Prosper借贷数据探索

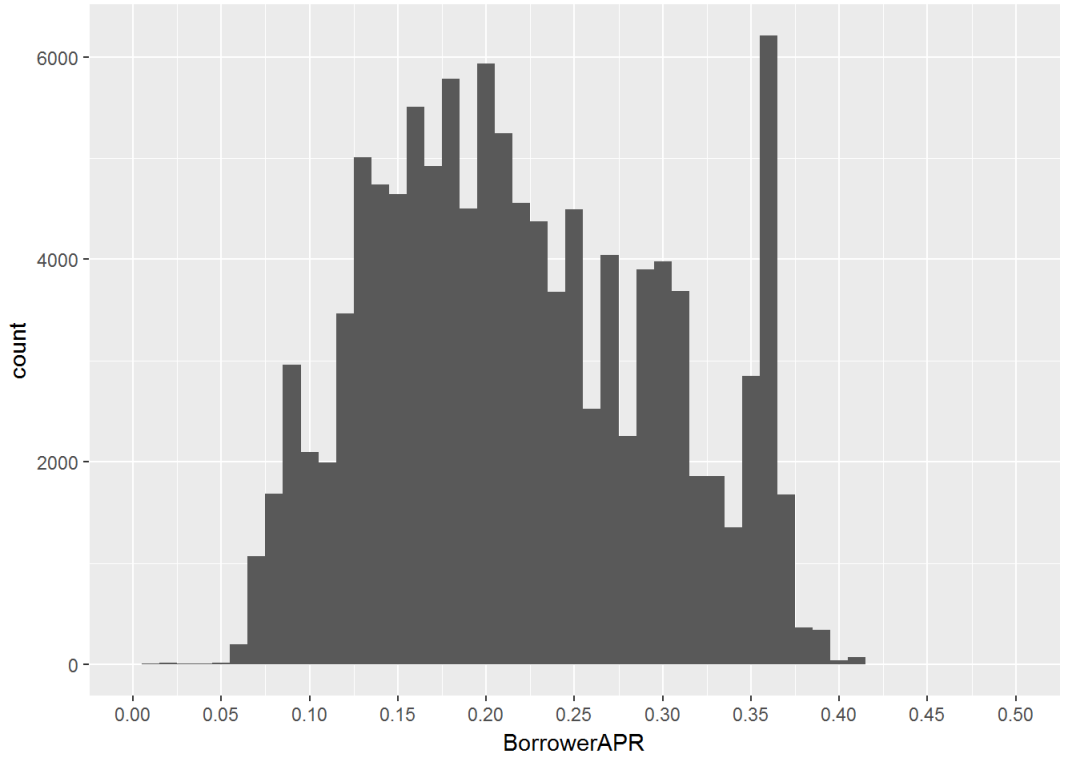
Univariate Plots Section

```
## [1] 113937     13
```

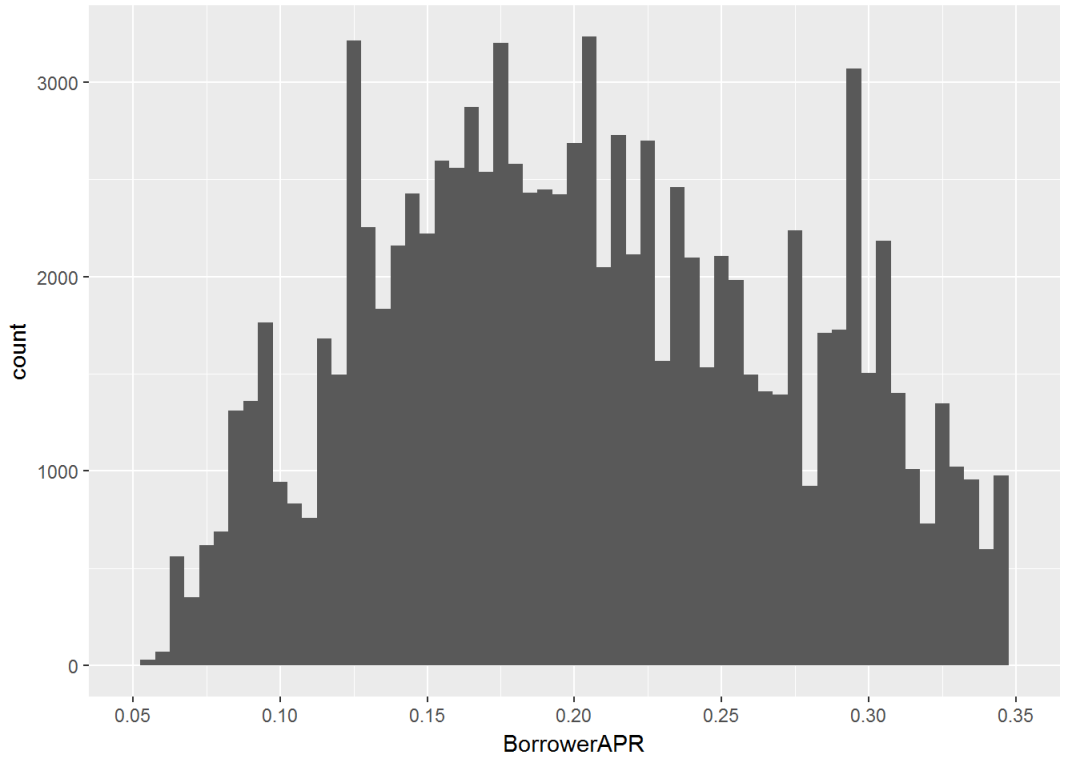
```
## 'data.frame':    113937 obs. of  13 variables:
##  $ Term                   : int  36 36 36 36 36 60 36 36 36 36 ...
##  $ BorrowerAPR            : num  0.165 0.12 0.283 0.125 0.246 ...
##  $ ProsperRating..Alpha.  : Factor w/ 8 levels "","A","AA","B",..: 1 2 1 2 6 4 7 5 3 3 ...
##  $ Occupation             : Factor w/ 68 levels "","Accountant/CPA",..: 37 43 37 52 21 43 50 29 24 24
...
##  $ EmploymentStatus       : Factor w/ 9 levels "","Employed",..: 9 2 4 2 2 2 2 2 2 2 ...
##  $ IsBorrowerHomeowner    : Factor w/ 2 levels "False","True": 2 1 1 2 2 2 1 1 2 2 ...
##  $ CreditScoreRangeLower  : int  640 680 480 800 680 740 680 700 820 820 ...
##  $ CreditScoreRangeUpper  : int  659 699 499 819 699 759 699 719 839 839 ...
##  $ DelinquenciesLast7Years: int  4 0 0 14 0 0 0 0 0 0 ...
##  $ DebtToIncomeRatio      : num  0.17 0.18 0.06 0.15 0.26 0.36 0.27 0.24 0.25 0.25 ...
##  $ IncomeRange            : Factor w/ 8 levels "$0","$1-24,999",..: 4 5 7 4 3 3 4 4 4 4 ...
##  $ StatedMonthlyIncome    : num  3083 6125 2083 2875 9583 ...
##  $ LoanOriginalAmount     : int  9425 10000 3001 10000 15000 15000 3000 10000 10000 10000 ...
```

```
##  Term        BorrowerAPR      ProsperRating..Alpha.
##  12: 1614   Min.   :0.00653          :29084
##  36:87778   1st Qu.:0.15629   C      :18345
##  60:24545   Median :0.20976   B      :15581
##             Mean   :0.21883   A      :14551
##             3rd Qu.:0.28381   D      :14274
##             Max.   :0.51229   E      : 9795
##             NA's   :25        (Other):12307
##                     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
##  IsBorrowerHomeowner CreditScoreRangeLower CreditScoreRangeUpper
##  False:56459         Min.   :  0.0         Min.   : 19.0
##  True :57478         1st Qu.:660.0         1st Qu.:679.0
##                      Median :680.0         Median :699.0
##                      Mean   :685.6         Mean   :704.6
##                      3rd Qu.:720.0         3rd Qu.:739.0
##                      Max.   :880.0         Max.   :899.0
##                      NA's   :591           NA's   :591
##  DelinquenciesLast7Years DebtToIncomeRatio         IncomeRange
##  Min.   : 0.000          Min.   : 0.000    $25,000-49,999:32192
##  1st Qu.: 0.000          1st Qu.: 0.140    $50,000-74,999:31050
##  Median : 0.000          Median : 0.220    $100,000+     :17337
##  Mean   : 4.155          Mean   : 0.276    $75,000-99,999:16916
##  3rd Qu.: 3.000          3rd Qu.: 0.320    Not displayed : 7741
##  Max.   :99.000          Max.   :10.010    $1-24,999     : 7274
##  NA's   :990             NA's   :8554      (Other)       : 1427
##  StatedMonthlyIncome LoanOriginalAmount
##  Min.   :      0     Min.   : 1000
##  1st Qu.:   3200     1st Qu.: 4000
##  Median :   4667     Median : 6500
##  Mean   :   5608     Mean   : 8337
##  3rd Qu.:   6825     3rd Qu.:12000
##  Max.   :1750003     Max.   :35000
##
```

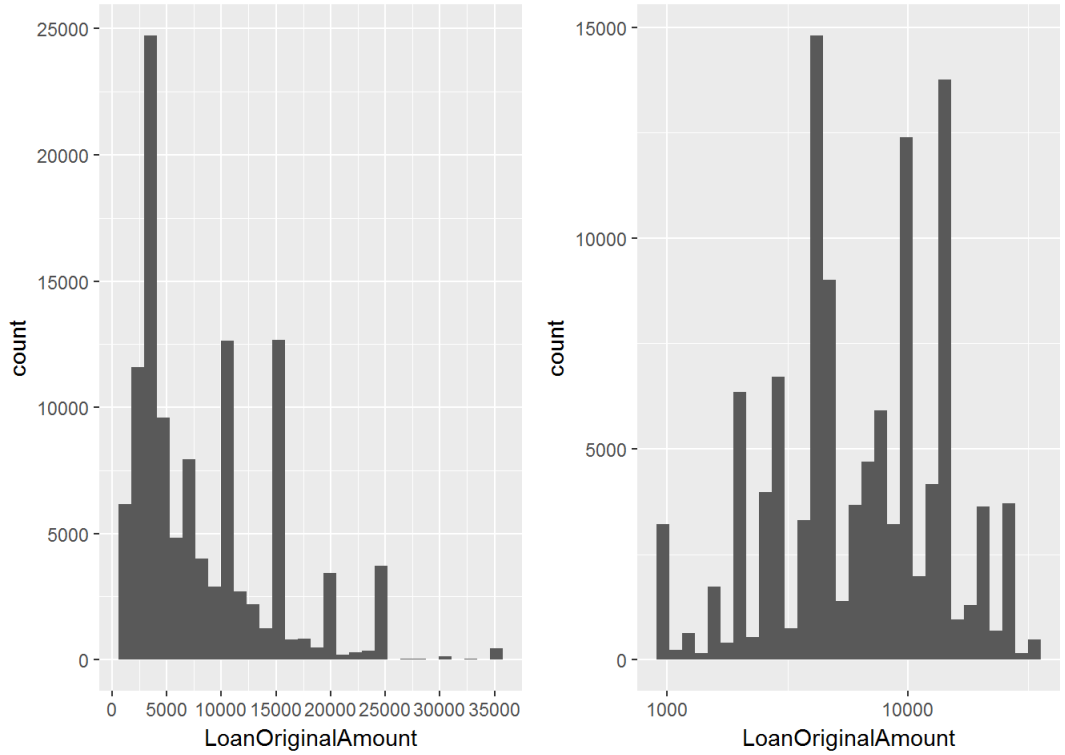
该数据集共有13个变量及113937条借款人数据。



BorrowerAPR存在缺失值，由于分布接近正态，故用均值来填充。

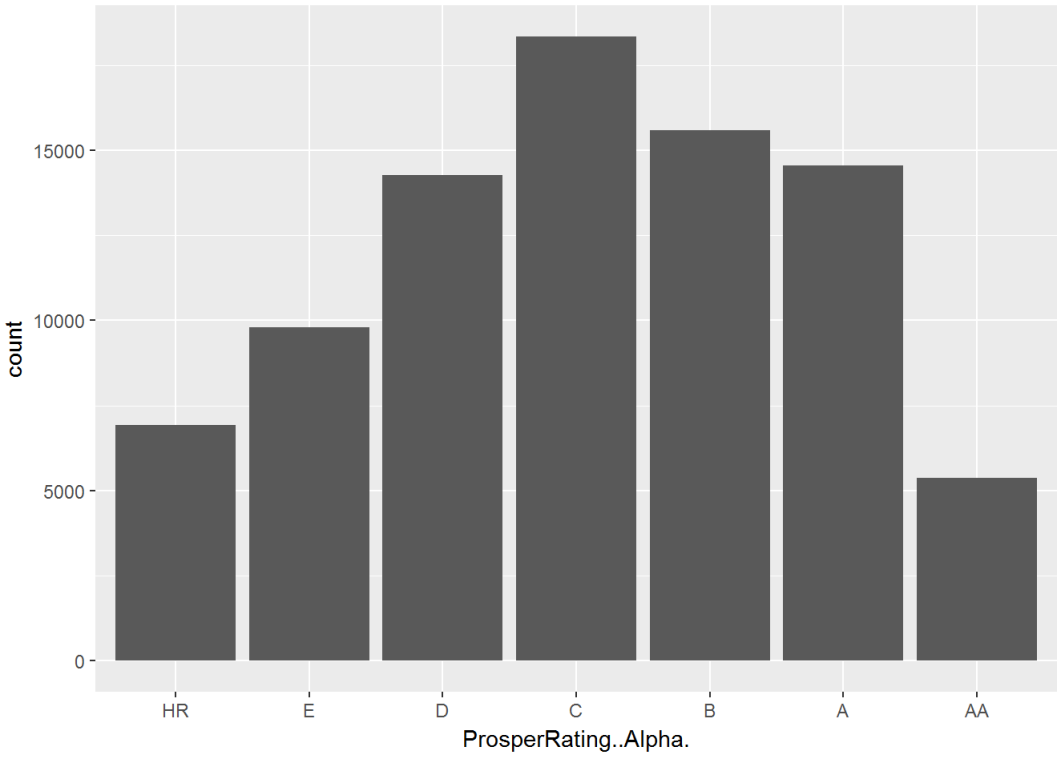


BorrowerAPR的值主要落在[0.05,0.35]，故仅观察该区间的分布，依然呈现正态，说明借款利率的确定不会过于激进。

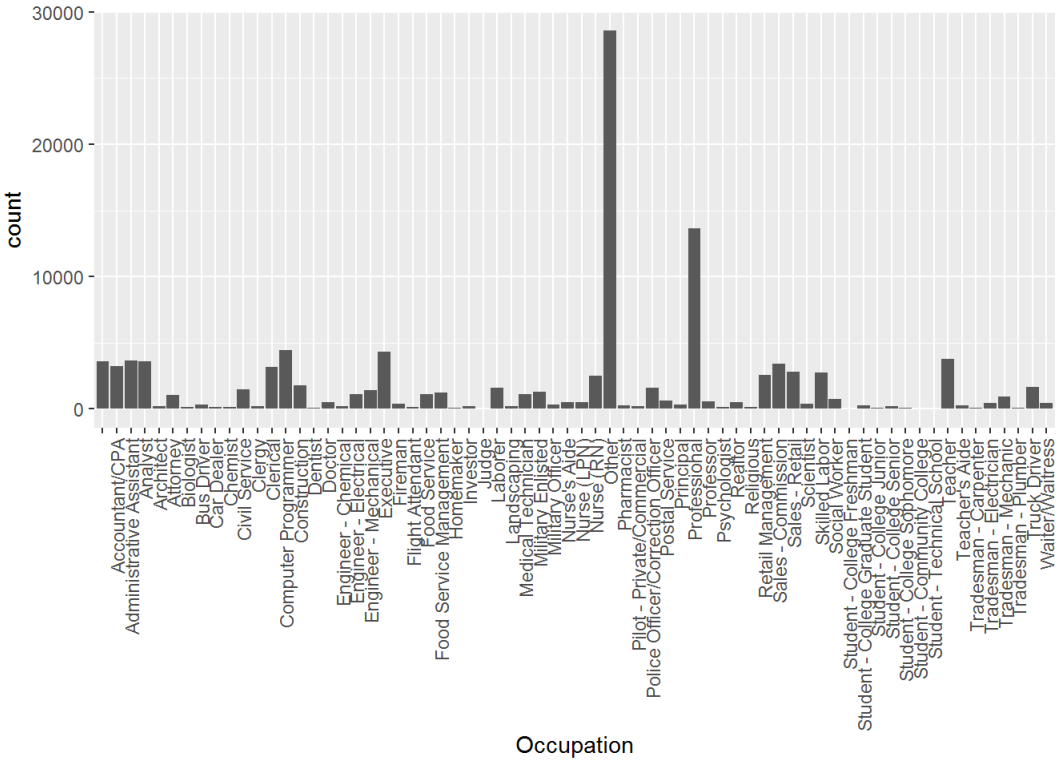


`LoanOriginalAmount`呈现正偏态分布，故进行对数转换来更好地观察分布，转换后呈现多峰形态，峰值分别为`4000`,`10000`

和15000左右。



`ProsperRating..Alpha.`呈现正态分布，`C`为最常见的等级，最高和最低等级的人数不多，说明资质的筛选较为稳健。

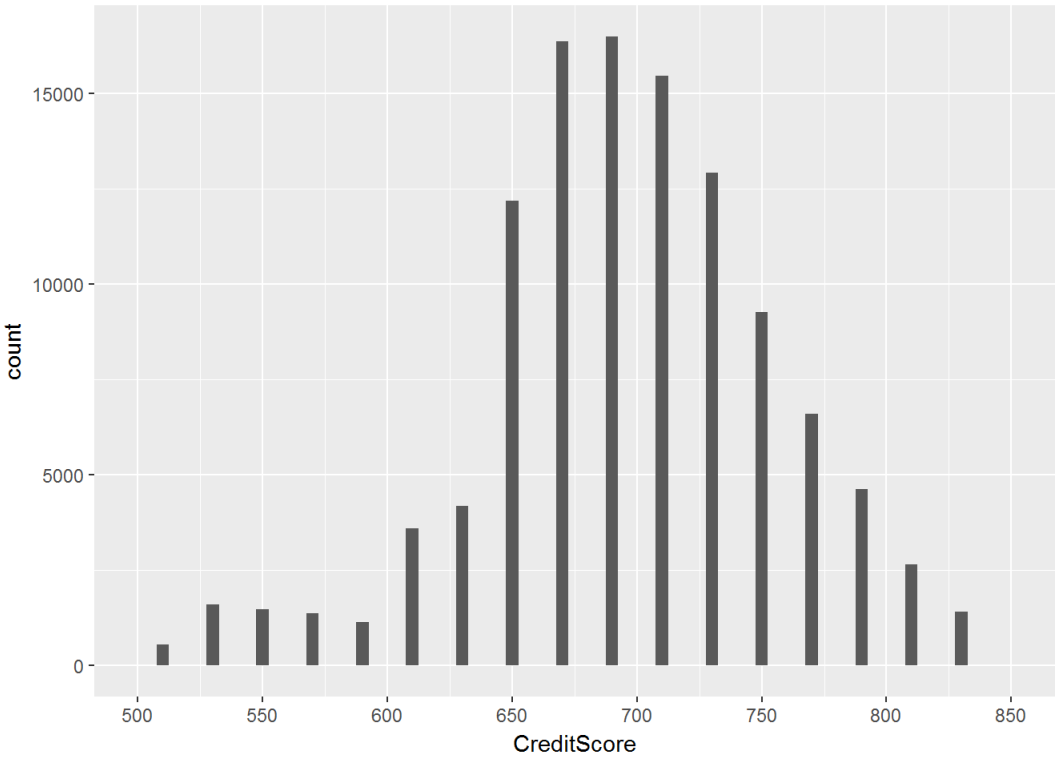


可以看到很多人在职业上选择了`other`，存在敷衍选择的可能，除此之外，职业最多的为专业人士，意味着他们的贷款需求最大

建立一个新变量`CreditScore`，以便能更为方便地对数据集进行探讨，它是对应`CreditScoreRangeLower`与`CreditScoreRangeUpper`的平均数。

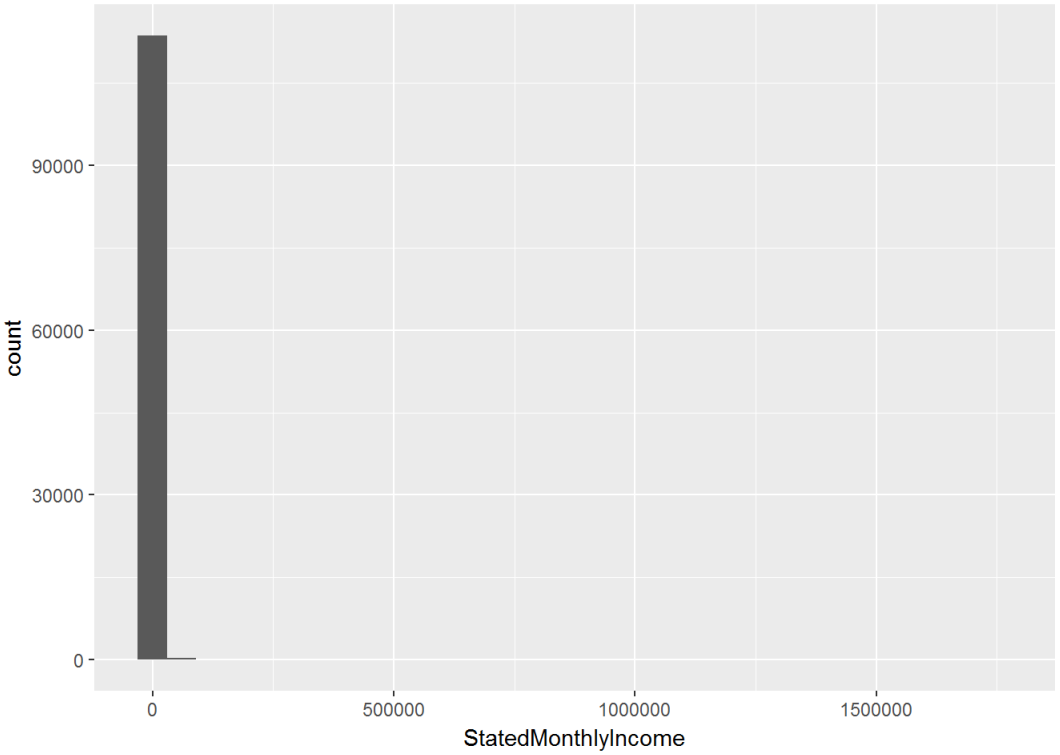
```
##
##   9.5 369.5 429.5 449.5 469.5 489.5 509.5 529.5 549.5 569.5 589.5 609.5
##   133     1     5    36   141   346   554  1593  1474  1357  1125  3602
## 629.5 649.5 669.5 689.5 709.5 729.5 749.5 769.5 789.5 809.5 829.5 849.5
##  4172 12199 16366 16492 15471 12923  9267  6606  4624  2644  1409   567
## 869.5 889.5
##   212    27
```

可以看到`CreditScore`是离散的，且`500`以下和`850`以上的取值很少，故对`x`轴进行相应限制。



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     9.5   669.5   689.5   695.1   729.5   889.5     591
```

CreditScore的值主要集中在650-750，且存在缺失值，由于分布接近正态，故用均值来填充。

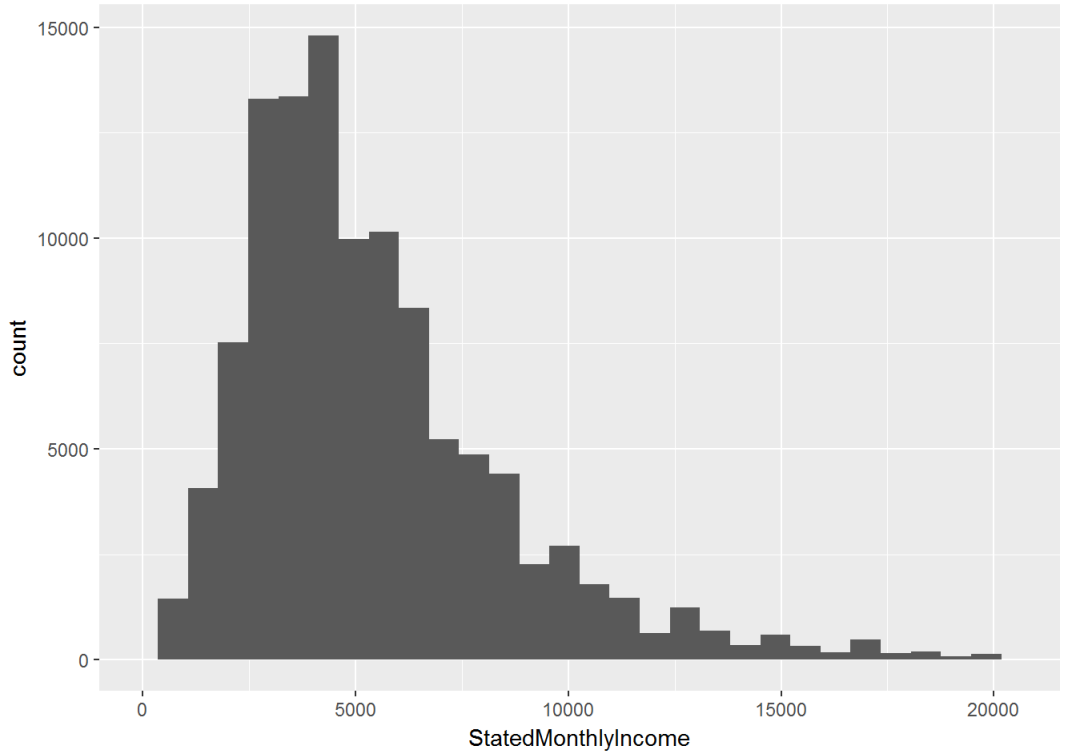


上图说明StatedMonthlyIncome存在极端值。

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    3200    4667    5608    6825 1750003
```

事实确实如此，月收入最高值为1750003美元，远远超过均值。

```
##  [1] 1750002.92  618547.83  483333.33  466666.67  416666.67  394400.00
##  [7]  250000.00  208333.33  185081.75  185081.75  158333.33  150000.00
## [13]  140416.67  120833.33  108750.00  108333.33  103334.08  100000.00
## [19]   96266.50   91666.67
```



可以看到并不是只有个别月收入超高的借款人，故去除分布中末端1%的数据，可以得到一个正偏态分布,StatedMonthlyIncome主要集中在2500-5000美元。

Univariate Analysis

What is the structure of your dataset?

数据集总共有113937条数据以及13个变量（包括BorrowerAPR、Term、Occupation等等），其中因子变量有Term、ProsperRating..Alpha.、IncomeRange、Occupation、EmploymentStatus和IsBorrowerHomeowner，对于ProsperRating..Alpha.，由好至坏可分为AA、A、B、C、D、E、HR七个等级。

其他观察：
1.借款人中除去填了其他的，专业人士比重最多。
2.CreditScore的中位数是689.5。
3.StatedMonthlyIncome主要集中在2500-5000美元。
4.BorrowerAPR和CreditScore都接近正态分布，即大多数人都有着平均水平的借款利率和信用评分，极端情况较少。

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

本项目主要研究对象为BorrowerAPR，旨在探索影响BorrowerAPR的因素，并以CreditScore等变量构建BorrowerAPR的预测模型。

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

Term、CreditScore、EmploymentStatus、IsBorrowerHomeowner、DebtToIncomeRatio、IncomeRange、DelinquenciesLast7Years、ProsperRating..Alpha.可能对BorrowerAPR具有影响，预计CreditScore、IncomeRange和ProsperRating..Alpha影响最大。

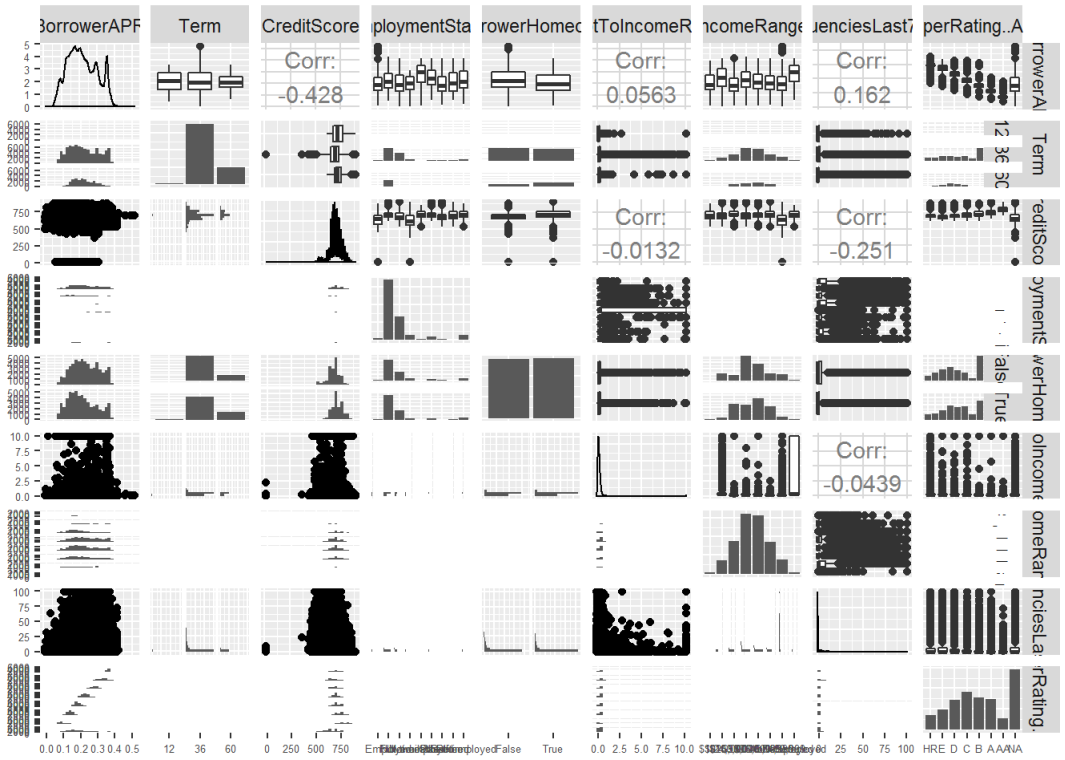
Did you create any new variables from existing variables in the dataset?

我创建了一个名为CreditScore的新变量，它是对应CreditScoreRangeLower与CreditScoreRangeUpper的平均数，以便能更为方便地对数据集进行探讨。

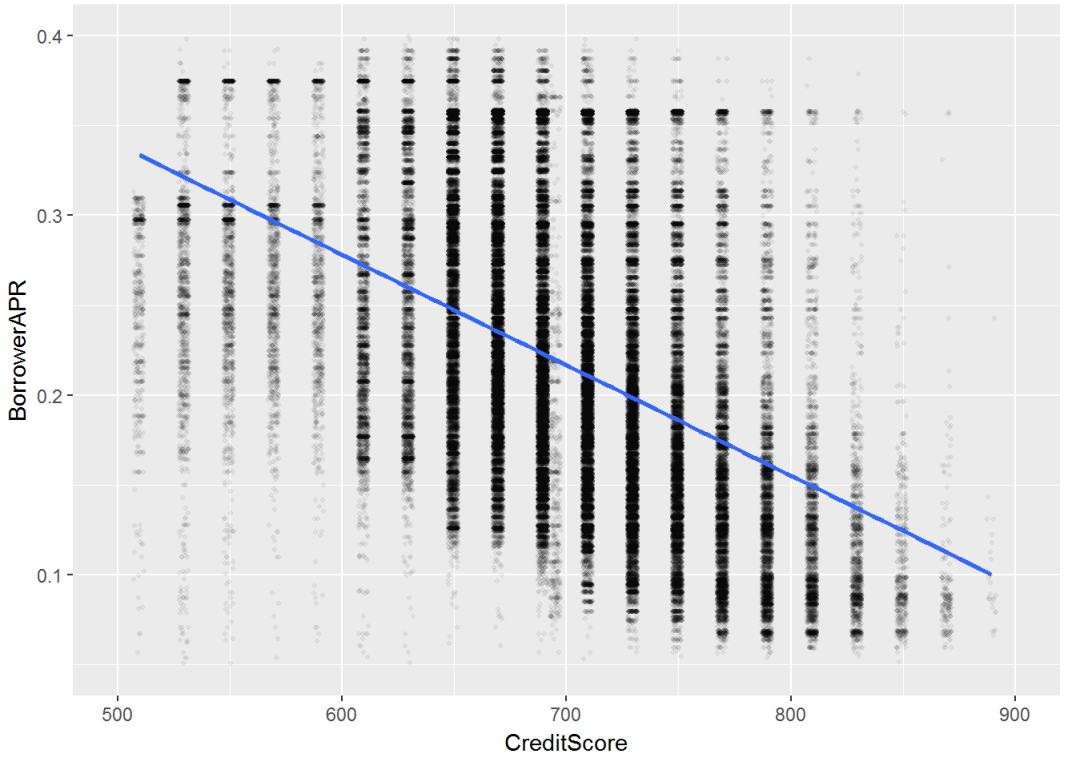
Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

1.LoanOriginalAmount呈正偏态分布，于是对其进行了对数转换，转换后呈多峰分布，峰值分别约为4000、10000和15000
2.BorrowerAPR和CreditScore存在缺失值，因为它们是研究的重要变量，故对其进行填充，二者的分布都接近正态，因此用均值进行填充。

Bivariate Plots Section



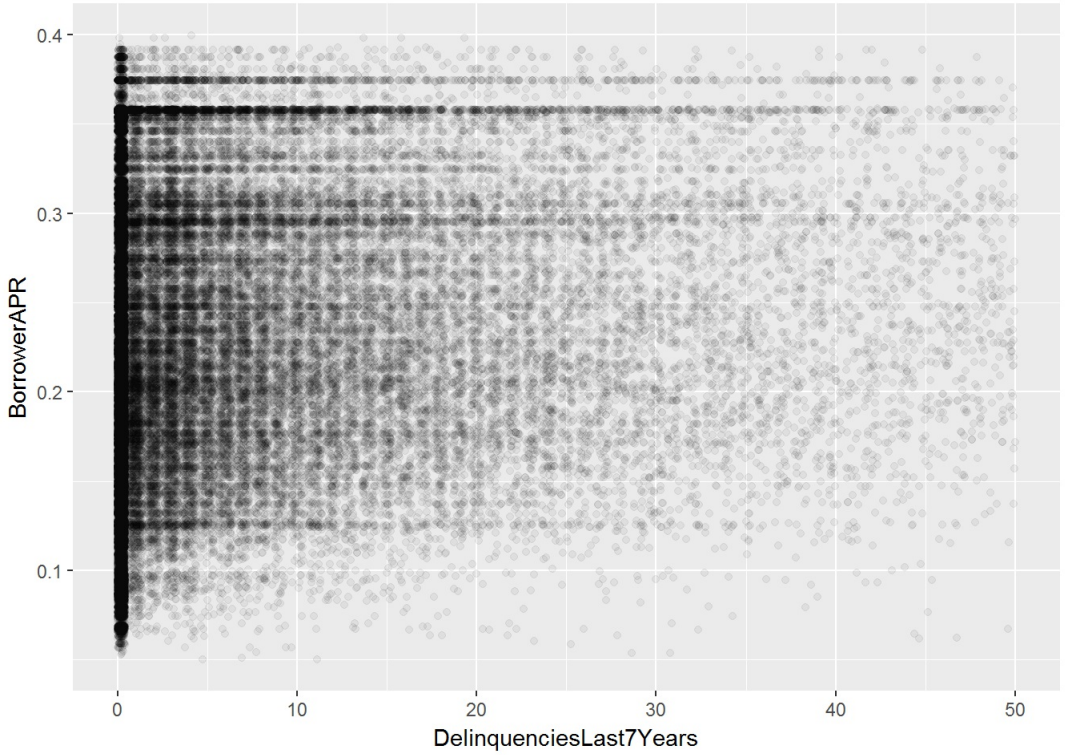
选取将探索的变量，先大致查看变量间关系，发现CreditScore和ProsperRating(Alpha)对BorrowerAPR相对影响大些，而 DelinquenciesLast7Years和DebtToIncomeRatio则影响很小。



首先观察CreditScore和BorrowerAPR的散点图，可以发现呈负相关

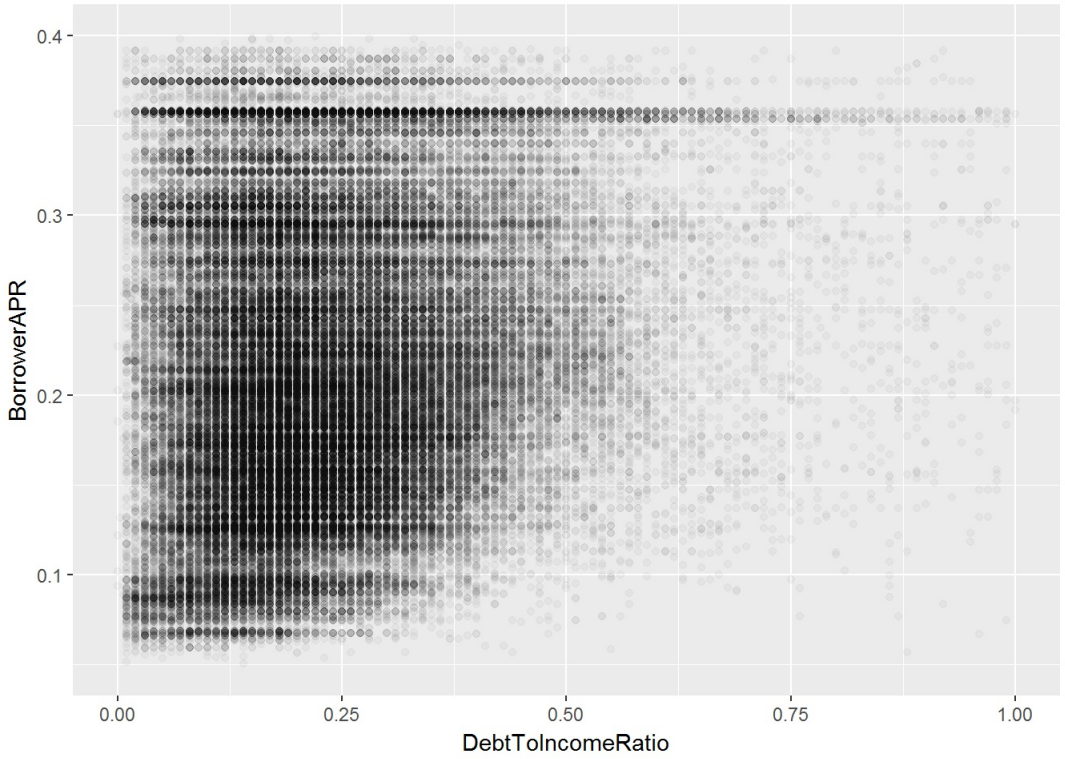
```
##
##  Pearson's product-moment correlation
##
## data:  pp$BorrowerAPR and pp$CreditScore
## t = -159.95, df = 113940, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4329459 -0.4234622
## sample estimates:
##        cor
## -0.4282158
```

由相关系数看出二者确实呈现一定程度的负相关。



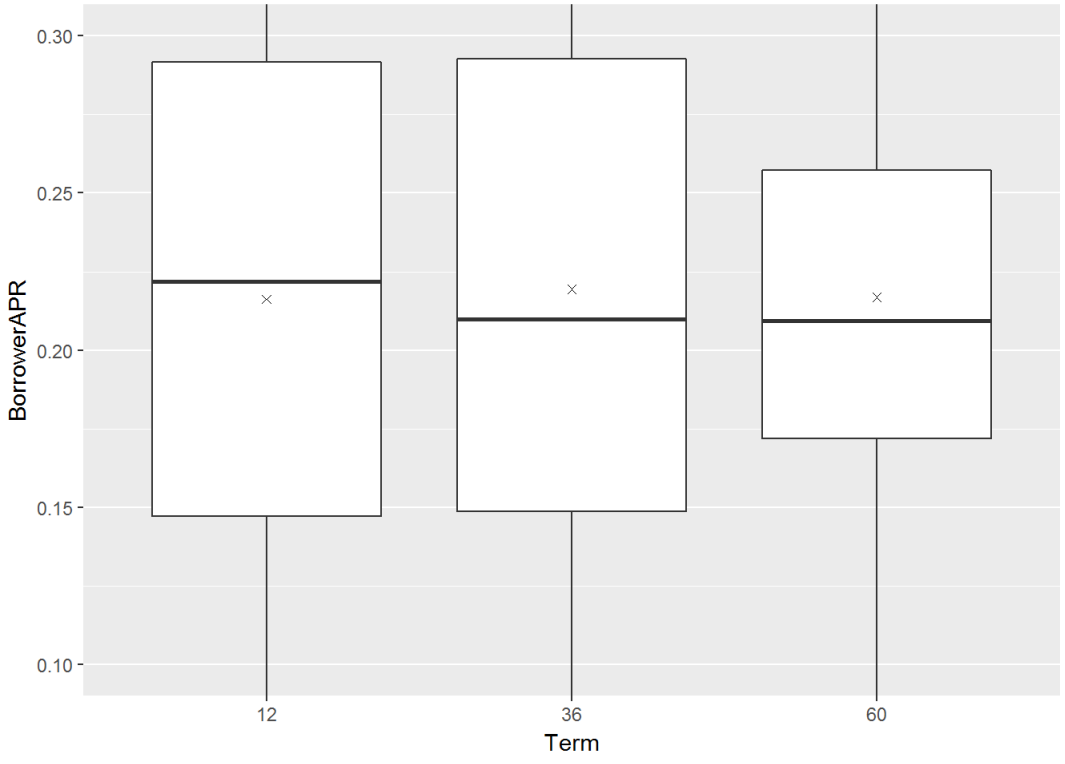

```
##
##  Pearson's product-moment correlation
##
## data:  pp$BorrowerAPR and pp$DelinquenciesLast7Years
## t = 55.251, df = 112940, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1565416 0.1678985
## sample estimates:
##       cor
## 0.1622254
```

由散点图和相关系数可以看出BorrowerAPR和DelinquenciesLast7Years相关程度不高。



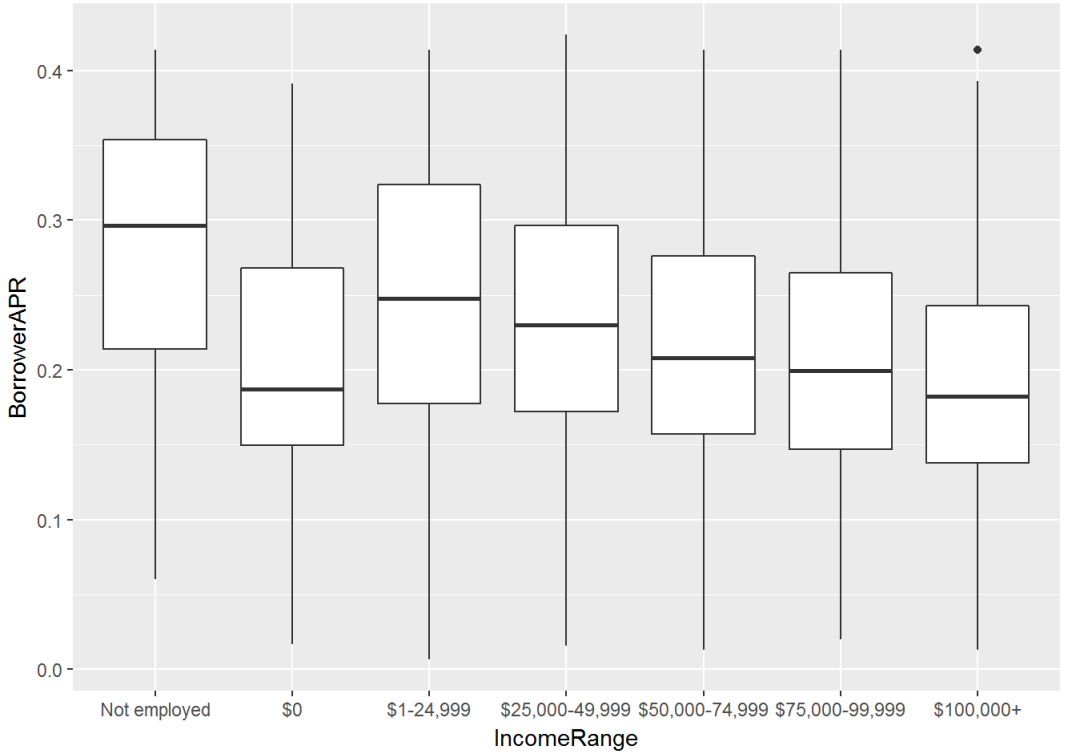

```
##
##  Pearson's product-moment correlation
##
## data:  pp$BorrowerAPR and pp$DebtToIncomeRatio
## t = 18.313, df = 105380, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.05030313 0.06234001
## sample estimates:
##       cor
## 0.05632362
```

同样由散点图和相关系数可以看出BorrowerAPR和DebtToIncomeRatio相关程度不高。



```
## pp$Term: 12
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.04935 0.14716 0.22189 0.21622 0.29167 0.35843
## ------------------------------------------------------------
## pp$Term: 36
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00653 0.14857 0.20979 0.21943 0.29265 0.51229
## ------------------------------------------------------------
## pp$Term: 60
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.07111 0.17184 0.20931 0.21684 0.25718 0.35838
```

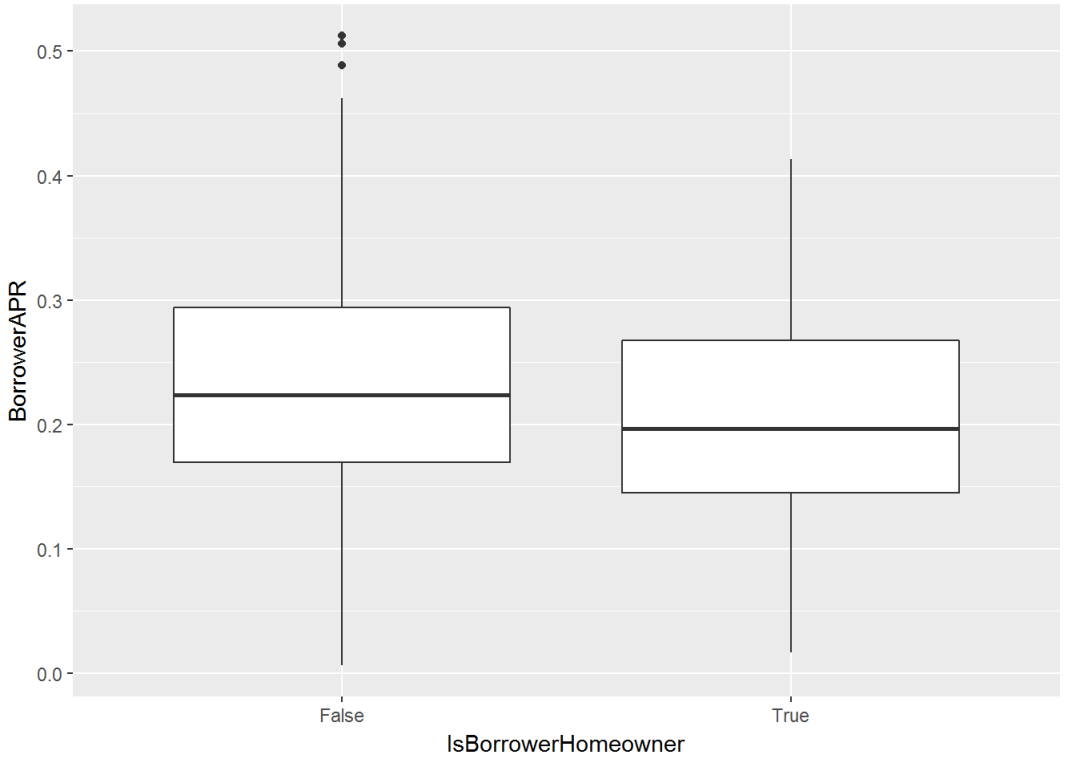
接下来观察不同期限下借款利率的箱线图以及描述统计，发现均值和中位数非常接近，看起来期限对借款利率影响不大。



按顺序对`IncomeRange`进行排序，并制作不同收入范围内借款利率的箱线图，除去收入填了`0`的情况外，就业的比未就业的利率低，收入高的比收入低的利率低。

```
## pp$IncomeRange: Not employed
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.06033 0.21372 0.29673 0.27921 0.35356 0.41355
## ------------------------------------------------------------
## pp$IncomeRange: $0
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01657 0.14960 0.18726 0.21035 0.26799 0.39153
## ------------------------------------------------------------
## pp$IncomeRange: $1-24,999
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00653 0.17730 0.24758 0.24797 0.32378 0.41355
## ------------------------------------------------------------
## pp$IncomeRange: $25,000-49,999
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01548 0.17192 0.22994 0.23494 0.29660 0.42395
## ------------------------------------------------------------
## pp$IncomeRange: $50,000-74,999
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01315 0.15713 0.20808 0.21774 0.27637 0.41355
## ------------------------------------------------------------
## pp$IncomeRange: $75,000-99,999
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01987 0.14714 0.19957 0.20807 0.26487 0.41355
## ------------------------------------------------------------
## pp$IncomeRange: $100,000+
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01315 0.13799 0.18224 0.19563 0.24282 0.41355
```

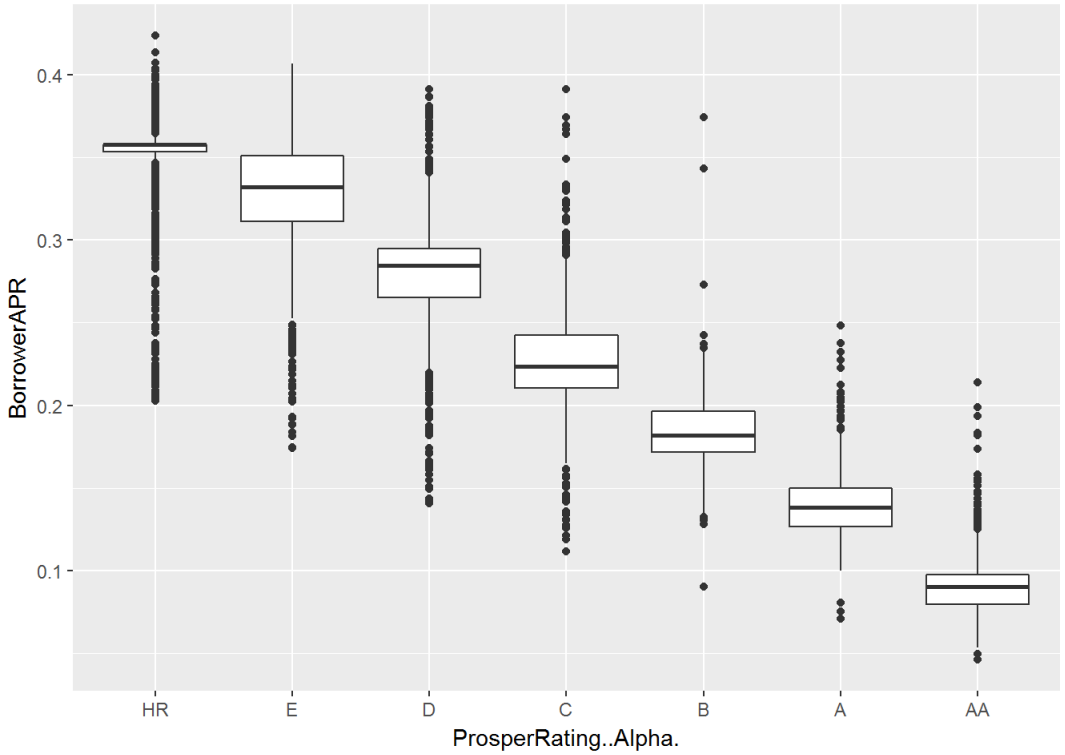
描述性统计同样能够印证。



总的来说，有房者比无房者利率低。

```
## pp$IsBorrowerHomeowner: False
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00653 0.16988 0.22362 0.22960 0.29371 0.51229
## ------------------------------------------------------------
## pp$IsBorrowerHomeowner: True
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01647 0.14494 0.19645 0.20825 0.26762 0.41355
```

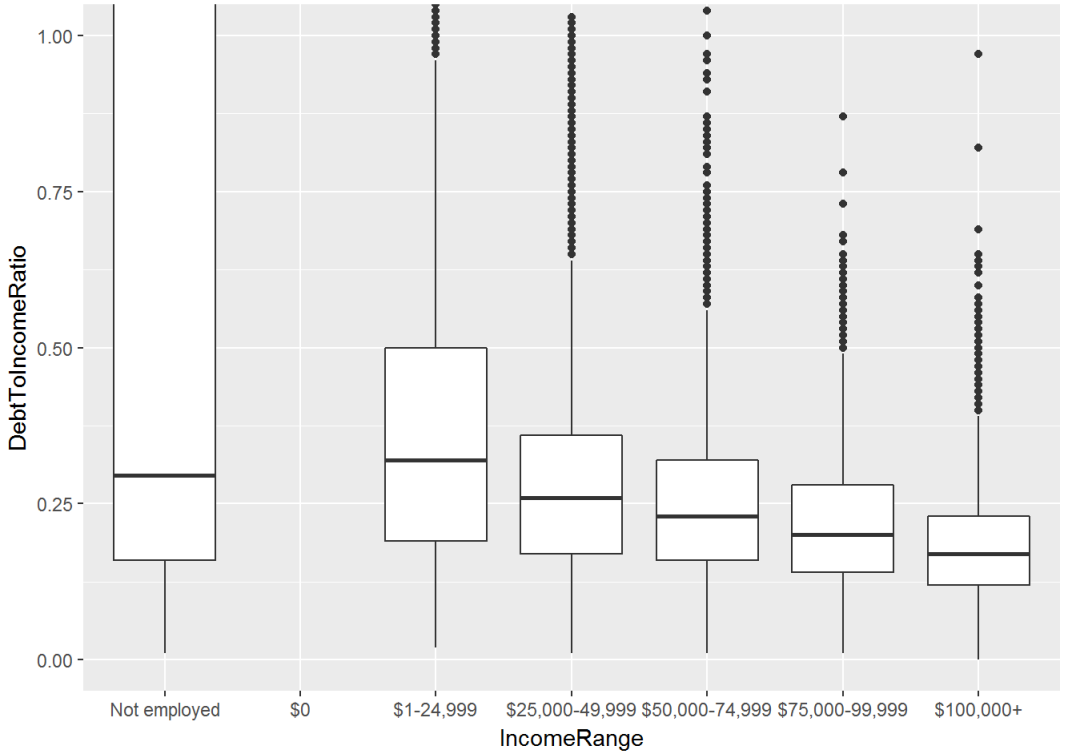
描述性统计同样能够印证。



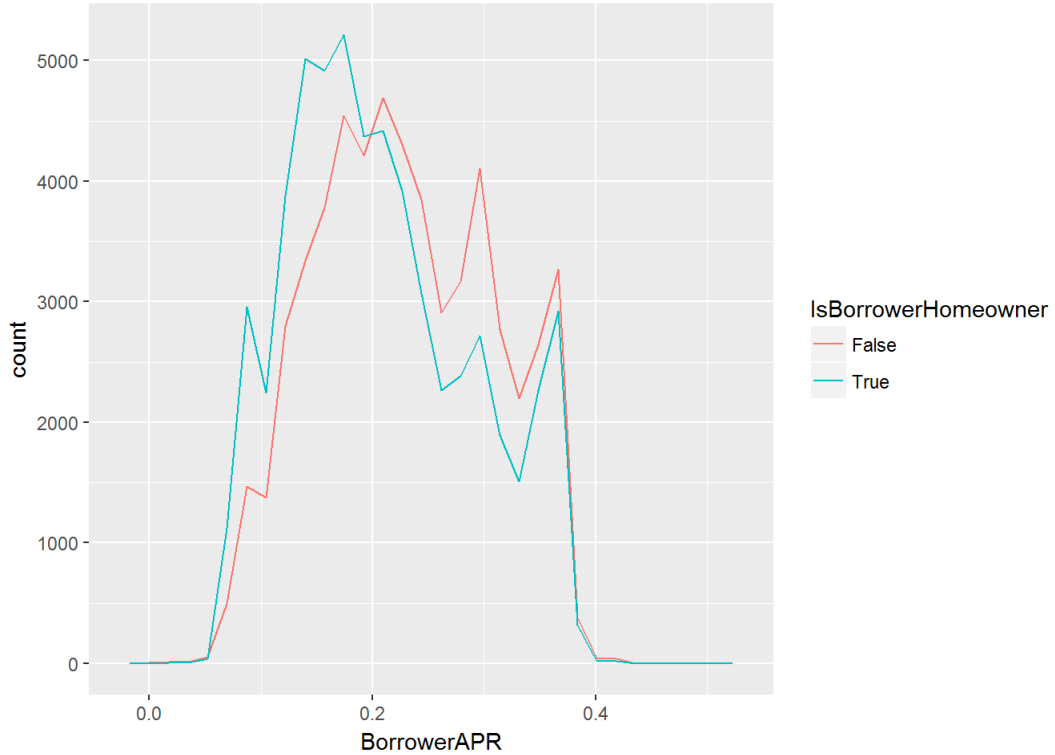
ProsperRating越高，借款利率明显越低。

```
## pp$ProsperRating..Alpha.: HR
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2026  0.3536  0.3580  0.3561  0.3580  0.4239
## ------------------------------------------------------------
## pp$ProsperRating..Alpha.: E
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1743  0.3116  0.3322  0.3306  0.3513  0.4068
## ------------------------------------------------------------
## pp$ProsperRating..Alpha.: D
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1406  0.2653  0.2849  0.2806  0.2951  0.3915
## ------------------------------------------------------------
## pp$ProsperRating..Alpha.: C
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1115  0.2102  0.2236  0.2261  0.2425  0.3915
## ------------------------------------------------------------
## pp$ProsperRating..Alpha.: B
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.08999 0.17151 0.18173 0.18403 0.19645 0.37453
## ------------------------------------------------------------
## pp$ProsperRating..Alpha.: A
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.07045 0.12626 0.13799 0.13891 0.14965 0.24807
## ------------------------------------------------------------
## pp$ProsperRating..Alpha.: AA
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.04583 0.07922 0.09000 0.09004 0.09736 0.21368
```

描述性统计同样能够印证。



可以看到收入为0的不存在收入负债比，就业者中收入越高，收入负债比的中位数越低，而未就业者的收入负债比的中位数和低收入人群相近。




```
## False  True
## 56459 57478
```

有房和无房人士的数量相近，但有房人士利率小于0.2的比重要高于无房人士，故平均来说利率低于无房人士，同时可以观察到当利率大于0.36时，二者的比重相近。

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?

1.CreditScore和BorrowerAPR呈现一定负相关。
2.BorrowerAPR和DelinquenciesLast7Years相关程度不高。
3.BorrowerAPR和DebtToIncomeRatio相关程度不高。
4.除去收入填了0的情况外，就业的比未就业的利率低，收入高的比收入低的利率低。
5.ProsperRating越高，借款利率明显越低。
6.当借款利率小于0.2时，借款人多为有房人士，当利率大于0.36时，二者的比重相近。

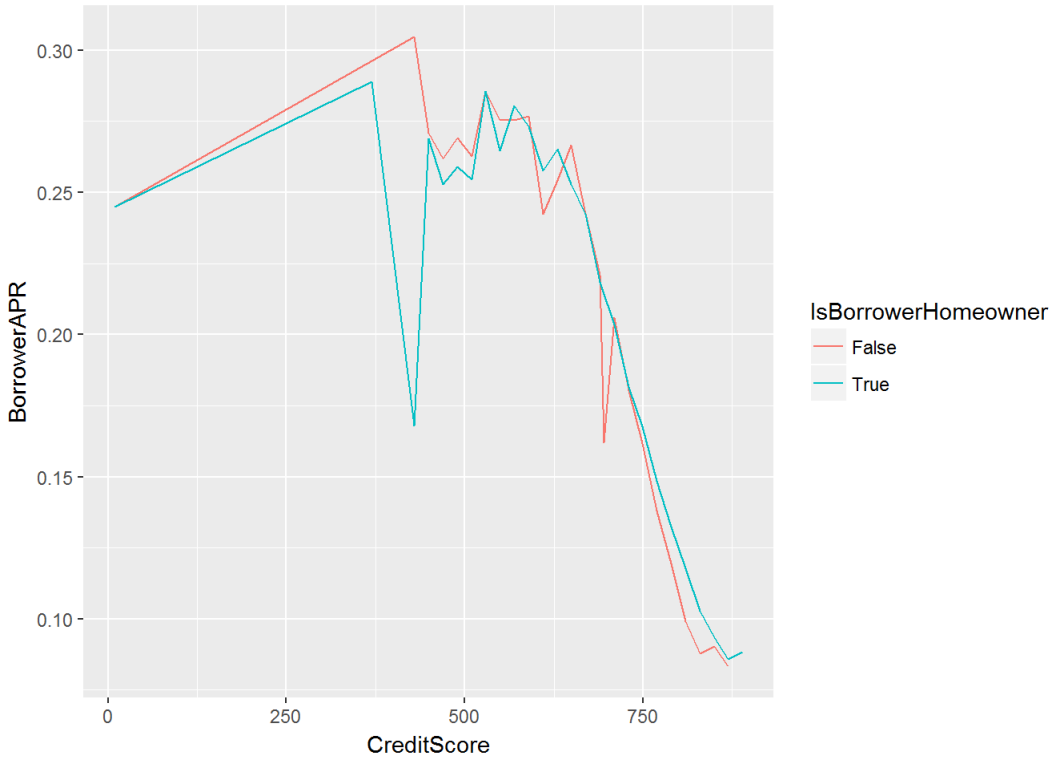
Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

收入为0的不存在收入负债比，就业者中收入越高，收入负债比的中位数越低，而未就业者的收入负债比的中位数和低收入人群相近。

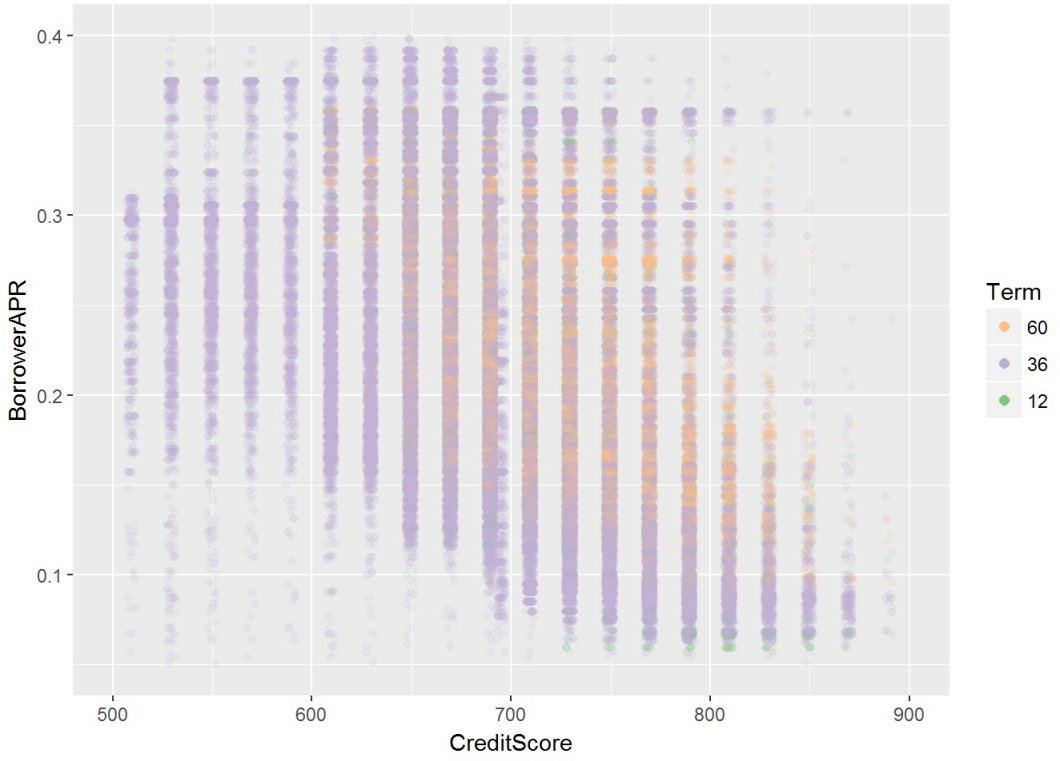
What was the strongest relationship you found?

CreditScore和ProsperRating对借款利率的影响最大。

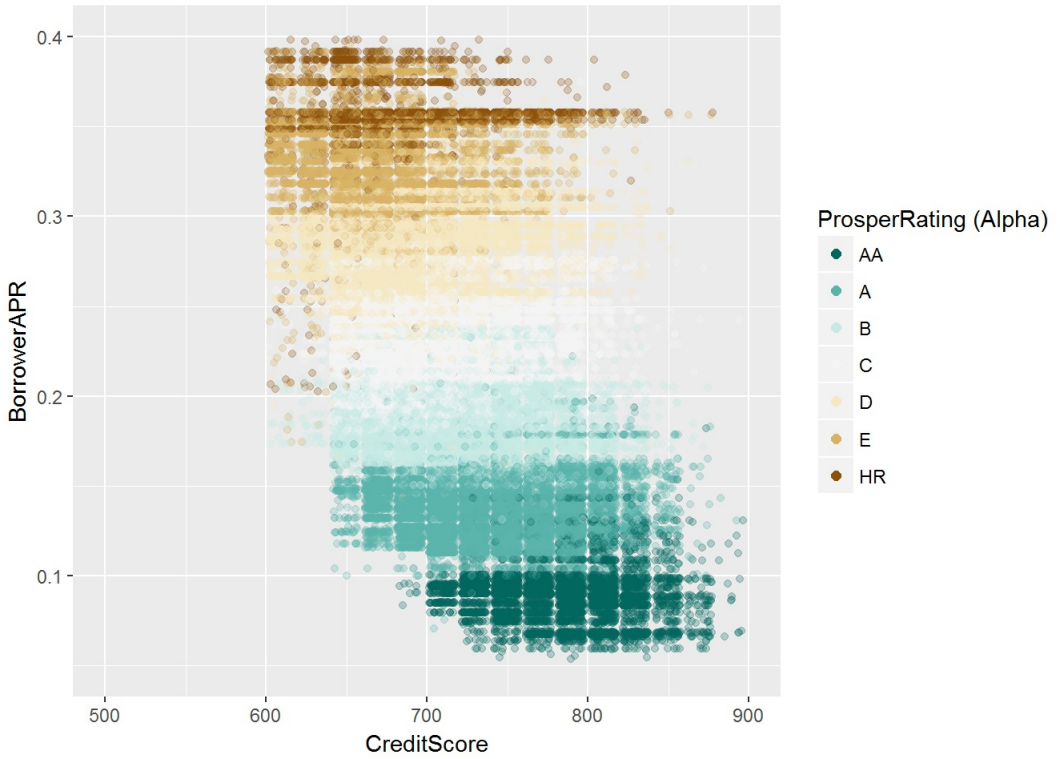
Multivariate Plots Section



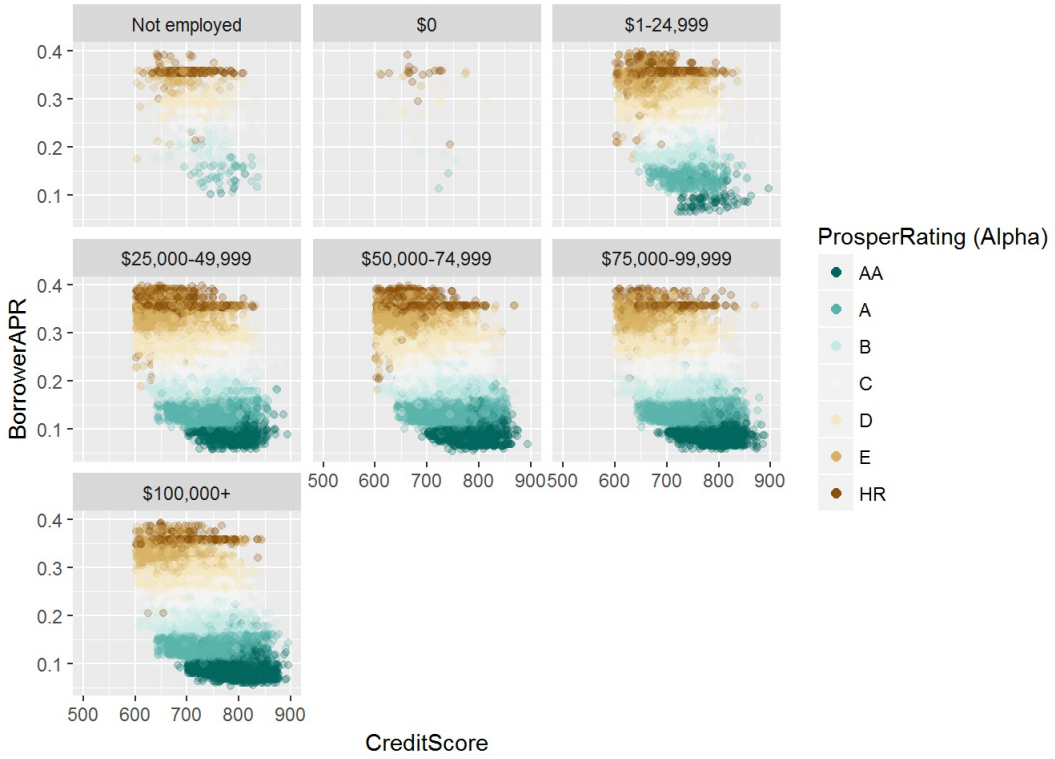
在CreditScore和BorrowerAPR关系图中加入了IsBorrowerHomeowner以作进一步区分，发现当CreditScore小于520时，无房者的利率中位数皆高于有房者，且随着CreditScore的增大有大幅波动；当CreditScore大于520时，二者的中位数高低交替，且大体上都随着CreditScore的增大而下降。



查看不同借款期限的CreditScore和BorrowerAPR的关系，可以看出借款12个月的比重很小，且期限对借款利率的影响也不明显



查看不同ProsperRating下CreditScore和BorrowerAPR的关系，可以明显发现同一CreditScore水平下，ProsperRating越好，借款利率越低。



以收入范围作分面，查看不同`ProsperRating`下`CreditScore`和`BorrowerAPR`的关系，可以发现除了收入为`0`的群体样本过少外，其余情况下越高的`ProperRating`都有着越低的借款利率。

```
##
## Calls:
## m1: lm(formula = BorrowerAPR ~ ProsperRating..Alpha., data = pp)
## m2: lm(formula = BorrowerAPR ~ ProsperRating..Alpha. + IsBorrowerHomeowner,
##     data = pp)
## m3: lm(formula = BorrowerAPR ~ ProsperRating..Alpha. + IsBorrowerHomeowner +
##     IncomeRange, data = pp)
## m4: lm(formula = BorrowerAPR ~ ProsperRating..Alpha. + IsBorrowerHomeowner +
##     IncomeRange + CreditScore, data = pp)
##
## ================================================================================================
##                                        m1             m2             m3             m4
## ------------------------------------------------------------------------------------------------
##   (Intercept)                        0.229***       0.230***       0.231***       0.214***
##                                     (0.000)        (0.000)        (0.000)        (0.001)
##   ProsperRating..Alpha.: .L         -0.241***      -0.241***      -0.241***      -0.243***
##                                     (0.000)        (0.000)        (0.000)        (0.000)
##   ProsperRating..Alpha.: .Q         -0.007***      -0.007***      -0.008***      -0.008***
##                                     (0.000)        (0.000)        (0.000)        (0.000)
##   ProsperRating..Alpha.: .C          0.009***       0.009***       0.009***       0.009***
##                                     (0.000)        (0.000)        (0.000)        (0.000)
##   ProsperRating..Alpha.: ^4         -0.010***      -0.010***      -0.010***      -0.010***
##                                     (0.000)        (0.000)        (0.000)        (0.000)
##   ProsperRating..Alpha.: ^5          0.002***       0.002***       0.002***       0.002***
##                                     (0.000)        (0.000)        (0.000)        (0.000)
##   ProsperRating..Alpha.: ^6          0.003***       0.003***       0.002***       0.002***
##                                     (0.000)        (0.000)        (0.000)        (0.000)
##   IsBorrowerHomeowner: True/False                  -0.000          0.000**       -0.000
##                                                    (0.000)        (0.000)        (0.000)
##   IncomeRange: .L                                                -0.007***      -0.006***
##                                                                  (0.001)        (0.001)
##   IncomeRange: .Q                                                 0.005***       0.004***
##                                                                  (0.000)        (0.000)
##   IncomeRange: .C                                                -0.003*        -0.003*
##                                                                  (0.001)        (0.001)
##   IncomeRange: ^4                                                 0.004*         0.004*
##                                                                  (0.002)        (0.002)
##   IncomeRange: ^5                                                -0.003*        -0.003*
##                                                                  (0.001)        (0.001)
##   IncomeRange: ^6                                                 0.001*         0.001*
##                                                                  (0.001)        (0.001)
##   CreditScore                                                                    0.000***
##                                                                                 (0.000)
## ------------------------------------------------------------------------------------------------
##   R-squared                           0.930          0.930          0.930          0.930
##   adj. R-squared                      0.930          0.930          0.930          0.930
##   sigma                               0.021          0.021          0.021          0.021
##   F                              187876.253     161035.242      87040.316      80971.994
##   p                                   0.000          0.000          0.000          0.000
##   Log-likelihood                 206827.945     206828.031     206980.280     207053.365
##   Deviance                           37.933         37.933         37.797         37.732
##   AIC                           -413639.890    -413638.062    -413930.559    -414074.731
##   BIC                           -413565.101    -413553.924    -413790.329    -413925.152
##   N                               84853          84853          84853          84853
## ================================================================================================
```

这里选取ProsperRating..Alpha.，IsBorrowerHomeowner、IncomeRange和CreditScore来构建线性模型，并对BorrowerAPR的变化有93%的解释，说明能够较好地用来预测借款利率。

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?

1.当CreditScore小于520时，无房者的利率中位数皆高于有房者，且随着CreditScore的增大有大幅波动；当CreditScore大于520时，二者的中位数高低交替，且大体上都随着CreditScore的增大而下降。
2.同一CreditScore水平下，ProsperRating越好，借款利率越低。
3.ProperRating越高，基本上借款利率也越低。

Were there any interesting or surprising interactions between features?

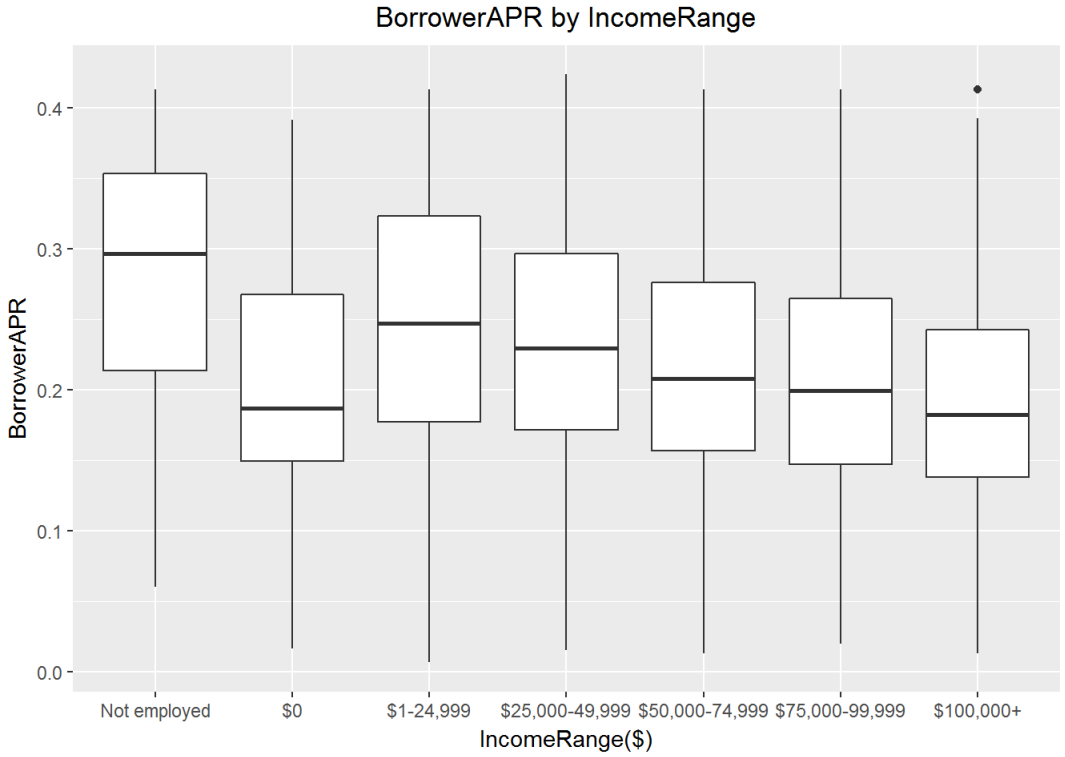
1.当CreditScore小于520时，无论有房还是无房者的借款利率都会有明显上升的过程，而预期是即使上升也不会这么明显
2.期限对借款利率的影响不明显，而当初预想的是期限越长，利率会越高。

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

我选取了ProsperRating..Alpha.，IsBorrowerHomeowner、IncomeRange和CreditScore来构建线性模型，亮点是R2为93%，说明能很好地解释利率的变化，缺点是ProsperRating..Alpha.之外的变量对模型的优化没有贡献。

Final Plots and Summary

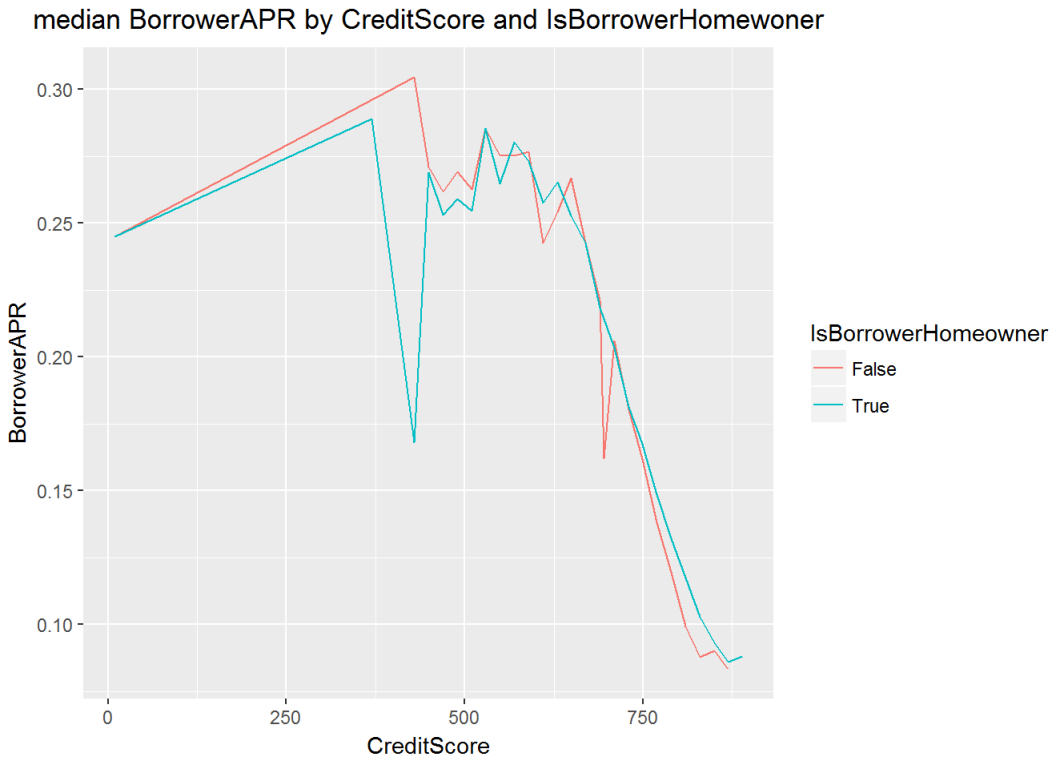
Plot One



Description One

可以看到就业的比未就业的利率低，收入高的比收入低的利率低，但是存在一个特例，即收入为0的情况，利率水平竟然和收入在10万美元以上的群体相当，这是个奇怪的现象，可能的解释是收入为0的记录很少，而刚好存在一部分低利率人群拉低了中位数，即样本量不足使得结果不具备代表性。

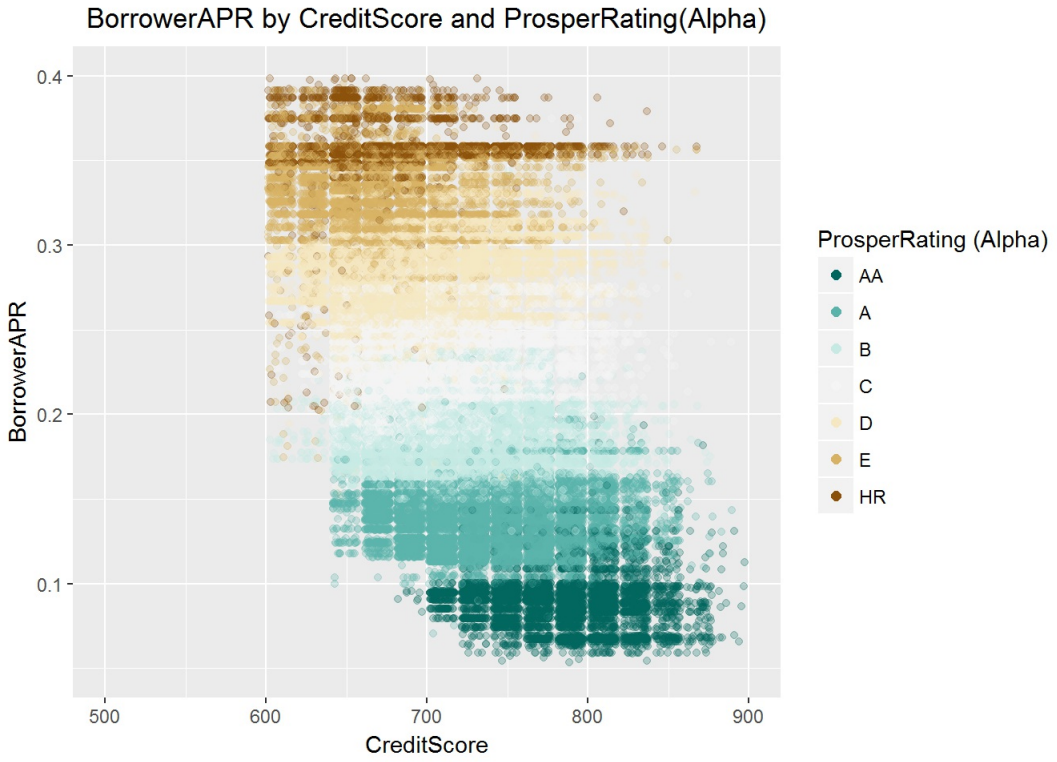
Plot Two



Description Two

当CreditScore小于520时，无房者的利率中位数皆高于有房者，且随着CreditScore的增大有大幅波动；当CreditScore大于520时，二者的中位数高低交替，且大体上都随着CreditScore的增大而下降。值得注意的是无论有房还是无房者的折线一开始都有明显的上升，意味着借款利率随着信用评分的增大而提高，这不符合常理，可能的原因同样是信用评分低的群体相对很少，样本量不足使得结果不具备代表性。

Plot Three



Description Three

这个图说明了在同一CreditScore水平下，ProsperRating越好，借款利率就越低，因为层次感异常明显，也直观地反映出为什么ProperRating能够如此好地解释借款利率的变动。

Reflection

项目中我依次进行了单变量、双变量和多变量分析，作图包含直方图、频数多边形、箱线图、散点图和折线图，主要探索了其余变量对贷款利率的影响，其中影响最为明显的是ProsperRating（Alpha），于是将其作为主导变量以构建线性模型，并取得了93%的R2，可以说模型具有较好的预测能力。
让我意外的是Term、DelinquenciesLast7Years和DebtToIncomeRatio对利率影响不大，所以他们只能作为参考而非主导因素。另外，收入为0的群体利率水平竟然和收入在10万美元以上的群体相当，这一开始让我困惑，后来发现收入为0的群体相对来说很少，不具备普遍代表性。
为了分析能够顺利进展，我对两个变量做了改动，一是将Term设置为因子变量以便进行分类讨论，二是将CreditScoreRangeLower和CreditScoreRangeUpper合并为CreditScore，这样使得探索更加便利。
在探索过程中也遇到了一些难题，例如作为主要探索因素的BorrowerAPR和CreditScore存在缺失值，后来观察分布后决定用平均值来填充。此外，在构建模型时，我发现基本上做贡献的只有ProsperRating..Alpha.，而其余因素在图形中又能明显解释差异，这让我非常困惑，这方面有赖于进一步学习统计知识来作解释。
在未来工作中，可以尝试对变量进行转换后再用以构建模型，说不定能取得意想不到的结果，EDA这个环节本身就是不断探索并不断优化的过程，所以一些创造性的想法也非常有必要！