Prosper Loan Exploratory Data Analysis

Mustafa Olomi

11 20, 2017

Table of Contents

Introduction:

This is an exploratory analysis of the data from the peer to peer lending company Prosper. The dataset contains over 100,000 borrowers with 84 variables. Prosper connects lenders, who are individual investors that can selectively invest their money in loans to borrowers based on a variety of criteria like credit score, prospers rating grade, profession, loan type & more.

In this analysis I aim to visualize and understand the characteristics of a good loan prospect and some of the qualities that make up bad loans, as well as Prosper's overall customer and loan profiles.

Univariate Plots Section

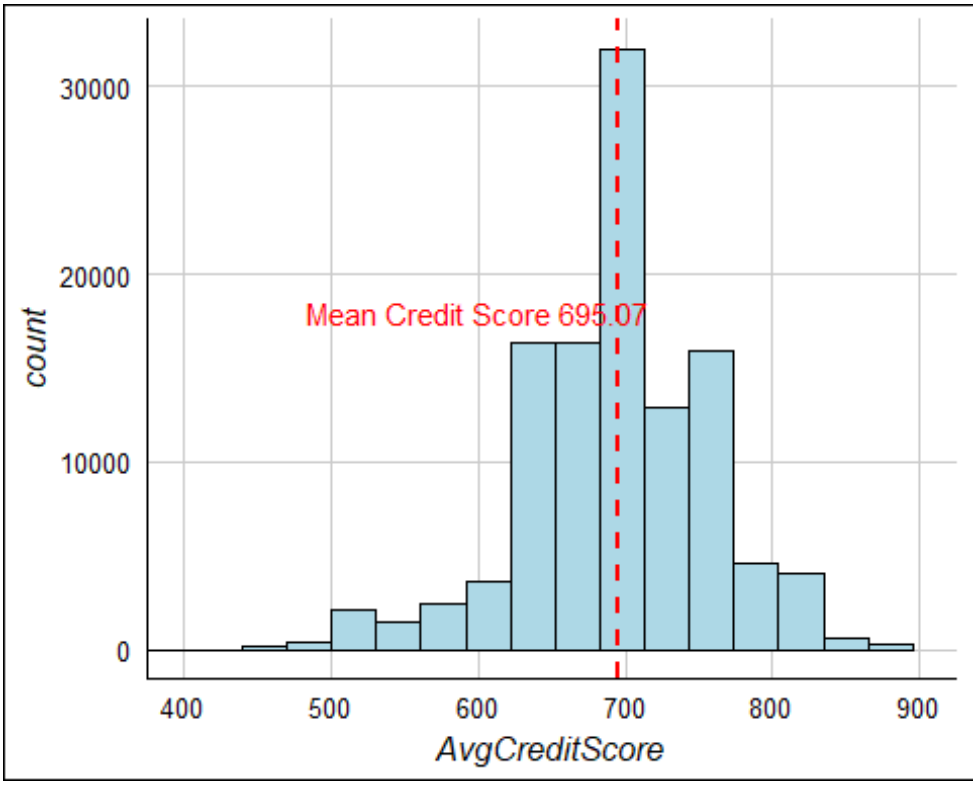
In this section, I perform some preliminary exploration of the dataset

Firstly I take out the cancelled loans because there are only a few cancelled loans and the loans were not initialized so we don't need them in our data

The avg credit score will give us a better datapoint to measure against

Plotting the distribution of avg credit scores.

Most traditional companies want to know the credit score of the portfolio, as it is a more universal measurement.



We can see the mean credit score for all loans is 695 This is considered "good" credit according to Experiean

300-579 is Very Poor (17% of people) may not be approved for credit at all

580-669 is Fair (20.2%) considered subprime

679-739 is Good (21.5%) 8% liklihood of serious delinquency

740-799 is Very Good (18.2%) likely to receive better than avg rates

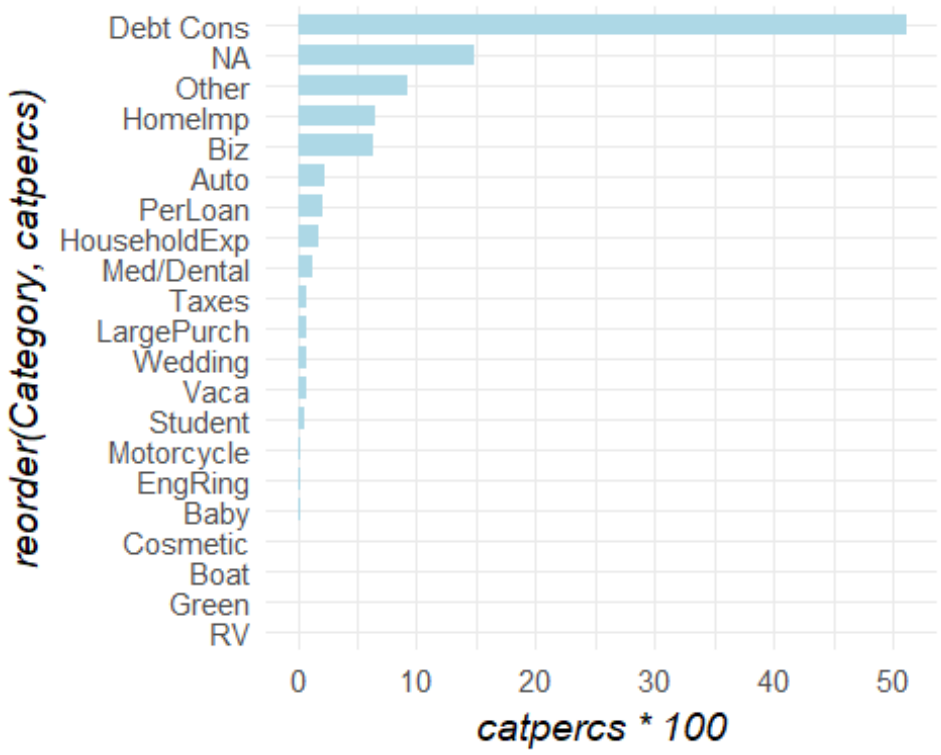
800-850 is Exceptional(19.9%) at the top of the list for best rates with lenders

https://www.experian.com/blogs/ask-experian/credit-education/score-basics

Creating labels for each of the categories

Debt consolidation is the highest category with over 50% of the loans

```
## # A tibble: 21 x 3
##        Category count    catpercs
##          <fctr> <int>       <dbl>
##  1    Debt Cons 58307 0.511770179
##  2           NA 16961 0.148869501
##  3        Other 10494 0.092107573
##  4      HomeImp  7433 0.065240670
##  5          Biz  7189 0.063099042
##  6         Auto  2572 0.022574869
##  7      PerLoan  2395 0.021021311
##  8 HouseholdExp  1996 0.017519222
##  9   Med/Dental  1522 0.013358846
## 10        Taxes   885 0.007767791
## # ... with 11 more rows
```



In the documentation it says: Before 2009 prosper generated grades with the variable CreditGrade

After 2009 they used the grade ProsperRating Alpha

Here we are creating a universal grading standard with all grades

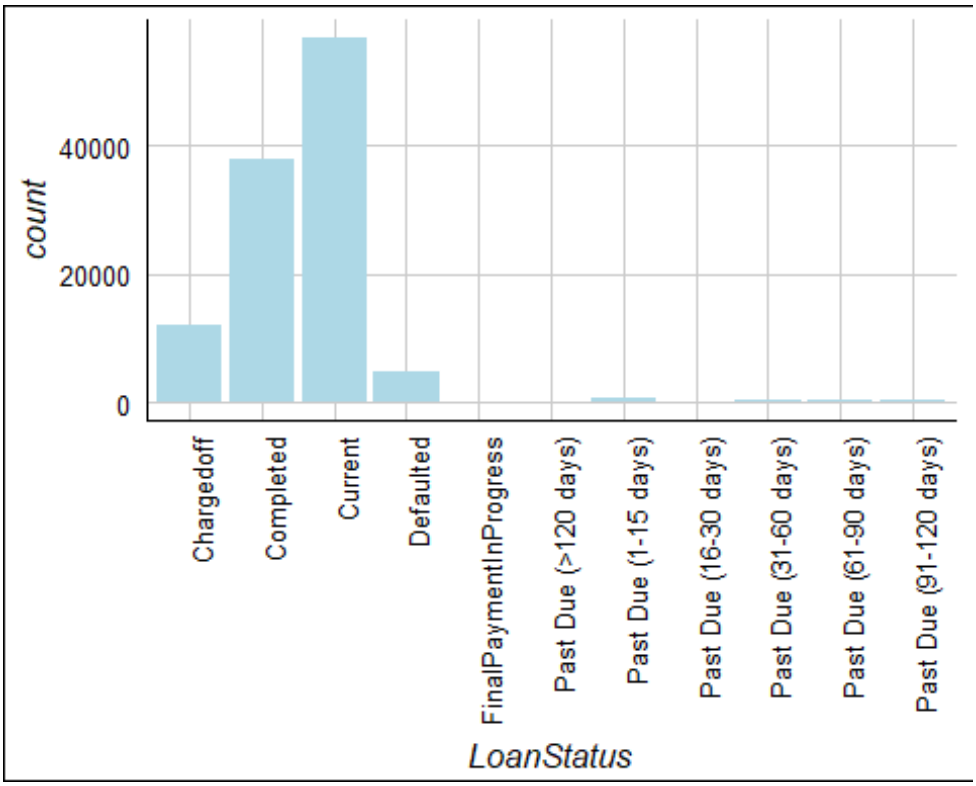
```
##
##          A         AA          B          C          D          E
## 0.15717931 0.07813655 0.17569945 0.21109449 0.17092205 0.11511526
##         HR
## 0.09185289

##    AA     A     B     C     D     E    HR
##  8881 17865 19970 23993 19427 13084 10440
```

We can see we have many categories, including 6 past due categories We could simplify this an better define the categories

```
##               Cancelled             Chargedoff              Completed
##                       0                  11951                  37910
##                 Current              Defaulted FinalPaymentInProgress
##                   56576                   4951                    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
```

We can see that most of the statuses are centered around being complete, current or chargedoff/defaulted

According to Prosper's website: A borrower loan is charged-off when it reaches 121 days past due.



Lets check to see where we should categorize the final payment loans based on their % funded

Because they are 99.65% funded we'll categorize them as 'Complete'

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.7055  1.0000  1.0000  0.9965  1.0000  1.0000
```

The data is centered around complete, current and chargedoff/defaulted so I am regrouping the data below into those categories

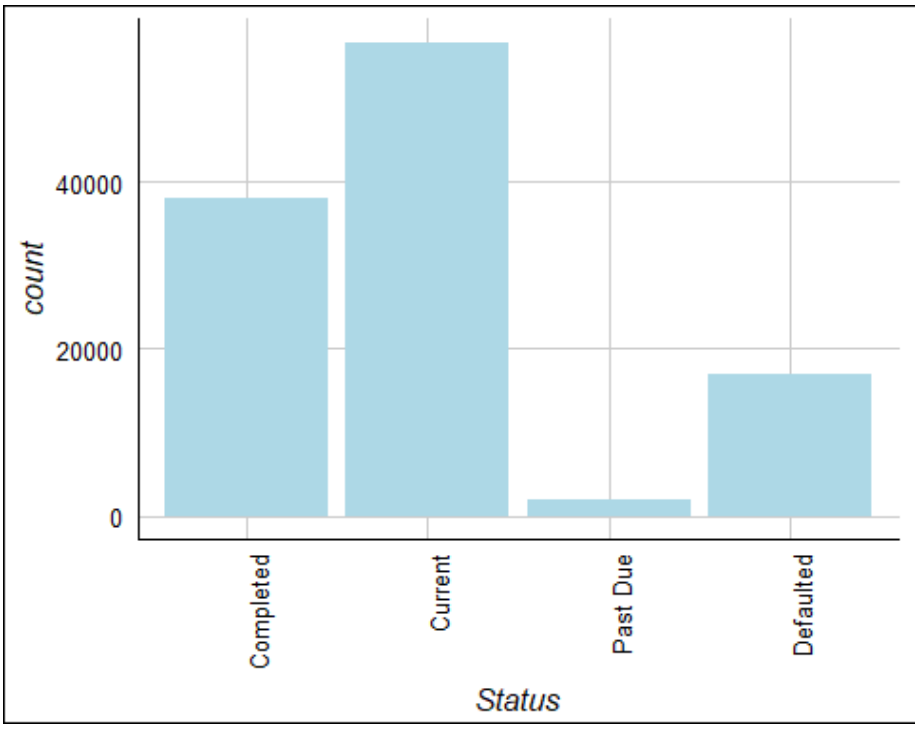
Summarizing all past dues as one category so we can easily see them together

Summarizing charged off & defaulted as one category because they are similar

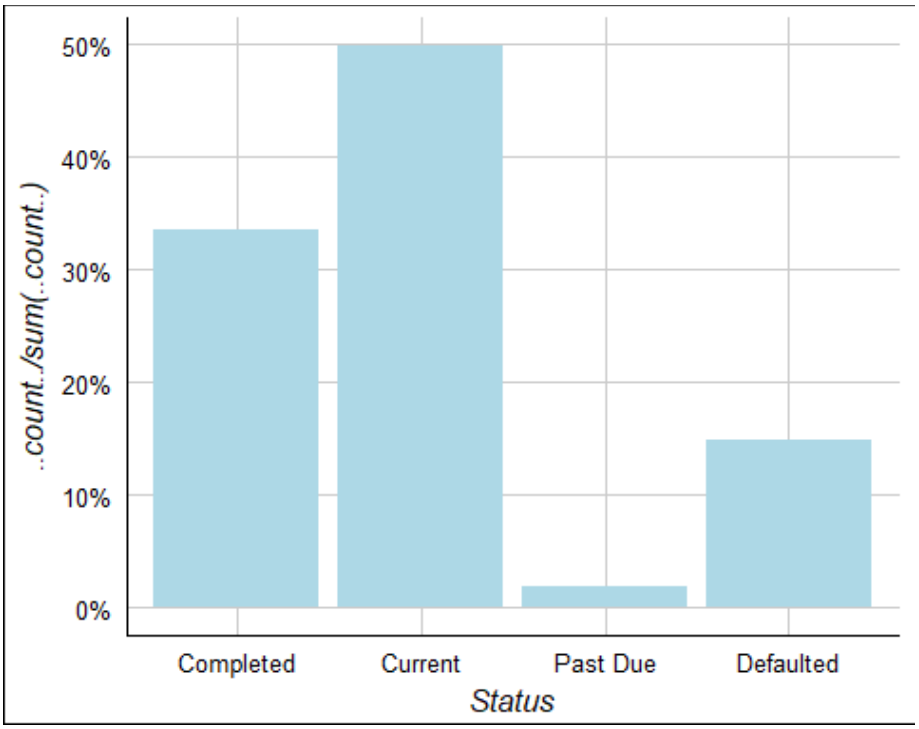
We can see the loans as a narrower category range & better view the loans and what standing they are in

```
## Completed   Current  Past Due Defaulted
##     38112     56576      2070     16902
```

This is the distribution of completed, current, past due and defaulted loans

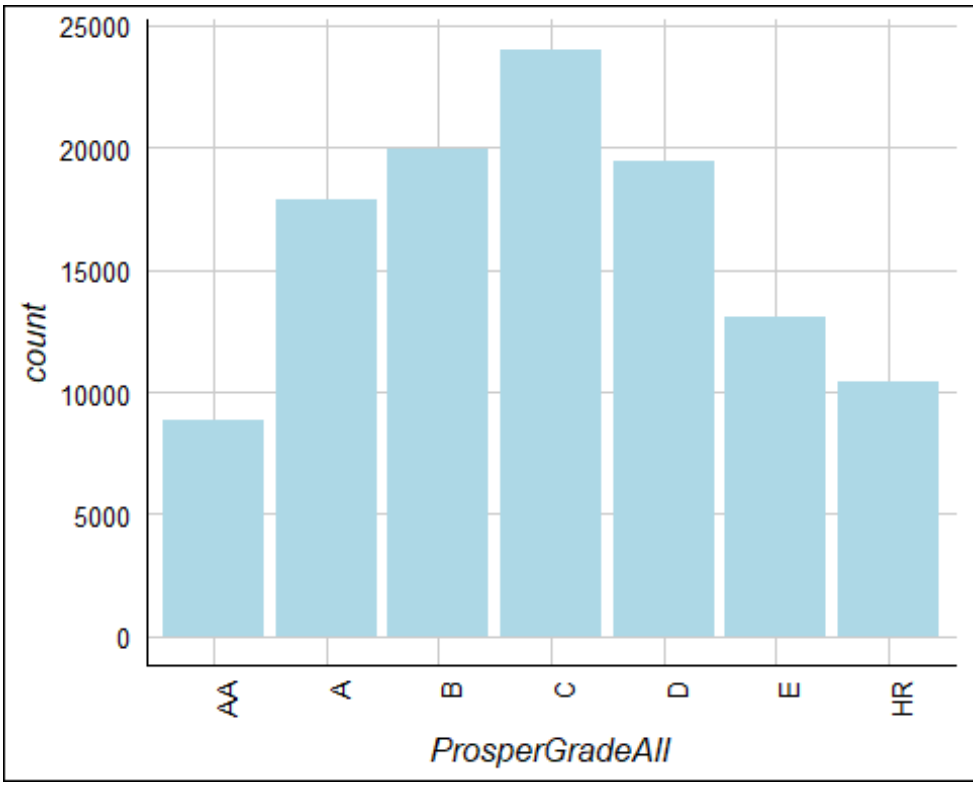


This is the same distribution but is a percentage distribution



```
##    AA     A     B     C     D     E    HR
##  8881 17865 19970 23993 19427 13084 10440
```

Interesting to see that it does look like a normal distribution curve Even though their credit ratings are not as normal

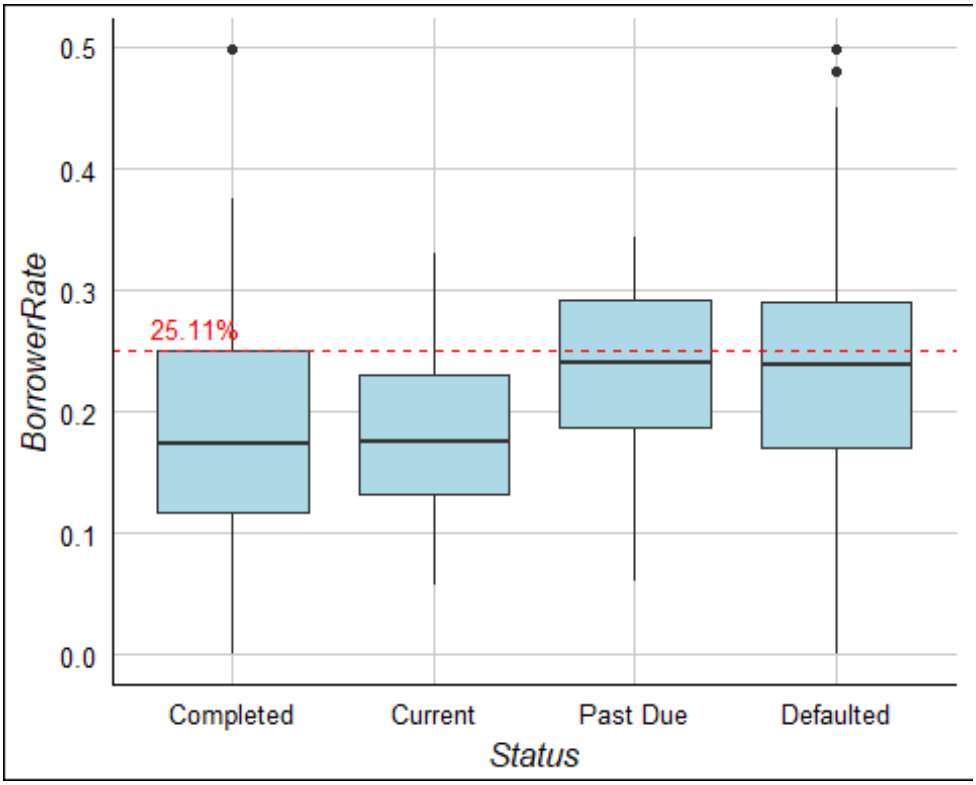


Bivariate Analysis

Now looking at the interest rate among the 4 main categories we see the top 75 percentile of completed loans were 25.11%.

It's also interesting to see that that is near the median of bad loans

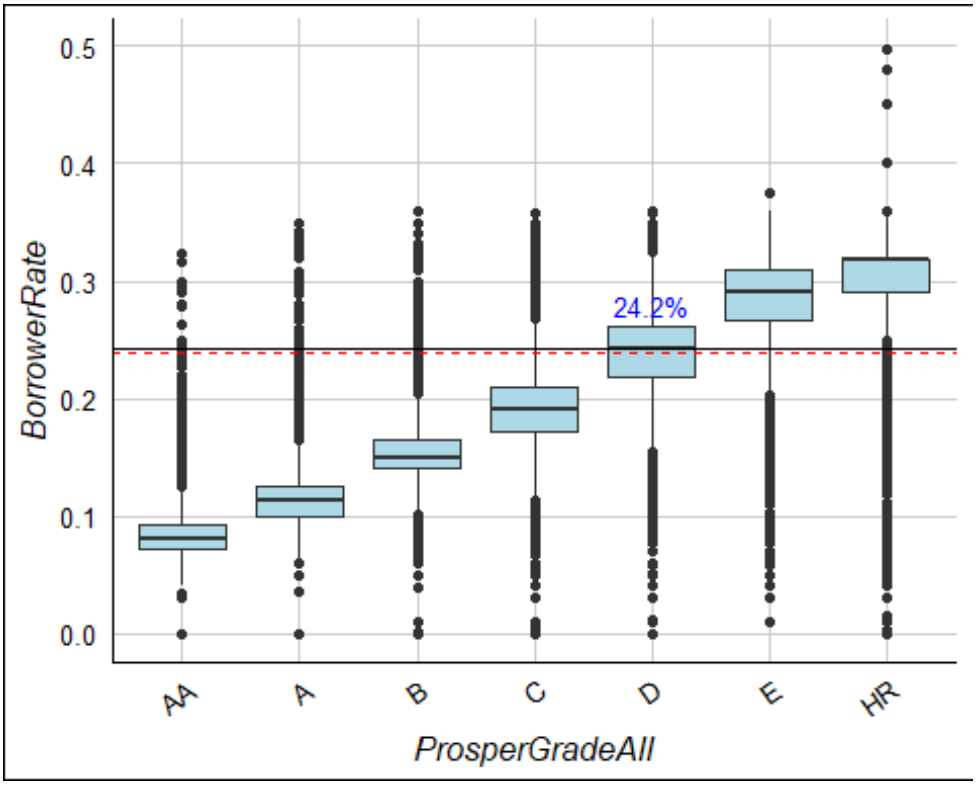
Past due and defaulted loans have a median just under 25.11% so this is a good indicator of when a loan may be too expensive or too high risk for clients.




```
##    75%
## 0.2511
```

This shows the distribution of prosper grades and interest rates and we can see that 24.2 is the median rate for D grade loans.

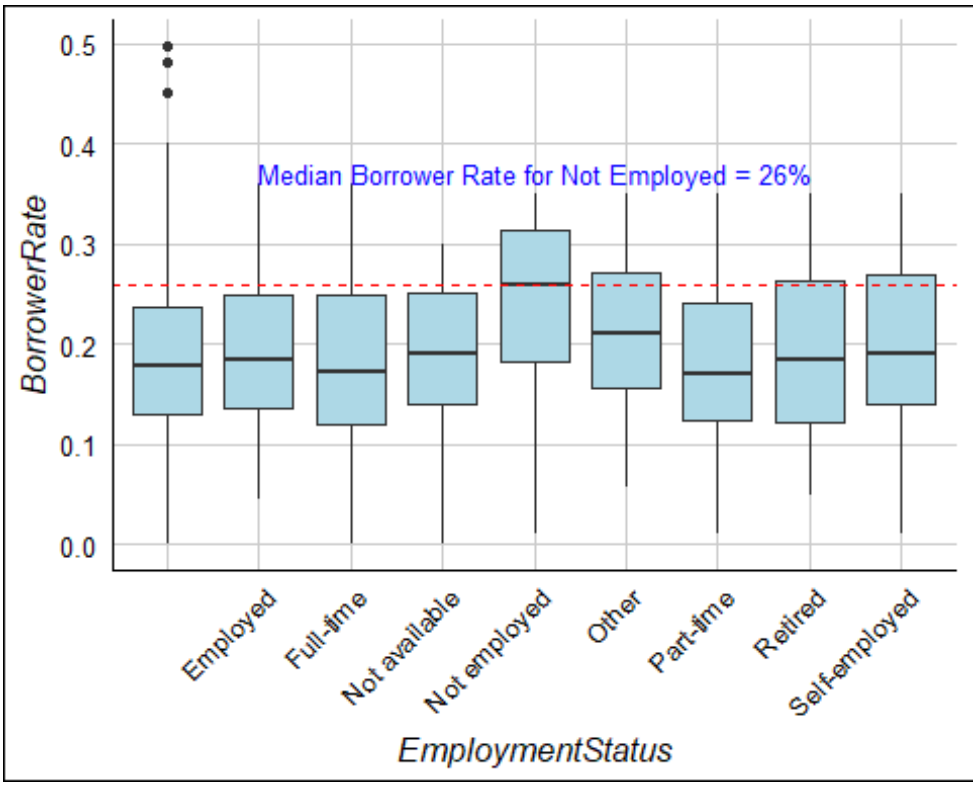
Interesting that this is close to the median of past due/defaulted loans 23.85 is the median for defaulted vs 24.19 for D grade loans




```
## [1] 0.2419
```

```
## [1] 0.23845
```

Unemployed borrowers had a median rate of 26%, substantially higher than loans to employed borrowers.




```
## [1] 0.2599
```

```
## [1] 0.23845
```

The most interesting relationship I saw was the fact that bad loans tended to be around the 25% interest rate level, that's the median for defaulted loans, and around the median for D grade loans and for unemployed borrowers

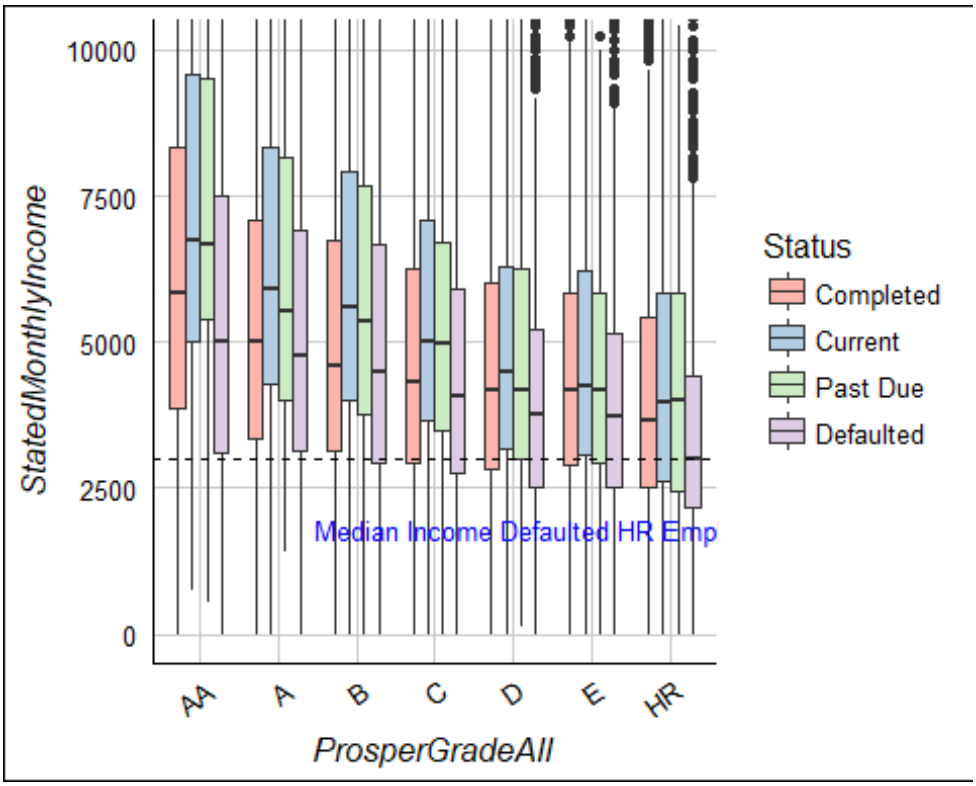
Whatever the case, borrowers willing to pay up to 25% are usually a sign of financial irresponsibility leading to bad performing loans

Multivariate Plots Section

Examining the income rates of employed individuals and their rates we can see that a good floor for income level is $3000

This is the median income for defaulted loans in the HR category

It is also the bottom 25% of most almost every other grades defaulted loan



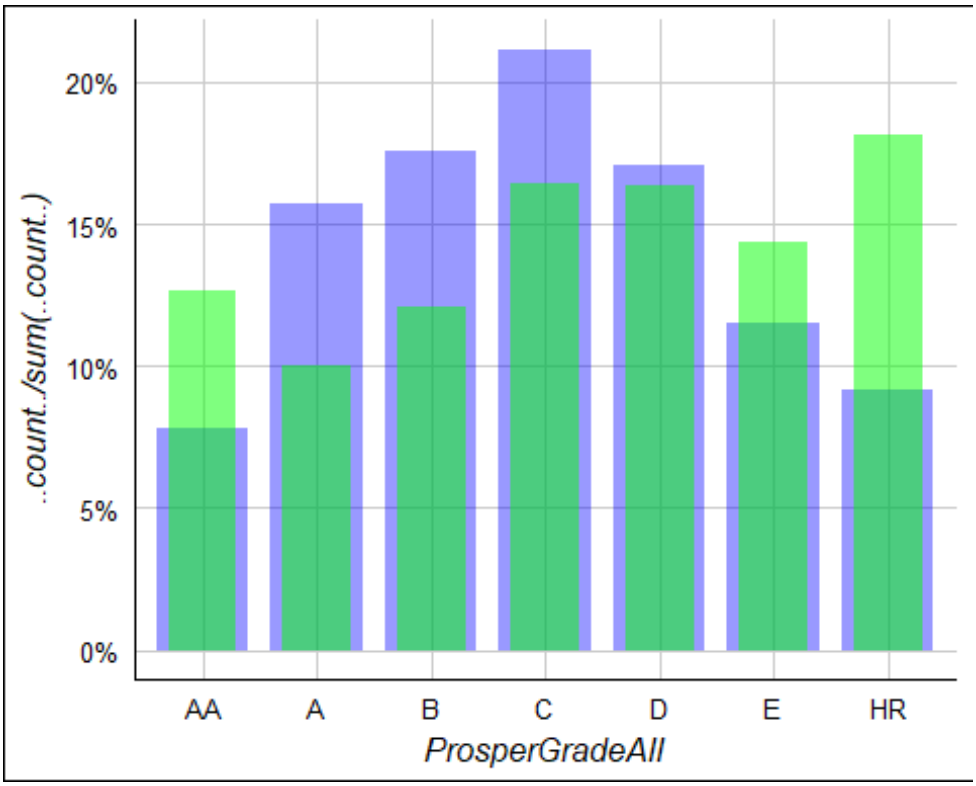

```
## [1] 3000
```

```
## [1] 4166.667
```

It looks like the Prosper grades and credit ratings shifted.

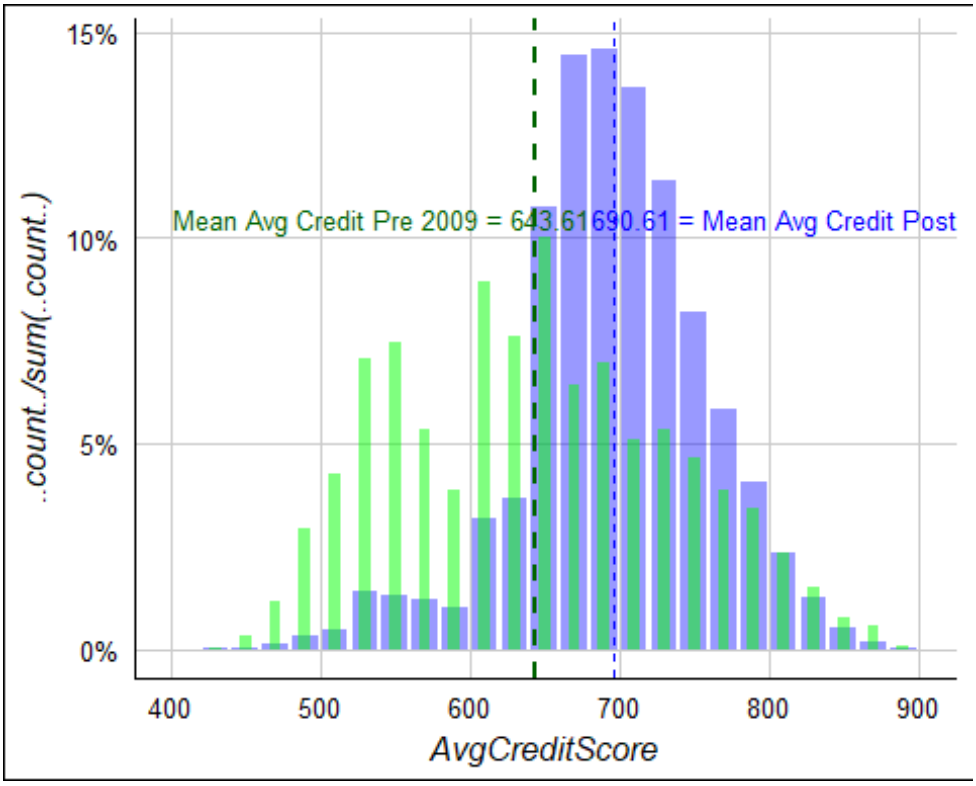
Just because the above graph resembles a normal distribution, doesn't mean that it is a normal distribtion

Credit standards changed after 2009



I am looking at the data before 2009 and after 2009

Prosper went through changes due to the credit crisis & we can see they changed their lending criteria before & after 2009



We can see that it looked like a fairly normal distribution prior to 2009 After 2009 the cut down on loans under 600 credit score Their average credit score also increased from 653.59 to 708.91

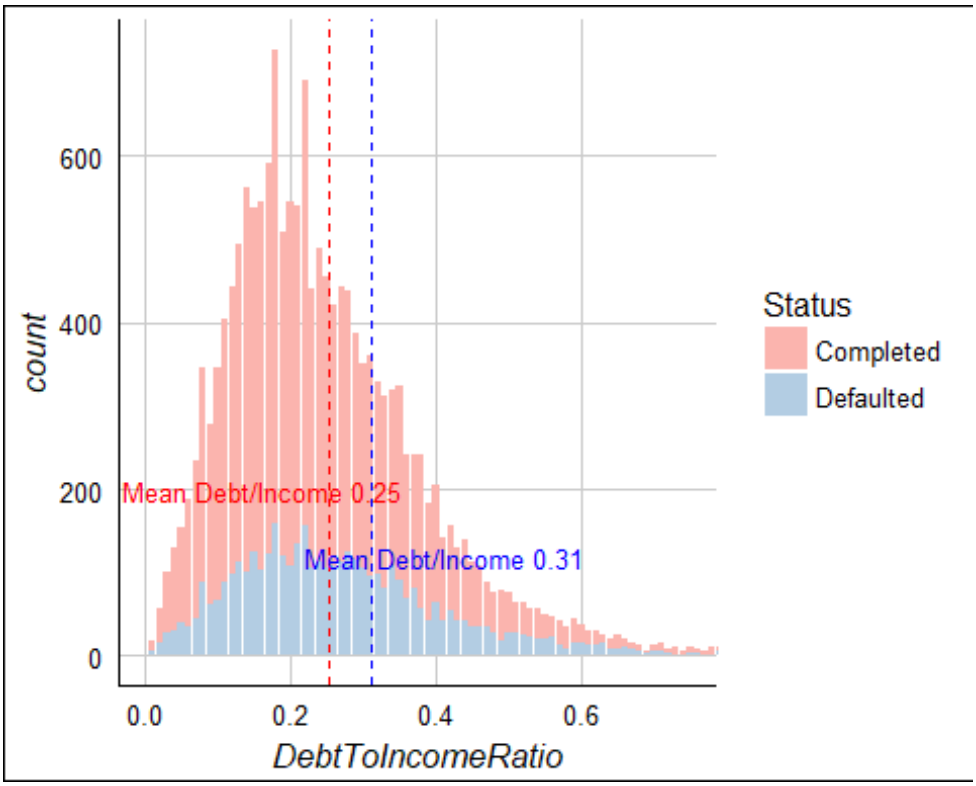
```
## [1] 695.8294
```

```
## [1] 643.104
```

Because debt consolidation is the most popular loan category it is worth exploring further this category more and seeing what trends make successful loans in this category.

We can see the debt to income of Completed vs Defualted loans.

Defaulted loans have 24% higher mean debt to income ratio 25% vs 31%

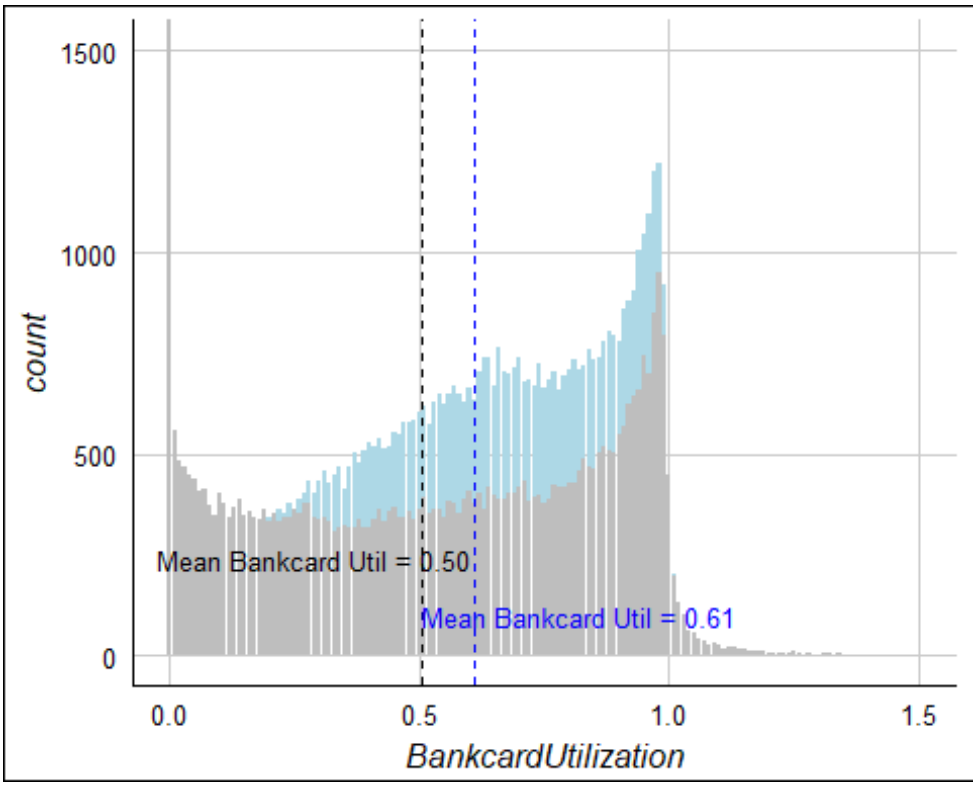



```
## [1] 0.2517811
```

```
## [1] 0.3103099
```

Looking at the bank card utilizatin rate at completed and defaulte loans

We can see that unsuccessful loans had 22% higher bank card utilization 50% vs 61%

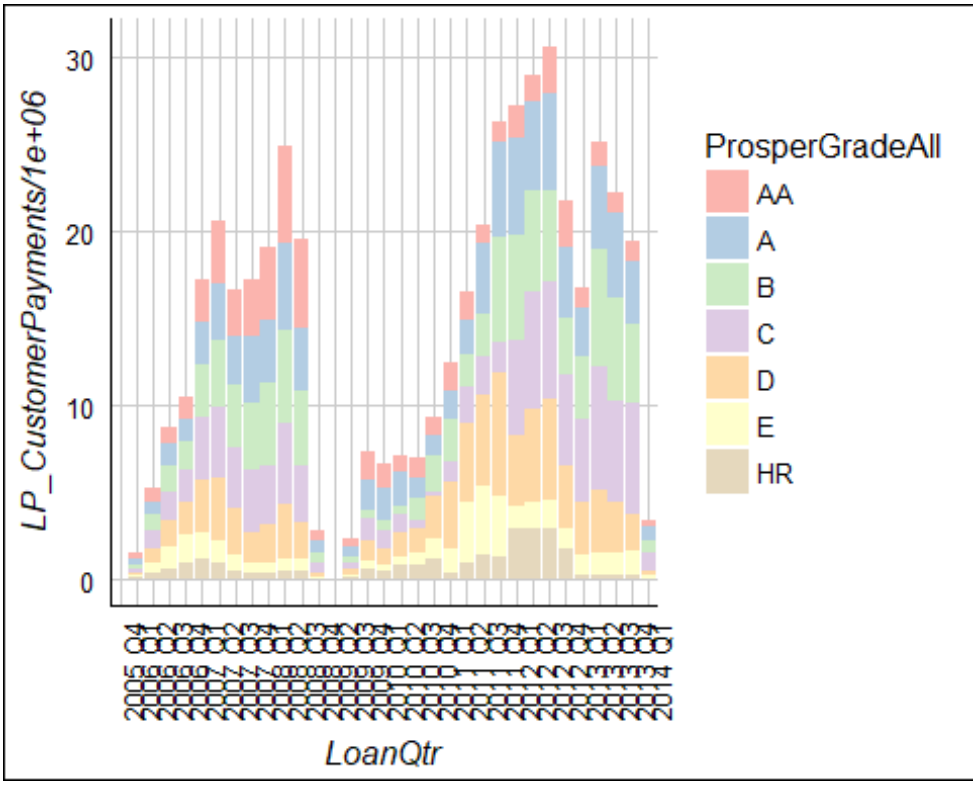



```
## [1] 0.6084715
```

```
## [1] 0.5043438
```

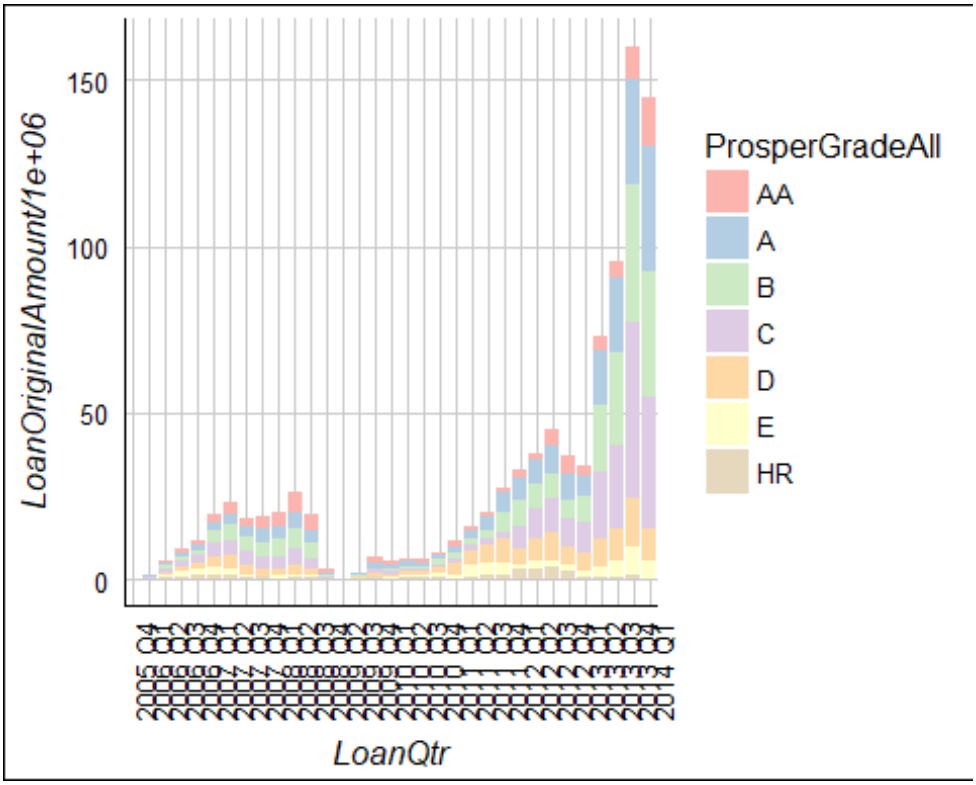
The loan quarter is formatted so that the Q is written first then the year, so we are transforming it so the year is frst then the Q

I visualized the loan payments by quarter by grade so we can see where their cash flows are coming from on a rolling quarter to quarter basis. We can see that C grade loans have grown more popular ad received more funding recently while B and A loans have remained consistent, and the number of AA loan flows dropped after 2009 perhaps because the credit market recovered and good borrowers have refinanced with better funding options.



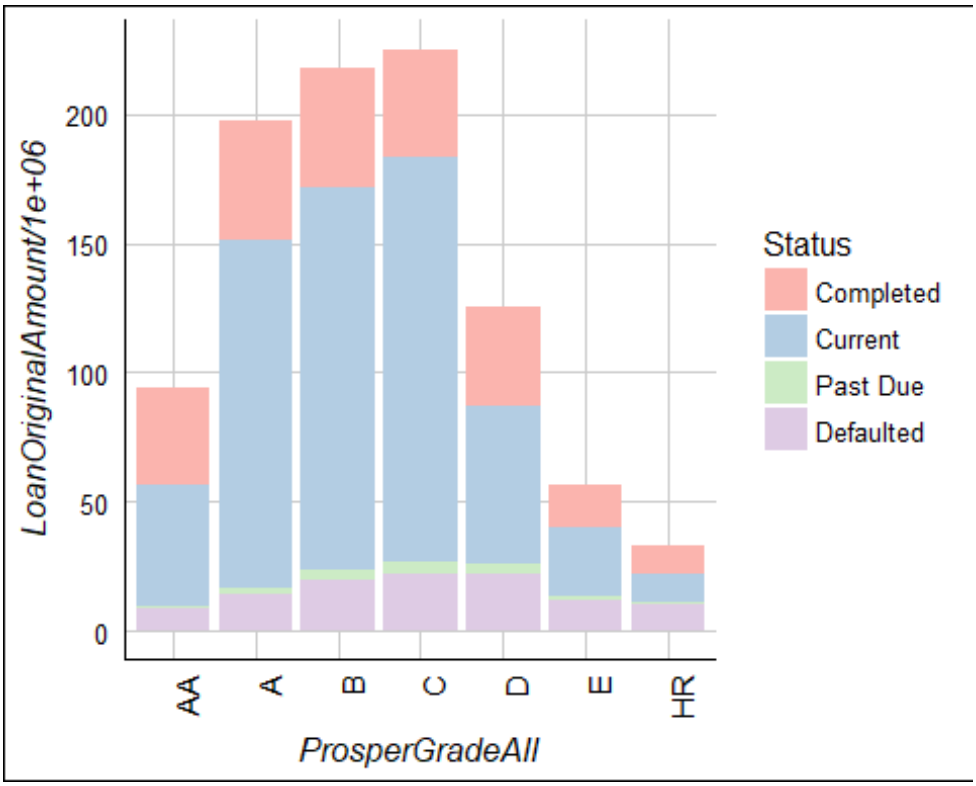
We can see the loan amounts (in millions) based on grade so we can see the quality of the loans originated as well as the amount of the loan.

We see that A, B, & C loans all grew in terms of loans originated near 2013



Now we can look at each loan category and see if they are in good standing

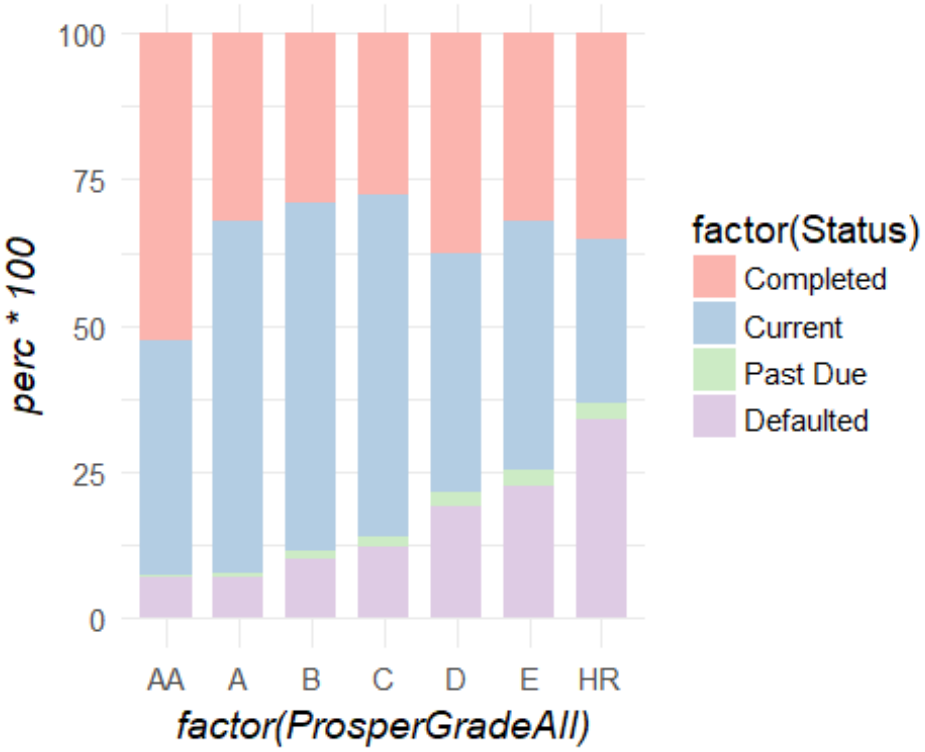
We see the distribution in millions of how many million are in each status



Below we can see what percentage of each status is complete, past due etc.

It looks like the largest incremental percentage jump is between C and D

This combined with our earlier analysis that showed D loans as having a median interest rate similar to the defaulted status tells us C is probably the lowest category of loan that is a good investment.

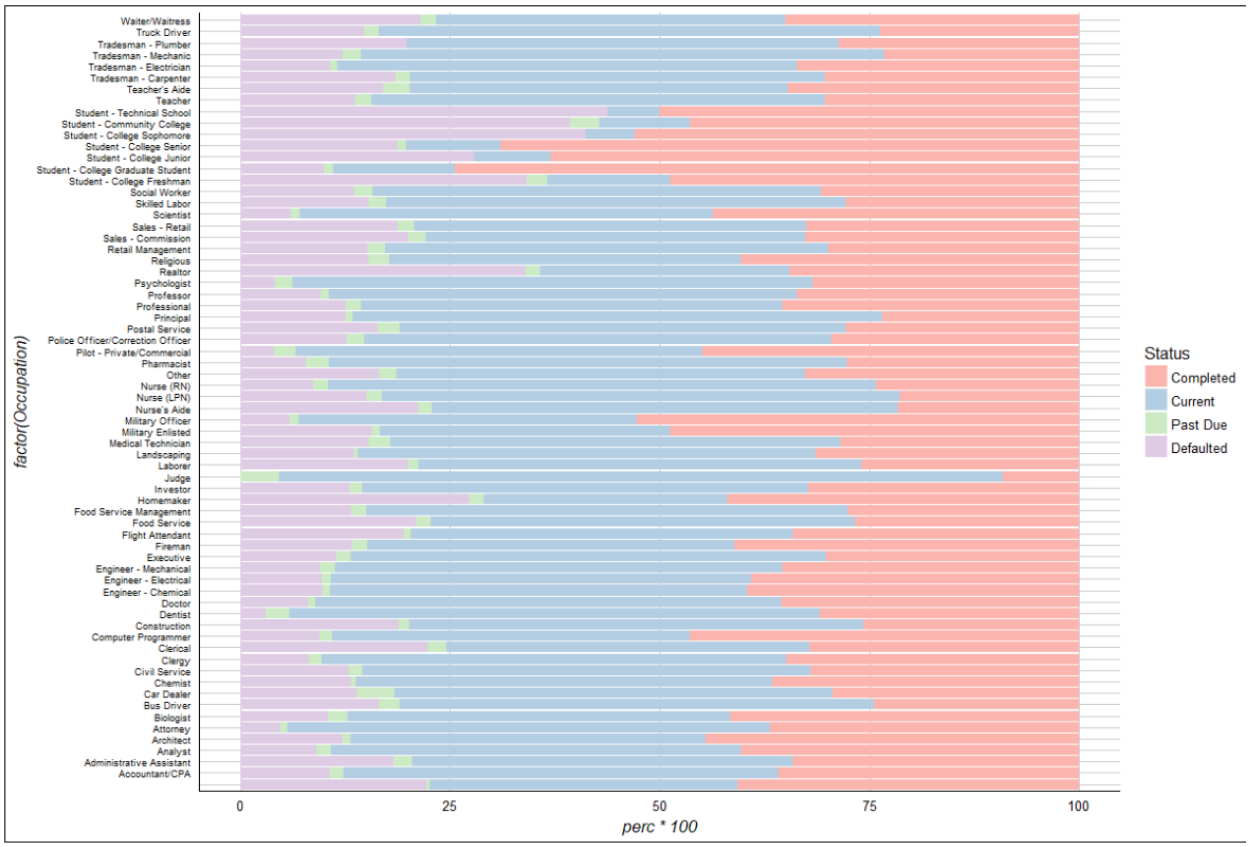


Below we see the different occupations of loan holders and what percentage of those loans are in good or bad standing

Students at technical schools have the worst performing loans of all

Judges, Psycholigst and Dentists have the lowest default ratings

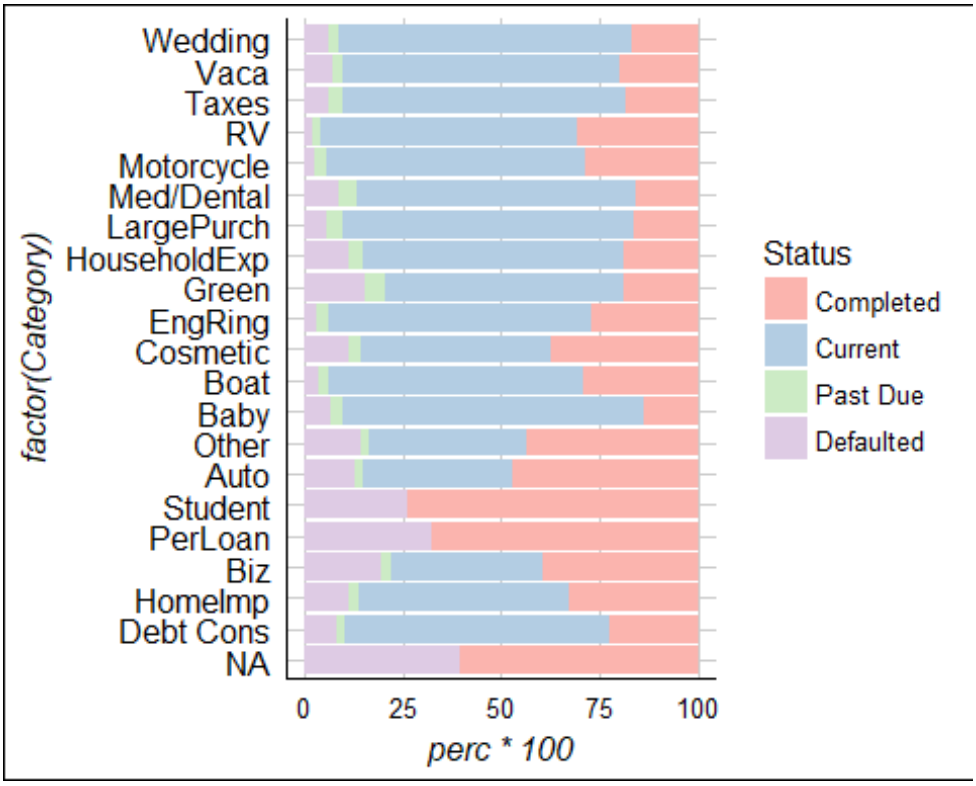
Further analysis could be done with students taking into account school info, graduation rates, grades, & job placement data



Now you can see the loan based on loan reason and the current standing

N/A and personal loans are the worst performing with students also continue to perform badly and have high defaults.

RV loans perform much better than all other categories, perhaps because RV customers are usually retired and older and trying to finance their luxury purchase longer term.

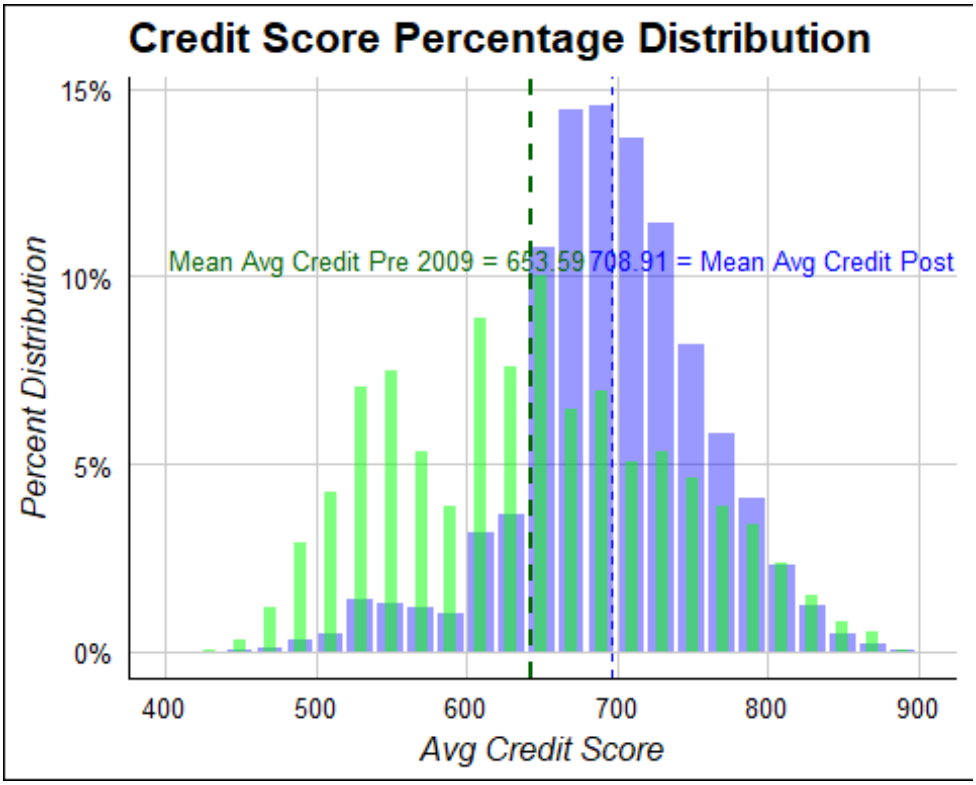


Final Plots and Summary

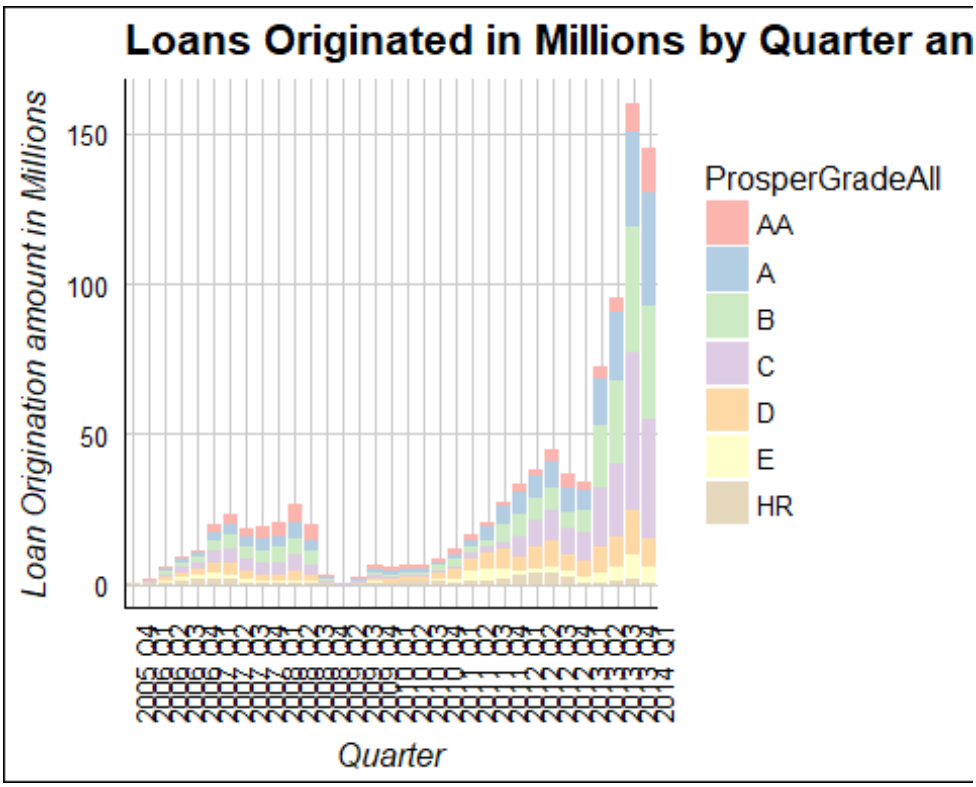
First I think it's important to see Prosper's credit rating distribution

This plot shows both the pre credit crisis and post credit crisis data It's good to see how Prosper adjusted their standards

Prosper's adjustments show us that they initially took on many loans of all different credit categories to gain marketshare but after 2009 they really tightened their lending policy away from bad credit scores



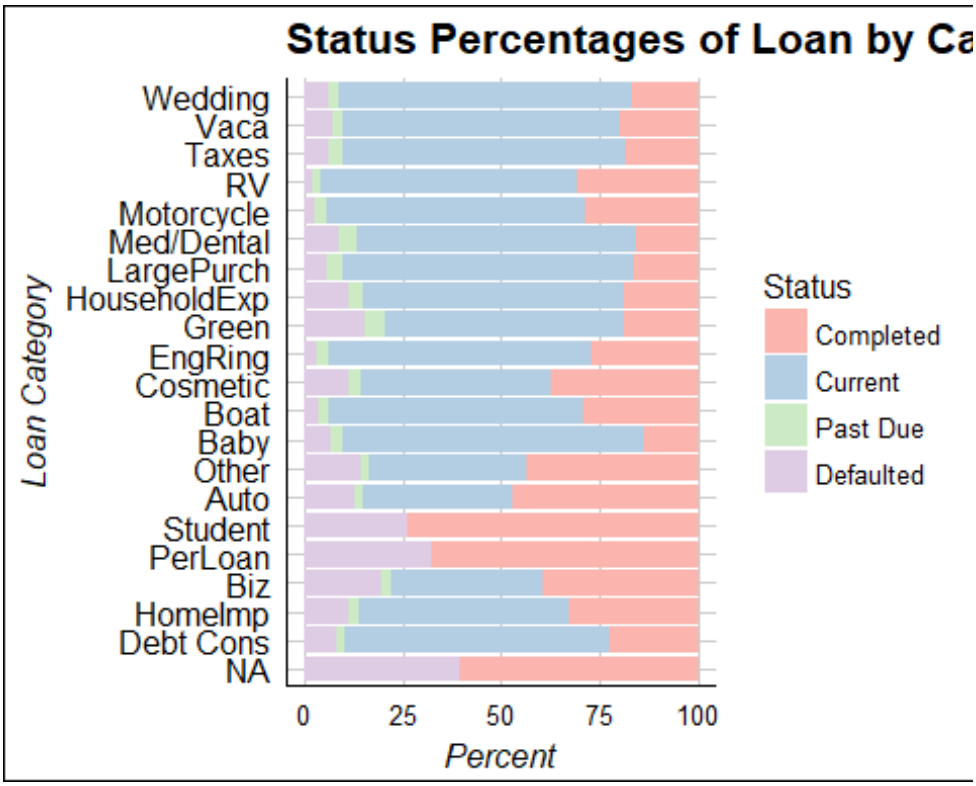
This chart shows how much in loans prosper was generating and by what grade Its good to see that they figured out what kind of loans are most successful Then after 2010 they ramped up their loans and we can see how C grade loans grew the most, this is consistent with our analysis in seeing the C is the lowest grade of loan that does not have a large incremental rise in defaults



Finally here we can see the performance of each loan category Its interesting to see different categories and how their loans perform NA and Personal loans perform the worst perhaps because the reason for the loan is not specified and because borrowers that do not have a specific reason for the loan.

Luxury items like RVs, Motorcycles & Boats perform the best because people may not buy these items unless they could afford luxury

Wedding, engagement rings, etc are all also times that require people to be a lot more responsible so these loans look like they are more successful



Reflection

Overall I learned a lot about the different criteria's that could be used to create a good portfolio of loans that avoids parameters that usually lead to defaults.

I would look for loans that are below the 25% rate and C grade or above Also people who are employed and make $3000 per month and certain categories

For debt consolidation loans I would look at credit card utilization rates and debt/income rates closer to those of complete or current loans

I was surprised to see that loans for luxury goods had such high completions Some of the struggles I had were that defaulted and completed loans had many of the same characteristics, and if I were to go into more detail I would do a more thorough analysis within each type of loan

I think having more unique data points like what types of schools students were applying from and the students stats like their grades and majors would be helpful in finding successful loans

There is no reason to give N/A & personal loans. We need to have more of a description of what the funds are being used for.

To take this analysis further I would backtest the parameters I found for successful loans and come up with a factor analysis to see how the different returns are currently being calculated and look for possible arbitrage opportunities with underpriced loans. I would create a model portfolio and see show the risk level and what the overall percentage return would be

It was great to see that we could get such detailed analysis of the loans I would think that many investors just pick loans based on categories they think are successful. I would not have lent to luxry items like boats because I live with a high level of financial responsibility and know that boats are a bad investment, but the loan data suggests otherwise.

I think the skills I learned here will really shape how I look at data and decisions in the future. Scraping the data with python for any decision and then exploring it with R will lead me to make much better decisions in my personal and professional life and there is no excuse to make an uninformed and non data-driven decision in the future.