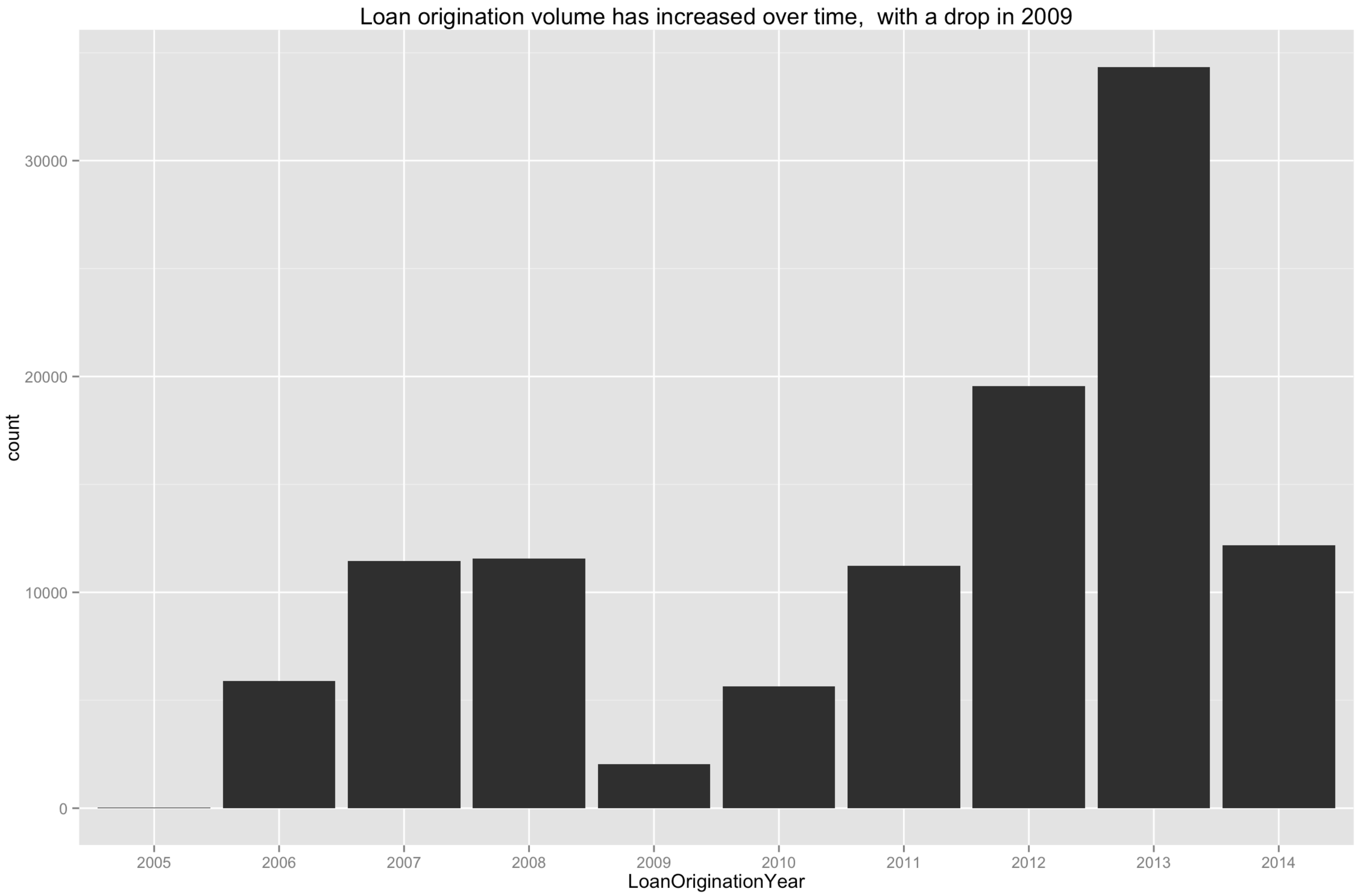
Loan Data Exploration by Eddy Shyu

Univariate Plots Section

Data subsetting According to the variable descriptions, many fields are only populated for a loanOriginationDate after July 2009. Consider looking at just data where these fields exist. First see distribution of data by date I can also spit the LoanOriginationDate into LoanOriginationYear, LoanOriginationMonth (I will ignore day and time)

The number of loan originations increases over time, with a drop in 2009 to 2010 after the financial crisis

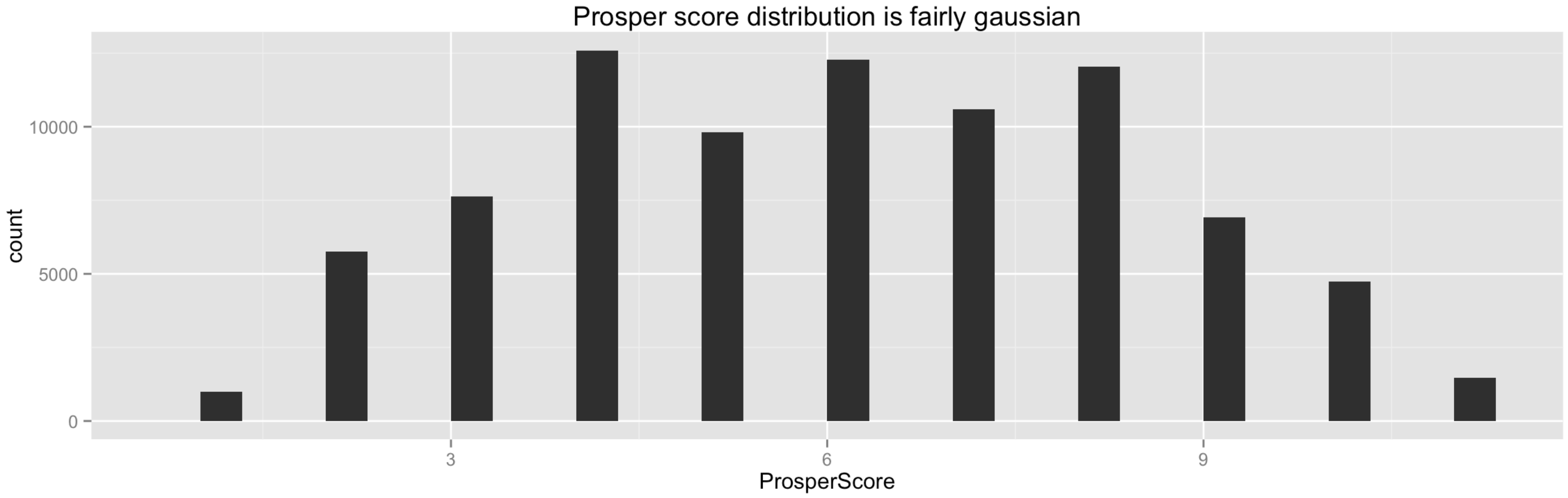


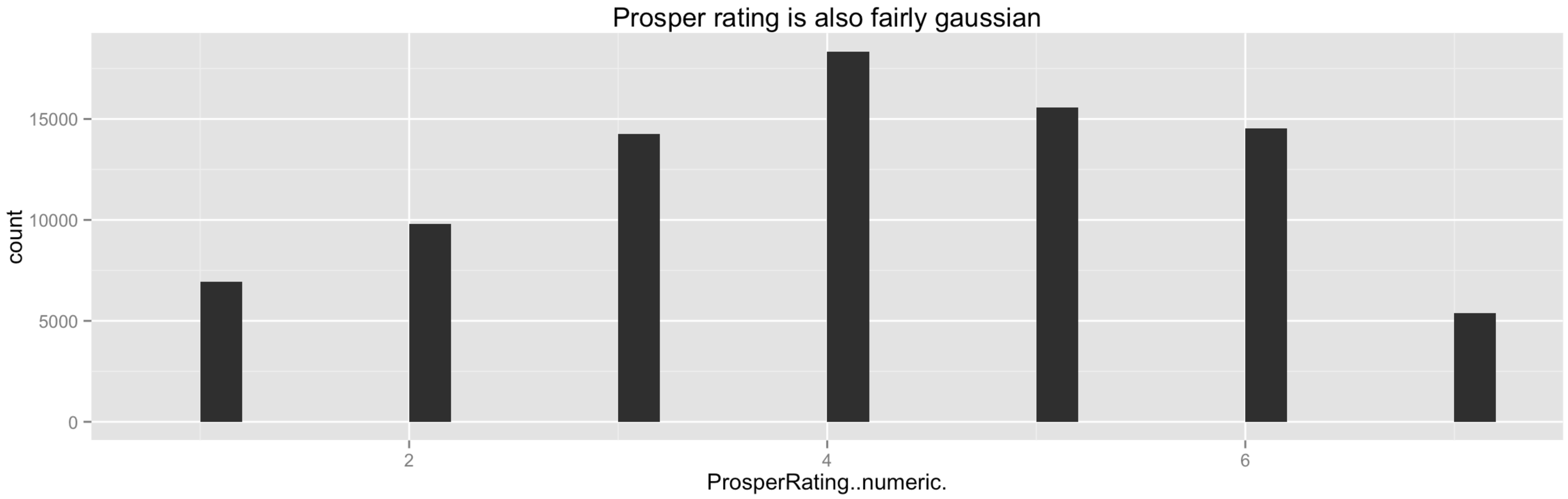
Possible independent variables

I'll first look at variables that would be known before the loan origination, so these are possible independent variables.

PropserScore and ProsperRating..numeric.

The ratings look gaussian.



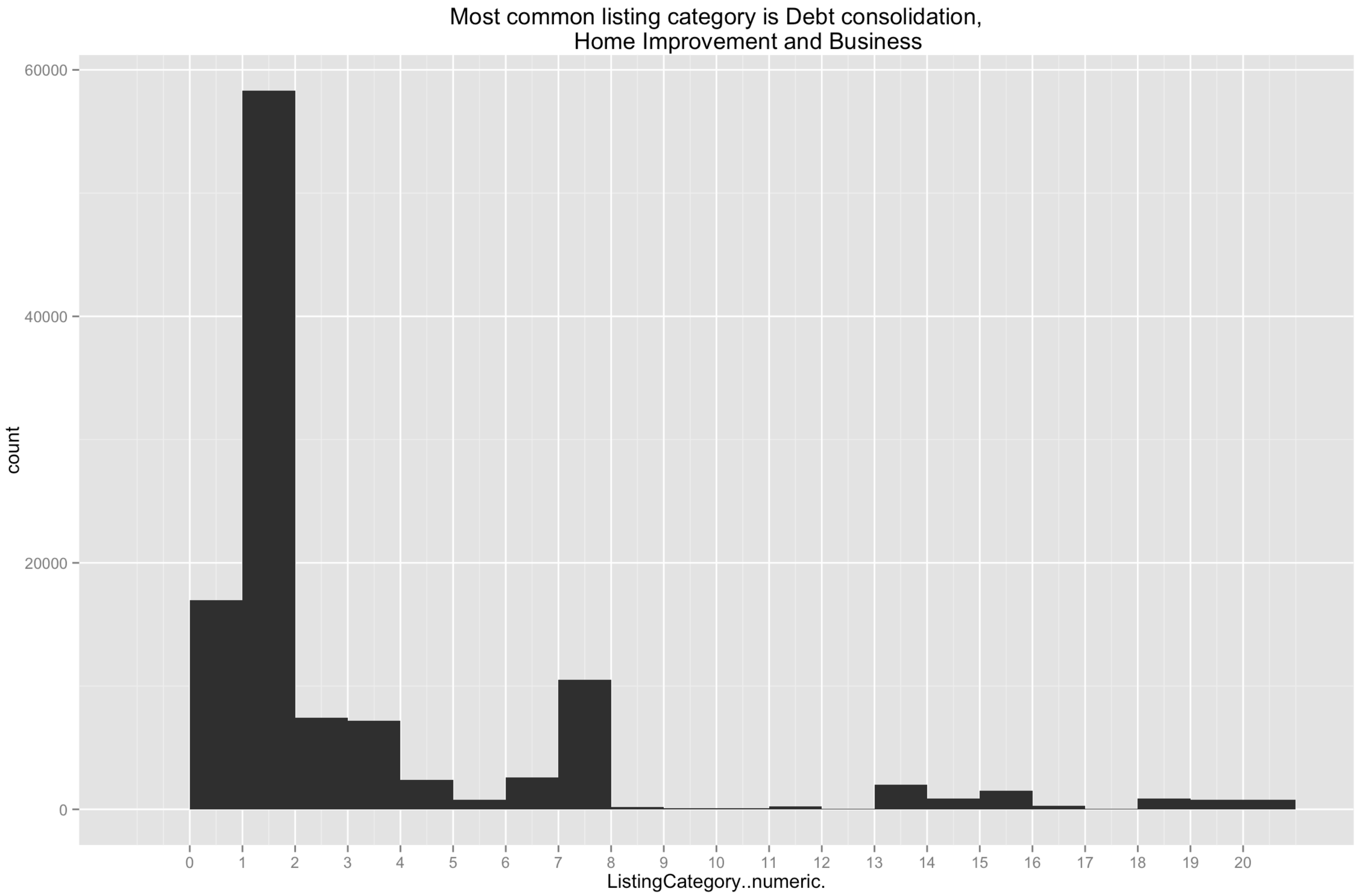


Listing Category

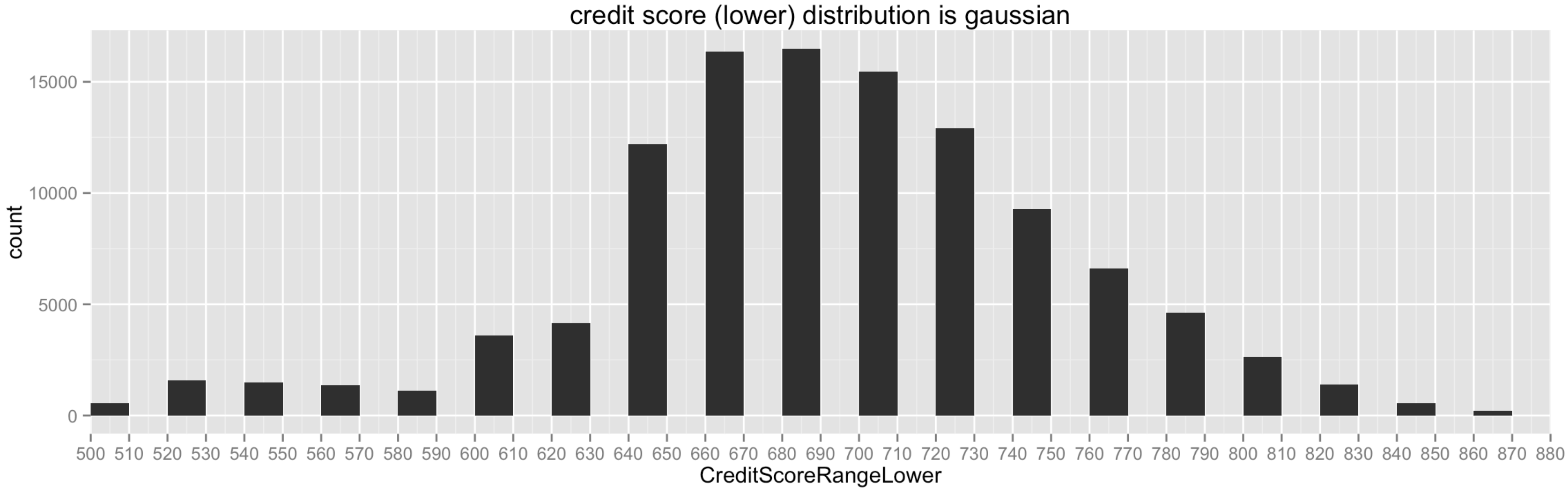
The most common category by far is 1 - Debt Consolidation, followed by 0 - Not Available, 7 - Other, 2 - Home Improvement, 3 - Business.

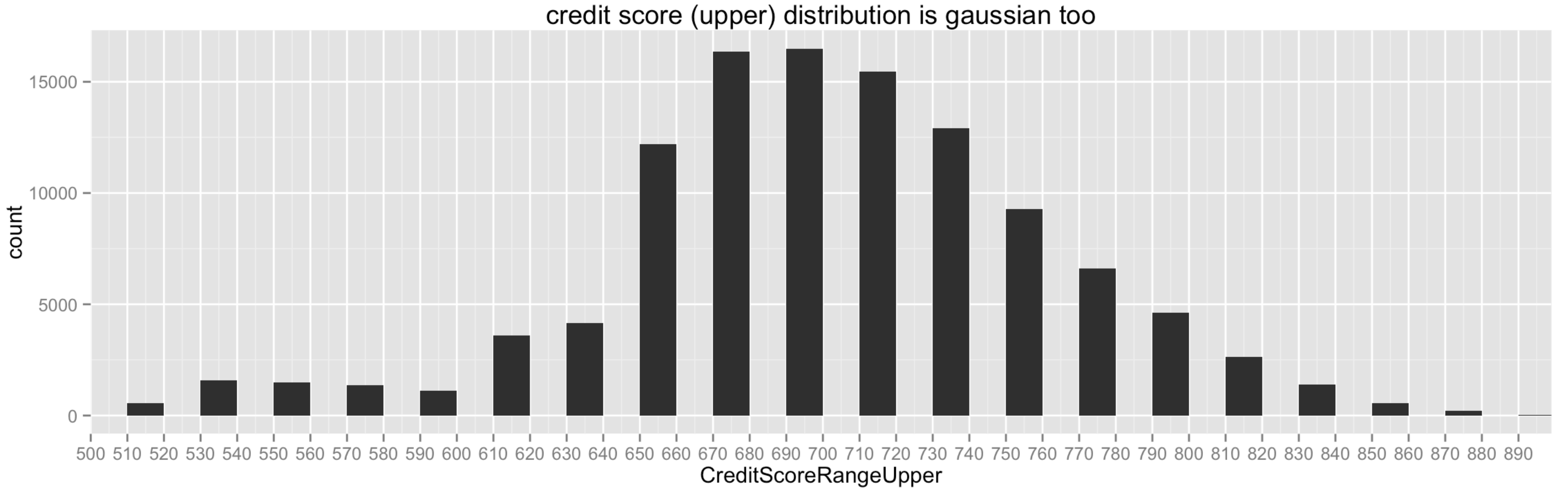
The category of the listing that the borrower selected when posting their listing: 0 - Not Available, 1 - Debt Consolidation, 2 - Home Improvement, 3 - Business, 4 - Personal Loan, 5 - Student Use, 6 - Auto, 7- Other, 8 - Baby&Adoption, 9 - Boat, 10 - Cosmetic Procedure, 11 - Engagement Ring, 12 - Green Loans, 13 -

Household Expenses, 14 - Large Purchases, 15 - Medical/Dental, 16 - Motorcycle, 17 - RV, 18 - Taxes, 19 - Vacation, 20 - Wedding Loans



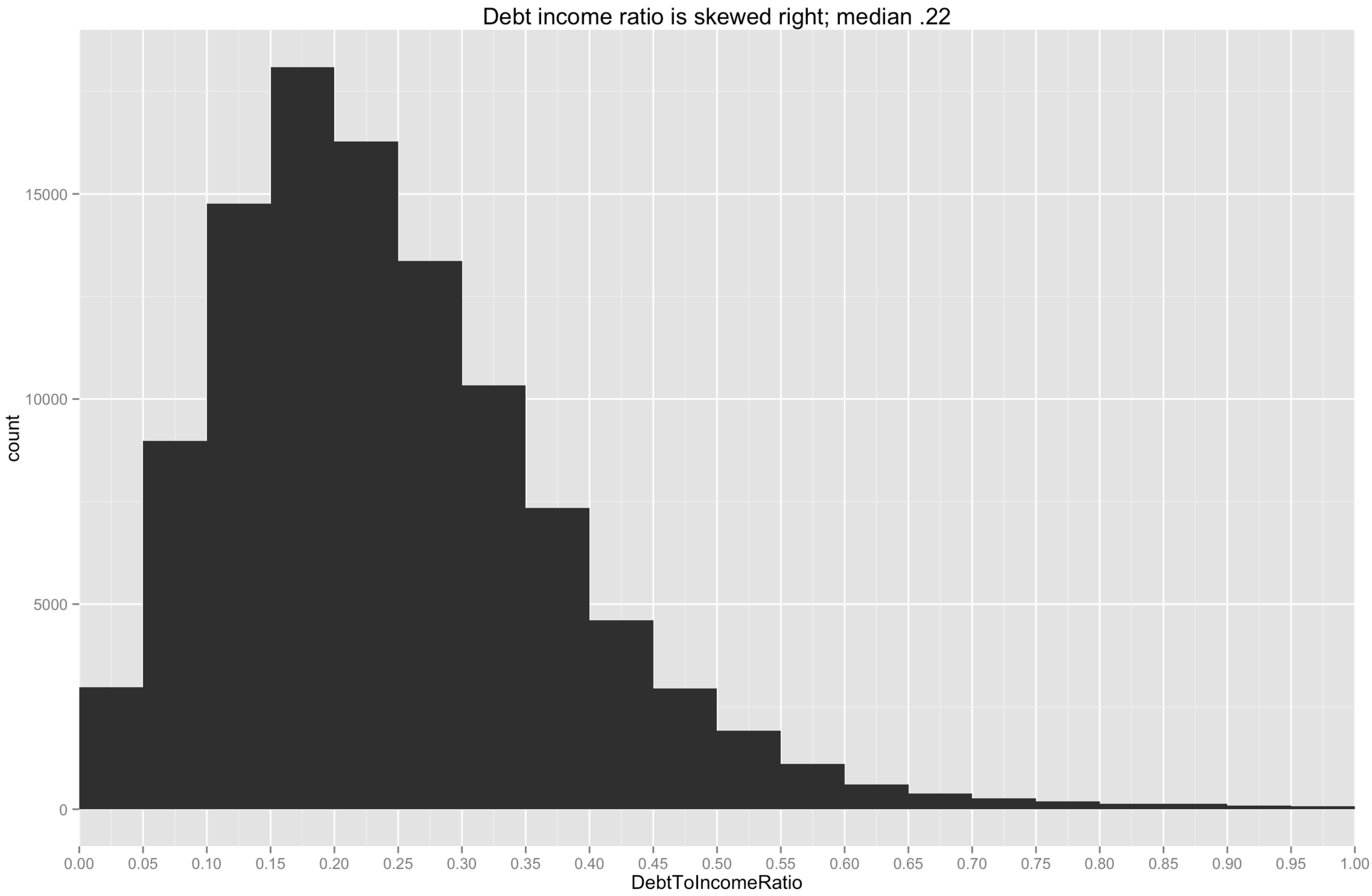
Credit Score Distribution looks gaussian, with most data between 500 and the max of 880 (for range lower), and 510 to 80 for range uper. There is an outlier of 0, which we may want to remove.





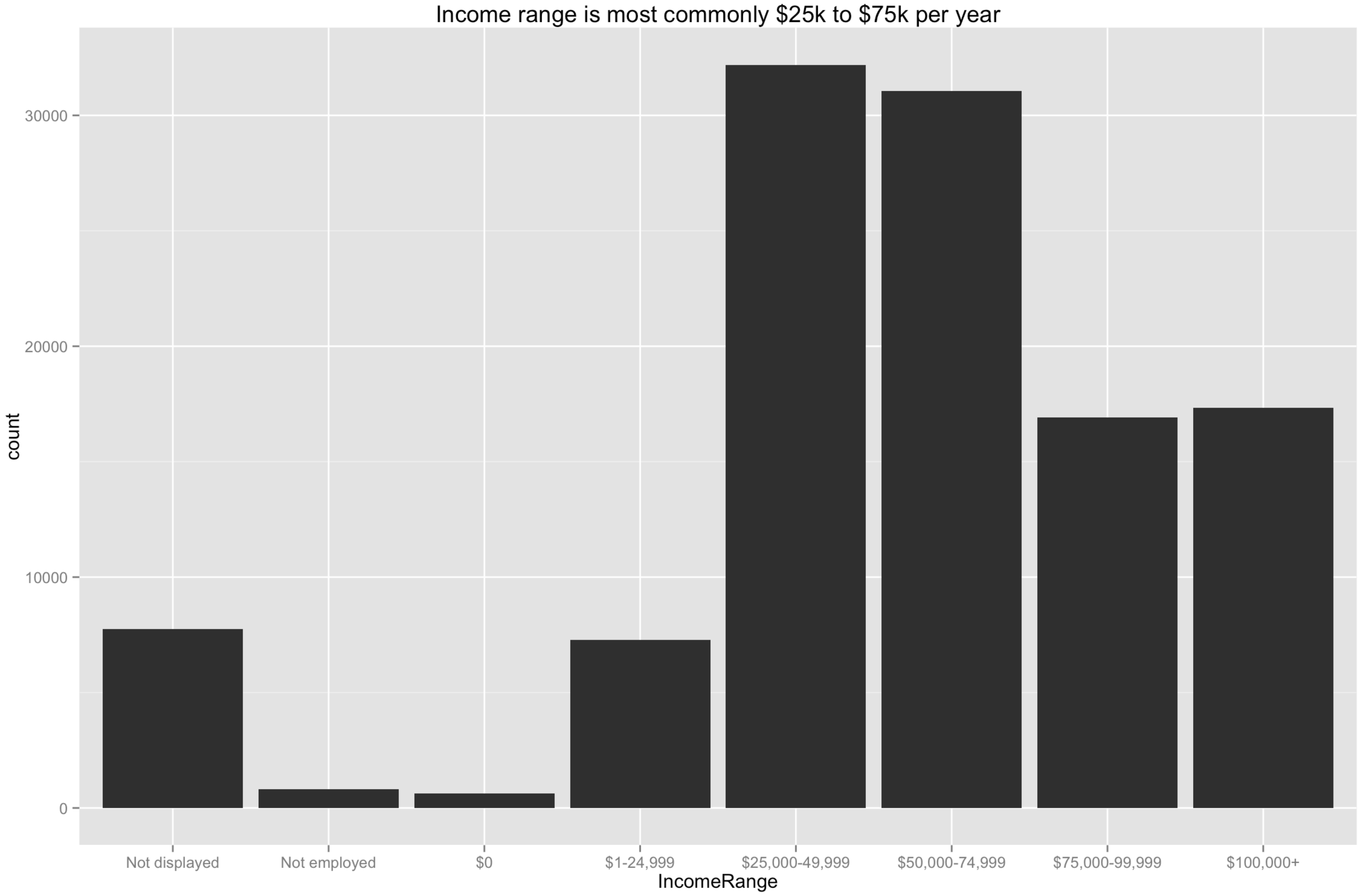
DebtToIncomeRatio

There is one outlier of 10+, but the rest are less than 1. Distribution is skewed to the right (more extreme values are high debt to income ratios). Most ratios are around the median of .22



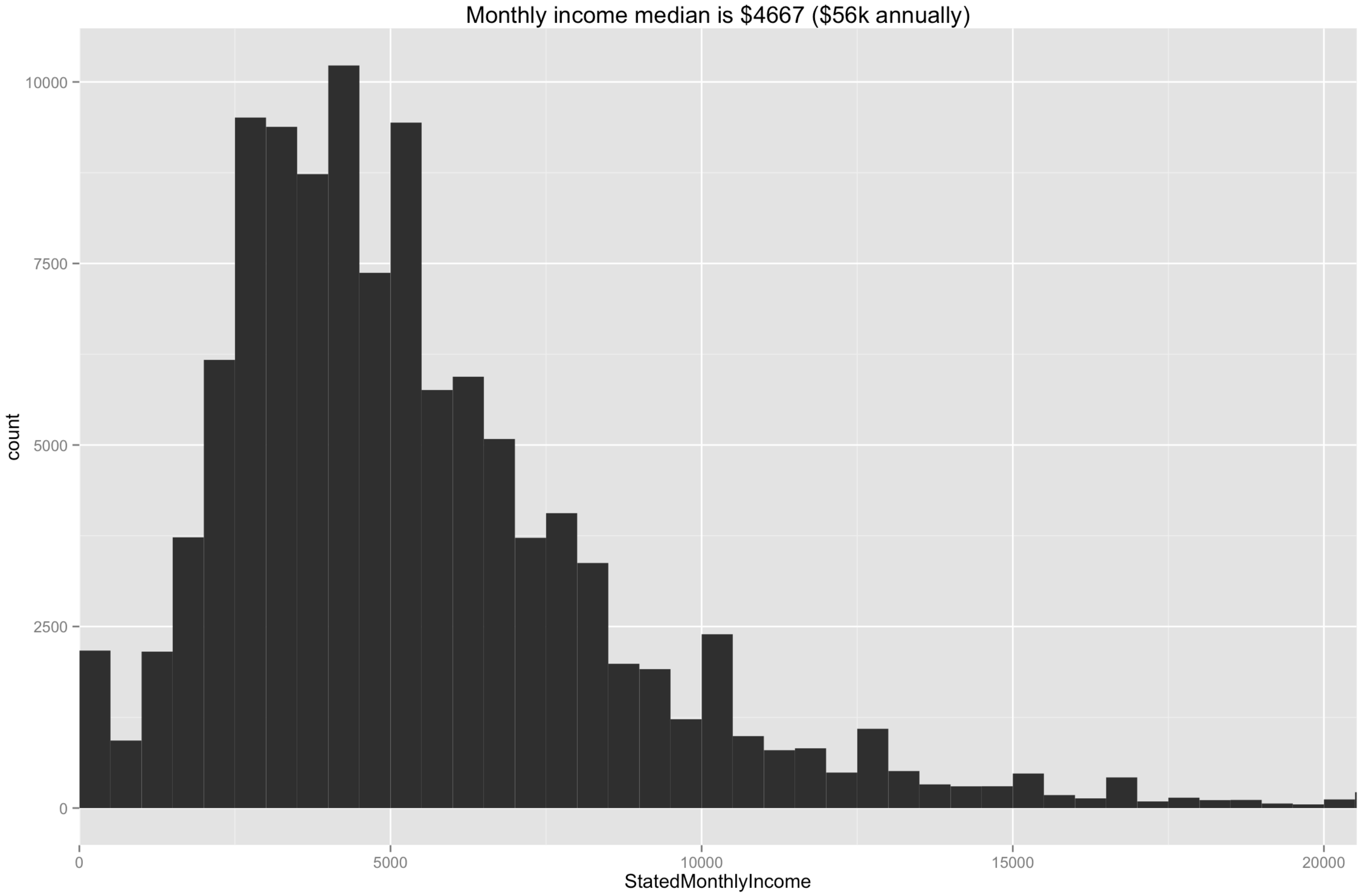
IncomeRange

Most common inomes are between 25k to 50k, and 50k to 75k.



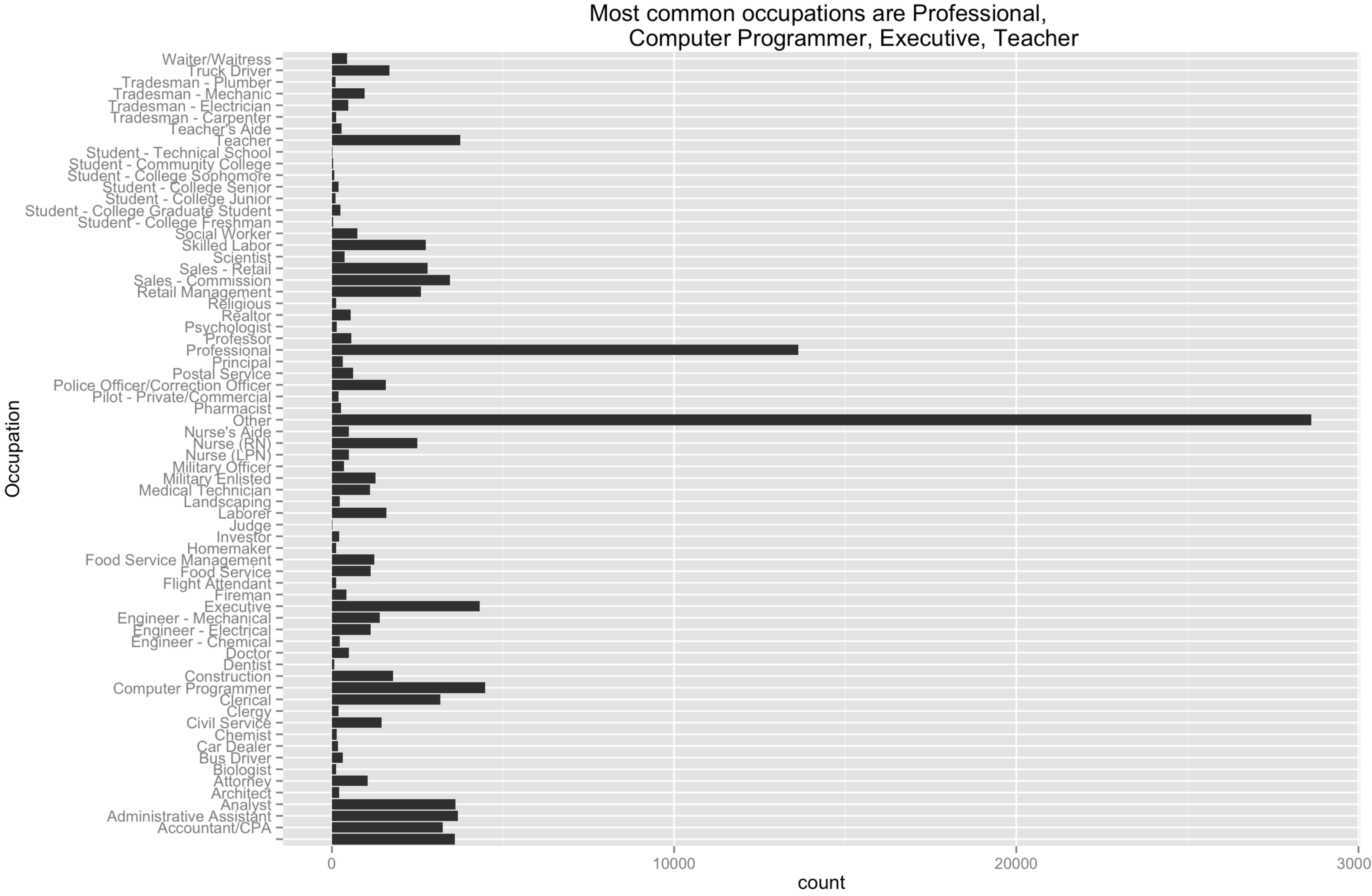
StatedMonthlyIncome

Might be more specific than IncomeRange The distrubtion is skewed right, with median at 4667 and IQR of 3200 to 6825.



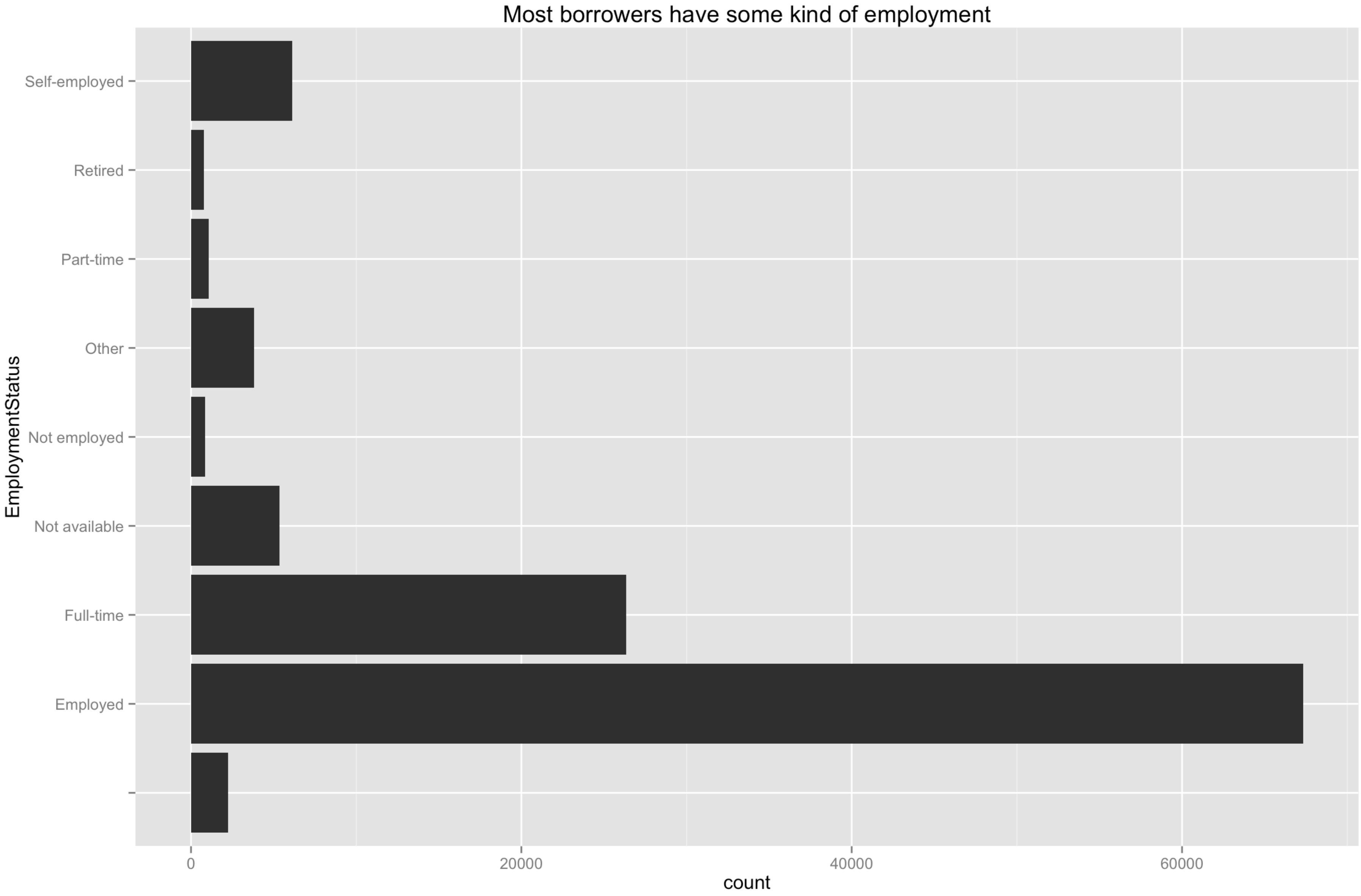
Occupation

This is related to income range, and income range might be a more useful predictor. Most common are ‘Other’ and Professional, followed by Computer Programmer, Executive, Teacher. Since the most common occupations are too general, it might be better to focus on income range.



EmploymentStatus

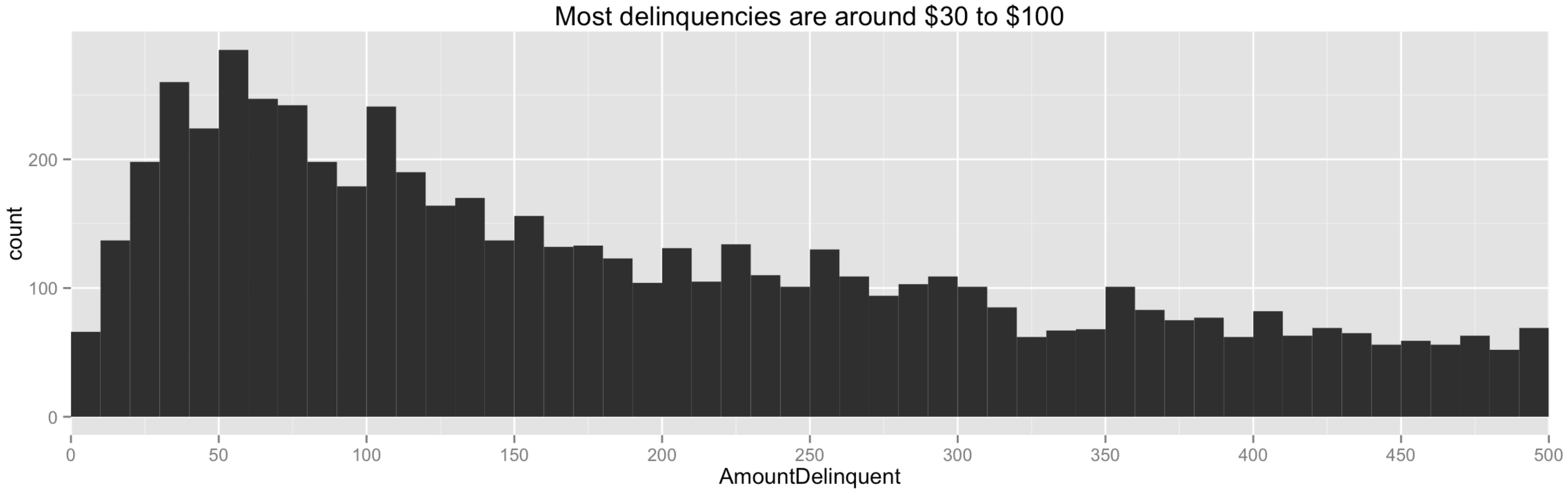
Most people are employed (Employed, Full-time, Self-employed, Part-time). Very few are “Not employed”

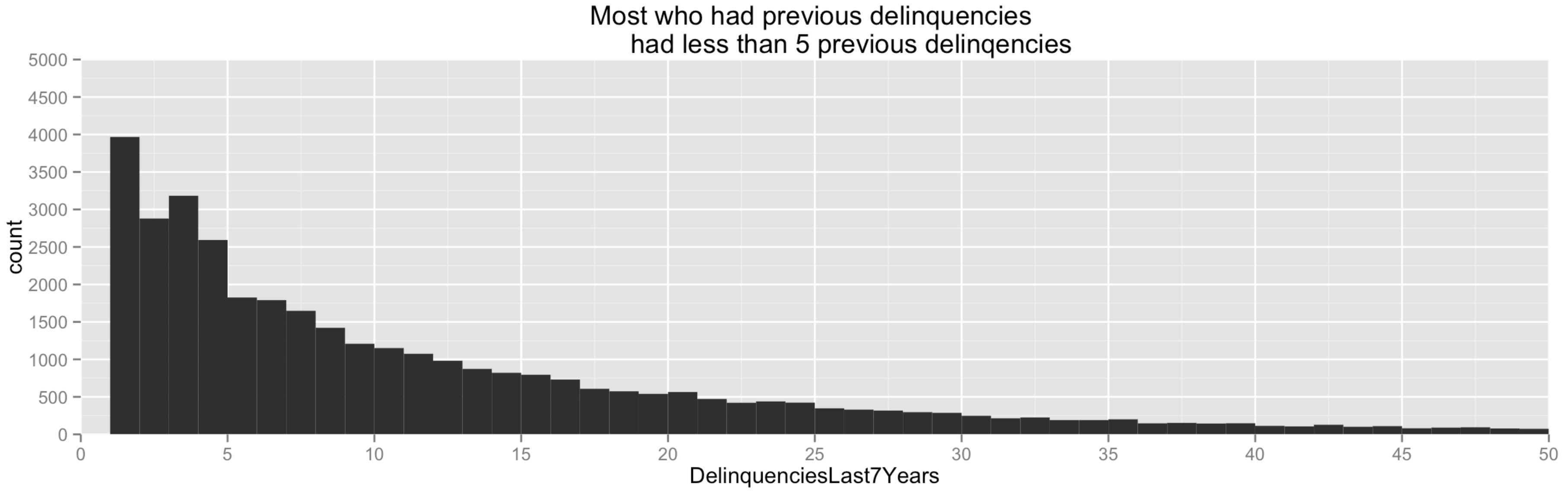


AmountDelinquent, DelinquenciesLast7Years

Most borrowers have zero or N/A delinquent amount at the time of application. So it does not apply to most records.

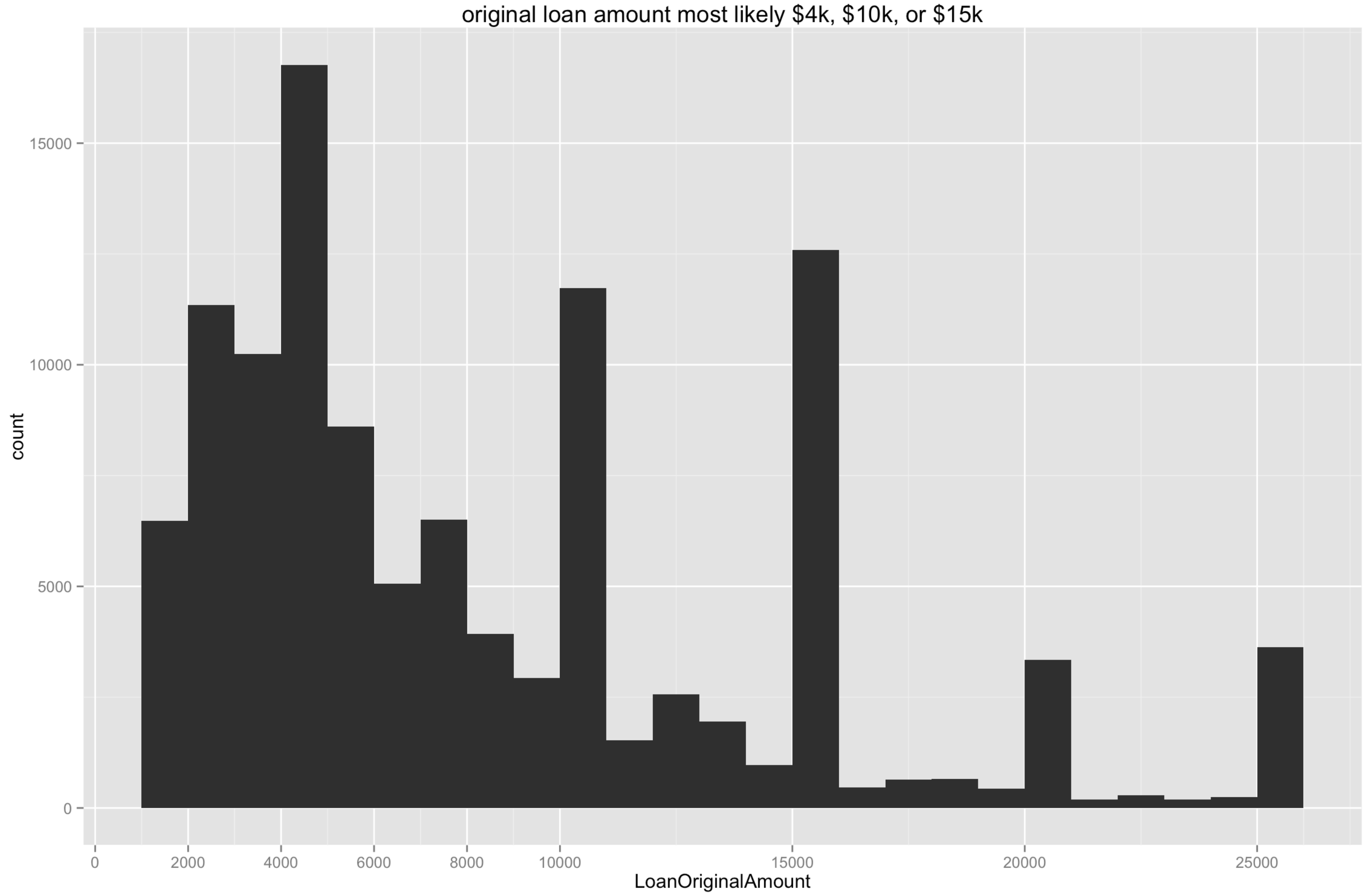
Most borrowers had zero delinquencies in past 7 years, but a couple thousand had 1 or more delinquencies.





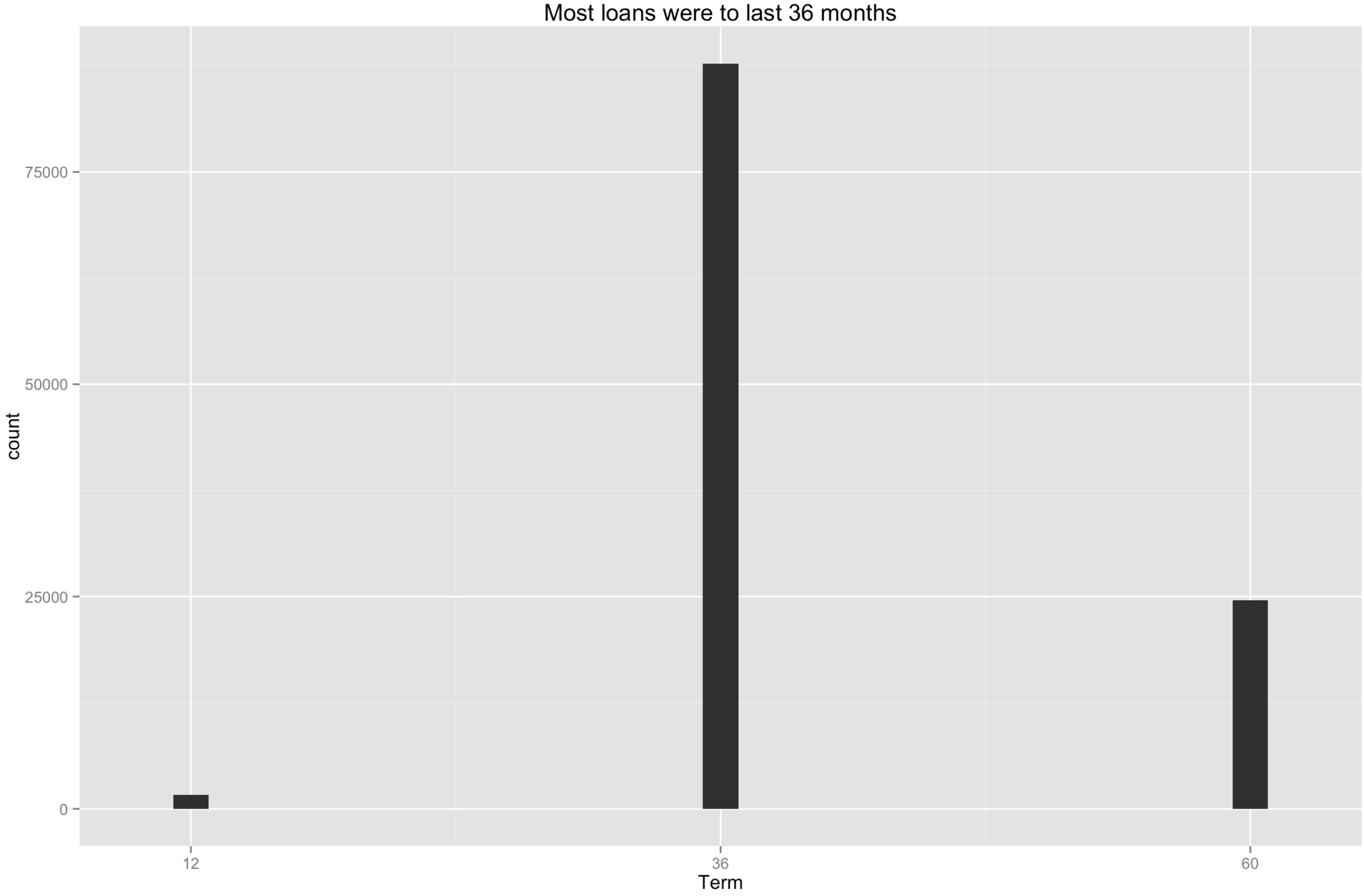
LoanOriginalAmount

The median loan is 6500, with highest values around 4k, 10k, 15k, and 20k. It seems that borrowers choose a clean number that rounds to one of these common values. They also round to the nearest thousand or 500.



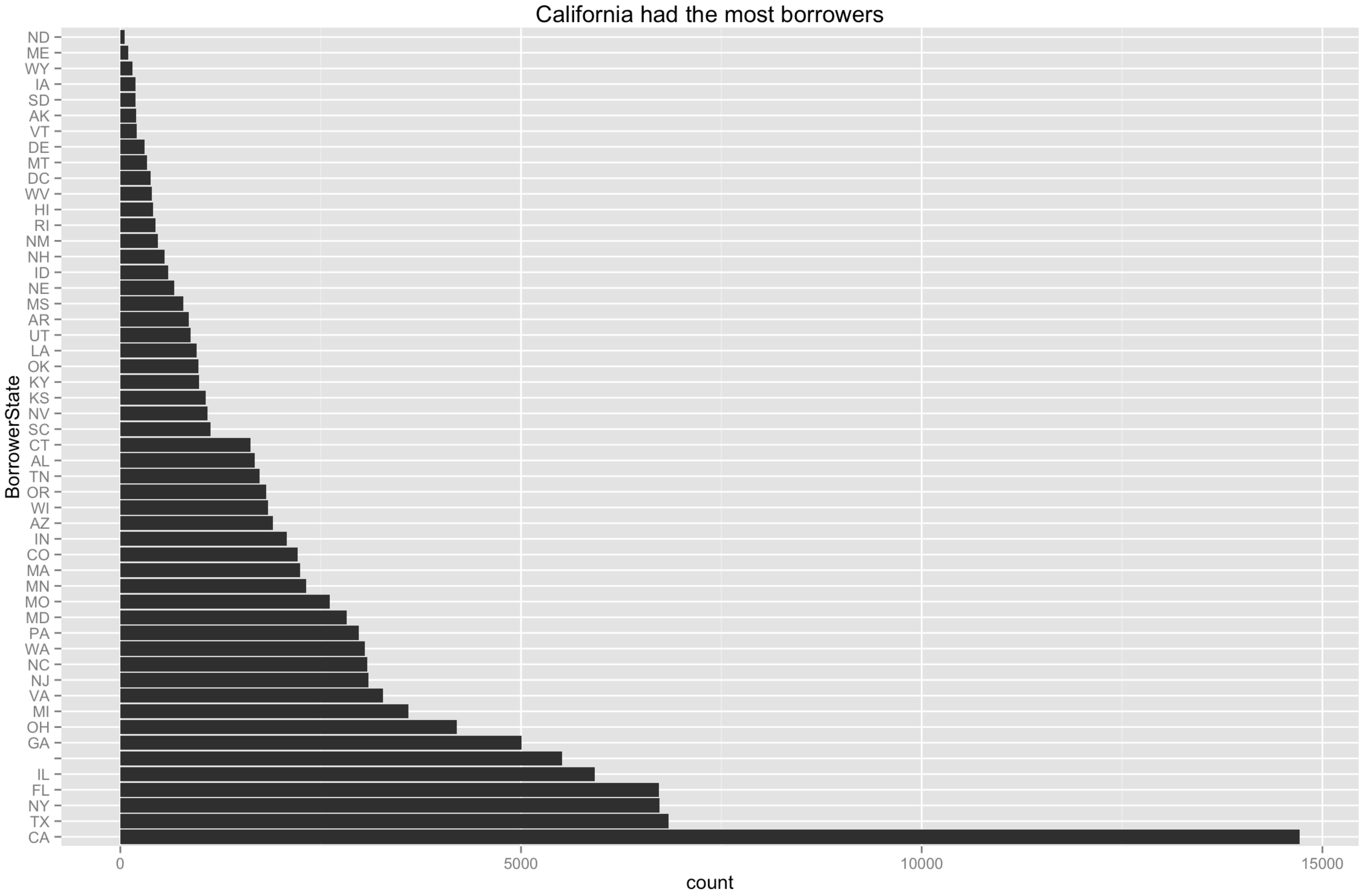
Term

Number of months to pay off the loan from start date Most loans are for 36 months (87,778), otherwise 60 months (24,545). Very few are for 12 months (1,614).



Borrower state

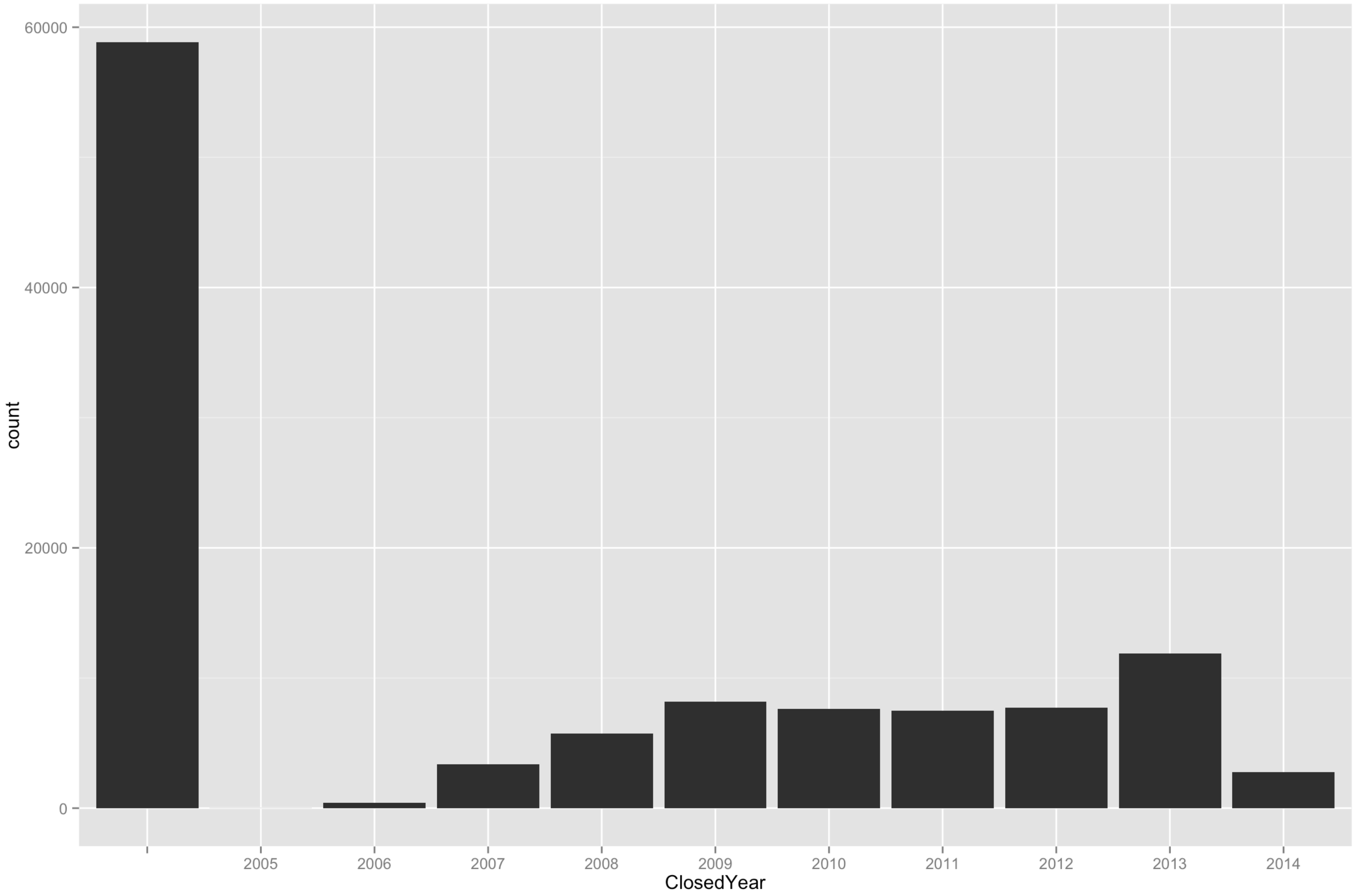
The states with large populations have the most loans (CA, NY, TX, FL, IL). I ordred states by number of observations per state. California has twice as many observations as the second state, Texas. About 5,500 observations do not have a state specified.



Variables after loan origination (possible dependent variables)

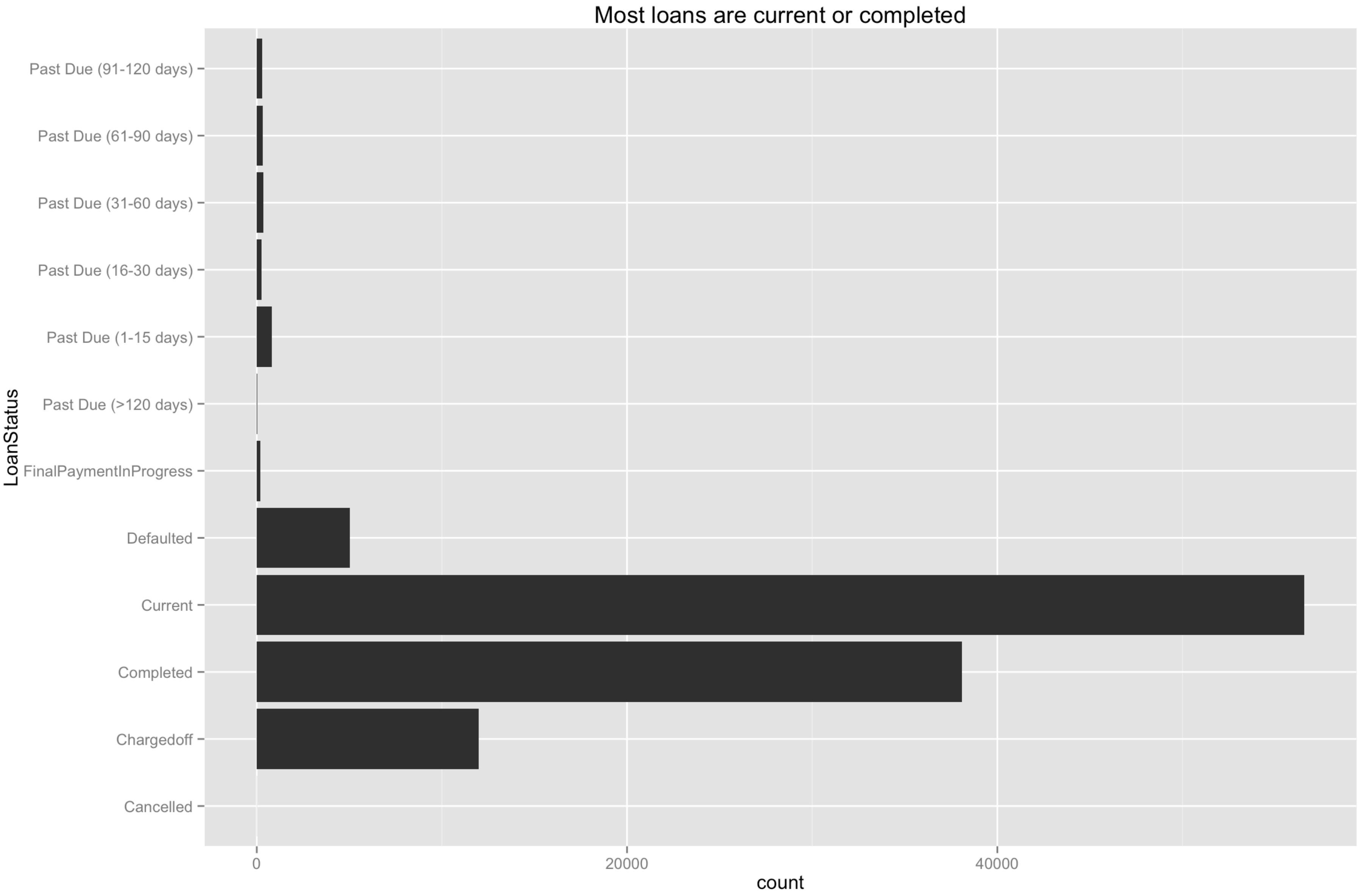
ClosedDate

If loans are still current, the value is NA. Most loans are still current.



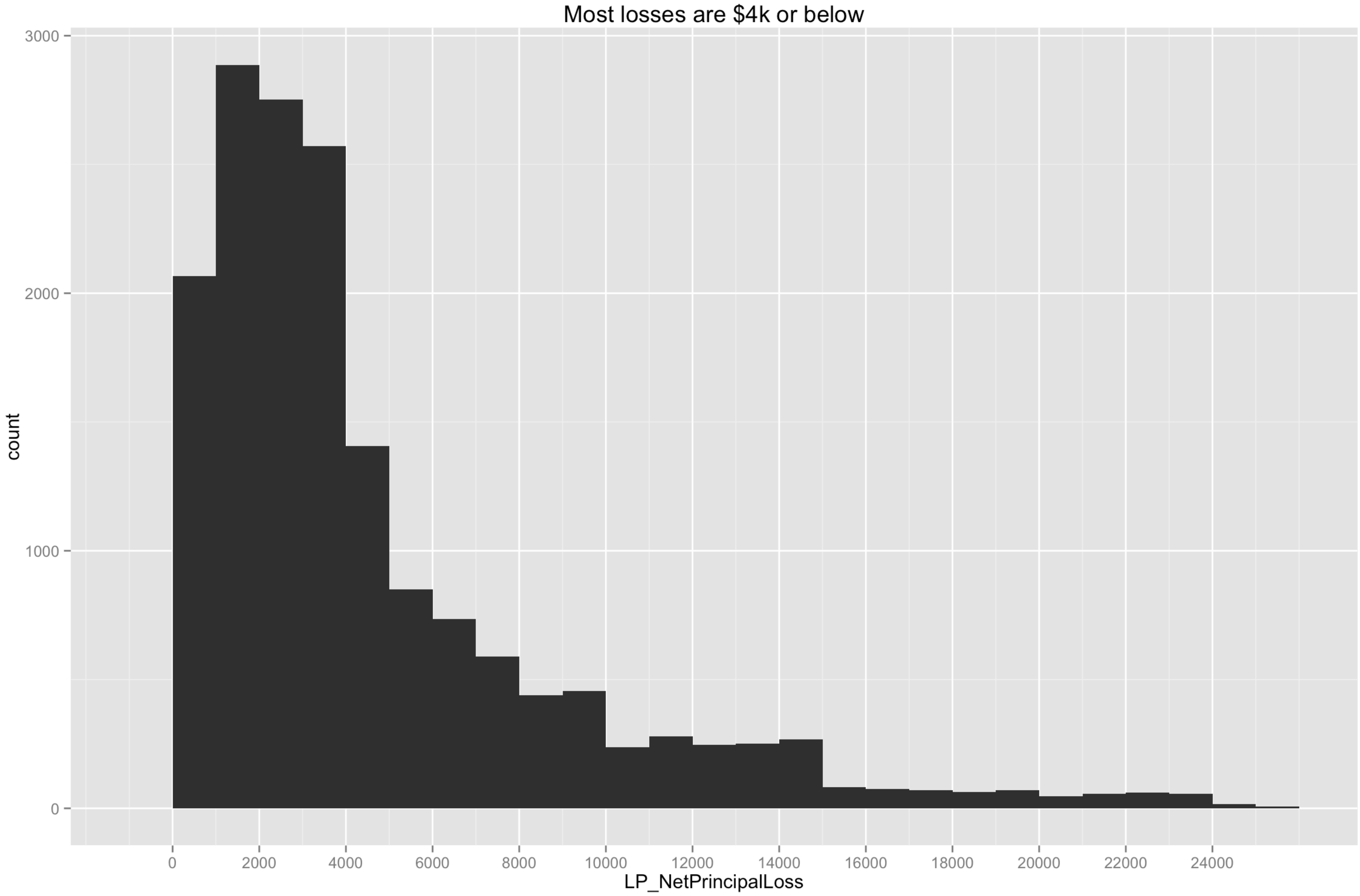
LoanStatus

Most loans are current. Many are completed or charged off.The past due loans are fairly evenly spread between the days past due (1-15, 16-30, 31-60, 61-90, 91-120)



LP_NetPrincipalLoss

The median net loss is zero. Subsetting on records where loss is > 0, most losses were between 1k to 4k.



Payment amounts

These relate to how much the borrower paid. I might want to show these as a fraction of the LoanOriginalAmount. LP_CustomerPayments : I guess this equals principle + interest and fees LP_CustomerPrincipalPayments, LP_InterestandFees, LP_ServiceFees, LP_CollectionFees, LP_GrossPrincipalLoss,LP_NetPrincipalLoss, LP_NonPrincipalRecoverypayments

Create variables for customer payments divided by loan original amount: LP_CustomerPayments_pc, LP_CustomerPrincipalPayments_pc, LP_InterestandFees_pc, LP_ServiceFees_pc, LP_CollectionFees_pc, LP_GrossPrincipalLoss_pc, LP_NetPrincipalLoss_pc, LP_NonPrincipalRecoverypayments_pc

Transform payment information as a fraction of the original loan amount

Create a subset of data where loan is no longer active. We can either find records where loan status is cancelld, chargedoff, completed, or defaulted; equivalently, I can find records where the ClosedDate (I separated into ClosedYear, ClosedMonth) is not NA and not the empty string. Both give the same result. This has 55,0089 observations.

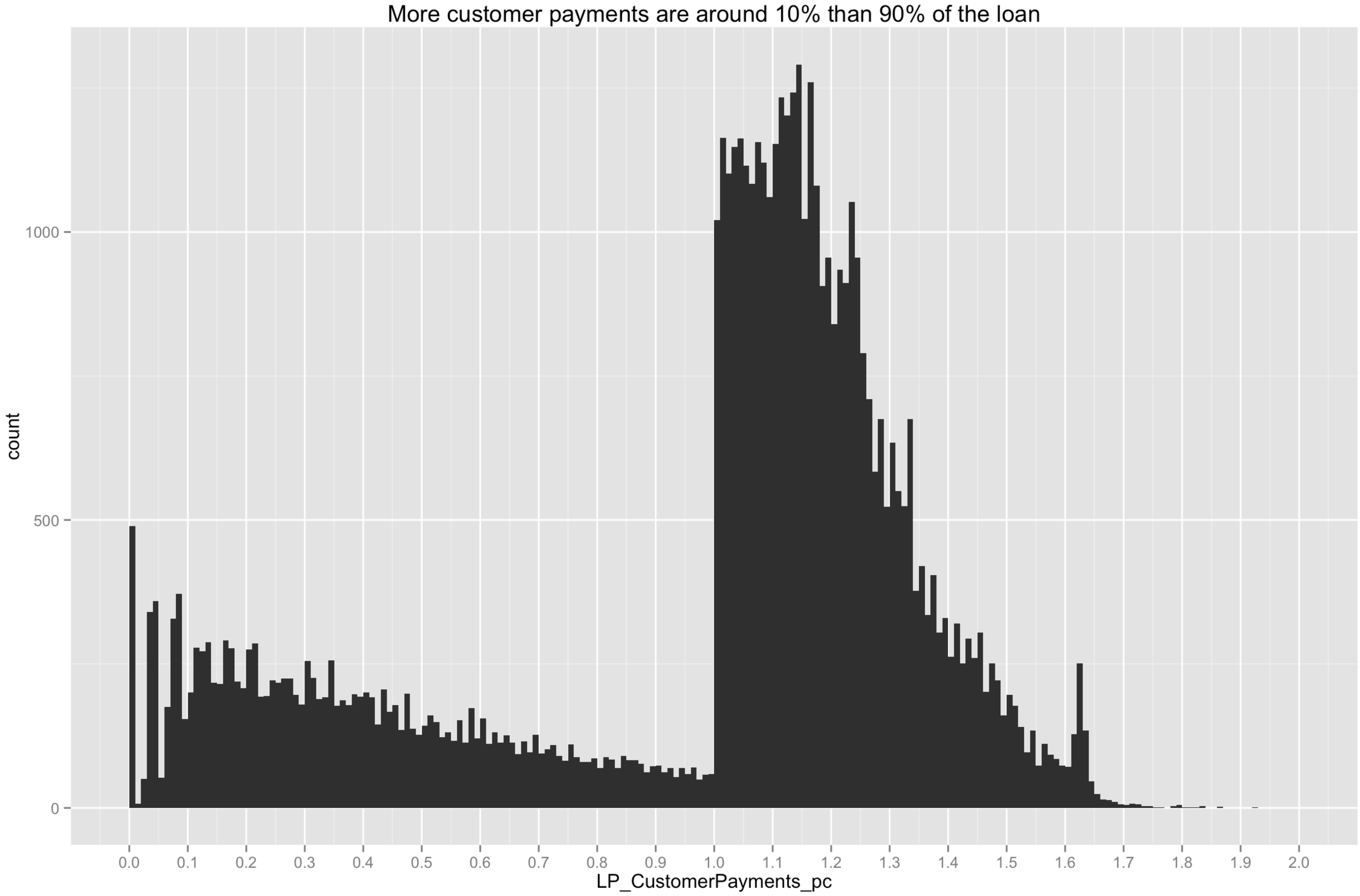
Further subset the data to include when “ProsperRating..numeric.” and other variables are not NA nor empty srings. These include EstimatedReturn, EstimatedLoss, EstimatedEffectiveYield, ProsperScore, “ProsperRating..alpha.” This has 26,005 observations

Look at completed loans only

When looking at customer payments, look at just the subset of loans that are completed, because in-progress loans will show only payments up to the point when data was collected, and will make the results hard to compare.

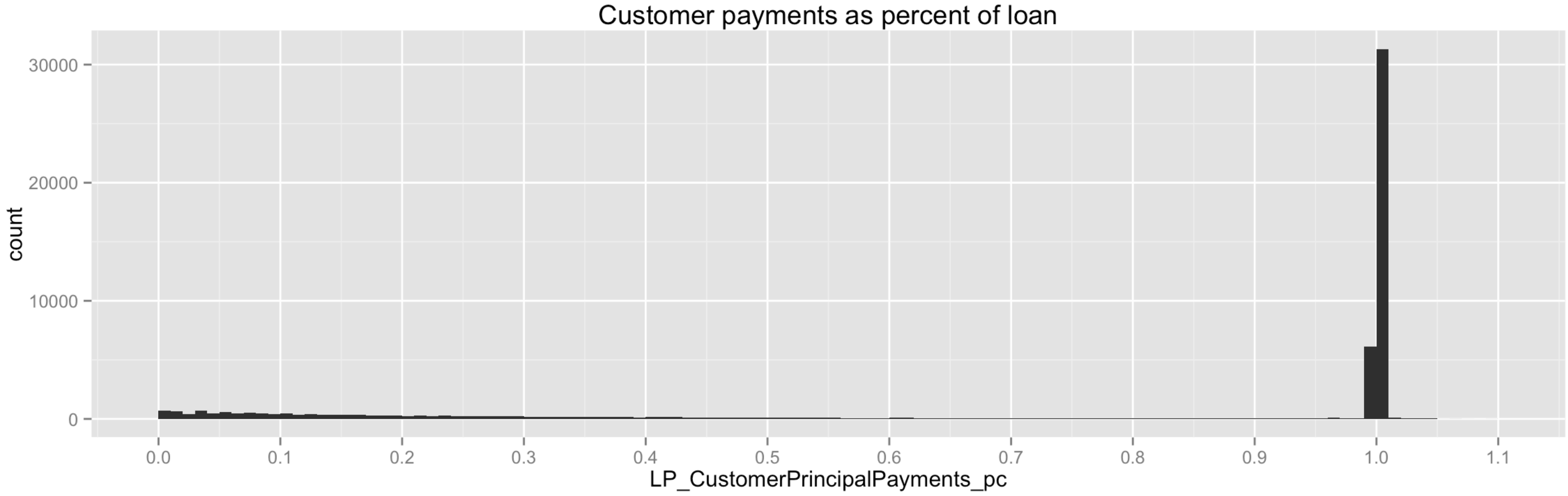
Customer payments / Loan original amount show that for those that paid less than the original amount borrowed, they tended to pay 20% or less of what they borrowed. It was less likely for someone to pay 50% to 99% and then default.

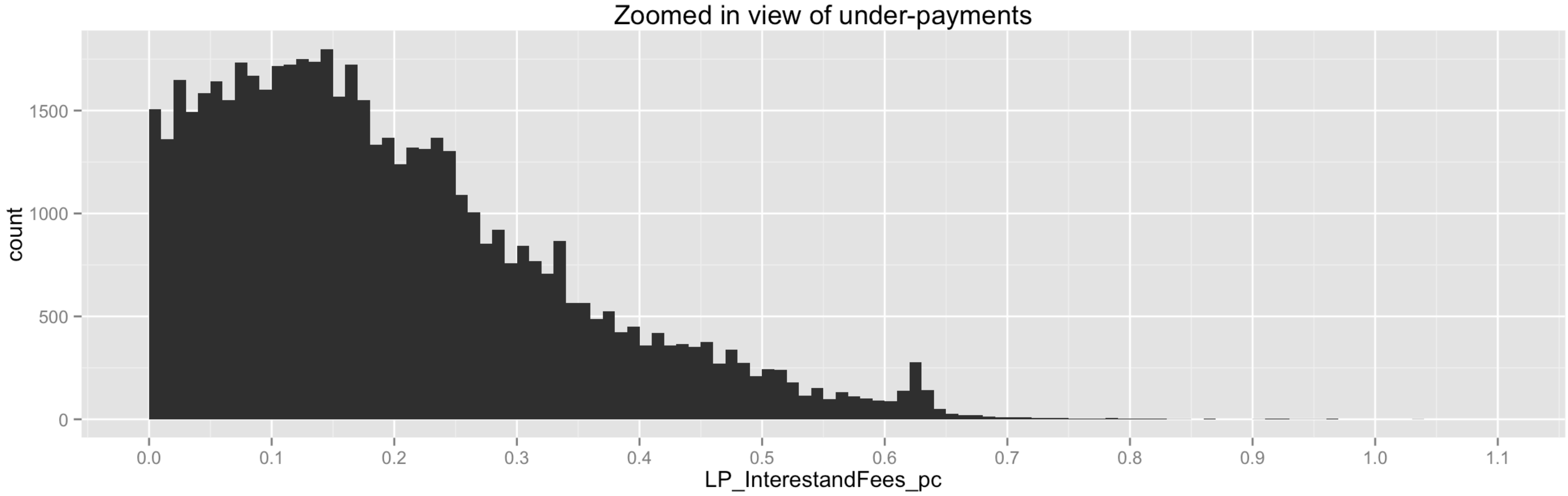
For those who paid at least the amount borrowed, most paid between 1 and 1.2 times the amount borrowed.



PrincipalPayments and Interest and Fees

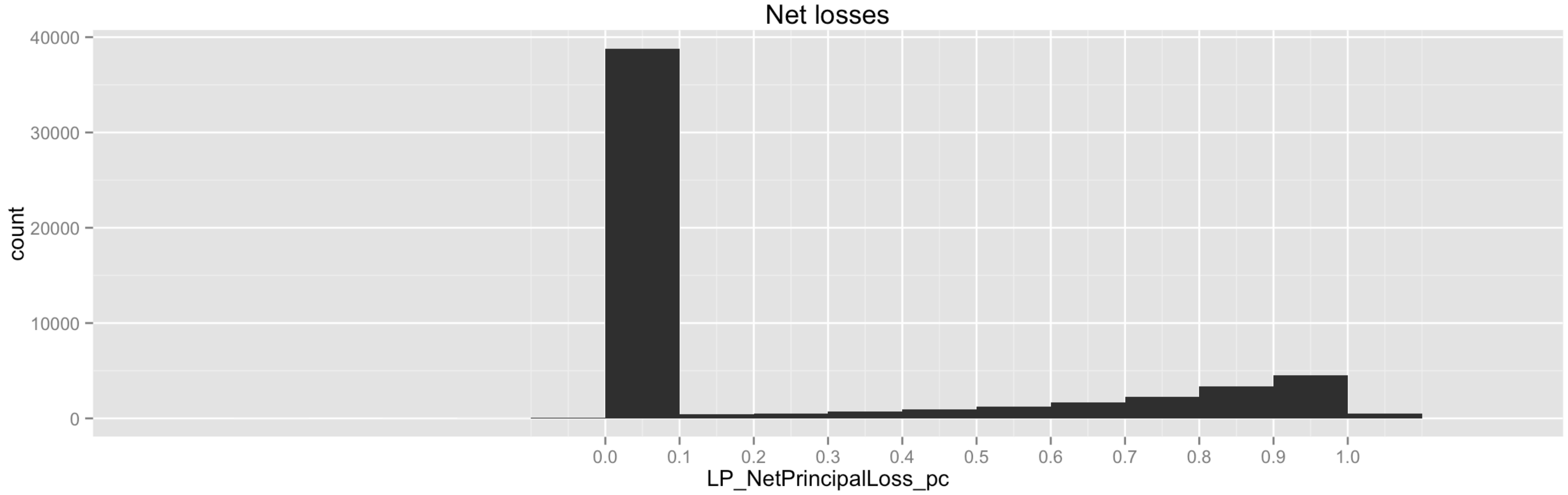
Principal payments are either a full 100% of the original loan, otherwise 10% or less of the original loan. Interest and Fees are mostly 10% or less (the distribution is skewed right)

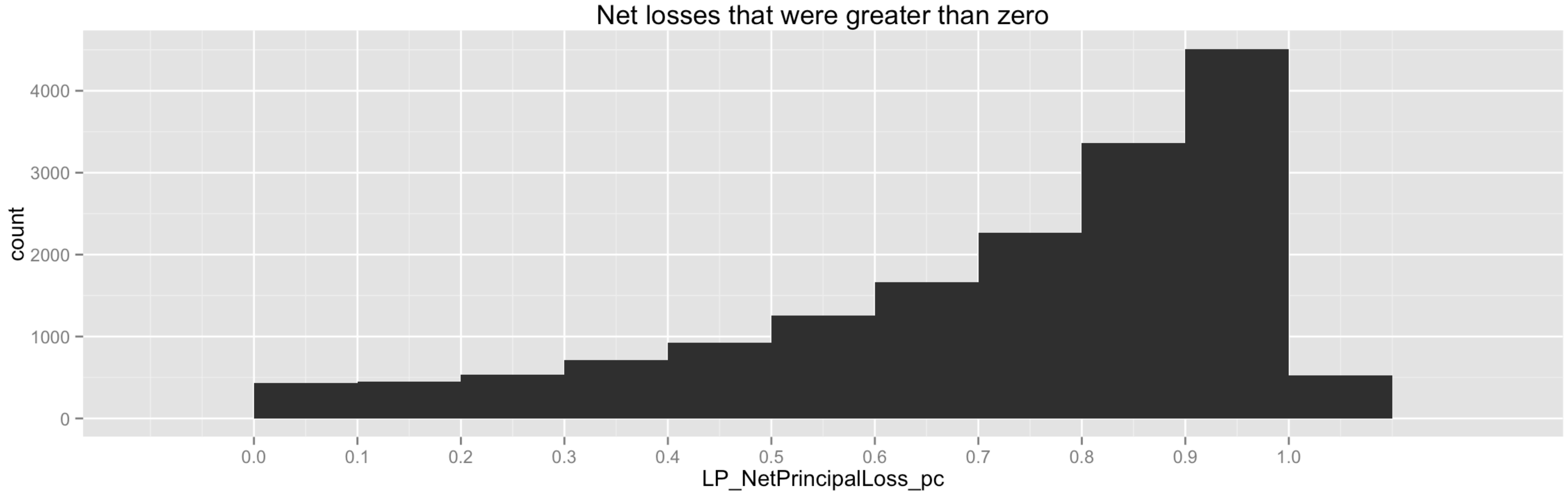




LP_NetPrincipalLoss

The principal that remains uncollected after any recoveries. Most losses are zero, but losses greater than zero tended to be closer to 100% of the original loan, rather than 50% or less of the loan.





```
# Univariate Analysis

### What is the structure of your dataset?
The data is a snapshot of data for loans that originated from 2005 to 2014, with some
loans curret and other completed or defaulted.  Fields are either known at the start
of the loan origination, and may have predicted the likelihood of the borrower paying
all pricipal plus interest.  Other fields are collected as a snapshot of the current
loan, and others are known at the termination of the loan, when it's either paid in f
ull or defaulted.  Some field are only used up to July 2009, and others are only pres
enta after July 2009, when one form of scoring a borrower's creditworthiness was repl
aced by another score.

### What is/are the main feature(s) of interest in your dataset?
Features that may have determined the crediworthiness of the borrower included the em
ployment status, stated monthly income, debt to income ratio, credit score, Prosper's
own credit rating of the borrower.

Features that help assess the quality of the investment are the amount of paid, amoun
t of total payments, and net loss on principal, all as a fraction of the original loa
n amount.
```

### What other features in the dataset do you think will help support your investigation into your feature(s) of interest?
Features that help to divide up the data into comparable subgroups are the date of loan origination and the loan status, which denotes whether a loan is current, completed, defaulted, or charged off. In particular, it will help to look at loans that are no longer current, to get the final total payments and losses. We can also determine which loans are closed by checking if the closed date exists (current loans have a null value for closed date). It will also help to use data for which loan origination occurred after July 2009, when Prosper replaced a rating system with a different one.

### Did you create any new variables from existing variables in the dataset?
I split the loan origination date and closed date into year and month (LoanOriginationYear, LoanOriginationMonth, ClosedYear, ClosedMonth). For customer payment related fields, I divided by the loan original amount and appended "_pc" to denote a percentage of the original loan (LP_CustomerPayments_pc, LP_CustomerPrincipalPayments_pc, LP_InterestandFees_pc, LP_ServiceFees_pc, LP_CollectionFees_pc, LP_GrossPrincipalLoss_pc, LP_NetPrincipalLoss_pc, LP_NonPrincipalRecoverypayments_pc).

### 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?
When I created box plots of customer payments as a fraction of the original loan, and then split them up by their original Prosper Rating, I found it odd that the higher rated loans had a lower median compared to lower rated borrowers. I realized that I was comparing loans that were still in progress with loans that had been completed. To make comparisons more clear, I filtered the observations to include only those that have completed, so I can compare only loans that have a full history of cumulative payments. Also, since some ratings did not become available until July 2009, I also filtered data to include only records where those fields have values.

# Bivariate Plots Section
###Payment amounts
These relate to how much the borrower paid as fraction of the original loan: LP_CustomerPayments_pc, LP_CustomerPrincipalPayments_pc, LP_InterestandFees_pc, LP_ServiceFees_pc, LP_CollectionFees_pc, LP_GrossPrincipalLoss_pc, LP_NetPrincipalLoss_pc, LP_NonPrincipalRecoverypayments_pc

I hypothesize that borrowers with lower credit scores made up more of the lower payment and upper payment ranges (they had higher interest rates), and higher credit score borrowers were closer to the 1.0 to 1.2 payment ratio.

ProsperRating
A custom risk score built using historical Prosper data. The score ranges from 1-10, with 10 being the best, or lowest risk score. Applicable for loans originated after July 2009.
Note, since this is based on historical Prosper data, I think it means that it refers to customers who had previous loans with Prosper, or refers to their payment history of their existing loan. I prefer to look at the Prosper Rating numeric, which was se

```
t at the time that the loan started.


### Unexpected results led me to subset the data for better comparisons of loan data
with full histories only
Surprisingly, those who had the worst rating (1,2,3) paid more than those with a high
er rating (4,5,6,7) even though the median for all was less than 100% of the original
loan amount.  Only the ratings 1 and 3 had a median payment of more than 50% of the o
riginal loan.

Looking at histograms faceted by ProsperRating, there were fewer borrowers with low r
atings compared to higher ratings, but the number of those paying nearly zero is simi
lar to those paying in full (distribution is relatively uniform).  For higher rated b
orrowers, especially (4,5,6), the distribution is bimodal, with more paying near zero
compared to paying in full.

This is probably because many of these loans are still current, so borrowers have mor
e time to pay more.  I could focus on just the loans that are completed.  After subse
tting on just loans that are completed, charged off, or defaulted, It higher rated bo
rrowers are much more likely to pay in full compared to defaulting. Those that pay at
least 100% of original also pay more when they have a lower rating (they have  higher
interest rates). Note that I went back to the univariate section to subset the data.
Correlation is .05, which is weak.

<img src="Exploration_of_loan_data_files/figure-html/unnamed-chunk-23-1.png" title=""
alt="" width="1152" />

#### LP_InterestandFees_pc
correlation is -0.44, which is fairly strong; lower rated customers paid more in inte
rest.
<img src="Exploration_of_loan_data_files/figure-html/unnamed-chunk-24-1.png" title=""
alt="" width="1152" />

#### LP_CustomerPrincipalPayments_pc
Correlation is .2425, which is not strong.
<img src="Exploration_of_loan_data_files/figure-html/unnamed-chunk-25-1.png" title=""
alt="" width="1152" />

#### LP_NonPrincipalRecoverypayments_pc
<img src="Exploration_of_loan_data_files/figure-html/unnamed-chunk-26-1.png" title=""
alt="" width="1152" />

#### LP_GrossPrincipalLoss_pc
Corr is -0.24 because lower rated borrowers had larger losses
<img src="Exploration_of_loan_data_files/figure-html/unnamed-chunk-27-1.png" title=""
alt="" width="1152" />

#### LP_NetPrincipalLoss_pc
Corr is -0.24 because lower rated borrowers had larger net losses
<img src="Exploration_of_loan_data_files/figure-html/unnamed-chunk-28-1.png" title=""
```

```
alt="" width="1152" />

### Compare income, amount borrowed, debt to income ratio to the final customer payme
nt

Monthly income vs. customer payments for completed loans.  Using a scatter plot, it's
hard to see a relation between income and payments, in part because the denser areas
are due to more borrowers being in the income range from 2800 to 6200, whether they p
aid 100% of the loan or defaulted.

I also tried cutting monthly income into 10 ranges by quantile, to spread the data po
ints more evenly when plotting.  Using a boxplot, I see that lower income ranges have
a wider spread of payments, but the medians for any income range are very similar.

Correlation between stated monthly income and customer payments is .04, which is weak
.
<img src="Exploration_of_loan_data_files/figure-html/unnamed-chunk-29-1.png" title=""
alt="" width="1152" />

#### DebtToIncomeRatio
Correlation with debt to income ratio vs. payments is -0.04, which is weak.
<img src="Exploration_of_loan_data_files/figure-html/unnamed-chunk-30-1.png" title=""
alt="" width="1152" />

#### CreditScoreRangeLower
Correlation between credit score and payments is 0.11, which is weak.
<img src="Exploration_of_loan_data_files/figure-html/unnamed-chunk-31-1.png" title=""
alt="" width="1152" />

#### LoanOriginalAmount
Those that borrowed the most (15,000 to 35000) had a higher risk of under-paying, but
so did the borrowers who borrowed between 3000 to 4000.  I will try to look at this a
gain with additional variables, to see if the amount borrowed is also related to inco
me.  Generally, higher income borrowers also borrow more, but I still don't see a rea
son for the increased risk for borrowers in the 3000 to 4000 loan amount range.

Correlation between original loan amount and customer payments is -0.07, which is wea
k.
```

$title

[1] “Defaults appear more common among those that borrowed 5000 or less”

attr(,“class”)

[1] “labels”

$title

[1] “Incomes under 6,800 per month borrowed $5,000 or less”

attr(,“class”)

[1] “labels”

$title

[1] “The lower 70% of loans ($7,450 or less) had incomes at <= $6,000”

attr(,“class”)

[1] “labels”

```

<img src="Exploration_of_loan_data_files/figure-html/unnamed-chunk-32-1.png" title=""
alt="" width="1152" />

#### LoanOriginalAmount / Term
<img src="Exploration_of_loan_data_files/figure-html/unnamed-chunk-33-1.png" title=""
alt="" width="1152" />

#### LoanOriginalAmount per Term / StatedMonthlyIncome
When normalizing original loan amount by the number of terms to pay, and also by mont
hly income, the distribution looks similar, with a higher risk at 9.5% or higher of i
ncome, and again at 3.17% to 3.92% of income.  Risk is lower in-between, from 3.92% t
o 9.58% of income.
<img src="Exploration_of_loan_data_files/figure-html/unnamed-chunk-34-1.png" title=""
alt="" width="1152" />

#### BorrowerState
```

Of states with at least 1000 observations (where loans were completed), New York, Virginia and Colorado had less downside risk compared to the others, such as California, Florida, Illinois, Georgia, Texa and Ohio.
<img src="Exploration_of_loan_data_files/figure-html/unnamed-chunk-35-1.png" title="" alt="" width="1152" />

#### Prosper rating and credit score

Prosper's rating is related to borrowers' credit scores, (0.6 correlation) but Prosper appears to give lower ratings to some borrowers with high credit scores, based on other information.
<img src="Exploration_of_loan_data_files/figure-html/unnamed-chunk-36-1.png" title="" alt="" width="1152" />

# 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?

Indicators of creditworthiness, such as income, credit score, and prosper's rating show that for lower creditworthiness, borrowers have more varied payment outcomes. However, the median outcomes are still pretty similar across different levels of creditworthiness.

When comparing the absolute loan amount to customer payments as a fraction of the the loan, there is more variance in payments for the highest loan amounts (15,000 to 35,000), and also at the 3000 to 4000 dollar range. So there is higher risk in those two ranges, but lower risk in-between (400 to 15,000). Lowest risk is for loans 1,500 and less. Even when I normalize the loan amount by dividing by terms (months to pay off the loan) and then dividing by monthly income, I still see two peaks where there is more downside risk, with lower risk in-between.

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

Payments as a fraction of the loan, when grouped by state, showed that, of the states with at least 1,000 observations (where loans had a closed date), New York, Virginia and Colorado had less downside risk compared to the others, such as California, Florida, Illinois, Georgia, Texa and Ohio.

### What was the strongest relationship you found?

The strongest relationship I found was between the prosper rating and the amount of interest and fees paid (as a fraction of the original loan), at -0.44. The lower rated borrowers had higher interest rates, and paid more interest than those with higher ratings.

# Multivariate Plots Section

#### Monthly Income, Prosper Rating, and Customer Payments

```
I want to look at how customer payments as a fraction of the loan vary by income, and
split up by Prosper Rating.  In order to make the plots more clear, I bucket the inco
me into $500 ranges, by rounding the income to the largest multiple of $500 that is l
ess than the income; for example, 1,501 and 1999 are both rounded to 1500.  For each
income range, I want to see the median customer payment ratio, and I use colors to se
parate the prosper ratings (1-7), where red is the lowest rating (1) and blue is the
best rating (7).  It looks like for lower rated borrowers with lower incomes (less th
an $2500), they pay less than higher rated borrowers of the same income level.  Howev
er, for higher incomes, especially around $8000 to $9000, lower rated borrowers pay m
ore than do higher rated borrowers. I think that the lower rating required these borr
owers to pay more interest, and their higher incomes enabled them to complete most or
all of their payments.

I also plotted credit score; first removing NA's and zeros, since minimum credit scor
es are usually around 300.  Credit scores that are not the worst, but still on the lo
wer end (560 to 680 range), appear to have higher median payments than for credit sco
res that are higher and lower.  Higher incomes appear to reduce the risk of underpaym
ent even for those with lower credit scores.




<img src="Exploration_of_loan_data_files/figure-html/unnamed-chunk-37-1.png" title=""
alt="" width="1152" />

<img src="Exploration_of_loan_data_files/figure-html/unnamed-chunk-38-1.png" title=""
alt="" width="1152" />




<img src="Exploration_of_loan_data_files/figure-html/unnamed-chunk-40-1.png" title=""
alt="" width="1152" />

#### Credit score, income, and payments
Higher credit scores resulted in better payments at any income.  For lower credit sco
res (360 to 560) and income below $3000, payments were lower compared to incomes betw
een $3000 and $9000.  Otherwise, payments do not seem to vary much with income levels
.  The graph shows more variation at higher incomes, but there are fewer observations
at these income levels, which makes it harder to generalize the the population in gen
eral.
<img src="Exploration_of_loan_data_files/figure-html/unnamed-chunk-41-1.png" title=""
alt="" width="1152" />



#### Stated Monthly Income, Prosper Rating, customer payments by quantile

Instead of boxplots, I'm trying line graphs for a more granular look at monthly incom
e.  At lower incomes and lower ratings, 1st quartile (25% quantile) customer payments
fall below 100% of the original loan.
```

int [1:26005] 1 3 2 6 6 7 3 2 3 2 …

‘data.frame’: 26005 obs. of 1 variable:

$ ProsperRating..numeric.: int 1 3 2 6 6 7 3 2 3 2 …

```


<img src="Exploration_of_loan_data_files/figure-html/unnamed-chunk-43-1.png" title=""
alt="" width="1152" />





<img src="Exploration_of_loan_data_files/figure-html/unnamed-chunk-45-1.png" title=""
alt="" width="1152" />

#### Stated Monthly Income, Prosper Rating, and interest & fees by quantile

<img src="Exploration_of_loan_data_files/figure-html/unnamed-chunk-46-1.png" title=""
alt="" width="1152" />


#### Stated Monthly Income, Prosper Rating, and principal by quantile

<img src="Exploration_of_loan_data_files/figure-html/unnamed-chunk-47-1.png" title=""
alt="" width="1152" />

#### Median payments versus income, grouped by rating, facetted by term
In the subset of borrowers whose loans are completed, and that have a Prosper rating,
most are for 36 months (20,778), followed by 60 months (3696) and 12 months (1531).
Some large variations occur in 12 and 60 term graphs, possibly due to the few number
of samples.  To see if this is the case, I also include a histogram to show the numbe
r of observations for each income bucket.  It does appear that much of the jagged-nes
s in the lines occurs when there are fewer obsevations.  However, even when looking a
t areas with at least 250 observations, it appears that in the 36 term category, borr
owers with incomes below $4000 are at higher risk of paying less than 100% of the ori
ginal loan, compared to borrowers within incomes between $4000 and $8000.  This appli
es to both good and bad ratings.  For the 60 term category, any income level and any
rating appears to have higher risk of paying less than the original loan.
```

```
<img src="Exploration_of_loan_data_files/figure-html/unnamed-chunk-50-1.png" title=""
alt="" width="1152" />




#### Loan status with delinquencies
I want to look at in-progress loans, where loan status is delinquent.  I also want to
look at how far along the loan is from the origination date.  When loans default, the
y tend to do sooner (within 12 months) rather than later (after 24 months).  Loan del
inquencies appeared to be more common for loans that originated 25 months prior.

Histograms of:
LoanCurrentDaysDelinquent
LoanFirstDefaultedCycleNumber
LoanMonthsSinceOrigination
<img src="Exploration_of_loan_data_files/figure-html/unnamed-chunk-51-1.png" title=""
alt="" width="1152" />
<img src="Exploration_of_loan_data_files/figure-html/unnamed-chunk-52-1.png" title=""
alt="" width="1152" />

LoanFirstDefaultedCycleNumber vs. customer payments
I'm trying to see if defaulted cycle is only for defaulted and chargedoff loan status
, or if it also applies to completed or current loans.  There are a few (46) where lo
an status is Completed and the first defaulted cycle is greater than zero.  There are
also very few where the loan is past due, and also has a first defaulted cycle number
greater than zero.  Most loans with a first default cycle are chargedoff or defaulted
.
```

Cancelled Chargedoff Completed

0 11985 46

Current Defaulted FinalPaymentInProgress

2 4843 1

Past Due (>120 days) Past Due (1-15 days) Past Due (16-30 days)

0 2 1

Past Due (31-60 days) Past Due (61-90 days) Past Due (91-120 days)

0 2 0

```


Interestingly, for those that default, the ones with lower ratings paid more as a per cent of the original loan compared to borrowers with higher ratings.
<img src="Exploration_of_loan_data_files/figure-html/unnamed-chunk-54-1.png" title="" alt="" width="1152" />



#### First defaulted cycle number and monthly income
There doesn't appear to be a clear relation between income and the first default cycle
<img src="Exploration_of_loan_data_files/figure-html/unnamed-chunk-55-1.png" title="" alt="" width="1152" />

#### first defaulted cycle versus debt income ratio
Debt ratio doesn't show a relationship with how early a borrower defaults.
<img src="Exploration_of_loan_data_files/figure-html/unnamed-chunk-56-1.png" title="" alt="" width="1152" />

####  First default cycle vs. Loan original amount

<img src="Exploration_of_loan_data_files/figure-html/unnamed-chunk-57-1.png" title="" alt="" width="1152" />

#### first default cycle v. LoanOriginalAmount_per_Term_pc_Income
I don't see a relation between loan amount per term as a percent of monthly income, to when the loan first defaulted.  Even when I bucket the loan per income into 0.1 sections, I still don't see a relationship with how early the borrower defaults.
<img src="Exploration_of_loan_data_files/figure-html/unnamed-chunk-58-1.png" title="" alt="" width="1152" />

#### MonthlyLoanPayment vs. default cycle
I don't see much of a change in default cycle based on the size of monthly loan payments.
<img src="Exploration_of_loan_data_files/figure-html/unnamed-chunk-59-1.png" title="" alt="" width="1152" />

#### Credit rating vs. default cycle
<img src="Exploration_of_loan_data_files/figure-html/unnamed-chunk-60-1.png" title="" alt="" width="1152" />



# Multivariate Analysis
```

### Talk about some of the relationships you observed in this part of the investigation. Were there features that strengthened each other in terms of looking at your feature(s) of interest?
I first looked at the relation between income and median of payments, and separated by their Prosper rating. It appeared that for incomes lower than $2500 a month, lower rated borrowers paid less, while above $2500 incomes, lower rated borrowers paid more, relative to higher rated borrowers. I posit that the lower ratings required borrowers to pay higher interest, while higher incomes made it more likely that they could make those payments.

To get a better sense of the distribution of payments and risk, I looked at the 25% quartile and 75% quartile of customer payments, versus income. The 25% quantile showed the lower paying borrowers, and clearly show that lower rated borrowers paid much less than higher rated ones. Higher rated borrowers appeared to pay similar ammounts regardless of income for the 25%, 75% and median quantiles. At the 75% quantile, lower rated borrowers paid more than the others at most income levels, reflecting their higher interest rates. So the lower rated borrowers' payments reflected higher risks and higher rewards.

I also looked at how early a borrower defaulted, compared with their final payments. As expected the earlier a borrower defaulted, the less of the original loan they paid off. The relationship is fairly linear between default cycle and customer payments.

### Were there any interesting or surprising interactions between features?
When I faceted the plots by term (12, 36, or 60 month terms), I found it interesting that the borrowers who choose a 60 month time period seem to be higher risk for under-paying, even for higher rated borrowers. For example, the median payments for 60 term borrowers was mostly less than 100% of the loan for most borrowers who made $4,000 or less, for any rating (high, medium, low).

When plotting the defaulted term against the payment amount, I noticed that at most income levels, lower rated borrowers paid more than higher rated borrowers by the time of default. This makes sense, given higher interest rates for lower rated borrowers. It is interesting that if you have a borrower default on a loan, it's worse for the lender when the borrower had a higher rating.

# Final Plots and Summary

### Plot One
<img src="Exploration_of_loan_data_files/figure-html/Plot_One-1.png" title="" alt="" width="1152" />

#### Description One
I plotted customer payments and dividing monthly income into 10 quantiles, and divided credit score (lower bound) into 10 quantiles. We can see that lower rated borrowers have more risk than higher rated borrowers at any income level, but payments for low credit scored borrowers were better with higher incomes. Moreover, the lower rated borrowers (560 to 680) have higher median payments than those rated above and below. For higher incomes, the 25% quantile of payments is higher (there is less risk), but

```
the median does not change as much as we compare by income level.

### Plot Two
<img src="Exploration_of_loan_data_files/figure-html/Plot_Two-1.png" title="" alt=""
width="1152" />

#### Description Two

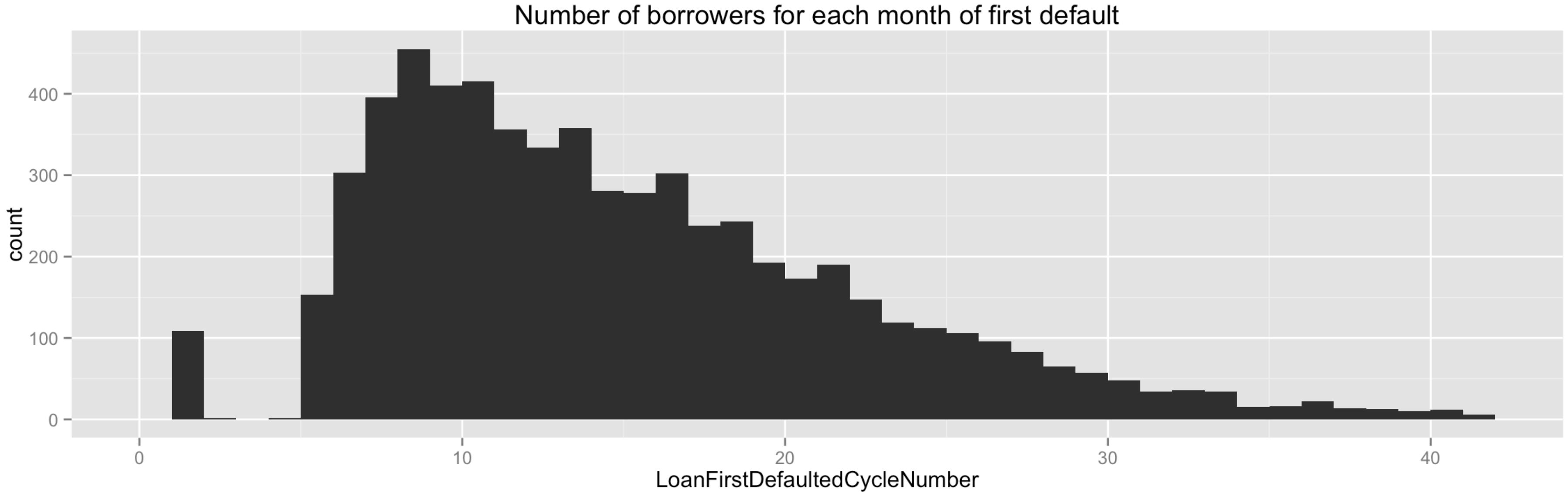
Customer payments are generally more stable at or above 100% of the original loan at
the 25% to 75% quantile.  Lower rated borrowers are at risk of under-paying, especial
ly when their incomes are below $2,500 per month.  Lower rated borrowers tend to pay
somewhat more than higher rated borrowers at incomes above $2,500.  Also, for those w
ho chose 60 months to pay off the loan, instead of 36 months, there was a higher like
lihood of under-paying at any rating, and at most income levels.

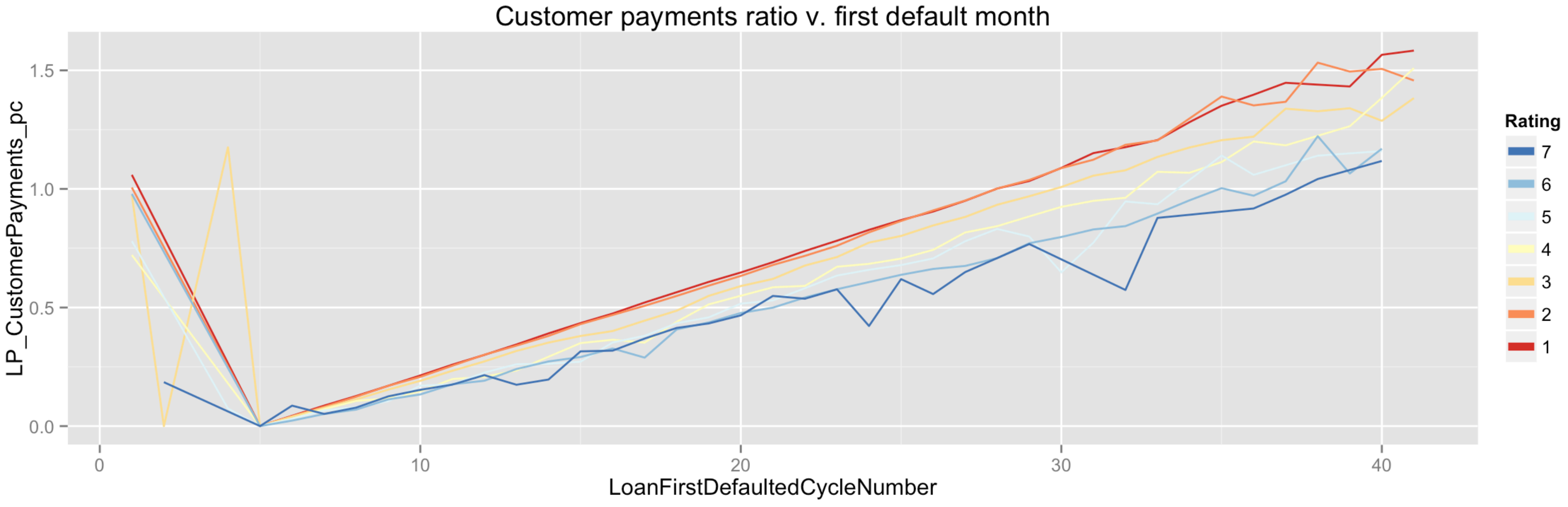
### Plot Three
```

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's

1.00 9.00 13.00 14.48 19.00 41.00 19769

```




```

Description Three

When plotting the default cycle (how many months after the loan began the borrower defaulted on the loan), the 25% to 75% quantiles are 9 to 19 months, with more borrowers defaulting earlier rather than later. There is a fairly linear relationship between early defaults and lower cumulative payments. Interestingly, for two borrowers who default at the same cycle, the one with a higher rating ends up paying less, likely because they have lower interest rates.

Reflection

This data is a snapshot of both in-progress and completed (or defaulted) loans. When I first attempted to look at all of the data, it appeared that better-rated borrowers were showing lower cumulative payments. This helped me realize that I was comparing in-progress loans with completed/defaulted loans, when I really should be looking at completed/defaulted loans as a subset. I also decided that to compare cumulative payments among borrowers, I should normalize payments by the original loan size, to get a payment as a percent of the loan.

I found some indication that better credit ratings, higher incomes, and 3-year terms showed better payment outcomes (similar median but smaller inter-quartile range). Also, borrowers with lower credit ratings (excluding the lowest), end up paying slightly more (as a median) than those with higher credit, so it seems that their higher interest rates compensate for their higher risk of default.

One thing that makes it harder to draw conclusions is the uneven distribution; for instance, there three types of terms (1 year, 3 year, 5 year), but most loans are 3 year terms, with much fewer loans in the other terms. So when there is more variation in the groups that have fewer observations, it's hard to know if this would still be the case if there were more observations.

It would be helpful for the geographic breakdown to show both the state and county level. Although there are differences when comparing loans at the state level, some of the larger states may have fairly distinct regions that might account for why some populous states had a fairly wide range of payment outcomes.

Also, it would be helpful to not only see the monthly income, but get a sense of monthly expenses. This could either be data on monthly rent or mortgage, or even monthly payments for other debts. Perhaps a net disposable income based on income and expenses would show a clearer relationship with cumulative loan payments.