Prosper Loan Exploratory Data Analysis by Mustafa Olomi

```
ld <- read.csv('prosperLoanData.csv')
```

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

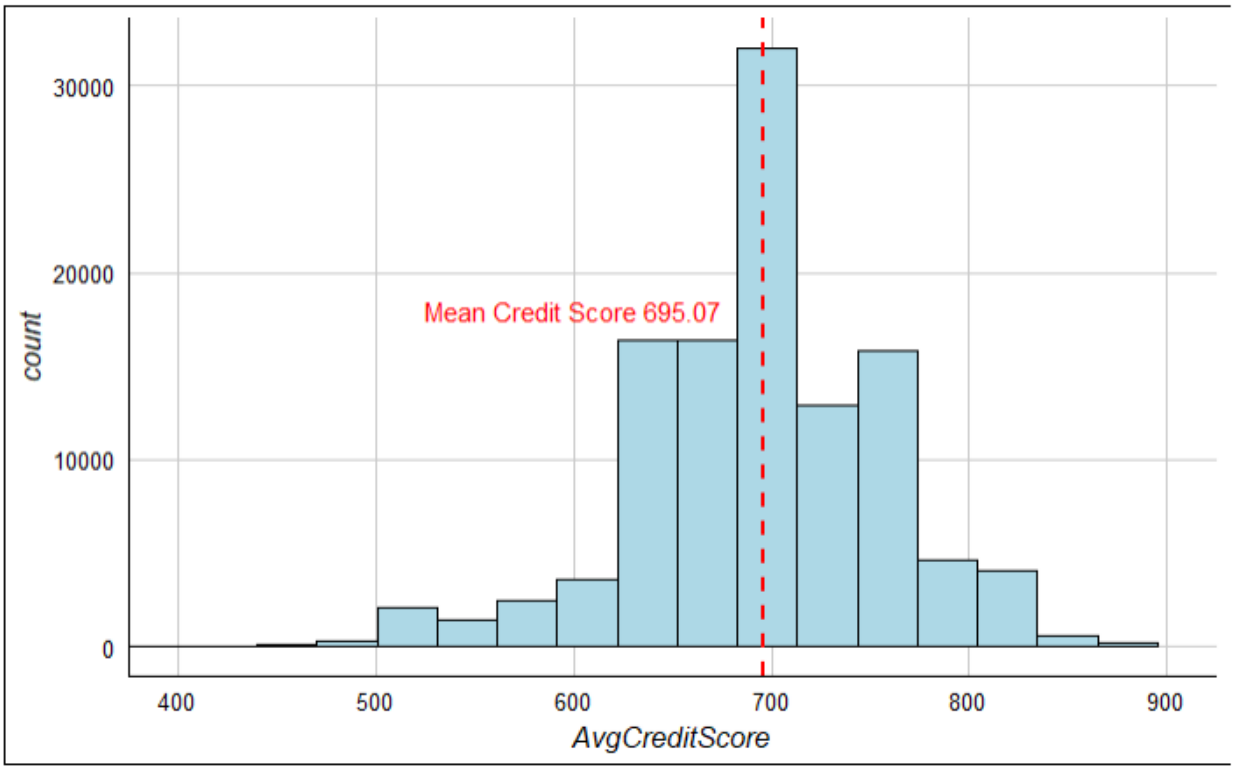
loans and the loans were not initialized so we don't need them in our data

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

```
#Averaging lower and upper credit score
ld$AvgCreditScore <- (ld$CreditScoreRangeUpper+ld$CreditScoreRangeLower) / 2
```

Plotting the distribution of avg credit scores. The middle line is the mean

Most traditional companies would want to know the credit score of the portfolio as it is a more universal measuring yardstick



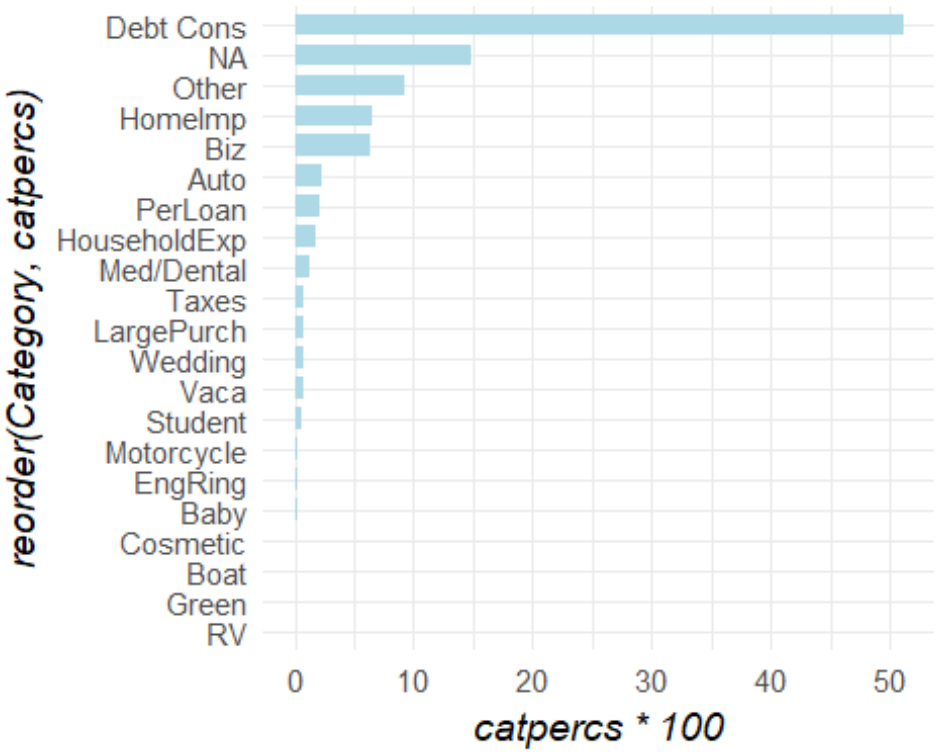
We can see the mean credit score for all loans is 695
This is considered "good" credit according to Experiean
300-579 is Very Poor (17% of people) may not be approved for credit at all
580-669 is Fair (20.2%) considered subprime
679-739 is Good (21.5%) 8% liklihood of serious delinquency
740-799 is Very 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 w lenders
https://www.experian.com/blogs/ask-experian/credit-education/score-basics

Creating labels for each of the categories

```
labs <- c('NA', 'Debt Cons', 'HomeImp', 'Biz', 'PerLoan', 'Student', 'Auto',
          'Other', 'Baby', 'Boat', 'Cosmetic', 'EngRing', 'Green',
          'HouseholdExp', 'LargePurch', 'Med/Dental', 'Motorcycle', 'RV',
          'Taxes', 'Vaca', 'Wedding')

ld$Category <- factor(ld$ListingCategory..numeric.,
                      labels = labs)
```

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

```
ld.cats <- ld %>%
    group_by(Category) %>%
    summarise(count = n()) %>%
    mutate(catpercs =count/sum(count))
    ld.cats %>% arrange(desc(catpercs))

## # 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

ggplot(ld.cats, aes(x = reorder(Category, catpercs), y = catpercs*100)) +
  geom_bar(stat="identity", width = 0.7, fill = 'light blue') +
  theme_minimal(base_size = 14) +
  coord_flip()
```



```
ld$LoanDate <- as.Date(ld$LoanOriginationDate, "%Y-%m-%d")
```

In the documentation it says:

Pre 2009 prosper generated grades with the variable CreditGrade

Post 2009 they used the grade ProsperRating Alpha

Here we are creating a universal grading standard with all grades

```
ld <- ld %>%
mutate(ProsperGradeAll = ifelse(ProsperRating..Alpha. != '',
                                as.character(ProsperRating..Alpha.),
                                as.character(CreditGrade)))

ld <- subset(ld, ProsperGradeAll != 'NC' & ProsperGradeAll != '')

prop.table(table(ld$ProsperGradeAll))

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

ld$ProsperGradeAll <- factor(ld$ProsperGradeAll,
                             levels = c('AA','A','B', 'C', 'D', 'E', 'HR'))
summary(ld$ProsperGradeAll)

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

#Looking at the distributions of the pre 2009 credit grades
prop.table(table(ld$CreditGrade))

##
##                      A         AA          B          C          D
## 0.74655112 0.02915714 0.03087278 0.03861517 0.04969206 0.04533697
##          E         HR         NC
## 0.02893718 0.03083759 0.00000000

#First lets adjust the close date to be just the date
#(since time is not in it anyway)
#First Recorded credit line also goes up to date level
ld$ClosedDate <- as.Date(ld$ClosedDate, "%Y-%m-%d")
ld$FirstRecordedCreditLine <- as.Date(ld$FirstRecordedCreditLine, "%Y-%m-%d")
```

We can see we have many categories, including 6 past due categories

We could simplify this an better define the categories

```
#Lets check what loan statuses we have
summary(ld$LoanStatus)

##               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
```

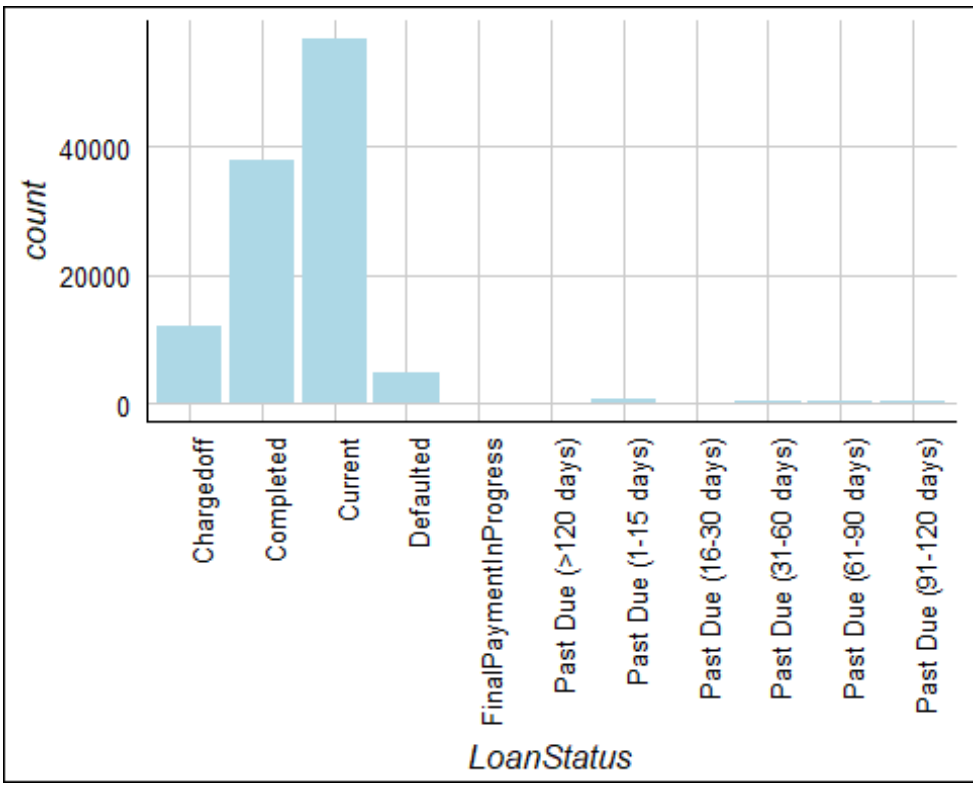
Based on the existing categories we can see that most of the status's are

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.

```
ggplot(ld, aes(LoanStatus)) + geom_histogram(stat='count', fill = 'light blue') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Lets check to see where we should categorize the final payment loans

Looking at their % funded

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

```
summary(subset(ld, LoanStatus=='FinalPaymentInProgress')$PercentFunded)
```

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

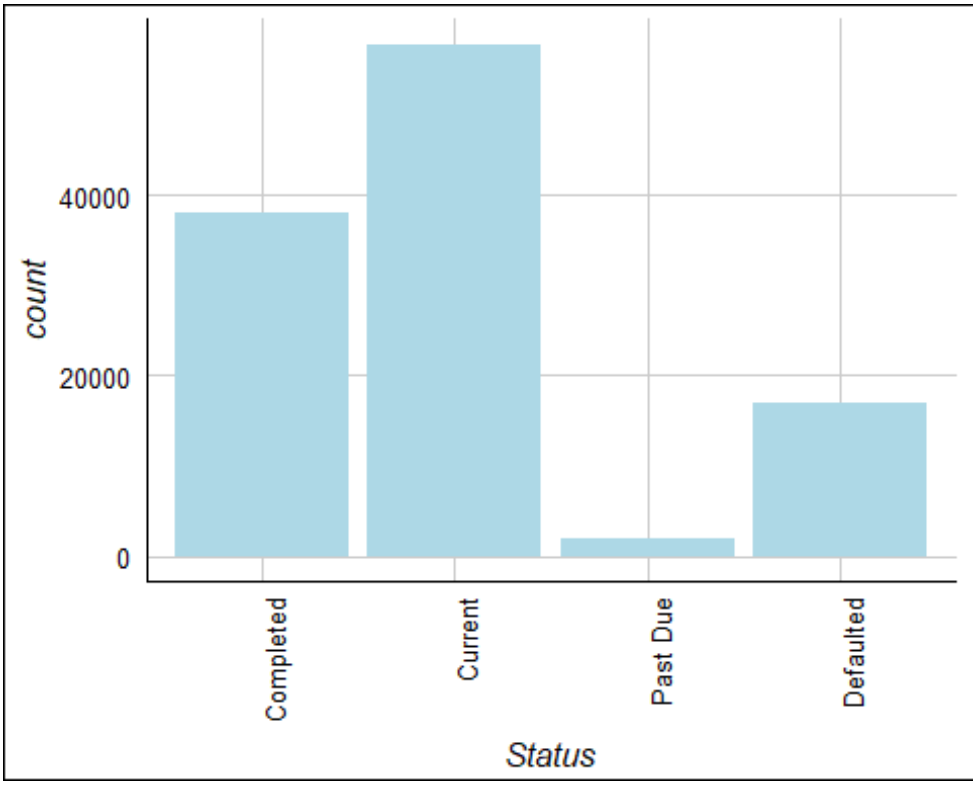
Summarizing All Past Dues as one category

Summarizing Charged Off & Defaulted as one category

```
#When I used the case_when, it does this as a character instead of a vector
#So changing to vector after case_when
ld <- mutate(ld, Status = case_when (LoanStatus == 'Current' ~ 'Current',
                                     LoanStatus == 'Completed' ~ 'Completed',
                                     LoanStatus == 'FinalPaymentInProgress' &
                                       PercentFunded >= 0.95 ~ 'Completed',
                                     LoanStatus == 'FinalPaymentInProgress' &
                                       PercentFunded < 0.95 ~ 'Past Due',
                                     LoanStatus %in% c('Chargedoff', 'Defaulted')
                                     ~ 'Defaulted',
                                     LoanStatus %in% c('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)')
                                     ~ 'Past Due'))

ld$Status <- factor(ld$Status, levels = c('Completed', 'Current',
                                          'Past Due', 'Defaulted'))
summary(ld$Status)

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

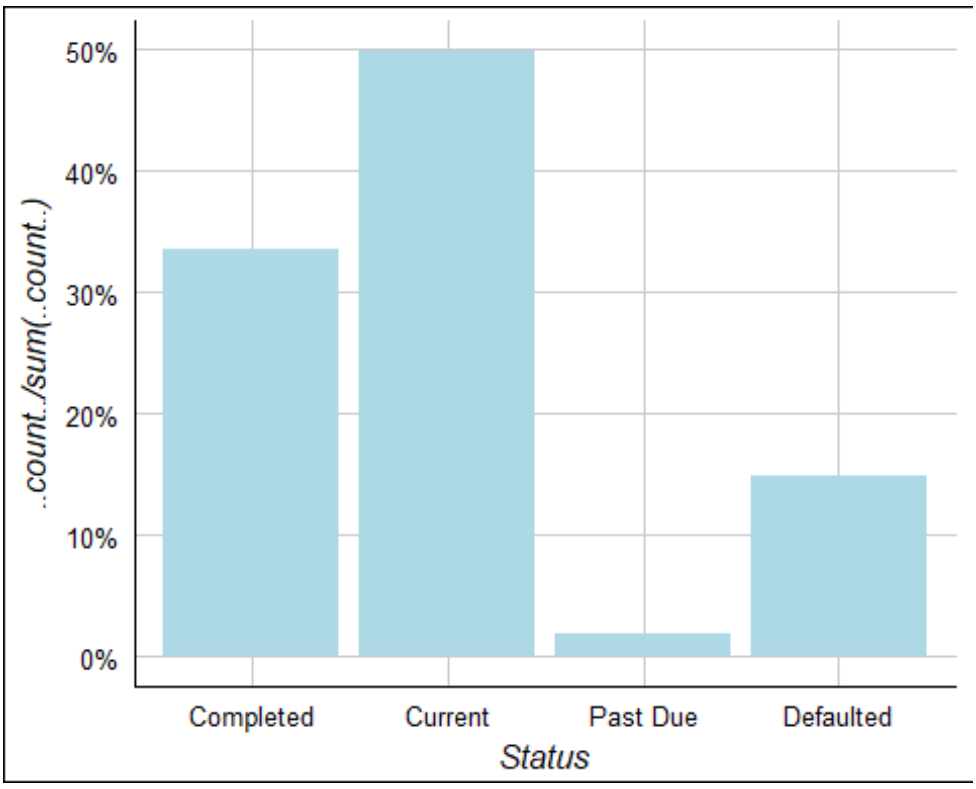
This is the distribution of completed, current, past due and defaulted loans

```
ggplot(ld, aes(Status)) + geom_histogram(stat='count', fill = 'light blue') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



This is the same distribution but is a percentage distribution

```
ggplot(ld, aes(x=Status)) +
  geom_histogram(data = ld, stat = 'count',
                 aes(y = ..count.. / sum(..count..)), fill = 'light blue') +
  scale_y_continuous(labels = scales::percent)
```



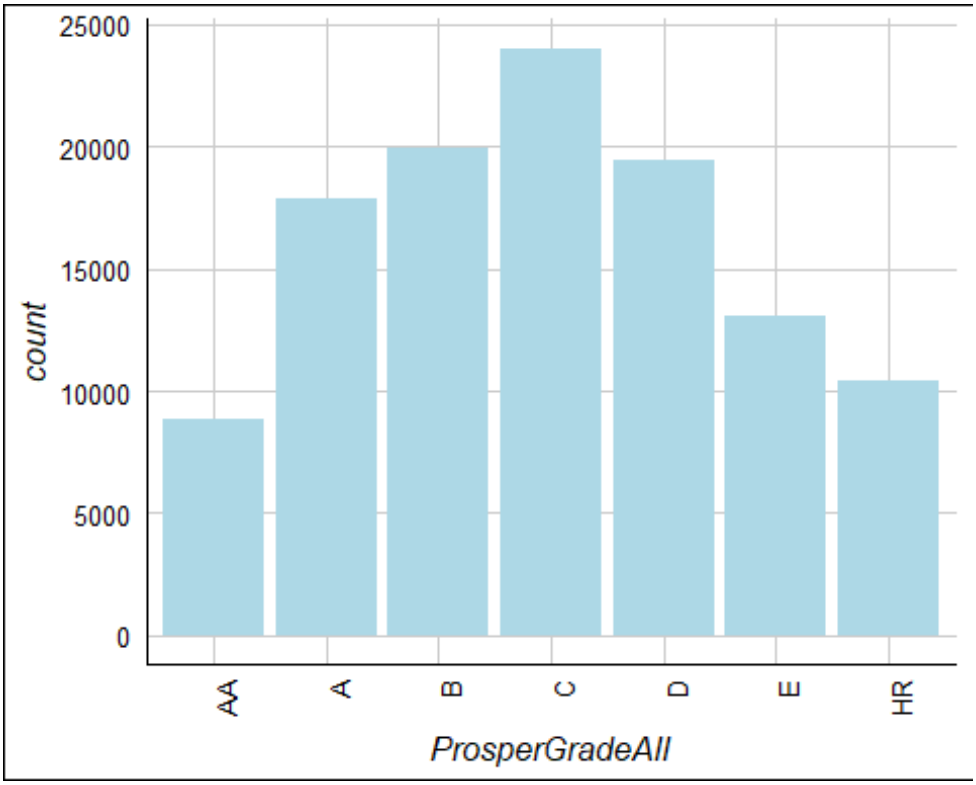

```
summary(ld$ProsperGradeAll)
```

```
##    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

```
ggplot(ld, aes(ProsperGradeAll)) + geom_histogram(stat = 'count',
                                                  fill = 'light blue') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Bivariate Analysis

Now looking at the interest rate among the 4 main categories we set

The top 75 percentile of completed loans were 25.11%

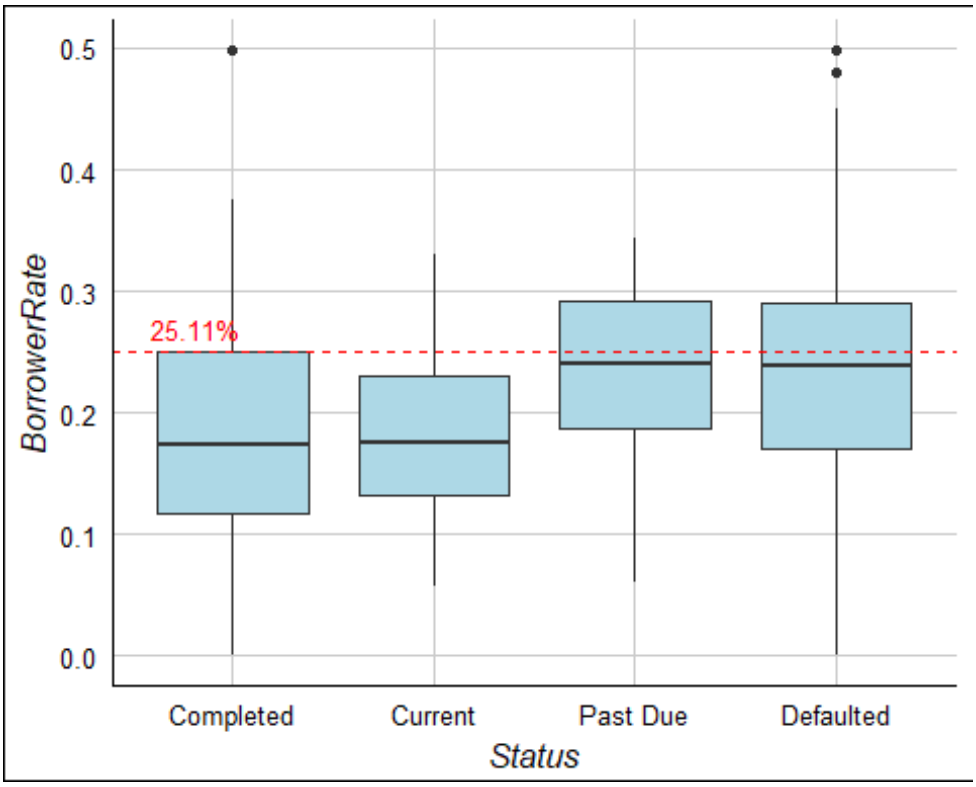
It's also interesting to see that that is near the median of bad loans

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.

```
ggplot(ld, aes(Status, BorrowerRate)) +geom_boxplot(fill = 'light blue') +
  geom_hline(yintercept = quantile(subset(ld, Status=='Completed')$BorrowerRate,
                                   probs = 0.75, na.rm = TRUE), color = 'red',
             linetype='dashed') +
    annotate("text", label = " 25.11%",
           x = 0.8, y = 0.27, color = "red", size=3.5)
```




```
quantile(subset(ld, Status=='Completed')$BorrowerRate,
                                   probs = 0.75, na.rm = TRUE)

##    75%
## 0.2511
```

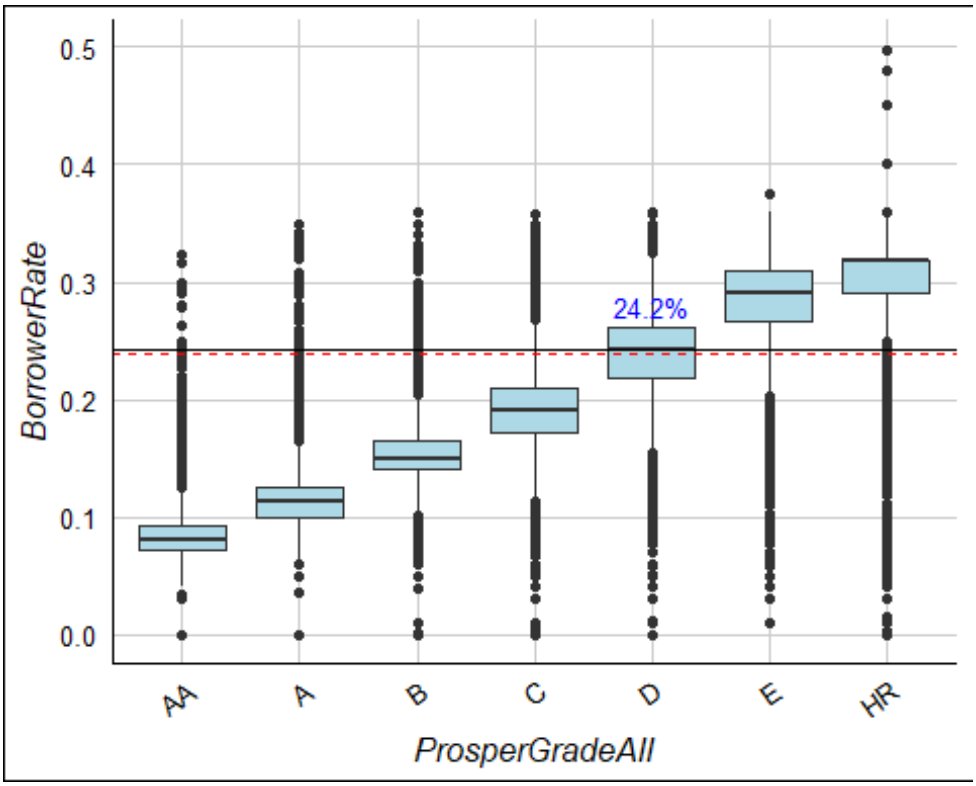
This shows the distribution of prosper grades and interest rates

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

```
ggplot(ld, aes(ProsperGradeAll, BorrowerRate)) +
  geom_boxplot(fill = 'light blue') +
  theme(text = element_text(size = 12),
        axis.text.x = element_text(angle = 35, hjust = 1)) +
  #coord_cartesian(ylim = c(0,10000)) +
  geom_hline(yintercept=median(subset(ld, ProsperGradeAll=='D')$BorrowerRate)) +
    #geom_hline(yintercept=median(ld$StatedMonthlyIncome)) +
  scale_fill_brewer(palette="Pastel1") +
  annotate("text", label = "24.2%",
           x = 'D', y = 0.28, color = "blue", size=3.5) +
    geom_hline(yintercept = median(subset(ld, Status=='Defaulted')$BorrowerRate)
               , color = 'red',
             linetype='dashed')
```




```
median(subset(ld, ProsperGradeAll=='D')$BorrowerRate)
```

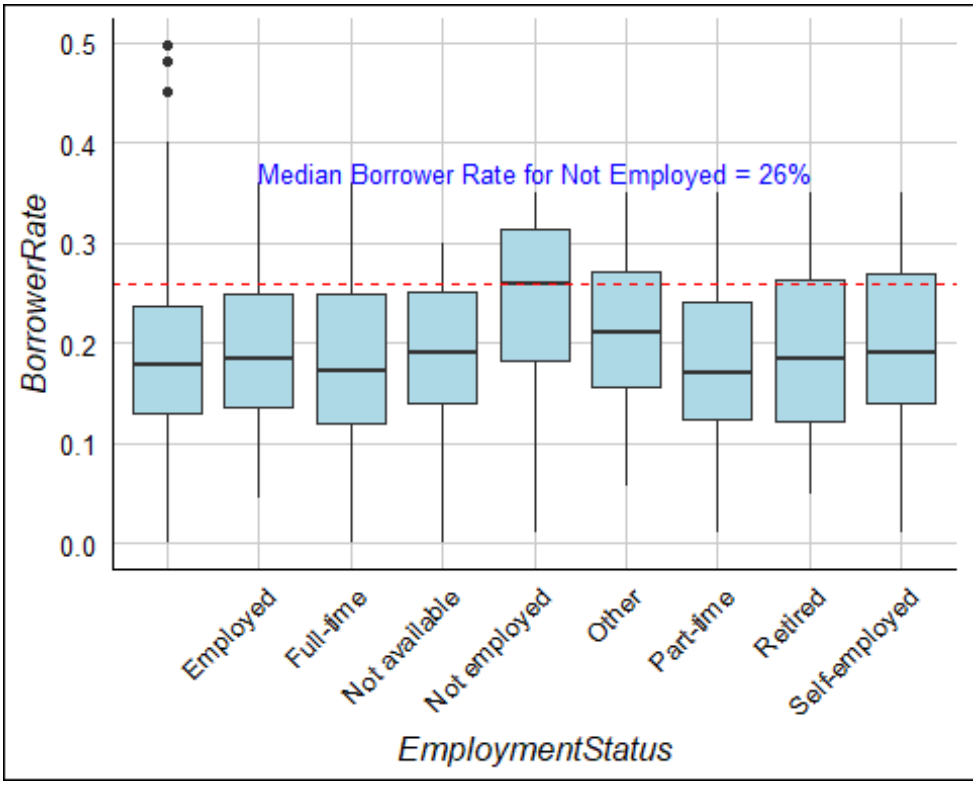
```
## [1] 0.2419
```

```
median(subset(ld, Status=='Defaulted')$BorrowerRate)
```

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

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

```
ggplot(ld, aes(EmploymentStatus, BorrowerRate),) +
  geom_boxplot(fill = 'light blue') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  geom_hline(aes(yintercept=median(subset(ld, EmploymentStatus=='Not employed')
                                   $BorrowerRate)),
             color='red', linetype='dashed') +
  annotate("text", label = "Median Borrower Rate for Not Employed = 26%",
           x = 'Not employed', y = 0.37, color = "blue", size=3.5)
```




```
median(subset(ld, EmploymentStatus=='Not employed')$BorrowerRate)
```

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

```
median(subset(ld,  Status=='Defaulted')$BorrowerRate)
```

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

The most interesting relationship I saw was the fact that bad loans tended
to be around the 25% interest rate level, thats the median for defaulted
loans, and around the median for D grade loans and for unemployed borrowers
Whatever the case, a borrower that is willing to pay close to 25% is usually
a sign of financial irresponsibility that leads to bad performing loans

Multivariate Plots Section

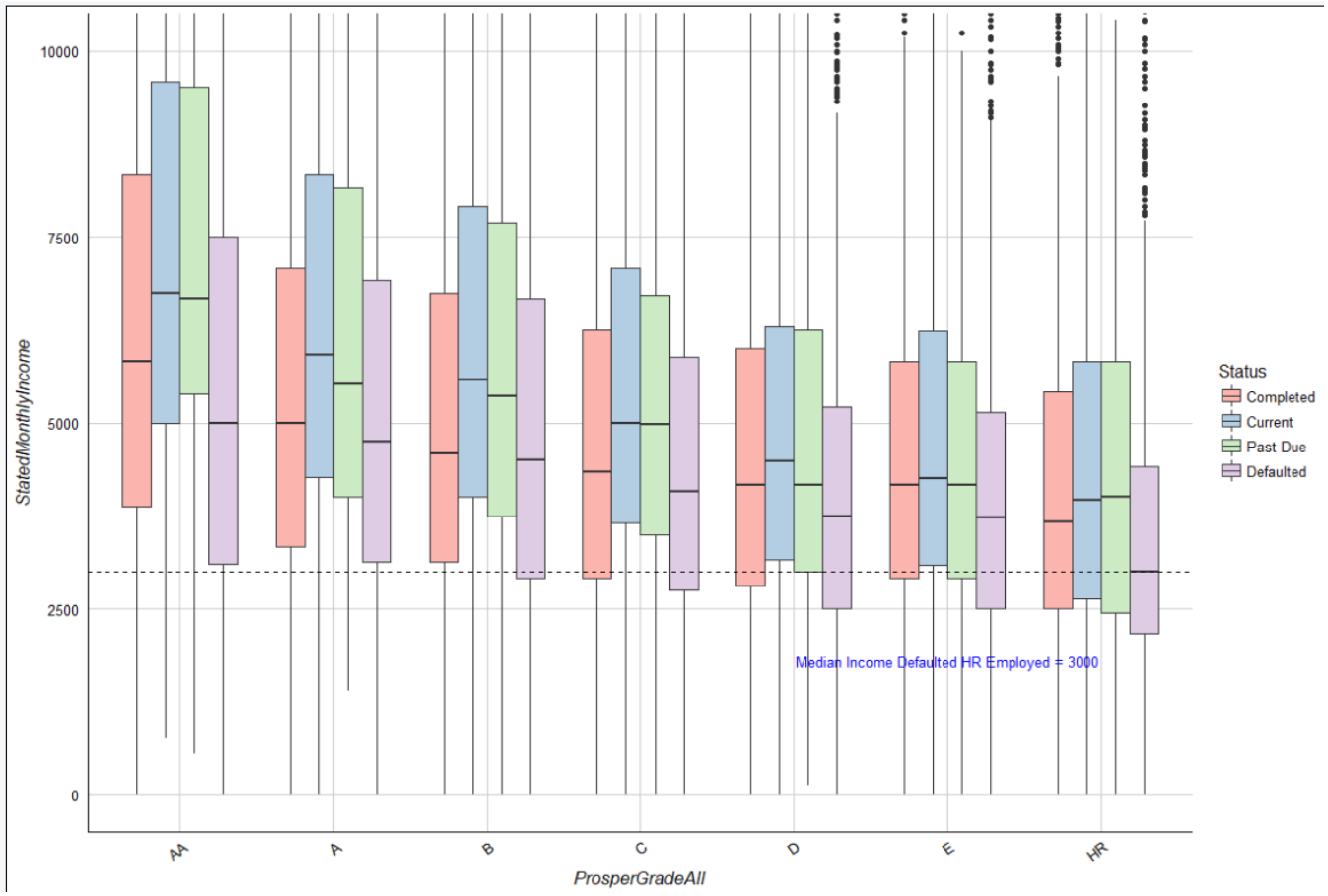
When examining the income rates of employed individuals and their rates

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

```
ggplot(subset(ld, EmploymentStatus != 'Not employed'),
       aes(ProsperGradeAll, StatedMonthlyIncome, fill=Status)) +
  geom_boxplot() +
  theme(text = element_text(size = 12),
        axis.text.x = element_text(angle = 35, hjust = 1)) +
  coord_cartesian(ylim = c(0,10000)) +
  geom_hline(yintercept=median(subset(ld, Status == 'Defaulted' &
                                      ProsperGradeAll == 'HR' &
                                      EmploymentStatus != 'Not employed')
                               $StatedMonthlyIncome), linetype = 'dashed') +
  scale_fill_brewer(palette="Pastel1")  +
  annotate("text", label = "Median Income Defaulted HR Employed = 3000",
           x = 'E', y = 1800, color = "blue", size=3.5)
```




```
median(subset(ld, Status == 'Defaulted' & ProsperGradeAll == 'HR' &
                                      EmploymentStatus != 'Not employed')
       $StatedMonthlyIncome)

## [1] 3000

median(subset(ld, ProsperGradeAll=='C' & IncomeVerifiable=='False')
       $StatedMonthlyIncome)

## [1] 4166.667
```

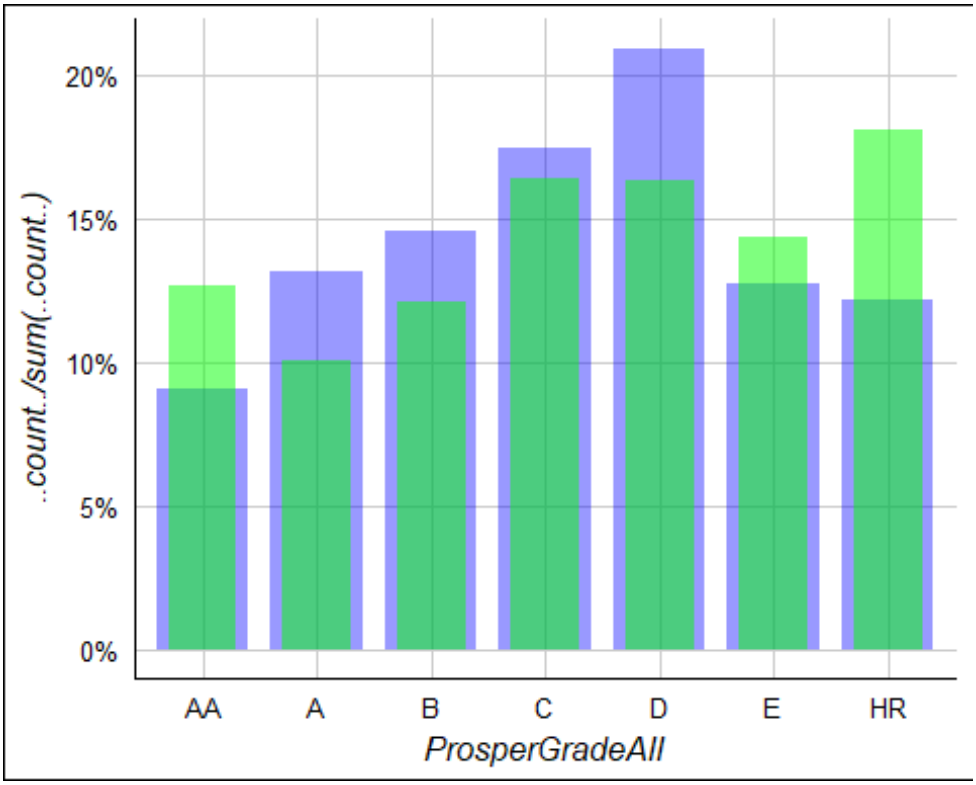
It looks like the Prosper grades also shifted like the way credit scores did

Just because the above graph resembles a normal distribution, doesn't mean

that it is a normal distribtion, credit standards changed after 2009

We will explore this in the bivariate plot section also

```
ggplot(ld, aes(x=ProsperGradeAll)) +
  geom_histogram(data = subset(ld, ClosedDate>= '2009-01-01'), stat = 'count',
                 aes(y = ..count.. / sum(..count..)), fill = 'blue',
                 alpha = 0.4, width = 0.8) +
  geom_histogram(data = subset(ld, ClosedDate< '2009-01-01'), stat = 'count',
                 aes(y = ..count.. / sum(..count..)), fill = 'green',
                 alpha = 0.5, width = 0.6)+
  scale_y_continuous(labels = scales::percent)
```

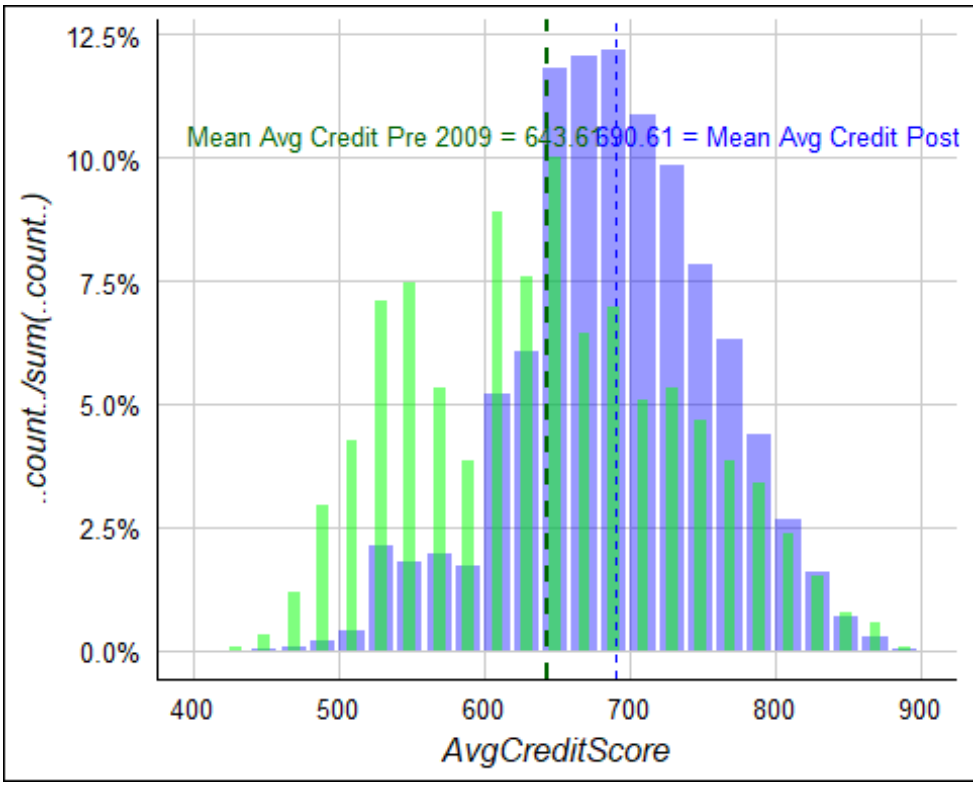


I am looking at the data before 2009 and after 2009

Prosper went through some changes because of the credit crisis and we can

see how they changed their lending criteria before and after 2009

```
ggplot(ld, aes(x=AvgCreditScore)) +
  geom_histogram(data = subset(ld, ClosedDate>= '2009-01-01'), stat = 'count',
                 aes(y = ..count.. / sum(..count..)), fill = 'blue',
                 alpha = 0.4, width = 17) +
  geom_histogram(data = subset(ld, ClosedDate< '2009-01-01'), stat = 'count',
                 aes(y = ..count.. / sum(..count..)), fill = 'green',
                 alpha = 0.5, width = 8)+
  scale_y_continuous(labels = scales::percent) +
    coord_cartesian(xlim = c(400, 900)) +
  geom_vline(data = subset(ld, ClosedDate>= '2009-01-01'),
             aes(xintercept = mean(AvgCreditScore, na.rm = T)),
             color = 'blue', linetype='dashed', size=0.5) +
  geom_vline(data = subset(ld, ClosedDate < '2009-01-01'),
             aes(xintercept = mean(AvgCreditScore, na.rm = T)),
             color = 'dark green', linetype='dashed', size=1) +
  annotate("text", label = "Mean Avg Credit Pre 2009 = 643.61",
           x = 540, y = 0.105, color = "darkgreen", size=3.5) +
  annotate("text", label = "690.61 = Mean Avg Credit Post 2009",
           x = 825, y = 0.105, color = "blue", size=3.5)
```



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

```
mean(subset(ld, ClosedDate >= '2009-01-01')$AvgCreditScore, na.rm = T)
```

```
## [1] 690.6151
```

```
mean(subset(ld, ClosedDate < '2009-01-01')$AvgCreditScore, na.rm = T)
```

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

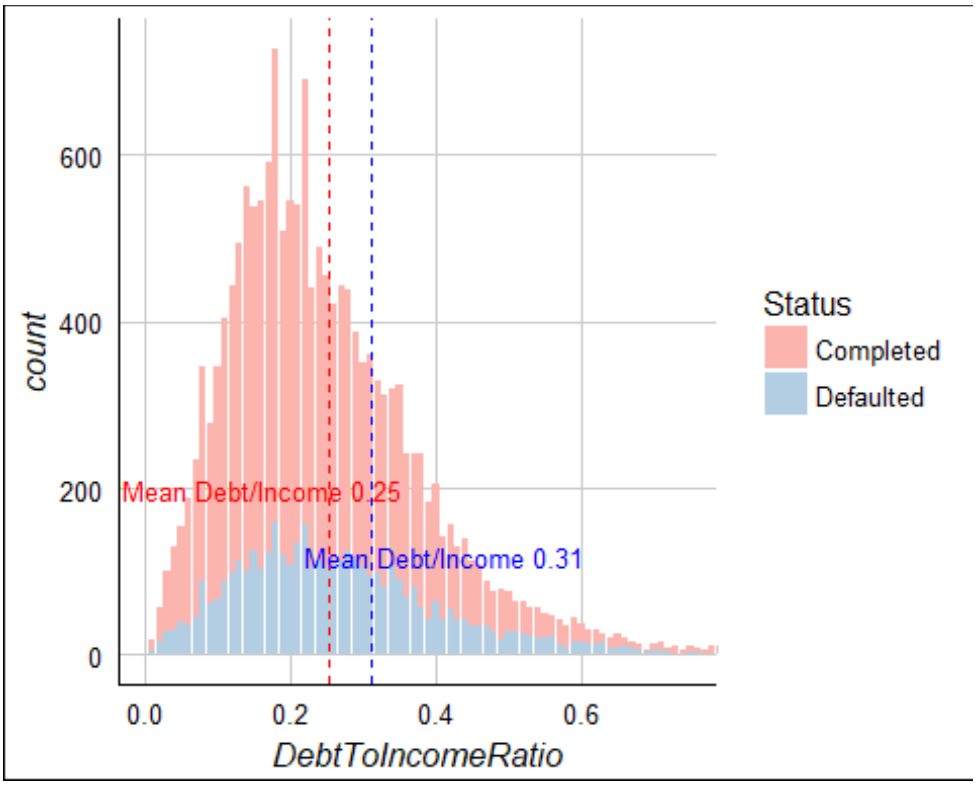
Because debt consolidation is the most popular loan category it is worth

exploring this category more and seeing what trends make successful loans

Below 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%

```
ggplot(subset(ld, Category == 'Debt Cons' & Status %in%
                c('Completed', 'Defaulted')),
       aes(DebtToIncomeRatio, fill = Status)) +
  geom_bar(stat = 'count') +
  coord_cartesian(xlim = c(0,0.75)) +
  scale_fill_brewer(palette="Pastel1") +
  geom_vline(aes(xintercept=mean(subset(ld, Category == 'Debt Cons'
                                        & Status == 'Completed')$
                                   DebtToIncomeRatio, na.rm = TRUE)),
             color='red', linetype='dashed') +
  annotate("text", label = "Mean Debt/Income 0.25", x = 0.16, y = 200,
           color = "red", size=3.5) +
  geom_vline(aes(xintercept=mean(subset(ld, Category == 'Debt Cons'
                                        & Status == 'Defaulted')$
                                   DebtToIncomeRatio, na.rm = TRUE)),
             color='blue', linetype='dashed') +
  annotate("text", label = "Mean Debt/Income 0.31", x = 0.41, y = 120,
           color = "blue", size=3.5)
```




```
mean(subset(ld, Category == 'Debt Cons'
                                        & Status == 'Completed')
     $DebtToIncomeRatio, na.rm = TRUE)
```

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

```
mean(subset(ld, Category == 'Debt Cons'
                                        & Status == 'Defaulted')
     $DebtToIncomeRatio, na.rm = TRUE)
```

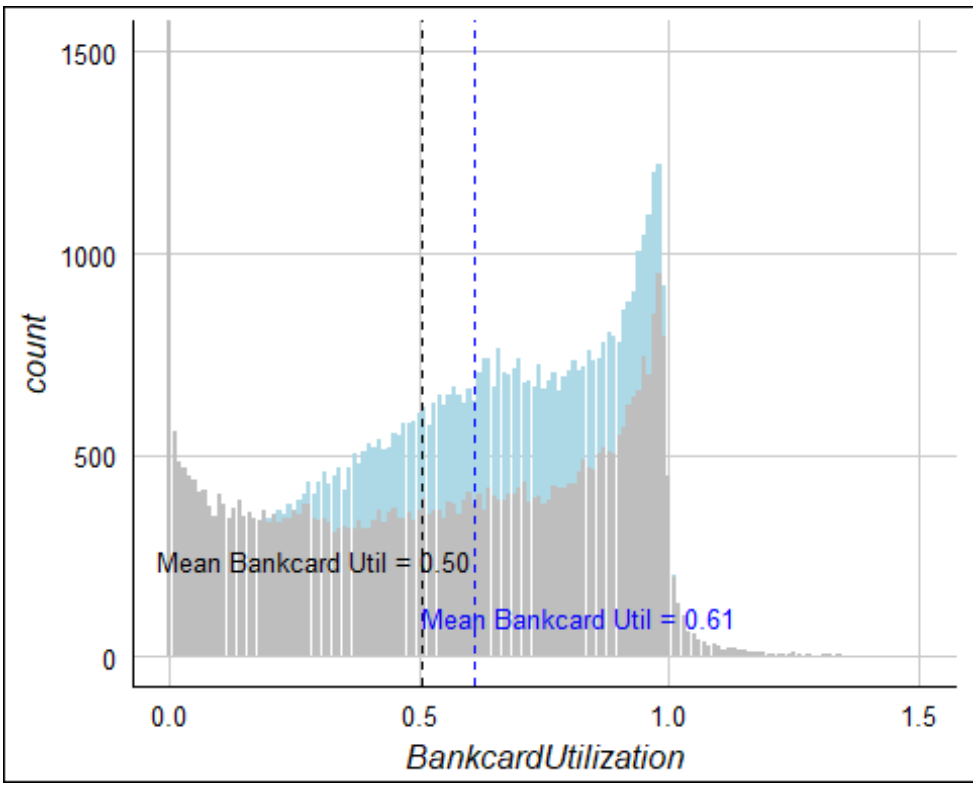
```
## [1] 0.3103099
```

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%

```
ggplot() +
  geom_bar(data = subset(ld, Category == 'Debt Cons'), aes(BankcardUtilization),
           stat = 'count', fill = 'light blue') +
 geom_bar(data = subset(ld, Category != 'Debt Cons'), aes(BankcardUtilization),
          stat = 'count', fill = 'gray') +
      geom_vline(aes(xintercept=mean(subset(ld, Category == 'Debt Cons')$
                                     BankcardUtilization, na.rm = TRUE)),
             color='blue', linetype='dashed') +
      geom_vline(aes(xintercept=mean(subset(ld, Category != 'Debt Cons')$
                                     BankcardUtilization, na.rm = TRUE)),
                 color='black', linetype='dashed') +
  coord_cartesian(xlim = c(0,1.5), ylim = c(0,1500)) +
    annotate("text", label = "Mean Bankcard Util = 0.50", x = 0.29, y = 240,
           color = "black", size=3.5) +
    annotate("text", label = "Mean Bankcard Util = 0.61", x = 0.82, y = 100,
           color = "blue", size=3.5) +
  scale_fill_brewer(palette="Pastel1")
```




```
mean(subset(ld, Category == 'Debt Cons')$BankcardUtilization, na.rm = TRUE)
```

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

```
mean(subset(ld, Category != 'Debt Cons')$BankcardUtilization, na.rm = TRUE)
```

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

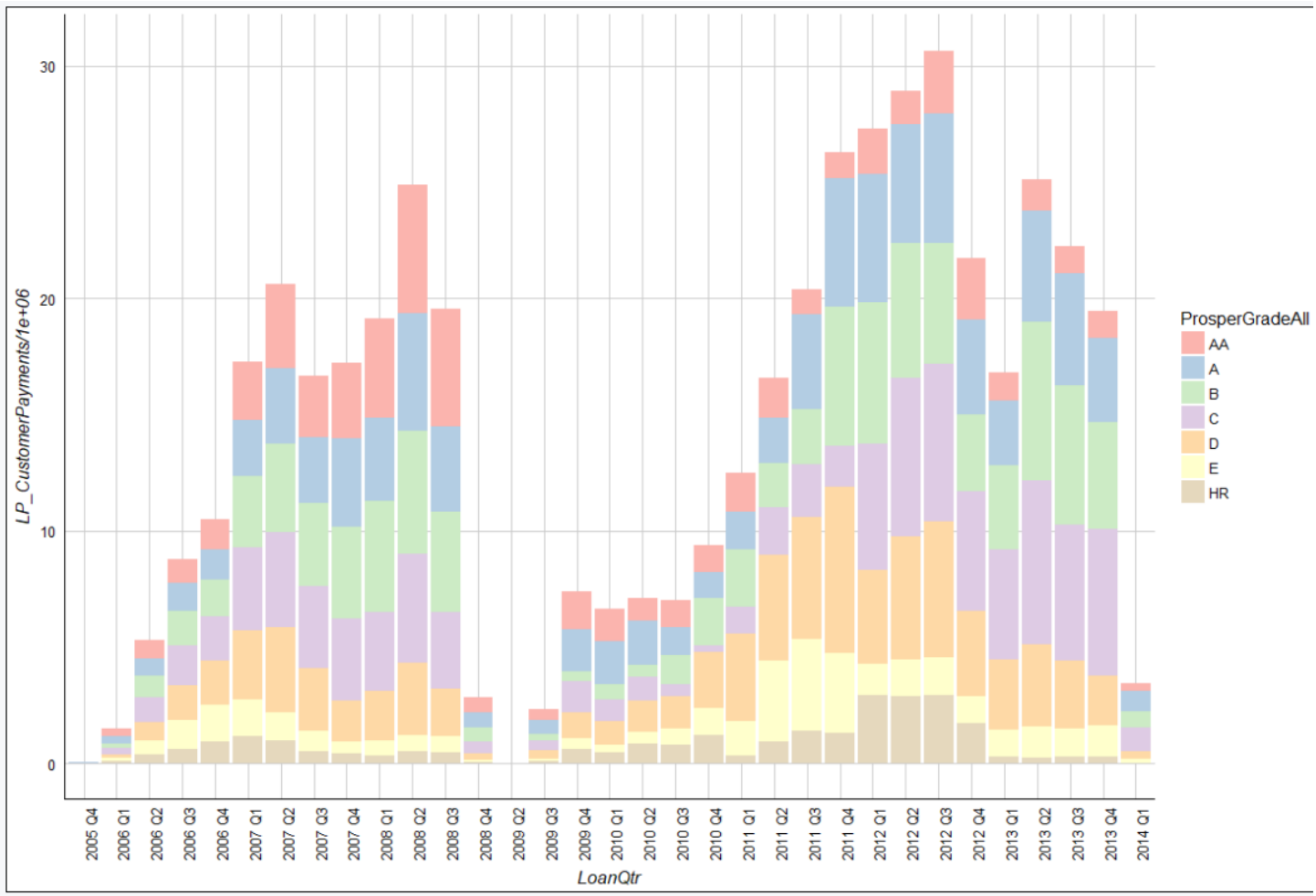
The loan quarter is formatted so that the Q is written first then the year

Thus below we are transforming it so the year is frst then the Q

```
ld <- mutate(ld, LoanQtr = paste(substring(LoanOriginationQuarter, 4,8),
                                 substring(LoanOriginationQuarter, 1,2)))
```

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 droppedd after 2009 perhaps because the credit market recovered and good borrowers have refinanced or found other funding options.

```
ggplot(ld, aes(LoanQtr, LP_CustomerPayments/1000000, fill = ProsperGradeAll)) +
  geom_bar(stat = 'identity') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, size = 10)) +
  scale_fill_brewer(palette="Pastel1")
```

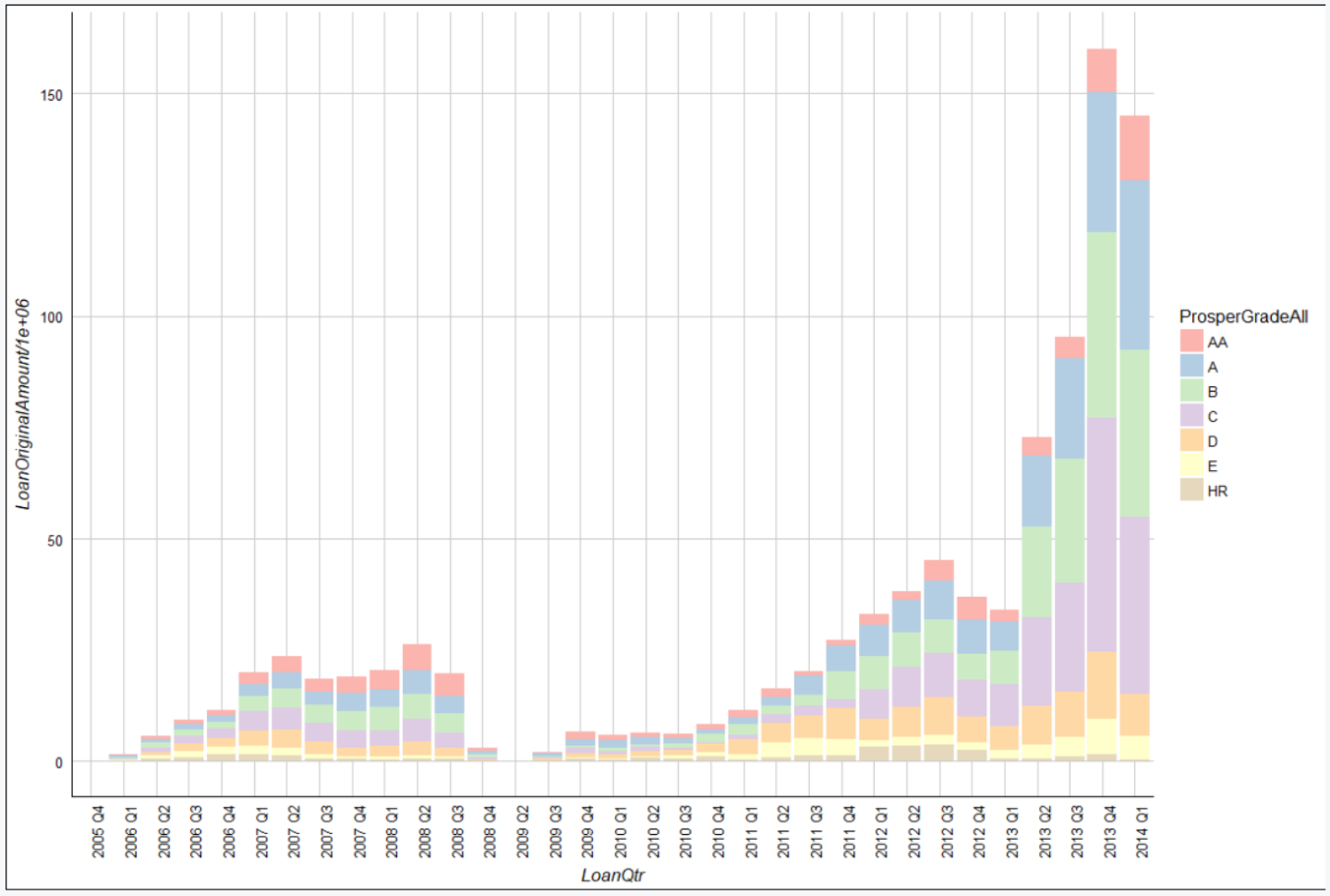


Below we can see the loan amounts (in millions) based on grade so we can see

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

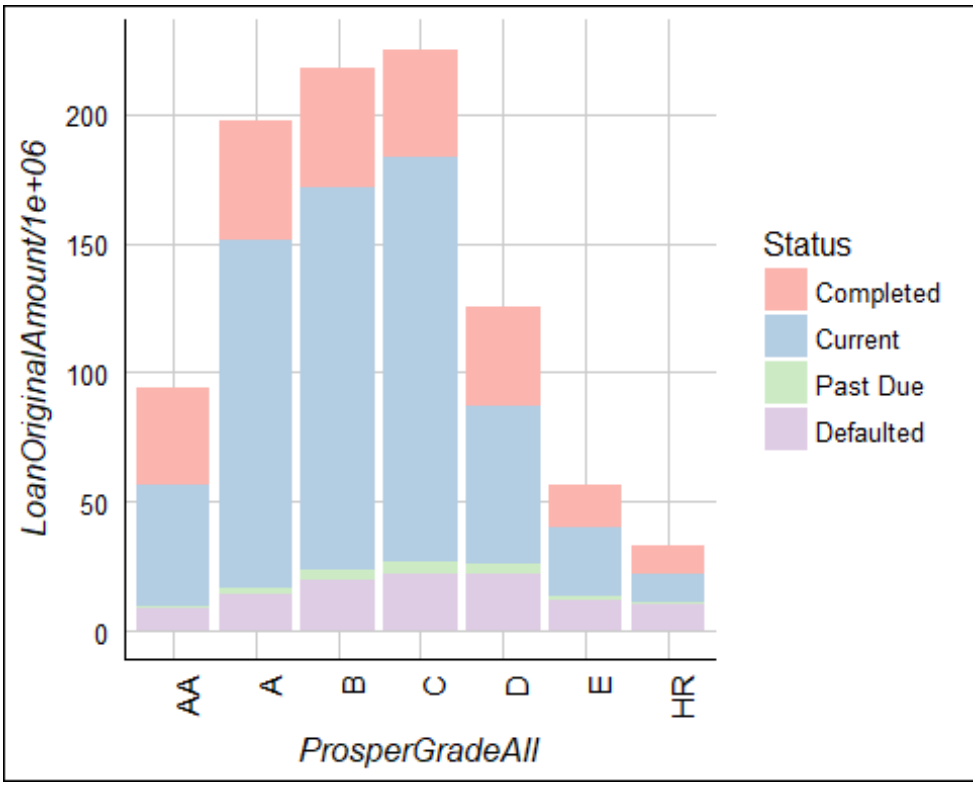
```
ggplot(ld, aes(LoanQtr, LoanOriginalAmount/1000000, fill = ProsperGradeAll)) +
  geom_bar(stat='identity') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, size=10)) +
  scale_fill_brewer(palette="Pastel1")
```



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

```
ggplot(ld, aes(ProsperGradeAll, LoanOriginalAmount/1000000, fill = Status)) +
  geom_bar(stat='identity') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, size=11)) +
  scale_fill_brewer(palette="Pastel1")
```



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

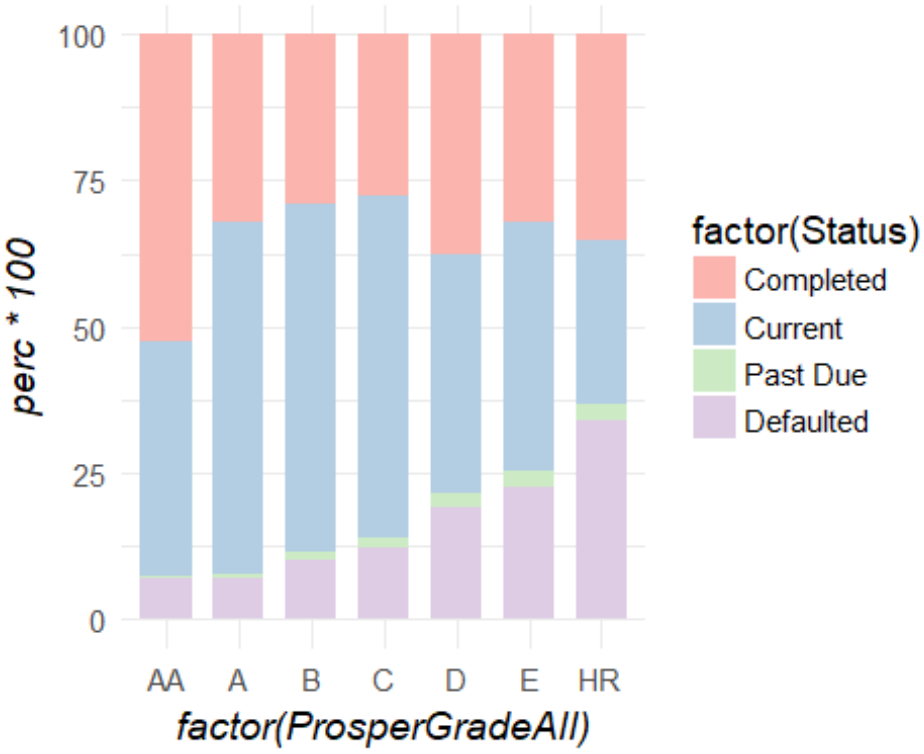
It looks like the largest incremental percentage jump is between C and D

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.

```
ld2 <- ld %>%
  group_by(ProsperGradeAll,Status) %>%
  summarise(count=n()) %>%
  mutate(perc=count/sum(count))


ggplot(ld2, aes(x = factor(ProsperGradeAll), y = perc*100, fill = factor(Status))) +
  geom_bar(stat="identity", width = 0.7) +
  theme_minimal(base_size = 14) +
  scale_fill_brewer(palette="Pastel1")
```



Below we see the different occupations of loan holders and what percentage

of those loans are in good or bad standing

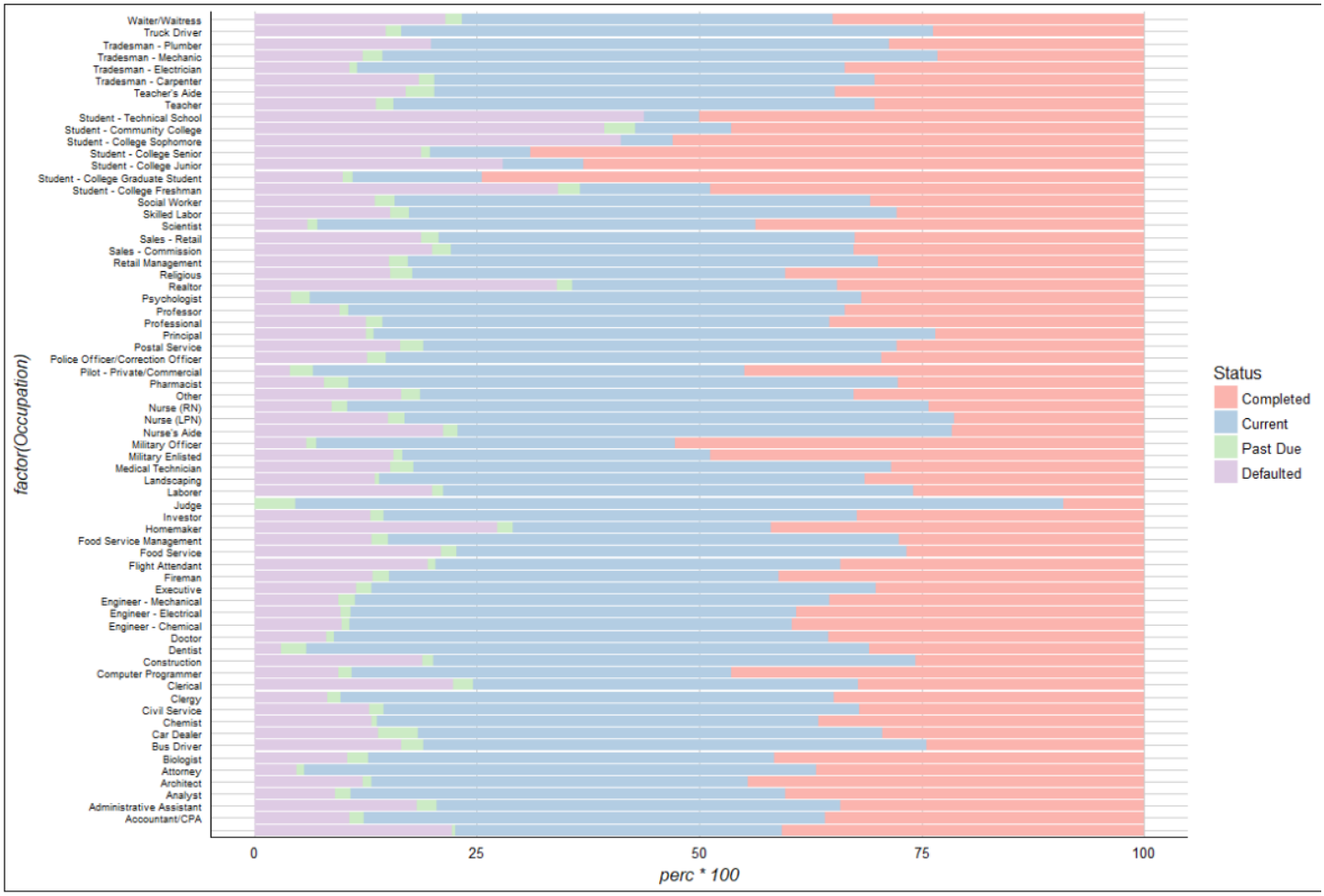
Students at technical schools have the worst performing loans of all

Judges, Psycholigst and Dentists have the lowest default ratings

Further analysis could be done with students where we take into account

Schools they attend, graduation rates, grades, and job placement data

```
ld3 <- ld %>%
  group_by(Occupation,Status) %>%
  summarise(count=n()) %>%
  mutate(perc=count/sum(count))

ggplot(ld3, aes(factor(Occupation), y = perc*100, fill = Status)) +
  geom_bar(stat='identity') +
  theme(axis.text.y = element_text(size=7)) +
  coord_flip() +
  scale_fill_brewer(palette="Pastel1")
```



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

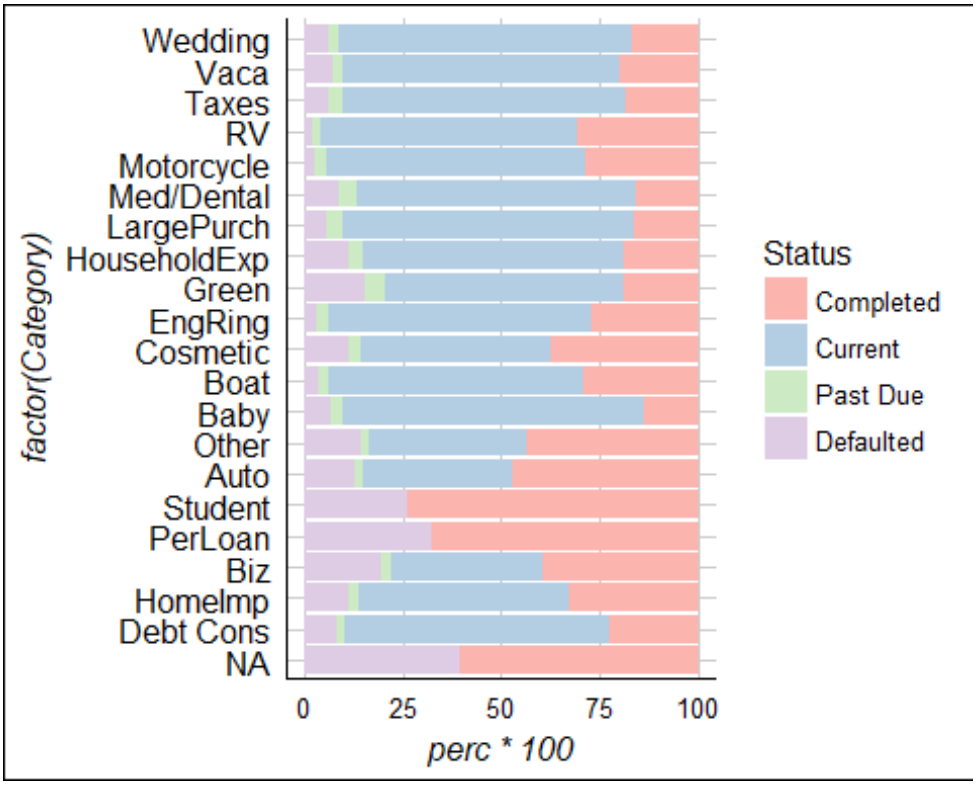
Continue to perform badly and have high defaults.

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.

```
ld4 <- ld %>%
  group_by(Category,Status) %>%
  summarise(count=n()) %>%
  mutate(perc=count/sum(count))

ggplot(ld4, aes(factor(Category), y = perc*100, fill = Status)) +
  geom_bar(stat='identity') +
  coord_flip()+
  theme(axis.text.y = element_text(size=12)) +
  scale_fill_brewer(palette="Pastel1")
```



Final Plots and Summary

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

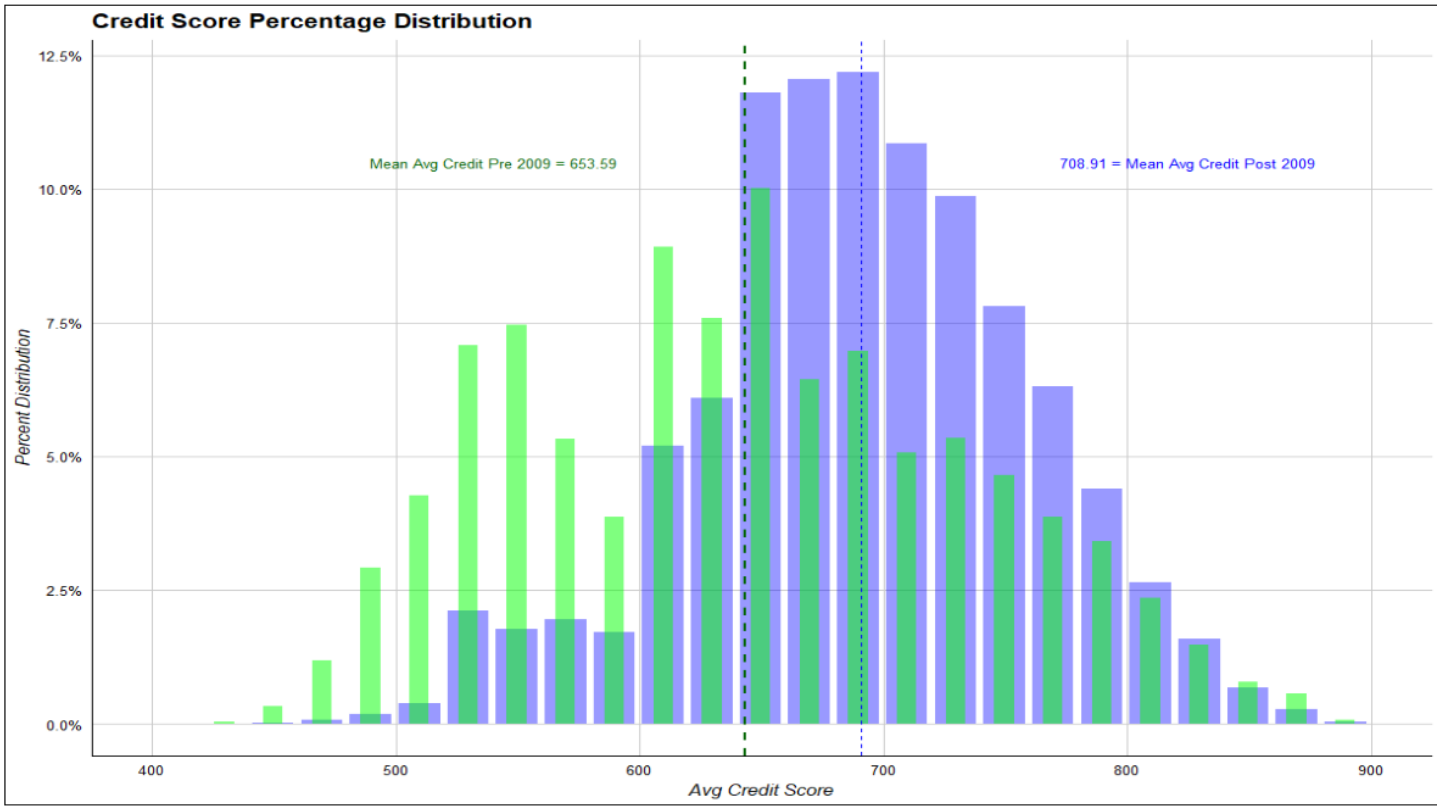
This plot shows both the pre credit crisis and post credit crisis data

It's good to see how Prosper adjusted their standards afterwards

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



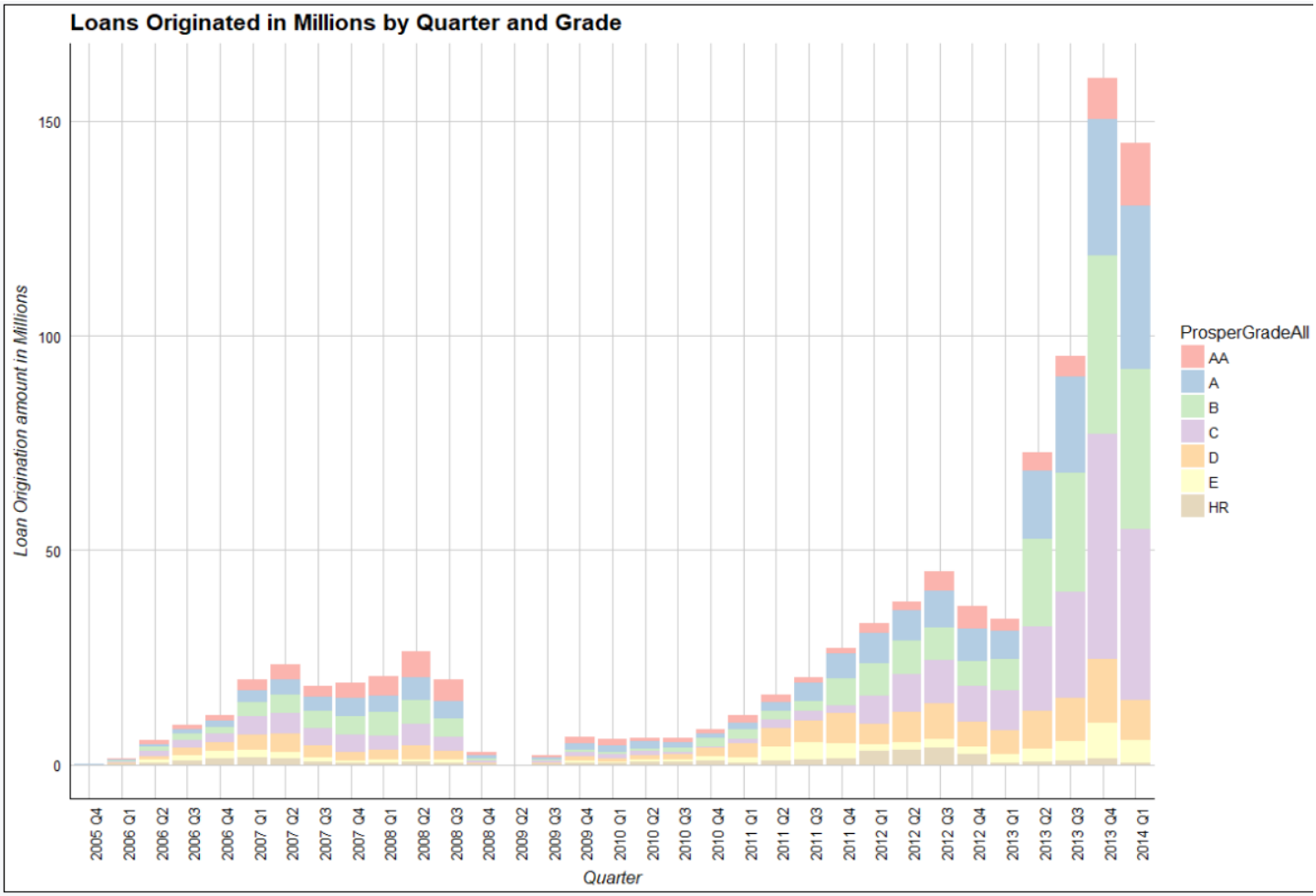
This chart shows how much in loans prosper was generating and by what grade

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

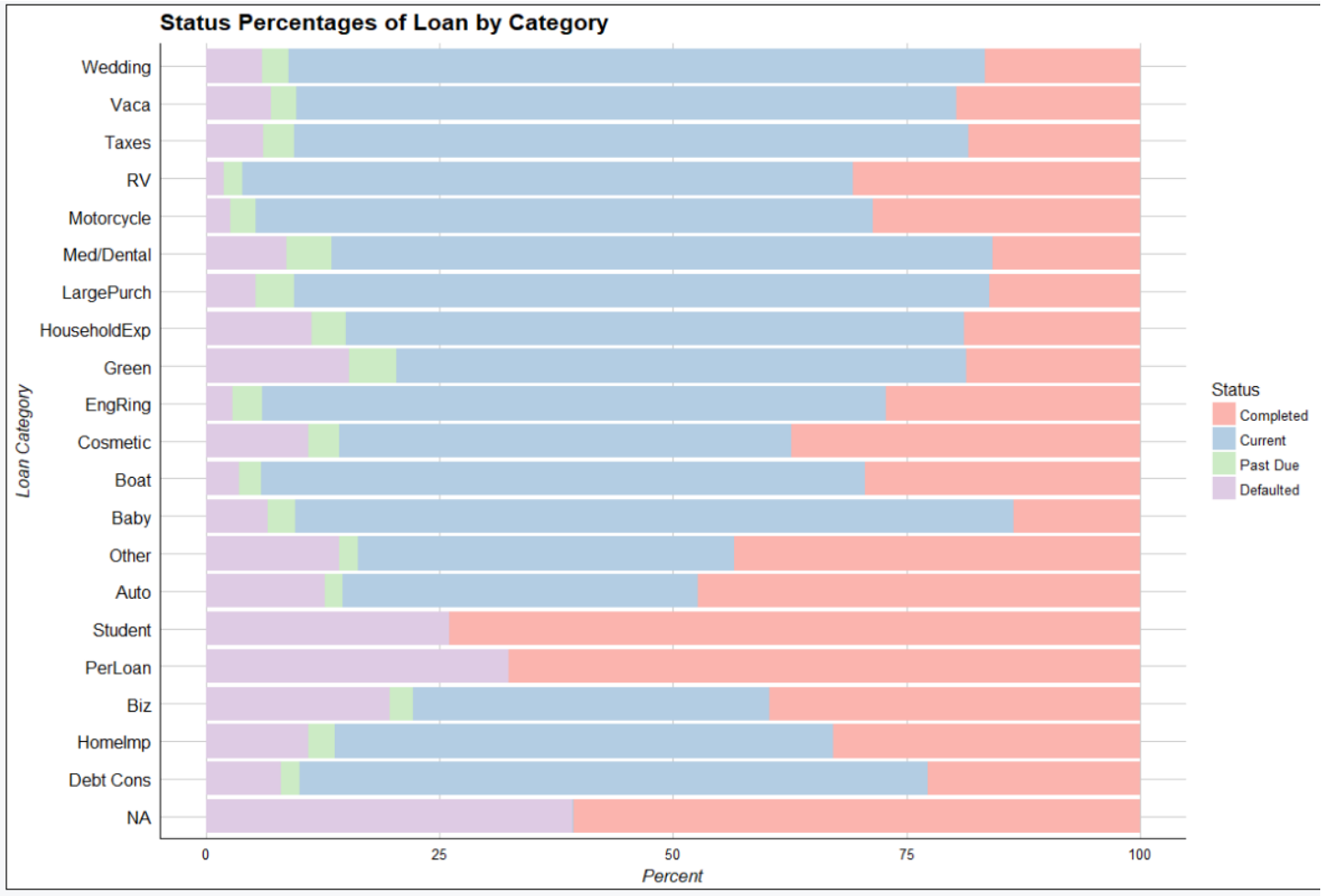
grew the most, this is consistent with our analysis in seeing the C is the

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 RV, Motorcycle, and Boats perform the best because people would not buy these items unless they could afford the 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 to income rate that are closer to that 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

I think there is no reason to give N/A loans, and personal loans 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. There is no excuse to make an uninformed
and non data-driven decision in the future.