

# Prosper Loan Data

## Univariate Plots Section

```
## [1] "ListingKey"
## [2] "ListingNumber"
## [3] "ListingCreationDate"
## [4] "CreditGrade"
## [5] "Term"
## [6] "LoanStatus"
## [7] "ClosedDate"
## [8] "BorrowerAPR"
## [9] "BorrowerRate"
## [10] "LenderYield"
## [11] "EstimatedEffectiveYield"
## [12] "EstimatedLoss"
## [13] "EstimatedReturn"
## [14] "ProsperRating..numeric."
## [15] "ProsperRating..Alpha."
## [16] "ProsperScore"
## [17] "ListingCategory..numeric."
## [18] "BorrowerState"
## [19] "Occupation"
## [20] "EmploymentStatus"
## [21] "EmploymentStatusDuration"
## [22] "IsBorrowerHomeowner"
## [23] "CurrentlyInGroup"
## [24] "GroupKey"
## [25] "DateCreditPulled"
## [26] "CreditScoreRangeLower"
## [27] "CreditScoreRangeUpper"
## [28] "FirstRecordedCreditLine"
## [29] "CurrentCreditLines"
## [30] "OpenCreditLines"
## [31] "TotalCreditLinespast7years"
## [32] "OpenRevolvingAccounts"
## [33] "OpenRevolvingMonthlyPayment"
## [34] "InquiriesLast6Months"
## [35] "TotalInquiries"
## [36] "CurrentDelinquencies"
## [37] "AmountDelinquent"
## [38] "DelinquenciesLast7Years"
## [39] "PublicRecordsLast10Years"
## [40] "PublicRecordsLast12Months"
## [41] "RevolvingCreditBalance"
## [42] "BankcardUtilization"
## [43] "AvailableBankcardCredit"
## [44] "TotalTrades"
## [45] "TradesNeverDelinquent..percentage."
## [46] "TradesOpenedLast6Months"
## [47] "DebtToIncomeRatio"
## [48] "IncomeRange"
## [49] "IncomeVerifiable"
## [50] "StatedMonthlyIncome"
## [51] "LoanKey"
## [52] "TotalProsperLoans"
## [53] "TotalProsperPaymentsBilled"
## [54] "OnTimeProsperPayments"
## [55] "ProsperPaymentsLessThanOneMonthLate"
## [56] "ProsperPaymentsOneMonthPlusLate"
## [57] "ProsperPrincipalBorrowed"
```

```
## [58] "ProsperPrincipalOutstanding"  
## [59] "ScorexChangeAtTimeOfListing"  
## [60] "LoanCurrentDaysDelinquent"  
## [61] "LoanFirstDefaultedCycleNumber"  
## [62] "LoanMonthsSinceOrigination"  
## [63] "LoanNumber"  
## [64] "LoanOriginalAmount"  
## [65] "LoanOriginationDate"  
## [66] "LoanOriginationQuarter"  
## [67] "MemberKey"  
## [68] "MonthlyLoanPayment"  
## [69] "LP_CustomerPayments"  
## [70] "LP_CustomerPrincipalPayments"  
## [71] "LP_InterestandFees"  
## [72] "LP_ServiceFees"  
## [73] "LP_CollectionFees"  
## [74] "LP_GrossPrincipalLoss"  
## [75] "LP_NetPrincipalLoss"  
## [76] "LP_NonPrincipalRecoverypayments"  
## [77] "PercentFunded"  
## [78] "Recommendations"  
## [79] "InvestmentFromFriendsCount"  
## [80] "InvestmentFromFriendsAmount"  
## [81] "Investors"
```

```

##          ListingKey      ListingNumber      ListingCreationDate
## 17A93590655669644DB4C06:       6      Min.    : 4      02:03.5:    12
## 349D3587495831350F0F648:       4     1st Qu.: 400919    20:28.5:    12
## 47C1359638497431975670B:       4     Median : 600554    22:05.3:    12
## 8474358854651984137201C:       4     Mean    : 627886    29:50.4:    12
## DE8535960513435199406CE:       4     3rd Qu.: 892634    00:34.6:    11
## 04C13599434217079754AEE:       3     Max.    :1255725    04:49.3:    11
## (Other)                      :113912                               (Other):113867
##          CreditGrade      Term           LoanStatus
## :84984      Min.    :12.00      Current      :56576
## C          : 5649      1st Qu.:36.00      Completed    :38074
## D          : 5153      Median :36.00      Chargedoff   :11992
## B          : 4389      Mean    :40.83      Defaulted    : 5018
## AA         : 3509      3rd Qu.:36.00      Past Due (1-15 days): 806
## HR         : 3508      Max.    :60.00      Past Due (31-60 days): 363
## (Other): 6745                               (Other)      : 1108
##          ClosedDate      BorrowerAPR      BorrowerRate
## :58848      Min.    :0.00653      Min.    :0.0000
## 3/4/14 0:00 : 105      1st Qu.:0.15629      1st Qu.:0.1340
## 2/19/14 0:00 : 100      Median :0.20976      Median :0.1840
## 2/11/14 0:00 :  92      Mean    :0.21883      Mean    :0.1928
## 10/30/12 0:00:  81      3rd Qu.:0.28381      3rd Qu.:0.2500
## 2/26/13 0:00 :  78      Max.    :0.51229      Max.    :0.4975
## (Other)      :54633      NA's    :25
##          LenderYield      EstimatedEffectiveYield      EstimatedLoss
## Min.    :-0.0100      Min.    :-0.183      Min.    :0.005
## 1st Qu.: 0.1242      1st Qu.: 0.116      1st Qu.:0.042
## Median : 0.1730      Median : 0.162      Median :0.072
## Mean    : 0.1827      Mean    : 0.169      Mean    :0.080
## 3rd Qu.: 0.2400      3rd Qu.: 0.224      3rd Qu.:0.112
## Max.    : 0.4925      Max.    : 0.320      Max.    :0.366
## NA's    :29084      NA's    :29084      NA's    :29084
##          EstimatedReturn      ProsperRating..numeric.      ProsperRating..Alpha.
## Min.    :-0.183      Min.    :1.000      :29084
## 1st Qu.: 0.074      1st Qu.:3.000      C      :18345
## Median : 0.092      Median :4.000      B      :15581
## Mean    : 0.096      Mean    :4.072      A      :14551
## 3rd Qu.: 0.117      3rd Qu.:5.000      D      :14274
## Max.    : 0.284      Max.    :7.000      E      : 9795
## NA's    :29084      NA's    :29084      (Other):12307
##          ProsperScore      ListingCategory..numeric.      BorrowerState
## Min.    : 1.00      Min.    : 0.000      CA     :14717
## 1st Qu.: 4.00      1st Qu.: 1.000      TX     : 6842
## Median : 6.00      Median : 1.000      NY     : 6729
## Mean    : 5.95      Mean    : 2.774      FL     : 6720
## 3rd Qu.: 8.00      3rd Qu.: 3.000      IL     : 5921
## Max.    :11.00      Max.    :20.000      : 5515
## NA's    :29084      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      IsBorrowerHomeowner      CurrentlyInGroup

```

```

## Min. : 0.00          Mode :logical      Mode :logical
## 1st Qu.: 26.00       FALSE:56459        FALSE:101218
## Median : 67.00       TRUE :57478         TRUE :12719
## Mean   : 96.07
## 3rd Qu.:137.00
## Max.   :755.00
## NA's   :7625

##           GroupKey          DateCreditPulled
##                 :100596  11/4/13 14:12:     8
## 783C3371218786870A73D20: 1140  07:54.9      : 6
## 3D4D3366260257624AB272D: 916   12/23/13 9:38:    6
## 6A3B336601725506917317E: 698   14:22.9      : 6
## FEF83377364176536637E50: 611   33:37.0      : 6
## C9643379247860156A00EC0: 342   34:47.1      : 6
## (Other)            : 9634 (Other)       :113899

## CreditScoreRangeLower CreditScoreRangeUpper FirstRecordedCreditLine
## Min.   : 0.0          Min.   : 19.0          : 697
## 1st Qu.:660.0         1st Qu.:679.0         12/1/93 0:00: 185
## Median :680.0         Median :699.0         11/1/94 0:00: 178
## Mean   :685.6         Mean   :704.6         11/1/95 0:00: 168
## 3rd Qu.:720.0         3rd Qu.:739.0         4/1/90 0:00 : 161
## Max.   :880.0         Max.   :899.0         3/1/95 0:00 : 159
## NA's   :591           NA's   :591           (Other)       :112389

## CurrentCreditLines OpenCreditLines TotalCreditLinespast7years
## Min.   : 0.00          Min.   : 0.00          Min.   : 2.00
## 1st Qu.: 7.00          1st Qu.: 6.00          1st Qu.: 17.00
## Median :10.00          Median : 9.00          Median : 25.00
## Mean   :10.32          Mean   : 9.26          Mean   : 26.75
## 3rd Qu.:13.00          3rd Qu.:12.00          3rd Qu.: 35.00
## Max.   :59.00           Max.   :54.00           Max.   :136.00
## NA's   :7604            NA's   :7604           NA's   :697

## OpenRevolvingAccounts OpenRevolvingMonthlyPayment InquiriesLast6Months
## Min.   : 0.00          Min.   : 0.0          Min.   : 0.000
## 1st Qu.: 4.00          1st Qu.: 114.0        1st Qu.: 0.000
## Median : 6.00          Median : 271.0        Median : 1.000
## Mean   : 6.97          Mean   : 398.3        Mean   : 1.435
## 3rd Qu.: 9.00          3rd Qu.: 525.0        3rd Qu.: 2.000
## Max.   :51.00           Max.   :14985.0       Max.   :105.000
## NA's   :697             NA's   :697           NA's   :697

## TotalInquiries CurrentDelinquencies AmountDelinquent
## Min.   : 0.000          Min.   : 0.0000          Min.   : 0.0
## 1st Qu.: 2.000          1st Qu.: 0.0000          1st Qu.: 0.0
## Median : 4.000          Median : 0.0000          Median : 0.0
## Mean   : 5.584          Mean   : 0.5921          Mean   : 984.5
## 3rd Qu.: 7.000          3rd Qu.: 0.0000          3rd Qu.: 0.0
## Max.   :379.000          Max.   :83.0000          Max.   :463881.0
## NA's   :1159            NA's   :697            NA's   :7622

## DelinquenciesLast7Years PublicRecordsLast10Years
## Min.   : 0.000          Min.   : 0.0000
## 1st Qu.: 0.000          1st Qu.: 0.0000
## Median : 0.000          Median : 0.0000
## Mean   : 4.155          Mean   : 0.3126
## 3rd Qu.: 3.000          3rd Qu.: 0.0000
## Max.   :99.000           Max.   :38.0000
## NA's   :990              NA's   :697

## PublicRecordsLast12Months RevolvingCreditBalance BankcardUtilization
## Min.   : 0.000          Min.   : 0          Min.   :0.000
## 1st Qu.: 0.000          1st Qu.: 3121        1st Qu.:0.310

```

```

## Median : 0.000      Median : 8549      Median : 0.600
## Mean   : 0.015      Mean   : 17599      Mean   : 0.561
## 3rd Qu.: 0.000      3rd Qu.: 19521      3rd Qu.: 0.840
## Max.   :20.000      Max.   :1435667     Max.   :5.950
## NA's   :7604        NA's   :7604        NA's   :7604
## AvailableBankcardCredit TotalTrades
## Min.   : 0           Min.   : 0.00
## 1st Qu.: 880         1st Qu.: 15.00
## Median : 4100        Median : 22.00
## Mean   : 11210        Mean   : 23.23
## 3rd Qu.: 13180        3rd Qu.: 30.00
## Max.   :646285       Max.   :126.00
## NA's   :7544         NA's   :7544
## TradesNeverDelinquent..percentage. TradesOpenedLast6Months
## Min.   :0.000          Min.   : 0.000
## 1st Qu.:0.820          1st Qu.: 0.000
## Median :0.940          Median : 0.000
## Mean   :0.886          Mean   : 0.802
## 3rd Qu.:1.000          3rd Qu.: 1.000
## Max.   :1.000          Max.   :20.000
## NA's   :7544           NA's   :7544
## DebtToIncomeRatio      IncomeRange      IncomeVerifiable
## Min.   : 0.000          $25,000-49,999:32192    Mode :logical
## 1st Qu.: 0.140          $50,000-74,999:31050    FALSE:8669
## Median : 0.220          $100,000+:17337       TRUE :105268
## Mean   : 0.276          $75,000-99,999:16916
## 3rd Qu.: 0.320          Not displayed : 7741
## Max.   :10.010          $1-24,999   : 7274
## NA's   :8554            (Other)      : 1427
## StatedMonthlyIncome      LoanKey        TotalProsperLoans
## Min.   : 0              CB1B37030986463208432A1: 6  Min.   :0.00
## 1st Qu.: 3200          2DEE3698211017519D7333F: 4  1st Qu.:1.00
## Median : 4667          9F4B37043517554537C364C: 4  Median :1.00
## Mean   : 5608          D895370150591392337ED6D: 4  Mean   :1.42
## 3rd Qu.: 6825          E6FB37073953690388BC56D: 4  3rd Qu.:2.00
## Max.   :1750003         0D8F37036734373301ED419: 3  Max.   :8.00
##                   (Other)          :113912    NA's   :91852
## TotalProsperPaymentsBilled OnTimeProsperPayments
## Min.   : 0.00          Min.   : 0.00
## 1st Qu.: 9.00          1st Qu.: 9.00
## Median : 16.00          Median : 15.00
## Mean   : 22.93          Mean   : 22.27
## 3rd Qu.: 33.00          3rd Qu.: 32.00
## Max.   :141.00          Max.   :141.00
## NA's   :91852           NA's   :91852
## ProsperPaymentsLessThanOneMonthLate ProsperPaymentsOneMonthPlusLate
## Min.   : 0.00          Min.   : 0.00
## 1st Qu.: 0.00          1st Qu.: 0.00
## Median : 0.00          Median : 0.00
## Mean   : 0.61          Mean   : 0.05
## 3rd Qu.: 0.00          3rd Qu.: 0.00
## Max.   :42.00          Max.   :21.00
## NA's   :91852           NA's   :91852
## ProsperPrincipalBorrowed ProsperPrincipalOutstanding
## Min.   : 0              Min.   : 0
## 1st Qu.: 3500          1st Qu.: 0
## Median : 6000          Median : 1627
## Mean   : 8472          Mean   : 2930

```

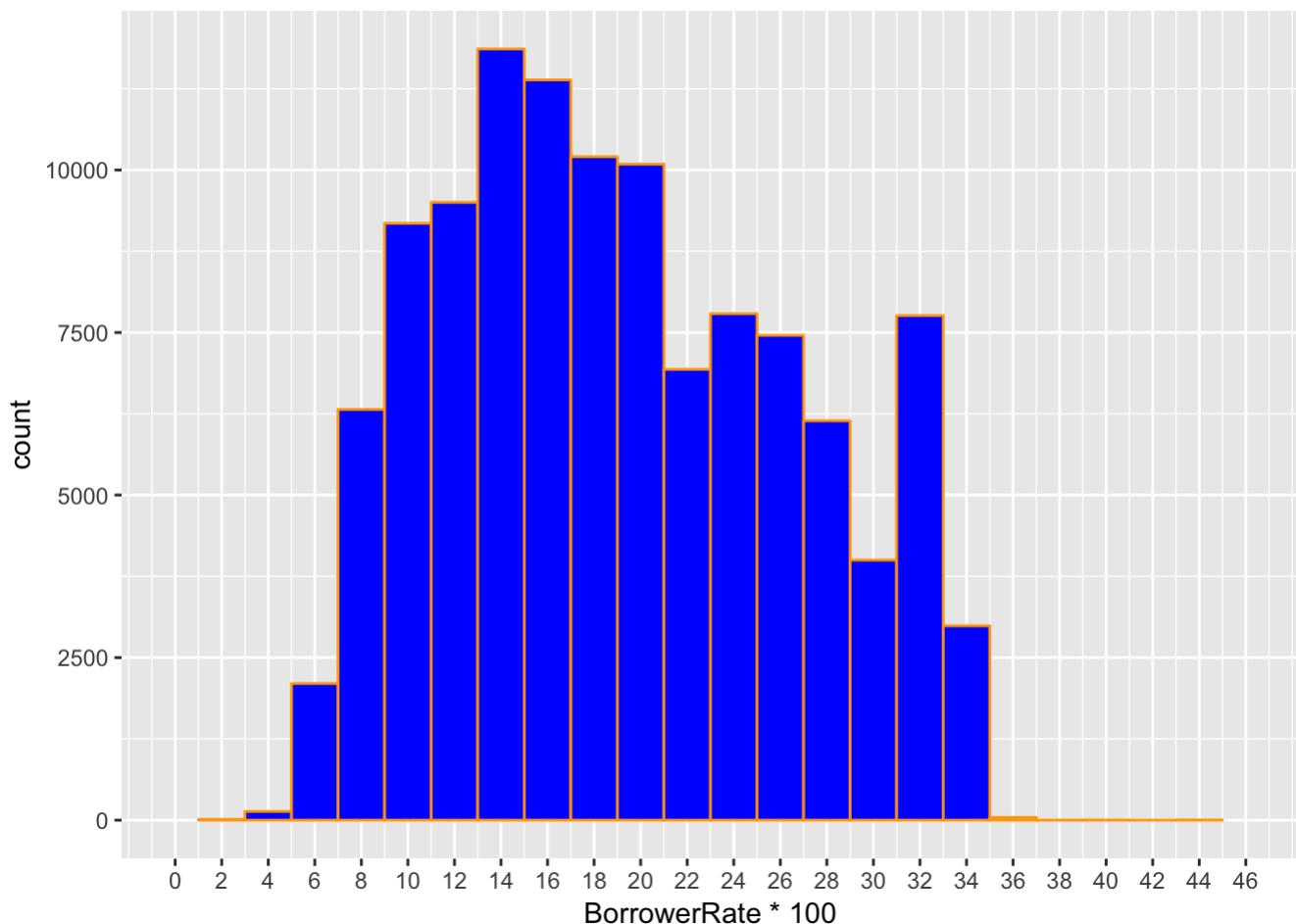
```

## 3rd Qu.:11000          3rd Qu.: 4127
## Max.    :72499          Max.    :23451
## NA's    :91852          NA's    :91852
## ScorexChangeAtTimeOfListing LoanCurrentDaysDelinquent
## Min.   :-209.00         Min.   : 0.0
## 1st Qu.:-35.00          1st Qu.: 0.0
## Median :-3.00           Median : 0.0
## Mean   :-3.22           Mean   : 152.8
## 3rd Qu.: 25.00          3rd Qu.: 0.0
## Max.   :286.00          Max.   :2704.0
## NA's   :95009
## LoanFirstDefaultedCycleNumber LoanMonthsSinceOrigination   LoanNumber
## Min.   : 0.00            Min.   : 0.0                  Min.   : 1
## 1st Qu.: 9.00            1st Qu.: 6.0                  1st Qu.: 37332
## Median :14.00            Median : 21.0                 Median : 68599
## Mean   :16.27            Mean   : 31.9                 Mean   : 69444
## 3rd Qu.:22.00            3rd Qu.: 65.0                 3rd Qu.:101901
## Max.   :44.00            Max.   :100.0                 Max.   :136486
## NA's   :96985
## LoanOriginalAmount      LoanOriginationDate  LoanOriginationQuarter
## Min.   :1000             1/22/14 0:00 : 491   Q4 2013:14450
## 1st Qu.:4000             11/13/13 0:00: 490   Q1 2014:12172
## Median :6500             2/19/14 0:00 : 439   Q3 2013: 9180
## Mean   :8337             10/16/13 0:00: 434   Q2 2013: 7099
## 3rd Qu.:12000            1/28/14 0:00 : 339   Q3 2012: 5632
## Max.   :35000            9/24/13 0:00 : 316   Q2 2012: 5061
## (Other)                :111428   (Other):60343
## MemberKey               MonthlyLoanPayment LP_CustomerPayments
## 63CA34120866140639431C9: 9   Min.   : 0.0      Min.   : -2.35
## 16083364744933457E57FB9: 8   1st Qu.: 131.6    1st Qu.: 1005.76
## 3A2F3380477699707C81385: 8   Median : 217.7    Median : 2583.83
## 4D9C3403302047712AD0CDD: 8   Mean   : 272.5    Mean   : 4183.08
## 739C338135235294782AE75: 8   3rd Qu.: 371.6    3rd Qu.: 5548.40
## 7E1733653050264822FAA3D: 8   Max.   :2251.5    Max.   :40702.39
## (Other)                :113888
## LP_CustomerPrincipalPayments LP_InterestandFees LP_ServiceFees
## Min.   : 0.0              Min.   : -2.35    Min.   : -664.87
## 1st Qu.: 500.9            1st Qu.: 274.87   1st Qu.: -73.18
## Median :1587.5            Median : 700.84   Median : -34.44
## Mean   :3105.5            Mean   :1077.54   Mean   : -54.73
## 3rd Qu.: 4000.0            3rd Qu.: 1458.54  3rd Qu.: -13.92
## Max.   :35000.0            Max.   :15617.03  Max.   : 32.06
##
## LP_CollectionFees  LP_GrossPrincipalLoss LP_NetPrincipalLoss
## Min.   :-9274.75          Min.   : -94.2     Min.   : -954.5
## 1st Qu.: 0.00              1st Qu.: 0.0      1st Qu.: 0.0
## Median : 0.00              Median : 0.0      Median : 0.0
## Mean   :-14.24             Mean   : 700.4    Mean   : 681.4
## 3rd Qu.: 0.00              3rd Qu.: 0.0      3rd Qu.: 0.0
## Max.   : 0.00              Max.   :25000.0   Max.   :25000.0
##
## LP_NonPrincipalRecoverypayments PercentFunded   Recommendations
## Min.   : 0.00              Min.   :0.7000    Min.   : 0.00000
## 1st Qu.: 0.00              1st Qu.:1.0000    1st Qu.: 0.00000
## Median : 0.00              Median :1.0000    Median : 0.00000
## Mean   : 25.14              Mean   :0.9986    Mean   : 0.04803
## 3rd Qu.: 0.00              3rd Qu.:1.0000    3rd Qu.: 0.00000
## Max.   :21117.90            Max.   :1.0125    Max.   :39.00000

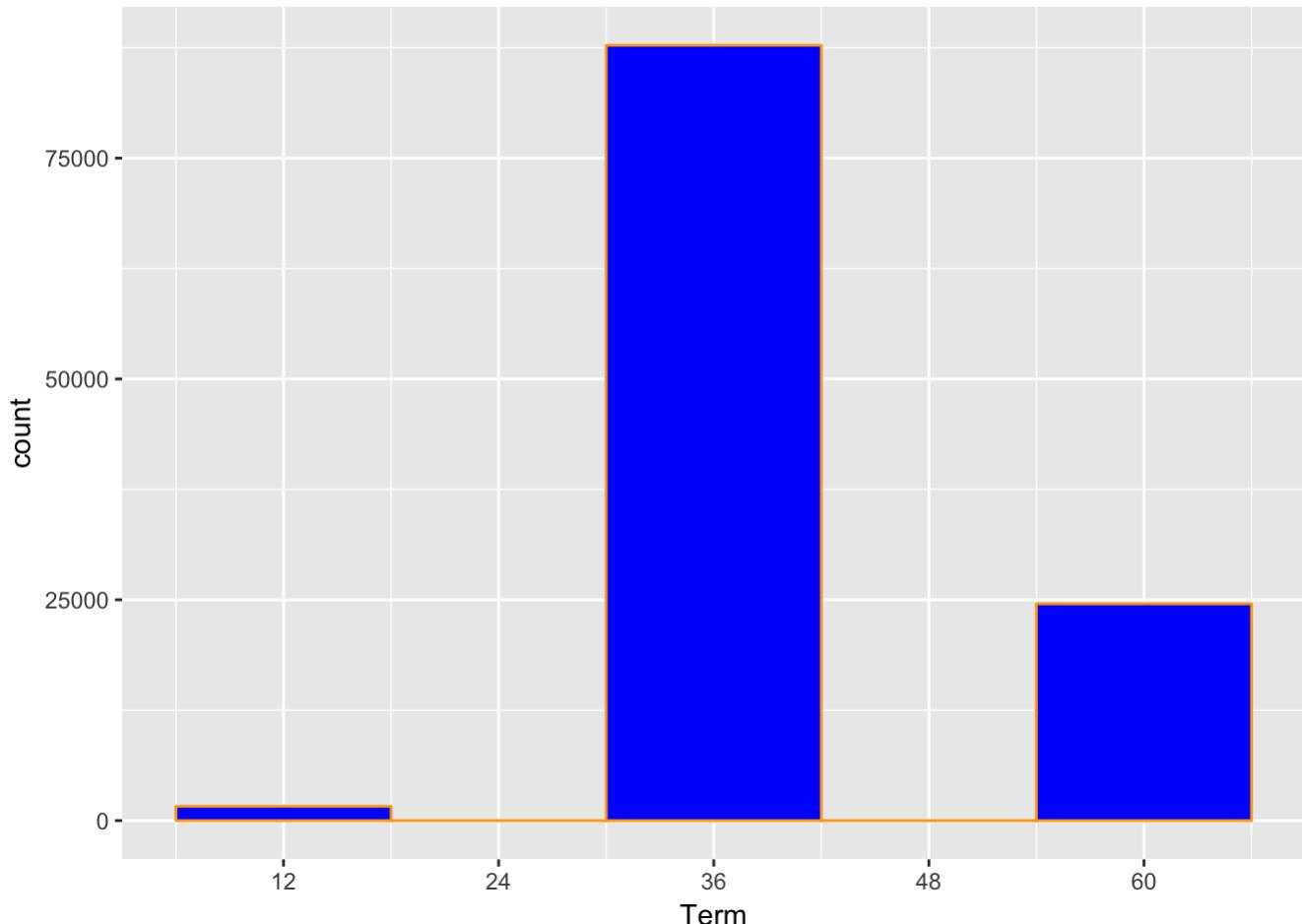
```

```
## 
##   InvestmentFromFriendsCount InvestmentFromFriendsAmount     Investors
##   Min.   : 0.00000           Min.   :  0.00               Min.   : 1.00
##   1st Qu.: 0.00000          1st Qu.:  0.00               1st Qu.: 2.00
##   Median : 0.00000          Median :  0.00               Median : 44.00
##   Mean    : 0.02346          Mean   : 16.55              Mean   : 80.48
##   3rd Qu.: 0.00000          3rd Qu.:  0.00               3rd Qu.:115.00
##   Max.    :33.00000          Max.   :25000.00             Max.   :1189.00
## 
```

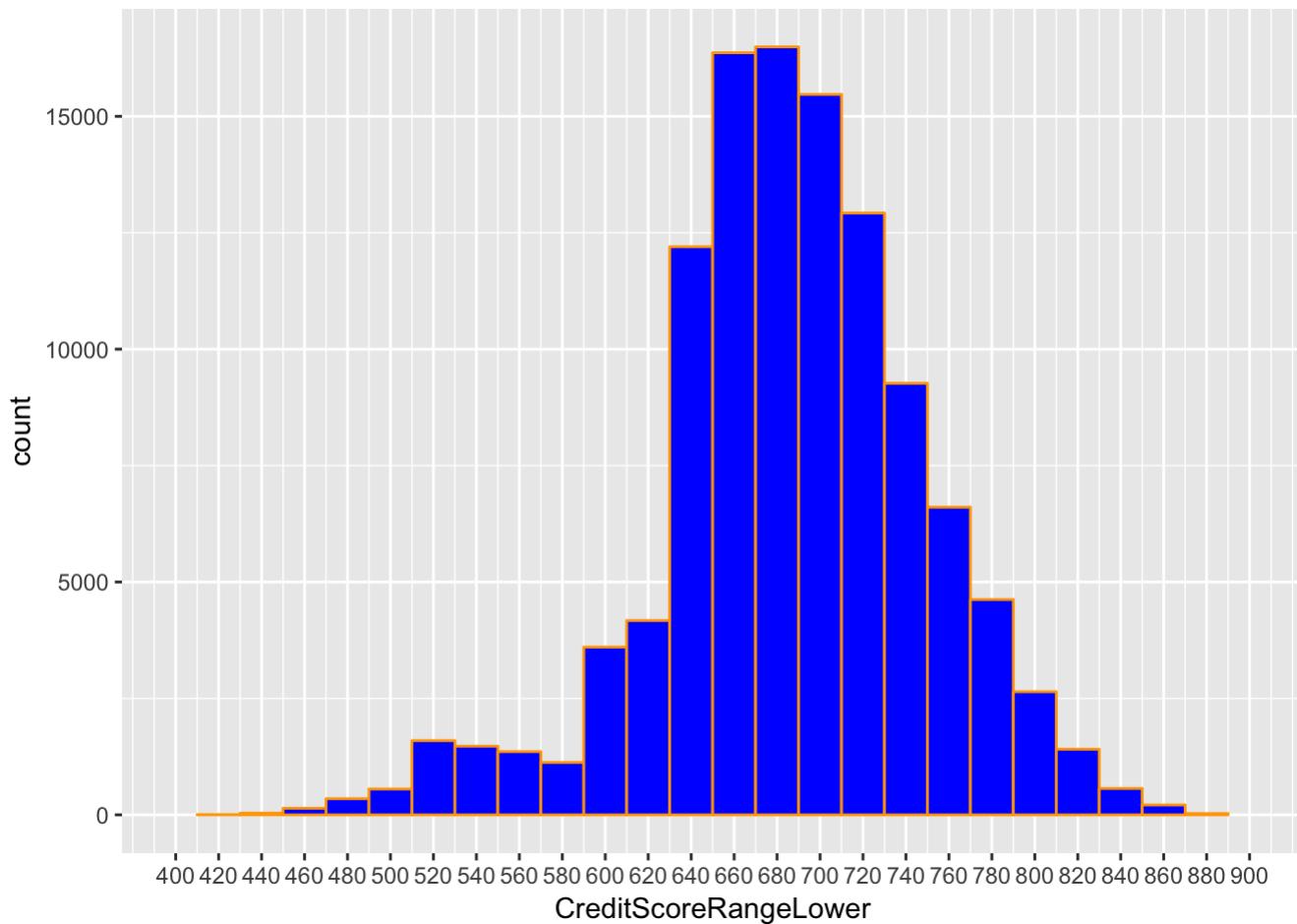
This data set contains 113,937 loans with 81 variables on each loan, including loan amount, borrower rate (or interest rate), current loan status, borrower income, borrower employment status, borrower credit history, and the latest payment information.



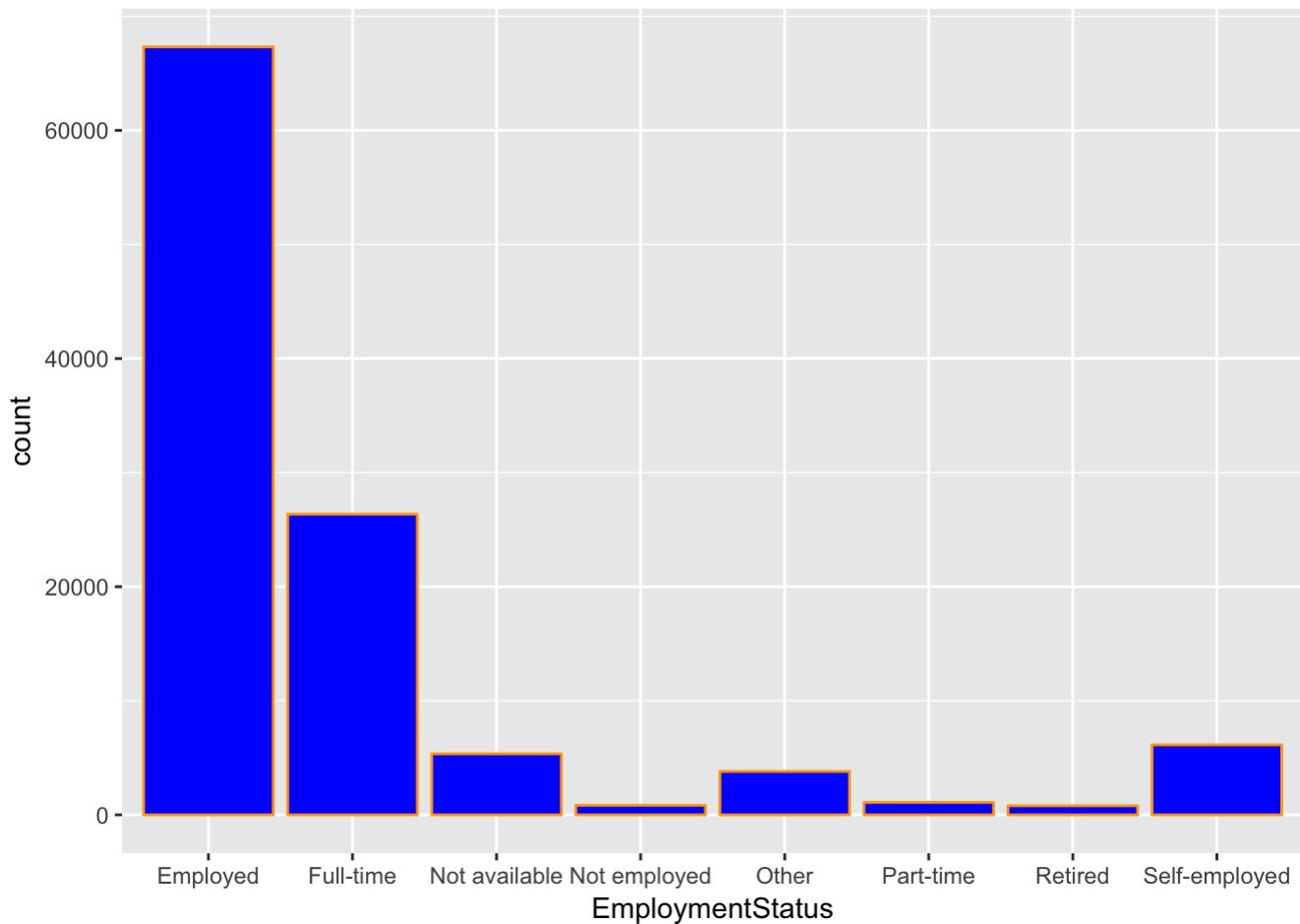
Plotting Borrower Rate histogram reveals that most of Prosper borrowers have their loans' rate (APR) of around 14%.



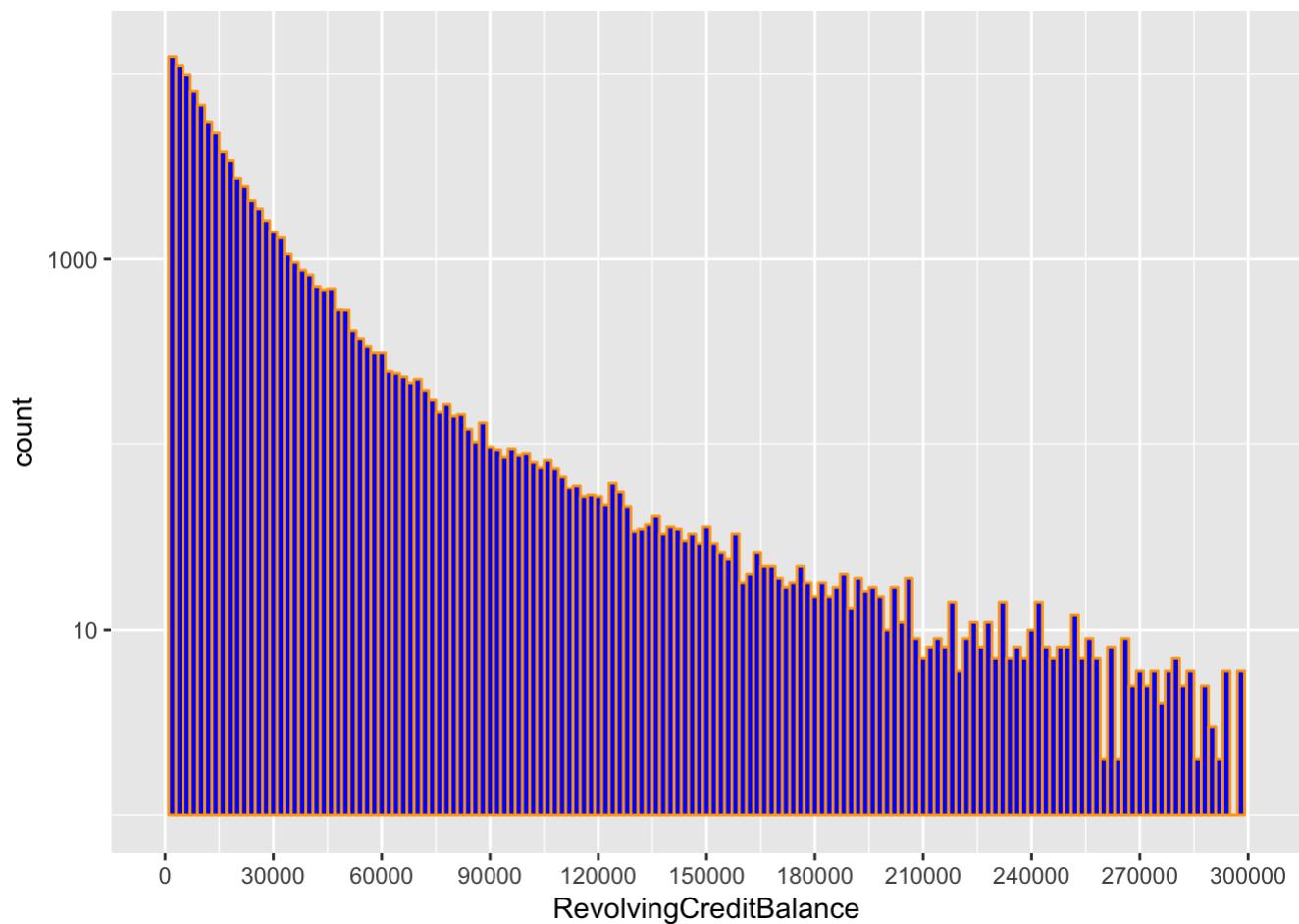
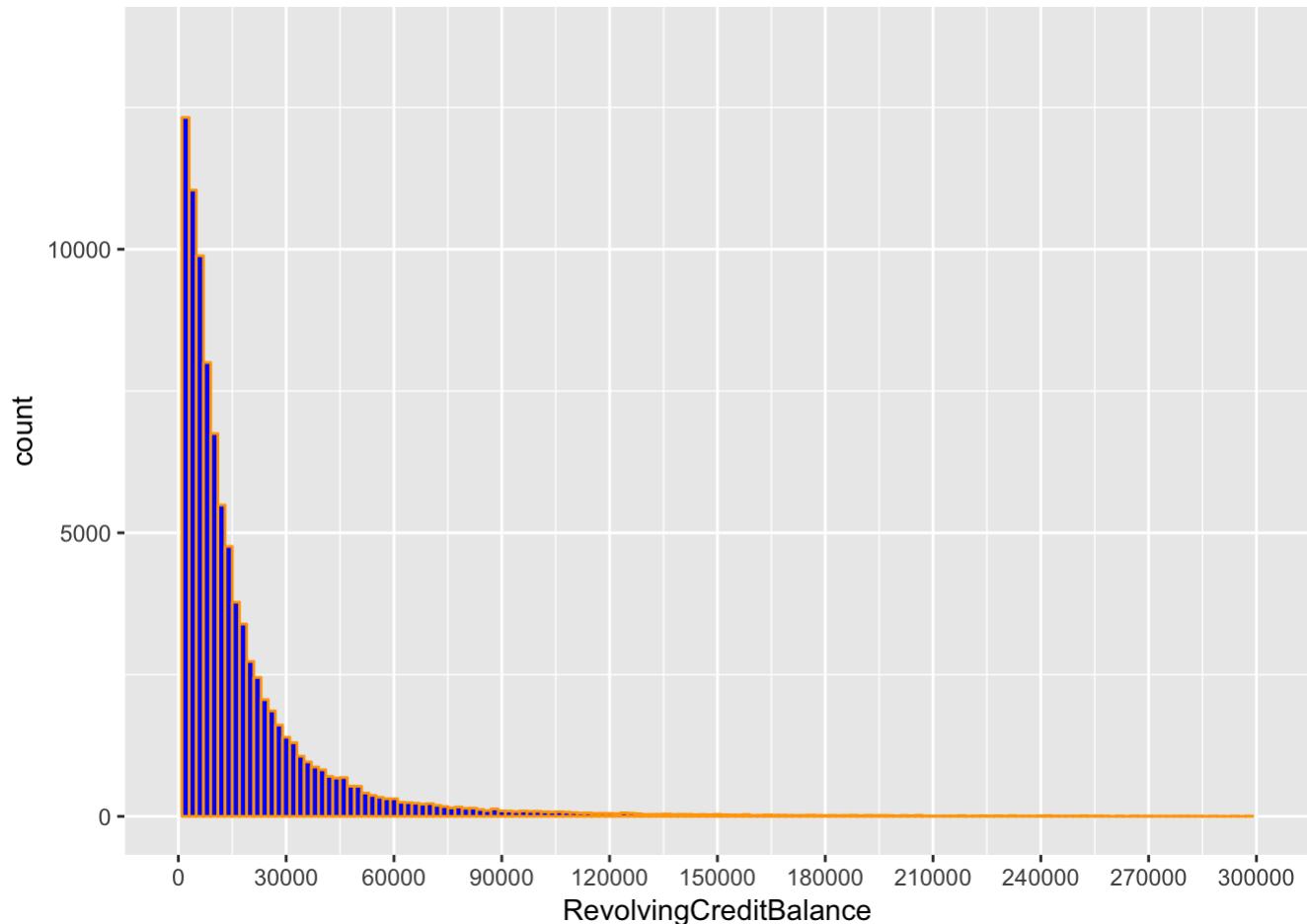
Only 3 types of loan terms offered by Prosper: 12, 36 and 60 months. As we can see from the chart above, most of the borrowers select 36 months term. More than 75% borrowers prefer 36 months or 3 years terms, probably because it gives enough length to repay the loan but not too long to accumulate unnecessary interests.



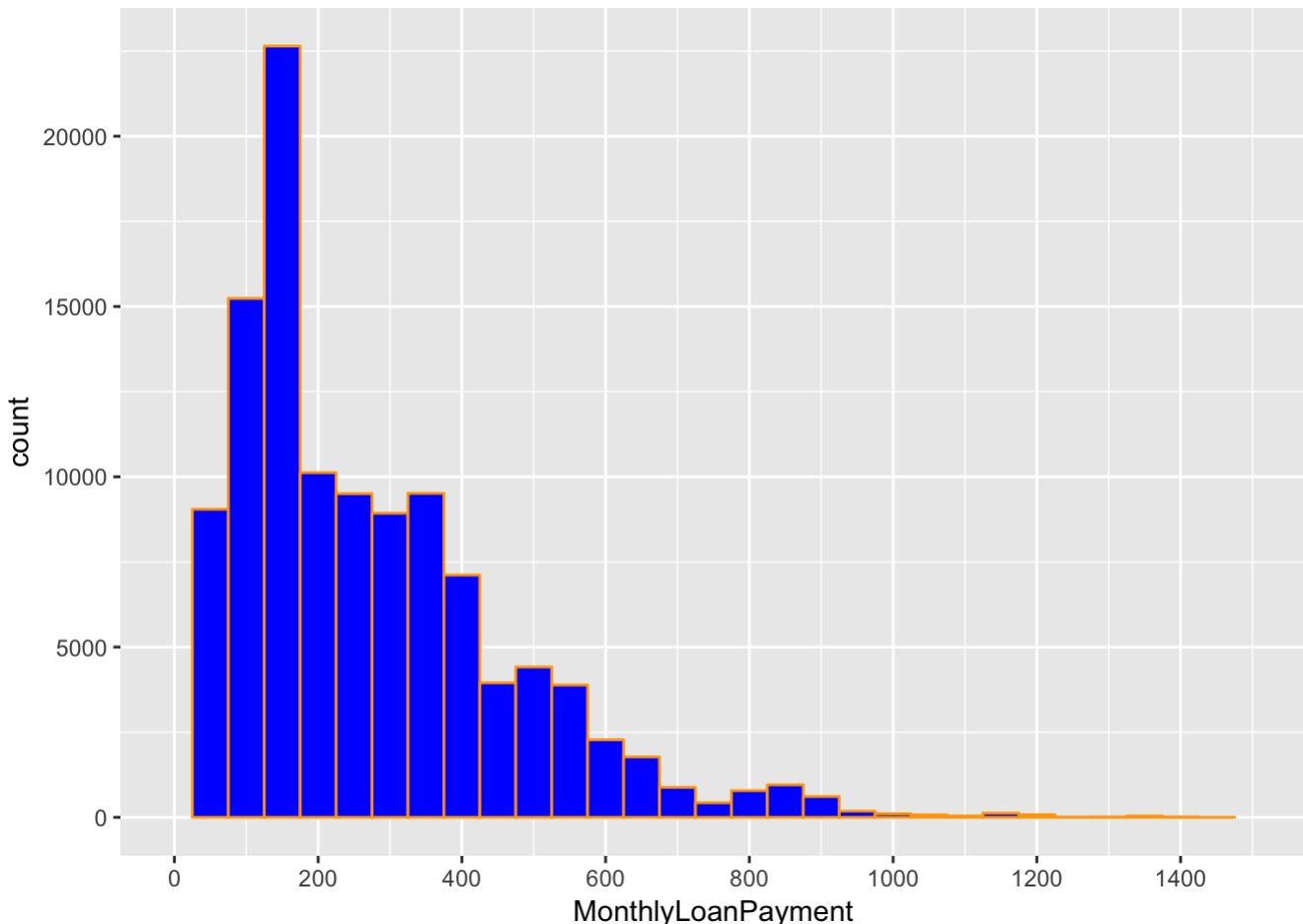
The plot above illustrates the lower range of borrowers' credit scores. It follows normal curve distribution. Most borrowers' credit scores fall in around 670 and 680. Credit score in this range is considered from Fair to Good.



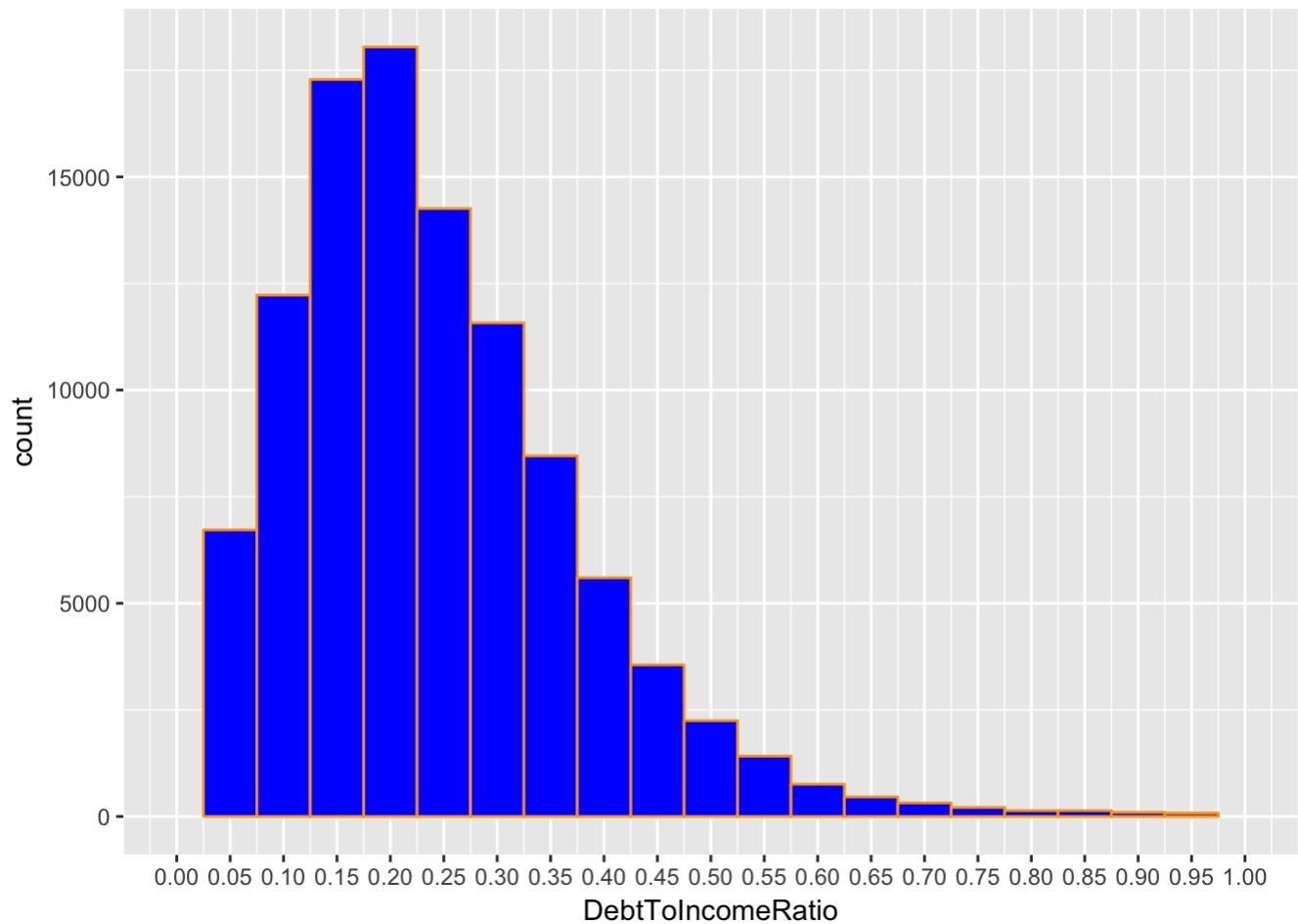
It is not surprising that the majority of the borrowers have regular income from employment. However, the data have ambiguity since employed category can be further divided into part-time and full-time, and these two sub-categories also exist as options.



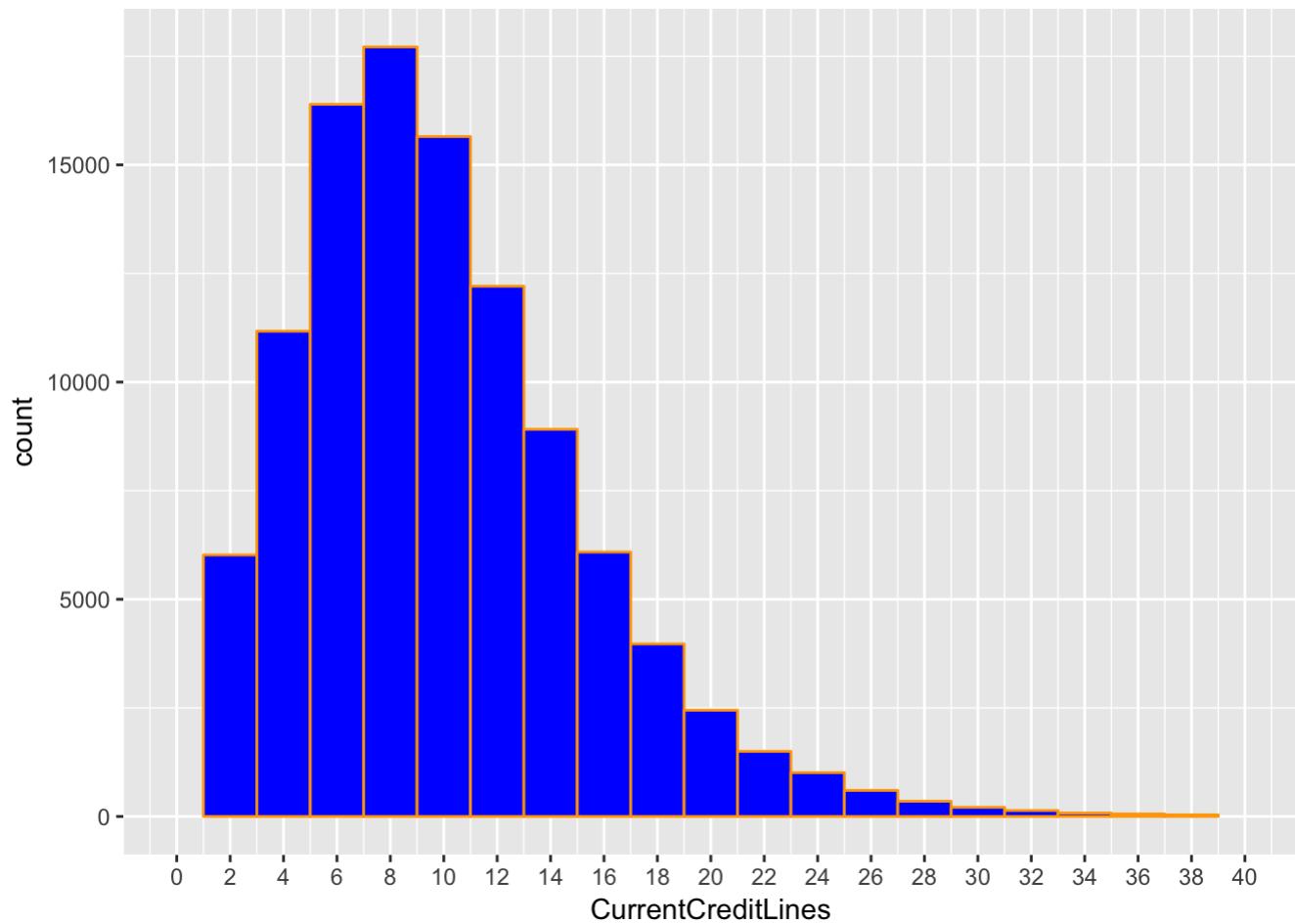
The first Revolving Credit Balance plot indicates positive skew distribution. The Mean is higher than the Median as it is “pulled” to the right. Because the plot is a long tailed one, it is transformed to log10 to better understand the distribution of the data. As we can see from the log the count peaks at around 0 (no revolving balance) and gradually decreases as the revolving balance increases. However, there is no significant difference in the distributions shown by log10, thus it is confirmed that the plot has postive skew distribution.



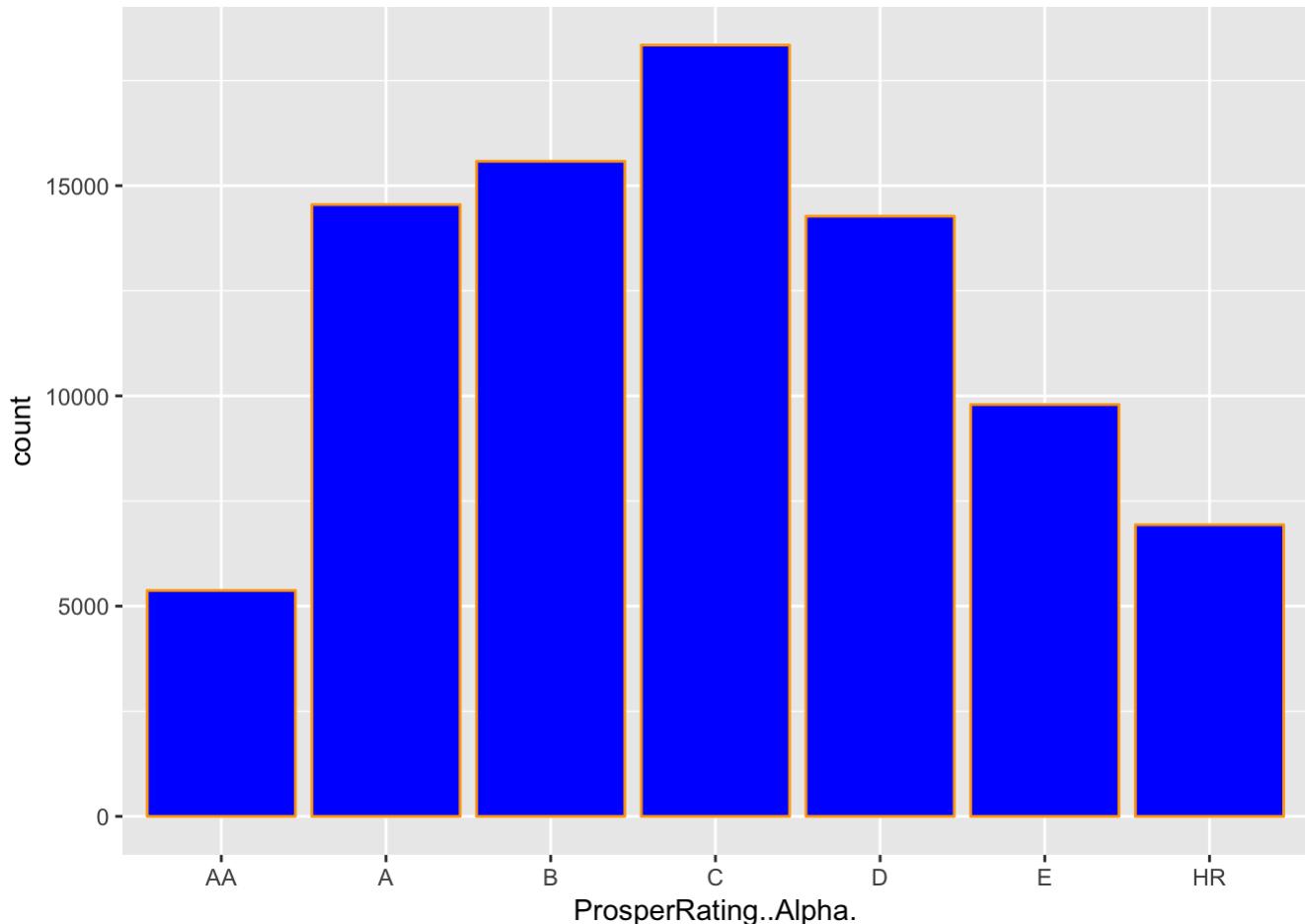
Similar to Revolving Credit Balance, Monthly Loan Payment plot also follows positive skew distribution with Mean value is higher than the Median. Most borrowers have monthly loan payment around \$150.



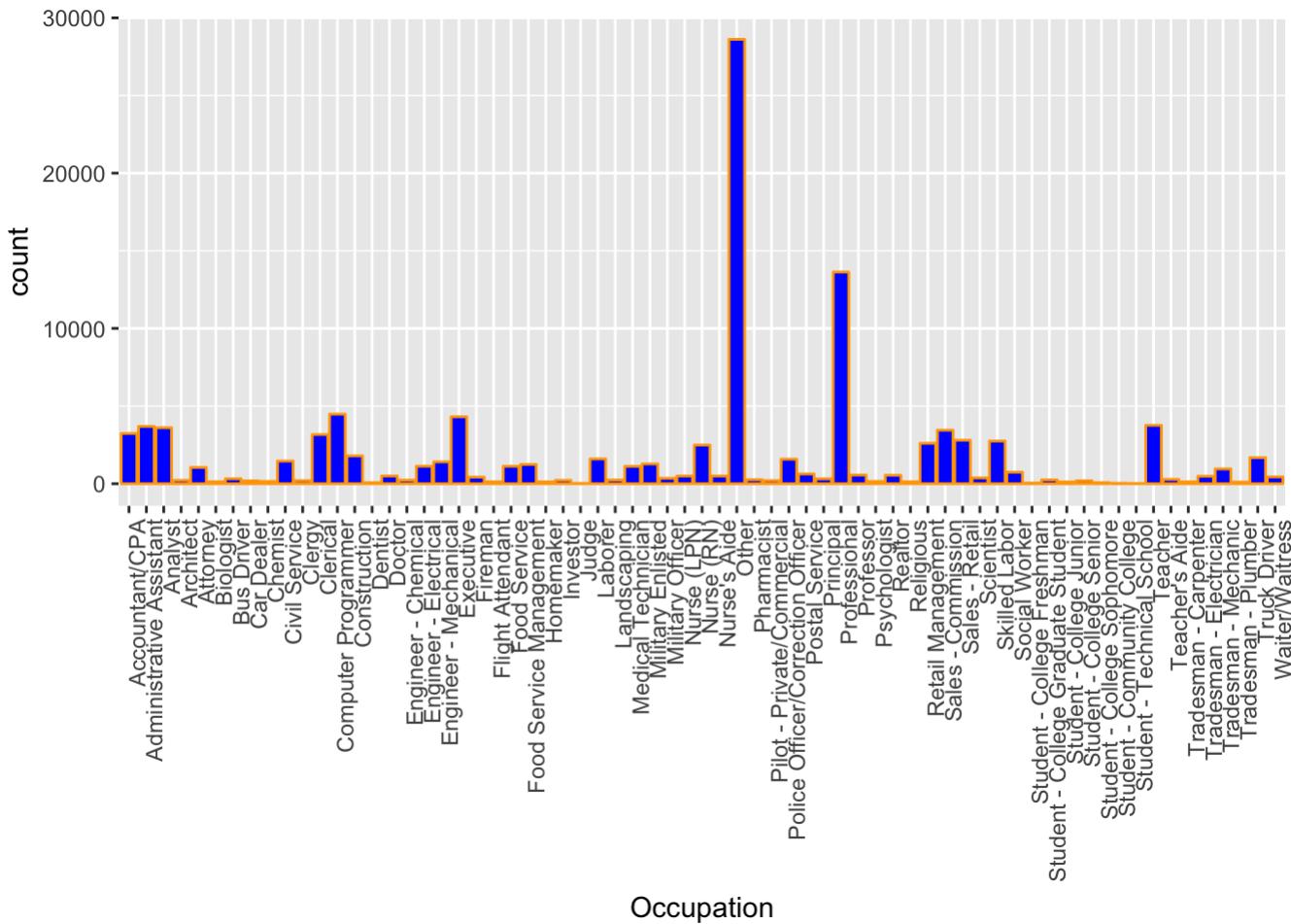
Debt to Income Ratio plot also has positive skew distribution in which majority of borrowers' ratios are around 0.2.



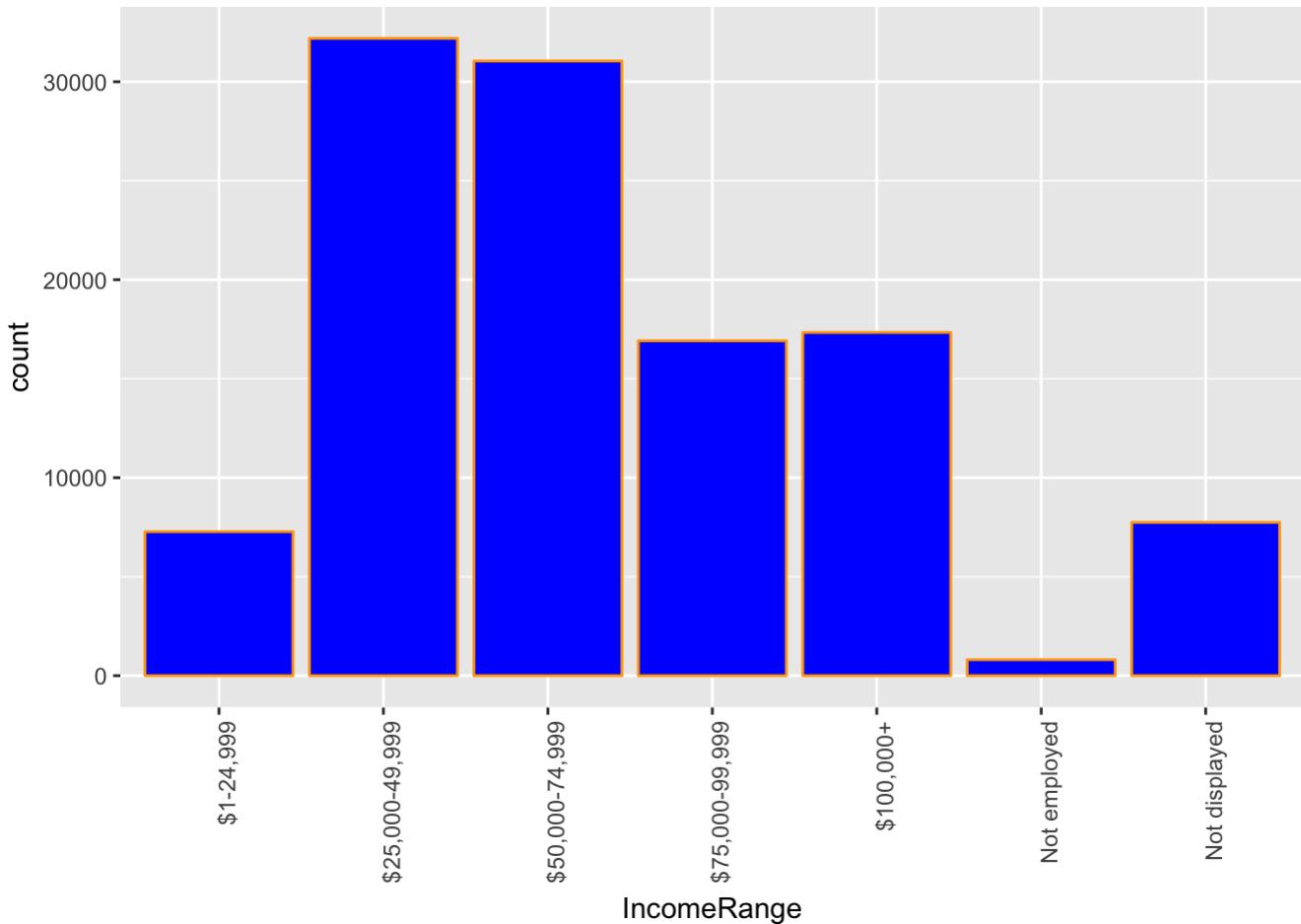
Again, Current Credit Lines plot also shows positive skew distribution. While there are outliers data such as 59 (Max), most borrowers have around 8 credit lines.



ProsperRating..Alpha. plot follows normal distribution. The X axis shows the credit grades from AA, which is the highest credit grade and has the lowest probability of default, to HR, which is the lowest credit grade. HR credit grade also means that there is no credit history or history of defaults. Most debtors have C grade which right in the middle.



In this plot we are looking at how the borrowers are distributed based on their occupations. Unfortunately, by a huge margin, most of the records have “Other”. Prosper needs to improve the data entry to have more specific values. Moreover, the second highest occupation is Professional, which can also be broken down into more specific profession.



Most of Prosper's debtors are in \$25,000-49,999 income range. No one in the data has income greater than \$100,000. There are records having "Not displayed". The distribution of income range might change if values of each Not displayed record is known.

## Univariate Analysis

### What is the structure of your dataset?

There are 113,937 records in the dataset with 81 variables. Variables ProsperRating(Alpha) and IncomeRange are ordered factor variables with the following levels.

Highest to Lowest; ProsperRating(Alpha): "AA", "A", "B", "C", "D", "E", "HR" IncomeRange: "\$100,000+", "\$75,000-99,999", "\$50,000-74,999", "\$25,000-49,999", "\$1-24,999", "\$0".

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

The main features of interest in this dataset are the BorrowerRate, CreditScoreRangeLower and DebtToIncomeRatio.

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

EmploymentStatus, MonthlyLoanPayment and LoanStatus are the features that will be useful during investigation of the features of the interest.

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

I did not create any new variables from the existing variables in the dataset since the existing ones are already self explanatory. Some of the existing variables even have very high correlation, for example BorrowerRate and BorrowerAPR, thus I only use one variable to represent the f.

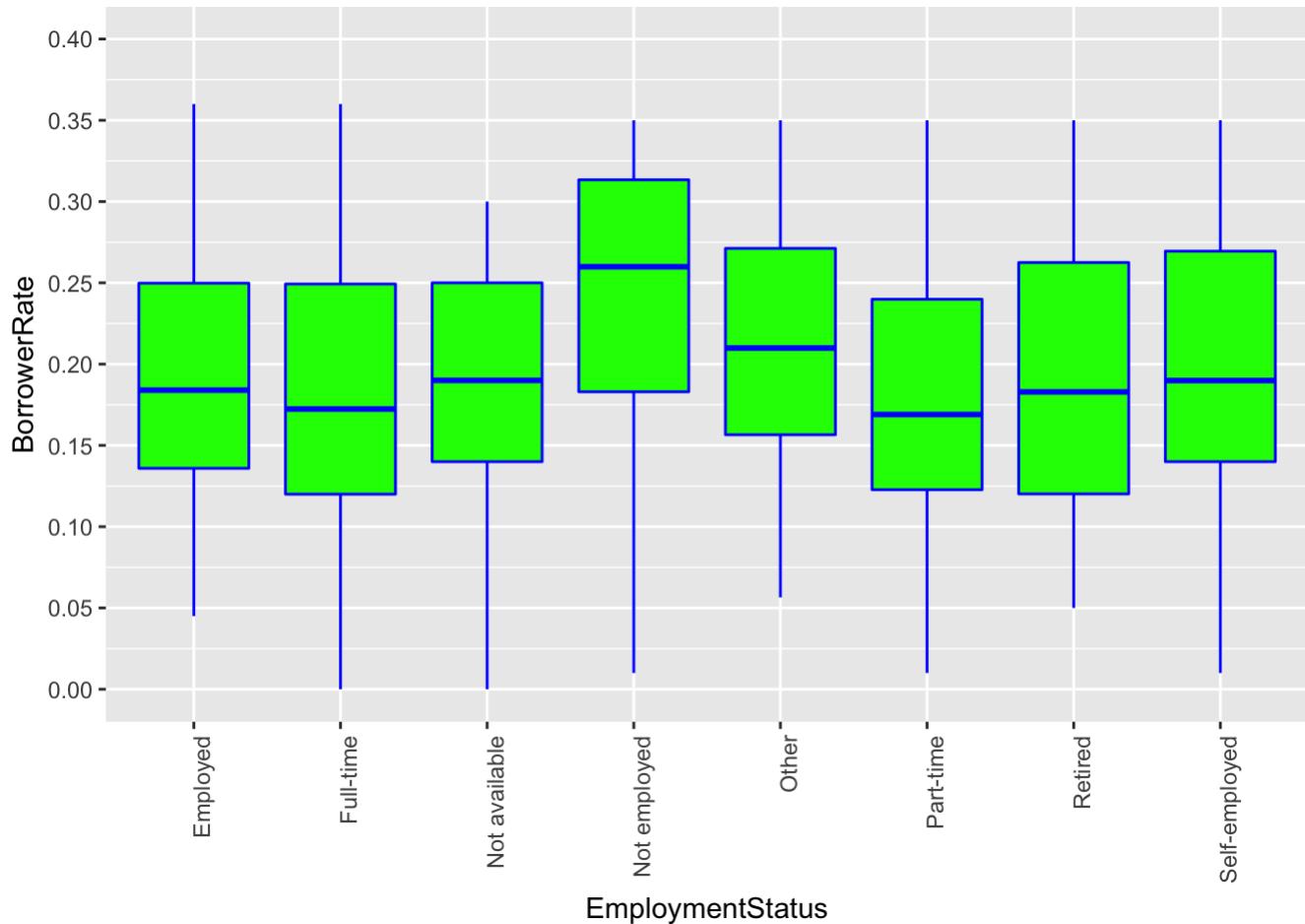
## Of the features you investigated, were there any unusual distributions?

## 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?

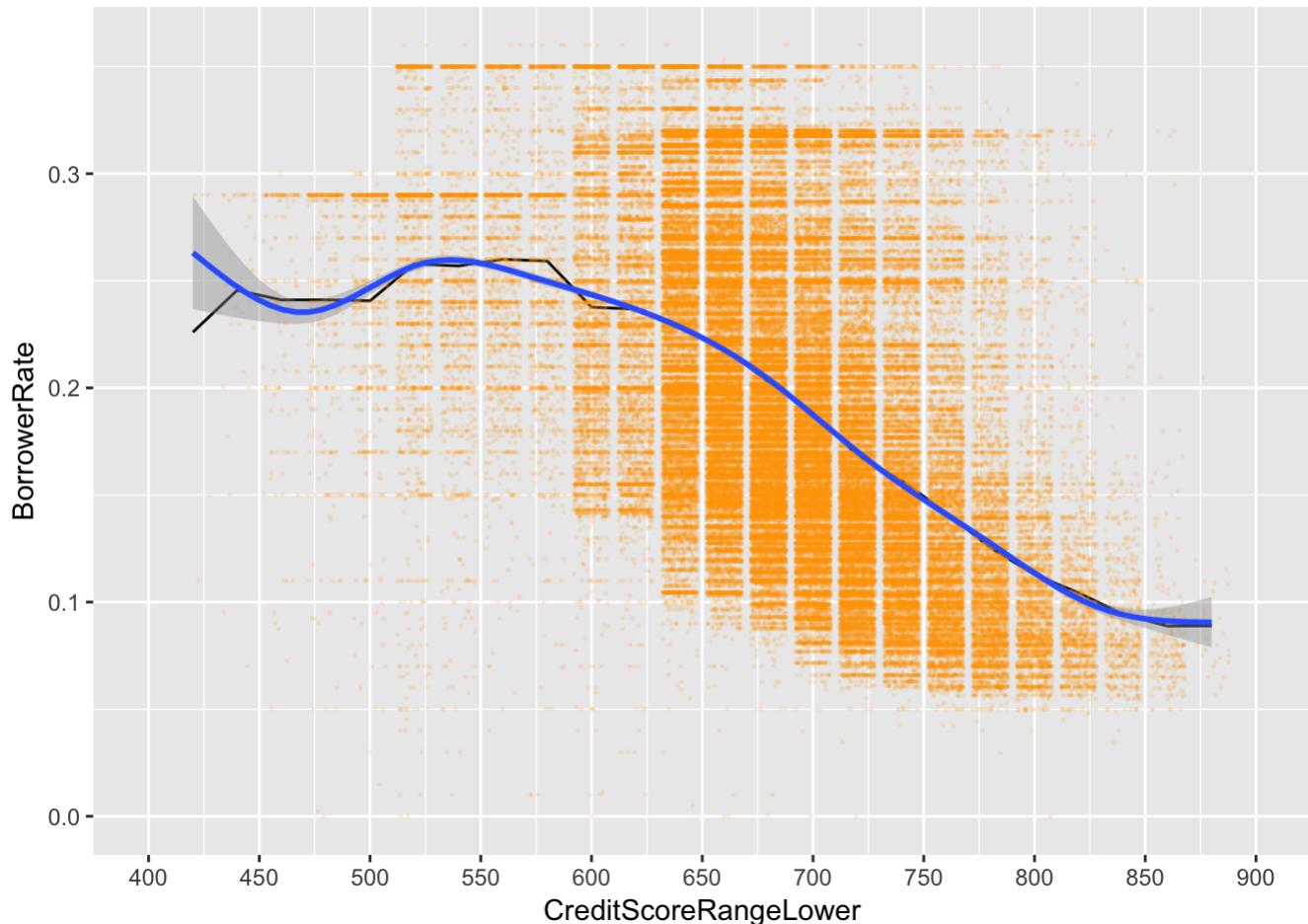
The Revolving Credit Balance plot is a very long tailed one. To see if there is a hidden distribution unseen, the plot is transformed using log10. However, there is no significant difference in the distributions shown by log10, thus the original distribution which is positive skewed is confirmed.

## Bivariate Plots Section



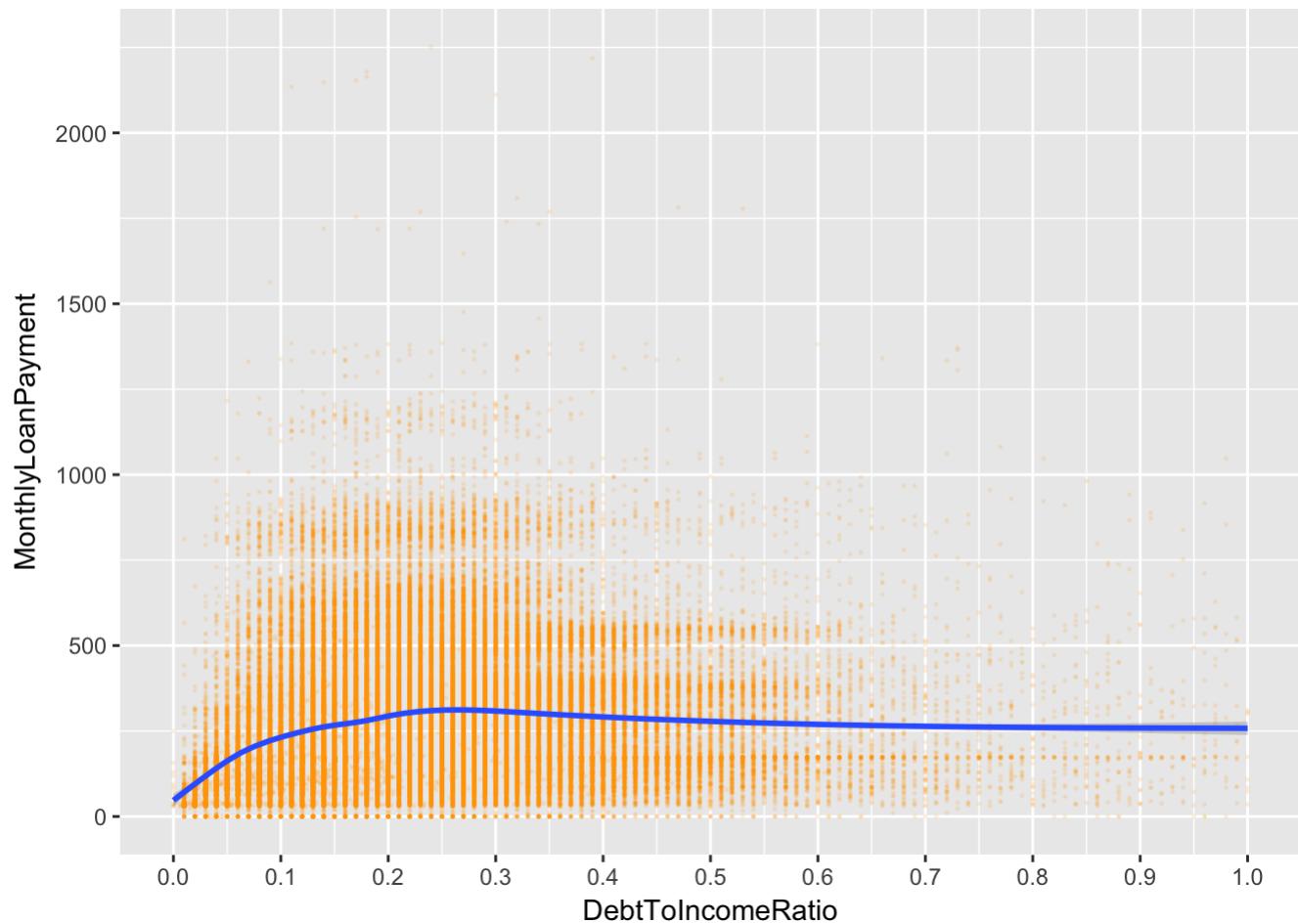


The boxplot above clearly indicates that Not Employed borrowers receive higher loan rate. This is not surprising considering the fact that not employed borrowers have higher risks. On the other hand, full-time and part-time employees have the lowest median loan rate among others.

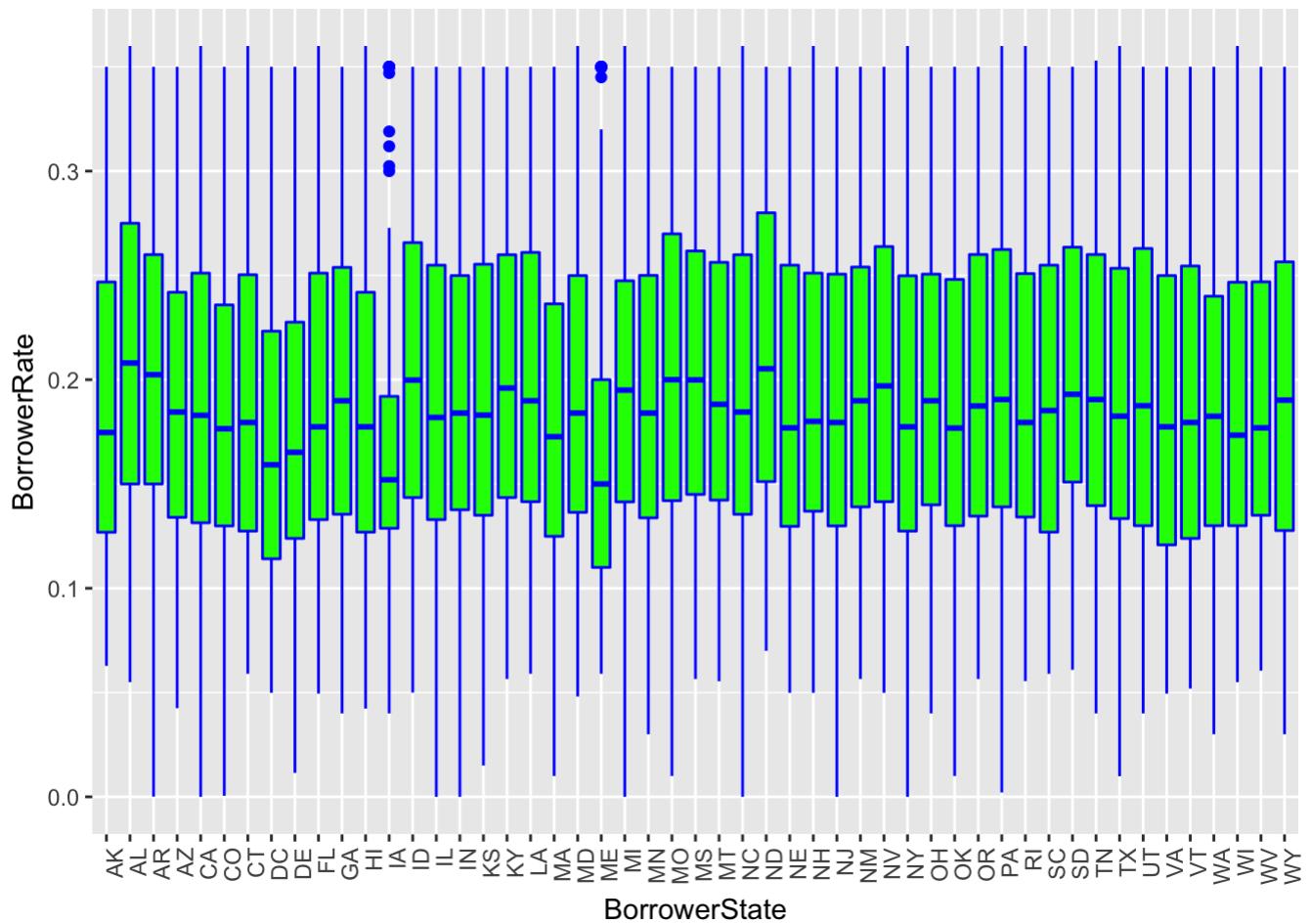
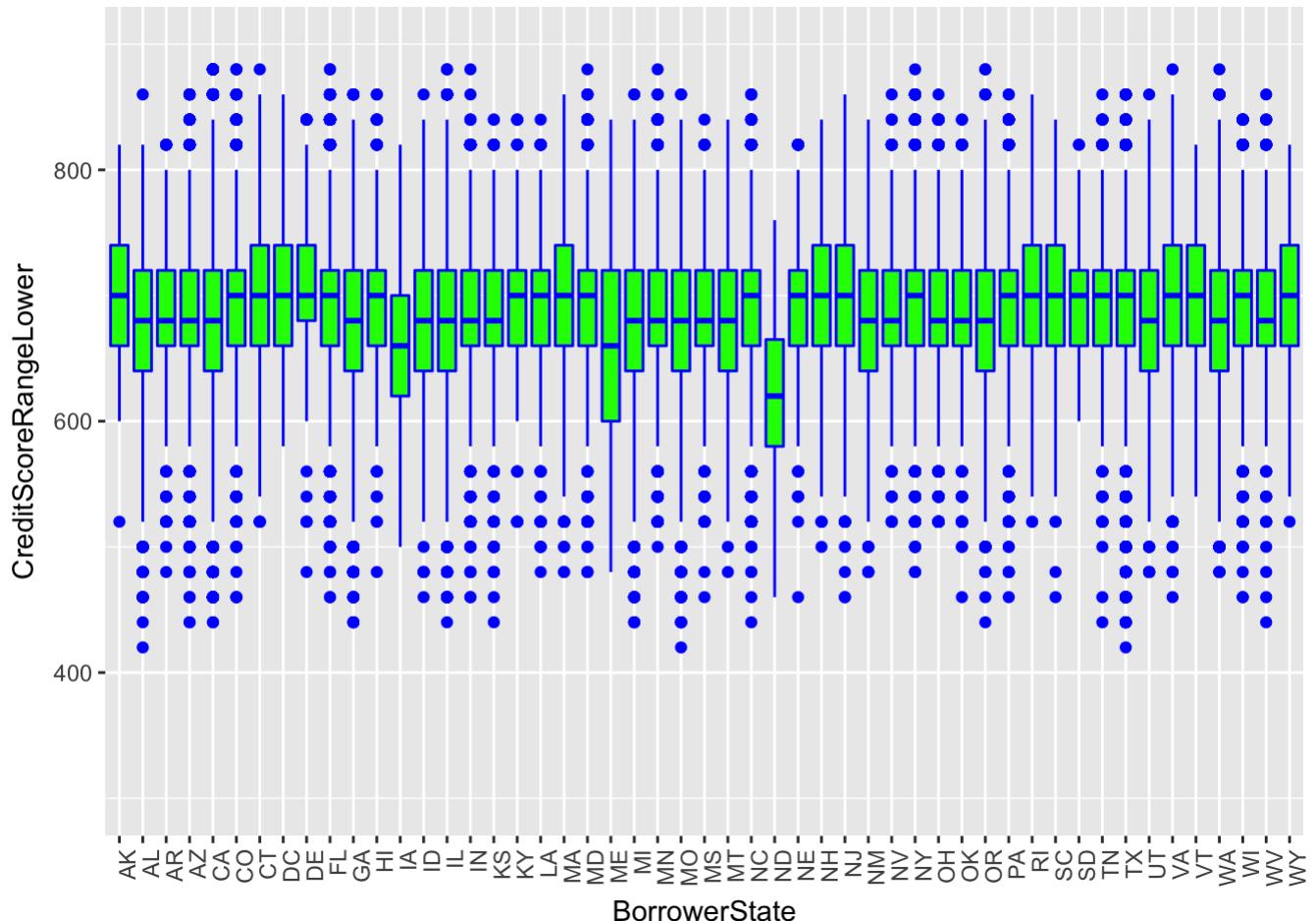


```
## 
## Pearson's product-moment correlation
##
## data: ld$CreditScoreRangeLower and ld$BorrowerRate
## t = -175.17, df = 113340, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## -0.4661358 -0.4569730
## sample estimates:
##      cor
## -0.4615667
```

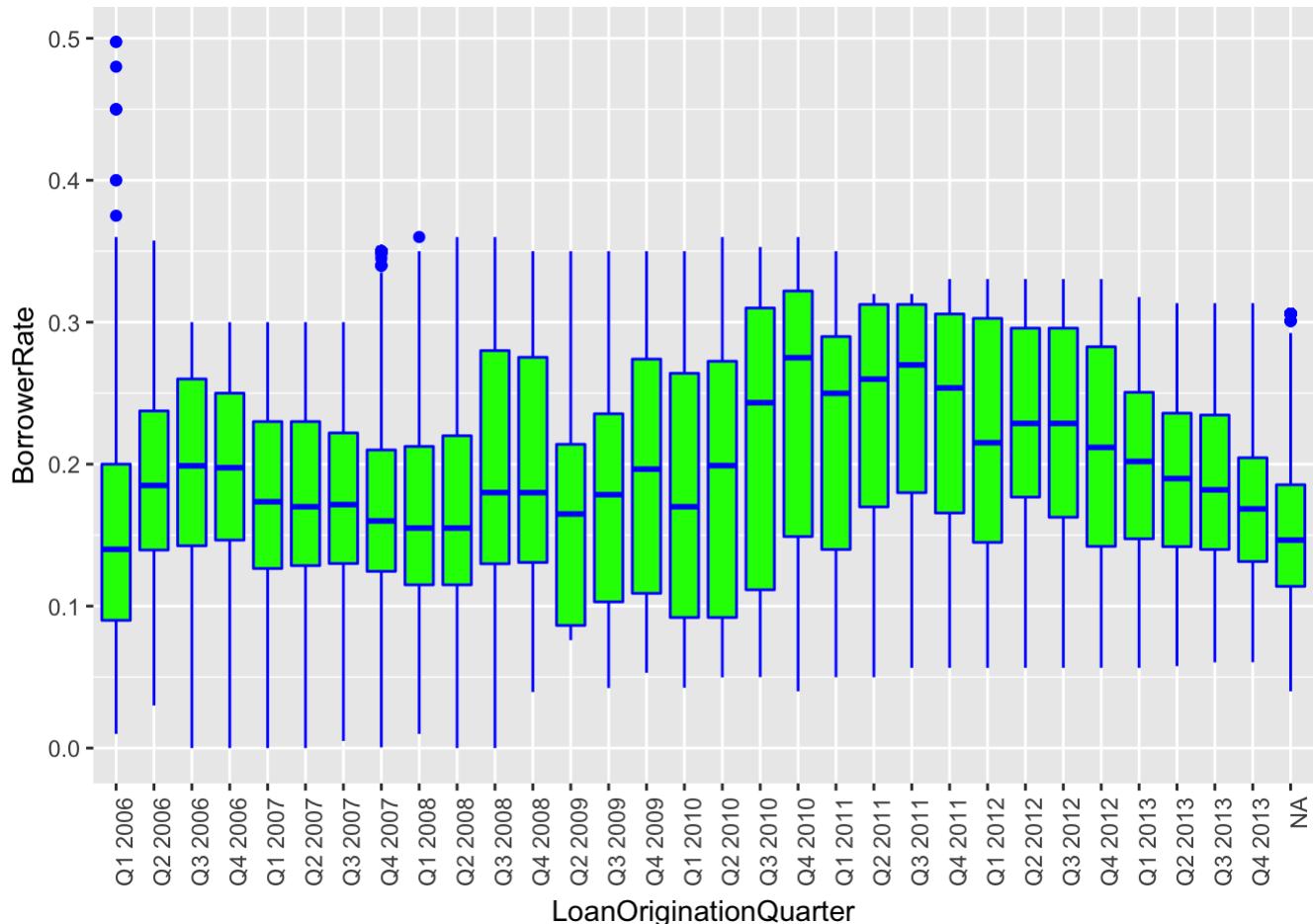
As the borrowers' credit scores higher, the loan rates are lower. This is inline with the assumption that people with higher credit score tend to have lower risk in terms of loan default.



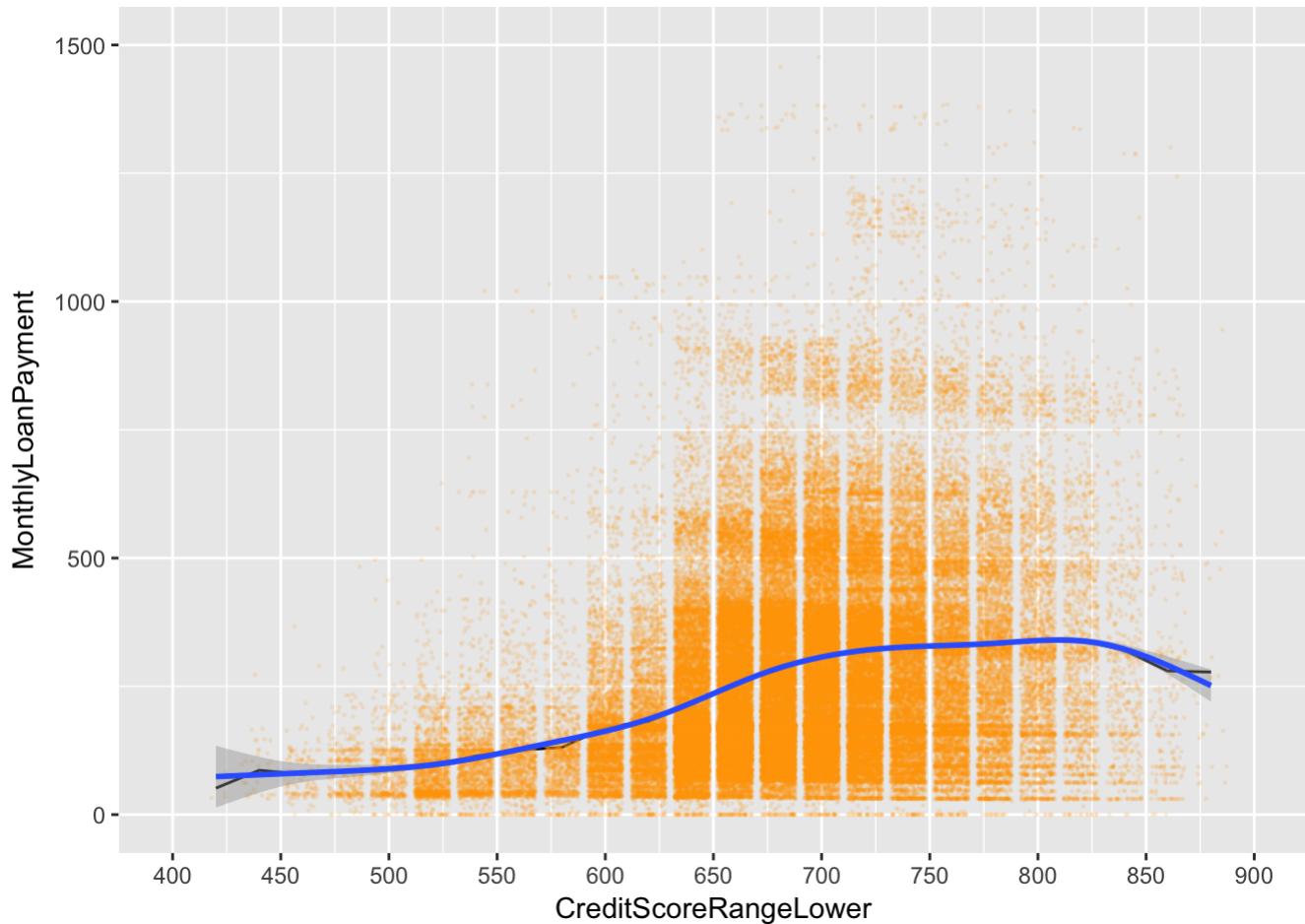
In the plot above the correlation between Debt to Income Ratio and Monthly Loan Payment is investigated. The plot is limited to max debt ratio = 1 since data with ratio greater than 1 is considered as outliers. Plot shows that from ratio 0.05 to 0.4, the monthly loan payment varies greatly. From 0.4 onward, the loan payment variations become more stable with slightly inclining trend.



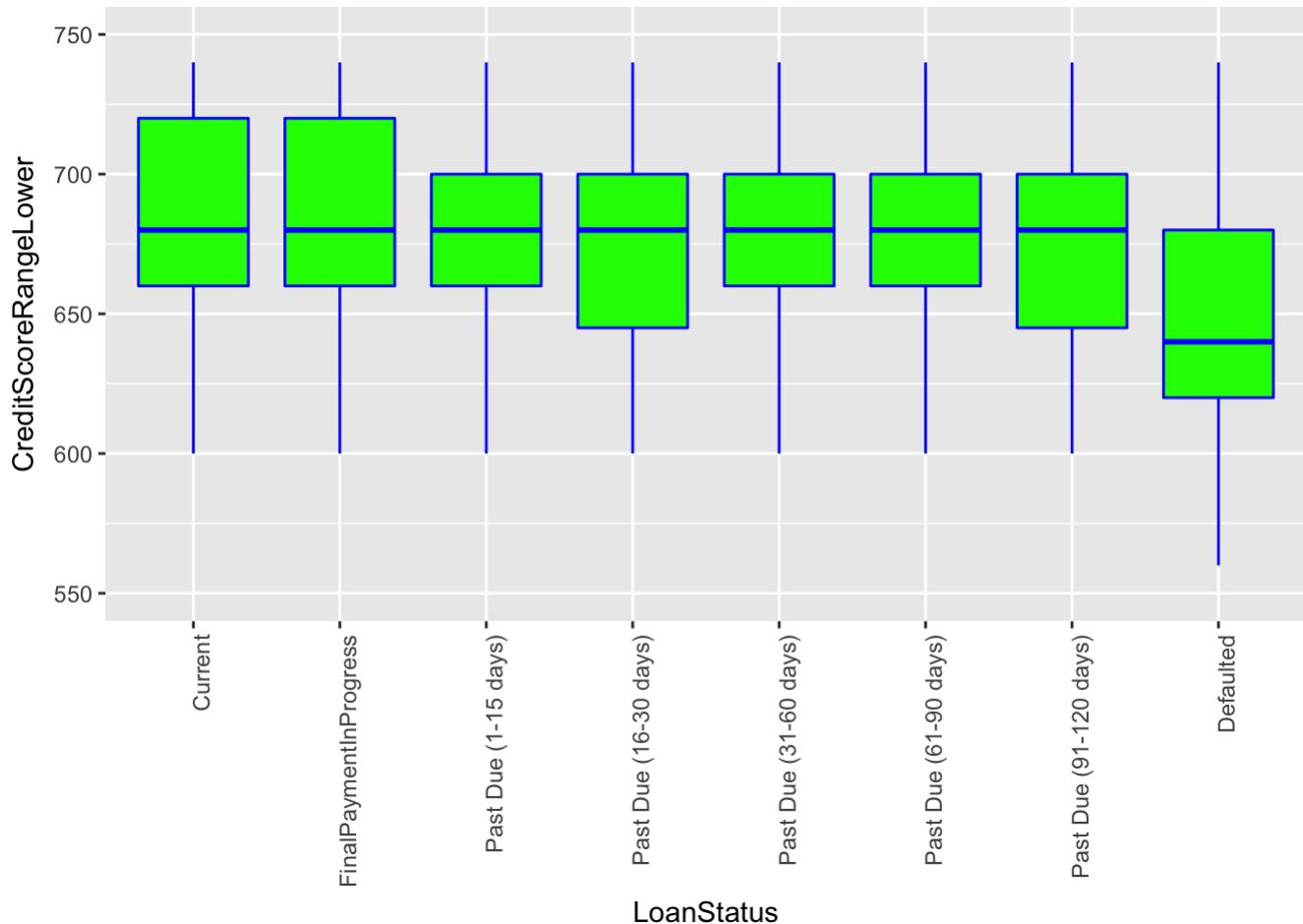
Two boxplots above are created to investigate the correlation between Borrower rate, Credit Score and Borrower State. In the second plot, borrowers from Maine enjoy the lowest rate. Interestingly, the first plot shows that borrowers from Maine have one of the lowest credit score medians among other states. On the other hand, borrowers from North Dakota have the lowest credit score median and consequently receive one of the highest borrower rate medians.



The above boxplot is used to investigate the correlation between borrower rate and when the loan was originated. Q3 2010 - Q4 2011 is the period when the medians of the borrower rate at the highest level. Since Q3 2012 it continuously decline.



Monthly payment values increase as the borrowers credit scores increase gradually. However, at credit score around 825, the monthly payment decreases. One possibility is that the Prosper's customers with credit score of 825 and above tend to have lower loan rate and/or borrow less money, bringing down the monthly payment.



As previously mentioned, there is assumption that borrowers with lower credit score carry higher risk of loan default. The LoanStatus boxplot above confirms this assumption. While there are no significant differences between current and past due loans, defaulted loans indicates that they occur more on borrowers with lower range of credit scores.

## 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?**

Borrower rate correlates with employment status and credit score. Borrowers who have full-time and part-time employment typically have the lowest loan rate. On the opposite, Prosper's customers who are not employed receive the highest rate.

The loan rate is negatively correlates with the credit score. The higher a customer's credit score, the lower the rate she or he gets.

The higher loan rates for the borrowers who have lower credit scores is based on the assumption that the lower credit scores indicate higher risk of the loans being defaulted. As we can see from the last bivariate plot, the median of the defaulted borrowers is significantly lower than other categories.

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

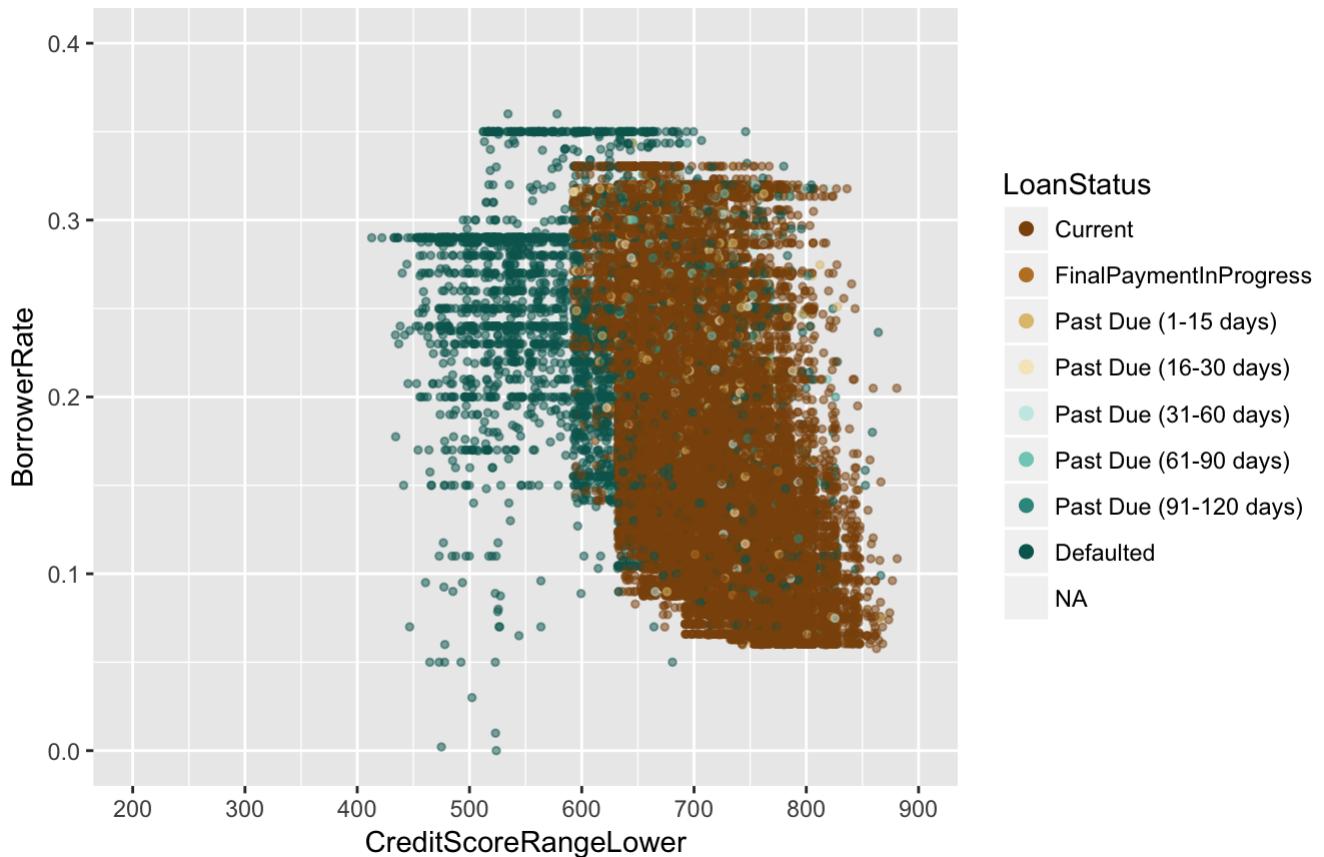
Loans originated from Q3 2010 to Q4 2011 carry highest loan rates in the dataset. However, there is no obvious seasonal pattern, thus we cannot conclude that there is correlation between borrower rate and origination quarter.

## What was the strongest relationship you found?

Borrower rate correlates negatively with the credit score range.

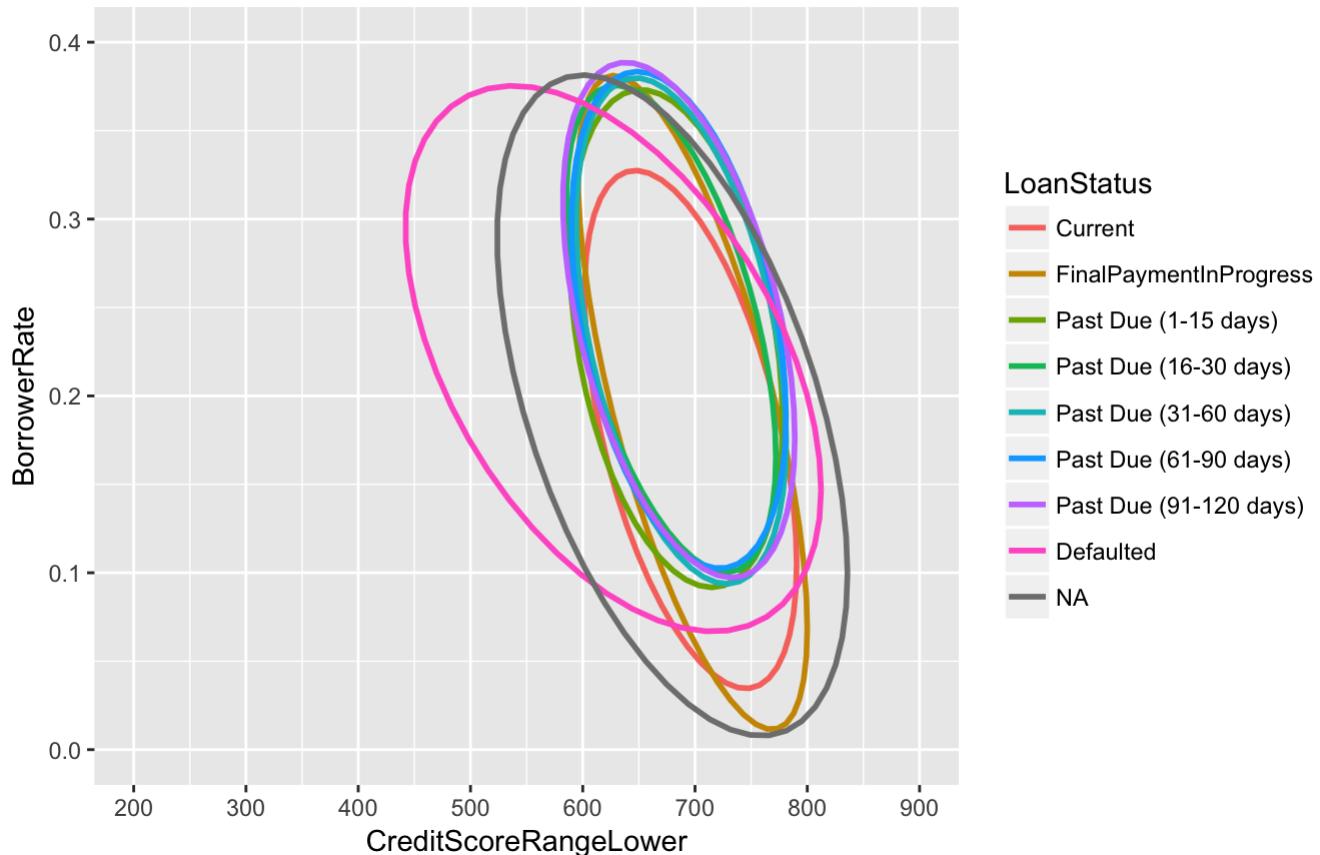
## Multivariate Plots Section

BorrowerRate by CreditScoreRangeLower  
and LoanStatus



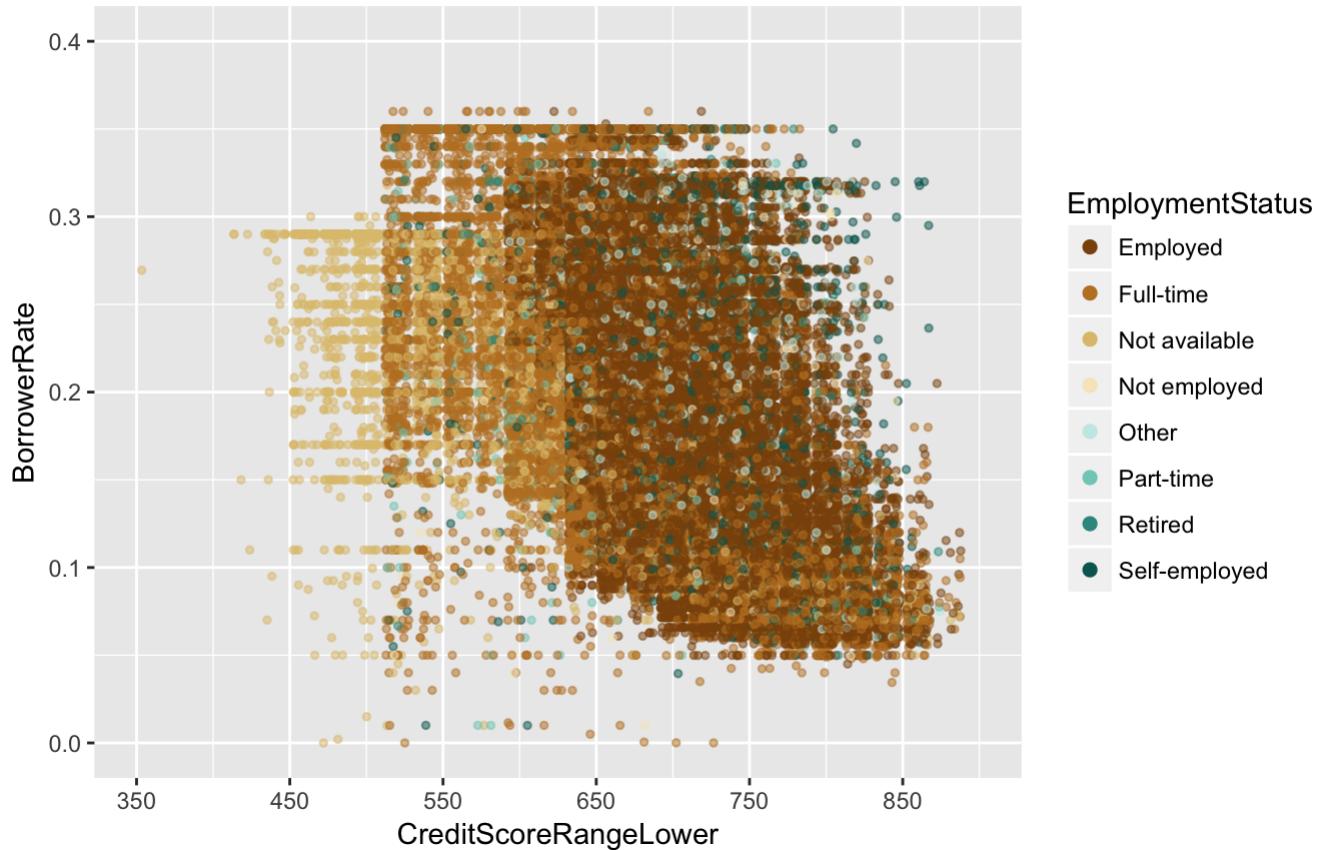
The scatter plot above may seem overplotted, hence we can also use ellipse to depict the operation.

### BorrowerRate by CreditScoreRangeLower and LoanStatus

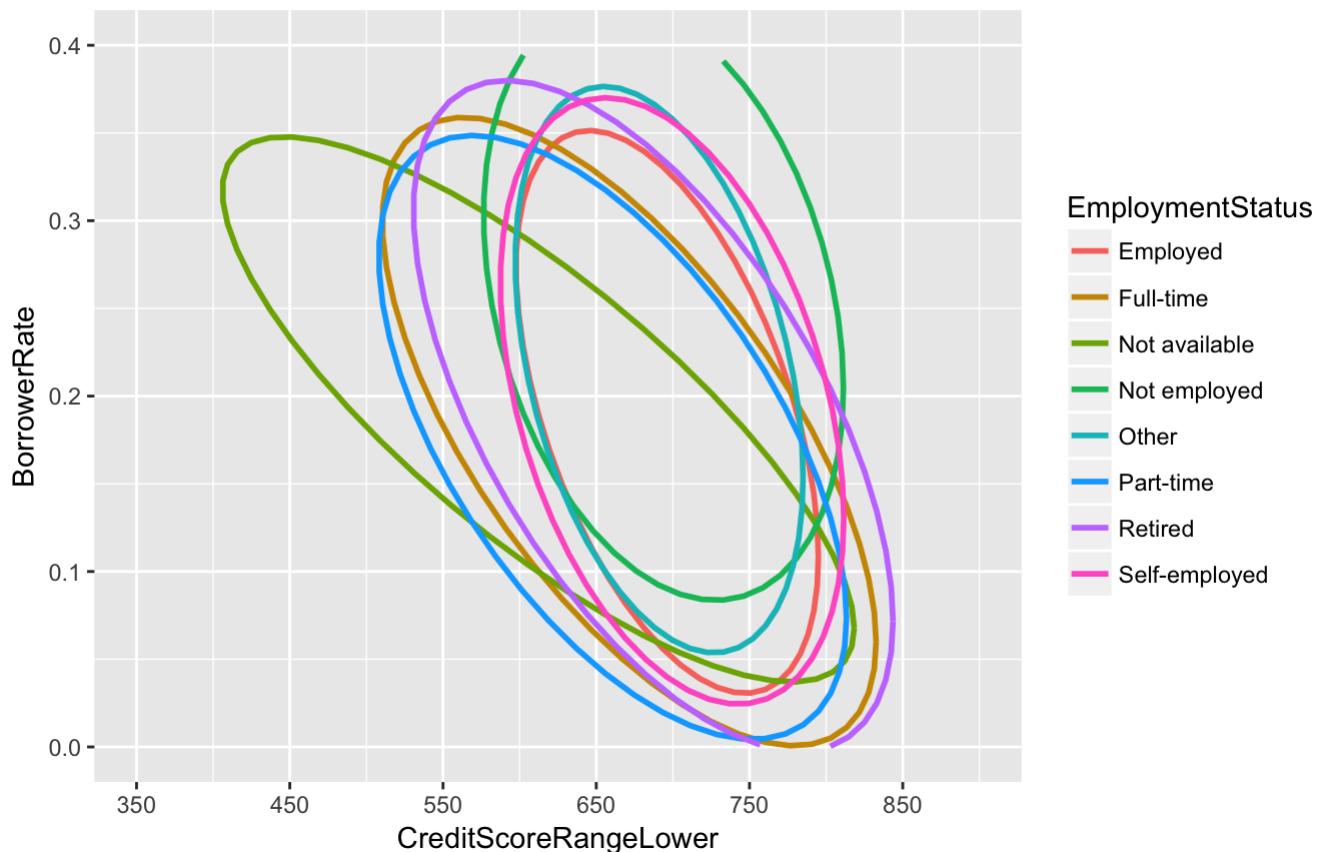


There is no discernible pattern between the rate, credit score and loan status. Most of the loans are in current status and they scatter in all combinations of rate and credit score. One thing to notice is that there are very few loans in past due status for borrowers having credit scores above 600.

### BorrowerRate by CreditScoreRangeLower and EmploymentStatus

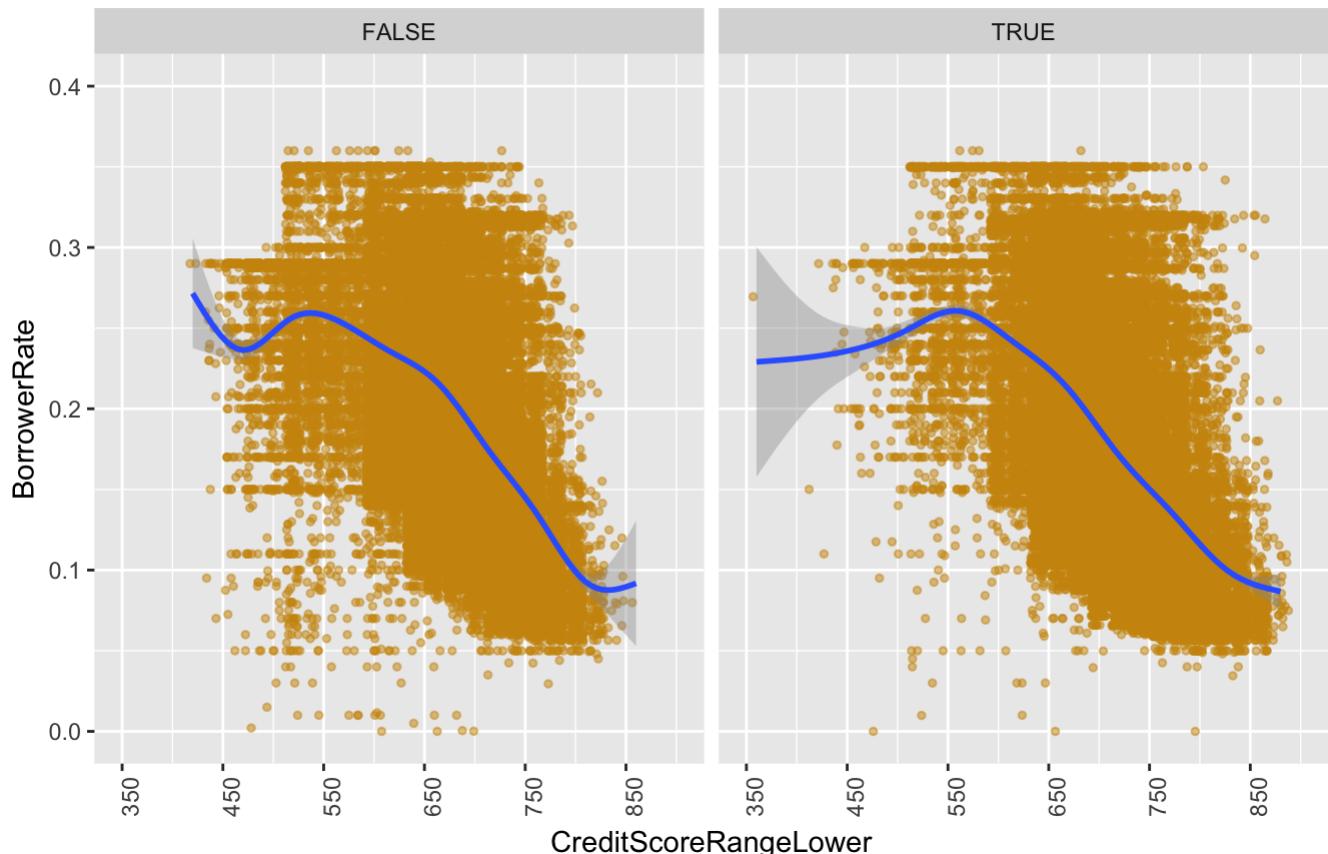


### BorrowerRate by CreditScoreRangeLower and EmploymentStatus



There are two observations from the chart above: 1. Employed and Full-time borrowers dominate the employment status 2. Employed and Full-time borrowers are leaning toward the high end of credit score ranges; however their employment status and credit score don't seem to be correlated with the loan rate.

BorrowerRate by CreditScoreRangeLower  
and Homeowner Status



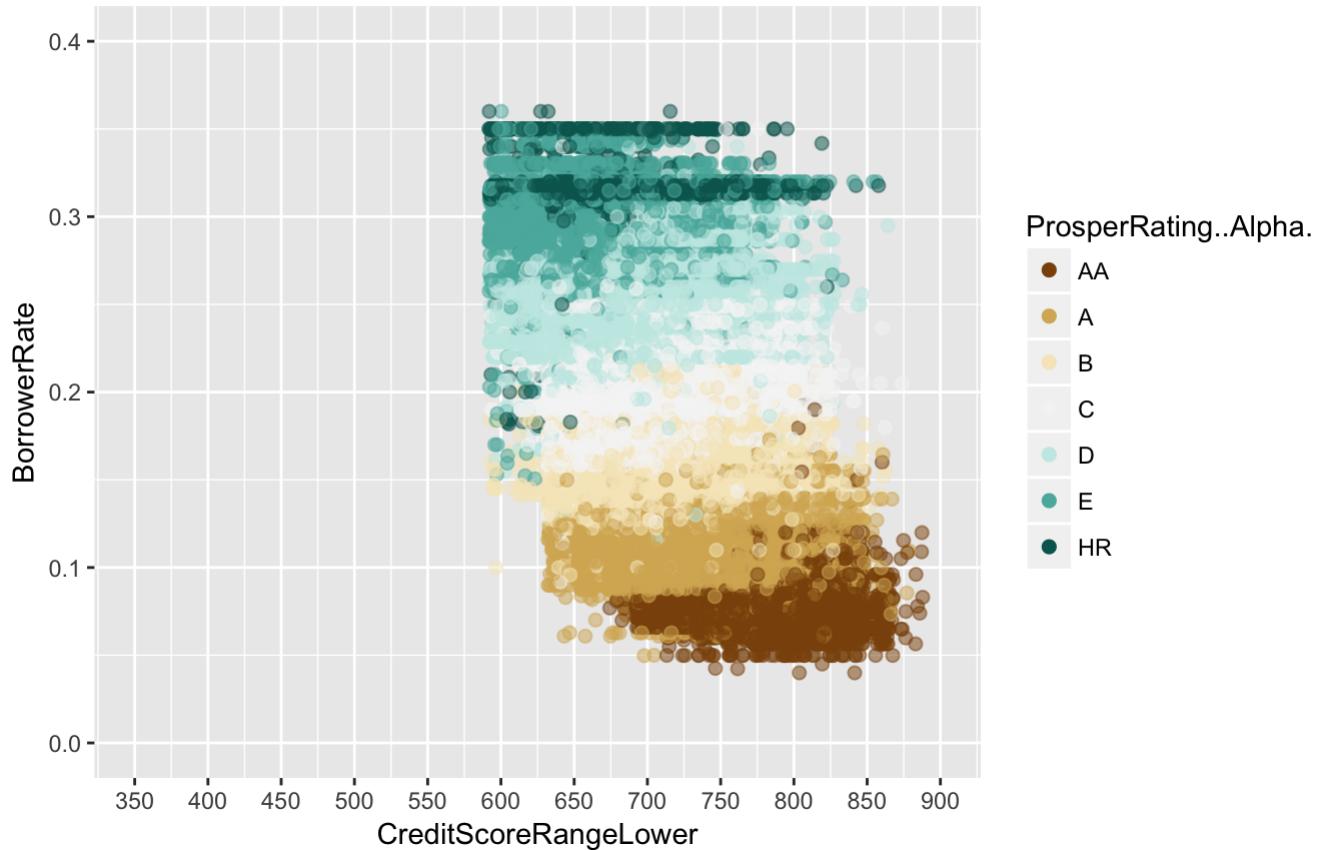
Both of the plots show similar pattern. However, the chart on the left that is for non-homeowners is skinier than the one for the homeowners. We can see that borrower rate is not really affected by the homeowner status, given the same credit score. However, we can also see that homeowners has wider credit score, with majority are in 650-850 (as opposed to 600-800 of the non homeowners)

## BorrowerRate by CreditScoreRangeLower and IncomeRange



Regardless the income ranges, all plots look to have similar pattern. Majority of borrowers are in \$25,000-49,999 and \$50,000-74,999 ranges. Again, we don't see strong correlation between income range and credit score to the borrower rate.

### BorrowerRate by CreditScoreRangeLower and ProsperRating..Alpha.



### BorrowerRate by CreditScoreRangeLower and ProsperRating..numeric.



```
## 
## Pearson's product-moment correlation
##
## data: ld$ProsperRating..numeric. and ld$BorrowerRate
## t = -917.37, df = 84851, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## -0.9537172 -0.9524846
## sample estimates:
## cor
## -0.9531049
```

```
## 
## Call:
## lm(formula = ld$BorrowerRate ~ ld$ProsperRating..numeric.)
## 
## Coefficients:
## (Intercept) ld$ProsperRating..numeric.
##                 0.36914                  -0.04251
```

Customers with the highest Prosper rating (AA and A or 6 and 7 in numerical scale) have the lowest loan rate. Interestingly, the plot shows that even though the customers' credit score vary, as long as they have the highest rating, they receive the lowest borrowing. In this case Prosper rating is better predictor for borrower rate than the credit score alone. The Pearson R test between numerical rating and borrower rate is -0.95, which is very strong.

## 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?**

Low loan rate is strongly correlated with the customer's credit rating assigned by Prosper. Credit score alone is not a determining factor for customers in getting the ideal rate.

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

Income range does not seem to affect the loan rate. Two different customers with same credit score, say 650, and one with \$25,000 income while the other has \$100,000, they could get the same loan rate.

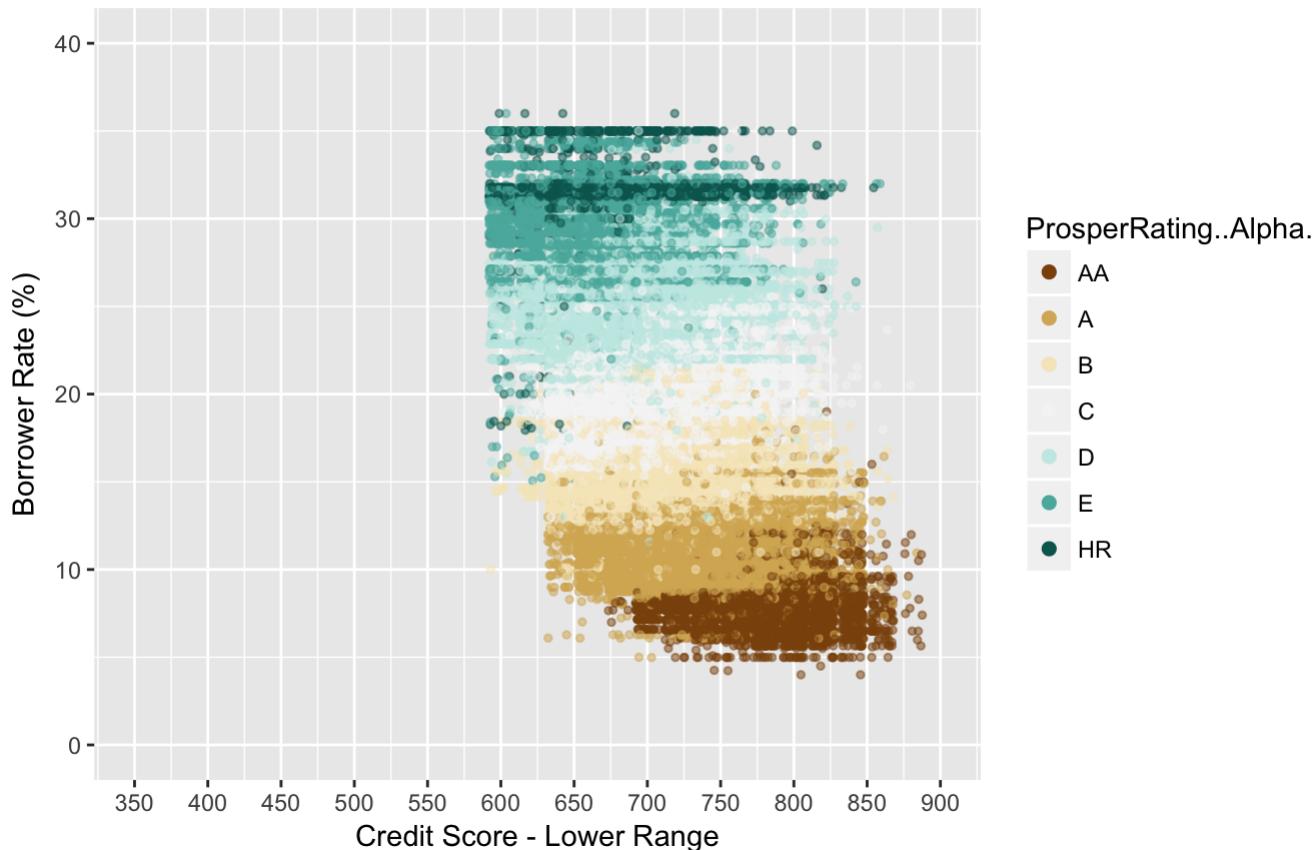
**OPTIONAL: Did you create any models with your dataset? Discuss the strengths and limitations of your model.**

Yes. A linear model is created from ProsperRating..numeric and BorrowerRate variables. The strength of this model is that there is strong correlation between these 2 variables as illustrated by the high R value. On the other side, for new data points, ProsperRating..numeric values may not be assigned immediately as other raw variables, rendering the linear model less effective.

# Final Plots and Summary

## Plot One

BorrowerRate by CreditScoreRangeLower  
and ProsperRating..Alpha.

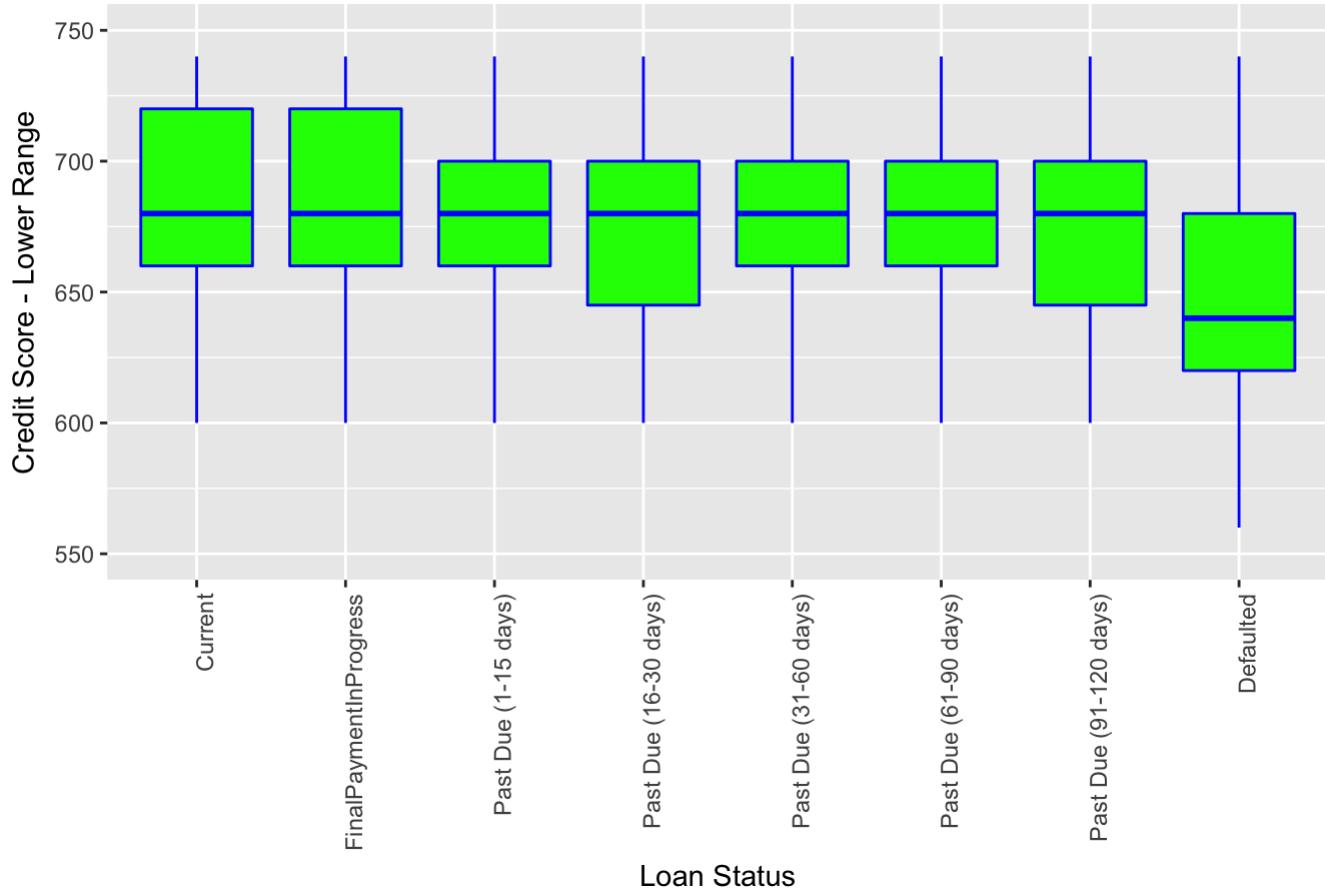


## Description One

The loan rate is shown as strongly correlated with the alphanumeric ProsperRating variable. AA is the highest rate, and HR is the lowest. What interesting with this plot is it shows that the loan rate for each rating is independent of the credit score and the rate depends almost exclusively on the rating. For example, a borrower with credit core of 700 and rating AA may be assigned loan rate between 6-9%. Borrowers with the same credit score but have A rating may have to accept higher rate between 9-14%.

## Plot Two

### Credit Score by Loan Status

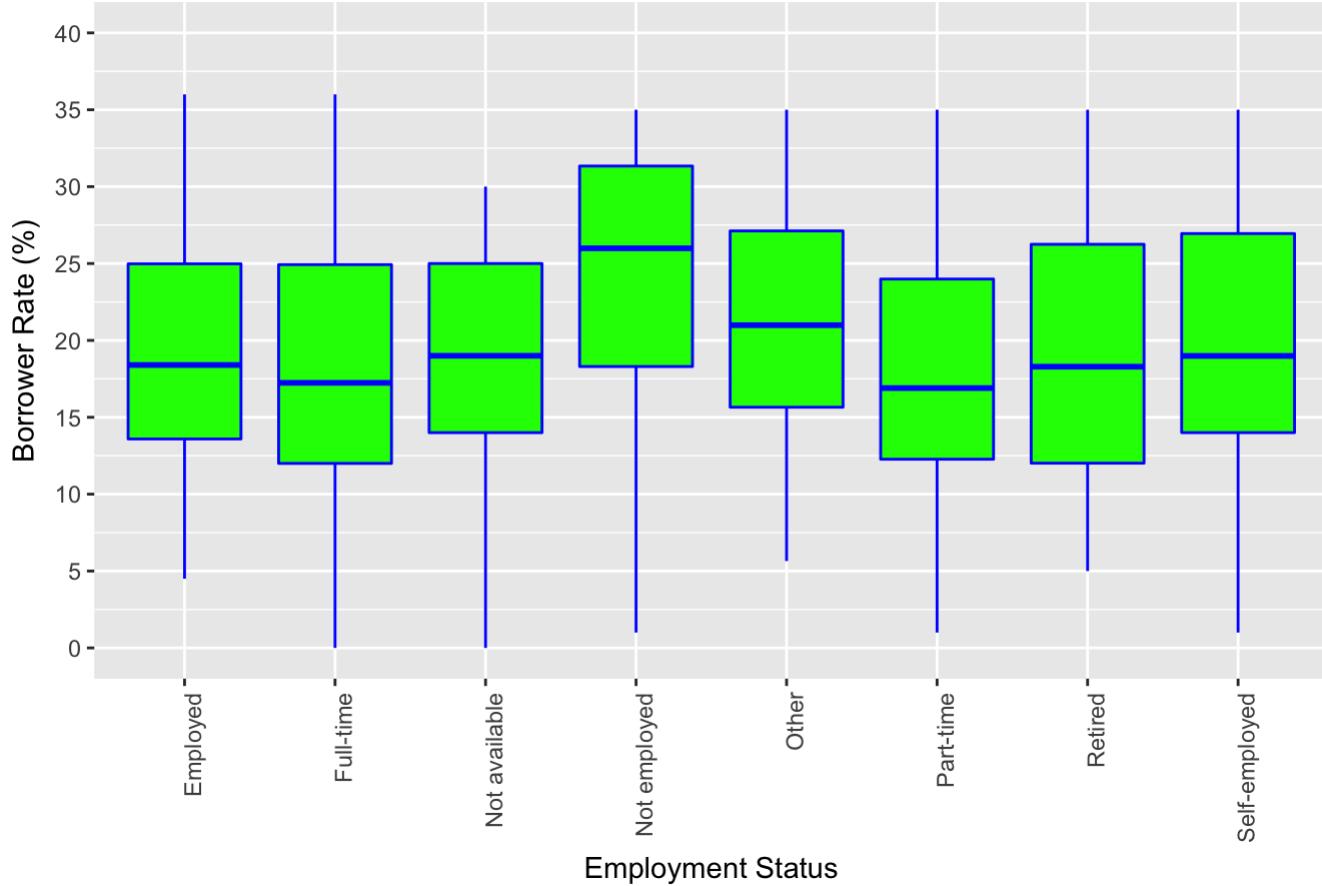


### Description Two

This plot illustrates that defaulted loans are more likely to occur on borrowers with lower credit scores. I choose this plot because it confirms the assumption that borrowers with lower credit scores have higher risks. On the other hand, the past due loans do not differ significantly to the current ones, perhaps due to the fact that most of past due loans will go back to current once the borrower pay the owned payment and only small amount of past due loans actually become defaulted.

### Plot Three

### Borrower Rate by Employment Status



### Description Three

The boxplot above clearly indicates that Not Employed borrowers receive higher loan rate. This is not surprising considering the fact that not employed borrowers have higher risks due to not having regular income. The higher rate is to compensate this higher risk. Other categories having higher borrower rates are Other and Not Available. I suspect these categories are result of Employment Status data are not filled or identified by some borrowers. As employment status is an important consideration when approving the loan, people with unidentified employment status have to settle with higher loan rate.

## Reflection

Prosper loan dataset contains huge amount of variables in which I suspect is a consolidation from various sources of data. Some of the variables though are telling the same thing in slightly different way. Among all variables, I am mainly interested in borrower rate and what are the factors that have impacts on it.

While there is correlation between credit score and employment status to the borrower rate, the strongest ones are shown by the ProsperRating..numeric. and ProsperRating..alpha. variables. However, it is not likely that Prosper derives ProsperRating from other variables by following certain formula, and then use the rating as the direction to assign certain loan rate to their customers.

Some limitations of this model includes the missing of critical variables such as prime interest rate from the dataset. The prime rate is the underlying index for most credit cards, home equity loans and lines of credit, auto loans, and personal loans and it varies over the time. The information regarding the prime rate will be useful when analyzing the loan rate based on the loan originations time.