

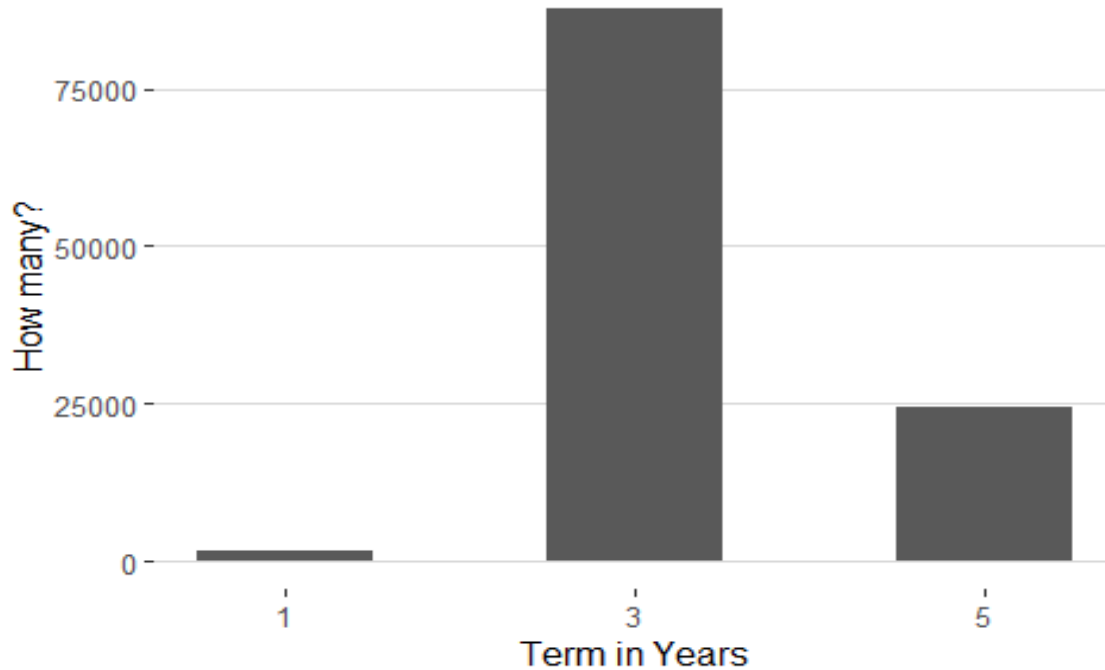
The Following is an analysis of data made on Prosper Company's Loan Prediction data set. Prosper is America's first marketplace lending platform. Get a personal loan at a low rate. Prosper personal loans require generally good credit; this peer-to-peer lender grades your loan so investors can decide whether to fund it.

SOME IMPORTANT TERMS->

1. **Term** : Amount of month customers opted for loan
2. **LoanStatus** : Current status of the loan like chargedoff, completed, defaulted etc...
3. **EstimatedEffectiveYield** : Yield of lenders from borrowers minus the processing fee and late fines
4. **ProsperScore** : Risk Factor score from 1 to 10. 10 being least risky
5. **BorrowerAPR** : The Borrower's Annual Percentage Rate (APR) for the loan.
6. **BorrowerRate** : The Borrower's interest rate for this loan.
7. **ListingCategory..numeric.** : Prosper rating for borrowers in numbers
8. **EmploymentStatus** : Current type of employment
9. **Occupation** : Occupation of borrower at the time of listing
10. **EmploymentStatusDuration** : How long the employee has been employed
11. **IsBorrowerHomeowner** : Does the borrower owns house at the time of listing (True & False)
12. **ProsperRating..Alpha.** : Prosper rating for borrowers in alphabets
13. **IncomeVerifiable** : If the income of the borrower is verifiable at the time of listing (True & False)
14. **StatedMonthlyIncome** : Monthly income of the borrower
15. **MonthlyLoanPayment** : Monthly loan payment amount
16. **Recommendations** : Recommendations the borrowers has at the time of listing
17. **DebtToIncomeRatio** : The debt to income ratio of the borrower at the time the credit profile was pulled.
18. **LoanOriginalAmount** : Original amount of the loan
19. **LoanOriginationQuarter** : Quarter of the month when loan was originated

HOW LONG PEOPLE USUALLY OPT FOR LOAN?

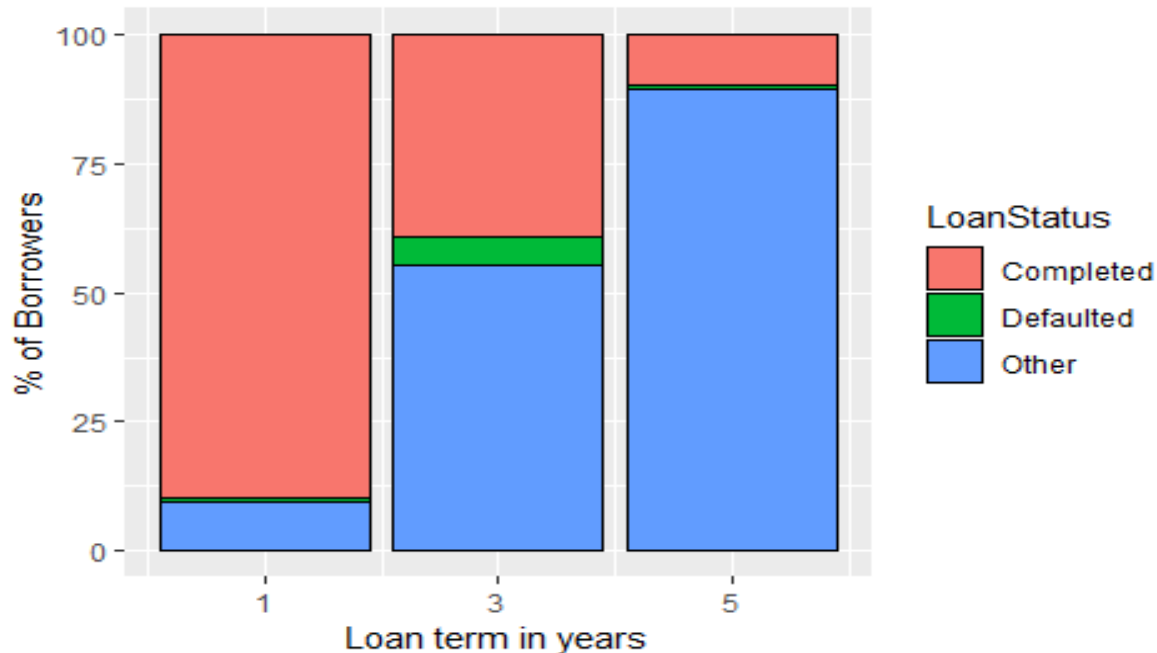
Distribution Loan Terms



I plotted the histogram between 'How many people and Term in years' and what I observed was people don't really loan any amount for less than 1 year and the most popular loan amount is of 3 years , although some people do choose for 5 years. I assume that people who opted for 1 year would fail to repay their loan more as compared to the people who opted for 3 or 5 years.

Plotting the trend of different customer types

LoanStatus: Completed vs Defaulted for each loan Term

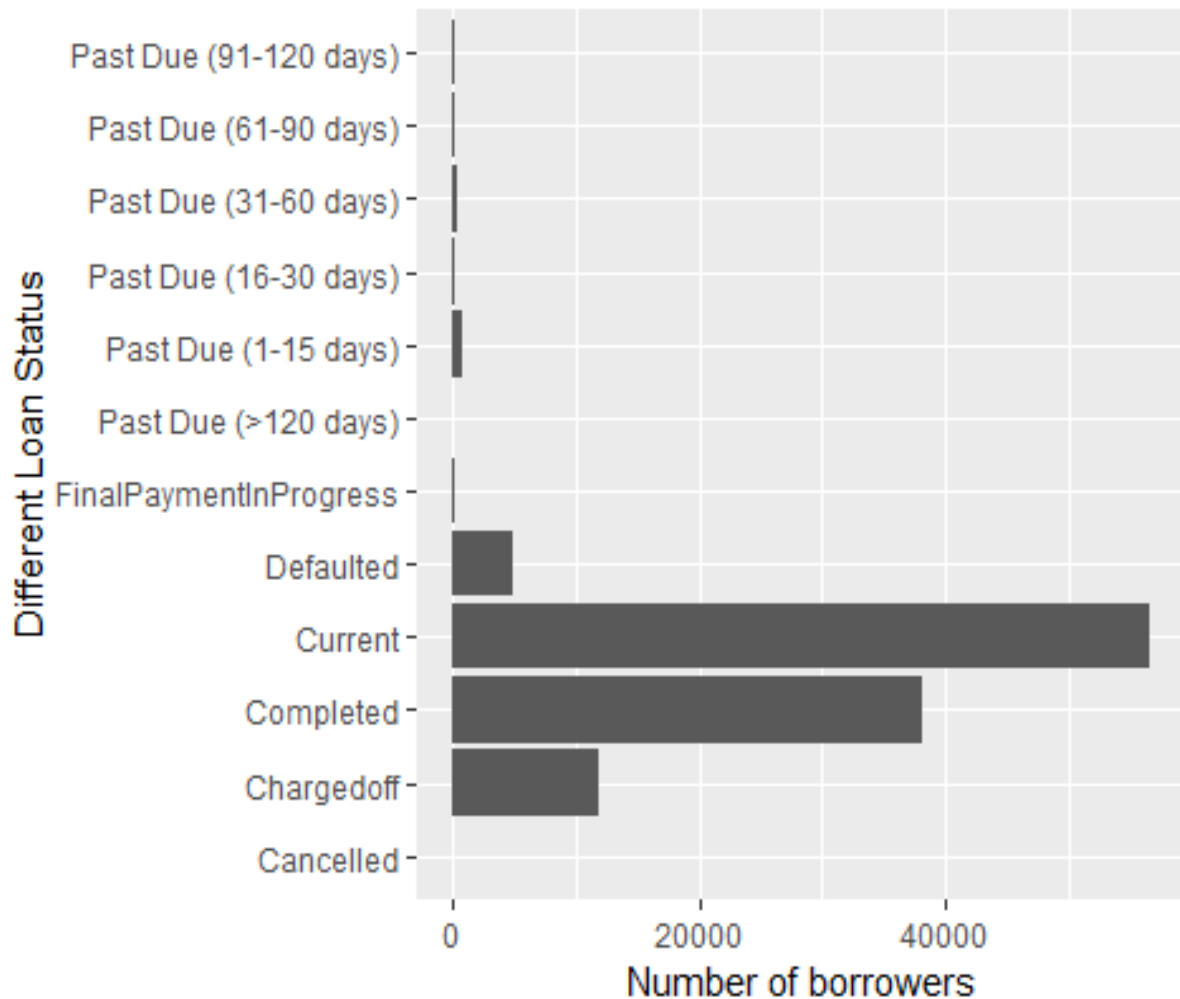


Here, I plotted the bargraph between 'Loan term in years and % of Borrowers'. From this graph, what I get was unusual trend because for loan status of 'completed', I saw a trend but not in loan status of 'defaulted' though. Now just because some customer's loan status is not Completed, doesn't mean his/her loan status is Defaulted. The reason why I explored this is because I wanted to find if BANKS SHOULD FOCUS ON CUSTOMERS OPTING LOANS FOR SMALLER TERMS OR NOT? For this getting data of only two loan status won't be enough. So I divided the customers into two groups-

1. Good Customer
2. Bad Customer

But, first I checked the distribution of Loan status variable by plotting bar graph between "Different Loan status & Number of Borrowers" to get more insights. In this, what I observed was loan status for 'completed' is higher than loan stats for 'defaulted'.

Distribution of Loan Status variable



Then , I plotted the graph between two types of customers- Good and Bad (between ' % of Borrowers and Loan term in years')

The bigger Picture

Good vs Bad

Number of good borrowes vs bad borrowers
in each loan Term

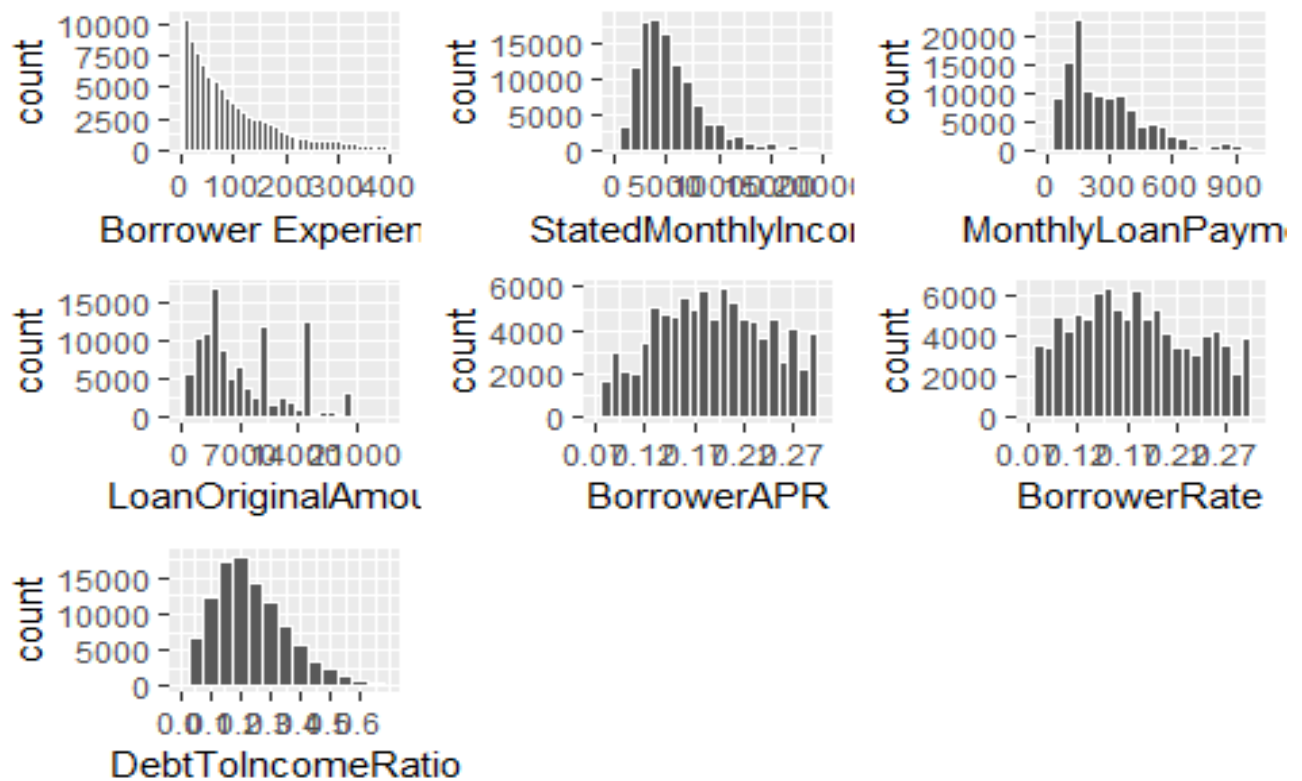


What I observed from this plot was that that the short and long term prospects of the banks who issues the loan seems good because both in 1 and 5 years category, I saw more number of good customers as compared to the 3 years. However , I also noticed the number of good customers is decreased a little from 1 year to 5 years but as compared to 3 years it is nothing. This may mean that banks should focus in long and short term customers as compared to medium term customers.

I've also analysed the distribution of some of categorical features that might be useful for this dataset. I removed some of the outliers to get better understanding of the features.

Distribution for some continuous/categorical features

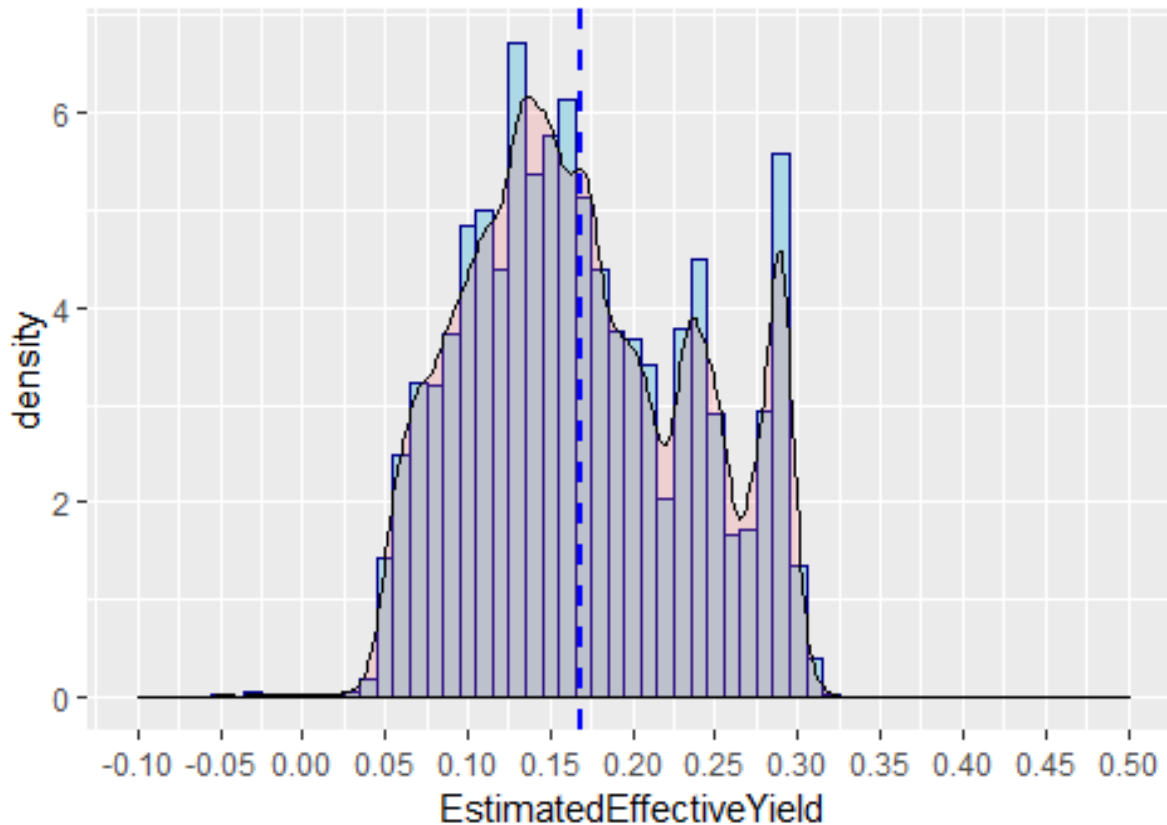
Distribution of feature variables



After that I plot the bar graph between 'density and Estimated Effective Yield(EEY) to understand this distribution'.

EstimatedEffectiveYield - A better measure for a successful Lender

Distribution of EstimatedEffectiveYield with the mean axis



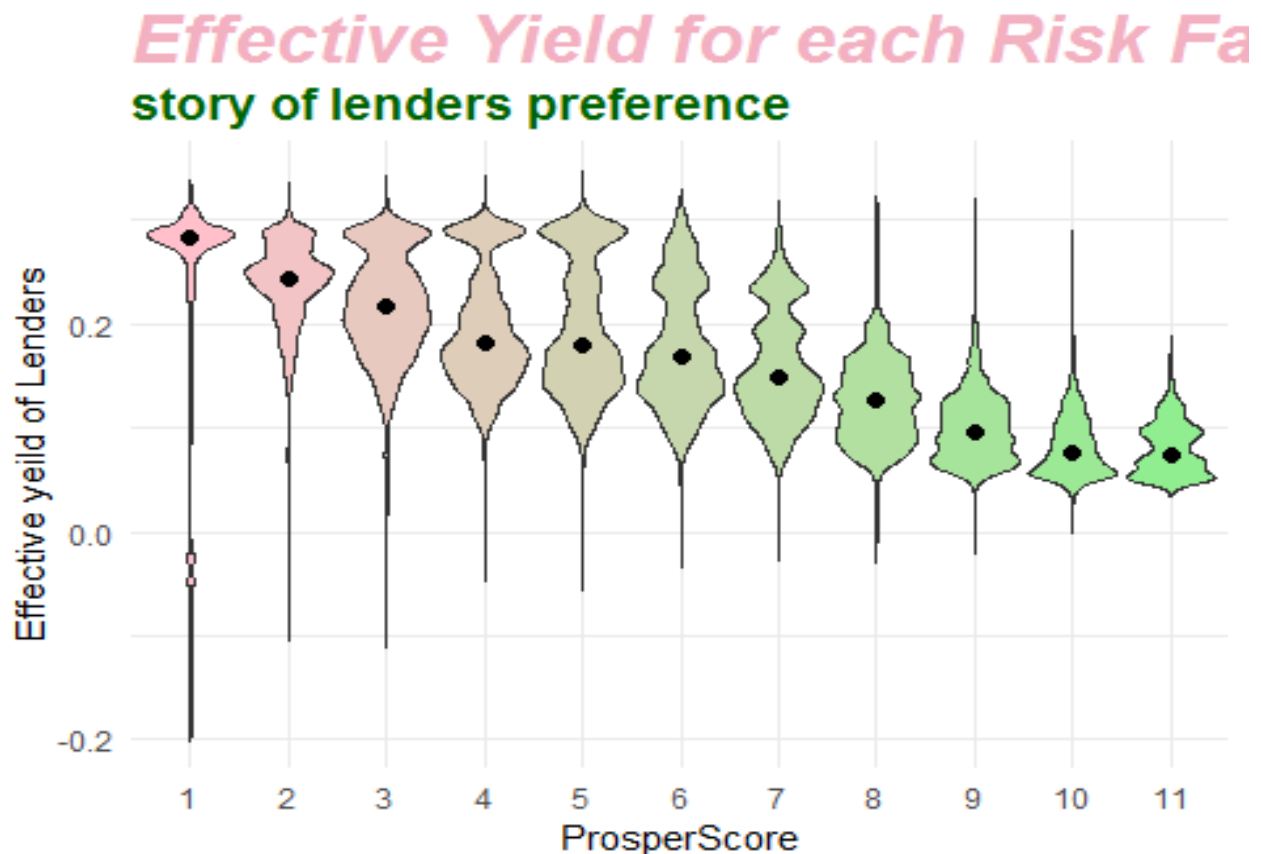
EstimatedEffectiveYield(EEY) is said to be better estimate for the lenders than the interest rate because the interest includes *processing fees, uncollected interest due to borrower being chargedoff*. Plus it also doesn't include *late fines*.

Hence, EstimatedEffectiveYield takes account for all these things and it is thus a better measure. From plotted graph, what I observed the distribution of EEY and saw that it is multimodal(*having several modes or maxima*). I saw that the most popular EstimatedEffectiveYield was around 0.3 while the mean was around 0.17 represented by the blue dotted line.

The multimodal pattern shows that there are multiple EstimatedEffectiveYield that is popular. Strangely I also noticed that the some customers have negative EstimatedEffectiveYield. This may mean a lot of other things like ,their BorrowerRate is much lower than their *service fee rate* or these customer's *uncollected interest on chargeoff* is lot more or they just never payed the late fee and payed back the loans along with the interest always on time.

Does Lenders prefer borrowers with better ProsperScore?

For this, I plot the violin plot to represent the continuous distribution. `geom_violin` (it's a blend of `geom_boxplot` and `geom_density`). The graph is between 'Effective yield of lenders and ProsperScore'. I plot this graph to have a better understanding on a question like, "if lenders get more EY, if they have a better ProsperScore."



I saw a wonderful trend here. Here more score for the risk factor means better the borrower and lesser score for risk factor means poor prospects from the borrowers.

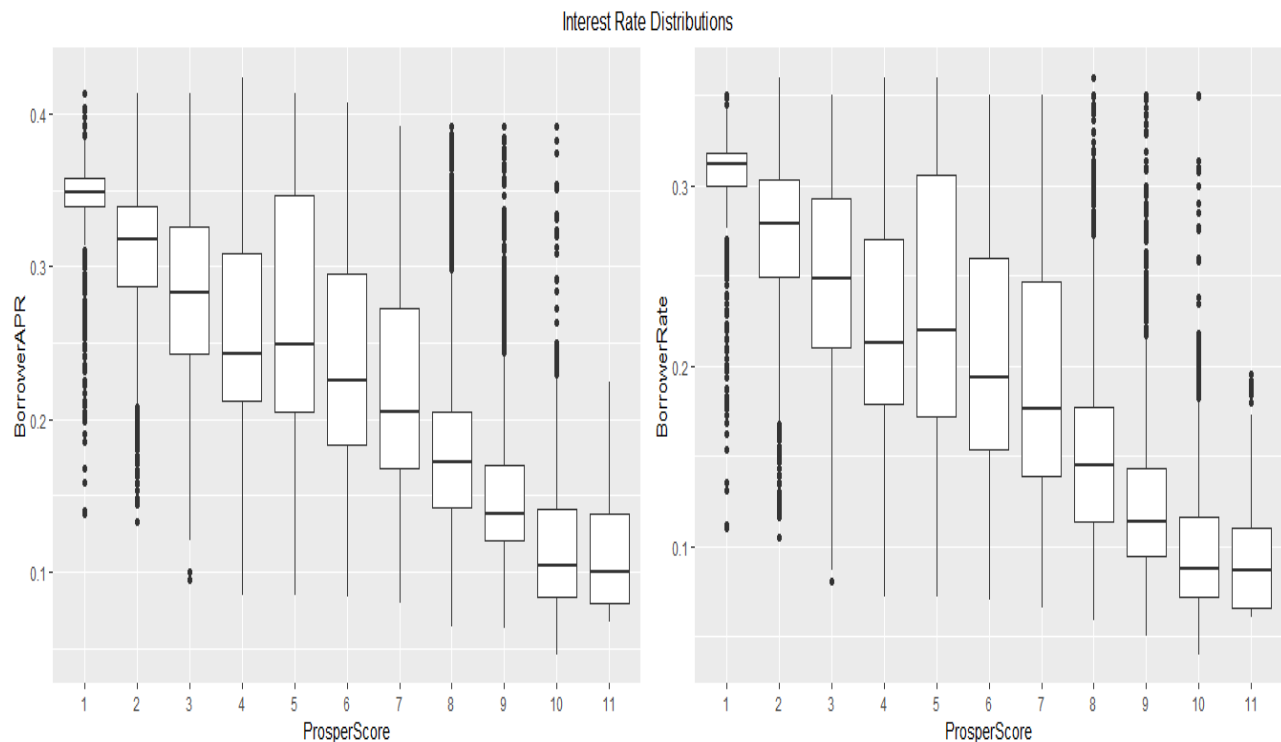
What I saw that for lower ProsperScore distribution of effective yield is a lot more than the higher ProsperScore. This may mean that lenders charge a variety of interest rate from the borrower with poor prospects as compared to borrowers with better prospect. We can also notice how median (*represented by the black dot*) is decreasing as ProsperScore is increasing. This may mean that lenders give more relaxations to borrowers with better ratings as compared to borrowers with poor rating.

I explored more to further analyse if lenders trust and like borrowers with better ProsperScore, since Estimated Effective Yield includes more things such as late fee and

doesn't include processing fee and others. So, more EstimatedEffectiveYield for lesser ProsperScore borrowers may be due to high late fines because lesser ProsperScore borrowers are more prone to fail to repay their loan on time each month.

To see if borrower's interest rate shows the same trend for each ProsperScore categories or not, because interest rates don't include late fines. I plotted the box-plot for Interest rate distributions between 'BorrowerAPR & ProsperScore' and between 'BorrowerRate & ProsperScore'.

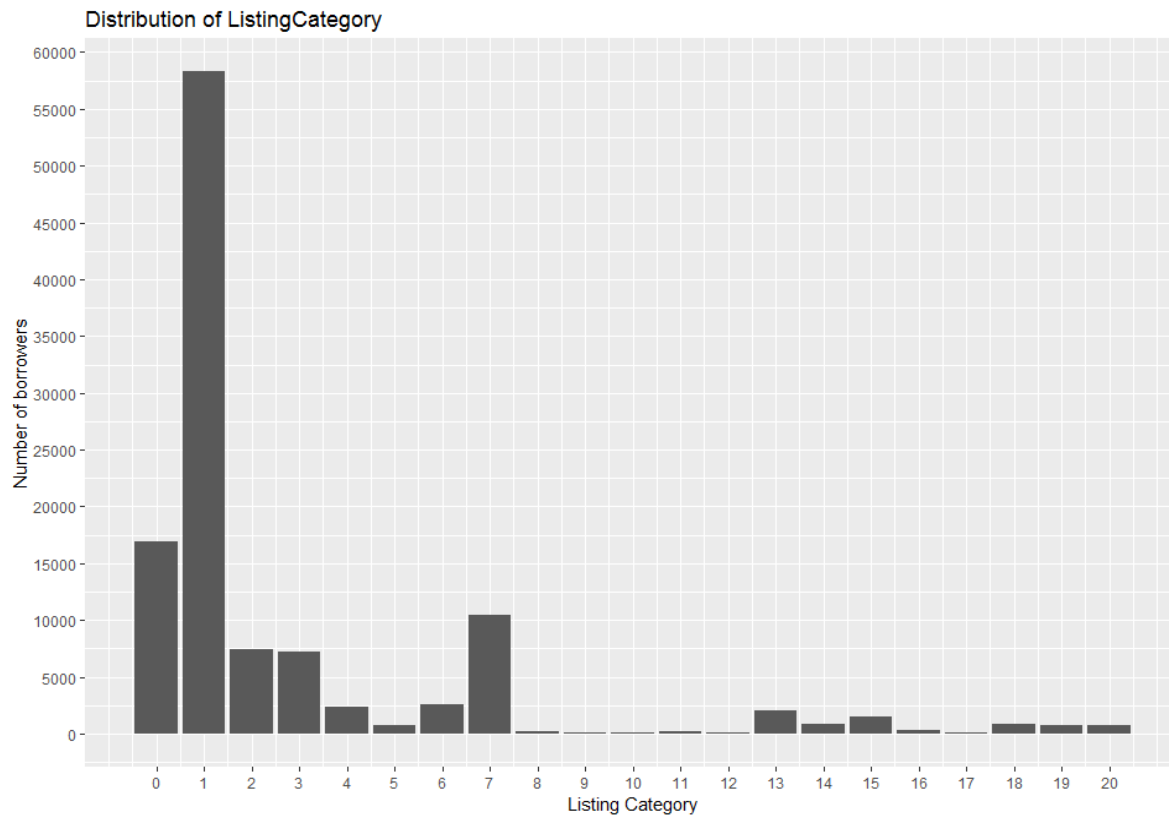
We got an interesting trend



From this boxplot, what I observed that for both BorrowerAPR and BorrowerRate, which are metric for interest rates. I saw a declining trend as the ProsperScore increased, that justifies the fact even more that lenders somehow prefers to charge less for all the borrowers with better ProsperScore rather than with lower score ones.

The whole idea behind this analysis on the ProsperLoan data was to answer the most important thing for a loan ? And its **BORROWERS** .

Distribution of Listing Category



In this, we plot the barchart between 'Number of borrowers and Listing category'. I observed after section indicated by number 7 with more than 10,000 of borrowers remained unknown. The question arises like does it indicate any illegal reason to opt for a loan?

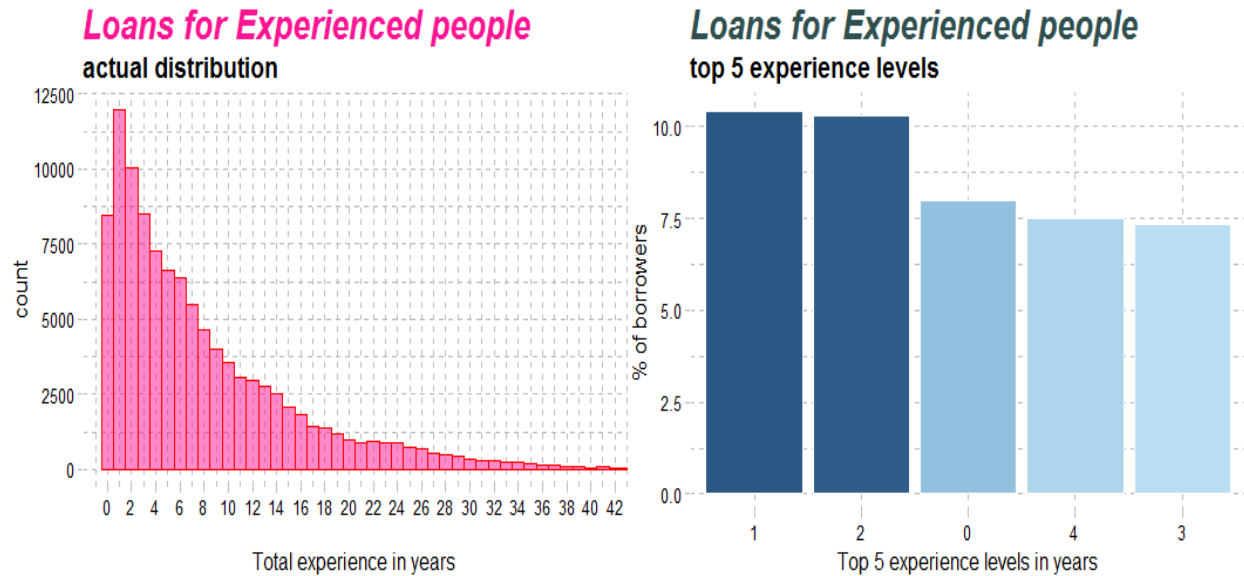
To answer such questions, I did some more research on what the occupation and employment status of borrowers for all the loans falling into the other section.

Who are others??

What I saw from ProsperLoan dataset , there were some dubious borrowers in the others table, where there were around 352 cases where borrowing reasons, employment status, occupation all are *Others* for *Occupation* and 95 cases where *Occupation* is left blank.

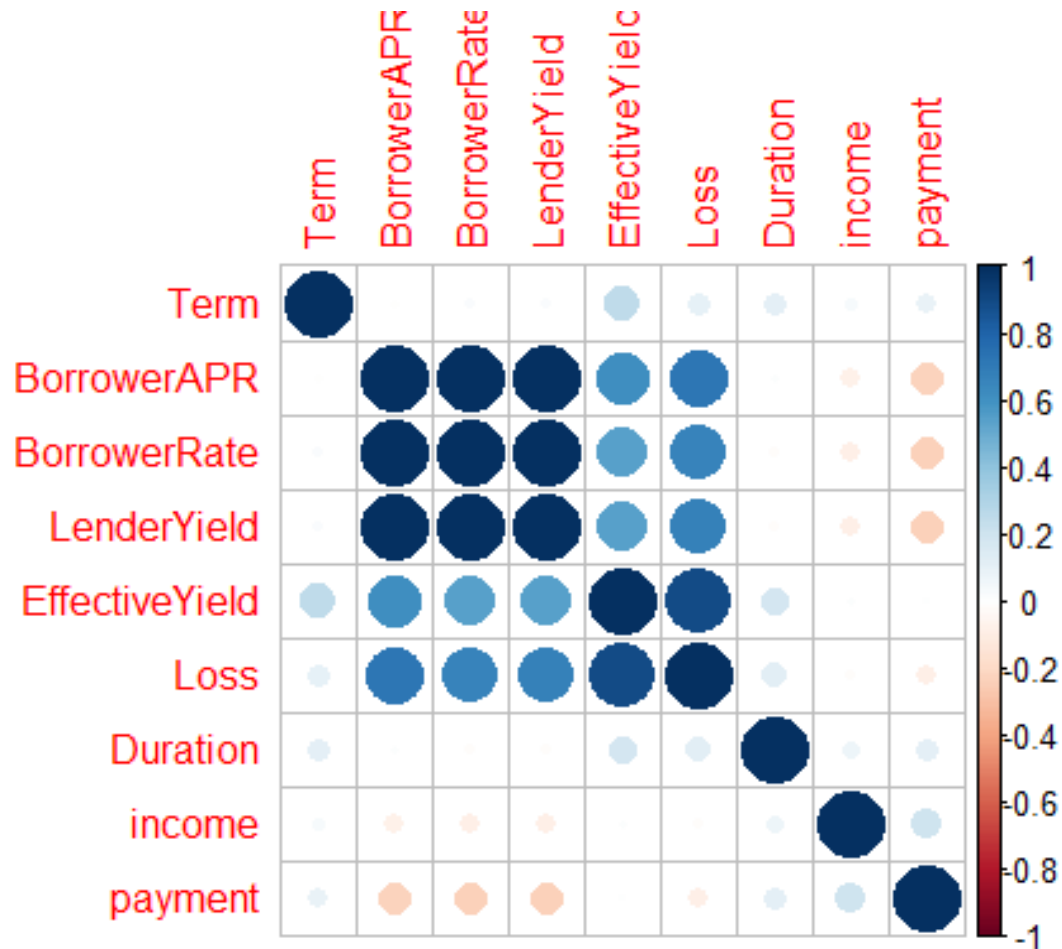
I suppose it's a partial dead end for us to see why people mention *Other* as a reason, while opting for loan. There could be chances that some of the borrower's information be fake too.

Does experienced people opt for loan lesser ?



Well , what I noticed here that my assumption is as people gain experience in their jobs, lesser they opt for loans. This may be due to fact that as people gain experience ,their salary also increases and hence, the lesser the reason they find to opt for loans. I saw that the right histogram of 95% quarterly, most people opt for loans when they've almost 2 years of experience. People with more job experience should have more potential to repay their loan better because they have higher paying jobs and hence their ProsperScore would be higher. And as it can be seen that borrowers with better prosper score pay lesser to the lenders and lenders somehow prefer them.

Other Correlations



I also plotted the graph to check other correlation using corrpplot.

How people take loan for their Homes ?

Here I divided the people for two categories->

1. First those who are opting for loan for renovation of their home (when they have a house).
2. Second those who opt for home loans even though they don't have house.

Home Improvement for All even if you dont have home



Are Investors partial to Borrowers with better rating ??

Here was an interesting trend. I looked for people with ProsperRating of AA, A & B, number of people opted for loan mentioning the reason for loan as **Home Improvement** but still has no home is more than the people who has home and opted for loan for Home Improvement and truly a home owner. But for people with poor rating, the trend is opposite and expected also. People who don't have any house should not get any loan mentioning the loan reason of House Improvement. These whole thing shows not only lenders give much preference to Rating over Verification and KYC (Know Your Customer) but this also shows irrespective of rating of borrowers, lenders don't care much about loan reason as a whole. The following code proves it even more. We can see that there are not a single borrowers who mentioned their loan purpose to be Home Improvement when they didn't have their own house and their loan was not approved. That's a strong evidence to show that investors don't care much about how much borrowers fudge or fake their loan purpose.

```
a <- loanData %>%  
  filter(ListingCategory..numeric. == 2,  
         LoanStatus == 'Cancelled')  
print.numeric_version(dim(a)[0])
```

SUMMARY

- In most of the cases people take loans for their own personal needs. In 10% of the case people borrow loans to buy automotives.
- Loans are borrowed to make payments.
- Some loans are taken for Home Improvement.
- Remaining loans are taken of other specific reasons mentioned in the graph.

Thus , an analysis has been made on the Prosper Loan Data Set. We saw on what basis the bank grants loans, the occupation of customers, the reason why they need loans and when the banks make a profit out of it.