Prosper Loan Data Exploration by Aloke K Das

Aloke K Das

June 29, 2017

Univariate Plots

Size of the Prosper loan dataset:

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

This is somewhat of a large dataset, especially with a large number of variables. I am not planning to analyze based on all the variables, The approach I am taking is:

1. Create a subset of the dataset with a reduced set of variables.
2. Find out and eliminate the outliers. Create a further reduced dataset. Use this reduced dataset for any further analysis.

I am creating a subset with a fewer and interesting set of variables that I will plan to analyze.

```
## [1] 113937     30
```

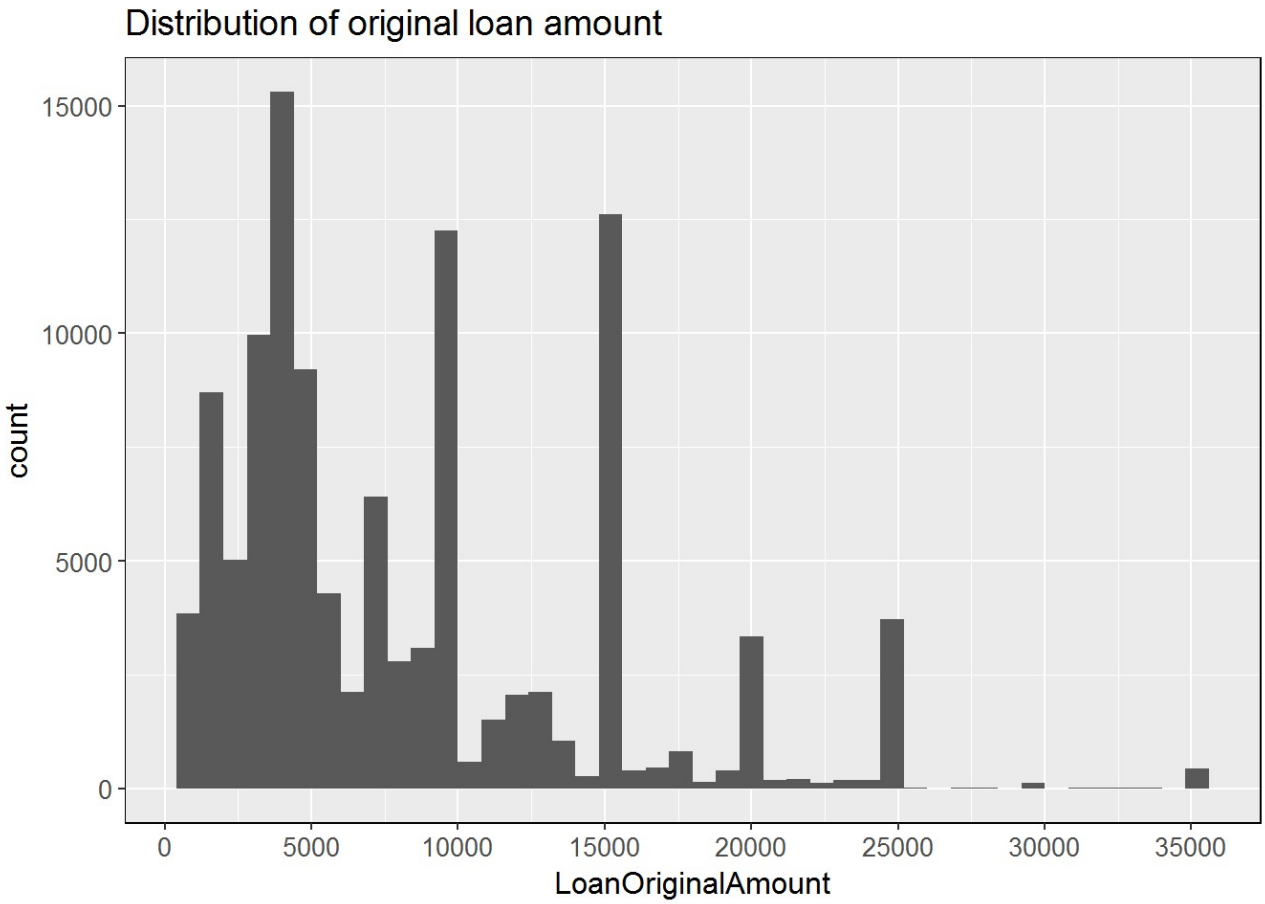
The variable ‘CreditGrade’ is available before 2009 and the variable ‘ProsperRating’ is available after 2009. I will merge these two variables into a single variable ‘Rating.merged’ such that they can be considered for the entire dataset period.

I have changed the variable name form “ListingCategory (numeric)” to “MyListingCategory” to make it convenient to handle.

Get a summary sense of the reduced loan dataset:

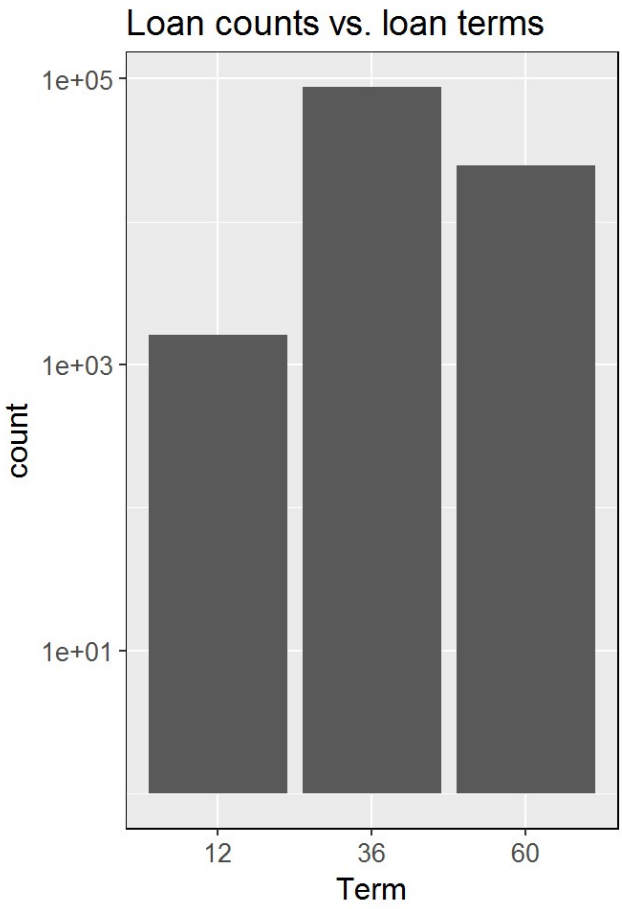
```
##                     ListingCreationDate  CreditGrade    ProsperRating
##  2013-10-02 17:20:16.550000000:     6          :84984          :29084
##  2013-08-28 20:31:41.107000000:     4   C      : 5649   C      :18345
##  2013-09-08 09:27:44.853000000:     4   D      : 5153   B      :15581
##  2013-12-06 05:43:13.830000000:     4   B      : 4389   A      :14551
##  2013-12-06 11:44:58.283000000:     4   AA     : 3509   D      :14274
##  2013-08-21 07:25:22.360000000:     3   HR     : 3508   E      : 9795
##  (Other)                       :113912   (Other): 6745   (Other):12307
##  Term                        LoanStatus     BorrowerAPR
##  12: 1614   Current              :56576   Min.   :0.00653
##  36:87778   Completed            :38074   1st Qu.:0.15629
##  60:24545   Chargedoff           :11992   Median :0.20976
##             Defaulted            : 5018   Mean   :0.21883
##             Past Due (1-15 days) :  806   3rd Qu.:0.28381
##             Past Due (31-60 days):  363   Max.   :0.51229
##             (Other)              : 1108   NA's   :25
##   BorrowerRate     ProsperScore   BorrowerState
##  Min.   :0.0000   Min.   : 1.00   CA     :14717
##  1st Qu.:0.1340   1st Qu.: 4.00   TX     : 6842
##  Median :0.1840   Median : 6.00   NY     : 6729
##  Mean   :0.1928   Mean   : 5.95   FL     : 6720
##  3rd Qu.:0.2500   3rd Qu.: 8.00   IL     : 5921
##  Max.   :0.4975   Max.   :11.00          : 5515
##                   NA's   :29084   (Other):67493
##                     Occupation         EmploymentStatus
##  Other                    :28617   Employed     :67322
##  Professional             :13628   Full-time    :26355
##  Computer Programmer      : 4478   Self-employed: 6134
##  Executive                : 4311   Not available: 5347
##  Teacher                  : 3759   Other        : 3806
##  Administrative Assistant: 3688                : 2255
##  (Other)                  :55456   (Other)      : 2718
##  EmploymentStatusDuration CreditScoreRangeLower CreditScoreRangeUpper
##  Min.   :  0.00           Min.   :  0.0         Min.   : 19.0
##  1st Qu.: 26.00           1st Qu.:660.0         1st Qu.:679.0
##  Median : 67.00           Median :680.0         Median :699.0
##  Mean   : 96.07           Mean   :685.6         Mean   :704.6
##  3rd Qu.:137.00           3rd Qu.:720.0         3rd Qu.:739.0
##  Max.   :755.00           Max.   :880.0         Max.   :899.0
##  NA's   :7625             NA's   :591           NA's   :591
##         FirstRecordedCreditLine CurrentCreditLines OpenCreditLines
##                       :   697     Min.   : 0.00      Min.   : 0.00
##  1993-12-01 00:00:00:   185     1st Qu.: 7.00      1st Qu.: 6.00
##  1994-11-01 00:00:00:   178     Median :10.00      Median : 9.00
##  1995-11-01 00:00:00:   168     Mean   :10.32      Mean   : 9.26
##  1990-04-01 00:00:00:   161     3rd Qu.:13.00      3rd Qu.:12.00
##  1995-03-01 00:00:00:   159     Max.   :59.00      Max.   :54.00
##  (Other)             :112389     NA's   :7604       NA's   :7604
```

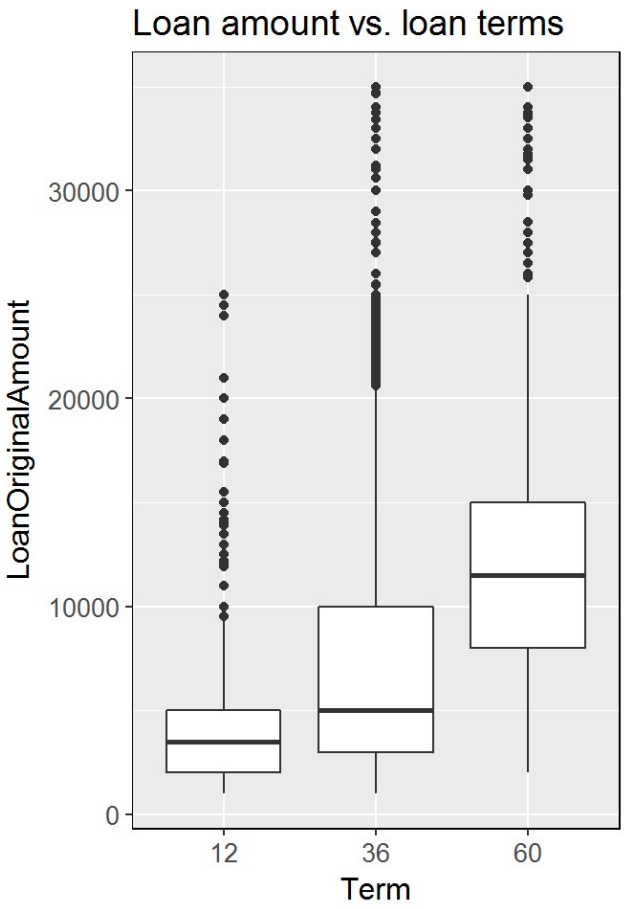
```
##  TotalCreditLinespast7years OpenRevolvingAccounts
##  Min.   :  2.00             Min.   : 0.00
##  1st Qu.: 17.00             1st Qu.: 4.00
##  Median : 25.00             Median : 6.00
##  Mean   : 26.75             Mean   : 6.97
##  3rd Qu.: 35.00             3rd Qu.: 9.00
##  Max.   :136.00             Max.   :51.00
##  NA's   :697
##  OpenRevolvingMonthlyPayment InquiriesLast6Months TotalInquiries
##  Min.   :    0.0             Min.   :  0.000      Min.   :  0.000
##  1st Qu.:  114.0             1st Qu.:  0.000      1st Qu.:  2.000
##  Median :  271.0             Median :  1.000      Median :  4.000
##  Mean   :  398.3             Mean   :  1.435      Mean   :  5.584
##  3rd Qu.:  525.0             3rd Qu.:  2.000      3rd Qu.:  7.000
##  Max.   :14985.0             Max.   :105.000      Max.   :379.000
##                              NA's   :697          NA's   :1159
##  CurrentDelinquencies DebtToIncomeRatio         IncomeRange
##  Min.   : 0.0000      Min.   : 0.000    $25,000-49,999:32192
##  1st Qu.: 0.0000      1st Qu.: 0.140    $50,000-74,999:31050
##  Median : 0.0000      Median : 0.220    $100,000+     :17337
##  Mean   : 0.5921      Mean   : 0.276    $75,000-99,999:16916
##  3rd Qu.: 0.0000      3rd Qu.: 0.320    Not displayed : 7741
##  Max.   :83.0000      Max.   :10.010    $1-24,999     : 7274
##  NA's   :697          NA's   :8554      (Other)       : 1427
##  StatedMonthlyIncome          LoanOriginationDate LoanOriginalAmount
##  Min.   :      0     2014-01-22 00:00:00:   491   Min.   : 1000
##  1st Qu.:   3200     2013-11-13 00:00:00:   490   1st Qu.: 4000
##  Median :   4667     2014-02-19 00:00:00:   439   Median : 6500
##  Mean   :   5608     2013-10-16 00:00:00:   434   Mean   : 8337
##  3rd Qu.:   6825     2014-01-28 00:00:00:   339   3rd Qu.:12000
##  Max.   :1750003     2013-09-24 00:00:00:   316   Max.   :35000
##                      (Other)            :111428
##  MyListingCategory MonthlyLoanPayment Rating.merged
##  Min.   : 0.000    Min.   :   0.0     C      :23994
##  1st Qu.: 1.000    1st Qu.: 131.6     B      :19970
##  Median : 1.000    Median : 217.7     D      :19427
##  Mean   : 2.774    Mean   : 272.5     A      :17866
##  3rd Qu.: 3.000    3rd Qu.: 371.6     E      :13084
##  Max.   :20.000    Max.   :2251.5     HR     :10443
##                                       (Other): 9153
```



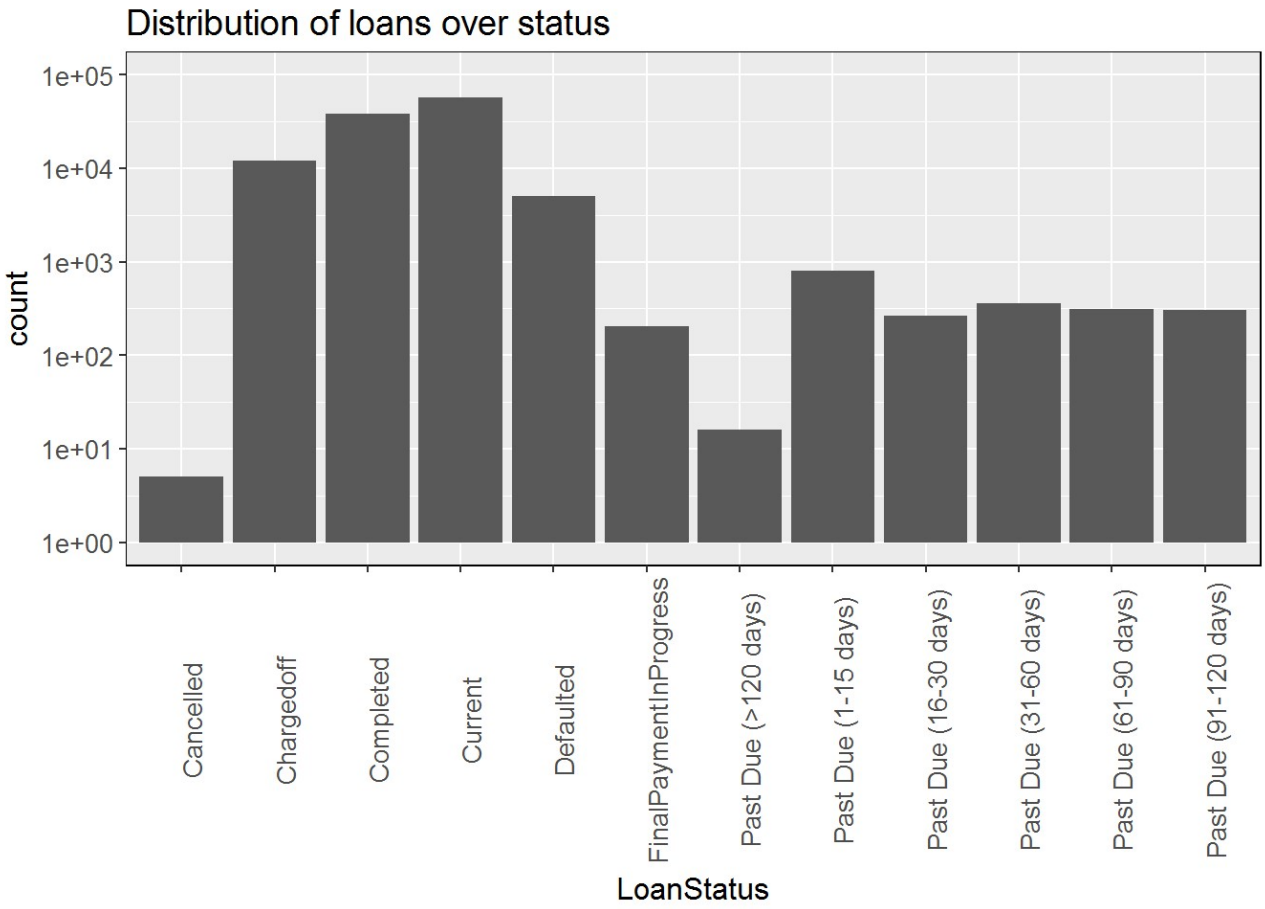
We observe that most of the loans are small in the sub $20000 range.

What loan terms these small borrowers are borrowing for? How are the loans distributed over various terms?





We see that the majority of the loans were issued for 36 month term. A relatively small percentage of loans have been issued for 12 moth term. Also, in general, we see that the higher the loan amount the longer is the term of the loan.



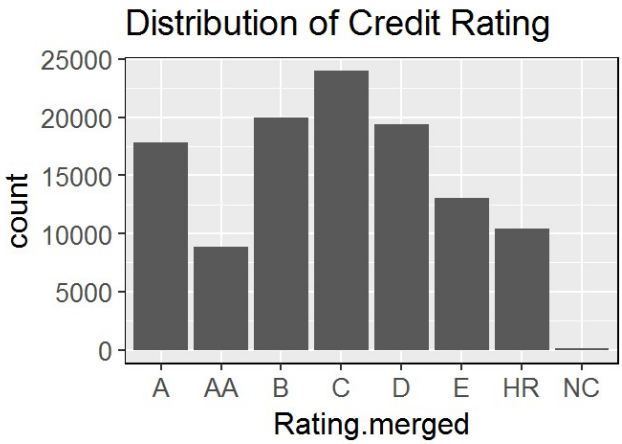
We see that most of the loans are either completed or current.However, there are a few defaults, chargedoff and delinquent loans. We will analyze them later.

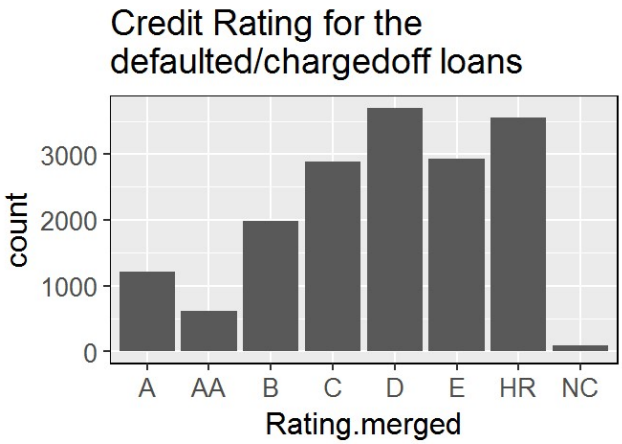
Lets find out the distribution of Credit Rating.

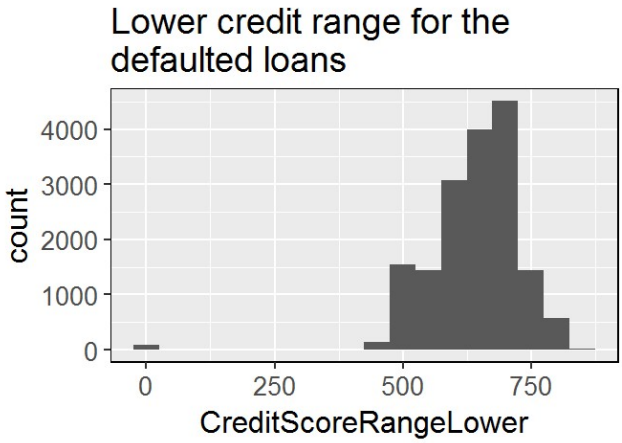
```
##    x  freq
## 1      131
## 2  A 17866
## 3 AA  8881
## 4  B 19970
## 5  C 23994
## 6  D 19427
## 7  E 13084
## 8 HR 10443
## 9 NC   141
```

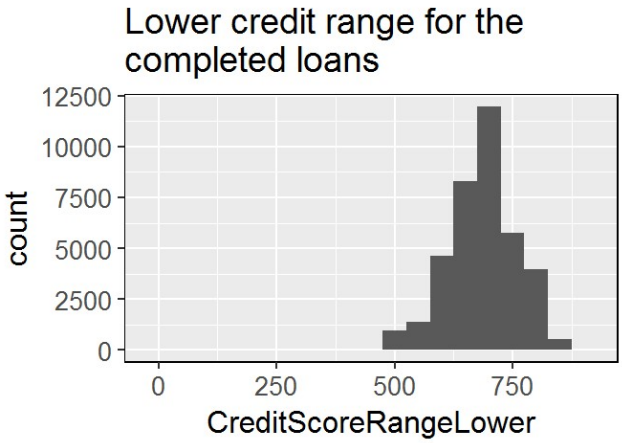
The first plot below shows the count of the loans with Credit Rating. The second plot shows, as expected, that the high risk (HR) Credit Rating results into most loan defaults.

I also wanted to find out the relationship of the credit score for the defaulted and completed loans. The plot shows a higher number for defaulted loans for lower credit score. On the similar line, we see the borrowers with higher credit score were likely to have completed the loan.

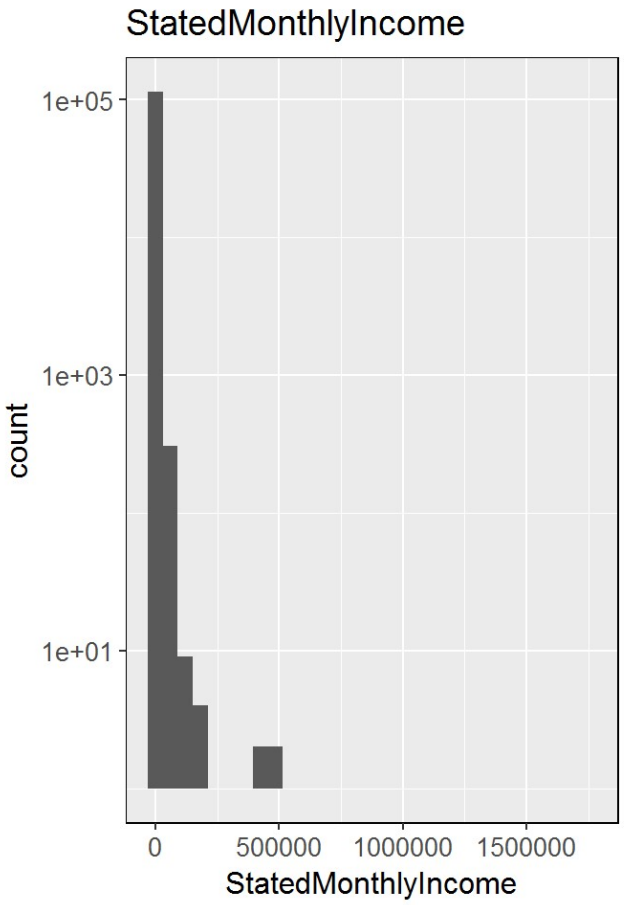


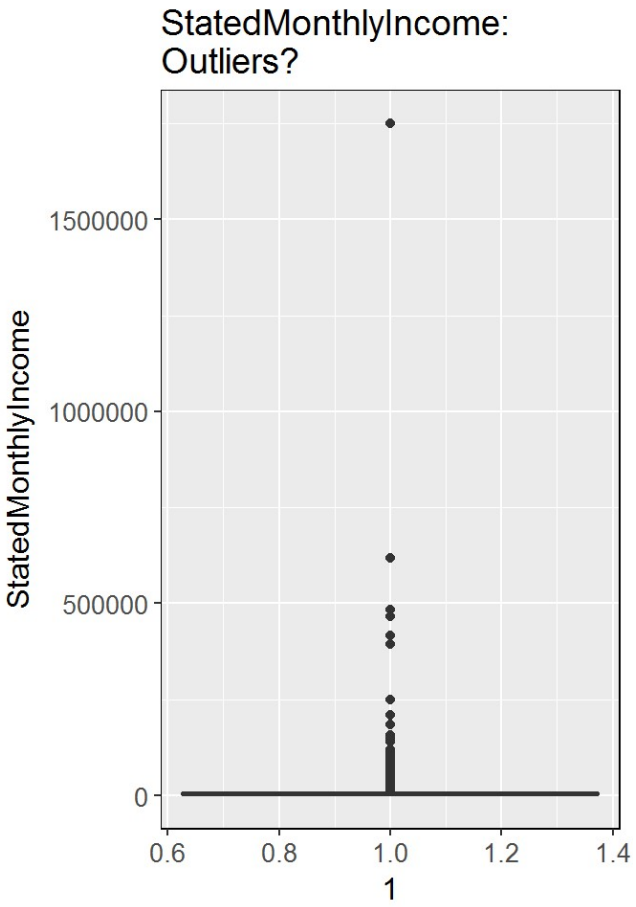






Lets take a look at the stated monthly income of the borrowers.



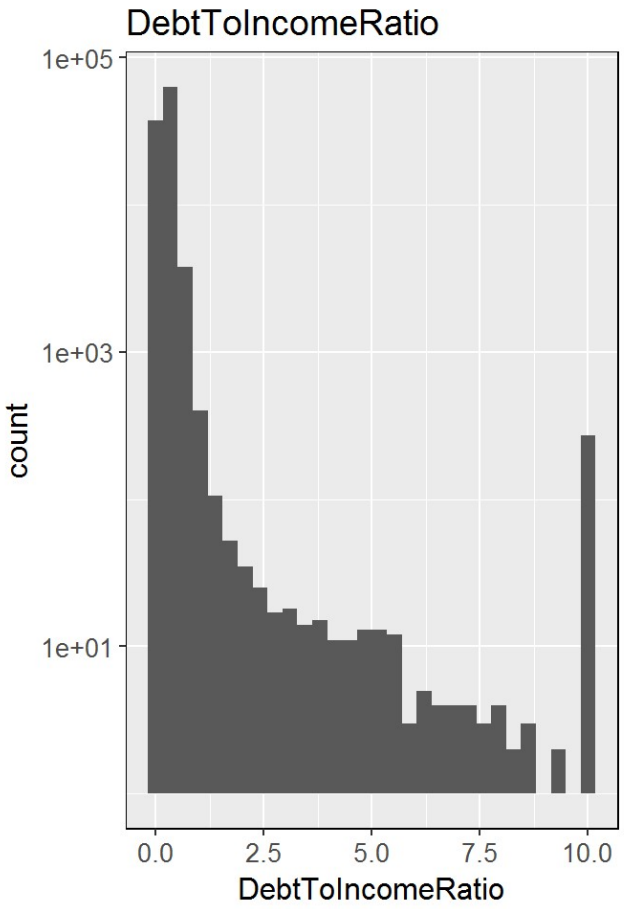


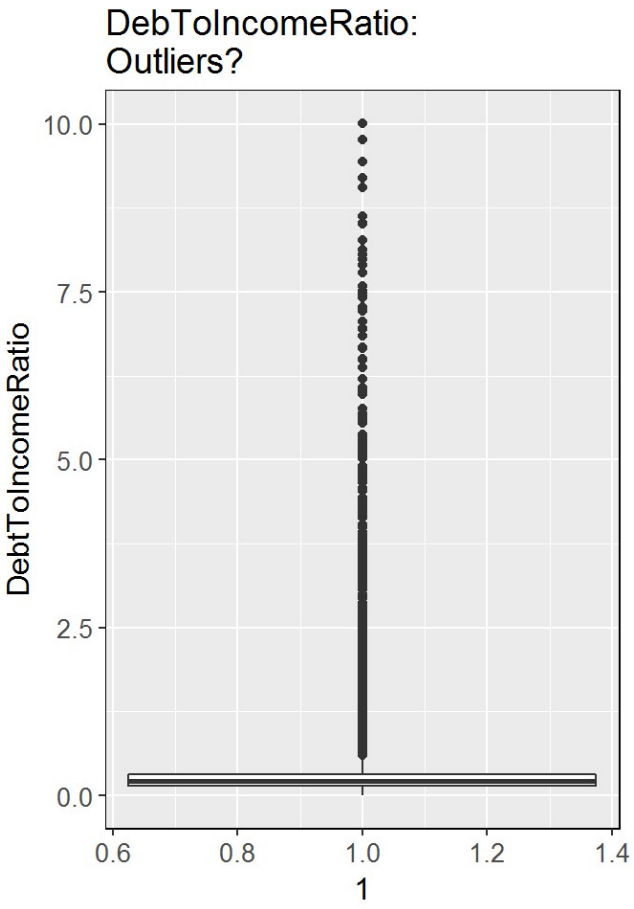
We see a huge skew in the data. There are borrowers who earn disproportionally higher than the others.

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

We will remove these outliers and consider only the monthly incomes that are less than $20000.

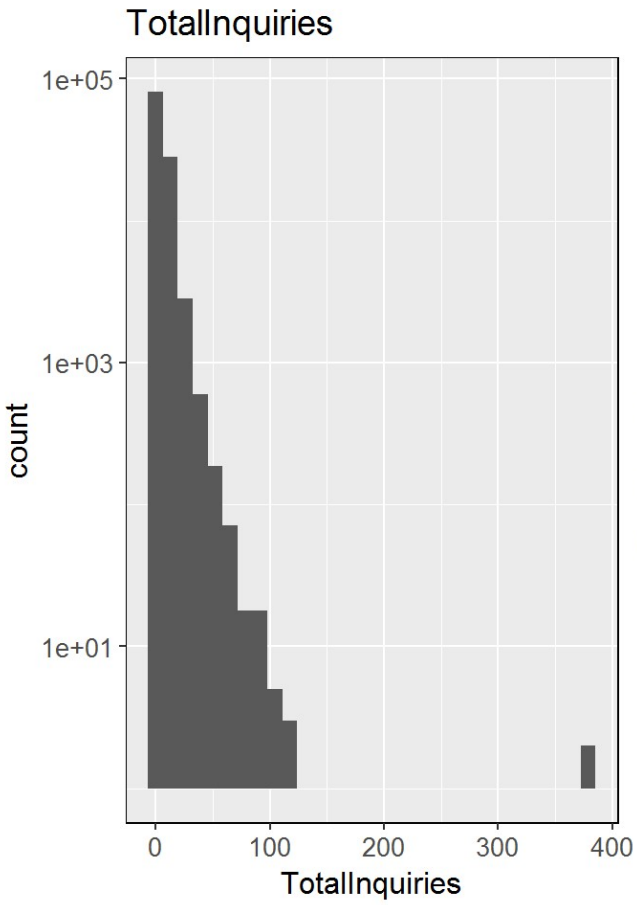
Taking a look at the DebtToIncomeRatio.

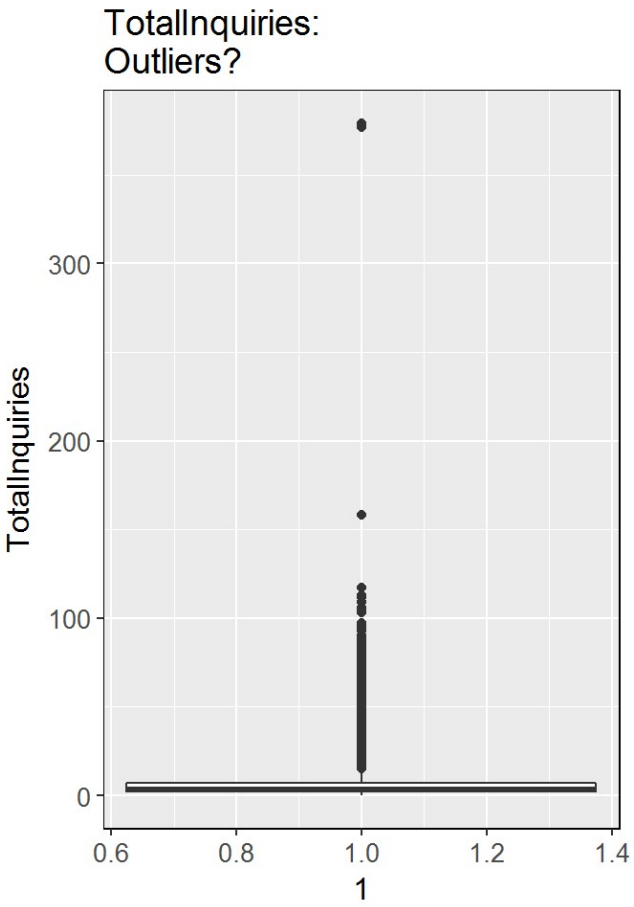




The above plots show some outliers with very high DTI. I will only consider the observations with DTI leass than 2.5.

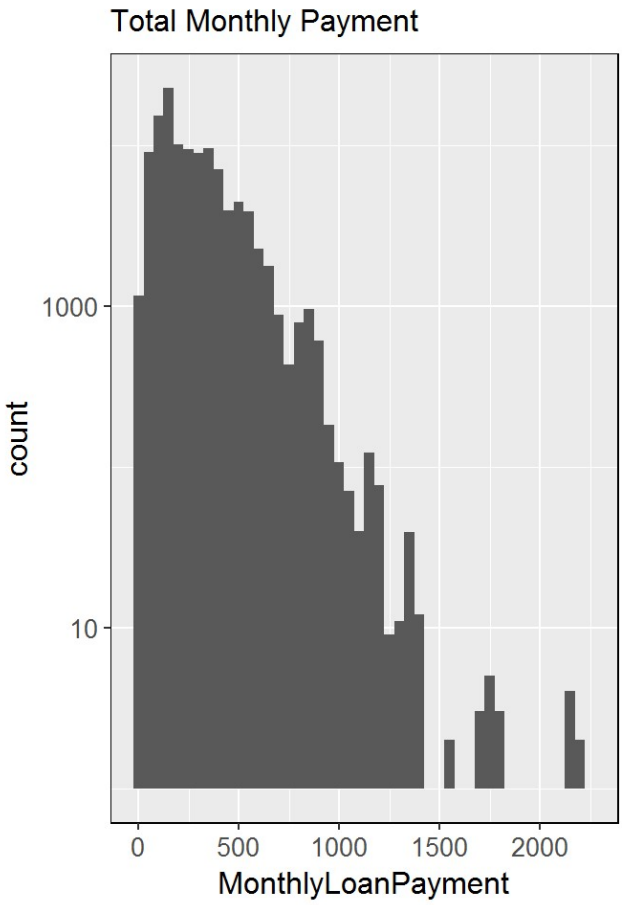
More about TotalInquiries:

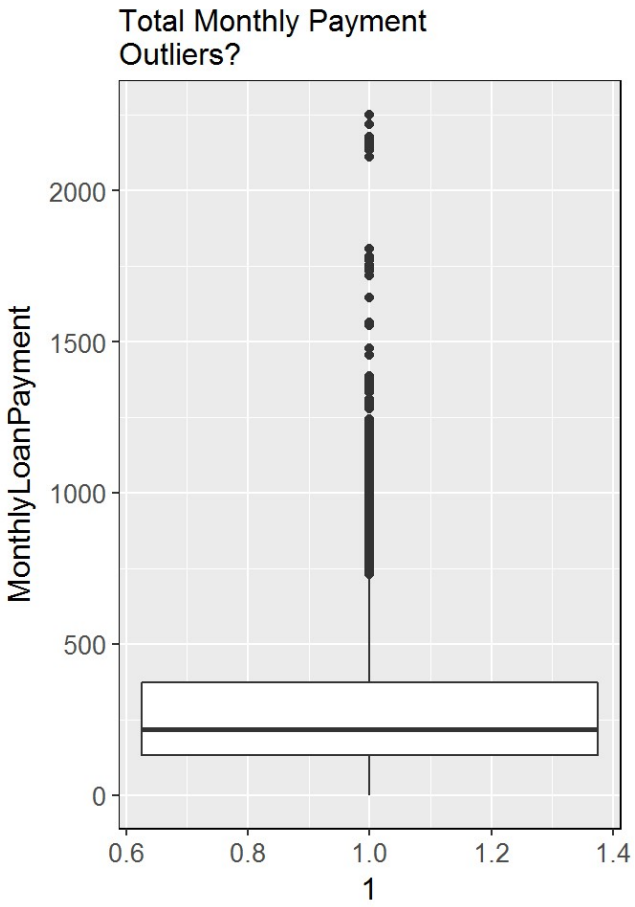




The above plots shows a few outliers for TotalInquiries. I will consider only those observations that have less than 100 inquiries in total.

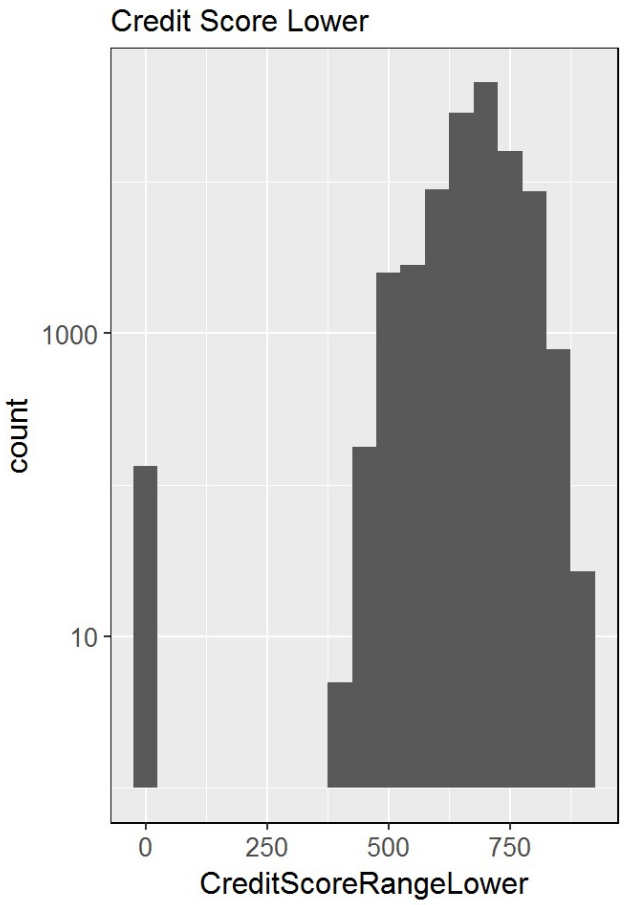
Let find out if there is anything unusual abou the monthly loan payment.

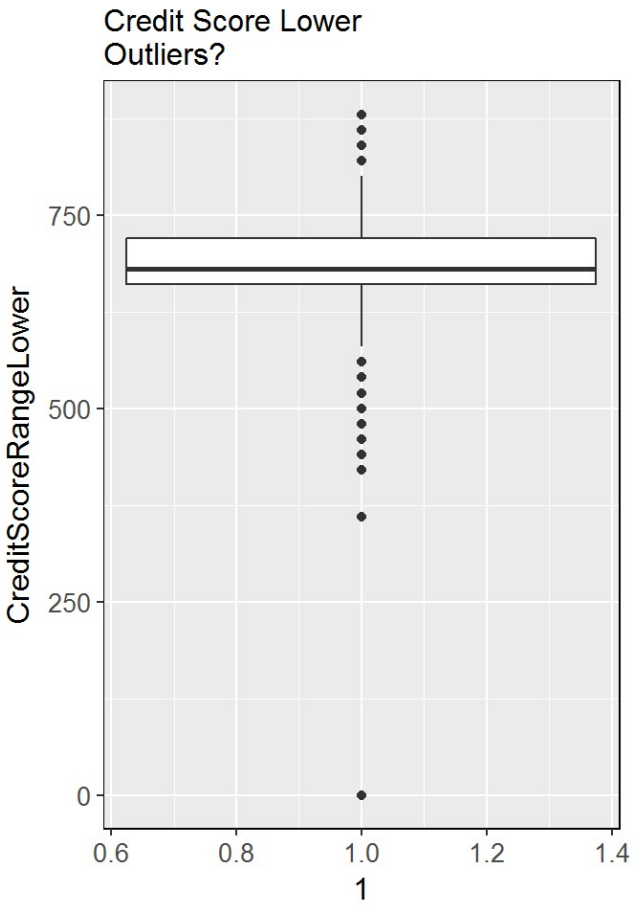




There are a few borrowers who are paying monthly loan payment much more than others. However, these higher values do not appear to be extraordinarily large. I will let them remain in the dataset.

Lets look at the available credit scrore data and find out if there is anything unusual.



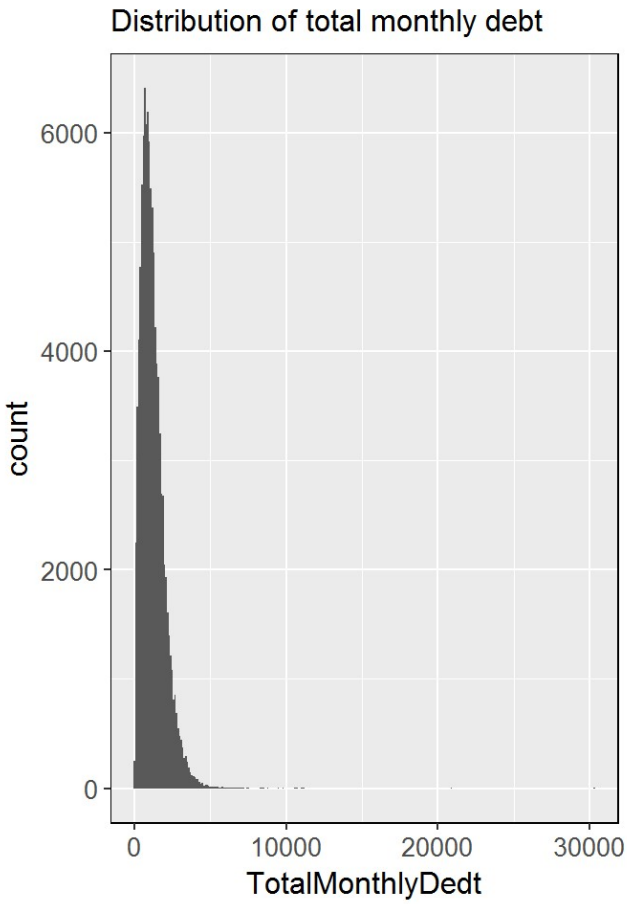


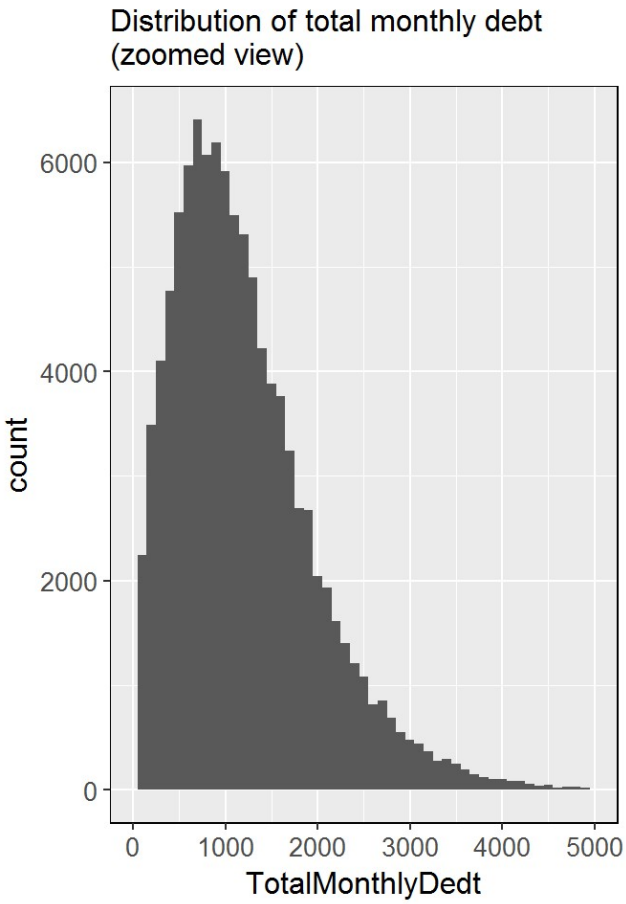
The above plots show that there are a few observations in which the credit score has been recorded as '0'. I would like to consider those borrowers with credit score > 300 only.

We will now prepare a subset of the dataset that will have the outliers removed. This subset of data will be analyzed in the subsequent sections.

```
## [1] 102653     31
```

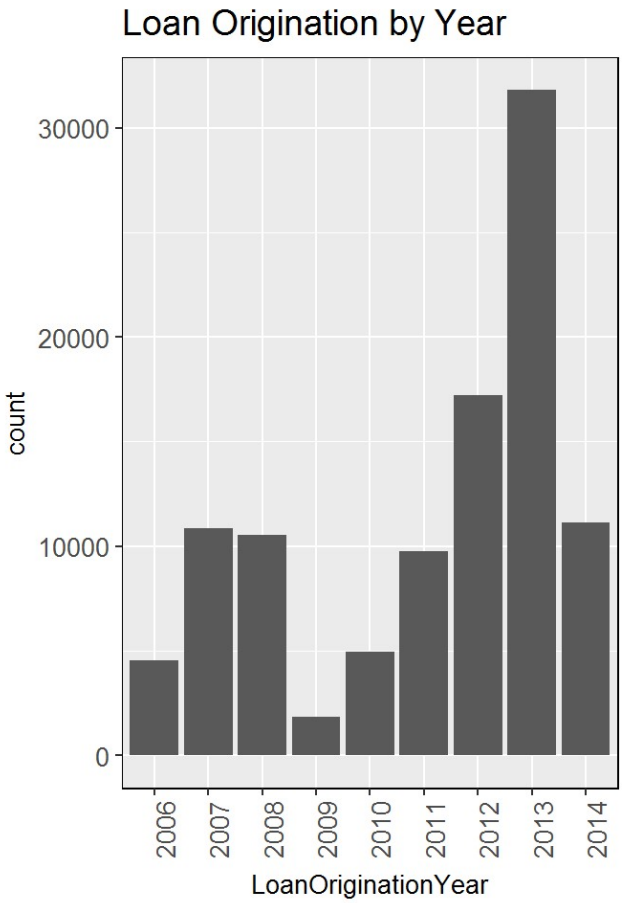
How are the borrowers doing w.r.t their total montly debt? I create a new variable to represent the total monthly debt.

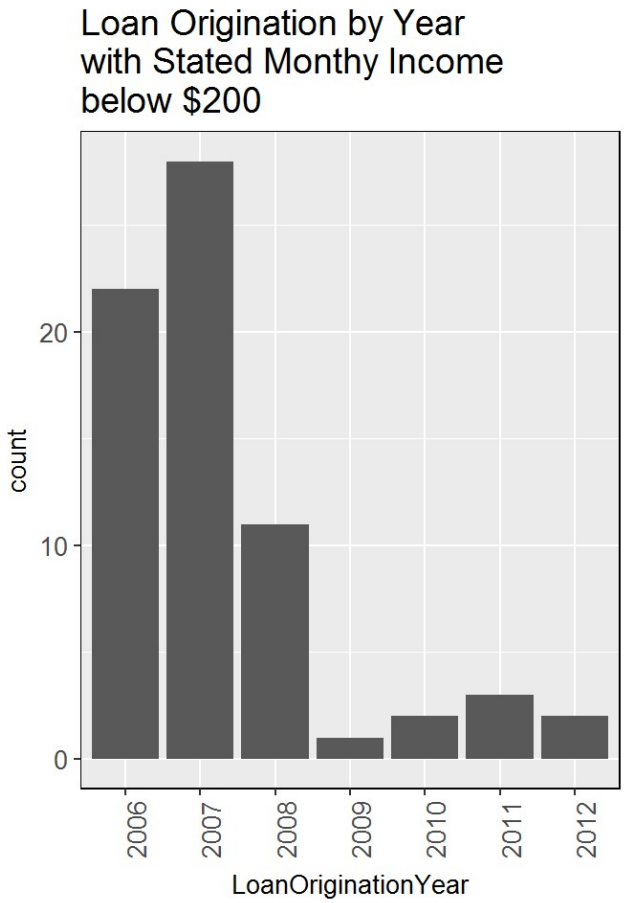




It looks like the majority of the Prosper loans was issued to the people who carry less total monthly debt. Zooming in to the debt values, we see that the total monthly dedts are even less than 5000.

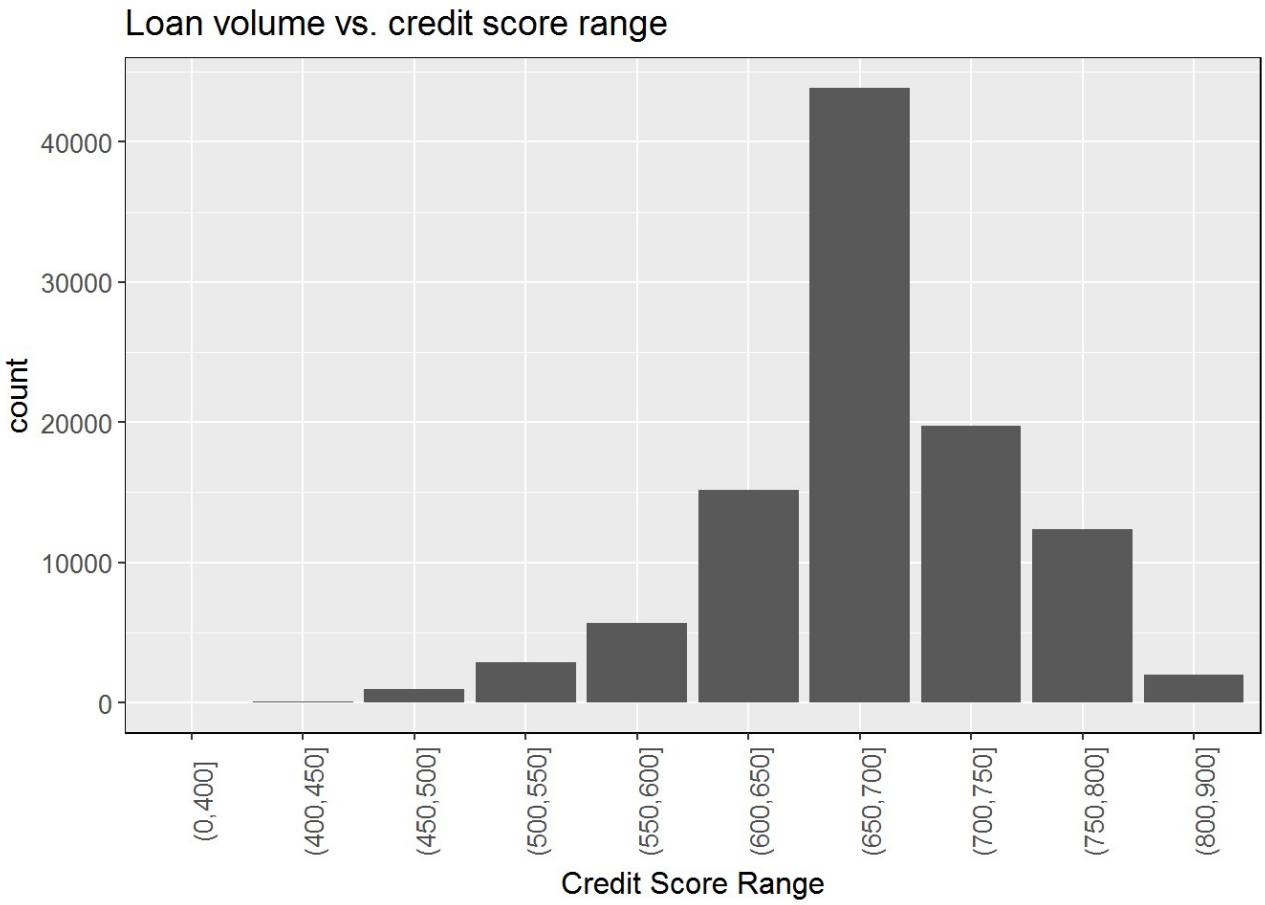
I wanted to explore the characteristics of the Prosper loans issued over the years. Given that we now know the housing market meltdown, can we confirm anything from the dataset?





The plots show the the loans issued over the years where the data is available. The loan industry became very strict on the onset of the credit market meltdown which is relevant from the plot. It shows that in 2009 loan volumes bottomed out. It is intersting to note that some loans where issued to borrowers with less than stated monthly income of $200.

We would also like to see the distribution of loans over various credit score ranges.



The plot shows a normal distribution of loan volumes over credit scores. Prosper was neither too restrictive, nor too liberal about issueing the loans while considering such scores.

Univariate Analysis

What is the structure of your dataset?

This is a loan database with 113937 individual loan records with 81 variables. I selected a set of 30 variables to work with.

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

Directly experienced in the years of credit market meltdown, I was interested to know how disciplined this lender was, what segment of the market it addressed, how was its business growth. I was curious to know what was the basis of the APR they charged to the borrowers. Finally, I was wondering if there is any information that can be extracted from the dataset to help predict if the loan will be completed or delinquent.

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

Various ratings including the credit scores and internal rating by Prosper will be helpful. Employment status, monthly income, loan amount, number of credit enquiries in the past are all useful information for the company. I will consider these variables for analysis.

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

Of sepecific interest to me was the year-wise profiling of this company's business and underwriting practices. Hence, I have used the 'lubridate' package to extract the information about the 'Year' in which any loan was issued. I have created a 'LoanOutcome' variable to classify the datasent into two broad categories: '1' = loan is current, completed or final payment in progress, '0' = for all others. I have also created a new variable 'cr_range' to represent various buckets of credit score ranges.

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?

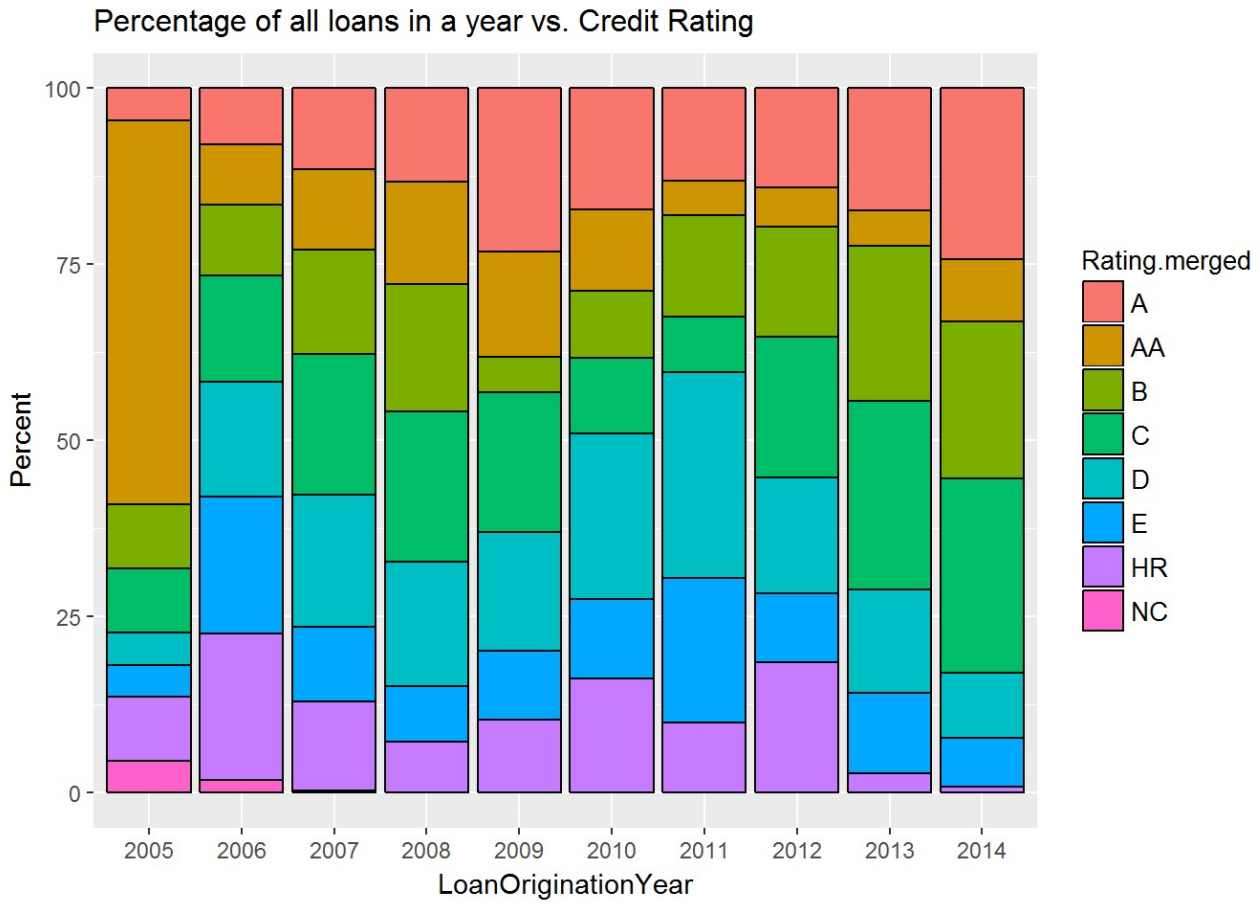
I noticed a few outliers in the dataset, e.g. StatedMonthlyIncomet, DebtToIncomeRatio, and TotalInquiries.

I conclude that these outliers may be a few mis-statements and therefore, created a datatset to eliminate those outliers. I will use this dataset for subsequent analysis.

I stayed with the given shape of the dataset and did not use any reshaping tools to change its form.

Bivariate Plots

We want to investigate the loans issued in a particular year spread across various credit ratings. We will also find out the loan counts as a percentge of the total loans issued in that year. This is to find out if there is any credit ratings that were more common in that year.

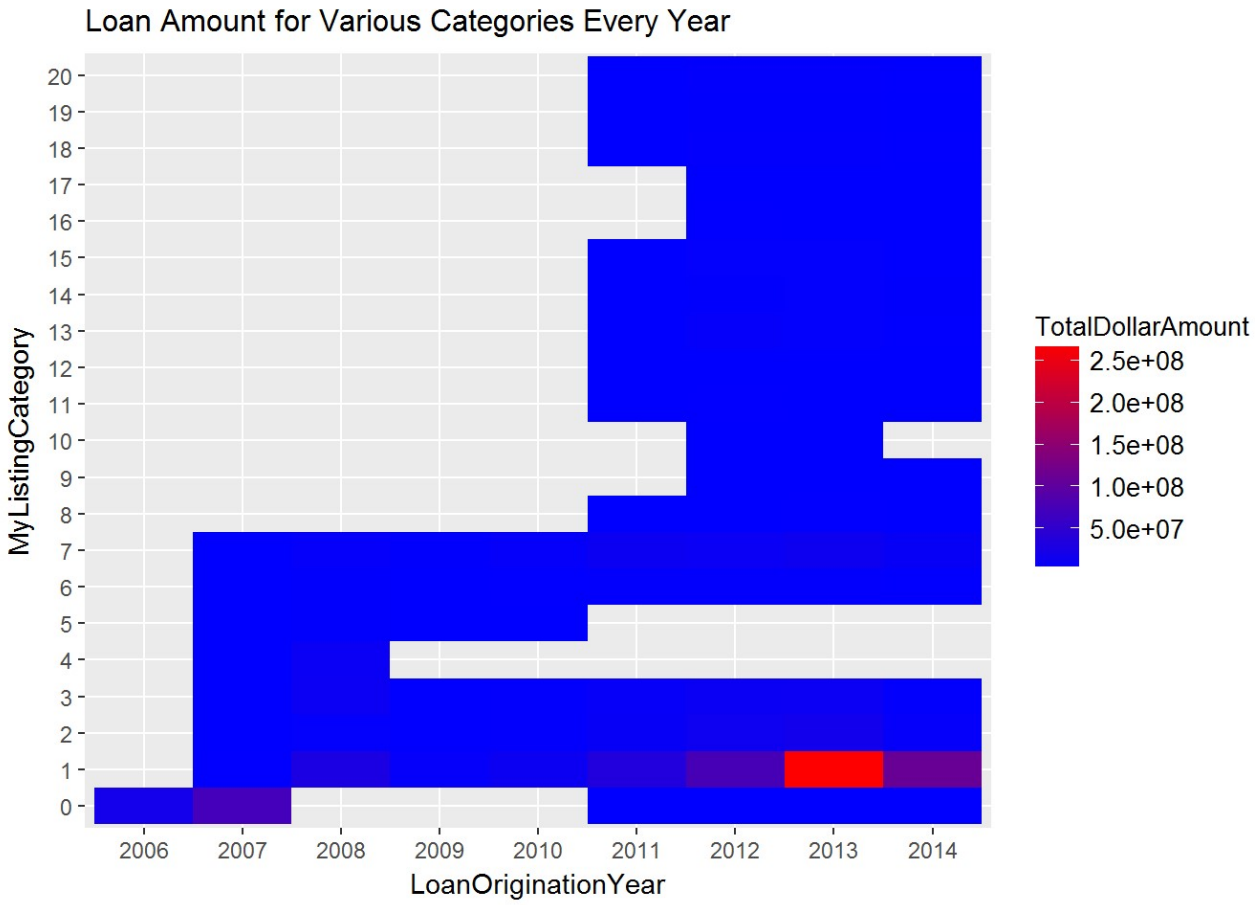


The plot does not indicate any bias to issuing any loan in favor of any specific credit rating all across the years.

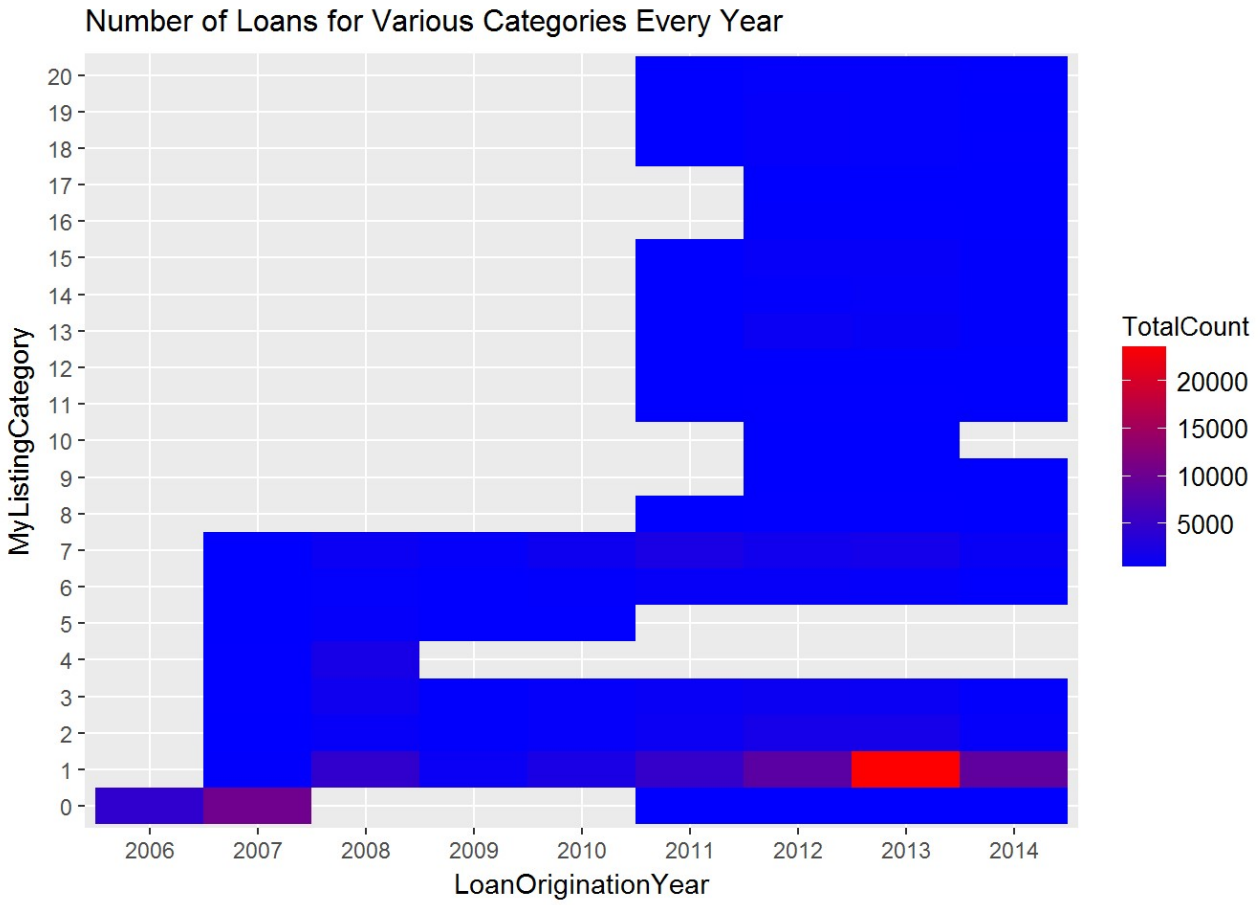
Lets analyze the loan data for various categories of loans over the period of time.

The following dataframe shows the total dollar amount and the count of loans for various categories.

```
## 'data.frame':    99 obs. of  4 variables:
##  $ LoanOriginationYear: Factor w/ 9 levels "2006","2007",..: 1 2 6 7 8 9 2
3 4 5 ...
##  $ MyListingCategory  : Factor w/ 21 levels "0","1","2","3",..: 1 1 1 1 1 1
2 2 2 2 ...
##  $ TotalDollarAmount  : int  21647092 72491978 44505 12000 22500 29000 13158
39 28873387 4582419 13261308 ...
##  $ TotalCount         : int  4561 10429 9 3 5 2 189 4615 883 2427 ...
```

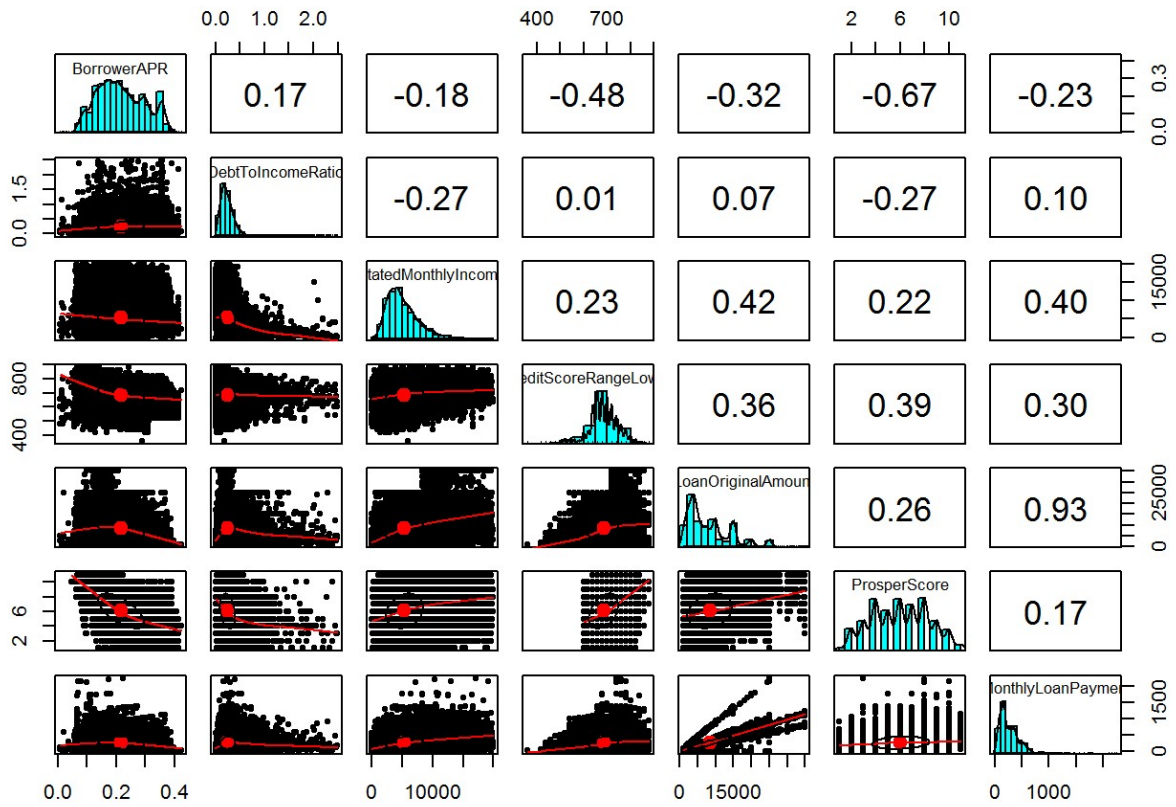


We see that, the most $ loan amount seems to be taken for debt consolidation. Significantly higher $ amount of debt consolidation loan was taken in 2013.



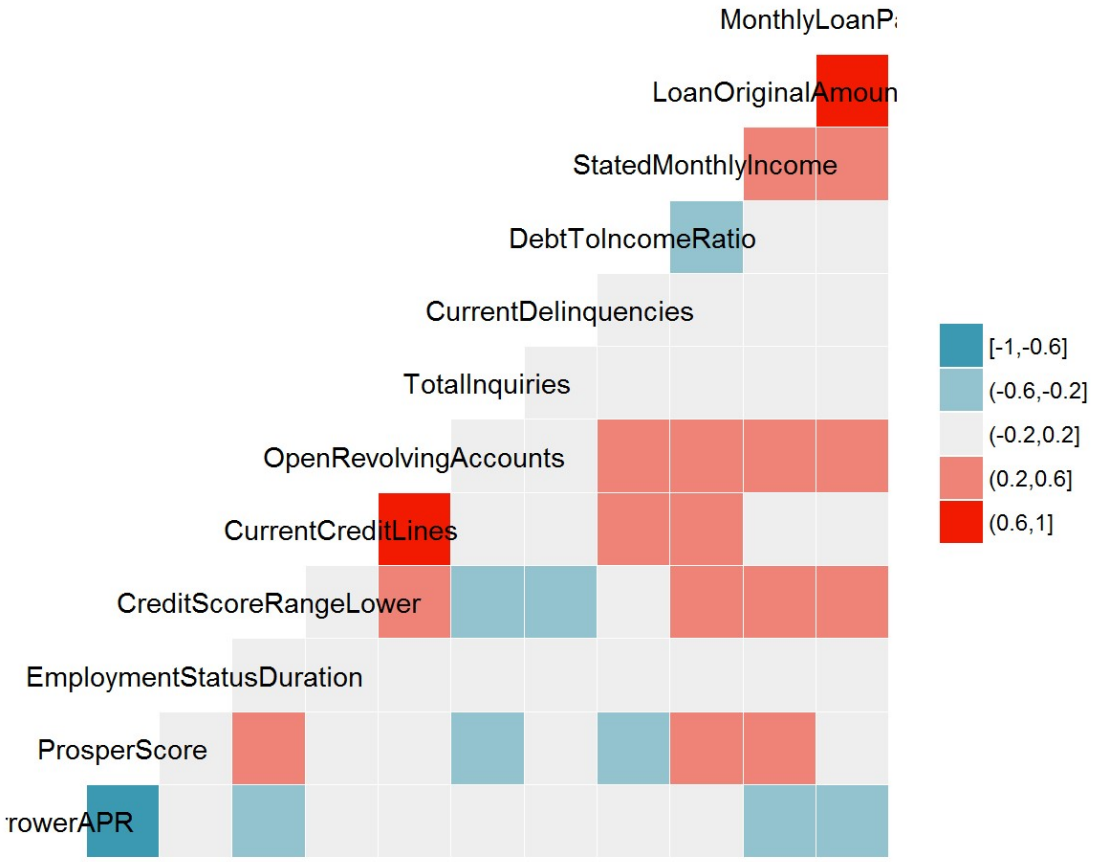
This plot is in line with the previous plot. I wanted to understand the total number of loans issued per category each year. We see that the most number of loans were issued for debt consolidation. Maximum such number of loans were issued in the year 2013.

Next I will try to find out any obvious correlation among variables. I will start with Scatterplot matrices.

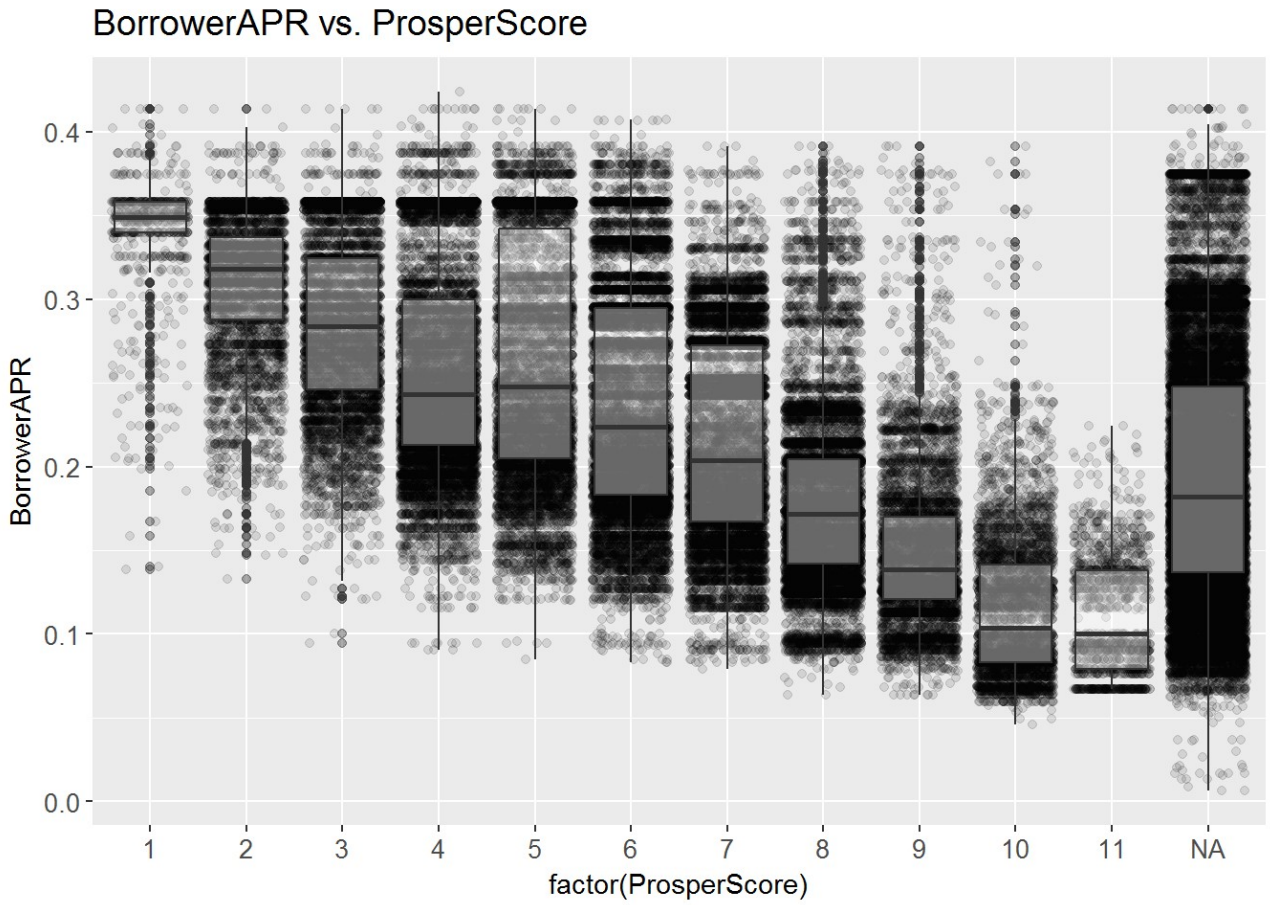


The scatterplot matrices with a few limited number of variables show some obvious relationshipship, e.g. original loan amount and the monthly payment.

In order to accomoate a few more variables like credit score, number of past enquiries, number of credit lines, etc. I went ahead to generate a correlation matrix visualization using the ‘ggcorr2’ function.



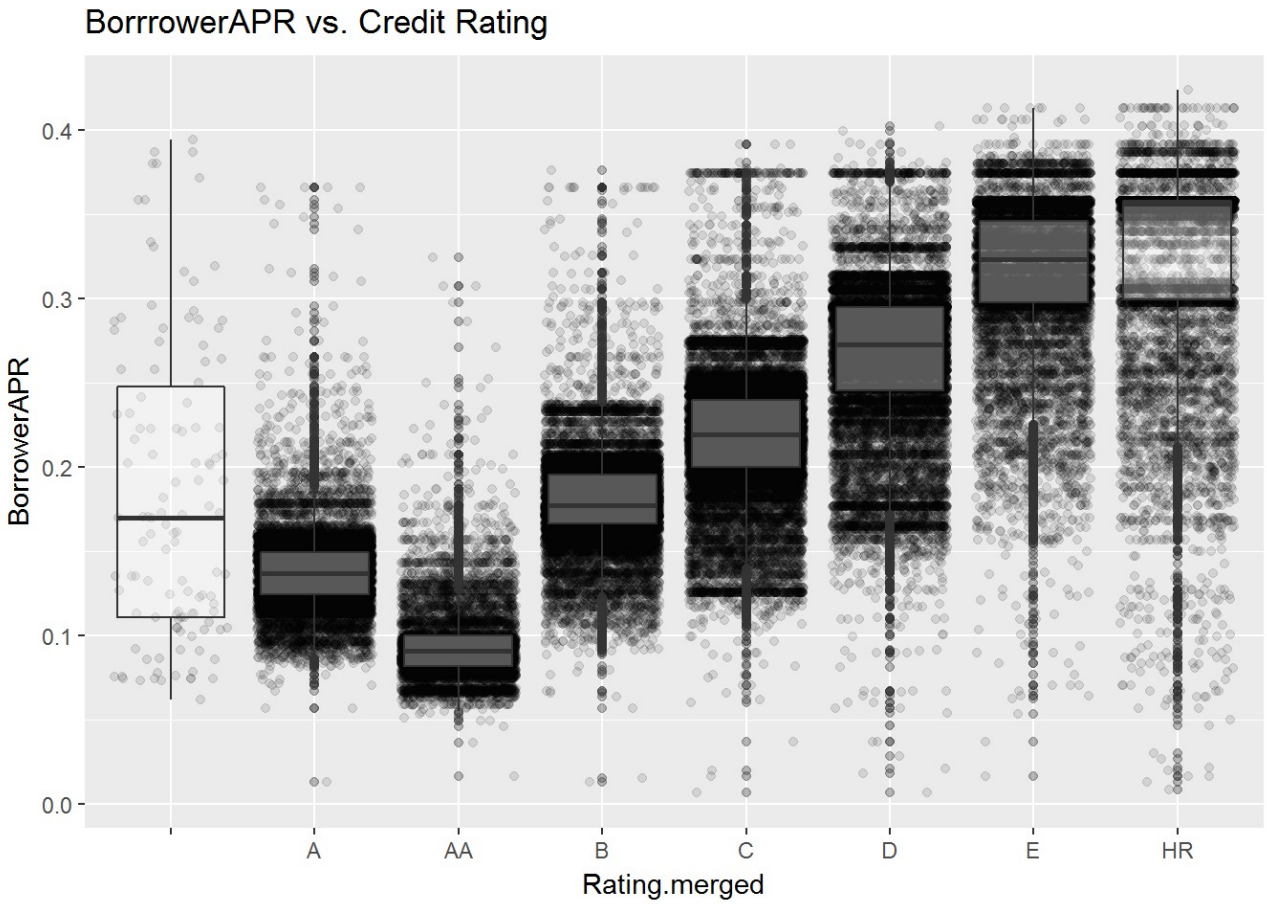
This visualization indicates that there is a strong negative correlation (between -1 and -0.6) between BorrowerAPR and ProsperScore. Lets try to plot this relationship with some more details



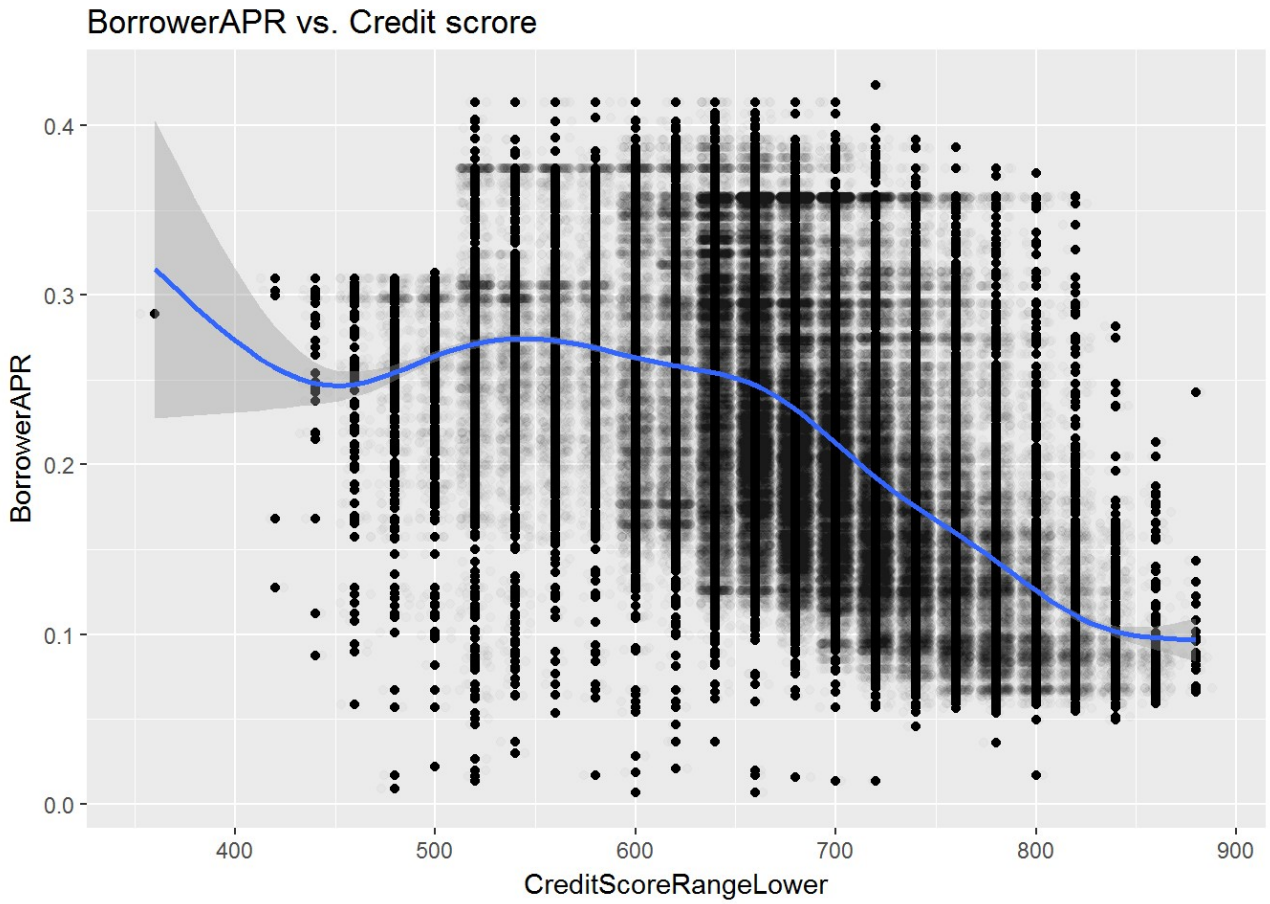
As expected, the higher the ProsperScore, the lower is the BorrowerAPR.

Now I will create a few plots to understand how BorrowerAPR is related to a few relevant variables.

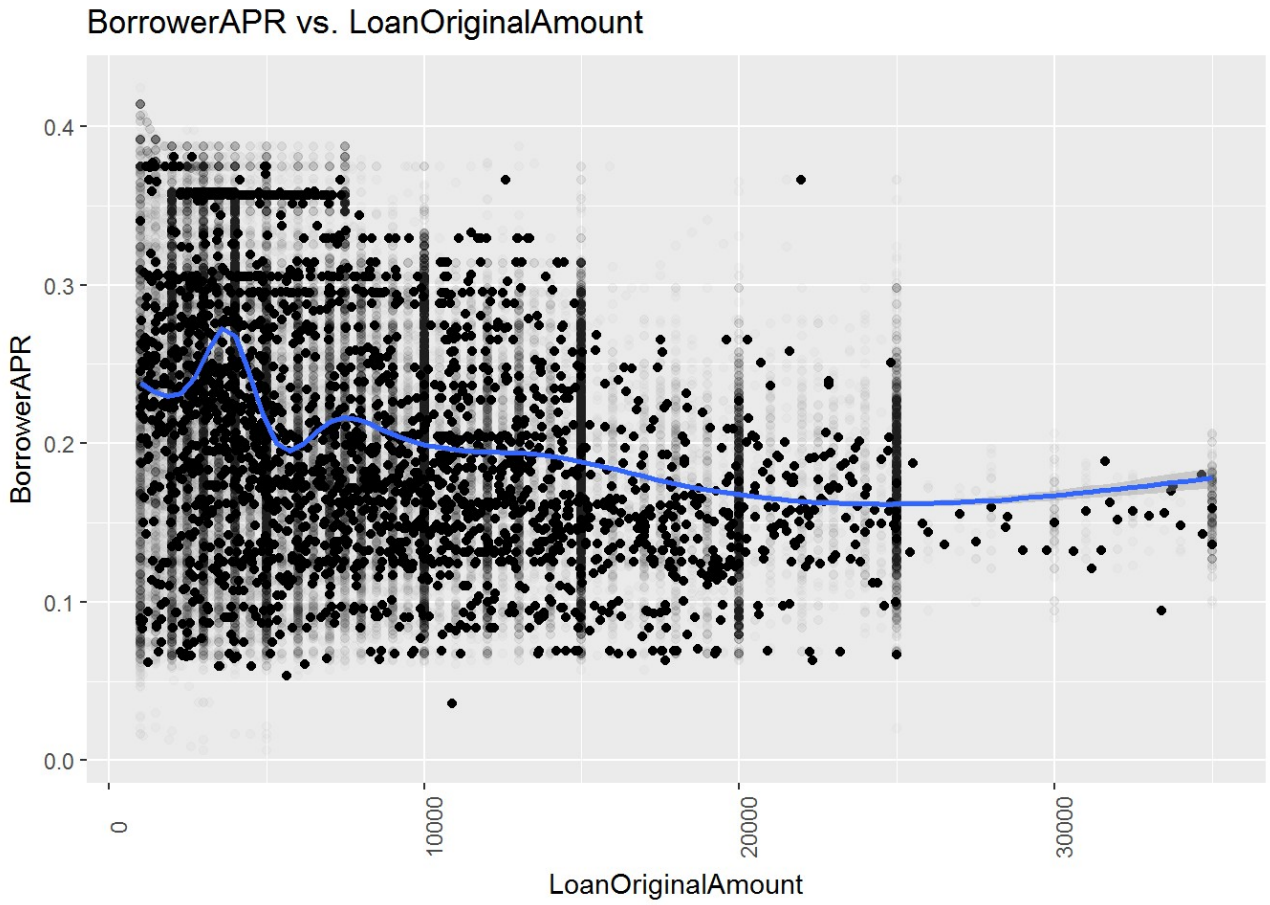
We did not find it, but presumably there is an expected strong corrleation between the credit rating and the APR. We will find it in the next plot.



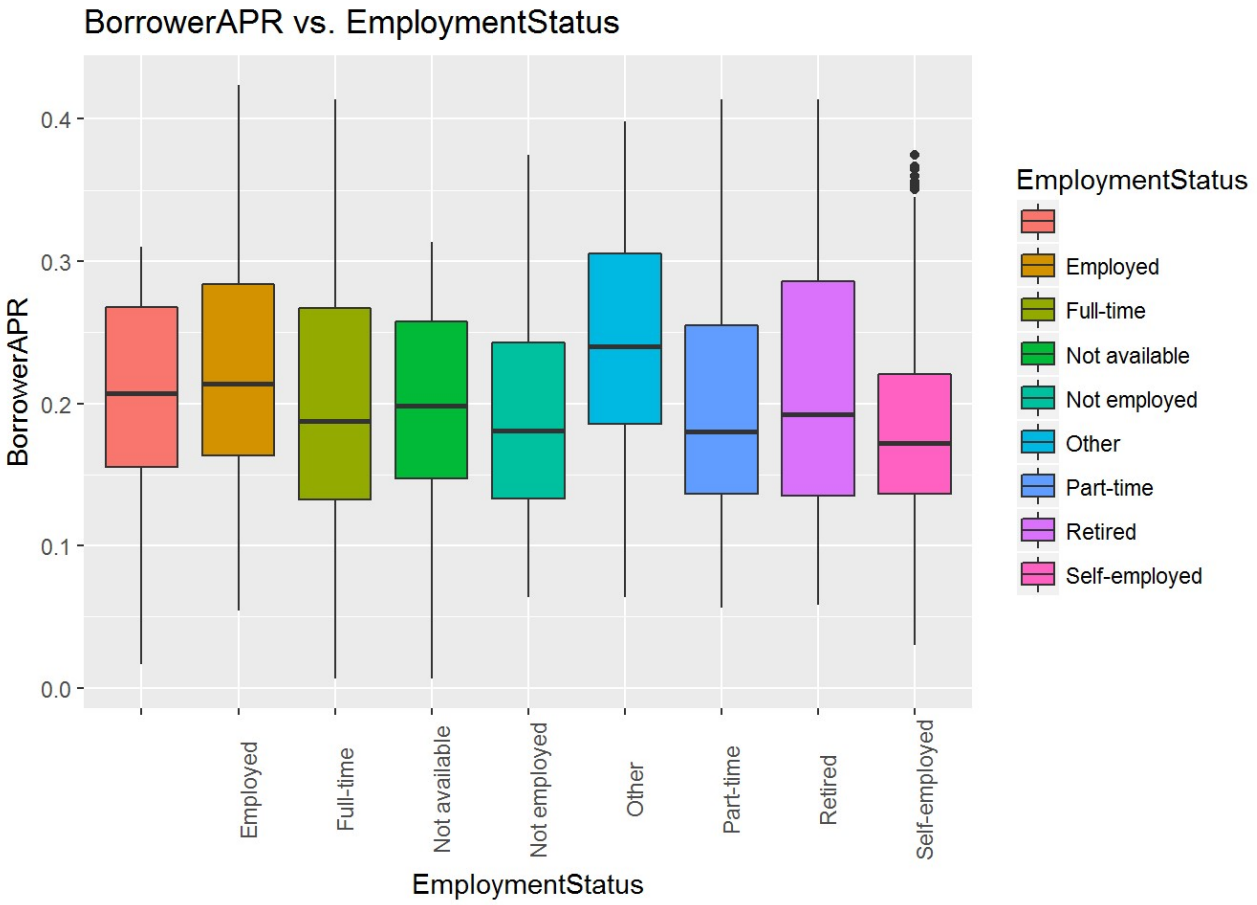
In general, as expected, we see that the higher the credit rating, the lower the borrower APR is. Do we expect the similar relationship with the credit score?

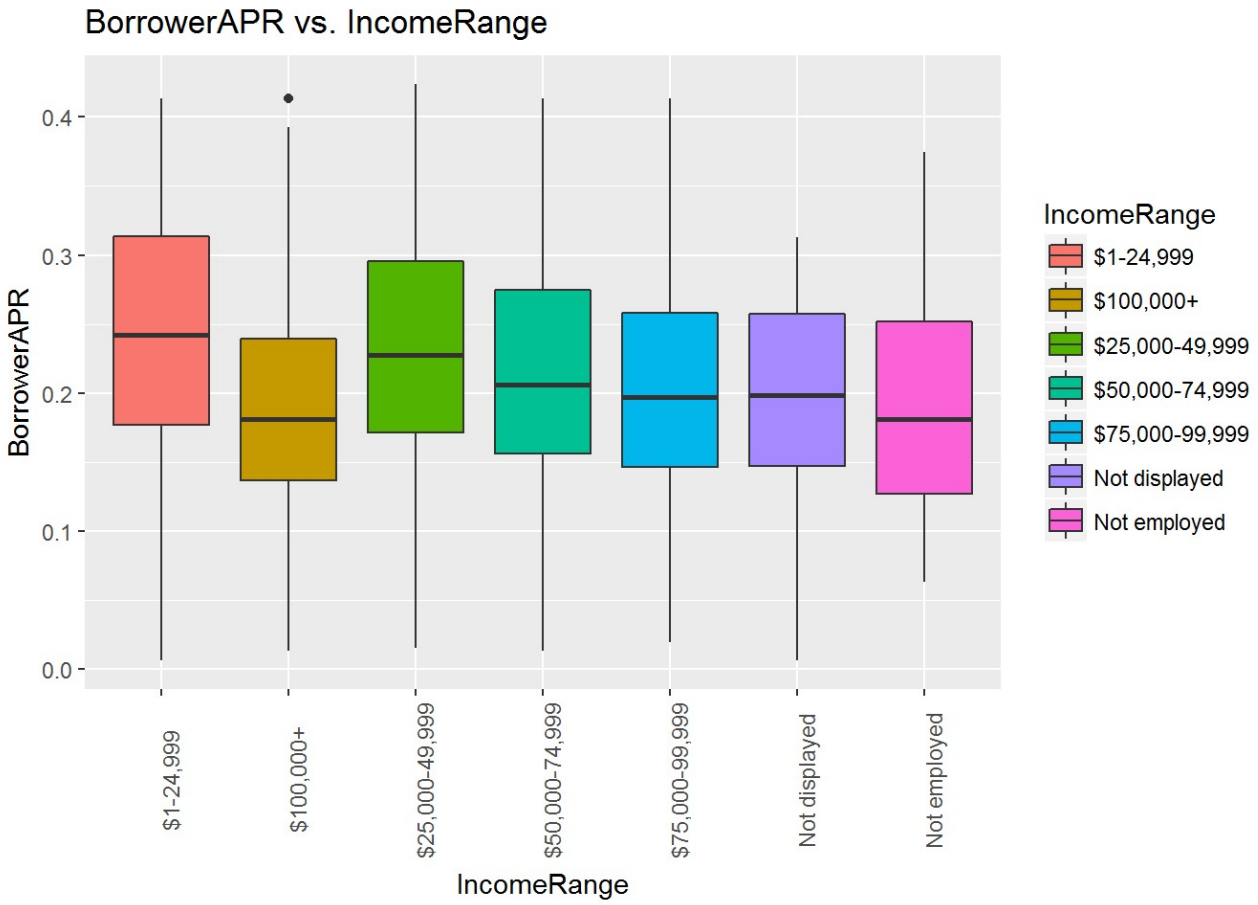


In general, the better the credit score, the lower the borrower APR is. How does the relationship work with the amount of the loan?



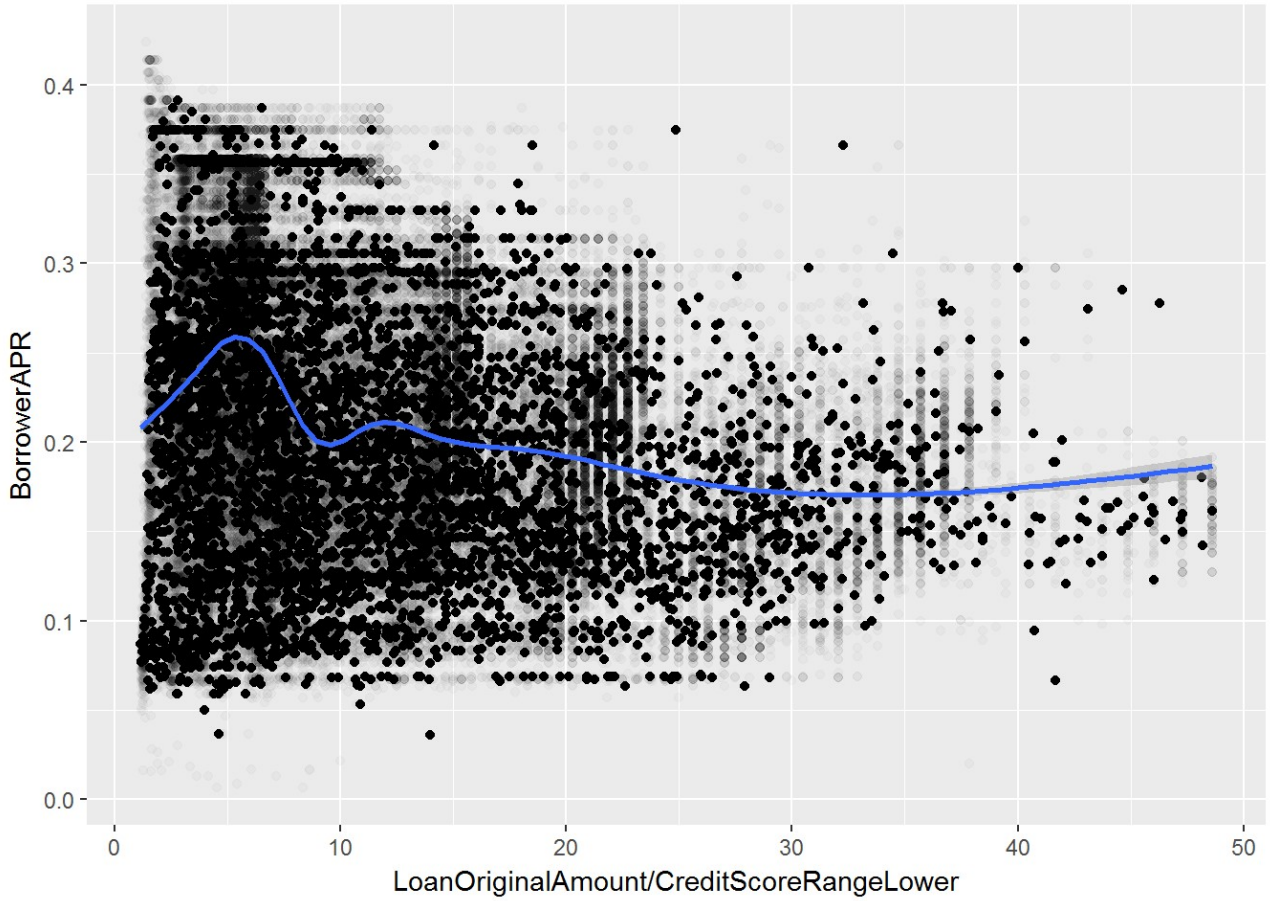
There is a slight downward trend for the borrower APR for the larger loans. But, there are probably other variables that impact the APR more than the loan alone. Maybe there are some dependencies on the employment status and income range?


BorrowerAPR vs. EmploymentStatus
BorrowerAPR
0.4
0.3
0.2
0.1
0.0
Employed
Full-time
Not available
Not employed
Other
Part-time
Retired
Self-employed
EmploymentStatus
EmploymentStatus
Employed
Full-time
Not available
Not employed
Other
Part-time
Retired
Self-employed



As the above two plots indicate, the employment status is rather less weeky related to the APR than the income range. Higher income range borrowers enjoy lower APRs. However, we cannot say similary for the employed borrowers. It is possible that some of the borrowers who are not employed may have higher monthly income and therefore enjoy lower APR than some of the employed people.

Can I explore any complex relationship involving multiple variables to the APR? I create one composite variable involving the loan amount and the credit score. Lets plot to explore.



The distribution of the APR w.r.t this composite variable does not show any prominent trend. I would think that there exists more compex relationship among the variables that determines the APR and we do not know that yet.

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?

We see that the borrower's APR varies with credit rating, credit score range, loan amount and the term of the loan. However, such relationships are not linear. Some more analysis is due, involving more variables together before we can try to predict any outcome.

Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

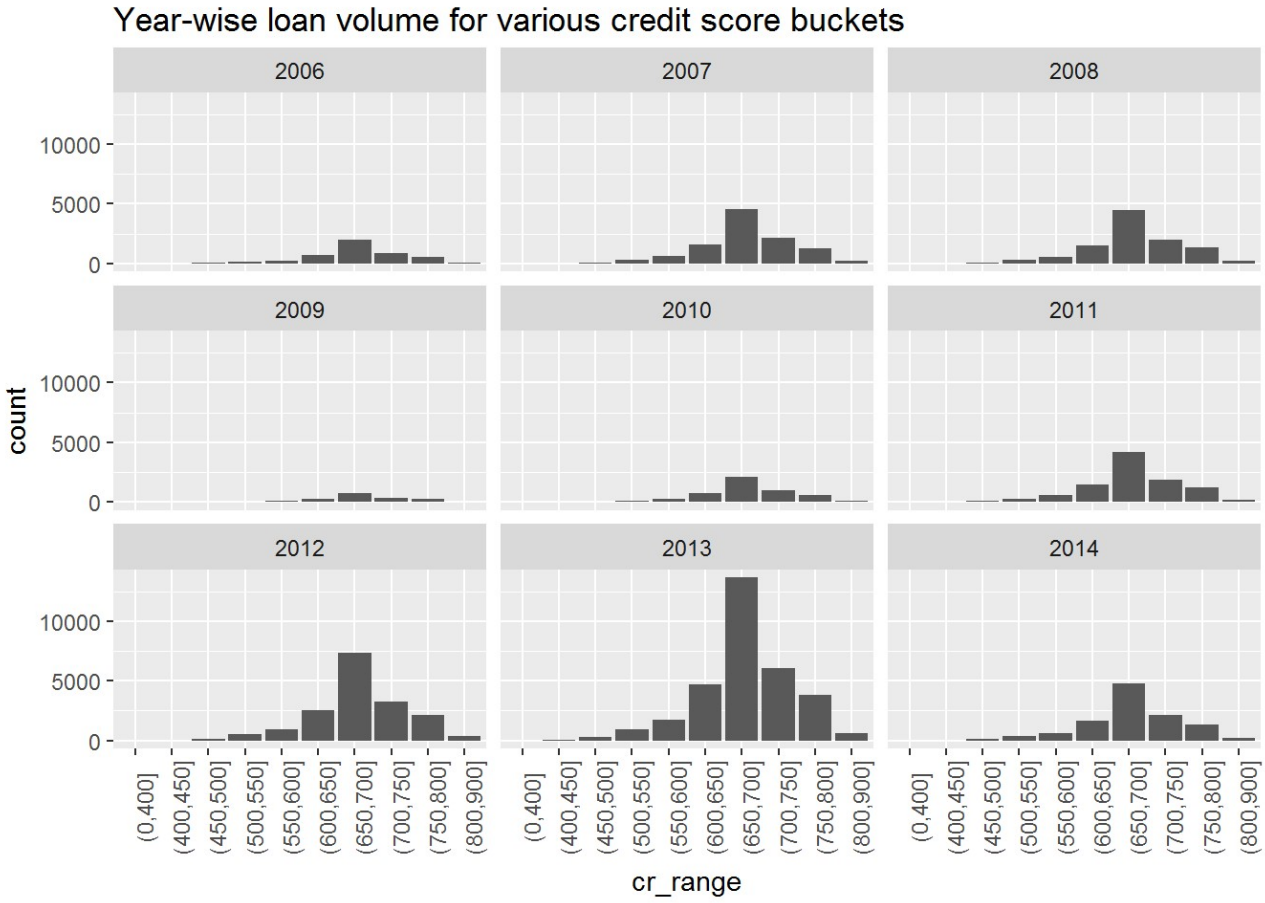
As expected, we see a positive correlation between the monthly loan payment and the total original loan amount. The plots also reveal that the majority of the loans (both number and the $ amount) were taken for debt consolidation.

What was the strongest relationship you found?

The strongest relationship is probably between the BorrowerAPR and the ProsperScore. However, I understand that it is difficult to consider this observation for predicting the borrower's APR since, ProsperScore itself is an outcome. Anyway this is an observation that came out strongly from the correlation analysis without having any knowledge of the ProsperScore.

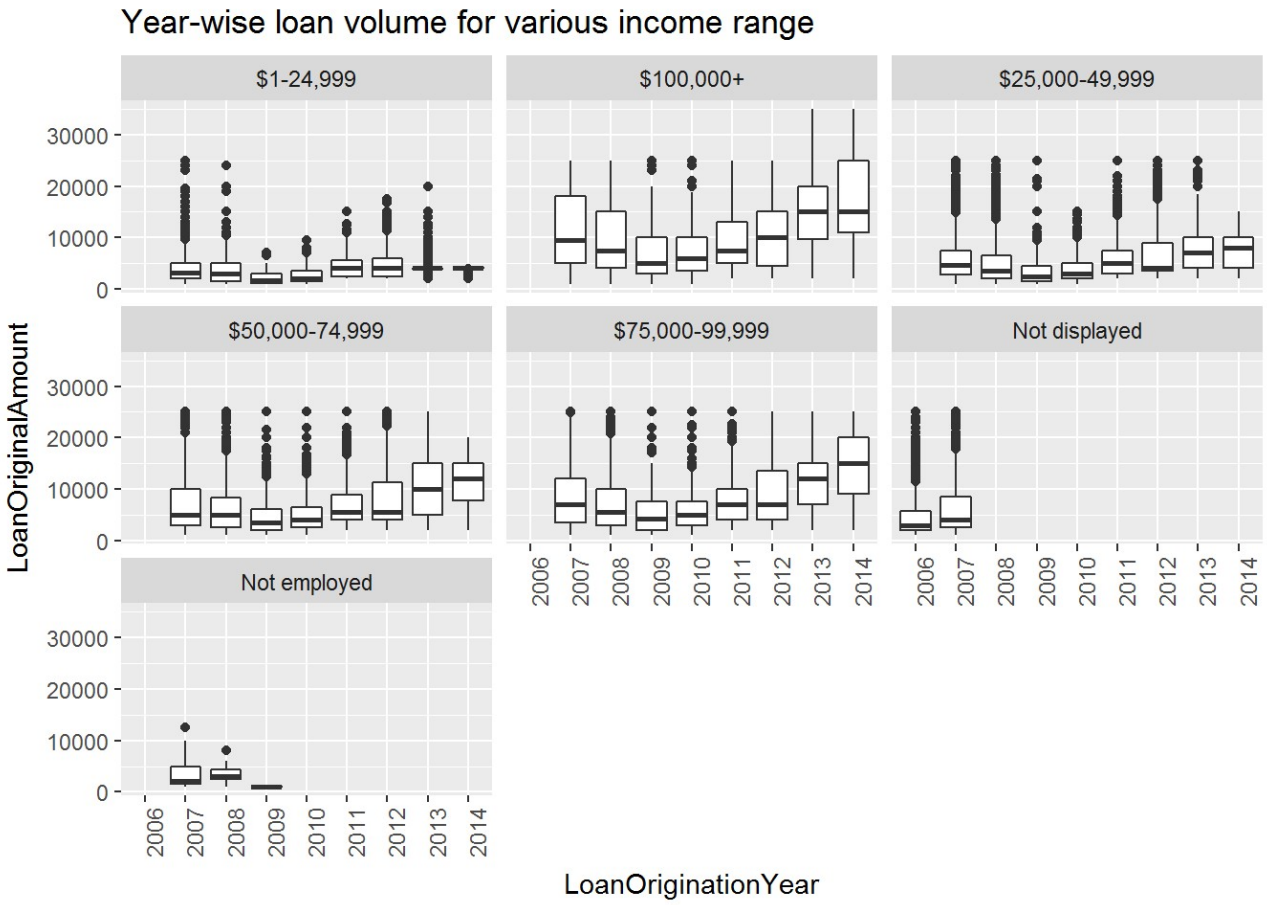
Multivariate Plots Section

Curious to know if there has been any shift in lending practice over the years that considered the credit scores differently. Lets see how the loans were issued over the years.



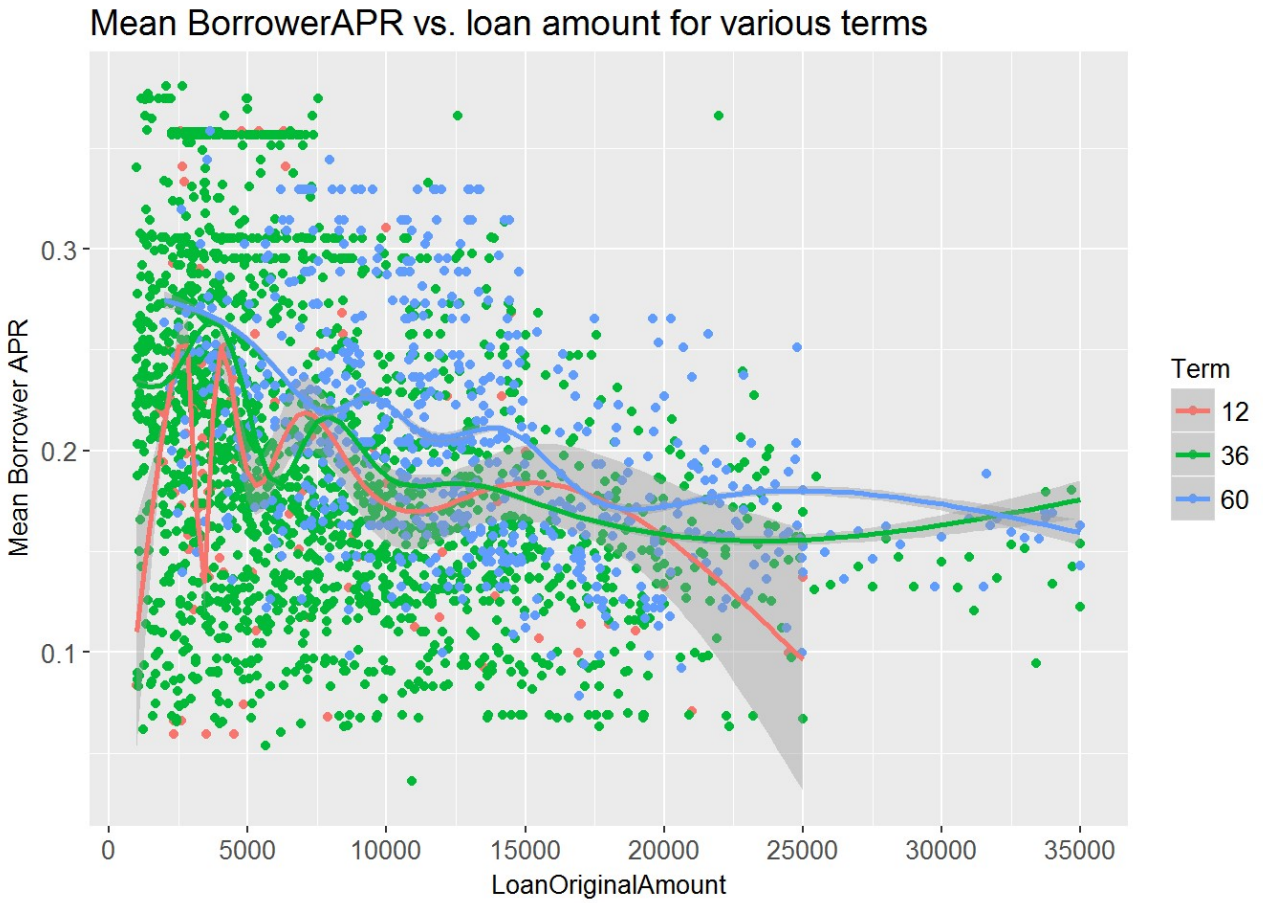
The distribution of the loans w.r.t the credit scores continue to be normal over all the years, including 2009 when the business was at its bottom. It appears that Prosper maintained a uniform strategy of issueing the loans and neither becoming too strict nor libraral about the credit scores.

Similar to the credit scores, I am curious to know if there was any shift in lending practice over the years for issuing loans that considered the income range differently. I will try to visualize in the following plot.



We see a pattern of the mean loan volume among the earning group ($25000+). There is a dip in 2009, but has a positive upward trend afterwards.It appears that there is a uniform practice followed for handling the loans for various income categories.

As our conventional wisdom says, the shorter the term of the loan is, the better becomes the APR. I would like to visualize this in the next plot.

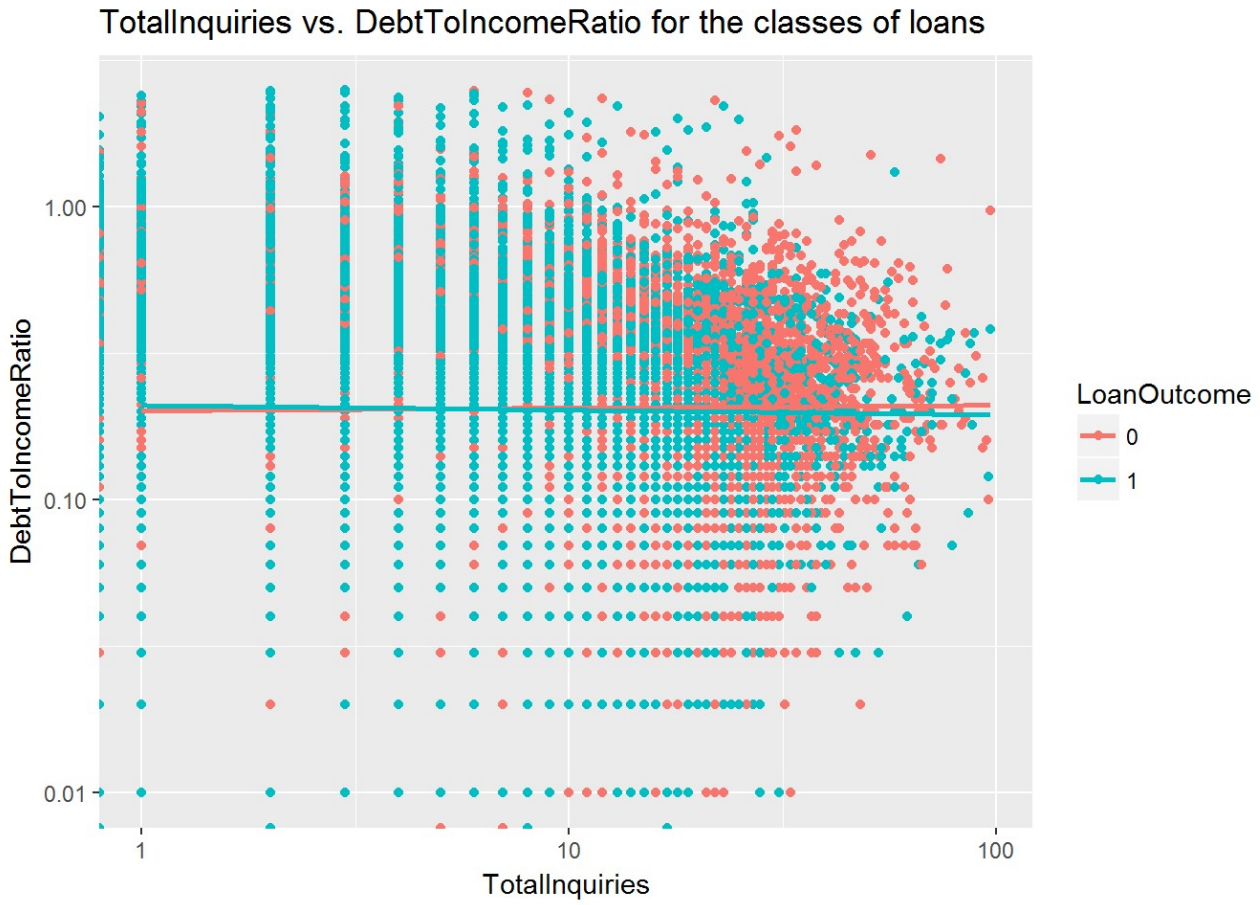


The above plot shows a distribution of the APR for loans with various terms. There is a lot of variance in the data, however, there is a clear trend that the loans with lower terms resulted into the borrowers having lower APRs.

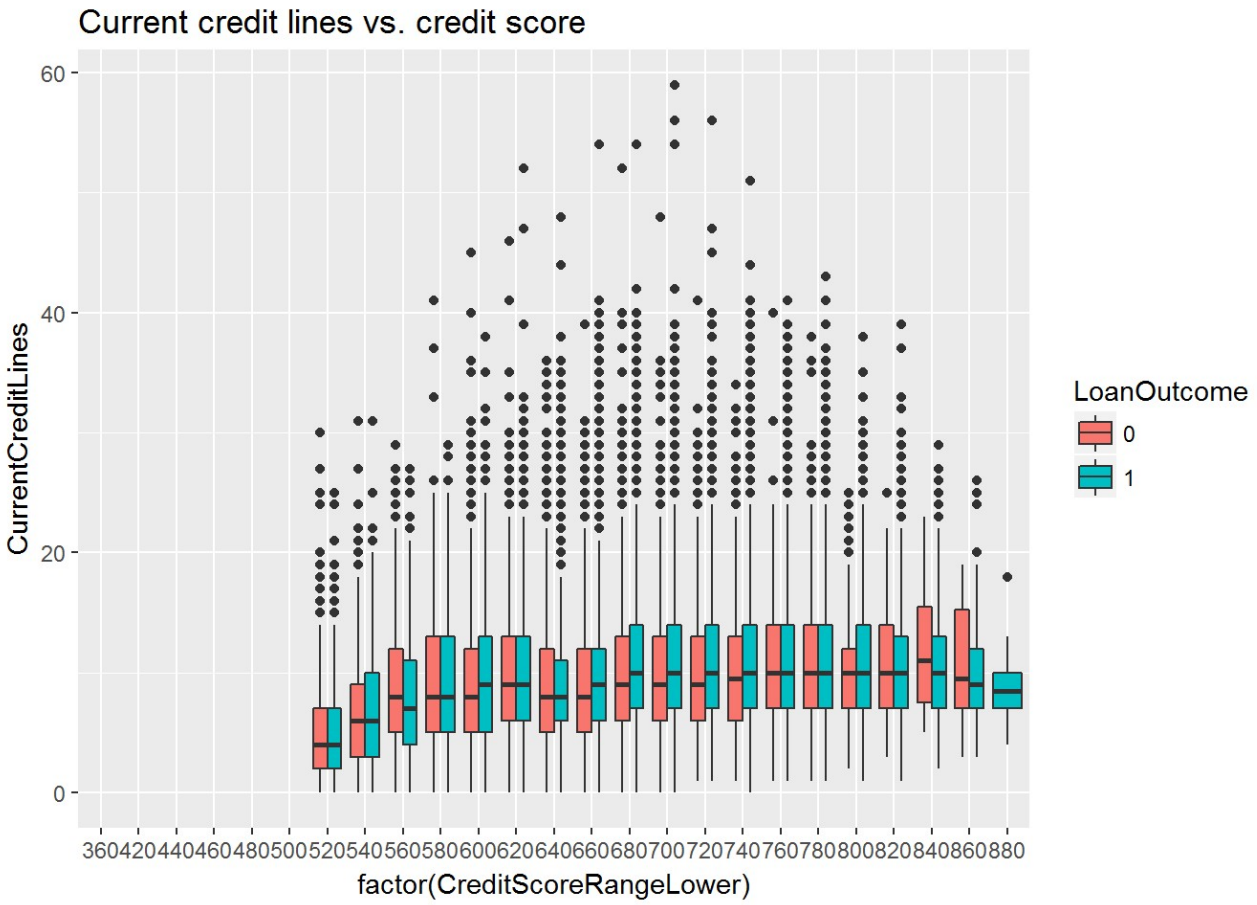
In this section I am going to work with the loans that are (i) ChrgedOff, (ii) Defaulted or (iii) Delinquent.Such loans are represented by the newly introduced variable loandata$LoanOutcome = 0. (The value is '1' for all the loans with a different status). Lets start by finding out the distribution of the defaulted or delinquent loans.

```
##                          x  freq
## 1                Cancelled     1
## 2              Chargedoff 10416
## 3                Defaulted  4413
## 4   Past Due (>120 days)    14
## 5   Past Due (1-15 days)   716
## 6  Past Due (16-30 days)   239
## 7  Past Due (31-60 days)   325
## 8  Past Due (61-90 days)   266
## 9 Past Due (91-120 days)   276
```

I am curious to know the relationship of the total number of credit inquiries, DTI and the outcome of the loan. Here I try to plot them to understand the trend in the loan outcome.

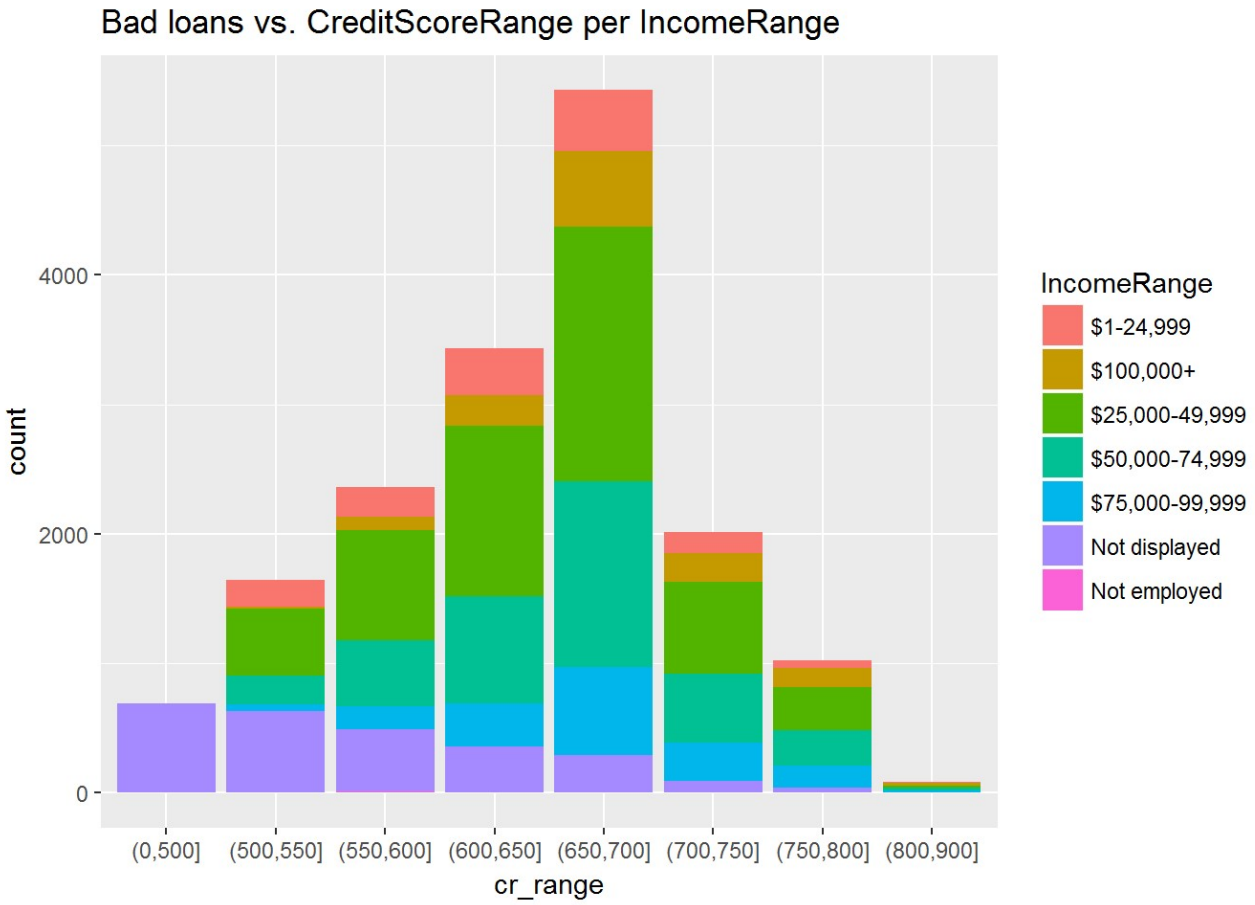


The above plot displays higher concentration of defauled borrowers with higher number of inquiries and DTI. What could we see as a similar relationship between credit scores of the borrowers and the number of credit lines they carry. The more they credit lines, the more dedt the borrowers have and need to repay. Does that result into the loan outcome one way or the other?



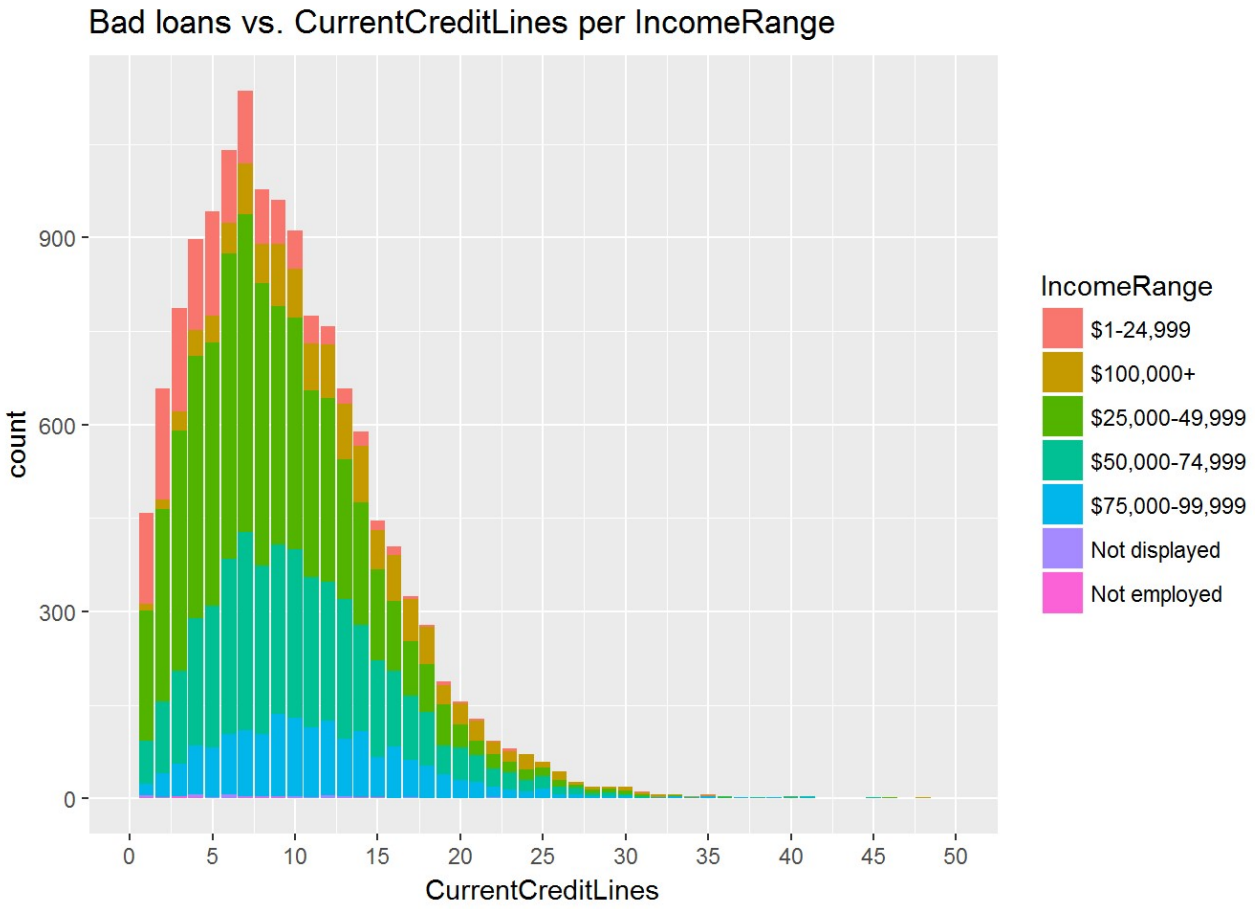
In general the number of credit lines have no prominent effect on the outcome of the loans as appears from the above plot.

Now focusing only on the ‘bad’ loans, I wanted to know if any specific credit score range with specific income range is more common to result in to loan default or not.



We notice from the plot above, the highest number of loan defaults belong to the borrowers in the income bracket of 25K - 50K irrespective of the credit scores.

In the same line as above - are there borrowers with specific income ranges, carrying specific number of credit lines, who are more prone to defaults?



The above plot shows the highest number of loan defaults are due to the borrowers belonging to the income range 25K - 50K irrespective of the number of credit lines any borrower carries. This is similar to our observation from the previous plot where we examined the credit scores.

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?

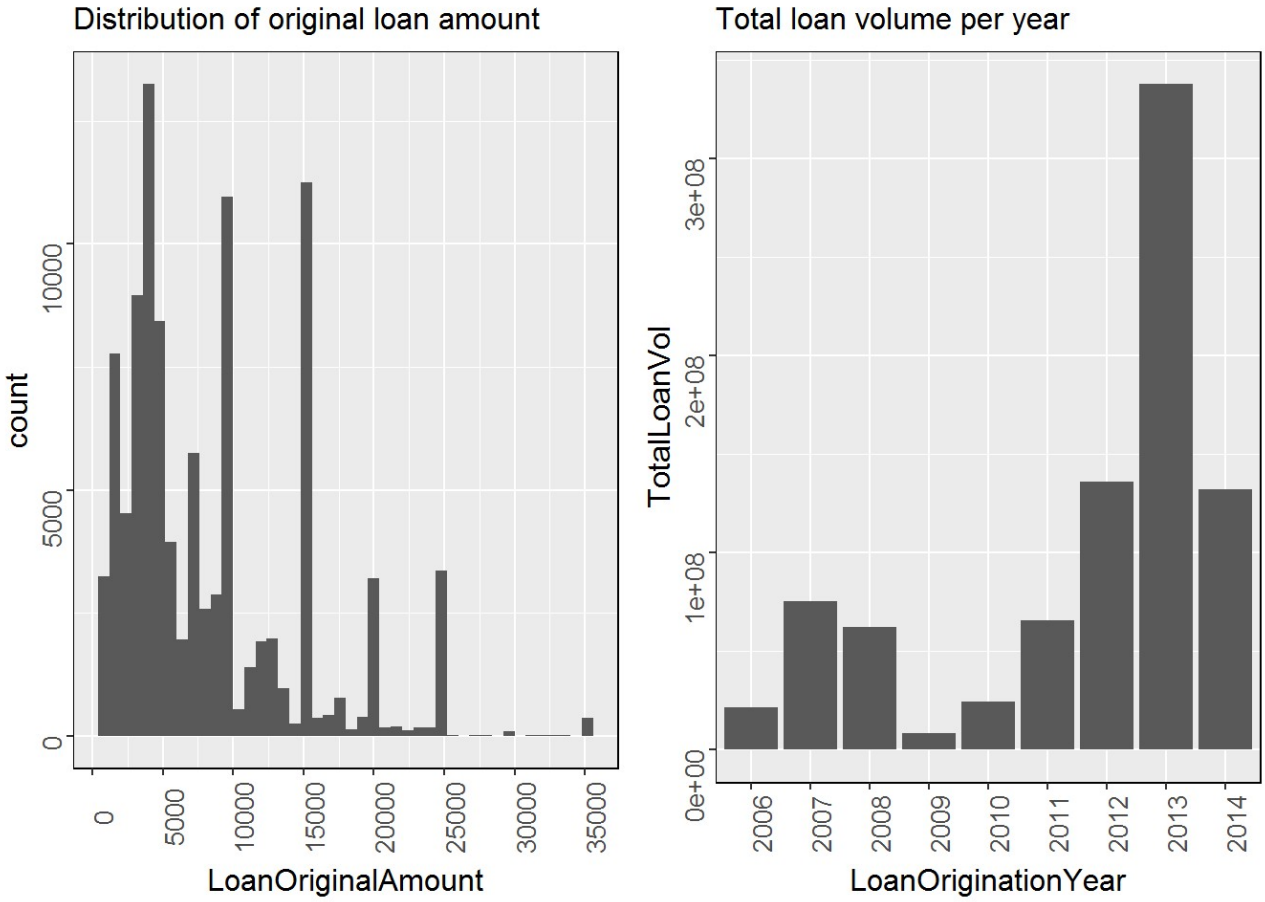
The loan distribution pattern over the creadit scores remain pretty much the same over the years. This is the same as I see for the loan distribution over the income range for all the years. The higher income bracket consistently borrowed larger loans for all the years. I would have liked to see more prominent trend for the BorrowerAPR for various terms, though.

Were there any interesting or surprising interactions between features?

The classification of the loans into two different classes, however, was an interesting way to analyze the dataset. some characteristics are distincly different for the borrowers who completed the loans vs. those who defaulted. I wish I could leverage this classification strategy more to be able to predict the outcome of the loan. Maybe the next Machine Learning course will be helpful towards that direction.

Final Plots and Summary

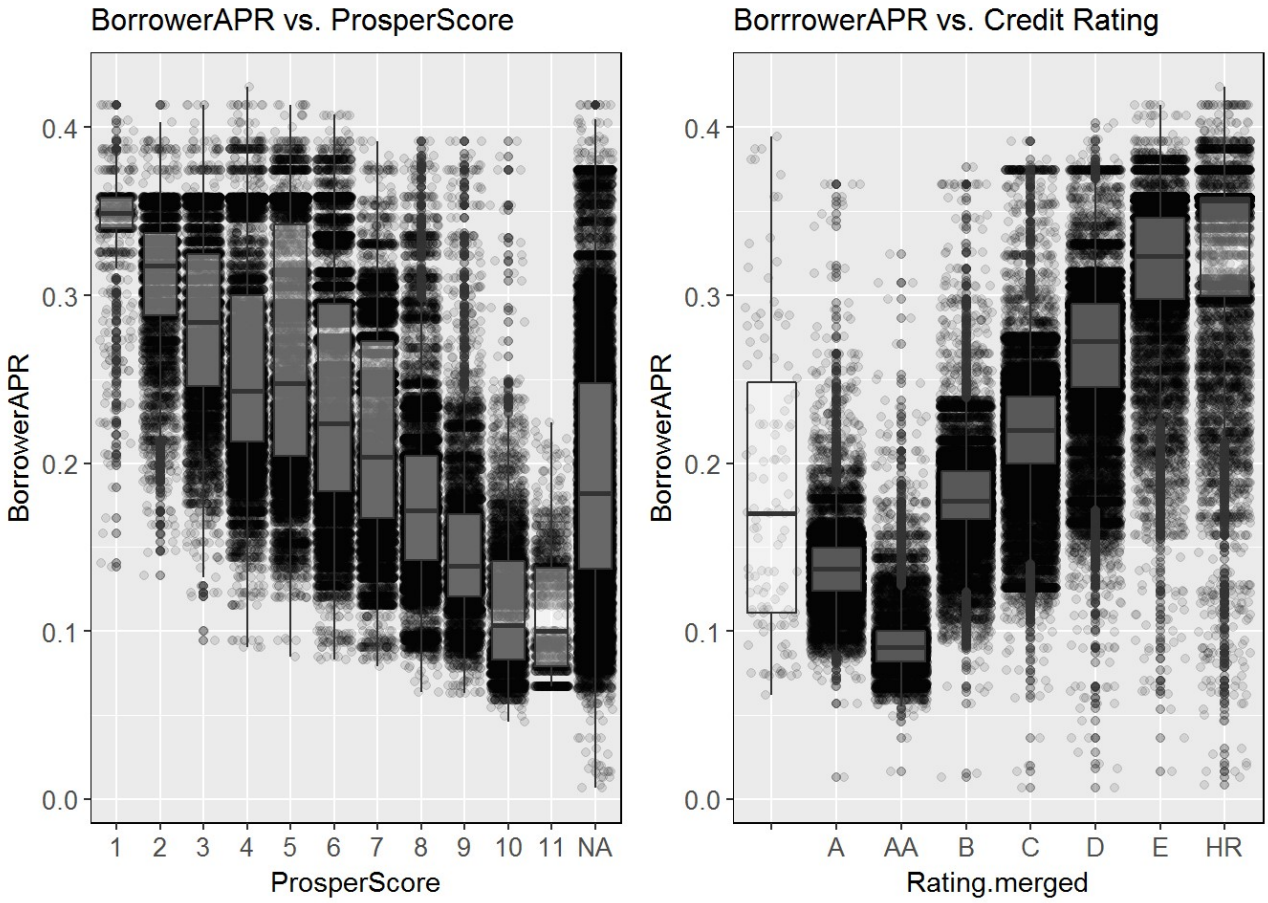
Plot One



Description One

This company primarily issues small loans around or below $25000. Surely, it focused on this segment of the market for small borrowers for non-residential personal loans. In general, we see a positive upward trend in the business procured by Prosper over the years. However, the 2014 data may be incomplete and collected only for part of the year.

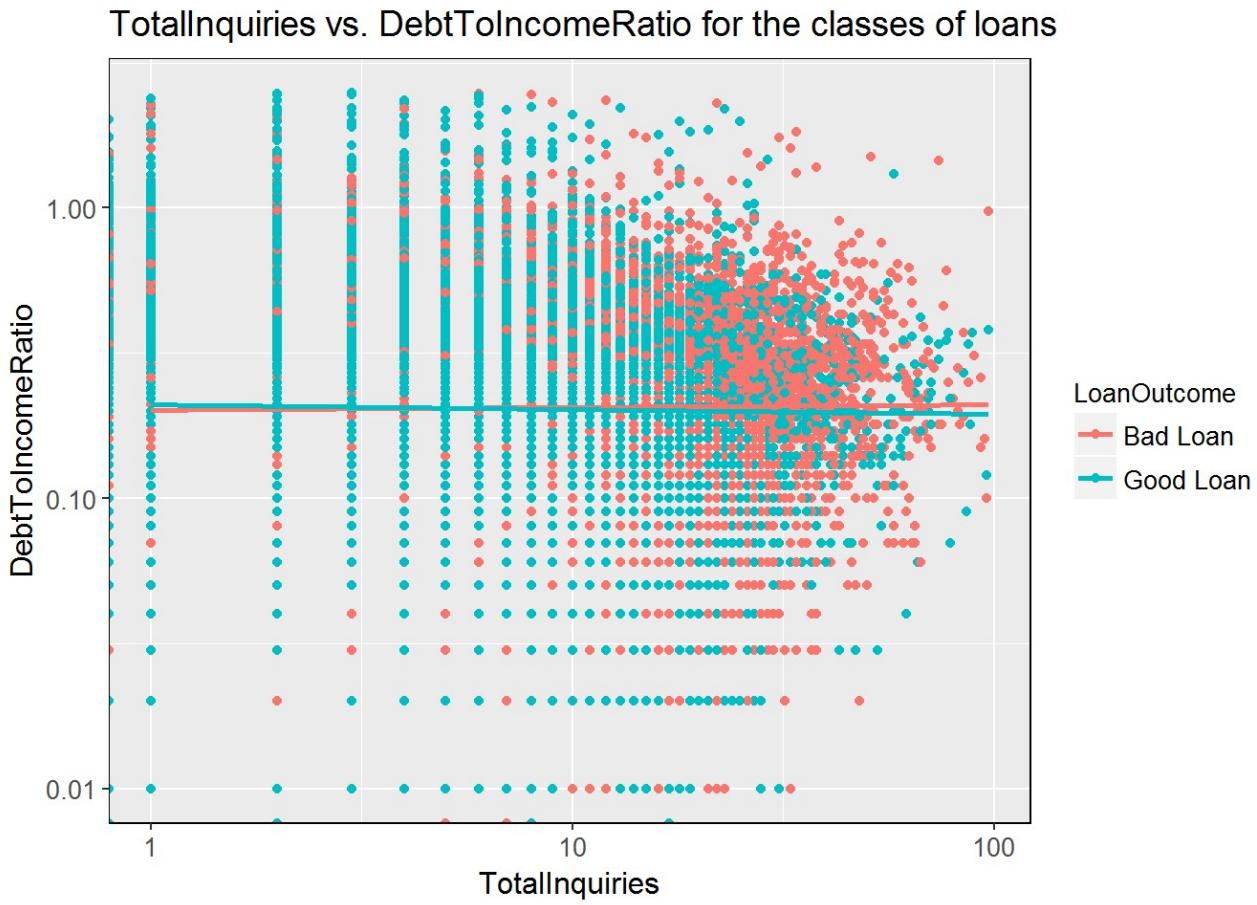
Plot Two

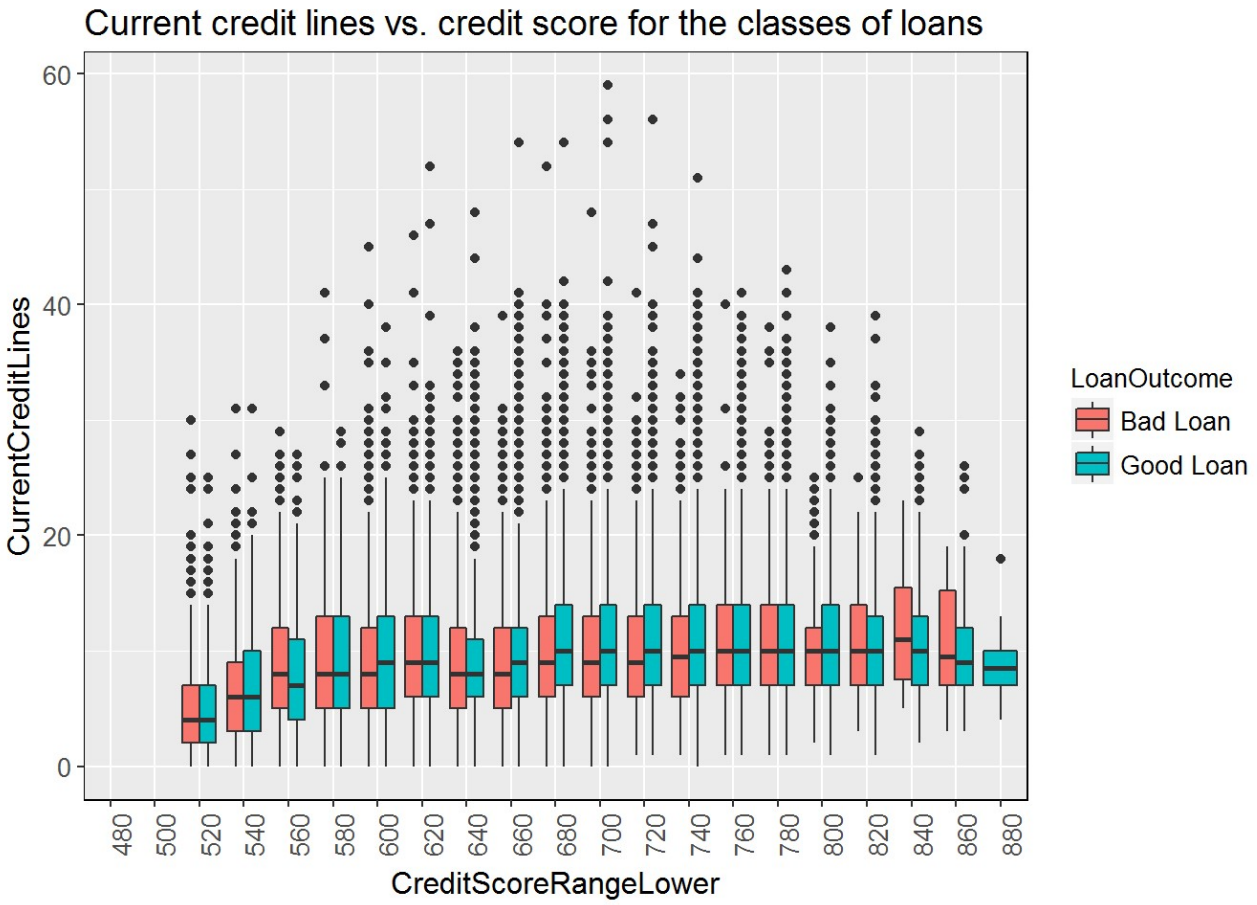


Description Two

In this section I tried to understand any relationship between the the BorrowerAPR and the remaining variables. I could find the strongest correlation with the ProsperScore. However, I understand that the prosperScore itself is an outcome variable for this company which is based on their proprietary mathematical formulation of variables. I have tried to plot several graphs to display how the APR varies. I could see a definitive trend as is evident in the graph that plots the Credit rating.

Plot Three


TotalInquiries vs. DebtToIncomeRatio for the classes of loans
DebtToIncomeRatio
1.00
0.10
0.01
1
10
100
TotalInquiries
LoanOutcome
Bad Loan
Good Loan



Description Three

I wanted to work with the two classes of loans - good and bad. Although I could not find a definitive relationship of the outcome of the loan (good/bad), I wanted to go ahead and explore some characteristics of the variables as they are displayed for these two classes of loans.

The first plot shows a tendency of the bad loan borrowers to have more enquiries than the good loan borrowers. The DTI also tends to be more for such borrowers. The second plot shows the the bad loan borrowers are carrying comparable current credit lines with the good loan borrowers having the same credit scores.

These two plots show some expected tendencies that match our intuition. I am sure these variables will have a place in the formula that can predict the outcome of the loan.

Reflection

This loan dataset posed a great opportunity to audit the data in multiple dimensions. A few correlations were somewhat obvious, e.g. APR vs. term of the loan, or, the credit score. However, there were many more that surfaced only after some exploration.

The loan volume is usually small for this peer-to-peer lending company. The APR is usually high for such loans and even higher for the longest term which is 36 months. I eliminated a few outliers from the dataset.

It was not an easy project for me to complete. I struggled to decide which relations are good to showcase and explore. I kept of plotting various variables and looked to find out some patterns. I picked up two goals - determining BorrowerAPR and the ultimate outcome (defaulted/completed) of the loan. I have presented the correlations/trends in this submission. However, I am certainly looking forward to the next Machine Learning course where I am expecting to carry this project further and learn the techniques to predict the outcome of the loan.