Exploration of Prosper Loan Data, by Max Sydow

> Prosper Loans is a private personal loan issuer. The overall data set has 82 variables, but here we will look at 18 of them. They are as follows:

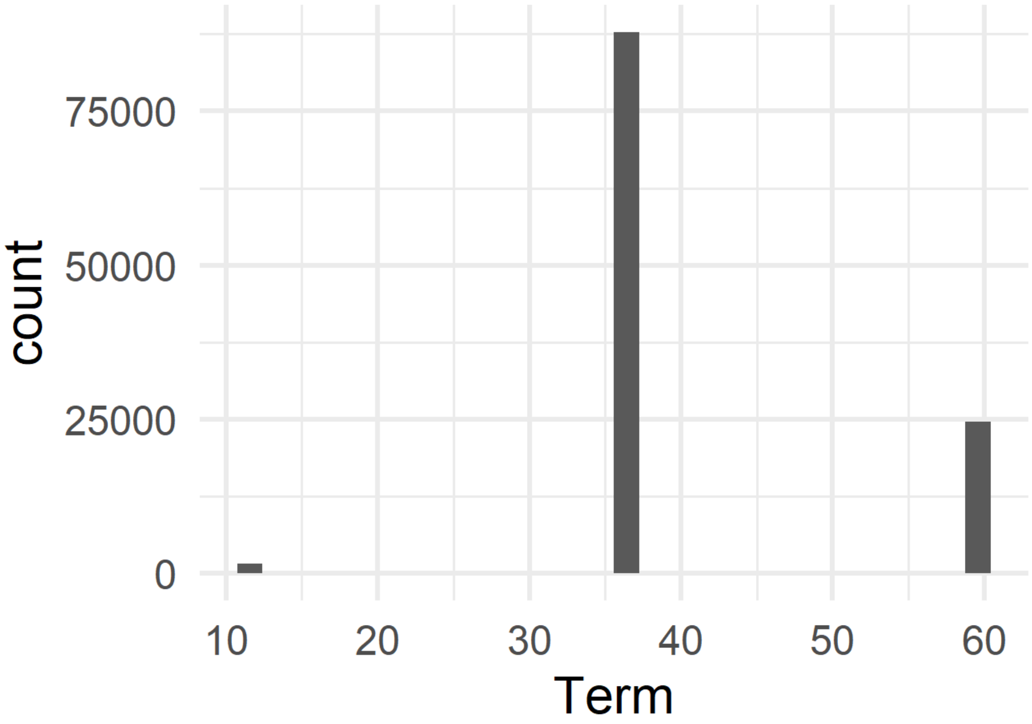
```
str(loan)
```

```
## 'data.frame':    113937 obs. of  18 variables:
##  $ 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 ...
##  $ BorrowerRate             : num  0.158 0.092 0.275 0.0974 0.2085 ...
##  $ BorrowerAPR              : num  0.165 0.12 0.283 0.125 0.246 ...
##  $ ProsperScore             : int  NA 7 NA 9 4 10 2 4 9 11 ...
##  $ BorrowerState            : Factor w/ 52 levels "","AK","AL","AR",..: 7 7 12 12 25 34 18 6 16 16 ...
##  $ 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 ...
##  $ IncomeRange              : Factor w/ 8 levels "$0 ","$1-24,999",..: 4 5 7 4 3 3 4 4 4 4 ...
##  $ LoanOriginalAmount       : int  9425 10000 3001 10000 15000 15000 3000 10000 10000 10000 ...
##  $ DebtToIncomeRatio        : num  0.17 0.18 0.06 0.15 0.26 0.36 0.27 0.24 0.25 0.25 ...
##  $ IncomeVerifiable         : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
##  $ TotalProsperLoans        : int  NA NA NA NA 1 NA NA NA NA NA ...
##  $ Investors                : int  258 1 41 158 20 1 1 1 1 1 ...
##  $ Occupation               : Factor w/ 68 levels "","Accountant/CPA",..: 37 43 37 52 21 43 50 29 24 24 ..
.
##  $ CreditScoreRangeLower    : int  640 680 480 800 680 740 680 700 820 820 ...
##  $ CreditScoreRangeUpper    : int  659 699 499 819 699 759 699 719 839 839 ...
##  $ ListingCategory          : int  0 2 0 16 2 1 1 2 7 7 ...
```

> The data can be found at https://s3.amazonaws.com/udacity-hosted-downloads/ud651/prosperLoanData.csv. A description of each variable is included in this link: https://docs.google.com/spreadsheets/d/1gDyi_L4UvIrLTEC6Wri5nbaMmkGmLQBk-Yx3z0XDEtI/edit#gid=0.

Univariate Plots Section

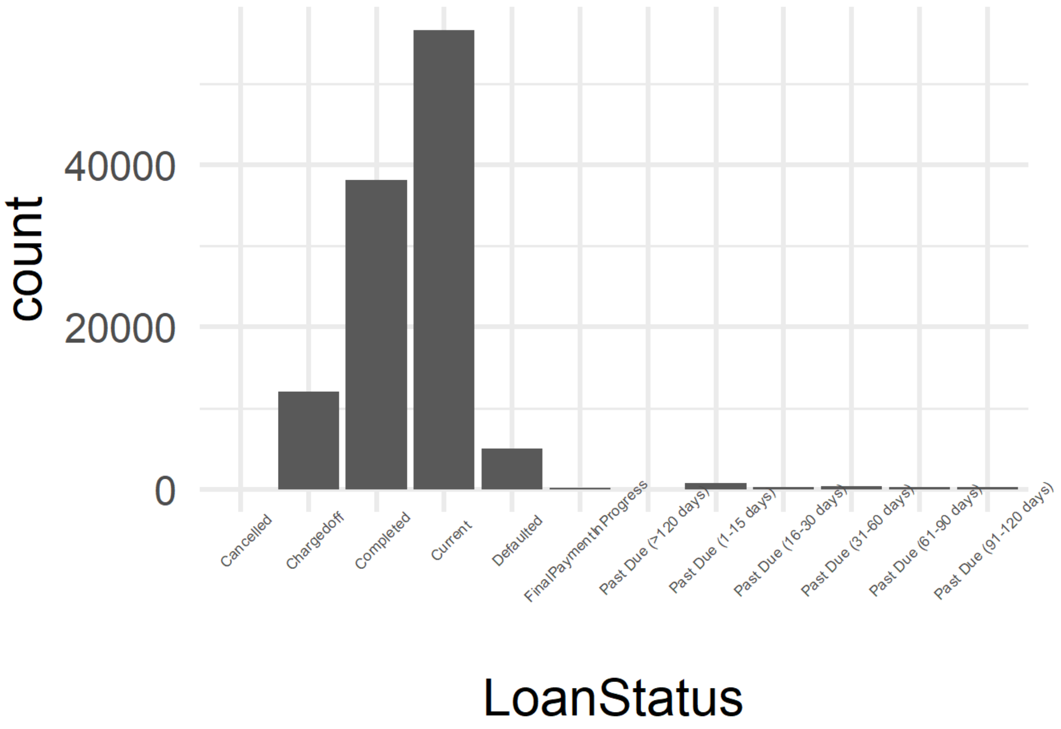
Loan Term Lengths



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.00   36.00   36.00   40.83   36.00   60.00
```

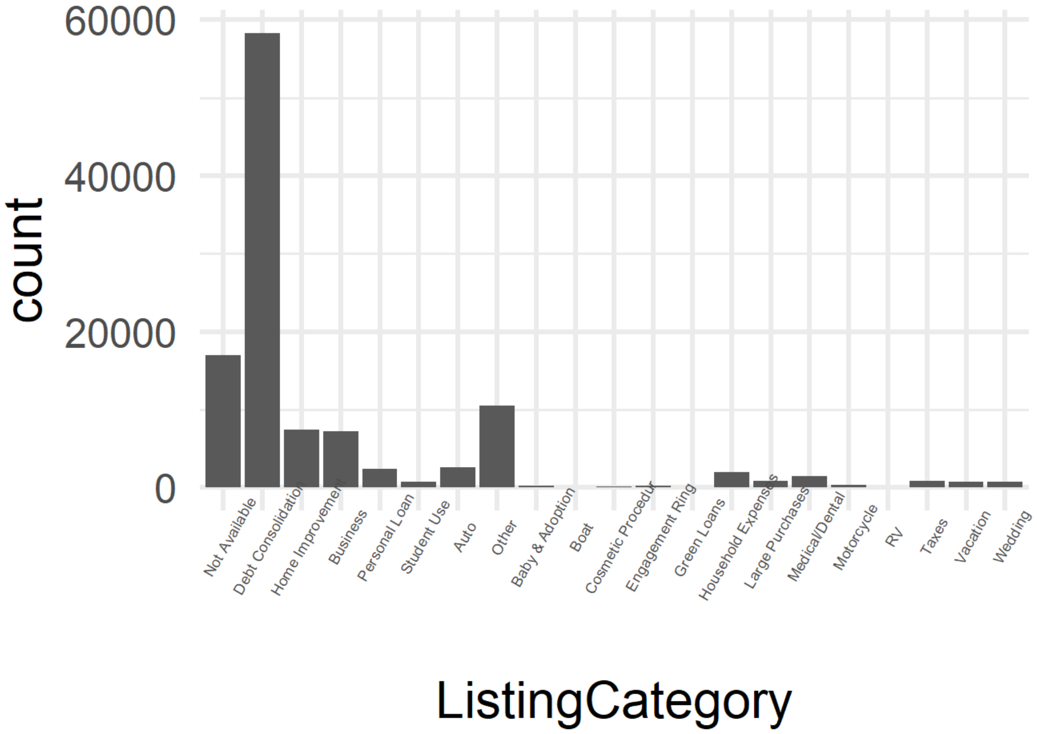
> It looks like there are only 3 lengths of loan terms. Most are for 36 months, fewer have a duration of 60, and a small number were payed off in 12 months.

Loan Status



> The majority of loans are current or completed. More loans were cancelled than defaulted, and there’s a slight decrease in past due loans as time goes on.

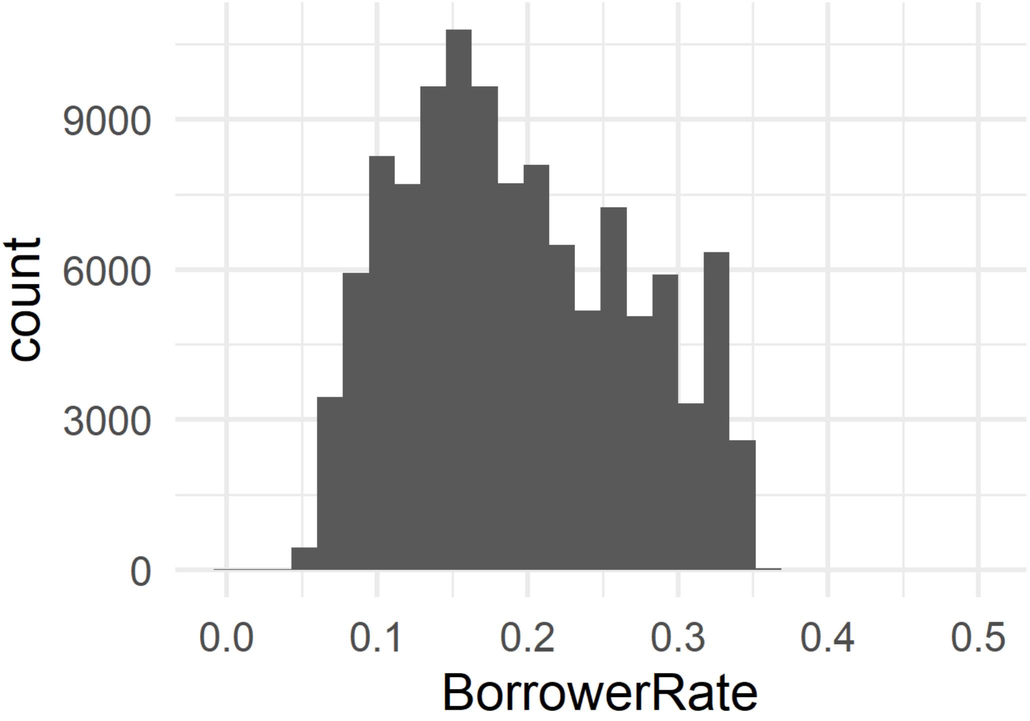
Listing Category



> By far most of the loans are for debt consolidation. There’s also a large portion with no description. Home improvement and business loans are relatively prevelant

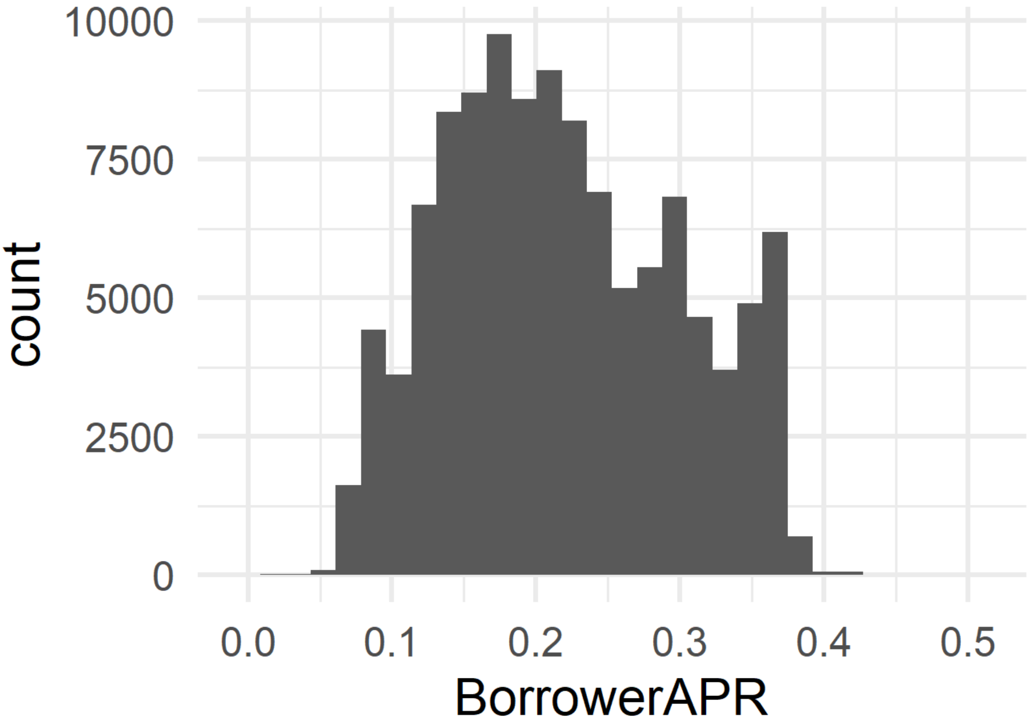
compared to other types.

Borrower Rate



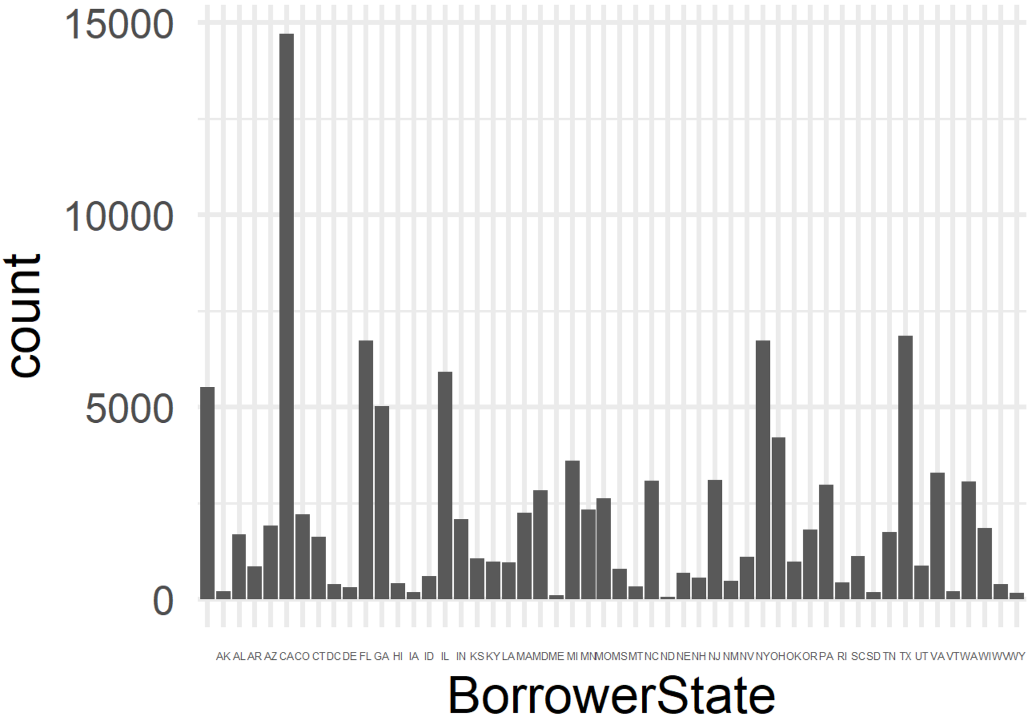
> I chose to look at Borrower Rate instead of APR for simplicity. This is just the overall total amount of interest on the loan. The distribution looks fairly normal with some rates more common than others. For example, 0.325 is far more common than 0.31.

Borrower APR



> The spikes in the Borrower Rate distribution made me want to look at APR as well, so I went back and added it to my data frame. Spikes appear here too. I'm sure there's a reason for this, and maybe accountant for Prosper Loans could give an explanation. For now, I'll just leave it as an interesting observation.

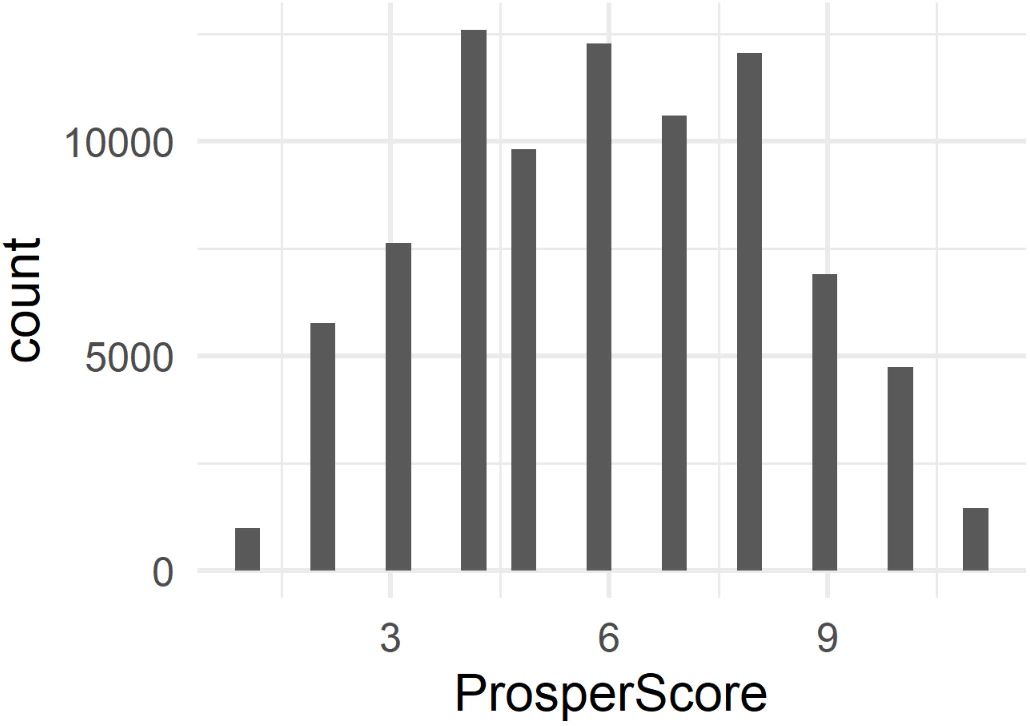
Borrower State



> California has the most

loans. This makes sense, since it is the most populous state. It also makes sense that other populous states like New York, Texas, Florida, and Illinois have higher counts. The unlabled spike might be where state info is not available, or maybe loans from US territories. No description for that category was given in the variable definitions document.

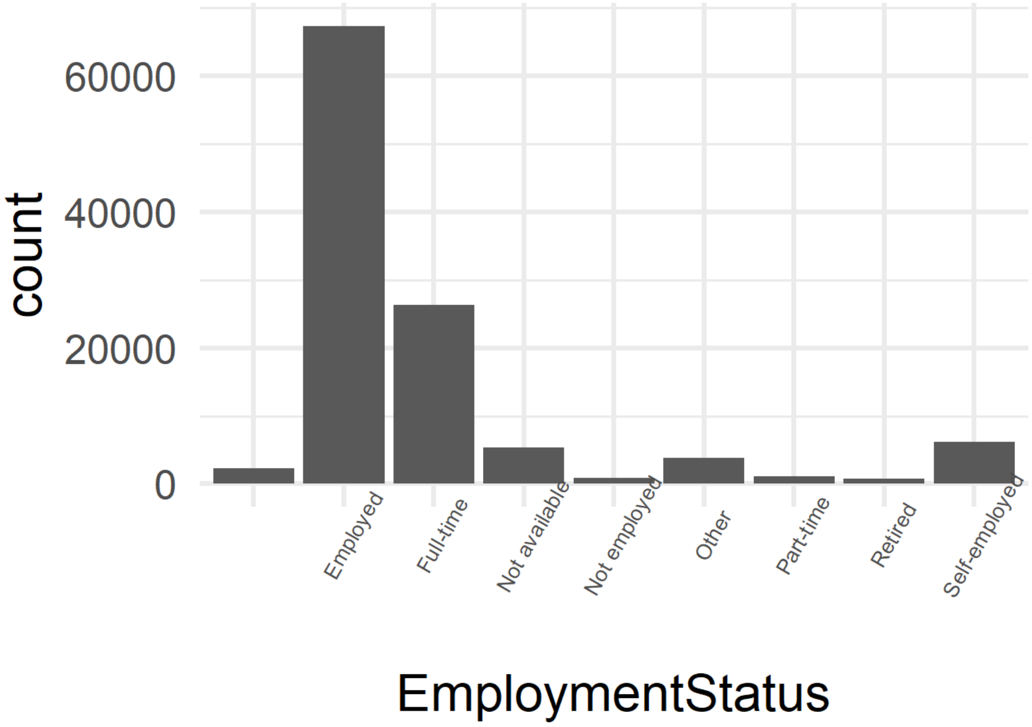
ProsperScore



> Prosper score is a custom

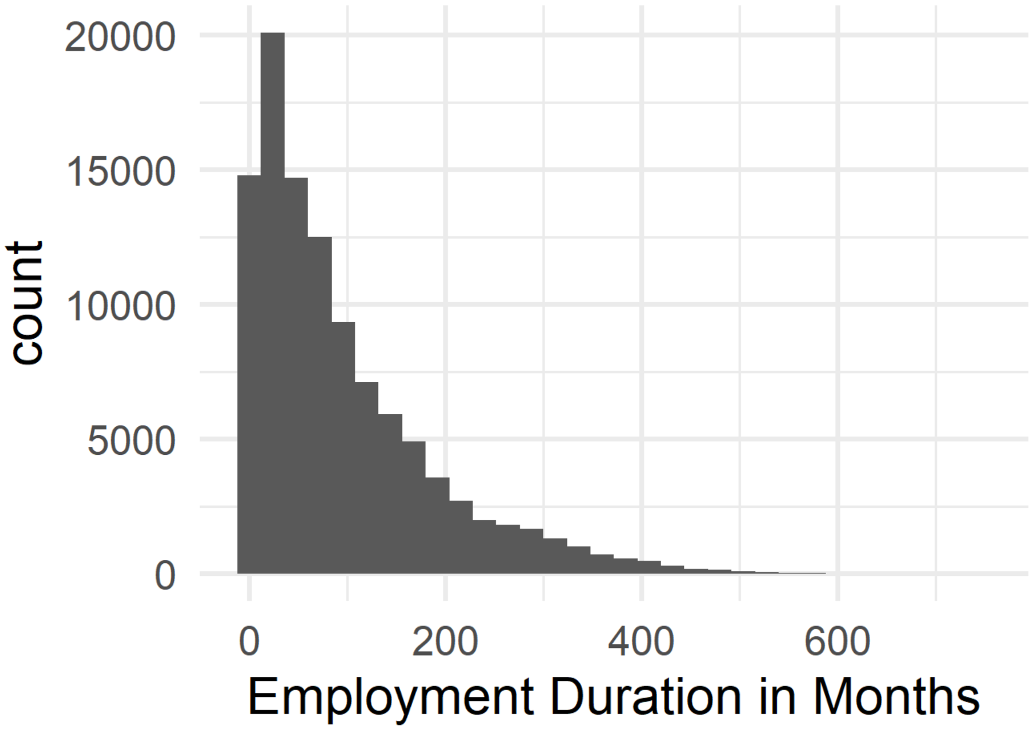
risk score based on in house data, 10 being the best. This distribution is generally normal with some variance in peaks towards center. I’ll leave this as an observation, as I don’t think I could guess why those are occurring.

Employment Status



> Most loan takers are employed, or employed full-time. This seems reasonable. Another significant portion are self-employed. This seems to fit with the independent nature of Prosper Loans and may be related to the occurance of small business loans observed above.

Employment Status Duration



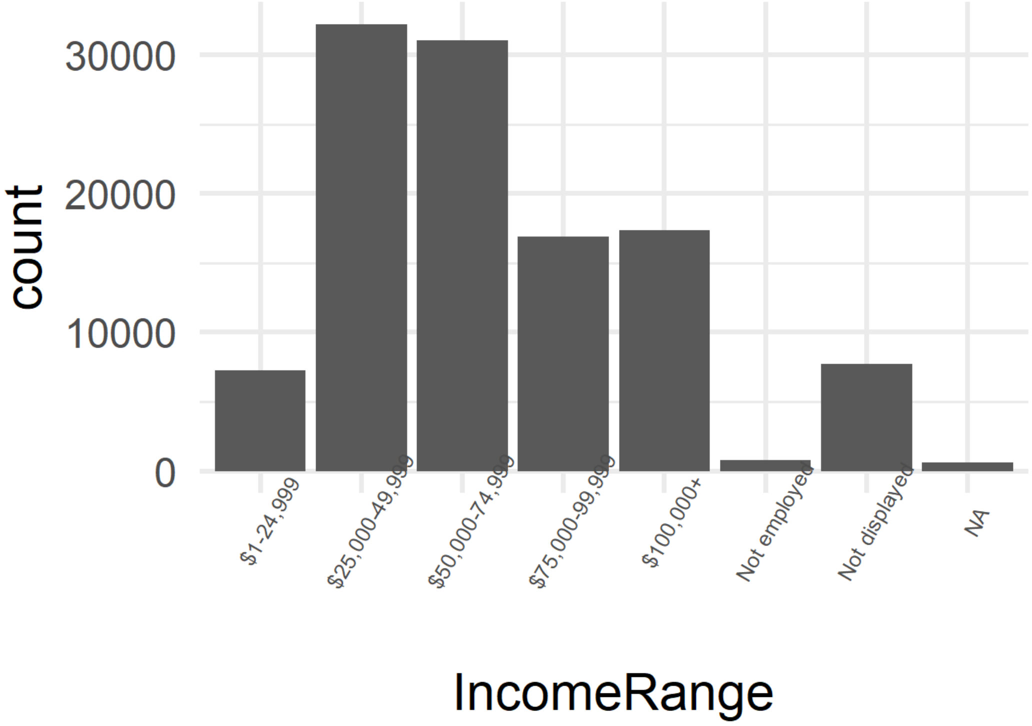
> With a bin width of 24 months it appears that a large number of loan recipients have been employed between 0 and 2 years, with a peak between 2-4 years, then trailing off for longer durations of employemnt. People who have been employed longer may have more savings and less need for loans. Those who are starting their careers may need or want some help with money for some of the categories for loans.

Income Range

```
levels(loan$IncomeRange)
```

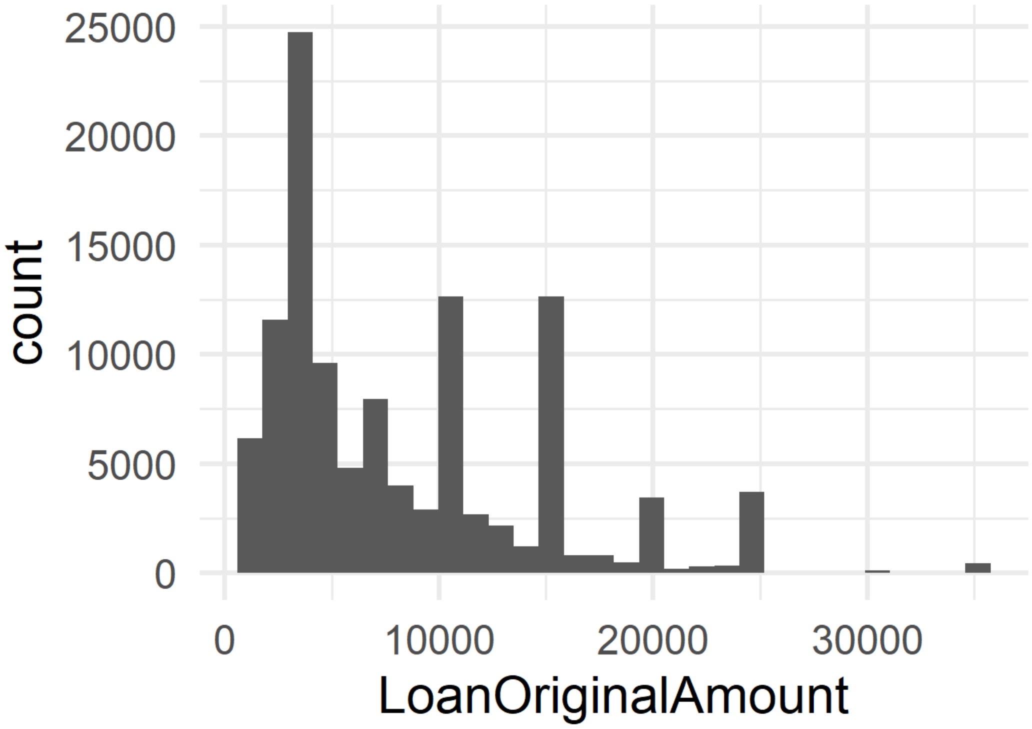
```
## [1] "$0 "            "$1-24,999"      "$100,000+"      "$25,000-49,999"
## [5] "$50,000-74,999" "$75,000-99,999" "Not displayed"  "Not employed"
```

```
loan$IncomeRange <- ordered(loan$IncomeRange, levels = c('$0', '$1-24,999', '$25,000-49,999',
'$50,000-74,999', '$75,000-99,999',
'$100,000+', 'Not employed', 'Not displayed'))
```



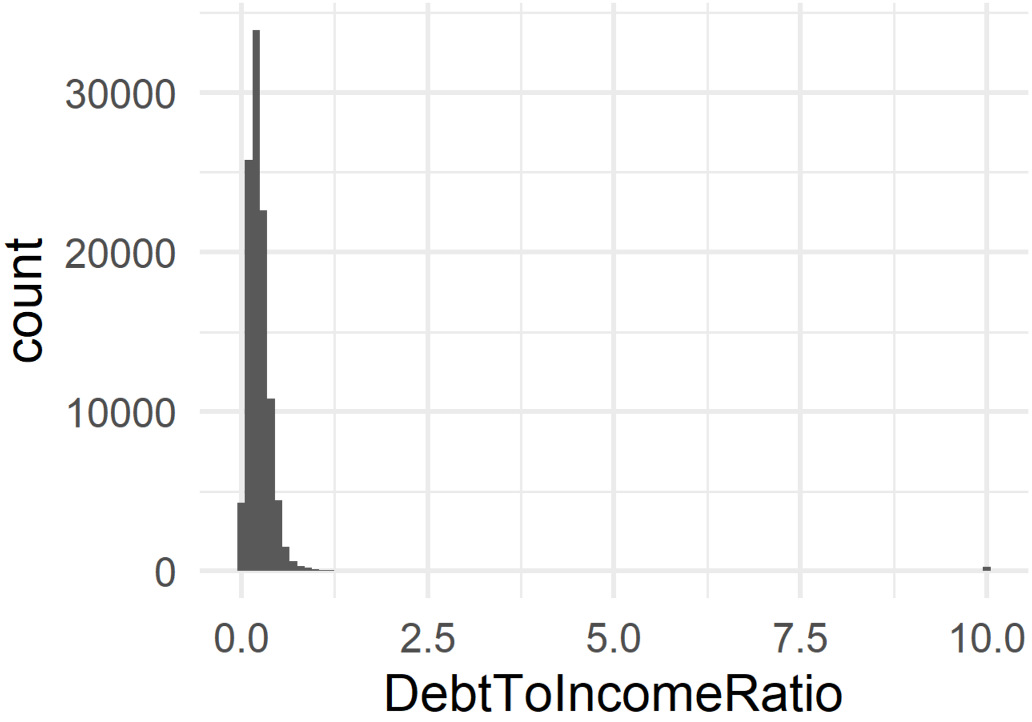
> Most loan takers are in the middle income range from $25k - $74,999k per year. Slightly more people are in the lower half of this income range. I should point out that the original array was out of order, with $100,000+ appearing between $1-24,999 and $25,000-49,999. The array was re-ordered before making the plot seen here.

Original Loan Amount



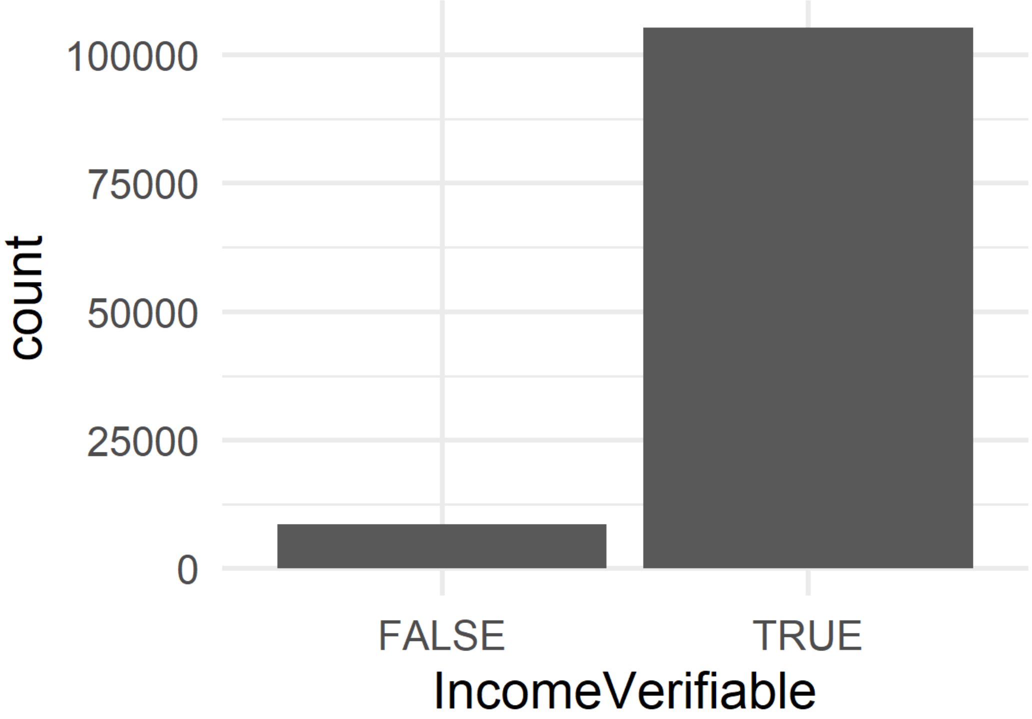
> Overall the distribution is skewed to the right, but there appears to be more common amounts taken in multiples of $10,000.

Debt to Income Ratio



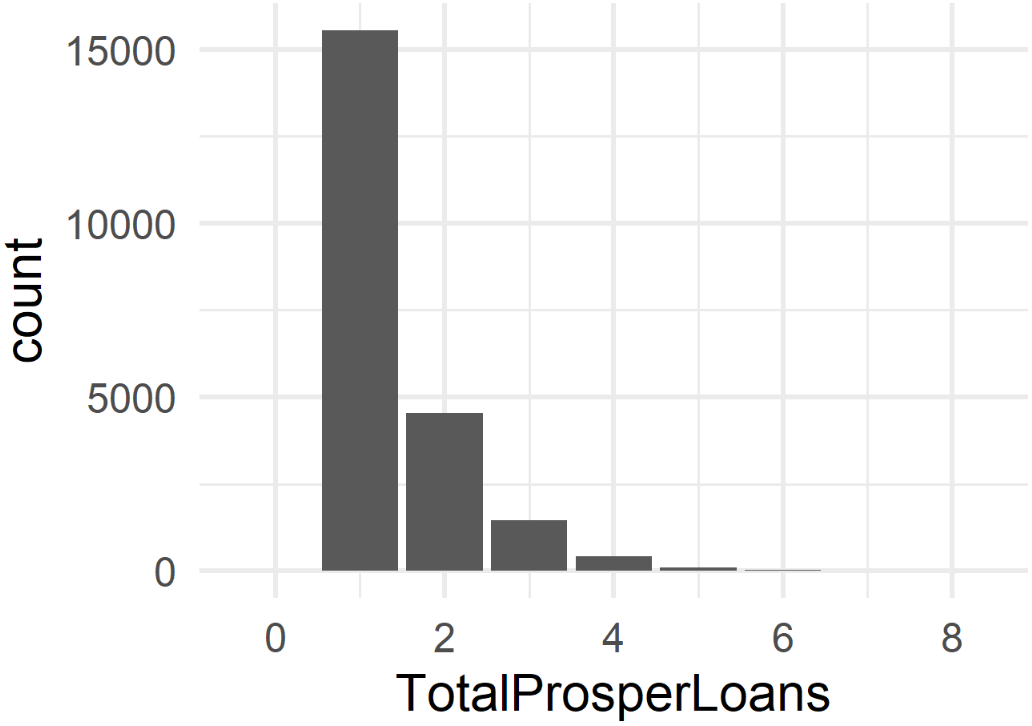
> Most recipients have lower debt to income ratios (DTI) with a peak at 0.3. According to credit.org a good DTI is below 36%, or 0.36. Most Prosper Loan takers have a good DTI, except for the outlier at 10.

Verifiable Income



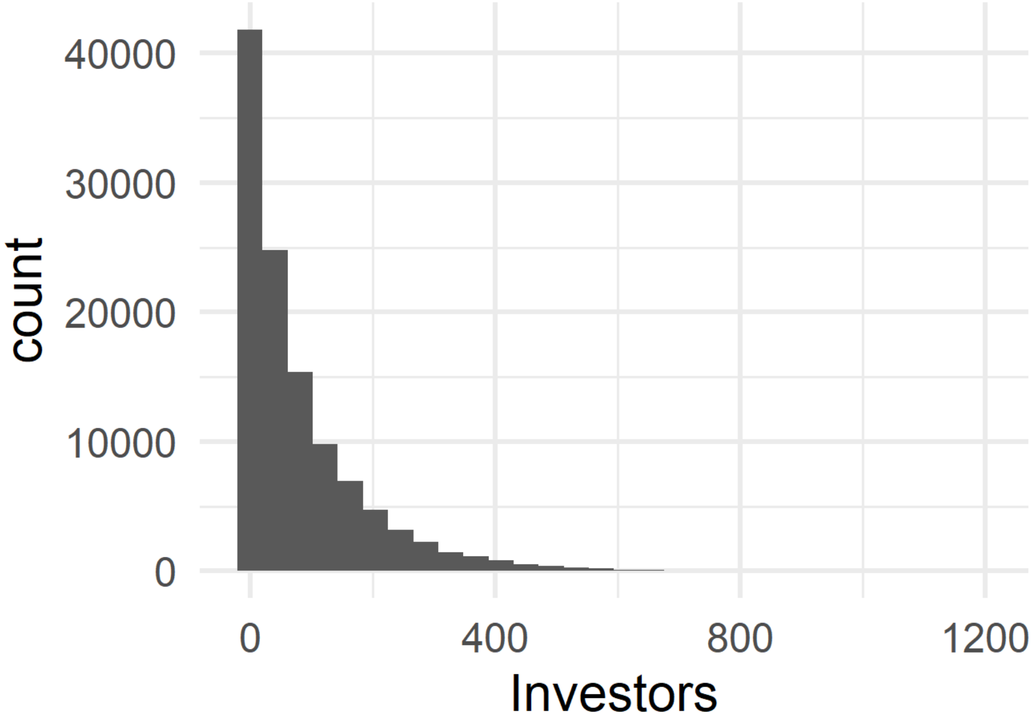
> While the majority of recipients have documentation to verify income there are a significant number who don't. I wander what other criteria are accepted in these cases.

Total Prosper Loans



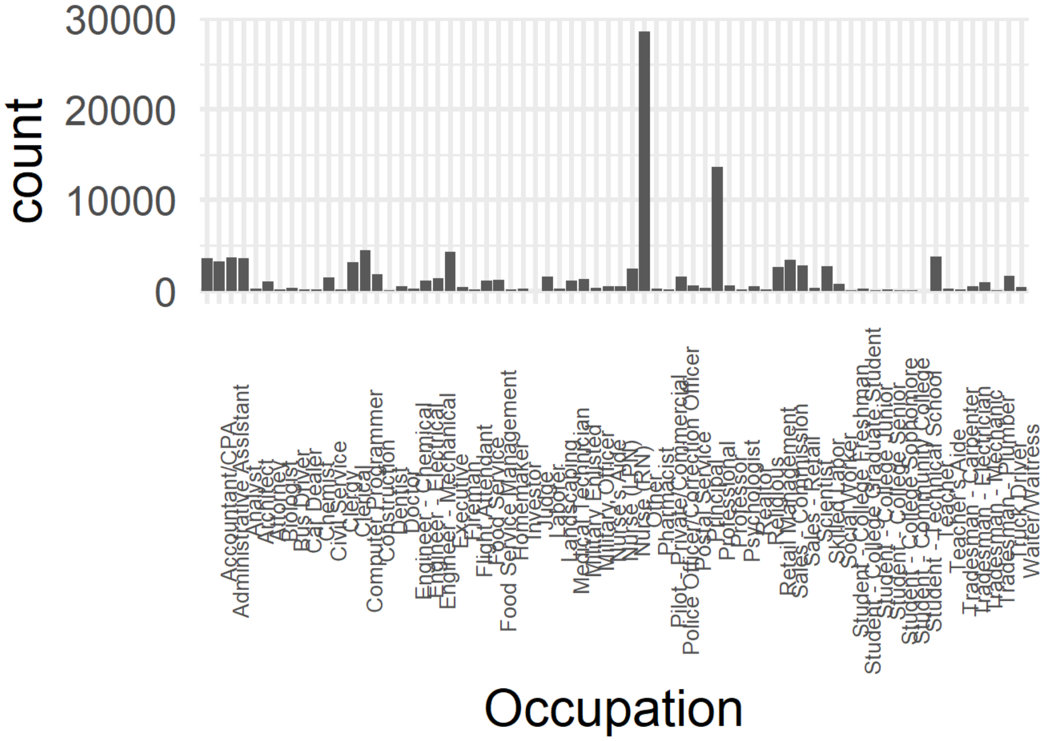
> Most people have taken only one loan, with a tapering off for subsequent loans.

Investors



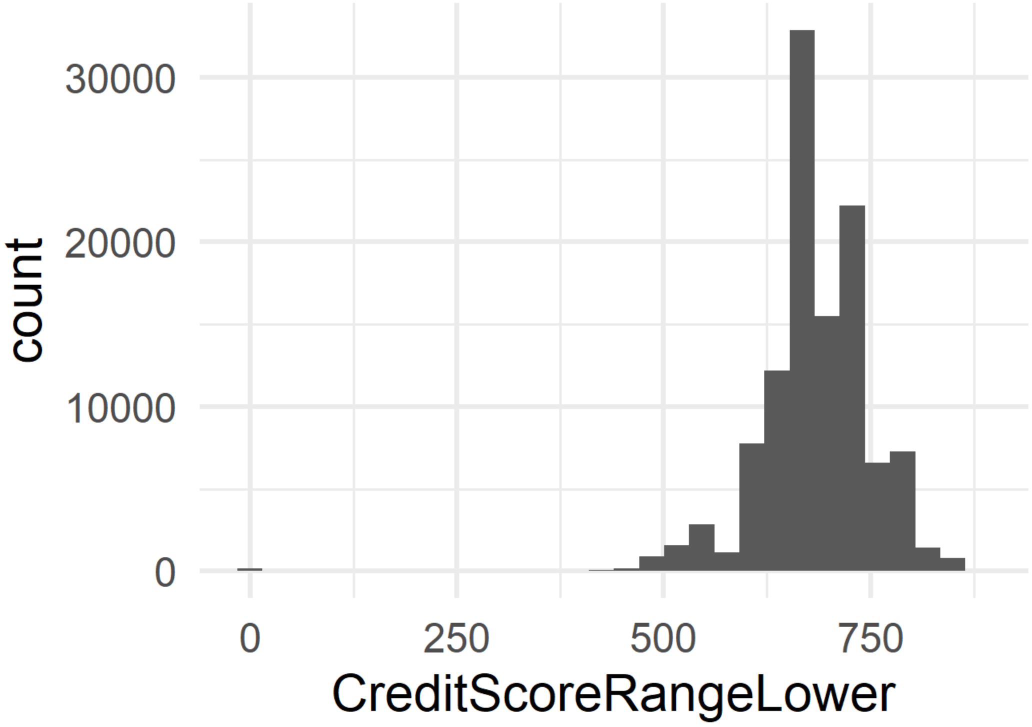
> Most loans have fewr investors, with a diminishing number having increasing number of investors.

Occupation

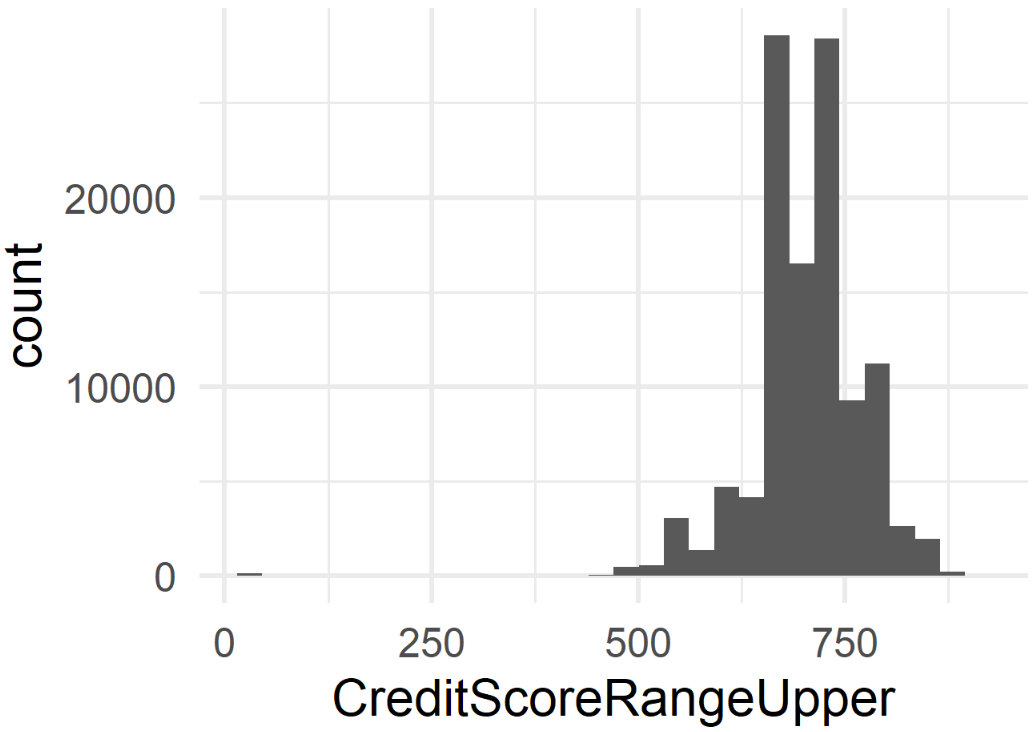


> Most loan recipients describe their occupation as professional, but the largest portion do not have their occupation listed. Teachers, engineers, accountants, and other professional occupations are common as well. I wander of data for other loan sources would show similar distributions for occupation types.

Lower Credit Score Range



Upper Credit Score Range



> Both lower and upper credit score ranges follow a general normal distribution. Both have dips in their peak regions though.

Univariate Analysis

What is the structure of your dataset?

> There are 84 variables and 113,937 observations in the Prosper Loan data set. I created a new data frame called “loan” to only include 18 of those variables. Each one is explored in a plot above.

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

> The non-categorical variables appear follow normal or Poisson distributions. Most findings seem resonable, such as most loans being current, more populous states having greater portion of loans, most recipients being employed, middle income ranges for most recipients, and credit scores that are fairly good around 700.

What other features in the dataset do you think will help support your

> I think it would be interesting to further compare upper and lower credit scores. Side by side box plots can make the comparison more visible, while comparing the standard deviations of the 2 can provide more quantifiable comparison.

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

> Listing category was originally had the awkward name “ProsperLoan$ListingCategory..numeric.” I reassigned that to the new variable ListingCategory. The IncomeRange variable was also re-ordered.

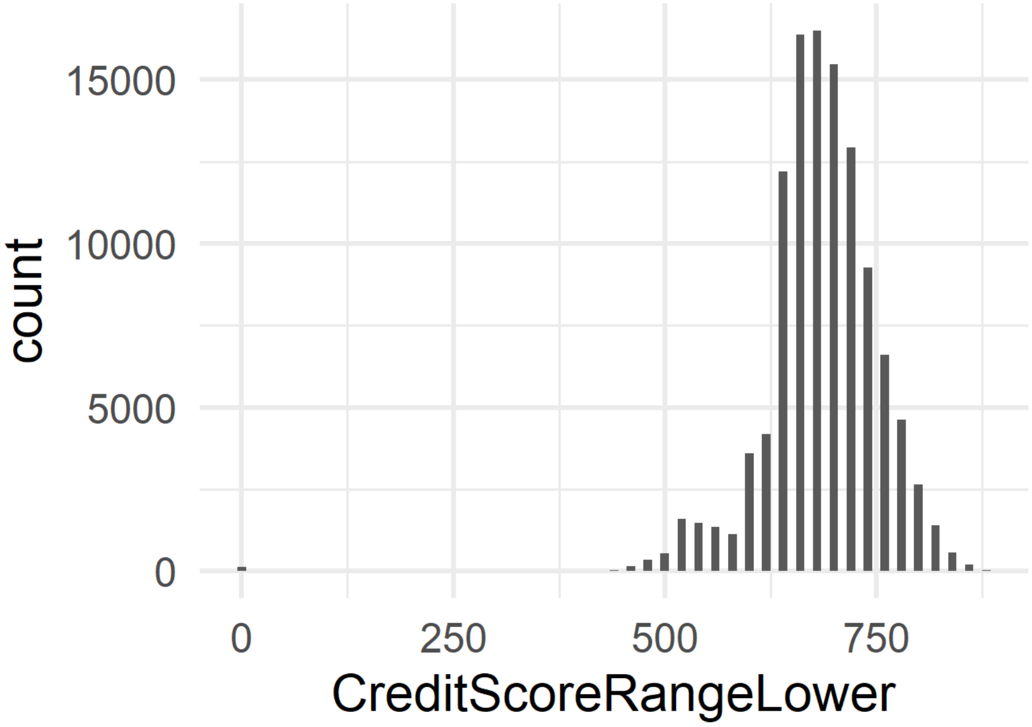
Of the features you investigated, were there any unusual distributions?

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?

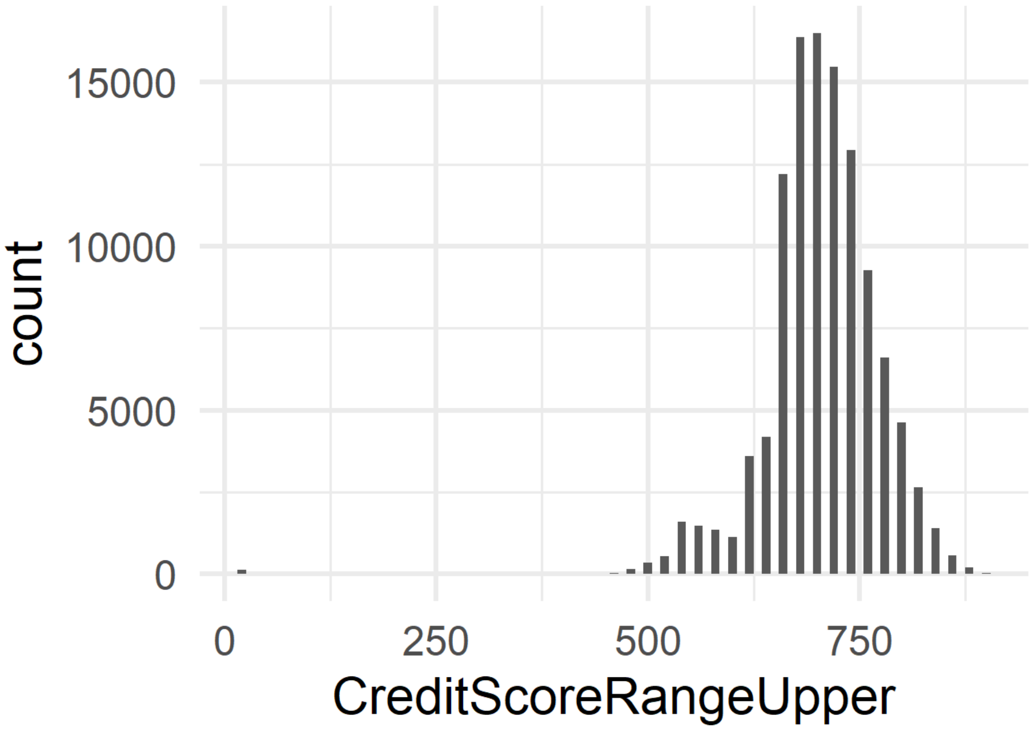
> The listing category variable had a numeric key, so I used the description in the Variable Definitions document to create a descriptive factor varaible for this array. The x-axis variables did not appear in a very legible form, so for many plots I had to adjust the format using the ggplot theme layer.

> By adjusting the binwidth the peaks observed in the credit score distributions become irrelevant. However, there is still an increase in number of loans for lower credit scores that drops off before increasing again.

Lower Credit Score Range (adjusted binwidth = 10)

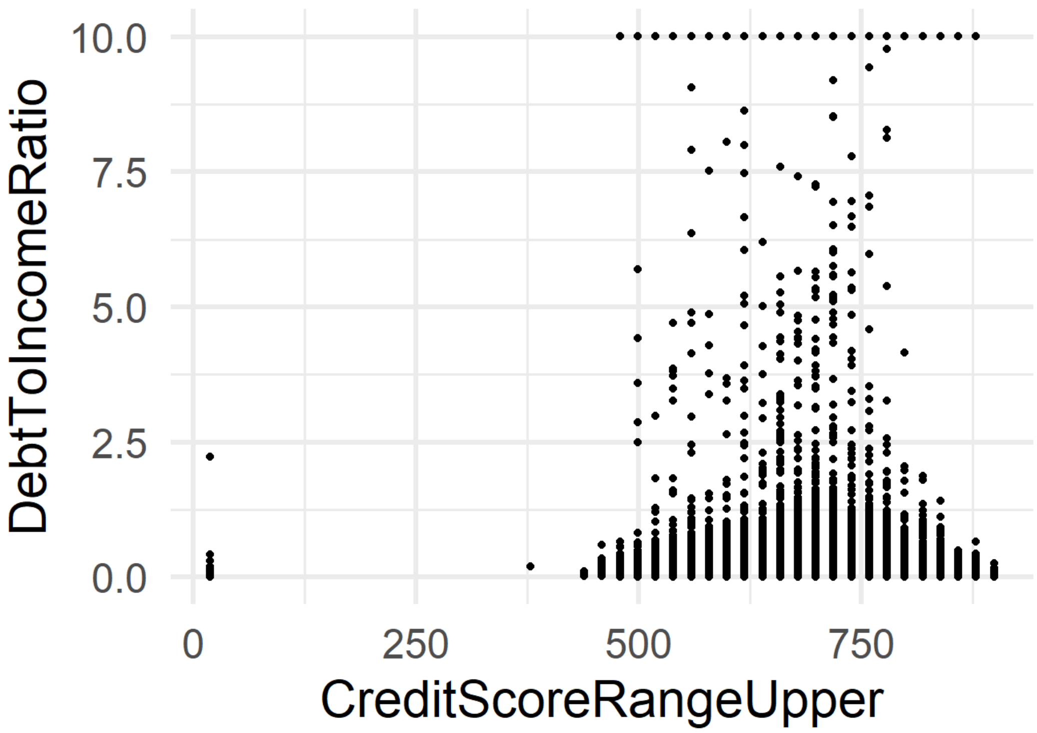


Upper Credit Score Range (adjusted binwidth = 10)

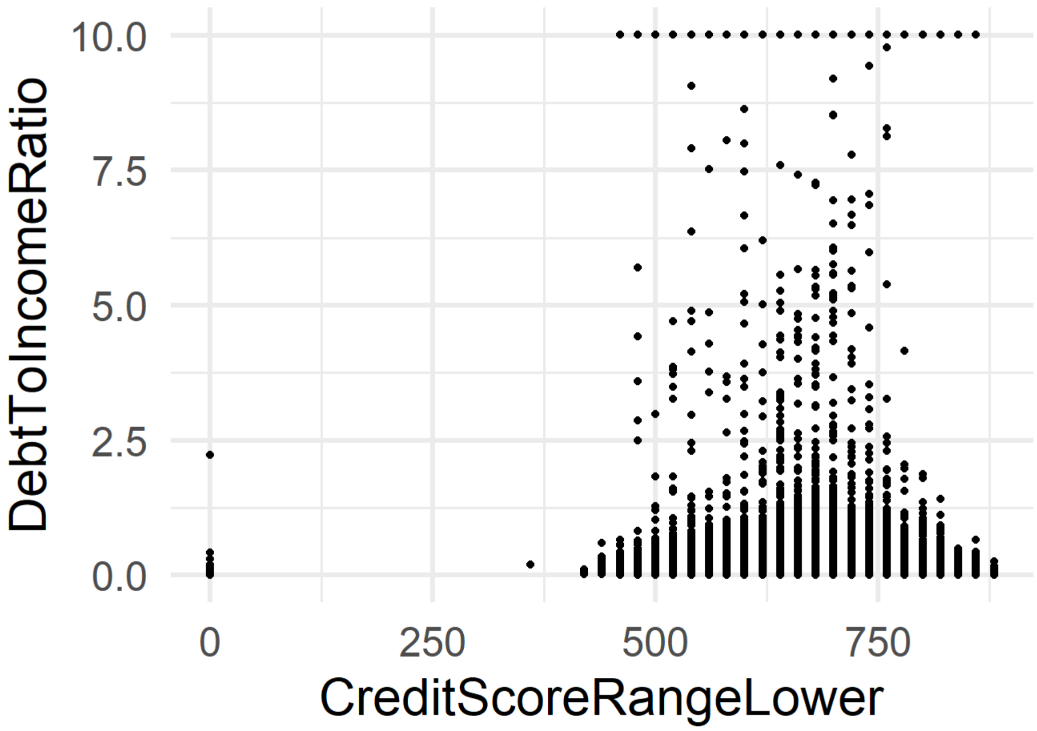


Bivariate Plots Section

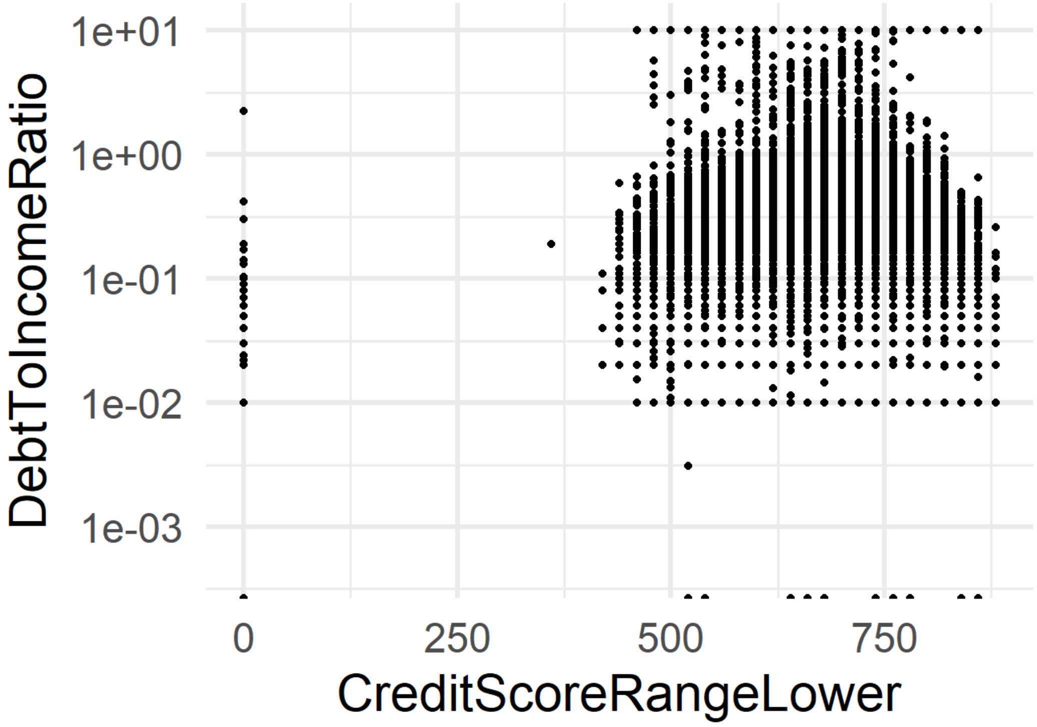
> Debt to Inome Ratio vs. Upper Credit Score Range



> Debt to Inome Ratio vs. Lower Credit Score Range



> Debt to Inome Ratio vs. log of Lower Credit Score Range

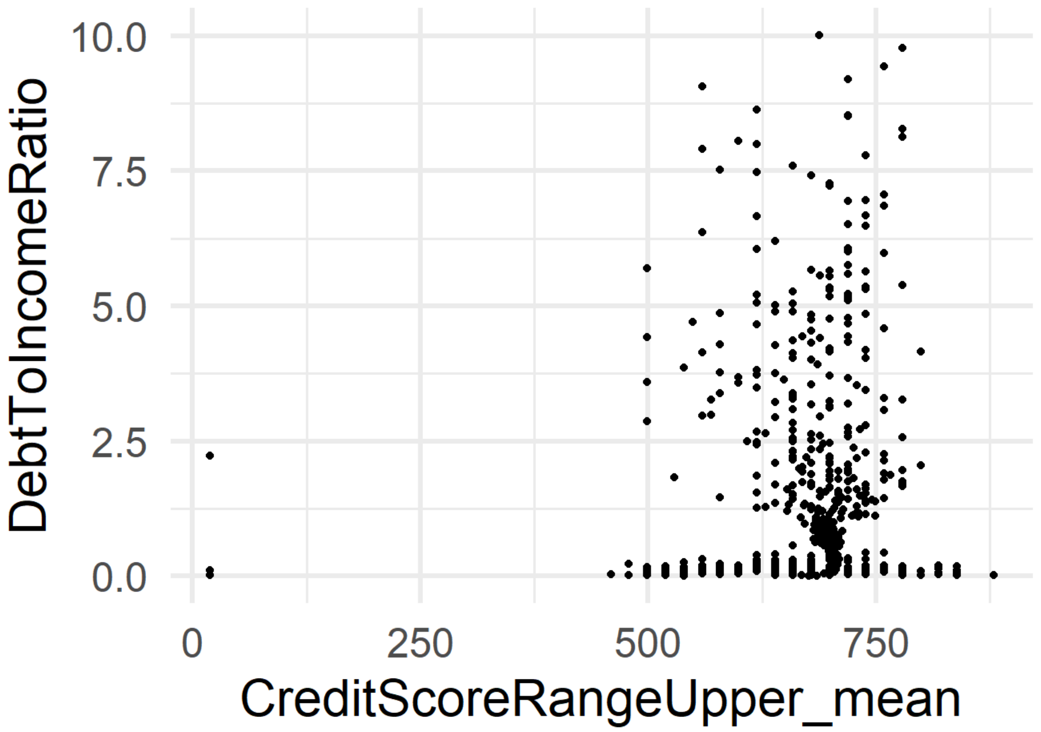


```
##
##  Pearson's product-moment correlation
##
## data:  loan$CreditScoreRangeLower and log(loan$DebtToIncomeRatio + 1)
## t = -0.7235, df = 104796, p-value = 0.4694
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.008289246  0.003819533
## sample estimates:
##          cor
## -0.002234938
```

> With an r of -0.002 there is hardly a correlation.

> Plotting mean of upper credit score range vs. DTI

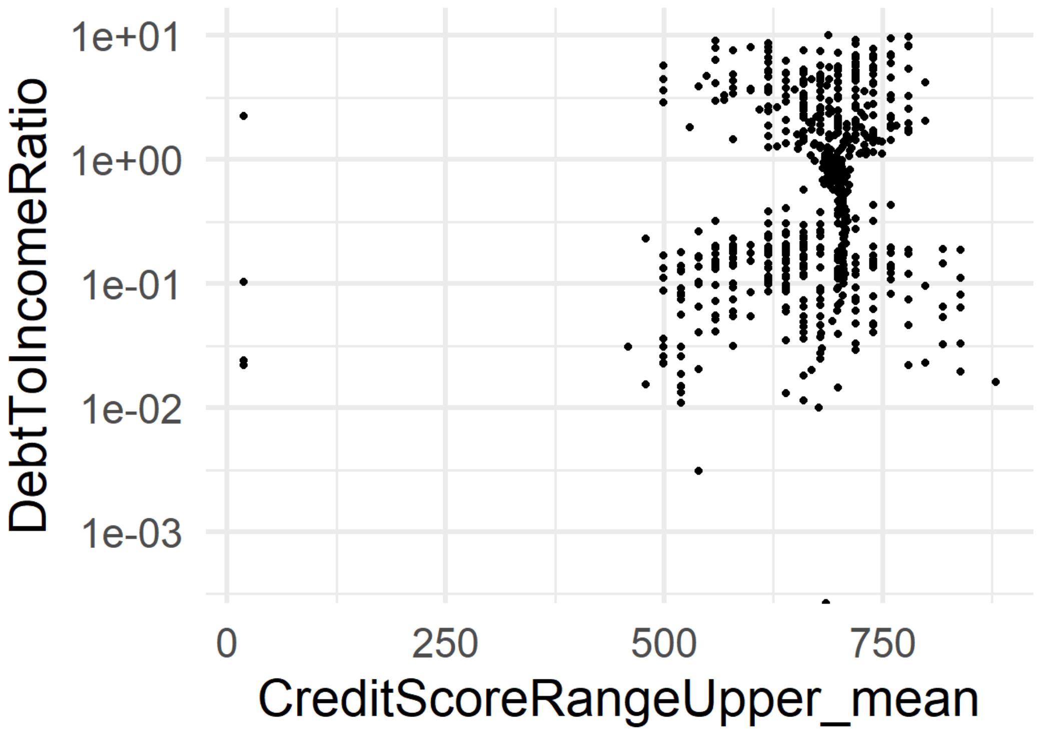
```
## # A tibble: 6 x 3
##   DebtToIncomeRatio CreditScoreRangeUpper_mean     n
##               <dbl>                      <dbl> <int>
## 1           0                             685.    19
## 2           0.00044                        NA      1
## 3           0.0031                        539      1
## 4           0.00611                        NA      1
## 5           0.00647                        NA      1
## 6           0.00677                        NA      1
```



```
##
##  Pearson's product-moment correlation
##
## data:  log(loan.crsc_by_dti$CreditScoreRangeUpper_mean) and loan.crsc_by_dti$DebtToIncomeRatio
## t = 1.9022, df = 623, p-value = 0.05761
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.002451367  0.153500471
## sample estimates:
##        cor
## 0.07598925
```

> Even when using the mean of the Upper Credit Score Ranges there still isn’t much of a correlation with an r value of 0.08.

> Plotting mean of upper credit score range vs. log of DTI



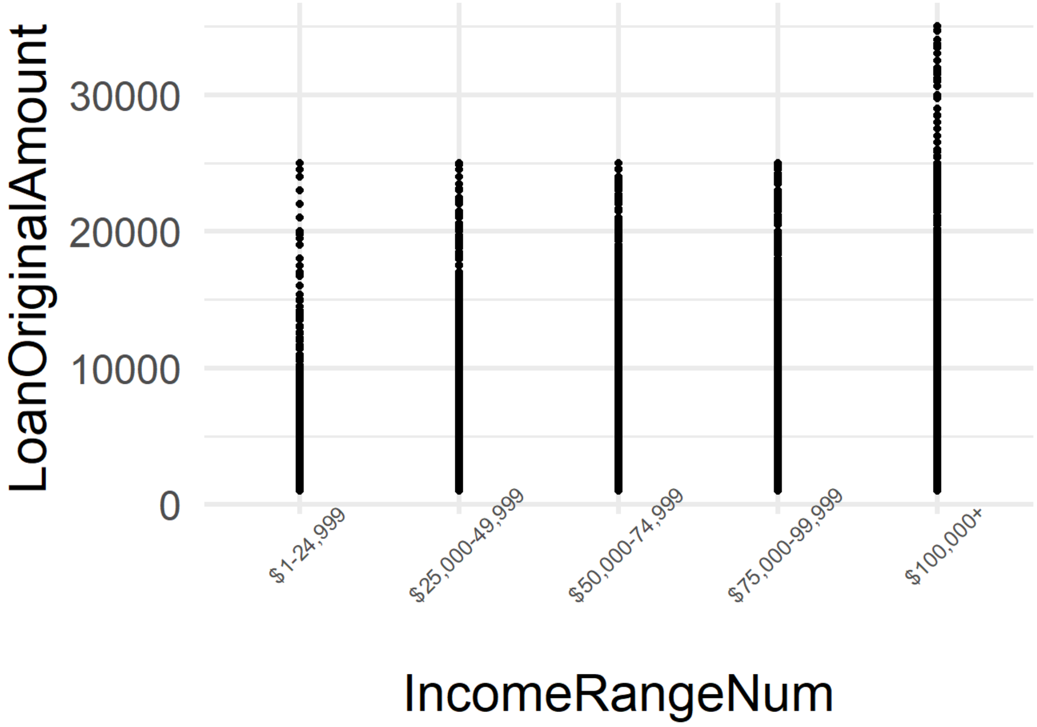
```
##
##  Pearson's product-moment correlation
##
## data:  loan.crsc_by_dti$CreditScoreRangeUpper_mean and log(loan.crsc_by_dti$DebtToIncomeRatio + 1)
## t = 3.7907, df = 623, p-value = 0.0001649
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.07257642 0.22591357
## sample estimates:
##       cor
## 0.1501478
```

> With an r value of 0.15 there still isn’t a good correlation between the log of DTI and mean upper credit score.

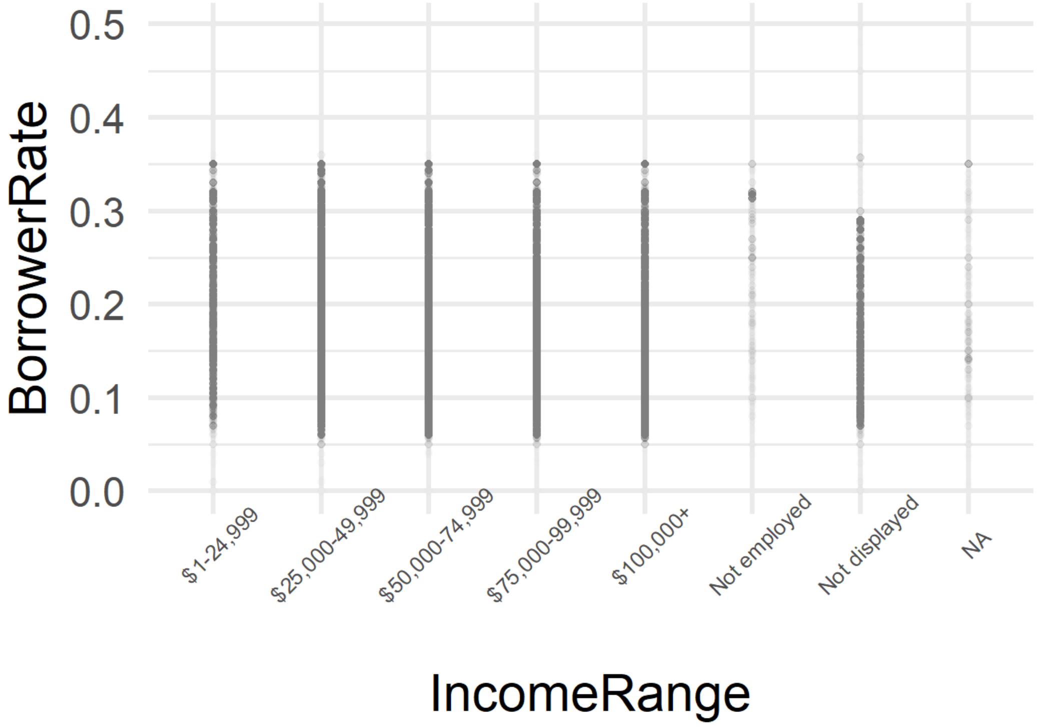
> I thought there might be a relationship between credit score and debt to income ratio, but there doesn’t seem to be. Plotting the log of DTI doesn’t seem to change that much, although the correlation coefficient is higher but still not very good. Plotting the average upper credit score range vs. DTI did not lead to better results. The outlier in DTI is noticable, as are higher concentrations of lower DTIs.

> Dropping Not employed and not displayed categories from IncomeRange in order to plot only numeric ranges.

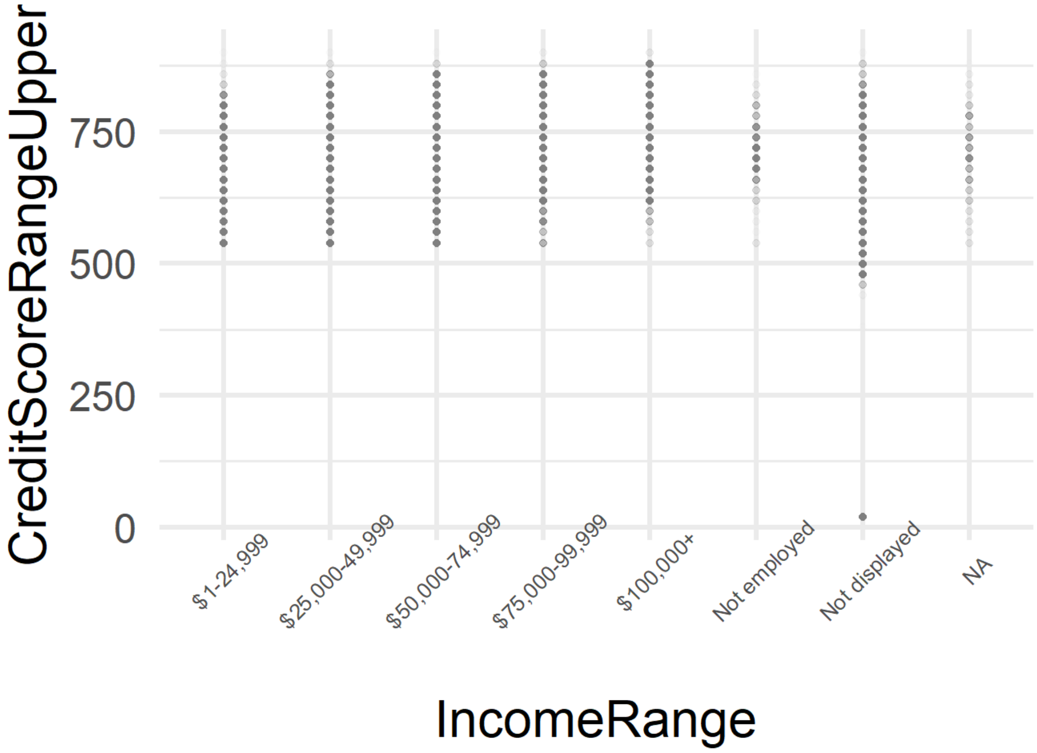
```
loan$IncomeRangeNum <- ordered(loan$IncomeRange, levels = c('$0', '$1-24,999', '$25,000-49,999',
'$50,000-74,999', '$75,000-99,999',
'$100,000+'))
```



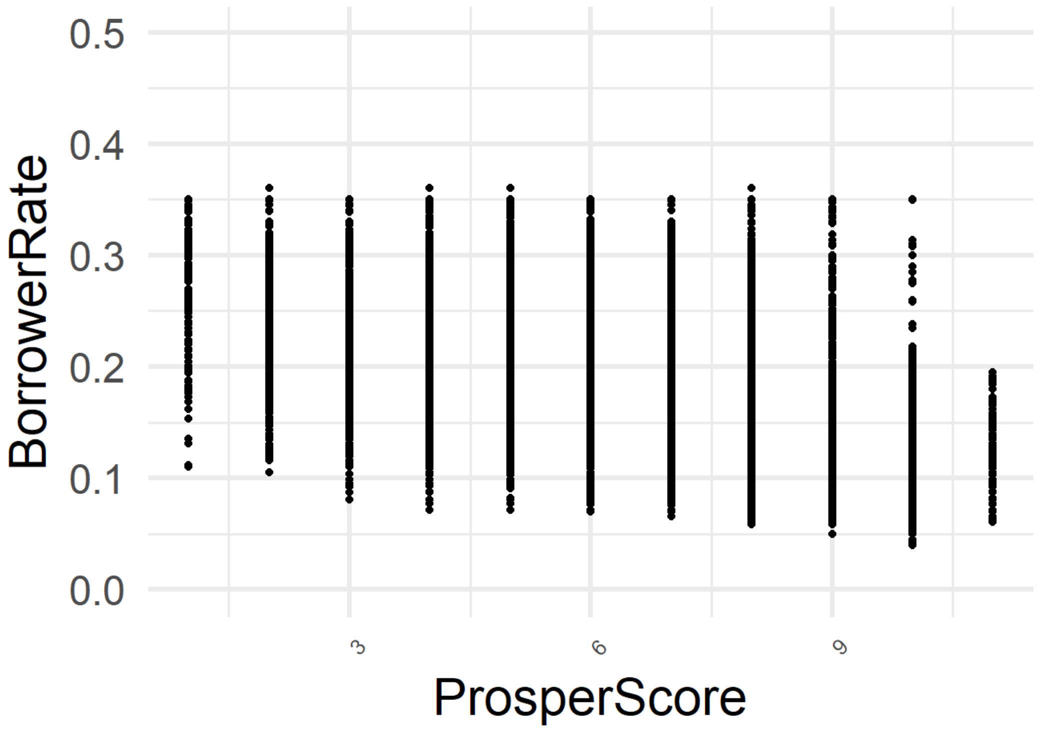
> Those who make 100k/yr or more take out a broader range of loan amounts, some higher than in other income ranges. Those who make more money would probably qualify for higher loan amounts. Those who are not employed tend to have lower loan amounts.



> There doesn’t appear to be much of a trend when comparing income range to borrower rate, even when plotting only 0.5% of the original data points.

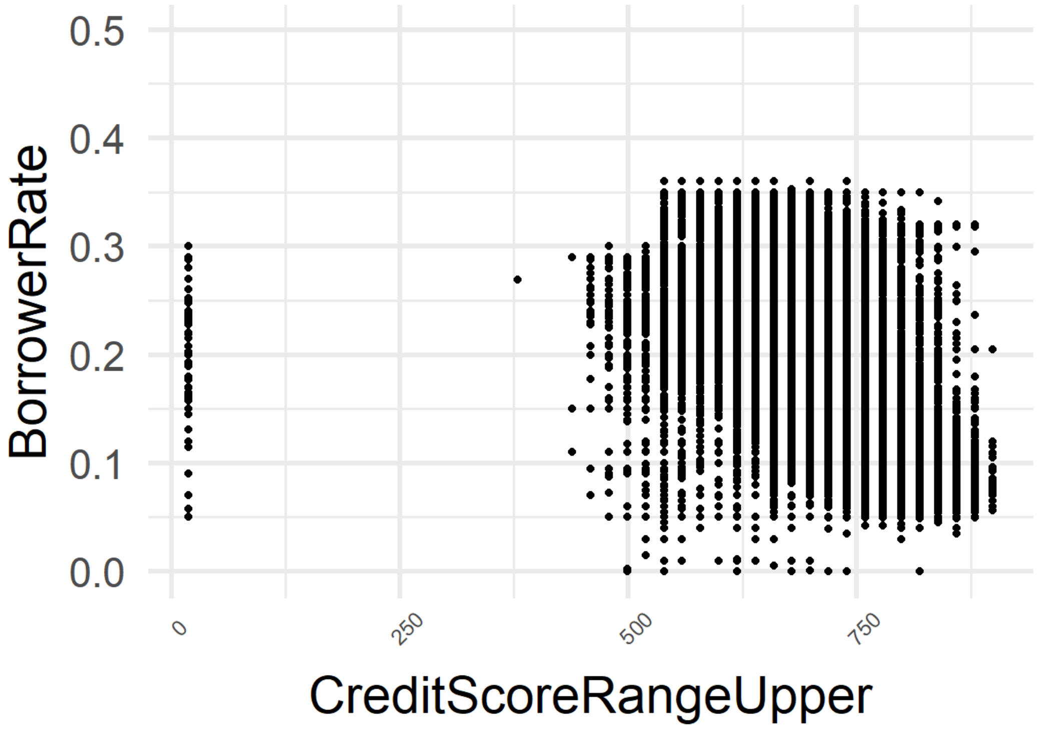


> By limiting the concentration of points plotted there does seem to be a slight increase in credit score as income rises.

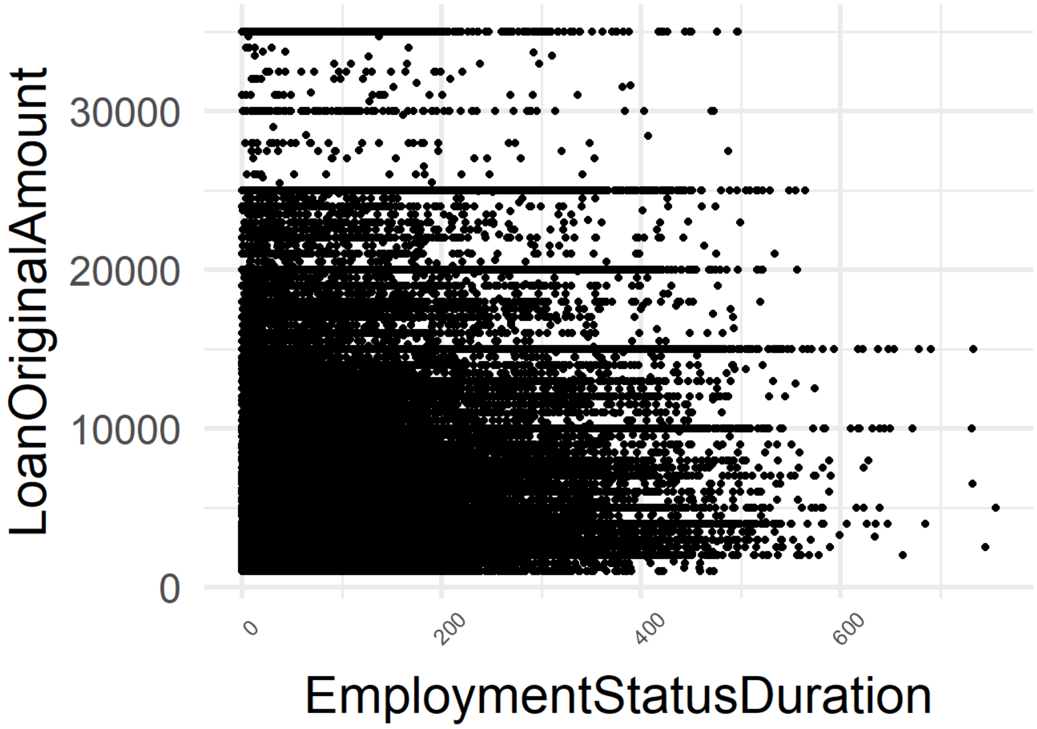


```
##
##  Pearson's product-moment correlation
##
## data:  loan$ProsperScore and loan$BorrowerRate
## t = -248.98, df = 84851, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.6536072 -0.6458311
## sample estimates:
##       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

> There does seem to be a relationship between Prosper Score and borrower rate. The lower the Prosper Score the higher the borrower rate is. However there is only an r correlation coefficient of -0.65 which isn’t very strong.

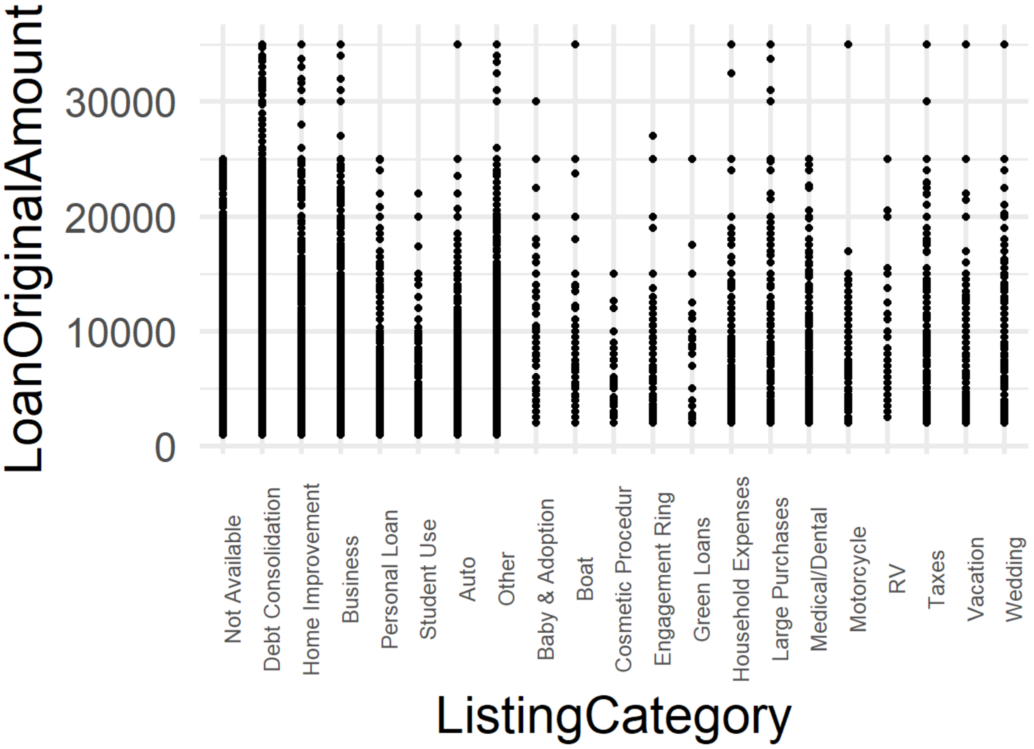


> There appears to be somewhat of a trend for lower borrower rates for higher credit scores. I wander if those outliers with near 0 upper credit scores obtained thier rates using some other factor in the Prosper Score.



```
##
##  Pearson's product-moment correlation
##
## data:  loan$EmploymentStatusDuration and loan$LoanOriginalAmount
## t = 32.157, df = 106310, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.09219259 0.10409908
## sample estimates:
##        cor
## 0.09814935
```

> It appears that those who have been employed longer are taking out lower amounts for loans in general. Correlation is very weak though.



> Debt consolidation loans occur more frequently in higher amounts than others, while those for cosmetic procedures occur in lower amounts than other types.

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?

> I was expecting to see a relationship between income and credit score, and credit score and borrower interest rates. I thought a relationship between credit score and DTI would stand out, but the results make sense in that regardless of how much debt one has some are better at making timely payments than others. It also made sense that those with lower incomes tended to take smaller loan amounts. I was expecting to see a noticable trend in income range vs. borrower rate, but people with lower incomes can still have good credit and vice versa so I shouldn’t be surprised.

Did you observe any interesting relationships between the other features

(not the main feature(s) of interest)?

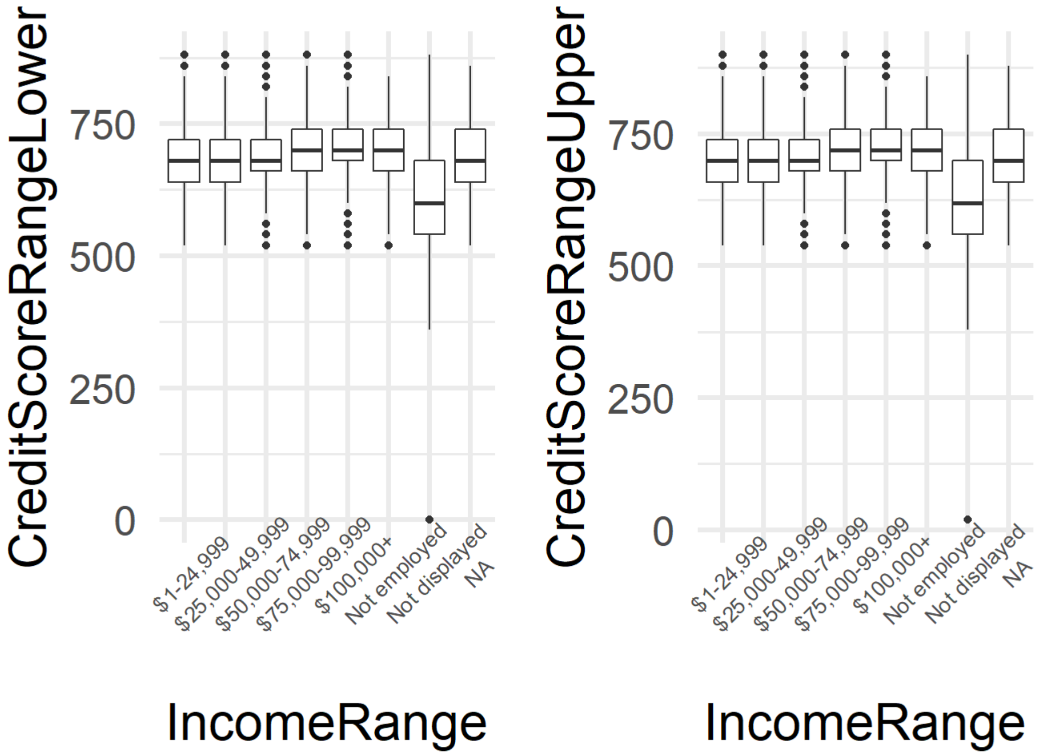
> The outliers in the credit score vs borrower rate plot threw me off. Some loan recipients have extremely low credit scores. There must be some other criteria that would explain the wide range in borrower rates. Perhaps the credit scores of co-signers are not included here, or maybe there are some other criteria in Prosper Score that determines these rates.

What was the strongest relationship you found?

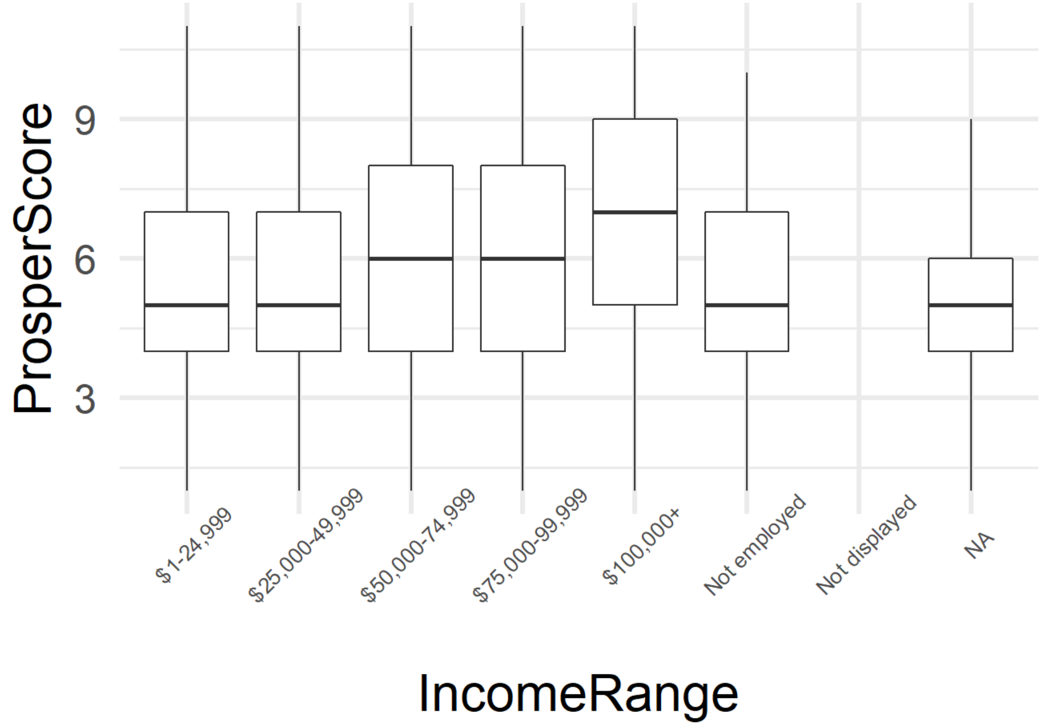
> The strongest relationship is between Prosper Score and Borrower Rate, as the absolute value of r is the closest to 1. It would’ve been nice if incomes were continuous, since it looks like there may be some correlation in those plots.

Multivariate Plots Section

> Comparing box plot of lower and upper credit score ranges.

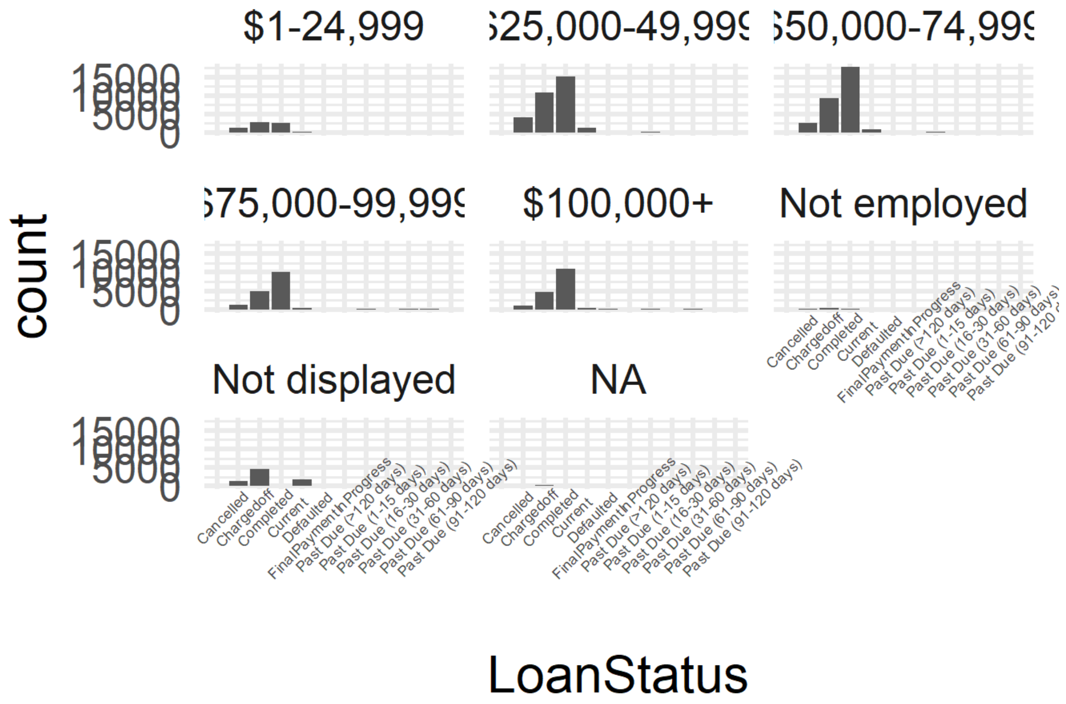


> The IQR spread for lower and upper credit score ranges are very close to each other. The median credit score generally rises with income, with the greatest IQR in the middle income range of $50,000-$70,000.



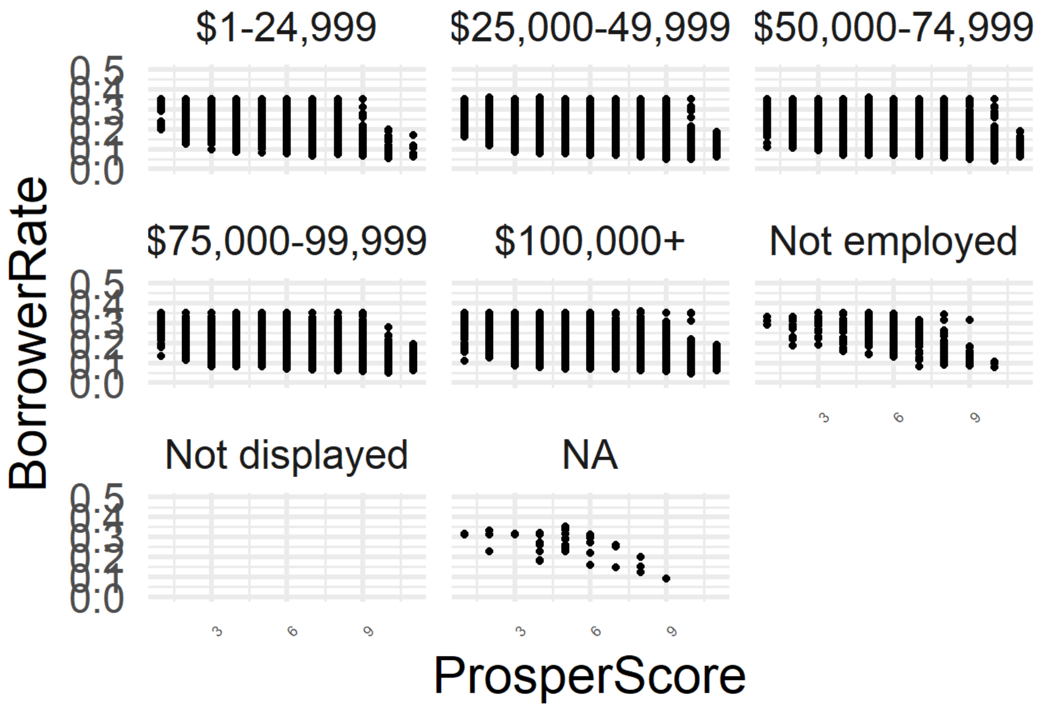
> The trend for Prosper Score rising with income is more noticable. Perhaps Prosper weighs income more heavily in their score than in credit score. The scale for Prosper Score is much narrower, so it’s not wise to jump to this conclusion just yet.

> I want to see if there are any trends regarding loan status in relation to income range.



```
## loan$IncomeRange: $0
## NULL
## --------------------------------------------------------
## loan$IncomeRange: $1-24,999
##              Cancelled             Chargedoff              Completed
##                      0                   1329                   2908
##                Current              Defaulted FinalPaymentInProgress
##                   2536                    334                     14
##   Past Due (>120 days)   Past Due (1-15 days)  Past Due (16-30 days) 
##                      0                     52                     20 
##  Past Due (31-60 days)  Past Due (61-90 days) Past Due (91-120 days) 
##                     32                     22                     27 
## -------------------------------------------------------- 
## loan$IncomeRange: $25,000-49,999
##              Cancelled             Chargedoff              Completed 
##                      1                   4162                  10891 
##                Current              Defaulted FinalPaymentInProgress 
##                  15111                   1290                     49 
##   Past Due (>120 days)   Past Due (1-15 days)  Past Due (16-30 days) 
##                      7                    266                     80 
##  Past Due (31-60 days)  Past Due (61-90 days) Past Due (91-120 days) 
##                    119                    102                    114 
## -------------------------------------------------------- 
## loan$IncomeRange: $50,000-74,999
##              Cancelled             Chargedoff              Completed 
##                      0                   2633                   9282 
##                Current              Defaulted FinalPaymentInProgress 
##                  17615                    874                     64 
##   Past Due (>120 days)   Past Due (1-15 days)  Past Due (16-30 days) 
##                      7                    212                     83 
##  Past Due (31-60 days)  Past Due (61-90 days) Past Due (91-120 days) 
##                     89                     92                     99 
## -------------------------------------------------------- 
## loan$IncomeRange: $75,000-99,999
##              Cancelled             Chargedoff              Completed 
##                      0                   1153                   4914 
##                Current              Defaulted FinalPaymentInProgress 
##                  10139                    375                     26 
##   Past Due (>120 days)   Past Due (1-15 days)  Past Due (16-30 days) 
##                      1                    131                     37 
##  Past Due (31-60 days)  Past Due (61-90 days) Past Due (91-120 days) 
##                     55                     52                     33 
## -------------------------------------------------------- 
## loan$IncomeRange: $100,000+
##              Cancelled             Chargedoff              Completed 
##                      0                    968                   4774 
##                Current              Defaulted FinalPaymentInProgress 
##                  10916                    322                     52 
##   Past Due (>120 days)   Past Due (1-15 days)  Past Due (16-30 days) 
##                      0                    134                     40 
##  Past Due (31-60 days)  Past Due (61-90 days) Past Due (91-120 days) 
##                     65                     38                     28 
## -------------------------------------------------------- 
## loan$IncomeRange: Not employed
##              Cancelled             Chargedoff              Completed 
##                      0                    182                    325 
##                Current              Defaulted FinalPaymentInProgress 
##                    248                     25                      0 
##   Past Due (>120 days)   Past Due (1-15 days)  Past Due (16-30 days) 
##                      0                      9                      5 
##  Past Due (31-60 days)  Past Due (61-90 days) Past Due (91-120 days) 
##                      3                      6                      3 
## -------------------------------------------------------- 
## loan$IncomeRange: Not displayed
##              Cancelled             Chargedoff              Completed 
##                      4                   1380                   4610 
##                Current              Defaulted FinalPaymentInProgress 
##                      0                   1747                      0 
##   Past Due (>120 days)   Past Due (1-15 days)  Past Due (16-30 days) 
##                      0                      0                      0 
##  Past Due (31-60 days)  Past Due (61-90 days) Past Due (91-120 days) 
##                      0                      0                      0
```

> Past due loans are less frequent for higher income ranges.



> Along income ranges the same trend of Borrower Rate generally decreasing with Prosper Score seems to hold.

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?

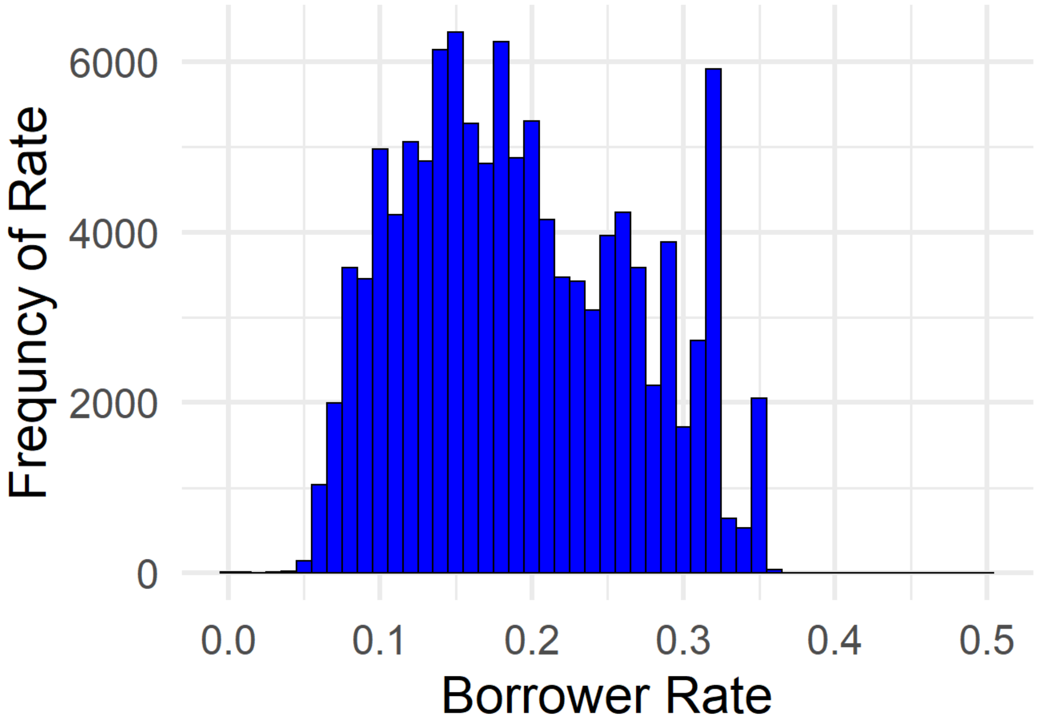
> Prosper Score generally increased with income range, while borrorer rate seemed to decrease as income range rose. The IQR spreads of lower and upper credit scores were very similar.

Were there any interesting or surprising interactions between features?

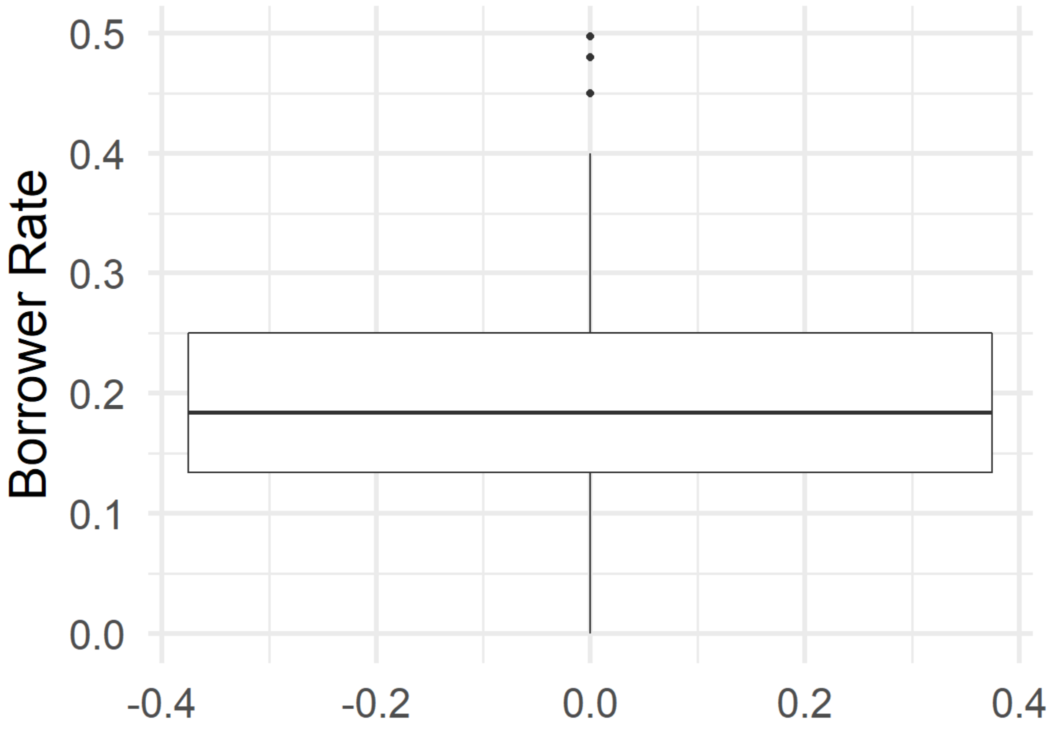
Final Plots and Summary

> Of the variables examined so far, there seems to be a more prevalent relationship between Prosper Score and Borrower Rate. These variables will be examined more closely.

Plot One: Histogram of Borrower Rates



> Box plot of Borrower Rates



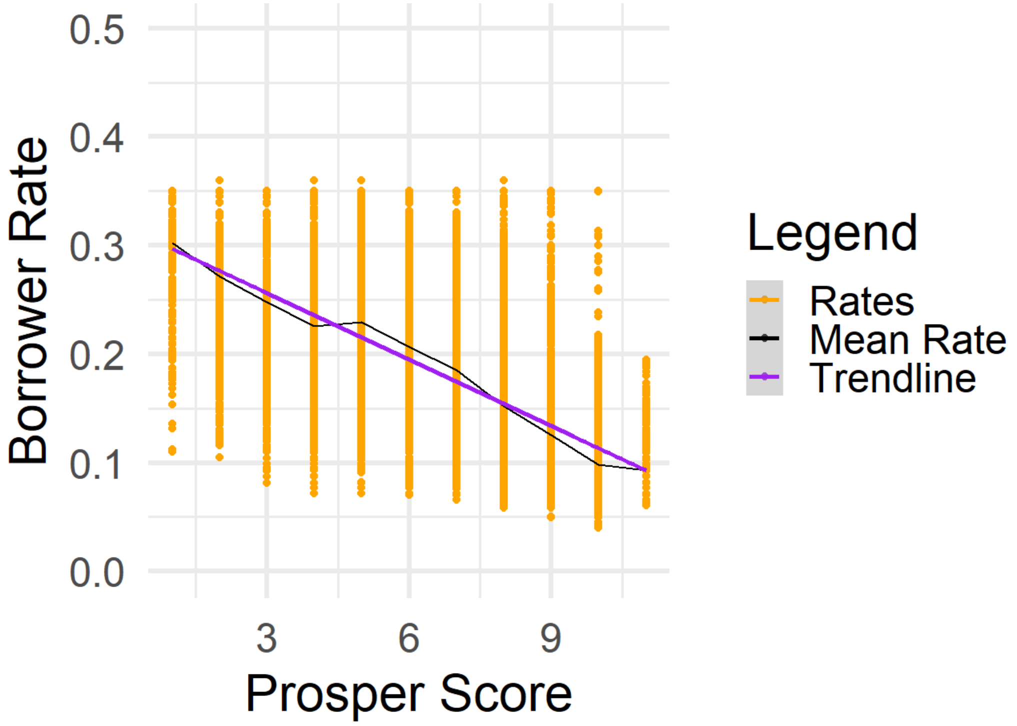
```
summary(loan$BorrowerRate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.1340  0.1840  0.1928  0.2500  0.4975
```

Description One

> Adjusting the binwidth on the histogram for Borrower Rate, we still see the increasing trend from 0.25 to 0.35. This suggests a bi-modal distribution. The range of rates is from 0 to 0.4975, with an IQR of 0.116. There are some outliers above rates of 0.4.

Plot Two: Borrower Rate vs. Prosper Score



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

Description Two

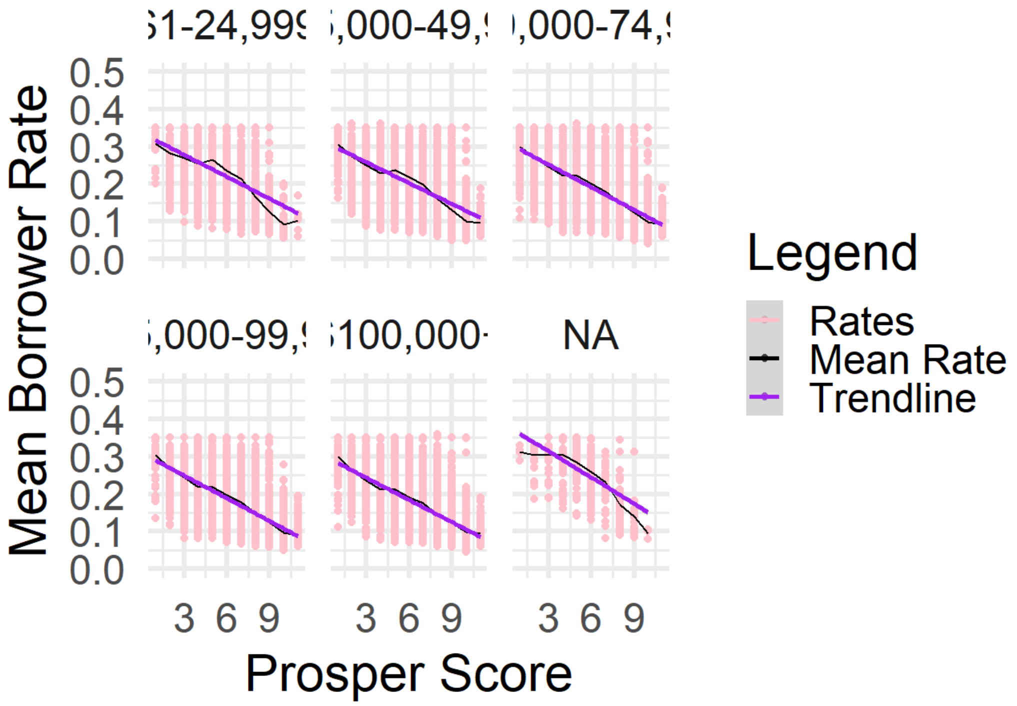
> Here we have Borrorwe Rate vs. Prosper Score, overlayed with mean Borrower Rate, and a trendline. The absolute value of r is 0.65, which isn’t very strong, but stronger than other correlations found in the Bivariate Plots section. With the plots of mean Borrower Score and the trendline, it is easy to see the rates decreasing as Prosper Score increases. This makes sense if higher Prosper Scores reflect loan takers general credit worthiness.

Plot Three: Borrower Rates vs. Prosper Score for each Income Range

```
levels(loan$IncomeRange)
```

```
## [1] "$0"              "$1-24,999"      "$25,000-49,999" "$50,000-74,999"
## [5] "$75,000-99,999" "$100,000+"      "Not employed"   "Not displayed"
```

```
loan$IncomeRangeMon <- ordered(loan$IncomeRange, levels = c('$0', '$1-24,999', '$25,000-49,999',
'$50,000-74,999', '$75,000-99,999',
'$100,000+'))
```



```
##
##  Pearson's product-moment correlation
##
## data:  BorrowerRate and ProsperScore
## t = -46.155, df = 4652, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5798373 -0.5404132
## sample estimates:
##        cor
## -0.5604427
```

```
##
##  Pearson's product-moment correlation
##
## data:  BorrowerRate and ProsperScore
## t = -106.56, df = 24173, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5738542 -0.5566996
## sample estimates:
##       cor
## -0.565338
```

```
##
##  Pearson's product-moment correlation
##
## data:  BorrowerRate and ProsperScore
## t = -133.98, df = 25625, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.6489805 -0.6345803
## sample estimates:
##       cor
## -0.641837
```

```
##
##  Pearson's product-moment correlation
##
## data:  BorrowerRate and ProsperScore
## t = -109.11, df = 14496, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.6803685 -0.6624915
## sample estimates:
##        cor
## -0.6715277
```

```
##
##  Pearson's product-moment correlation
##
## data:  BorrowerRate and ProsperScore
## t = -120.14, df = 15203, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7059275 -0.6896176
## sample estimates:
##       cor
## -0.697863
```

```
r <- c(-0.56, -0.56, -0.64, -0.67, -0.7)
income <- c('$1-24,999', '$25,000-49,999', '$50,000-74,999', '$75,000-99,999', '$100,000+')
corr.df <- data.frame('IncomeRange' = income, 'r' = r)

corr.df
```

```
##      IncomeRange     r
## 1      $1-24,999 -0.56
## 2 $25,000-49,999 -0.56
## 3 $50,000-74,999 -0.64
## 4 $75,000-99,999 -0.67
## 5      $100,000+ -0.70
```

Description Three

> Here I plotted Brorrower Rate vs. Prosper Score for each non-zero income range. The patterns are similar in that the rate decreases as score increases. Looking at Pearson's correlation coefficient for each, we see that the correlation becomes stronger as income range rises. Finally in the $100,000+ range the correlation of -.7 is strong enough to say that there is a correlation. It would seem that whichever other factors that go into determining the borrower rate don't play as much of a role for higher earning loan takers.

Reflection

> I ended up looking at this more from a consumer stand point, examining aspects such as credit score, income range, loan amount, employment status, and interest rates. I was hoping to find a relationship between credit score and interest rate, but ended up finding better correlation using Prosper Score instead. These correlations weren't very strong though.

> Out of the 82 variables in this data set, I only looked at 18. Perhaps other relationships exist among such variables as Estimated Return, LenderYield, IsBorrowerHomeowner, OpenCreditLines that weren't included in my subset. Expanding on the analysis already performed I think that the IsBorrowerHomeowner variable would've been a good one to look at in terms of credit scores and interest rates.