Loan Credit Investigation by Rivers, Xiao

Abstract

During last several hundreds of years, mutiple Industrial Revolutions happened, yet, financial industry has been stubbornly resilient to new technology inventions. But in recent years,P2P lending has become a quite hot topic. I???m curious to know whether P2P lending will democratize the financial industry or fall short of its promises. In this report, I???ll perform some data analysis to explore different variables in the Prosper dataset(a dataset from a P2P firm called Prosper) to get a better understanding of this business and try to predict the prospect of this industry.

Introduction

Since the beginning of the modern financial system, borrowing and lending hasn???t changed much. The loan is usually covered by collaterals and issued by a commercial bank. If the loan is not repaid in time, banks would use the backup collaterals to cover the exposure of the default risk. This traditional method, however, doesn???t fully utilize the potential of personal credit. Since credit worthy people won???t be able to borrow any funds if the borrower doesn???t have any asset as collateral for the loan.

With the advent of P2P lending, P2P firms claimed to offer lower rates to borrowers, and reduced risk to lenders. By examining the credit worthy of a potential borrower based on some algorithms and models, the P2P firm is able to approve a certain amount of loans to the borrower without any collateral, thus reduces the cost faced by the borrower. Besides, by applying advanced models and algorithms, the P2P firm is capable of filtering out high risk borrowers, which in turn reduced the risk faced by lenders.

Without further ado, let???s take a look on the data provided by Prosper- a leading lending company, and get a glimps about this industry.

First, I'll load all the libraries used in the analysis. ggplot2 is used for drawing the graph, lubridate is used to deal with the datetime type data, maps is used to get Geo data, ggthemes is used to decorate the graph, corrplot is used to draw correlation between variables, gridExtra is used to display the graph, reshape is used to reshape the data, and dplyr is used to manupulating the data.

Dateset

Before the analysis, I'll describe about the dataset I use. The dataset I'm using comes from Prosper - an online P2P firm. This data set contains 113,937 loans with 81 variables on each loan, including borrower profile, loan attributes, Prosper's risk system score assigned to the loan and etc.

The following operations give a brief summary of the dataset, and change the data type of a few columns for further investigation.

```
## 'data.frame':    113937 obs. of  81 variables:
##  $ ListingKey                       : Factor w/ 113066 levels "00003546482094282
EF90E5",..: 7180 7193 6647 6669 6686 6689 6699 6706 6687 6687 ...
##  $ ListingNumber                    : int  193129 1209647 81716 658116 909464 10
74836 750899 768193 1023355 1023355 ...
##  $ ListingCreationDate              : Factor w/ 113064 levels "2005-11-09 20:44:
28.847000000",..: 14184 111894 6429 64760 85967 100310 72556 74019 97834 97834 ...
##  $ 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","Chargedof
f",..: 3 4 3 4 4 4 4 4 4 4 ...
##  $ ClosedDate                       : Factor w/ 2803 levels "","2005-11-25 00:0
0:00",..: 1138 1 1263 1 1 1 1 1 1 1 ...
##  $ 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 6 16 16 ...
##  $ Occupation                       : Factor w/ 68 levels "","Accountant/CP
A",..: 37 43 37 52 21 43 50 29 24 24 ...
##  $ 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 "","00343376901312423
168731",..: 1 1 335 1 1 1 1 1 1 1 ...
##  $ DateCreditPulled                 : Factor w/ 112992 levels "2005-11-09 00:30:
04.487000000",..: 14347 111883 6446 64724 85857 100382 72500 73937 97888 97888 ...
##  $ CreditScoreRangeLower            : int  640 680 480 800 680 740 680 700 820 8
20 ...
##  $ CreditScoreRangeUpper            : int  659 699 499 819 699 759 699 719 839 8
39 ...
##  $ FirstRecordedCreditLine          : Factor w/ 11586 levels "","1947-08-24 00:0
0:00",..: 8639 6617 8927 2247 9498 497 8265 7685 5543 5543 ...
##  $ 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 ...
```

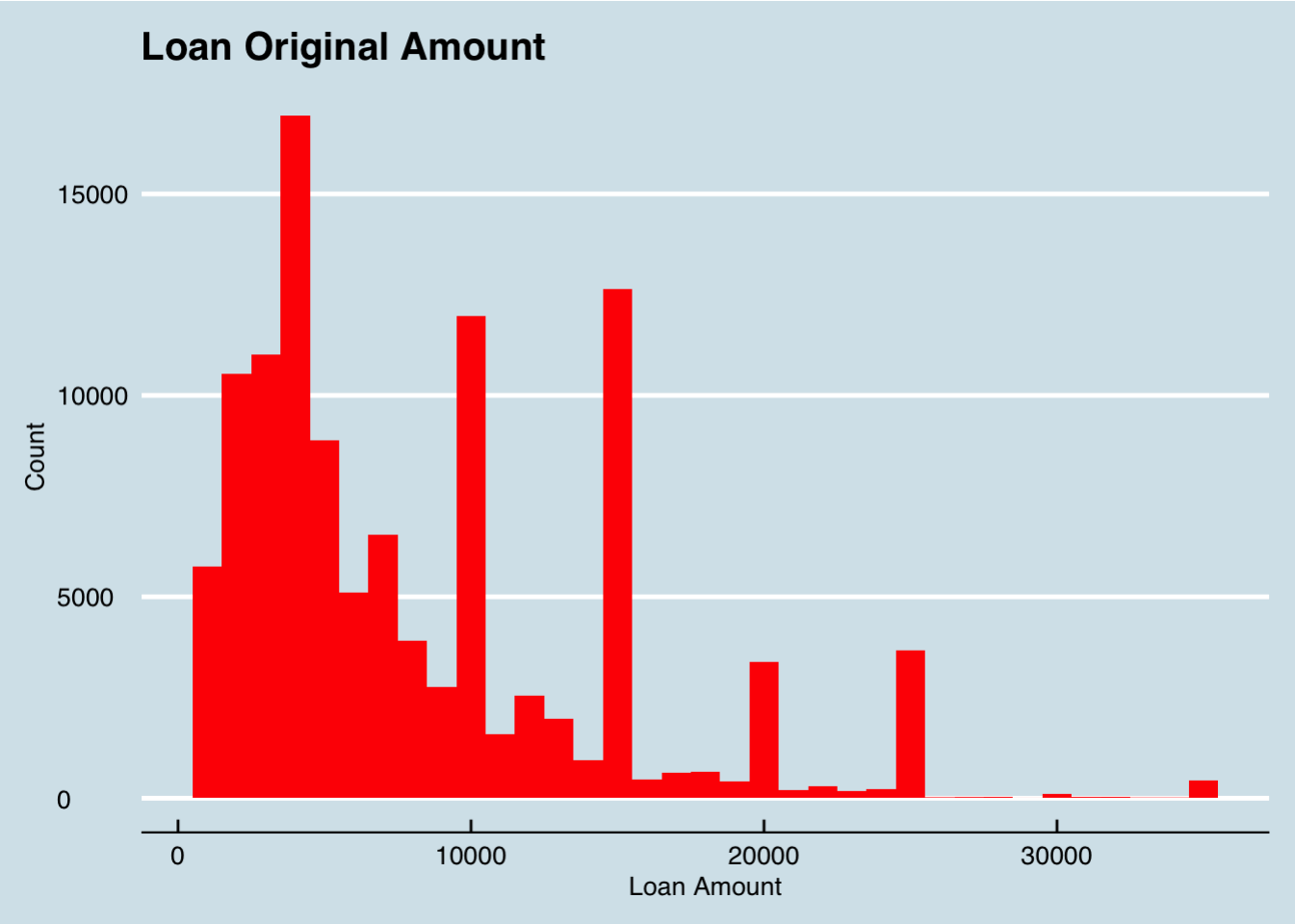
```
##  $ PublicRecordsLast10Years           : int  0 1 0 0 0 0 0 1 0 0 ...
##  $ PublicRecordsLast12Months          : int  0 0 NA 0 0 0 0 0 0 0 ...
##  $ RevolvingCreditBalance             : num  0 3989 NA 1444 6193 ...
##  $ BankcardUtilization                : num  0 0.21 NA 0.04 0.81 0.39 0.72 0.13 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 "00003683605746079
487FF7",..: 100337 69837 46303 70776 71387 86505 91250 5425 908 908 ...
##  $ 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 121268 ...
##  $ LoanOriginalAmount                 : int  9425 10000 3001 10000 15000 15000 300
0 10000 10000 10000 ...
##  $ LoanOriginationDate                : Factor w/ 1873 levels "2005-11-15 00:00:0
0",..: 426 1866 260 1535 1757 1821 1649 1666 1813 1813 ...
##  $ LoanOriginationQuarter             : Factor w/ 33 levels "Q1 2006","Q1 200
7",..: 18 8 2 32 24 33 16 16 33 33 ...
##  $ MemberKey                          : Factor w/ 90831 levels "00003397697413387C
AF966",..: 11071 10302 33781 54939 19465 48037 60448 40951 26129 26129 ...
##  $ 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 ...
```

Analysis for the single varible.

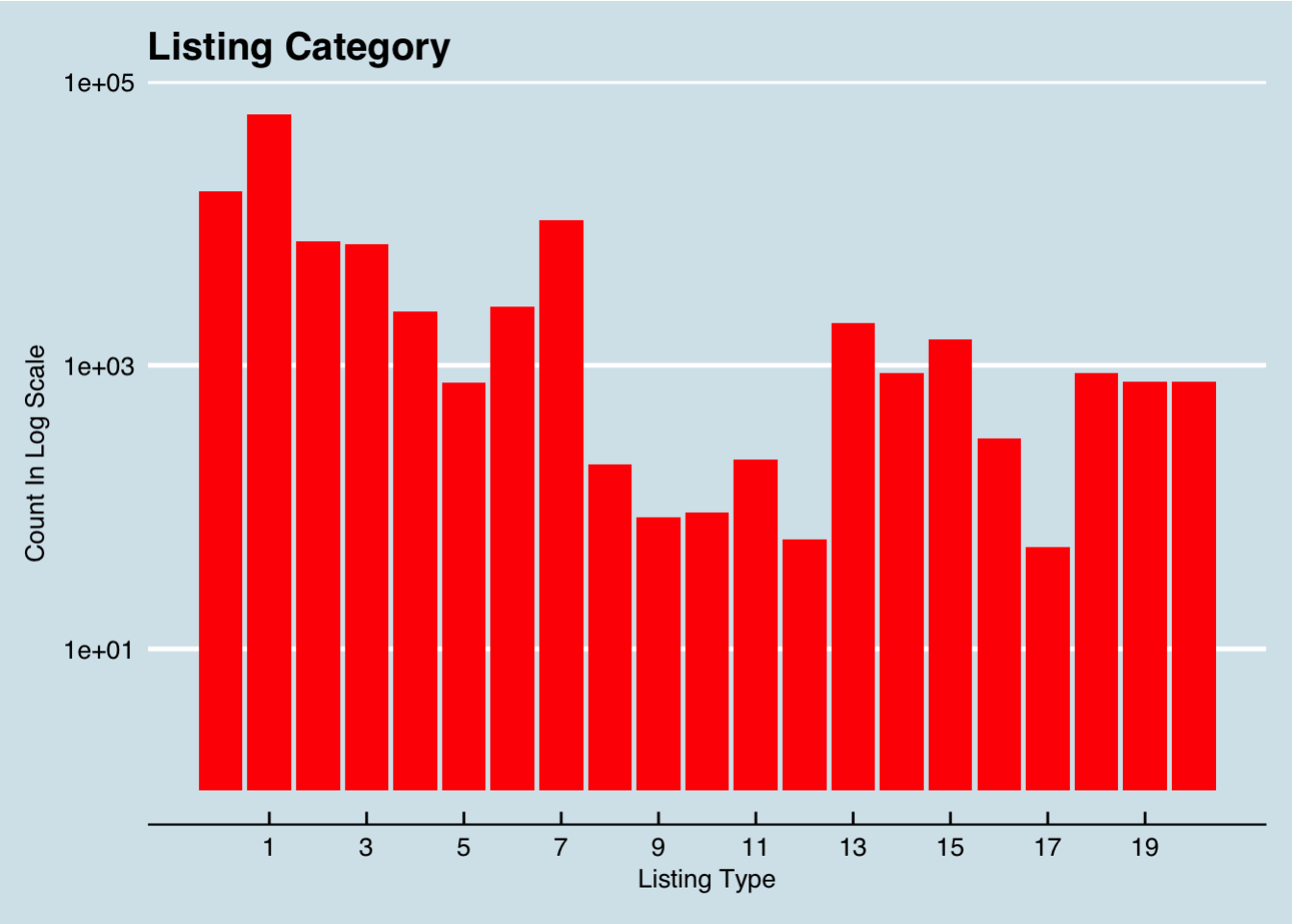
As we can see from above summary, there are too many variables to look at, so I will first look at variables which interest me most.

The first variable catched my eye is the listing category, since I'm curious about where the borrowers are going to use the funds.

graph 1-1



graph 1-2



There are 20 categories in total, And I listed all of them below for referrence.

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

As we can see from above graph, most listing are used for type 1(Debt Consolidation) reason. I don't feel surprised at seeing this result. I think there are probably two main reasons for this high ratio.

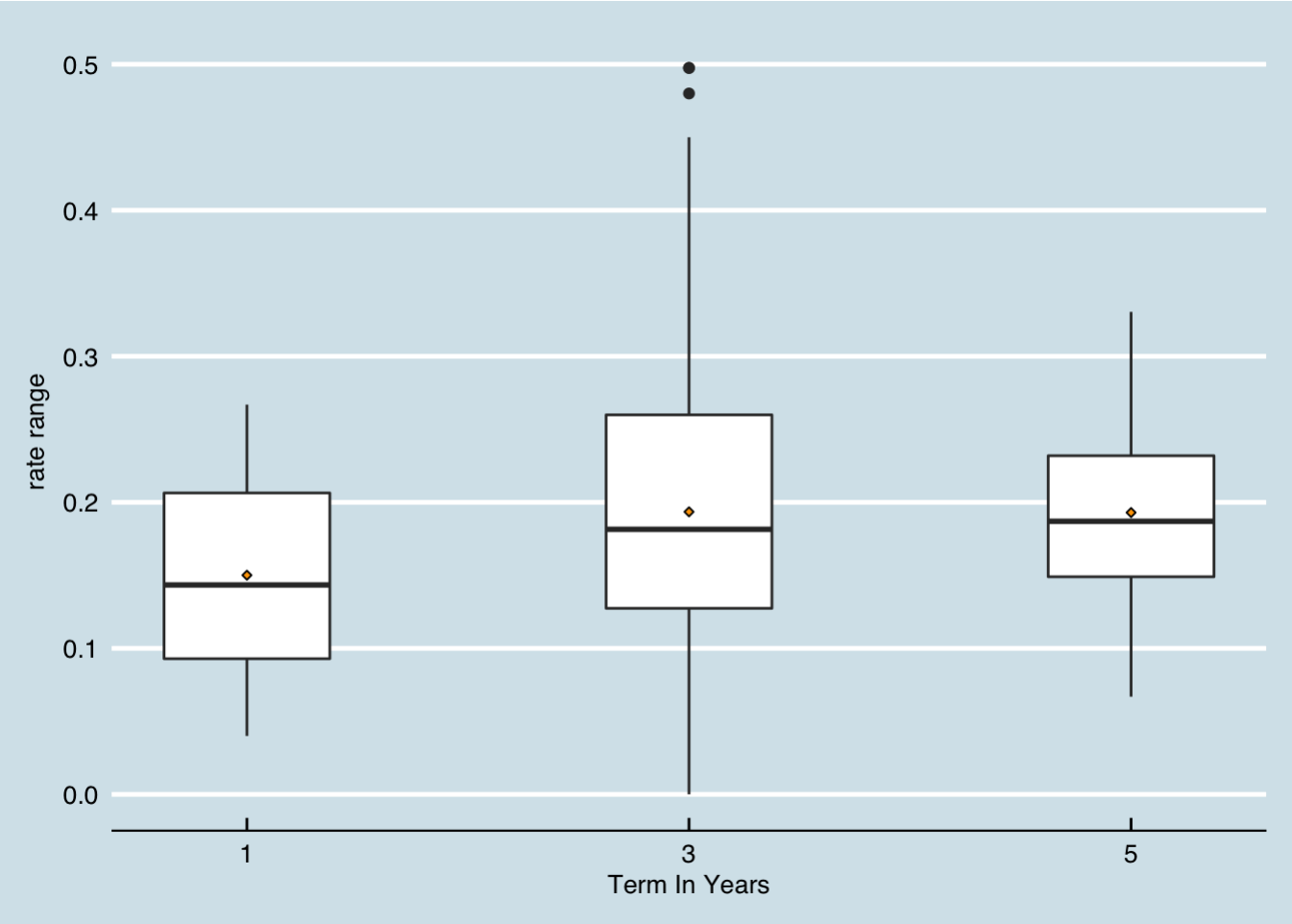
The first reason is as follow, a lot of borrowers are too lazy to check all available options to validate their loan type, so a lot of people would simply choose which ever comes to them first, namely the first option.

Now,Let's assume all borrowers will check each option carefully and faithfully choose their real reason, then type 1 will still occupy a large percentage. why? Because a lot of people have strong incentive to use the funds to repay their credit card. Often times, the loan rate offered by P2P firms is lower than that offered by credit union. Take China for intance, the credit card repayment EAR(effective annual rate) is around 18% if user is not able to repay before payable day. Hence ,as long as P2P rate is lower than 18%, borrowers would use P2P loans to consolidate their loan. To confirm my assumption, I'll take a look at the interest rate in the next plot.

Other options are significantly dwarfed by type 1 option. I was think that since US households typically don't save much, "Large Purchases" could also occupy a large percentage, but it turns out little people choose that type. I guess one reason could be that they are using specific reason like auto, house expense, or home improvement instead.

Now let's look at the borrower rate as it determines the attractiveness of the P2P lending.

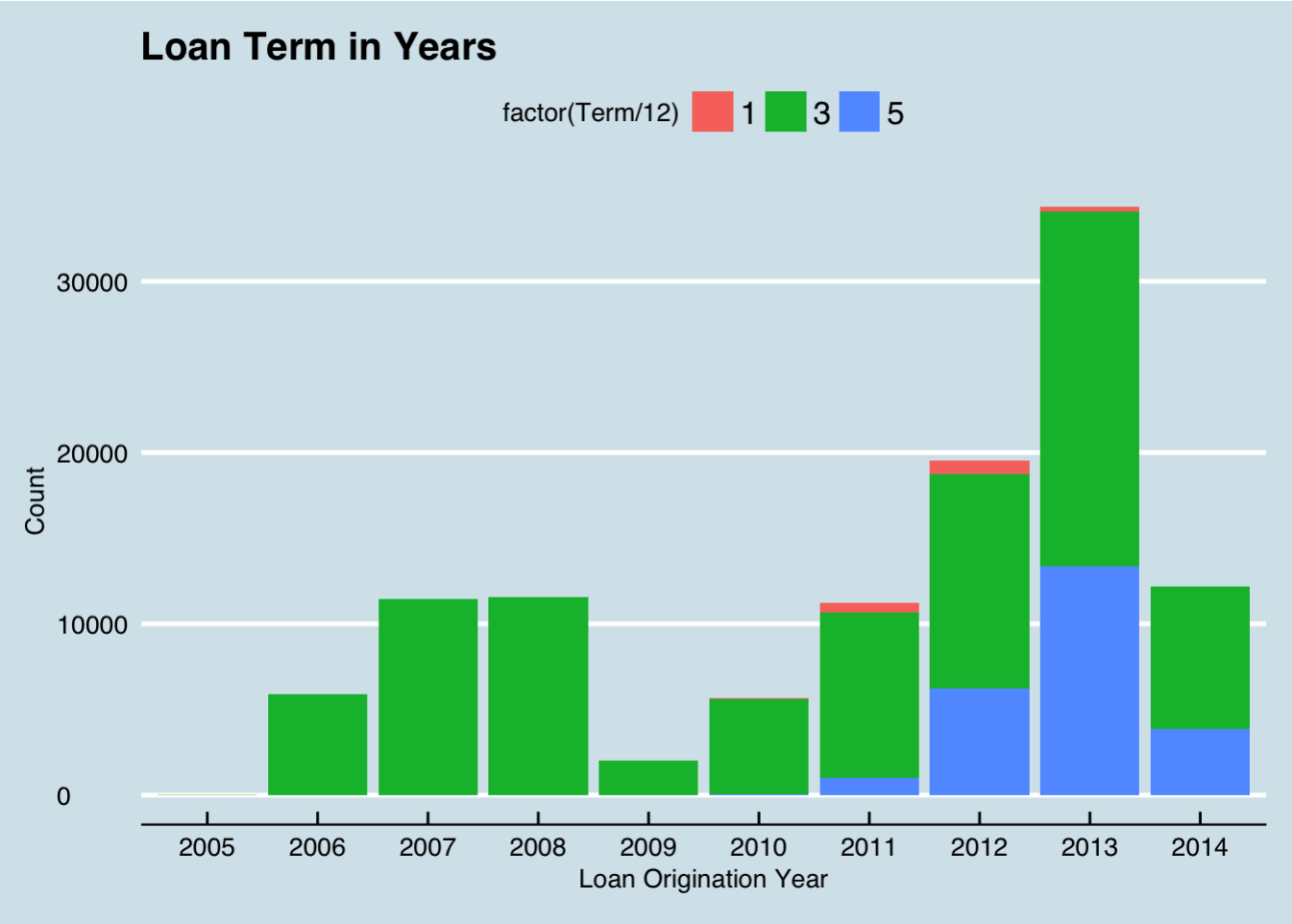
graph 2



From above boxplot,we can clear see that borrowing rate range varies a lot. The majority loans have a upper bound below 25% and a lower bound above 10%, and as Term increases, uncertainty also increases, thus a higher yield is required to compensate the risk involved. As we can see, the mean value among these three terms are between 15%-20%, which is lower than that of the credit union. This will definitely serve as a huge incentive for people to borrow from P2P firm.

The next variable that interests me is the term of the loan. Because the longer the term the higher the exposure. To properly handle this type of risk, Prosper would have to device its loan term very carefully to avoid long term deliquency issue.

graph 3

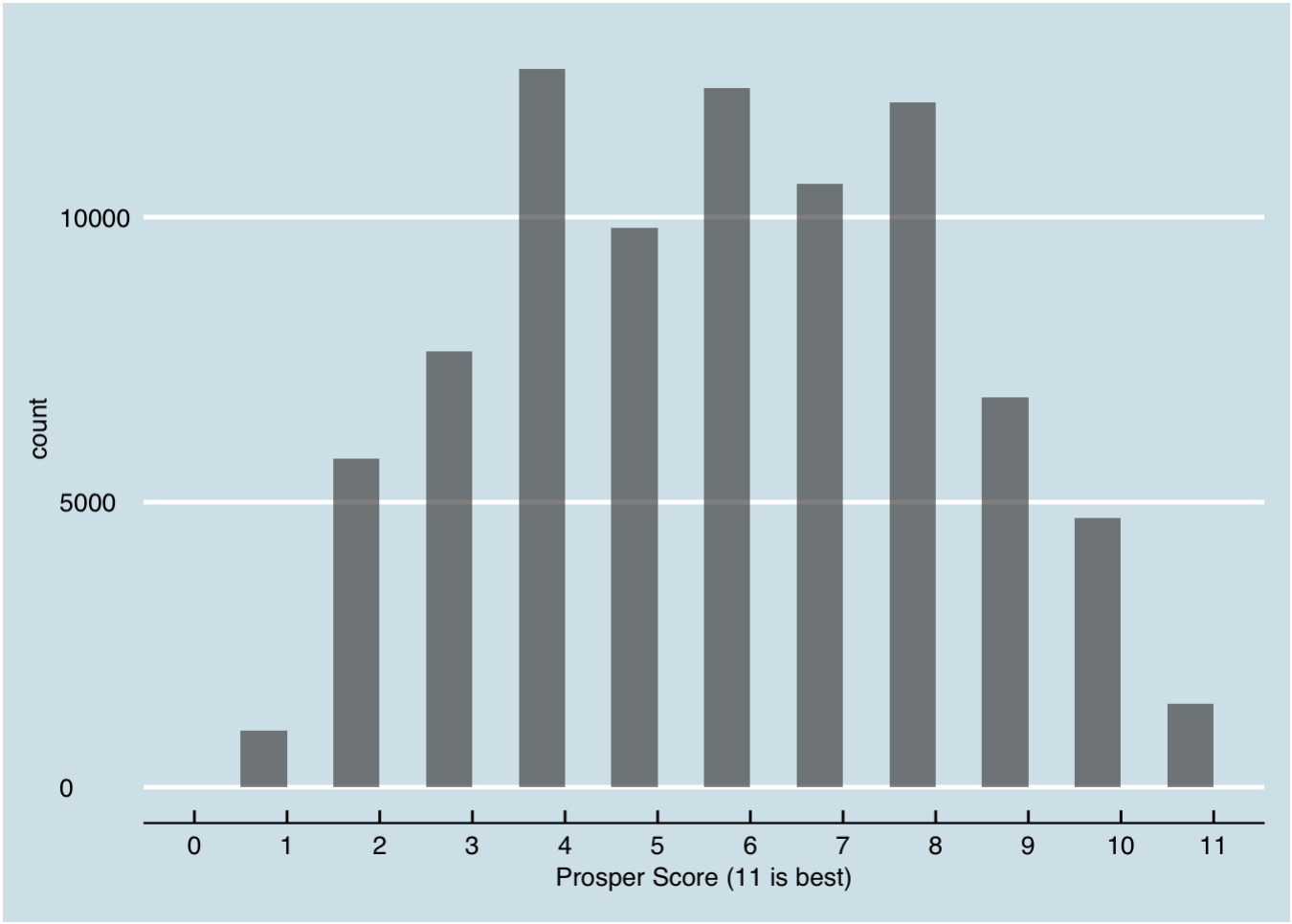


From the above table and chart, we can observe a clear trend-As loan Term increases, loan amount is also tend to increase.

There are three loan terms in Prosper-the minimum period of the loan is 1 year and maximum period is 5 year. The median is 3 years. The majority loans are listed with three year period, and few people take loan of one year and five year peroid. It is easy to handle the deliquency issue for simple term structure loans. The prosper company starts with 3 year loan and later, when the company gain more experience, they added 1 year loan and 5 year loan.

The next variable to look into is the rating of borrower given by Prosper

graph 4

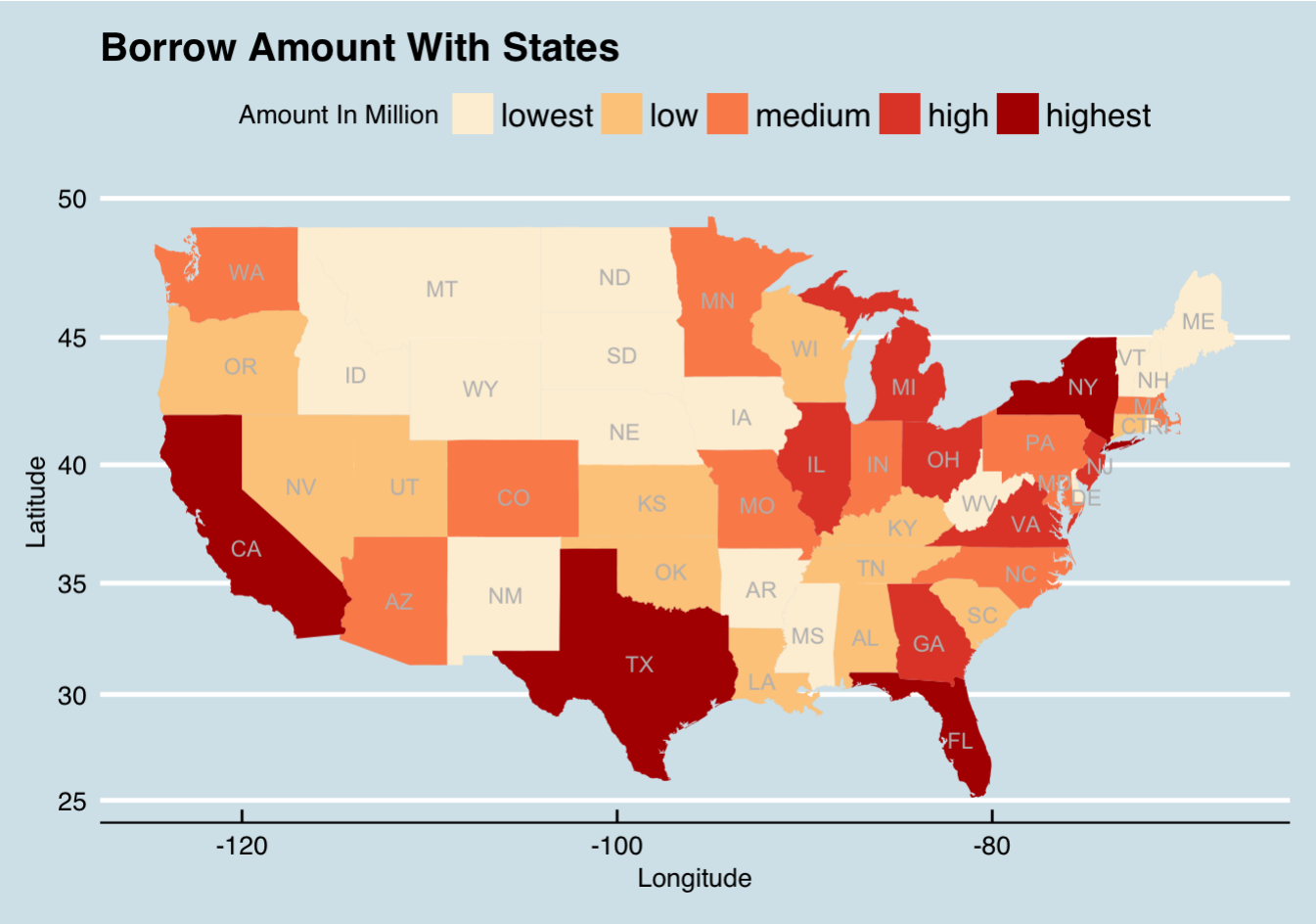


Prosper rates it borrowers from 1 to 11,with 11 being the best.

As the above chart shows, the ratings of the borrowers falls into a bell shape curve, with large portion of borrwers in the middle range and a few borrower falls in the range of good rating or bad rating. Rating is a key indicator of borrower's credit worthiness. Normally, borrowers with low credit worthiness-namely higher default rate would need to offer a higher rate to attract potential investors, I'll look into the relationship between these two in the bivariate analysis section.

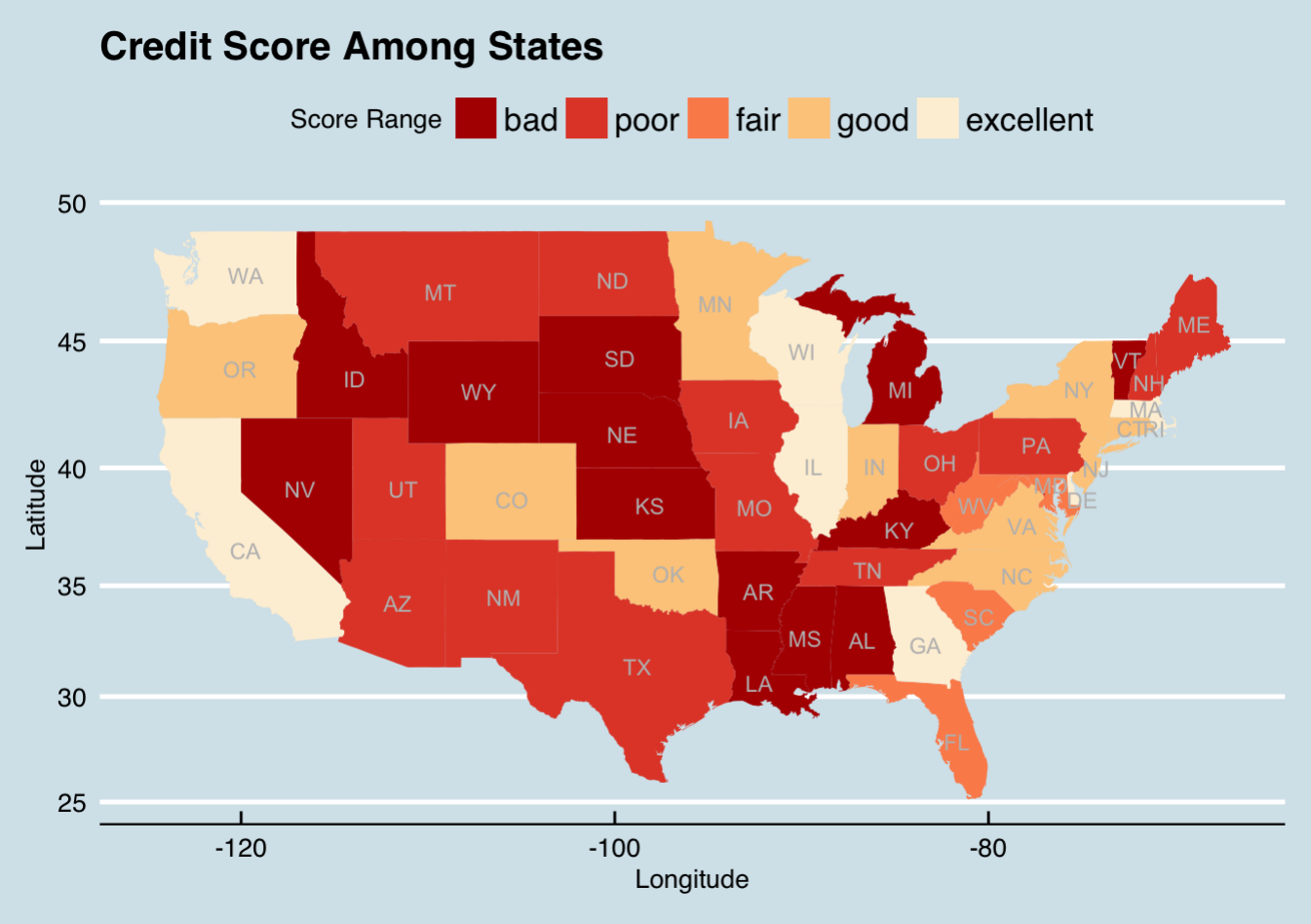
Last, I'm going to look at which states borrows most from P2P firm.

graph 5



As we can see, regions with higher population tend to have higher borrowing. for example, the four states with highest population - CA, NY, TX, FL are the ones borrow the most.

which state has the poorest credit score



As we can see from the above picture, some states have relatively low average credit score.

Univariate Analysis Summary

In the above plots, I analysed 5 parts of the data.

1. Listing Categories
2. Rate given by the borrowers
3. Term Structure of the loans
4. Rating given by Prosper
5. loan based on State

From the above analysis, here are some of the findings worth noting.

Firstly, P2P lending lower the borrowing rate thus helps borrowers save the cost. As a report shows , average US credit card rate is around 18%, while as our stats shows ,the borrowing rate of P2P is around 15%, this three percent gap will surely make P2P firms more appealling to customer than traditional funding method.

Secondly, P2P firms are at their early stage. Term structure lacks variation. Most loans are three years, though there are signs that five years and one year loan are become more frequent. Yet, compared with traditonal funding method, it still has long way to catch up.

Last but not the least, normally, borrower's credit rating should fall into a bell curve shape as it shows in the graph.It appears Prosper has a good scoring model, however, I think there are potential problems with the rating.

As we all know, there are lots of frausters out there trying to take advantage of Prosper lending business. With these fraudulent loans adding to the data, score system should give a right tailed curve rather than a normal distributed one.

To further investigate the data, I'll take a look at the relationship between variables.

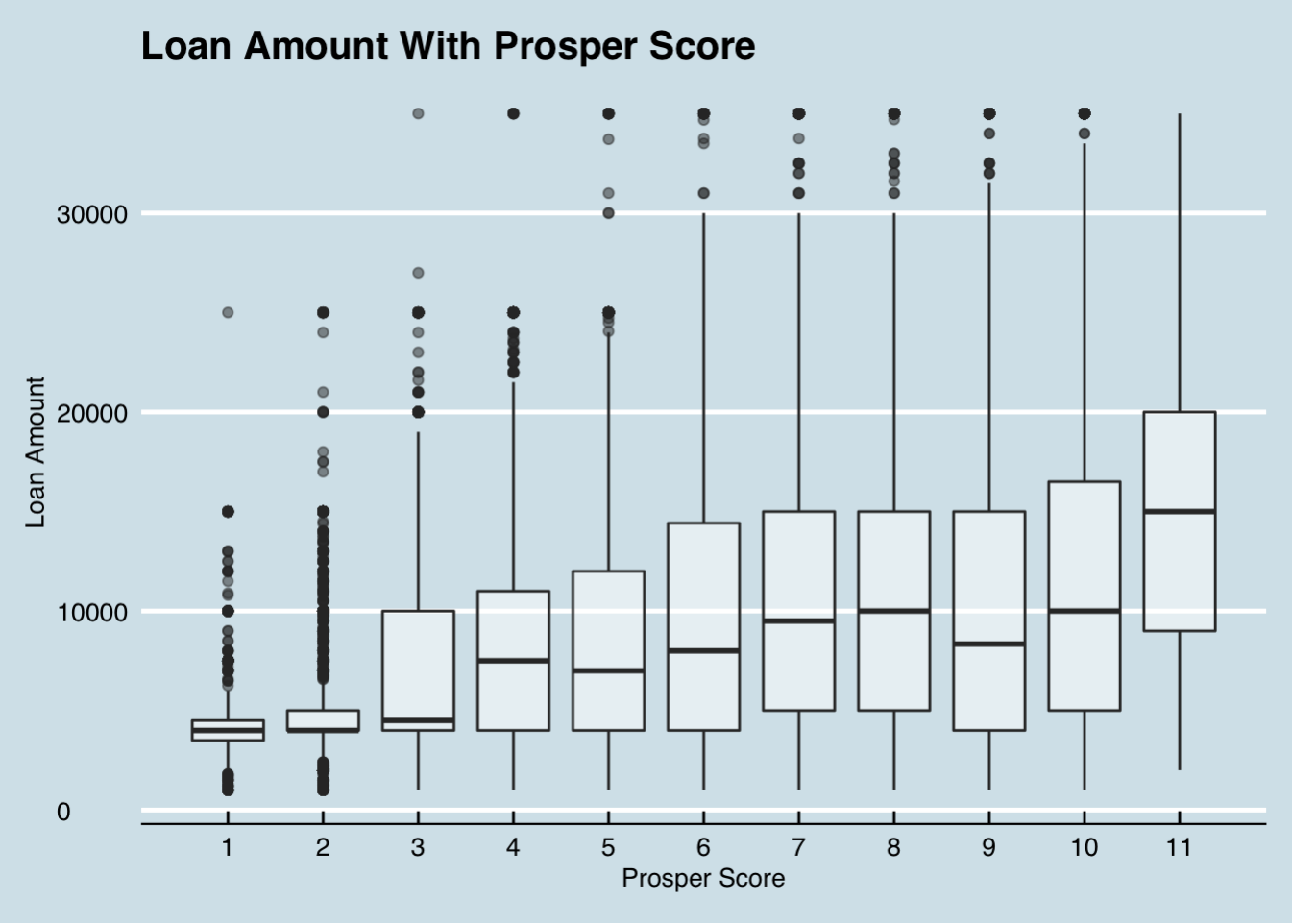
Bivariate Plots

From above analysis, we have a general idea of where these loans are spent, loans types and credit worthy of the investors. Lenders will definitely look into these indicators to measure the loan quality, yet, there are a lot more factors affecting people's choice on which loan to invest. There is a popular theory in finance and economics call risk preference theory. According to the theory, most investors are risk averse, which states that when faced with two investments with a similar expected return (but different risks),investors will prefer the one with the lower risk.

Here, I would like to use the emperical data to examine this theory. My logic goes like this 1. If investors are risk averse, then they will prefer borrowers with higher credit. 2. If investors are risk averse, With similar expected return, people will choose low defauld borrowers. 3. Since there are limited amount of high credit worthy borrowers, they would have upper hand in negotiagting the rate with lenders, thus, the higher score the borrowers ,the lower the borrow rate will be.

To test the assumption, Let's first take a look at how credit score affecting investors' choice.

graph 6



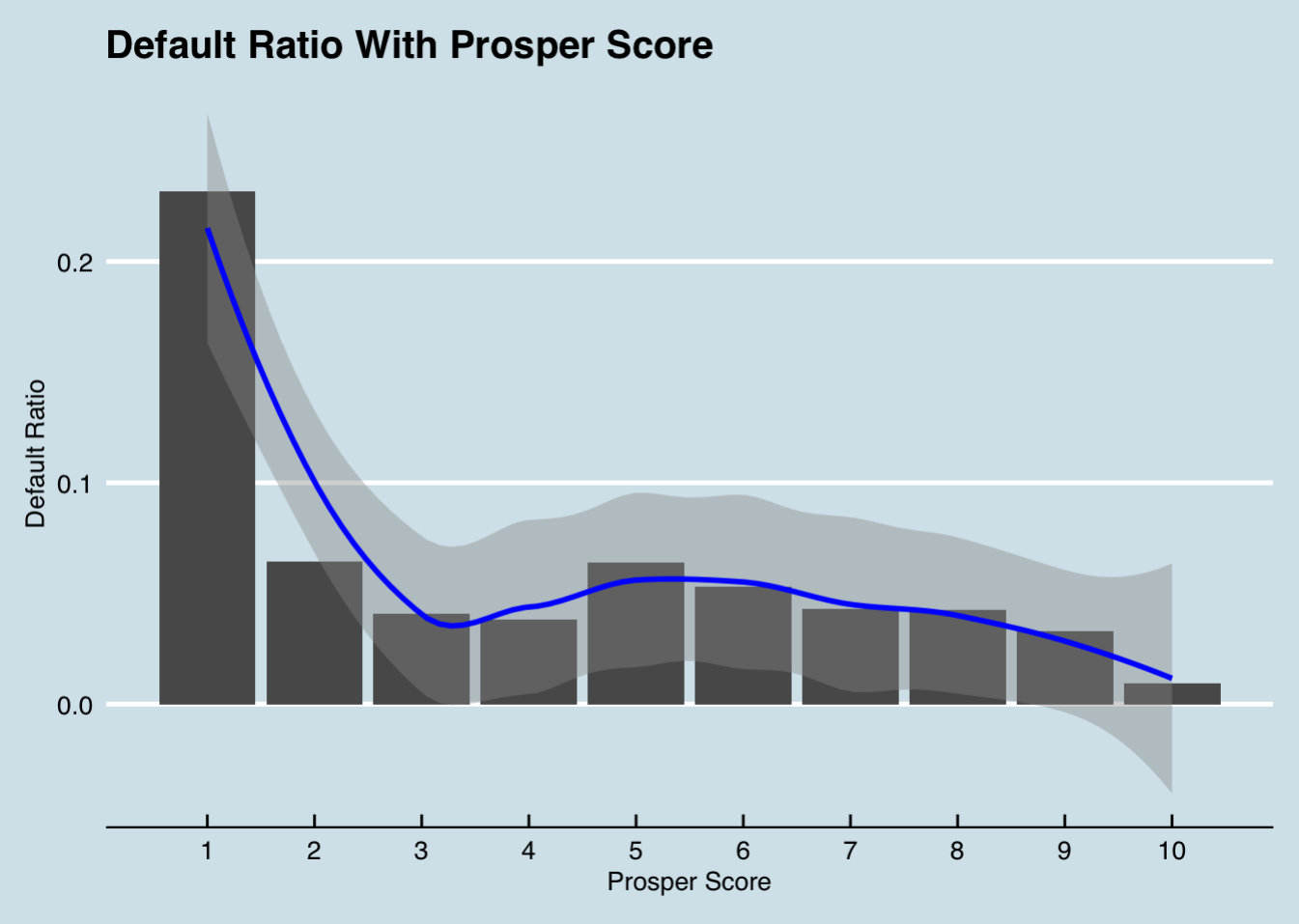
The x axies represent the score given by Prosper, the higher the score the better the borrower. The y axies is the loan amount given to each loan. As we can see from above graph, borrowers who have higher prosper scores are able to borrow more loans compared with those who have lower prosper scores. I previously briefly discussed the possible bias of the Prosper score system in graph 4, and it may not reflect the real credit worthiness of a customer.In the long run, if a P2P firm wants to develop sustainably and make some profit, everything boils done to its risk control ability, Prosper score definitely shows how well its risk control ability is.

Since we have several years' data of the borrowers. I'll draw a graph to meature the correlation of the default rate and Prosper Score.

In the dataset, there are two terms associated with default risk. They are “Chargedoff” and “Defaulted”. The definitions of these two terms differs from our common sense. According to defination on Prosper website, when a loan is 121 days past due, it is considered charge off. When a loan is in one of the following situations ,it is considered default. Delinquency Bankruptcy Deceased Repurchased Paid in full Settled in full

Here, I’ll combine Charge off loans and Default loans as our indicator to represent our loan default rate. the following graph shows the relationship of default rate and prosper score. If the Prosper system is good, these two terms should have a very high correlation.

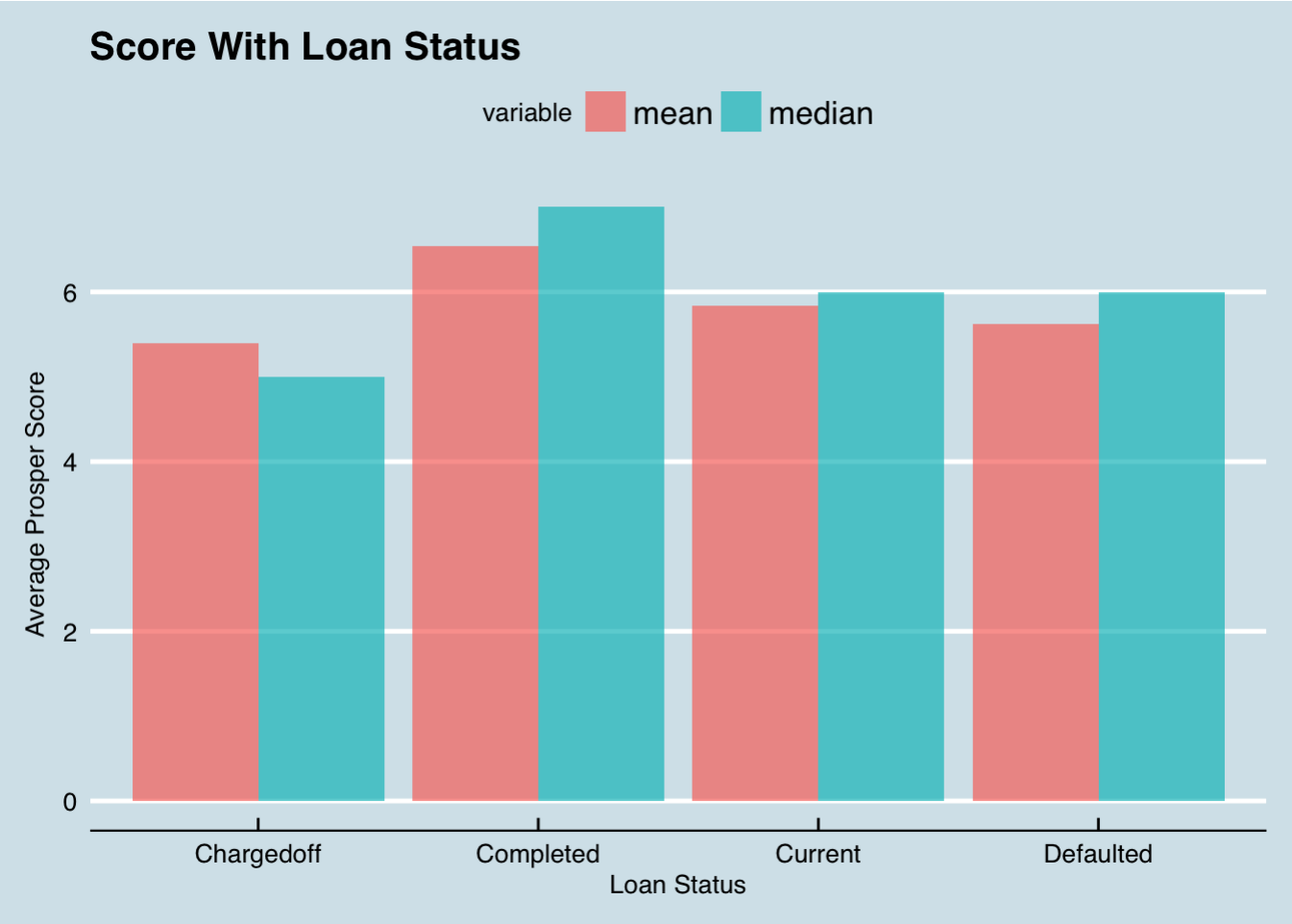
graph 7



From above graph,we can clearly see a strong correlation between these two variables- the higher the Prosper Score, the lower the default rate, but there exist some adversaries- borrowers with prosper score of 4 and 3 has a lower default rate than that of borrowers with prosper score of 5 and 6. Combining result from graph 4 and 7, I believe there are exist some problems in the prosper rating system. Some low credit borrower(fraudster) took the advantage of these flaws, got themselves a higher score. That’s why we see a higher default rate in credit score of 5 and 6 compared with that of 3 and 4.

The following graph took a further look at the score system

graph 8



As we can see from above graph ,even though the score system won't be able to give a precise measurement of the credit worthiness of the borrowers, it does give a good fair measurement of the borrowers in general. On average, borrowers who completed the loan repayment has one point higher than those who didn't.

Now, let's get back to our 'risk aversion theory' test. I'll measure the relationship between expected return on the loan and prosper score.

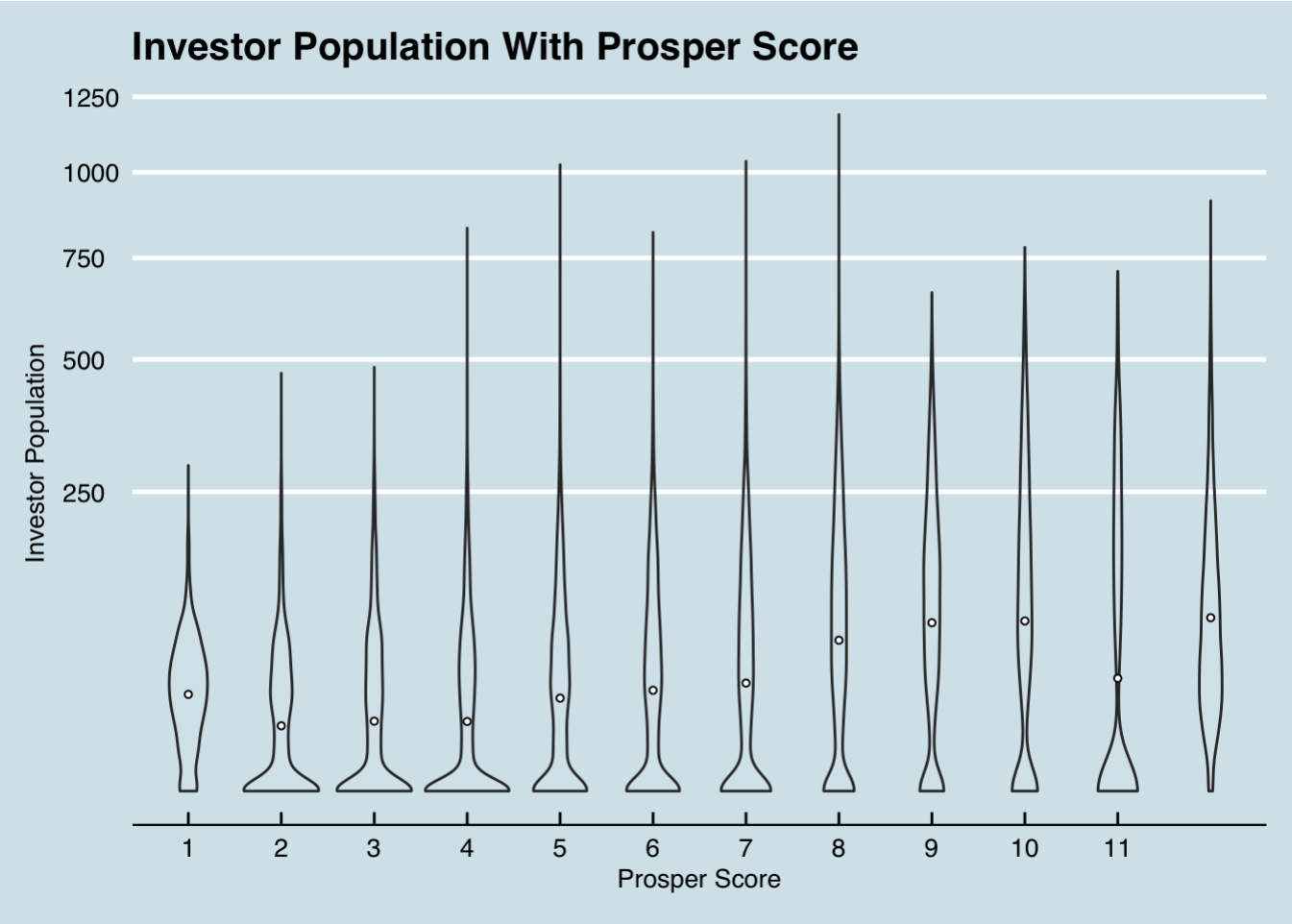
graph 9

As we can see from the above graph, most loans with score below 8 offer similar expected return, yet they bear different risks. For instance, for borrowers with score 1, you would probably lose 20% of your principle, while for borrowers with a higher score , let's say score 7, the maximum amount you would lose is around 2% of your principle. This observation fits our theory exactly. Further more, since the number of prime loans are limited, investors would like to sacrifice return for a lower risk loan.

Investor's preference

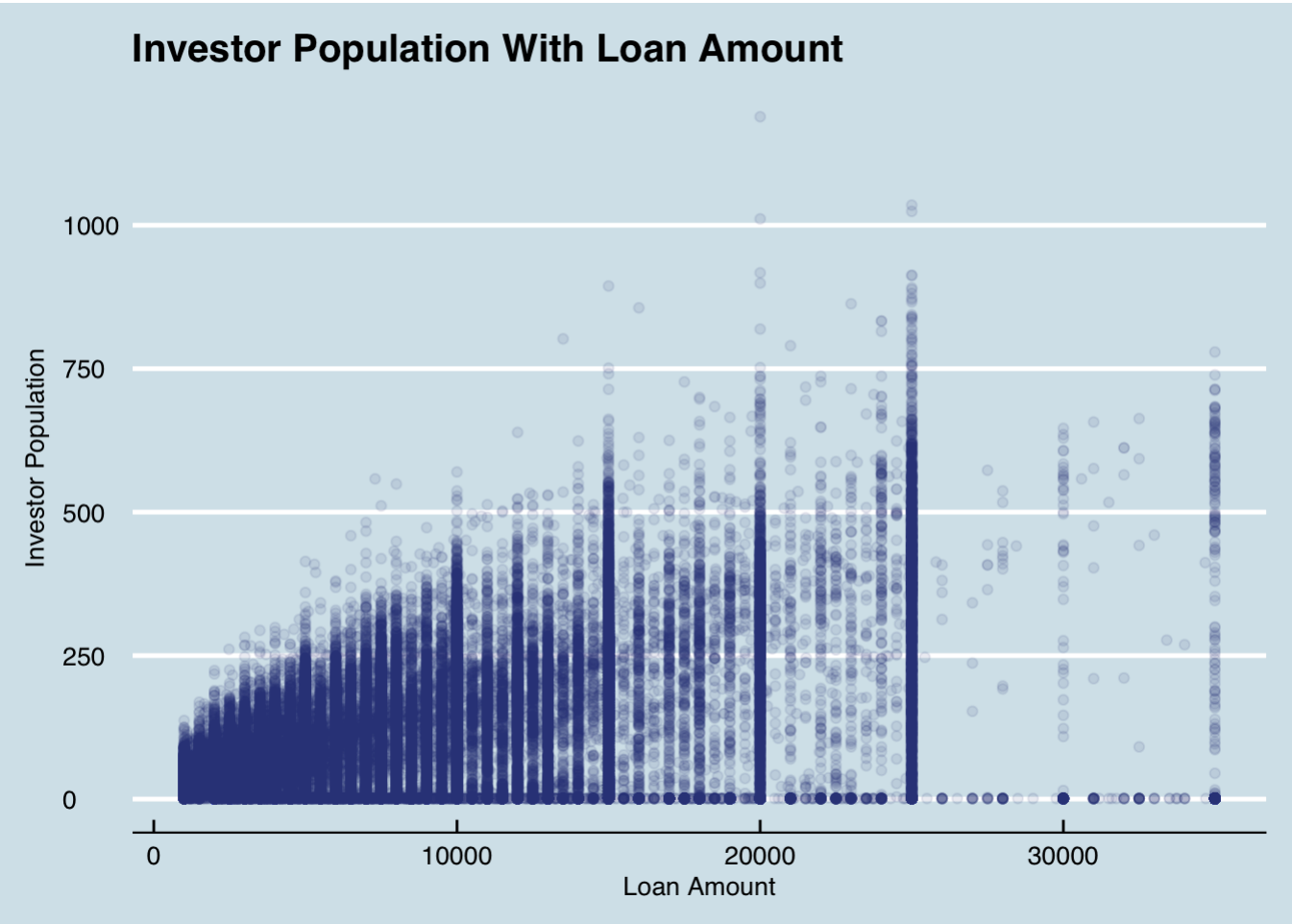
Next, I'll draw several graph to take a dip into investors' preference The below graph shows the same risk aversion theory from the perspective of investor population.

graph 10



The below graph show the relationship between Investor population and loan amount

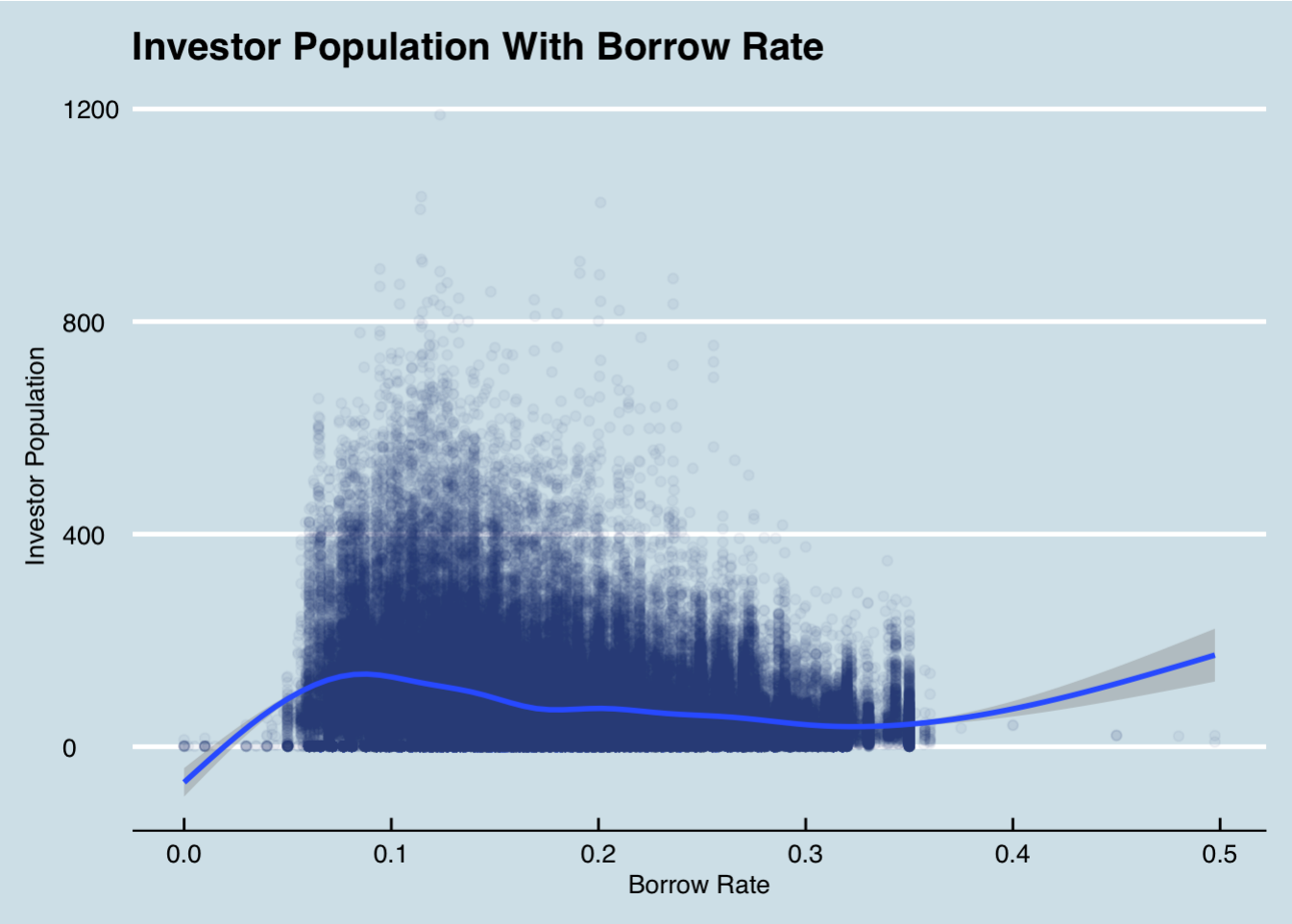
graph 11



As we can see, most loans are in the area where loan amount are below 10,000, and investor population are below 250. And we can also observe that, most borrowers like to borrow money in whole amount such as 10,000, 15,000, 20,000, 30,000, 35,000.

Next I'll take a look at the relationship between Investor population with Borrow Rate.

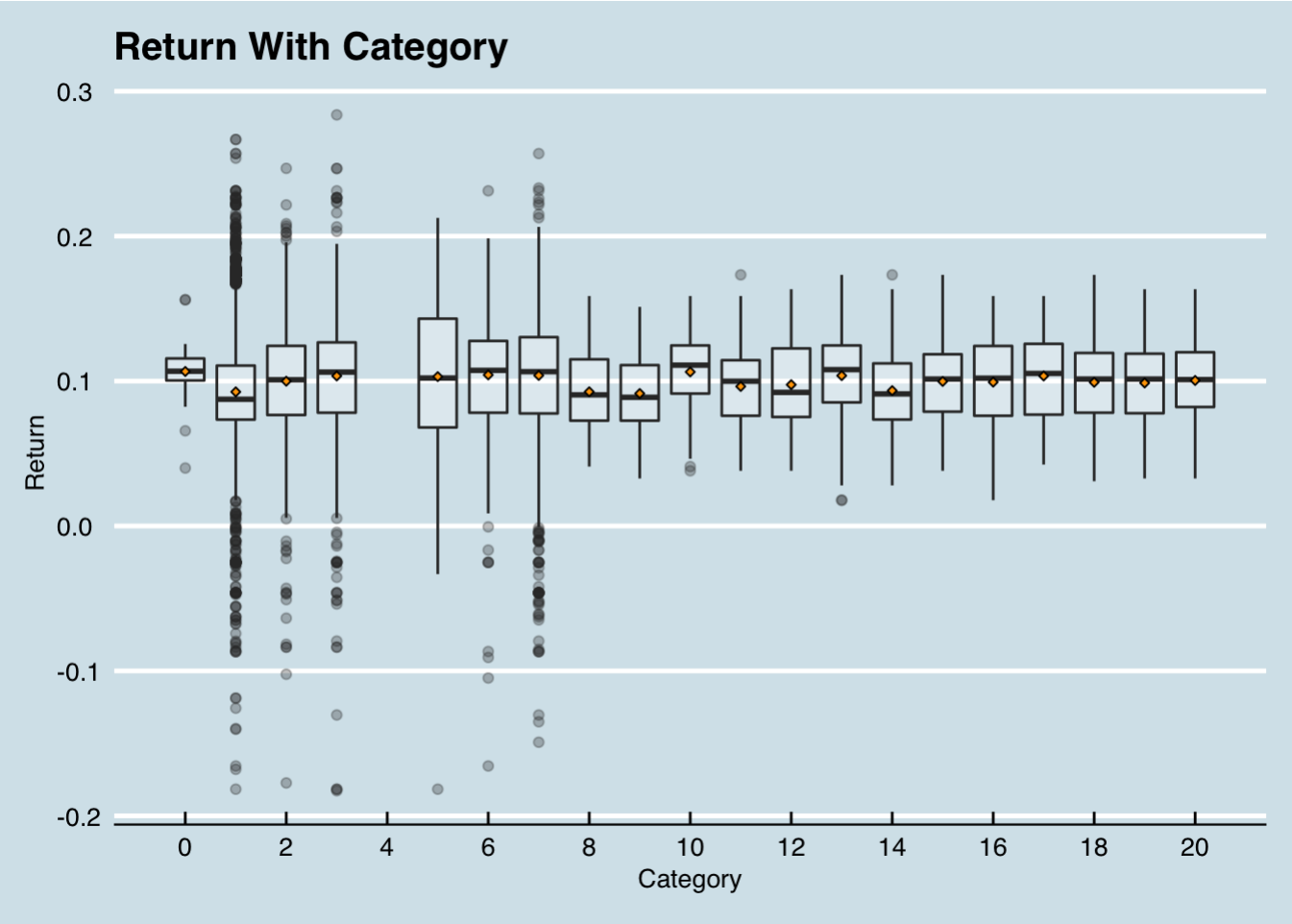
graph 12



This graph is quite intesting, the majority investors prefer loans within range from 10-20, there do exist some risk lovers who dare to invest in loans with rate higher than 30%.

After all these investigations, I'm wondering which type of investment would give the best return, after all, as a risk averse investor, I would definitely love to find the investment which gives high return but low risk.

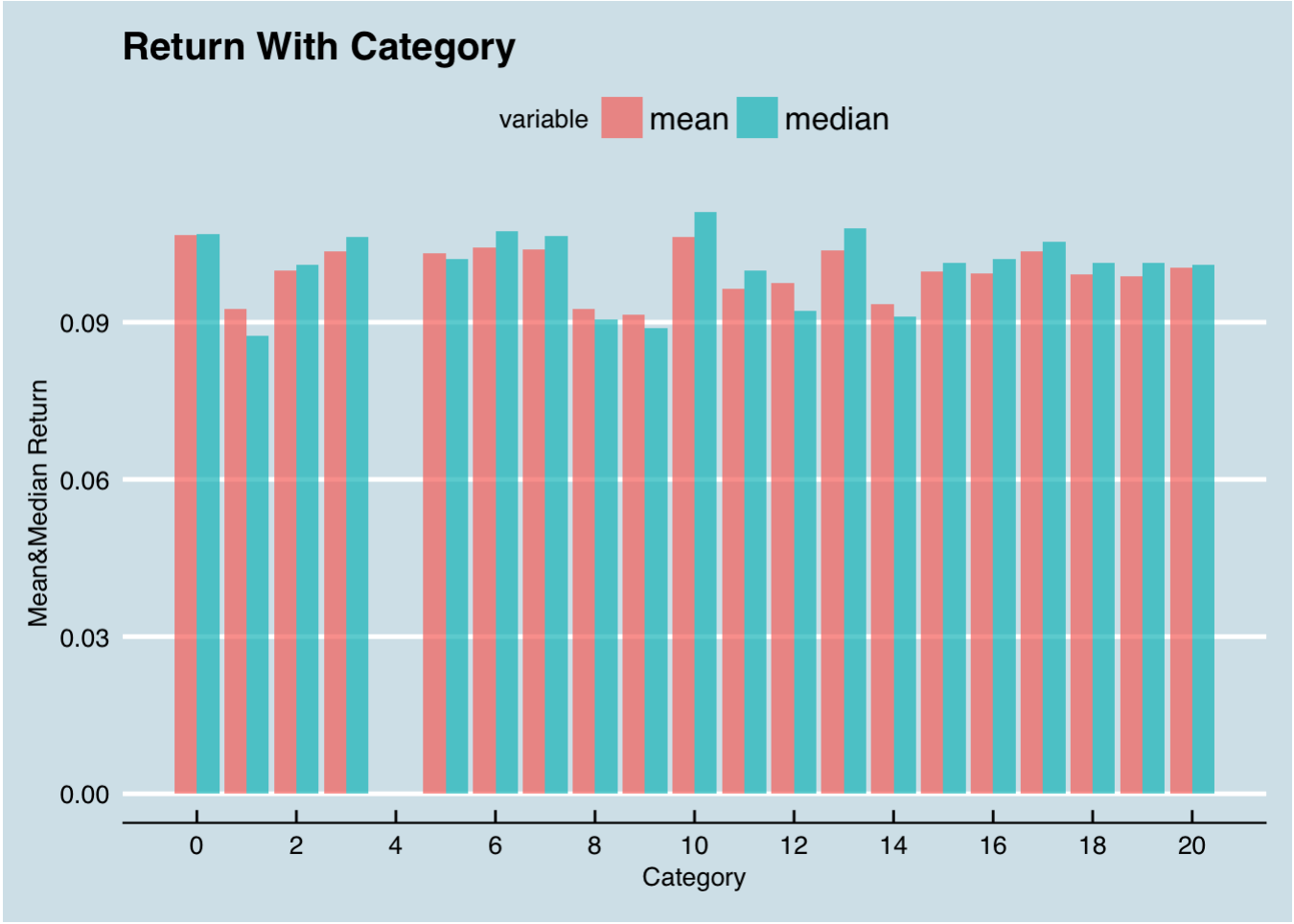
graph 13



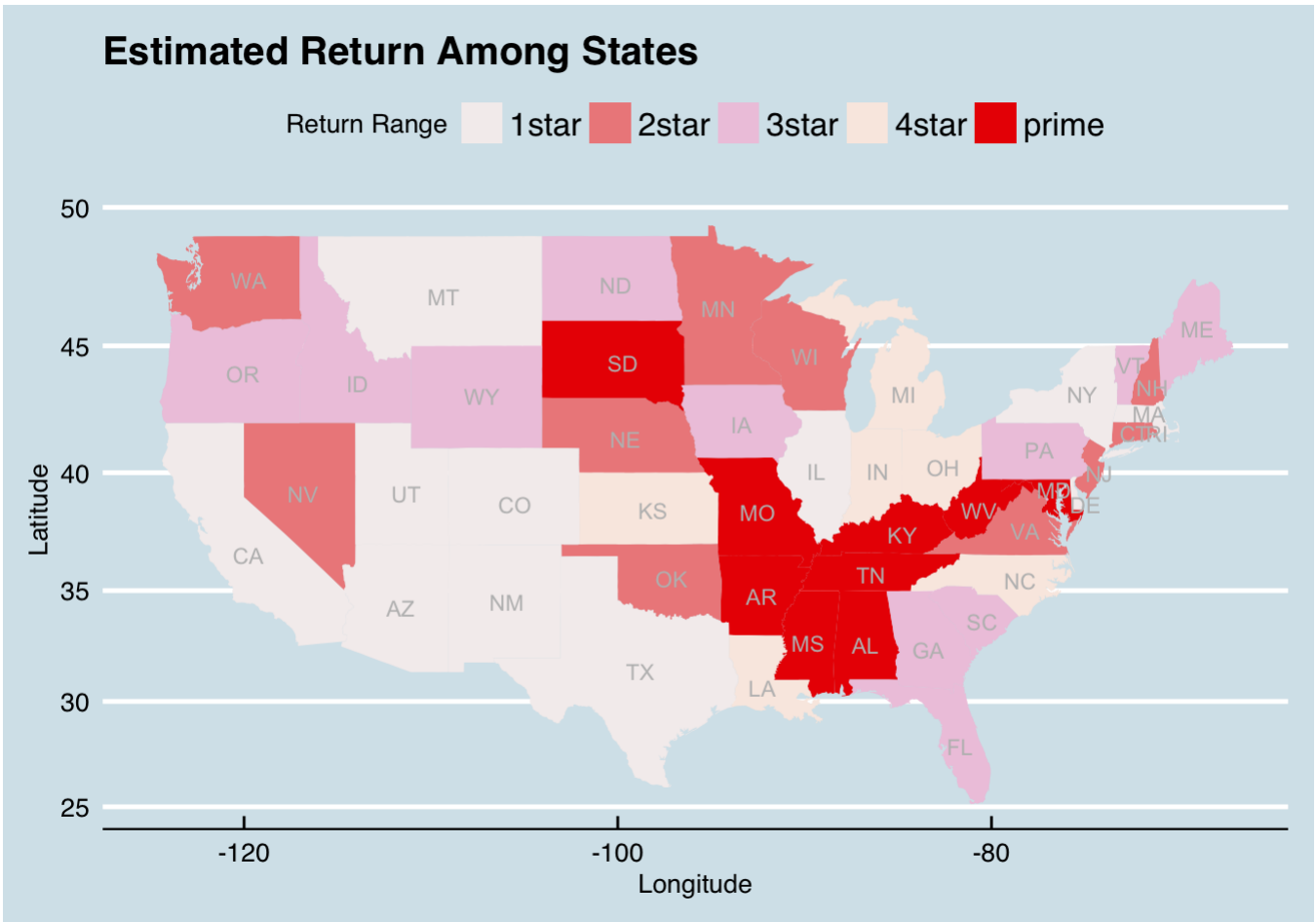
As we can see from the graph, most categories give roughly the similar return, the difference is within 1%, yet, their risk is drastically different, type 10(Cosmetic Procedure) gives a good reward and a small variation.So I conclude - The future belongs to those who believe in the beauty :).

Here is a mean and median return graph similar to the above graph,but only focus on the mean and median.

graph14


Return With Category
variable
mean
median
Mean&Median Return
0.09
0.06
0.03
0.00
0
2
4
6
8
10
12
14
16
18
20
Category

graph 15


Estimated Return Among States
Return Range
1star
2star
3star
4star
prime
Latitude
50
45
40
35
30
25
-120
-100
-80
Longitude
WA
MT
ND
MN
OR
ID
SD
WI
MI
WY
NY
VT
NH
ME
MA
RI
CT
NE
IA
PA
NJ
IL
IN
OH
NV
UT
CO
KS
MO
WV
MD
DE
KY
VA
CA
OK
AR
TN
NC
AZ
NM
MS
AL
SC
GA
TX
LA
FL

Bivariate plot Summary

In the above plots, I analysed 2 major area, reward and risk and briefly discussed how good is the Prosper scoring system.

Here are some of the findings worth noting.

Firstly,faced with two investments with a similar expected return ,investorsprefer the one with the lower risk,and majority investors are shy away from high rewards loans since they bear signifcantly higher risk. thus, most investors are risk averse. And some risk averse investors would like to sacrifice return for getting a lower risk loan. Neverthless,there do exist some risk lovers who dare to invest high risk loans for higher rewards.

Secondly, most categories give roughly the similar return, the difference is within 1%, yet, their risk is drastically different, Cosmetic Procedure gives a good reward and a small variation. And I conclude that The future belongs to those who believe in the beauty.

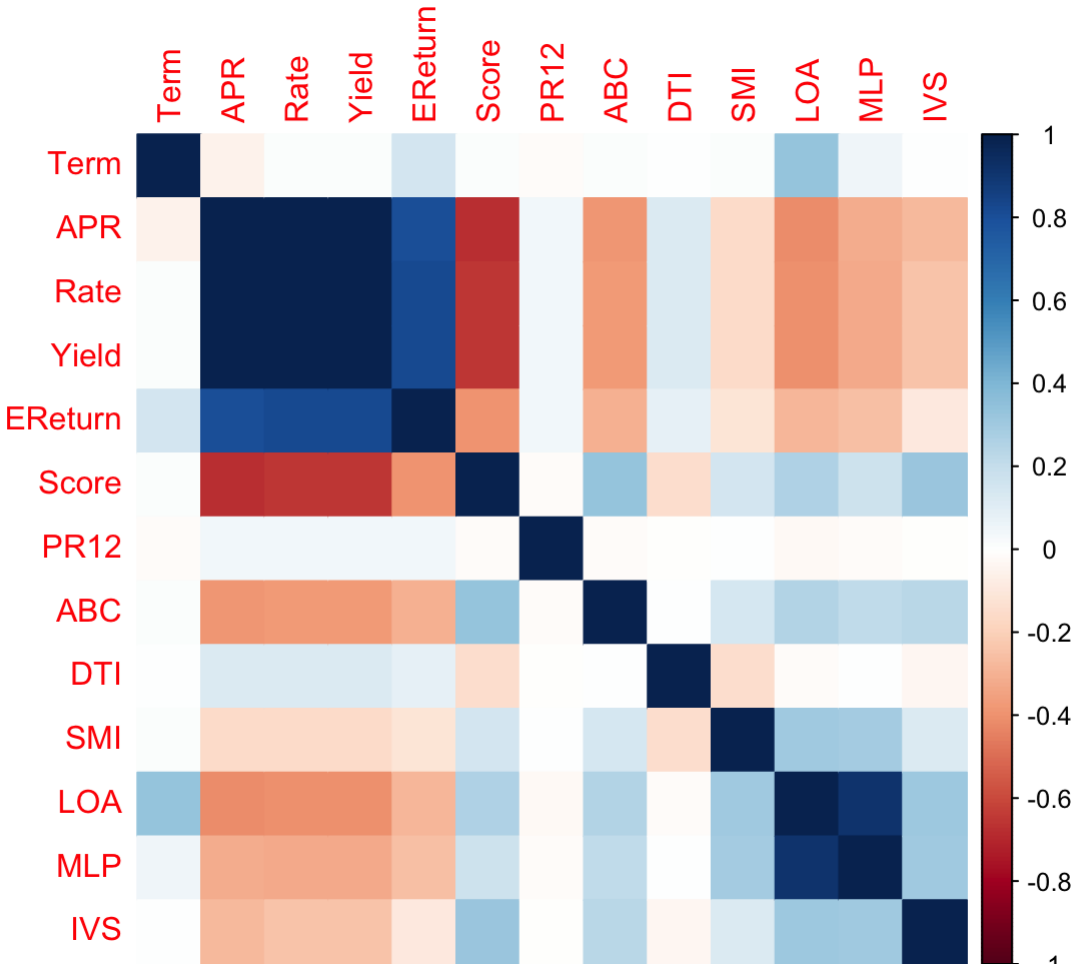
Last but not the least, there do exits some flaw in Prosper scoring system which need further improvement whether in the model or algorithm or other aspect such as policy or operations to reduce the amount of fraudster's attack.

Multivariate Plots

Lastly, I'll take a look at more variables to see if any interesting pattern can be detected

first I'll draw a correlation map between some selected variables.

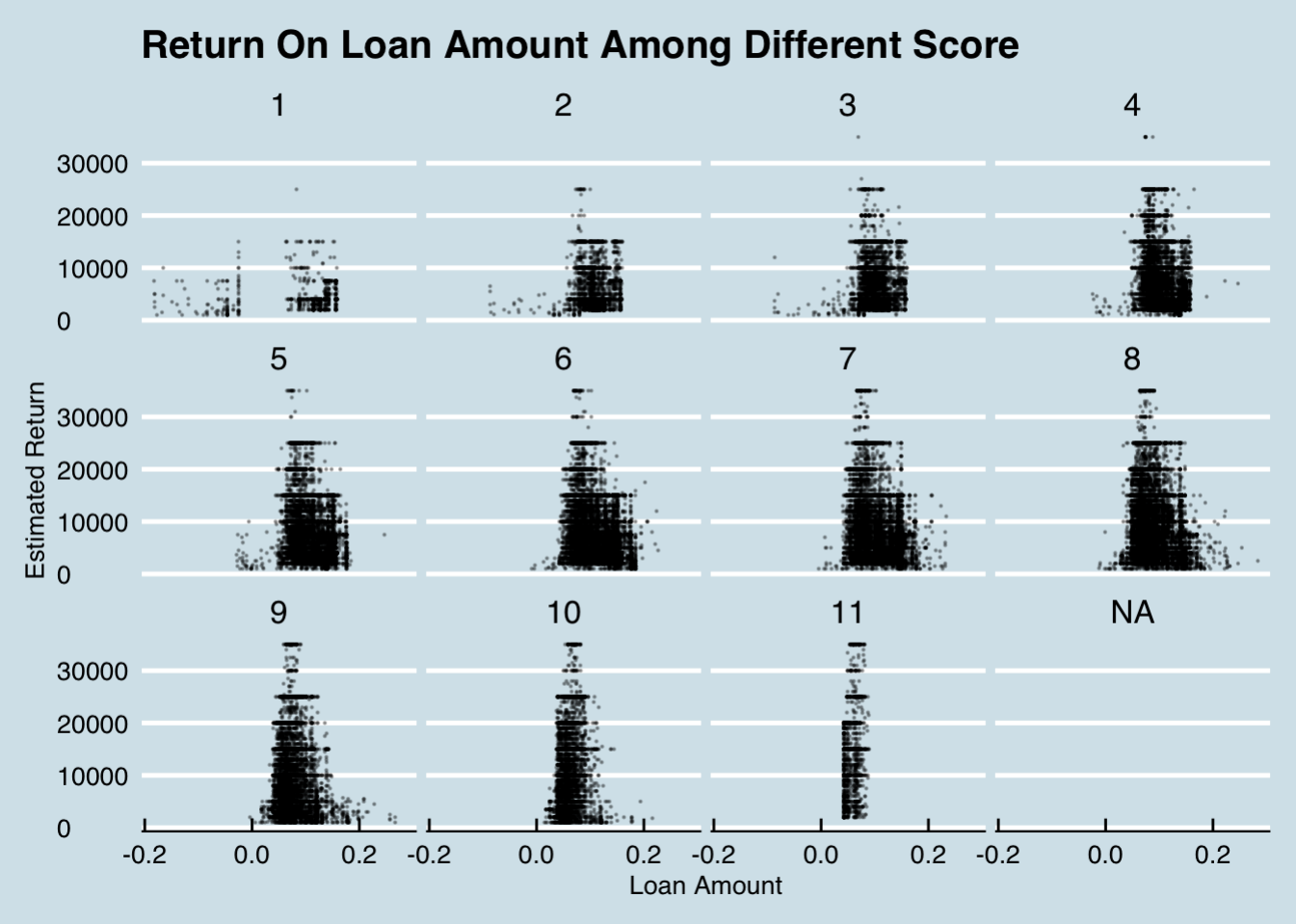
graph 16



As we can see from above picture, LOA(LoanOriginalAmount) has strong negative relationship with Rate,Prosper score has a even higher negative relationship with Borrower Rate, these two terms greatly influenced investors' choice on which loan to invest. Estimated Return may be a better measurement for potential return,since it's an adjusted term, yet we see less correlation between Estimated return and terms such as LOA(LoanOriginalAmount), Prosper Score,SMI(Stated Monthly Income).

The below graph shows the relationship between Estimated Return and Loan Amount based on different Prosper Score.

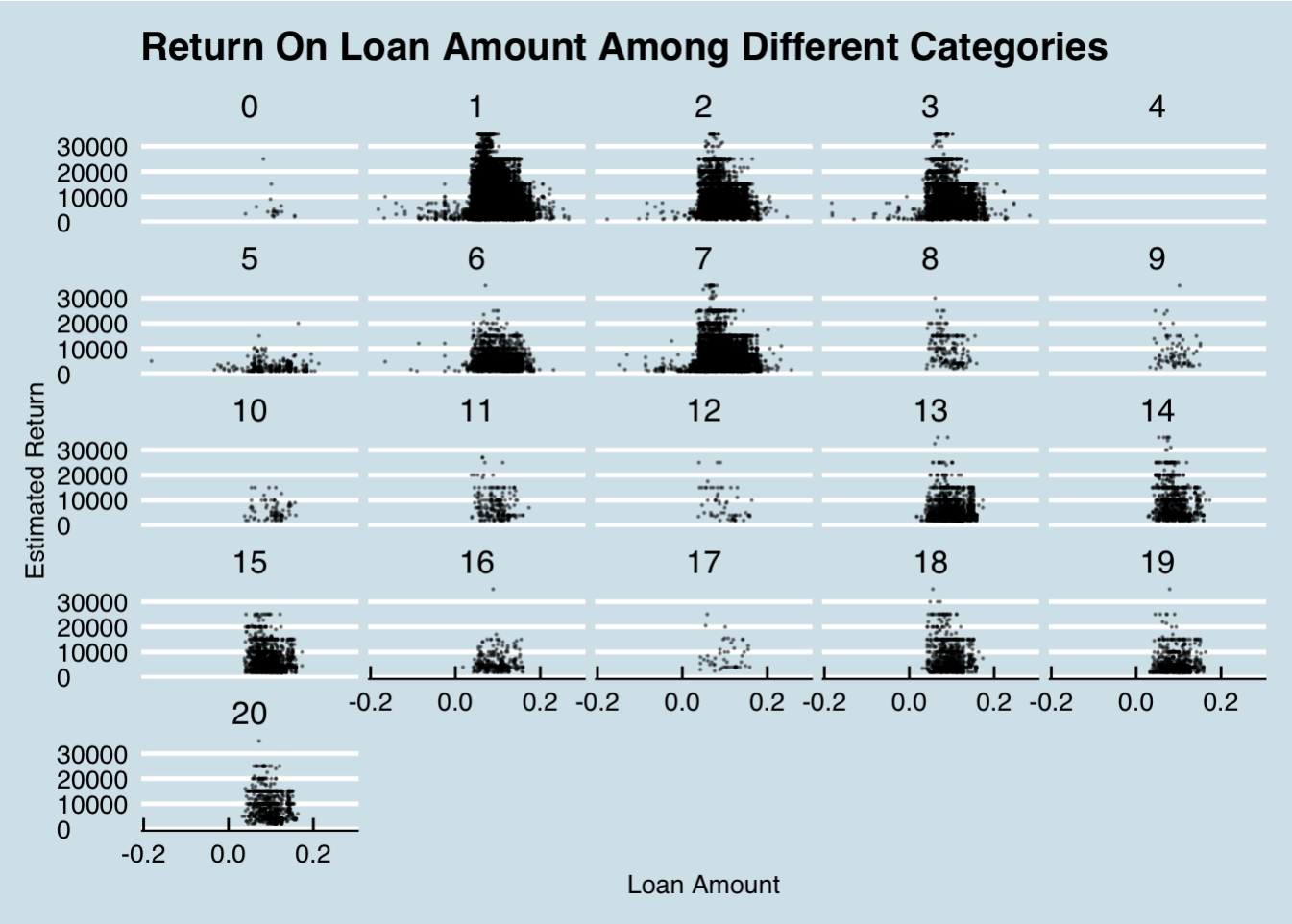
graph 17



As we can clearly see from the above chart, graphs with higher scores have a isosceles triangle shape and tends to have a long tail on the right, meanwhile graphs with lower scores have isosceles triangle shape and tends to have a long tail on the left.

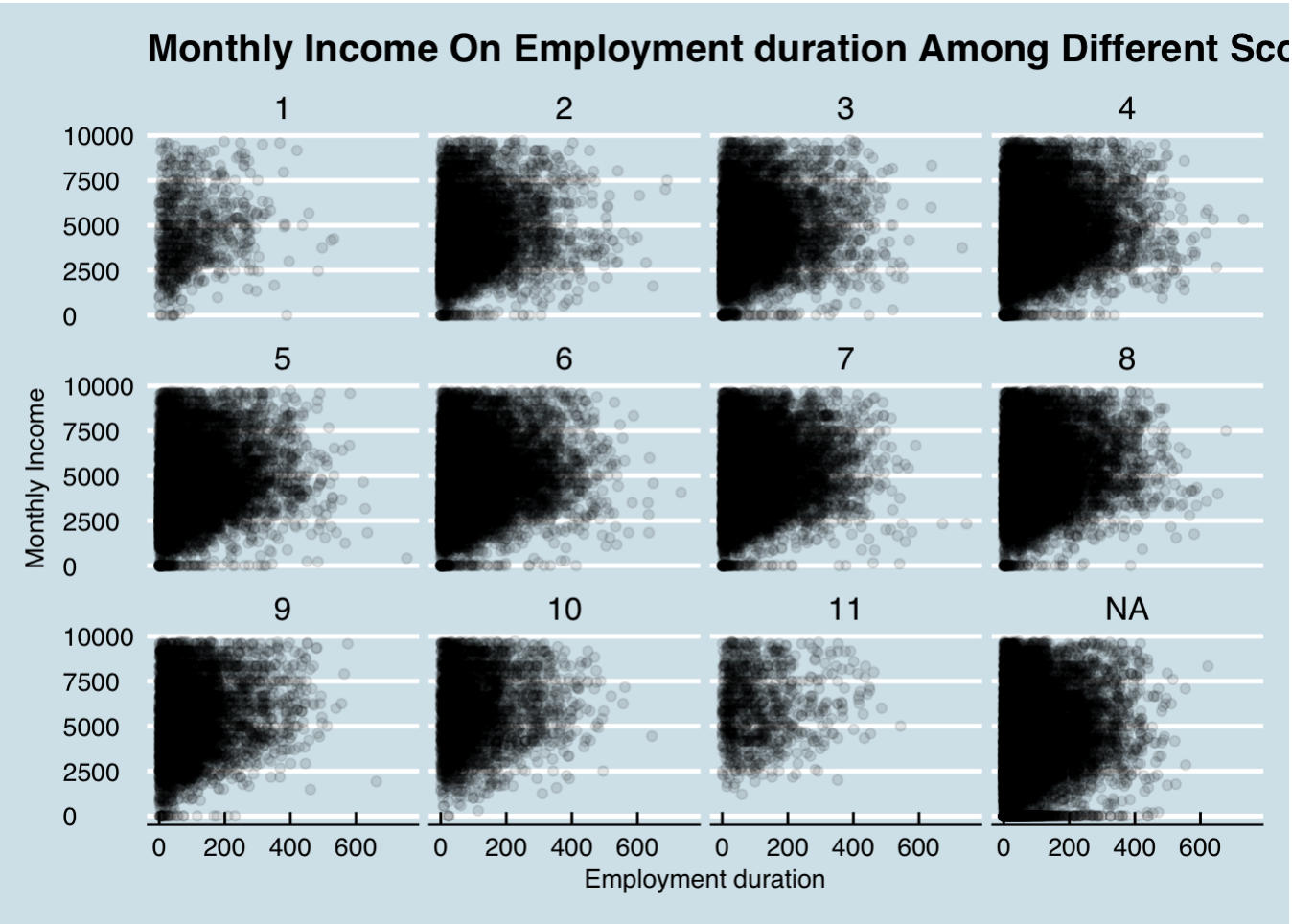
Return On Loan Amount Among Different Categories

graph 18



Monthly Income On Employment duration Among Different Score

graph 19



Multivariate Analysis

In the above plots, I analysed the correlations between multiple variables and dive deeper in the relationship of Estimated Return and Loan Amount based on different factors, as well as the relationship between income and eployment duration based on different factors.

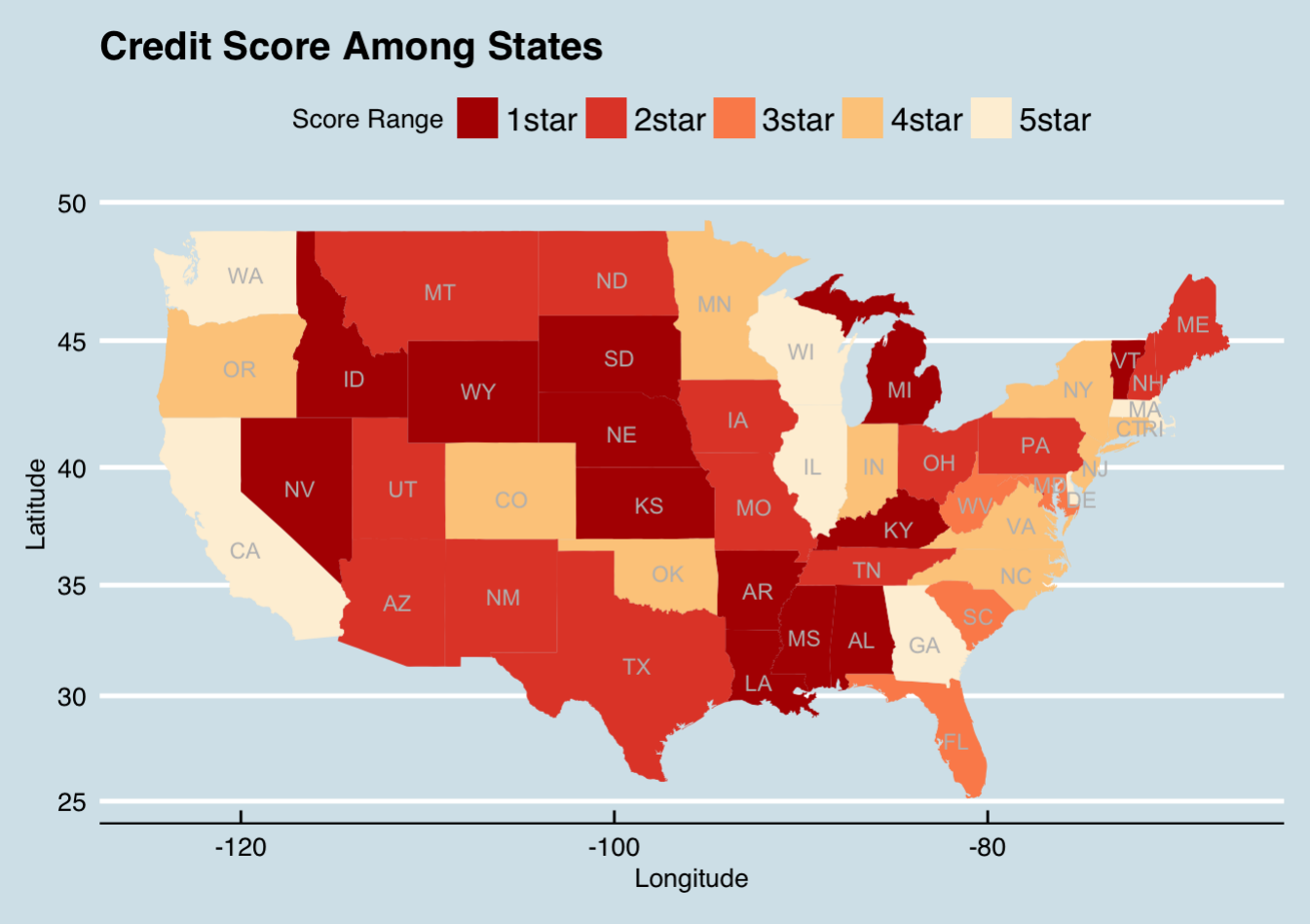
Here are some findings related to the above graph.

Firstly, Loan Amount and Prosper score have strong negative correlation with Borrower Rate, and these two terms greatly influenced investors' choice on which loan to invest.

Secondly, In the return on loan amount chart, graphs with higher scores have a isosceles triangle shape and tends to have a long tail on the right, meanwhile graphs with lower scores tends to have a long tail on the left.

Final Plots and Summary

Plot One



Description One

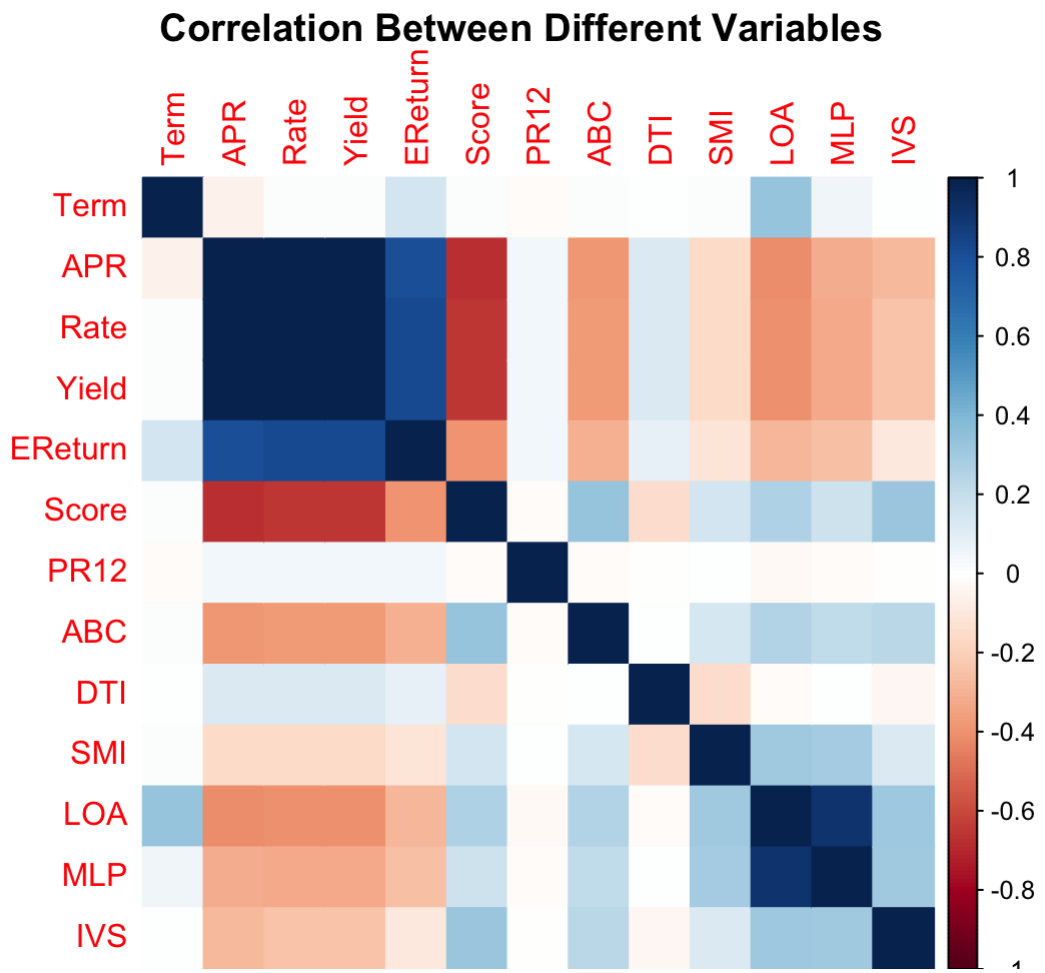
The above picture show the average credit score among states, even though credit score doesn't vary much among states and it has nothing to do with personal loans, we can still get a general idea about the overall performance of these states. States along side the east coast and west coast have better credit score in general than those states in the middle region.

Plot Two

Description Two

As we can see from the above graph, most loans with score below 8 offer similar expected return, yet they bear different risks. For instance, for borrowers with score 1, you would probably lose 20% of your principle, while for borrowers with a higher score , let's say score 7, the maximum amount you would lose is around 2% of your principle. This observation fits our theory exactly. Further more, since the number of prime loans are limited, investors would like to sacrifice return for a lower risk loan.

Plot Three



Description Three

As we can see from above picture, LOA(LoanOriginalAmount) has strong negative relationship with Rate,Prosper score has a even higher negative relationship with Borrower Rate, these two terms greatly influenced investors' choice on which loan to invest. Estimated Return may be a better measurement for potential return,since it's an adjusted term, yet we see less correlation between Estimated return and terms such as LOA(LoanOriginalAmount), Prosper Score,SMI(Stated Monthly Income).

Reflection

In this report, I analyzed the influence of multiple variables on a loan issuing in P2P industry and got a good understanding of different variables and their relationship between each other. Meanwhile, I also have lots of interesting observations and finding during the analysis.

Apparently, P2P industry is at its early age, though it is a small industry, it advances dramatically and owns multiple merits compared with banking system and credit union. Its score system takes advanced machine learning algorithms and models , utilize the potential of big data and well measured the credit worthiness of their customers. But, there are still weakness and flaws presented in its system and these flaws are actively used by fraudulent appliers. This will be a big threat to the development of the industry.

Judging by its Term structure, state applying distribution and categories distribution, we could expect a huge diversification and variation emerging in this industry. However, before it grows in to a behemoth, numerous new fraud schemes will appeal and spoil the industry in the cradle. In addition, banks and credit union will also join the competition and try to topple this little brother. Further more, regulation and new laws will also pose potential thread to the industry.

The advent of this new business model is much welcomed from the perspective of an investor or a borrower , as it drastically brings down the cost of raising a funds and provides a better option to the investors, neediness to mention these convenient services it gives. Though the road forward is tough and arduous, I still have great faith in this industry simply because it brings down the transaction cost.

The dataset has rich information to discover, limited by the time and imagination, I could only scratch the surface, If I have more time, I would like to build the model myself and predict the loan outcome and join the whole process of the loan applying and investing process to see if any further interesting things can be revealed.