Exploratory Analysis Project - Prosper Loan

==========

1. Introduction

Prosper Marketplace,Inc., which was founded in 2005, is America's first peer-to-peer lending marketplace, Borrowers request personal loans on Prosper and investors (individual or institutional) can fund anywhere from $2,000 to $35,000 per loan request.

This report explores the Prosper's loan data, which contains 113,937 loans with 81 variables on each loan, and it was Last updated on 03/11/2014. The dataset will be used for exploring two main ideas 1) Investors' preferences 2)The relationship between borrowers'characteristics and the risks.

```
## [1] 113937     36
```

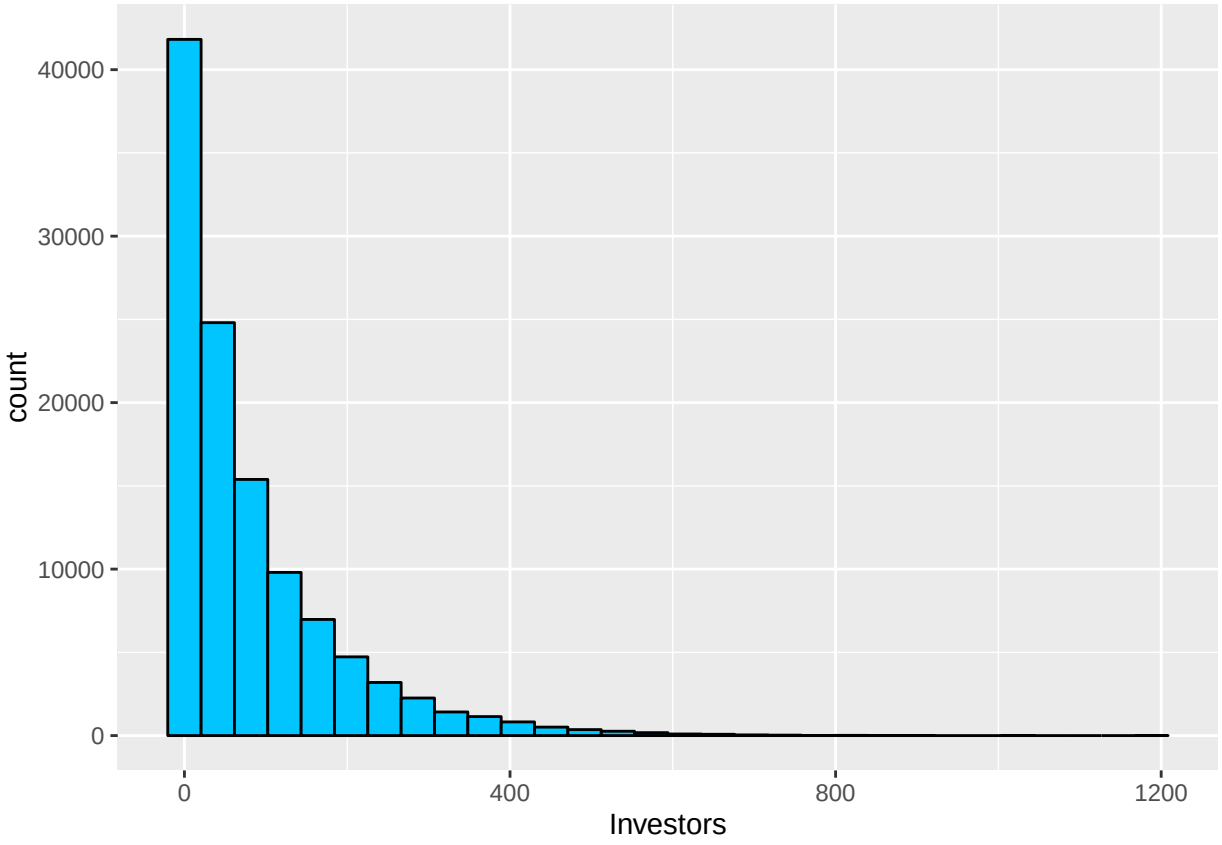
```
## 'data.frame':    113937 obs. of  36 variables:
##  $ ListingCreationDate               : Factor w/ 113064 levels "2005-11-09 20:44:28.847000000",..: 14
##  $ CreditGrade                       : Factor w/ 9 levels "","A","AA","B",..: 5 1 8 1 1 1 1 1 1 1 ..
##  $ Term                              : int  36 36 36 36 36 60 36 36 36 36 ...
##  $ LoanStatus                        : Factor w/ 12 levels "Cancelled","Chargedoff",..: 3 4 3 4 4 4 4
##  $ BorrowerRate                      : num  0.158 0.092 0.275 0.0974 0.2085 ...
##  $ ProsperRating..Alpha.             : Factor w/ 8 levels "","A","AA","B",..: 1 2 1 2 6 4 7 5 3 3 ..
##  $ ListingCategory..numeric.         : int  0 2 0 16 2 1 1 2 7 7 ...
##  $ BorrowerState                     : Factor w/ 52 levels "","AK","AL","AR",..: 7 7 12 12 25 34 18 0
##  $ Occupation                        : Factor w/ 68 levels "","Accountant/CPA",..: 37 43 37 52 21 43
##  $ EmploymentStatus                  : Factor w/ 9 levels "","Employed",..: 9 2 4 2 2 2 2 2 2 2 ...
##  $ EmploymentStatusDuration          : int  2 44 NA 113 44 82 172 103 269 269 ...
##  $ IsBorrowerHomeowner               : Factor w/ 2 levels "False","True": 2 1 1 2 2 2 1 1 2 2 ...
##  $ CreditScoreRangeLower             : int  640 680 480 800 680 740 680 700 820 820 ...
##  $ CreditScoreRangeUpper             : int  659 699 499 819 699 759 699 719 839 839 ...
##  $ FirstRecordedCreditLine           : Factor w/ 11586 levels "","1947-08-24 00:00:00",..: 8639 6617
##  $ CurrentCreditLines                : int  5 14 NA 5 19 21 10 6 17 17 ...
##  $ OpenCreditLines                   : int  4 14 NA 5 19 17 7 6 16 16 ...
##  $ TotalCreditLinespast7years        : int  12 29 3 29 49 49 20 10 32 32 ...
##  $ OpenRevolvingAccounts             : int  1 13 0 7 6 13 6 5 12 12 ...
##  $ InquiriesLast6Months              : int  3 3 0 0 1 0 0 3 1 1 ...
##  $ CurrentDelinquencies              : int  2 0 1 4 0 0 0 0 0 0 ...
##  $ AmountDelinquent                  : num  472 0 NA 10056 0 ...
##  $ DelinquenciesLast7Years           : int  4 0 0 14 0 0 0 0 0 0 ...
##  $ PublicRecordsLast10Years          : int  0 1 0 0 0 0 0 1 0 0 ...
##  $ BankcardUtilization               : num  0 0.21 NA 0.04 0.81 0.39 0.72 0.13 0.11 0.11 ...
##  $ AvailableBankcardCredit           : num  1500 10266 NA 30754 695 ...
##  $ TotalTrades                       : num  11 29 NA 26 39 47 16 10 29 29 ...
##  $ TradesNeverDelinquent..percentage.: num  0.81 1 NA 0.76 0.95 1 0.68 0.8 1 1 ...
##  $ TradesOpenedLast6Months           : num  0 2 NA 0 2 0 0 0 1 1 ...
##  $ DebtToIncomeRatio                 : num  0.17 0.18 0.06 0.15 0.26 0.36 0.27 0.24 0.25 0.25 ...
##  $ IncomeRange                       : Factor w/ 8 levels "$0","$1-24,999",..: 4 5 7 4 3 3 4 4 4 4 .
##  $ IncomeVerifiable                  : Factor w/ 2 levels "False","True": 2 2 2 2 2 2 2 2 2 2 ...
##  $ StatedMonthlyIncome               : num  3083 6125 2083 2875 9583 ...
##  $ LoanOriginalAmount                : int  9425 10000 3001 10000 15000 15000 3000 10000 10000 10000
```

```
##  $ MonthlyLoanPayment                : num  330 319 123 321 564 ...
##  $ Investors                         : int  258 1 41 158 20 1 1 1 1 1 ...
```

2. Univariate Plots Section

2-1 Investors

The investors can use the invest tool to browse loans by multiple criteria and fund the selected loan requests, and there might be several investors per loan request, so let's begin with the 'Investors' variable.

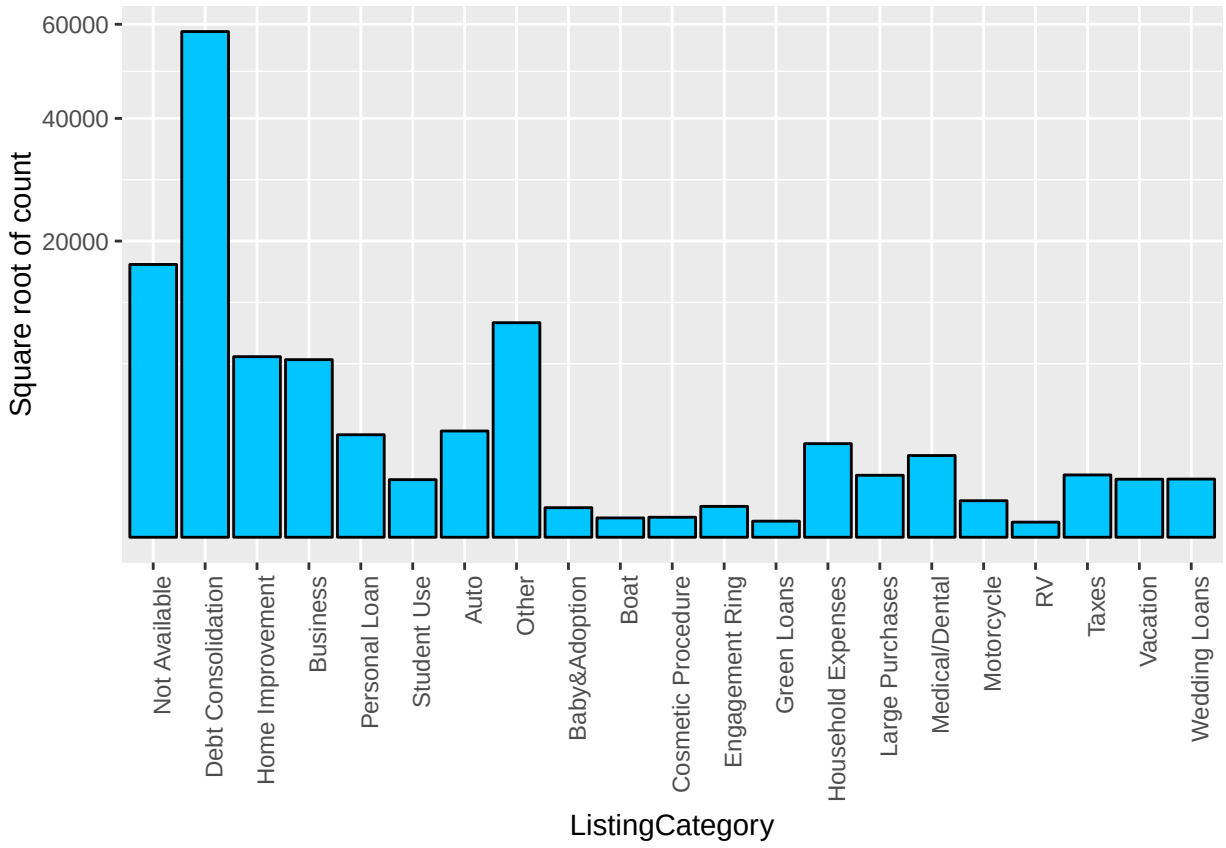



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    2.00   44.00   80.48  115.00 1189.00
```

Most loans (75%) received the funding from less than 115 investors. Now, let's dig deeper into the characteristics of loans and borrowers.

2-2 Listing Category

The category of the listing is what the borrower selected as the purpose of the loan when posting their listing. In order to make the plot readable, I converted the numeric values to the corresponding factors(labels). Moreover, I also take the square root of the count.

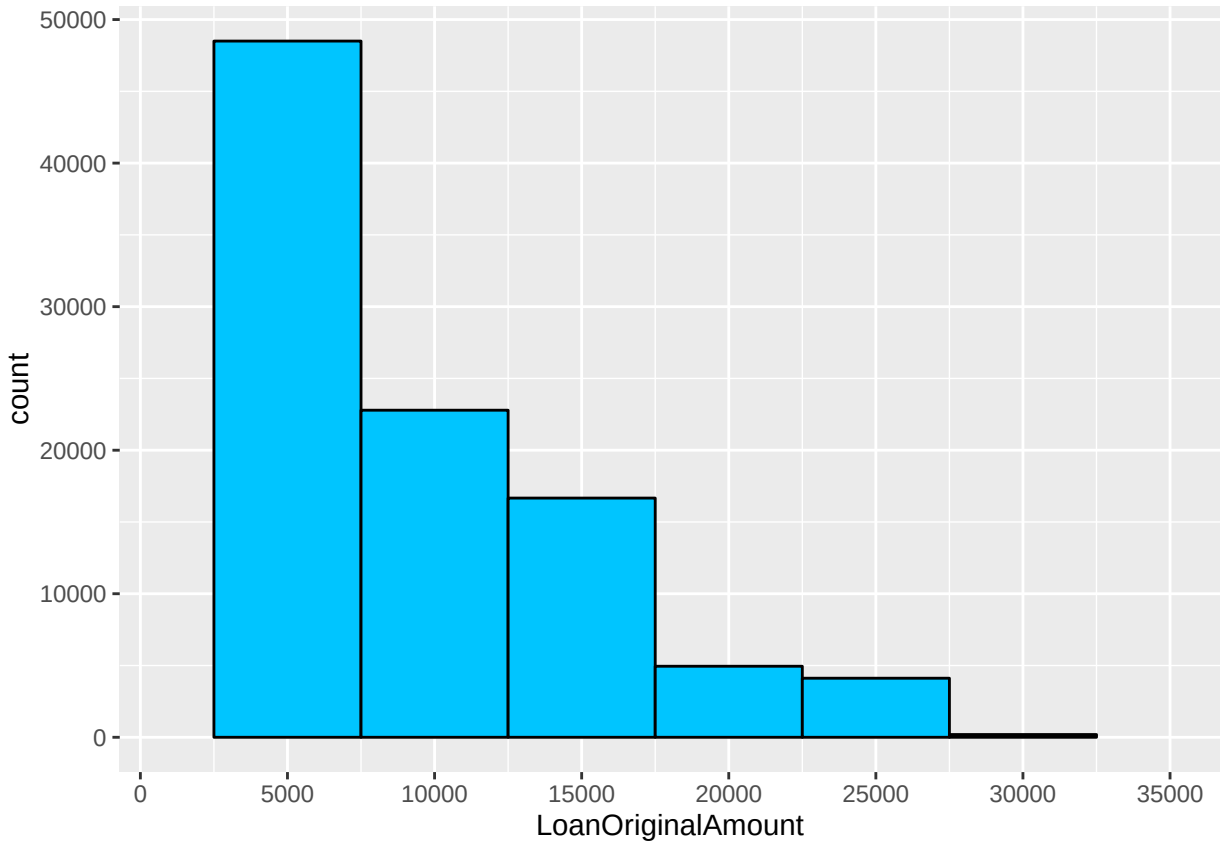



```
##
##      Not Available Debt Consolidation   Home Improvement
##       0.1488980753       0.5117564970       0.0652378069
##           Business      Personal Loan        Student Use
##       0.0630962725       0.0210203885       0.0066352458
##               Auto              Other      Baby&Adoption
##       0.0225738785       0.0921035309       0.0017465792
##               Boat Cosmetic Procedure    Engagement Ring
##       0.0007460263       0.0007986870       0.0019045613
##        Green Loans Household Expenses    Large Purchases
##       0.0005178300       0.0175184532       0.0076884594
##     Medical/Dental         Motorcycle                 RV
##       0.0133582594       0.0026681412       0.0004563926
##              Taxes            Vacation      Wedding Loans
##       0.0077674504       0.0067405672       0.0067668975
```

Debt Consolidation is the commonest purpose of the loan requests, more than 50% of the borrowers use the funding for debt consolidation.

2-3 LoanOriginalAmount

How about the loan amounts? Let's taka a look at the origination amount of the loans.



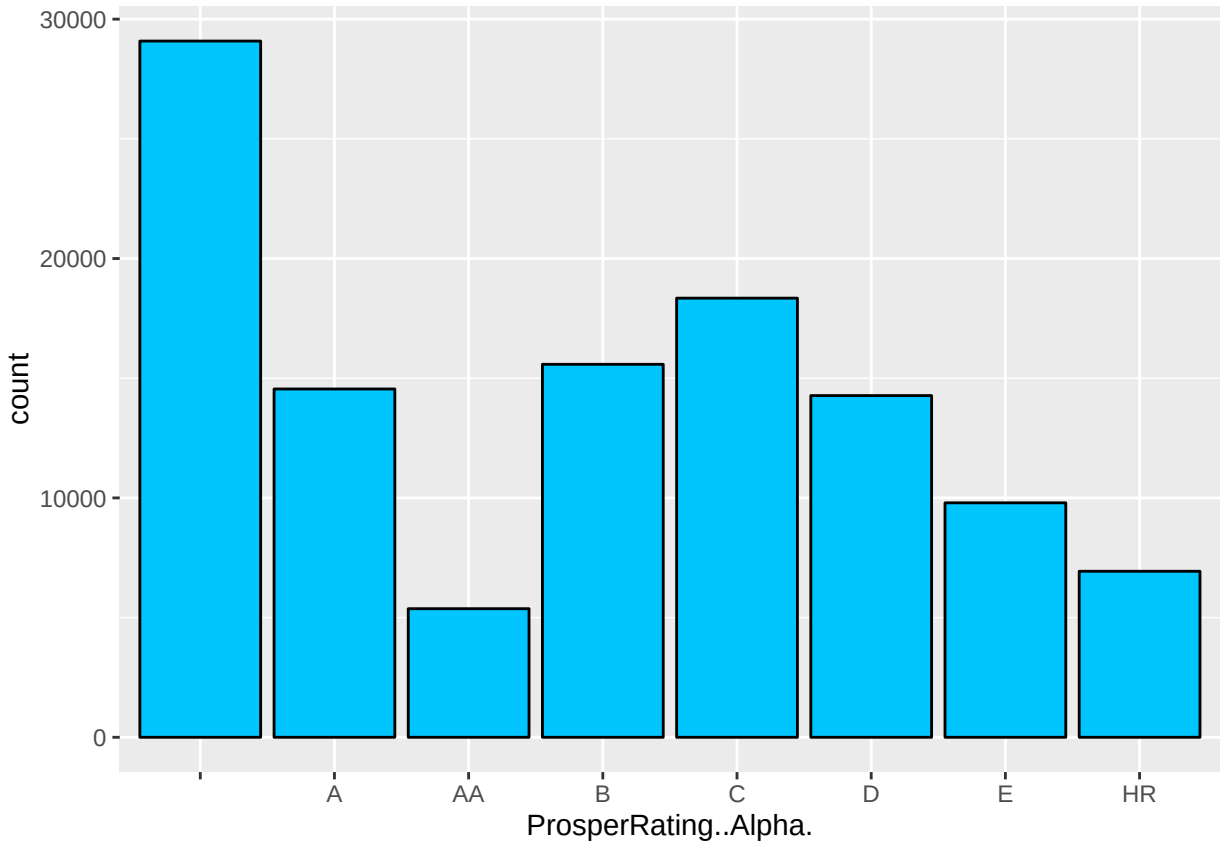
Most of the loan original amount values are between $1,000 and $15,000

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

The mininmum is $1,000, the maximum is $35,000, and the median as $6.500

2-4 Prosper Ratings

Prosper Ratings, from lowest-risk to highest-risk, are labeled AA, A, B, C, D, E, and HR ("High Risk"). Applicable for loans originated after July 2009.

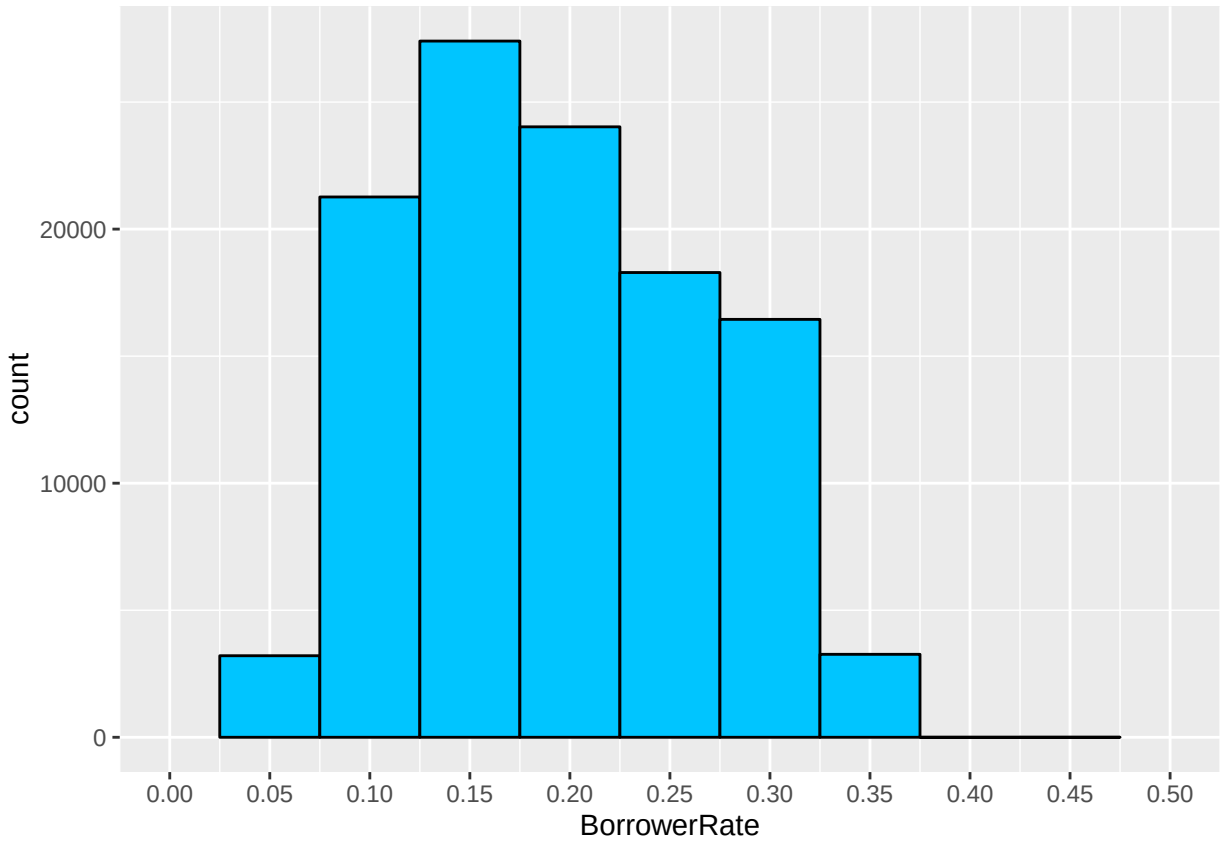



```
##           A    AA     B     C     D     E    HR
## 29084 14551  5372 15581 18345 14274  9795  6935
```

Prosper Ratings of A,B and C are three major ratings among the loans.

2-5 Interest Rate

The Borrower's interest rate for this loan.



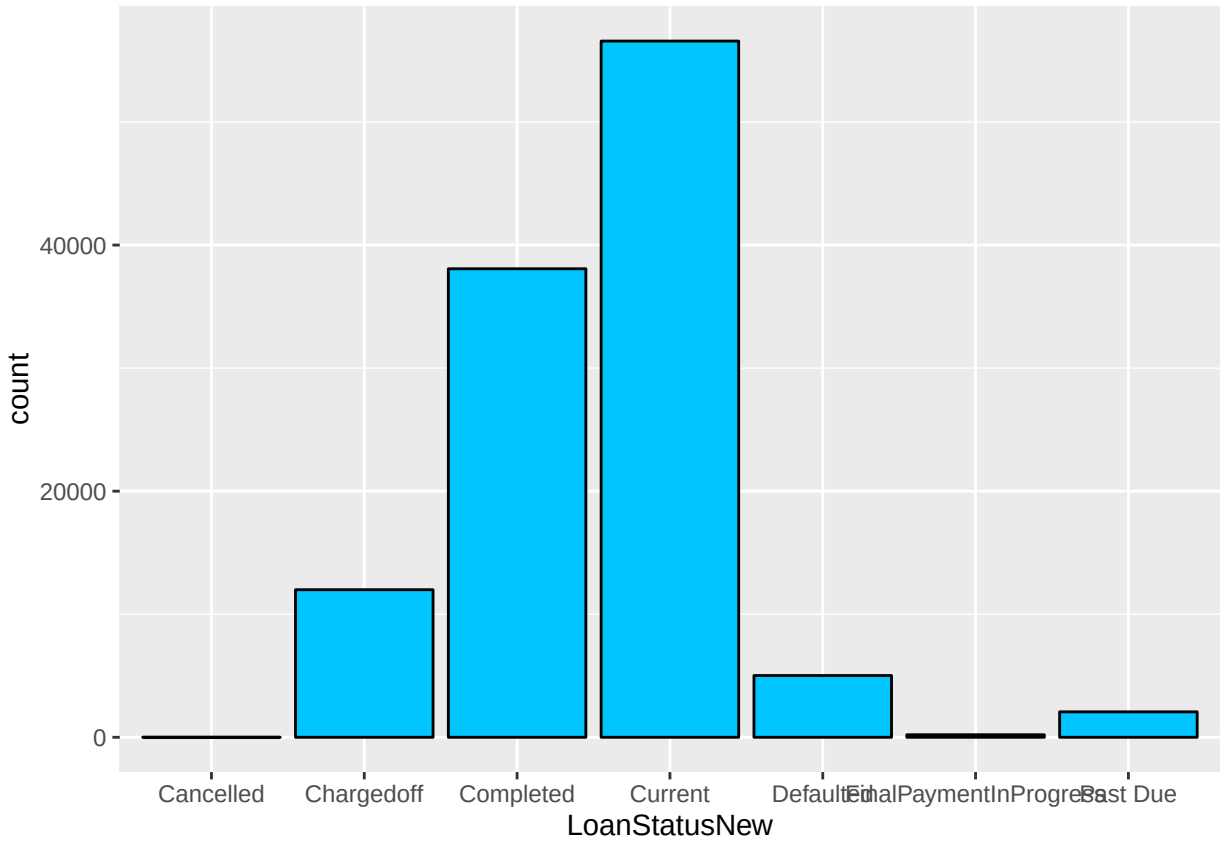

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.1340  0.1840  0.1928  0.2500  0.4975
```

2-6 Loan Status

The current status of the loan: Cancelled, Chargedoff, Completed, Current, Defaulted, FinalPaymentInProgress, PastDue.

```
##
##              Cancelled             Chargedoff              Completed
##                      5                  11992                  38074
##                Current              Defaulted FinalPaymentInProgress
##                  56576                   5018                    205
##   Past Due (>120 days)   Past Due (1-15 days)  Past Due (16-30 days)
##                     16                    806                    265
##  Past Due (31-60 days)  Past Due (61-90 days) Past Due (91-120 days)
##                    363                    313                    304
```

The PastDue Status was subdivided into six factors by the delinquency periods, and I would like to add a new row "LoanStatusNew" which all the PastDue Status will be grouped by only one factor so that I can use it to find out the correlations between risk and other variables.




```
##
##             Cancelled              Chargedoff               Completed
##                     5                   11992                   38074
##               Current               Defaulted FinalPaymentInProgress
##                 56576                    5018                     205
##              Past Due
##                  2067
```

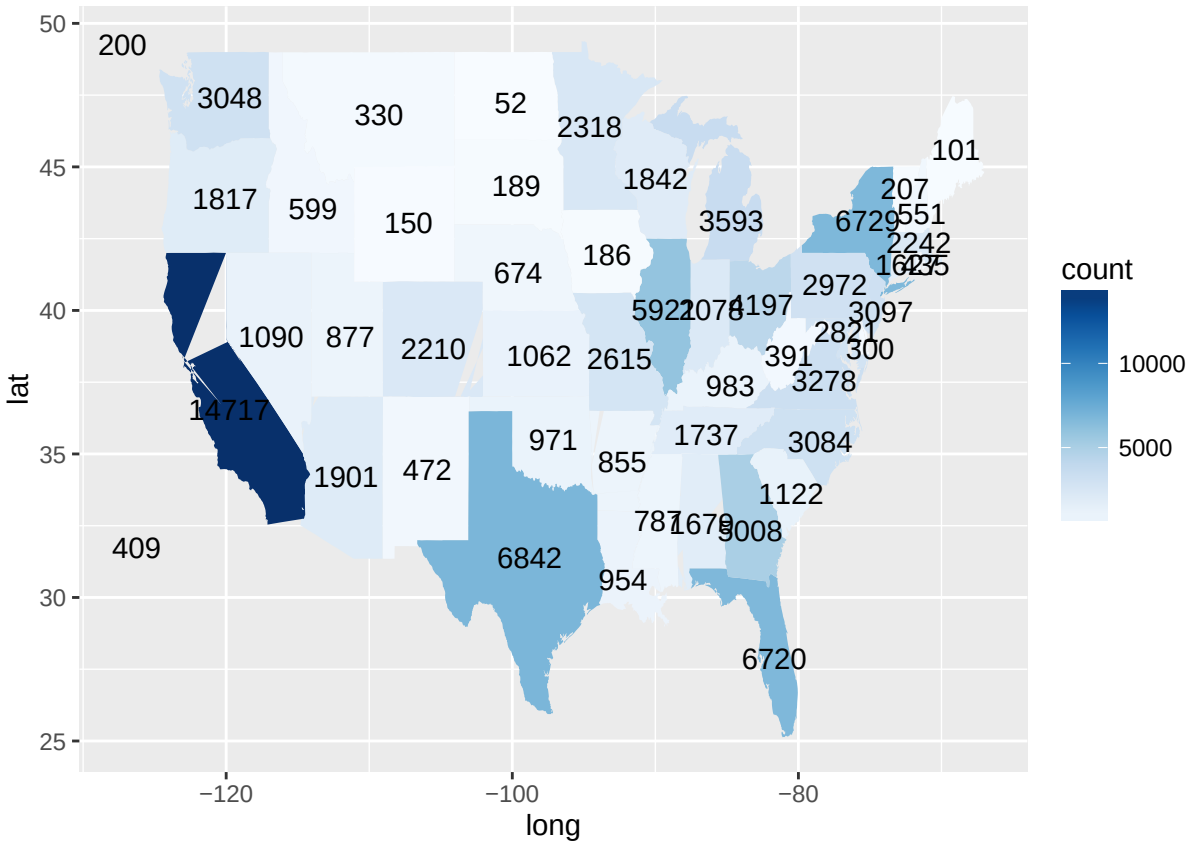
```
##
##             Cancelled              Chargedoff               Completed
##          0.0000438839          0.1052511476          0.3341671274
##               Current               Defaulted FinalPaymentInProgress
##          0.4965551138          0.0440418828          0.0017992399
##              Past Due
##          0.0181416046
```

More than 10% of the loans are charge-off, 4% are defaulted, and nearly 2% of the loans are past due.

Next, I would like to know more about the borrowers.

2-7 Borrowers' States

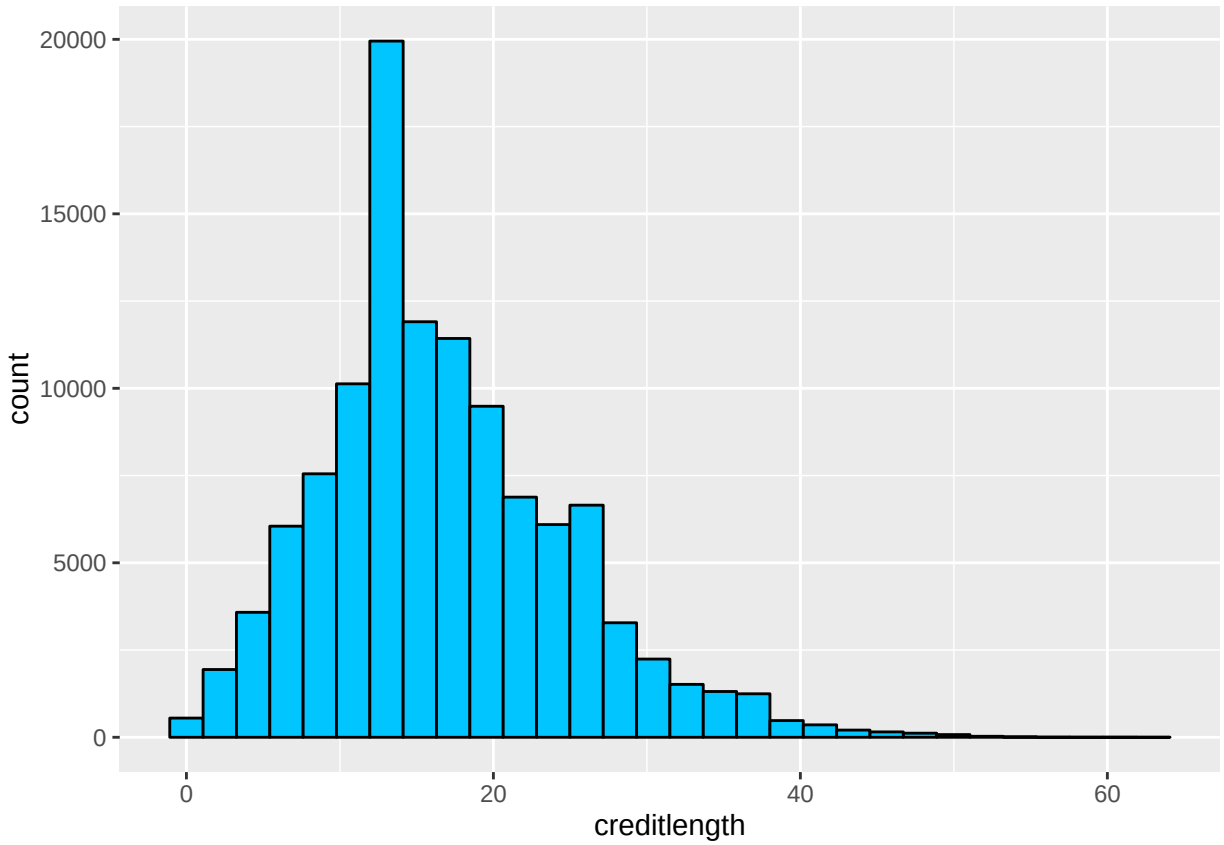
The state of the address of the borrower at the time the Listing was created, and a heatmap might be a perfect fit for this geographical variable.



With the heatmap above, we can see that California has the largest number of borrowers.

2-8 Length of Credit History

To have a roughly understanding about the length of borrowers' credit history, I would like to create a new variable, creditlength, by using the difference between two available variables-FirstRecordedCreditLine and ListingCreationDate. As mentioned before, I just want a roughly understanding about the length and see if there is any value for further exlporing, so I only use the year in the variables to calculate. By doing so, we can know how many years the borrowers'lengths of credit history are.

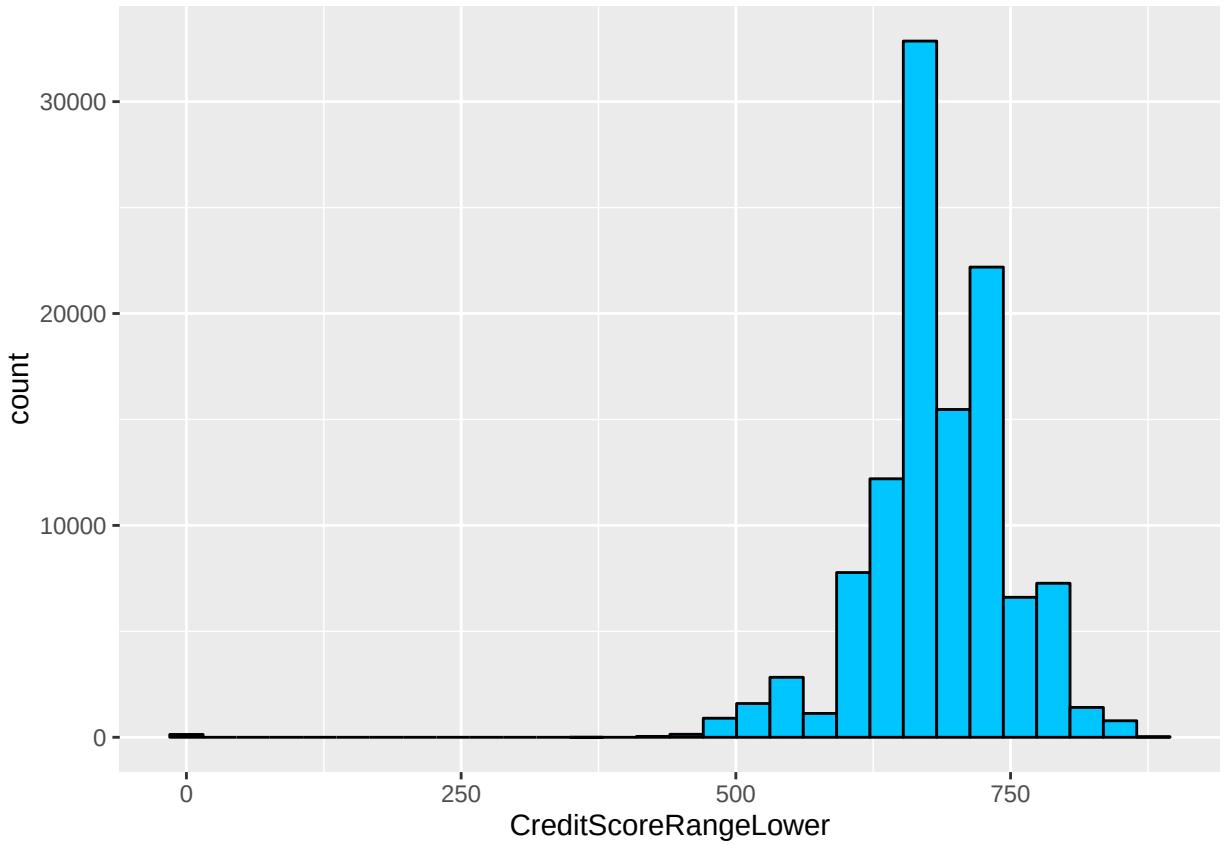



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00   11.00   16.00   16.68   21.00   63.00     697
```

The length of credit history shows a normal distribution , and more than 50% of the borrowers have more than 16-year-long credit history.

2-9 Credit Scores

Credit score represents the creditworthiness of a borrower, and it is also used by banks or other lenders to evaluate the potential risk of a loan. Here, I am going to use CreditScoreRangeLower, the lower value representing the range of the borrower's credit score as provided by a consumer credit rating agency, to see the borrowers' credit scores.



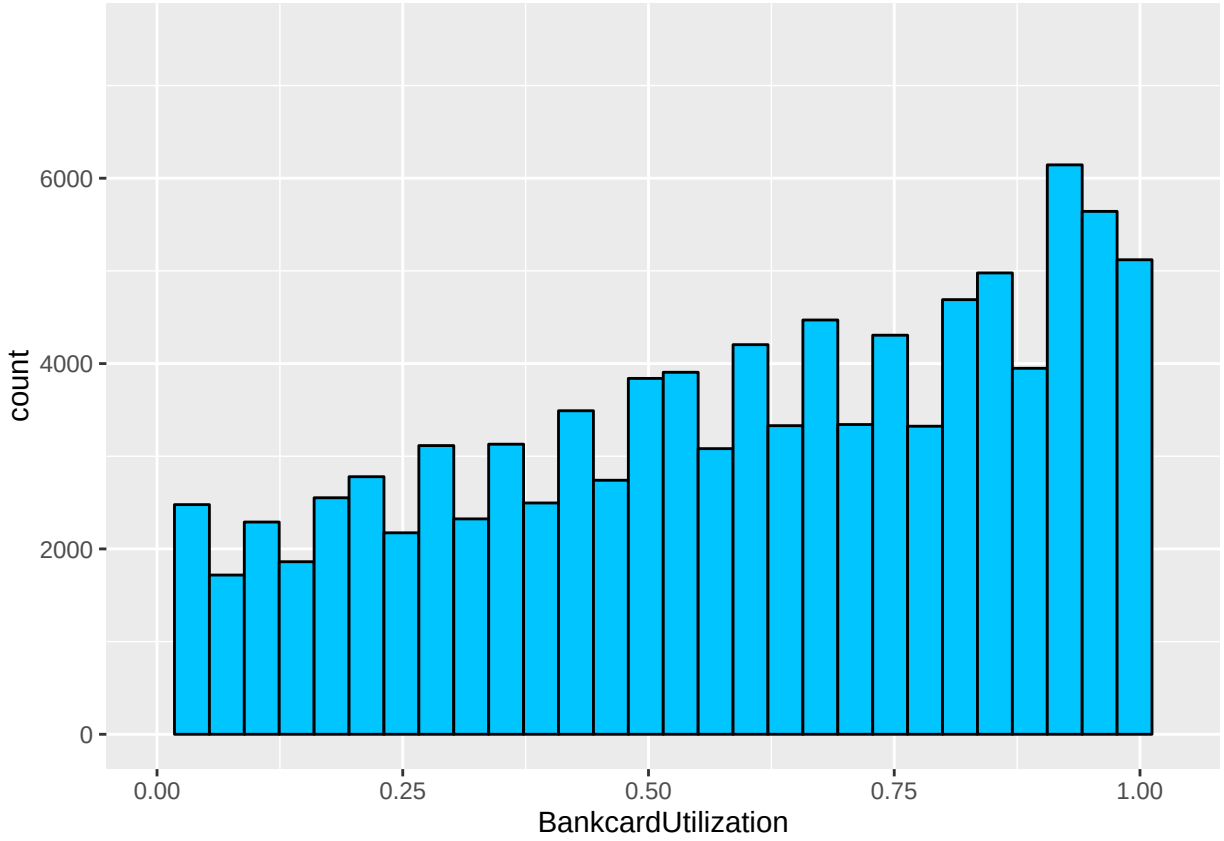

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0   660.0   680.0   685.6   720.0   880.0     591
```

The credit scores are nearly normally distributed, with means of 685.6, which is much better than I expected.

2-10 BankcardUtilization

Now, I would like to look into the credit lines closer. Let's begin with the utilization of the revolving credits. The variable shows the percentage of available revolving credit that is utilized at the time the credit profile was pulled.

To remove the 1% outliers, I use the quantile function and create the plot with 99% of the data.

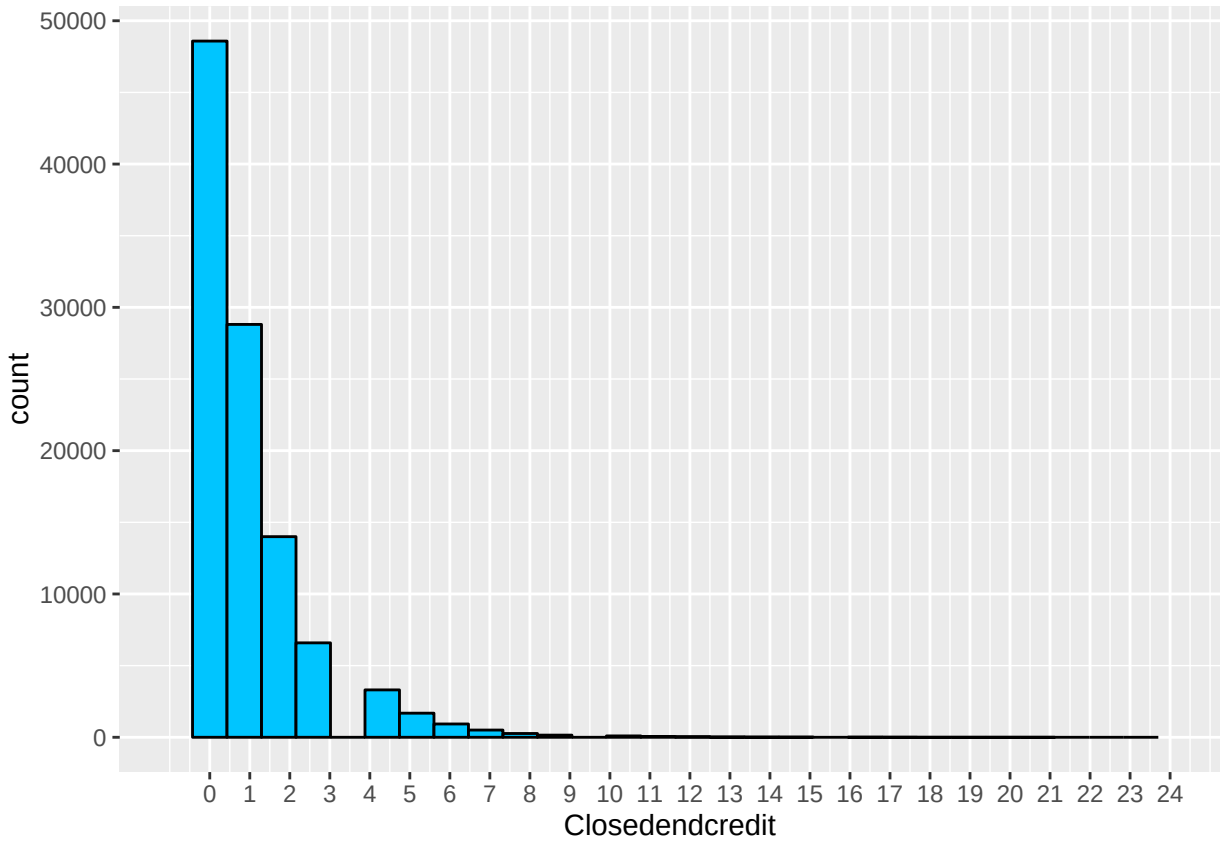



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   0.310   0.600   0.561   0.840   5.950    7604
```

More than a half of the borrowers use 60% of the revolving credits.

2-11 Closed-end credit line

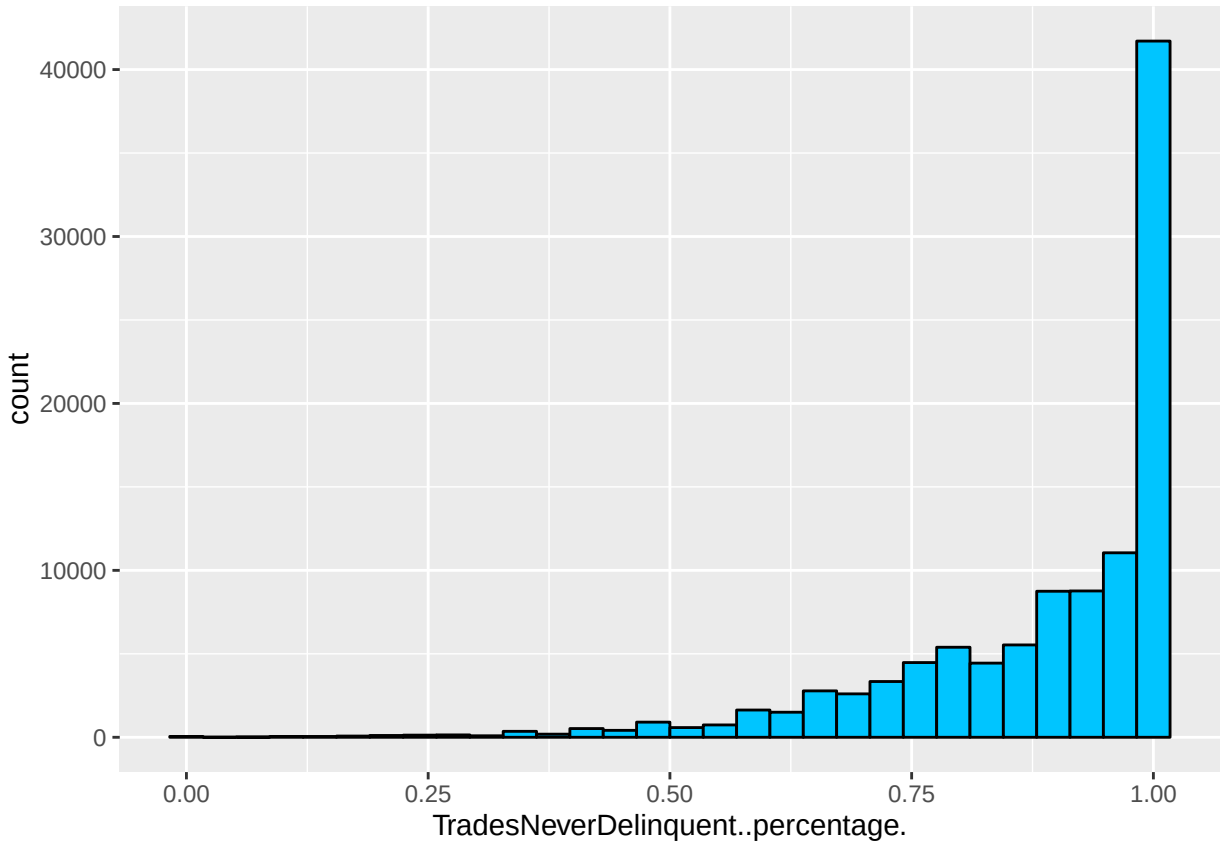
Unlike open credit lines, closed-end credit provides a fixed amount of money to finance a specific purpose and period of time. The loan may require periodic principal and interest payments,or payment of the entire principal at the end of the loan term. Many financial institutions refer to closed-end credit as an installment loan or a secured loan. Generally, real estate and auto loans are closed-end credit, and credit cards are revolving lines of credit or open-end. I think closed-end credit line might be a good variable to predict the risk later, so I will create a new variable 'Closedendcredit' by two on hand variables, 'CurrentCreditLines' and 'OpenCreditLines'. The diffrence between the variables will be the quantities of the closed-end credit line.



Not all the borrowers have a closed-end credit line, and most closed-end credit lines owners have 1 or 2, but still, we can examine if the closed-end credit lines oweners would cause less charge-off or delinquencies later.

2-12 TradesNeverDelinquent

In the next plot, I will explore the percentages of each borrower 's trades that have never been delinquent at the time the credit profile was pulled.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   0.820   0.940   0.886   1.000   1.000    7544
```

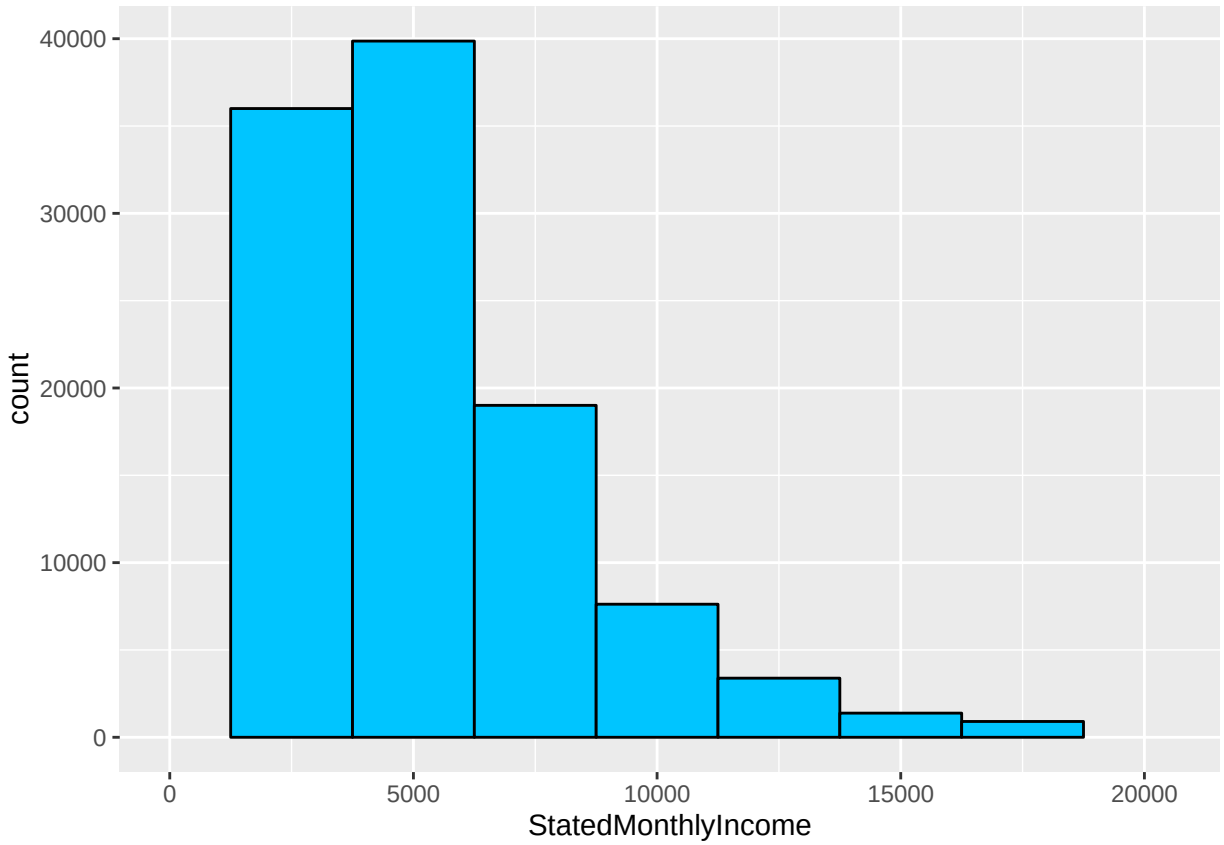
It seems that most of the borrowers are successful in repayment.

2-13 StatedMonthlyIncome

As for the borrowers' income, I created the plot below with the variable 'StatedMonthlyIncome', which is the monthly income the borrower stated at the time the listing was created.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    3200    4667    5608    6825 1750000
```

The minimum is $0, the median is $4,667, and the maximum is $1,750,000. There are some significant outliers, so I Limited the values to 99% quantile to avoid the outlier issue.



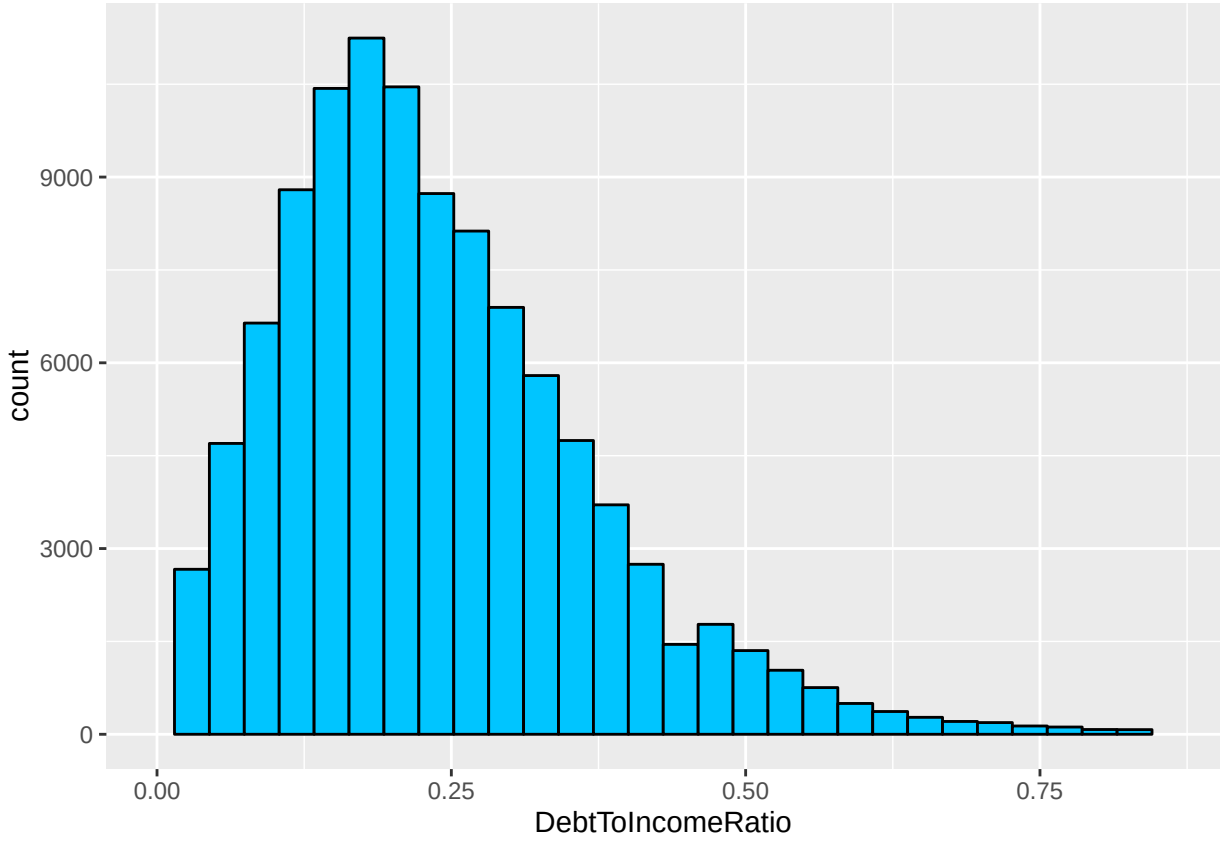
From the plot and summary above, we learn that most of the borrowers' annuel income is less than $100,000. However, knowing the income is not enough, the debt to income ration will give us a better understanding about the borrowers' financial status.

2-14 DebtToIncomeRatio

The debt to income ratio of the borrower at the time the credit profile was pulled. This value is Null if the debt to income ratio is not available. This value is capped at 10.01 (any debt to income ratio larger than 1000% will be returned as 1001%).

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   0.140   0.220   0.276   0.320  10.010    8554
```

The minimum is 0, the median is 0.22, and the maximum is 10.010. Again, there are some significant outliers, so I Limited the values to 99% quantile to avoid the outlier issue.



The debt to income ratio histogram shows a positively skewed, normal distribution, and more than 75% of the borrowers have the debt to income ratio less than 0.32.

3. Univariate Analysis

What is the structure of your dataset?

There are 113,937 observations in the Prosper loan data with 81 variables. However, I have narrowed the variables down to 35 by removing the similar information. In the univariate analysis section, I employed 19 variables, including 10 continuous Variables and 9 categorical variables.

Continuous Variables: Investors,LoanOriginalAmount,BorrowerRate,BankcardUtilization, TradesNeverDelinquent,StatedMonthlyIncome,DebtToIncomeRatio,PublicRecordsLast10Years, CurrentCreditLines and OpenCreditLines.
Categorical Variables:ListingCategory,ProsperRatings,Term,LoanStatus,BorrowerState, IsBorrowerHomeowner, CreditScoreRangeLower,FirstRecordedCreditLine and ListingCreationDate.

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

The main features in the dataset is loan status and investors. I position myself as an analyst in Prosper Marketplace and there are two main investigations in this report. First, I'd like to determine which features are best for predicting the risk of charge-off or default, so loan status is the dependent variable of this analysis. Second, I would like know about investors' preferences, so I will use investors to figure out which variables are the criteria that the investor would value more while choosing the loan requests.

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

ListingCategory,CreditScoreRangeLower, ProsperRatings, LoanOriginalAmount,Term, and BorrowerRate likely contribute to the average numbers of the investors. As for loan status, I think CreditScoreRangeLower, DebttoIncomeRatio, BorrowerRate, LoanOriginalAmount, StatedMonthlyIncome, TradesNeverDelinquent and Bankcard Utilization probably contribute to the the loss or risk.

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

I created two new variables, 'Length of Credit History' and 'Closed-end credit line'.

I use the exsting variables, FirstRecordedCreditLine and ListingCreationDate to calculate the borrowes' lengnth of credit history. Moreover, I also use 'CurrentCreditLines' and 'OpenCreditLines' to create the new variable 'Closed-end credit line'.Both of the new variables might be helpful for the risk analysis later.

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

1. PublicRecordsLast10Years was a continuous numeric variable, but I coverted it to boolean value, so I can see how many borrowers have public records clearly.
2. LoanStatus had 12 factors, and 6 of the factors are Pastdue status (different delinquency period), but I want the data to be simplier, so I combine the 6 pastdue factors as one.
3. ListingCategory was a set of numbers, which represent different categories: 0 - Not Available, 1 - Debt Consolidation, 2 - Home Improvement, 3 - Business, 4 - Personal Loan, 5 - Student Use, 6 - Auto, 7- Other, 8 - Baby&Adoption, 9 - Boat, 10 - Cosmetic Procedure, 11 - Engagement Ring, 12 - Green Loans, 13 - Household Expenses, 14 - Large Purchases, 15 - Medical/Dental, 16 - Motorcycle, 17 - RV, 18 - Taxes, 19 - Vacation, 20 - Wedding Loans. In order to make the plot readable, I converted the numeric factors to string factors.
4. Instead of histogram, an US map of heatmap can be a more graphic visualization for BorrowerState. To create the heatmap, I needed to seperate the BorrowerState to two row: State.name and numbers in the fisrt place. Later, I have transferred the states' abbreviation to all lowercase full names and merge both of the rows to the set called "map" in library(map).

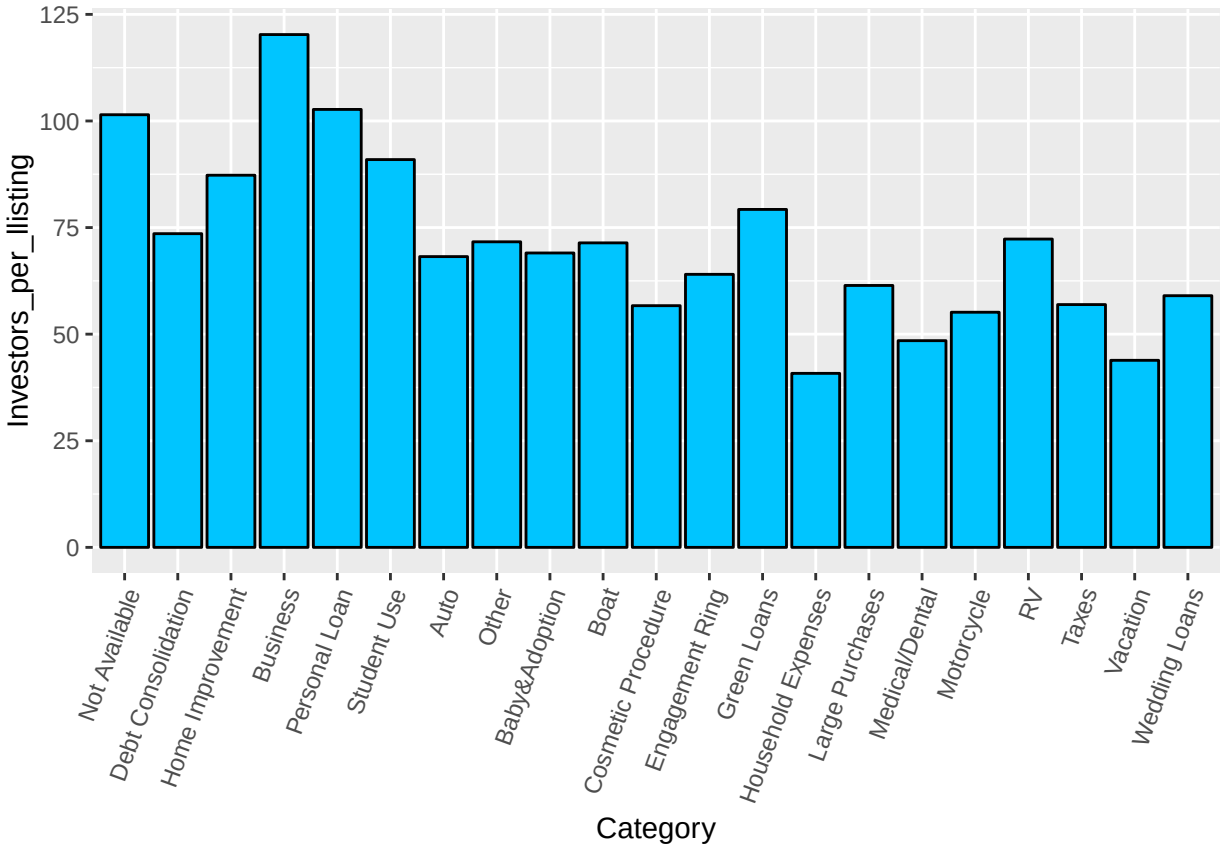
4. Bivariate Plots Section

In the section, I will focus on two main features of interest in the report-Investors and Loan Status. Unlike the univariate plots above, , I will also start using multiple colors in the bivariate plots for better visualization.

What attract investors?

4-1 Investors & Listing Category

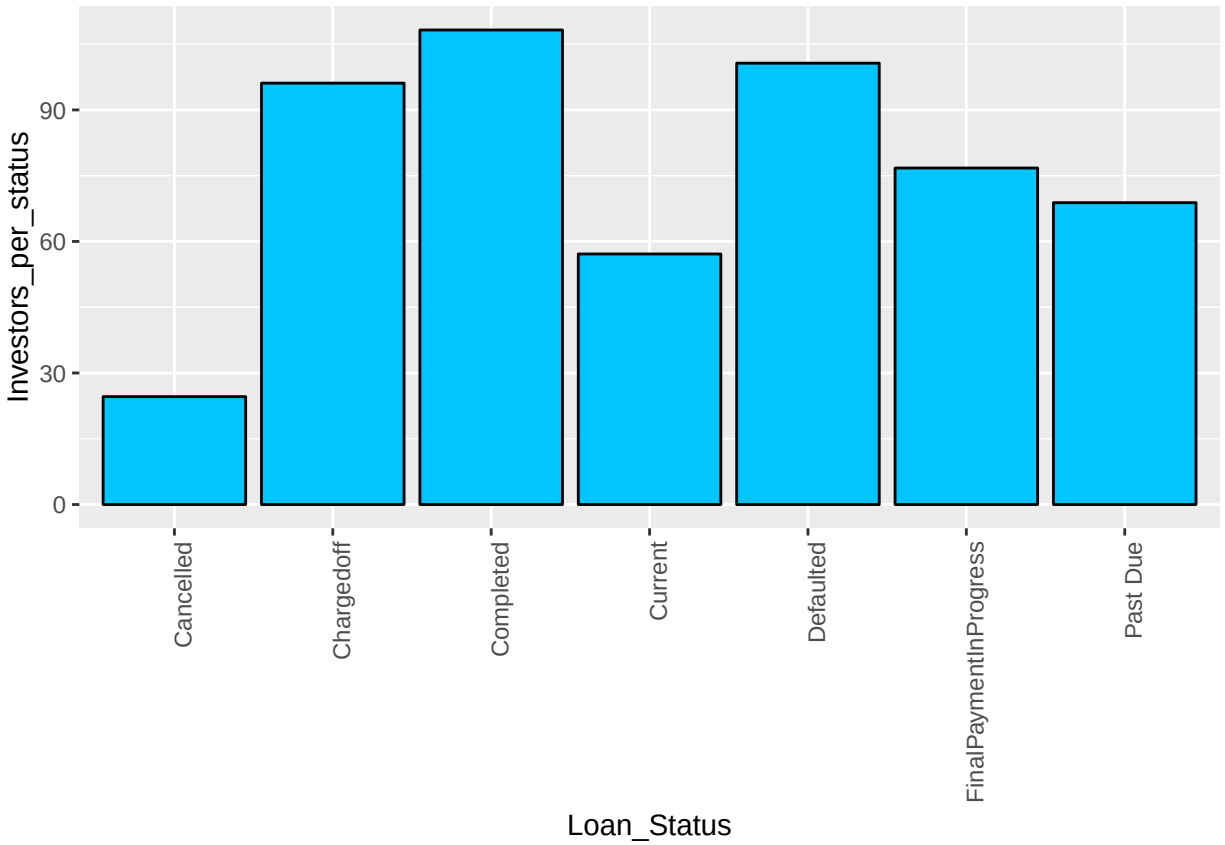
First, what purpose of loans are more popular among investors? There are 20 categories that borrowers can select and we already know that 'Debt Consolidation' accounts for 50% based on the univariate plot above. However, to answer the question, the numbers I need here are the average investors for each categories. Therefore, I used the 'group by' function to calculate the mean investors of all the categories.



'Business' accounts for only 6% of the all listings, but it is the the most popular category among investors. Every 'Business' loan can attract more than 120 investors.

4-2 Investors & Loan Satus

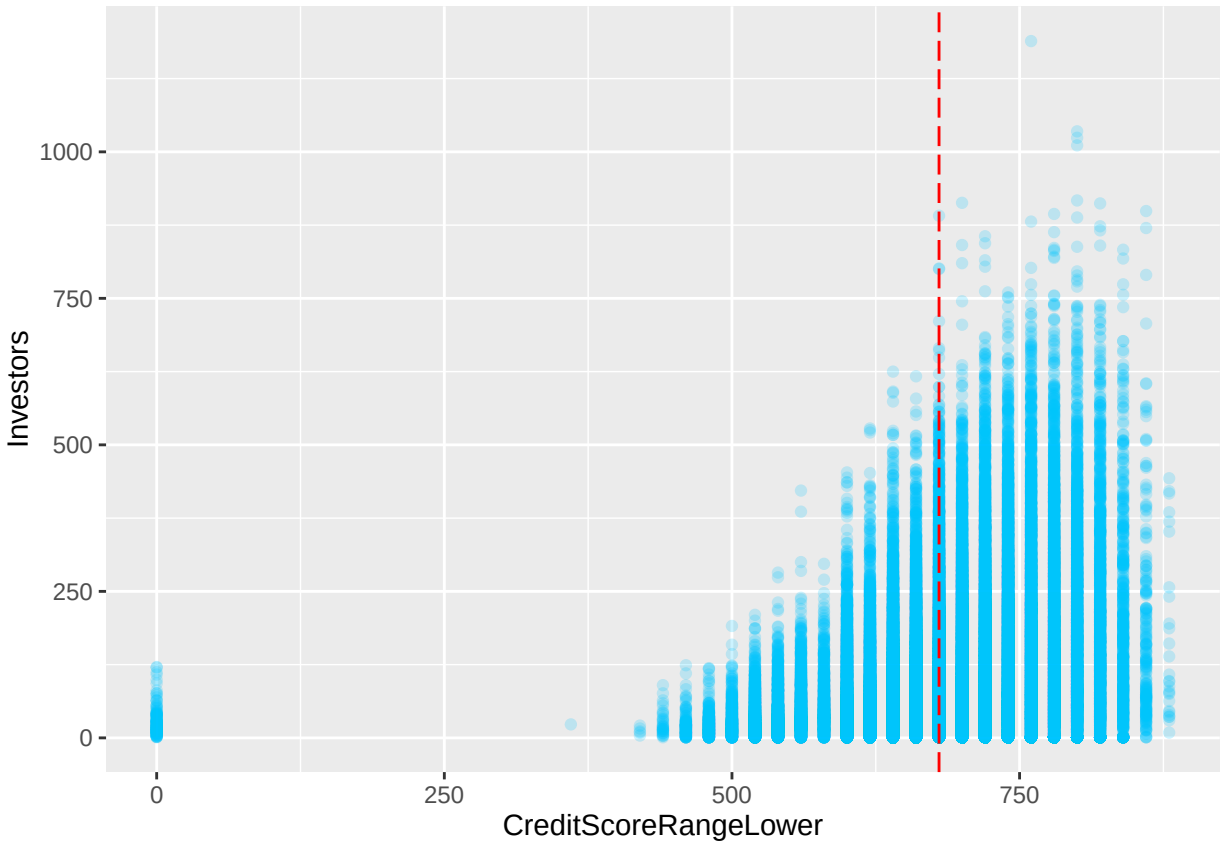
Did the investment pay off? Let's see if the numbers of average investor very in different loan status.



From the plot and the table above, we can learn that the average investors of chargedoff and defaulted loan are more than the loans lebeled with'FinalPaymentInProgress' and 'current' but less tan the completed loans.

4-3 Investors & Credit Scores

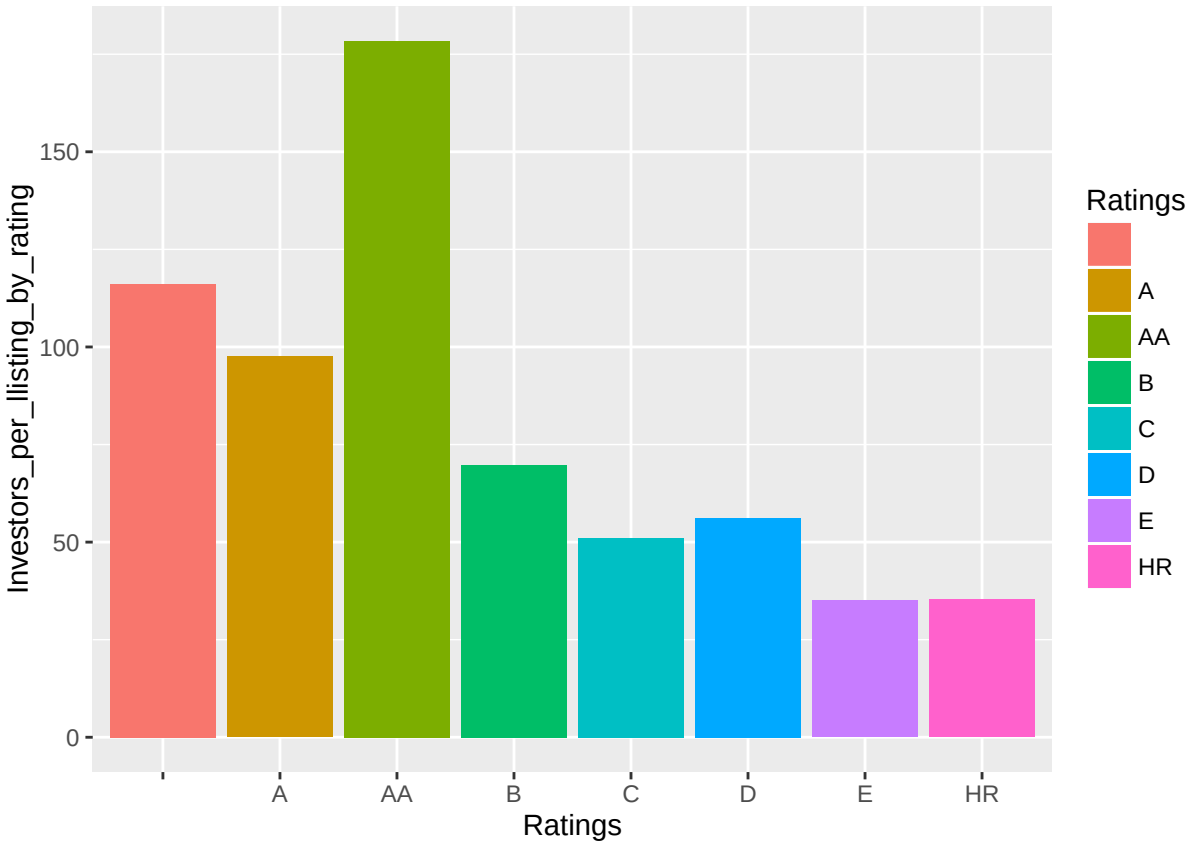
Dose the credit scores matter? I use 'CreditScoreRangeLower' here as what I did in univariate section.



The black dotted line is the median of the borrowers' credit score, 680, and it seems that most investors would choose the loan requsts more than the median scores.

4-4 Investor & Prosper Ratings

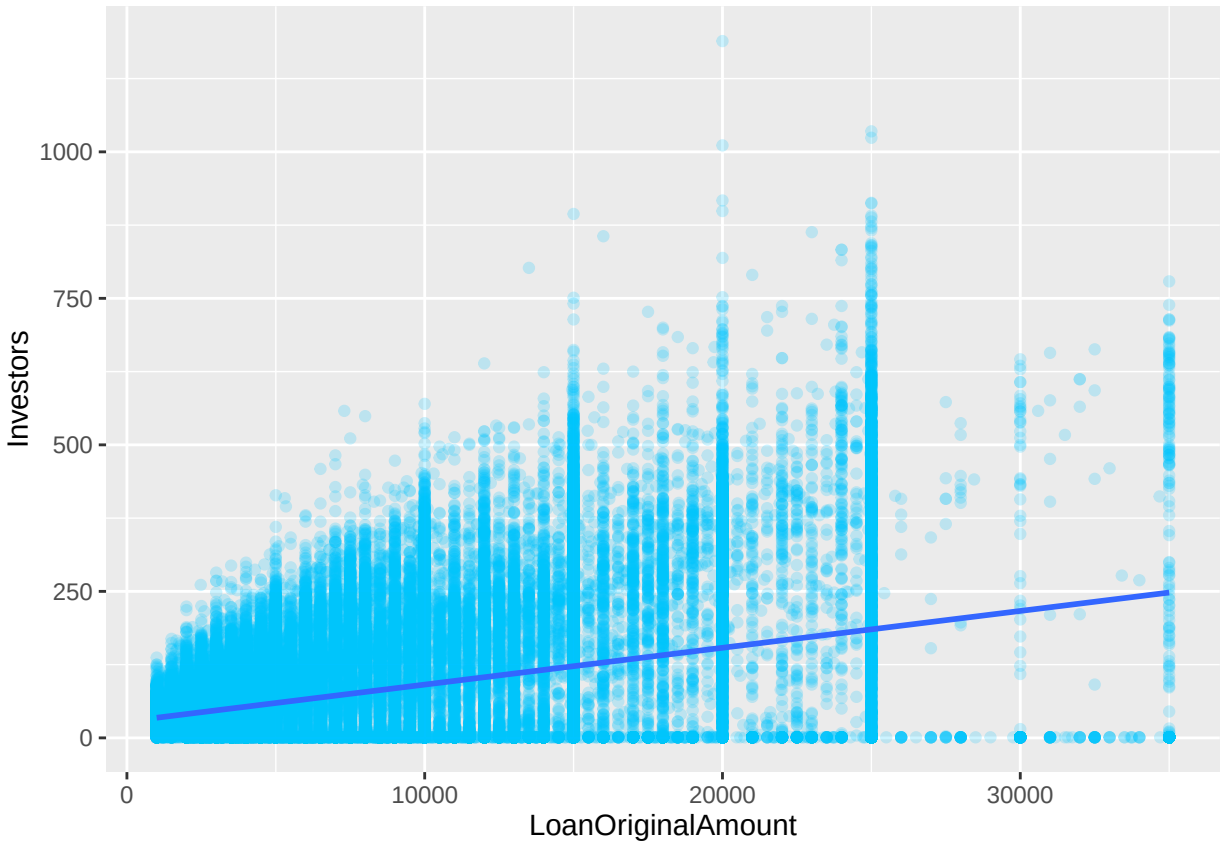
Do the investors trust Prosper Rating system? Prosper Rating composed by 7 levels,from lowest-risk to highest-risk, are labeled AA, A, B, C, D, E, and HR ("High Risk"). According to the statistics in the univariate section, we learn that most loan were rated 'C'. To answer the question, I will use the average investors for each rating. Therefore, I used the 'group by' function to calculate the mean investors of all the ratings.



Even though 'C' rating is the majority, 'AA' is the most popular rating among the investors; 'E' and 'HR' are undesirable. It seems that most investors think Prosper Rating System is trustworthy and they are conservative with investments.

4-5 Investor & Loan Original Amount

Do the investors prefer the loan with larger amount or less?

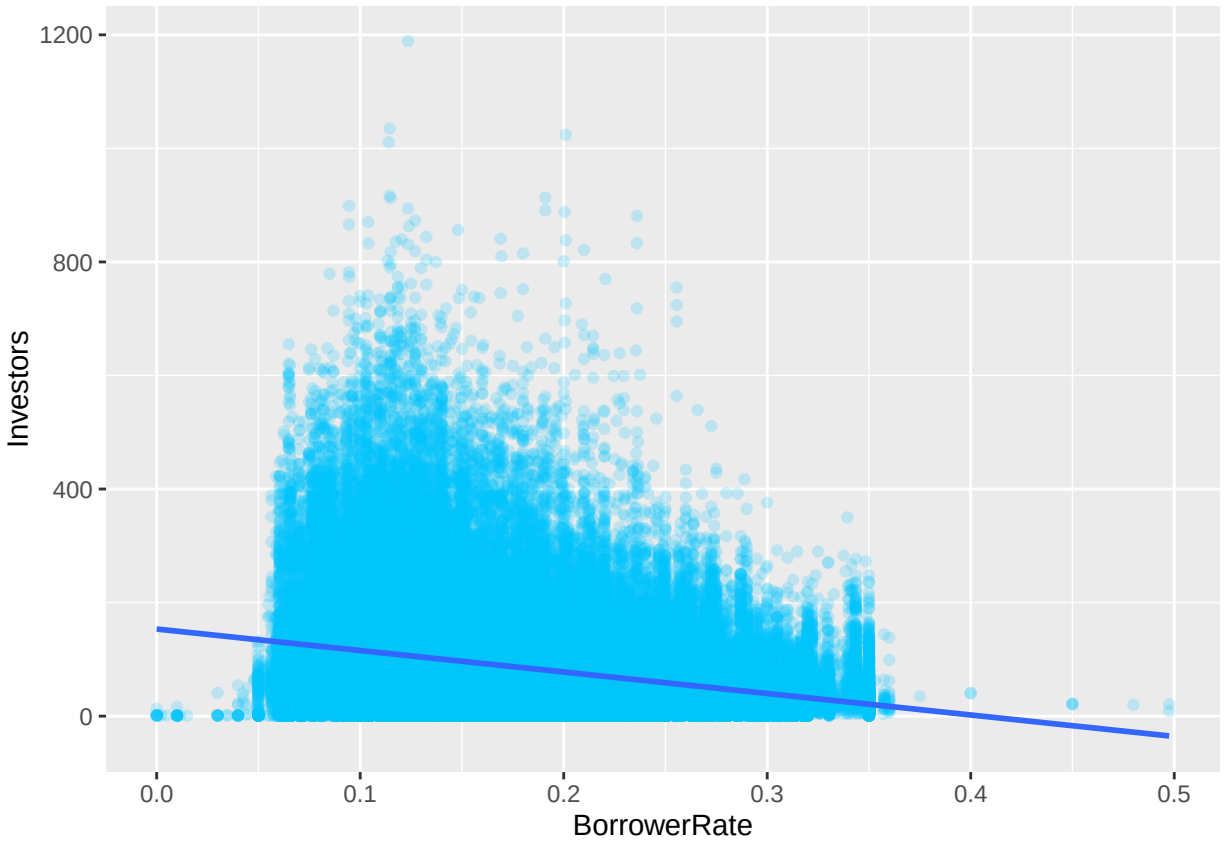



```
##
##  Pearson's product-moment correlation
##
## data:  df$LoanOriginalAmount and df$Investors
## t = 138.7077, df = 113935, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3751140 0.3850494
## sample estimates:
##       cor
## 0.3800926
```

Under the amount of $25,000, we can see that there's a positive linear correlation between investors abd the loan amount, and the correlation coefficient is 0.38.

4-6 Investor & Interest Rate

Does the interest rate motivate investors?



No! The plot above tell us again that the investors on Prosper are conservative. In fact,higher interest rate, less investors.

```
##
##  Pearson's product-moment correlation
##
## data:  df$BorrowerRate and df$Investors
## t = -96.2493, df = 113935, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2795783 -0.2688385
## sample estimates:
##        cor
## -0.2742169
```

Let's calculate the correlation coefficient to look into it closer, the correlation coefficient is -0.274; there's slightly negative correlation between investors and the interest rate.

What variables might be correlated with risky loan?

Now, I would like to transfer my attention to loan status,especially these status-chargedoff, defaulted and pastdue.

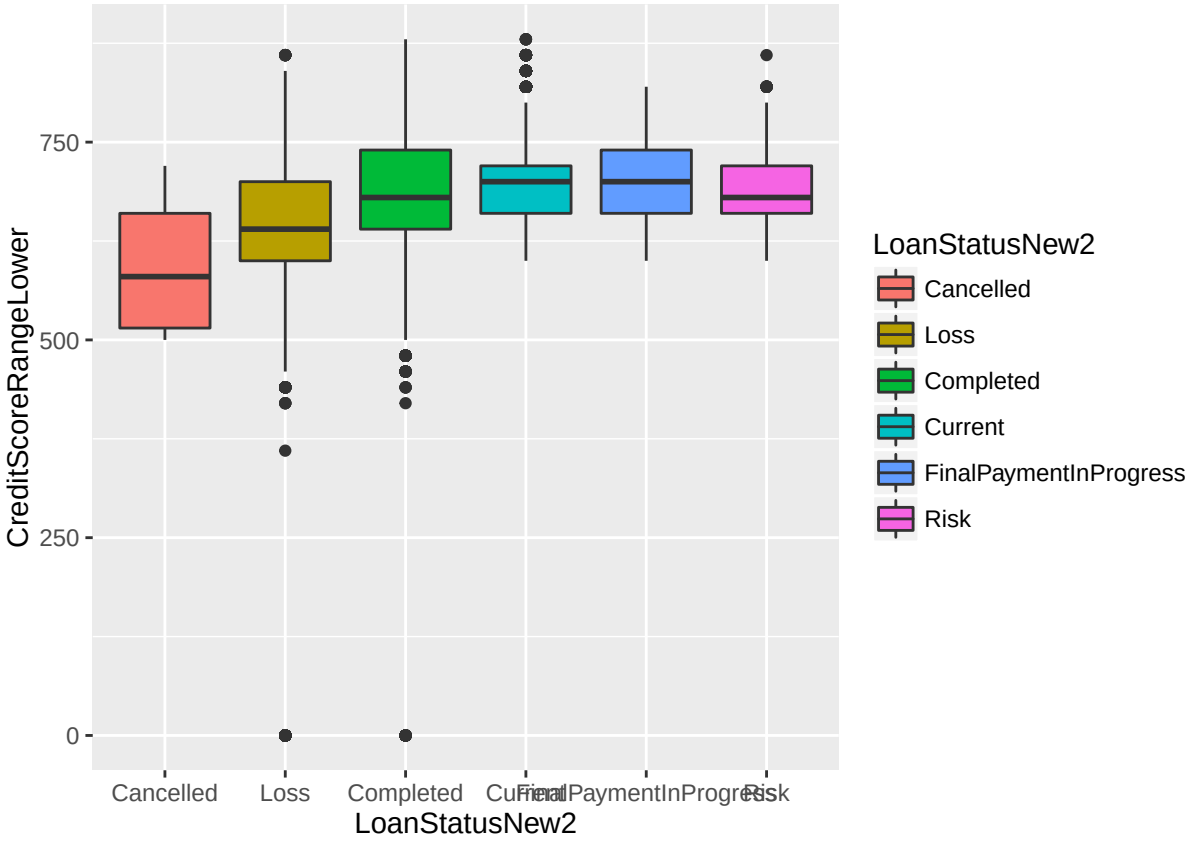
In general, a note goes into Default status when it is 121 or more days past due. When a note is in Default status, Charge Off occurs no later than 150 days past due (i.e. No later than 30 days after the Default status

is reached) when there is no reasonable expectation of sufficient payment to prevent the charge off. However, bankruptcies may be charged off earlier based on date of bankruptcy notification.

Thus, Chargedoff and defaulted loan could cause loss for the lenders, and the past due implies the upcoming risk. I would comapre the loan status and other variables to see if there's any correlation.

4-7 Loan Status & Credit Scores

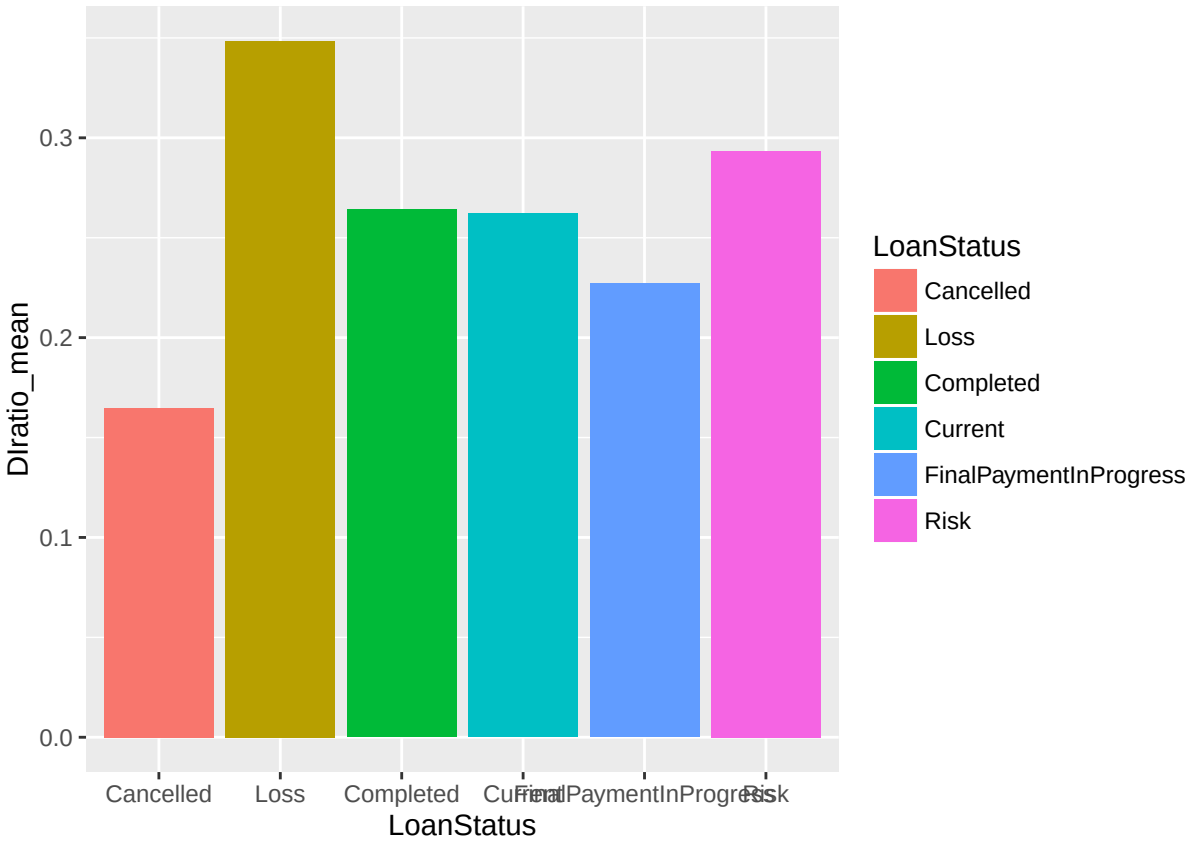
Can we observe differences in borrowers' credit scores with different loan status? The answer is yes.



We can see median borrower credit scores for the cancelled loan are the lowest, followed by loss, risk, and other normal loans. Cancelled loans have small sample size so the results might not be significant. It will be ignored in the following sections. The rests are aligned with my intuition: people with lower credit score tend not to pay back. Credit scores are an accurate measure of people's credibilities.

4-8 Loan Status & Debt to Income Ratio

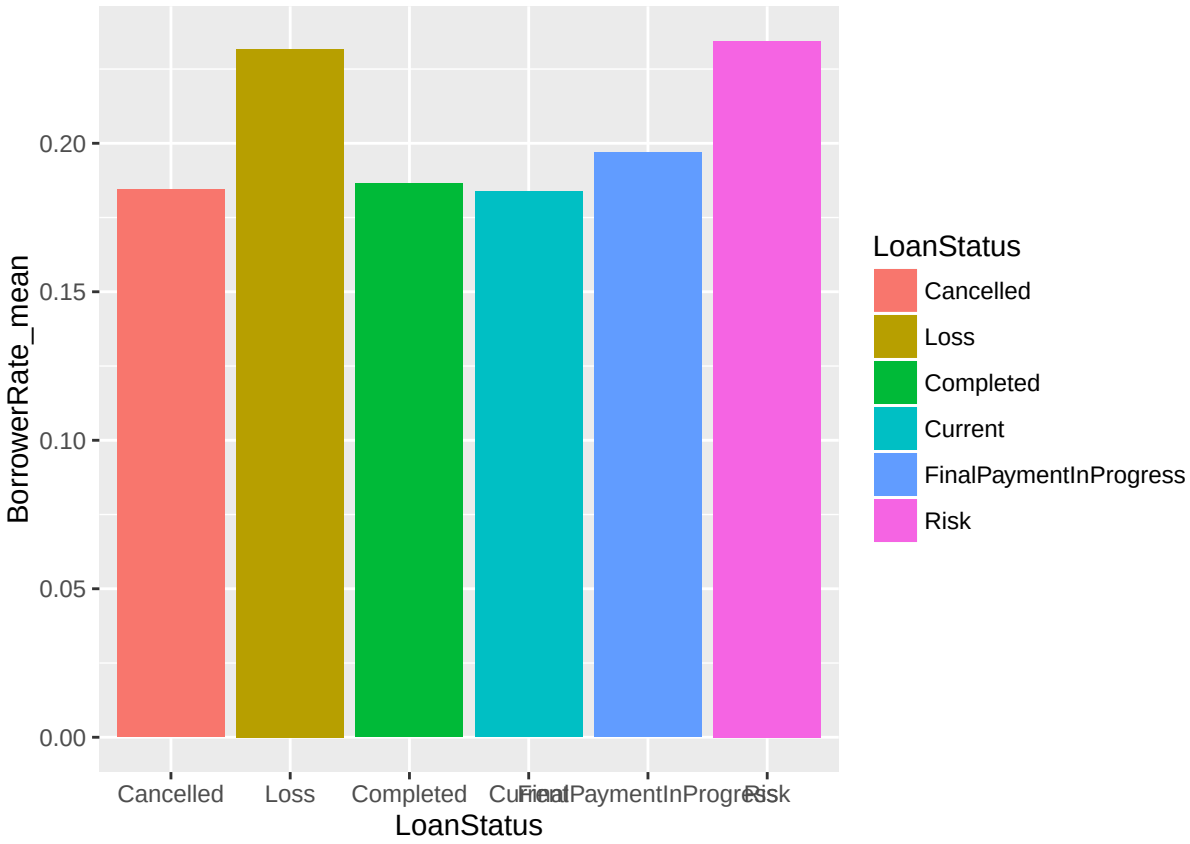
Do people who owe debts more tend to default?



From the above graph, we can clearly see that borrowers whose loans have loss (defaulted or chargedoff) have an average debt-to-income-ratio of 0.35, followed by risky loans (around 0.3). Borrowers with higher leverage on debt tend to default on their loans.

4-9 Loan Status & Interest Rate

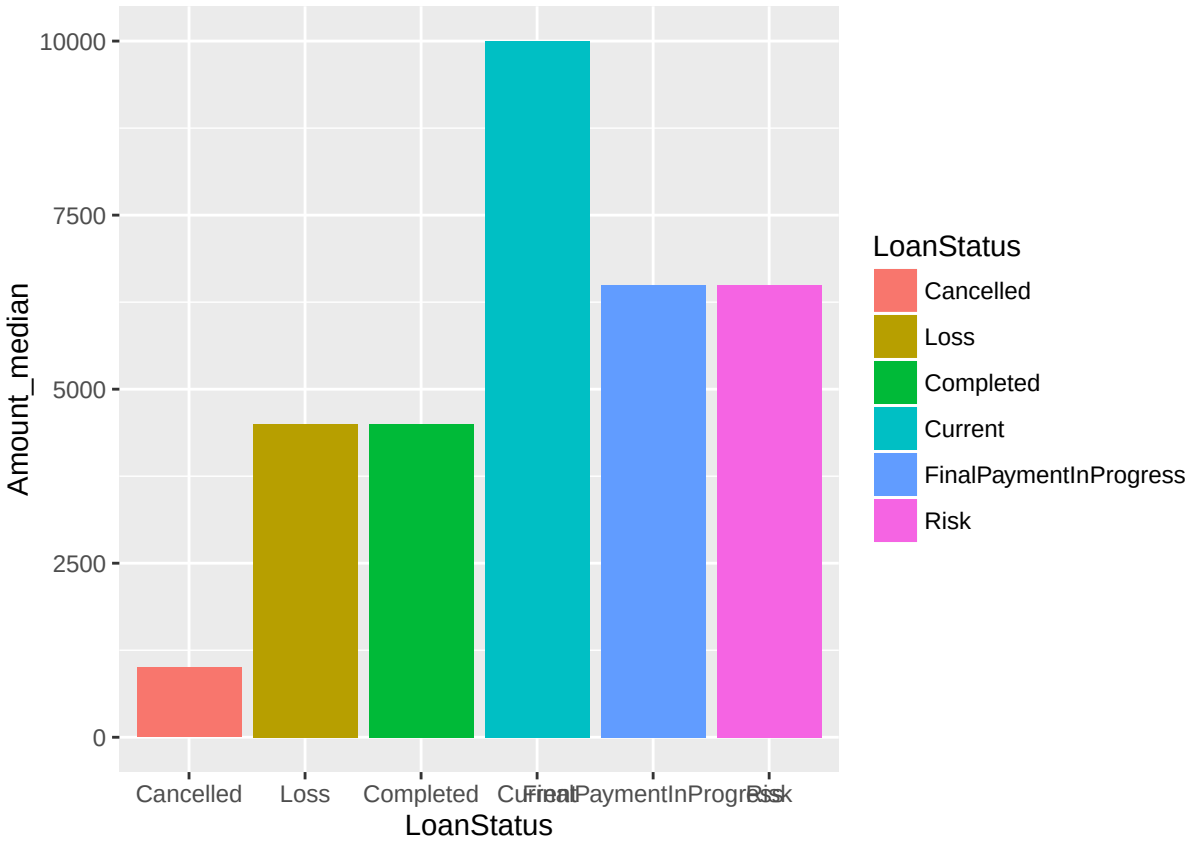
Loan status is the outcome of investment in terms of risk. Do risk metrics like borrower rates reflect the outcome?



Loss and risk categories have the highest of average borrower rates (more than 20%). Those borrowers are viewed by investors as riskier investment, and the outcomes do reflect their expectations.

4-10 Loan Status & Loan Original Amount

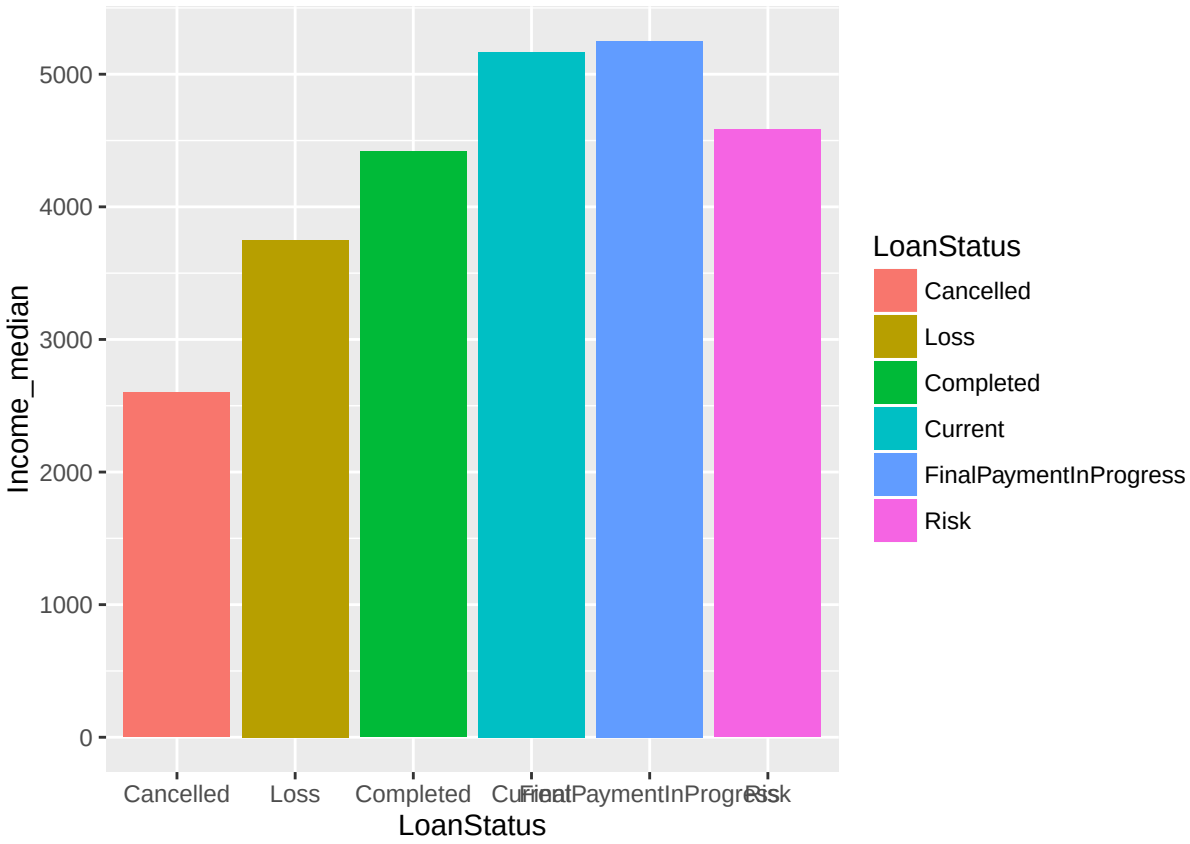
Do risky borrowers not get trust from investors so they can't borrow as much money as other less risky borrowers?



Although median loan amount for current loans is $10000, but the median for completed loans is less than $5000 (same amount as loss loan)Those current loans can also turn into loss or risky loans. From this graph, we can't say there is a strong pattern for sure.

4-11 Loan Status & Stated Monthly Income

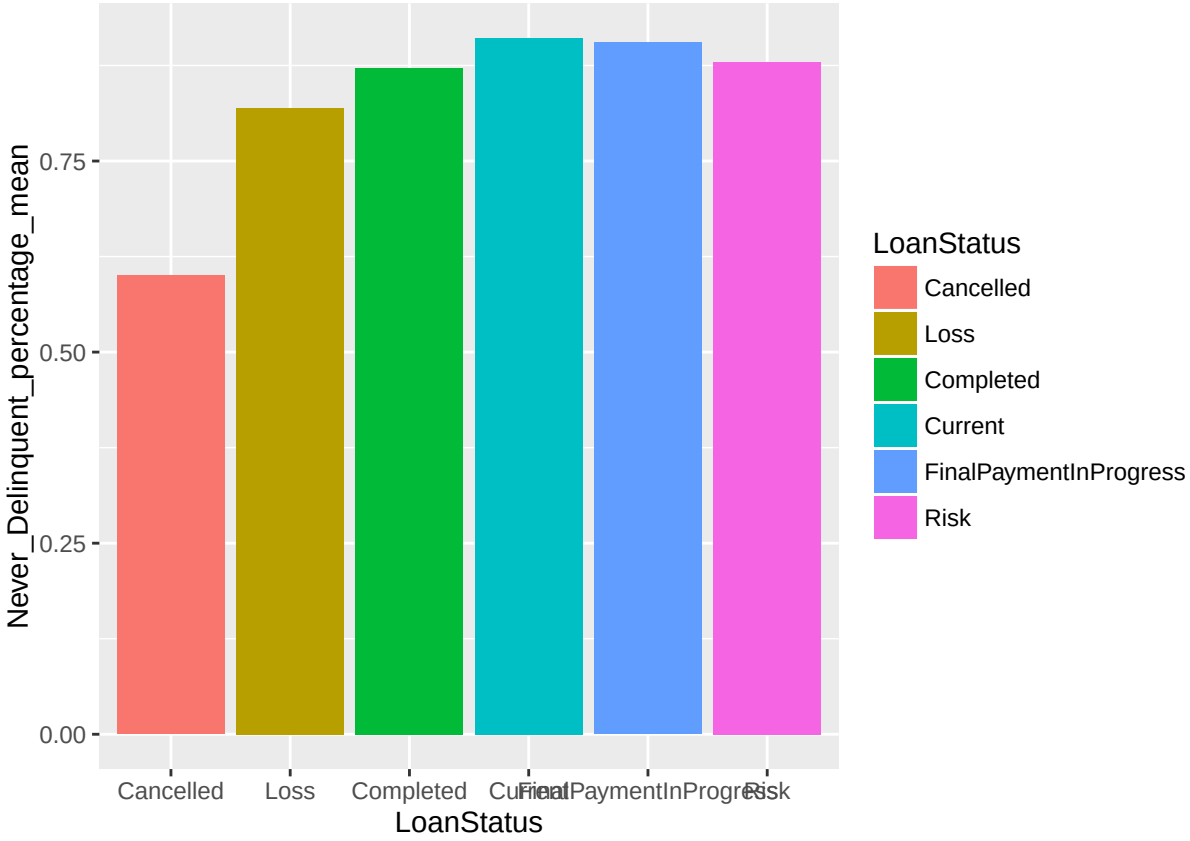
Is there any relation between loans' risk categories and borrowers' monthly income?



We can see pretty interesting findings here. Borrowers whose loan status are current or payment in progress have the highest income. Borrowers whose loans are risk or completed have about the same income, followed by loss. People with least monthly income have hard time paying back the loans. People with monthly income around $4500 have quite different behavior - some of them pay back on time but some don't.

4-12 Loan Status & Trades Never Delinquent

When borrowers are never delinquent on their trades, will they also pay back the loans on time?

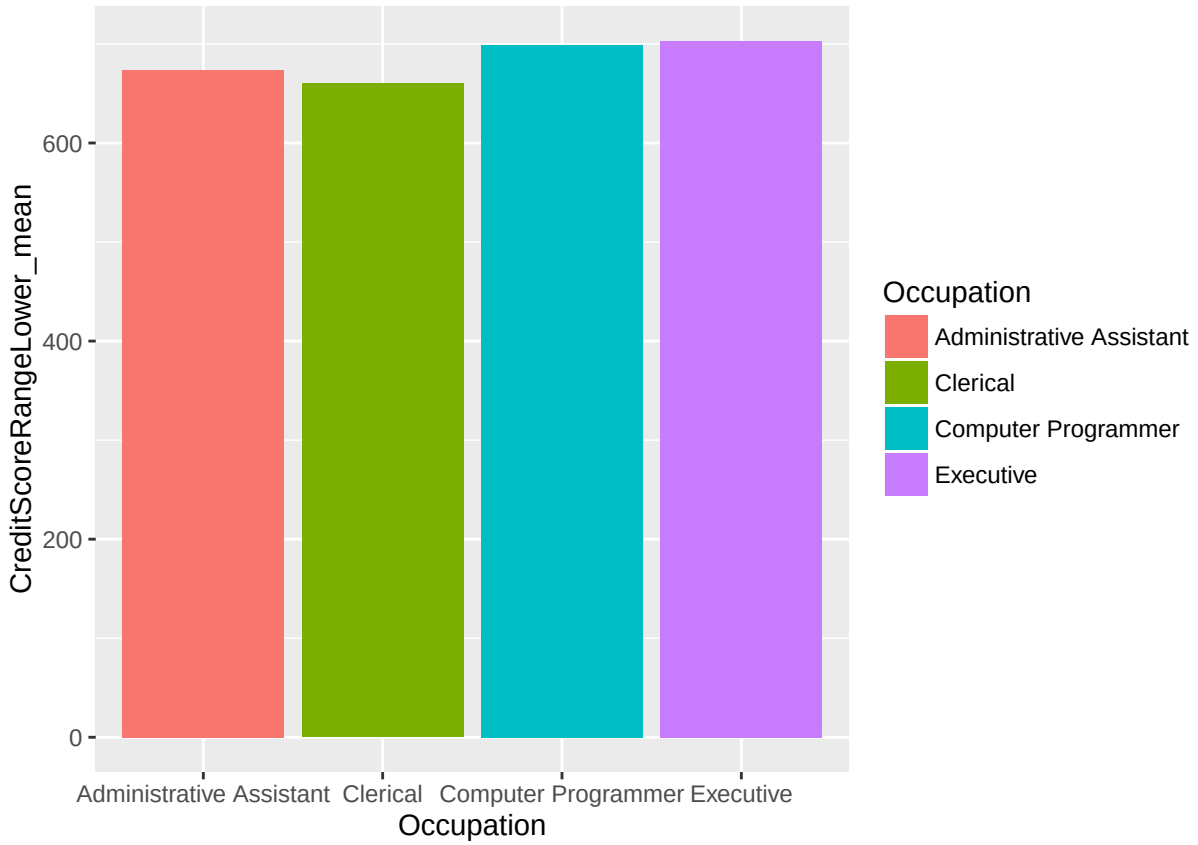



```
##               LoanStatus Never_Delinquent_percentage_mean
## 1              Cancelled                         0.6000000
## 2                   Loss                         0.8191688
## 3              Completed                         0.8718497
## 4                Current                         0.9108597
## 5 FinalPaymentInProgress                         0.9052195
## 6                   Risk                         0.8793662
```

Definitely yes. We can see that difference in never_delinquent_percentage between loss and current loans are almost 10% (91.08% - 81.91%)

4-13 Occupation & Credit Score

Except for the main features, I am also curious about the correlation between occupation with the credit score, so I choose 4 occupations based on the amount and the title differences-Executive, Computer Programmer, Administrative Assistant and Clerical.




```
##                 Occupation CreditScoreRangeLower_mean
## 1 Administrative Assistant                   673.6388
## 2                 Clerical                   660.0759
## 3      Computer Programmer                   698.7182
## 4                Executive                   702.8810
```

The finding is quite interesting and intuitive; the Executive and Computer programmer have relatively high credit scores.

5. 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 set?

For the 'Investors' variable, there are relatively strong correlation with 'ListingCategory', 'ProsperRating' 'LoanOriginalAmount' and 'Term'. Based on the analysis, we can learn that investors prefer the 3-year business loan with AA ratings, and if the amount of the loan is under $25,000, the larger amoumt will attract more investors.

As for the other main feature 'LoanStatus' variable, there are relatively strong correlation with 'DebtToIncomeRatio', 'BorrowerRate', 'StatedMonthlyIncome' and 'TradesNeverDelinquent'. Based on the analysis, we can learn that the borrowers with higher debt-income ratio, higher interest rate, less monthly income and less trades that never delinquent are more likely to fail in repayment.

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

Yes, the relationship between occupation and the credit score fits the common stereotype well. The decent job titles implies the higher credit scores.

What was the strongest relationship you found?

The relationship between Investors and ProsperRating is strong.

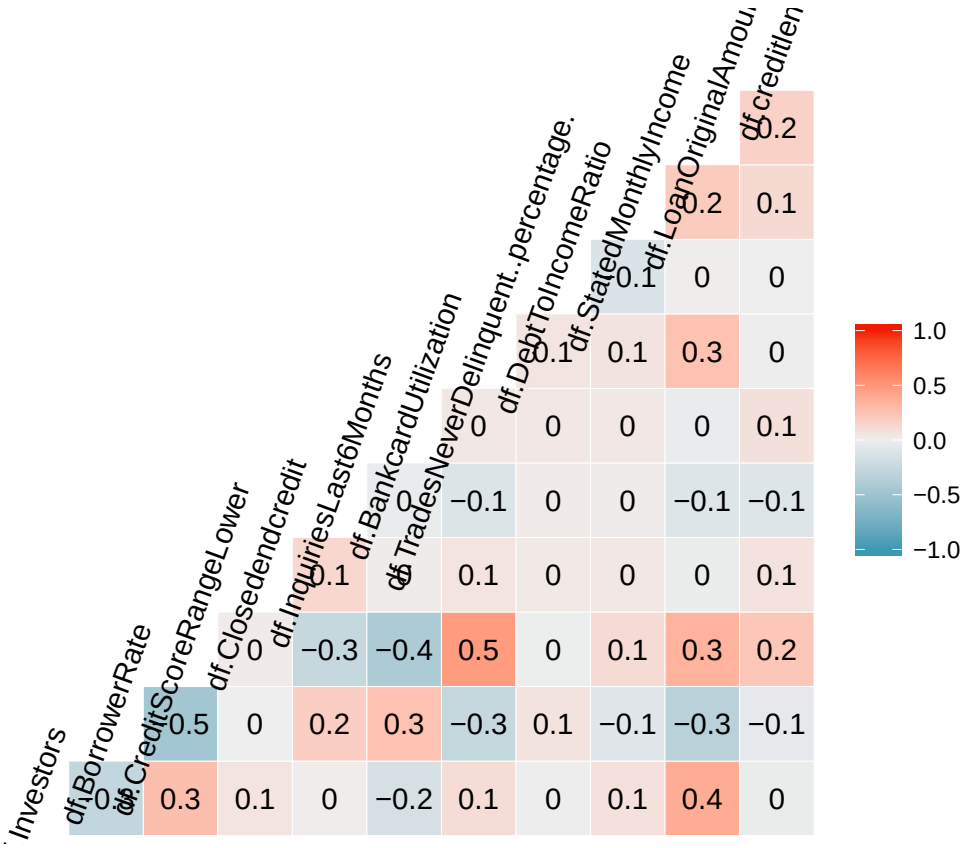
```
##   Ratings Investors_per_llisting_by_rating
## 1                                116.09971
## 2       A                         97.52869
## 3      AA                        178.36783
## 4       B                         69.80245
## 5       C                         51.09054
## 6       D                         56.23371
## 7       E                         35.01419
## 8      HR                         35.27585
```

The gap between the average investors of 'AA' and 'HR' is almost 143, which means that each AA rating loan can attract 143 more investors than the 'HR' loan. The relationship is not only strong but also significant.

6. Multivariate Plots Section

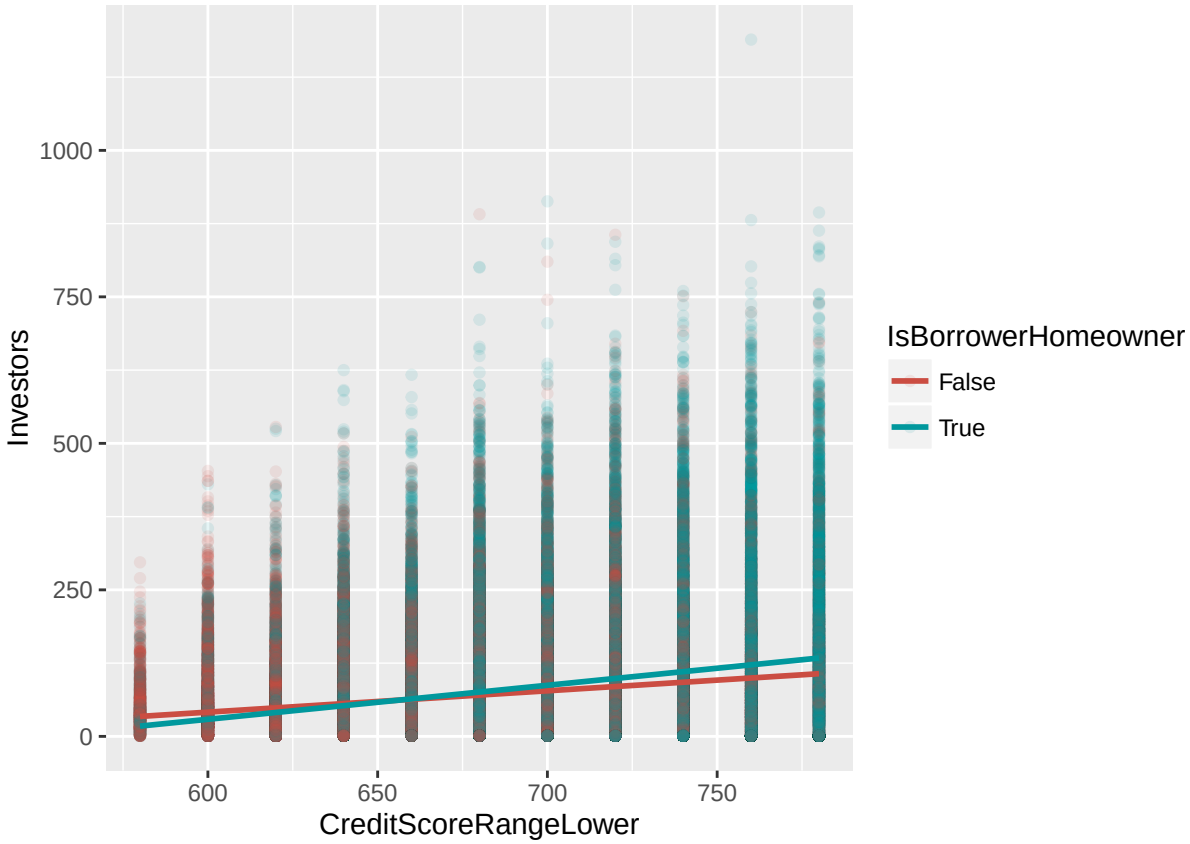
6-1 Correlation Plot of 'Investors'

'Investors' is one of the most important features of the report, so I would like to see the correlation coefficient between 'Investors' and other numerical continuous variables of the data.



From the plot, we can see the pearson's r between the variables and 'Investors.' Next,I will choose the variables which have comparatively stronger correlation with 'Investors' and also apply the other main feature in the report 'Loan Status' and some other categorical variables to create some multivariate plots.

6-2 'Investors' VS. 'IsBorrowerHomeowner' VS.'CreditScoreRangeLower'

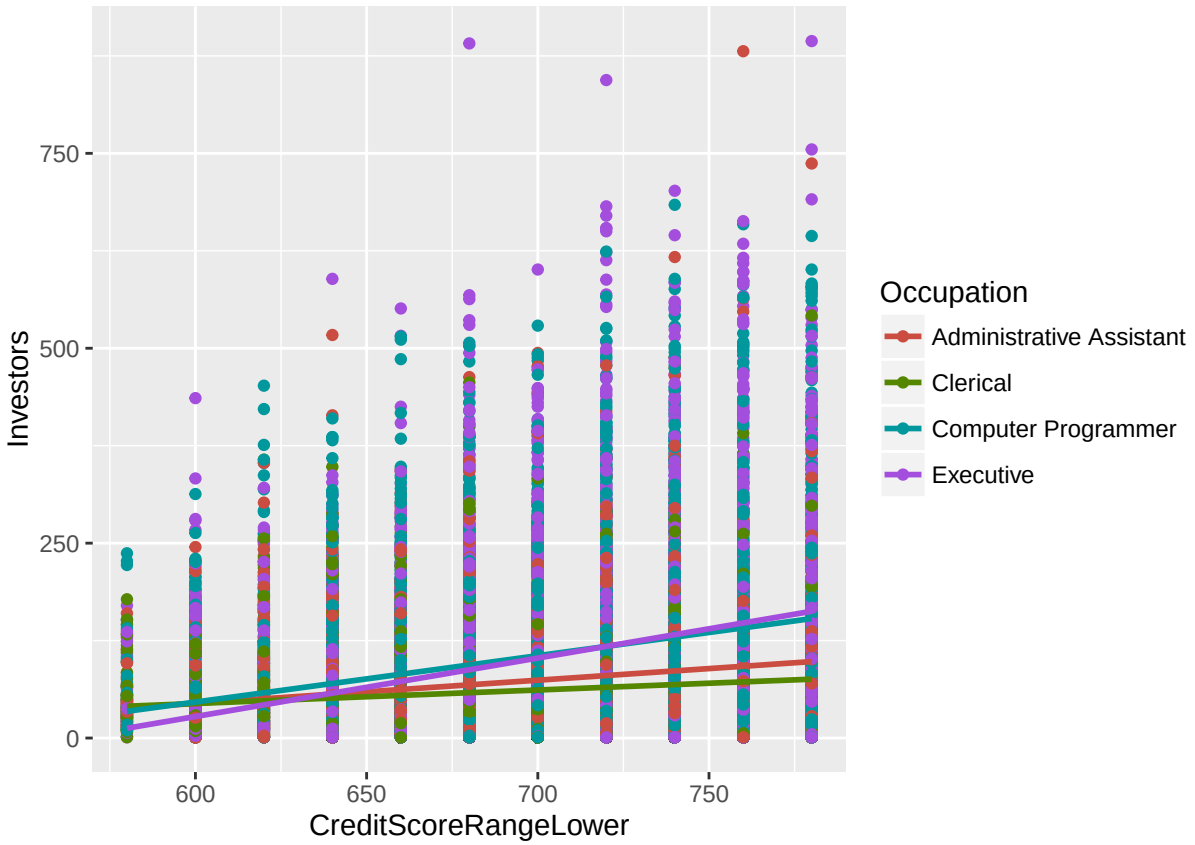


The interesting finding in the plot above is that starting from the scores 650, even the borrowers have the same scores, the investors are more likely to fund the homeowner's loans.

6-3 'Investors' VS. 'Occupation' VS.'CreditScoreRangeLower'

As the analysis in 4-18, I will use 4 occupations-Executive, Computer Programmer, Administrative Assistant and Clerical here and I would like to know if the investors have a bias against the borrowers who are not from management level or Computer Programmer.

```
##                         Var1  Freq
## 37                     Other 28617
## 43              Professional 13628
## 14      Computer Programmer  4478
## 21                 Executive  4311
## 61                   Teacher  3759
## 3  Administrative Assistant  3688
```



The finding here is interesting but cruel: Computer Programmer is the most popular occupation from the scores 600 to 700, and after 700, Executive become the most popular occpation gradually .
Even the borrowers have the same credit scores, the investors prefer the borrowers with decent job titles.

6-4 'LoanStatus' VS. 'BorrowerState' VS.'IncomeRange'

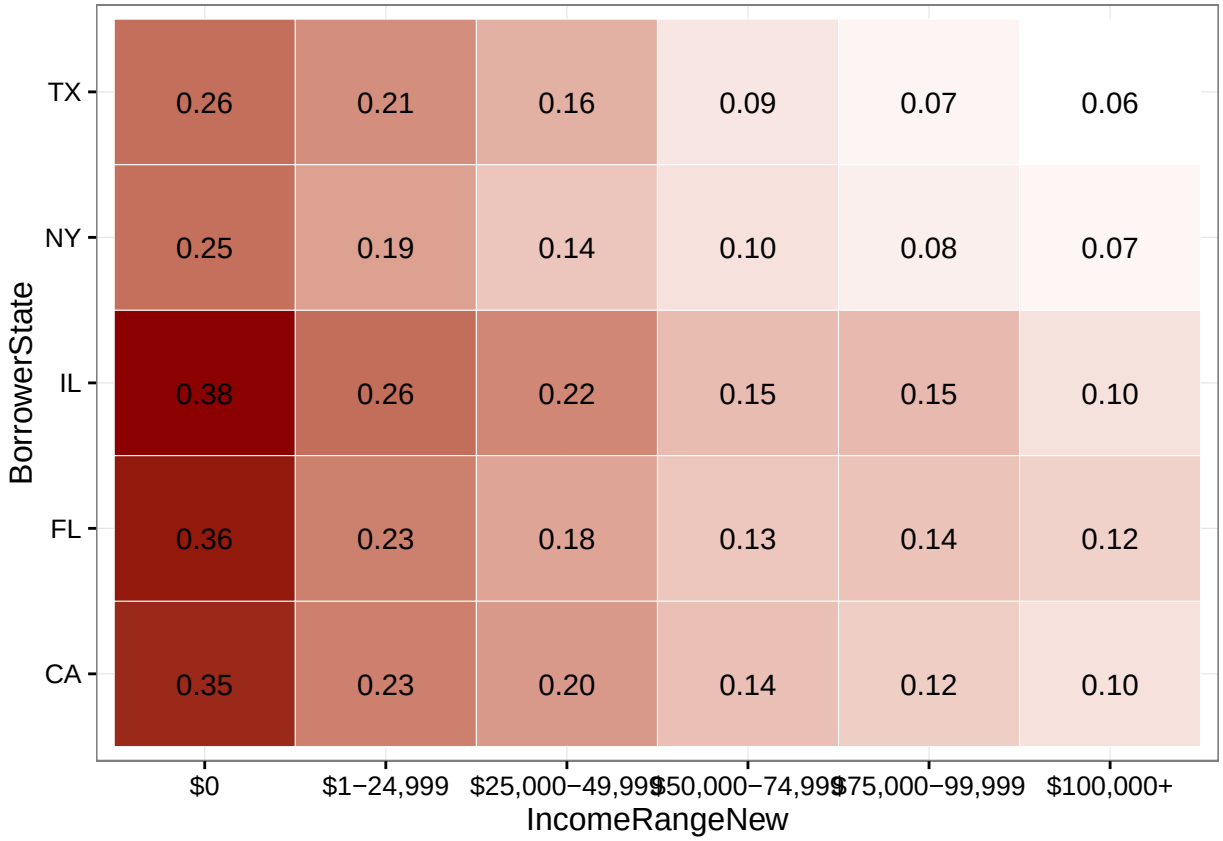
Last, let's add more demographic variables to the analysis. I would like to use the other main feature of the report 'LoanStatus'to see the borrowers from which State are more likely to successfully repay to the loans.

First, I'd like to narrow the 50 States down to top 5 States by the amount of the borrowers based on the analysis in the univariate section. The top 5 States include California, Texas, Florida, New York and Illinois.

In order to calculate the risky loan rate , I would also simplify the levels of the loan status, so I will combine the status 'Chargedoff' and 'Defualted' as a new level 'Loss' and the rename the 'Past Due' level 'Risk'. Here, I will use the sum of 'Loss' and 'Risk' to calculate the risky loan rate.

To make the analysis more accurate, I also use the 'IncomeRange' as another variable here. Thus, we will learn that if the borrowers in the same income range will have different repayment behaviour by different residential State.

```
##    Var1  Freq
## 6    CA 14717
## 45   TX  6842
## 36   NY  6729
## 11   FL  6720
## 16   IL  5921
## 1        5515
```



The informative plot above shows us that in the lowest income range, the borrowers from New York are more likely to have better repayment behaviour; in the highest income range, the borrowers from Taxes are more likely to have better repayment behaviour.

7. 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?

'Investors' VS. 'Occupation' VS.'CreditScoreRangeLower' is extended from the 4-18 plot in the bivariate section. I never thought that the 'Occupation' is so important to the investors, and the magnitude of the relationship between investors and occupation become more significant while we control the 'credit score' variable.

Were there any interesting or surprising interactions between features?

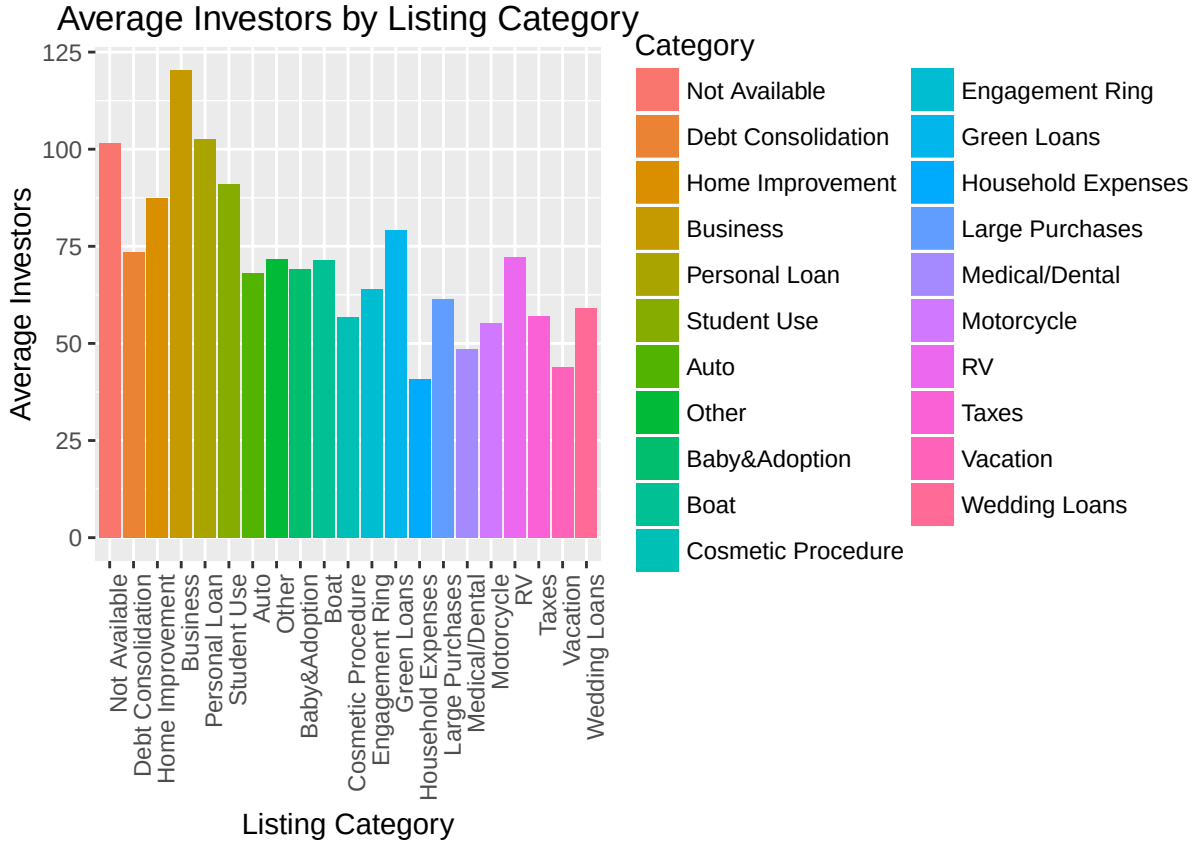
In the 6-4 'LoanStatus' VS. 'BorrowerState' VS.'IncomeRange', I found that in general, the borrowers from Taxes and New York have better repayment behaviors, and the borrowers from Illinois have the worst one. It's quite interesting that the geographic variable might also be a predictor variable in the analysis.

OPTIONAL: Did you create any models with your set? Discuss the strengths and limitations of your model.

---

8. Final Plots and Summary

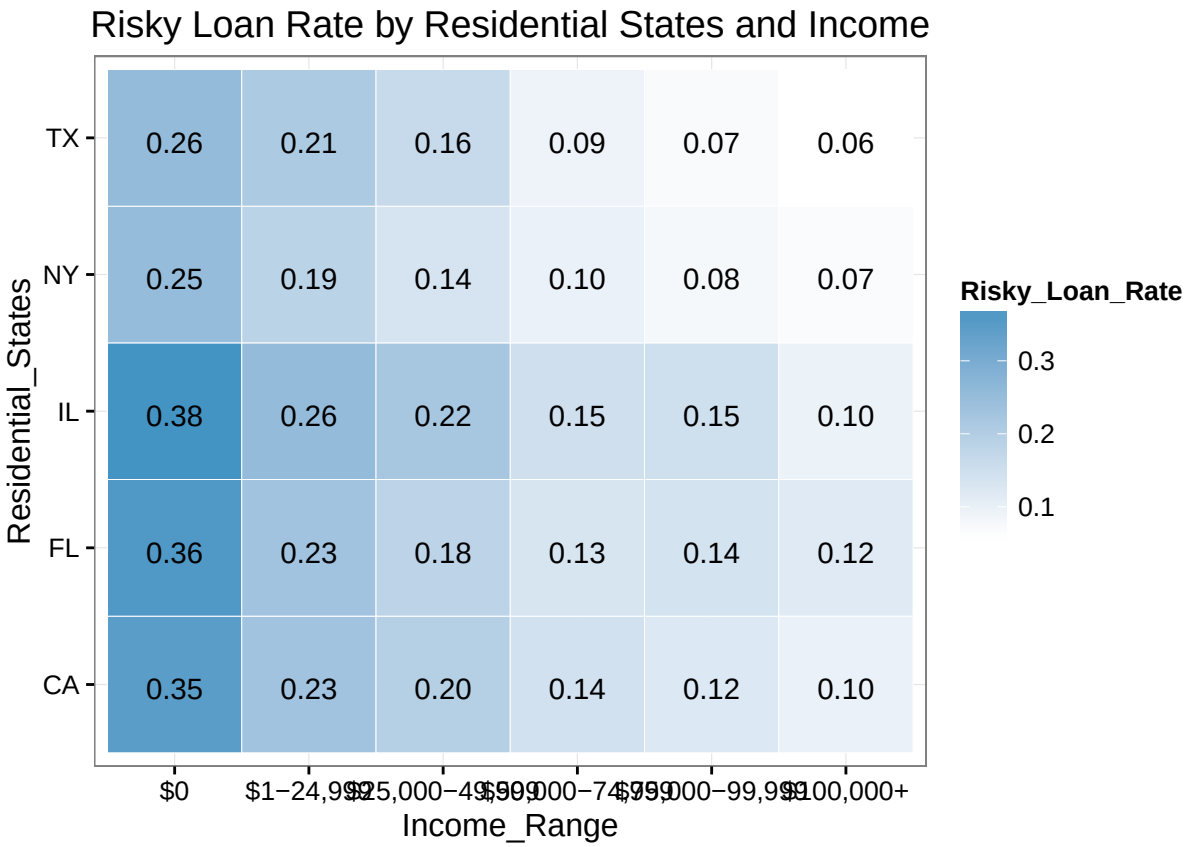
8-1 Plot One



Description One

Laon listings under the category 'Business' have the most mean investors; and loan listings under the category 'Household Expenses' have the least mean investors. Investors incline to trust athe loan specifically intended for business purposes instead of a personal fixed expense.

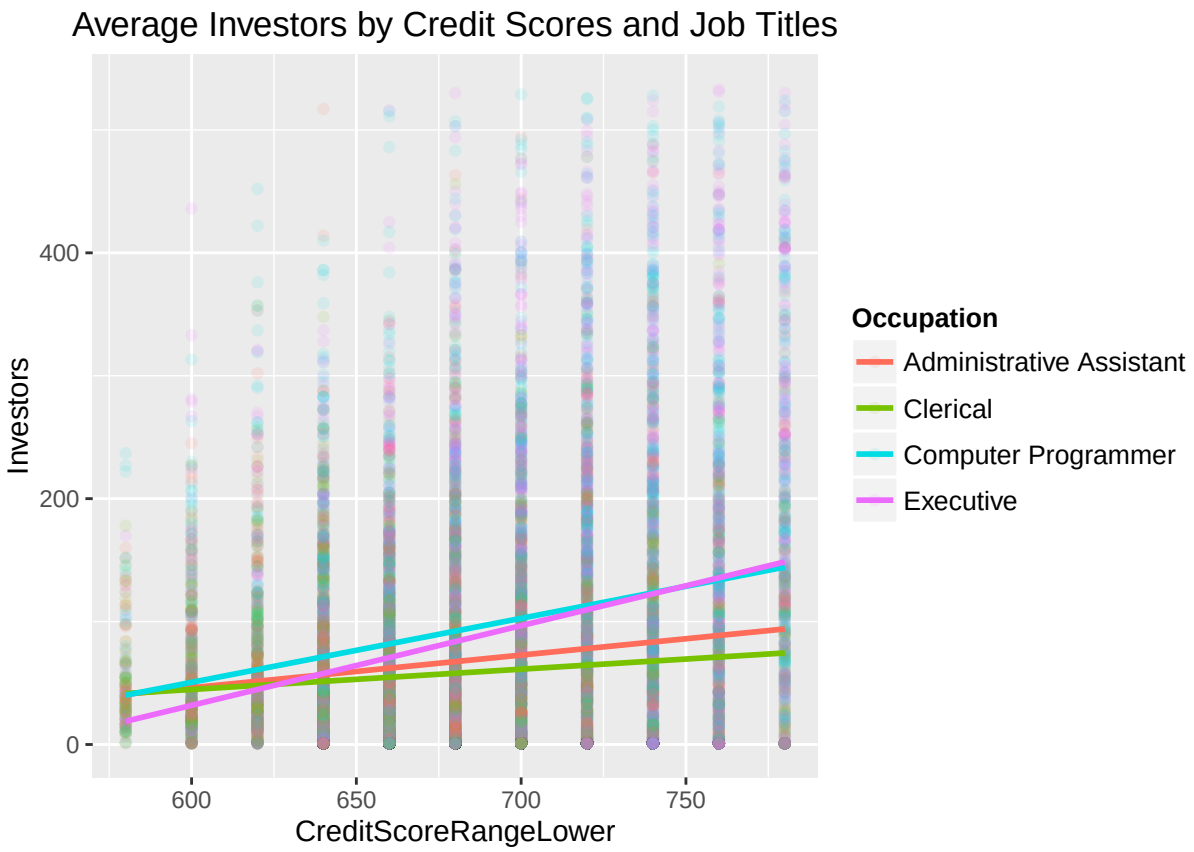
8-2 Plot Two



Description Two

The income range is positively and strongly correlated with risky loan rate, the higher income borrowers have less risky loan rate. The variables Borrower's State also correlate with the risky loan rate, we can see the borrowers in some State have relatively higher risky loan rate in various income range. Thus, income range and Borrower's State can be used in a model to predict the risk of a loan.

8-3 Plot Three



Description Three

The plot indicates that a linear model could be constructed to predict the average investors of variables using 99.5% investors as the outcome variable and credit scores as the predictor variable. Holding credit scores constant, the borrowers with more respectable job titils, such as computer programmer can attract more investors than borrowers with less decent job titles to account for additional variability in investors.

Reflection

The Prosper loan dataset contains information on 113,937 observations across 81 variables. I started by understanding the individual variables in the data set, and then I picked up 36 variables to explored interesting questions and leads as I continued to make observations on plots. Eventually, I explored the investors and loan status across several variables.

While analysing the dataset, one thing made me anxious: the correlations between main features and other variables are not so significant, the highest Pearson's R I've seen is around 0.4. I expected the correlations between variables can be as strong as the tutorial materials, and the plots can easily be interpreted by themselves, so the fact After consulting with my data scientist friends, they told me that for the data in real world, Pearson's R around 0.4 is relatively high, and it's really difficult to see a 0.8 or 0.9 ones. Thus, I feel much more relieved about the dataset and have the confidence to exploring the data.

I found success while creating multivariate plots. As I mentioned above, I found the correlation between two variables are not strong and hard to be awared. However, when I added one more variable into the analysises, holding one of the variables constant, the correlation become much easier to be awared. For example, the correlation with 'IsBorrowerHomeowner' and 'Investors'are not notable in the bivariate section, however, when I added the credit score as the third variables, I can clearly see that homeowners somehow can attract more investors.

With the analysis above, I would like to know if the local ecoomy will affect the risky loan rate. In the multivariate section, I found that the borrowers in certain States are likely to have worse repayment behaviors, which might be caused by local economy. If I have some local economic indicators, such as as Consumer confidence index, I would like to apply the data to the analysis and examine the correlation between the local economy and the repayment behaviors there.