

Prosper Loan Data Analysis

Mayukh Sarkar

Introduction

This dataset is financial dataset and this is related to the loan, borrowers, lenders, interest rates and stuffs like that. Prosper or Prosper Marketplace Inc. is a San Francisco, California based company specializing in loans at low interest rates to the borrowers. In this dataset, we are using the data from the Posper to analyse it and trying to find the pattern in the Prosper data. This may be tedious because of the sheer size of the dataset and the complicated nature of all the financial datasets. We are using R, an advanced high level programming language of the analysis with some of its most popular graphic package `ggplot`. The codes are not visible in the HTML/PDF export for the simlicity but the codes can be reviewed from the RMD file.

The dataset is comprised of 81 variables and contains 113937 entries. The variable that are explored in the dataset are the following Term : Amount of month customers opted for loan

LoanStatus : Current status of the loan like chargedoff, completed, defaunted etc...

EstimatedEffectiveYield : Yield of lenders from borrowers minus the processing fee and late fines

ProsperScore : Risk Factor score from 1 to 10. 10 being least risky

BorrowerAPR : The Borrower's Annual Percentage Rate (APR) for the loan.

BorrowerRate : The Borrower's interest rate for this loan.

ListingCategory..numeric. : Prosper rating for borrowers in numbers

EmploymentStatus : Current type of employment

Occupation : Occupation of borrower at the time of listing

EmploymentStatusDuration : How long the employee has been employed

IsBorrowerHomeowner : Does the borrower owns house at the time of listing (True & False)

ProsperRating..Alpha. : Prosper rating for borrowers in alphabets

IncomeVerifiable : If the income of the borrower is verifiable at the time of listing (True & False)

StatedMonthlyIncome : Monthly income of the borrower

MonthlyLoanPayment : Monthly loan payment amount

Recommendations : Recommendations the borrowers has at the time of listing

DebtToIncomeRatio : The debt to income ratio of the borrower at the time the credit profile was pulled.

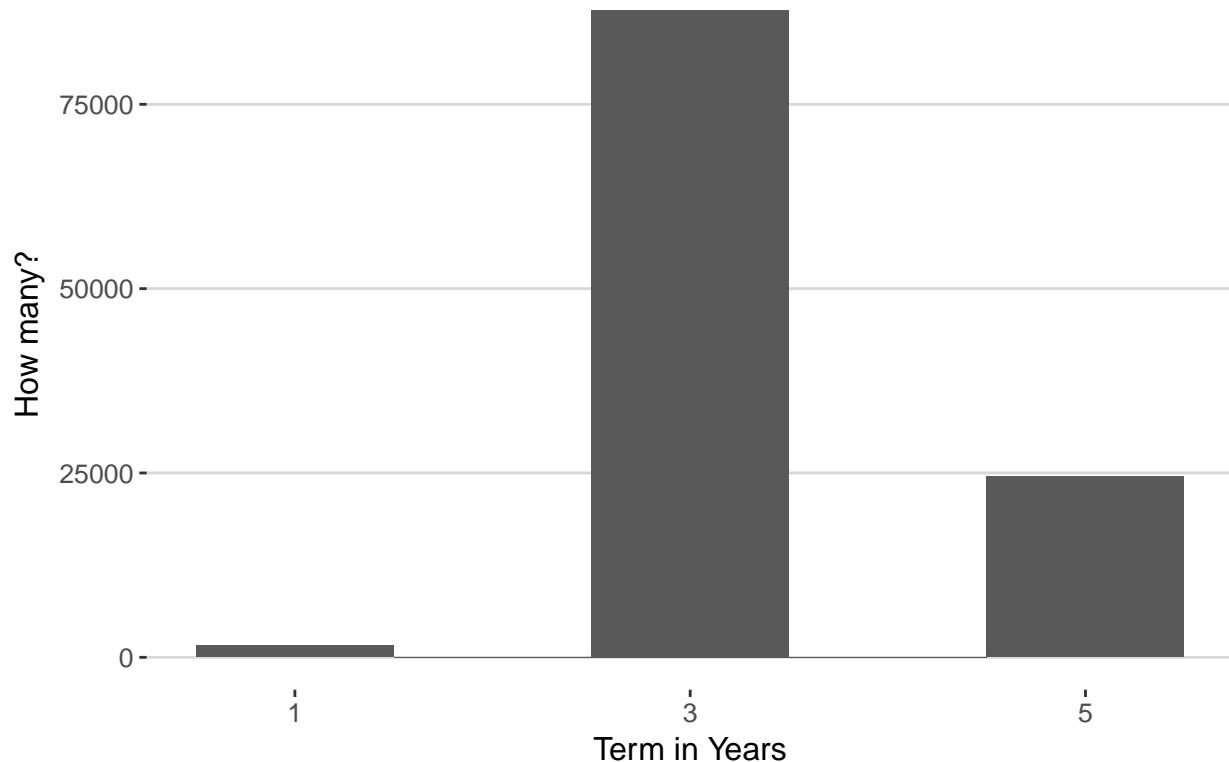
LoanOriginalAmount : Original amount of the loan

LoanOriginationQuarter : Quarter of the month when loan was originated

The first Qustion

The first question that needs to be asked is HOW LONG PEOPLE USUALLY OPT FOR LOAN? Let's answer this question with a histogram

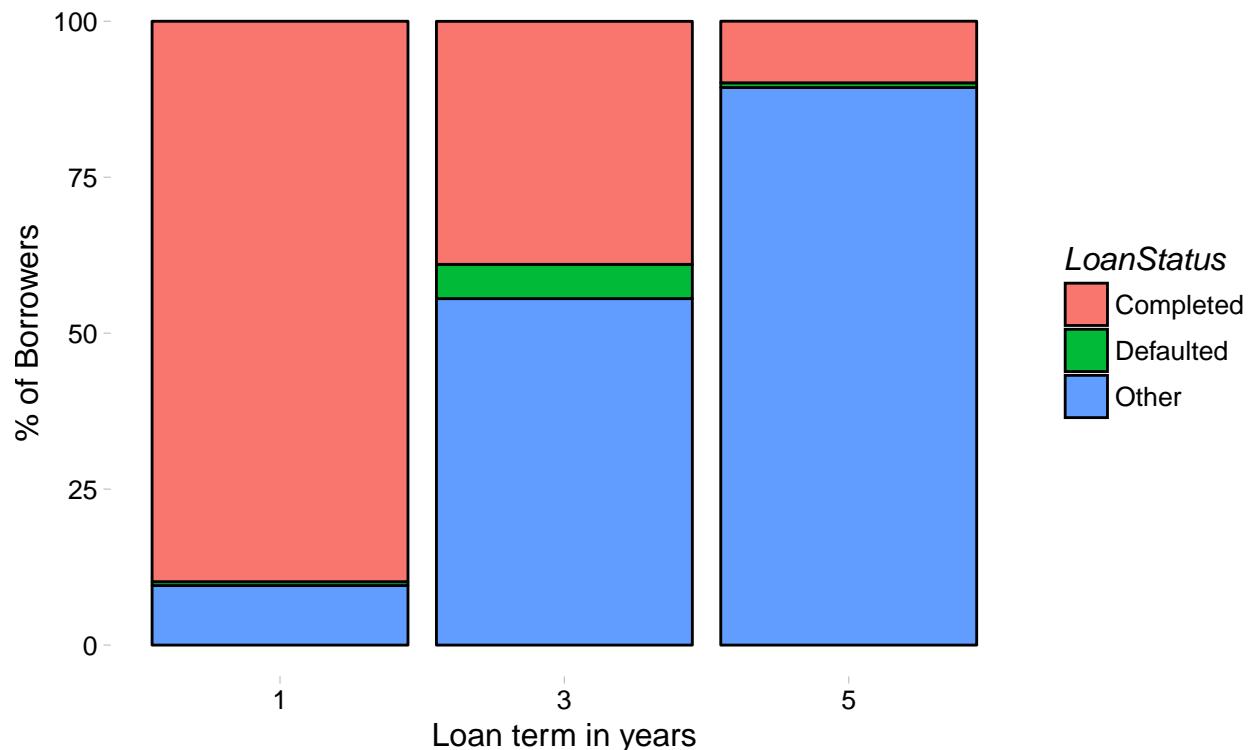
Distribution Loan Terms



We can see that people don't really loan any amount for less than one year and the most popular loan amount is of 3 years although some people do choose for 5 years. Now lets assume something and check if it is correct or not. 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. But is it true? So the next question is DOES PEOPLE REPAY THE LOANS BETTER WHEN THEY ARE GIVEN MORE TIME?

Plotting the trend of different customer types

LoanStatus: Completed vs Defaulted for each loan Term



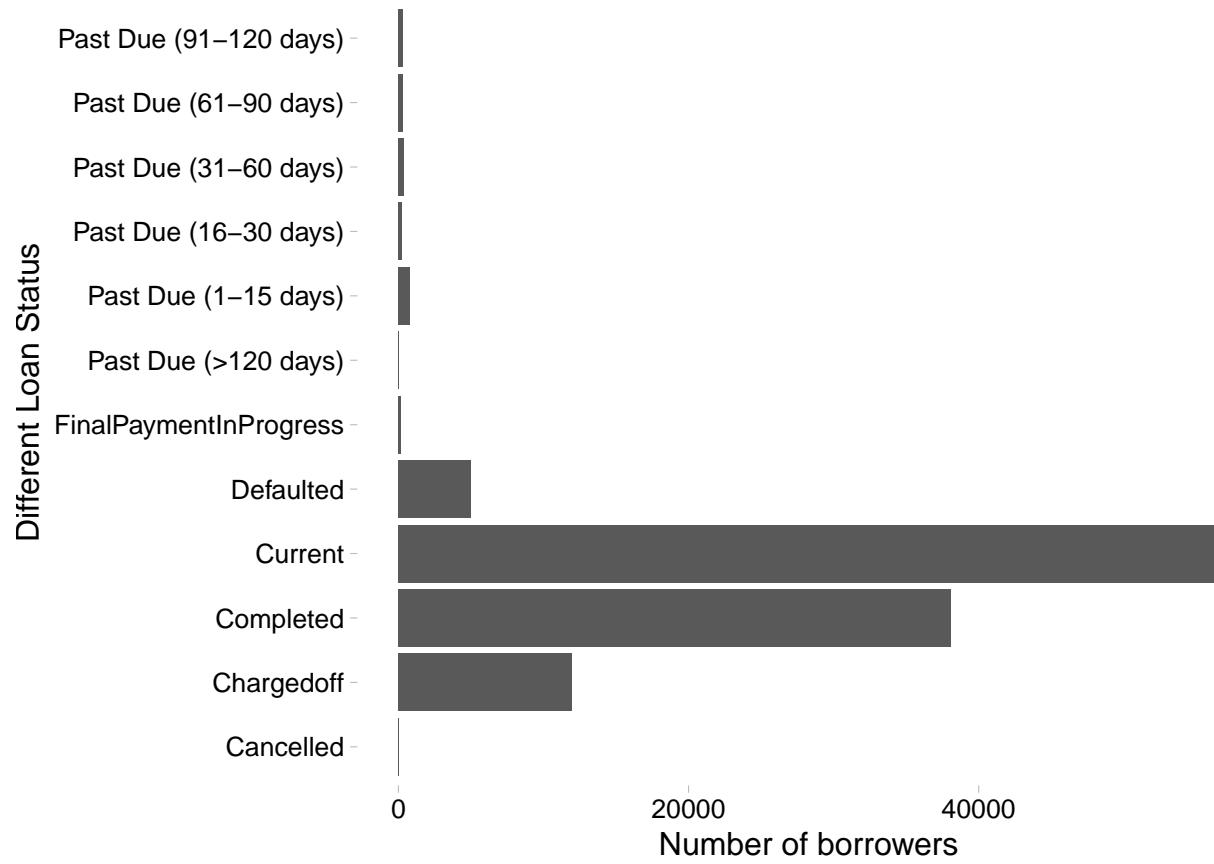
That's unusual because for **LoanStatus** of *Completed*, we see a trend but not in **LoanStatus** of *Defaulted* though. Now just because some customer's loan status is not *Completed*, doesn't mean his/her loan status is *Defaulted*.

We can see that there is a trend in the of **LoanStatus** of *Completed* and *Other* but we see any trend in *Defaulted* though. Now just because some customer's loan status is not *Completed*, doesn't mean his/her loan status is *Defaulted*. So let's explore the *Defaulted* section more.

The reason we are exploring this is because we want to find if BANKS SHOULD FOCUS ON CUSTOMERS OPTING LOANS FOR SMALLER TERMS OR NOT? AND HOW SOME OF THE MOST COMMON BORROWER LOANSTATUS VARIES FOR DIFFERENT LOAN TERMS? For this getting data of only two loan status won't be enough. So let's divide the customers into two groups

1. Good Customer
2. Bad Customer

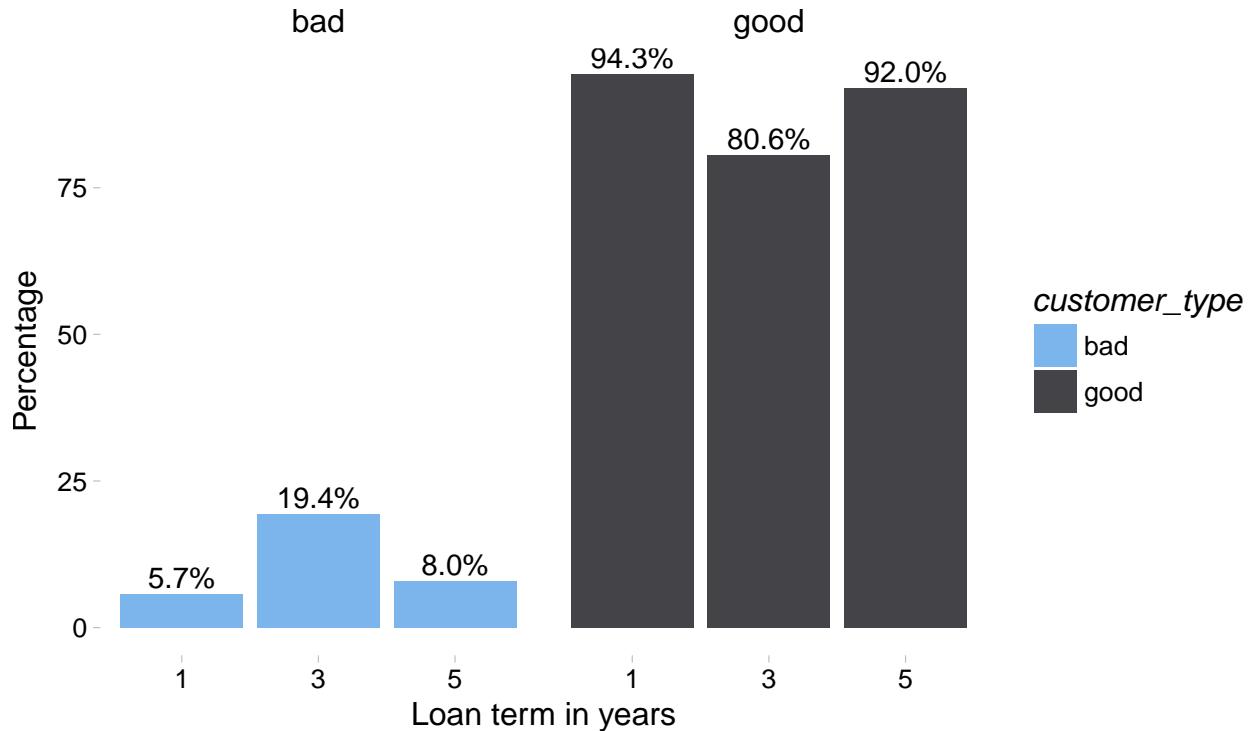
But before that lets see the distribution of LonaStatus variable



The bigger Picture

Good vs Bad

Number of good borrows vs bad borrowers
in each loan Term

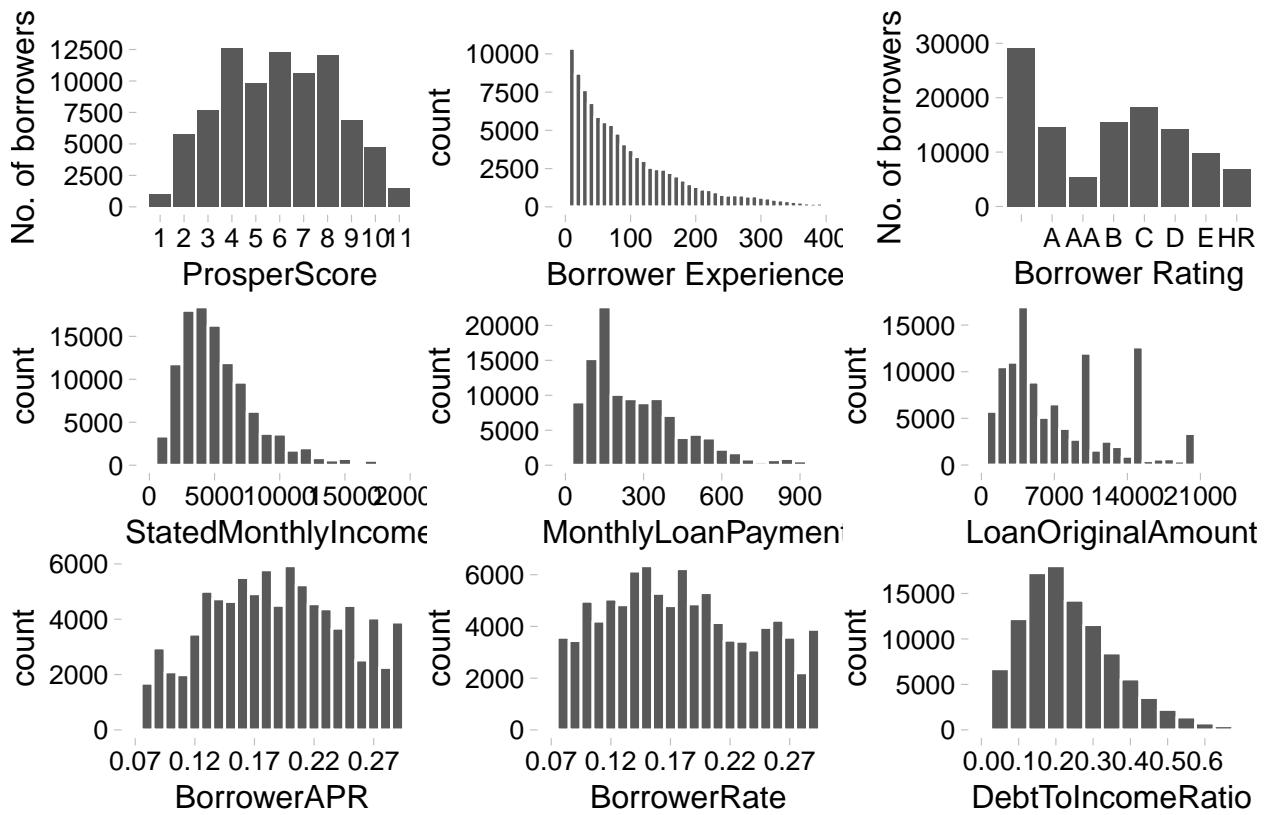


As we can see that the short and long term prospects of the banks who issues the loan seems good because both in 1 and 5 year category, we see more number of good customers as compared to the 3 years. The number of good customers is however decreased a little from 1 year to 5 years but as compared to 3 years it is nothing. This may mean that banks should focus in long and short term customers as compared to medium term customers. But we should take this fact with a grain of salt because there might be other factors that may affect this trend. Only further exploration of the data would reveal that.

But before we move ahead, lets analyse the distribution of some of the features that might be useful for us in this data set. We would also remove some the outliers to better understand the features and may dive into some descriptive statistics too.

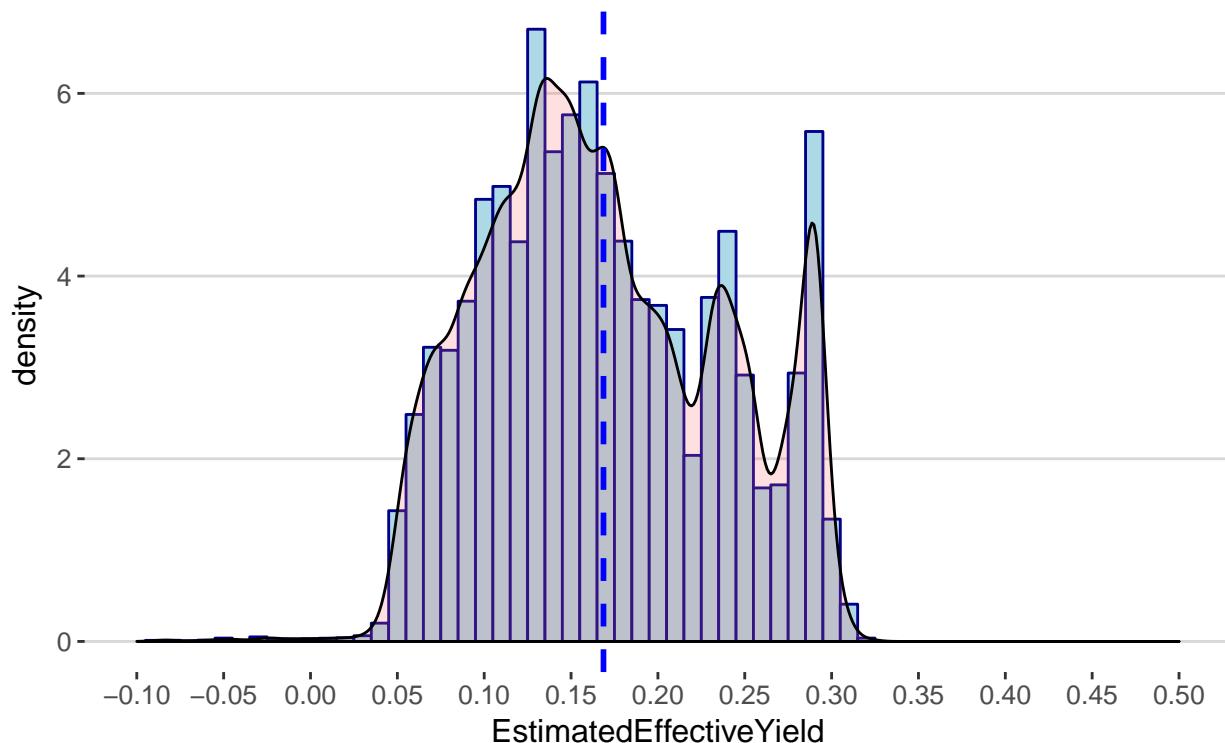
Distribution for some continuous/categorical features

Distribution of feature variables



EstimatedEffectiveYield - A better measure for a successful Lender

Distribution of EstimatedEffectiveYield with the mean axis



EstimatedEffectiveYield 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. Above we are trying to see the distribution of the EstimatedEffectiveYield and we can see that it is multimodal. We see the most popular EstimatedEffectiveYield is around 0.3 while the mean is around 0.17 represented by the blue dotted line. The multimodal pattern shows that there are multiple EstimatedEffectiveYield that is popular. Strangely we can also see that the some customers have negative EstimatedEffectiveYield. This may mean a lot of things. This may mean that their BorrowerRate is a lot 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.

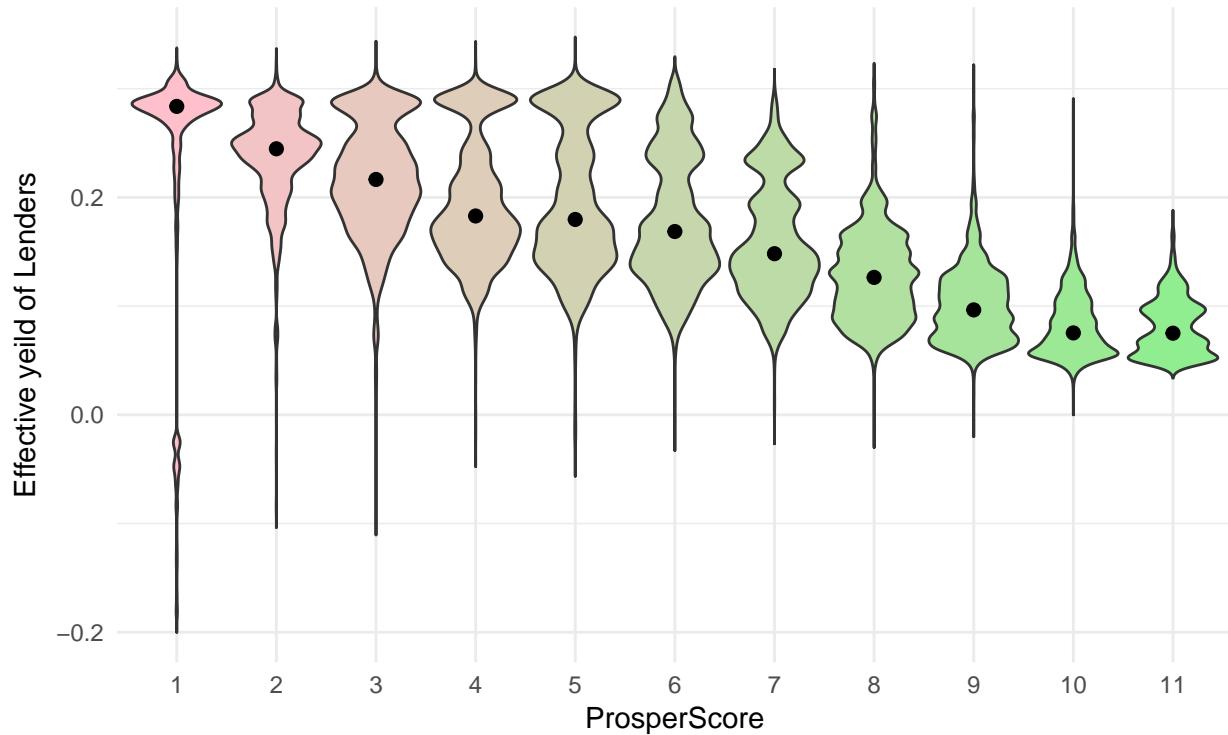
Does Lenders prefer borrowers with better Prosper Score ?

Now lets see what is the distribution of EstimatedEffectiveYield depending on the different **ProsperScore**. This is important because we want to answer a question, i.e., IF LENDERS GET MORE EstimatedEffectiveYield IF THEY HAVE BETTER ProsperScore ?

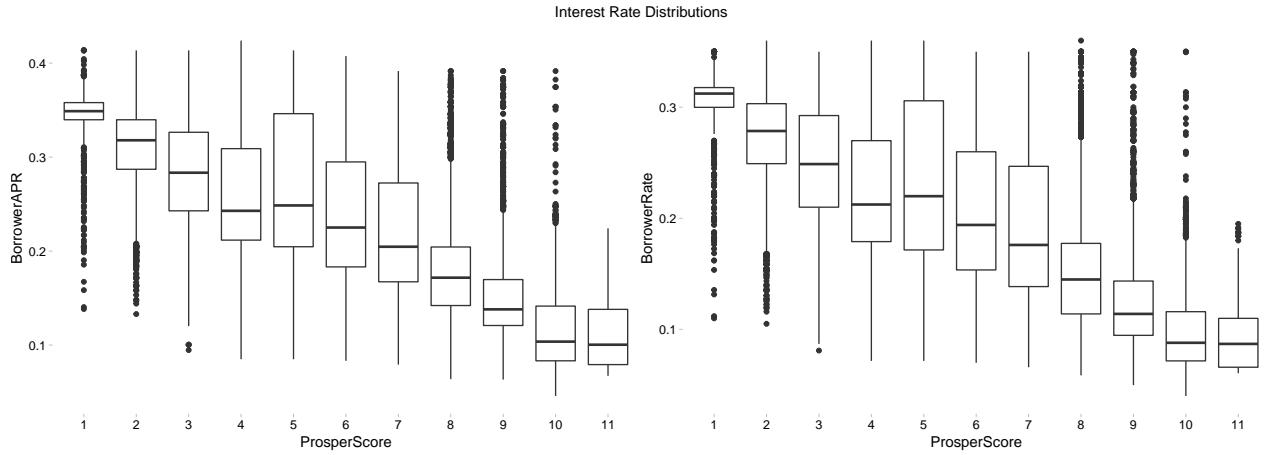
We are using violin plot instead of box plot for this.

Effective Yield for each Risk Factors

story of lenders preference



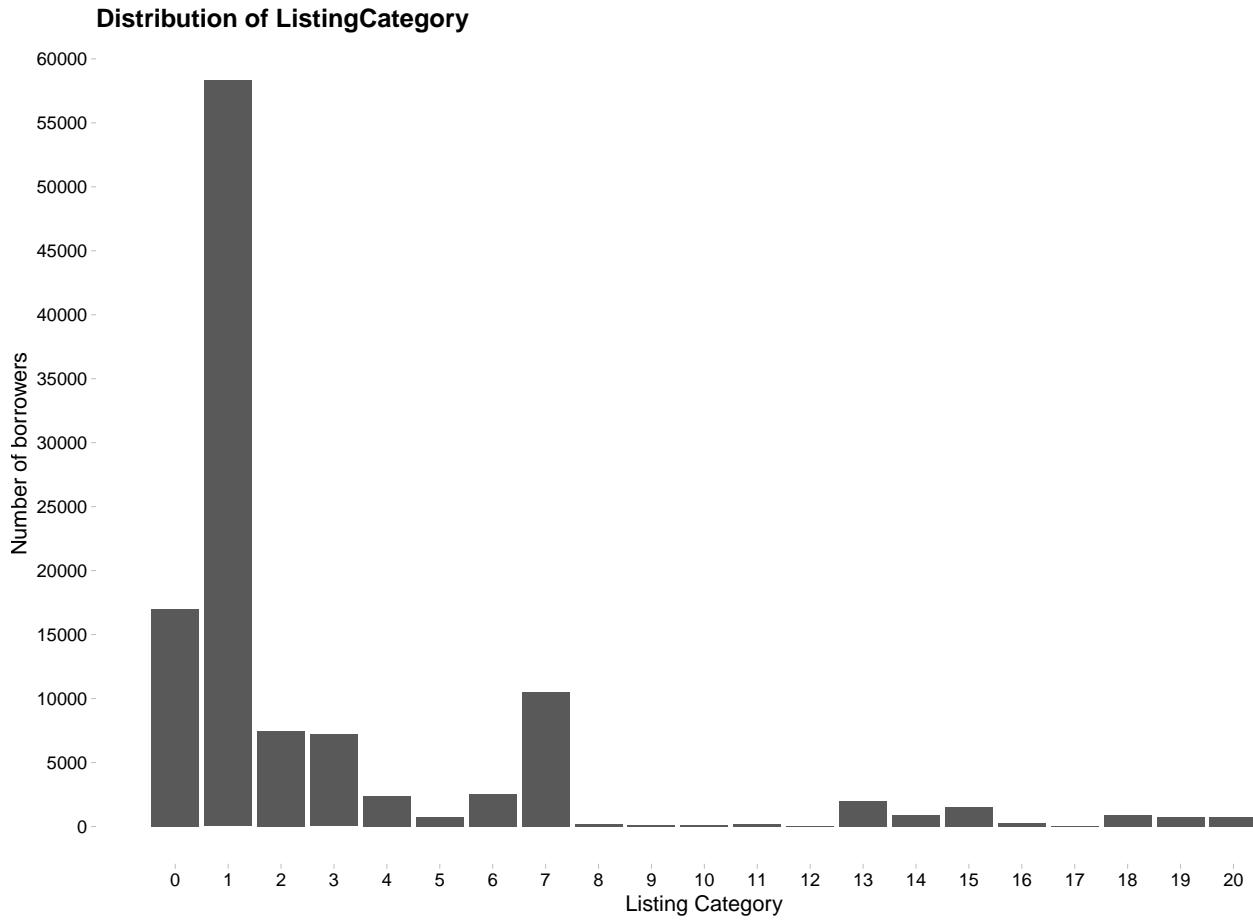
Well this is interesting. We see an wonderful trend here. Here more score for the risk factor means better the borrower and lesser score for risk factor means poor prospects from the borrowers. We can see that for lower ProsperScore distribution of effective yield in a lot more than the higher ProsperScore. This may mean that lenders charges 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. Does that mean lenders trust and like borrowers with better ProsperScore! Let's do a little more analysis to reveal it more. The reason we need more exploration on this is because EstimatedEffectiveYield includes more things such as late fine 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. So, Let's see if borrower's interest rate shows the same trend for each ProsperScore categories or not because interest rates doesn't include late fines.



Finally things are revealed much better now! We can clearly observe that for both *BorrowerAPR* and *BorrowerRate* which are metric for interest rates, we see a declining trend as the *ProsperScore* is increasing. This justifies the fact even more that lenders somehow prefers to charge less for all the borrowers with better *ProsperScore* as compared to borrowers with inferior *ProsperScore*. Here food for thought is that is it moral too? Well I guess no amount of EDA can reveal it.

But are we missing the real question here? We are performing all these analysis on the Prosper loan data to answer several questions but what is the most important thing for a loan ? It's **BORROWERS** and what is the most important question related to borrowers?

Distribution of Listing Category



The other section indicated by number 7 with more than 10000 of the borrowers would remain unknown. Why the loan reason would be unknown? Does it indicate any illegal reason to opt for loan! Let's explore this more. Let's see the Occupation and Employment status of Borrowers for all the loans falling into the *others* section.

Who are others??

We can see some really strange things happening here. In fact there are so many strange things that I have to list them down

Firstly what is the difference between *Employed* and *Full-time*?

Secondly there are too many others. The complete sample of the data is taken for those borrowers whose purpose for borrowing is *Others*. But there are *Others* for *Occupation* and *EmploymentStatus* too.

Thirdly how can someone be *Retired* but still has *Occupation Others* ?

Lastly there are some dubious borrowers in the *Others* table. We can see that there are 352 cases where the borrowing reason, *EmploymentStatus*, *Occupation* all are *Others* and 95 cases where *Occupation* is left blank.

These above points, specially the last one says not all the borrower's information is revealed or in some cases they seemed to be fake too. I guess it is a partial dead end for us to see why people mention *Other* as a reason while opting for loan.

Table 1: Employed

Occupation	n
Other	1579
Professional	789
Administrative Assistant	280
Teacher	256
Computer Programmer	247

Table 2: Full-time

Occupation	n
Other	622
Professional	326
Computer Programmer	139
Teacher	125
Administrative Assistant	115

Table 3: Part-time

Occupation	n
Other	23
Sales - Retail	13
Administrative Assistant	8
Food Service	7
Nurse (RN)	4
Student - College Graduate Student	4
Student - College Senior	4

Table 4: Self-employed

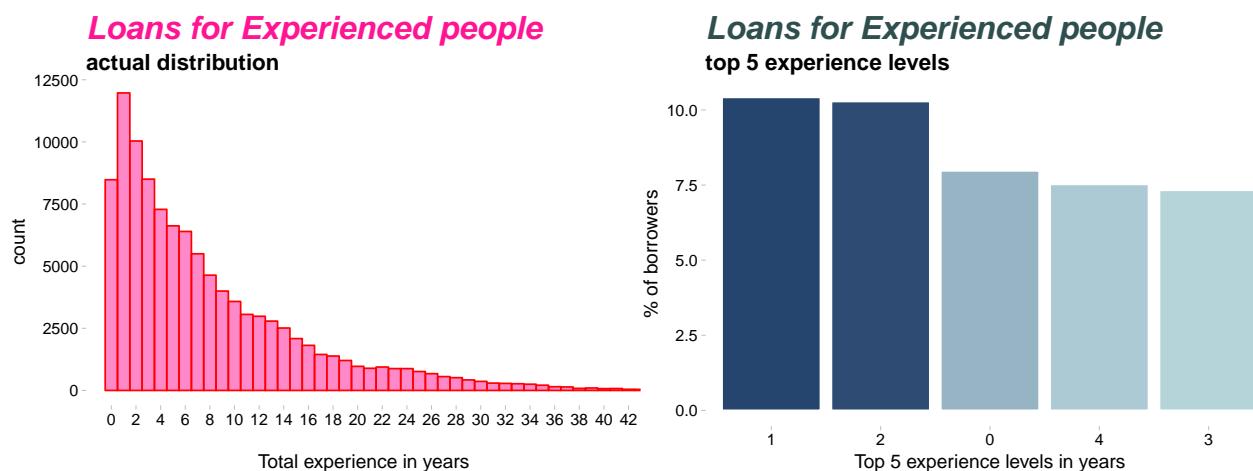
Occupation	n
Other	152
Professional	32
Investor	19
Construction	16
Sales - Commission	16

Table 5: Others

EmploymentStatus	Occupation	n
Other	Other	352
Other		95
Retired	Other	127
Retired	Engineer - Electrical	2

Let's now move into something more computational. Let's first see after working for how many years, people like to opt for a loan? We can then move into some deeper level of analysis.

Does experienced people opt for loan lesser ?



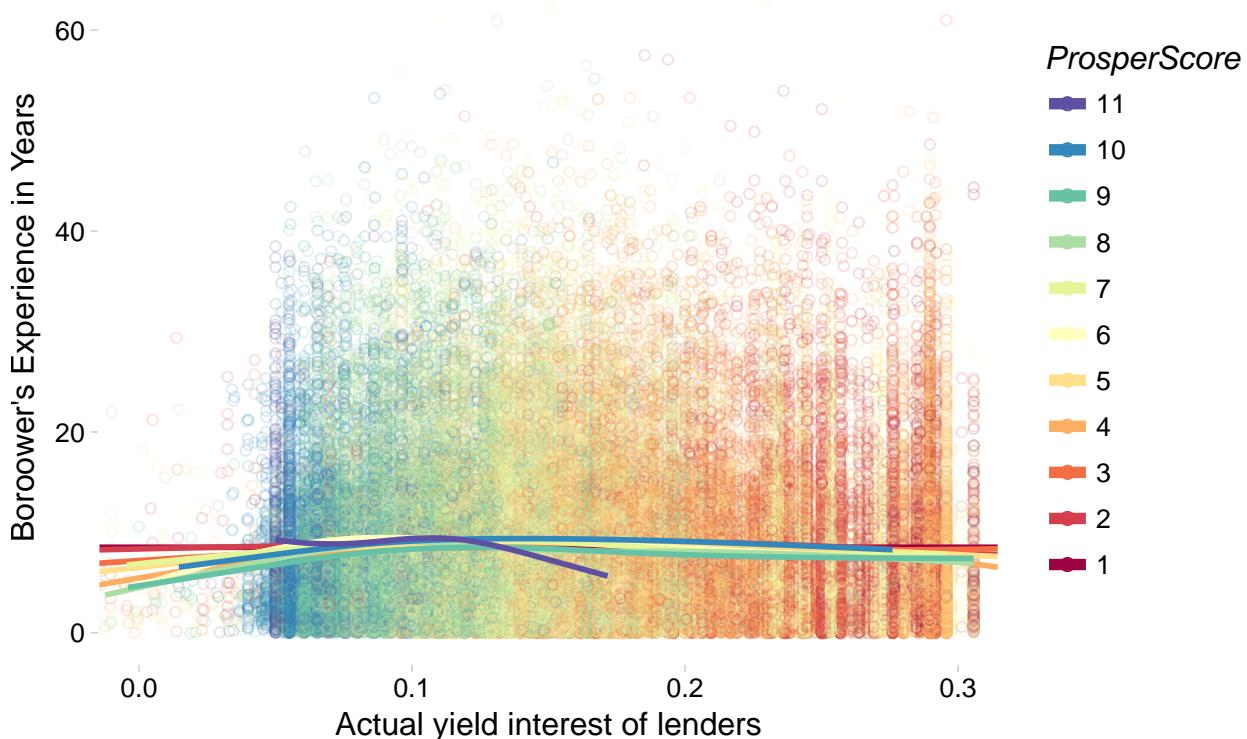
I was right ...)

Well this was really interesting. We can see that our assumption was overall correct because as people gain experience in their jobs, lesser they opt for the loans. This may be due to the fact that as people gain experience their salary also increases and hence the lesser reason they find to opt for loans or the reason can be something different. We can also see in the right histogram of 95% quarterly, most people opt for loans when they have almost 2 years of experience.

Lets explore it even more let's see the correlation between number of month a person has experience and the EstimatedEffectiveYield variable that we have explored earlier. The question that we want to explore - DOES LENDERS ASK FOR LESS INTEREST FORM THE BORROWERS WHO ARE MORE EXPERIENCED? This can be true because 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 we have seen that borrowers with better prosper score pay lesser to the lenders and lenders somehow prefer them.

Correlation of Borrower's Experience

with Actual Yield of Lenders



As it seems from the scatter plot that the pattern seems to have no correlation. It means our assumption was not correct. Borrowers with better **EmploymentStatusDuration** don't seem to get any special relaxation from lenders in terms of interest each month. This can be further confirmed by checking the *Pearson's correlation Coefficient*. We see that the correlation coefficient is almost 0 (-0.0233).

```
##  
## Pearson's product-moment correlation  
##  
## data: EmploymentStatusDuration/12 and EstimatedEffectiveYield  
## t = -6.7924, df = 84832, p-value = 1.11e-11  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.03003902 -0.01658791
```

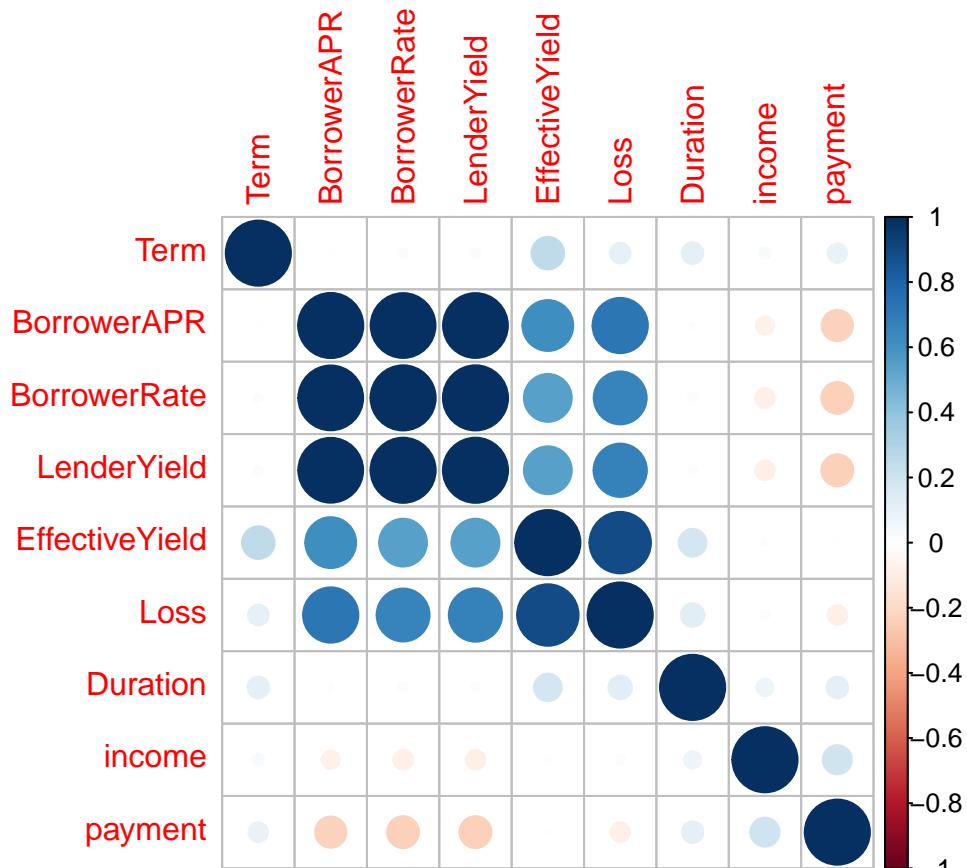
```

## sample estimates:
##          cor
## -0.02331452

```

This also says that even though the true correlation is not true and alternative hypothesis is accepted, there is some **serious statistical evidence of significance**. But if we look into the CI, it is within the range of -0.03 to -0.016 which is very small. Good R value is said a value < -0.3 or value > 0.3 . This value is definitely not that large. Judging from the context latest it is not. So we can say that there is **no practical significance**. Hence we can not tell with any confirmation that More Experienced Lenders end up paying Less/More interest to the Lenders.

Other Correlations



How people loan for their Homes ?

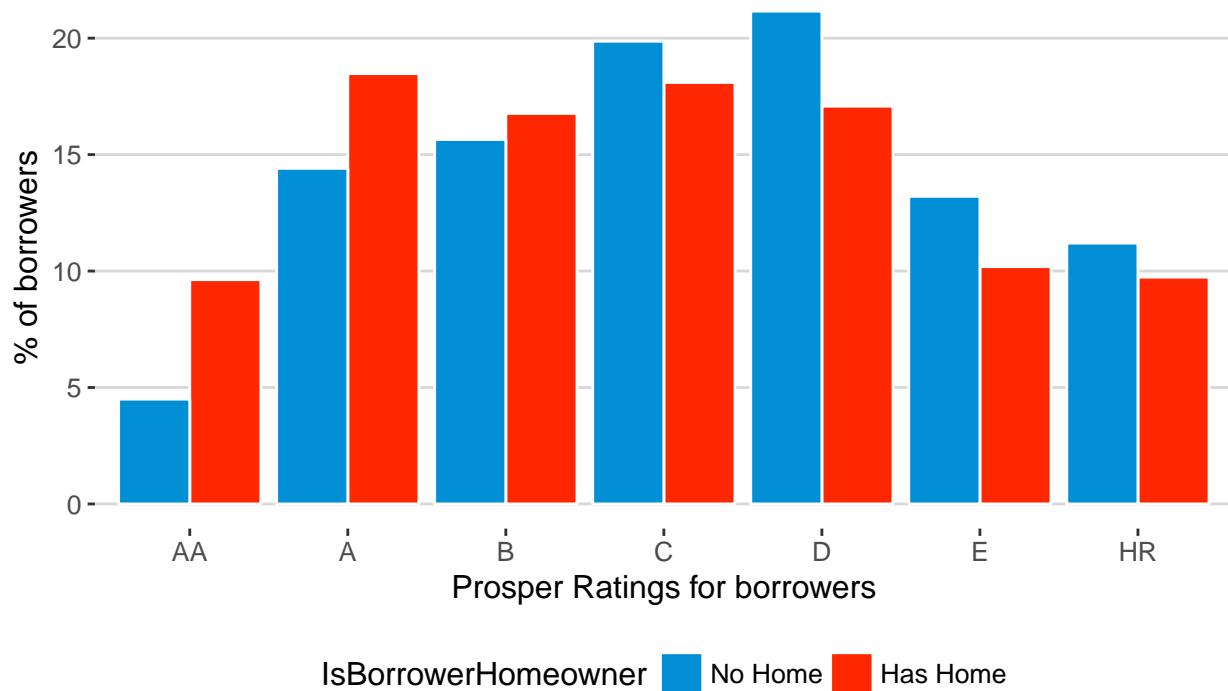
Here we are going to explore the people for two category.

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

Now let's plot the data

Home Improvement for All

even if you dont have home



Are Investors partial to Borrowers with better rating ??

Here we visualizing an interesting trend. We are seeing 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 in 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 font 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.

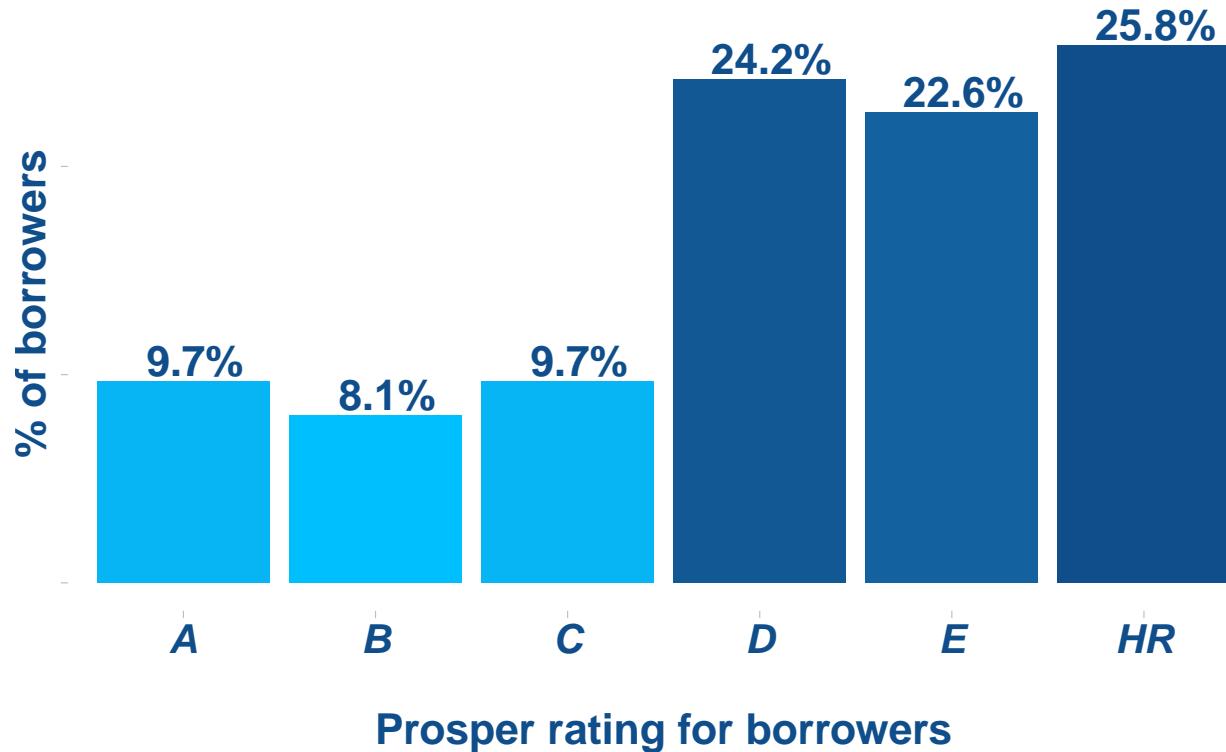
```
a <- loanData %>%
  filter(ListingCategory..numeric. == 2,
        LoanStatus == 'Cancelled')
print.numeric_version(dim(a)[0])

## <0 elements>
```

What's up with the people having dubious income sources ?

Story of Dubious borrowers

by different borrower's category

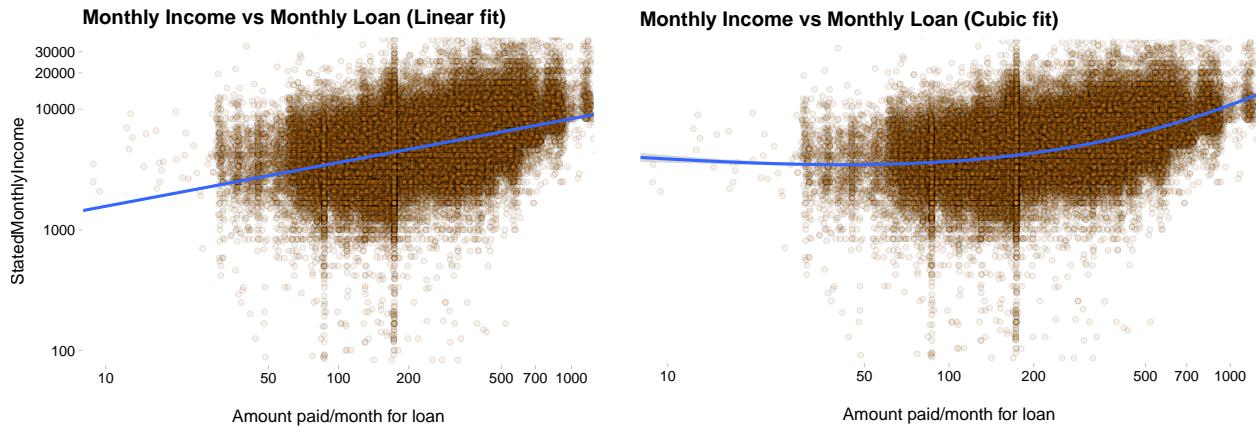


Prosper rating system seems legit !!

Although we can't see any specific pattern among the dubious borrowers but we can see one thing very well. Overall borrowers with better ratings tends to be less dubious than borrowers with poor ratings. One more interesting thing is that the above bar chart does not include borrowers with AA rating. It means that the best borrowers are really best and the prosper rating really works.

Are Lenders greedy ?

I always wanted to do this. But from this dataset we can answer this question a lot of way. One of the way is to check if the lenders asked for money if the borrowers income was high. Let's see if the correlation is substantial.



Here I am trying to see the relationship of *MonthlyLoanPayment* with *StatedMonthlyIncome*. We can see from the above statterplot that most of the *MonthlyLoanPayment* is distributed from 10 to 1000 and the *StatedMonthlyIncome* is distributed from 100 to 30000. Both the scales are transformed in log scale and we can clearly observe two things, namely,

1. There is definite a strong positive correlation between monthly income and monthly loan amount.
2. Cubic relationship seems a better fit than simple linear model.

Are we sure that they are greedy ?

Now we can see that there was definitely a strong correlation between the two variables but are we sure? Let's find the **Correlation Coefficient** to analyse it more.

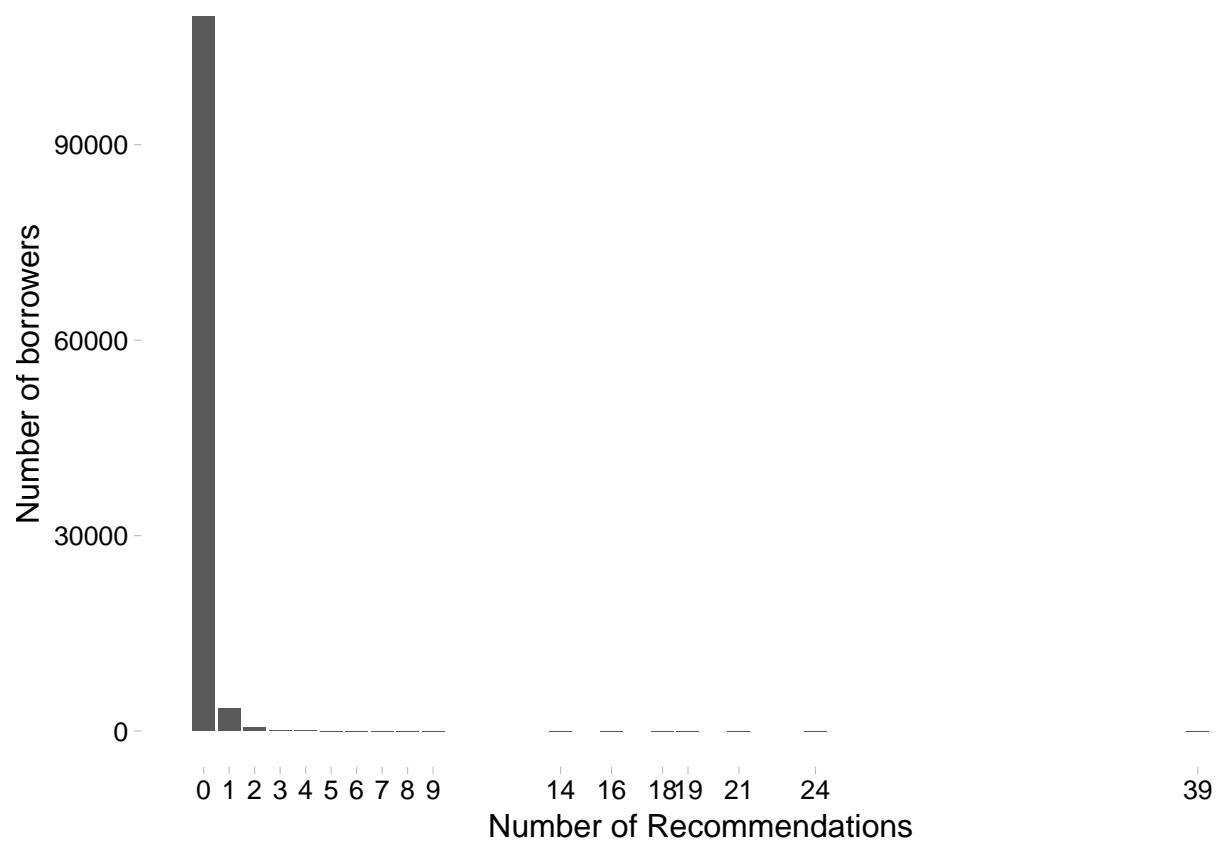
```
## 
## Pearson's product-moment correlation
## 
## data: MonthlyLoanPayment and StatedMonthlyIncome
## t = 67.764, df = 113940, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1912423 0.2024055
## sample estimates:
##        cor
## 0.1968303
```

Well we can't really say that there is a strong correlation looking at the value of R which is almost 0.2. Usually it is said to be of high statistical importance if it is more than 0.3 or less than -0.3. But we can see that the value is still acceptable with somewhat positive correlation with the population Confidence Interval being more than 0. The strong t-statistics of 67.76 and small p-value shows that the statistical significance of alternative hypothesis is very strong. But this can also be somewhat practical significance too, at least to be included in a linear model.

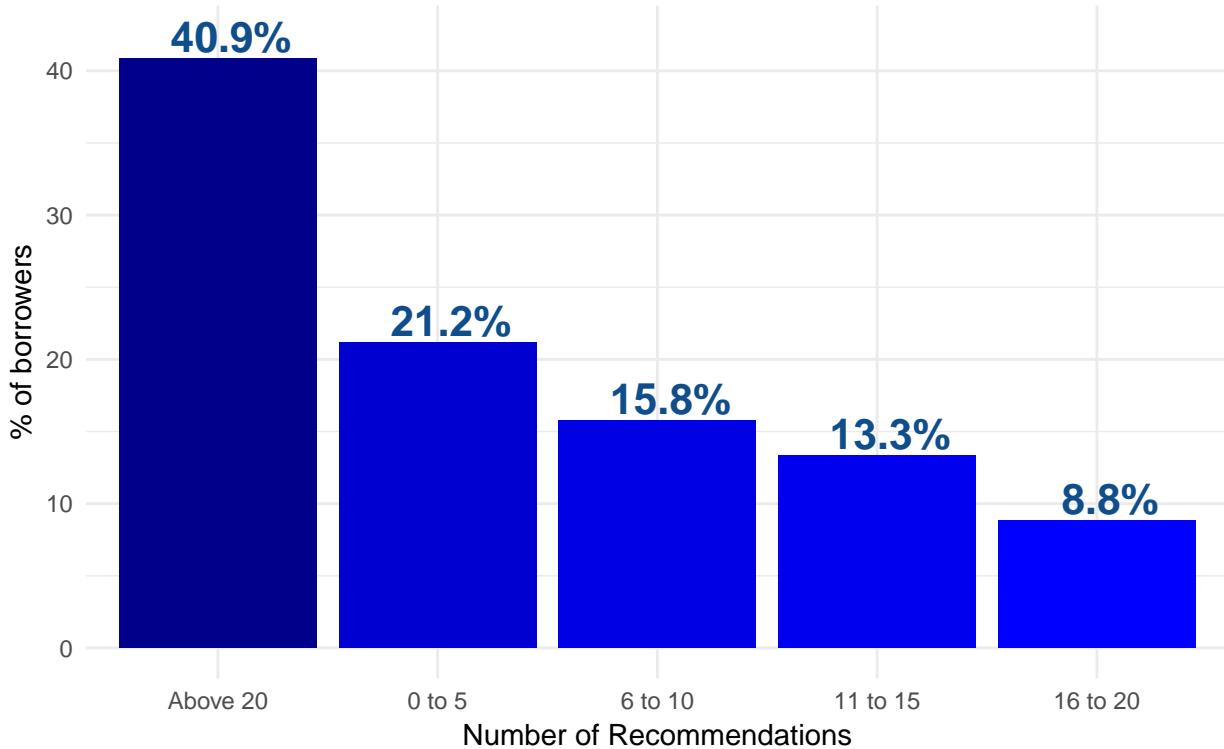
Does Recommendations matter ??

Before we move ahead let's review the *Recommendations* variable and understand it's impact and distribution.

Distribution of Recommendation



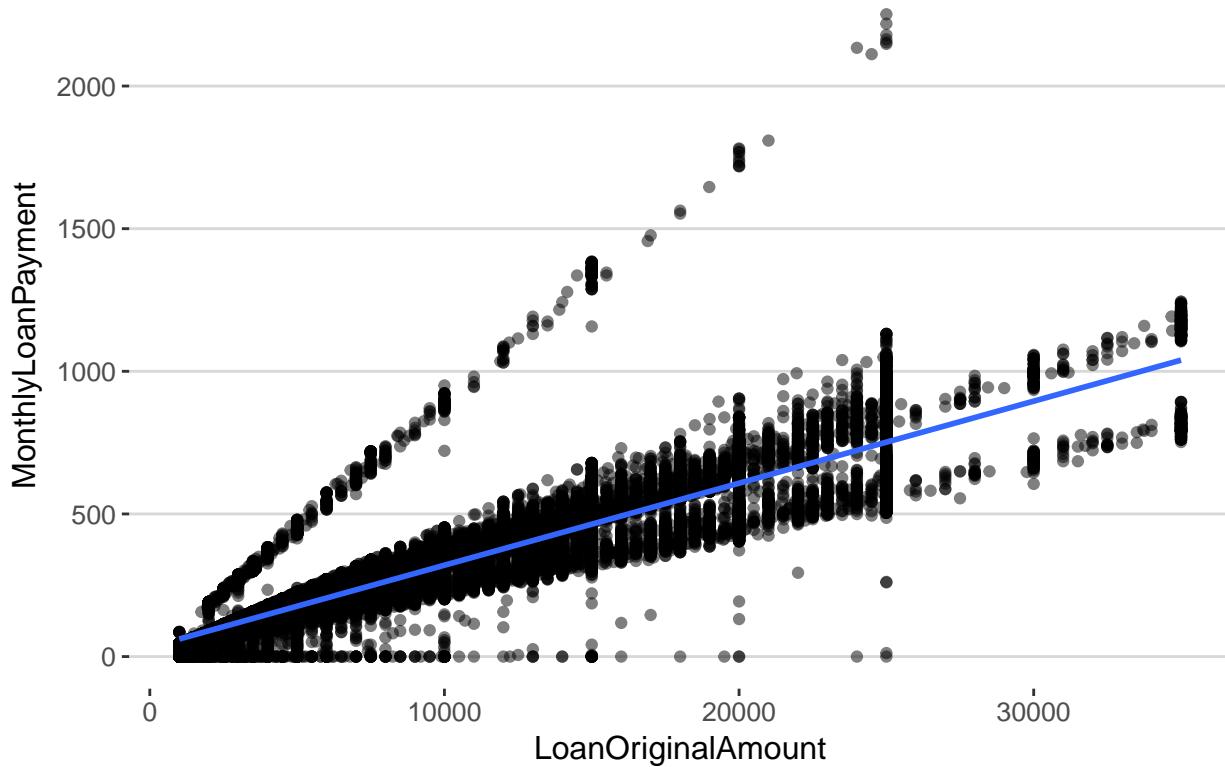
Borrowers with different recommendations in terms of monthly loan payment



As we can see that people with *Above 20* recommendations has the highest number of *MonthlyLoanPayment* percentage as compared to any other group. But this still can't assure us that people with higher recommendations pay more loan each month. This question was relevant because we want to find out IF PEOPLE WITH MORE RECOMMENDATION ARE POWERFUL AND RICH ENOUGH TO PAY MORE LOAN MONTHLY? The answer is we can't be really sure.

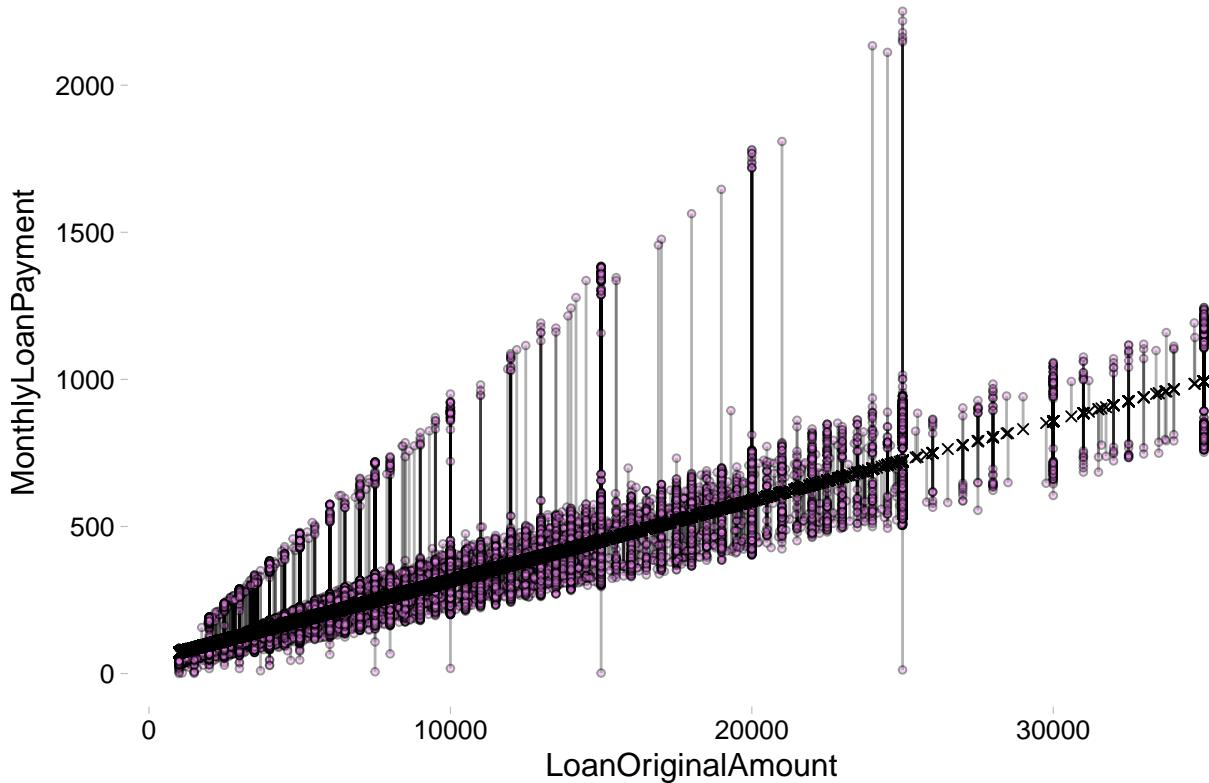
Let's predict the Monthly Loan Payment

Correlation of MonthlyLoanPayment vs LoanOriginalAmount



As we can see the a very good linear relationship here. Let's explore this even further. There are multiple number of models that we can create with different number of input or independent variables. Let's try to create a linear model and see how we can improve it.

Residuals plot for initial model (fit1)



```
##
## Call:
## lm(formula = MonthlyLoanPayment ~ LoanOriginalAmount, data = model)
##
## Residuals:
##    Min      1Q   Median      3Q     Max 
## -710.32  -33.57   -3.00   27.64 1528.86 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 4.556e+01 4.716e-01  96.61 <2e-16 ***
## LoanOriginalAmount 2.708e-02 4.177e-05 648.33 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 74.23 on 77112 degrees of freedom
## Multiple R-squared:  0.845, Adjusted R-squared:  0.845 
## F-statistic: 4.203e+05 on 1 and 77112 DF,  p-value: < 2.2e-16
```

LoanOriginalAmount should be the most important factor to decide what would be the *MonthlyLoanPayment*. As we can see that the correlation has very good *Adjusted R-squared* value of 0.845 or almost 85%. The *Residual Standard Error* is however quite high of almost 75 which indicates the average performance of the linear model. We can visualize how the residuals are spread from the above plot. Their distribution is quite high. So the bigger question is that can we improve this model. Improving the model would ensure the following things

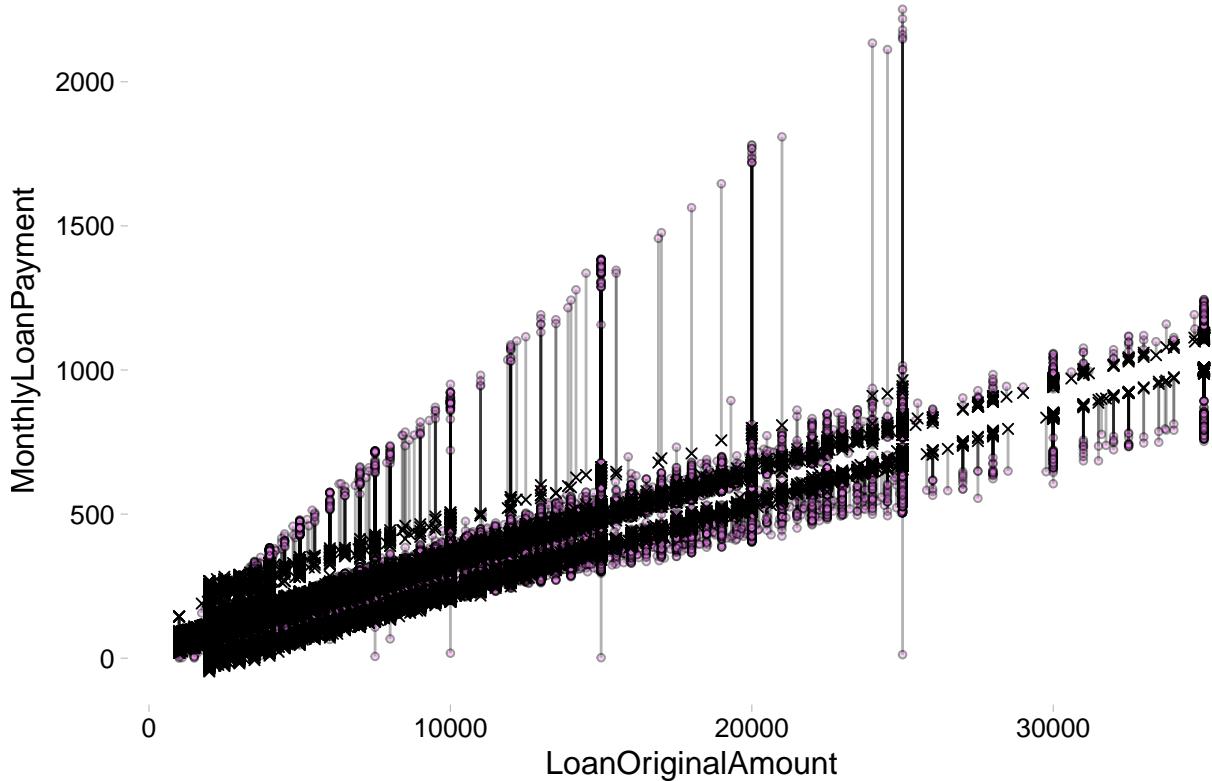
1. Higher Adjusted R-squared value

2. Lower Residual Standard Error
3. Less variation/ Less dispersed residual graph

The key to improve any model liner/non-linear is to increase the number of independent variables. The choice of independent variable should be very logical. We should not include the variable which are dependent on the predicting variable. The independent variables should be truly independent. Here are the following variables that are chosen to improve the model.

1. `LoanOriginalAmount`: This should be the most important one because how much the borrower has to pay each month should be dependent on the total amount that is loaned
2. `DebtToIncomeRatio`: This is important because lenders may ask more money each month who has a record of high `DebtToIncomeRatio` because they are considered borrowers with poor prospect and lenders might charge them more if giving loans
3. `ProsperRating..numeric.`: This is obvious because we have shown already that lenders trust the prosper rating a lot over anything else and hence borrowers with poor ratings should pay more each month as compared to better ratings.
4. `Term`: Term decides how long the borrower is opting for the loan and hence longer the Term shorter should be monthly payment.
5. `LoanOriginationQuarter`: Lenders may ask for different monthly loan payment on different time of the year because of dynamic interest rate and changing macro economic factors on different time of the year.

Residuals plot of improved model (fit2)



We can see a lot less variation in the residual. Moreover when we summarize the fit2 for our improved model, we see that the value of *Residual standard error* decreased a lot to almost 49 and *Adjusted R-squared* increased to almost 94%. This was a huge jump from our previous model. Hence we can say that fit2 predicts the monthly loan amount better than the fit1 model. This also proves that *LoanOriginalAmount* is not only the deciding factor of *MonthlyLoanPayment*. There are other factors too that contribute to this. From the `summary(fit2)`, we observe that the slope values like below

1. `LoanOriginalAmount`: `3.124e-02`
2. `DebtToIncomeRatio`: `3.211e+00`
3. `ProsperRating..numeric.`: `-1.050e+01`
4. `Term`: `-4.993e+00`

We can see that as we might have assumed, *LoanOriginalAmount* & *DebtToIncomeRatio* has positive slope which is obvious because if these two increase, lenders ask for more money from the borrowers. On the other hand, *Term* has a negative slope which is logical because if term increases, monthly loan payment should decrease.

We can also see that the both maximum and minimum residual for our improved model is lesser than our previous model.

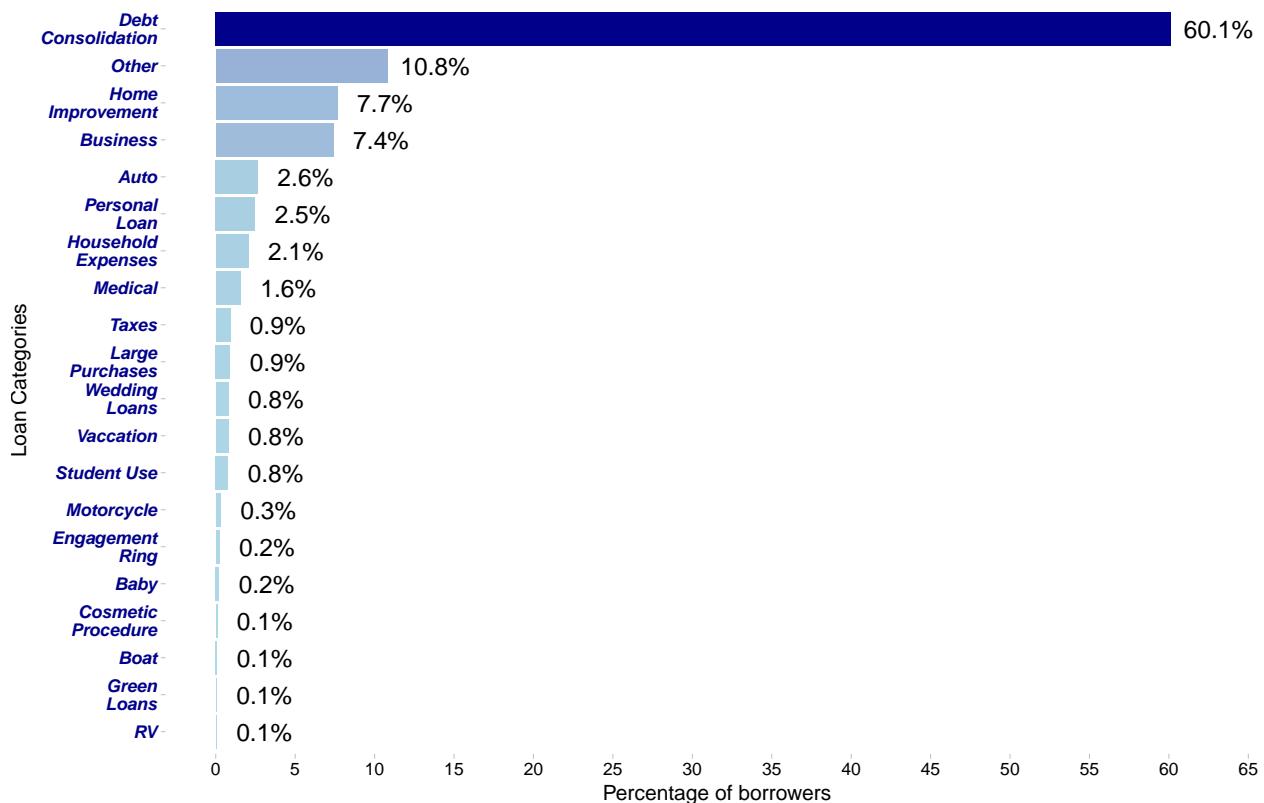
Some Final Thoughts

Let's select 3 plots from what we have discussed and elaborate them bit further.

What people loans for ?

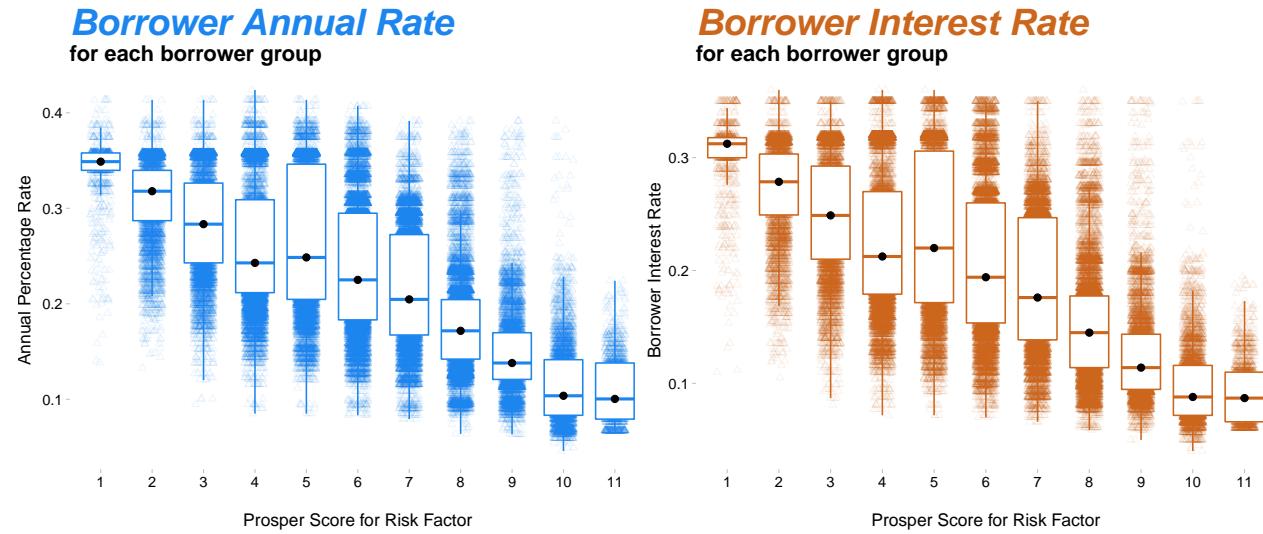
People Loan to repay Loans

this is why people loan!!



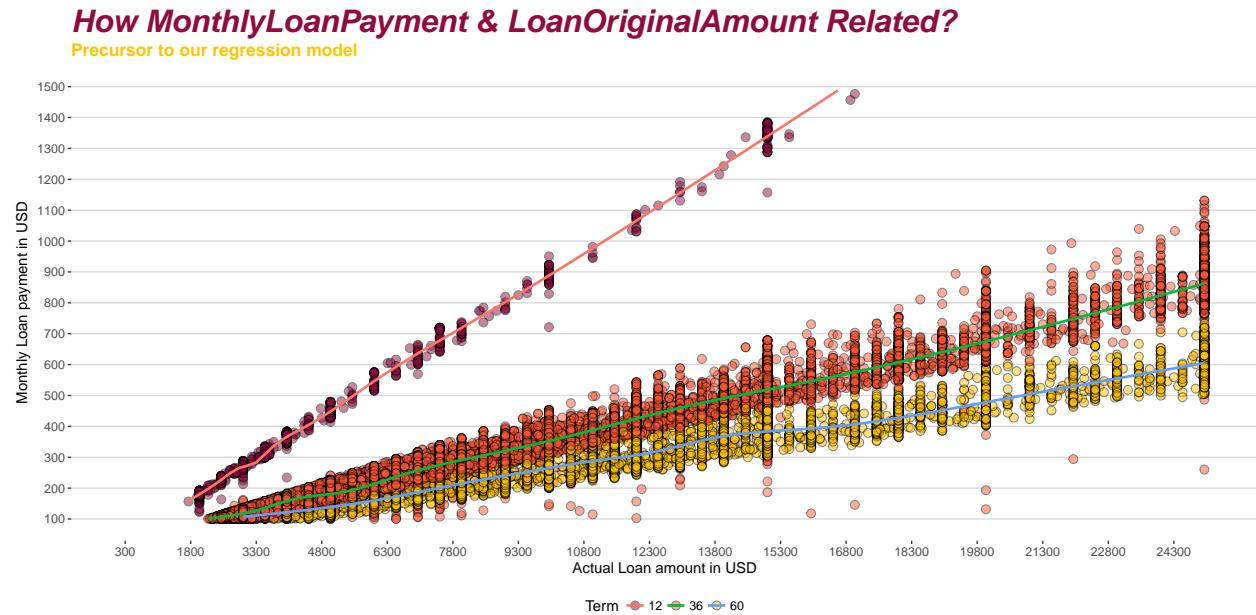
It is really strange that people took loans to reimburse loans/debts and lenders also issued loans to these people. Isn't it a bad idea to take loans to pay for loans? Aren't we falling into a vicious cycle? We can also see that Home Improvement is above Medical which may mean well insured population or negligence toward health.

Does lenders prefer borrowers with better ProsperScore ?



This depicts how we interest rates are affected by the Prosper Score for risk factor. As the score improves, the interest rate shows a declining trend visible from the black points. This indeed proves that the lenders like to charge less from borrowers with better prosper score.

Correlation of MonthlyIncome vs MonthlyLoanPayment



This created the base for the Linear regression model that was already discussed. This shows that *LoanOriginalAmount* and *MonthlyLoanPayment* has very good linear relationship for each *Term*. We have improved the model and by adding multiple input variables and achieved a very good *Adjusted R squared* value of almost 94% and a very low *Standard Residual Error*.

Reflection

Struggle with this data set were many. I was at first intimidated by its size of number of rows. I honestly have never seen any CSV file which is this big and contains this many variables. However the real relief was the CSV file that contained the meaning of all the variables in the data set. But the biggest struggle that I faced in this data set was actually common in most of the data sets that contain this many variables. Some of the variables seem to have the same meaning for example *BorrowerAPR*, *BorrowerRate*, *LenderYield*, *EstimatedEffectiveYield*. These terms seemed to have similar meaning but not exactly the same. Now the question was which one to include in our analysis and which is more important than other? Although with little bit more understanding I solved it but it took some time. Another struggle was to establish right combination of variables to create accurate linear regression model for minimum residual. This problem could have been solved using *Gradient Decent* but because it being out of scope I had to try different combination manually. This also took some time.

Successes were achieved even after the above mentioned struggle because of some detailed analysis of the data set and the CSV file supplied that explains the variables in the data set. Linear Regression model's accuracy was checked by carefully taken the variables and then checking the *Adjusted R-squared* values and the *Residual standard error* and plotting the Residual standard error. By doing this I could actually improve the regression model substantially already discussed earlier.

Although some **ideas for future exploration** can improve this analysis substantially. I really think regression model can be improved further using *Gradient Decent* to better approximate the slope and the intercept of the line. I also believe that there is some information hidden inside the delinquency variables and late payment variables in *CurrentDelinquencies*, *AmountDelinquent*, *DelinquenciesLast7Years* and *OnTimeProsperPayments*, *ProsperPaymentsLessThanOneMonthLate*, *ProsperPaymentsOneMonthPlusLate* with respect to the Monthly loan amount which can be explored further.

However one thing is confirmed, Financial Data set is really intensive stuff.