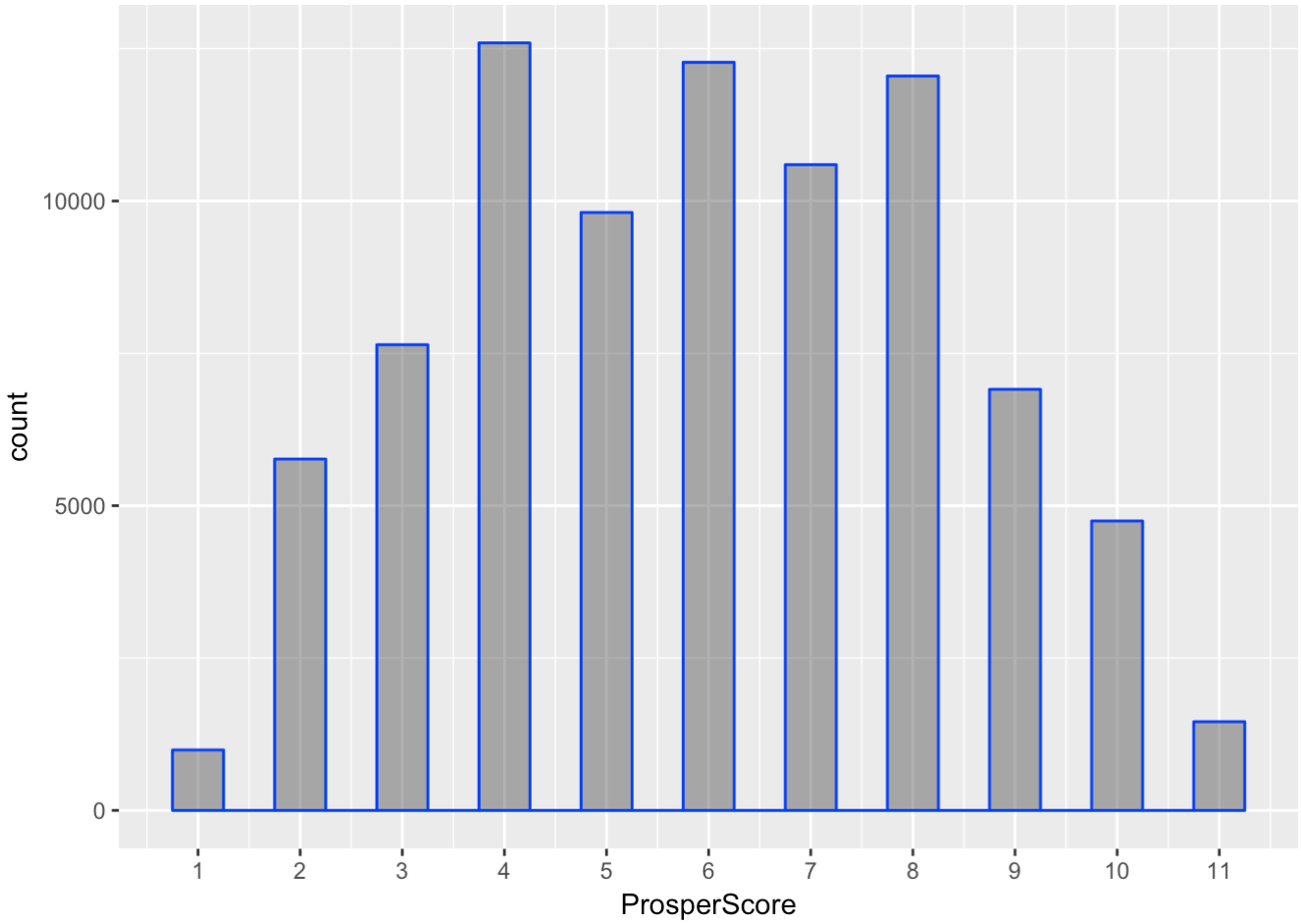
Majority of loans granted are between 2 and 5K, also around 11K and 15K. This is clearly a cashloan as the amounts are pretty small. The summary statistics (see below graph) is pretty disbursed with a difference between mean and median and minimum and maximum values. But the overall the loan amounts are not very big and as such, they appear to be normal.




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

Prosper score

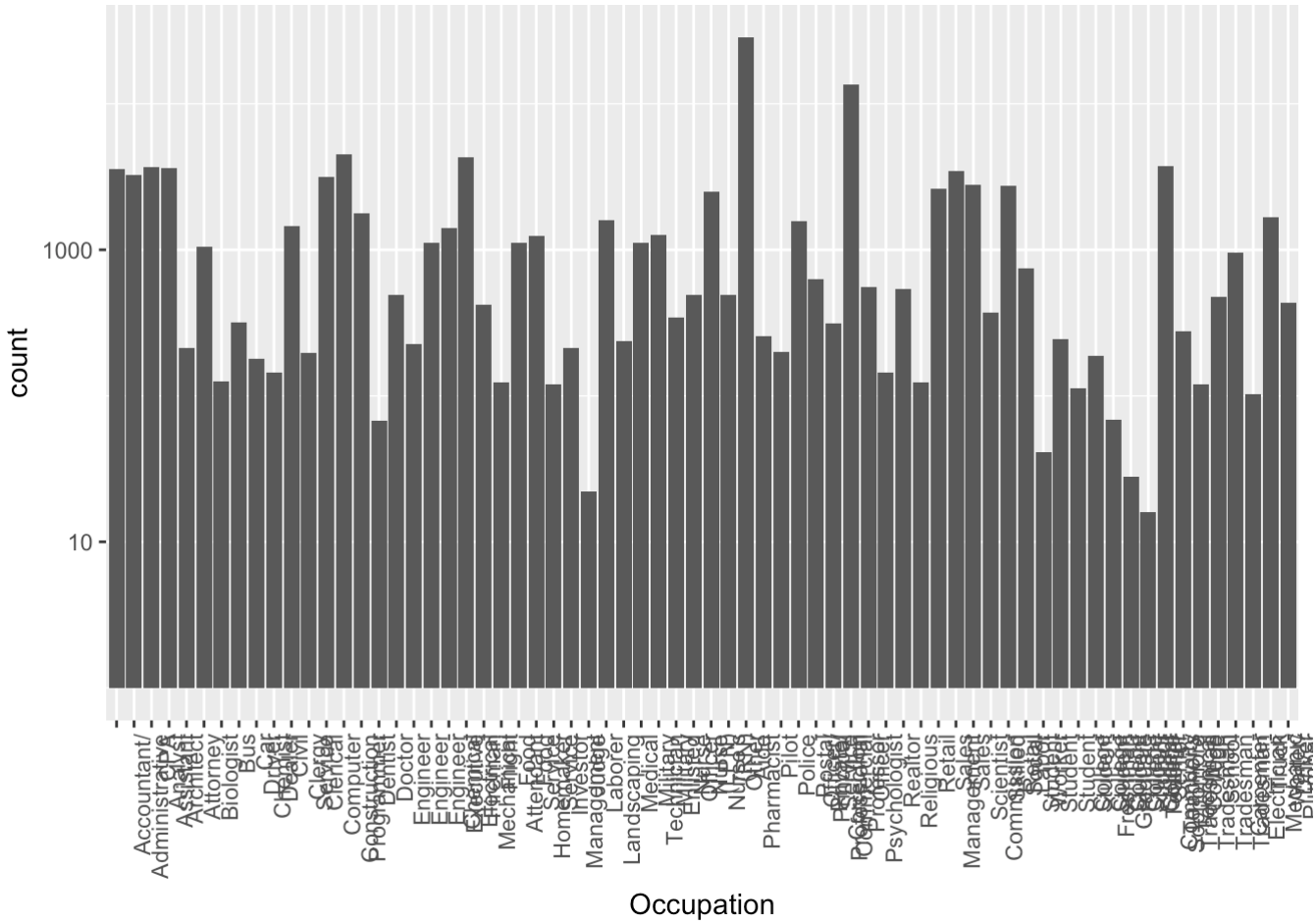
Prosper score is distributed around mean and median data, majority is scored at 4, 6 and 8. Around 30K data points are at NA, whichi is around 27% of all applicants. It is a rather significant number.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00    4.00    6.00    5.95    8.00   11.00   29084
```

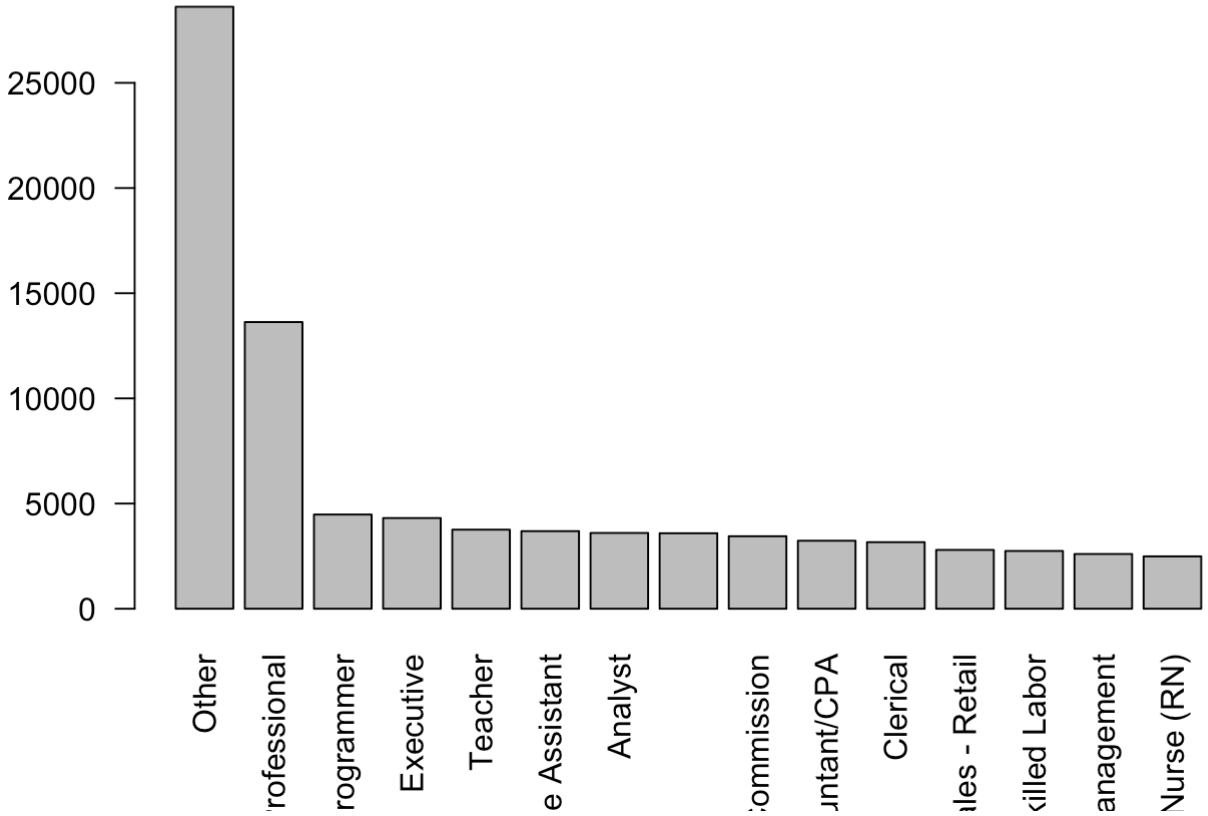
Occupation

It is hard to read the occupation variable as there 68 levels. So we will need to manipulate a data.



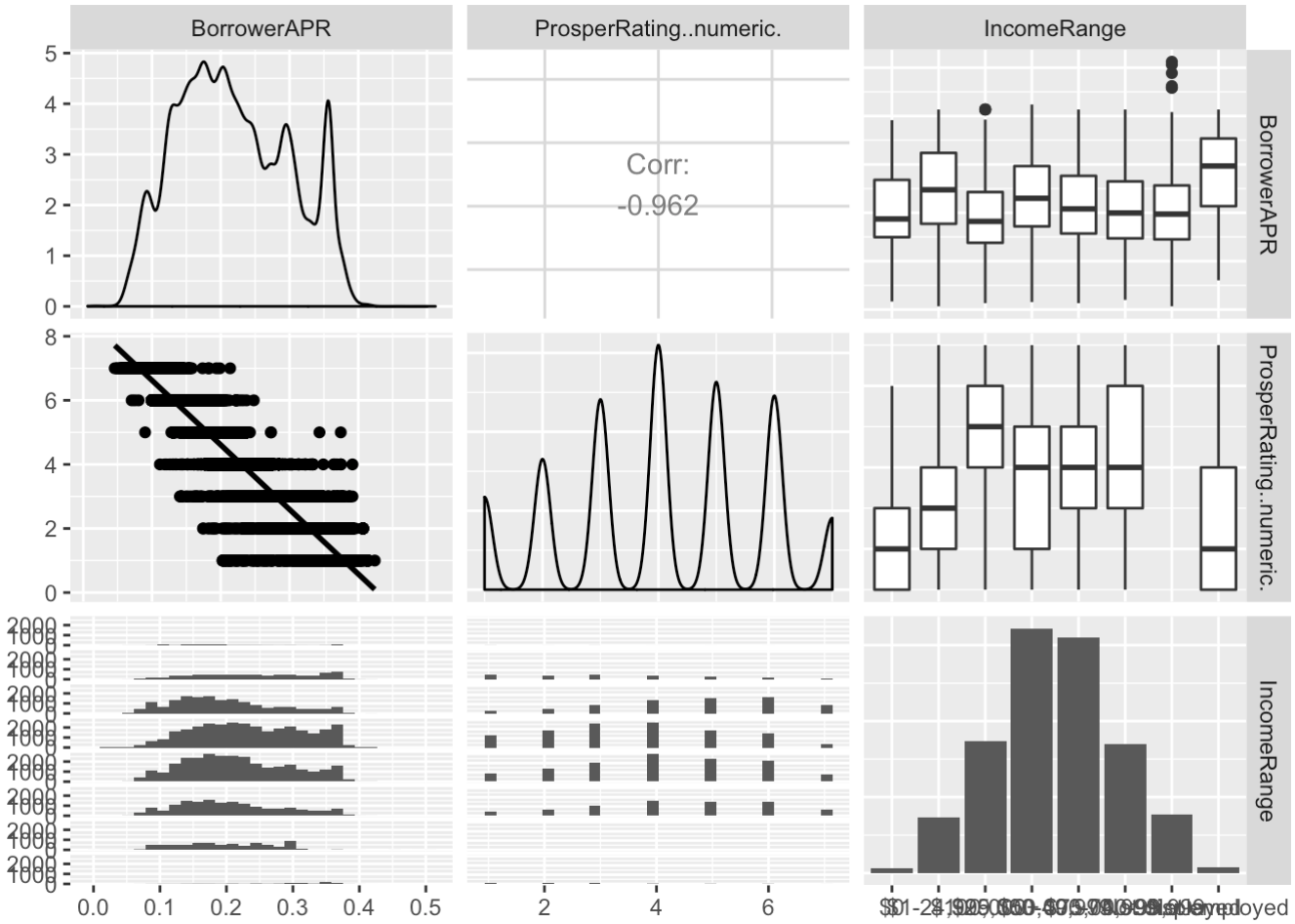
Top occupation

As we can see the top occupations amongst applicants are Other, Professional, Programmer, Executive etc.



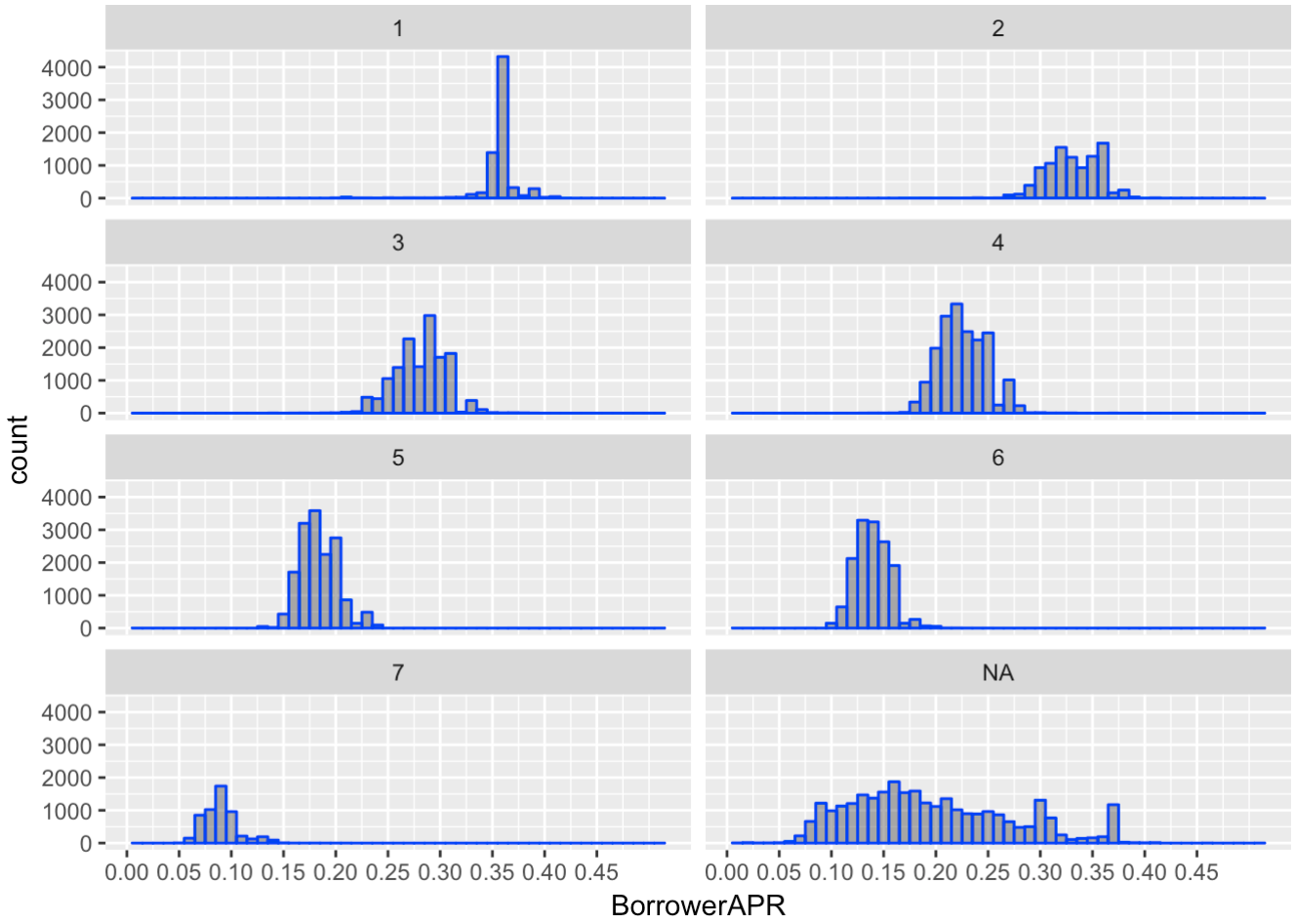
Initial research on loans - Bivariable

Below is an attempt to figuring out dependencies between two variables by plotting few variables against one another. There is a dependency between Borrower APR and Prosper Rating both in graph and correlation, and there is an interesting distribution of Borrower APR by Income range. We will investigate those further.

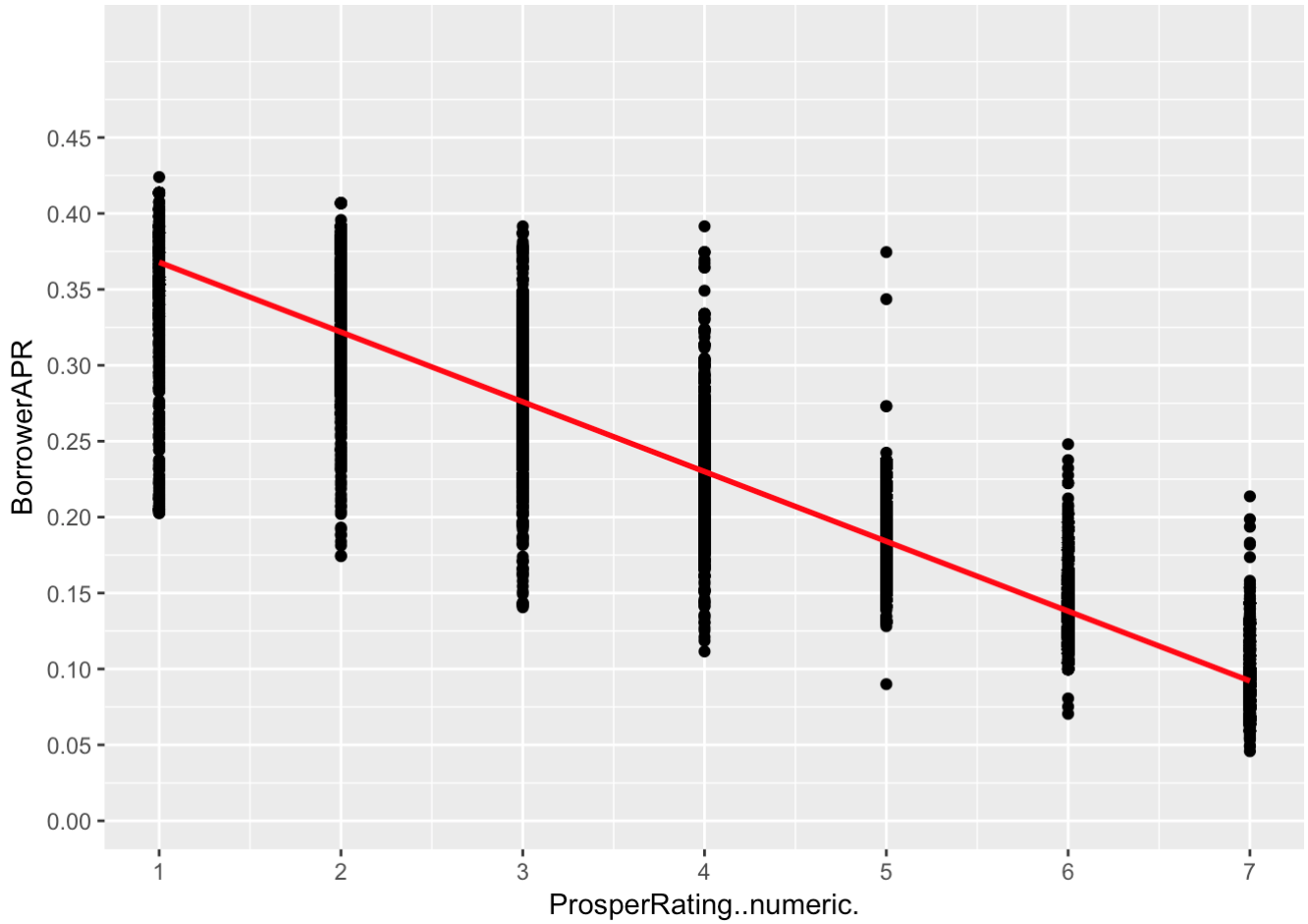


Bivariant analysis - APR and Ratings

Now it's a time to combine to variables - APR and Ratings. Below are histograms for Borrower APR divided by ProsperRating. Most widely distributed are loans APR within ratings 4 and 6 while the narrow distribution is for rating 1. Probably, borrowers with this rating are getting the maximum possible rate or not getting the loan at all.

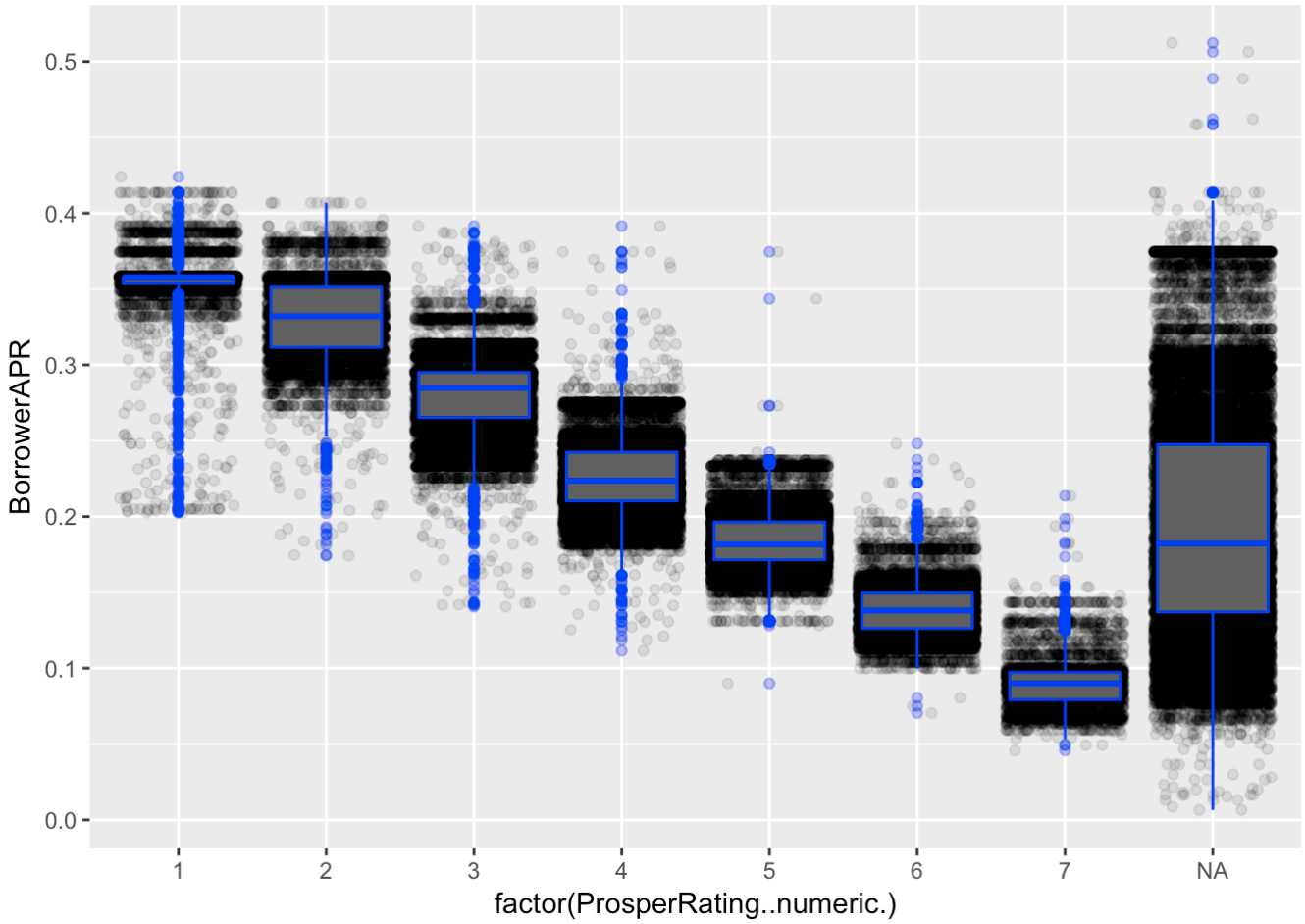


We are exploring the dependency of ProsperRating vs BorrowerAPR and clearly the higher the rate the lowever is APR. The correlation line proves it as well as the correlation coefficient which is -0.96




```
## [1] -0.9621513
```

Boxplot will help to explore further this dependency. As we can see from below, median APR are surely decreasing with the Rating increase. There is wide variety of rates in the lowest rating category 1. However, overall, borrowers with rating 1 can obtain loans with the same APR as borrowers all ratings, even at the highest 7 rate. The biggest variance is in NA data - clearly APR does not only depend on Prosper rating and many loans have been given to applicants with Prosper rating of NA. Overall - median of NA is at rating 5 level.

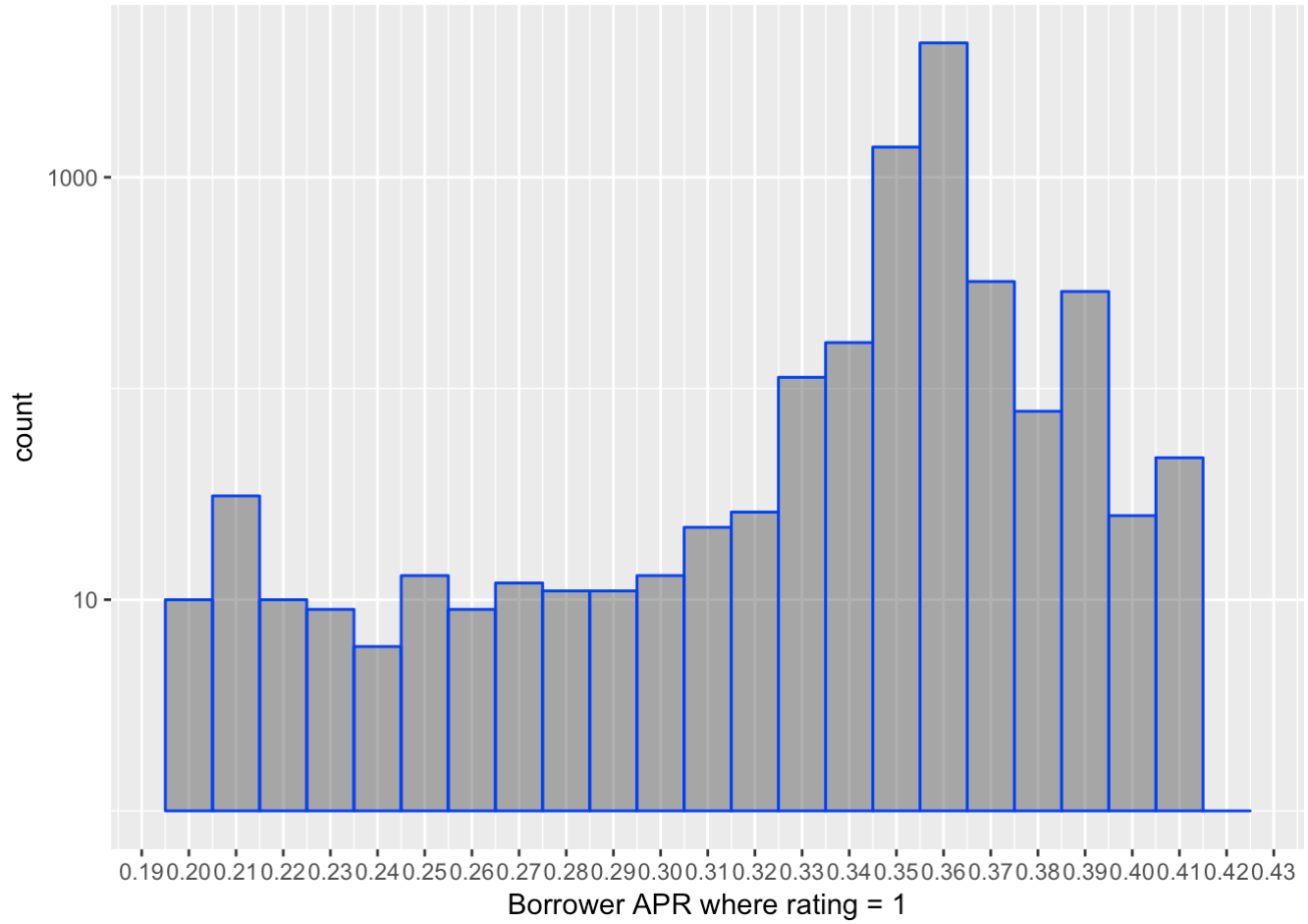


Loan amount vs income range

We can see that the majority of loans are given to borrowers with income range between 75 -100K, with very small amount of borrowers with 100K+ income range. It is logical and higher earning borrowers probably don't need cash loans or have better deals at traditional banks. We will leave this exploration as it is.

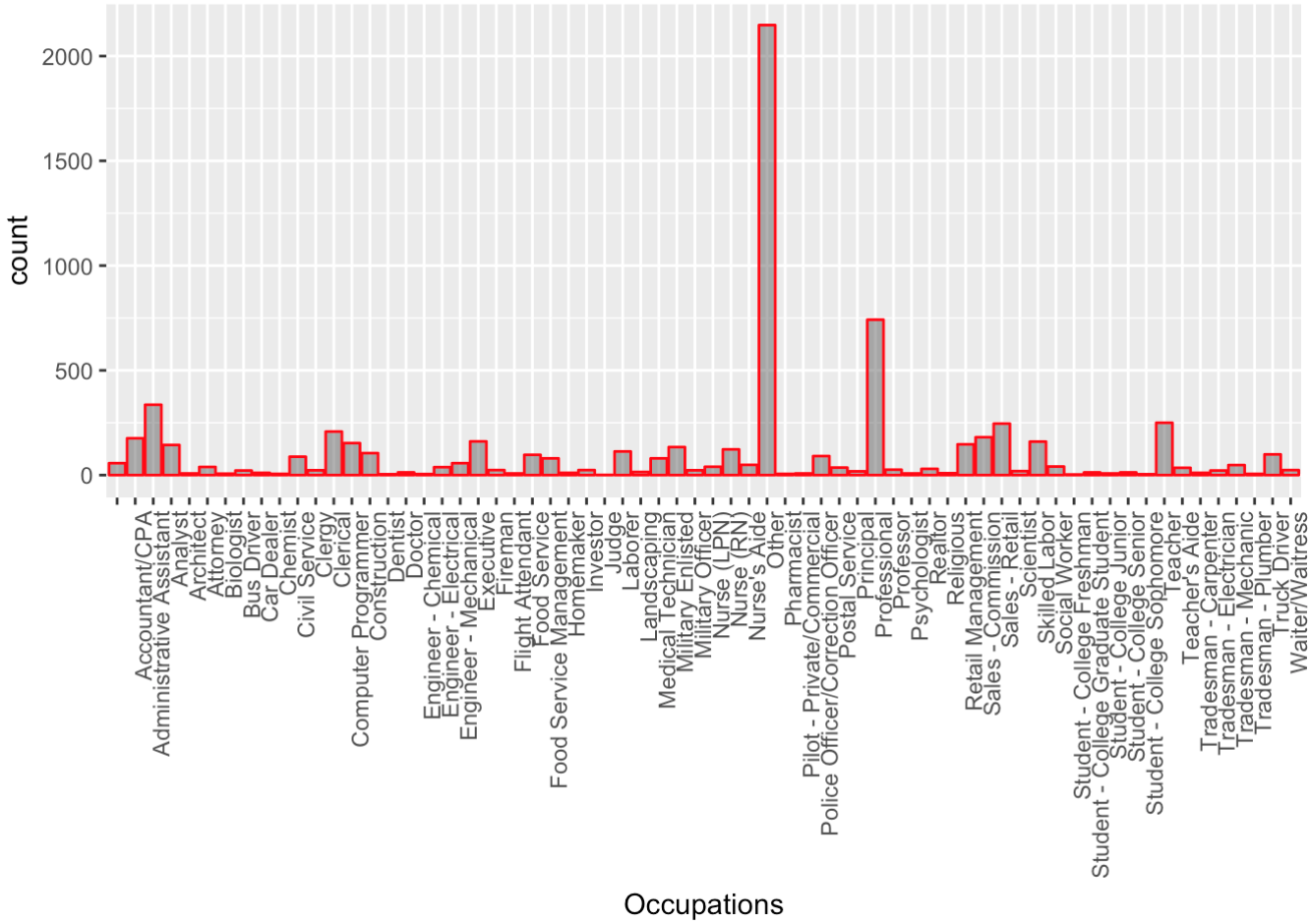
Rating 1 investigation

What is going on with rating 1, what states, professions, income level and other characteristics it represents? First we will create a table to focus on some of the most interesting variables. Then we will build a APR histograph on a log scale to show all the values, as APR rate for rating



Occupations in Rating 1

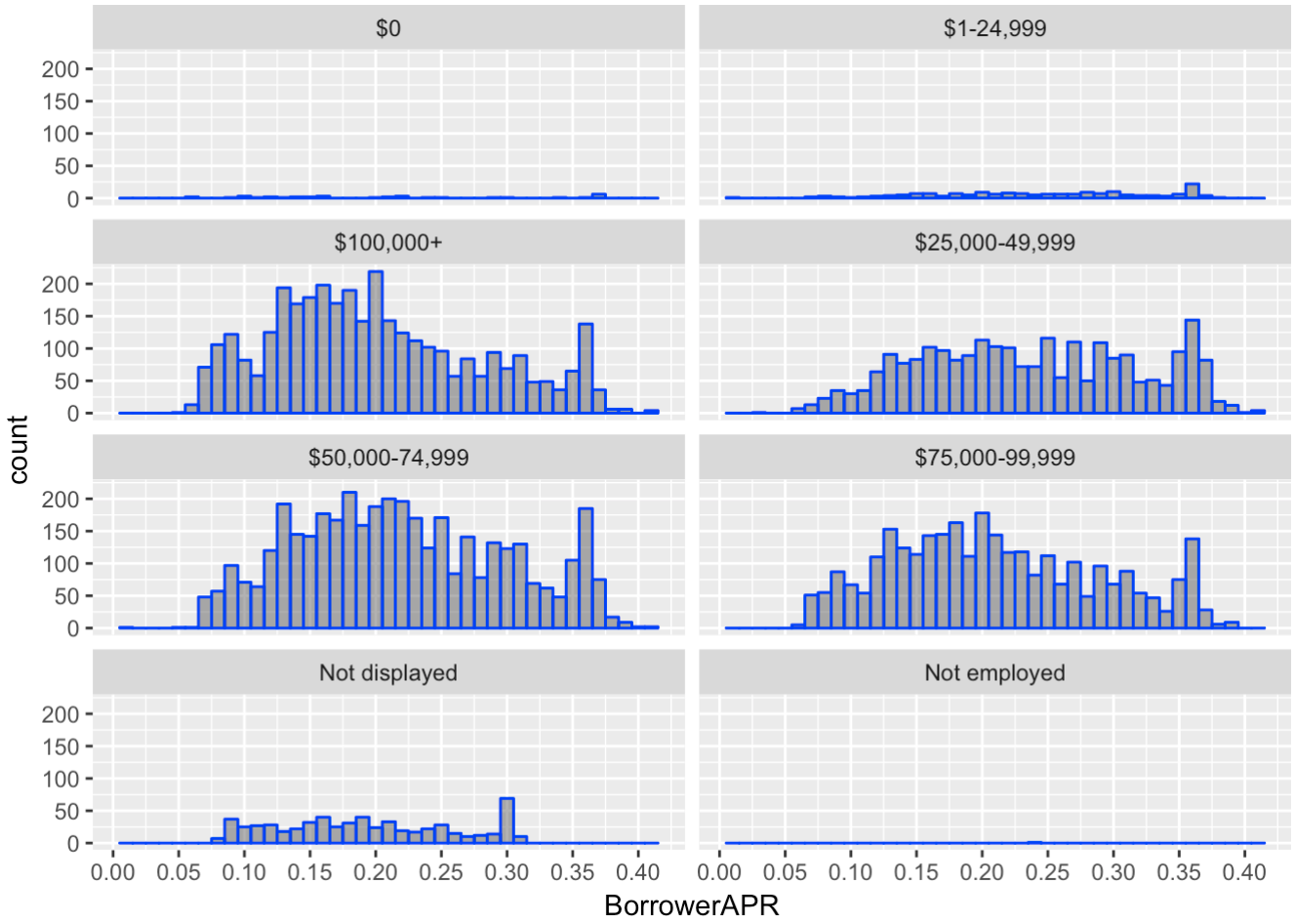
We will explore the most frequent occupation which were rated the lowest rating 1. Clearly there is a leader.



Leader is unfortunately “Other”, so we will get rid of that.

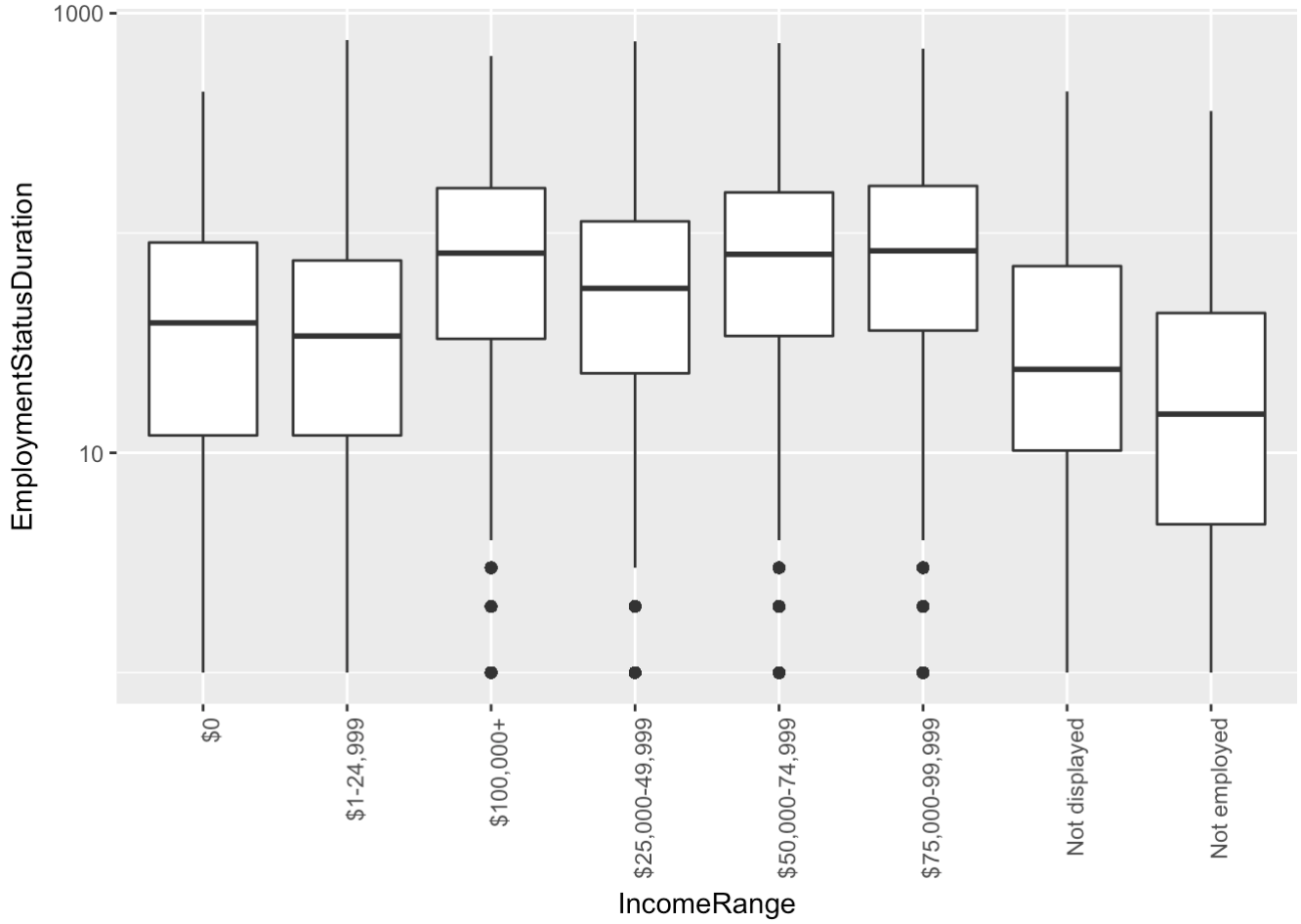
The top occupation in rating 1 is “Professional” followed by Adminstrative Assistant and Teacher.

I want to explore further the Occupation “Professional” as it very ambiguous. As you can see below majority of this occupation earns from 50 to above 100K. It means that this occupation has a varied pay based probably on firm and experience.



Income range vs Employment Status Duration

I want to see the distribution of income between Income Range and Employement Status duration and there is not really some meaningful distribution - if you see the States table, you will see the employment duration is much higher for income level $50K+

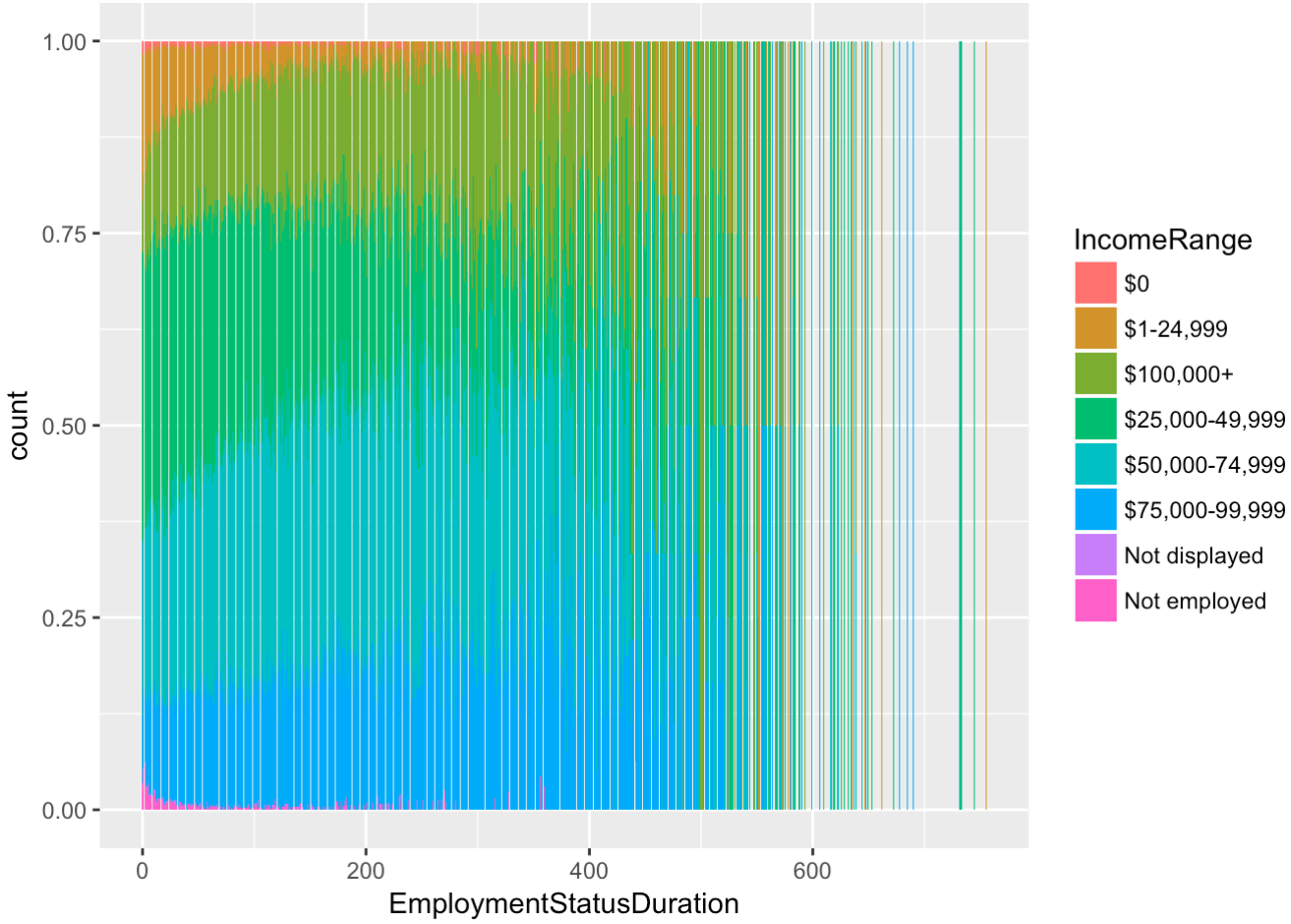


Stats for income range vs Occupation duration

Applicants with salaries above 50K stay much longer than those with lower salary.

```
##                $0 $1-24,999 $100,000+ $25,000-49,999 $50,000-74,999
## min             0       0.0         0              0             0
## lower quartile 10      10.5        33             21            33
## median         37      32.0        80             55            79
## upper quartile 87      73.0       159            111           152
## max           198     166.0       348            246           330
##               $75,000-99,999 Not displayed Not employed
## min                        0           0.0            0
## lower quartile            35           7.5            3
## median                    82          21.0           13
## upper quartile           163          67.0           41
## max                      355         152.0           96
## attr(,"class")
##        $0
## "integer"
```

I have done the above but in percentage terms. There is still not clear dependency of higher salary vs Employment Status Duration.



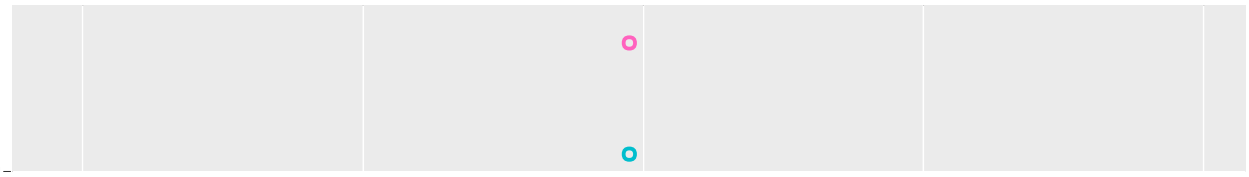
New table with Occupation data

Now I am going to create a long table with the following colums - Occupation, median_BorrowerAPR, mean_ProsperRating..numeric., median_CreditScoreRangeUpper. By aggregating data to tables and then merging the data. I would like to explore dependencies of different variables on Occupation.

```
## 'data.frame':    68 obs. of  4 variables:
##  $ Occupation   : Factor w/ 68 levels "","Accountant/CPA",..: 1 2 3 4 5 6 7 8 9 10
 ...
##  $ CreditScore  : num  679 719 699 719 719 719 719 699 699 719 ...
##  $ ProsperRating: num  4 4 4 5 5 5 4 4 4 4 ...
##  $ BorrowerAPR  : num  0.199 0.2 0.237 0.189 0.184 ...
```

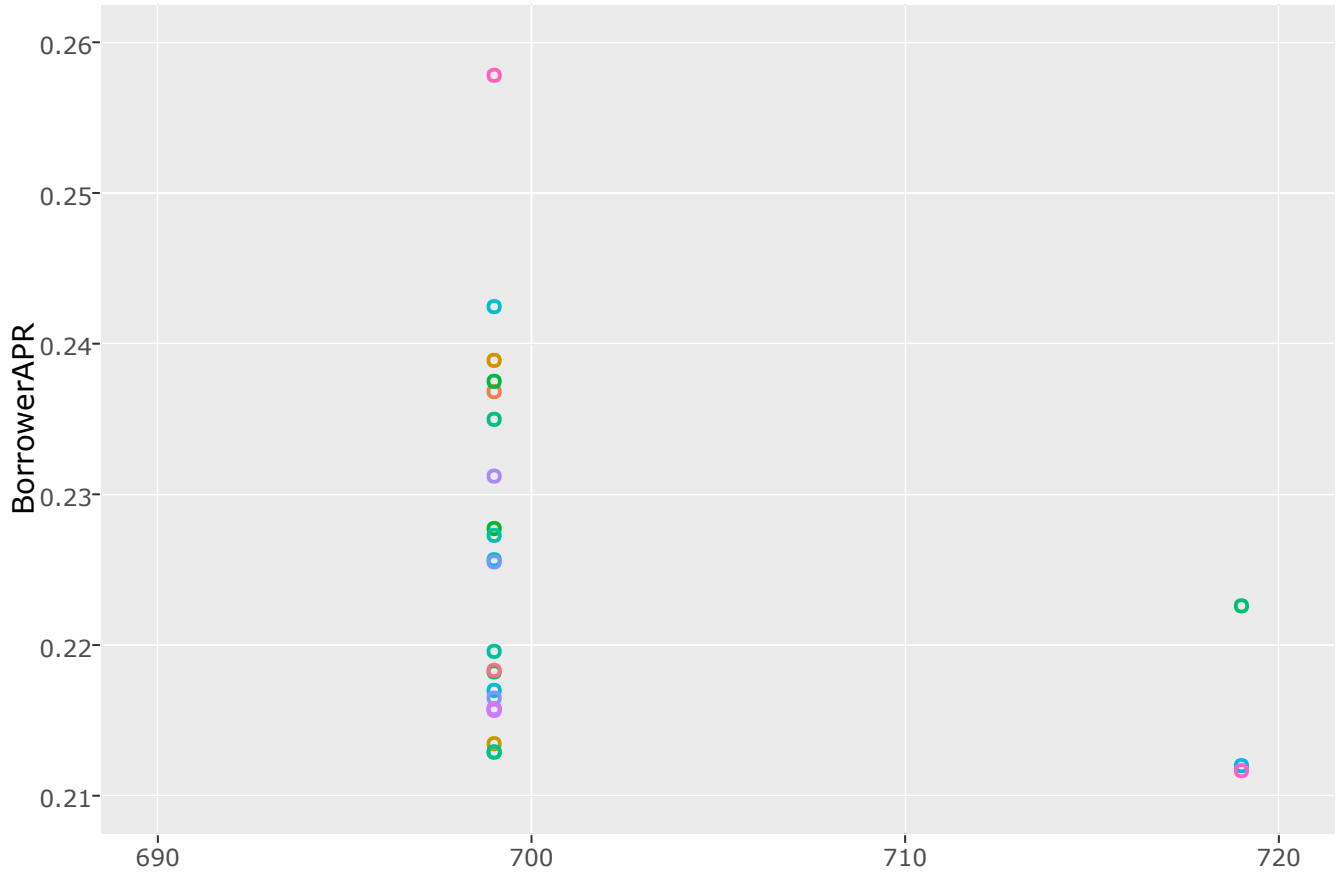
APR vs CreditScore vs Occupation

Majority of occupations are scored under 700 and 720 where the highest APR was granted to Teacher's and Nurse's Aide. To see the profession, please point with a mouse at a dot and the description will appear. Students had lower score but lower APR as well, reflecting on their good potential for lenders.



Focusing on high APR and high Credit scores

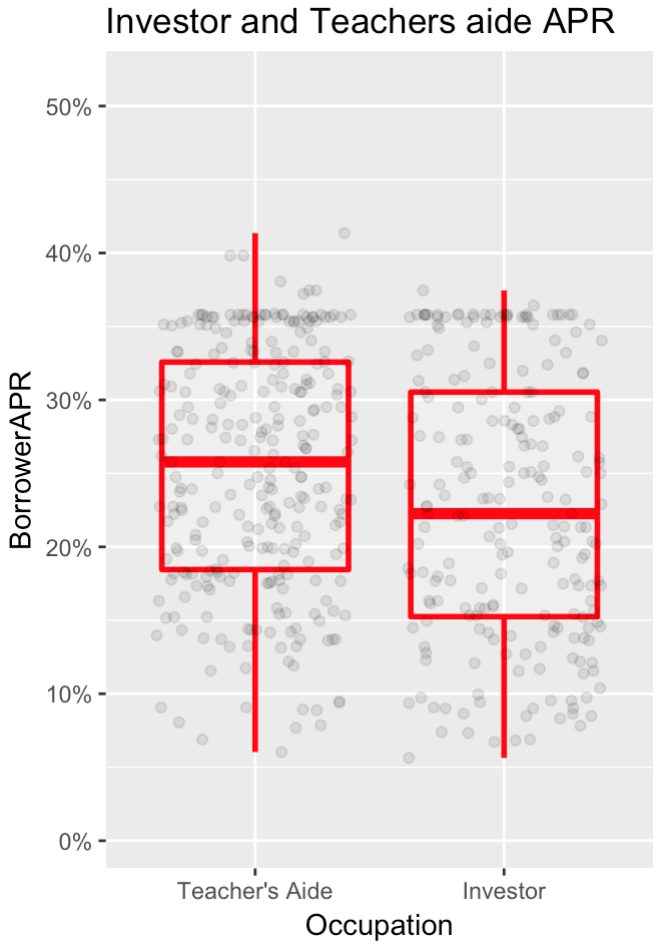
Surprising, an occupation Investor had a pretty high Credit Score but high APR as well. This indicates that even thought this person must be wealthy and have good credit score, probably his or her other parameters were risky. Based on the below graph I selected to investigate Occupations with relatively high Credit score but high APR as well like Teachers Aide and Investor.

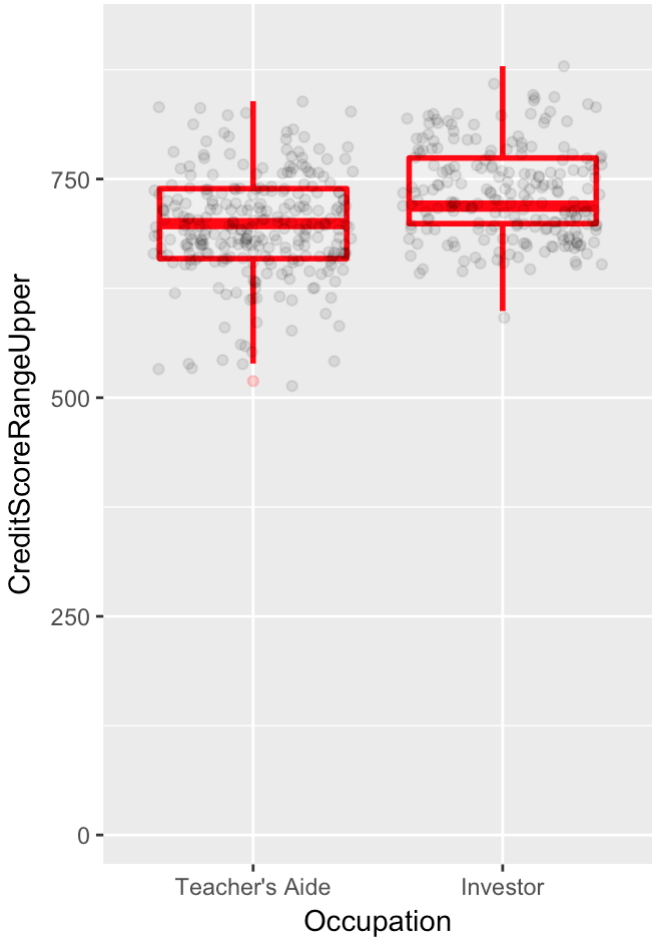


CreditScore

Further exploration of some of occupations vs Borrower APR

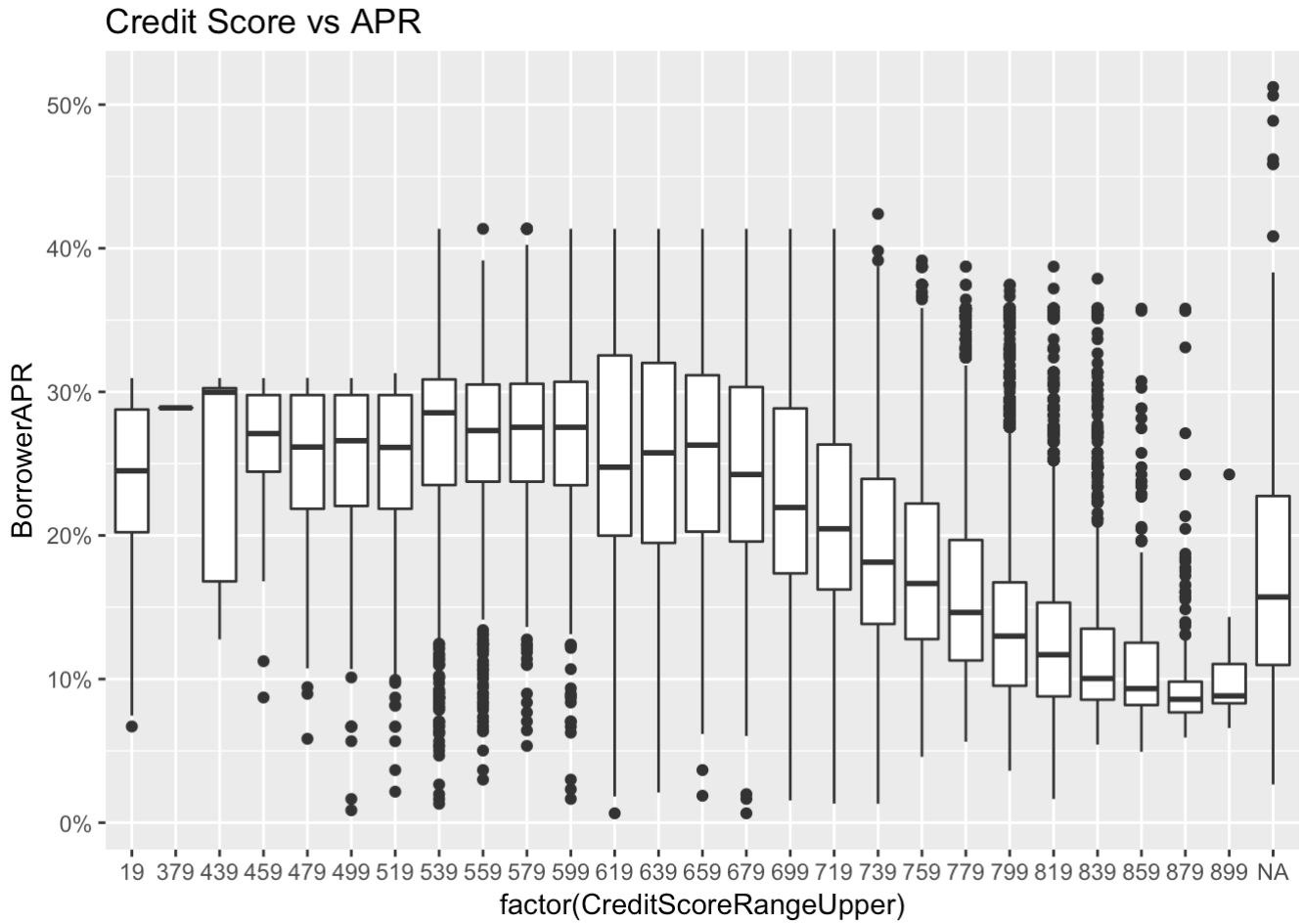
Occupations like Teacher's aid and Investor had high credit scores but high median APR as well. On average investors showed better Credit score and lower APR and Teacher's Adie but were not very much even though one would expect large differences between the both.





Credit score and APR correlation

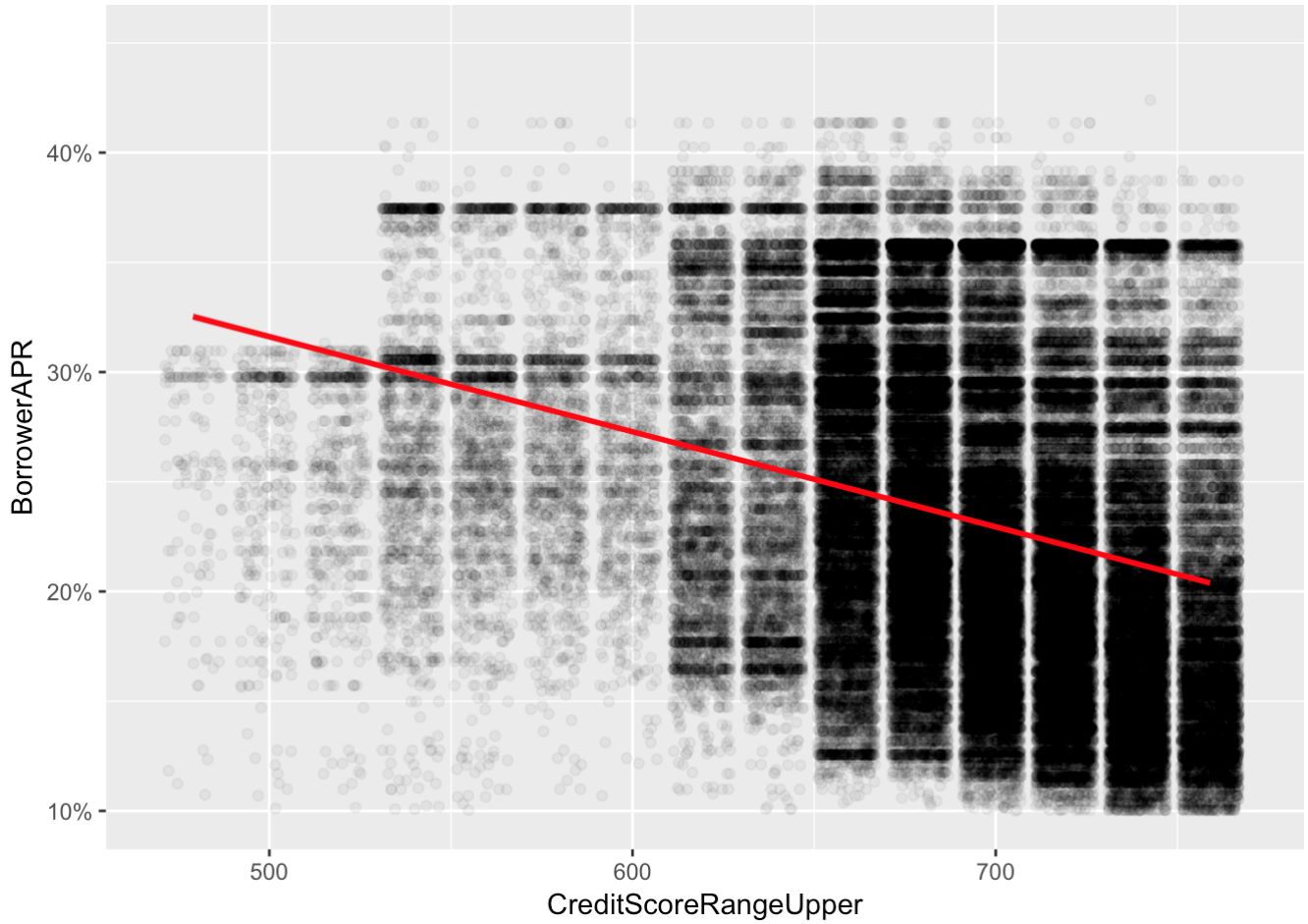
I want to see what is the correlation between - credit score and APR as below. There is low correlation (-0.43) between BorrowerAPR and Credit score which is somewhat surprising but given that it is a cash based loan, given to borrowers who exhausted other more traditional banking options it makes sense. The boxplot does not depict the picture in full, and I will use jitterplot with correlation line below.




```
## [1] -0.4297073
```

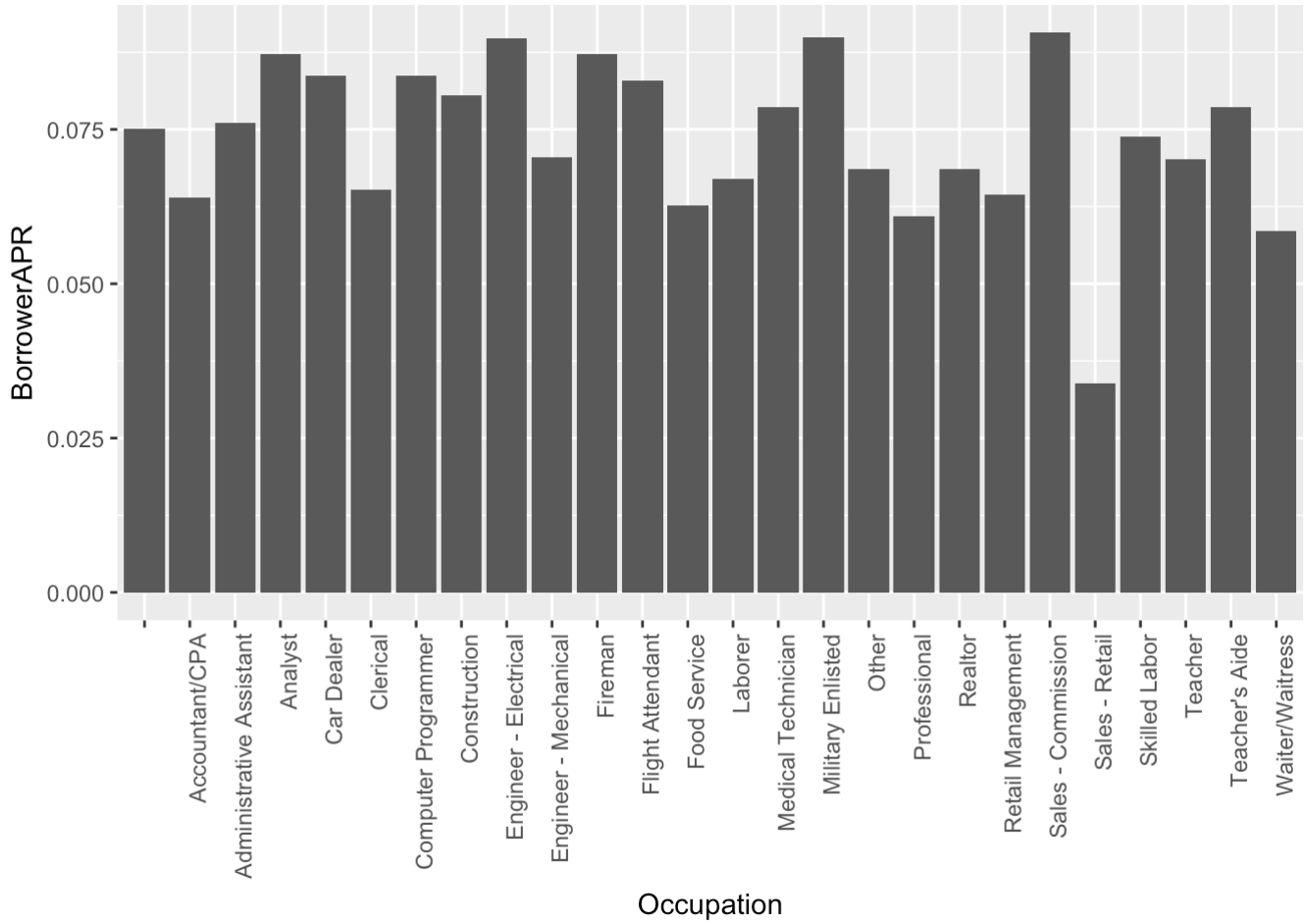
Credit score vs APR - different graph with correlation line

We can see the there is clear negative correlation but there is not much data points in credit scores below 600. And as discussed above correlation coefficient is not very high.



Occupations of low APR and low Credit Score

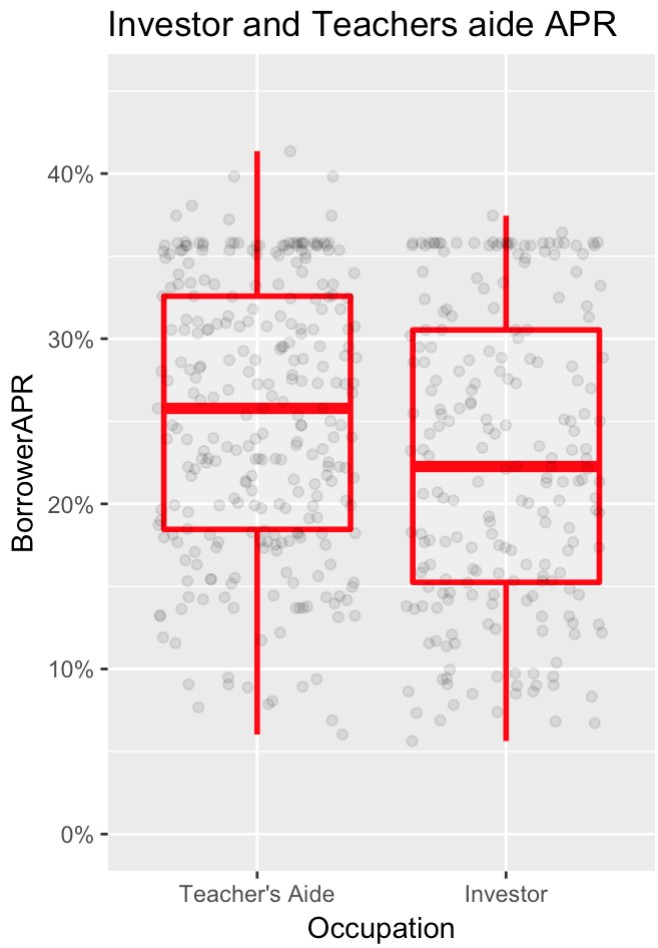
There are few occupations which get low APR and have low credit score suprisingly. Sales-reatil obtained the lowest APR.

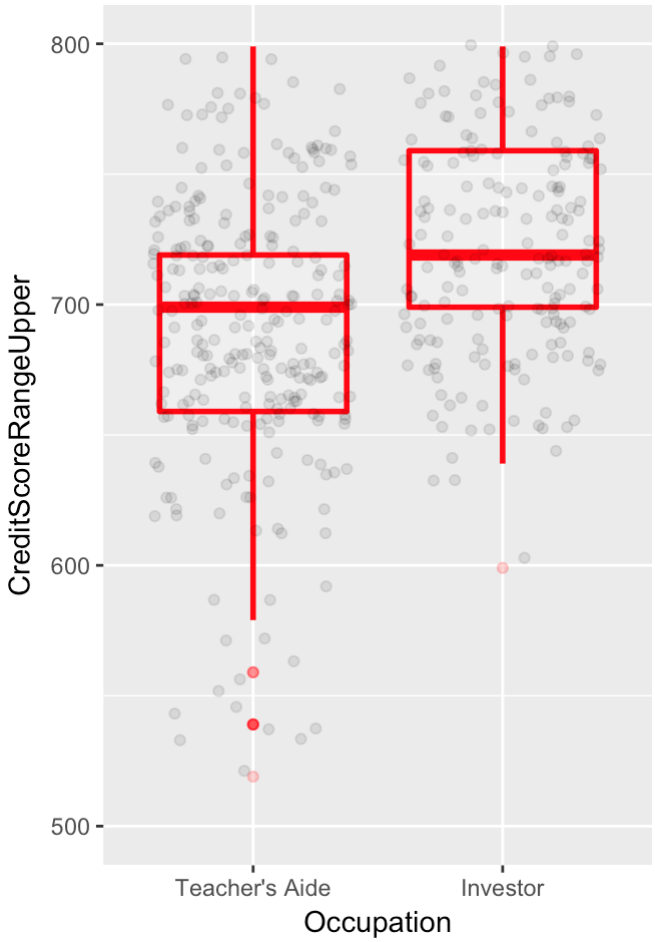


Final Plots and Summary

APR for drastically different occupations

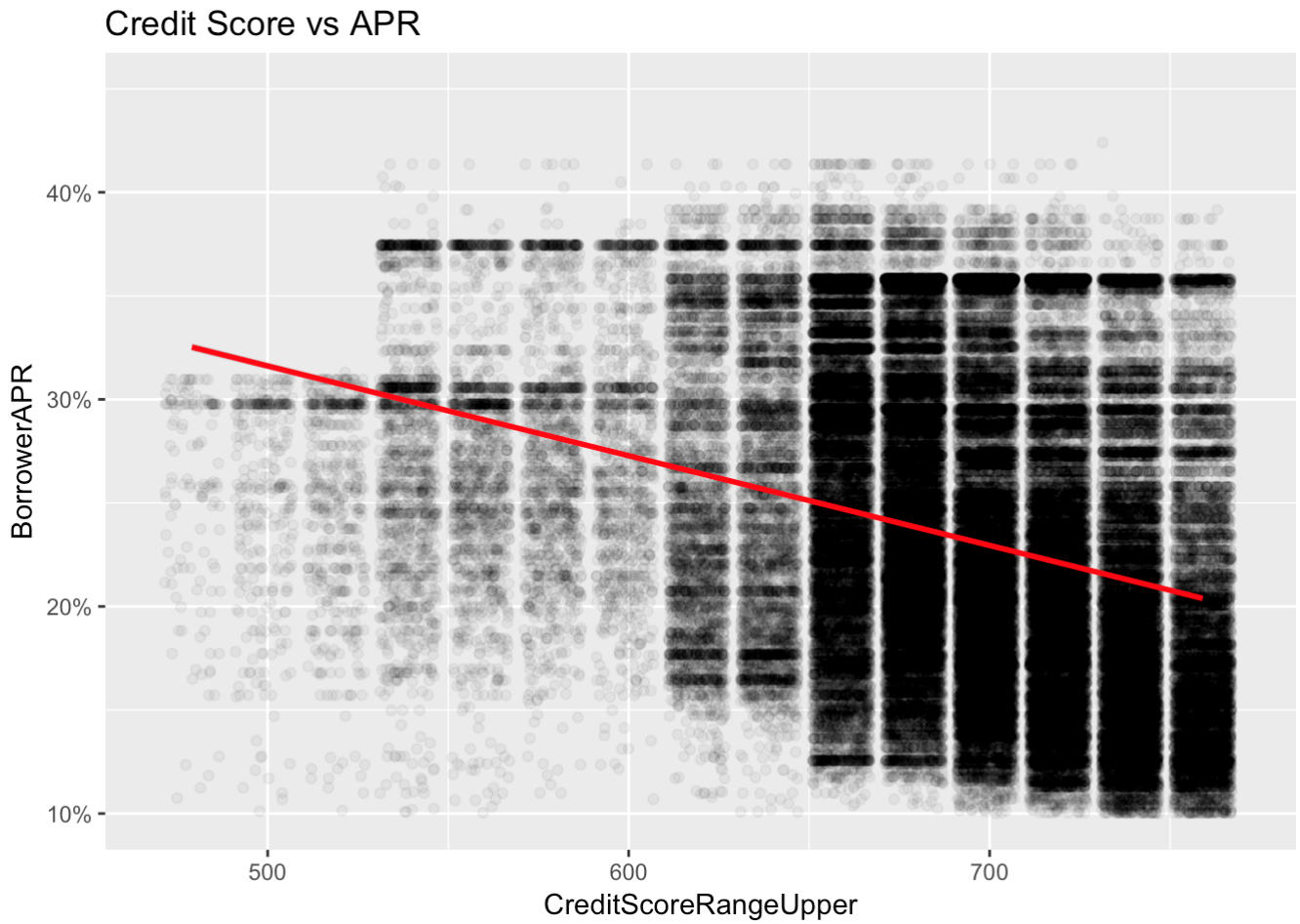
Occupations like Teacher's aid and Investor had high credit scores but high median APR as well. On average investors showed better Credit score and lower APR and Teacher's Adie but were not very much even though one would expect large differences between the both. These kind of data makes sense for relatively small loans, where an applicant either needs a quick cash or/and exhausted banking options.





Credit score vs APR

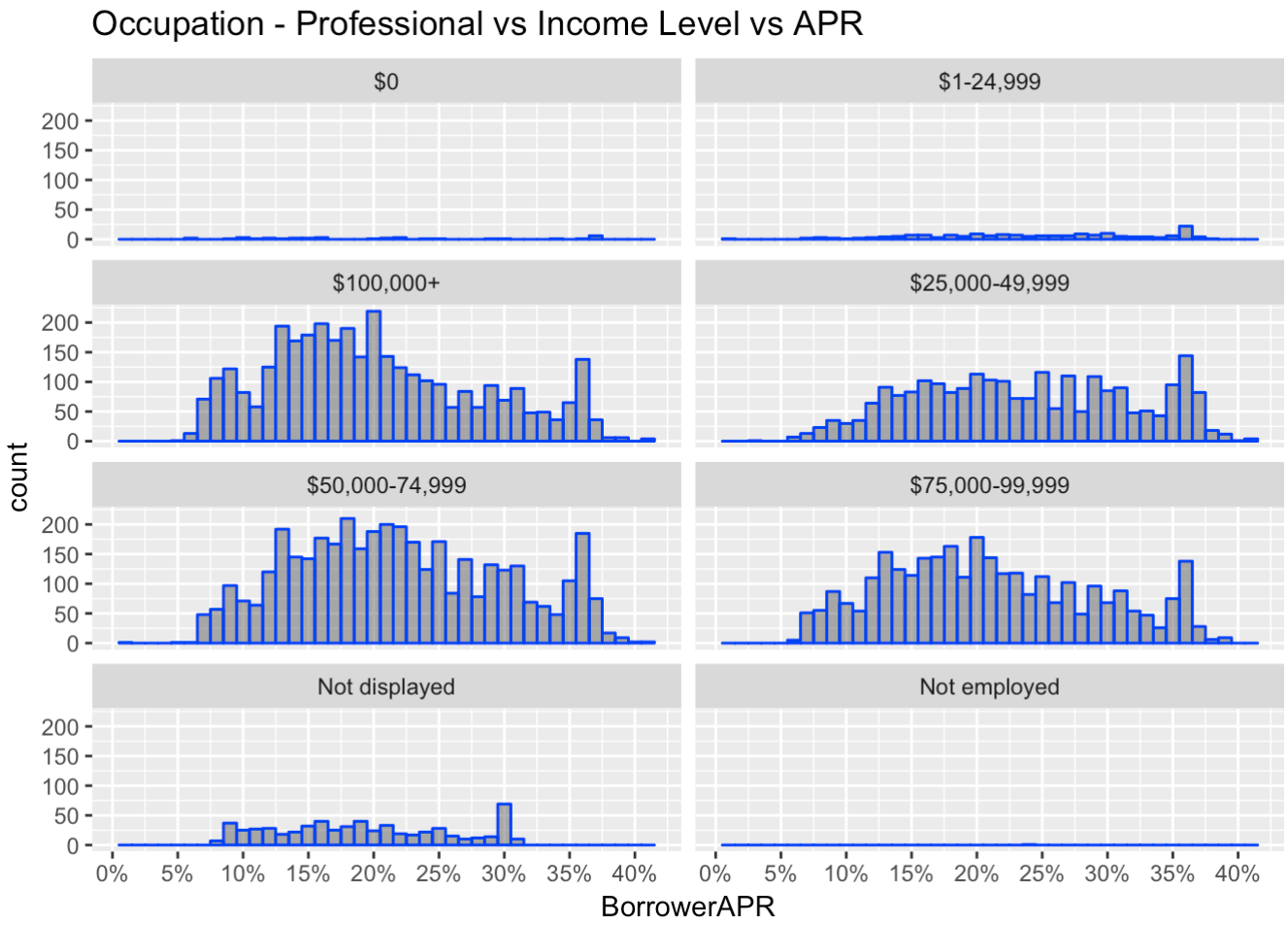
There is low correlation (-0.43) between BorrowerAPR and Credit score which is somewhat surprising as usually one would expect a more clear correlation. However, this is a small value loan, given to borrowers who exhausted other more traditional banking options. Traditional credit score parameters do not work any more in such cases, and cash lenders need to look for other data points to be able to lend borrowers at attractive rates.




```
## [1] -0.4297073
```

Occupation Professional and APR

The occupation “Professional” as it very ambiguous. As you can see below majority of this occupation earns from 50 to above 100K. It means that this occupation has a varied pay based probably on type of work and experience. It is not a great indicator of the borrower’s credibility though. I would probably recommend Prosper to omit this description from its database if they want to have a clearer picture of occupations and instead break it down to several occupations.



Reflection

It is a a large set with many factor like variables. The context of loan related information like APR, Credit Score, Lender Credit Score is familiar to many, however, it should be noted that this is a different type of lending. As such, we cannot just assume the correlations between APR and credit scores like in traditional banking. There must be proprietary scoring of some sort done by cash lenders.

We saw it clearly in this case, as Prosper specializes in short term risky cash loans and usual assuptions may not always apply there. Like there is not clear correlation between credit score and APR. Also some income levels and professions were not getting low APRs as expected. Prosper is a niche lender serving the market not covered by traditional banking - i.e. those who exhausted other options or who were not granted loans by banks.

There were many factor like variables which needed to be thouroughly analyzed and few of those factors needed to be focused on. I had to specifically choose two professions to see if there any similarity between them and weather Prosper grants loans based on profession. As such, this type of work by factor variables takes some time.

It was interesting to explore occupation vs credit ratings (both Prosper and Credit Score) and it gave lots of insights to the data. I would further research on what Prosper bases its rating - what is the most important factor? Is it income level, number of loans, debt to income ratio? Does state location have any bearing on the Prosper rating and APR? How the credit score and prosper rating correlate?

Also, very interesting would be to predict the quality of Prosper rating - wether they were able to properly grant the loan and an appropriate rate based. I would focus on overdue loans and try to dissect it further by occupation, Prosper rating, APR, debt to income ratio, number of previous and other loans etc.