Exploration of Prosper Loan Data

Christopher Ivanovich

June 6, 2017

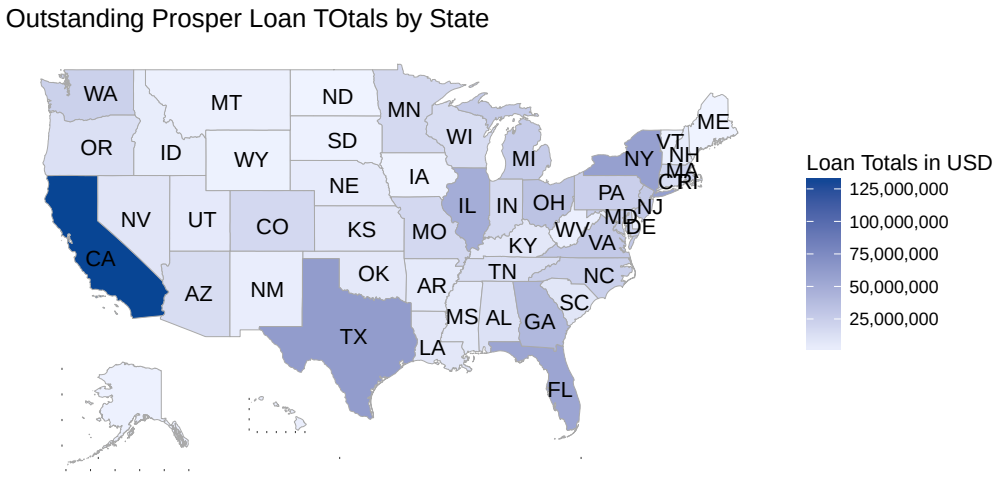
```
knitr::opts_chunk$set(echo = F, warning = F, message = F, fig.width = 7, fig.align = "center", out.widt
```

```
## 'data.frame':    113937 obs. of  16 variables:
##  $ ProsperScore               : num  NA 7 NA 9 4 10 2 4 9 11 ...
##  $ ProsperRating..Alpha.      : Factor w/ 7 levels "A","AA","B","C",..: NA 1 NA 1 5 3 6 4 2 2 ...
##  $ 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 ...
##  $ CreditScoreRangeUpper      : int  659 699 499 819 699 759 699 719 839 839 ...
##  $ CreditScoreRangeLower      : int  640 680 480 800 680 740 680 700 820 820 ...
##  $ IsBorrowerHomeowner        : Factor w/ 2 levels "False","True": 2 1 1 2 2 2 1 1 2 2 ...
##  $ IncomeRange                : Factor w/ 8 levels "$0","$1-24,999",..: 4 5 7 4 3 3 4 4 4 4 ...
##  $ StatedMonthlyIncome        : num  3083 6125 2083 2875 9583 ...
##  $ DebtToIncomeRatio          : num  0.17 0.18 0.06 0.15 0.26 0.36 0.27 0.24 0.25 0.25 ...
##  $ MonthlyLoanPayment         : num  330 319 123 321 564 ...
##  $ LoanOriginalAmount         : int  9425 10000 3001 10000 15000 15000 3000 10000 10000 10000 ...
##  $ AmountDelinquent           : num  472 0 NA 10056 0 ...
##  $ BorrowerState              : Factor w/ 51 levels "AK","AL","AR",..: 6 6 11 11 24 33 17 5 15 15 ..
##  $ ProsperPrincipalOutstanding: num  NA NA NA NA 9948 ...
```

In this report, we are concerned with the Prosper Loans dataset, containing 113,937 observations and 16 variables chosen from the full set of 81.

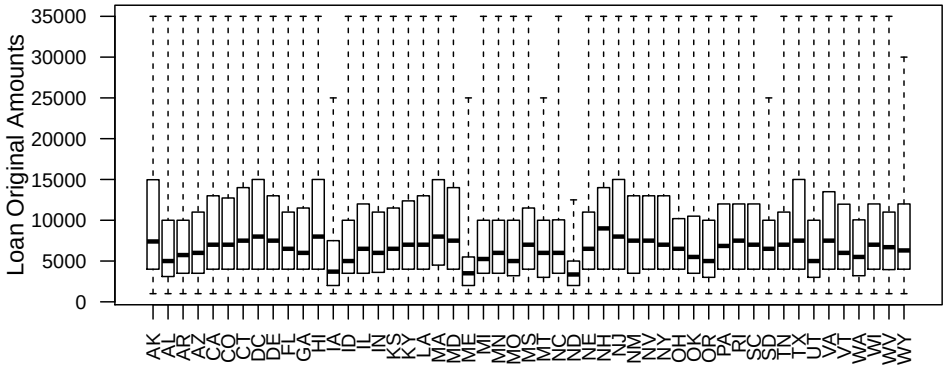
Nota bene: Many fields in the source CSV file are missing values. When importing the data-set into R Studio, I assign NA values to these empty fields.

Out of Curiosity, What is the Outstanding Prosper Indebtedness of a Given State?



Clearly, California leads the way, at least among this set of borrowers for which state values were listed. This may have something to do with Prosper's popularity among the Silicon Valley set as a microlending platform, since it is rooted in the startup culture of that place, or it may have more to do with state populations, as other stand-out states include Texas, Illinois, and New York.

Another way to look at this is with a barchart faceted across all of the provided states:



There were many empty values for borrowers' states of residency, which somewhat boggles the mind–I can't imagine one caould take out a loan from Prosper without providing at least this modicum of location data.

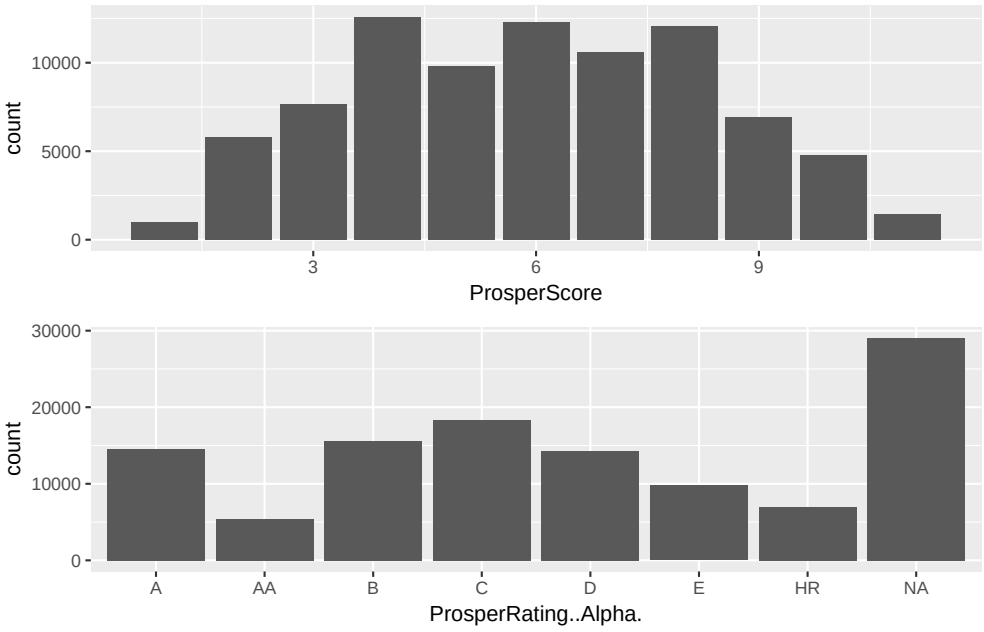
This begs the question, how riddled with NAs is this dataset? How do NA-values vary across the different fields of the data set? Let's find out!

```
##                                round.na_count.nrow.loans_sub...4.
## ProsperScore                                               0.2553
## ProsperRating..Alpha.                                      0.2553
## BorrowerAPR                                                0.0002
## BorrowerRate                                               0.0000
## LenderYield                                                0.0000
## CreditScoreRangeUpper                                      0.0052
## CreditScoreRangeLower                                      0.0052
## IsBorrowerHomeowner                                        0.0000
## IncomeRange                                                0.0000
## StatedMonthlyIncome                                        0.0000
## DebtToIncomeRatio                                          0.0751
## MonthlyLoanPayment                                         0.0000
## LoanOriginalAmount                                         0.0000
## AmountDelinquent                                           0.0669
## BorrowerState                                              0.0484
## ProsperPrincipalOutstanding                                0.8062
```

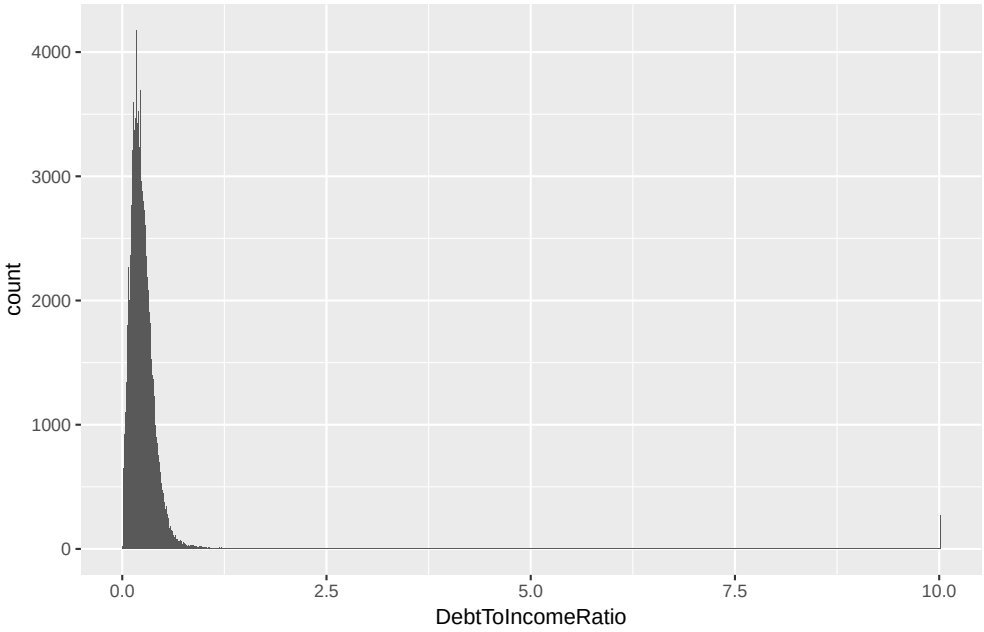
These values are proportions of the total number of observations, and we see that the ProsperPrincipalOutstanding field is missing for 80% of the entries. I suspect this may reflect loans now fully paid-off or otherwise written-off.

Now, Some Univariate Plots with Analysis

Prosper's Assigned Borrower Scores and Ratings



Debt to Income Ratio



We note the mode around a value of 0.50, as well as the distant outliers in which borrowers on Prosper apparently have 10.01-times as much debt as income–one wonders what their Prosper scores are, and how many investors they could get with such a scary ratio of debt to income. . .

```
##   ProsperScore   ProsperRating..Alpha.         IncomeRange  StatedMonthlyIncome DebtToIncomeRatio Mo
##  Min.   :1.000   HR     : 19           $1-24,999     :187   Min.   :    0.000   Min.   :10.01     Mi
##  1st Qu.:3.000   E      : 11           Not displayed : 59   1st Qu.:    0.083   1st Qu.:10.01     1s
##  Median :4.500   B      :  6           Not employed  : 24   Median :    0.083   Median :10.01     Me
##  Mean   :4.478   D      :  5           $100,000+     :  1   Mean   :  133.299   Mean   :10.01     Me
##  3rd Qu.:6.000   C      :  3           $50,000-74,999:  1   3rd Qu.:    0.396   3rd Qu.:10.01     3r
##  Max.   :8.000   (Other):  2           $0            :  0   Max.   :17083.333   Max.   :10.01     Ma
##  NA's   :226     NA's   :226           (Other)       :  0
##  LoanOriginalAmount
##  Min.   : 1000
##  1st Qu.: 3000
##  Median : 6000
##  Mean   : 8296
##  3rd Qu.:10000
##  Max.   :25000
##
```

Quite a surprising amount of variation in this group of

```
## [1] 272
```

borrowers. Apparently Prosper is willing to assign decent scores to people in this group, perhaps weighting income and homeownership more heavily than debt-to-income ratio.

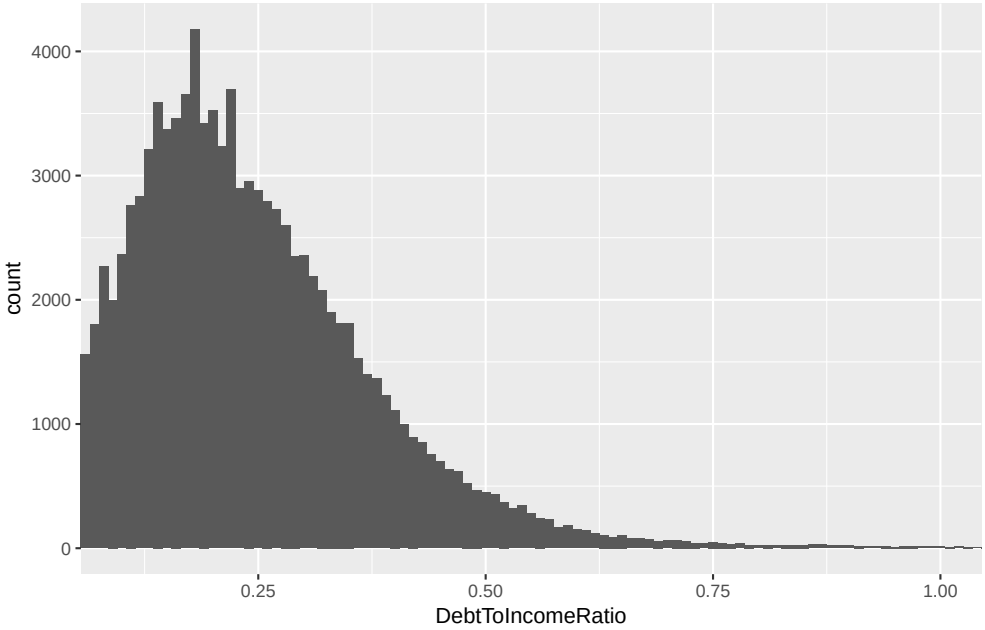
A quick call to the `table` function, to take a look at the income ranges of those with such extreme debt-burdens, reveals:

```
##
##      $1-24,999  Not displayed   Not employed      $100,000+ $50,000-74,999            $0 $25,000-49
##            187             59             24              1              1             0
```

These results likely scuttle my pet theory that a significant number of these highly indebted individuals are upper-class individuals using Prosper as a source of speculative capital. It looks like the majority are quite low-income, or else lying about their income, as so many have reported $0, and the mean stated income is only ~$133 USD per month.

Debt to Income Historgram, non-special cases

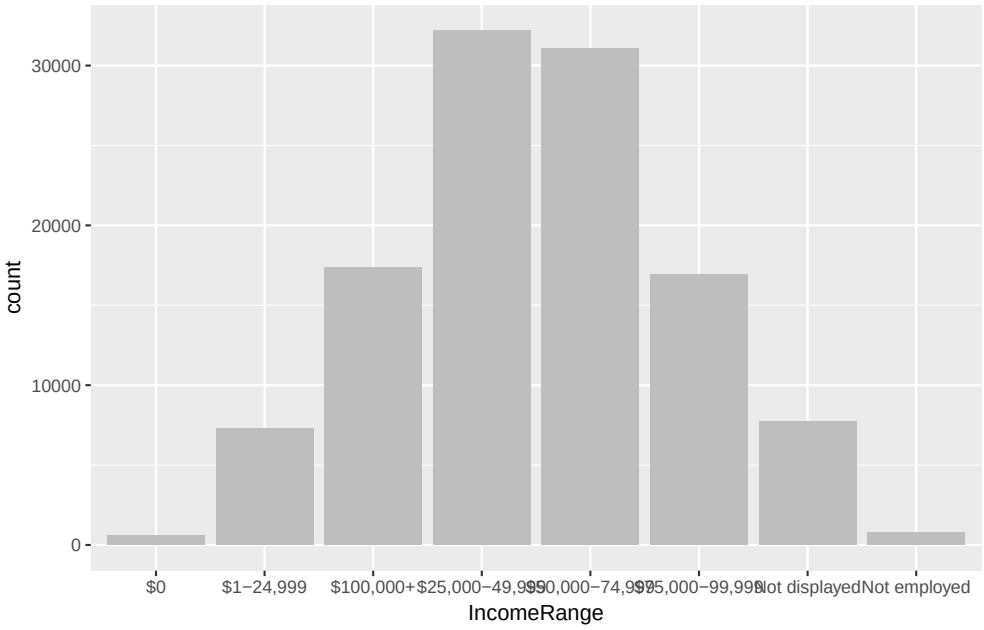
In our *Debt:Income histogram*, the vast majority of the data is clustered within a very narrow range of about [0. 0.8], so we cut out the extreme outliers by setting reasonable limits of 0.1 and 1.0:



Here we see a normal, positively-skewed distribution. Our earlier summary data for *DebtToIncomeRatio* tells us that the mean value is 0.276.

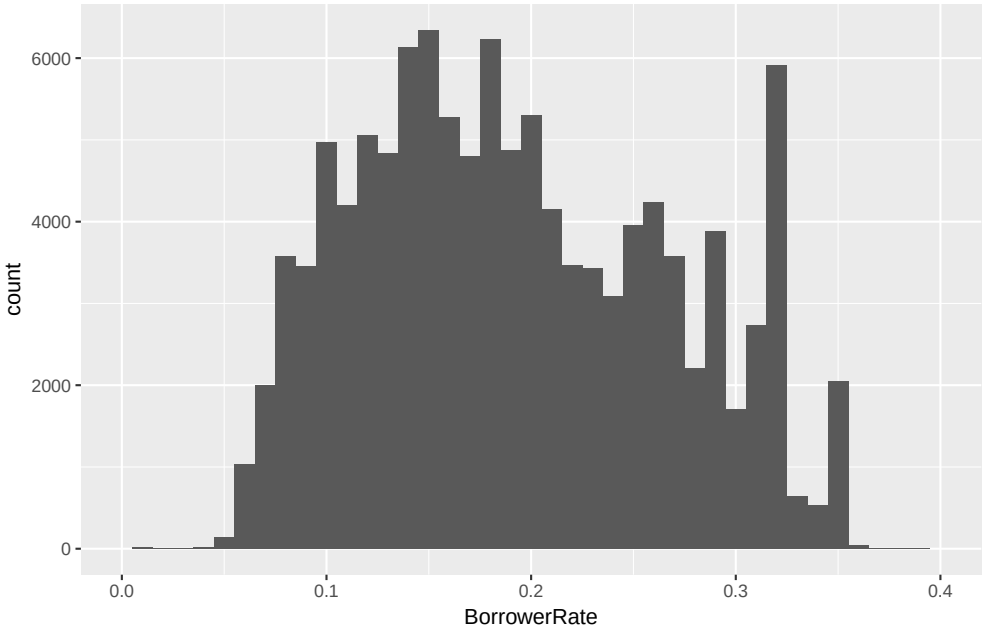
Income Range

For our next plot, we look at *Income Range*:



This barchart looks wonderfully normal as well, though the granularity of such wide intervals is probably quite low.

Next, we look at *Borrower Rate*, which is the interest rate assigned to the borrower by Prosper:



These are percentage-rates, and we see a nearly normal distribution with an interesting secondary peak around 32% interest, which is quite a high rate. This makes me wonder what the Prosper scores are for these borrowers, that so many of them would be dealt such a punishing interest rate.

How many borrowers are in this rate bracket, and what do their other variables look like? We subset for borrowers with rates in the open interval of (0.31, 0.32), which contains 7427 loan data entries, and we take a summary of a selection of variables:

```
##  ProsperRating..Alpha. StatedMonthlyIncome  BorrowerRate     LenderYield     CreditScoreRangeUpper
```

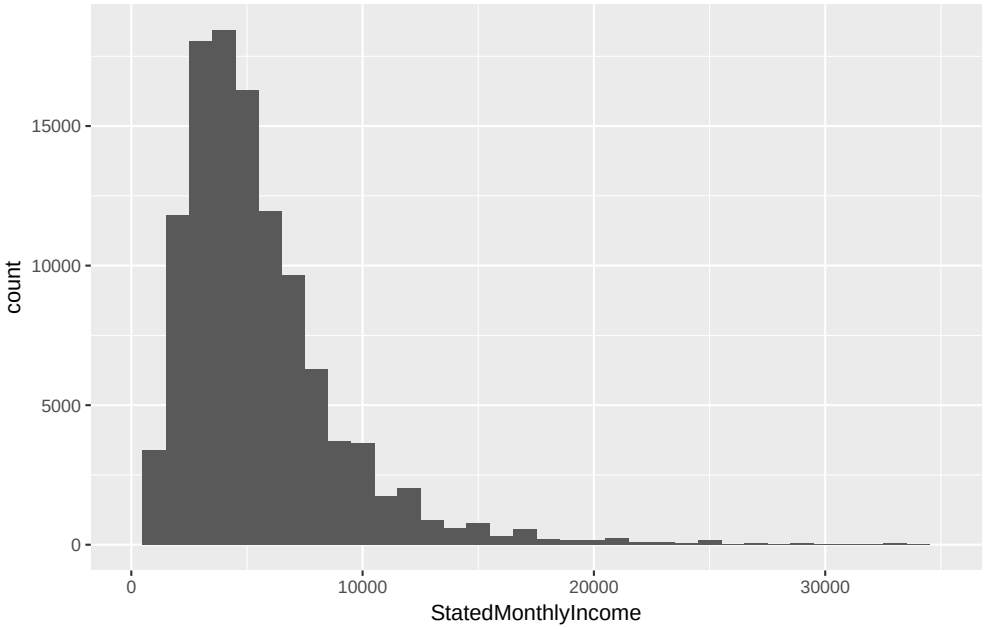
```
##  HR     :5562          Min.   :      0     Min.   :0.3100   Min.   :0.2900   Min.   :539.0
##  E      :1654          1st Qu.:   2500     1st Qu.:0.3160   1st Qu.:0.3060   1st Qu.:679.0
##  D      :  35          Median :   3875     Median :0.3177   Median :0.3077   Median :699.0
##  C      :   3          Mean   :   4992     Mean   :0.3171   Mean   :0.3071   Mean   :698.8
##  B      :   1          3rd Qu.:   5833     3rd Qu.:0.3177   3rd Qu.:0.3077   3rd Qu.:719.0
##  (Other):   0          Max.   :1750003     Max.   :0.3200   Max.   :0.3100   Max.   :879.0
##  NA's   : 172                                                                 NA's   :5
##  CreditScoreRangeLower
##  Min.   :520.0
##  1st Qu.:660.0
##  Median :680.0
##  Mean   :679.8
##  3rd Qu.:700.0
##  Max.   :860.0
##  NA's   :5
```

As we see, most of these people have the worst possible Prosper Rating of "HR", i.e. high-risk, with on average higher debt:income ratios, but a surprisingly high monthly income mean and median. We note that the max stated income value in our subset matches that in the summary of the full dataset, and clearly this high-earner is distorting the mean. Yet, the median monthly income remains solidly high, even for this HR subset, which might help explain why the *LenderYield* parameters do not deviate substantially from the interest rate.

From the perspective of Prosper, it would make sense to only offer these high-risk, high interest rate loans to individuals with high incomes.

The distribution of the credit scores is a bit of a mysterious–we would expect somewhat uniformly low values for such high interest rates and terrible ratings. We also note that there are only 5 NA values for credit scores in this entire data set–this suggests that Prosper deems these values as of great importance, and they may be the best predictor of loan size, interest rate, and other fields (to be explored in the multivariate section further below).

For now, let's take a look at the distribution of monthly incomes with the self-professed millionaire outlier removed, to see if anything else emerges.

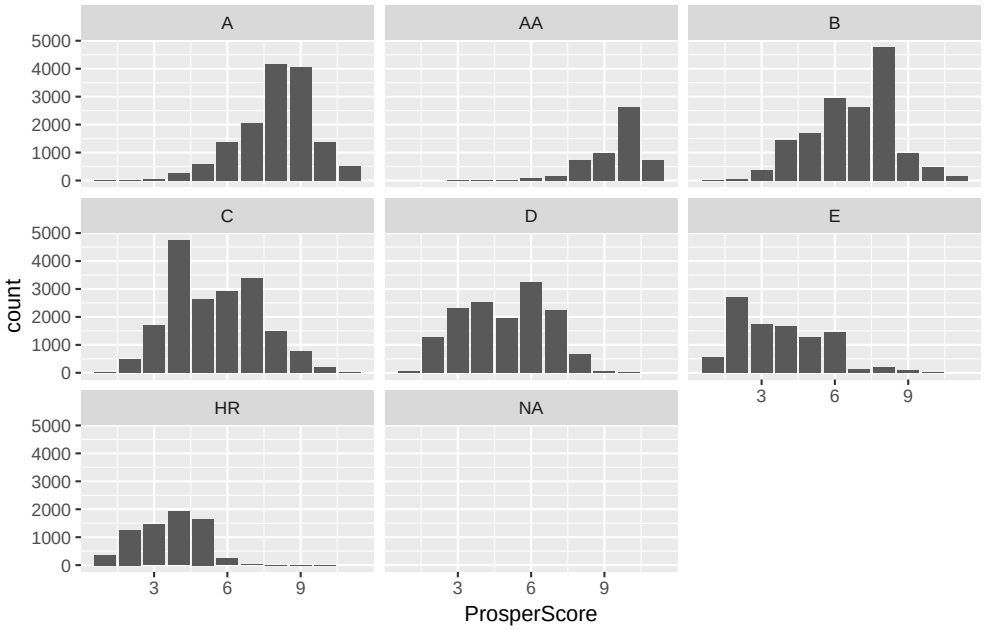


It looks like the bulk of the borrowers here inhabit the American middle-class income range, and now the intervals chosen by the data curator in the IncomeRange variable make more sense.

We could log10 transform our counts to flatten the peaks a bit and better visualize the right tail of this positively skewed distribution, but no good reason to do so comes to mind just yet. Given the non-trivial amounts of high-earners here, though, one wonders what their occupations and other indicators of wealth and financial activity might look like (or are they simply exaggerating?).

Assigning Loan Risk: Prosper Ratings versus Prosper Scores

Let's take a look at the letter-score Prosper uses to rate the risk of a particular loan to a particular lender, alongside the numerical score.



These histograms count up Prosper Scores, separated by so-called Prosper Ratings.

At first glance, this appears quite strange. Why is there so much divergence between the two scoring methods? Prosper is pretty mum on the definition of the *ProsperRating* (AA-HR), beyond calling it a proprietary scoring method for evaluating risk, with AA being the best, HR being the worst.

ProsperScores (not to be confused with Prosper's (Numeric) Ratings) are also defined as a risk-measuring tool in the range:

1 -> 11, i.e., Riskiest -> Safest

If these methods have any kind of parity, we would expect a very high density of 10-11 Scores to be in the "AA" rating graph above, and we'd only expect to see scores of 1-2 or so under the "HR" facet of our graph.

The power of the `table` function allows us to quickly tabulate the frequencies of ProsperScore vs. Prosper-Rating:

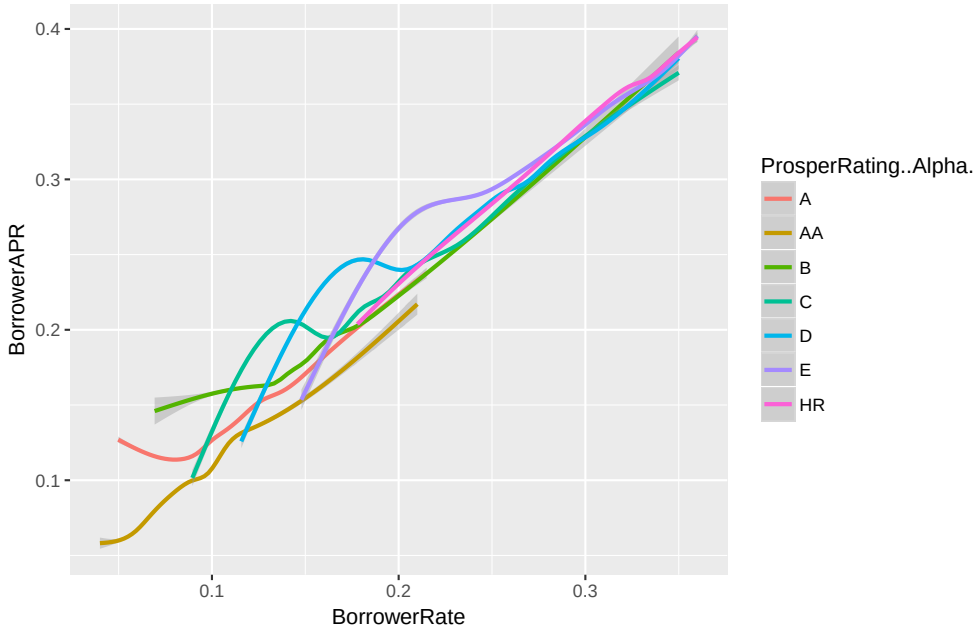
```
##              ProsperRating..Alpha.
## ProsperScore    A   AA    B    C    D    E   HR
##            1     2    0    3   21   49  552  365
##            2     6    0   61  490 1255 2692 1262
##            3    53    4  367 1711 2295 1750 1462
##            4   285   14 1451 4738 2506 1681 1920
```

```
##            5   608   11 1695 2642 1940 1262 1655
##            6  1392   94 2968 2900 3256 1438  230
##            7  2050  164 2624 3370 2251  118   20
##            8  4170  718 4781 1482  679  214    9
##            9  4070  983  968  770   36   75    9
##            10 1394 2645  492  196    7   13    3
##            11  521  739  171   25    0    0    0
```

We see that there does appear to be some consistency in the frequencies between these two methods, but that the relationship appears weaker than we might hope or expect. That is, apart from the large spread of the data, the modal letter Ratings appear to correspond to particular numerical Scores, but we see some oddities: scores of 10-11 peak at the best rating of AA, a score of 9 peaks at A and 8 peaks at B, but 6 peaks at D while 4-5 peak at C.

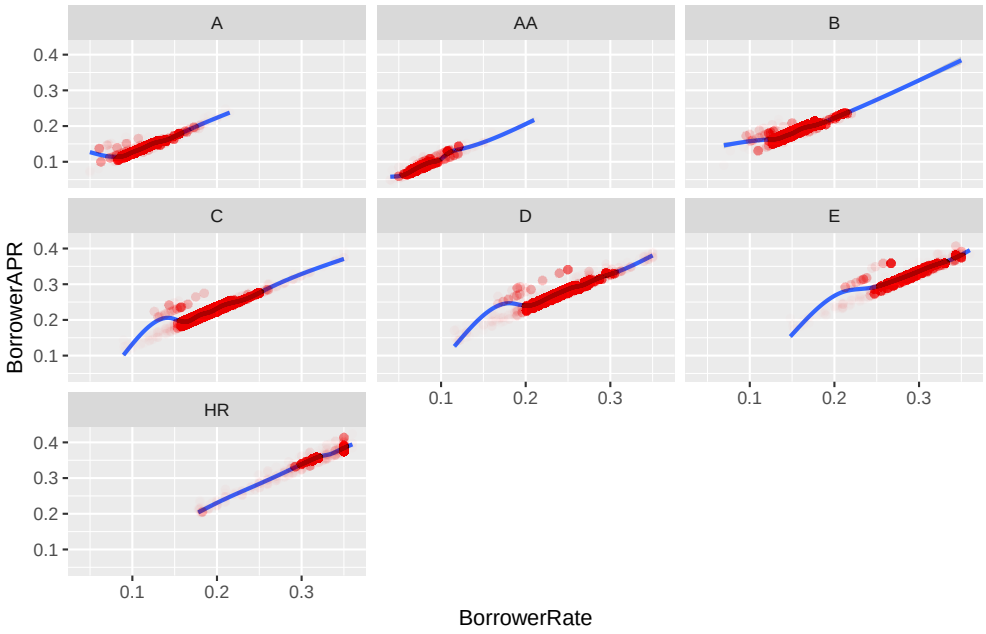
This divergence tells us that there are some significant differences in how risk gets assigned using these two different "proprietary" representations.

Multivariate Plots, Grouping, and Analysis



Why have I plotted a borrower's APR against the interest rate on their loan? Simply to confirm that Prosper is using a consistent, flat method for servicing loans (unlike an interest rate, an APR includes things like service and transaction fees, a.k.a. Prosper's share of the loan pie).

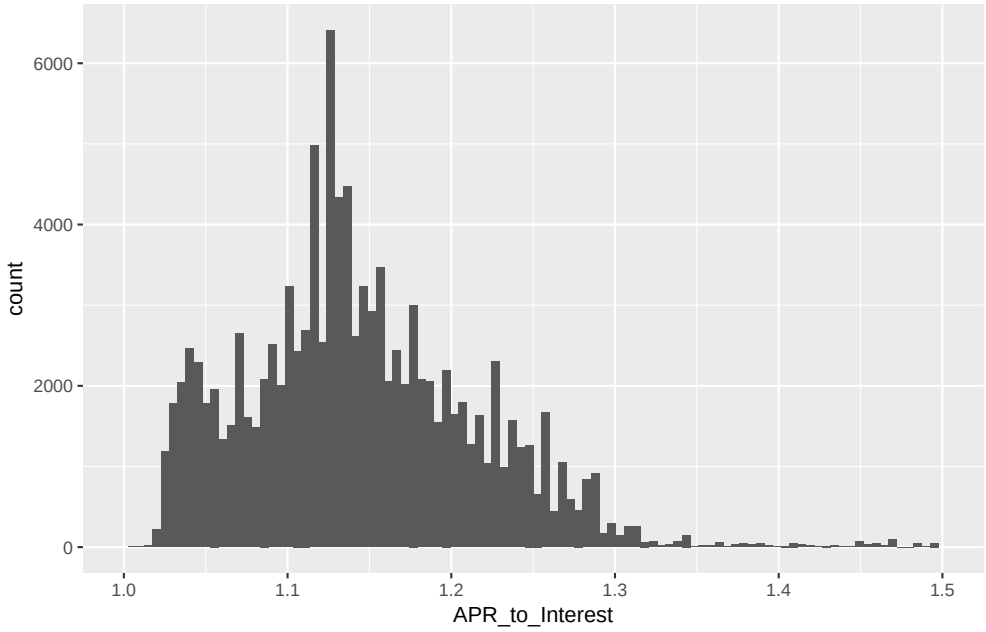
We note some off-trend points, though they are few in number. We can facet these plots to get a clearer look at the Ratings in which the two rates diverge a bit.



Without surprise we note that well-rated loans enjoy generally lower interest rates, though there is quite a wide range of borrowing rates across the Prosper ratings. Is there a graduated interest rate related to the size of the loan? We shelve that question for a moment, to be pursued with our next plot–for now, we note that there are different APRs assigned to the same borrower interest rates even within Rating categories. What's that about?

For the most part, these varying APRs seem to range discretely, i.e., there are simply categories of APR, which we infer by the fact that each facet graph seems to be composed of four or five distinct linear trends, representing the relationship between APR and loan interest rate. This would suggest that certain categories of loans get assigned a certain percentage on top of the loan interest rate to yield an APR–i.e. the percent-increase of the additional interest is not simply a function of a Prosper Rating.

We can further decompose this distinction to see how it varies by creating a new variable, in which we represent the ratio between borrower interest rate and APR.

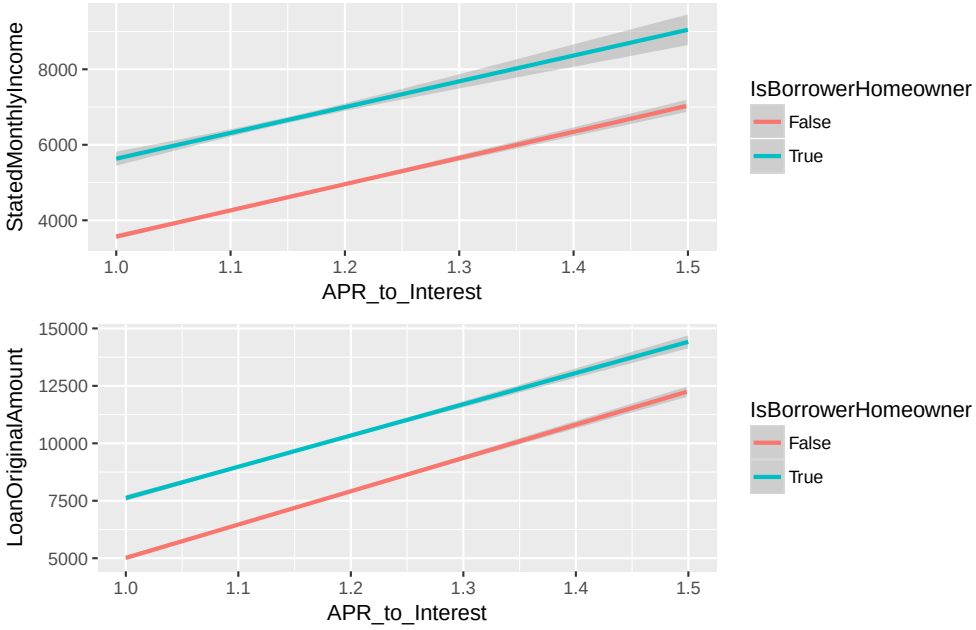


Okay, what is this madness? Why do we see so much variation, one might even say normally distributed variation, including some painfully highl ratios of APR over interest rate?

(It needs to be noted that creation of this ratio variable introduced exactly eight infinite values into our dataset, due to zero-division error. We will remove these as needed for linear modeling.)

Perhaps, more than Prosper Rating, reported income and some binary marker of financial stability, like homeownership or employment, could better predict this discrepancy?

APR-Interest Rate Ratio versus Reported Income, versus Loan Amount, Faceted for Home Ownership Status



Our geometric smoothing function has done a good job of minimizing the effect of those with $0 reported income.

Apart from the obvious fact that homeowners tend to report higher monthly incomes in this dataset (the graph marked True on the rigth), we see, with surprisingly small variation, a solidly positive linear trend between the size of one's APR against the interest rate and one's monthly income. We see essentially the same trend for LoanOriginalAmount and the rate ratio. Probably this is a common-sense finding, but it's nice to see it well-visualized here.

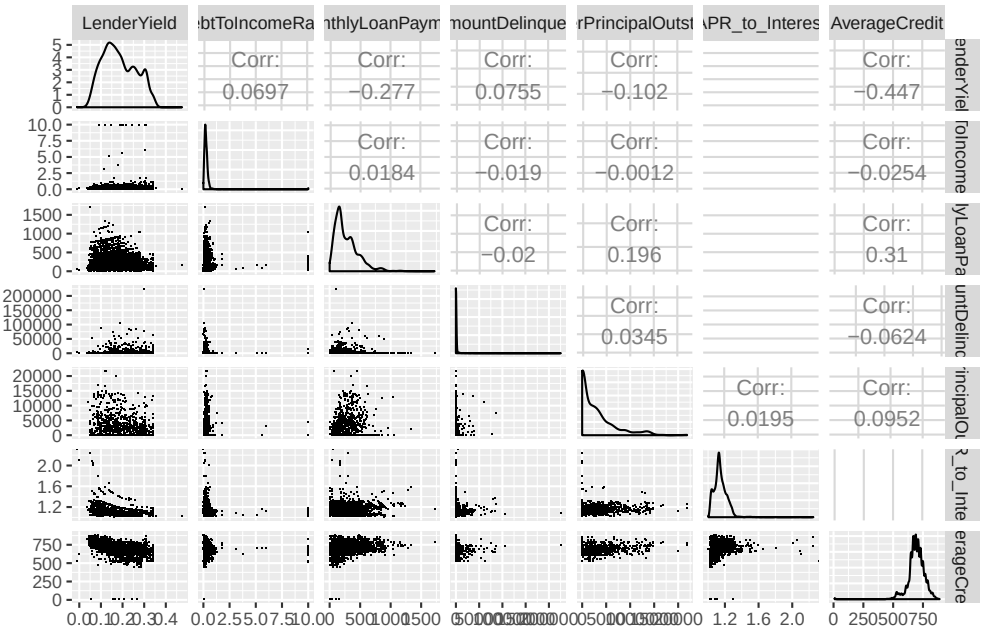
Homeowners seem to enjoy substantially higher minimum loans as well as an average higher income, though naturally, we don't take visual confirmation at face-value: we need to run a linear regression model to evaluate the fit of these lines.

```
##
## Call:
## lm(formula = LoanOriginalAmount ~ APR_to_Interest, data = subset(loans_sub,
##     IsBorrowerHomeowner == "True" & !is.infinite(APR_to_Interest)))
##
## Coefficients:
##     (Intercept)  APR_to_Interest
##          -841.2           9091.7
```

```
##
## Call:
## lm(formula = StatedMonthlyIncome ~ APR_to_Interest, data = subset(loans_sub,
##     IsBorrowerHomeowner == "True" & !is.infinite(APR_to_Interest)))
##
## Coefficients:
##     (Intercept)  APR_to_Interest
##           305.1           5512.4
```

Average Credit Scores, Differences, and a Correlation Matrix

As noted some time ago, credit score seemed to have a major impact on other aspects of a given loan entry–this makes intuitive senese, since Prosper is not likely to take things like reported income, home ownership, and so on into great account, as all of those factors would be accounted for in a credit score.



In the above correlation matrix, we see, not surprisingly, that high credit scores correlate negatively with lender yield–good scores mean lower interest payments.

Other noteworthies include the fairly strong (-o.458) creditnegative coefficient between a Credit Score Rating and the %-gains for the lenders (lender yield). This indicates that those with lower credit scores are more likely to fail to make payments. But, it must be noted, these are the highest recorded credit scores for users–we would likely see a tighter negative correlation if we took an average of highest and lowest scores, or just used the lowest scores.

Average Credit Score Variable and Correlation Coefficients