Exploratory Data Analysis - Prosper Loans (By Kris Harmon)

For this project, we are analyzing the Prosper Loan Data. Prosper is a company that basically matches borrowers with investors. Investors can determine the level / percentage of investment into a loan as they see fit. It contains 113,937 observations of 81 variables. The first thing I want to do is familiarize myself with the data to see where I might want to begin my analysis. I'll first run code to see what the structure of the data is.

Univariate Plots and Analysis Section

Please note that I prefer to iterate my analysis as I go along plotting data via various visualizations. I believe it flows better and helps the reviewer better understand the analysis and reflection occurring at the time the visualizations are being run.

```
## 'data.frame':    113937 obs. of  81 variables:
##  $ ListingKey                       : Factor w/ 113066 levels "00003546482094282EF90E5",..:
7180 7193 6647 6669 6686 6689 6699 6706 6687 6687 ...
##  $ ListingNumber                    : int  193129 1209647 81716 658116 909464 1074836 75089
9 768193 1023355 1023355 ...
##  $ ListingCreationDate              : Factor w/ 113064 levels "2005-11-09 20:44:28.84700000
0",..: 14184 111894 6429 64760 85967 100310 72556 74019 97834 97834 ...
##  $ CreditGrade                      : Factor w/ 9 levels "","A","AA","B",..: 5 1 8 1 1 1 1
1 1 1 ...
##  $ Term                             : int  36 36 36 36 36 60 36 36 36 36 ...
##  $ LoanStatus                       : Factor w/ 12 levels "Cancelled","Chargedoff",..: 3 4
3 4 4 4 4 4 4 4 ...
##  $ ClosedDate                       : Factor w/ 2803 levels "","2005-11-25 00:00:00",..: 11
38 1 1263 1 1 1 1 1 1 1 ...
##  $ 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 ...
##  $ EstimatedEffectiveYield          : num  NA 0.0796 NA 0.0849 0.1832 ...
##  $ EstimatedLoss                    : num  NA 0.0249 NA 0.0249 0.0925 ...
##  $ EstimatedReturn                  : num  NA 0.0547 NA 0.06 0.0907 ...
##  $ ProsperRating..numeric.          : int  NA 6 NA 6 3 5 2 4 7 7 ...
##  $ ProsperRating..Alpha.            : Factor w/ 8 levels "","A","AA","B",..: 1 2 1 2 6 4 7
5 3 3 ...
##  $ ProsperScore                     : num  NA 7 NA 9 4 10 2 4 9 11 ...
##  $ ListingCategory..numeric.        : int  0 2 0 16 2 1 1 2 7 7 ...
##  $ BorrowerState                    : Factor w/ 52 levels "","AK","AL","AR",..: 7 7 12 12 2
5 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 ...
##  $ EmploymentStatusDuration         : int  2 44 NA 113 44 82 172 103 269 269 ...
##  $ IsBorrowerHomeowner              : Factor w/ 2 levels "False","True": 2 1 1 2 2 2 1 1 2
2 ...
##  $ CurrentlyInGroup                 : Factor w/ 2 levels "False","True": 2 1 2 1 1 1 1 1 1
1 ...
##  $ GroupKey                         : Factor w/ 707 levels "","0034337690131242316873 1",..:
1 1 335 1 1 1 1 1 1 1 ...
##  $ DateCreditPulled                 : Factor w/ 112992 levels "2005-11-09 00:30:04.48700000
0",..: 14347 111883 6446 64724 85857 100382 72500 73937 97888 97888 ...
##  $ CreditScoreRangeLower            : int  640 680 480 800 680 740 680 700 820 820 ...
##  $ CreditScoreRangeUpper            : int  659 699 499 819 699 759 699 719 839 839 ...
##  $ FirstRecordedCreditLine          : Factor w/ 11586 levels "","1947-08-24 00:00:00",..: 8
639 6617 8927 2247 9498 497 8265 7685 5543 5543 ...
##  $ CurrentCreditLines               : int  5 14 NA 5 19 21 10 6 17 17 ...
##  $ OpenCreditLines                  : int  4 14 NA 5 19 17 7 6 16 16 ...
##  $ TotalCreditLinespast7years       : int  12 29 3 29 49 49 20 10 32 32 ...
##  $ OpenRevolvingAccounts            : int  1 13 0 7 6 13 6 5 12 12 ...
##  $ OpenRevolvingMonthlyPayment      : num  24 389 0 115 220 1410 214 101 219 219 ...
##  $ InquiriesLast6Months             : int  3 3 0 0 1 0 0 3 1 1 ...
##  $ TotalInquiries                   : num  3 5 1 1 9 2 0 16 6 6 ...
##  $ CurrentDelinquencies             : int  2 0 1 4 0 0 0 0 0 0 ...
##  $ AmountDelinquent                 : num  472 0 NA 10056 0 ...
```

```
##  $ DelinquenciesLast7Years           : int  4 0 0 14 0 0 0 0 0 0 ...
##  $ PublicRecordsLast10Years          : int  0 1 0 0 0 0 0 1 0 0 ...
##  $ PublicRecordsLast12Months         : int  0 0 NA 0 0 0 0 0 0 0 ...
##  $ RevolvingCreditBalance            : num  0 3989 NA 1444 6193 ...
##  $ 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 ...
##  $ TotalTrades                       : num  11 29 NA 26 39 47 16 10 29 29 ...
##  $ TradesNeverDelinquent..percentage. : num  0.81 1 NA 0.76 0.95 1 0.68 0.8 1 1 ...
##  $ TradesOpenedLast6Months           : num  0 2 NA 0 2 0 0 0 1 1 ...
##  $ DebtToIncomeRatio                 : num  0.17 0.18 0.06 0.15 0.26 0.36 0.27 0.24 0.25 0.2
5 ...
##  $ 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 ...
##  $ StatedMonthlyIncome               : num  3083 6125 2083 2875 9583 ...
##  $ LoanKey                           : Factor w/ 113066 levels "00003683605746079487FF7",..:
100337 69837 46303 70776 71387 86505 91250 5425 908 908 ...
##  $ TotalProsperLoans                 : int  NA NA NA NA 1 NA NA NA NA NA ...
##  $ TotalProsperPaymentsBilled        : int  NA NA NA NA 11 NA NA NA NA NA ...
##  $ OnTimeProsperPayments             : int  NA NA NA NA 11 NA NA NA NA NA ...
##  $ ProsperPaymentsLessThanOneMonthLate: int  NA NA NA NA 0 NA NA NA NA NA ...
##  $ ProsperPaymentsOneMonthPlusLate   : int  NA NA NA NA 0 NA NA NA NA NA ...
##  $ ProsperPrincipalBorrowed          : num  NA NA NA NA 11000 NA NA NA NA NA ...
##  $ ProsperPrincipalOutstanding       : num  NA NA NA NA 9948 ...
##  $ ScorexChangeAtTimeOfListing       : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ LoanCurrentDaysDelinquent         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ LoanFirstDefaultedCycleNumber     : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ LoanMonthsSinceOrigination        : int  78 0 86 16 6 3 11 10 3 3 ...
##  $ LoanNumber                        : int  19141 134815 6466 77296 102670 123257 88353 9005
1 121268 121268 ...
##  $ LoanOriginalAmount                : int  9425 10000 3001 10000 15000 15000 3000 10000 100
00 10000 ...
##  $ LoanOriginationDate               : Factor w/ 1873 levels "2005-11-15 00:00:00",..: 426 1
866 260 1535 1757 1821 1649 1666 1813 1813 ...
##  $ LoanOriginationQuarter            : Factor w/ 33 levels "Q1 2006","Q1 2007",..: 18 8 2 32
24 33 16 16 33 33 ...
##  $ MemberKey                         : Factor w/ 90831 levels "00003397697413387CAF966",..:
11071 10302 33781 54939 19465 48037 60448 40951 26129 26129 ...
##  $ MonthlyLoanPayment                : num  330 319 123 321 564 ...
##  $ LP_CustomerPayments               : num  11396 0 4187 5143 2820 ...
##  $ LP_CustomerPrincipalPayments      : num  9425 0 3001 4091 1563 ...
##  $ LP_InterestandFees                : num  1971 0 1186 1052 1257 ...
##  $ LP_ServiceFees                    : num  -133.2 0 -24.2 -108 -60.3 ...
##  $ LP_CollectionFees                 : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ LP_GrossPrincipalLoss             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ LP_NetPrincipalLoss               : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ LP_NonPrincipalRecoverypayments   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ PercentFunded                     : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ Recommendations                   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ InvestmentFromFriendsCount        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ InvestmentFromFriendsAmount       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Investors                         : int  258 1 41 158 20 1 1 1 1 1 ...
```

Now, I want to see a summary of the data to see if anything stands out to me.

```
##                   ListingKey     ListingNumber
##  17A93590655669644DB4C06:     6   Min.   :      4
##  349D3587495831350F0F648:     4   1st Qu.: 400919
##  47C135963849743197567OB:     4   Median : 600554
##  8474358854651984137201C:     4   Mean   : 627886
##  DE8535960513435199406CE:     4   3rd Qu.: 892634
##  04C13599434217079754AEE:     3   Max.   :1255725
##  (Other)                :113912
##                    ListingCreationDate  CreditGrade         Term
##  2013-10-02 17:20:16.550000000:     6          :84984   Min.   :12.00
##  2013-08-28 20:31:41.107000000:     4   C      : 5649   1st Qu.:36.00
##  2013-09-08 09:27:44.853000000:     4   D      : 5153   Median :36.00
##  2013-12-06 05:43:13.830000000:     4   B      : 4389   Mean   :40.83
##  2013-12-06 11:44:58.283000000:     4   AA     : 3509   3rd Qu.:36.00
##  2013-08-21 07:25:22.360000000:     3   HR     : 3508   Max.   :60.00
##  (Other)                       :113912   (Other): 6745
##                 LoanStatus                  ClosedDate
##  Current              :56576                      :58848
##  Completed            :38074   2014-03-04 00:00:00:  105
##  Chargedoff           :11992   2014-02-19 00:00:00:  100
##  Defaulted            : 5018   2014-02-11 00:00:00:   92
##  Past Due (1-15 days) :  806   2012-10-30 00:00:00:   81
##  Past Due (31-60 days):  363   2013-02-26 00:00:00:   78
##  (Other)              : 1108   (Other)            :54633
##   BorrowerAPR       BorrowerRate     LenderYield
##  Min.   :0.00653   Min.   :0.0000   Min.   :-0.0100
##  1st Qu.:0.15629   1st Qu.:0.1340   1st Qu.: 0.1242
##  Median :0.20976   Median :0.1840   Median : 0.1730
##  Mean   :0.21883   Mean   :0.1928   Mean   : 0.1827
##  3rd Qu.:0.28381   3rd Qu.:0.2500   3rd Qu.: 0.2400
##  Max.   :0.51229   Max.   :0.4975   Max.   : 0.4925
##  NA's   :25
##  EstimatedEffectiveYield EstimatedLoss   EstimatedReturn
##  Min.   :-0.183          Min.   :0.005   Min.   :-0.183
##  1st Qu.: 0.116          1st Qu.:0.042   1st Qu.: 0.074
##  Median : 0.162          Median :0.072   Median : 0.092
##  Mean   : 0.169          Mean   :0.080   Mean   : 0.096
##  3rd Qu.: 0.224          3rd Qu.:0.112   3rd Qu.: 0.117
##  Max.   : 0.320          Max.   :0.366   Max.   : 0.284
##  NA's   :29084           NA's   :29084   NA's   :29084
##  ProsperRating..numeric. ProsperRating..Alpha.  ProsperScore
##  Min.   :1.000                  :29084         Min.   : 1.00
##  1st Qu.:3.000           C      :18345         1st Qu.: 4.00
##  Median :4.000           B      :15581         Median : 6.00
##  Mean   :4.072           A      :14551         Mean   : 5.95
##  3rd Qu.:5.000           D      :14274         3rd Qu.: 8.00
##  Max.   :7.000           E      : 9795         Max.   :11.00
##  NA's   :29084           (Other):12307         NA's   :29084
##  ListingCategory..numeric. BorrowerState
##  Min.   : 0.000            CA     :14717
##  1st Qu.: 1.000            TX     : 6842
##  Median : 1.000            NY     : 6729
##  Mean   : 2.774            FL     : 6720
```

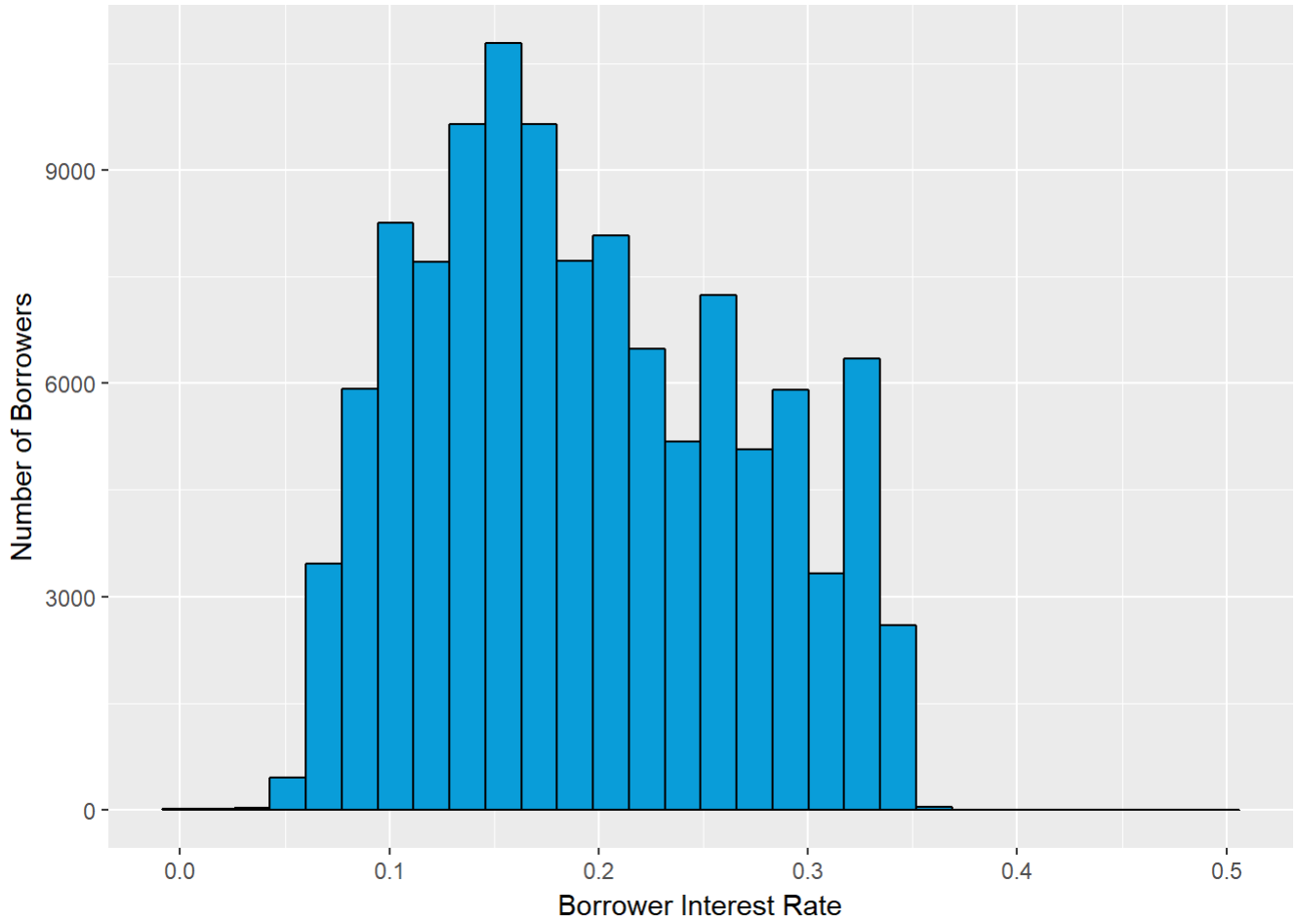
```
##  3rd Qu.: 3.000            IL     : 5921
##  Max.   :20.000                   : 5515
##                           (Other):67493
##                     Occupation         EmploymentStatus
##  Other                   :28617   Employed     :67322
##  Professional            :13628   Full-time    :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           False:56459         False:101218
##  1st Qu.: 26.00           True :57478         True : 12719
##  Median : 67.00
##  Mean   : 96.07
##  3rd Qu.:137.00
##  Max.   :755.00
##  NA's   :7625
##                     GroupKey                 DateCreditPulled
##                         :100596   2013-12-23 09:38:12:     6
##  783C3371218786870A73D20:  1140   2013-11-21 09:09:41:     4
##  3D4D3366260257624AB272D:   916   2013-12-06 05:43:16:     4
##  6A3B336601725506917317E:   698   2014-01-14 20:17:49:     4
##  FEF83377364176536637E50:   611   2014-02-09 12:14:41:     4
##  C9643379247860156A00EC0:   342   2013-09-27 22:04:54:     3
##  (Other)                 :  9634   (Other)            :113912
##  CreditScoreRangeLower CreditScoreRangeUpper
##  Min.   :  0.0         Min.   : 19.0
##  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
##         FirstRecordedCreditLine CurrentCreditLines OpenCreditLines
##                       :   697     Min.   : 0.00      Min.   : 0.00
##  1993-12-01 00:00:00:   185     1st Qu.: 7.00      1st Qu.: 6.00
##  1994-11-01 00:00:00:   178     Median :10.00      Median : 9.00
##  1995-11-01 00:00:00:   168     Mean   :10.32      Mean   : 9.26
##  1990-04-01 00:00:00:   161     3rd Qu.:13.00      3rd Qu.:12.00
##  1995-03-01 00:00:00:   159     Max.   :59.00      Max.   :54.00
##  (Other)              :112389     NA's   :7604       NA's   :7604
##  TotalCreditLinespast7years OpenRevolvingAccounts
##  Min.   :  2.00             Min.   : 0.00
##  1st Qu.: 17.00             1st Qu.: 4.00
##  Median : 25.00             Median : 6.00
##  Mean   : 26.75             Mean   : 6.97
##  3rd Qu.: 35.00             3rd Qu.: 9.00
##  Max.   :136.00             Max.   :51.00
##  NA's   :697
##  OpenRevolvingMonthlyPayment InquiriesLast6Months TotalInquiries
##  Min.   :    0.0             Min.   :  0.000      Min.   :  0.000
##  1st Qu.:  114.0             1st Qu.:  0.000      1st Qu.:  2.000
```

```
##  Median :  271.0            Median :  1.000      Median :  4.000
##  Mean   :  398.3            Mean   :  1.435      Mean   :  5.584
##  3rd Qu.:  525.0            3rd Qu.:  2.000      3rd Qu.:  7.000
##  Max.   :14985.0            Max.   :105.000      Max.   :379.000
##                             NA's   :697          NA's   :1159
##  CurrentDelinquencies AmountDelinquent   DelinquenciesLast7Years
##  Min.   : 0.0000      Min.   :     0.0   Min.   : 0.000
##  1st Qu.: 0.0000      1st Qu.:     0.0   1st Qu.: 0.000
##  Median : 0.0000      Median :     0.0   Median : 0.000
##  Mean   : 0.5921      Mean   :   984.5   Mean   : 4.155
##  3rd Qu.: 0.0000      3rd Qu.:     0.0   3rd Qu.: 3.000
##  Max.   :83.0000      Max.   :463881.0   Max.   :99.000
##  NA's   :697          NA's   :7622       NA's   :990
##  PublicRecordsLast10Years PublicRecordsLast12Months RevolvingCreditBalance
##  Min.   : 0.0000          Min.   : 0.000            Min.   :      0
##  1st Qu.: 0.0000          1st Qu.: 0.000            1st Qu.:   3121
##  Median : 0.0000          Median : 0.000            Median :   8549
##  Mean   : 0.3126          Mean   : 0.015            Mean   :  17599
##  3rd Qu.: 0.0000          3rd Qu.: 0.000            3rd Qu.:  19521
##  Max.   :38.0000          Max.   :20.000            Max.   :1435667
##  NA's   :697              NA's   :7604              NA's   :7604
##  BankcardUtilization AvailableBankcardCredit  TotalTrades
##  Min.   :0.000       Min.   :     0          Min.   :  0.00
##  1st Qu.:0.310       1st Qu.:   880          1st Qu.: 15.00
##  Median :0.600       Median :  4100          Median : 22.00
##  Mean   :0.561       Mean   : 11210          Mean   : 23.23
##  3rd Qu.:0.840       3rd Qu.: 13180          3rd Qu.: 30.00
##  Max.   :5.950       Max.   :646285          Max.   :126.00
##  NA's   :7604        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   False:  8669
##  1st Qu.: 0.140    $50,000-74,999:31050   True :105268
##  Median : 0.220    $100,000+     :17337
##  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
```

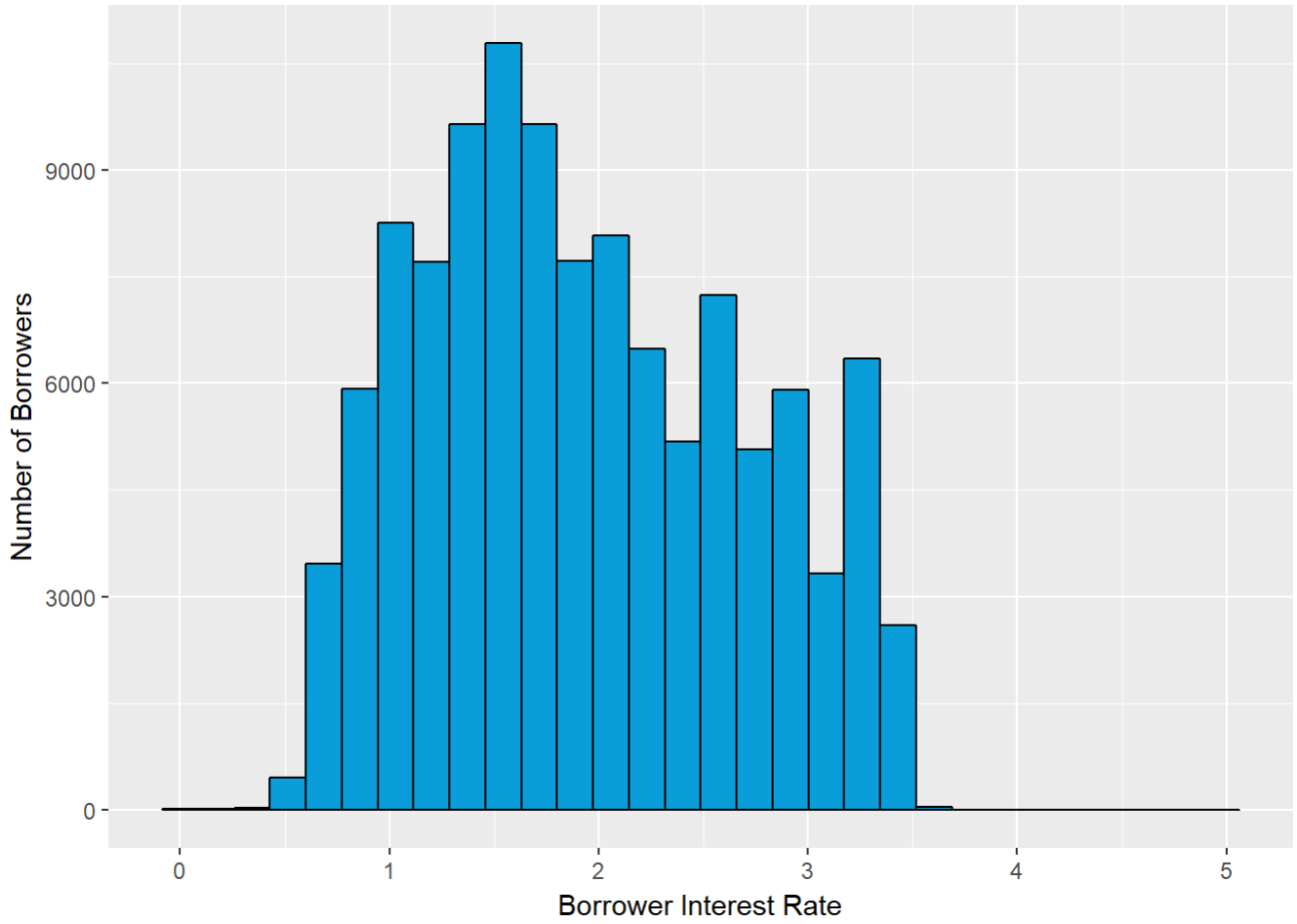
```
##  Min.   :  0.00             Min.   :  0.00
##  1st Qu.:  9.00             1st Qu.:  9.00
##  Median : 16.00             Median : 15.00
##  Mean   : 22.93             Mean   : 22.27
##  3rd Qu.: 33.00             3rd Qu.: 32.00
##  Max.   :141.00             Max.   :141.00
##  NA's   :91852              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      2014-01-22 00:00:00:   491   Q4 2013:14450
##  1st Qu.: 4000      2013-11-13 00:00:00:   490   Q1 2014:12172
##  Median : 6500      2014-02-19 00:00:00:   439   Q3 2013: 9180
##  Mean   : 8337      2013-10-16 00:00:00:   434   Q2 2013: 7099
##  3rd Qu.:12000      2014-01-28 00:00:00:   339   Q3 2012: 5632
##  Max.   :35000      2013-09-24 00:00:00:   316   Q2 2012: 5061
##                     (Other)            :111428   (Other):60343
##                    MemberKey      MonthlyLoanPayment LP_CustomerPayments
##  63CA34120866140639431C9:     9   Min.   :   0.0     Min.   :   -2.35
##  160833364744933457E57FB9:     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
##
```

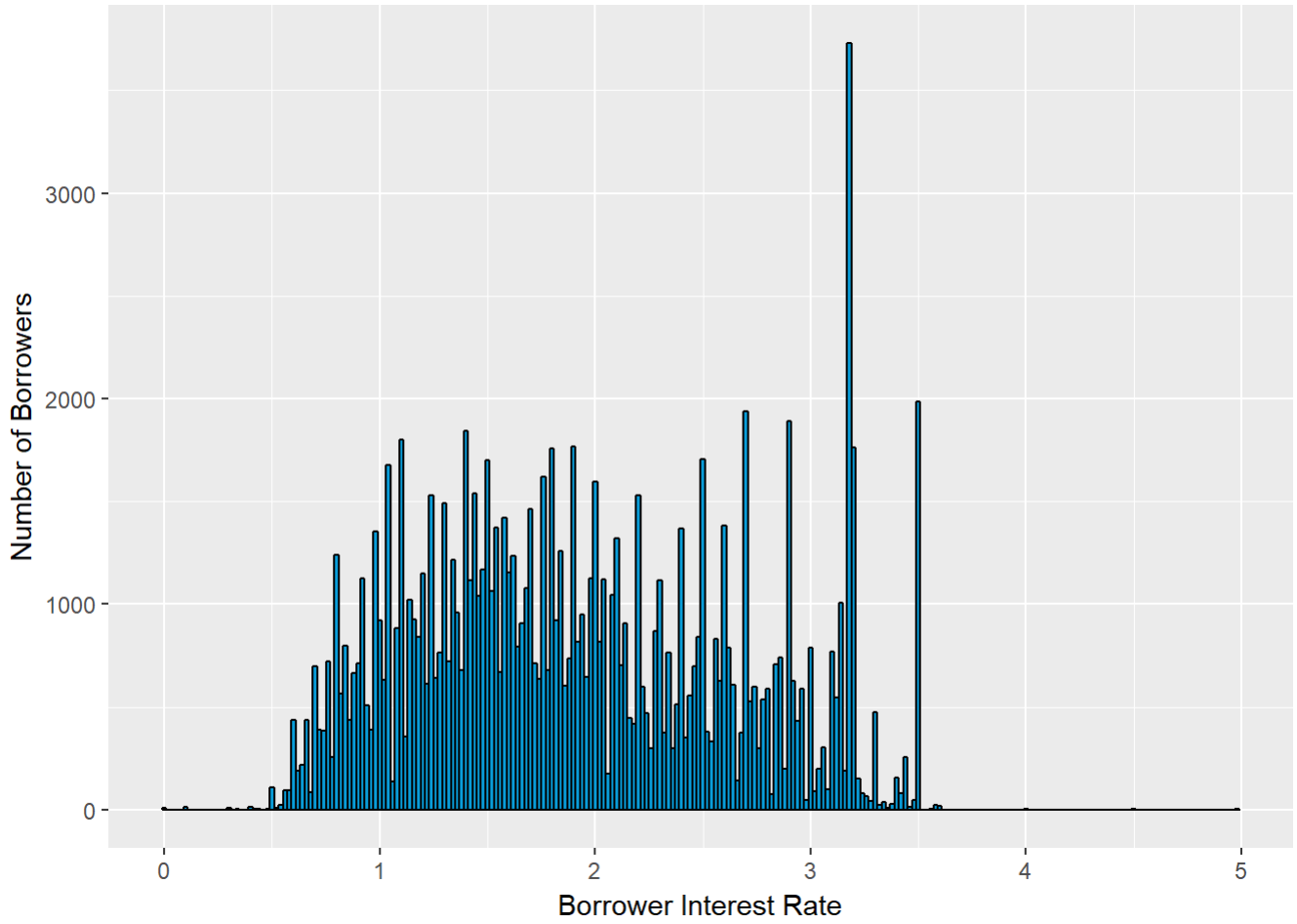
At this point, there are some variables that I'd like to chart out and see what story they may tell.



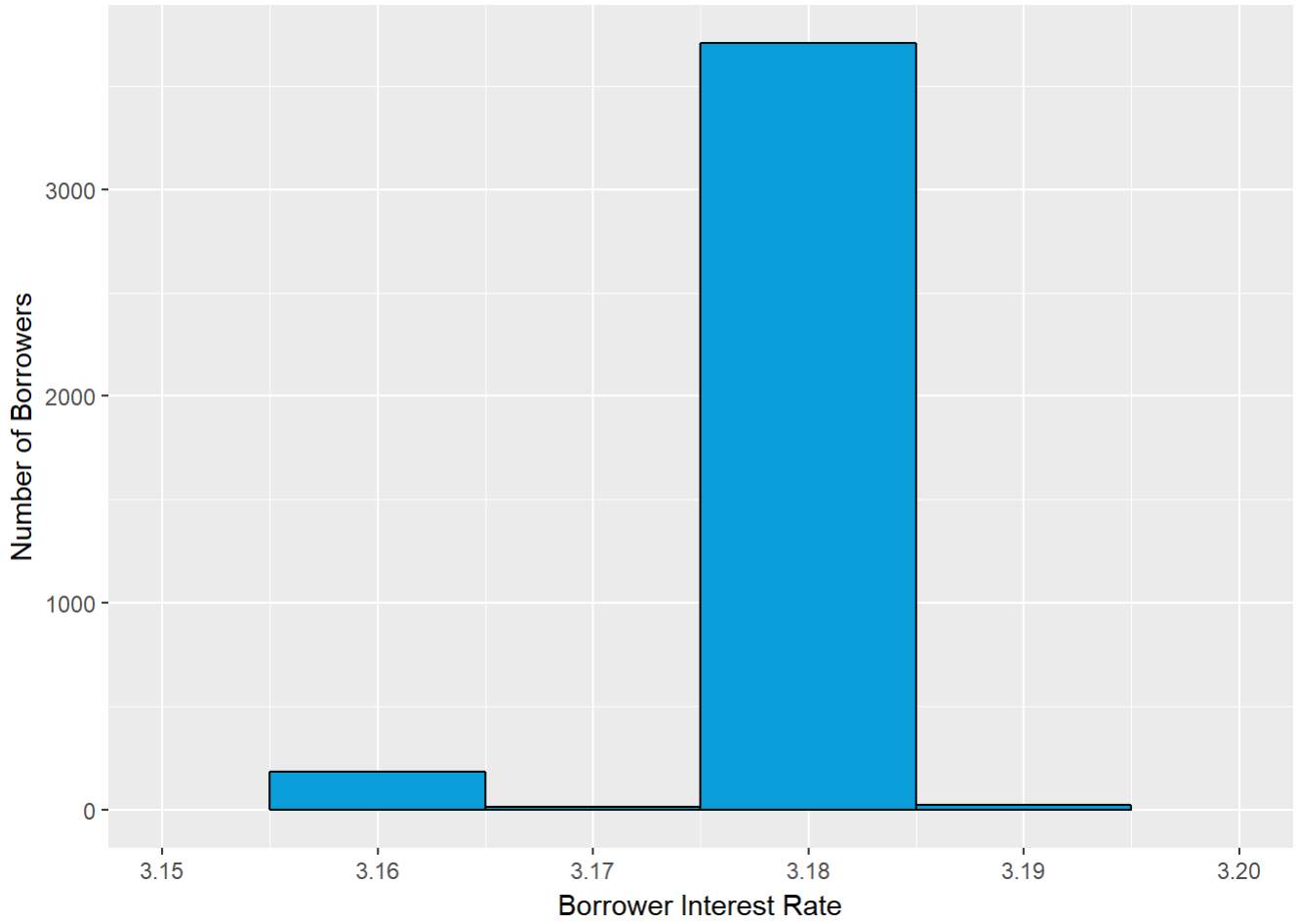
I see that the interest rates aren't converted to percentages, so I'll need to account for that in the chart development.



Now, I can easily see that we have the majority of interest rates sitting between 1 and 2 percent. I am curious if I adjust down the binwidth if it'll tell a different storry or uncover some oddities in the data.



While I still believe the majority sit between 1 and 2 percent, There's a spike just above the 3 percent mark. I want to drill further down into that area to see if I can get a better understanding of things.



Now that I have drilled down, I can see that it's a range of 3.175 to 3.185 percent that causes the spike. As low as the surrounding values are, however, that explains to me why the median isn't skewed to the right. Let's summarize the rates just to double-check myself here.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.340   1.840   1.928   2.500   4.975
```

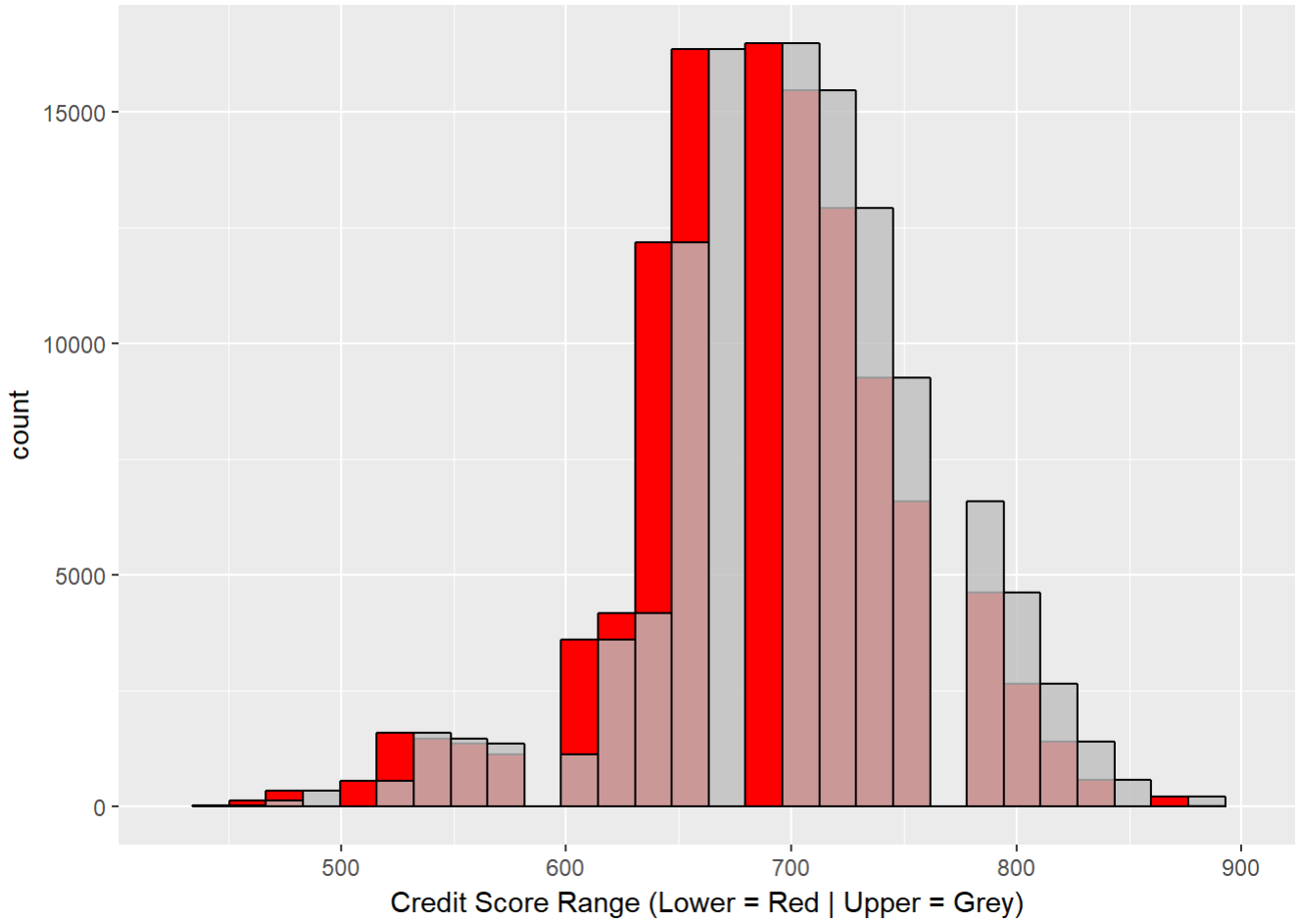
So this confirms my thinking. The median and mean are both within the 1 to 2 percent range. The 3rd quartile doesn't even get into the range of this oddity "spike" if you will. With that said, I believe I'm at a dead end on this pursuit regarding interest rate analysis.

At this point, I want to turn my attention towards credit scores. I am curious what the ranges (upper and lower) are for Prosper's customers. I will summarize the ranges and then plot them to see if anything stands out to me.

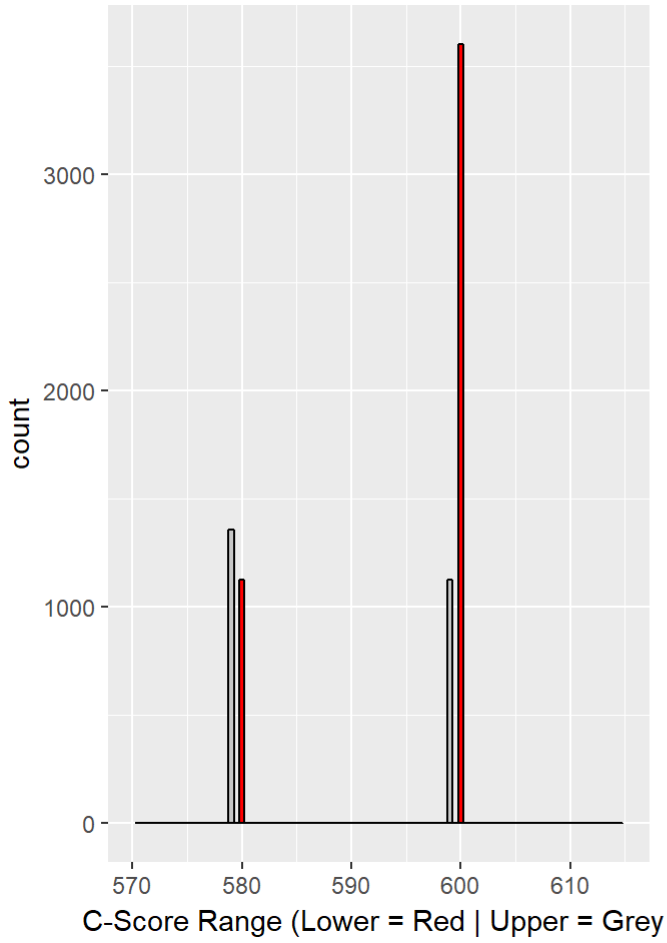
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0   660.0   680.0   685.6   720.0   880.0     591
```

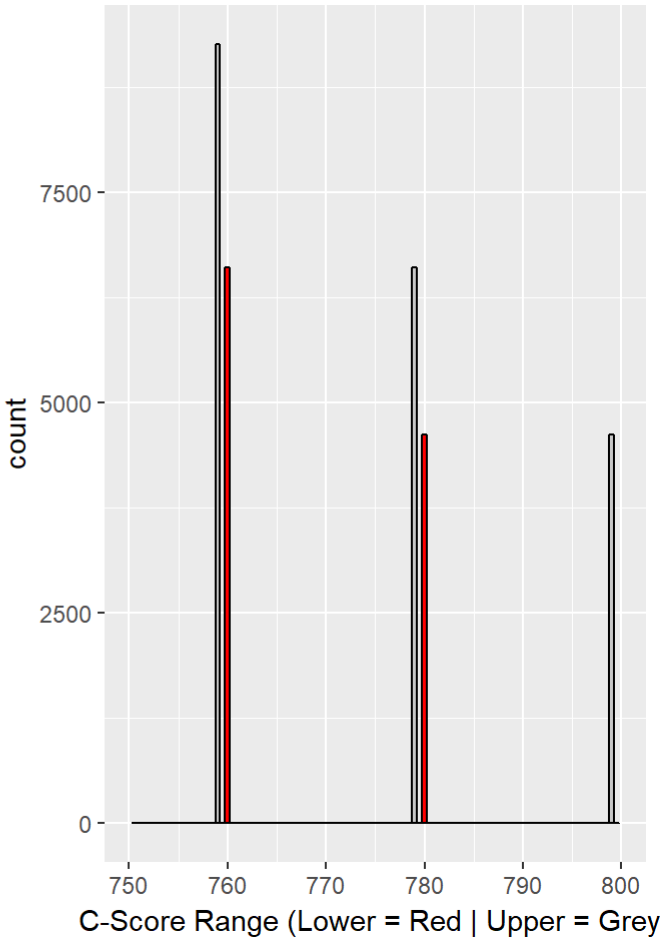
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    19.0   679.0   699.0   704.6   739.0   899.0     591
```

I see there's a 19-point spread between the average upper and lower range (as an example). Based on personal experience, that seems fairly normal. I'll proceed to charting them out to see how we fair quantity-wise.



This looks to be a fairly normal distribution of both variables (e.g. - Upper and Lower Credit Ranges). I notice two areas that have neither variable in them. This indicates to me there's either a lack of data in the fields or simply that none of their customers' credit scores fall in this bin. For grins, I will run this again with a much smaller x-axis range.





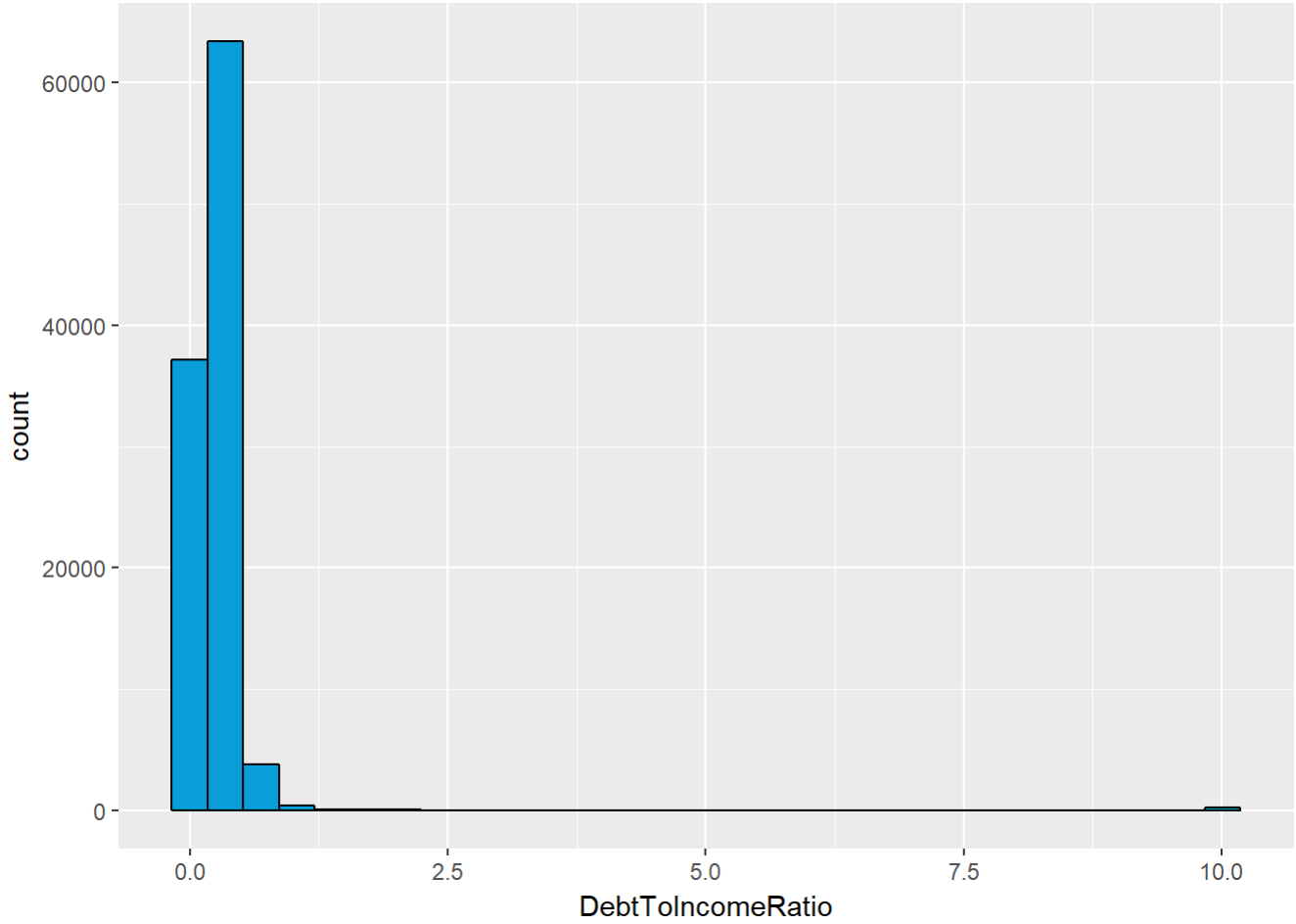
Now that I have seen this at a more granular level, I believe the data is indicating a lack of values in the blank spaces identified. There's nothing additional I am electing to pursue here.

Although we've already been turn there's a correlation between the interest rate we receive for a loan and the credit score we have, I want to confirm that by testing the respective correlations.

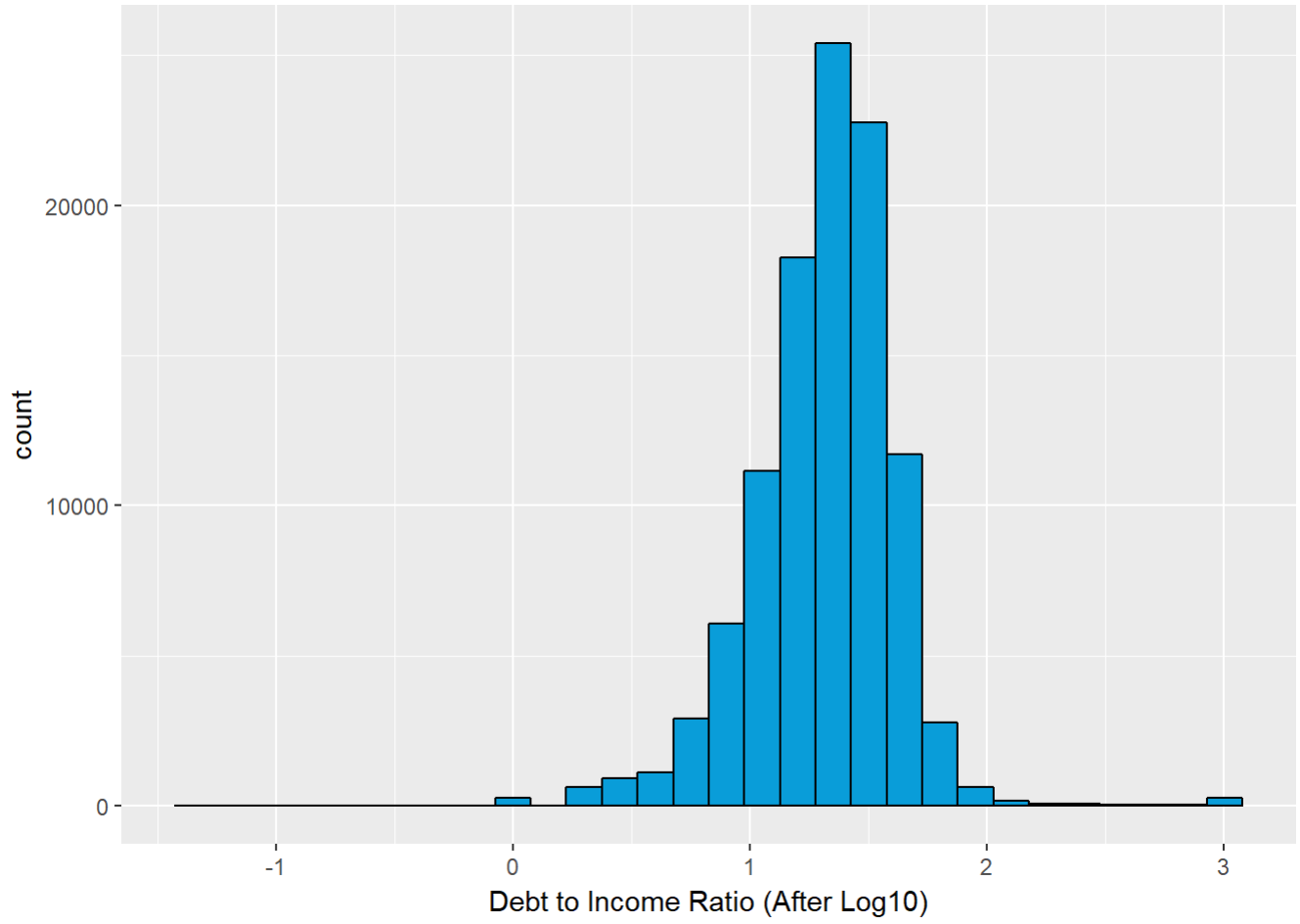
```
##
##  Pearson's product-moment correlation
##
## data:  loanData$BorrowerRate and loanData$CreditScoreRangeLower
## t = -175.17, df = 113340, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4661358 -0.4569730
## sample estimates: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

```
##
##  Pearson's product-moment correlation
##
## data:  loanData$BorrowerRate and loanData$CreditScoreRangeUpper
## 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
```

Both show a good / strong p-value and confirm our assumptions. Now I'd like to plot out the Debt-To-Income Ratio. I am curious how this company looks at this field as it pertains to loan approvals. It may also show some interesting data when combined with other data later in this report.



This graph shows a ton of data points in a narrow bin range to me. I also notice it's positive-skewed with a very long tail. I'd like to perform a logarithm against this data field in an attempt to even out the distribution and get a better feel for where the ratio is spread across the loan customers.



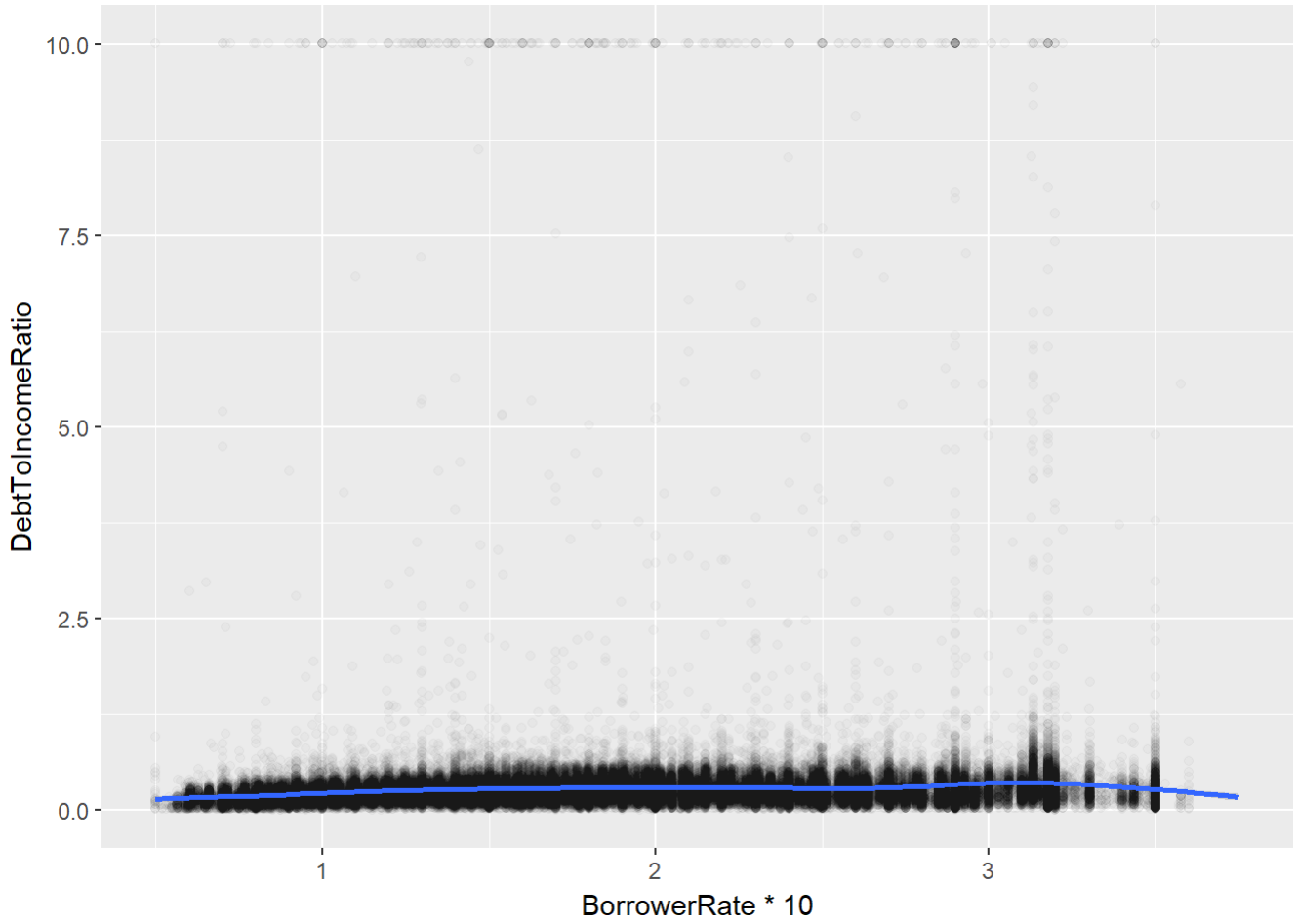
This looks better. Please note that I also multiplied the variable by 100 to positively-adjust the data into a manner that makes more sense. The data shows me the vast majority of folks holding loans have a Debt-to-Income ratio between 1 - 2.

Bivariate Plots and Analysis Section

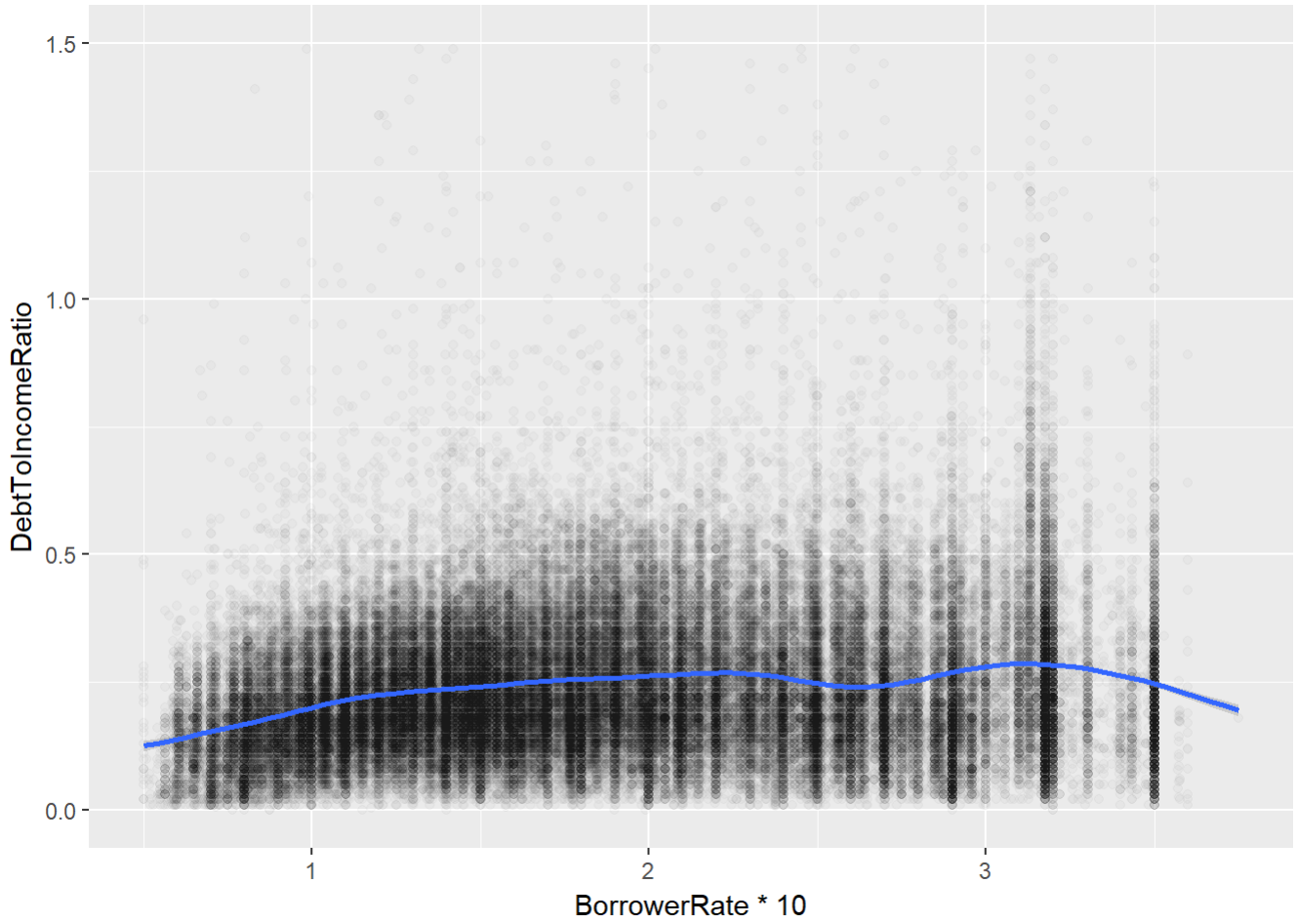
Please note that I prefer to iterate my analysis as I go along plotting data via various visualizations. I believe it flows better and helps the reviewer better understand the analysis and reflection occurring at the time the visualizations are being run.

In this section, I plan to dive deeper into some relationships in the data that peaked my interest during (a) my initial review of the raw data and (b) the univariate plotting / analysis I performed. The following relationships will be explored:

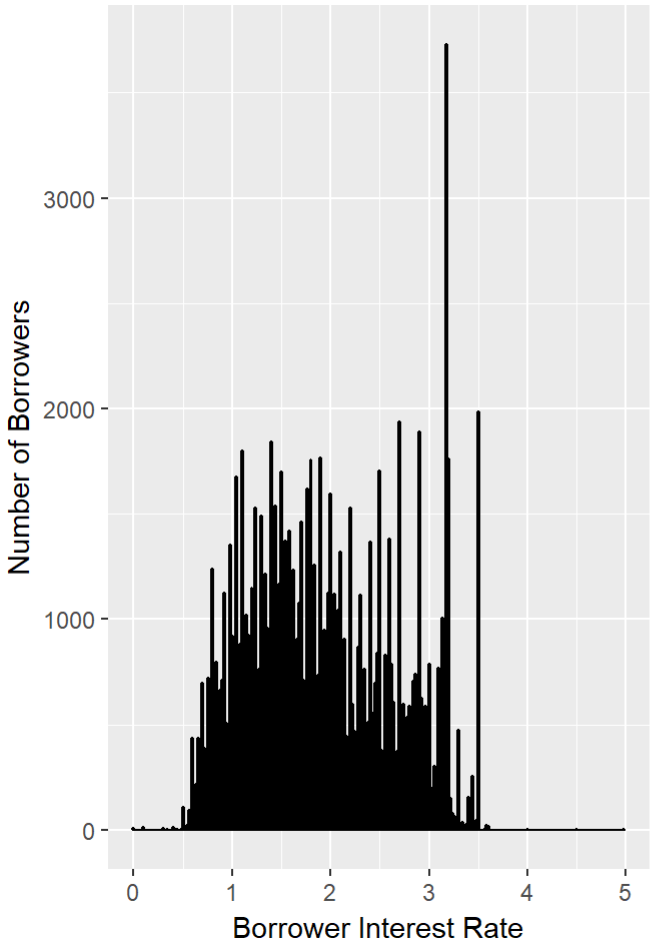
- Borrower's Interest Rate vs. Debt-to-Income Ratio
- Original Loan Amount vs. Loan Terms
- Income Range vs. Borrower's Interest Rate
- Amount Delinquent vs. Credit Score Range (Lower)
- Stated Monthly Income vs. Debt-to-Income Ratio
- Stated Monthly Income vs. Original Loan Amount
- Delinquencies over the past 7 Years vs. Credit Score Range (Lower)
- Prosper Score vs. Loan Interest Rate

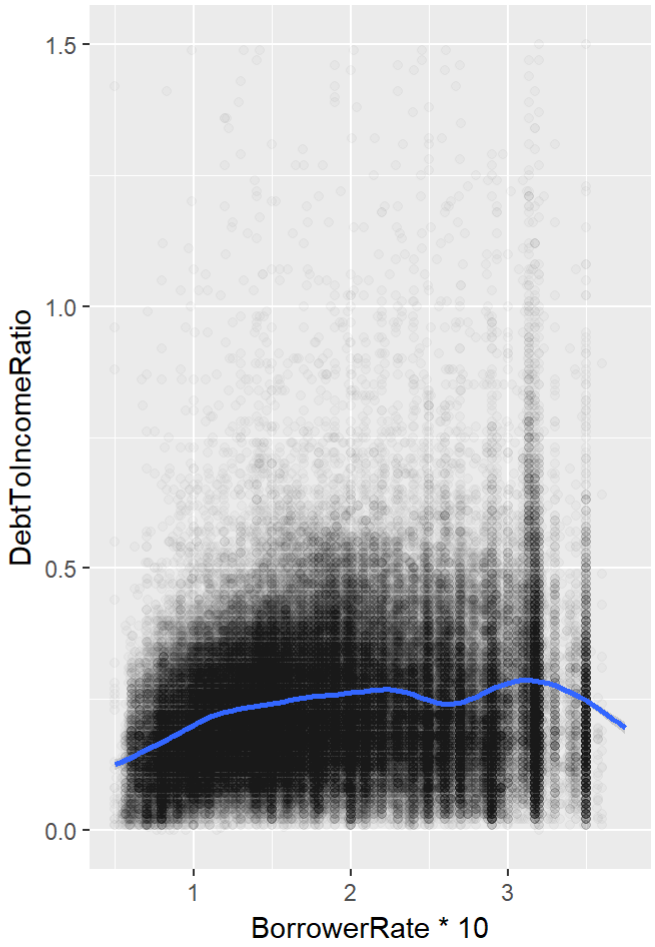


I found this very interesting. I had to apply jitter to the data to clear the “fog” if you will. With that said, I think this chart it’s a positive. By that I mean regardless of the interest rate (generally speaking), customers’ debt-to-income ratio is fairly low. I threw in a geom_smooth layer to the scatter plot to help visualize where the focus of the data points are. I do want to zoom into the data to see if there’s an additional story that’s being told.

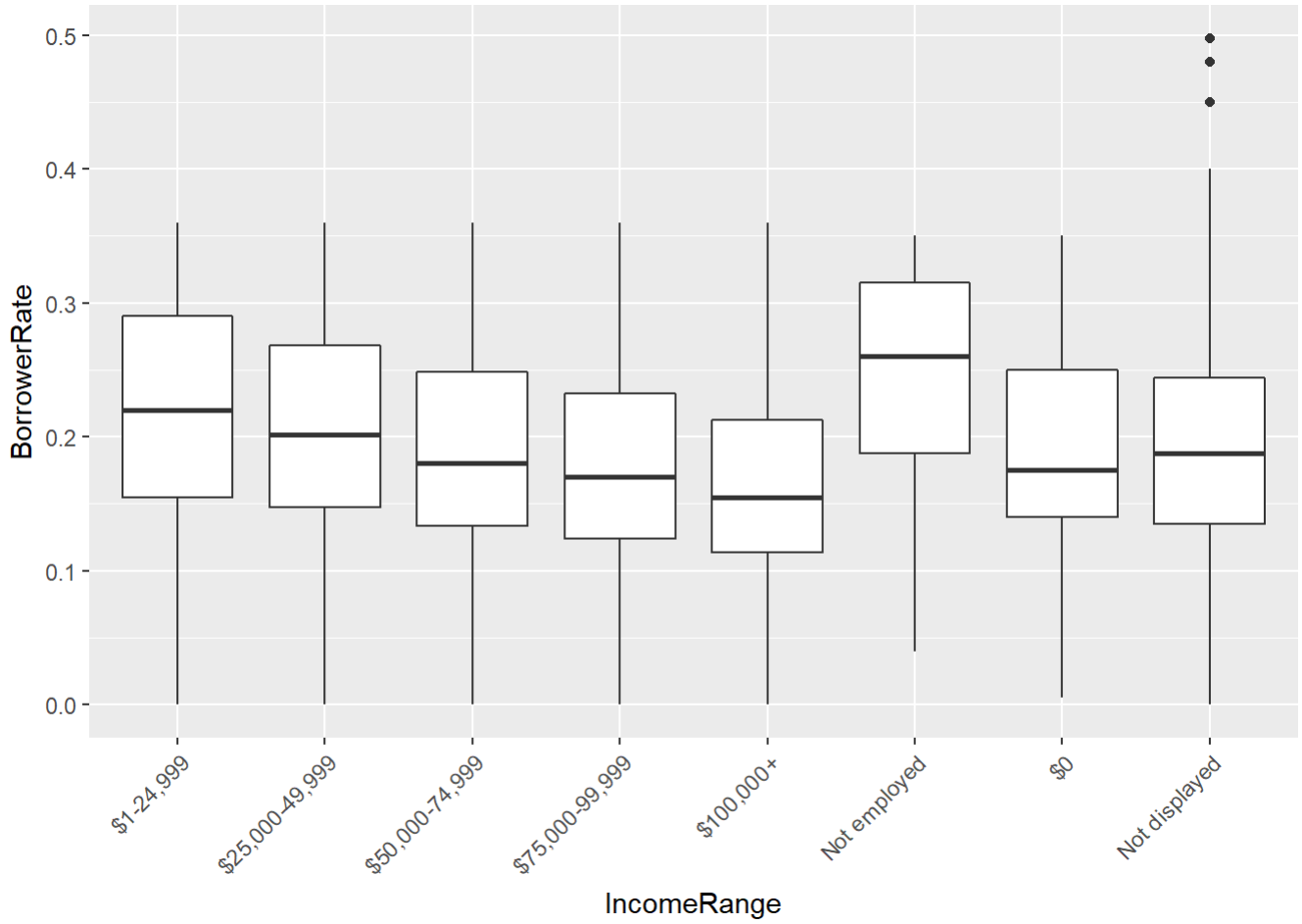


Now that I've put y-parameters on the visual, I can better see what's going on. We have a minor spike in the annual interest rate (north of 3 percent), but the majority of borrowers' rates are fairly steady in the 1-2 range. I have also noticed the debt-to-income ratio seems to peak at 0.5 with the majority between 0.25 - 0.35. Overall, this would give me confidence (as an investor) in the company's drive to provide loans to "healty" candidates (fiscally-speaking).

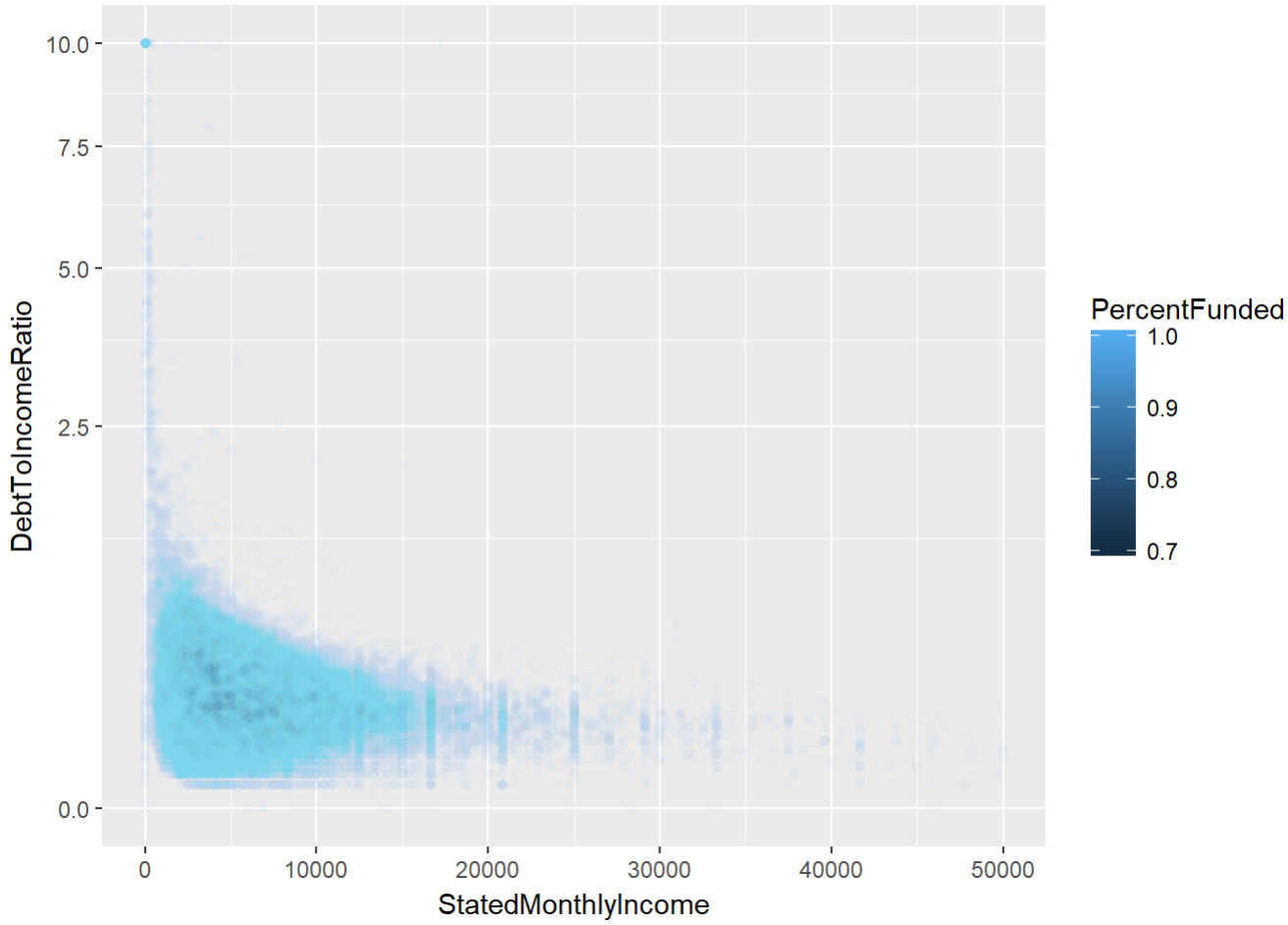




I didn't want to let go of analyzing the borrower's interest rate. I was curious how a comparison of the scatter plot side-by-side a bar-chart of the interest rates would look and whether or not it confirms my thoughts on the subject. Although I'm not using the debt-to-income ratio on the bar chart, it does show a general similarity between the two charts.



I used this chart to help me see what the data point ranges are within each salary range regarding interest rates. Given we know the lower income-to-debt ratio correlates to a lower (generally speaking) interest rate, This box plot chart confirms it for me. More specifically, we see the higher the salary range, the lower the interest rate is, which tells me they likely have a lower income-to-debt ratio as well.

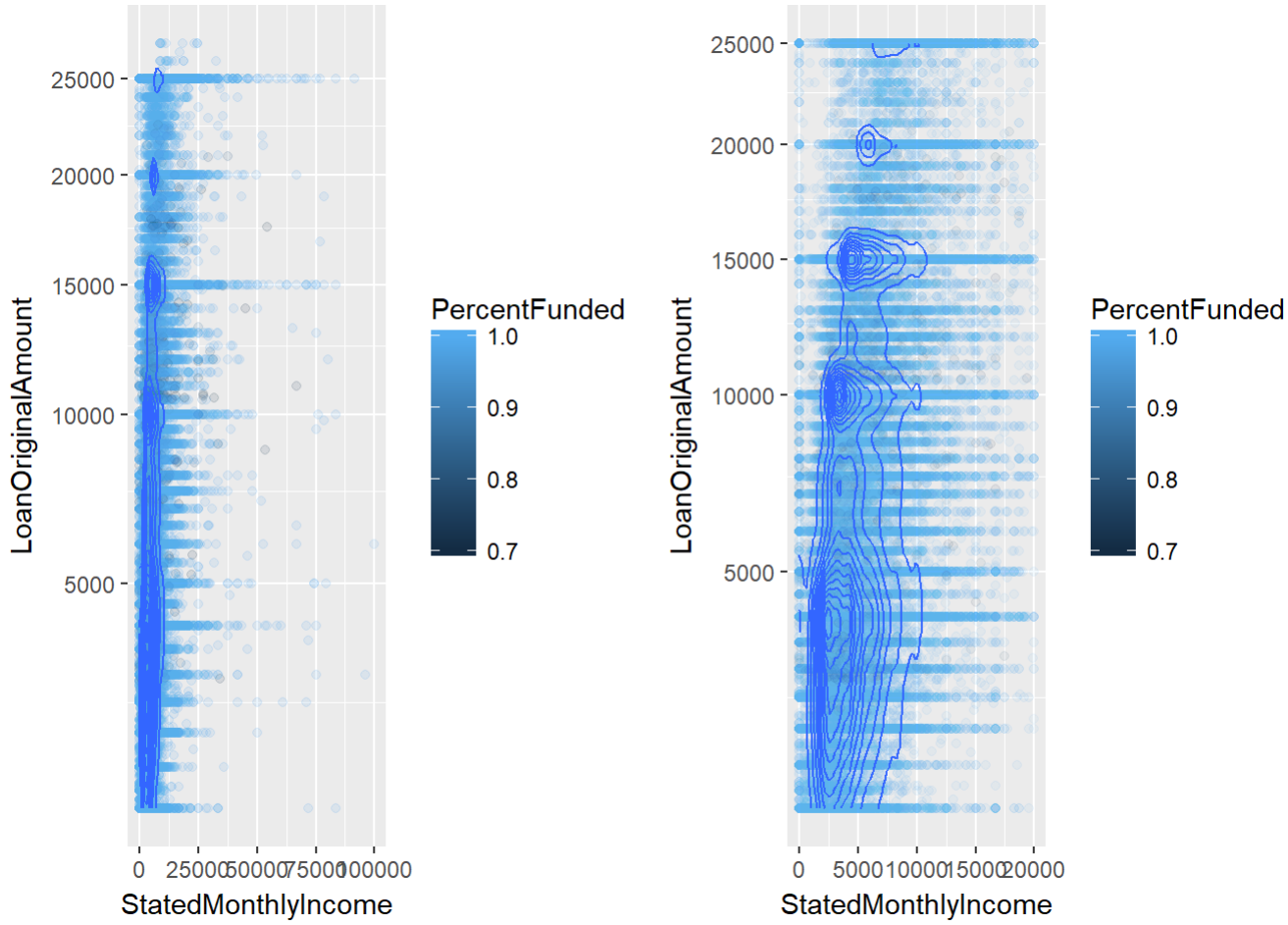


While we're on the subject of income to debt ratio and incomes, I wanted to see if I can uncover how much of the loan was funded. My thought was to compare the stated income and income-to-debt ratio.. then color in the plots based on percentage funded.

After removing the noise original shown in the chart, I can easily see the following:

- the majority of loans funded occur below a 2.5 debt-to-income ratio and for folks who state their income is less than $20,000.
- some loans show partially-funded in the lower-tier stated income range

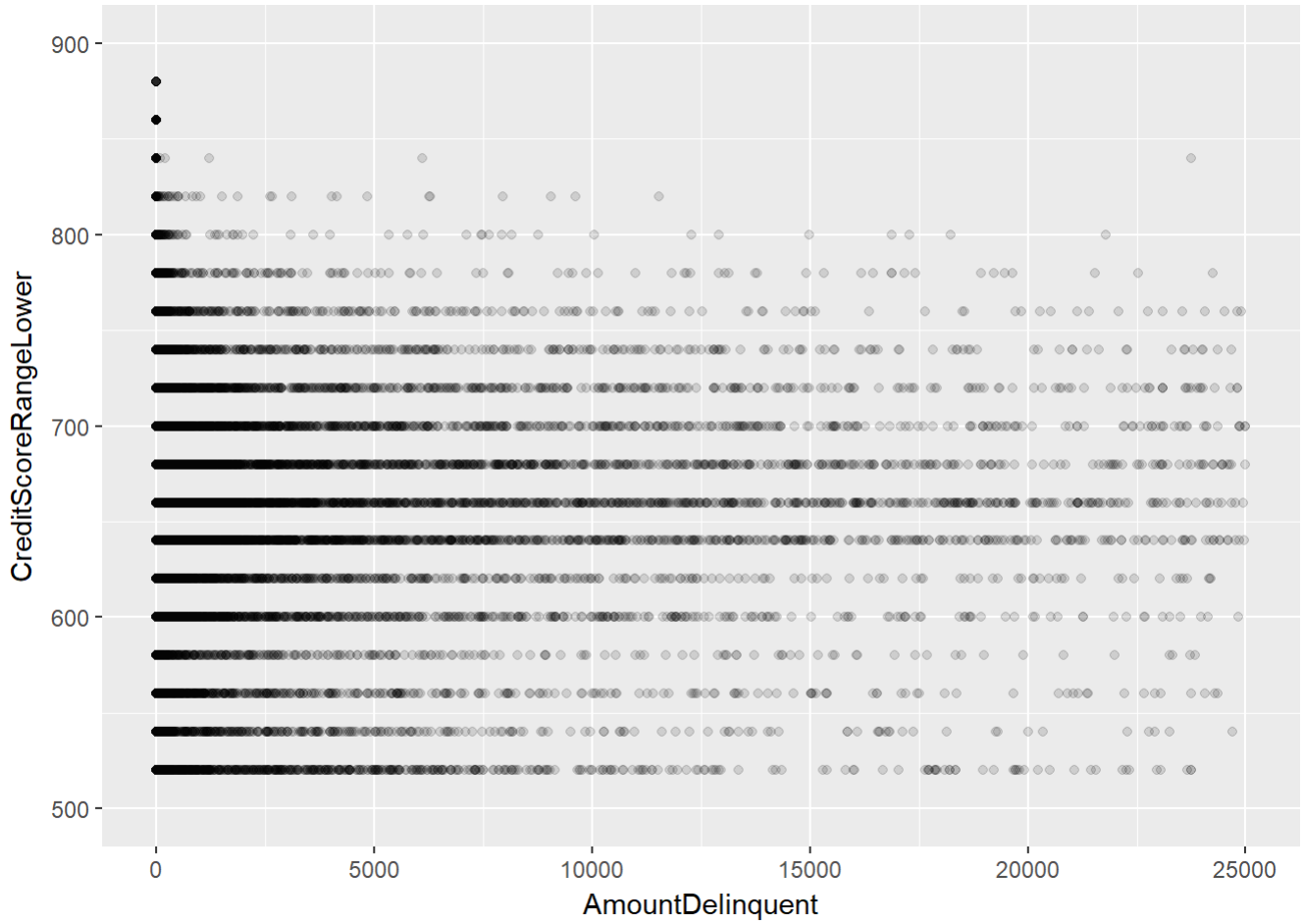
While we can see there's an obvious discrepency between stated and income range, the bigger value (to me) is placed on the income-to-debt ratio as a means to determine confidence in proceeding with loan funding. With that said, I do want to further investigate the stated income and loan amounts.



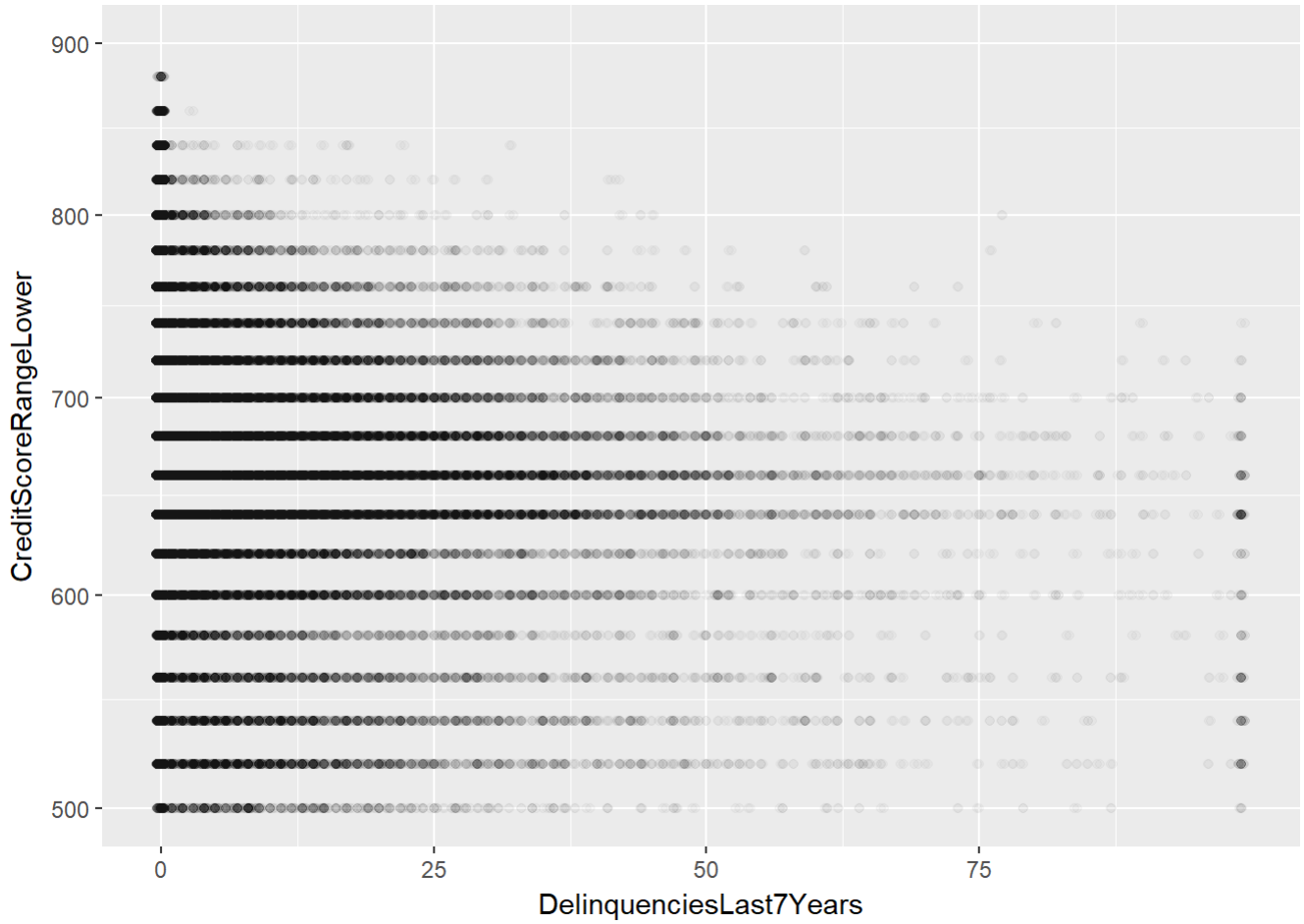
The two graphs above show the overall stated salary range (left) and then the focused salary ranges(right). I added in a density layer to these graphs to help me see where we're at level playing fields with the data. This shows to me what the loan amounts are compared to stated incomes.

```
##
##  Pearson's product-moment correlation
##
## data:  loanData$Term and loanData$LoanOriginalAmount
## t = 121.6, df = 113940, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3337778 0.3440569
## sample estimates:
##       cor
## 0.3389275
```

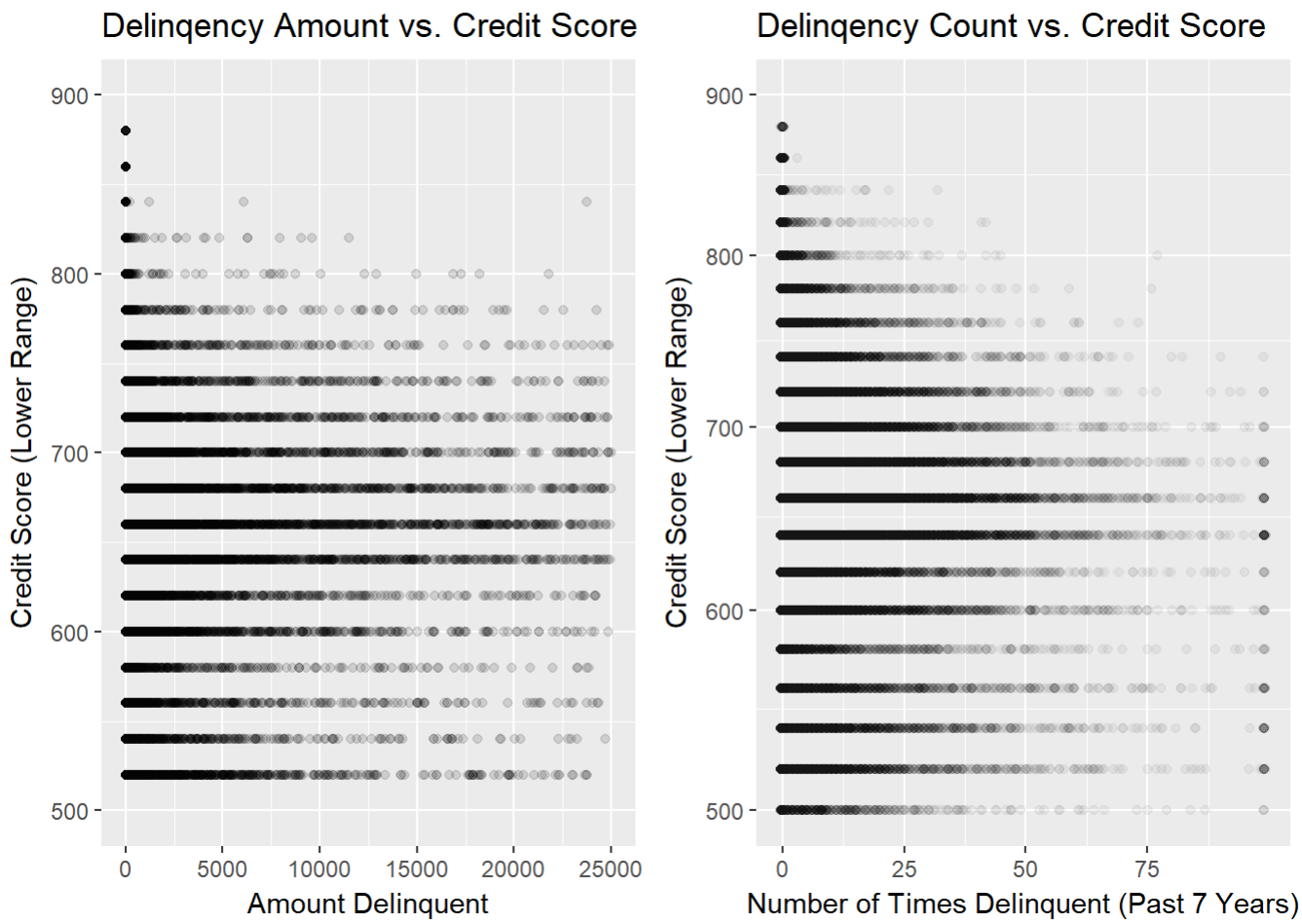
This confirms for me there is in fact a correlation here. The bigger the loan amount, the longer the term is. This also can warrant an assumption that the motivation behind this is to keep the monthly payments manageable.



Being that delinquent accounts are obviously present with Prosper, I want to better understand what's going on there. My thinking led me to review if there's any pattern between the amount of delinquency vs. the borrower's credit score. I would figure the higher the credit score, the lower the probability that loans would be delinquent and/or have a higher amount that is past due. After trimming down the dollar amount to reasonable levels (aka removing the outlying "noise"), I was surprised to see that the mid-600 range up to 700 credit score borrowers have the widest-ranging amounts that are past due. As I recall in the previous chart breaking down income salary ranges, the bigger picture / explanation here would be the higher limit loan amounts folks are signing onto. Additionally, we know there's a correlation between credit scores and loan amounts, so that would explain what we're seeing here as well.

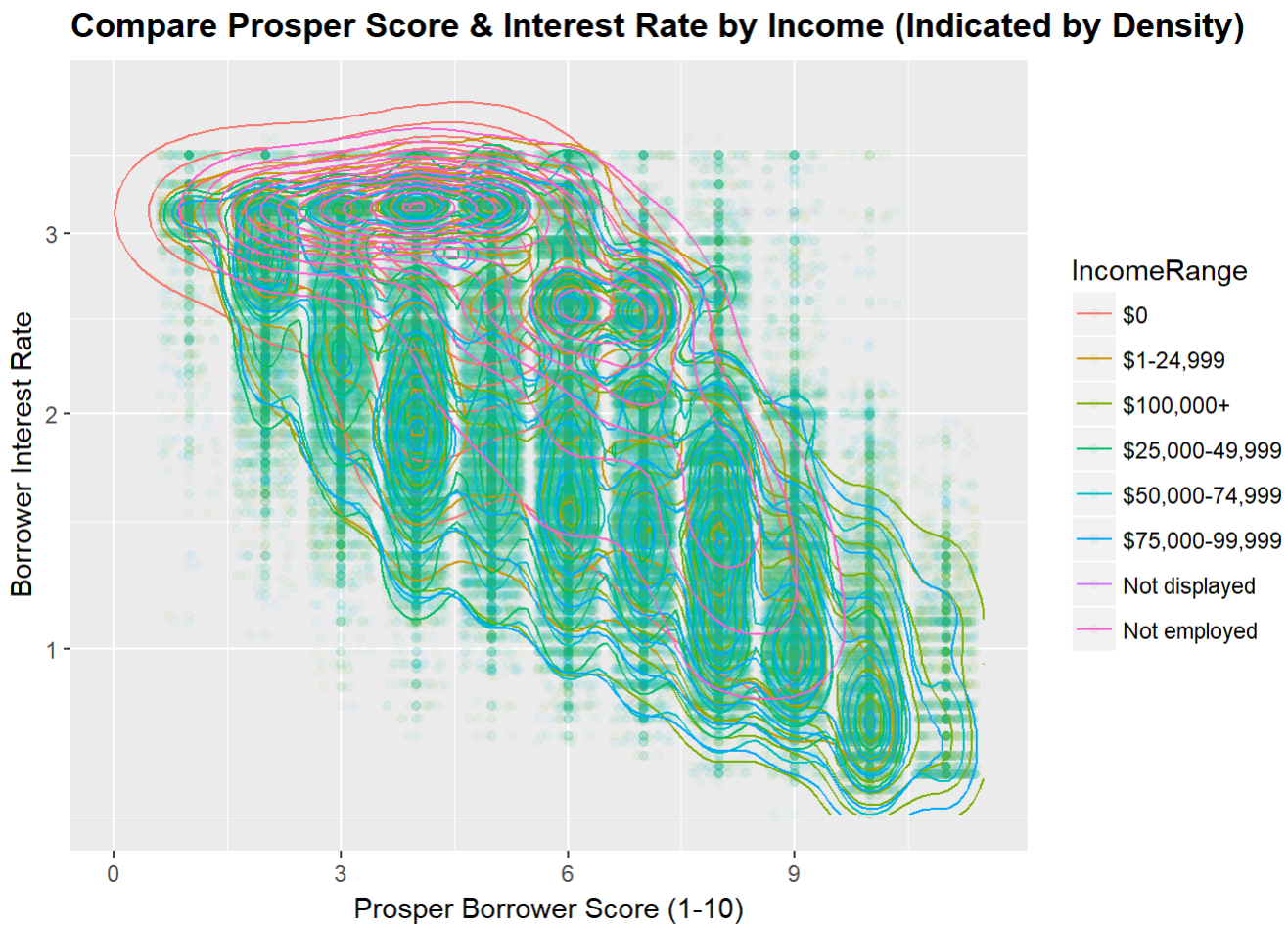


This visualization is interesting to me. Over the last 7 years, we can see how many times folks have missed a payment (based upon credit score). I know we've already talked about the correlation between loan amounts and credit scores, but I want to see these two charts side-by-side because I think there's some overlap here.



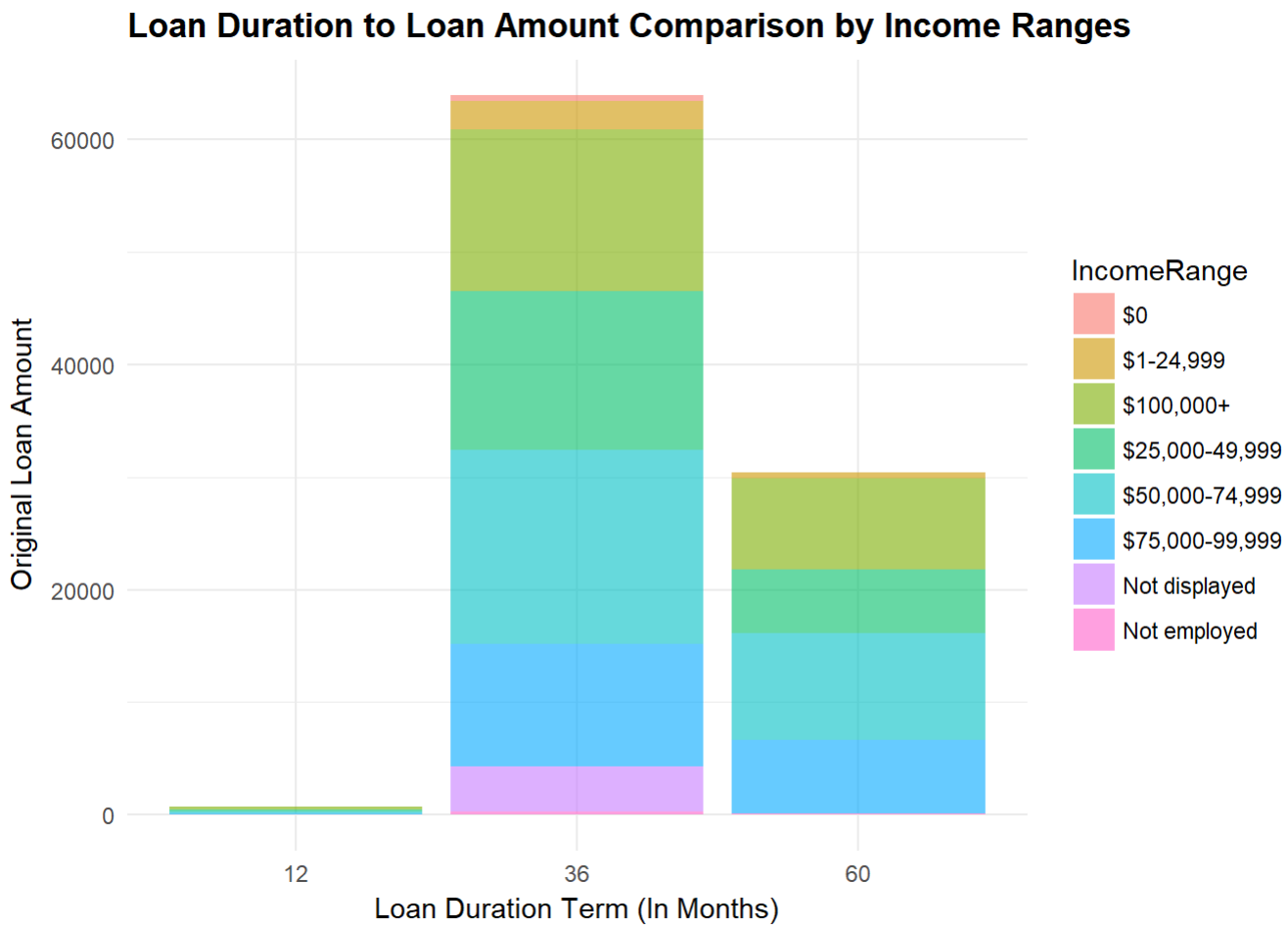
Interesting finding here. So we can say with confidence that the lower-mid tier 600 to 700 credit score range struggles the most in keeping their payments current. This is a red-flag to me because the loan amount correlates to the quantity of delinquencies. This tells me there's a higher probability of someone (in this credit range) defaulting on their loan.

Multivariate Plots and Analysis Section



The final angle of my analysis is how Prospers customer scoring correlates to the interest rate (broken down by salary range). I applied noise reduction techniques (jittering) and added a density layer to help me see where groups are concentrated at. This gives me a wealth of information:

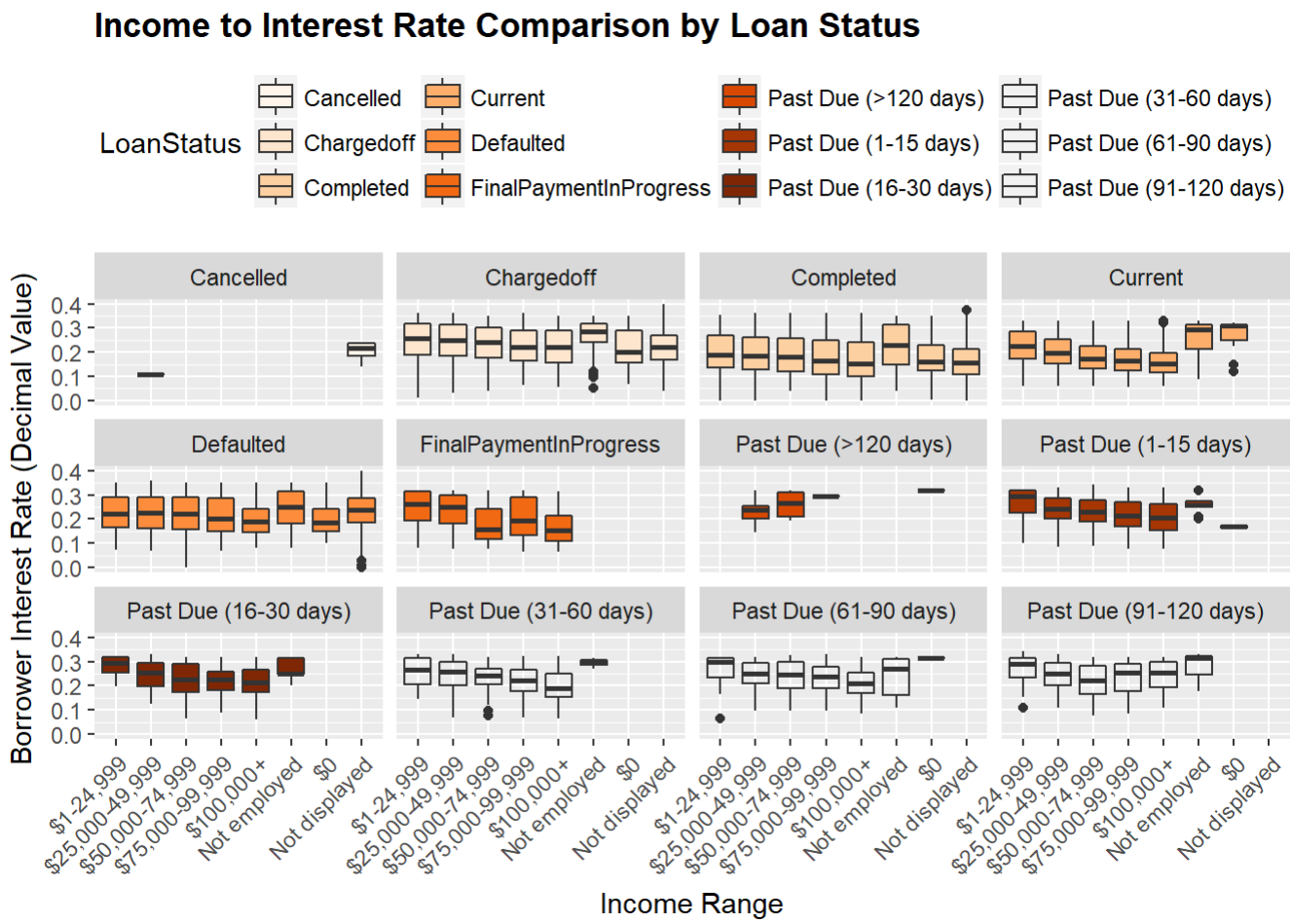
- We have a concentration of folks either showing $0 or “Not employed” at the upper spectrum of interest rates
- We have a concentration of folks ranging from 50,000 to $100,000+ that are on the higher Prosper Score and lower interest rate.
- Lower-tiered salary ranges have a lower Prosper score



Having looked at the debt-to-income ratio and the borrowers' credit scores, I was curious about a couple of things:

- Knowing we have a good correlation between credit score and interest rate, does that tend to give the borrowers more confidence to take on bigger / longer-term debt?
- How exactly do the loan terms lay down as far as the dollar amount loaned and a person's salary range?

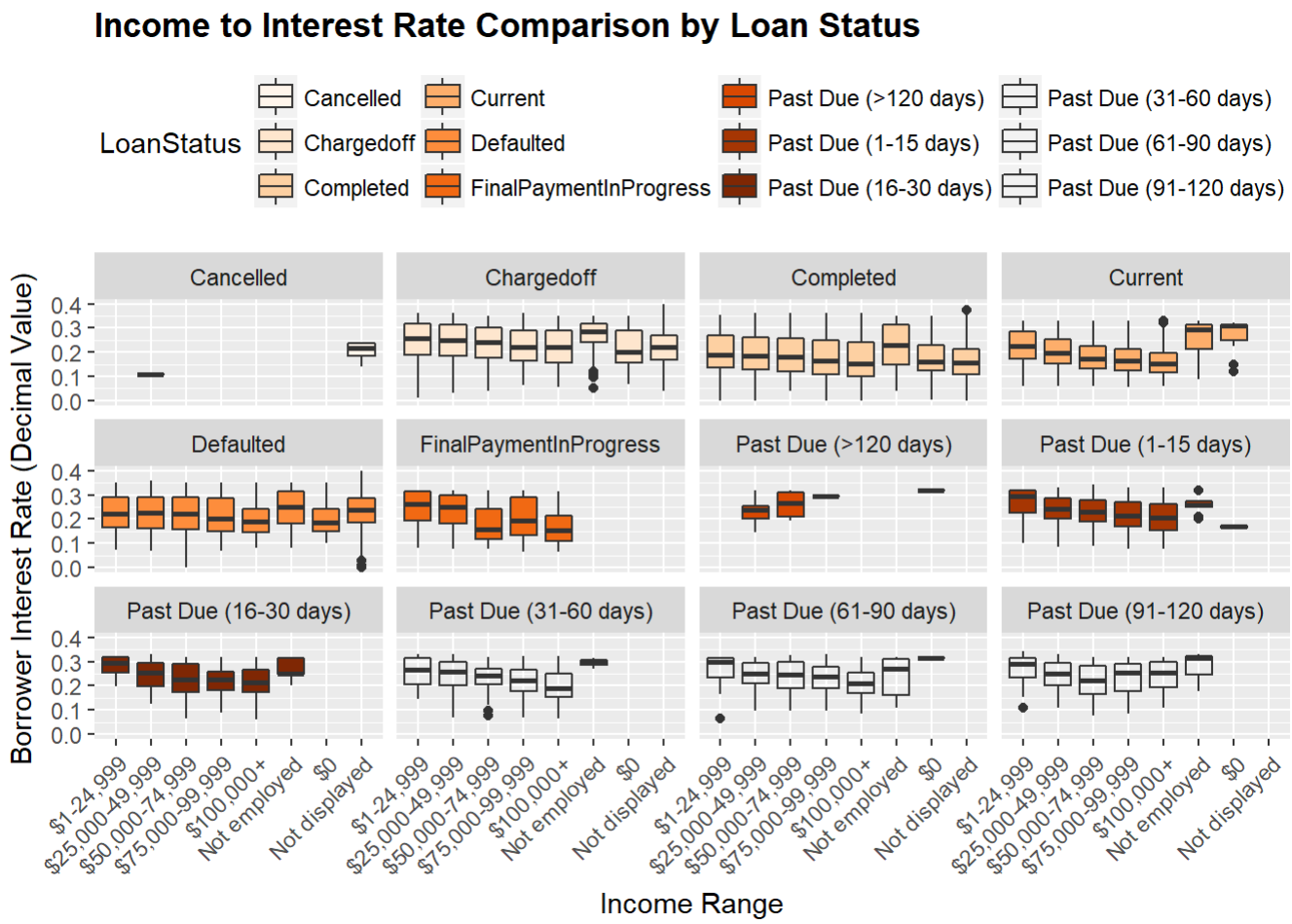
I think this graph does a nice job of giving me the answers. This graph shows a breakdown of loan terms and the amount on the loan… all by the different income ranges. I did have to multiply the loan amount value to get it to display correctly on the visual.



I still had lingering questions about the income range. More specifically, what's the breakdown of loan status' look like based upon the salary range? My reasoning / thinking is that, depending on the combo of salary and interest rate, there's a lower likelyhood that higher wage earners get past due and/or default on their loan. I am surprised to see that the various salary ranges (generally speaking) do NOT appear to have a major bearing on whether or not the loan either defaulted or was charged off.

Final Plots and Summary

Plot One



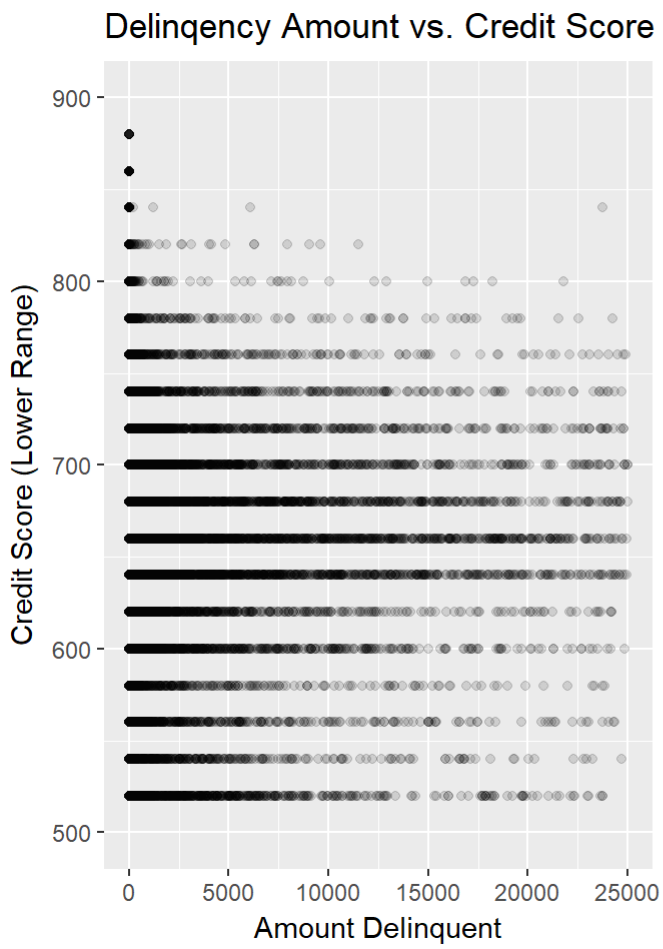
Description One

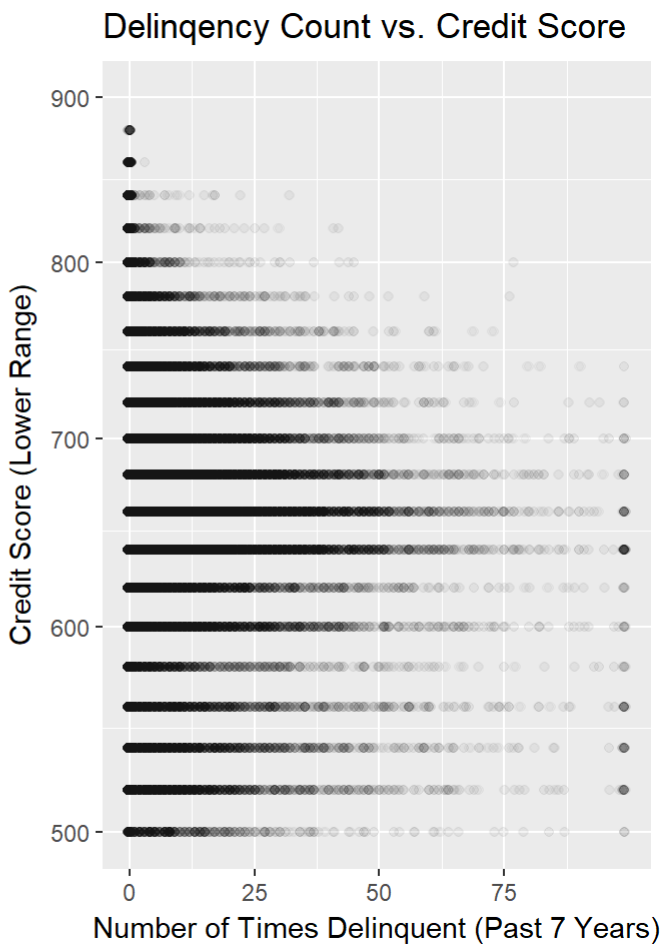
This was one of the most interesting visualizations for me. I took this from the viewpoint of an investor and asked myself:

- How can I quickly size up what I may or may not be getting myself into by working with Prosper?

With this chart, I can quickly assess the different income ranges across the spectrum of historical loan status' and also see how it compares to their respective interest rates. I use interest rates to gauge Prosper's "comfort level" with borrowers, so I can use that information to compare against the previously-mentioned variables. As an investor, I can see that I would want to focus my investments towards the $100,000+ salary range.

Plot Two

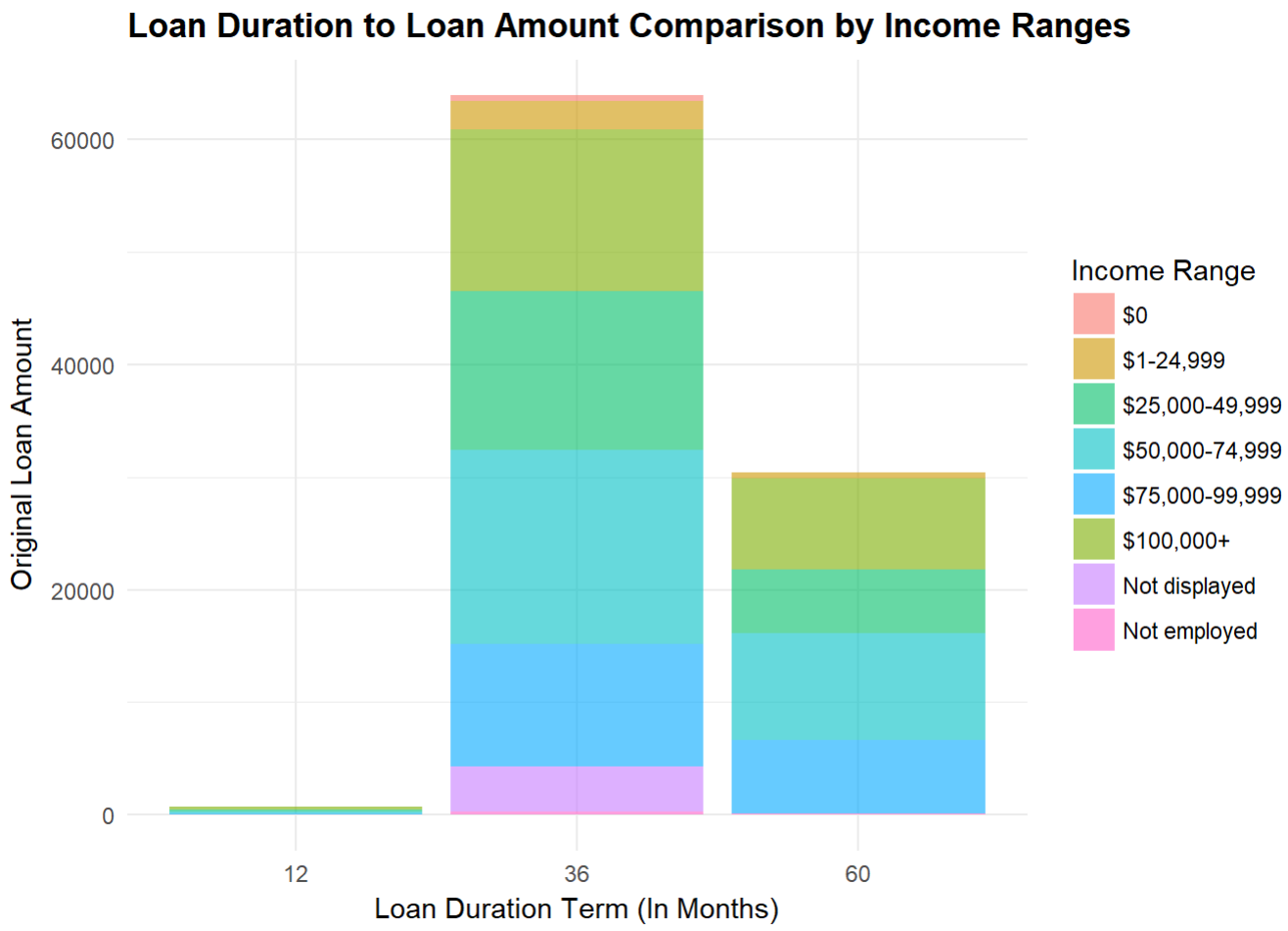




Description Two

Again, I like to look at this data from the eyes of the investor. As an investor, this comparison between deliquency information and credit scores (lower range) help me gauge where I'd like to specifically target and/or avoid borrowers to help hedge my bets against defaulting loans. With this type of analysis, I can see that I'd want to (at least in the beginning) avoid the 600 - 700 credit score range and likely target the 701+ credit score range to minimize risk to my investment(s).

Plot Three



Description Three

Part of an investor's mindset in evaluating their next venture is determining and/or understanding the full picture before making a decision. Part of this is understanding the size of investment(s) they would be facing. As an investor, I can use this multivariant chart to quickly see the vast majority of incomes are (approximately) ranging in the 30-month to 50-month loan terms. Given my leaning-interest towards the 100,000+ salary range borrowers, I can also see (using this visual) that the majority of those folks are in the 30-month to 50-month range. Knowing this kind of information, along with the dollar amount range of approximately 50,000 - 60,000, helps me to determine things like, "how much / what percentage of their loan request would I feel comfortable funding?" or "knowing the approximate range of their interest rate, will I make enough Return On Investment (ROI) with the borrowers I'm wanting to target?"

Reflection

This was an enjoyable project to work on. There were so many data points and possible packages I could have used to explore, size up, and make decisions upon my findings that it was difficult selecting a path forward. I chose to take a lot that was taught from this course and augment a few approaches (e.g. - custom labels, jittering, etc.) to come away from this with insightful information.

Generally speaking, my findings were within my expectations given my past personal and professional experiences in the financial industry. One surprise to me was my finding in the dollar amount and quantity of delinquencies in the 600 - 700 credit score range. That highlighted a red flag for potential investors regarding certain loan requests to avoid and frankly was not expecting that credit range to have such a negative history with Prosper.

Some of the struggles I went through was selecting the right visual to tell the story that I saw in my mind. I spent significant time comparing different visuals, especially for multivariant plotting, to make sure the story being told was fair and accurate. Although the struggles existed, I found myself learning much more in building the visuals out and uncovering / confirming theories in my mind. So you could say it was bittersweet for me. One final element I struggled with was the initial review of the raw data. I tend to look at that first to get a feel for what I'm dealing with rather than just throwing charts up and seeing what sticks if you will. On a positive note, I thought throwing the bivariate and univariate plots together went fairly well and quickly. It afforded me more time to focus on understanding what the data was telling me and helped me to step my way through the exploration efforts needed for this project. By the time I was done with bivariate plotting, I knew where I wanted to combo data points to get a "bigger picture" angle of Prosper's data.