ProsperLoanData Analysis

Robin

September 9, 2016

Prosper Loan Data Exploration by Robin Garrow

This report explores a dataset containing amounts, APR, and personal financial information for approximately 85,000 loans over 5 years.

```
## [1] 113937     81
```

The data contains eighty one variables. I hope to determine what other factors besides Prosper Rating may indicate whether or not a borrower will default or miss payments on a loan. Prosper Rating was not added to the data until 2009 so for this analysis I will take a subset of the data to include only records where the Propser Rating exists and loans that have not been cancelled. I am intentionally displaying data manipulations here per the submission guidelines.

```
# Intentionally display this section in the report so data manipulations can be
# observed.
# subset the Prosper Data to capture data of interest and exclude cancelled
#loans

spd <- subset(pd, LoanStatus != "Cancelled" & !is.na(ProsperRating..numeric.),
              select = c(Term, LoanStatus, BorrowerAPR, ProsperRating..numeric.,
                         ProsperRating..Alpha., ListingCategory..numeric.,
                         Occupation, IsBorrowerHomeowner, CreditScoreRangeLower,
                         CreditScoreRangeUpper, CurrentCreditLines,
                         CurrentDelinquencies, DebtToIncomeRatio, IncomeRange,
                         LoanOriginalAmount, LoanOriginationDate))
dim(spd)
```

```
## [1] 84853    16
```

```
summary(spd)
```

```
##       Term                       LoanStatus     BorrowerAPR
##  Min.   :12.00   Current              :56576   Min.   :0.04583
##  1st Qu.:36.00   Completed            :19664   1st Qu.:0.16328
##  Median :36.00   Chargedoff           : 5336   Median :0.21945
##  Mean   :42.49   Defaulted            : 1005   Mean   :0.22666
##  3rd Qu.:60.00   Past Due (1-15 days) :  806   3rd Qu.:0.29254
##  Max.   :60.00   Past Due (31-60 days):  363   Max.   :0.42395
##                  (Other)              : 1103
##  ProsperRating..numeric. ProsperRating..Alpha. ListingCategory..numeric.
##  Min.   :1.000           C      :18345         Min.   : 0.000
##  1st Qu.:3.000           B      :15581         1st Qu.: 1.000
##  Median :4.000           A      :14551         Median : 1.000
                             D      :14274         Mean   : 3.313
##  3rd Qu.:5.000           E      : 9795         3rd Qu.: 3.000
##  Max.   :7.000           HR     : 6935         Max.   :20.000
##                          (Other): 5372
##                Occupation    IsBorrowerHomeowner CreditScoreRangeLower
##  Other              :21317   False:40005         Min.   :600.0
##  Professional       :10542   True :44848         1st Qu.:660.0
##  Executive          : 3468                        Median :700.0
##  Computer Programmer: 3236                       Mean   :699.4
##  Teacher            : 2888                        3rd Qu.:720.0
##  Analyst            : 2735                        Max.   :880.0
##  (Other)            :40667
##  CreditScoreRangeUpper CurrentCreditLines CurrentDelinquencies
##  Min.   :619.0         Min.   : 0.00      Min.   : 0.0000
##  1st Qu.:679.0         1st Qu.: 7.00      1st Qu.: 0.0000
##  Median :719.0         Median :10.00      Median : 0.0000
##  Mean   :718.4         Mean   :10.51      Mean   : 0.3225
##  3rd Qu.:739.0         3rd Qu.:13.00      3rd Qu.: 0.0000
##  Max.   :899.0         Max.   :59.00      Max.   :51.0000
##
##  DebtToIncomeRatio         IncomeRange    LoanOriginalAmount
##  Min.   : 0.000    $50,000-74,999:25627   Min.   : 1000
##  1st Qu.: 0.150    $25,000-49,999:24175   1st Qu.: 4000
##  Median : 0.220    $100,000+     :15205   Median : 7500
##  Mean   : 0.259    $75,000-99,999:14498   Mean   : 9083
##  3rd Qu.: 0.320    $1-24,999     : 4654   3rd Qu.:13500
##  Max.   :10.010    Not employed  :  649   Max.   :35000
##  NA's   :7296      (Other)       :   45
##           LoanOriginationDate
##  2014-01-22 00:00:00:  491
##  2013-11-13 00:00:00:  490
##  2014-02-19 00:00:00:  439
##  2013-10-16 00:00:00:  434
##  2014-01-28 00:00:00:  339
##  2013-09-24 00:00:00:  316
##  (Other)            :82344
```

```
# Convert Loan Origination Date from factor to more usable date
spd$LoanDate <- as.Date(ymd_hms(spd$LoanOriginationDate))
```

```
# Rearrange the order of the Loan status rather than leaving it in alphabetical
# order
spd$LoanStatus <- ordered(spd$LoanStatus, levels = c("Completed",
                                                     "FinalPaymentInProgress",
                                                     "Current",
                                                     "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)",
                                                     "Chargedoff",
                                                     "Defaulted" ))

# Add an numeric field and set the value to a corresponding LoanStatus value,
# 0 is the best and 8 is the worst
spd$StatusCode <- NA
spd <- within(spd, {
  StatusCode[LoanStatus == "Completed" |
               LoanStatus == "FinalPaymentInProgress"] <- 0
  StatusCode[LoanStatus == "Current"] <- 1
  StatusCode[LoanStatus == "Past Due (1-15 days)"] <- 2
  StatusCode[LoanStatus == "Past Due (16-30 days)"] <- 3
  StatusCode[LoanStatus == "Past Due (31-60 days)"] <- 4
  StatusCode[LoanStatus == "Past Due (61-90 days)"] <- 5
  StatusCode[LoanStatus == "Past Due (91-120 days)"] <- 6
  StatusCode[LoanStatus == "Past Due (>120 days)"] <- 7
  StatusCode[LoanStatus == "Chargedoff" | LoanStatus == "Defaulted"] <- 8
  }
  )


# Rearrange order of income range so it makes sense when displayed graphically
spd$IncomeRange <- ordered(spd$IncomeRange, levels = c("Not employed",
                                                       "Not displayed", "$0",
                                                       "$1-24,999",
                                                       "$25,000-49,999",
                                                       "$50,000-74,999",
                                                       "$75,000-99,999",
                                                       "$100,000+"))

# Rearrange order of Prosper rating since AA is better than A
spd$ProsperRating..Alpha. <- ordered(spd$ProsperRating..Alpha.,
                                     levels = c("AA","A", "B", "C", "D",
                                                "E", "HR"))

# If the range of Credit scores exists calculate the mean of the Lower and
# Upper Credit score for each borrower, otherwise set it to NA.

spd$CreditScoreMean <- ifelse(is.na(spd$CreditScoreRangeLower), NA,
                              (spd$CreditScoreRangeLower +
                                 spd$CreditScoreRangeUpper)/2)

# For the rest of our analysis we only want to consider records with a
# CreditScoreMean greater than 9.5 since during Univariate analysis
# we determined these records did not make any sense.
spd <- subset(spd, CreditScoreMean > 9.5)

# Let's just select data with status code of either default or complete
# and save that in a new dataframe for faster processing
spd.sts <- spd %>%
  filter((StatusCode == 0 | StatusCode == 8))

# and lets change it to a factor variable for displaying nicely on graphs
spd.sts$StatusCode <- factor(spd.sts$StatusCode, order = TRUE, levels = c(0, 8),
                             labels = c("Closed", "Defaulted"))
```

Univariate Plots Section

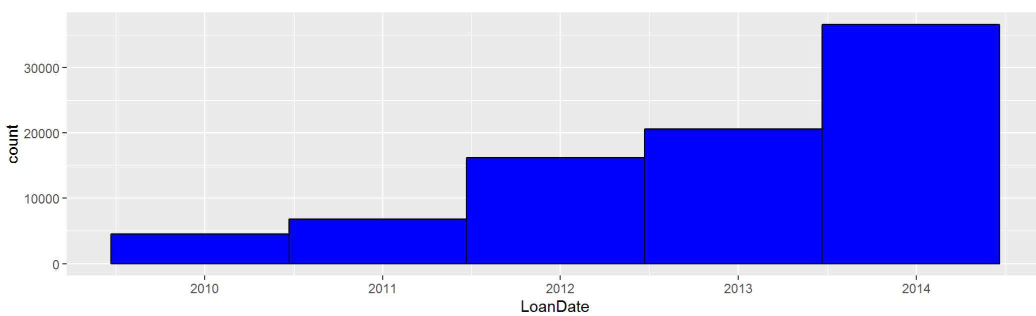
Two additional variables were added in the above code to give 19 variables.

```
## [1] 84853    19
```

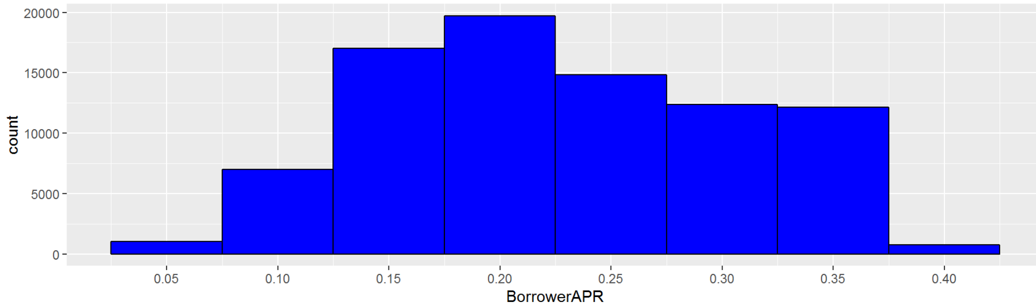
```
## 'data.frame':    84853 obs. of  19 variables:
##  $ Term                     : int  36 36 36 60 36 36 36 36 60 36 ...
##  $ LoanStatus               : Ord.factor w/ 11 levels "Completed"<"FinalPaymentInProgress"<..: 3 3 3 3 3 3 3
3 3 4 ...
##  $ BorrowerAPR              : num  0.12 0.125 0.246 0.154 0.31 ...
##  $ ProsperRating..numeric.  : int  6 6 3 5 2 4 7 7 4 5 ...
##  $ ProsperRating..Alpha.    : Ord.factor w/ 7 levels "AA"<"A"<"B"<"C"<..: 2 2 5 3 6 4 1 1 4 3 ...
##  $ ListingCategory..numeric.: int  2 16 2 1 1 2 7 7 1 1 ...
##  $ Occupation               : Factor w/ 68 levels "","Accountant/CPA",..: 43 52 21 43 50 29 24 24 22 50 ...
##  $ IsBorrowerHomeowner      : Factor w/ 2 levels "False","True": 1 2 2 2 1 1 2 2 1 1 ...
##  $ CreditScoreRangeLower    : int  680 800 680 740 680 700 820 820 640 680 ...
##  $ CreditScoreRangeUpper    : int  699 819 699 759 699 719 839 839 659 699 ...
##  $ CurrentCreditLines       : int  14 5 19 21 10 6 17 17 2 9 ...
##  $ CurrentDelinquencies     : int  0 4 0 0 0 0 0 0 1 0 ...
##  $ DebtToIncomeRatio        : num  0.18 0.15 0.26 0.36 0.27 0.24 0.25 0.25 0.12 0.18 ...
##  $ IncomeRange              : Ord.factor w/ 8 levels "Not employed"<..: 6 5 8 8 5 5 5 5 7 5 ...
##  $ LoanOriginalAmount       : int  10000 10000 15000 15000 3000 10000 10000 10000 13500 4000 ...
```

```
##  $ LoanOriginationDate      : Factor w/ 1873 levels "2005-11-15 00:00:00",..: 1866 1535 1757 1821 1649 1666 1
813 1813 1419 1829 ...
##  $ LoanDate                 : Date, format: "2014-03-03" "2012-11-01" ...
##  $ StatusCode               : num  1 1 1 1 1 1 1 1 1 2 ...
##  $ CreditScoreMean          : num  690 810 690 750 690 ...
```

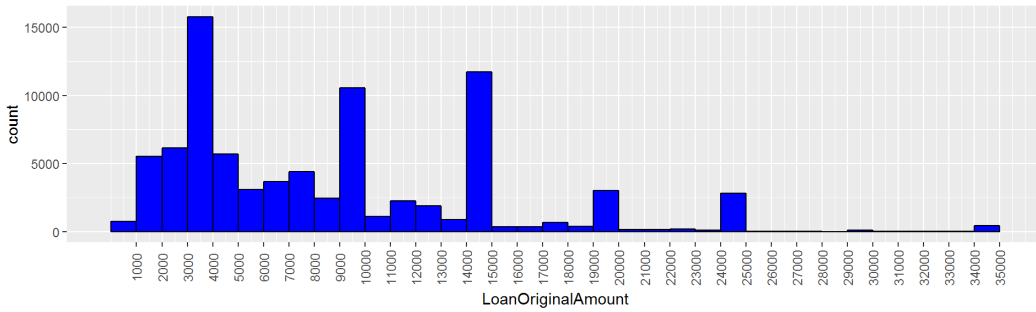
```
##       Term                        LoanStatus     BorrowerAPR
##  Min.   :12.00   Current              :56576   Min.   :0.04583
##  1st Qu.:36.00   Completed            :19664   1st Qu.:0.16328
##  Median :36.00   Chargedoff           : 5336   Median :0.21945
##  Mean   :42.49   Defaulted            : 1005   Mean   :0.22666
##  3rd Qu.:60.00   Past Due (1-15 days) :  806   3rd Qu.:0.29254
##  Max.   :60.00   Past Due (31-60 days):  363   Max.   :0.42395
##                  (Other)              : 1103
##  ProsperRating..numeric. ProsperRating..Alpha. ListingCategory..numeric.
##  Min.   :1.000           AA: 5372              Min.   : 0.000
##  1st Qu.:3.000           A :14551              1st Qu.: 1.000
##  Median :4.000           B :15581              Median : 1.000
##  Mean   :4.072           C :18345              Mean   : 3.313
##  3rd Qu.:5.000           D :14274              3rd Qu.: 3.000
##  Max.   :7.000           E : 9795              Max.   :20.000
##                          HR: 6935
##                Occupation    IsBorrowerHomeowner CreditScoreRangeLower
##  Other              :21317   False:40005         Min.   :600.0
##  Professional       :10542   True :44848         1st Qu.:660.0
##  Executive          : 3468                       Median :700.0
##  Computer Programmer: 3236                       Mean   :699.4
##  Teacher            : 2888                       3rd Qu.:720.0
##  Analyst            : 2735                       Max.   :880.0
##  (Other)            :40667
##  CreditScoreRangeUpper CurrentCreditLines CurrentDelinquencies
##  Min.   :619.0         Min.   : 0.00      Min.   : 0.0000
##  1st Qu.:679.0         1st Qu.: 7.00      1st Qu.: 0.0000
##  Median :719.0         Median :10.00      Median : 0.0000
##  Mean   :718.4         Mean   :10.51      Mean   : 0.3225
##  3rd Qu.:739.0         3rd Qu.:13.00      3rd Qu.: 0.0000
##  Max.   :899.0         Max.   :59.00      Max.   :51.0000
##
##  DebtToIncomeRatio          IncomeRange    LoanOriginalAmount
##  Min.   : 0.000    $50,000-74,999:25627   Min.   : 1000
##  1st Qu.: 0.150    $25,000-49,999:24175   1st Qu.: 4000
##  Median : 0.220    $100,000+     :15205   Median : 7500
##  Mean   : 0.259    $75,000-99,999:14498   Mean   : 9083
##  3rd Qu.: 0.320    $1-24,999     : 4654   3rd Qu.:13500
##  Max.   :10.010    Not employed  :  649   Max.   :35000
##  NA's   :7296      (Other)       :   45
##            LoanOriginationDate    LoanDate             StatusCode
##  2014-01-22 00:00:00:  491    Min.   :2009-07-20   Min.   :0.000
##  2013-11-13 00:00:00:  490    1st Qu.:2012-02-23   1st Qu.:1.000
##  2014-02-19 00:00:00:  439    Median :2013-04-09   Median :1.000
##  2013-10-16 00:00:00:  434    Mean   :2012-11-15   Mean   :1.351
##  2014-01-28 00:00:00:  339    3rd Qu.:2013-11-05   3rd Qu.:1.000
##  2013-09-24 00:00:00:  316    Max.   :2014-03-12   Max.   :8.000
##  (Other)            :82344
##  CreditScoreMean
##  Min.   :609.5
##  1st Qu.:669.5
##  Median :709.5
##  Mean   :708.9
##  3rd Qu.:729.5
##  Max.   :889.5
##
```



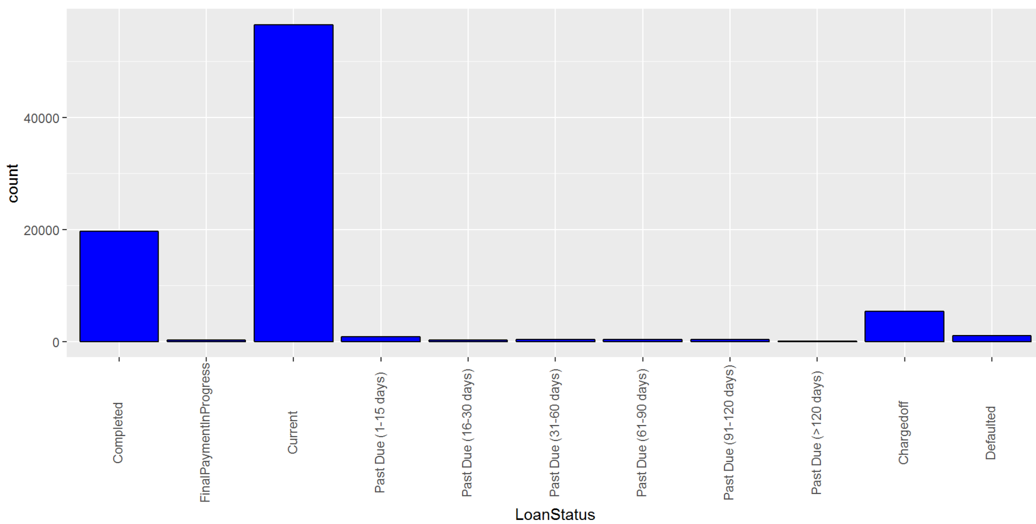
The histogram of loan origination dates is skewed to the left indicating more loans have been funded recently



At first glance APR values appear to have a fairly symmetric, unimodal distribution

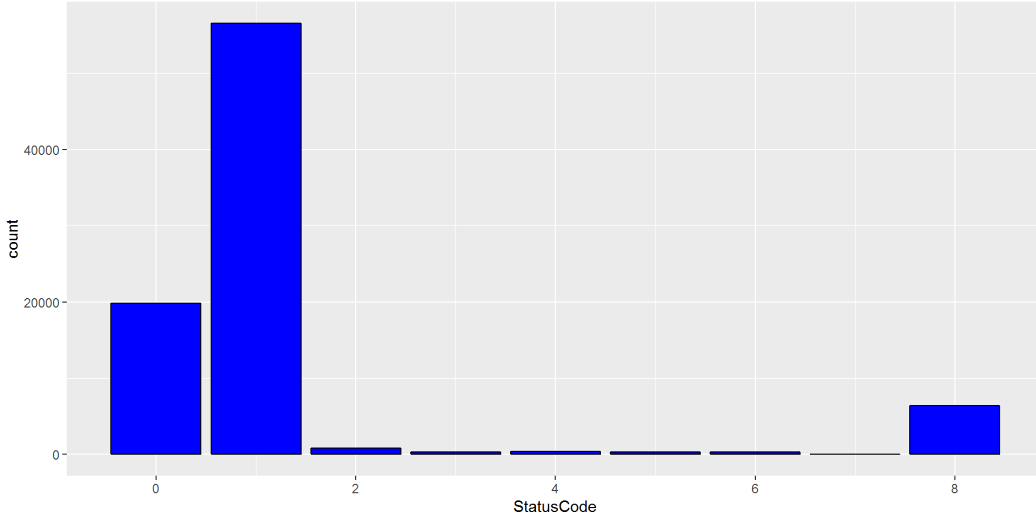


Loan amounts are skewed to the right indicating higher loan amounts are less common.



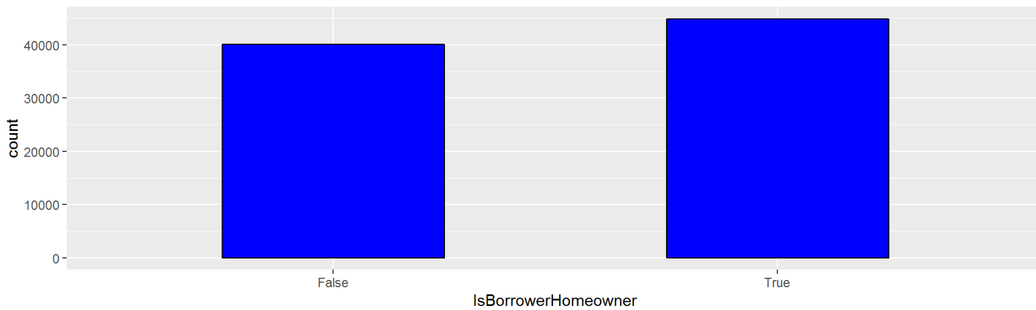

```
##
##              Completed FinalPaymentInProgress                Current
##                  19664                    205                  56576
##   Past Due (1-15 days)  Past Due (16-30 days)  Past Due (31-60 days)
##                    806                    265                    363
##  Past Due (61-90 days) Past Due (91-120 days)   Past Due (>120 days)
##                    313                    304                     16
##             Chargedoff              Defaulted
##                   5336                   1005
```

From the bar graph and table we see over 6,000 loans are in a default or charged off status

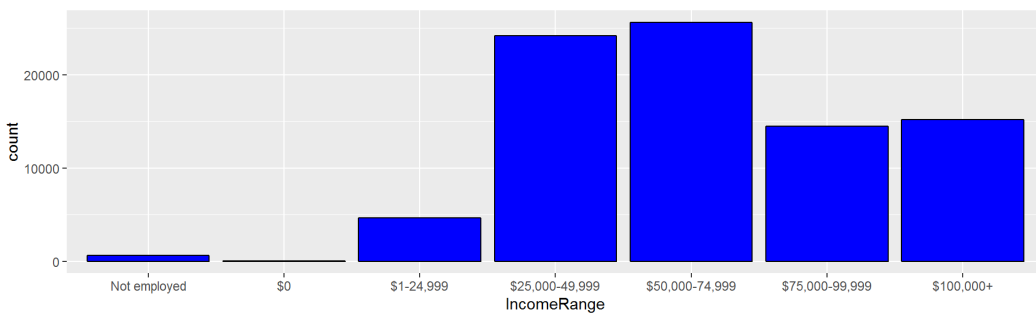



```
## [1] 0.2419306
```

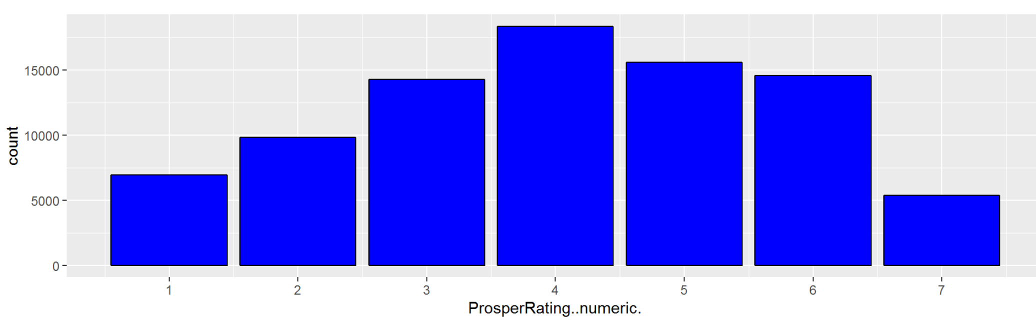
This graph uses the newly created variable StatusCode where Loan Status values Completed & FinalPaymentinProgress were combined as were Chargedoff and Defaulted. Approximately 24% of the finished loans are in a default status.



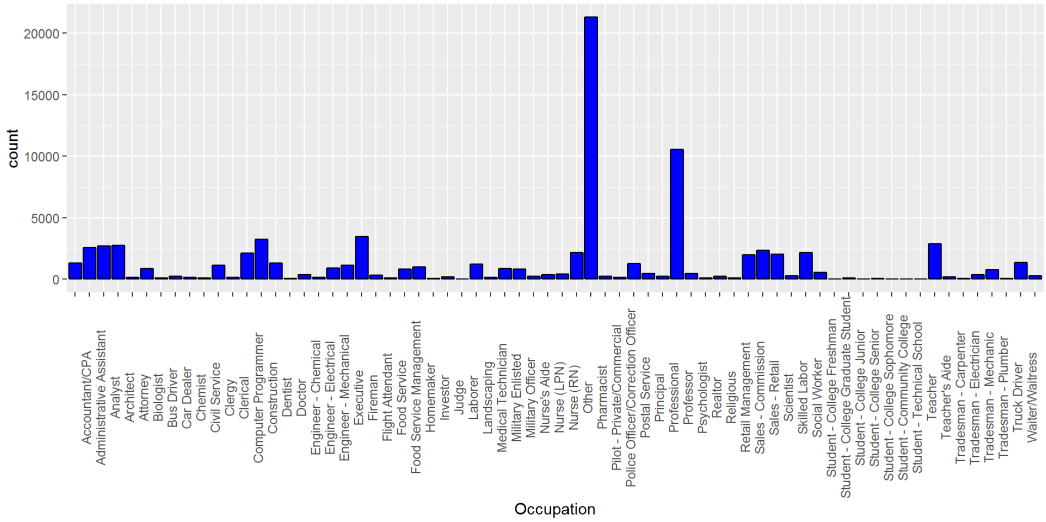
From this graph we can see about 10,000 more borrowers own homes than those who do not.



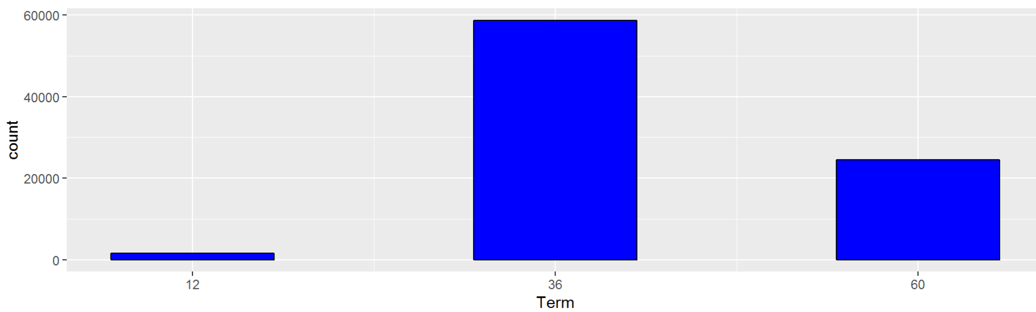
It appears a majority of borrowers have income in the $25,000 to $75,000 range. This is a self report field so the actual earnings may be different and we will need to keep this in mind when attempting to draw conclusions.



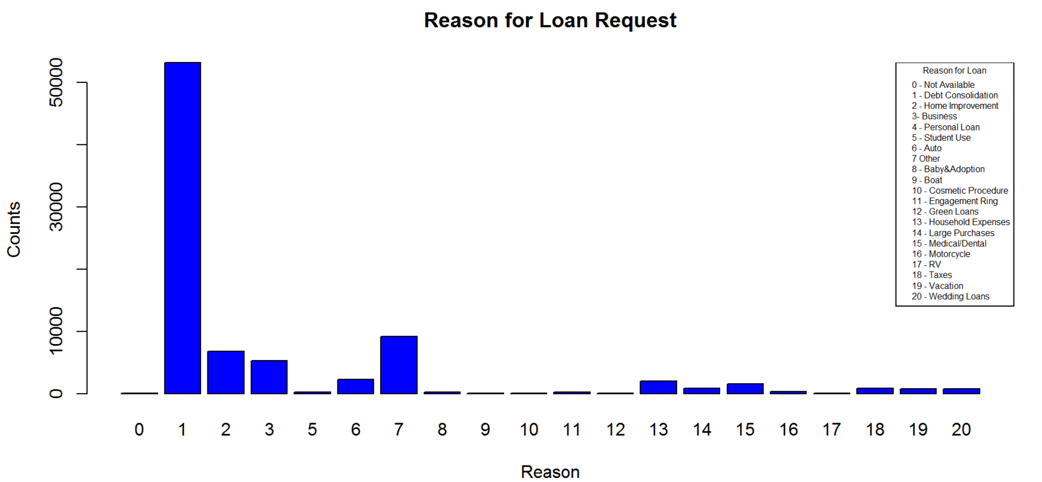
The distribution of scores determined by Prosper are symmetrical but with the mode, or most common score being 4.



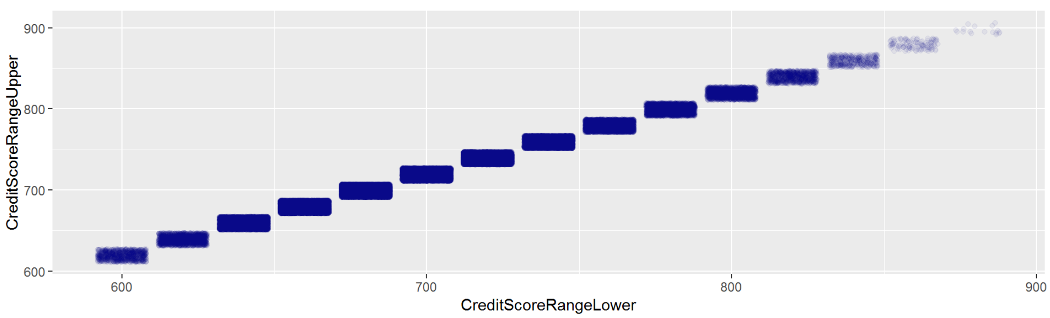
Occupation is selected by the borrower at the time they created the listing which could account for the fact that "Other" is the most common occupation.



Three years or 36 months is the most common loan term length.



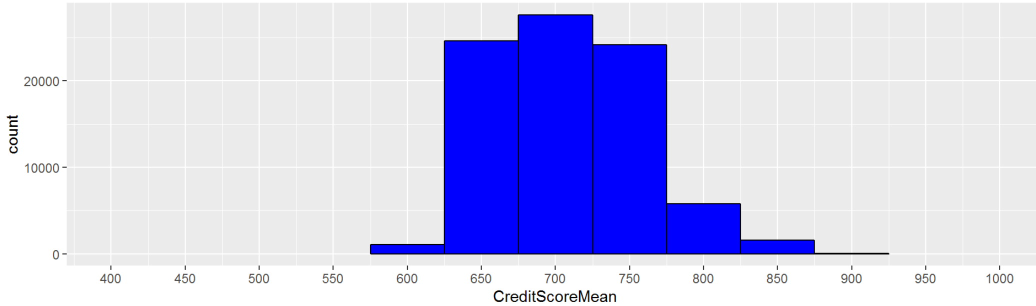
Reason "Not Available" is the most common response. I wonder if this will affect the APR rate assigned to the loan and will analyze later in the bivariate analysis.



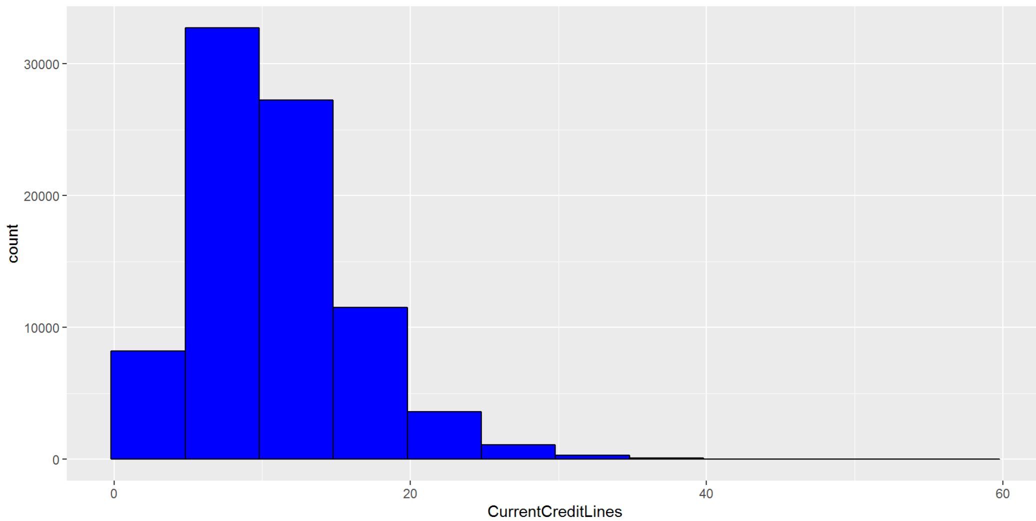

```
## [1] 84853
```

I realize this is technically a bivariate plot but I want to see if creating one variable from them is appropriate to to then create a univariate plot.

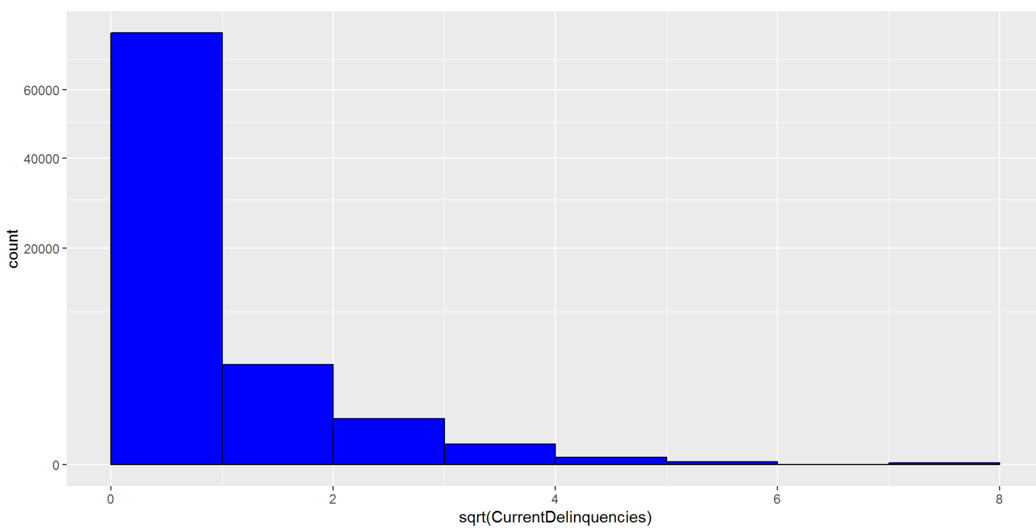
This scatterplot shows the Upper and Lower Credit Scores of individuals as reported by consumer credit rating agencies. Since the pattern of these values form a line axis it indicates the reporting agencies return about the same score for individuals with similar personal financial information. Due to this it seems using the mean of these two scores is appropriate.



The graph showing the average credit score for each borrower is skewed somewhat to the right with a peak around 700.

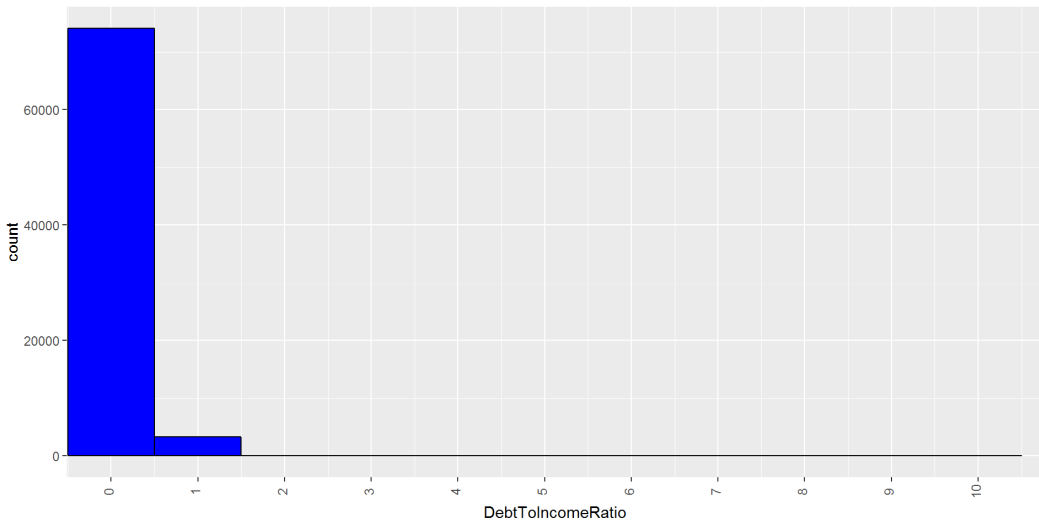


At the time the credit profile was pulled most people had between 5 and 10 Current Credit Lines. We can again see the data is skewed to the right



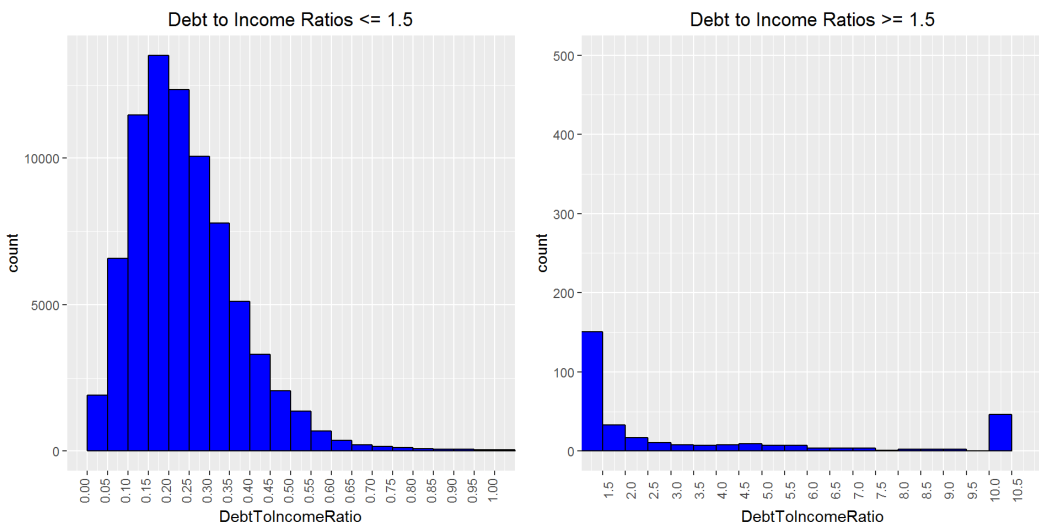

```
##
##     0     1     2     3     4     5     6     7     8     9    10    11
## 71252  8223  2631  1039   611   310   232   167   106    79    54    43
##    12    13    14    15    16    17    18    19    20    21    22    24
##    23    26    11    12     9     9     2     1     1     5     1     2
##    27    32    51
##     2     1     1
```

I took the square root of the number of Delinquencies and transformed the y scale by taking the square root. The majority of borrower's had 0 delinquencies at the time of requesting a loan. In rare cases we see 4 or more delinquencies. Surprisingly we we see a borrowers with over 20 delinquencies. It will be very interesting to see how delinquencies might relate to loan status.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   0.150   0.220   0.259   0.320  10.010    7296
```

According to the Prosper documentation the possible value is capped at 10.01. I decided to split the graph into two sections in order to zoom in on the values.



The most common debt to income ratio is between .15 and .20. I find it quite interesting that the Debt to Income ratio from 0 to 0.5 almost fits the normal model. It is quite unusual for borrowers to have a ratio above .5.

Univariate Analysis

What is the structure of your dataset?

There were 113,937 loans in the dataset with 81 variables. I considered a subset of the variables(Term, LoanStatus, BorrowerAPR, ProsperRating..numeric., ProsperRating..alpha., ListingCategory..numeric.,Occupation, IsBorrowerHomeowner, CreditScoreRangeLower, CreditScoreRangeUpper, CurrentCreditLines, CurrentDelinquencies, DebtToIncomeRatio, IncomeRange, LoanOriginalAmount, LoanOriginationDate) and excluded any records that had a loan status of “Cancelled”or did not have a prosper rating. This resulted in 84,853 loans left for analysis.

In the original dataset LoanStatus, Occupation, IsBorrowerHomeowner, IncomeRange, and LoanOriginationDate were all factor variables. LoanStatus and IncomeRange were not ordered and I decided to apply an order to them which is reflected below along with the factors of the other variables.

LoanStatus - Ordered “Completed”, “FinalPaymentInProgress”, “Current”, “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)”, “Chargedoff”, “Defaulted”

Occupation, 68 levels - no Order

Occupations	Occupations	Occupations
”	Accountant/CPA	Administrative Assistant
Analyst	Architect	Attorney
Biologist	Bus Driver	Car Dealer
Chemist	Civil Service	Clergy
Clerical	Computer Programmer	Construction
Dentist	Doctor	Engineer - Chemical
Engineer - Electrical	Engineer - Mechanical	Executive
Fireman	Flight Attendant	Food Service
Food Service Management	Homemaker	Investor

Judge	Laborer	Landscaping
Medical Technician	Military Enlisted	Military Officer
Nurse's Aide	Nurse (LPN)	Nurse (RN)
Other	Pharmacist	Pilot - Private/Commercial
Police Officer/Correction Officer	Postal Service	Principal
Professional	Professor	Psychologist
Realtor	Religious	Retail Management
Sales - Commission	Sales - Retail	Scientist
Skilled Labor	Social Worker	Student - College Freshman
Student - College Graduate Student	Student - College Junior	Student - College Senior
Student - College Sophomore	Student - Community College	Student - Technical School
Teacher	Teacher's Aide	Tradesman - Carpenter
Tradesman - Electrician	Tradesman - Mechanic	Tradesman - Plumber
Truck Driver	Waiter/Waitress	

IsBorrowerHomewowner "False", "True"

IncomeRange - Ordered "Not employed", "Not displayed", "$0", "$1-24,999", "$25,000-49,999", "$50,000-74,999", "$75,000-99,999", "$100,000+"

LoanOriginationDate Factor variable which was not particularly useful for my analysis so I created the new variable LoanDate using this data. Please see below for additional information.

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

The main features in the data set are ProsperRating..numeric. and Loan Status. I hope to determine which features are best for predicting if the loan associated with a borrower's loan status. Since Prosper provides a service to both borrowers and investors it would be useful for investors to be able to predict which loans are most likely to be completed. Prosper's website for available loans to invest in gives the Loan Category which is character value corresponding to ProsperRating..numeric. AA or numerically 7 is the best rating with HR(i.e. 1) the worst. This field is only applicable for loans originated after July 2009. In order to see what other information is available regarding the borrower's particulars you must register as an investor with Prosper. It will be intersting to see if the Prosper Score is truly the best predictor or if other features are better.

What other features in the dataset do you think will help support your investigation into your feature(s) of interest?

APR, Occupation, Listing Category, IsBorrowerHomeowner, Credit Scores, CurrentDelinquencies, DebtToIncomeRatio, and IncomeRange are potential indicators of whether or not a loan will end in default or charge off.

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

I used the variable LoanOriginationDate to create the variable **LoanDate** which is a date variable in YYYY-MM-DD format. The original variable was a factor variable and contained a timestamp as well which is not useful for my analysis.

LoanStatus was used to create the variable **StatusCode**, a numeric field in dataframe SPD, which I consider to be my primary response(dependent) variable. I considered Completed & FinalPaymentinProgress essentially the same status but due to timing are in different categories and assigned 0 to StatusCode. The same is true of Chargedoff and Defaulted so they were assigned status code 8.

CreditScoreMean was created by taking the average of a borrower's CreditScoreRangeLower and CreditScoreRangeUpper. Multiple agencies may be contacted for a borrower's credit scrore. During the Univariate section I did do a scatterplot of these two scores which I know is bivariate analysis. I thought this was the appropriate time to see what the relationship was between these two scores so that I would know which one to use during the true bivariate analysis phase when comparing Credit Score to Loan Status. Since their relationship was linear it seemed appropriate to use the average of Lower and Upper Credit Scores.

spd.sts is a dataframe which is a subset of spd and includes only records that are in a default(8) or closed(1) status. I then changed the StatusCode in this dataframe to a factor variable with two levels. This file is used in the bivariate and multivariate analysis.

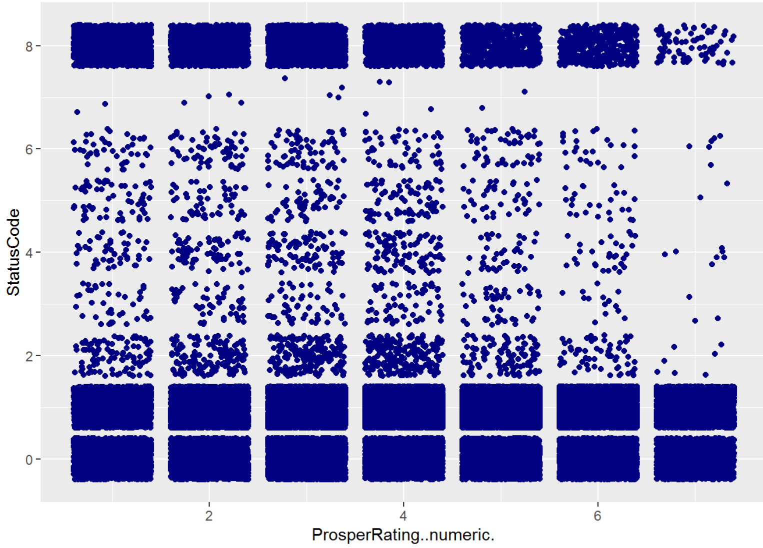
Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

I did use lubridate to create the new field Loan Date to make this feature more useful as a time stamp if not necessary for my analysis.

For the number of delinquencies to standout graphically I decided to take the square root of both the CurrentDelinquencies and y scale.

I thought it was interesting that the number of homeowners and non-homeowners is about the same. It will be interesting to see if one or the other is more likely to default on a loan.

Bivariate Plots Section

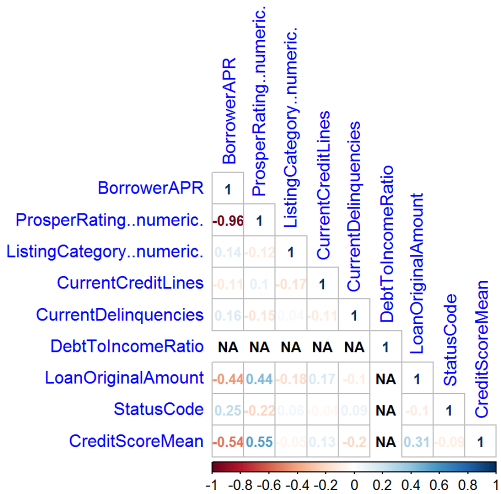



```
##
##  Pearson's product-moment correlation
##
## data:  spd$ProsperRating..numeric. and spd$StatusCode
## t = -55.129, df = 84851, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1924422 -0.1794506
## sample estimates:
##        cor
## -0.1859545
```

There is not a linear relationship between Prosper Risk Rating and the status of a loan as indicated by the correlation coefficient of -.19 as well as the plot. Let's see if a stronger relationship between other variables exists.

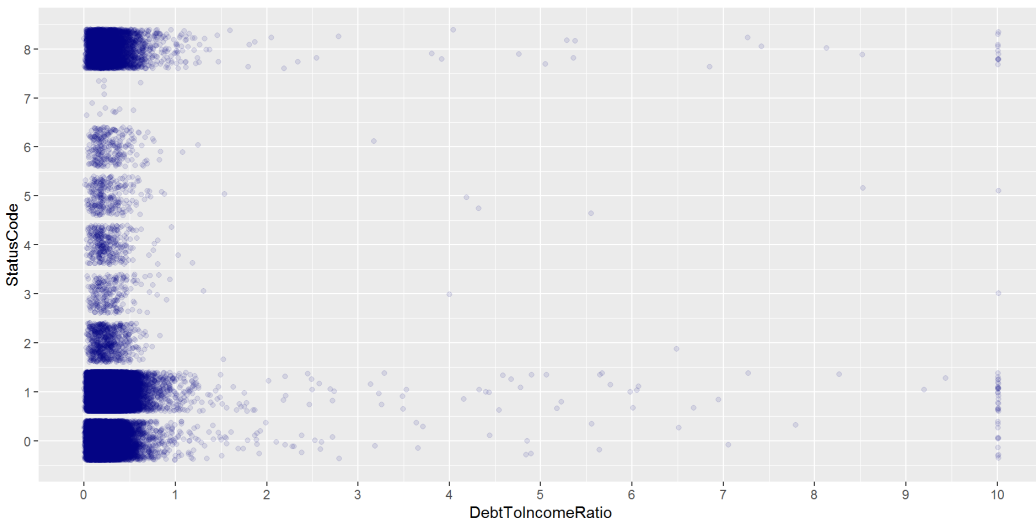
```
##                             BorrowerAPR ProsperRating..numeric.
## BorrowerAPR                  1.0000000             -0.96215126
## ProsperRating..numeric.     -0.9621513              1.00000000
## ListingCategory..numeric.    0.1087837             -0.09447405
## CurrentCreditLines          -0.1095986              0.09237706
## CurrentDelinquencies         0.1538153             -0.14520526
## DebtToIncomeRatio            0.1288220             -0.13534359
## LoanOriginalAmount          -0.4263610              0.42855722
## StatusCode                   0.2166196             -0.18595454
## CreditScoreMean             -0.5258881              0.54887385
##                             ListingCategory..numeric. CurrentCreditLines
## BorrowerAPR                               0.108783740        -0.10959863
## ProsperRating..numeric.                  -0.094474047         0.09237706
## ListingCategory..numeric.                 1.000000000        -0.13339337
## CurrentCreditLines                       -0.133393374         1.00000000
## CurrentDelinquencies                      0.062200585        -0.13152708
## DebtToIncomeRatio                        -0.041342845         0.14661499
## LoanOriginalAmount                       -0.202322201         0.19296190
## StatusCode                                0.031855085        -0.06747953
## CreditScoreMean                          -0.007928902         0.09335291
##                             CurrentDelinquencies DebtToIncomeRatio
## BorrowerAPR                           0.15381529        0.12882198
## ProsperRating..numeric.              -0.14520526       -0.13534359
## ListingCategory..numeric.             0.06220058       -0.04134284
## CurrentCreditLines                   -0.13152708        0.14661499
## CurrentDelinquencies                  1.00000000       -0.03839116
## DebtToIncomeRatio                    -0.03839116        1.00000000
## LoanOriginalAmount                   -0.11111509       -0.01783746
## StatusCode                            0.05701356        0.04589815
## CreditScoreMean                      -0.16013458       -0.01370880
##                             LoanOriginalAmount  StatusCode CreditScoreMean
## BorrowerAPR                        -0.42636102  0.21661957    -0.525888129
## ProsperRating..numeric.             0.42855722 -0.18595454     0.548873850
## ListingCategory..numeric.          -0.20232220  0.03185508    -0.007928902
## CurrentCreditLines                  0.19296190 -0.06747953     0.093352909
## CurrentDelinquencies               -0.11111509  0.05701356    -0.160134579
## DebtToIncomeRatio                  -0.01783746  0.04589815    -0.013708795
## LoanOriginalAmount                  1.00000000 -0.07535678     0.277918466
## StatusCode                         -0.07535678  1.00000000    -0.089583397
## CreditScoreMean                     0.27791847 -0.08958340     1.000000000
```

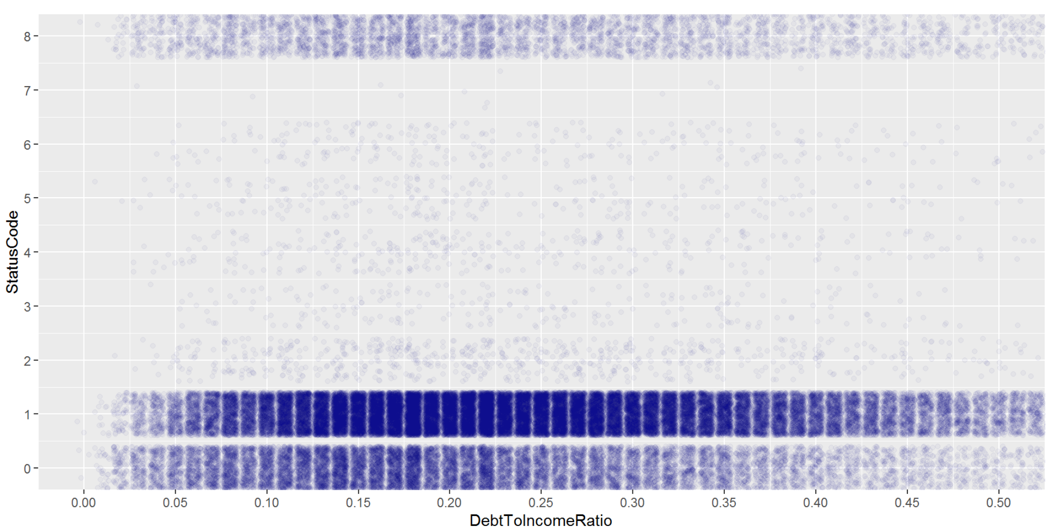
The correlation between the loan risk rating assigned by Prosper and the offered APR is quite strong with r = -.96. This is not surprising since one would expect borrowers with good(high ratings) credit history to be offered a lower APR than one who presents as bad risk(low rating) Additionally it appears that the APR in turn has a slightly stronger relationship r = .22 with loan status code than the Prosper rating. I wonder how allowing for the categorical variables affects the relationship.



Proper Risk Rating and Credit Score Mean are only moderately correlated with r = .55. This indicates Prosper does not use only Credit Scores to determine the risk rating of a loan

I want to look closer at scatter plots involving Status Code as it relates to Debt/Income, Number of Credit Lines, Reason, Number of Delinquencies, and Mean Credit Score. I don't believe we will see much of a linear relationship but it will let me get quick idea of where values are clustered.

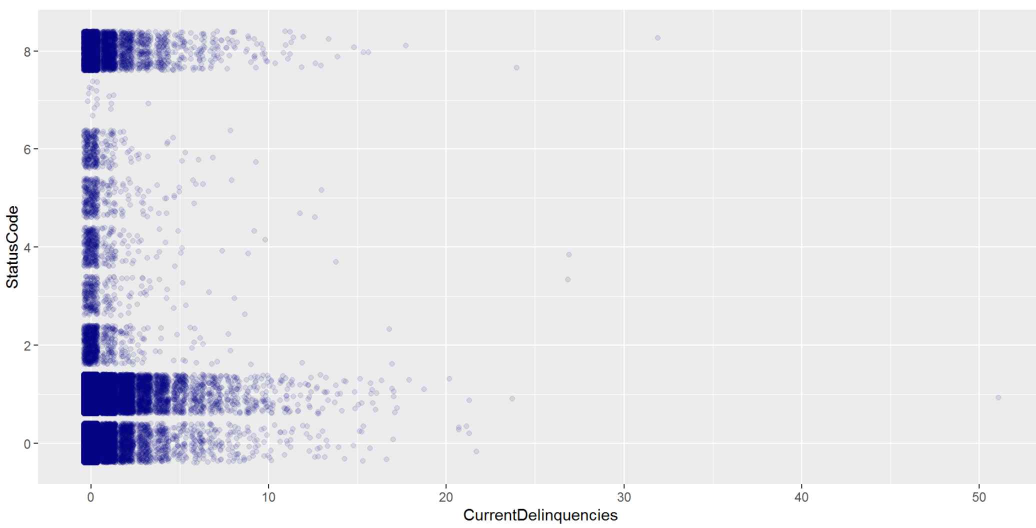




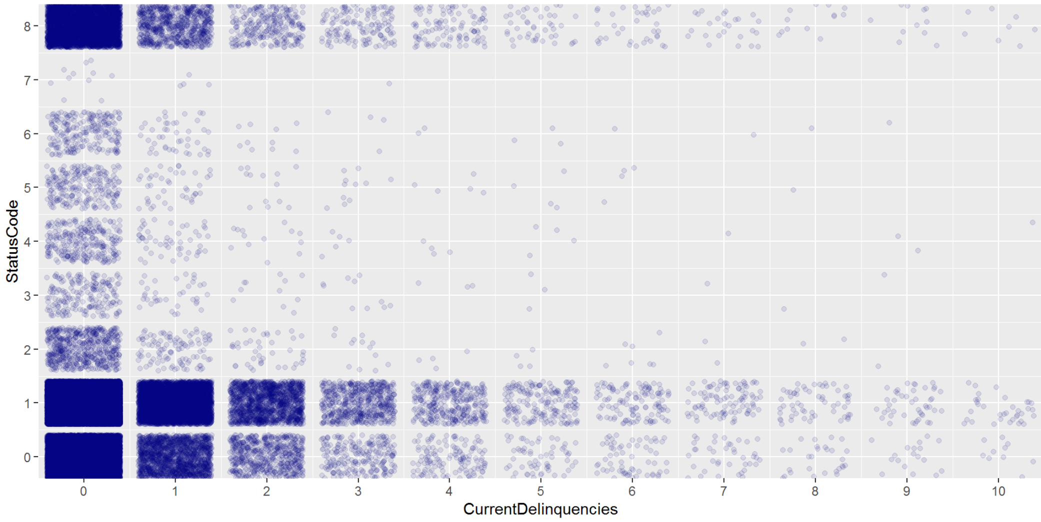

```
##       StatusCode
##          0    1    2    3    4    5    6    7    8
##   0      3    2    0    0    0    0    0    0    1
##   0.01  31   12    0    0    0    1    0    0    2
##   0.02 115   63    1    0    0    1    0    0   23
##   0.03 187  153    2    0    3    3    0    1   53
##   0.04 235  210    4    1    3    1    1    0   54
##   0.05 301  340   11    3    4    4    5    0   70
##   0.06 351  484   13    2    6    3    3    0   75
```

```
##    0.07  410  633    6    6    3    4    2    0   93
##    0.08  525  866   17    2   10    5    4    0  134
##    0.09  403  756   19    4    5    8    4    1   86
##    0.1   508  975   12    5    4    5    3    0  117
##    0.11  547 1240   25    3    8   11   10    0  127
##    0.12  586 1258   14    9    7    5    9    0  145
##    0.13  651 1489   12    7    6    8   11    0  154
##    0.14  778 1672   30    6   12   11    5    0  189
##    0.15  608 1617   16    8    9    5   11    0  159
##    0.16  659 1688   24    7   11    4   11    1  162
##    0.17  710 1789   27   10   15    7    7    1  183
##    0.18  835 2066   29    7   13   20   12    0  231
##    0.19  518 1740   18    2   17   11   11    0  127
##    0.2   566 1789   14    5   15   10    7    0  143
##    0.21  551 1794   28    6   10   10    6    1  172
##    0.22  713 2003   24   13   10    9   13    2  212
##    0.23  438 1655   17    4    7    2    2    1  116
##    0.24  479 1656   24    6    5    2    3    0  122
##    0.25  435 1649   18    8    7    2    8    0  107
##    0.26  392 1627   22    9    6    5    6    0  112
##    0.27  389 1566   20    2   11    7    6    0  123
##    0.28  397 1515   17    6   11    6    3    0  119
##    0.29  357 1341   19    5    8   10    7    0  115
##    0.3   329 1353   22    2    6    6    3    0  100
##    0.31  326 1275    6    3    4    3    4    0  116
##    0.32  300 1186   14    7    8    6    6    1  110
##    0.33  275 1104   19    4    7    2    4    0   98
##    0.34  258 1042   15    5    4    6    5    1   91
##    0.35  296 1031    9   13    4    4    8    1   97
##    0.36  188  908    7    6   11    2    4    0   65
##    0.37  189  786    8    3    3    7    5    0   77
##    0.38  178  777   11    2    3    3    5    0   61
##    0.39  163  701   14    2    2    6    8    1   42
##    0.4   152  623   13    3    7    2    2    0   67
##    0.41  108  633    5    6    1    1    3    0   46
##    0.42  117  506    8    1    4    2    4    0   62
##    0.43   98  500    5    5    1    3    2    0   43
##    0.44  106  428    4    1    3    2    3    0   41
##    0.45   94  383   12    3    4    5    3    0   41
##    0.46   83  347    7    2    5    2    2    0   42
##    0.47   72  358    4    3    2    1    2    0   41
##    0.48   52  296    3    1    1    1    2    0   35
##    0.49   65  249    8    0    3    2    0    0   23
##    0.5    51  241    7    0    0    2    4    0   33
```

Between .05 and .35 we see an increase in the number of defaults and late payments.



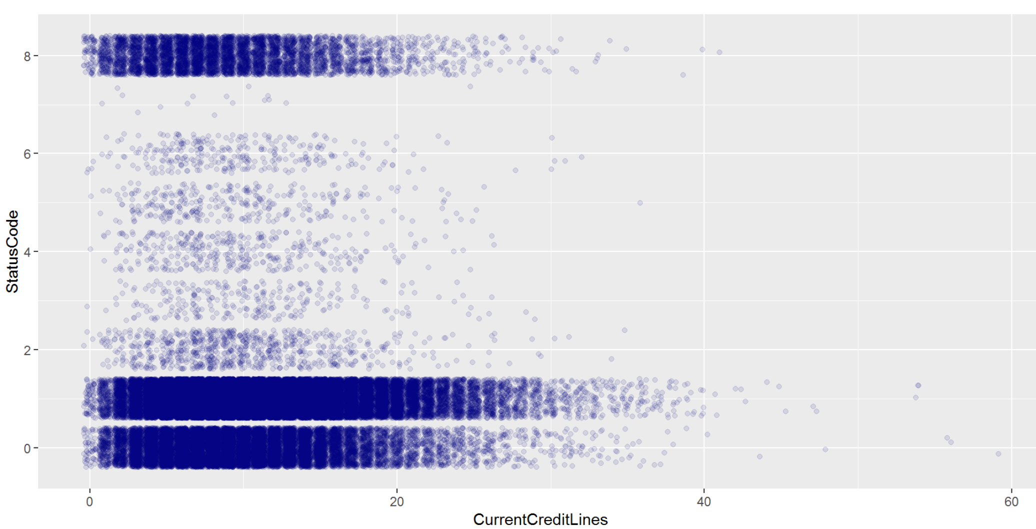
This doesn't give me much information about how the number of Current Delinquencies at the time the profile was pulled affects default status. I want to zoom in to see if their is a number range that might indicate a borrower will default.



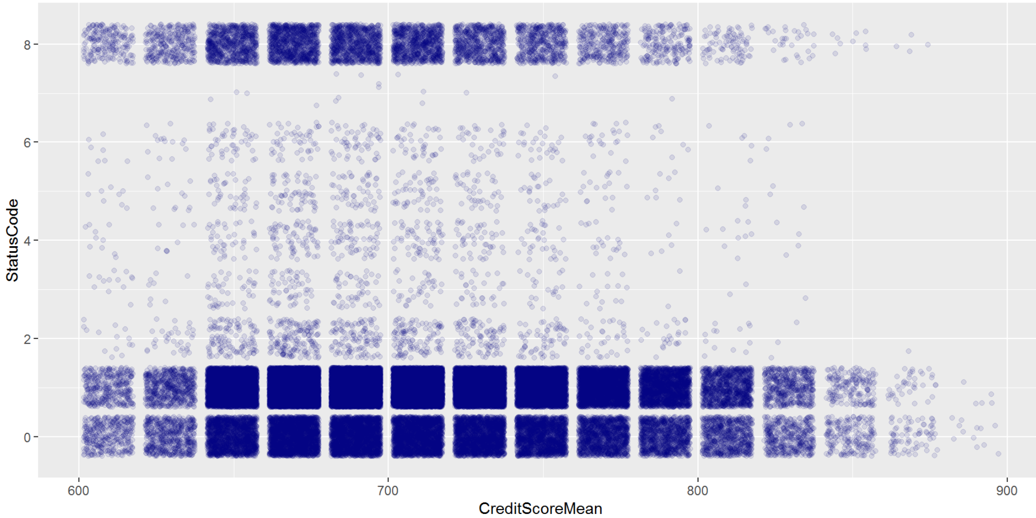

```
##                        StatusCode
## CurrentDelinquencies      0     1     2     3     4     5     6     7     8
##                    0  16802 48033   619   198   281   230   237    11  4841
##                    1   1824  5295   109    36    50    42    44     4   819
##                    2    614  1652    42    11    13    13    10     0   276
##                    3    236   615    13    10     5    11     4     1   144
##                    4    134   355     4     3     4     5     2     0   104
##                    5     60   188     4     3     4     4     3     0    44
##                    6     63   121     7     0     0     4     1     0    36
##                    7     40    97     2     1     1     0     1     0    25
##                    8     28    60     2     1     0     1     1     0    13
##                    9     18    45     1     1     2     0     1     0    11
##                   10     10    35     0     0     1     0     0     0     8
```

From the graph and table at and above 0 Delinquencies the difference between the number of Completed vs Defaulted loans begins to decrease.

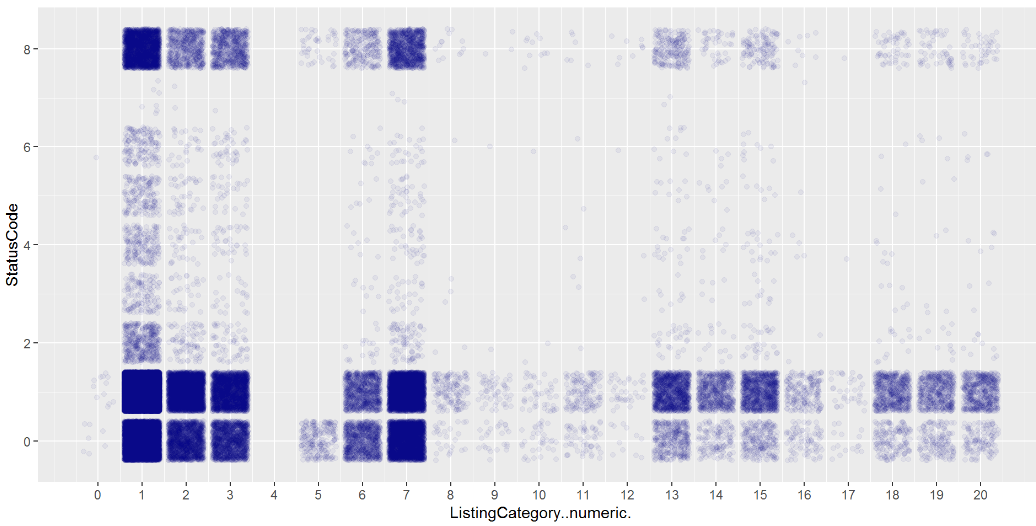
```
##           CurrentCreditLines
## StatusCode    0    1    2    3    4    5    6    7    8    9   10   11
##          0   55  254  500  824 1105 1383 1561 1662 1784 1620 1451 1399
##          1   74  245  630 1114 1768 2992 3652 4867 4993 4973 4765 4407
##          2    3   15   22   35   38   49   47   73   75   78   60   48
##          3    1    1    6    9   12   15   19   25   16   19   18   28
##          4    1    4   15   13   23   20   28   38   26   32   31   18
##          5    1    3    8   21   18   29   28   29   14   21   25   22
##          6    4    4    6    7   14   24   32   31   16   27   28   13
##          7    0    1    2    1    0    1    1    1    1    2    1    1
##          8   80  196  295  352  444  443  518  520  474  468  410  385
##           CurrentCreditLines
## StatusCode   12   13   14   15   16   17   18   19   20   21   22   23
##          0 1150  994  841  668  593  440  343  261  221  155  133  106
##          1 3816 3304 2862 2356 2038 1515 1348 1073  809  643  486  380
##          2   50   36   37   21   31   22   14   10   10    6    5    4
##          3   16   15   17    8    9    3    3    4    5    5    0    1
##          4   23   15   17   14   10   11    6    1    6    3    2    1
##          5   15   13   11   10   11    6    7    2    5    2    0    6
##          6   20   12   14   16    6    6    5    4    4    2    1    2
##          7    2    1    0    0    0    0    0    0    0    0    0    0
##          8  326  291  241  185  167  124  109   73   57   46   27   26
```



Now this is quite interesting. For borrower's with 0 & 1 credit lines a similar number of defaults compared to complete loans. And then as the number of credit lines increases a great deal more completed loans than defaulted. So it seems having existing credit lines is a slight indication that the loan will be completed rather than defaulted.



For borrower's with credit scores near 600 about the same number of borrower's defaulting vs borrower's completing their loans while for those above 600 more do complete their loan but quite a few people with scores between 650 and 750 default.




```
## [1] "Reason x Status Code table below"
```

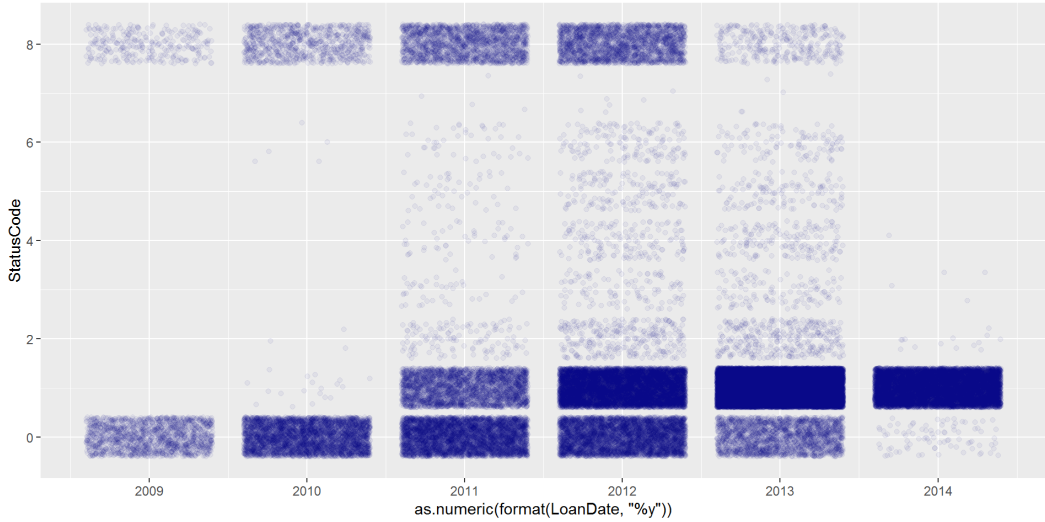
```
##
##        0    1    2    3    4    5    6    7    8
##   0  0.35 0.60 0.00 0.00 0.00 0.00 0.05 0.00 0.00
##   1  0.19 0.74 0.01 0.00 0.00 0.00 0.00 0.00 0.06
##   2  0.30 0.58 0.01 0.00 0.00 0.01 0.00 0.00 0.09
##   3  0.32 0.52 0.01 0.00 0.01 0.01 0.01 0.00 0.13
##   5  0.84 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.16
##   6  0.43 0.44 0.01 0.00 0.00 0.00 0.00 0.00 0.11
##   7  0.40 0.46 0.01 0.00 0.00 0.00 0.00 0.00 0.12
##   8  0.14 0.77 0.01 0.02 0.00 0.00 0.01 0.00 0.07
##   9  0.29 0.65 0.01 0.00 0.00 0.00 0.01 0.00 0.04
##   10 0.37 0.48 0.02 0.00 0.00 0.00 0.01 0.00 0.11
##   11 0.27 0.67 0.00 0.01 0.01 0.00 0.00 0.00 0.03
##   12 0.19 0.61 0.00 0.02 0.02 0.00 0.02 0.00 0.15
##   13 0.19 0.66 0.01 0.00 0.01 0.00 0.01 0.00 0.11
##   14 0.16 0.74 0.02 0.01 0.01 0.01 0.00 0.00 0.05
##   15 0.16 0.71 0.02 0.01 0.01 0.00 0.01 0.00 0.09
##   16 0.29 0.66 0.01 0.00 0.01 0.00 0.00 0.00 0.03
##   17 0.31 0.65 0.00 0.00 0.02 0.00 0.00 0.00 0.02
##   18 0.18 0.72 0.01 0.01 0.01 0.00 0.00 0.00 0.06
##   19 0.20 0.71 0.01 0.00 0.00 0.00 0.01 0.00 0.07
##   20 0.17 0.75 0.01 0.00 0.01 0.00 0.01 0.00 0.06
```

```
##
##          0     1     2     3     4     5     6     7     8
##   0      7    12     0     0     0     0     1     0     0
##   1   9868 39194   457   138   196   172   161     7  2987
##   2   2016  3963    81    30    31    39    27     2   612
##   3   1679  2738    56    21    33    29    40     1   701
##   5    231     0     0     0     0     0     0     0    43
##   6    964   975    18     8     7     6    10     0   249
##   7   3645  4218    86    26    34    38    25     3  1143
```

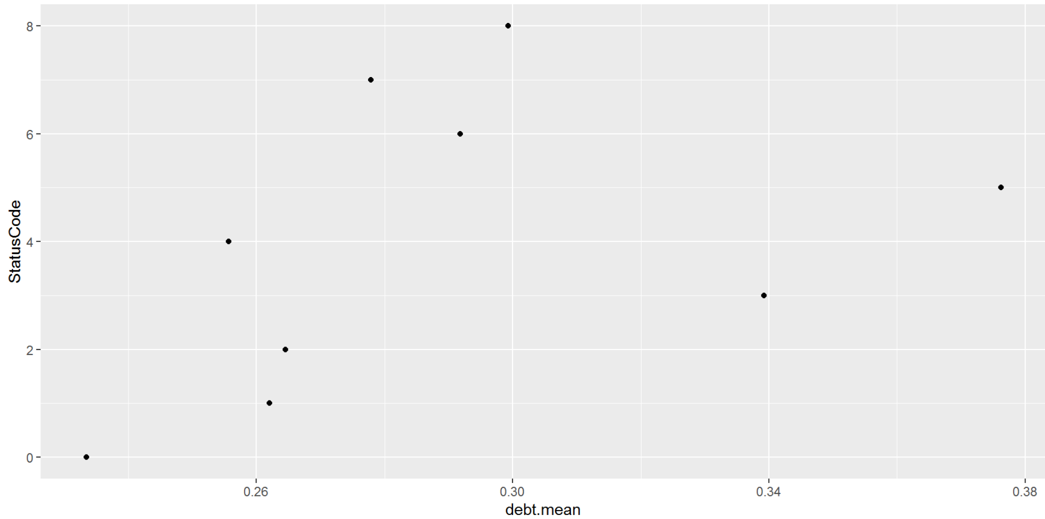
```
##    8     27   153     2     3     0     0     1     0    13
##    9     25    55     1     0     0     0     1     0     3
##    10    34    44     2     0     0     0     1     0    10
##    11    59   145     1     2     2     1     1     0     6
##    12    11    36     0     1     1     0     1     0     9
##    13   377  1321    26     8    17     9    10     2   226
##    14   141   652    14     5     7     7     4     0    46
##    15   240  1078    28    11    17     7    10     0   131
##    16    87   201     2     0     4     1     0     1     8
##    17    16    34     0     0     1     0     0     0     1
##    18   163   639    13     6     6     2     2     0    54
##    19   151   543    10     3     2     2     4     0    53
##    20   128   575     9     3     5     0     5     0    46
```

The most common reason code is 1(Loan Consolidation) yet uthas a lower percentage of defaulted loans than reason code 5 which has 16% of it's loans in a default status. I'm probably not going to pursue whether or not Listing Category affects loan status any further as this self select category and the borrower may not be giving the true purpose of the loan.

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.
## "2009-07-20" "2012-02-23" "2013-04-09" "2012-11-15" "2013-11-05"
##         Max.
## "2014-03-12"
```



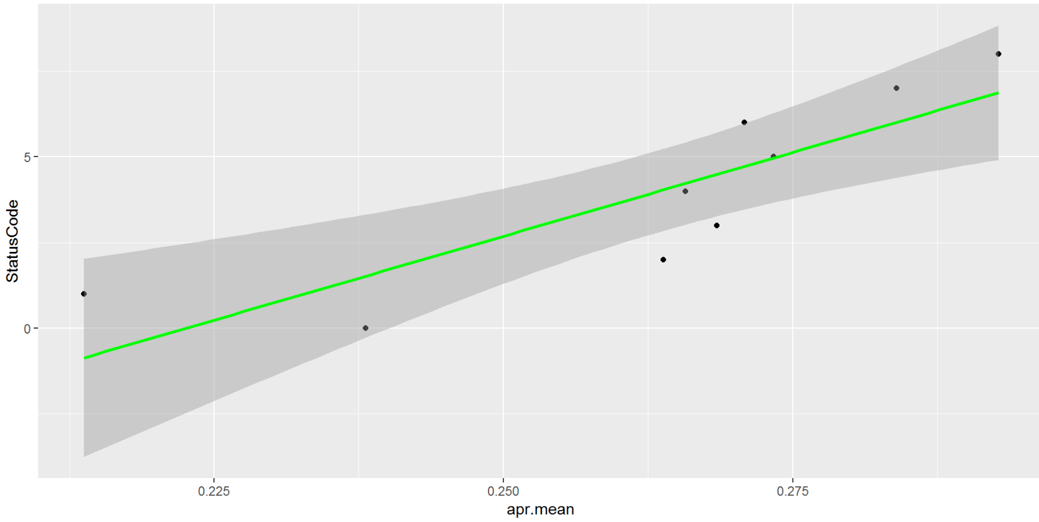
Loans with an origination date of 2011 and later could still have been in progress as of 2014(as indicated by loan status 1-7) so the final outcome of the loan(i.e. Completed or Defaulted) is unknown. I am not going to do further analysis on year and Loan Status I just wanted to see the yearly trend for completion vs. default.




```
## [1] "Correlation between Status Code and Mean Debt"
```

```
## [1] 0.4123365
```

Status Code vs. DebtToIncomeRatio Mean graph shows reasonable scatter from which a linear model could be created & positive correlation of .41 indicating a slight positive linear relationship



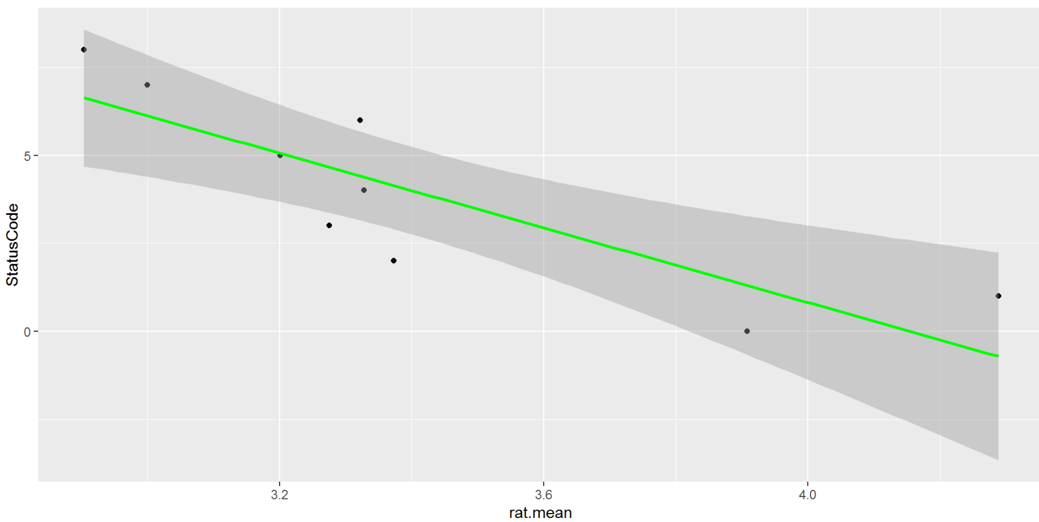

```
## [1] "Correlation betweeen Status Code and APR mean"
```

```
## [1] 0.8554174
```

```
## [1] "Mean Borrower's APR"
```

```
## [1] 0.2266582
```

Since our univariate analysis showed that the Borrower's APR was fairly symmetrical with a single peak near the center of the data it is realistic to use the mean of .23 as the center of the distribution. That combined with the fact that Status Code and the APR's mean have a strong correlation with r = .86 it appears that, for this data, the higher the borrower's APR is above the mean of .23 the more likely the borrower had late payments or defaulted.



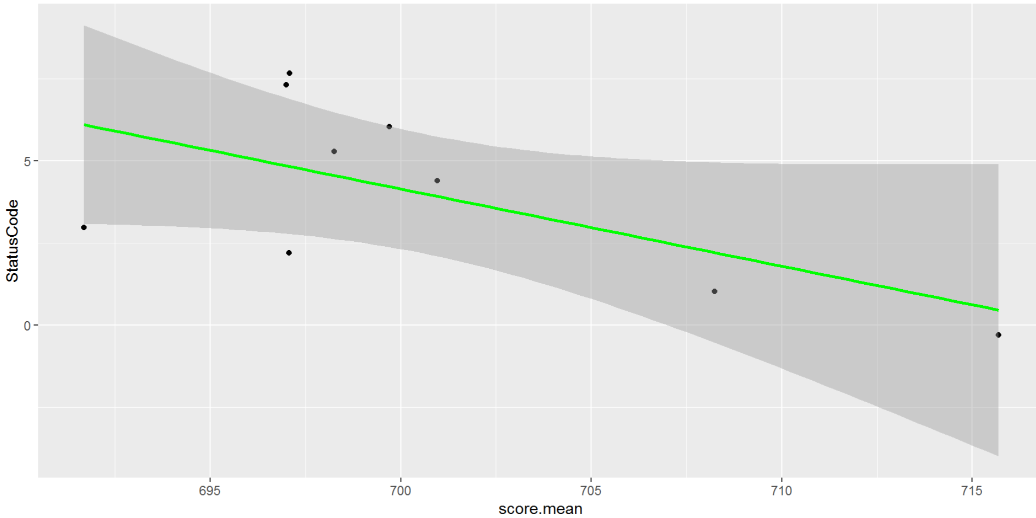

```
## [1] "Corrrelation betweeen Status Code and Prosper Rating Mean"
```

```
## [1] -0.8436804
```

```
## [1] "Mean Prosper Rating"
```

```
## [1] 4.072243
```

The relationship between Prosper Rating mean vs Status code is strong and negative (-.84). In other words the lower the prosper rating the more likely borrower's had loans that were in a late or defaulted status.




```
## [1] "Corrrelation betweeen Status Code and Credit Score Mean"
```

```
## [1] -0.6143982
```

```
## [1] "Mean Credit Score"
```

```
## [1] 708.8902
```

The relationship between Credit Score Mean vs Status Code is moderately strong and negative with r = -.61. Credit Score used in conjunction with the Prosper Rating could help an investor select loans most likely to end up closed rather than defaulted. Most points are clustered around 698.

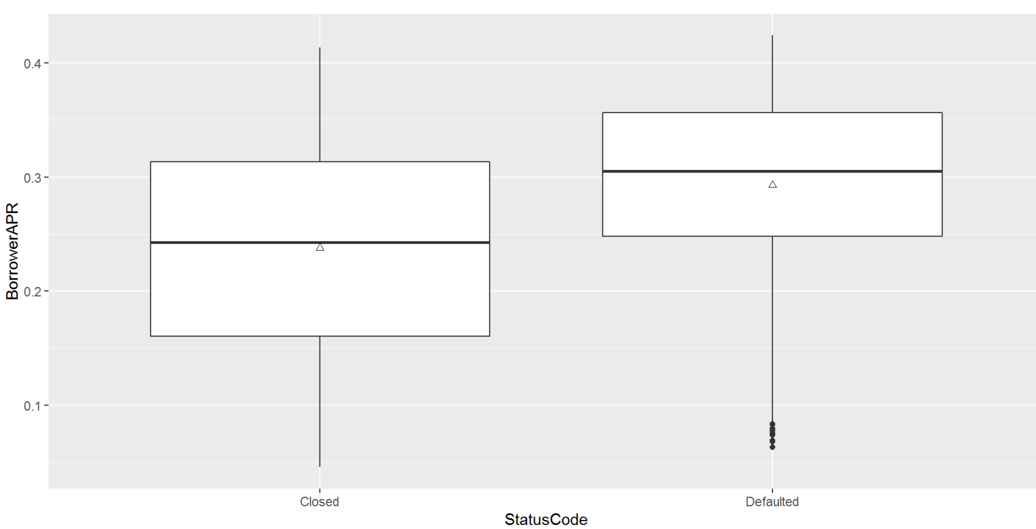
```
dim(spd.sts)
```

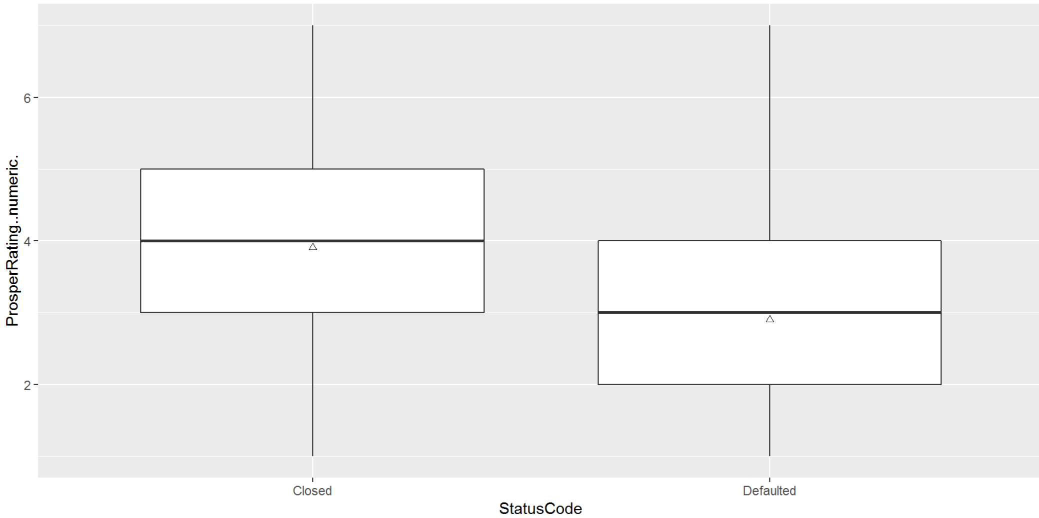
```
## [1] 26210    19
```

```
table(spd.sts$StatusCode)
```

```
## 
##    Closed Defaulted 
##     19869      6341
```

For the rest of the analysis I'm only interested in Status Code 0 & 8. What will happen with Statuscodes 1-7 is unknown as they are in progress. This means I am now treating Status Code as a qualitative variable. Reducing the data set gives me 26,210 records for analysis. 6,341 of theses are in a default status.

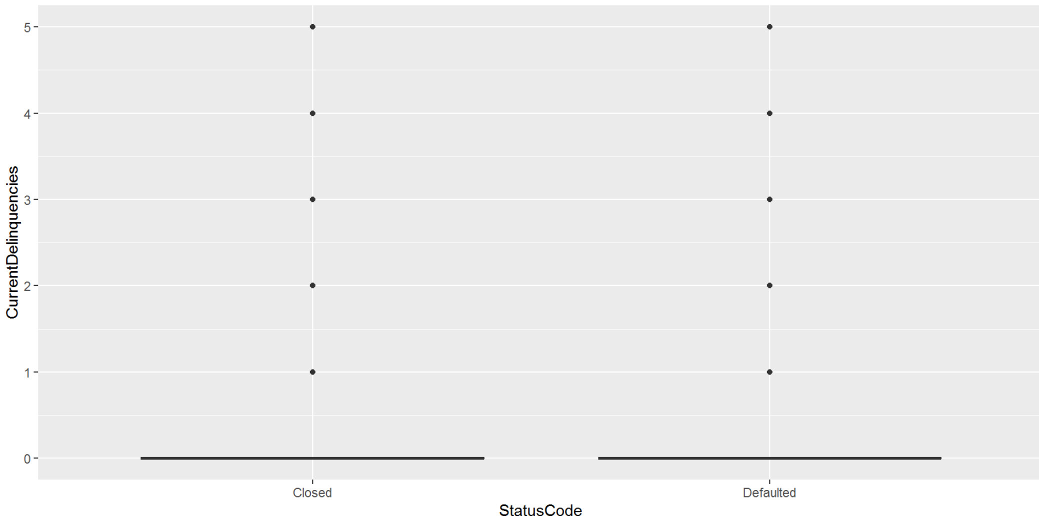




The above two box plots examine APR and Prosper Rating for closed and defaulted loans.

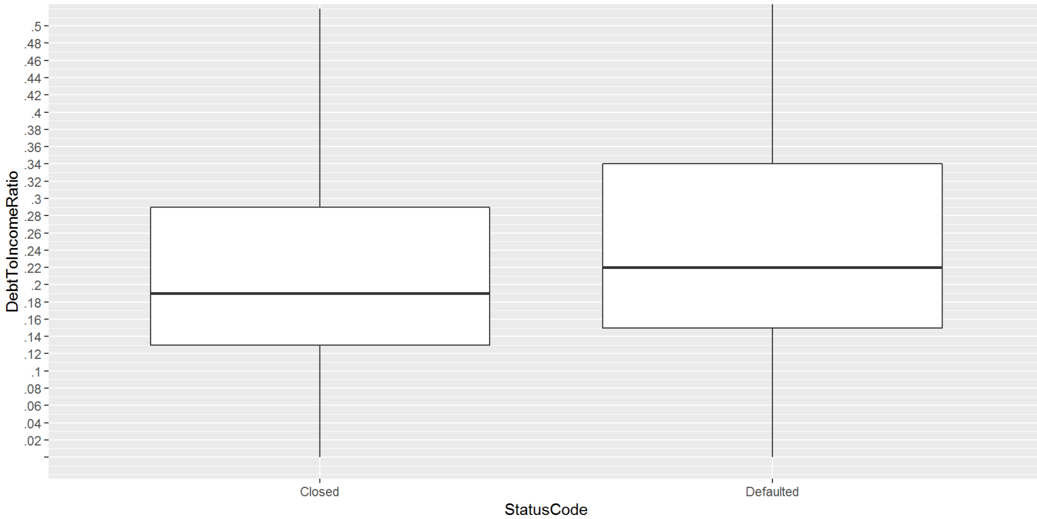
The borrower's APR 50th percentile is approximately 25% for completed loans while it is at about 30% for defaulted loans. Note the triangle on the graphs represents the mean of each. Since these medians and means are different this could be an indication that knowing a borrower's APR in addition to Prosper Rating could be useful.

Similarly the Prosper Rating's 50th percentile is four for completed loans while for defaulted loans it is three and 75% of borrower's who default on a loan have a Prosper Rating of four or less.

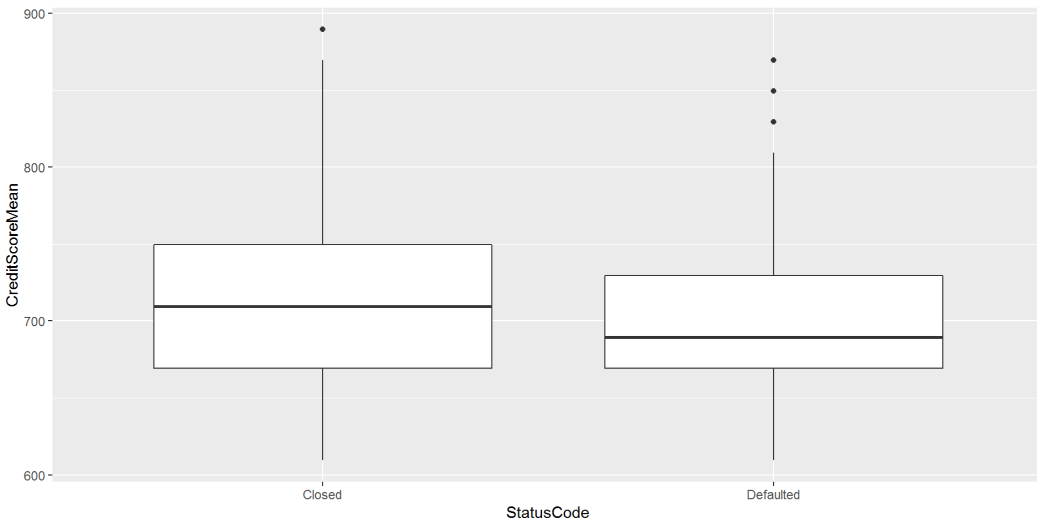



```
##            CurrentDelinquencies
## StatusCode     0    1    2    3    4    5    6    7    8    9   10
##   Closed    0.78 0.69 0.69 0.62 0.56 0.58 0.64 0.62 0.68 0.62 0.56
##   Defaulted 0.22 0.31 0.31 0.38 0.44 0.42 0.36 0.38 0.32 0.38 0.44
```

The distributions of Current Delinquencies are quite similar for completed and defaulted loans. They both share the same 75th percentile of 0. Outliers exist at 1 & up so I decided to focus on them. Borrower's with one or more delinquencies have a greater percent of defaults compared to those with 0 delinquencies. For example for borrower's with just 1 delinquency the default rate is about 31% but for those with 0 delienquencies the default rate is about 22%.



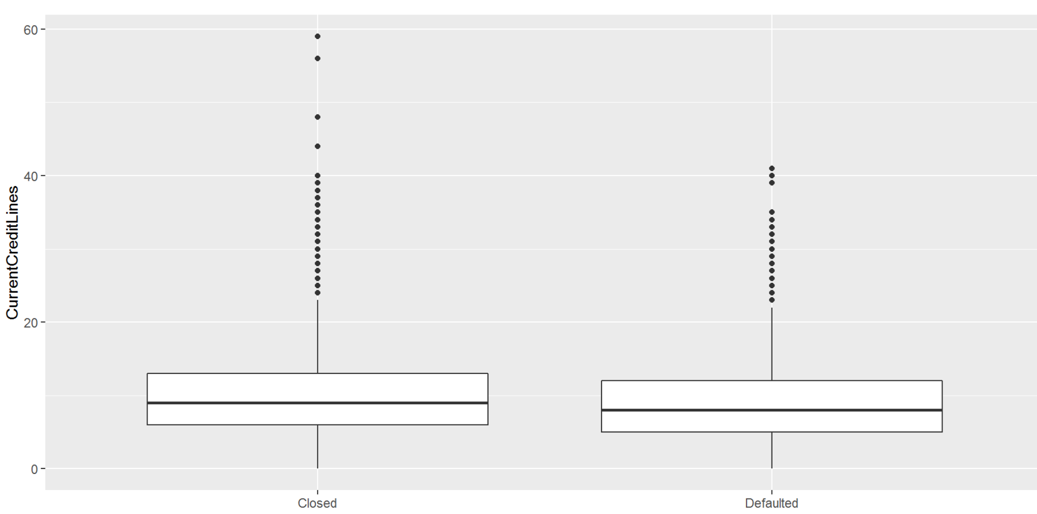
Outliers with values between .5 and 10 exist for both closed and defaulted loans but don't occur with high frequency. I am zooming in on debt to income ratios between 0 and .5 For loans in a closed status 75% of borrowers had a Debt to Income Ratio of .29 or less while for loans in a Defaulted Status 75% of borrowers had a Debt to Income Ratio of .34. The difference between medians of the two status is only 5% but it does appear that having a lower debt to Income Ratio may be related to Status.

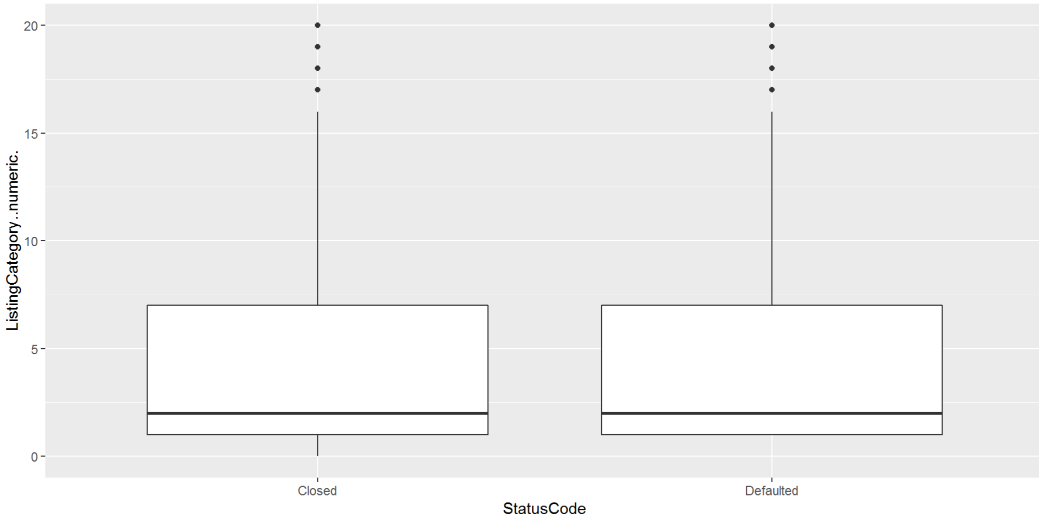


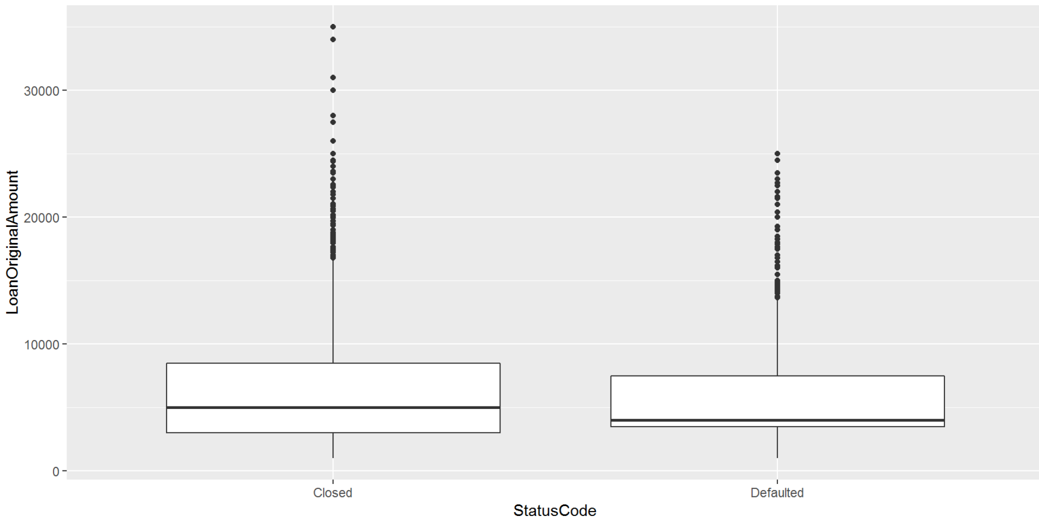

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   609.5   669.5   709.5   715.7   749.5   889.5
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   609.5   669.5   689.5   697.1   729.5   869.5
```

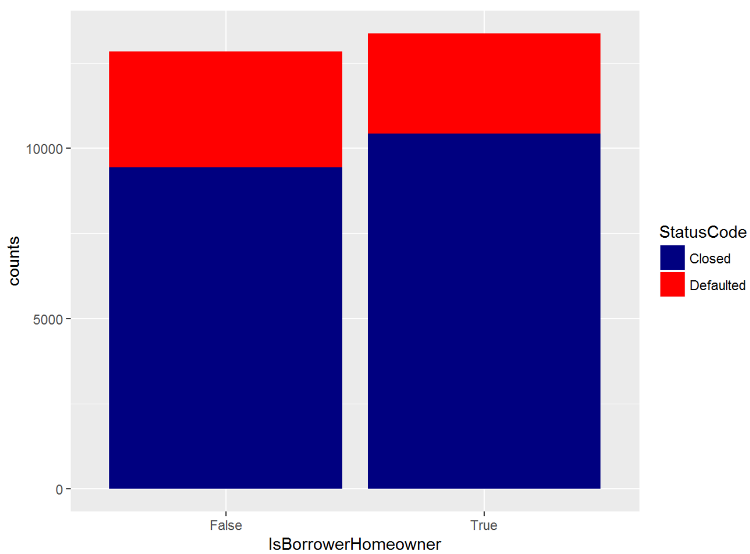
75% of borrowers with closed loans had a mean credit score of 750 or less while for defaulted loans the 75th percentile is about 730.
While a difference of 20 points does not seems important perhaps it is large enough to indicate a relationship between Credit Score Mean and Loan Status.



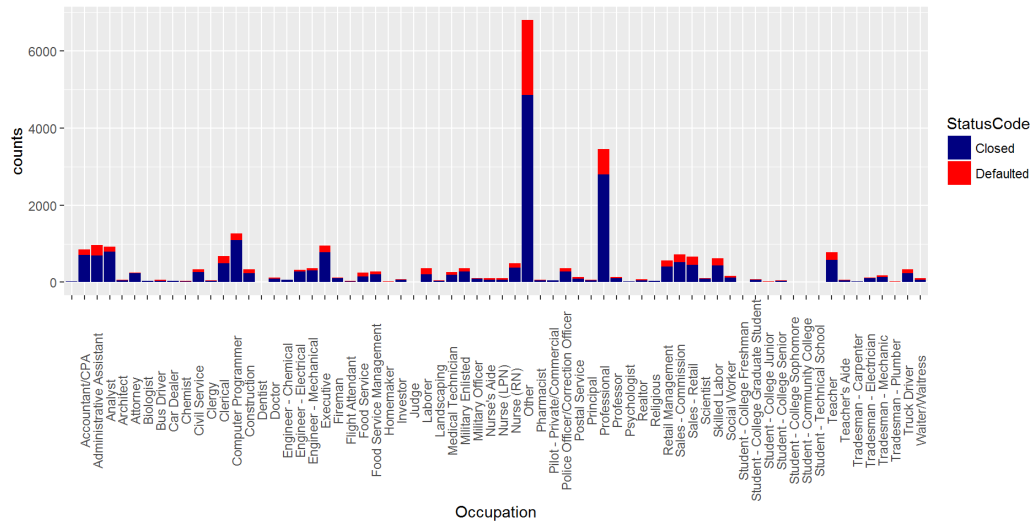




Current Credit Lines, Reason(ie. Listing Category), and Original Loan Amount have similar distributions for both completed and defaulted loans so it does not appear that their is a relationship between them and default status.



Whether or not someone is a homeowner does not seem to impact whether or not they will default on a loan. Approximately the same number of homeowners and non homeowners have loans in closed status and while non homeowners may have defaulted on loans in greater number it is not by that much as depicted in the graph.



Again this is a self selected category and no Occupation stands out more than another as having more or less defaults compared to closed. It doesn't appear a relationship exists between Status Code and Occupation. # Bivariate Analysis

Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

slight differences in Debt to Income Ratio, Credit Score, and number of delinquencies for closed vs. defaulted loans exist. I was hoping to find a more dramatic relationship but nothing is standing out yet. Perhaps a variable will stand out more in the multivariate analysis.

Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

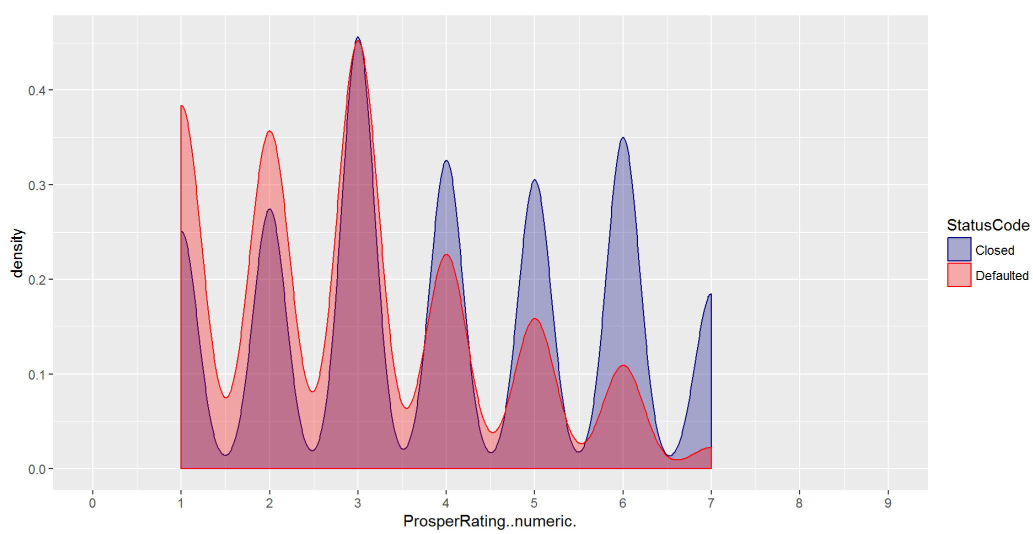
During this analysis it became clear that I could not treat Loan Status as a quantitative variable and decided to just focus on closed vs defaulted status. What is most surprising is the seemingly lack of a strong relationship. For example the distribution between home owners/non owners and whether a person defaulted on a loan is about the same. I started this project believing homeowers would be more likely to complete loans than non-homeowners which was an incorrect assumption.

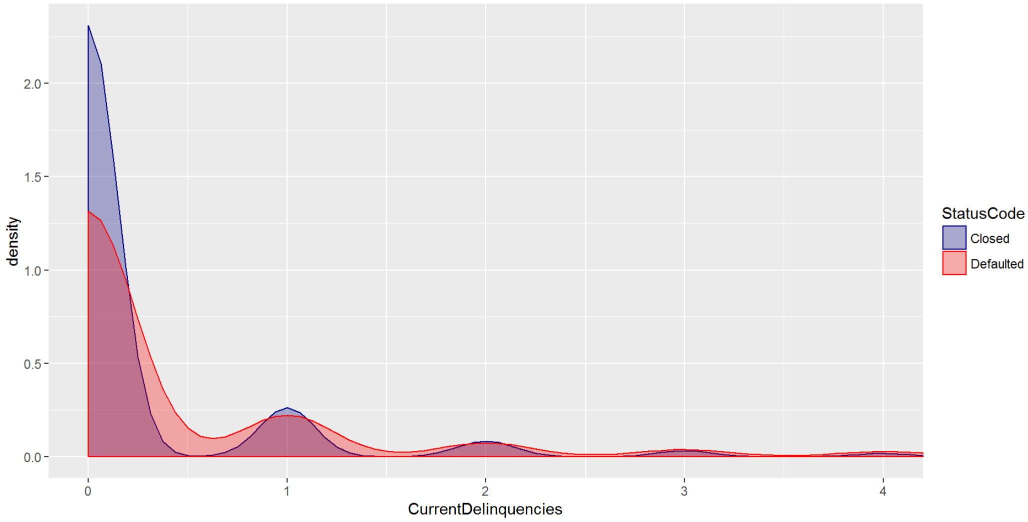
What was the strongest relationship you found?

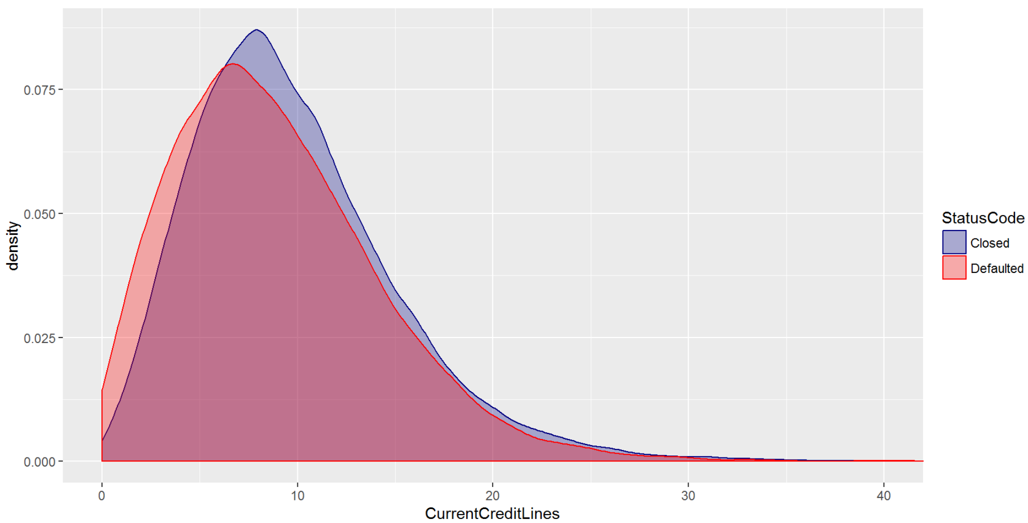
The strongest linear relationship is between Prosper Rating Mean vs. Status Code but a linear model here just doesn't seem appropriate. Treating Status code as a qualitative(Closed vs Default) variable I am seeing a strong relationship between Status Code vs Borrower's APR as indicated by the box plots and scatterplot of Status code vs. Borrower's APR mean.

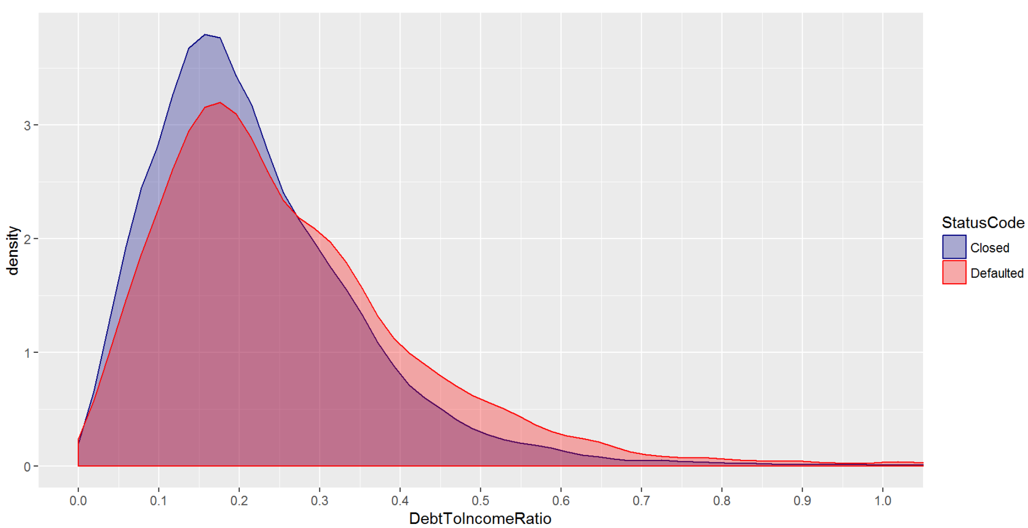
Multivariate Plots Section

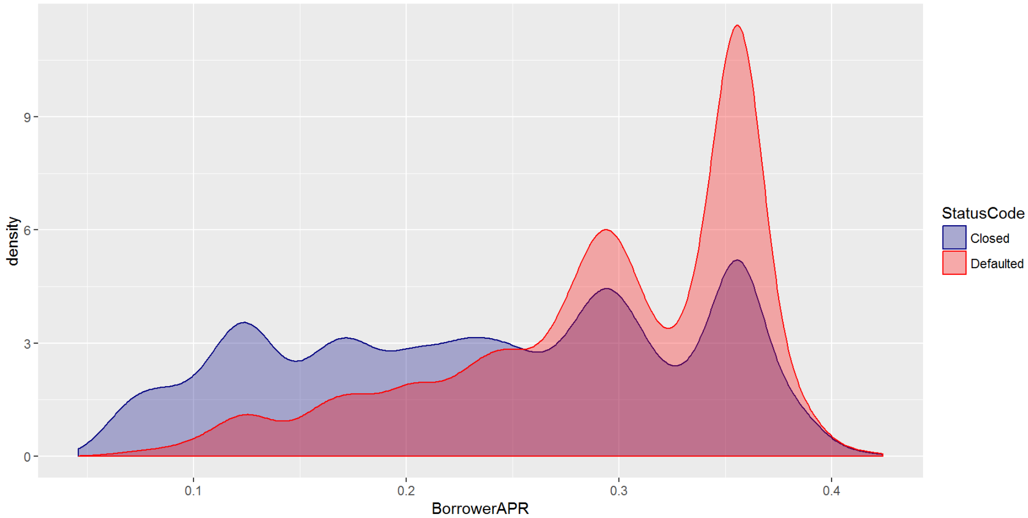
For this section I have decided to just work with the dataframe containing status code 0 (Loan in Closed Status or pending closure) and status code 8 (Defaulted). Also please note that numeric Prosper Ratings range from 1(Higher Risk - "HR") to 7(Lower Risk - "AA")

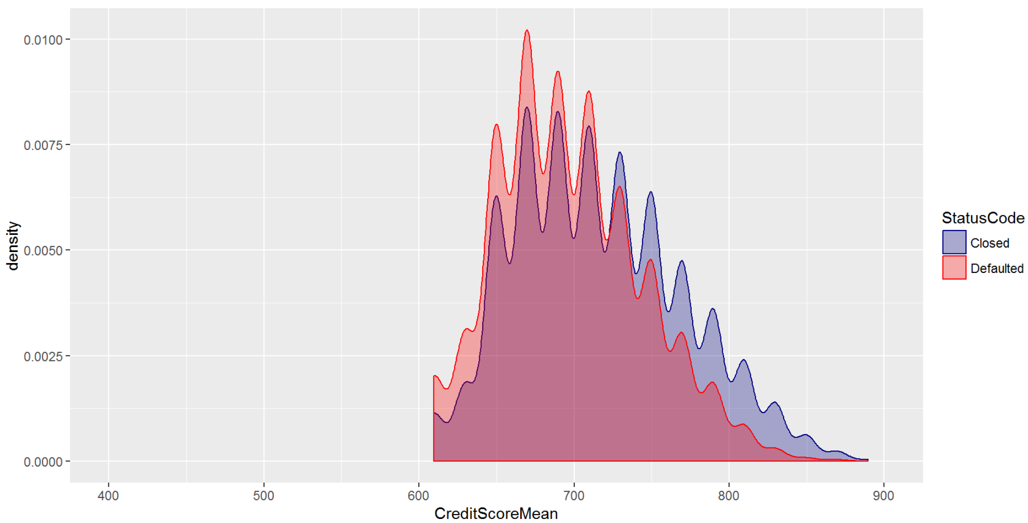



density
2.0
1.5
1.0
0.5
0.0
0
1
2
3
4
CurrentDelinquencies
StatusCode
Closed
Defaulted


density
0.075
0.050
0.025
0.000
0
10
20
30
40
CurrentCreditLines
StatusCode
Closed
Defaulted


density
3
2
1
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
DebtToIncomeRatio
StatusCode
Closed
Defaulted





The Prosper Rating density graph makes it very clear that investing in a loan that has been rated below 4 is a risky investment but I want to see what else might help us predict the likelihood of a borrower defaulting.

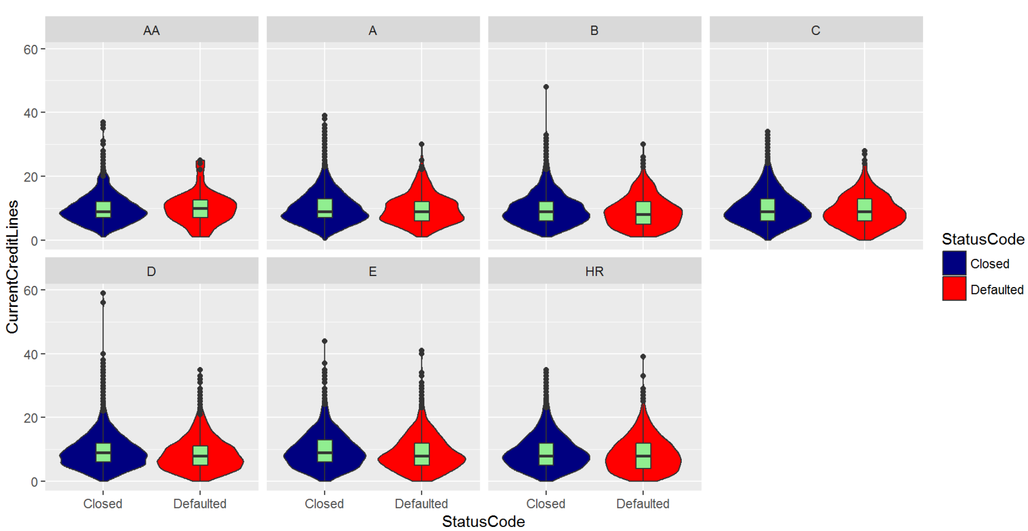
For borrowers with zero Delinquencies, the percent of closed loans is greater than defaulted but as soon as 1 or more delinquencies are found, loans in default status are almost equal to those in closed.

Borrowers with 0 to 5 credit lines actually have more loans in a default status and those with 5 to about 18 have more loans in a close status but these densities are similar and have significant overlap.

The distributions of Debt to Income Ratio's for Closed and Defaulted loans also have very similar shapes. Borrower's with a Debt To Income Ratio between approximately 5% and 25% have a greater number of loans in closed status and then once the Debt To Income Ratio exceeds 25% the occurrence of defaulted loans is greater than closed loans.

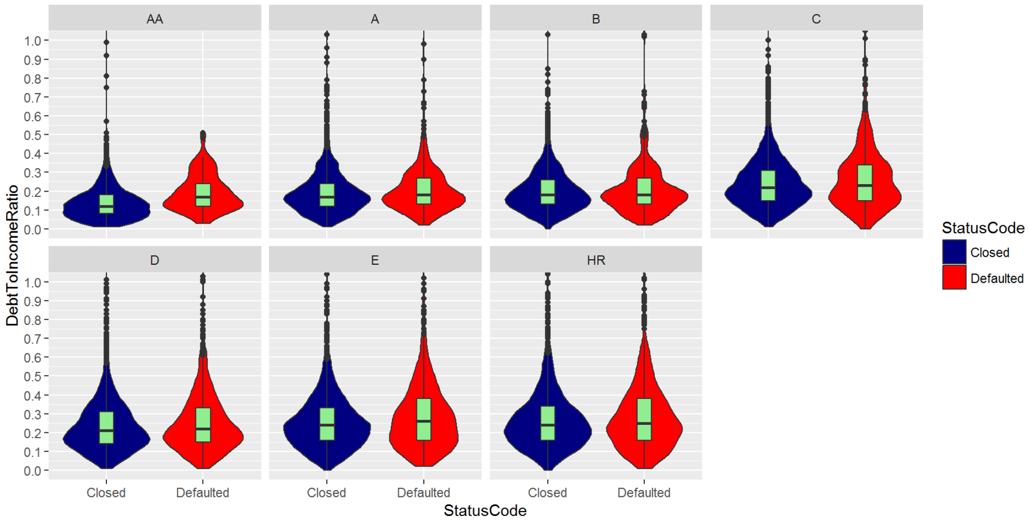
An APR of 25% seems to be a tipping point for the likelihood of a loan being in default status. Below 25% more loans are closed than defaulted. Above 25% and defaults are more common.

The last density indicates borrowers with a credit score less than approximately 710 had a greater percentage of loans in a default status



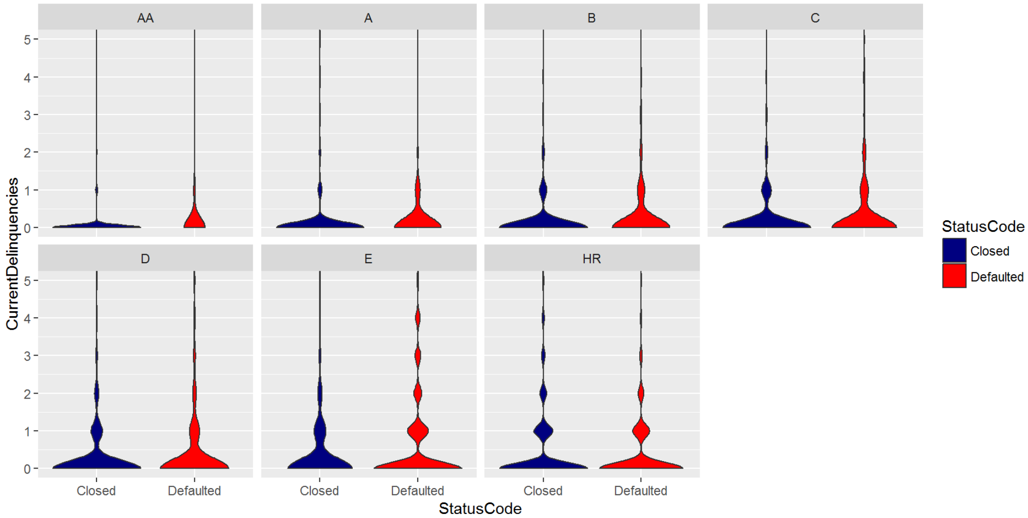
```
##    Rating StatusCode Min Q1 Median   Mean   Q3 Max
## 1      AA     Closed   1  7      9  9.871 12.0  37
## 2      AA  Defaulted   1  7     10 10.080 12.5  25
## 3       A     Closed   0  7      9 10.180 13.0  39
## 4       A  Defaulted   1  6      9  9.849 12.0  30
## 5       B     Closed   1  6      9  9.784 12.0  48
## 6       B  Defaulted   1  5      8  8.981 12.0  30
## 7       C     Closed   0  6      9 10.210 13.0  34
## 8       C  Defaulted   0  6      9  9.238 13.0  28
## 9       D     Closed   0  6      9  9.417 12.0  59
## 10      D  Defaulted   0  5      8  8.603 11.0  35
## 11      E     Closed   0  6      9  9.731 13.0  44
## 12      E  Defaulted   0  5      8  9.071 12.0  41
## 13     HR     Closed   0  5      8  9.411 12.0  35
## 14     HR  Defaulted   0  4      8  8.545 12.0  39
```

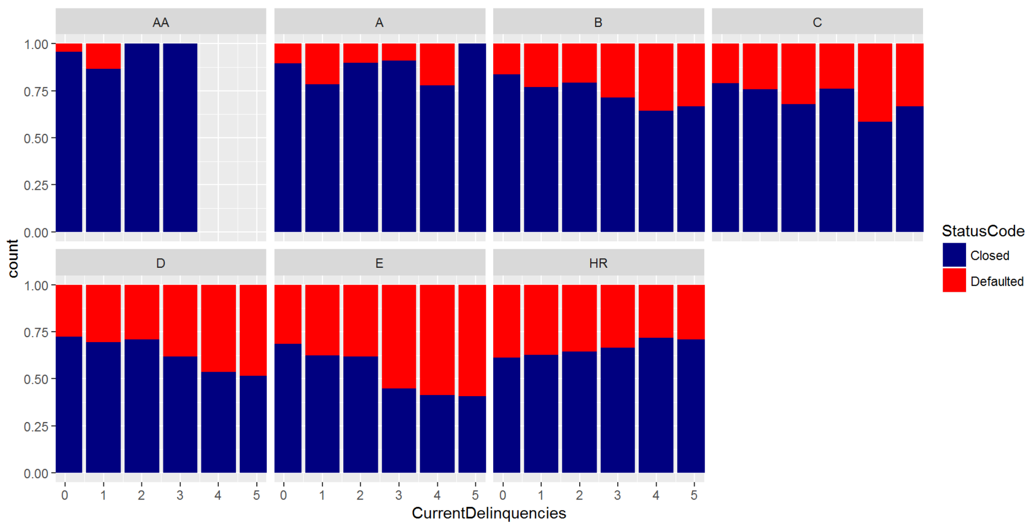
Since the above distributions are so similar knowing the number of current credit lines on top of Prosper Rating does not give us any more information about whether or not the borrower will have a loan in a default status




```
##    Rating StatusCode  Min   Q1 Median   Mean   Q3   Max
## 1      AA     Closed 0.01 0.08   0.12 0.1453 0.18 10.01
## 2      AA  Defaulted 0.03 0.12   0.17 0.1940 0.24  0.51
## 3       A     Closed 0.01 0.12   0.17 0.1957 0.24  5.64
## 4       A  Defaulted 0.02 0.13   0.18 0.2263 0.27  2.79
## 5       B     Closed 0.00 0.13   0.18 0.2094 0.26 10.01
## 6       B  Defaulted 0.02 0.13   0.18 0.2125 0.27  1.32
## 7       C     Closed 0.01 0.15   0.22 0.2466 0.31  5.56
## 8       C  Defaulted 0.00 0.15   0.23 0.3093 0.34 10.01
## 9       D     Closed 0.01 0.14   0.21 0.2421 0.31 10.01
## 10      D  Defaulted 0.01 0.15   0.22 0.2870 0.33 10.01
## 11      E     Closed 0.00 0.16   0.24 0.2699 0.33  4.89
## 12      E  Defaulted 0.02 0.16   0.26 0.3353 0.38 10.01
## 13     HR     Closed 0.00 0.16   0.24 0.3271 0.34 10.01
## 14     HR  Defaulted 0.01 0.16   0.25 0.3469 0.38 10.01
##    DebtToIncomeRatio.NA's
## 1                      65
## 2                       6
## 3                     186
## 4                      30
## 5                     158
## 6                      47
## 7                     260
## 8                     117
## 9                     500
## 10                    276
## 11                    388
## 12                    234
## 13                    420
## 14                    301
```

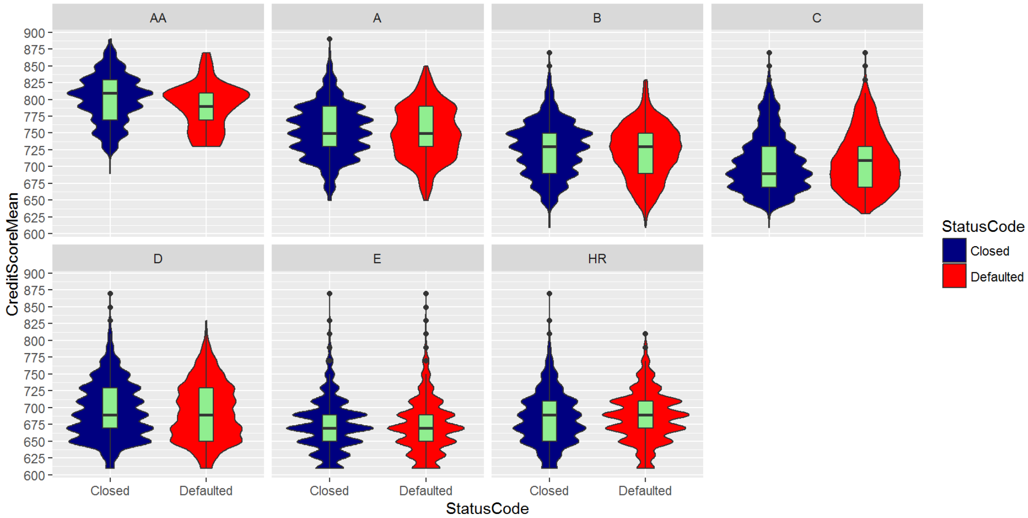
Debt To Income ratio gives a bit more information. It appears not all loans given “AA” ratings are equivalent. The medians differ by 5% and the 3rd quartiles by 6%. With the exception of rating “AA” the medians are quite similar for “Closed” & “Defaulted” in each of the Prosper Ratings. A bit more difference exists between the 3rd Quartiles(75th Percentile) for “EE”, and “HR”. This indicates to me that knowing the DebtToIncome ratio of borrower and the median Debt to Income ratio of all closed loans for each Rating would be very useful.





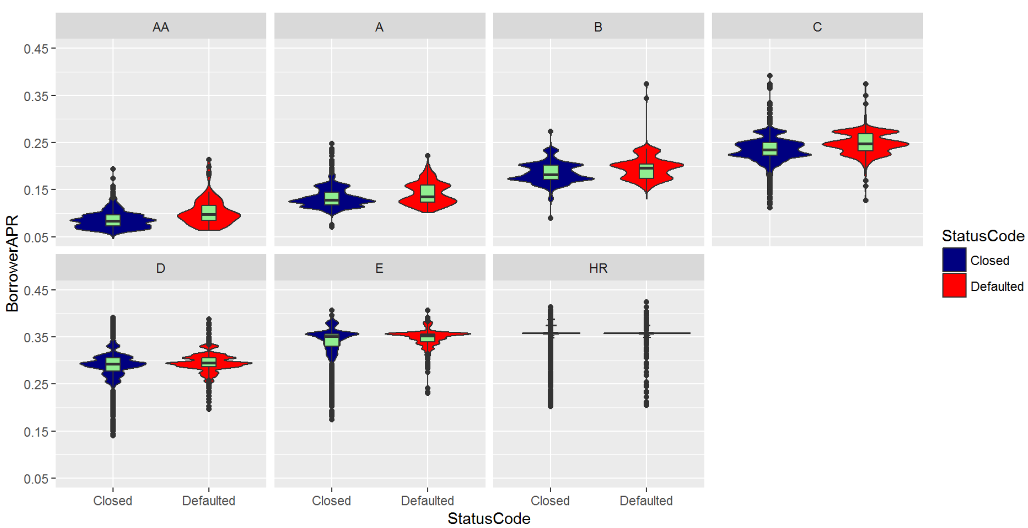

```
##    Rating StatusCode Min Q1 Median    Mean Q3 Max
## 1      AA     Closed   0  0      0 0.03736  0  10
## 2      AA  Defaulted   0  0      0 0.14460  0   6
## 3       A     Closed   0  0      0 0.13100  0  21
## 4       A  Defaulted   0  0      0 0.21230  0  12
## 5       B     Closed   0  0      0 0.23020  0  15
## 6       B  Defaulted   0  0      0 0.33840  0   9
## 7       C     Closed   0  0      0 0.27990  0  21
## 8       C  Defaulted   0  0      0 0.37620  0  13
## 9       D     Closed   0  0      0 0.38550  0  21
## 10      D  Defaulted   0  0      0 0.48060  0  14
## 11      E     Closed   0  0      0 0.50320  0  21
## 12      E  Defaulted   0  0      0 0.85880  1  32
## 13     HR     Closed   0  0      0 0.60770  1  22
## 14     HR  Defaulted   0  0      0 0.59130  1  15
```

Because number of delinquencies are clustered around 0 & 1 I wanted to see a barchart in addition to the boxplots. With the exception of the “HR” rating, borrower's with 0 delinquencies have more closed loans than defaulted across ratings. As soon as 1 delinquency is found the percentage in a default status exceeds those in closed. Just having access to whether or not a borrower has any delinquencies exist may be useful.



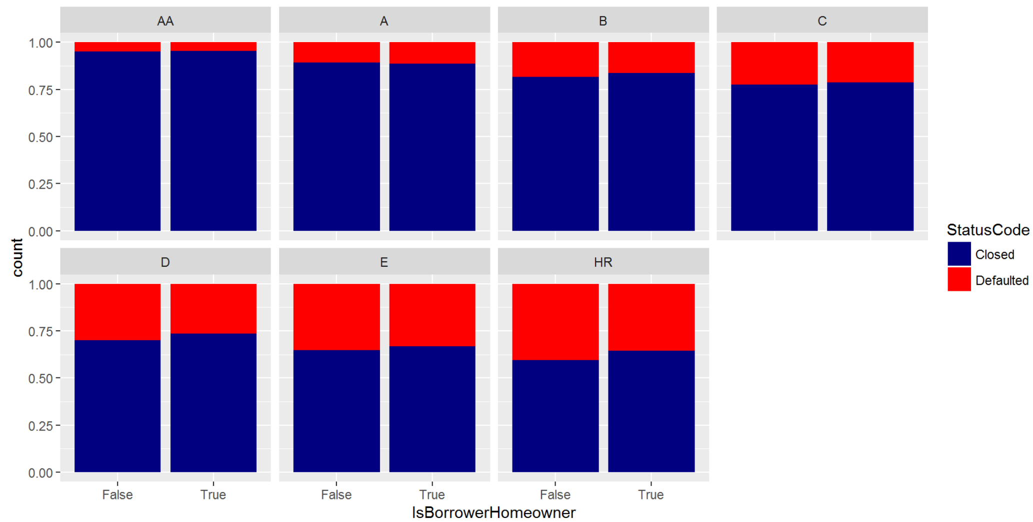

```
##    Rating StatusCode   Min    Q1 Median  Mean    Q3   Max
## 1      AA     Closed 689.5 769.5  809.5 800.9 829.5 889.5
## 2      AA  Defaulted 729.5 769.5  789.5 790.0 809.5 869.5
## 3       A     Closed 649.5 729.5  749.5 753.4 789.5 889.5
## 4       A  Defaulted 649.5 729.5  749.5 750.4 789.5 849.5
## 5       B     Closed 609.5 689.5  729.5 726.4 749.5 869.5
## 6       B  Defaulted 609.5 689.5  729.5 722.0 749.5 829.5
## 7       C     Closed 609.5 669.5  689.5 706.1 729.5 869.5
## 8       C  Defaulted 629.5 669.5  709.5 709.3 729.5 869.5
## 9       D     Closed 609.5 669.5  689.5 694.9 729.5 869.5
## 10      D  Defaulted 609.5 649.5  689.5 694.0 729.5 829.5
## 11      E     Closed 609.5 649.5  669.5 673.8 689.5 869.5
## 12      E  Defaulted 609.5 649.5  669.5 672.2 689.5 869.5
## 13     HR     Closed 609.5 649.5  689.5 683.1 709.5 869.5
## 14     HR  Defaulted 609.5 669.5  689.5 685.8 709.5 809.5
```

This one shows for “AA” ratings median score differs by about 20 points with closed loans being higher, yet for “C” Prosper Ratings the median Credit score is actually higher for loans in a default status. For the other Prosper Ratings the distributions for closed & defaulted distributions are fairly similar.




```
##    Rating StatusCode     Min      Q1  Median    Mean      Q3    Max
## 1      AA     Closed 0.04583 0.07339 0.08341 0.08569 0.09643 0.1936
## 2      AA  Defaulted 0.06327 0.08466 0.09736 0.10310 0.11670 0.2137
## 3       A     Closed 0.07045 0.11770 0.12780 0.13270 0.14470 0.2481
## 4       A  Defaulted 0.10080 0.12400 0.13520 0.14160 0.15940 0.2224
## 5       B     Closed 0.08999 0.17160 0.18190 0.18530 0.20200 0.2731
## 6       B  Defaulted 0.13110 0.17360 0.19650 0.19570 0.20490 0.3745
## 7       C     Closed 0.11160 0.22280 0.23510 0.23550 0.24970 0.3915
## 8       C  Defaulted 0.12720 0.23250 0.24760 0.24760 0.26890 0.3745
## 9       D     Closed 0.14060 0.27770 0.29260 0.29010 0.30530 0.3915
## 10      D  Defaulted 0.19690 0.28700 0.29510 0.29540 0.30530 0.3872
## 11      E     Closed 0.17430 0.33040 0.35090 0.34290 0.35650 0.4068
## 12      E  Defaulted 0.23120 0.33970 0.35240 0.34860 0.35640 0.4068
## 13     HR     Closed 0.20260 0.35640 0.35800 0.35610 0.35800 0.4136
## 14     HR  Defaulted 0.20500 0.35640 0.35800 0.35860 0.35800 0.4240
```

The plot that stands out the most is for “AA”. A greater percentage of loans are in a default status when the APR rate is higher, even by just a percentage point. The median for those in a default status is about 9.7% while for those in closed is about 8.3%. The 3rd quartiles differ by 2% Having access to the borrower's APR and the median APR rate for closed loans as a reference point would be useful for investors.



Once again it does not appear that being a homeowner influences status of the loan.

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?

Certainly taking into account the Prosper Rating in conjunction with the loan status of closed or defaulted helped to better identify variables which may help predict whether or not a borrower will default on a given loan. Besides the risk rating assigned by Prosper also knowing a borrower's Mean Credit Score, Debt To Income Ratio, and whether or not the borrower has any delinquencies reported would be helpful to investor's when trying to determine which loans they wish to fund in order to gain the most profit. Of course for these values to be useful investor's would also need the corresponding information for all closed loans to use as a reference point.

Were there any interesting or surprising interactions between features?

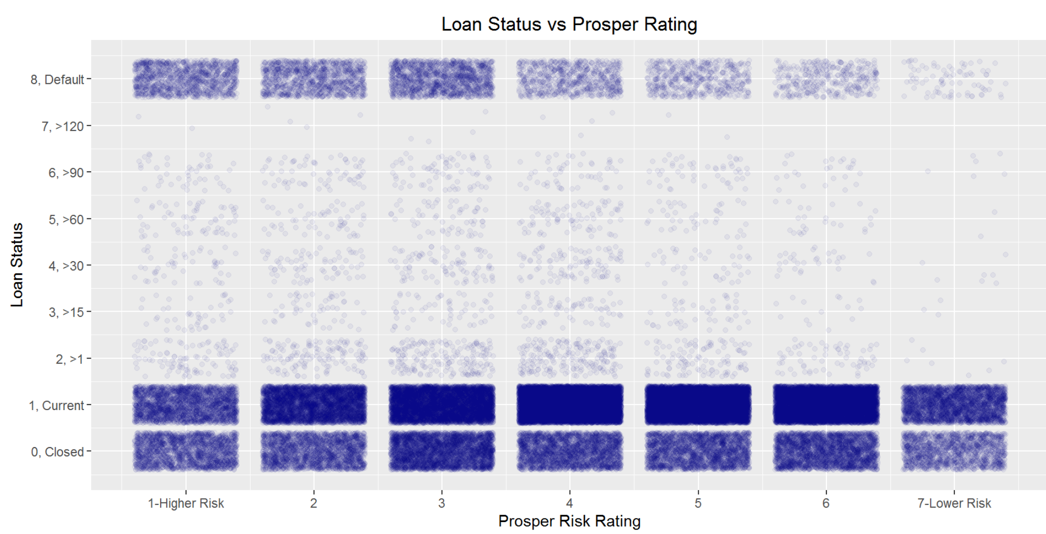
The density graph of number of delinquencies hilited that even if a person has only one delinquency it can affect the ultimate loan status. I was surprised to see when faceting by Prosper risk rating that their rating system seems to reflect this. While prosper rating does seem fairly accurate knowing a bit more of the details and paying attention to seemingly small differences in financial information an investor may increase their chances of selecting a loan that is more likely to result in closure.

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

I did not create any models.

Final Plots and Summary

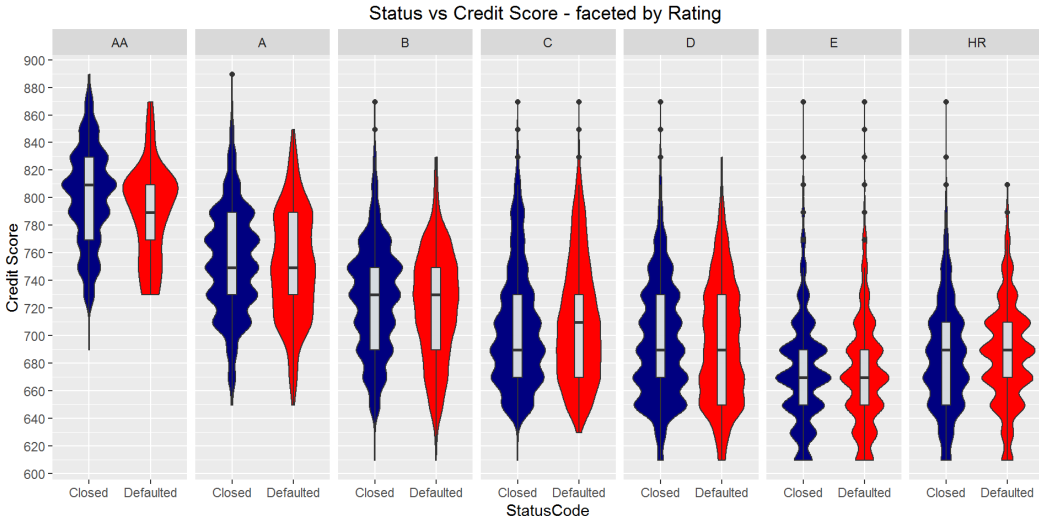
Plot One



Description One

While Prosper Risk Rating and Loan status are not linearly correlated the number of observations decreases overall as the rating increases from higher risk(0) to lower risk(7). Since lower risk ratings with defaults and late payments exist I tried to identify other features that may help predict whether or not a loan will be defaulted on.

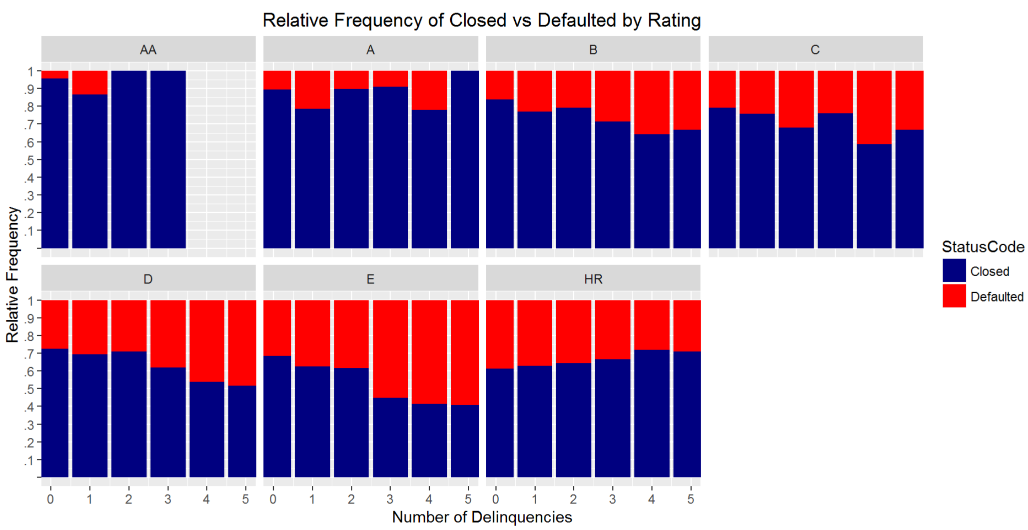
Plot Two



Description Two

The rating for which knowing the borrower's Credit Score seems most critical is "AA", Prosper's highest rating. For "AA" closed loans the 75th percentile is about 20 points higher than for defaulted loans. The same is true of the medians. This tells me that an investor selecting a "AA" rating should also select a loan whose borrower has a credit score of 810 or higher to help increase the chance of a closed loan. The other rating with a strange anomaly is "C". While the 75th percentiles are about the same the median of closed loans is actually about 20 points less than for defaulted loans.

Plot Three



Description Three

About 95% of "AA" rated loans result in a closed status if a borrower has zero delinquencies and if they have one delinquency about 85% of loans result in closed. Interestingly if they had three or four, 100% of the loans were closed. This indicates to me that Prosper's calculation for assigning their rating does a good job taking into account delinquencies.

Reflection

Prosper provides a risk rating on a scale from HR(higher risk, 1) to AA(lower risk, 7) to help investors select an investment. I wanted to see if there were other financial indicators associated with a borrower that influenced loan status. As I progressed through the analysis it became apparent that treating loan status as a quantitative variable was not appropriate and I changed my focus to examining which variables might have the biggest affect on a loan be defaulted on vs. closed.

I was surprised that overall the Prosper Rating by itself is a very reasonable predictor and had to look carefully for subtleties that in conjunction with the rating might help predict the completion or default of a loan. At first I thought a borrower having any delinquencies was a definite indicator but as I looked more carefully at the data and took a subset of the data to only include records with a prosper rating which was added in 2009, I saw that the risk rating does seem to take delinquencies into account in the formula for calculating AA. My analysis indicates knowing the borrower's APR and Credit Score and associated medians for closed AA ratings could be useful in selecting a loan that has a greater chance of being closed. For all ratings knowing the borrower's debt to income ratio could also help an investor select a loan that is more likely to be closed.

Future work could include actually doing analysis of variance(ANOVA) to help determine which of the predictor variables, Debt to income ratio, APR, or Credit Score has the greatest influence on loan default. For example the proportion of borrowers who default given the loan has an APR of .23 or less to the proportion of borrowers who default given the loan has an APR greater than .23 could be compared.