# Prosper Loan Data EDA

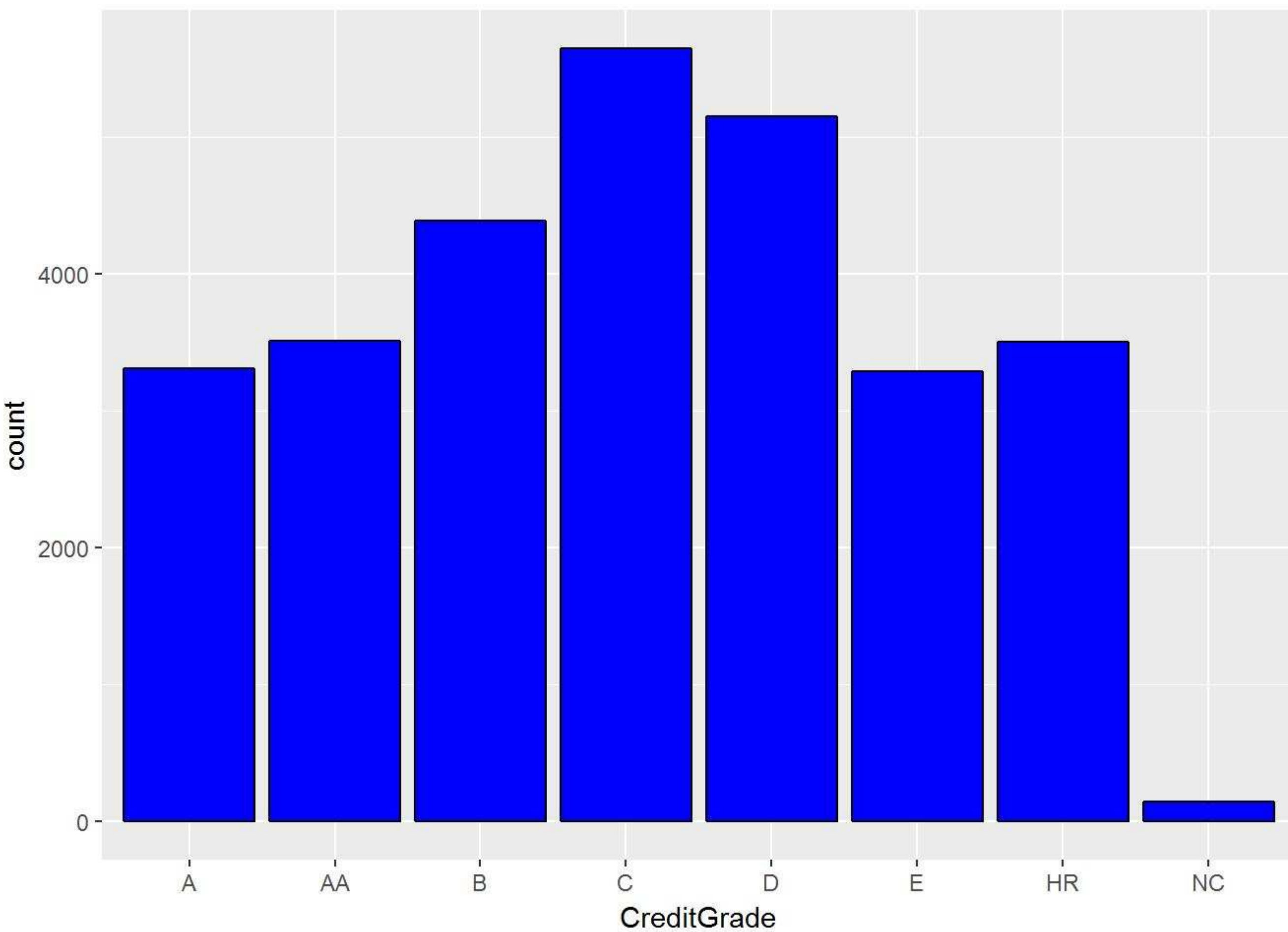
*Shahrooz Govahi*

*March 12, 2017*

============================================================

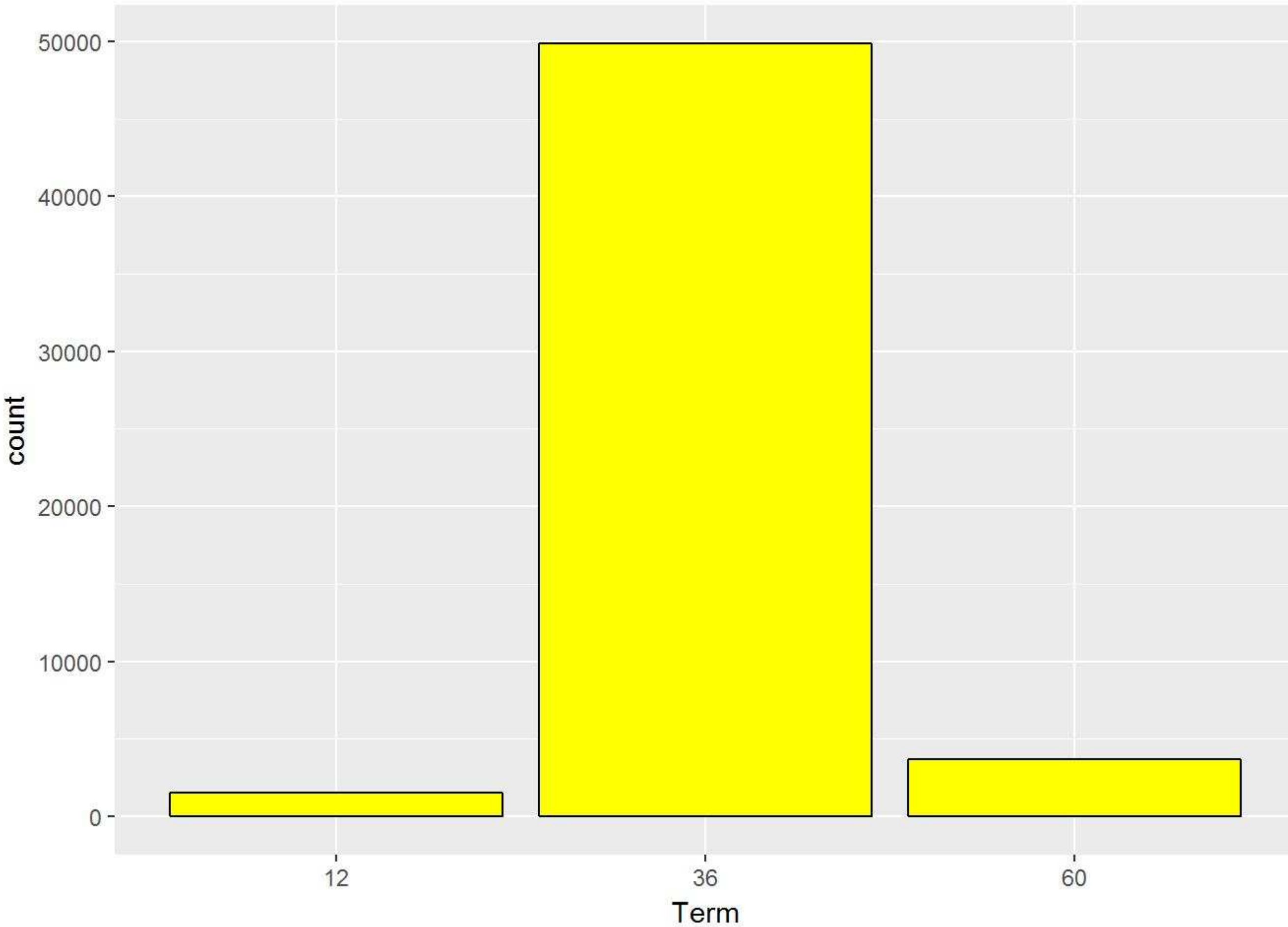
Base on LoanStatus categories; I think one possible goal for this EDA should be looking at different features and predict either that loan will be Completed or Failed. By Failed loans I am referring to chargedoff or Defaulted loans. The other categories are not useful for this purpose. So I am defining a new variable LoanStatusLabel. I will put it's value as Completed if it's Completed and Failed if it's Defaulted or Chargedoff and NA for other categories (e.g.Cancelled, Current).

# Univariate Plots Section



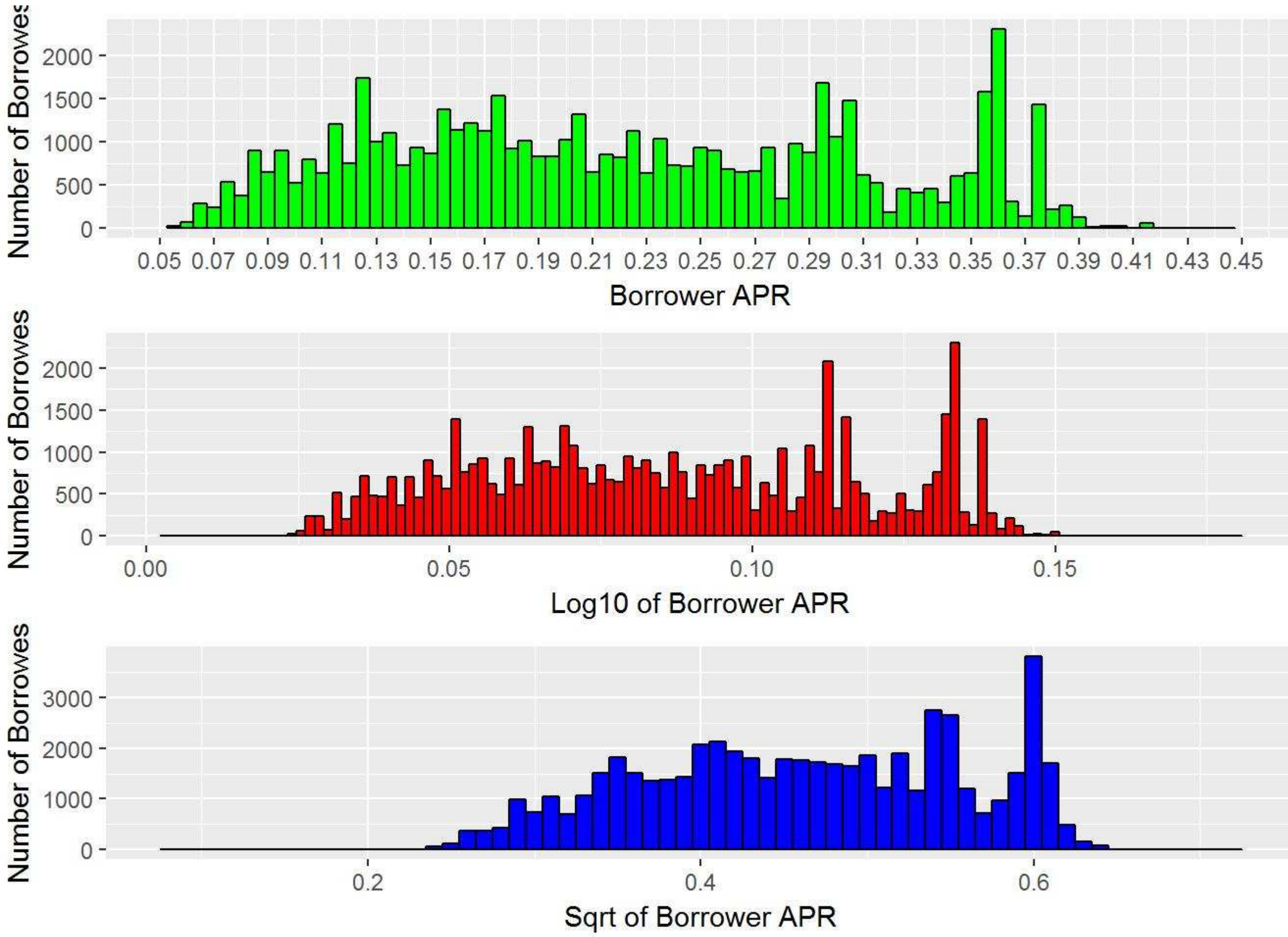

```
##     A    AA     B     C     D     E    HR    NC  NA's 
##  3314  3509  4389  5648  5153  3289  3505   141 26136
```

The majority of Credit Grades in our data set are NAs; Most of Credit Grades are C and the least number of them are NC.



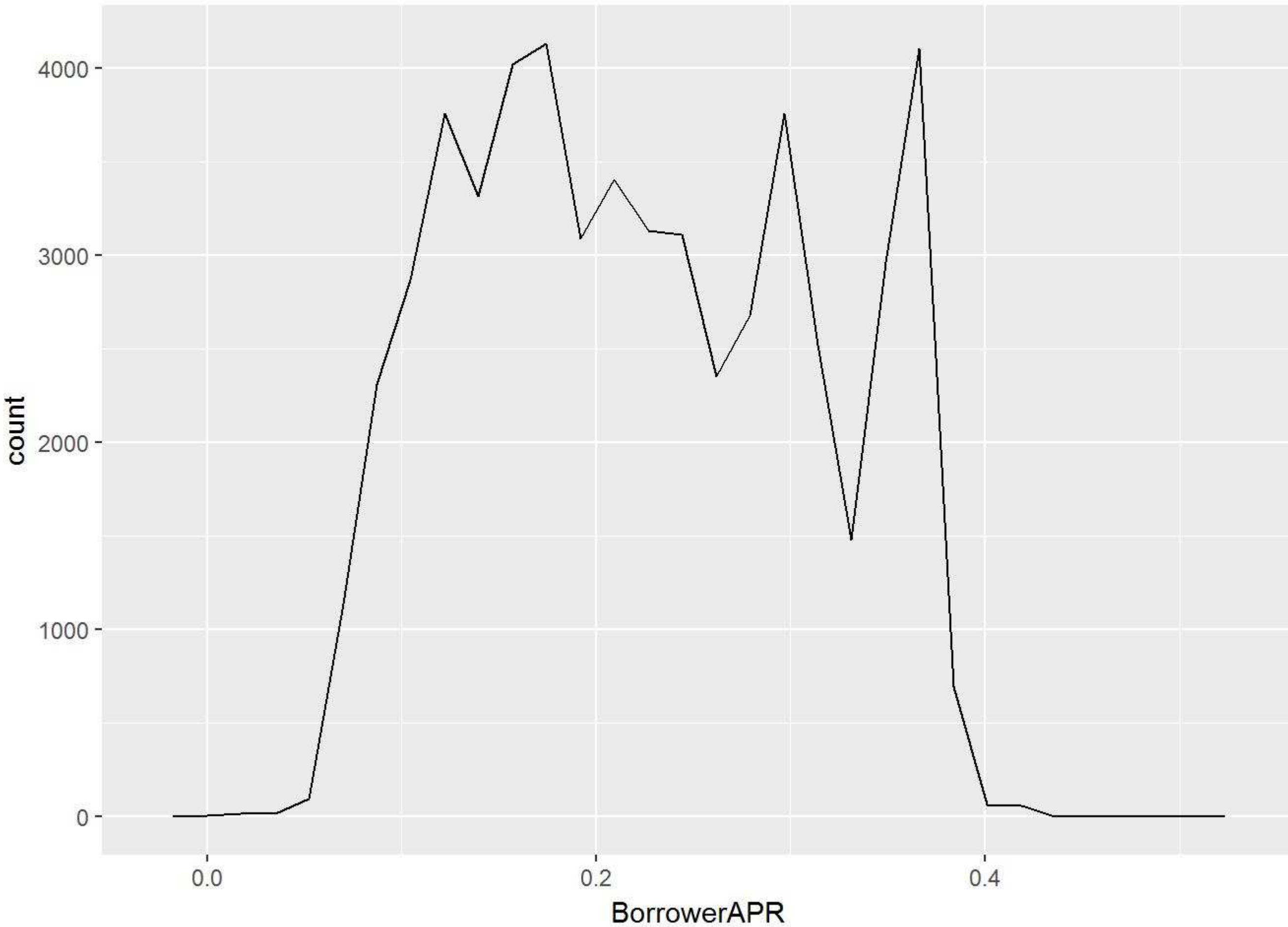

```
##    12    36    60
##  1532 49856  3696
```

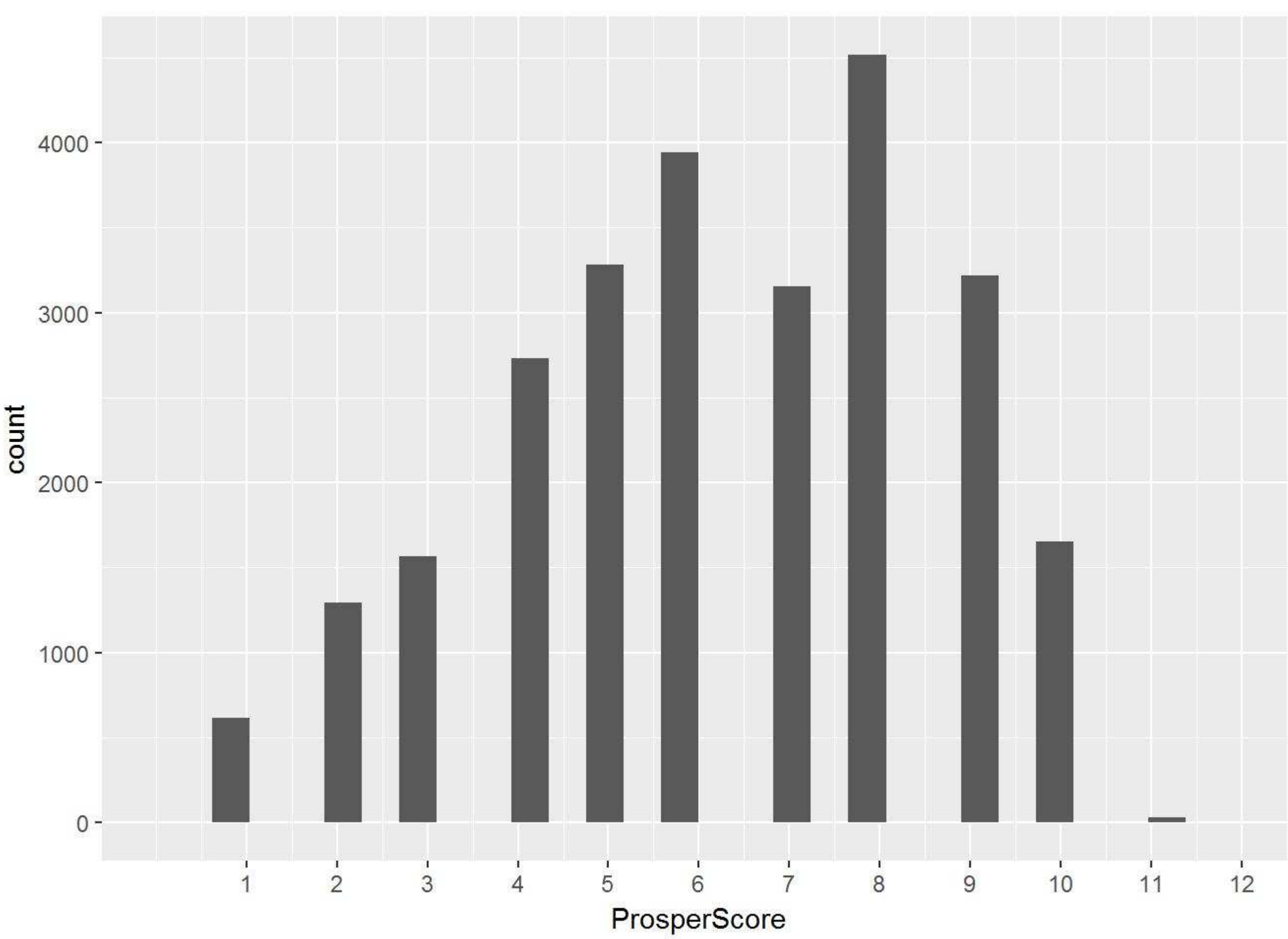
The majority of loan terms are 36 months and less than 2% of them have 12 months term.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
## 0.00653 0.14970 0.21430 0.22220 0.29510 0.51230      25
```

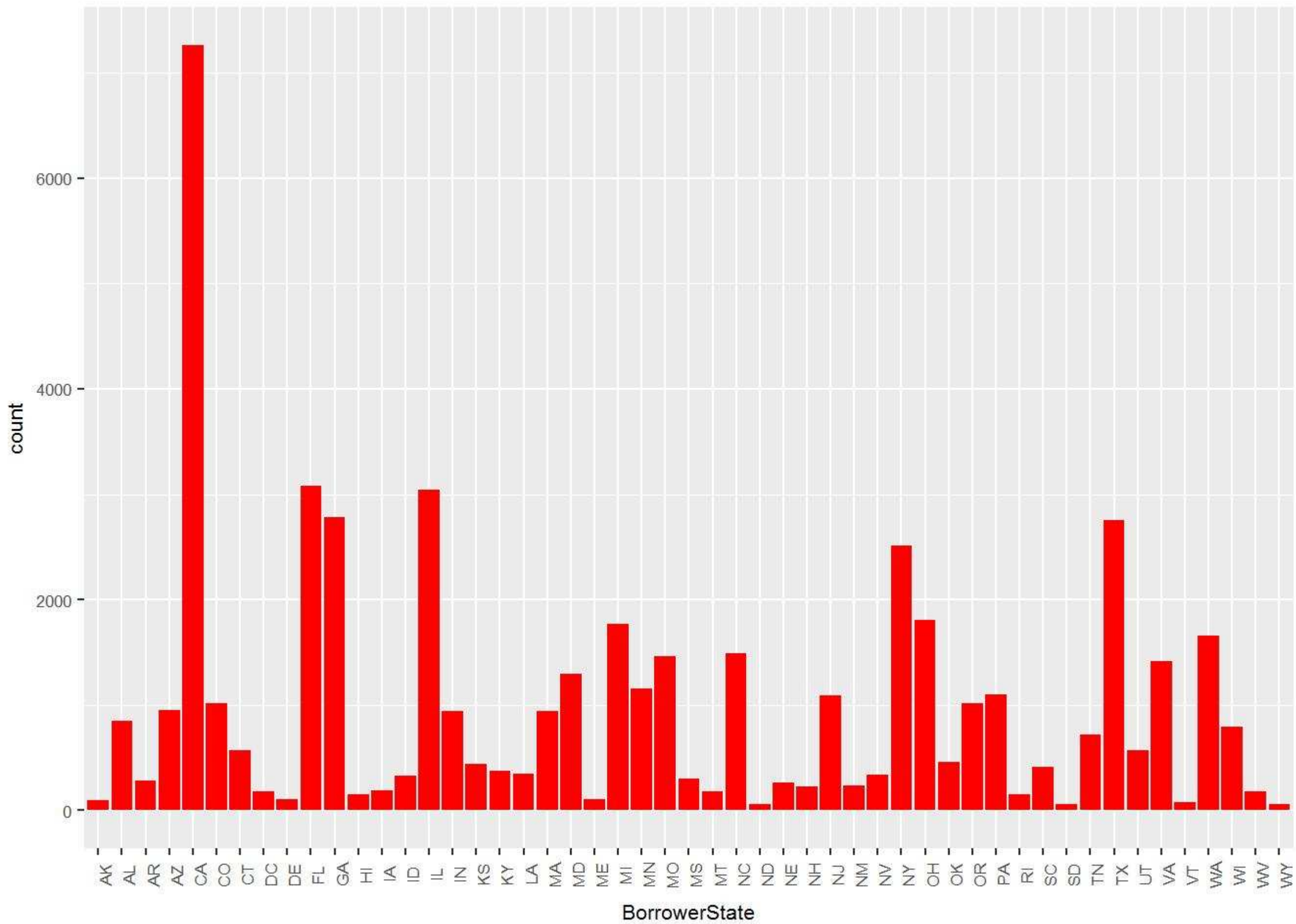
The most frequent Borrower APR is 0.36; By transforming data with Square root, the distribution result is more smoother.


4000
3000
2000
1000
0
count
0.0
0.2
0.4
BorrowerAPR


4000
3000
2000
1000
0
count
1
2
3
4
5
6
7
8
9
10
11
12
ProsperScore

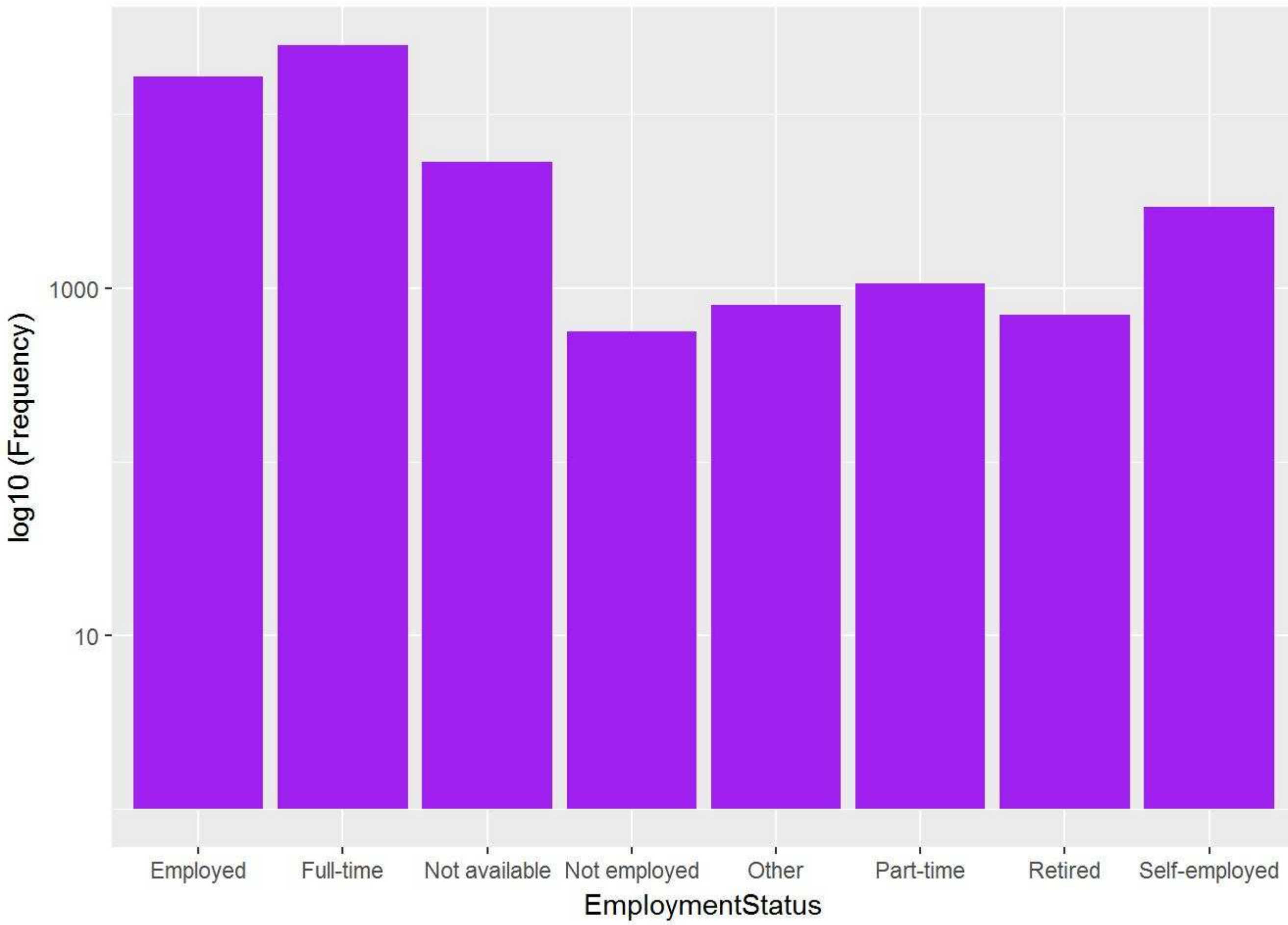
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##   1.000   5.000   6.000   6.266   8.000  11.000   29079 
```

The most frequesnt ProspectScores are 4, 6 and 8; On the other hand, 1 and 11 are the least frequent ones.



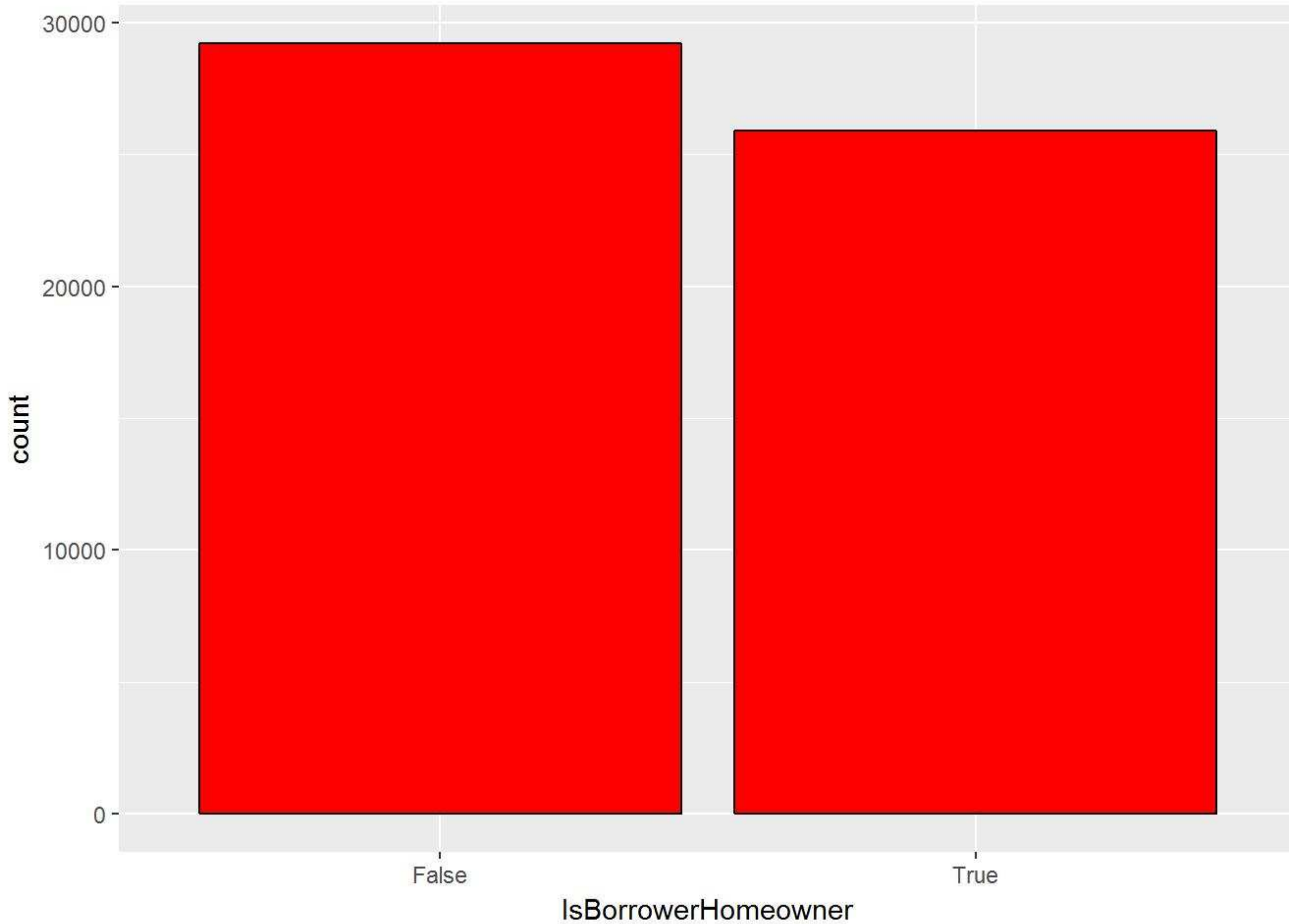

```
##   AK   AL   AR   AZ   CA   CO   CT   DC   DE   FL   GA   HI   IA   ID   IL 
##   93  848  279  946 7263 1013  569  174  103 3077 2783  153  186  328 3039 
##   IN   KS   KY   LA   MA   MD   ME   MI   MN   MO   MS   MT   NC   ND   NE 
##  944  441  370  348  944 1289  101 1767 1151 1459  295  181 1487   52  262 
##   NH   NJ   NM   NV   NY   OH   OK   OR   PA   RI   SC   SD   TN   TX   UT 
##  228 1090  230  339 2515 1808  460 1018 1098  153  414   61  716 2752  569 
##   VA   VT   WA   WI   WV   WY NA's 
## 1417   76 1655  793  178   57 5512 
```

California has the most users of Prosper loan services.



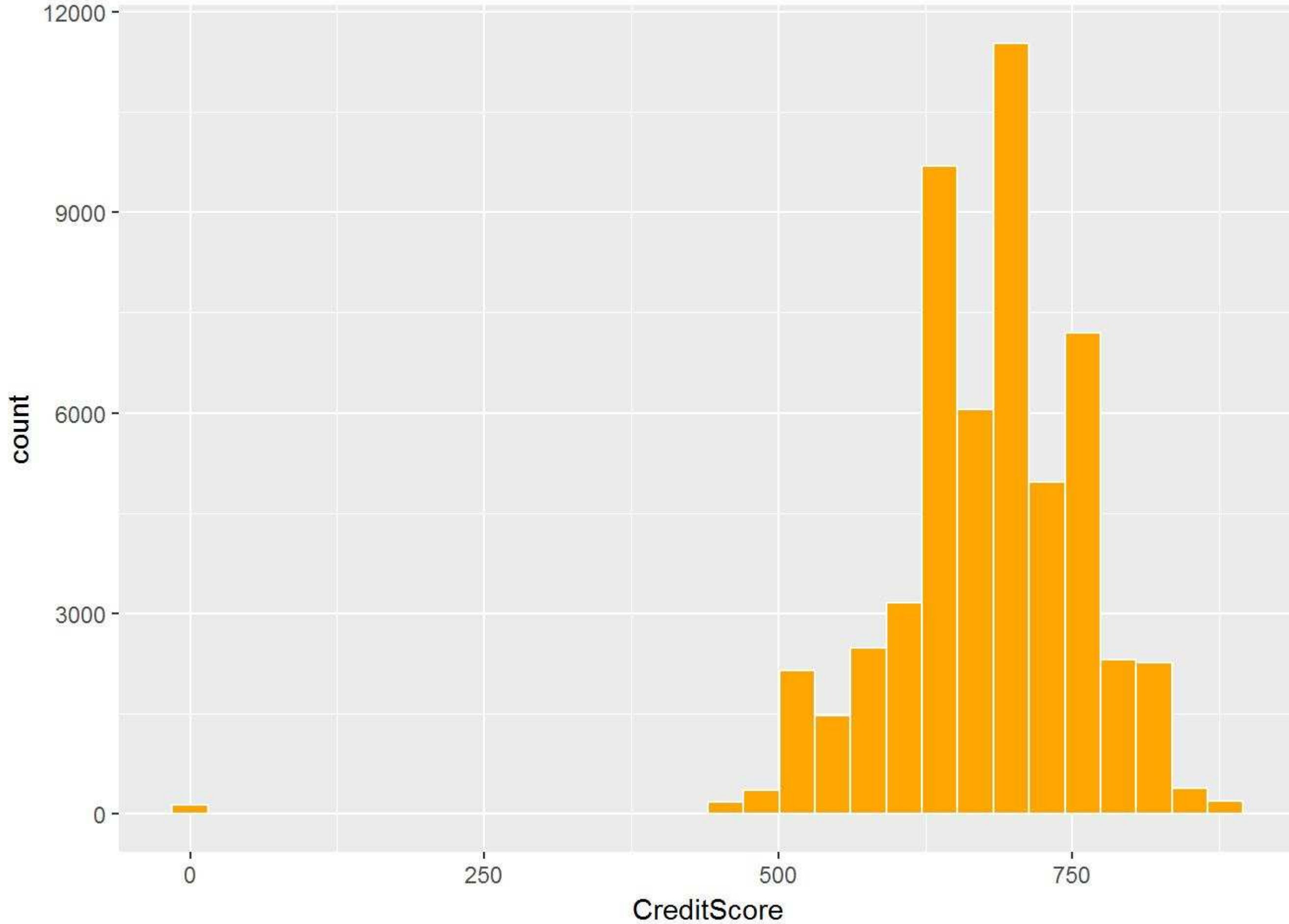

```
##      Employed     Full-time Not available  Not employed         Other
##         16491         24957          5346           561           798
##     Part-time       Retired Self-employed          NA's
##          1056           697          2926          2252
```

Employed category is the most frequent employment status.



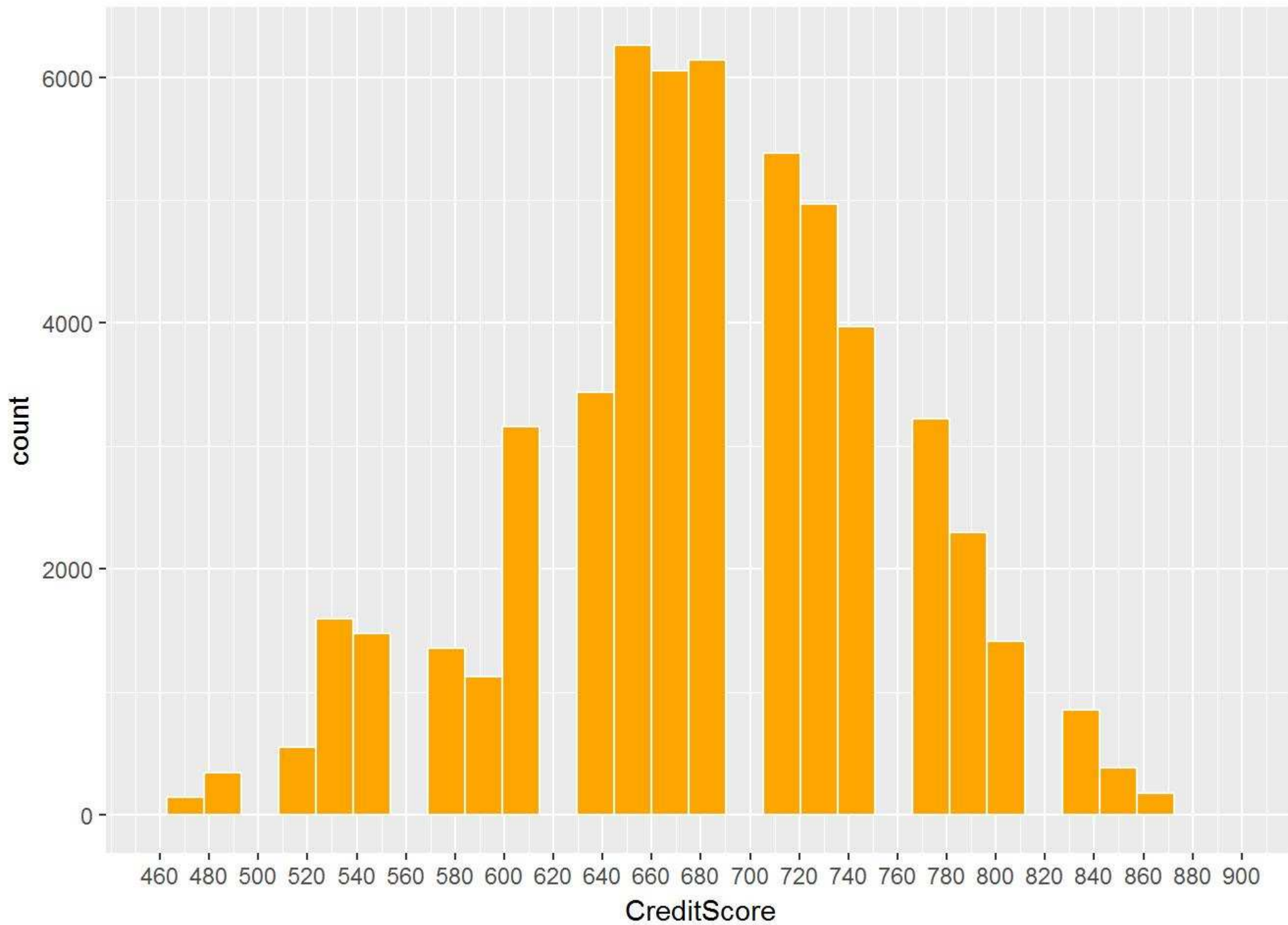

```
## False  True
## 29199 25885
```

Home owners and not home owners numbers are very close to each other;

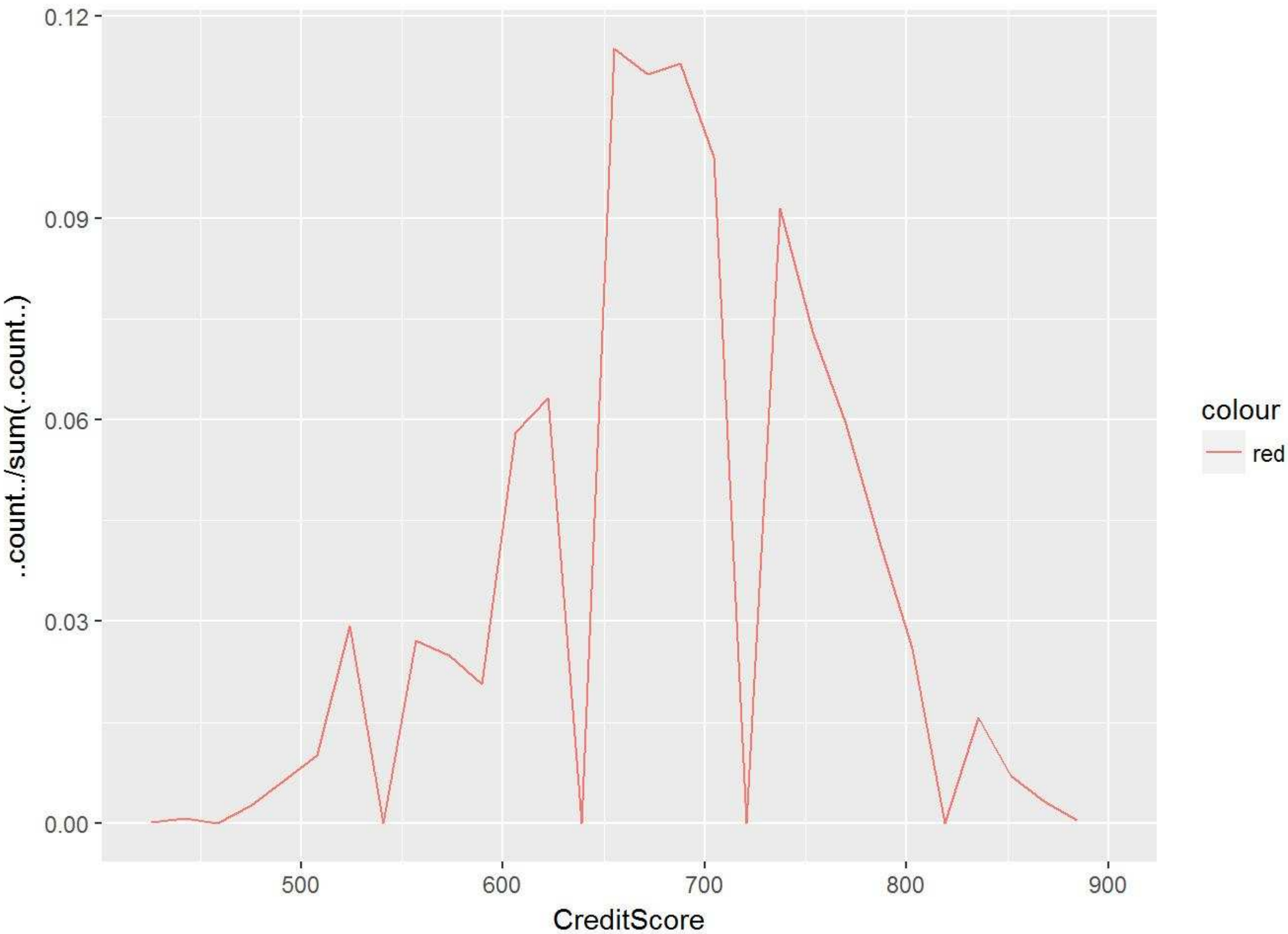



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    10.0   650.0   690.0   681.7   730.0   890.0     590
```

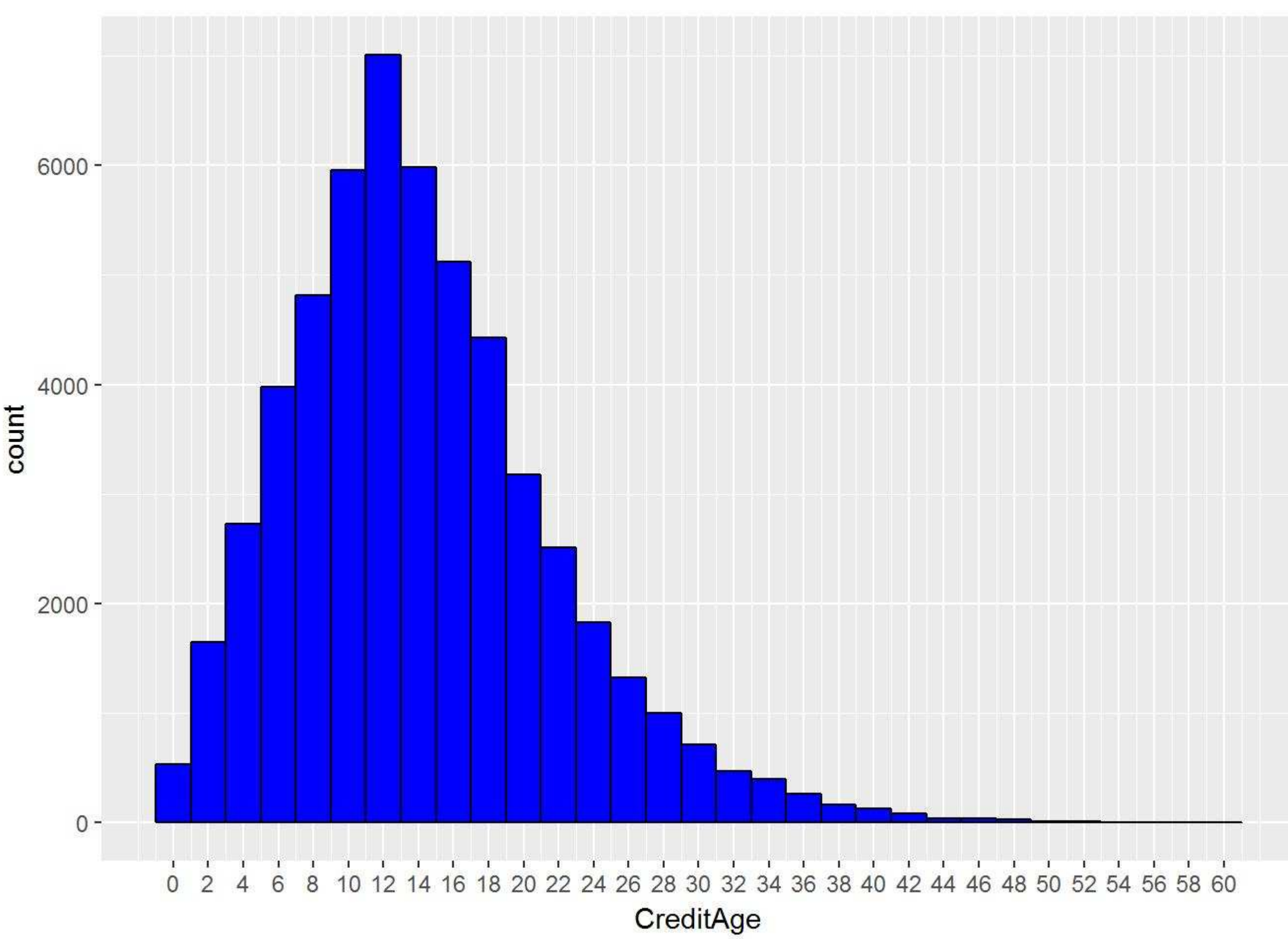
The CreditScore = 10 should be an error in our data; So I will focus on scores more than 460;




There are some gaps between scores. e.g. There is no person with credit score equal to 700 or 560. It seems like credit scores has 7 clusters.

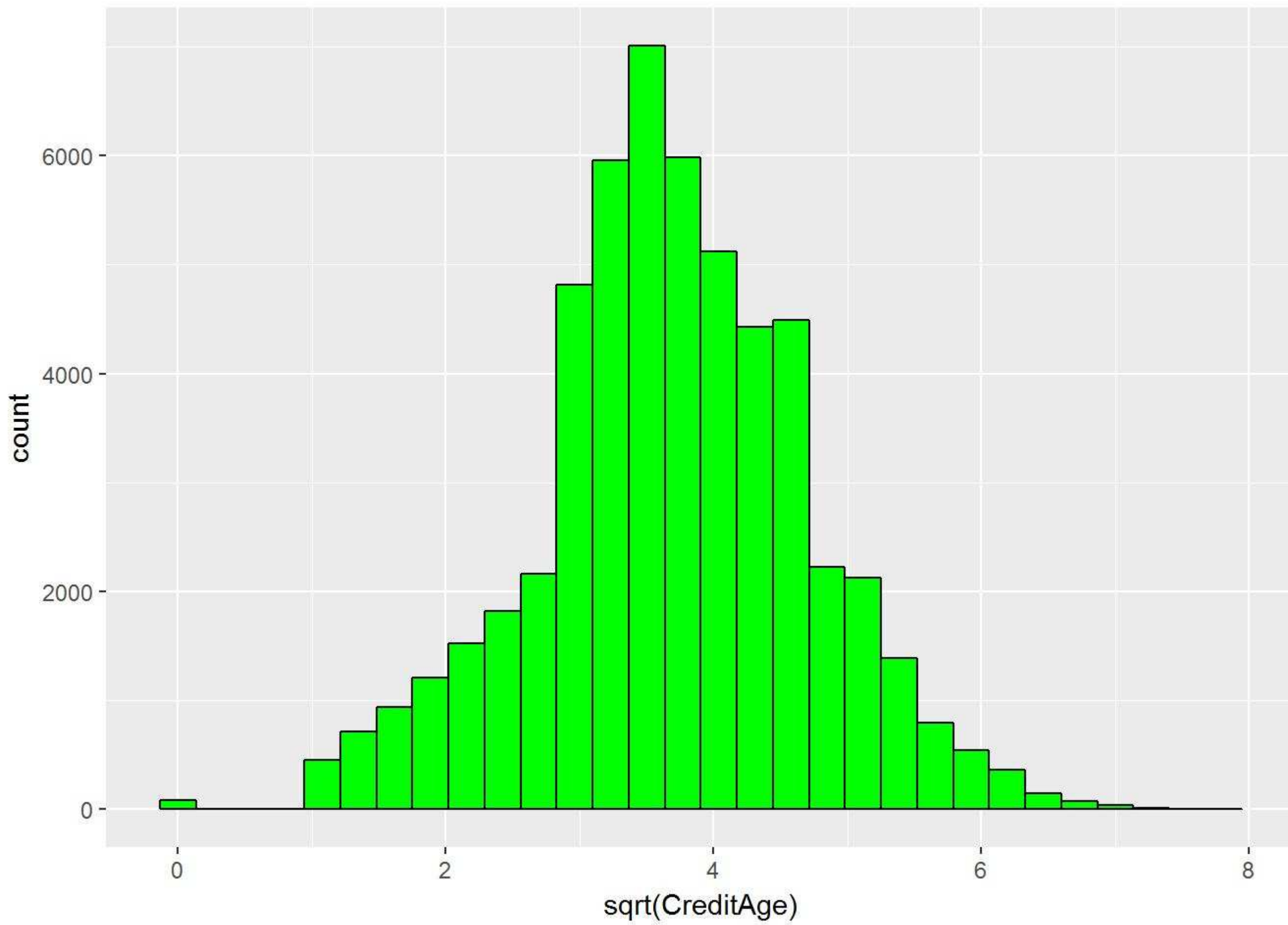


0.12
0.09
0.06
0.03
0.00
..count../sum(..count..)
500
600
700
800
900
CreditScore
colour
red

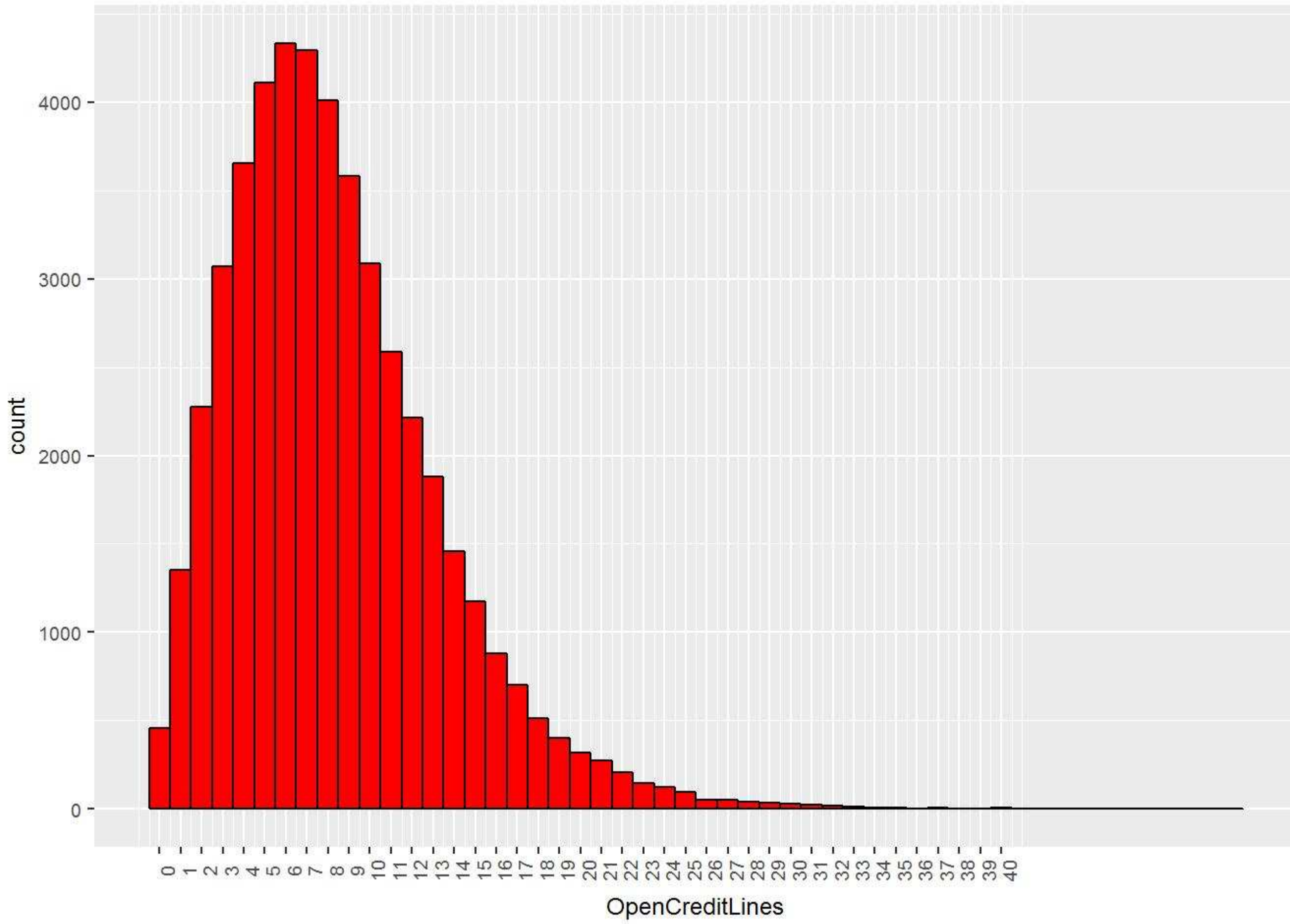



6000
4000
2000
0
count
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54 56 58 60
CreditAge

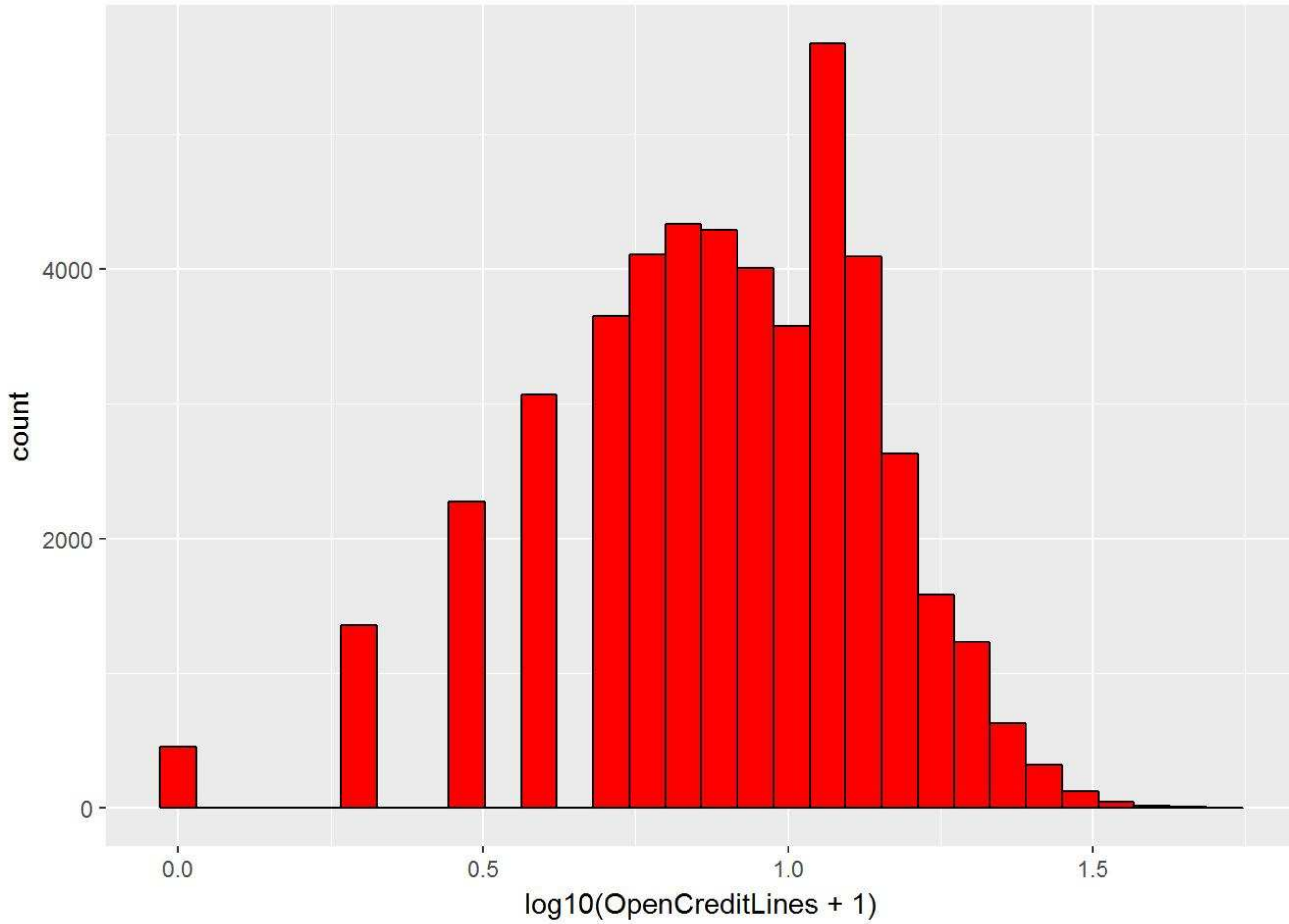
The CreditAge is a little right skewed. So I will look at Square root of the data as well.



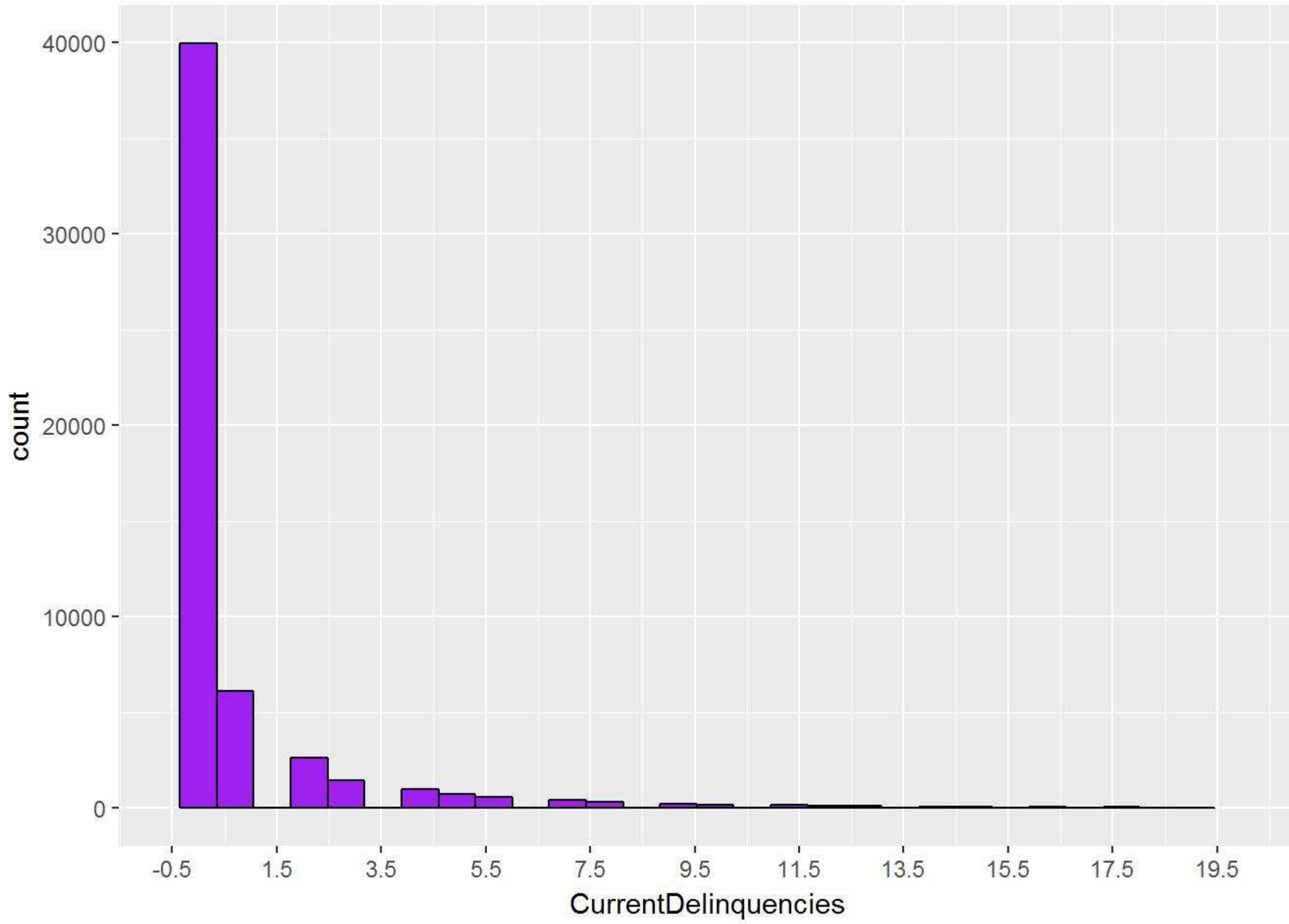

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00    9.00   14.00   14.68   19.00   61.00     696
```




The OpenCreditLines is a little skewed. I tried transforming it with log10 bot the result is not very useful.

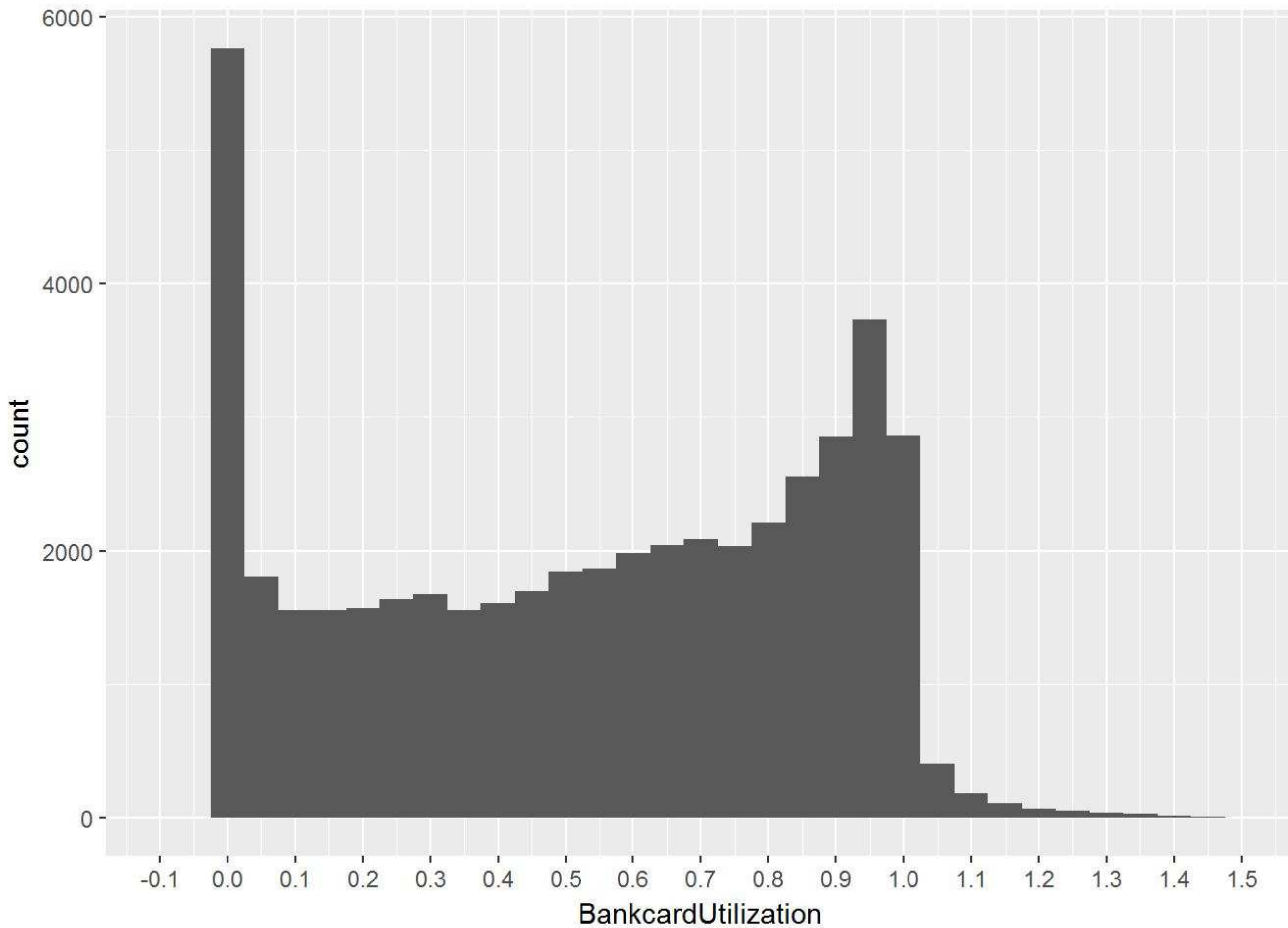



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##   0.000   5.000   8.000   8.338  11.000  51.000    7600
```



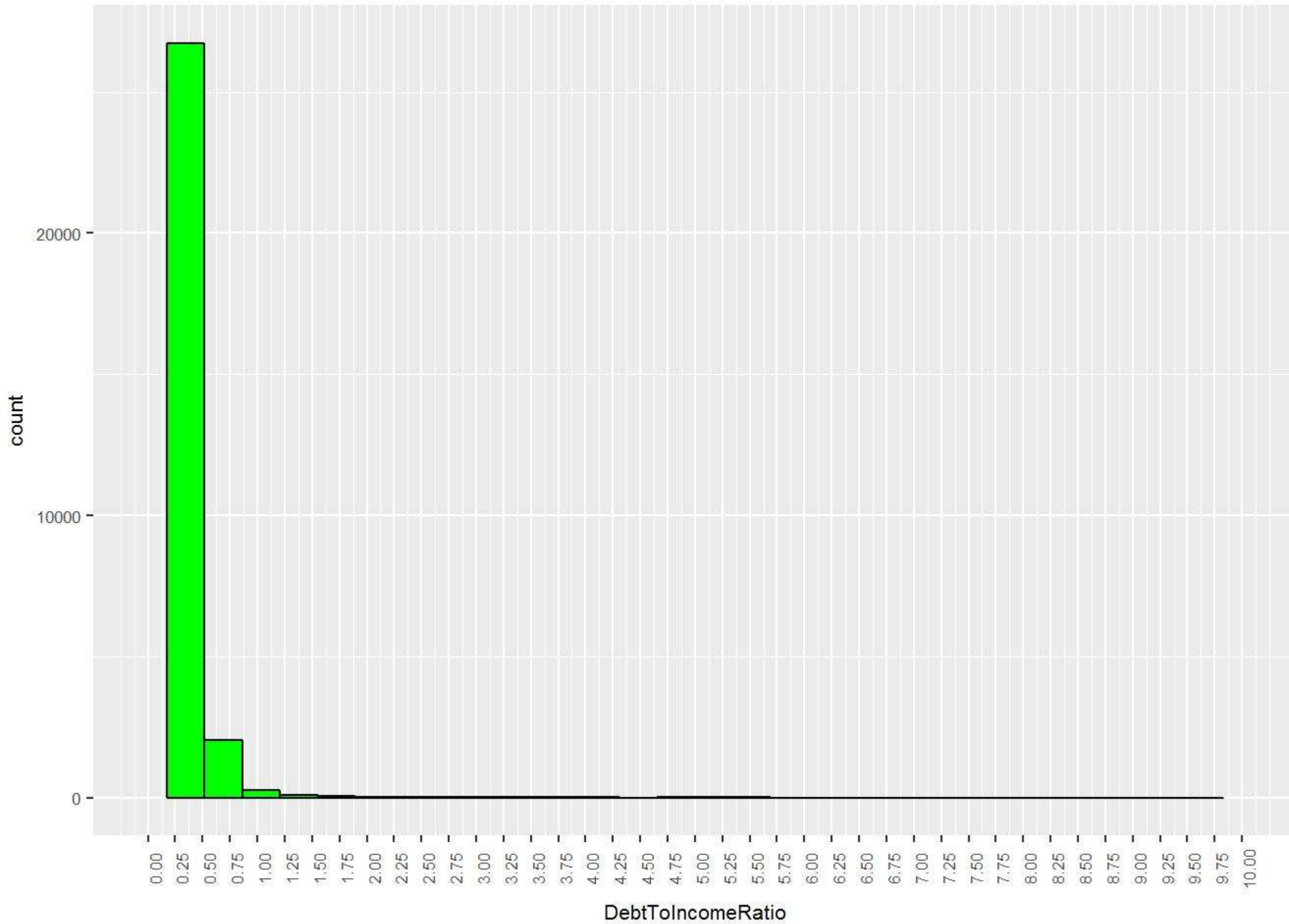

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.0000  0.0000  0.9062  1.0000 83.0000     696
```

Most of the users has no Current Delinquency.



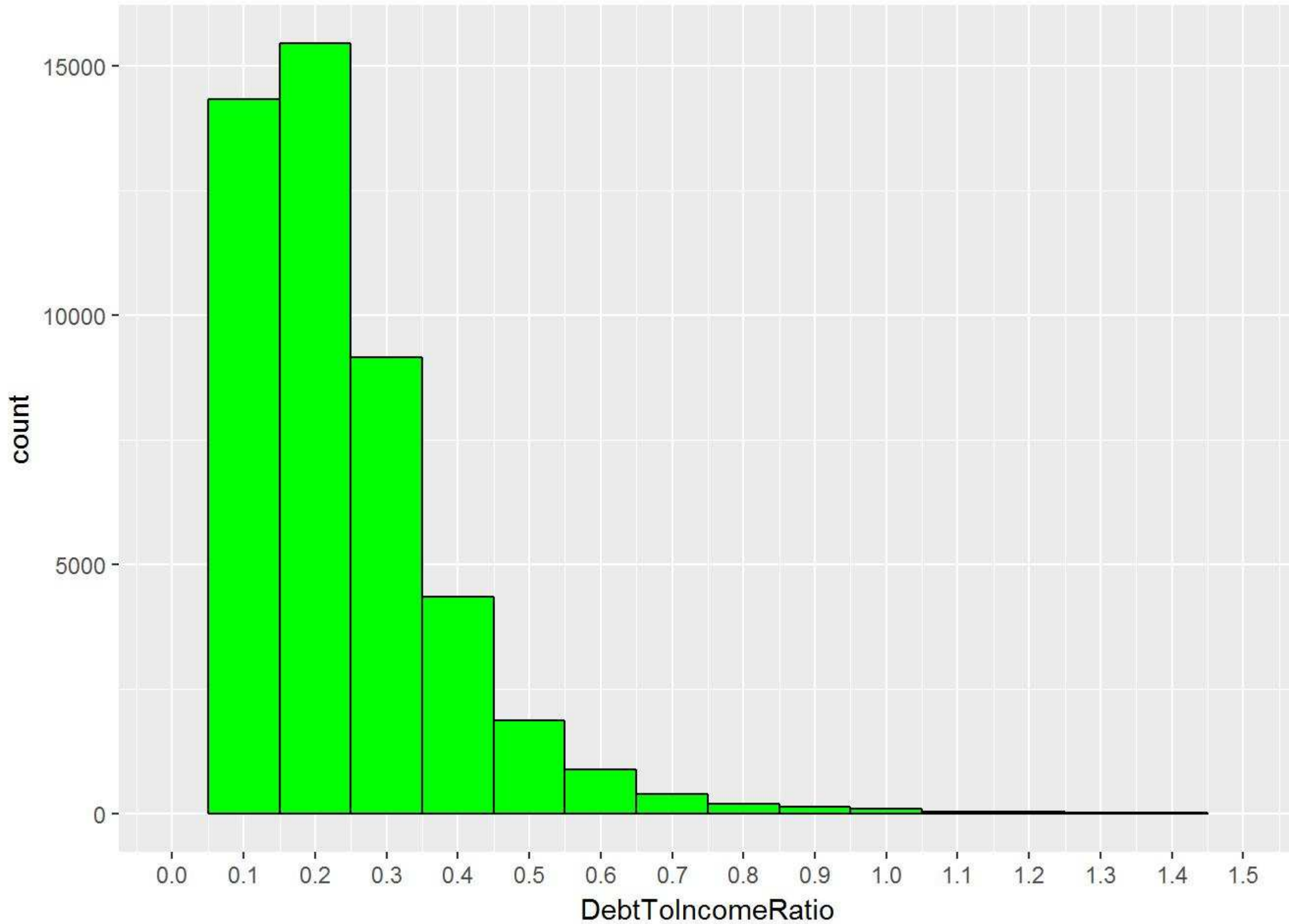

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00    0.21    0.56    0.53    0.85    5.95    7600
```

BankcardUtilization has a bimodal distribution. Most users utilize 0 percent of their credit cards and next the majority of users utilize 95% of their bank cards.

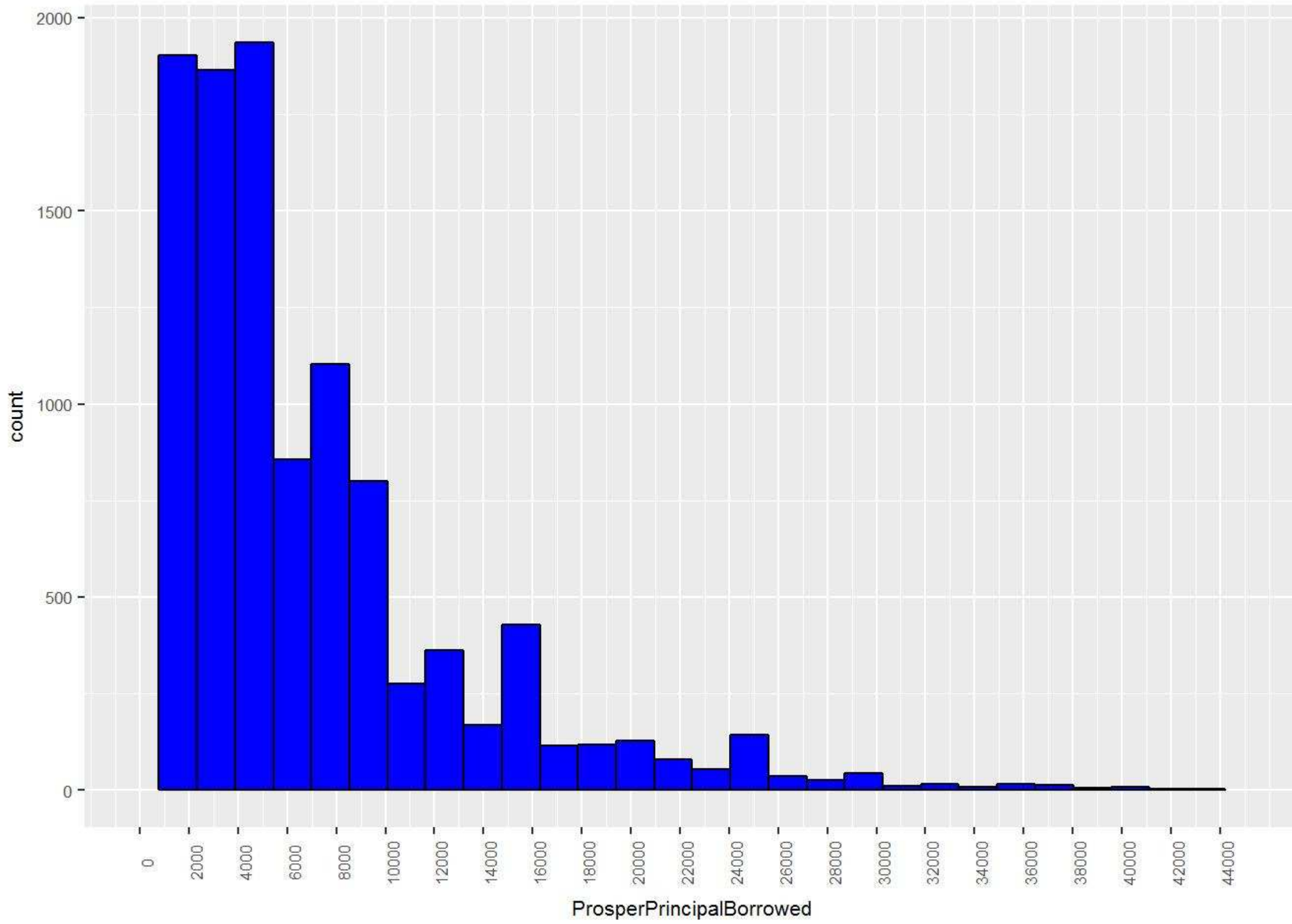



```
## [1] 289
```

I think the DebtToIncomeRatio greater than 10 is a problem in dataset. Because there are totally 270 points with this ratio greater 6 and 247 of them are greater than 10; There are a few outliers in DebtToIncomeRatio, so I will change the scales to avoiding them;



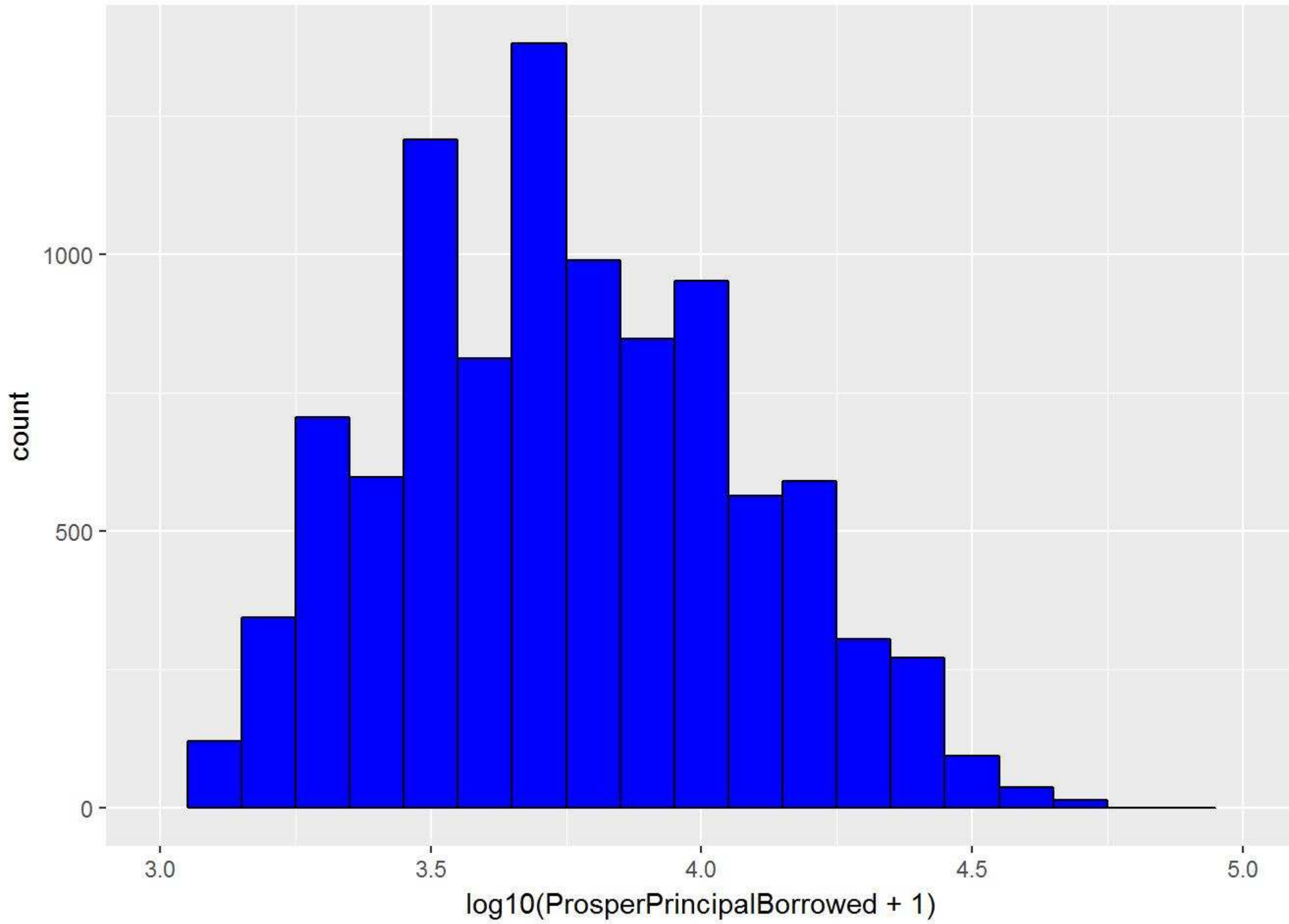

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00    0.13    0.20    0.29    0.30   10.01    4230
```



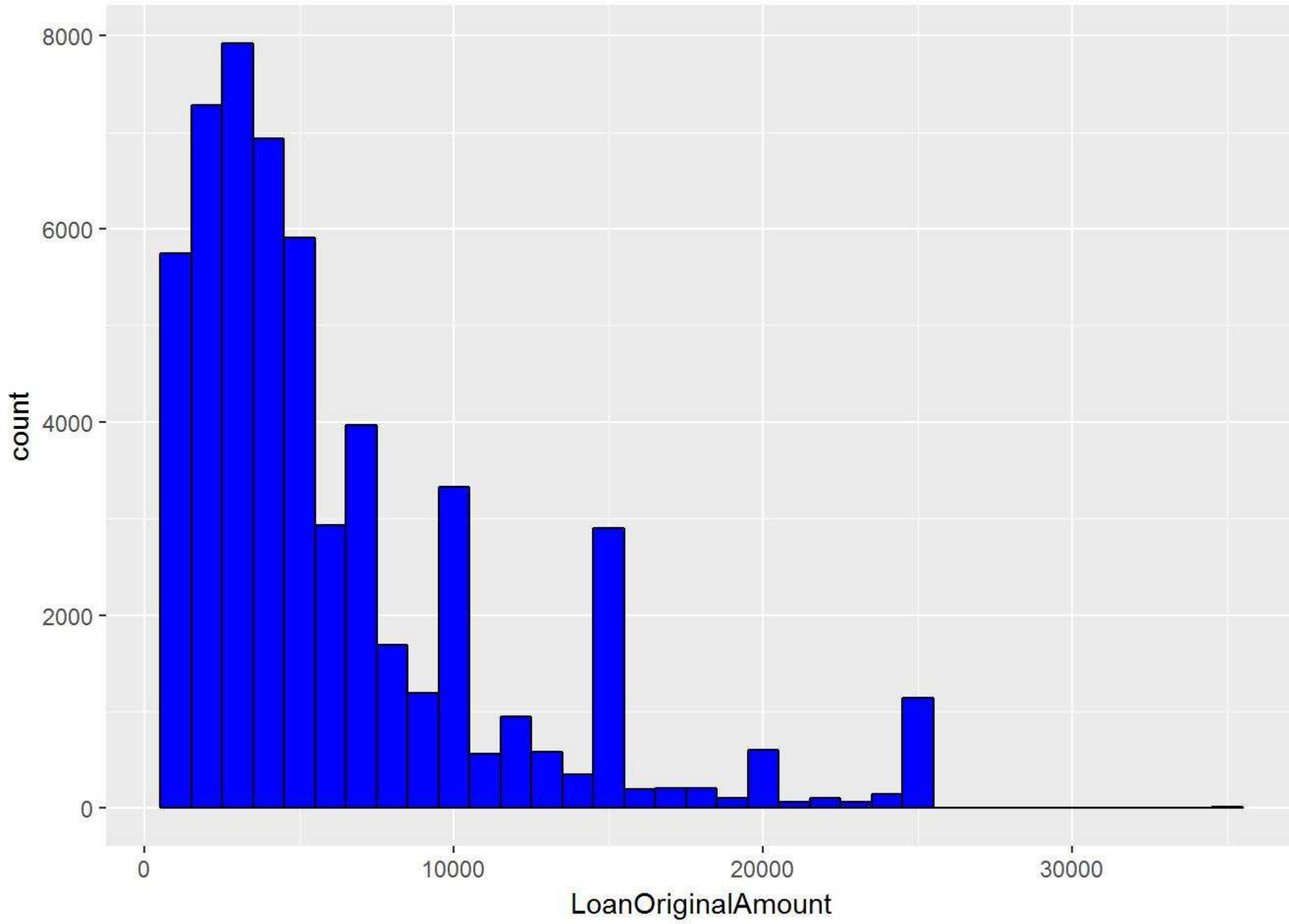

Most of the loans are under $6000. The distribution is right skewed;

```
loan_under_6000 <- subset(loan, ProsperPrincipalBorrowed <= 6000)
nrow(loan_under_6000)
```

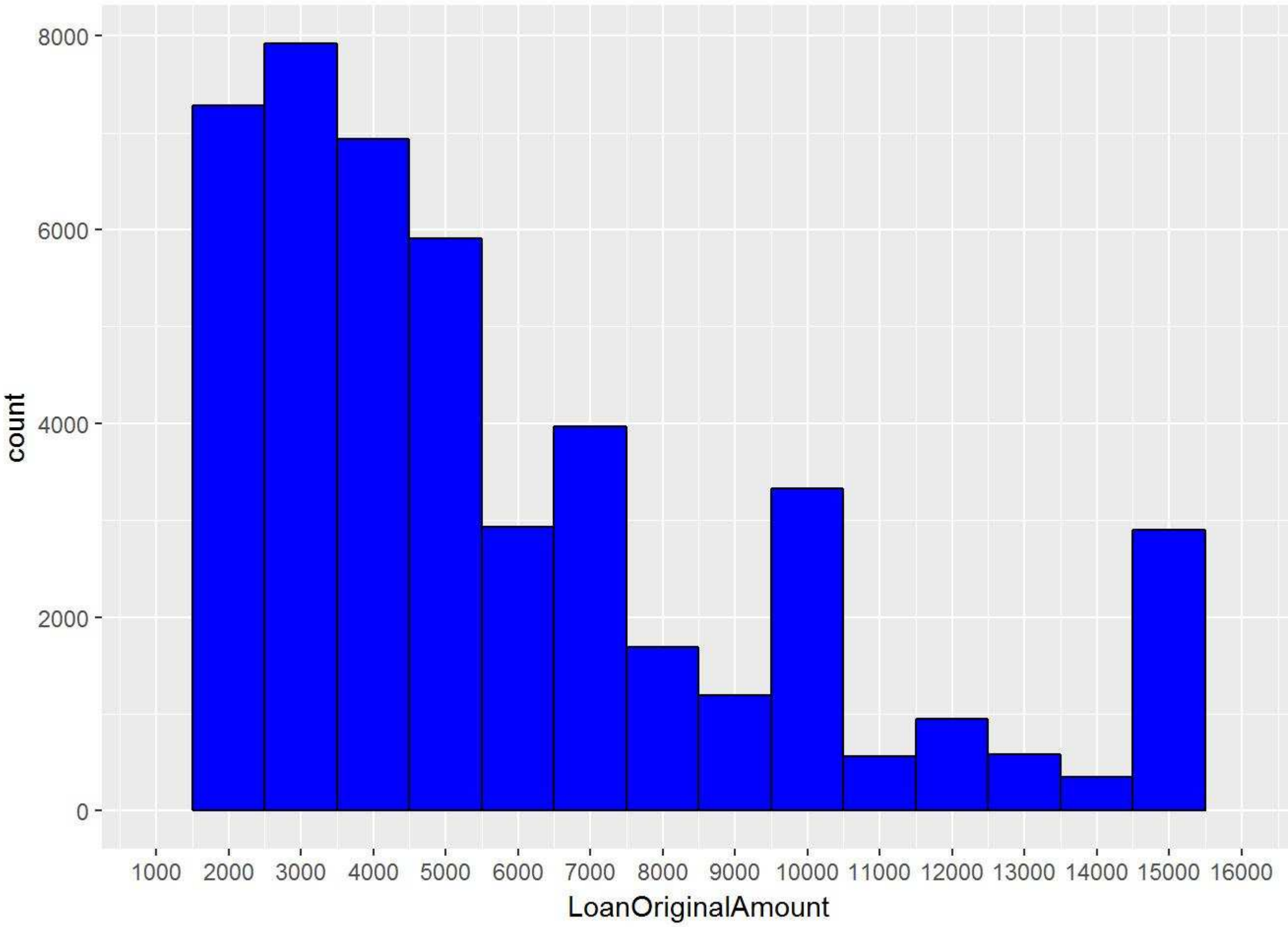
```
## [1] 6235
```




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##       0    3000    5000    7105    9500   60000   44545
```

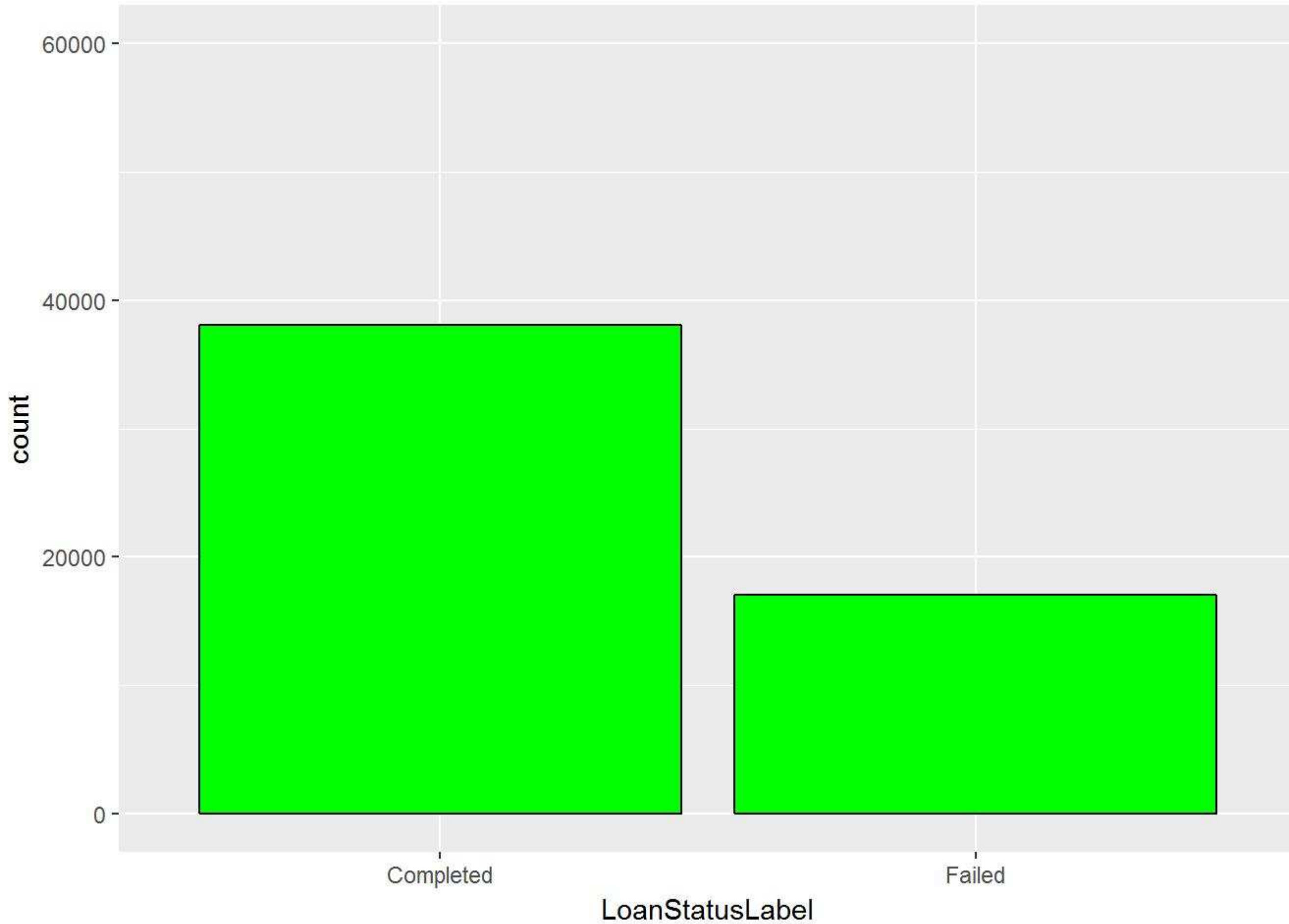



There are some ouliers in LoanOriginalAmount, so I changed the scales.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1000    2600    4500    6262    8000   35000
```

The most popular LoanOriginalAmounts are between $2500 and $3500.




```
##
## Completed    Failed
##     38074     17010
```

# Univariate Analysis

## What is the structure of your dataset?

The original data set contains 113,937 observations with 81 variables. I choose some of those variablels and focusing only on Completed or failed loans for doing EDA; It means that I am working on 55084 observations with 16 variables.

```
## [1] 55084    16
```

```
## 'data.frame':    55084 obs. of  16 variables:
##  $ CreditGrade            : Factor w/ 8 levels "A","AA","B","C",..: 4 7 4 NA 2 5 NA NA NA NA
...
##  $ Term                   : Factor w/ 3 levels "12","36","60": 2 2 2 2 2 2 2 3 2 2 ...
##  $ BorrowerAPR            : num  0.165 0.283 0.15 0.358 0.132 ...
##  $ ProsperScore           : num  NA NA NA 5 NA NA 5 3 9 9 ...
##  $ BorrowerState          : Factor w/ 51 levels "AK","AL","AR",..: 6 11 NA 10 NA 23 15 35 6
 5 ...
##  $ EmploymentStatus       : Factor w/ 8 levels "Employed","Full-time",..: 8 3 2 5 3 2 1 1 2
 1 ...
##  $ IsBorrowerHomeowner    : Factor w/ 2 levels "False","True": 2 1 1 2 2 1 1 1 2 1 ...
##  $ CreditScore            : num  650 490 650 710 770 630 690 670 710 750 ...
##  $ CreditAge              : num  6 5 7 13 16 4 15 38 10 24 ...
##  $ OpenCreditLines        : int  4 NA 2 9 NA 4 7 6 16 4 ...
##  $ CurrentDelinquencies   : int  2 1 3 0 2 1 0 0 0 1 ...
##  $ BankcardUtilization    : num  0 NA 0.32 0.97 NA 0.08 0.84 0.3 0.09 0.13 ...
##  $ DebtToIncomeRatio      : num  0.17 0.06 0.27 0.49 0.12 0.09 0.39 0.11 0.26 0.11 ...
##  $ ProsperPrincipalBorrowed: num  NA NA NA NA NA NA NA NA NA NA ...
##  $ LoanOriginalAmount     : int  9425 3001 1000 4000 10000 3000 2000 4000 4000 10000 ...
##  $ LoanStatusLabel        : Factor w/ 2 levels "Completed","Failed": 1 1 1 2 2 1 2 1 1 1 ...
```

```
##   CreditGrade    Term        BorrowerAPR       ProsperScore
##  C      : 5648   12: 1532   Min.   :0.00653   Min.   : 1.000
##  D      : 5153   36:49856   1st Qu.:0.14974   1st Qu.: 5.000
##  B      : 4389   60: 3696   Median :0.21434   Median : 6.000
##  AA     : 3509              Mean   :0.22219   Mean   : 6.266
##  HR     : 3505              3rd Qu.:0.29510   3rd Qu.: 8.000
##  (Other): 6744              Max.   :0.51229   Max.   :11.000
##  NA's   :26136              NA's   :25        NA's   :29079
##  BorrowerState        EmploymentStatus IsBorrowerHomeowner  CreditScore
##  CA     : 7263   Full-time    :24957   False:29199         Min.   : 10.0
##  FL     : 3077   Employed     :16491   True :25885         1st Qu.:650.0
##  IL     : 3039   Not available: 5346                       Median :690.0
##  GA     : 2783   Self-employed: 2926                       Mean   :681.7
##  TX     : 2752   Part-time    : 1056                       3rd Qu.:730.0
##  (Other):30658   (Other)      : 2056                       Max.   :890.0
##  NA's   : 5512   NA's         : 2252                       NA's   :590
##    CreditAge     OpenCreditLines  CurrentDelinquencies BankcardUtilization
##  Min.   : 0.00   Min.   : 0.000   Min.   : 0.0000      Min.   :0.00
##  1st Qu.: 9.00   1st Qu.: 5.000   1st Qu.: 0.0000      1st Qu.:0.21
##  Median :14.00   Median : 8.000   Median : 0.0000      Median :0.56
##  Mean   :14.68   Mean   : 8.338   Mean   : 0.9062      Mean   :0.53
##  3rd Qu.:19.00   3rd Qu.:11.000   3rd Qu.: 1.0000      3rd Qu.:0.85
##  Max.   :61.00   Max.   :51.000   Max.   :83.0000      Max.   :5.95
##  NA's   :696     NA's   :7600     NA's   :696          NA's   :7600
##  DebtToIncomeRatio ProsperPrincipalBorrowed LoanOriginalAmount
##  Min.   : 0.00     Min.   :    0            Min.   : 1000
##  1st Qu.: 0.13     1st Qu.: 3000            1st Qu.: 2600
##  Median : 0.20     Median : 5000            Median : 4500
##  Mean   : 0.29     Mean   : 7105            Mean   : 6262
##  3rd Qu.: 0.30     3rd Qu.: 9500            3rd Qu.: 8000
##  Max.   :10.01     Max.   :60001            Max.   :35000
##  NA's   :4230      NA's   :44545
##   LoanStatusLabel
##  Completed:38074
##  Failed   :17010
##
##
##
##
##
```

# What is the main feature of interest in your dataset?

In this investigation I want to consider final result of a loan: LoanStatusLabel as the main feature of interest and I am looking at other features as predictors of this target.

# What other features in the dataset do you think will help support your

# investigation into your feature(s) of interest?

I choose these 15 features to help my investigation: CreditGrade, Term, ProsperScore, BorrowerAPR, BorrowerState, EmploymentStatus, IsBorrowerHomeowner, CreditScore, CreditAge, OpenCreditLine, CurrentDelinquencies, BankcardUtilization, DebtToIncomeRatio, ProsperPrincipalBorrowed, LoanOriginalAmount.

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

I created these variables: 1. CreditAge, base on FirstRecordedCreditLine and DateCreditPulled 2. CreditScore, base on CreditScoreRangeLower and CreditScoreRangeUpper I calculated the average of these two ranges to use one variable. 3. LoanStatusLabel, base on LoanStatus I create 2 categories: Completed, Failed.

## Of the features you investigated, were there any unusual distributions?

The EmploymentStatus is right skewed. The reason should be that, the prosper company only focus on Employed or Full_time or Self-employed applicants. Therefore the majority of it's clients are in these three categories and very few of them are in retired or other kinds of employment types. BankcardUtilization has a bimodal distribution. One peak occurs at 0 utilization and the other is at 95% utilization.

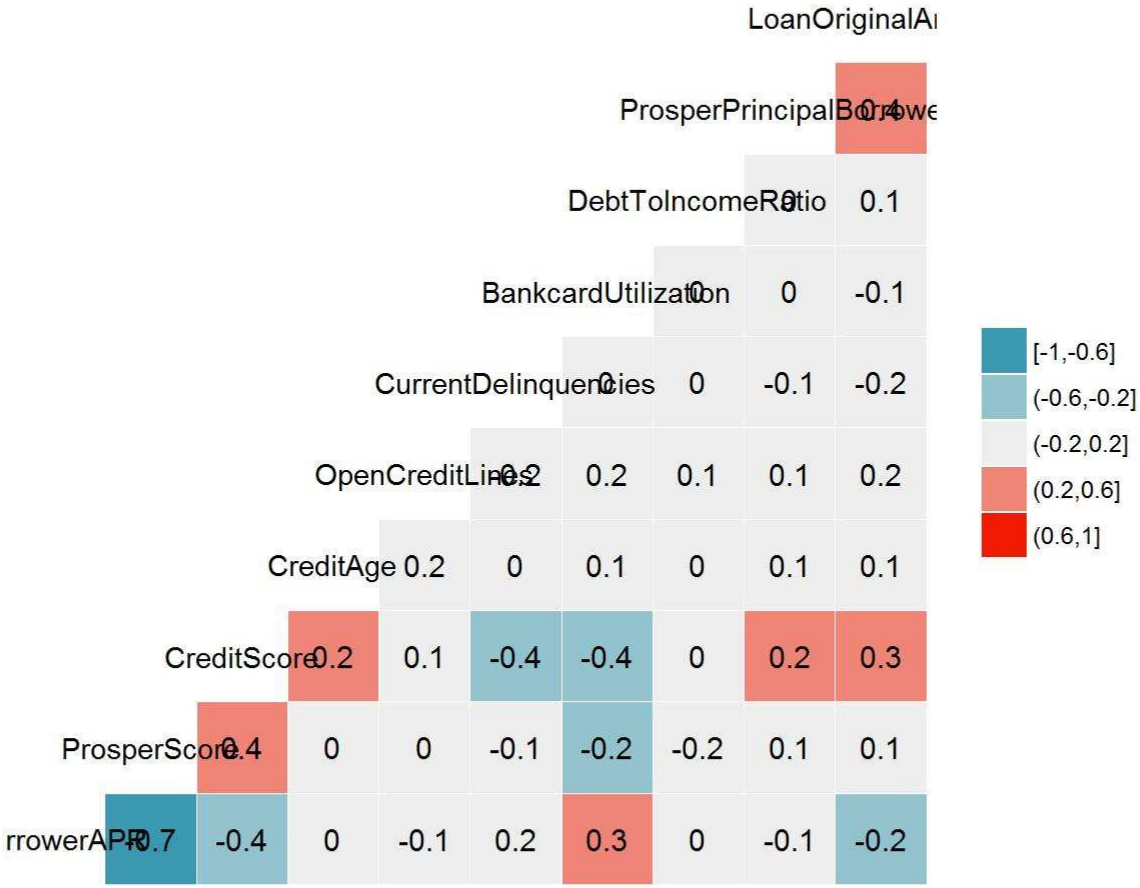
## Did you perform any operations on the data to tidy, adjust, or change the

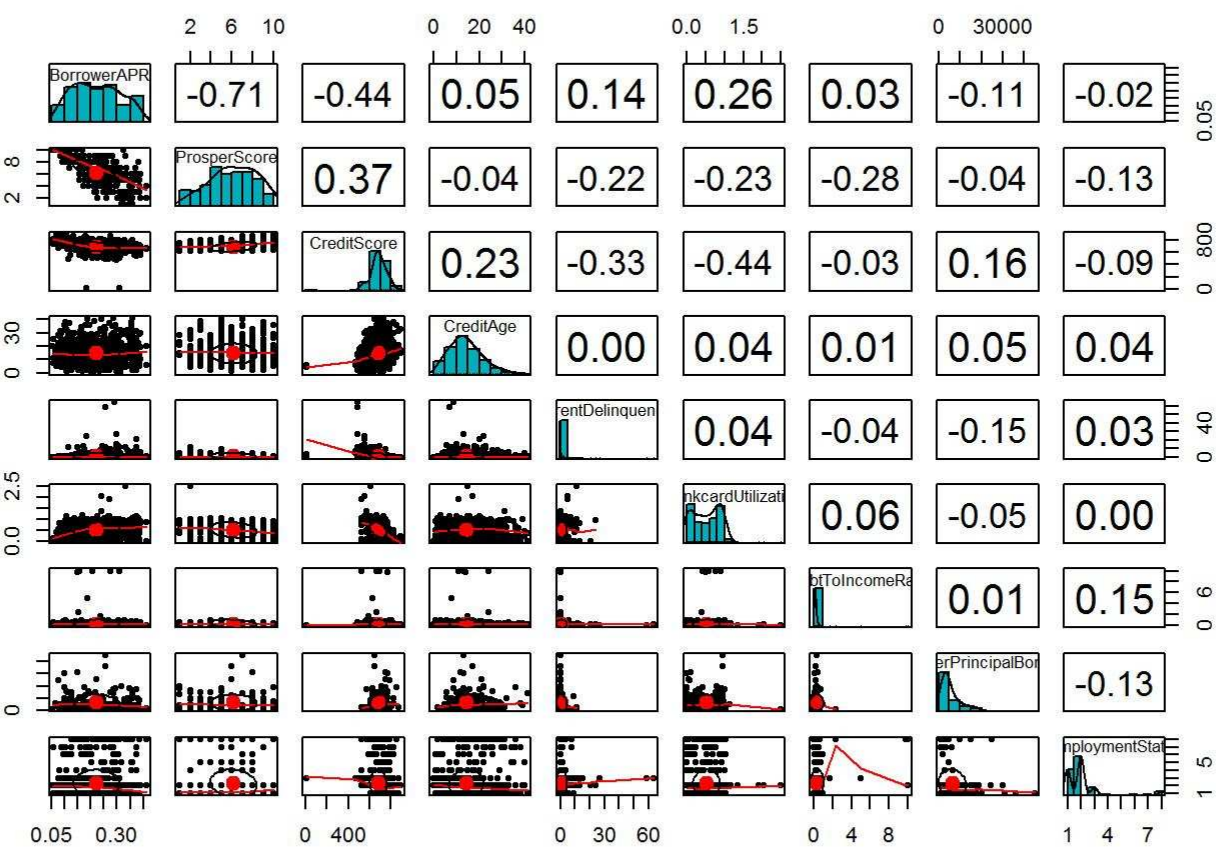
## form of the data? If so, why did you do this?

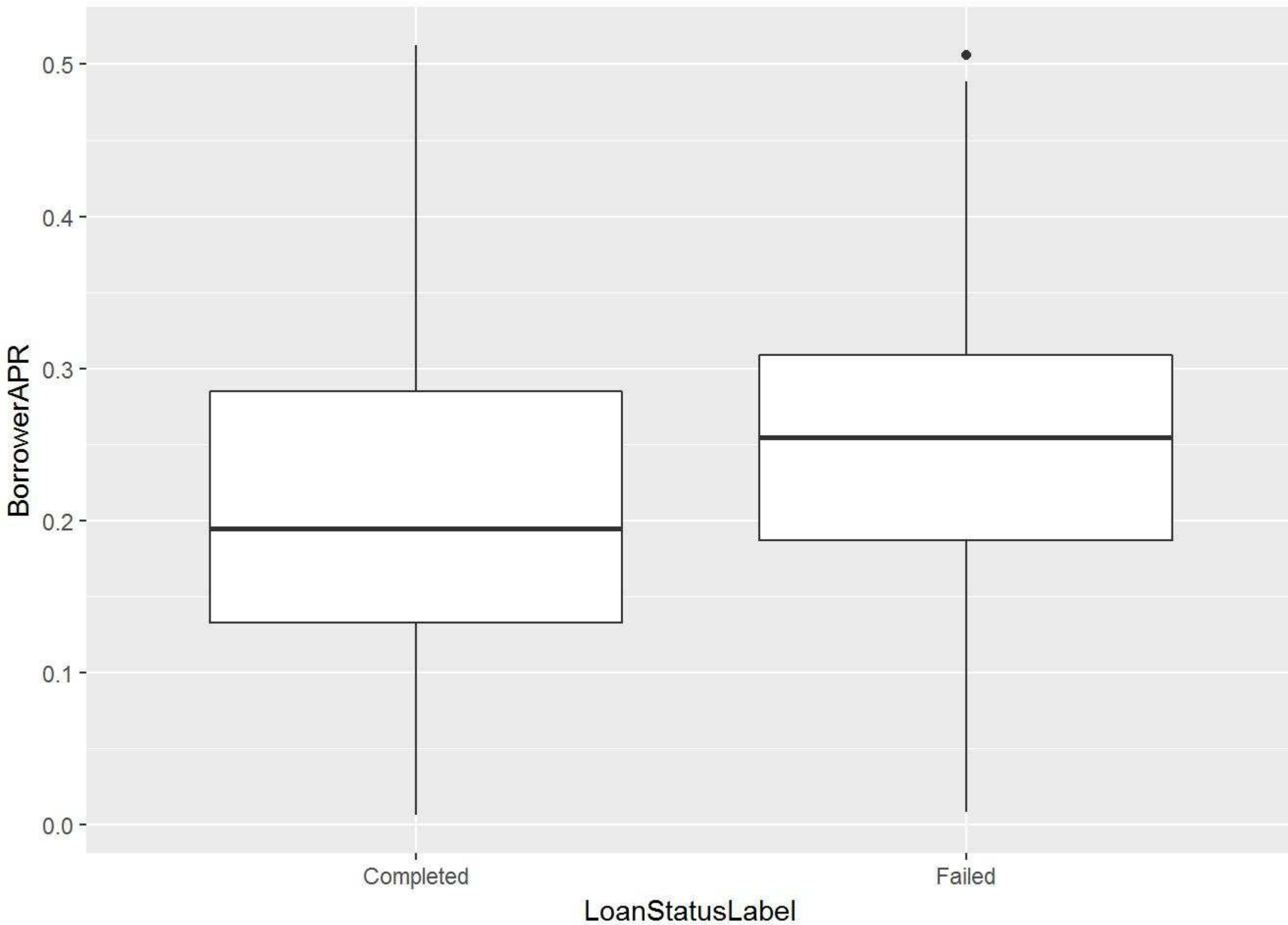
*I only look at two kinds of LoanStaus, because I am trying to find out what's the most effective features for predicting a Completed loan. So I cleaned all other LoanStatuses.* I changed the Term variable as factor. Because it has only 3 fixed values: 12, 36, 60. *For all graphs, I put NA's aside for looking at values only.* Another issue was distribution type for some of the features, like BorrowerAPR. It does not have a normal distribution. So I tried log10 and sqrt of this variable, and I think sqrt of values is smoother and more similar to normal distribution.

# Bivariate Plots Section

## Correlations

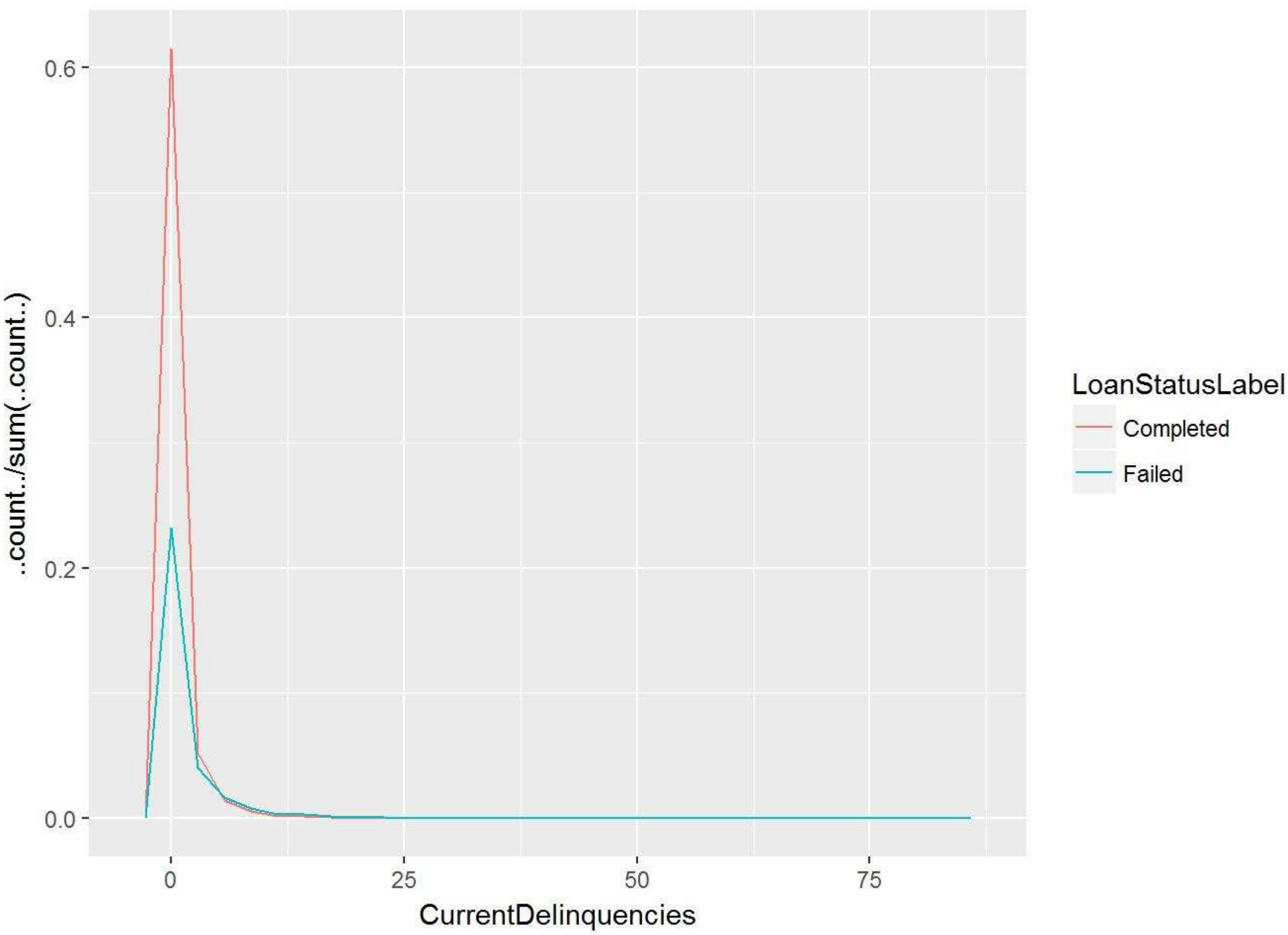

LoanOriginalAı
ProsperPrincipalBorrowe
DebtToIncomeRatio
BankcardUtilization
CurrentDelinquencies
OpenCreditLines
CreditAge
CreditScore
ProsperScore
rrowerAPR
[-1,-0.6]
(-0.6,-0.2]
(-0.2,0.2]
(0.2,0.6]
(0.6,1]


BorrowerAPR
ProsperScore
CreditScore
CreditAge
rentDelinquen
nkcardUtilizati
btToIncomeRa
erPrincipalBor
nploymentStat
-0.71
-0.44
0.05
0.14
0.26
0.03
-0.11
-0.02
0.37
-0.04
-0.22
-0.23
-0.28
-0.04
-0.13
0.23
-0.33
-0.44
-0.03
0.16
-0.09
0.00
0.04
0.01
0.05
0.04
0.04
-0.04
-0.15
0.03
0.06
-0.05
0.00
0.01
0.15
-0.13

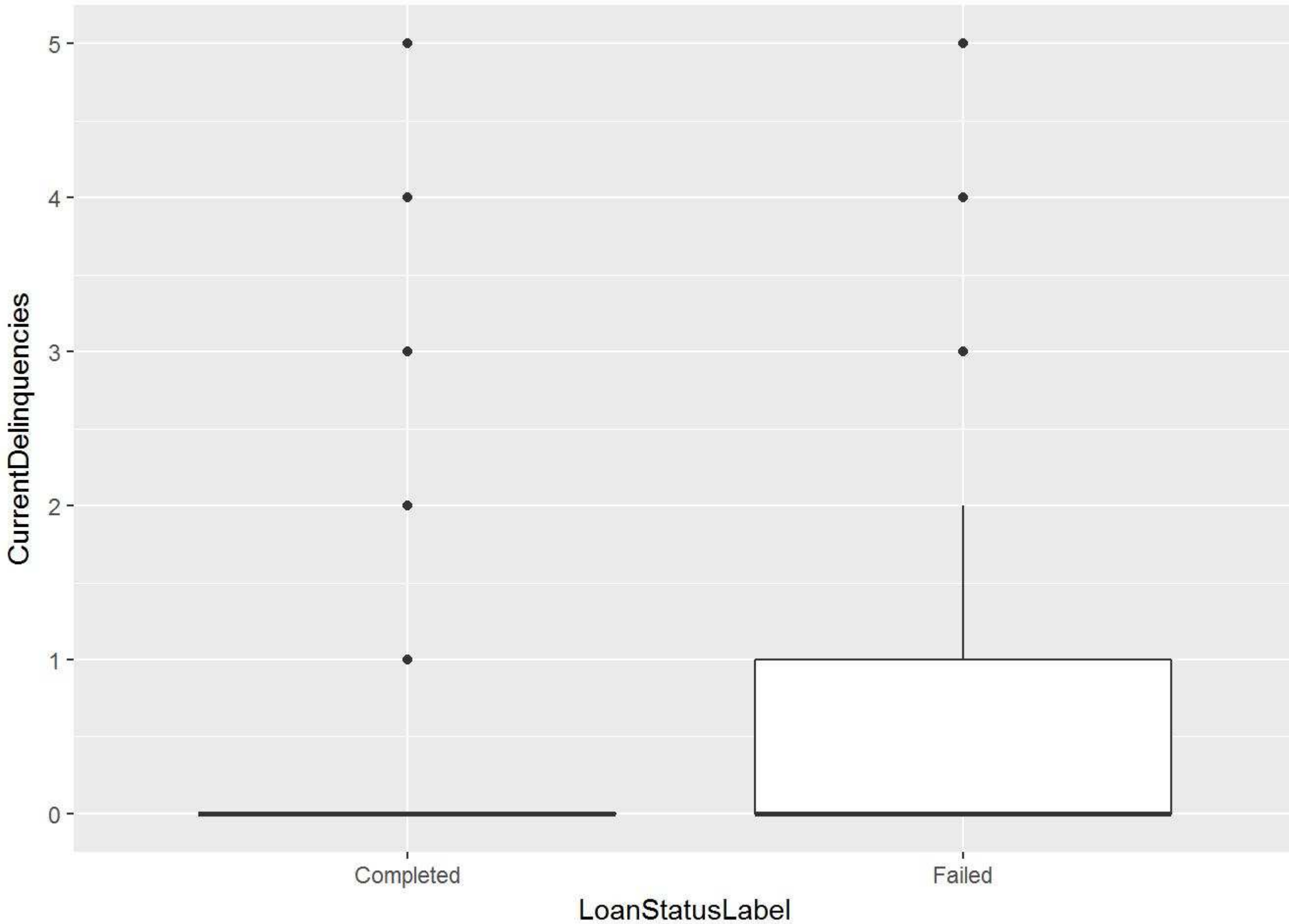



```
## loan$LoanStatusLabel: Completed
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
## 0.00653 0.13270 0.19480 0.20880 0.28500 0.51230      25 
## -------------------------------------------------------- 
## loan$LoanStatusLabel: Failed
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## 0.00864 0.18690 0.25440 0.25220 0.30910 0.50630
```

The APR mean for failed loans is 4.5% more than completed loans.

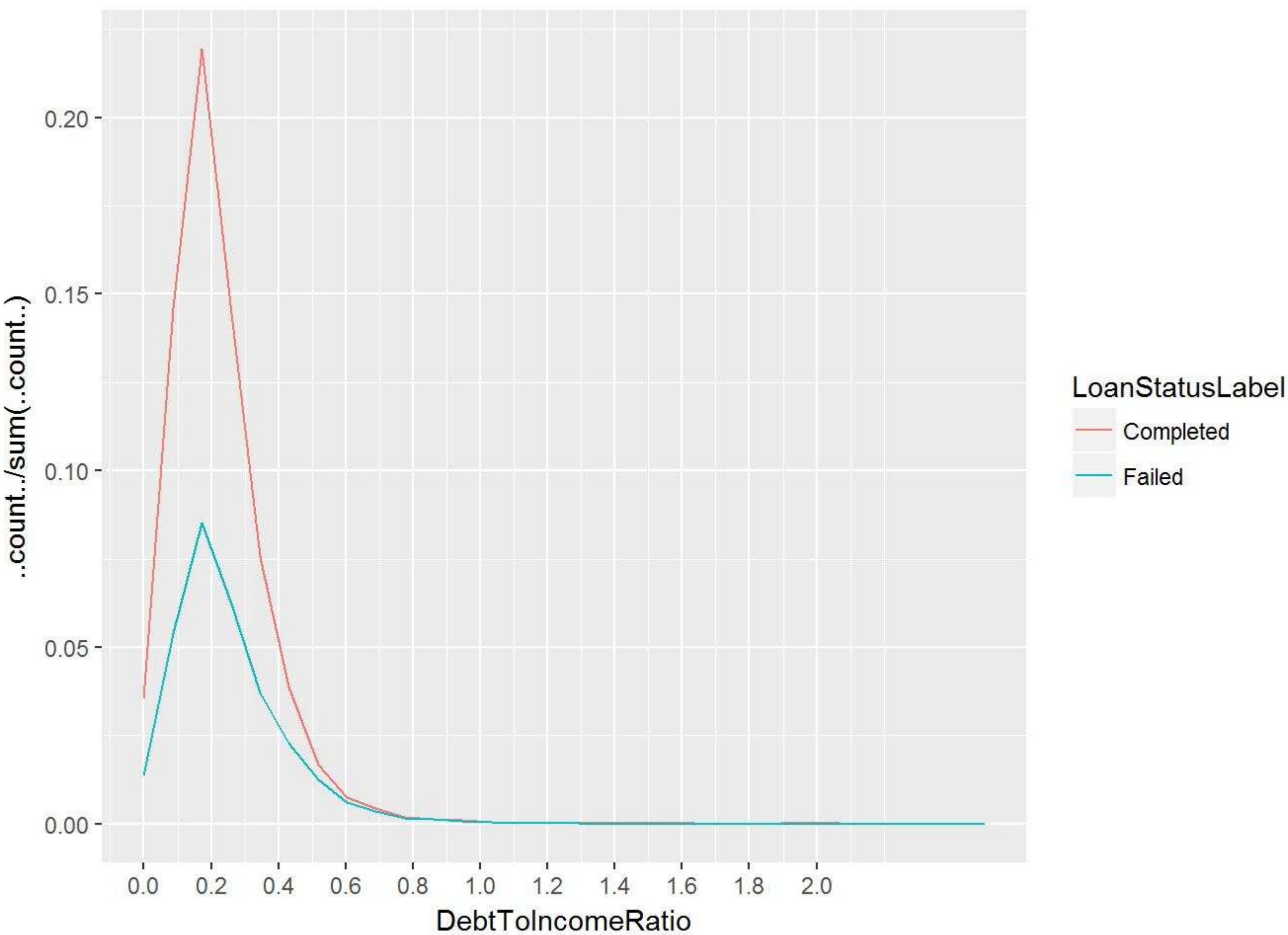



Around 60% of completed loans has zero CurrentDelinquencies and this number is 20% for failed loans.

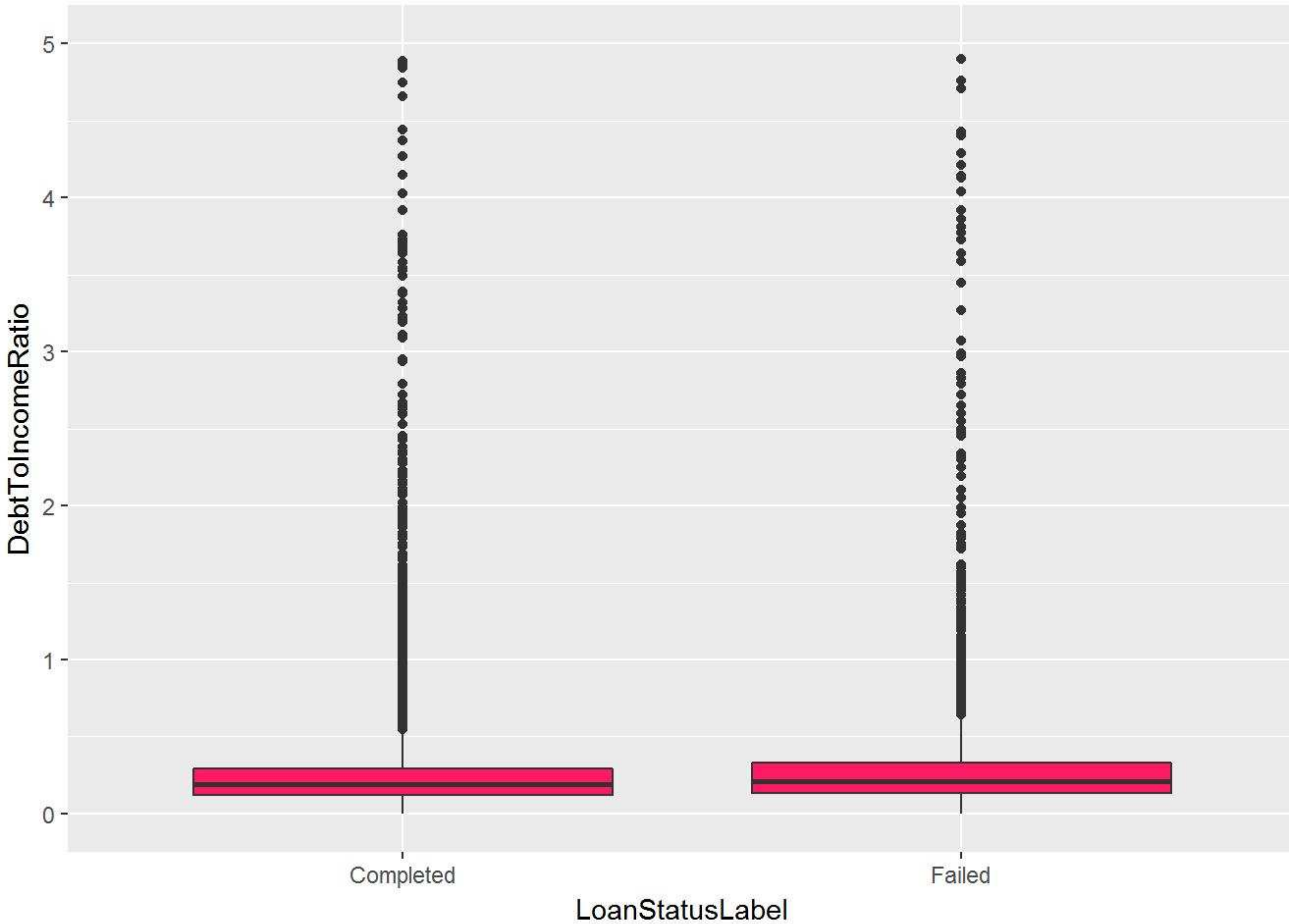



```
## loan$LoanStatusLabel: Completed
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##  0.0000  0.0000  0.0000  0.5958  0.0000 50.0000     463 
## -------------------------------------------------------- 
## loan$LoanStatusLabel: Failed
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##   0.000   0.000   0.000   1.602   1.000  83.000     233
```

Mean of CurrentDelinquencies for Colpleted loans is 0.6 and this number is 1.6 for failed loans. CurrentDelinquencies is a powerful predictor for failing a loan.

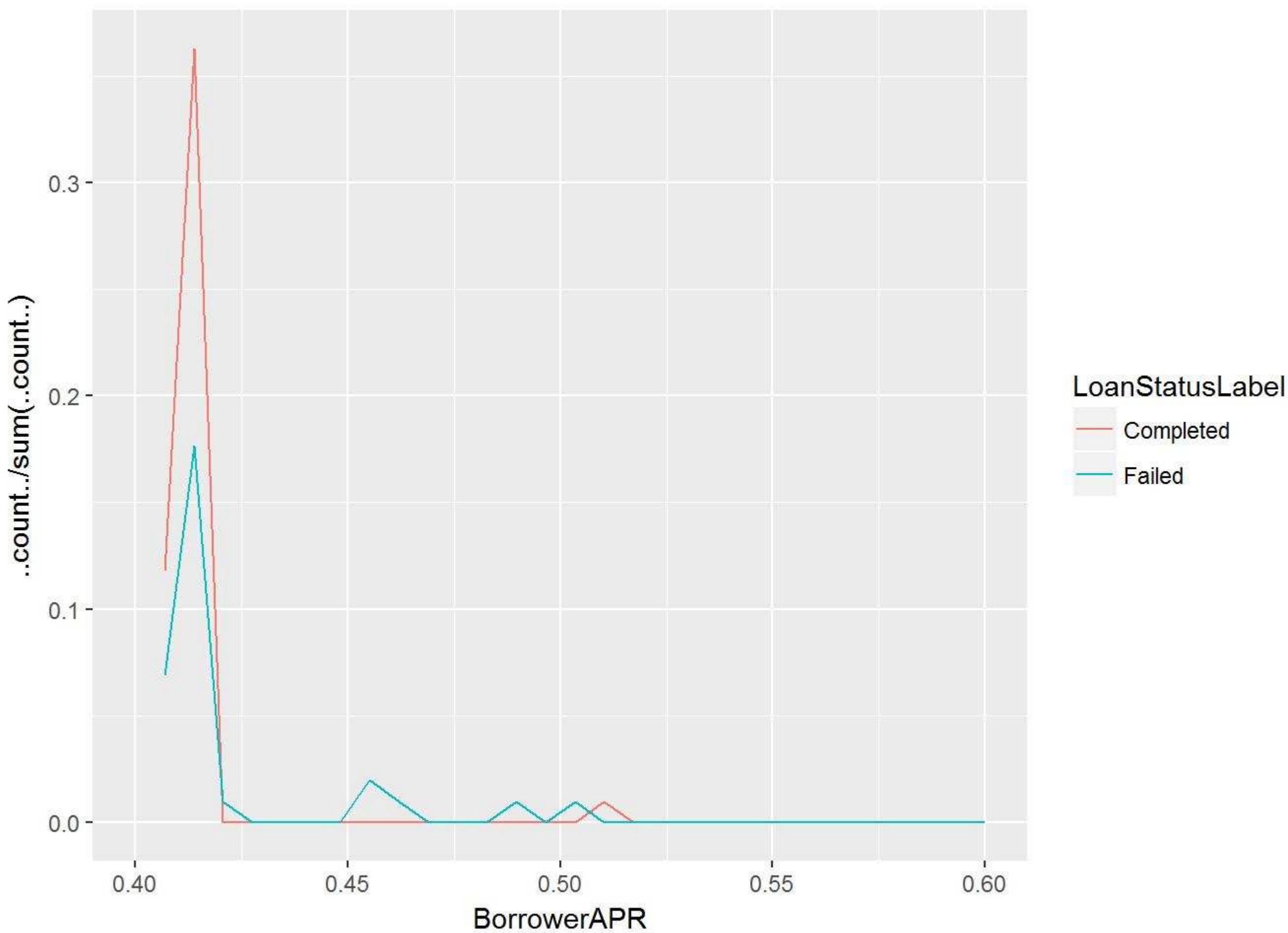



DebtToIncomeRatio is less than 20% for 24% of completed loans and this number is around 8% for failed loans.

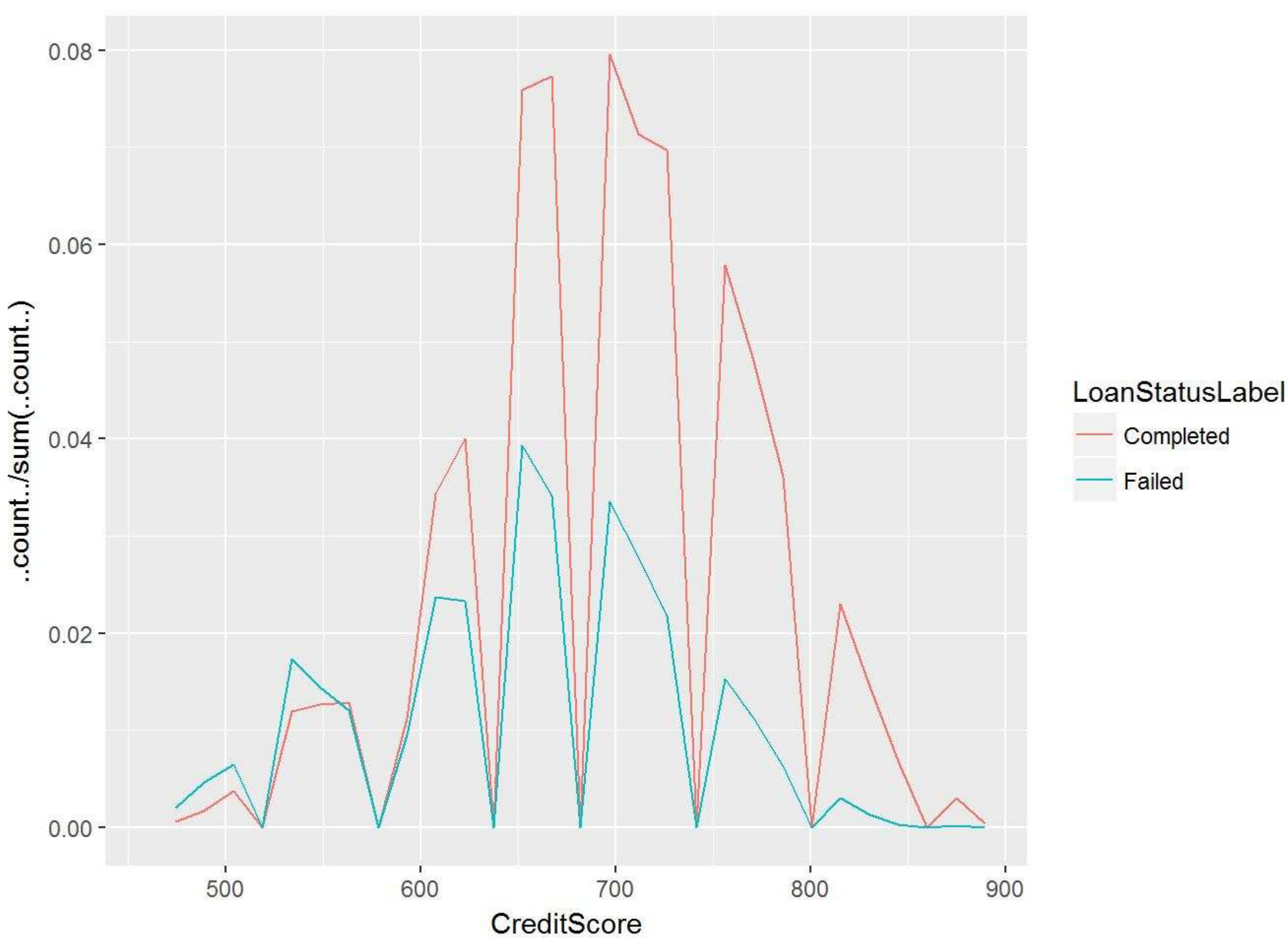



```
## loan$LoanStatusLabel: Completed
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.1200  0.1900  0.2642  0.2900 10.0100    2734
## --------------------------------------------------------
## loan$LoanStatusLabel: Failed
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.1393  0.2200  0.3484  0.3300 10.0100    1496
```

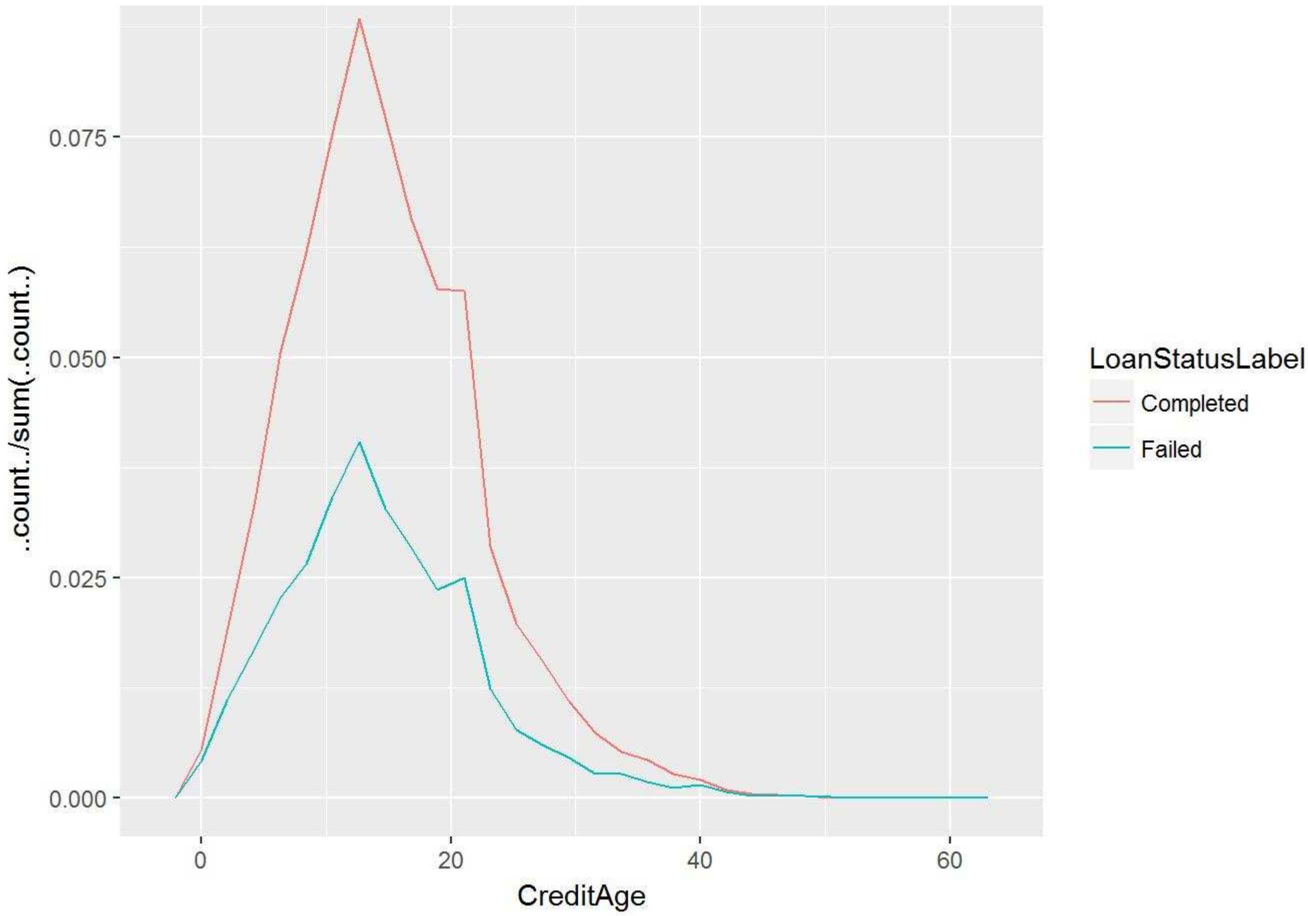
So DebtToIncomeRatio is another good predictor for loan status.



0.3
0.2
0.1
0.0
..count../sum(..count..)
0.40
0.45
0.50
0.55
0.60
BorrowerAPR
LoanStatusLabel
Completed
Failed

0.08
0.06
0.04
0.02
0.00
..count../sum(..count..)
500
600
700
800
900
CreditScore
LoanStatusLabel
Completed
Failed




```
##
## Completed    Failed
##         1         6
```

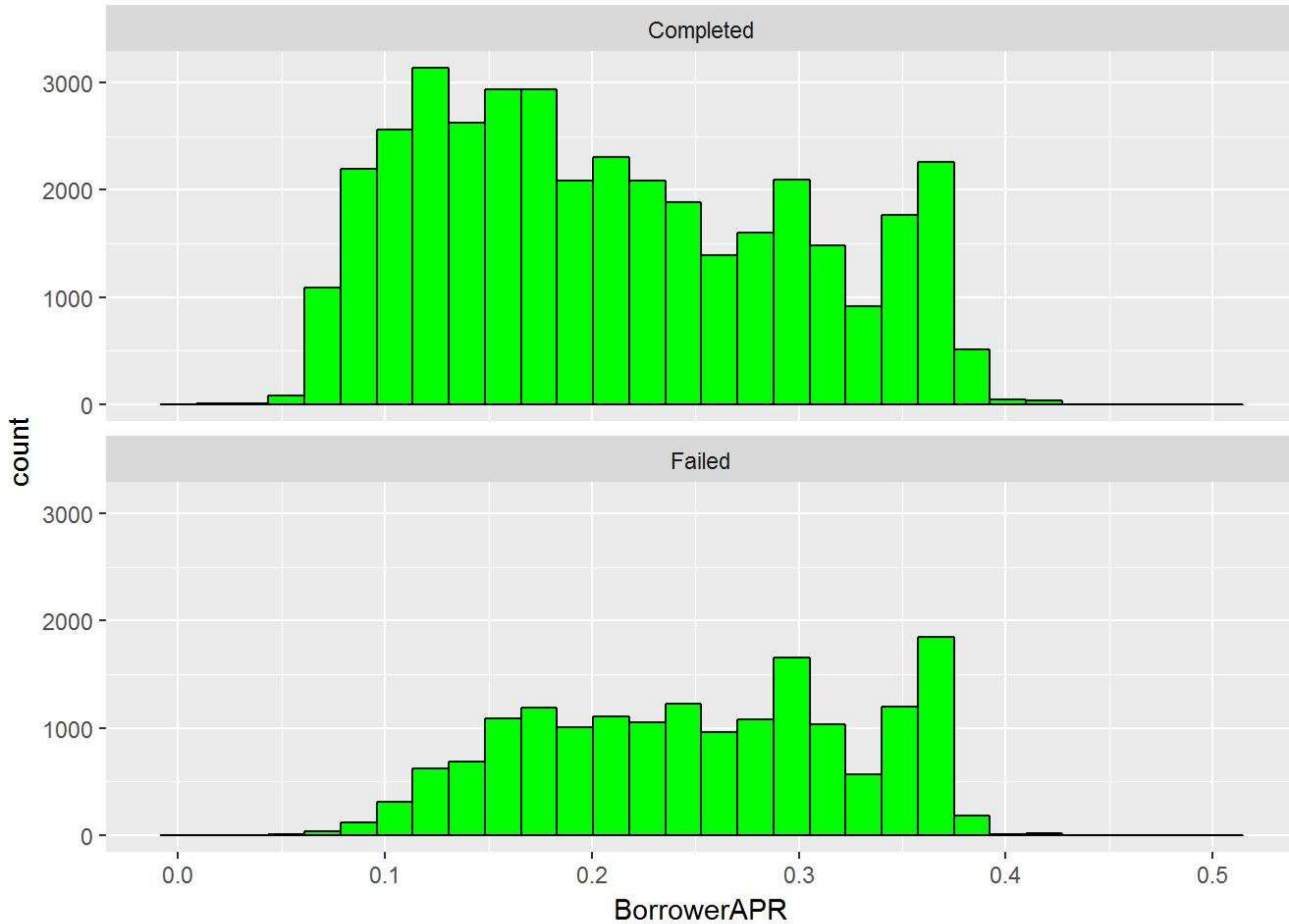
It seems like most of loans with APR more than 0.45 failed.

```
## # A tibble: 2 x 4
##   LoanStatusLabel CreditAge_mean CreditAge_median     n
##            <fctr>          <dbl>            <dbl> <int>
## 1       Completed       14.77661               14 38074
## 2          Failed       14.46605               13 17010
```

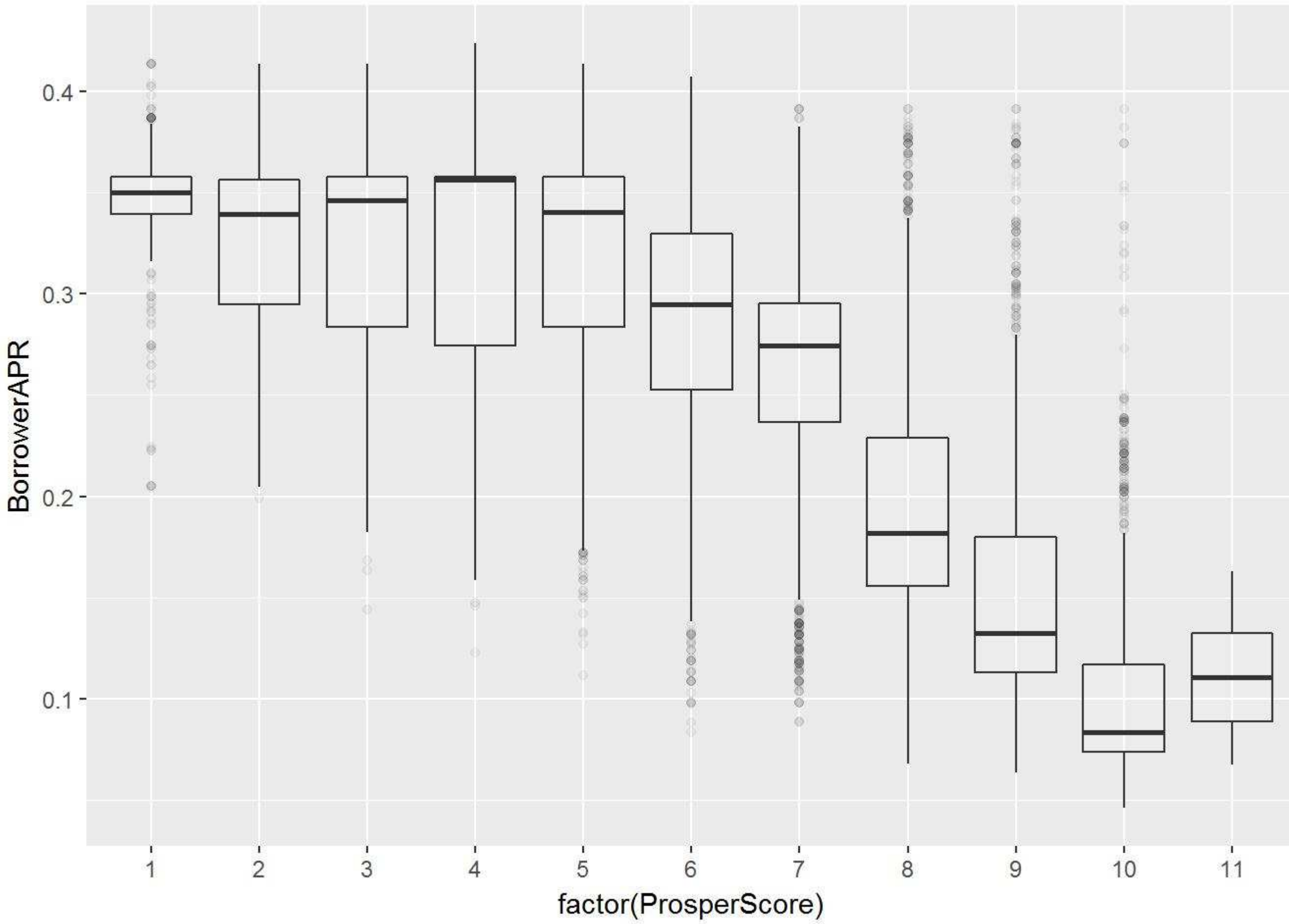
The difference between CreditAge mean and median for 2 group of LoanStatusLabel of interest (Completed and Failed) is not very significant.

```
## # A tibble: 2 x 4
##   LoanStatusLabel creditScore_mean CreditScore_median     n
##            <fctr>            <dbl>              <dbl> <int>
## 1       Completed         695.5871                690 38074
## 2          Failed         650.7876                650 17010
```

The median CreditScore for Completed and Failed LoanStatusLables has 40 points difference.

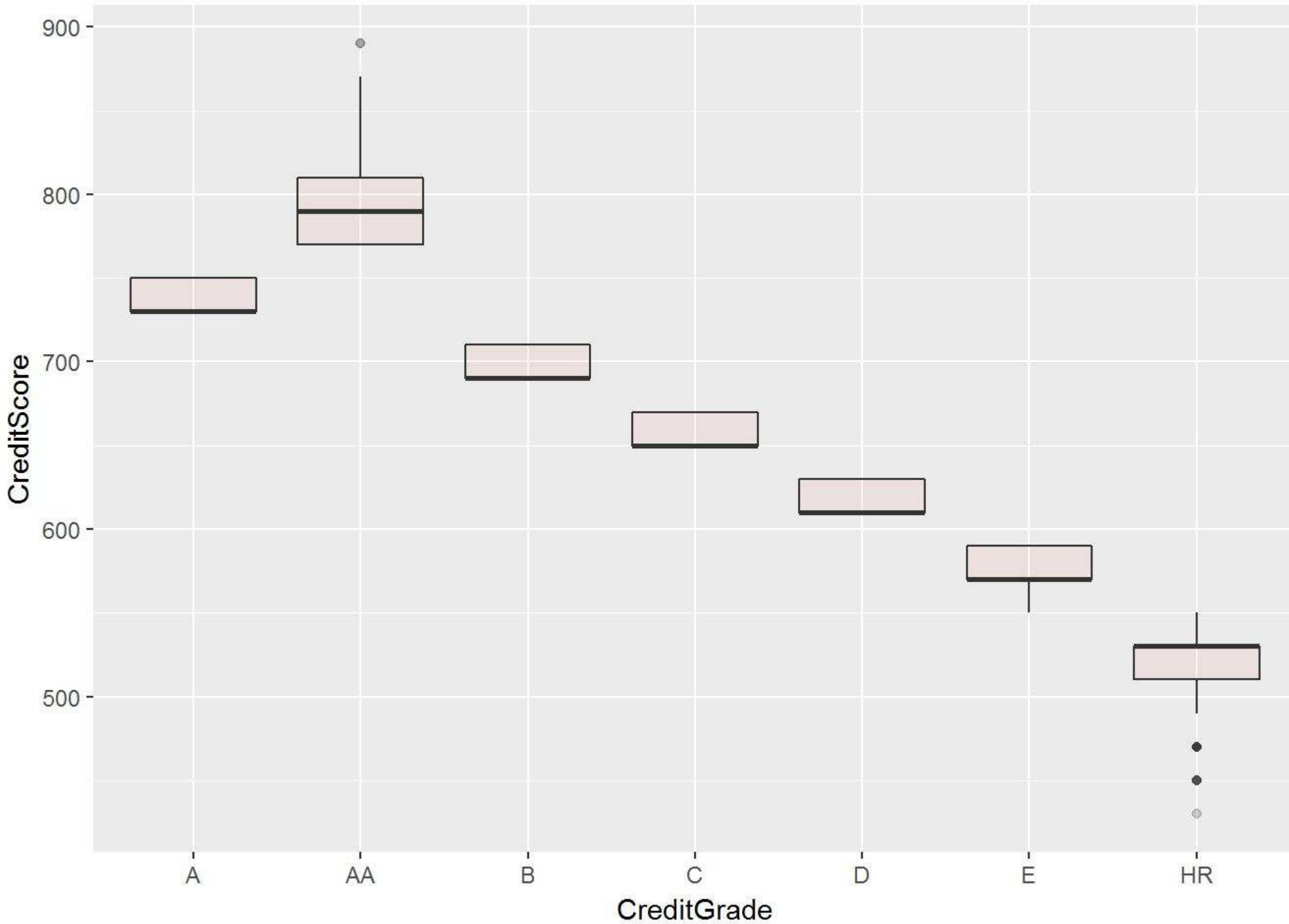



The distribution of different result for LoanStatusLabel in BorrowerAPR shows that Completed loans are a little right skewed and Faild loans are roughly uniformly distributed. It means most of the failed loans relativly had larger APRs.

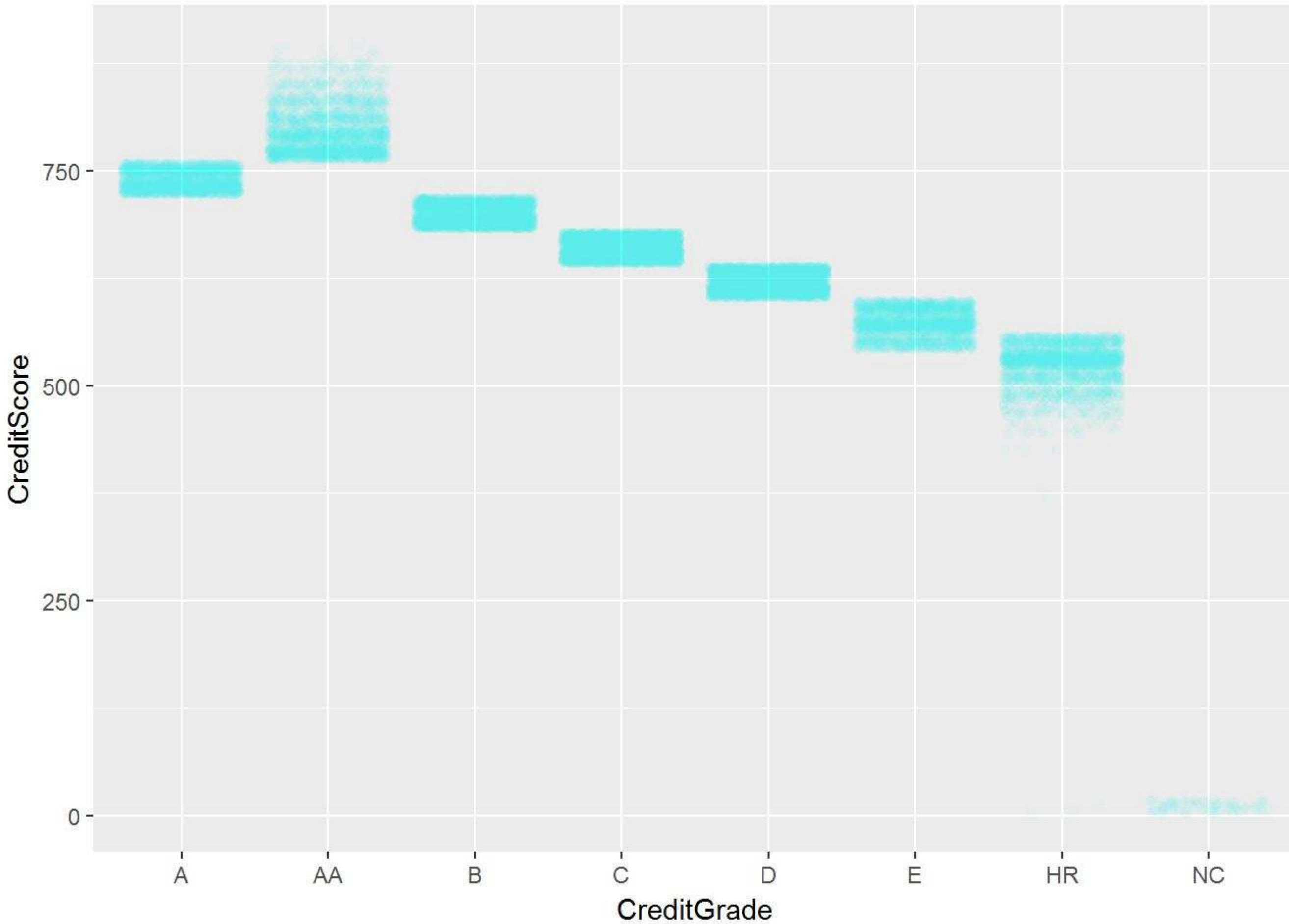



The ProsperScore and BorrowerAPR are negatively correlated with -0.74 Correlation Coefficient. That’s a strong correlation.

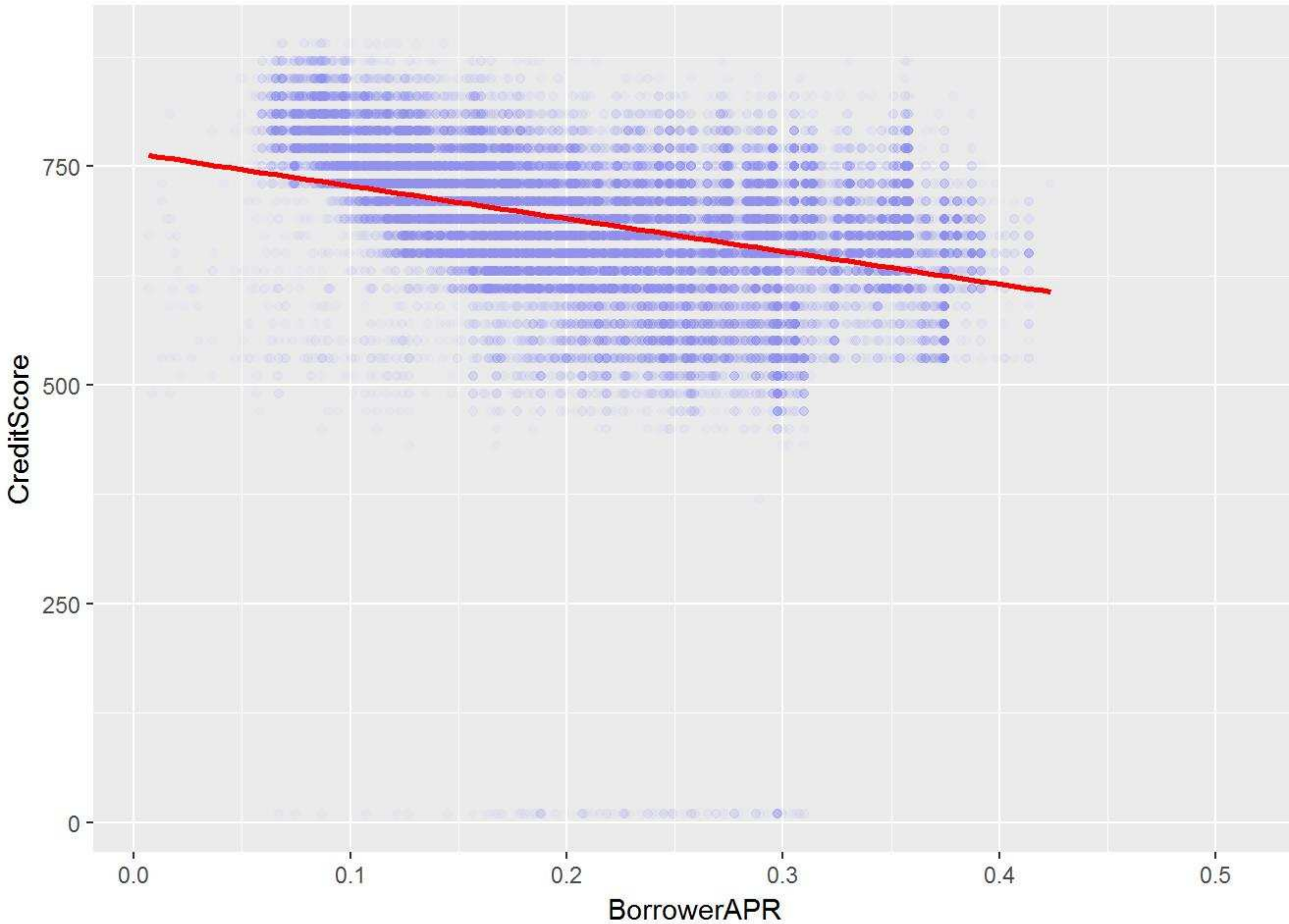
```
## [1] -0.7380842
```




There are many overlaps, so I am using Jitter to see more clearly.

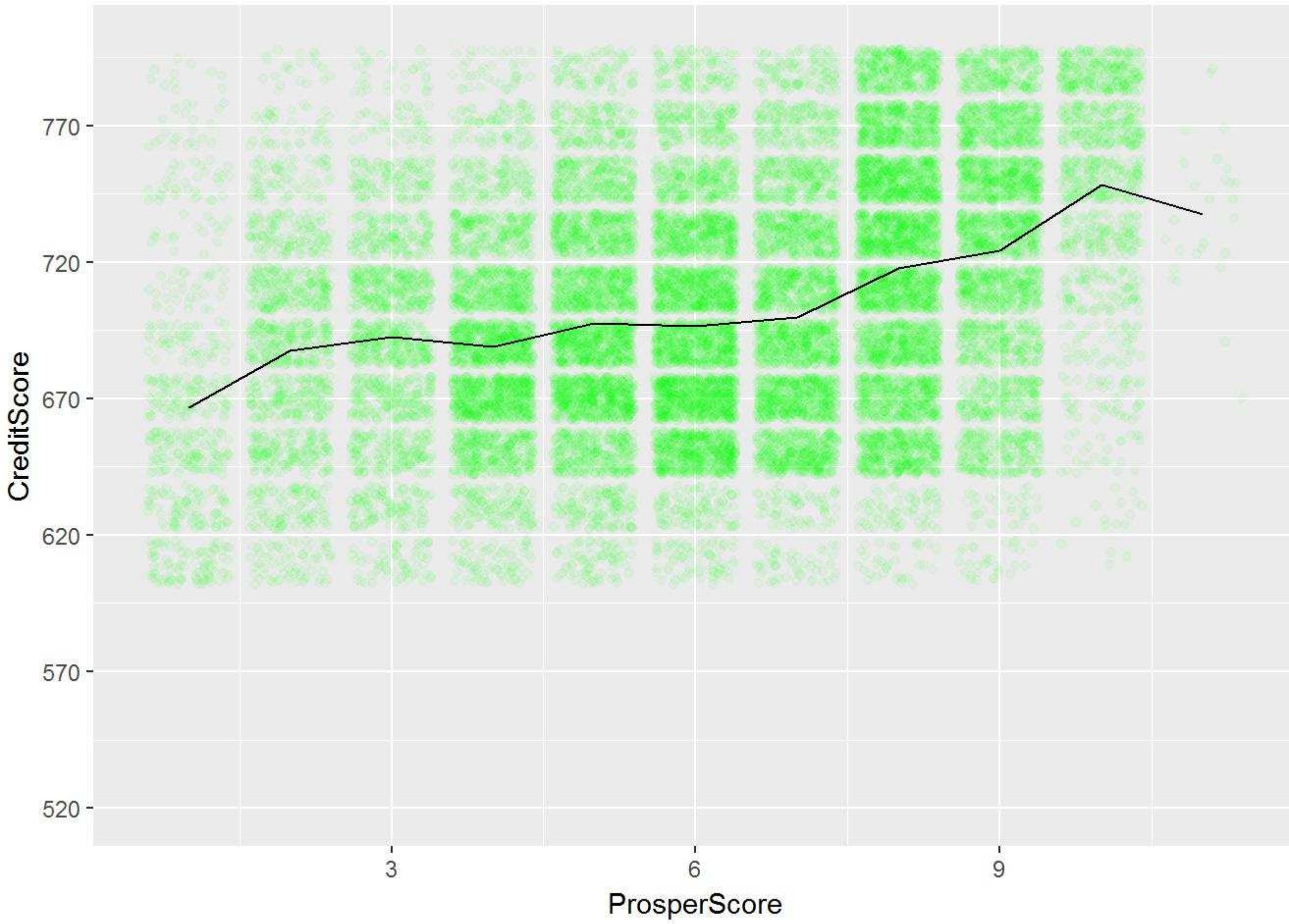



It seems like CreditGrades are implemented base on CreditScore.



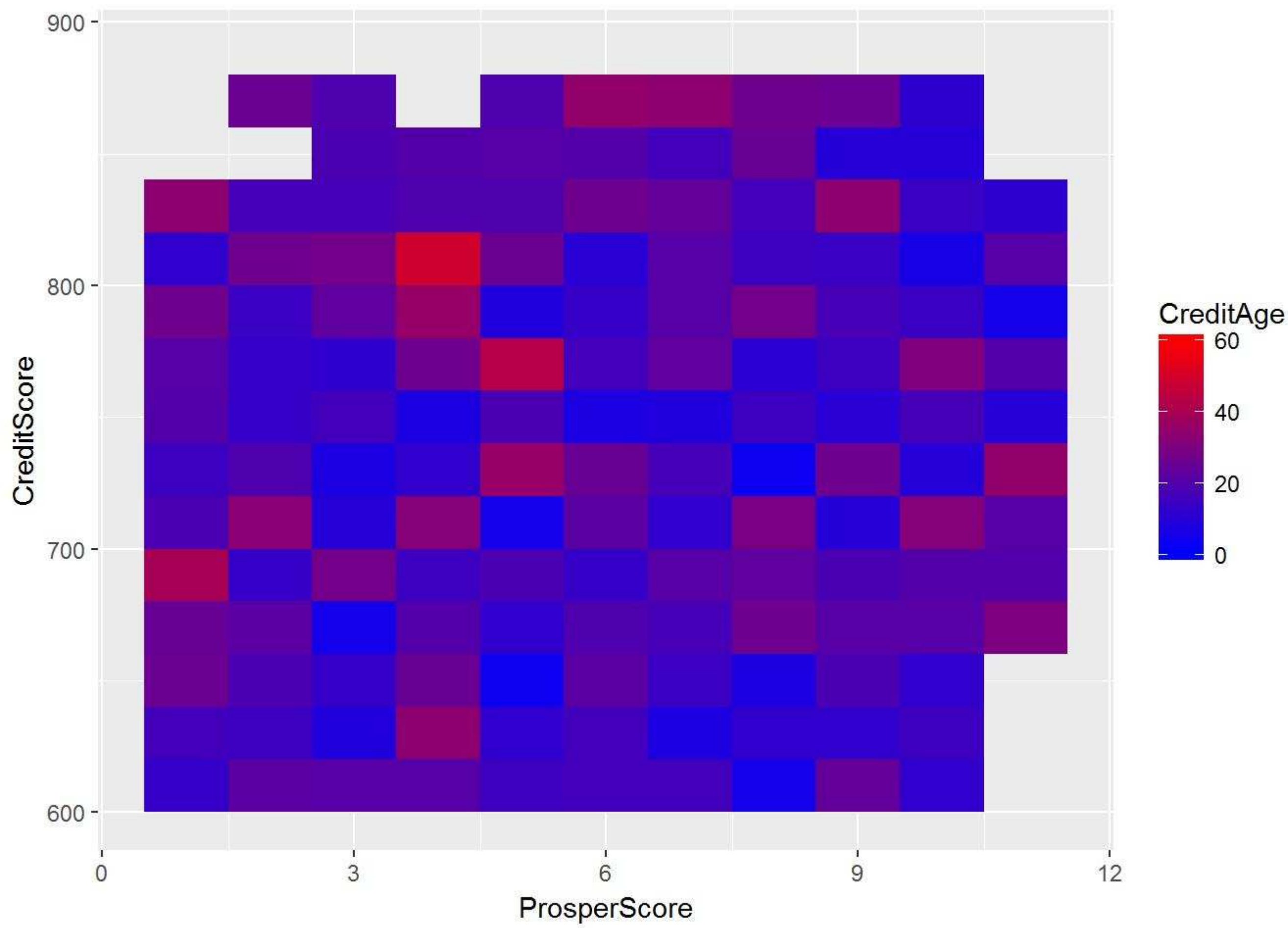

There is a negative correlation between BorrowerAPR and CreditScore:

```
## [1] -0.4022054
```

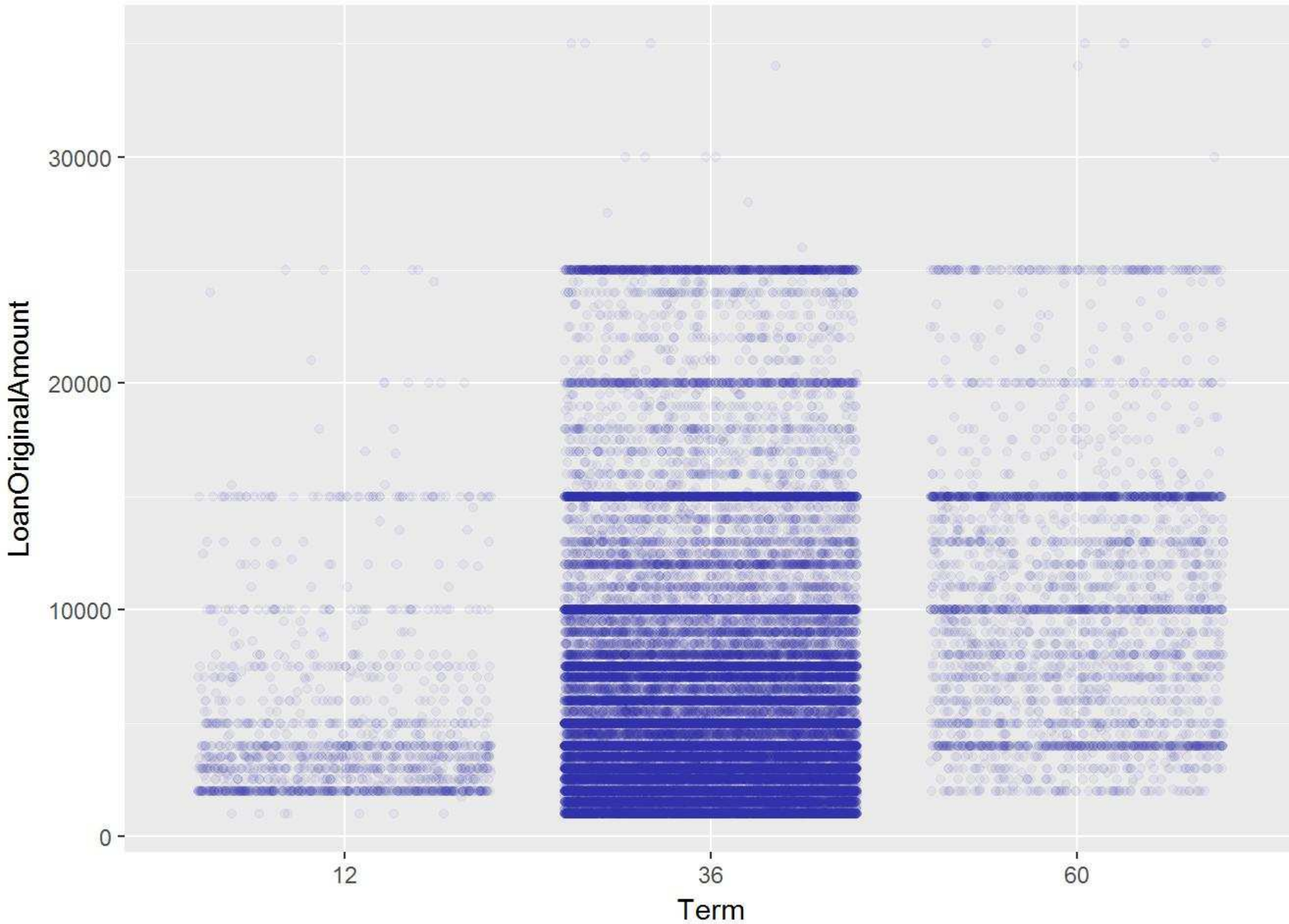



ProsperScore and CreditScore variables are positively correlated; The correlation coefficient is:

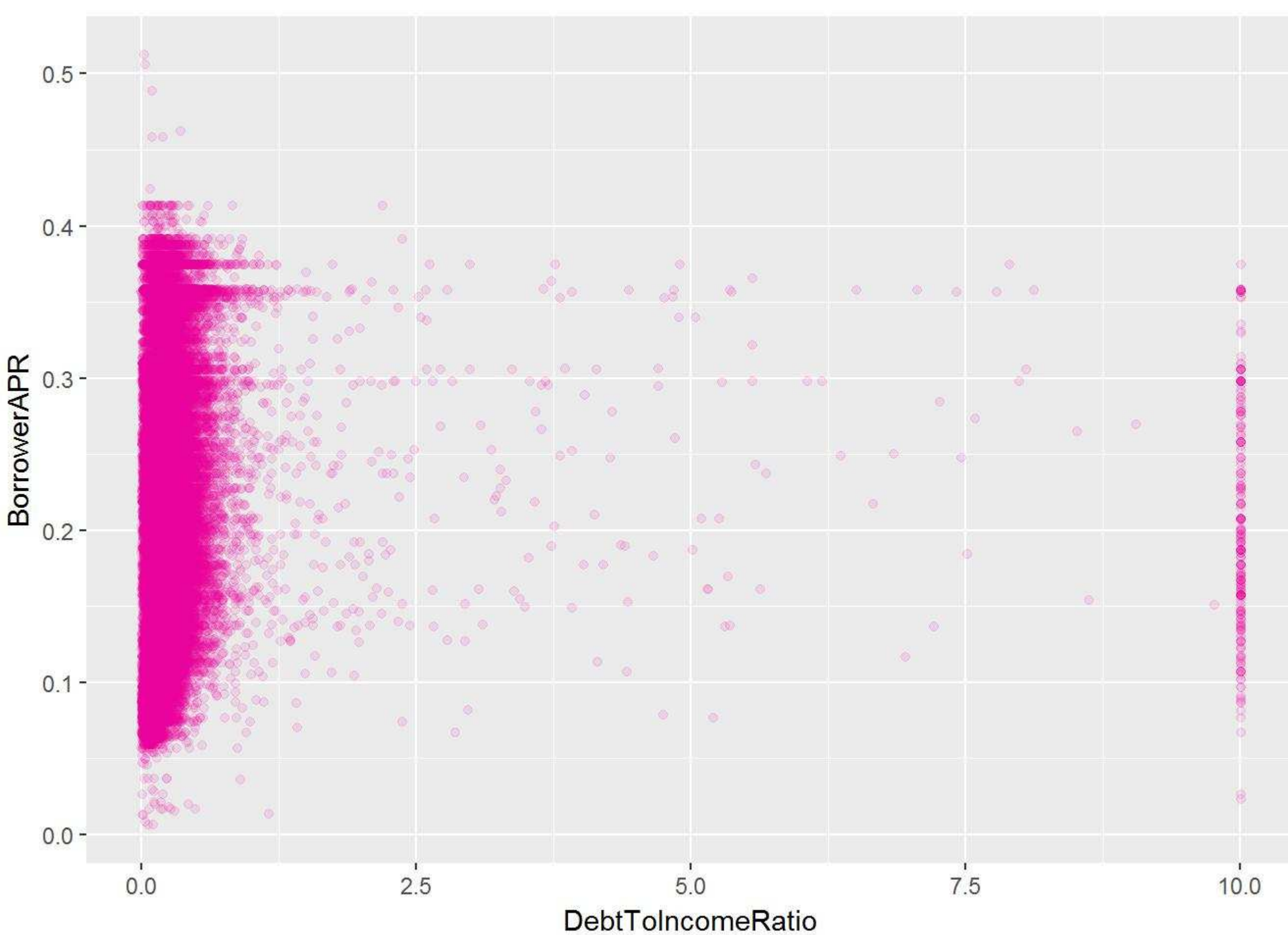
```
## [1] 0.3967875
```




Term and LoanOriginalAmount:



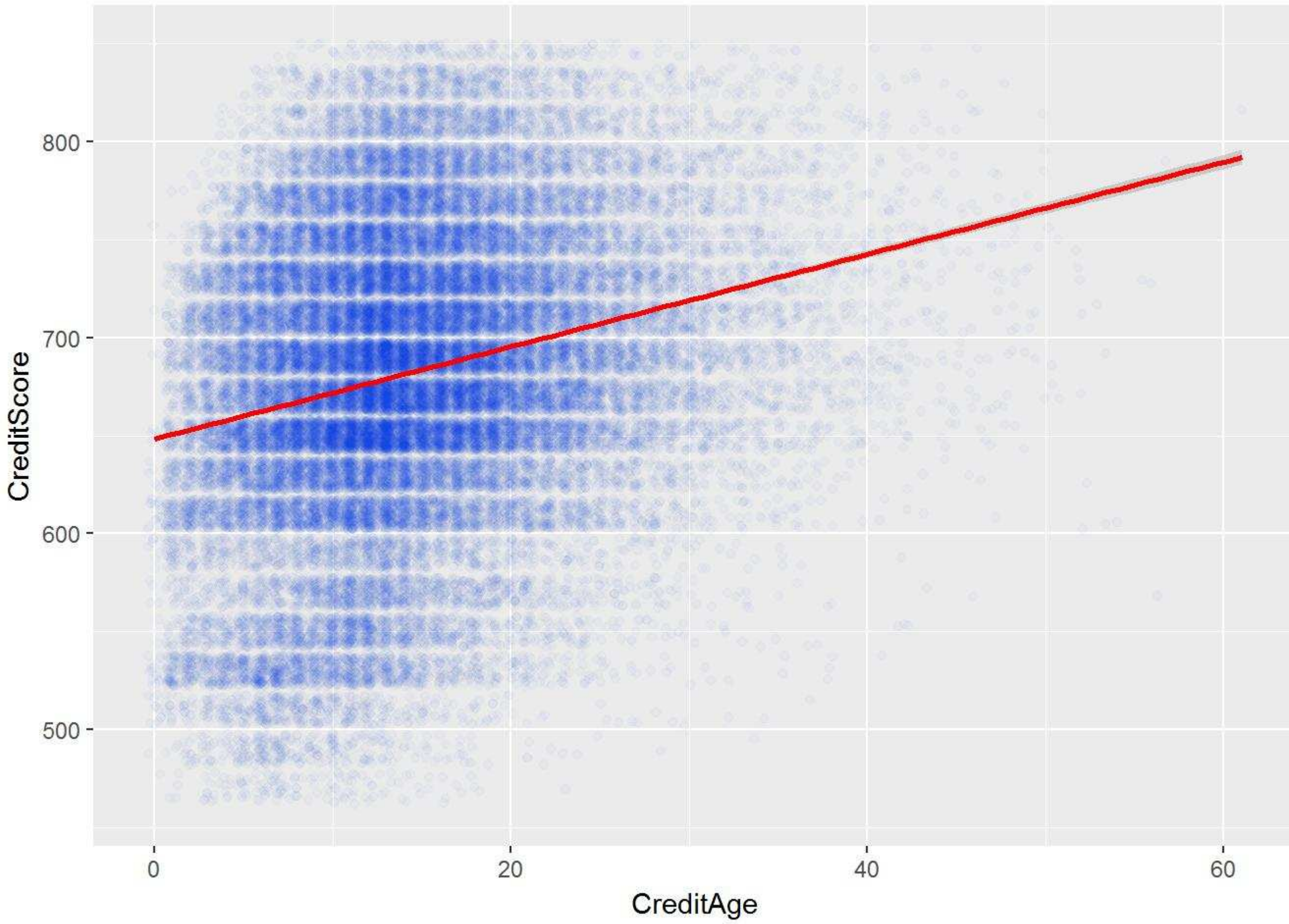
LoanOriginalAmount
30000
20000
10000
0
12
36
60
Term

BorrowerAPR
0.5
0.4
0.3
0.2
0.1
0.0
0.0
2.5
5.0
7.5
10.0
DebtToIncomeRatio

Correlation between DebtToIncomeRatio and BorrowerAPR:

```
## [1] 0.03053818
```




Correlation between CreditAge and CreditScore:

```
## [1] 0.2445919
```

# Bivariate Analysis

## Talk about some of the relationships you observed in this part of the

## investigation. How did the feature of interest vary with other features

## in the dataset?

- Lower BorrowerAPR has a bigger change of being a Completed loan rather than the larger BorrowerAPRs. The majority of completed loans have BorrowerAPR less than 0.2
- Borrowers with more CurrentDelinquencies are more likely to fail a loan rather than borrowers with zero number of CurrentDelinquencies.

- Borrower that complete their loan, has a bigger proportion of larger CreditScore rather than failed loan borrowers.

## Did you observe any interesting relationships between the other features

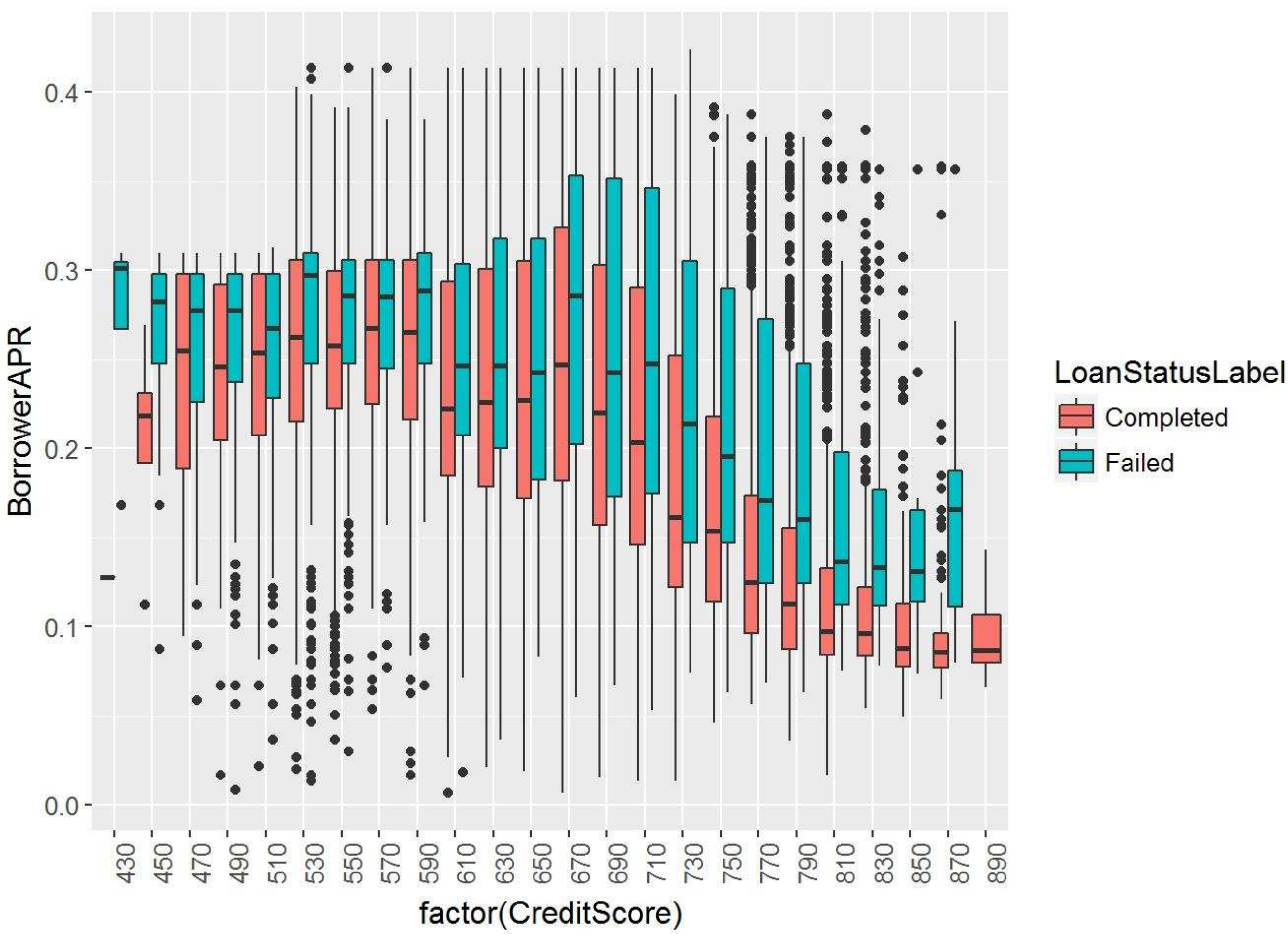
## (not the main feature of interest)?

There are some strong relationships between CreditScore and many other variables like: BankcardUtilization, DebtToIncomeRatio, CurrentDelinquencies. It makes sense. Because CreditScore is a function of many of these variables. On the other hand ProsperScore has good correlation with CreditScore, and it should be because they both have same features for creation.

## What was the strongest relationship you found?

The strongest correlation is between ProsperScore and BorrowerAPR with correlation coefficient equal to -0.73; That’s maybe related to this fact that one of the main patameters for determining BorrowerAPR is ProsperScore;

# Multivariate Plots Section




That’s interesting; Without exception, average APR for borrowers with same CreditScore is more for Failed loans!

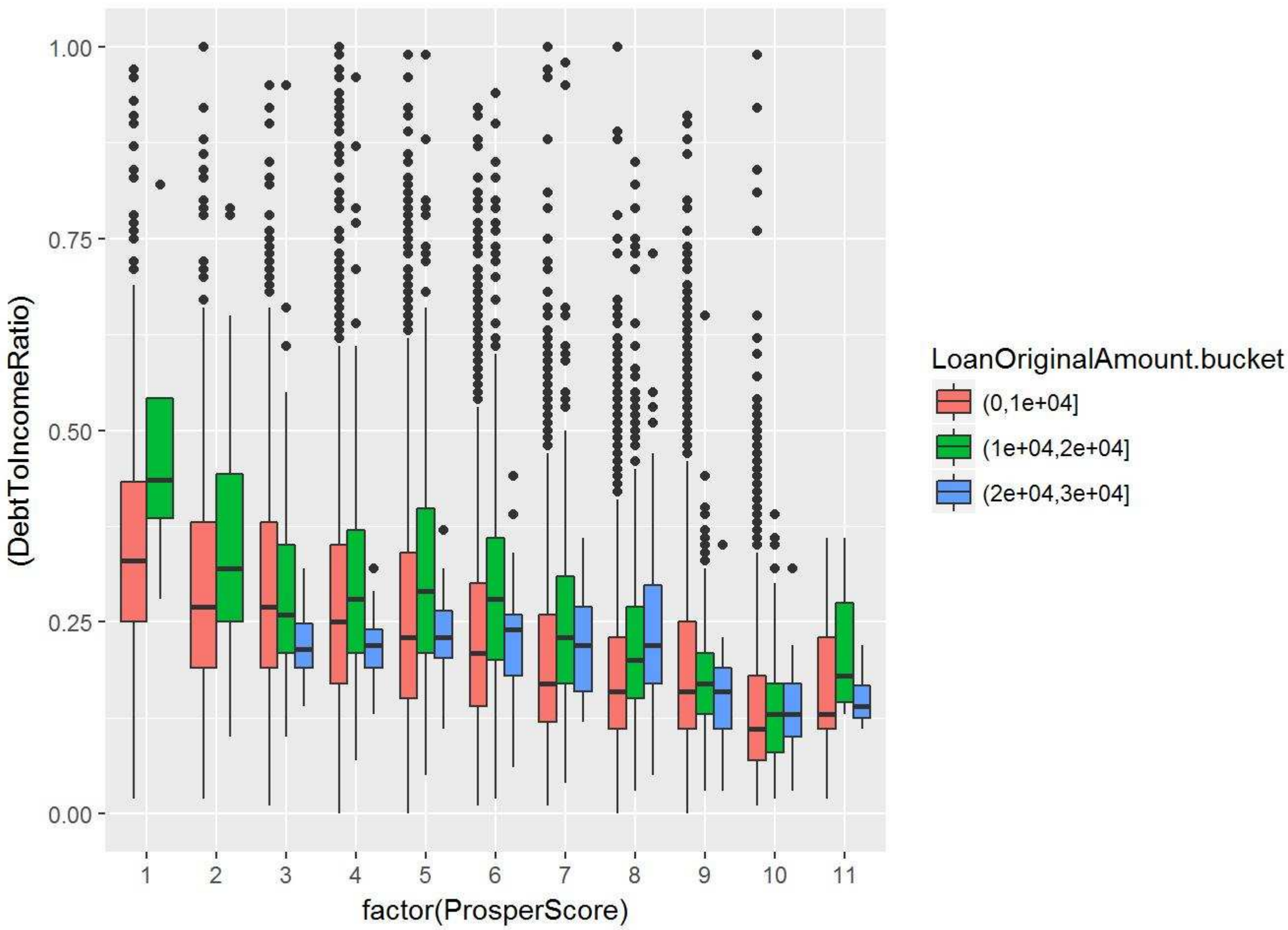
```
##
## Completed    Failed
##      2584       256
```

```
##
## Completed    Failed
##       373       842
```

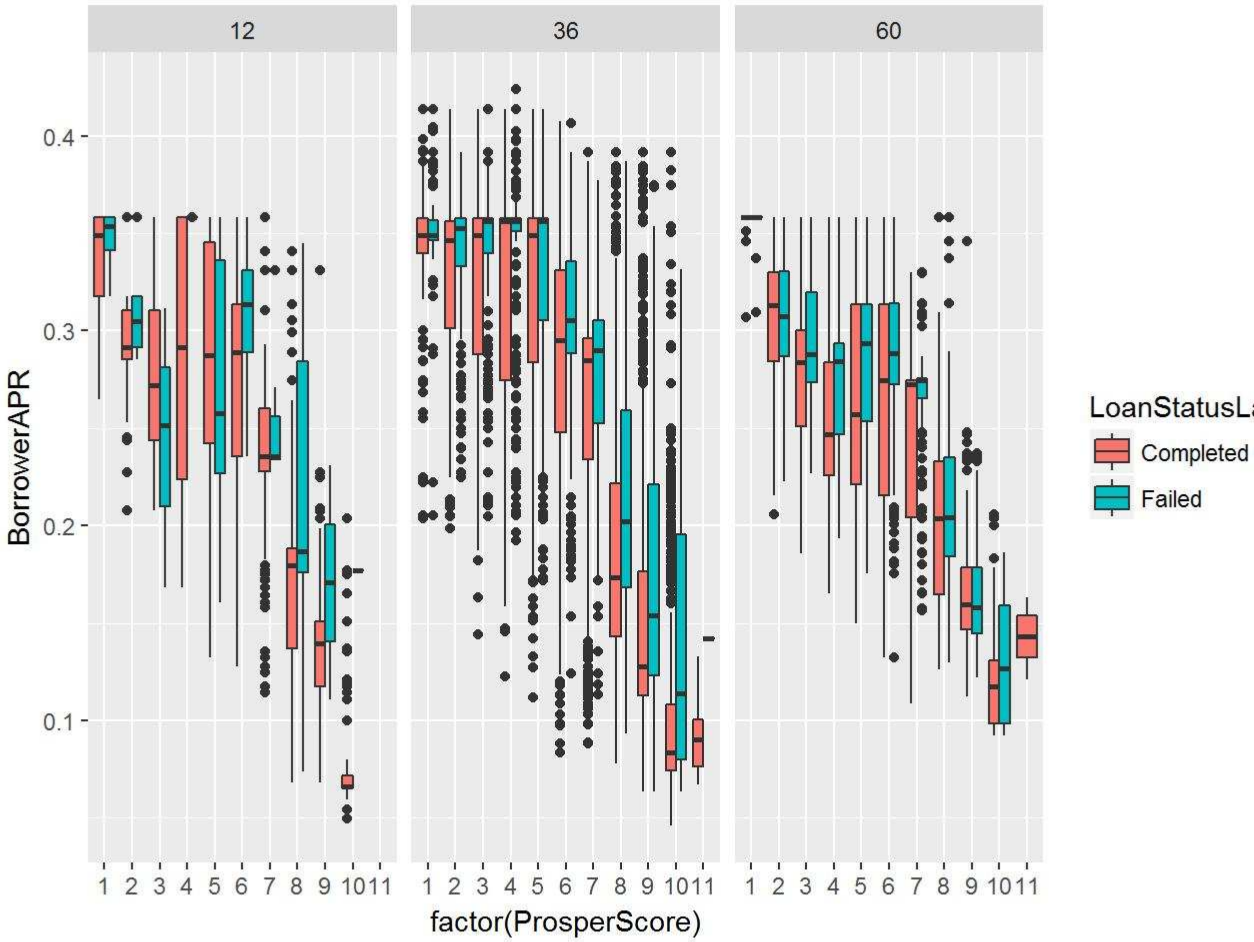
Above 90% of debtors with creditScore more than 810, are Completed their Loan and the majority(69%) of debtors with CreditScore under 510, failed their loan.

```
##
## Completed    Failed
##      4166       412
```

Loans with DebtToIncomeRatio less than 0.5 with ProsperScore more than 8 is very likely to complete(91%).

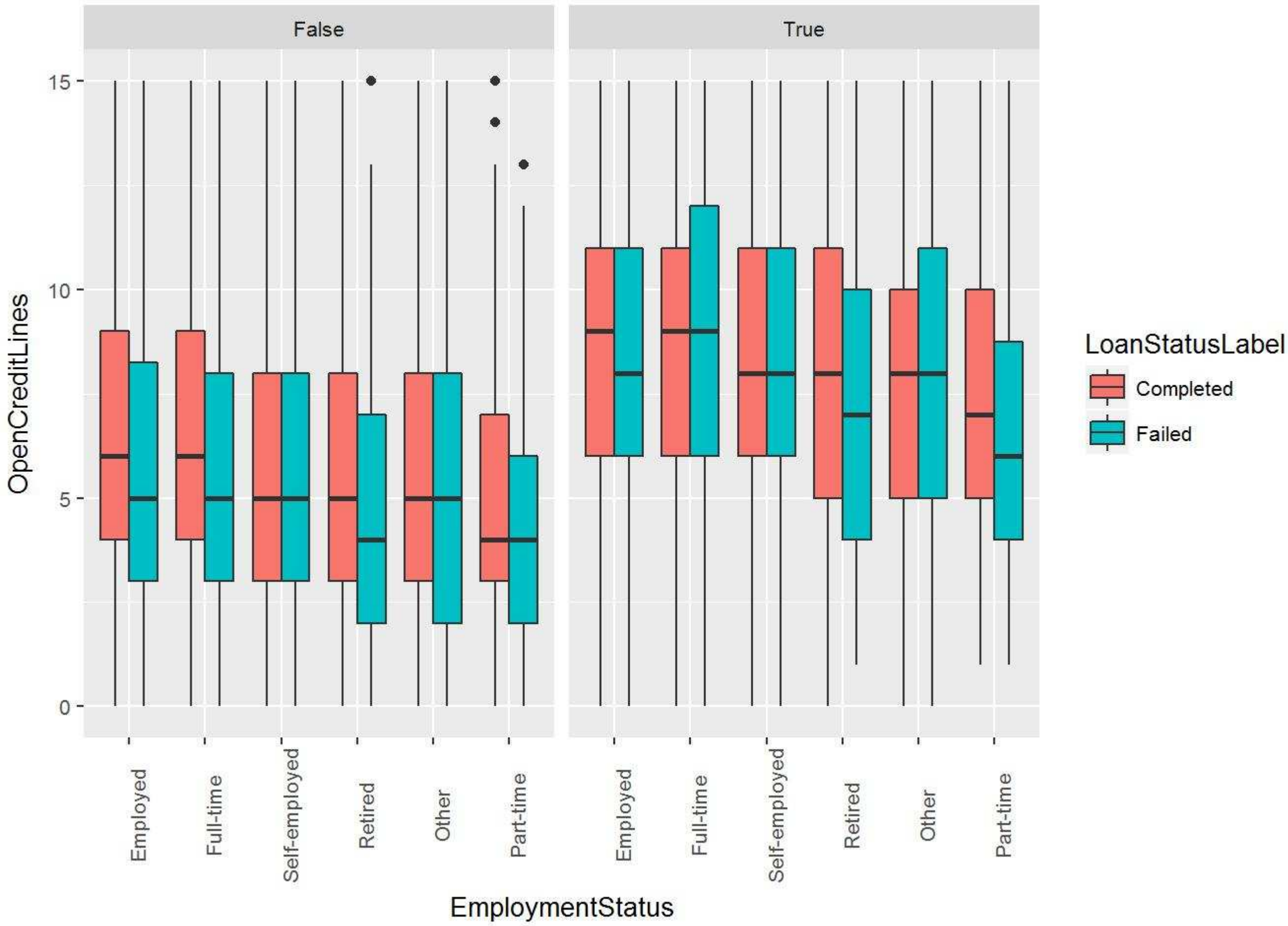


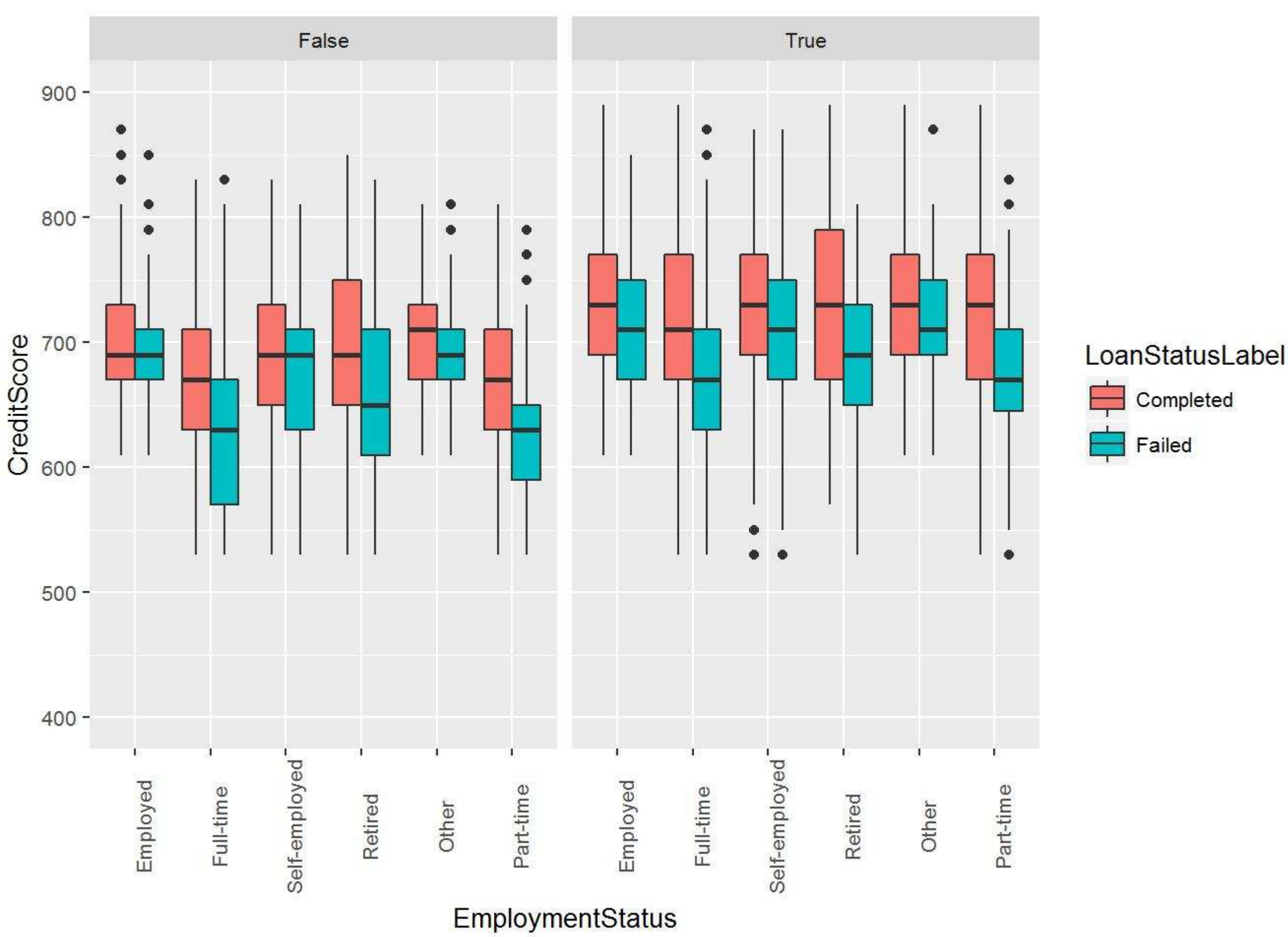

When DebtToIncomeRatio decreases, the ProsperScore increases; For ProsperScore less than 3, there is no LoanOriginalAmount greater than 20000.



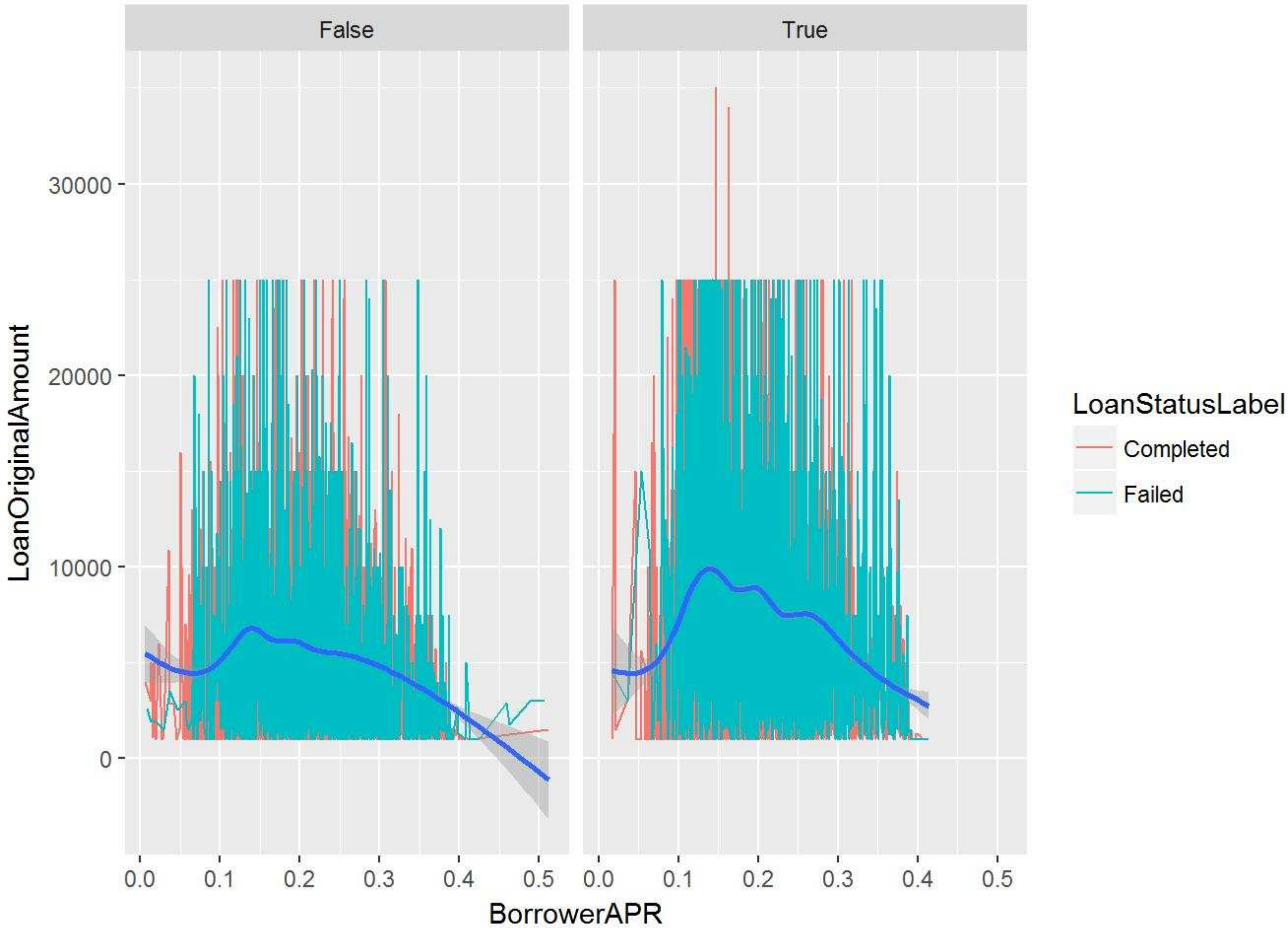

```
##
## Completed    Failed
##       650        12
```

The most successful loans has 12 months Term with BorrowerAPR less than 0.2 and Employed or Fulltime job borrower; 98% of these loans are Completed.


False
True
OpenCreditLines
0
5
10
15
Employed
Full-time
Self-employed
Retired
Other
Part-time
EmploymentStatus
LoanStatusLabel
Completed
Failed


False
True
CreditScore
400
500
600
700
800
900
Employed
Full-time
Self-employed
Retired
Other
Part-time
EmploymentStatus
LoanStatusLabel
Completed
Failed

HomeOwners has larger margings for CreditSCore and if they are Employed or Fulltime job, it’s more likely to complete their loan.




```
##
## False  True
##   880  1829
```

In general HomeOwners are qualified for larger loans and the percentage of their failure is more than not HomeOwners.

# 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 of interest?

Terms of loans and BorrowerAPR and EmploymentStatus are creat a good predictor for LoanStatusLabel. e.g. a loan with 12 months Term and APR less than 0.2 for a borrower that is Employed or has Fulltime job is very likely to complete.

## Were there any interesting or surprising interactions between features?

HomeOwners with Original Loan amount more than 10000 are more likely to fail thier prosper loan rather than not HomeOwners the ratio is 67%.

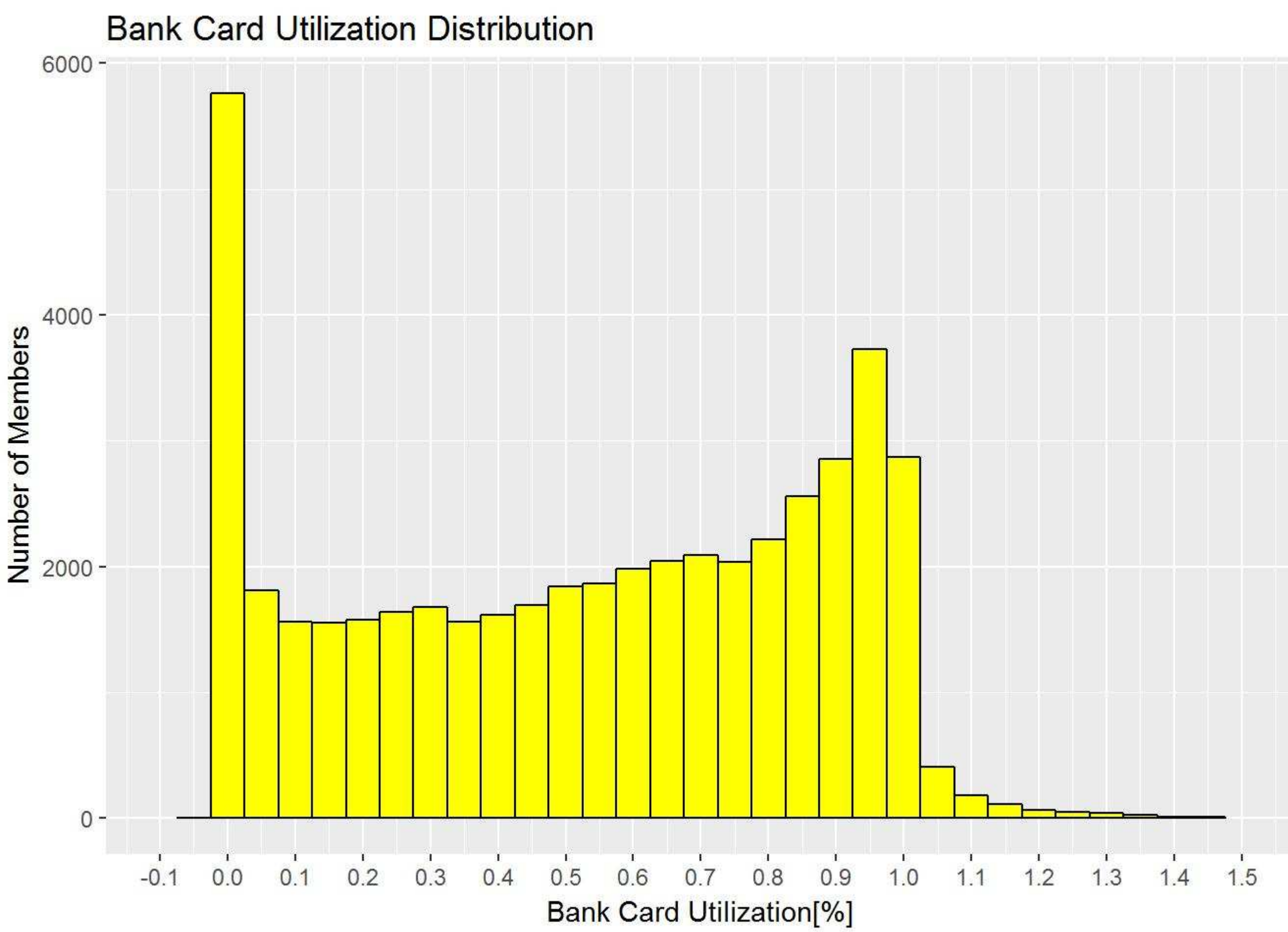
## OPTIONAL: Did you create any models with your dataset? Discuss the

## strengths and limitations of your model.

I think with some of these features we can train a classifier for predicting Completed of Failed loans. ——

# Final Plots and Summary

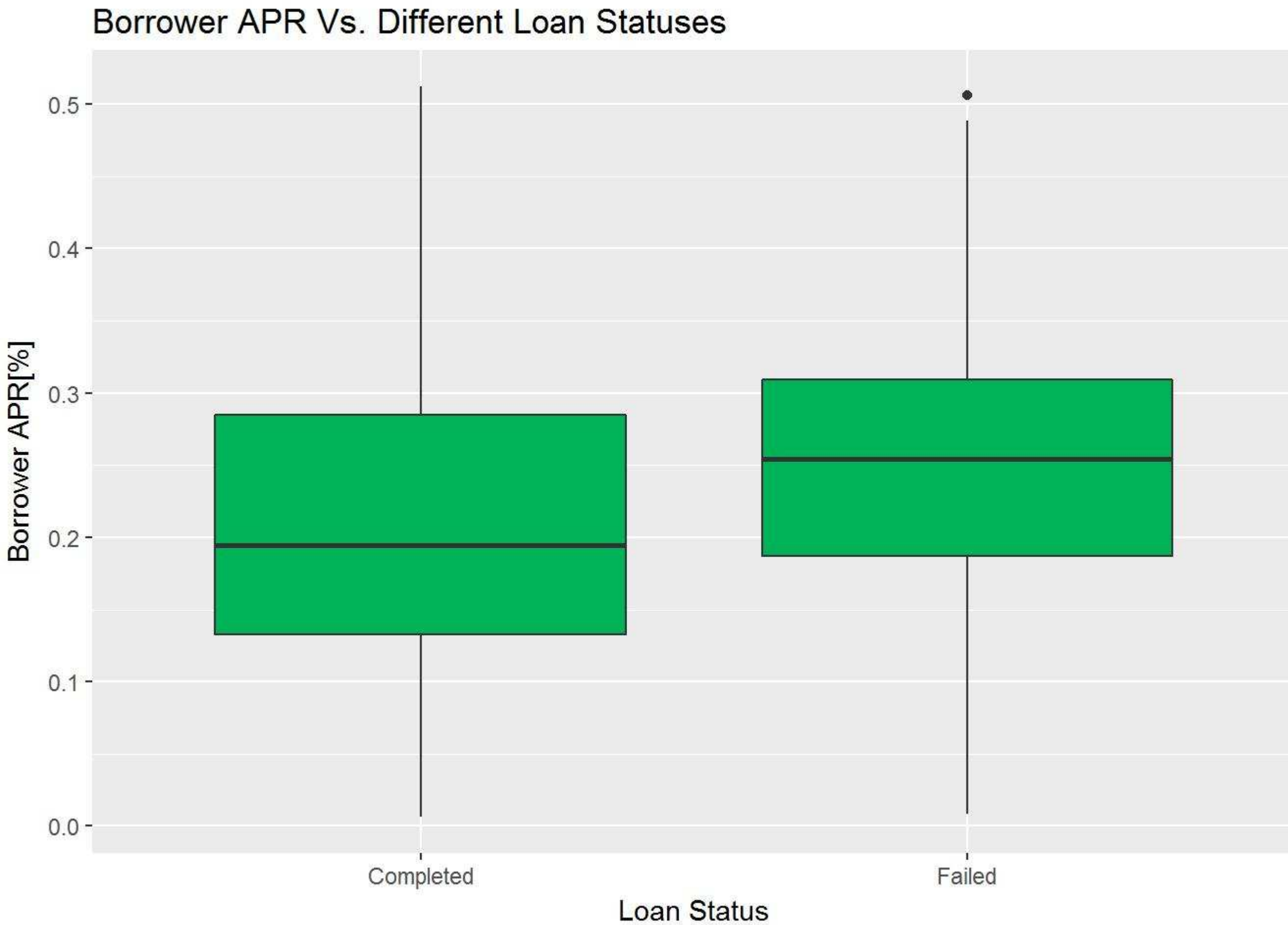
## Plot One




## Description One

It’s a bimodal distribution. Many of members have 0 revolving bank card utilization and the second crowded point is 0.95 utilization. It’s interesting because there are two peaks in two sides of the spectrum. The first quartile is 0.21 and mean of this variable is 0.53. The maximun value of this variable is 5.95;

# Plot Two

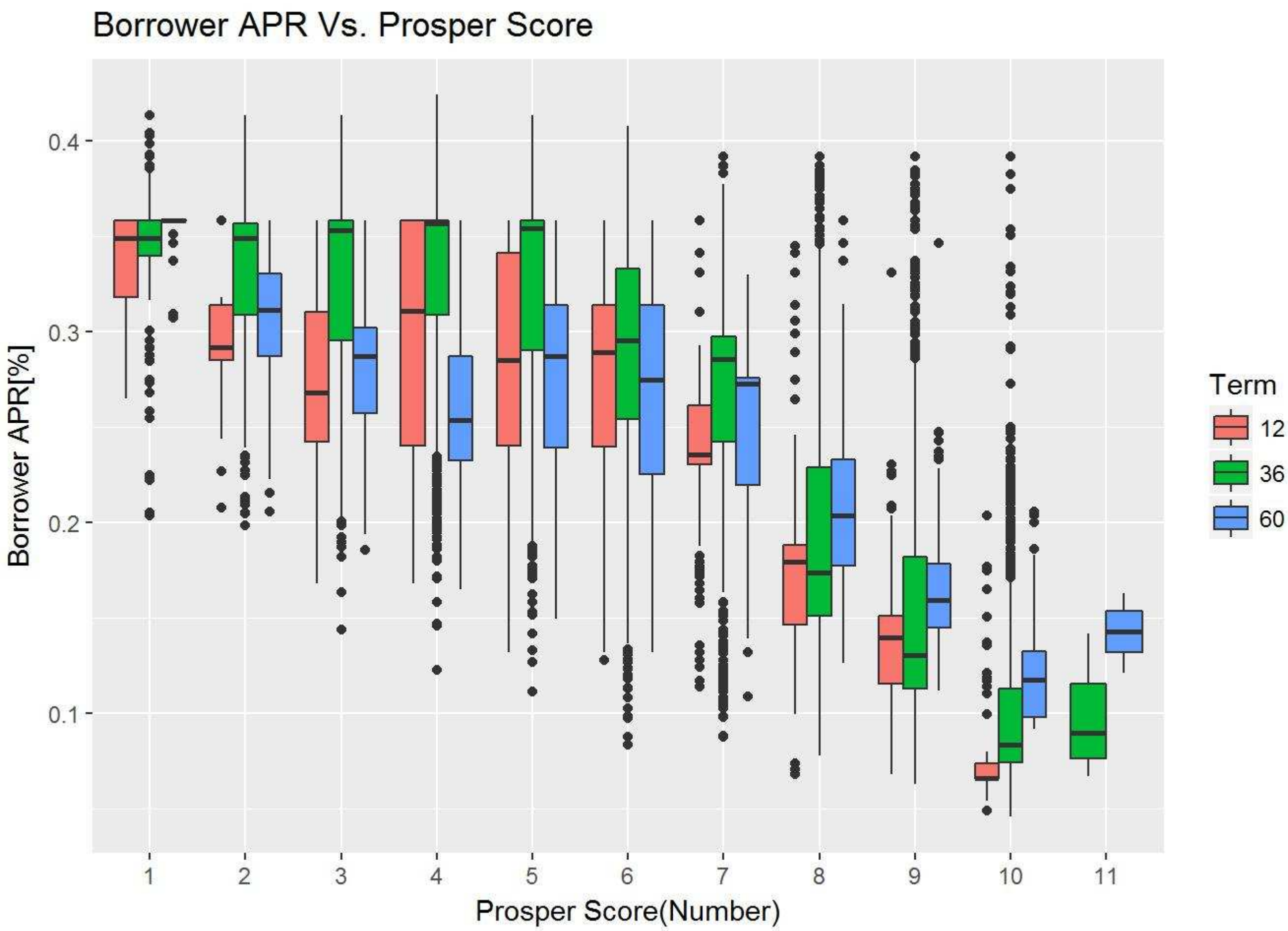



```
##
## Completed    Failed
##     19638      5062
```

# Description Two

The BorrowerAPR is a strong predictor for the loan status. Mean of APR for Completed loans is roughly 0.05 less than mean of APR for Failed loans. The majority of loans with the APR more than 0.42 are failed: 6 faild in total 7. The distribution of APR for Completed and Failed loans are very different. The Completed loans have a right skewed distribution and the Failed loans have a left skewed distribution. It means most of Completed loans have lower APRs and most of the Failed loans have larger APRs. 80% of loans with APR less than 0.2 are completed. Minimun APR in this dataset is 0.0063 and maximum value is 0.512; The average of APR is 0.222;

# Plot Three




```
##
## Completed    Failed
##       650        12
```

## Description Three

The most successful loans have 12 months Term. On the other hand 80% of loans with Borrower APR less than 0.2 are completed. Another important parameter is Employment status. Best categories for Employment Status are Employed and Full-time. If we add up all these conditions together the loans will have 98% chance of being completed. That’s a really high chance. There are 16491 borrower that are Employed and 24957 borrower has Full-time job; Very few of borrowers are not employed: 561; It makes sense because when Prosper choose somebody to offer a loan, they prefer that person has some kind of job; ——

# Reflection

The Prosper loan data set contains 113,937 clients with 81 variables. I choose 16 variables and focus on specific categories and I found 55084 clients in the new dataset. Some variables are related to scoring clients. Variables like CreditScore, ProsperScore, CreditGrade. These scores has close relationship together.

I was struggling how I choose a handful of variables between 81
variables from the original dataset. I choose the most independent and not IDs and gradually decrease them till I reach 16 of them. Another problem was the dataset has null values for NAs and it takes a while to understand why

I can not filterout NAs with !is.na(variable). Another thing that I made decision about was LoanStatus variable. It contain many different categories and after considering them, I understand I only need two labels for loans: Completed or Failed and the other categories are not finished and I can not use them as label.

I believe finding main features for predicting the loan status is a great success. And my graphs show how strong these features predict the loan status. This is important for implementing the classifier for future research. Defining CreditAge and CreditScore as two new variables are another success of this study. Finding the importance of Borrower APR from different aspects, is another success that I personally like it.

The process I went through has these stages: First I investigate each variable distribution and used some transformations for changing some non-normal distributions. Next I investigate the relationship and correlations between different features. Gradually I defined the project goal, finding the most important predictors for successfull loans or Completed ones. I realize that The BorrowerAPR and Terms of loan and ProsperScore and Employment Status are good predictors for categorizing loans in Completed and Failed groups.

The next step base on these findings could be training a decision tree or naive bayes classifier base on mentioned fetures for Loas Status as target.

Some limitation of this proccess should be the size of the filtered dataset. Because I only focus on Completed and Failed loan statuses and I can not more than 50% of the original dataset because those loans results are not finished and they are in progress. If I have access to the updated dataset, I should use all original datapoints for training and testing purposes and that really help when I double size of dataset.