Prosper Loan Investment Strategy by Mike Chan

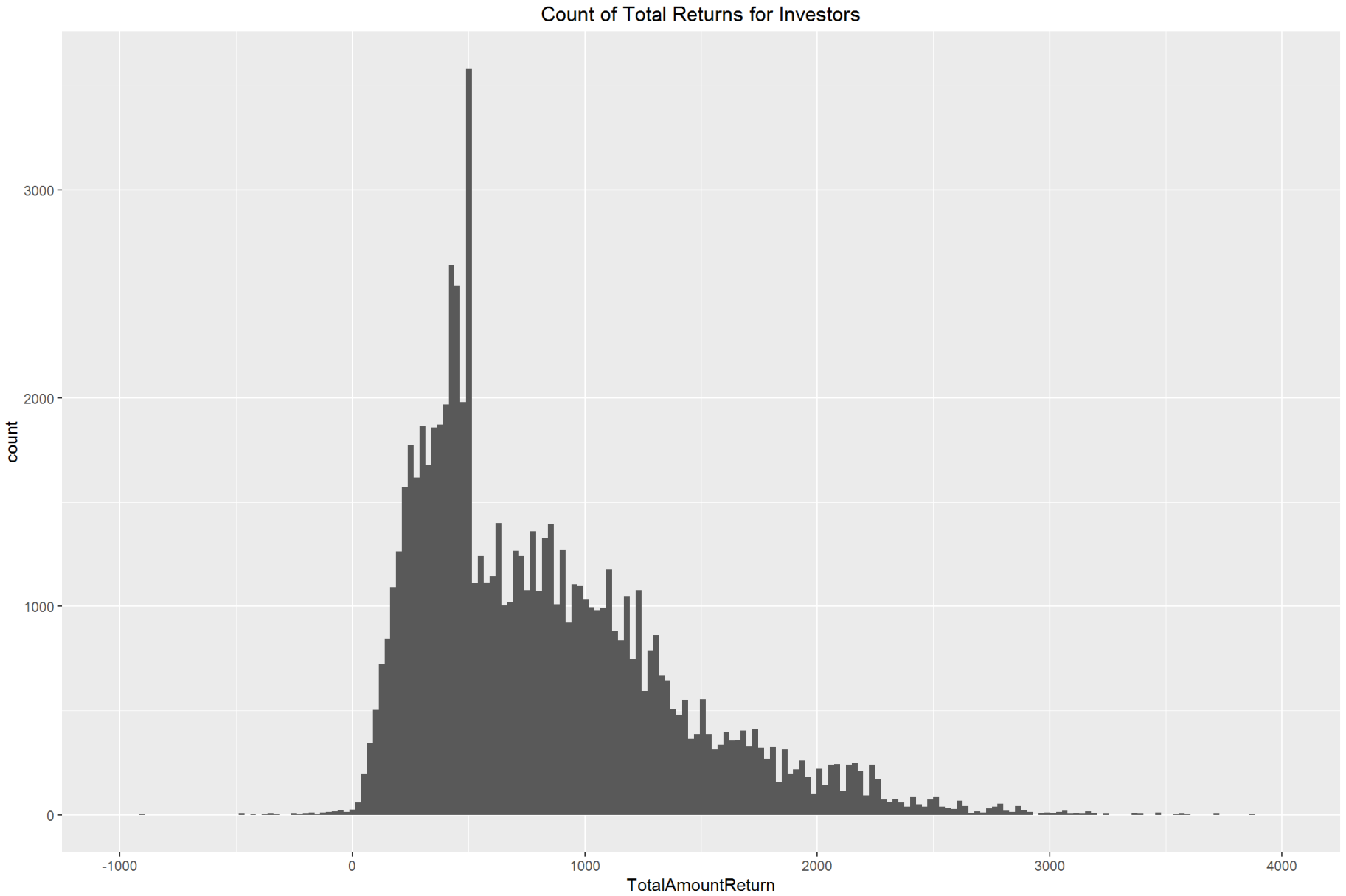
Univariate Plots Section

I would like to display the data of all my data set and what it is like…

```
##       Term                      LoanStatus     BorrowerAPR
##  Min.   :12.00   Current              :52478   Min.   :0.04583
##  1st Qu.:36.00   Completed            :17703   1st Qu.:0.16175
##  Median :36.00   Chargedoff           : 4445   Median :0.21566
##  Mean   :42.72   Defaulted            :  885   Mean   :0.22373
##  3rd Qu.:60.00   Past Due (1-15 days) :  722   3rd Qu.:0.28780
##  Max.   :60.00   Past Due (31-60 days):  327   Max.   :0.42395
##                  (Other)              :  997
##  EstimatedLoss    EstimatedReturn    ProsperRating CreditScoreRangeLower
##  Min.   :0.0049   Min.   :-0.18160   AA: 5151      Min.   :600
##  1st Qu.:0.0424   1st Qu.: 0.07408   A :13705      1st Qu.:660
##  Median :0.0699   Median : 0.09060   B :14631      Median :700
##  Mean   :0.0786   Mean   : 0.09529   C :16890      Mean   :699
##  3rd Qu.:0.1080   3rd Qu.: 0.11500   D :12824      3rd Qu.:720
##  Max.   :0.3660   Max.   : 0.26670   E : 8617      Max.   :880
##                                      HR: 5739
##  DebtToIncomeRatio         IncomeRange    LoanOriginalAmount
##  Min.   : 0.0000   $0             :    0   Min.   : 1000
##  1st Qu.: 0.1500   $1-24,999      : 3840   1st Qu.: 4000
##  Median : 0.2200   $25,000-49,999:22023   Median : 8000
##  Mean   : 0.2588   $50,000-74,999:24030   Mean   : 9269
##  3rd Qu.: 0.3200   $75,000-99,999:13644   3rd Qu.:14500
##  Max.   :10.0100   $100,000+      :14019   Max.   :35000
##                    NA's           :    1
##  TotalAmountReturn
##  Min.   :-1052.2
##  1st Qu.:  414.6
##  Median :  702.5
##  Mean   :  828.6
##  3rd Qu.: 1133.0
##  Max.   : 4117.5
##
```

Comment: General idea of what to expect of the data set

I would like to see how much investors have earned for using the Prosper Loan system…

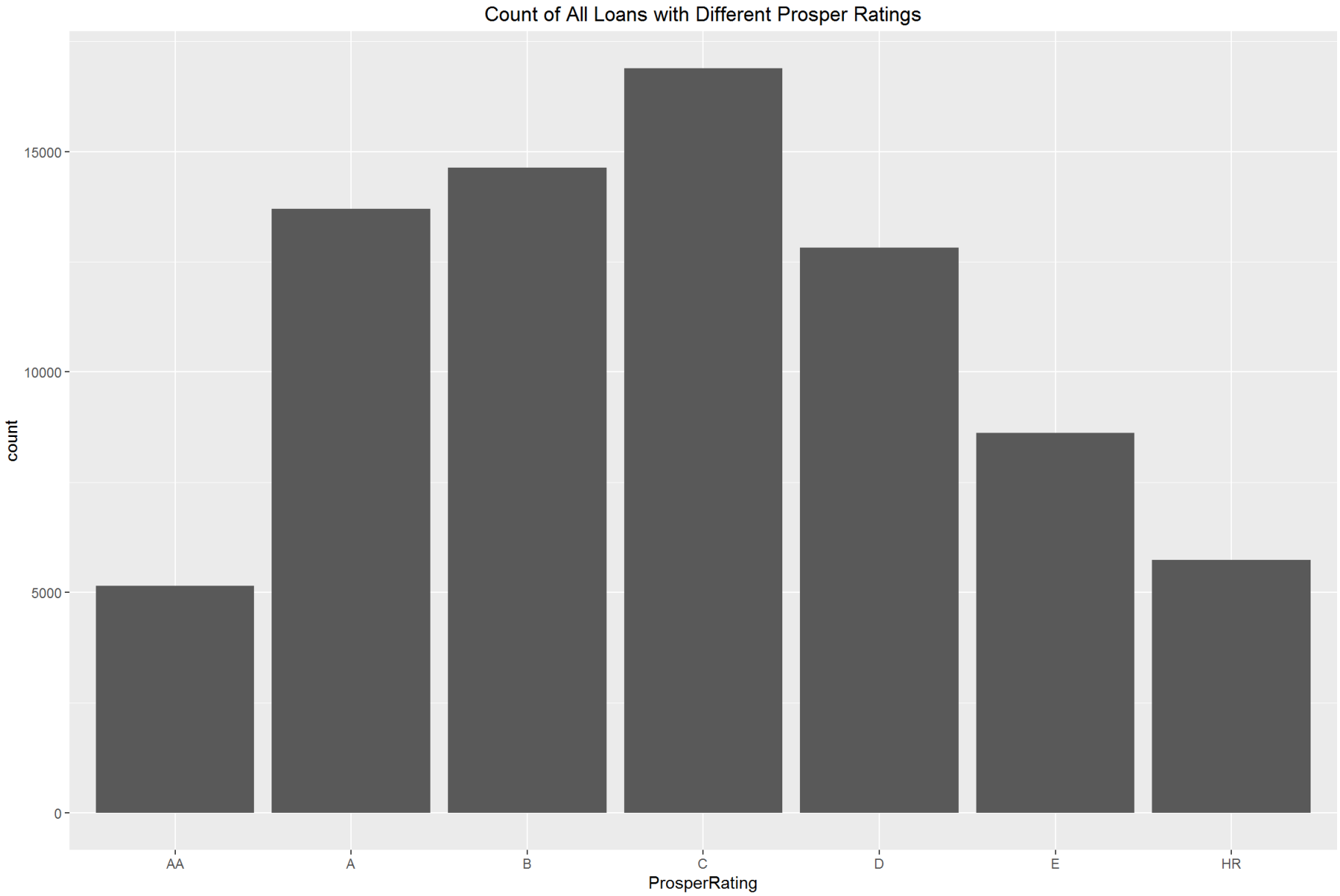


Comment: Investors have nearly received on average, $828 in return per loan

Investors can possibly earn as high as $4,000. Interestingly, I'm seeing a large spike of investors earning $500.

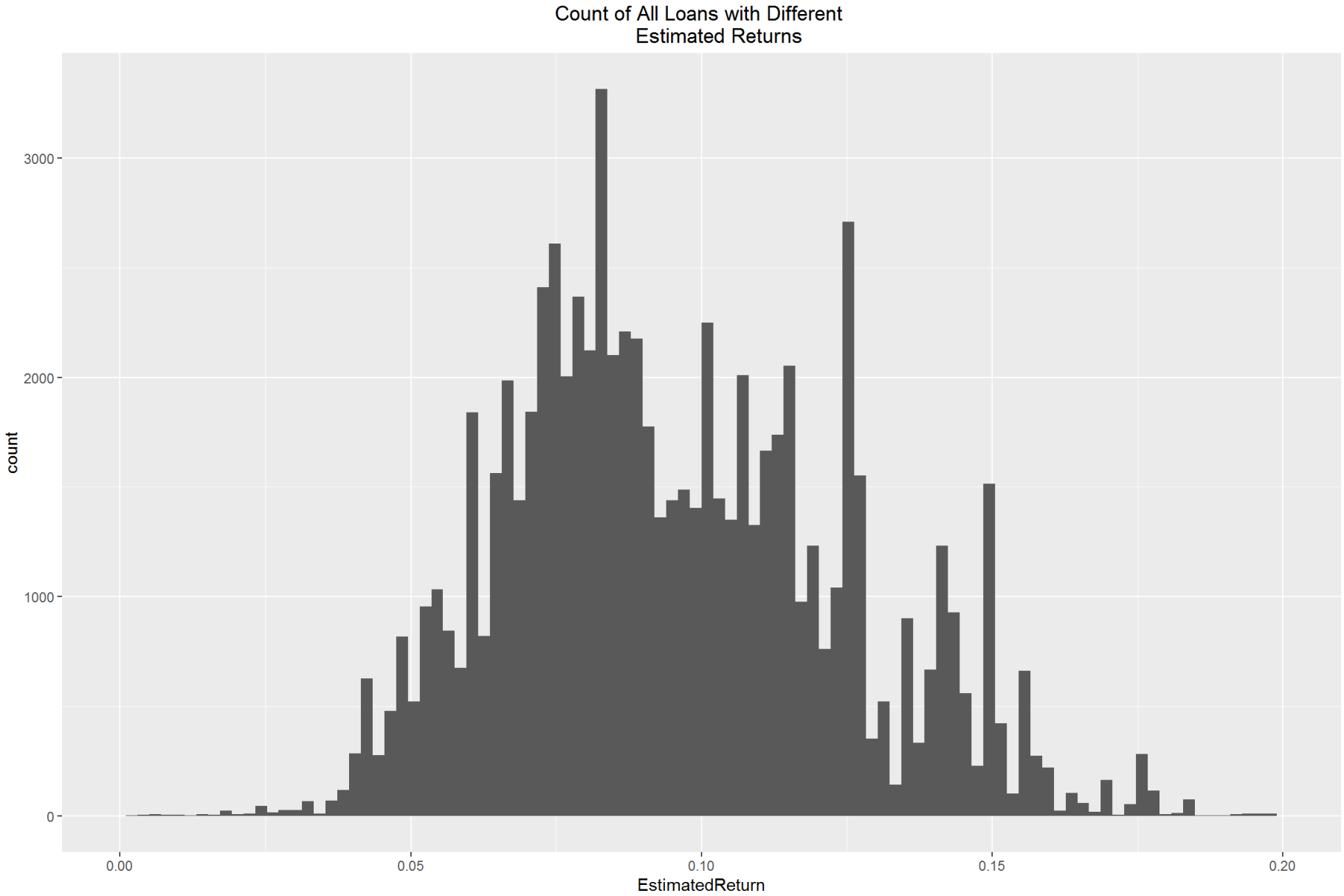
Notice though, that investors still can lose money as we see negative returns.

I would like to show the count of all loans of different prosper ratings…



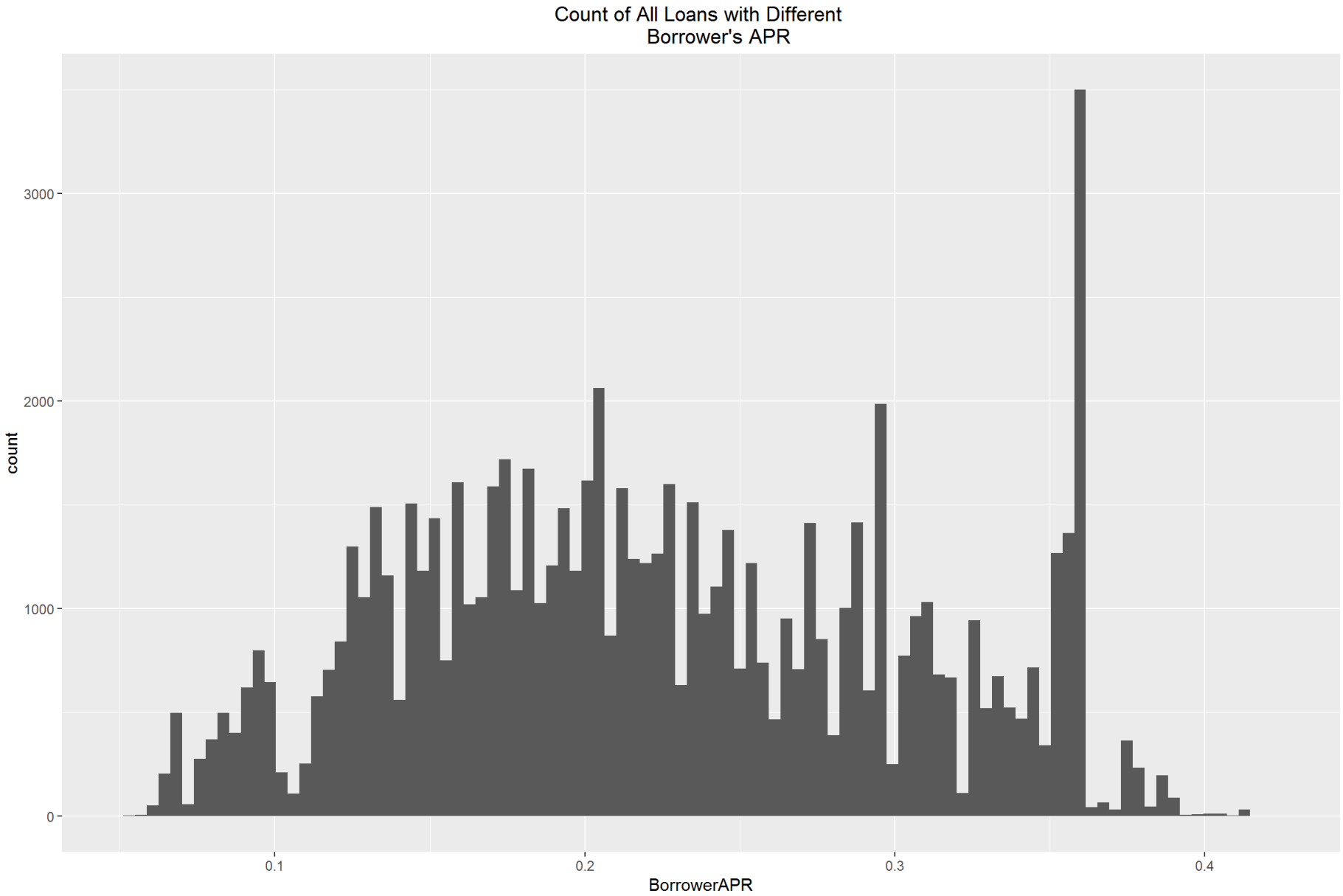
Comment: Most loans have a prosper rating of C

I would like a count of all loans with different estimated returns…



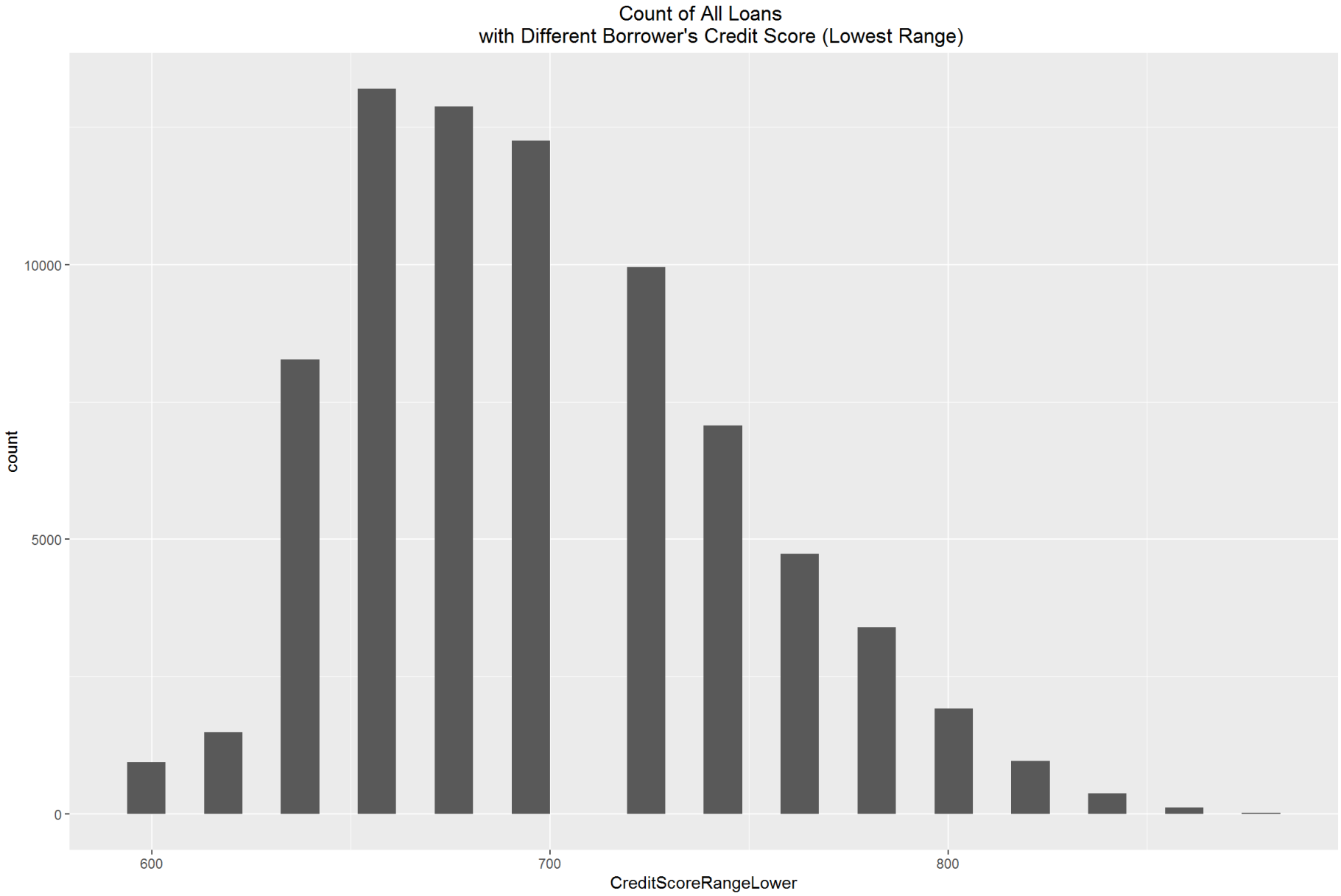
Comment: The average estimated return looks to be around 0.09

I would like a count of al loans with different numbers of borrower APRs…



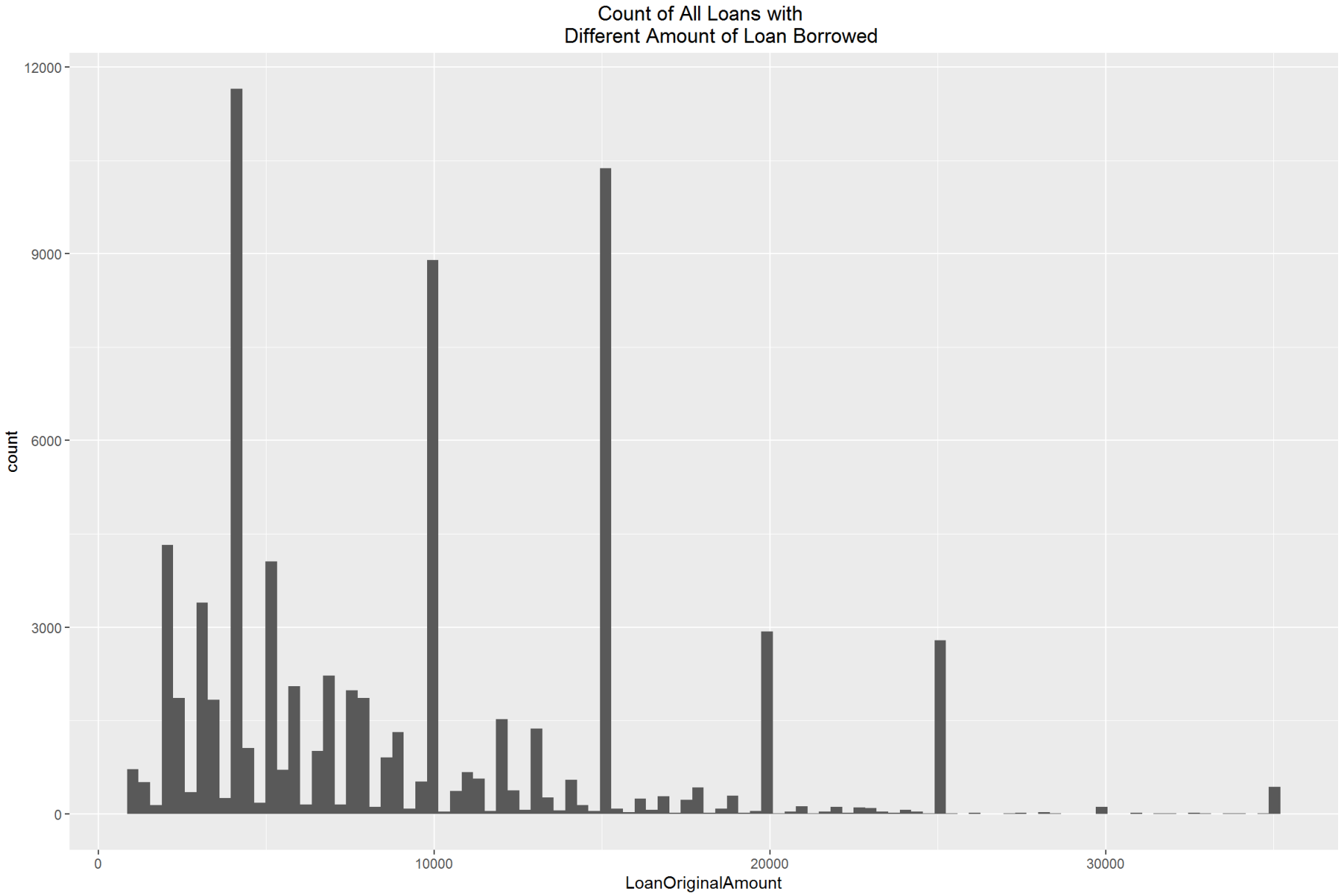
Comment: The average normal distribution looks to be at mean of 0.2.
There is however a sudden big spike at around 0.36 APR.

I would like a count of all loans with different borrower's credit score..



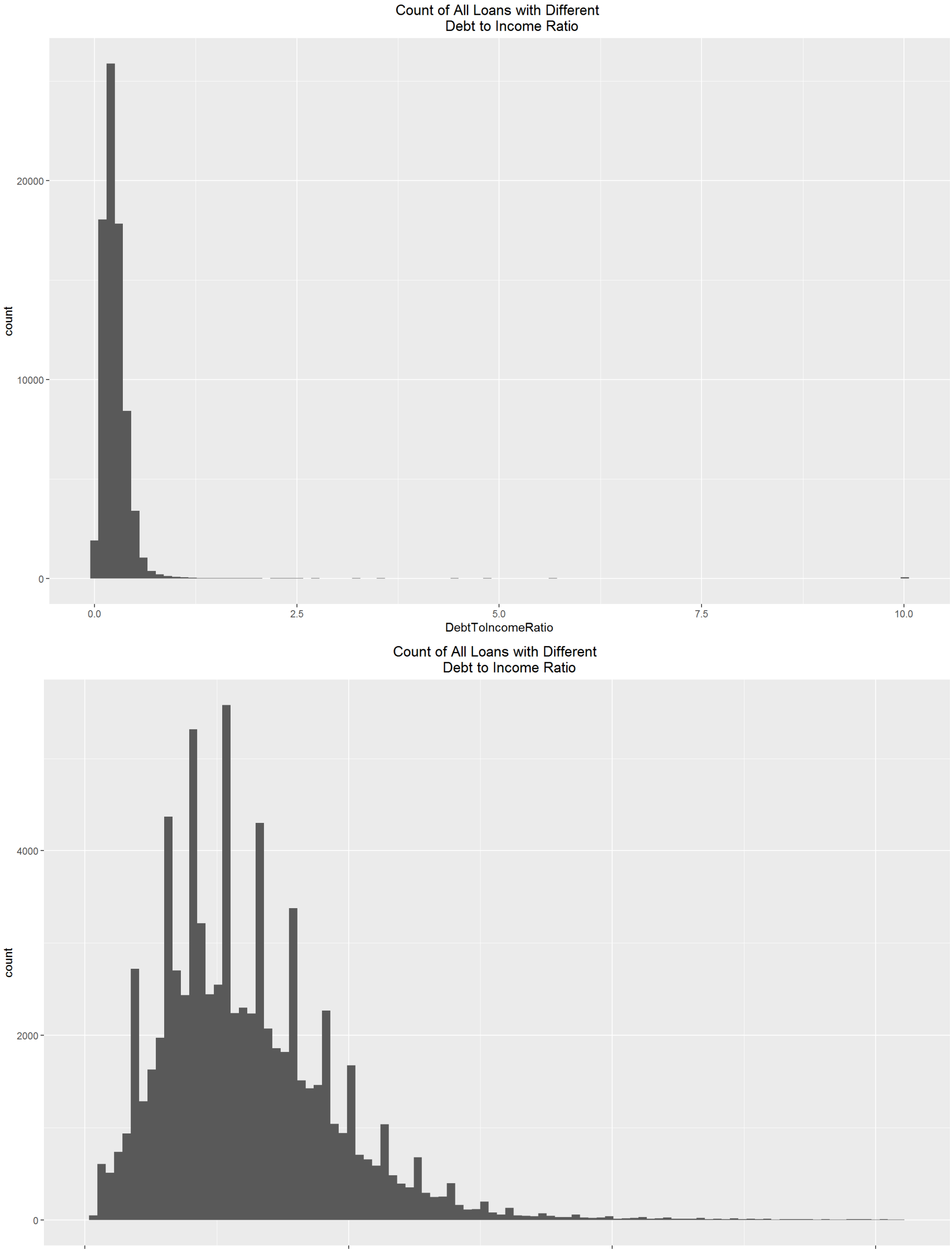
Comment: The avearge Credit Score Range is about 675.

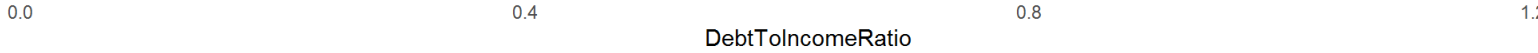
I would like a count of loan of all different amounts borrowed…



Comment: The graph looks skewed to the left. So the average amount borrowers borrow is $8,000. I notice that borrows tend to borrow in sets of 500s and 1000s because we can see big spikes at 4,000, 5000, 10000, 15000, 20000, etc.

I would like a count of all loans with different debt to income ratio…

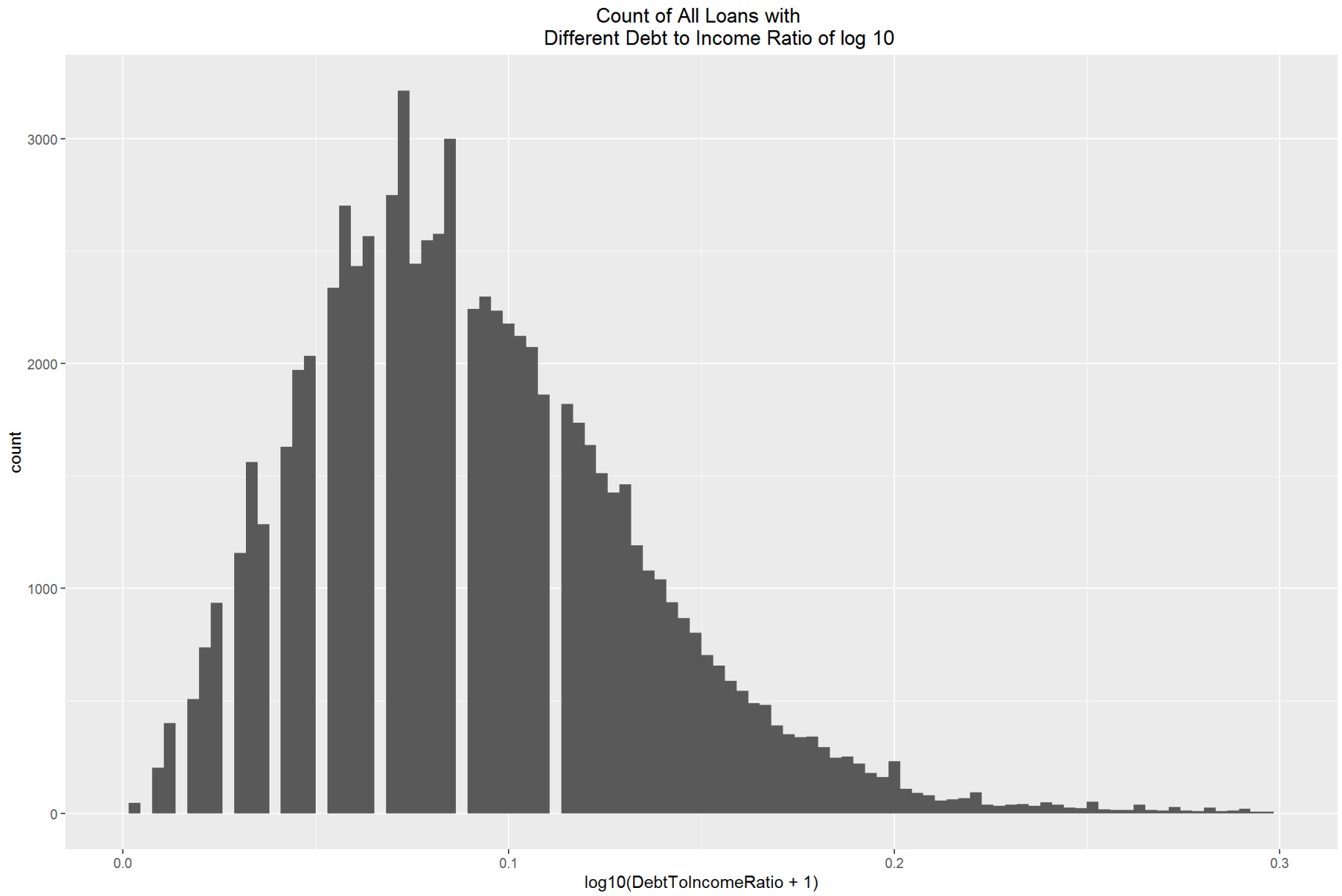




Comment: First, I want to point out the big outlier at the Debt to Income Ratio of 10, and anything in between around 1.2 to 10 looks to absolutely nothing.
I then removed the bias by changing my x limit to between 0 and 1.25.
I see that the average Debt to Income Ratio is about 0.2 . I also notice the large spikes in which my best guess is that borrowers just put in random numbers like 0.1, 0.2, etc.

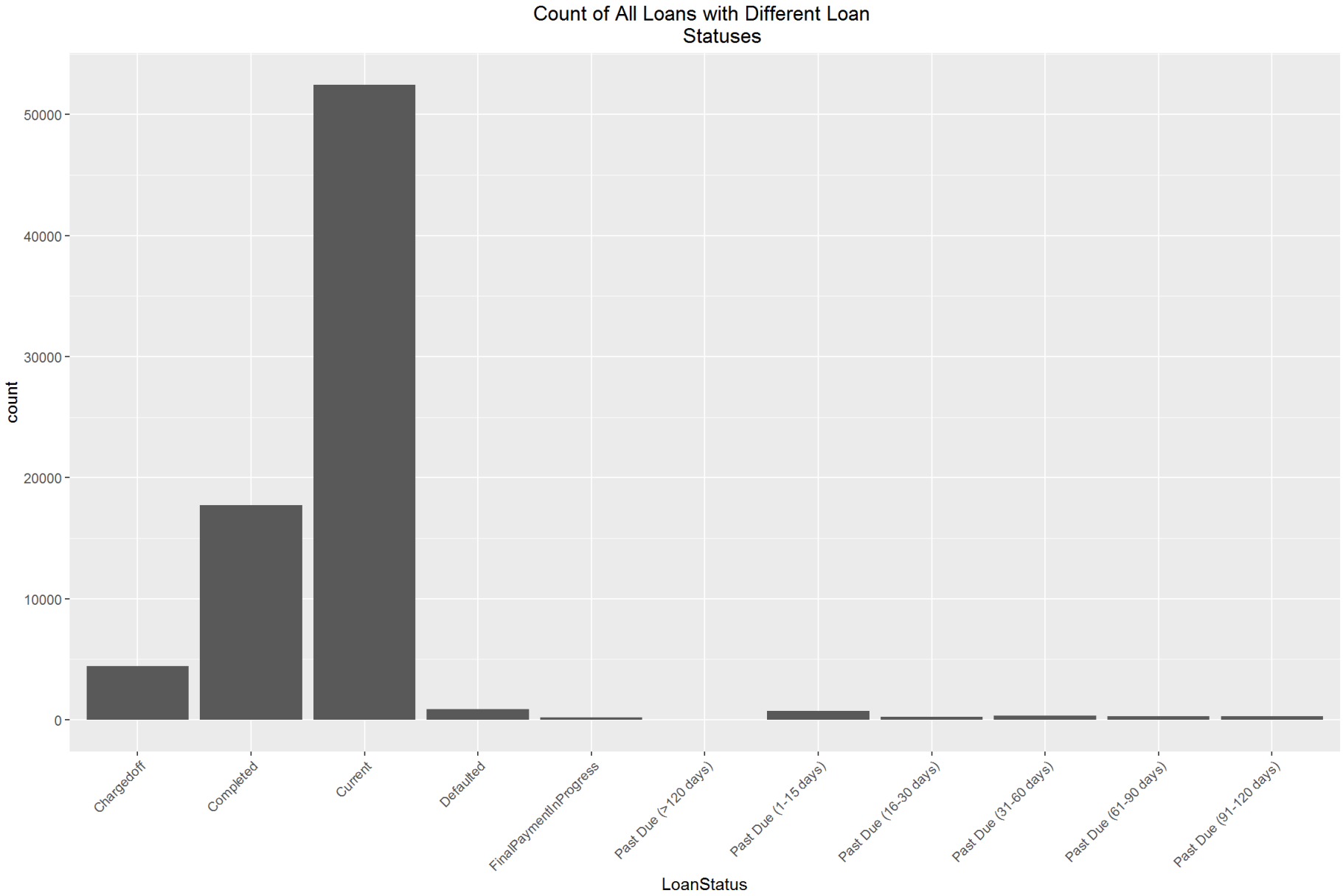
In the next graph we will use log transform to see it more normally distributed

I would like a count of all loans with different debt to income ratio of log 10 because it looked very skewed before…



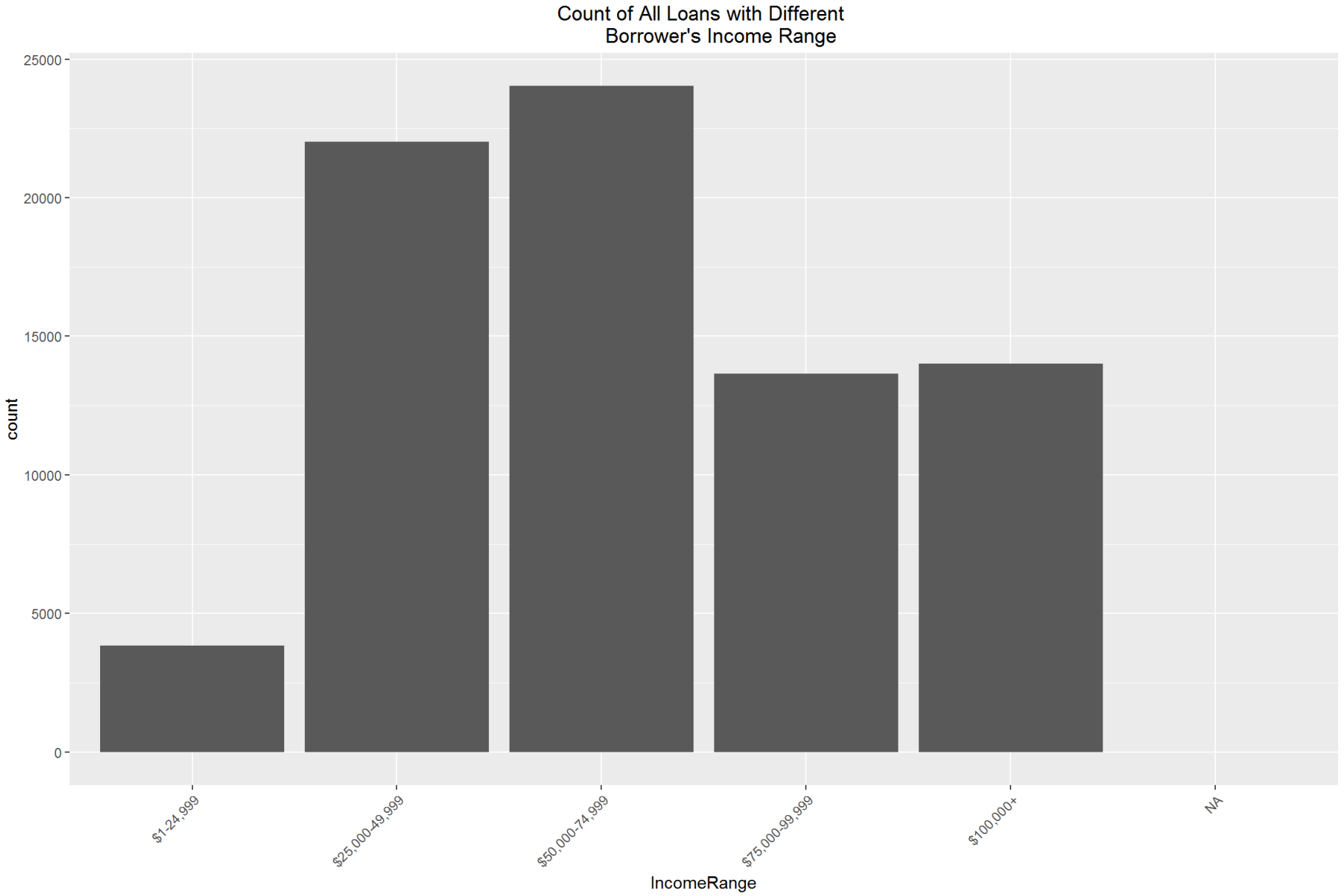
Comment: The average debt to income ratio appears to be around log of 0.09 but as the ratio becomes larger, the value of that ratio becomes more continuous (Left half of the graph has more break lines than right half).

I would like a count of all loans with different loan statuses…



Comment: Most loans are currently going on and a majority of them are completed.

I would like a count of all loans with different borrower's income range…



Comment: Most borrowers are between the income range of $50,000 and $74,999 I was hoping to see borrowers of income between $1 and $24,999 to be the highest because, they need to borrow more.

I would like a count of all loans with different terms

```
## Source: local data frame [3 x 2]
##
##    Term     n
##   (int) (int)
## 1    12  1414
## 2    36 53000
## 3    60 23143
```

Comment: A lot of people borrow a loan for 36 months.

Univariate Analysis

What is the structure of your dataset?

There are 11,3937 prosper loans with 10 different features.

The PowerPoint Slide "Explanation of the Data Set" explains why I chose the features and explaining in depth.

I stated what I found through different codes.

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

The main features in the data set are

EstimatedReturn: Because as Investors, we want to use the data to find the best return for our investment.

ProsperRating: And to determine what may be the way to invest for a return on our loans, the risk of a loan may have the most influence on a return out of all features.

What other features in the dataset do you think will help support your

investigation into your feature(s) of interest?

All the 10 features I picked of the 81 different features on slide “Explanation of the Data Set” will support my investigation.
I explain in detail why I chose those features.

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

I created the total amount returned variable (“TotalAmountReturn”) using the code:

ld$TotalAmountReturn <- with(ld,EstimatedReturn*LoanOriginalAmount)

This function helps us to determine what an investor can expect to be returned from investing in that loan. We’re not only interested in percentage but the amount investors will receive.

Of the features you investigated, were there any unusual distributions?

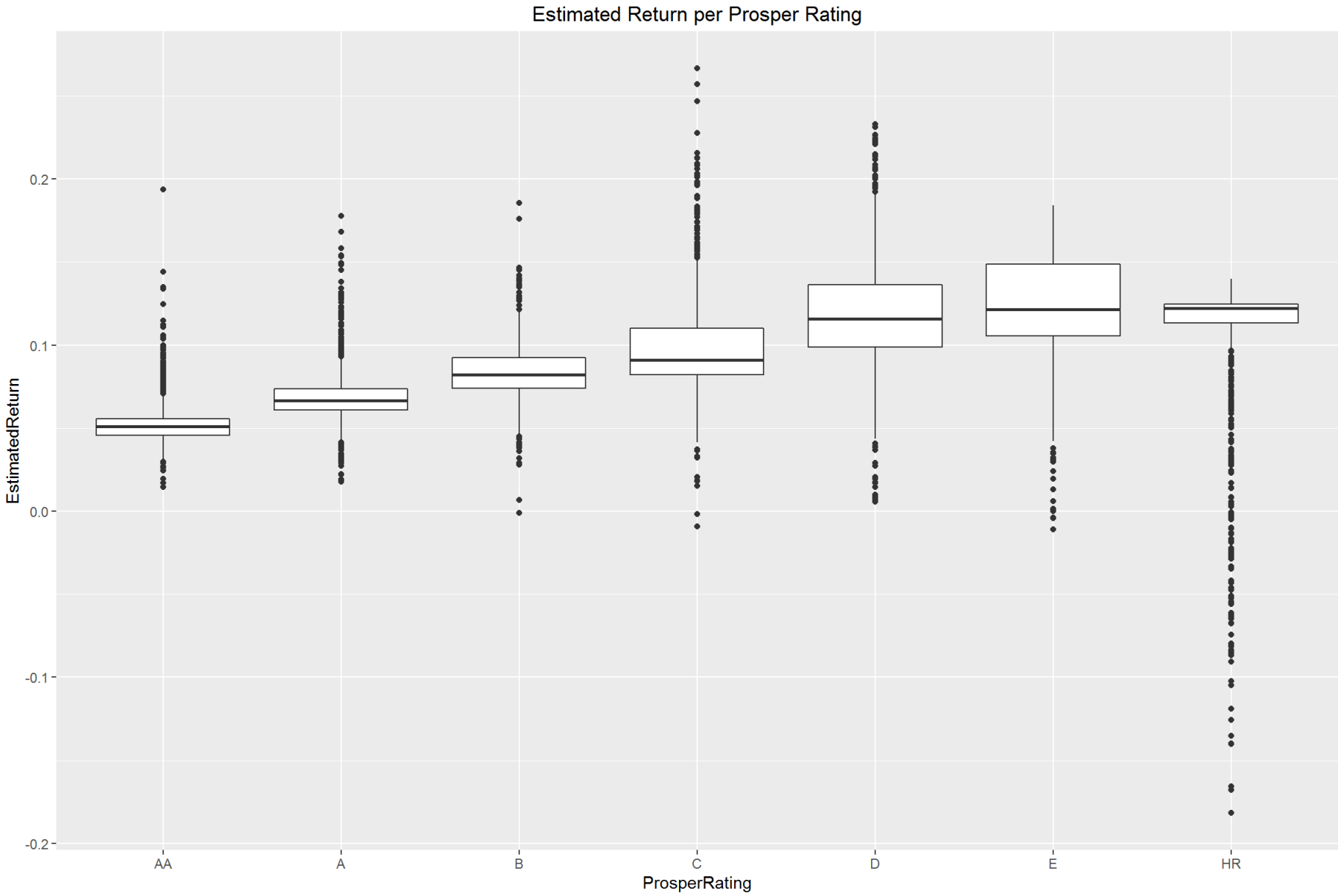
Did you perform any operations on the data to tidy, adjust, or change the

form of the data? If so, why did you do this?

Yes, the count for each Debt To Income ratio was greatly skewed to the left and I wanted to make it more normally distributed.

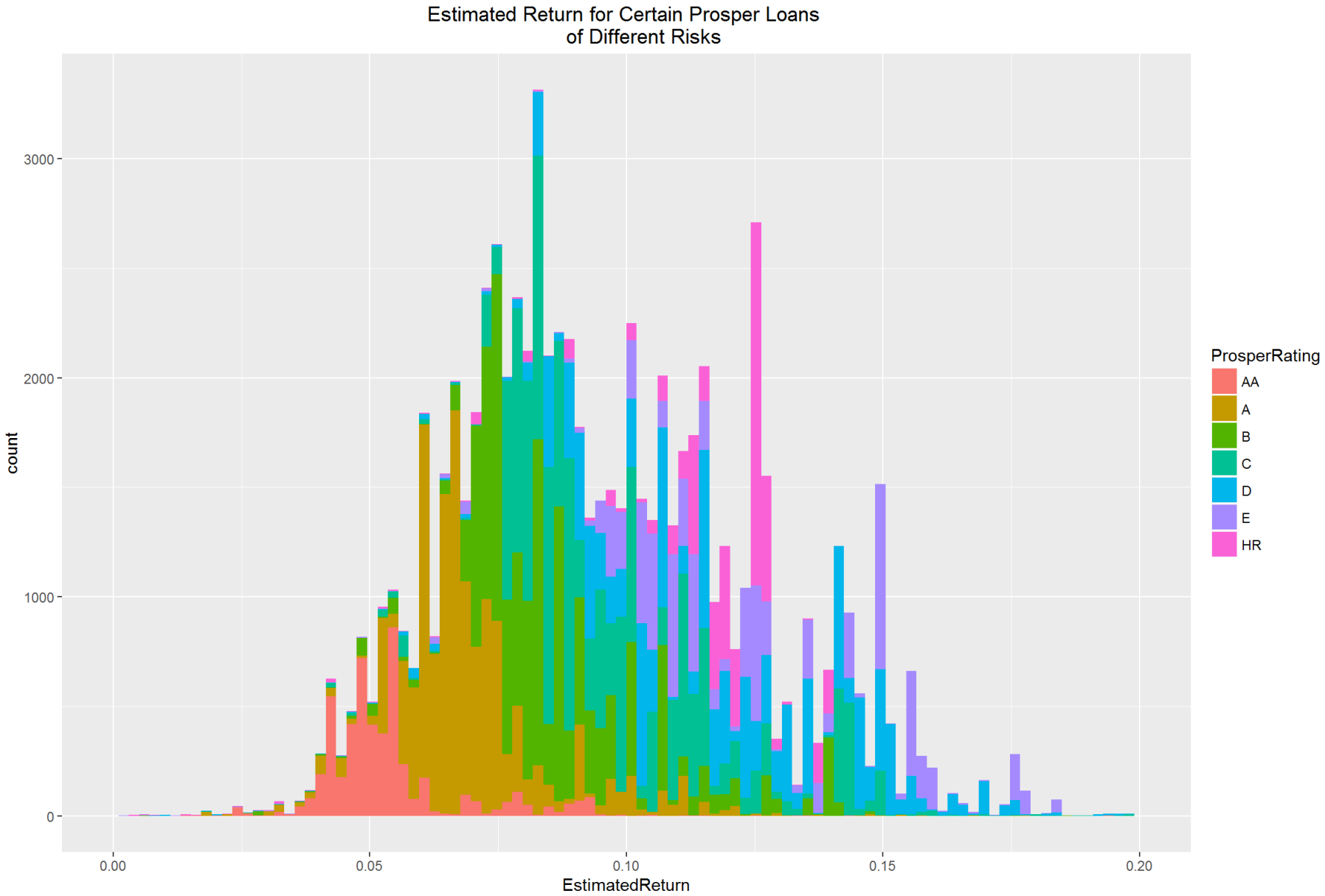
Bivariate Plots Section

I would like to see a box plot of all estimated return for each loans of different prosper ratings…



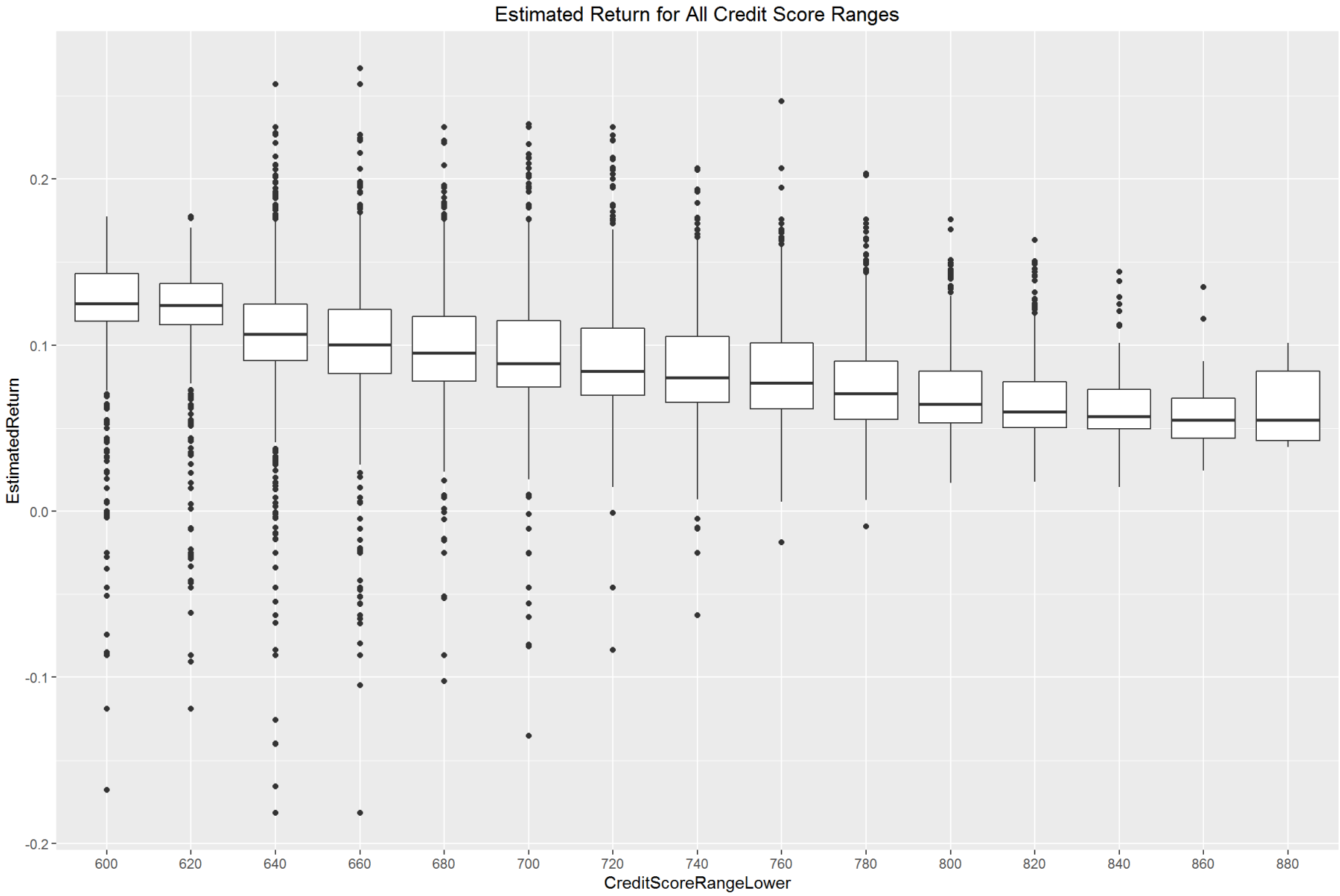
Comment:The average estimated return appears to increase as the Prosper Rating Risk increases. However, note the tails from one end changes amongst different risks. For example, we see more tails in “A” and they are above the 3rd Quantile and we see more tails in “HR” much below the 1st Quantile.

I would like a count of all estimated returns of differnet prosper ratings…



Comment:That we see a pattern. As the higher higher the estimated return, the higher the prosper rating risk showing up.

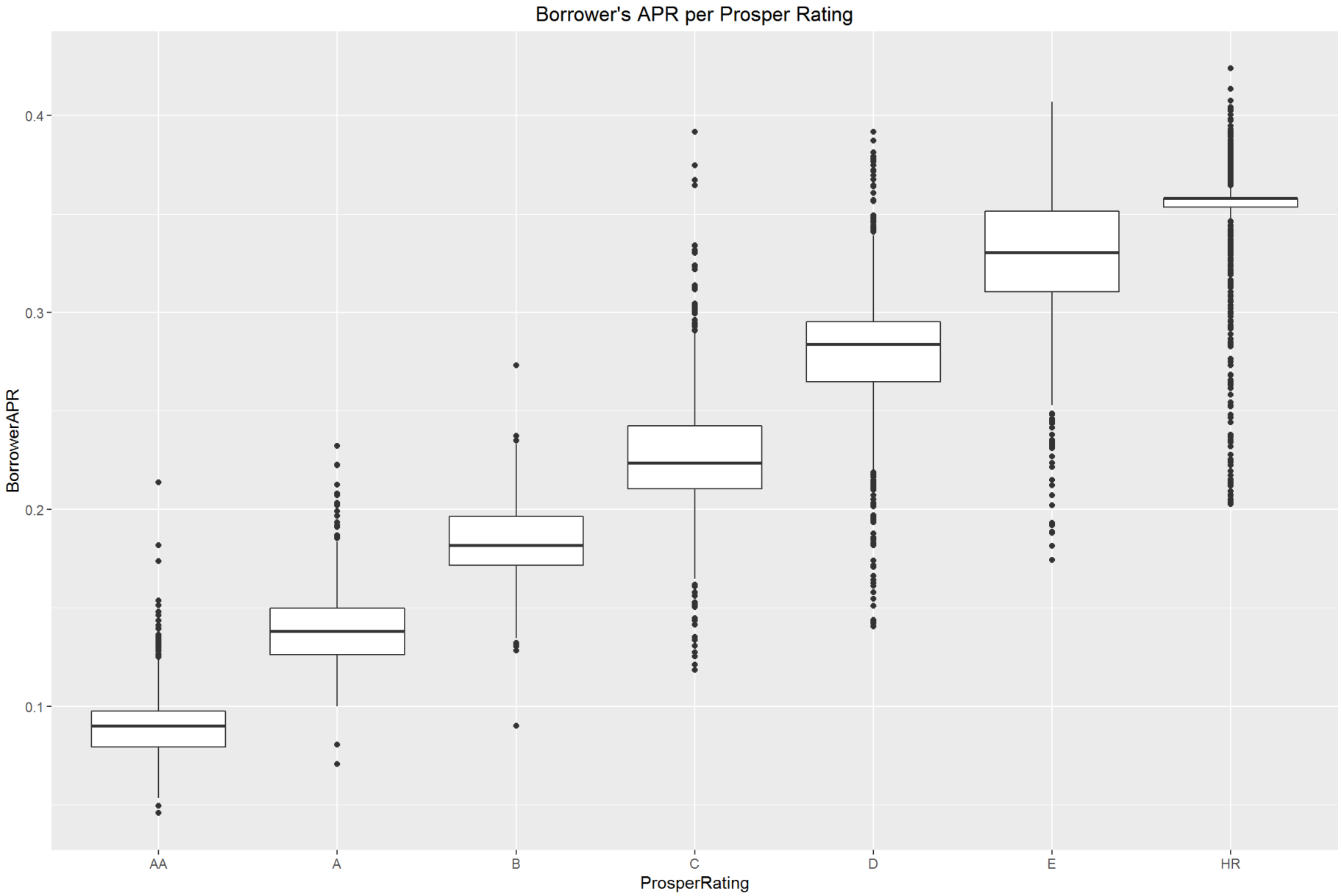
I would like to see the estimated returns of all borrowers of different credit score ranges…



Comment: The investor's estimated return decreases as borrower's credit score increases. The Credit Score Range also follows the Prosper Rating.

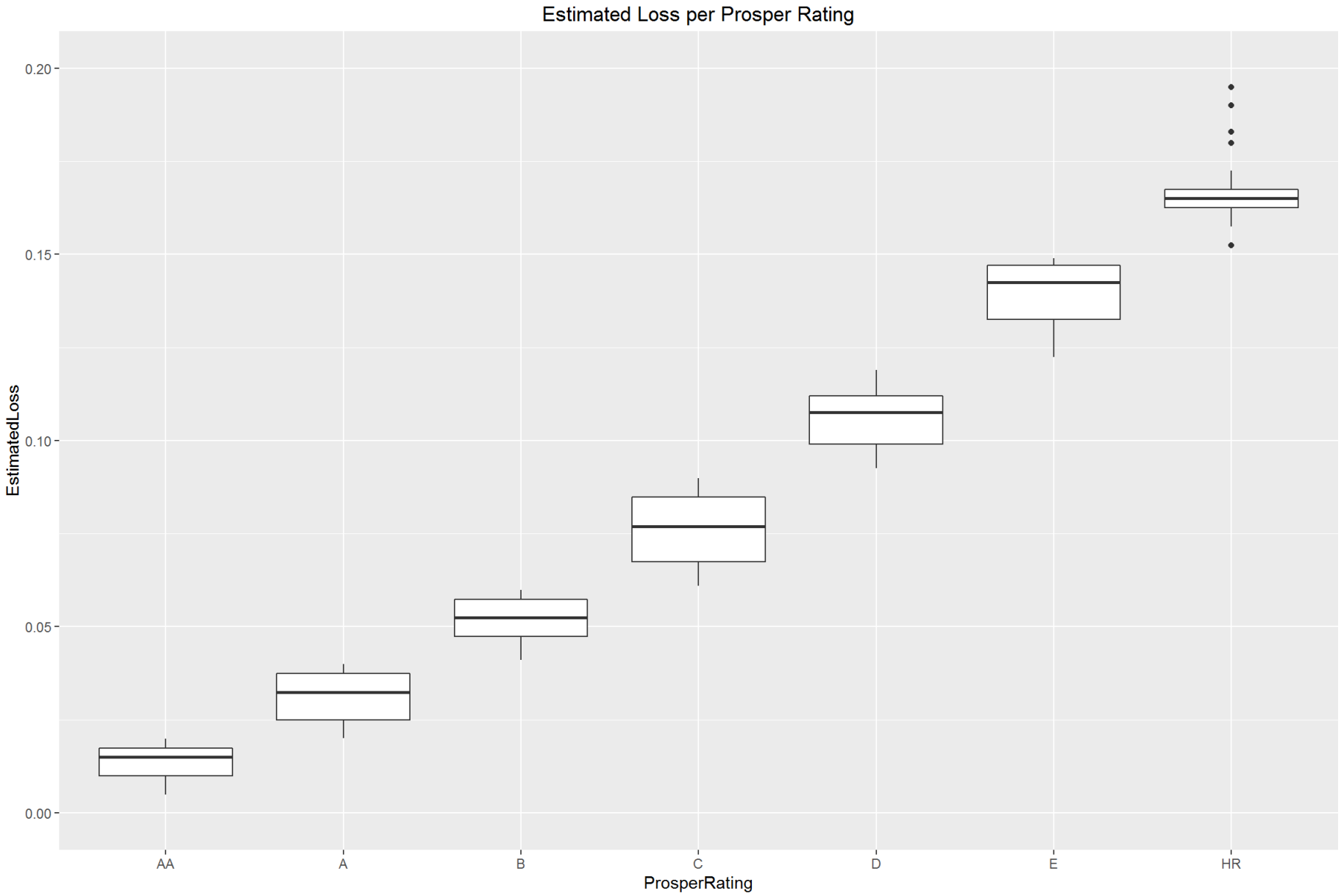
The lower the Credit Score, the higher the prosper rating score risk tends to be.

I would like to see how the typical borrower's APR for each loan



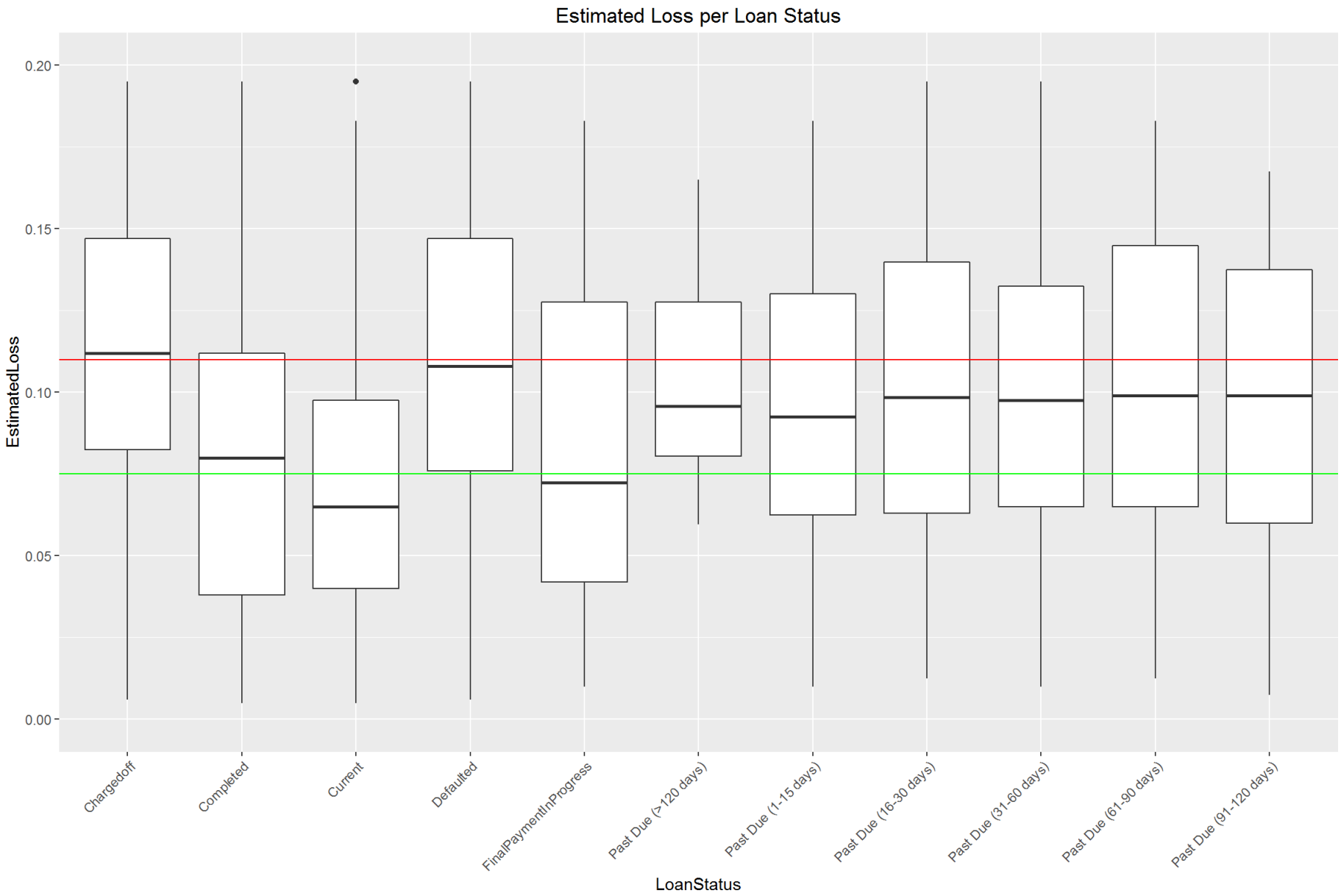
Comment:This box plot is a decision making tool for investors to invest in a certain APR in each Prosper Rating for greater gains. For example, in "C" rated loans, investor should only invest in those loans when Borrower APR is between 0.21 and 0.24.

I would like to see a boxplot of different estimated losses per each loan of different prosper ratings…



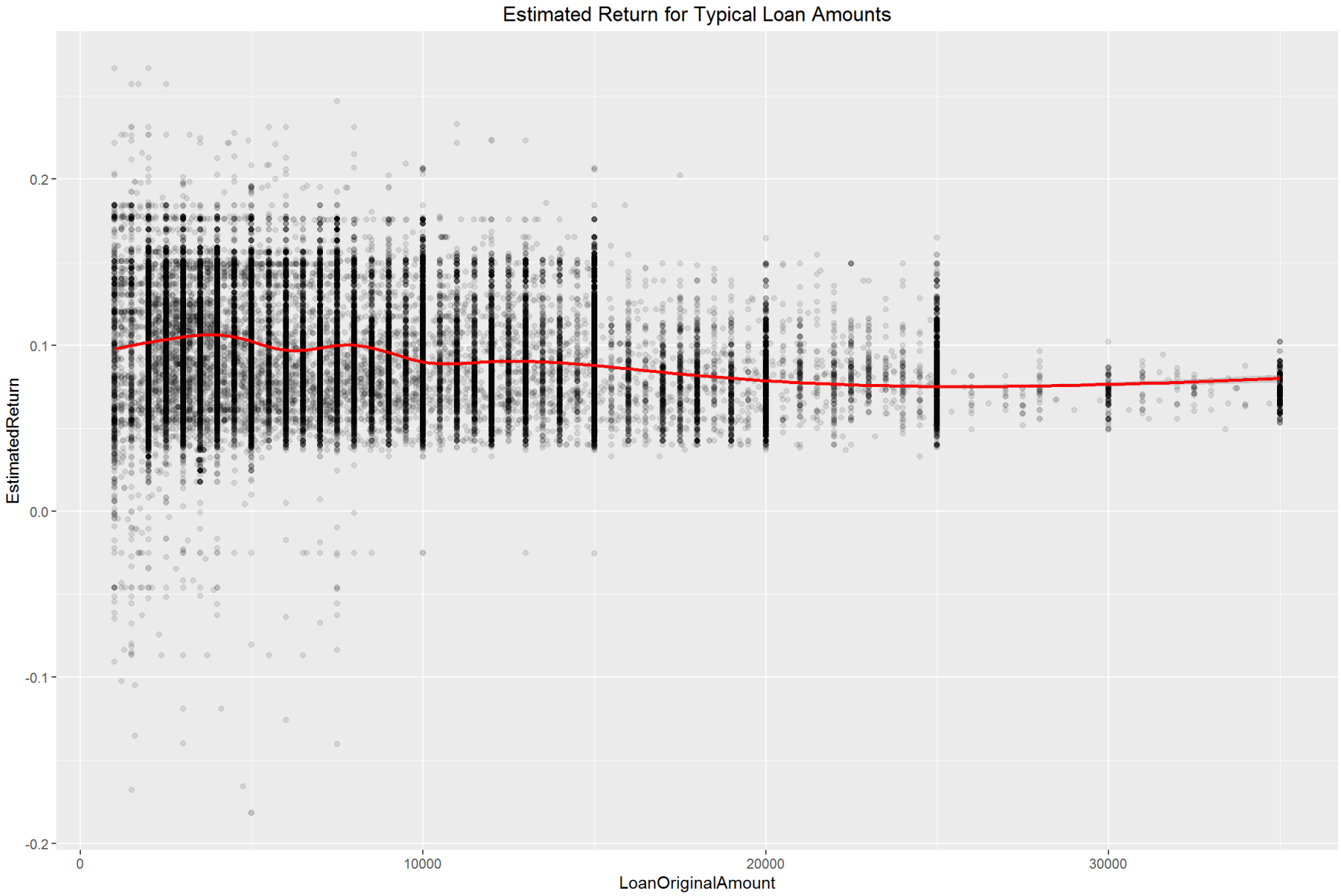
Comment: The higher one's borrower rating is, the higher the estimated loss (unrecoverable money through charge offs) will be for us as investors

I would like to see a boxplot of different estimated losses for each loan status..



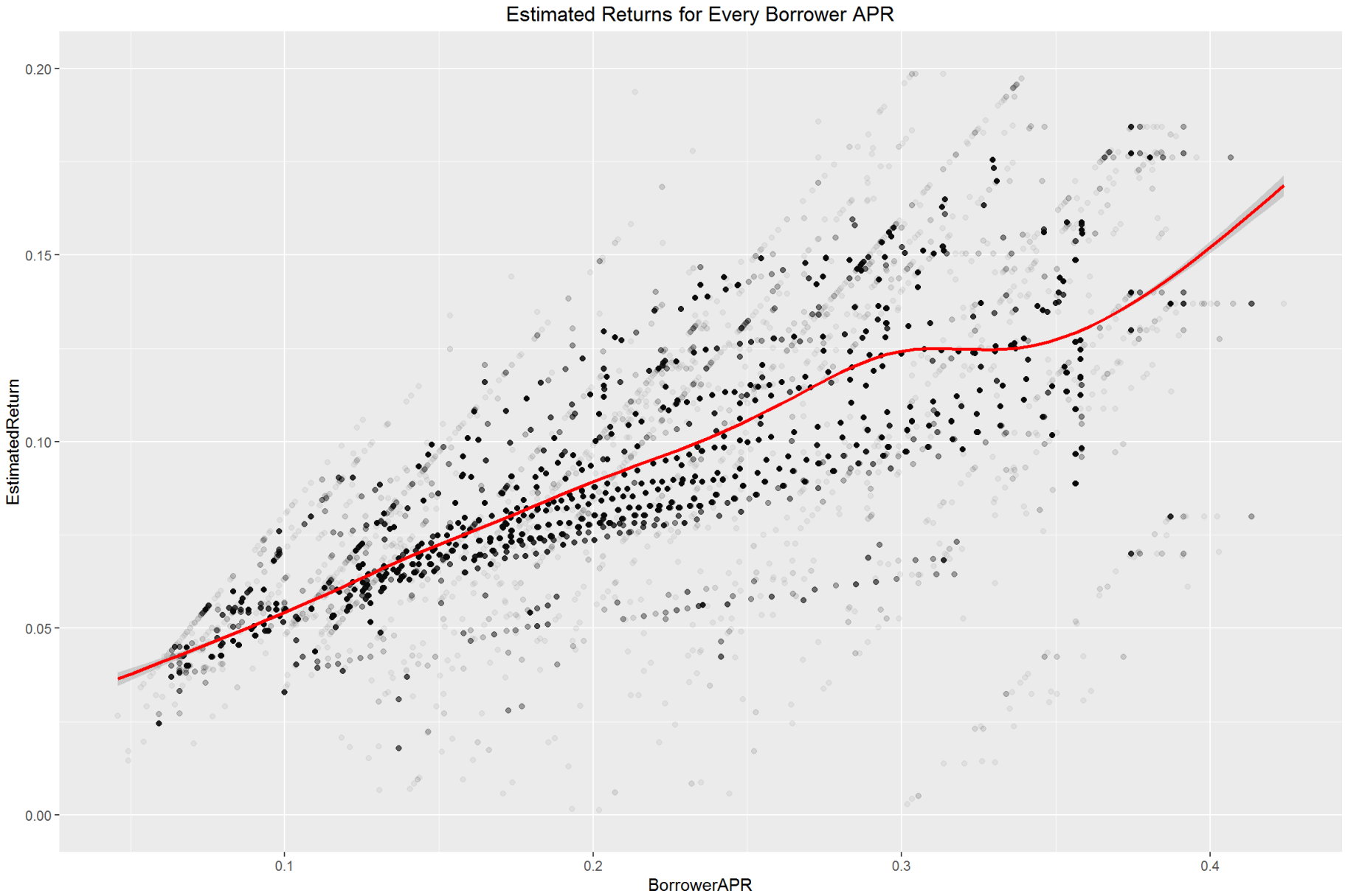
Comment:Loans that are charged off, defaulted, or past due usually have higher estimated losses to the investor Loans that are completed, current, and in the final payment progress are loans with less estimated losses

I would like to see a scatterplot of all different estimated returns of all different loan amounts…



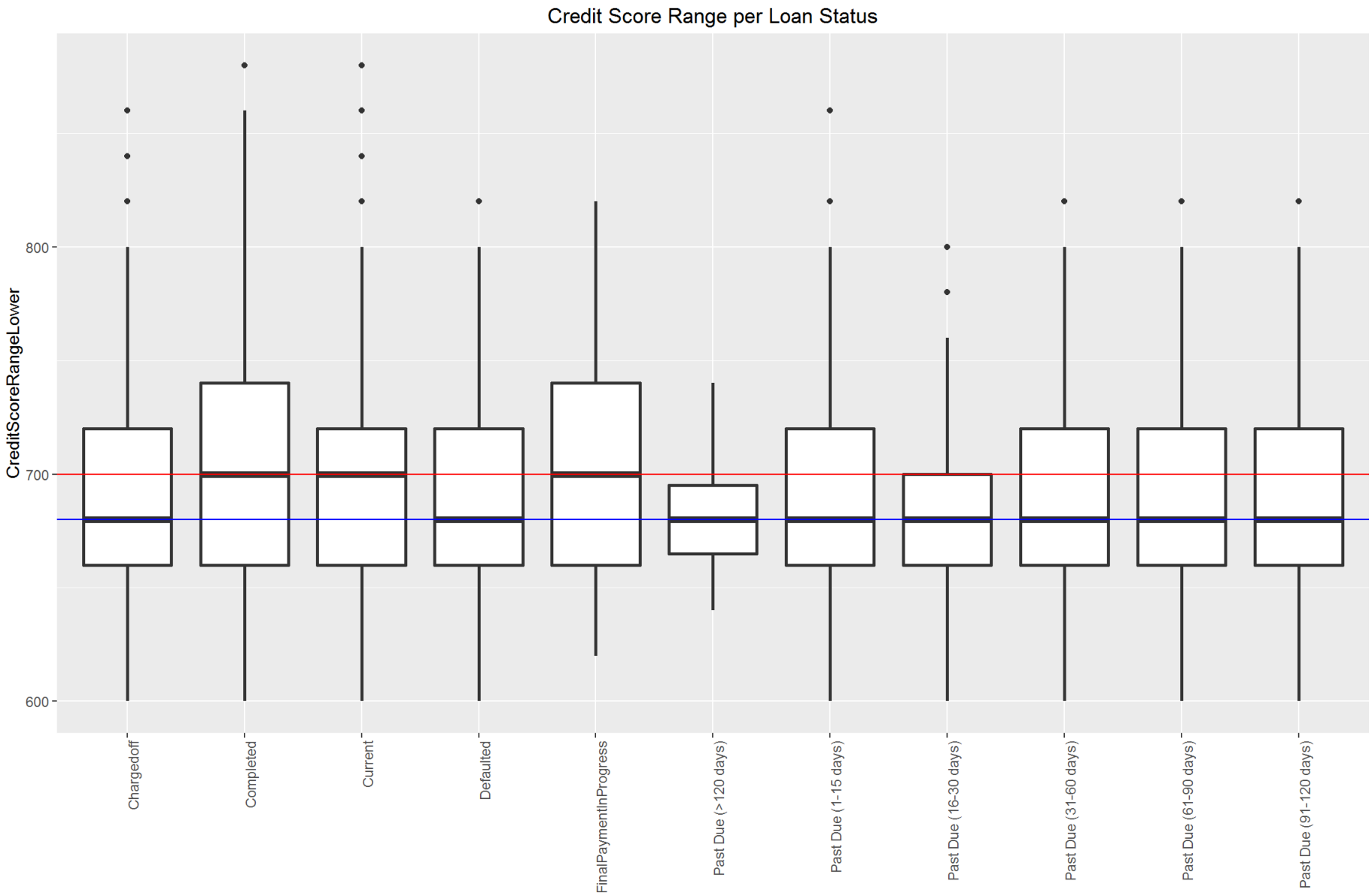
Comment: There is too much noise to conclude that the balance of the loan has any impact on the estimated return.

I would like to see a scatter plot of all different esimated returns from all different borrower APR.



Comment:The estimated return does increase as the BorrowerAPR increases but variation also increases as well.

I would like to see a box plot of different credit score ranges of different loan statuses…



Comment:loans that were cancelled, chargedoff, or defaulted usually are borrowers with lower average Credit Scores to begin with.
Loans that are currently going on, completed, or in the final payment progress are loans with borrowers whose credit scores are much higher.
The summary here is that the lower one's credit score, the higher the chance of having a borrower not being able to finish paying off their loans (or not receiving in the first place).

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?

The higher the Prosper Rating, the more likely higher the return (but more volatile)

The higher one's Credit Score, the higher the Prosper Rating.

The lower the Credit Score Range, the lower the estimated return

The higher the Prosper Rating, the higher the estimated loss percentage.

The higher a borrower's APR, the higher the investor's individual return (but more volatile).

Loans that have been completed, currently going , and payments in final progress are loans in which borrowers have higher credit score than loans that have been canceled, charged off, and defaulted.

Did you observe any interesting relationships between the other features

(not the main feature(s) of interest)?

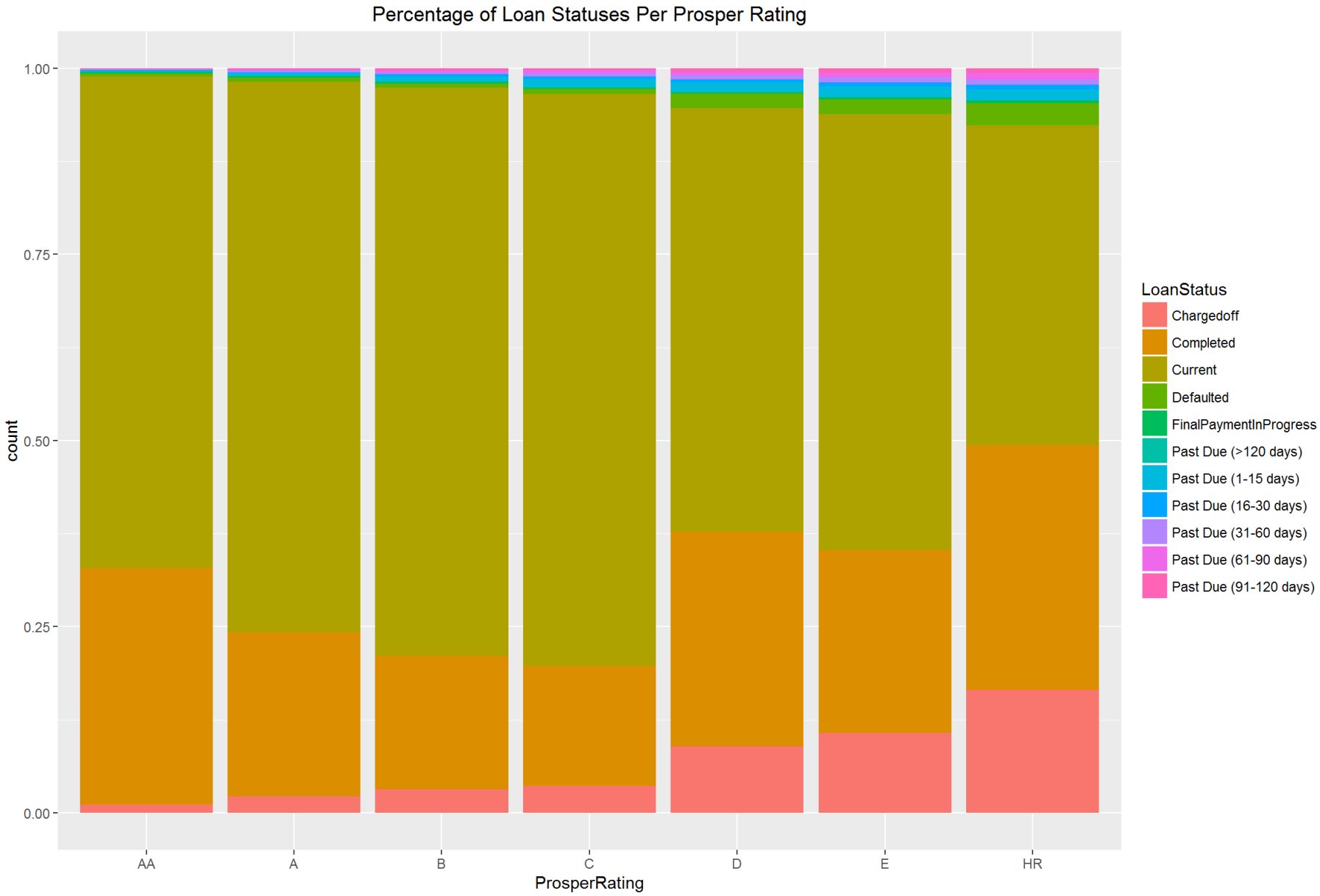
I notice that as the Prosper Rating risk increases, the spread for estimated return and loss increases. This makes sense because the higher the risk, the more likely a borrower can fail to repay back the loan thus resulting in charge off or default, and thus resulting in investors losing money.

What was the strongest relationship you found?

The strong relationship I found usually involves between risks and rewards, the higher the risk of a borrower (lower credit core and higher prosper rating) , the higher the return but also higher the loss.

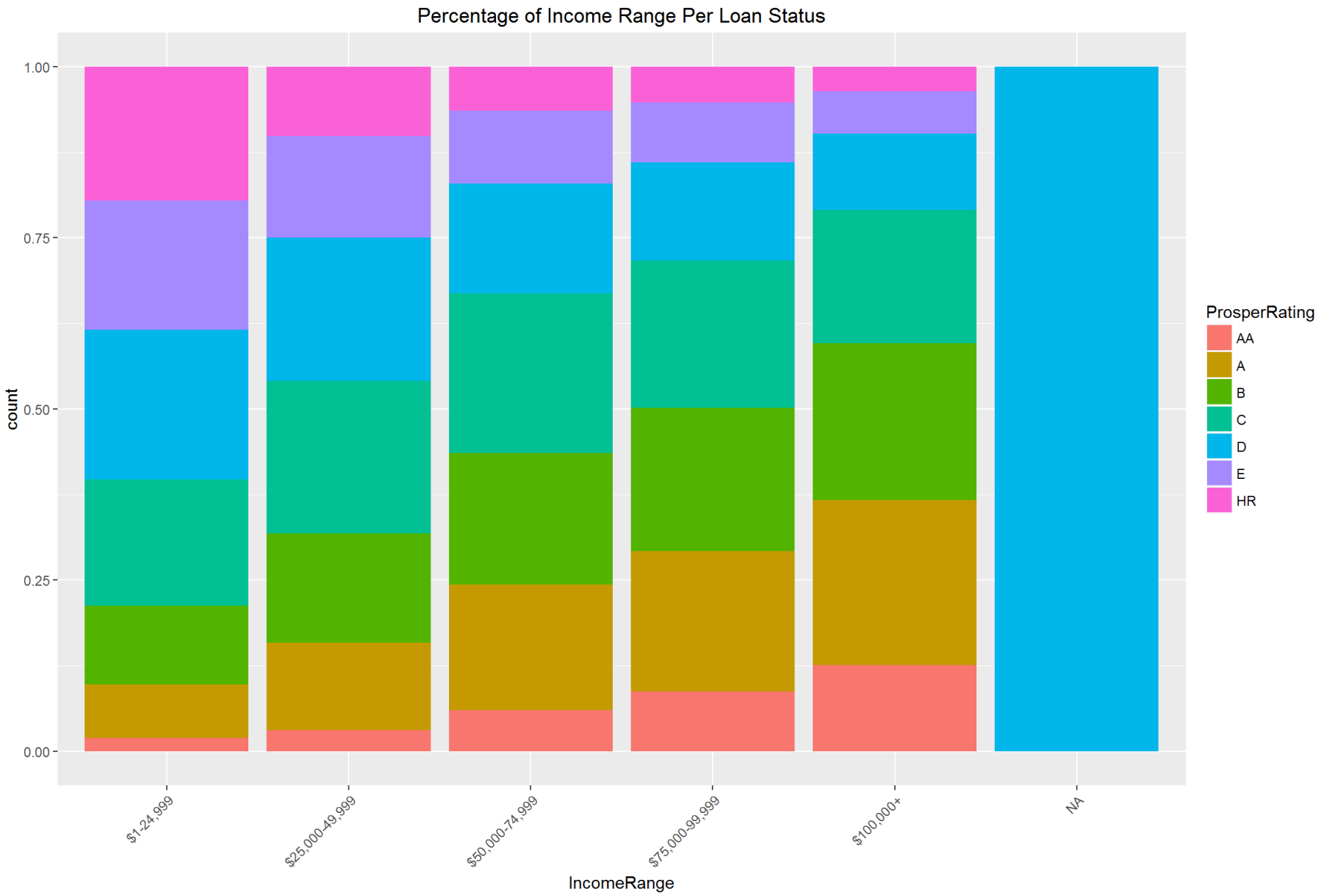
Multivariate Plots Section

I would like to see the percentage of counts of loan statuses per every prosper rating loans…



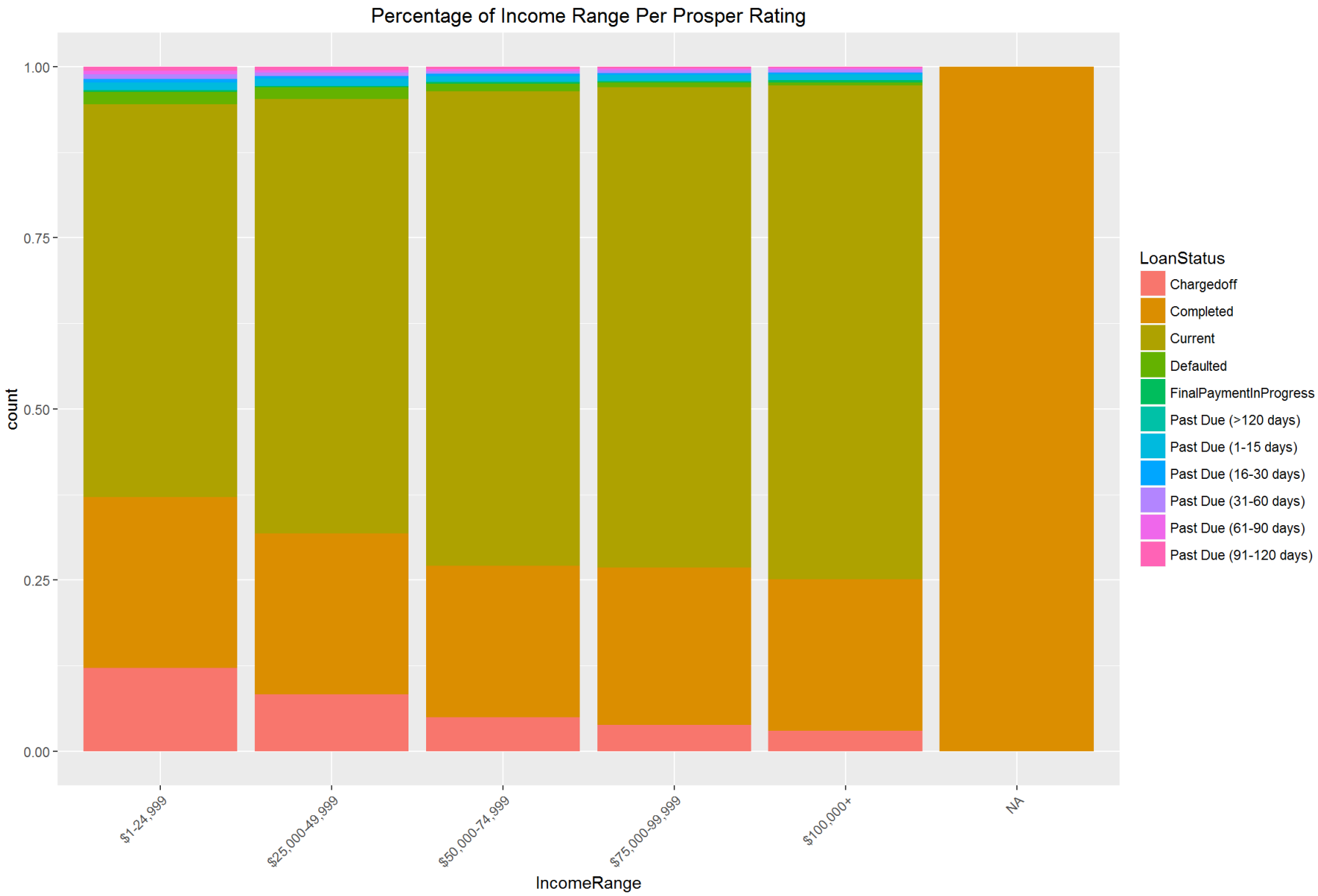
Comment:The higher the loan risk, the more likely the loan will end up being charged off and defaulted. And also more likely borrowers will fail to pay on time.

I would like to see the percentage of borrower's income range for all prosper rating loans…



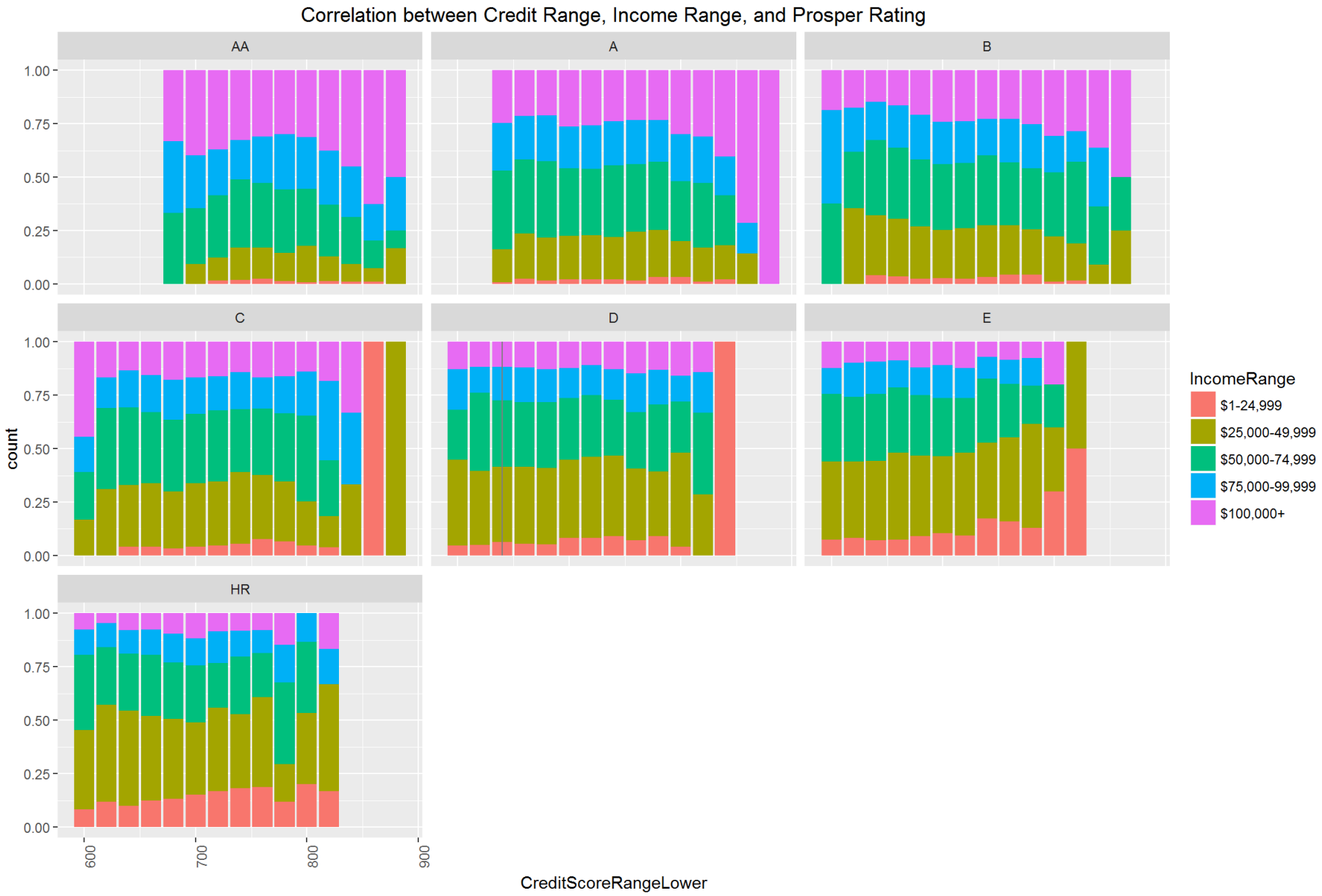
Comment:The higher the borrower's income, the more likely their loans would be lower risk. The lower one's income range, the more percentage their loans would be higher risk. It's probably because a borrower with in the lower income range would most likely have sufficient income to pay back the loan thus putting that borrower at a higher risk.

I would like to see the percentage of income range for different loan statuses..



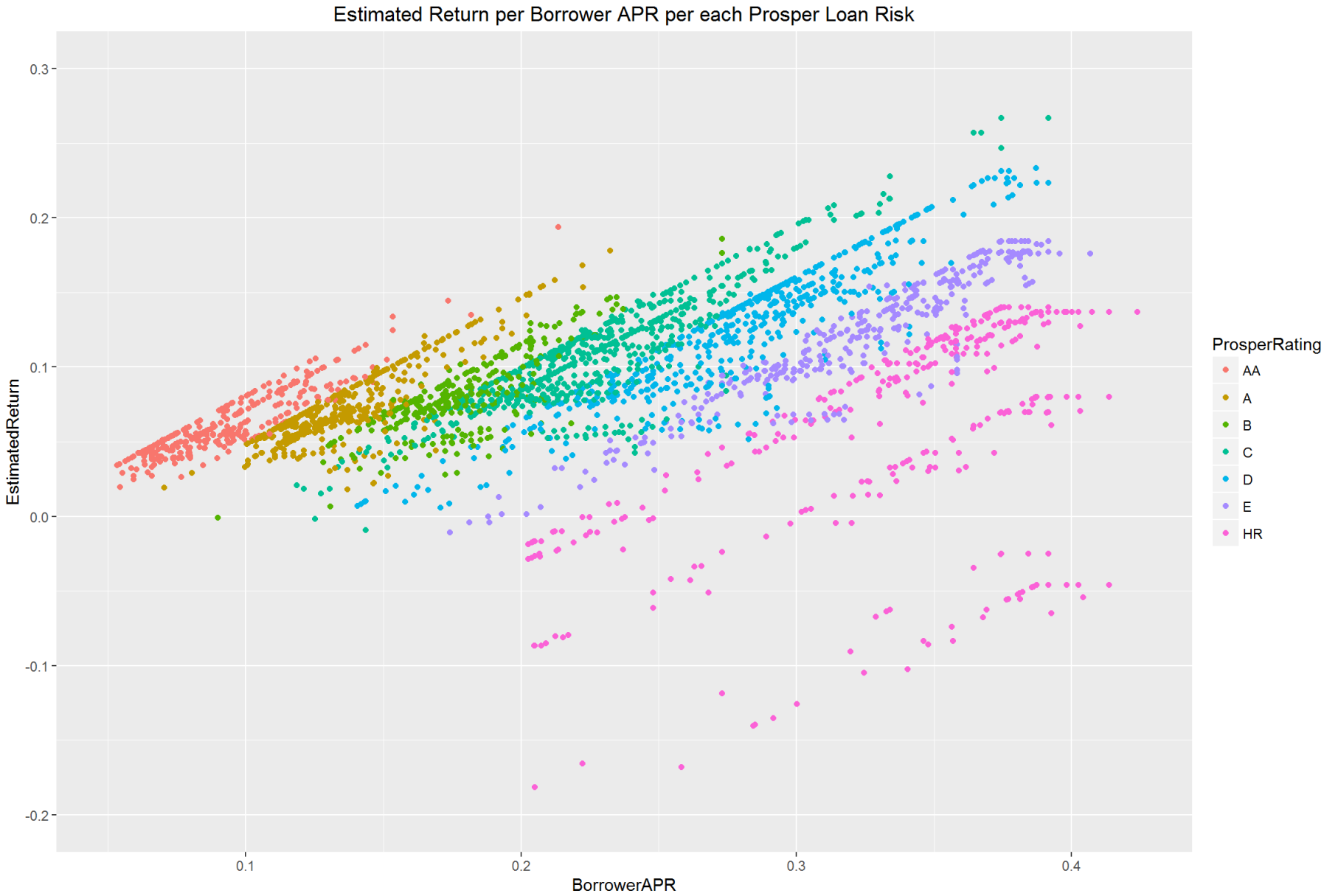
Comment:The lower one’s income range, the more likely their loans would be charged off, defaulted and have late payments

I would like to see different graphs of different correlations between, borrower's credit score, income range, and prosper rating..



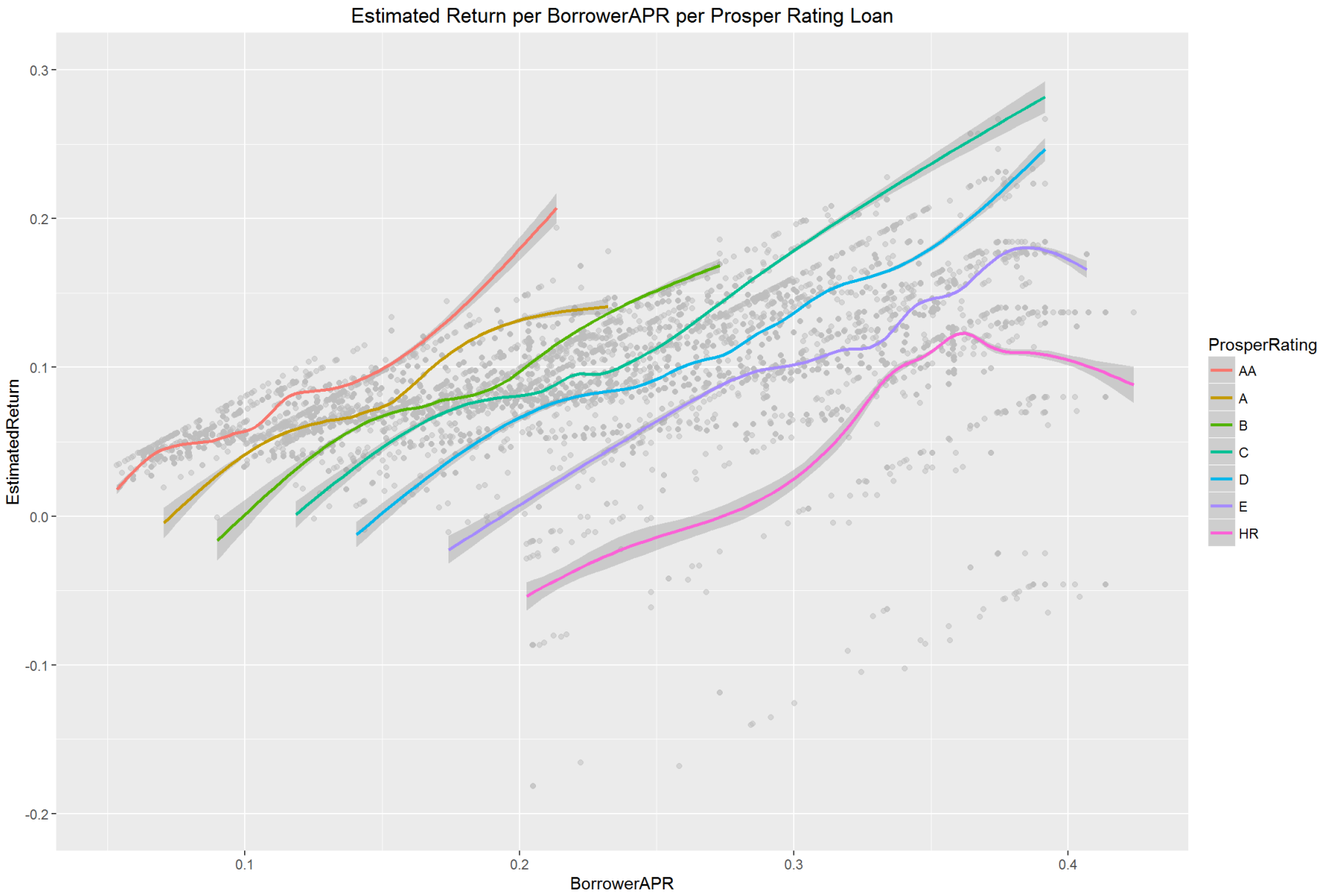
Comment:The higher the risk of a prosper loan, the more likely the borrower's income range will be lower (as we start to see more brown, pink, and brown bars) and the borrower's credit score will be lower (the bar chart seem to be shifting to the left as the prosper rating increases).

I would like to see the estimated return of from each loan of different borrower's APR and prosper rating…



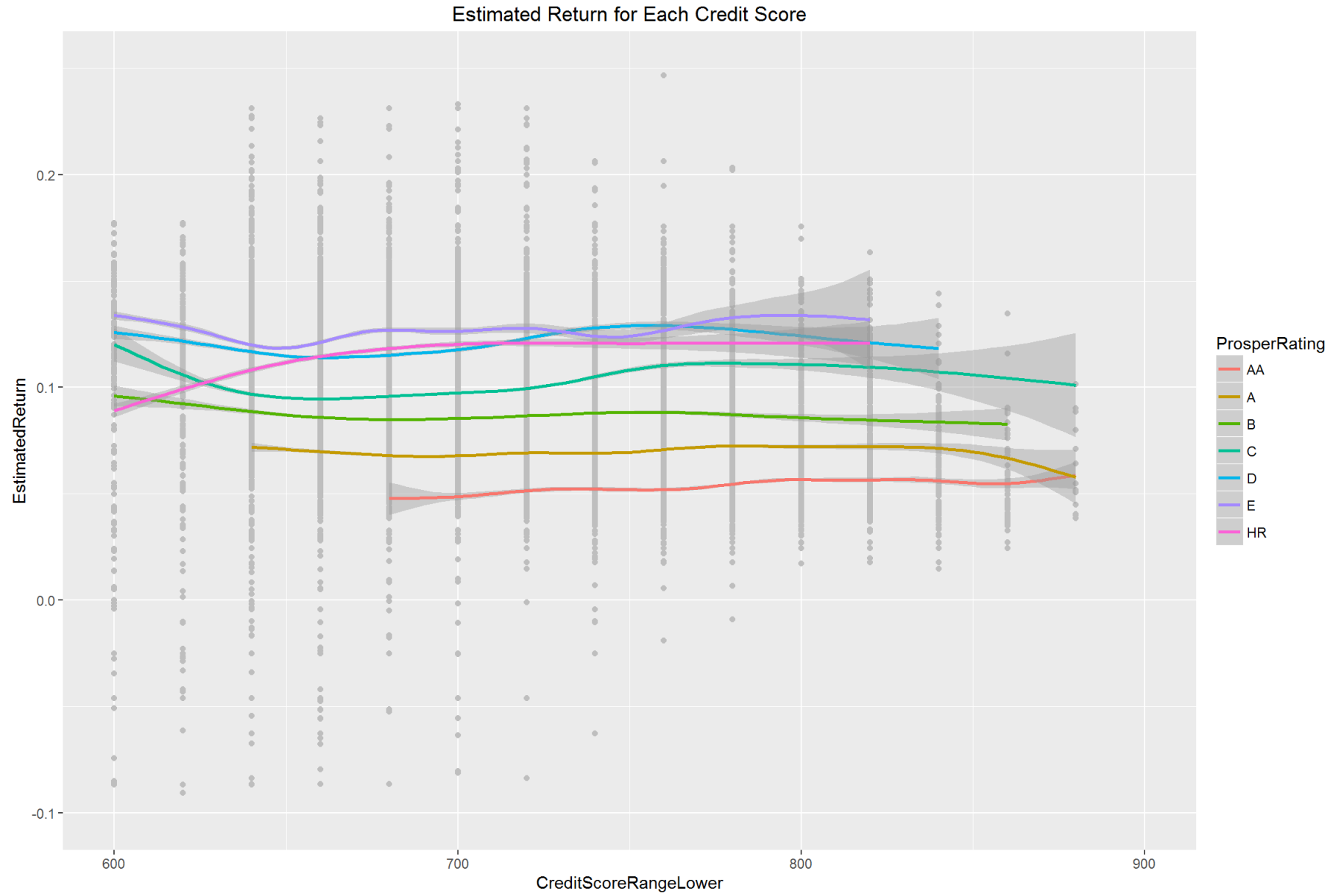
Comment:We see that different prosper ratings are directly stacked on top of each other. We can use this data as an advantage for investors to decide on which loans for optimal returns.

I would like to see an average moving line through different estimated returns through borrower's APR and prosper rating loans…



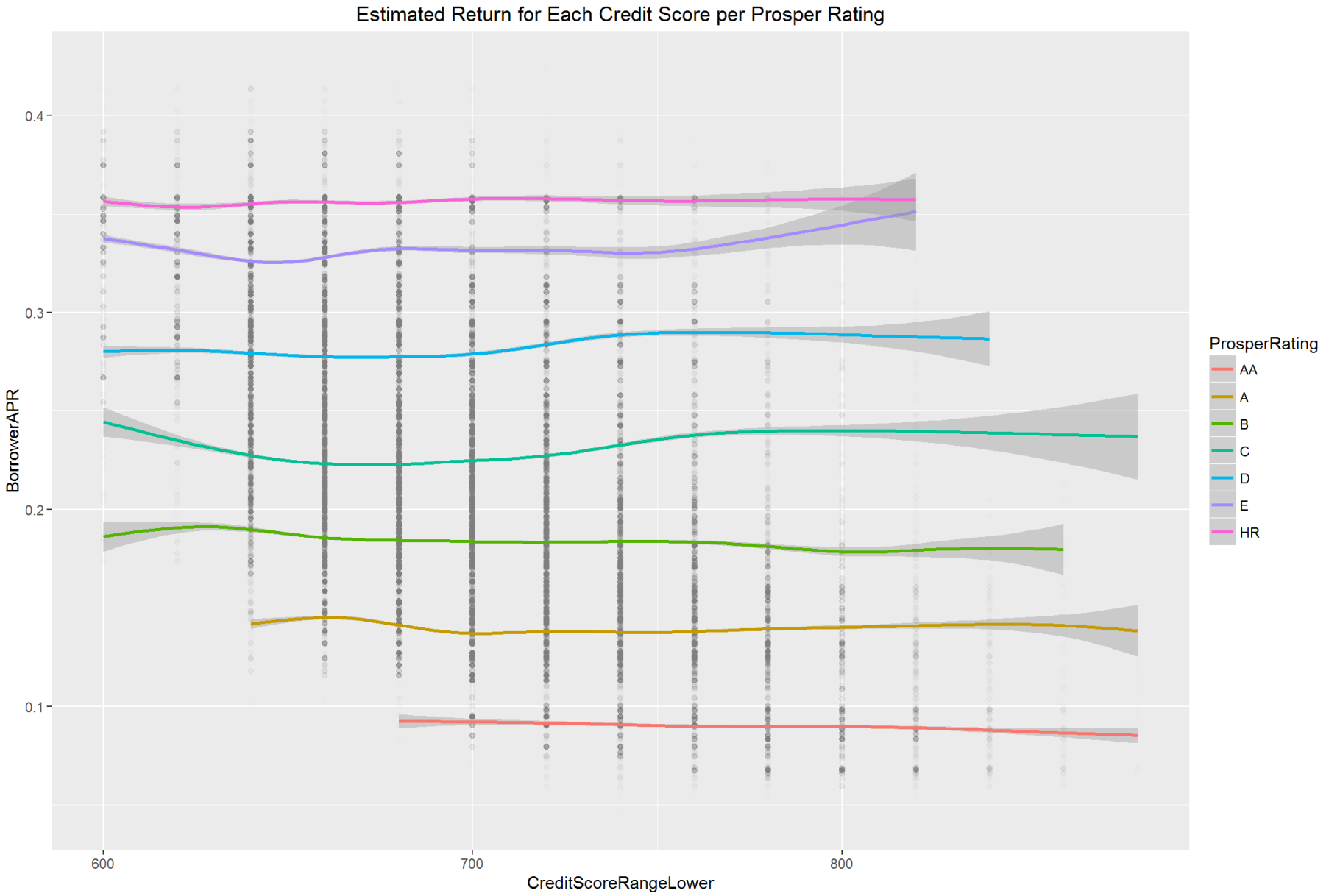
Comment:That we see the averages of each ratings do not intersect one another. Therefore, I see a pattern..

I would like to see an average line through all loans of different estimated returns of different borrowers APR with different prosper rating…



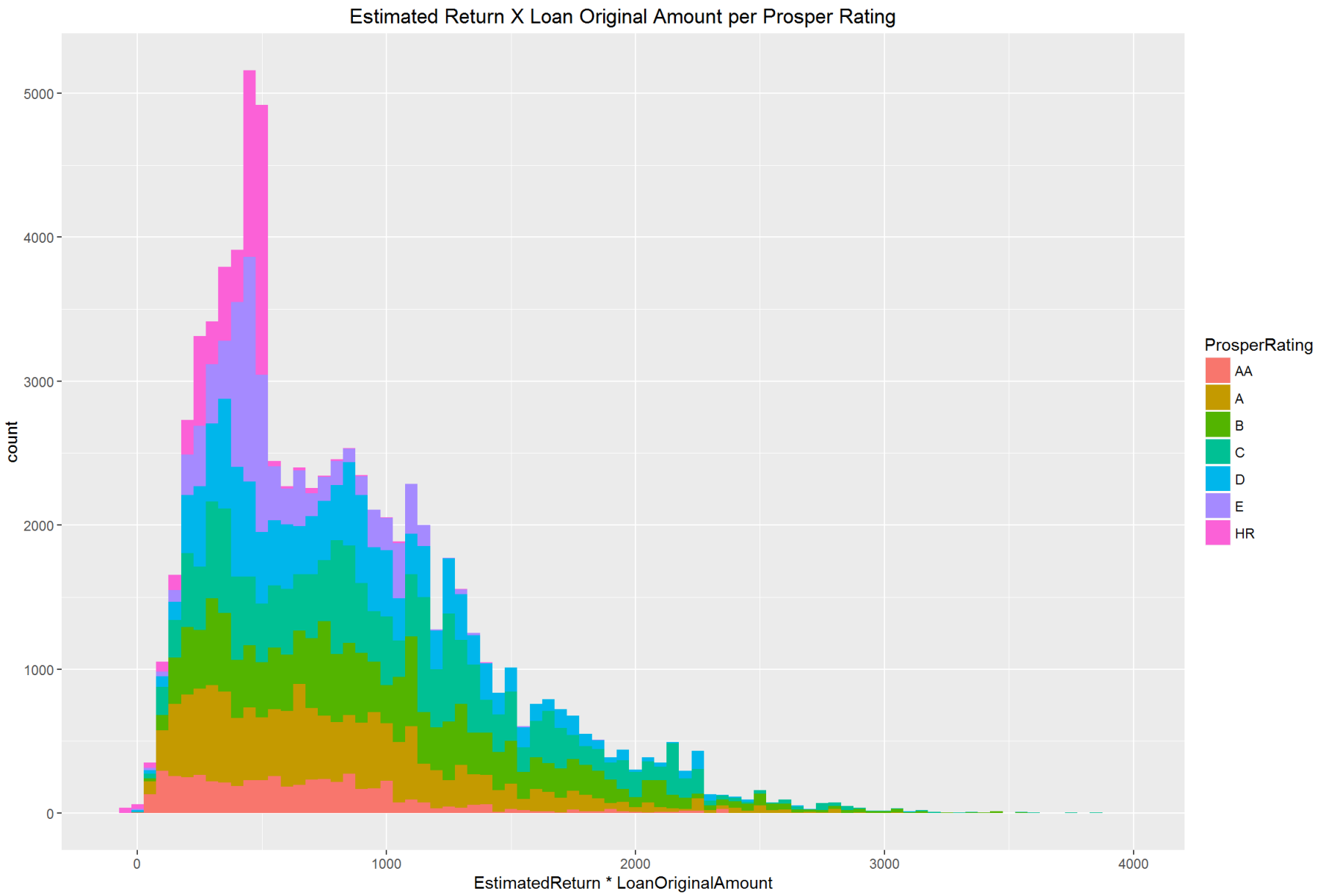
Comment:It’s difficult to determine which Prosper Rating to invest at different Credit Scores. The variance is also too high and other factors are more important to contribute the decision.

I would like to see estimated returns for each credit score per prosper rating…



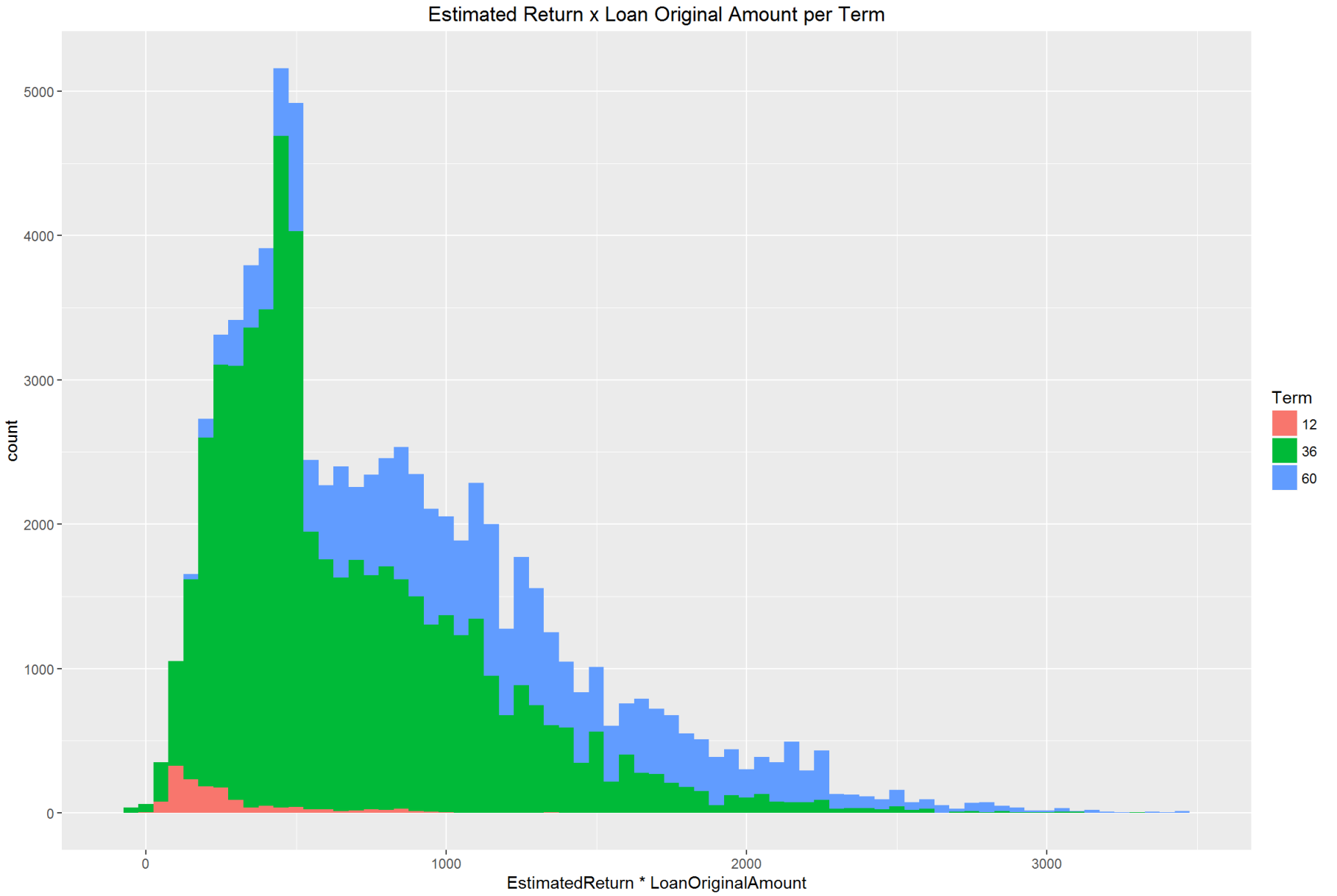
Comment:There appears to be a consistent horizontal pattern for each Prosper Rating. By consistency, the higher the rating of a loan, the higher and consistent the borrower's APR.

I would like to see total amount returned (estimated return times loan original amount) for each prosper rating…



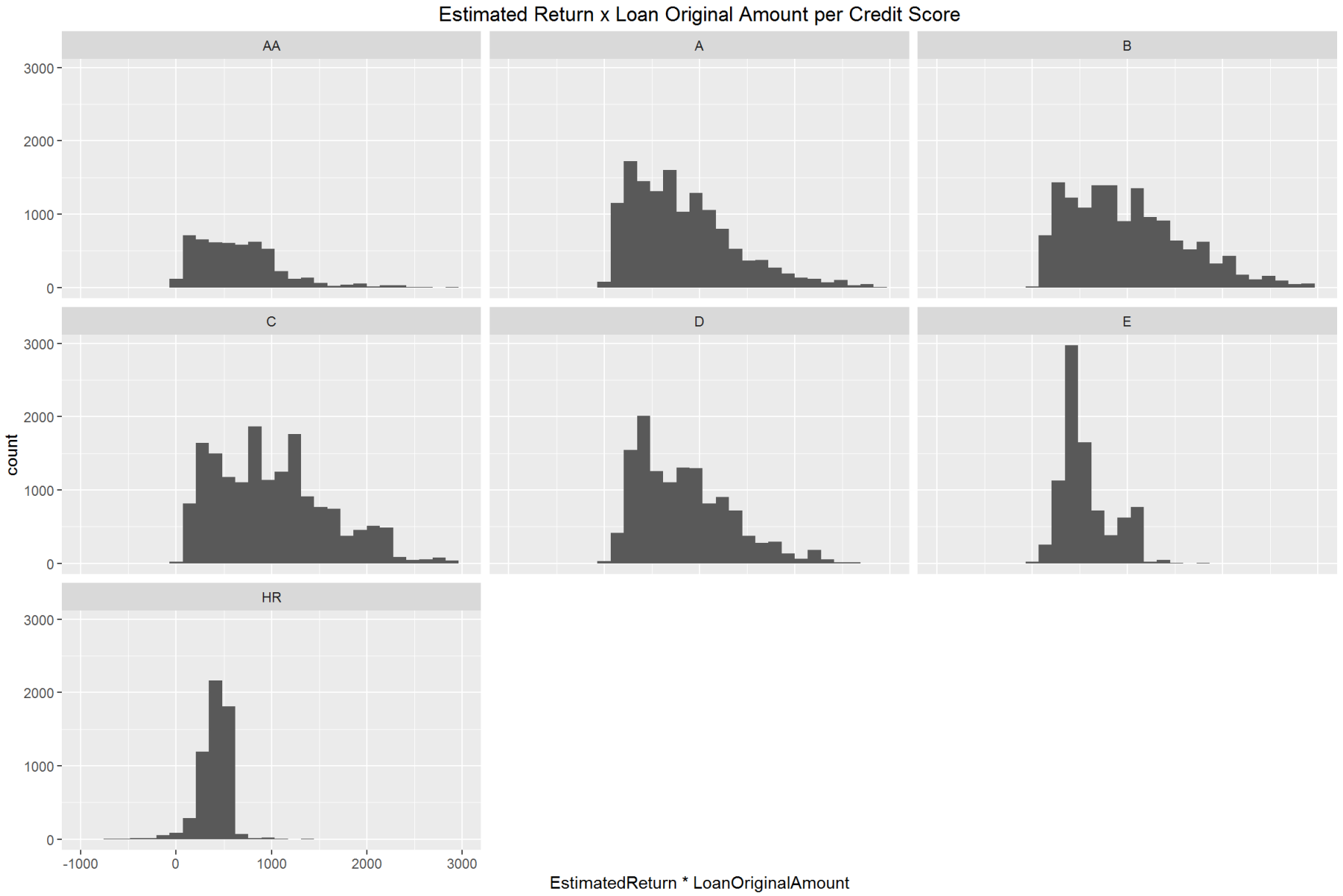
Comment:At most, an investor can make almost $4000 on a loan. We see that an invstor investing in HR can struggle to make more than $1000 in return.

I would like to see investor's total return (estimated return times loan original amount) for each term…



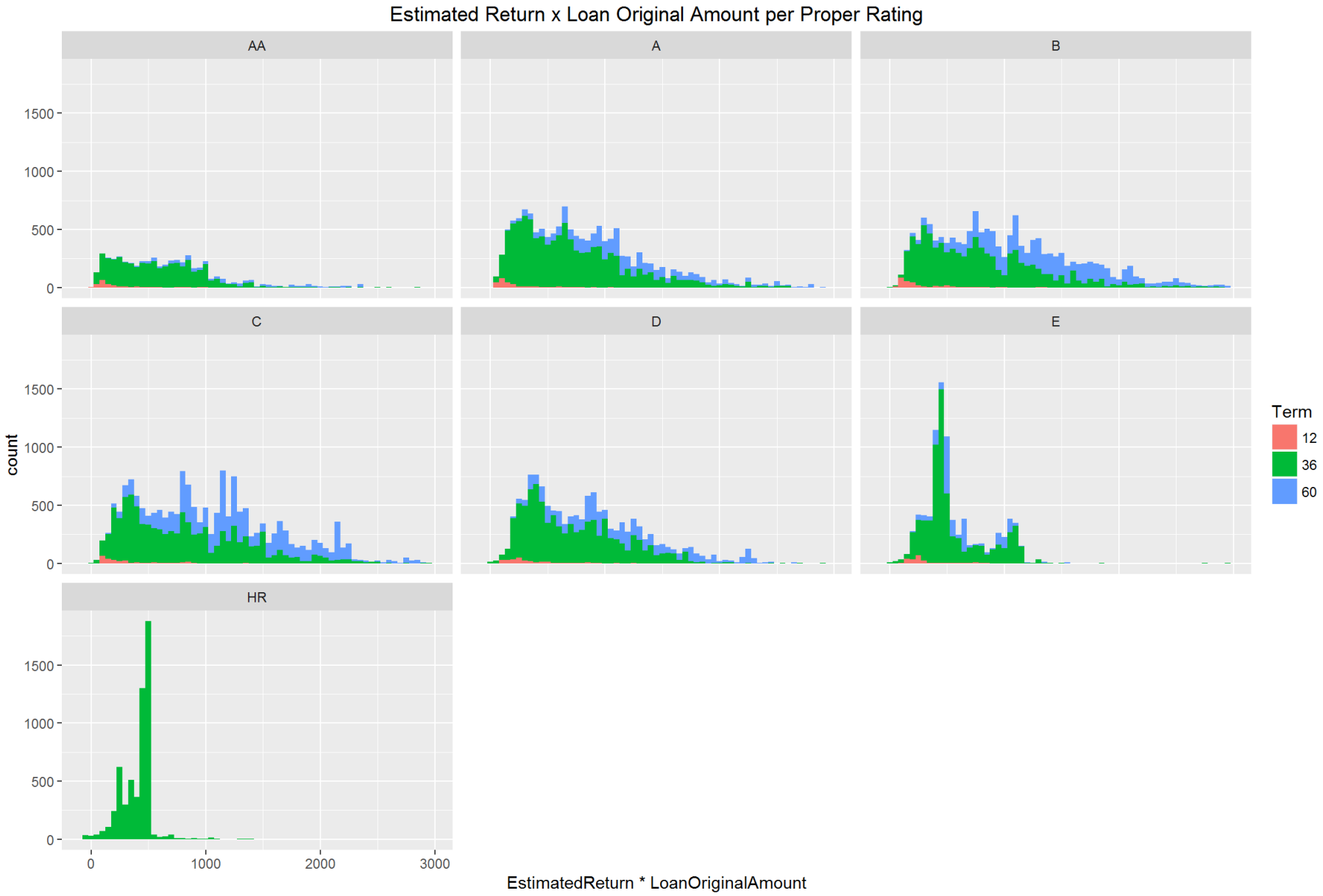
Comment: Investing in longer term can yield more earnings to the investor

I would like to see all investor's total return (estimated return times loan original amount) at each prosper rating at different credit score…



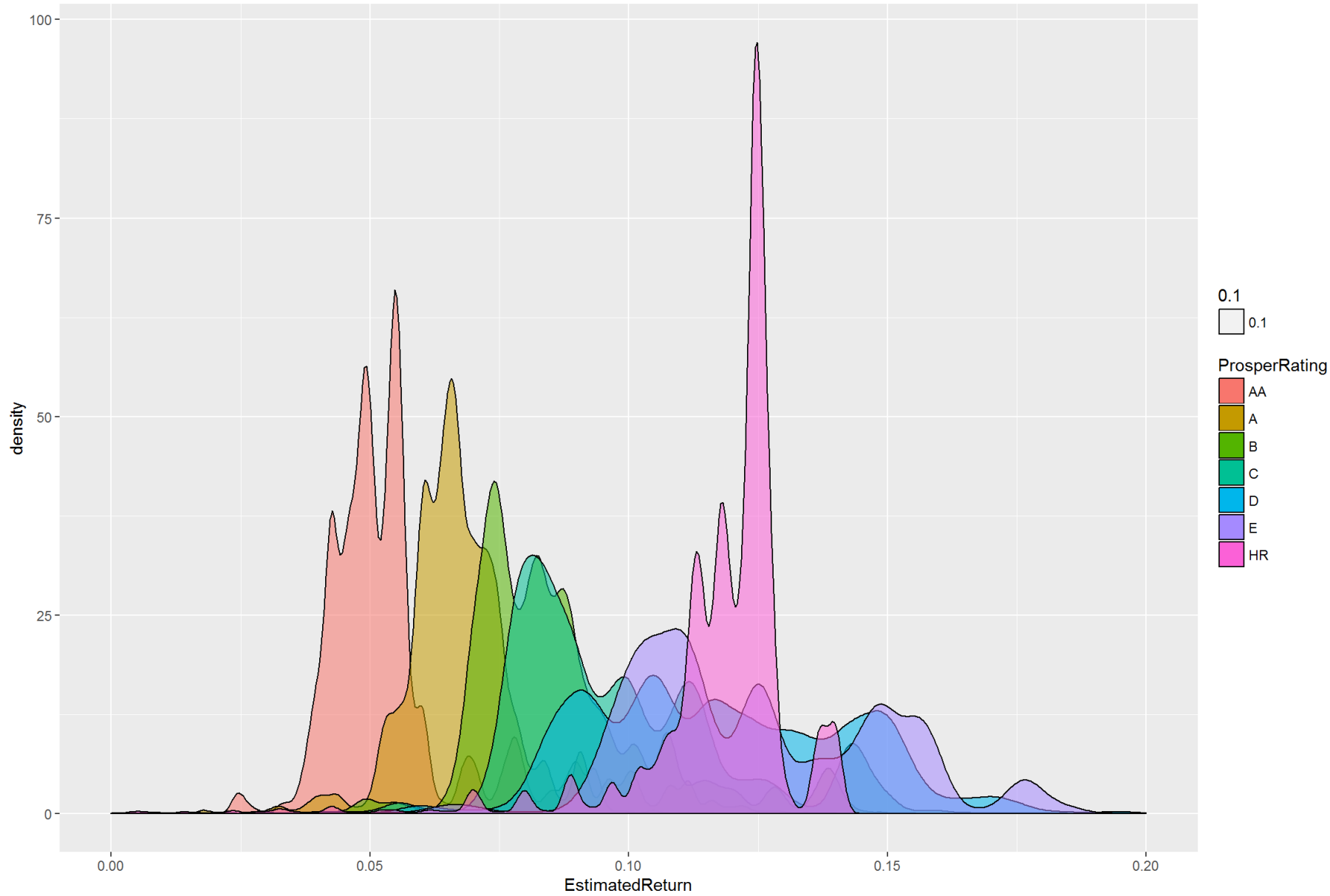
Comment:An investor on average can achieve the highest return from loan "E".
Loan "E" also has appears to have the least variance so it makes achieving that higher return much easier for the investor. Investors can benefit earnings from borrowers with higher Credit Scores in safer loans: "AA" and "A" where we see some slight "pink and purple".

I would like to see all investor's total return (estimated return times loan original amount) at each prosper rating at different loan terms…



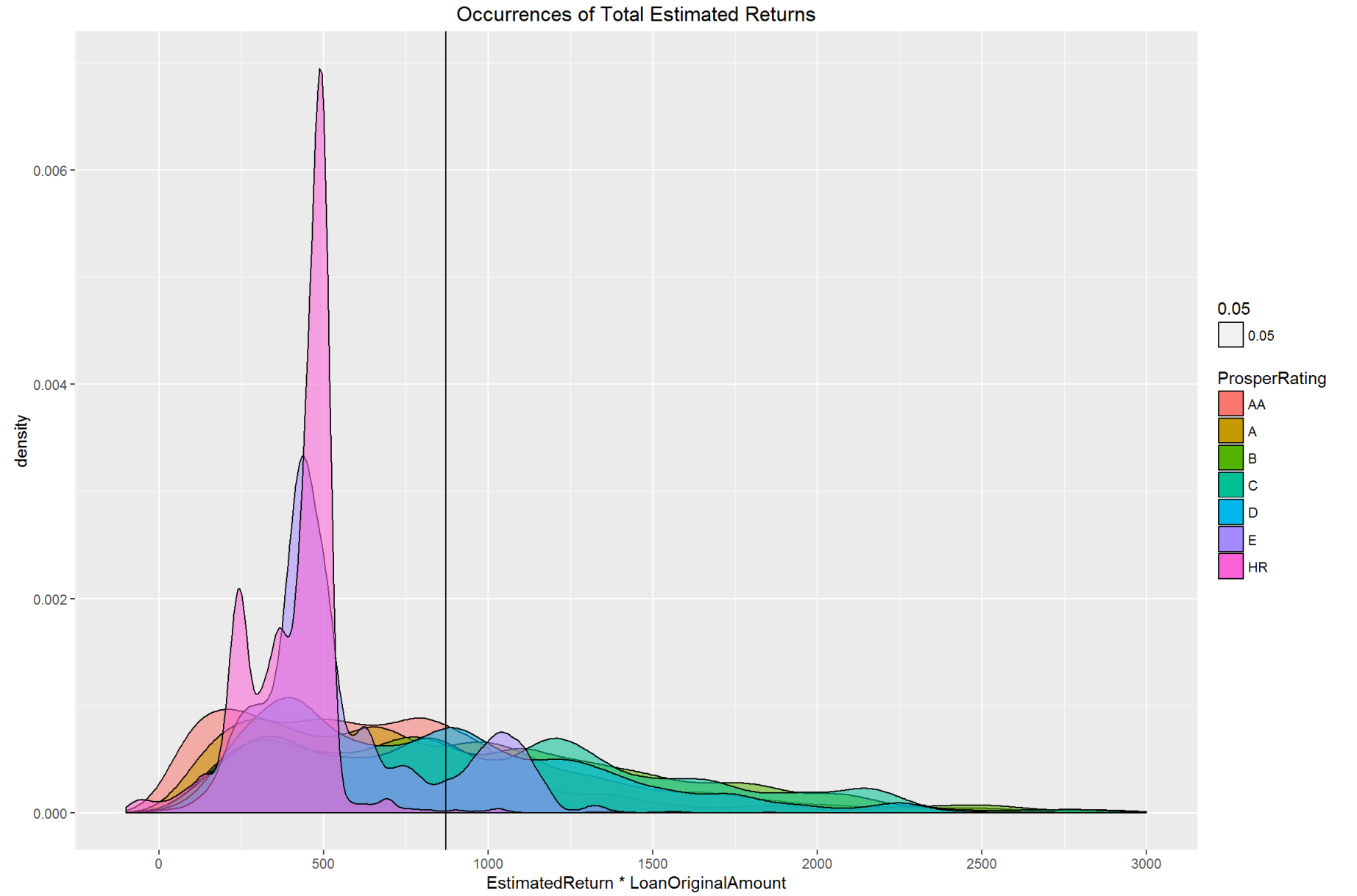
Comment:An Investor will not see the benefit of a loan borrowed for 60 months at "HR" (as we cannot see any blue spots in that graph). However, an investor may benefit greatly from that "HR" loan because of the spike we see returning almost $25,00. We can hardly see any great gains from 12 month term loans.

I would like to see estimated returns for each prosper rating..



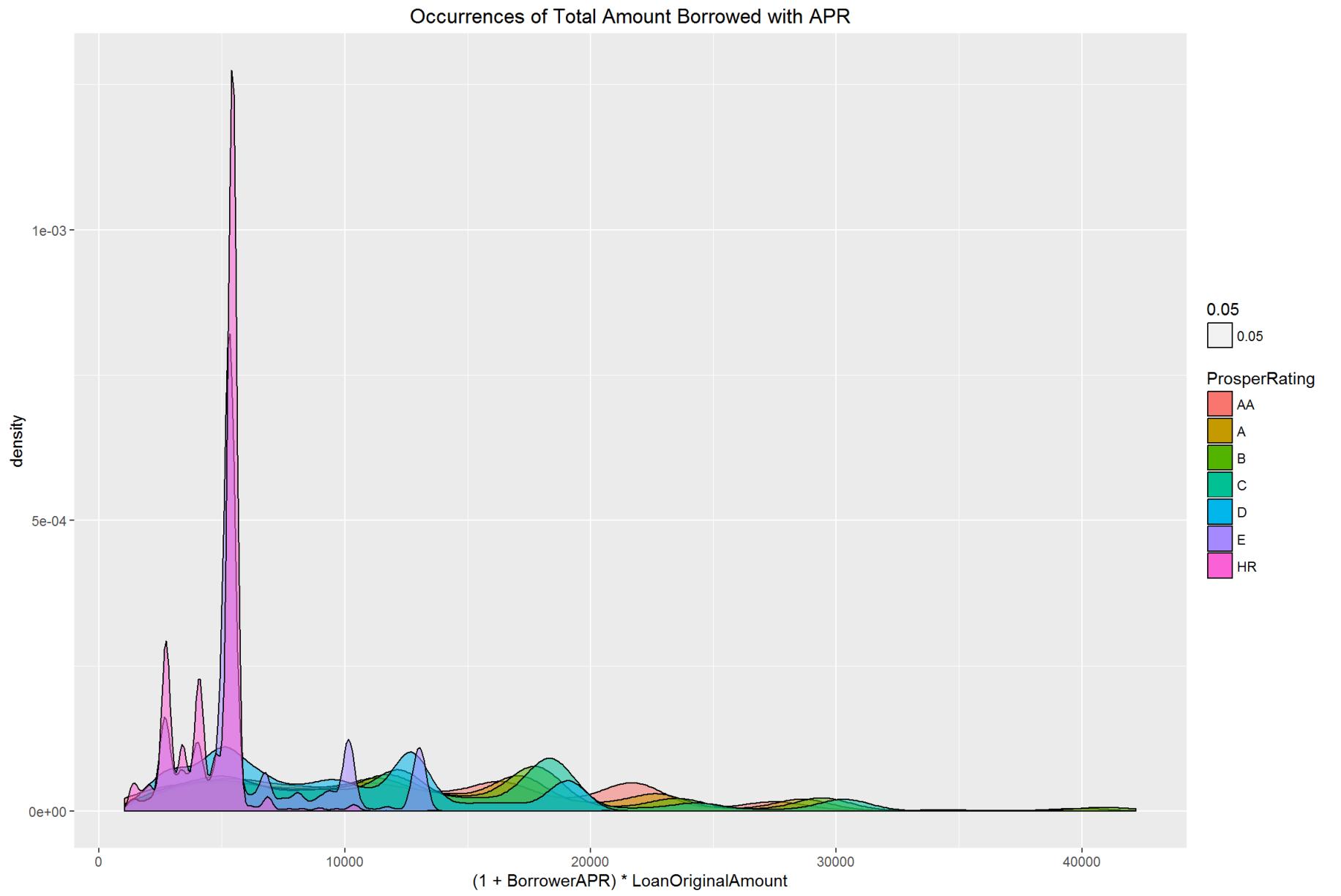
Comment: Indeed, the average estimated return does increase as the risk of the Prosper rating increases. However, a loan of “D” or “E” can achieve higher returns (at and above 0.15) than loan “HR”. Estimated Returns for “HR” do not appear pass 0.13.

I would like to see a density chart of all investor's total estimated return for each prosper rating…



Comment:The average investor return per loan is $872.63. It looks like that can most likely be achieved with a "AA" or "B" loan. Investing in the "HR" loan, it is very unlikely an investor can get an investment of $872.63. In fact it looks to be the average return of the "HR" loan is $500

I would like to see a density chart of all investor's total amount borrowered with APR for each loans of different prosper rating…



Comment: The lower the risk rating of the borrower, the more the borrower will borrow. The maximum amount for borrowers to borrow on loans rated "HR" appears to be at about 1,000, then the risk levels of Prosper decreases as borrowers seek to borrow more.

Probably the reason why it doesn't look like borrowers of loan HR can borrow more than 1,000 is because of high risk of Prosper Loan company to not want to lose. You obviously do not want to lend out a whole lot of money to risky people and have a high chance of not be able to have them pay back fully. This so on for the other less riskier loans.

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?

There are certain percentage of Borrower's APR in which the investor should not invest in because the previous less risky loans will tend to have a better return than.

The longer the term, the higher the estimated return.

Sometimes, it is not worth investing in an HR Loan between in an APR less than 0.3 when a much return can be invested in other less riskier loans less than the APR of 0.2.

Were there any interesting or surprising interactions between features?

Yes on average an investor who invests in the "HR" loan will have high return but the highest return can be achieved through E or D loan.
It is low probability an investor can achieve that sort of return in the upper quartiles of loans E and D but it is something that mostly unlikely to be achieved through loan HR.

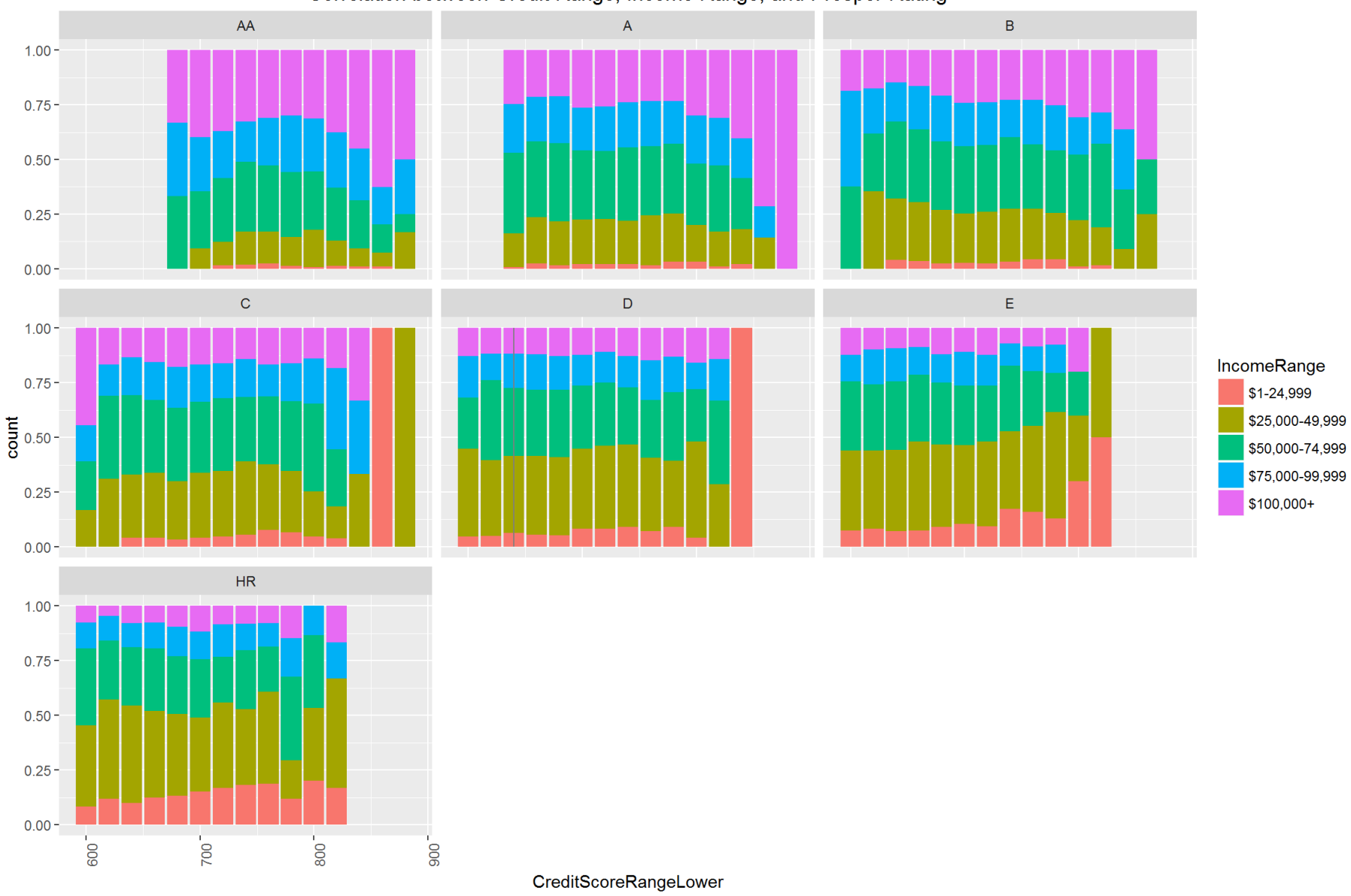
OPTIONAL: Did you create any models with your dataset? Discuss the

strengths and limitations of your model.

N/A

Final Plots and Summary

Plot One



Description One

Looking at the graphs below starting with the lowest risk (“AA”) and examining up to the higher risk (“HR”). A few observations. We start to see more colors of the lower income ranges (see more red, brown, and pink emerging)

We start to see the graph shifting to the left (meaning the borrower’s credit score decreases)

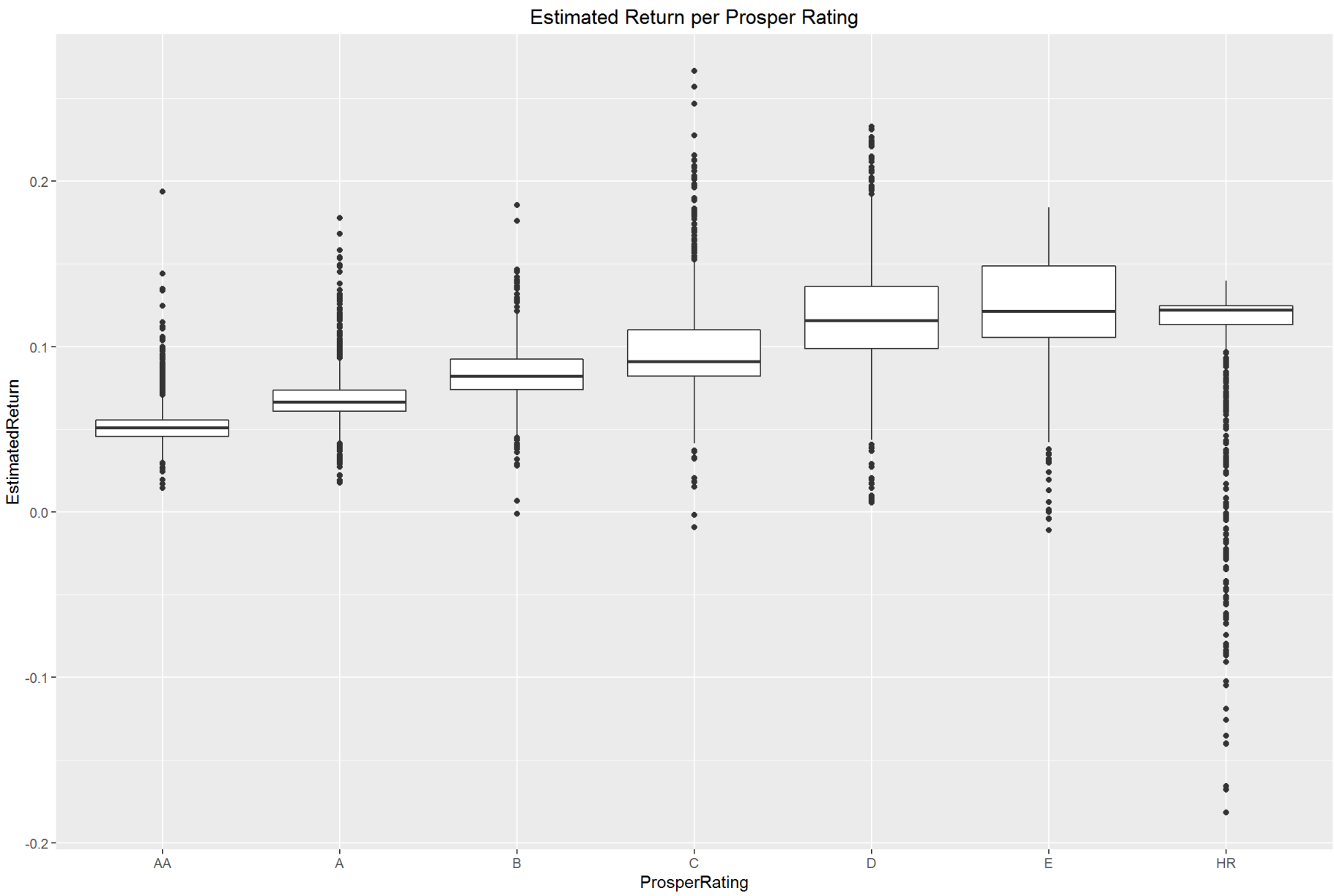
This concludes that a prosper loan is most likely determined by a borrower’s income range and credit score.

The lower a borrower’s income range, the less likely that borrower would have sufficient balance to pay back, thus higher the Prosper Rating would be.

The lower a borrower’s credit score, the less likely the borrower would have the ability to pay back the loan. Most likely because the borrower doesn’t have a stable job or have bad habits of making late payments.

This is important to the investor to understand why a prosper loan risk rating is rated the way it is and the variables that may contribute to that rating causing investors to either have gains or losses whe investing.

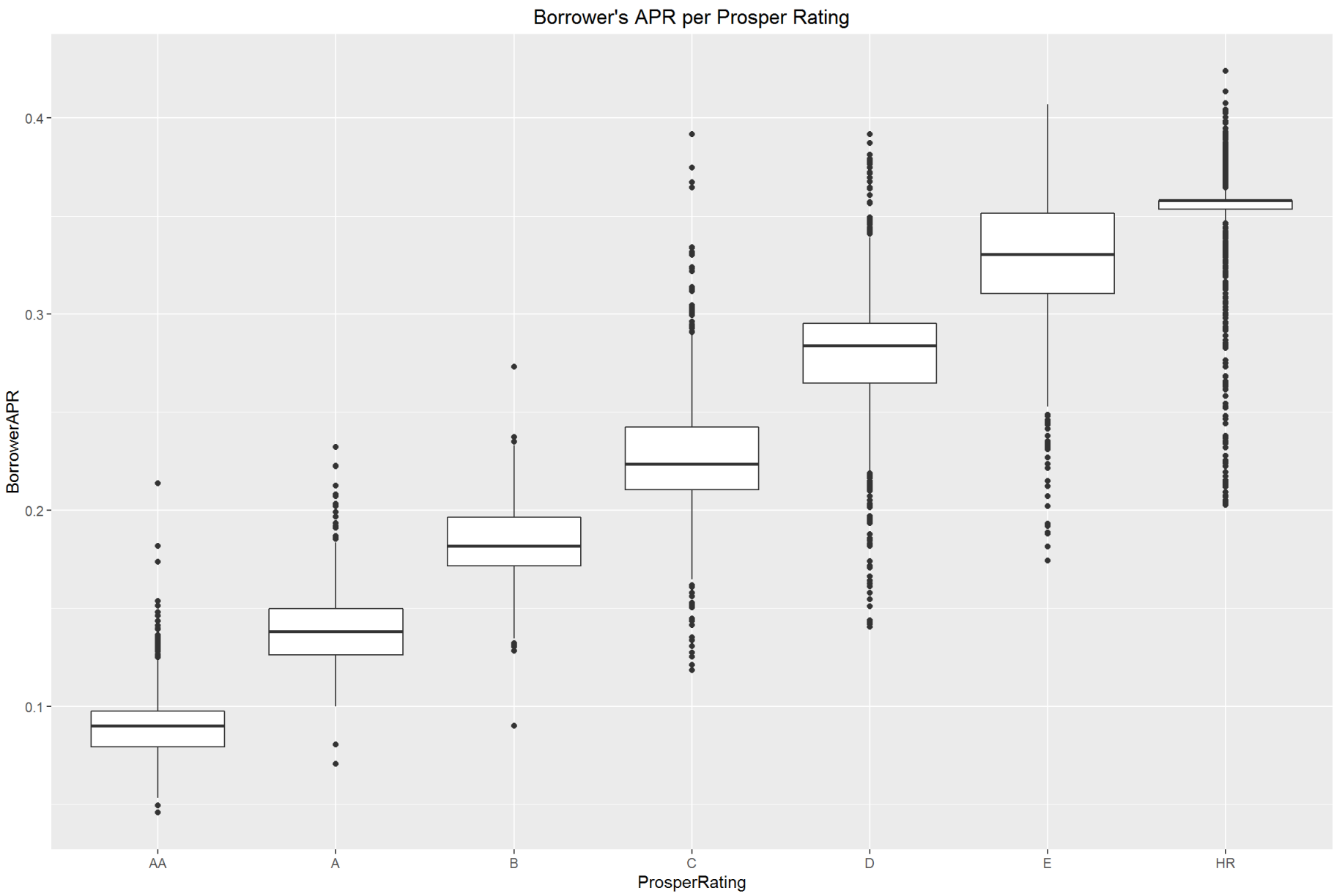
Plot Two



Description Two

We always tend to think that the higher the risk of a loan, the higher chance on average we will receive a higher estimated return than the other lower risk loans. However, that is not the case by looking at loan HR and comparing that with loans E and D. The average estimated return for E and D look to be identical to the average return for loan HR. And also, looking at the upper quartiles of loans E, D, and HR, one can achieve a highest return on loans E and D that may unlikely to be achieved on loan HR. Also HR has a lot of potential risk of losses for investors. For this, an investor may be better off looking to invest in loans E and D for higher returns and ignoring investing at all in HR if possible.

Plot Three



Description Three

This plot helps investors decide on which typical Prosper Rating loan to invest in based on the borrower's APR. Investor should target the filled color boxes and avoid the "tail ends" if possible (the plots below the lower quartile and the plots above the upper quartile).

For example, in loan B, an investor should invest in that loan when the APR is between 0.17 and 0.19. Anything less than an APR 0.17, an investor would be better off investing in loan A and anything higher than 0.19 an investor should be better off investing in loan C.

Reflection

Where did I run into difficulties in the Analysis?

One of the great difficulties I had was before my analysis was greatly trying to understand the purpose of each feature. With the 81 different features, it's difficult to pick which ones would offer great relationships between each other and insights to what investors and borrowers want. I was surprised that an investor debt to income ratio did not result in much correlation to any features. I was hoping the more debt/income ratio a borrower have, the more likely he/she will struggle to pay it back thus resulting in higher prosper loan rating.

Where did I find success?

I found success after plotting bivariate and multivariate charts thus finding patterns between certain features. Some of the features I chose were ones that would give investors good information as to how they would choose where to invest. And after doing a little bit of exploring between data of each desired feature, that's where I found success. I succeeded knowing for investors that loan HR should not be invested at all. I found this out by plotting graphs of the trends between Borrower APR, Prosper Rating, and the Estimated Return. As an investor, we would usually want information such as the borrower's credit score, income range, and APR of a loan. Those 3 key information would allow investors see if investing in that certain loan would probably have good returns/losses. In general, I found success by thinking what an investor would think: What are we most concerned about from a given loan? And we want to know about the Borrower's history as much as possible to minimize risk and maximize returns.

How could the analysis be enriched in future work?

Most of the data collected are loans currently being paid off slowly by the investor. It's better to revisit the data after all current loans have ended in either completed or incomplete (defaulted or charged off). In the end, what matters to an investor is a completely paid off loan will always result in a return for the investor and an incomplete loan will always result in some loss to the investor. From the looks of the count of complete or incomplete loans, it looks like most loans in the end will be completed.

What would be more enriching is to get to know about the investor's situation. Of course an investor would want to the best possible return but there are factors like how soon an investor would like all his money receive back all his money including earnings. Therefore, we should analyze with the term being 12 months if an investor wants everything returned as soon as possible. Or what if the investor is very emotional and anxious about losses, then we should probably calculate the best possible return from maybe loans A, B, and C even though loans D and E offer the highest average return. These are variables that needs to be taken into account to see how we can best diversify the choosing of the loans for best returns suitable for each investor's unique situation.