Exploratory Data Analysis on Loan Data from Prosper

LinhNguyen

Thursday, November 17, 2016

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

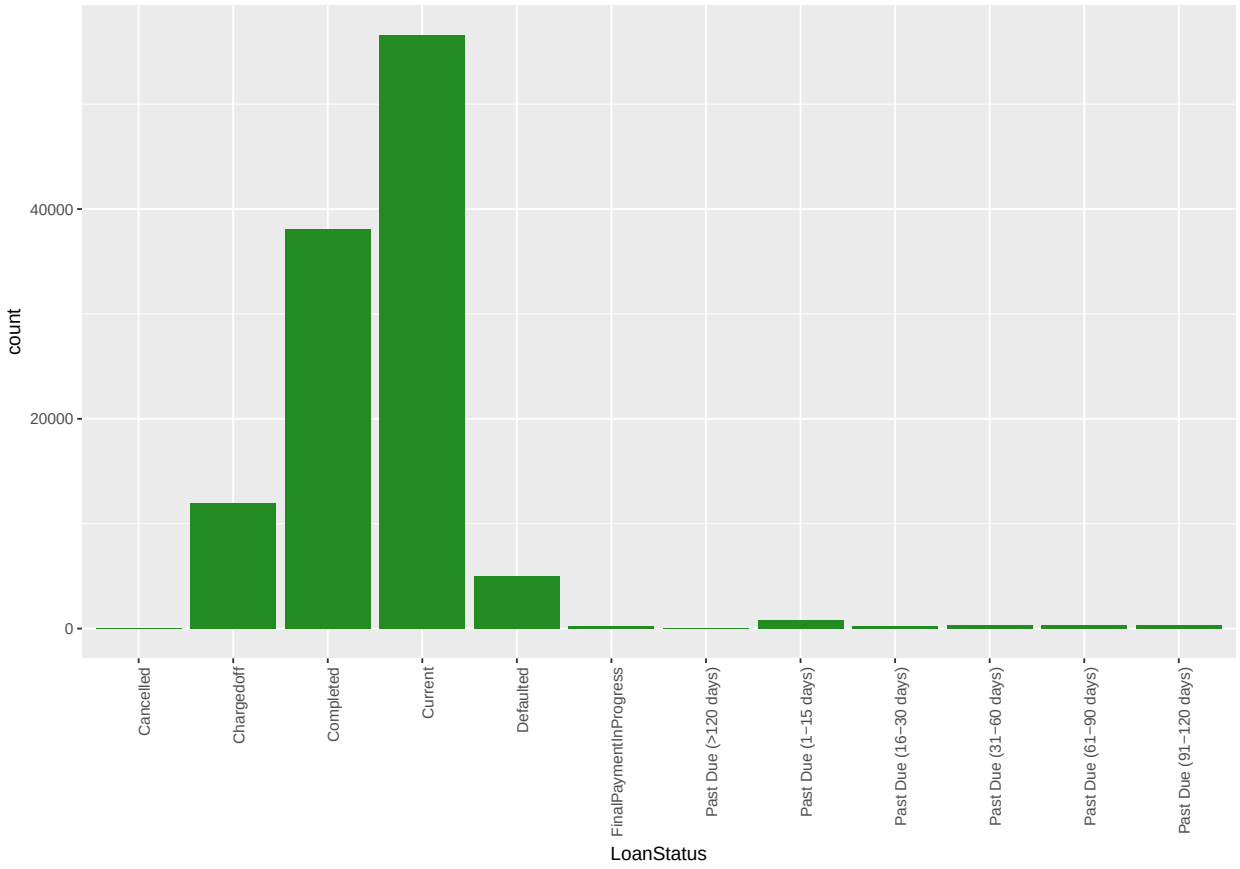
The data set is big, with 81 variables and 113,937 observations.

This analysis has 2 parts:

- Part 1: I poked around, asking random questions, trying to get a sense of the data. Because the data set are too big, and without a proper question serving as some kind of anchor, one could easily get lost in the sea of data.
- Part 2: The main theme of the analysis, where I tried to understand how the risk and prospect affect the investor's amount of investment?

PART I

1. What are the differences in term of profitability between status in LoanStatus?



The graph above shows the distribution of LoanStatus. There is a total of 12 factors, too many for pattern recognization using color.

Therefore, before solving the question, I will first try to group related factors with each other, and put them under the same color palette:

The good scenario:

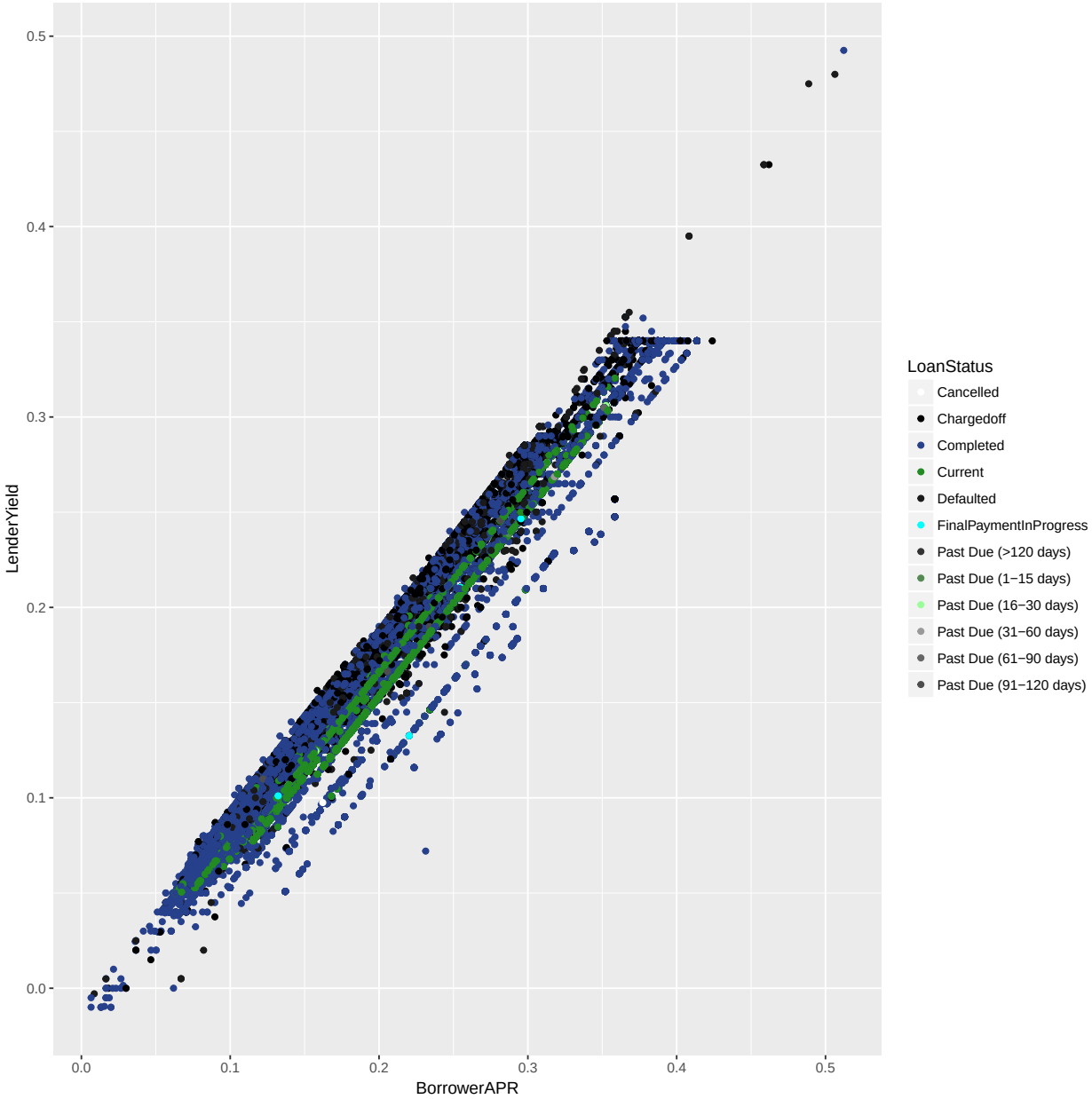
- Completed: Represent the best scenario, where the loan was paid back in full
- FinalPaymentInProgress: The near best scenario, a little more and the loan will be paid back in full

The neutral scenario:

- Current: The loan is being paid, and nothing has gone wrong, yet.
- Past Due (1-30 days): Some small delay, probably nothing.

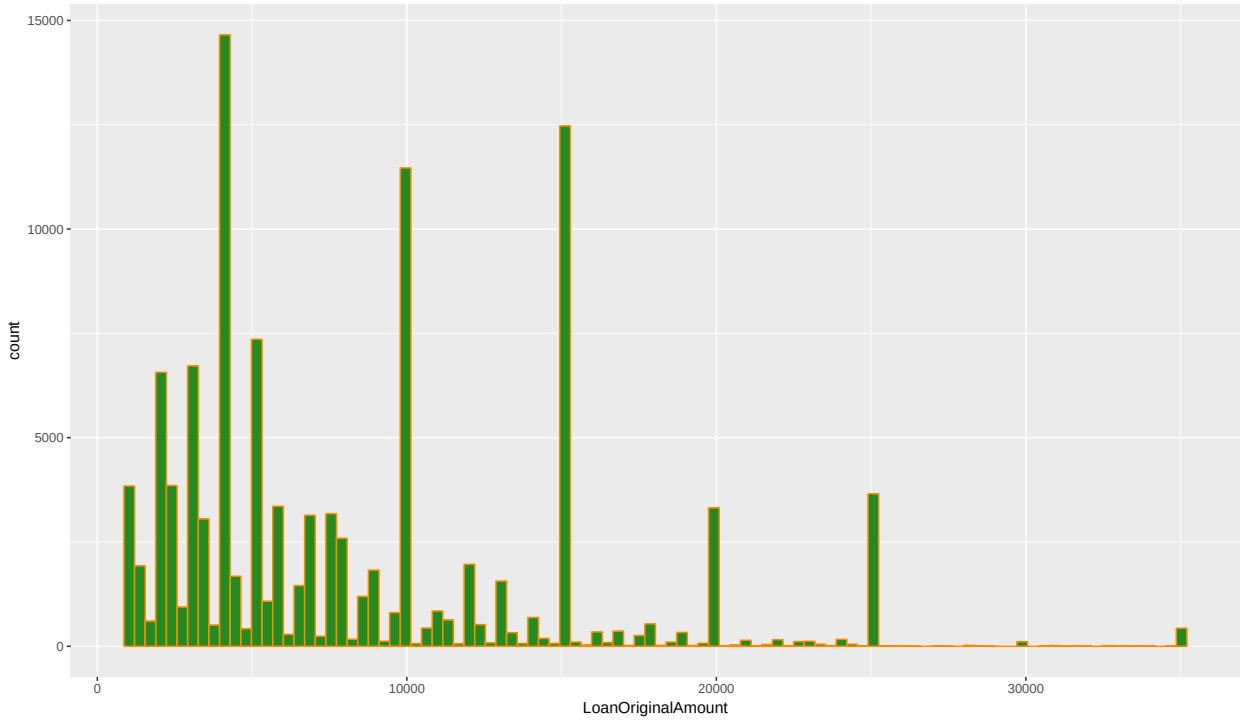
The bad scenario:

- Past Due (30 days onward): The longer, the more likely a charged off will occur.

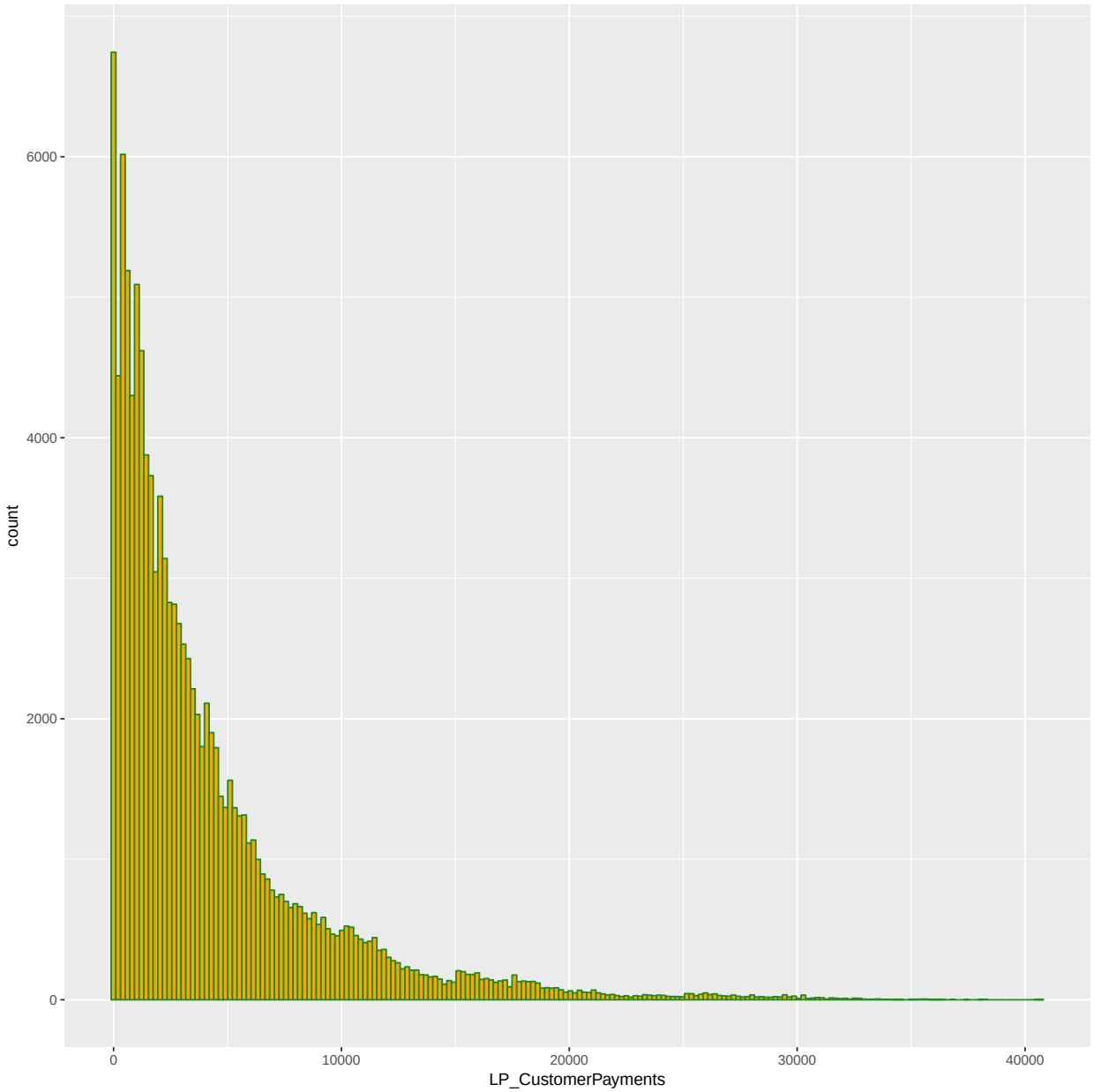


A random scatter plot that to check the effect of the grouping. It's good now, time to start answering the real question.

First I will try to solve this the simplest way possible by looking at the differences between the money given out (LoanOriginalAmount) and the money that the borrower paid back (LP_CustomerPayments).



People love rounded numbers, notice how the majority of loan amount are rounded to the nearest thousands. There are a lot of big loans (10,000 and above).

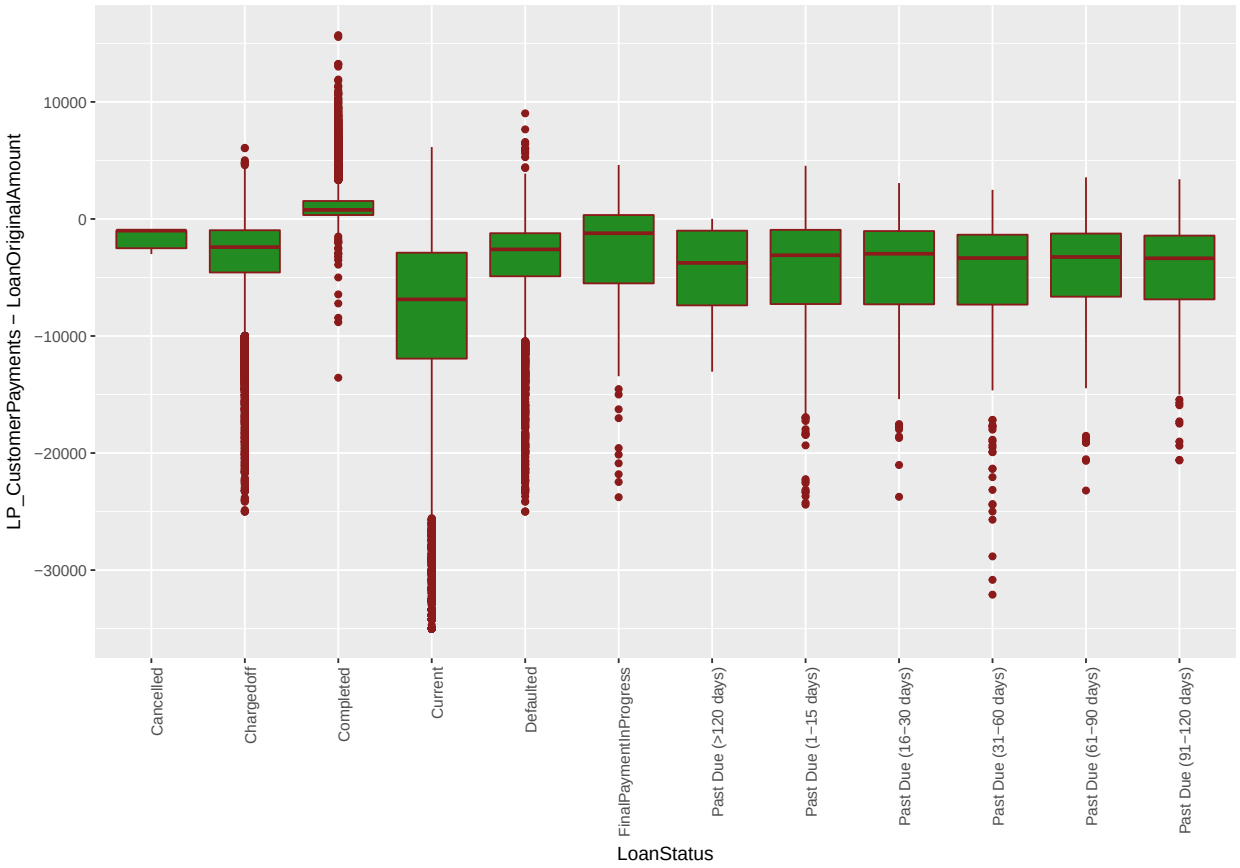


The majority of borrower's payments are below 10,000, that's a bad news, considering the number of 10,000 and above original loan amount are not that far away from the below 10,000 one.

Perhaps something is missing, I will come back later and dig this up.

```
summary(loan.data$LoanStatus)
```

```
##              Cancelled            Chargedoff              Completed
##                      5                11992                  38074
##                Current             Defaulted FinalPaymentInProgress
##                  56576                  5018                    205
##   Past Due (>120 days)   Past Due (1-15 days)  Past Due (16-30 days)
##                     16                    806                    265
##  Past Due (31-60 days)  Past Due (61-90 days) Past Due (91-120 days)
##                    363                    313                    304
```



The lenders will get a green cashflow for a completed loan, and red for everything else.

2. Is there a differences in the borrower's payments between small loans and big loans? Why?

Initially, I noticed a different in the borrower's payments between the 2 group, above 10000 and below 10000. Now I'm going to dig deeper, see if there's anything strange in here.

I added 2 more variables:

- Unpaid balance (total principal payments - original loan amount).

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -35000   -9036   -3000   -5231       0    1424
```

- Categorized original loan amount, I wanted to see if and how the borrower's payments change with differences loan size. Everything else (interest, fees, etc) was discounted for the sake of simplicity.

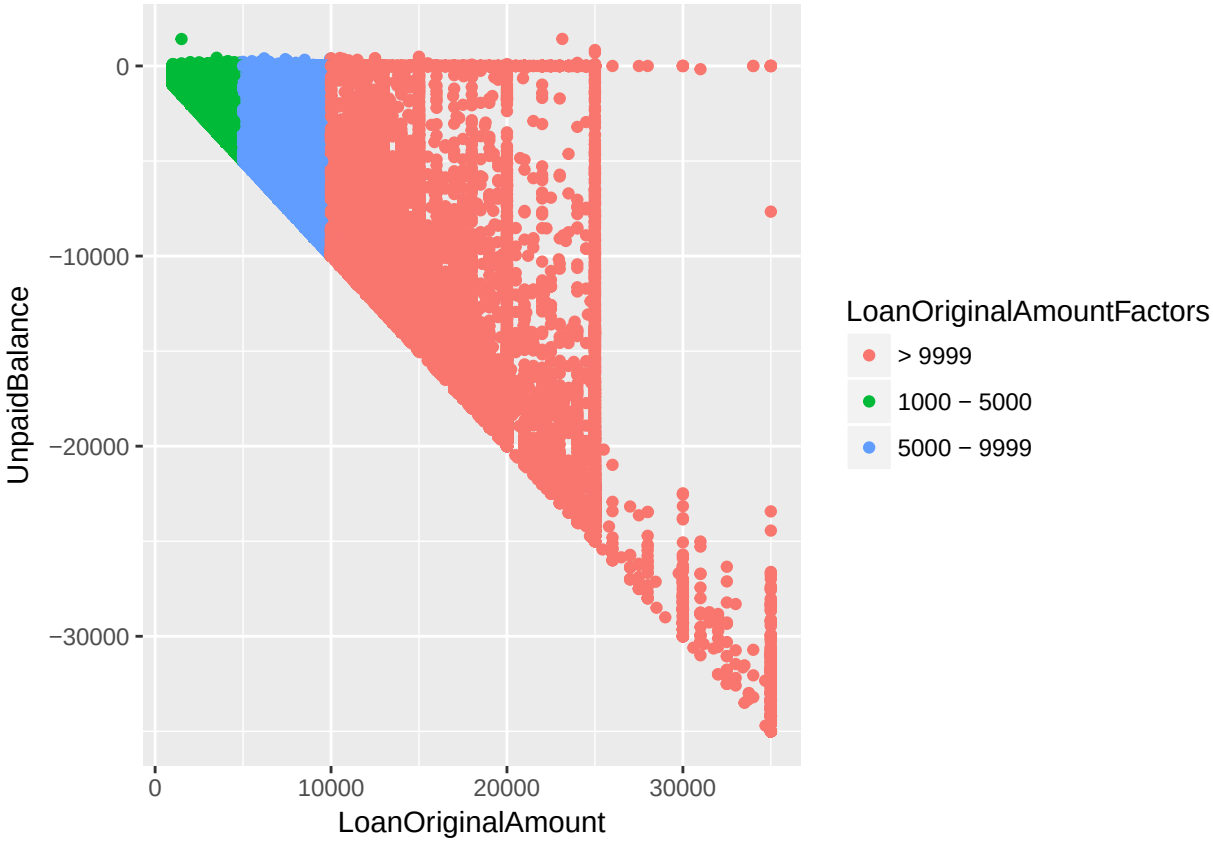
```
##      > 9999 1000 - 5000 5000 - 9999
##       42083       44834       27020
```

```
## Original Loan Amount: 1000 - 5000
## Number of loans: 44834.000000
## Average unpaid balance: -1399.991262
## Average Payments over Average Original Loan Amount: 0.533
```

```
## Original Loan Amount: 5000 - 9999
## Number of loans: 27020.000000
## Average unpaid balance: -3089.830642
## Average Payments over Average Original Loan Amount: 0.540
```

```
## Original Loan Amount: > 9999
## Number of loans: 42083.000000
## Average unpaid balance: -10688.505872
## Average Payments over Average Original Loan Amount: 0.290
```

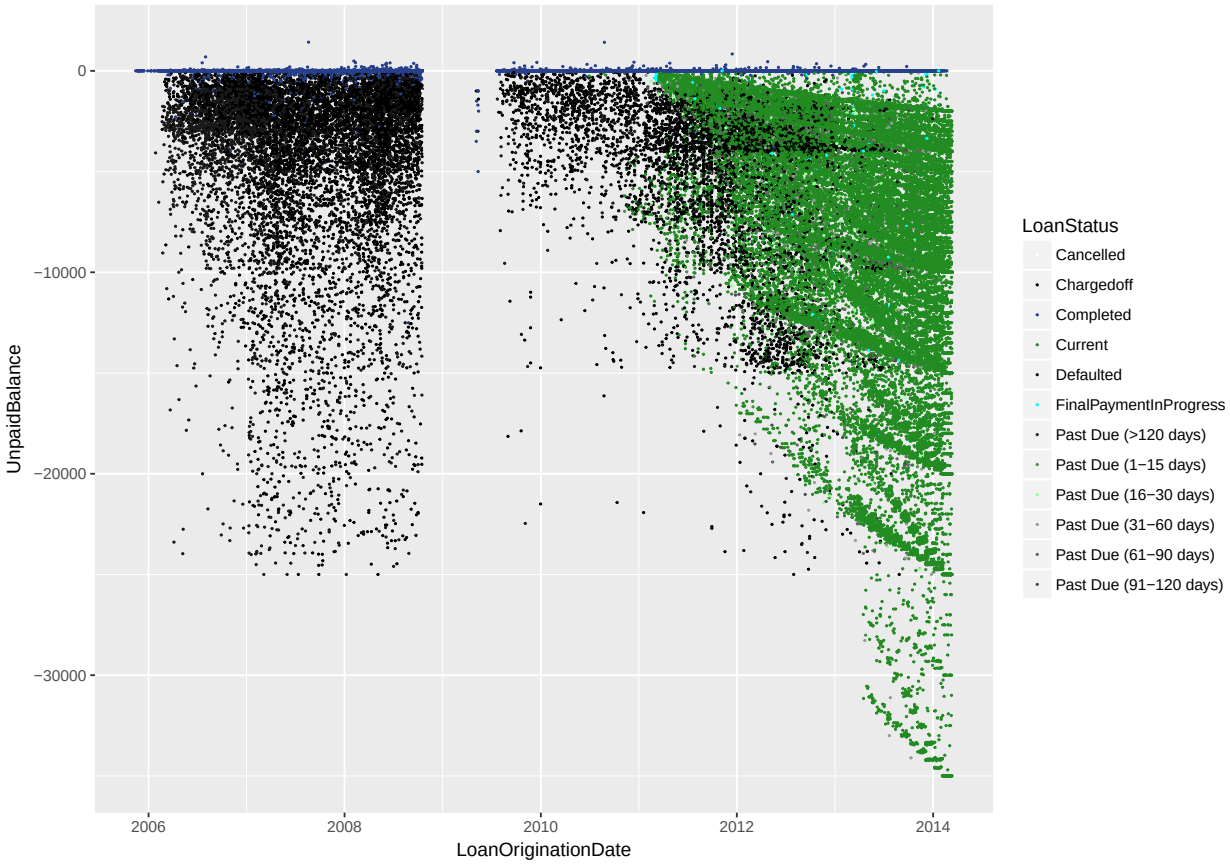
The average unpaid balance dropped dramatically when the loans get bigger, from -1399 to over -10000, the same as the percentage of average payments over average original loan amount, from 0.533 down to 0.29



The larger the loan, the more difficult to pay back, that's a fact.

But this is totally unexpected, for this graph show a massive loss, and it told us that not only should we stop going for large loans, but we should quit the industry, entirely.

It could be because these loans are incompleted. Now, to avoid bias, I will separate the completed loan, and see if the situation is really this terrible.



There are a lot of current loans, and perhaps it is these loans that distorted the picture. Before 2011, most of the loans are either completed, charged off or defaulted, making it an ideal timeframe for examination.

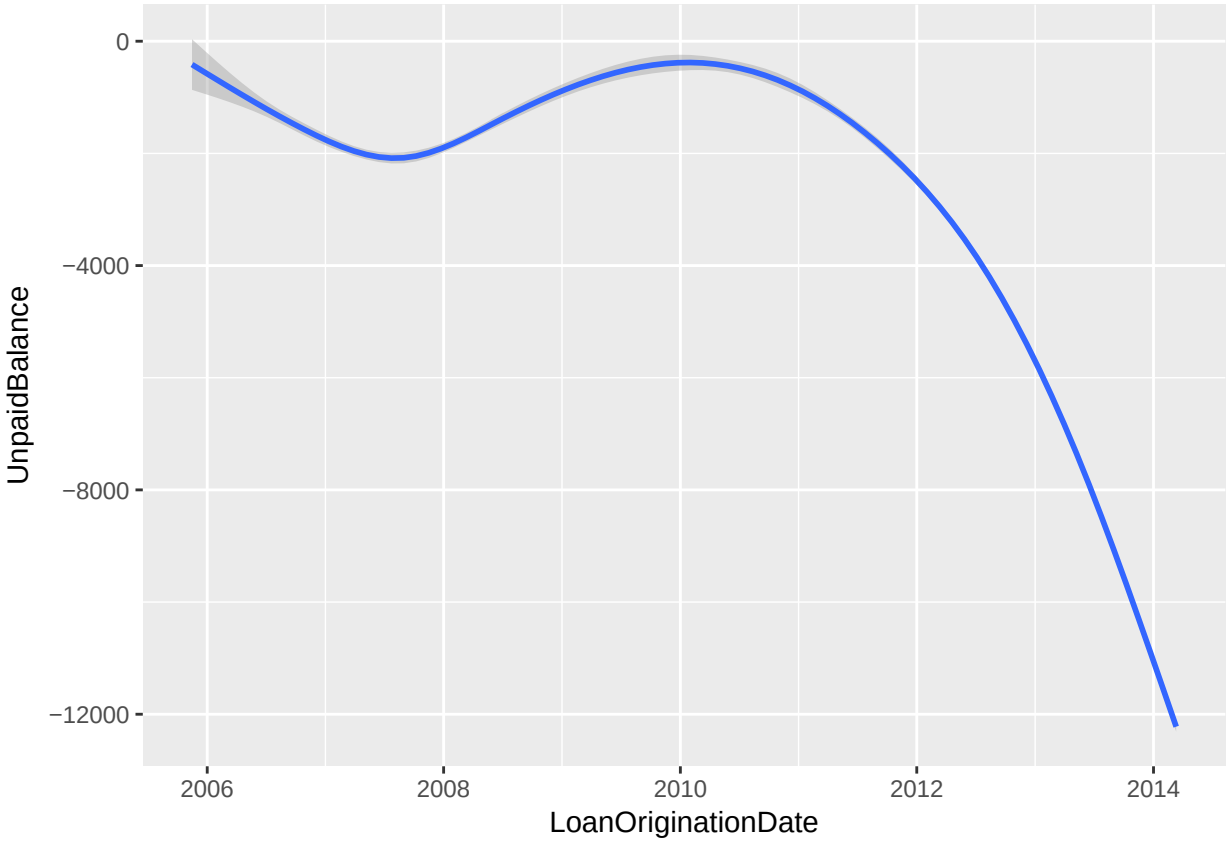
There is also a gap between 2008 and 2010, I wonder why, perhaps there was something wrong with the data, or perhaps it was because of some real world event that affected the entire financial world, which is not quite impossible. I will probably go back and see if I could answer this question later on.

```
## Original Loan Amount: 1000 - 5000
## Number of loans: 19833.000000
## Average unpaid balance: -567.704980
## Average Payments over Average Original Loan Amount: 0.770
```

```
## Original Loan Amount: 5000 - 9999
## Number of loans: 10485.000000
## Average unpaid balance: -1369.467763
## Average Payments over Average Original Loan Amount: 0.789
```

```
## Original Loan Amount: > 9999
## Number of loans: 6321.000000
## Average unpaid balance: -4143.092534
## Average Payments over Average Original Loan Amount: 0.731
```

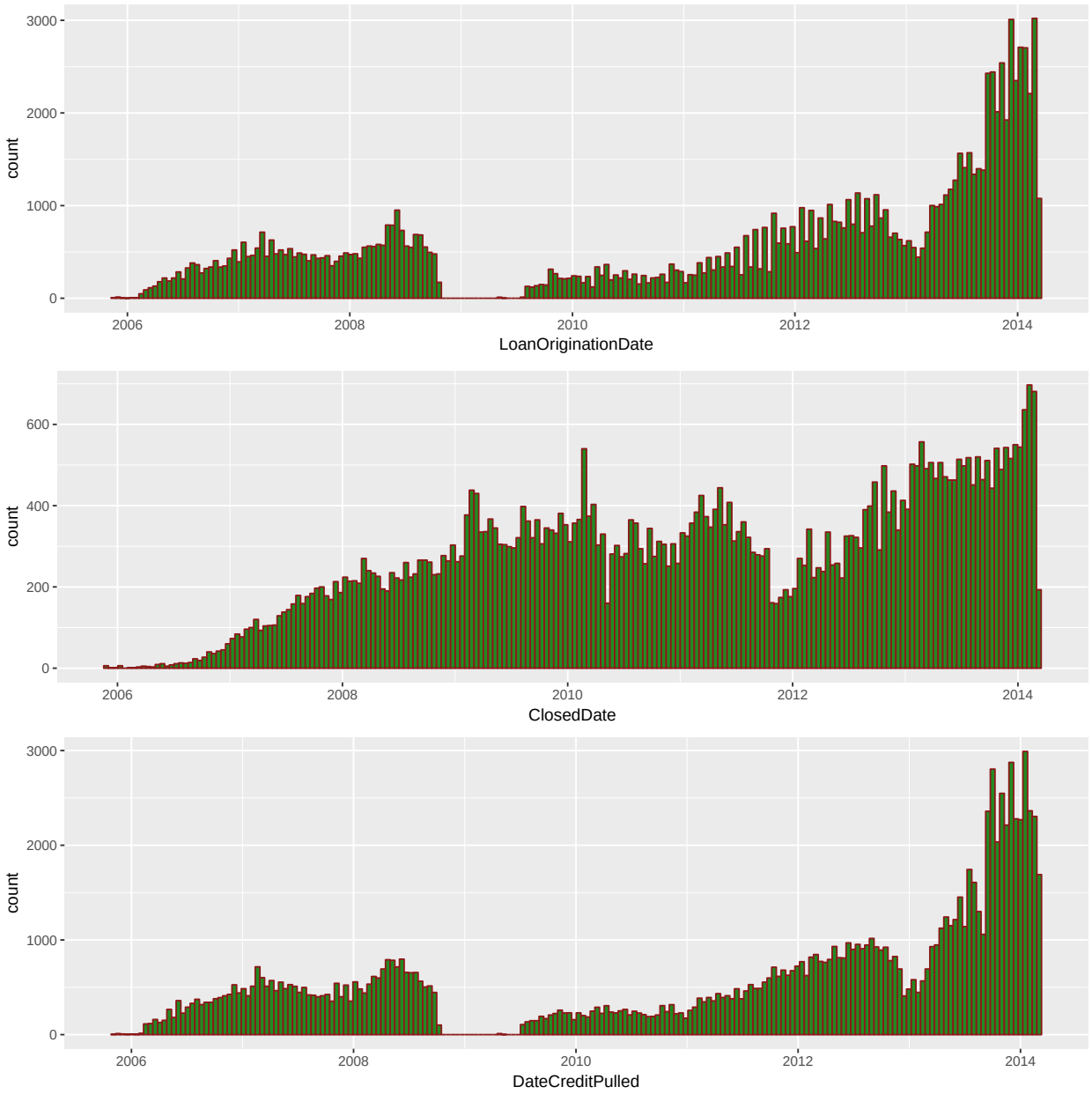
The percentage of average payments over average original loan amount are at the same level for all 3 groups, while the average unpaid balances still get bigger in parallel with the size of loans, but with a much lower rate.



It looks quite promising when interest and fees are taken into consideration. It took a while for a loan to mature, that's why the payments balance between the period of 2011 and 2014 are so low.

The loans didn't do that great in the period between 2006 and 2008, this, again, I suspected that it has to do with the gap between 2008 and 2010.

3. Why is there a gap in data between 2008 and 2010?



There were no new loans in the gap between 2008 and 2010, not a single one, while other already funded loans were still functional as is.

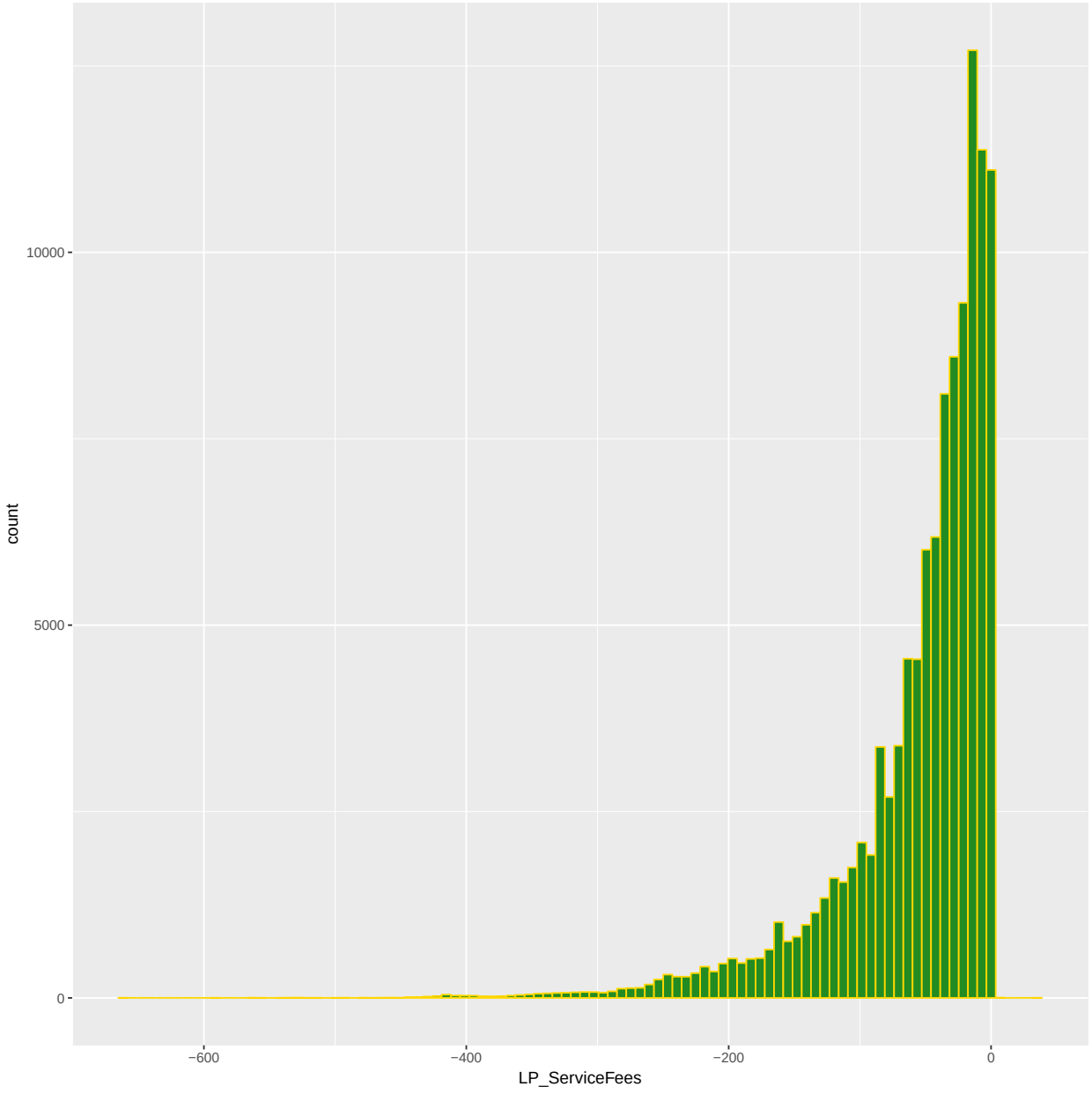
Sadly, there's a limit to how useful data can be, and in this case, it stopped here. There's no longer any kind of data that could explain what happened in that period.

So I did a quick google search, as it turned out, it was the result of a law suit, in which SEC temporarily forced Prosper.com to stop brokering new loans while it determined whether Prosper's loans should be classified as securities.

4. How does Prosper benefit?

Before, I thought that this data set was from some kind of bank, turned out, I was wrong. This data set originally came from Prosper, which is a platform for peer to peer lending service.

Prosper doesn't give people money. It only runs the platform, and charge people for using it, both the lenders and the borrowers.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -664.90  -73.18  -34.44  -54.73  -13.92   32.06
```

On average, Prosper charges around $50 for the service. Notice that this service fee is divided proportionally upon investors of a particular loan.

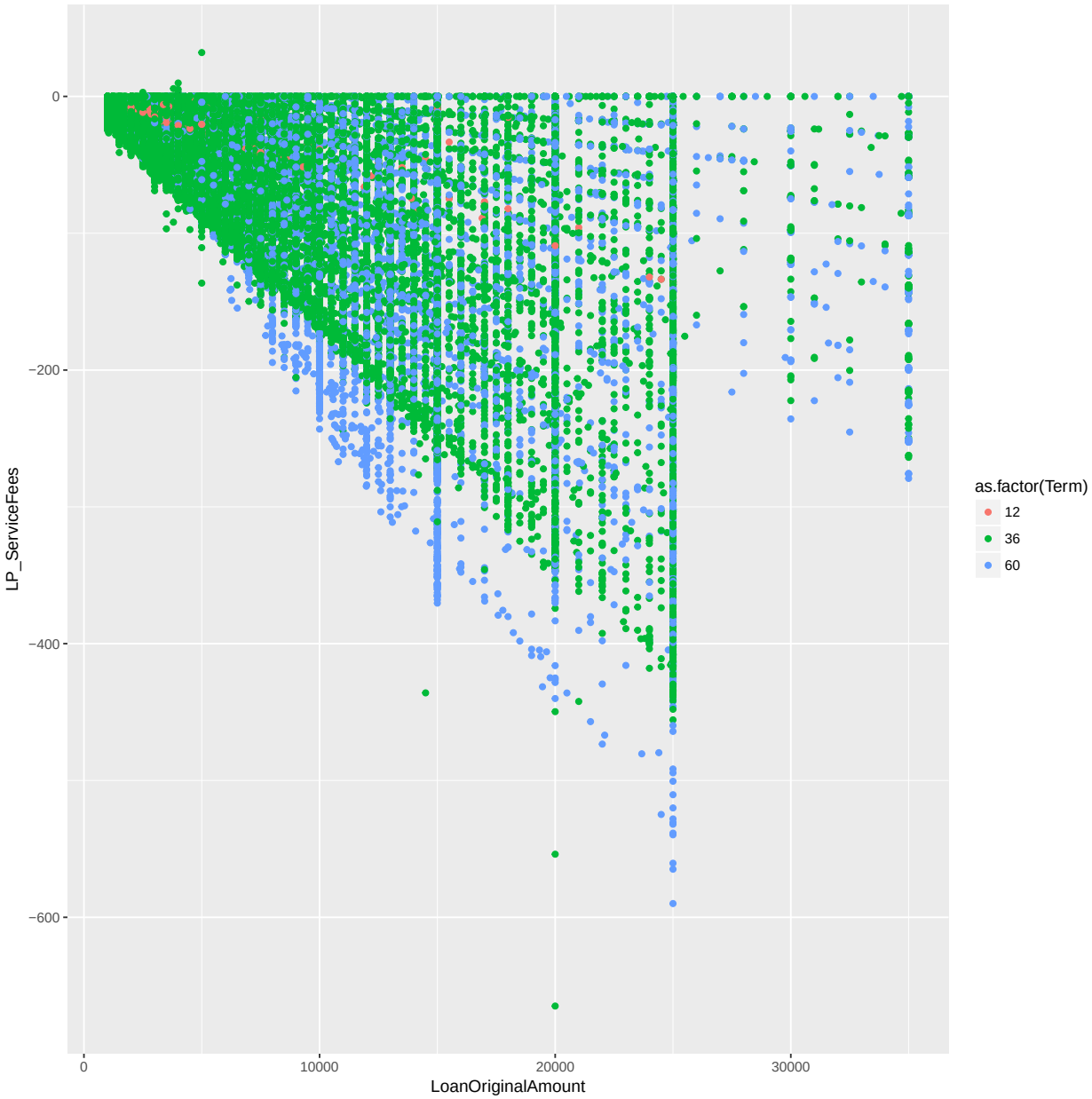
The service fees were meant to be subtracted from every borrower loan payment received, perhaps that's why they put it in the form of negative numbers, so it's kind of strange to spot some positive numbers here,

perhaps something is wrong.

```
##        LP_ServiceFees
## 3393             5.07
## 8378             9.81
## 40835            5.70
## 47933           32.06
## 67544            0.77
## 68401            2.87
## 102726           3.01
## 102736           0.05
```

A total of 8 occurrences with a positive service fees. On one hand, this might just be an engineering bug, otherwise, it meant that special circumstances happened and that Prosper was the one paying the lenders in these 8 instants.

I'd love to go to the end of this, but to do that, I'd need detailed data of transactions, which apparently, doesn't exist in this data set.

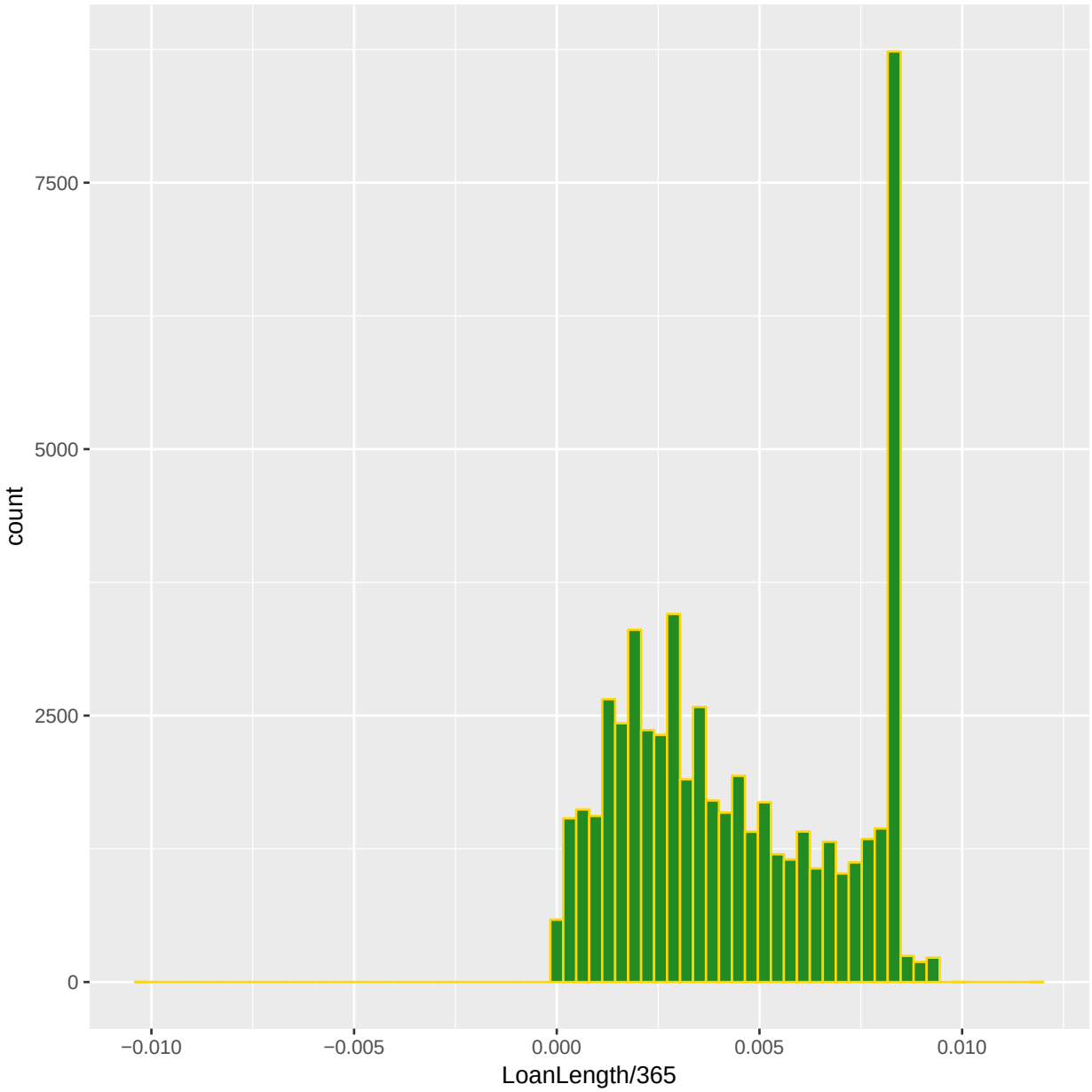


The annual loan service fee for lenders is 1%, so the bigger the loan, the larger the fee.

Another contributor is the length of the loan.

In theory, loans can either have a 12, 36 or 60 months term and that each month the borrower will make a certain fixed amount of payment until the loan is matured.

In reality, however, the loan length is not fixed and borrowers could always pay back earlier if they could. That's why I will also look at the length of the loan. Ideally, the data would be the number of payments that the borrowers made, but that kind of data wasn't available, so I will make do with the differences in days between the closed date and the listing date. It won't be as accurate as the number of payments, but it will be something.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   -0.01    0.00    0.00    0.00    0.01    0.01   58848
```

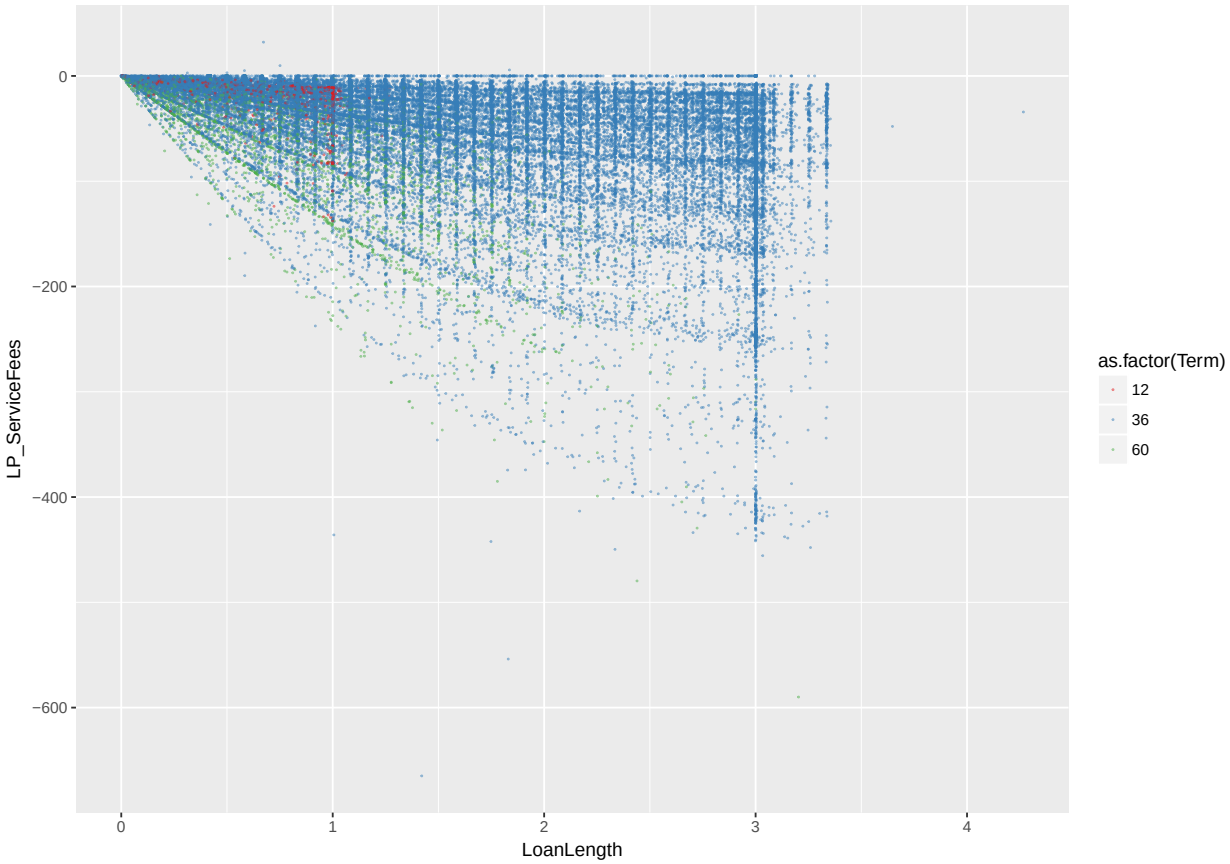
Only the completed, defaulted or chargedoff loans would get a closed date. For loans that are still being paid, the closed date is blank, therefore will produce a missing value, which is neat.

Most of the borrowers pay back in 3 years time.

There is, however, a negative value.

```
## 'data.frame':    2 obs. of  2 variables:
##  $ ClosedDate         : Date, format: "2010-03-16" "2008-10-16"
##  $ ListingCreationDate: Date, format: "2013-12-27" "2008-10-07"
```

I'm pretty sure that this is an engineering problem, perhaps the date was reversed. Again, like earlier, the transaction data wasn't available, so there's no way to get to the bottom of this.

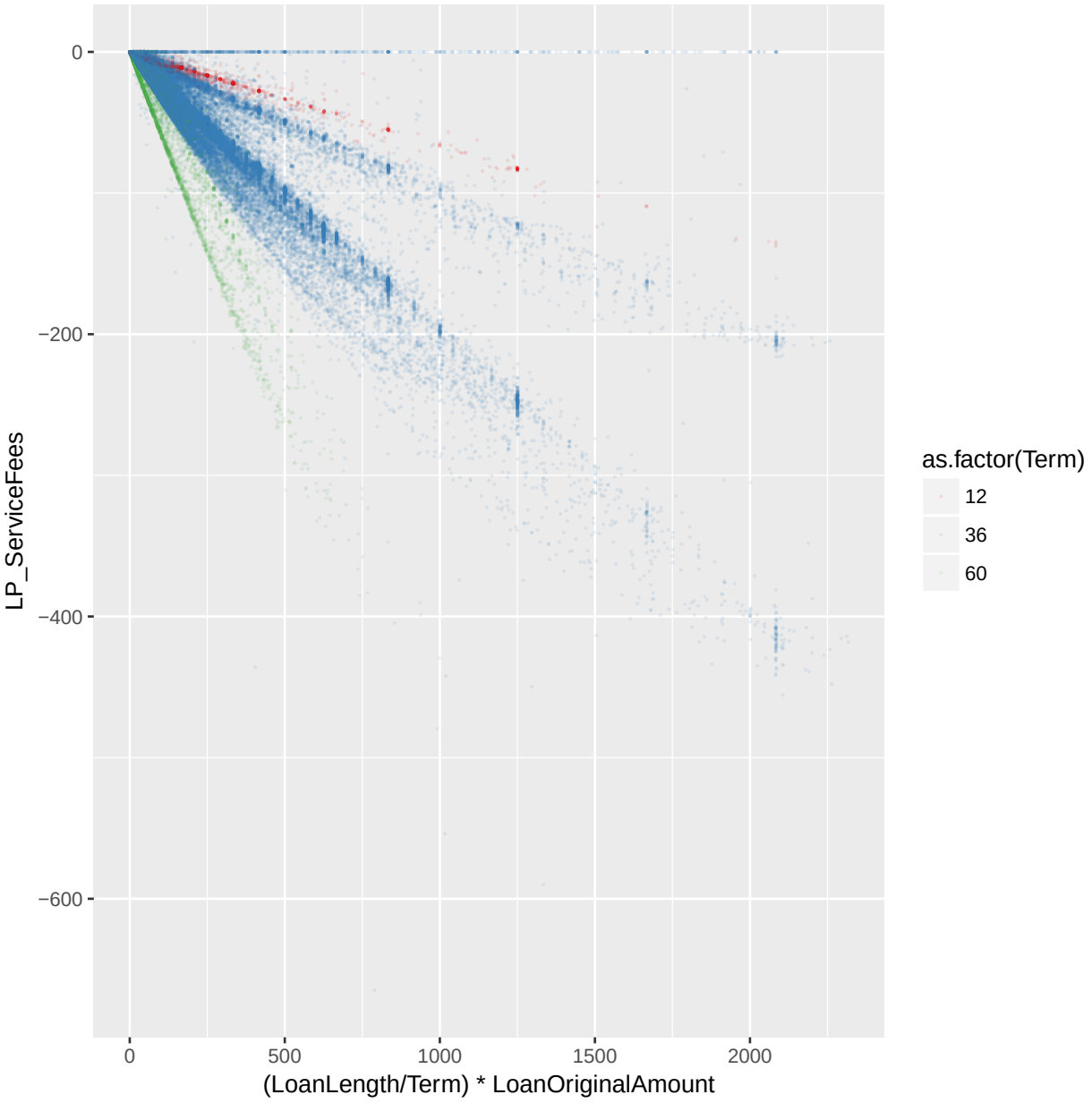


It's true, the longer the length, the bigger the service fee. There are different patterns for loans with different terms, tho not very clear.

Now I will combine them together, the loan amount and the loan length, to get a bigger picture.

Different loan terms will have different loan length, therefore i will divide loan length to term, to find the ratio of real loan length to scheduled loan length.

And then, because both the loan length and the loan amount are numeric data, I will simply multiply it with each other and observe it there's anything to it.



The longer the term, the harder the borrower tried to pay it back early, for the longer they dragged out, the bigger the interest. This graph, however, didn't have the space for interest rate, so later.

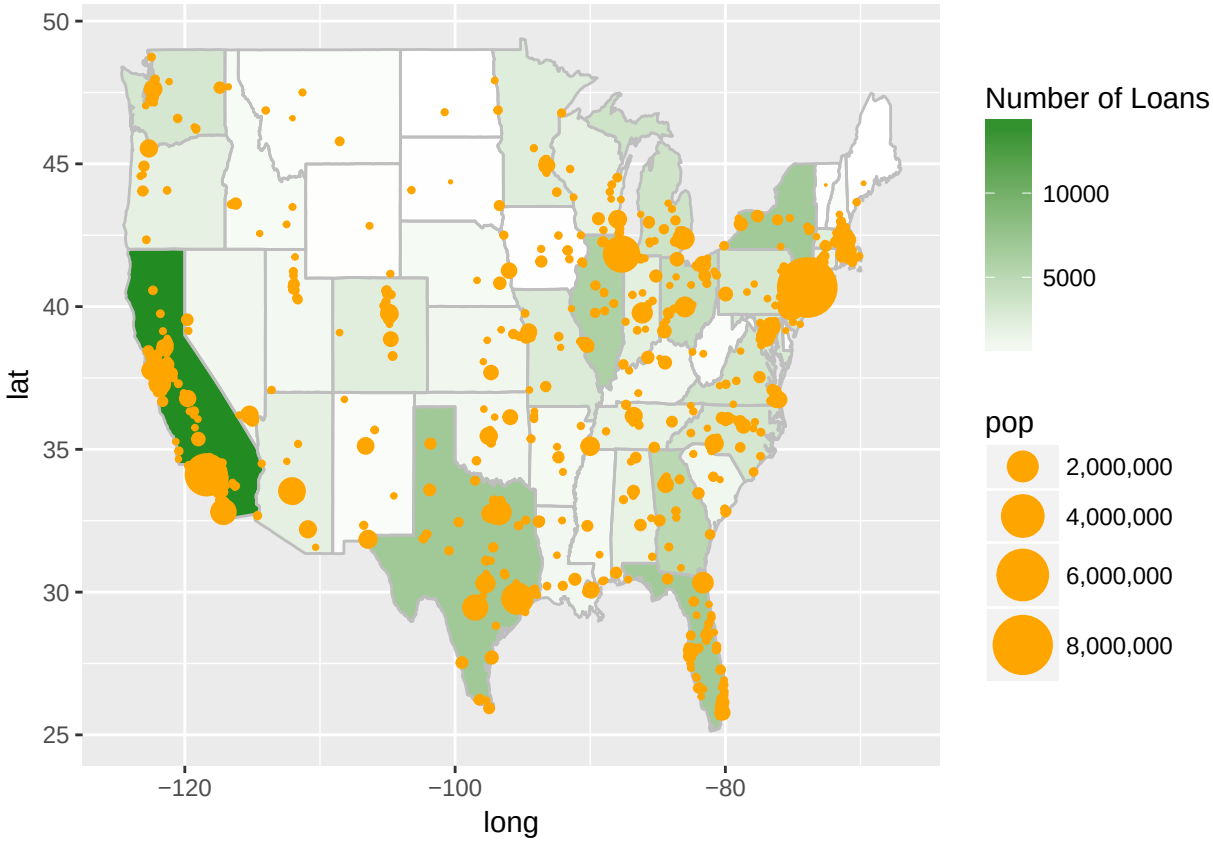
Another part of the service fee is through Borrower's fee.

Prosper can charge borrower with 2 type of fees:

- Closing fee: A percentage is taken directly in the beginning based on prosper rating, the fee associated with rating changed with time and the data set didn't come with this kind of data. While it was possible to cook something up to follow it, it could be a pain.
- Failed payment fee: one time per payment record, no data available.

So given the circumstances, I will not follow through on this one.

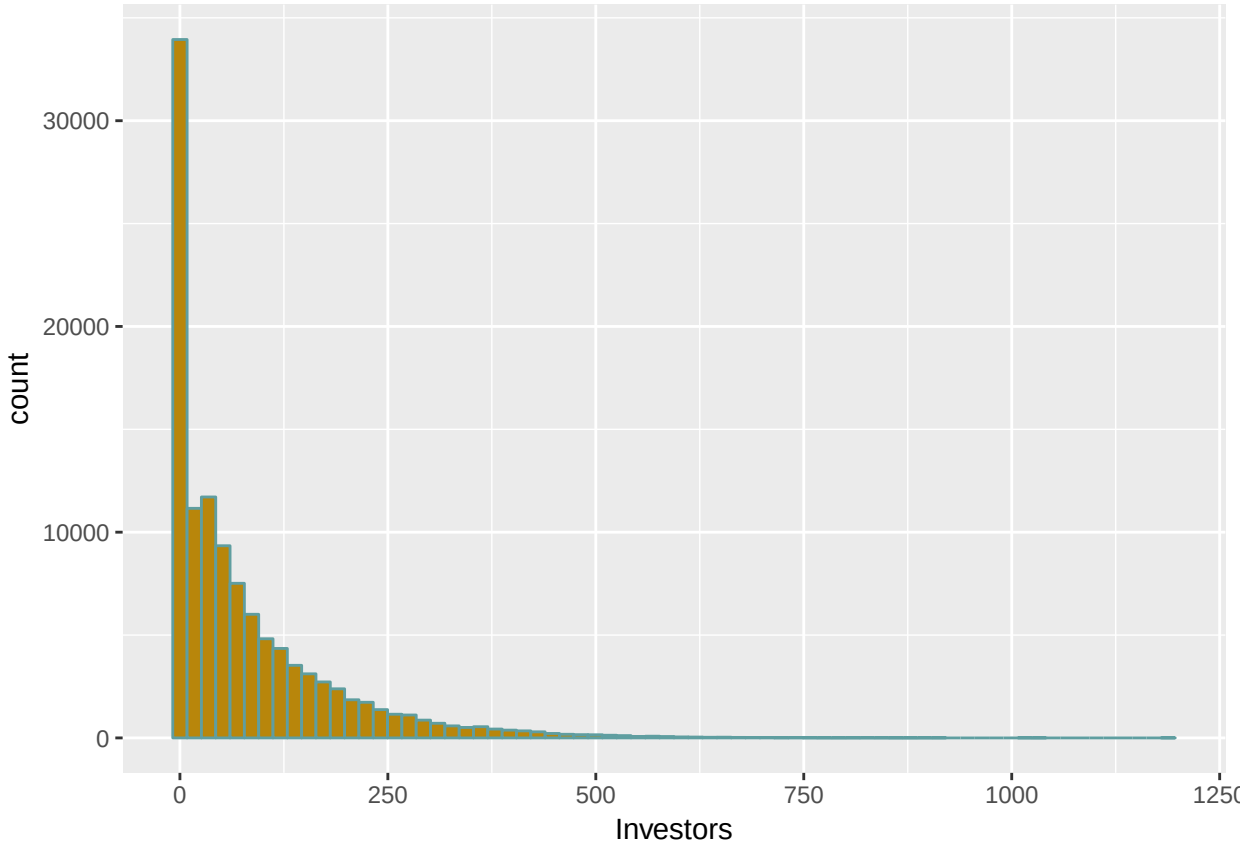
5. How are loans distributed geographically?



Bigger States like California or Texas would have a greater number of loans. California has the highest number, partly because it's where Prosper put their headquarters, the other part, I believe, is because of California's strong economies.

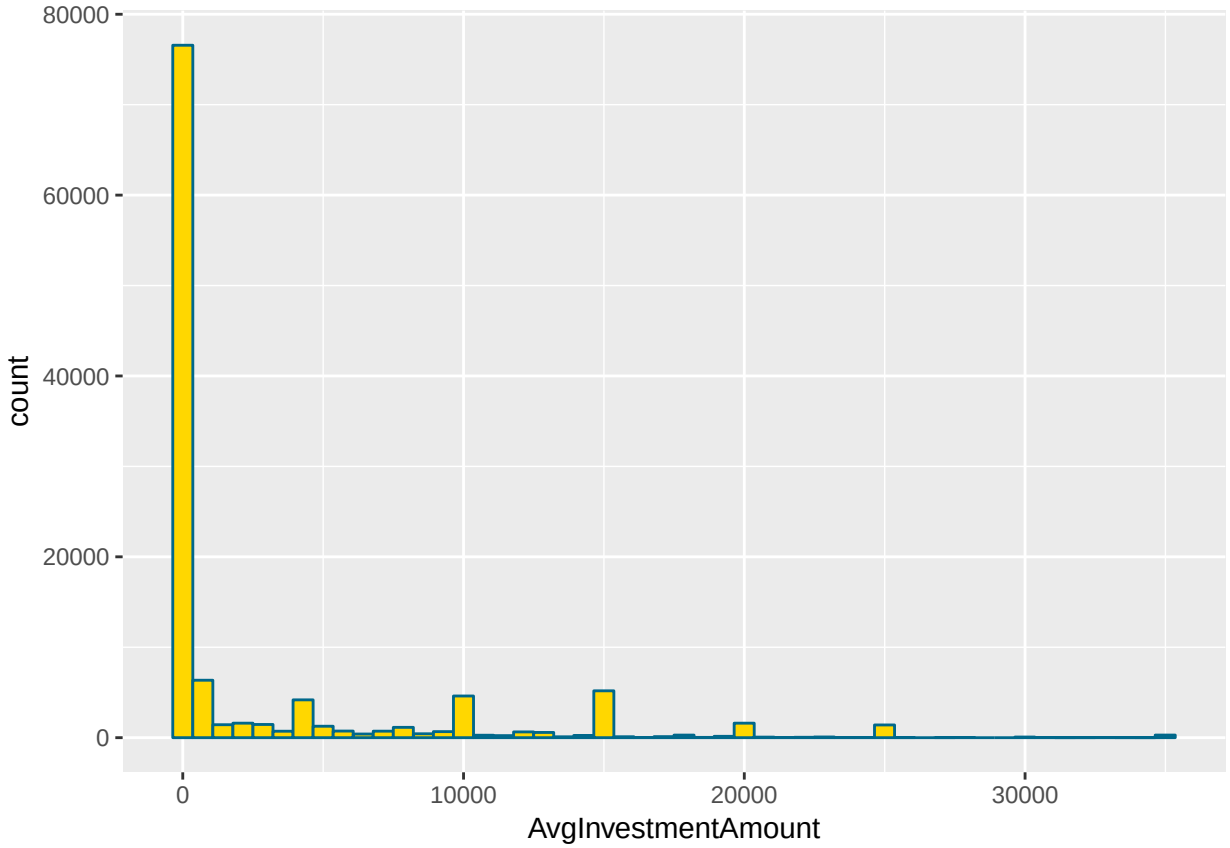
PART II:

In this part of the analysis, I want to see if and how investors rationalize the amount of investment, to be more specific, how will the prospect and the risk of the loan influence the lender's investment amount?




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

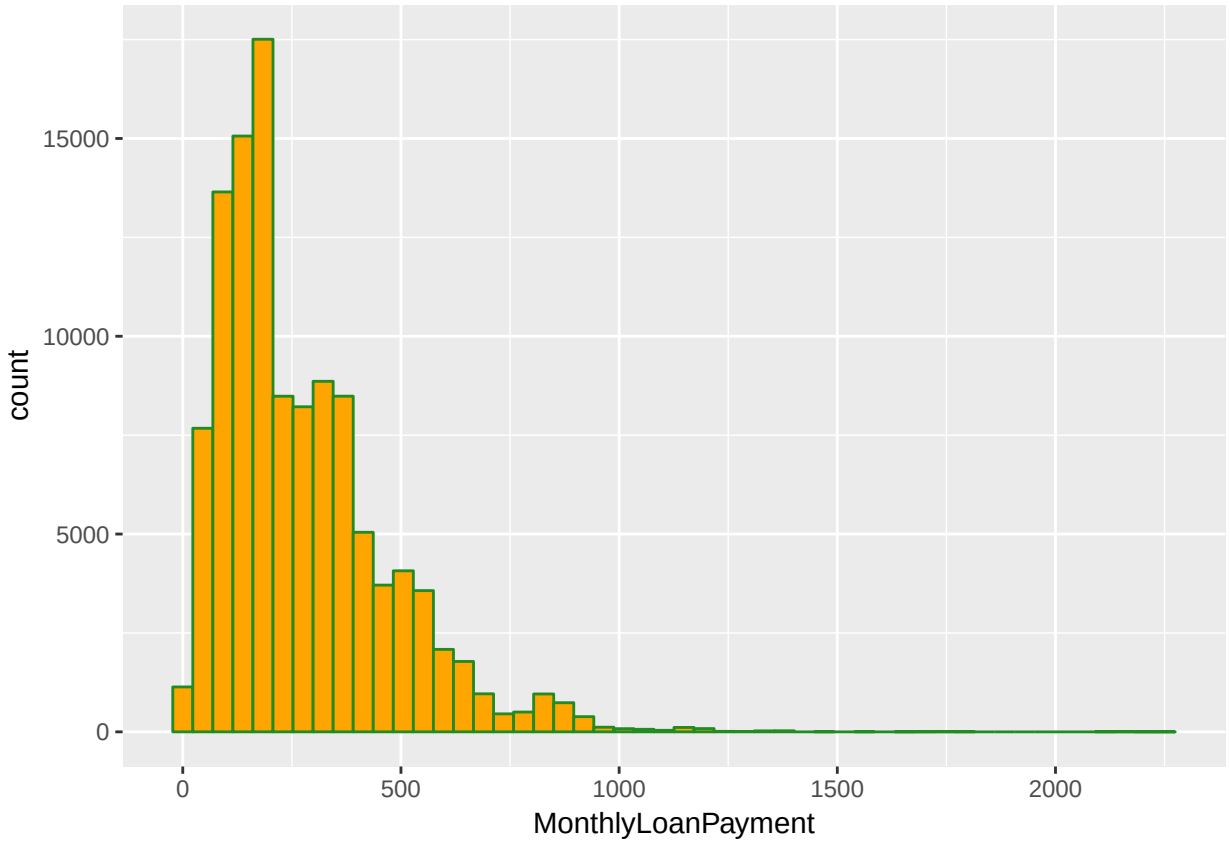
The investors distribution is skewed to the right, with the majority of loans having less than 100 investors.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     7.3    54.4    97.4  2854.0  2000.0 35000.0
```

The lender's investment decision will be represented by the average investment amount.

Investors generally prefer to play it safe, with the majority of average investment at below 2000. Some loans, however, were able to convince investors to put up a larger investment. Perhaps, these loans promised greater return at lower risk.



The prospect of the loan is represented by the lender's expected payment, which is the result of the multiplication of Term and MonthlyLoanPayment.

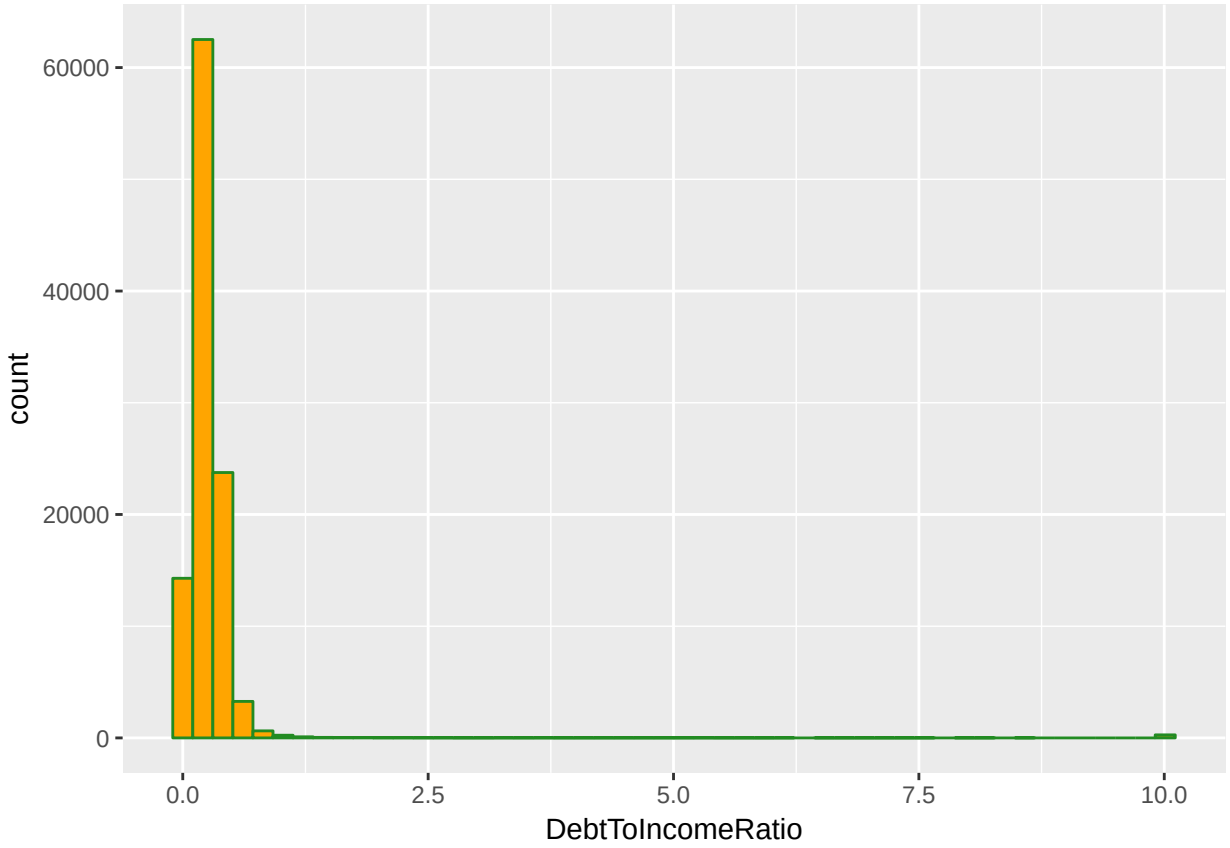
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   131.6   217.7   272.5   371.6  2252.0
```

```
## [1] 935
```

There is a problem, however. The MonthlyLoanPayment data is supposed to be bigger than 0, no matter how I look at it, a 0 is just unnatural.

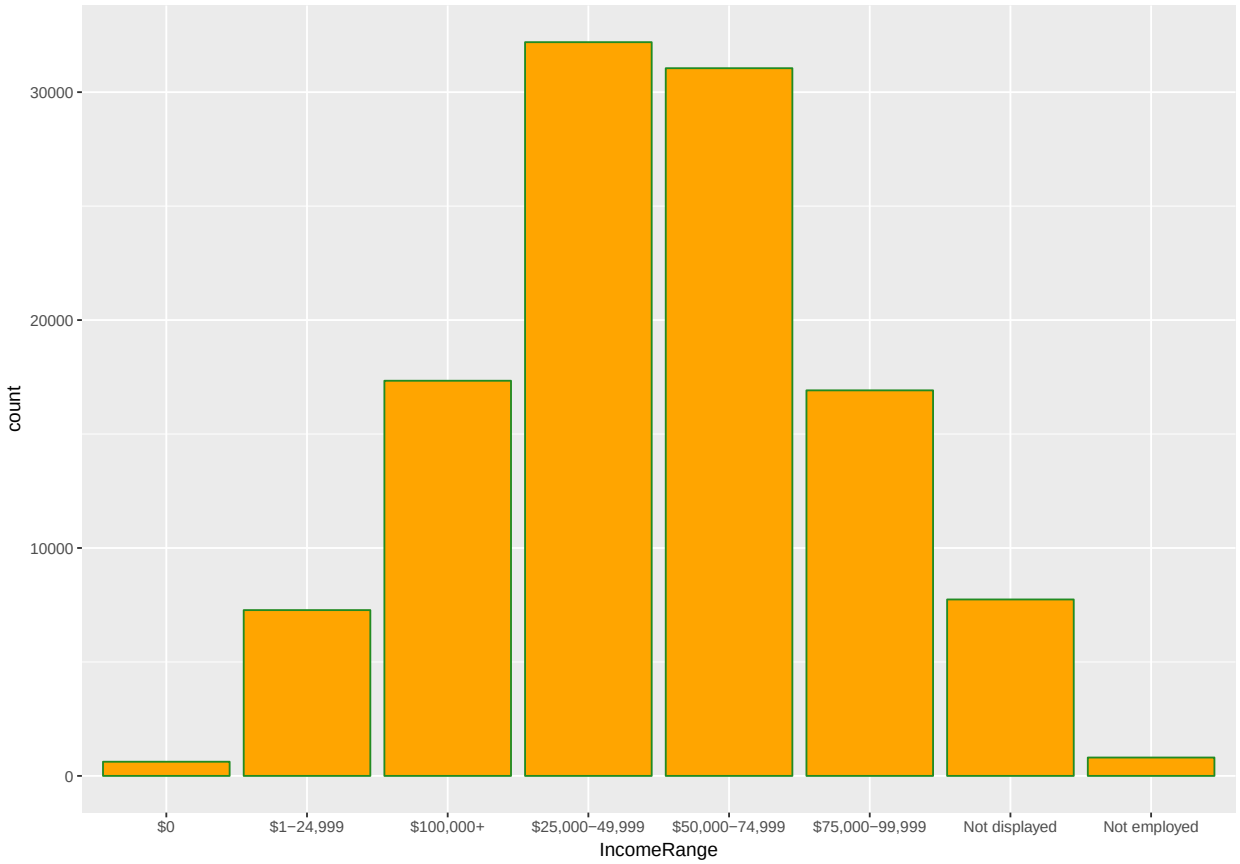
I looked for the source of it, but there was nothing. It could be another kind of engineering error, where the loan payment data was simply not processed. That said, I will still press forward with MonthlyLoanPayment, since there's no such thing as perfect, especially when it came to data.

Another factor that will influence the investor is the risk. The first variable the came to my mind is the DebtToIncomeRatio.

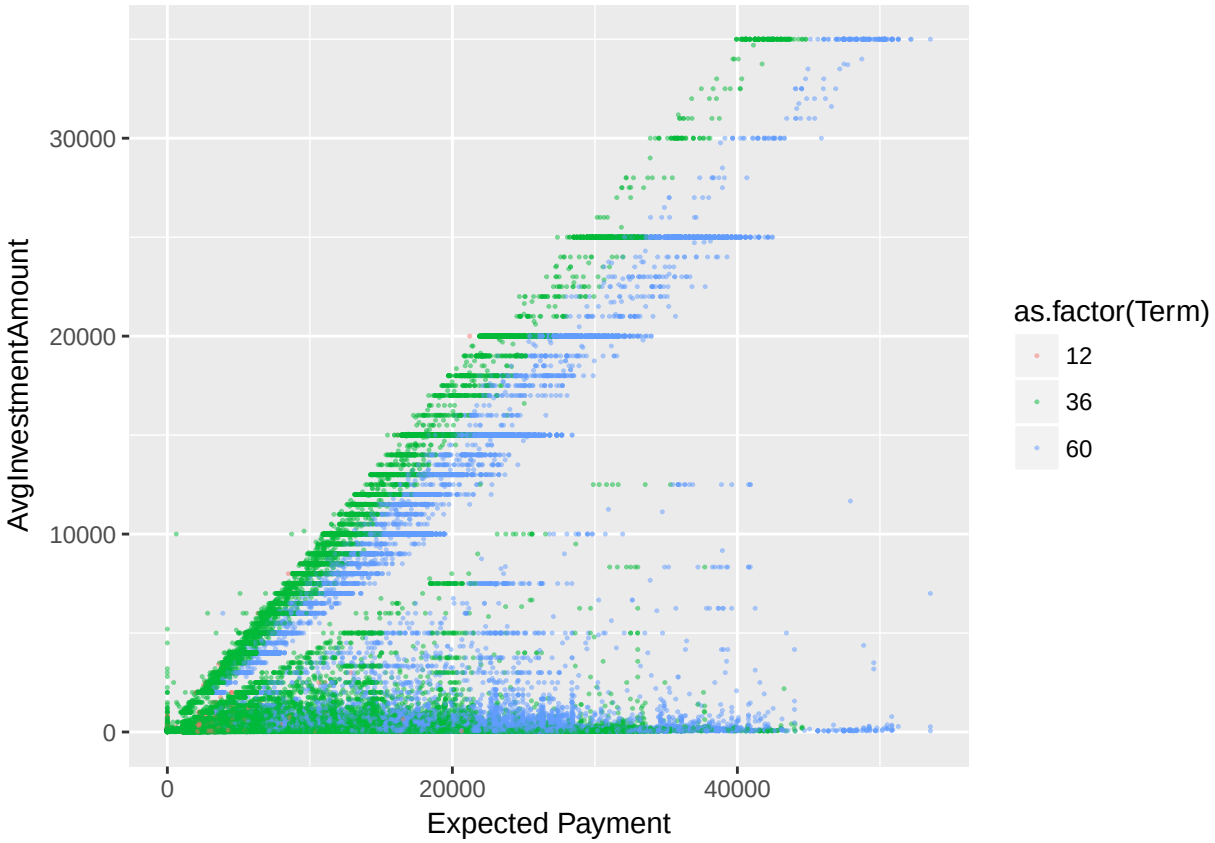



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   0.140   0.220   0.276   0.320  10.010    8554
```

People took the debt to income ratio seriously, most investors will not go for loans in which the debt to income ratio is higher than 1%, and no one will push higher than 10%.

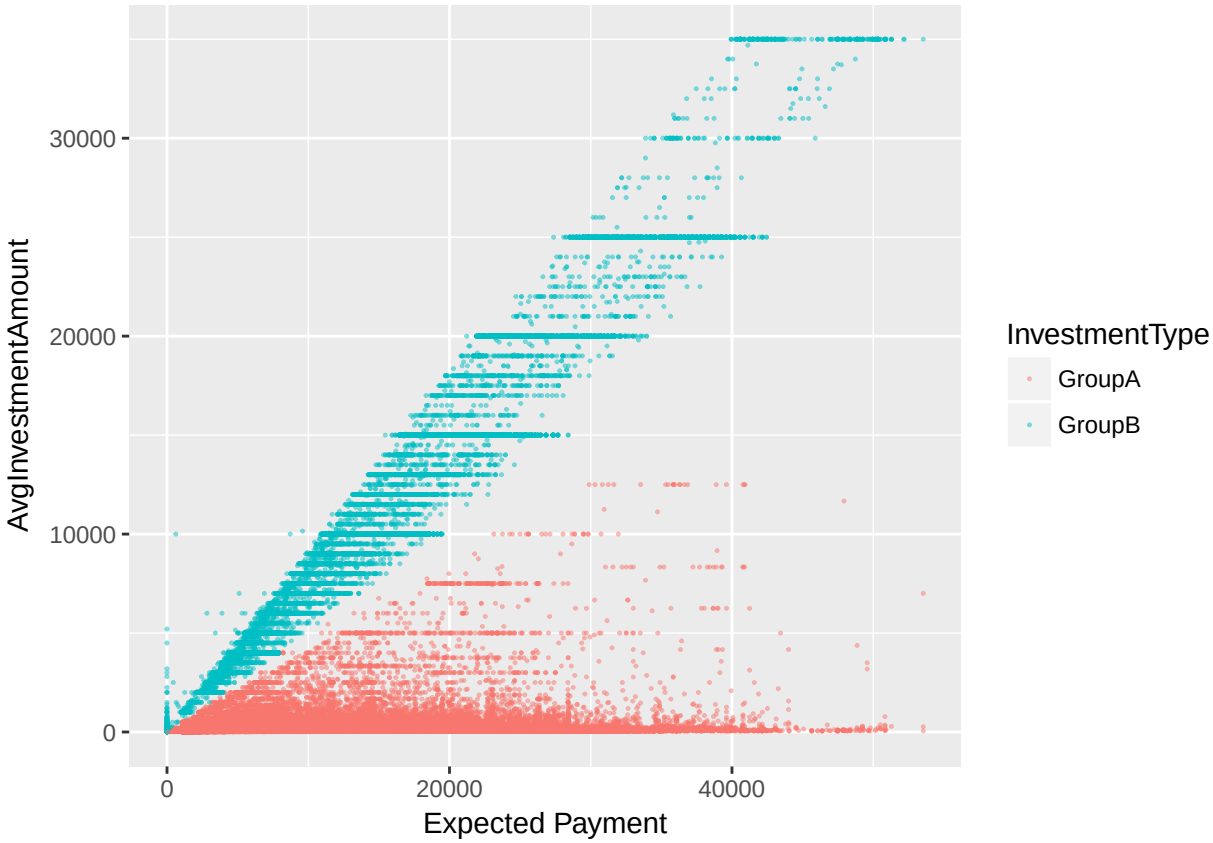


So instead of factoring in the DebtToIncomeRatio, I will incorporate the IncomeRange, because I know, no matter where I look at, chances are that the loan amount will only be around 1% of the borrower's income.

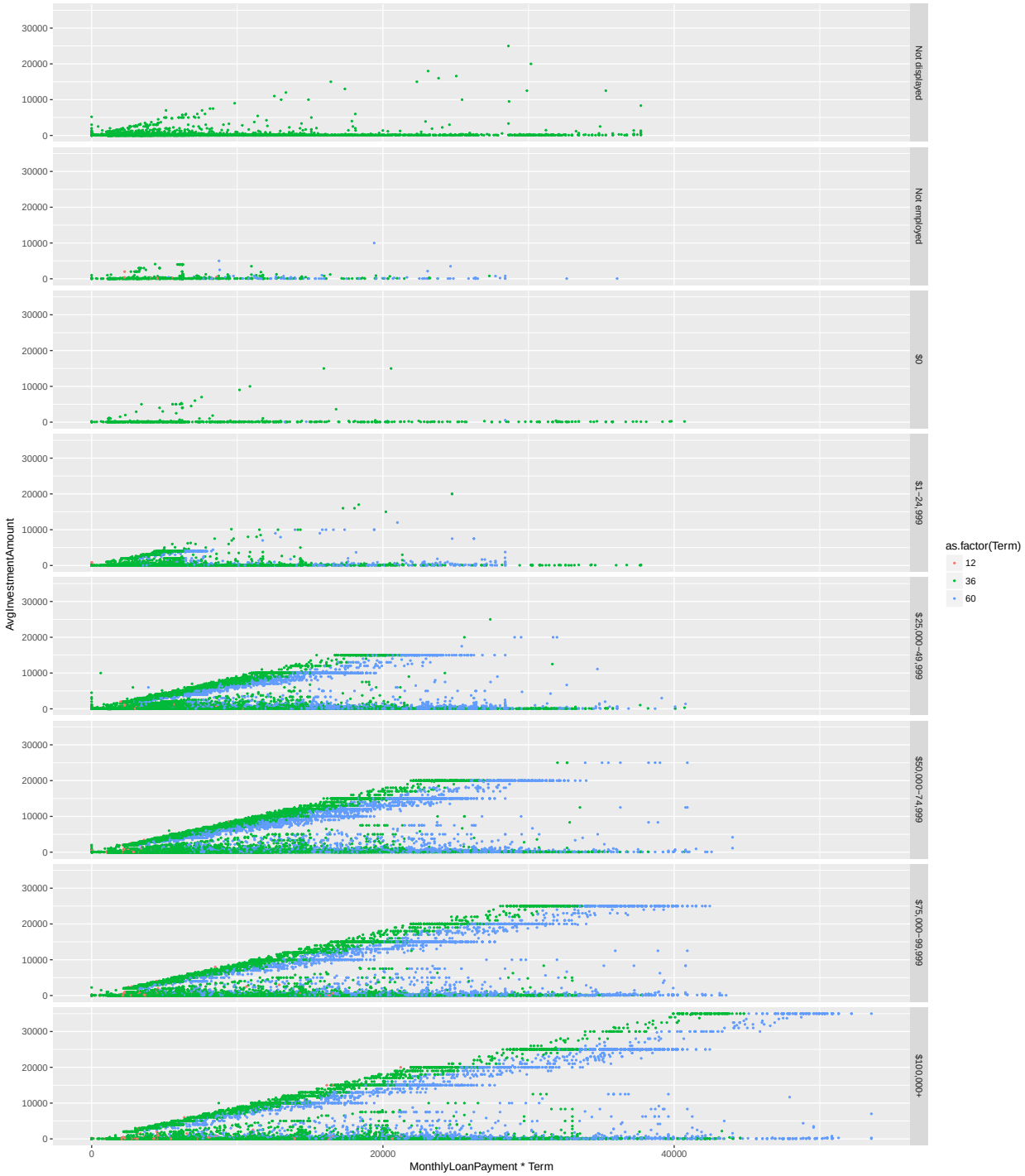


The longer the terms, the higher the expectation. There are 2 kinds of investment practice depicted in this graph:

- Low average investment amount: Don't put all your eggs in one basket, the standard practice in the finance world, identify as many prosperous investments as possible and put in small amount while expecting a large return on investment. Investors who made this kind of investments will be referred to as group A investors.
- High average investment amount: Occasionally, some golden deals will appear, investors who are experienced enough to identify them will want to grab up a bigger bite, sometimes the whole pie. Investors who made this kind of investments will be referred to as group B investors.



I visualized them so that viewers could get a clear picture of the 2 groups.



Group B investors are more sensitive to the income level, and they generally prefer to lend to people with higher income.

```
## Source: local data frame [8 x 4]
## Groups: IncomeRange [8]
##
##      IncomeRange GroupA GroupB AboveAvgRatio
##           <fctr>  <int>  <int>         <dbl>
## 1  Not displayed   7477    264    0.03410412
```

```
## 2   Not employed    749     57    0.07071960
## 3             $0    585     36    0.05797101
## 4      $1-24,999   6141   1133    0.15576024
## 5 $25,000-49,999  24692   7500    0.23297714
## 6 $50,000-74,999  21979   9071    0.29214171
## 7 $75,000-99,999  11696   5220    0.30858359
## 8      $100,000+  11848   5489    0.31660610
```

The higher the income level, the more confidence the people from Group B.

```
## # A tibble: 2 × 3
##   InvestmentType CreditScoreRangeUpper CreditScoreRangeLower
##            <chr>                 <dbl>                 <dbl>
## 1         GroupA              702.6453              683.6453
## 2         GroupB              710.2243              691.2243
```

Another way to look at this is through the credit score. Borrowers financed by Group A investors typically have a lower credit score than people from group B.

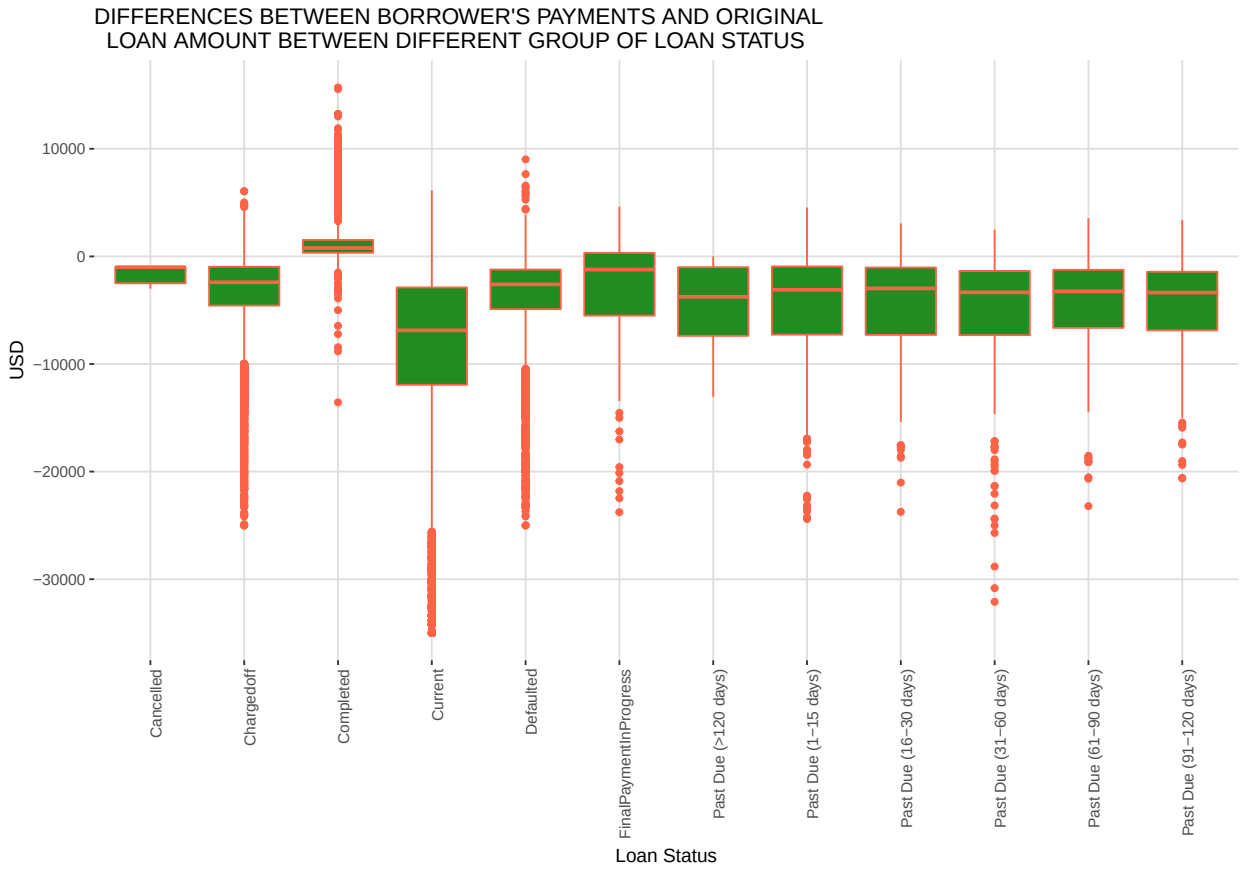
```
## # A tibble: 2 × 4
##   InvestmentType LoanOriginalAmount ExpectedPayment EarningRatio
##            <chr>              <dbl>           <dbl>        <dbl>
## 1         GroupA           7519.496        10245.51     1.362526
## 2         GroupB          11004.851        14940.02     1.357585
```

Group A's investors generally go for smaller loan, with higher expectation.

FINAL PLOTS AND SUMMARY

Plot 1

I tried to ascertain the meaning of loan status. I wanted to know about the thing that set these status apart.

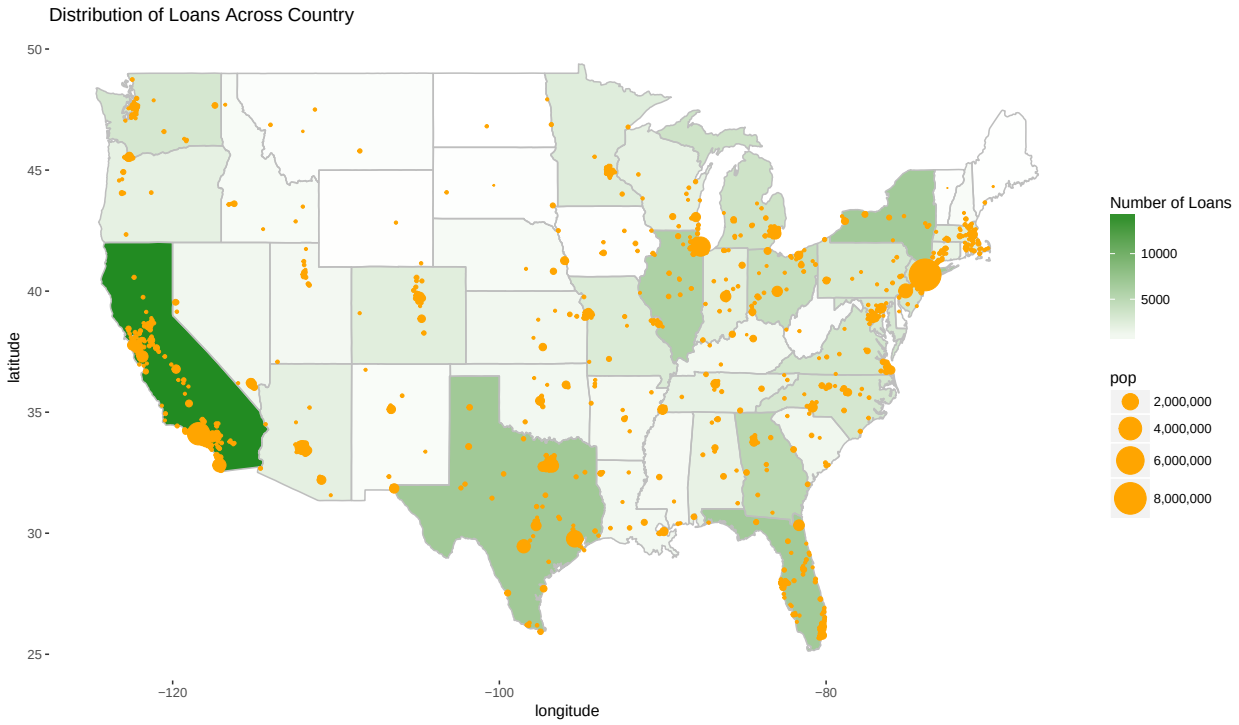


The lenders will get a green cashflow for a completed loan, and red for everything else.

```
## # A tibble: 12 × 2
##                LoanStatus AvgCashFlow
##                    <fctr>       <dbl>
## 1               Cancelled   -1700.000
## 2             Chargedoff    -3510.592
## 3               Completed    1133.481
## 4                 Current   -7861.496
## 5               Defaulted   -3879.319
## 6  FinalPaymentInProgress   -3260.333
## 7    Past Due (>120 days)   -4505.566
## 8    Past Due (1-15 days)   -4451.419
## 9   Past Due (16-30 days)   -4413.277
## 10  Past Due (31-60 days)   -5067.413
## 11  Past Due (61-90 days)   -4406.525
## 12 Past Due (91-120 days)   -4541.186
```

Plot 2

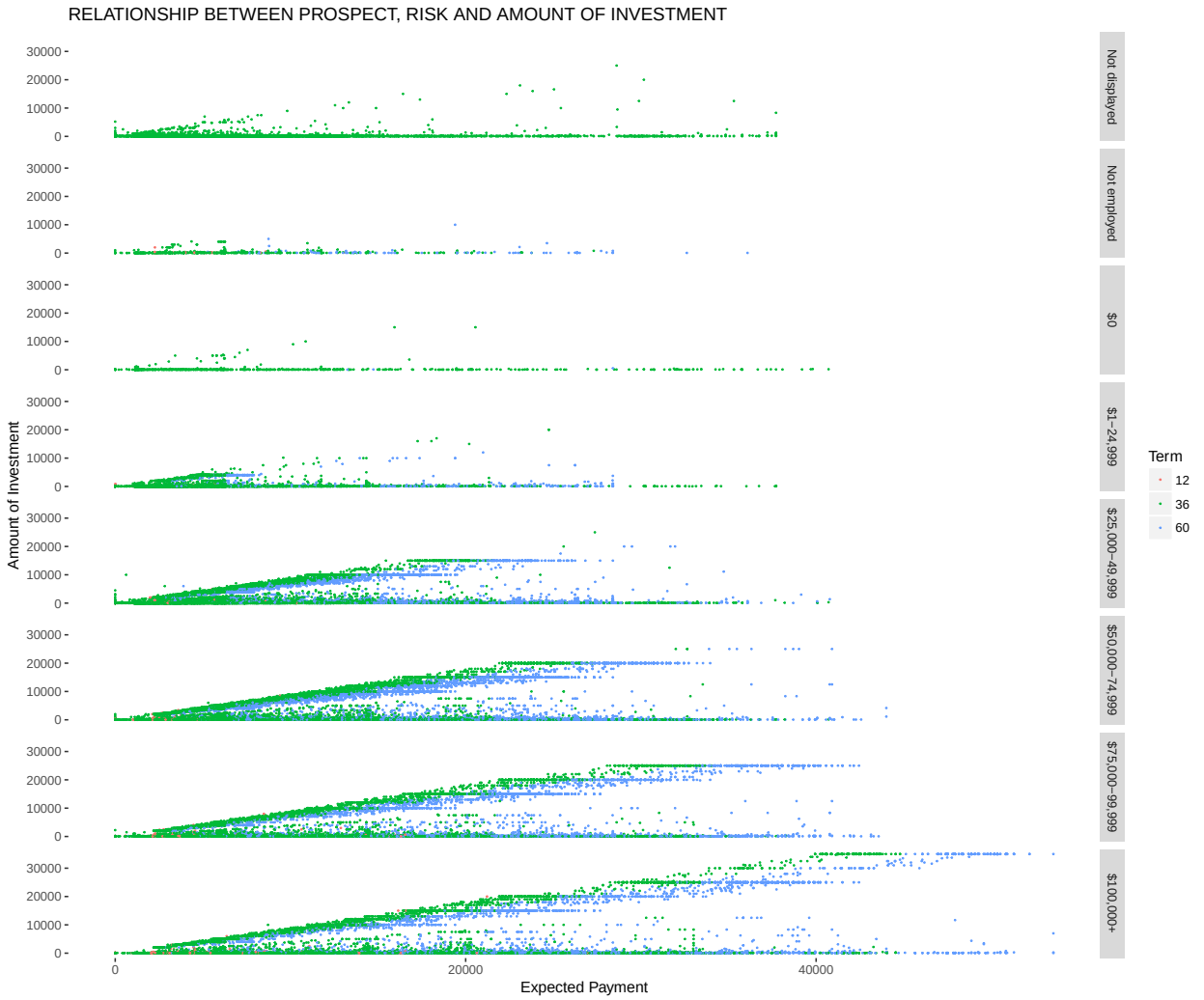
Here I was curious about the geographic distribution of loans. I wanted to see if loans were distributed evenly across country.



Bigger States like California or Texas would have a greater number of loans. California has the highest number, partly because it's where Prosper put their headquarters, the other part, I believe, is because of California's strong economies.

Plot 3

I wanted to see if and how the prospect and the risk of the loan influenced the lender's amount of investment.



There is 2 type of investments:

- Higher risk, higher expectation, spreaded across many small invesments.
- Lower risk, lower expectation, focused on key loans.

The results are questionable, of course, because there wasn't any recording the investor's investment, making it impossible to determine if these 2 mentioned above type of investments are individual-based or situation-based.

Reflection

The data set contains 113937 observations, with 81 variables. It was just huge, there was no direction at all, no nothing. But thanks to that, I was able to understand just how important would the domain knowledge is.

I started by exploring the LoanStatus, I wanted to check if these loan status reflect what they were meant to reflected, meaning completed loan should give out a positive balance, and anything else shouldn't.

Then I noticed some strange pattern with the loan amount and the date of the loan, with some hacks, I found the answer to the loan amount, as for the date, there was simply no data that could explain it. Found my answer on wiki.

Next, it was a real struggle, I've already explored half of the data set, and I still couldn't figure out any interesting questions, took me a while to educated myself on the finance field, and thanks to that, I was able

to come up with some interesting things.

Some investors would like to take it safe, some would like a bigger return, some try to lower the risk by placing lots of investments, while some try to identify some lower risk loan, and put all their money in that. This finding is still questionable, because of the lack of data.

I was able to get something going for the analysis of Proper's benefit. However, the lack of data hampered my movements, and it took too much of effort to extract the borrower's fee from the available data.

Some future direction:

- Some kind of model that will predict the number of investors that a loan might attract.
- See how investor's investment strategy played out, between spreading the risk to a lot of loans, and putting large amount of money on some particular safe loan.