Prosper Loan Data Exploration

Chris Young

Initial Dataset Analysis

```
## [1] "/Users/dcyoung23/Documents/Project P4 - Explore and Summarize Data"
```

Load data:

```
pl <- read.csv('prosperLoanData.csv')
```

113,937 observations of 81 variables in the dataset.

Isolate specific colums for further data exploration:

```
##  [1] "Term"
##  [2] "LoanStatus"
##  [3] "BorrowerRate"
##  [4] "EstimatedLoss"
##  [5] "EstimatedReturn"
##  [6] "ProsperRating..numeric."
##  [7] "ProsperRating..Alpha."
##  [8] "ProsperScore"
##  [9] "ListingCategory..numeric."
## [10] "BorrowerState"
## [11] "Occupation"
## [12] "EmploymentStatusDuration"
## [13] "IsBorrowerHomeowner"
## [14] "CreditScoreRangeLower"
## [15] "OpenCreditLines"
## [16] "OpenRevolvingAccounts"
## [17] "OpenRevolvingMonthlyPayment"
## [18] "AmountDelinquent"
## [19] "DelinquenciesLast7Years"
## [20] "PublicRecordsLast10Years"
## [21] "RevolvingCreditBalance"
## [22] "BankcardUtilization"
## [23] "TradesNeverDelinquent..percentage."
## [24] "DebtToIncomeRatio"
## [25] "IncomeRange"
## [26] "StatedMonthlyIncome"
## [27] "LoanOriginalAmount"
## [28] "LoanOriginationDate"
## [29] "LoanOriginationQuarter"
## [30] "MonthlyLoanPayment"
## [31] "LP_InterestandFees"
## [32] "LP_GrossPrincipalLoss"
## [33] "LP_NetPrincipalLoss"
## [34] "LP_NonPrincipalRecoverypayments"
```

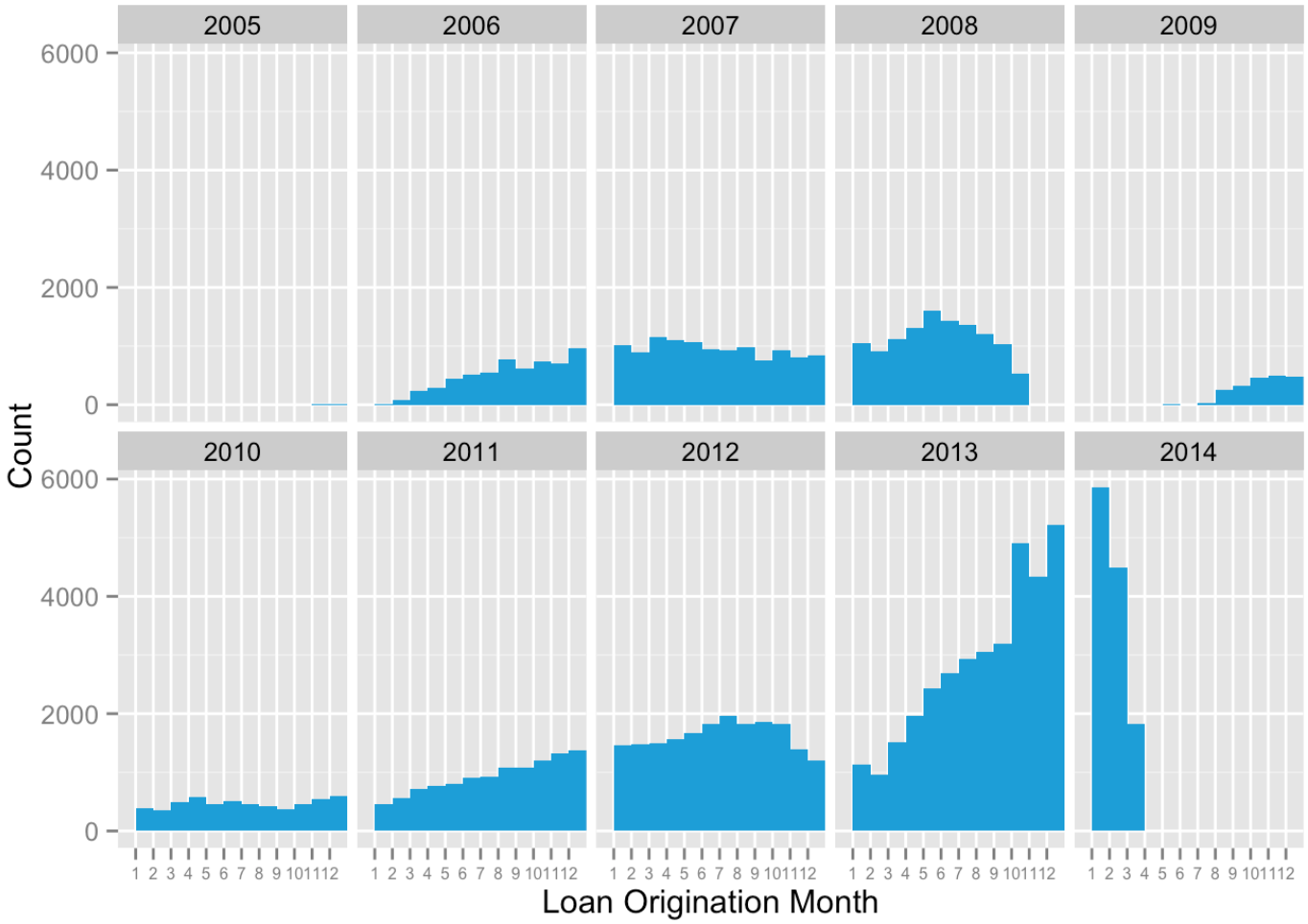
Summary statistics for specified columns:

```
##       Term                       LoanStatus     BorrowerRate
##  Min.   :12.00   Current              :56576   Min.   :0.0000
##  1st Qu.:36.00   Completed            :38074   1st Qu.:0.1340
##  Median :36.00   Chargedoff           :11992   Median :0.1840
##  Mean   :40.83   Defaulted            : 5018   Mean   :0.1928
##  3rd Qu.:36.00   Past Due (1-15 days) :  806   3rd Qu.:0.2500
##  Max.   :60.00   Past Due (31-60 days):  363   Max.   :0.4975
##                  (Other)              : 1108
##  EstimatedLoss   EstimatedReturn  ProsperRating..numeric.
##  Min.   :0.005   Min.   :-0.183   Min.   :1.000
##  1st Qu.:0.042   1st Qu.: 0.074   1st Qu.:3.000
##  Median :0.072   Median : 0.092   Median :4.000
##  Mean   :0.080   Mean   : 0.096   Mean   :4.072
```

```
##  3rd Qu.:0.112   3rd Qu.: 0.117   3rd Qu.:5.000
##  Max.   :0.366   Max.   : 0.284   Max.   :7.000
##  NA's   :29084   NA's   :29084    NA's   :29084
##  ProsperRating..Alpha.  ProsperScore   ListingCategory..numeric.
##         :29084         Min.   : 1.00   Min.   : 0.000
##  C      :18345         1st Qu.: 4.00   1st Qu.: 1.000
##  B      :15581         Median : 6.00   Median : 1.000
##  A      :14551         Mean   : 5.95   Mean   : 2.774
##  D      :14274         3rd Qu.: 8.00   3rd Qu.: 3.000
##  E      : 9795         Max.   :11.00   Max.   :20.000
##  (Other):12307         NA's   :29084
##  BorrowerState                      Occupation    EmploymentStatusDuration
##  CA     :14717   Other                    :28617   Min.   :  0.00
##  TX     : 6842   Professional             :13628   1st Qu.: 26.00
##  NY     : 6729   Computer Programmer      : 4478   Median : 67.00
##  FL     : 6720   Executive                : 4311   Mean   : 96.07
##  IL     : 5921   Teacher                  : 3759   3rd Qu.:137.00
##         : 5515   Administrative Assistant: 3688   Max.   :755.00
##  (Other):67493   (Other)                  :55456   NA's   :7625
##  IsBorrowerHomeowner CreditScoreRangeLower OpenCreditLines
##  False:56459         Min.   :  0.0         Min.   : 0.00
##  True :57478         1st Qu.:660.0         1st Qu.: 6.00
##                      Median :680.0         Median : 9.00
##                      Mean   :685.6         Mean   : 9.26
##                      3rd Qu.:720.0         3rd Qu.:12.00
##                      Max.   :880.0         Max.   :54.00
##                      NA's   :591           NA's   :7604
##  OpenRevolvingAccounts OpenRevolvingMonthlyPayment AmountDelinquent
##  Min.   : 0.00         Min.   :    0.0             Min.   :     0.0
##  1st Qu.: 4.00         1st Qu.:  114.0             1st Qu.:     0.0
##  Median : 6.00         Median :  271.0             Median :     0.0
##  Mean   : 6.97         Mean   :  398.3             Mean   :   984.5
##  3rd Qu.: 9.00         3rd Qu.:  525.0             3rd Qu.:     0.0
##  Max.   :51.00         Max.   :14985.0             Max.   :463881.0
##                                                    NA's   :7622
##  DelinquenciesLast7Years PublicRecordsLast10Years RevolvingCreditBalance
##  Min.   : 0.000          Min.   : 0.0000          Min.   :      0
##  1st Qu.: 0.000          1st Qu.: 0.0000          1st Qu.:   3121
##  Median : 0.000          Median : 0.0000          Median :   8549
##  Mean   : 4.155          Mean   : 0.3126          Mean   :  17599
##  3rd Qu.: 3.000          3rd Qu.: 0.0000          3rd Qu.:  19521
##  Max.   :99.000          Max.   :38.0000          Max.   :1435667
##  NA's   :990             NA's   :697              NA's   :7604
##  BankcardUtilization TradesNeverDelinquent..percentage. DebtToIncomeRatio
##  Min.   :0.000       Min.   :0.000                      Min.   : 0.000
##  1st Qu.:0.310       1st Qu.:0.820                      1st Qu.: 0.140
##  Median :0.600       Median :0.940                      Median : 0.220
##  Mean   :0.561       Mean   :0.886                      Mean   : 0.276
##  3rd Qu.:0.840       3rd Qu.:1.000                      3rd Qu.: 0.320
##  Max.   :5.950       Max.   :1.000                      Max.   :10.010
##  NA's   :7604        NA's   :7544                       NA's   :8554
```

```
##          IncomeRange    StatedMonthlyIncome LoanOriginalAmount
##  $25,000-49,999:32192   Min.   :      0     Min.   : 1000
##  $50,000-74,999:31050   1st Qu.:   3200     1st Qu.: 4000
##  $100,000+     :17337   Median :   4667     Median : 6500
##  $75,000-99,999:16916   Mean   :   5608     Mean   : 8337
##  Not displayed : 7741   3rd Qu.:   6825     3rd Qu.:12000
##  $1-24,999     : 7274   Max.   :1750003     Max.   :35000
##  (Other)       : 1427
##           LoanOriginationDate LoanOriginationQuarter MonthlyLoanPayment
##  2014-01-22 00:00:00:   491   Q4 2013:14450          Min.   :   0.0
##  2013-11-13 00:00:00:   490   Q1 2014:12172          1st Qu.: 131.6
##  2014-02-19 00:00:00:   439   Q3 2013: 9180          Median : 217.7
##  2013-10-16 00:00:00:   434   Q2 2013: 7099          Mean   : 272.5
##  2014-01-28 00:00:00:   339   Q3 2012: 5632          3rd Qu.: 371.6
##  2013-09-24 00:00:00:   316   Q2 2012: 5061          Max.   :2251.5
##  (Other)            :111428   (Other):60343
##  LP_InterestandFees LP_GrossPrincipalLoss LP_NetPrincipalLoss
##  Min.   :   -2.35   Min.   :  -94.2       Min.   : -954.5
##  1st Qu.:  274.87   1st Qu.:    0.0       1st Qu.:    0.0
##  Median :  700.84   Median :    0.0       Median :    0.0
##  Mean   : 1077.54   Mean   :  700.4       Mean   :  681.4
##  3rd Qu.: 1458.54   3rd Qu.:    0.0       3rd Qu.:    0.0
##  Max.   :15617.03   Max.   :25000.0       Max.   :25000.0
##
##  LP_NonPrincipalRecoverypayments
##  Min.   :    0.00
##  1st Qu.:    0.00
##  Median :    0.00
##  Mean   :   25.14
##  3rd Qu.:    0.00
##  Max.   :21117.90
##
```

Univariate Plots

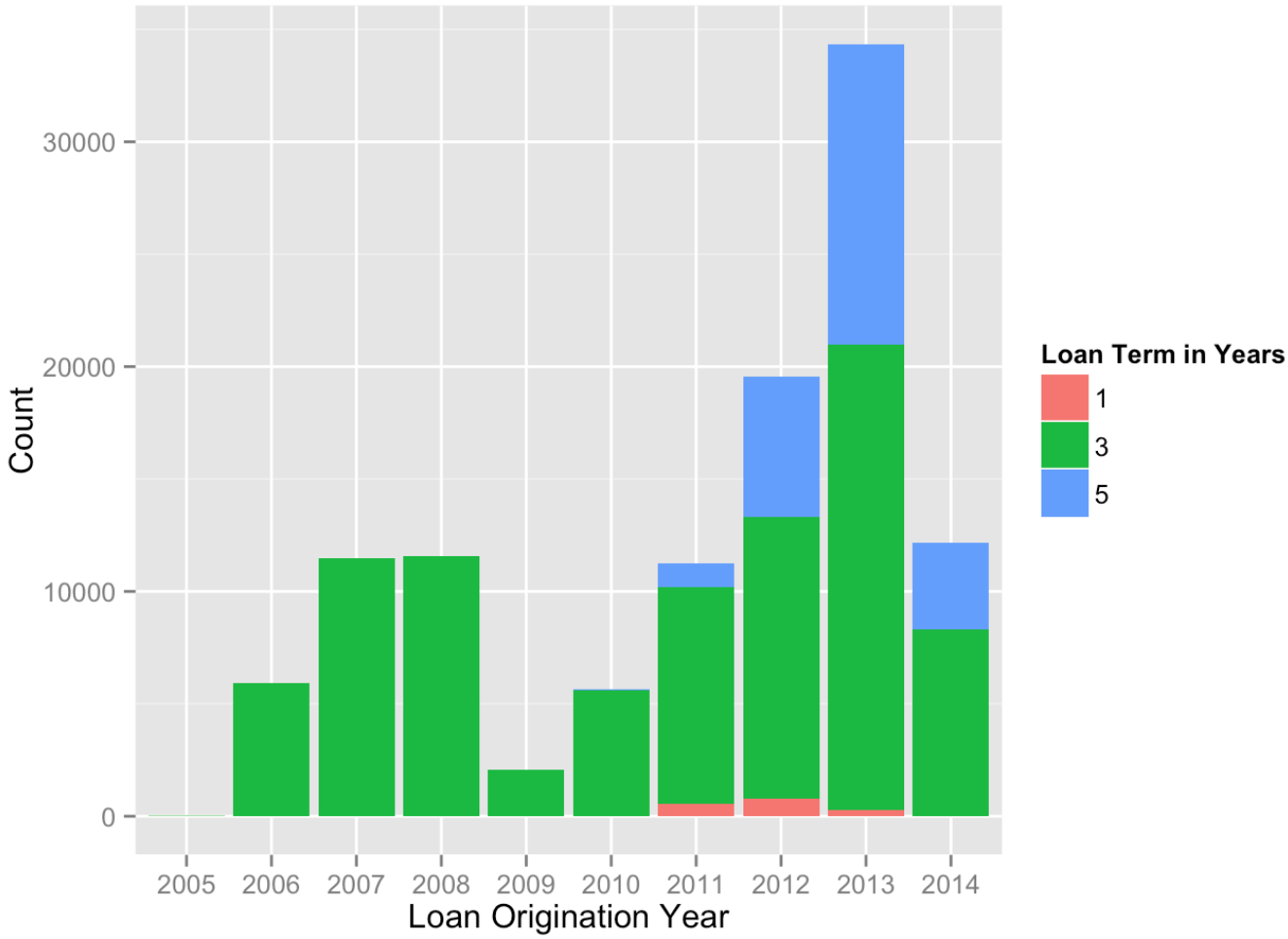


Created new variables for loan origination month and loan origination year and a bar graph by month with a facet by year. The dataset ranges from November 2005 to March 2014. There is a gap in data between November 2008 to June 2009.

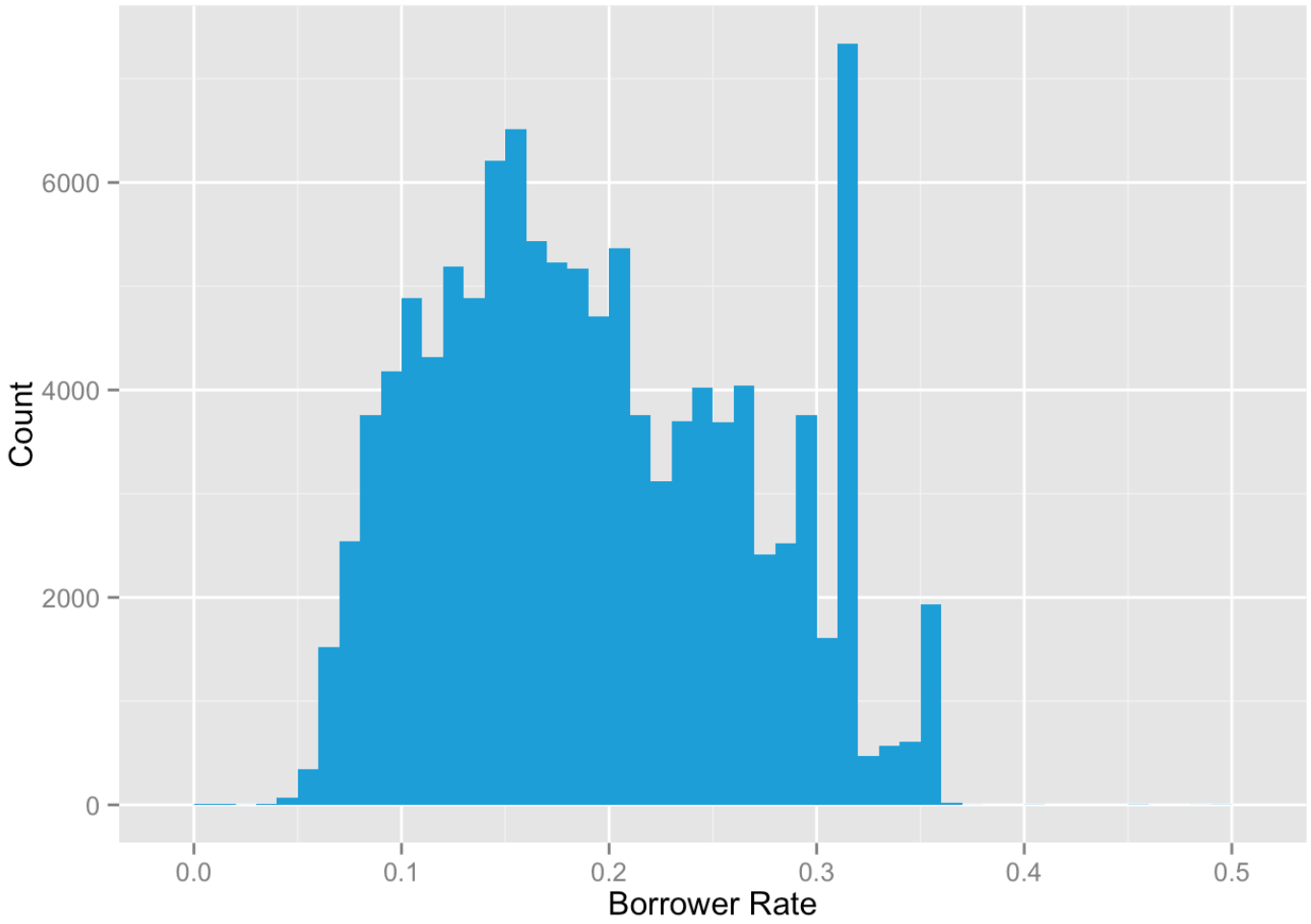
A google search of “Prosper loan November 2008” and the 1st hit is a TechCrunch article (http://techcrunch.com/2008/11/26/sec-outlines-its-reasoning-for-shutting-down-p2p-lender-prosper/) regarding the SEC shut down of peer-to-peer lender Prosper that stopped all lending.

The metadata makes references to several columns only available after July 2009. CreditGrade is applicable for listings pre-2009 and all of the estimated credit yields as well as what appears to be a new prosper rating/scoring system starting in July 2009. This is key information that will be considered in subsequent data exploration and creation of predictive models.

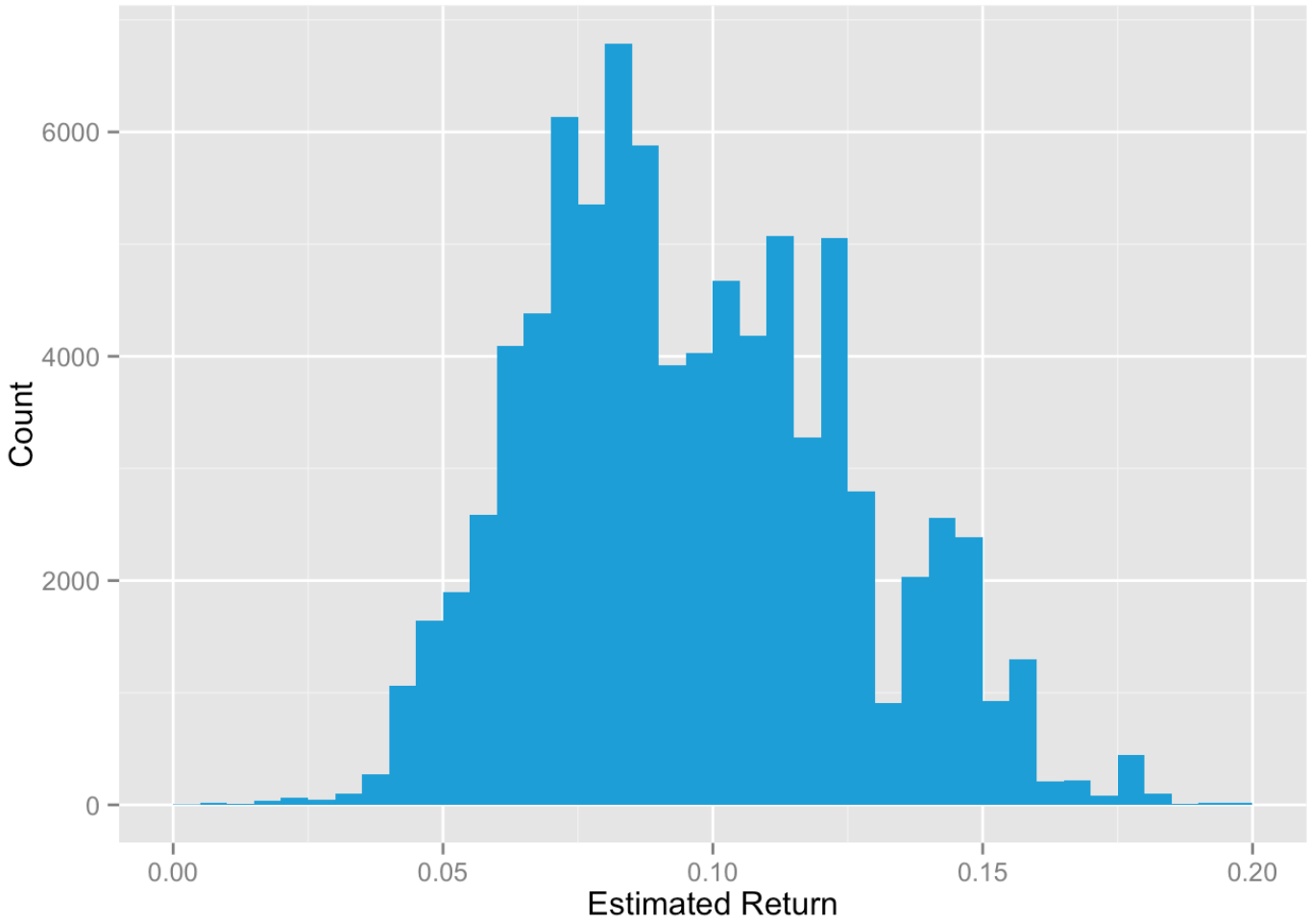
```
## Source: local data frame [3 x 5]
##
##   Term LoanAmtVolume LoanAmtMean LoanAmtMedian LoanCnt
## 1   36     638686342    7276.155          5000   87778
## 2   60     303631410   12370.398         11500   24545
## 3   12       7576595    4694.297          3500    1614
```



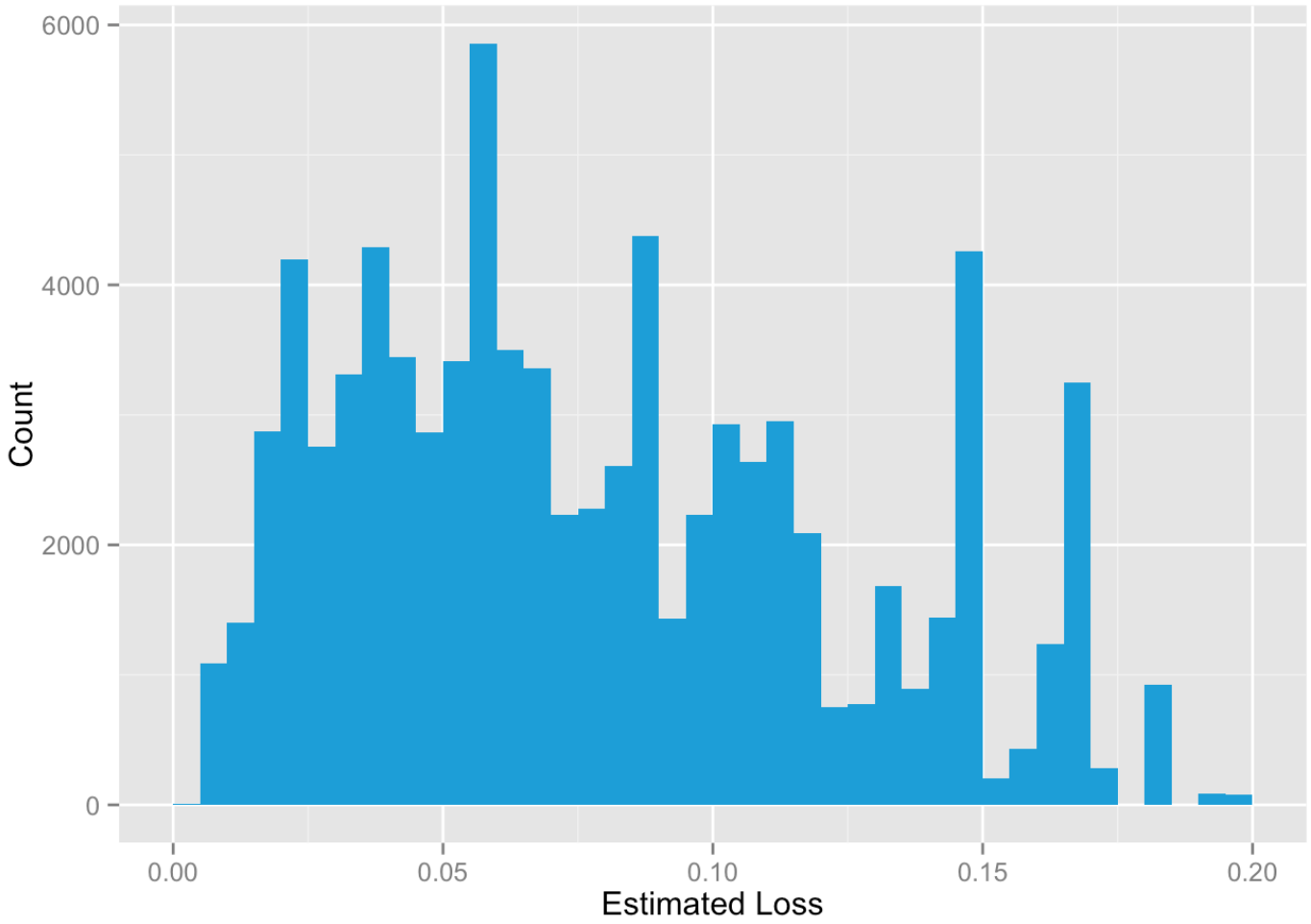
Loan term options are 1, 3 and 5 years. 1 and 5 year loans were not made until 2011. 3 year loans remained the most popular choice.



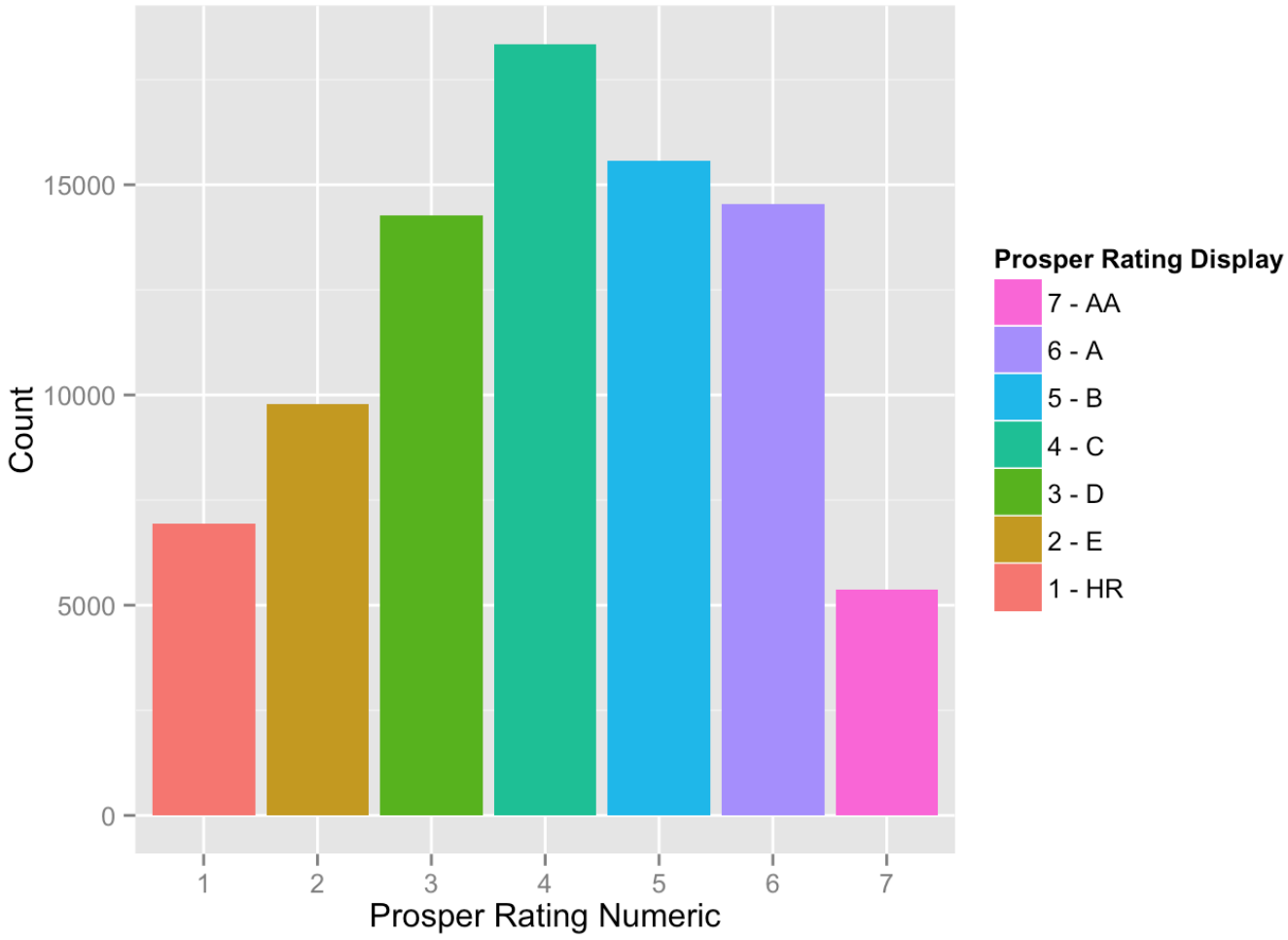
Data is normally distributed but with a spike in loan volume around .31. Median of .1840 and Mean of .1928.



Data is normally distributed. Median of .092 and Mean of .096. Added x limits of 0 and .2 to remove long tail primarily for negative returns to provide a better visualization of the distribution.

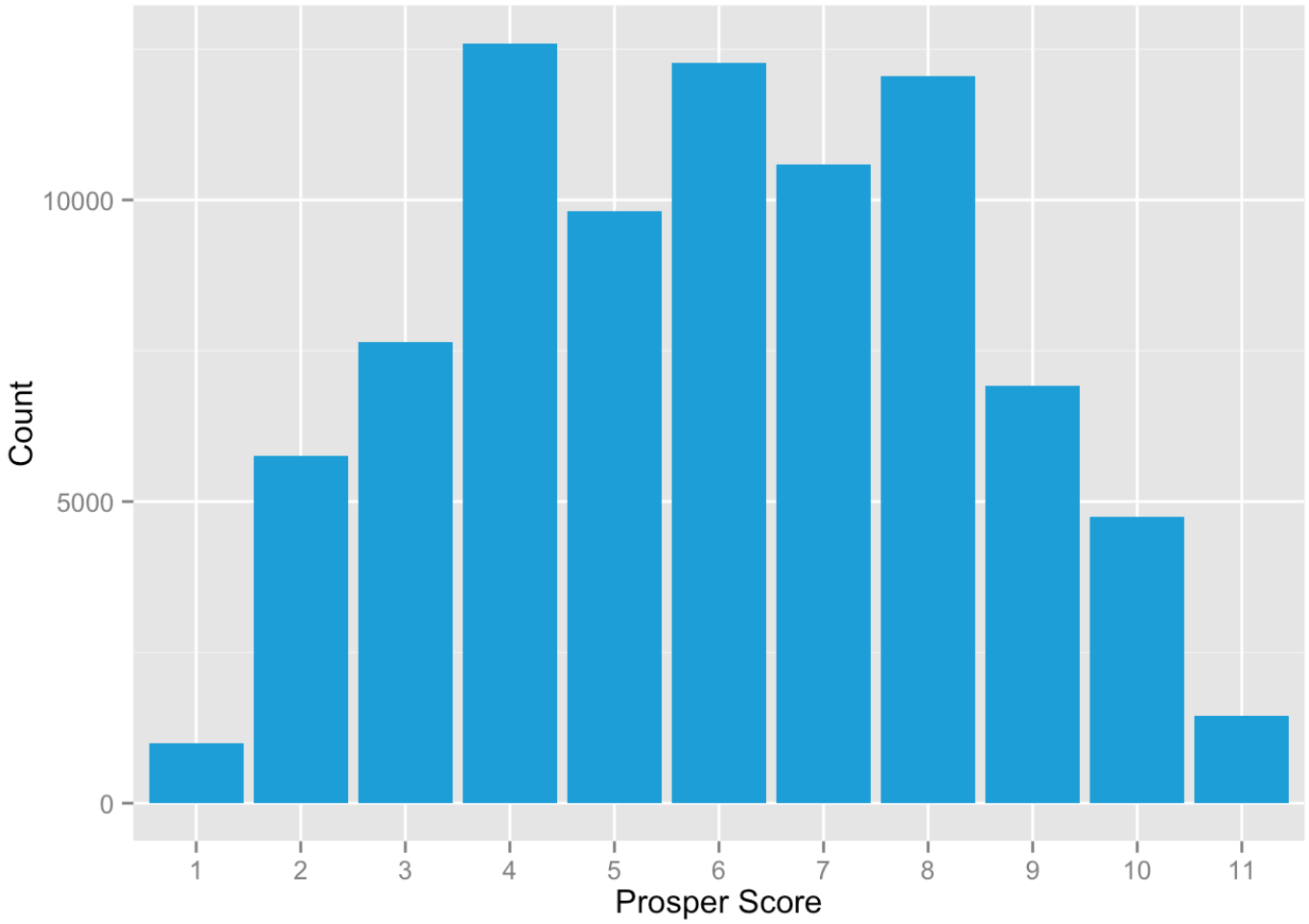


Data closely resembles a plateau distribution with multiple peaks and valleys. Median of .072 and Mean of .080.

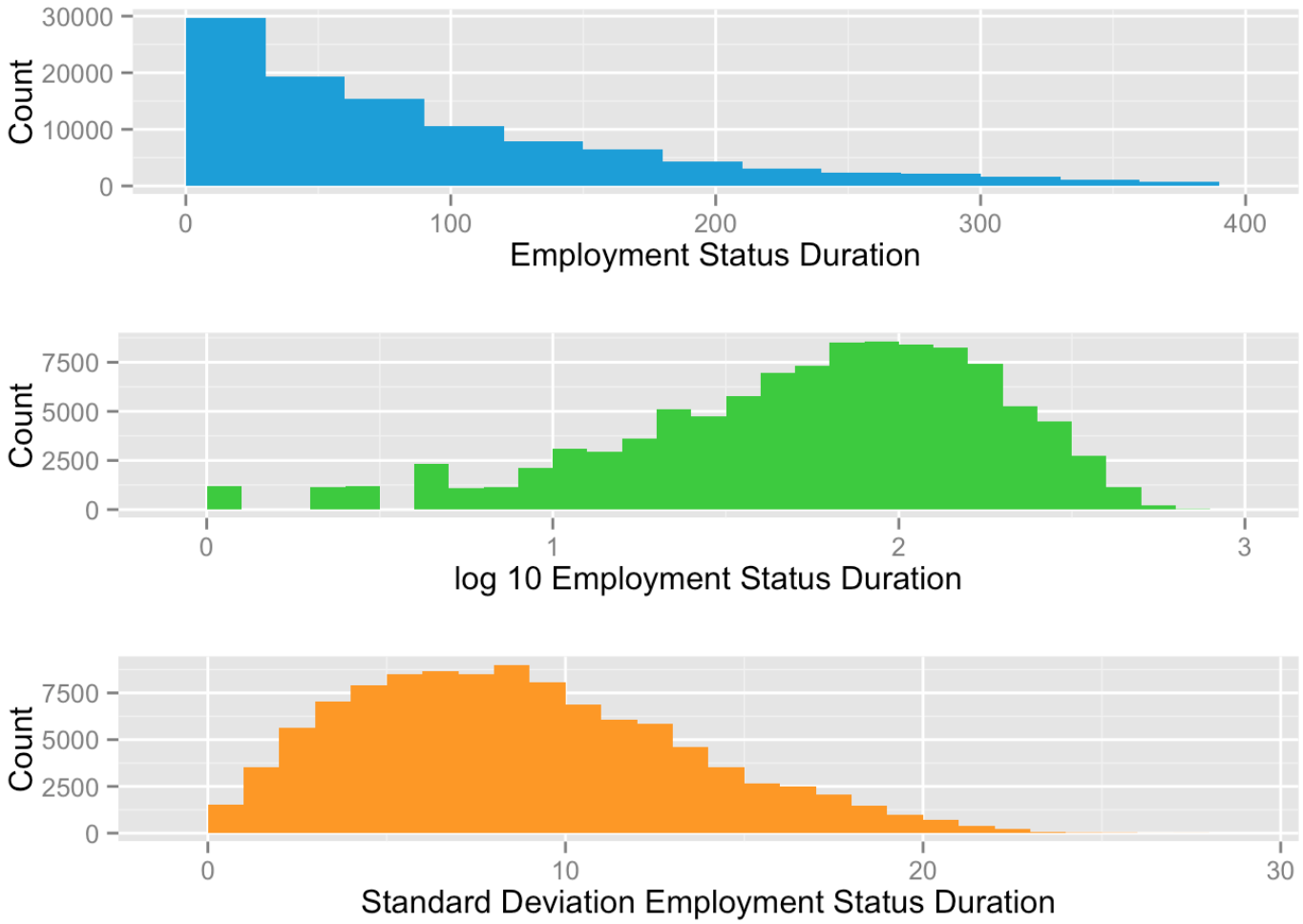


Data is normally distributed. Median of 4.00 and Mean of 4.072. The minimum and maximum rating is 1 and 7 respectively. From worst to best, the alpha scale ranges from HR to AA.

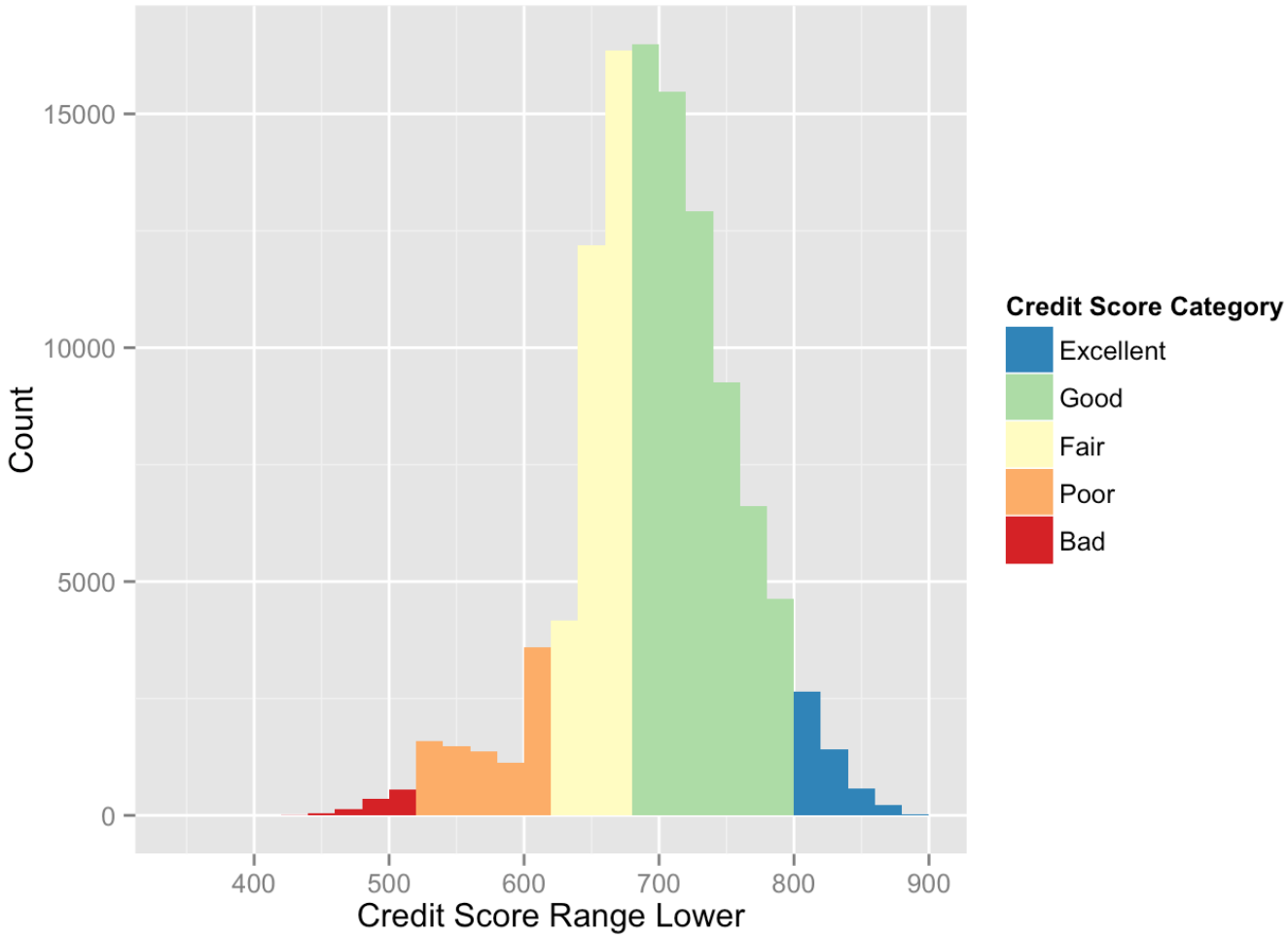
```
## Source: local data frame [11 x 5]
##
##    ProsperScore LoanAmtMean LoanAmtMedian LoanAmtVolume LoanCnt
## 1             1    4570.955          4000       4534387     992
## 2             2    5279.778          4000      30443202    5766
## 3             3    7062.552          4500      53972021    7642
## 4             4    8401.920          7500     105822181   12595
## 5             5    8400.081          7000      82429995    9813
## 6             6    9222.604          8000     113235137   12278
## 7             7   10097.153          9500     106999534   10597
## 8             8   10487.978         10000     126411602   12053
## 9             9   10055.976          8300      69496847    6911
## 10           10   11742.895         10000      55778753    4750
## 11           11   14858.186         15000      21633519    1456
```



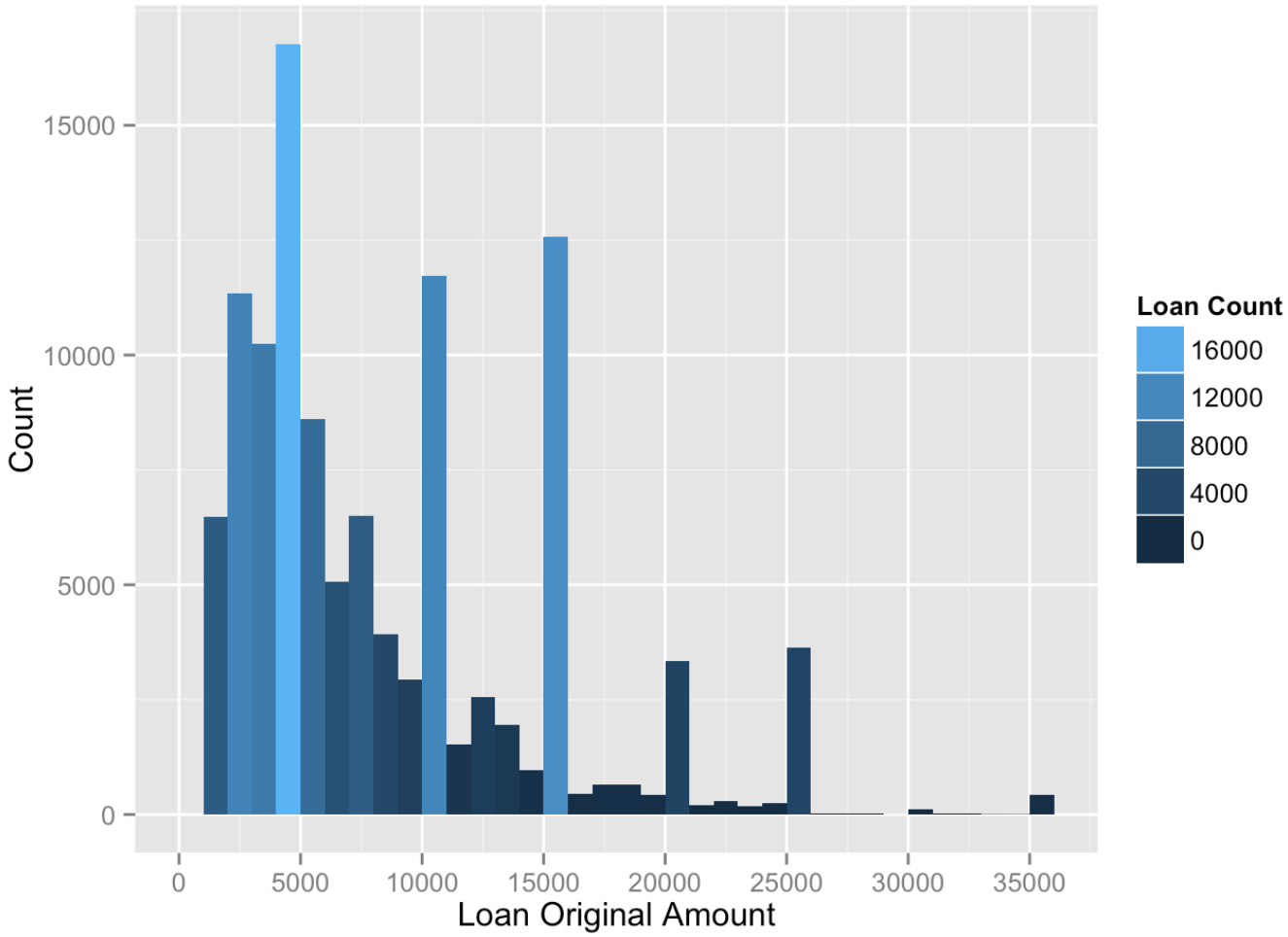
Data is normally distributed with similar peaks at 4, 6 and 8. The maximum score is 11 which is inconsistent with the metadata that indicates score ranges from 1 to 10, with 10 being the best.



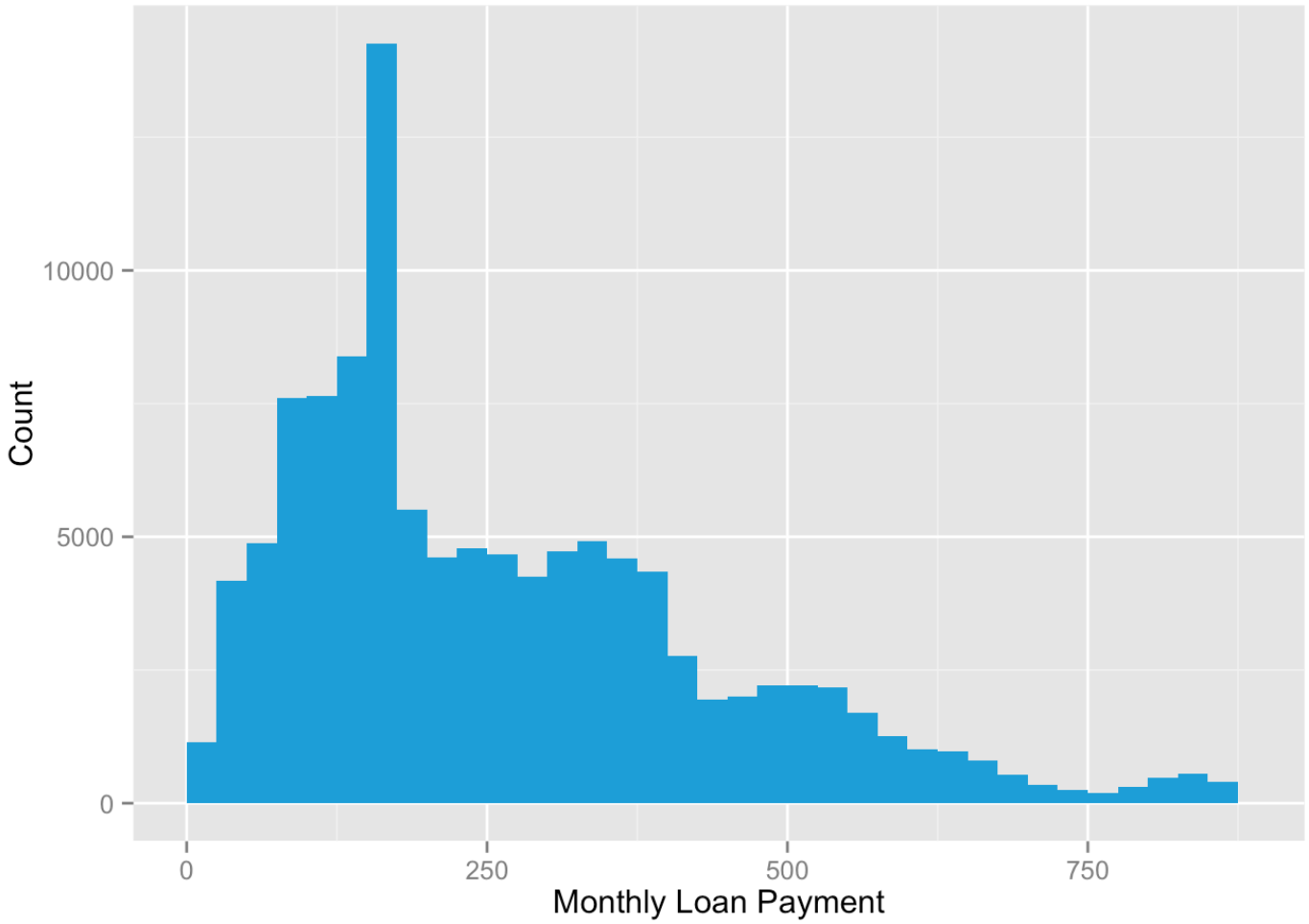
Data is positively skewed for shorter employment lengths. Added a log10 and square root transformation arranged in a 1 column grid. Median of 67.00 and Mean of 96.07.



Data is normally distributed. Median of 680.0 and Mean of 685.6. Filtered out records with invalid credit score = 0 to remove left long tail. Used credit score range lower since rate qualification is typically based on the lower score in multi-credit scoring pricing model. Added spectral color pallette for Bad to Excellent credit score ranges.



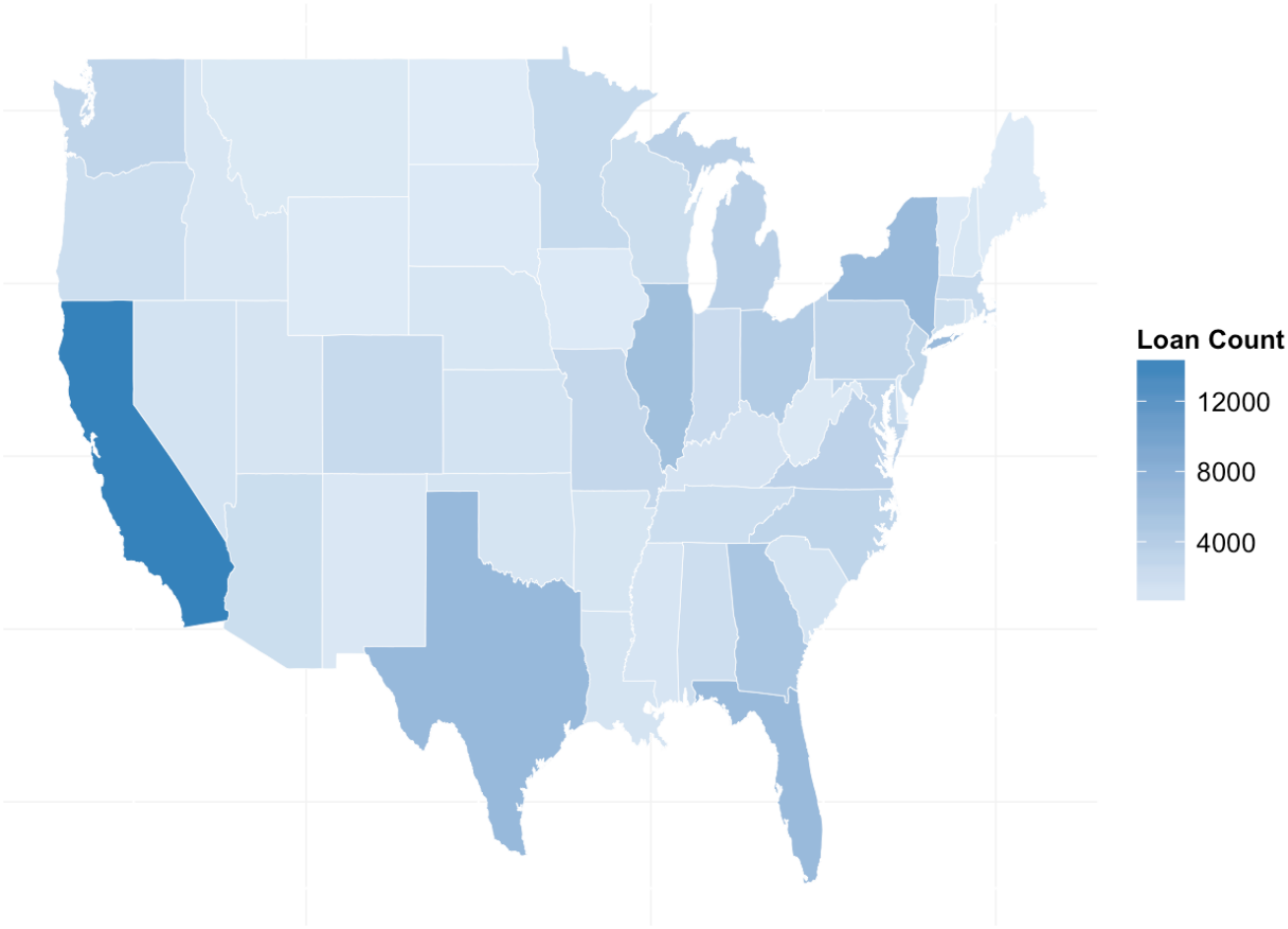
Data is positively skewed. Added tick marks every 5000 since peaks show loan amounts appear to be more common in 5K increments (5K - 25K). Median of 6500.0 and Mean of 8337.0.



Data is positively skewed. Added x axis limit to exclude long tail for monthly payments greater than 99% quantile. Median of 217.7 and Mean of 272.5.

Top 10 loan volume by BorrowerState:

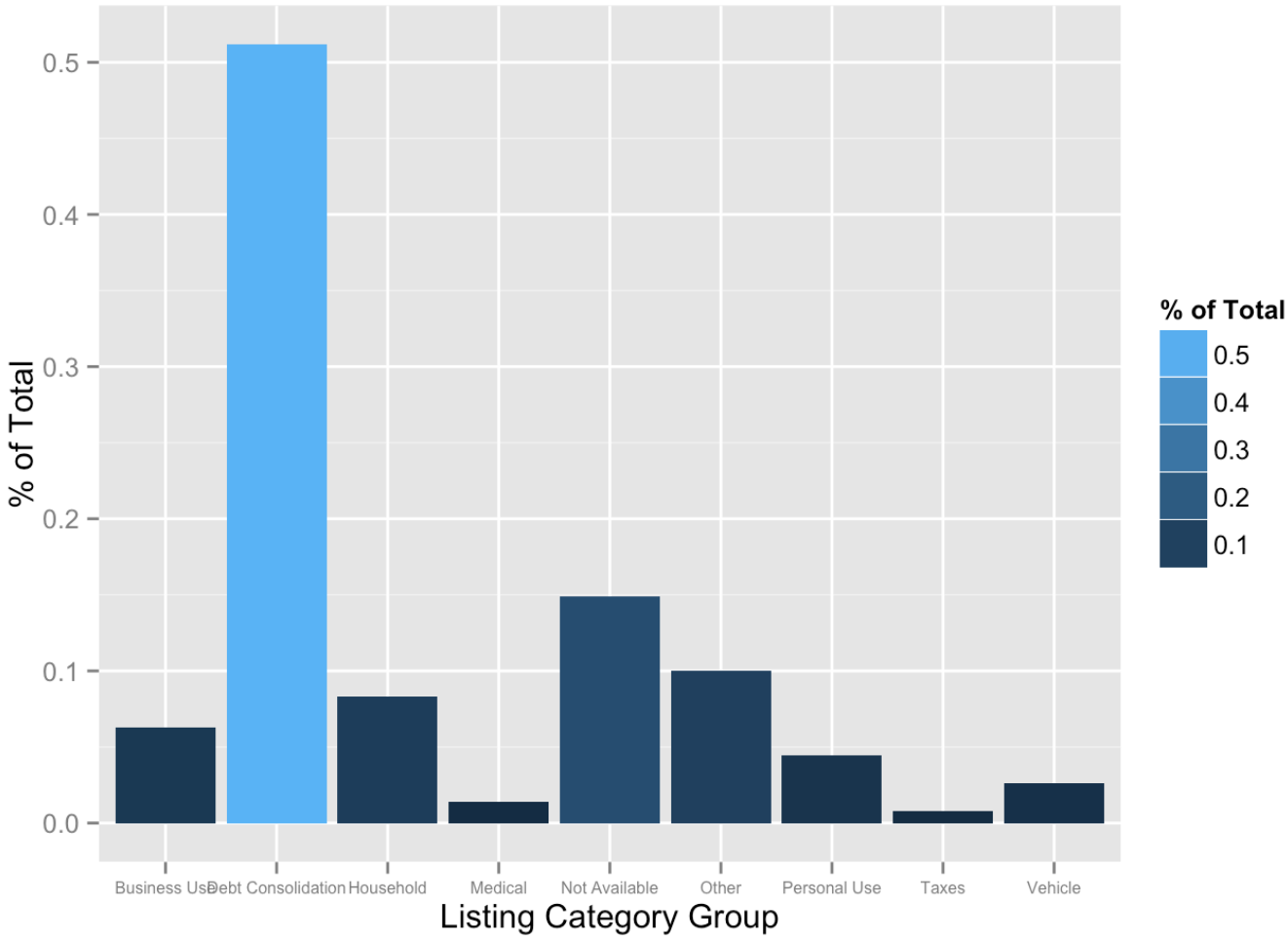
```
## Source: local data frame [51 x 5]
##
##    BorrowerState LoanAmtVolume LoanAmtMean LoanAmtMedian LoanCnt
## 1             CA     132075153    8974.326          7000   14717
## 2             TX      62179088    9087.853          7500    6842
## 3             NY      59437488    8833.034          7000    6729
## 4             FL      55154135    8207.461          6500    6720
## 5             IL      49712307    8395.931          6500    5921
## 6             GA      41881214    8362.862          6000    5008
## 7             OH      33904448    8078.258          6500    4197
## 8             MI      27469230    7645.207          5250    3593
## 9             VA      29408372    8971.437          7500    3278
## 10            NJ      29511373    9529.019          8000    3097
## ..           ...           ...         ...           ...     ...
```



Added map plot to show dominant CA market.

Loan volume by custom ListingCategoryGroup variable:

```
## Source: local data frame [9 x 5]
##
##   ListingCategoryGroup LoanAmtVolume LoanAmtMean LoanAmtMedian LoanCnt
## 1   Debt Consolidation     577736197    9908.352          9500   58308
## 2        Not Available     106096621    6253.853          4500   16965
## 3                Other      69719962    6131.923          4000   11370
## 4            Household      71197236    7503.925          5000    9488
## 5         Business Use      64175191    8926.859          7279    7189
## 6         Personal Use      28095040    5502.358          4000    5106
## 7              Vehicle      15718287    5216.823          4000    3013
## 8              Medical      10447136    6476.836          4000    1613
## 9                Taxes       6708677    7580.426          5000     885
```



Debt consolidation loans are the most common loan type representing more than 50% of the total.

Summary of LoanStatus and new calculated LoanStatusBucket variable:

```
##               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
```

```
## Cancelled    Closed      Open
##         5     55084     58848
```

New variable PrincipalLoss flag:

```
##
##     0     1
## 97291 16646
```

The principal Loss variable will be used to identify loans that have defaulted and any principal amount was charged off.

New variable BorHomeowner flag:

```
##
##     0     1
## 56459 57478
```

Converted True/False text field.

Univariate Summary

What is the structure of your dataset?

There are 113,937 loans in the dataset with 81 total columns. Based on existing domain knowledge of the lending industry, I immediately isolated specific columns for further analysis. However, as I conducted the single variable analysis I went back and made revisions to my column list. For example, once I made the determination that the credit rating system changed in July 2009 I added each of the Prosper Rating/Score variables and removed Credit Grade.

Loan origination volume range from November 2005 to March 2014. There is no data between November 2008 to June 2009 due to the SEC shut down of Prosper. Loan term options are 1, 3 and 5 years. Prosper did not make 1 and 5 year loans until 2011 but the 3 year loan remained the most popular choice.

Prosper has an ordered factor variable for credit rating as well as an accompanying numeric variable and custom risk score.

Worst -> Best
Prosper rating alpha: HR, E, D, C, B, A, AA
Prosper rating numeric: 1 - 7
Prosper score: 1 - 11

All rating/risk score fields are NA pre-2009 so the datset will be filtered accordingly. Based on the variable definitions the maximum risk score was expected to be 10.

Median rate is 18.4% with spike in volume at 31%.
Median estimated return and loss is 9.2% and 7.2% respectively.
Median employment duration is 67 months.
Median credit score lower is 680. Median loan amount is $6500.

CA has the highest volume of loans and Debt Consolidation is the top listing category.

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

The main features of interest in the dataset rate, loan amount and prosper rating My perspective is from an investor point of view and investigating the relationship of customer profiles and probability of the loan defaulting and loss of principal.

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

The features that will be explored further are:

- Term
- EstimatedLoss
- EstimatedReturn
- EmploymentStatusDuration
- BorHomeowner
- CreditScoreRangeLower
- AmountDelinquent
- DelinquenciesLast7Years
- PublicRecordsLast10Years
- RevolvingCreditBalance
- OpenCreditLines
- OpenRevolvingAccounts
- OpenRevolvingMonthlyPayment
- BankcardUtilization
- TradesNeverDelinquent..percentage.
- DebtToIncomeRatio
- StatedMonthlyIncome
- PrincipalLoss

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

The dataset includes estimated return and loss rates but these are assigned at the time the loan listing was created. The dataset does not have average daily balance data to calculate actual rates so net principal loss was used to calculate a principal loss flag.

In addition, new variables were created for loan origination month and year, credit score category (Bad, Poor, Fair, Good, Excellent), listing category group and loan status bucket (Cancelled, Open, Closed).

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 positive skew of employment status duration was transformed using log 10 and standard deviation. Due to the gap in data between November 2008 to June 2009 and introduction of the prosper rating and risk score metrics, the dataset will be filtered for loans >= 2009 for analysis on these fields.

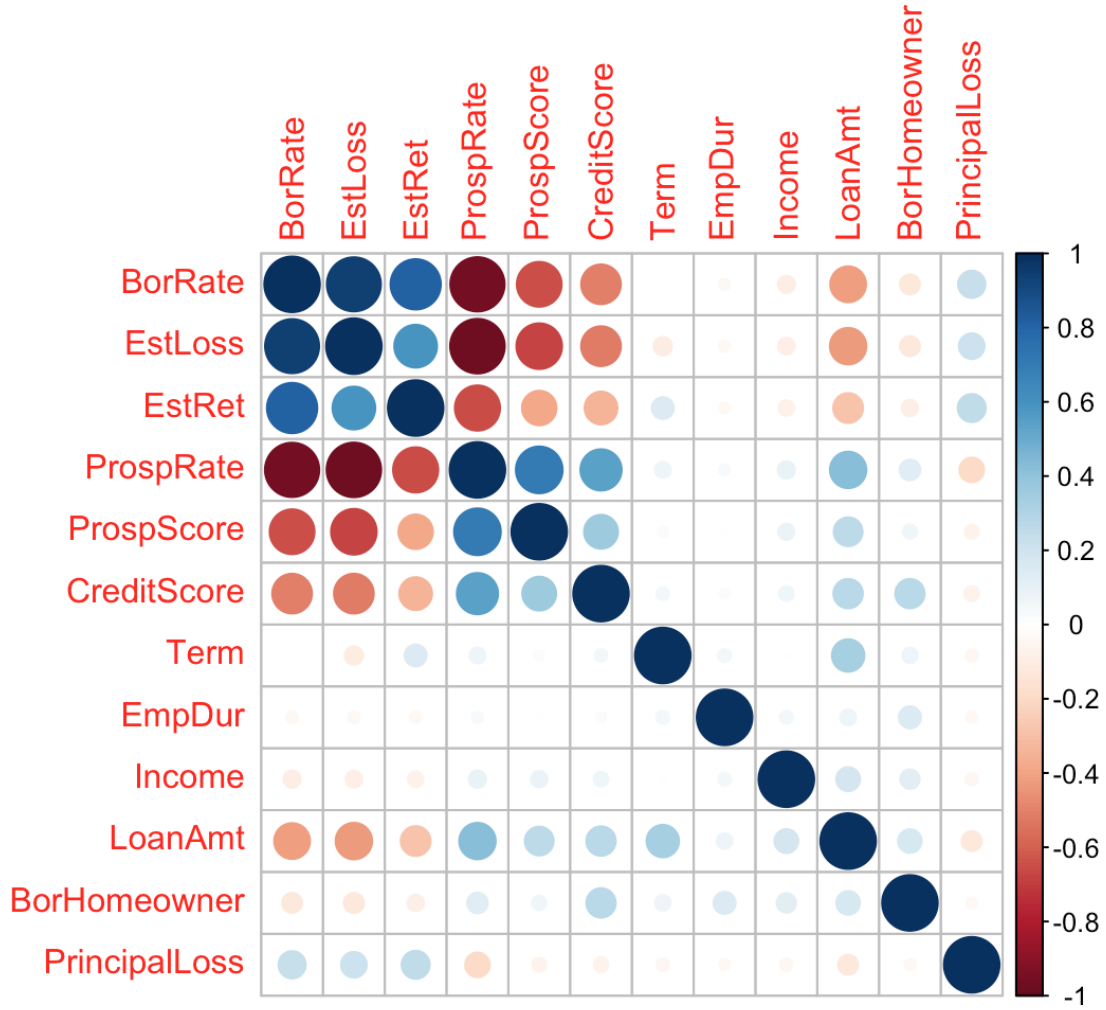
Rate has an unusual spike in loan counts at 31% with median values of 18.40% and Mean of 19.28%. Estimated loss has what closely resembles a plateau distribution with multiple peaks at similar heights.

Prosper rating and prosper score are normally distributed but the prosper score has unusual peaks at 4, 6 and 8. This will be explored further but will focus on prosper rating in subsequent analysis.

Bivariate Plots

Correlation matrix 1 for rate, credit rating and key loan fields:

```
##                    BorRate     EstLoss      EstRet   ProspRate
## BorRate        1.0000000000  0.94529248  0.81767854 -0.95310541
## EstLoss        0.9452924831  1.00000000  0.59105633 -0.96418058
## EstRet         0.8176785371  0.59105633  1.00000000 -0.65998874
## ProspRate     -0.9531054149 -0.96418058 -0.65998874  1.00000000
## ProspScore    -0.6497455552 -0.67371541 -0.38326305  0.70522276
## CreditScore   -0.5086563005 -0.51123953 -0.34620303  0.54884709
## Term          -0.0000762274 -0.10712738  0.15258177  0.07915118
## EmpDur        -0.0391820185 -0.03916047 -0.03648651  0.03607306
## Income        -0.0934771070 -0.08925538 -0.07500608  0.09430718
## LoanAmt       -0.4135014440 -0.42995566 -0.28608272  0.42861533
## BorHomeowner  -0.1261559744 -0.12712601 -0.08613725  0.13645771
## PrincipalLoss  0.2387775340  0.21135582  0.25191666 -0.19520939
##                 ProspScore CreditScore          Term       EmpDur
## BorRate       -0.649745555 -0.50865630 -0.0000762274 -0.039182018
## EstLoss       -0.673715407 -0.51123953 -0.1071273760 -0.039160473
## EstRet        -0.383263051 -0.34620303  0.1525817714 -0.036486506
## ProspRate      0.705222758  0.54884709  0.0791511793  0.036073061
## ProspScore     1.000000000  0.36960569  0.0289477016 -0.007302109
## CreditScore    0.369605692  1.00000000  0.0502557933  0.029313163
## Term           0.028947702  0.05025579  1.0000000000  0.052555496
## EmpDur        -0.007302109  0.02931316  0.0525554963  1.000000000
## Income         0.083777108  0.06770702  0.0092381215  0.051380442
## LoanAmt        0.266386099  0.27786723  0.3390362058  0.078216651
## BorHomeowner   0.064430345  0.27692463  0.0760532815  0.155986596
## PrincipalLoss -0.062738593 -0.06963978 -0.0459869079 -0.036297552
##                     Income     LoanAmt BorHomeowner PrincipalLoss
## BorRate       -0.093477107 -0.41350144  -0.12615597    0.23877753
## EstLoss       -0.089255381 -0.42995566  -0.12712601    0.21135582
## EstRet        -0.075006076 -0.28608272  -0.08613725    0.25191666
## ProspRate      0.094307176  0.42861533   0.13645771   -0.19520939
## ProspScore     0.083777108  0.26638610   0.06443034   -0.06273859
## CreditScore    0.067707020  0.27786723   0.27692463   -0.06963978
## Term           0.009238121  0.33903621   0.07605328   -0.04598691
## EmpDur         0.051380442  0.07821665   0.15598660   -0.03629755
## Income         1.000000000  0.18283792   0.12027047   -0.04688319
## LoanAmt        0.182837920  1.00000000   0.17783717   -0.12620284
## BorHomeowner   0.120270472  0.17783717   1.00000000   -0.03704747
## PrincipalLoss -0.046883194 -0.12620284  -0.03704747    1.00000000
```

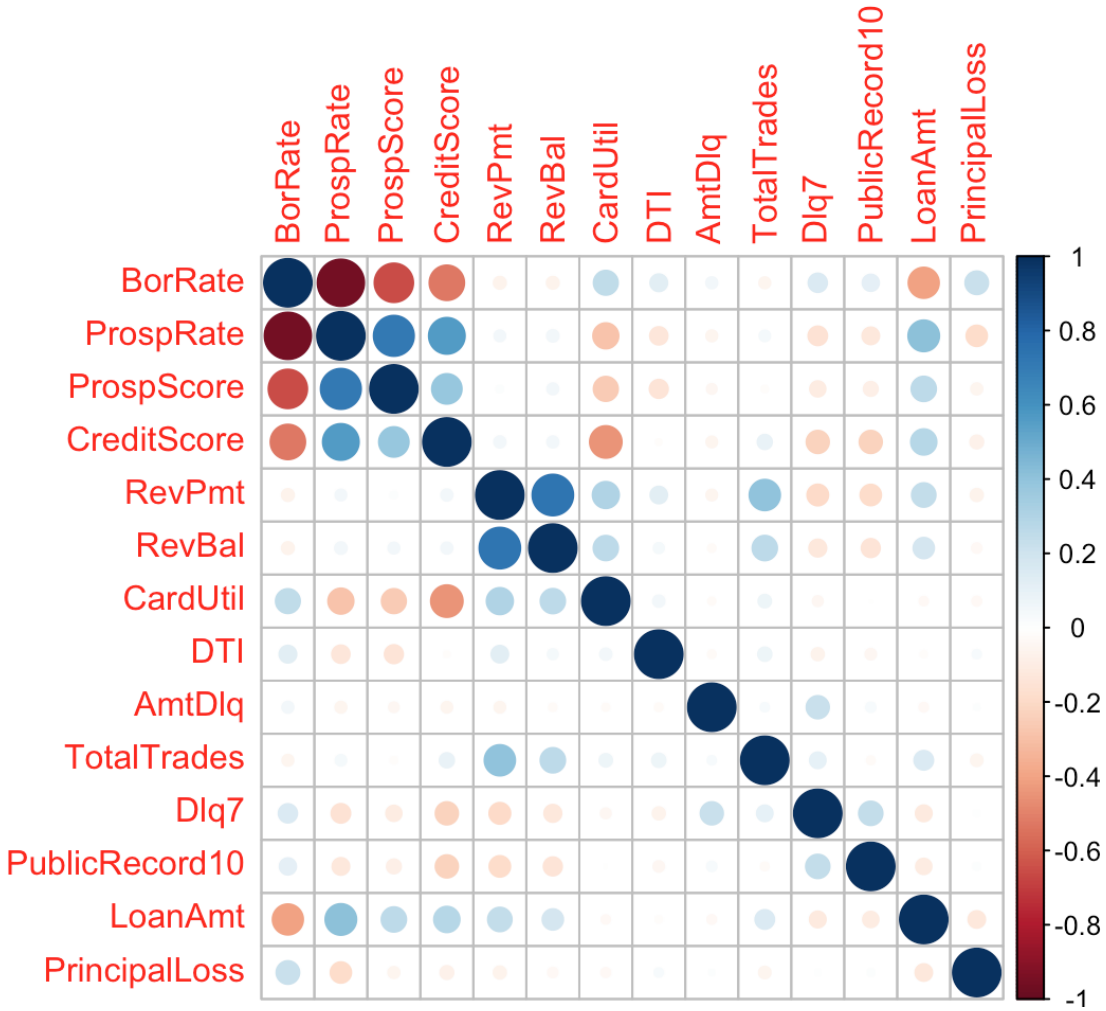


In the top left of the correlation matrix, rate, estimated loss, estimated return, prosper rating, prosper score and credit score all have high correlation and will be explored further. Loan amount and principal loss are the additional fields of interest.

Correlation matrix 2 for rate, credit rating and key credit reporting fields:

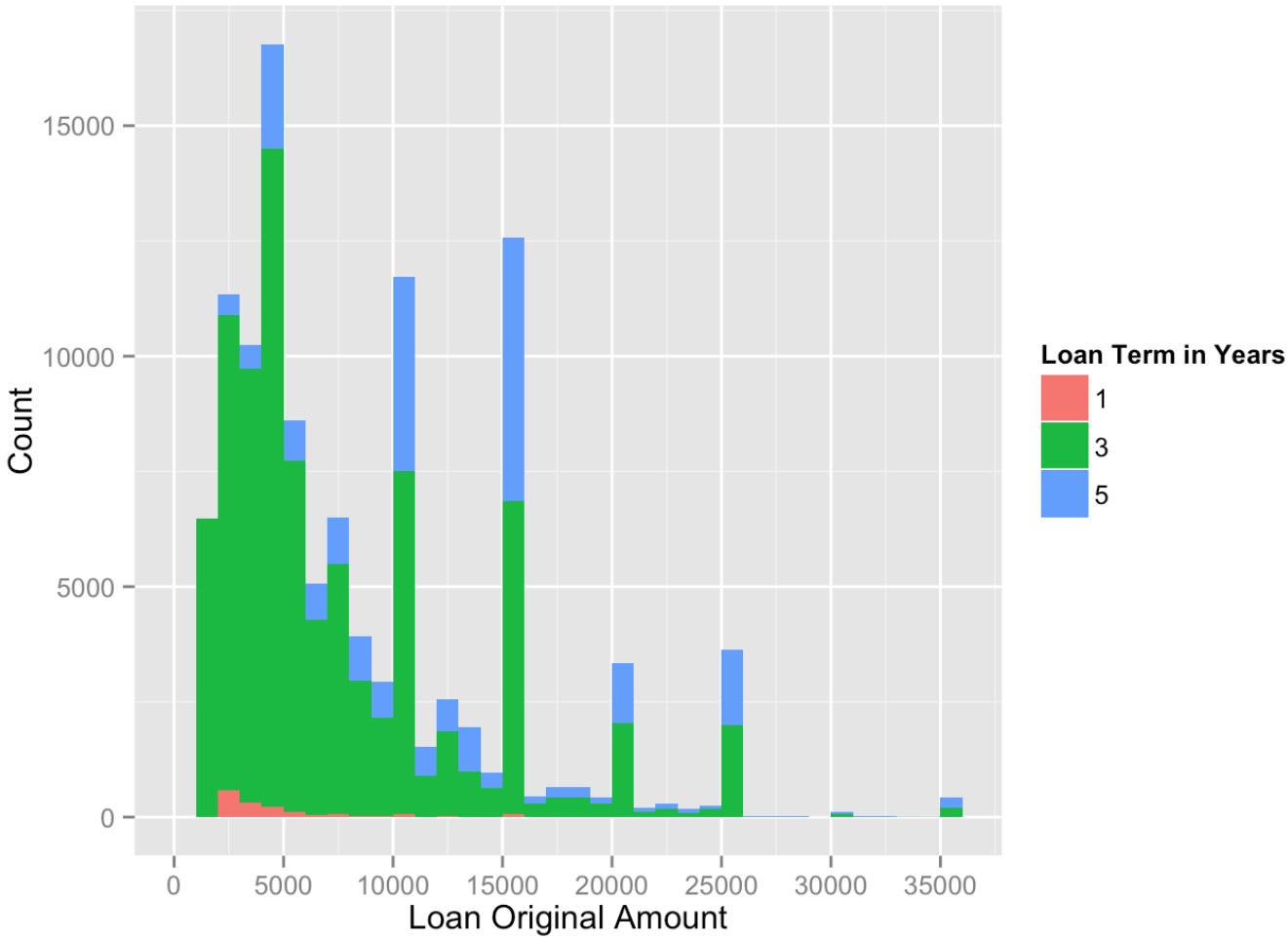
```
##                    BorRate   ProspRate  ProspScore CreditScore      RevPmt
## BorRate         1.00000000 -0.95346747 -0.65657254 -0.52854934 -0.06348382
## ProspRate      -0.95346747  1.00000000  0.71059189  0.56864137  0.05193759
## ProspScore     -0.65657254  0.71059189  1.00000000  0.38700058  0.01799715
## CreditScore    -0.52854934  0.56864137  0.38700058  1.00000000  0.05658146
## RevPmt         -0.06348382  0.05193759  0.01799715  0.05658146  1.00000000
## RevBal         -0.06298067  0.05991414  0.05360869  0.05534997  0.73248926
## CardUtil        0.25520817 -0.28093930 -0.25744548 -0.44309234  0.30342525
## DTI             0.12642797 -0.13534359 -0.14533589 -0.01370880  0.12308564
## AmtDlq          0.05485593 -0.05344622 -0.04223792 -0.05137617 -0.05162808
## TotalTrades    -0.05316765  0.04959010 -0.01792112  0.09049613  0.40908969
## Dlq7            0.15018934 -0.15500180 -0.10399108 -0.22023679 -0.19232792
## PublicRecord10  0.11825403 -0.12504463 -0.08678053 -0.22029099 -0.18770705
## LoanAmt        -0.40541402  0.41975417  0.26446056  0.28580939  0.24613633
## PrincipalLoss   0.22883249 -0.18520392 -0.05485596 -0.07422851 -0.06615066
##                     RevBal     CardUtil          DTI      AmtDlq
## BorRate        -0.06298067  0.255208165  0.12642797  0.05485593
```

```
## ProspRate       0.05991414 -0.280939297 -0.13534359 -0.05344622
## ProspScore      0.05360869 -0.257445475 -0.14533589 -0.04223792
## CreditScore     0.05534997 -0.443092336 -0.01370880 -0.05137617
## RevPmt          0.73248926  0.303425250  0.12308564 -0.05162808
## RevBal          1.00000000  0.262865719  0.04485603 -0.02471830
## CardUtil        0.26286572  1.000000000  0.05744962 -0.02237036
## DTI             0.04485603  0.057449615  1.00000000 -0.02694389
## AmtDlq         -0.02471830 -0.022370358 -0.02694389  1.00000000
## TotalTrades     0.26572616  0.070371237  0.07937152  0.03118406
## Dlq7           -0.12852593 -0.044270080 -0.06634706  0.22074222
## PublicRecord10 -0.14576905 -0.001275848 -0.04997642  0.03996097
## LoanAmt         0.18488691 -0.031467564 -0.01783746 -0.03381422
## PrincipalLoss  -0.03966827 -0.036125729  0.03149852  0.01147444
##                TotalTrades         Dlq7 PublicRecord10     LoanAmt
## BorRate        -0.05316765  0.150189336    0.118254033 -0.40541402
## ProspRate       0.04959010 -0.155001803   -0.125044632  0.41975417
## ProspScore     -0.01792112 -0.103991082   -0.086780531  0.26446056
## CreditScore     0.09049613 -0.220236790   -0.220290987  0.28580939
## RevPmt          0.40908969 -0.192327918   -0.187707048  0.24613633
## RevBal          0.26572616 -0.128525932   -0.145769051  0.18488691
## CardUtil        0.07037124 -0.044270080   -0.001275848 -0.03146756
## DTI             0.07937152 -0.066347055   -0.049976416 -0.01783746
## AmtDlq          0.03118406  0.220742216    0.039960974 -0.03381422
## TotalTrades     1.00000000  0.109393082   -0.027460673  0.15630683
## Dlq7            0.10939308  1.000000000    0.243622496 -0.11199231
## PublicRecord10 -0.02746067  0.243622496    1.000000000 -0.10524710
## LoanAmt         0.15630683 -0.111992312   -0.105247095  1.00000000
## PrincipalLoss  -0.05948305  0.009173272    0.017404246 -0.12288087
##                PrincipalLoss
## BorRate          0.228832488
## ProspRate       -0.185203923
## ProspScore      -0.054855959
## CreditScore     -0.074228510
## RevPmt          -0.066150663
## RevBal          -0.039668271
## CardUtil        -0.036125729
## DTI              0.031498522
## AmtDlq           0.011474438
## TotalTrades     -0.059483047
## Dlq7             0.009173272
## PublicRecord10   0.017404246
## LoanAmt         -0.122880873
## PrincipalLoss    1.000000000
```



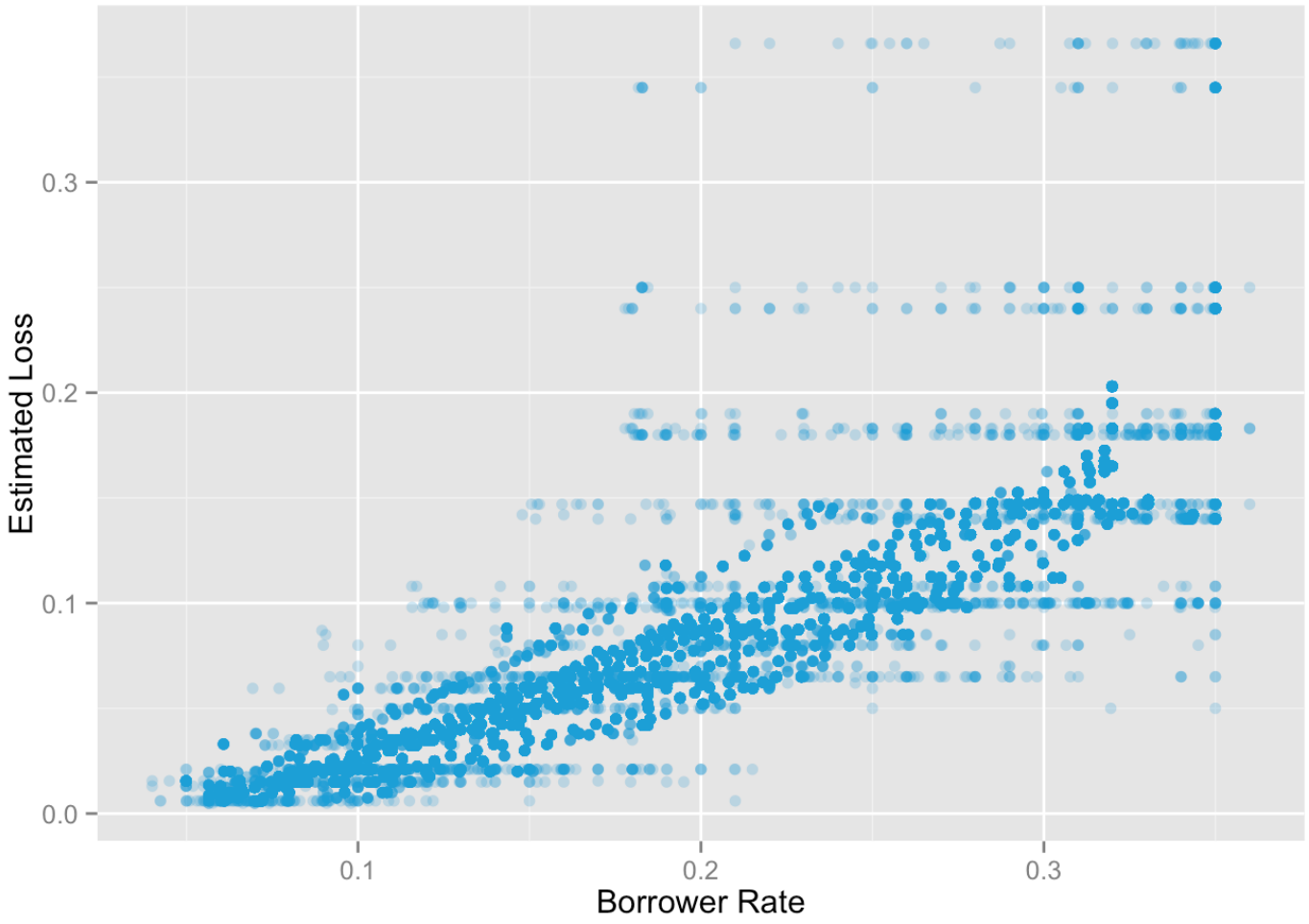
Due to number variables in the dataset, this correlation matrix now explores the various credit reporting fields and relationship with rate, prosper rating, prosper score and credit score. Card utilization and credit score as well as total trades/revolving balance and revolving payment have a good relationship but no other variables stand out for further exploration.

It was assumed that most of these credit reporting fields would have some factor in the credit score so it is interesting to see that card utilization has the highest correlation of all credit related fields.

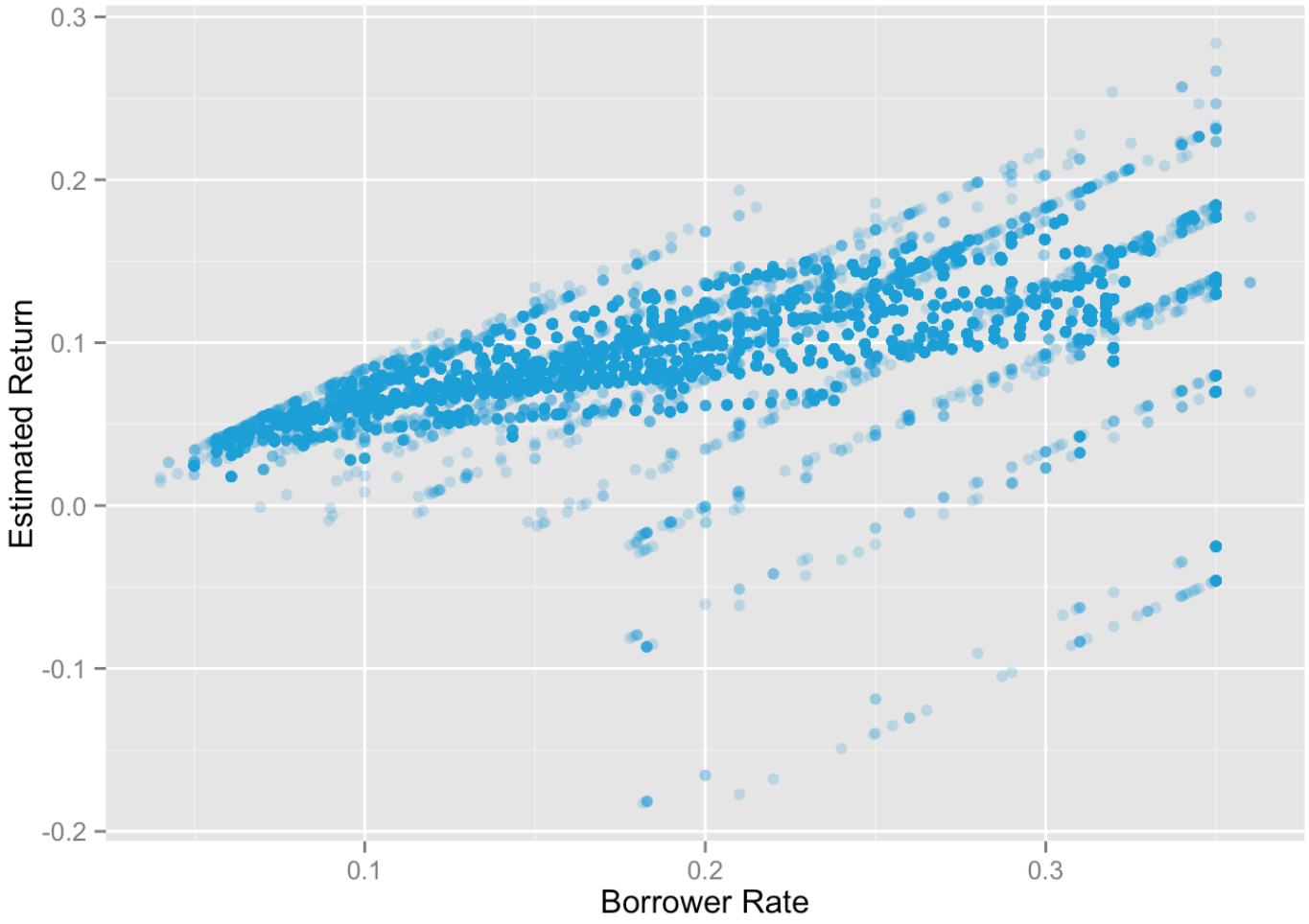


3 year loans are the most common across all loan amounts.

I will now explore rate and estimated loss, estimated return, prosper rating, credit score, loan amount and income.

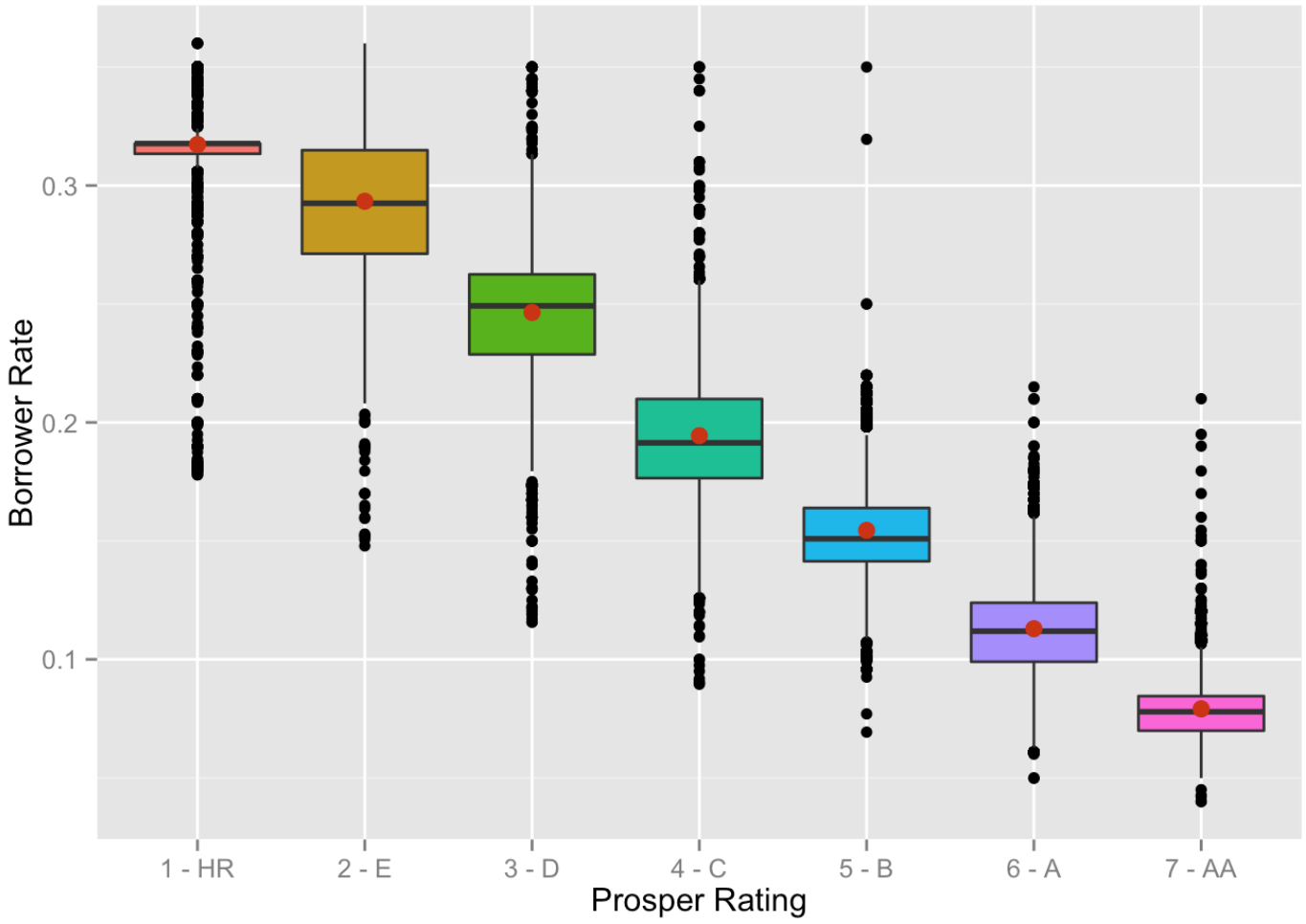


As rate increases, estimated loss for the loan increases. At rates at 20% and higher there are vertical bands for higher loss estimates. This will be explored in the multivariate analysis section to identify the types of loans that represent these higher loss estimates.

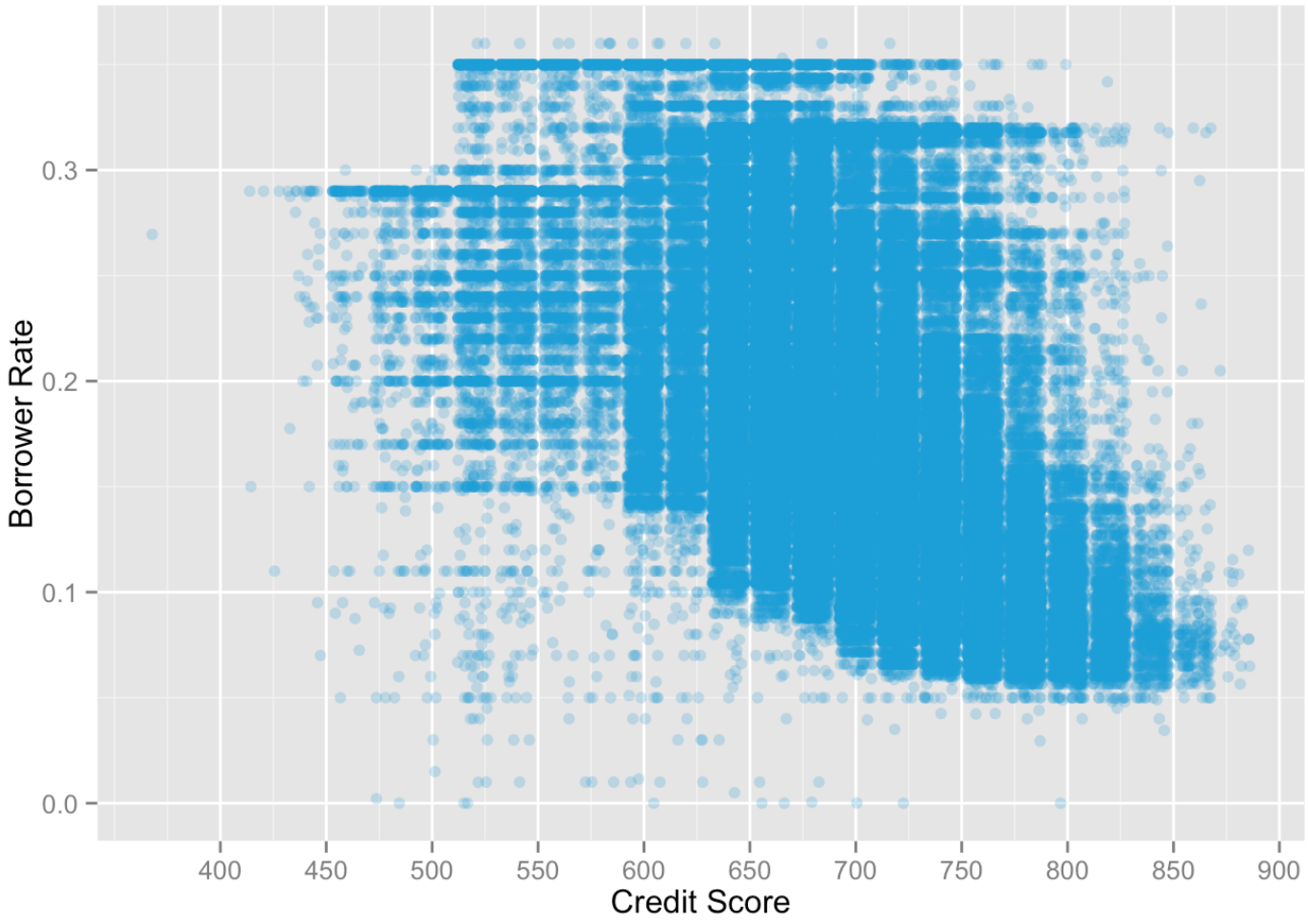


As rate increases, estimated return for the loan increases. At rates at 20% and higher there are vertical bands for lower return estimates. This is consistent since as the estimated loss increases, estimated return decreases. This will be explored in the multivariate analysis section to identify the types of loans that represent these lower return estimates. It is noted some loans have estimated negative returns.

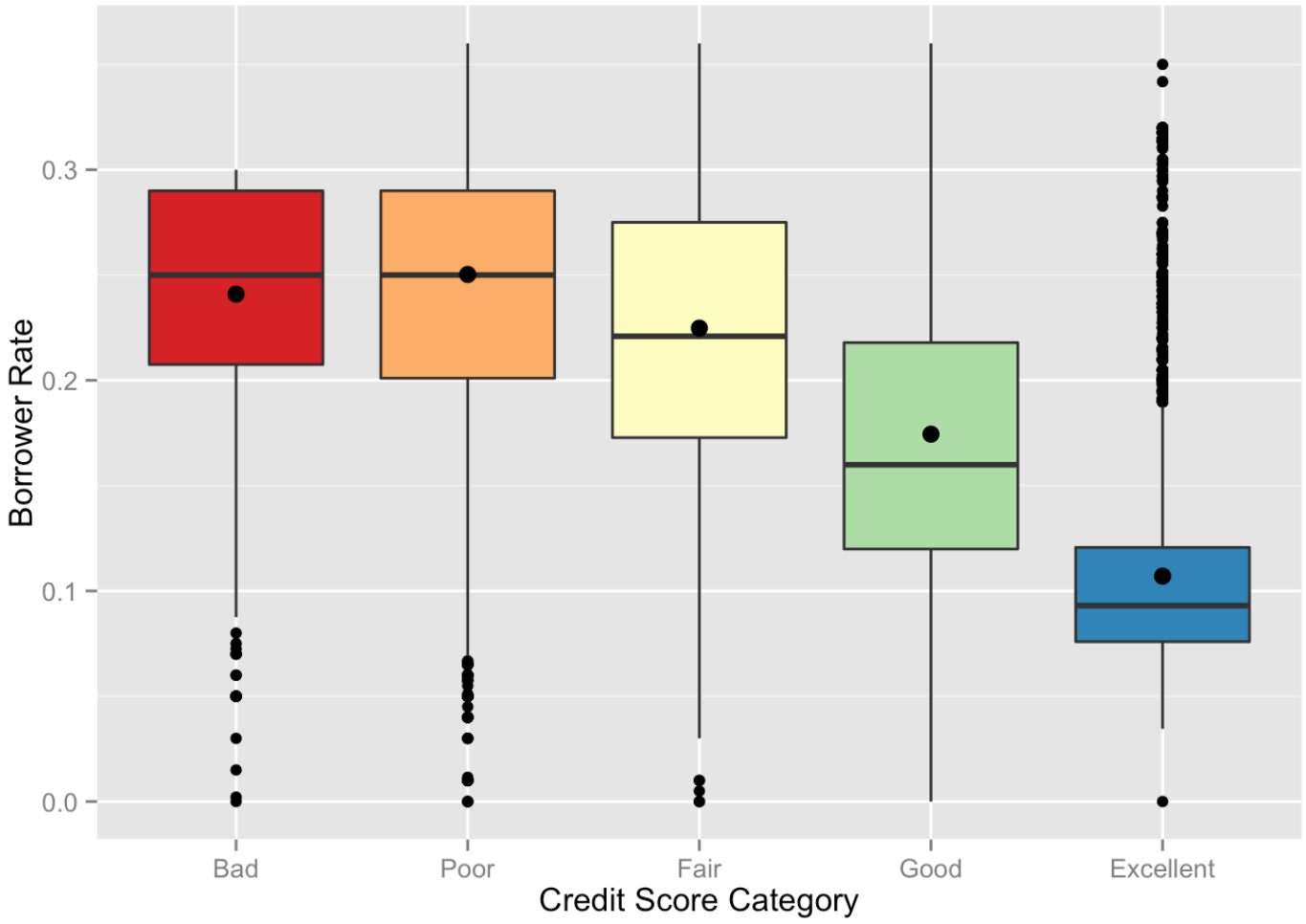
Prosper rating scale is a discrete range from 1 to 7 so you get overplotting on the vertical bands for each rating. There is a negative linear relationship with higher amount of outliers for lower ratings. This plot is not very effective due to discrete nature of the rating scale.



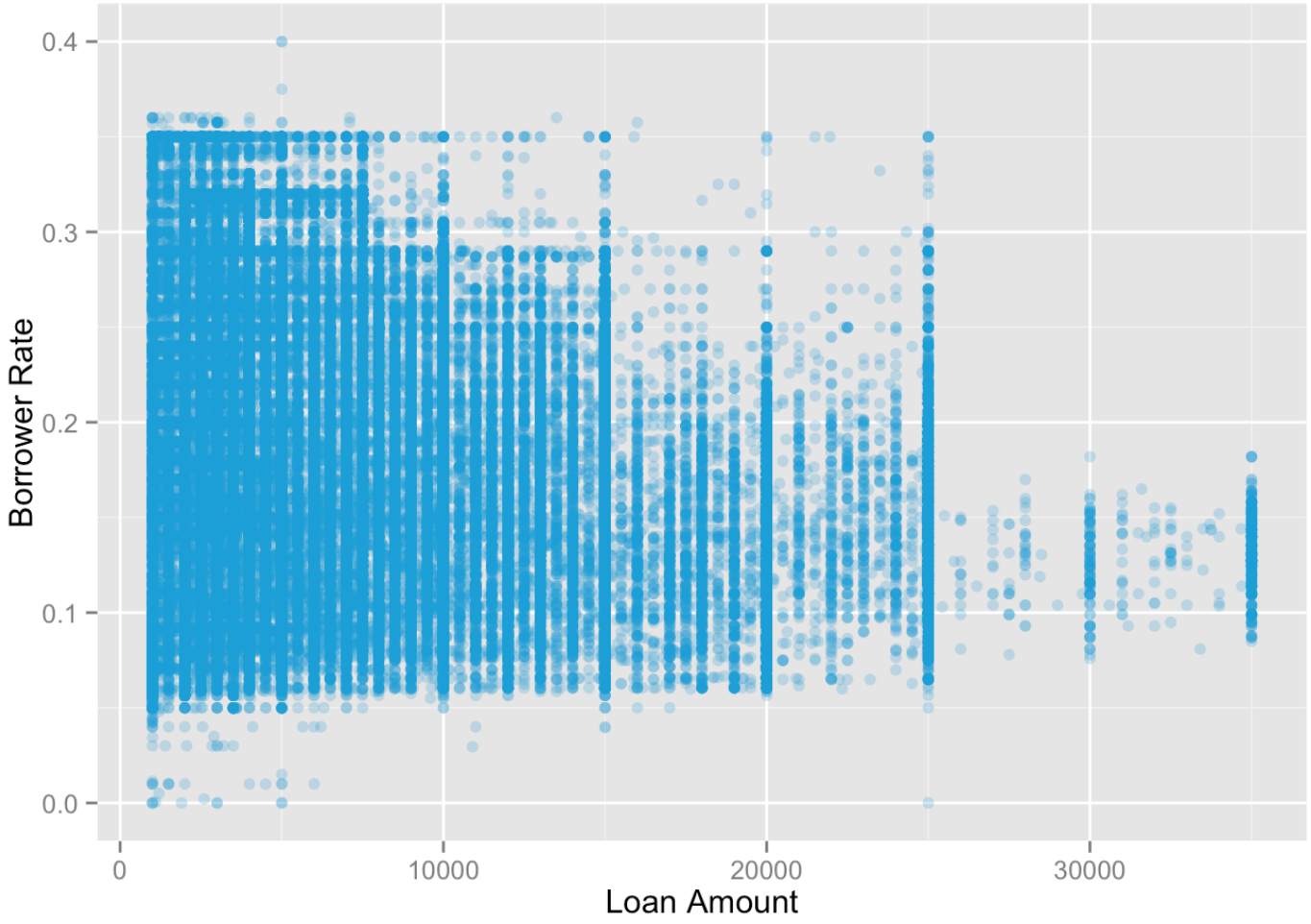
The boxplot is a better plot to visualize the data for rating and rate due to the discrete rating scale.



The credit score range lower variable is represented in increments of 20 so you get overplotting on the vertical bands for each score. Although there is a negative non-linear relationship between credit score and rate, the data really shows that credit score most likely is not the sole factor in the customers rate.



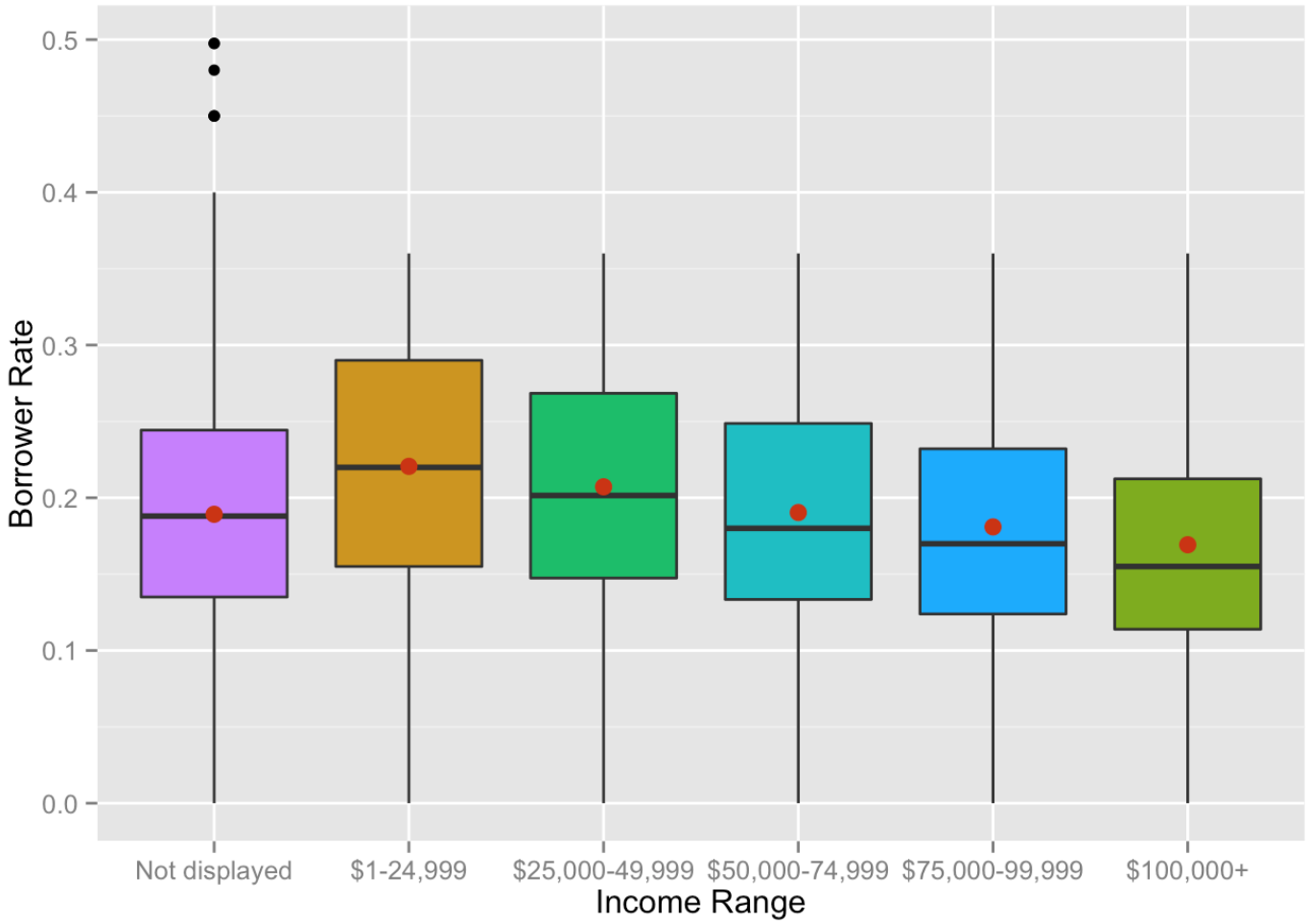
Using the custom credit score category field it buckets the credit scores to get a more natural visual instead of the vertical bands in the scatterplot. The boxplot wiskers also show the wide ranging rates for each credit score category.



There is a non-linear relationship between loan amount and rate. As loan amount increases, the density and range of rates at the top end decreases.

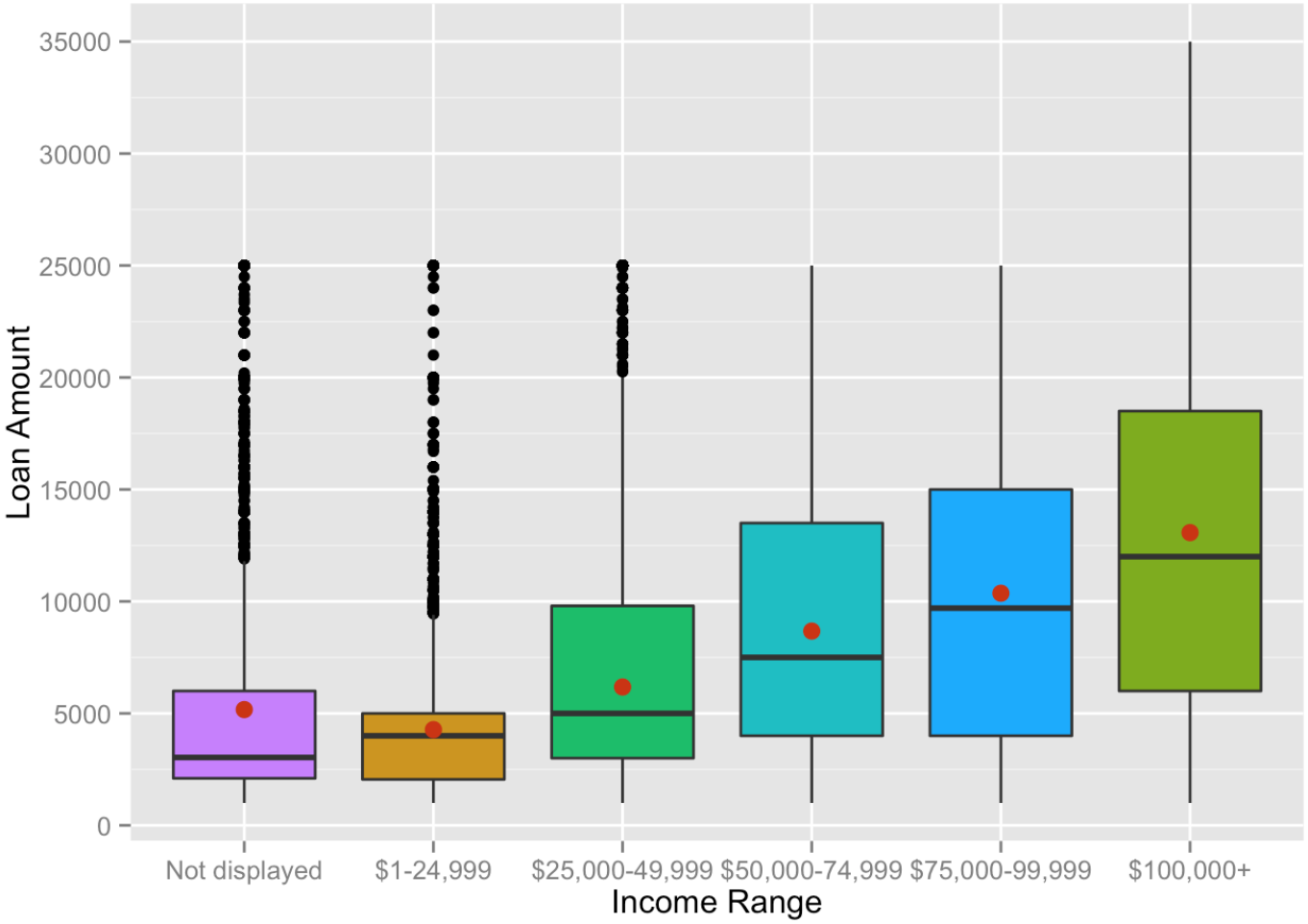
Summary data for Income Range:

```
## Source: local data frame [8 x 5]
##
##       IncomeRange LoanAmtMean LoanAmtMedian LoanAmtVolume LoanCnt
## 1   Not displayed    5169.649          3033      40018253    7741
## 2       $1-24,999    4273.974          4000      31088885    7274
## 3    Not employed    4884.829          4000       3937172     806
## 4              $0    7410.931          5000       4602188     621
## 5 $25,000-49,999    6177.987          5000     198881762   32192
## 6 $50,000-74,999    8675.276          7500     269367313   31050
## 7 $75,000-99,999   10365.924          9700     175349966   16916
## 8       $100,000+   13073.127         12000     226648808   17337
```



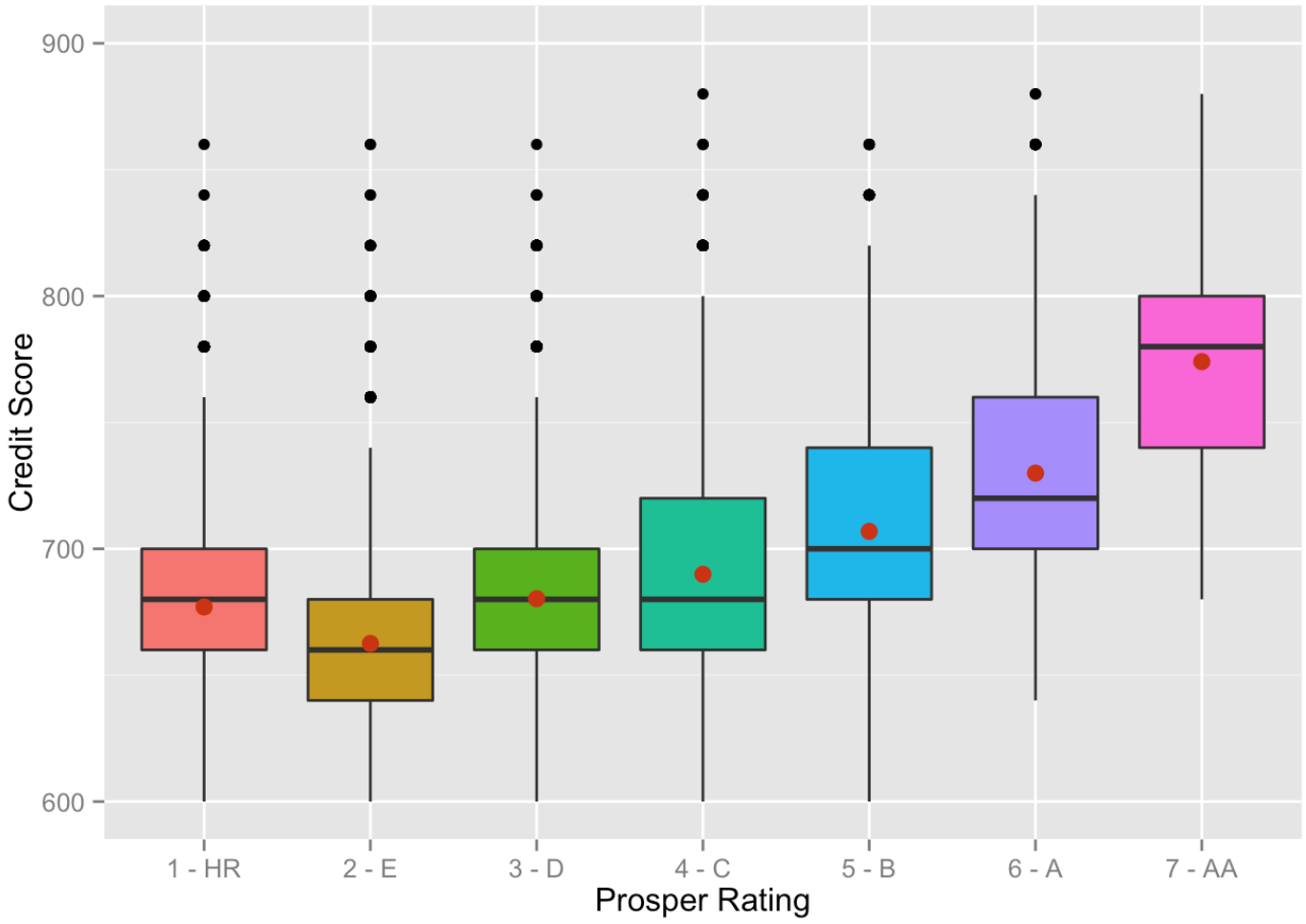
Median rate decreases slightly as income range increases. Income level is unknown for Not displayed so unable to determine where it would truly fall in the range. There does not appear to be a significant relationship between income level and rate.

I'll take a look at the relationship of income and loan amount now.



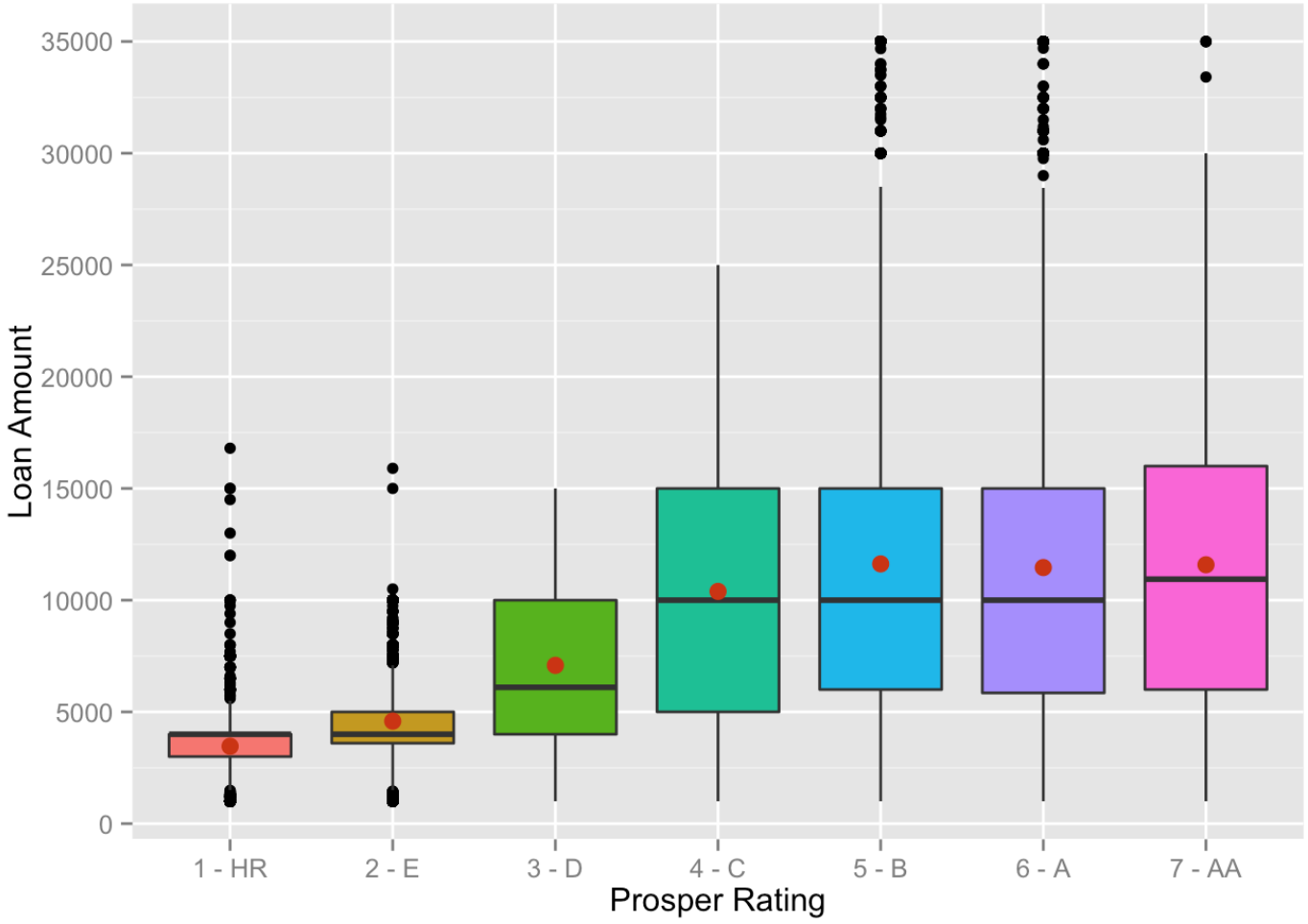
Median loan amount increases as income range increases. Excluded outliers $0 and Not employed from box plots. Stated monthly income Median of 4667 and Mean of 5608.

Next I will explore prosper rating and credit score and loan amount.



Median credit score is higher for the lowest 1 - HR credit rating than 2 - E. This suggests some other credit factor could be the determining factor between these ratings. Median credit score does increase as credit rating increases from credit rating 2 - E to 7 - AA.

```
## Source: local data frame [7 x 6]
##
##   ProsperRating..numeric. LoanAmtMean LoanAmtMedian LoanAmtMax
## 1                       1    3463.114          4000      16800
## 2                       2    4586.405          4000      15900
## 3                       3    7083.439          6100      15000
## 4                       4   10391.940         10000      25000
## 5                       5   11622.355         10000      35000
## 6                       6   11459.886         10000      35000
## 7                       7   11583.539         10940      35000
## Variables not shown: LoanAmtVolume (dbl), LoanCnt (int)
```



Similar to income range, median loan amount increases for higher credit rating. The boxplot shows the maximum loan amounts by credit rating. For credit ratings >= 5 max loan amount is 35,0000, Credit rating = 4 max loan amount is 25,000 and <= 3 betwee 15,000 - 17,000. Loan amount caps can minimize credit loss exposure in the event of default. I'm really interested in exploring the benefits of loan amount limits based on the borrowers credit rating.

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?

Rate has a strong positive linear relationship with estimated loss and estimated return with a correlation coefficient of .95 and .82 respectively. It has a strong negative linear relationship with prosper rating at -.95. The relationship with the custom risk prosper score was not as strong as expected at -.65 and confirmed some suspicion about the distribution of this field in the univariate analysis section. It was a little surprising the relationship with credit score was not as strong at -.51.

Loans with the highest stated monthly income have the highest median loan amount and lowest median rate. This is consistent with expectations that people that make more money can afford larger loan payments. Although the disparity in rates is not large it also shows the more money you make the better rate you will get. Is this a product of making more money or better credit ratings?

Loans with the highest credit rating have the highest median loan amount and lowest median rate. What would be the driver for lower loan amounts for lower credit ratings? The disparity in rates is much larger for credit rating.

Loan amount has the strongest relationship with prosper rating at .43 and is the only field that had any significant relationship with the Term of the loan at .34.

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

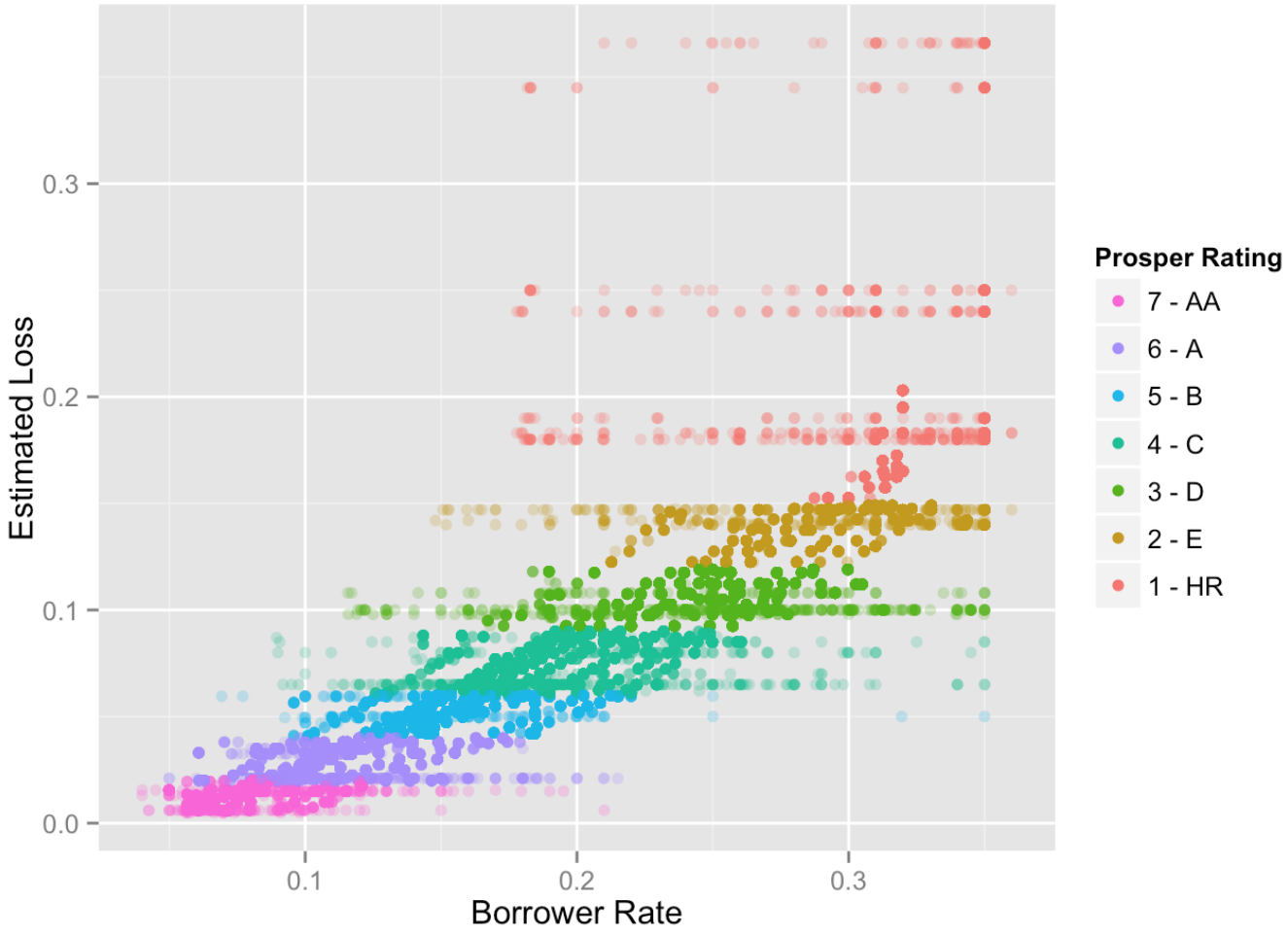
In the other features, I really was expecting stronger relationships with the selected credit reporting fields and credit score. Card utilization was the strongest at -.44. delinquency last 7 and public record last 10 were the next at -.22 for both. This really moved my focus to the prosper rating field since it inherently represented the credit worthiness of the borrower.

What was the strongest relationship you found?

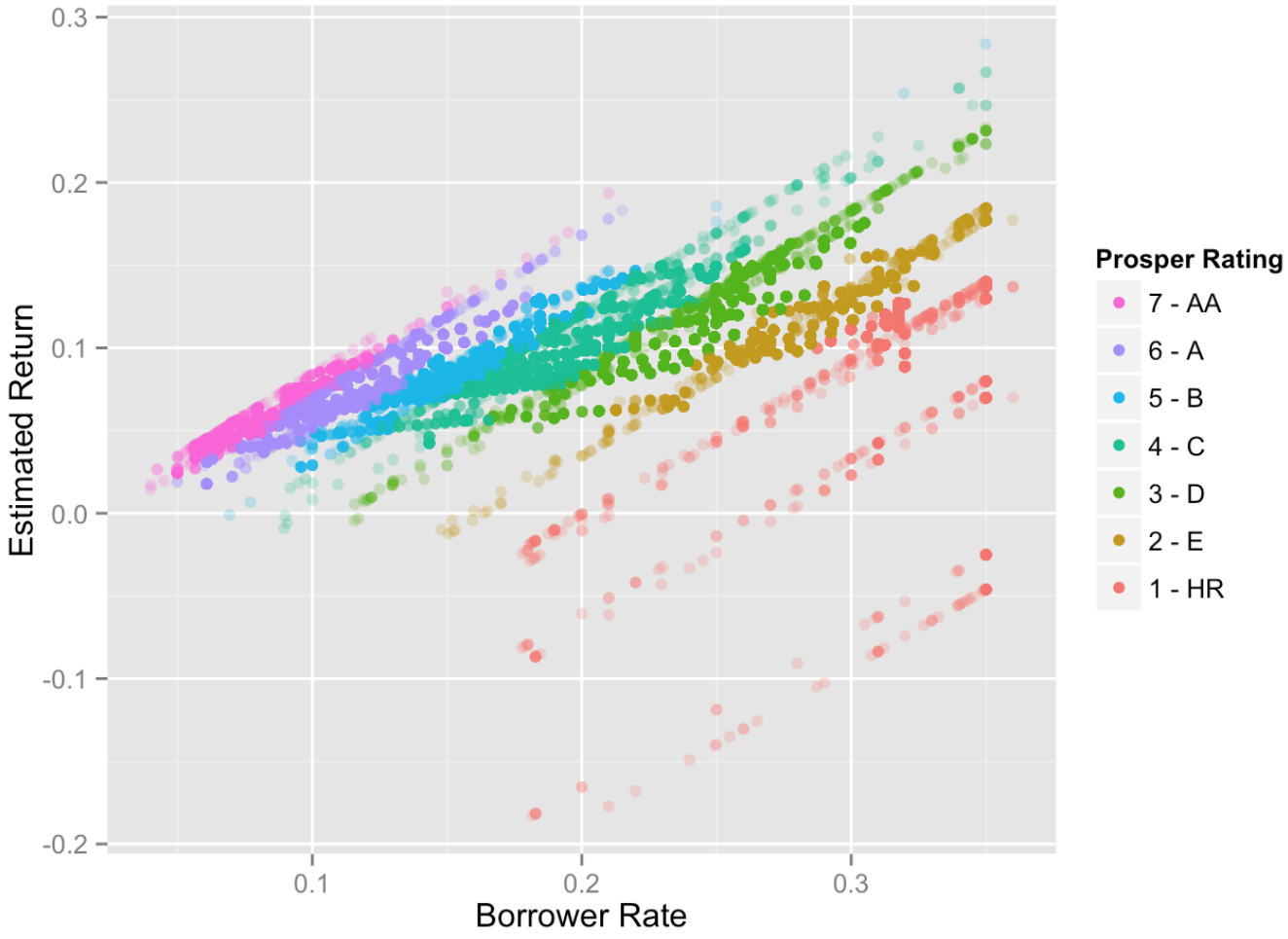
The strongest relationship was prosper rating and rate and estimated loss at -.95 and -.96 respectively. Due to limitations in the dataset, I could not calculate actual loss rates so I focused on defaulted loans that had any amount of principal amount charged off.

Multivariate Plots

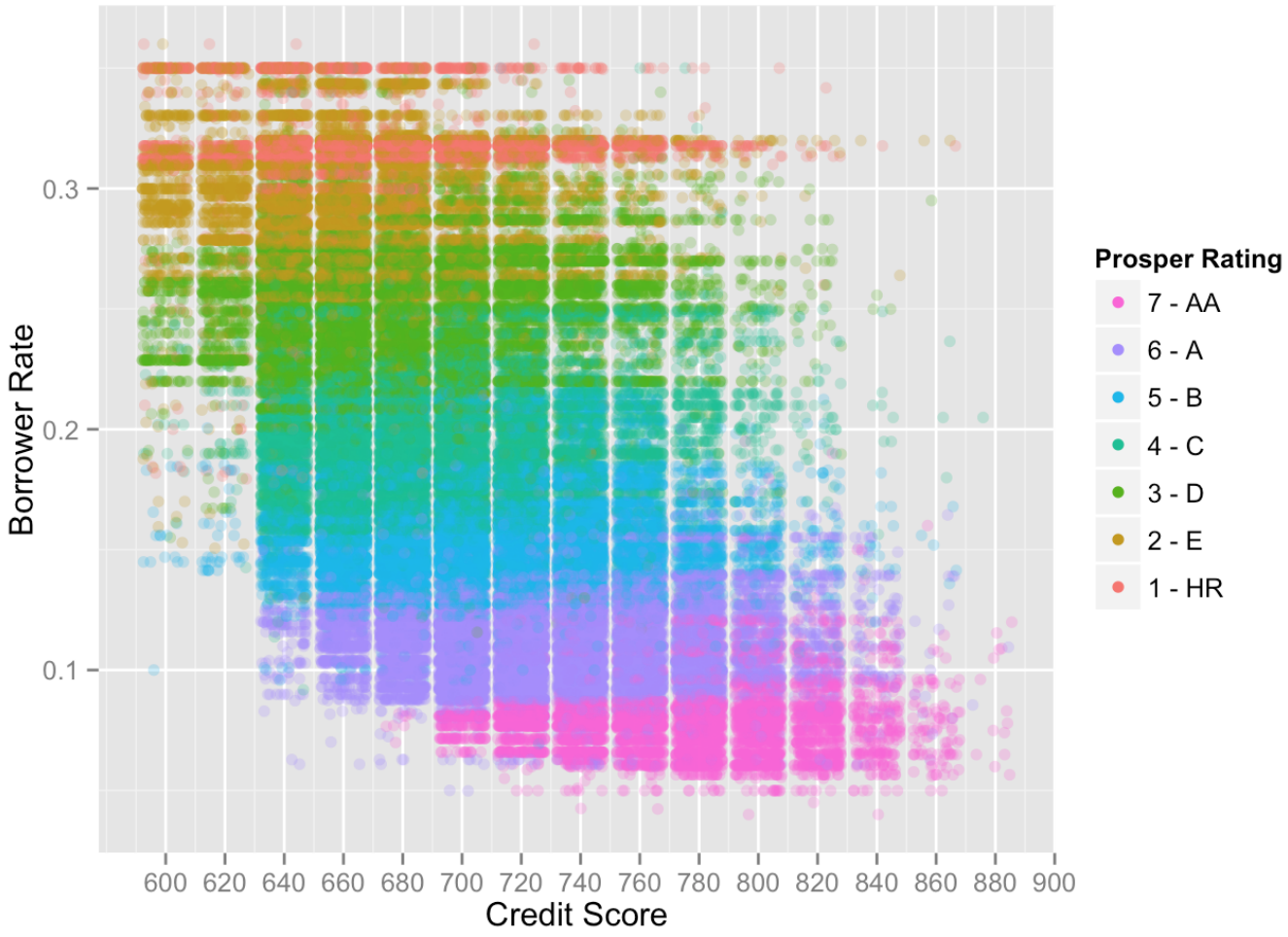
I will now revisit the scaterplots in the Bivariate Plots section and add color by prosper rating.



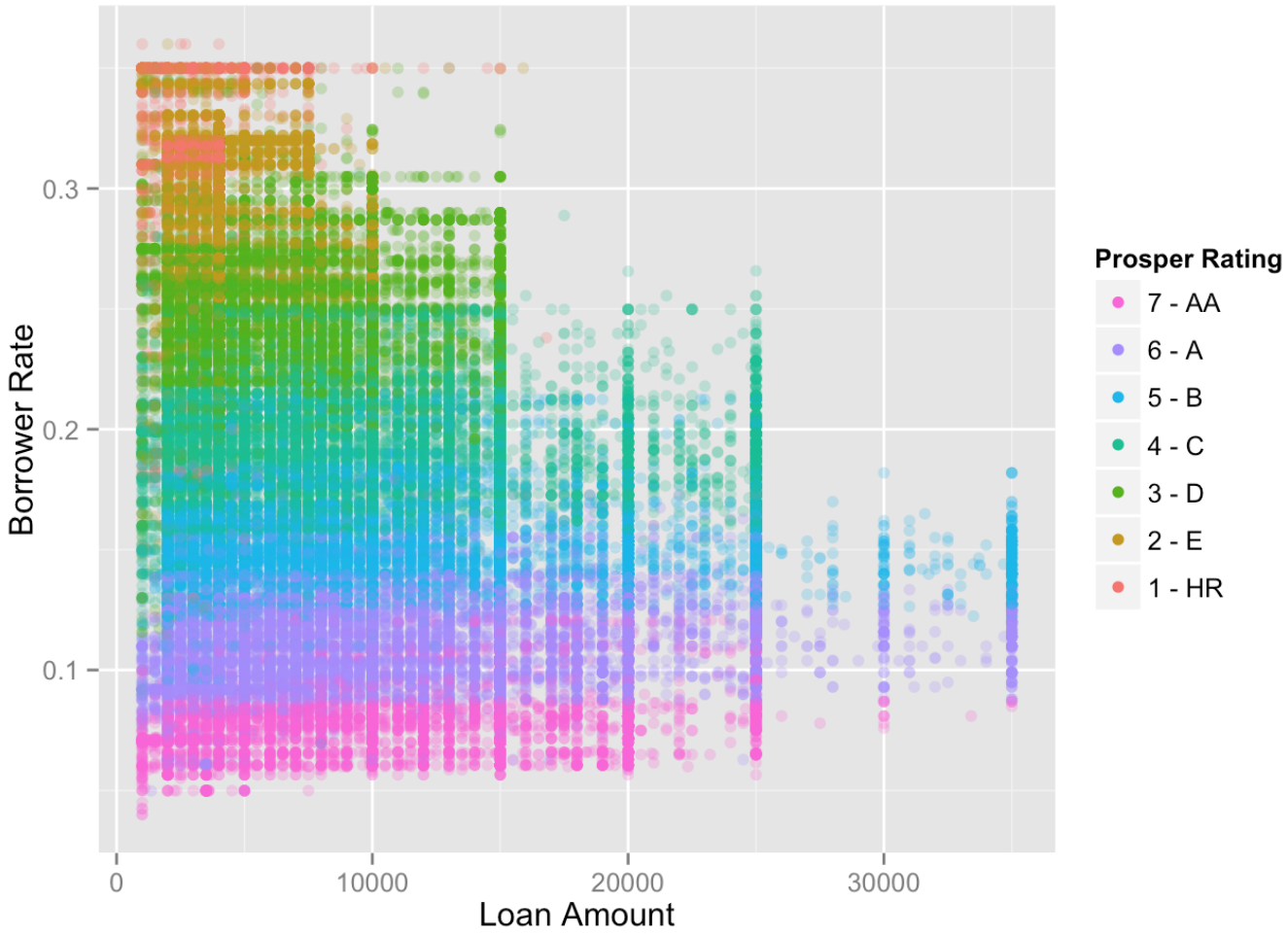
The vertical bands are clearly associated with higher loss estimates for the worst HR credit rating.



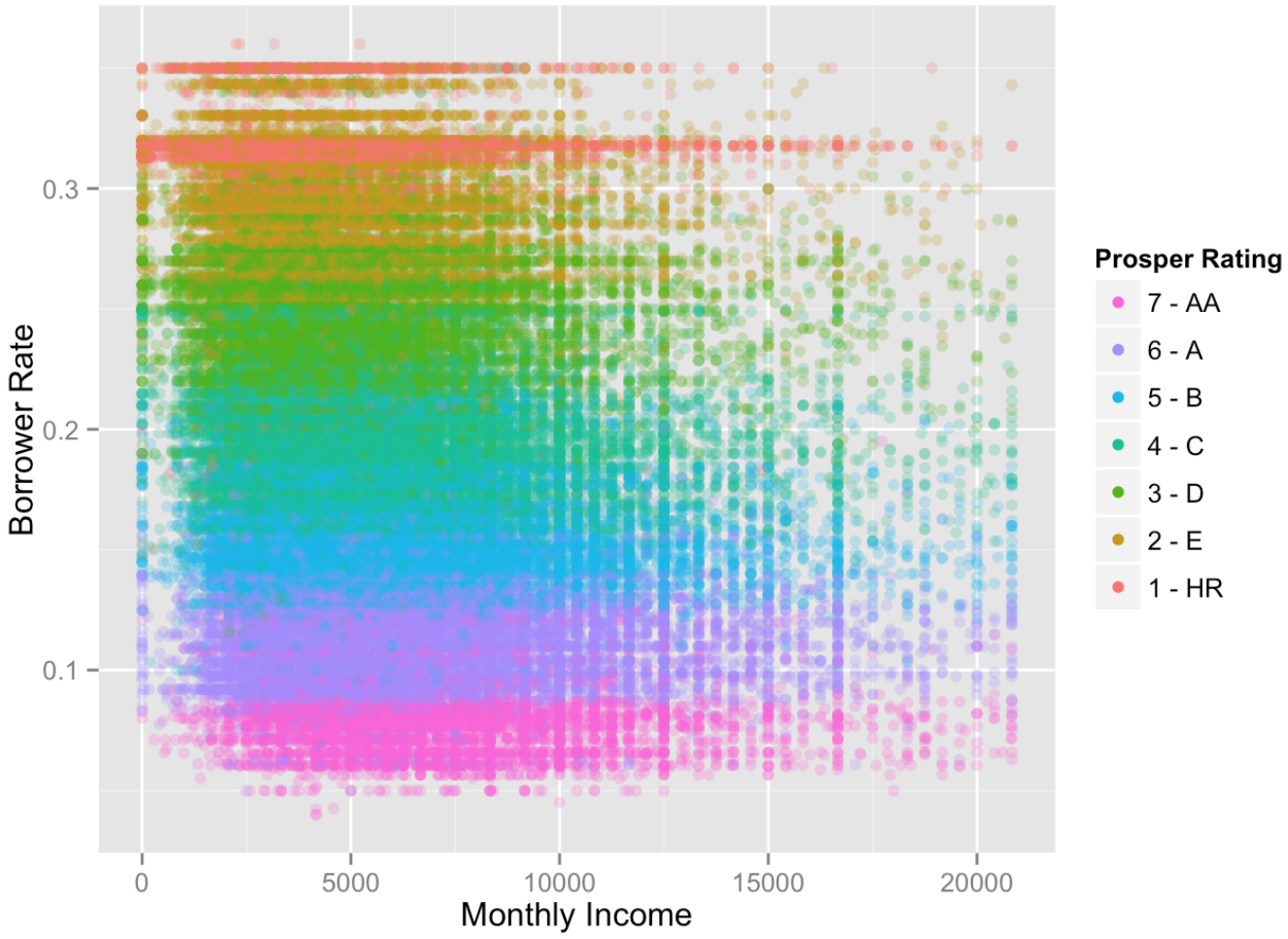
Inverting the plot now for return estimates shows the same relationship. The really interesting part about this plot is some HR credit rating loans have negative estimated returns.



Adding prosper rating to the credit score and rate scatterplot really highlights the relationship of the credit rating system and interest rates across ranges of credit scores.

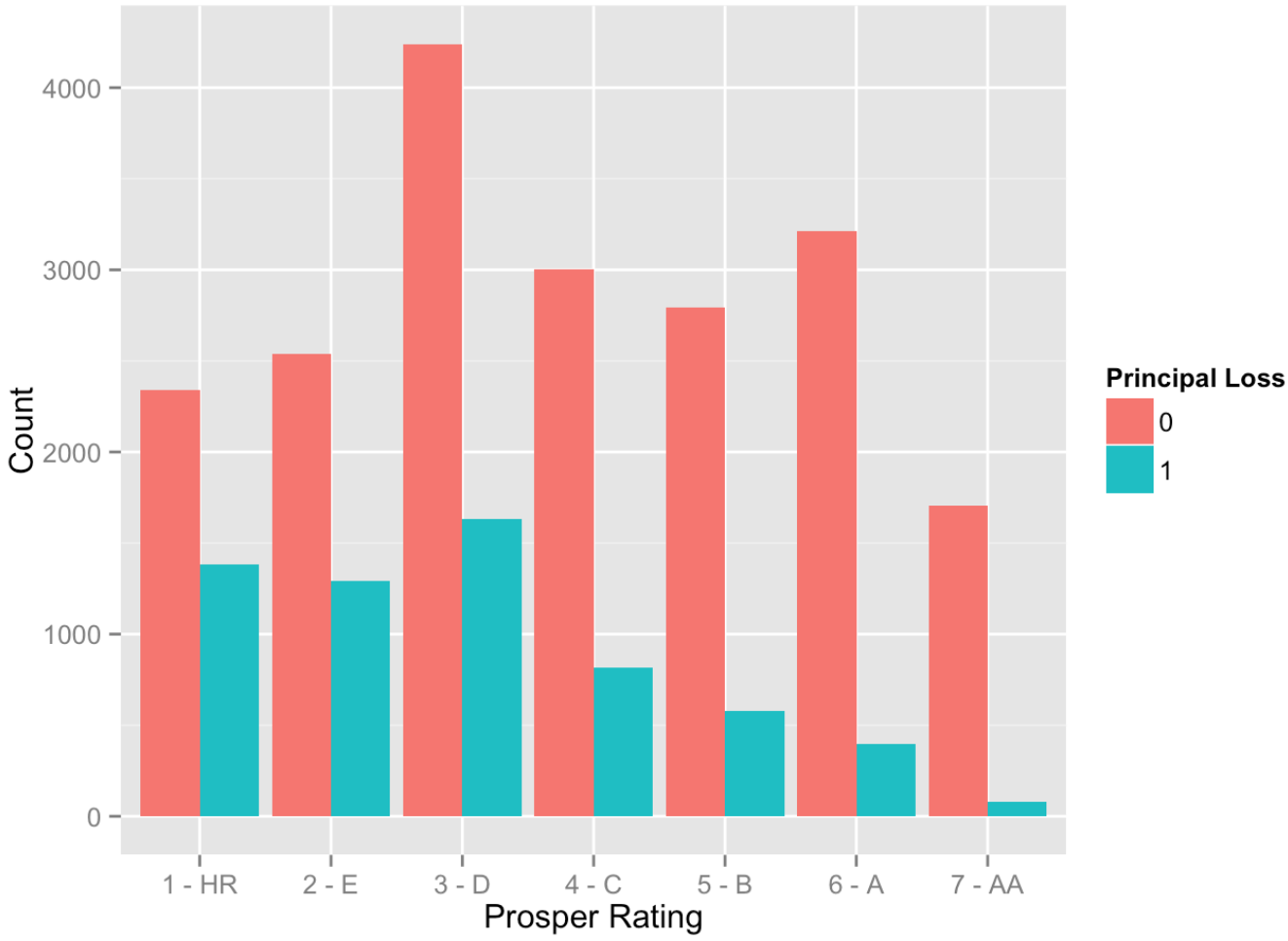


This plot is another visual that highlights at lower credit ratings loan amounts have maximum limits. The plot shows only ratings >=5 have loan amounts up to the maximum of 35,000. I'm interested in exploring this further in a predictive model to identify how loan amount limits for each credit rating can minimize the probability of loss.

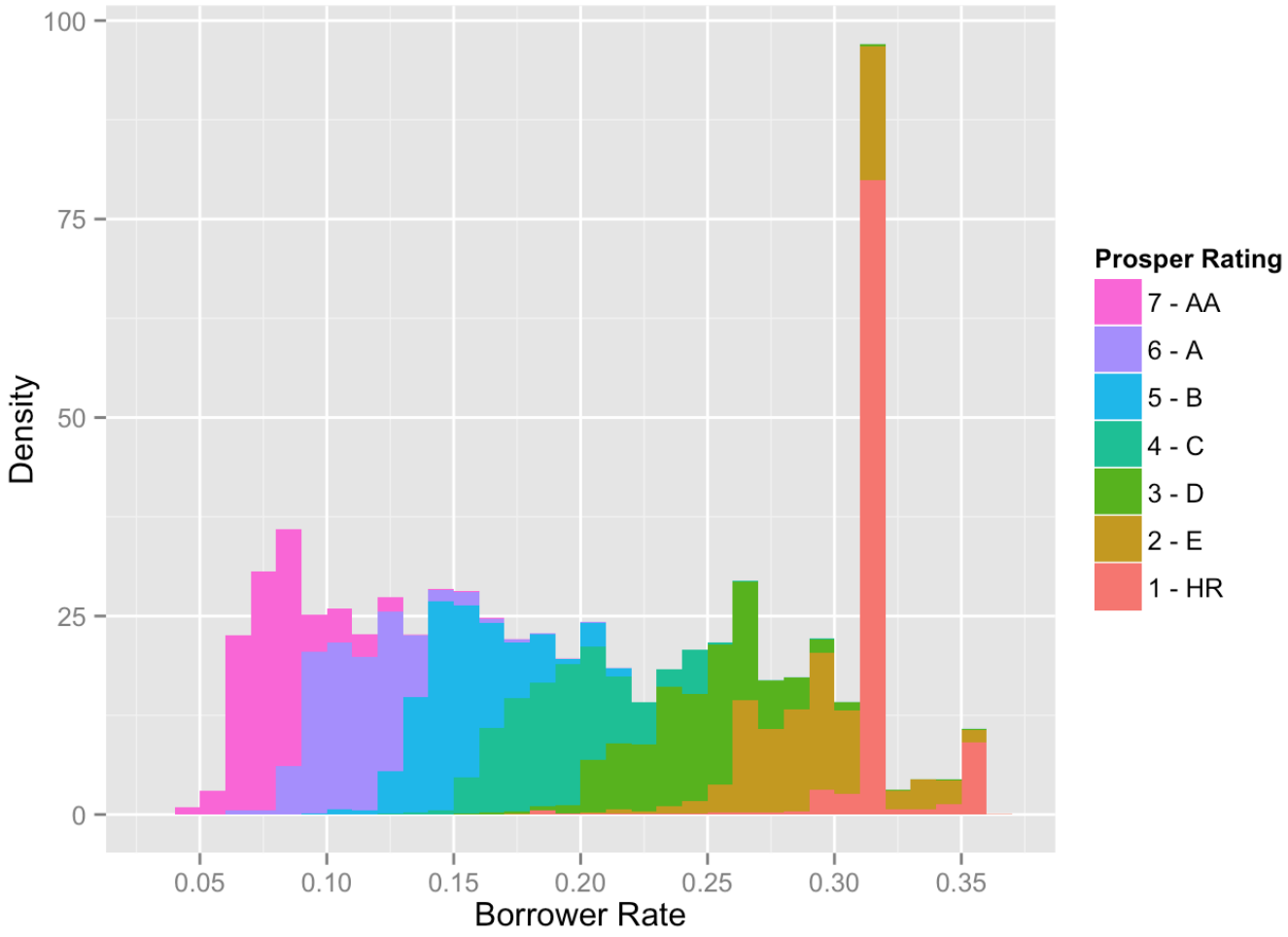


Rate and monthly income have a weak relationship but adding the credit rating tells the story about rates across similar income levels.

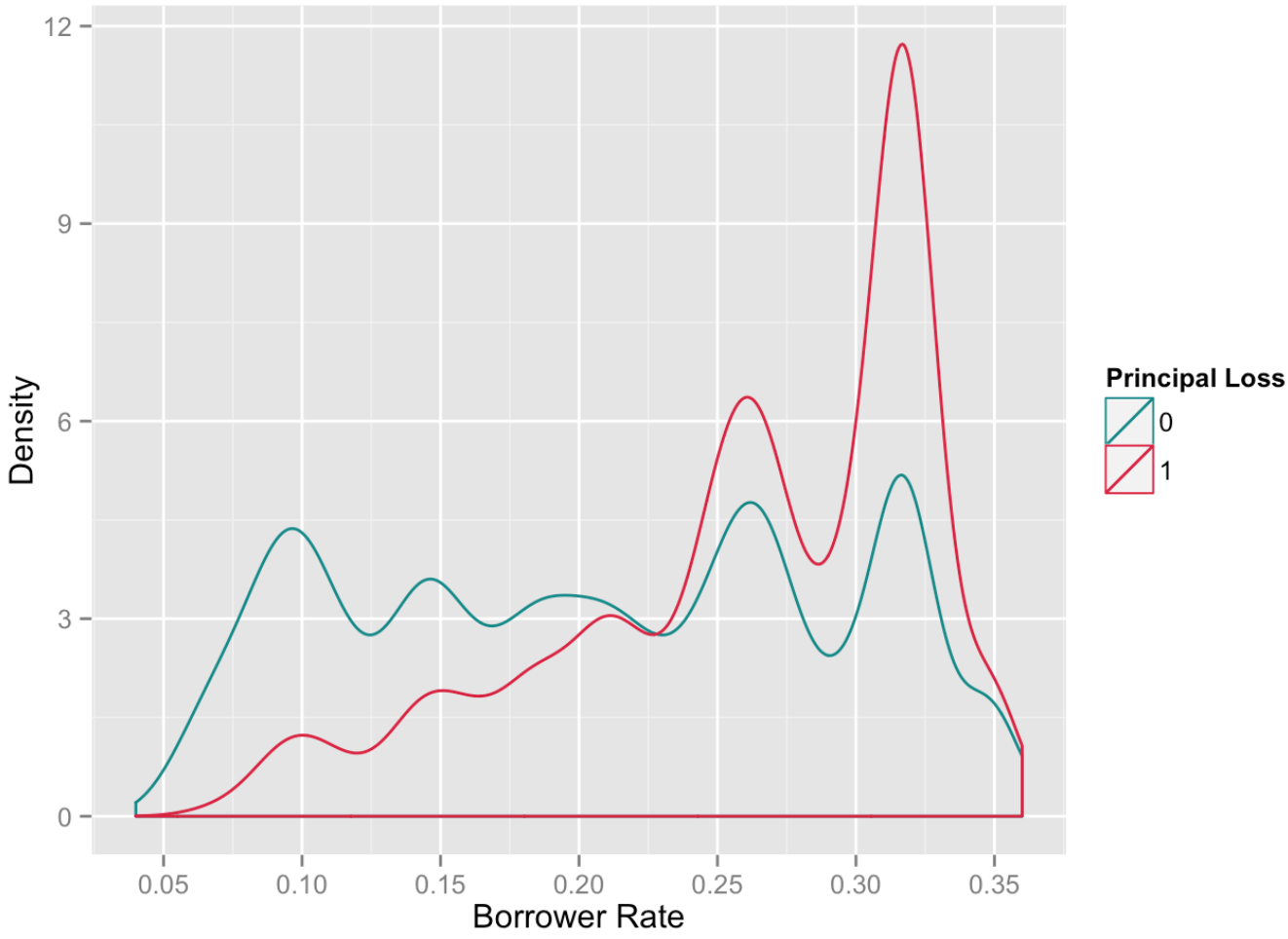
I will now explore relationships with the calculated variable principal loss. I have narrowed down my variables of interest to borrower rate, prosper rating and loan amount.



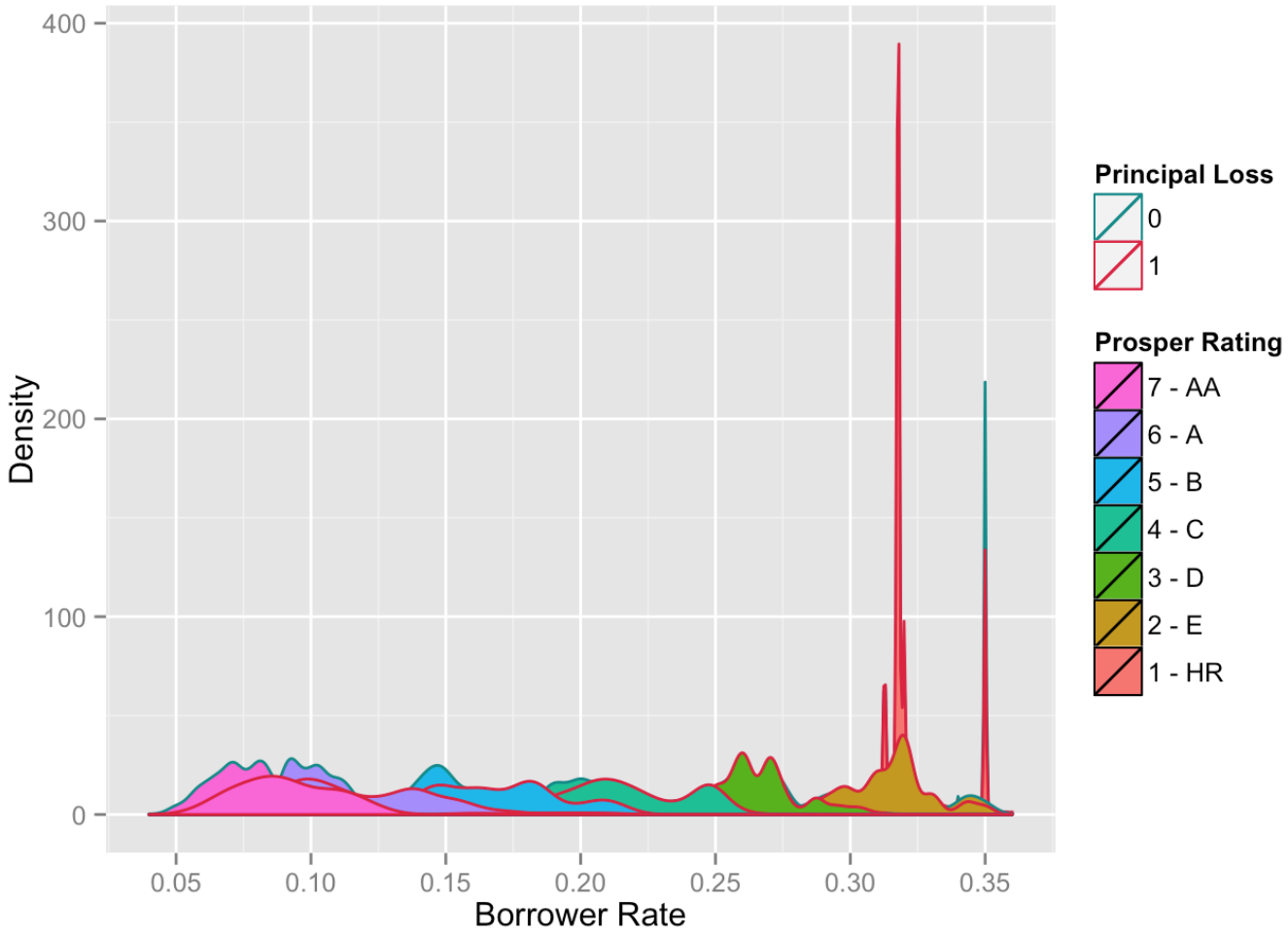
Prosper rating histogram by loans with a principal loss. Filtered for all loans originated >= 2009 and closed. Including loans that are still open would skew the analysis.



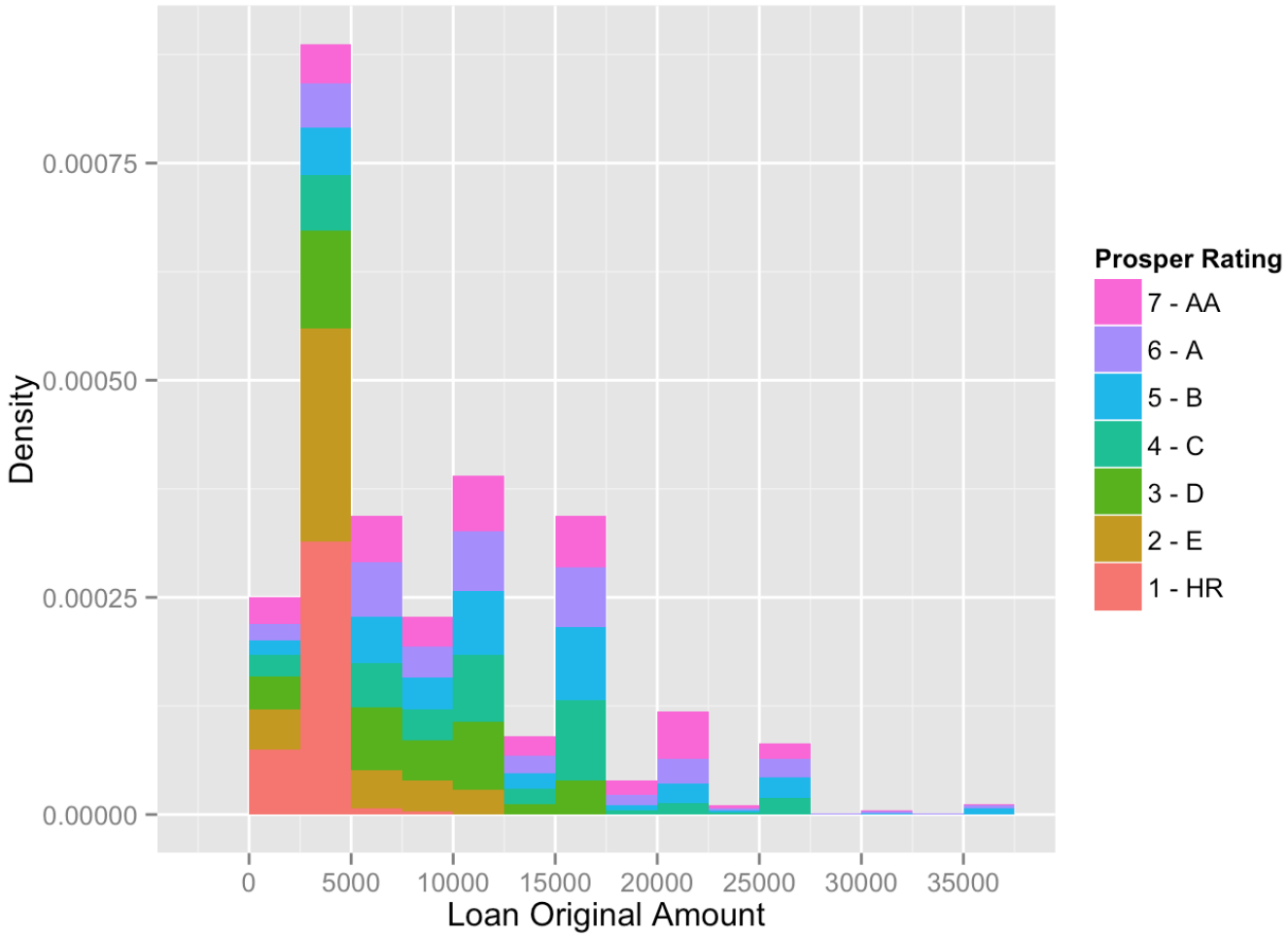
The rate spike at .31 is predominantly HR credit rating loans.



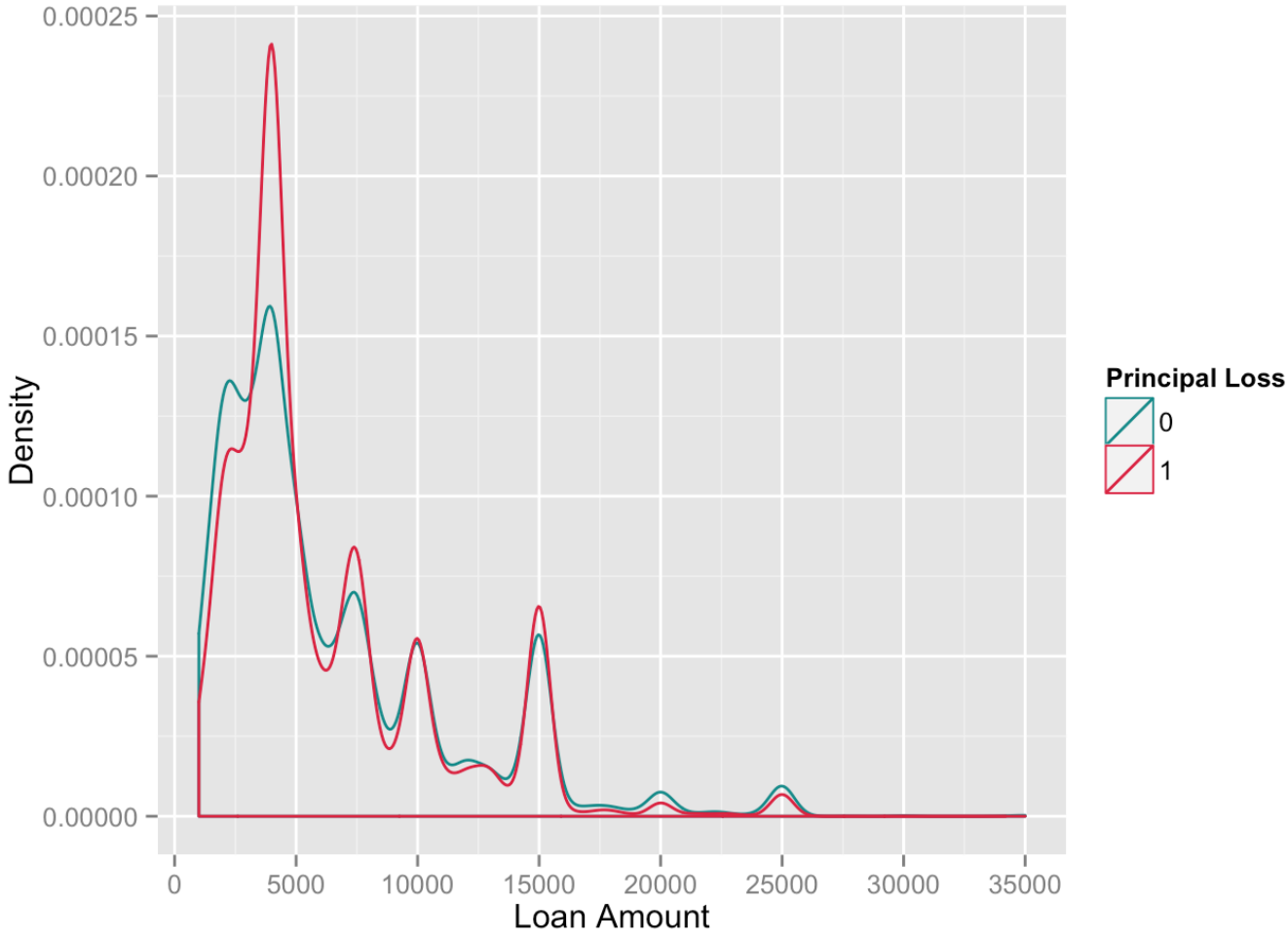
The density of loans with a principal loss is much higher with rates >= .225.



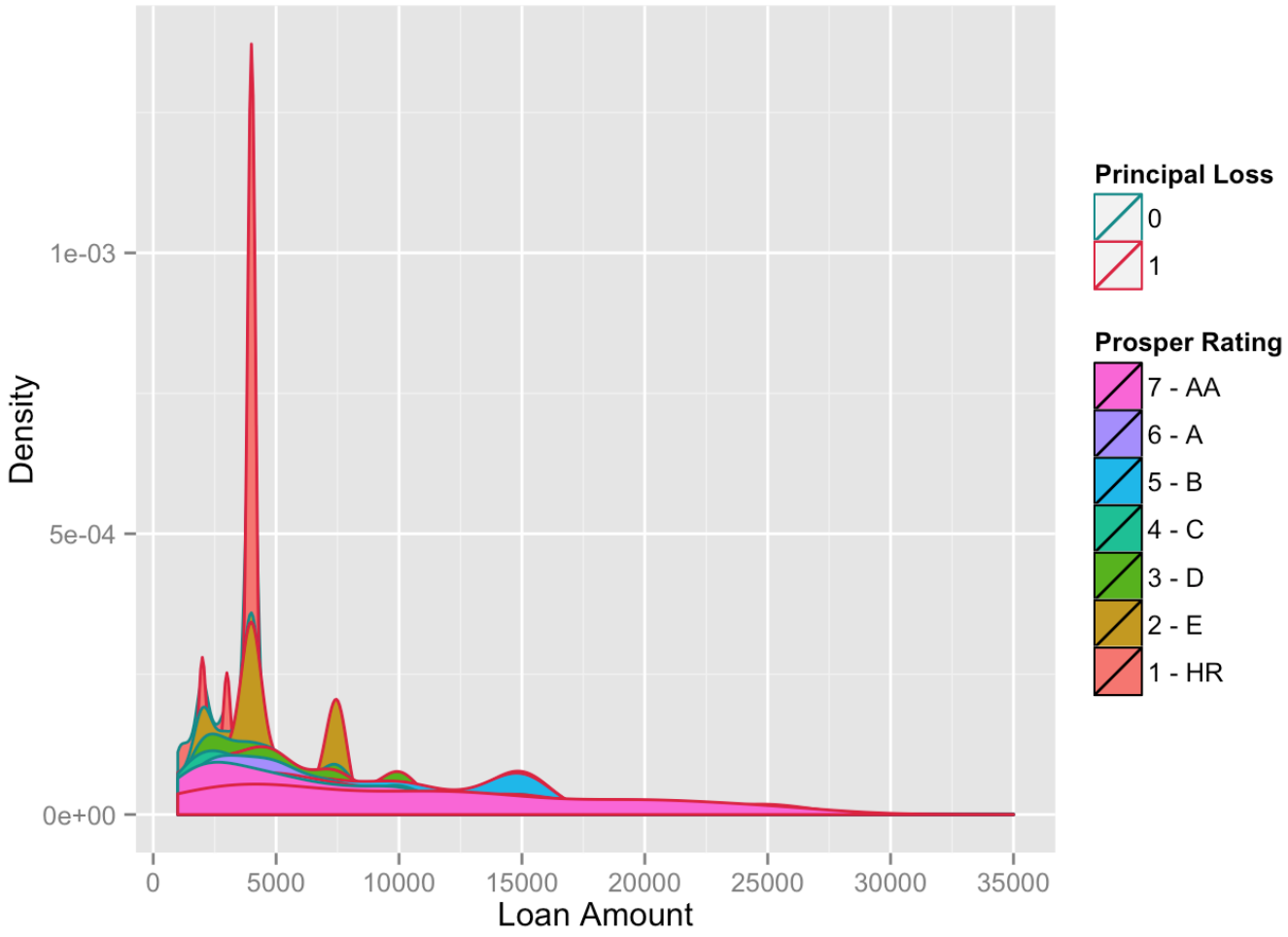
Added fill for prosper rating to the density plot for borrower rate and principal loss. This plot is a great visual to show the relationship ofrating and rate. The density curves are distinct along the axis for each rating. It is interesting to see the spike for no principal loss loans at .35 but a significant spike for 1 - HR credit rating at .31.



Loans with lower credit ratings 1 - 3 have a higher distribution in the loan amount <= 5000 bins.



The density of loans with a principal loss is much higher with loan amounts between 2500 - 5000 and then follows a very similar curve as loan amount approaches 35000.

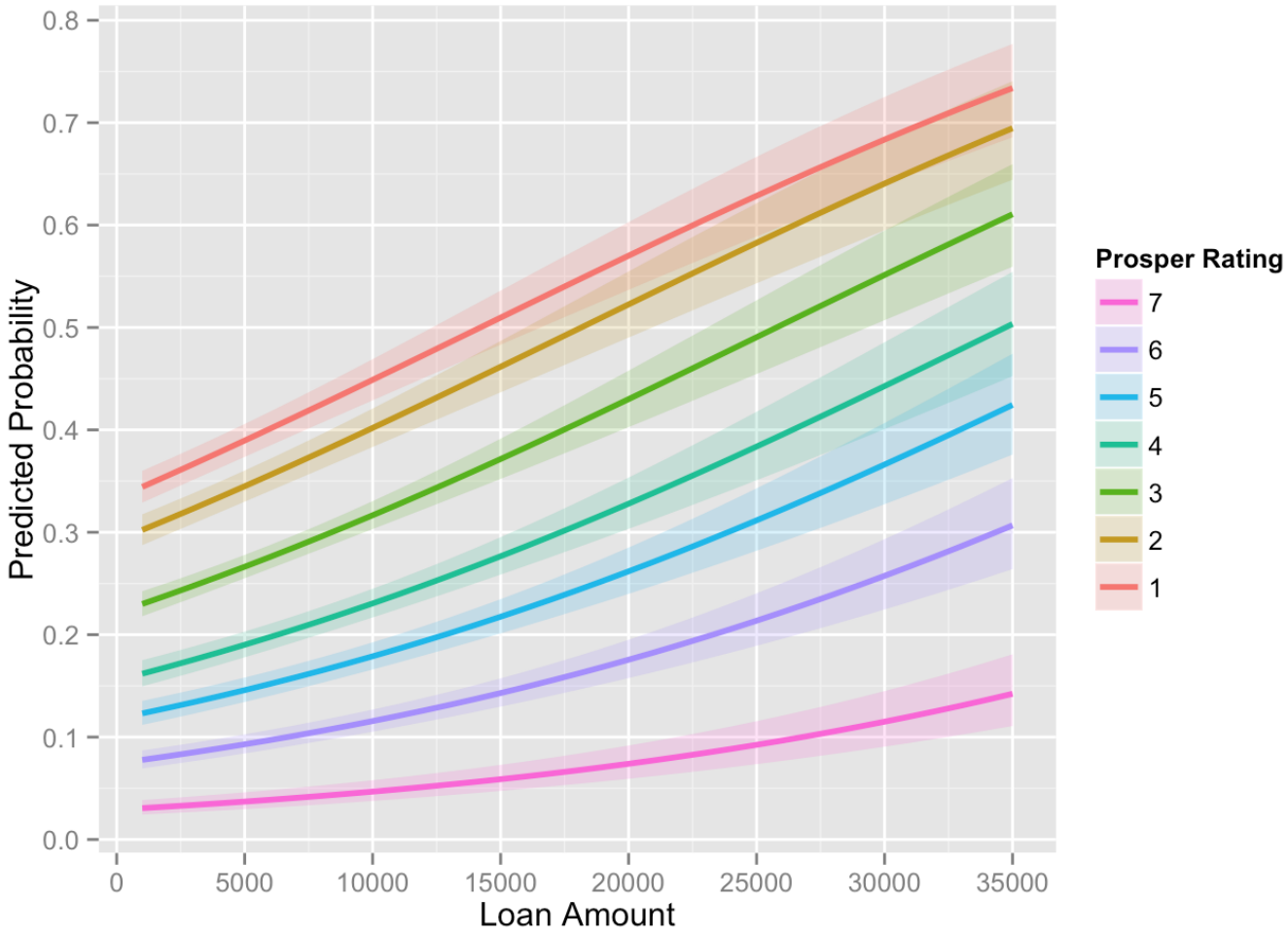


Added fill for prosper rating to the density plot for loan amount and principal loss. The spike around 5K for 1 - HR credit rating loans with a principal loss really stands out in this plot.

```
##
## Call:
## glm(formula = PrincipalLoss ~ ProsperRating..numeric. + LoanOriginalAmount,
##     family = "binomial", data = myData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2315  -0.8205  -0.6077  -0.2586   2.6404
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)               -6.929e-01  3.611e-02 -19.191  < 2e-16 ***
## ProsperRating..numeric.2 -1.923e-01  4.830e-02  -3.981 6.87e-05 ***
## ProsperRating..numeric.3 -5.642e-01  4.601e-02 -12.263  < 2e-16 ***
## ProsperRating..numeric.4 -1.000e+00  5.522e-02 -18.110  < 2e-16 ***
## ProsperRating..numeric.5 -1.319e+00  6.112e-02 -21.575  < 2e-16 ***
## ProsperRating..numeric.6 -1.829e+00  6.709e-02 -27.266  < 2e-16 ***
## ProsperRating..numeric.7 -2.811e+00  1.221e-01 -23.027  < 2e-16 ***
## LoanOriginalAmount        4.877e-05  3.615e-06  13.490  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28508  on 26004  degrees of freedom
## Residual deviance: 26768  on 25997  degrees of freedom
##   (144 observations deleted due to missingness)
## AIC: 26784
##
## Number of Fisher Scoring iterations: 5
```

```
##                                OR      2.5 %     97.5 %
## (Intercept)               0.50011510 0.46586568 0.53670266
## ProsperRating..numeric.2 0.82508646 0.75053635 0.90698774
## ProsperRating..numeric.3 0.56883903 0.51978321 0.62251279
## ProsperRating..numeric.4 0.36785391 0.33001927 0.40978487
## ProsperRating..numeric.5 0.26747524 0.23715001 0.30136256
## ProsperRating..numeric.6 0.16052769 0.14060745 0.18291358
## ProsperRating..numeric.7 0.06017423 0.04703614 0.07594495
## LoanOriginalAmount       1.00004877 1.00004167 1.00005584
```

```
##   LoanOriginalAmount ProsperRating..numeric.       Pred
## 1               4500                       1 0.38379544
## 2               4500                       2 0.33945219
## 3               4500                       3 0.26160811
## 4               4500                       4 0.18640534
## 5               4500                       5 0.14280350
## 6               4500                       6 0.09089479
## 7               4500                       7 0.03612486
```



The line plot with the confidence interval ribbon shows how the predicted probability for loss increases as loan amount increases and credit rating decreases. In addition the confidence interval gets wider.

See the multivariate and final plots section for additional comments on this model.

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?

I first expanded the scatterplots from the bivariate section to include prosper rating. estimated loss and estimated return had a clear correlation with credit rating. As credit rating decreases, estimated loss increases and return decreases due to their direct relationship (return is the difference between effective

yield and loss). In the rate and loan amount by prosper rating scatterplot, all credit rating levels have loan amounts between 0 and 10000 but the lower the credit rating the higher the rate. Only the top 4 credit ratings have loan amounts >= 25000.

For the calculated column principal loss, the strongest relationship was with rate, estimated loss, estimated return, prosper rating and loan amount. The density plot for rate and principal loss shows a higher density for rates >= .225. I added a fill by credit rating and this really popped with the curves higher for no principal loss and credit rating 5 - 7 and spikes significantly at a rate of .31 for credit rating 1.

Were there any interesting or surprising interactions between features?

Credit score and credit rating relationship was interesting in that I expected a more distinct line between rating and the credit score ranges. Although median credit scores are higher for higher credit ratings, the scatterplot showed a lot of overlap in scores across ratings.

Income amount was really all over the place with no clear relationship but there is the appearance of a higher representation of higher income amounts for higher credit ratings.

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

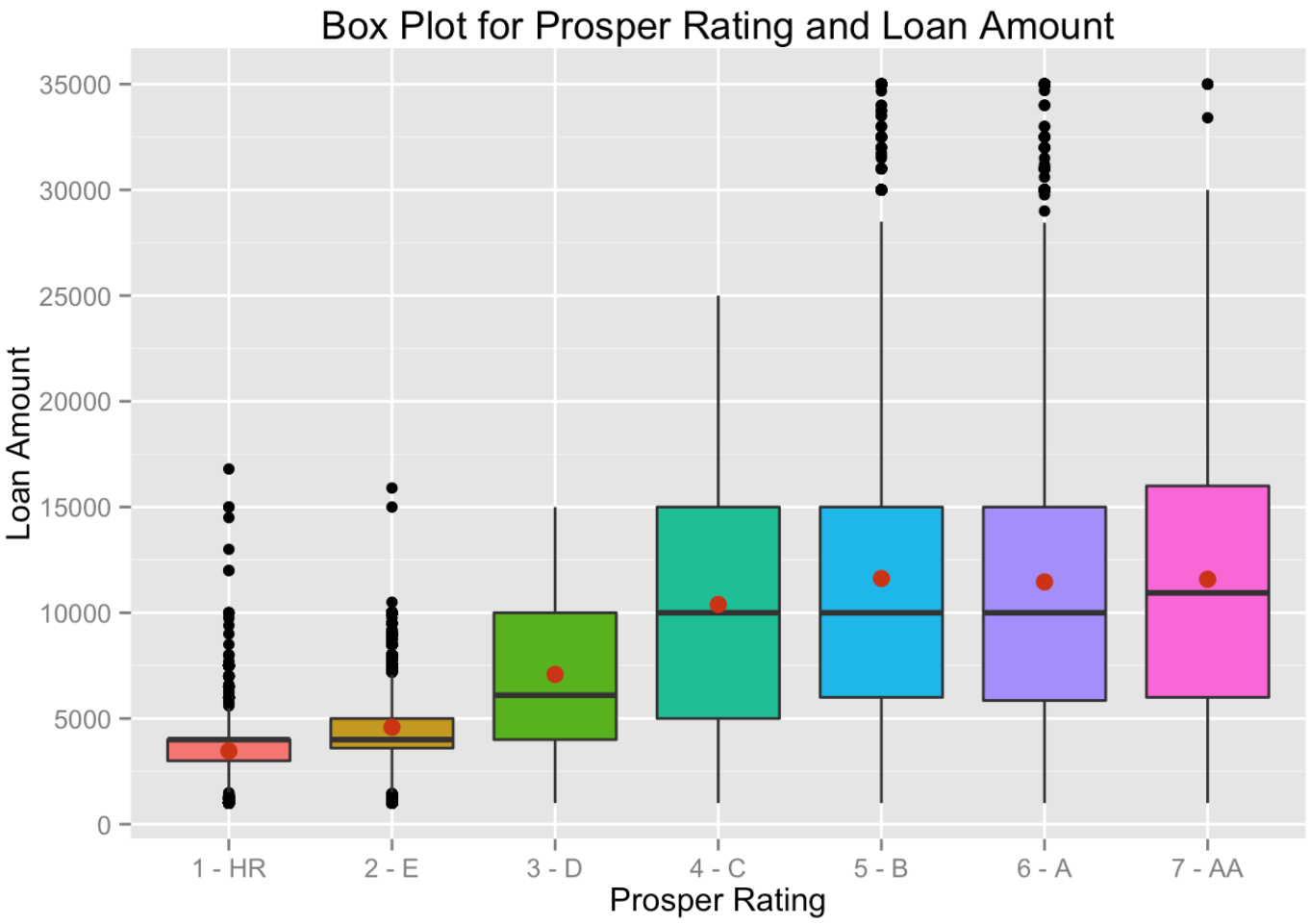
I created a logistic regression model for principal loss as the dependent variable and independent variables prosper Rating and loan amount. I did not include borrower rate as well since it has such strong negative linear relationship with prosper rating.

I initially used a random sample of 500 loans to get a better sense of the true p values for the variables. The first dataset that I ran through the model was for the median loan amount of 4500 for all 7 credit ratings. The probability of a loan defaulting and any amount of principal being charged off is 38.38% and 3.61% for credit ratings 1 - HR and 7 - AA respectively.

The second dataset was simulated for loan amounts from 1000 to 35000 to build a 95% confidence interval plot for the predicted probabilities.

Final Plots and Summary

Plot 1



Description 1

This boxplot provides the best visual for median loan amounts for each credit rating and the whiskers of the boxplots show the maximum loan amounts.

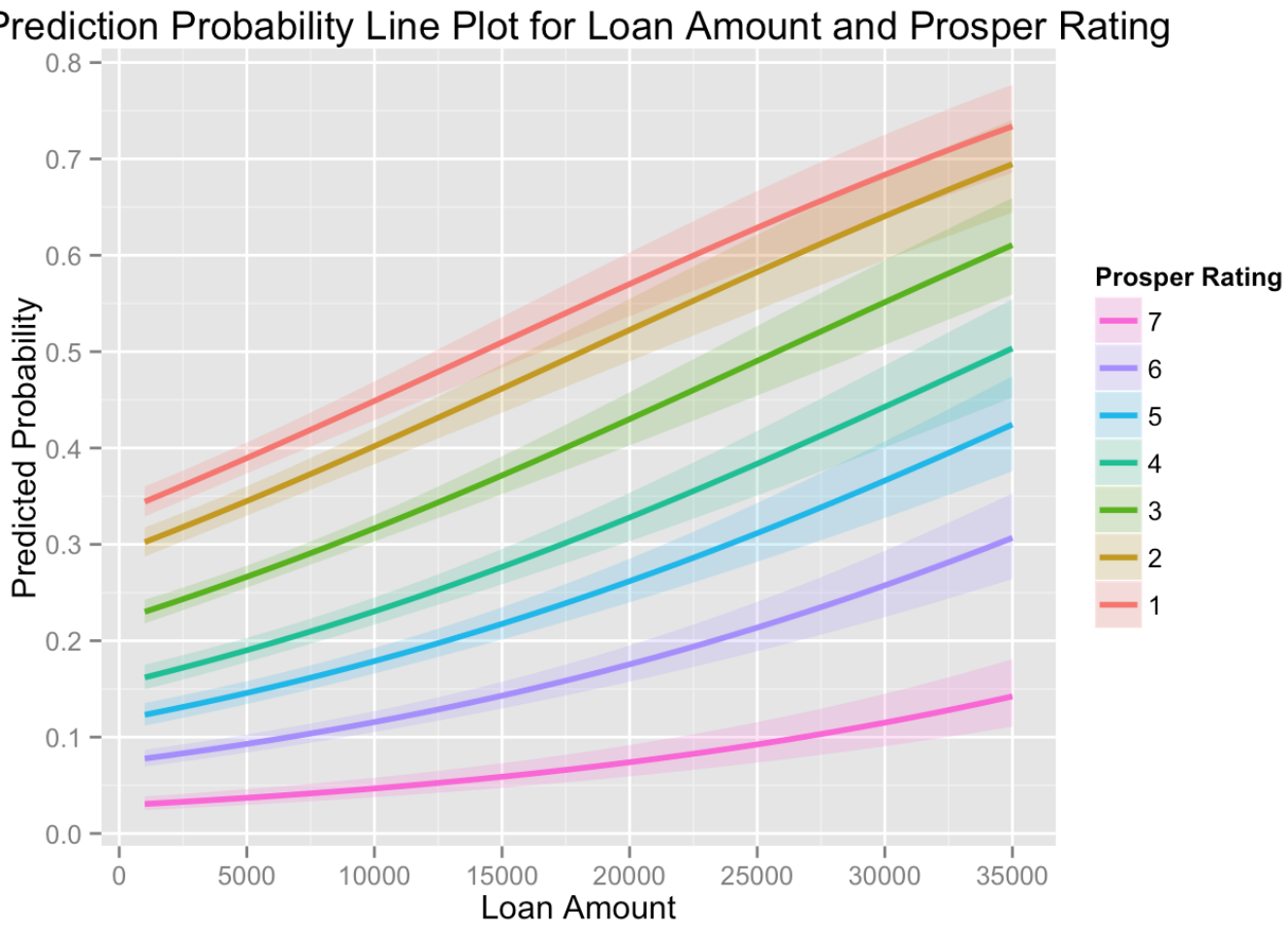
Plot 2

Description 2

Initially I planned on calculating actual return and loss rates but due to the absence of daily average balance data you really cannot calculate actual rates. You can calculate simple rates based on the Interest and Fees and Non-Principal Recovery Payments variables and Loan Amount. However, I decided to focus my analysis on the derived binary variable for Principal Loss. These plots provide a great visual of the linear relationship between Borrower Rate and Estimated Loss/Return with the outliers for the 1 - HR credit rating.

I did some customization of the final plot to arrange the plots together with only 1 x axis label and legend guide positioned at the bottom.

Plot 3



Description 3

The line plot with the prediction probability confidence interval ribbon for the logistic regression model was the culmination of my analysis. This model could be the basis for establishing loan amount limits for each credit rating. From an investor perspective, listings could be run through the model to identify the probability of loss of principal to diversify a portfolio of peer to peer loan investments.

Reflection

The Prosper loan dataset is pretty large with 113,937 loans with 88 variables ranging in loan originations from 2005 - 2014. Based on the number of variables my first step was to review the variable definitions and header data to reduce the columns down to a more manageable list. I then discovered the gap in data between November 2008 - June 2009 due to the SEC shutdown. This was an important observation since it was critical for my subsequent data subsets since the credit rating system changed after they reopened in 2009. Once I explored the estimated loss and return variables my immediate thought was to calculate actual loss and return rates for closed loans. However, I decided not to go down that path since the dataset only had original loan amount and lacked time series daily average loan balance data. I then focused my attention on exploring the variables that had the strongest relationship to rate and loan amount and ultimately which loans defaulted and had any amount of principal charged off. I eliminated fields such as term, monthly income and employment duration since they did not have strong relationships

with my features of interest. I really honed in on the prosper credit rating in the bivariate section based on the output of each correlation matrix. At that point, I really established my direction and brought it all together in the multivariate section and creation of the predictive model.

Initially I was not creating values for each plot and during the course of my analysis I struggled a little keeping everything straight on what plots I had already created. By using a specific naming convention this really helped keep my project organized and I could review what values I had in the environment already.

My plan is to continue on with this analysis but on Prospers main competitor Lending Club to see if there is any measurable performance differences between the two peer to peer lenders.