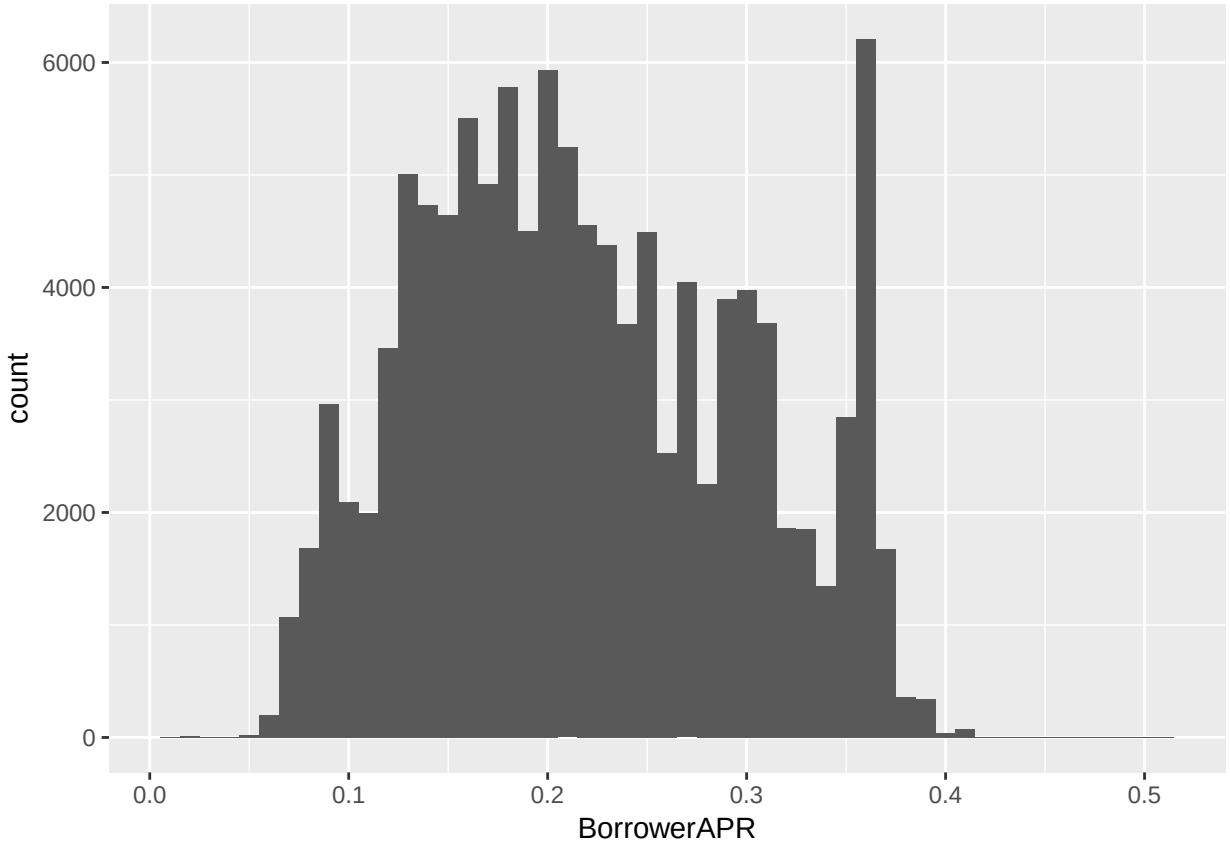
Loan Analysis by Ryan Neal

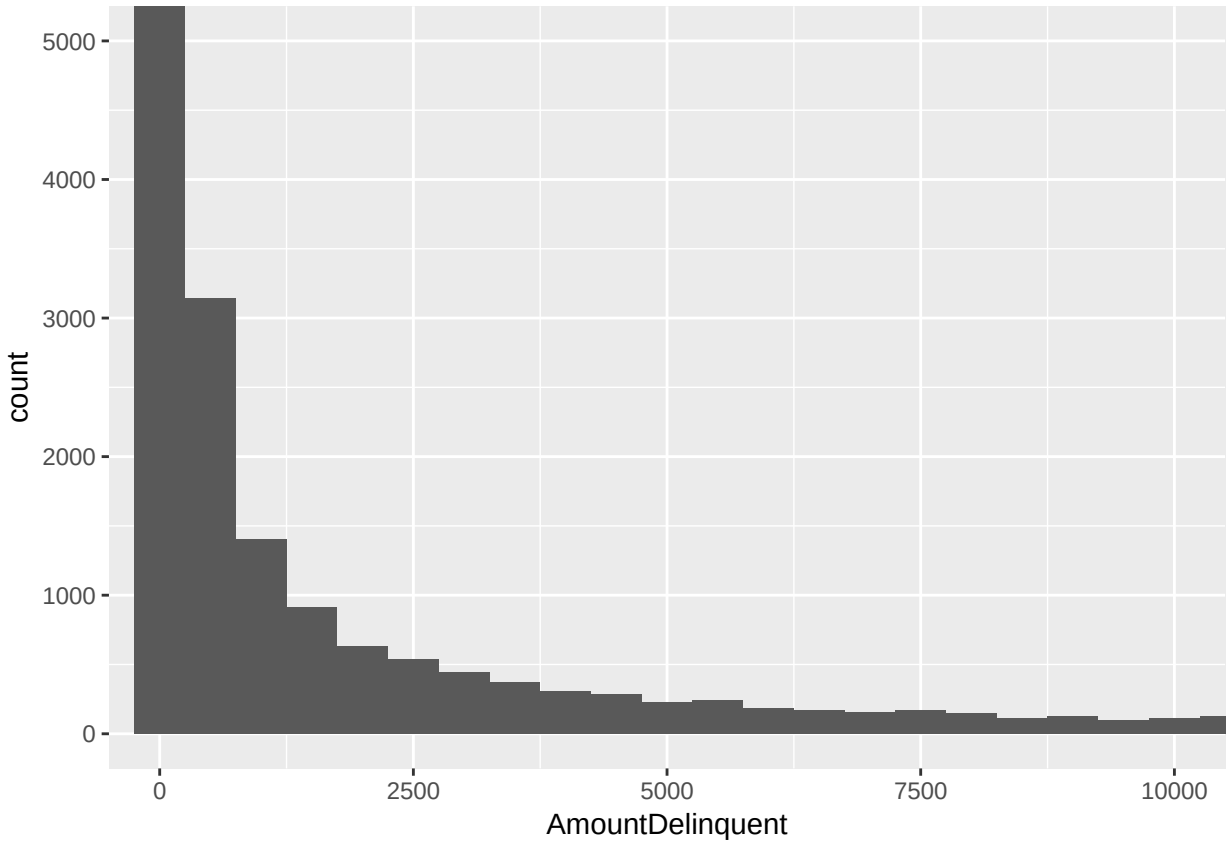
This project analyzes the prosper loan data set. In particular, I examine how multiple variables affect BorrowerAPR. The dataset contains 113937 observations of 87 variables.

Univariate Plots Section

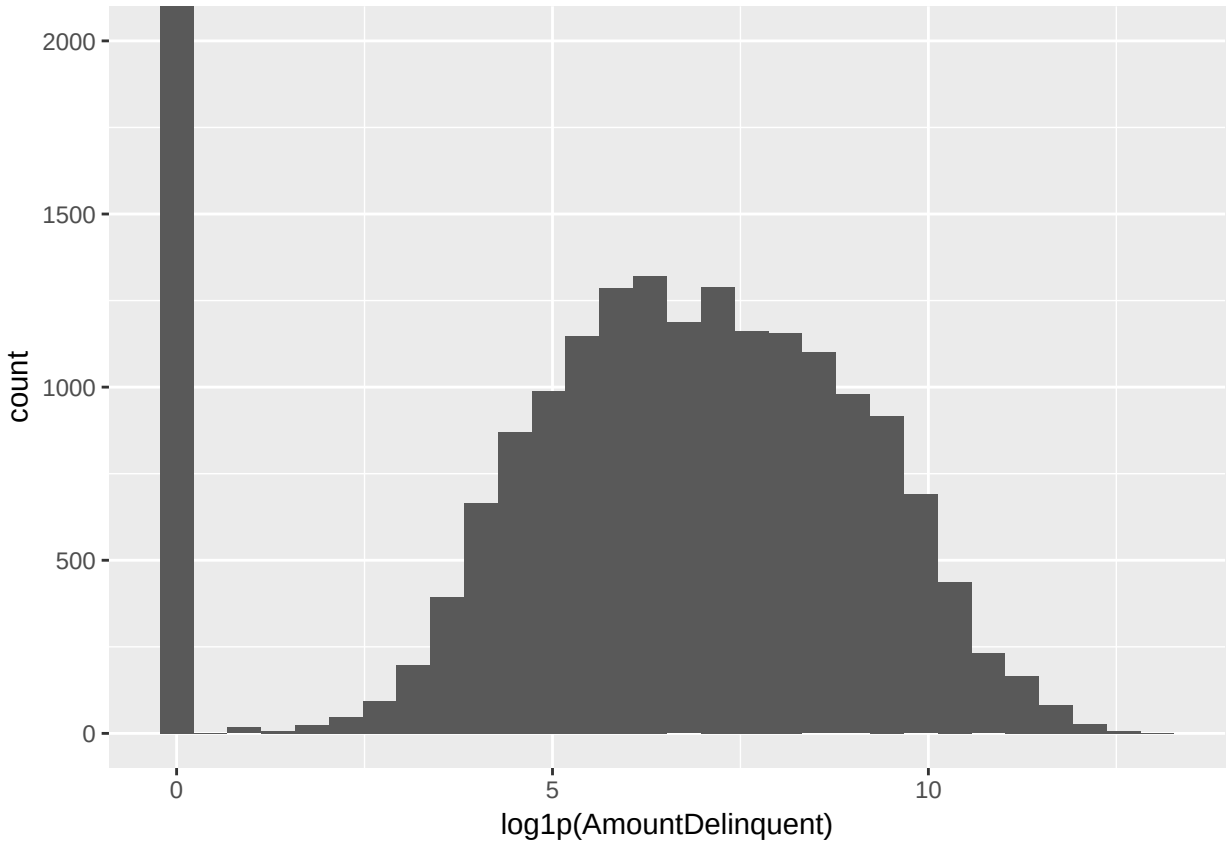
First, I want to look at different variables to better understand their distributions. As previously mentioned, I hope to explore the relationship of these variables with BorrowerAPR. Thusly, it makes since to first examine BorrowerAPR.



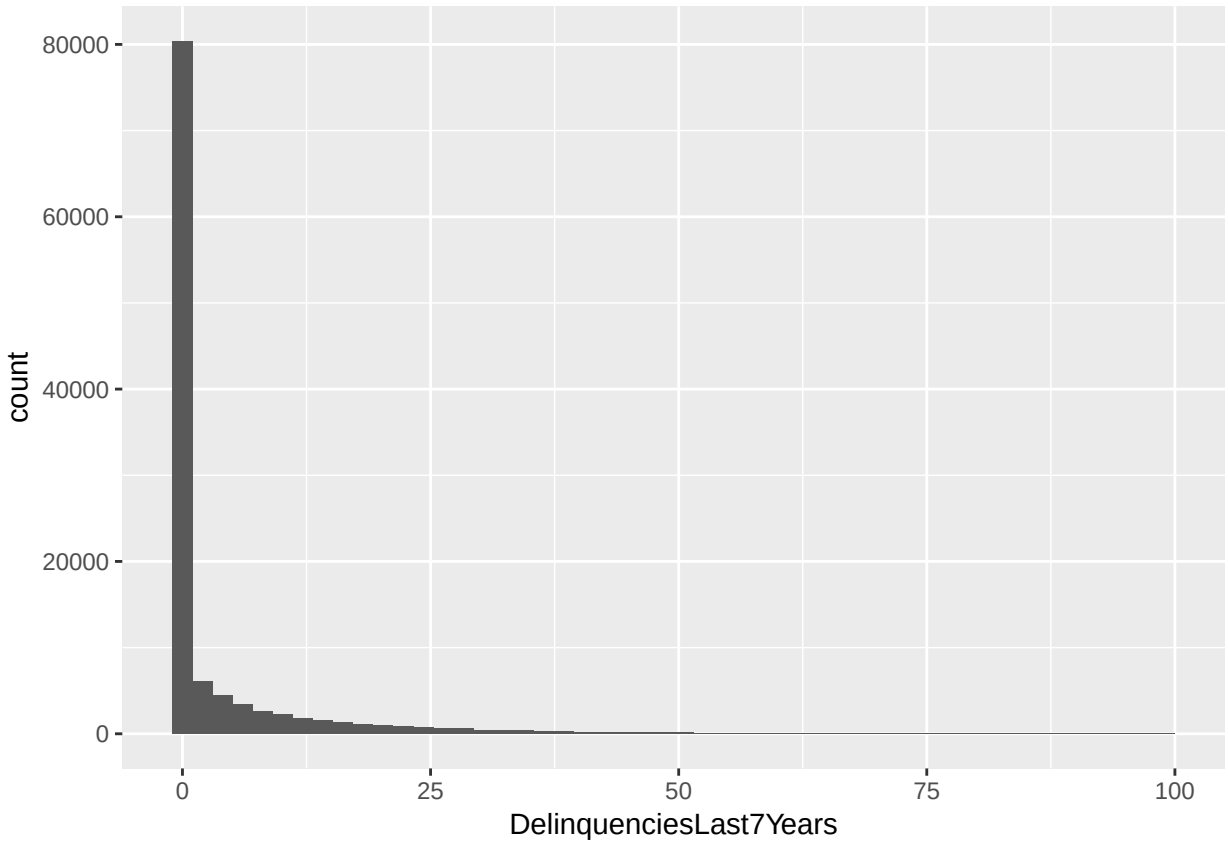
The distribution of BorrowerAPR is approximately normally distributed with a spike around 3.5. It appears there is large amount of people who receive a high-interest rate. Next let us examine Amount delinquent.



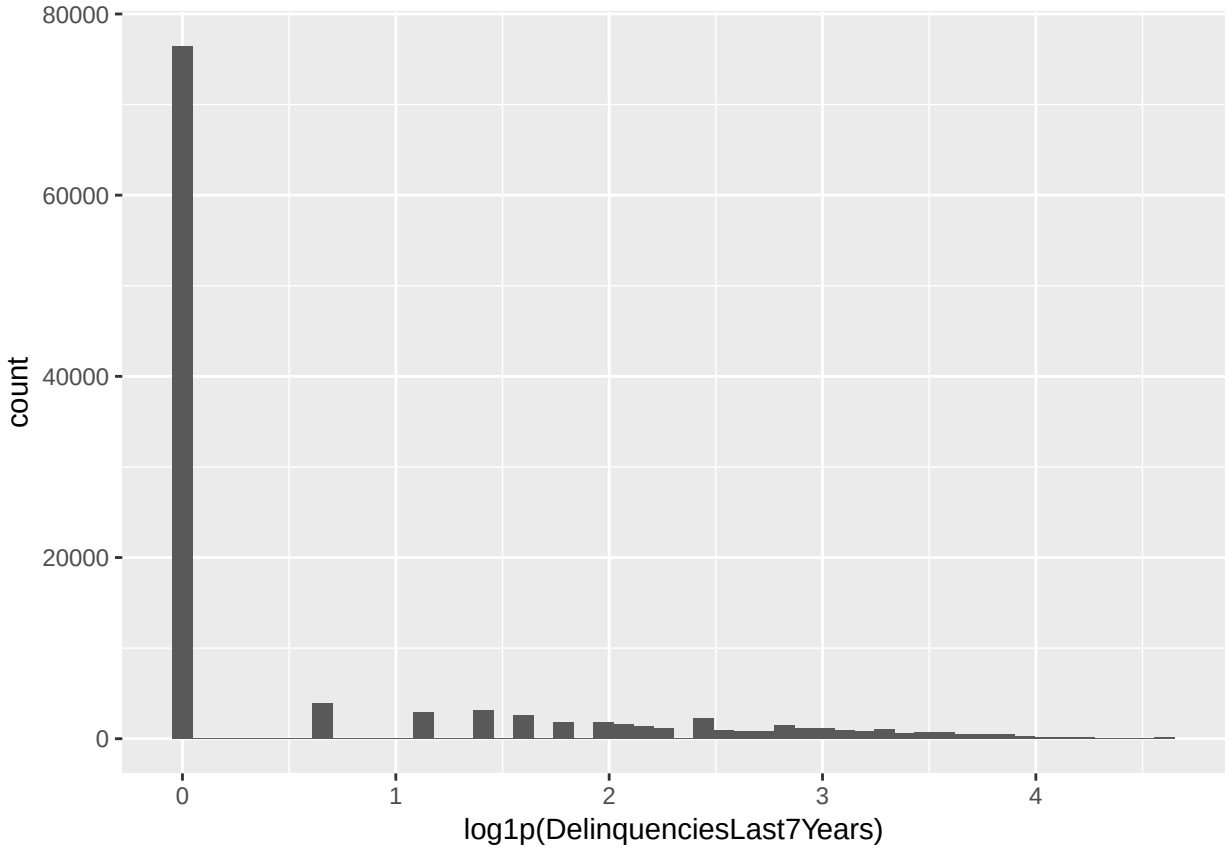
The vast majority of lenders seem to have no delinquencies. Let's look at a log transformation to be sure.



This is interesting. The log transformation of amount delinquent is normally distributed. We may be able to use this in a model later. Let us now see if there is a different pattern between delinquencies and delinquencies in the last 7 years.



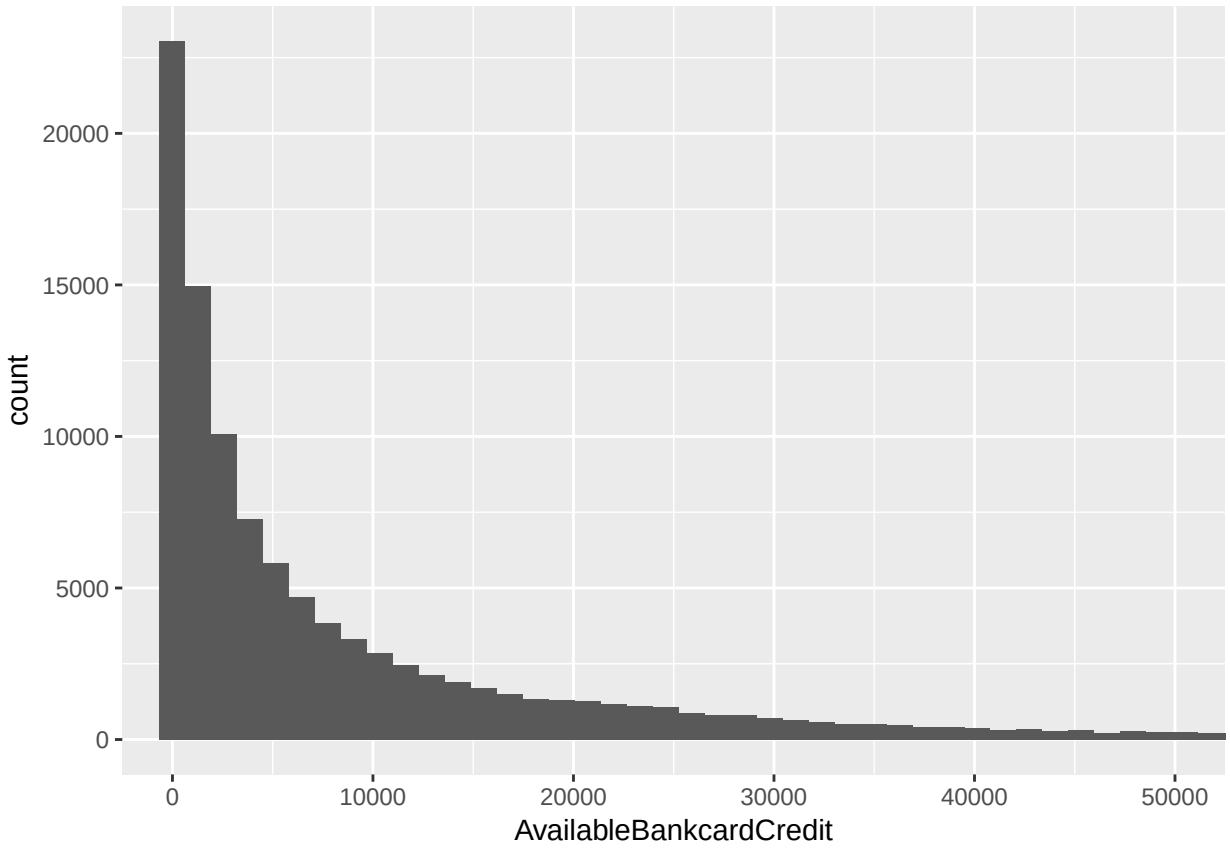
Delinquencies over the last 7 years is even more skewed than amount delinquent. It appears most people do not have delinquencies. Let us log tranform and try again.



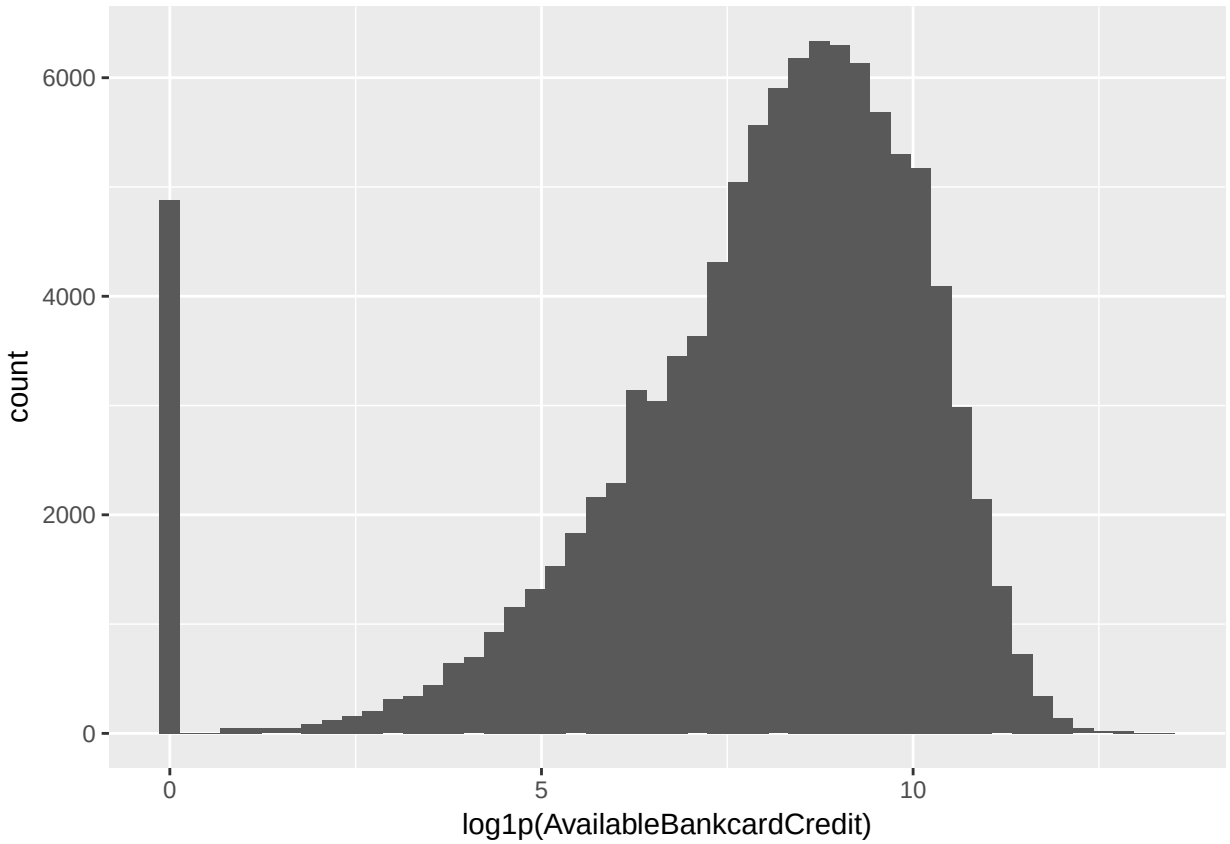

```
##
##  Pearson's product-moment correlation
##
## data:  clean_prosper$DelinquenciesLast7Years and clean_prosper$AmountDelinquent
## t = 78.217, df = 106310, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2275783 0.2389463
## sample estimates:
##       cor
## 0.2332703
```

This graph really drives home to point. Delinquencies over the last 7 years is more skewed than amount delinquent. We will keep that in mind going forward.

Now I am curious about total credit. Let us look at that distribution now.



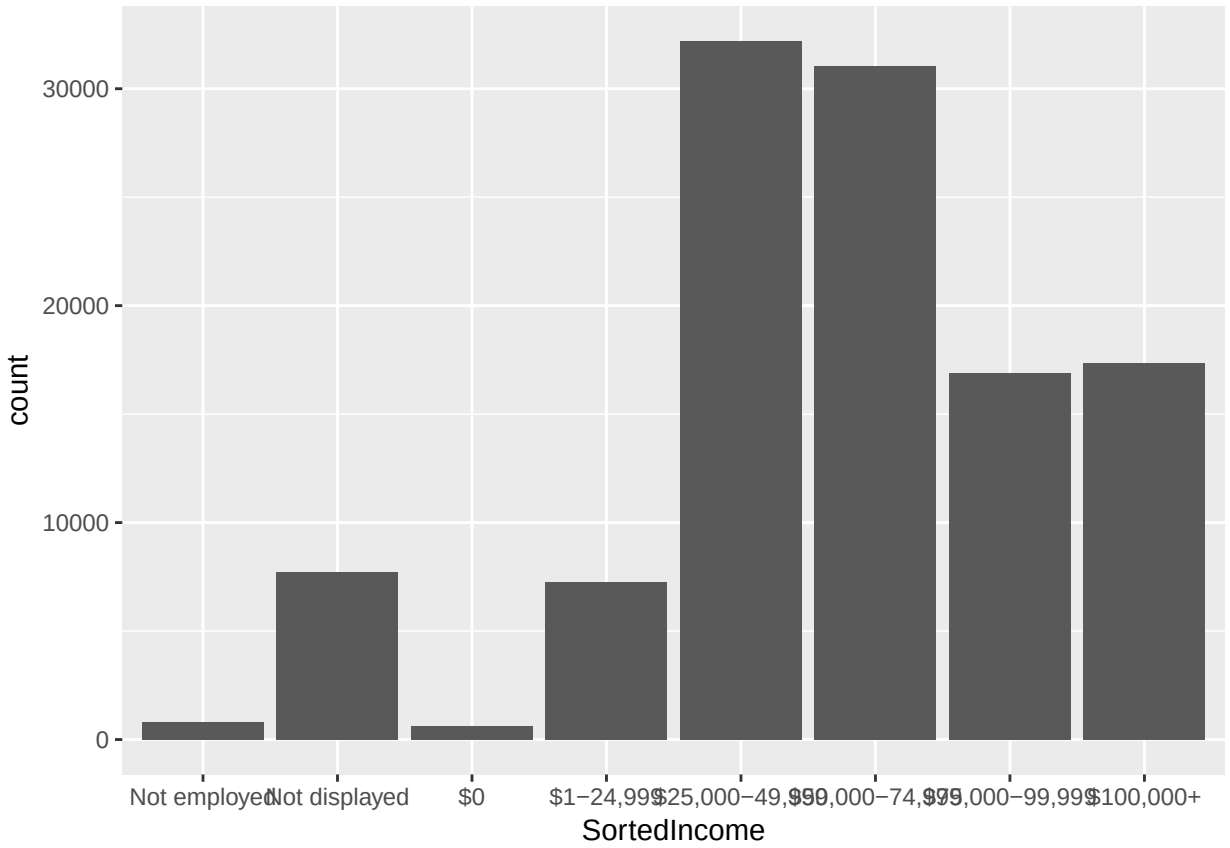
This is an interesting distribution. It drops precipitously around 0 and approaches a uniform distribution around 3000. Thus far we are seeing a lot of right-skewed distributions. Let us examine the log of bank credit next.



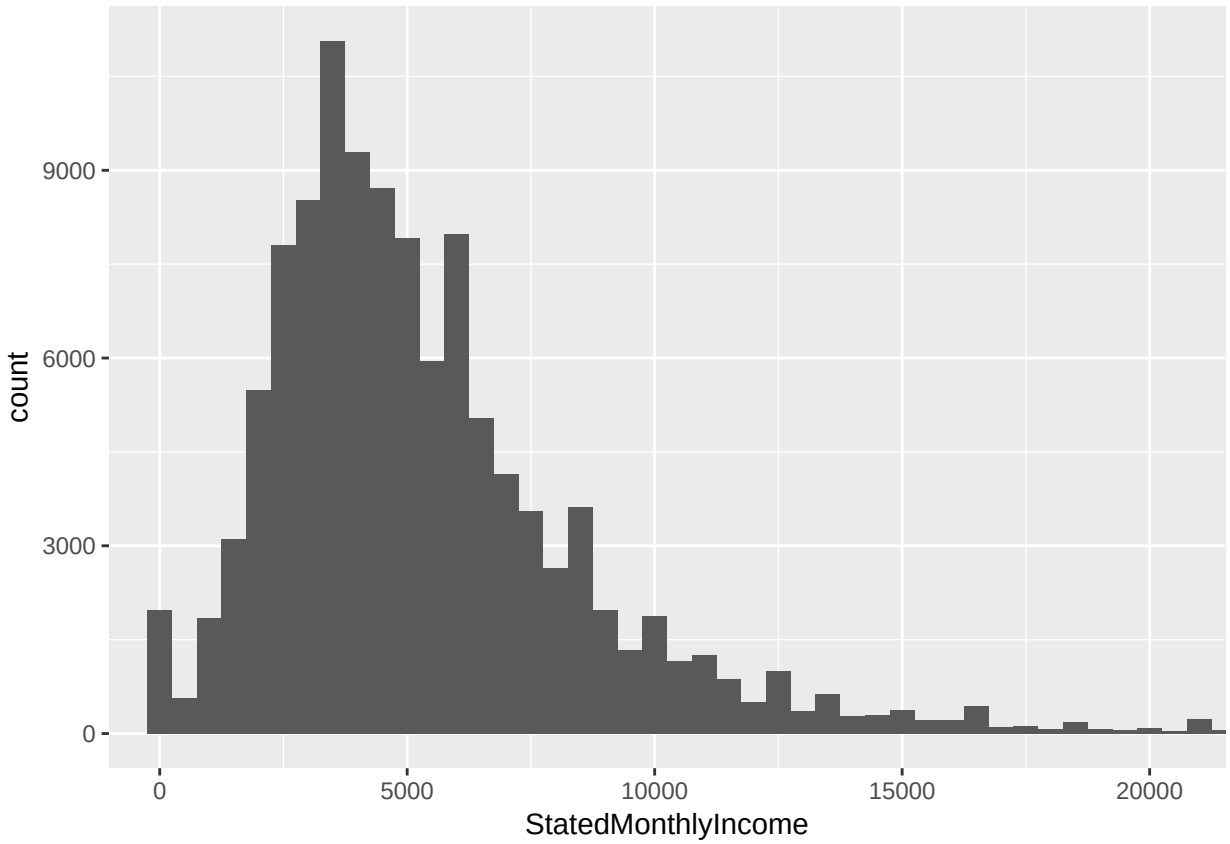
Now that is interesting! Bank credit appears to be a bi-modal distribution.A very large portion of the population has 0 available bank credit.Then there is a slow scale up before a sharp decline in credit.

Let us look at income range next

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

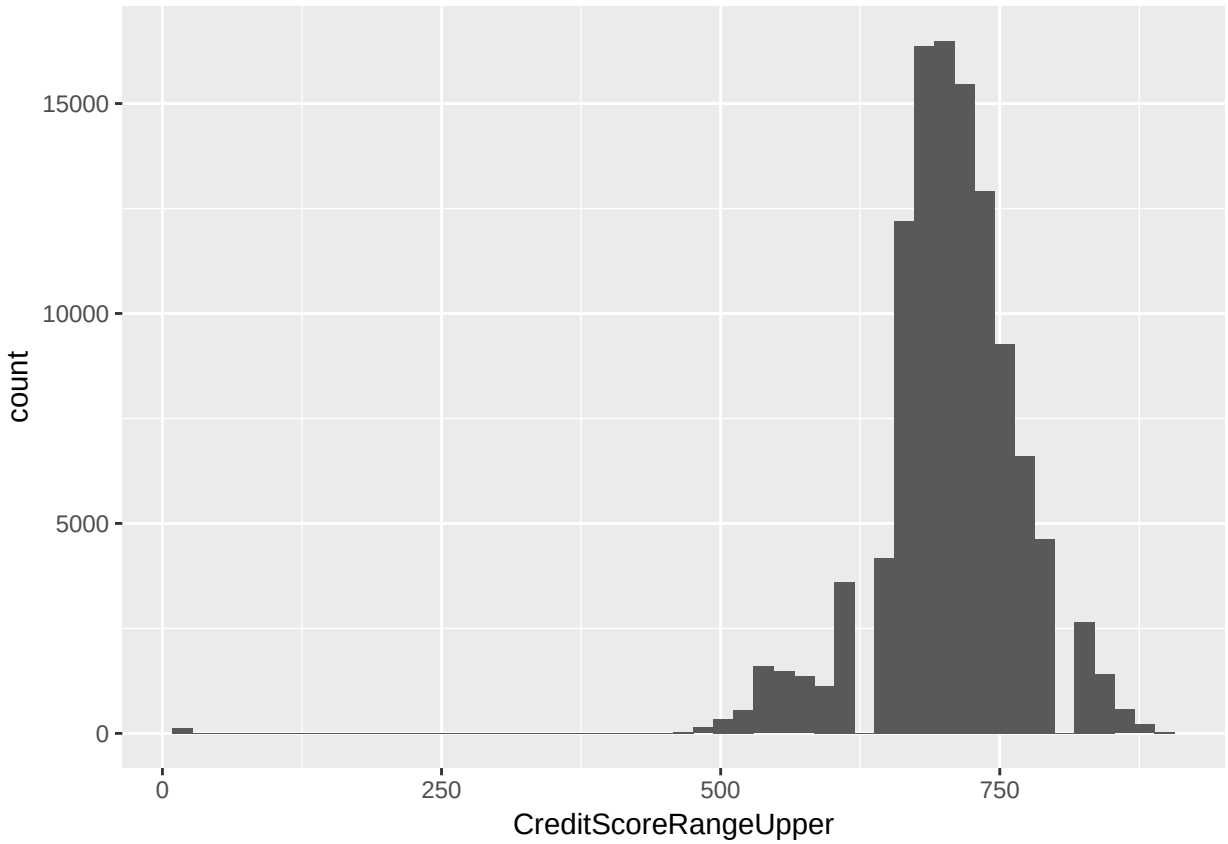


This is pretty interesting. It appears there are more people with higher incomes than I would have expected. For example, there are more people with \$100,000 incomes than <\$25,000 incomes. Could this be a way the bank clientele differ from the general population? (We will not answer this question). I am curious if a more granular examination of income will be more normally distributed. Let us examine monthly stated income.



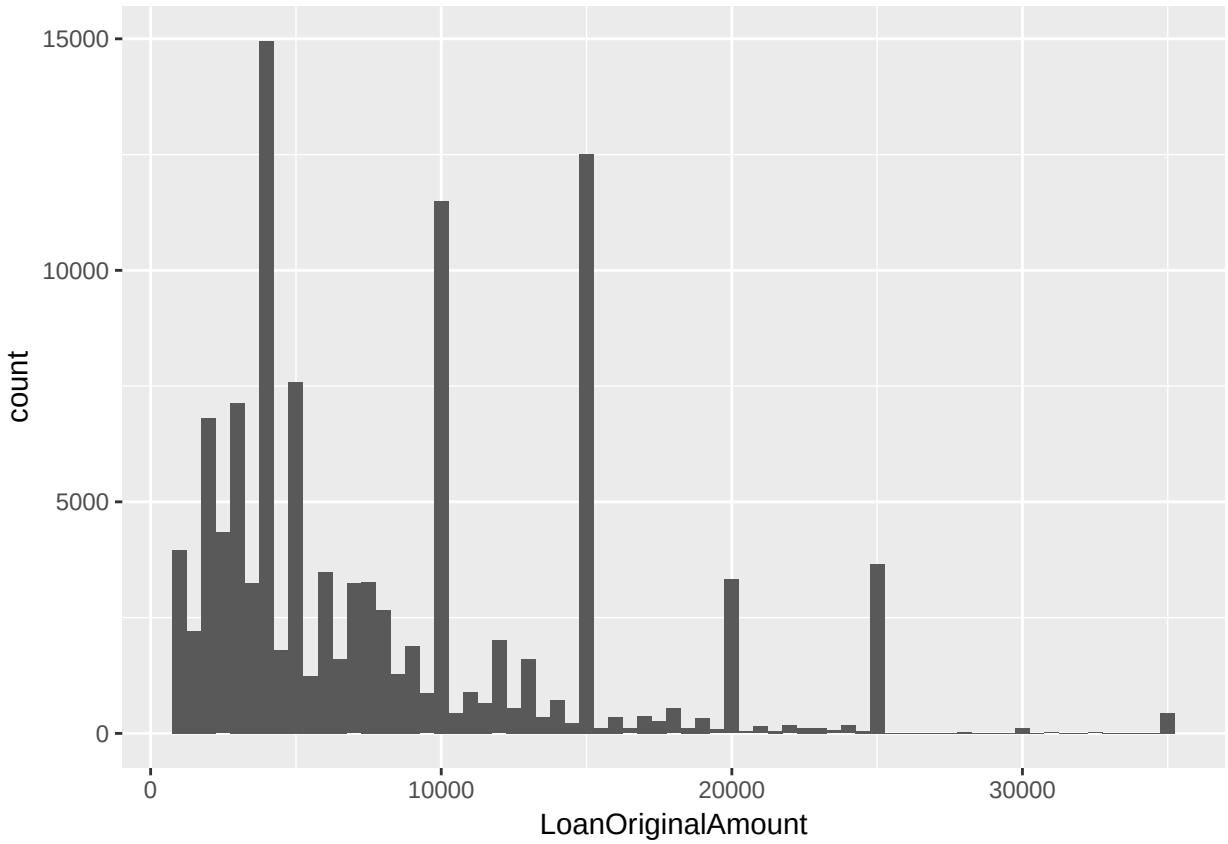
The data is once again right skewed. There are also some spikes. My guess is most lenders report hard numbers (e.g. \$5,000 or \$2,500), which causes spikes.

Let us now look at credit score

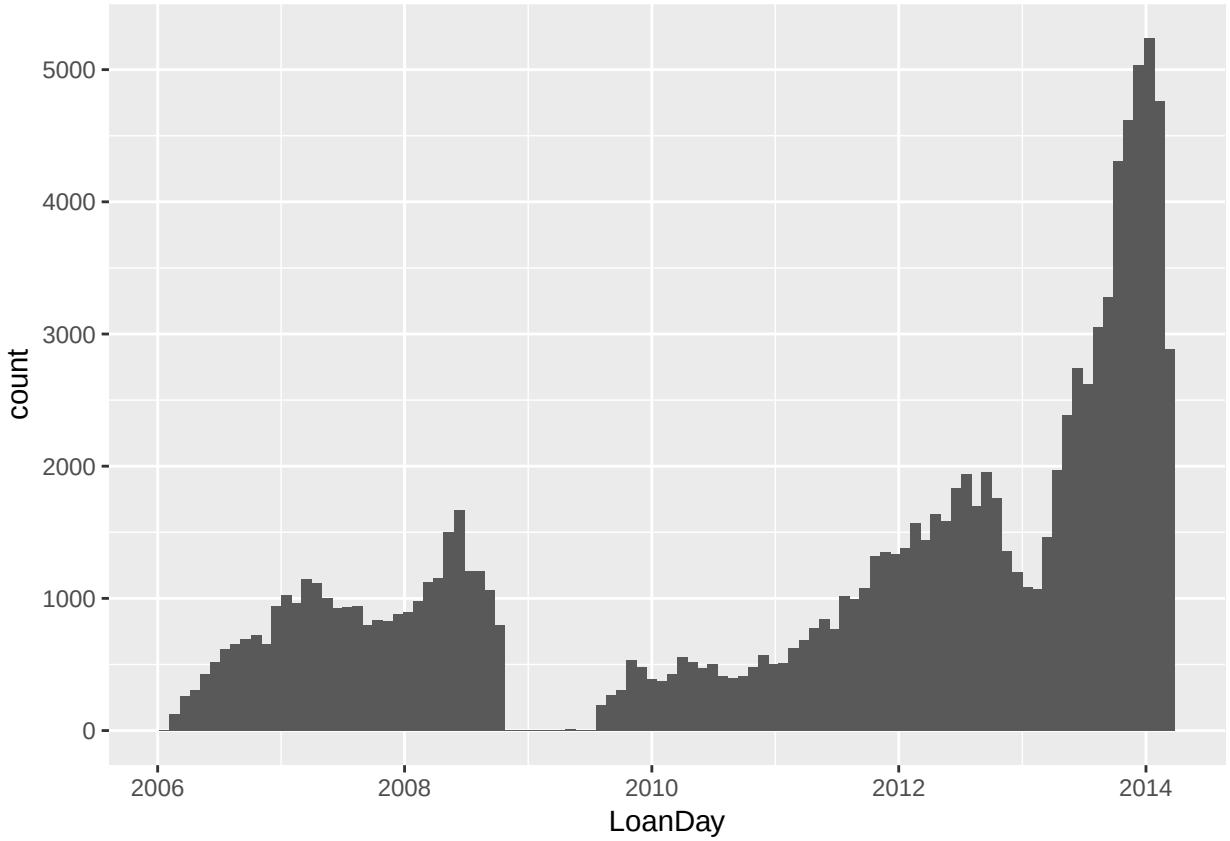


Once again, we have a cohort at 0 and then a normal distribution on the right. In this instance, that makes sense. I did not know credit could be 0. For the most part, it looks like credit scores are between 450 - 850

Now, let us check out the distribution for original loan amount



It looks like it is very common to take out loans in $5,000 increments. $4,000 is the most common and $15,000 is the next most common. Now, I want to see how loan count changes by year. I created a variables for loans per yer and loans per day for this.



This is not surprising, but it is illuminating. Loans dropped precipitously in 2009 - following the financial crisis. Loans reached their pre-crisis level in 2012 and increased until 2014.

Univariate Analysis

What is the structure of your dataset?

Many of the plots in this data set have a bi-modal distribution with many people having very bad credit and then a normal distribution of credit. One really disappointing characteristic of this data set is many features have so many null values that they can not be used.

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

I am most interested in predicting the interest rate, so BorrowerAPR.

What other features in the dataset do you think will help support your
I will examine how credit score, income, past delinquencies, and loan date affect BorrowerAPR

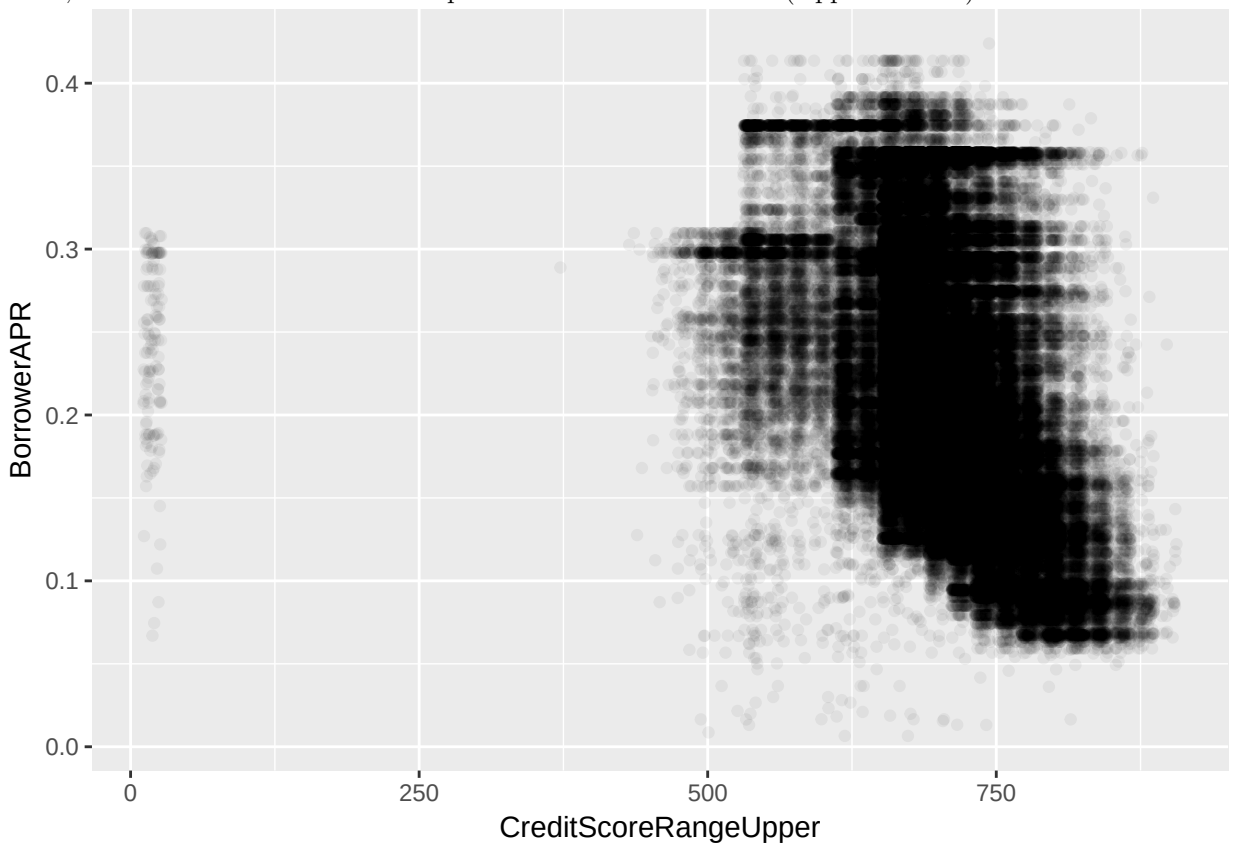
Did you create any new variables from existing variables in the dataset?

I created new variables for loan year and loan day. I also created a variable, SortedIncome, that creates a factor from IncomeRange.

Of the features you investigated, were there any unusual distributions?
Oh yes. As touched upon earlier, I expected a normal distribution on most variables. Instead, it appears as if there are a large number of negative (and positive outliers). I will explore this more with bivariate analysis.

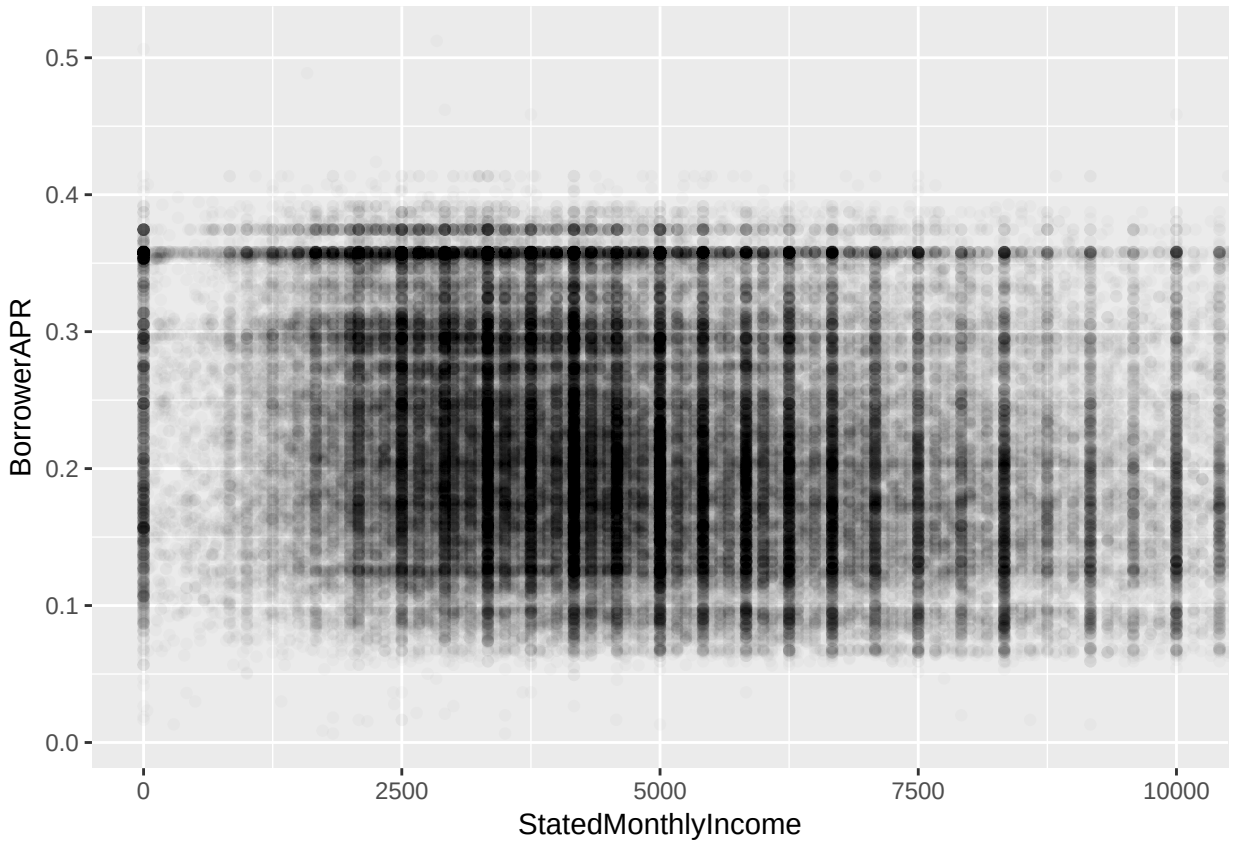
Bivariate Plots Section

First, I examined the relationship between Credit Score (Upper bound) and Borrower APR.



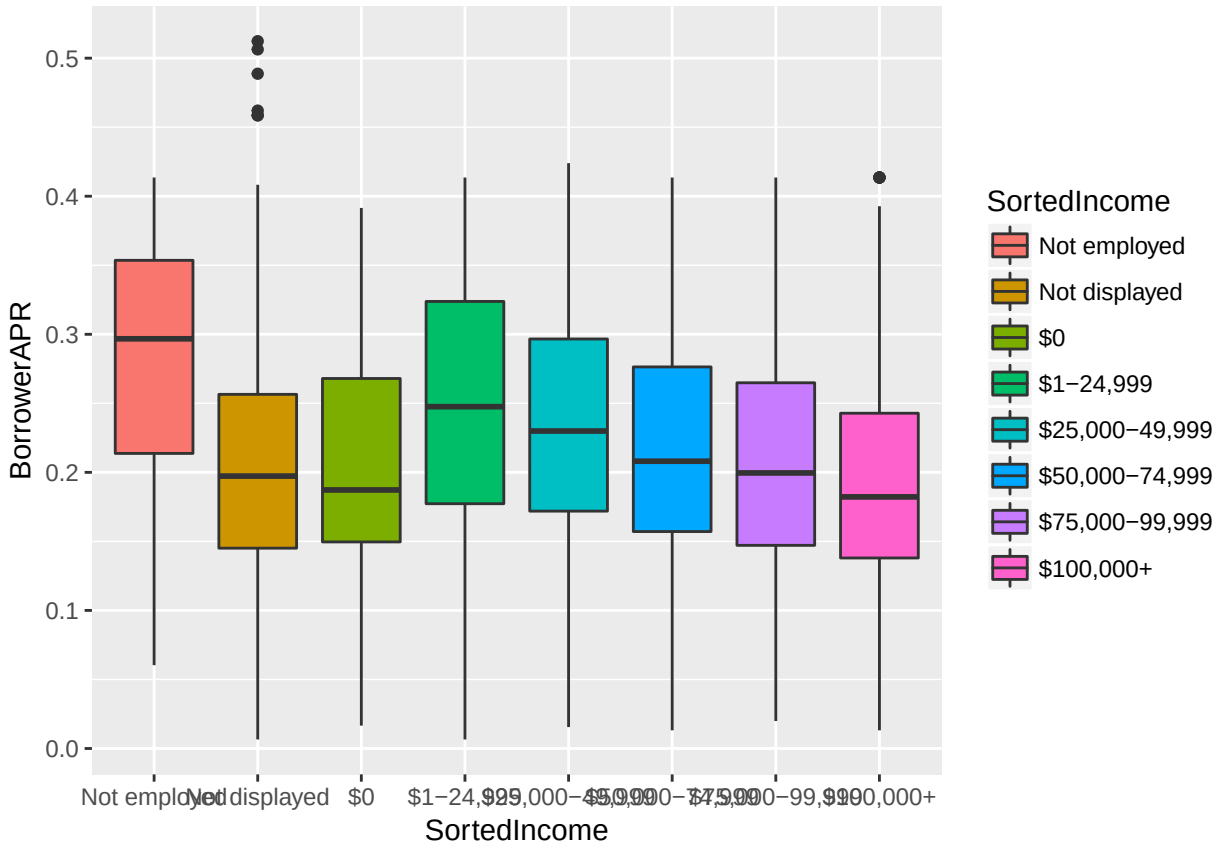

```
##
##  Pearson's product-moment correlation
##
## data:  clean_prosper$CreditScoreRangeUpper and clean_prosper$BorrowerAPR
## t = -160.21, df = 113340, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4344422 -0.4249487
## sample estimates:
##        cor
## -0.4297073
```

The correlation is around -.4, so there is a moderately strong relationship. However, there are a lot of people with high-credit scores who pay a high APR. I wonder if this is due to a lack of negotiation. Next, I examined stated monthly income to see how it affects BorrowerAPR.



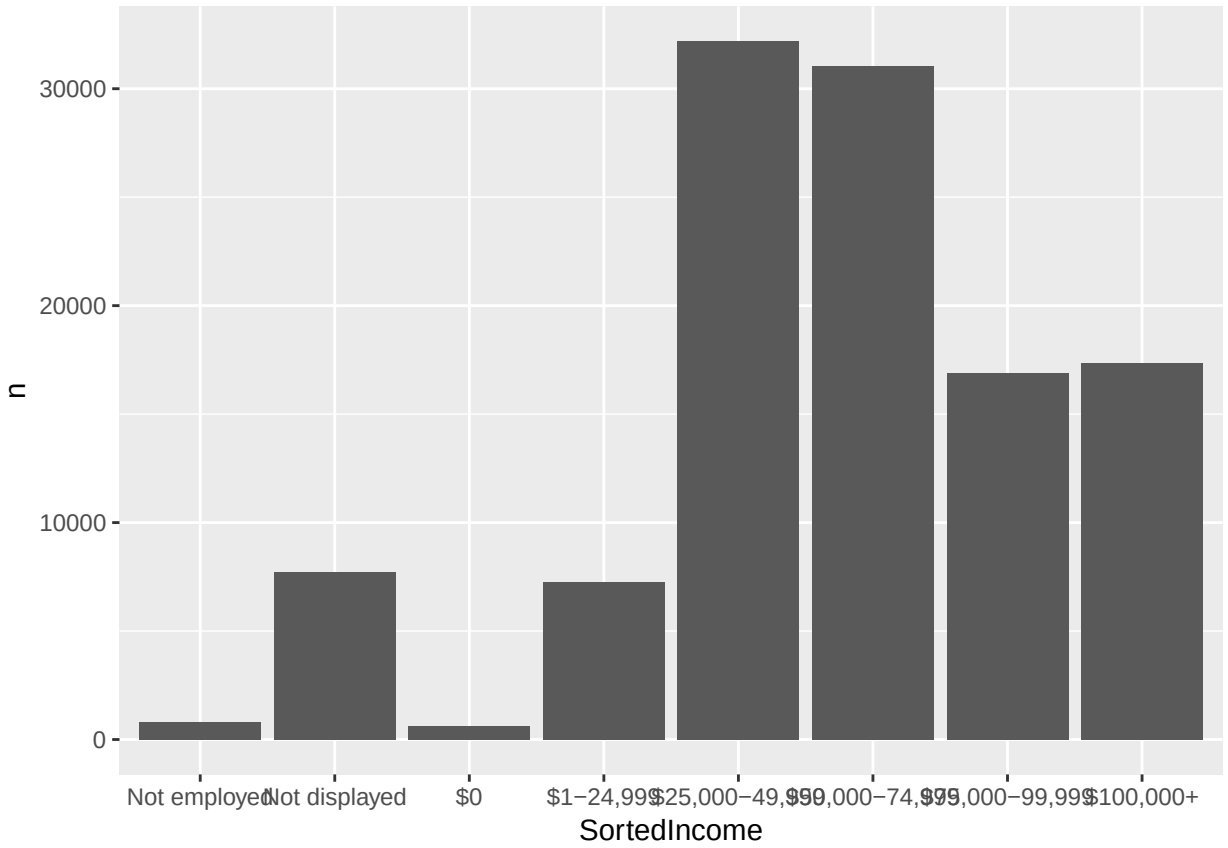

```
##
##  Pearson's product-moment correlation
##
## data:  clean_prosper$StatedMonthlyIncome and clean_prosper$BorrowerAPR
## t = -27.884, df = 113910, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.08810353 -0.07656794
## sample estimates:
##         cor
## -0.08233849
```

The relationship is not very strong. Let us check out income range. Perhaps that relationship is stronger.

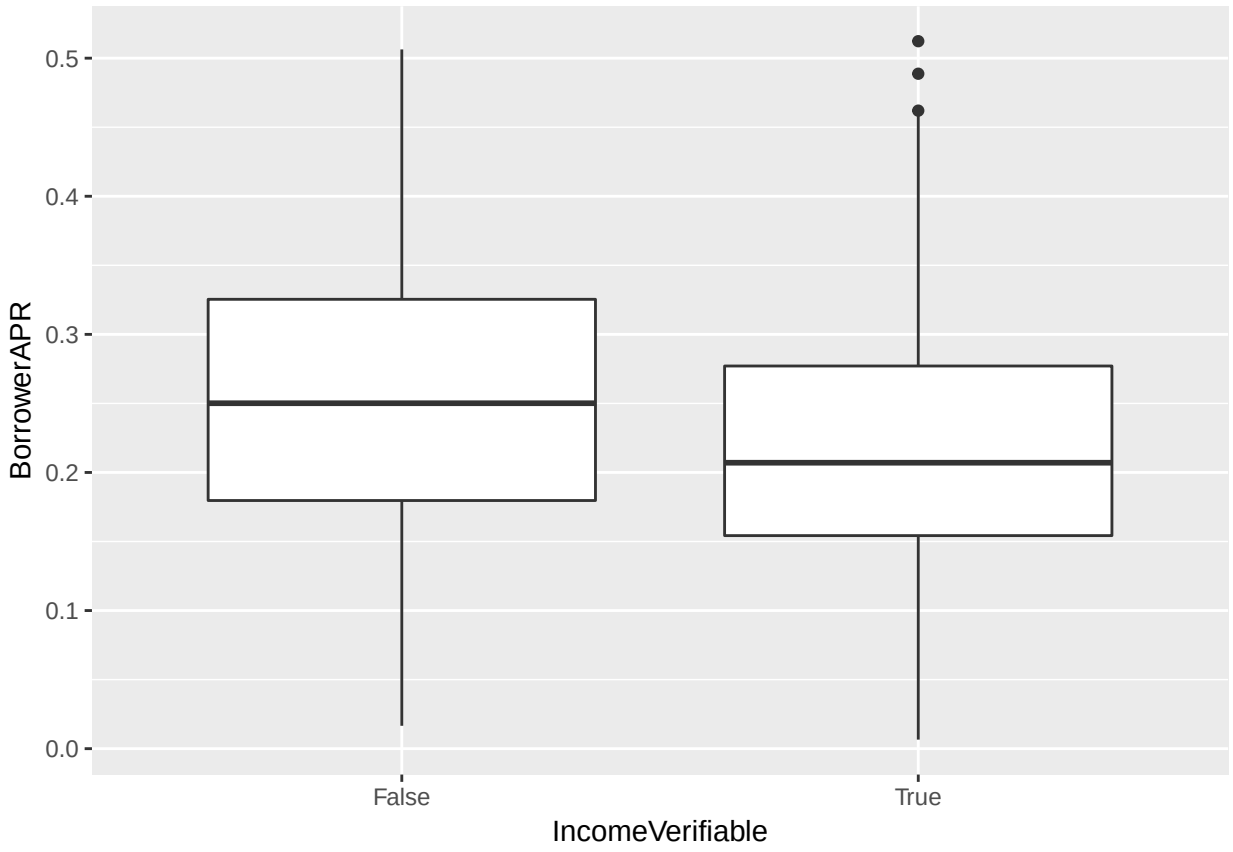


This looks a lot better. The weird thing here is $0 income has the lowest median APR. However, I as illustrated in the graph below there are far fewer loans for people with no income. Keeping in mind the small sample size, there appears a steady decreasing trend in borrowing APR as income increases. This is exactly what I would expect.

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

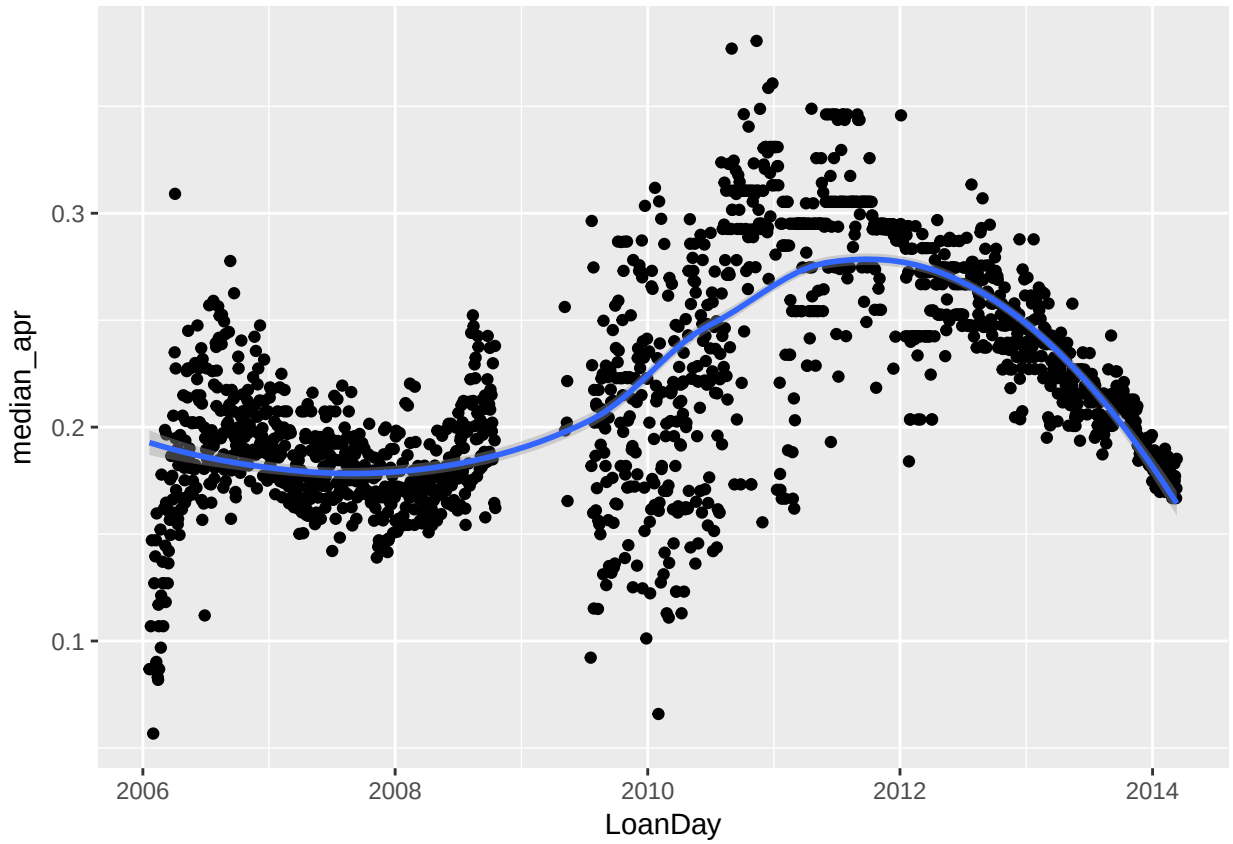


Finally, I am curious why income range and stated monthly income differ so much. Let us examine APR by verified income.

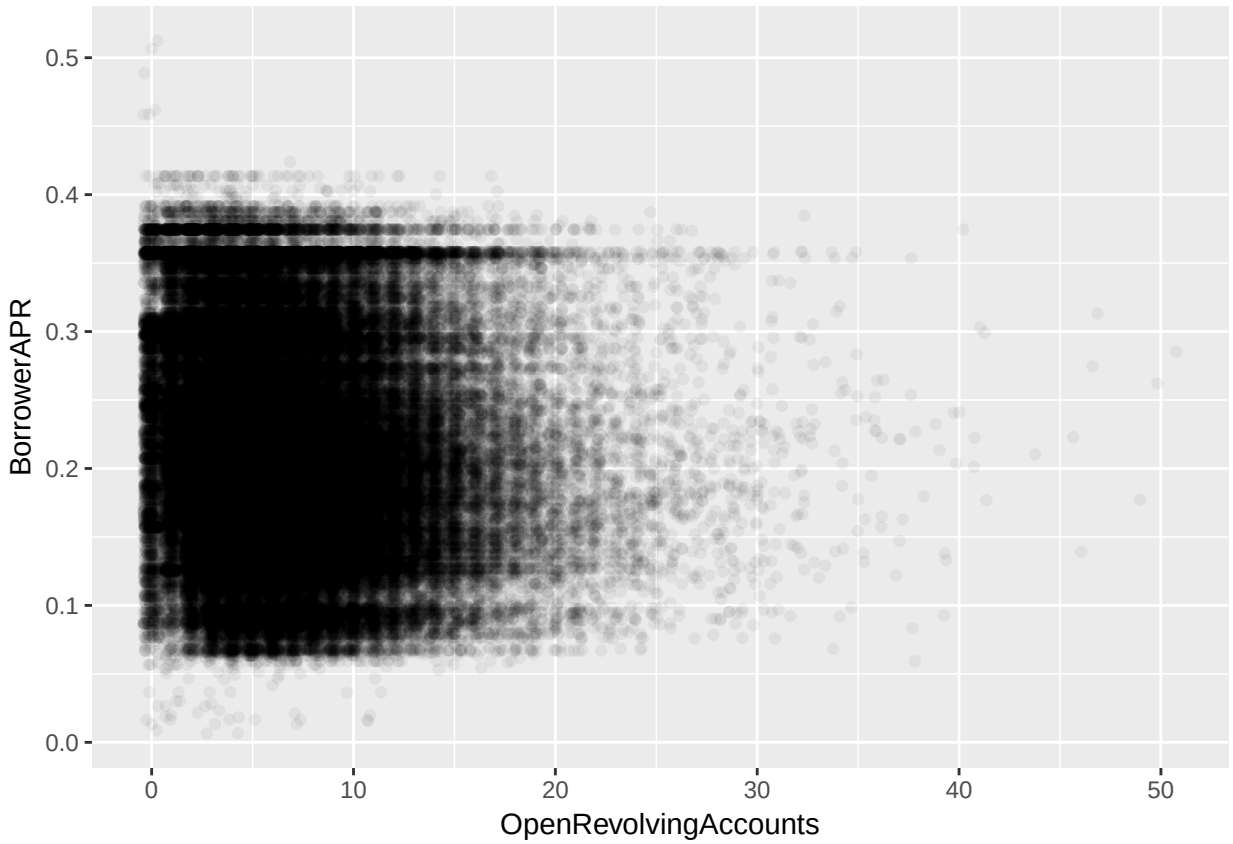


It does appear like verifying income decreases APR. Could it be the case that many of the high stated monthly incomes are unverified? Going forward we will use the income range as it seems more reliable.

Next I am curious how APR changes over time. My hypothesis is APR will increase following the financial crisis.

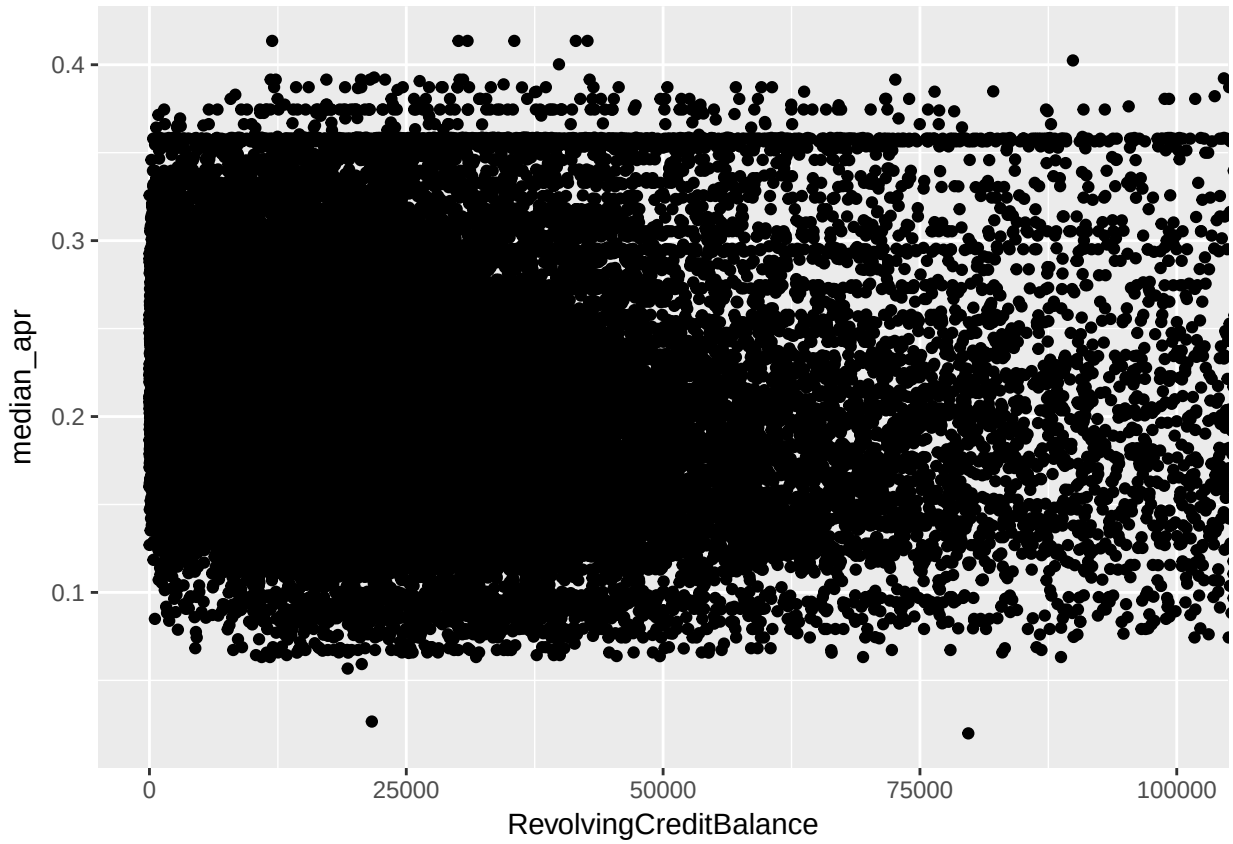


This seems to confirm our hypothesis. There is a steep increase in APR following the financial crisis. Now, I am curious how credit usage affects APR. It is also really interesting how tightly grouped the data is with the exception of 2010. Could it be the case that the bank made major business changes following the crisis and 2010 was a year of figuring things out?

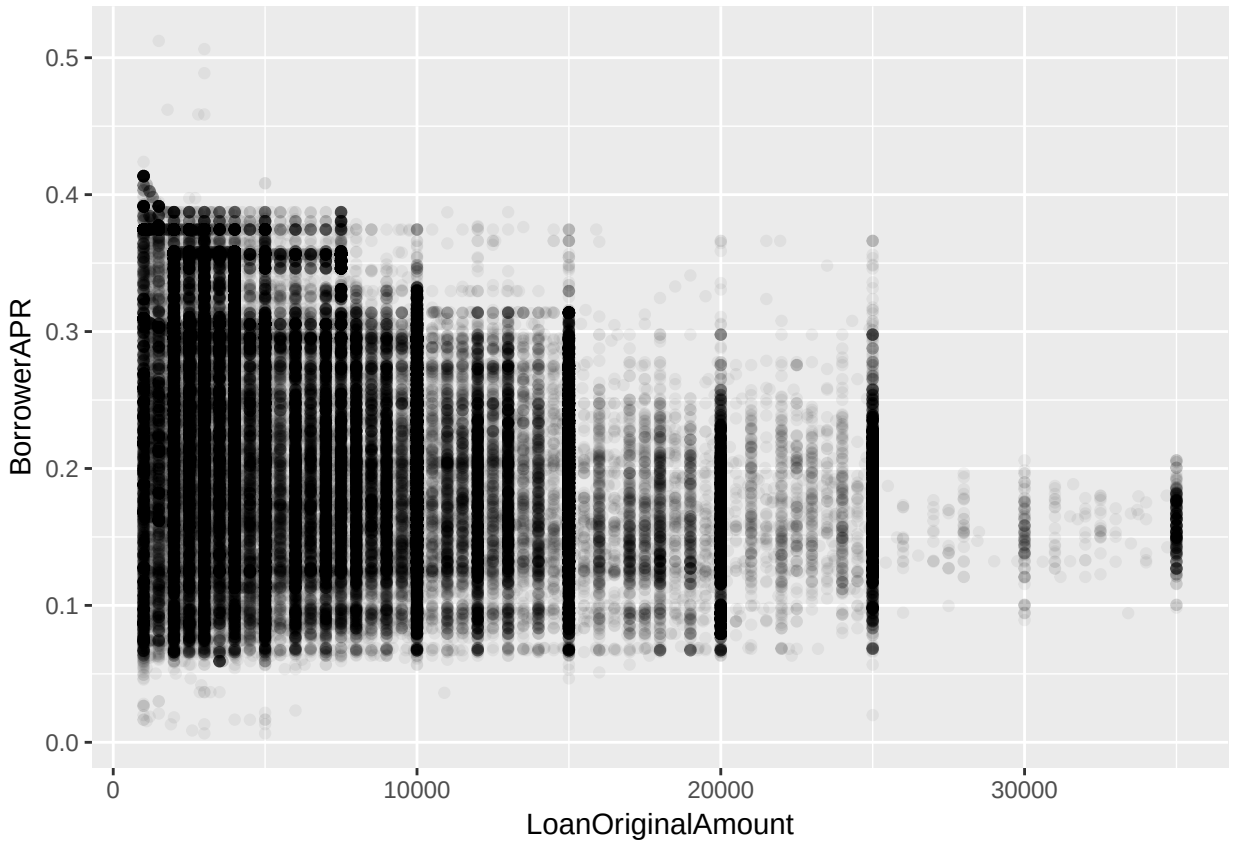



```
##
##  Pearson's product-moment correlation
##
## data:  clean_prosper$OpenRevolvingAccounts and clean_prosper$BorrowerAPR
## t = -37.422, df = 113910, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1159353 -0.1044620
## sample estimates:
##        cor
## -0.1102023
```

Not a strong relationship. Let us try account balance instead.

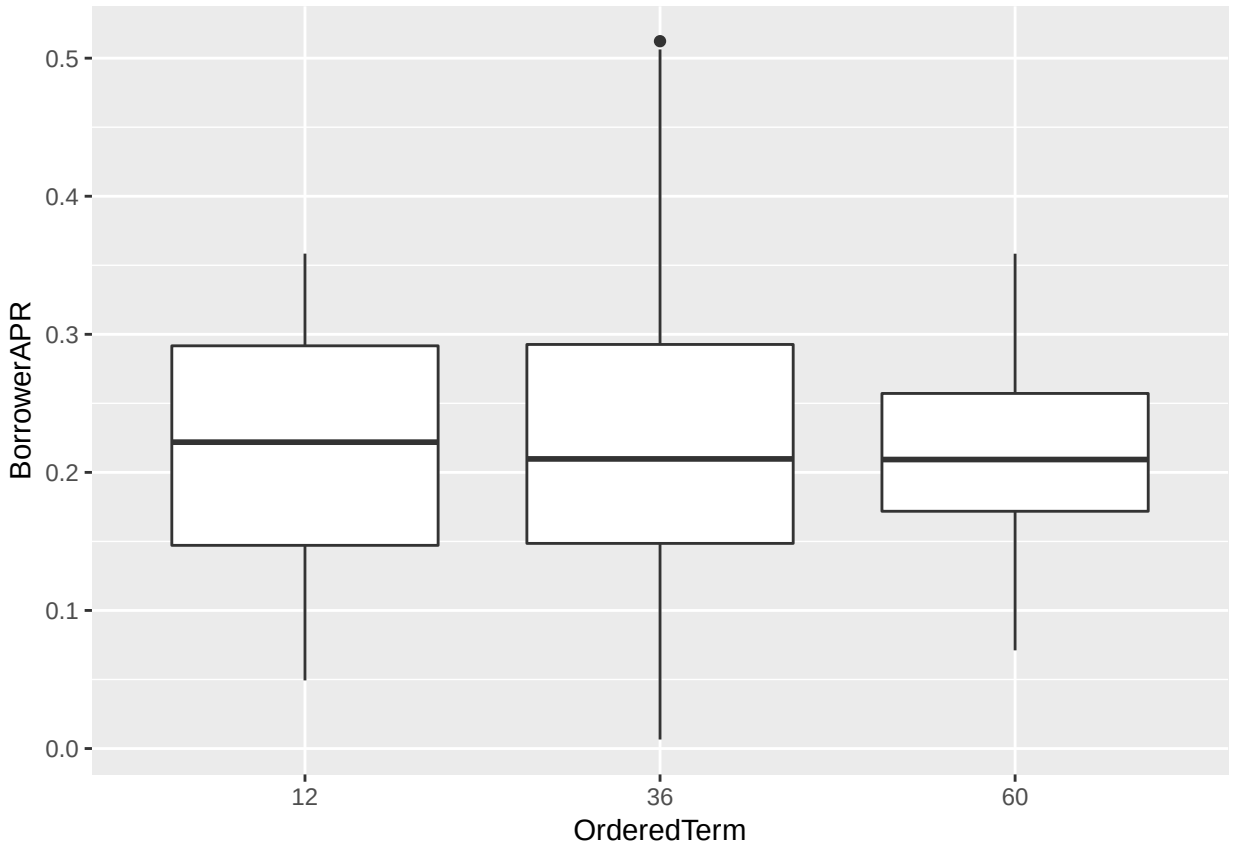


It does not appear like there is a much of a relationship between credit balance and APR. Let us check out how loan amount affects APR.

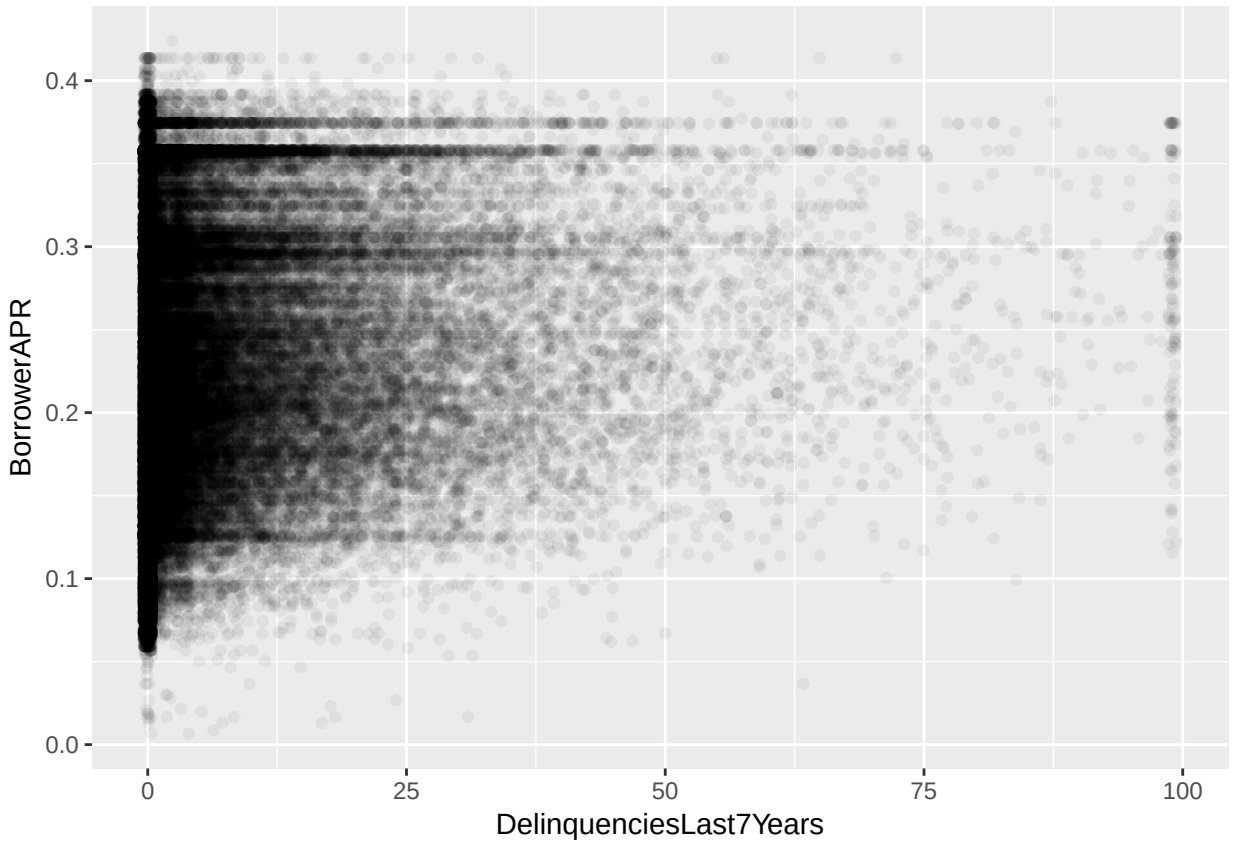



```
##
##  Pearson's product-moment correlation
##
## data:  clean_prosper$LoanOriginalAmount and clean_prosper$BorrowerAPR
## t = -115.14, df = 113910, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3280787 -0.3176752
## sample estimates:
##        cor
## -0.3228867
```

It looks like there is a moderately negative relationship. Higher loans tend to have lower APRs. Maybe it is also the case that longer terms have lower APRs.

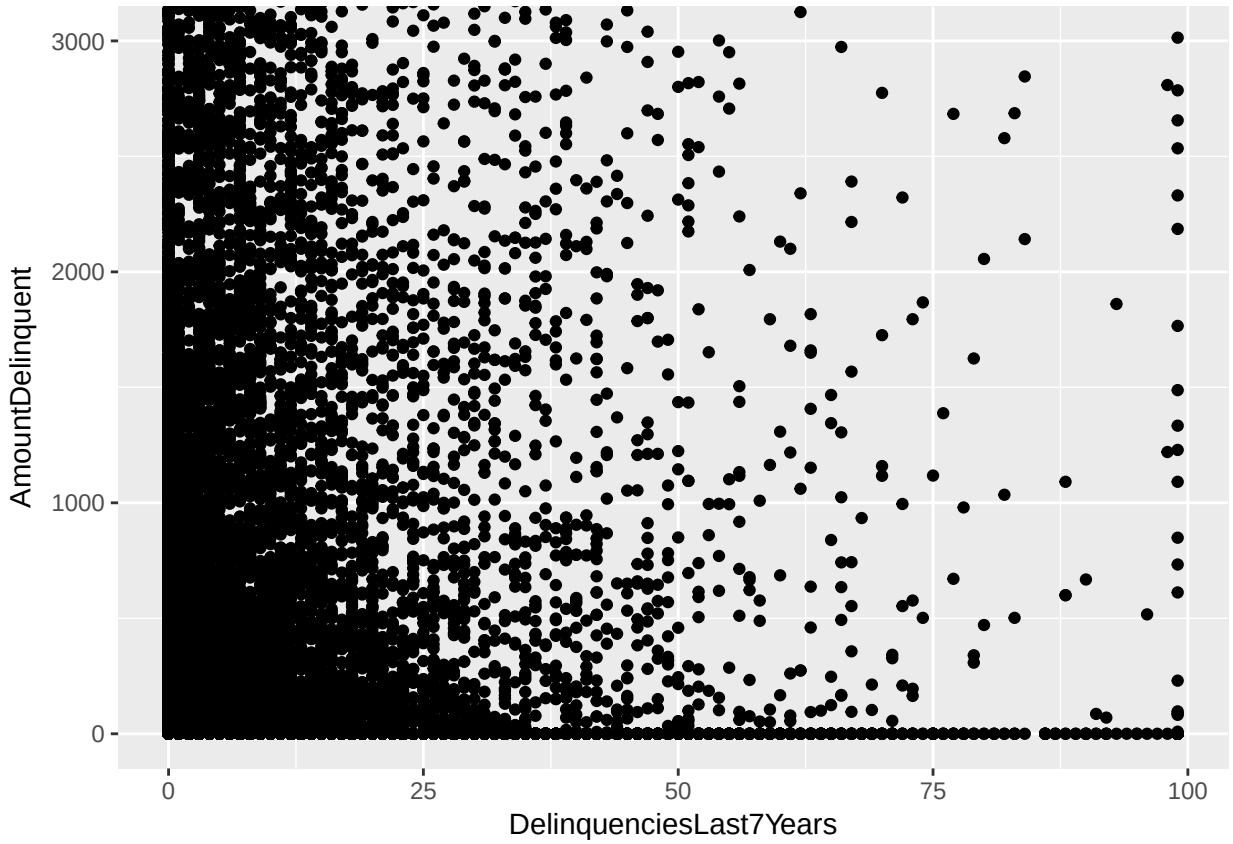


It does not look like it. 36 months is the most common term and has a lot of outliers. Finally let us see how delinquency affects APR



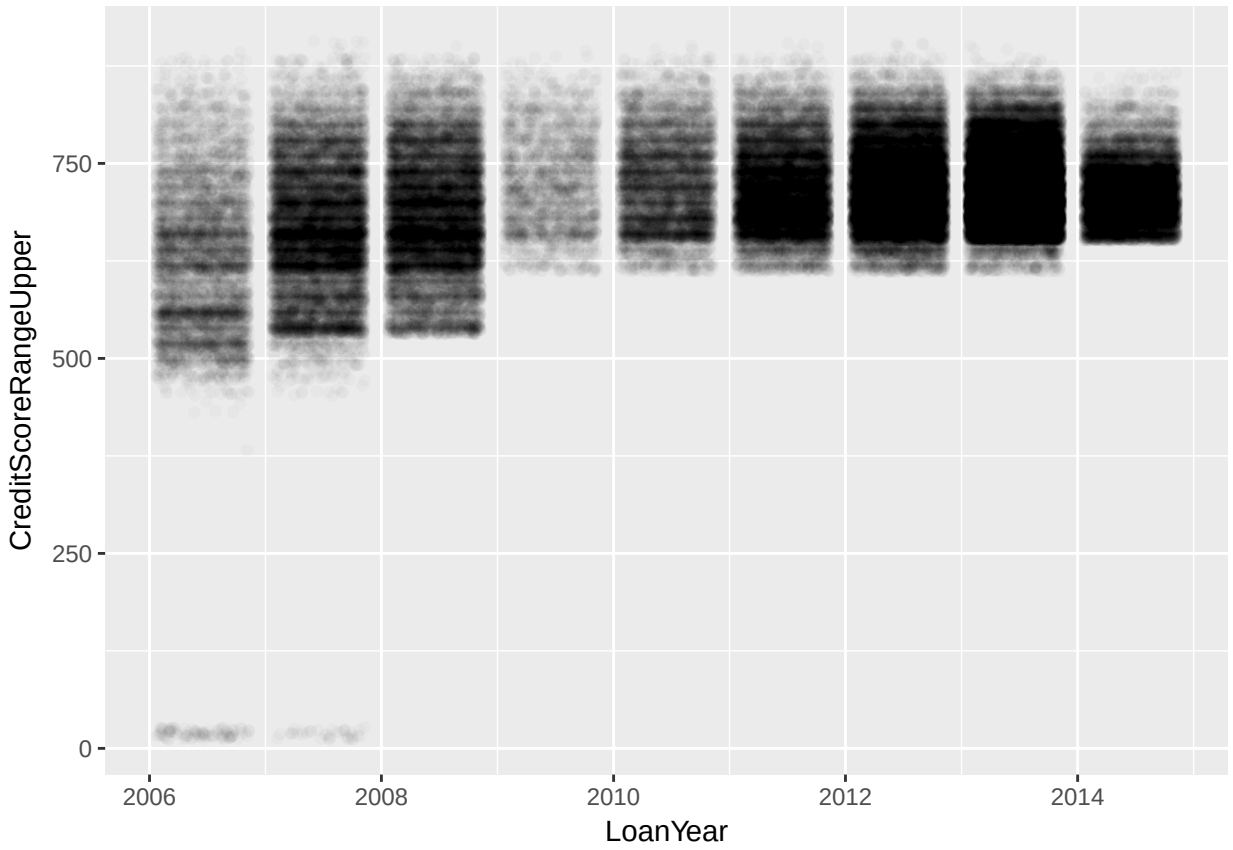

```
##
##  Pearson's product-moment correlation
##
## data:  clean_prosper$DelinquenciesLast7Years and clean_prosper$BorrowerAPR
## t = 55.251, df = 112940, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1565416 0.1678985
## sample estimates:
##       cor
## 0.1622254
```

This is not real strong. I am actually very surprised by this. Out of curiosity I wonder how count of delinquencies correlates with amount.



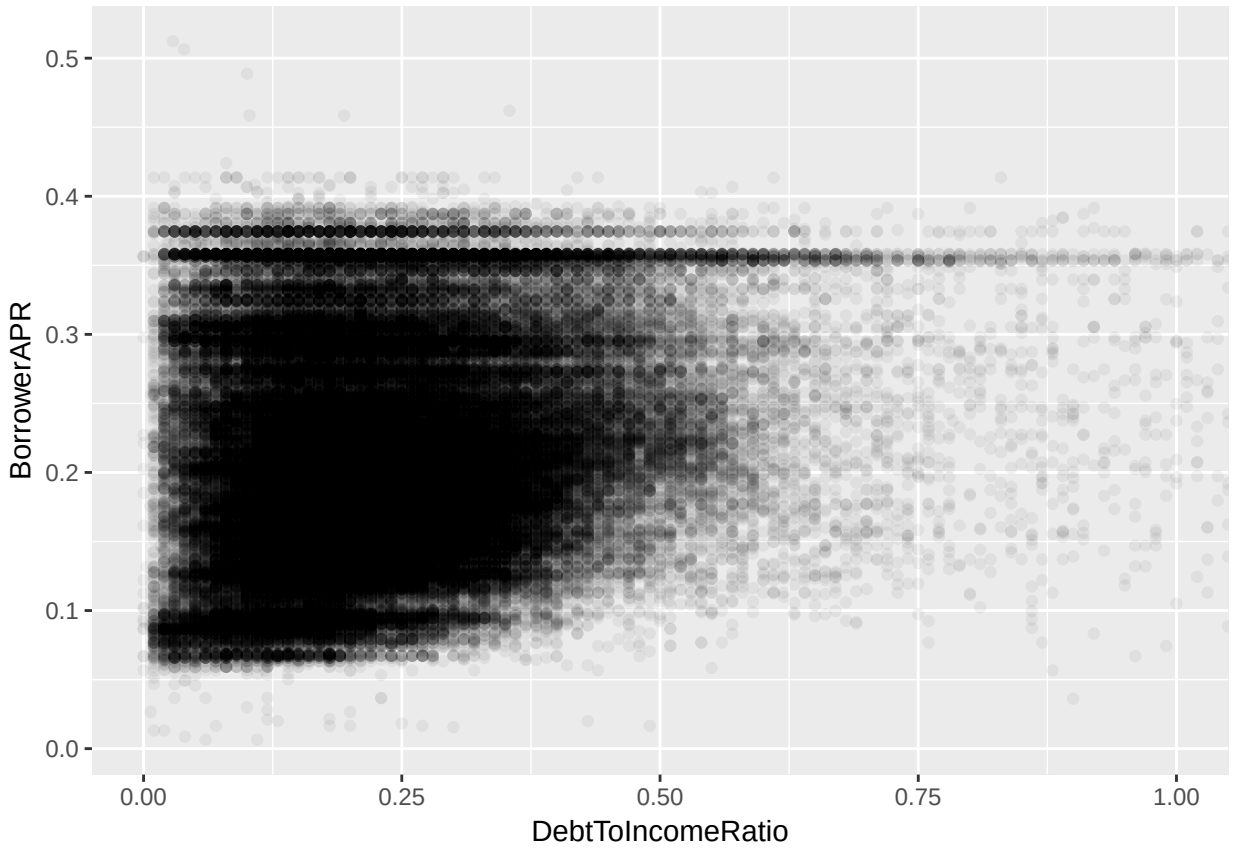

```
##
##  Pearson's product-moment correlation
##
## data:  clean_prosper$DelinquenciesLast7Years and clean_prosper$AmountDelinquent
## t = 78.217, df = 106310, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2275783 0.2389463
## sample estimates:
##       cor
## 0.2332703
```

This is interesting there is a relationship between total delinquences and amount delinquent, but it is not as strong as I would expect. Out of curiosity, I would like to examine the relationship between Credit Score and Loan Year.



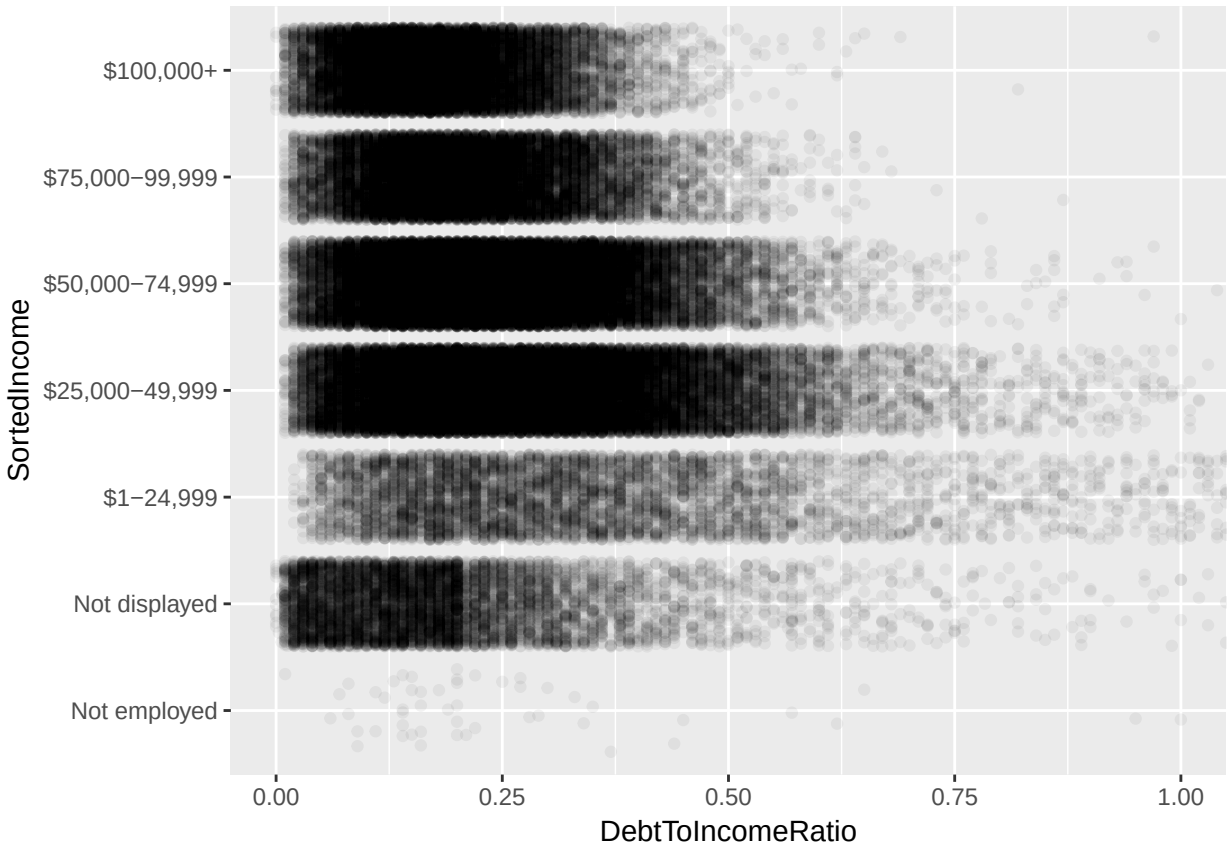
Interestingly, it seems like credit score rises over time. My guess is following the financial crisis, the company required higher credit scores.

```
ggplot(aes(x=DebtToIncomeRatio, y=BorrowerAPR),
       data=subset(clean_prosper, !is.na(DebtToIncomeRatio))) +
  geom_jitter(stat="identity", alpha=1/20) +
  coord_cartesian(xlim=c(0, 1))
```



Finally, I examined the Debt to income ratio versus APR. This is an interesting relationship. Perhaps not surprisingly individuals with debt as a higher percentage of their income receive higher interest rates.

```
ggplot(aes(x=DebtToIncomeRatio, y=SortedIncome),
       data=subset(clean_prosper, !is.na(DebtToIncomeRatio))) +
  geom_jitter(stat="identity", alpha=1/20) +
  coord_cartesian(xlim=c(0, 1))
```



Out of curiosity, I compared the relationship between debt to income and income. In general, higher income individuals have a lower percentage of debt compared to income.

Bivariate Analysis

Talk about some of the relationships you observed in this part of the project

In this section, we noticed some interesting relationships between variables and APR. In particular, we noticed APR varies greatly based on macroeconomic trends (i.e. APR changes based on year). In addition, APR also varies based on individual differences. Higher income individuals generally have a lower APR. We also learned that verifying income can lower APR.

However, we also learned some things we might imagine affect APR do not. For example, neither open revolving accounts or account balance have a large impact. Term length also did not have much effect. Whether delinquencies or loan amount affected APR was not conclusive

Did you observe any interesting relationships between the other features

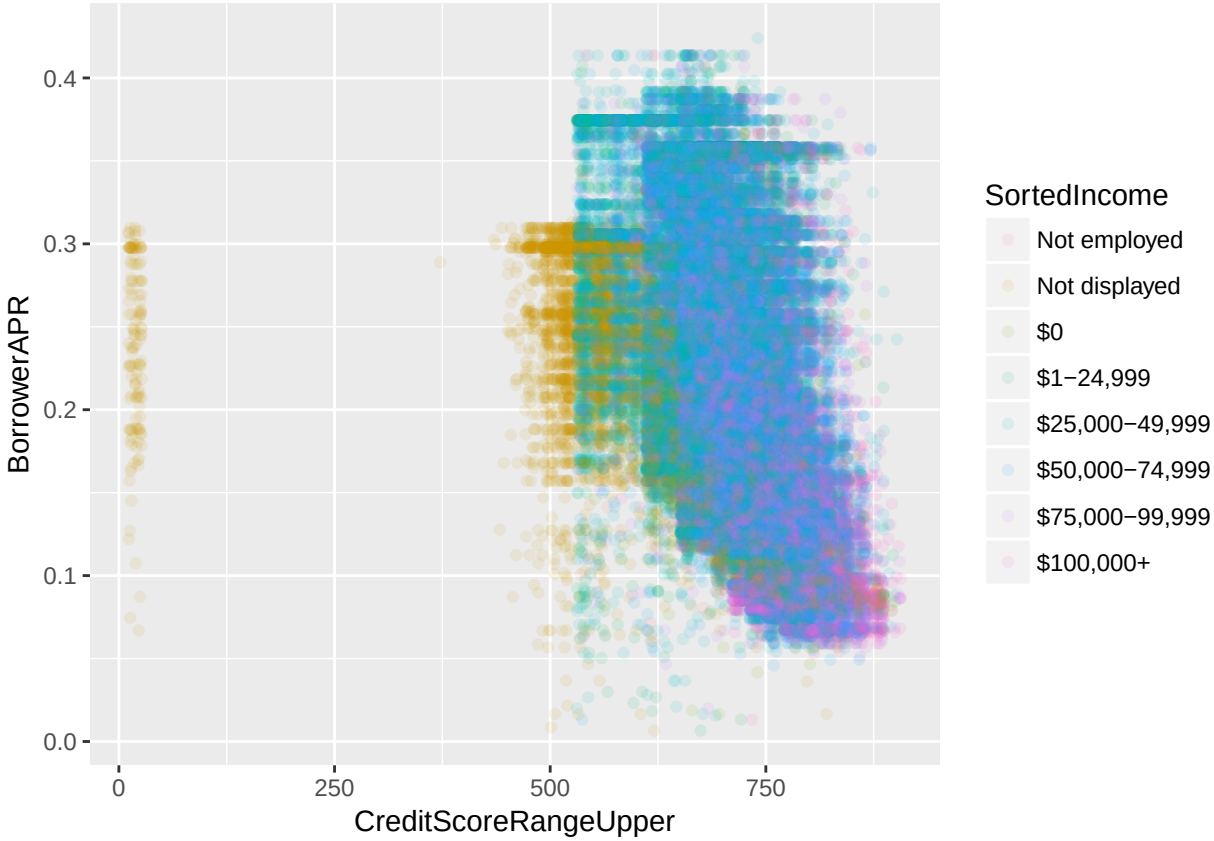
I was surprised the relationship between delinquencies and amount delinquent was not stronger.

What was the strongest relationship you found?

The relationship between median APR and loan year was very strong. It goes to show even with perfect credit and high income, market conditions have a large affect on interest rates.

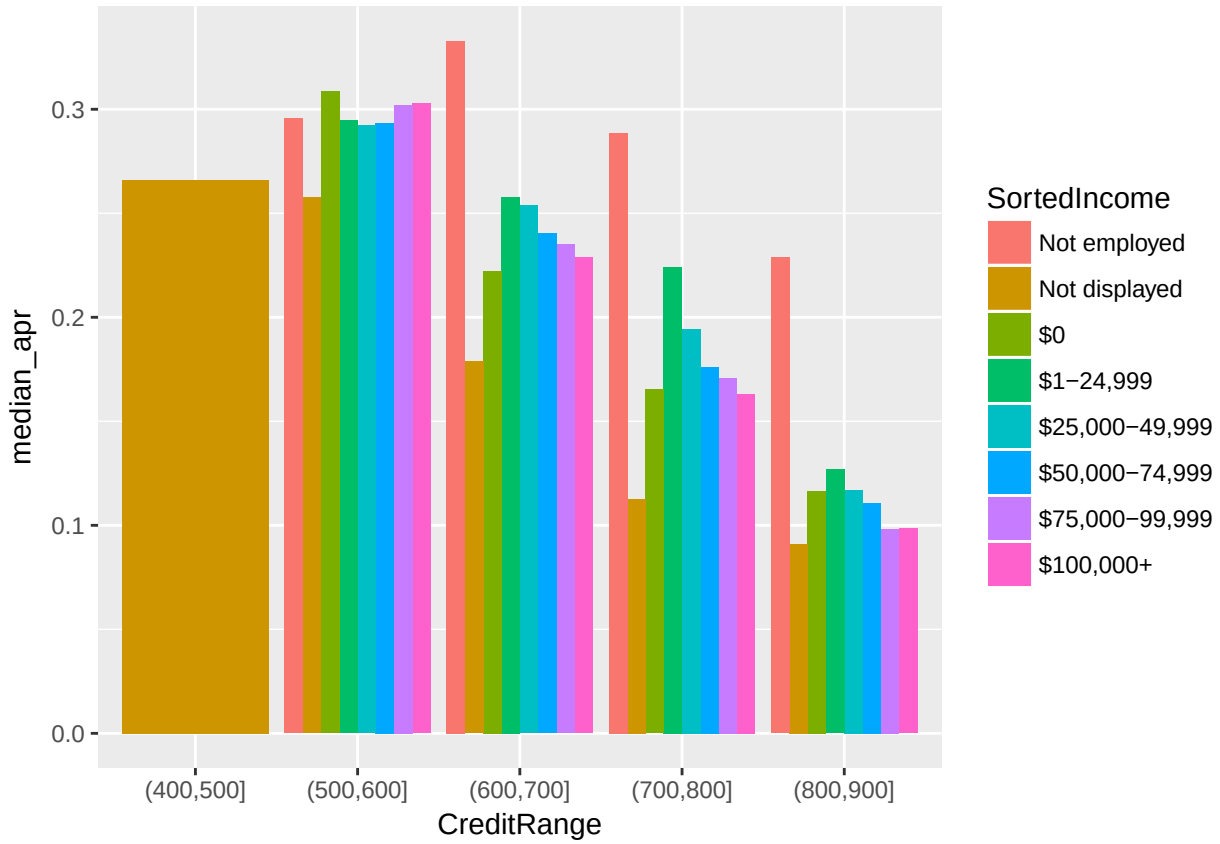
Multivariate Plots Section

Based on previous analysis, it seems like the most important variables are credit, income, and loan year. We will examine each of these in greater detail.

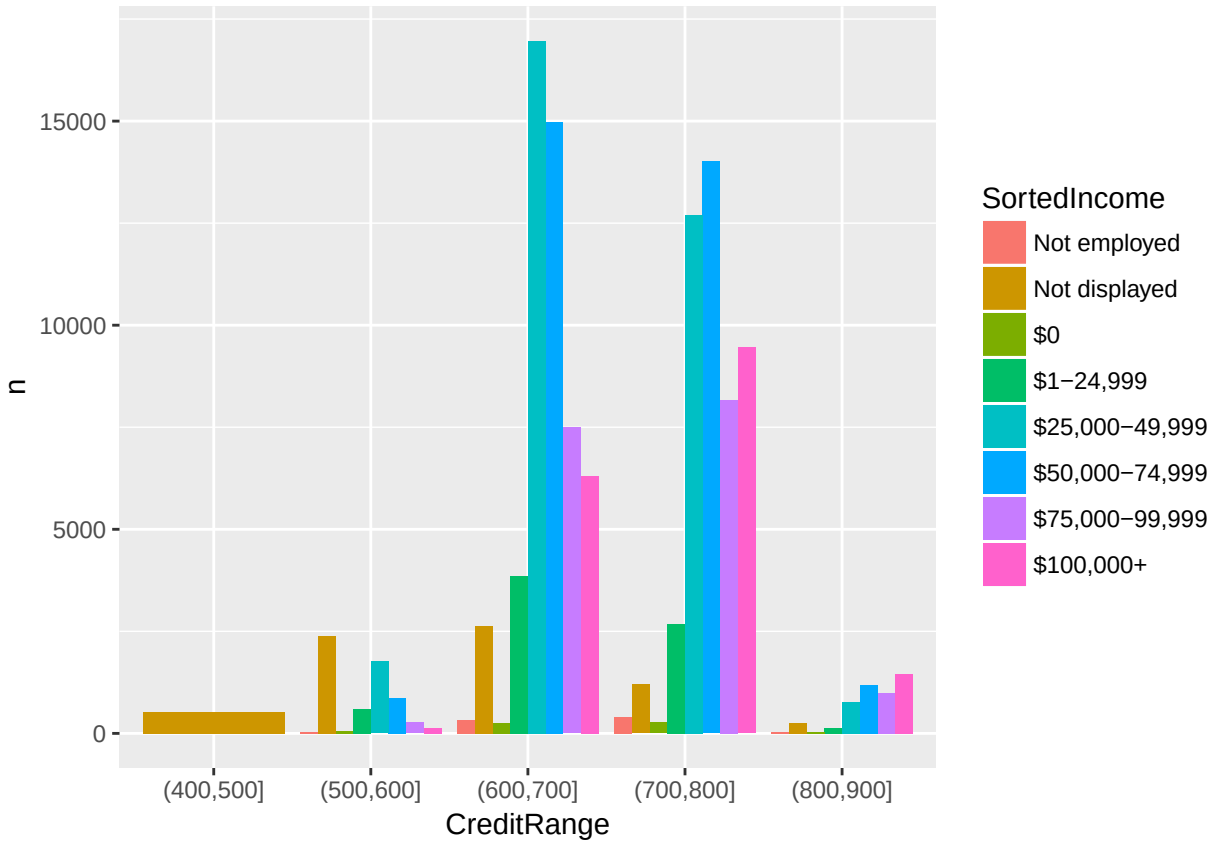


The graph above examines APR by credit score and income. One thing that stands out is all of the sub 500 credit scores have undisplayed incomes. There also seems to be a relationship between these 3 variables - the bottom right has more higher incomes, but the graph is noisy.

```
clean_prosper$CreditRange =
  cut(clean_prosper$CreditScoreRangeUpper, breaks=c(400, 500, 600, 700, 800, 900), right=T)
clean_prosper$APRRange = cut(clean_prosper$BorrowerAPR, breaks=c(.05, .1, .15, .2, .25, .3, .35, .4, .45
```

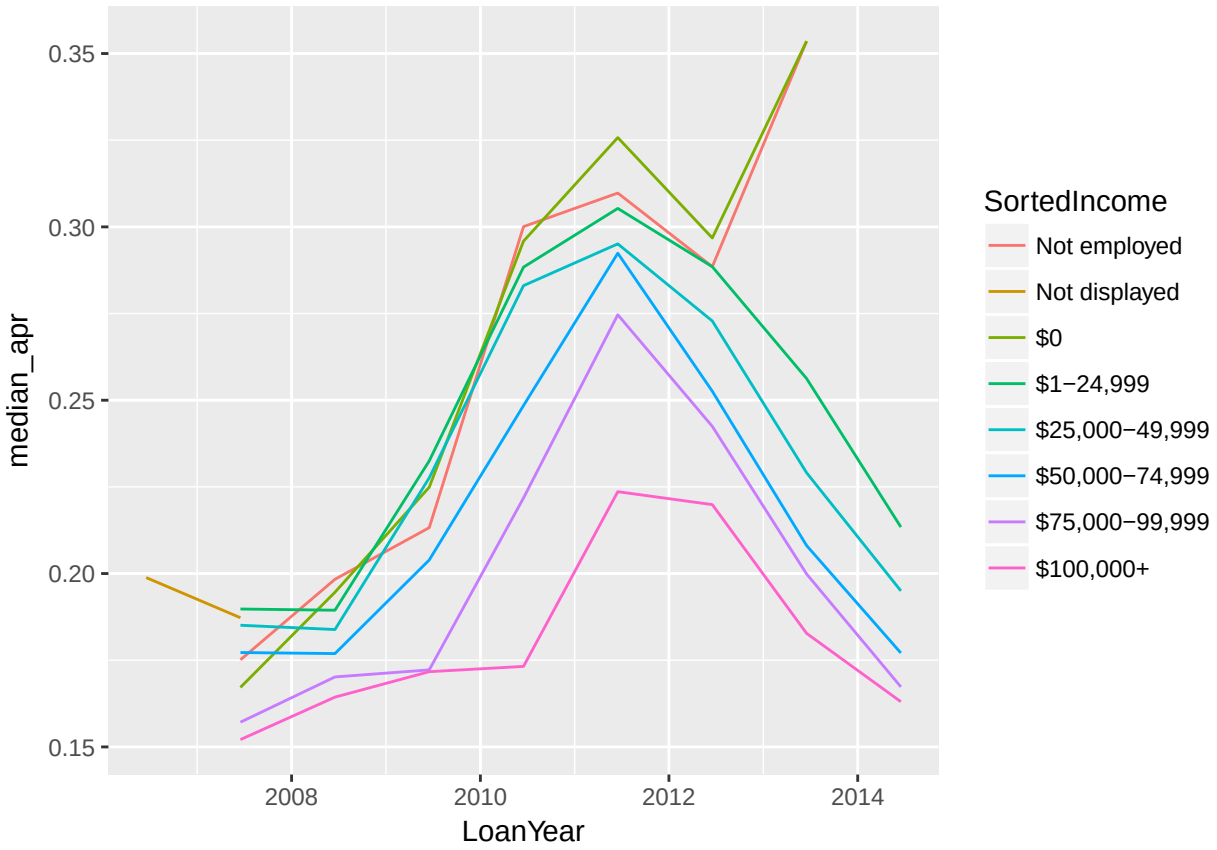


This graph much of the same information, but is more interpretable. It is clearly the case that as income and credit score increase interest rate decreases.

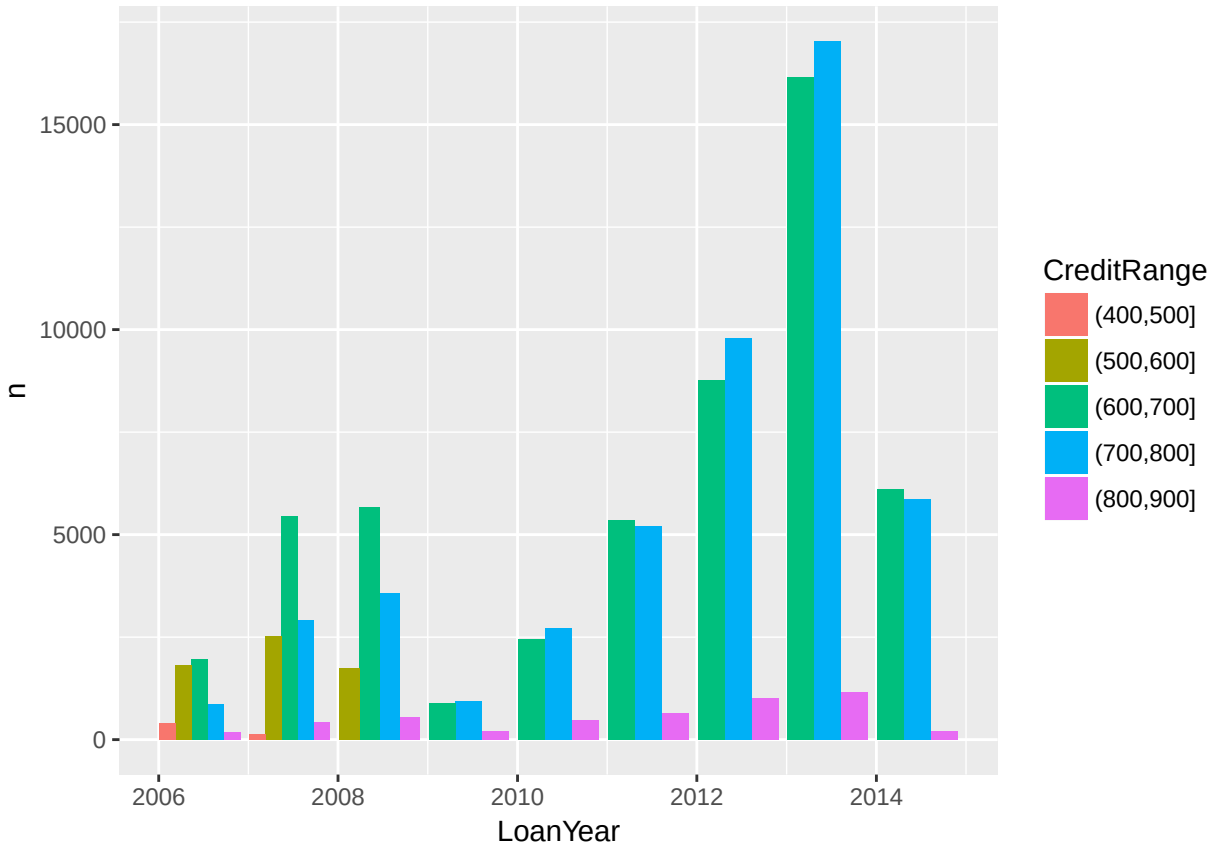


Furthermore, the bank also grants more loans to higher incomes, so higher incomes are more likely to receive loans and get better rates. Let's examine the time series data now.

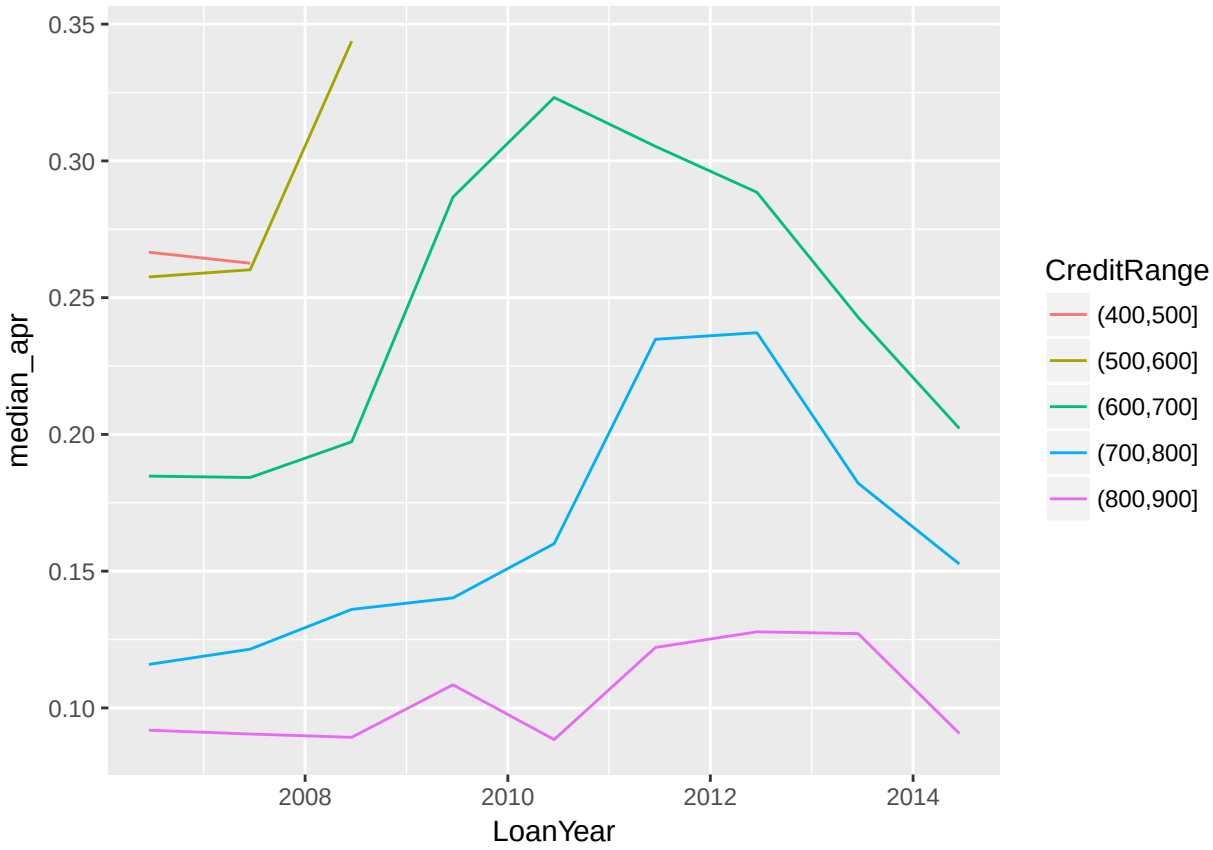
This is very interesting. It looks like all of the undisplayed incomes occur before the financial crisis. Could the bank have had lax standards that got them in trouble? It also appears total loans for people making greater than $25,000 have increased and loans for everyone else have decreased.



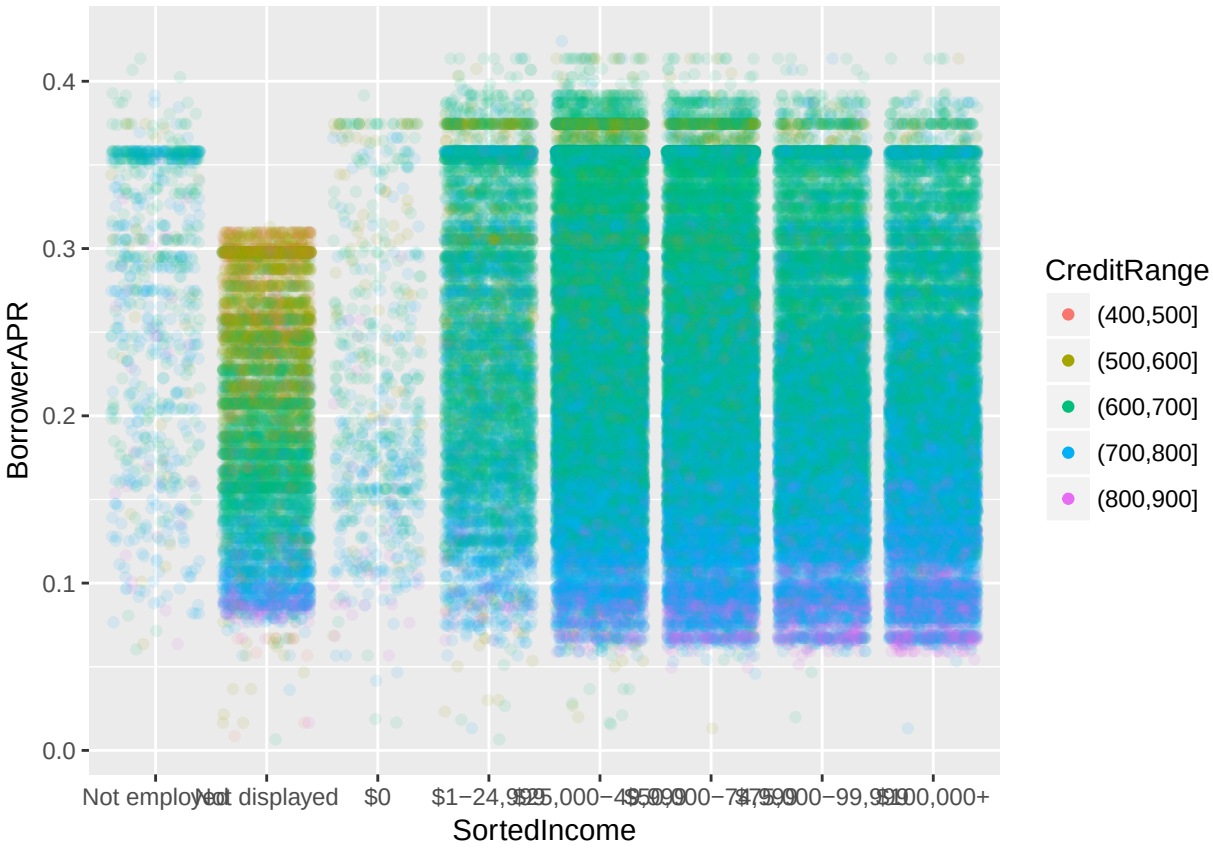
The graph above illustrates APR over time. There is a big jump for all income brackets following 2011, but for everyone making greater than $25,000 there is a steep decline around 2013. The graph below explores credit range over time.



It appears that after 2008 or 2009, the bank stopped granting loans for anyone with credit scores less than 500. The vast majority of loans are from people with credit scores between 600-800.

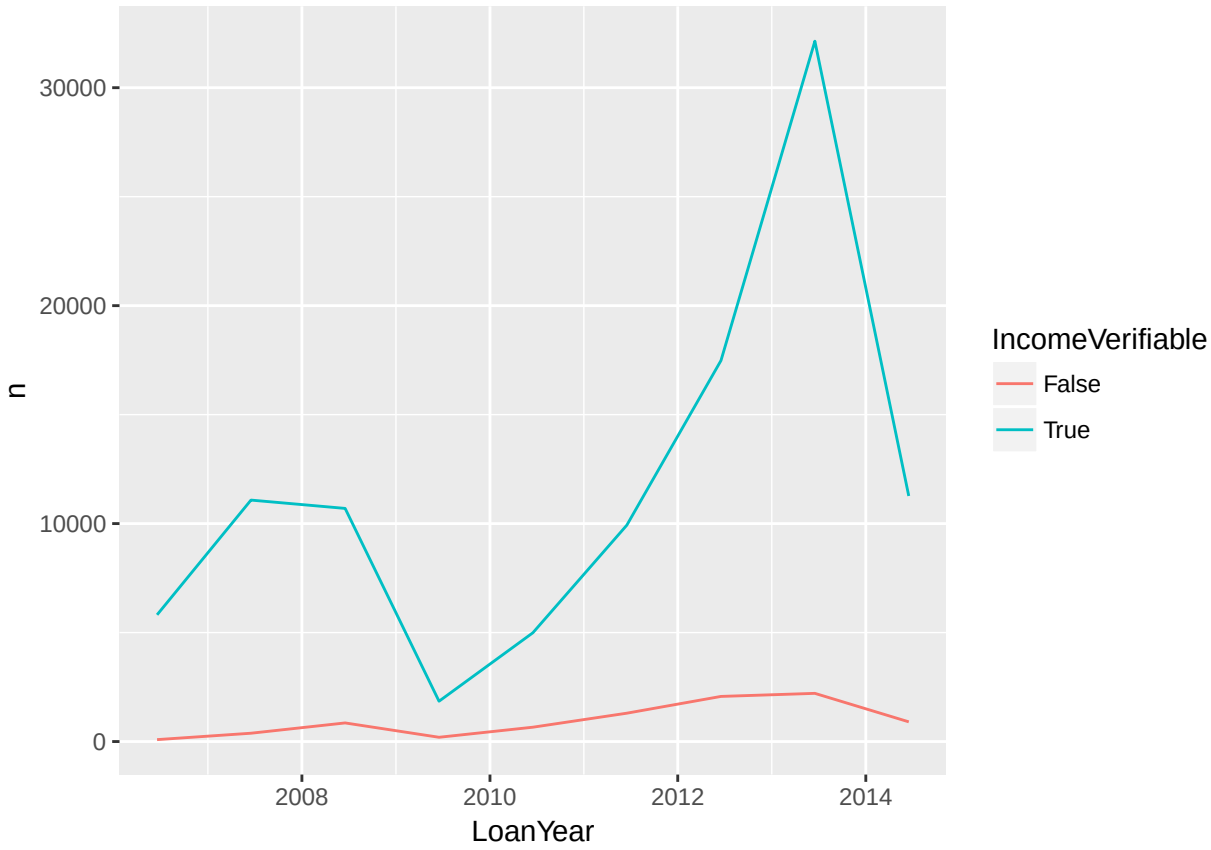


There is also an interesting APR trend over time. Improving credit by 100 points leads to a significantly lower APR.

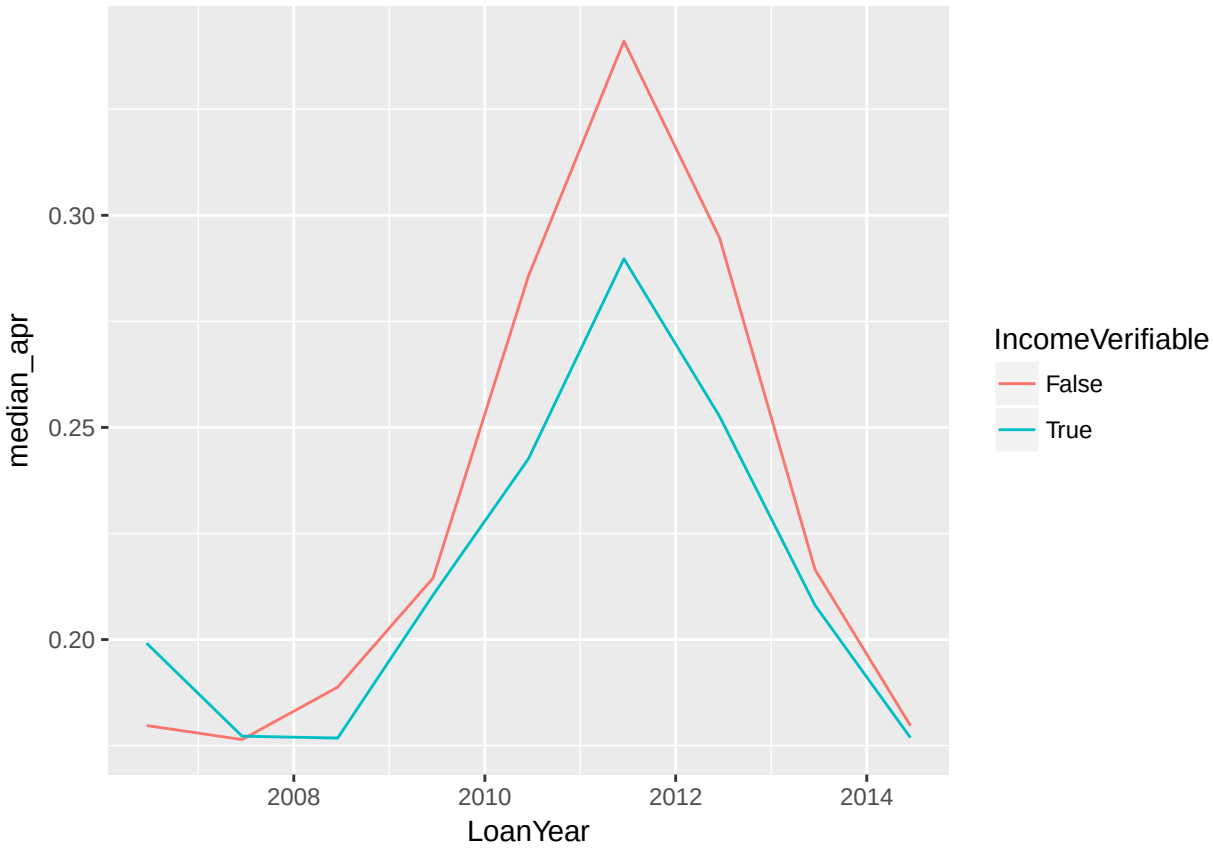


Now this is interesting! It appears unemployed people often have a credit rating around 600-700 to get loans and their interest rate is around .35. We also see a clear decrease in interest rates as income and credit increase.

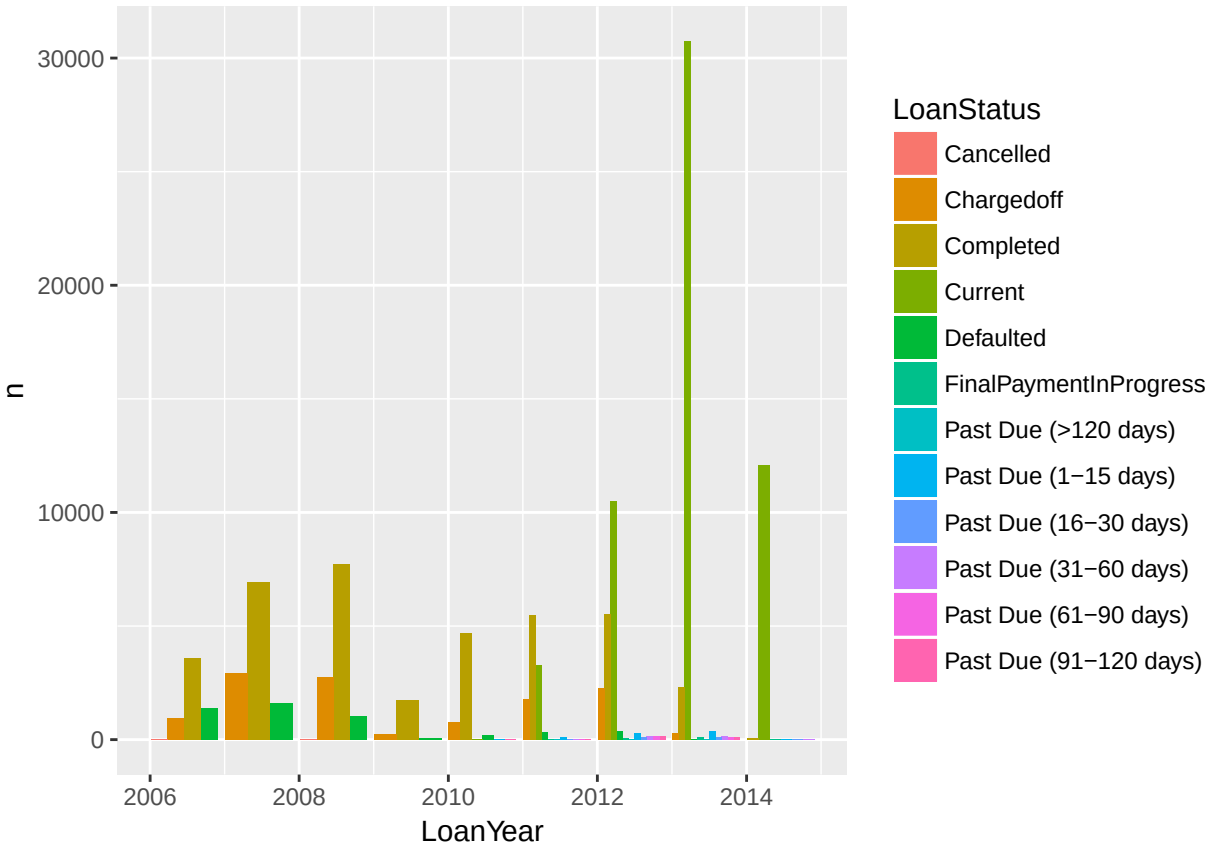
Early we mentioned verified income as a feature. Let us now examine how the bank has verified income over time.



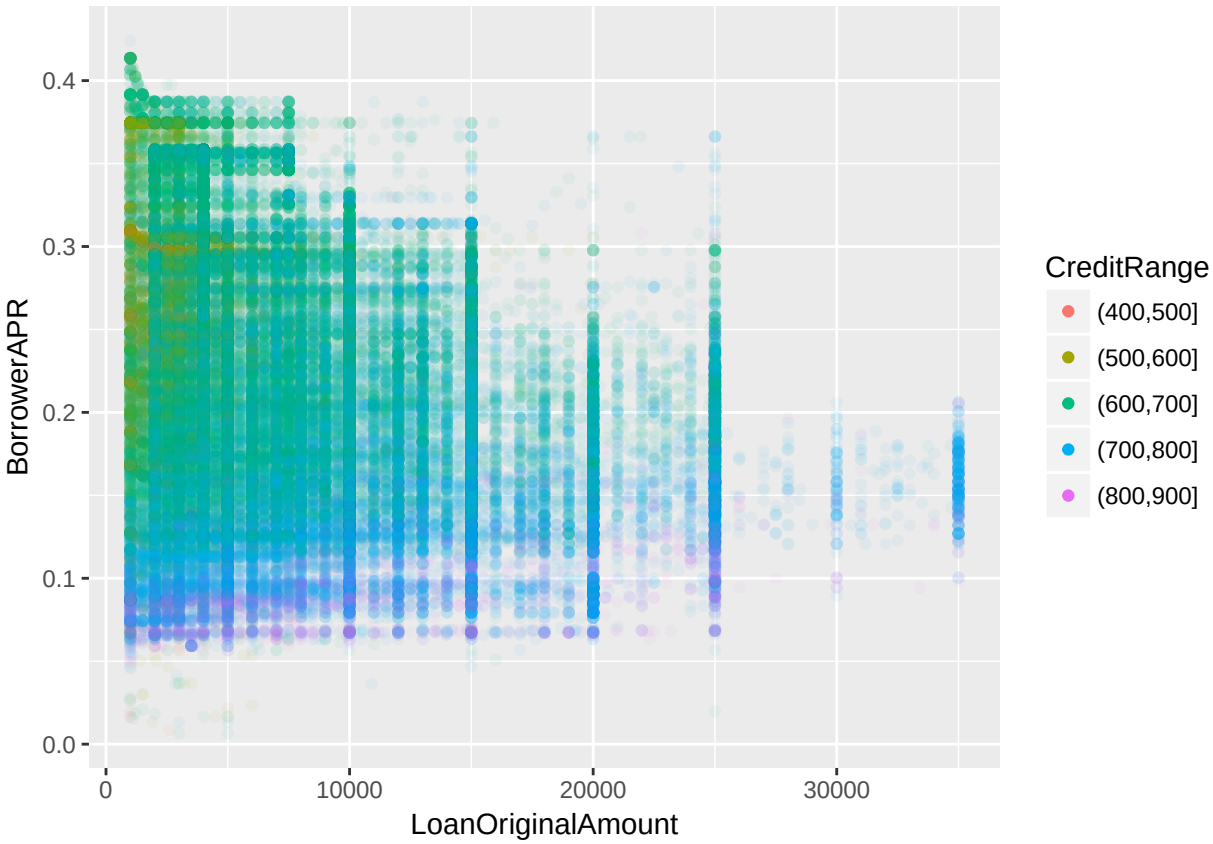
Overall, the number of unverified income loans is steady, but the amount of verified income loans grew sharply. In general, it appears the bank is verifying more incomes following 2009.



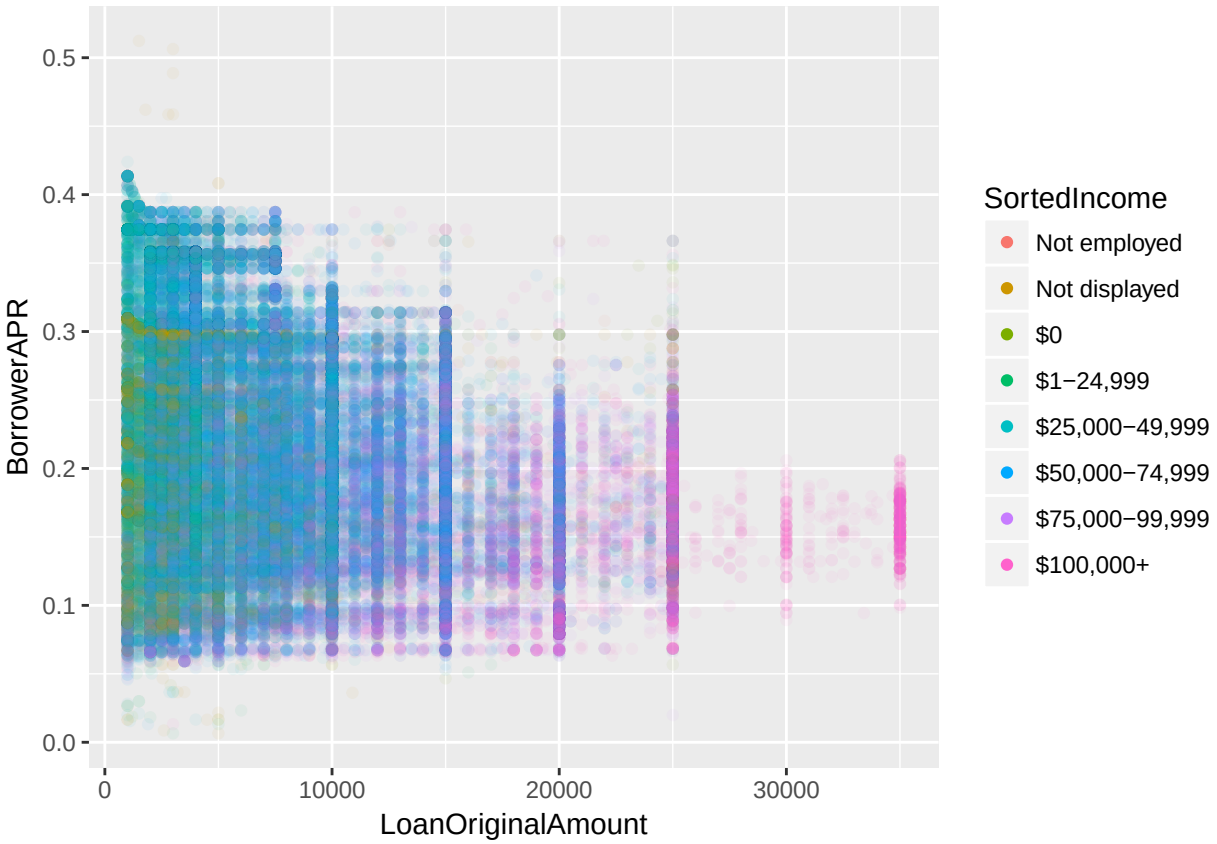
There is also some difference in APR for verified and unverified incomes, but it is not great.



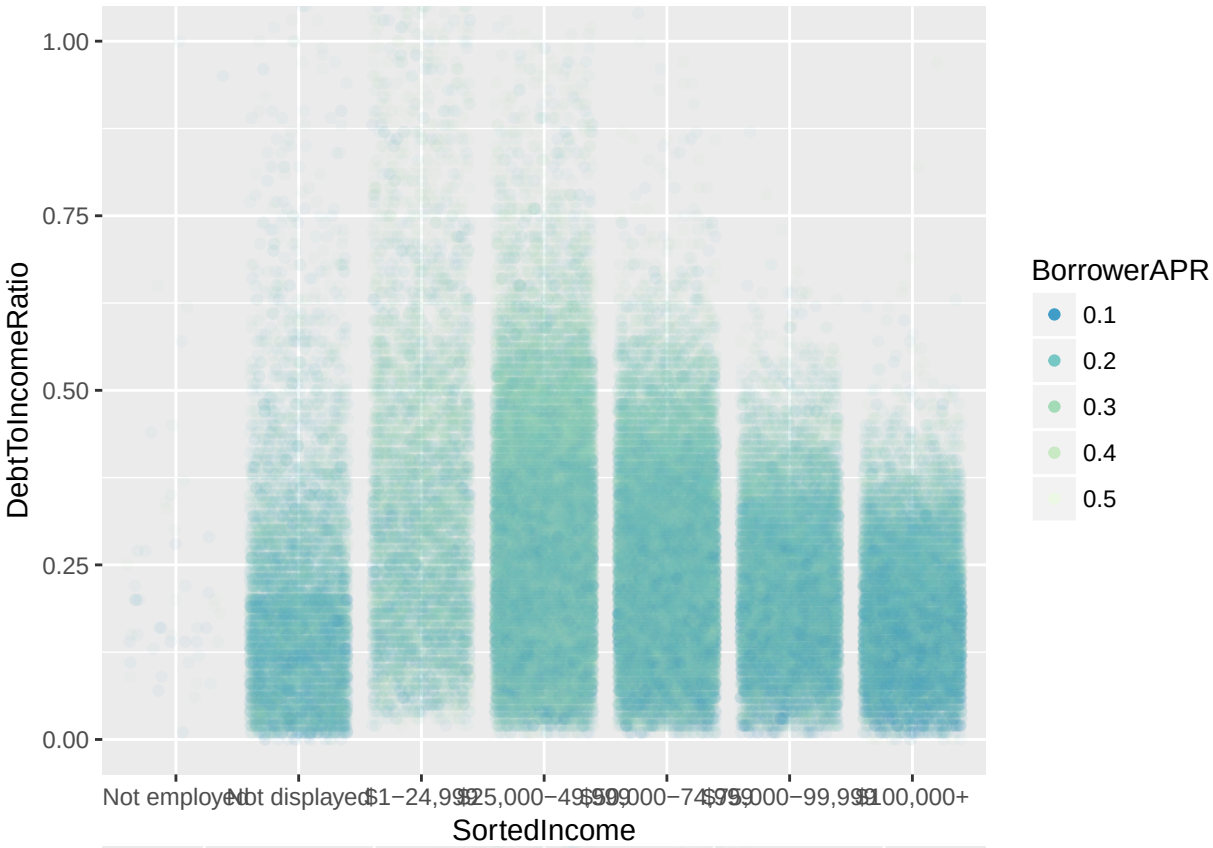
I also examined loan status over time. The bank chargedoff and defaulted a lot of loans prior to 2008. Overall, it appears the bank changed its strategy considerably following the financial crisis with a greater volume of loans to higher income and credit individuals.

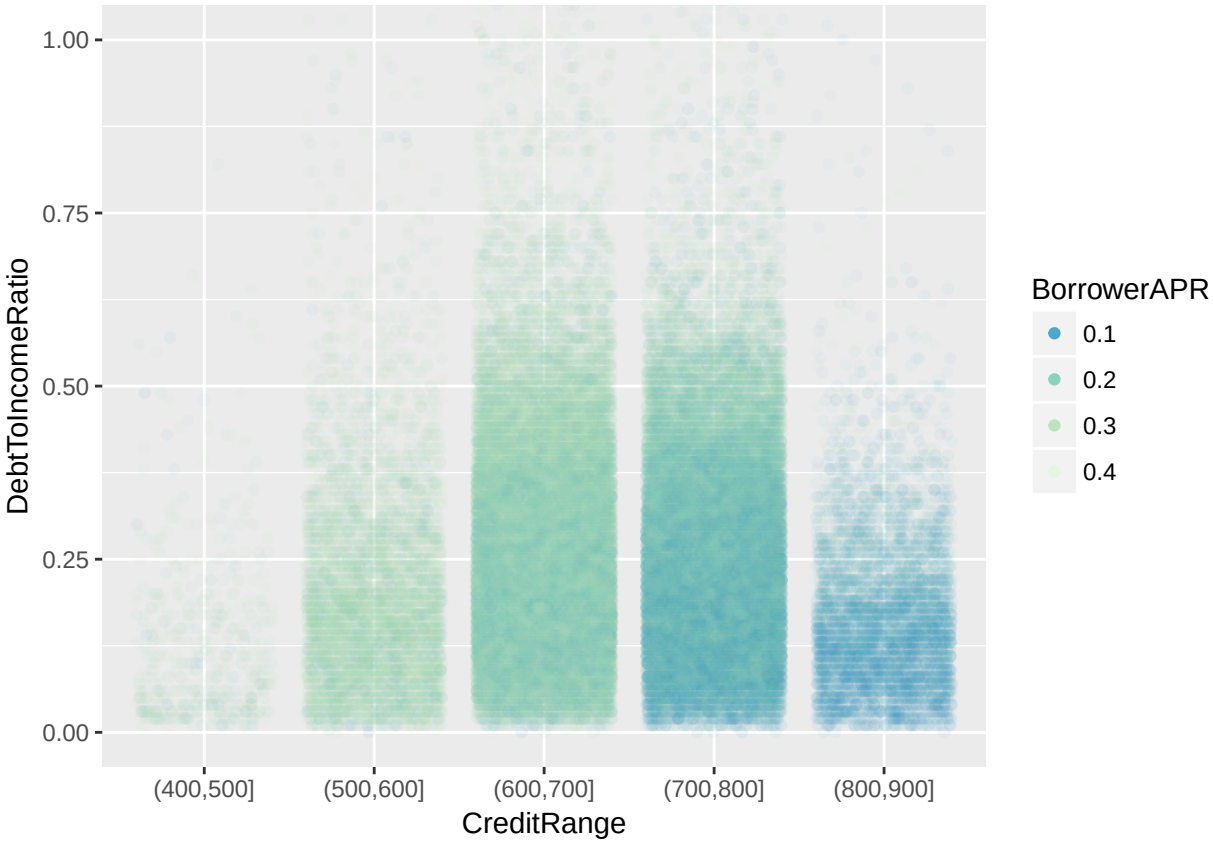


Earlier, we noticed a relationship between apr and loan amount. I reexamined this relationship from the perspective of income and credit. Unsurprisingly, the vast majority of loans for values greater than $30,000 come from 700+ credit ratings.



And most people who get loans greater than $30,000 have incomes above $100,000. Finally, I compared income and credit range to Borrower APR with debt to income ratio.





Overall, the results are interesting. It appears individuals with high income, high credit, and low debt ratios

can expect lower interest rates and higher dollar value loans. To confirm these thoughts, I built a simple linear model with 5 variables - Credit score, day of loan, income range, loan amount, and debt to income ratio. These 5 variables account for 1/3 of the variance in BorrowerAPR.

```
summary(m5)
```

```
##
## Call:
## lm(formula = BorrowerAPR ~ CreditScoreRangeUpper + LoanDay +
##     SortedIncome + LoanOriginalAmount + DebtToIncomeRatio, data = clean_prosper)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51083 -0.04746 -0.00521  0.04339  0.24655
##
## Coefficients:
##                          Estimate Std. Error  t value Pr(>|t|)
## (Intercept)            3.088e-01  4.412e-03   69.988  < 2e-16 ***
## CreditScoreRangeUpper -6.055e-04  3.461e-06 -174.938  < 2e-16 ***
## LoanDay                2.270e-05  2.677e-07   84.787  < 2e-16 ***
## SortedIncome.L         4.297e-02  4.276e-03   10.049  < 2e-16 ***
## SortedIncome.Q        -3.763e-02  4.059e-03   -9.271  < 2e-16 ***
## SortedIncome.C         1.307e-02  3.069e-03    4.260 2.04e-05 ***
## SortedIncome^4         9.747e-03  1.891e-03    5.155 2.54e-07 ***
## SortedIncome^5        -1.348e-02  1.054e-03  -12.782  < 2e-16 ***
## SortedIncome^6         5.302e-03  6.046e-04    8.769  < 2e-16 ***
## LoanOriginalAmount    -2.837e-06  3.727e-08  -76.139  < 2e-16 ***
## DebtToIncomeRatio      8.953e-03  3.766e-04   23.771  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06466 on 104787 degrees of freedom
##   (9114 observations deleted due to missingness)
## Multiple R-squared:  0.3374, Adjusted R-squared:  0.3374
## F-statistic:  5337 on 10 and 104787 DF,  p-value: < 2.2e-16
```

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?

Credit score and income work well together. Having both will likely lead to a lower APR. Time and credit score and time and income are also interesting relationships. It appears the bank has asked for higher credit scores and incomes as a result of the financial crisis.

Were there any interesting or surprising interactions between features?

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

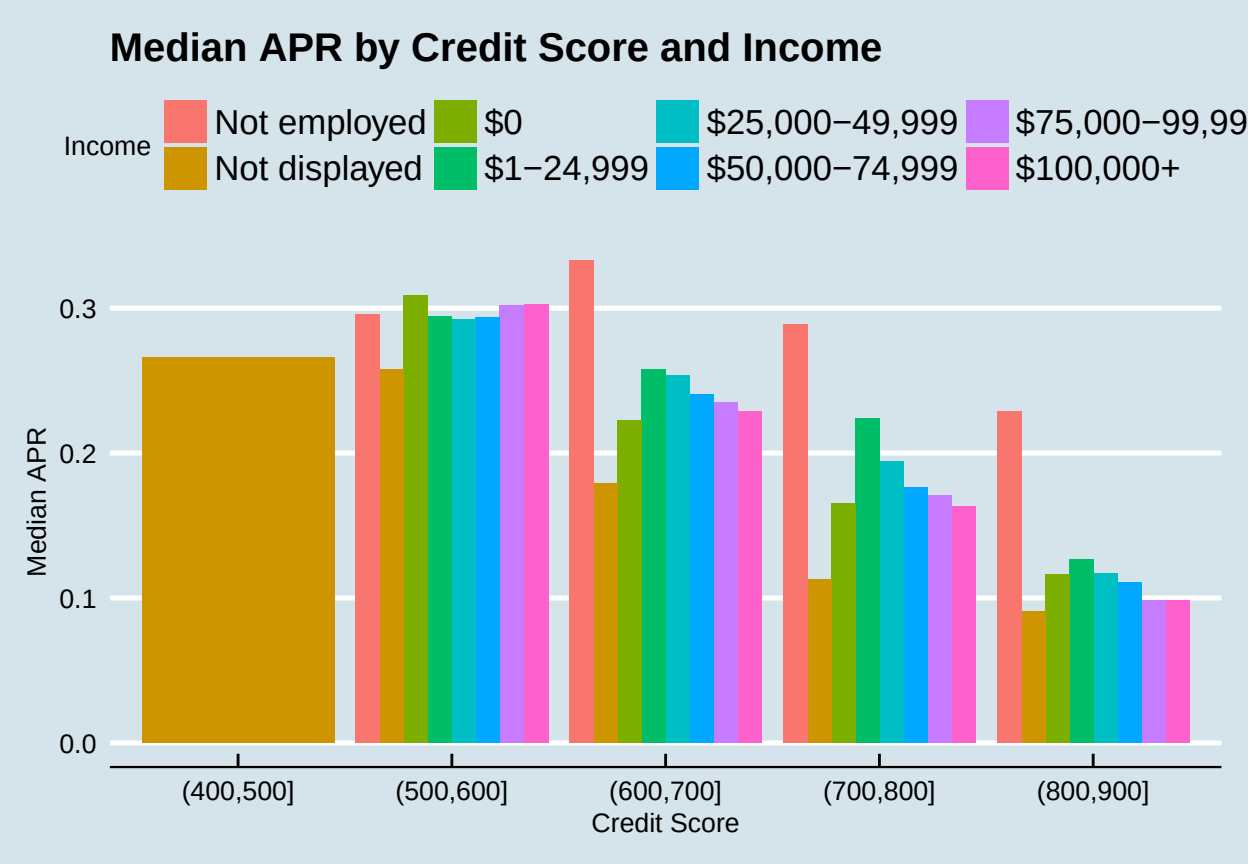
One limitation the model is some important variables, such as DebtToIncomeRatio, have no data. Improving the data integrity could increase the model's R squared. The other big limitation is the model only explains 1/3 of the variance. We need to identify features that are not included. For example, how might location play a role? What about occupation? We have both of these features, but did not include them in the model.

The biggest strength of the model is its simplicity. A banker could bank an approximate decision quickly with only a couple variables.

Final Plots and Summary

> **Tip**: You've done a lot of exploration and have built up an understanding of the structure of and relationships between the variables in your dataset. Here, you will select three plots from all of your previous exploration to present here as a summary of some of your most interesting findings. Make sure that you have refined your selected plots for good titling, axis labels (with units), and good aesthetic choices (e.g. color, transparency). After each plot, make sure you justify why you chose each plot by describing what it shows.

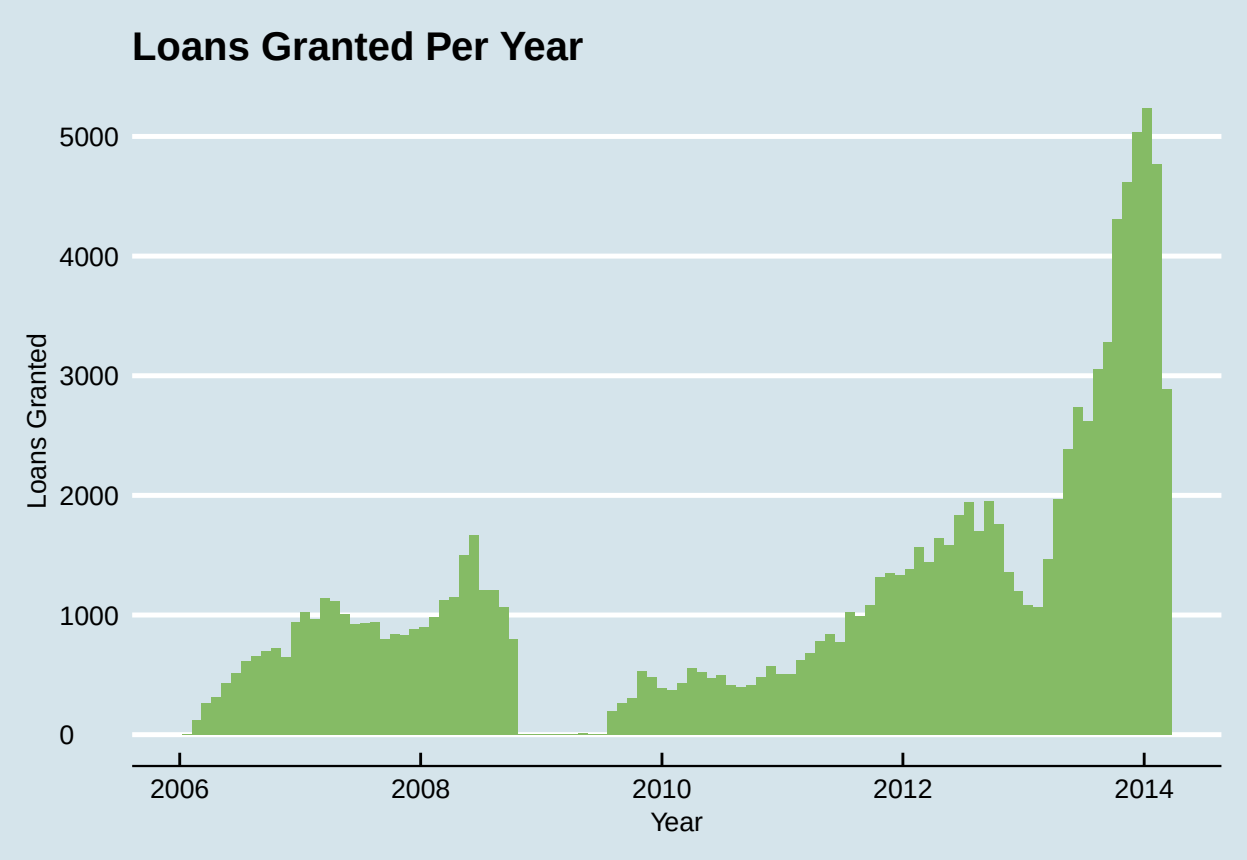
Plot One



Description One

I love this graph because it shows two things. First wee see non displayed incomes are all the 400-500 credit scores. I was curious what not displayed meant and this graph provided some information. But it also pains a clear picture of how APR decreases as credit and income increase. For example, someone with an 800+ credit rating who makes greater than $100,000 can expect an interest rate nearly a point lower than someone with the same income and a 700+ credit rating. Interestingly, their interest rate is about the same with someone who has a $75,000+ income and an 800+ credit rating. Reaching an 800+ credit rating is extremely significant.

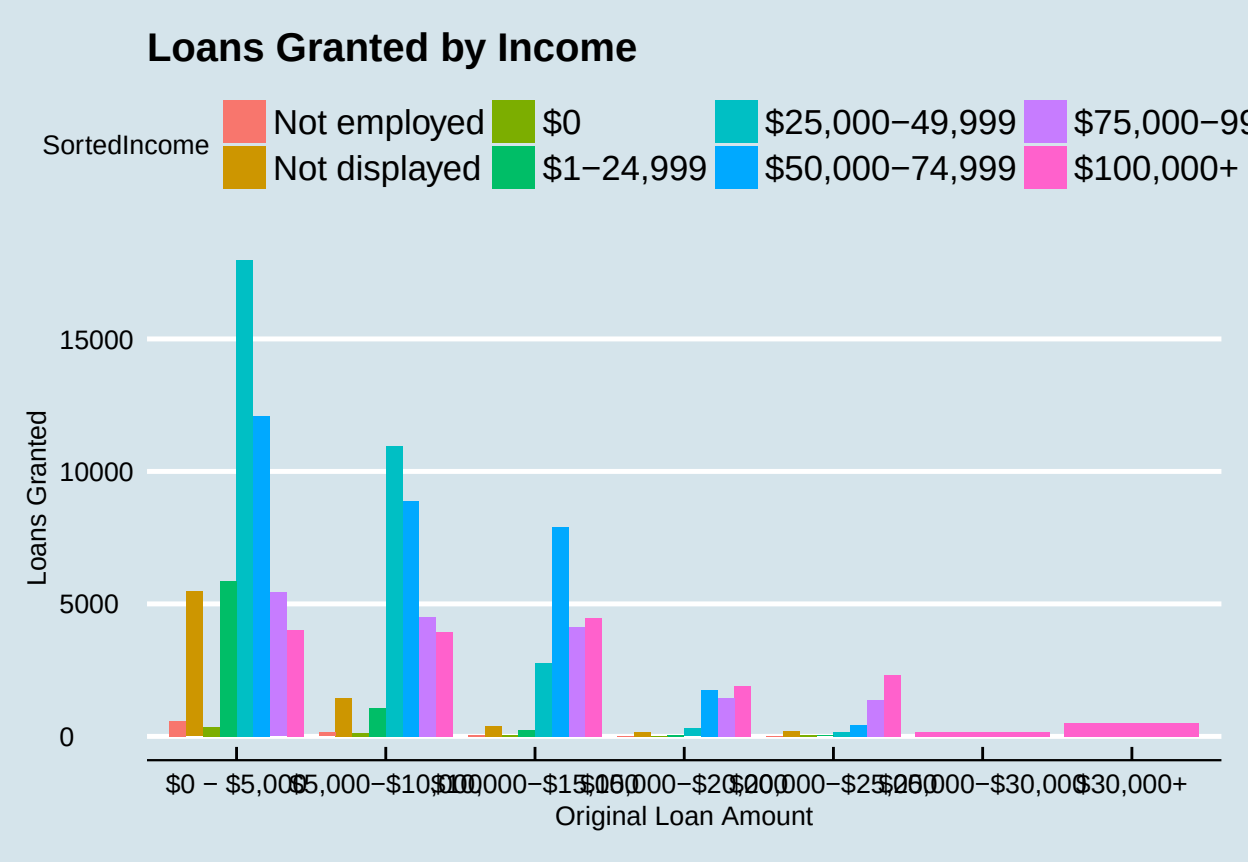
Plot Two



Description Two

I like this chart because it visualises the magnitude of the financial crisis. 2009 barely even registers on the graph. Someone from the future could see this graph and wonder - what happened in 2009?

Plot Three



Description Three

This chart is interesting because it tells a slighter different story than the first chart. In that chart, it appeared an increase in credit score was more important than an increase in income to lower interest rates. While that might be true, income has a larger effect on the ability to receive loans greater than $20,000

Reflection

The prosper loan dataset contains 11937 observations and 82 variables. I started analyzing the data by exploring the distribution and correlations of different variables in the hopes of predicting a loan recipients interest rate.

I indentified debt to income ratio, loan day, loan amount, credit score, and income as predictive variables of interest rate. From these variables, I was able to build a model that accounts for 33 percent of the variance in APR. I also identified some interesting trends. For example, the bank granted very few loans during the financial crisis. In fact, it appears the financial crisis might have lead to a complete change in strategy for the bank. It offered more loans to higher,verified income individuals with greater than 600 credit scores.