Explore and Summarize Loan Data

Hye Joo Han

April 27, 2017

Table of Contents

# Load all of the packages that you end up using in your analysis in this code  
# chunk.  
  
library(ggplot2)  
library(tidyr)  
library(dplyr)   
library(knitr)  
library(GGally)  
library(scales)  
library(RColorBrewer)  
library(gridExtra)  
#library(memisc) # Will be loaded later for mtable()  
 # since this makes an error from select() in dplyr

# Load the Data  
LoanDataRaw = read.csv('prosperLoanData.csv', header = TRUE)

# Check the dimension of data and column names  
dim(LoanDataRaw)

## [1] 113937 81

#colnames(LoanDataRaw)

The dataset contains 113,937 loans each with 81 variables. The variables include loan amounts and prosper ratings measuring loan's level of risk, borrower's information such as their interest rate, Prosper rating, occupation, credit score, income and etc.

# Univariate Plots Section

## Selecting and overviewing variables

#Unload these two packages if select() does not work  
#detach("package:MASS", unload=TRUE)  
#detach("package:memisc", unload=TRUE)  
  
# Selelct variables to be investigated  
LoanData = select(LoanDataRaw, Term, LoanStatus, BorrowerRate,  
 ProsperRating..numeric., ProsperScore,   
 ListingCategory..numeric., BorrowerState, Occupation,  
 EmploymentStatus, CreditScoreRangeLower,   
 CreditScoreRangeUpper, DelinquenciesLast7Years,  
 AvailableBankcardCredit, IncomeRange, StatedMonthlyIncome,   
 LoanOriginalAmount, LoanMonthsSinceOrigination)  
# or LoanData = subset(LoanDataRaw, select = c(Term, LoanStatus, ... ))   
  
dim(LoanData)

## [1] 113937 17

colnames(LoanData)

## [1] "Term" "LoanStatus"   
## [3] "BorrowerRate" "ProsperRating..numeric."   
## [5] "ProsperScore" "ListingCategory..numeric."   
## [7] "BorrowerState" "Occupation"   
## [9] "EmploymentStatus" "CreditScoreRangeLower"   
## [11] "CreditScoreRangeUpper" "DelinquenciesLast7Years"   
## [13] "AvailableBankcardCredit" "IncomeRange"   
## [15] "StatedMonthlyIncome" "LoanOriginalAmount"   
## [17] "LoanMonthsSinceOrigination"

First, I selected 17 variables to be investigated and make a new data frame. Above are dimension of the new data frame and the selected variable names. Some of the variables give redundant information:

* ProsperRating..numeric. vs. ProsperScore
* IncomeRange vs. StatedMonthlyIncome
* CreditScoreRangeLower vs. CreditScoreRangeUpper

For these pairs, I will create a new variable from a pair or choose one from each pair for prediction models.

# Check variables  
str(LoanData)

## 'data.frame': 113937 obs. of 17 variables:  
## $ Term : int 36 36 36 36 36 60 36 36 36 36 ...  
## $ LoanStatus : Factor w/ 12 levels "Cancelled","Chargedoff",..: 3 4 3 4 4 4 4 4 4 4 ...  
## $ BorrowerRate : num 0.158 0.092 0.275 0.0974 0.2085 ...  
## $ ProsperRating..numeric. : int NA 6 NA 6 3 5 2 4 7 7 ...  
## $ ProsperScore : num NA 7 NA 9 4 10 2 4 9 11 ...  
## $ ListingCategory..numeric. : int 0 2 0 16 2 1 1 2 7 7 ...  
## $ BorrowerState : Factor w/ 52 levels "","AK","AL","AR",..: 7 7 12 12 25 34 18 6 16 16 ...  
## $ Occupation : Factor w/ 68 levels "","Accountant/CPA",..: 37 43 37 52 21 43 50 29 24 24 ...  
## $ EmploymentStatus : Factor w/ 9 levels "","Employed",..: 9 2 4 2 2 2 2 2 2 2 ...  
## $ CreditScoreRangeLower : int 640 680 480 800 680 740 680 700 820 820 ...  
## $ CreditScoreRangeUpper : int 659 699 499 819 699 759 699 719 839 839 ...  
## $ DelinquenciesLast7Years : int 4 0 0 14 0 0 0 0 0 0 ...  
## $ AvailableBankcardCredit : num 1500 10266 NA 30754 695 ...  
## $ IncomeRange : Factor w/ 8 levels "$0","$1-24,999",..: 4 5 7 4 3 3 4 4 4 4 ...  
## $ StatedMonthlyIncome : num 3083 6125 2083 2875 9583 ...  
## $ LoanOriginalAmount : int 9425 10000 3001 10000 15000 15000 3000 10000 10000 10000 ...  
## $ LoanMonthsSinceOrigination: int 78 0 86 16 6 3 11 10 3 3 ...

Next, I checked variable types using str() function. I noticed there are some blank entries for categorical variables and NA's for numerical variables.

# Summaries for each column  
summary(LoanData)

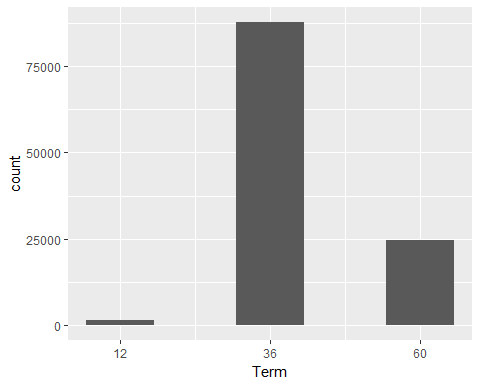
## Term LoanStatus BorrowerRate   
## Min. :12.00 Current :56576 Min. :0.0000   
## 1st Qu.:36.00 Completed :38074 1st Qu.:0.1340   
## Median :36.00 Chargedoff :11992 Median :0.1840   
## Mean :40.83 Defaulted : 5018 Mean :0.1928   
## 3rd Qu.:36.00 Past Due (1-15 days) : 806 3rd Qu.:0.2500   
## Max. :60.00 Past Due (31-60 days): 363 Max. :0.4975   
## (Other) : 1108   
## ProsperRating..numeric. ProsperScore ListingCategory..numeric.  
## Min. :1.000 Min. : 1.00 Min. : 0.000   
## 1st Qu.:3.000 1st Qu.: 4.00 1st Qu.: 1.000   
## Median :4.000 Median : 6.00 Median : 1.000   
## Mean :4.072 Mean : 5.95 Mean : 2.774   
## 3rd Qu.:5.000 3rd Qu.: 8.00 3rd Qu.: 3.000   
## Max. :7.000 Max. :11.00 Max. :20.000   
## NA's :29084 NA's :29084   
## BorrowerState Occupation EmploymentStatus  
## CA :14717 Other :28617 Employed :67322   
## TX : 6842 Professional :13628 Full-time :26355   
## NY : 6729 Computer Programmer : 4478 Self-employed: 6134   
## FL : 6720 Executive : 4311 Not available: 5347   
## IL : 5921 Teacher : 3759 Other : 3806   
## : 5515 Administrative Assistant: 3688 : 2255   
## (Other):67493 (Other) :55456 (Other) : 2718   
## CreditScoreRangeLower CreditScoreRangeUpper DelinquenciesLast7Years  
## Min. : 0.0 Min. : 19.0 Min. : 0.000   
## 1st Qu.:660.0 1st Qu.:679.0 1st Qu.: 0.000   
## Median :680.0 Median :699.0 Median : 0.000   
## Mean :685.6 Mean :704.6 Mean : 4.155   
## 3rd Qu.:720.0 3rd Qu.:739.0 3rd Qu.: 3.000   
## Max. :880.0 Max. :899.0 Max. :99.000   
## NA's :591 NA's :591 NA's :990   
## AvailableBankcardCredit IncomeRange StatedMonthlyIncome  
## Min. : 0 $25,000-49,999:32192 Min. : 0   
## 1st Qu.: 880 $50,000-74,999:31050 1st Qu.: 3200   
## Median : 4100 $100,000+ :17337 Median : 4667   
## Mean : 11210 $75,000-99,999:16916 Mean : 5608   
## 3rd Qu.: 13180 Not displayed : 7741 3rd Qu.: 6825   
## Max. :646285 $1-24,999 : 7274 Max. :1750003   
## NA's :7544 (Other) : 1427   
## LoanOriginalAmount LoanMonthsSinceOrigination  
## Min. : 1000 Min. : 0.0   
## 1st Qu.: 4000 1st Qu.: 6.0   
## Median : 6500 Median : 21.0   
## Mean : 8337 Mean : 31.9   
## 3rd Qu.:12000 3rd Qu.: 65.0   
## Max. :35000 Max. :100.0   
##

The summary of this dataset shows some factor levels and the number of blank entries for each of categorical variables (e.g., LoanStatus, IncomeRange) and the number of NA's and summary statistics for each of numerical variables (e.g., BorrowerRate, LoanOriginalAmount).

## Exploring each variable

### Term

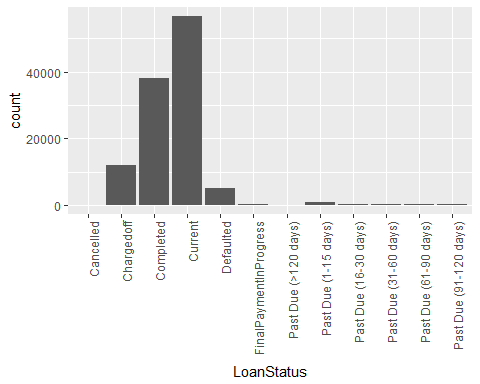
# Term (length of the loan expressed in months)  
#table(LoanData$Term)  
ggplot(data = LoanData, aes(x=Term)) + geom\_bar() +  
 scale\_x\_continuous(breaks = c(12,36,60))



The variable 'Term' contains lengths of loans in months. The plot shows there are only 3 kinds of length, 12, 36, and 36 months (i.e., 1, 3, 5 years). The most frequent length is 3 years. I wonder which variables are related to lengths of loans. For example, the length of a loan can be related to a loan amount 'LoanOriginalAmount' or its status 'LoanStatus'.

### LoanStatus

#LoanStatus (current status of the loan)  
#table(LoanData$LoanStatus)  
qplot(data = LoanData, x=LoanStatus) +   
 # for vertical x-axis labels  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))



The variable 'LoanStatus' contains the current status of a loan. The majority of loans are completed or current. I wonder if borrowers with higher Prosper ratings and lower interest rates are more likely to have loan status without issues.

To predict whether a loan has issues or not, I will make a new variable "GoodLoanStatus". In this variable, "1" stands for a loan that is completed, current, or with final payment in progress and "0" stands for a loan with all other levels (with issues). I will also remove loans with "Cancelled" status because we are not interested in those loans.

# Romove cancelled loans from dataset  
LoanData <- subset(LoanData, !(LoanStatus =="Cancelled"))

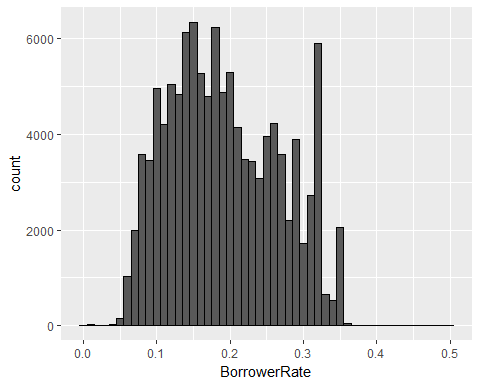
# Make a new variable  
# 1\* makes True and False into 1 and 0, respectively  
LoanData$GoodLoanStatus <- 1\*((LoanData$LoanStatus == "Completed") |   
 (LoanData$LoanStatus == "Current") |   
 (LoanData$LoanStatus == "FinalPaymentInProgress"))   
# Check some values of the new variable   
subset(LoanData, select = c(LoanStatus, GoodLoanStatus))[11:20,]

## LoanStatus GoodLoanStatus  
## 11 Current 1  
## 12 Completed 1  
## 13 Past Due (1-15 days) 0  
## 14 Current 1  
## 15 Current 1  
## 16 Defaulted 0  
## 17 Current 1  
## 18 Chargedoff 0  
## 19 Current 1  
## 20 Current 1

The above table shows some values of the new variable 'GoodLoanStatus' with the original variable 'LoanStatus'.

### BorrowerRate

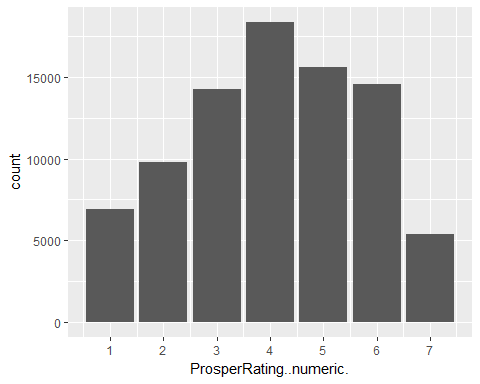
#BorrowerRate (Borrower's interest rate for this loan)  
qplot(data = LoanData, x=BorrowerRate, color = I('black'), binwidth =.01)



The variable 'BorrowerRate' of each loan shows borrower's interest rates for the loan. The plot shows most interest rates are between 5% and 35%. I expect a borrower rate is related to many other variables in this dataset since an interest rate is likely to be influenced by credit or Prosper scores or lengths of loans. We will have some limitations on predicting interest rates because interest rates of loans also depend on other factors not included in this dataset such as government's directives or the market.

### ProsperRating..numeric.

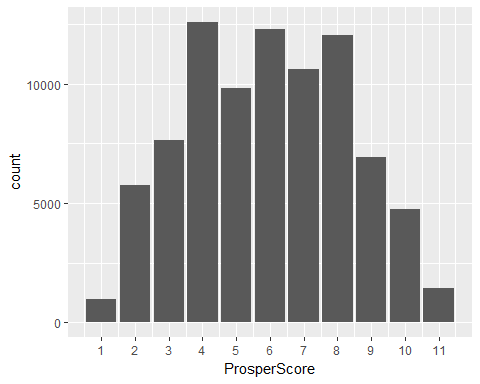
#ProsperRating..numeric.   
# NA - N/A, 1 - HR, 2 - E, 3 - D, 4 - C, 5 - B, 6 - A, 7 - AA  
# NA's are not shown in the plot  
qplot(data = LoanData, x=ProsperRating..numeric.) + geom\_bar() +  
 scale\_x\_continuous(breaks = 1:7)



'ProsperRating..numeric.' contains the level of risk for each loan. There are levels1 through 7 and NA with 7 being the lowest level of risk and 1 being the highest risk. The plot shows the Prosper rating has a bell-shaped distribution with one mode at the middle Prosper rating 4.

### ProsperScore

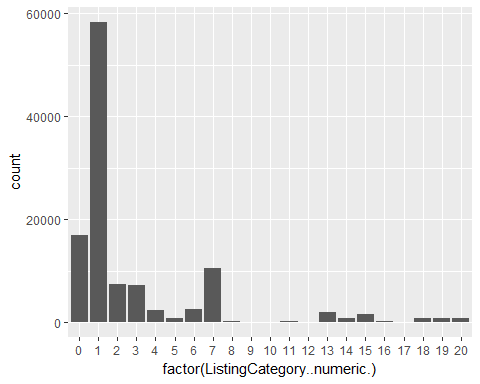
#ProsperScore (A custom risk score built using historical Prosper data.   
# The score ranges from 1-11, with higher score being the lower risk score.  
# NA's are not shown in the plot  
qplot(data = LoanData, x=ProsperScore) + geom\_bar() +  
 scale\_x\_continuous(breaks = 1:11)



'ProsperScore' is a custom risk score measured using historical Prosper data. It is similar to 'ProsperRating..numeric.' and they indeed have similar distributions. I will choose a better variable from the two eventually. As mentioned, these variables are likely related to loan status and interest rates.

### ListingCategory..numeric.

#ListingCategory..numeric. (category of the listing that the borrower selected   
# when posting their listing)  
qplot(data = LoanData, x=factor(ListingCategory..numeric.))



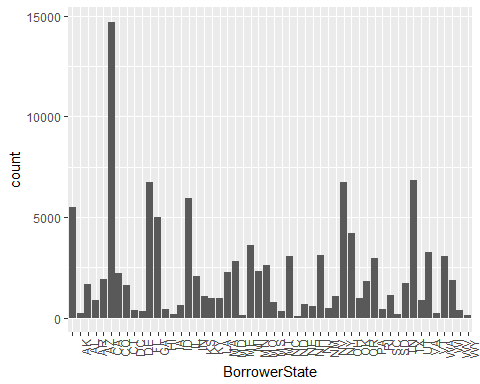
'ListingCategory..numeric.' contains categories of the listing selected by borrowers. Each number stands for as followings.

0 - Not Available, 1 - Debt Consolidation, 2 - Home Improvement, 3 - Business, 4 - Personal Loan, 5 - Student Use, 6 - Auto, 7- Other, 8 - Baby&Adoption, 9 - Boat, 10 - Cosmetic Procedure, 11 - Engagement Ring, 12 - Green Loans, 13 - Household Expenses, 14 - Large Purchases, 15 - Medical/Dental, 16 -Motorcycle, 17 - RV, 18 - Taxes, 19 - Vacation, 20 - Wedding Loans

The majority of loans were used for debt consolidation (1) and the second and third largest counts were found in categories "Not Available" and "Other". Thus, I will not investigate this variable further with other variables.

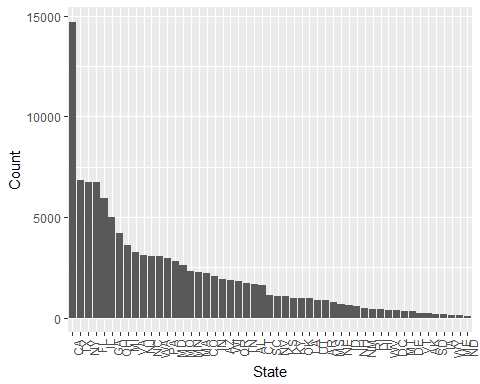
### BorrowerState

#BorrowerState (state of the address of the borrower)  
qplot(data = LoanData, x=BorrowerState) +   
 # for vertical x-axis labels  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))



'BorrowerState' shows a state abbreviation of borrower's addresses. I wanted to check which states have more loans, but this graph is somewhat hard to check, so I sorted states by their counts for loans below.

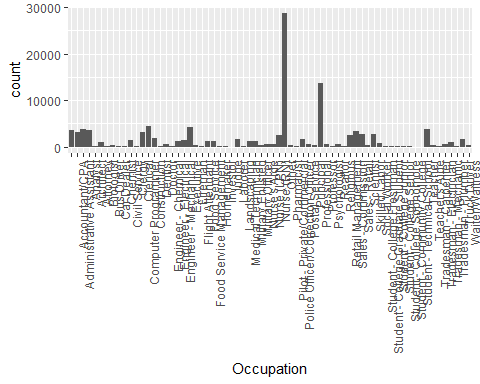
# Sort by counts  
State\_count <- data.frame(table(LoanData$BorrowerState))  
colnames(State\_count)[1] <- 'State'  
colnames(State\_count)[2] <- 'Count'  
State\_count\_orderd <- State\_count[-1,] %>% #removed count for blanks  
 arrange(desc(Count))   
# Keep the order of levels sorted by counts  
State\_count\_orderd$State <- factor(State\_count\_orderd$State,   
 levels = State\_count\_orderd$State)  
# Plot again with states ordered by counts  
ggplot(data=State\_count\_orderd, aes(x=State, y = Count)) +  
 geom\_bar(stat="identity") +   
 # for vertical x-axis label  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))



The graph shows California is the state with the biggest number of loans. I noticed that the order of top states seems to be similar to the order of states with top populations (<http://worldpopulationreview.com/states/>). The number of loans and population of states look correlated, but I will not look into this further since state populations are not in our dataset.

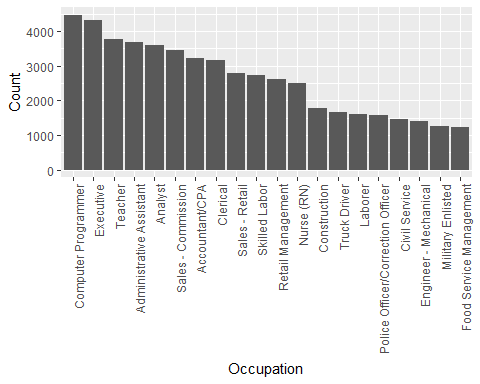
### Occupation

# Occupation (Occupation selected by the Borrowe)  
qplot(data = LoanData, x=Occupation) +   
 # for vertical x-axis labels  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))



I also wanted to check borrower's occupations, but the above plot has too many categories to find frequent occupations. Thus, I will sort occupations by their counts to make a plot.

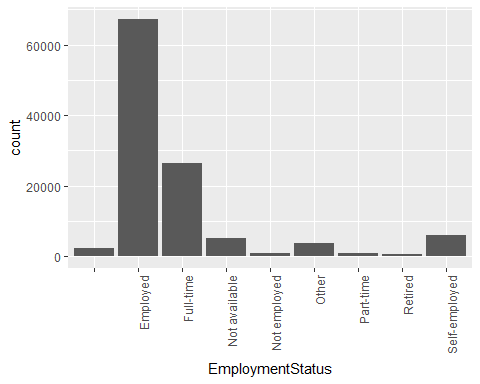
# Sort by counts  
Occupation\_count <- data.frame(table(LoanData$Occupation))  
colnames(Occupation\_count)[1] <- 'Occupation'  
colnames(Occupation\_count)[2] <- 'Count'  
Occupation\_count\_orderd <- Occupation\_count[-1,] %>% #removed count for blanks  
 arrange(desc(Count))  
Occupation\_count\_orderd$Occupation<-factor(Occupation\_count\_orderd$Occupation,   
 levels = Occupation\_count\_orderd$Occupation)  
# Barplot of top 20 occupations   
ggplot(data=Occupation\_count\_orderd[3:22,], #omitting 'Other' & 'Professional'  
 # in the 1st & 2nd rows from the plot  
 aes(x=Occupation, y = Count)) +  
 geom\_bar(stat="identity") +   
 # for vertical x-axis label  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))



This barplot shows the top 20 occupations of borrowers. I omitted the top two categories, 'Other' & 'Professional' from the plot since they are too ambiguous. I wonder how these categories are related to other variables. For example, I wonder which occupations have lower interest rates on average.

### EmploymentStatus

# EmploymentStatus   
qplot(data = LoanData, x=EmploymentStatus) +   
 # with vertical x-axis labels  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))



This plot shows the majority of borrowers are employed as expected.

### CreditScoreRangeLower and CreditScoreRangeUpper

str(subset(LoanData, select = c(CreditScoreRangeLower,CreditScoreRangeUpper)))

## 'data.frame': 113932 obs. of 2 variables:  
## $ CreditScoreRangeLower: int 640 680 480 800 680 740 680 700 820 820 ...  
## $ CreditScoreRangeUpper: int 659 699 499 819 699 759 699 719 839 839 ...

CreditScoreRangeLower and CreditScoreRangeUpper are always 19 scores apart. To make a simpler variable, a new variable 'CreditScore' was made using the number CreditScoreRangeLower plus 10, which is in the middle (not exactly) of lower and upper bounds.

LoanData$CreditScore <- LoanData$CreditScoreRangeLower +10  
summary(LoanData$CreditScore)

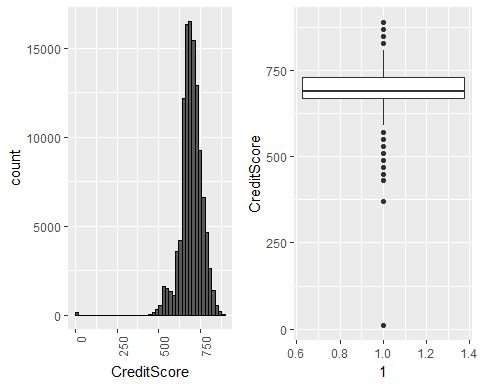
## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 10.0 670.0 690.0 695.6 730.0 890.0 590

table(LoanData$CreditScore)

##   
## 10 370 430 450 470 490 510 530 550 570 590 610   
## 133 1 5 36 141 346 553 1592 1474 1357 1125 3602   
## 630 650 670 690 710 730 750 770 790 810 830 850   
## 4172 12198 16366 16492 15471 12922 9267 6606 4624 2644 1409 567   
## 870 890   
## 212 27

Above are the summary of the new variable 'CreditScore' and its frequency table.

# CreditScore   
# NA's are not shown in the plot  
grid.arrange(qplot(data = LoanData, x=CreditScore, color = I('black'),   
 binwidth =20, boundary =0) +  
 #vertical x-axis labels  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1)) ,  
 ggplot(aes(x=1, y=CreditScore), data = LoanData) +  
 geom\_boxplot( ) , nrow =1)



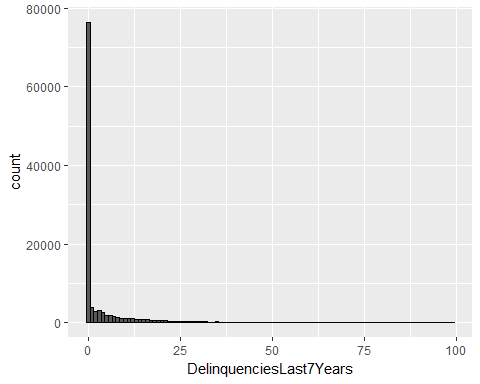
The plots and frequency table show that most credit scores are between 450 and 890. There are 133 borrowers with extremely low credit score 10 (minimum) and no borrower with scores between 10 and 370. I wonder who are those people with credit score 10.

### DelinquenciesLast7Years

# DelinquenciesLast7Years (Number of delinquencies in the past 7 years)   
# NA's are not shown in the plot  
head(data.frame(table(LoanData$DelinquenciesLast7Years)))

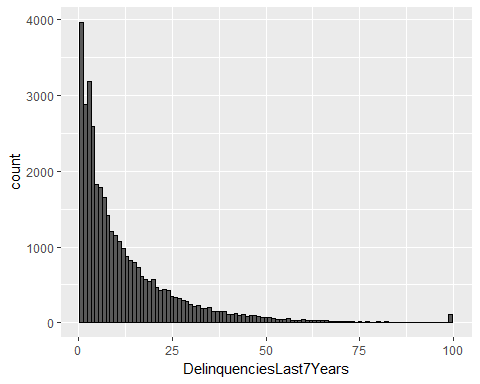
## Var1 Freq  
## 1 0 76438  
## 2 1 3967  
## 3 2 2879  
## 4 3 3182  
## 5 4 2592  
## 6 5 1826

qplot(data = LoanData, x=DelinquenciesLast7Years, color = I('black'),  
 boundary = -0.5, binwidth =1)



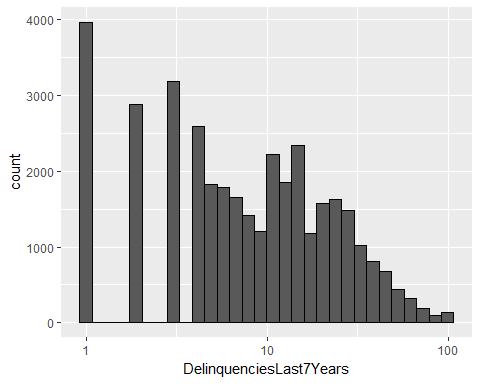
Zero is the most and extremely frequent value for the number of delinquencies in the past 7 years. The histogram seems right-skewed. For a better look, two further histograms were made (below). The first plot is made without zeros and and the second plot is created after omitting zeros and taking log 10 transformation to the number of delinquencies.

# NA's are not shown in the plot  
qplot(data = LoanData, x=DelinquenciesLast7Years, color = I('black'),  
 boundary = -0.5, binwidth =1) +  
 xlim(0.5,100) #removing the first bar with zeros



The histogram shows frequency decreases as the number of delinquencies increases, but frequency suddenly increase at 99, the maximum. It seems this is simply because the highest number set for this variable was 99. I will investigate what other variables this variable is related to.

# Histrogram of Log10 of Number of Delinquencies   
# Zero's are not shown in the plot since log10 of zero is undefined.   
qplot(data = subset(LoanData,DelinquenciesLast7Years > 0), # removing zeros  
 x=DelinquenciesLast7Years, color = I('black')) +  
 scale\_x\_log10()



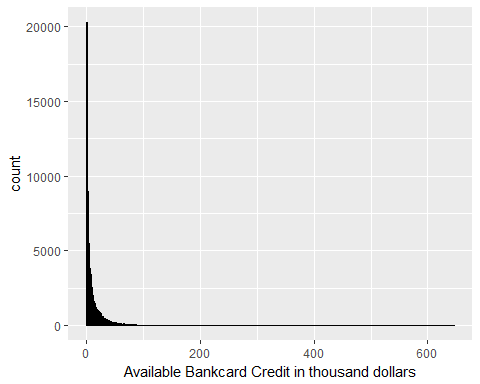
The above plot shows frequency for log 10 of number of delinquencies. Zeros were removed in the plot since log 10 of zero is undefined.

### AvailableBankcardCredit

#AvailableBankcardCredit (total available credit via bank card)   
# NA's are not shown in the plot  
head(data.frame(table(LoanData$AvailableBankcardCredit)))

## Var1 Freq  
## 1 0 4881  
## 2 1 47  
## 3 2 50  
## 4 3 44  
## 5 4 45  
## 6 5 41

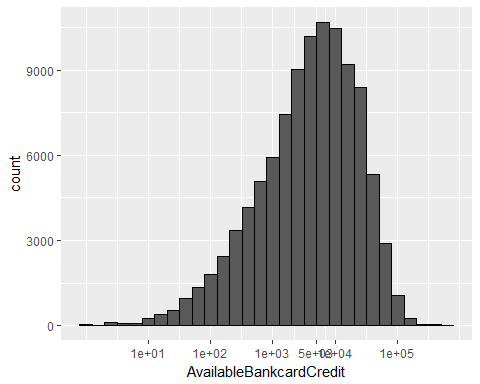
qplot(data = LoanData, x = AvailableBankcardCredit/1000, color = I('black'),   
 binwidth = 1) + xlab("Available Bankcard Credit in thousand dollars")



The plot is a histogram of available bank card credit in thousand dollars. The extremely high frequency for $0 and extremely high credit in several hundred thousand dollars make it hard to read the plot. Thus, I log-transformed the variable to make the smaller values visible (below).

# AvailableBankcardCredit   
LoanData\_noNA\_BankCredit <- subset(LoanData, !is.na(AvailableBankcardCredit))  
  
#qplot(data =LoanData\_noNA\_BankCredit,   
# x=AvailableBankcardCredit,   
# color = I('black'), binwidth = 1000) +  
# scale\_x\_continuous(limits =  
# c(quantile(LoanData\_noNA\_BankCredit$AvailableBankcardCredit,.01),  
# quantile(LoanData\_noNA\_BankCredit$AvailableBankcardCredit,.99)))

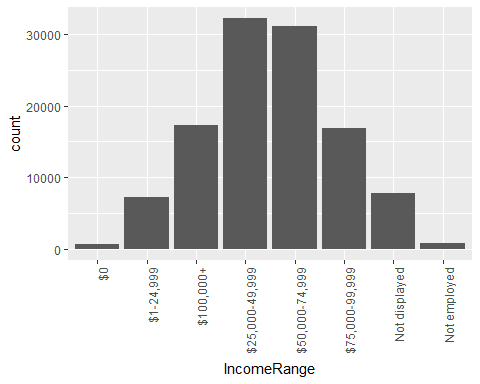
#AvailableBankcardCredit (total available credit via bank card)   
# NA's are not shown in the plot  
qplot(data = subset(LoanData,AvailableBankcardCredit > 0),   
 x = AvailableBankcardCredit, color = I('black')) +  
 scale\_x\_log10(breaks = c(10,100,1000,5000,10000,100000))



The most of bank card credit is less than 100,000 dollars and the mode is around 5000.

### IncomeRange

#IncomeRange   
qplot(data = LoanData, x=IncomeRange) +   
 # with vertical x-axis labels  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))



The above barplot shows income ranges of borrowers (per year), but the order of factor levels is not well arranged.

levels(LoanData$IncomeRange)

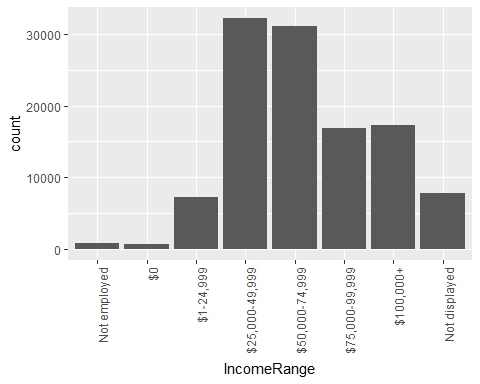
## [1] "$0" "$1-24,999" "$100,000+" "$25,000-49,999"  
## [5] "$50,000-74,999" "$75,000-99,999" "Not displayed" "Not employed"

# Arrange the factor levels  
LoanData$IncomeRange = factor(LoanData$IncomeRange, levels =  
 levels(LoanData$IncomeRange)[c(8,1,2,4,5,6,3,7)])  
levels(LoanData$IncomeRange)

## [1] "Not employed" "$0" "$1-24,999" "$25,000-49,999"  
## [5] "$50,000-74,999" "$75,000-99,999" "$100,000+" "Not displayed"

Above are the factor levels before and after I changed the order of levels for income ranges.

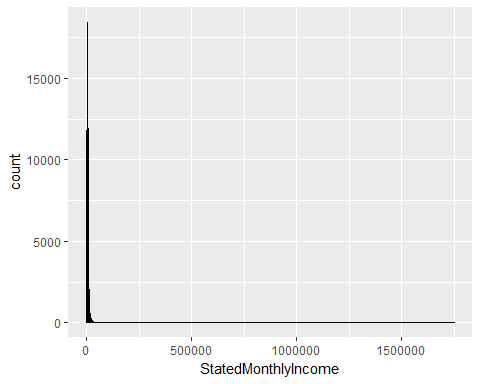
#IncomeRange   
qplot(data = LoanData, x=IncomeRange) +   
 # with vertical x-axis labels  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))



The factor levels of income ranges were reordered and the barplot was constructed with the new order. The plot shows the most common income ranges are are 25,000-49,999 and 50,000-74,999 dollars, but there are also substantially many borrowers with income ranges 75,000-99,999 and $100,000+. I want to investigate how other variables differ for different income ranges.

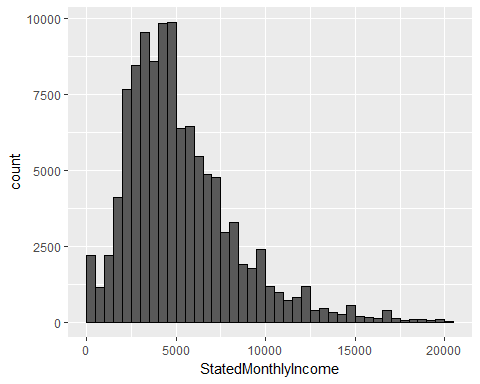
### StatedMonthlyIncome

#StatedMonthlyIncome   
qplot(data = LoanData, x=StatedMonthlyIncome, color = I('black'),  
 binwidth = 1000)



StatedMonthlyIncome is a similar variable to IncomeRange, but the amount is for each month (not year) and it is a numerical variable. Because the extremely high incomes make it hard to check the distribution, I omitted the top 1% incomes in the next plot.

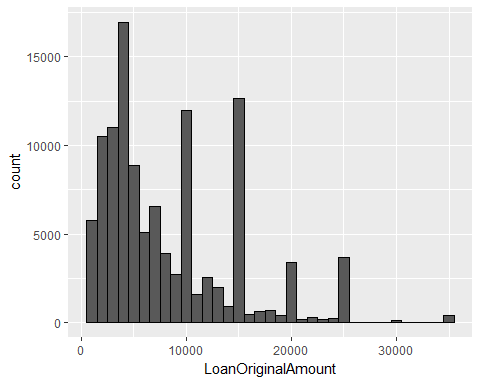
#StatedMonthlyIncome   
qplot(data = LoanData, x=StatedMonthlyIncome, color = I('black'),  
 binwidth = 500, boundary = 0) +   
 xlim(0, quantile(LoanData$StatedMonthlyIncome,.99))



Even without the top 1%, the histogram of monthly incomes still looks pretty right-skewed. The peak is around $4500.

### LoanOriginalAmount

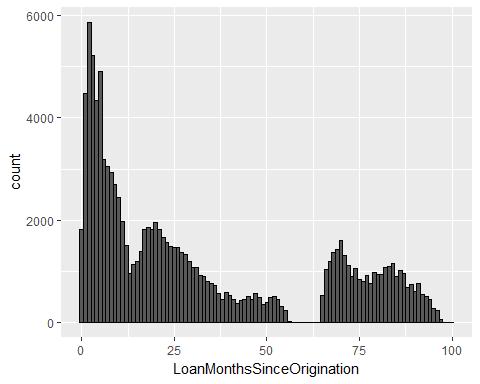
#LoanOriginalAmount   
qplot(data = LoanData, x=LoanOriginalAmount, color = I('black'),  
 binwidth = 1000)



The histogram shows the distribution of original loan amounts. The graph has a long right tail. Small loan amounts are more frequent with a peak at 4000 and amounts over 20,000 are very rare except for 25,000. There are some loan amounts much more frequent than their neighboring amounts; they are $4000 and multiples of $5000. I wonder how this variable is related to the interest rate.

### LoanMonthsSinceOrigination

# LoanMonthsSinceOrigination   
qplot(data = LoanData, x=LoanMonthsSinceOrigination, color = I('black'),  
 binwidth=1)



The histogram shows the distribution of the number of months since loan origination. The distribution is right-skewed and there are no records of loans around 60 and 65 months old. This variable will be useful when looking at some changes over time.

# Univariate Analysis

### What is the structure of your dataset?

The dataset contains 113,937 loans each with 17 variables I selected. Through univariate analysis I decided to drop some of the variables for further analysis and created some new variables.

Categorical variables:

* LoanStatus
* BorrowerState (to be dropped)
* Occupation
* ListingCategory..numeric. (to be dropped)
* EmploymentStatus
* IncomeRange
* GoodLoanStatus (Created)

Numerical variables

* Term
* BorrowerRate
* ProsperRating..numeric.
* ProsperScore
* CreditScoreRangeLower (to be dropped)
* CreditScoreRangeUpper (to be dropped)
* CreditScore (Created)
* DelinquenciesLast7Years
* AvailableBankcardCredit
* StatedMonthlyIncome
* LoanOriginalAmount
* LoanMonthsSinceOrigination

None of the categorical variables have completely ordered levels, but I reordered levels of 'IncomeRange' to make levels with dollar amounts ordered.

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

A main feature of interest is 'BorrowerRate' that contains interest rates of loans. This is likely to be predicted well using variables for loan's level of risk i.e., 'ProsperScore' or 'ProsperRating..numeric.'.

'GoodLoanStatus' I created could be another main feature of interest I would like to predict. Although we can predict a categorical variable like 'GoodLoanStatus' using logistic regressions, this prediction could be more challenging than predicting 'BorrowerRate'. Thus, I will only explore how GoodLoanStatus are related to other variables and stop there. Loan's level of risk could be again a good predictor for this variable.

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

The followings are other features that can help predicting 'BorrowerRate':

* Term
* DelinquenciesLast7Years
* AvailableBankcardCredit
* StatedMonthlyIncome
* LoanOriginalAmount

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

I created two new variables:

* GoodLoanStatus

I created 'GoodLoanStatus'. The variable contains "1" for a loan that is completed, current, or with final payment in progress and "0" for a loan with all other bad status, charged off, defaulted, or past due. Note that cancelled loans were removed from the dataset.

* CreditScore

To make a simpler variable than 'CreditScoreRangeLower' and 'CreditScoreRangeUpper' , 'CreditScore' was created using the number in the middle (not exactly) of lower and upper bounds of credit scores.

### 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?

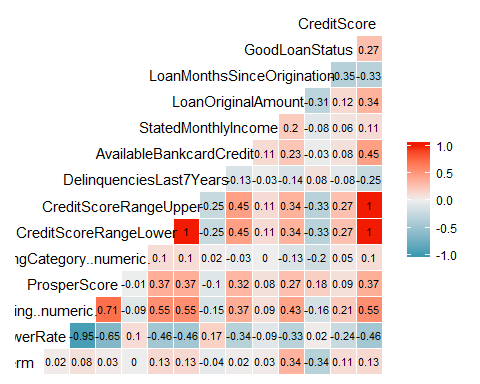
'AvailableBankcardCredit' and 'StatedMonthlyIncome' are too right-skewed to read their histograms. When I removed the highest 1% values of the variables, 'StatedMonthlyIncome' was looking good. Since 'AvailableBankcardCredit' was still very right-skewed, so I log-transformed the variable after omitting zeros.

'DelinquenciesLast7Years' contains too many zero values to read its histogram well. First, I removed the bar with zeros from the histogram, but it was still very right-skewed. Thus, next I log-transformed the variable for a better investigation.

'IncomeRange' (income ranges of borrowers per year) did not have well-ordered factor levels, so I arranged the factor levels to make levels with dollar amounts ordered.

# Bivariate Plots Section

ggcorr(LoanData,   
 label = TRUE, # corr. coefficients  
 hjust = 0.9, # change variable name positions  
 label\_round = 2, # corr. digits  
 label\_size = 3) # corr. letter size

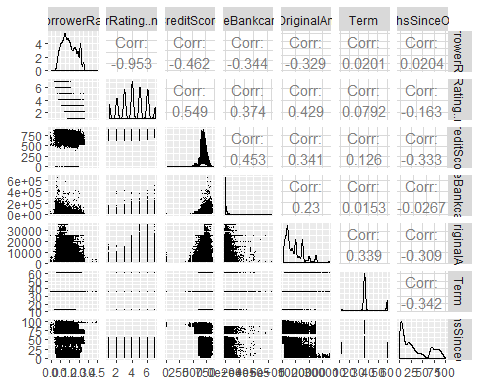


# Correlations between selected variables  
#round(cor(subset(na.omit(LoanData), select = c(Term,  
# BorrowerRate,  
# ProsperRating..numeric.,  
# ProsperScore,  
# CreditScore,  
# DelinquenciesLast7Years,  
# AvailableBankcardCredit,  
# StatedMonthlyIncome,  
# LoanOriginalAmount,  
# GoodLoanStatus,  
# LoanMonthsSinceOrigination  
# ))),  
# 3)

The above output shows correlation coefficients between numerical variables. The strongest correlation is found between 'BorrowerRate' and 'ProsperRating..numeric.' (correlation r = -0.95). Moreover, 'ProsperScore', 'CreditScore', 'AvailableBankcardCredit' and 'LoanOriginalAmount' have moderate to high correlations to 'BorrowerRate'.

ProsperScore will be omitted from here since it is less correlated with other variables than ProsperRating..numeric.

# Overall plots using subset data   
ggpairs(subset(LoanData, select = c(BorrowerRate,  
 ProsperRating..numeric.,  
 CreditScore,  
 AvailableBankcardCredit,  
 LoanOriginalAmount,  
 Term,  
 LoanMonthsSinceOrigination  
 )),  
 lower = list(continuous = wrap("points", shape = I('.'))),   
 upper = list(combo = wrap("box", outlier.shape = I('.'))))



The above set of graphs are created using ggpairs(). It shows overall relationships between some variables (the correlations are slightly different from the above table because of the different ways of removing NAs).

The following are the variables to be investigated. I decided these using the above correlations and plots.

Between the main and each of supporting variables

* Interest rate vs. Prosper Rating
* Interest rate vs. Credit score
* Interest rate vs. Bank card credit
* Interest rate vs. Loan original amount
* Interest rate vs. Loan term
* Interest rate vs. Loan months since origination

Between supporting variables

* Prosper Rating vs. Credit score
* Prosper Rating vs. Bank card credit
* Prosper Rating vs. Loan original amount

Additionally, I will also check how the following categorical variables are related to the main variable, interest rate.

* Loan status (Good loan status)
* Occupation
* Income range

Finally, I will investigate how the proportion of good loan status varies over different Prosper ratings and interest rates.

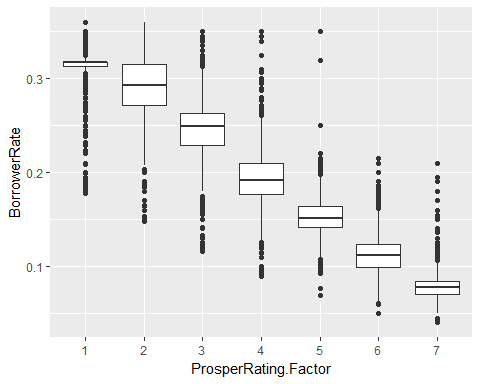
I will now look into relationships between the main variable and each of supporting variables.

### Interest rate vs. Prosper Rating

'ProsperRating..numeric.' has been treated as a numeric variable, but it can be considered as a categorical variable with order levels. If we change it to a categorical variable, we can easily make boxplots for all levels.

# Change numeric to categorical variable  
LoanData$ProsperRating.Factor <- factor(LoanData$ProsperRating..numeric.,  
 ordered = TRUE)  
#levels(LoanData$ProsperRating.Factor)

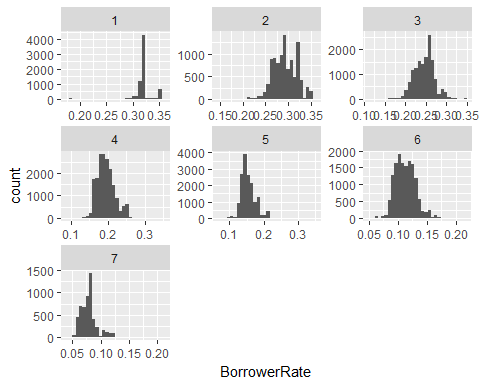
qplot(data = subset(LoanData,!is.na(ProsperRating.Factor)),   
 x=ProsperRating.Factor, y=BorrowerRate, geom = 'boxplot')



The boxplots show that interest rates decrease as a Prosper rating increases. The boxes of different Prosper ratings never overlap with each other and this means interest rates for different Prosper ratings are different significantly.

The distribution of interest rates for each Prosper Rating looks different.

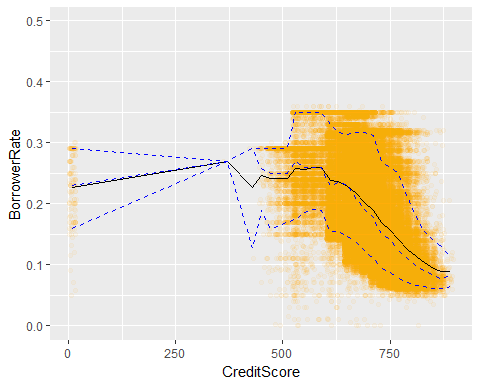
qplot(x = BorrowerRate, data = subset(LoanData,!is.na(ProsperRating.Factor))) +   
 facet\_wrap(~ProsperRating.Factor, scales = "free")



The distribution of interest rates for the lowest Prosper rating (highest loan risk) is very left-skewed while the distribution of interest rates for the highest Prosper rating (lowest loan risk) is very right-skewed. The distributions for all other moderate Prosper ratings are more symmetric. This means some borrowers in the highest risk group still receive much lower interest rates than others in that group and some borrowers in the lowest risk group still receive much higher interest rates than others in the same group for some reasons. I would like to find out what other variables make these exceptional interest rates.

### Interest rate vs. Credit Score

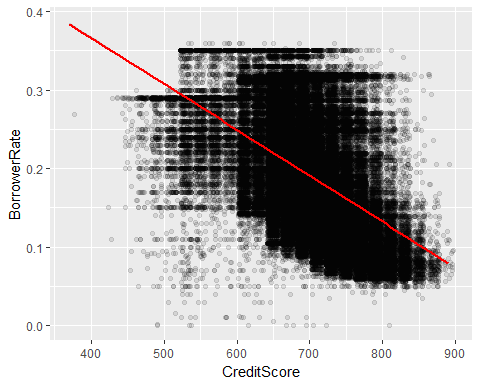
ggplot(aes(x = CreditScore, y = BorrowerRate), data = LoanData) +  
 geom\_jitter(alpha=1/20, color=I('orange')) +   
 # Add summary lines  
 geom\_line(stat='summary', fun.y=mean) +  
 geom\_line(stat='summary', fun.y=quantile, fun.args = list(probs = .1),   
 linetype = 2, color='blue') +  
 geom\_line(stat='summary', fun.y=quantile, fun.args = list(probs = .5),  
 linetype = 2, color='blue') +  
 geom\_line(stat='summary', fun.y=quantile, fun.args = list(probs = .9),  
 linetype = 2, color='blue')



The black curve on the graph is connecting the mean of interest rates for each credit score. The 3 dotted blue curves are representing 10, 50 (median), 90 percentiles of interest rates for each credit score.

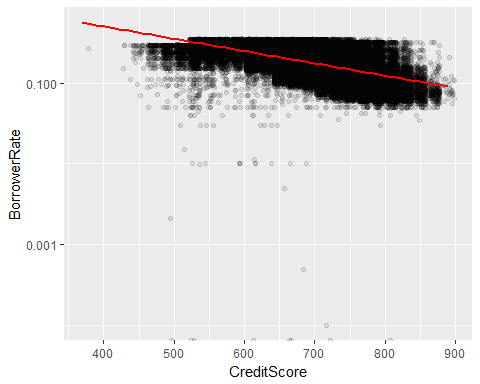
The graph shows interest rates seem to decrease as credit scores increase. I will add a linear regression line to see this more clearly. There is no point between 10 and 370 credit scores, so I will zoom the graph by removing the points with credit score 10.

ggplot(aes(x = CreditScore, y = BorrowerRate),   
 # Removing low credits score outliers  
 data = subset(LoanData, CreditScore > 10)) +   
 geom\_jitter(alpha=1/10) +  
 # Add a regression line  
 geom\_smooth(method = 'lm', color='red')



As expected, this graph with the regression line shows the negative relationship between interest rates and credit scores. The variance of y tends to be larger for larger y, so I will take a log transformation for y in the next graph.

ggplot(aes(x = CreditScore, y = BorrowerRate),   
 data = subset(LoanData, CreditScore > 10)) +  
 geom\_jitter(alpha=1/10) +  
 geom\_smooth(method = 'lm', color='red') +  
 scale\_y\_continuous(trans = log10\_trans()) #log10 transformation



It seems the points with log 10 of interest rates stay much closer to the linear regression line now.

# omit all NAs (more points are removed, but for a quick check)  
LoanData\_NoNA <- na.omit(LoanData)   
# Correlations  
c("corr before log-transforming y:",  
 round(cor(LoanData\_NoNA$CreditScore, LoanData\_NoNA$BorrowerRate),3))

## [1] "corr before log-transforming y:" "-0.509"

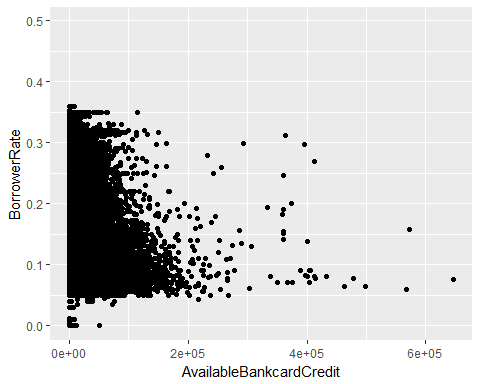
c("corr after log-transforming y:",   
 round(cor(LoanData\_NoNA$CreditScore, log10(LoanData\_NoNA$BorrowerRate)),3))

## [1] "corr after log-transforming y:" "-0.544"

The log transformation indeed improved the correlation between the variables! I also tried many powers for x, but no further improvements were evident with transforming x.

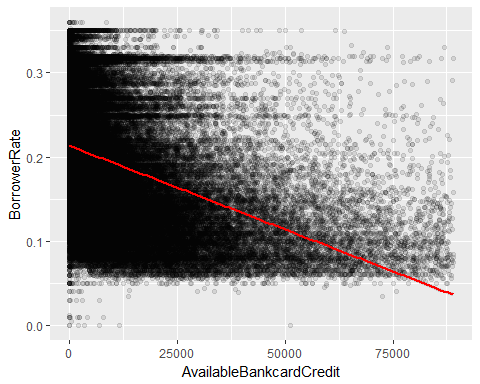
### Interest rate vs. Bank card credit

ggplot(aes(x = AvailableBankcardCredit, y = BorrowerRate), data = LoanData) +  
 geom\_point()



As I found that there are extremely outliers for bank card credits from my univariate analysis, the outliers make it hard to see the overall relationship between interest rates and bank card credits. I will improve this scatter plot by applying alpha and removing points with extreme bank card credits (above 99 percentile) and NAs.

ggplot(aes(x = AvailableBankcardCredit, y = BorrowerRate),   
 data = subset(LoanData,!is.na(AvailableBankcardCredit))) +  
 geom\_point(alpha= 1/10) +  
 # Omit extreme values  
 scale\_x\_continuous(limits =  
 c(0, quantile(LoanData\_noNA\_BankCredit$AvailableBankcardCredit,.99))) +   
 # Add a regression line  
 geom\_smooth(method = 'lm', color='red')



This graph shows the relationship between interest rates and bank card credits better. The red regression line shows interest rates tend to decrease as credit card credits increase. However, the flat bottom part of points around y = 0.05 shows lowest interest rates starts around 5% regardless of bank card credits. Moreover, the exceptionally low interest rates were usually given to borrowers with very low bank card credits.

# omit all NAs (more points are removed, but for a quick check)  
LoanData\_NoNA <- na.omit(LoanData)   
# Correlations  
c("corr between x and y:",  
 round(cor(LoanData\_NoNA$AvailableBankcardCredit,   
 LoanData\_NoNA$BorrowerRate),3))

## [1] "corr between x and y:" "-0.36"

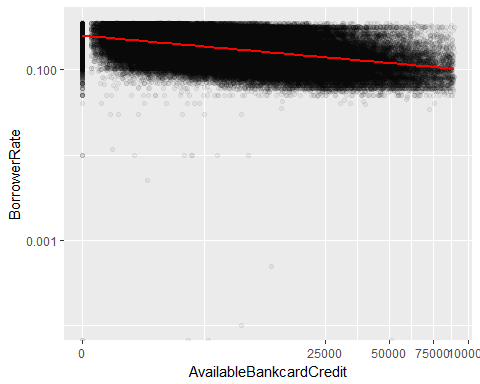
c("corr between x and log y:",  
 round(cor(LoanData\_NoNA$AvailableBankcardCredit,   
 log10(LoanData\_NoNA$BorrowerRate)),3))

## [1] "corr between x and log y:" "-0.4"

c("corr between cube root of x and log y:",  
 round(cor((LoanData\_NoNA$AvailableBankcardCredit)^(1/3),   
 log10(LoanData\_NoNA$BorrowerRate)),3))

## [1] "corr between cube root of x and log y:"  
## [2] "-0.486"

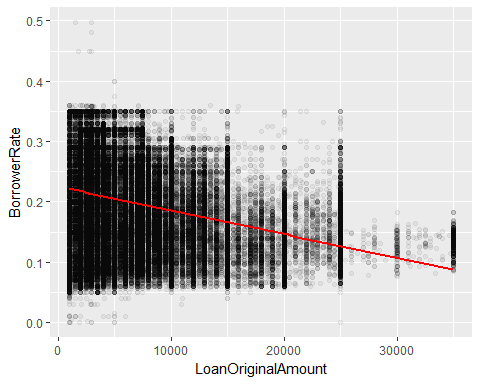
# Make cuberoot\_trans function  
cuberoot\_trans <- function() trans\_new('cuberoot',  
 transform = function(x) x^(1/3),  
 inverse = function(x) x^3)  
  
ggplot(aes(x = AvailableBankcardCredit, y = BorrowerRate),   
 data = subset(LoanData,!is.na(AvailableBankcardCredit))) +  
 geom\_point(alpha= 1/20) +  
 scale\_x\_continuous(trans = cuberoot\_trans(), #cube root transformation  
 limits =  
 c(0, quantile(LoanData\_noNA\_BankCredit$AvailableBankcardCredit,.99))) +   
 # Add a regression line  
 geom\_smooth(method = 'lm', color='red') +  
 scale\_y\_continuous(trans = log10\_trans()) # log10 transformation



To stabilize variance, log transformation was used again for y and it worked as before and the cube root of x improved the fit even more!

### Interest rate vs. Loan original amount

ggplot(aes(x = LoanOriginalAmount, y = BorrowerRate), data = LoanData) +  
 geom\_jitter(alpha=1/20) +  
 geom\_smooth(method = 'lm', color='red')



This shows an interest rate tends to be smaller for a loan with higher original amount. No loans with original amounts higher than 25000 were given high interest rates; the rates are around between 5% and 20%. As loan amounts decrease, the variance of interest rates tends to increase. Thus, I will transform x and y again.

# omit all NAs (more points are removed, but for a quick check)  
LoanData\_NoNA <- na.omit(LoanData)   
# Correlations  
c("corr between x and y:",  
 round(cor(LoanData\_NoNA$LoanOriginalAmount, LoanData\_NoNA$BorrowerRate),3))

## [1] "corr between x and y:" "-0.413"

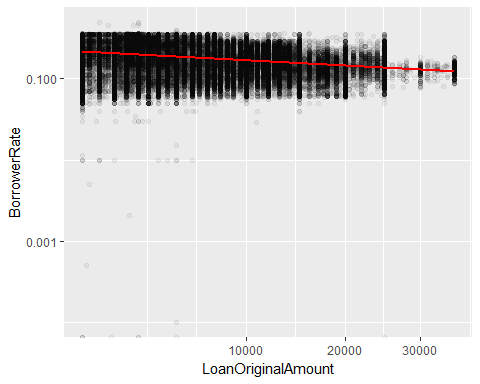
c("corr between x and log y:",   
 round(cor(LoanData\_NoNA$LoanOriginalAmount,   
 log10(LoanData\_NoNA$BorrowerRate)),3))

## [1] "corr between x and log y:" "-0.366"

c("corr between square root of x and log y:",  
 round(cor((LoanData\_NoNA$LoanOriginalAmount)^(1/2),   
 log10(LoanData\_NoNA$BorrowerRate)),3))

## [1] "corr between square root of x and log y:"  
## [2] "-0.373"

ggplot(aes(x = LoanOriginalAmount, y = BorrowerRate), data = LoanData) +  
 geom\_point(alpha= 1/20) +  
 scale\_x\_continuous(trans = sqrt\_trans()) + # square root transformation  
 scale\_y\_continuous(trans = log10\_trans()) + # log10 transformation  
 geom\_smooth(method = 'lm', color='red')

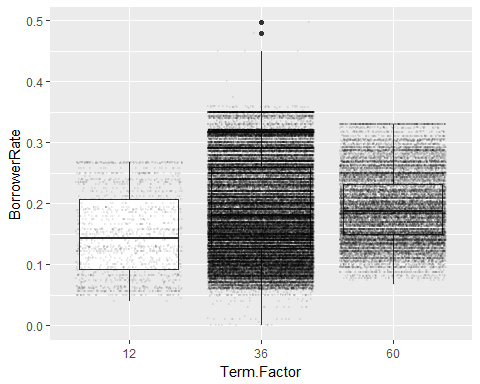


The log transformation of y and the square root transformation of x made the variance more constant over different x's and a better fit to the regression line.

### Interest rate vs. Term

#is.factor(LoanData$Term)  
LoanData$Term.Factor <- factor(LoanData$Term, ordered = TRUE)  
#levels(LoanData$Term.Factor)

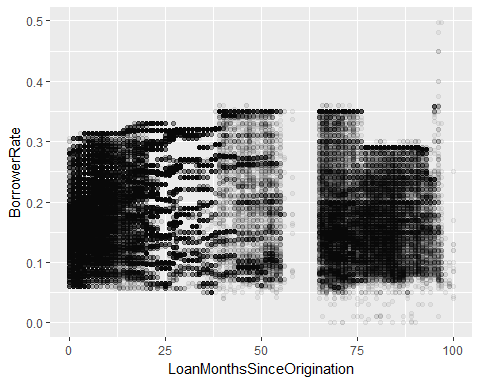
qplot(x = Term.Factor, y = BorrowerRate, data = LoanData,   
 geom = 'boxplot') +  
 geom\_point(position = 'jitter',alpha=1/20,size=1/2)



I made 'Term.Factor' variable to make a categorical variable from the numeric one 'Term' since there are only 3 kinds for loan terms. This graph shows an interest rate of a loan is related to the length of the loan. Interest rates tend to be higher if lengths of loans are longer, but the loans with 36 months term are much more scattered interest rates than loans with other lengths. This pattern in 36 months loans can be possibly from their much higher frequency.

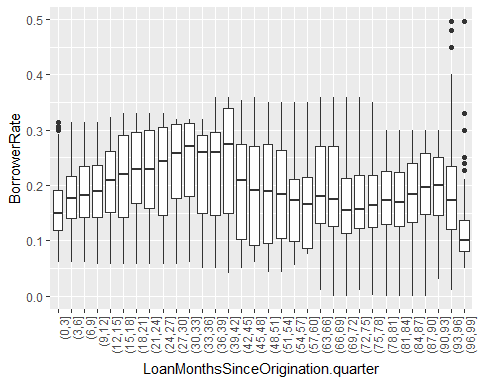
### Interest rate vs. Loan months since origination

ggplot(aes(x = LoanMonthsSinceOrigination, y = BorrowerRate), data = LoanData)+  
 geom\_point(alpha =1/20)



'BorrowerRate' and 'LoanMonthsSinceOrigination' are correlated as 0.257. There seem to be some longitudinal patterns for interest rates. The patterns are possibly from the factors not included in this dataset such as government's directives or the market.

# Make buckets for LoanMonthsSinceOrigination  
LoanData$LoanMonthsSinceOrigination.quarter <-   
 cut(LoanData$LoanMonthsSinceOrigination, breaks = seq(0,100, by=3))  
  
qplot(x = LoanMonthsSinceOrigination.quarter, y = BorrowerRate,  
 data = subset(LoanData,!is.na(LoanMonthsSinceOrigination.quarter)),   
 geom = 'boxplot') +  
 #vertical x-axis labels  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))

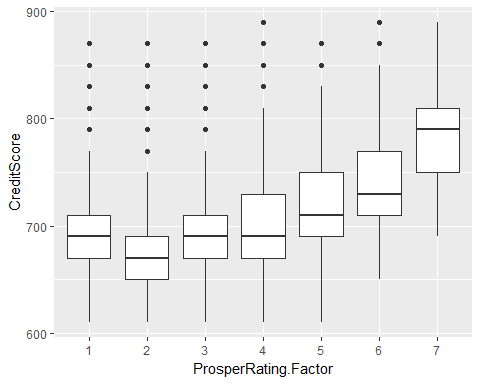


I made a bucket variable 'LoanMonthsSinceOrigination.quarter' from 'LoanMonthsSinceOrigination' to see the patterns in interest rates over time we saw in the previous graph. Each bucket is 3 months long (a quarter year). This plot shows how median interest rates and their variances for each quarter change over time. The boxes (interquartile ranges) seem to be wider for the months in the middle, but the overall ranges tend to be larger for older loans, which have longer whiskers. There were outliers only during the first and oldest two quarters.

Now I will investigate relationships between supporting variables.

### Prosper Rating vs. Credit Score

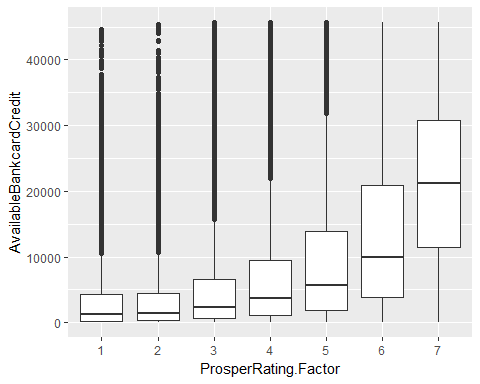
qplot(data = subset(LoanData,!is.na(ProsperRating.Factor)),   
 x=ProsperRating.Factor, y=CreditScore, geom = 'boxplot')



Credit scores tend to increase as a Prosper rating increases (i.e., as the level of loan risk decreases). I wonder what makes the revered order of credit scores for Prosper rating 1 and 2.

### Prosper Rating vs. Bank card credit

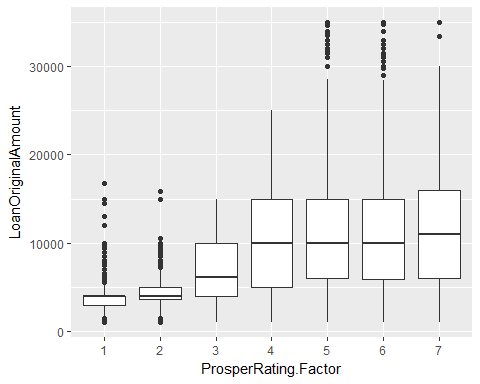
qplot(data = subset(LoanData,!is.na(ProsperRating.Factor)),   
 x=ProsperRating.Factor, y=AvailableBankcardCredit, geom = 'boxplot') +  
 scale\_y\_continuous(  
 limits=c(0,quantile(LoanData\_noNA\_BankCredit$AvailableBankcardCredit,.95)))



The points with the top 5% of bank card credit are removed in this box plot to see the boxes better. This shows bank card credits for those with higher Prosper ratings are larger as expected. The box heights for bank card credits increase as a Prosper rating increases. This could be because the people with higher Prosper ratings can have more options for bank card credits from low to high.

### Prosper Rating vs. Loan original amount

qplot(data = subset(LoanData,!is.na(ProsperRating.Factor)),   
 x=ProsperRating.Factor, y=LoanOriginalAmount, geom = 'boxplot')



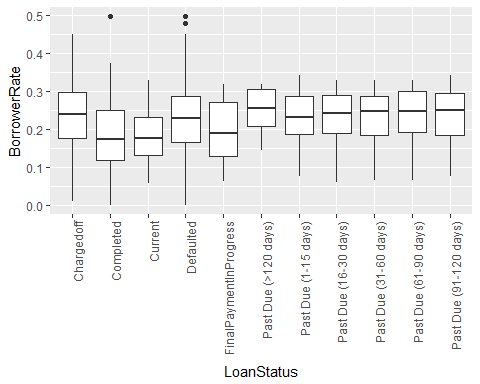
Loan original amounts for lower Prosper ratings tend to increase as Prosper rating increases, but they seem to stop increasing after Prosper rating 4. It looks loan amounts are no more restricted if a loan has a Prosper rating above the average.

The supporting variables for credit scores, bank card credits, and loan original amount correlate with Prosper ratings. Thus, they might be redundant predictor variables in the linear model between interest rates (y) and Prosper ratings (x). Adding them to the liner model as additional predictors might not help much.

Here are the additional analyses planned.

### Interest rate vs. LoanStatus

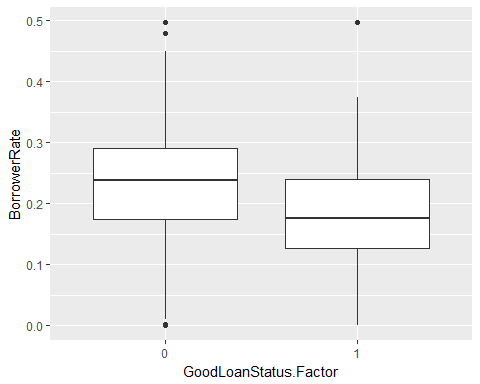
qplot(data = subset(LoanData,!is.na(LoanStatus)),   
 x=LoanStatus, y=BorrowerRate, geom = 'boxplot') +   
 # for vertical x-axis labels  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))



This graph shows loans in good loan status (i.e., Completed, Current, and FinalPaymentInProgress) had lower borrower rates than other status with some issues. The next graph with GoodLoanStatus shows this more clearly. Again, 1 stands for loans in good loan status and 0 for other status.

# making GoodLoandStatus a factor variable  
LoanData$GoodLoanStatus.Factor <- factor(LoanData$GoodLoanStatus,  
 ordered = TRUE)  
#levels(LoanData$GoodLoanStatus.Factor)

qplot(data = LoanData,   
 x=GoodLoanStatus.Factor, y=BorrowerRate, geom = 'boxplot')



### Interest rate vs. Occupation

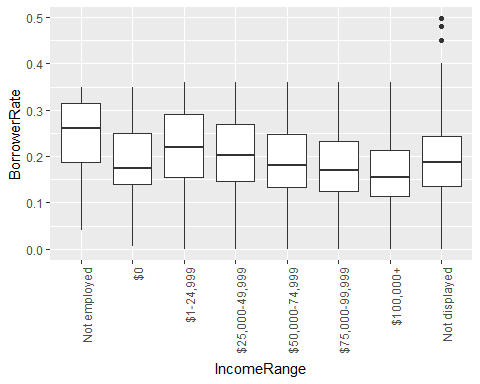
# Order occupations of borrowers by mean interest rates   
BorrowerRateByOccupation <- LoanData %>%  
 group\_by(Occupation) %>%  
 summarise(Mean\_BorrowerRate = mean(BorrowerRate)) %>%  
 arrange(Mean\_BorrowerRate)  
  
head(BorrowerRateByOccupation,10)

## # A tibble: 10 × 2  
## Occupation Mean\_BorrowerRate  
## <fctr> <dbl>  
## 1 Judge 0.1518864  
## 2 Doctor 0.1606737  
## 3 Pharmacist 0.1640292  
## 4 Engineer - Chemical 0.1669853  
## 5 Computer Programmer 0.1679992  
## 6 Attorney 0.1680181  
## 7 Pilot - Private/Commercial 0.1686739  
## 8 Engineer - Electrical 0.1692284  
## 9 Scientist 0.1704323  
## 10 Professor 0.1706077

This table shows the top 10 occupations of borrowers with the lowest interest rates.

### Interest rate vs. Income range

qplot(data = LoanData,   
 x=IncomeRange, y=BorrowerRate, geom = 'boxplot') +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))



by(LoanData$BorrowerRate, LoanData$IncomeRange, summary)

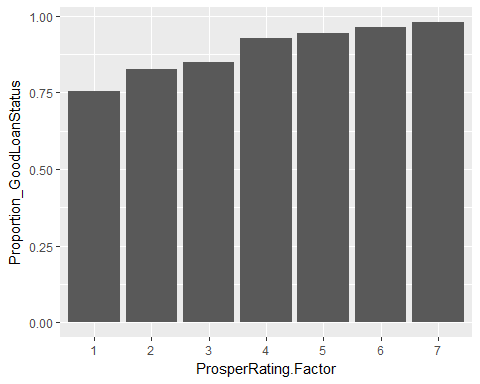
## LoanData$IncomeRange: Not employed  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0400 0.1874 0.2600 0.2467 0.3149 0.3500   
## --------------------------------------------------------   
## LoanData$IncomeRange: $0  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0050 0.1400 0.1750 0.1952 0.2500 0.3500   
## --------------------------------------------------------   
## LoanData$IncomeRange: $1-24,999  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.1550 0.2199 0.2206 0.2900 0.3600   
## --------------------------------------------------------   
## LoanData$IncomeRange: $25,000-49,999  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.1474 0.2015 0.2072 0.2684 0.3600   
## --------------------------------------------------------   
## LoanData$IncomeRange: $50,000-74,999  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.1334 0.1800 0.1903 0.2487 0.3600   
## --------------------------------------------------------   
## LoanData$IncomeRange: $75,000-99,999  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.1239 0.1699 0.1809 0.2321 0.3600   
## --------------------------------------------------------   
## LoanData$IncomeRange: $100,000+  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.1139 0.1550 0.1692 0.2124 0.3600   
## --------------------------------------------------------   
## LoanData$IncomeRange: Not displayed  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.1350 0.1875 0.1892 0.2445 0.4975

The graph shows interest rates tend to be lower for those with higher incomes with one exception; The group with $0 does not have the highest interest rates. The mean interest rate of the $0 group is lower than that of the groups with income ranges $1-24,999 and 25,000-49,999. The median interest rate of the group is even lower than that of the group with income range $50,000-74,999. I wonder what factors make this exception. The borrowers not employed tend to have the highest interest rates.

Finally, I will check how the proportion of good loan status changes over different Prosper rating and interest rates.

### Prosper rating vs. Good loan status

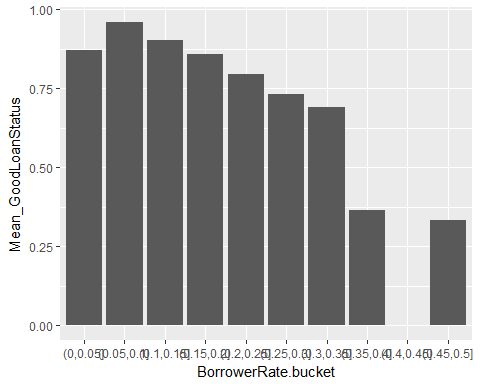
# Calculate the proportion of good loan status for each Prosper rating  
LoanData\_by\_ProsperRating <- group\_by(  
 subset(LoanData, !is.na(ProsperRating.Factor)), ProsperRating.Factor)  
  
LoanData\_meanGLS\_by\_ProsperRating <-   
 summarise(LoanData\_by\_ProsperRating,  
 Proportion\_GoodLoanStatus = mean(GoodLoanStatus))  
  
ggplot(LoanData\_meanGLS\_by\_ProsperRating,   
 aes(x=ProsperRating.Factor, y =Proportion\_GoodLoanStatus)) +   
 geom\_bar(stat="identity")



This barplot shows how the proportion of good loan status (i.e., completed, current, and final payment in progress) increases as Prosper ratings increases. The proportion of good status is about 75% for Prosper rating 1 and it increases up to 98% for the highest Prosper rating (7). This suggests that the level of loan risk was well rated using Prosper ratings.

### Interest rate vs. Good loan status

# Make bucket variable from borrower rates  
LoanData$BorrowerRate.bucket <- cut(LoanData$BorrowerRate,   
 breaks = seq(0,5, by=.05))  
  
# Calculate the proportion of good loan status for each bucket  
LoanData\_by\_BorrowerRate.bucket <- group\_by(  
 subset(LoanData,!is.na(BorrowerRate.bucket)), BorrowerRate.bucket)  
LoanData\_meanGLS\_by\_BorrowerRate.bucket <-   
 summarise(LoanData\_by\_BorrowerRate.bucket,  
 Mean\_GoodLoanStatus = mean(GoodLoanStatus))  
  
ggplot(LoanData\_meanGLS\_by\_BorrowerRate.bucket,   
 aes(x=BorrowerRate.bucket, y =Mean\_GoodLoanStatus)) +   
 geom\_bar(stat="identity")



Interest rates for loans are also related to the proportion of good loan status. Loans are more likely to be in good status if interest rates of the loans are lower with only one exception; the group with the lowest range of interest rates does not have the highest proportion of good loan status. The group has lower proportion of good loan status than the groups with higher interest ranges (0.05, 0.10] and (0.15, 0.20]. I wonder if this exception is related to the exception I found in the analysis Interest rate vs. Income range (the group with $0 income does not tend to have a higher interest rate than the groups with higher income ranges). There were also loans with low credit scores and bank card credits, but with very low interest rates (see those points on the graphs in the analyses for Interest rate vs. Credit score and Interest rate vs. Bank card credit). I will investigate this in the multivariate section.

# 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?

The main feature of interest 'BorrowerRate' (interest rate of loan) is highly correlated with Prosper ratings. The interest rate of a loan tends to decrease as a Prosper rating increases. In other words, a loan with less risk is likely to get a lower interest rate.

Moreover, an interest rate of a loan tends to decrease as credit scores and bank card credits of the borrower, and the original amount of loan increase.

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

Credit scores and bank card credits of borrowers are likely to be higher if their loans have higher Prosper ratings as expected. Loan original amounts seem to increase as Prosper ratings increase, but having more than Prosper rating 4 did not seem to help borrowing more money.

The more interesting relationships I found were those points with very low interest rates that do not follow the overall patterns mentioned above. These points have low credit scores and bank card credits. I wonder these points made the two exceptions I found. First exception was the group with the lowest range of interest rates (0, 0.05] that did not have the highest proportion of good loan status. The second exception was the group with $0 income that did not have the higher interest rates than the groups with higher income ranges.

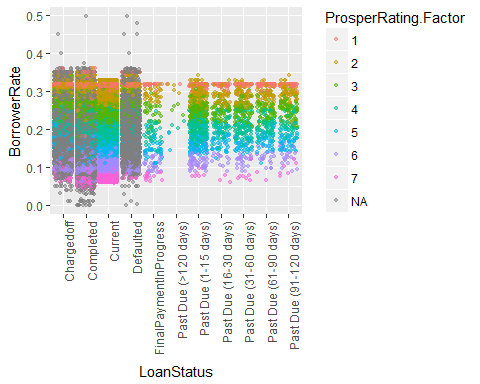
### What was the strongest relationship you found?

The strongest relationship I found was between interest rates and Prosper ratings. Their correlation is -0.95. Their boxplot showed the median of interest rates strictly decreases as Prosper ratings increase and the boxes of different Prosper ratings never overlap with each other.

# Multivariate Plots Section

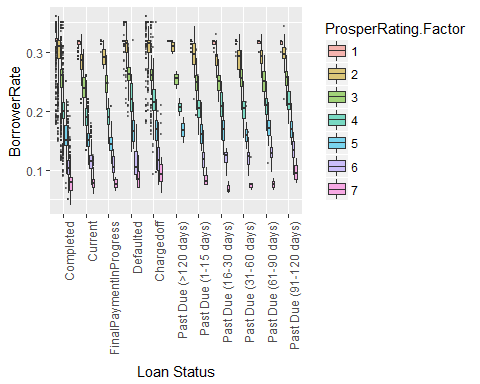
### Borrower rate vs. Loan status & Prosper rating

ggplot(aes(x = LoanStatus, y = BorrowerRate), data = LoanData) +  
 geom\_point(alpha =.5, size = 1, position='jitter',   
 aes(color=ProsperRating.Factor)) +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))



This graph shows again how strong relationship borrower rates and Prosper ratings have. I also noticed that NA Prosper ratings (grey) were only in charged off, completed, and defaulted loan status. Thus, I checked the description for the Prosper rating and realized that the Prosper rating is applicable only for loans originated after July 2009.

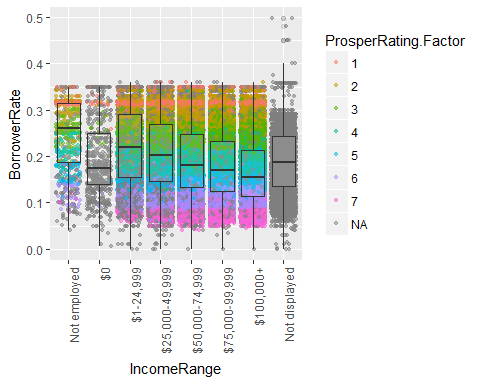
#levels(LoanData$LoanStatus)  
  
ggplot(aes(x = factor(LoanStatus,  
 levels =   
 levels(LoanData$LoanStatus)[c(3,4,6,5,2,7,8,9,10,11,12)]),   
 y = BorrowerRate,   
 fill= ProsperRating.Factor),   
 data = subset(LoanData, !is.na(ProsperRating.Factor))) +  
 geom\_boxplot(alpha =.5, outlier.size =.5) +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1)) +  
 xlab('Loan Status')



I created another plot with the same variables as the previous graph. I made boxplots to separate loans with different Prosper ratings. I also changed the order of factor levels for loan status and removed points with NAs in Propser ratings. This shows that interest rates tend to be lower for the good loan statuses, completed, current, and final payment in progress for the same prosper ratings, but the differences do not seem to be significant (i.e., no significant interaction between loan status and Prosperratings is evident in this graph).

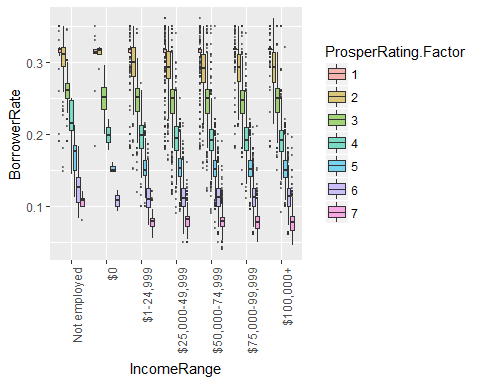
### Borrower rate vs. Income range s & Prosper rating

ggplot(aes(x = IncomeRange, y = BorrowerRate), data = LoanData) +  
 geom\_point(alpha =.5, size = 1, position='jitter',   
 aes(color=ProsperRating.Factor)) +  
 geom\_boxplot(alpha = .1) + # Add boxplot   
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))



This graph also shows the strong relationship between borrower rates and Prosper ratings and what possibly went wrong with the data. The majority of loans in the category $0 and "Not displayed" are the loans with NA Prosper ratings. These NA points seem to make those exceptions I have seen in the bivariate analyses.

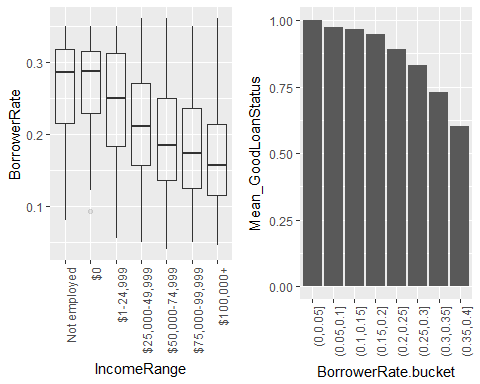
ggplot(aes(x = IncomeRange, y = BorrowerRate,   
 fill= ProsperRating.Factor),   
 data = subset(LoanData, !is.na(ProsperRating.Factor))) +  
 geom\_boxplot(alpha =.5, outlier.size =.5) +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))



I created another plot with the same variables as the previous graph. I made boxplots to separate loans with different Prosper ratings. I also removed points with NAs in Prosperratings. The interest rates in the same Prosperratings do not seem to change much as income ranges increase or decrease. i.e., No significant interaction between income ranges and Prosper ratings is evident in this graph

### Does removing those points with NA Prosper ratings remove the exceptions?

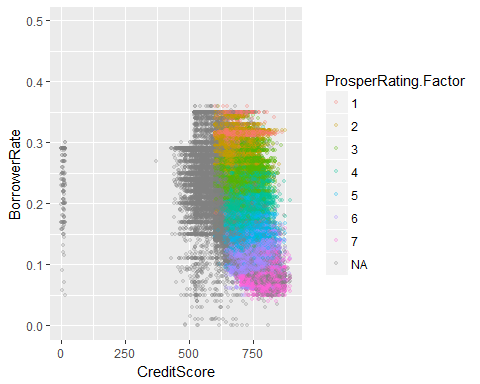
# Borrower rate vs. Income range  
p1 = ggplot(aes(x = IncomeRange, y = BorrowerRate),   
 data = subset(LoanData, !is.na(ProsperRating.Factor))) +  
 geom\_boxplot(alpha = .1) +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))  
  
LoanData$BorrowerRate.bucket <- cut(LoanData$BorrowerRate,   
 breaks = seq(0,5, by=.05))  
  
LoanData\_by\_BorrowerRate.bucket <- group\_by(  
 subset(subset(LoanData,!is.na(ProsperRating.Factor)),  
 !is.na(BorrowerRate.bucket)), BorrowerRate.bucket)  
LoanData\_meanGLS\_by\_BorrowerRate.bucket <-   
 summarise(LoanData\_by\_BorrowerRate.bucket,   
 Mean\_GoodLoanStatus = mean(GoodLoanStatus))  
  
# Proportions of good loan status for each borrower rate bucket   
p2 = ggplot(LoanData\_meanGLS\_by\_BorrowerRate.bucket,   
 aes(x=BorrowerRate.bucket, y =Mean\_GoodLoanStatus)) +   
 geom\_bar(stat="identity") +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))  
  
# Plot them together  
grid.arrange(p1,p2, ncol=2)



Yes!!! Removing those points removed the exceptions we have seen in the bivariate analysis.

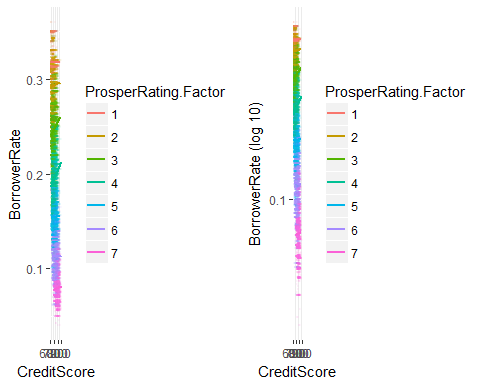
### Borrower rate vs. Credit Score s & Prosper rating

ggplot(aes(x = CreditScore, y = BorrowerRate), data = LoanData) +  
 geom\_point(alpha =.2, size = 1, position='jitter',   
 aes(color=ProsperRating.Factor))



This graph also shows the strong relationship between interest rates and Prosper ratings. As it was shown before, this also shows the negative relationships between interest rates and credit scores. Those points with low credits, but with very low interest rates are indeed NA Prosper rating points (grey).

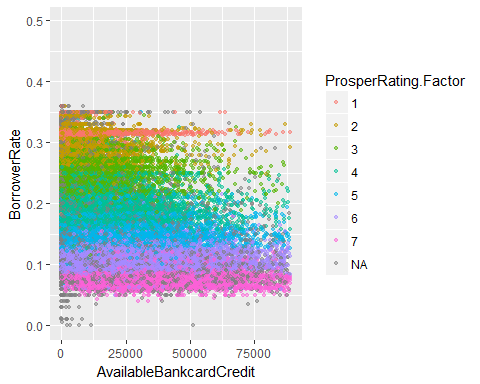
#Borrower rate vs. Credit score (without NAs in Prosper ratings)  
p1 = ggplot(aes(x = CreditScore, y = BorrowerRate,color=ProsperRating.Factor),   
 data = subset(LoanData, !is.na(ProsperRating.Factor))) +  
 geom\_point(alpha =1/20, size = .5, position='jitter') +  
 geom\_smooth(method = "lm", se = FALSE,size=1)  
  
#Borrower rate (log10) vs. Credit score (without NAs in Prosper ratings)  
p2 = ggplot(aes(x = CreditScore, y = BorrowerRate,color=ProsperRating.Factor),   
 data = subset(LoanData, !is.na(ProsperRating.Factor))) +  
 geom\_point(alpha =1/20, size = .5, position='jitter') +  
 scale\_y\_continuous(trans = log10\_trans(), name = 'BorrowerRate (log 10)') +  
 geom\_smooth(method = "lm", se = FALSE,size=1)  
  
grid.arrange(p1,p2, ncol=2)



I removed the points with NA Prosper ratings from the previous graph (left) and made another graph with log 10 of interest rates (right) for the better linear model (as found in the bivariate analysis).

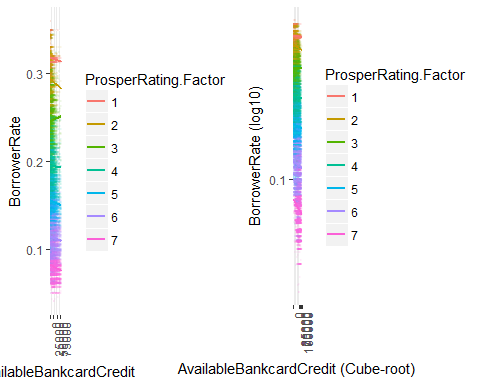
### Borrower rate vs. Bank card credits s & Prosper rating

ggplot(aes(x = AvailableBankcardCredit, y = BorrowerRate),   
 data = LoanData) +  
 geom\_point(alpha =0.5, size = 1, position='jitter',   
 aes(color=ProsperRating.Factor)) +  
 scale\_x\_continuous(limits =  
 c(0, quantile(LoanData\_noNA\_BankCredit$AvailableBankcardCredit,.99)))



As we have seen before in the bivariate analysis, this graph shows that interest rates and available bank card credits are related. This again also shows that points are ordered by Prosper ratings and those points with low bank card credits and low interest rates are those with NA Prosper rating points (grey).

# Borrower rate vs. Bank card credits (without NAs in Prosper ratings)  
p1 = ggplot(aes(x = AvailableBankcardCredit, y = BorrowerRate,  
 color=ProsperRating.Factor),   
 data = subset(LoanData,!is.na(ProsperRating.Factor))) +  
 geom\_point(alpha =1/20, size = .5, position='jitter') +  
 scale\_x\_continuous(limits =  
 c(0, quantile(LoanData\_noNA\_BankCredit$AvailableBankcardCredit,.99)))+  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))+  
 geom\_smooth(method = "lm", se = FALSE,size=1)  
  
# Borrower rate (log10) vs. Bank card credits (cube-root)  
# without NAs in Prosper ratings  
p2 = ggplot(aes(x = AvailableBankcardCredit, y = BorrowerRate,  
 color=ProsperRating.Factor),   
 data = subset(LoanData,!is.na(ProsperRating.Factor))) +  
 geom\_point(alpha =1/20, size = .5, position='jitter') +  
 scale\_x\_continuous(trans = cuberoot\_trans(), limits =  
 c(0, quantile(LoanData\_noNA\_BankCredit$AvailableBankcardCredit,.99)),  
 name = "AvailableBankcardCredit (Cube-root)" ) +  
 scale\_y\_continuous(trans = log10\_trans() , name = "BorrowerRate (log10)")+  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))+  
 geom\_smooth(method = "lm", se = FALSE,size=1)  
  
  
grid.arrange(p1,p2, ncol=2)



I removed the points with NA Prosper ratings from the previous graph (left) and made another graph with log 10 of interest rates and cube-root of bank card credits (right) for the better linear model (as found in the bivariate analysis) .

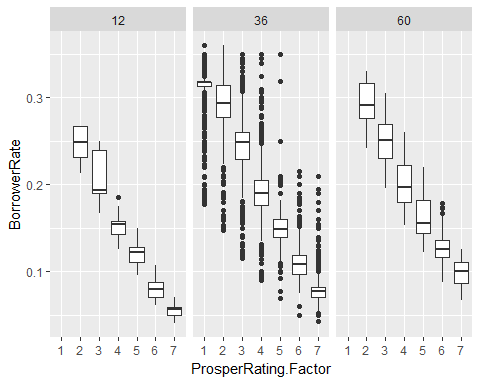
The almost flat color stripes ordered by Prosper ratings are in the previous graphs for both "Borrower rate vs. credit scores & Prosper rating" and "Borrower rate vs. Bank card credits s & Prosper rating". The horizontal color stripes and flat regression lines show two important things:

* Prosper ratings explain the variance in interest rates very well.
* Both credit scores and bank card credits do not seem to explain the additional variance in interest rates.

These findings will be checked in the linear model section.

### Borrower rate vs. Prosper rating: Facet wrap by Term

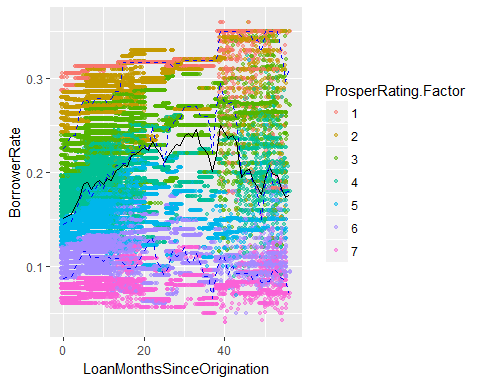
qplot(data = subset(LoanData,!is.na(ProsperRating.Factor)),   
 x=ProsperRating.Factor, y=BorrowerRate, geom = 'boxplot') +  
 facet\_wrap(~Term)



All of the 3 graphs facet wrap by loan terms (12, 36, or 60 months) show the strong negative relationship between interest rates and Prosper ratings. However, they have somewhat different patterns. The interest rates for the 36 month term are more scattered and have many more outliers than those for the 12 and 60 month terms. Moreover, no loans with the lowest Prosper rating (1) had the 12 or 60 month terms. If a loan term is shorter, the interest rate tends to be lower for a given Prosper rating.

### Borrower rate vs. Months Since Origination & Prosper rating

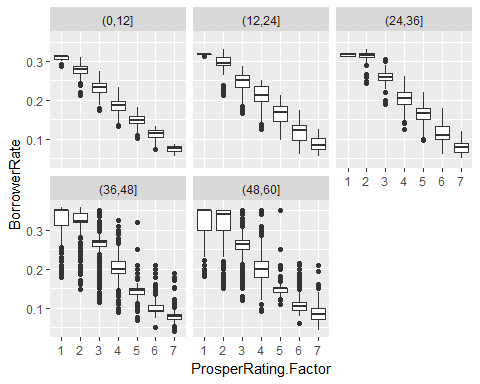
ggplot(aes(x = LoanMonthsSinceOrigination, y = BorrowerRate),   
 data = subset(LoanData,!is.na(ProsperRating.Factor))) +  
 geom\_point(alpha =0.5, size = 1, position='jitter',   
 aes(color=ProsperRating.Factor)) +  
 #Add summary lines  
 geom\_line(stat='summary', fun.y=mean) +  
 geom\_line(stat='summary', fun.y=quantile, fun.args = list(probs = .1),   
 linetype = 2, color='blue') +  
 geom\_line(stat='summary', fun.y=quantile, fun.args = list(probs = .5),  
 linetype = 2, color='blue') +  
 geom\_line(stat='summary', fun.y=quantile, fun.args = list(probs = .9),  
 linetype = 2, color='blue')



This graph shows that overall interest rates and their patterns changed over time. The black line is connecting the mean interest rates for each month. The 3 dotted blue lines are representing 10, 50 (median), 90 percentiles of interest rates for each month. The newer loans have smaller ranges of interest rates, which are more systematically ordered by Prosper ratings. The loans older than 40 months have more scattered interest rates with mixed orders of Prosper ratings.

### Borrower rate vs. Prosper rating: Facet wrap by LoanMonthsSinceOrigination

# Make buckets for LoanMonthsSinceOrigination  
LoanData$LoanMonthsSinceOrigin.bucket <-   
 cut(LoanData$LoanMonthsSinceOrigination, breaks = seq(0,60, by=12))  
  
qplot(data = subset(LoanData,!is.na(ProsperRating.Factor)&  
 !is.na(LoanMonthsSinceOrigin.bucket)),   
 x=ProsperRating.Factor, y=BorrowerRate, geom = 'boxplot') +  
 facet\_wrap(~LoanMonthsSinceOrigin.bucket)



The interest rate vs. Prosper rating graphs are separated by 12 months of loan durations since origination. These support what I found in the previous graph. The recent loans have interest rates more strictly decided by Prosper ratings, but the older loans (over 36 months) have much more outliers and interest rates are much more overlapping between different Prosper ratings.

### Linear Models

We have seen relationships between interest rates of loans and many other variables. We also found better transformations that work for each pair of variables. Using these findings, I will make linear models that predict interest rates of loans.

m1 <- lm(I(log10(BorrowerRate)) ~ ProsperRating.Factor,   
 data = LoanData) # NAs are automatically omitted  
m2 <- update(m1, ~ . + I(AvailableBankcardCredit^(1/3)))  
m3 <- update(m2, ~ . + I(LoanOriginalAmount^(1/2)))  
m4 <- update(m3, ~ . + I(CreditScore))  
  
library(memisc) # This is loaded here because it conflicts with dplyr select()  
mtable(m1, m2, m3, m4, sdigits = 4)

##   
## Calls:  
## m1: lm(formula = I(log10(BorrowerRate)) ~ ProsperRating.Factor, data = LoanData)  
## m2: lm(formula = I(log10(BorrowerRate)) ~ ProsperRating.Factor +   
## I(AvailableBankcardCredit^(1/3)), data = LoanData)  
## m3: lm(formula = I(log10(BorrowerRate)) ~ ProsperRating.Factor +   
## I(AvailableBankcardCredit^(1/3)) + I(LoanOriginalAmount^(1/2)),   
## data = LoanData)  
## m4: lm(formula = I(log10(BorrowerRate)) ~ ProsperRating.Factor +   
## I(AvailableBankcardCredit^(1/3)) + I(LoanOriginalAmount^(1/2)) +   
## I(CreditScore), data = LoanData)  
##   
## ============================================================================================  
## m1 m2 m3 m4   
## --------------------------------------------------------------------------------------------  
## (Intercept) -0.748\*\*\* -0.742\*\*\* -0.754\*\*\* -0.826\*\*\*   
## (0.000) (0.000) (0.001) (0.004)   
## ProsperRating.Factor: .L -0.542\*\*\* -0.537\*\*\* -0.542\*\*\* -0.548\*\*\*   
## (0.001) (0.001) (0.001) (0.001)   
## ProsperRating.Factor: .Q -0.099\*\*\* -0.097\*\*\* -0.095\*\*\* -0.098\*\*\*   
## (0.001) (0.001) (0.001) (0.001)   
## ProsperRating.Factor: .C 0.005\*\*\* 0.006\*\*\* 0.007\*\*\* 0.007\*\*\*   
## (0.001) (0.001) (0.001) (0.001)   
## ProsperRating.Factor: ^4 -0.011\*\*\* -0.010\*\*\* -0.011\*\*\* -0.012\*\*\*   
## (0.001) (0.001) (0.001) (0.001)   
## ProsperRating.Factor: ^5 0.005\*\*\* 0.005\*\*\* 0.004\*\*\* 0.005\*\*\*   
## (0.000) (0.000) (0.000) (0.000)   
## ProsperRating.Factor: ^6 0.007\*\*\* 0.007\*\*\* 0.008\*\*\* 0.007\*\*\*   
## (0.000) (0.000) (0.000) (0.000)   
## I(AvailableBankcardCredit^(1/3)) -0.000\*\*\* -0.000\*\*\* -0.001\*\*\*   
## (0.000) (0.000) (0.000)   
## I(LoanOriginalAmount^(1/2)) 0.000\*\*\* 0.000\*\*\*   
## (0.000) (0.000)   
## I(CreditScore) 0.000\*\*\*   
## (0.000)   
## --------------------------------------------------------------------------------------------  
## R-squared 0.9119 0.9121 0.9126 0.9130   
## adj. R-squared 0.9119 0.9121 0.9126 0.9130   
## sigma 0.0538 0.0537 0.0535 0.0534   
## F 146305.3822 125733.7822 110769.6051 98987.1585   
## p 0.0000 0.0000 0.0000 0.0000   
## Log-likelihood 127642.3180 127744.2011 128008.5274 128215.0478   
## Deviance 245.2423 244.6541 243.1346 241.9540   
## AIC -255268.6361 -255470.4022 -255997.0547 -256408.0957   
## BIC -255193.8467 -255386.2642 -255903.5680 -256305.2602   
## N 84853 84853 84853 84853   
## ============================================================================================

As expected (mentioned above), adding the variables, bank card credits, credits scores, and loan original amounts do not help the linear model a lot; the improvement made by each additional variable, is less than 0.0005 in R-squared. Prosper ratings, available bank card credits, credit scores and original loan amounts seem to explain the similar variance of the interest rates. Thus, I tried other variables that are related to interest rates in different ways to explain other kinds of the variance in the interest rates.

m1 <- lm(BorrowerRate~ ProsperRating.Factor,   
 data = LoanData) # NAs are automatically omitted  
m2 <- update(m1, ~ . + Term.Factor)  
m3 <- update(m2, ~ . + LoanMonthsSinceOrigin.bucket)  
  
mtable(m1, m2, m3, sdigits = 4)

##   
## Calls:  
## m1: lm(formula = BorrowerRate ~ ProsperRating.Factor, data = LoanData)  
## m2: lm(formula = BorrowerRate ~ ProsperRating.Factor + Term.Factor,   
## data = LoanData)  
## m3: lm(formula = BorrowerRate ~ ProsperRating.Factor + Term.Factor +   
## LoanMonthsSinceOrigin.bucket, data = LoanData)  
##   
## ==========================================================================================  
## m1 m2 m3   
## ------------------------------------------------------------------------------------------  
## (Intercept) 0.200\*\*\* 0.190\*\*\* 0.177\*\*\*   
## (0.000) (0.000) (0.000)   
## ProsperRating.Factor: .L -0.221\*\*\* -0.221\*\*\* -0.215\*\*\*   
## (0.000) (0.000) (0.000)   
## ProsperRating.Factor: .Q 0.000 0.004\*\*\* -0.001\*\*\*   
## (0.000) (0.000) (0.000)   
## ProsperRating.Factor: .C 0.014\*\*\* 0.015\*\*\* 0.015\*\*\*   
## (0.000) (0.000) (0.000)   
## ProsperRating.Factor: ^4 -0.007\*\*\* -0.008\*\*\* -0.008\*\*\*   
## (0.000) (0.000) (0.000)   
## ProsperRating.Factor: ^5 0.003\*\*\* 0.002\*\*\* 0.005\*\*\*   
## (0.000) (0.000) (0.000)   
## ProsperRating.Factor: ^6 0.003\*\*\* 0.003\*\*\* 0.001\*\*\*   
## (0.000) (0.000) (0.000)   
## Term.Factor: .L 0.032\*\*\* 0.041\*\*\*   
## (0.000) (0.000)   
## Term.Factor: .Q -0.010\*\*\* -0.012\*\*\*   
## (0.000) (0.000)   
## LoanMonthsSinceOrigin.bucket: (12,24]/(0,12] 0.019\*\*\*   
## (0.000)   
## LoanMonthsSinceOrigin.bucket: (24,36]/(0,12] 0.023\*\*\*   
## (0.000)   
## LoanMonthsSinceOrigin.bucket: (36,48]/(0,12] 0.021\*\*\*   
## (0.000)   
## LoanMonthsSinceOrigin.bucket: (48,60]/(0,12] 0.018\*\*\*   
## (0.000)   
## ------------------------------------------------------------------------------------------  
## R-squared 0.9138 0.9224 0.9396   
## adj. R-squared 0.9138 0.9224 0.9396   
## sigma 0.0219 0.0208 0.0184   
## F 149953.3005 126057.6550 107681.2500   
## p 0.0000 0.0000 0.0000   
## Log-likelihood 203812.3479 208257.9474 214101.3735   
## Deviance 40.7276 36.6760 27.9929   
## AIC -407608.6958 -416495.8948 -428174.7470   
## BIC -407533.9064 -416402.4080 -428044.1695   
## N 84853 84853 83031   
## ==========================================================================================

I tried several models with different combinations of variables and m3 here is the one of the best models without too many predictor variables. The three variables in m3 account for 93.96% of the variance in the interest rates of loans. Adding the variables used in the previous table or other variables can improve the model only slightly.

# 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?

* Prosper ratings (1-7) and loan terms (12, 36, or 60 months) strengthened each other in terms of predicting interest rates. If a loan term is shorter, the interest rate tends to be lower for a given Prosper rating.
* Prosper ratings and the number of months since origination of loan also strengthened each other in terms of predicting interest rates since the overall interest rates changed over time.

### Were there any interesting or surprising interactions between features?

* Different loan terms have different patterns for interest rates vs. prosper ratings. The interest rates for the 36 month term are more scattered and have many more outliers than those for the 12 and 60 month terms. Moreover, no loans with the lowest Prosper rating had the 12 or 60 month terms.
* The recent loans have interest rates more strictly decided by Prosper ratings, but the older loans (over 36 months) have much more outliers. The interest rates of older loans are much more scattered and overlapping between different Prosper ratings.

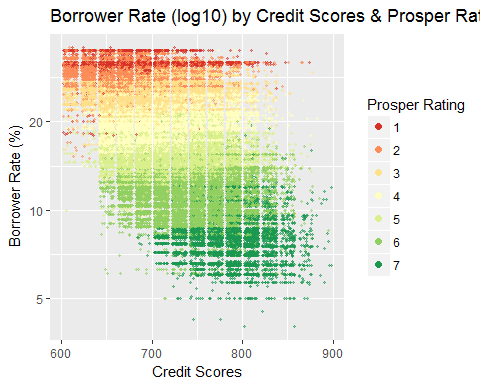
### OPTIONAL: Did you create any models with your dataset? Discuss the strengths and limitations of your model.

Yes, I tried many liner models and the final model I chose was predicting interest rates with 3 predictors, Prosper rating, loan term, and the number of months since loan origination. With only the 3 variables, the model account for 93.96% of the variance in the interest rates of loans. Adding other variables such as credit scores, bank card credits, and loan original amounts only improved the model very little, so they were omitted from the final model. There could be some other columns not included in my data frame 'LoanData' that can further improve the linear model. Moreover, we might need more information like government's directives or the market to improve the prediction on interest rates. They were not in the data set, but they might have created the patterns of the interest rates shown in the plot for "Borrower rate vs. Months Since Origination & Prosper rating".

# Final Plots and Summary

### Plot One

ggplot(aes(x = CreditScore, y = BorrowerRate\*100, color = ProsperRating.Factor),   
 data = subset(LoanData, !is.na(ProsperRating.Factor))) +  
 geom\_point(alpha =.5, size = .7, position='jitter') +  
 # Change colors  
 scale\_color\_brewer(type = 'div', palette ='RdYlGn',  
 guide = guide\_legend(title = 'Prosper Rating',  
 override.aes = list(alpha = 1, size = 2))) +  
 scale\_y\_continuous(trans = log10\_trans(), name = 'Borrower Rate (%)',  
 breaks = c(.05,.1,.2,.5)\*100) +  
 xlab('Credit Scores') +  
 ggtitle('Borrower Rate (log10) by Credit Scores & Prosper Rating')



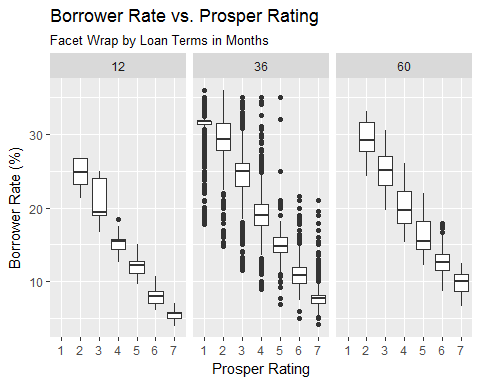
### Description One

This graph shows how interest rates of a loan change as credit scores of its borrower and Prosper rating of the loan change. First of all, this plot shows the negative relationship between interest rates and credit scores, and also the strong negative relationship between interest rates and Prosper ratings. In other words, loans with higher prosper ratings (i.e., lower risk) and higher credit scores of their borrowers are likely to receive lower interest rates.

Secondly, the horizontal color stripes ordered by Prosper ratings show that the variance in interest rates are well explained by Prosper ratings, but credit scores do not seem to explain the variance in interest rates additionally. These findings were confirmed using linear models in the model section.

### Plot Two

qplot(data = subset(LoanData,!is.na(ProsperRating.Factor)),   
 x=ProsperRating.Factor, y=BorrowerRate\*100, geom = 'boxplot') +  
 facet\_wrap(~Term) +  
 xlab('Prosper Rating') +  
 ylab('Borrower Rate (%)') +  
 ggtitle('Borrower Rate vs. Prosper Rating',  
 subtitle = 'Facet Wrap by Loan Terms in Months') +  
 # Change colors  
 scale\_color\_brewer(type = 'div', palette ='RdYlGn',  
 guide = guide\_legend(title = 'Prosper Rating',  
 override.aes = list(alpha = 1, size = 2)))



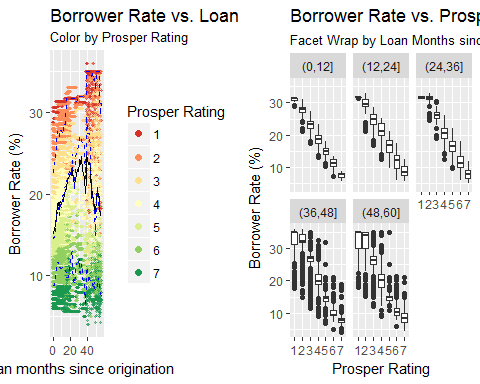
### Description Two

These graphs show the relationship between interest rates and Prosper ratings of loans for each of 12, 36, or 60 month loan terms. All of the 3 graphs again show the strong negative relationship between interest rates and Prosper ratings.

More importantly, interest rates tend to be lower for shorter loans for a given Prosper rating. There are also interesting patterns found across different loan terms. The interest rates for 36 month loans are much more varied than those for 12 or 60 month loans. Moreover, 12 or 60 month loans were not made for the lowest Prosper rating. The interactions between Prosper rating and loan terms found here strengthened each other in predicting interest rates.

### Plot Three

# Borrower Rate vs. Loan Months (Color by Prosper Rating)  
p1= ggplot(aes(x = LoanMonthsSinceOrigination, y = BorrowerRate\*100),   
 data = subset(LoanData,!is.na(ProsperRating.Factor))) +  
 geom\_point(alpha =0.5, size = .8, position='jitter',   
 aes(color=ProsperRating.Factor)) +  
 # Add summary lines  
 geom\_line(stat='summary', fun.y=mean) +  
 geom\_line(stat='summary', fun.y=quantile, fun.args = list(probs = .1),   
 linetype = 2, color='blue') +  
 geom\_line(stat='summary', fun.y=quantile, fun.args = list(probs = .5),  
 linetype = 2, color='blue') +  
 geom\_line(stat='summary', fun.y=quantile, fun.args = list(probs = .9),  
 linetype = 2, color='blue') +  
 # Change colors  
 scale\_color\_brewer(type = 'div', palette ='RdYlGn',  
 guide = guide\_legend(title = 'Prosper Rating',  
 override.aes = list(alpha = 1, size = 2))) +  
 xlab('Loan months since origination') +  
 ylab('Borrower Rate (%)') +  
 ggtitle('Borrower Rate vs. Loan Months',  
 subtitle = 'Color by Prosper Rating')  
  
# Borrower Rate vs. Prosper Rating (Facet Wrap by Loan Months)  
p2=qplot(data = subset(LoanData,!is.na(ProsperRating.Factor)&  
 !is.na(LoanMonthsSinceOrigin.bucket)),   
 x=ProsperRating.Factor, y=BorrowerRate\*100, geom = 'boxplot' ) +  
 facet\_wrap(~LoanMonthsSinceOrigin.bucket) +  
 scale\_color\_brewer(type = 'div', palette ='RdYlGn',  
 guide = guide\_legend(title = 'Prosper Rating',  
 override.aes = list(alpha = 1, size = 2)))+  
 xlab('Prosper Rating') +  
 ylab('Borrower Rate (%)') +  
 ggtitle('Borrower Rate vs. Prosper Rating',  
 subtitle = 'Facet Wrap by Loan Months since origination')  
  
grid.arrange(p1,p2, ncol = 2)



### Description Three

The left graph shows that overall changes of interest rates over time. The black line is the mean interest rate and the 3 dotted lines are 10, 50 (median), 90 percentiles of interest rates for each month. Loans were also colored by their Prosper ratings from red (rating = 1, highest risk) to green (rating = 7, lowest risk). The overall interest rates of loans fluctuate over time and their patterns also change. The newer loans have narrower ranges of interest rates and interest rates seem to be more systematically decided according to Prosper ratings. The older loans (say, over 40 months) have more scattered interest rates and their mixed colors show their interest rates are not ordered by Prosper ratings.

To see this pattern more clearly, I made the interest rate vs. Prosper rating graphs separated by 12 months of loan durations since origination (right). The interest rates of recent loans are more strictly ordered by Prosper ratings while interest rates of older loans (over 36 months) have much more outliers. These are consistent with what I found in the left graph. This multivariate analysis suggested that loan months since origination would account for the extra variability in interest rates and it was found to be true in the model section.

# Reflection

My data set contained 113,937 loans each with 81 variables and I selected 17 variables to investigate. First, I explored each variable and decided what variables to drop and made some new variables. After the univariate analysis, I decided to make the interest rate of a loan ('BorrowerRate') the main feature of interest. As expected, I found the Prosper rating that measures the level of loan risk has the strongest relationship with interest rates. Interest rates are lower if Prosper ratings are higher (i.e., if loans have lower risk). I also found credit scores, bank card credits, and original loan amounts, are correlated with interest rates, but they are also highly correlated with Prosper ratings. For this reason, these 3 variables found to be predictors redundant with Prosper rating in a linear model predicting interest rates. They did not account for extra variability in interest rates. These were the variables I first expected to support Prosper rating when predicting interest rates, so I had to find other variables that can strengthen Prosper ratings in terms of looking at interest rates. Finding out such variables were the struggles I had since most of variables I tried could not explain variabilities in interest rates that Prosper ratings cannot. However, I finally noticed loan terms and loan months since origination have interesting relationships with interest rates in the bivariate analysis. In the multivariate analysis, I also found they explain some extra variabilities in interest rates not explained by Prosper ratings alone. These findings were my successes since they were successfully confirmed by linear models. They together with Prosper rating account for almost 94% of the variance in the interest rates of loans.

It was surprising to find that the recent loans have interest rates more strictly ordered by Prosper ratings and interest rates of older loans are much more scattered and overlapping between different Prosper ratings. I wonder what would make those patterns changing over time. Did banks have more freedom to choose interest rates of loans in the past or did they consider some features more importantly than Prosper ratings when deciding interest rates?

In the beginning of my analysis, I was interested in predicting whether a loan will be in a good status. I made a variable 'GoodLoanStatus' that contains value 1 for completed, current, or with final payment in progress loans and value 0 for all other loans. Prosper ratings would be again a good predictor for this variable and I checked this using the plot for "Prosper rating vs. Good loan status" in the bivariate analysis. I showed the proportion of good loan status is higher for higher Prosper ratings (the proportion of good loan status increases up to 98% for the highest Prosper rating). We can predict a categorical variable like 'GoodLoanStatus' using logistic regressions. It would be fascinating if we can well predict whether a borrower can successfully make loan payments or not in the future using current information.