Exploratory Data Analysis for Loan Data

Barbara Joebstl

March, 12 2017

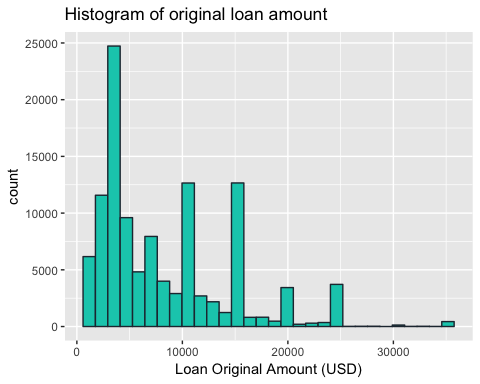
Table of Contents

# Loan data

In this project we are analysing a dataset from the company Prosper, who is part of the peer-to-peer lending industry.

# Univariate Plots Section

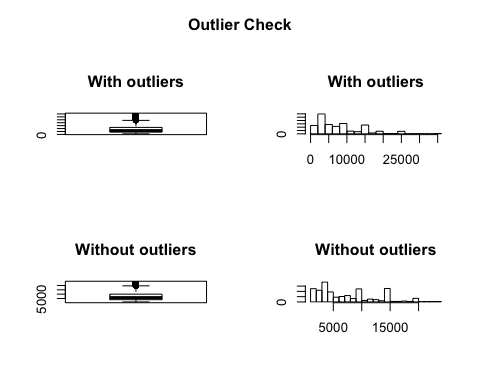
In this section we will preform prelaiminary exploration of the dataset to get an understanding ot the structure and the indivual variables in the loan dataset.



## [1] "Summary of Loan Original Amaount(USD):"

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

The loan orginial amount is the amount that was bid. The median of the loans is 6500. I suggest the money is needed for extra expenses due to unexpected problems, like home improvements or taking a small loan for a holiday. It seems the loans above 25000 are not often needed. So we will perform an outlier check.

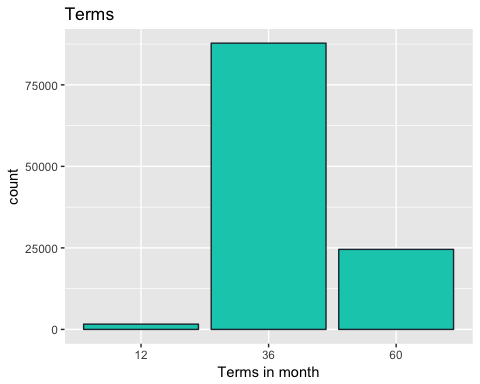


## Outliers identified: 4395 nPropotion (%) of outliers: 4 nMean of the outliers: 26253.53 nMean without removing outliers: 8337.01 nMean if we remove outliers: 7618.17 nDo you want to remove outliers and to replace with NA? [yes/no]:   
## Nothing changed n

There are 4395 outliers identified from the 113937 data objects. We will first replace the outliers with NA values and then create a new filtered data frame without the outliers.

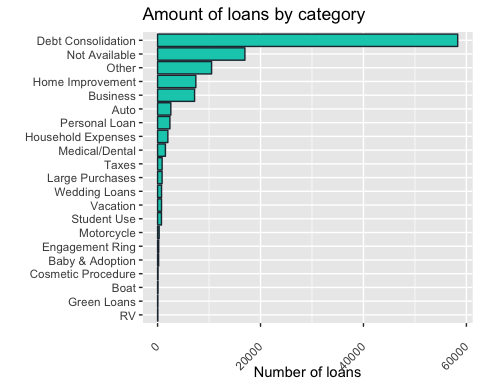
## [1] "Summary of Terms:"

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



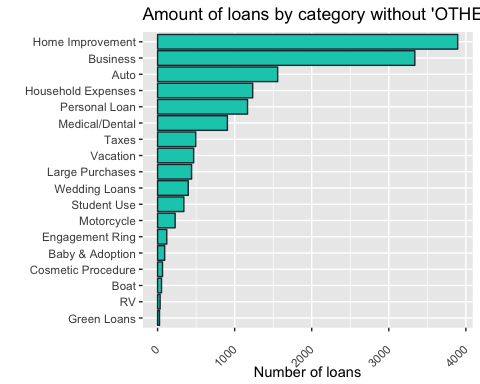
The loan takers can choose between a 12, 36 or 60 month long term. In the plot above we can see that most of the loans are taken with a term of 36 month.

To get a better readability, we are going to map the numeric values to better readable strings according to this site: <https://www.prosper.com/Downloads/Services/Documentation/ProsperDataExport_Details.html>



As we can see in the plot above the most loans are needed for the categories "Debt Consolidation", "Not Available" and "Other". So it seems loan takers do not want to tell the purpose of their loans.

We will make another plot where we skip those categories, to take a closer look at the other categories.



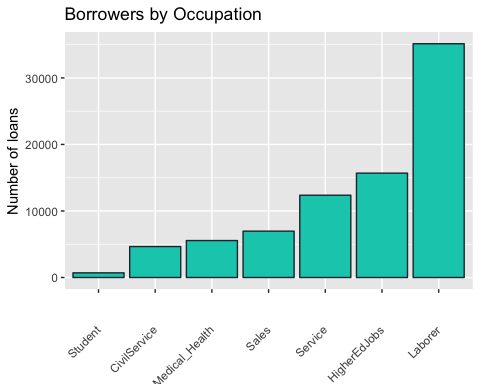
We can see that now the most popular purposes are "Home Improvement", "Business" and "Auto".

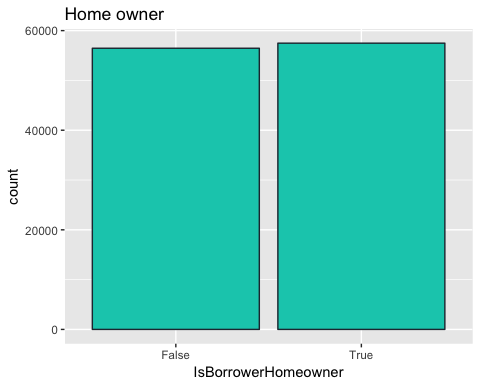
### Occupation of loan takers

## [1] "Head of Occupation Data:"

## [1] Other Professional Other Skilled Labor Executive   
## [6] Professional   
## 68 Levels: Accountant/CPA Administrative Assistant Analyst ... Waiter/Waitress

Because there are 68 different types of occupation we are going to combine groups into a new data frame into bigger occupation groups.

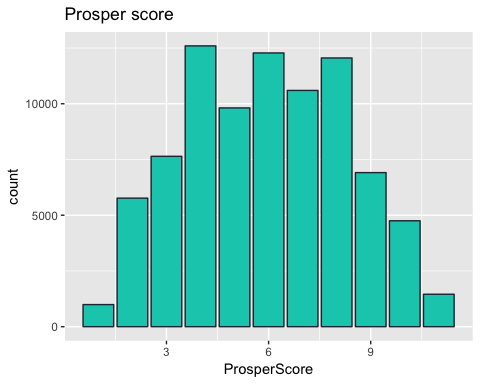




## [1] "Summary of Home Owners:"

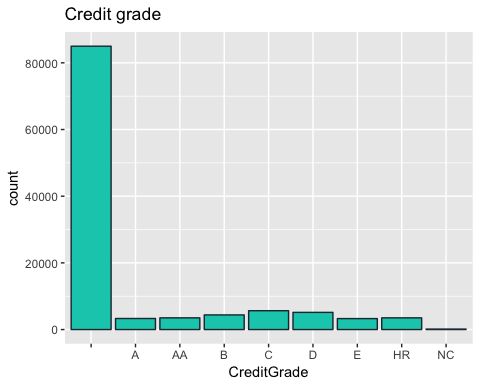
## False True   
## 56459 57478

Almost the same amount of loan takers ar home owners.

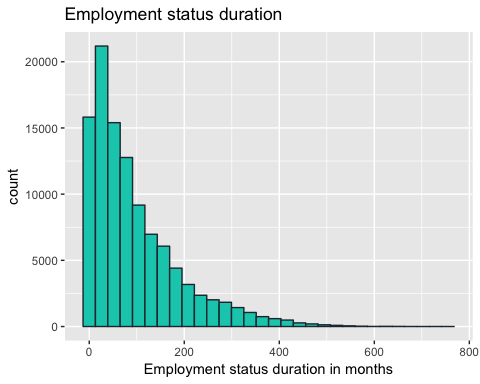


## [1] "Summary of Prosper Score:"

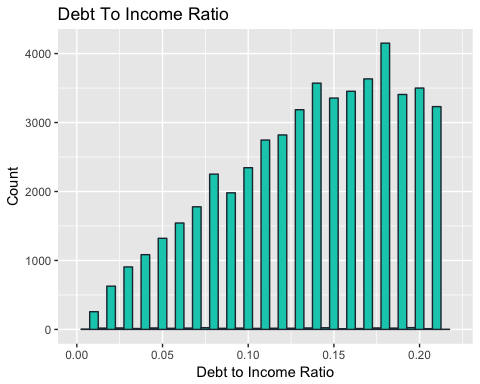
## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 1.00 4.00 6.00 5.95 8.00 11.00 29084

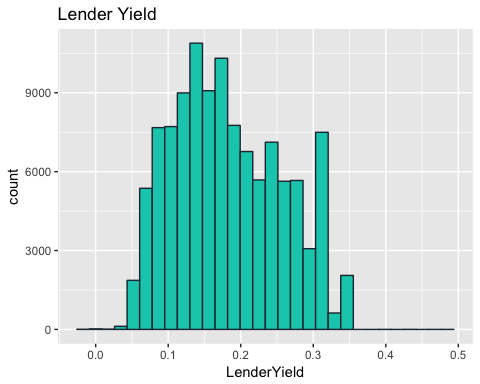


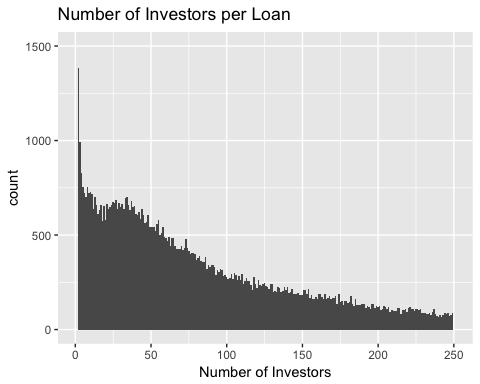
The plot above shows that there is most of the data for credit grade missing.



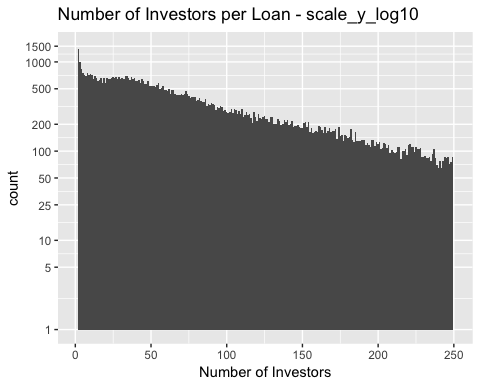
We can see that the longer people have an employment the less they need a loan.







Now we are going to transform this plot using the scale\_y\_log10 function to have a better handling for the outliers.



# Univariate Analysis

### Structure of the dataset

This data set contains 113,937 loans with 82 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 explanation of the variables can be found there: <https://www.prosper.com/Downloads/Services/Documentation/ProsperDataExport_Details.html>

### Main features of interest in the dataset

We want to know how much money is needed, when and why. So the most important variables are 'OriginalLoanAmount', 'LoanOriginationDate' and 'Category'. We think it is also interesting to see if there is a difference in loan taking between home owners and non home owners. Then we are also interested to see if the credit grade and the prosper score are related to other variables.

### More features

Another point of interest is the lender yield. And we are alos curious about the fees that the company takes.

### New variables

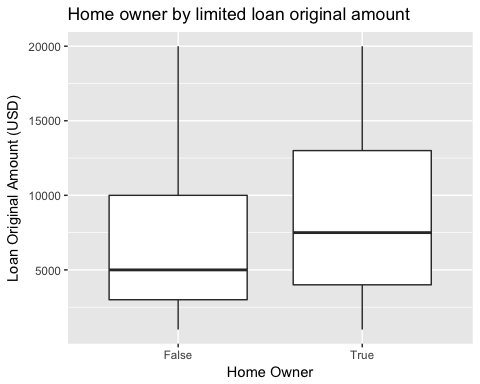
To provide a better readability we created the variable "Category", where we mapped the categories to the numbers of the 'ListingCategory..numeric'. We also created a new variable named 'GroupedOccupation'. Because the original variable 'Occupation' consits of 68 levels, we wanted to group those to get a better overview. The new variable consits of 8 levels representing our occupational groups.

### Changed variables

In the LoanOriginalAmount variable we perforemed an outlier check and removed those values in order to make the following analysis more robust.

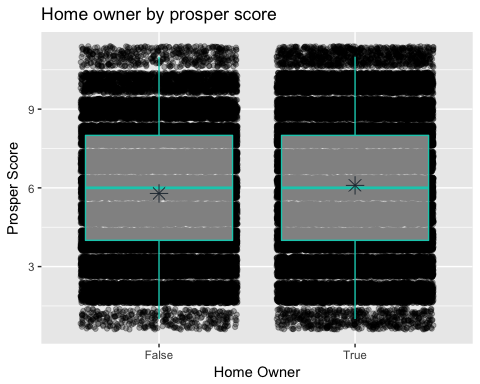
# Bivariate Plots Section

No we want to take a look at the home owners. In the next step we want to proof that people who aren t home owners need more often small loans for vacation, home improvement or household expenses than home owners. Our suggestion that house owners need less loans, seem to be wrong. Almost half the amount of borrowers are house owners. We limit the loan amount to a smaller range, because we want to know if house owners also need smaller loans, for home improvments or others.

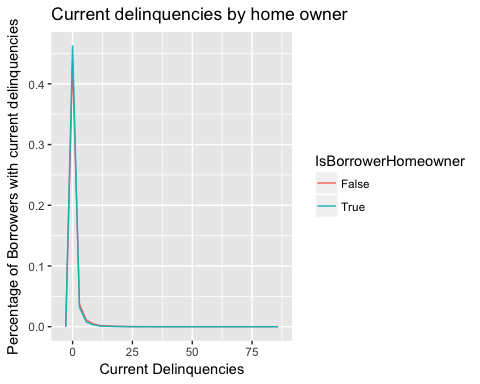


In the plot above we can see that house owners need bigger amounts of money than non house owners.

Next we want to see, if there is a relation between the prosper rate and the fact that the borrower is a house owner. Normally a house owner has a better rating, due to more financial security.



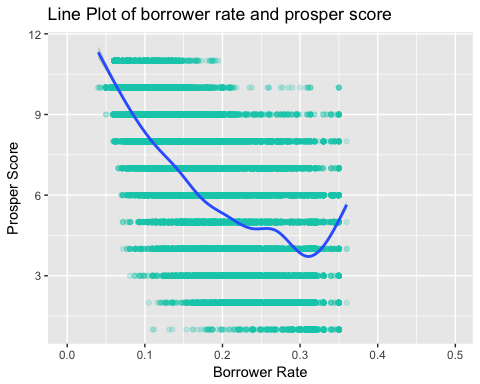
Another surprise here. The Prosper score for non home owners is just a little bit lower.



In the plot above we can not notice any difference between home owners and non home owners in current delinquencies.

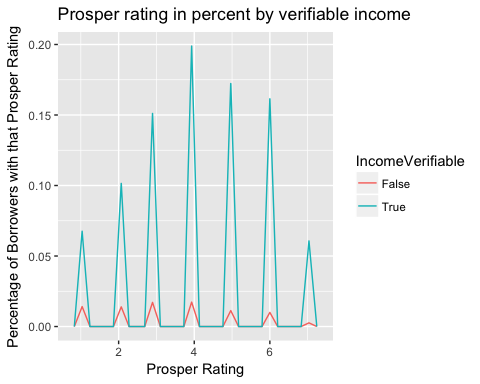
## [1] "Credits By Prosper Rating:"

## # A tibble: 8 <U+00D7> 6  
## ProsperRating..numeric. mean\_amount median\_amount min\_amount max\_amount  
## <int> <dbl> <dbl> <int> <int>  
## 1 1 3463.114 4000 1000 16800  
## 2 2 4586.405 4000 1000 15900  
## 3 3 7083.439 6100 1000 15000  
## 4 4 10391.940 10000 1000 25000  
## 5 5 11622.355 10000 1000 35000  
## 6 6 11459.886 10000 1000 35000  
## 7 7 11583.539 10940 1000 35000  
## 8 NA 6159.303 4500 1000 25000  
## # ... with 1 more variables: n <int>

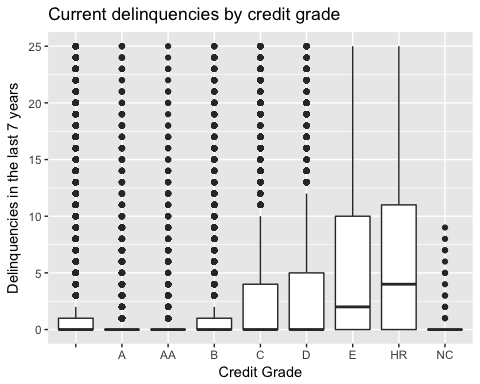


No surprises here, the better the Prosper Score, the better the borrower rate.

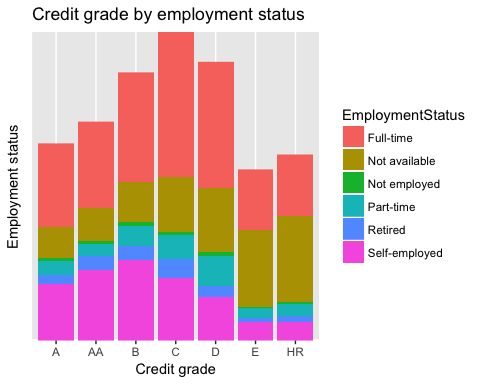
We suggest that the prosper rating is better if the income is verifiable.



Yes, our suggestion is right.

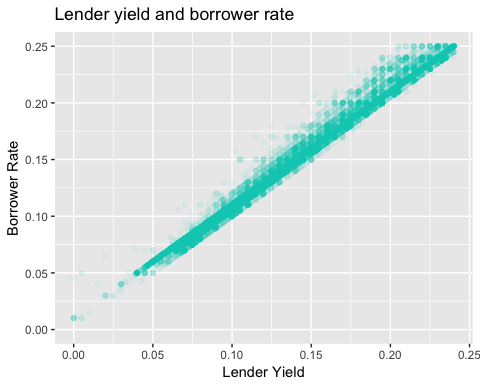


As we can see in the plot above, the better the credit grade, less delinquenices.



The majority of the loan takers are full time employees. The worse the credit grade gets the more likely the employment status is "not available". We exclude missing data in our plot, to get a good picture.

We suggest that there is a higher lender yield, if the borrower rate is higher.



Yes, we can see clearly the higher the borrower rate, the higher the lender yield.

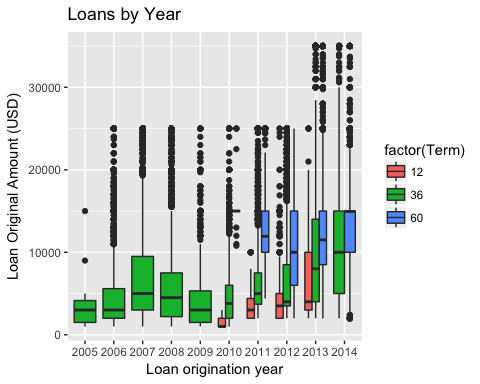
# Bivariate Analysis

We found out that there is almost no difference between home owners and non home owners by prosper score and delinquencies. There is a slight difference as home owners tend to need bigger loans.

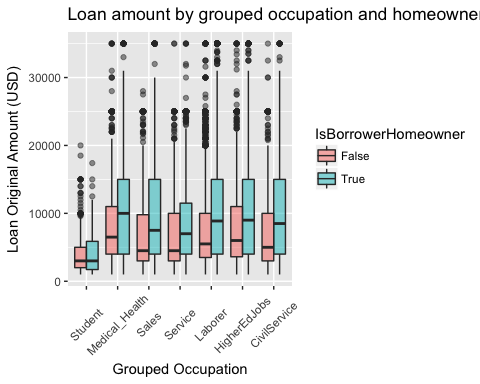
The prosper rating is better if the income is verifiable. The worse the credit grade the more often occured delinquencies in the last 7 years.

We notice a strong relation between the lender yield and the borrower rate. The higher the borrower rate, the better the lender yield.

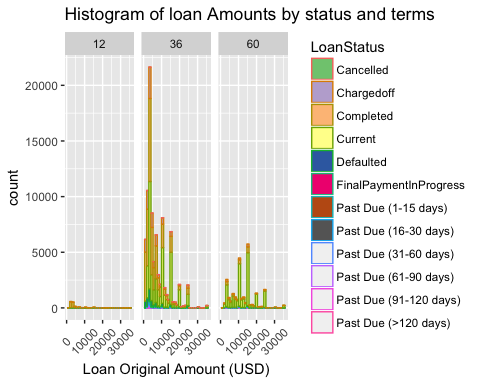
# Multivariate Plots Section



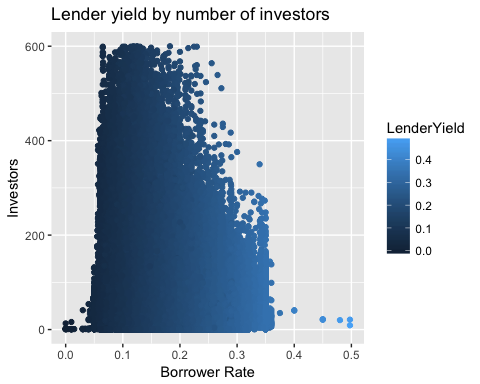
In the plot above we can see that from 2006 to 2010 the loans where taken for 36 month. Maybe back than this was the only term available. From 2010 on if the loan amount was higher borrowers selected the 60 month term. In 2011 the loans increased a lot.



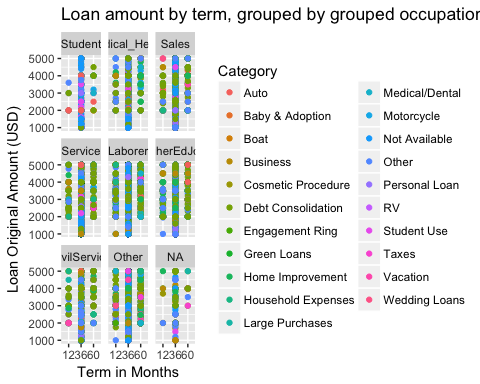
In the factor plot above, we can see that students need the smaller amounts of money. The higher the loan gets, the more homeowners are the borrowers.



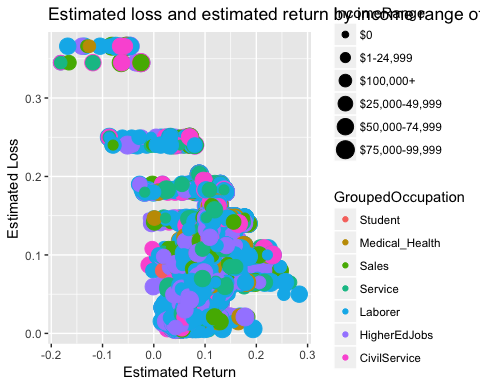
In the plot above we can see that most of the loans are mid-term.



In the plot above we can see quite well, that the investors yield gets higher the higher the borrower rate gets.



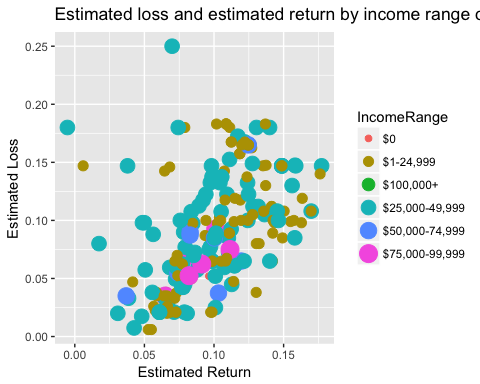
The plot above gives us a nice overview. As we can see most of the loans are mid term loans with a duration of 36 month. The usages of the loans are well mixed.

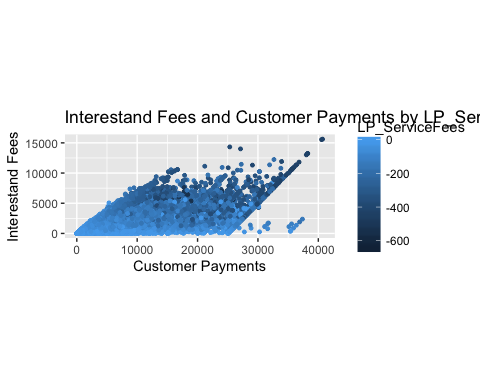


The plot above gives an overview of the estimated return and estimated loss by grouped occupation and income range. As there is to much information in the plot, we hardly can see anything. So we are going to split this into two plots.

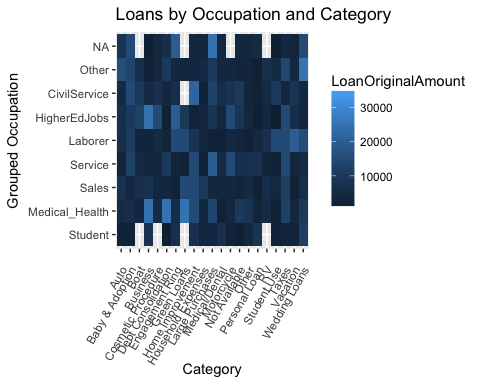


The plots above are still not readable. So in the next plot we want to focus on specific occupation group and the estimated losses and returns.





This plot shows that the lower the customer payments are the lower the service fees and interest fees are.



In the plot above we get an overview of the loans by category and grouped occupation. It is quite hard to find a pattern on the first sight, so maybe a normal list would have done a better job.

# Multivariate Analysis

### Estimated loss and estimated return by income range and grouped occupation

We ploted an overview of the estimated return and the estimated loss by grouped occupation and income range. Because the plot was to dense and unreadable, we split it up into two plots. One for the income range and one for the occupation group. But this plot is not very readable either. So we focused on one occupationl group - the students. This plot is well readable. And it was quite surprising that there is one outlier with a negative estimated return and a really high estimated loss.

### Heatmap of grouped occupation and category by loan original amount

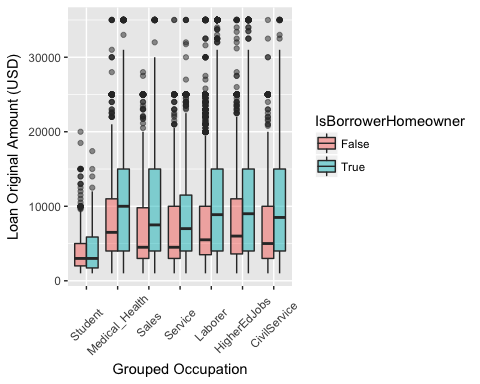
It was expected that this heatmap gives a nive overview of the occupations and categories. We were hoping to find patterns with one look. But this is not the case, you have to elaborate this plot. Maybe it would have been better to provide a list in this case.

### Loans and Fees

From 2006 to 2010 the loans where taken for 36 month. Maybe back than this was the only term available. From 2010 on if the loan amount was higher borrowers selected the 60 month term. In 2011 the loans increased a lot. We found out that the lower the customer payments are the lower the service fees and interest fees are.Which is not a surprise.

# Final Plots and Summary

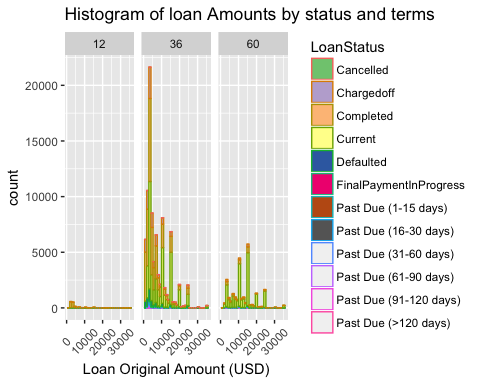
### Loan amount by grouped occupation and homeowner status



### Description

In the factor plot above, we can see that students need the smaller amounts of money. Surprisingly there are a lot of home owners in the group student. The higher the loan gets, the more homeowners are the borrowers. In some occupational groups there are a lot more home owners. This may be caused by our occupational grouping, which contains professions with a wide income range. For example in the group 'Medical\_Health', there are doctors and nurses etc.

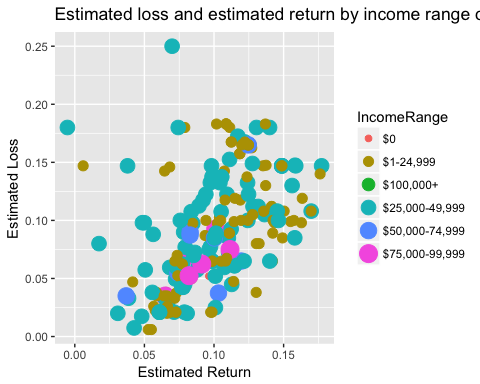
### Histogram of loan amounts by term and loan status



### Description

In the plot above we can see that most of the plots are taken with a terma of 36 month. Even though the amounts are not that big. This may be because until 2009 there where only loans with a term of 36 month. This may also explain why the loan status is completed or charged off for the most of the loans under 5000. We can also see that short term loans are not often used at the moment. People prefer mid or long term loans.

### Estimated return and estimated loss for students



### Description

We can see that most of the students earn between $1- 24,999. There are some students that a earn more than $75,000. There are some outliers where the estimated loss is higher than the estimated return. The majority of the estimated returns and estimated losses are between 0.05 and 0.15. Surprisingly there is one outlier with a negative estimated return and a really high estimated loss.

## Summary

Because of the big amount of variables it took some time, to read through the explanations of the prosper loan data. To get started we explored some different variables. In order to get nice plots, we had to convert some values. For example the origin date to year, the numeric categories into readable categories and the job duration months where summarized in buckets It was interresting to see that from 2011 on borrowers needed higher loans with longer terms. It was quite a surprise that there is no big difference between home owners and non home owners, because we suggested that home owners are financially more strong and don't need small loans. For the other variables I could not find a lot of surprising facts. For example the worse the credit grade is, the higher the delinquencies in the last 7 years are or that the lender yield gets lower the higher the numbe of investors get. It would be nice if there were not big groups like 'na' or 'other' in the occupation and category group. Maybe there could be also data about the age and gender of the borrower provided, which may lead to interesting findings.