Prosper Loan Analysis by Travis Dickey

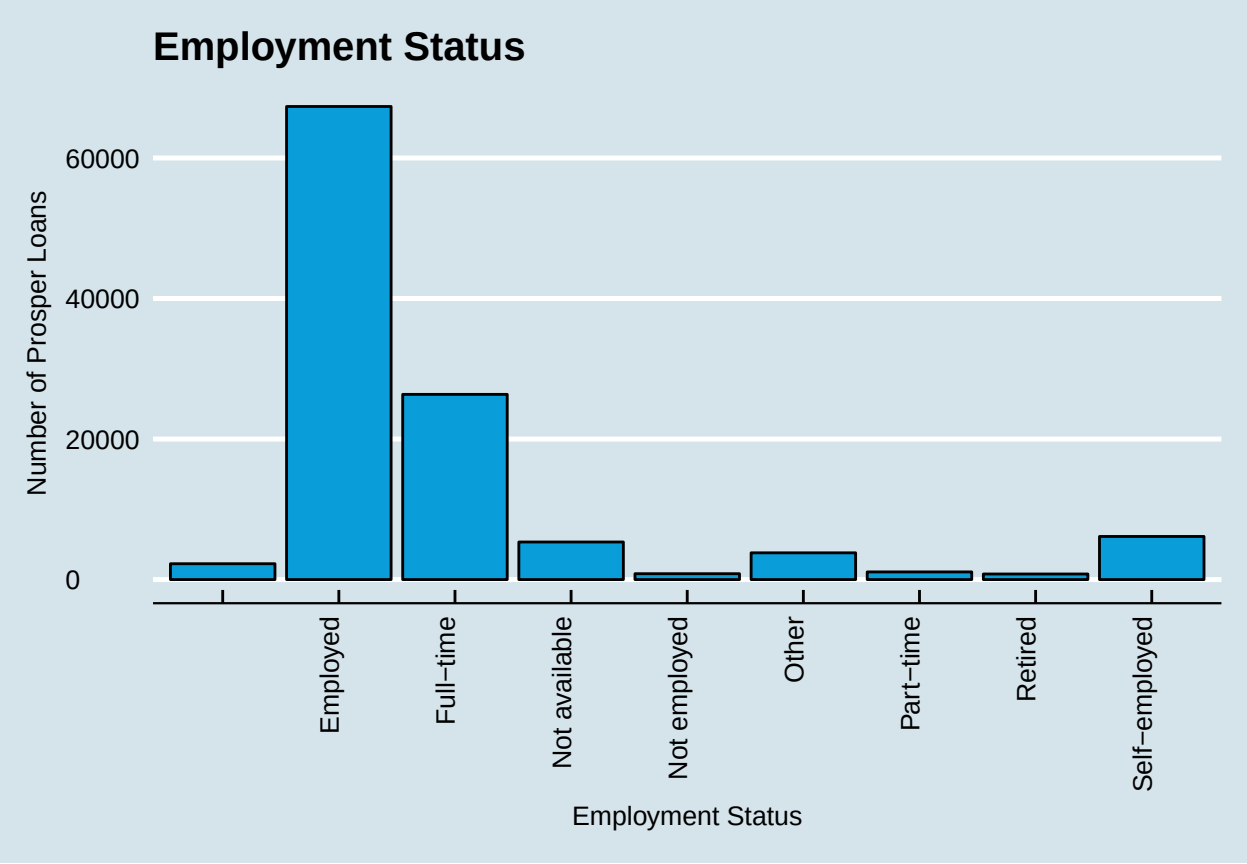
Prosper Loan Data: The data set contains 113,937 loans with 81 variables on each loan, including loan amount, borrower rate (or interest rate), current loan status, borrower income, borrower employment status, borrower credit history, and the latest payment information.

Univariate Plots Section

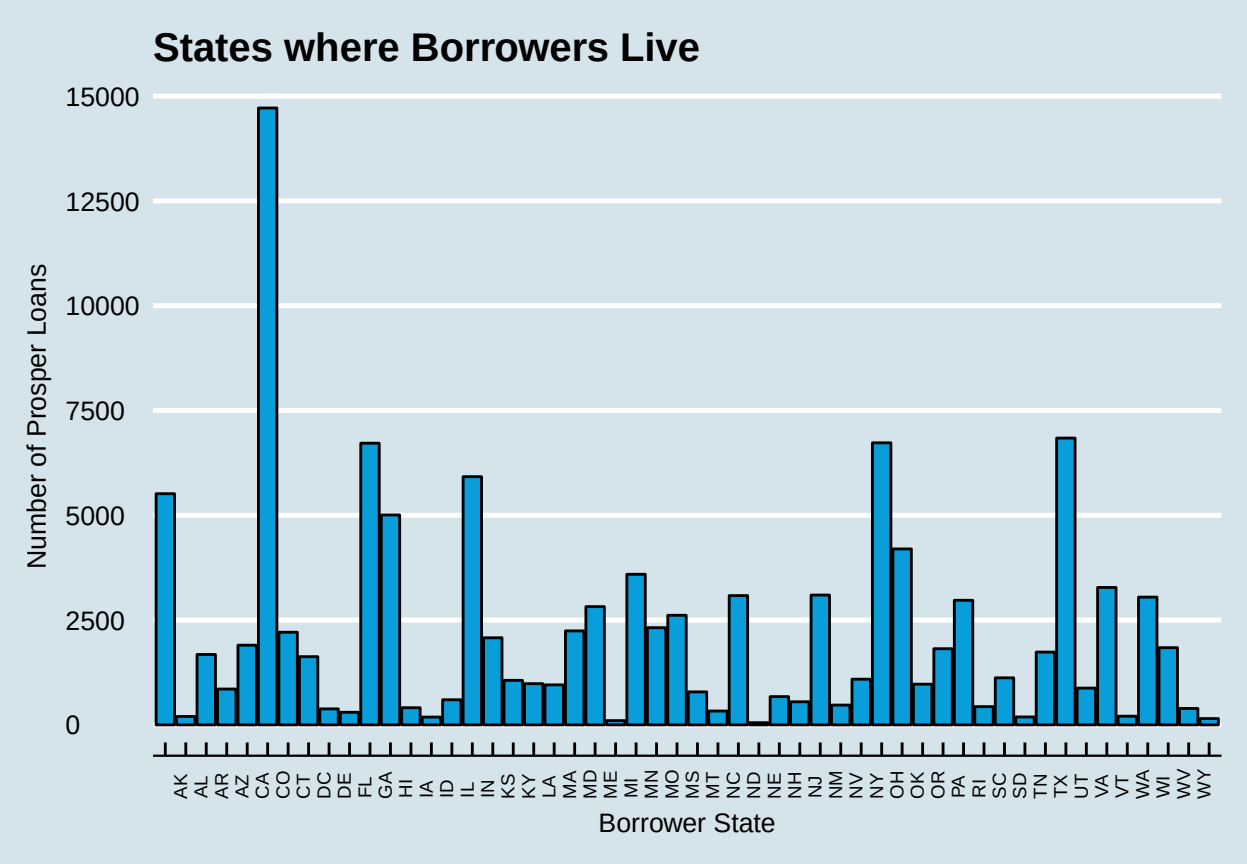
```
## [1] 113937     81
```

```
## 'data.frame':    113937 obs. of  10 variables:
##  $ CreditGrade            : Factor w/ 9 levels "","A","AA","B",..: 5 1 8 1 1 1 1 1 1 1 ...
##  $ 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 ...
##  $ ProsperRating..Alpha.  : Factor w/ 8 levels "","A","AA","B",..: 1 2 1 2 6 4 7 5 3 3 ...
##  $ 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 ...
##  $ CreditScoreRangeLower  : int  640 680 480 800 680 740 680 700 820 820 ...
##  $ DelinquenciesLast7Years: int  4 0 0 14 0 0 0 0 0 0 ...
##  $ DebtToIncomeRatio      : num  0.17 0.18 0.06 0.15 0.26 0.36 0.27 0.24 0.25 0.25 ...
##  $ IncomeRange            : Factor w/ 8 levels "$0","$1-24,999",..: 4 5 7 4 3 3 4 4 4 4 ...
```

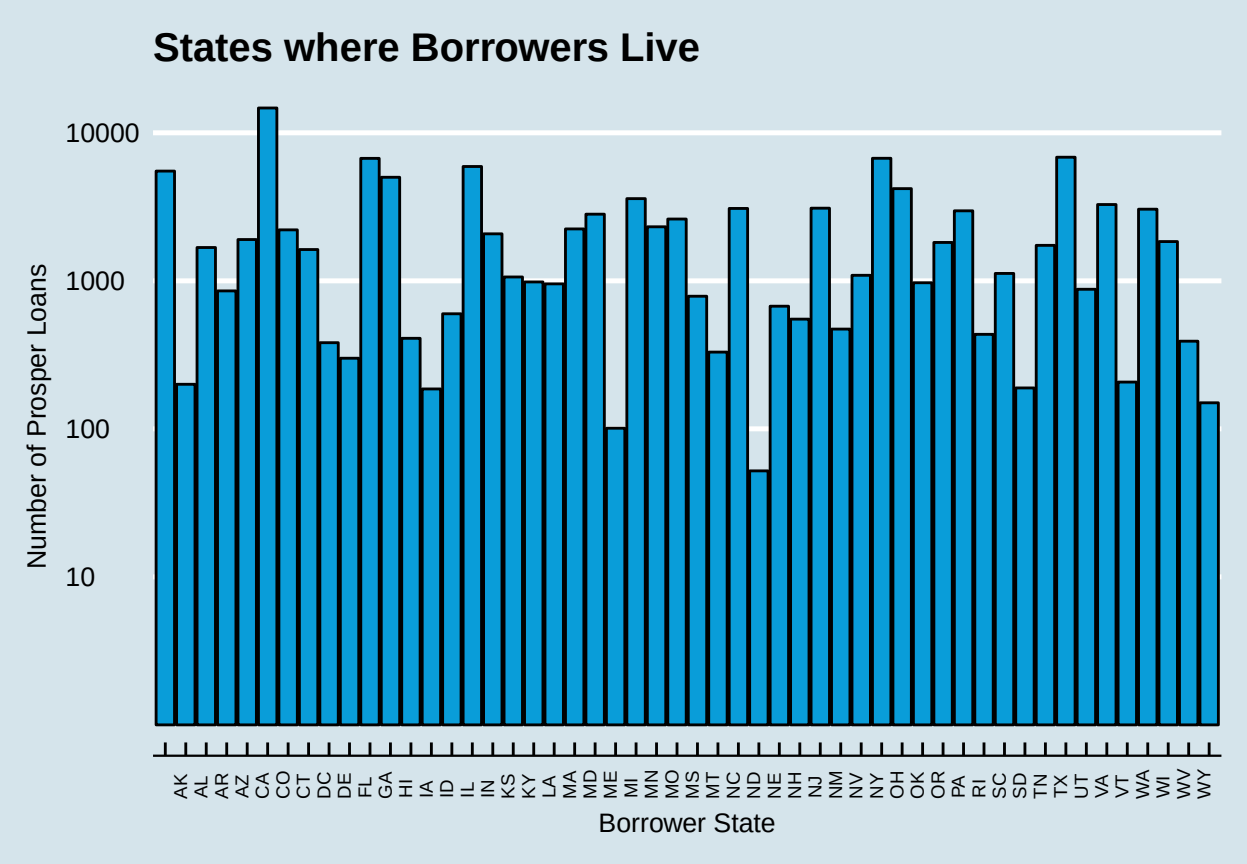
The full dataset contains 81 variables. I will focus on 10 variables. The structure of those variables includes 2 int, 2 num, and 6 factor variables.



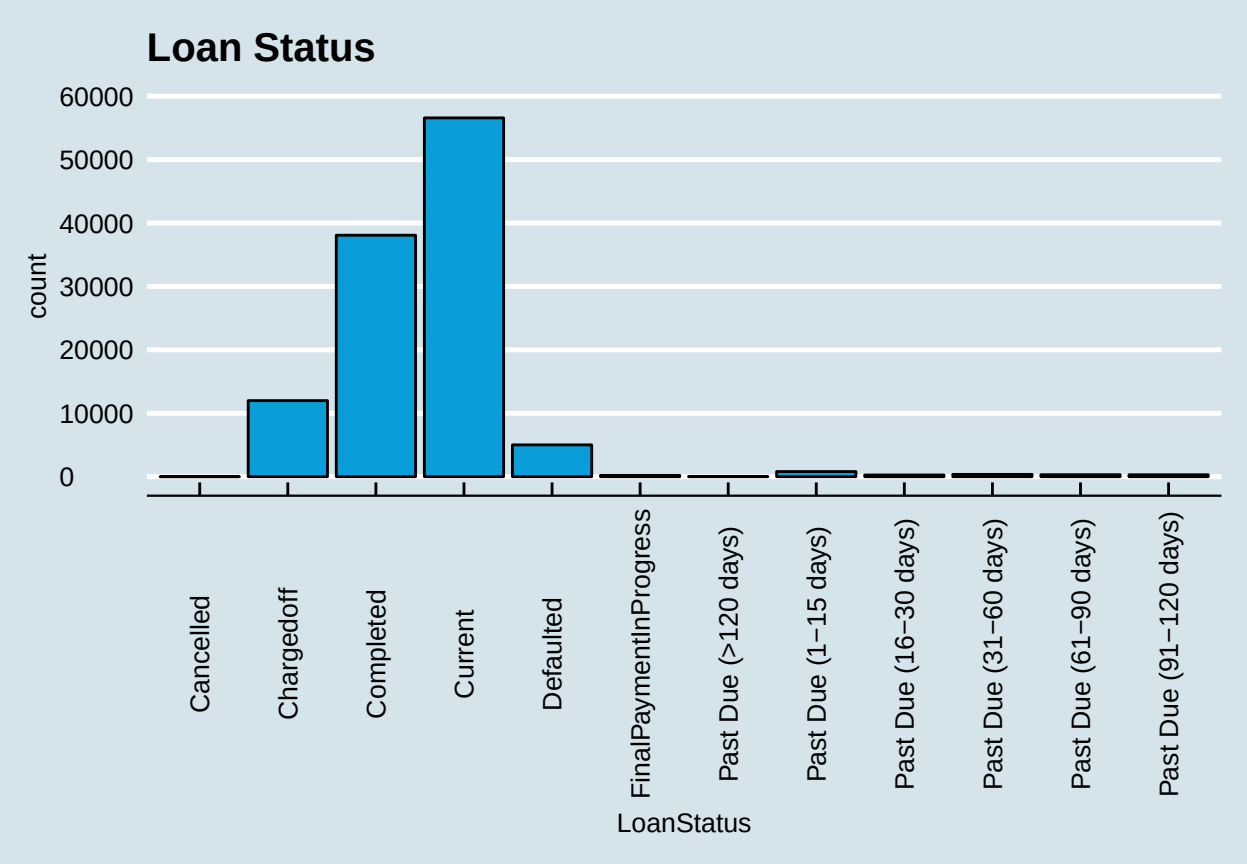
The vast majority of loans are for employed people, which is no surprise.



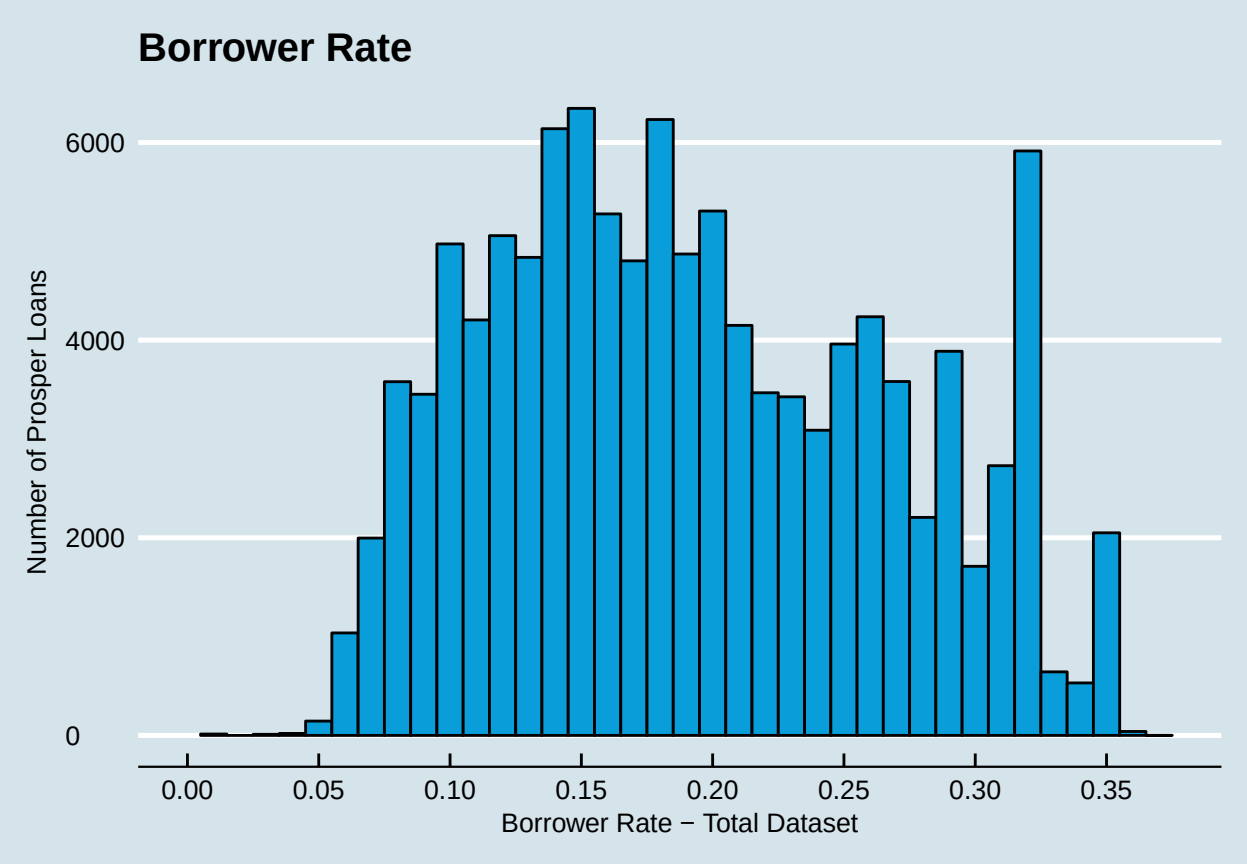
California, at nearly 15,000, has by far the most loan recipients. Florida, New York, and Texas all follow with about 7,000 recipients. It's no surprise that the most populous states have the most loan recipients.



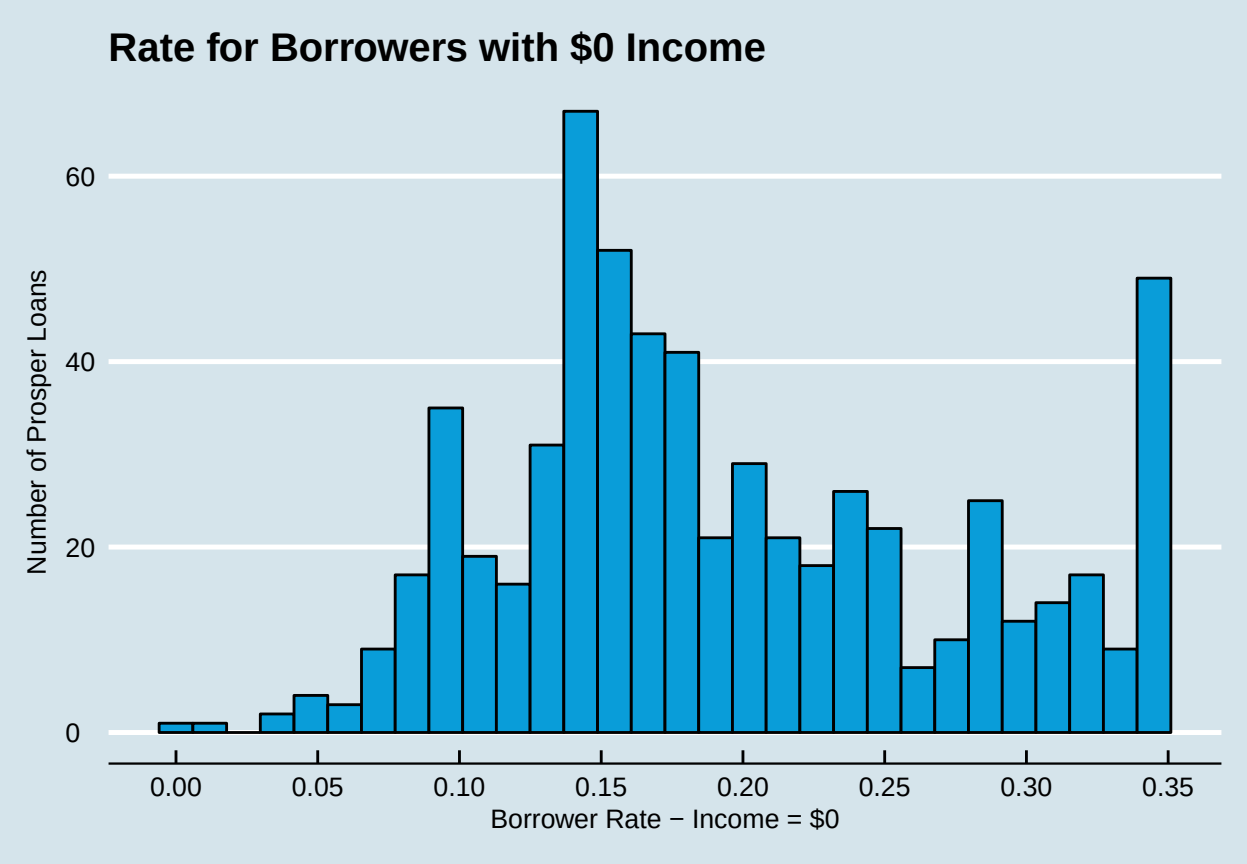
North Dakota has the least number of loan recipients with less than 100. Maine is next with right at 100.



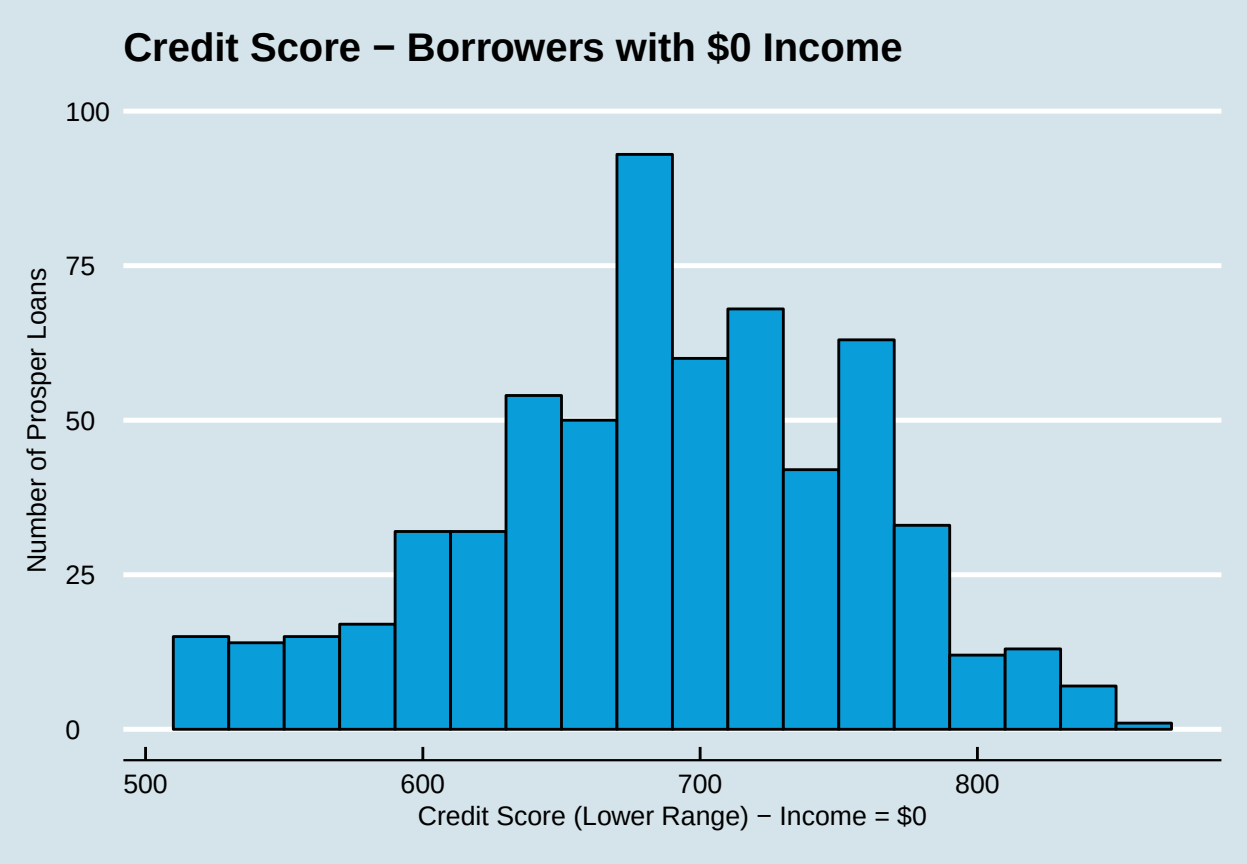
The majority of loans are current or completed, with just over 10,000 charged off and about 5,000 defaulted. I'm curious about the characteristics of the chargeoffs and defaulted loans. Are there common features in the creditworthiness of borrowers who end up not repaying the loan?



Borrower Rate follows a fairly normal distribution with an abnormal jump at about 32%. I wonder what criteria Prosper uses to set the borrower rate.

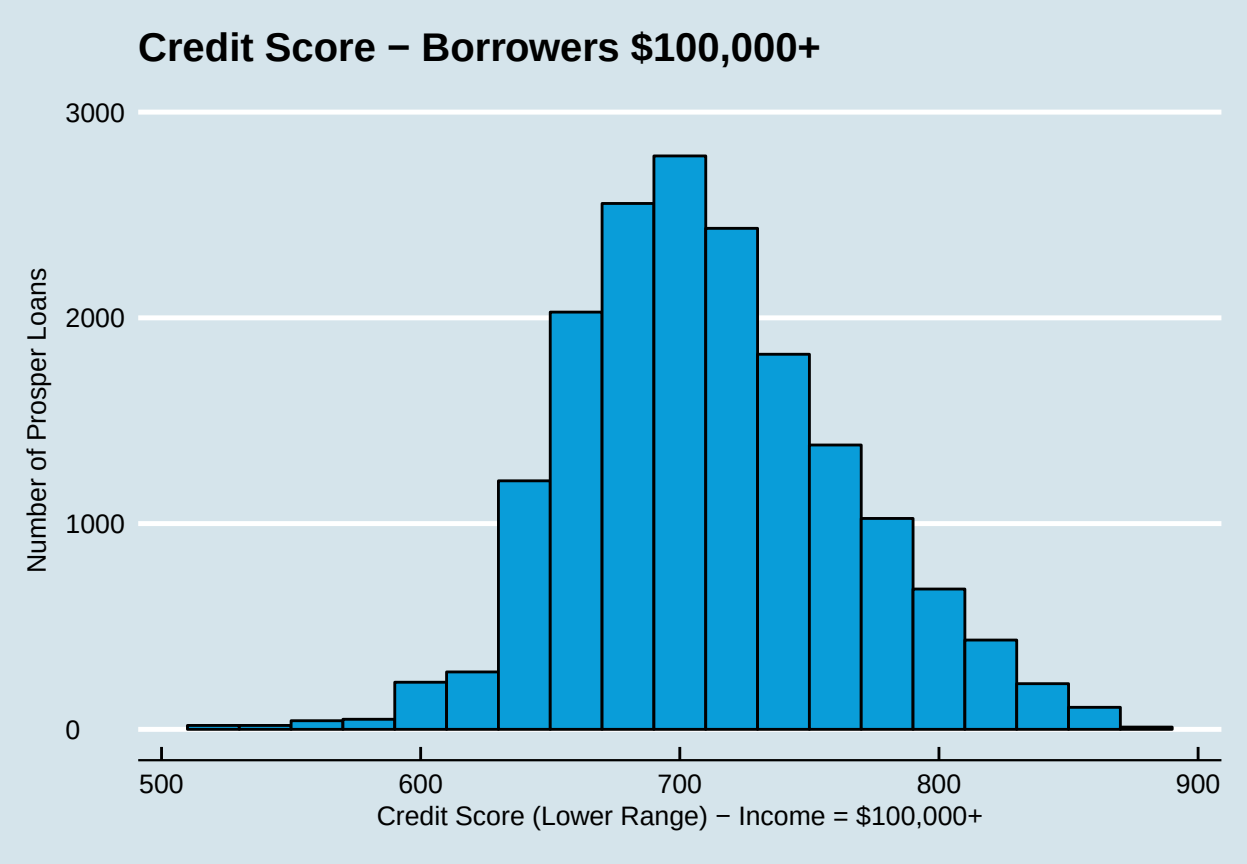


Why would Prosper make loans to people with no income?



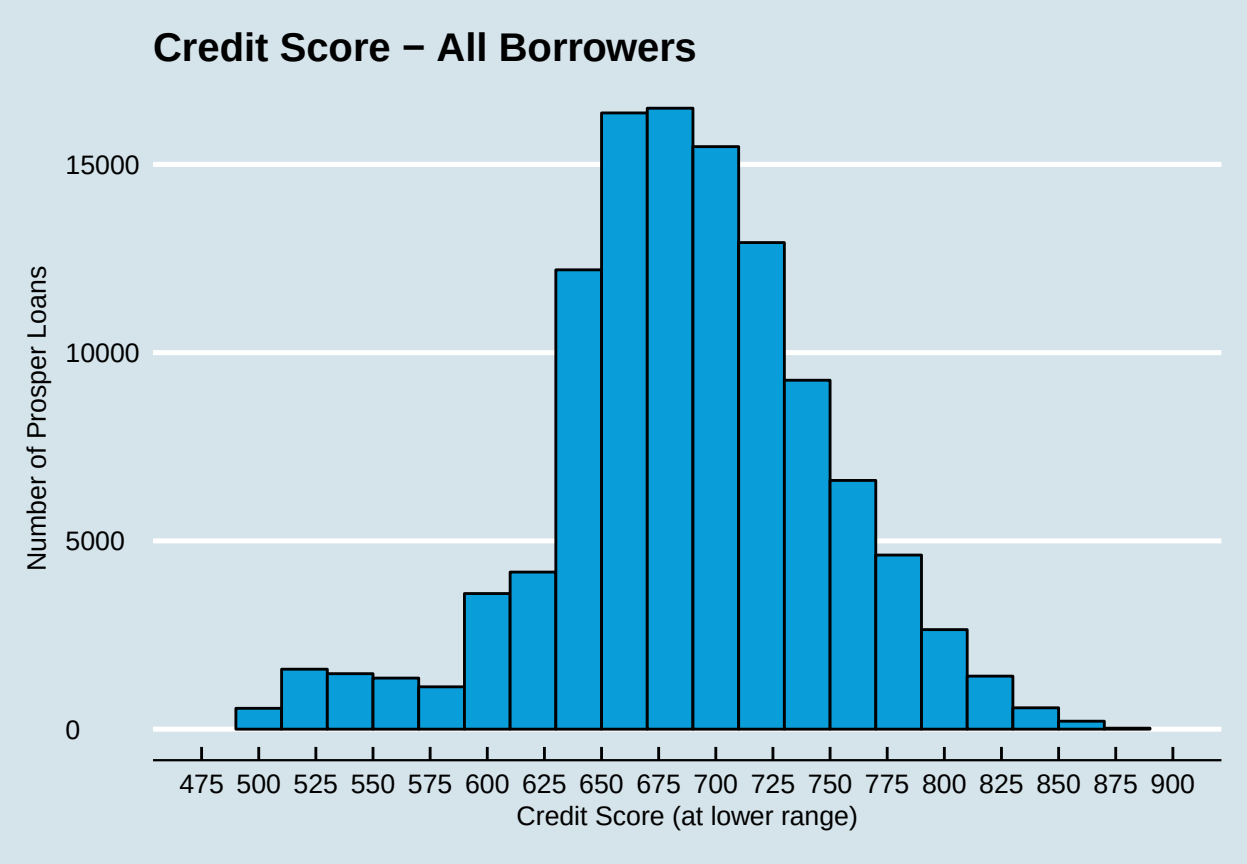

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   520.0   640.0   680.0   686.2   740.0   860.0
```

Credit scores for borrowers with $0 income follows a non-normal distribution with median at 680.



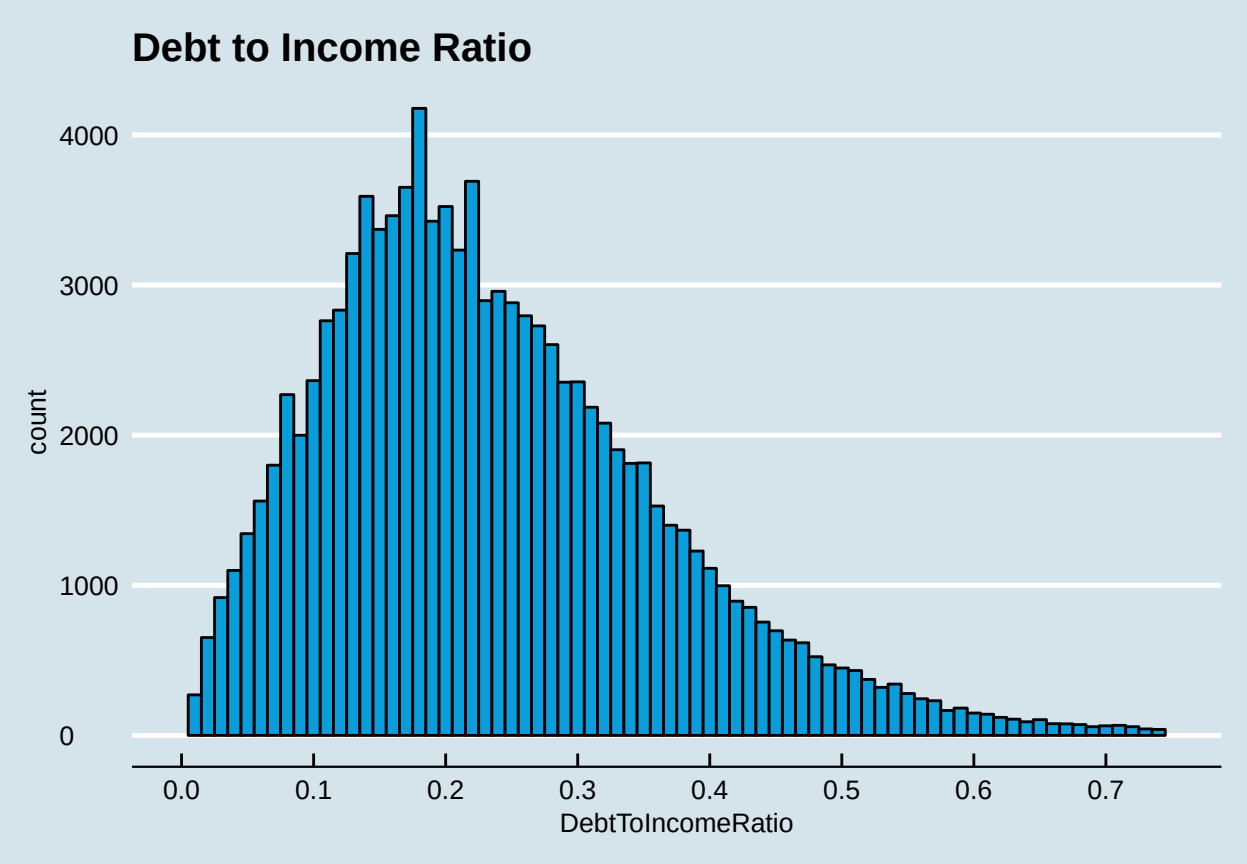

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   520.0   680.0   700.0   710.9   740.0   880.0
```

Credit score for $100,000+ borrowers follows a generally normal distribution, with median at 700 and inter-quartile range between 680 and 740.

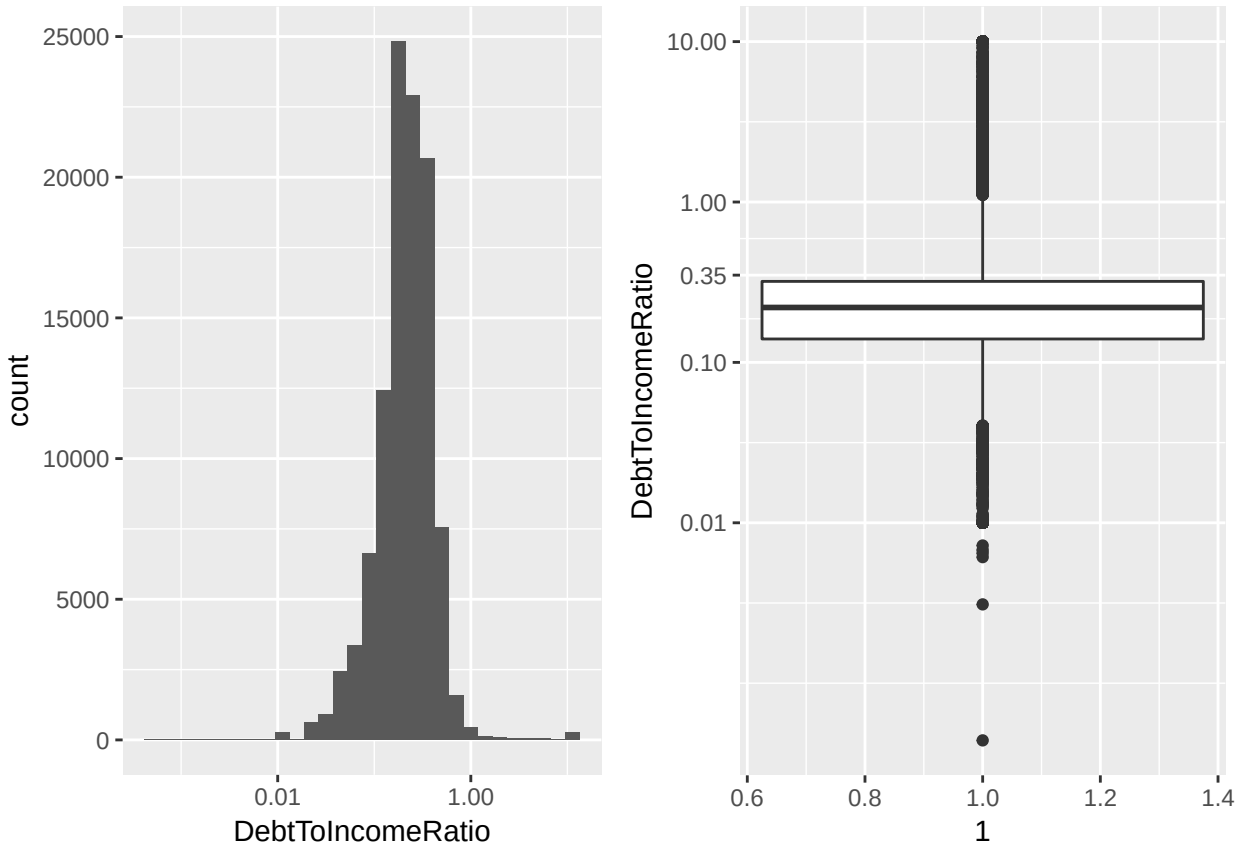



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0   660.0   680.0   685.6   720.0   880.0     591
```

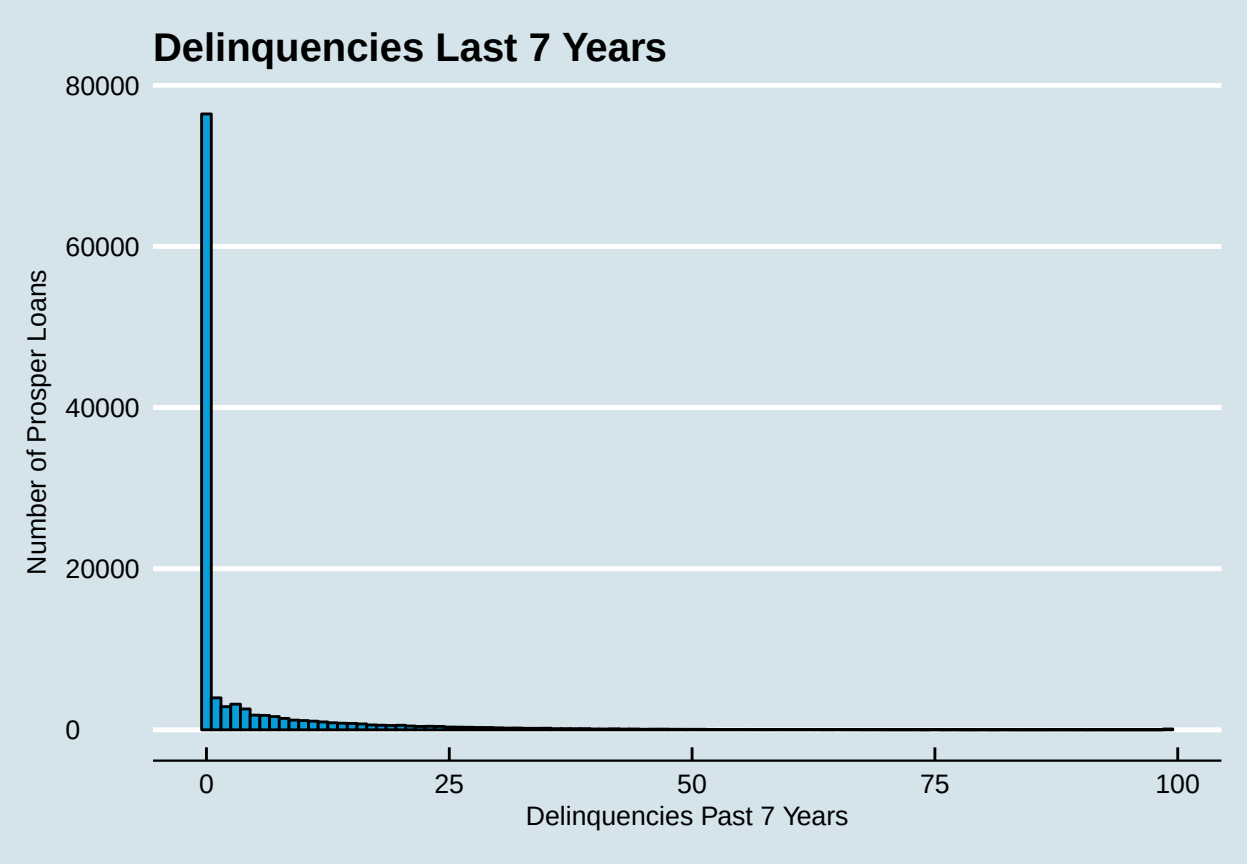
The credit score for all borrowers follows a fairly normal distribution with median of 680 and inter-quartile range between 660 and 720.



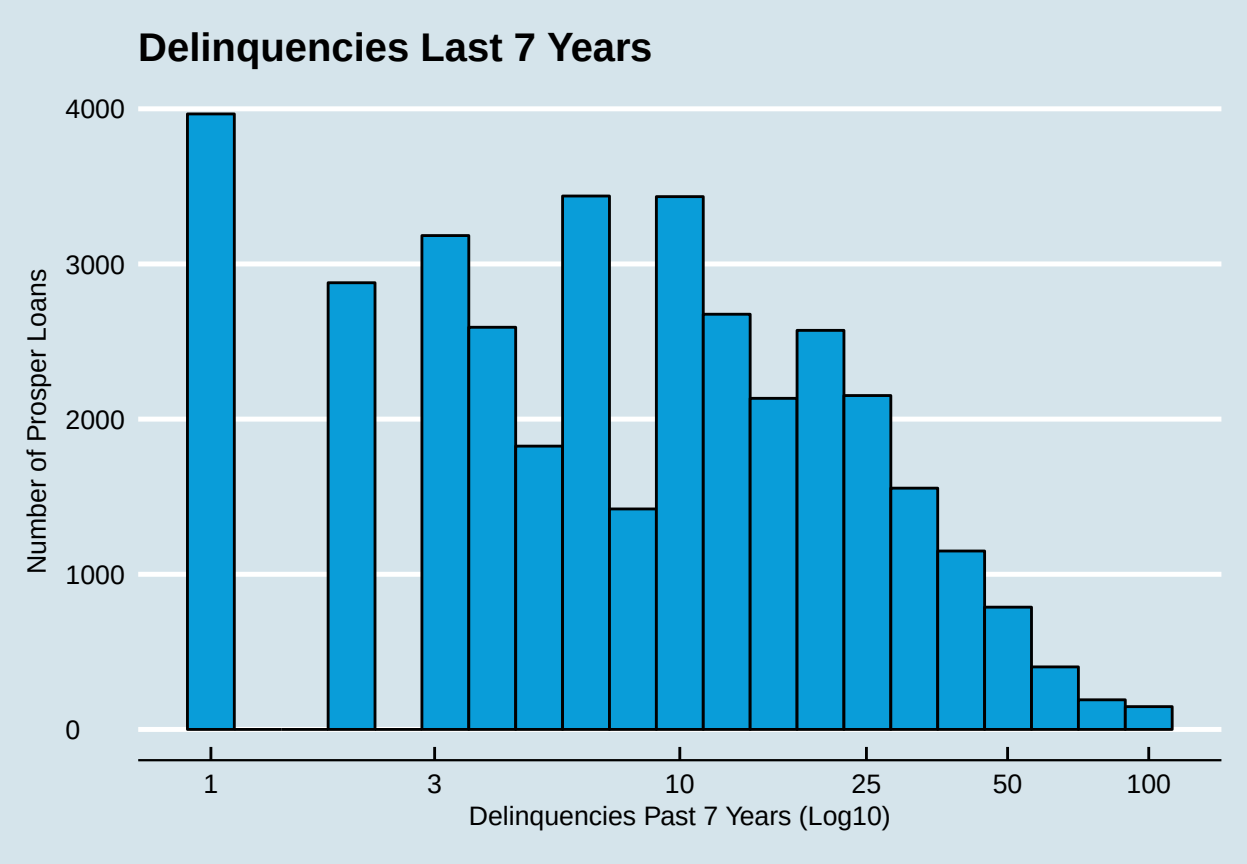
Most borrowers have a debt-to-income-ratio less than 0.4; however, there is a long tail on this chart. It's not visisble here because I limited the x-axis, but some borrowers have as much as 10 times debt to income.



From the boxplot, we see most borrowers have a debt-to-income-ratio somewhere between .1 and .35. However, the tails extend far away from center, with the lowest near 0 and the highest at 10 times debt-to-income.



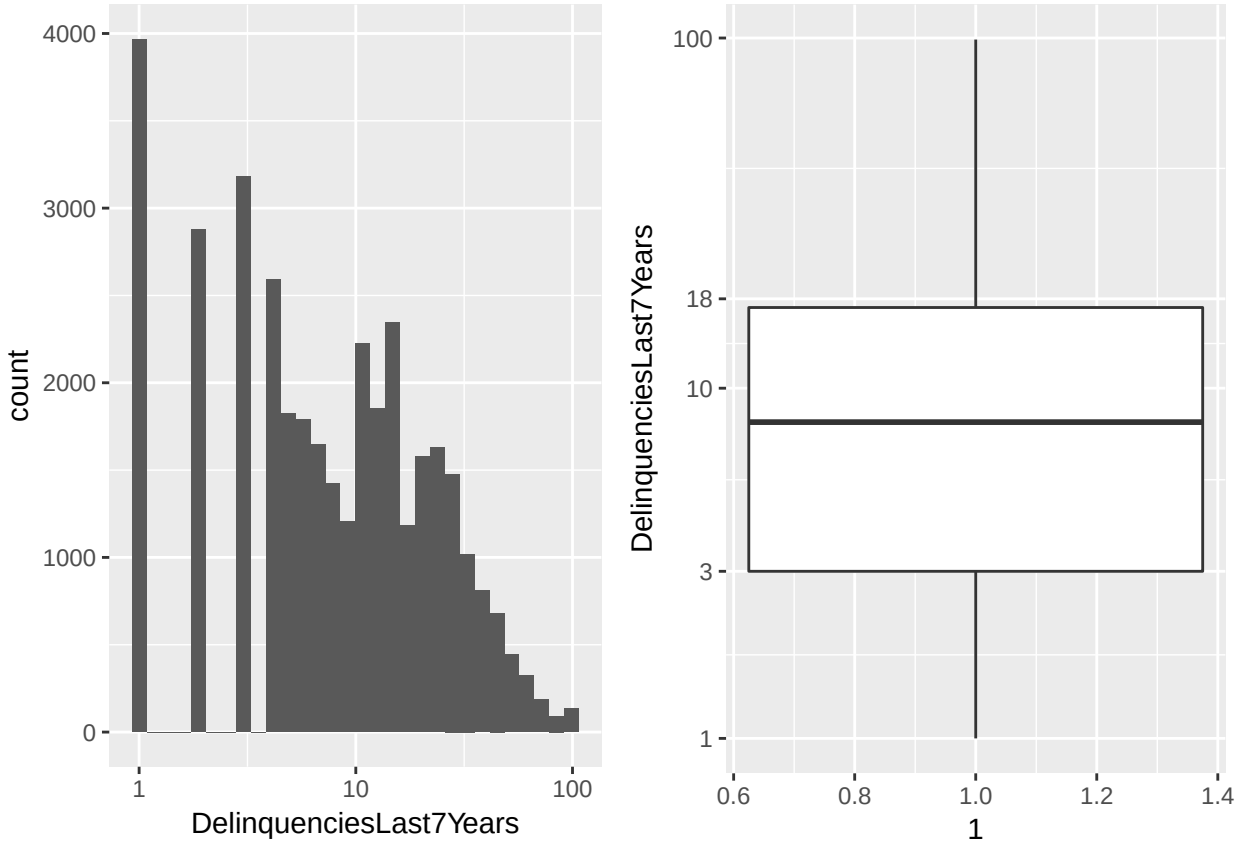
By far, the majority of borrowers have no delinquencies in the past 7 years, but there is a very long tail on this chart, with some borrowers having as many as 100 delinquencies. Let's zoom in.



Number of loans to borrowers with 25 or more delinquencies in past 7 years:

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1  5177
```

It's remarkable that over 5,000 loans were made to borrowers with more than 25 delinquencies in the past 7 years. Why would Prosper make loans to people with such bad credit histories?



From the boxplot, we see the median number of delinquencies for all borrowers is near 10, with inter-quartile range somewhere between 3 and 18. Again, the outliers extend far away from center, with some borrowers having as many as 100 delinquencies.

Univariate Analysis

What is the structure of your dataset?

There are 113,937 loans in the dataset with 81 variables, including borrower rate, borrower credit grade, loan status, borrower income, employment status, credit history, and loan payment information. The dataset contains two variables for borrower credit grade, one through 2008 called "CreditGrade" and another after 2008 called "ProsperRating..Alpha." Both are ordered factor variables with the following levels:

(best) ————————-> (worst)
CreditGrade: AA, A, B, C, D, E, HR
ProsperRating..Alpha.: AA, A, B, C, D, E, HR, NC

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

The main features that I will explore are borrower rate, loan status, income range, and creditworthiness, as determined by Prosper's proprietary rating system.

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

The credit history, credit score, and debt to income ratio likely influence Prosper's overall determination of creditworthiness. However, their exact formula is is not made public, so it may be difficult to determine precisely how they assign loans to a certain credit grade.

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

I created the variables Income_f, CreditGrade_f , ProsperRating_f, which are used to rearrange the order in which the variables of similar names display in plots.

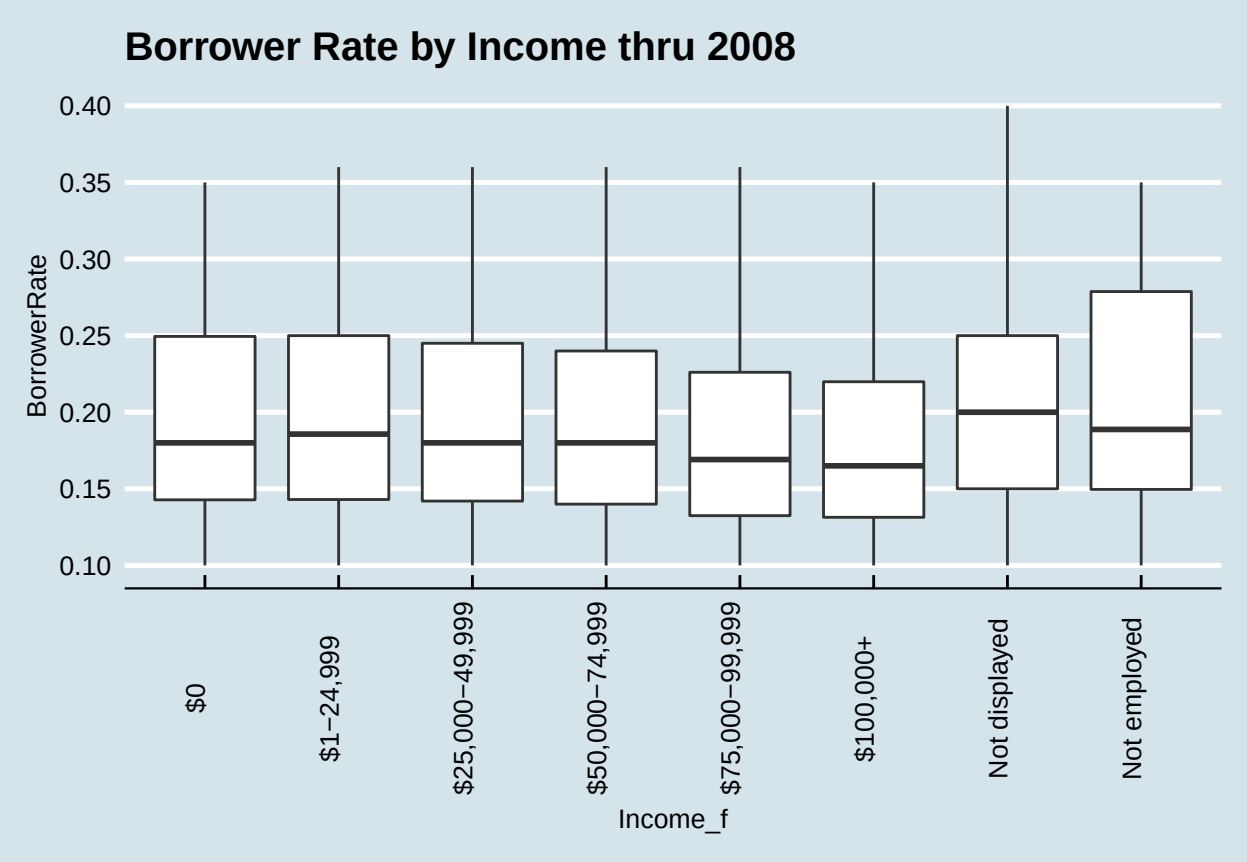
Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

I log transformed the right skewed plot for Delinquencies in the Last 7 years. The vast majority of borrowers had no delinquencies. However, log transforming the plot reveals a surprising number of borrowers with many delinquencies. In fact, there were over 5,000 loans made to borrowers who had 25 or more delinquencies in the past 7 years. This data becomes even more clear in the log transformed boxplot of delinquencies, revealing an inter-quartile range of delinquencies somewhere between 3 and 18.

Similarly, I log transformed the right skewed plot for Debt to Income Ratio. The log transormed boxplot of this data reveals that most borrowers have a debt-to-income-ratio somewhere between .1 and .35. However, the tails extend far away from center, with the lowest near 0 and the highest at 10 times debt-to-income.

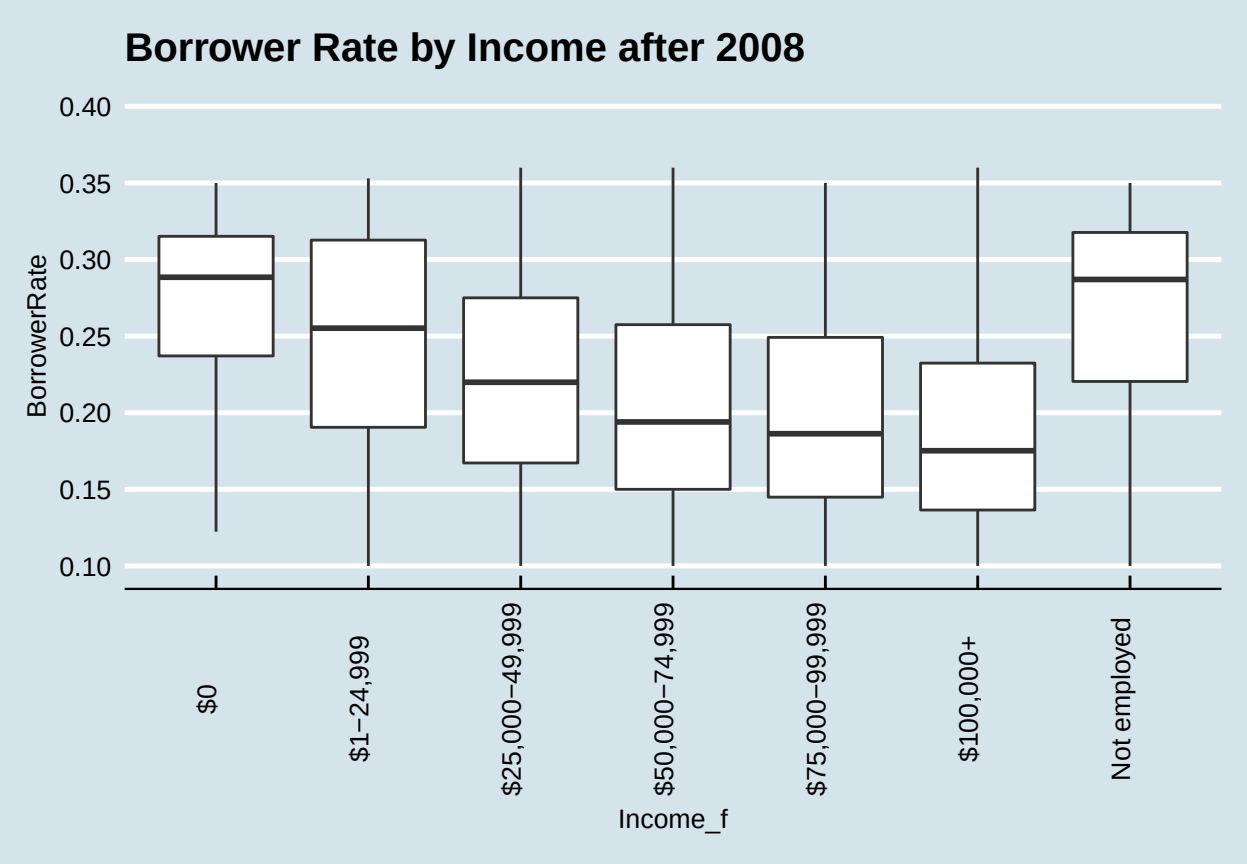
Finally, I also log transformed the y-axis for the Borrower State plot to get a look at the states who had the fewest number of borrowers.

Bivariate Plots Section



Not much variability in median interest rate across income ranges prior to 2009. The median interest rate decreases only slightly as income increases, and surprisingly, borrowers with no income have a similar interest rate to that of borrowers in the $50,000-74,999 & $75,000-99,999 ranges. Why is that?

Furthermore, the median borrower rate for each income range is higher than I would have thought. The lowest is over 15%. I wonder what the rates look like after 2008, when the economy went in the tank.



Much more variability in median borrower rate across income ranges after 2008, with a clear decline in median rate as income increases. This is more in line with what one would expect.

Summary of Borrower Rate by Income thru 2008

```
## pld$Income_f: $0
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0050  0.1400  0.1750  0.1952  0.2500  0.3500
## ------------------------------------------------------------
## pld$Income_f: $1-24,999
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.1550  0.2199  0.2206  0.2900  0.3600
## ------------------------------------------------------------
## pld$Income_f: $25,000-49,999
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.1474  0.2015  0.2072  0.2684  0.3600
## ------------------------------------------------------------
## pld$Income_f: $50,000-74,999
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.1334  0.1800  0.1903  0.2487  0.3600
## ------------------------------------------------------------
## pld$Income_f: $75,000-99,999
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.1239  0.1699  0.1809  0.2321  0.3600
## ------------------------------------------------------------
```

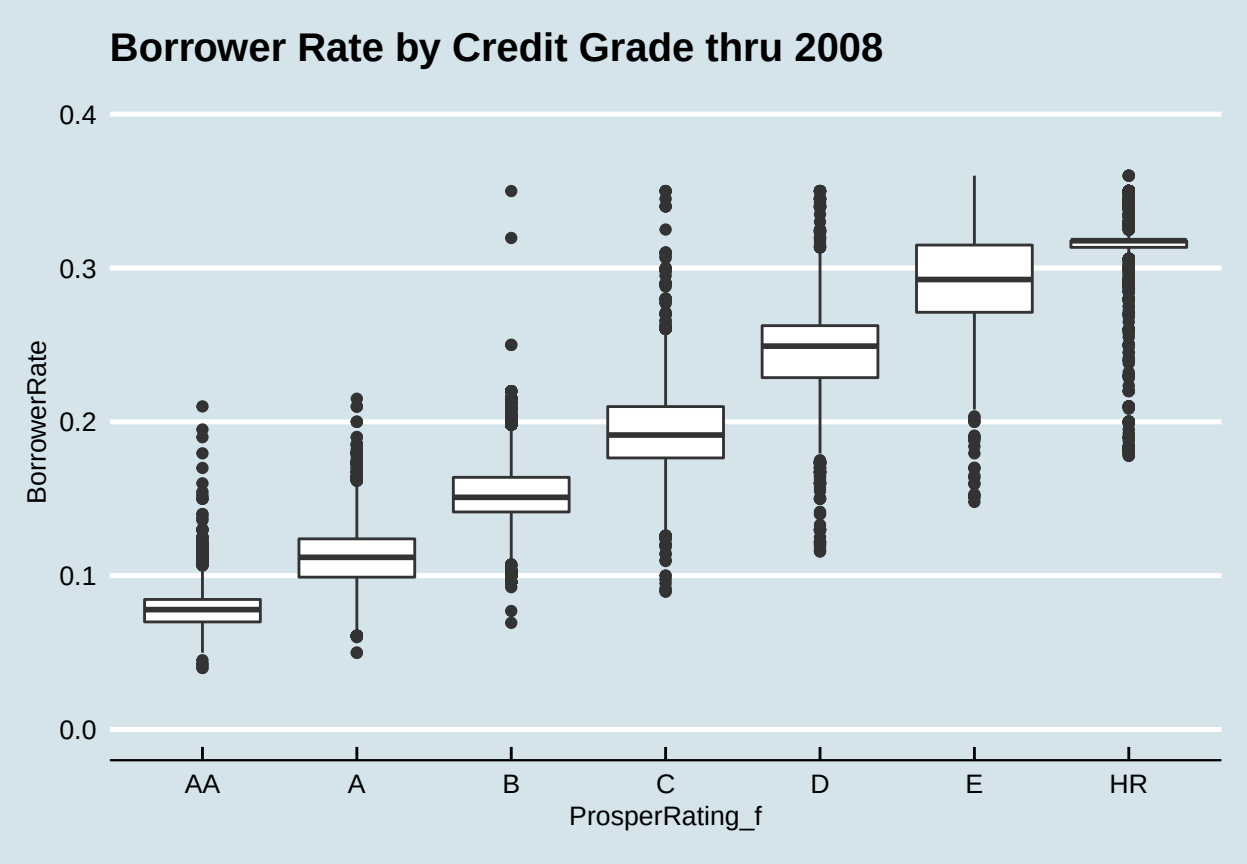
```
## pld$Income_f: $100,000+
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.1139  0.1550  0.1692  0.2124  0.3600
## ------------------------------------------------------------
## pld$Income_f: Not displayed
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.1350  0.1880  0.1892  0.2443  0.4975
## ------------------------------------------------------------
## pld$Income_f: Not employed
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0400  0.1874  0.2600  0.2467  0.3149  0.3500
```

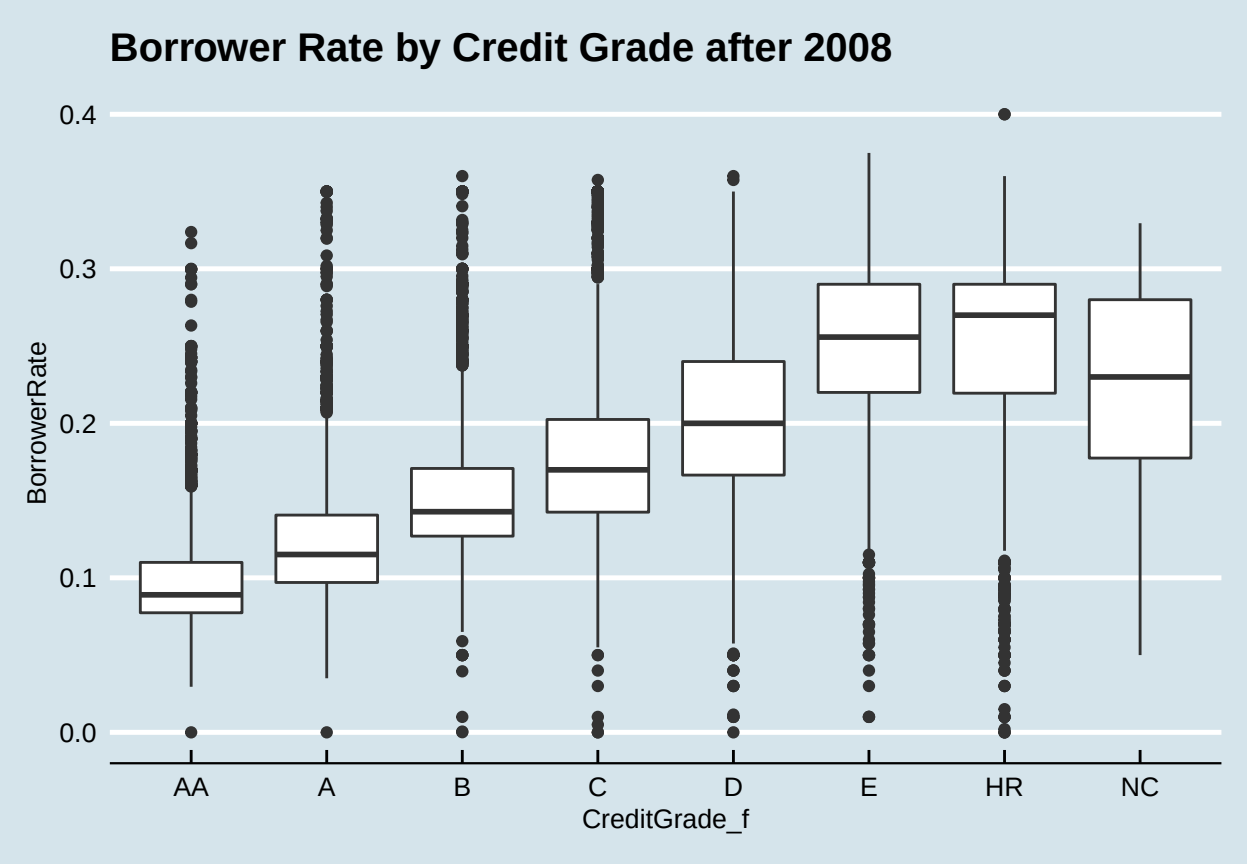
Summary of Borrower Rate by Income after 2008

```
## pld_a_2008$Income_f: $0
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0920  0.2287  0.2870  0.2655  0.3149  0.3500
## ------------------------------------------------------------
## pld_a_2008$Income_f: $1-24,999
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0550  0.1824  0.2493  0.2374  0.3123  0.3530
## ------------------------------------------------------------
## pld_a_2008$Income_f: $25,000-49,999
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0498  0.1559  0.2100  0.2136  0.2699  0.3600
## ------------------------------------------------------------
## pld_a_2008$Income_f: $50,000-74,999
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0400  0.1355  0.1840  0.1926  0.2499  0.3600
## ------------------------------------------------------------
## pld_a_2008$Income_f: $75,000-99,999
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0499  0.1249  0.1734  0.1831  0.2359  0.3500
## ------------------------------------------------------------
## pld_a_2008$Income_f: $100,000+
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0450  0.1153  0.1559  0.1703  0.2139  0.3600
## ------------------------------------------------------------
## pld_a_2008$Income_f: Not displayed
## NULL
## ------------------------------------------------------------
## pld_a_2008$Income_f: Not employed
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0800  0.2148  0.2859  0.2614  0.3177  0.3500
```

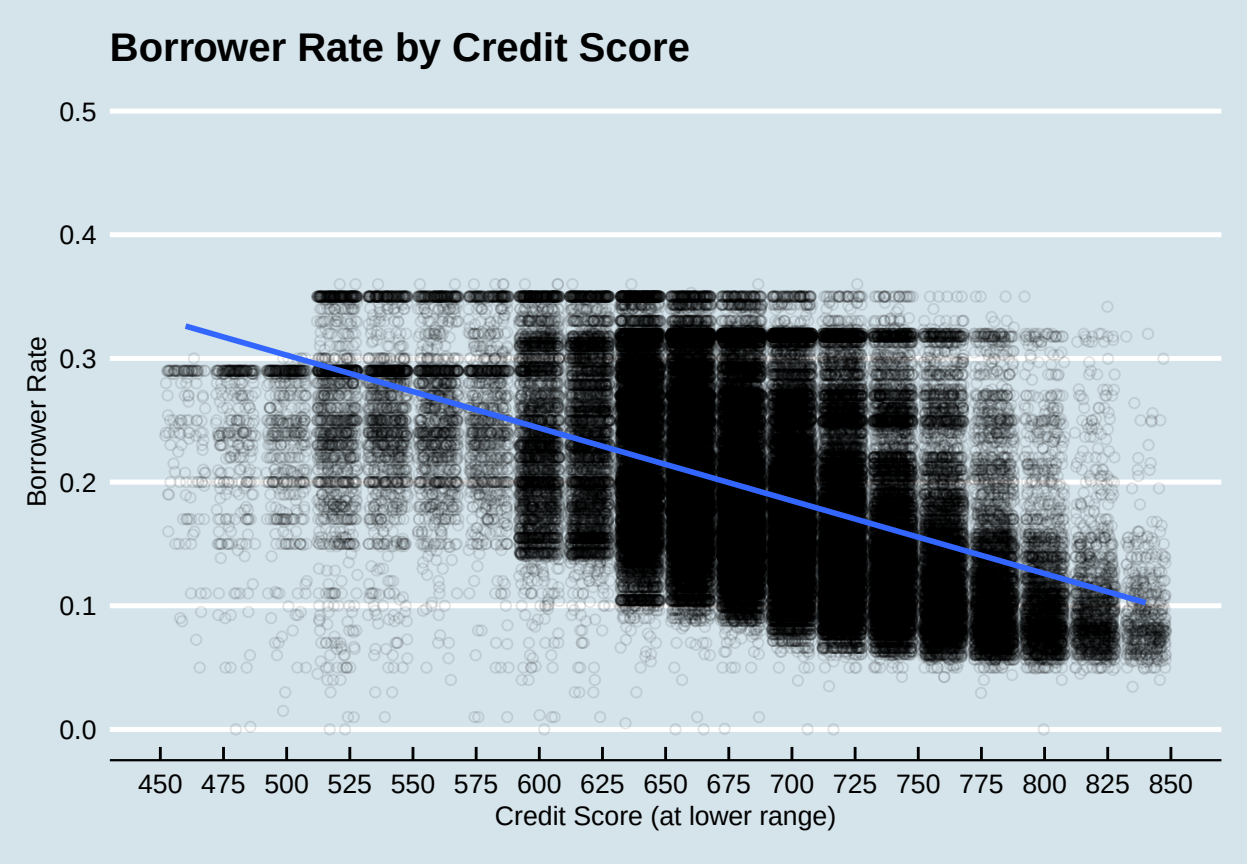
Despite the Fed's reduction of the Prime Rate to nearly zero following the 2008 financial crisis, the median borrower rate across income categories generally increases after 2008. The rate especially increases in the low income range. Borrowers earning $1-$24,999 saw a median increase in interest rate of 2.94 percentage points, whereas the $100,000+ borrowers saw a median increase of .09 points.

Surprisingly, borrowers through 2008 with $0 income had a median interest rate lower than the borrowers in the next three income ranges. Their rate fell between that of borrowers in the $50,000-74,999 and $75,000-99,999 ranges. This anomaly corrected after 2008, with the median rate for $0 income borrowers jumping to 28.70%, the highest for all categories.


Borrower Rate by Credit Grade thru 2008
0.4
0.3
0.2
0.1
0.0
BorrowerRate
AA
A
B
C
D
E
HR
ProsperRating_f



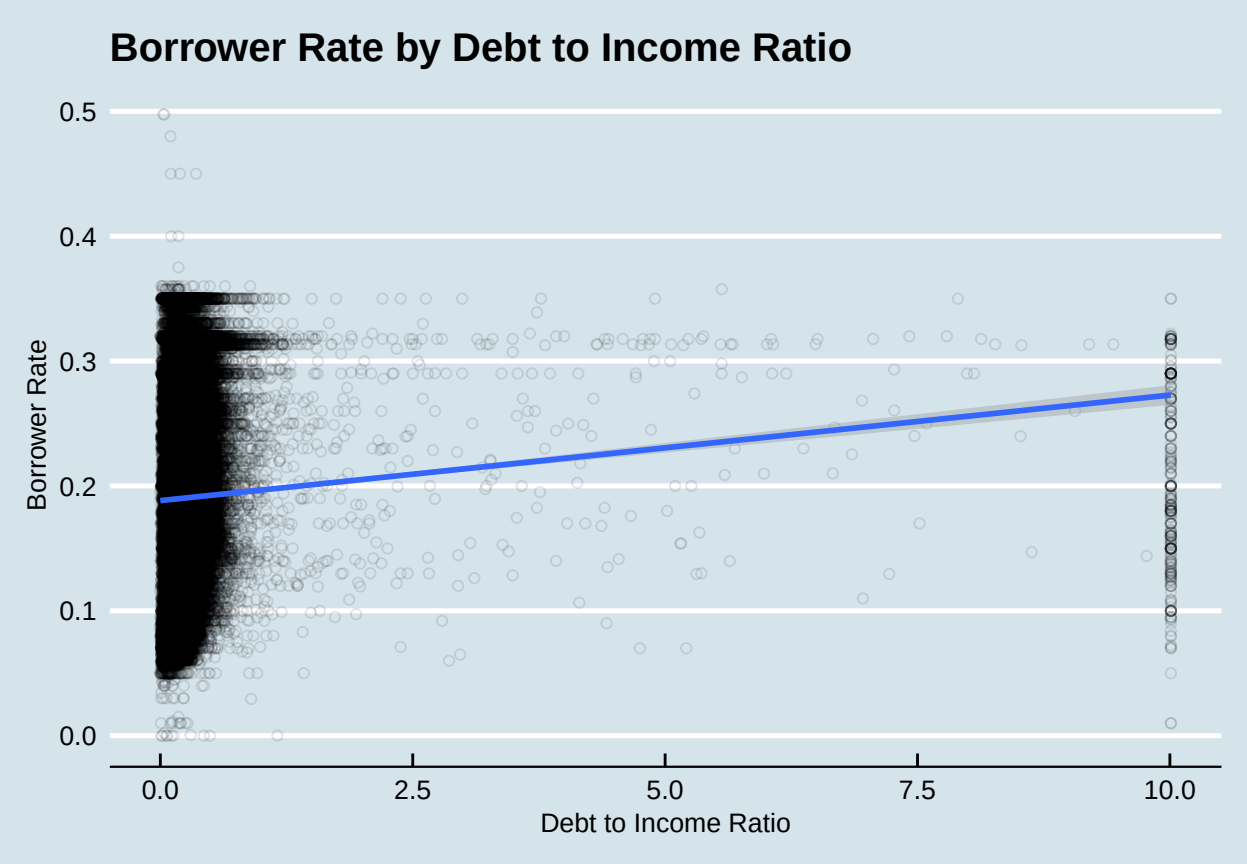
Clearly the borrower rate is tied closely to a borrower's creditworthiness as determined by Prosper. As the rating gets weaker, the median interest rate goes up. I wonder what criteria Prosper uses to determine the credit rating.



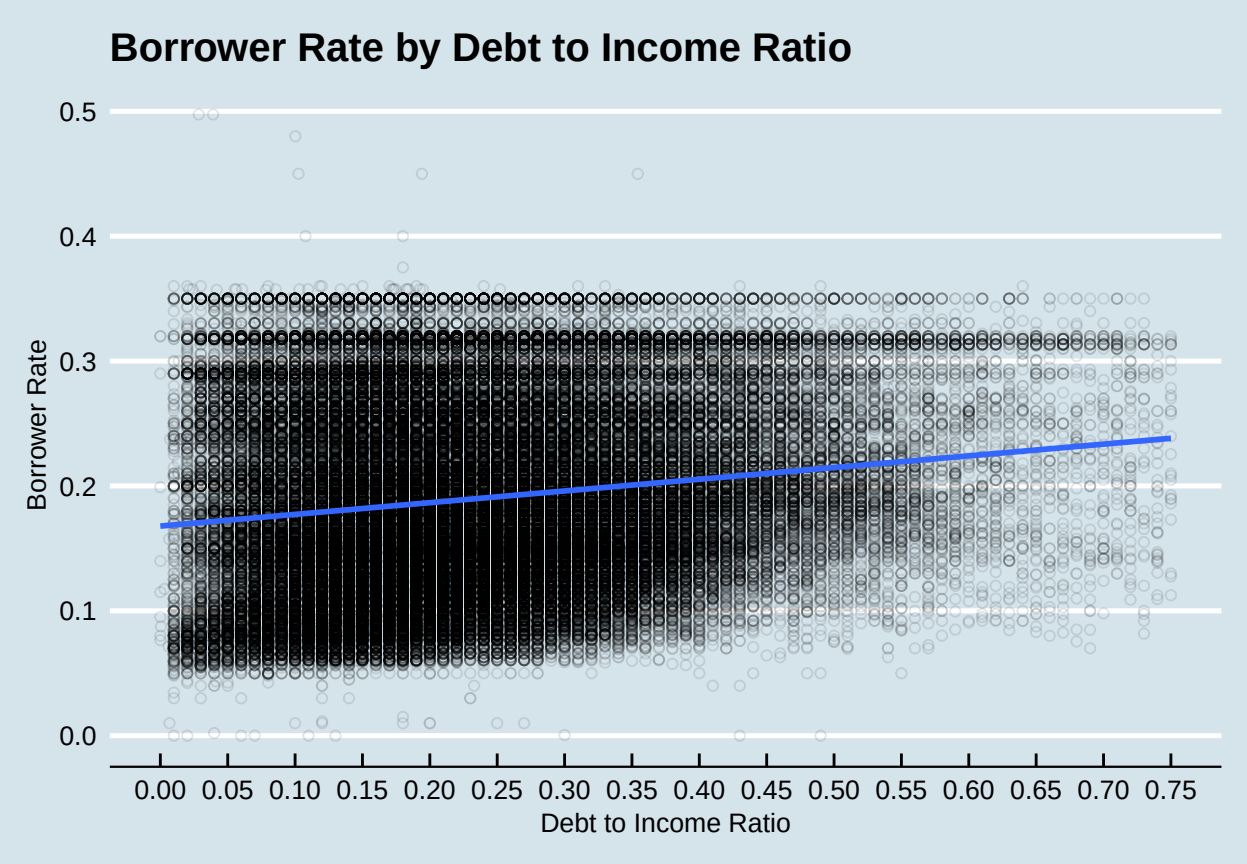
Correlation coefficient:

```
## [1] -0.4615667
```

Weak negative correlation between credit score and borrower rate. Certainly more low-rate loans are made for borrowers with high credit scores, but interest rates run the gamut for all credit scores.



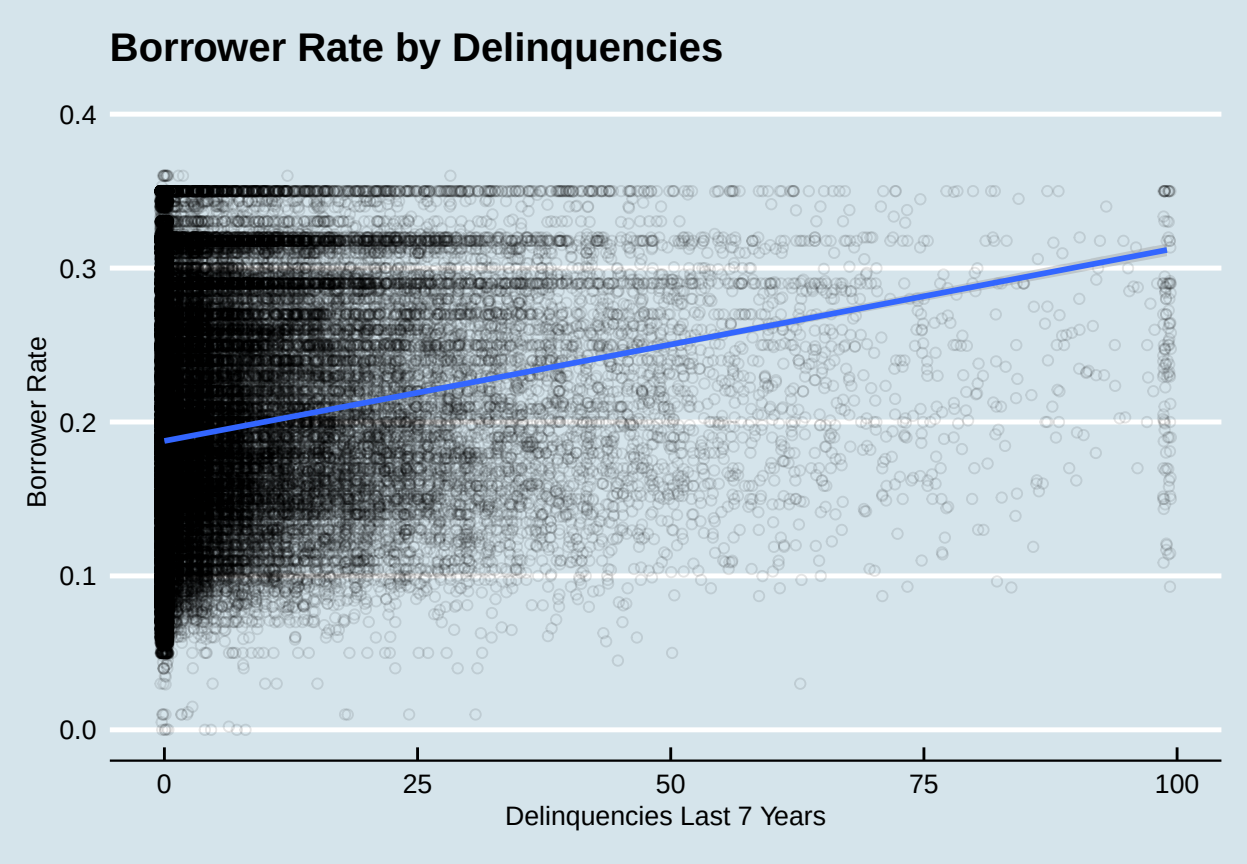
Understandably, most loans are made for borrowers with less than 0.75 debt-to-income-ratio. Remarkably, some loans are made (some even with rates in the teens) for borrowers with 5 times debt to income, or more. Let's zoom in to get a closer look at the typical borrower.



Correlation coefficient:

```
## [1] 0.06291678
```

Almost no correlation between debt-to-income-ratio and borrower rate. Interest rates run the gamut for each ratio.

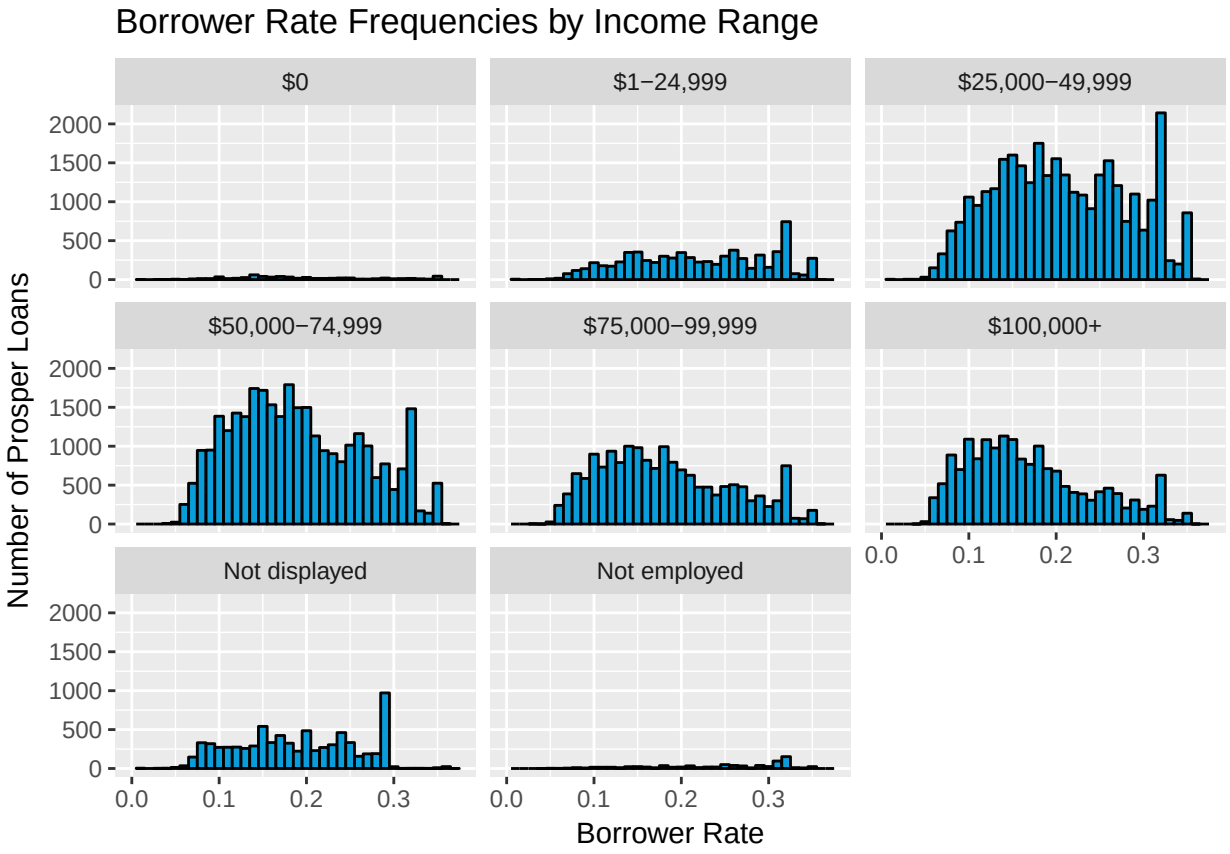


Correlation coefficient:

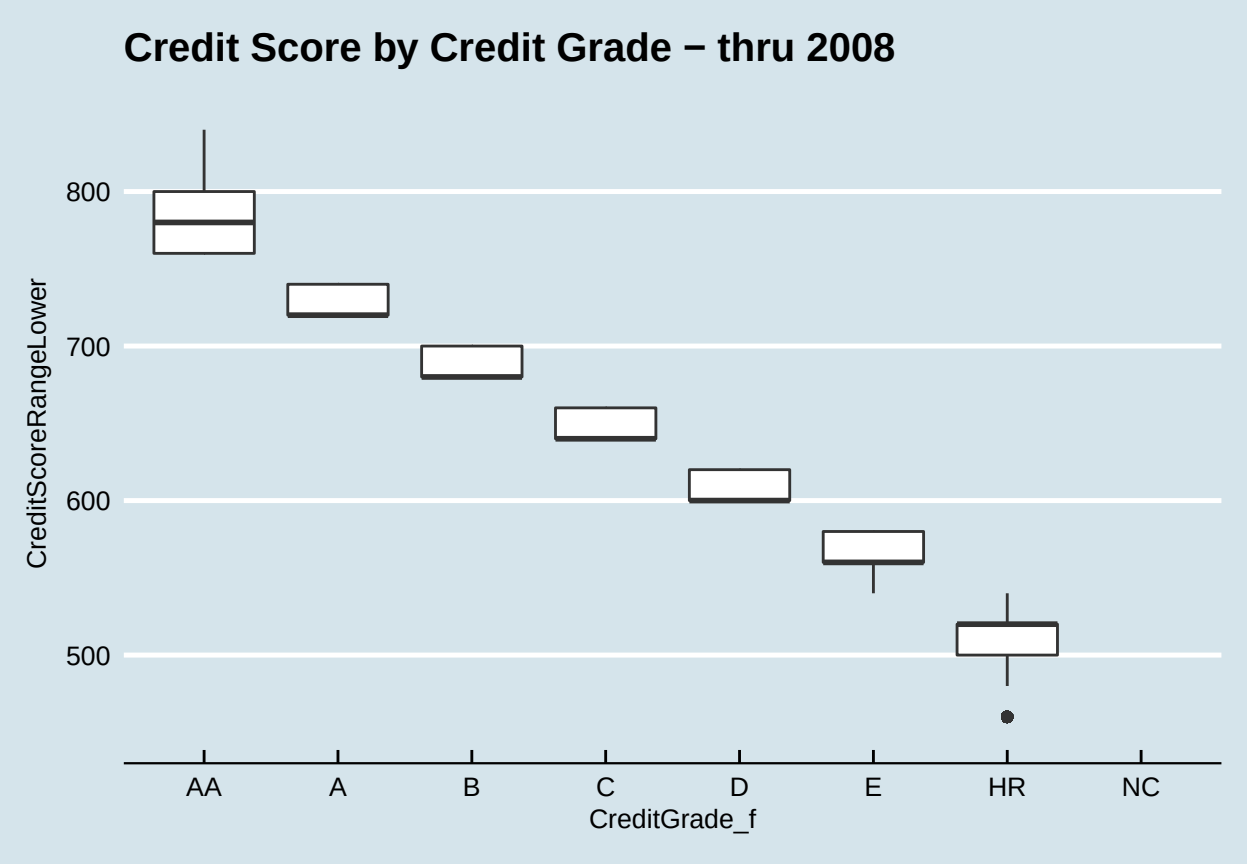
```
## [1] 0.1702787
```

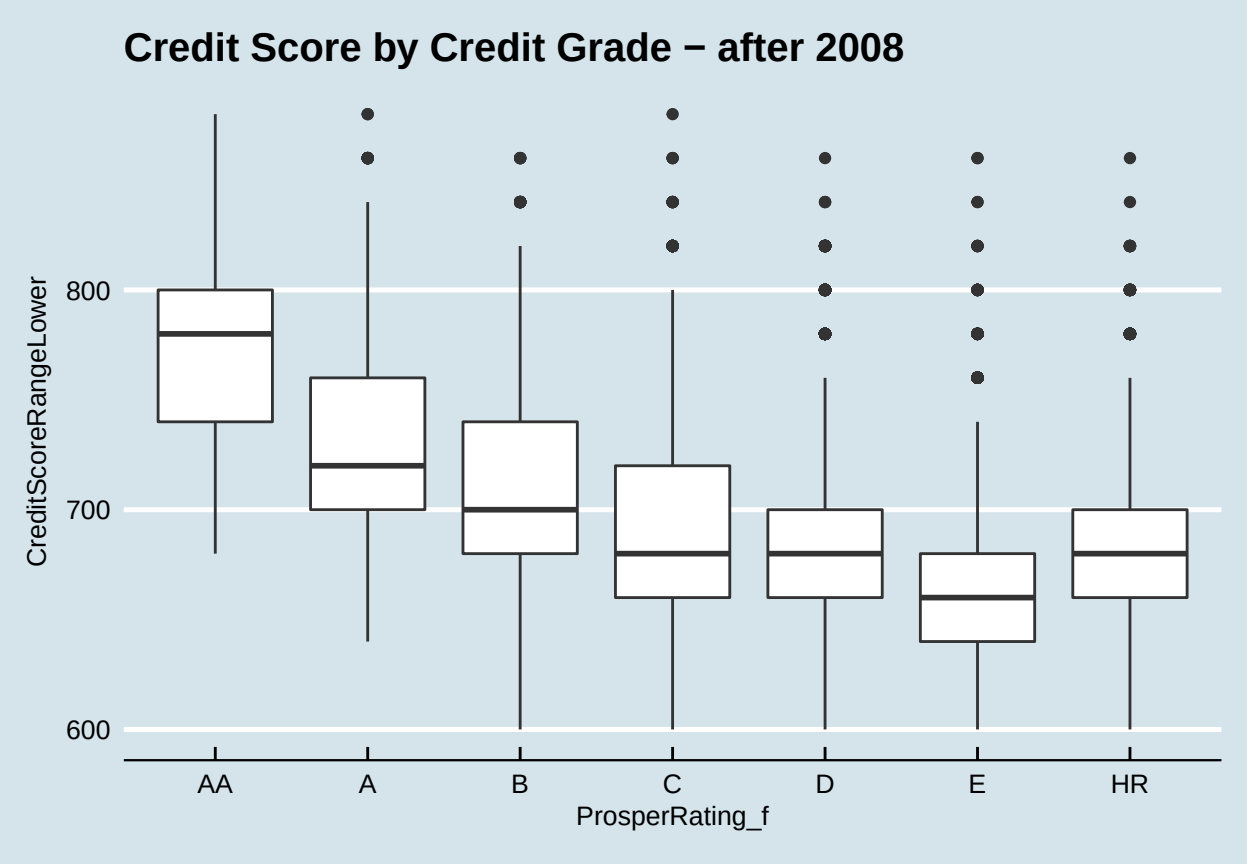
Very weak correlation between delinquencies and borrower rate.

As a side note, it is remarkable that Prosper would make so many loans to people more than 25 delinquencies in the past 7 years, especially with some of them at rates below 10%.



The most loans are made for income ranges $25,000-49,999 and $50,000-74,999. The fewest loans are made for borrowers who have no income or are not employed, which is not surprising.


Credit Score by Credit Grade − thru 2008
CreditScoreRangeLower
800
700
600
500
AA
A
B
C
D
E
HR
NC
CreditGrade_f



Very little variability in credit score by credit grade up through 2008. However, after 2008 we see significantly more variability in credit scores across the credit grades, which reflects what we've already seen with regard to changes in how Prosper determined credit grade beginning in 2009.

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?

Prior to 2009, median borrower rate varied only slightly across income ranges, and remarkably, borrowers earning $0 income enjoyed a rate similar to that of borrowers in the $50,000-100,000 dollar range. Prosper changed its rating variable beginning in 2009. Then median rates began to fall more in line with what one would expect, with a clear decrease in borrower rate as income increases.

As indicated in the boxplots of the relationship between credit grade and borrower rate, it is clear that Prosper sets the borrower rate based on their proprietary credit rating system, which they changed beginning in 2009. To try to understand how Prosper determines the credit rating, I plotted several variables against the borrower rate to see if there were possible correlations. I could not find any obvious correlations.

After doing some research into how Prosper determines the rating, I learned that their analysis may include hundreds of variables, not all of which may be included in the dataset. Therefore, I shifted my focus to comparing the quality of their rating system up through 2008 and after 2008, which occurs primarily in the multivariate section.

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

There is a distinct change in credit scores before and after 2008 when compared across credit grade. Through 2008, Prosper seems to have based the credit grade largely on credit score. After 2008, they changed their method and we begin to see a great deal more variability in the credit scores across credit grade. Therefore, presumably Prosper began to add many more variables to their assessment of credit risk.

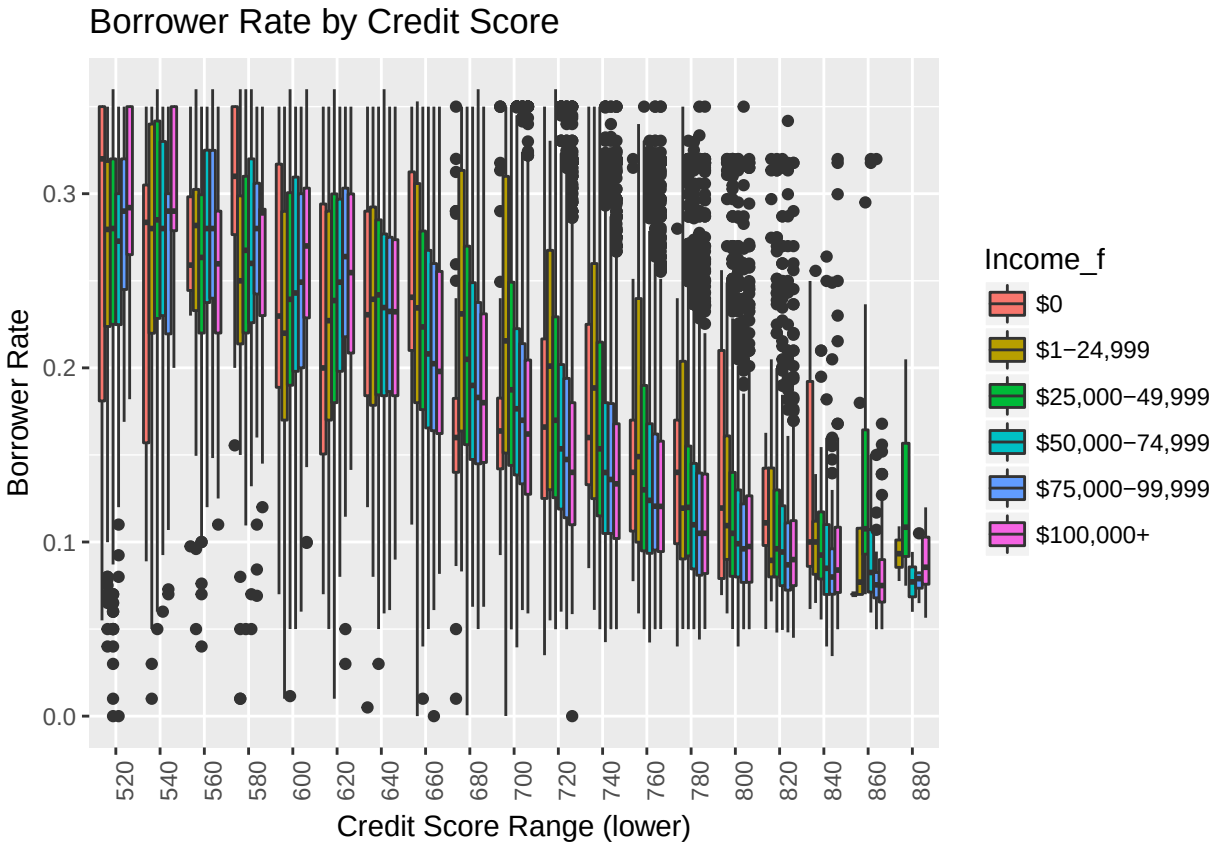
What was the strongest relationship you found?

The strongest relationship was borrower rate vs Prosper credit grade. The boxplots indicated a clear increase in median borrower rate as the borrowers' credit grades get worse.

Multivariate Plots Section

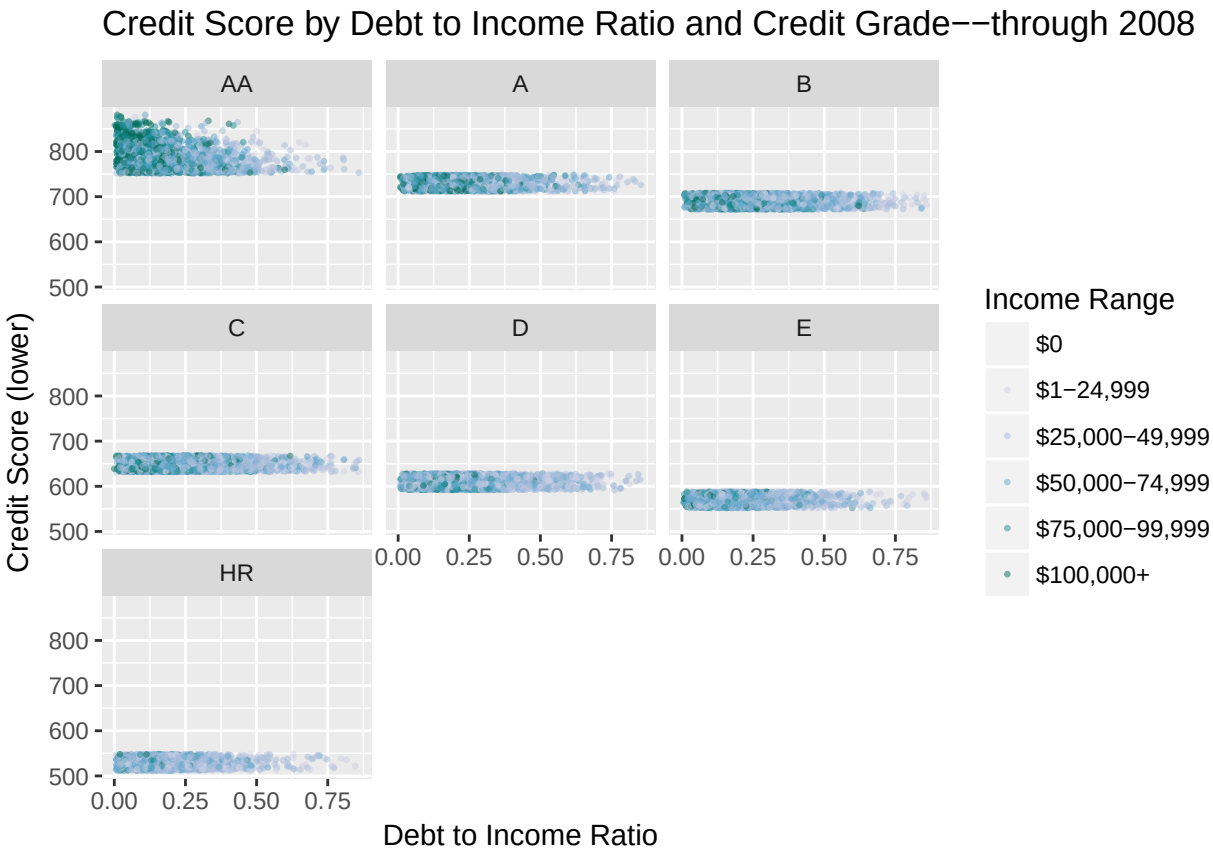
Again, no strong relationship between borrower rate and debt-to-income-ratio. There seems to be a weak relationship between income range and debt-to-income-ratio, as would be expected. But as for borrower rate, neither income range nor debt-to-income-ratio seem to have a significant influence on the rate Prosper sets.

Not income range, nor debt-to-income-ratio, nor borrower rate seem to be good indicators of whether the loan will remain current.



In general, the higher the credit score, the lower the borrower rate. However, there are many outliers to this trend. Furthermore, for borrowers with high credit scores but low income, there is wide variability in interest rate. For example, at credit score 840, borrowers with $0 income have a huge inter-quartile range compared with borrowers from other income categories.

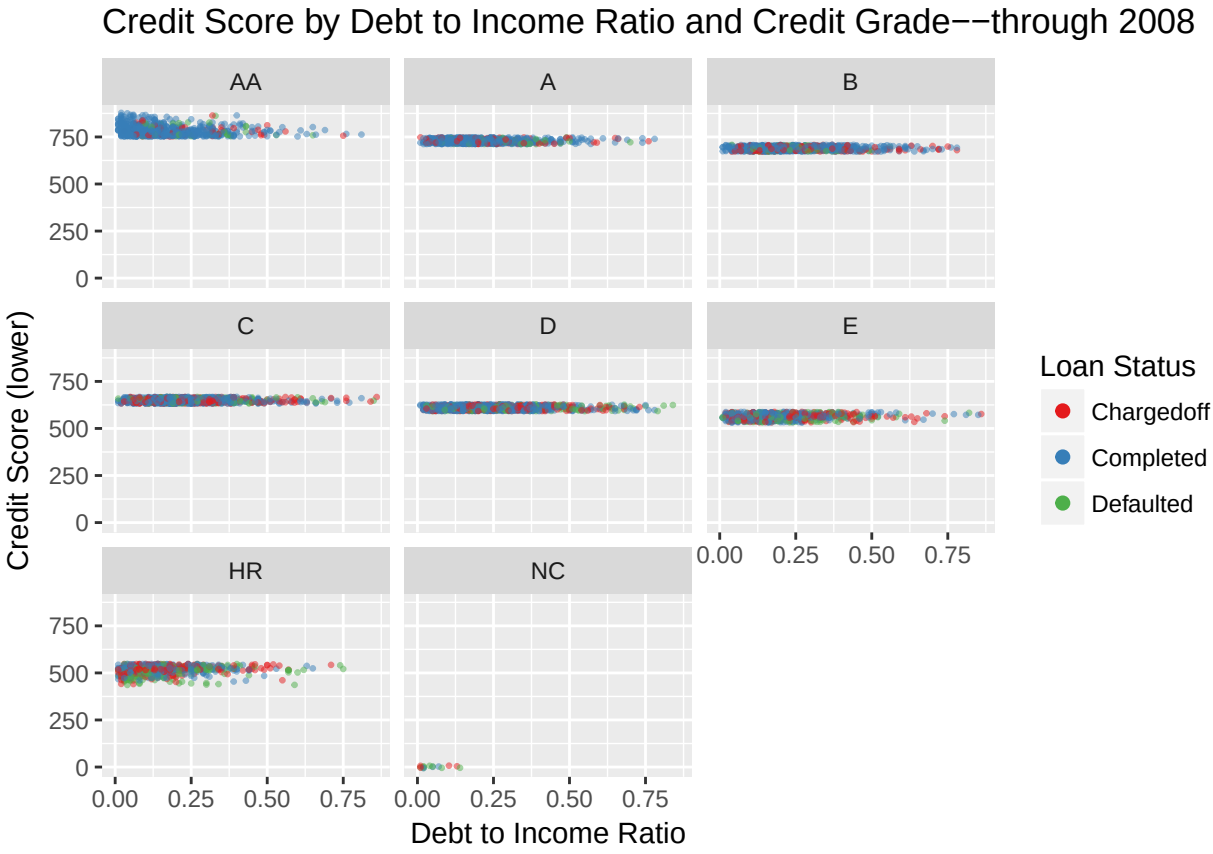
Higher credit scores have a higher concentration of high income borrowers; nevertheless, there are still many high income borrowers with low debt-to-income-ratios who have less than ideal credit scores.



Up through 2008, given the clear delineaions at certain credit scores, Prosper's determination of credit grade seemed to be based primarily on credit score.

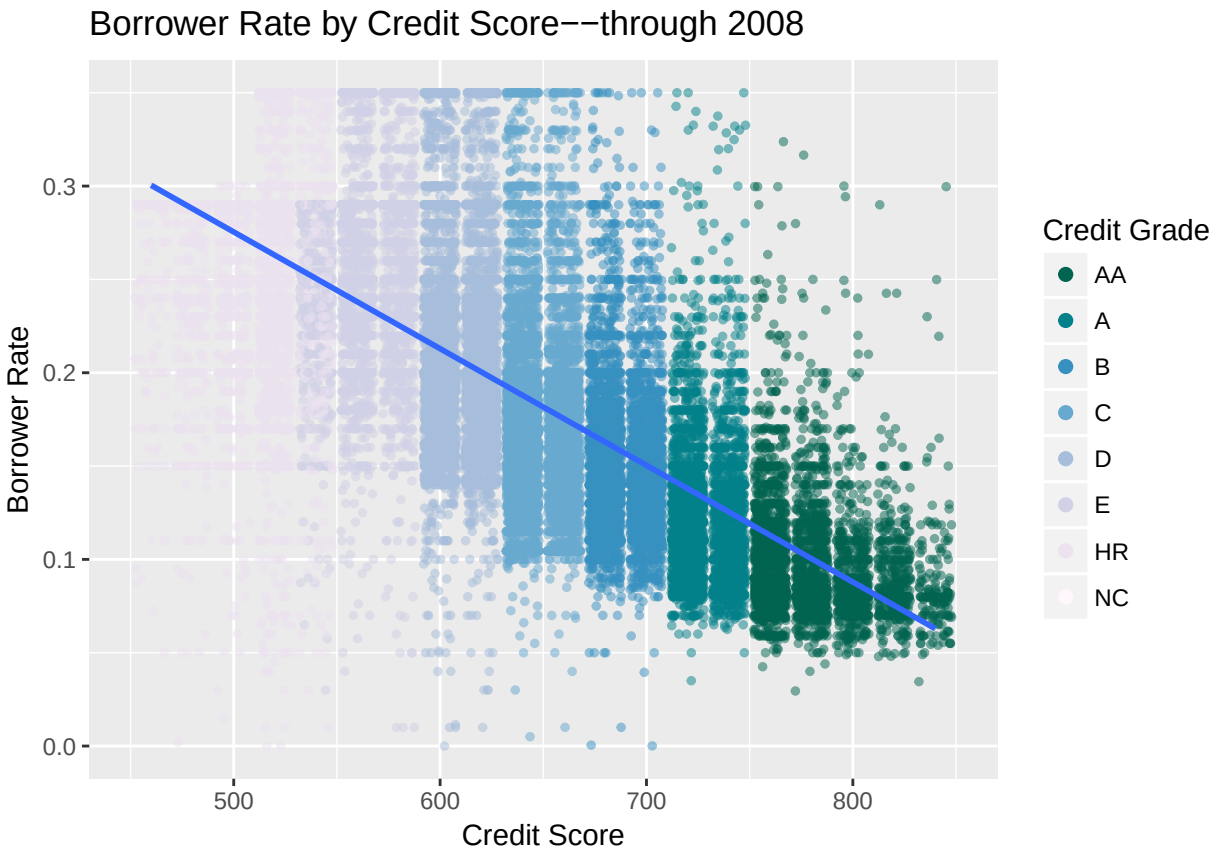
After 2008, Prosper seems to have developed a more nuanced approach. As expected, high income borrowers with higher credit scores tend to make up the best Prosper ratings, but no longer are there clear delineations by credit score.

Furthermore, as we've seen in previous plots, the correlations between debt-to-income-ratio and credit score are weak at best, with the strongest in this plot being a negative correlation in the "AA" category.



This chart looks at Loan Status across the various credit grades. It reveals a fairly significant number of chargeoffs and defaults, even in the higher credit grades.

The quality of Prosper's risk assessment does seem to improve after 2008. Up through 2008, a fairly significant number of chargeoffs and defaults occur in the higher grade loans. After 2008, the chargeoffs and defaults in grades AA-B appear to be a bit less.

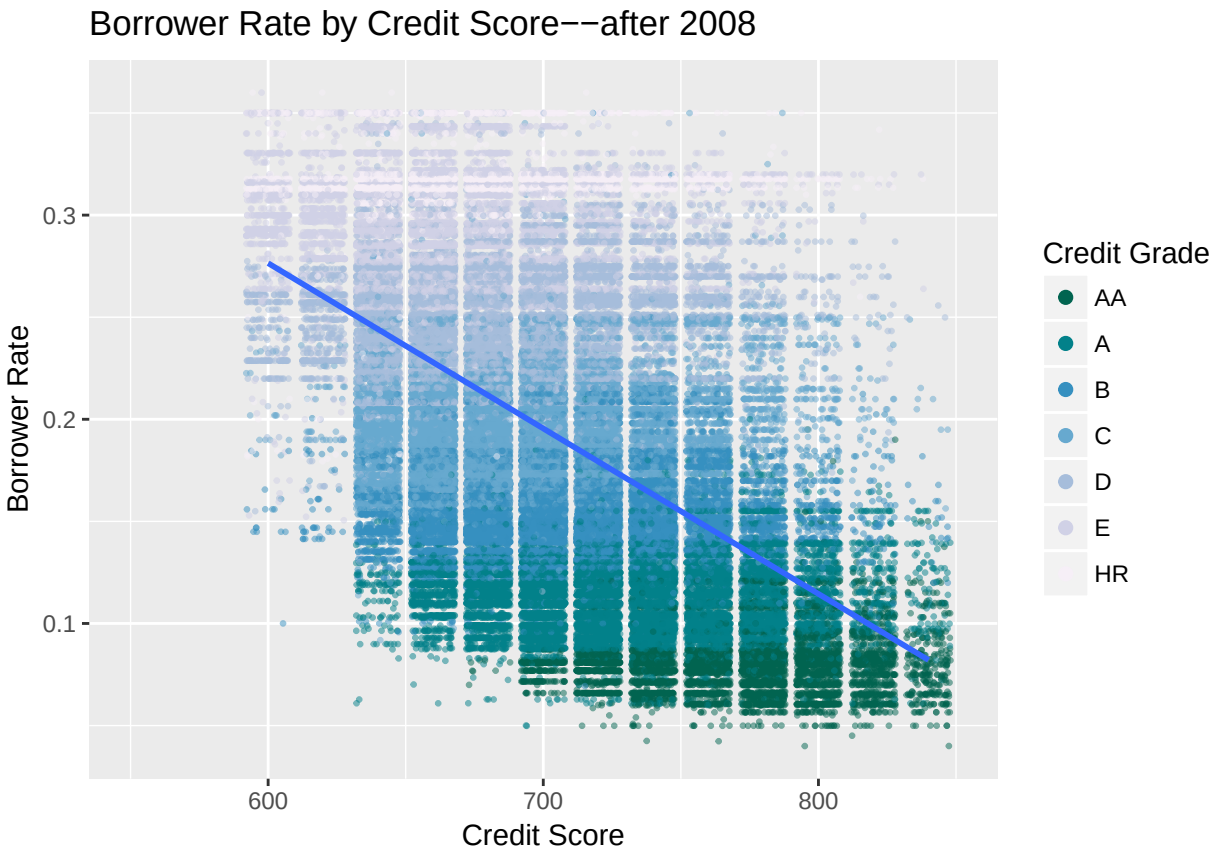


Correlation coefficient:

```
## [1] -0.6219264
```

This is the strongest correlation we've seen yet, which is understandable given that Prosper apparently based its assessment of credit risk up through 2008 primarily on credit score.

Looking at the colored variable, we see definitive cut scores for credit grade based on credit score up through 2008, while the borrower rate varies significantly within the credit risk levels.



Correlation coefficient:

```
## [1] -0.5090406
```

After 2008, the correlation between credit score and borrower rate gets weaker, and the chart colors sort of flip horizontally. Now the borrower rate varies only slightly within the credit risk levels, but the credit score ranges within the levels much more. The definitive cut scores by credit score are no longer there, but borrower rate is more closely tied to credit grade. Also, apparently Prosper stopped making loans to people with credit scores below about 575.

Percentage of Loans by Loan Status:

```
##                            N Percent
## Cancelled                  5    0.00
## Chargedoff             11992   10.53
## Completed              38074   33.42
## Current                56576   49.66
## Defaulted               5018    4.40
## FinalPaymentInProgress   205    0.18
## Past Due (>120 days)      16    0.01
## Past Due (1-15 days)     806    0.71
## Past Due (16-30 days)    265    0.23
## Past Due (31-60 days)    363    0.32
## Past Due (61-90 days)    313    0.27
## Past Due (91-120 days)   304    0.27
```

Overall, the percentage of chargeoffs is 10.53, with a default rate 4.4, totalling to 14.93% for all chargeoffs and defaults.

Percentage of Loans by Loan Status through 2008:

```
##                            N Percent
## Cancelled                  5    0.02
## Chargedoff              6650   22.97
## Completed              18288   63.16
## Current                    0    0.00
## Defaulted               4010   13.85
## FinalPaymentInProgress     0    0.00
## Past Due (>120 days)       0    0.00
## Past Due (1-15 days)       0    0.00
## Past Due (16-30 days)      0    0.00
## Past Due (31-60 days)      0    0.00
## Past Due (61-90 days)      0    0.00
## Past Due (91-120 days)     0    0.00
```

Through 2008, the percentage of chargedoffs was quite high, 22.97%, and those defaulted was 13.85%.

Percentage of Loans by Loan Status after 2008:

```
##                            N Percent
## Cancelled                  0    0.00
## Chargedoff              5342    6.29
## Completed              19786   23.28
## Current                56576   66.57
## Defaulted               1008    1.19
## FinalPaymentInProgress   205    0.24
## Past Due (>120 days)      16    0.02
## Past Due (1-15 days)     806    0.95
## Past Due (16-30 days)    265    0.31
## Past Due (31-60 days)    363    0.43
## Past Due (61-90 days)    313    0.37
## Past Due (91-120 days)   304    0.36
```

After 2008, the percentage of chargeoffs is much lower, 6.29%, with the percentage of loans in default at 1.19%.

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?

The difference in how Prosper determined credit risk up through 2008 vs. after 2008 became quite clear in this section. Several of the plots highlight clear delineations in credit score when compared across credit grade through 2008, but that delineation disappears after 2008. Furthermore, it appears that Prosper's assessment of credit risk improved when it made the change in 2009. The percentage of loans that end up being chargedoff becomes significantly less.

Were there any interesting or surprising interactions between features?

The most striking figures in this section are the plots that show clear delineation lines up through 2008 for Prosper's credit risk assessment. Credit scores seem to be the primary determination for assignment of a loan to a particular credit risk level. However, after 2008, the delineation lines disappear.

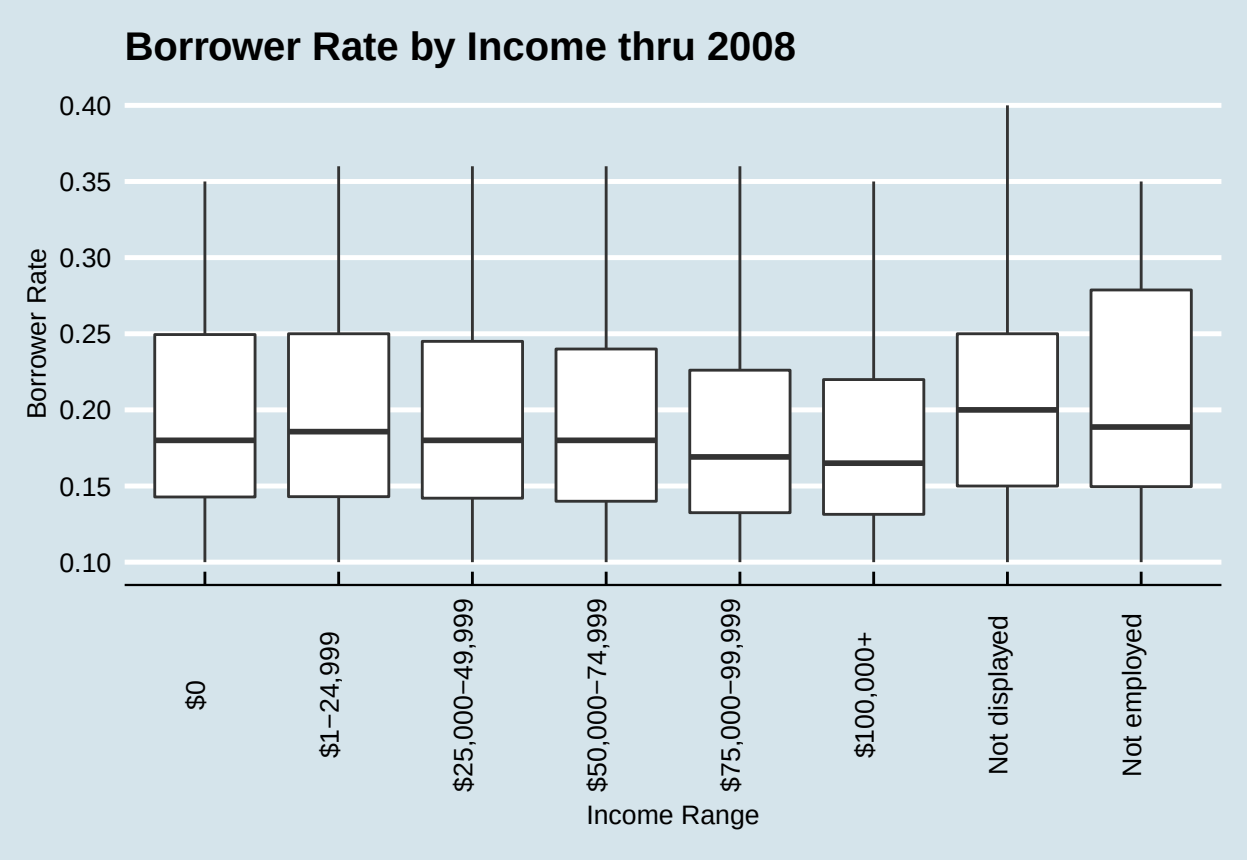
The colors in the scatterplots comparing borrower rate to credit score, which are colored by Prosper's credit grade, are particularly striking. It seems clear that through 2008, Prosper assigned cut scores for the different credit risk levels based on credit score. However, after 2008, they developed a more nuanced approach, taking more variables into account, and thereby improving their credit risk assessment.

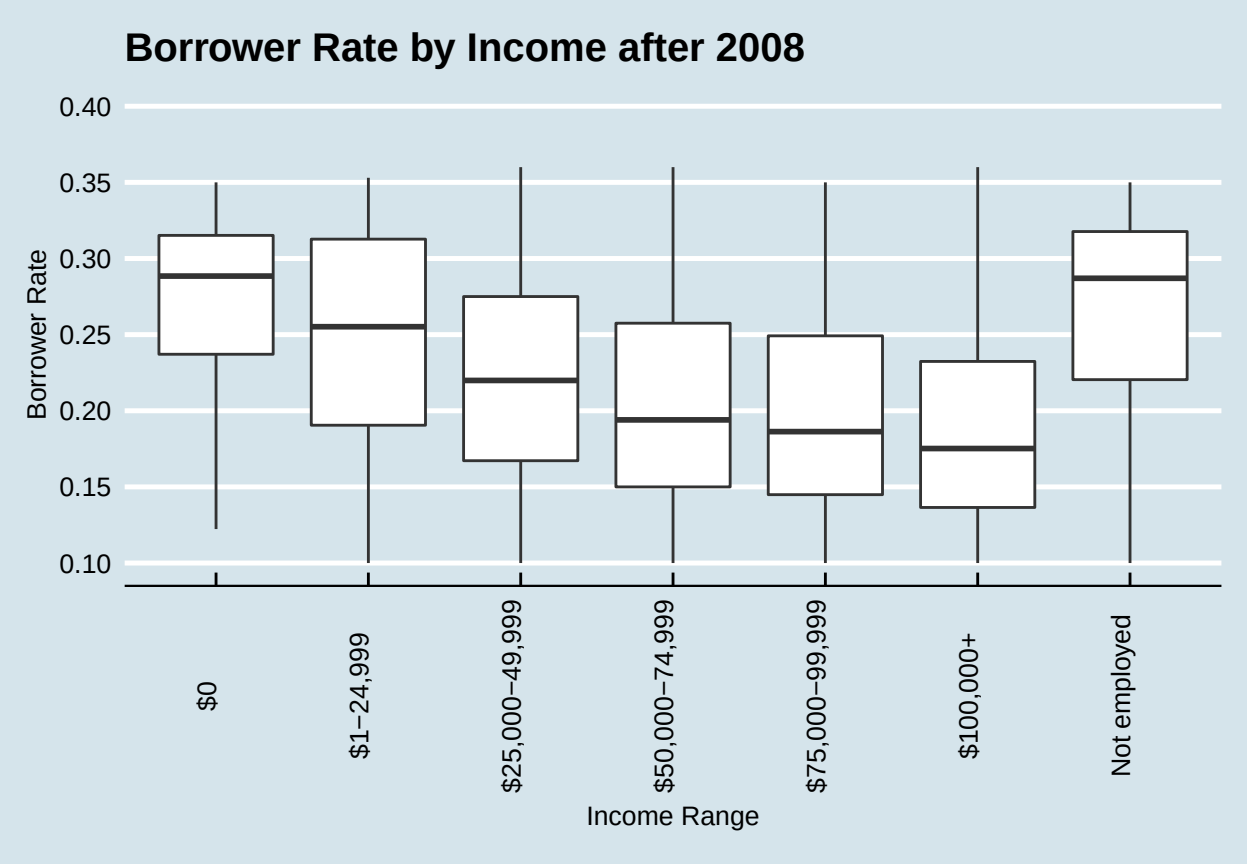
OPTIONAL: Did you create any models with your dataset? Discuss the strengths and limitations of your model.

I did not create any models with my dataset.

Final Plots and Summary

Plot One





Description One

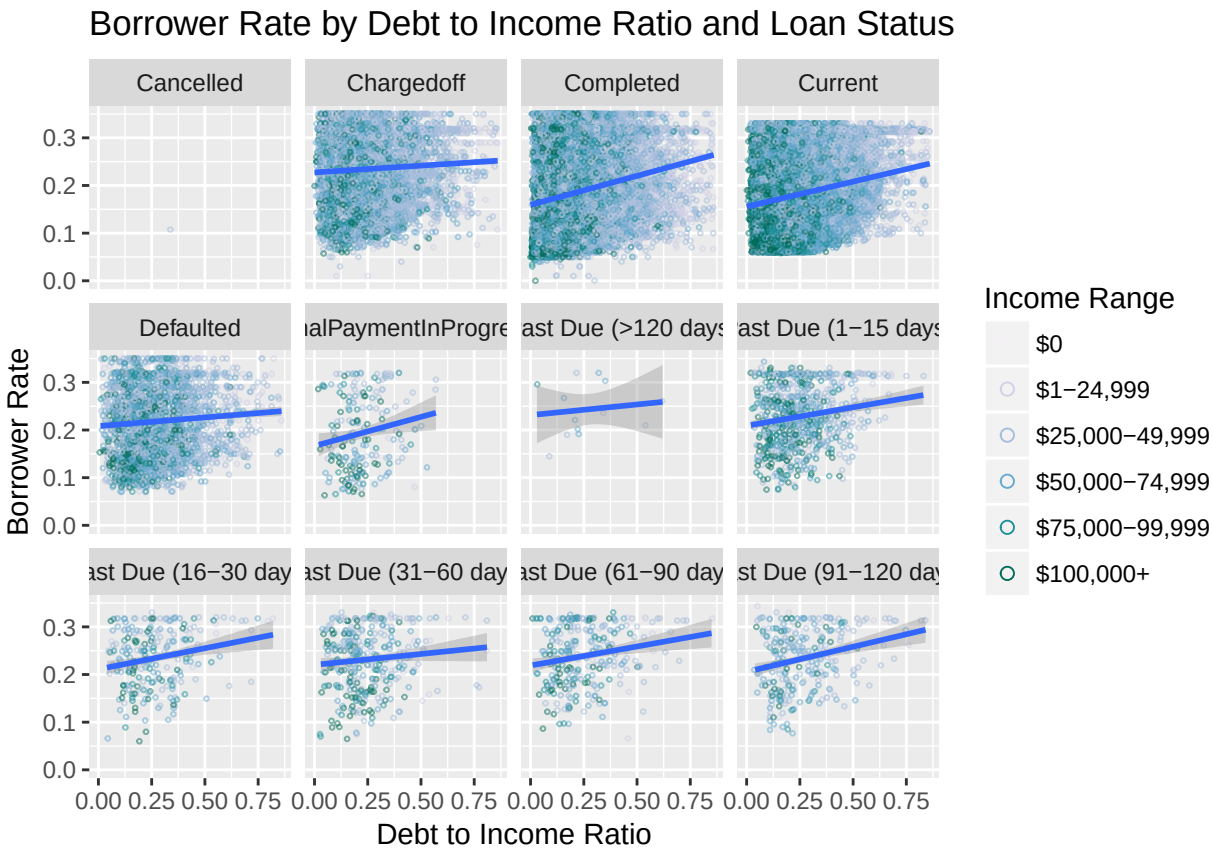
These plots show the difference in borrower rate across income range through 2008 and after 2008. There was very little difference in median borrower rate across income ranges through 2008. After 2008, we see a change, with a steady decline in median borrower as income increases.

Plot Two

Description Two

These charts look at borrower rate by credit score colored by Prosper credit grade. The first chart is for data through 2008. The second one covers data since 2008. These charts clearly show the difference in Prosper assessment of credit risk beginning in 2009. In the first chart, clear delineation lines at certain credit scores mark the categories of credit risk as determined by Prosper. In the second chart, the colors sort of flip horizontally, indicating how the criteria for Prosper's assessment of credit risk changed after 2008. The credit grades include a much wider variety of credit scores, but the borrower rate is much less varied across credit grade. Furthermore, we see a slightly weaker correlation between credit score and borrower rate after 2008.

Plot Three



Description Three

This chart looks at borrower rate by debt-to-income-ratio, faceted across loan status and colored by income range, with regression lines for each loan status category. The likelihood that a particular loan will default was not a primary area of focus for my analysis. However, this chart is interesting because it suggests that the characteristics of loans that eventually default are widely varied. Neither borrower rate, nor debt-to-income-ratio, nor income range seem to be significant indicators of whether a loan will eventually default.

```
# Reflection

The Prosper Loan Dataset tracks nearly 114,000 loans across 81 variables. I began with a basic explorati
```

individual variables, and then I started to explore key questions, particularly how Prosper determines the borrower rate for a given a loan. I had quite a bit of difficulty finding clear correlations between borrowers' credit history and the rates they earned. I also was surprised to find that Prosper was willing to make a large number of loans to borrowers with poor credit histories. Over 5,000 loans were made to borrowers with more than 25 delinquencies in the past 7 years, and prior to 2009, some borrowers had credit scores in the 400's.

Eventually, I discovered that Prosper assesses the credit risk of each loan based on a proprietary method and assigns borrower rate according to that assessment. Furthermore, I discovered that Prosper changed its method of assessment beginning in 2009. Prior to 2009, credit risk was tied very closely with credit score, but that turned out to be a relatively inaccurate assessment of risk. Beginning in 2009, it appears Prosper

added a number of other variables to the credit risk calculation and as a result seems to have improved their assessment. The percentage of chargeoffs and defaults after 2008 decreased significantly.

In future study, it would be interesting to see if the improved credit risk assessment holds true over time. It also would be interesting to track the rate of default over the life-cycle of the loan for the various credit risk levels, i.e., to use the data to gauge at what point in the loan cycle the defaults tend to occur. Finally, future exploration into common characteristics of loans that eventually default could help improve the overall assessment of credit risk.