Prosper Loan Data Analysis by Preston Hall

> We will be using the data provided by Prosper Bank loans. Prosper Bank has facilitated more than $12 billion in loans to more than 770,000 people. You can find more information about Prosper Bank at their website. (www.prosper.com)

Variables I will be analyzing

I have worked in the financial industry for close to 10 years and have serviced and reviewed many loan documents. I am excited to analyze this data and discover new information.

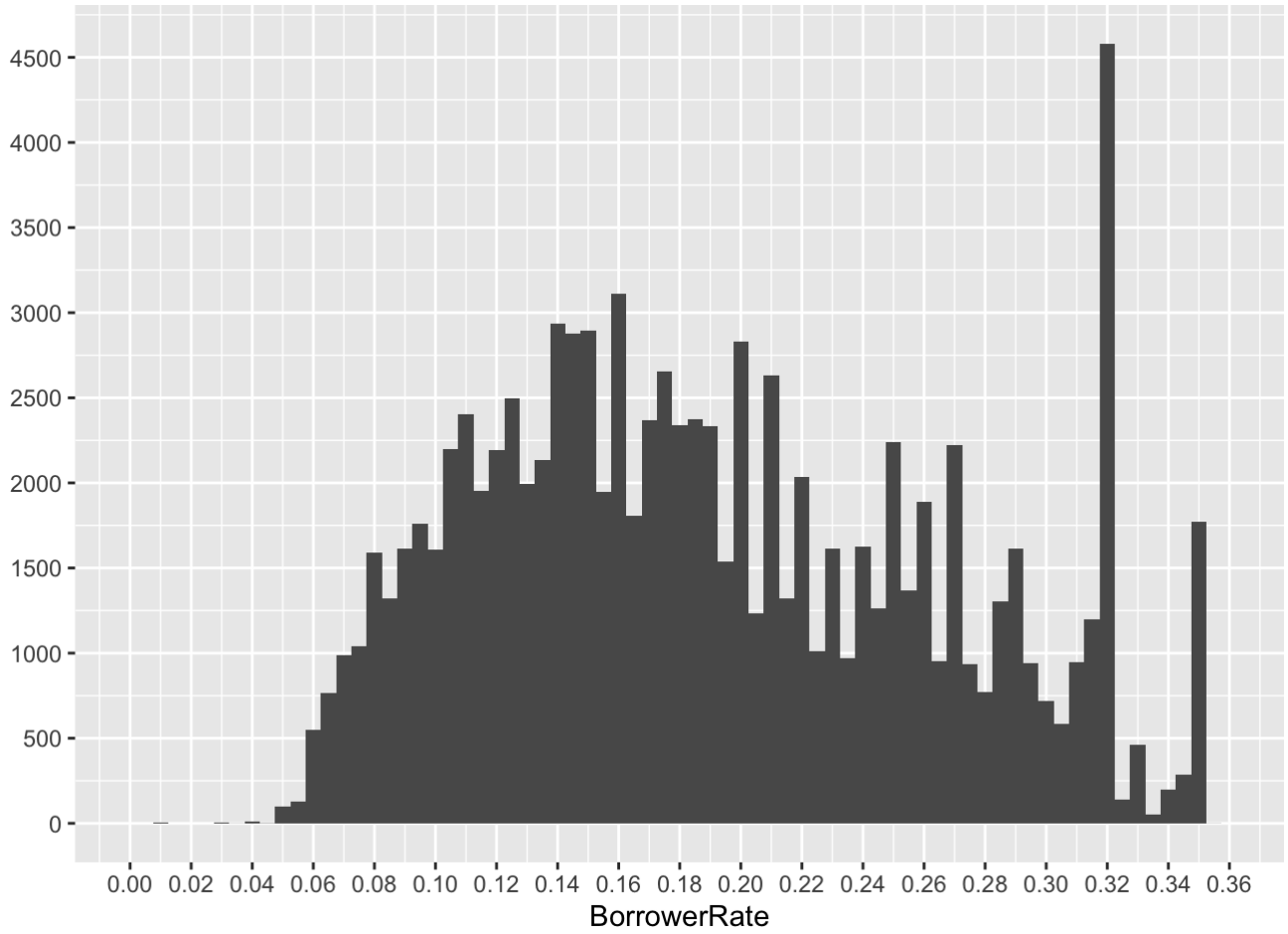
The data that we are exploring is provided in a .csv file 86.5 MB. There are a total of 81 variables and 113,937 observations. There is more information than what is needed for my analysis so I will subest it out to 17 different varaiables and remove all of the NA values. Keeping all of the information will slow down my code and make it difficult to parse through all of the unnecessary information.

After I have created the subset that I want to work with, I will clean it up a bit by removing a few of the NA values and adding factors to IncomeRange, Term and CreditGrade. IncomeRange and CreditGrade will be ordered factors. I will focussing on the following variables:

- Term
- ListingCreationDate
- LoanStatus
- BorrowerRate
- BorrowerState
- EmploymentStatus
- CreditScoreRangeLower
- CreditScoreRangeUpper
- CreditGrade
- OpenCreditLines
- CurrentDelinquencies
- AmountDelinquent
- DebtToIncomeRatio
- StatedMonthlyIncome
- IncomeRange
- MonthlyLoanPayment
- LoanOriginalAmount

Univariate Plots Section

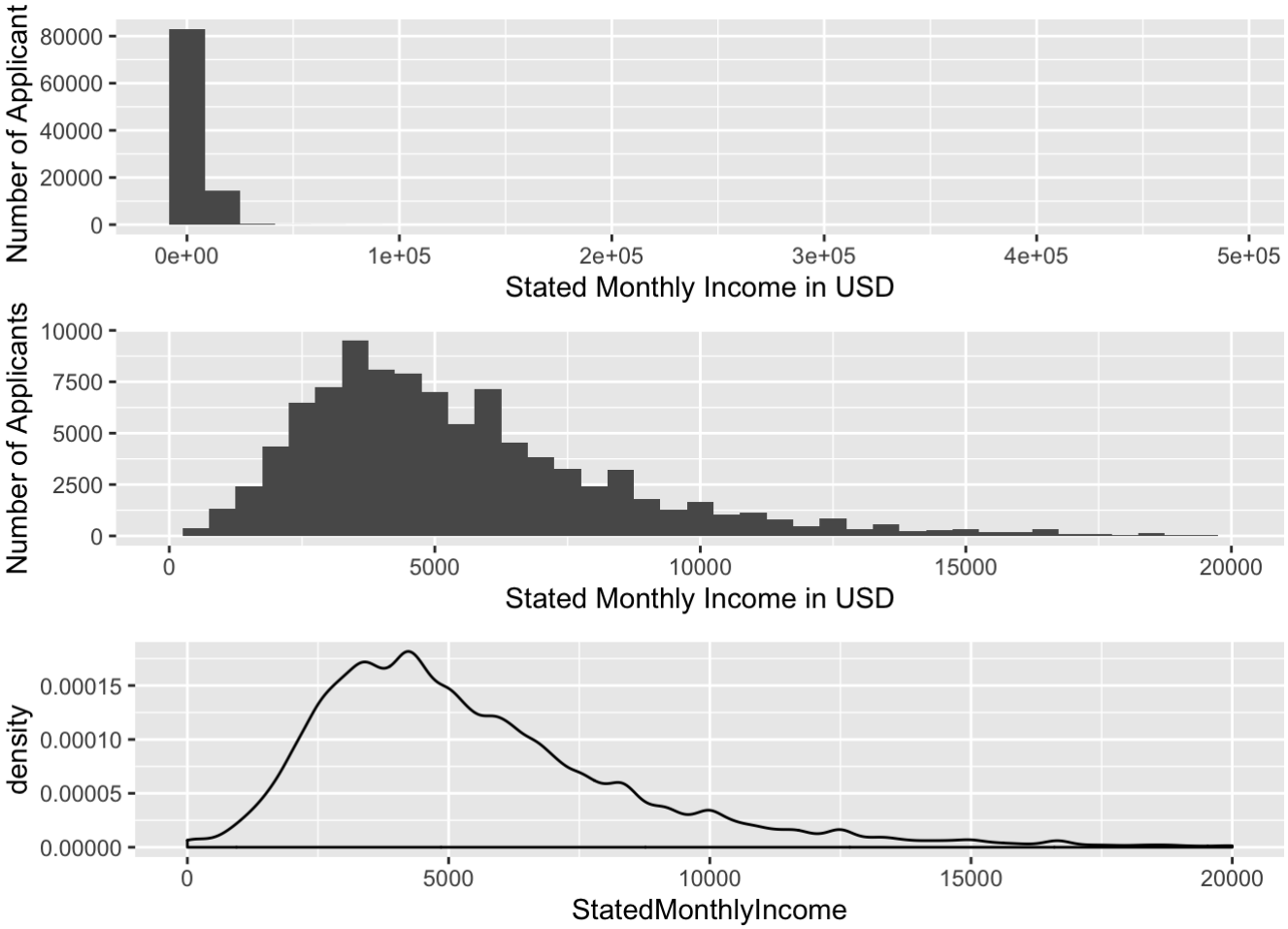
Borrower rate




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.1314  0.1800  0.1907  0.2492  0.3600
```

This graph is interesting as it is has a fairly normal distribution save for the rate at 32% the amount jumps up to over 4500. I am interested in the lull between 32% and 35% and wonder why the numbers drop so low for this rate.

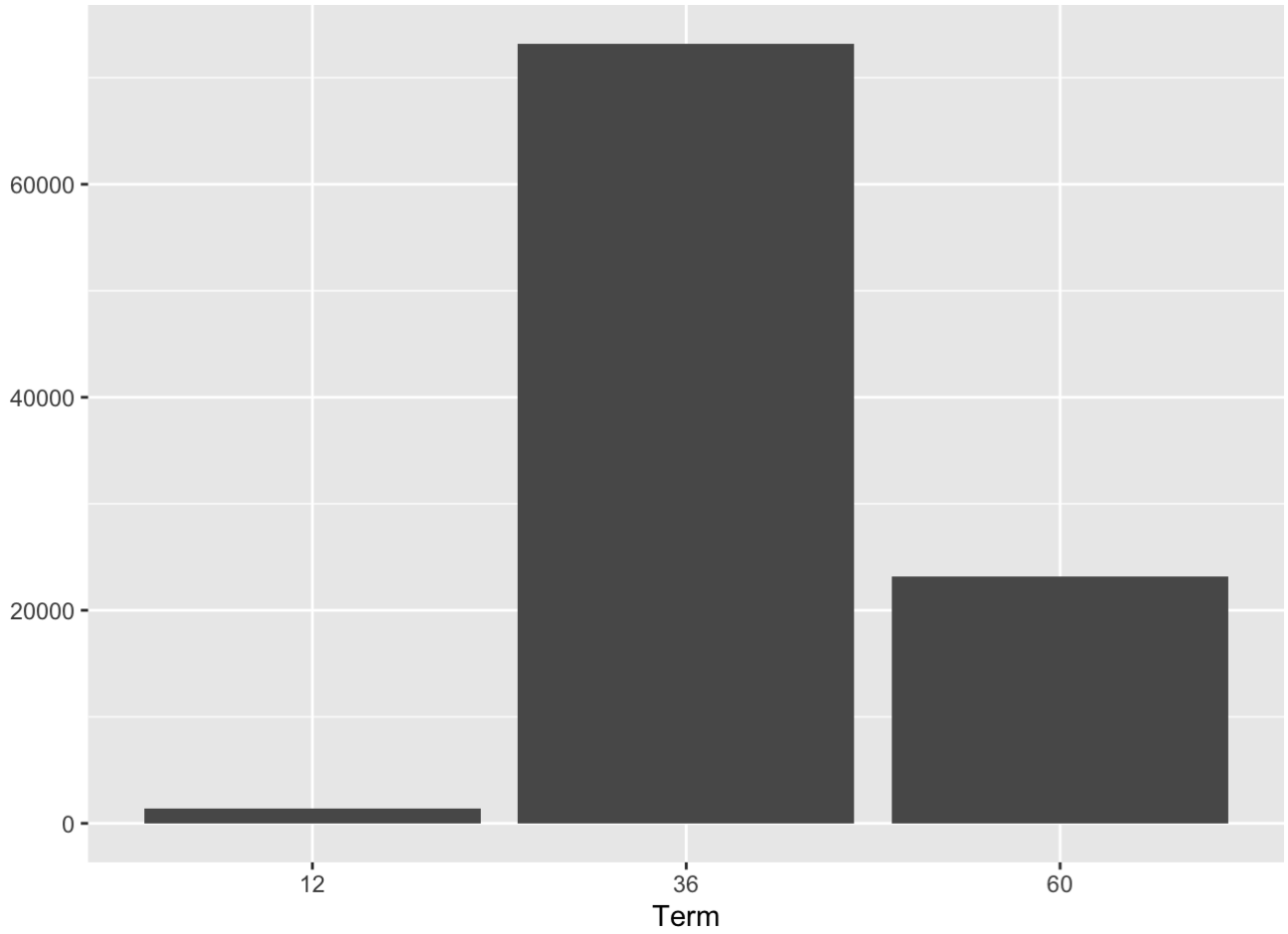
Stated monthly income




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    3333    4833    5720    6986  483333
```

There is an outlier that extends the tail of the graph too far to the right and the graph becomes unreadable. I added a limit to the x axis to better view the information. You can see on this graph that the plot peaks at about $3,000 per month and then decreases. I had to limit the x axis as there is an outlier that skews the graph to a point that is unreadable.

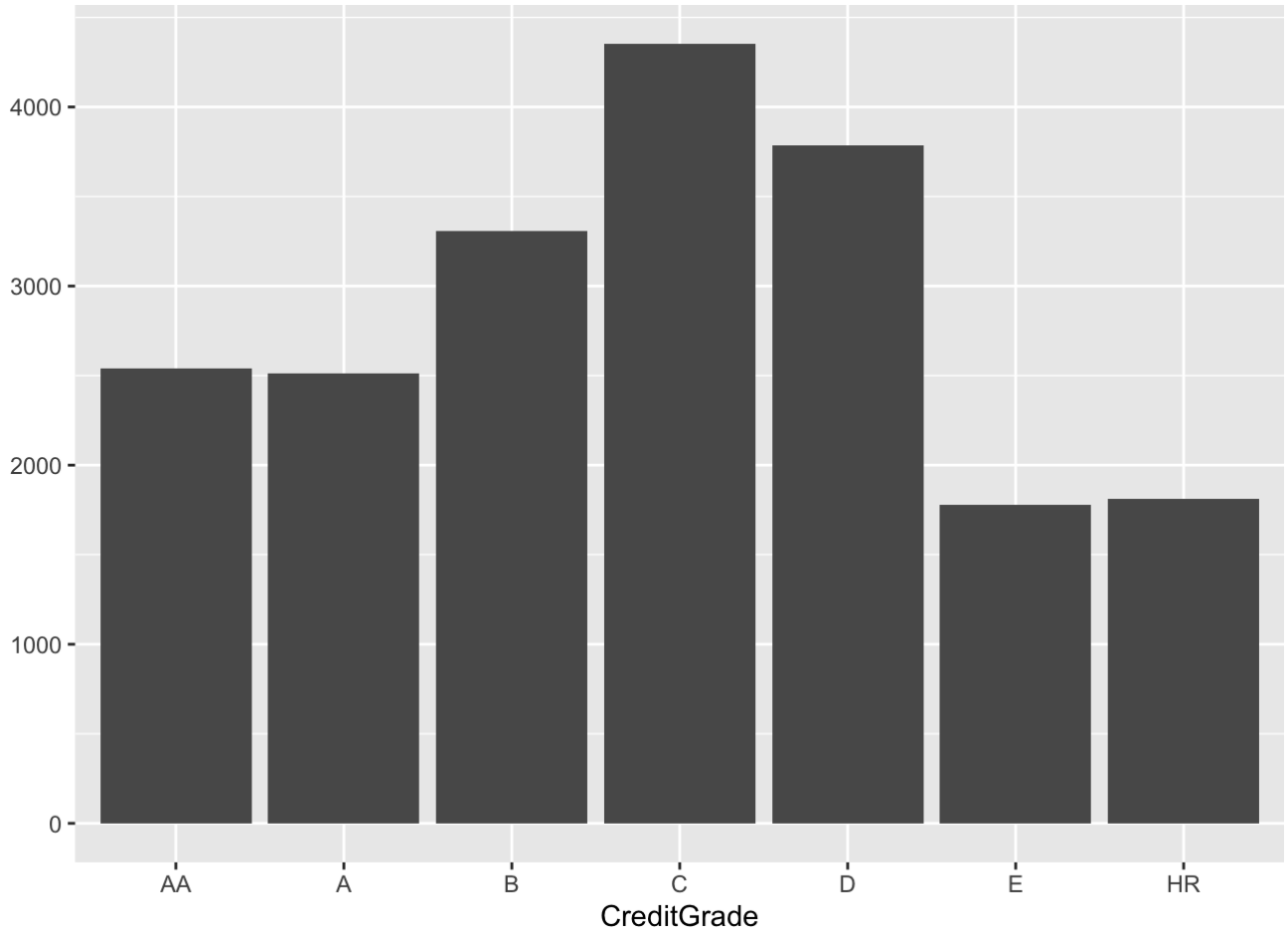
Loan Term




```
##    12    36    60
##  1415 73207 23143
```

The types of loans fall into 3 different terms 12, 36, or 60 months. Or 1, 3, or 5 years respectively. The most common term is 3 years with over 73,000 loans.

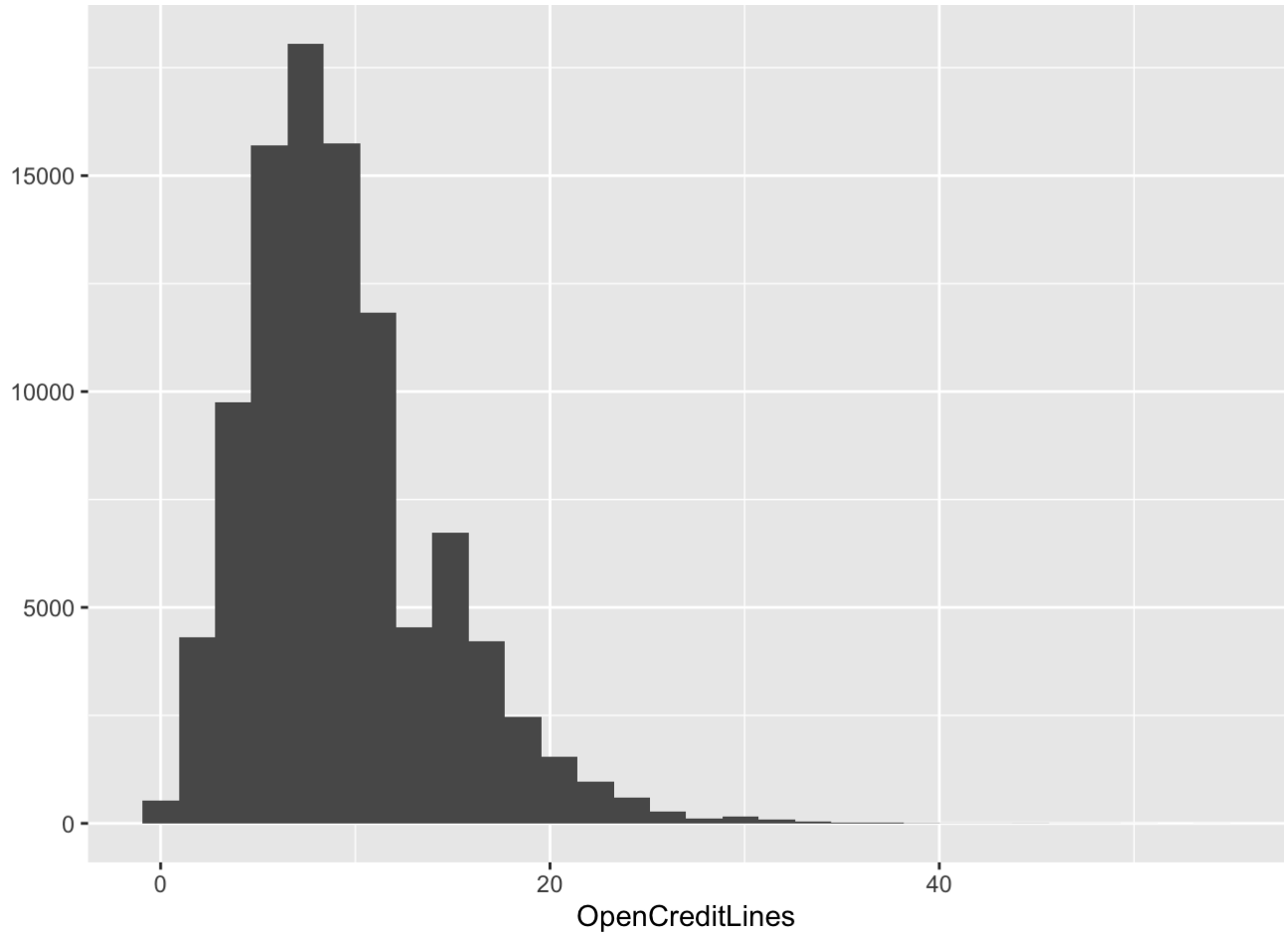
Credit Grades

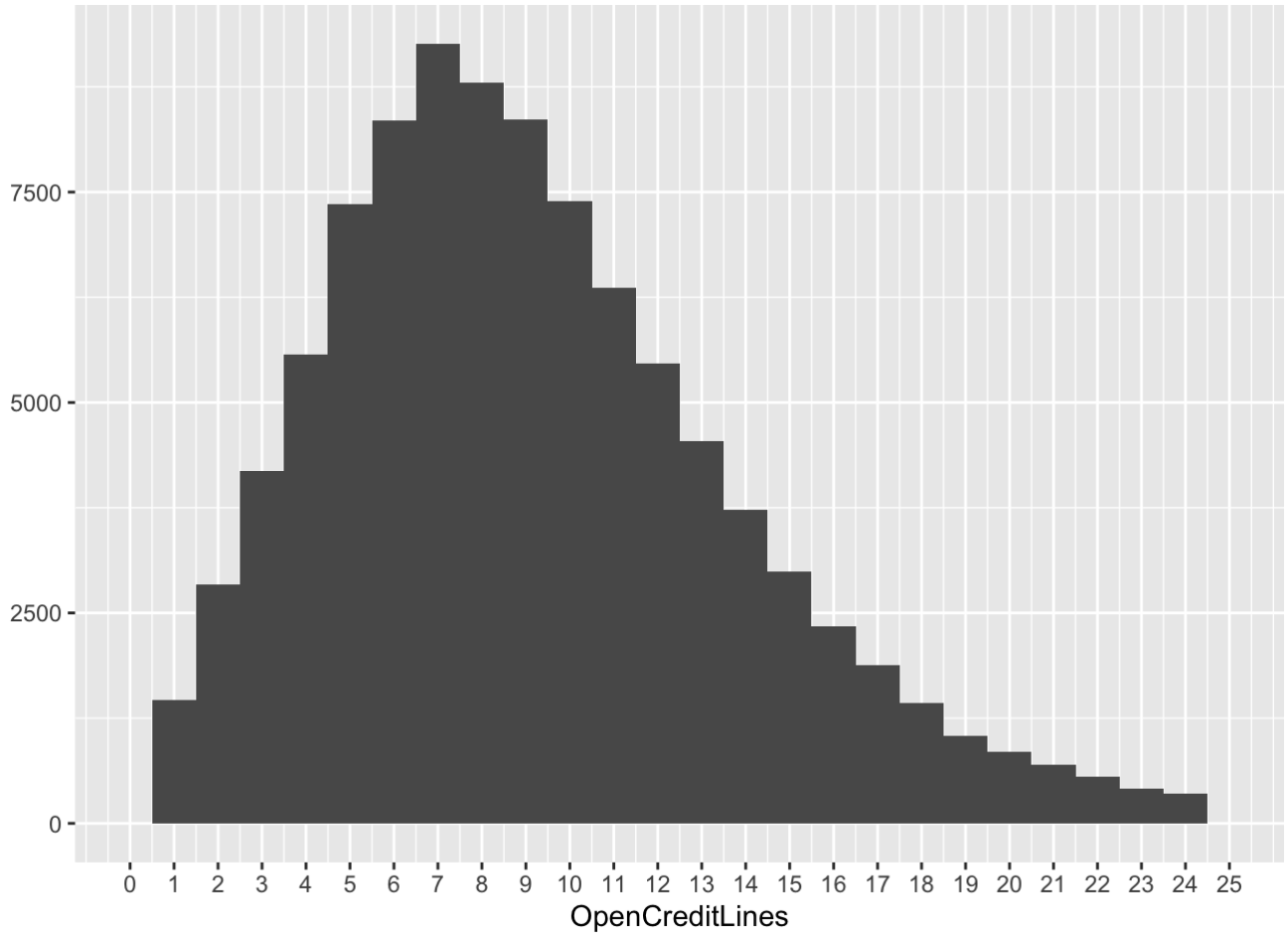



```
##    AA     A     B     C     D     E    HR    NC  NA's
##  2542  2510  3307  4354  3788  1776  1811     0 77677
```

Customer are given a credit grade when provided a loan by Prosper Bank. AA would be the highest score with HR (High Risk) being at the lower end. Most customs have a score between B and D. I am not including NA values in this graph.

Open Credit lines

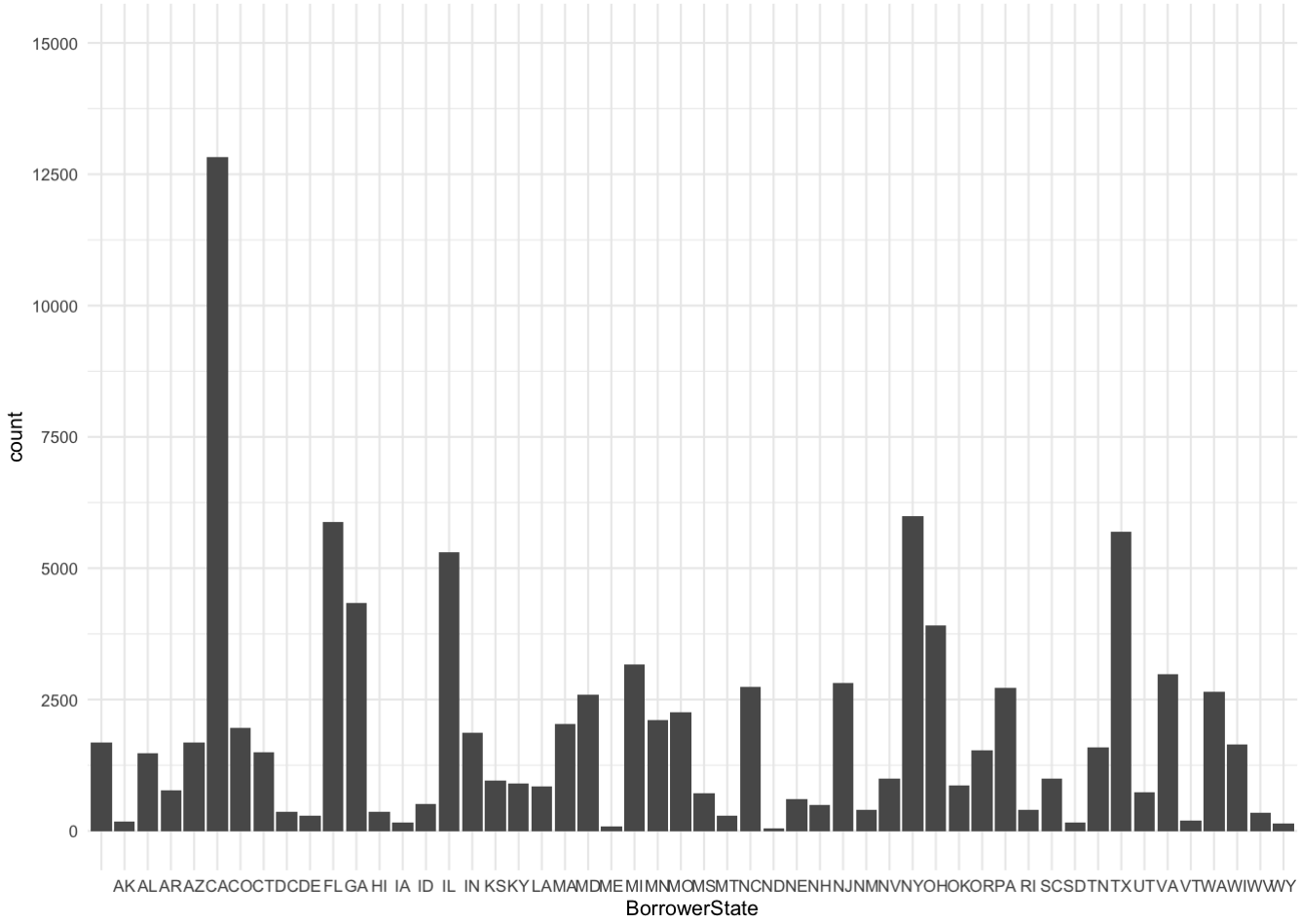

15000
10000
5000
0
0
20
40
OpenCreditLines


7500
5000
2500
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
OpenCreditLines

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   6.000   9.000   9.314  12.000  54.000
```

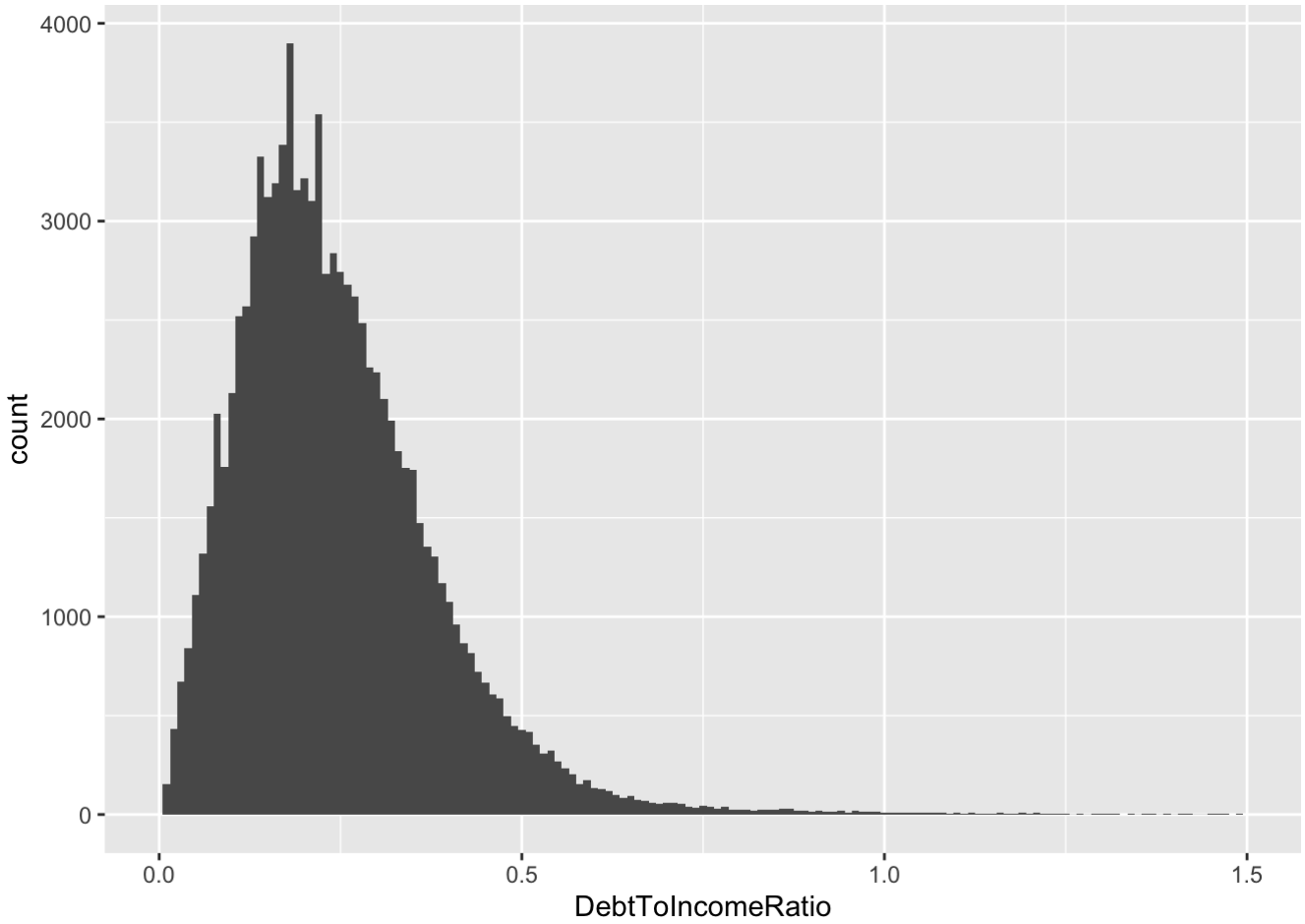
This graph is right tailed distribution. I limited the number of open credit lines so to better see the data. You can see from this graph that most people have around 7 open credit lines with a median of 9.

Borrower state



You can see that Prosper Bank operates out of each of the 50 United States but has most of its loans located in California. This makes sence as Prosper Bank is headquartered out of California.

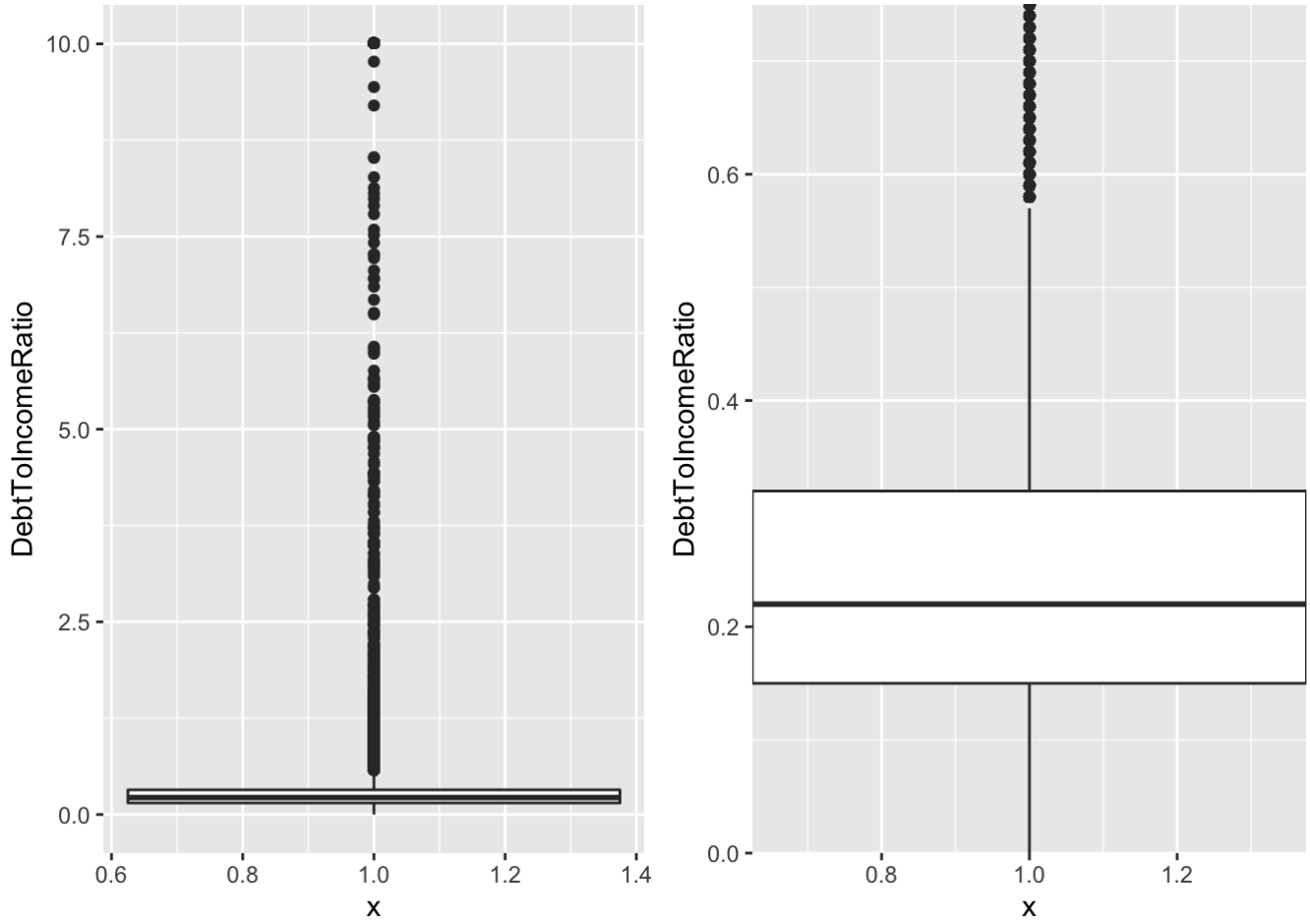
Debt to income ratio Histogram




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.1500  0.2200  0.2743  0.3200 10.0100
```

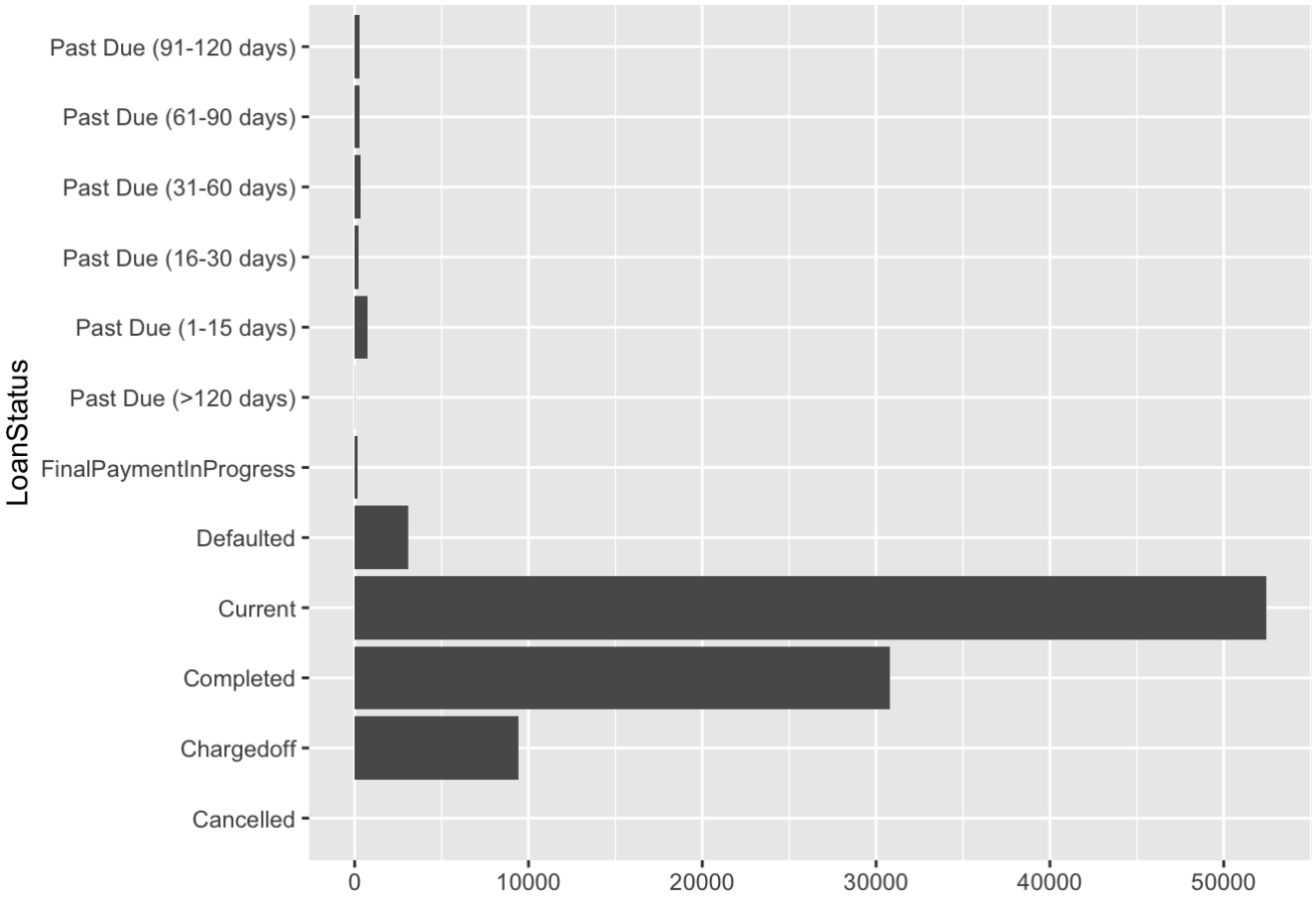
This graph shows a right tailed distribution with a peak in DTI ratio around 20%. Borrowers are less likely to be approved for a loan with a DTI over 32% based off of this information.

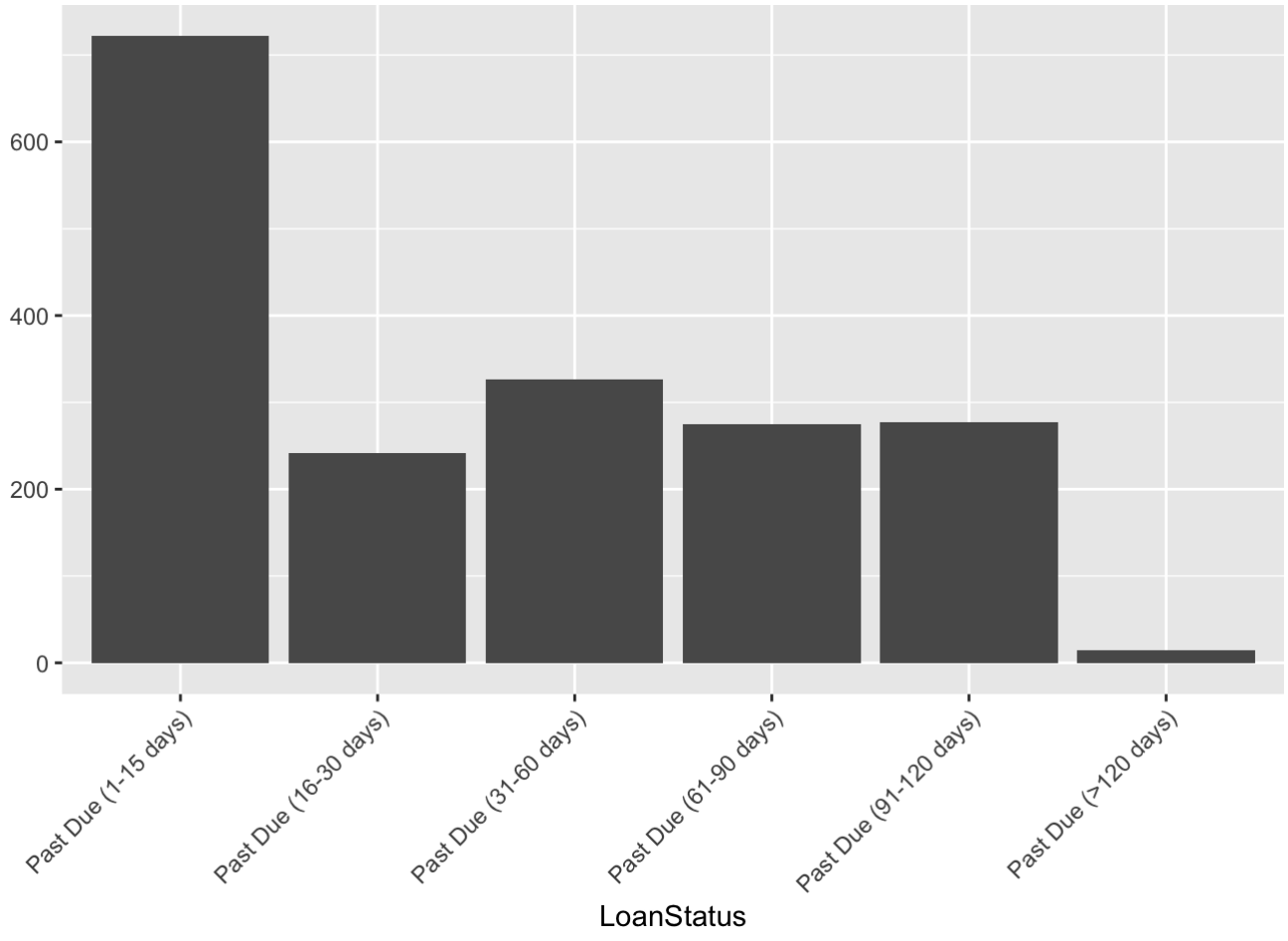
Debt to income ratio Box Plot



The first graph is difficult to see as there are several outliers that stretch the graph too much. By using coord_cartesion with expand = FALSE allows us to zoom in on the graph. We can see that the median is 22%.

Loan Status

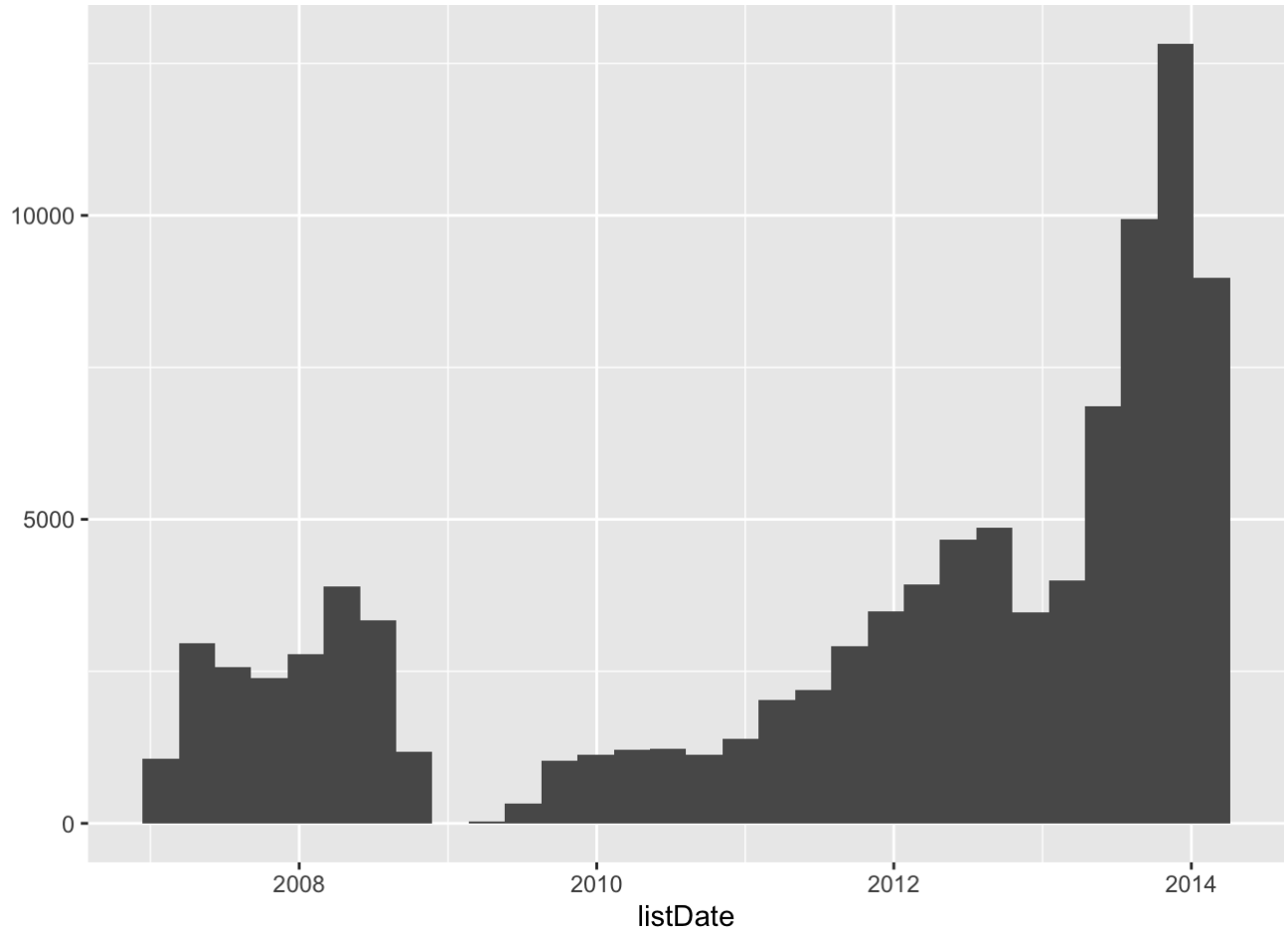

Past Due (91-120 days)
Past Due (61-90 days)
Past Due (31-60 days)
Past Due (16-30 days)
Past Due (1-15 days)
Past Due (>120 days)
FinalPaymentInProgress
Defaulted
Current
Completed
Chargedoff
Cancelled
LoanStatus
0
10000
20000
30000
40000
50000


600
400
200
0
Past Due (1-15 days)
Past Due (16-30 days)
Past Due (31-60 days)
Past Due (61-90 days)
Past Due (91-120 days)
Past Due (>120 days)
LoanStatus

```
##             Cancelled             Chargedoff              Completed
##                     1                   9399                  30787
##               Current              Defaulted FinalPaymentInProgress
##                 52478                   3054                    189
##   Past Due (>120 days)   Past Due (1-15 days)  Past Due (16-30 days)
##                    14                    722                    242
##  Past Due (31-60 days)  Past Due (61-90 days) Past Due (91-120 days)
##                   327                    275                    277
```

It seems almost 3 times more likely to be 15 days late on the payment then it is to be 30 days or more. This makes sense as people are more likely to not let there loans go too far past due. What seems odd is that the amount stay similar from 16 days to 120 and then dropped significantly after that.

By Year




```
##         Min.      1st Qu.       Median         Mean      3rd Qu.
## "2007-02-12" "2010-06-17" "2012-08-28" "2011-11-08" "2013-09-23"
##         Max.
## "2014-03-10"
```

The dates range from February of 2007 until March of 2014. It is interesting to see that there is a huge decrease to 0 around the end of 2008 and the graph slowly begins to climb back up. This is very likely due to the housing market crash of 2008. It looks like it took the bank several years to recover and then in 2014 loan amounts spike to almost 13,000.

Univariate Analysis

113,937 observations with 81 different variables. This was a bit larger than what I was wanting to work with right away, so I created a subset of 16 variables and stripped the NA values and that brought the amount of observations down to 97,765. This was still plenty of information to draw conclusions from.
You can see from the chart that credit score had a large and direct effect on the borrower's interest rate. Rates were higher than 30% with borrower's whose credit score was below 600 and the rate dropped to well below 10% for better credit. Prosper bank grouped the interest rates into their own credit scoring system.

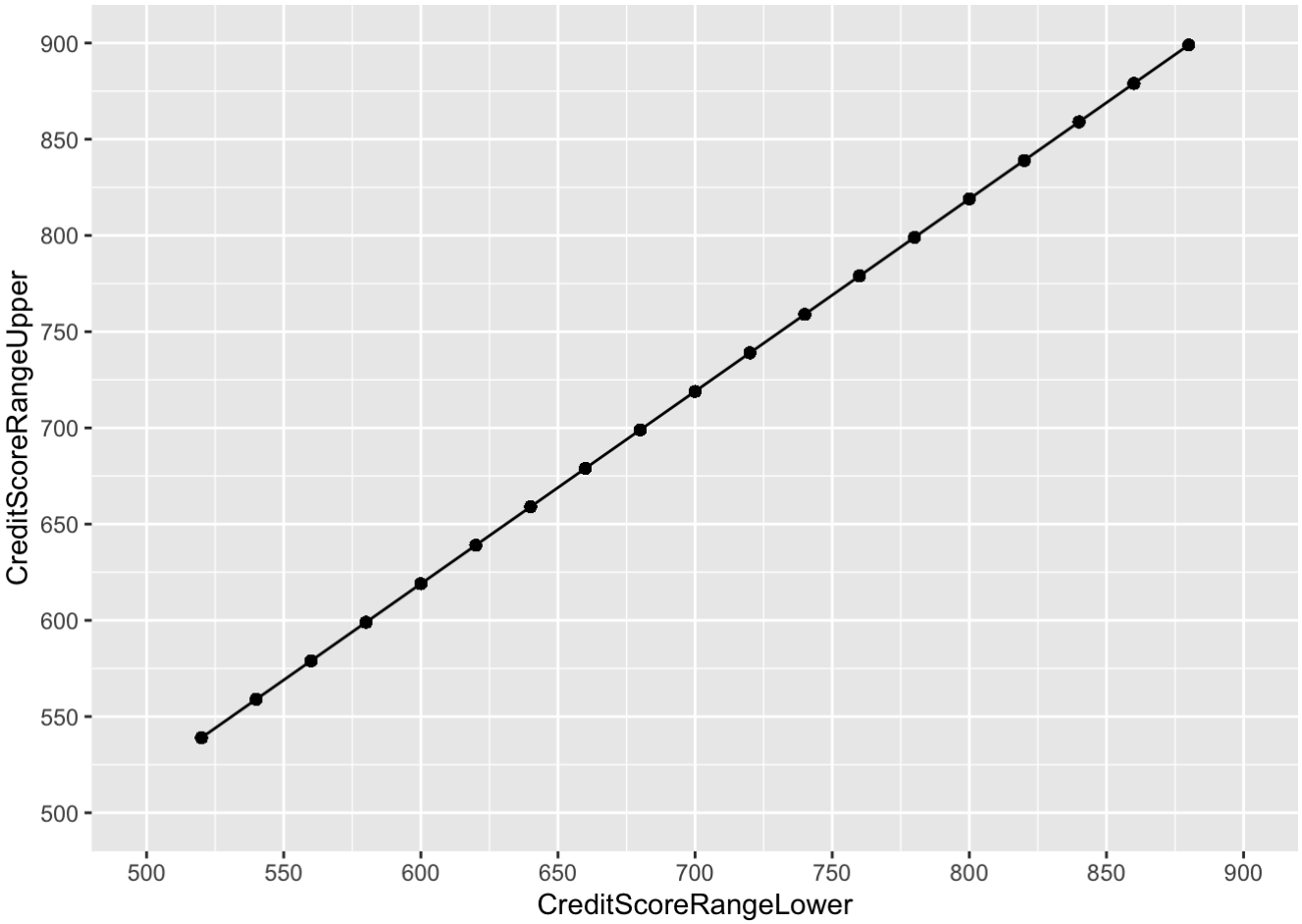
- AA Median score of 780
- A Median score of 720
- B Median score of 680
- C Median score of 640
- D Median score of 620
- E Median score of 560
- HR Median score of 520

Out of the 16 variables used, I created graphs with BorrowerRate, StatedMonthlyIncome, Term, CreditGrade, LoanStatus, AmountDelinquent, DebtToIncomeRatio, BorrowerState, and OpenCreditLines. I wanted to get a bigger picture of the type of data I was working with.

It was interesting to see the box plot of the Debt to income ratio. There were several borrowers that had DTI well above 200% and some reached as high as 1000%. Looking closer at the graph, it was easier to see that the median was 22%

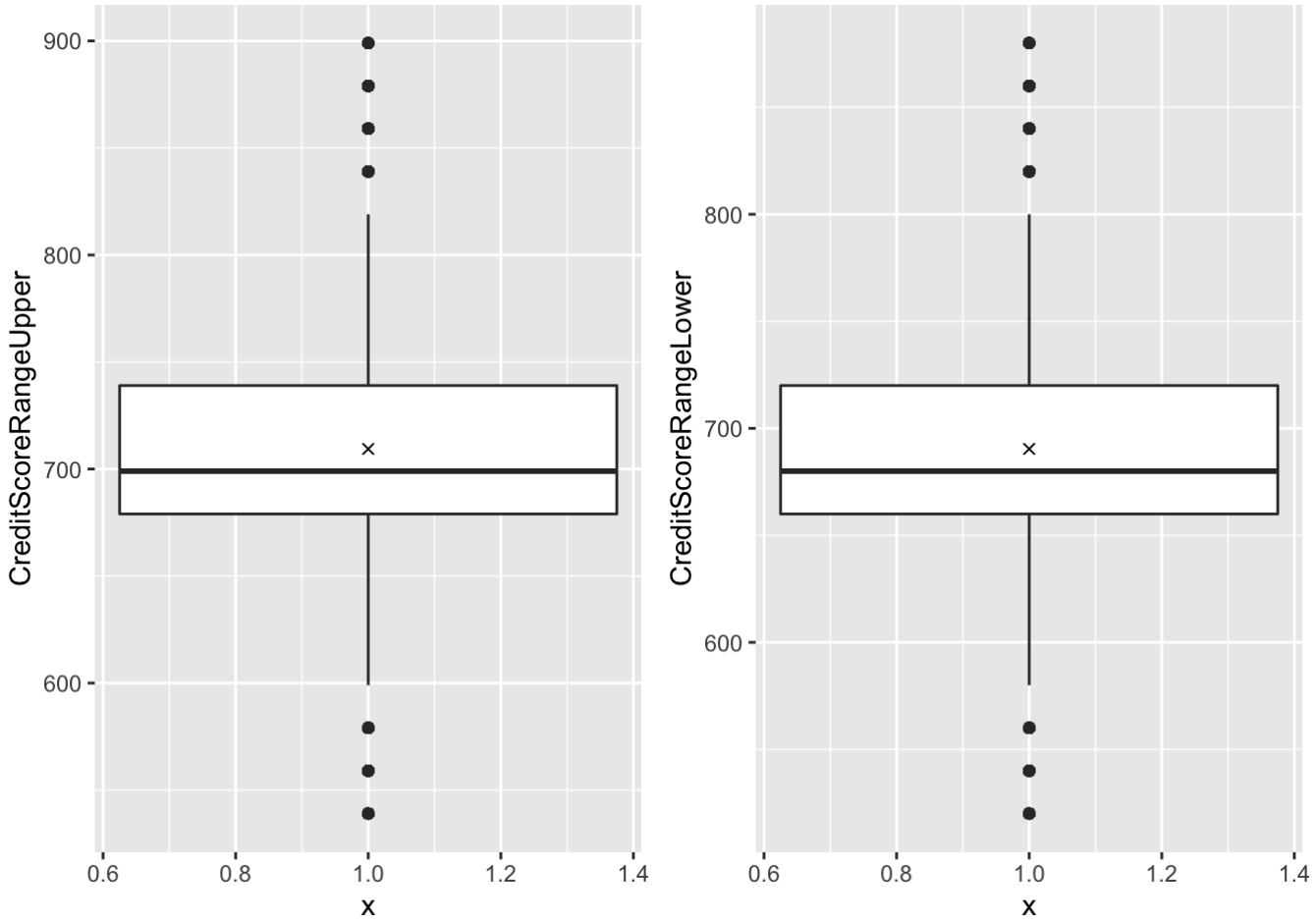
Bivariate Plots Section

Credit Score Range Upper vs Lower



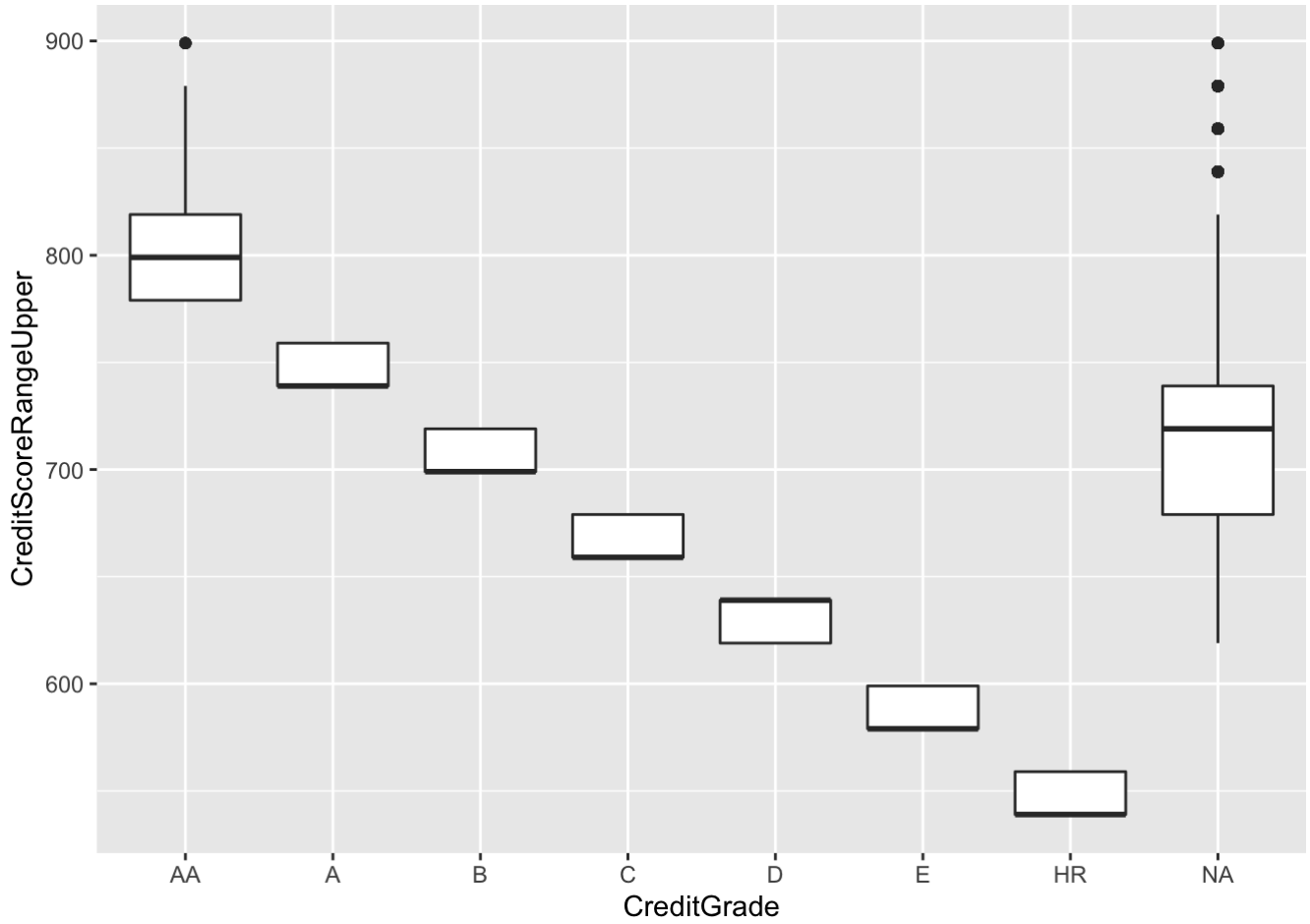

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   539.0   679.0   699.0   709.4   739.0   899.0
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   520.0   660.0   680.0   690.4   720.0   880.0
```



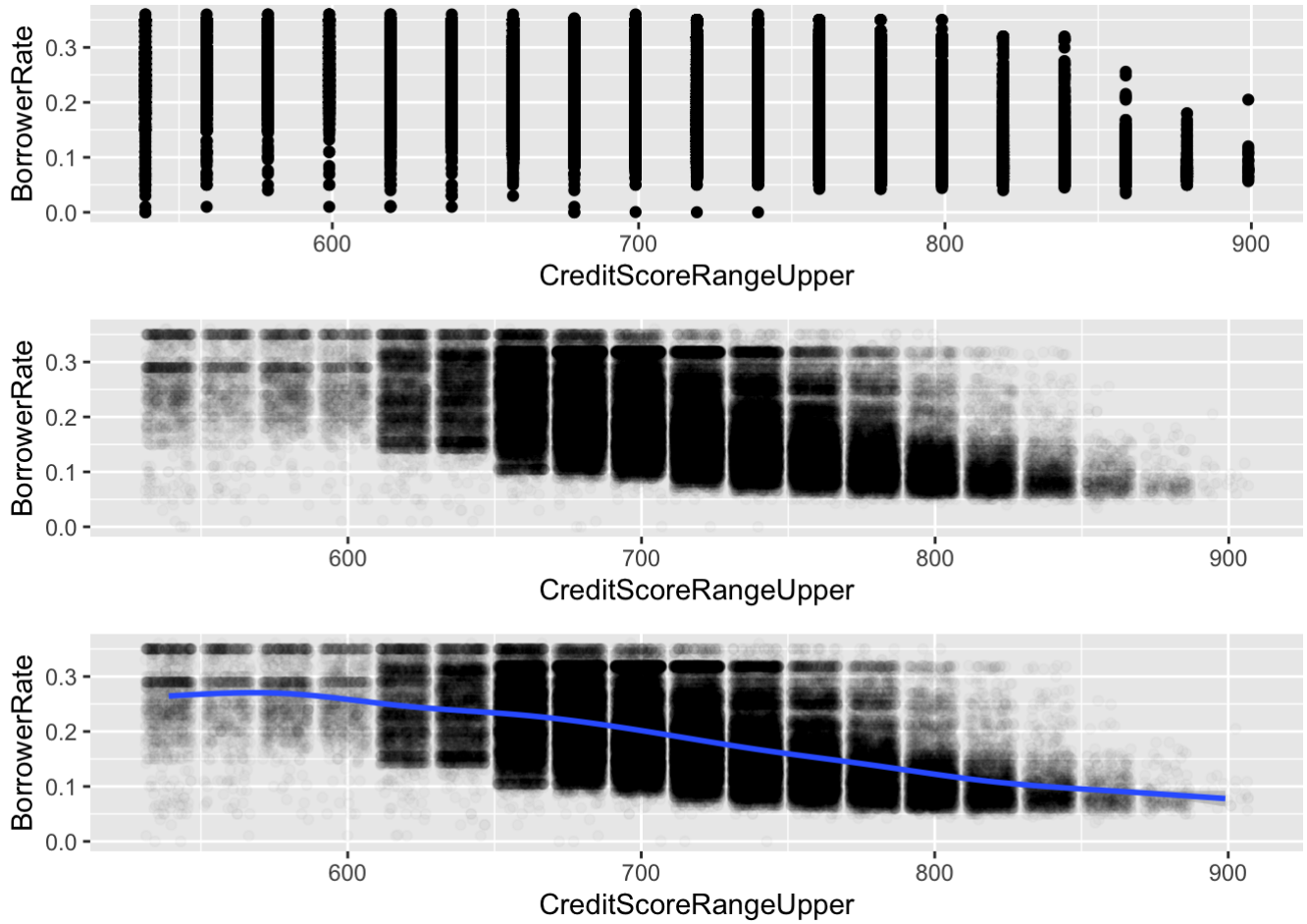
By reviewing the relationship between CreditScoreUpper and CreditScoreLower, you can see that there is only a difference of about 19 points accross the charts. The median upper credit score is 709 while the median lower credit score is 609.

Credit Grade / Credit Score Boxplot



You can see the scores evenly distributed within the Credit Grades based on their number. AA score has a mean score of 780 and goes all the up to 900, the highest available score.

Credit Score effect on Rate



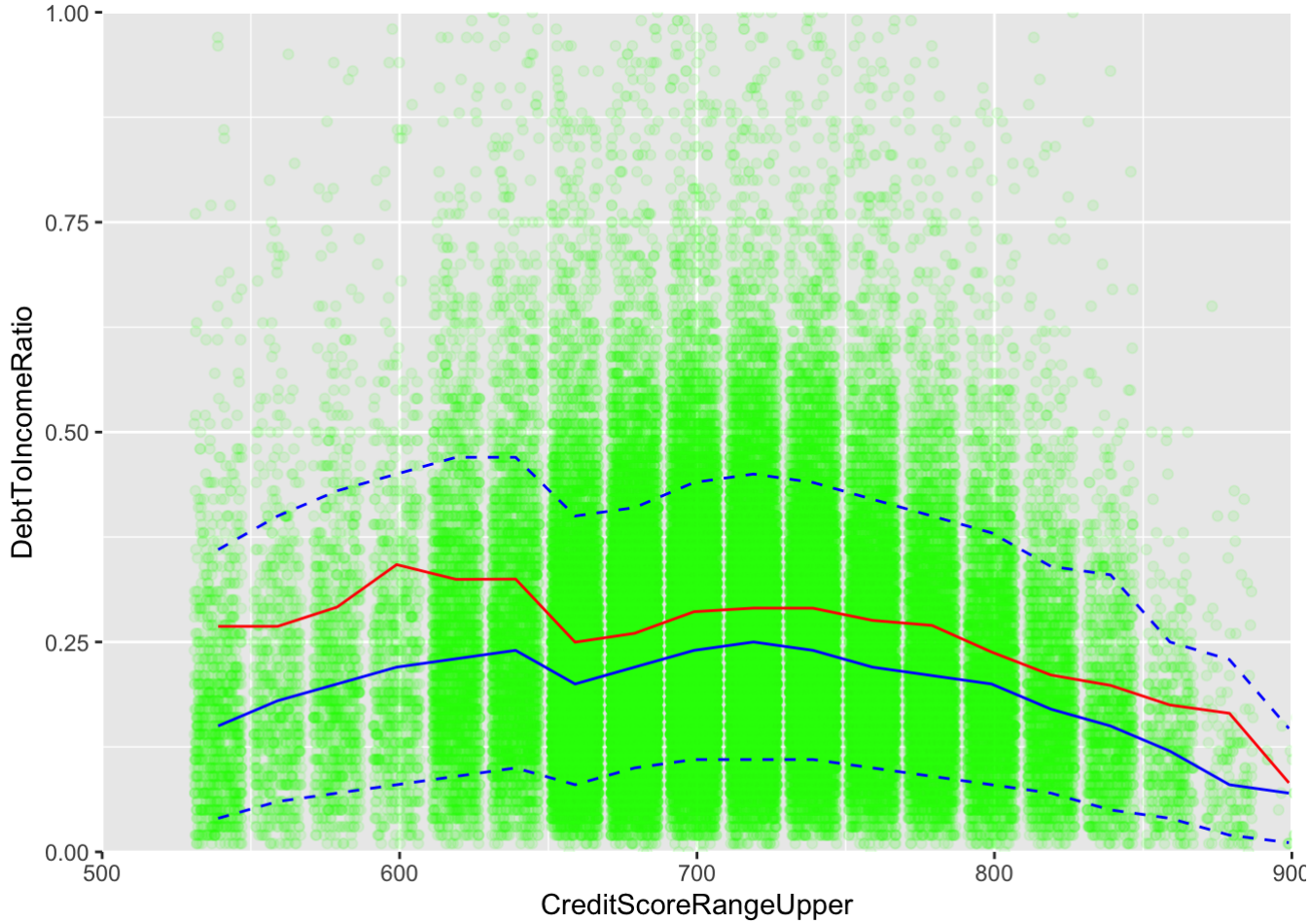
This scatter plot was very difficult to see any relation between Credit Score and Borrower Rate. I transformed the data so the plotted points were more transparent and added a linear regression line. You can see that the the interest rate decreases as the credit score improves.

Credit Grade vs Credit Score

```
## # A tibble: 8 x 4
##   CreditGrade credit_score_mean credit_score_median     n
##   <ord>                   <dbl>               <dbl> <int>
## 1 AA                       784.                 780  2542
## 2 A                        729.                 720  2510
## 3 B                        689.                 680  3307
## 4 C                        648.                 640  4354
## 5 D                        610.                 620  3788
## 6 E                        569.                 560  1776
## 7 HR                       528.                 520  1811
## 8 <NA>                     699.                 700 77677
```

I was curious to see if the credit grade is solely determined on credit score or if there were other factors that effected it. By looking at the table, it is clear that credit grade is only determined by the credit score of the borrower.

Credit Score Range and Debt To Income Ratio



It is interesting to see that there is a normal curve in this graph with a sudden dip in DTI right around a credit score of 650.

Bivariate Analysis

I analyzed the relationship of the credit score in this section by using different charts to visualize the relationship of the data. CreditScoreRangeUpper and CreditScoreRangeLower are similarly related and only varying by 20 points from each other. I wanted to dig deep into the credit analysis of this data and see what the relationships were like between the different variables and to see if there was any correlation.

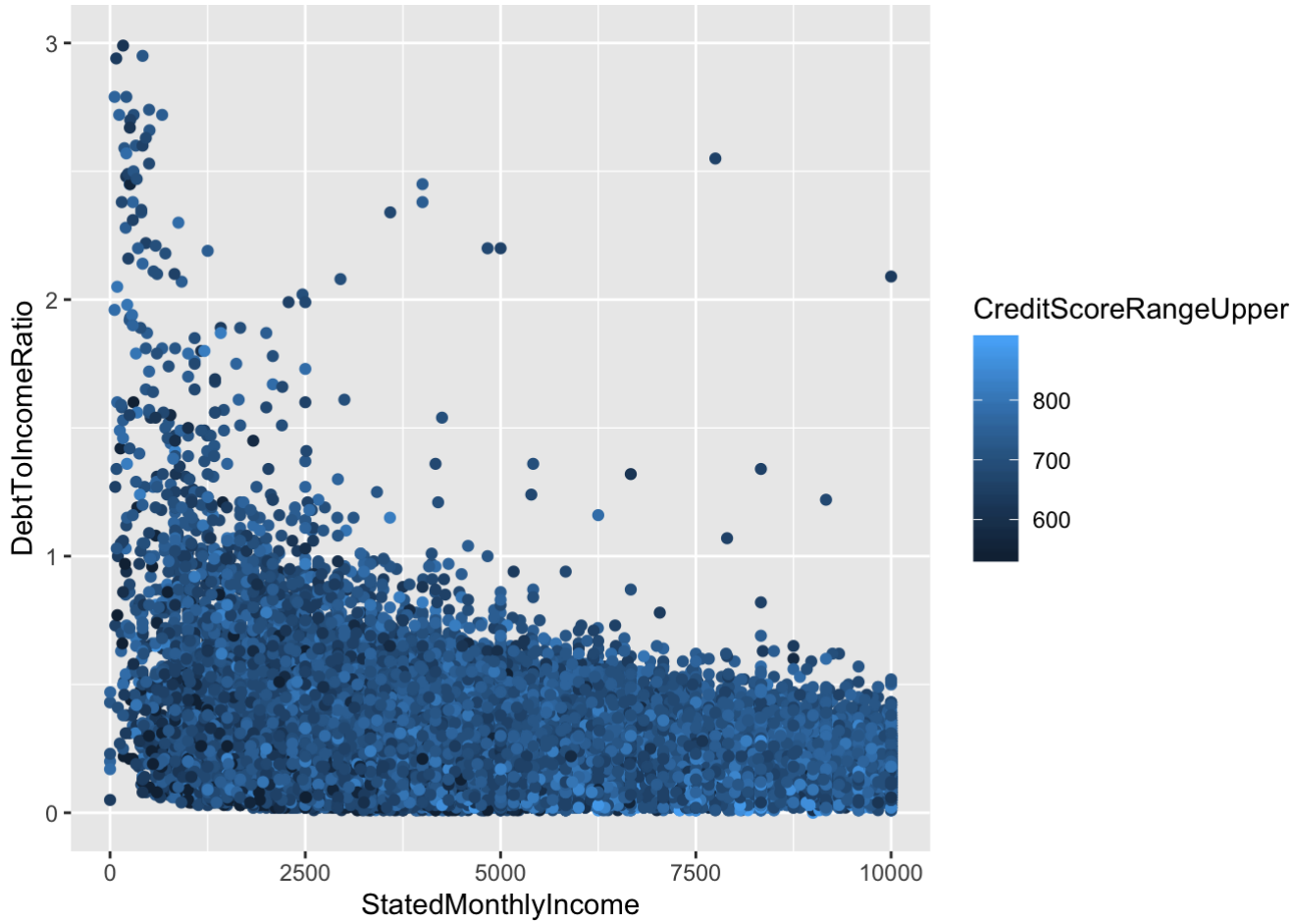
When comparing the credit score and DTI ratio, it showed an unforseen dip in the DTI by about 10% around a credit score of 650 and then went right back up when the credit score reaches 700.

The Credit score had a strong effect on the borrower rate. It was difficult to see the relationship at first but by adding the linear regression line, it was prominent that the better the credit score, the better the rate.

I created a Credit Grade Group variable to see see if the loans were grouped together by Credit Score only or if there may be another variable involved. Based on my analysis, the Credit Grade Group is only determined by credit score.

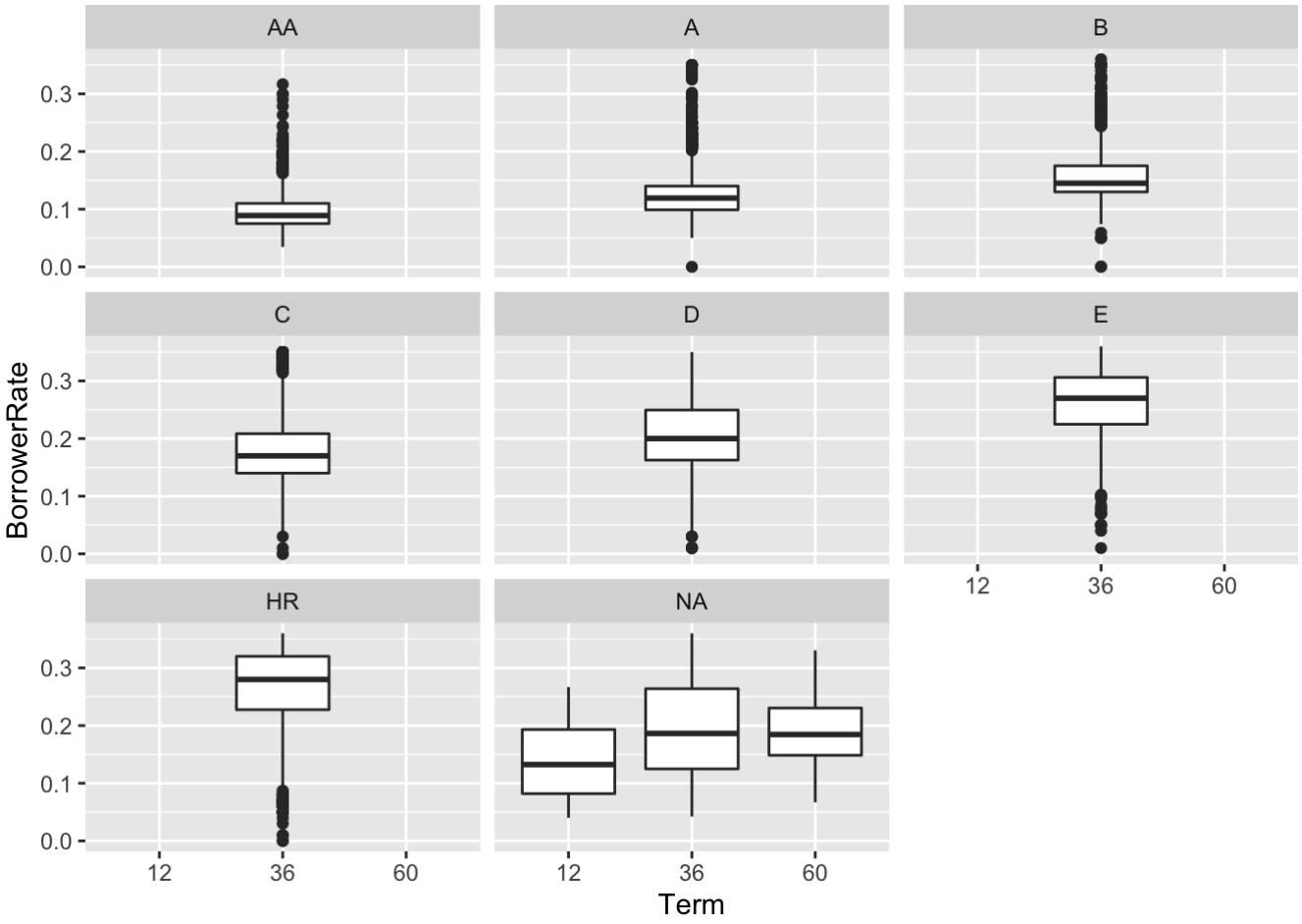
Multivariate Plots Section

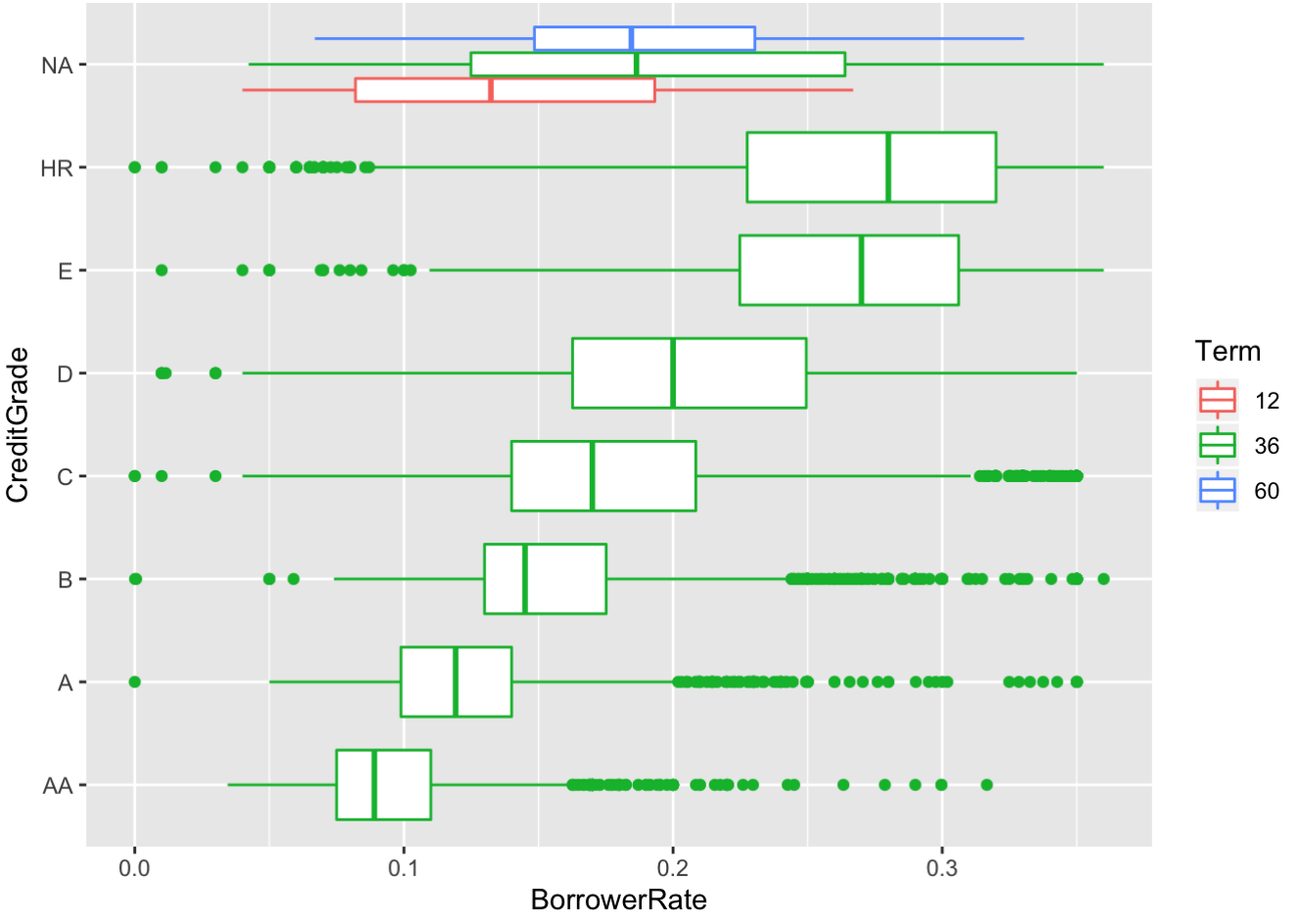
Stated monthly income, DTI and credit score?



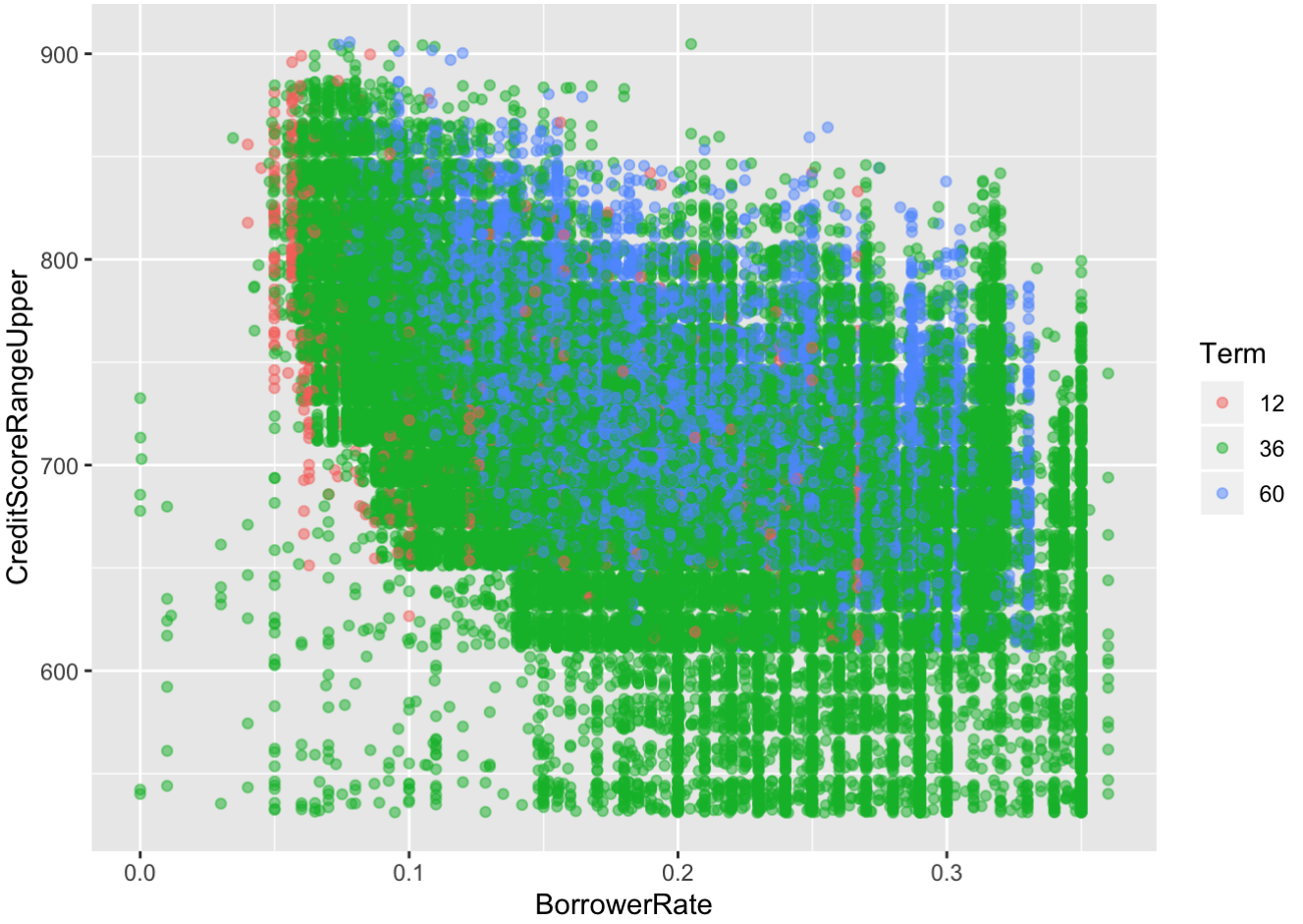
You can see on this graph that a large percentage of borrowers fall below a DTI of 1. However, it is far more likely for a borrower to have a higher debt to income ratio if their income is less than $2,500. By added a color attribute to Credit Score, it is easy to see that most of the lower credit scores are in the lower income range and usually have a higher DTI.

CreditGrade, BorrowerRate and Term


AA
A
B
C
D
E
HR
NA
BorrowerRate
0.0
0.1
0.2
0.3
12
36
60
Term

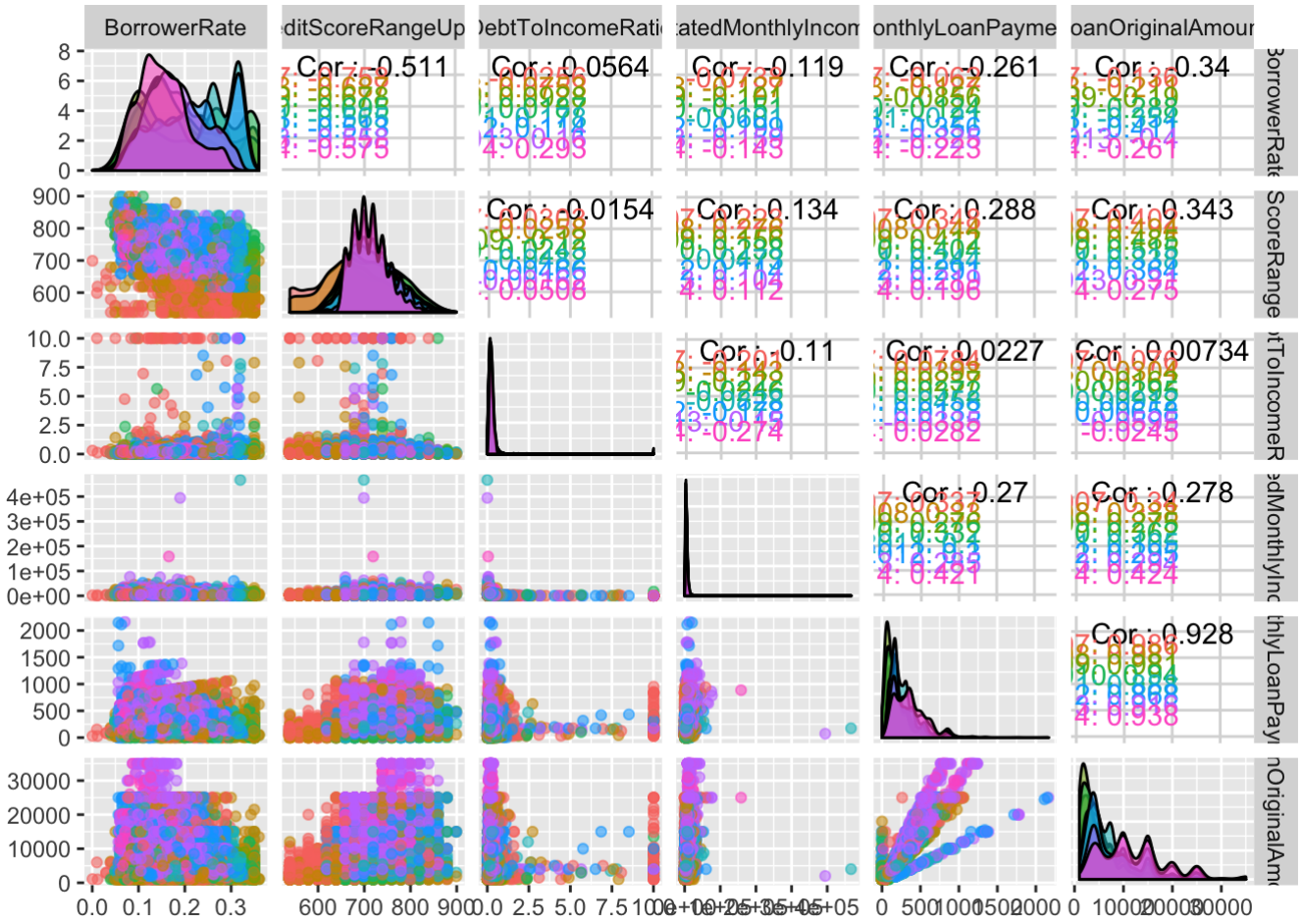

CreditGrade
NA
HR
E
D
C
B
A
AA
0.0
0.1
0.2
0.3
BorrowerRate
Term
12
36
60

I wanted to see the relationship between CreditGrade, BorrowerRate and Term to see if there was a major effect and the different variables and I mistakenly discovered that Prosper Bank has not provided credit grades to terms other than 36 months. Not the information I was expecting but certainly useful in getting a full picture of the data.



This graph shows a better relationship between the 3 variables that I was not able to see on the last graph by changing the CreditGrade variable to CreditScoreRangeUpper and plotting a scatterplot graph rathen than a box plot. You can see that the 12 month terms typically have higher credit scores and lower DTI ratios then longer terms. 36 month terms are far more common than others. 12 and 60 month terms seem to be reserved for those with credit scores of 600 or higher.

Scatterplot Matrix



Running a scatterplot matrix, I was able to graph a large amount of information into several table to compare. For this, I used a sample size of 25,000 that showed correlation coefficients for 6 of the quantitative variable and shows the year based on color.

Variables used: - BorrowerRate - CreditScoreRangeUpper - DebtToIncomeRatio - StateMonthlyIncome - MonthlyLoanPayment - LoanOriginalAmount

Multivariate Analysis

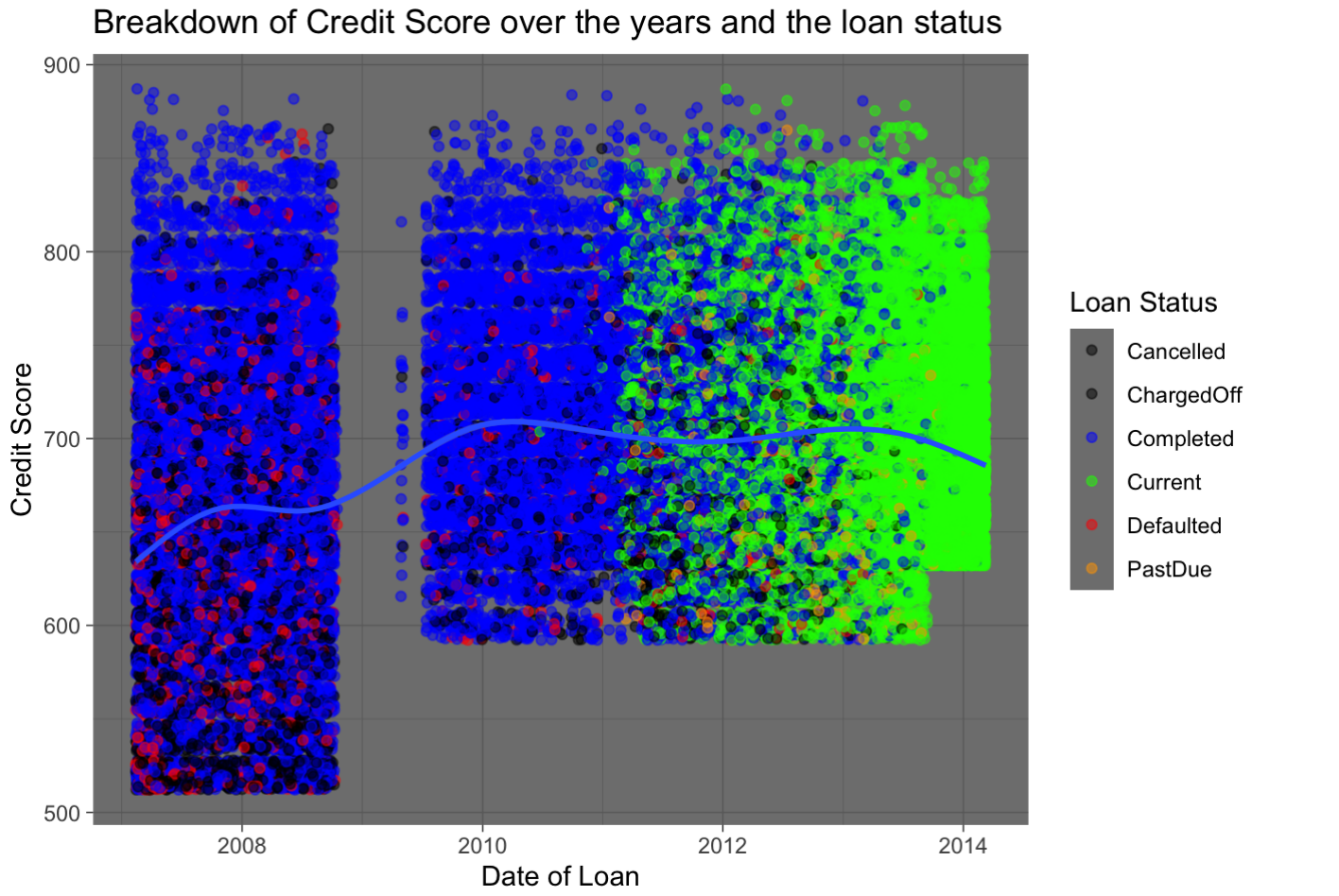
It was interesting to see the relationship between DTI, Credit Score and Monthly income. I had to restrict a lof of the outliers to see the data that I was interested in. By adding a color range to the credit score I could point out the high and low points in the graph and their location showed the DTI and Monthly income for that borrower.

The graph with CreditGrade, BorrowerRate and Term surprised me as it showed that Prosper Bank had not given Credit Grades to loans other than 36 months. This prompted the next graph where I replaced CreditGrade with CreditScoreRangeUpper. This provided more information as to what I was looking for.

The Scatterplot Matix was able to provide a lot of information on the variables analyzed and seeing how the different years were effect by each.

Final Plots and Summary

Plot One



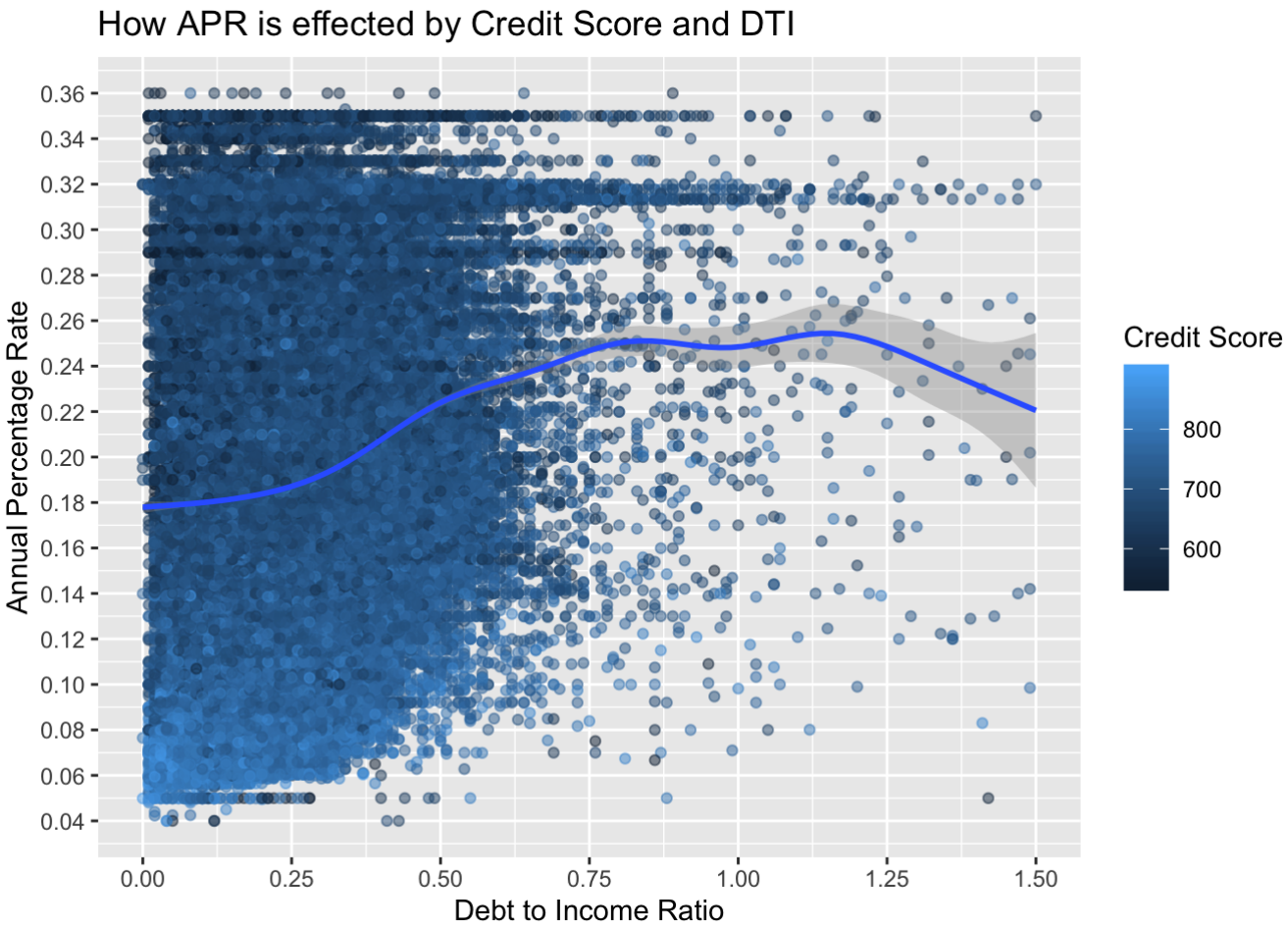
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   539.0   679.0   699.0   709.4   739.0   899.0
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   520.0   660.0   680.0   690.4   720.0   880.0
```

Description One

This graph shows the history of the loans given, the borrower's credit score at the date of the loan and the current status. The loans are grouped into 3 time periods, before 2009 Prosper Bank allowed borrower's with credit scores as low as 530. After 2009 the bank seems to have increased their minimum score to an mean credit score of 600. Around mid 2013, the minimum mean credit score was again increased to about 640.
The housing market crash of 2008 seemed to have effected a lot of borrowers as most of the Charged Off loans fall before 2009 and had a credit score of below 600. The policy change to have a minimum mean score of 600 decreased the chances of the defaulting.
The green points are most populated on the right of the graph. This makes sense as most of the loans are completed in about 3 years. After that timeframe, they are completed, charged off or defaulted.

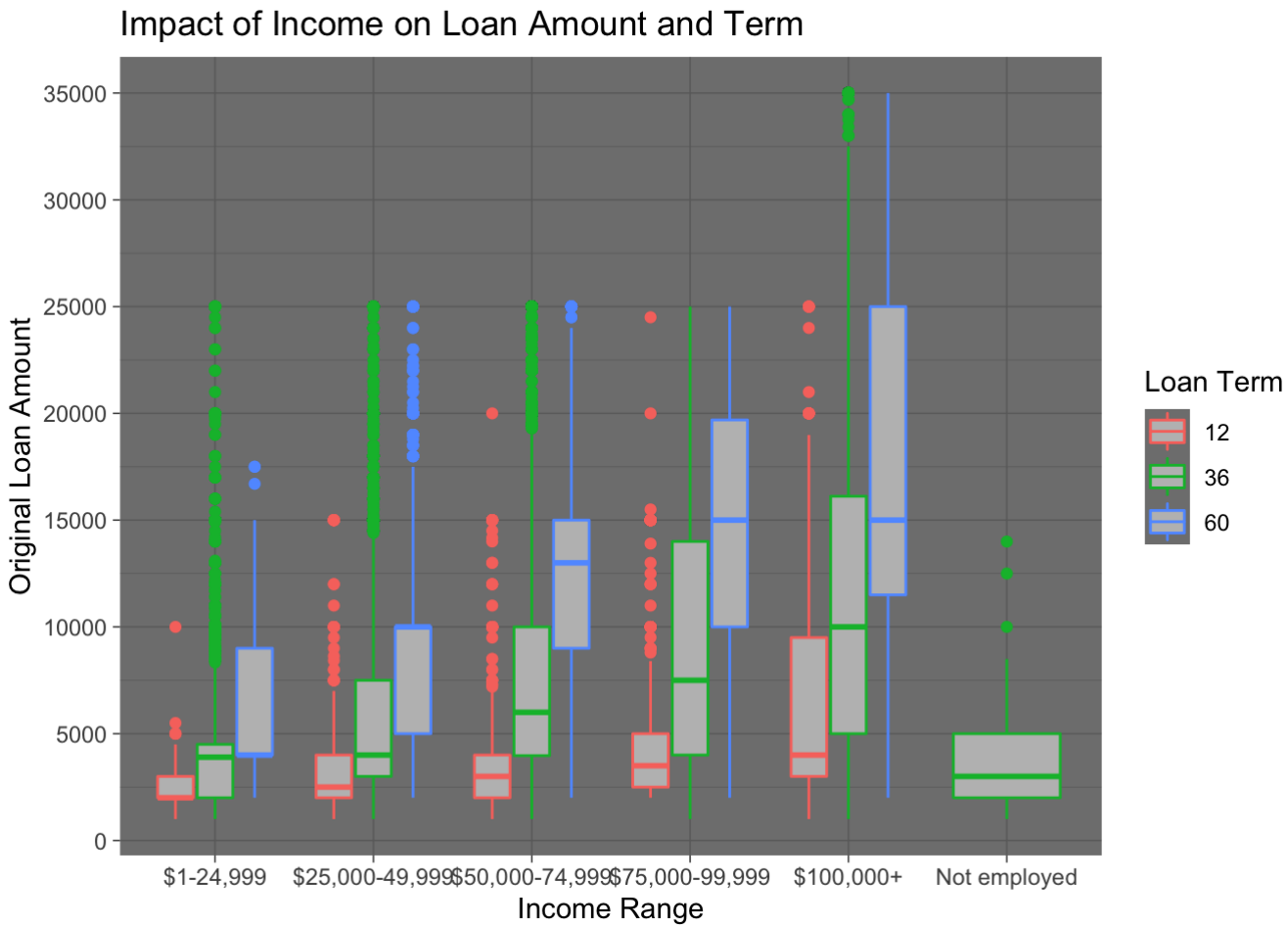
Plot Two



Description Two

You can see a great effect how credit score impacts interest rates as the plot shows lighter points (Higher Score) at the bottom and darker points (Lower Score) towards the top of the graph. There seems to be a few groupings along the 32% rate and the 35% rate. I can assume this is due to the different market changes and what the highest interest rate is during that period. Most of the groupings here seem to be darker as so would getting the worst rates. The DTI Ratio seems to have just a slight effect on Rate when where it will start to increase around 25% DTI but will level out around 75%.

Plot Three



Description Three

There is an obvious relation between income range and the amount lent to the borrower. There is a loan amount limit of $25,000 for those whose income is less that $100,000 and then increases to $35,000 for high income earners. About 90% of all 12 month terms are less than $10,000. It is plain to see that the loan term increases when the loan amount increase. This makes sense as it would take longer to pay back a larger loan with the same monthly payment.

Reflection

In our analysis of Prosper Loans, we started out with 81 variables and filtered most to bring the amount down to 17 to analyze. We ended up adding 4 more to help with some of the data we were exploring. It is difficult to work with missing data so we cleaned up the data by removing NAs. The brought the number of loans from 113,937 to 97,765. I cleaned up the data a little more by factoring the IncomeRange and the CreditGroup and ordering the variables to assist with plotting the graphs.

It was very difficult for me to split the data value of the Date time group under ListingDate into two separate variables. But using the strsplit() function, I was able to extract the data as needed for my analysis and then convert the data to a Date with as.Date(). And then to create another variable with just the Year to help with my

scatterplot matrix.

This was a very large dataset with a lot of data to work with. One of the limitation that I noticed and would have liked to explore is the credit score at the time of loan opening vs credit score when it closed, to see how the average score changes over time and if there is a increase, decrease or no movement. I would also like to investigate further to predict success of future loans based off of the data of past and current loans. Overall, it was a great dataset to work with and I learned a lot about R and the different packages available for use.