EDA_USING_R_LOAN_DATA

Deepak Singh

February 22, 2018

Univariate Plots

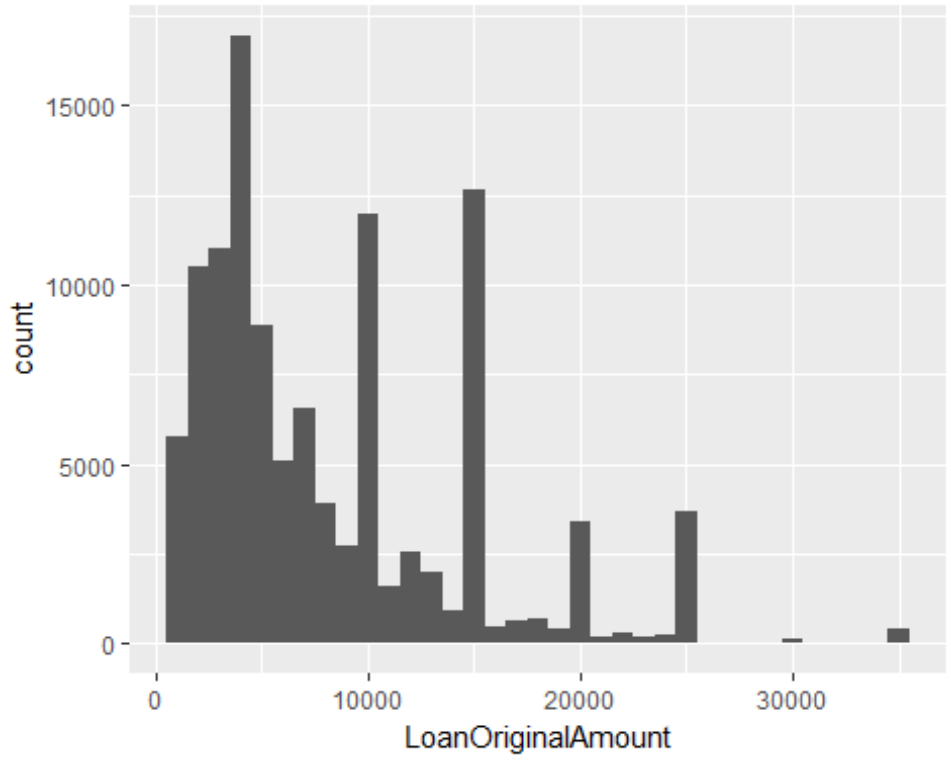
Reading the dataset and producing summary of Loan amount

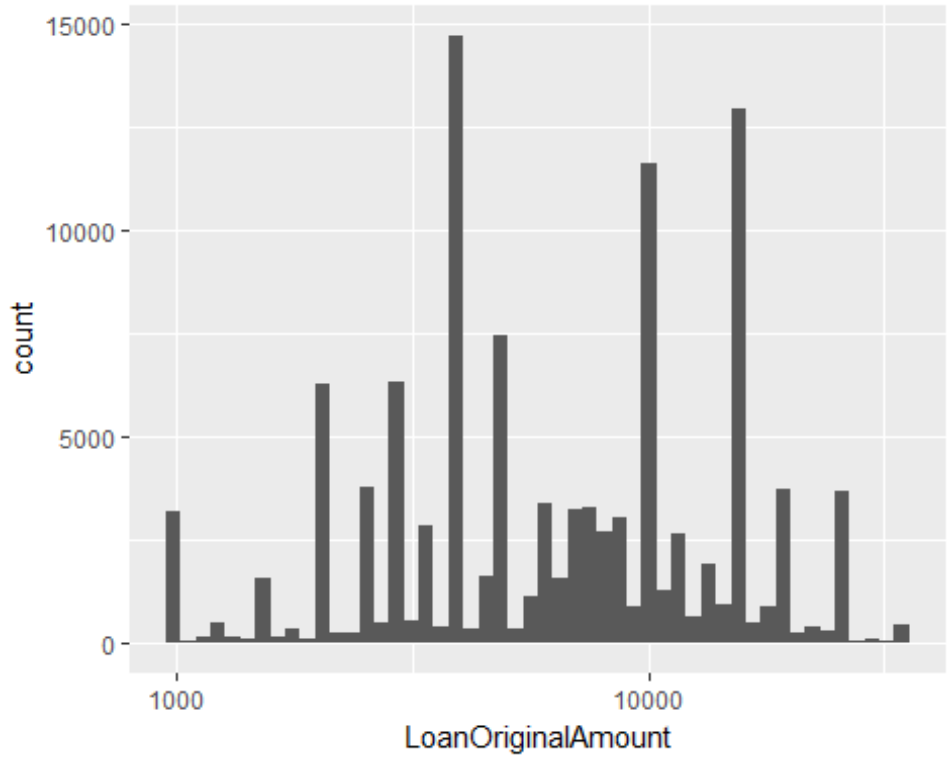
We will start with reading the loan proser data into the variable and produce summaries of some variables. Below is the summary of The origination amount of the loan.

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

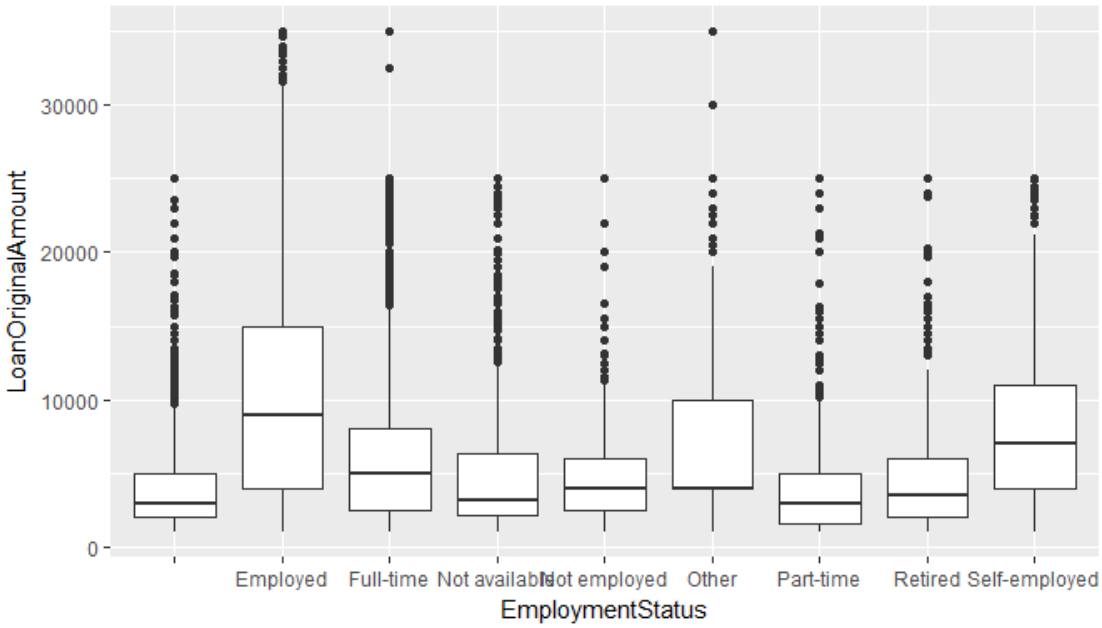
Producing histogram of Loan amount

The histogram of the loan amount reveals that there are few ouliers i the dataset. Also, the distribution of LoanOriginalAmount seems to skewed with few spikes.





Transformed the skewed distributed data for loan amount to better understand the the distribution of loan amount. The transformed distribution looks more like normaly distributed. Lets see if the employment status has an effect on the loan amount.



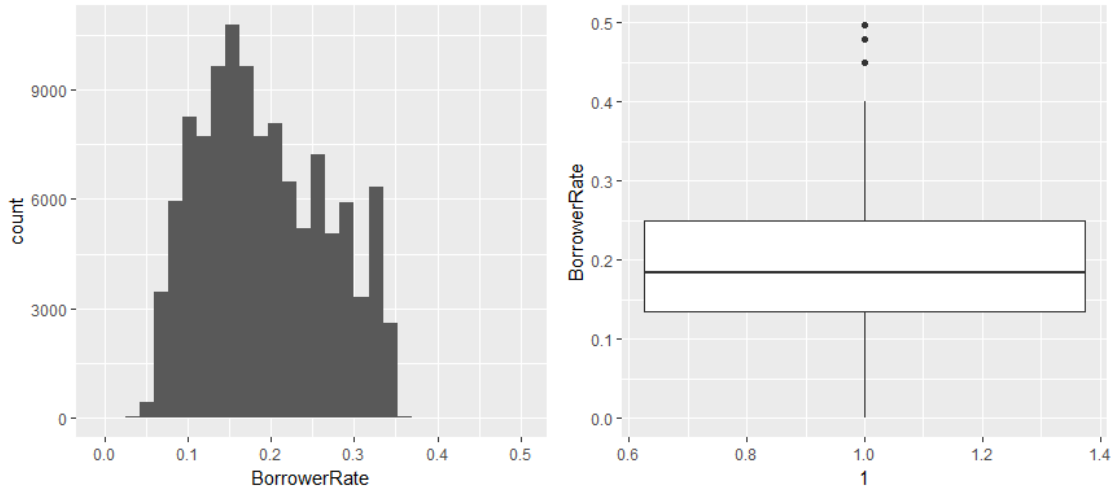

```
## EmploymentStatus:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1000    2000    3000    4563    5000   25000
```

```
## ------------------------------------------------------
## EmploymentStatus: Employed
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1000    4000    9000    9794   15000   35000
## ------------------------------------------------------
## EmploymentStatus: Full-time
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1000    2500    4950    6195    8000   35000
## ------------------------------------------------------
## EmploymentStatus: Not available
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1000    2138    3225    5373    6300   25000
## ------------------------------------------------------
## EmploymentStatus: Not employed
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1000    2500    4000    4873    6000   25000
## ------------------------------------------------------
## EmploymentStatus: Other
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1000    4000    4000    6862   10000   35000
## ------------------------------------------------------
## EmploymentStatus: Part-time
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1000    1600    3000    4089    5000   25000
## ------------------------------------------------------
## EmploymentStatus: Retired
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1000    2000    3500    4784    6000   25000
## ------------------------------------------------------
## EmploymentStatus: Self-employed
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1000    4000    7000    8123   11000   25000

## EmploymentStatus
##                     Employed     Full-time Not available  Not employed
##          2255         67322         26355          5347           835
##         Other     Part-time       Retired Self-employed
##          3806          1088           795          6134
```

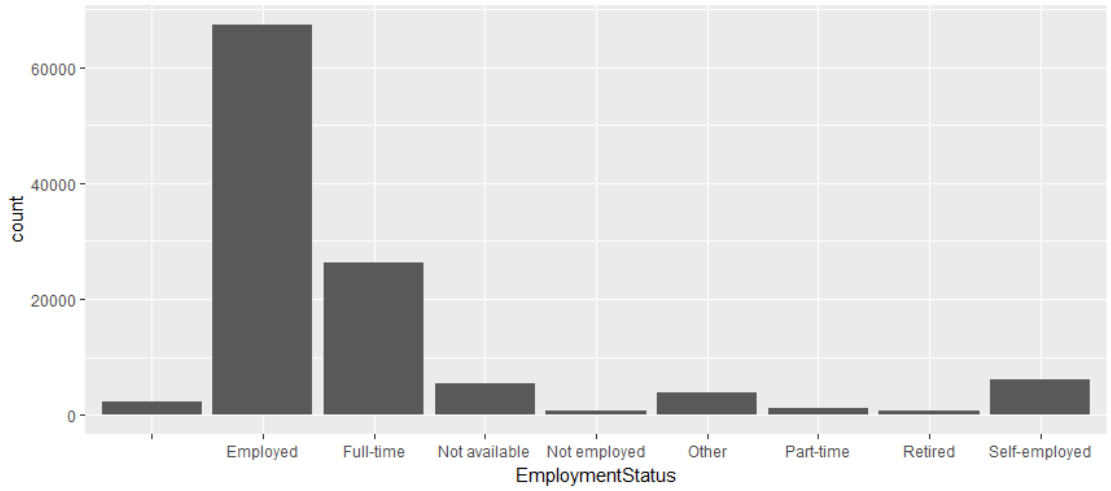
We can see from above that mean of loan amount is highest from the Employed followed by full-time then self-employed. Same is true for the number of loans issued for the respective employmentstatus

Analysing the Borrower Rate



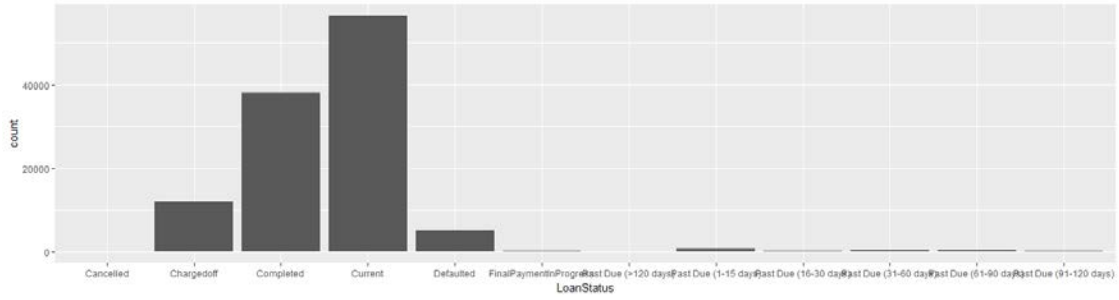
Interest rate are normally distributed mean/med = .19 However, there is a spike around .31. Also, the boxplot indicates that there are some outliers

Analysing the Employment Status



As shown from the plot above most of the Loans are taken by people who are employed.

Analysing the Loan Status



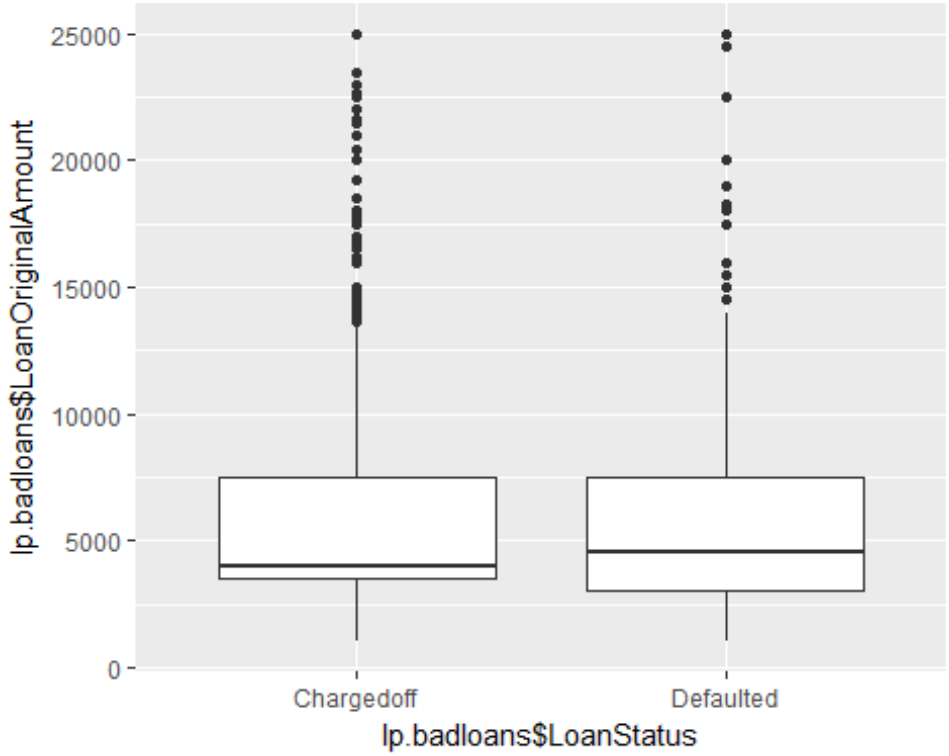
From the above plot we can see that most of the loans are current and completed followed by ChargedOff loans. This draws us towards the loans that have gone bad.

Analysing the bad loans

In this ananlysis we are going to see if there is any relationship of bad loans(loan status of chargedoff, defaulted) with credit score

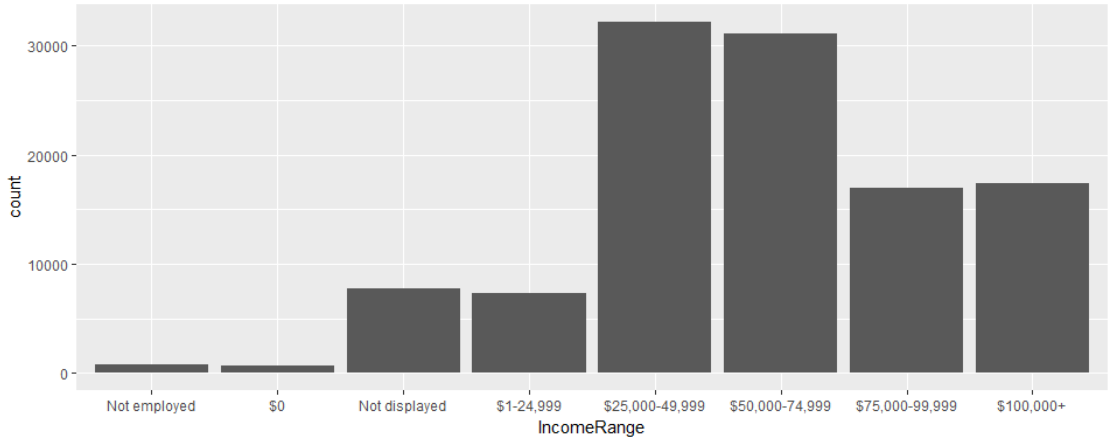
```
## lp.badloans$LoanStatus: Cancelled
## NULL
## ------------------------------------------------------------
## lp.badloans$LoanStatus: Chargedoff
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00    4.00    5.00    5.39    7.00   10.00       6
## ------------------------------------------------------------
## lp.badloans$LoanStatus: Completed
## NULL
## ------------------------------------------------------------
## lp.badloans$LoanStatus: Current
## NULL
## ------------------------------------------------------------
## lp.badloans$LoanStatus: Defaulted
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   4.000   6.000   5.619   7.000  11.000       3
## ------------------------------------------------------------
## lp.badloans$LoanStatus: FinalPaymentInProgress
## NULL
## ------------------------------------------------------------
## lp.badloans$LoanStatus: Past Due (>120 days)
## NULL
## ------------------------------------------------------------
## lp.badloans$LoanStatus: Past Due (1-15 days)
## NULL
## ------------------------------------------------------------
## lp.badloans$LoanStatus: Past Due (16-30 days)
## NULL
## ------------------------------------------------------------
## lp.badloans$LoanStatus: Past Due (31-60 days)
## NULL
```

```
## ------------------------------------------------------------
## lp.badloans$LoanStatus: Past Due (61-90 days)
## NULL
## ------------------------------------------------------------
## lp.badloans$LoanStatus: Past Due (91-120 days)
## NULL
```



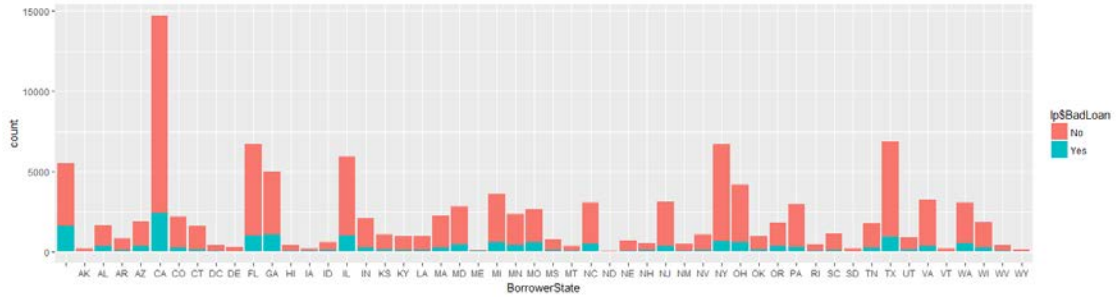
There are 6348 bad loans in the dataset after July 2009 and out of those the mean ProsperScore were 5.39 and 5.62 for both ChargedOff or Defaulted Loan status respectively

Analysis of Income ranges



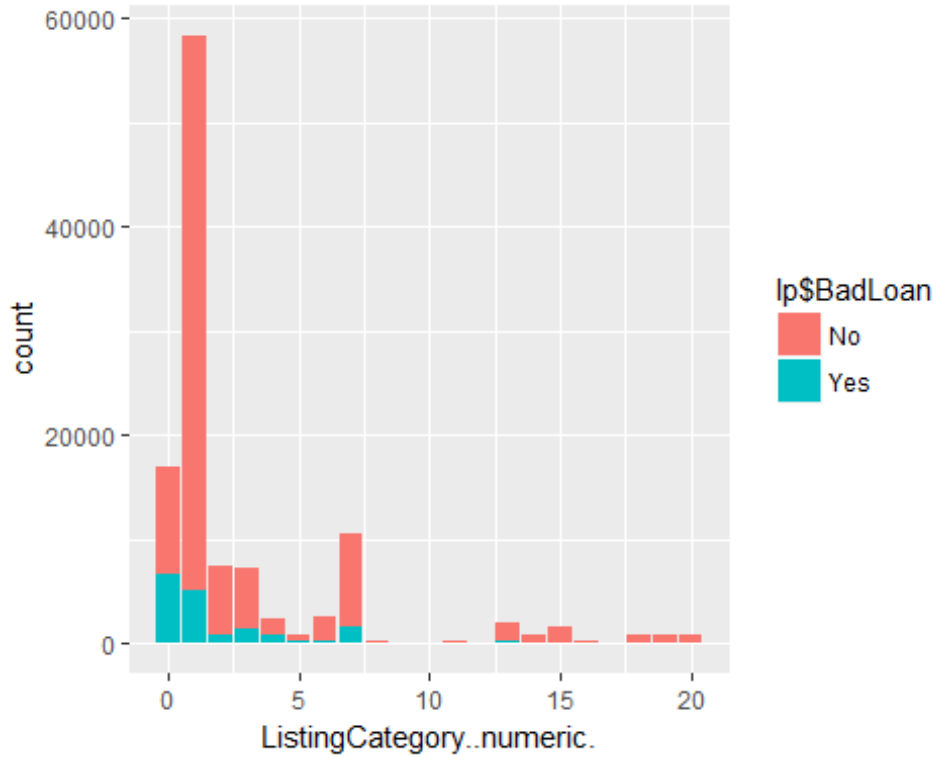
Income ranges are not normally distributed with peak for income ranging 25,000-49,999.

Analysis of Loans by state



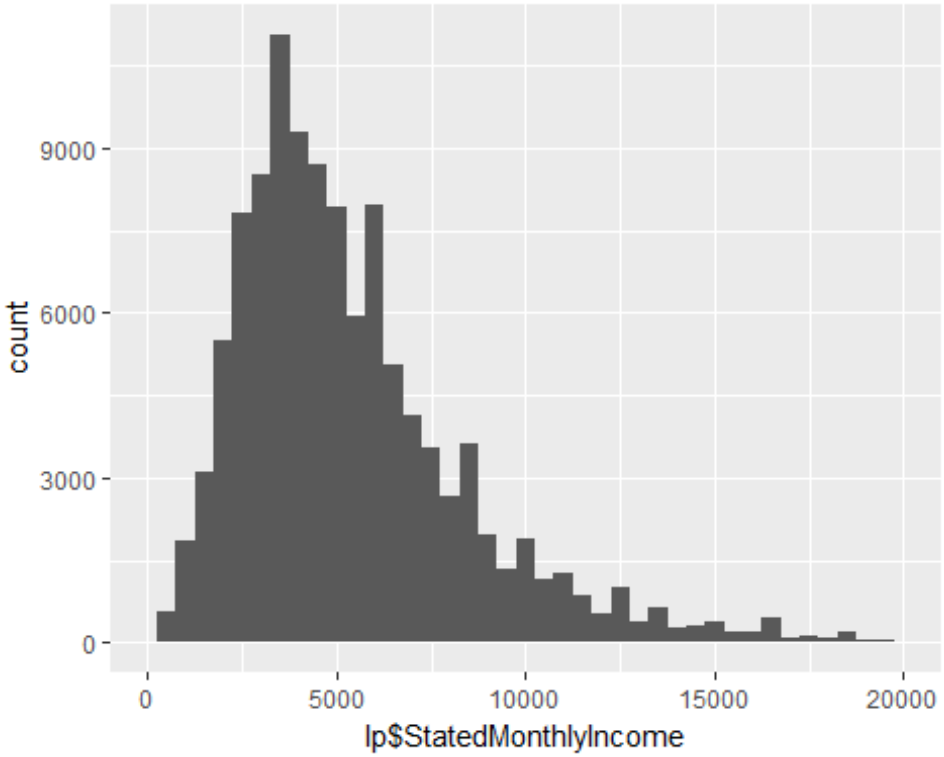
Largest number of borrowers are from California.

Analysis of Loans by Listing Category



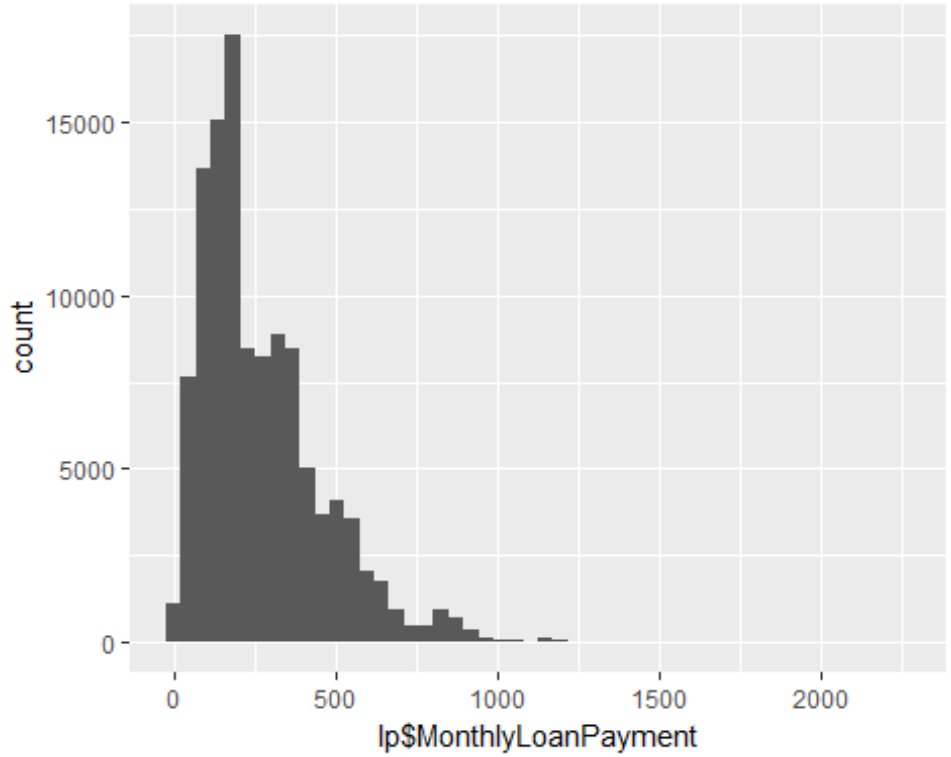
From the above plot most of the loans seems to be in Debt Consolidation category

Analysis of StatedMonthlyIncome



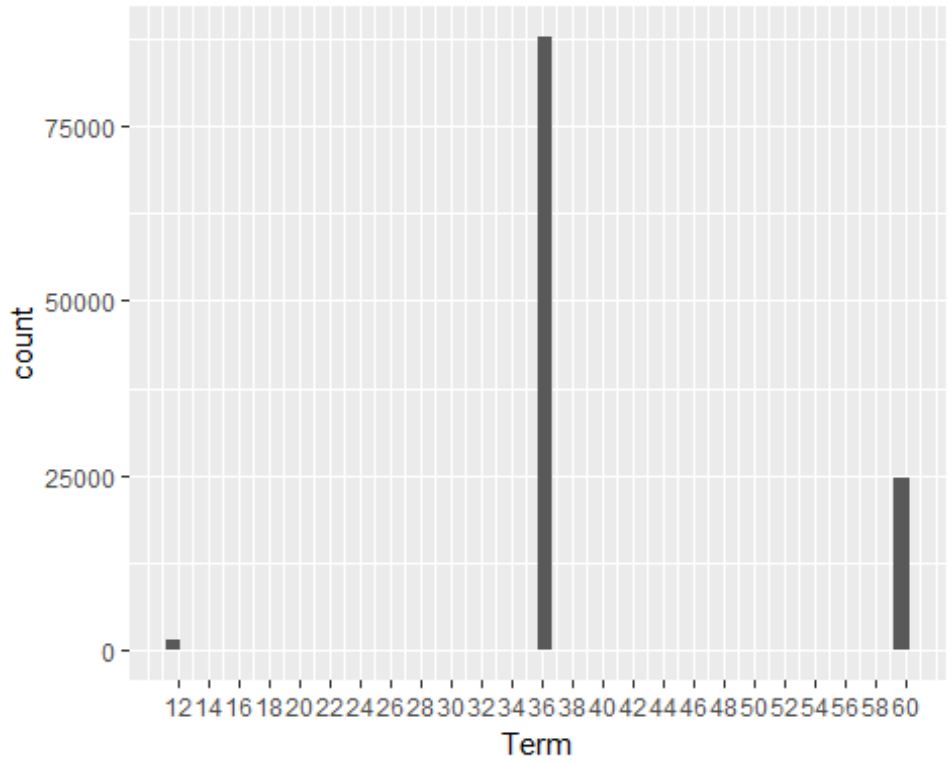
The distribution appears to be skewed

Analysis of MonthlyLoanPayment



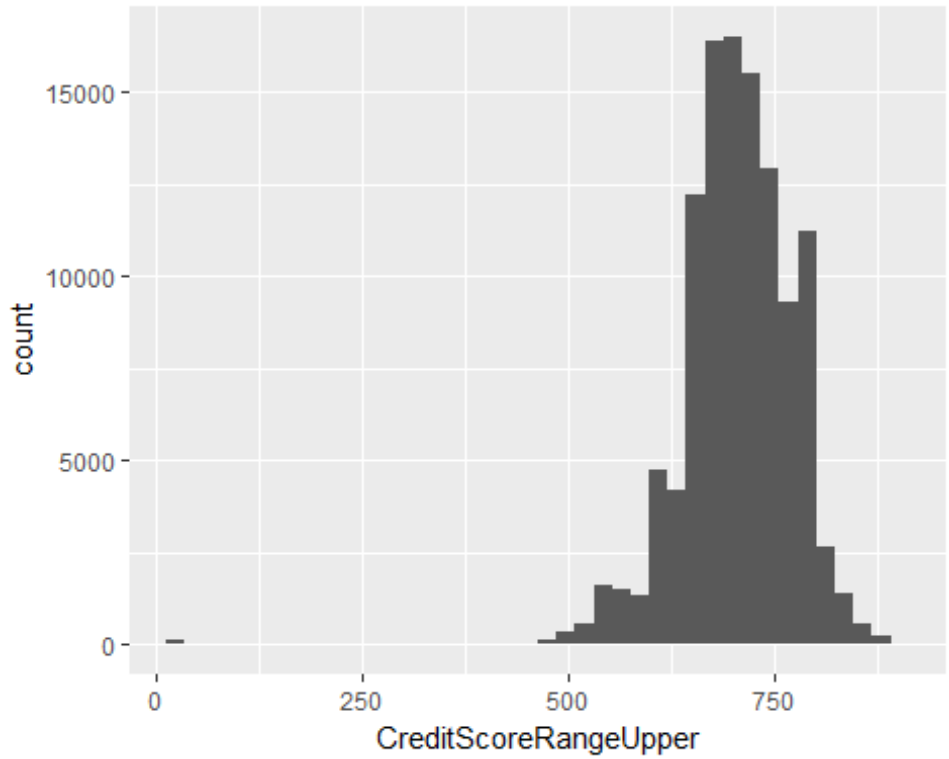
Monthly Loan payment appears to be positively skewed

Analysis of Loan Terms



As we can see above most number of loans are given with 36 month term.

Analysis of CreditScoreRangeUpper



From the above curve we are seeing somewhat normal distribution of the credit score range upper limit. Credit Score Range lower has the similar distribution

Univariate Analysis

What is the structure of your dataset?

There are 113,937 observations in the dataset with 81 variables.

Some of the Continuous Variables are:

EmploymentStatusDuration CreditScoreRangeLower CreditScoreRangeUpper CurrentCreditLines DebtToIncomeRatio StatedMonthlyIncome BorrowerAPR BorrowerRate LoanOriginalAmount MonthlyLoanPayment

Some of the Discrete Variables are:

BorrowerState Occupation EmploymentStatus IsBorrowerHomeowner IncomeRange IncomeVerifiable Term ListingCategory LoanStatus LoanOriginationDate

What is/are the main feature(s) of interest in your dataset?

There are a number of features that will give insights to the profile of the loan and borrower. I think the main ones for borrowers are credit score, income and the ones for the loans are loan amount, interest rate, and term.

What other features in the dataset do you think will help support your investigation into your feature(s) of interest?

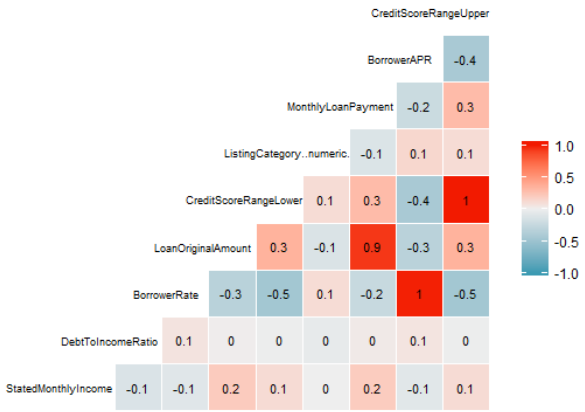
Loan category and status can help determine how the loans are being used and what loans are current, defaulted, delinquent etc (Bad Loans). Also, we can leverage Employment status and explore the impact on Loan.

Did you create any new variables from existing variables in the dataset?

I created a new categorical variables that determines a loan is a Bad Loan or not. I grouped the loans with status of "Defaulted" , "Chargedoff", "Past Due (61-90 days)" , "Past Due (91-120 days)" and "Past Due (>120 days)" as Bad loans

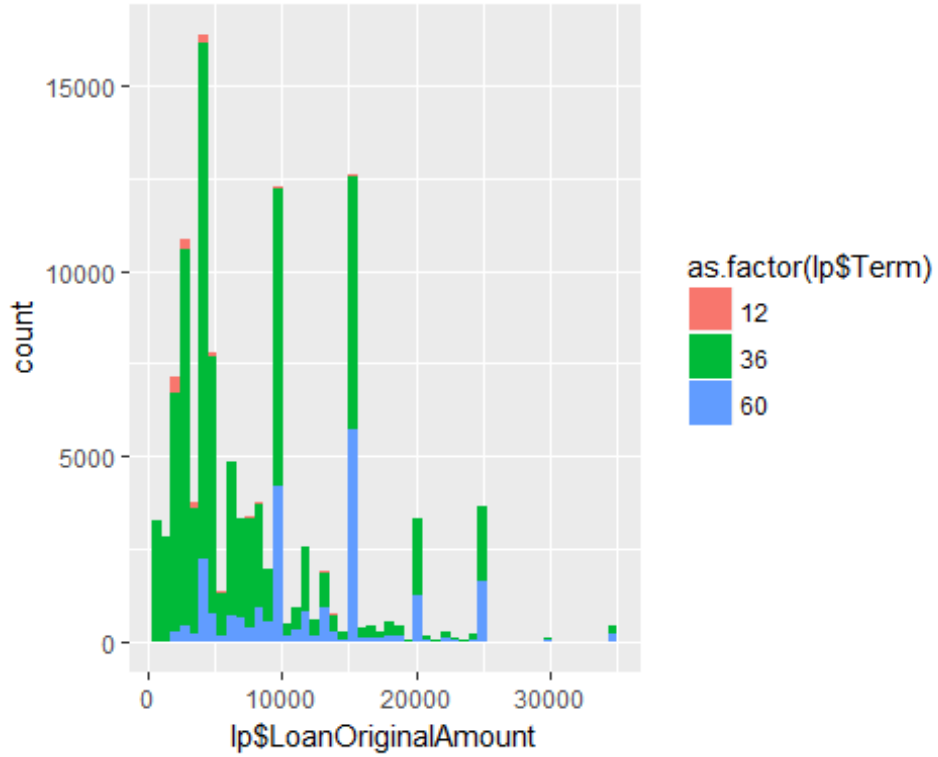
Bivariate plots

Correlation Matrix



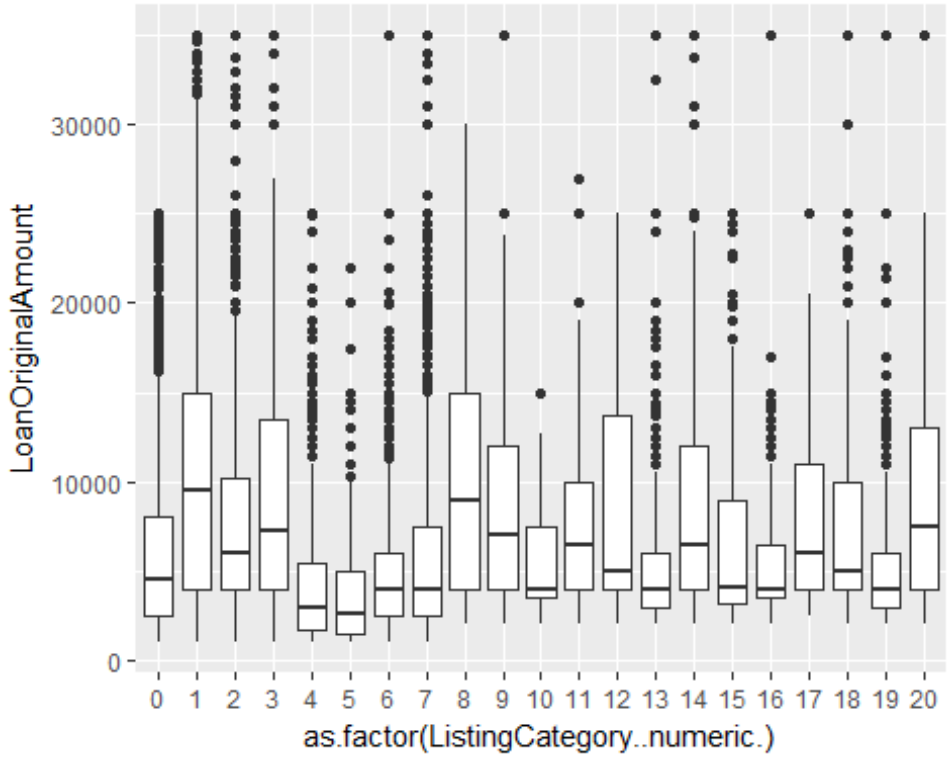
The correlation matrix revealed a few surprising things - I thought there would be a much stronger relationship between interest rate (BorrowerRate) and the credit score (CreditScoreRangeUpper/Lower). At a score of -0.5 it's the strongest correlation out of the selected variables.

Analysis of Loan term on loans



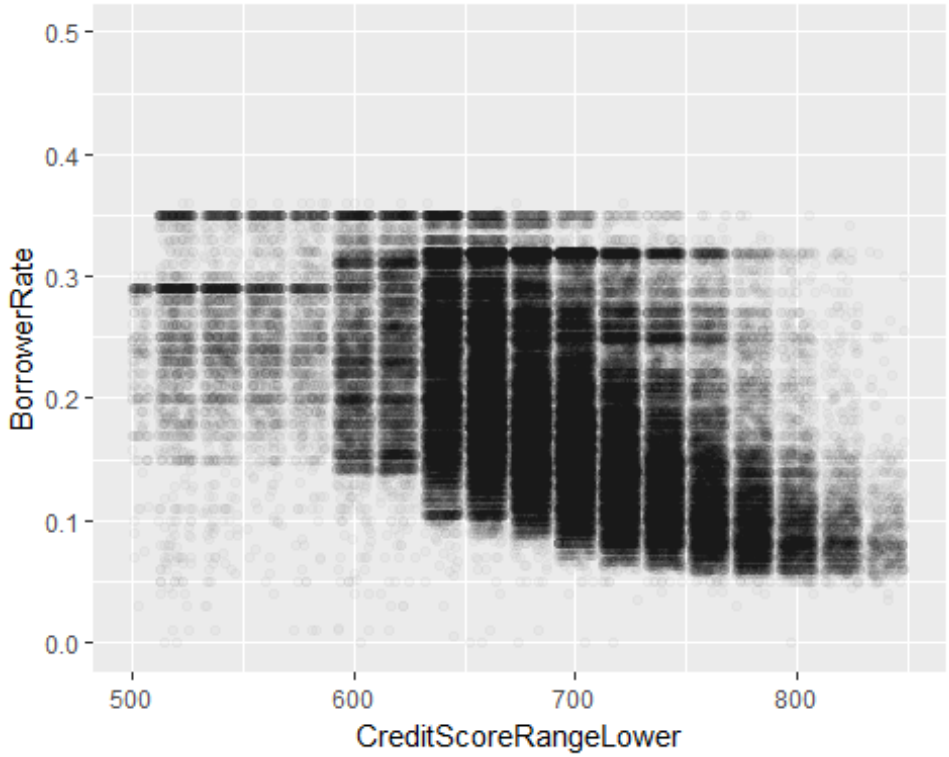
Most of the loan amount are termed at 36 months.

Analysis of Loan Amount with Listing Category



From the above plot we can see that Debt consolidation(1) as expected has the highest loan amounts across categories; however, unexpectedly “Baby and Adoption” (8) category is also the highest (tied) loan amount across categories

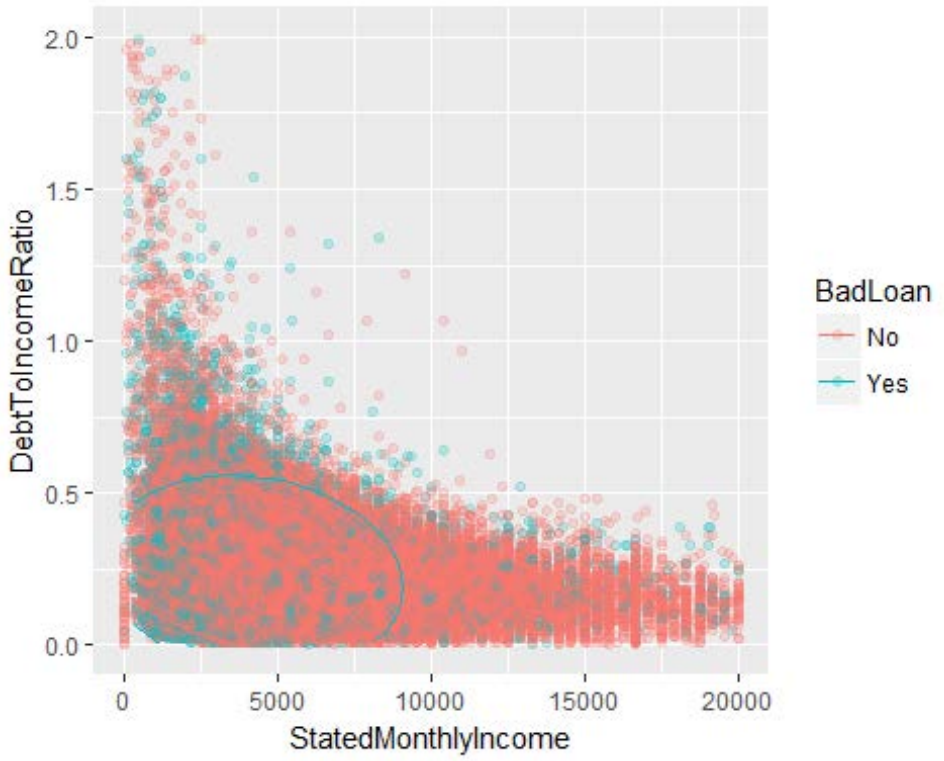
Analysis of Borrower rate with Credit score




```
##
##  Pearson's product-moment correlation
##
## data:  CreditScoreRangeLower and BorrowerRate
## t = -175.17, df = 113340, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4661358 -0.4569730
## sample estimates:
##       cor
## -0.4615667
```

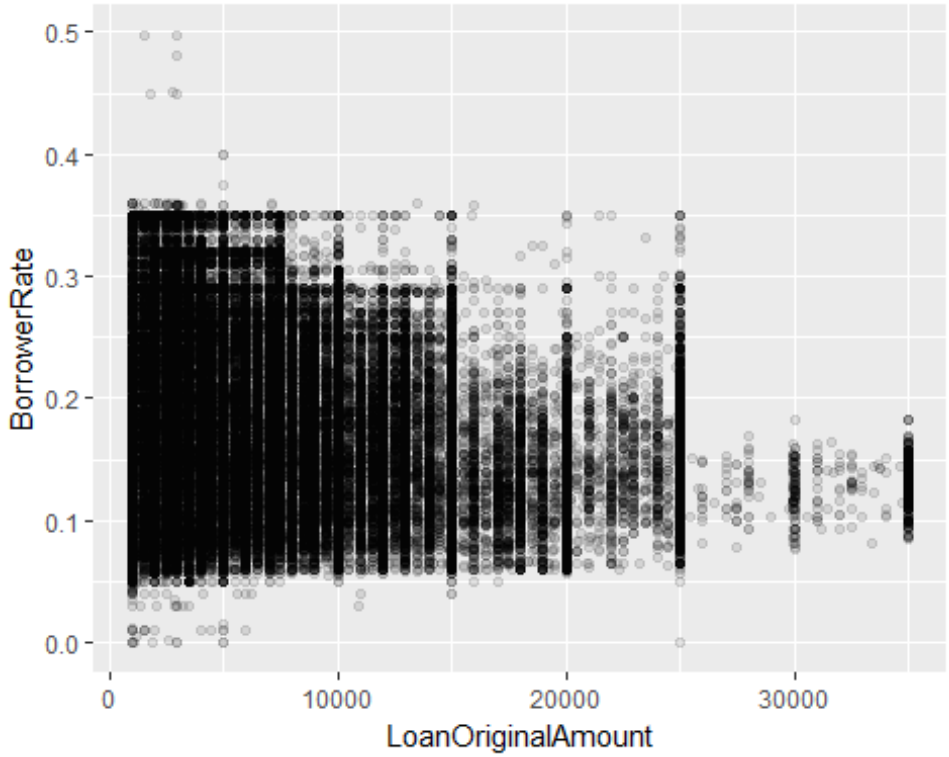
The above plot seems to have negative correlation and it follows the general understanding of - interest rate is lower for more creditworthy borrowers

Analysis of Debt-to-Income ratio with Monthly Income



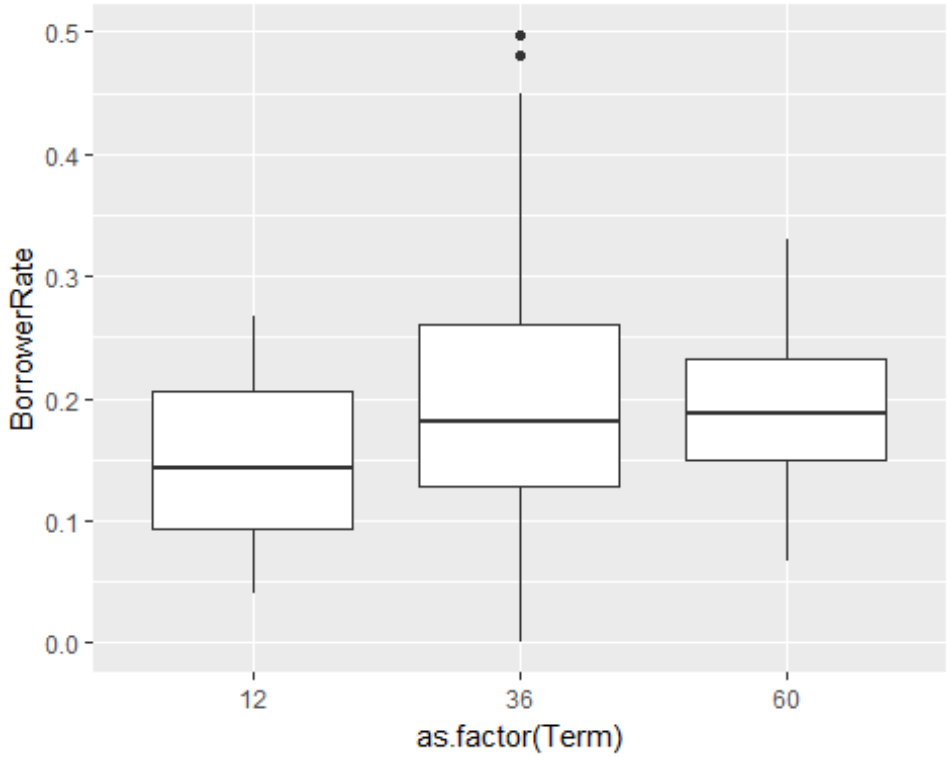
In comparing debt-to-income ratio with a borrower's stated monthly income I was expecting to see a trend that delinquent borrowers would have a lower monthly income and a higher debt-to-income ratio.

Analysis of Borrower rate with Loan amount



There appears to be a negative correlation. There are some large loan amounts with low Borrower rate.

Analysis of Loan term with Borrower rate



I was expecting to see low rate as the term increases but as shown from the above plot small terms have the lowest mean borrower rate. There are other factors in addition to the loan term affecting the borrower rate.

Bivariate Analysis

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?

There's a negative relationship between interest rates and loan amount, the larger the loan, the lower the rate on average. That was mostly due to them having higher credit scores.

Delinquent borrowers would have a lower monthly income and a higher debt-to-income ratio

There is a negative correlation between Credit score and Borrower rate - interest rate is lower for more creditworthy borrowers

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

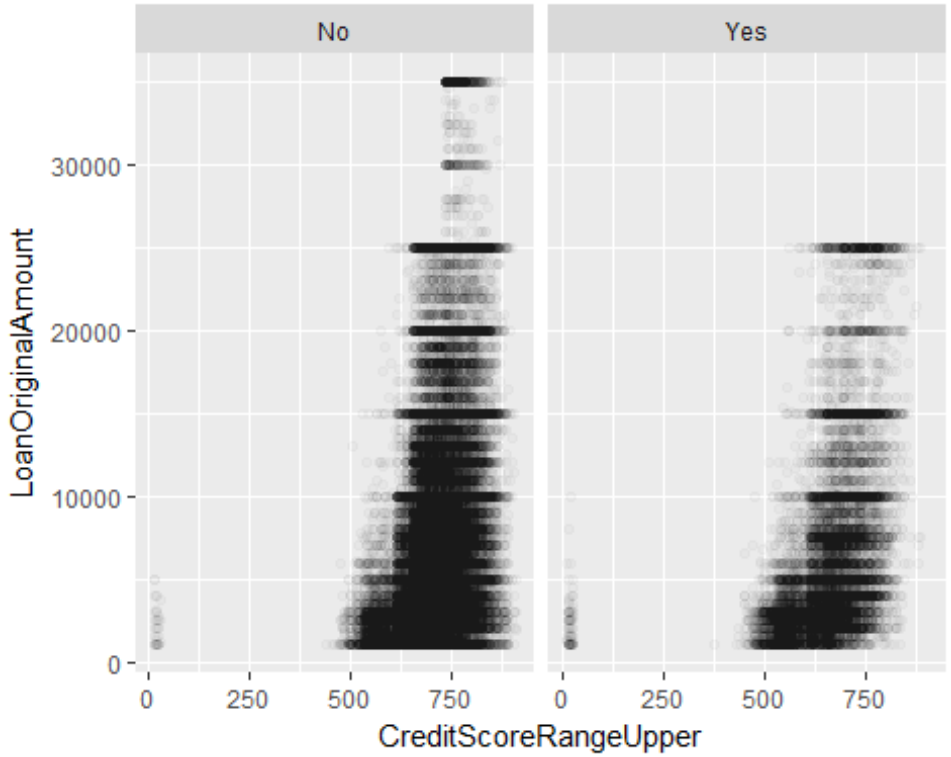
Debt consolidation as expected has the highest loan amounts across categories; however, unexpectedly "Baby and Adoption" category is also the highest (tied) loan amount across categories.

What was the strongest relationship you found?

The strongest relationship I found was between credit score and interest rate, with -0.46. This makes sense since credit score is a rating of the credit-worthiness of the borrower and that should be related to the cost of borrowing (interest rate).

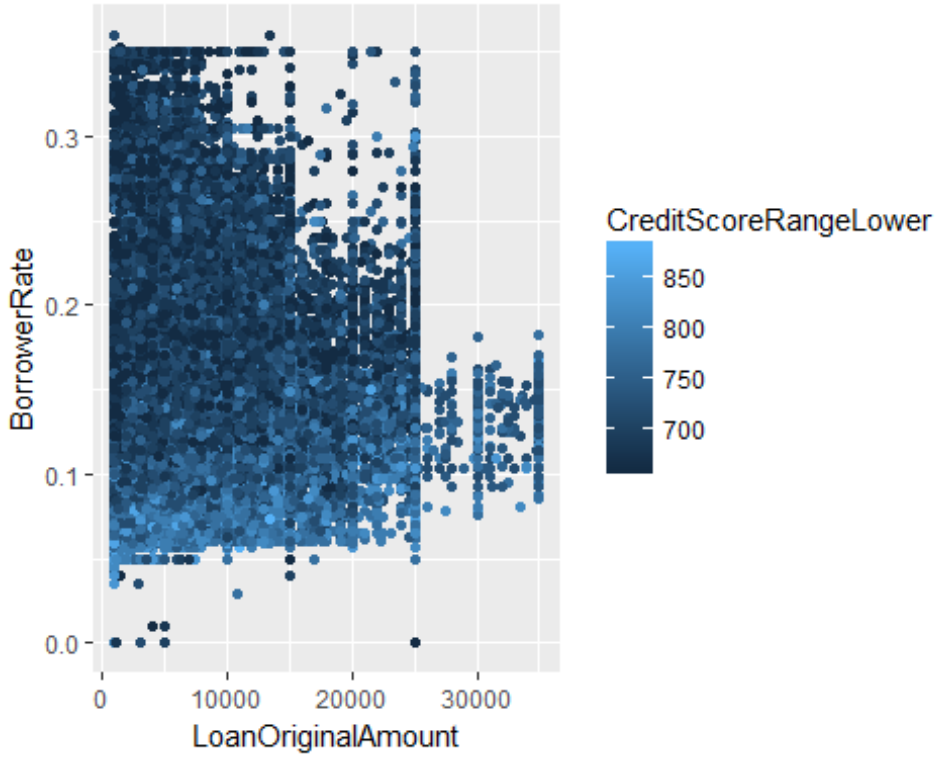
Multivariate Plots

Analysis of credit score with loan amount and BadLoans



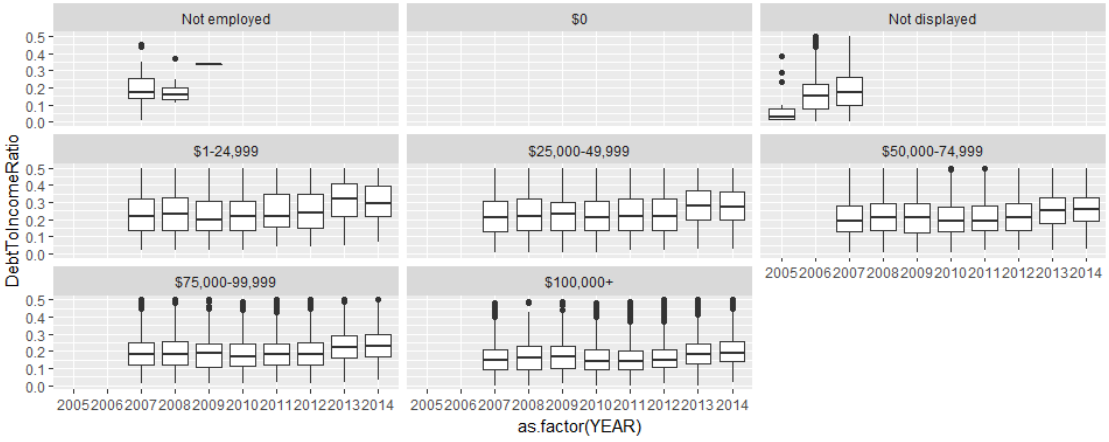
In the above comparisons, both scatterplots show a borrower's credit score against their loan amounts.The higher concentration of Bad Loans have lower credit scores, and they also tend to borrow less money, under $10,000.

Analysis of Loan Amount with Borrower rate and Credit score



The borrowers with high credit scores have lower interest rates and larger loan amounts. I subsetted only credit scores from 660 (1st quartile) and above for better visual presentation.

Analysis of Debt to Income ratio with income range and Year



Comparing DI ratio, most of the borrowers seem to have DI ratio close to 20-30% with a uptick in the most recent years. The $100k + income range have noticeably lower DI ratio at around 15-20%. The variance in DI ratio is reduced as incomes increase.

Multivariate Analysis

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?

When looking at the loan amounts vs cost (interest rate), the credit scores demarcated borrowers by credit worthiness.

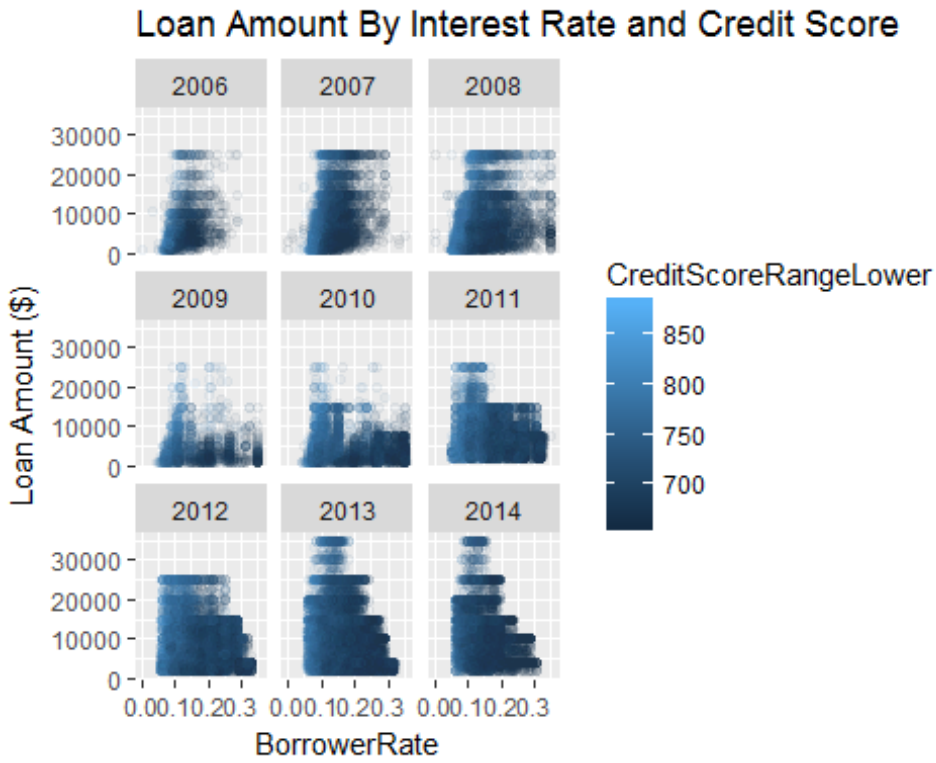
Investigating further, I looked at DI ratio and observed the higher the income, the lower the percentage of debt.

Were there any interesting or surprising interactions between features?

The higher concentration of Bad Loans have lower credit scores, and they also tend to borrow less money, under $10,000.Also, we can see that lower loan amount for borrowers with much higher credit scores. Based on previous analysis, higher scores provide lower rates, which would tend to allow the borrower to gain access to a higher loan amount. And although the average loan amount for loans in good standing are higher, there is still a high concentration of loans under $20,000 and with credit scores over 700.

Final Plots

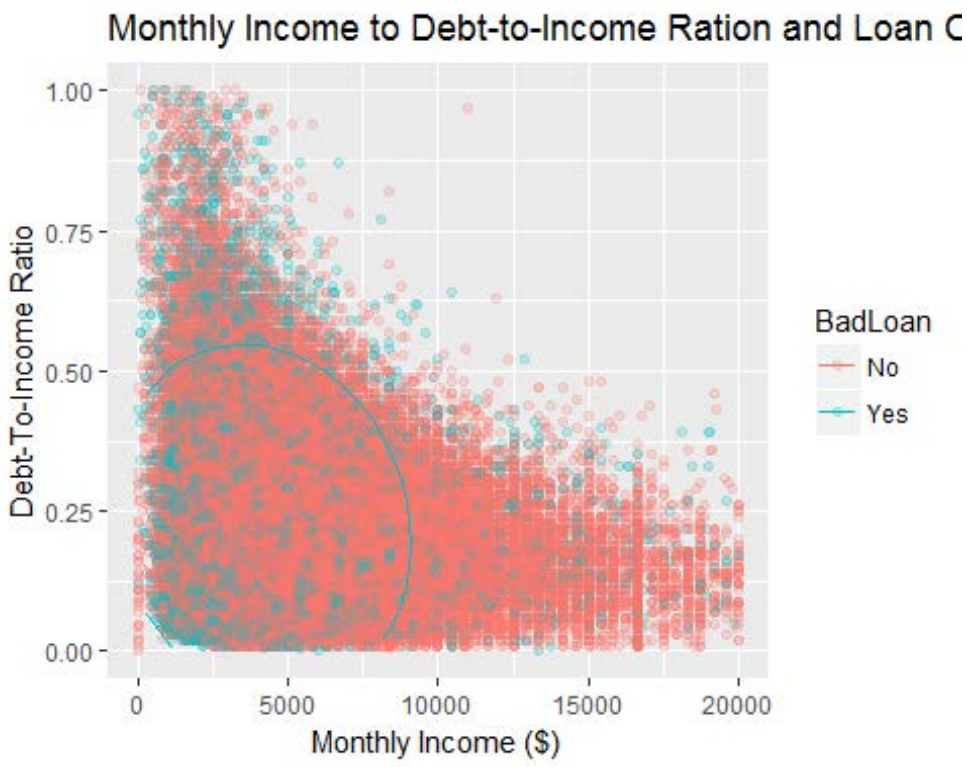
PLOT One



Description One

The borrowers with high credit scores have lower interest rates and larger loan amounts. I subsetted only credit scores from 650 (1st quartile) and above for better visual presentation. And this shows that as the lending platform matured, the overal risk exposure increased. In 2014, much more blue (credit score ~700) borrowers.

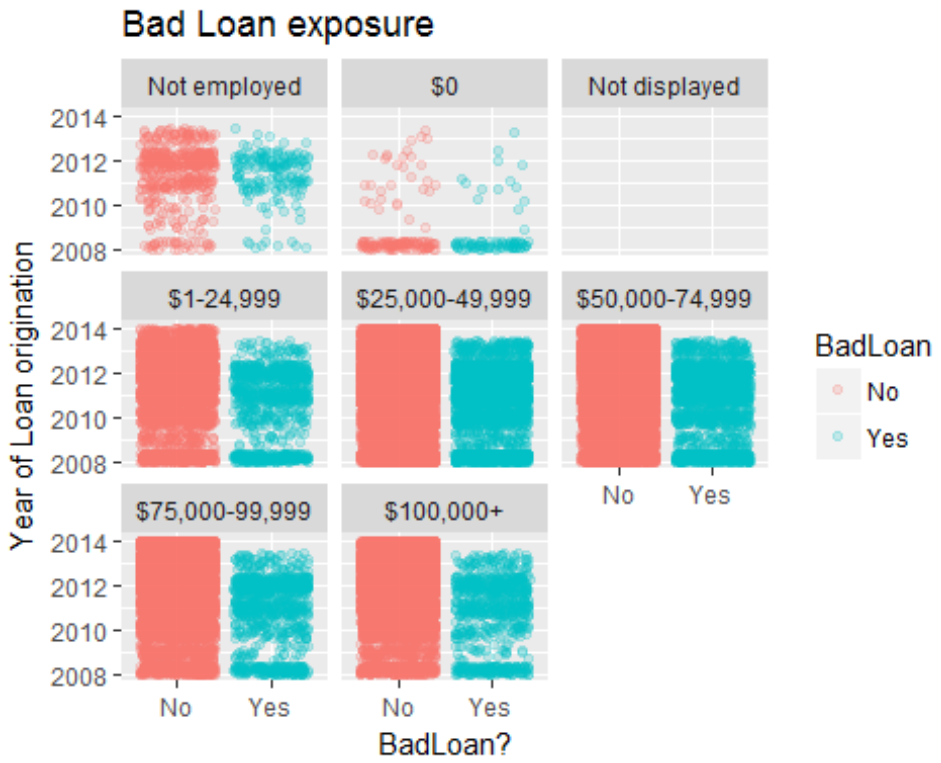
Plot two



In the above plot we examine the stated monthly incomes of borrowers and their debt-to-income ratio. This data was then visually categorized by Loan category (Good Loan or Bad Loan).

High concentration of Bad Loan borrowers earn less than $2500 a month but have a relatively low debt-to-income ratio of under 0.50 (or 50%). The plot also suggests a negative correlation between monthly income and debt-to-income ratio, i.e the more a borrower makes in monthly income the lower their debt-to-income ratio; However, this does not guarantee the borrower will not go into delinquency.

Plot Three



It appears the majority of bad loans are clustered in 2008 and 2011-2012 for the borrowers with income ranges $25k-50k and $50-75k. This leads us to beleive that recession would have impacted the Loan and would have increased delinquency.

Reflection

The data set had nearly 114,000 loans from Nov 2005 - March 2014. Over the course of those years, Prosper has made almost $1 trillion dollars in loans ($949,894,347 to be exact). The difficulties I had with the data mainly stemmed from understanding the variables and then selecting the appropriate ones to analyze. Another persistent issue was overplotting on scatterplots, a number of techniques were used across multiple plots.

The general analysis revealed areas of interests such as negative corelation between credit score and borrowing rate which brough up any curious questions concerning delinquent borrowers .Also, we came to know that he main loan exposures are in debt consolidation.

Additional data would also enhance this dataset. Having the borrower's age and sex would allow analysis to possibly discover trends among men and women or young and old.