Exploratory data analysis on Prosper Loan Dataset

by Joy Lal Chattaraj

Introduction

In statistics, exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. Exploratory data analysis was promoted by John Tukey to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments. EDA is different from initial data analysis (IDA), which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed. EDA encompasses IDA.

The Dataset

This dataset, provided by Udacity, contains 113,937 loans with 81 variables on each loan, including loan amount, borrower rate (or interest rate), current loan status, borrower income, borrower employment status, borrower credit history, and the latest payment information. It was last updated on 03/11/2014.

Structure of the data

```
## [1] 81
```

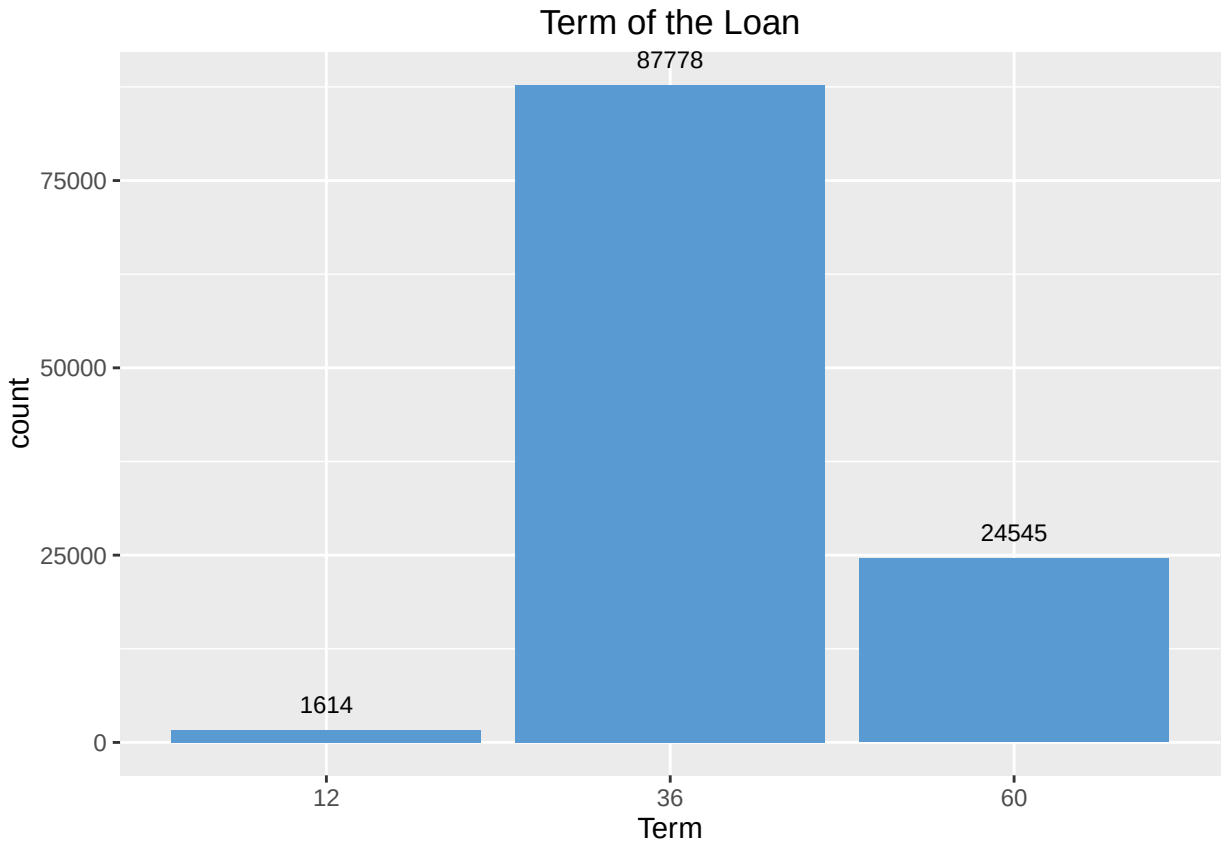
```
## [1] 113937
```

```
## 'data.frame':    113937 obs. of  81 variables:
##  $ ListingKey                          : Factor w/ 113066 levels "0000354648209428 2EF90E5",..: 7180 7
##  $ ListingNumber                       : int  193129 1209647 81716 658116 909464 1074836 750899 768193
##  $ ListingCreationDate                 : Factor w/ 113064 levels "2005-11-09 20:44:28.847000000",..:
##  $ 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
##  $ ClosedDate                          : Factor w/ 2803 levels "","2005-11-25 00:00:00",..: 1138 1 120
##  $ BorrowerAPR                         : num  0.165 0.12 0.283 0.125 0.246 ...
##  $ BorrowerRate                        : num  0.158 0.092 0.275 0.0974 0.2085 ...
##  $ LenderYield                         : num  0.138 0.082 0.24 0.0874 0.1985 ...
##  $ EstimatedEffectiveYield             : num  NA 0.0796 NA 0.0849 0.1832 ...
##  $ EstimatedLoss                       : num  NA 0.0249 NA 0.0249 0.0925 ...
##  $ EstimatedReturn                     : num  NA 0.0547 NA 0.06 0.0907 ...
##  $ ProsperRating..numeric.             : int  NA 6 NA 6 3 5 2 4 7 7 ...
##  $ ProsperRating..Alpha.               : Factor w/ 8 levels "","A","AA","B",..: 1 2 1 2 6 4 7 5 3 3 .
##  $ ProsperScore                        : num  NA 7 NA 9 4 10 2 4 9 11 ...
##  $ 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
##  $ 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 ...
##  $ CurrentlyInGroup                    : Factor w/ 2 levels "False","True": 2 1 2 1 1 1 1 1 1 1 ...
```

```
##  $ GroupKey                          : Factor w/ 707 levels "","00343376901312423168731",..: 1 1 335
##  $ DateCreditPulled                  : Factor w/ 112992 levels "2005-11-09 00:30:04.487000000",..: 1
##  $ 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 ...
##  $ OpenRevolvingMonthlyPayment       : num  24 389 0 115 220 1410 214 101 219 219 ...
##  $ InquiriesLast6Months              : int  3 3 0 0 1 0 0 3 1 1 ...
##  $ TotalInquiries                    : num  3 5 1 1 9 2 0 16 6 6 ...
##  $ 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 ...
##  $ PublicRecordsLast12Months         : int  0 0 NA 0 0 0 0 0 0 0 ...
##  $ RevolvingCreditBalance            : num  0 3989 NA 1444 6193 ...
##  $ 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 ...
##  $ LoanKey                           : Factor w/ 113066 levels "00003683605746079487FF7",..: 100337
##  $ TotalProsperLoans                 : int  NA NA NA NA 1 NA NA NA NA NA ...
##  $ TotalProsperPaymentsBilled        : int  NA NA NA NA 11 NA NA NA NA NA ...
##  $ OnTimeProsperPayments             : int  NA NA NA NA 11 NA NA NA NA NA ...
##  $ ProsperPaymentsLessThanOneMonthLate: int  NA NA NA NA 0 NA NA NA NA NA ...
##  $ ProsperPaymentsOneMonthPlusLate   : int  NA NA NA NA 0 NA NA NA NA NA ...
##  $ ProsperPrincipalBorrowed          : num  NA NA NA NA 11000 NA NA NA NA NA ...
##  $ ProsperPrincipalOutstanding       : num  NA NA NA NA 9948 ...
##  $ ScorexChangeAtTimeOfListing       : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ LoanCurrentDaysDelinquent         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ LoanFirstDefaultedCycleNumber     : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ LoanMonthsSinceOrigination        : int  78 0 86 16 6 3 11 10 3 3 ...
##  $ LoanNumber                        : int  19141 134815 6466 77296 102670 123257 88353 90051 121268
##  $ LoanOriginalAmount                : int  9425 10000 3001 10000 15000 15000 3000 10000 10000 10000
##  $ LoanOriginationDate               : Factor w/ 1873 levels "2005-11-15 00:00:00",..: 426 1866 260
##  $ LoanOriginationQuarter            : Factor w/ 33 levels "Q1 2006","Q1 2007",..: 18 8 2 32 24 33 1
##  $ MemberKey                         : Factor w/ 90831 levels "00003397697413387CAF966",..: 11071 10
##  $ MonthlyLoanPayment                : num  330 319 123 321 564 ...
##  $ LP_CustomerPayments               : num  11396 0 4187 5143 2820 ...
##  $ LP_CustomerPrincipalPayments      : num  9425 0 3001 4091 1563 ...
##  $ LP_InterestandFees                : num  1971 0 1186 1052 1257 ...
##  $ LP_ServiceFees                    : num  -133.2 0 -24.2 -108 -60.3 ...
##  $ LP_CollectionFees                 : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ LP_GrossPrincipalLoss             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ LP_NetPrincipalLoss               : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ LP_NonPrincipalRecoverypayments   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ PercentFunded                     : num  1 1 1 1 1 1 1 1 1 1 ...
```

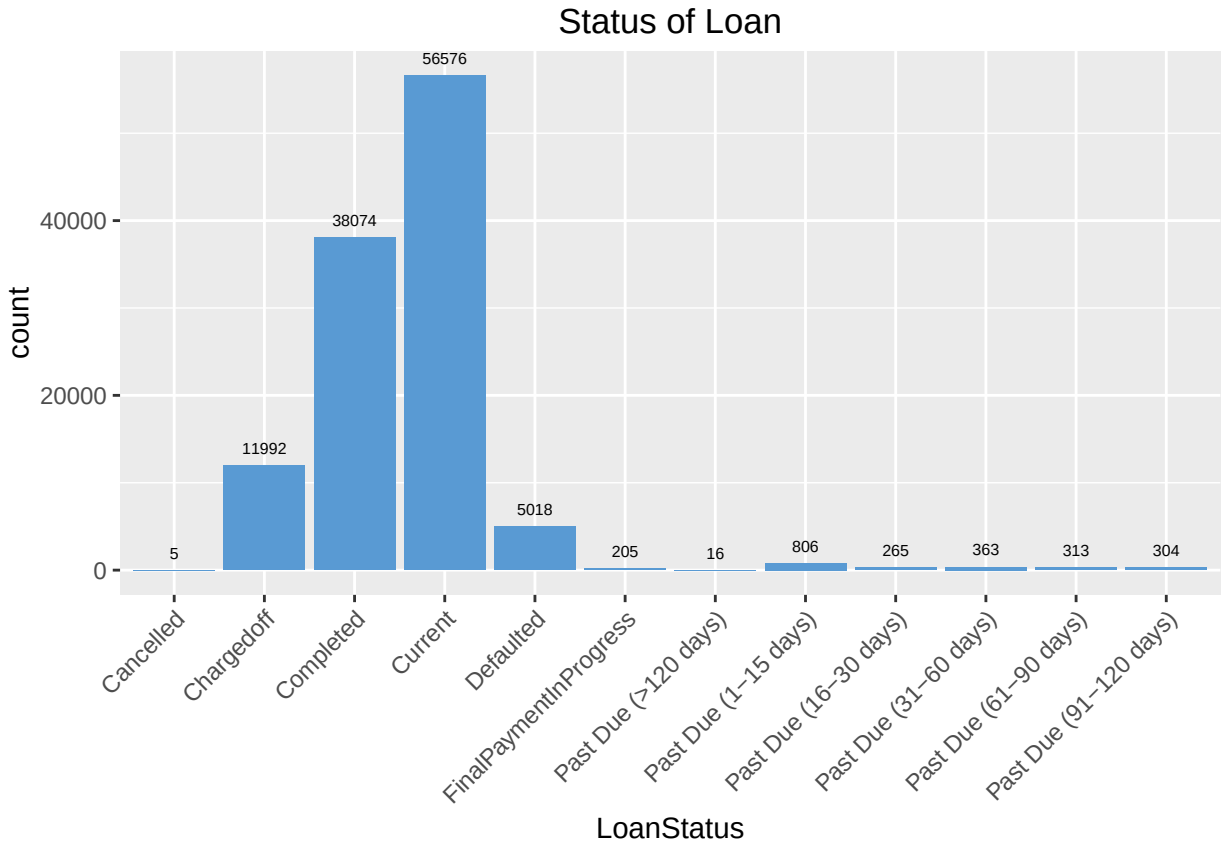
```
##  $ Recommendations                  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ InvestmentFromFriendsCount       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ InvestmentFromFriendsAmount      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Investors                        : int  258 1 41 158 20 1 1 1 1 1 ...
```

Univariate Plots Section



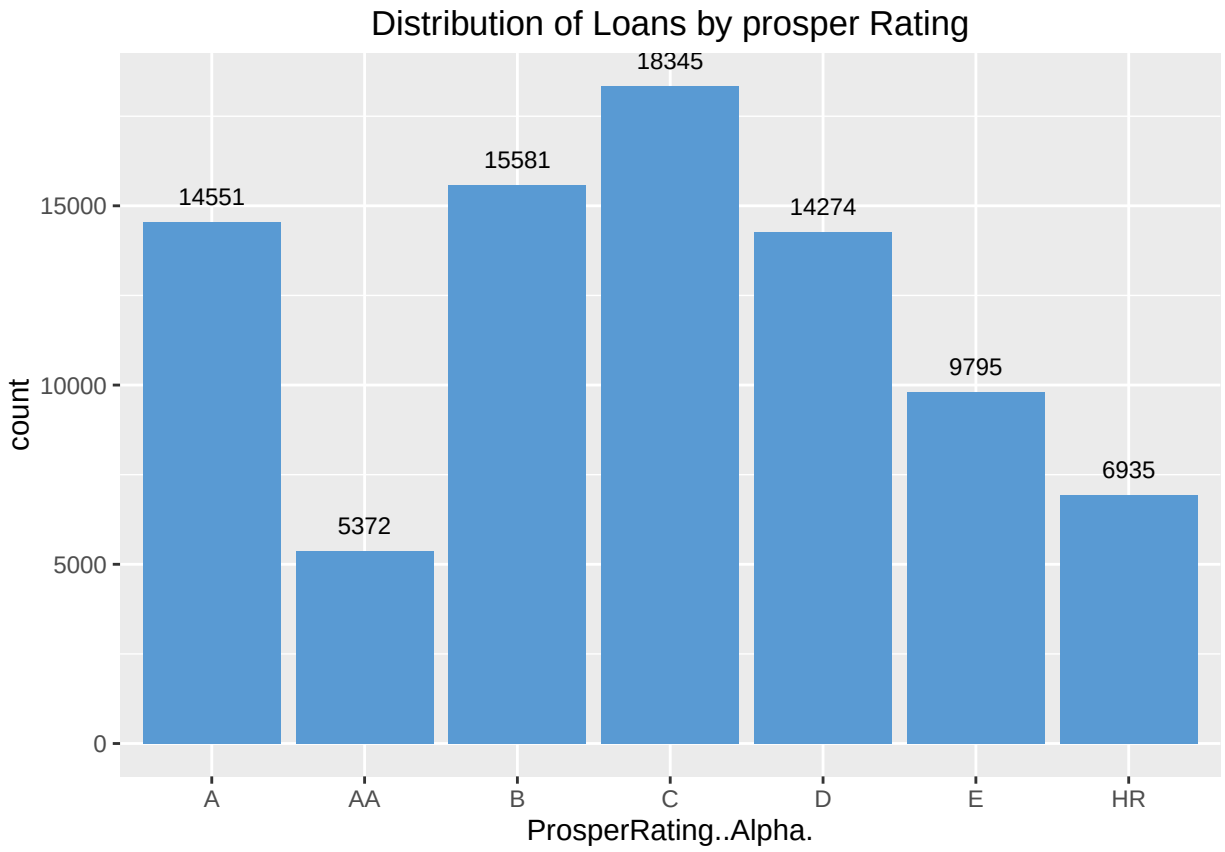

```
##    12    36    60
##  1614 87778 24545
```

Major part of the Loans belong to the Term of 36 Months.



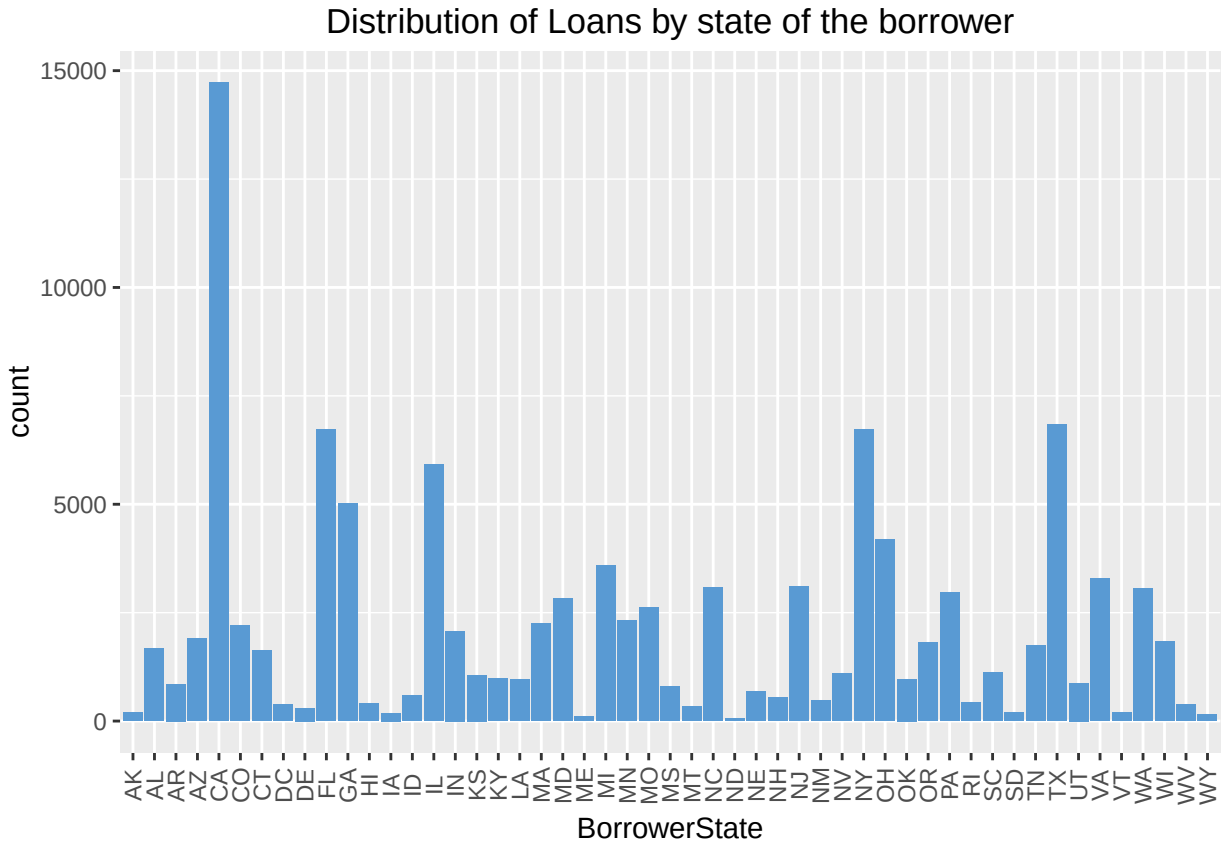

```
##             Cancelled            Chargedoff             Completed
##                     5                 11992                 38074
##               Current               Defaulted FinalPaymentInProgress
##                 56576                  5018                   205
##   Past Due (>120 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
```

Although a considerable amount of loans have been completed but most of them are yet to be completed.



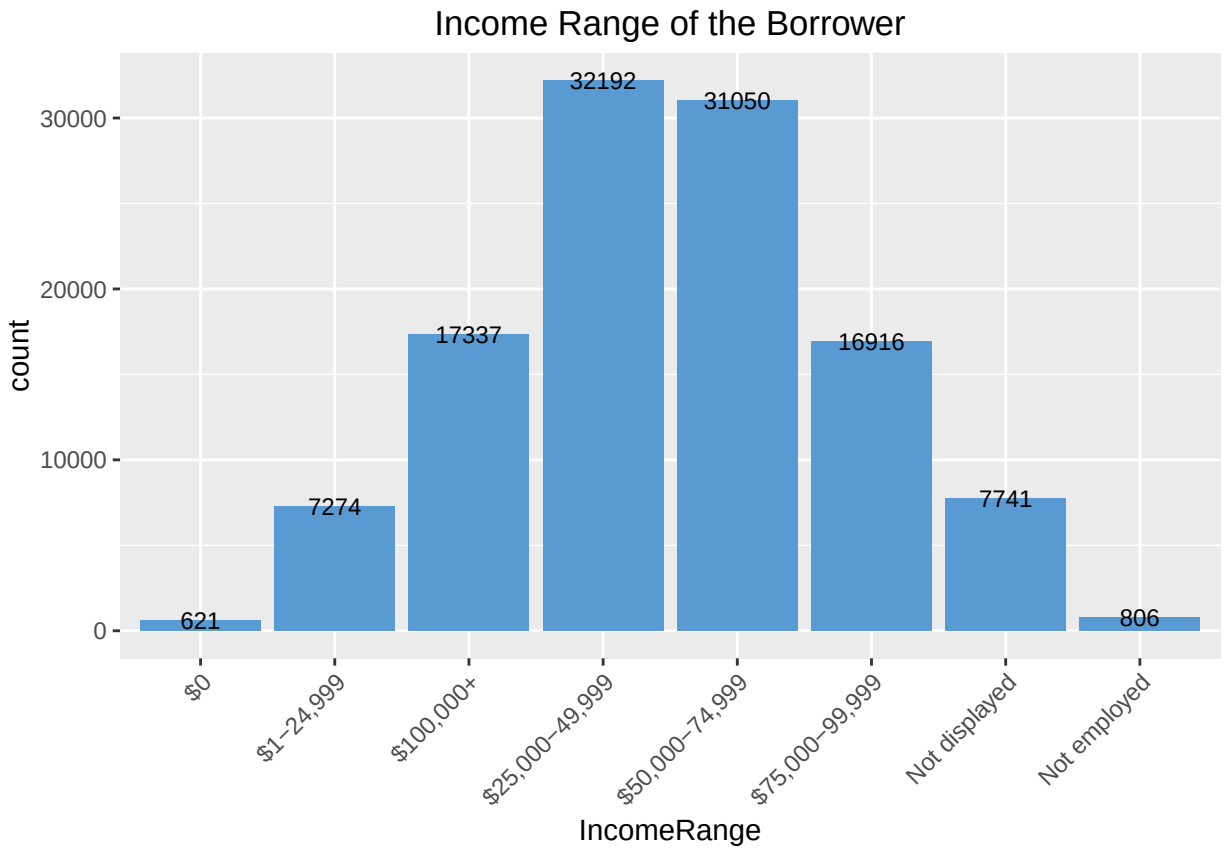

```
##           A    AA     B     C     D     E    HR
## 29084 14551  5372 15581 18345 14274  9795  6935
```

Most of the Loans are unrated but among the rated ones A, B, C and D ratings are most prevalent.



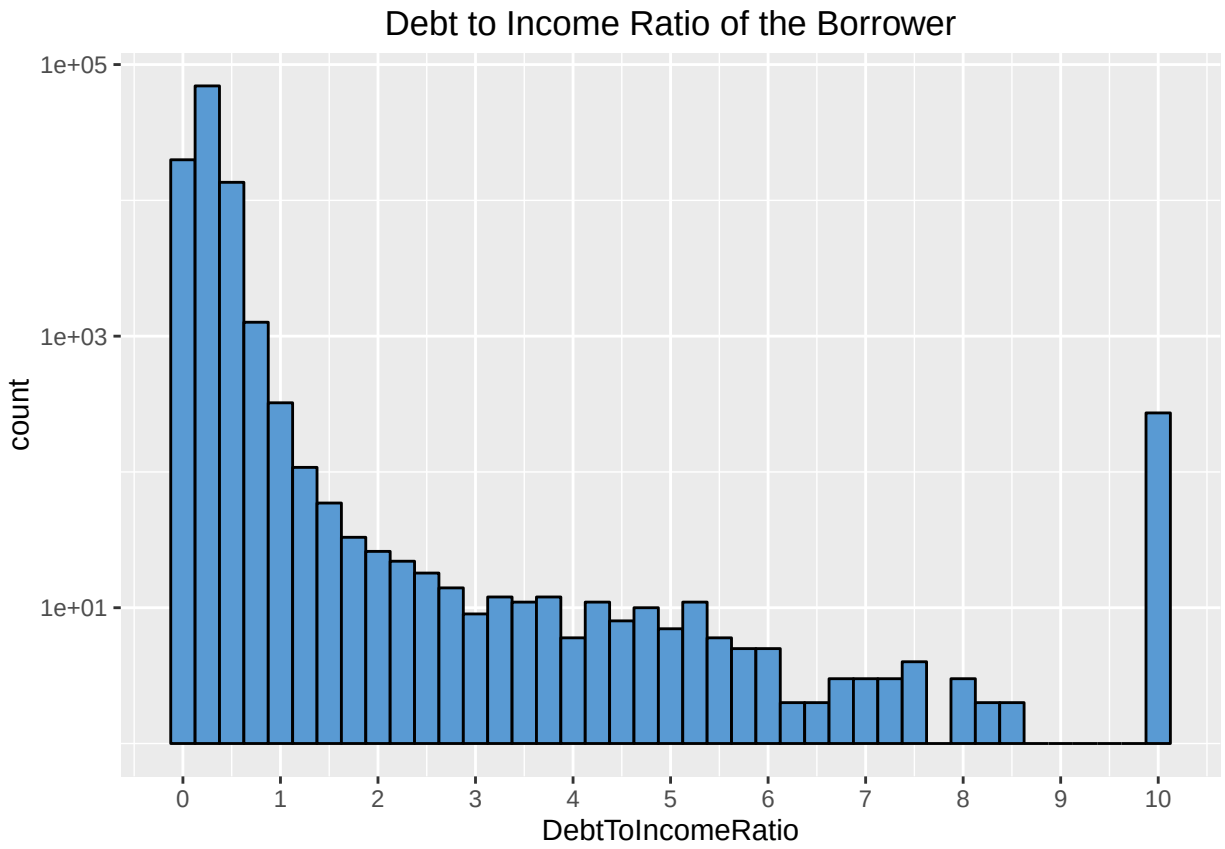

```
##          AK    AL    AR    AZ    CA    CO    CT    DC    DE    FL    GA
##  5515   200  1679   855  1901 14717  2210  1627   382   300  6720  5008
##    HI    IA    ID    IL    IN    KS    KY    LA    MA    MD    ME    MI
##   409   186   599  5921  2078  1062   983   954  2242  2821   101  3593
##    MN    MO    MS    MT    NC    ND    NE    NH    NJ    NM    NV    NY
##  2318  2615   787   330  3084    52   674   551  3097   472  1090  6729
##    OH    OK    OR    PA    RI    SC    SD    TN    TX    UT    VA    VT
##  4197   971  1817  2972   435  1122   189  1737  6842   877  3278   207
##    WA    WI    WV    WY
##  3048  1842   391   150
```

A few peaks can be observed showing the states where most of the borrowers live.



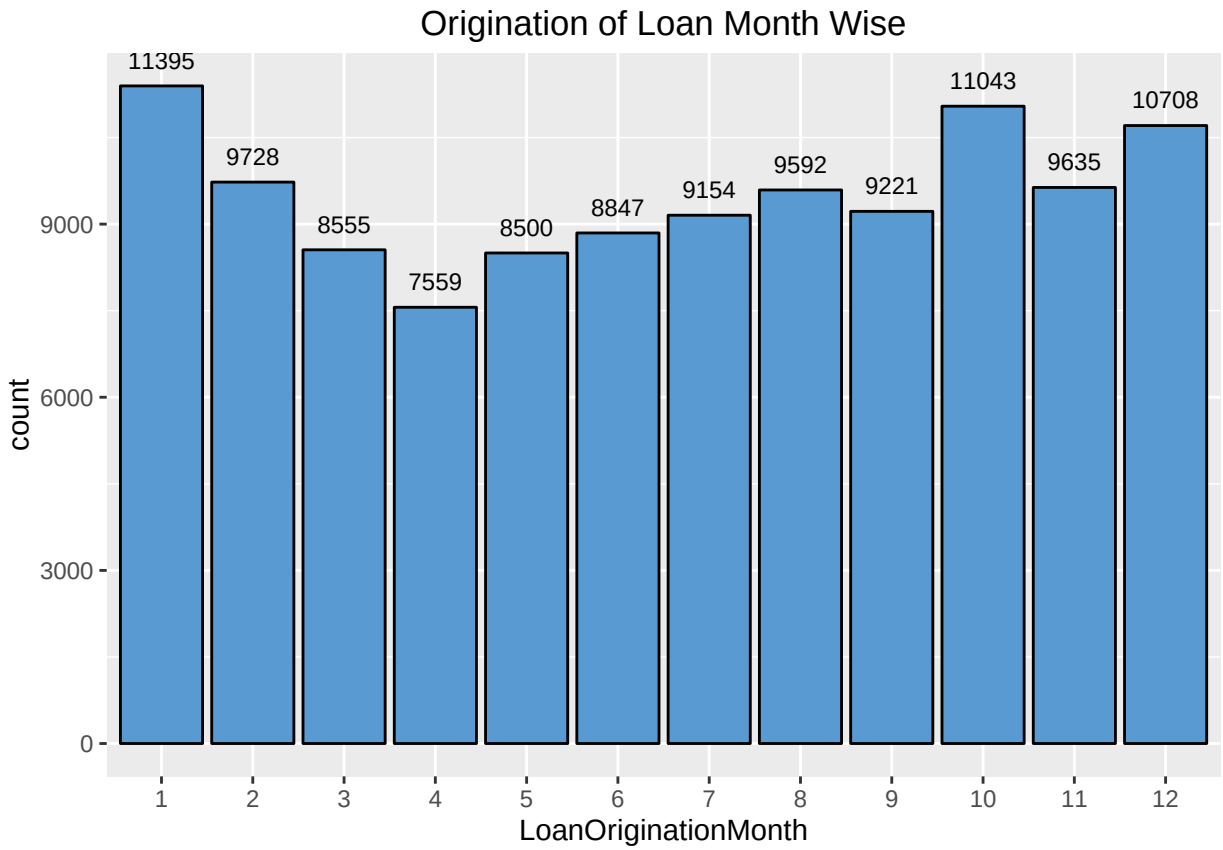

```
##             $0      $1-24,999      $100,000+ $25,000-49,999 $50,000-74,999
##            621           7274          17337          32192          31050
## $75,000-99,999  Not displayed   Not employed
##          16916           7741            806
```

Its a normal distribution with most of the borrowers lying in the range $25,000 to $75,000.

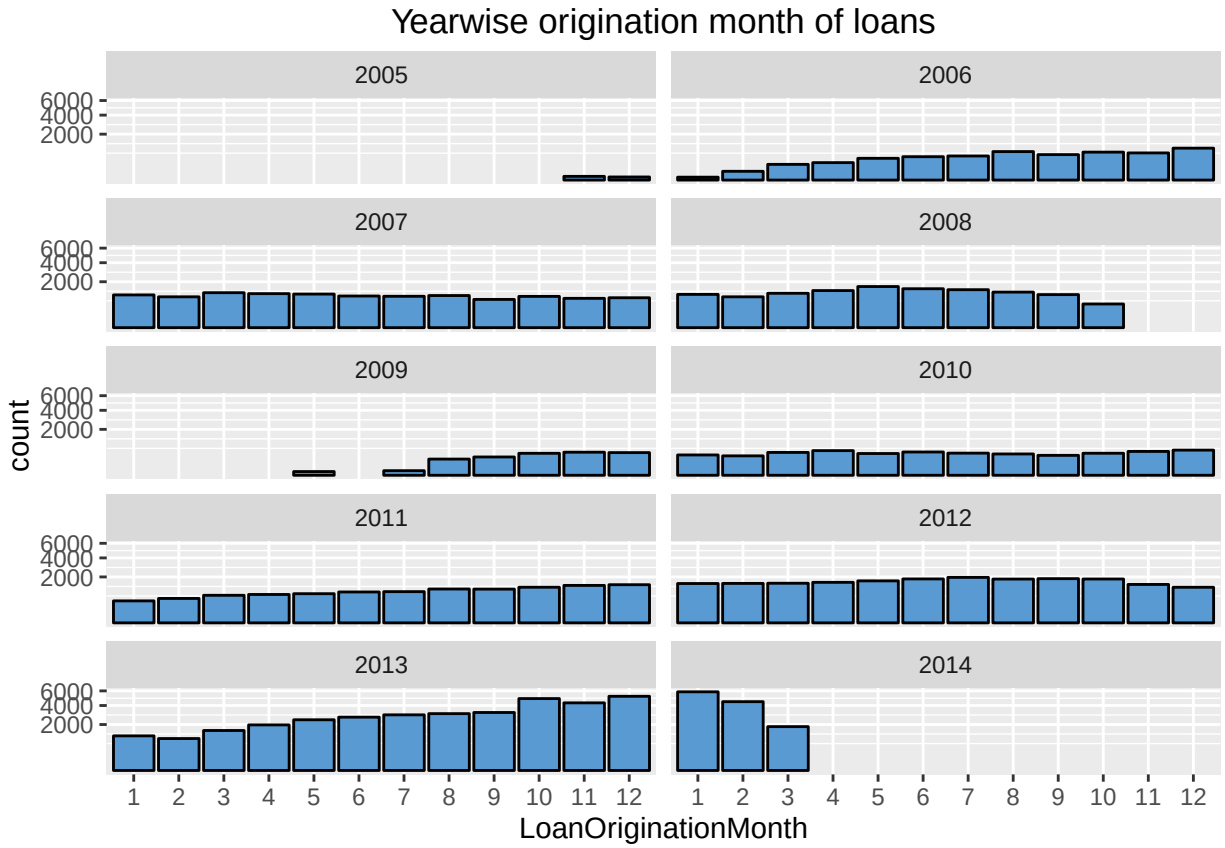



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   0.140   0.220   0.276   0.320  10.010    8554
```

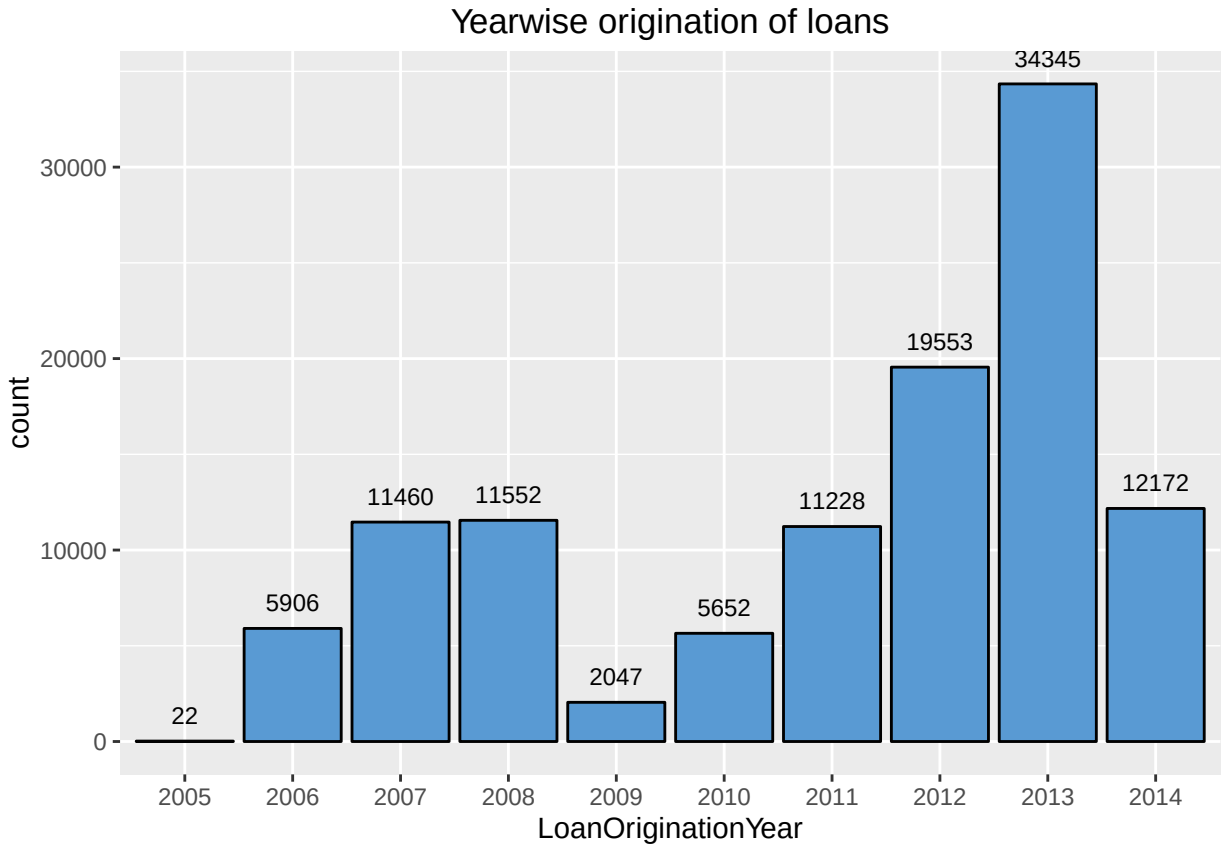
A Positively skewed distribution is observerd with most of the borrowers having a low debt to income ratio (Debt is not much higher than Income, thus can be repaid easily.)



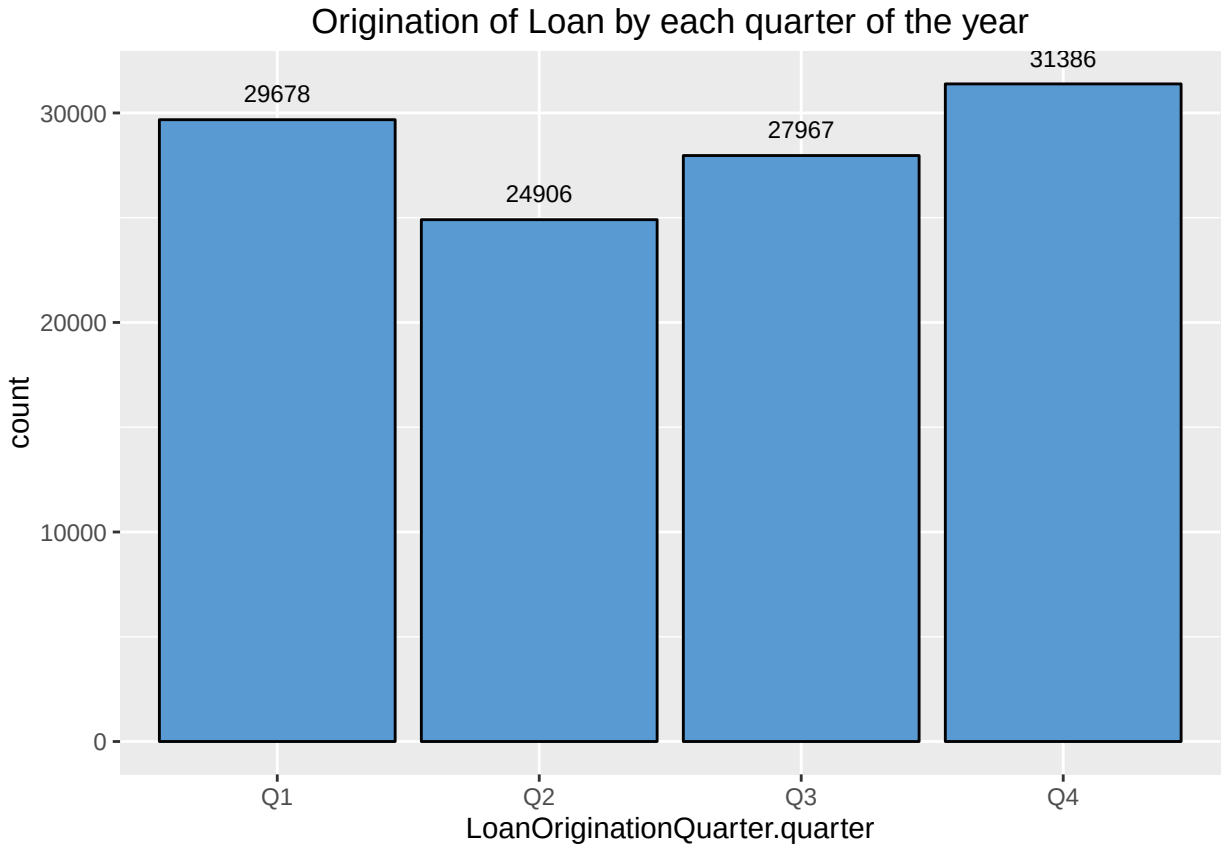
Nothing Interesting to observe except that March-April-May have relatively less Loans sanctioned.



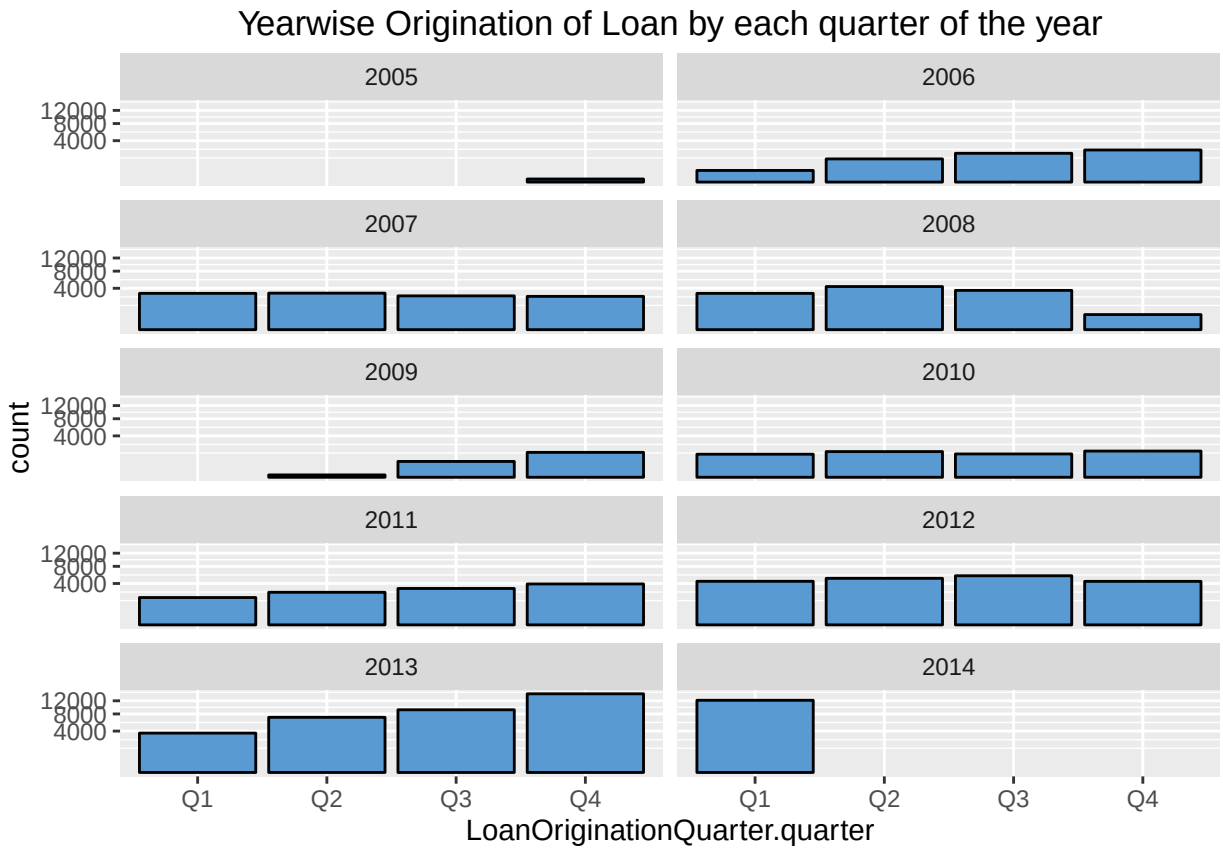
No pattern observed by the months but we can see that as years pass by, there is an increase in the Loans except in 2009 where the number of loans are significantly lower.



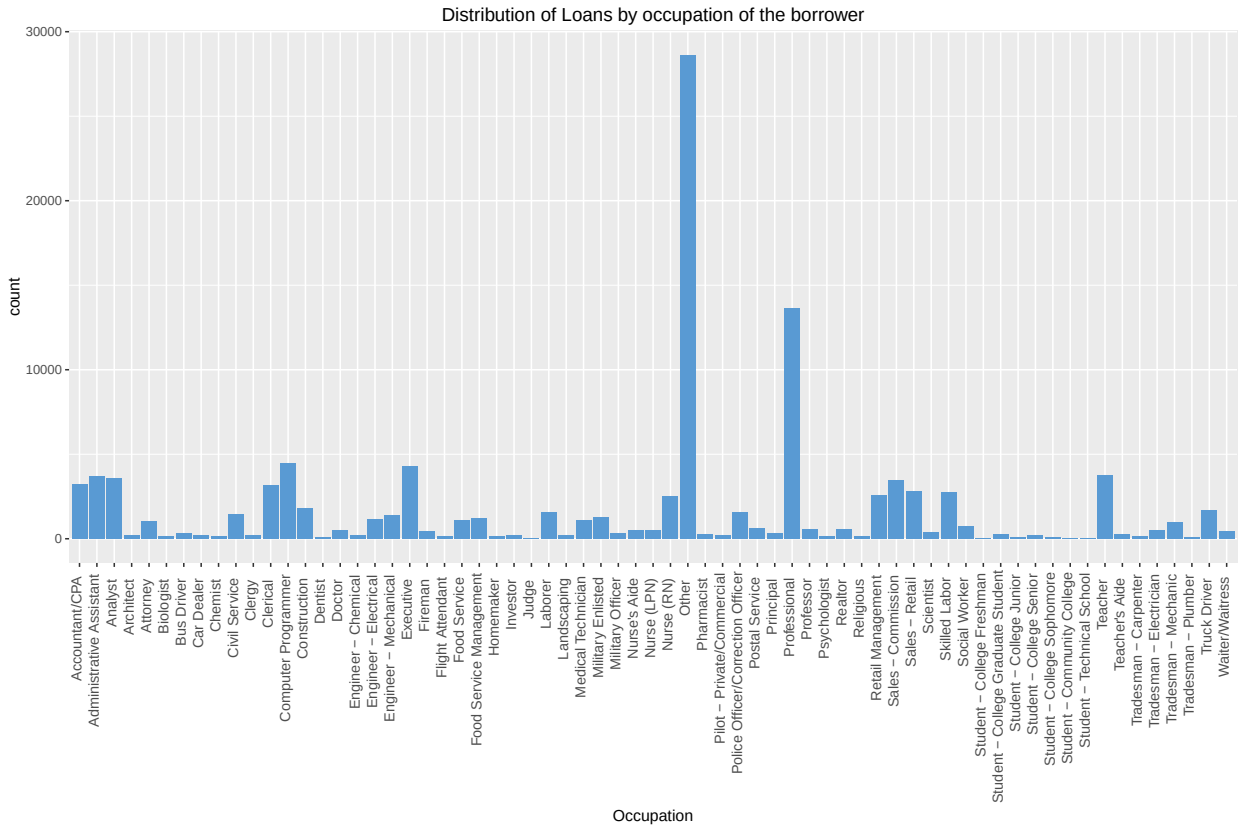
We see a sudden decrease in 2009.Doing a quick Google search, we now understand the period of October 15, 2008 to July 13, 2009 is Prosper's QuietSEC Period, from which they are suspended for lending activities upon SEC approval. Prosper relaunched in July 2009 and this can be seen in the previous plot.There's also a large number of loans sanctioned in 2013.



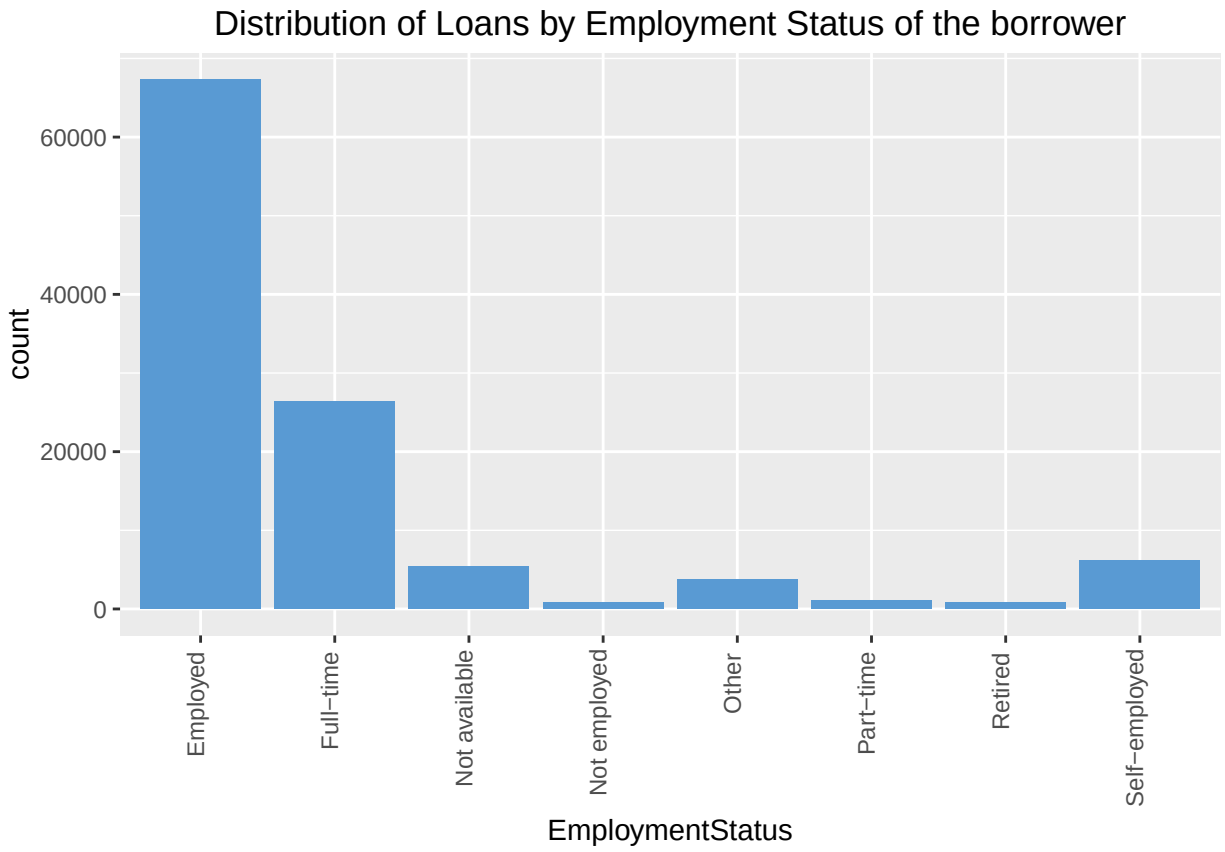
Not much to observe, its nearly an even distribution.



Loans sanctioned either remain same over all quarters or tend to increase after each quarter in some years.

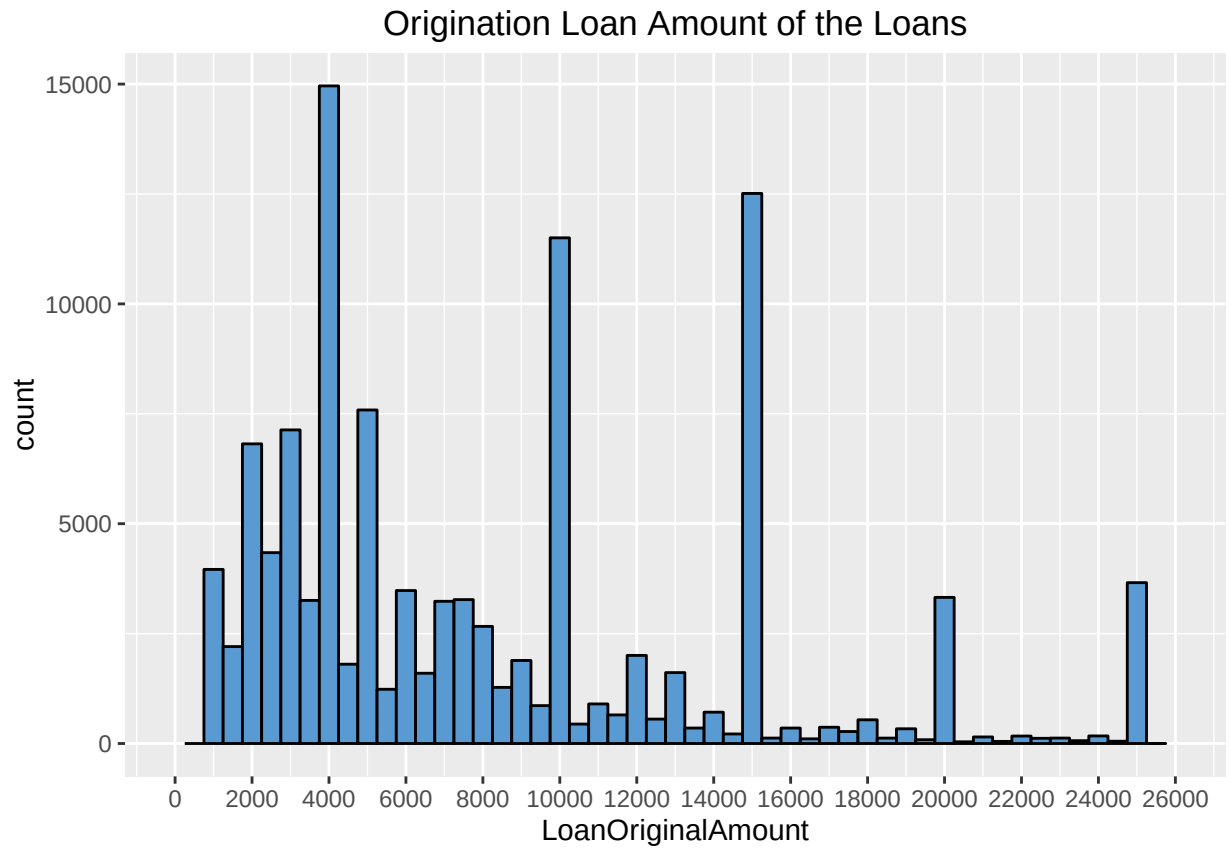


Most Loans fall into the category of others but the category "Professionals" has a lot of loans under its name.



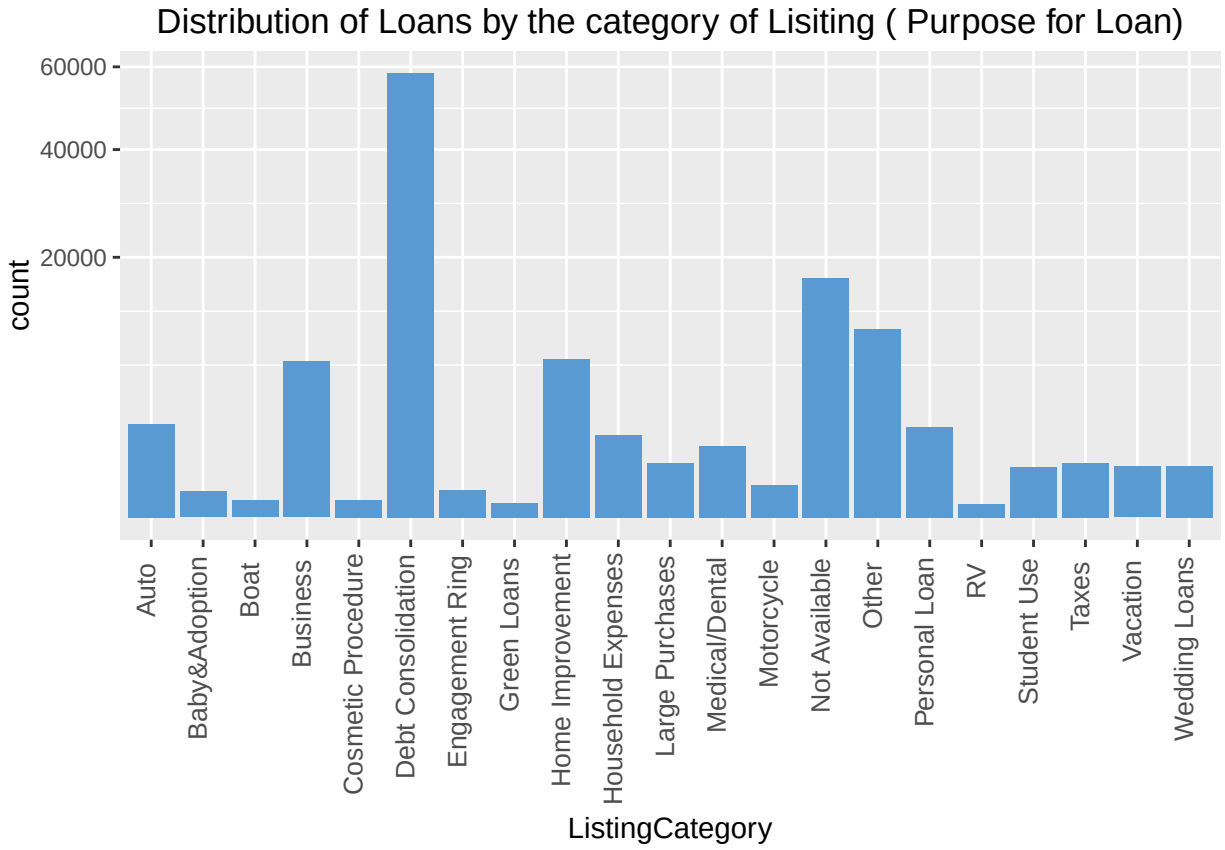

```
##                  Employed     Full-time Not available  Not employed
##          2255         67322         26355          5347           835
##         Other     Part-time       Retired Self-employed
##          3806          1088           795          6134
```

Most of the borrowers are earning which is obvious, a few are unemployed (student loans maybe) and a few retired.

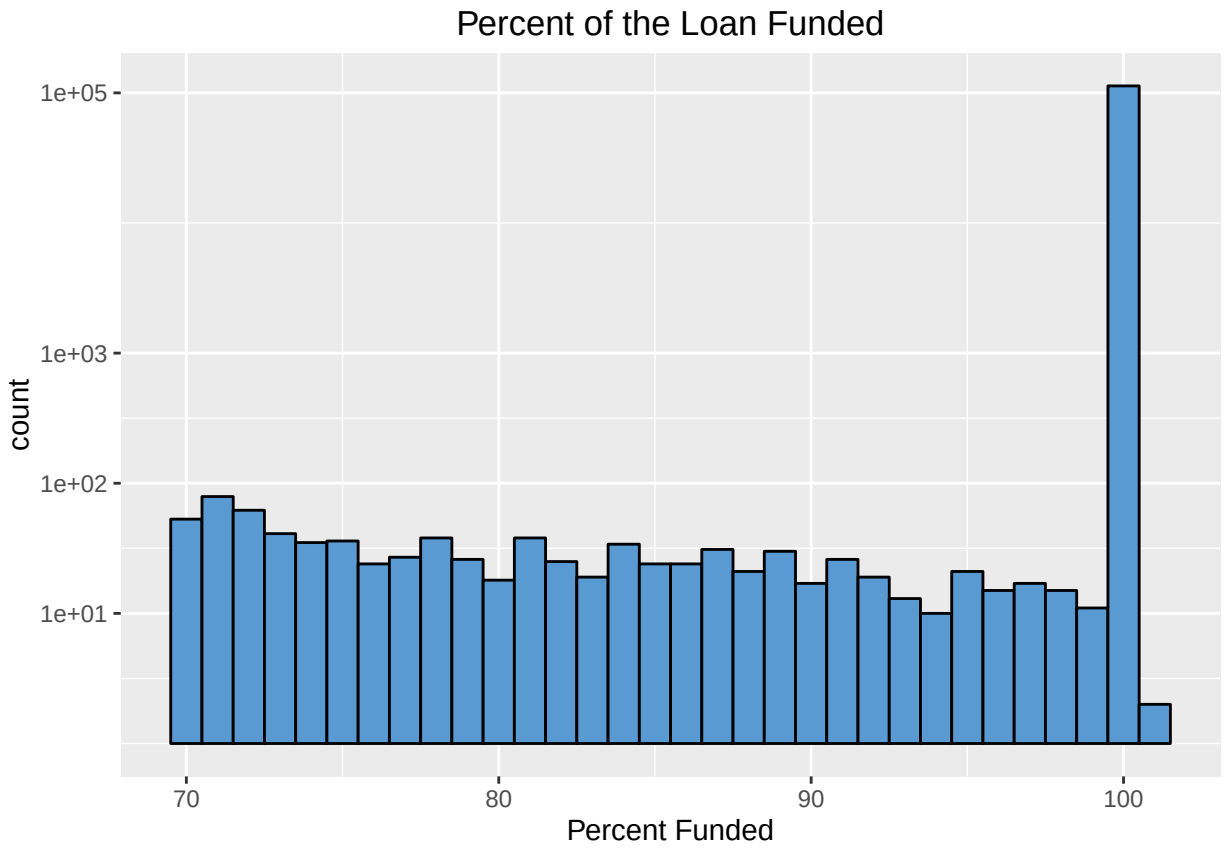



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1000    4000    6500    8337   12000   35000
```

The data seems to be quite discrete, peaks are observed at values 5000, 10,000, 15,000 20,000 and so on.. which is actually expected. Also, the minimum loan amount is $1,000.

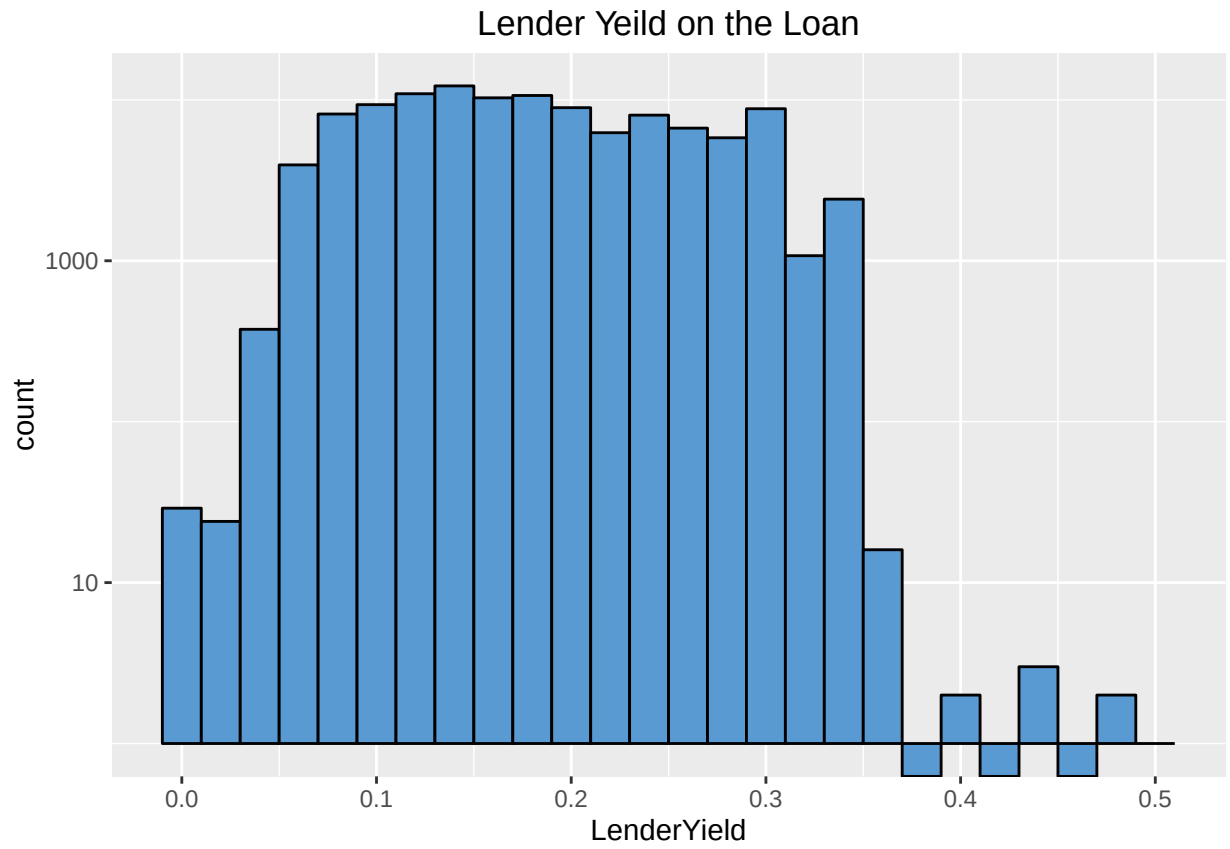


Purpose for most loans are unavailable whereas 'Debt Consolidation' is the purpose for most of the loans.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.7000  1.0000  1.0000  0.9986  1.0000  1.0120
```

Since its their business, most of the loans are entirely funded, very few loans have a funding % less than 90.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.0100  0.1242  0.1730  0.1827  0.2400  0.4925
```

The different Lender Yields must have been because of different borrower rates for different categories of loan. The Median is 18.27%, quite interesting.

Univariate Analysis

What is the structure of your dataset?

The dataset comprises of 81 variables with 113937 observations. The loans cover the period 2005 & 2014. Variables are of classes int, numeric, date, and factor. The variable contain information about the borrower, every detail about the loan and all the figures that effect the lender.

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

We can spit the variable into the categories mentioned above.

For Borrower, I believe the Prosper Score, Proser Rating are the main indicators of a quality of borrowers. Further bivariate analysis would help to learn about these 2.

For Lender, I now understand Lender Yield is the most important factor for investor.

And the date and period of the loans depict information about the buisness.

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

A better idea about the assets of the borrower would make it clear how the loans are rated.

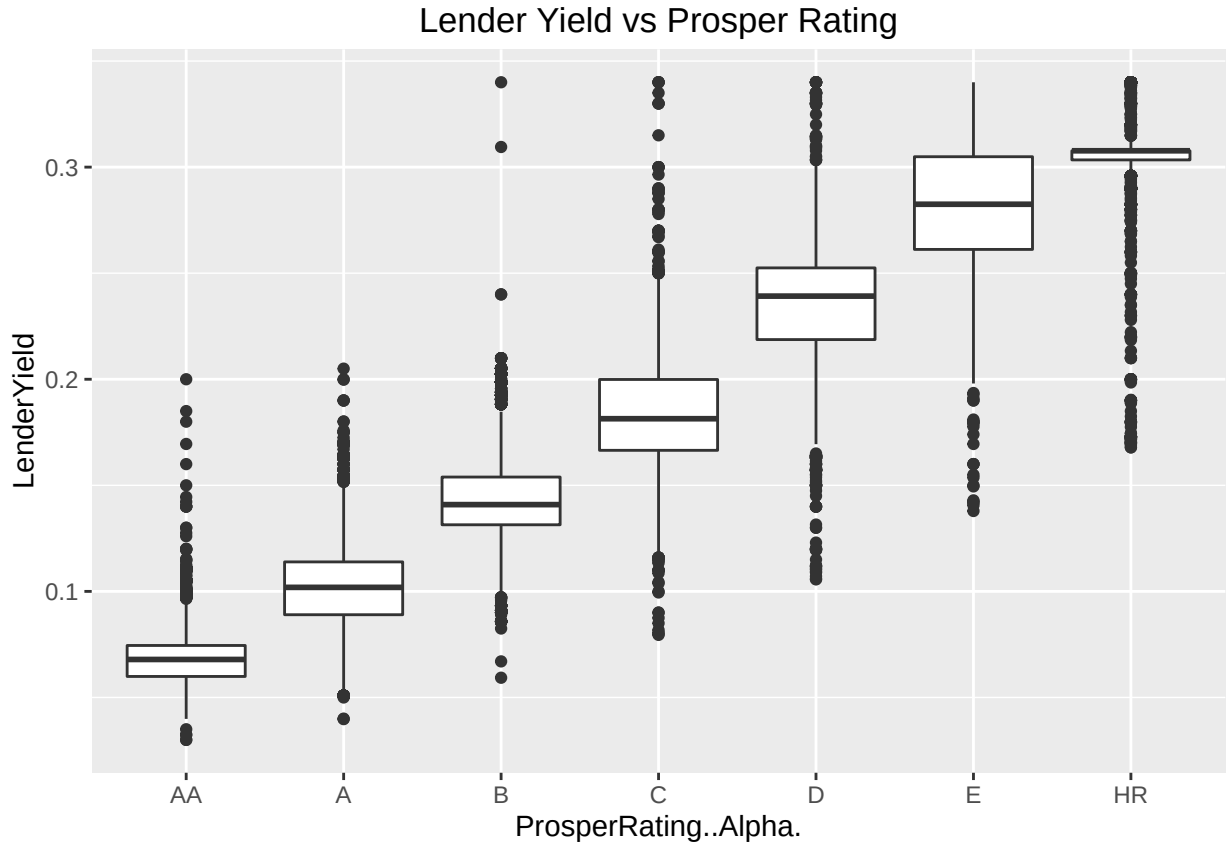
Did you create any new variables from existing variables in the dataset?

Yes I created some new variables and factorized a few others. I decided to split the date into month and year and quarter of the year into quarter and year seperately.

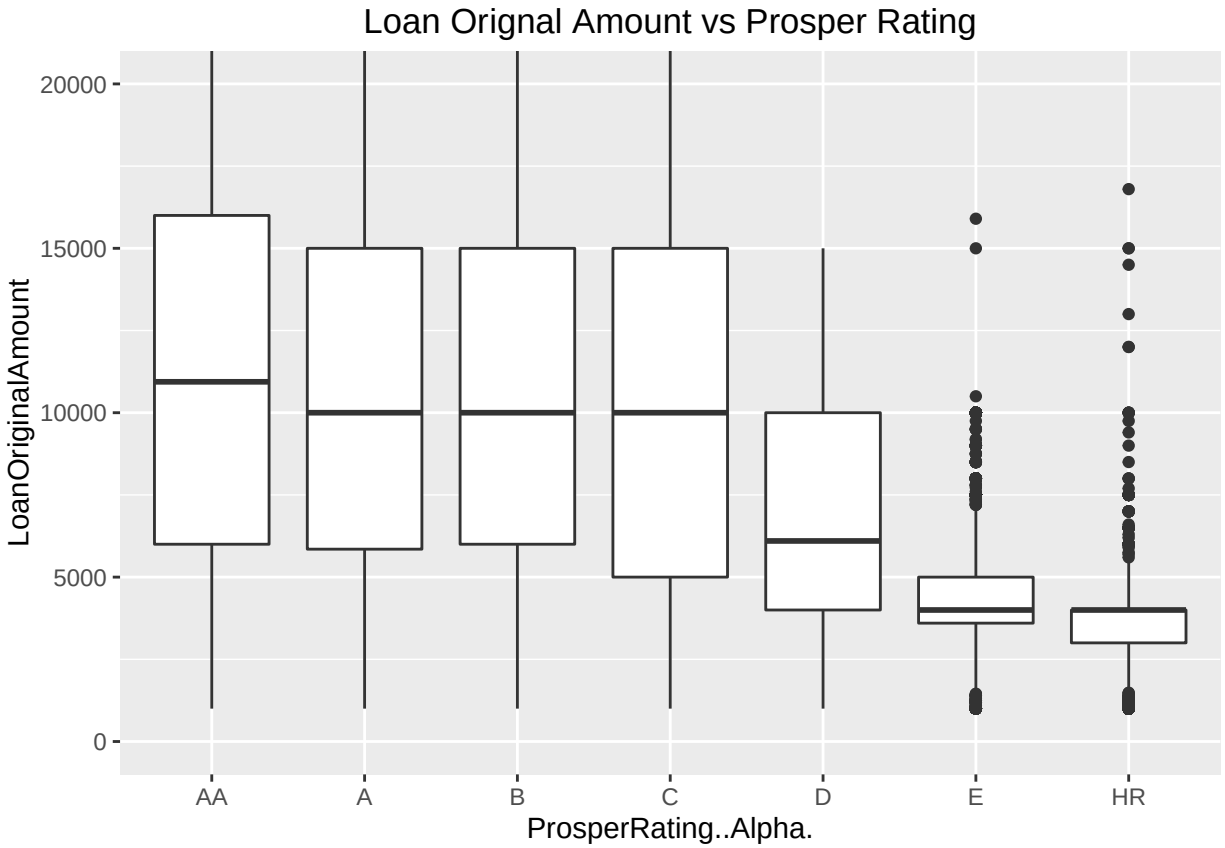
Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

Yes I saw some unusual spikes which I later found belonged to the category for which data was unavailable (N/A). The LoanOriginalAmount distribution had some spikes at regular intervals but I could figure out the reason. I also factorized some variable so that the plotting functions treat them accordingly as they are categorical variables.

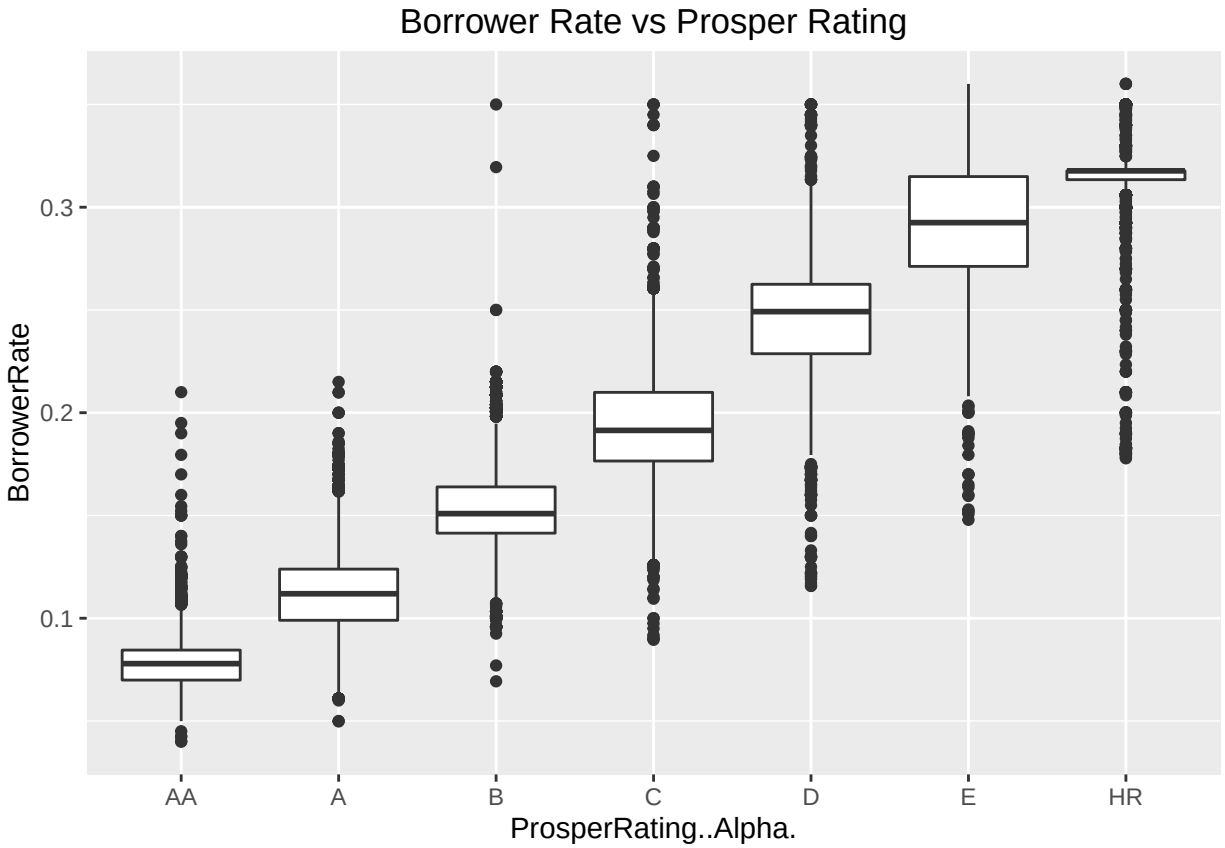
Bivariate Plots Section



There's a clear relation between the two, the higher the lender yield, poorer is the Prosper Rating.

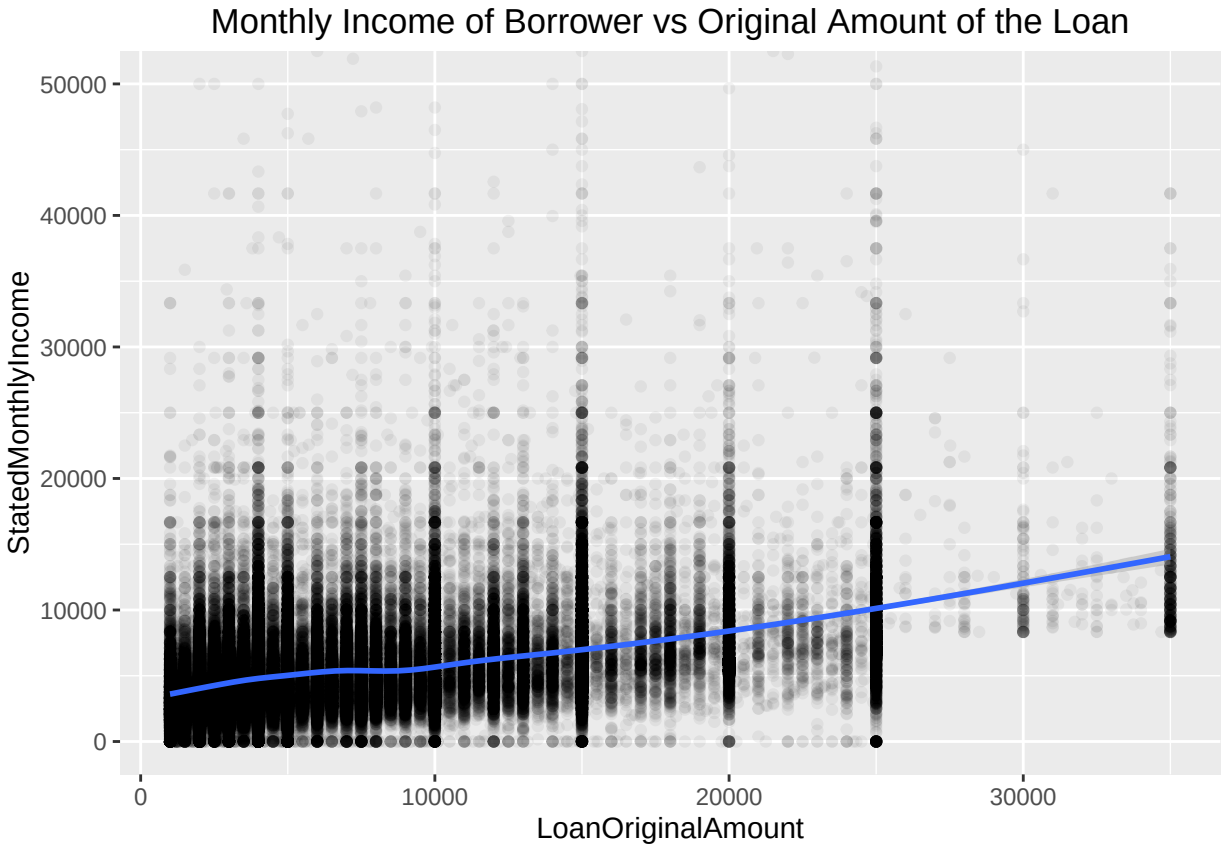


The Median Loan Amount appears to decrease as the rating gets poorer.

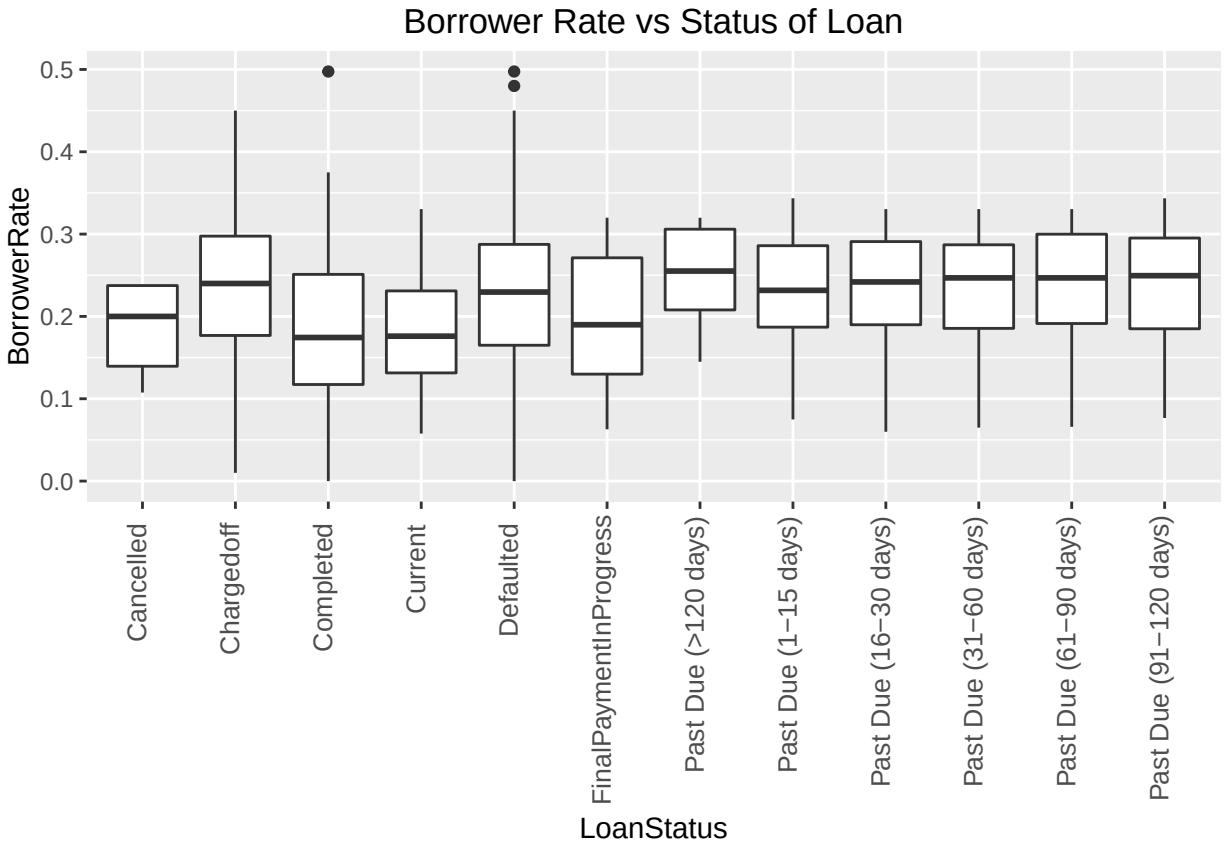


The borrower rate actually is quite similar to the Lender Yield. So, the relationship observed here is also similar to the lender yield.

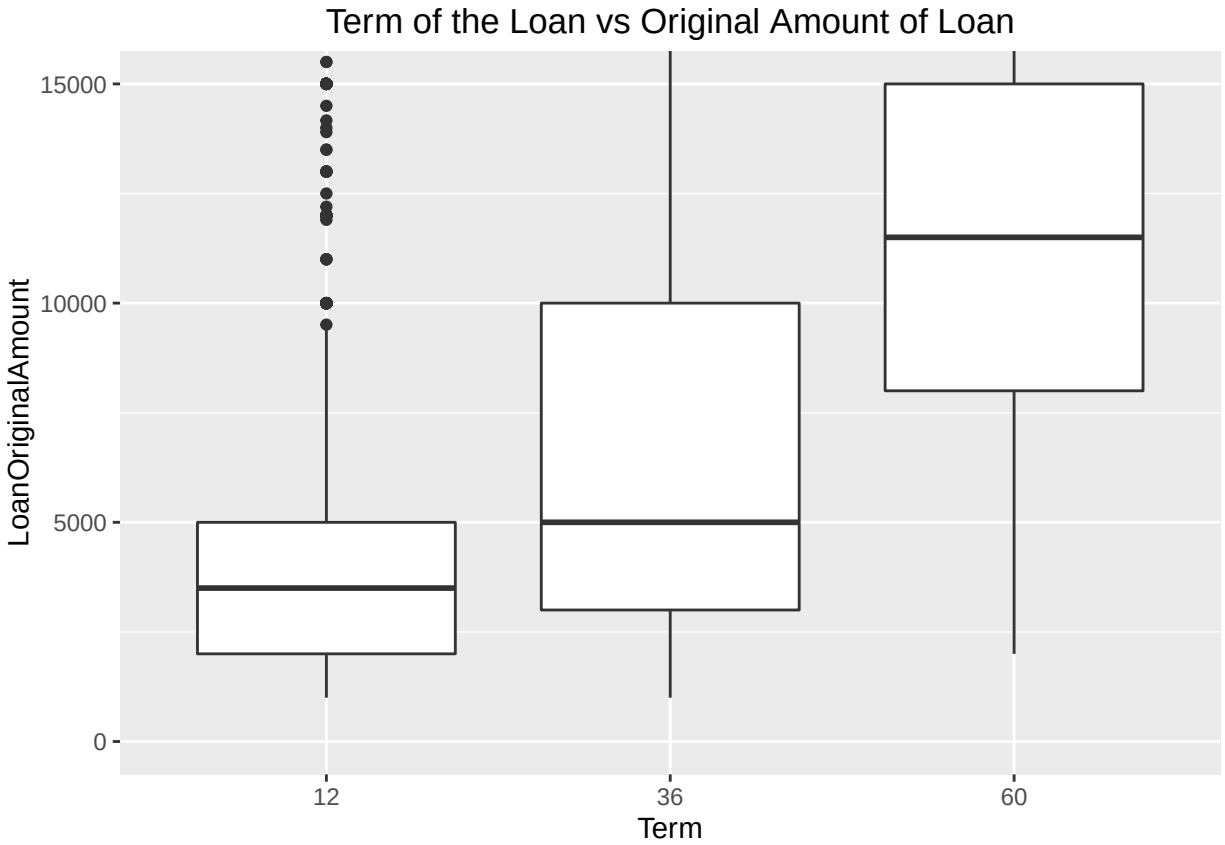
The Monthly Income of the borrower shows a slight decrease as the rating gets poorer.



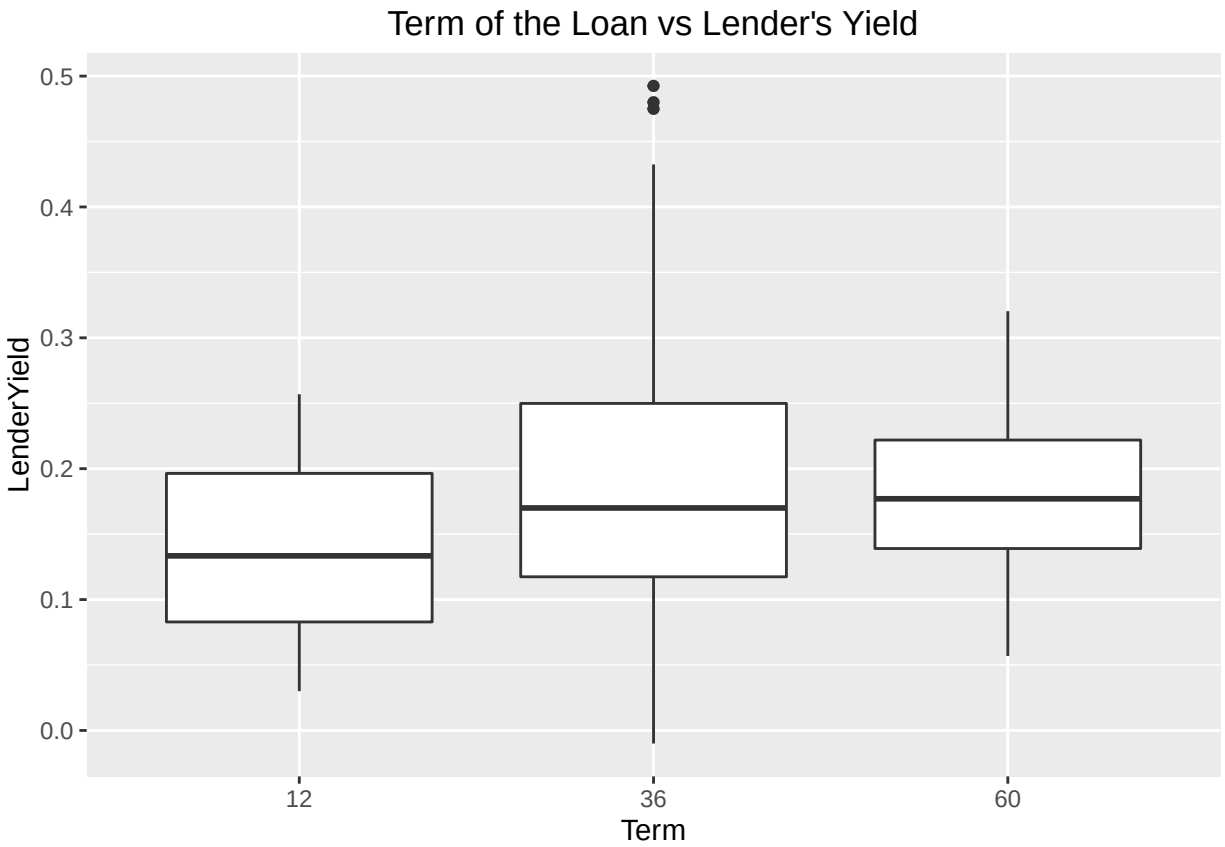
It appears that as the Loan Amount increases, the median monthly income also increases (which for obvious reasons should increase). Another interesting trend is that the Loan amount values appear to be discrete as the Loan amounts are usually round figures rather than random values.



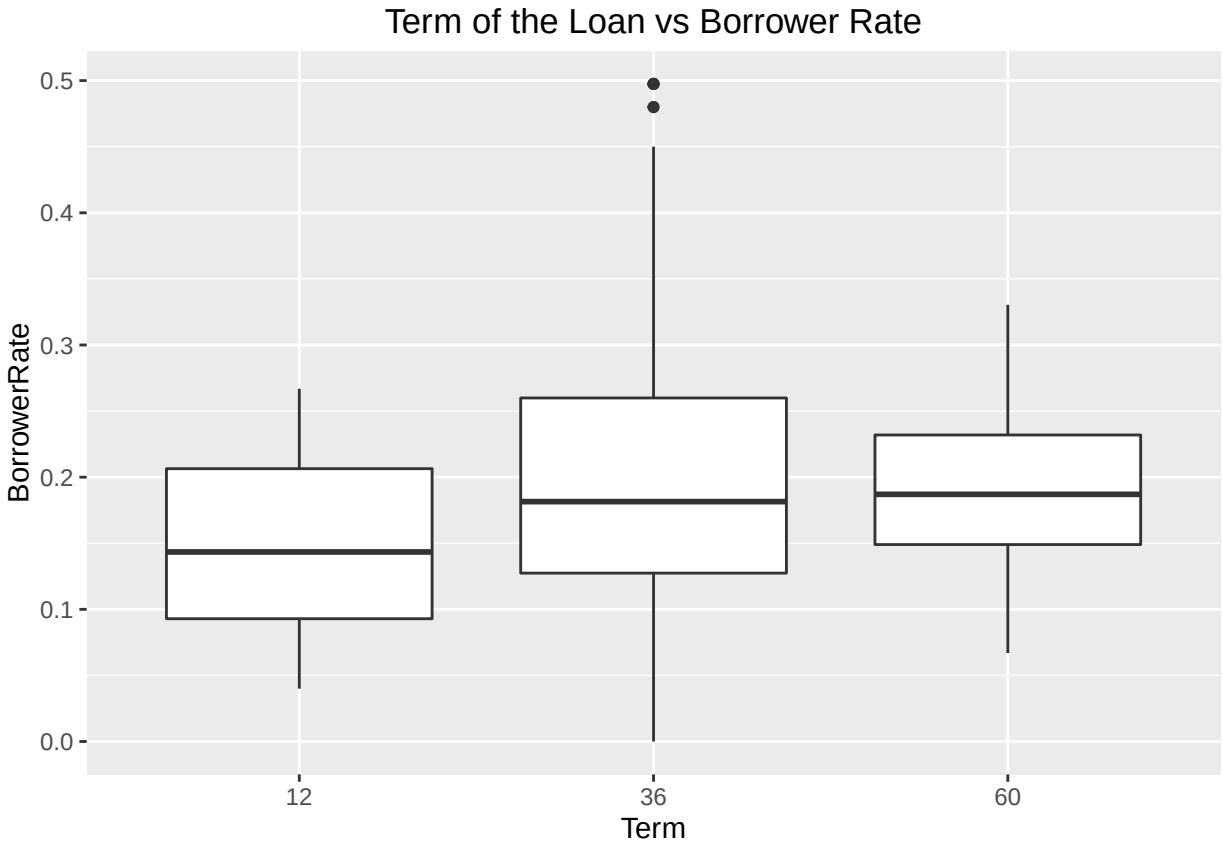
The median of Borrower Rate for defaulted and chargedoff loans appear to be a bit higher than the others.



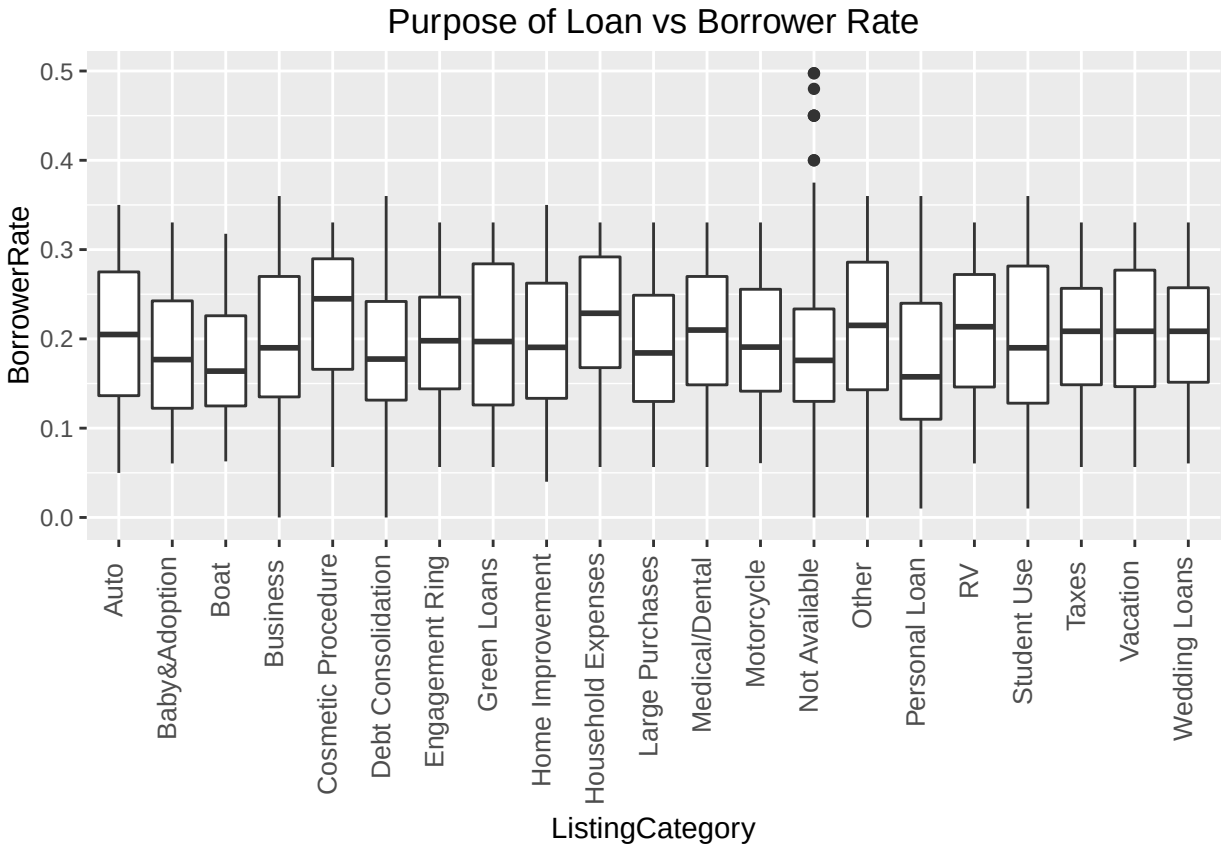
This is quite obvious the higher is the amount of loan, the period of loan must be higher except in some rare cases.



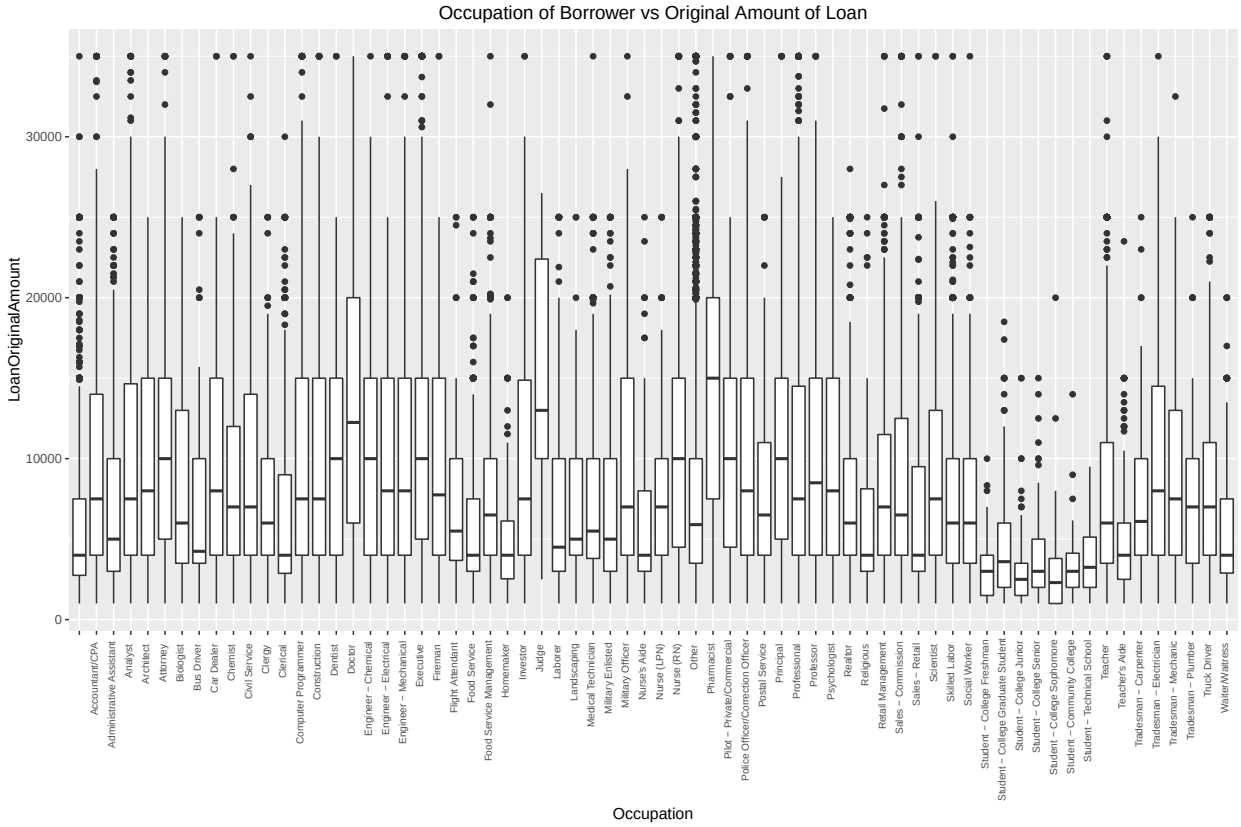
The Lender Yield is a bit higher for long term loans but the difference isn't much.



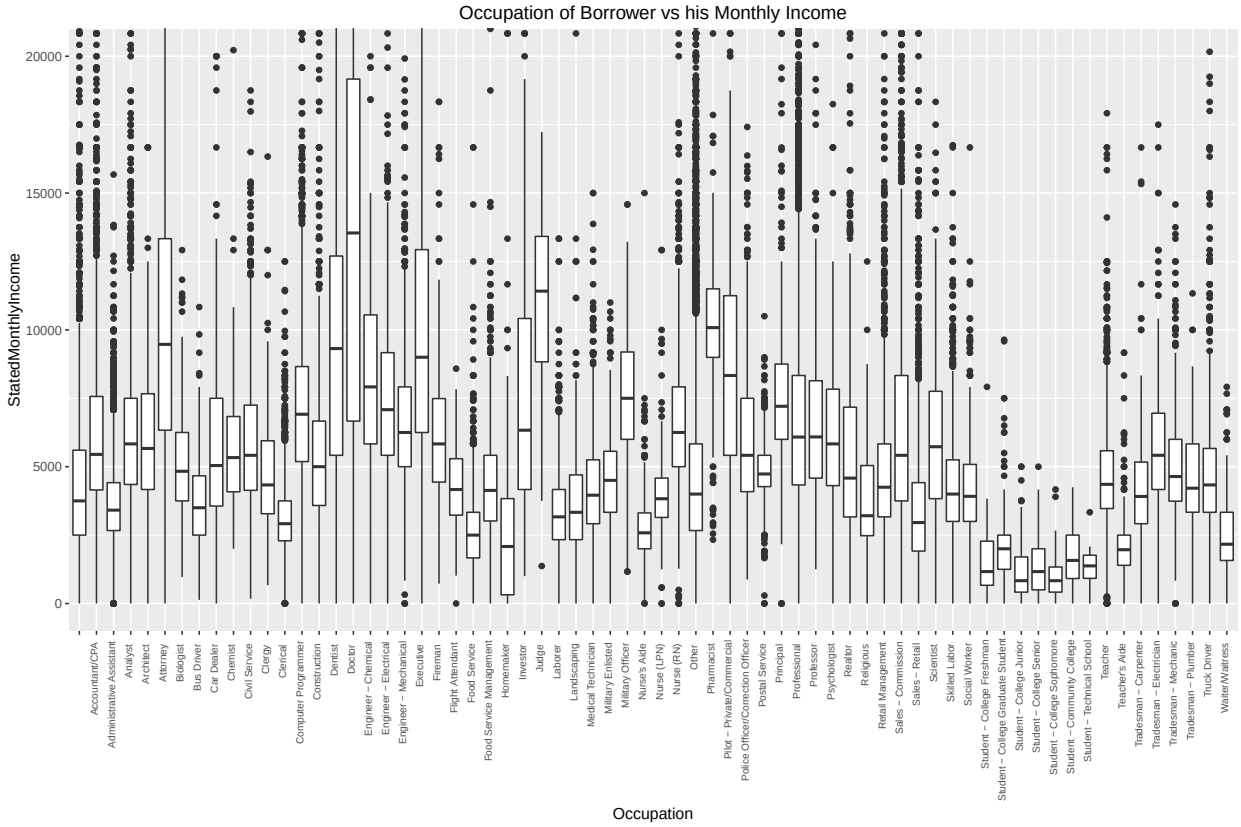
The borrower rate always shows same relation with a field as that shown by Lender Yield.



It is interesting to observe the rates for various types of loans. Some of them have a higher borrower rate than the others.



Quite Interesting. This shows the variations in the amount taken as loan by people from different working groups.



We can see the variation in monthly income for different working classes.

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?

I was interested to know how the loans are rated by Prosper. I found a clear relation between the Borrower Rate and the Rating. Lender Yeild, which is quite similar to borrowerRate also shows the same relation. Monthly Income of the borrower and the Loan amount may also have some effect on the rating but the trend observed wasn't enough to confirm the relationship.

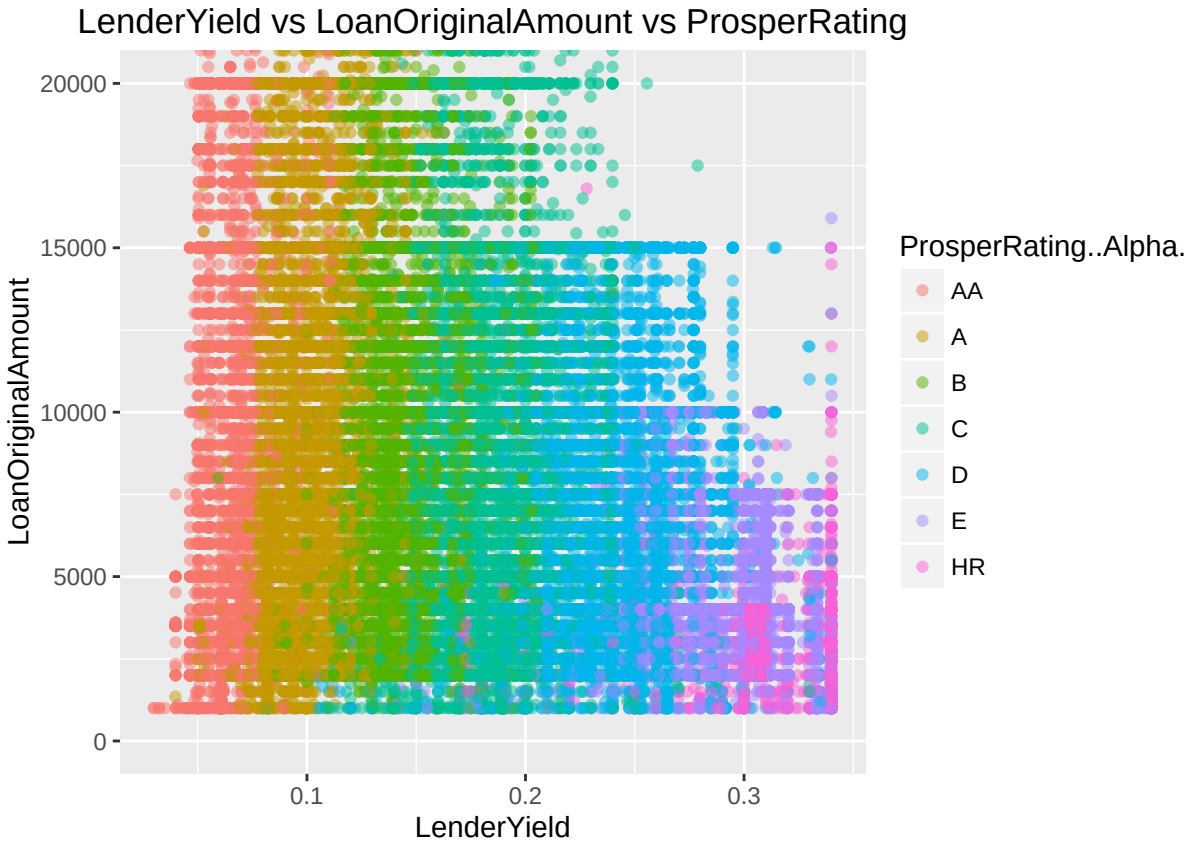
Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

The later plots show the incomes of various working classes, also the amount of loans that they take. I also observed that the loan amounts aren't continous. They are usually round figures thus we get discrete values such as 1k,2k and so on.

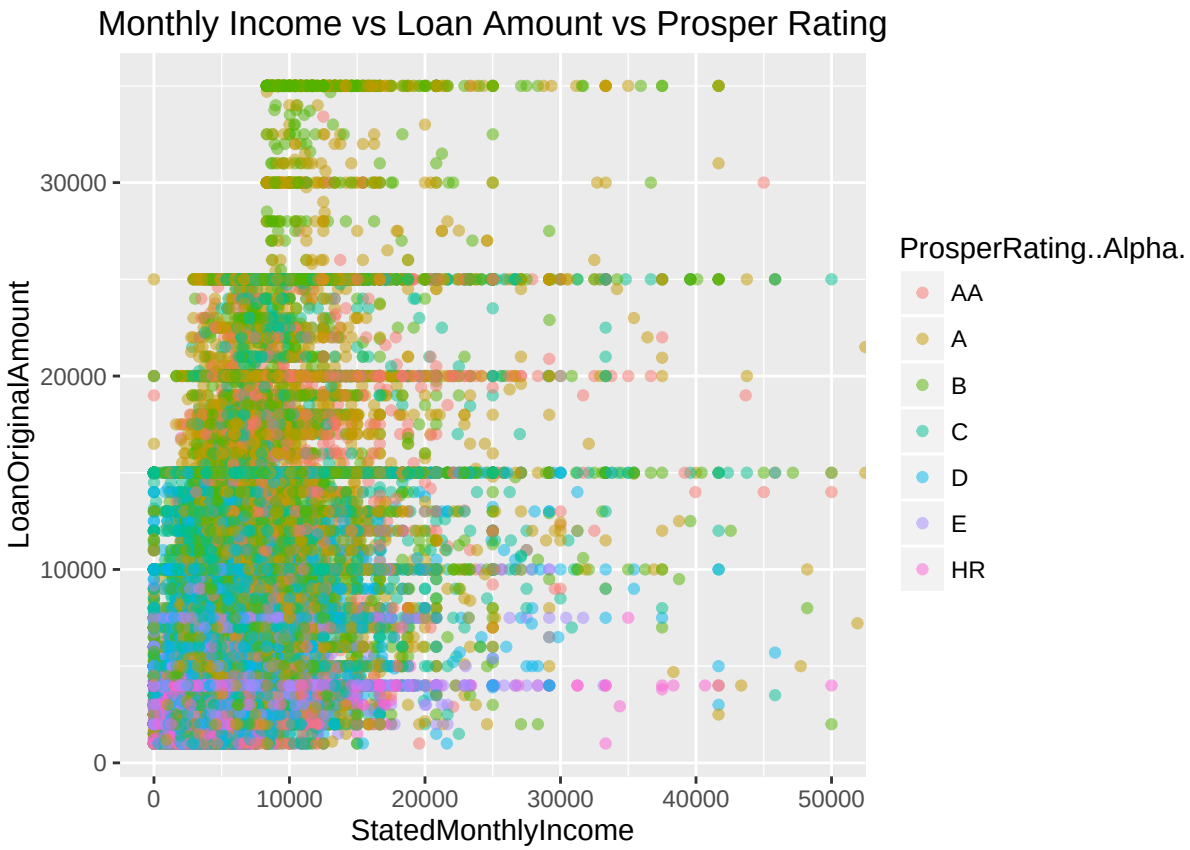
What was the strongest relationship you found?

The strongest relationship I found isn't mentioned in the above plots. Lender Yield and Borrower rate have the strongest relationship as they are quite similar which I could figure out after reading the variable's dictionary.

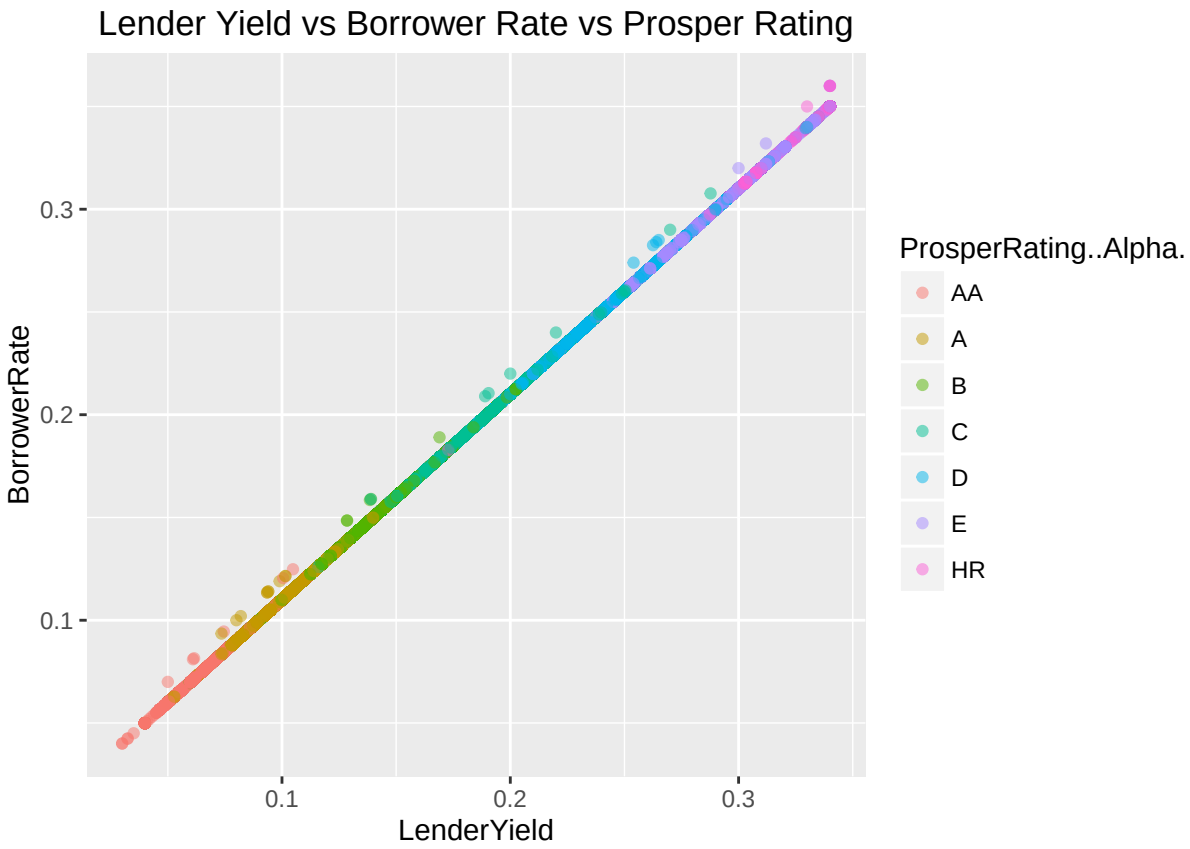
Multivariate Plots Section



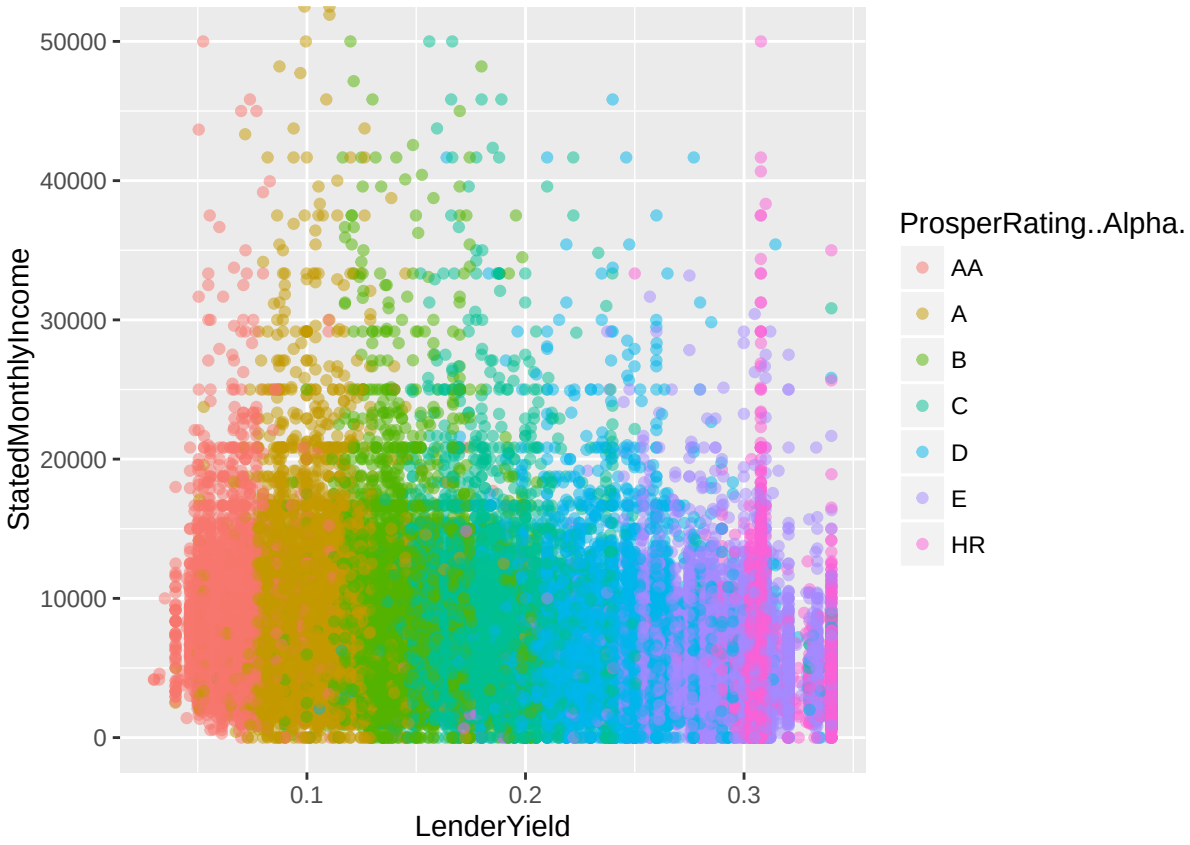
The trend observed here verifies that as the Lender Yield increases, rating gets poorer. We can also see that as the rating gets poorer, the median amount of the loan decreases



The trend observed isn't so clear but the loan amount and monthly income seem to have a relation with the prosper rating. The greater the lower the values of these two, poorer is the rating. But the opposite doesn't seem to hold good.



There's a clear relation between borrower rate and lender's yield which has been verified here (Their values are near equal or both lie on the same straight line). The rating gets poorer as these values increase.



Monthly income of the borrower doesn't seem to have much of a relation with the rating.

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?

I was interested in finding the reason behind ProsperRating. I was able to eshtablish clear relation between lender yield, borrower rate and loan amount (and Monthly Income to some extent) in predicting the rating of the Loan. The Lender Yield and Borrower Rate appear to highly related maybe only with some constant difference among the both.

Were there any interesting or surprising interactions between features?

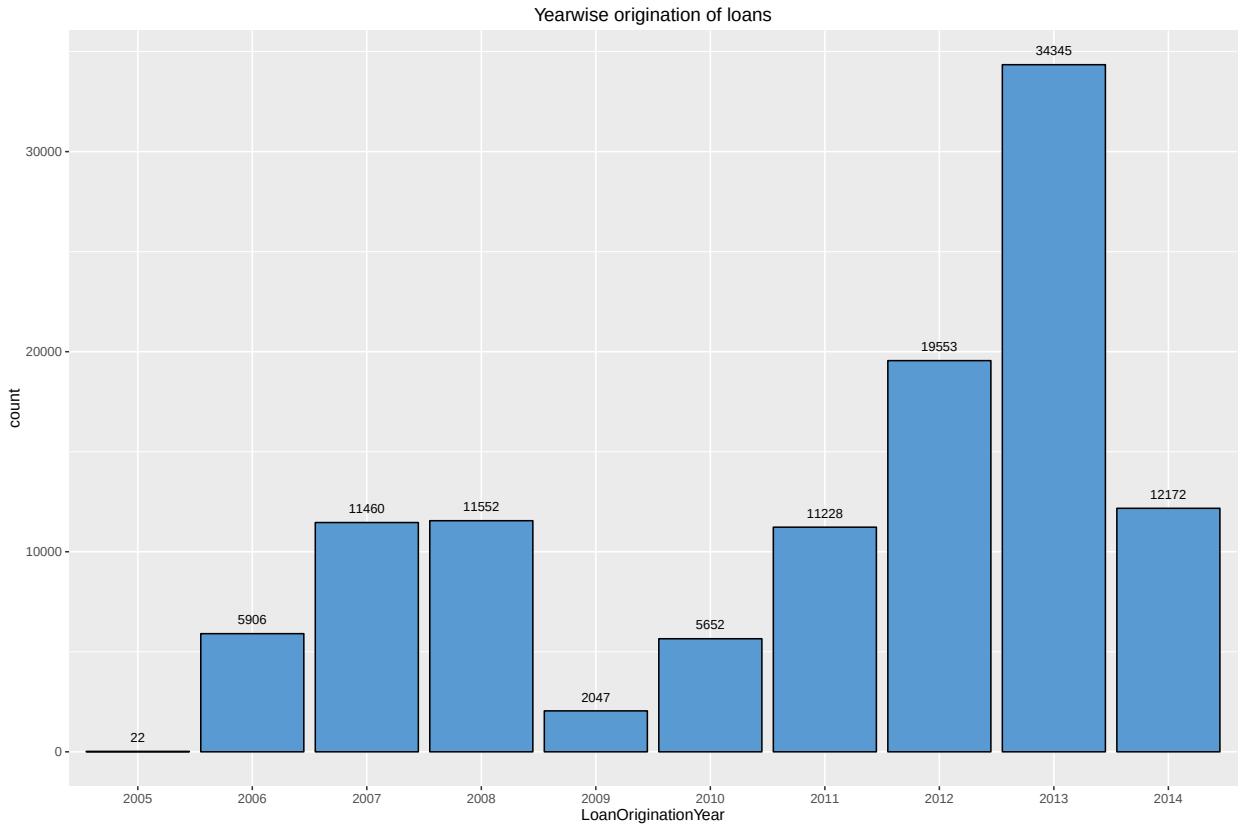
Borrower rate and Lender Yield have a linear relationship.

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

No, I didn't create any models. But a model to predict the Prosper Rating is obviously possible using the four variables mentioned above.

Final Plots and Summary

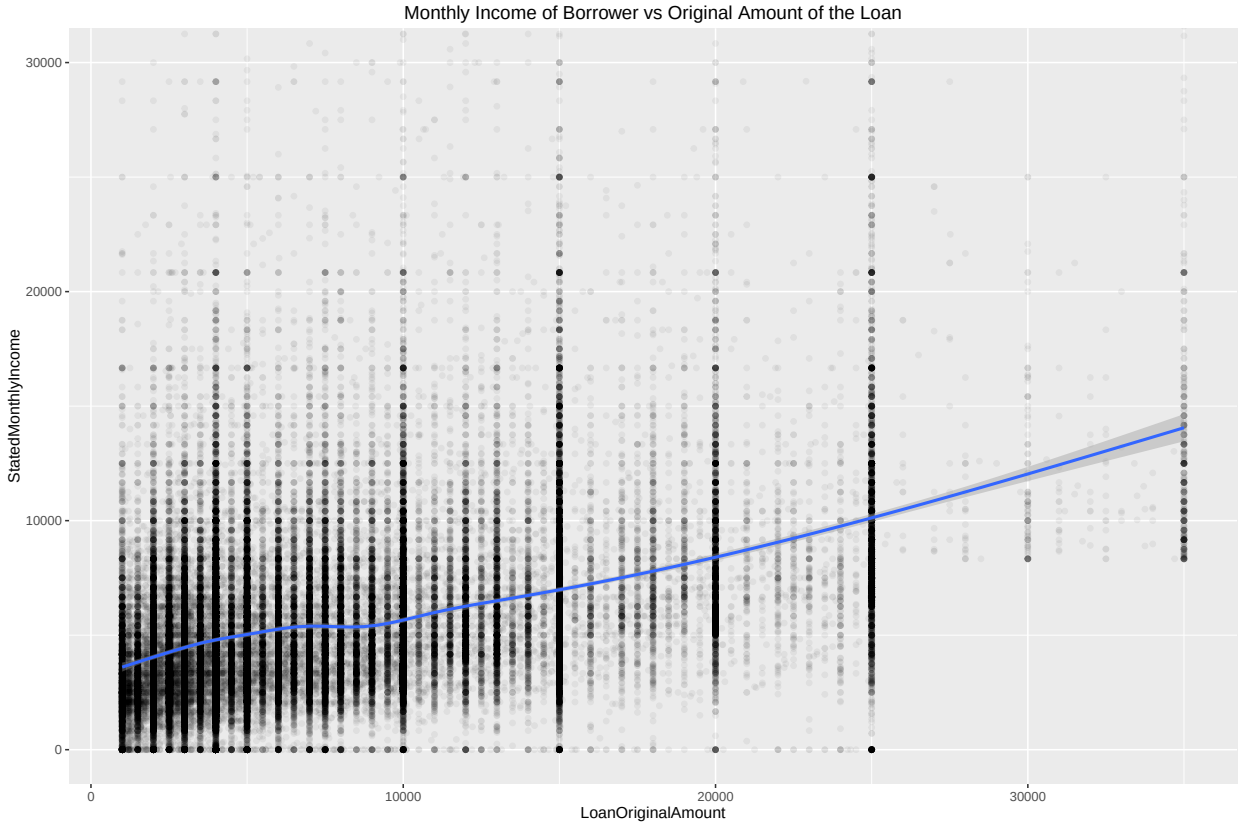
Plot One



Description One

We see a sudden decrease in 2009. Doing a quick Google search, we now understand the period of October 15, 2008 to July 13, 2009 is Prosper's QuietSEC Period, from which they are suspended for lending activities upon SEC approval. Prosper relaunched in July 2009. There's also a large number of loans sanctioned in 2013. Probably business kicked off well after 2009 with increase in Loans sanctioned every year. 2013 shows a peak but we have only 3 month's data for 2014 and the loans sanctioned in that year are quite high. Maybe 2014 ended up with much more loans being sanctioned than 2013.

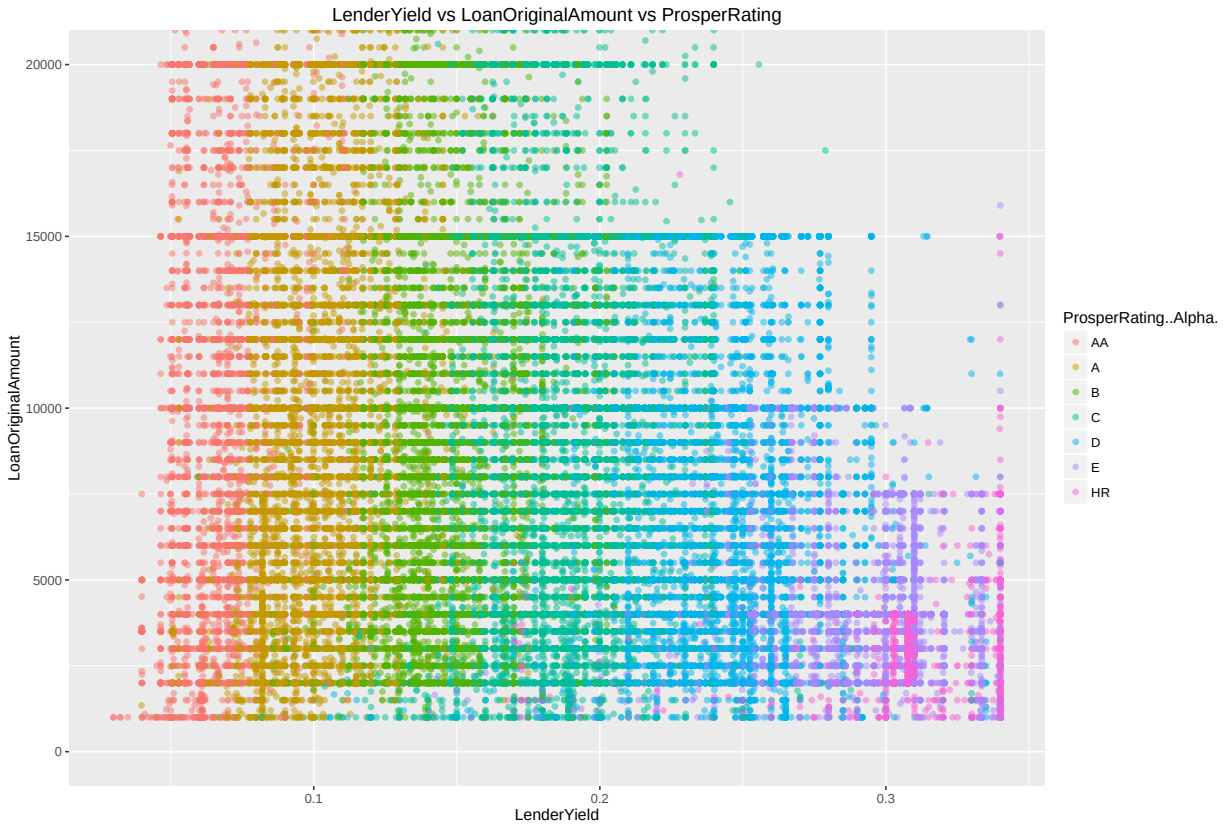
Plot Two



Description Two

It appears that as the Loan Amount increases, the median monthly income also increases (which for obvious reasons should increase). This shows that loans with higher amount are sanctioned to only to people with higher a income. Another interesting trend is that the Loan amount values appear to be discrete as the Loan amounts are usually round figures rather than random values.

Plot Three



Description Three

This is probably to most efficient plot towards the reason behind the Prosper Rating. The trend observed here verifies that as the Lender Yield increases, rating gets poorer. We can also see that as the rating gets poorer, the median amount of the loan decreases, so Loans having low amount and high lender Yield have a poor rating.

Reflection

The Data set has a lot of variables that have nothing to do with the trends and patterns. It took me some time to figure out the useful ones. Although there are more that 30 plots in this project, but there were many more plots that I plotted but chose not to mention in the final analysis as they didnot convey much information.

My point of interest was finding the features that are the reason for Prosper's rating of the loan. I feel, I could succesfully find three of them but there may be others that I have missed out for some reason. I was also able to see some general trends of the society like salaries of people with different occupations, the loan amount that most people take and many other features.

Finally, I tried to keep the visualizations as simple as I could, avoiding the use of unecessary colours wherever possible.