Prosper Loan Data Analysis

Uma Maheswari Raju

Introduction:

Prosper was founded in 2005 as the first peer-to-peer lending marketplace in the United States. Since then, Prosper has facilitated more than $13 billion in loans to more than 860,000 people. Prosper Marketplace is backed by leading investors including Sequoia Capital, Francisco Partners, Institutional Venture Partners, and Credit Suisse NEXT Fund. In this Exploratory Data Analysis, I explore a Prosper dataset containing loan information for over a 100,000 people between the years 2006 and 2013.

This data set contains 113,937 loans with 81 variables on each loan, including loan amount, borrower rate (or interest rate), current loan status, borrower income, and many others.

List of Variables:

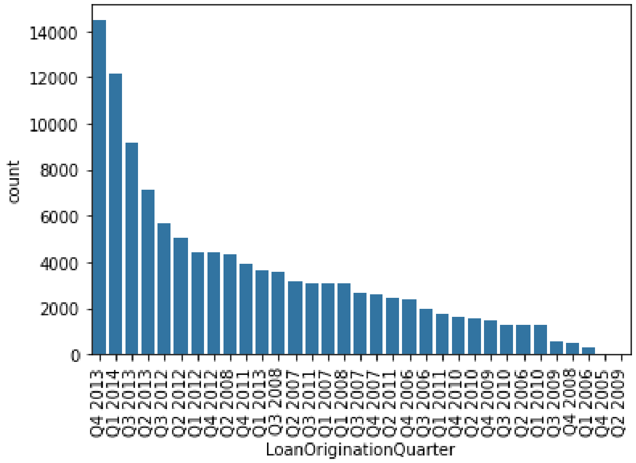
LoanStatus : Current status of the loan like chargedoff, completed, defauted etc.
EstimatedEffectiveYield : Yield of lenders from borrowers minus the processing fee and late fines
ProsperScore : Risk Factor score from 1 to 10. 10 being least risky
BorrowerAPR : The Borrower's Annual Percentage Rate (APR) 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

This project is divided into three parts such as Univariate plots, Bivariate Plots and Multivariate Plots and Summary Plots which summarize the final outputs of this exploratory analyses.

Univariate Plots:

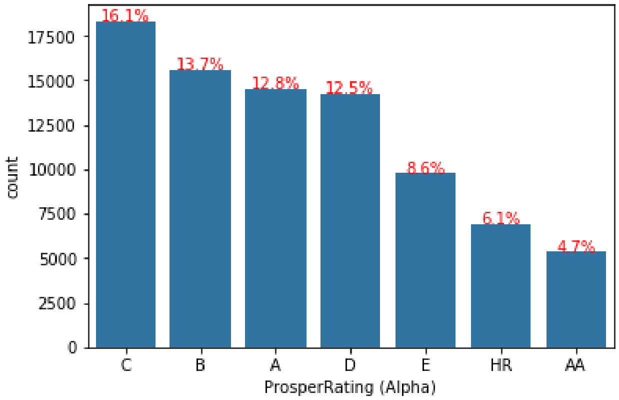
This section describes about individual variable.

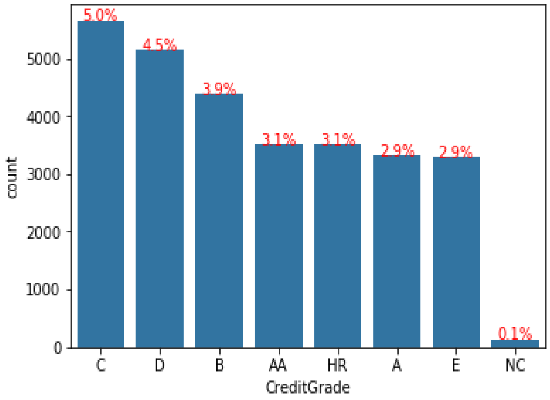
Loan Origination Quarter Plot:¶

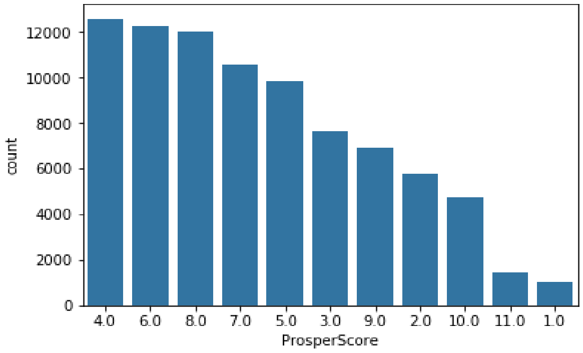


- The result shows loan origination year starts from 2005 to 2014 with number of listings. It indicates low listings during Q1 2006, Q4 2008 and Q3 2009. This time period coincides with the collapse of Lehman Brothers and the following fallout in the global financial system. Even though Prosper is an alternative to traditional loan models, it appears that its business was not immune to the global economic crisis. Perhaps Prosper's credit policies were much looser before the credit crisis? After all, it appears that Prosper only established its Prosper Rating and Prosper Score in July 2009.

Prosper Rating, Prosper Score and credit grade Plots:

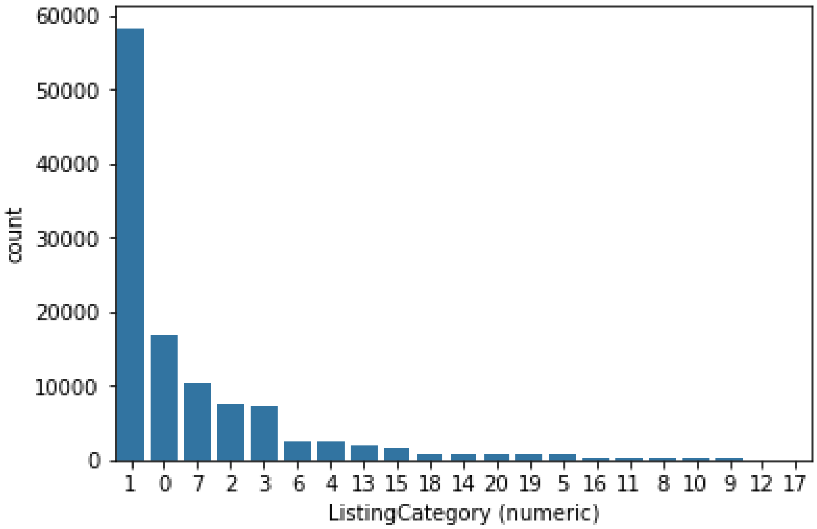






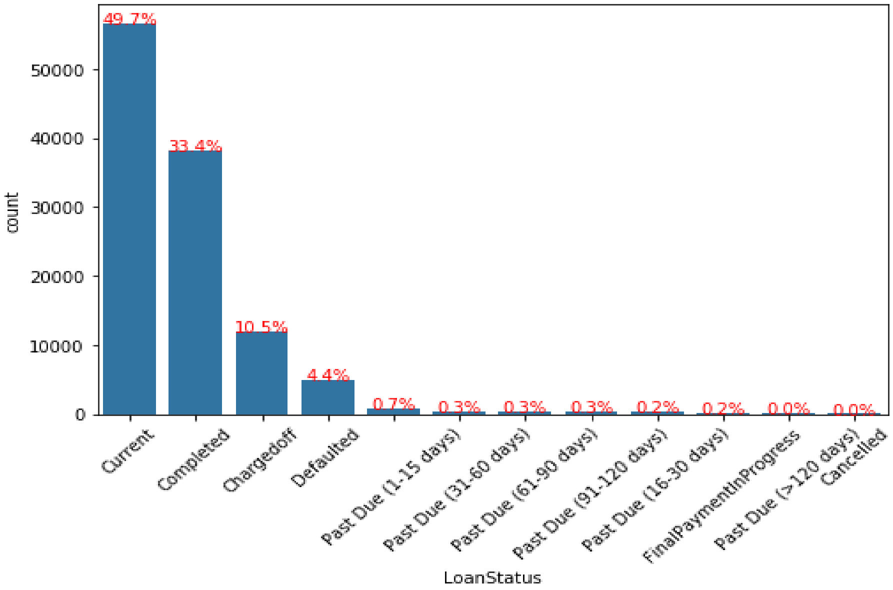
- Overall borrowers with better ratings tends to be less doubtful than borrowers with poor ratings. It means that the best borrowers are really best and the prosper rating really works.
- Prosper rating with C has high number of listings(16.1%) and AA rating has low number of listings (4.7%). Both plots(prosper rating and credit grade is that there are less borrowers in the both ends of the credit rating (AA,HR,A,E). This is because most people with super good credit are financially stable and do not usually take loans while people at the tail-end of the credit rating don't always get approved for credits.
- prosper score with 4.0 has high number of listings compared to prosper score with 1.0

Listing Category Plot:

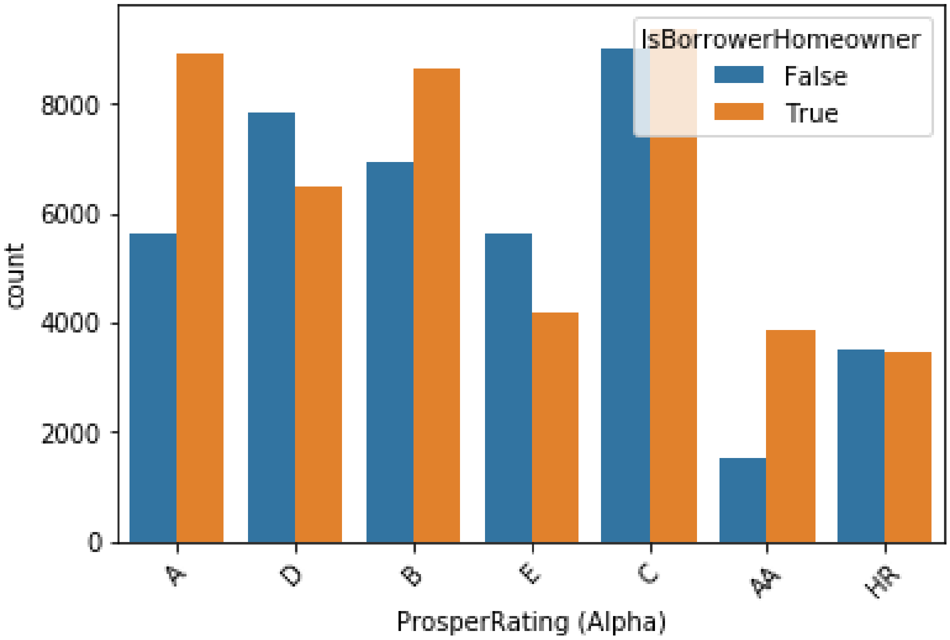


- Listing number 1(Debt Consolidation) has high number of listings compared to listing number 17. This shows People loan to repay loans

Loan Status Plot:

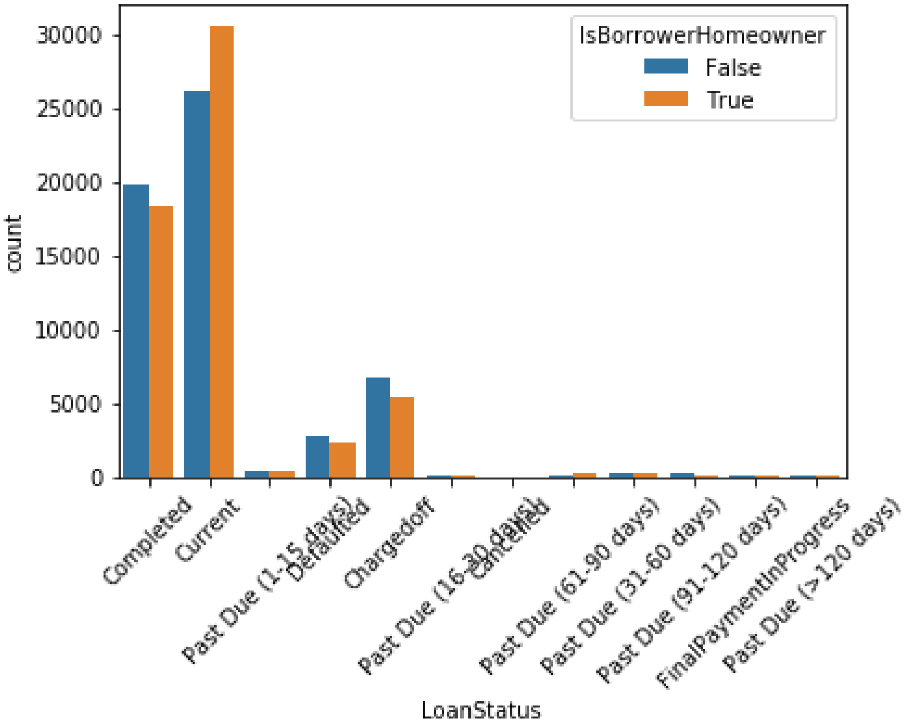


- Around 50% of loans are ongoing followed by completed.



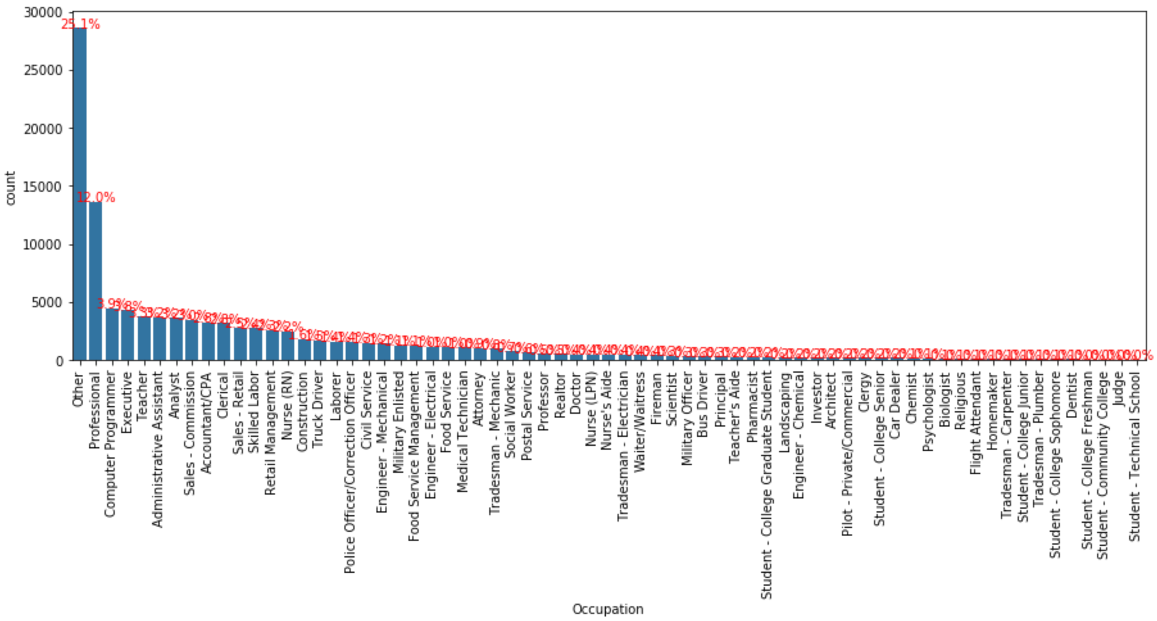
- Are Investors partial to Borrowers with better rating ??
- It is noticed that people with Prosper Rating of AA, A & B, more number of people opted for loan mentioning the reason for loan as Home Improvement. But for people with poor rating, the trend in opposite way. People who don't have any house 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.

Loan Status plot:

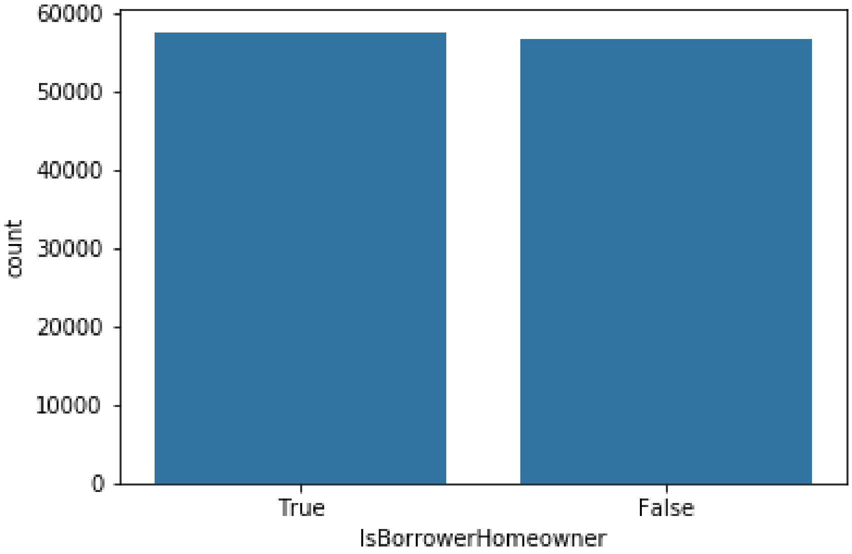


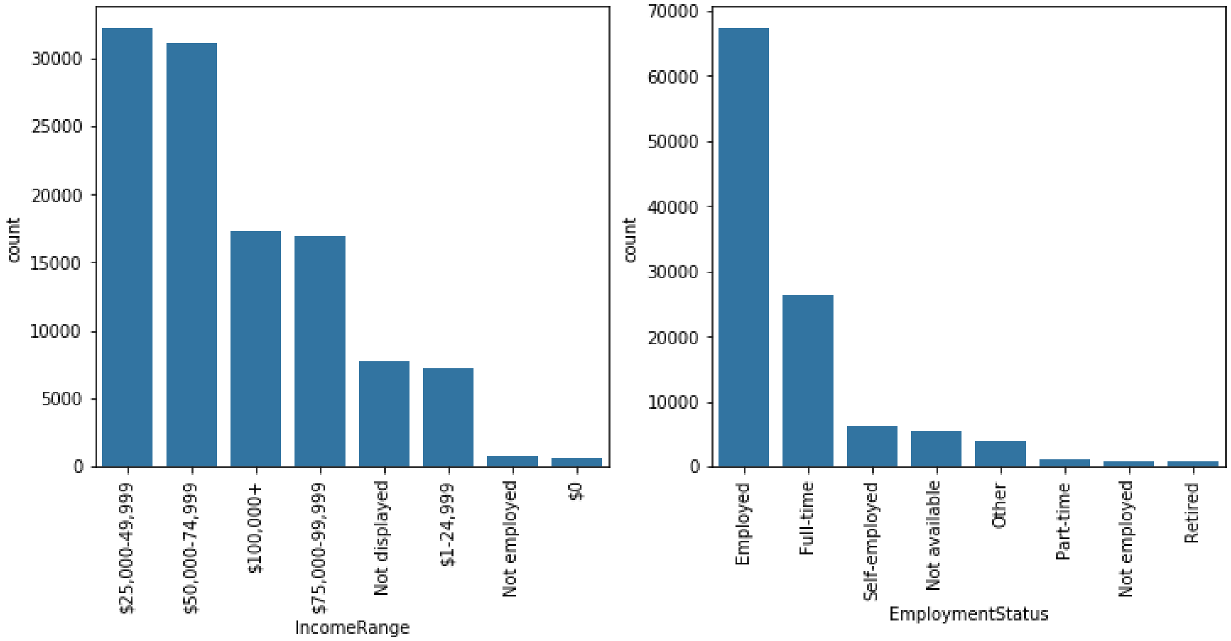
- above plot indicates that most of the current borrowers are homeowners

Occupation vs Count Plot:

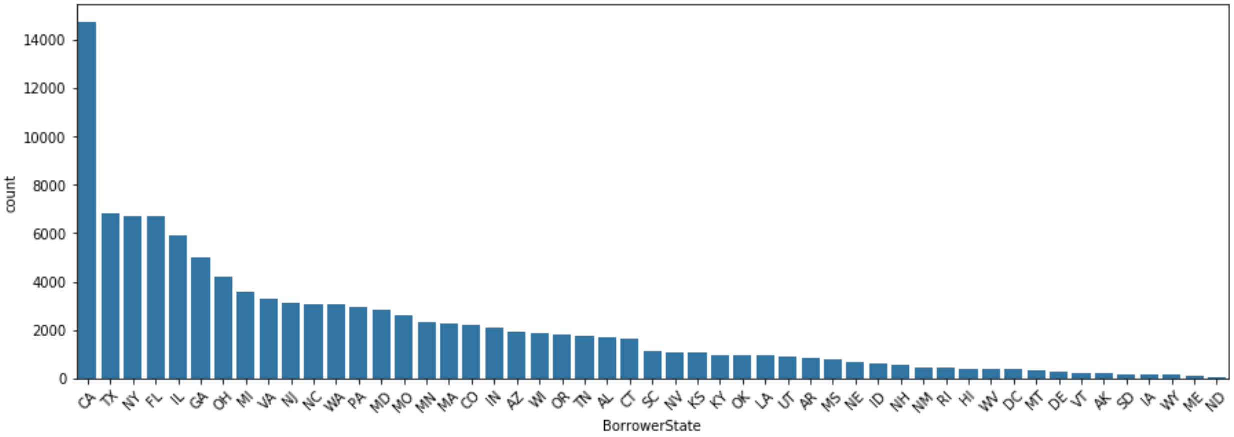


- above plot shows that most of the borrowers are professional occupation

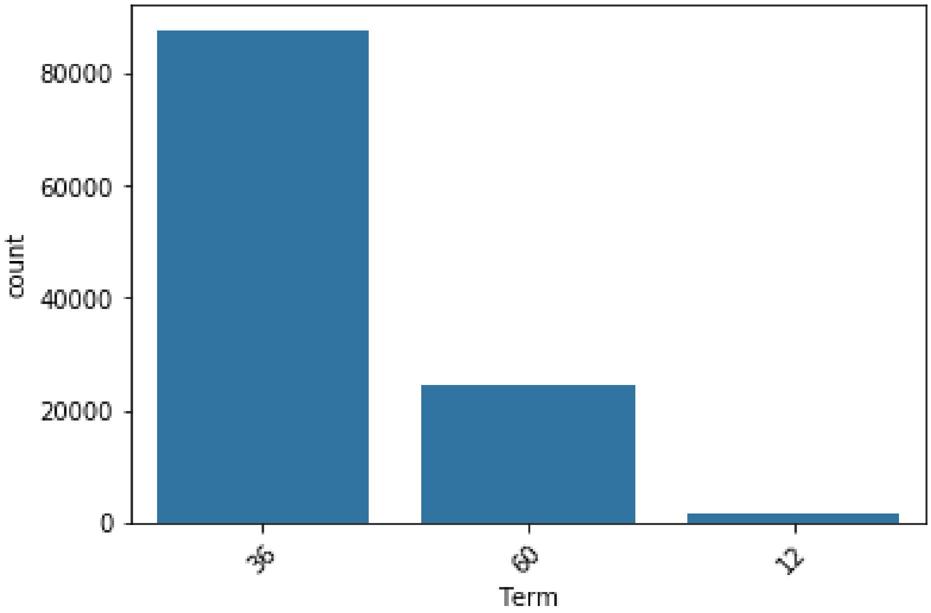




- above plot shows that income range around 25000 to 50000 are maximum borrowers and employed
- This plot approximate Normal distribution with borrowers with zero income and not employed constituting the lowest number of borrowers as expected. I do not think loans were given to $0 income earner or employed. This could be borrowers that lost their jobs after getting the loans or have loans secured by other assets.

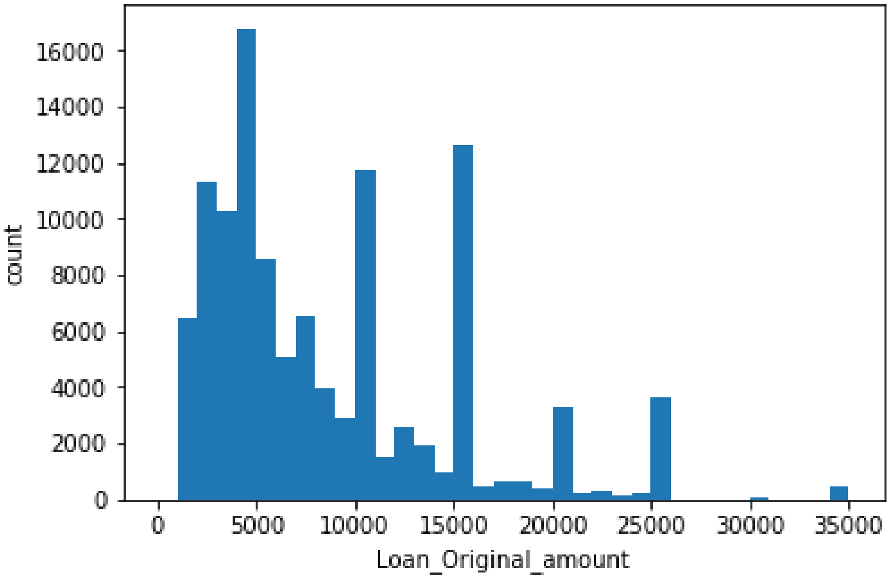


- California state has large number of borrowers followed by Texas. This could be because Prosper Loan was founded and local to CA and being the state with one of the highest state debt per capital. The other popular states include Florida, New York, Texas and Illinois.



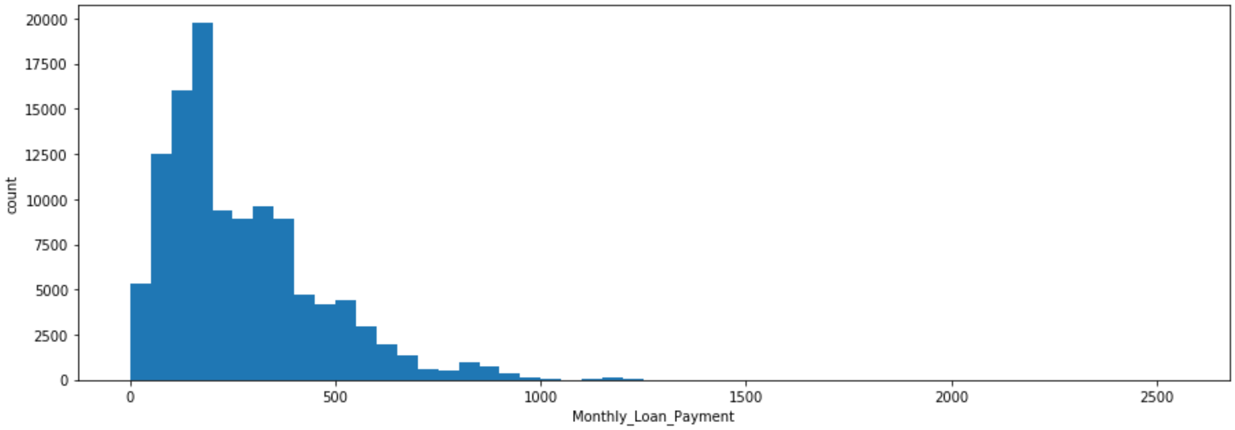
- 36 months term is the top in the list
- We can see that people don't really loan any amount for less than one year and the most popular loan amount is of 3 years although some people do choose for 5 years.

Loan original amount distribution:

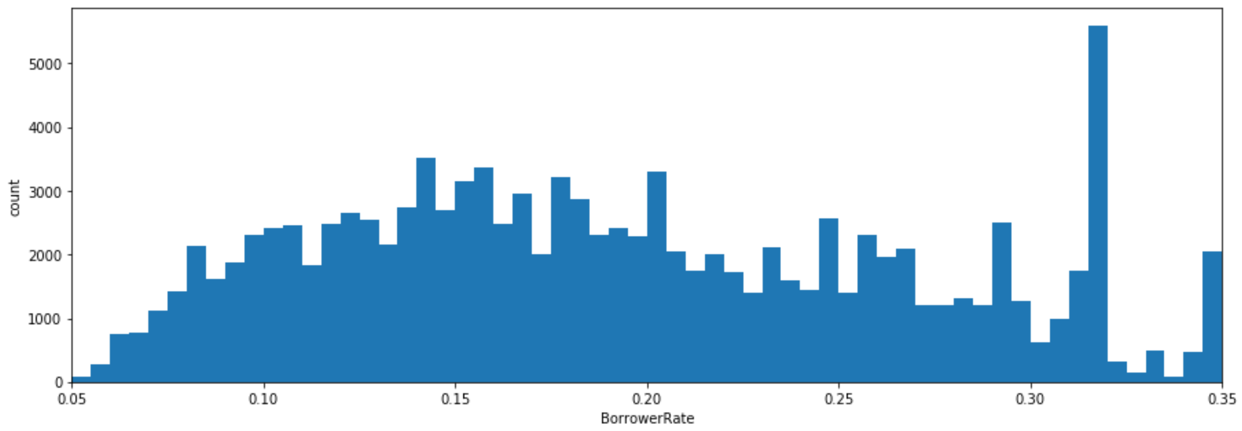


- mean loan original loan amount is around 8337−Bulkoftheloansgivenwerebelow15,000.Also,mostloanswereinmultiplesof8337–Bulkoftheloansgivenwerebelow15,000.Also,mostloanswereinmultiplesof5,000 as seen in number of loans at 5,000,5,000,10,000, 15,000,15,000,20,000, 25,000,25,000,30,000 and $35,000 points. It could also support the fact that loans were not granted based on values of underlying assets but mostly to refinance existing debts.

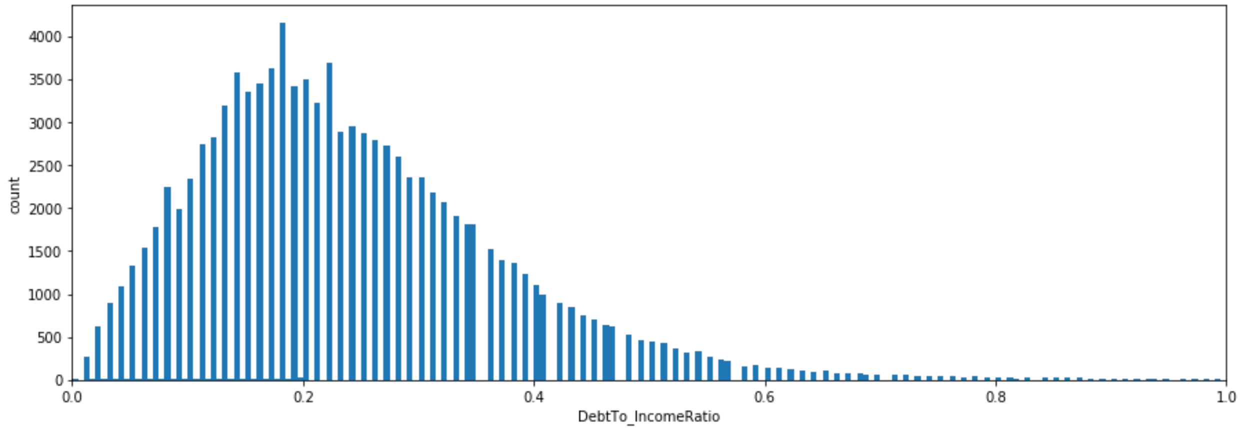
Monthly loan payment distribution:



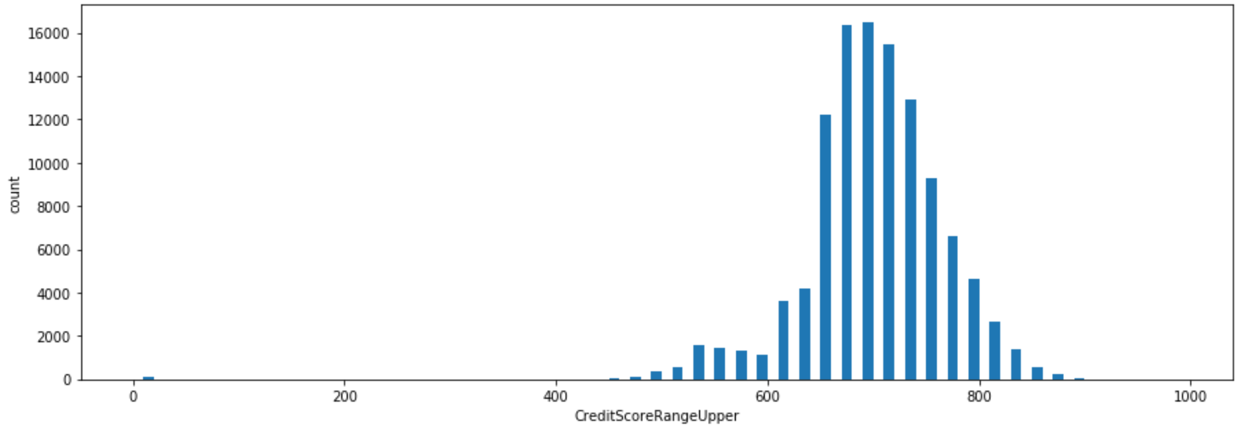
Borrower rate distribution:

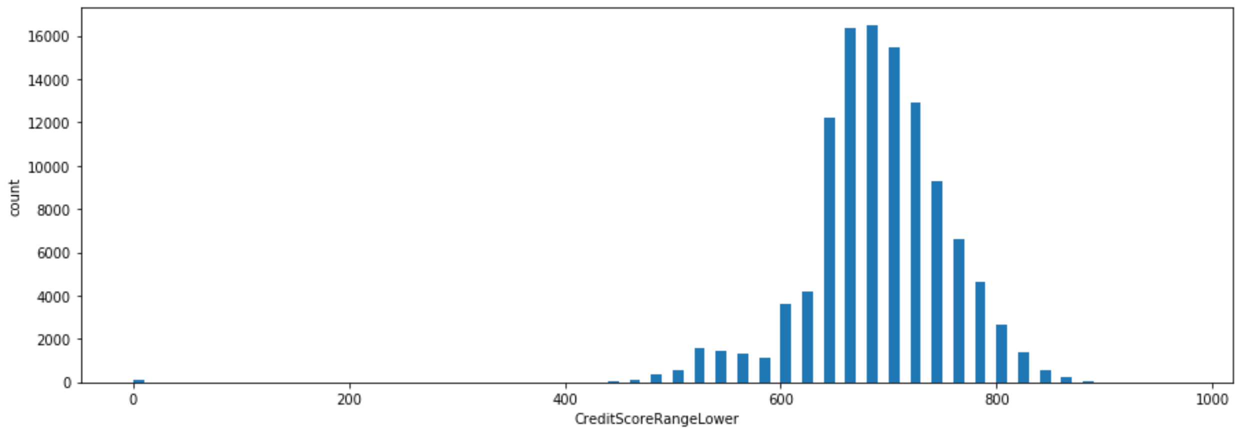


Debt to income ratio distribution:



Credit score distribution:





Discuss the distribution(s) of your variable(s) of interest. Were there any unusual points? Did you need to perform any transformations?

Variables such as credit grade, score, monthly loan payment, stated monthly income, debt to income ratio, prosper rating score, prosper rating, borrower rate, borrower APR, loan status and loan origination quarter were plotted for distribution. I did not see unusual distribution in those patterns. I did not do any transformation in this analysis

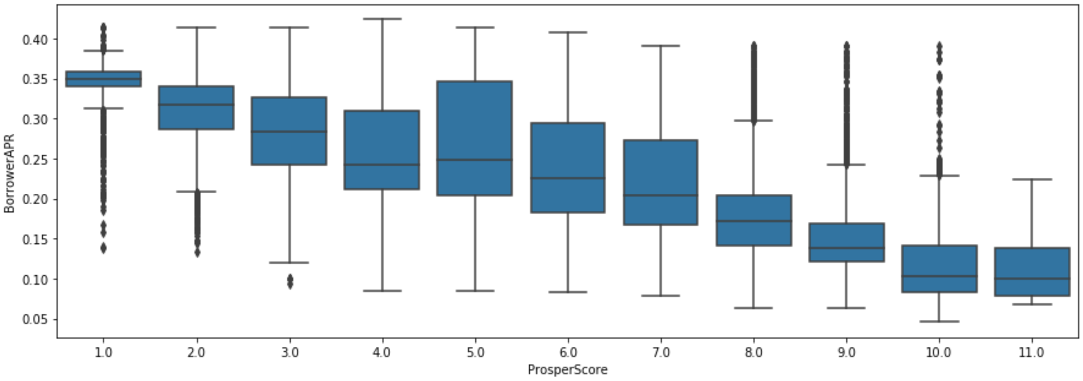
Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

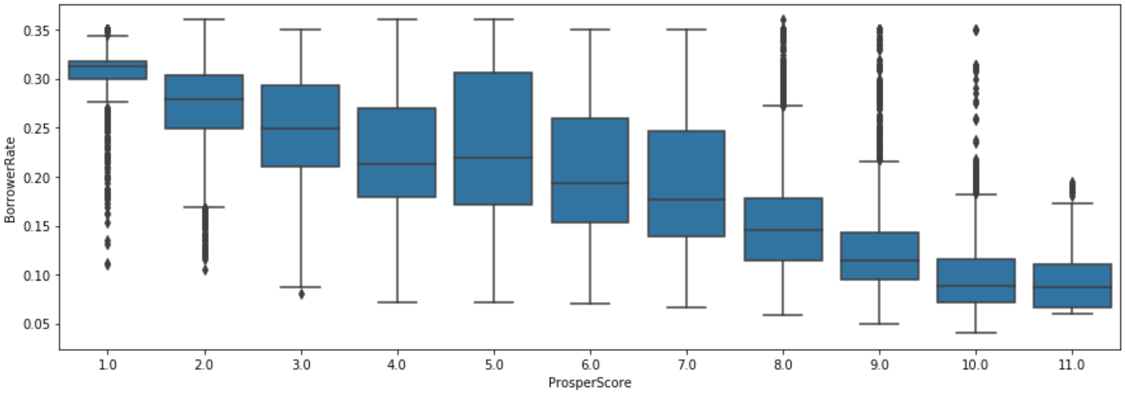
The main "unusual" distribution that stood out was the number of loans originated in each quarter, and how that number dipped to zero in early 2009.

Bivariate Exploration

In this section, investigation between pairs of variables in prosper loan data was considered for analysis. In this section, i would like to explore more on prosper score and rating on different variables such as credit grade, score, borrowers APR and borrowers rate.

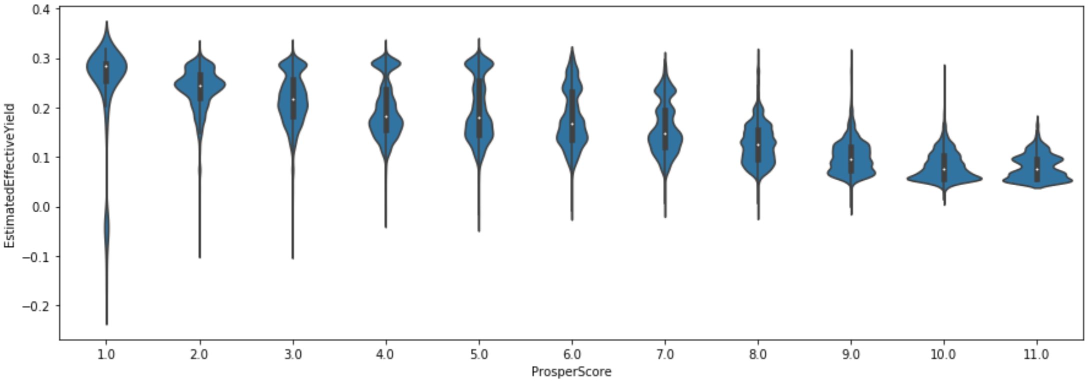
Prosper score vs borrower rate and Borrower APR:



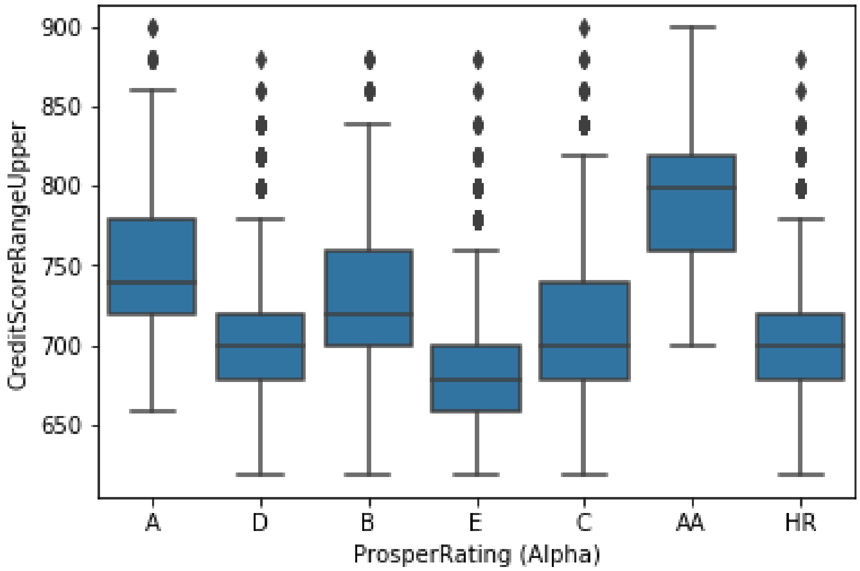


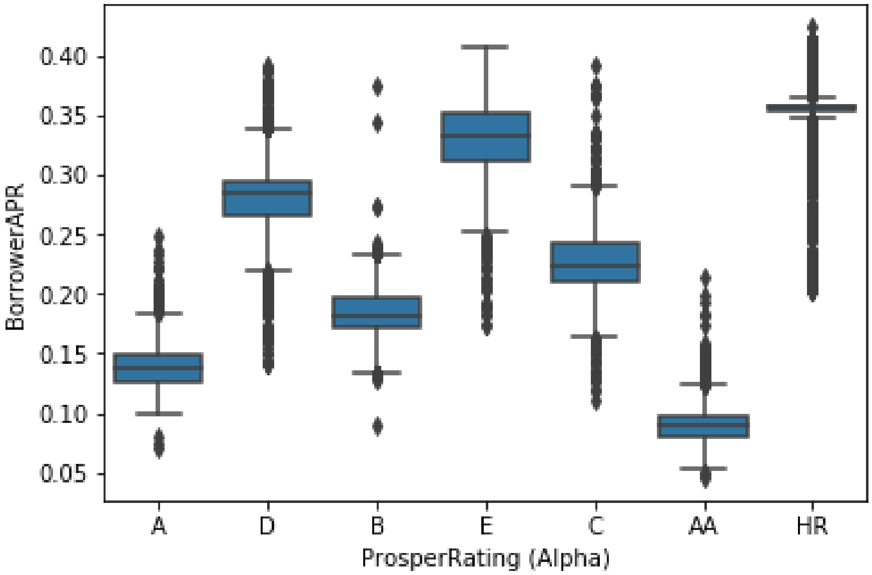
- Above chart shows how the interest rates are affected by the Prosper Score for risk factor. As the score improves, the interest rate shows a declining trend visible from the black points. This indeed proves that the lenders like to charge less from borrowers with better prosper score.
- We can clearly observe that for both Borrower APR and Borrower Rate which are metric for interest rates, we see a declining trend as the Prosper Score is increasing. This justifies the fact even more that lenders somehow prefers to charge less for all the borrowers with better Prosper Score as compared to borrowers with inferior Prosper Score.

Effective yield for each risk factors:

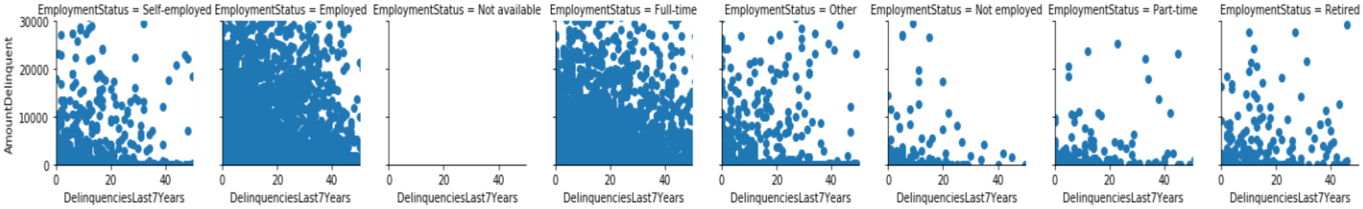


- Here more score for the risk factor means better the borrower and lesser score for risk factor means poor prospects from the borrowers. We can see that for lower Prosper Score distribution of effective yield in a lot more than the higher Prosper Score. This may mean that lenders charges a variety of 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 Prosper Score is increasing. This may mean that lenders give more relaxations to borrowers with better ratings as compared to borrowers with poor rating.



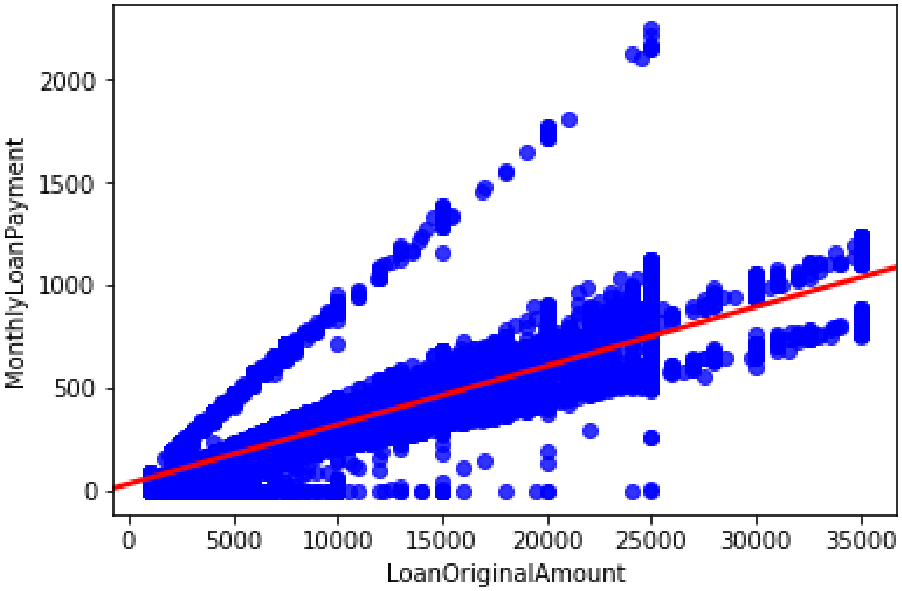


- The boxplots above show the relationship between borrower's Prosper rating and their assigned Annual Percentage Rate (APR). It's very clear that as we go down the ladder of risk - from a 'High Risk' to an 'AA' rating - the APR for the borrower reduces drastically.
- The variation in APRs also decreases as the loans get less riskier as displayed by the decreasing size of the boxes in the boxplots when going from 'HR' to 'AA'.

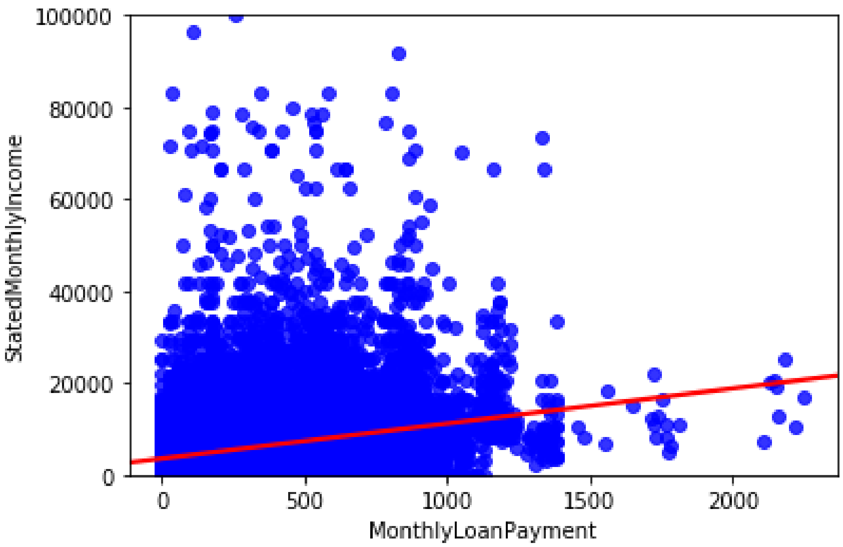


- Above plotting and general dispersion of data doesn't really reveal much trend in this plot.

Let's predict monthly loan payment:



- From above plot, we can see the positive linear relationship between loan original amount and monthly loan payment.



- We can see from the above scatterplot that most of the Monthly Loan Payment is distributed from 10 to 1000 and the Stated Monthly Income is distributed from 100 to 30000. There is definite a strong positive correlation between monthly income and monthly loan amount.

Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

- It's very clear that as we go down the ladder of risk - from a 'High Risk' to an 'AA' rating - the APR for the borrower reduces drastically.

- The variation in APRs also decreases as the loans get less riskier as displayed by the decreasing size of the boxes in the boxplots when going from 'HR' to 'AA'.
- We can clearly observe that for both Borrower APR and Borrower Rate which are metric for interest rates, we see a declining trend as the ProsperS core is increasing. This justifies the fact even more that lenders somehow prefers to charge less for all the borrowers with better Prosper Score as compared to borrowers with inferior Prosper Score.

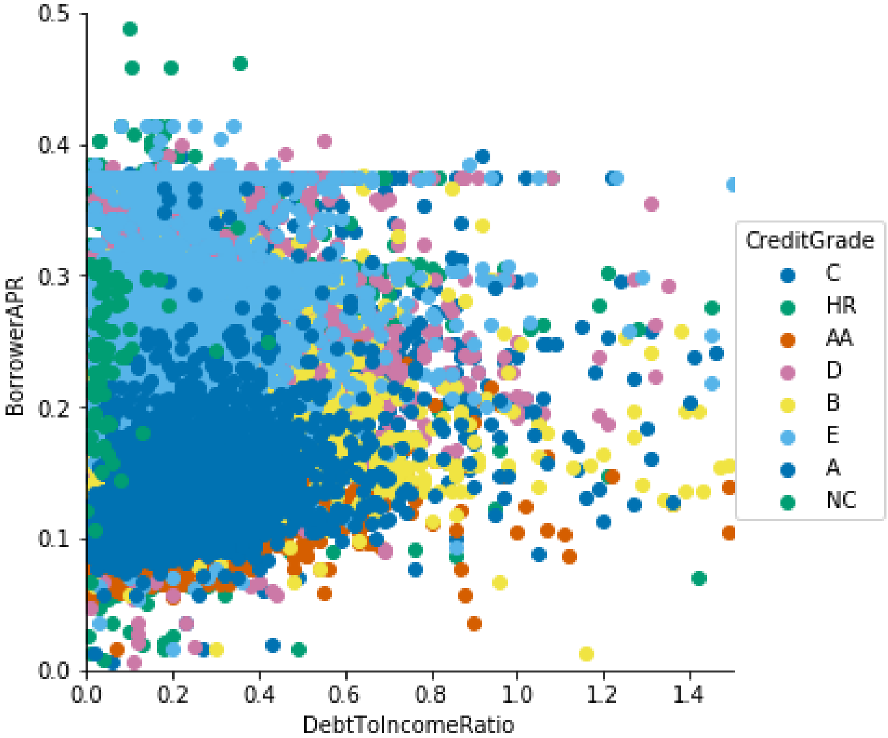
- Here more score for the risk factor means better the borrower and lesser score for risk factor means poor prospects from the borrowers. We can see that for lower Prosper Score distribution of effective yield in a lot more than the higher Prosper Score. This may mean that lenders charges a variety of 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 Prosper Score is increasing. This may mean that lenders give more relaxations to borrowers with better ratings as compared to borrowers with poor rating.

Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

- Yes, We can see from the above scatterplot that most of the Monthly Loan Payment is distributed from 10 to 1000 and the Stated Monthly Income is distributed from 100 to 30000. There is definite a strong positive correlation between monthly income and monthly loan amount.
- we can see the positive linear relationship between loan original amount and monthly loan payment.

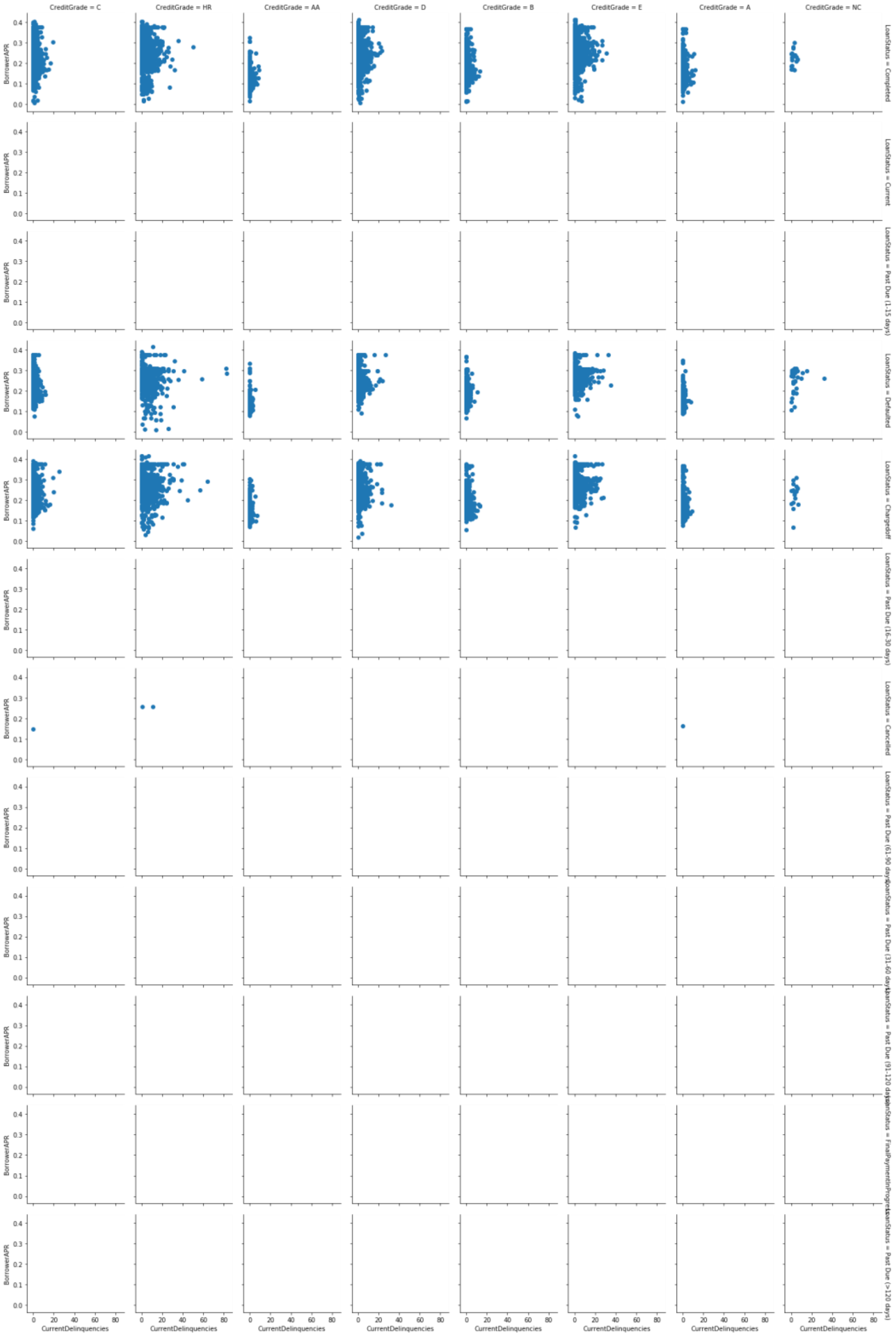
Multivariate Exploration

In this section three or more variables are considered for analysis. I would like to explore more on factors affecting borrowers APR such as credit score and debt to income ratio.



- This is a great plot with a lot of information. Here we have a scatter plot of borrower's APR and the debt to income ratio of the borrower, with the colors describing the risk category given to the particular loan.
- The first thing I noticed and found interesting was that 'A' category loans seem to have a lower APRs and a smaller range of debt-to-income ratios, both of which indicate less risk. The rest of the plot follow the color palette and APR increases as the rating gets riskier. Another thing is that most people tend to have debt-to-income ratios below 1, regardless of risk category. Lower ratings tend to be sparse in the 1.0+ debt-to-income ratio range.

Plot of current delinquencies and borrower APR:



- This plot was made to see if there were any distinct differences in terms of completing and defaulting loans when it came to current delinquencies. Unfortunately, there doesn't seem to be any tell-tale signs and both plots look pretty similar. However, I do notice that higher rated loans seem less diverse in terms of delinquencies and APR, and customarily lumped in the bottom left corner. As the loans gets riskier, the points get more varied and diverse, and tend to be all over the graph.

Talk about some of the relationships you observed in this part of the investigation. Were there features that strengthened each other in terms of looking at your feature(s) of interest?

- Scatter plot of borrower's APR and the debt to income ratio of the borrower, with the colors describing the risk category given to the particular loan.

Were there any interesting or surprising interactions between features?

- It was interesting that 'A' category loans seem to have a lower APRs and a smaller range of debt-to-income ratios, both of which indicate less risk. The rest of the plot follow the color palette and APR increases as the rating gets riskier. Another thing is that most people tend to have debt-to-income ratios below 1, regardless of risk category. Lower ratings tend to be sparse in the 1.0+ debt-to-income ratio range.

Summary

The prosper loans dataset contains over 100k observations with 81 variables distributing across 9 years. Initially, understanding the variables, terminology and general domain knowledge of financial peer-to-peer lending was the first obstacle in approaching this dataset. Another hurdle was which variable to consider for the analyses, not drifting too far off any one path of investigation and not pulling in new variables throughout the process. Below are key insights found through this analysis:

- It's very clear that as we go down the ladder of risk - from a 'High Risk' to an 'AA' rating - the APR for the borrower reduces drastically.
- The variation in APRs also decreases as the loans get less riskier as displayed by the decreasing size of the boxes in the boxplots when going from 'HR' to 'AA'.
- 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 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.
- We can see from the above scatterplot that most of the MonthlyLoanPayment is distributed from 10 to 1000 and the StatedMonthlyIncome is distributed from 100 to 30000. There is definite a strong positive correlation between monthly income and monthly loan amount.
- We can see the positive linear relationship between loan original amount and monthly loan payment.
- It was interesting that 'A' category loans seem to have a lower APRs and a smaller range of debt-to-income ratios, both of which indicate less risk. The rest of the plot follow the color palette and APR increases as the rating gets riskier. Another thing is that most people tend to have debt-to-income ratios below 1, regardless of risk category. Lower ratings tend to be sparse in the 1.0+ debt-to-income ratio range.