Project

Author: Qi Meng

Last modified date: 2018-03-01

==

Loan Data from Prosper

This data set contains 113,937 loans with 81 variables on each loan, including loan amount, borrower rate (or interest rate), current loan status, borrower income, borrower employment status, borrower credit history, and the latest payment information.

The diamention of Proserper Loan Dataset

```
dim(pr)

## [1] 113937     81
```

In total there are 81 variables that corresponding to each loan In order better understand the dataset, in this analysis, 19 variables will be selected.

```
pr <- subset(pr, select = c('LoanStatus',
                                'BorrowerAPR',
                                'BorrowerRate',
                                'LenderYield',
                                'ProsperScore',
                                'BorrowerState',
                                'Occupation',
                                'EmploymentStatus',
                                'IsBorrowerHomeowner',
                                'TotalCreditLinespast7years',
                                'TotalInquiries',
                                'BankcardUtilization',
                                'AvailableBankcardCredit',
                                'IncomeRange',
                                'IncomeVerifiable',
                                'LoanOriginalAmount',
                                'LoanOriginationDate',
                                'MonthlyLoanPayment',
                                'Investors'
))

str(pr)

## 'data.frame':    113937 obs. of  19 variables:
##  $ LoanStatus                : Factor w/ 12 levels
"Cancelled","Chargedoff",..: 3 4 3 4 4 4 4 4 4 4 ...
```

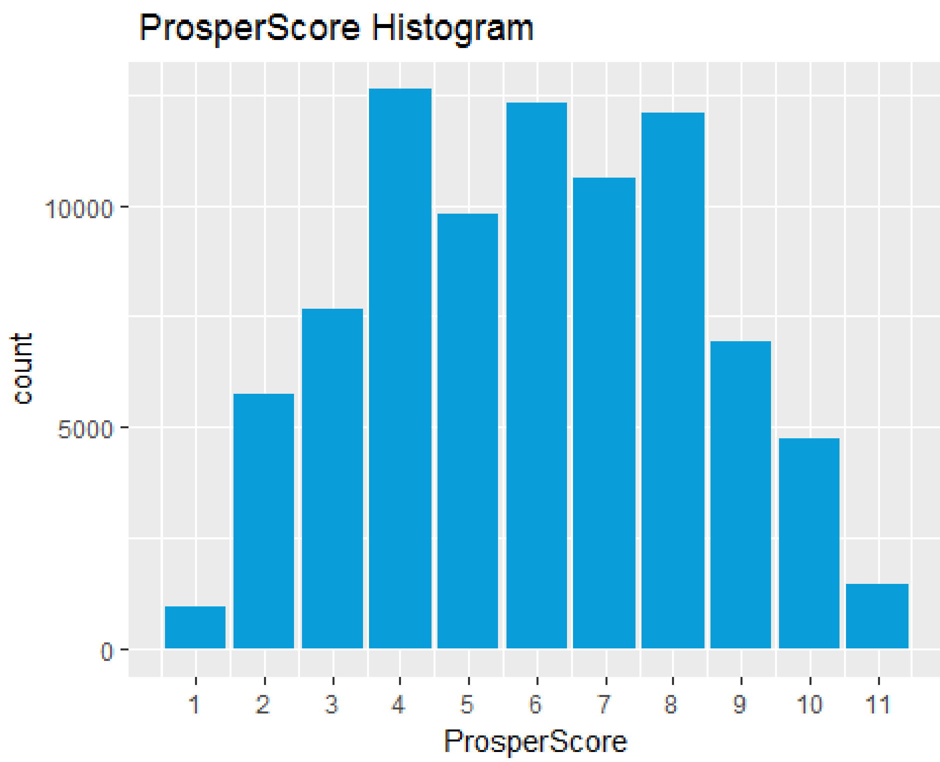
```
##  $ BorrowerAPR                : num  0.165 0.12 0.283 0.125 0.246 ...
##  $ BorrowerRate               : num  0.158 0.092 0.275 0.0974 0.2085 ...
##  $ LenderYield                : num  0.138 0.082 0.24 0.0874 0.1985 ...
##  $ ProsperScore               : num  NA 7 NA 9 4 10 2 4 9 11 ...
##  $ BorrowerState              : Factor w/ 52 levels "","AK","AL","AR",..: 7
7 12 12 25 34 18 6 16 16 ...
##  $ 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 ...
##  $ TotalCreditLinespast7years: int  12 29 3 29 49 49 20 10 32 32 ...
##  $ TotalInquiries             : num  3 5 1 1 9 2 0 16 6 6 ...
##  $ BankcardUtilization        : num  0 0.21 NA 0.04 0.81 0.39 0.72 0.13
0.11 0.11 ...
##  $ AvailableBankcardCredit    : num  1500 10266 NA 30754 695 ...
##  $ IncomeRange                : Factor w/ 8 levels "$0","$1-24,999",..: 4 5
7 4 3 3 4 4 4 4 ...
##  $ IncomeVerifiable           : Factor w/ 2 levels "False","True": 2 2 2 2
2 2 2 2 2 2 ...
##  $ LoanOriginalAmount         : int  9425 10000 3001 10000 15000 15000 3000
10000 10000 10000 ...
##  $ LoanOriginationDate        : Factor w/ 1873 levels "2005-11-15
00:00:00",..: 426 1866 260 1535 1757 1821 1649 1666 1813 1813 ...
##  $ MonthlyLoanPayment         : num  330 319 123 321 564 ...
##  $ Investors                  : int  258 1 41 158 20 1 1 1 1 1 ...
```

Univariate Plots Section

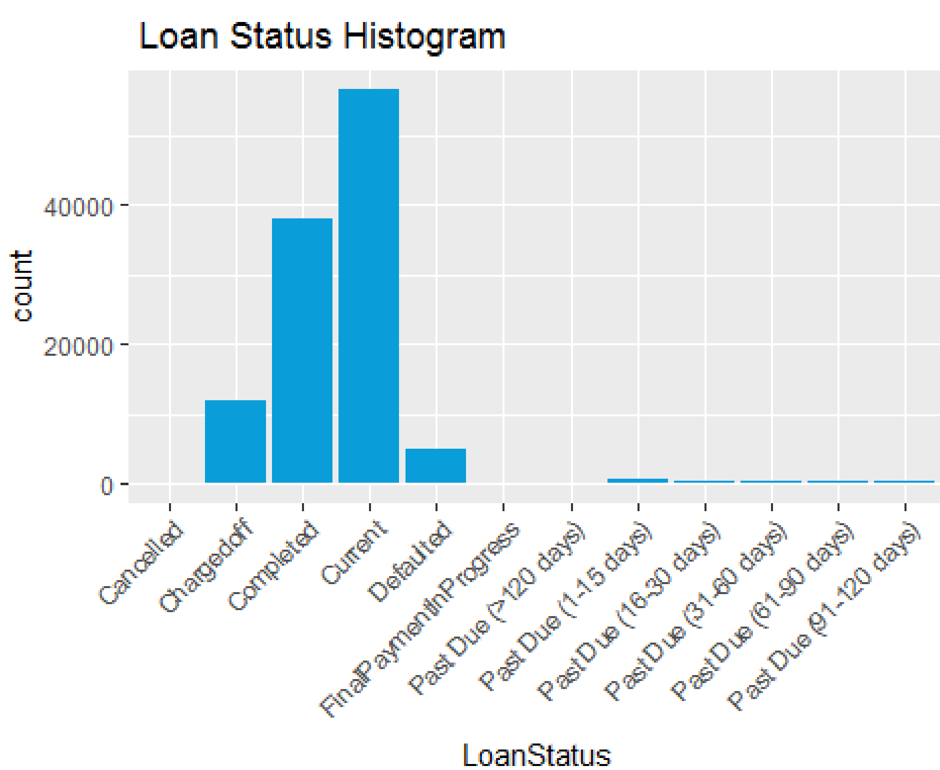
Prosper Score measures the loan applicant's risk level, the higer the score the lower the risk to lend.

```
summary(pr$ProsperScore)

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

table(pr$ProsperScore)

## 
##     1     2     3     4     5     6     7     8     9    10    11 
##   992  5766  7642 12595  9813 12278 10597 12053  6911  4750  1456
```

The Highest score is 11 and the score distribution would be better presented by the following chart.



From the histogram, the majority of the ProserScore are 4, 6, and 8

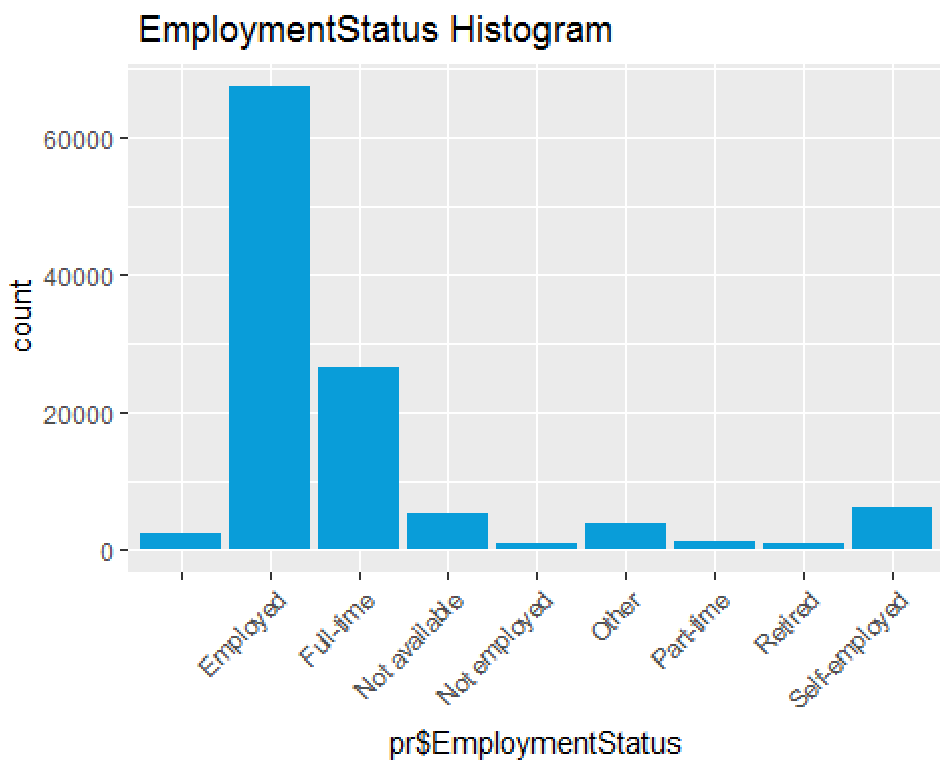


Most of the loan are current and the second is the completed loans

```r
summary(pr$LoanStatus)
```

```
##              Cancelled             Chargedoff              Completed
##                      5                  11992                  38074
##                Current              Defaulted FinalPaymentInProgress
##                  56576                   5018                    205
##   Past Due (>120 days)   Past Due (1-15 days)  Past Due (16-30 days)
##                     16                    806                    265
##  Past Due (31-60 days)  Past Due (61-90 days) Past Due (91-120 days)
##                    363                    313                    304
```

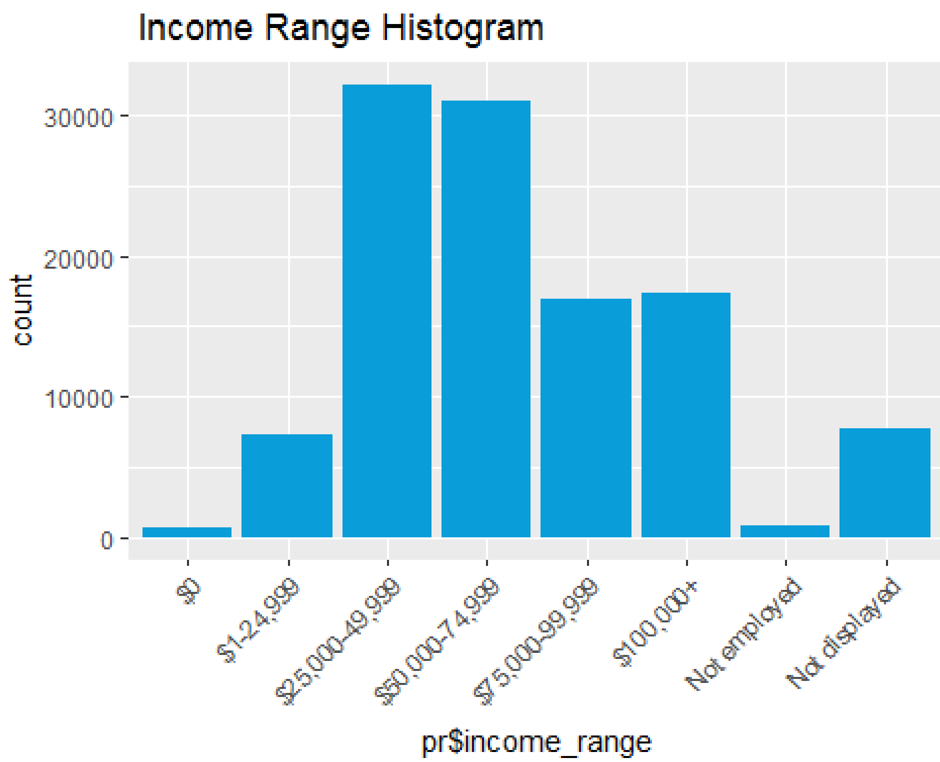
Creating a new variabel income_range to better format the IncomeRange variable



The loan data mostly coming from people who works employed or full time.

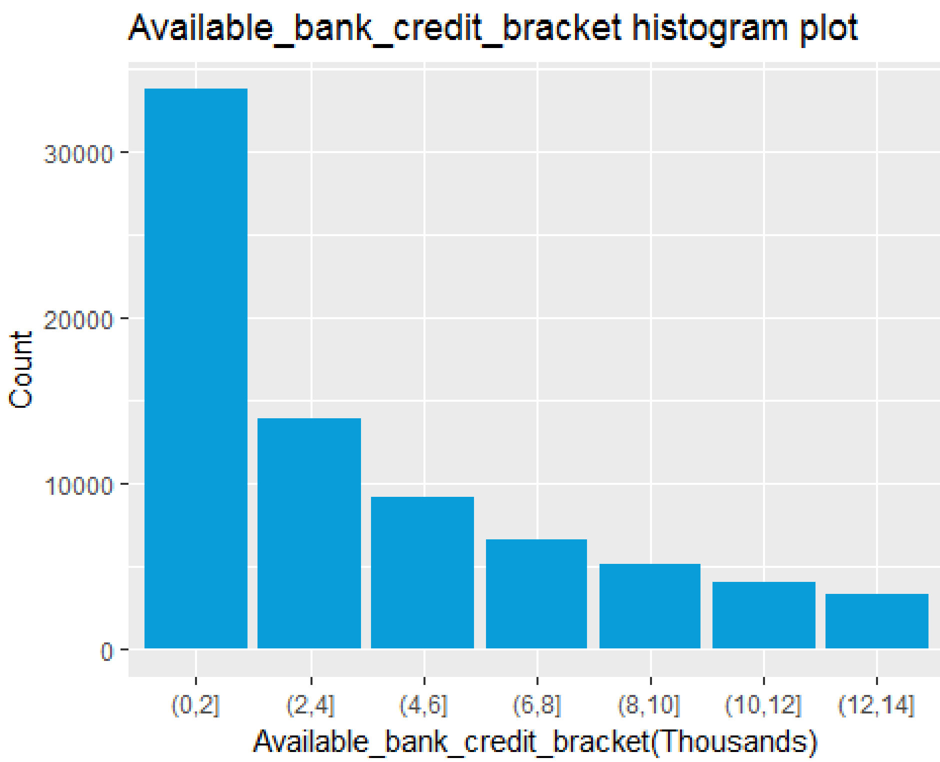
```r
summary(pr$EmploymentStatus)
```

```
##                    Employed     Full-time Not available  Not employed
##          2255         67322         26355          5347           835
##         Other     Part-time       Retired Self-employed
##          3806          1088           795          6134
```



People whose income is between $25,000 and $100,000 applied for the loans.

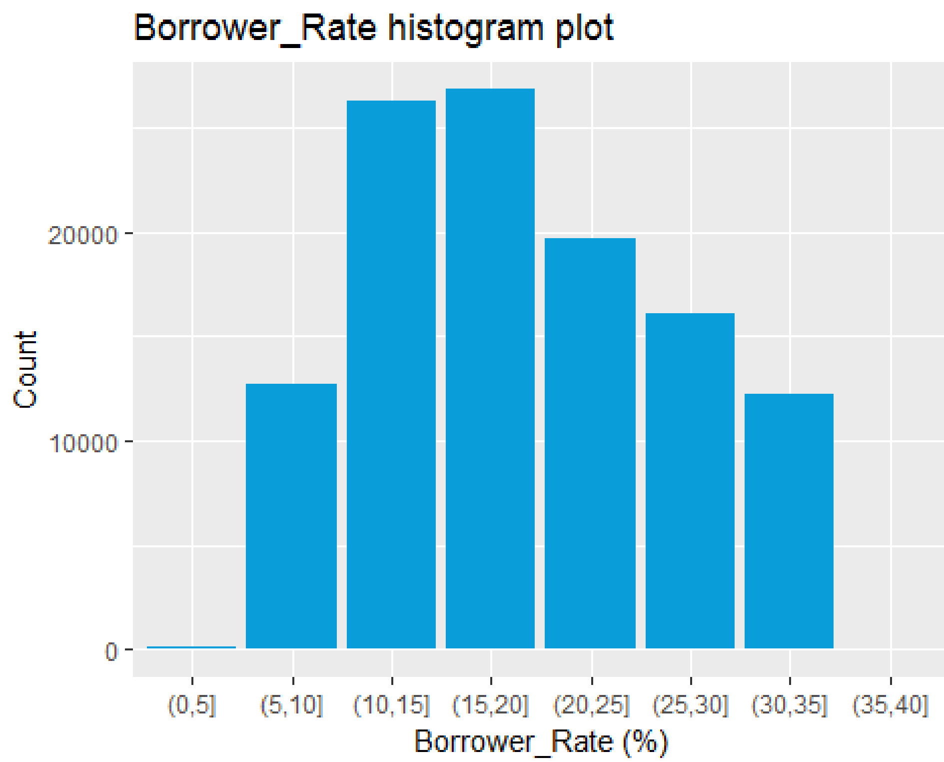
```
summary(pr$income_range)

##             $0      $1-24,999 $25,000-49,999 $50,000-74,999 $75,000-99,999 
##            621           7274          32192          31050          16916 
##       $100,000+   Not employed  Not displayed 
##          17337            806           7741
```



Based on the available bank credit breakdown, most of the borrowers have the credit less than 6,000.

```
summary(pr$available_bank_credit_bracket)

##   (0,2]   (2,4]   (4,6]   (6,8]  (8,10] (10,12] (12,14]    NA's 
##   33758   14002    9200    6676    5176    4055    3295   37775
```

Summary of available_bank_credit_bracket



Borrowers’ interest rates are between 0 tp 35%.

```
summary(pr$AvailableBankcardCredit)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0     880    4100   11210   13180  646300    7544
```

Summary of AvailableBankcardCredit

```
summary(pr$available_bank_credit_bracket)

##   (0,2]   (2,4]   (4,6]   (6,8]  (8,10] (10,12] (12,14]    NA's
##   33758   14002    9200    6676    5176    4055    3295   37775
```

Summary of available_bank_credit_bracket

Univariate Analysis

What is the structure of your dataset?

The ProsperLoan data have 113937 overvations and 19 variables ProsperSocre: A custom risk score built using historical Prosper data. The score ranges from 1-10, with 10 being the best, or lowest risk score.
Applicable for loans originated after July 2009. Loanstatus: Completed, Current, Past Due (1-15 days), Defaulted, Chargedoff, Past Due (16-30 days), Cancelled, Past Due (61-90 days), Past Due (31-60 days), Past Due (91-120 days)

EmploymentStatus:Self-employed, Employed, Not available, Full-time, Other, Not employed, Part-time, Retired

IncomeRange: $0, $1-24,999, $25,000-49,999, $50,000-74,999, $75,000-99,999, $100,000+, Not employed, Not displayed

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

The main feature is the ProsperSocre, which measue the risk ability of the loan itself, verus the BorrowerRate

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

The loan applicants' occupation,income, bankcard utilization, available bank card credit and other variables might impact the rick score when valued by th Prosper Company

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

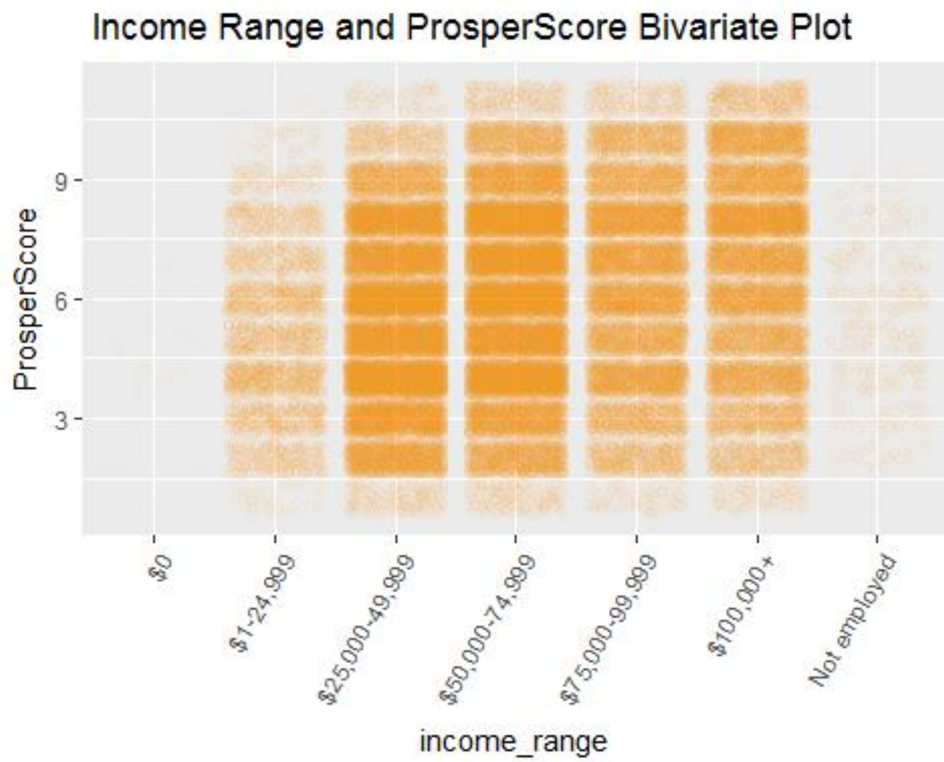
Yes, I created the new variable range_new to reorder the income range variable in a ascending order.

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?

Yes. I did select 19 variables out of 81 in total. The reason for this is that not all the variables are revelent in determining the Prosper Score.

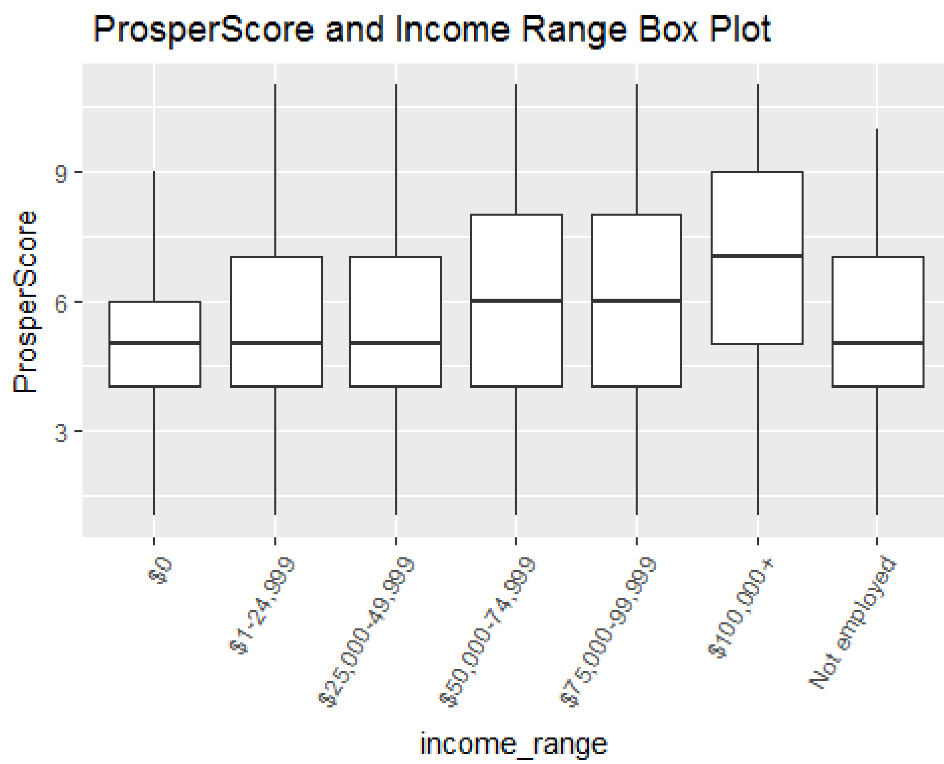
Bivariate Plots Section



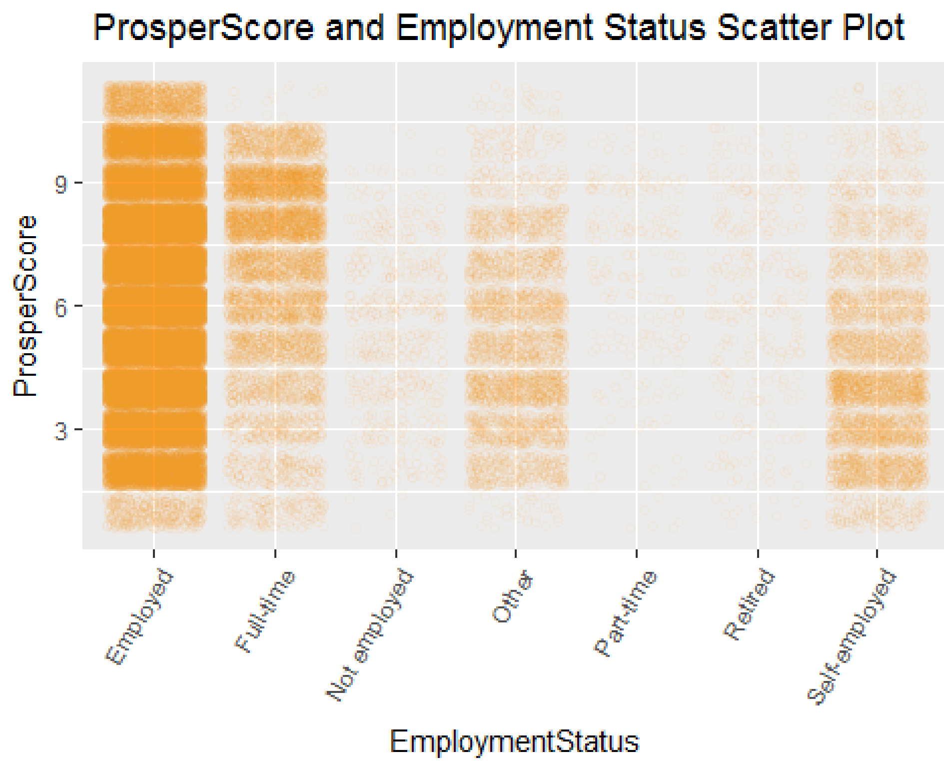
For applicants with lower income, the Prosper Score seems lower than 6, and for people with higher income, the ProsperScore seems higher, which means less risk.

```
tapply( pr$ProsperScore, pr$income_range,summary)

## $`$0`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     1.0     4.0     5.0     4.6     6.0     9.0     576
##
## $`$1-24,999`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   4.000   5.000   5.093   7.000  11.000    2620
##
## $`$25,000-49,999`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   4.000   5.000   5.424   7.000  11.000    8017
##
## $`$50,000-74,999`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   4.000   6.000   5.957   8.000  11.000    5423
##
## $`$75,000-99,999`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   4.000   6.000   6.297   8.000  11.000    2418
##
```

```
## $`$100,000+`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   5.000   7.000   6.738   9.000  11.000    2132
##
## $`Not employed`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   4.000   5.000   5.308   7.000  10.000     157
##
## $`Not displayed`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##      NA      NA      NA     NaN      NA      NA    7741
```



From the box plot we can see it more clearly, the median of people of income greater than $50,000 is much higher than people with income less than $50,000

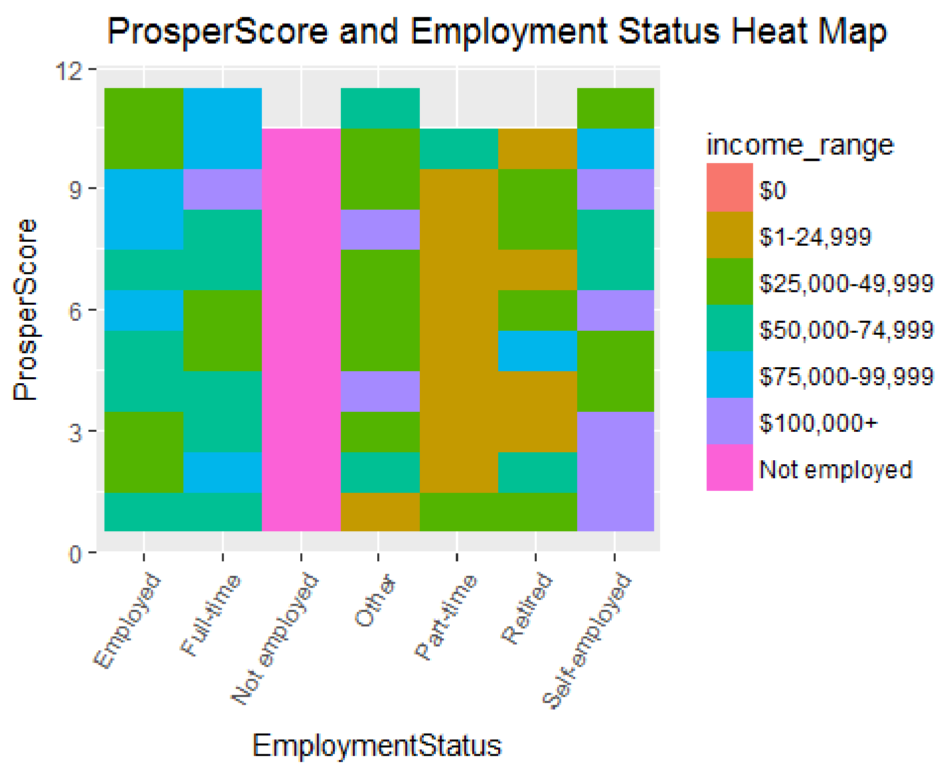


If we compare ProsperScore with employment status, clearly, full time employees will have greater ProperScore, and therefore, less risky to lend money to them

```
tapply( pr$ProsperScore, pr$EmploymentStatus,summary)
```

```
## [[1]]
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##      NA      NA      NA     NaN      NA      NA    2255 
## 
## $Employed
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##   1.000   4.000   6.000   5.973   8.000  11.000      12 
## 
## $`Full-time`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##   1.000   6.000   8.000   7.006   9.000  11.000   18428 
## 
## $`Not available`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##      NA      NA      NA     NaN      NA      NA    5347 
## 
## $`Not employed`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##   1.000   4.000   5.000   5.308   7.000  10.000     186 
## 
## $Other
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
##    1.000    4.000    5.000    5.167    7.000   11.000
##
## $`Part-time`
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.000    5.000    7.000    6.801    9.000   10.000     832
##
## $Retired
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.000    5.000    7.000    6.237    8.000   10.000     428
##
## $`Self-employed`
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.000    3.000    4.000    4.444    6.000   11.000    1596
```



It is clearer to see from the heat map that employed and full time workers who have above average income will have higher Prosper Score. Meanwhile, for some self employed applicants, even though the income range is above $100,000, the Prosper Score is extremely low, indicting higher risk than other applicants with considerably lower income.

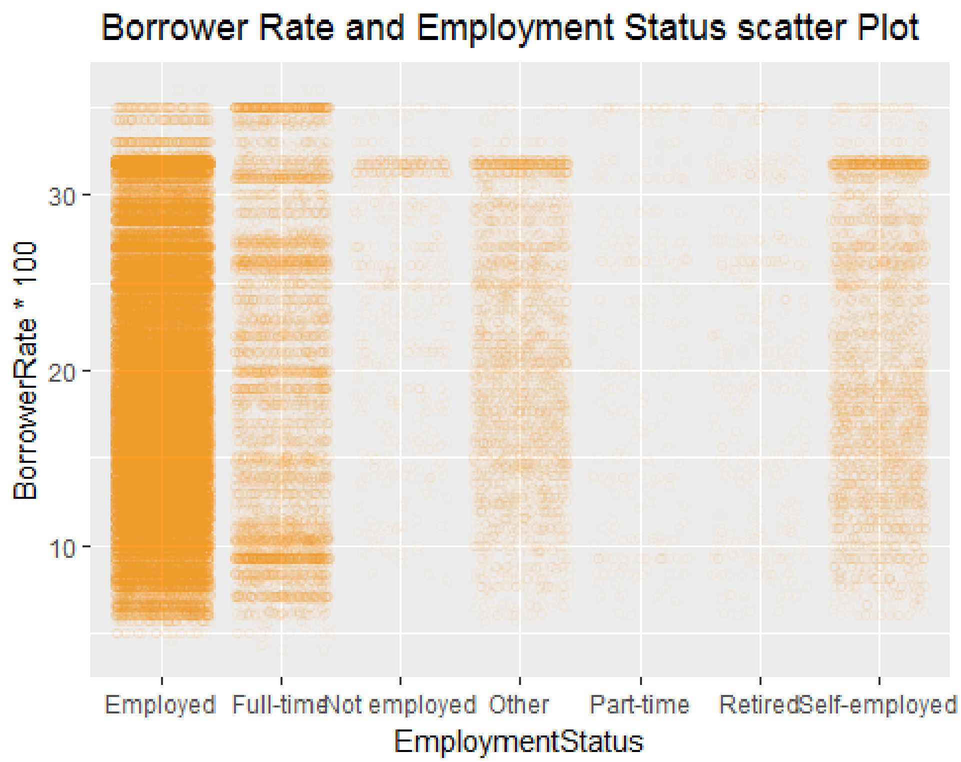
Unsperisingly, applicants with higher prosperscore seem to have lower borrow rate.

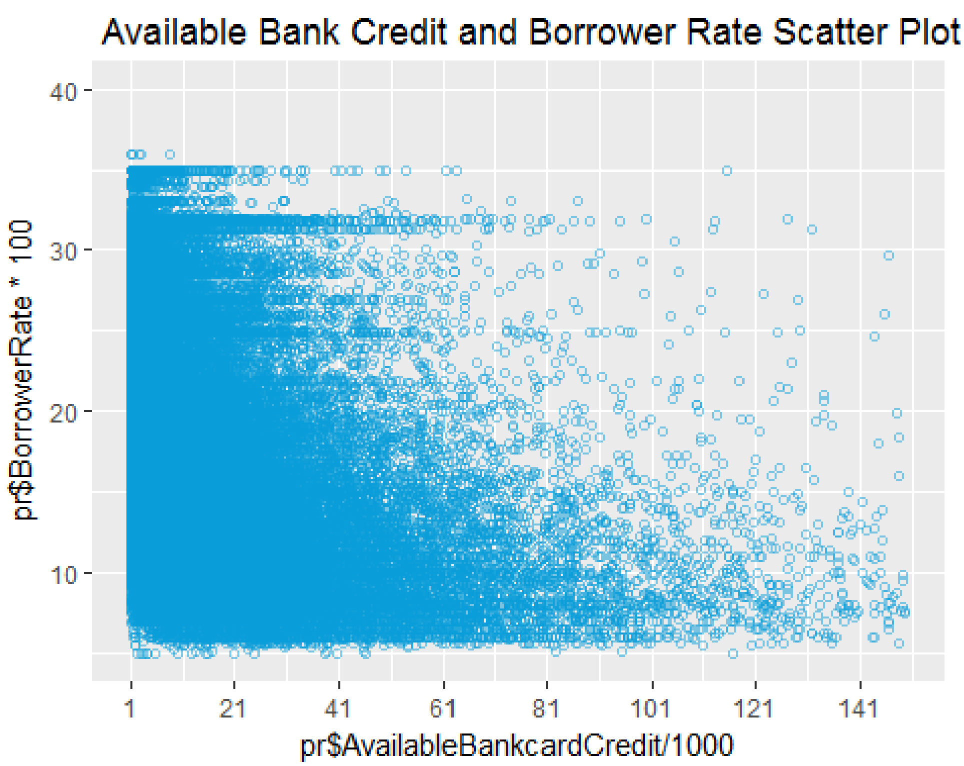
Summary for Prosper Score and Borrower Rate(%)

```
tapply(pr$BorrowerRate*100, pr$ProsperScore, summary)

## $`1`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.99   29.99   31.23   30.21   31.77   35.00
##
## $`2`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.50   24.92   27.86   27.12   30.32   36.00
##
## $`3`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.09   21.00   24.88   24.79   29.25   35.00
##
## $`4`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.16   17.90   21.24   22.54   26.99   36.00
##
## $`5`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.16   17.15   21.99   22.91   30.58   36.00
##
## $`6`
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    6.99   15.35   19.40   20.62   25.99   35.00 
## 
## $`7`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    6.59   13.85   17.60   18.51   24.68   35.00 
## 
## $`8`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    5.86   11.39   14.49   15.17   17.74   36.00 
## 
## $`9`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    4.98    9.46   11.39   12.51   14.35   35.00 
## 
## $`10`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   4.000   7.160   8.790   9.797  11.590  35.000 
## 
## $`11`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   6.050   6.590   8.690   9.328  10.990  19.500

## 
##  Pearson's product-moment correlation
## 
## data:  pr$ProsperScore and pr$BorrowerRate
## t = -248.98, df = 84851, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.6536072 -0.6458311
## sample estimates:
##        cor 
## -0.6497361
```

And also from the correlation, Prosper Score and Borrower Rate has a strong negative correlation.





Higer available bank card credit will also have lower borrow rate.

```
with(na.omit(pr), cor(BorrowerRate*100,AvailableBankcardCredit/1000))
```

```
## [1] -0.274097
```

Correlation between BorrowerRate and AvailableBankcardCredit

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?

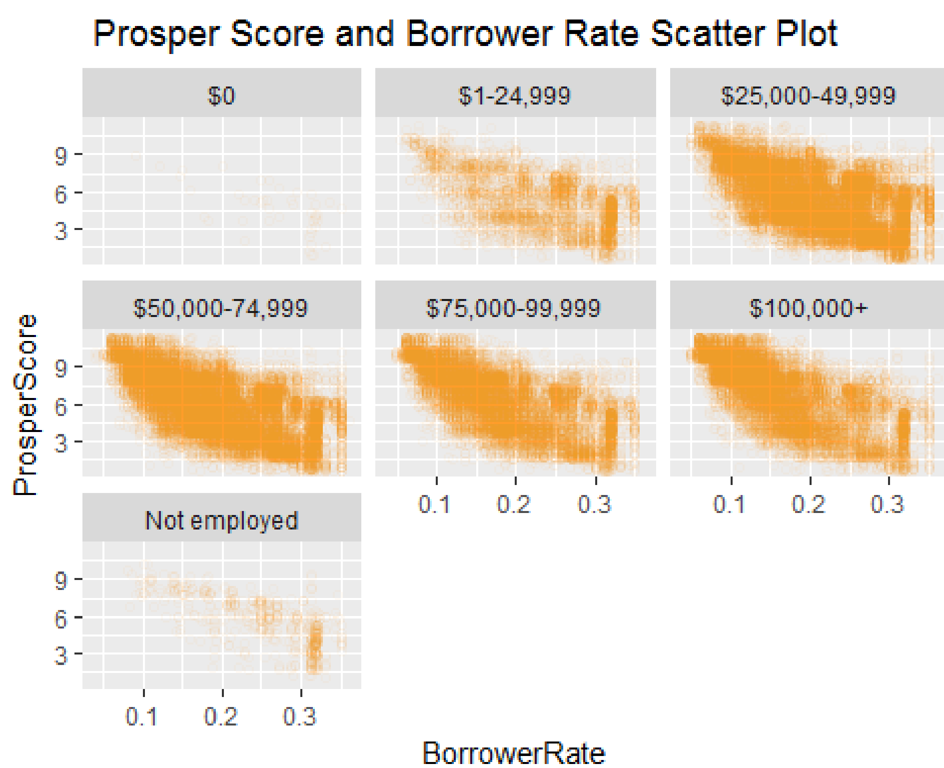
Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

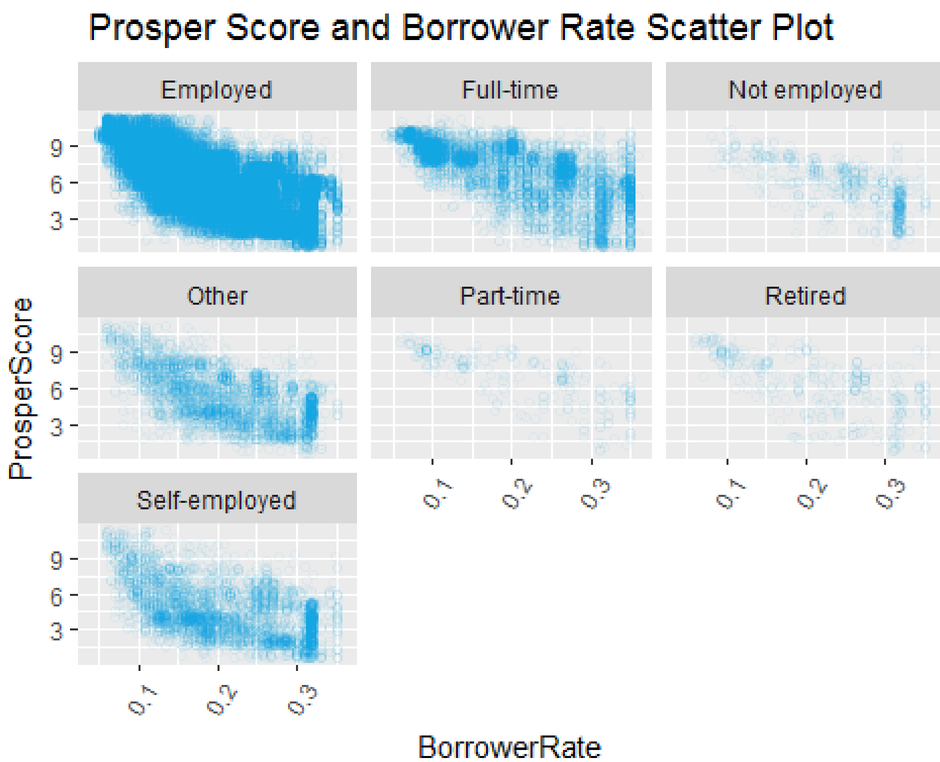
Other than main feature, I also noticed applicants with higher available bank card credit will also have lower borrow rate.

What was the strongest relationship you found?

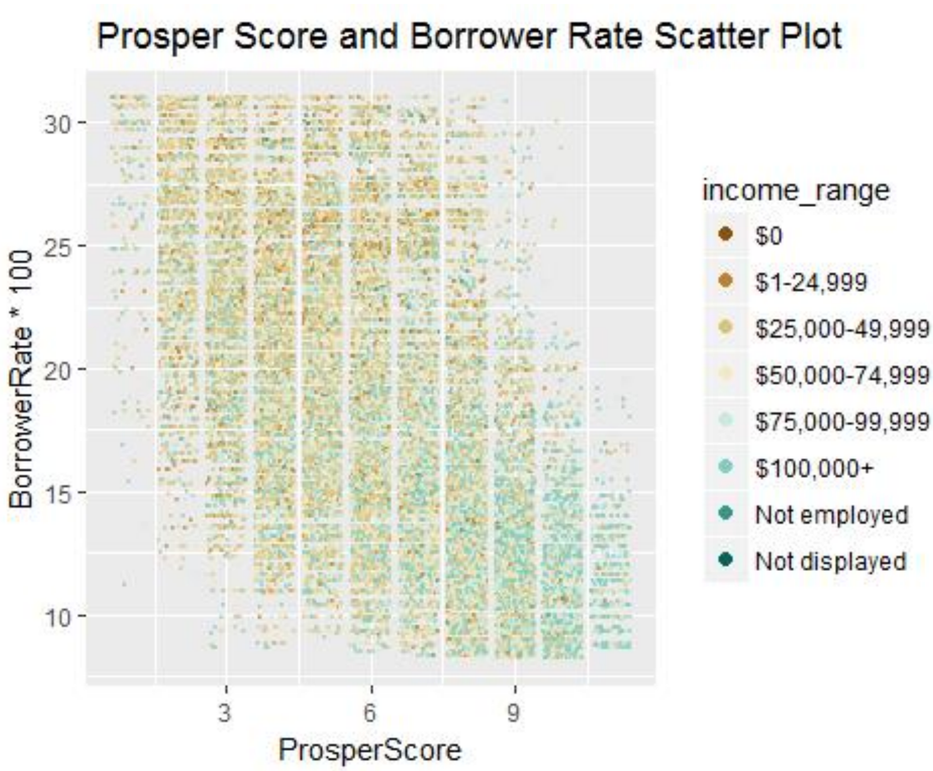
Borrow rate verus the properscore, The coefficient between them is -0.6682872.

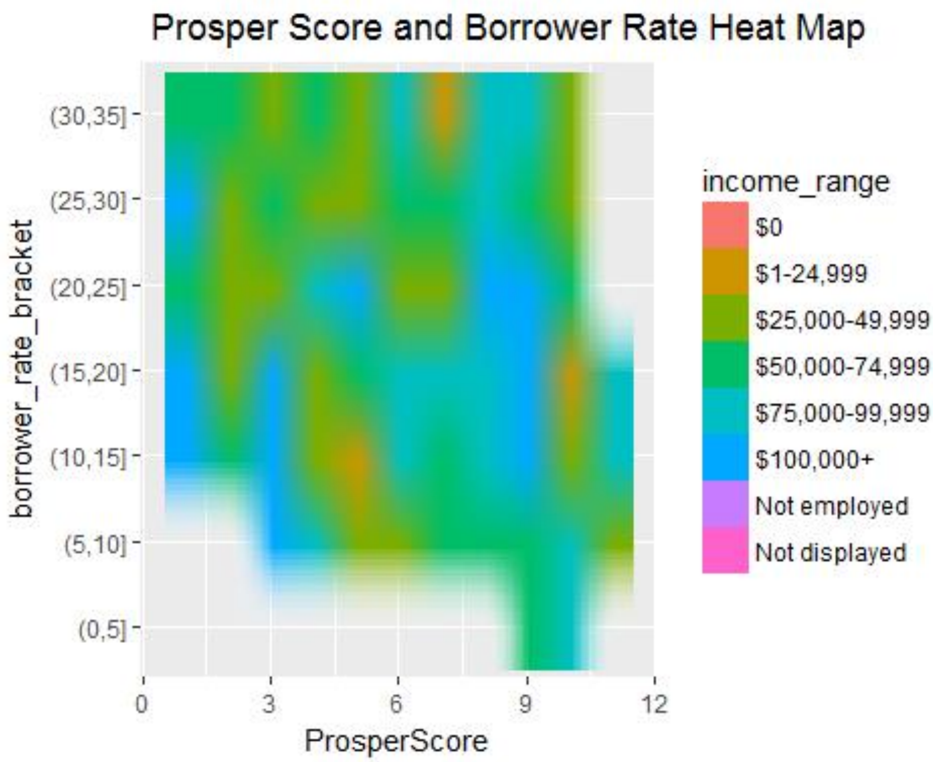
Multivariate Plots Section





The same pattern also occur for different employment status





When comparing with the main relationship ProsperScore and BorrowerRate, we also noticed a liner relationship adding the category variable income range

```
##
## Calls:
## m1: lm(formula = I(BorrowerRate) ~ 0 + I(ProsperScore), data = subset(pr,
##     !is.na(ProsperScore)))
## m2: lm(formula = I(BorrowerRate) ~ I(ProsperScore) +
AvailableBankcardCredit -
##     1, data = subset(pr, !is.na(ProsperScore)))
## m3: lm(formula = I(BorrowerRate) ~ I(ProsperScore) +
AvailableBankcardCredit +
##     IncomeRange - 1, data = subset(pr, !is.na(ProsperScore)))
## m4: lm(formula = I(BorrowerRate) ~ I(ProsperScore) +
AvailableBankcardCredit +
##     IncomeRange + EmploymentStatus - 1, data = subset(pr,
!is.na(ProsperScore)))
## m5: lm(formula = I(BorrowerRate) ~ I(ProsperScore) +
AvailableBankcardCredit +
##     IncomeRange + EmploymentStatus + TotalCreditLinespast7years -
##     1, data = subset(pr, !is.na(ProsperScore)))
##
##
============================================================================
==============================================
##                                                    m1              m2
m3              m4              m5
## ---------------------------------------------------------------------------
---------------------------------------------------
##   I(ProsperScore)                                  0.026***        0.028***
-0.018***       -0.019***       -0.019***
##                                                   (0.000)         (0.000)
(0.000)         (0.000)         (0.000)
##   AvailableBankcardCredit                                         -0.000***
-0.000***       -0.000***       -0.000***
##                                                                   (0.000)
(0.000)         (0.000)         (0.000)
##   IncomeRange: $0
0.352***        0.348***        0.348***
##
(0.008)         (0.008)         (0.008)
##   IncomeRange: $1-24,999
0.334***        0.336***        0.336***
##
(0.001)         (0.001)         (0.001)
##   IncomeRange: $100,000+
0.304***        0.307***        0.307***
##
(0.001)         (0.001)         (0.001)
##   IncomeRange: $25,000-49,999
0.317***        0.319***        0.318***
##
(0.001)         (0.001)         (0.001)
```

```
##   IncomeRange: $50,000-74,999
0.308***        0.310***        0.309***
##
(0.001)         (0.001)         (0.001)
##   IncomeRange: $75,000-99,999
0.306***        0.309***        0.308***
##
(0.001)         (0.001)         (0.001)
##   IncomeRange: Not employed
0.364***        0.368***        0.367***
##
(0.002)         (0.002)         (0.002)
##   EmploymentStatus: Full-time/Employed
0.025***        0.025***
##
(0.001)         (0.001)
##   EmploymentStatus: Other/Employed
-0.003**        -0.003**
##
(0.001)         (0.001)
##   EmploymentStatus: Part-time/Employed
0.016***        0.016***
##
(0.003)         (0.003)
##   EmploymentStatus: Retired/Employed
0.020***        0.020***
##
(0.003)         (0.003)
##   EmploymentStatus: Self-employed/Employed
-0.009***       -0.009***
##
(0.001)         (0.001)
##   TotalCreditLinespast7years
0.000
##
(0.000)
## -----------------------------------------------------------------------------
--------------------------------------------------
##   R-squared                                             0.612           0.621
0.932           0.933           0.933
##   adj. R-squared                                        0.612           0.621
0.932           0.933           0.933
##   sigma                                                 0.131           0.129
0.055           0.054           0.054
##   F                                                133697.431       69606.628
128923.549       84690.770       79044.868
##   p                                                     0.000           0.000
0.000           0.000           0.000
##   Log-likelihood                                    52266.161       53323.957
126092.434      126950.145      126950.686
```

```
##   Deviance                                      1449.358        1413.669
254.367         249.276         249.273
##   AIC                                         -104528.323     -106641.915
-252164.869     -253870.290     -253869.372
##   BIC                                         -104509.626     -106613.869
-252071.382     -253730.059     -253719.793
##   N                                             84853           84853
84853           84853           84853
##
=============================================================================
===============================================
```

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?

From multivariate analysis, the borrower's rate is also affacted by other variables such as income range, employment status, aviable bank credits.

Were there any interesting or surprising interactions between features?

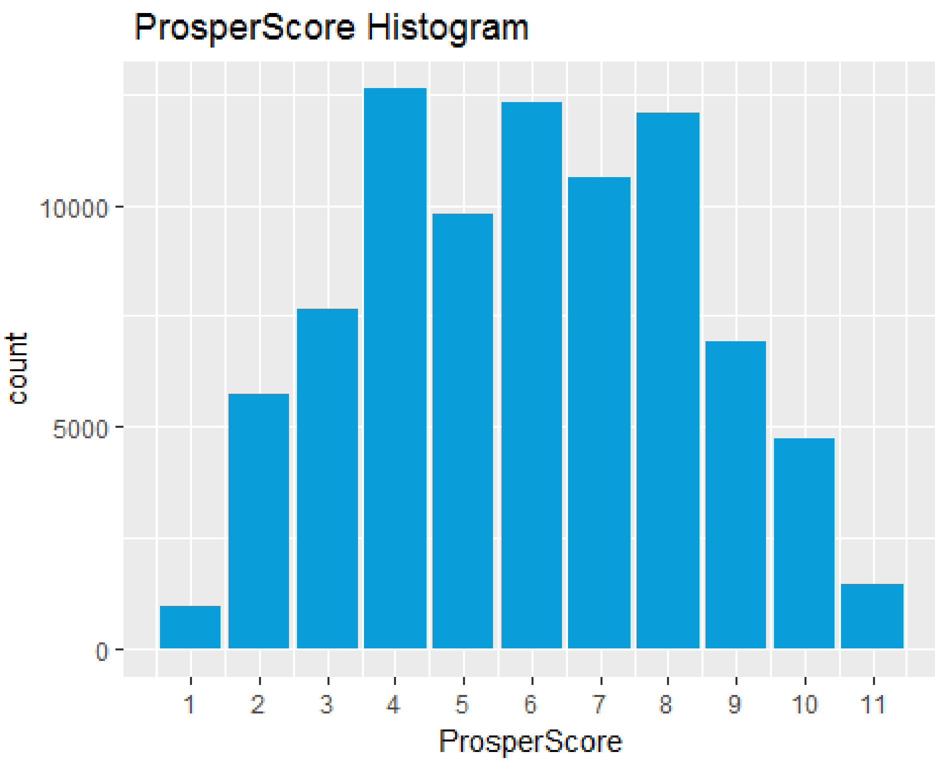
Holding the borrower's rate constant, the employed status will have lower prosper score.

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

Yes, I did. For the linear models I created, R square is about 93%, which means almost 93% percent of the variation can be explained by the model. I also excluded the intercept, which strenthed the linear relationship between the variables.

Final Plots and Summary

Plot One



Description One

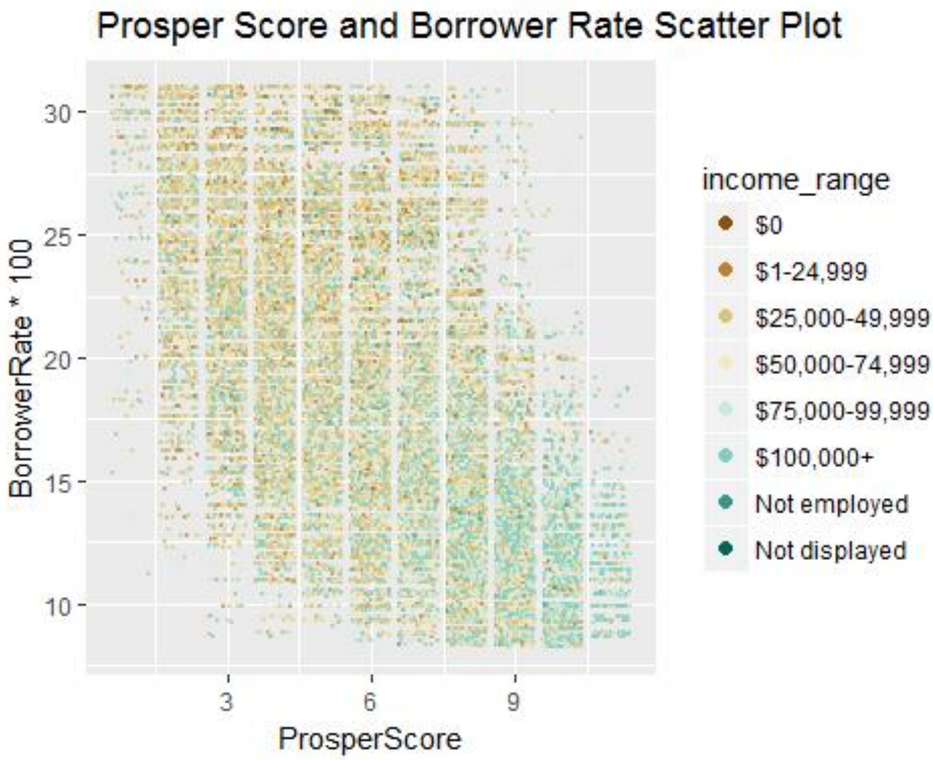
The loans were valued by the risk levels which bing called as ProsperScore, and the greater the score the lower the risk. Histogram showed us the counts for different ProsperScore. Majority of the loans have the score between 4 and 9.

Plot Two

Description Two

Ggplot gives us the negative correlation between ProsperScore and the borrower's rate, the higher the score seems to lead to a lower borrower's rate.

Plot Three



Description Three

When considering other variables, for example, income range, will also help us in predicting the borrower's rate.From the plot, the higer income leads also have higer ProsperScore, and revelvantly lower interest rate.

Reflection

Prosper Loan dataset has thorogh loan data regrading theri unique attributes, and when evaluating the loan applications, these variables could benefit the company in deciding the accurate rate. From the the analysis, it is shown that borrower's rate the highly correlated with borrower's ProsperScore, which measured the risk of the applicant. We also learned that other factors such as income level, employment status, avaible bank credits could also affect borrower's rate. More thorough demographic data can be included in the dataset, therefore, we can better the detailed attributs of the applicants.