Author : Farhad Ahmad Qureshi

Introduction

Prosper is a marketplace lending platform. Borrower’s send requests for loans and investors then decide how much to lend. Borrower rates are usually lesser than financial institutions such as banks and multiple investors contribute towards the loan reducing the risk of impacting only one investor if the borrower defaults.

Dataset

This 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.

The reason I selected this dataset was because I never worked on a data set like this and I wanted to see what factors lead borrowers to become defaulters and what factors lead an investor to contribute towards a loan. I will first explore the dataset and see distribution of a few selected variables that I found interesting. This will be the univariate analysis. Following this, I will explore relation of these variables with being a default non-default (good) borrower. This will be the bivariate analysis. At the end I will use these plots to to add more information in the multivariate plots section.

The exploratory analysis will begin with the univariate plots using the following variables:

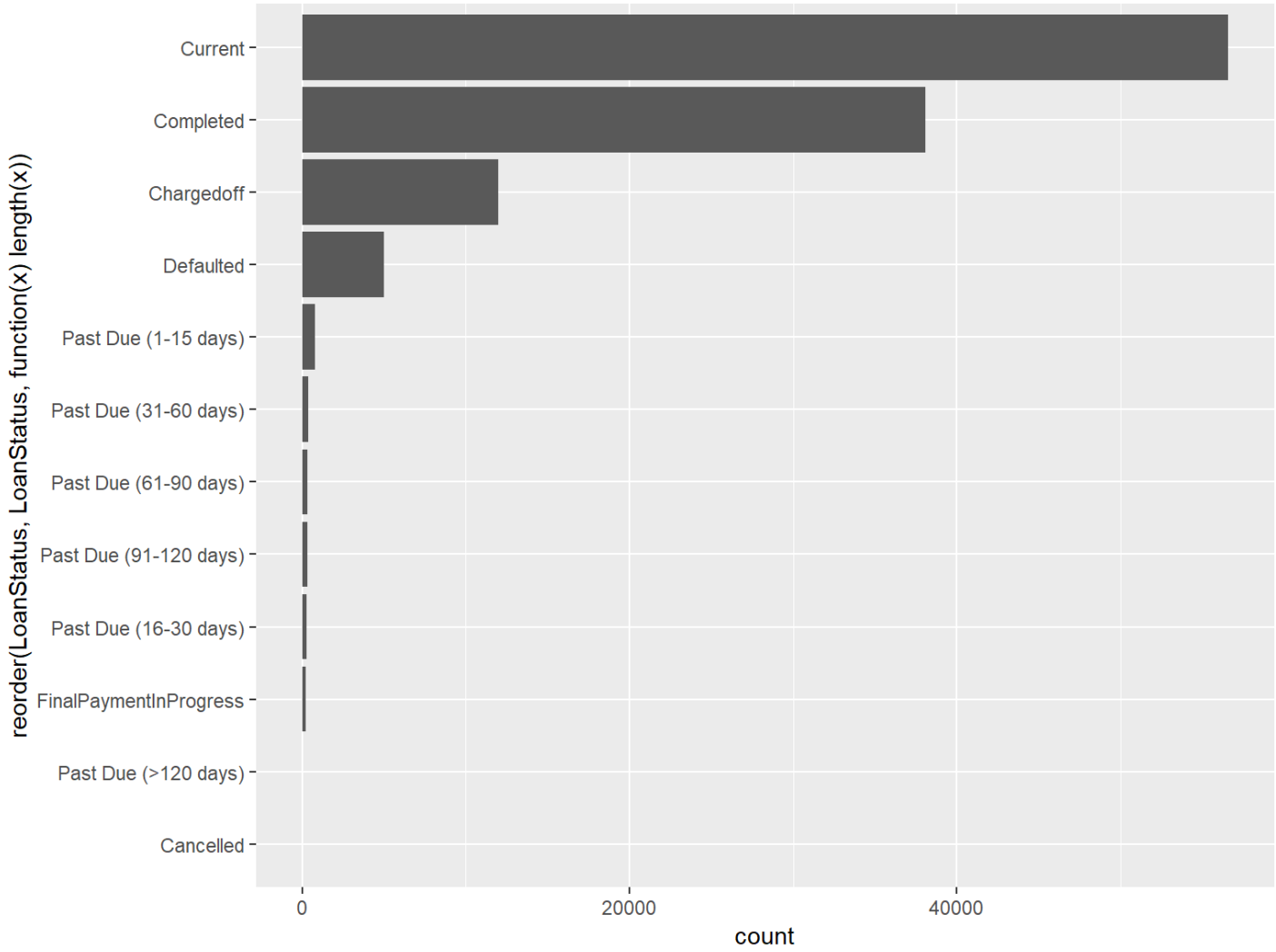
Distribution of variables selected for analysis.

```
## 'data.frame':    113937 obs. of  17 variables:
##  $ ListingCreationDate       : chr  "2007" "2014" "2007" "2012" ...
##  $ 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 ...
##  $ BorrowerAPR               : num  0.165 0.12 0.283 0.125 0.246 ...
##  $ BorrowerState             : Factor w/ 52 levels "","AK","AL","AR",..: 7 7 12 12 25 34 18 6
16 16 ...
##  $ Occupation                : Factor w/ 68 levels "","Accountant/CPA",..: 37 43 37 52 21 43
50 29 24 24 ...
##  $ EmploymentStatus          : Factor w/ 9 levels "","Employed",..: 9 2 4 2 2 2 2 2 2 2 ...
##  $ IsBorrowerHomeowner       : Factor w/ 2 levels "False","True": 2 1 1 2 2 2 1 1 2 2 ...
##  $ CreditScoreRangeUpper     : int  659 699 499 819 699 759 699 719 839 839 ...
##  $ InquiriesLast6Months      : int  3 3 0 0 1 0 0 3 1 1 ...
##  $ 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
...
##  $ StatedMonthlyIncome       : num  3083 6125 2083 2875 9583 ...
##  $ LoanOriginalAmount        : int  9425 10000 3001 10000 15000 15000 3000 10000 10000 10000
...
##  $ LP_CustomerPayments       : num  11396 0 4187 5143 2820 ...
##  $ Investors                 : int  258 1 41 158 20 1 1 1 1 1 ...
##  $ LoanMonthsSinceOrigination: int  78 0 86 16 6 3 11 10 3 3 ...
```

There are 6 factor variables 2 character variables and rest being numerical.

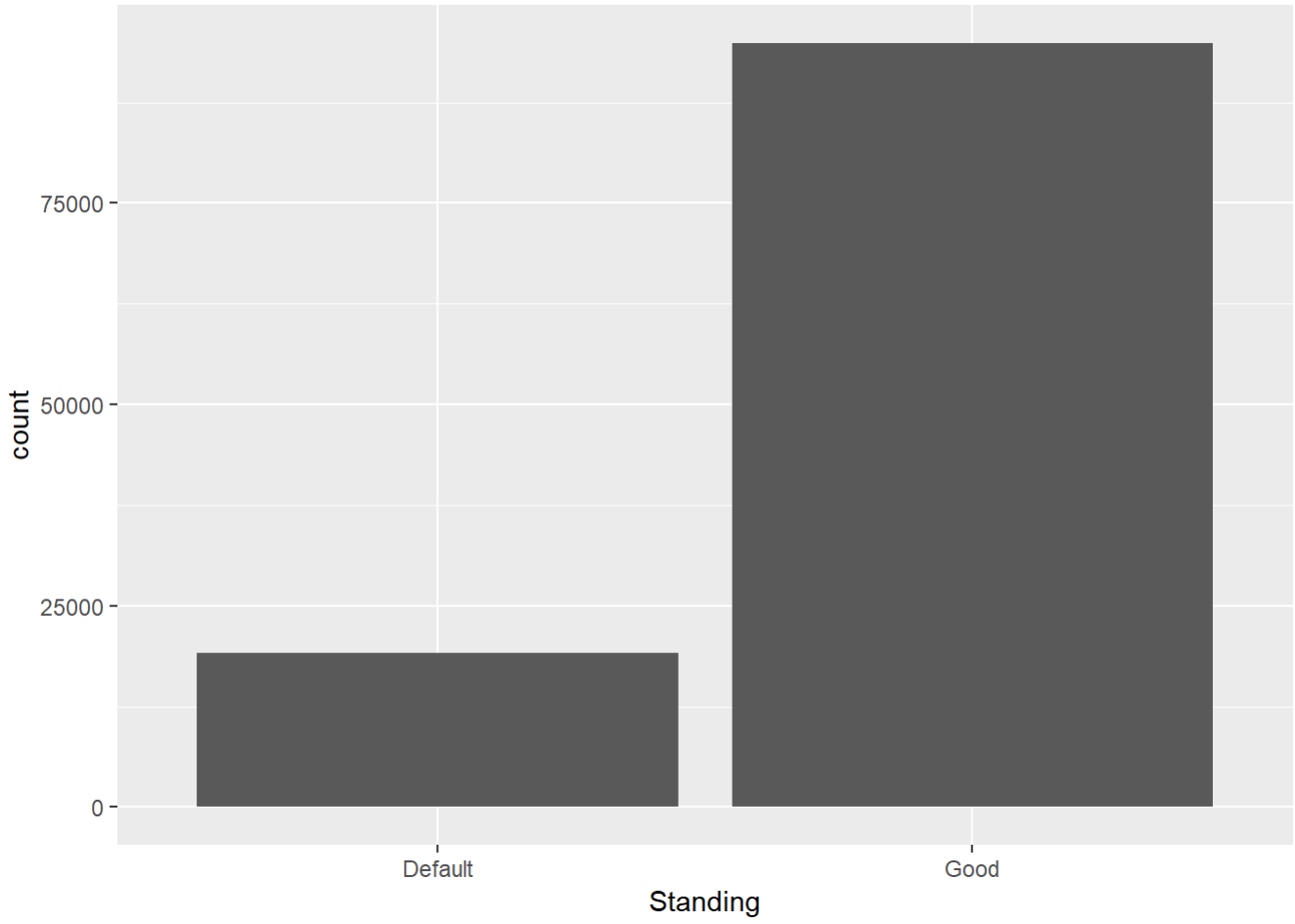
Univariate Plots Section

Loan Status.



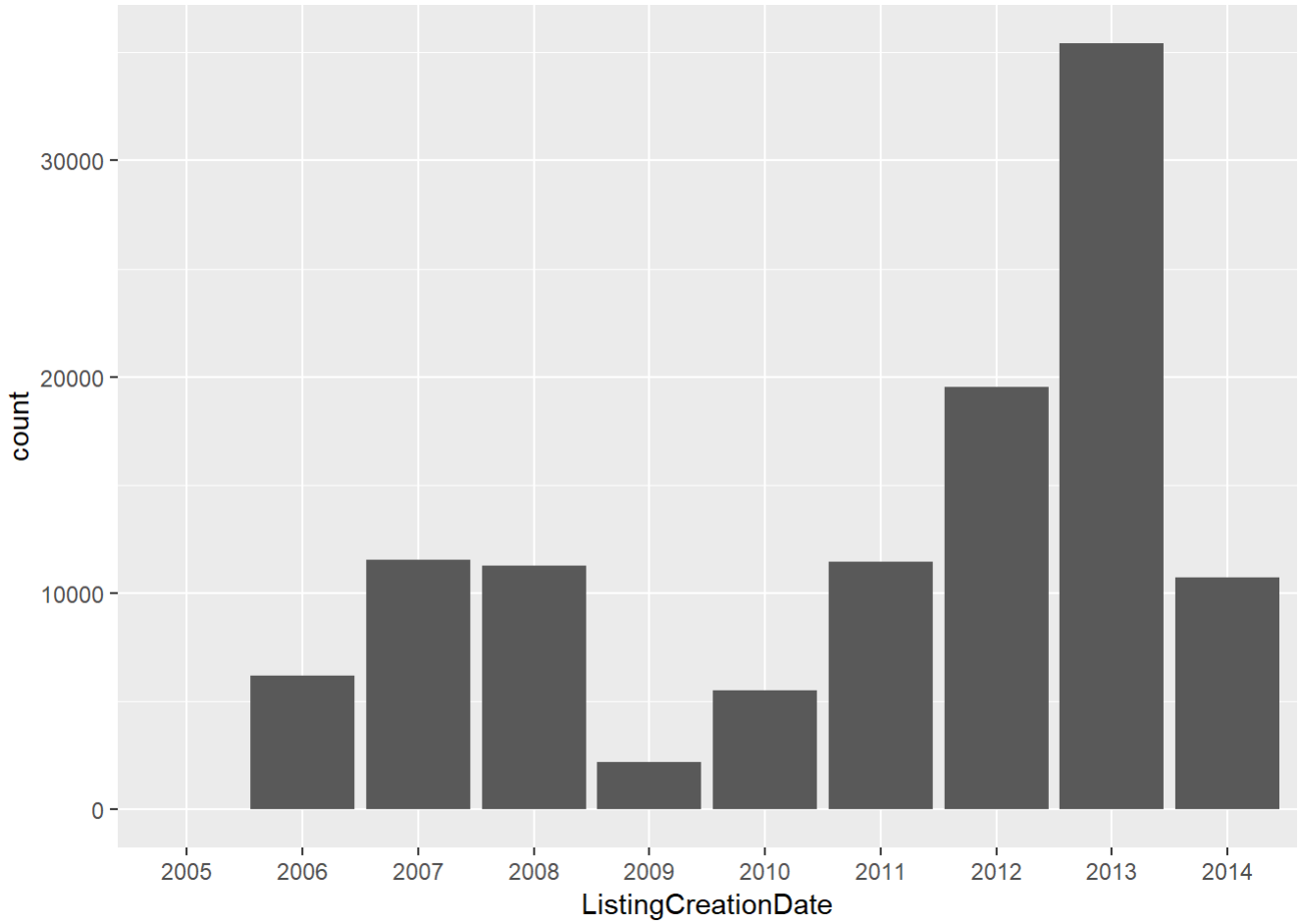
Distribution of the loan status. Majority are current borrowers. A good bit of borrowers have completed their terms. However there is a good amount of borrowers who have been marked as charged off. So after doing some reserach charged off is basically a defaulter in a really worst condition. In fact, all except current, completed and cancelled will be made into the ‘Good’ category and rest will be put in to the default category for analysis.

Checking the distribution of default and good standing borrowers:



Defaulters are usually less than in number from those in Good standing (including those who have completed their loans).

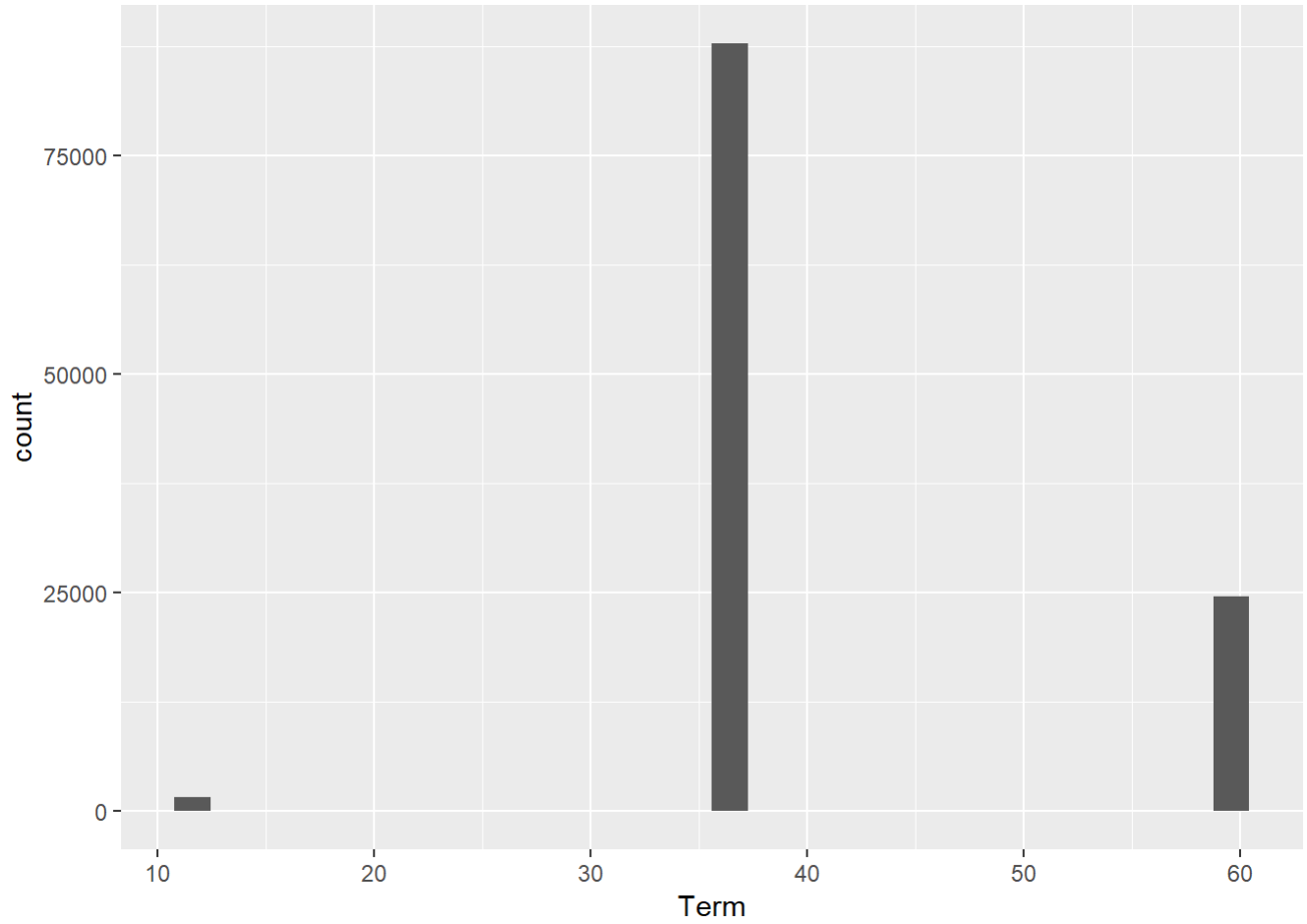
Creation Date



Number of loans were maximum in 2013. There is a reduction in loans in 2008 - 2011. It might be due to the global financial crisis. This will further be investigated to come to any conclusion.

Loan Term

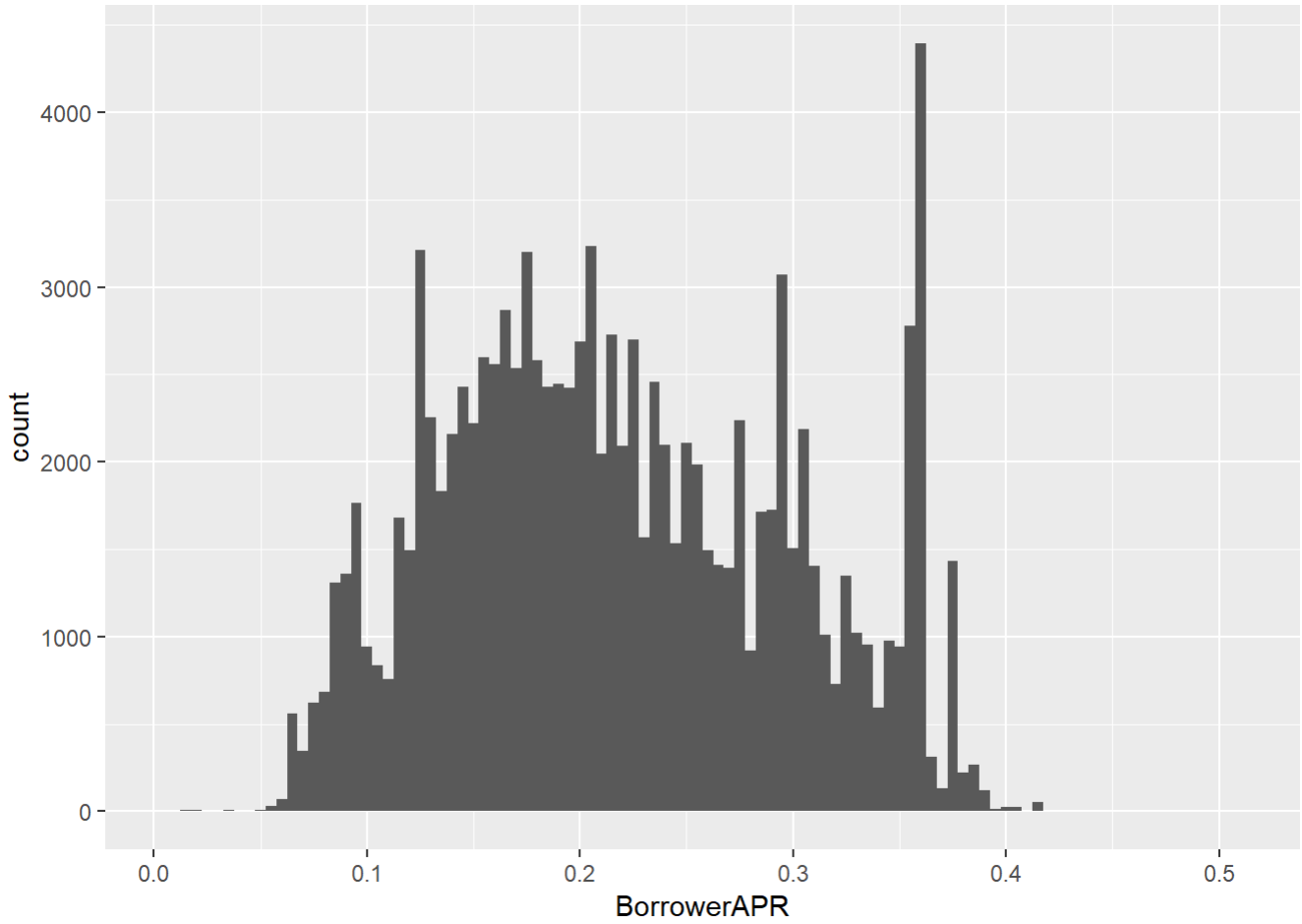
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Majority of the borrowers are given loan for a term of 36 months. The second highest being the 60 months term.

Borrower APR

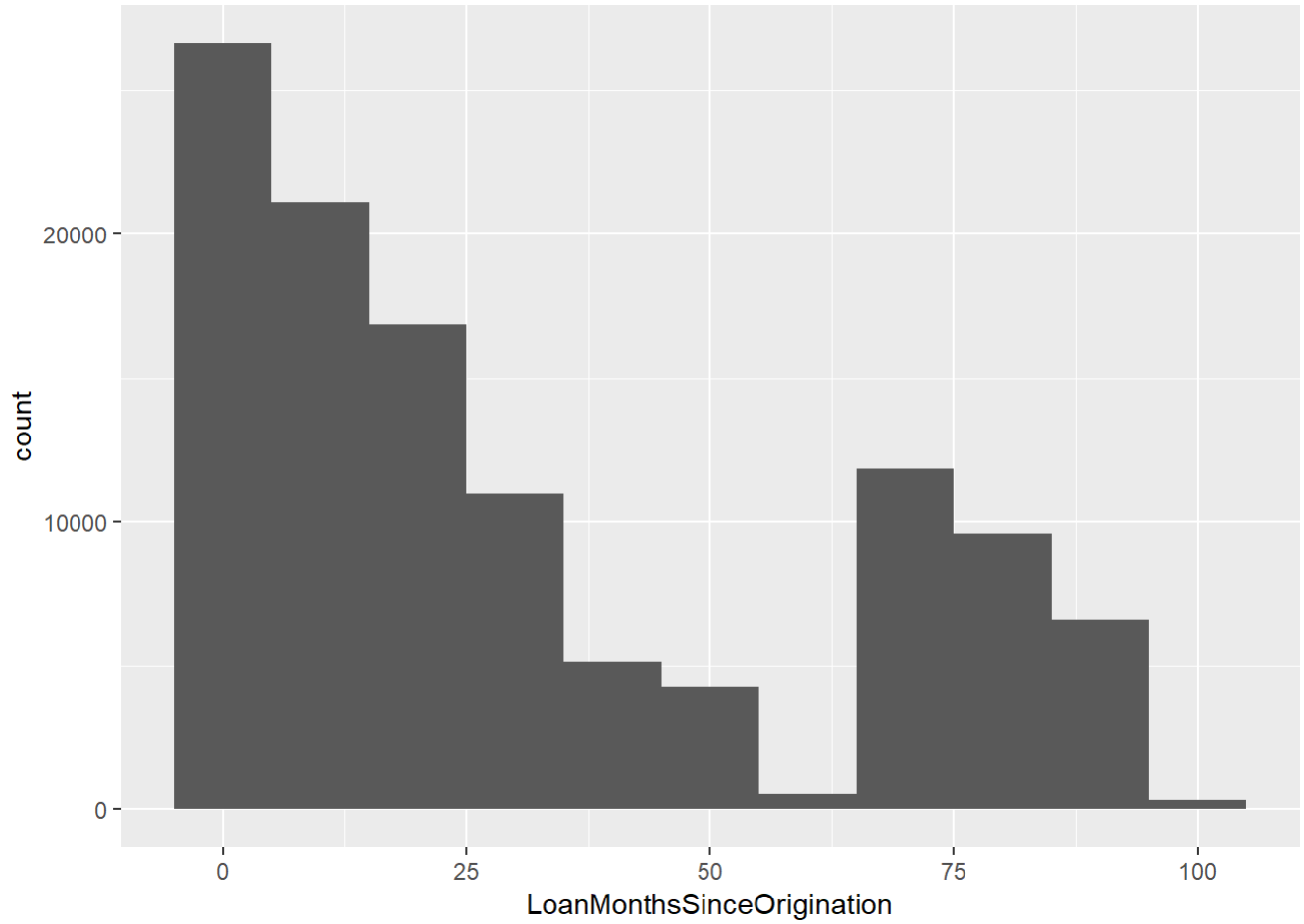
This is used instead of Borrower rate since it contains any additional service charges as well.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## 0.00653 0.15629 0.20976 0.21883 0.28381 0.51229      25
```

The lowest of Borrower's APR is 0.00653 and the maximum is upto 0.51. The typical range of this is between 0.05 to 0.38

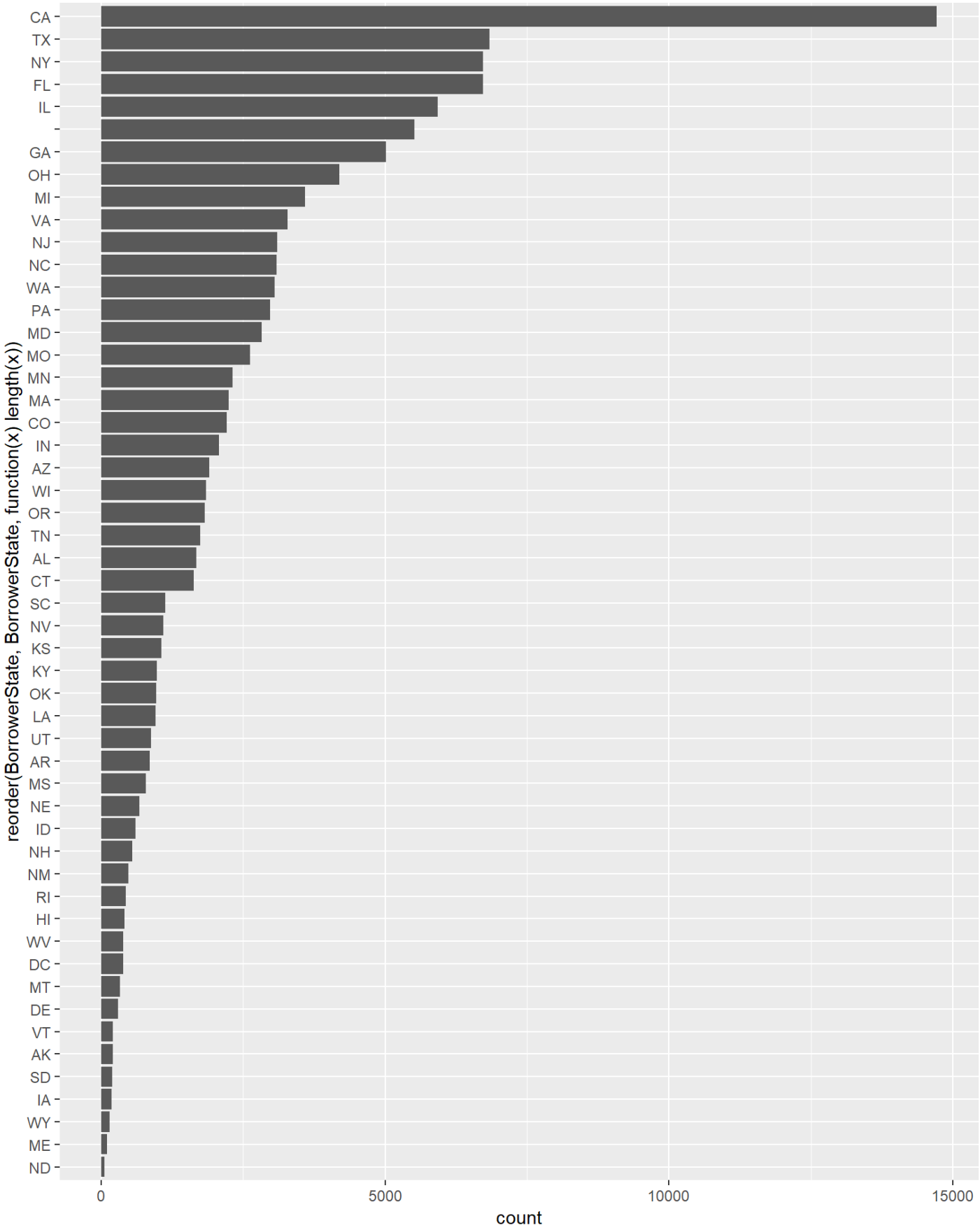
LoanMonthsSinceOrigination



Distribution of loans with respect to the months since they were originated.Most loans are between 0 - 25. This will be an interesting feature later to investigate default borrowers.

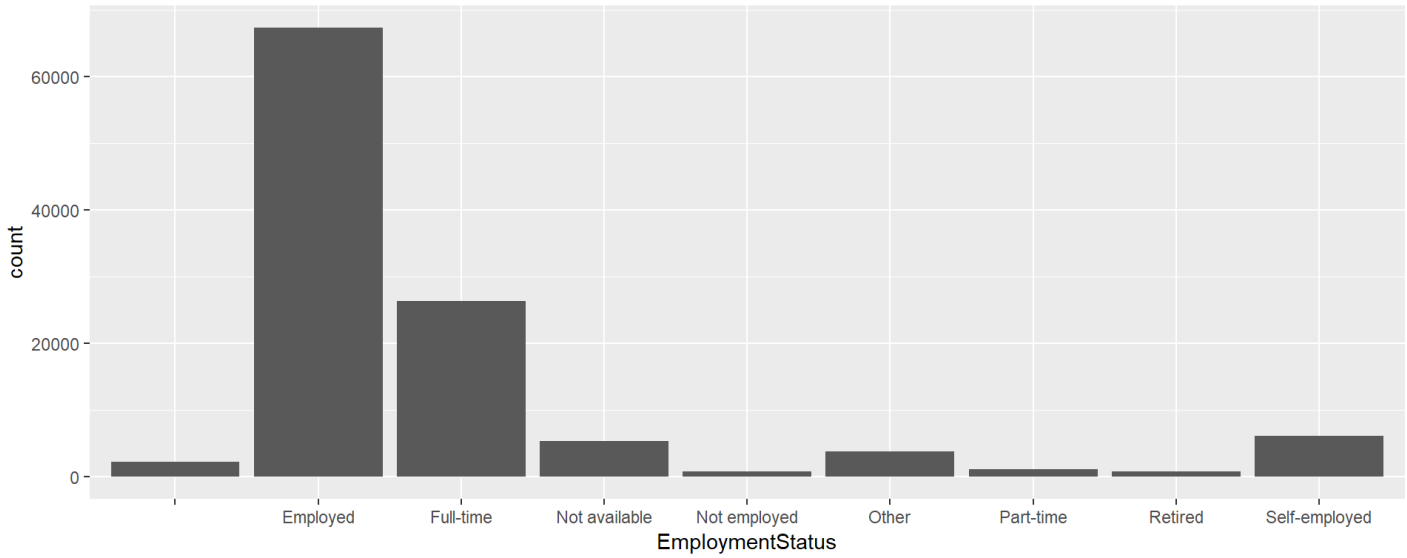
Borrower State

I was interested to see the demographics for each state and if there are any states with high number of borrowers.



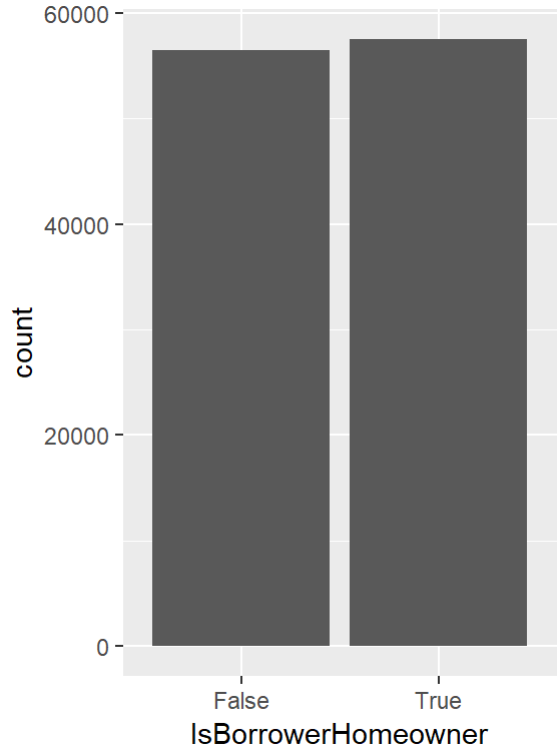
Majority of loans are from California. This might be because of its large population. This causes the plot to be right skewed so we will take this into consideration later and log transform the y axis.7 States have count more than or equal to 5000

Employment Status



Majority of the loans go to the the employed classes.

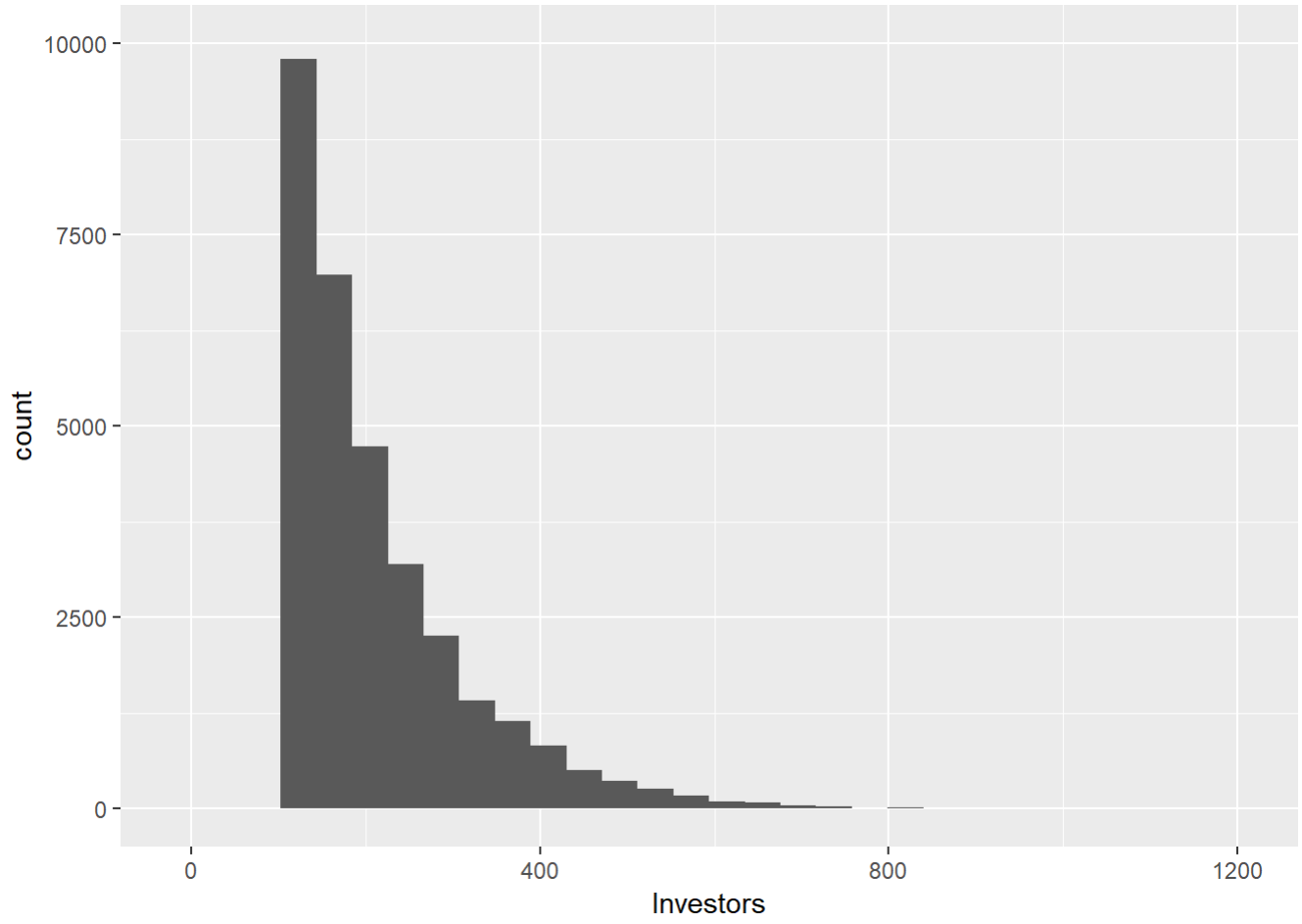
IsBorrowerHomeowner



Doesnt really matter if a borrower owns a house or not count is almost the same for both.

Investors

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



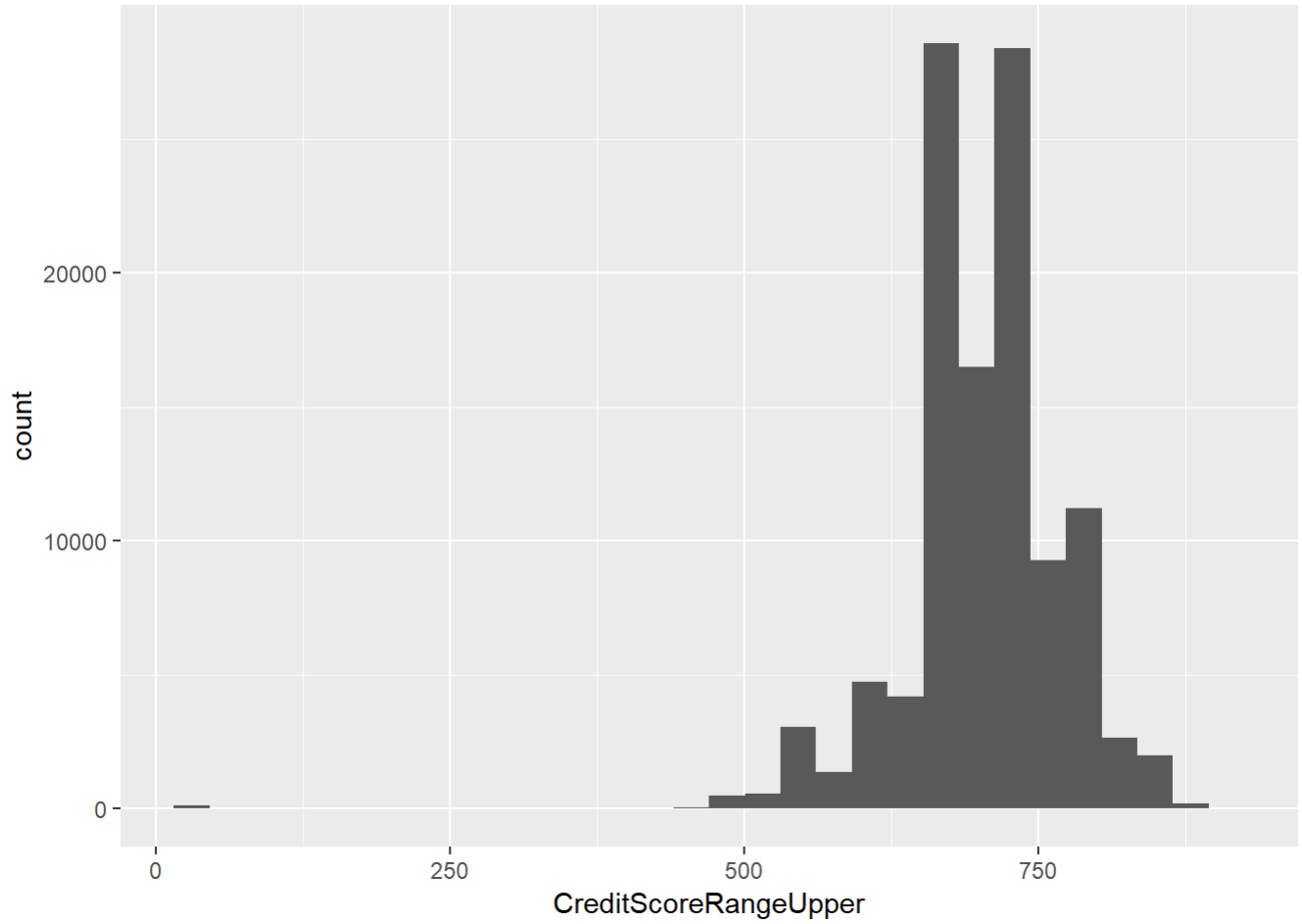
Distribution of investors majority being under 100 but it would be interesting to see the loan amount and how much investors contribute.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    2.00   44.00   80.48  115.00 1189.00
```

Median investors for the data is 44.

CreditScoreRangeUpper

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

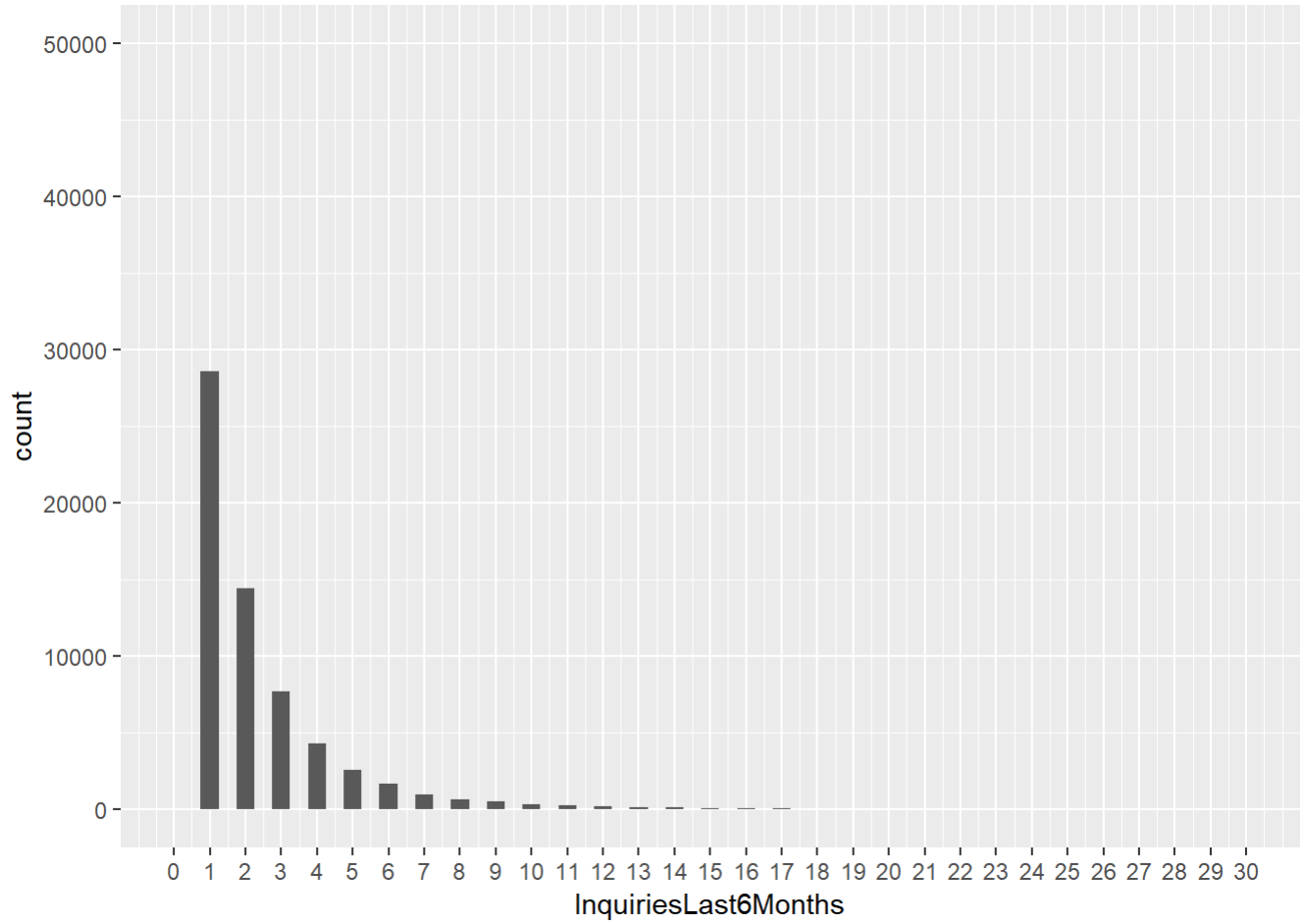


As expected most borrowers have high scores (peak around 700).

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    19.0   679.0   699.0   704.6   739.0   899.0     591
```

For upper range we see median rating is in the Good category as per the experience or FICO rating used by Prosper.

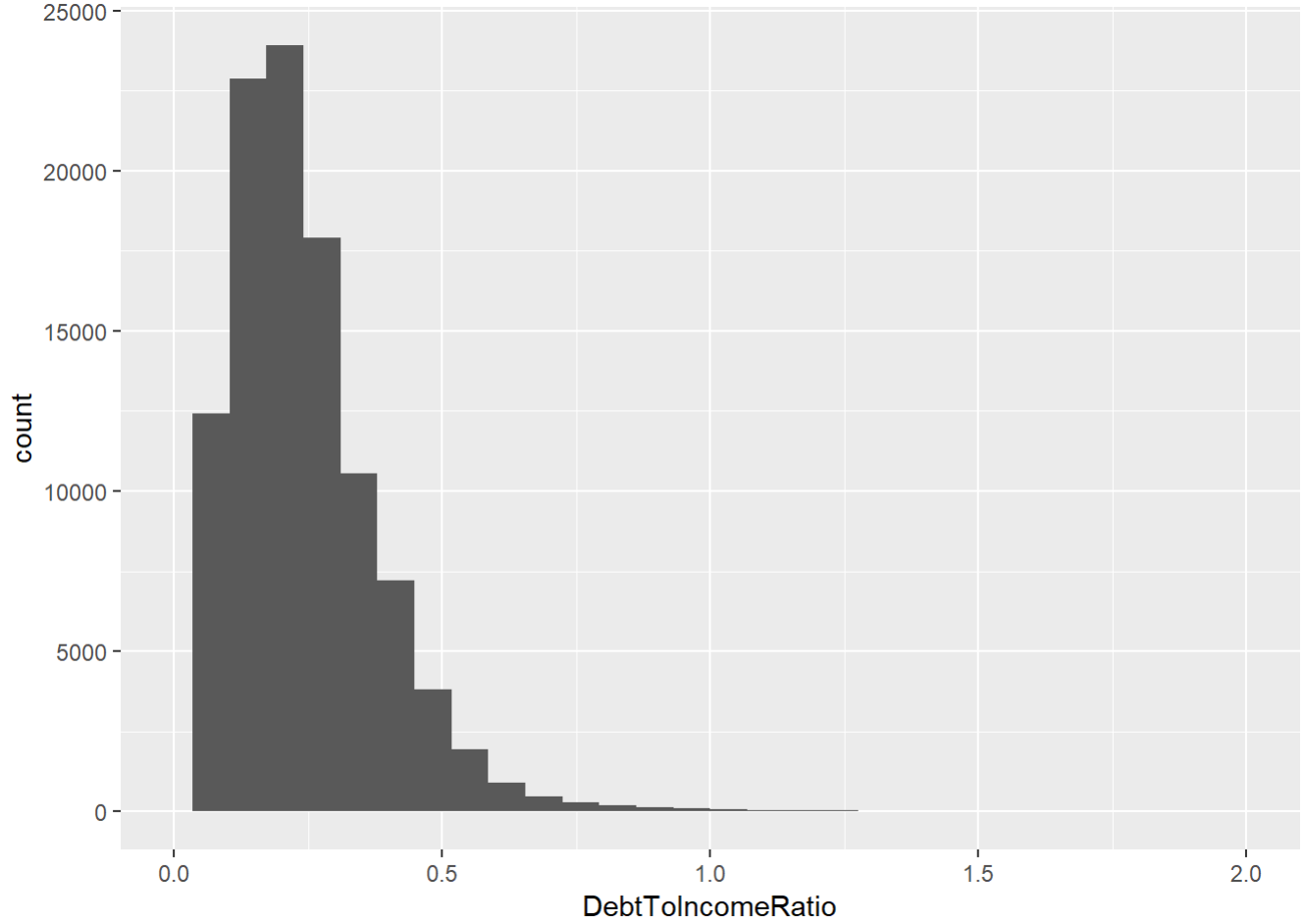
InquiriesLast6Months



Most borrowers had inquiries under 5 which as mentioned on their website is one of the criteria to give loans but there are outliers here as well.

DebtToIncomeRatio (DTI)

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

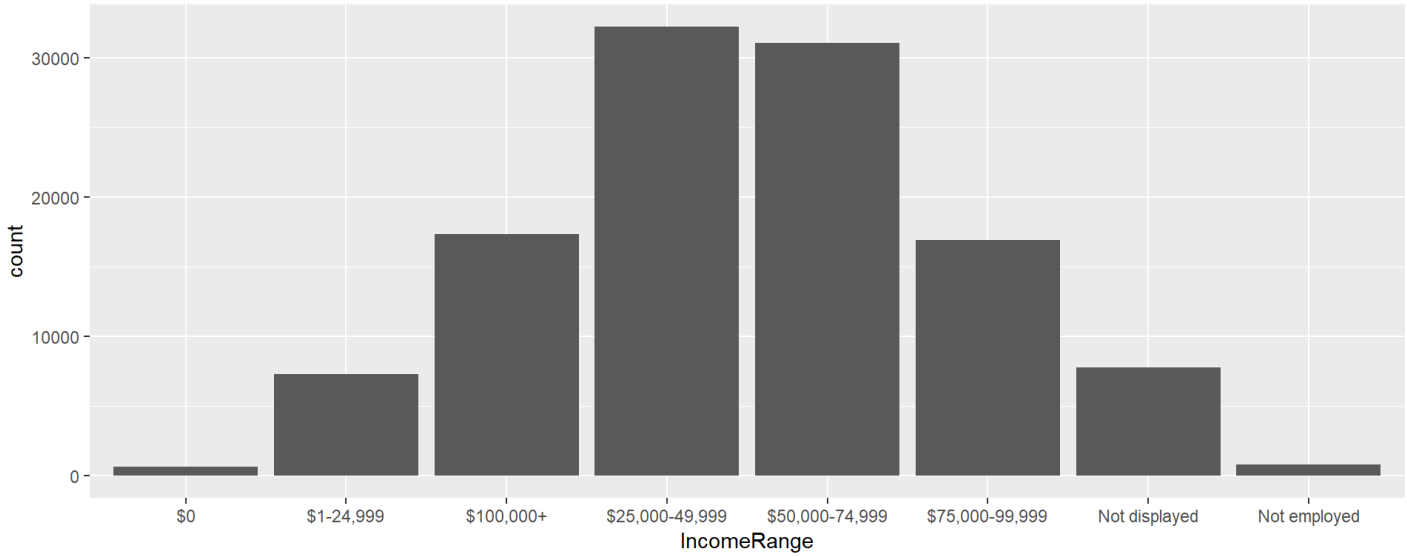


Borrowers in majority have this below 0.5 which means their income is more than their debt. There are outliers so only including till 1.

Debt to income ratio of majority of the masses s between 0 to 0.5 which is one of the criteria they follow to lend. However I do see outliers!

IncomeRange

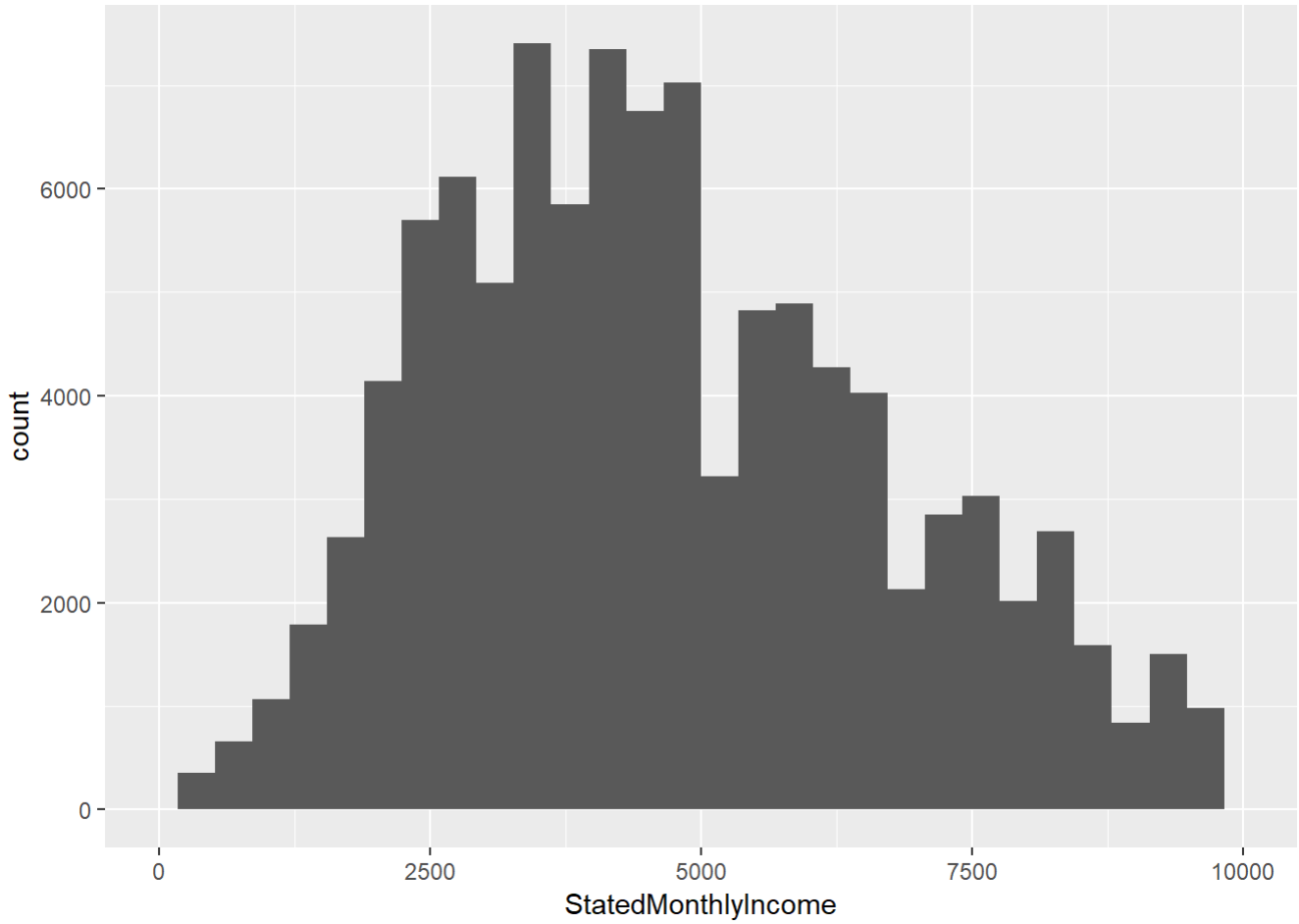
```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



Most of the loans given are for borrower's within the range of 25k - 100k.

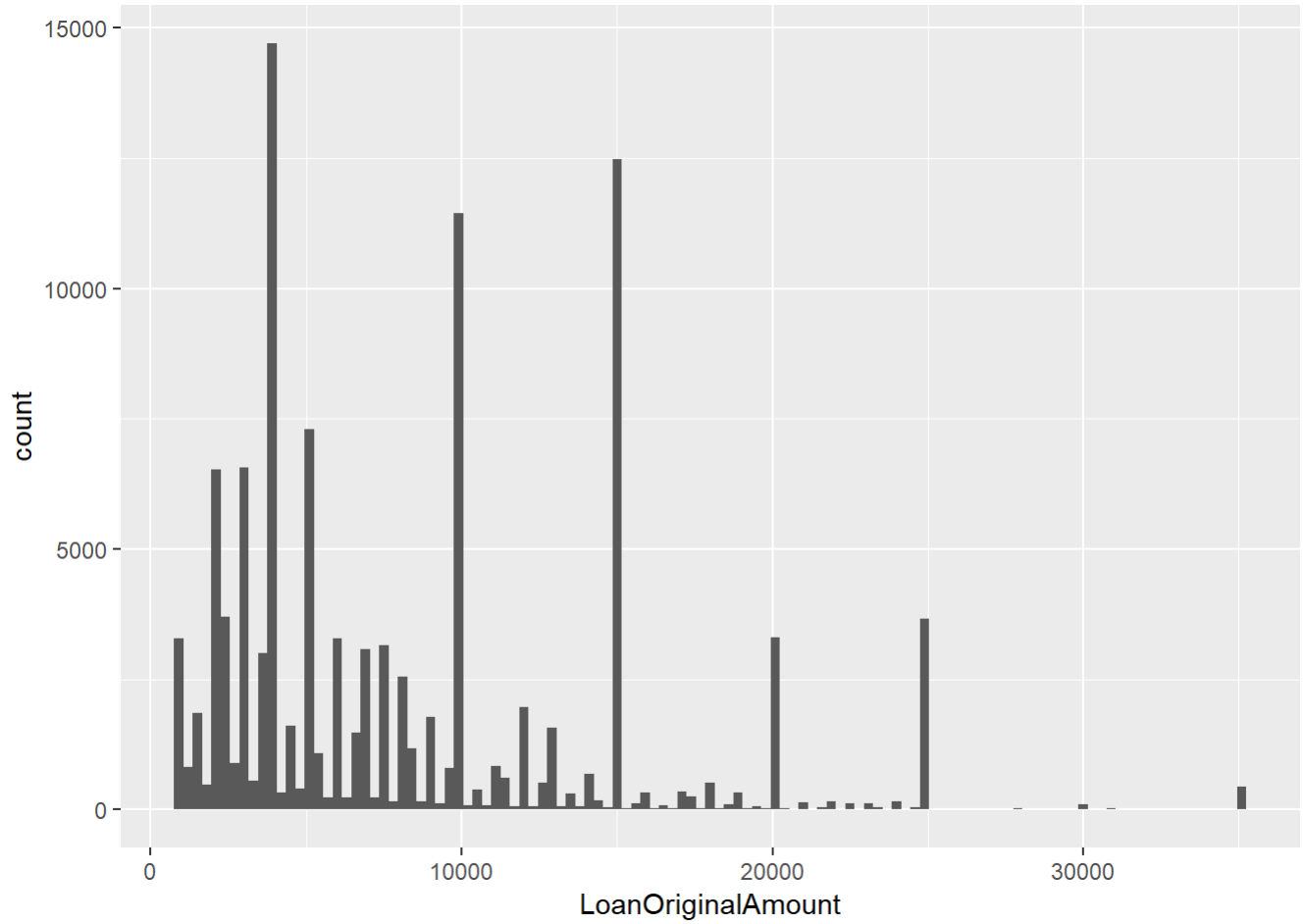
StatedMonthlyIncome

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



There were outliers so checking monthly payments withing $0 - 10000 only.

LoanOriginalAmount



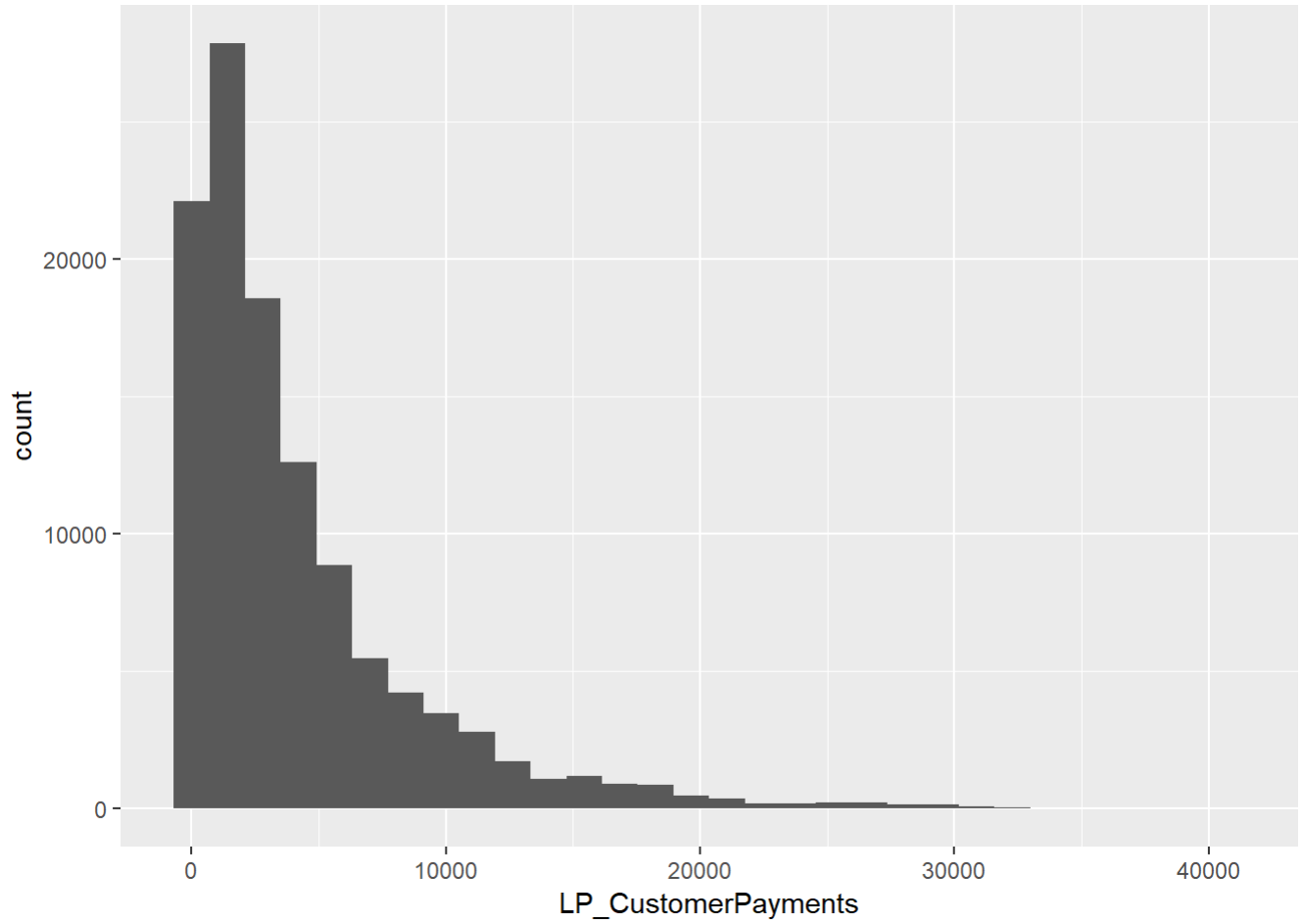
Loan amount varies between 0 to 30k. We can check the categories of each loan as well if we need to do further analysis.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1000    4000    6500    8337   12000   35000
```

Minimum amount lent is 1000 and max is 35000.

LP_CustomerPayments

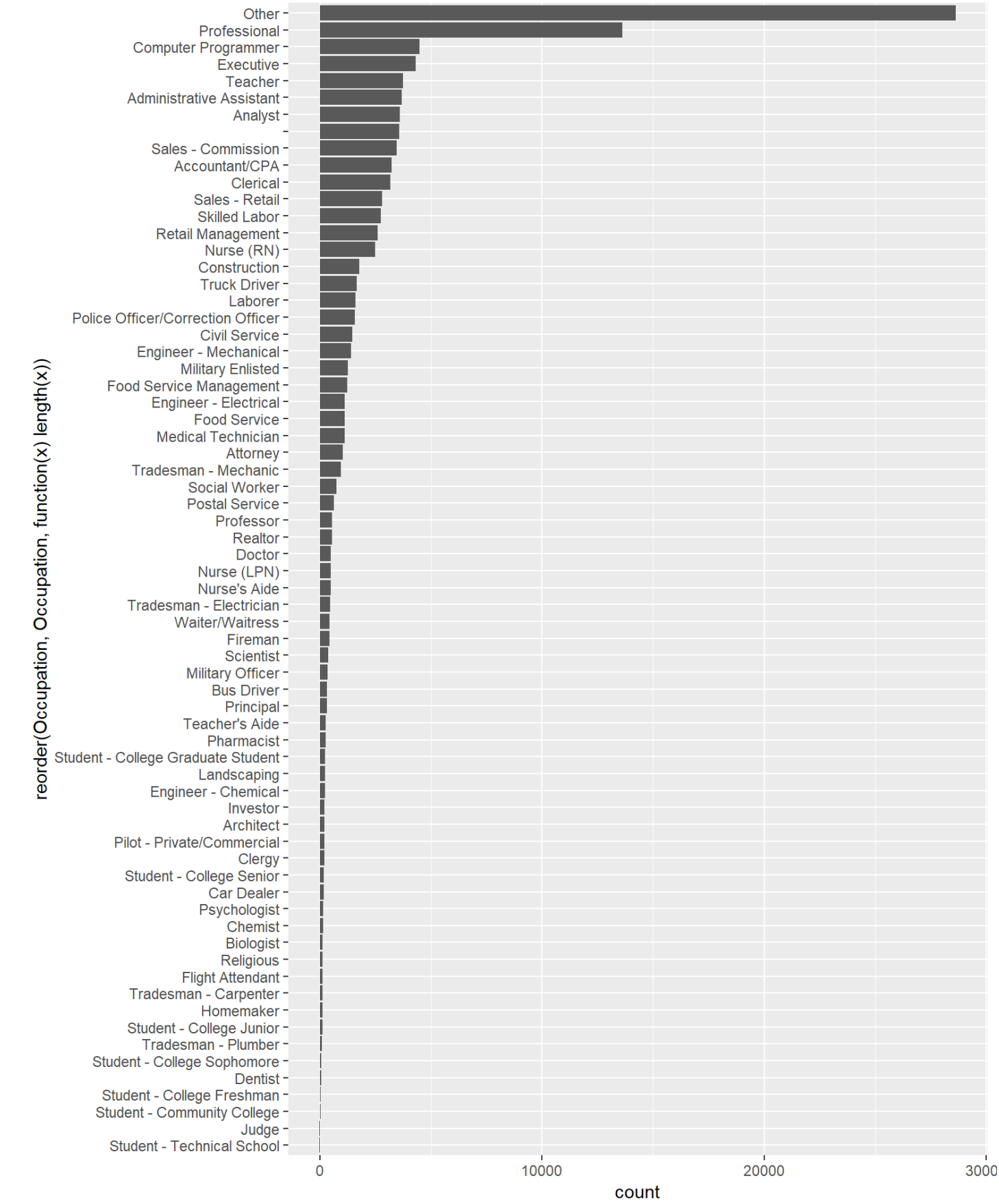
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Right skewed distribution. Most payments are between 0 to 10000. A summary of the payments made by borrowers:

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    -2.35  1005.76  2583.83  4183.08  5548.40 40702.39
```

Occupation



Occupation by loan count.

Summary of Univariate Analysis.

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

The main feature of interest is the LoanStatus variable that contains following categories:

Cancelled

Chargedoff

Completed

Current

Defaulted

FinalPaymentInProgress

Past Due (120 days)

Past Due (1-15 days)

Past Due (16-30 days)

Past Due (31-60 days)

Past Due (61-90 days)

Past Due (91-120 days)

It will be important to see what causes a borrower to become a defaulter. After doing some research the categorieschargedoff along with defaulted and past dues are all considered to be the defaulted borrowers and will be referred to as defaulters from now.

The main feature of interest would therefore be to investigate the variables associated with being a defaulter.

The main variables that will be used to investigate this are.

Term

BorrowerAPR

InquiriesLast6Months

DebtToIncomeRatio

IncomeRange

EmploymentStatus

CreditScoreRangeUpper

StatedMonthlyIncome

ListingCreationDate

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

The following will be used to see other trends within the dataset such as investors strategies to lend to borrowers and payments made by borrowers as compared to the original loan amount.

BorrowerState

LP_CustomerPayments

Investors

Occupation

IsBorrowerHomeowner

LoanOriginalAmount

LoanMonthsSinceOrigination

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

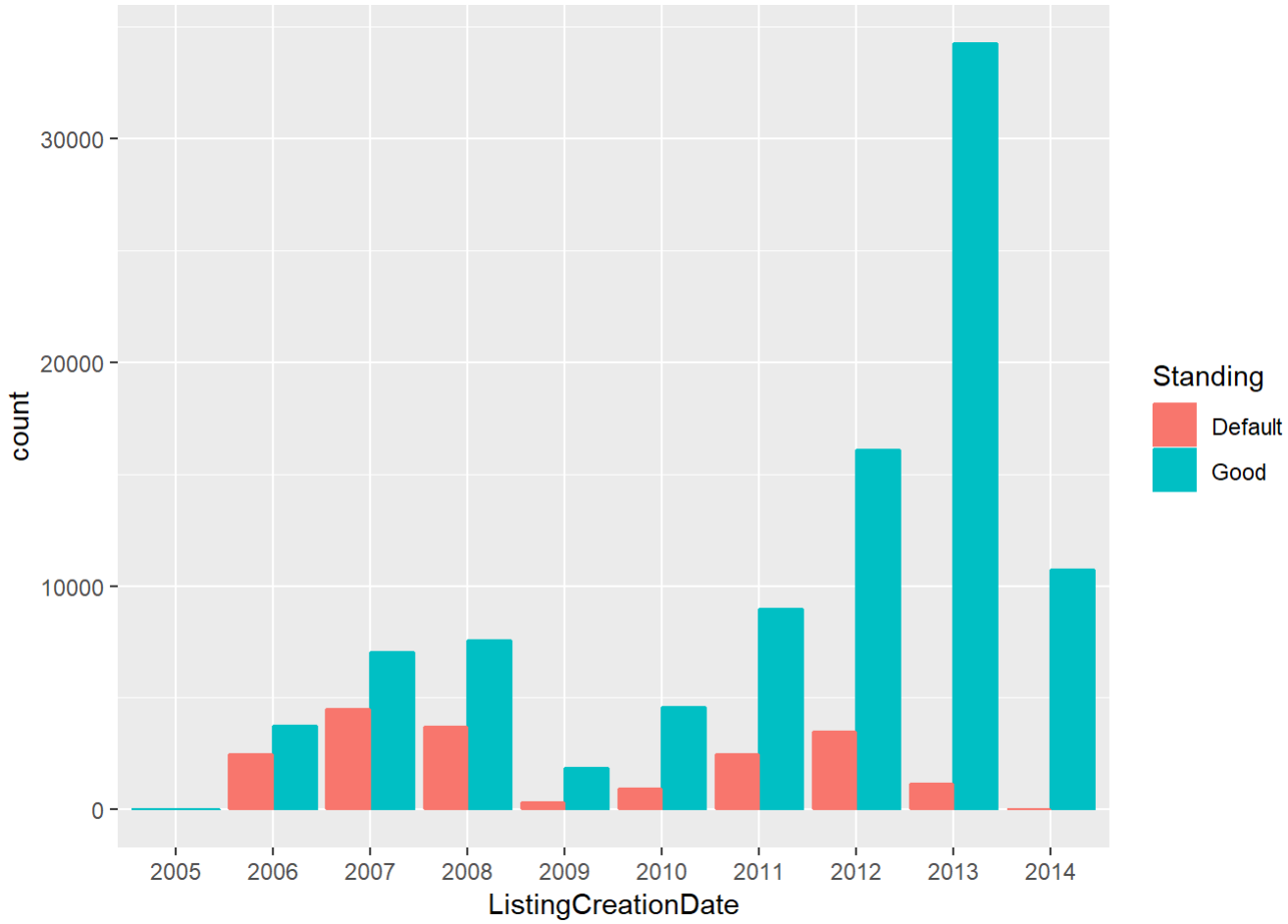
A new variable is created for the Default borrower using the LoanStatus variable.Listing creation year variable is also created from the original date format.

Of the features you investigated, were there any unusual distributions?
Most of the loans given are for borrower's within the range of 25k - 100k. However I do see a few borrower's

who have 0$ income. Not sure what that means but I also see a few in not employed (might be students). Most of the loans are given for 36 months.

Bivariate Plots Section

ListingCreationDate vs Standing

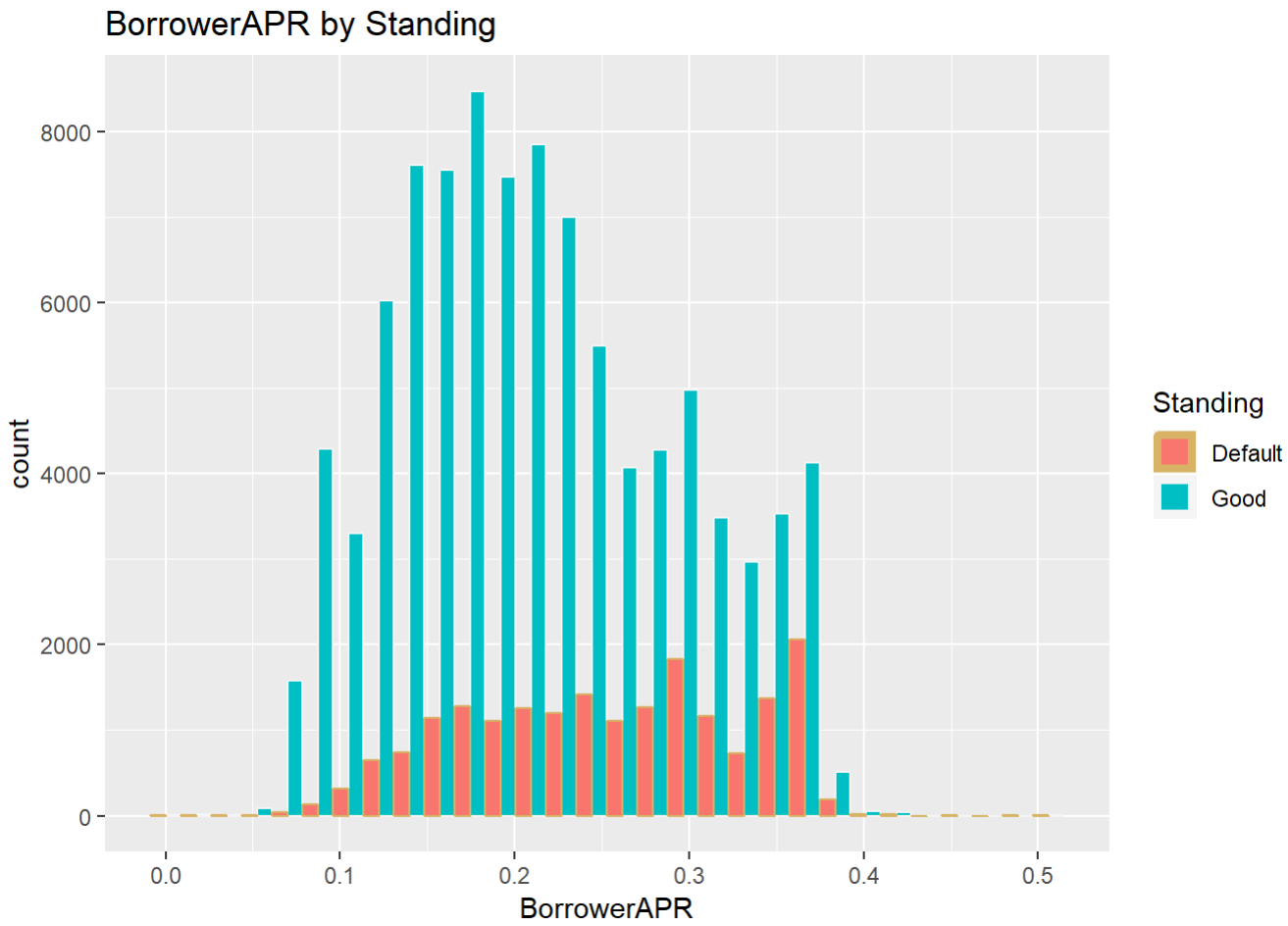


The number of loans are low during 2005 - 20111. Defaulters are mostly in 2007, the year before the financial crisis. Maybe they borrowed on a certain amount and during crisis were not able to pay their loans back. So, financial crisis might be a reason for the skewness seen here in defaulters and should be taken into account.

BorrowerAPR vs Standing

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

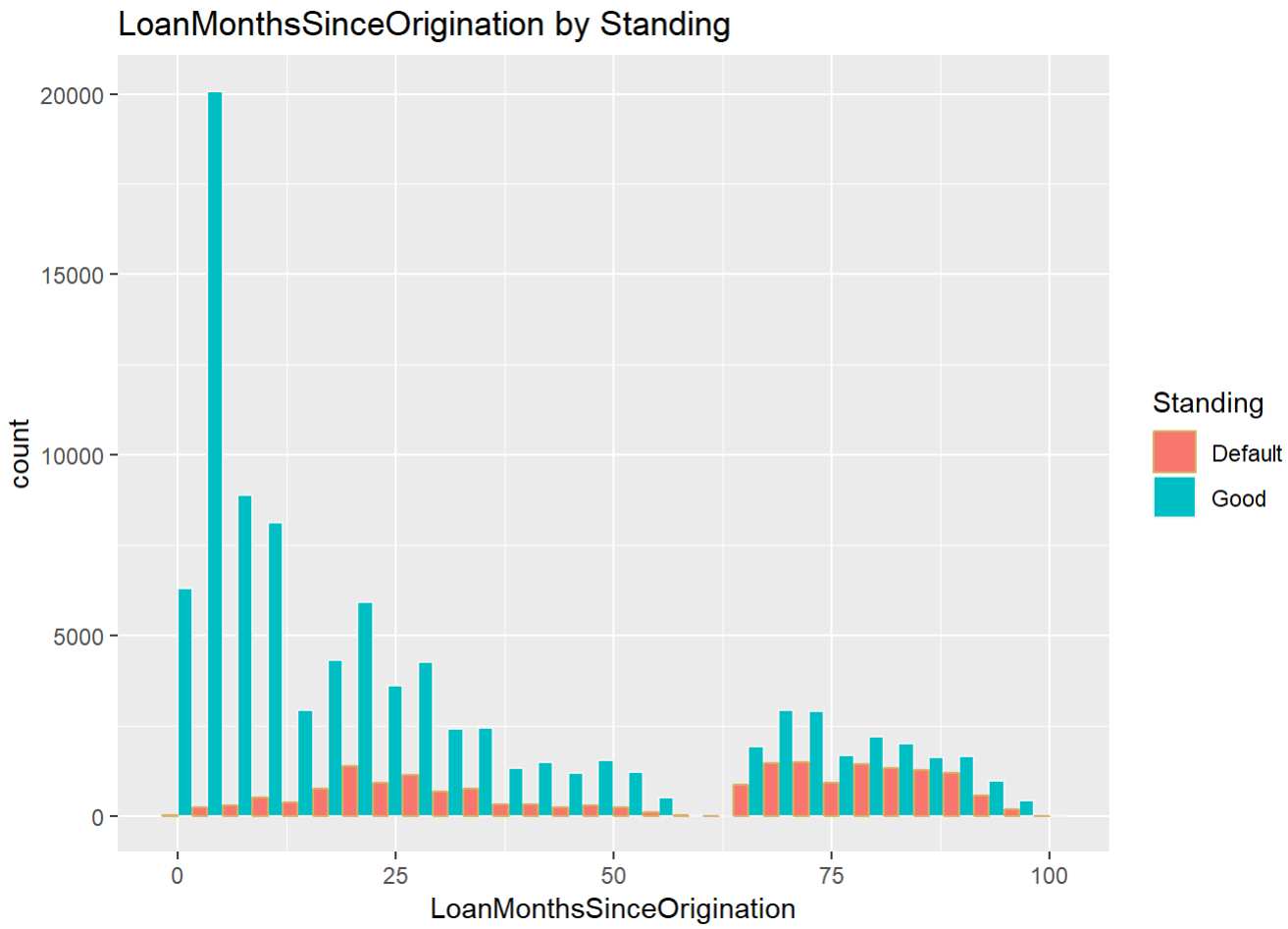
```
## Warning: Removed 25 rows containing non-finite values (stat_bin).
```



There is some trend seen between high APR and being a defaulter from 0.08 to 0.36 there is a direct relation.

LoanMonthsSinceOrigination vs Standing

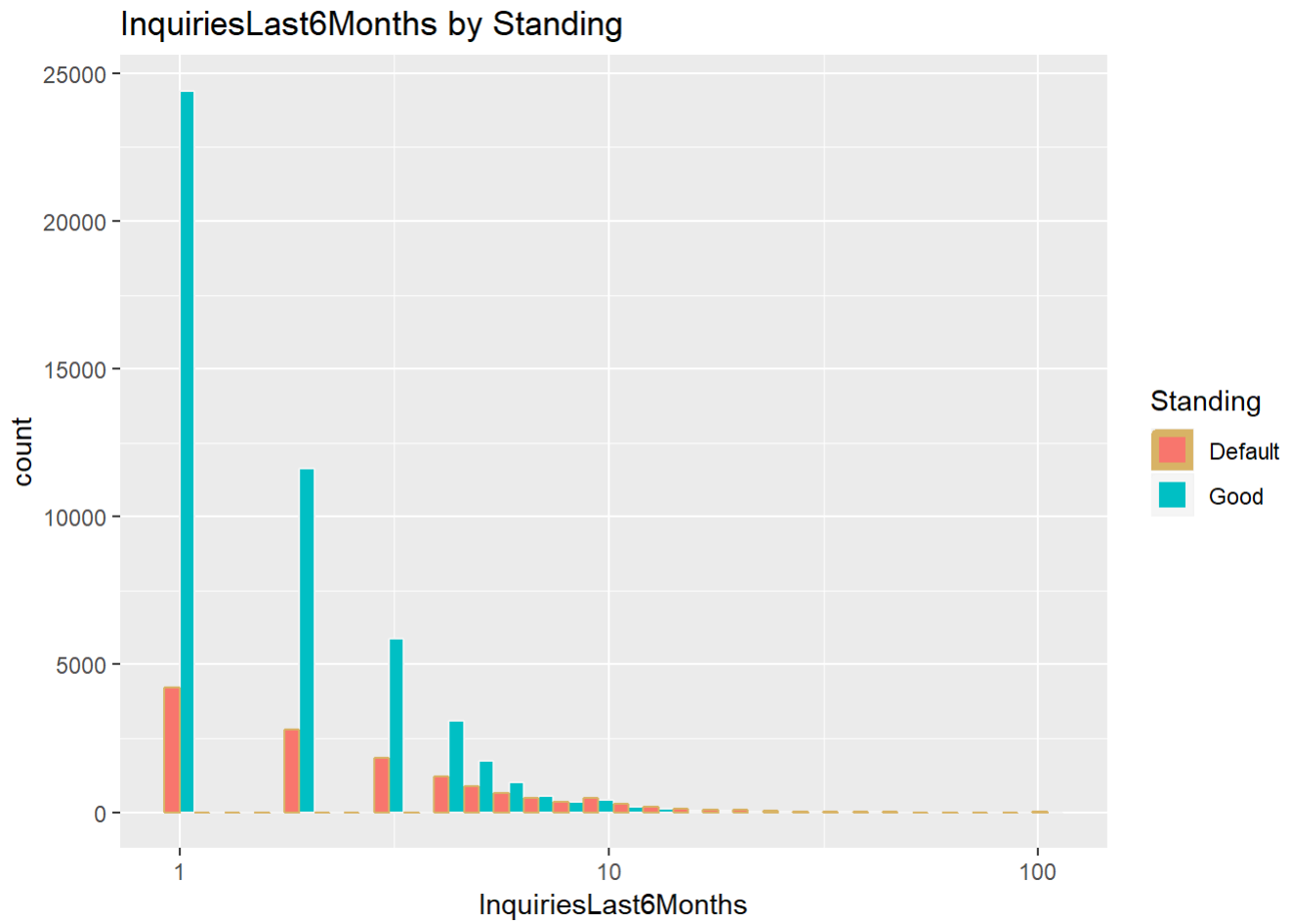
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



There is a spike around 24 and around 70 months. So I will later check their loan terms to see which loan terms usually go default

InquiriesLast6Months vs Standing

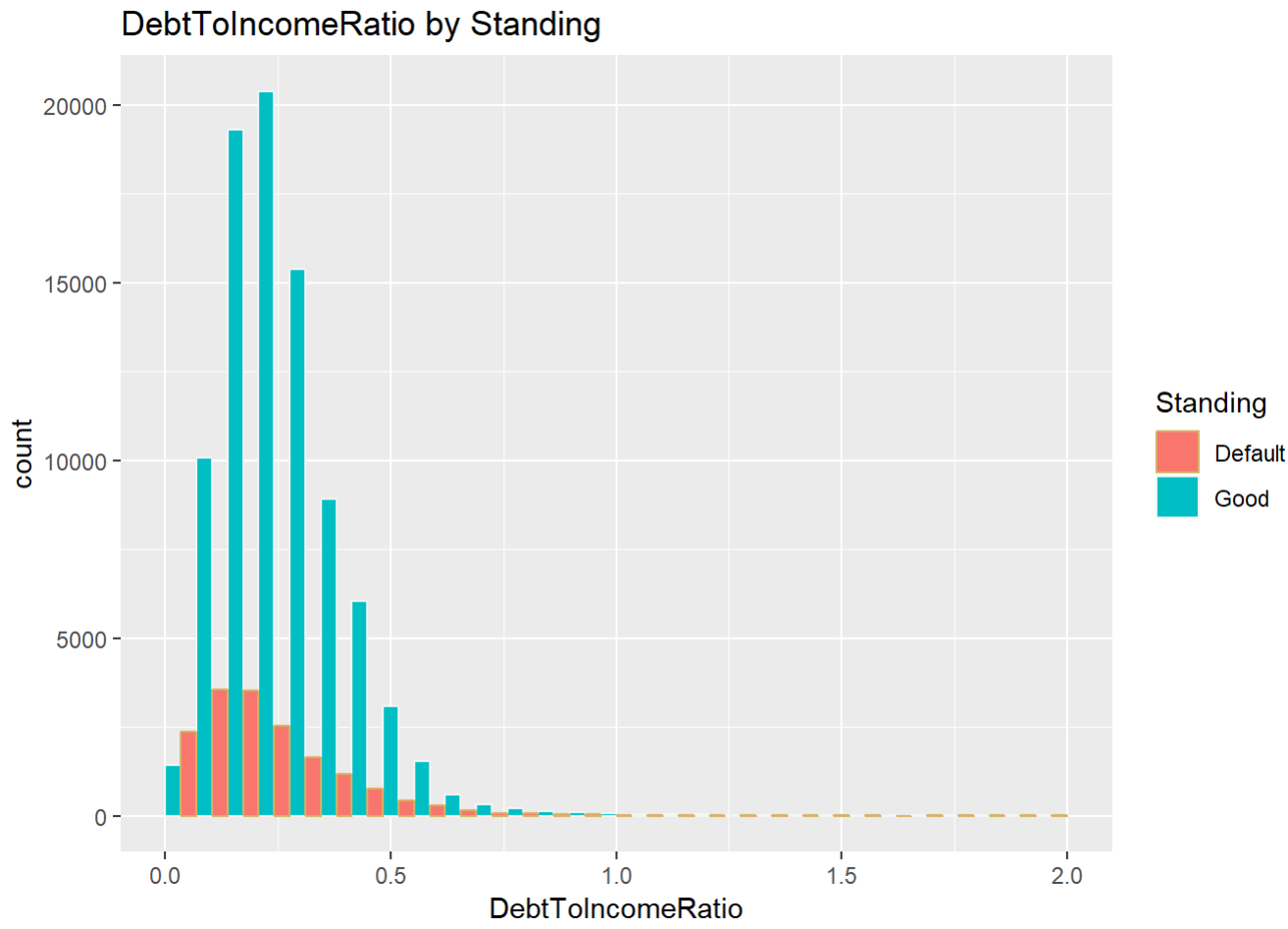
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Most of the defaulters have maximum 1 inquiries in last six months. So will check this later if this has something to do with credit score as well.

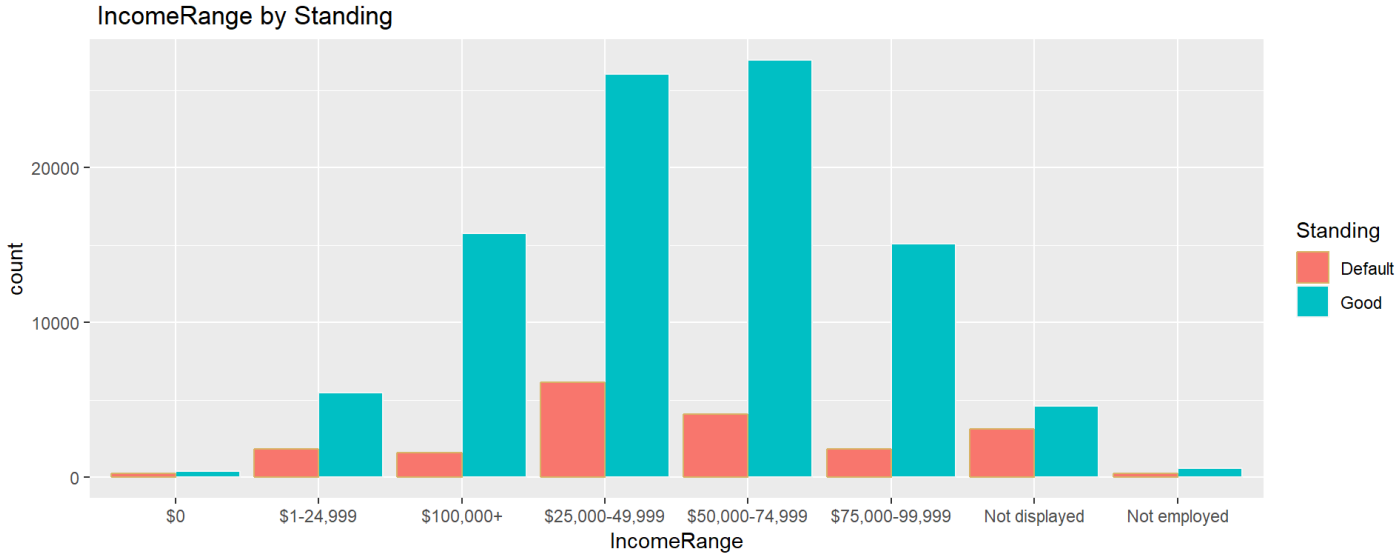
DebtToIncomeRatio vs Standing

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



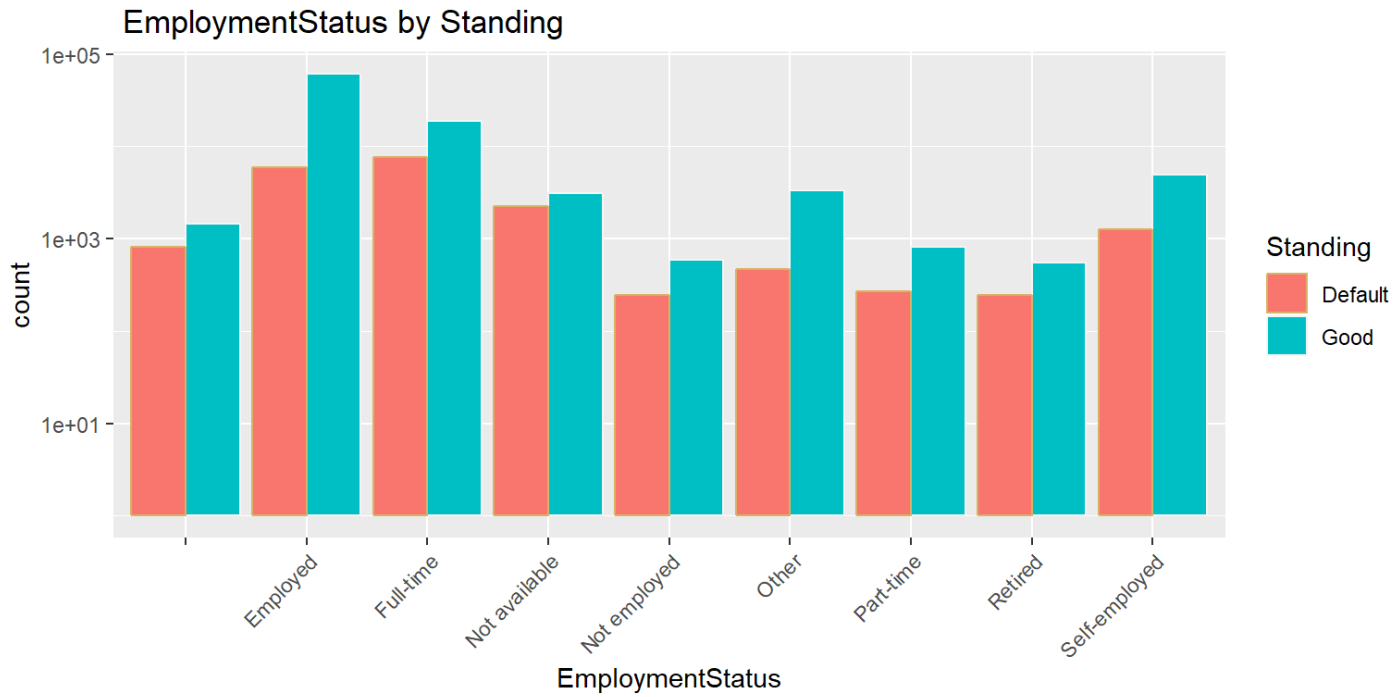
Similar trend of defaulters and those in good standing but obviously different in counts.

IncomeRange vs Standing



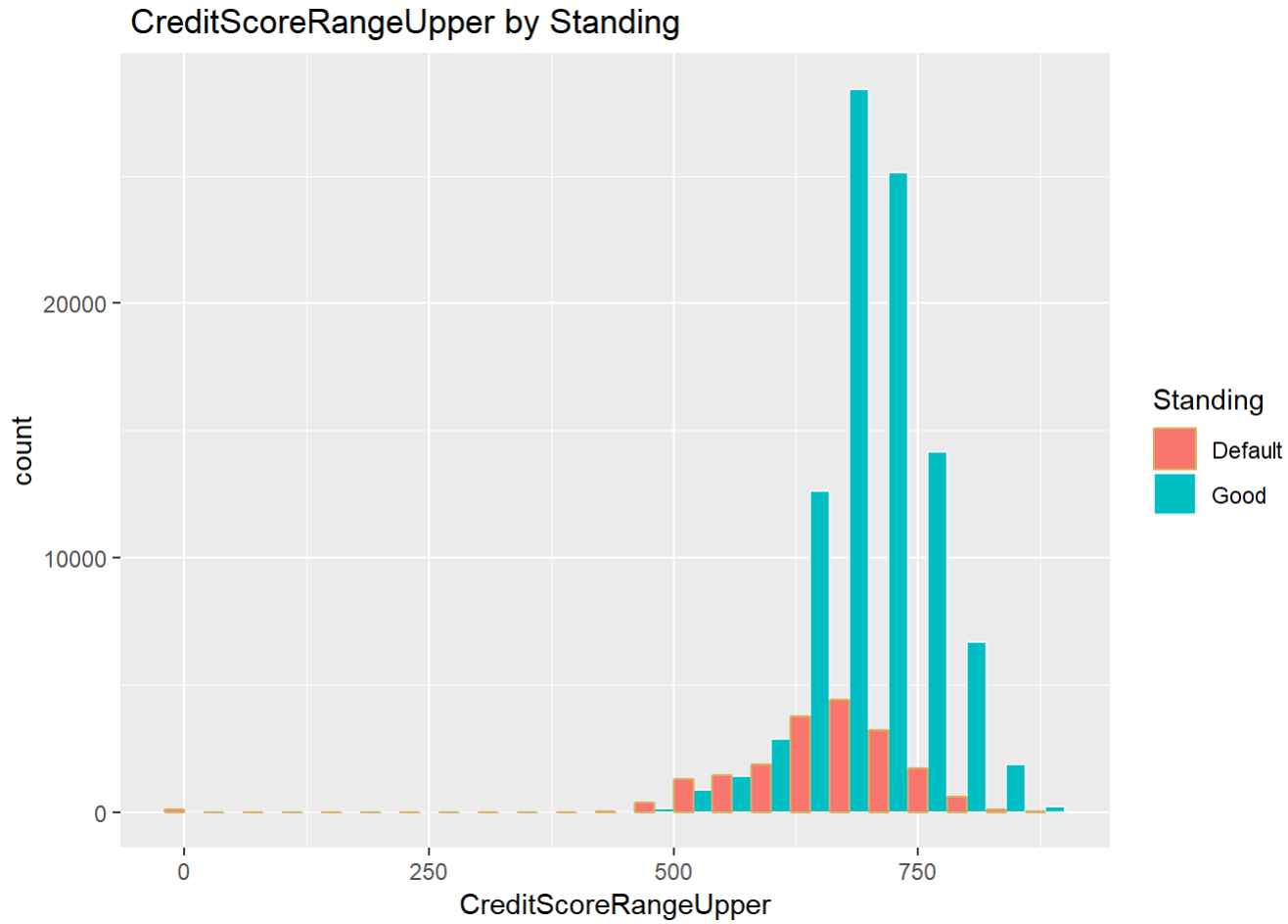
The $25k to 50k range borrower's have most number of defaulters but this might be due to the fact that this category has majority of borrowers.

EmploymentStatus VS Standing



Employment status vs. standing. All categories seem to be having almost equal distribution of defaulter as compared to non defaulters when log transformed the y axis. Data was skewed towards employed borrowers before. Also, defaulters are almost always less than borrowers in each category.

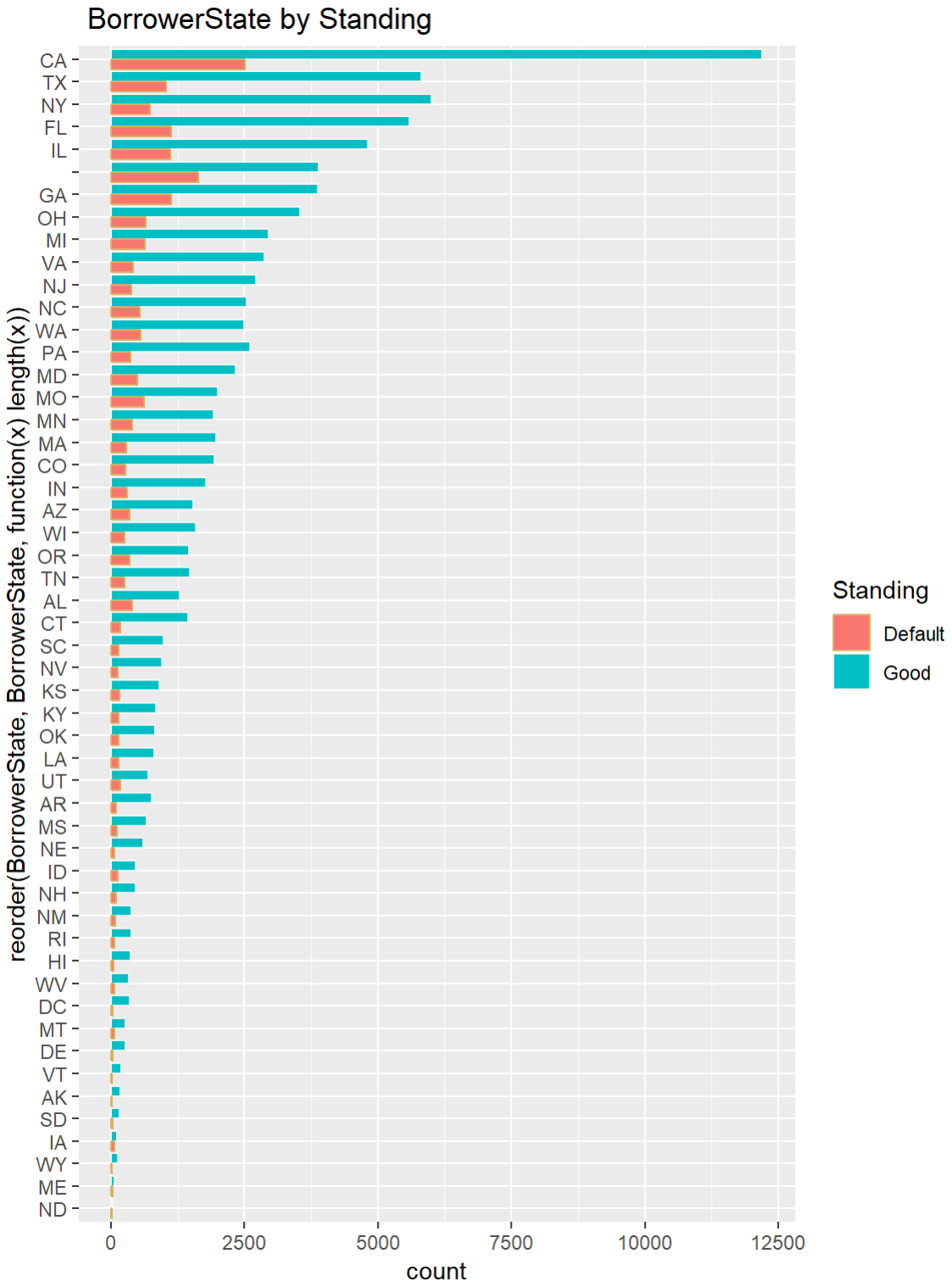
CreditScoreRangeUpper vs Standing



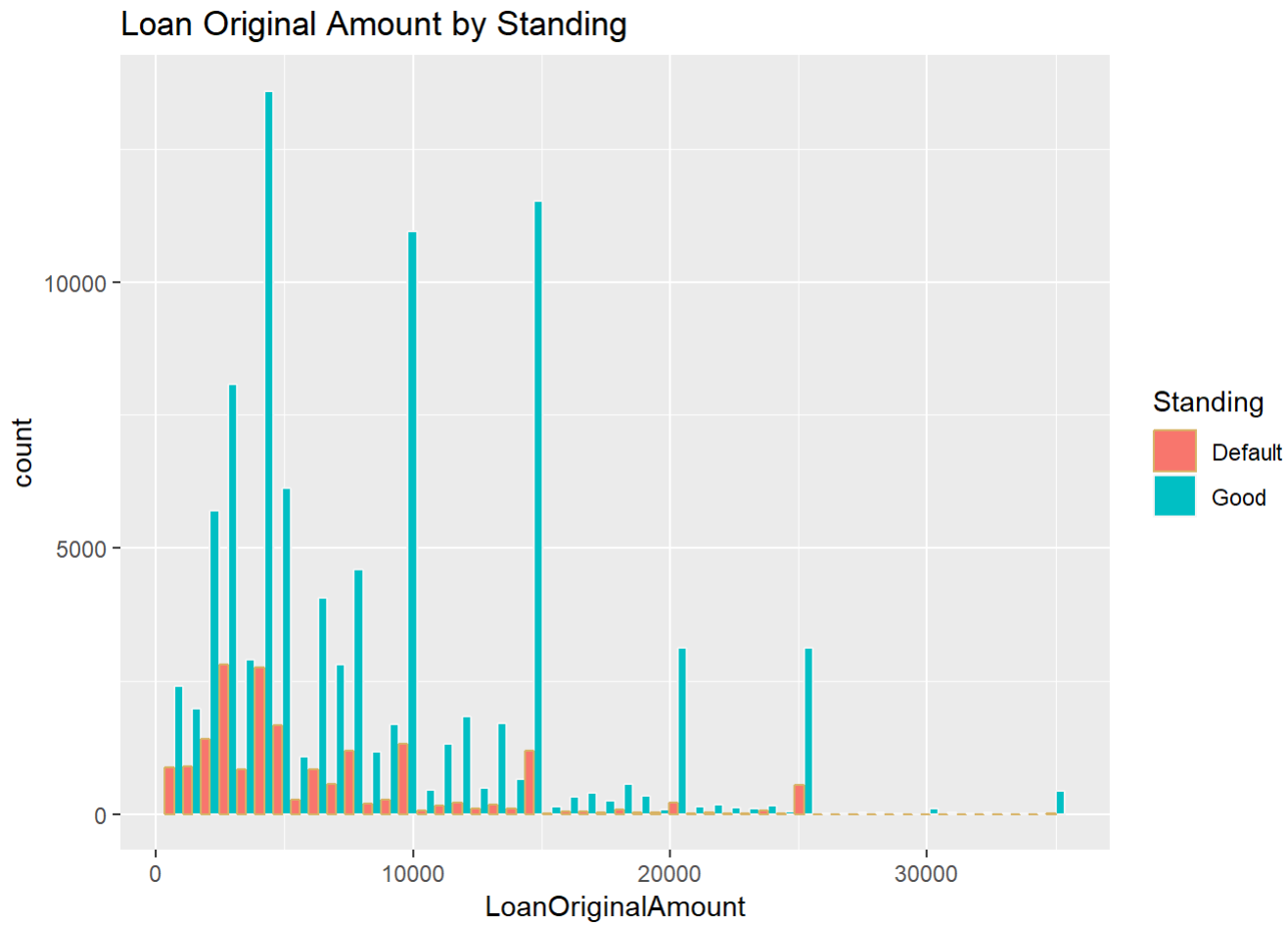
Less defaulters for high credit rating. Most defaulters are in the range 650 - 700 rating.

BorrowerState vs Standing

ggplot(aes(, y=Investors),data=plds) + geom_point() + coord_flip()



LoanOriginalAmount by Standing

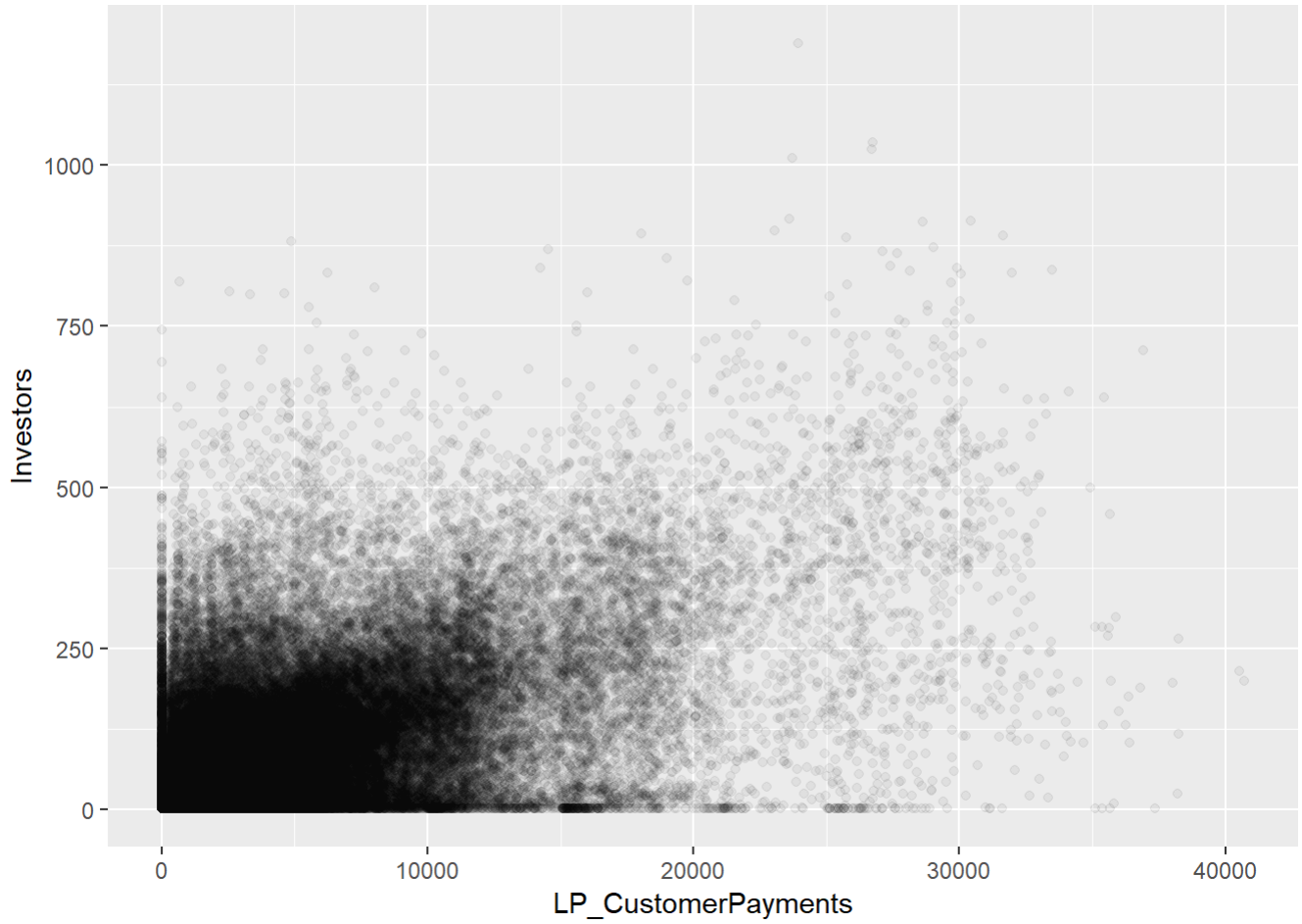


Loan amounts of defaulters is almost always less than the others for the same amount.

The distribution was right skewed so logtransformed the y axis. The skewness was due to the high population of CA now it looks more nice. No state has more defaulters as compared to good standings.

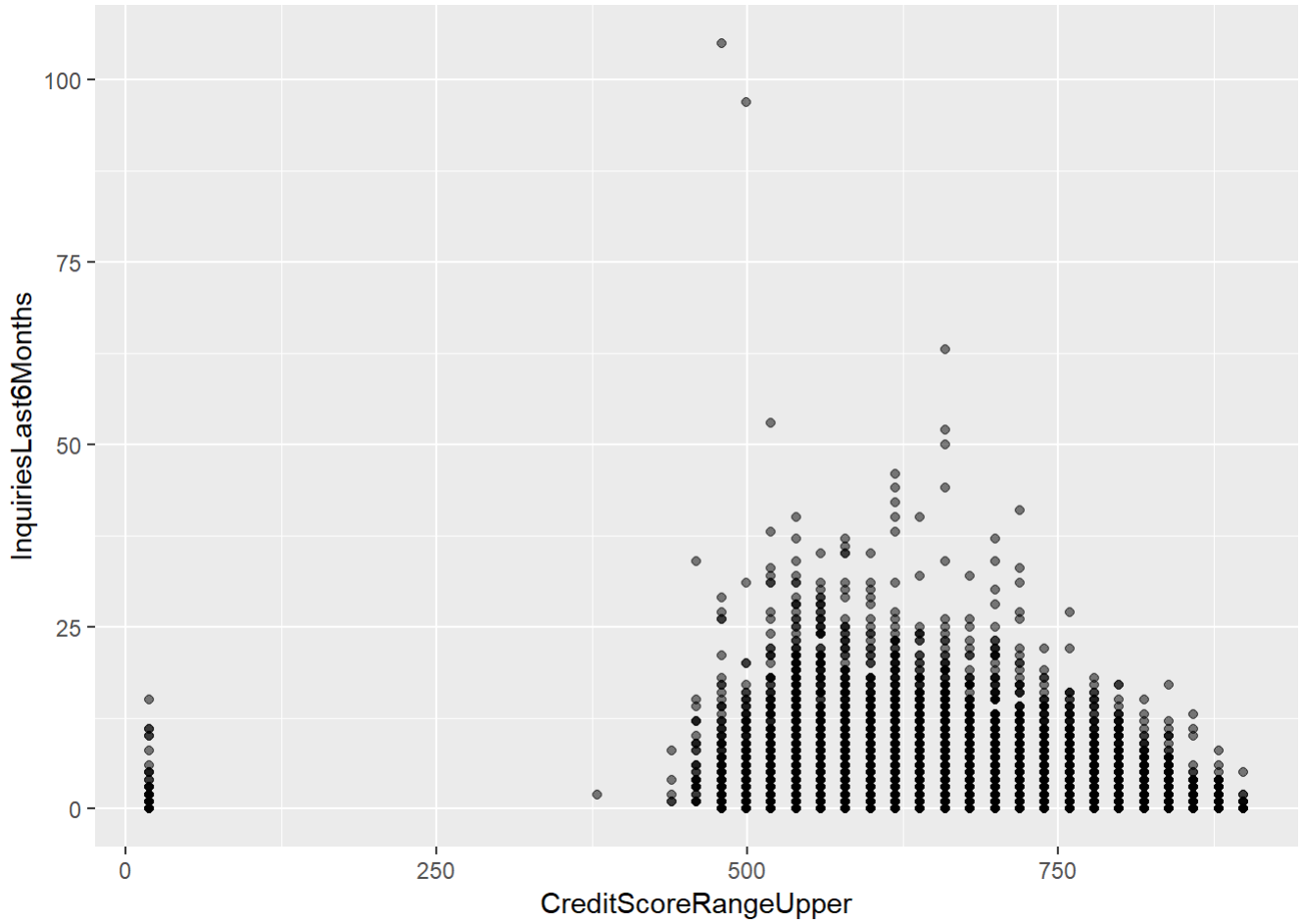
Exploring other trends in the data by seeing interactions between different variables.

LP_CustomerPayments vs Investors



There is some relation between these two so the distribution of good and default borrowers will be interesting to see in multivariate analysis.

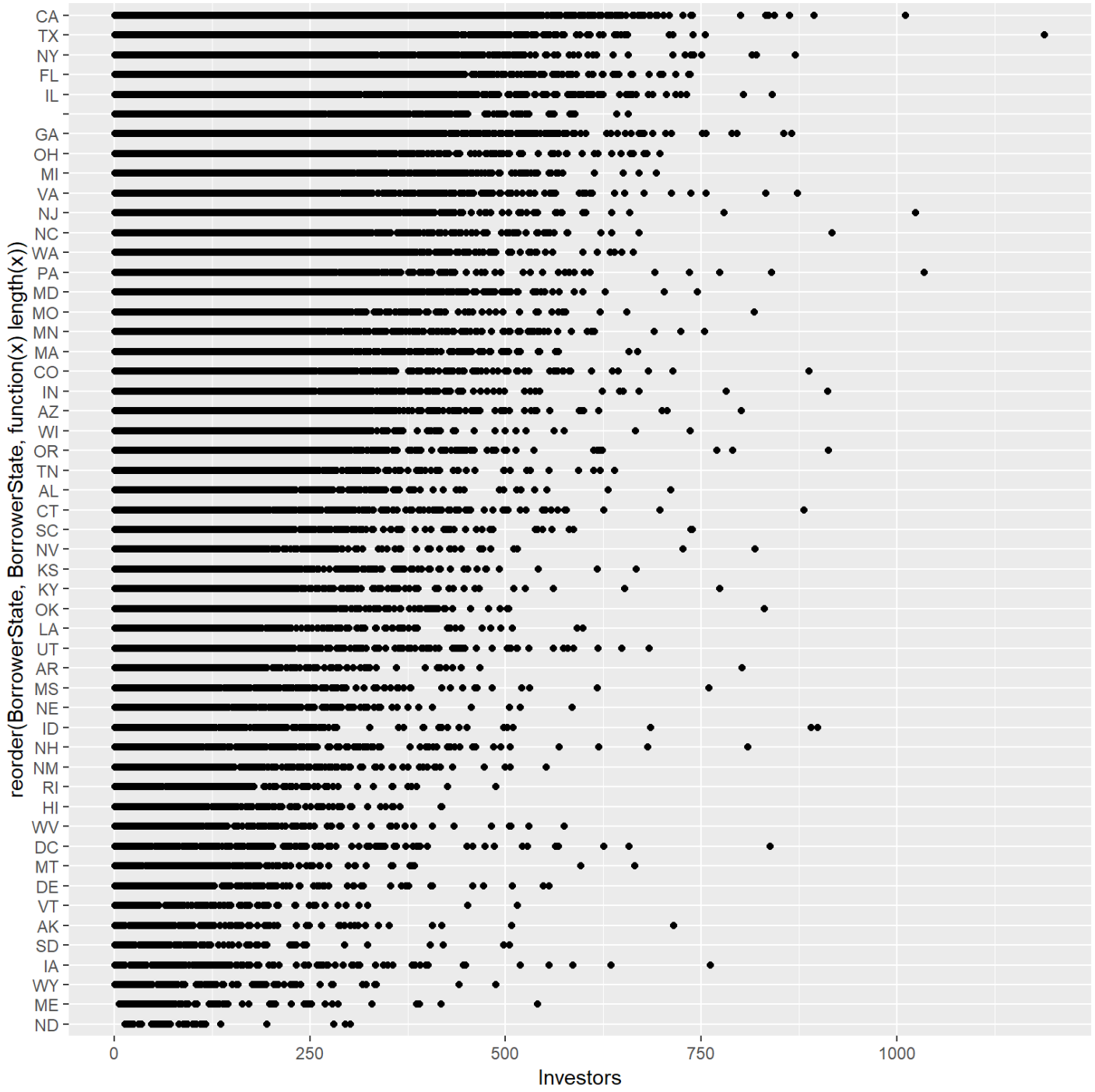
CreditScoreRangeUpper vs InquiriesLast6Months



Inquiries increase initially from 450 to 550 but then decreases so it is expected that defaulters are in the lower side of the rating and with more inquiries. Will be checked in multivariate section.

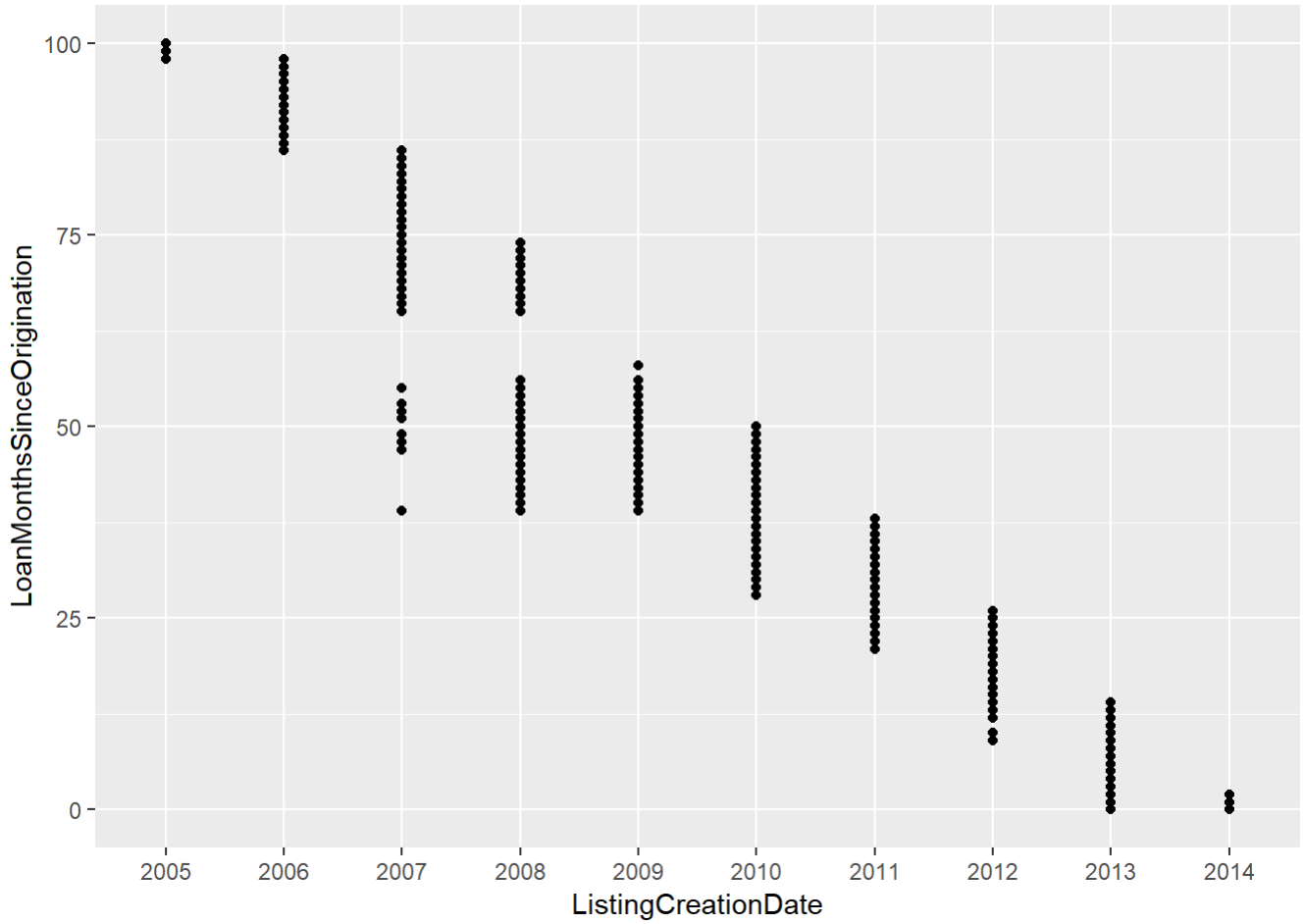
BorrowerState vs Investors

I was curious to know if investors prefer any states so checking this below:



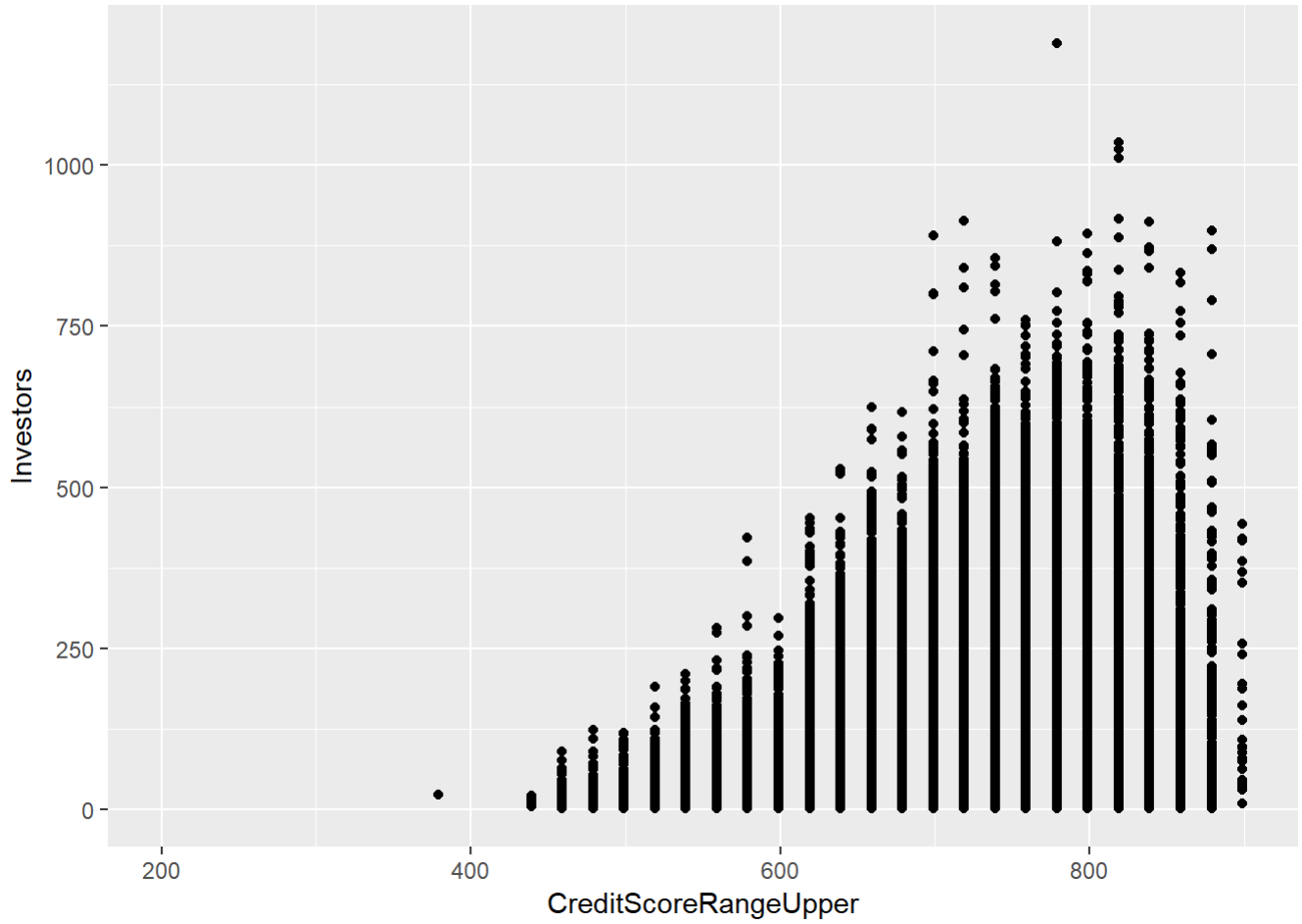
Seeing the relation between investors interest for certain states doesn't show any specific trends.

Investors vs LoanMonthsSinceOrigination



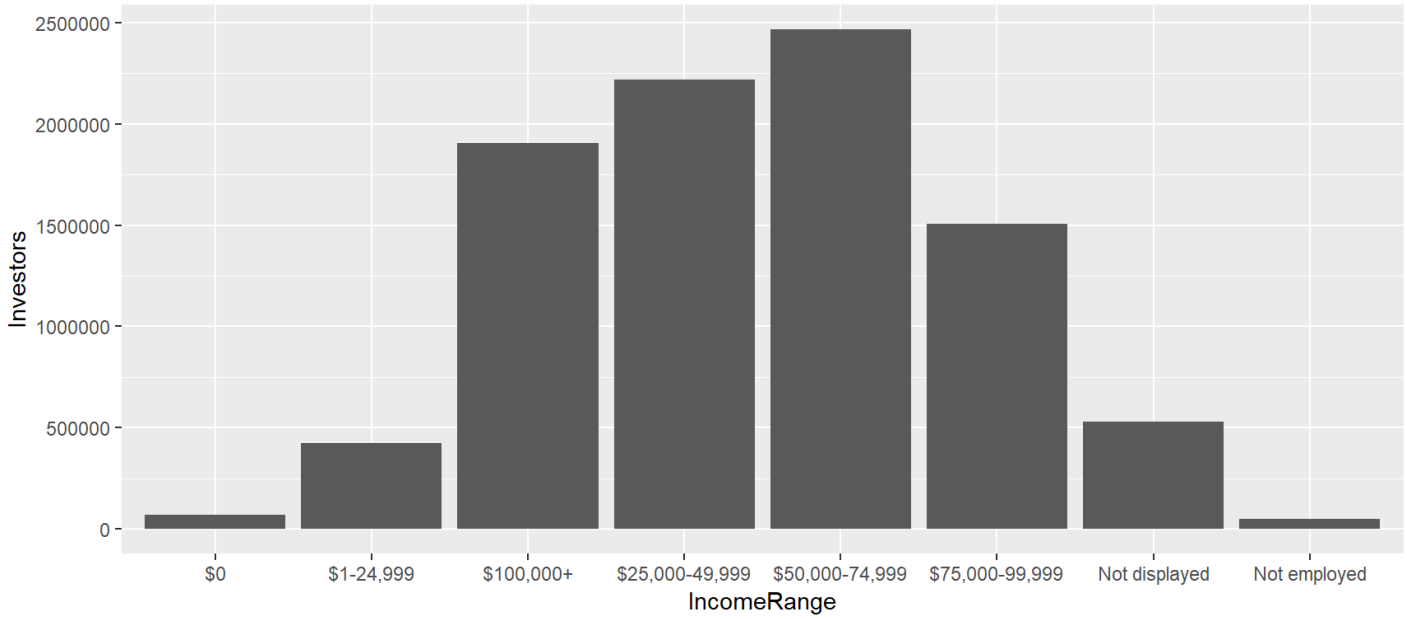
There seem to some more delinquicy from 2005 to 2010 so will be interesting to see the data for good and default borrowers in multivariate analysis for the same plot.

CreditScoreRangeUpper vs Investors



Strong relation seems to be existing between these. Investors seems to be trusting high credit scores.

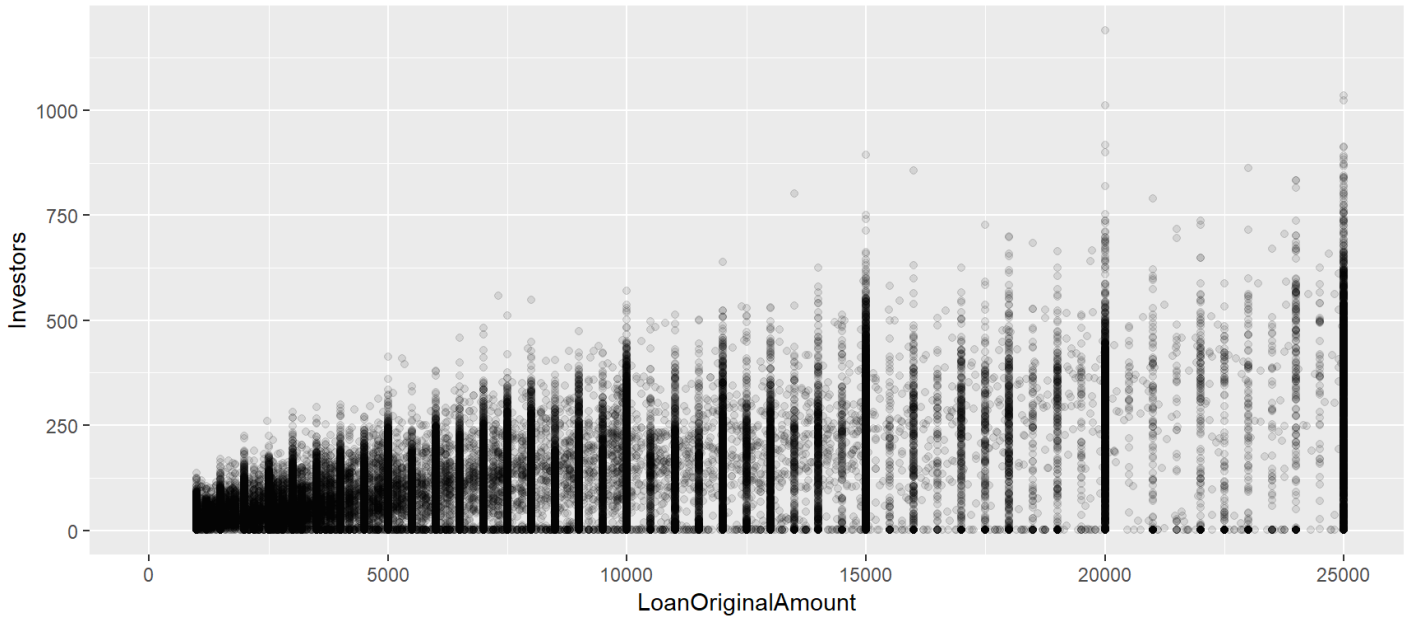
IncomeRange vs Investors



Most investors investing into the employed masses.

LoanOriginalAmount vs Investors

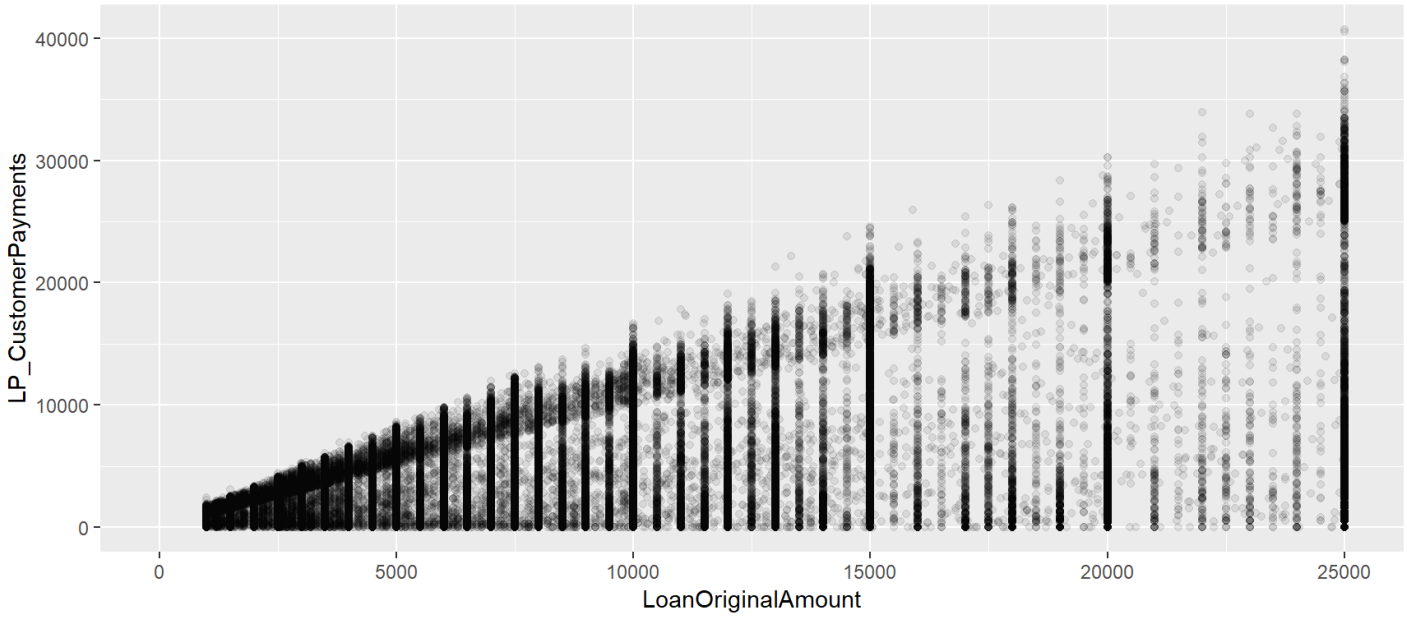
```
## Warning: Removed 680 rows containing missing values (geom_point).
```



As expected the more the loan amount the more investors. (Excluding top 1 % loan amount)

LoanOriginalAmount vs LP_CustomerPayments

```
## Warning: Removed 680 rows containing missing values (geom_point).
```



Higher the loan amount more payments are made by the borrower. (Excluding top 1 % loan amount)

Summary of Bivariate plots

The employed borrowers are mostly the highest in defaulters

Number of investors increase with high credit scores

Strong inverse relation between monthly income and debt to income ratio as expected

When analyzing loan months since origination, there was a spike around 24 and after 60 months. So I will later check their loan terms to see which loan terms usually has most defaulters.

What was the strongest relationship you found?

Investors and credit scores

The strong relation was between Loan months since origination and listing creation.

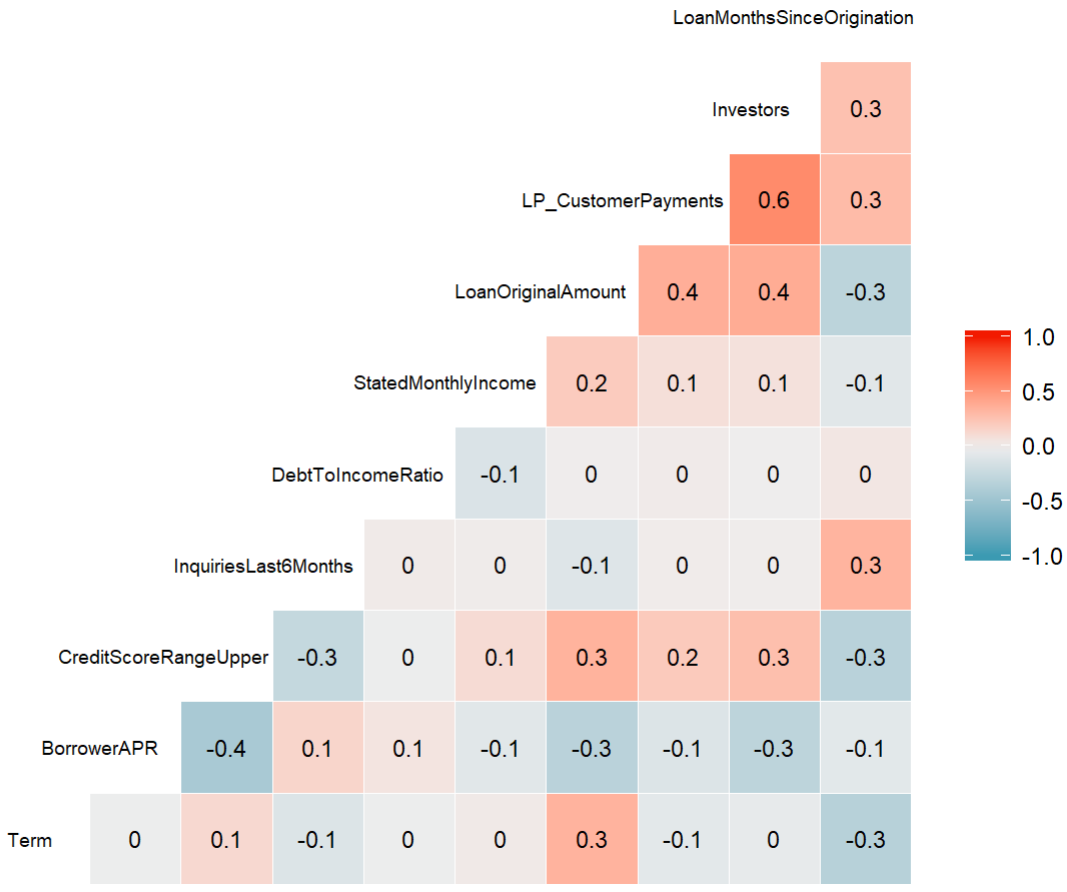
LP_customerpayments and investors

Investors and credit score

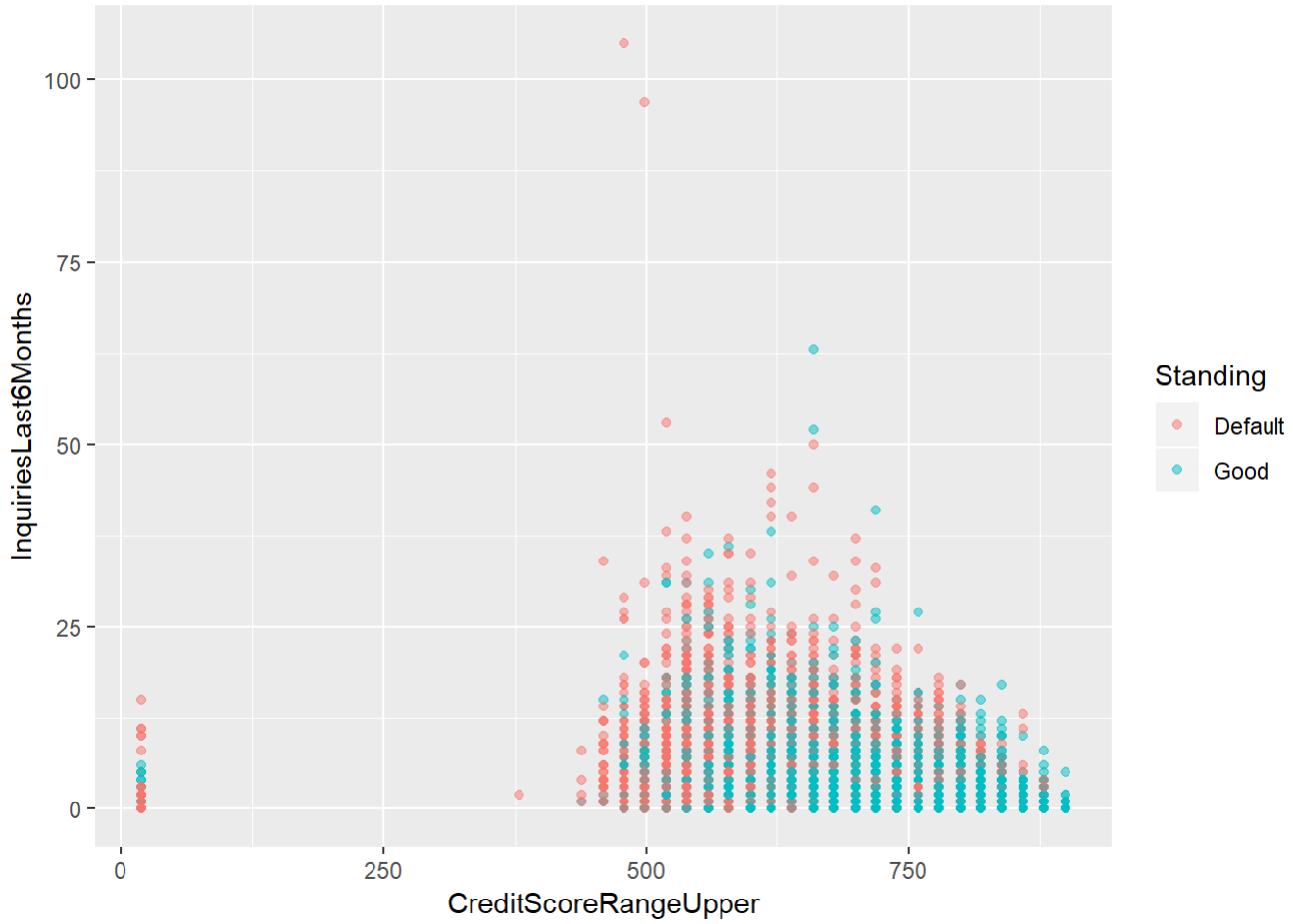
Income range and investors

Multivariate Plots Section

Correlation Analysis

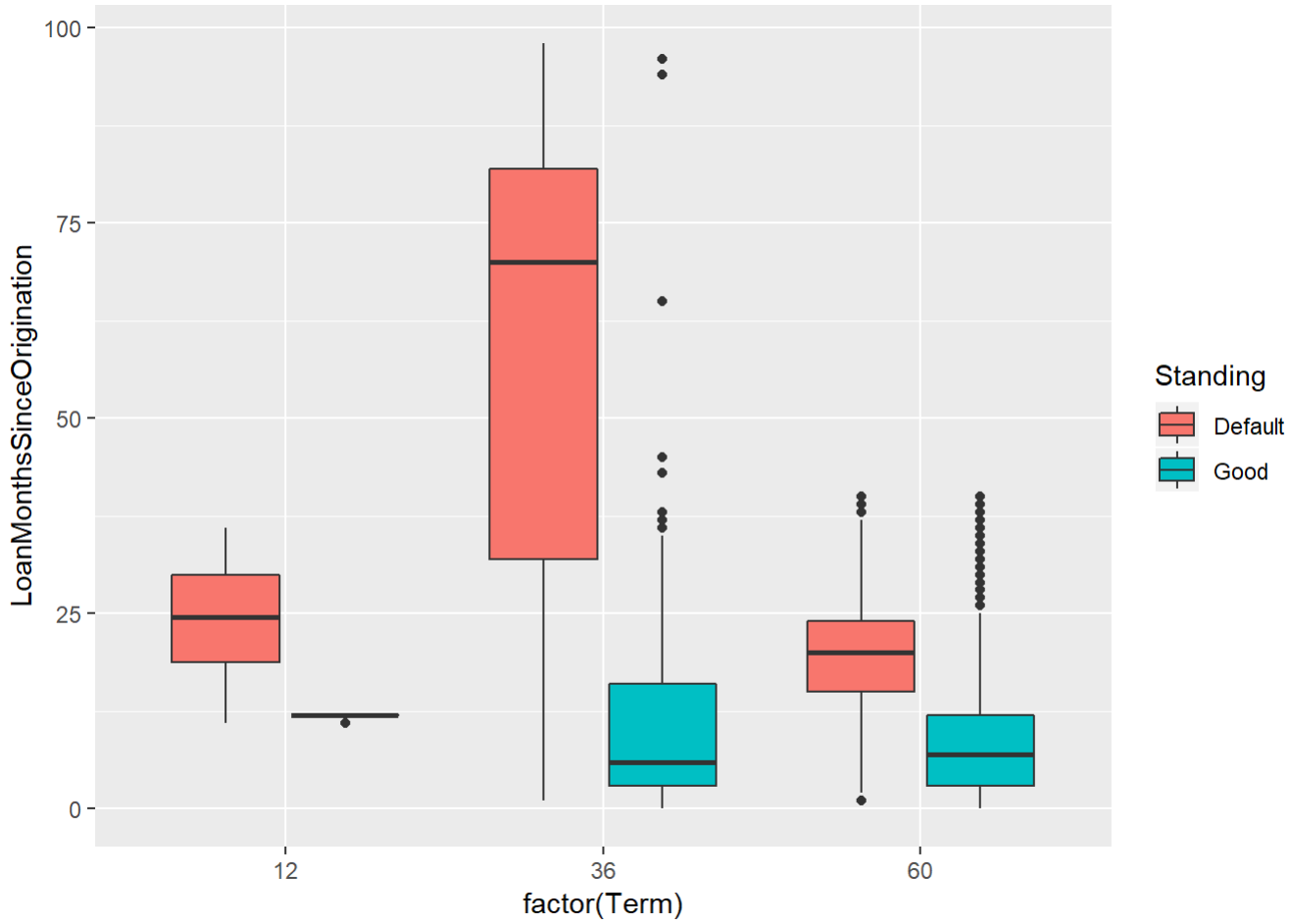


Looking at the correlation coefficients of quantitative variables only. There seem to be some relation of investors with lp_customerpayments as well as loan original amount. There is inverse relation between debt to income ratio and monthly income as expected.There is inverse relation between loan original amount and inquiries last 6 months. CreditScoreRangeUpper and InquiriesLast6Months does not seem to have any relation but still I am curious to see it by standing so beginning the analysis with these.



As expected more defaulters on lower side of the rating but there are also some on the higher side!

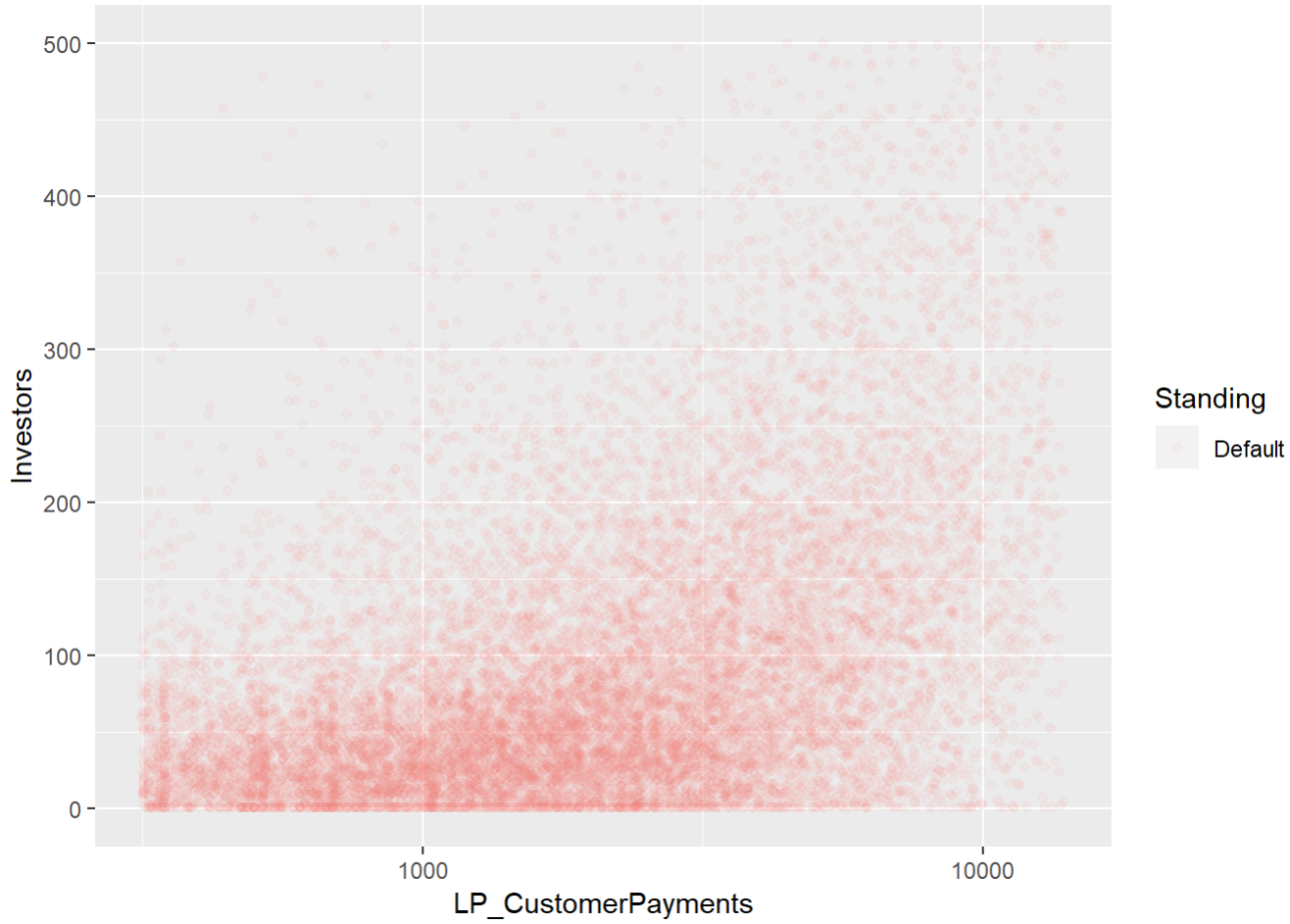
LoanMonthsSinceOrigination vs Term vs Standing



Excluding the borrowers who have completed their loans to make the comparison. Defaulters as expected spent more months than what they expressed initially to pay off their loans. Most defaulters are for the 36 term duration since most loans are also for this term. Median for loan months since origination for defaulters of 36 month term is around 70. Which means 50% are below this and 50% above this.

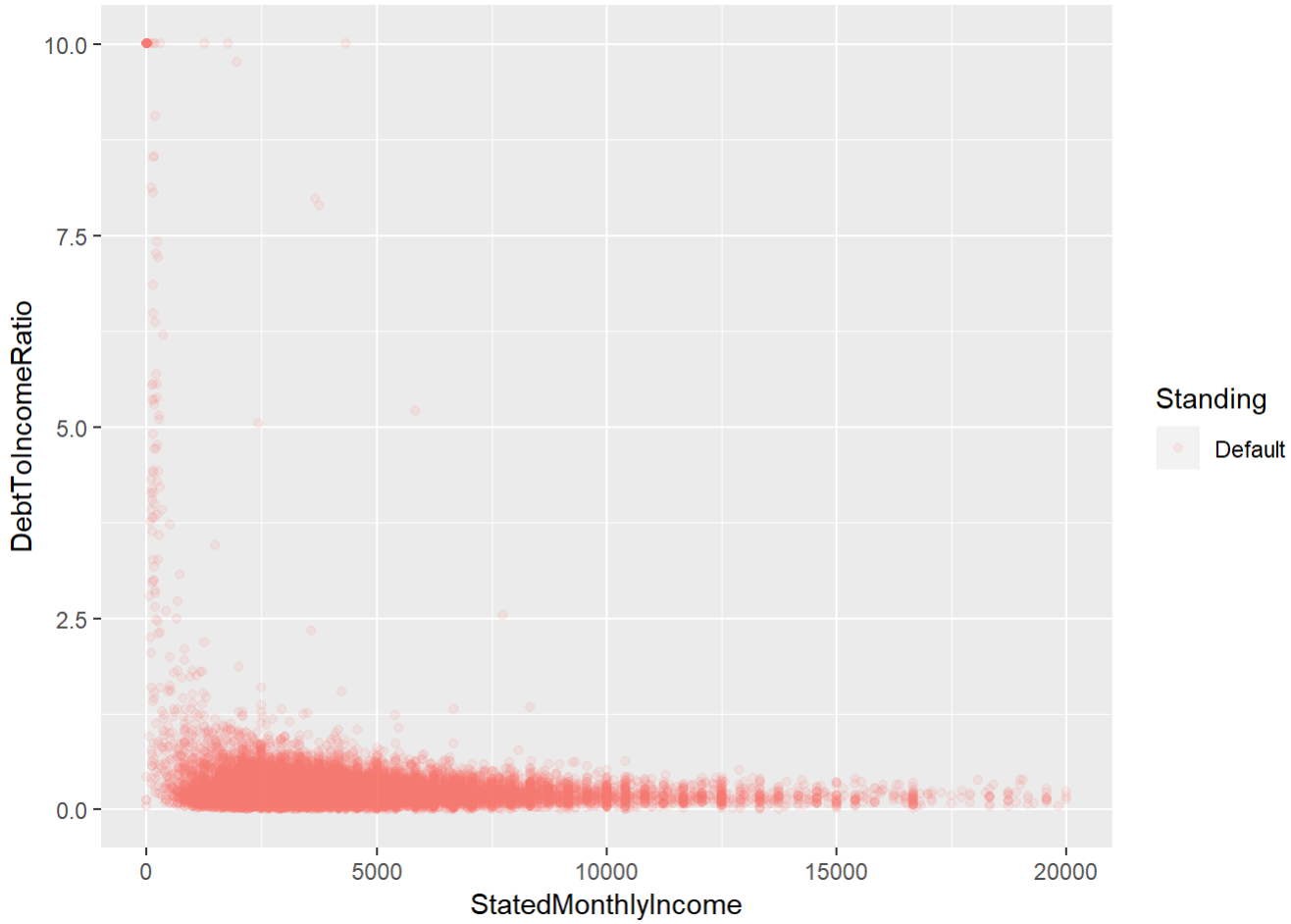
The 60 month duration term doesn't have any borrower exceeding the limit of 60 months. For 12 month term almost all defaulters have exceeded the duration of term. Months for defaulters is always high as compared to non defaulters.

Investors vs LP_CustomerPayments by Standing



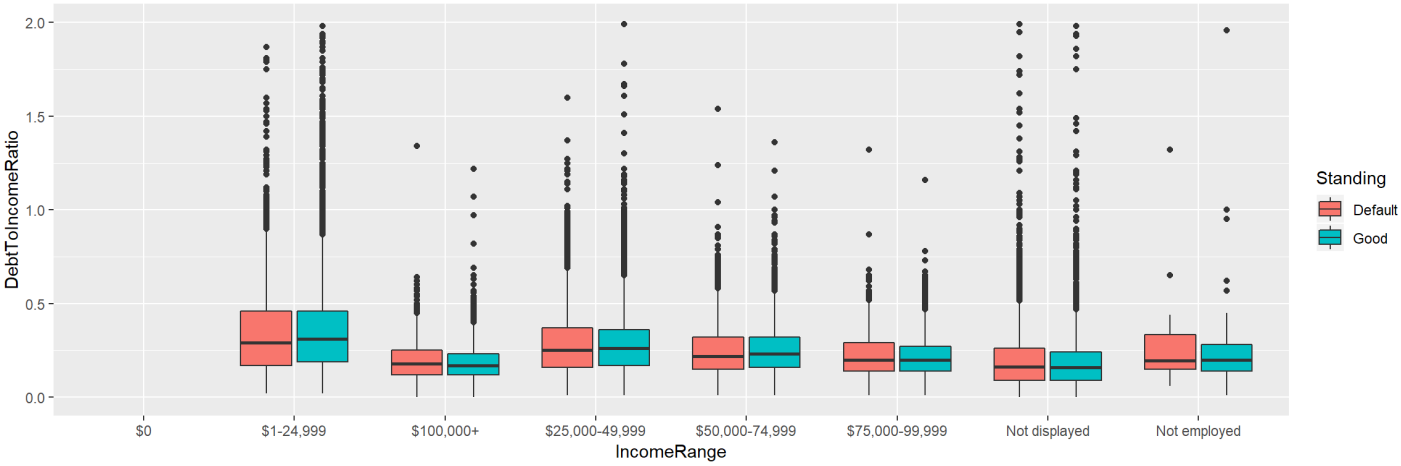
As shown by the correlation there was some relation between these two. So only looking at defaulted borrowers with x axis log transformed and not including bottom 10% LP_customerpayments. So there is some interaction between these two.

StatedMonthlyIncome vs DebtToIncomeRatio by Standing



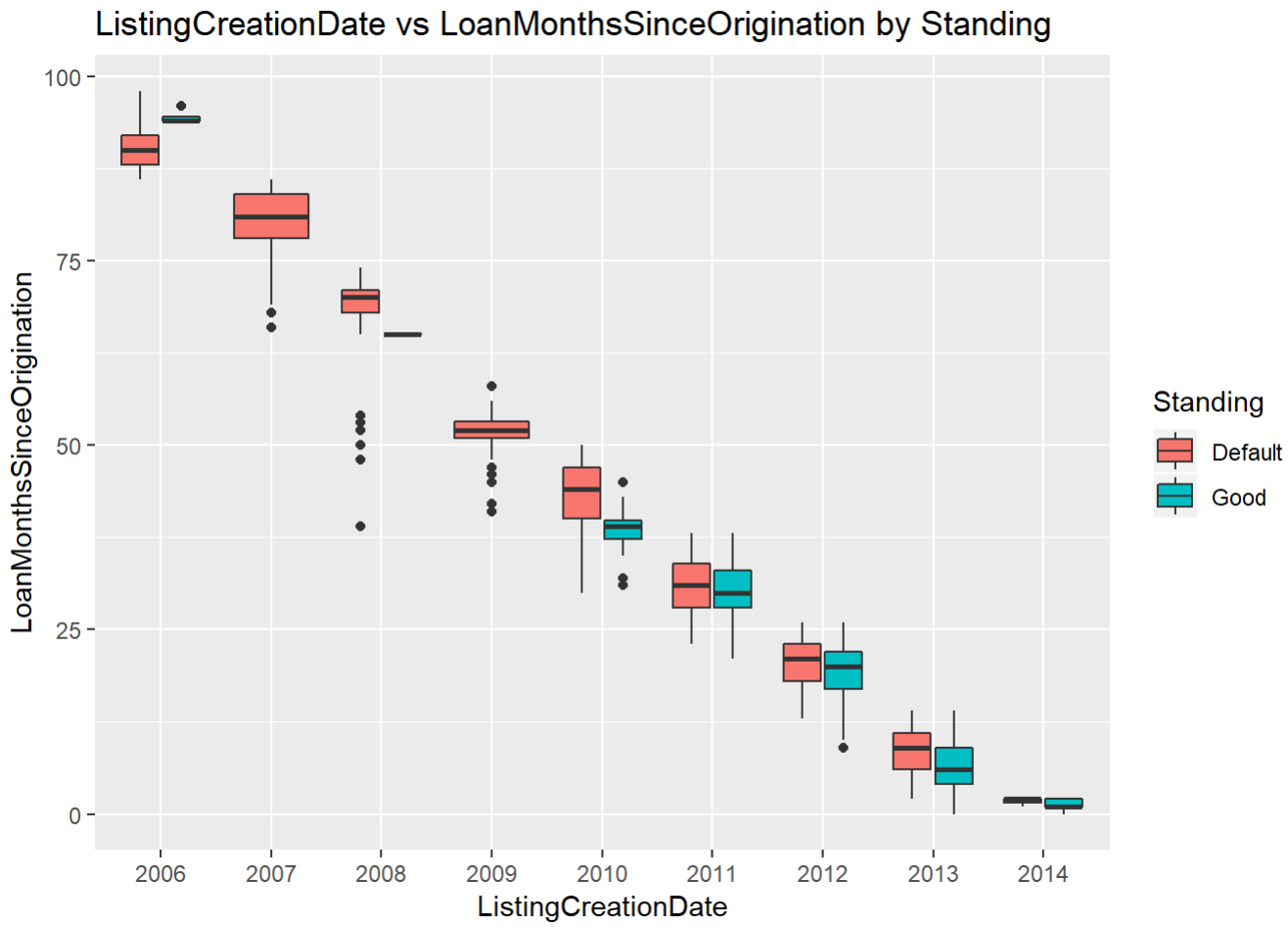
Only looking at the default borrowers. Most of them are within 0 - 10k range. There are also some with income under 2500 who have high debt to income ratios as expected. As seen from correlation coefficient there is inverse relation between these two which can be expected.

StatedMonthlyIncome vs DebtToIncomeRatio by Standing



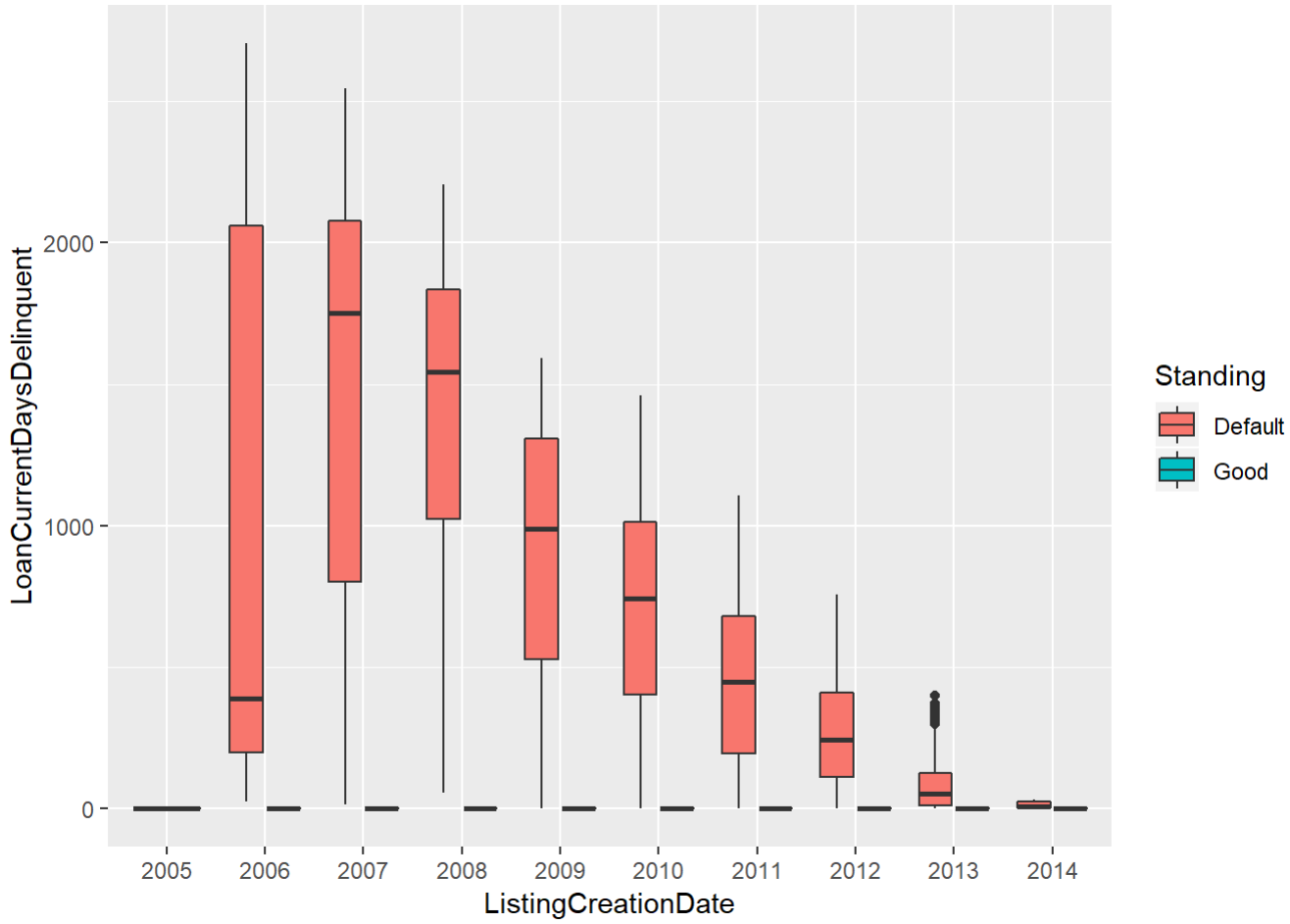
The average Prosper borrower according to this dataset has 27.60% DTI, although the maximum allowable is 50% as per their website the ratio is below 1 for all income range borrowers but I do see outliers in each income range.

ListingCreationDate vs LoanMonthsSinceOrigination by Standing



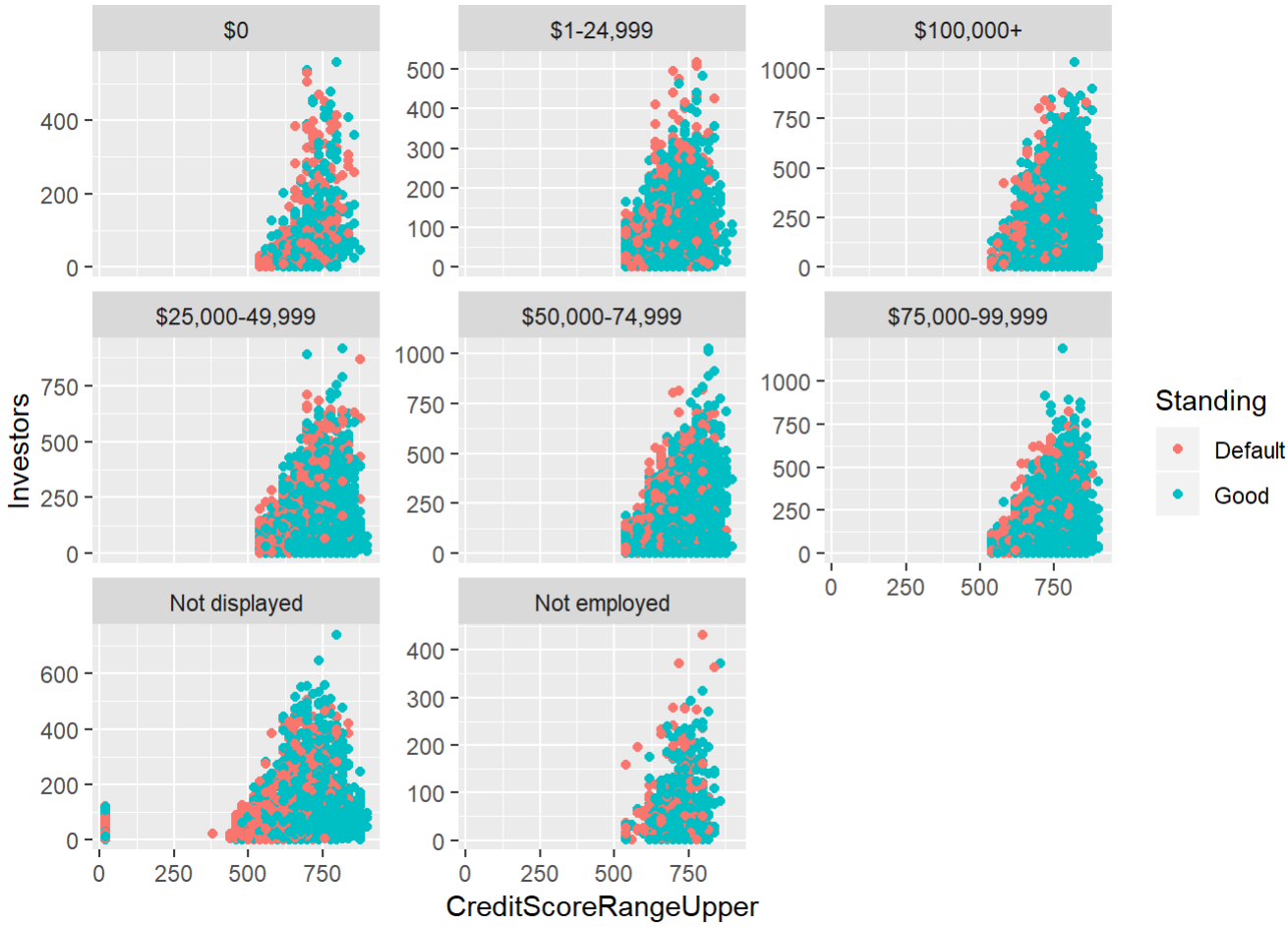
Most borrowers that went delinquent were from 2006 - 2008 (not including those who have completed their terms) that is seen by the increased number of months for these years. The skewness again could also be due to the financial crisis period (2008-09). From 2009 inwards the number of defaulters have almost always conceded more months than those in good standing which could be expected.

ListingCreationDate vs LoanCurrentDaysDelinquent by Standing



Loans created in 2006 - 2008 were delinquent for most of the days.

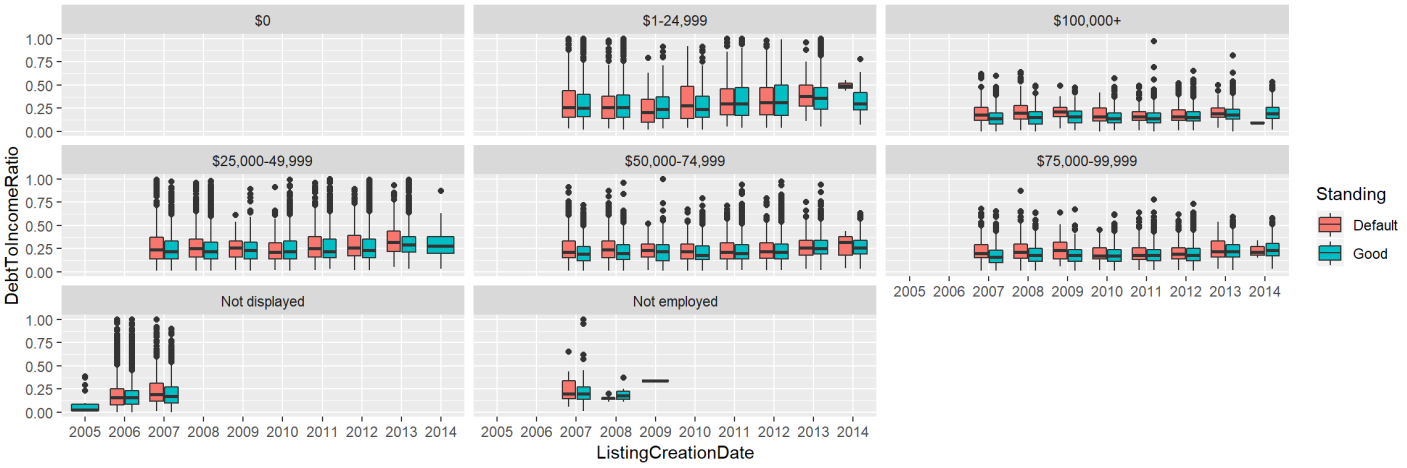
CreditScoreRangeUpper vs Investors by Standing



There is a direct relation between credit score and number of investors with defaulters being at lower credit rating usually as compared to good standing accounts.

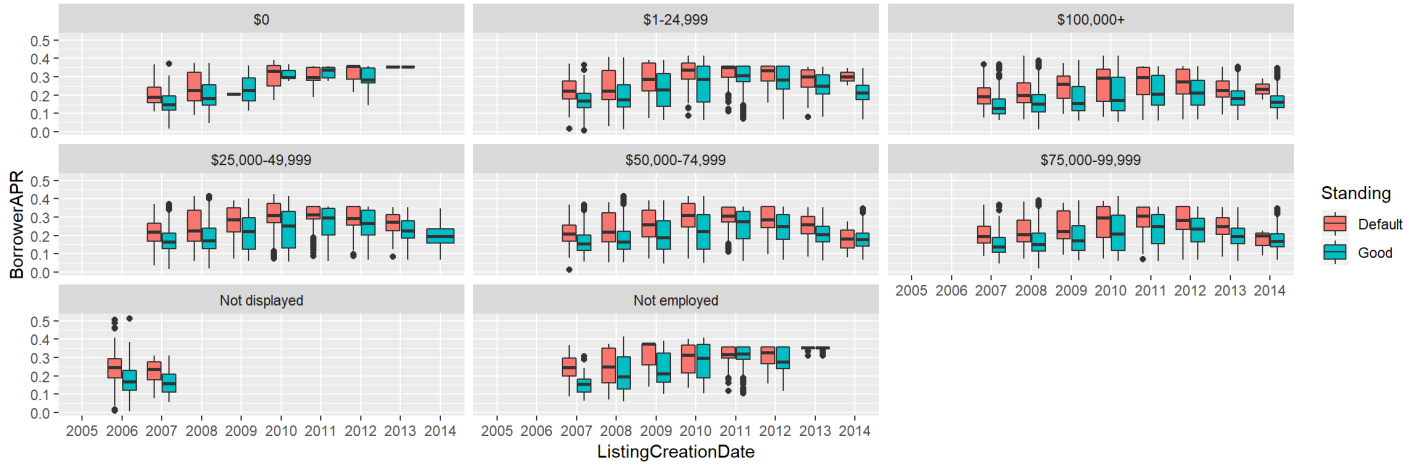
ListingCreationDate vs DebtToIncomeRatio by Standing

```
## Warning: Removed 9353 rows containing non-finite values (stat_boxplot).
```



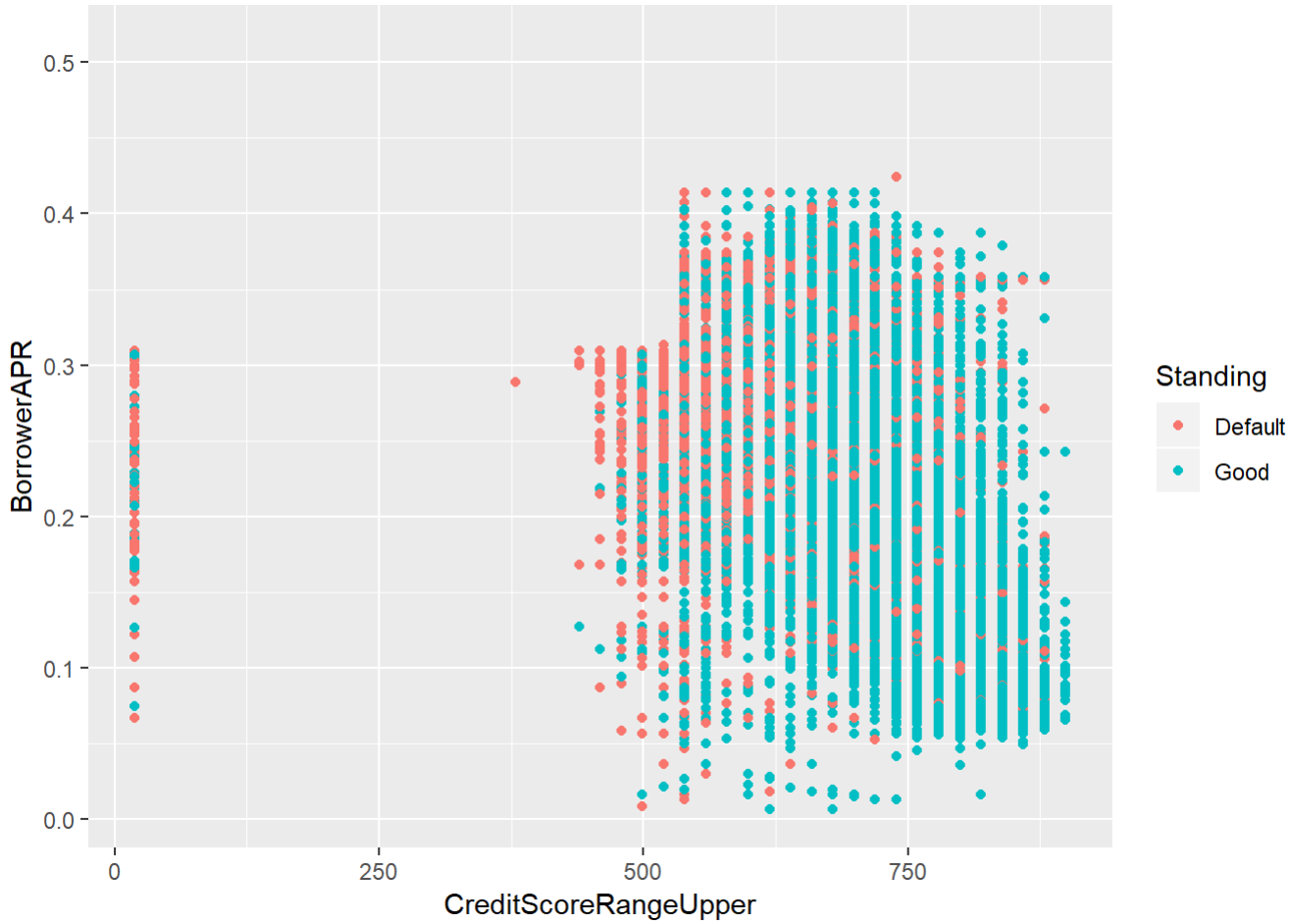
Not displayed category does not exist after 2007. There are a large number of outliers above a ratio of 1 so limited analysis to a ratio of 1 only.

ListingCreationDate vs BorrowerAPR by Standing



Defaulters have more APR almost always as compared to the others. This is not because they have taken high loan amounts as seen previously in the bivariate analysis section the number of defaulters were almost always less than the good standing borrowers for the same amount range.

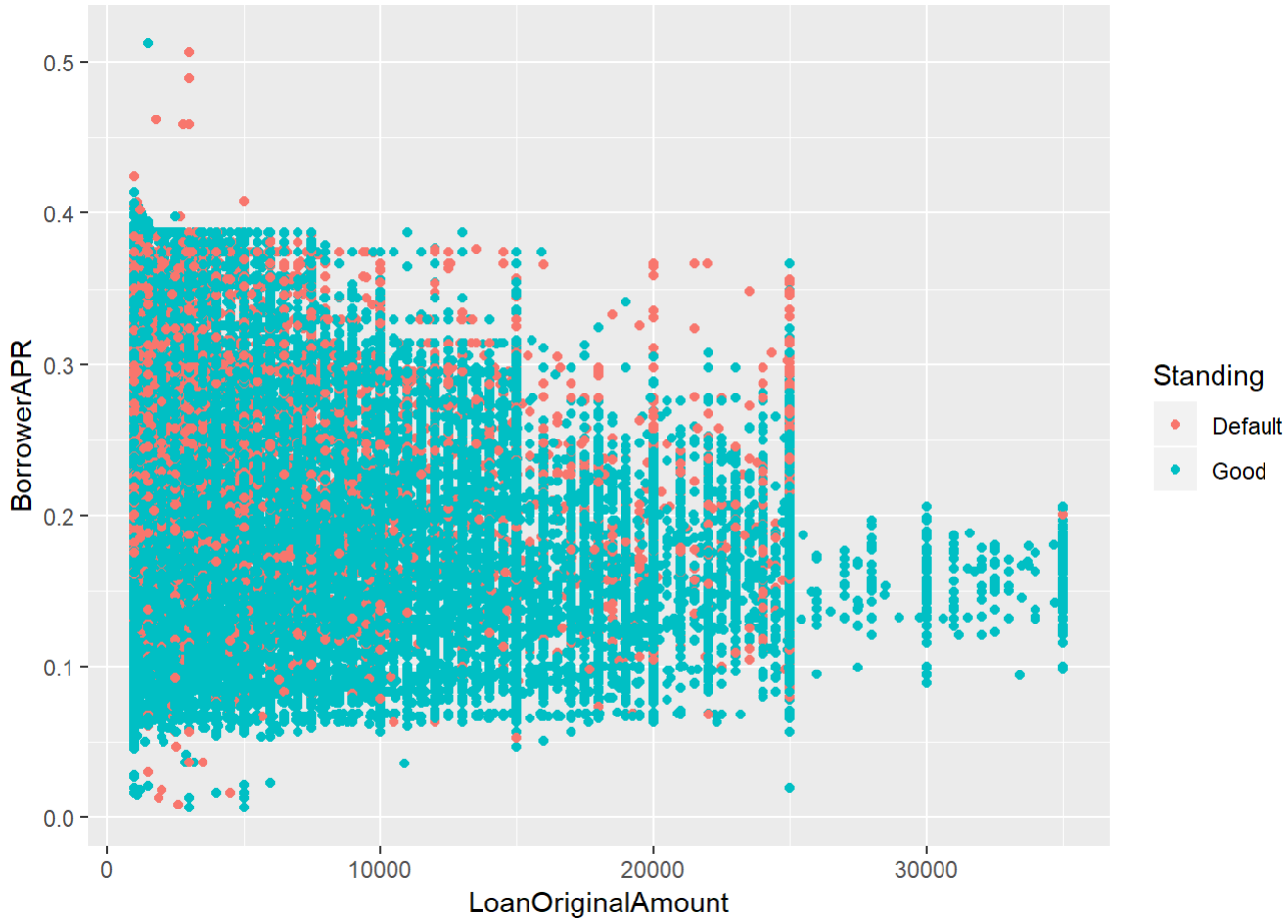
Analyzing borrower APR with credit score to see if this has something do with high rates for bad standing borrowers also APR with loan original amount will be checked.



The rates for defaulters seem to be within 0.3 from 450 to 600 range. They do seem to be on the lower side of the credit score something investors can keep into consideration.

```
by(plds$BorrowerAPR,plds$Standing,summary)
```

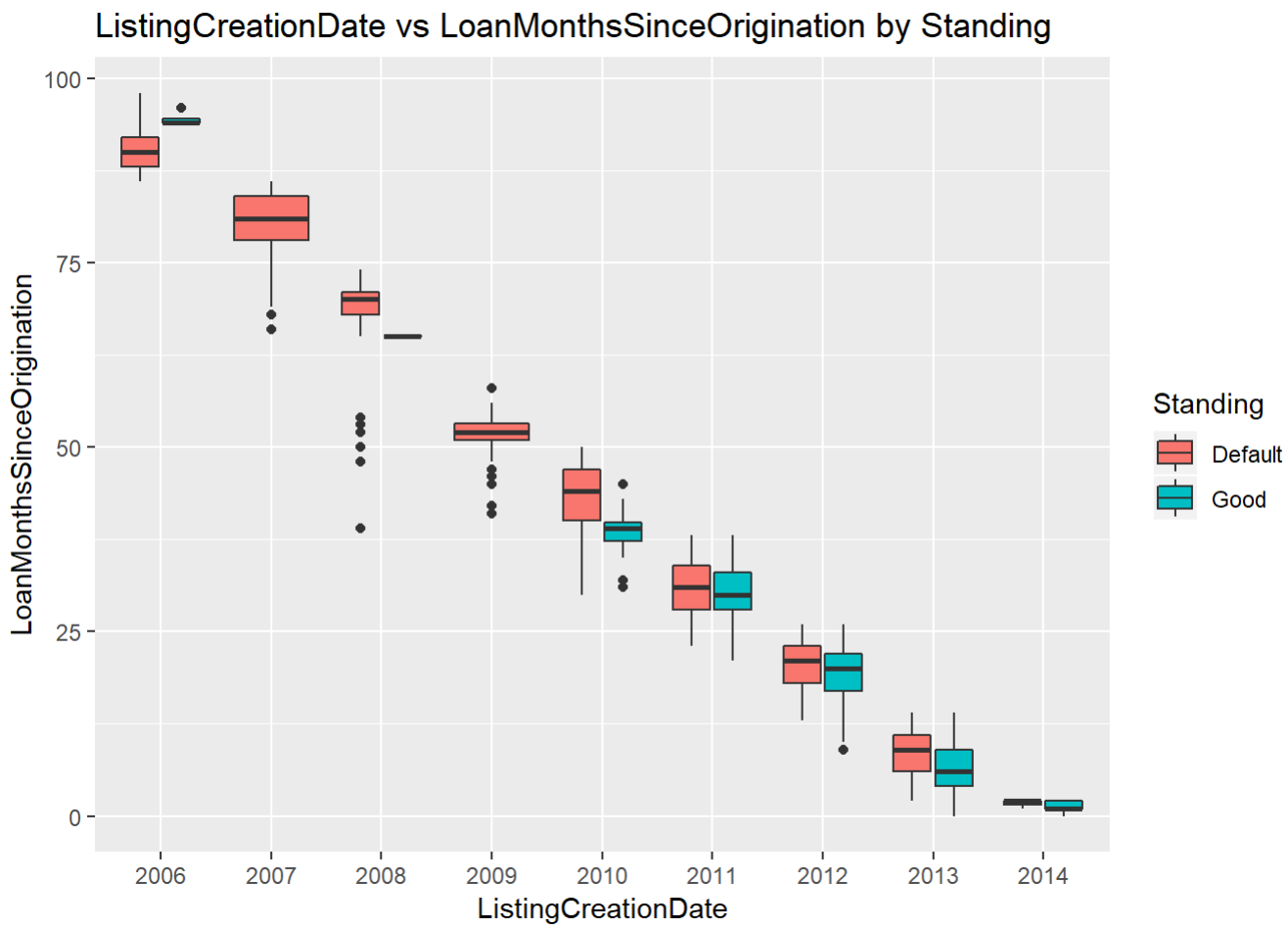
```
## plds$Standing: Default
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00864 0.18977 0.25627 0.25384 0.31033 0.50633
## ------------------------------------------------------------
## plds$Standing: Good
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## 0.00653 0.15016 0.20268 0.21178 0.27246 0.51229      25
```



Final Plots and Summary

Plot One

ListingCreationDate vs LoanMonthsSinceOrigination by Standing

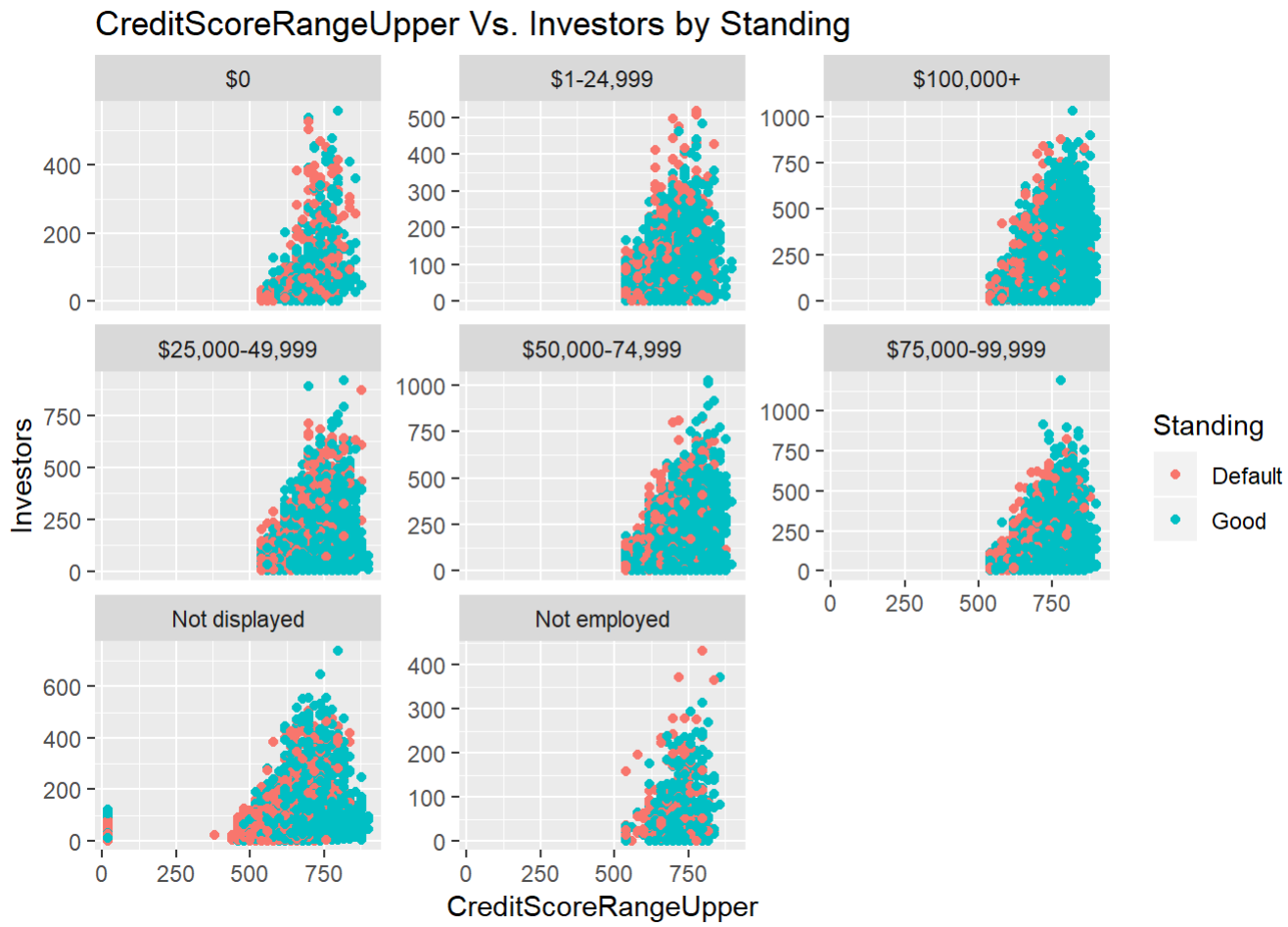


Description One

The plot does not include those borrowers who have completed their terms in order to see the current borrowers only as those who have completed their terms will be causing a bias. This plot shows borrowers with default standing had taken most number of months in almost every year. The skewness seen might be due to the financial crisis from 2008 - 09 but the number of defaulters since then has been on the fall.

Plot Two

CreditScoreRangeUpper vs Investors by Standing

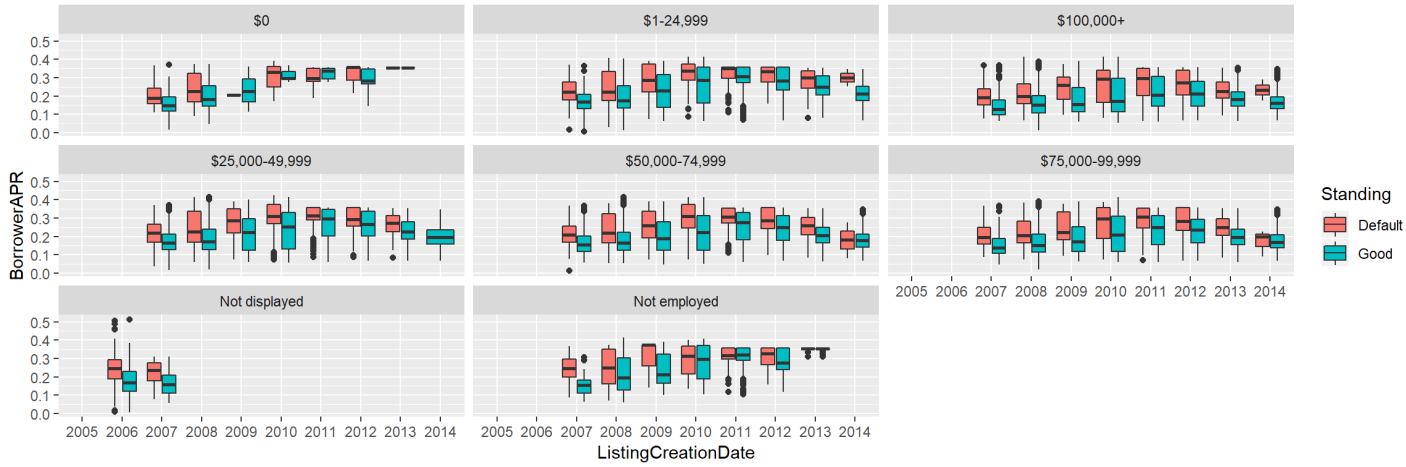


Description Two

This plot shows the trend of investors on selecting borrowers. There is almost always trust shown by investors above 500 score. Also there are a few outliers in Not Displayed maybe due to the fact that this category only existed until 2007 and after this more transparency was introduced and borrowers were made to declare their income.

Plot Three

ListingCreationDate vs BorrowerAPR by Standing



Description Three

Defaulters have more APR almost always as compared to the others. This is not because they have taken high loan amounts as seen previously in the bivariate analysis section the number of defaulters were almost always less than the good standing borrowers for the same amount range. Analyzing borrower APR with credit score also did not show any unexpected trends that could be associated with this.Also, there wasn't any specific relation between loan original amount and APR for the same loan amount. ——

Reflection

The dataset contained around 100,000 observations and 81 variables. Given the original range of keeping the analysis restricted to 10-15 variables was difficult since it was hard to decide which variables to keep. I tried to keep the variables within the prescribed range but some features required more analysis to come to any conclusion.

There were few financial terms that were hard to understand so first some reserach was conducted to get familiar with them.

Comparing the default borrowers with those in good standing revealed a lot of insights about what factors actually contribute towards delinquicy such as BorrowerAPR, Credit Score and inquiries to name a few. Strict policies could further be set up in order to further reduce the number of defaulters.

The dataset was full of interesting information such as the criteria followed by investors to lend to borrowers with high credit scores (500). There was a positive relation between loan original amount and number of investors showing more investors contribute towards more loan amount not causing a burden in case if only was lending.Also, payments made by customers including all service charges showed positive relation with the number of investors.

There is still a lot room for further analysis such as seeing the categories of the purpose loan was taken and average earning per state could be included as a feature.

References

https://prosper.zendesk.com/hc/en-us/articles/210013963 (https://prosper.zendesk.com/hc/en-us/articles/210013963)

https://www.investopedia.com/terms/ (https://www.investopedia.com/terms/)

https://stats.stackexchange.com/questions/11406/boxplot-with-respect-to-two-factors-using-ggplot2-in-r (https://stats.stackexchange.com/questions/11406/boxplot-with-respect-to-two-factors-using-ggplot2-in-r)

https://www.orchardplatform.com/blog/2014519lender-yield-prosper/ (https://www.orchardplatform.com/blog/2014519lender-yield-prosper/)