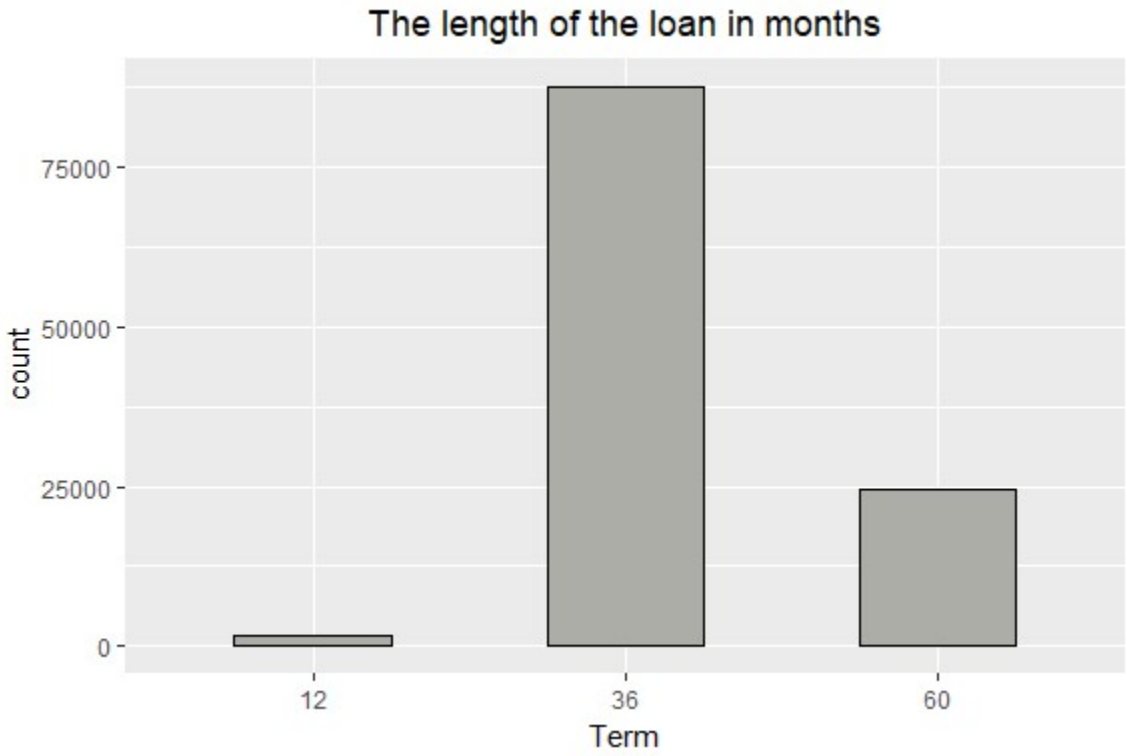
Exploratory Analysis of Prosper.com Loans

Bin Liu

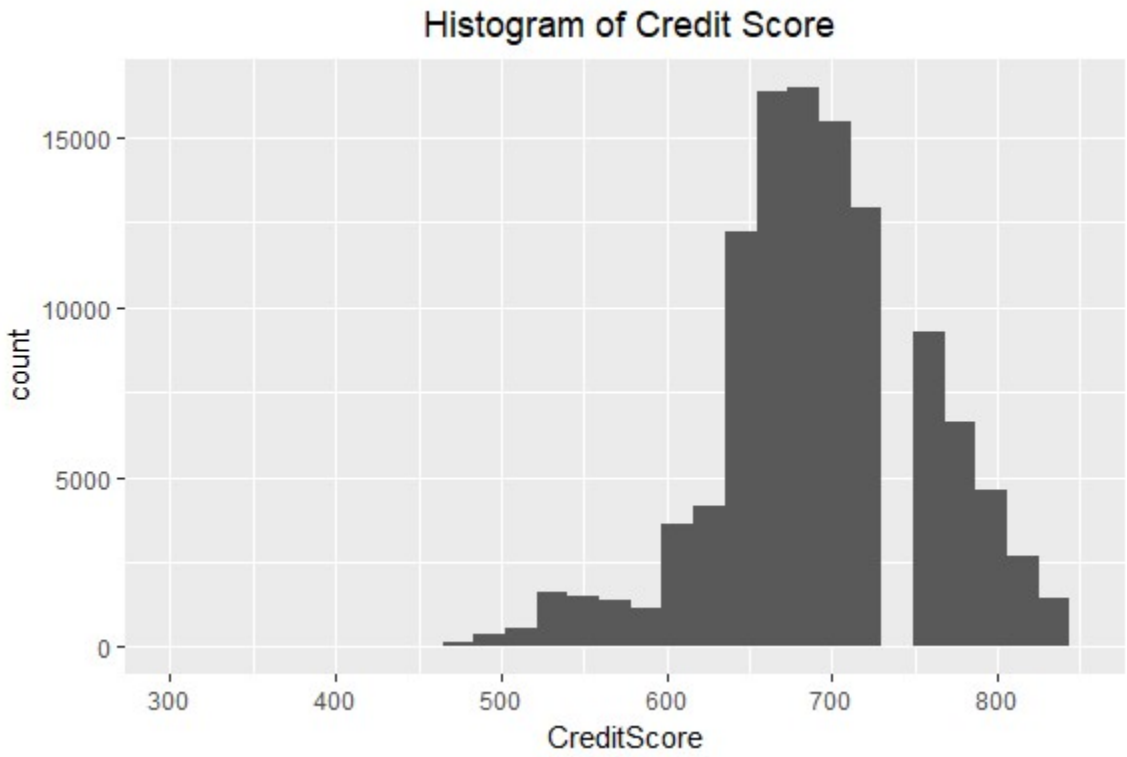
Udacity Data Analyst Nanodegree Project 4

Univariate Plots Section

What is the most frequent Term of loan?

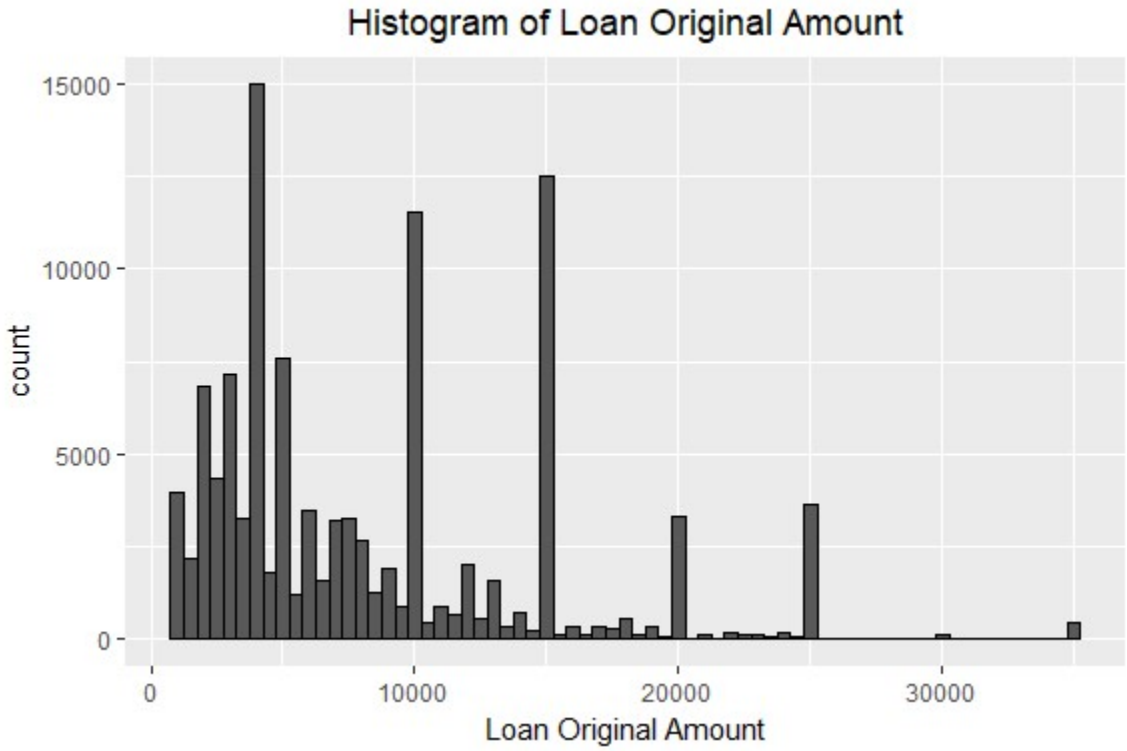


What's the borrower's credit score?



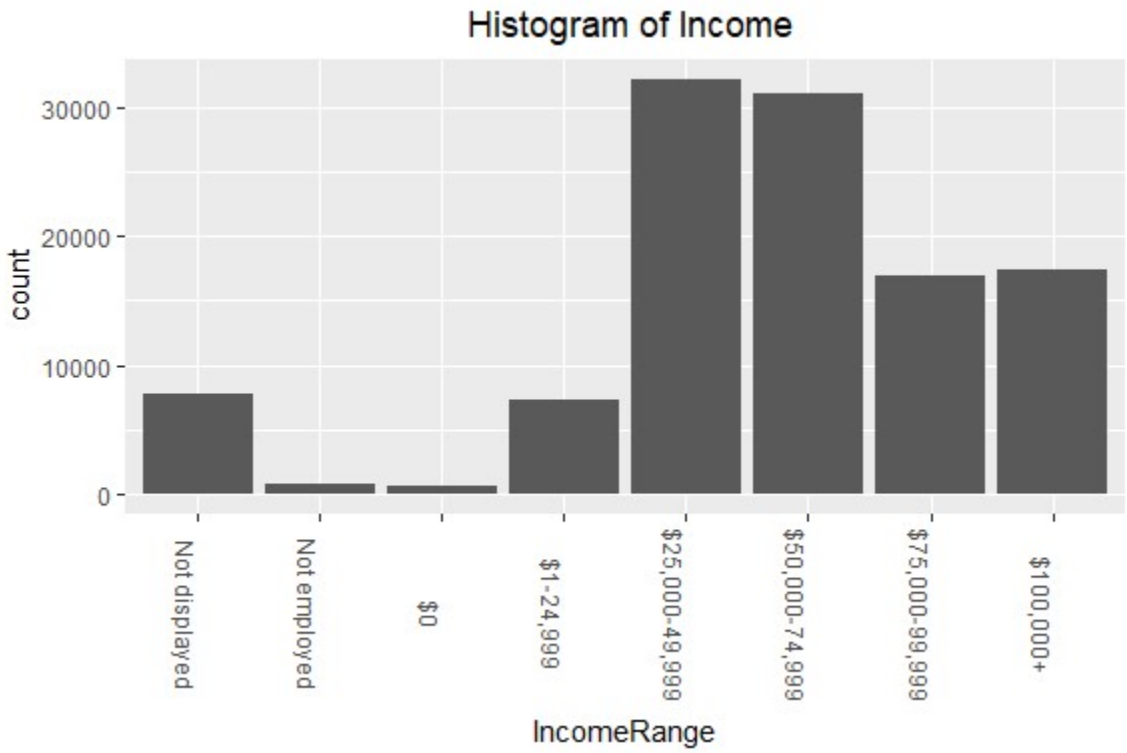
Credit scores range from 300 to 850. Prosper borrower have a median score of NA, which is considered good credit. Prosper now requires a minimum credit score of 640 for new borrowers or 600 for returning borrowers, but initially, subprime borrowers could also apply for loans.

How much people are borrowing?

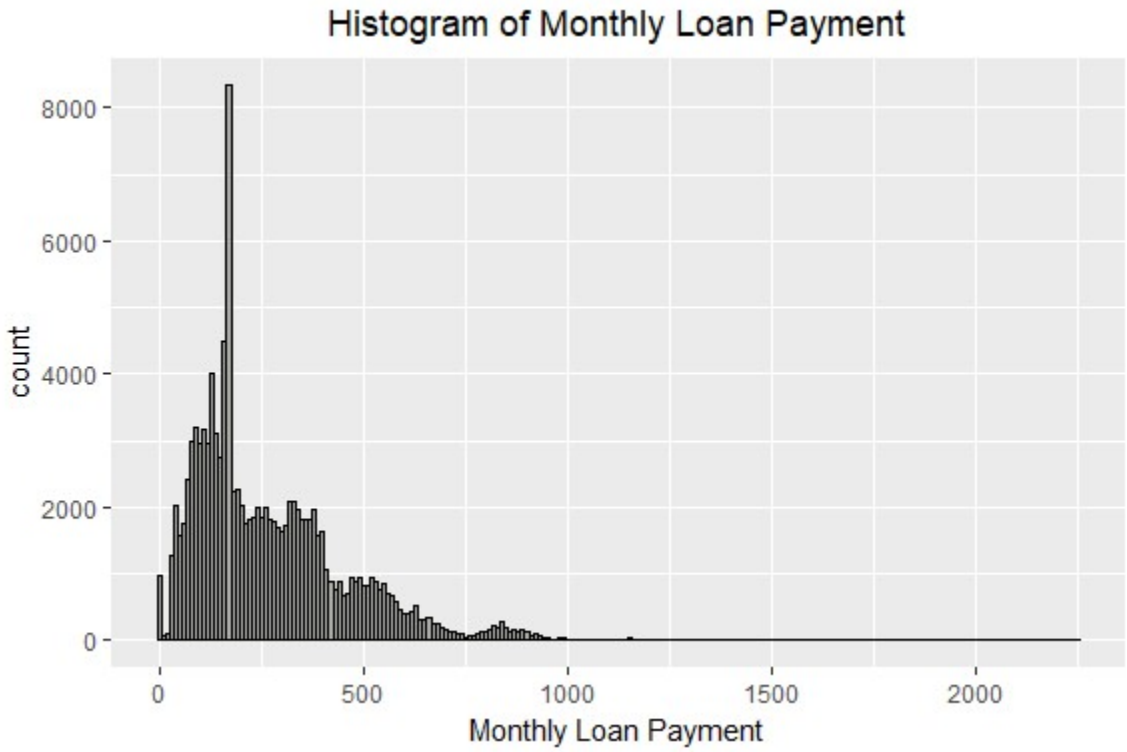



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1000    4000    6500    8337   12000   35000
```

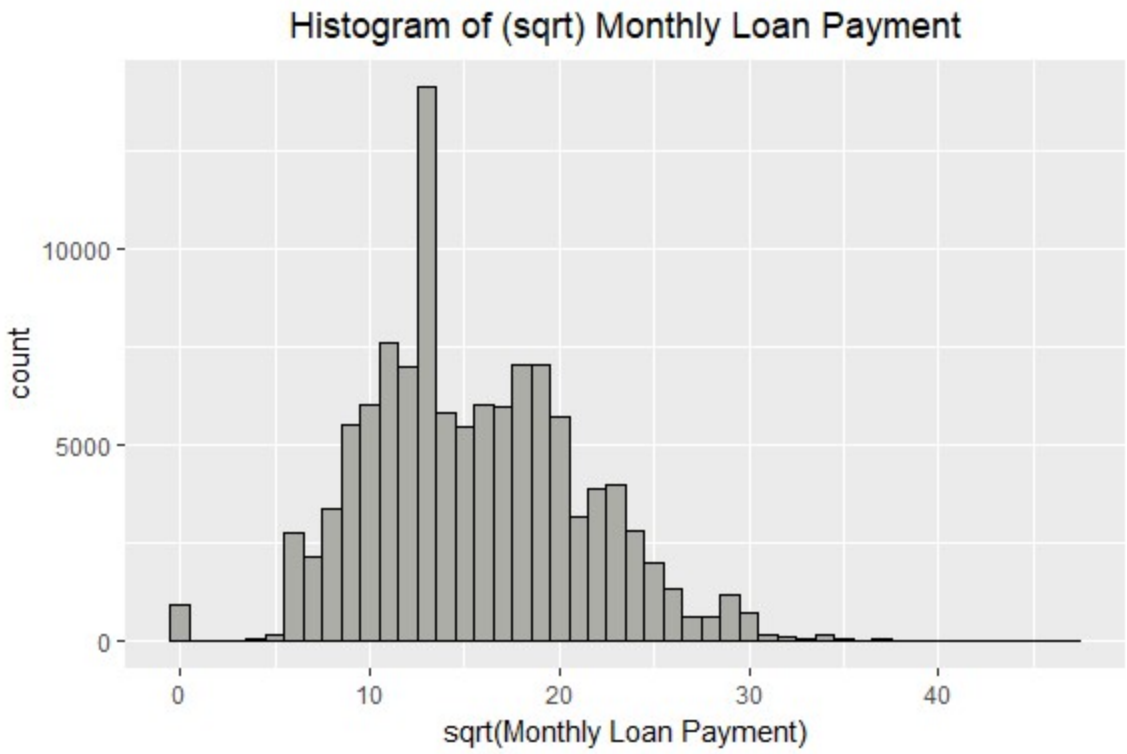
How much borrowers earn annually?



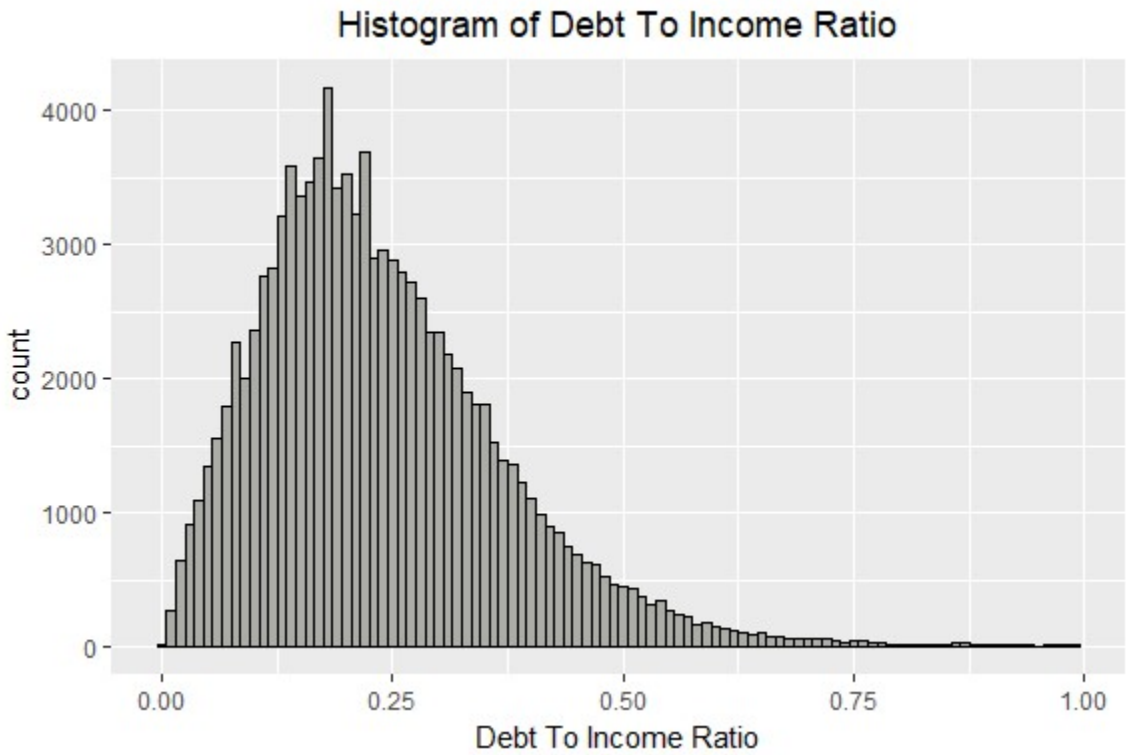
What's the distribution of Monthly Loan Payment?




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   131.6   217.7   272.5   371.6  2251.5
```



How the Debt to Income is distributed?

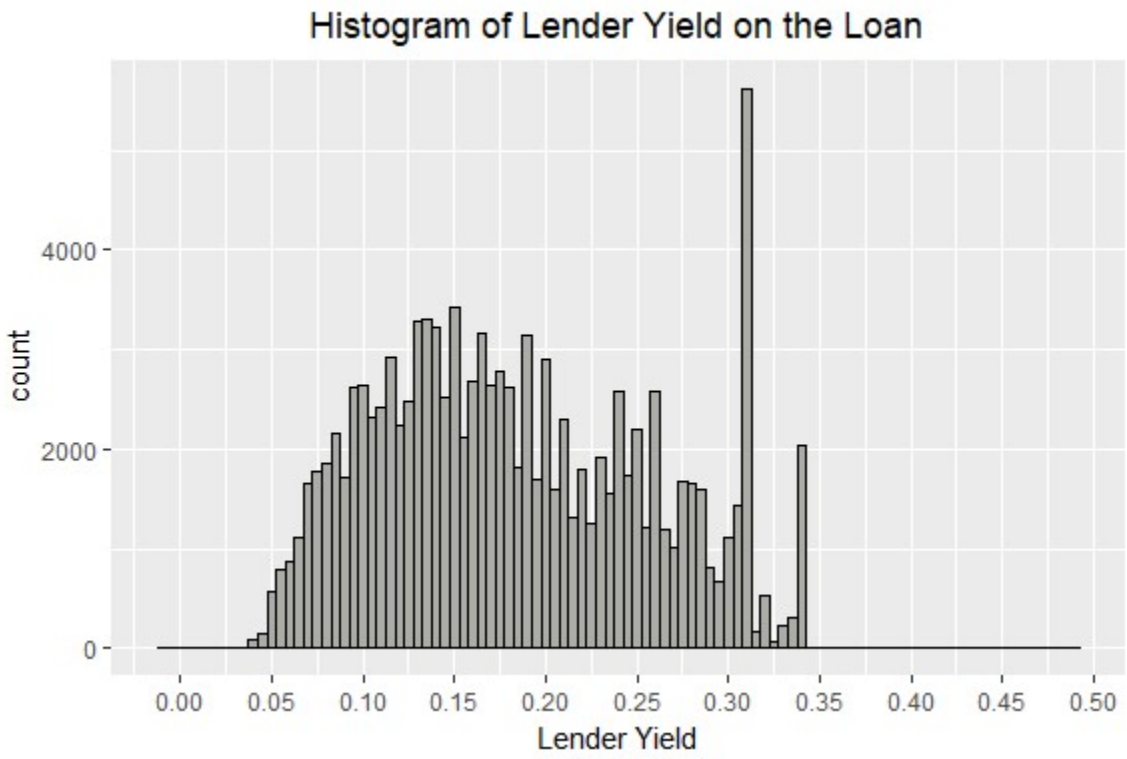



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   0.140   0.220   0.276   0.320  10.010    8554
```

Most of the borrowers are trying to keep their Debt to Income Ratio bellow 0.32 (3rd quarter value).

How lenders benefit from investing in loans?

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



Most lenders yields is between 0.05 and 0.35. The highest peak in the graphic is around 0.31.

Univariate Analysis

What is the structure of your dataset?

The ProsperLoan dataset contains 113937 observations and 81 variables. The loans cover the period 2005-11-15, 2014-03-12. Variables are of classes int, numeric, date, and factor.

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

There are two domain in the Prosper Loan model. Domain one is company domain (one role: Prosper), Domain two is customer domain (two role: Investor and borrower).

As a business, Prosper would be most concerned with LP_ServiceFees and LP_CollectionFees, which form their primary revenue source.

The main feature of the borrowers is ProsperRating, which is based on their credit score and history with Prosper loans.

For investors, the main features are the LenderYield (interest rate minus the service fee) and the LP_NetPrincipalLoss, which is the principal that remains uncollected after any recoveries.

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

Other variables that can help me to support my investigation are Debt To Income Ratio,

Occupation, Employment Status and Employment Duration.

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

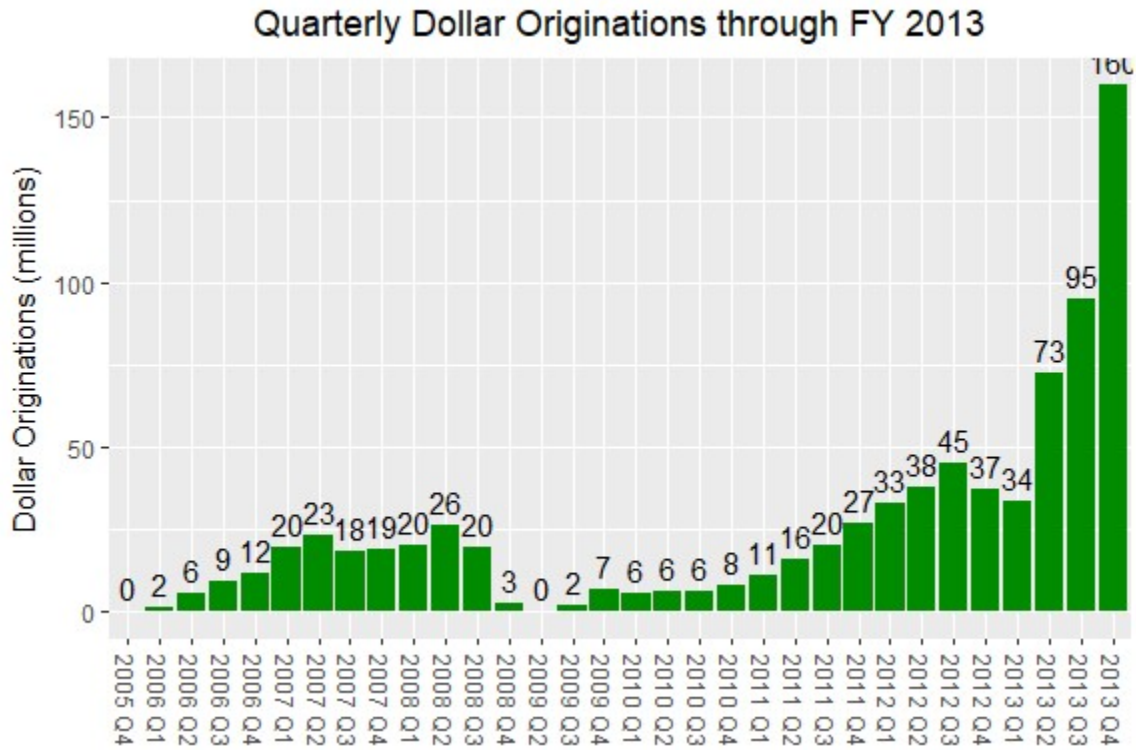
I created two new variables, The first is CreditScore, which is average number of the CreditScoreRangeUpper and CreditScoreRangeLower variables. Another is a factor variable Results, which simplify each loans status as “Current or Paid”, “Past Due”, or “Defaulted”.

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?

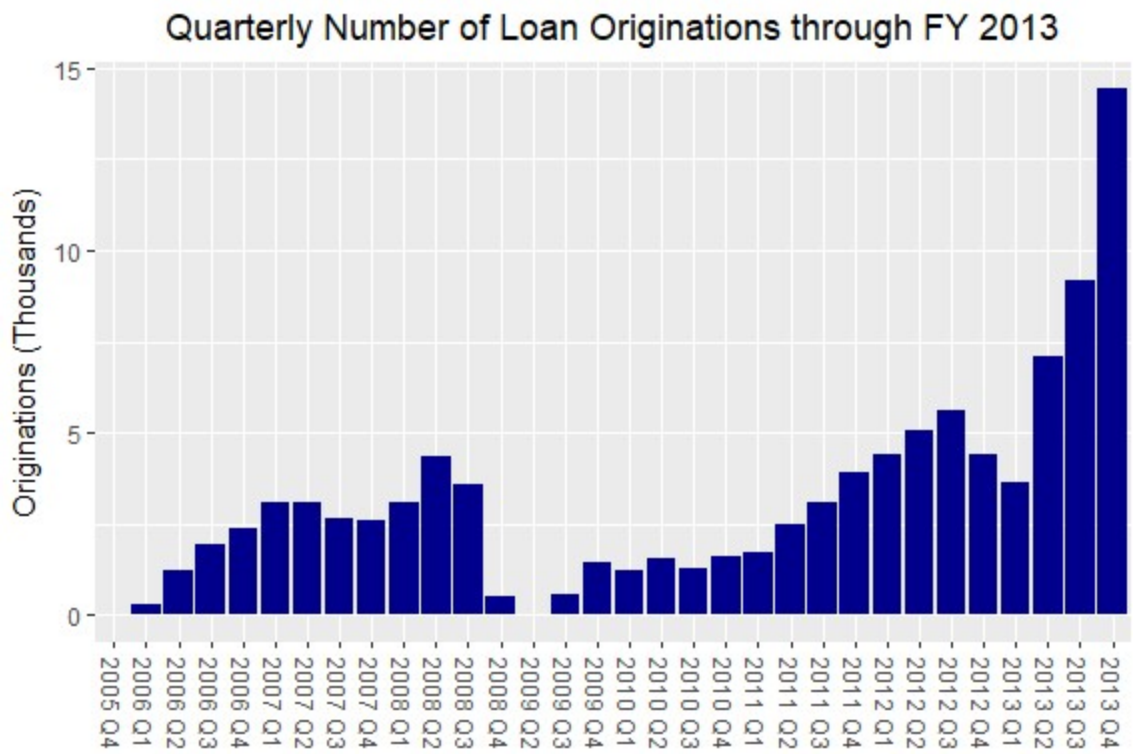
There is a high decrease of the number of the loans between the fourth quarter of 2008 and The fourth quarter of 2009. Cause is Prosper’s business model came under scrutiny by the US Securities and Exchange Commission at that time. If I was not aware of the event, this would be a unusual distributions.

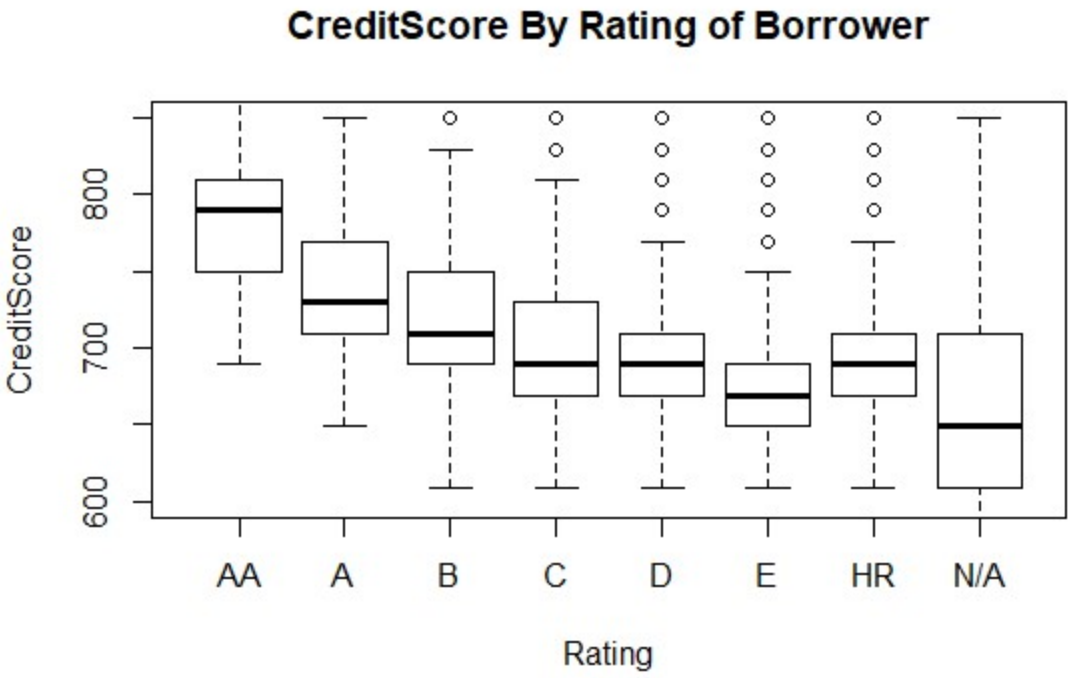
Bivariate Plots Section

The first thing I analyzed was the financial stability of Prosper business. How their business was growing and were there any ups and downs between 2005 and 2014. Let’s see how are loans distributed over the years.

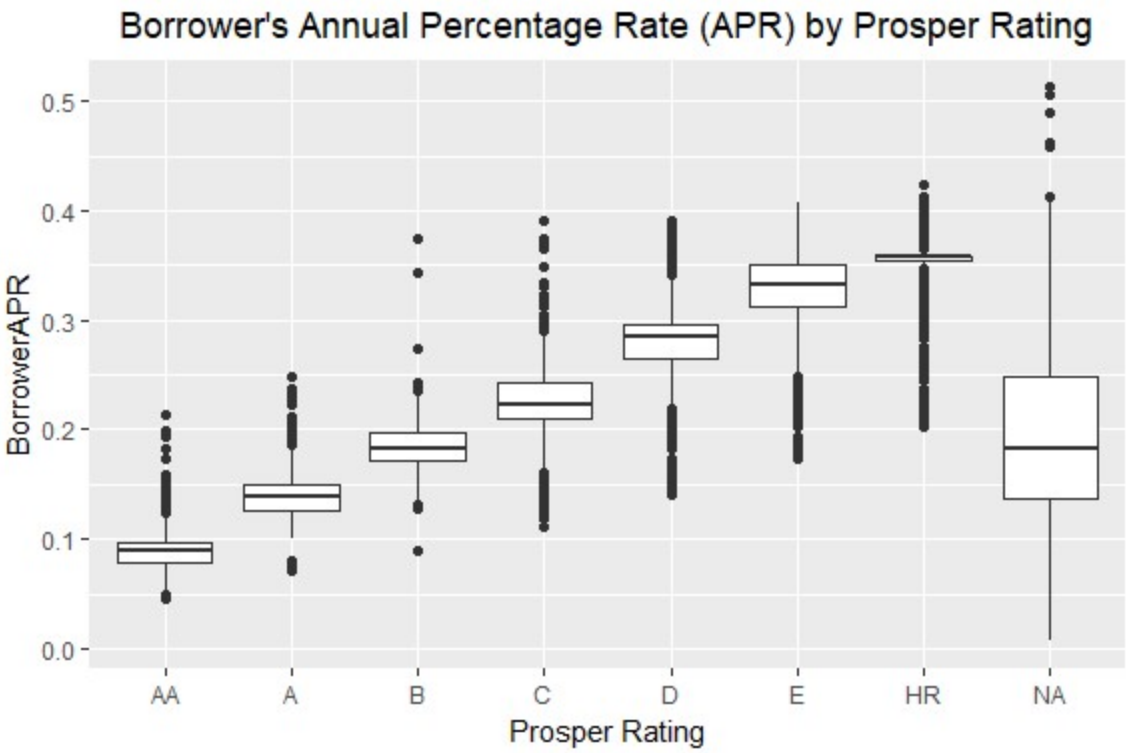


The chart in the annual report began in the third quarter of 2009. The period of October 15, 2008 to July 13, 2009 is known as Prosper’s Quiet Period (http://www.lendacademy.com/a-look-back-at-the-lending-club-andprosper-quiet-periods/ (http://www.lendacademy.com/a-look-back-at-the-lending-club-andprosper-quiet-periods/)) when they were required to suspend lending pending SEC approval. When they relaunched in July 2009, there were several changes to their lending process, so I’ll have to keep that in mind.

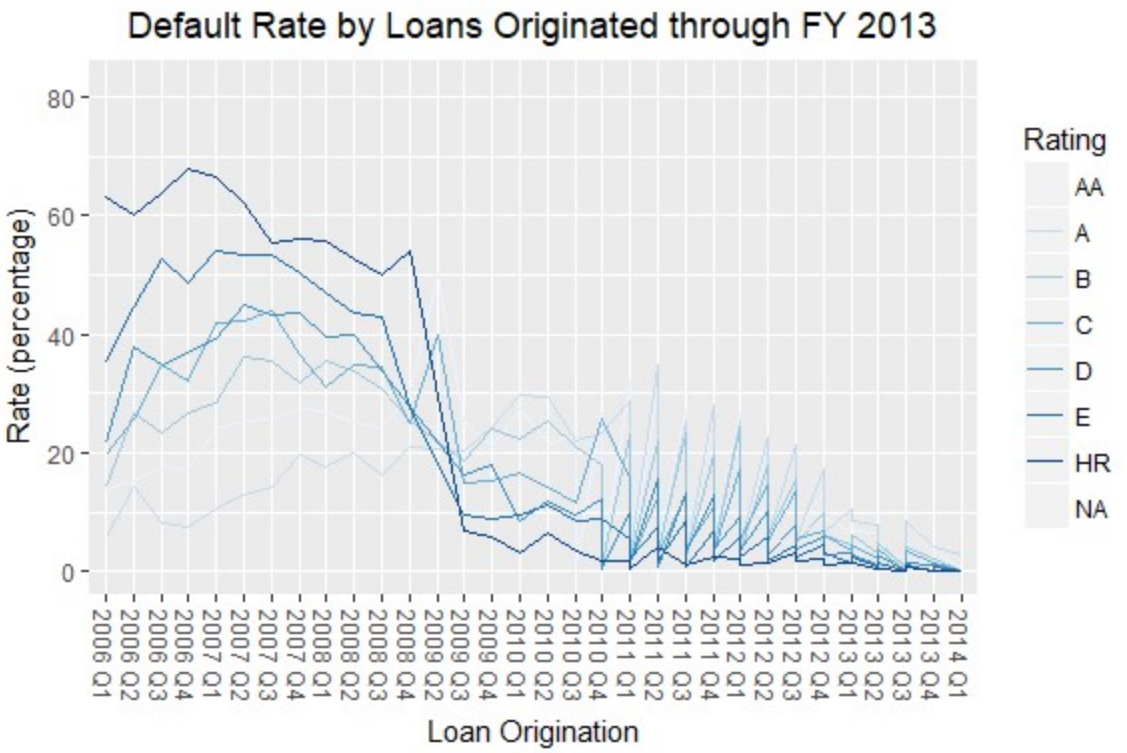




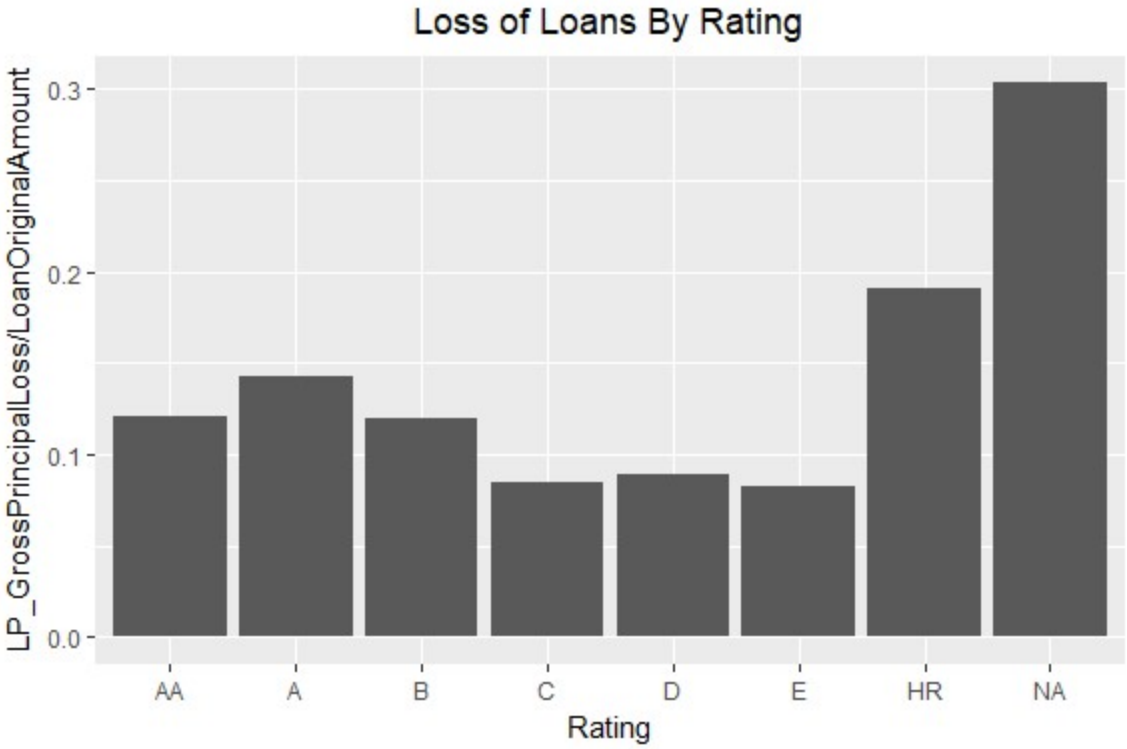
The trend is generally that higher ratings have higher credit scores, but Prosper clearly uses more than credit score, since there is a lot of overlap between the ratings.



There is much less overlap in the APR (interest rate) the borrowers are assigned.



Default rates track roughly with the risk categories. Default rates are much higher than I would have expected. The E and HR groups, even post-recession have 25-30% of loans default.



I want to see what are the main factors correlate with default. Based on this plot, I'm going to exclude everything from before July 2009 (the end of the "quiet period") and only include loans that have a ProsperRating. I'm only going to use loans that are completed, so I will exclude LoanStatus of Current or Past Due.

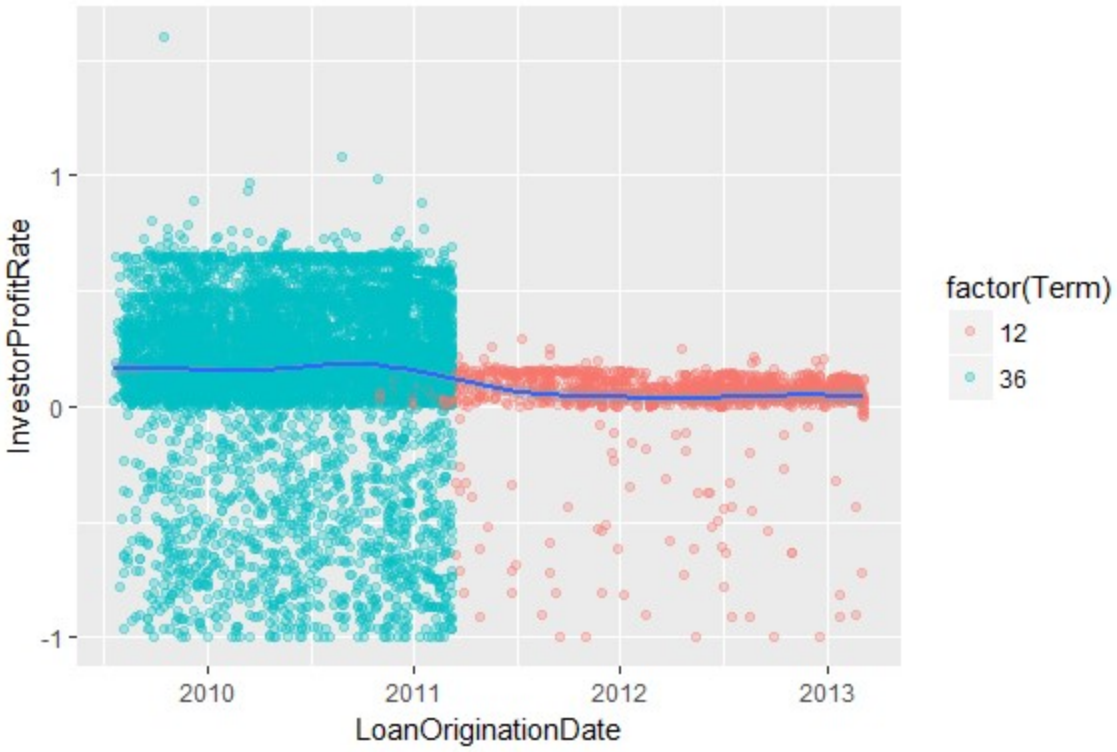
```
## [1] "Dimensions of new dataset"
```

```
## [1] 26210    14
```

```
## [1] "Loan results by rating"
```

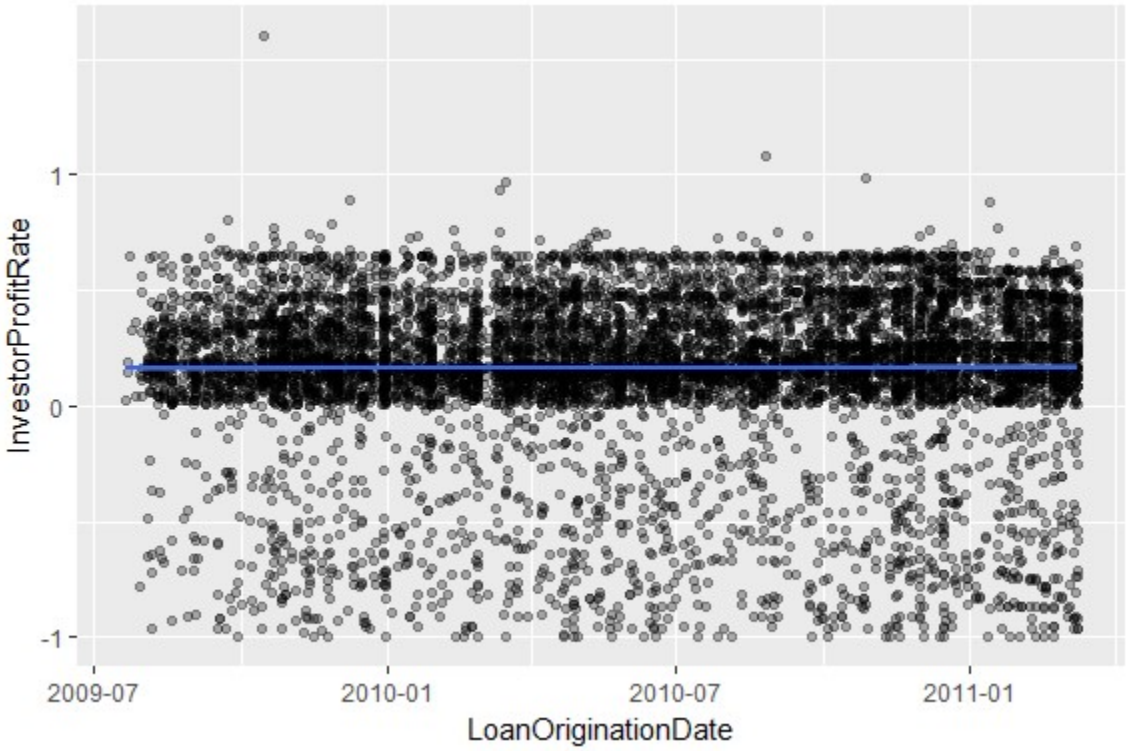
```
##
##       AA    A    B    C    D    E   HR
##   0 1324 1424 1677  840  588  405   83
##   2 2538 2322 4220 3015 2824 3237 1713
```

```
## [1] 10078    14
```



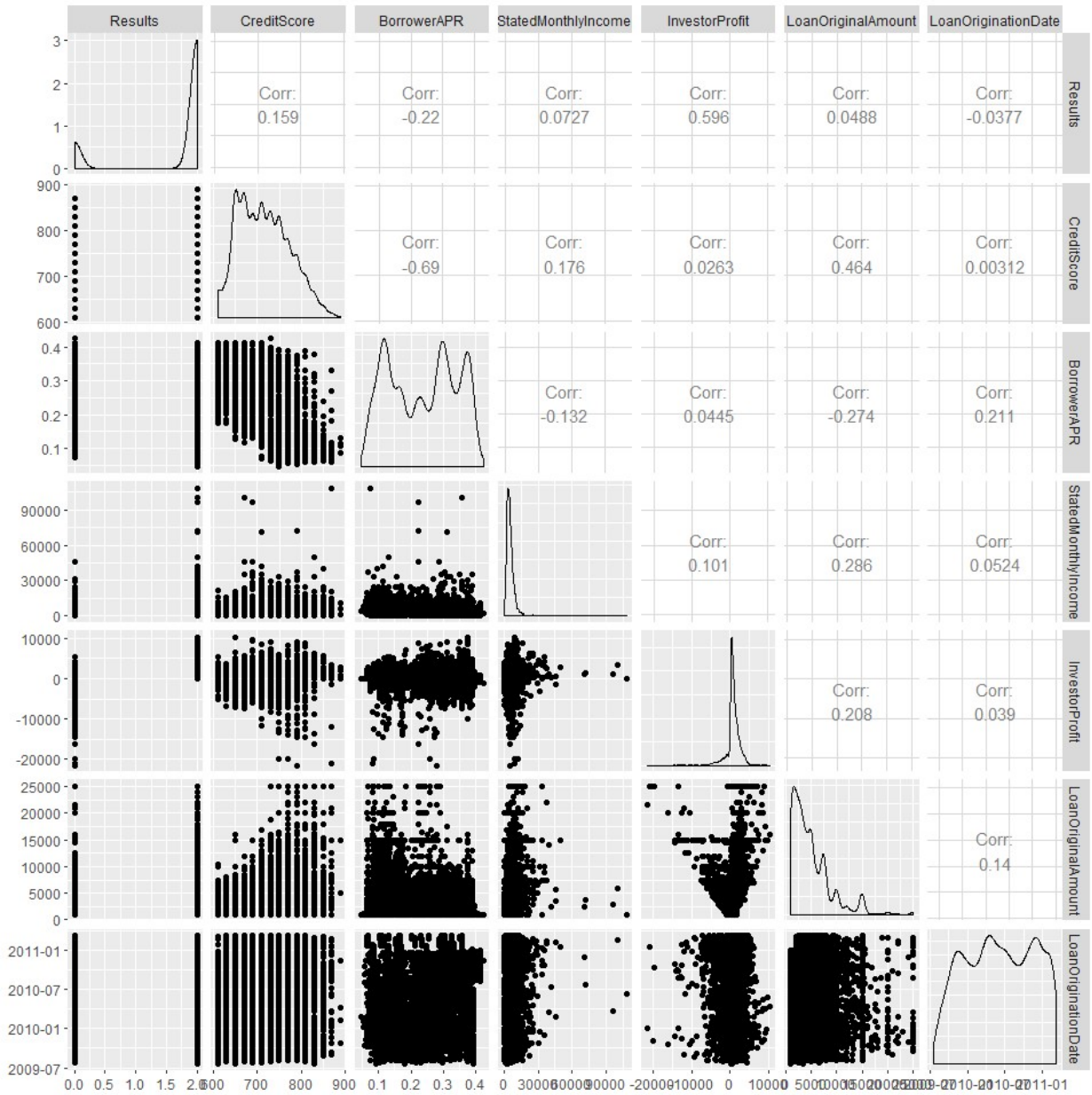
There is an average profit for investors, though not by much for the one year loans. There is also not much overlap between the two terms, only about six months. They look like two different populations. I think I'll just look at the 36 month loans. How many are there?

```
## [1] 8571   13
```



There is a regression line shows profit rate to be fairly constant over this time period.

Correlation matrix



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?

Investor profits depend on defaults (Results 0.596) and interest income(BorrowerAPR -0.22).

Loan defaults did not have a strong correlation with any of the expected variables. The largest correlations are with BorrowerAPR (-0.22) and CreditScore (-0.159).

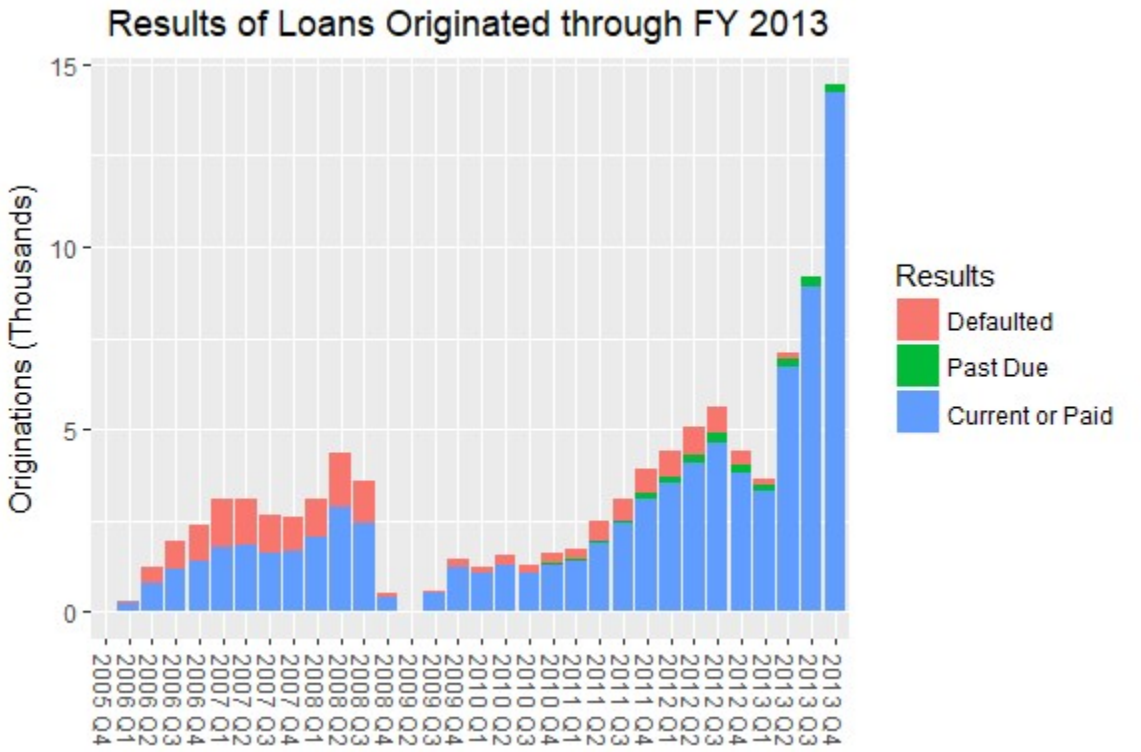
LoanOriginalAmount have a strong correlation with CreditScore (0.464), BorrowerAPR (-0.274), StatedMonthlyIncome (0.286) and InvestorProfit (0.208). #### Did you observe any interesting relationships between the other features (not the main feature(s) of interest)? StatedMonthlyIncome (0.286) was positively correlated with the loan amount.BorrowerAPR (0.211) was correlated with LoanOriginationDate, reflecting broader changes in interest rates.

What was the strongest relationship you found?

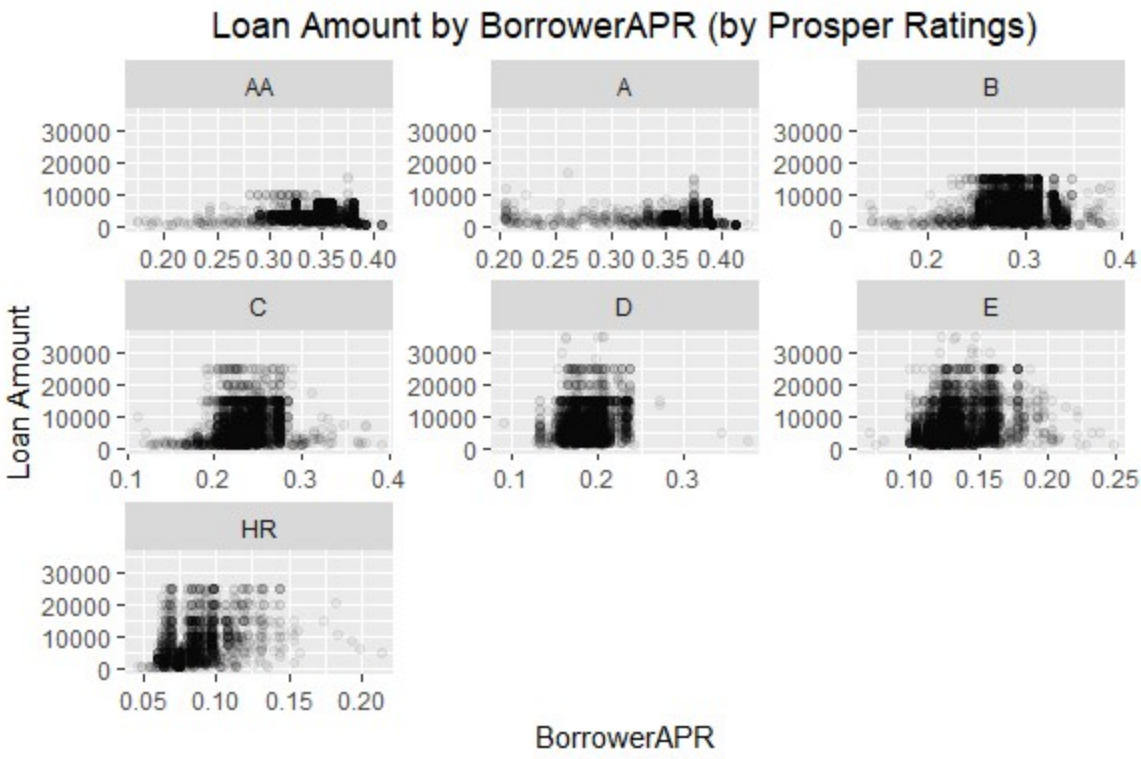
The strongest relationship among my variables was between CreditScore and BorrowerAPR (-0.69).

Multivariate Plots Section

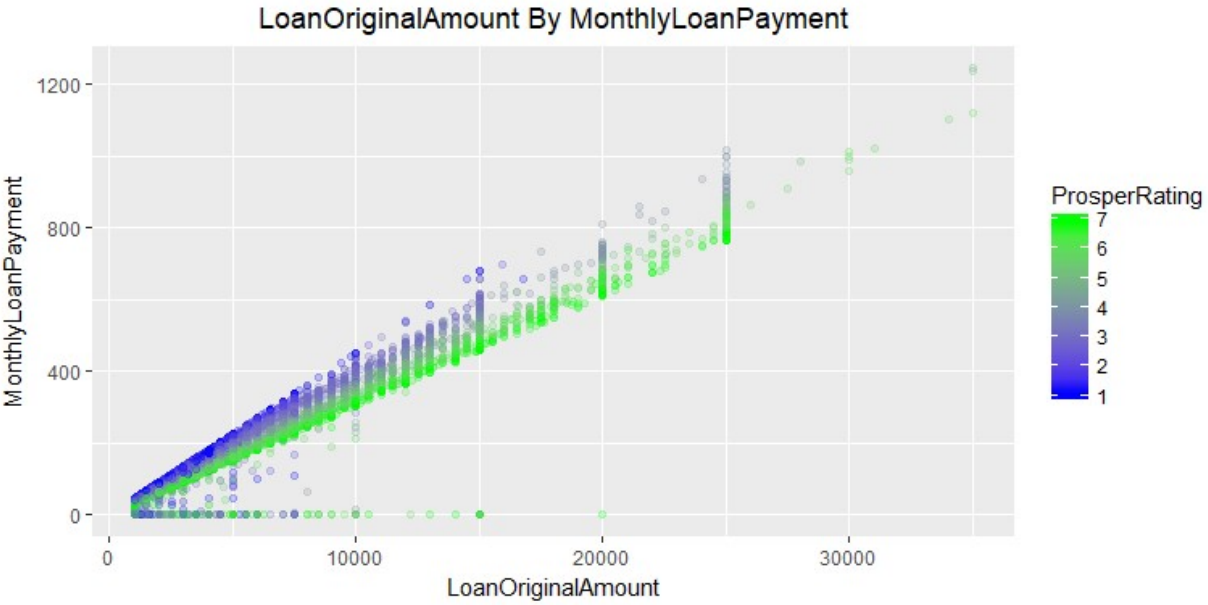
```
##               Cancelled             Chargedoff               Completed
##                       5                  11992                   38074
##                 Current              Defaulted FinalPaymentInProgress
##                   56576                   5018                     205
##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



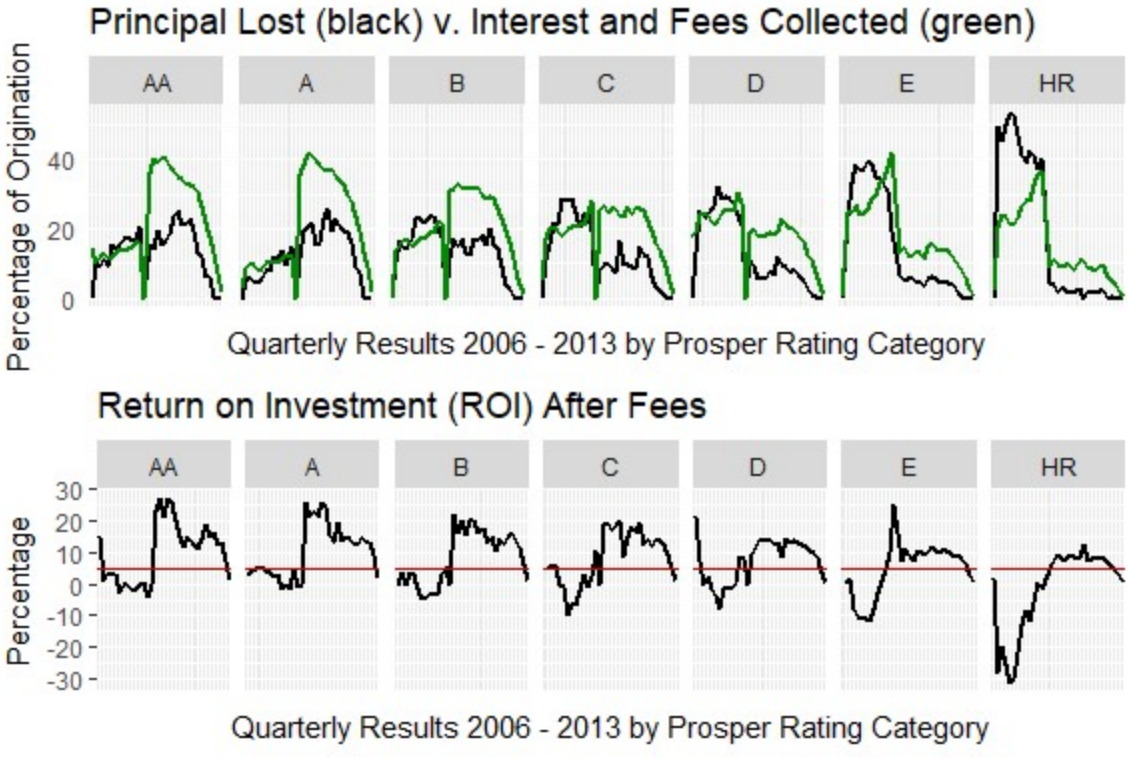
Default rates were high initially, but improved with the new standards implemented after the ‘quiet period’. Default rates drop in recent quarters because those loans have had less time to enter default. Proper models defaults with curves and notes that those recent loans have default rates below expectations.



Loans with higher prosper rating have lower Borrower Rate. Borrowers with lower Prosper Rating usually are borrowing smaller amounts with higher rate. For better visualization I've used only loans created after 7/1/2009 because Prosper Rating was not available for loans before that date.



We observe that the variance is explained by risk, which is represented by the ProsperScore. The bottom of the scatter plot is dominated by loans with a ProsoperScore equal to 11, which represents loans with low risks. The top of the scatter plot is dominated by loans with a ProsoperScore equal to 4, which represents loans with higher risks. Loans with a loan amount higher than $25000 are mostly dominated by a ProsperScore equal or superior to 6.



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?

- Lower borrower rates is usually related to higher credit score, higher prosper score, higher prosper rating and owning home.
- Borrowers with lower Prosper Rating usually are borrowing small amounts with very high rate.

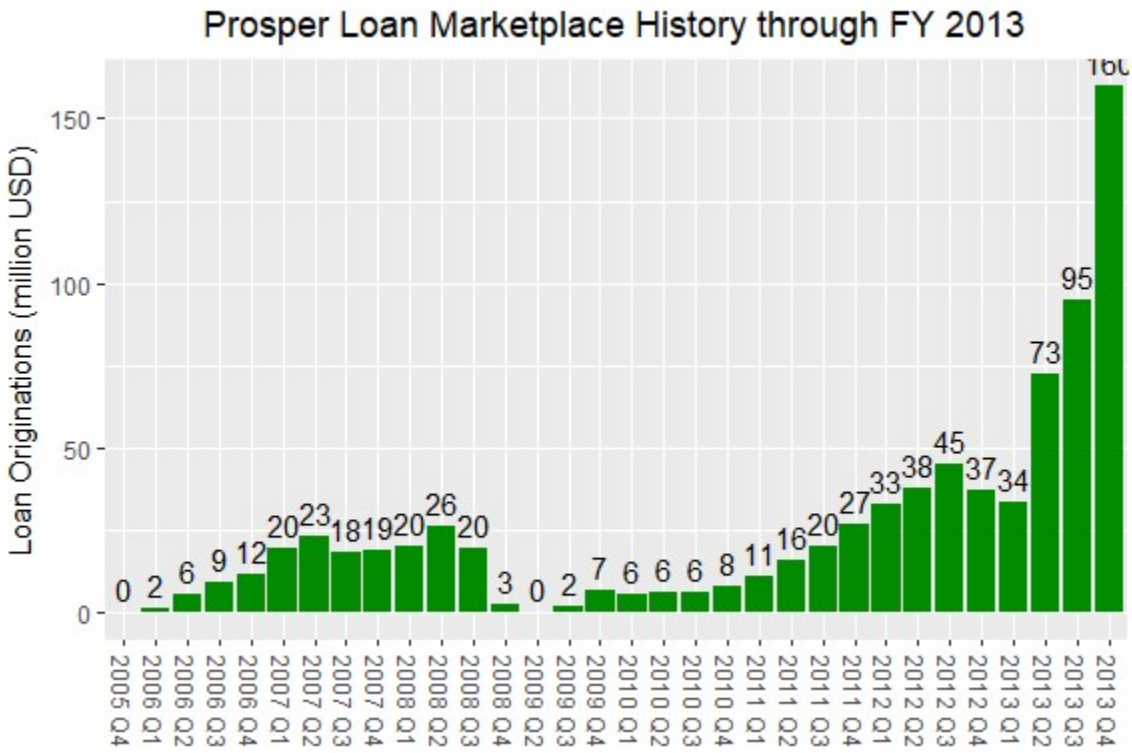
Were there any interesting or surprising interactions between features?

I found it interesting that larger loan requests were associated with higher credit scores.

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

Final Plots and Summary

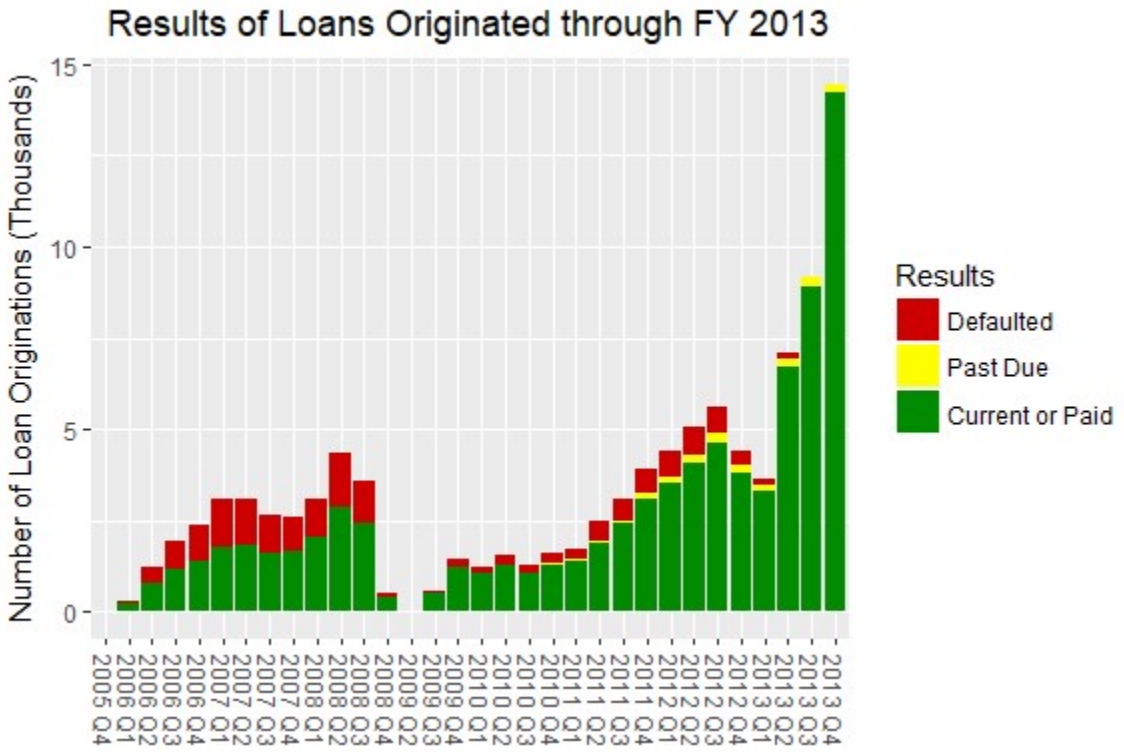
Plot One



Description One

Prosper Loan's business history is encoded the the dollar value of the loans originated through their online marketplace. Prosper was the first peer-to-peer lending marketplace, opening to the public February 5, 2006[1 (https://en.wikipedia.org/wiki/Prosper_Marketplace)]. Initially, lenders bid on loans by offering competing interest rates. Prosper's business model came under scrutiny by the US Securities and Exchange Commission, who issued a "cease and desist" letter November 24, 2008.[2 (https://www.sec.gov/litigation/admin/2008/33-8984.pdf)] In anticipation, Prosper filed for SEC registration, which required a "quiet period" from October 15, 2008 until July 13, 2009, during which time, no new loans were originated.[3 (http://www.lendacademy.com/a-look-back-at-the-lending-club-and-prosper-quiet-periods/)] Prosper attributes the decrease in originations at the end of 2012 to a decrease in liquidity and in January of 2013 undertook an equity financing [4 (https://www.prosper.com/Downloads/Legal/prosper10k12312013.pdf), p 74]. The increase in capital was used in part for a marketing campaign to attract more borrowers and to launch IRA accounts to attract institutional lenders.

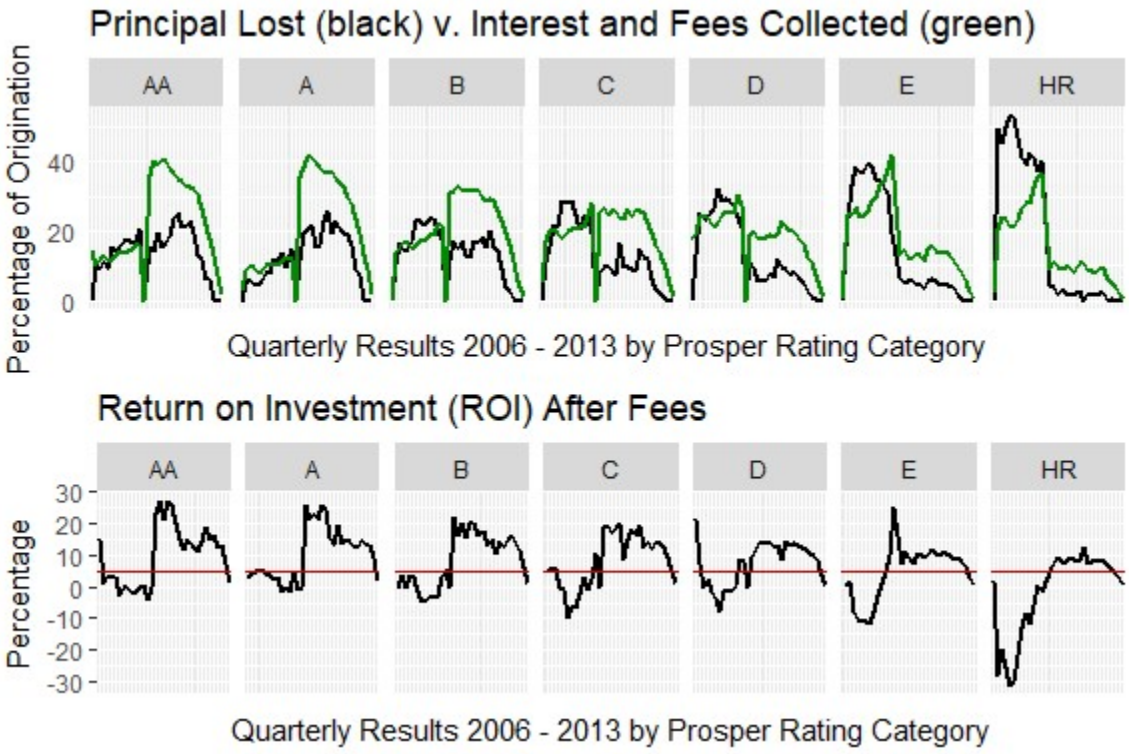
Plot Two



Description Two

In this plot, we switch from dollar amounts to number of new loans originated each quarter and the final disposition of those loans. The early days of Prosper were marked by very loose lending standards. Coupled with the global financial crisis, these early loans had very high default rates and many investors realized losses. After Prosper's relaunch in 2009, minimum credit scores were increased and Prosper made more of an effort to verify borrower's information[5 (http://www.wsj.com/articles/SB120525138644627455)]. Prosper's prospectus makes it clear that investors should expect some loans to default[6 (https://www.prosper.com/invest/marketplace-performance/)], and charges interest rates high enough to account for risk, but lower than a borrower would get from a credit card.

Plot Three



Description Three

Although every rating category have defaults, investors still make money by collecting more (on average) in interest and fees than the principal lost to defaulting borrowers. Here, principal lost, service fees, and collection fees are subtracted from the interest and (borrower) fees paid to investors. All rating categories have generated impressive profits since 2009 Q4, with generally higher volatility in riskier categories.

Reflection

Before this project I didn't know anything about R, ggplot, Prosper peer-to-peer lending business. But I selected the most difficult looking dataset and want challenge myself.

I tried to begin by read all of Prosper's website and their annual report, analyzing Prosper's business model. I divided the business model into two domain. Domain one is company domain (one role: Prosper), Domain two is customer domain (two role: Investor and borrower), Then classify the attributes of the dataset according to the three roles. And it started to tell a coherent story. The variable list made a lot more sense, and I could see what information wasn't included for public release. At last, I can explore. I chose these three plots because together they tell the history of Prosper Loan and how it works.

Through this project practice, I realized this truth, the third leg of data science is substantive experience. For the future work of the project, I would really like to make a predictive model to compare against the FY 2014 to FY 2017 data. I could treat loan result as label and train the model with selected features (CreditScore, Rating and BorrowerAPR, etc). Then use the trained model to predict the result of new loans. I am looking forward to learning more about modeling and predictions in coming the Machine Learning class.