LOAN DATA EXPLORATION by SAMIN EMAMI

In this report I am going to explore the loans dataset provided by Prosper, one of the pioneers of peer to peer lending. The dataset contains 113,937 loans in the US between 2005-2014 with 81 variables. I will look at 13 of these variables in the dataset and some of the relationships between them and determine if there is any noticeable pattern that can help us understand the attributes of loans and their borrowers.

```
library(ggplot2)
library(dplyr)
library(tidyr)
library(RColorBrewer)
library(forcats) #order bar_graph by length
library(scales)   #add percent/dollar sign to x-y axis
library(maps)  #plot a map of United States
library(gridExtra)
```

```
# Load the Data
setwd("/Users/saminemami/Documents/DAND/Exploratory Data Analysis")
loans <- read.csv("prosperLoanData.csv")
```

Univariate Plots Section

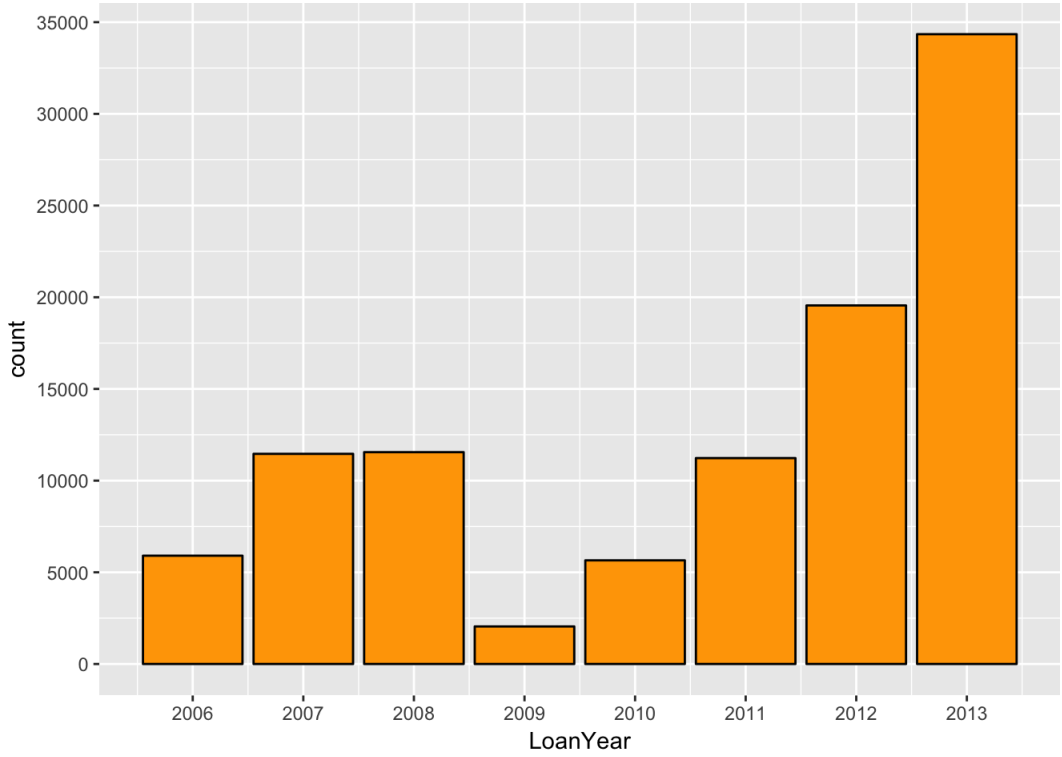
In the first section, I am going to look at the distribution of some of the individual variables.

Number of Loans per Year/Quarter

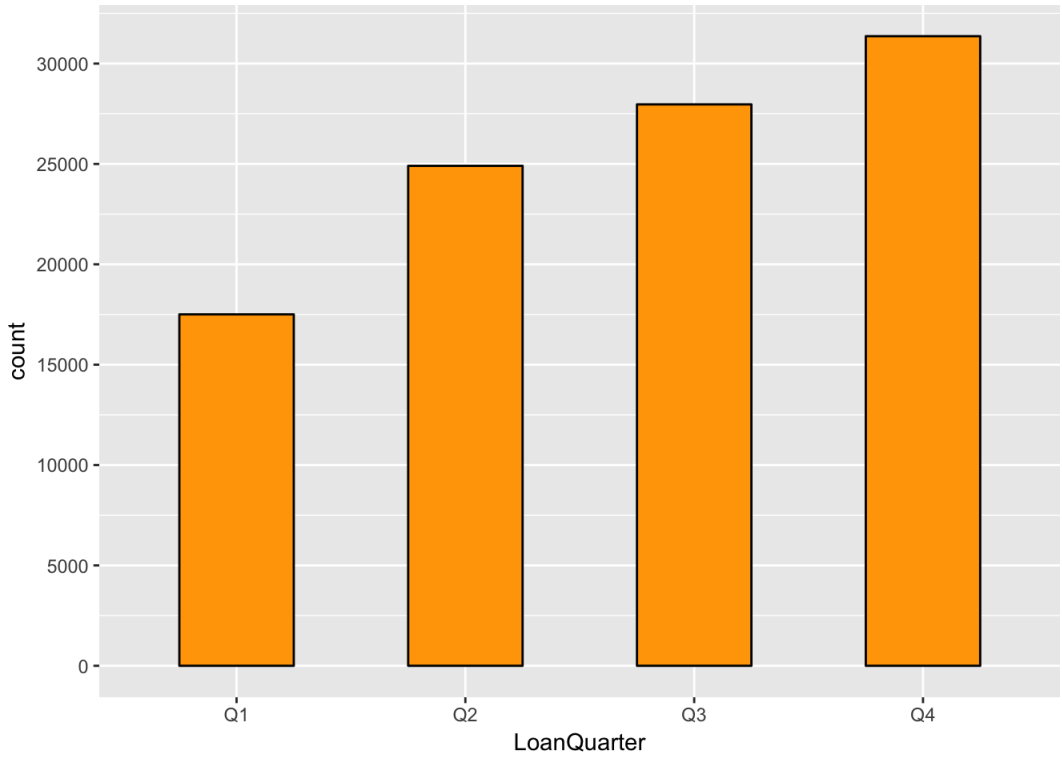
```
# split the Loan Qurater column into Year and Quarter columns
loans <- separate(loans, LoanOriginationQuarter, c("LoanQuarter", "LoanYear"),
                  sep = " ")
# convert year column from character to integer
loans$LoanYear <- as.integer(loans$LoanYear)
# finding the max and min of loan origination years
range(loans$LoanYear)
```

```
## [1] 2005 2014
```

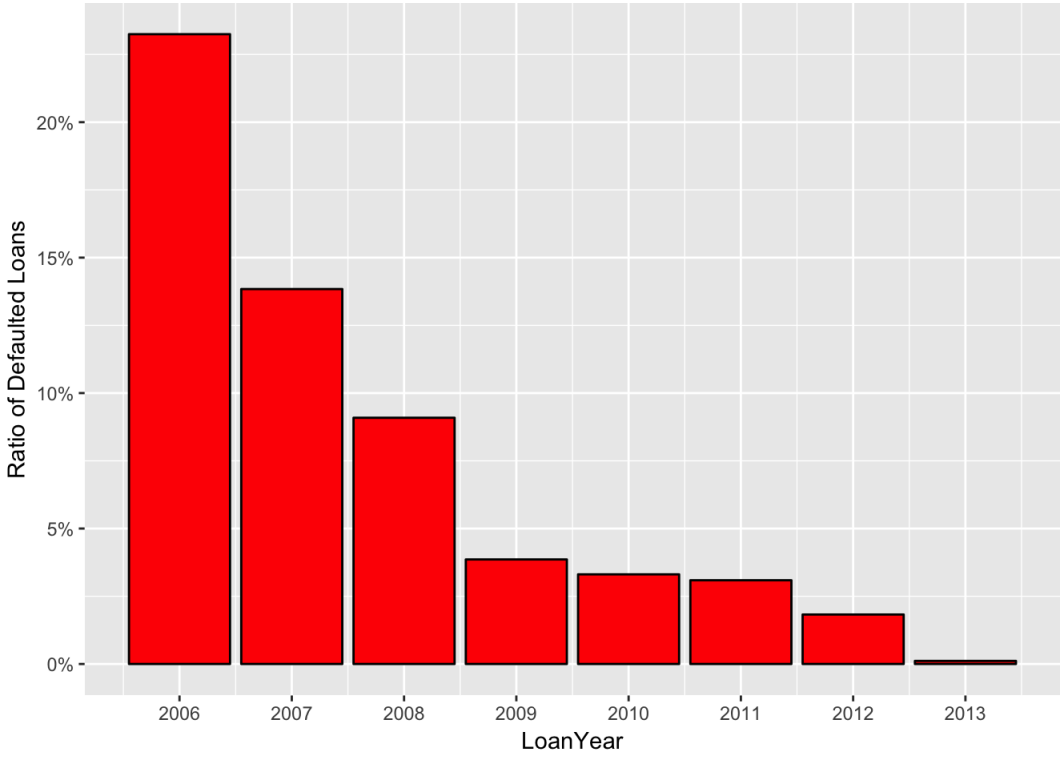
```
ggplot(aes(x=LoanYear), data = subset(loans, LoanYear > 2005 & LoanYear<2014))+
  geom_bar(color = "black", fill = "orange") +
  scale_x_continuous(breaks = seq(2006, 2013, 1)) +
  scale_y_continuous(breaks = seq(0,35000,5000))
```




```
ggplot(aes(x=LoanQuarter), data = subset(loans, LoanYear > 2005 & LoanYear<2014))+
geom_bar(width = 0.5, color = "black", fill = "orange") +
  scale_y_continuous(breaks = seq(0,35000,5000))
```




```
# Create a summary table of ratio of defaulted zones per year
Defaulted_loans <- loans %>% group_by(LoanYear) %>%
  summarise(percent_default = (sum(LoanStatus == "Defaulted")/n()))

# Plot the distribution of defaulted loans per year
ggplot(aes(x=LoanYear),
       data=subset(Defaulted_loans, LoanYear > 2005 & LoanYear < 2014))+
  geom_bar(aes(weight = percent_default), color = "black", fill = "red") +
  ylab("Ratio of Defaulted Loans") +
  scale_x_continuous(breaks = seq(2006, 2013, 1)) +
  scale_y_continuous(labels = percent)
```



Since the dataset is not complete for years 2005 and 2014, I looked at the number of loans between 2006-2013. As we can see, right after the financial crisis in 2008, the number of loans fell drastically (by almost 83%) in 2009. Since then, however, we see a nice and steady recovery year after year. In 2013 the number of loans was 16 times (!) more than 2009 which could be an indicator of a good recovery of the economy. One interesting point is the low number of loans in 2006, when the economy was doing OK (though there was a bubble) and the number of loans, one would think, should have been higher. Later in this report, we will look at the distribution of loans by cateogry (home improvement, auto, personal loan, etc.)

In the second chart, I looked at the distribution of loans across the four quraters of the year. It is interesting to see that people generally apply for more loans later in the year. They probably want to get an idea about how they are doing financially in a given year and if they are meeting their financial goals before applying for a loan.

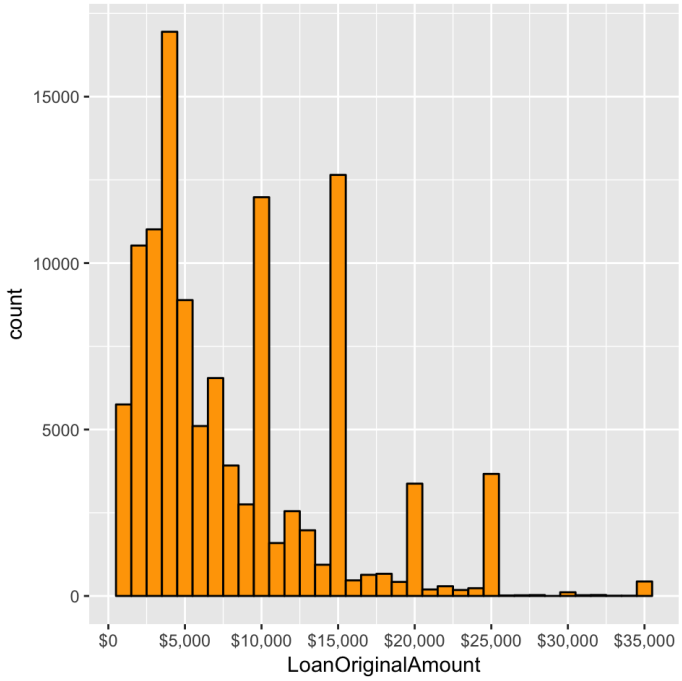
The third chart shows the percentage of "Defaulted" loans relative to total number of loans in different years. Again, as expected, we can see that the percentage of Defaulted loans was much higher during the financial crisis years and it gets lower after that. Although 2006 was before the fnancial crisis actually hit, we still see a high default rate. That could have been an indicator of something strange going on with the economy for people who pay attention!

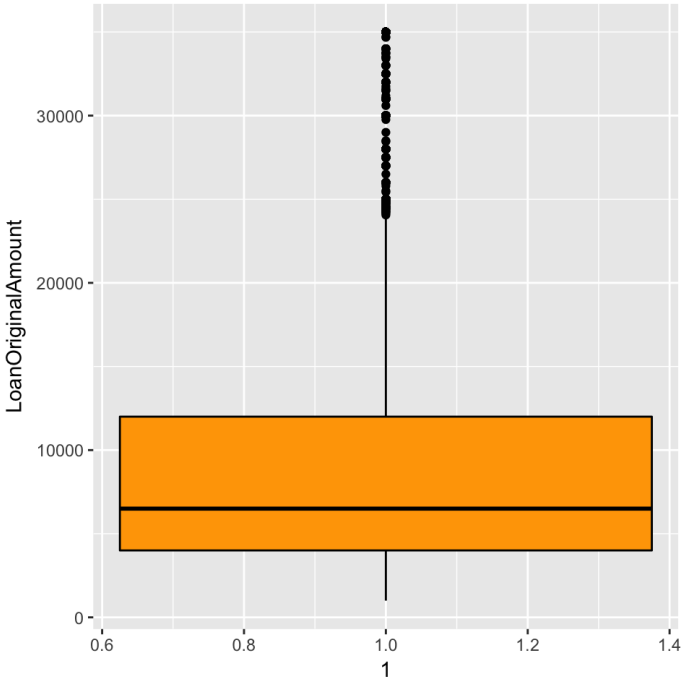
Loan Amount Distribution

```
summary(loans$LoanOriginalAmount)
```

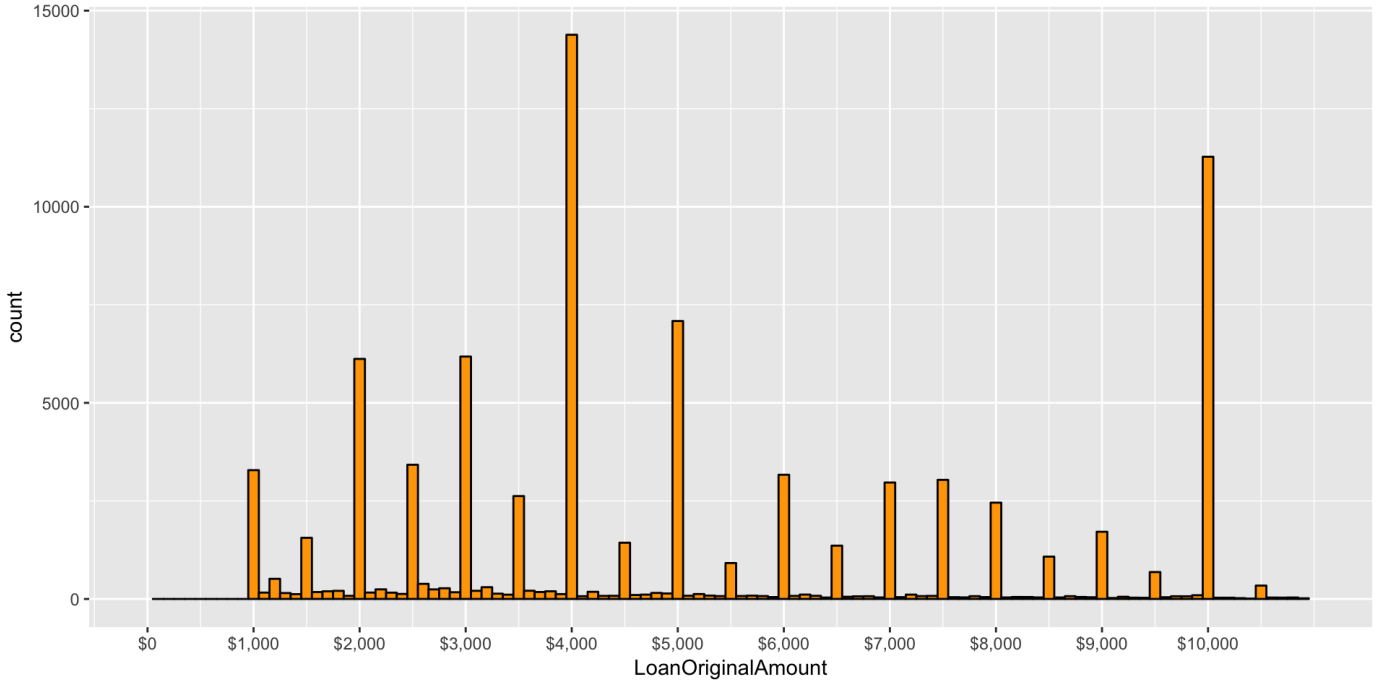
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1000    4000    6500    8337   12000   35000
```

```
p1 <- ggplot(aes(x=LoanOriginalAmount), data = loans) +
  geom_histogram(binwidth = 1000, color = "black", fill = "orange")+
  scale_x_continuous(breaks = seq(0,35000, 5000), labels = dollar)

p2 <- ggplot(aes(x=1, y=LoanOriginalAmount), data = loans) +
  geom_boxplot(color = "black", fill = "orange")

grid.arrange(p1, p2, ncol=2)
```






```
ggplot(aes(x=LoanOriginalAmount), data = loans) +
  geom_histogram(binwidth = 100, color = "black", fill = "orange")+
  scale_x_continuous(breaks = seq(0,10000, 1000), limits = c(0,11000),
                     labels = dollar)
```



The loans had a range of 1000-35000 dollars. The first thing that is noticed in this chart is the bumps that we see on some of the bins. Looking more closely, we can see that the bumps are due to the fact that most people apply for loans in round numbers and in intervals of $5000, specifically at $10000, $15000, $20000, and so forth. The only exception is that the number of loans at $4000 is much higher than $5000 (in fact $4000 loans had the highest count). Most of the loans (75% of them) were lower than $12000, which has caused the skewness that we see on the chart. Although the first chart and the boxplot above count loan amounts above $25000 as outliers, I am going to keep these values in the dataset for later analysis to gain a better understanding of loan amount relationship with other variables.

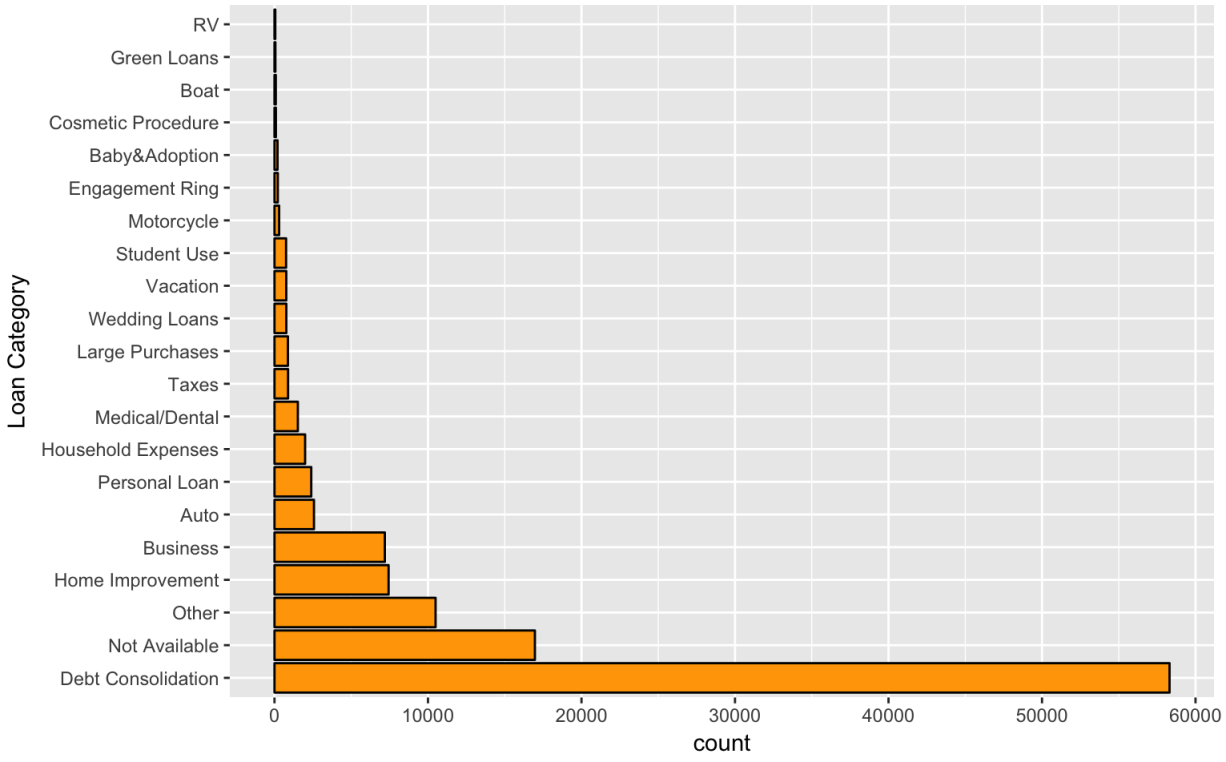
Zooming in on the loans less than $10000, we see that there are loans with all kinds of (non-round) amounts in between the ones with the highest count.

Loan Category

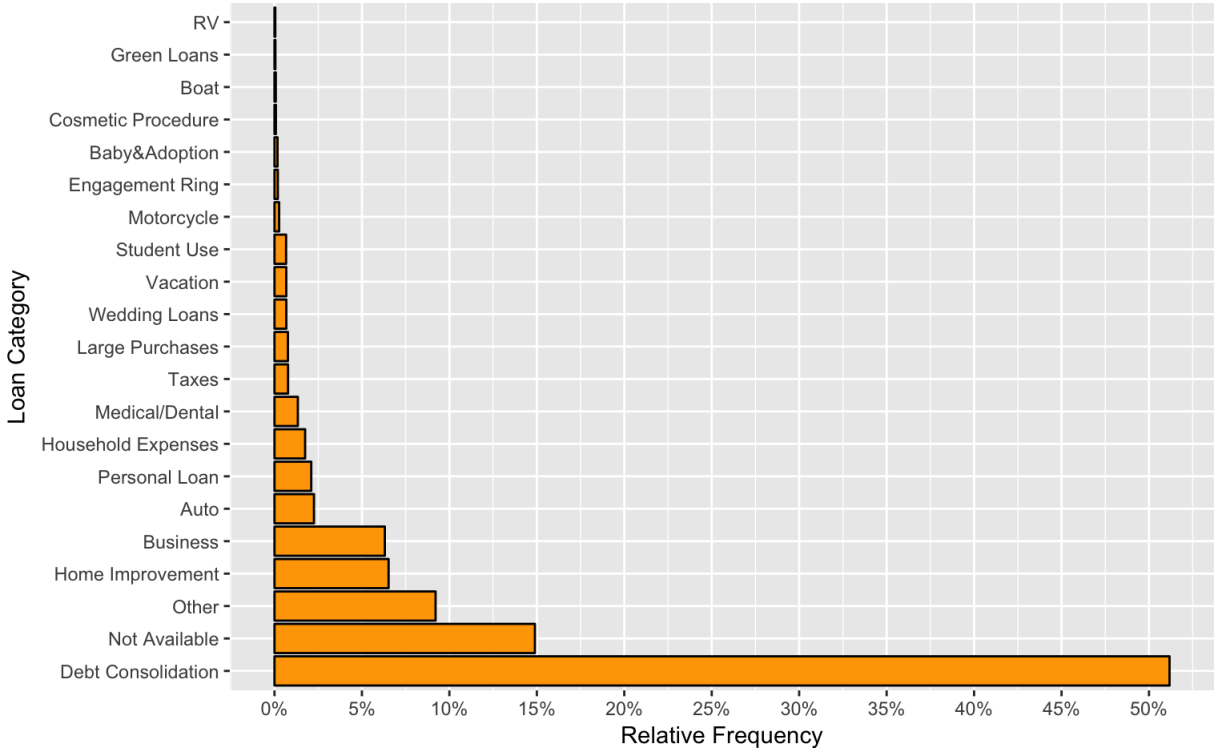
```
# create a vector to replace ListingCategory values with
loan_cat <- c('Not Available', 'Debt Consolidation', 'Home Improvement',
              'Business', 'Personal Loan', 'Student Use', 'Auto', 'Other',
              'Baby&Adoption', 'Boat', 'Cosmetic Procedure', 'Engagement Ring',
              'Green Loans', 'Household Expenses', 'Large Purchases',
              'Medical/Dental', 'Motorcycle', 'RV', 'Taxes', 'Vacation',
              'Wedding Loans')
# replace loan cateogry numbers with loan category names
loans$LoanCategory <- plyr::mapvalues(loans$ListingCategory..numeric.,
                                      c(0:20), loan_cat)

table(loans$LoanCategory)
```

```
##
##               Auto      Baby&Adoption               Boat
##               2572               199                 85
##           Business Cosmetic Procedure Debt Consolidation
##               7189                91              58308
##    Engagement Ring        Green Loans   Home Improvement
##                217                59              7433
## Household Expenses    Large Purchases     Medical/Dental
##               1996               876              1522
##         Motorcycle      Not Available              Other
##                304             16965              10494
##      Personal Loan                RV        Student Use
##               2395                52               756
##              Taxes           Vacation      Wedding Loans
##                885               768               771
```

```
# used fct_infreq from forcats library to order the category bars by length (Source: stackoverflow.com)
ggplot(loans, aes(fct_infreq(factor(LoanCategory)))) +
  geom_bar(color = "black", fill = "orange") + coord_flip() +
  xlab("Loan Category") + scale_y_continuous(breaks = seq(0,60000,10000))
```




```
# used relative frequency for y axis (Source: https://sebastiansauer.github.io/percentage_plot_ggplot2_V2/)
ggplot(loans, aes(fct_infreq(factor(LoanCategory)))) +
  geom_bar(aes(y = (..count..)/sum(..count..)), color = "black", fill = "orange") +
  coord_flip() + xlab("Loan Category") +
  scale_y_continuous(labels=percent, breaks = seq(0,1,0.05)) +
  ylab("Relative Frequency")
```



To make the Loan Cateogry chart more clear, I replaced the integer values of Listing Category variable with the *names* of the loan cateogries provided in the variable dictionary. The first chart shows the distribution of loans across the 21 different categories in the dataset (which includes “Not Available” and “Other” as well).

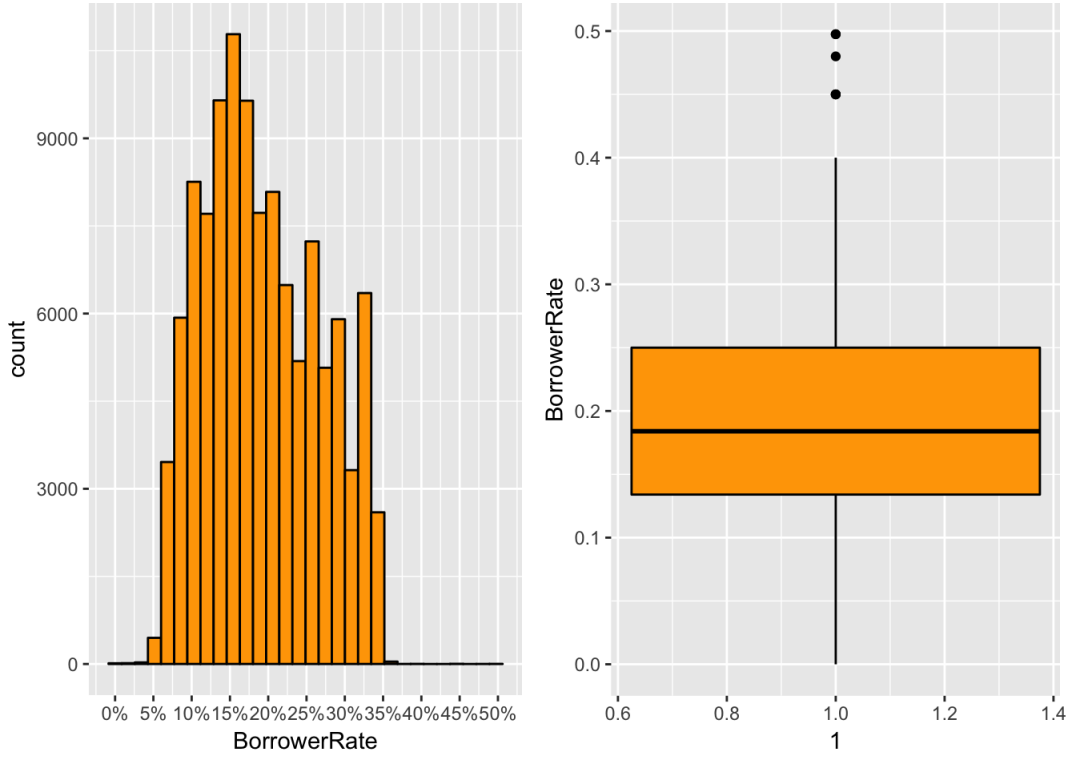
In the second chart (which shows the relative frequency of each category), we can see that “Debt Consolidation” loans account for more than 50% of the loans! Following that, “Home Improvement” and “Business” make up most of the loans applied for between 2005-2014 in the US (excluding “Not Available” and “Other”). At the bottom of the list, we can see that loans for “RV” (52 loans), “Green Loans” (59 loans), and “Boats” (85 loans) were the three least frequent loans.

Interest Rate

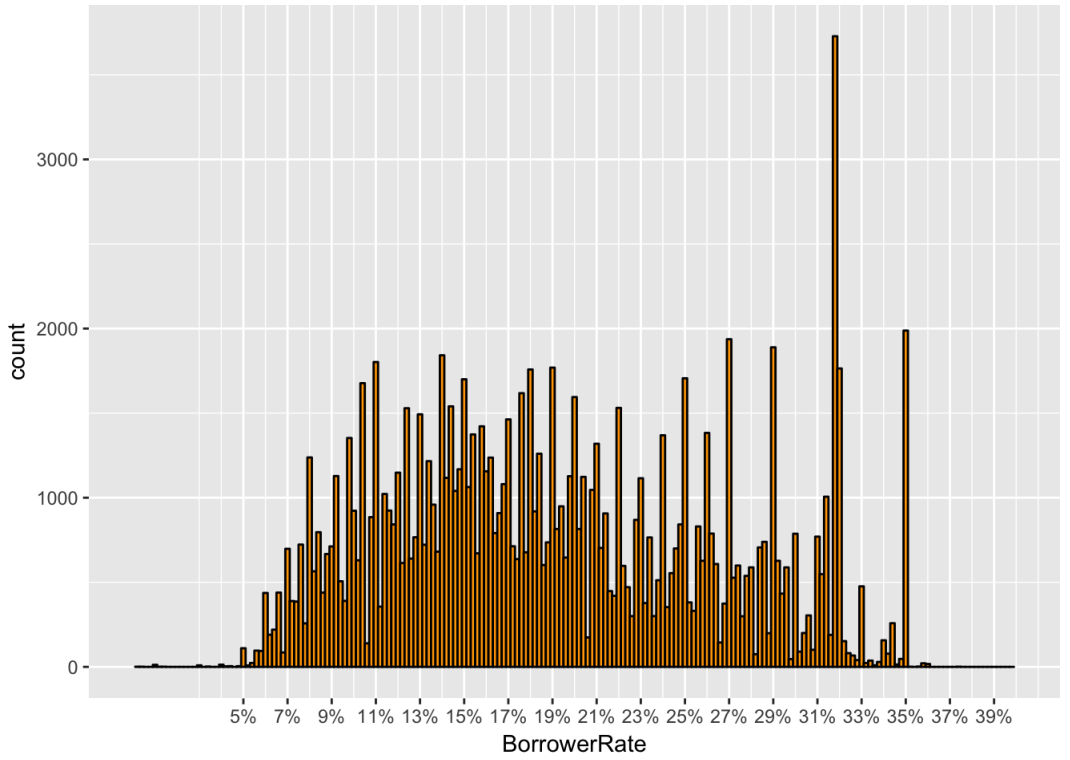
```
summary(loans$BorrowerRate*100)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   13.40   18.40   19.28   25.00   49.75
```

```
I1 <- ggplot(aes(x=BorrowerRate), data= loans)+
  geom_histogram(color = "black", fill = "orange") +
  scale_x_continuous(breaks = seq(0, 0.5, 0.05), labels = percent)

I2 <- ggplot(aes(x=1, y=BorrowerRate), data = loans) +
  geom_boxplot(color = "black", fill = "orange")

grid.arrange(I1, I2, ncol = 2)
```




```
ggplot(aes(x=BorrowerRate), data= loans)+
  geom_histogram(binwidth = 0.002, color = "black", fill = "orange")+
  scale_x_continuous(limits = c(0, 0.4),
                     breaks = seq(0.05, 0.4, 0.02), labels = percent)
```




```
table(subset(loans, BorrowerRate>0.317 & BorrowerRate<0.319)$BorrowerRate*100)
```

```
##
## 31.74 31.75 31.76 31.77 31.78 31.79  31.8 31.85 31.88 31.89
##     3     9     2  3672     1     2     1    27     5     2
```

```
table(subset(loans, BorrowerRate>0.349 & BorrowerRate<0.352)$BorrowerRate*100)
```

```
##
## 34.92 34.94 34.95 34.96 34.97 34.98 34.99    35
##     6     6    37     1     2     4    27  1905
```

Interest rates in the dataset vary from 0% all the way to 49.75%. The first chart shows that there some outlier rates above 40%. With the binwidth set at %0.5 we see that the most frequent bin is the one around 15%. However, limiting the x-axis to 0%-40% (since the boxplot above also counts values above 40% as outliers) and setting the binwidth to 0.2%, we can get a better clarity about the distribution of interest rates on the loans (the second chart). The second chart reveals an unusual frequency for the interest rates around 32%. Looking at the count table for interest rates betwen 31% and 36%, we can see that the interest rate with the highest count is 31.77%. The table also shows that the second most frequent interest rate is 35%.

Later in the report, we will look at how different variables are correlated with interest rate and which ones indicate a stronger correlation.

```
zero_interest <- subset(loans, BorrowerRate == 0)
zero_interest[, c("BorrowerRate", "LoanCategory", "LoanYear",
                  "CreditScoreRangeLower", "LoanOriginalAmount",
                  "BorrowerAPR", "LoanStatus")]
```

```
##        BorrowerRate       LoanCategory LoanYear CreditScoreRangeLower
## 29860             0      Not Available     2006                   800
## 46875             0      Not Available     2007                   520
## 65260             0           Business     2008                   520
## 76859             0      Not Available     2007                   660
## 78402             0      Not Available     2006                   600
## 78921             0 Debt Consolidation     2008                   660
## 90052             0      Not Available     2006                   480
## 112718            0              Other     2008                   720
##        LoanOriginalAmount BorrowerAPR LoanStatus
## 29860                1000     0.01650  Completed
## 46875                1900     0.01315  Defaulted
## 65260                1000     0.01987  Completed
## 76859                3000     0.00653  Completed
## 78402                5000     0.00653  Completed
## 78921               25000     0.01987  Completed
## 90052                1000     0.01650  Completed
## 112718               3000     0.01315  Completed
```

I also looked at the 8 loans with 0% interest rate to see if there is anything noticeable. The 8 loans vary from $1000 to $25000 and the credit score associated with the borrowers ranges from 480 all the way to 800. The loans are related to different loan categories (mostly “Not Available”). Excpet for one, all of them were completed. In general, there is no specific variable that can help us understand the reason behind 0% interest rate in the dataset. The only interesting thing is that all these loans were applied before 2009.

After exploring some of the variables regarding the loans, let’s look at some of the variables regarding the loan borrowers in the following.

Stated Annual Income

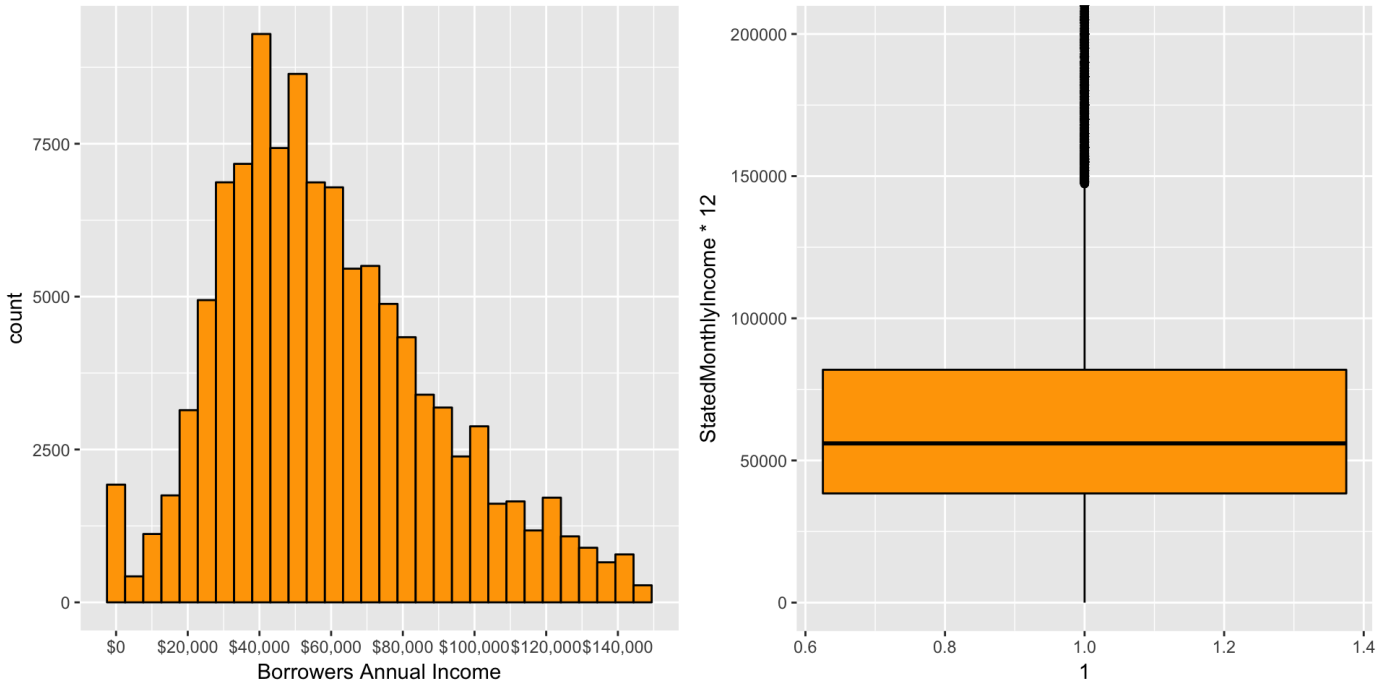
```
summary(loans$StatedMonthlyIncome*12)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##        0    38400    56000    67300    81900 21000000
```

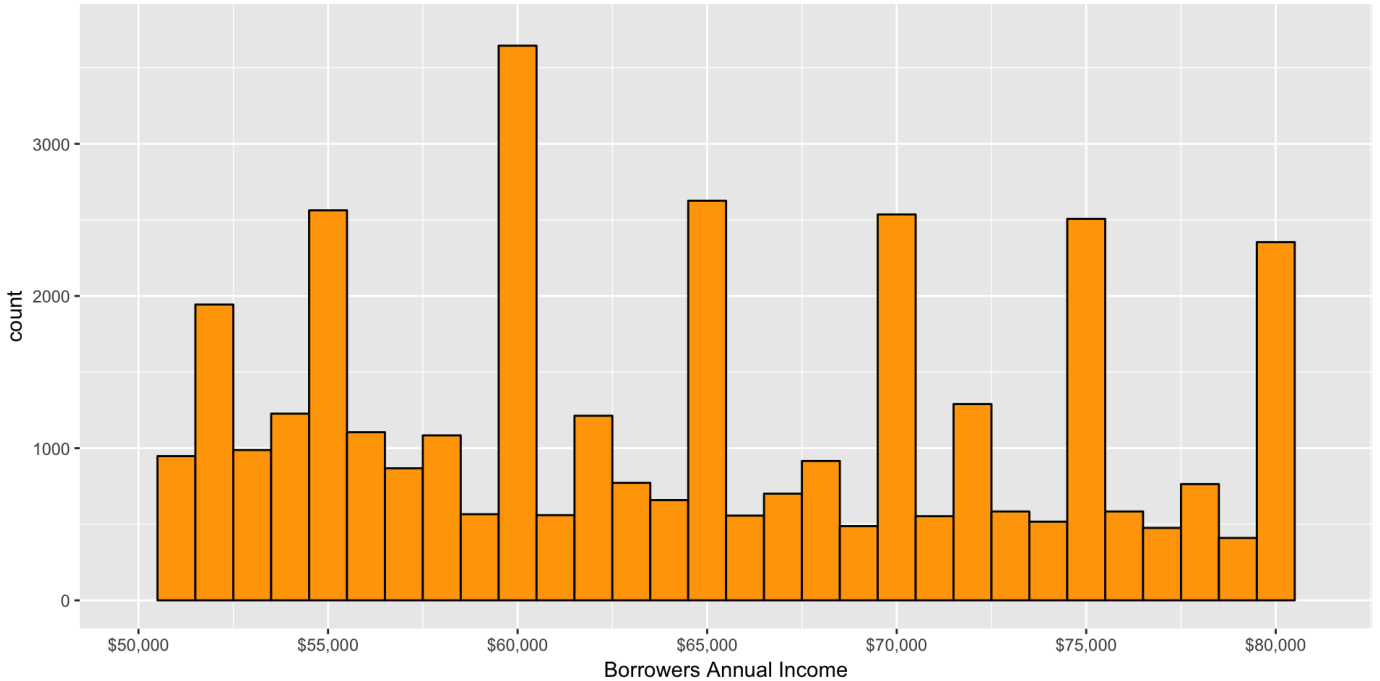
```
S1 <- ggplot(aes(x=1, y=StatedMonthlyIncome*12), data = loans) +
  geom_boxplot(color = "black", fill = "orange") +
  coord_cartesian(ylim = c(0, 200000))

quantile(loans$StatedMonthlyIncome*12, 0.95)
```

```
##    95%
## 147000
```

```
S2 <- ggplot(aes(x=StatedMonthlyIncome*12),
  data = subset(loans, StatedMonthlyIncome < quantile(StatedMonthlyIncome, 0.95))) +
  geom_histogram(color = "black", fill = "orange") +
  scale_x_continuous(breaks = seq(0, 150000, 20000), labels = dollar) +
  xlab("Borrowers Annual Income")

grid.arrange(S2, S1, ncol = 2)
```




```
ggplot(aes(x=StatedMonthlyIncome*12),
       data = subset(loans, StatedMonthlyIncome <
                       quantile(StatedMonthlyIncome, 0.95))) +
  geom_histogram(binwidth = 1000, color = "black", fill = "orange") +
  scale_x_continuous(limits = c(50000, 81000),
                     breaks = seq(50000, 810000, 5000), labels = dollar) +
  xlab("Borrowers Annual Income")
```



The dataset includes the data for stated monthly income, but I decided to annualize the number by multiplying it by 12, since I have a better idea about the annual income scale. As we can see from the stat summary of Stated Income, this variable is extremely skewed to the right. The upper quartile is $81,900 and the max income is $21,000,000! So for the histogram, I limited the x-axis to 95% quantile of the incomes, which is $147,000. Also, we can see from the boxplot that values above $147,000 are considered as outlier. In the analysis later in this report, I am only going to look at a subset of data with income less than the $147,000. Another outstanding point is that 1500 loans were

associated with the stated income of $0.

The first histogram shows that the annual income range of $40,000 - $55,000 has the highest frequency. The binwidth in the first chart is $5000. I set the binwidth at $1000 and zoomed in to get some more info about annual income. The second histogram shows an uneven distribution. Looking more closely at the income distribution in the second histogram, we can see that the bumps are mostly related to the fact that most incomes are at $5000 intervals (i.e. $50,000, $55,000, $60,000, and so forth).

Credit Score

Out of all the creidt rating variables in the dataset, Credit Score is the only one that covers the whole data. “CreditGrade” is for the listings before 2009 and “ProsperRating” and “ProsperScore” is for the listings after 2009. So here we only look at Credit Score as an indicator of the credit status of borrowers.

```
summary(loans$CreditScoreRangeLower)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0   660.0   680.0   685.6   720.0   880.0     591
```

```
table(loans$CreditScoreRangeLower)
```

```
##
##     0   360   420   440   460   480   500   520   540   560   580   600
##   133     1     5    36   141   346   554  1593  1474  1357  1125  3602
##   620   640   660   680   700   720   740   760   780   800   820   840
##  4172 12199 16366 16492 15471 12923  9267  6606  4624  2644  1409   567
##   860   880
##   212    27
```

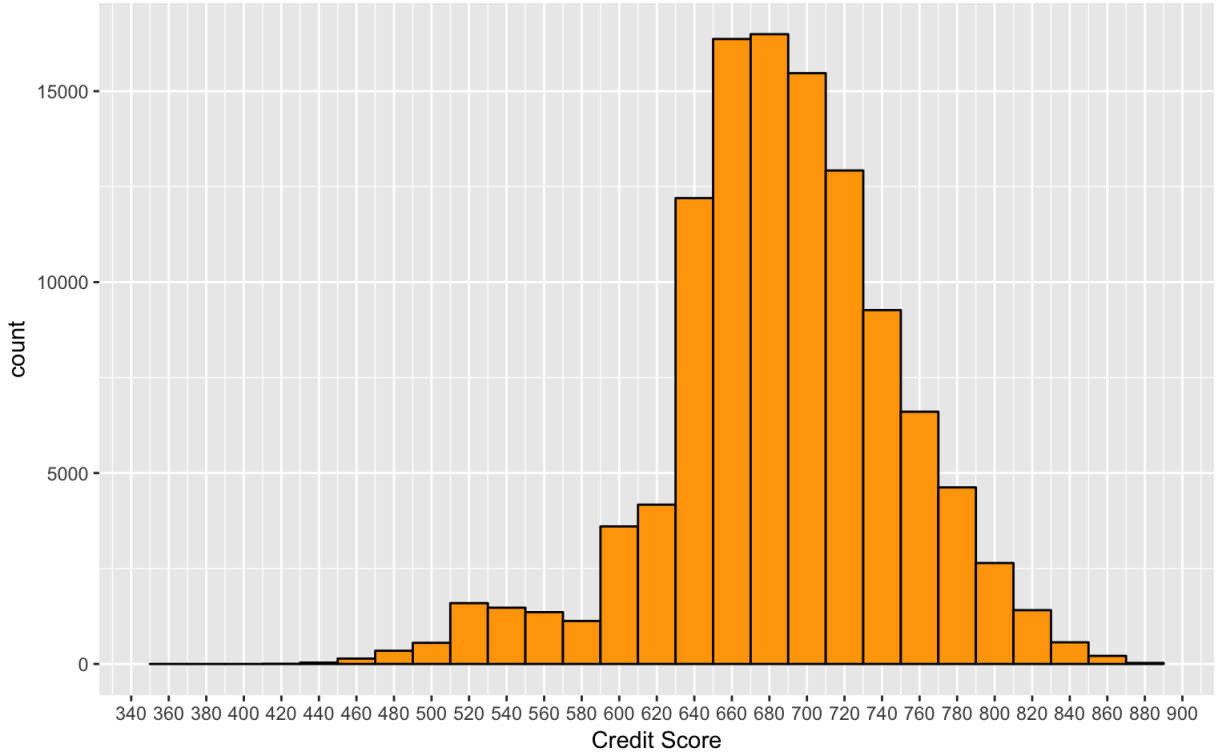
```
summary(loans$CreditScoreRangeUpper)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    19.0   679.0   699.0   704.6   739.0   899.0     591
```

```
table(loans$CreditScoreRangeUpper)
```

```
##
##    19   379   439   459   479   499   519   539   559   579   599   619
##   133     1     5    36   141   346   554  1593  1474  1357  1125  3602
##   639   659   679   699   719   739   759   779   799   819   839   859
##  4172 12199 16366 16492 15471 12923  9267  6606  4624  2644  1409   567
##   879   899
##   212    27
```

```
ggplot(aes(x=CreditScoreRangeLower),
       data = subset(loans, CreditScoreRangeLower>0))+
  geom_histogram(binwidth = 20, color = "black", fill = "orange") +
  scale_x_continuous(breaks = seq(300, 900, 20)) + xlab("Credit Score")
```



The first table shows the summary stat for lower range credit score and the second table shows the summary for upper range credit score. There are 133 loans with the credit score of 0-19 for the borrowers which were removed from the chart above. After removing those records, the credit score of 360 becomes the minimum value. Since the credit score is calculated in intervals of 20, I set the binwidth of the histogram at 20. (e.g. bin 660 represents a credit score within the 660-679 interval). Minimum credit score is 360-379 and maximum is 880-899. We see a somewhat normal distribution for credit score here (there is a bit of a long tail on the left side) where most loans were applied by people with a credit score between 660 to 720. It is interesting to see that the left side of the chart is less smooth than the right side. There is a drastic reduction in count from score 640 to 620 and then there is a slight increase in count from credit score 580 to 520. On the right side, we see a smooth gradual reduction in count from score 680 all the way to 880.

```
zero_creditscore <- subset(loans, CreditScoreRangeLower == 0)
head(zero_creditscore[,c("CreditScoreRangeLower", "EmploymentStatus")])
```

```
##      CreditScoreRangeLower EmploymentStatus
## 795                      0    Not available
## 913                      0    Not available
## 1686                     0    Not available
## 4291                     0
## 4372                     0    Not available
## 4431                     0    Not available
```

```
length(unique(zero_creditscore$MemberKey))
```

```
## [1] 133
```

```
mean(zero_creditscore$BorrowerRate)
```

```
## [1] 0.2260353
```

```
median(zero_creditscore$StatedMonthlyIncome*12)
```

```
## [1] 23000
```

I was curious about the 117 records with 0-19 credit score. So I further looked at a subset of the data with the value of 0 for CreditScoreRangeLower variable. Looking at the unique member keys, we find out that these credit scores belong to 133 different borrowers. One interesting (and kind of expected) point is that, although “Not available” accounts for only 4.7% of all Employment Status records in the original dataset, for this subset, %100 of the records show “Not available” as Employment Status. Also the interest rate for these loans has an average of 22.6% which is about 3.3 percentage points higher than the overall average interest rate. Also, none of of these borrowers were homeowners. The median annual income for these borrowers is $23000 which is about 60% lower than the overall median income ($56000).

Delinquent Accounts

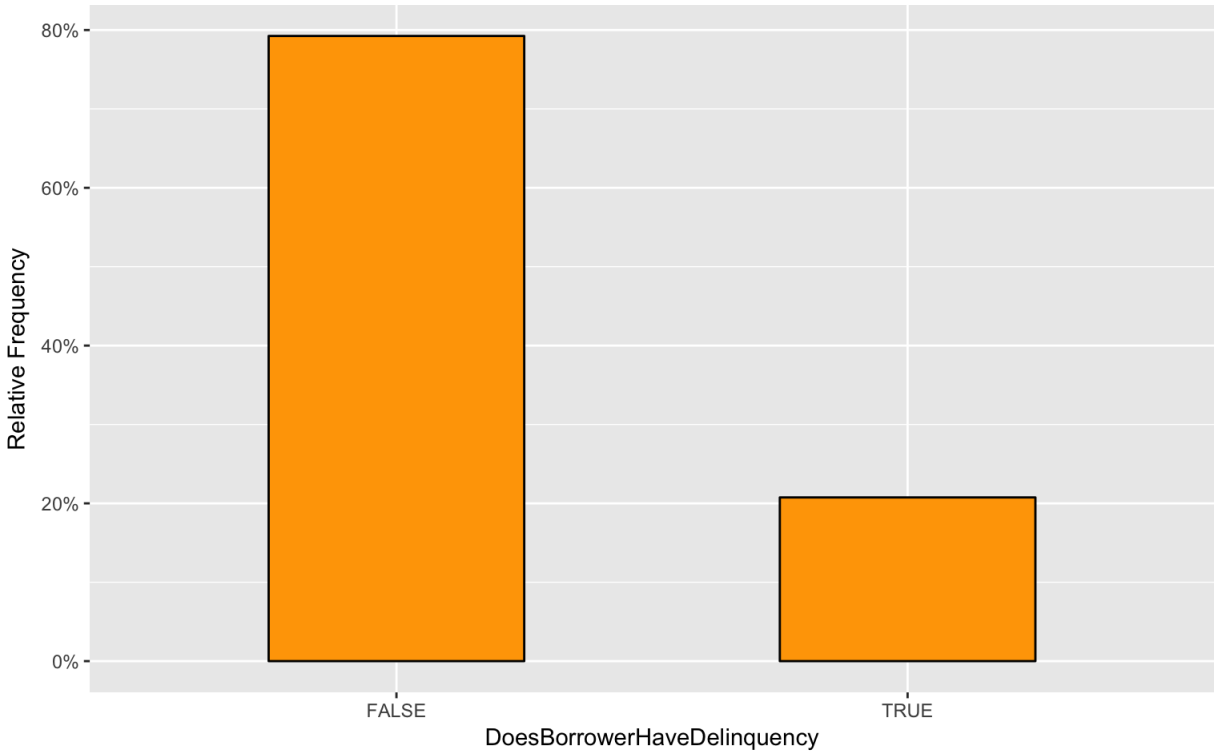
```
summary(loans$CurrentDelinquencies)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.0000  0.0000  0.5921  0.0000 83.0000     697
```

```
subset_delinquent <- subset(loans, !is.na(CurrentDelinquencies))
quantile(subset_delinquent$CurrentDelinquencies, seq(0.75, 0.99, 0.03))
```

```
## 75% 78% 81% 84% 87% 90% 93% 96% 99%
##   0   0   1   1   1   2   2   4  10
```

```
loans$DoesBorrowerHaveDelinquency <- ifelse(loans$CurrentDelinquencies > 0,
                                            TRUE, FALSE)

ggplot(aes(x=DoesBorrowerHaveDelinquency),
       data = subset(loans, !is.na(DoesBorrowerHaveDelinquency))) +
  geom_bar(aes(y = (..count..)/sum(..count..)), color = "black",
           fill = "orange", width = 0.5) + scale_y_continuous(labels = percent)+
  ylab("Relative Frequency")
```

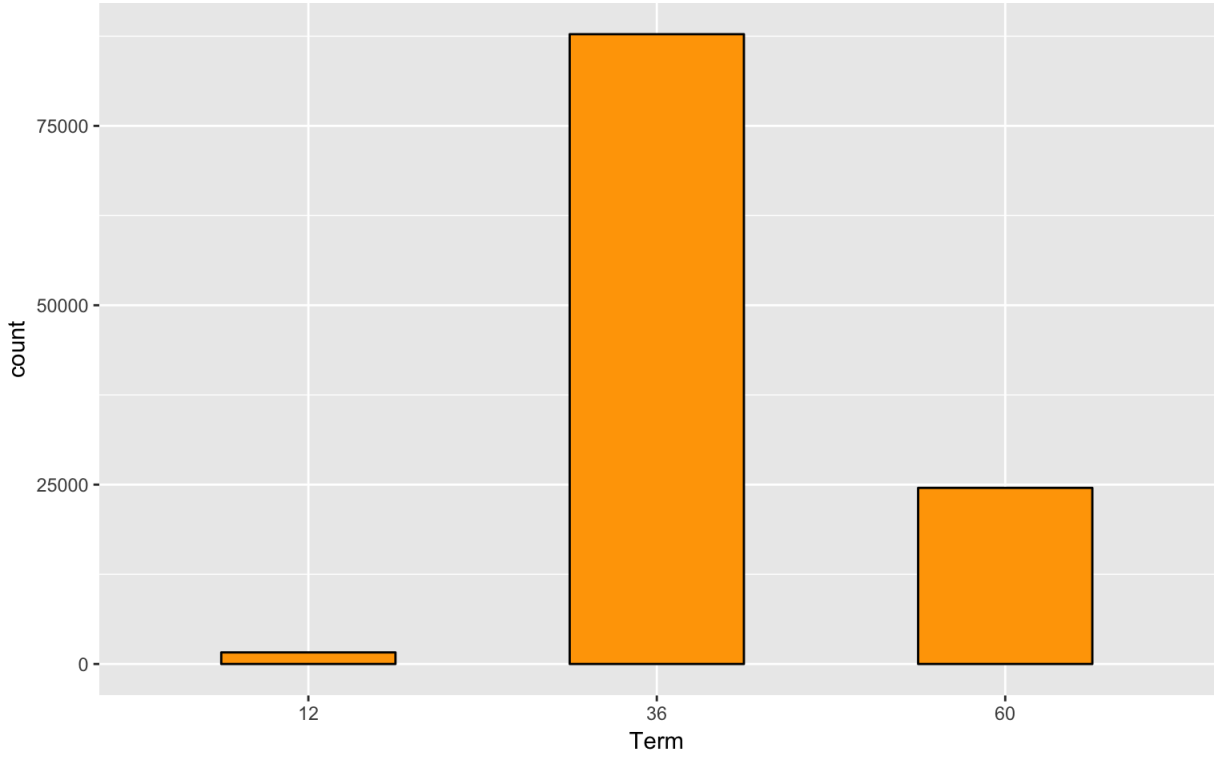


Looking at the summary table of the number of delinquent accounts that the borrower had at the time credit profile was pulled, we see that at least 75% of borrowers had no delinquent accounts. Closer look at the quantiles (second table) we find out that the percentage of borrowers with no delinquent accounts is about 80%. Given the skewed distribution and also nature of the data, I decided to split the borrowers into 2 groups: one with no delinquent accounts and one with delinquent accounts. The chart shows the relative frequency of borrowers with vs without delinquency.

```
table(loans$Term)
```

```
##
##    12    36    60
##  1614 87778 24545
```

```
ggplot(aes(x=factor(Term)), data = loans) +
  geom_bar(color = "black", fill = "orange", width = 0.5) + xlab("Term")
```

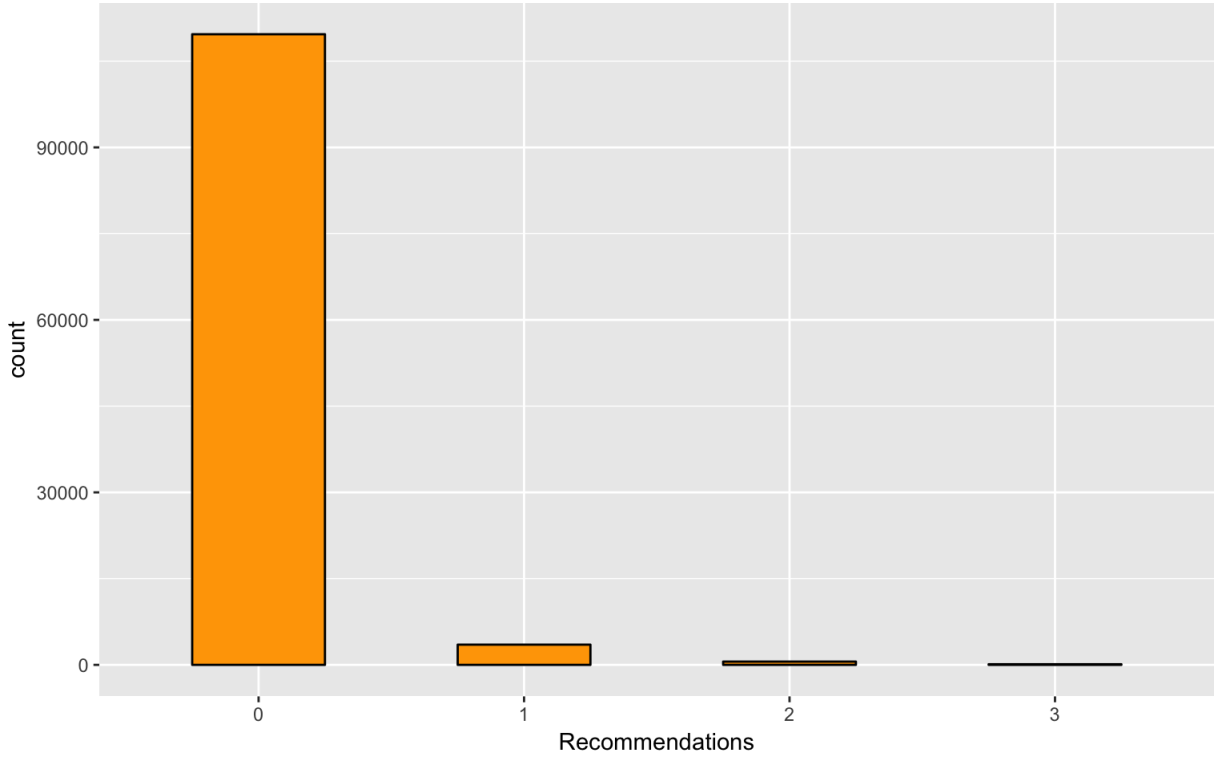


The chart above shows that the payment period of most of the loans (77% of the loans to be exact) were 36 months. 12 month loans had the lowest count.

```
table(loans$Recommendations)
```

```
##
##      0      1      2      3      4      5      6      7      8      9
## 109678   3516    568    108     26     14      4      5      3      6
##     14     16     18     19     21     24     39
##      1      2      2      1      1      1      1
```

```
ggplot(aes(x=factor(Recommendations)), data = subset(loans, Recommendations<4)) +
  geom_bar(color = "black", fill = "orange", width = 0.5) +
  xlab("Recommendations")
```



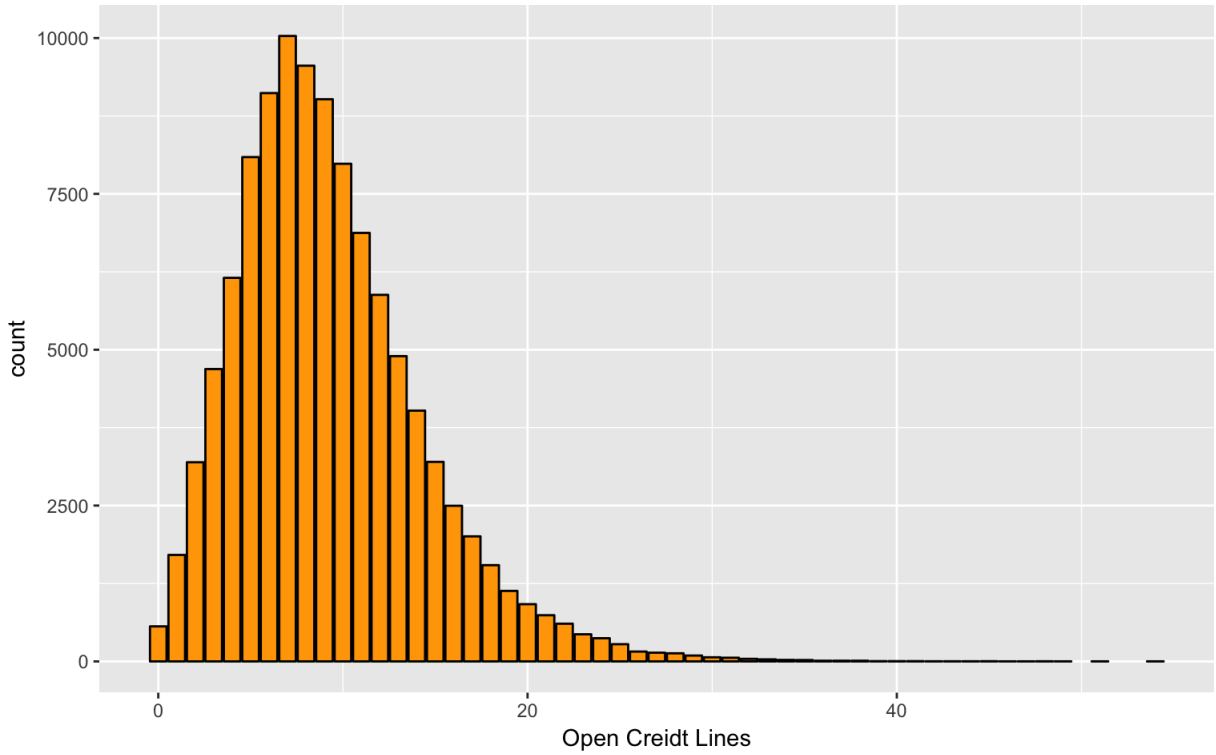
As we see from the table of recommendations above, the loans with more than 3 recommendatons had a very low count. So I created the

bar graph for loans with 0 to 3 recommendations. We can see from the chart that overwhelmingly most loans did not have any recommendations (more than 96%).

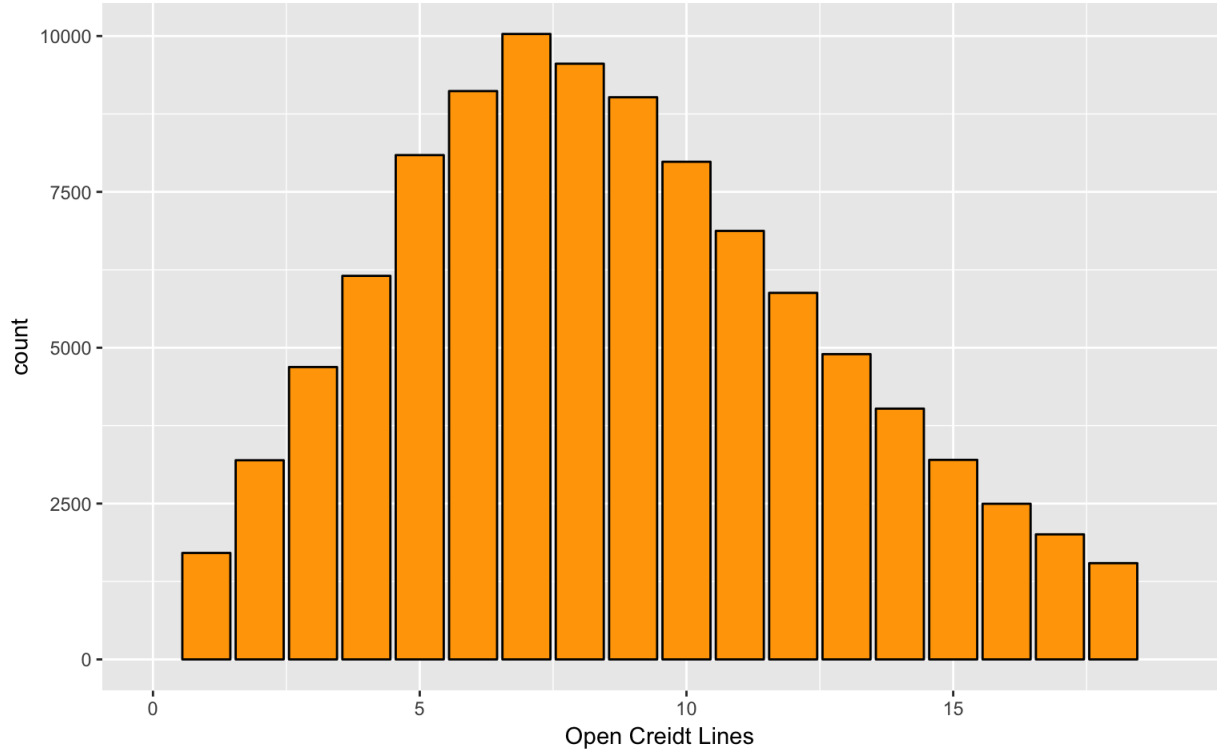
```
table(loans$OpenCreditLines)
```

```
##
##    0    1    2    3    4    5    6    7    8    9   10   11
##  562 1708 3195 4690 6153 8090 9117 10033 9556 9018 7983 6874
##   12   13   14   15   16   17   18   19   20   21   22   23
## 5880 4897 4023 3201 2497 2006 1544 1131  918  741  605  435
##   24   25   26   27   28   29   30   31   32   33   34   35
##  372  276  158  140  130   94   66   59   41   33   24   22
##   36   37   38   39   40   41   42   43   44   45   46   47
##   11   11   11    4    4    4    2    2    2    3    1    1
##   48   49   51   54
##    2    1    1    1
```

```
ggplot(aes(x=OpenCreditLines), data = loans) +
  geom_bar(color = "black", fill = "orange") +  xlab("Open Creidt Lines")
```




```
ggplot(aes(x=OpenCreditLines), data = loans) +
  geom_bar(color = "black", fill = "orange") +  xlab("Open Creidt Lines") +
  xlim(0,19)
```



The number of open credit lines at the time of applying for the loan has a distribution with a long tail to the right. However, if we exclude the loans with more than 19 open credite lines for the borrower (since the count for those loans is very low), we can see that the distribution of open credit lines is a normal one between 0-19 credit lines, which peaks at 7 credit lines.

Univariate Analysis

What is the structure of your dataset?

There are 113,937 loans in the Prosper dataset with 81 attributes about the loans. The variables provide information about the loans (amount, monthly payment, term, etc.) as well as information about the borrowers (stated income, credit score, occupation, etc.).

The loans' amounts range from $1000 to $35000 (with a median of $6500) and have been applied for in 19 different categories (excluding "Not Available" and "Other" categories), with "Debt Consolidation" making up over 50% of the loans.

The interest rate that has been applied to these loans ranges from 0% to 49.75% with a mean of 19.28%. From the data, it looks like that the rate of 31%-32% had the highest count.

Furthermore, Borrowers' credit score has a range of 360-900 with 133 records of borrowers with credit score in the range of 0-19. The median credit score observed in the dataset was the range of 680-699.

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

The main variables that I would like to investigate in this dataset are interest rate and credit score and their relationships with some of the other variables. I would like to better understand that out of the provided variables, which factors mostly impact the interest rate applied to a loan. Of course, credit score is one of the variables which most likely has a strong impact and so I would also like to understand the relationships between credit score and some of the other variables.

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

I suspect variables such as loan category, number of recommendations, and annual income are among the variables that have a strong impact on interest rate. These relationships will be investigated in the next section.

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

I added a variable to the dataset to help with data exploration and visualization. The "Listing Category" variable included the type of the loan in integer numbers from 0-20. Each number represents one type of loan which is explained in the variable dictionary (e.g. business, debt consolidation, home improvement). Using the "mapvalues" function from "plyr" library, I created another column "Loan Category" which mapped the integer values of the "Listing Category" column to their associated names of the loan category.

I also added a variable "DoesBorrowerHaveDelinquency" since 80% of the borrowers do not have any delinquent accounts, I was interested to see how just the fact of having a delinquent account will impact other variables in the next section.

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?

The Stated Income has an extremely skewed distribution, so in order to show the distribution in a proper way, I limited the income data at the 95% quantile and also annualized the data by multiplying the column by 12 (since annual numbers make more sense).

I also noticed some unusual spikes in the loan amount distribution, which upon closer look, turned out to be related to the fact that most loans are in round numbers with $5000 intervals (e.g. at $10000, $15000, $20000).

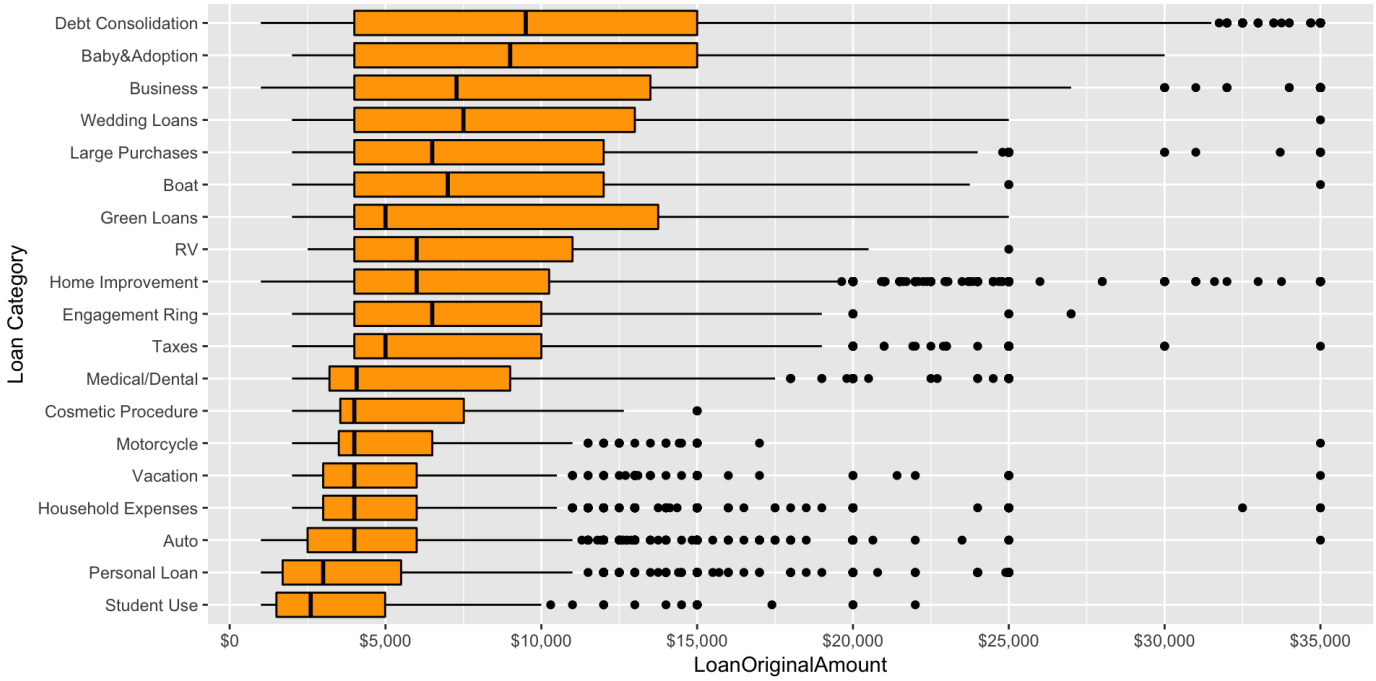
I also split “Loan Originiation Quarter” column into “Loan Quarter” and “Loan Year” so that I can use the “year” and “quarter” variable separately in my exploration.

In the next section, I will explore the relationship between some of the variables in the dataset.

Bivariate Plots Section

Loan Amount vs. Loan Category

```
# used reorder function to sort the box plot by median amount
#(Source: https://groups.google.com/forum/#!topic/ggplot2/8N0ofttOdcw)
ggplot(aes(x=reorder(LoanCategory, LoanOriginalAmount, fun=median),
           y=LoanOriginalAmount), data = subset(loans, !LoanCategory %in%
                                                c("Other", "Not Available")))+
  geom_boxplot(color = "black", fill = "orange")+ coord_flip()+
  scale_y_continuous(breaks = seq(0,35000, 5000), labels = dollar) +
  xlab("Loan Category")
```



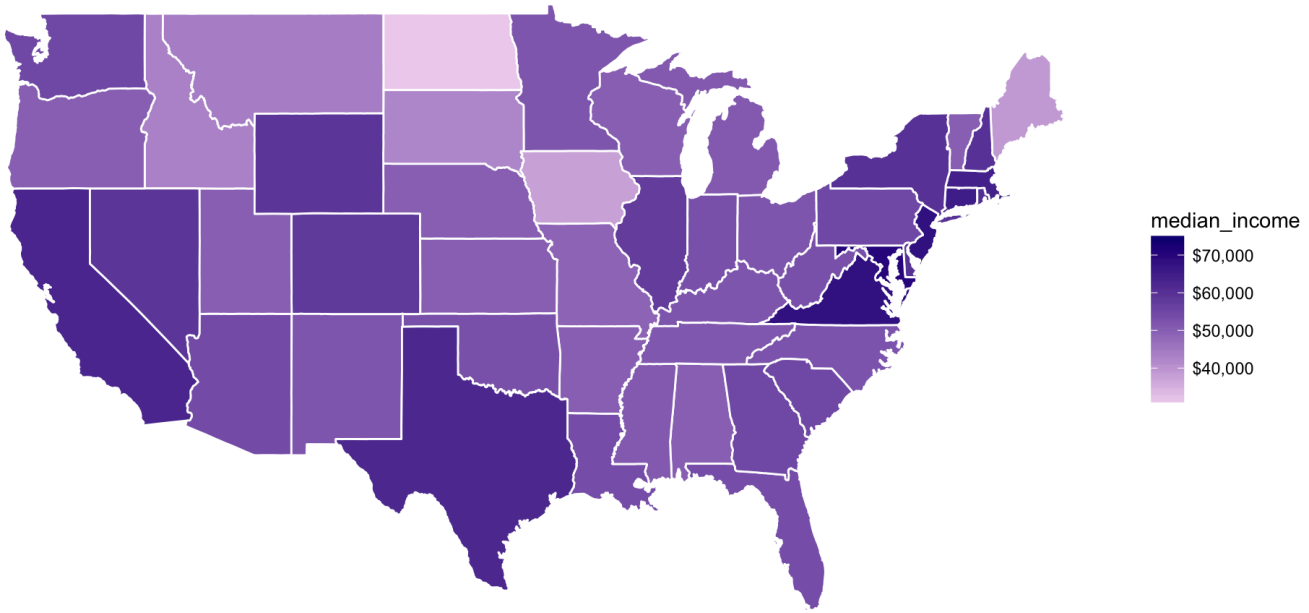
Looking at the amount of loan in different loan categories reveals some interesting info about the loans. First of all, Debt Consolidation loans are the loans with the highest variance and highest median (median of about $10000). The second highest loan median belongs to “Baby&Adoption” loans (median of $9000), which is one of the least frequent types of loan, followed by “Business” Loans (median of $7000). It is interesting that despite low counts, “Baby&Adoption” and “Green Loans” have the second and third highest interquartile range (indicating a wide spread) among all categories. At the other end of the chart, we can see that “Student Use” and “Personal Loan” have the lowest amount with the median of about $2500 and $3000 respectively. The other interesting point is that the top 11 categories have the same lower quartile at $4000 (which we saw in the previous section was the loan amount with the highest count).

Borrower State vs (Income-Loan Amount)

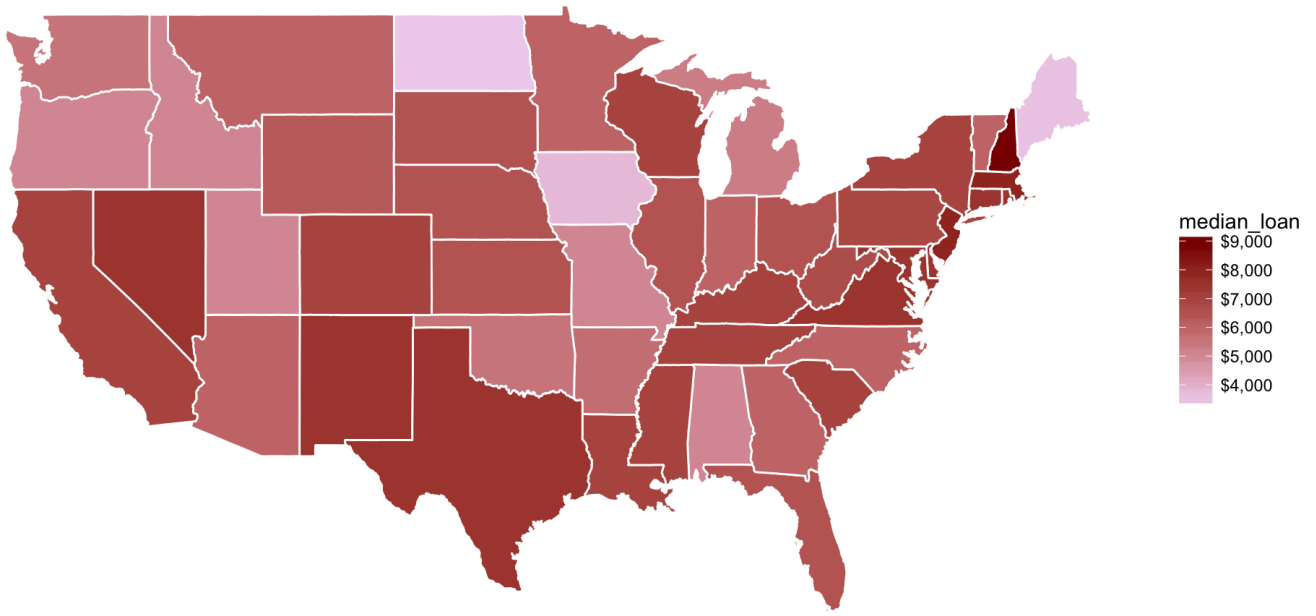
```
summary(loans$StatedMonthlyIncome*12)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##         0     38400     56000     67300     81900 21000000
```

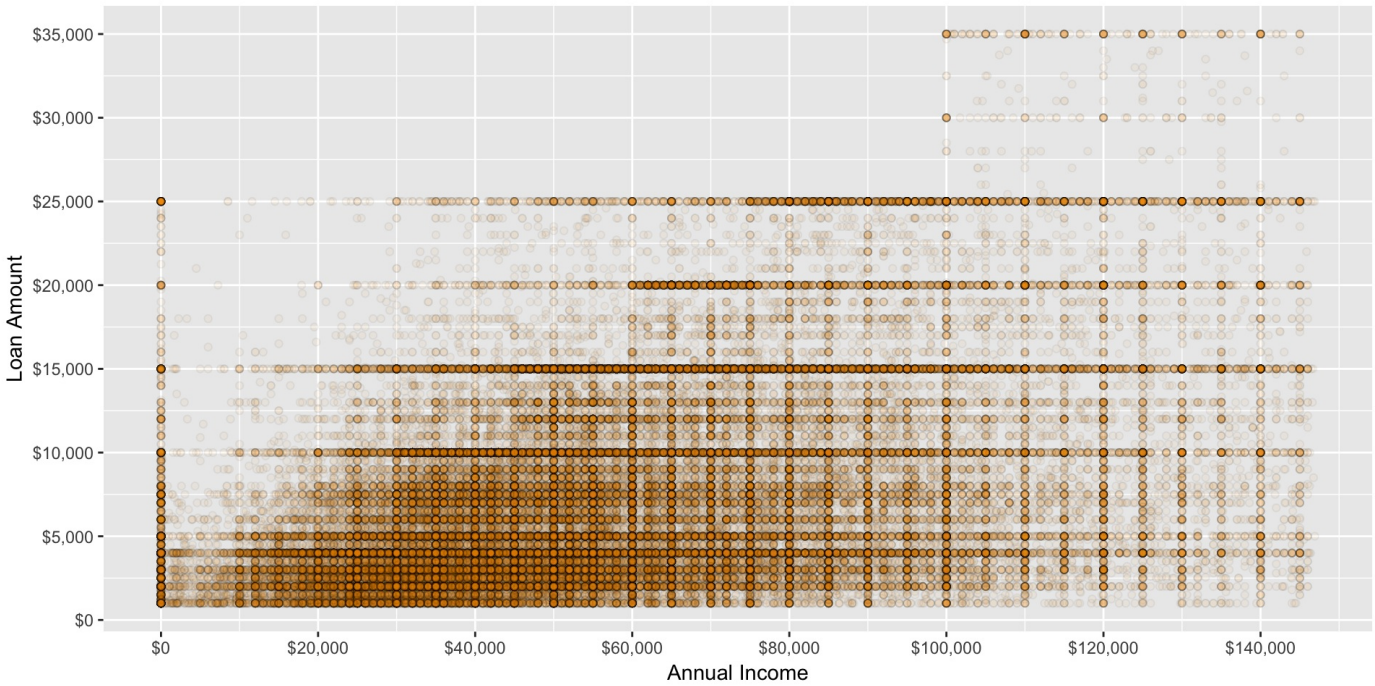
```
# Create a table of median income and median loan amount per state
state_income_loan <- subset(loans, BorrowerState != "") %>%
  group_by(BorrowerState) %>%
  summarise(median_income = median(StatedMonthlyIncome*12),
            median_loan = median(LoanOriginalAmount)) %>%
  arrange(desc(median_income))

#load US states map, capitalize the first letter of state names and
#replace them with abbreviated names to match loans BorrowerState column
#(Source: Stackoverflow.com)
all_states <- map_data("state")
all_states$region <- stringi::stri_trans_totitle(all_states$region)
all_states$region <- state.abb[match(all_states$region, state.name)]

ggplot()+geom_map(data=all_states, map=all_states,
                  aes(x=long, y=lat, map_id=region)) +
  geom_map(data=state_income_loan, map = all_states,
           aes(fill = median_income, map_id = BorrowerState), color="#ffffff") +
  labs(x= NULL, y = NULL) + theme(axis.text = element_blank()) +
  theme(panel.background = element_blank()) +
  theme(axis.ticks = element_blank()) +
  scale_fill_continuous(low = "thistle2", high = "navyblue", labels = dollar)
```




```
ggplot()+geom_map(data=all_states, map=all_states,
                  aes(x=long, y=lat, map_id=region)) +
  geom_map(data=state_income_loan, map = all_states,
           aes(fill = median_loan, map_id = BorrowerState), color="#ffffff") +
  labs(x= NULL, y = NULL) + theme(axis.text = element_blank()) +
  theme(panel.background = element_blank()) +
  theme(axis.ticks = element_blank()) +
  scale_fill_continuous(low='thistle2', high='darkred', labels = dollar)
```




```
ggplot(aes(x=StatedMonthlyIncome*12, y=LoanOriginalAmount),
       data=subset(loans, StatedMonthlyIncome <
                     quantile(StatedMonthlyIncome, 0.95))) +
  geom_point(color = I("black"), fill = I("#F79420"), shape = 21, alpha = 0.05) +
  scale_x_continuous(breaks = seq(0,150000, 20000), labels = dollar) +
  scale_y_continuous(breaks = seq(0,35000, 5000), labels = dollar) +
  labs(x= "Annual Income", y = "Loan Amount")
```



The first chart displays the median annual income in different states for people who applied for Prosper loans. One of the surprises here for me was the state of Virginia, which I did not know was among the top states in median_income ($68000). Of course, California ($63,000) and Texas ($62,300) with the huge tech and gas&oil industry have a high median income as well. At the bottom of the list, we have states like North Dakota and Iowa with $31,600 and $37,800 as median income.

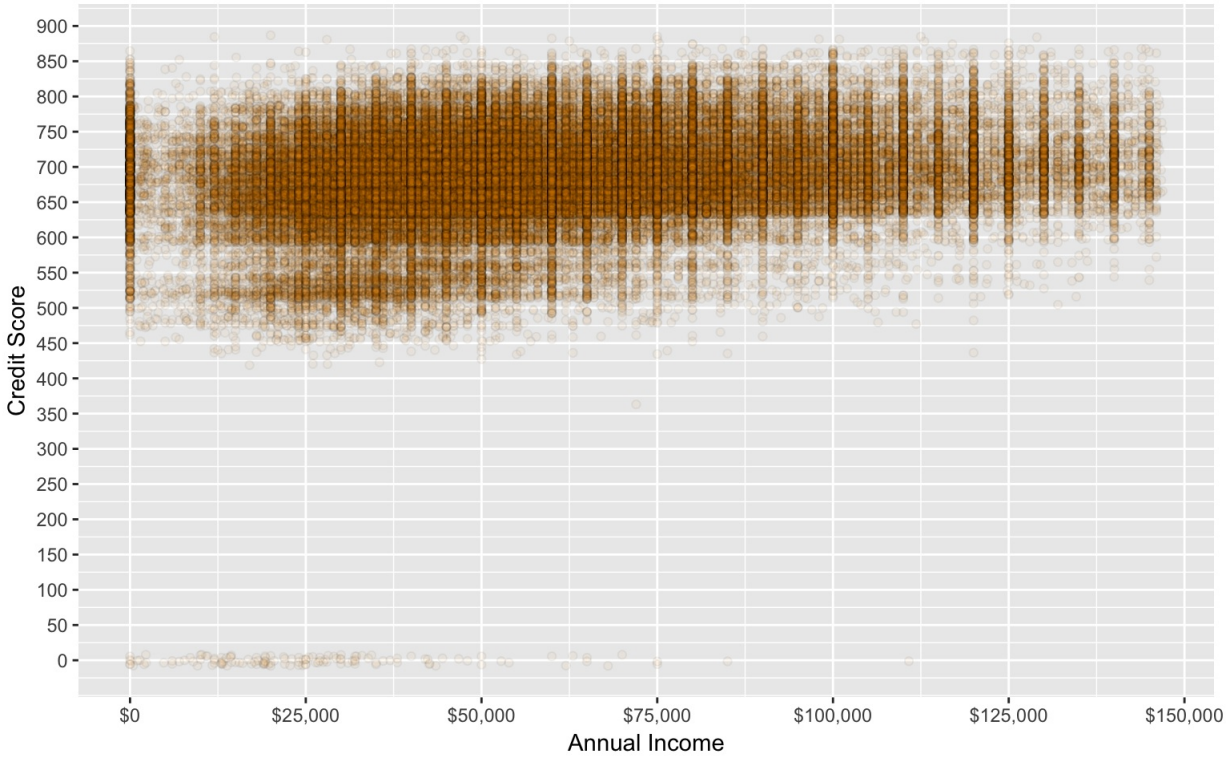
The second chart shows the median amount of loan applied in different states. Comparing the two charts side by side, We can see a very similar trend between income and amount of loan, which is interesting.

The scatter plot above explains the similarity in the trend between income and loan amount. First of all, as mentioned before, most loans are in round amounts with $5000 intervals. That is why we see horizanotal bands of data points at loans of $10000, $15000 and so forth. The chart clearly shows that at income levels below $80,000 most loans are below $10,000. At income levels below $60,000, we rarely see loans above $15000. On the other hand, for $25000 loans, we see that most loans are applied for by people with income above $80,000. The interesting point is that there is not a single loan of more than $25000 for an annual income of less than $100,000.

Credit Score Relationships

In this section, I am going to look at multiple variables and their relationships primarily with borrowers' credit score.

```
ggplot(aes(x=StatedMonthlyIncome*12, y=CreditScoreRangeLower),
       data = subset(loans, StatedMonthlyIncome <
                       quantile(StatedMonthlyIncome, 0.95))) +
  geom_jitter(alpha = 0.05, color = I("black"),
              fill = I("#F79420"), shape = 21) +
  scale_x_continuous(breaks = seq(0,150000,25000), labels = dollar) +
  scale_y_continuous(breaks = seq(0, 900, 50)) +
  labs(x="Annual Income", y = "Credit Score")
```




```
cor.test(loans$StatedMonthlyIncome, loans$CreditScoreRangeLower)
```

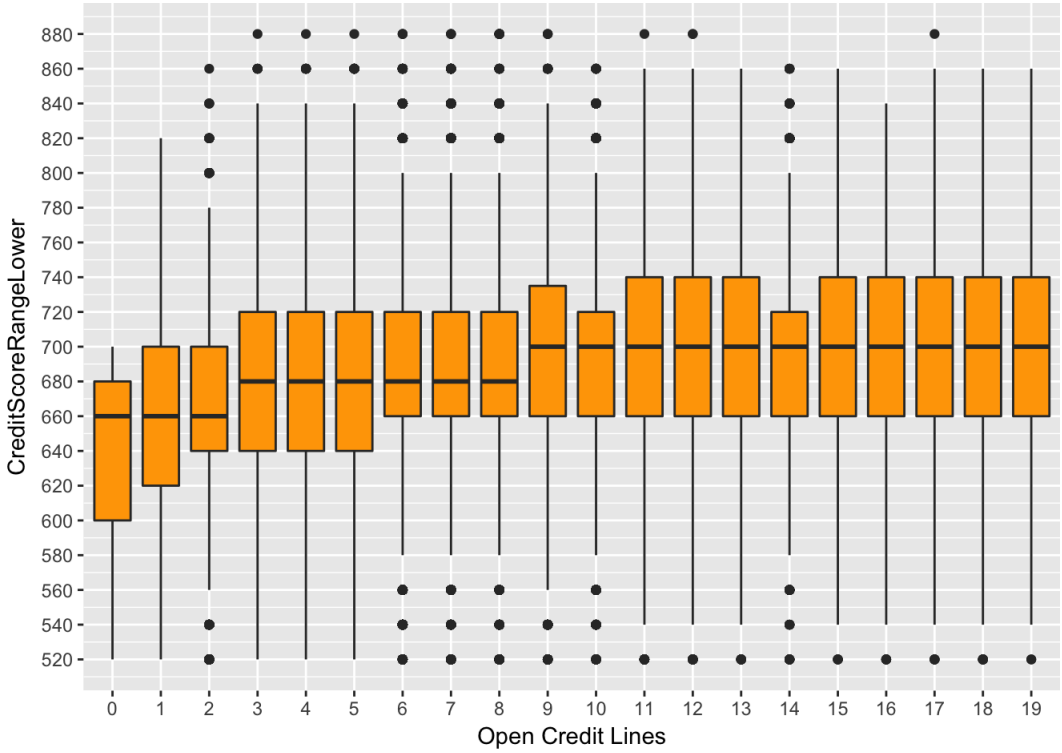
```
##
##  Pearson's product-moment correlation
##
## data:  loans$StatedMonthlyIncome and loans$CreditScoreRangeLower
## t = 36.54, df = 113340, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1021433 0.1136511
## sample estimates:
##       cor
## 0.1079008
```

From the scatter plot we can see that no one with more than $75,000 in annual income has a credit score in 0-19 range and most borrowers with the credit score lower than 600 earn less than $75,000 per year. However, within the range of 650-750 which makes up more than 60% of the borrowers, we see a very wide range of income from $25,000 to $100,000. The correlation number between the two variables does not reveal a very strong correlation between Income and Credit Score.

```
table(loans$OpenCreditLines)
```

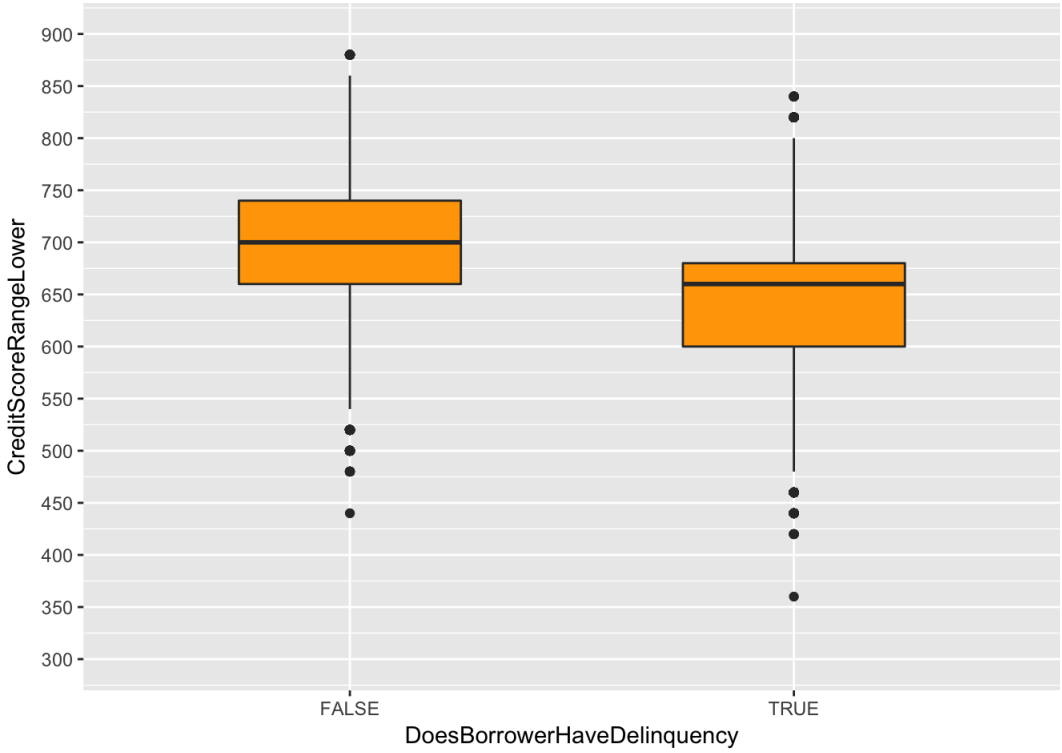
```
##
##     0     1     2     3     4     5     6     7     8     9    10    11
##   562  1708  3195  4690  6153  8090  9117 10033  9556  9018  7983  6874
##    12    13    14    15    16    17    18    19    20    21    22    23
##  5880  4897  4023  3201  2497  2006  1544  1131   918   741   605   435
##    24    25    26    27    28    29    30    31    32    33    34    35
##   372   276   158   140   130    94    66    59    41    33    24    22
##    36    37    38    39    40    41    42    43    44    45    46    47
##    11    11    11     4     4     4     2     2     2     3     1     1
##    48    49    51    54
##     2     1     1     1
```

```
ggplot(aes(x=factor(OpenCreditLines), y=CreditScoreRangeLower),
       data = subset(loans, OpenCreditLines < 20)) +
  geom_boxplot(fill = "orange") + xlab("Open Credit Lines") +
  scale_y_continuous(breaks = seq(400, 900, 20))
```



I suspected that the number of borrower's open credit lines at the time his/her credit profile was pulled might have some correlation with the credit score that they get. The boxplot shows that the lower quartile of credit score increasee by 20 between 0-2 open credit lines which is not a lot, and after 3 credit lines the chart shows no evidence for a strong relationship. I limited the number of open credit lines to less than 20 to get rid of the least frequent values in this variable and get a better clarity from the chart.

```
ggplot(aes(x=DoesBorrowerHaveDelinquency, y = CreditScoreRangeLower),
       data = subset(loans, !is.na(DoesBorrowerHaveDelinquency)))+
  geom_boxplot(fill="orange", width = 0.5)+coord_cartesian(ylim = c(300,900)) +
  scale_y_continuous(breaks = seq(300,900,50))
```



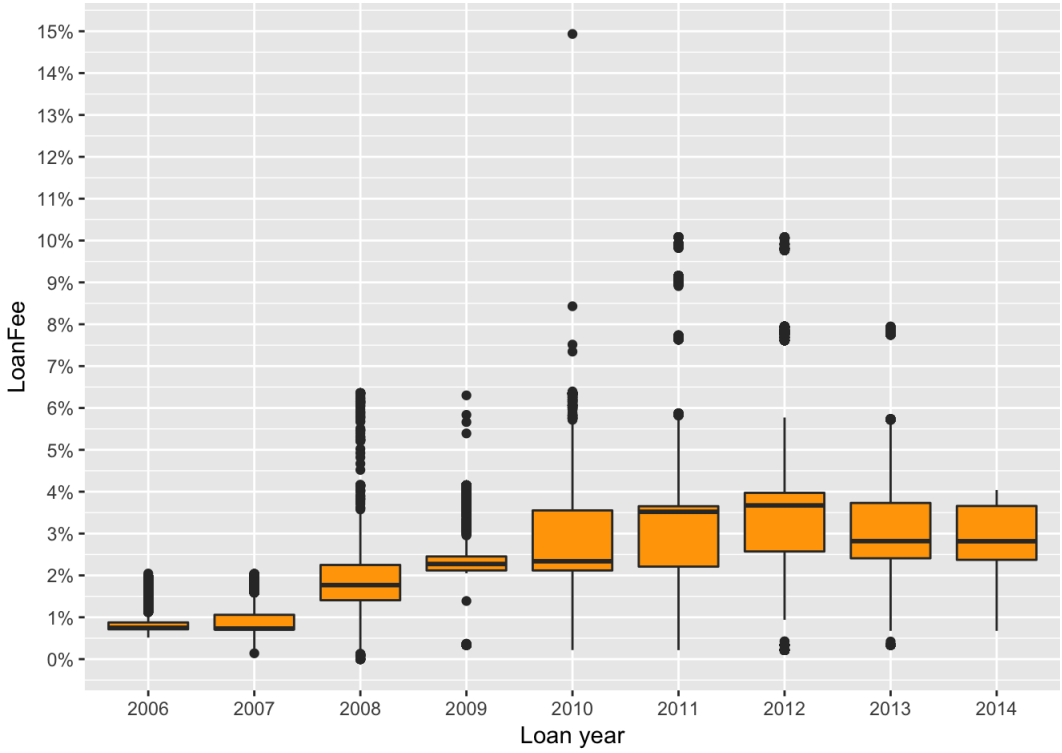
As expected, we can see from the data that in general, borrowers with no delinquent accounts have a higher credit score. In fact, the median score for borrowers with no delinquency (700) is higher than the upper quartile of borrowers with delinquency (about 660).

Time Series of Loan Fees (APR - Interest Rate) and Interest Rates

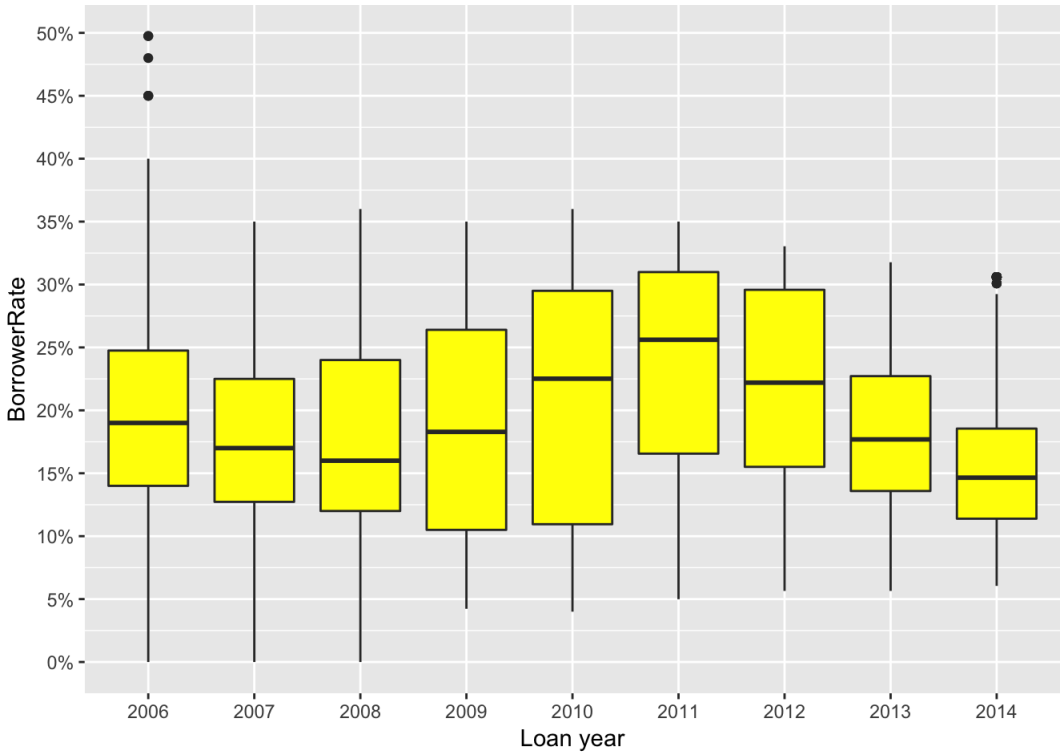
```
# add a column to the dataset to display loan fees in percentage
loans$LoanFee <- loans$BorrowerAPR - loans$BorrowerRate

summary(loans$LoanFee*100)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   1.900   2.502   2.604   3.653  14.930      25
```

```
ggplot(aes(x=factor(LoanYear), y=LoanFee), data = subset(loans, LoanYear > 2005))+
  geom_boxplot(fill = "orange")+scale_x_discrete(breaks = seq(2006,2014,1)) +
  xlab("Loan year") +
  scale_y_continuous(breaks = seq(0,0.15,0.01), labels = percent)
```




```
ggplot(aes(x=factor(LoanYear), y=BorrowerRate),
       data = subset(loans, LoanYear > 2005))+geom_boxplot(fill = "yellow")+
  scale_x_discrete(breaks = seq(2006,2014,1)) +xlab("Loan year")+
  scale_y_continuous(breaks = seq(0,0.5,0.05), labels = percent)
```



The first chart above shows the trend of loan fees, which is the difference between loan APR and loan interest rate in the dataset (presented as percentage). We can see that starting form 2009-2010, Prosper has been charging higher fees with a median of about 3-4 percent which is a significant spike compared to the fee of less than 1% in 2006-2007. Although we also see a much greater spread after 2010. This is somehow in contrast to what I have read before. According to Investopedia.com, "*Following the financial crisis of 2008, the government passed new laws circumscribing how lenders could be compensated, and public pressure provided an additional incentive for lenders to reign in the practices that had made them rich during the housing boom. Origination fees shrunk to an average of 1% or less, and since borrowers in 2016 come armed with much more information about going rates compared to borrowers in 2006, the days of making easy money from yield spread premium are over.*" Further research is required to determine if the statement only applies to mortgage fees or not. And if not, why have the fees gone up after 2010?

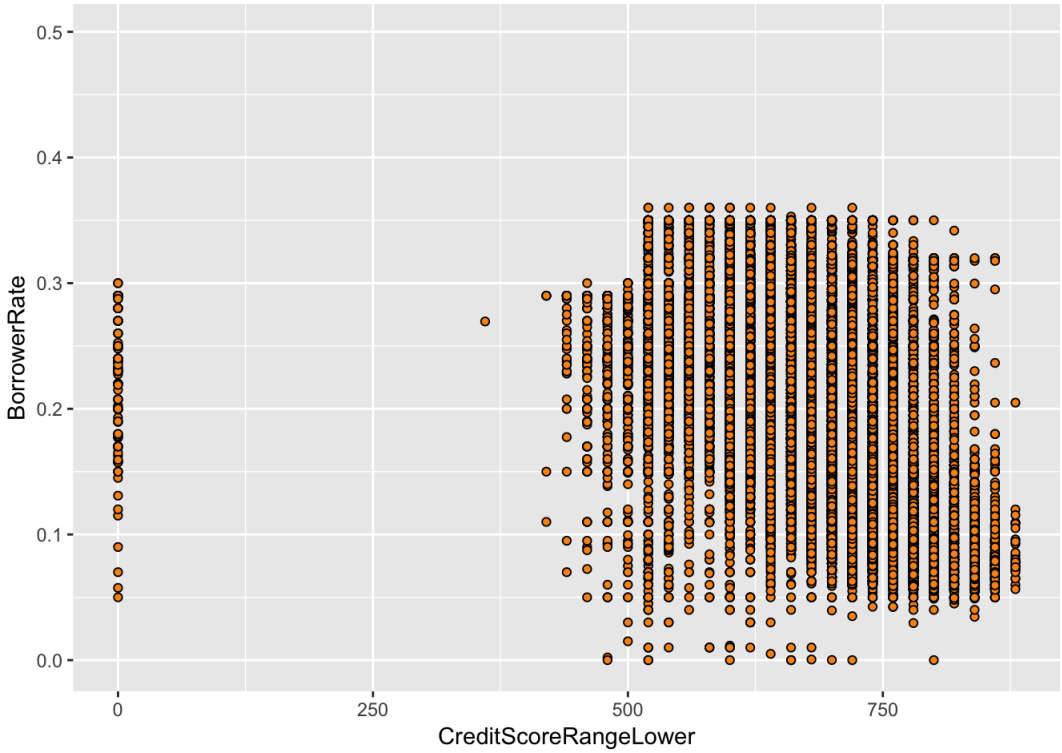
On the other hand, interest rates (displayed in the second chart), do not follow a similar trend. After some increase between 2009-2011, we

see a consistent decrease in interest rates and a decrease in variance between 2011-2014.

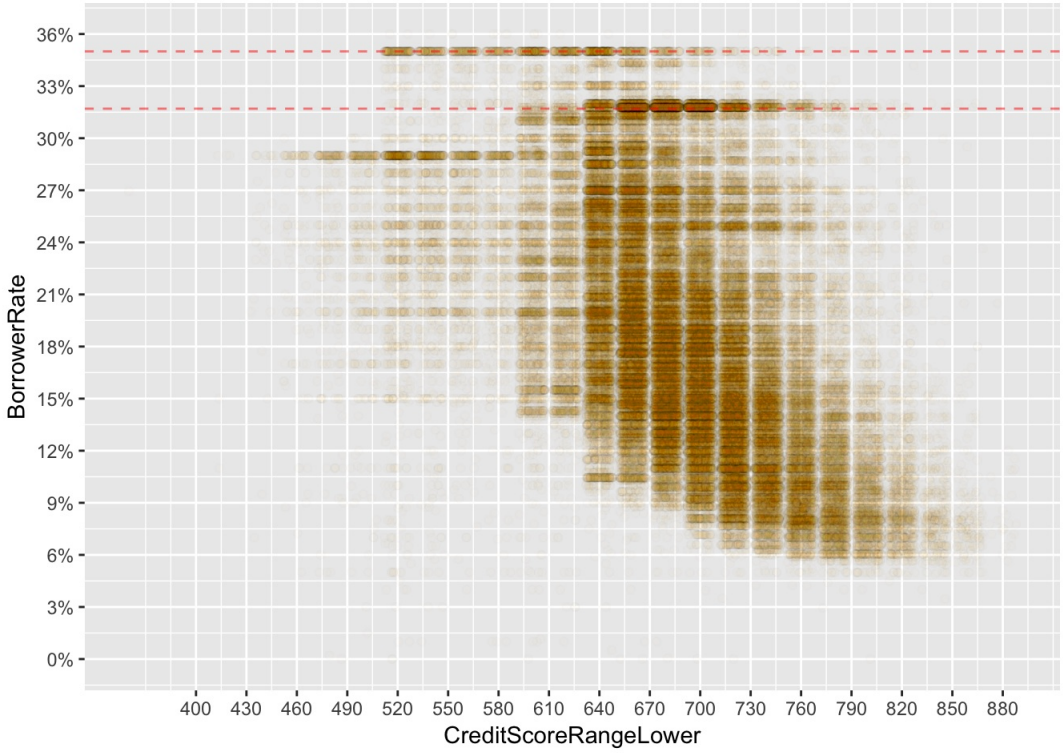
Interest Rate Relationships

Looking at Prosper website, we can see that the main factor that determines the interest rate on a loan is Prosper rating. For each rating, there is a range of 11% to 15% between the Min interest rate and Max interest rate. In this section, I will look at some other factors that might impact the interest rate within each rating.

```
ggplot(aes(x=CreditScoreRangeLower, y= BorrowerRate), data = loans)+
  geom_point(color = I("black"), fill = I("#F79420"), shape = 21)
```




```
ggplot(aes(x=CreditScoreRangeLower, y= BorrowerRate),
       data = subset(loans, CreditScoreRangeLower > 0))+
  geom_jitter(alpha = 0.01, color = I("black"), fill = I("#F79420"), shape = 21)+
  scale_x_continuous(breaks = seq(400,900, 30)) +
  scale_y_continuous(breaks = seq(0, 0.5, 0.03), labels = percent) +
  geom_hline(yintercept = c(0.35, 0.317), linetype =2, color = "red", alpha = 0.5)
```




```
with(subset(loans, CreditScoreRangeLower > 0),
     cor.test(CreditScoreRangeLower, BorrowerRate))
```

```
##
##  Pearson's product-moment correlation
##
## data:  CreditScoreRangeLower and BorrowerRate
## t = -188.07, df = 113210, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4923346 -0.4834578
## sample estimates:
##        cor
## -0.4879088
```

Looking at the first chart on credit score vs interest rate, we see a lot of overplotting, especially due to the integer nature and the intervals of 20 in credit score. So again I had to use jitter and transparency to make the chart easier to interpret.

The second chart shows a somewhat exponential relationship between these 2 variables in a negative direction within the range of 640-880 credit score. As expected, the higher scorecards have got lower interest rates. We see a horizontal band at 35% interest rate with a range of score cards between 520-700 and a horizontal band at 31.7% with a range of score cards mostly between 640-760. As mentioned before, these 2 interest rates were among the most frequent interest rates on Prosper loans.

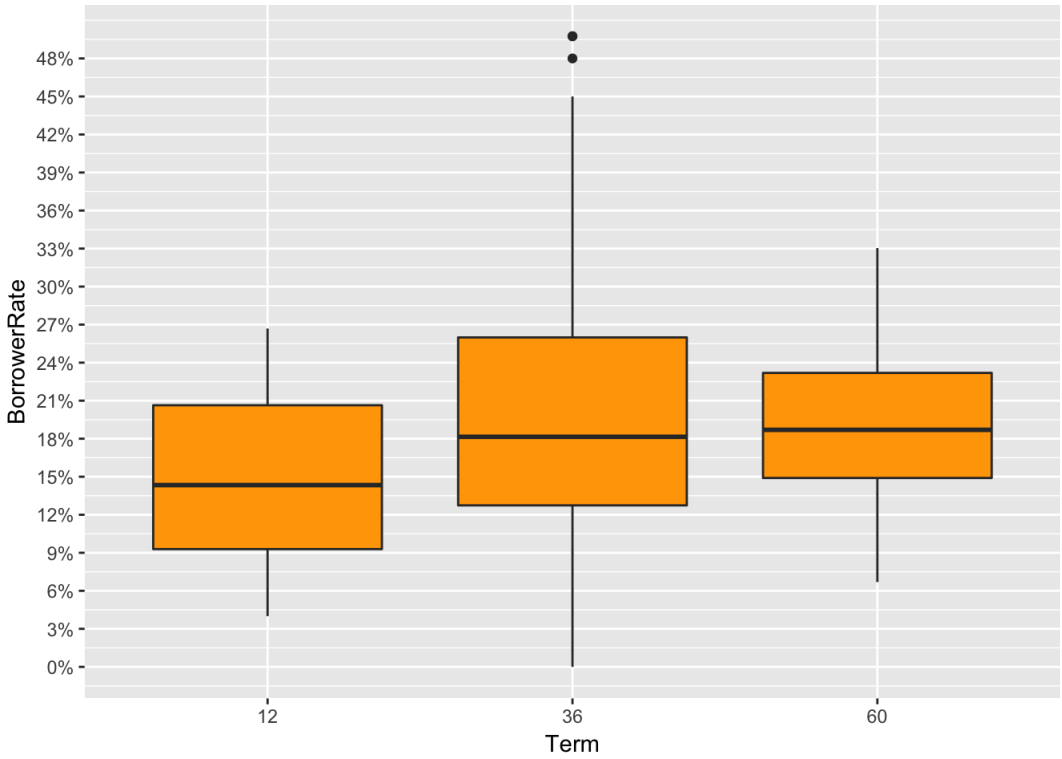
The Pearson's correlation test shows a moderate linear correlation between the 2 variables with the value of -0.49.

Interest Rates vs. Loan Term

```
by(loans$BorrowerRate, loans$Term, summary)
```

```
## loans$Term: 12
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0400  0.0929  0.1434  0.1501  0.2064  0.2669
## ------------------------------------------------------------
## loans$Term: 36
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.1274  0.1815  0.1935  0.2599  0.4975
## ------------------------------------------------------------
## loans$Term: 60
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0669  0.1490  0.1870  0.1930  0.2319  0.3304
```

```
#Converting Term variable to factor variable
loans$Term <- factor(loans$Term, levels = c(12, 36, 60))

ggplot(aes(x = Term, y = BorrowerRate), data = loans) +
  geom_boxplot(fill = "orange") +
  scale_y_continuous(breaks = seq(0, 0.5, 0.03), labels = percent)
```



We have 3 loan terms in the data set: 12, 36, and 60 month. First I converted this variable from int to factor variable. Looking at the interest rate at each term level, we do not see a clear correlation. While the median interest rate for 36 month loans (18%) is about 3.7 percentage points higher than 12 month loans (which have a count of less than 2% of 36 month loans), we do not see any significant difference between the interest rates for 36 month and 60 month loans other than the fact that the variance is much higher on 36 month loans (the high number of 36 months loans can explain this high variance). So we cannot really conclude any meaningful impact from loan term on interest rate.

Interest Rates vs. Recommendations

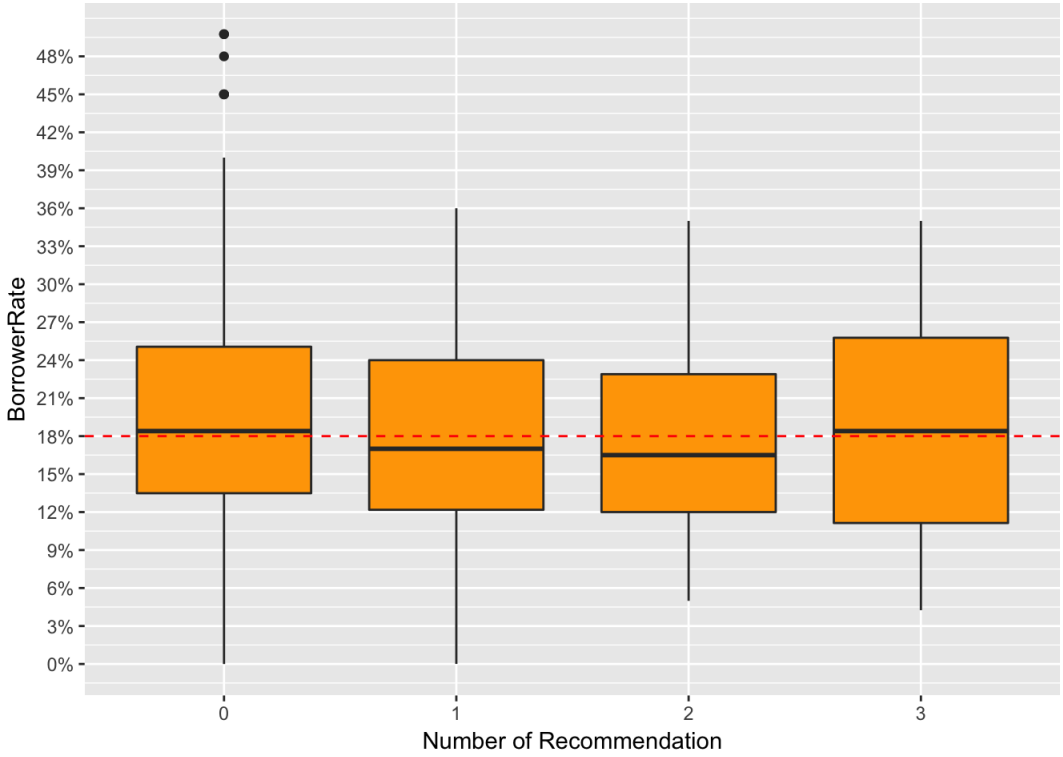
```
summary(loans$Recommendations)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.00000  0.00000  0.00000  0.04803  0.00000 39.00000
```

```
table(loans$Recommendations)
```

```
##
##      0      1      2      3      4      5      6      7      8      9
## 109678   3516    568    108     26     14      4      5      3      6
##     14     16     18     19     21     24     39
##      1      2      2      1      1      1      1
```

```
ggplot(aes(x=factor(Recommendations), y= BorrowerRate),
       data = subset(loans, Recommendations < 4))+ geom_boxplot(fill = "orange") +
  scale_y_continuous(breaks = seq(0, 0.5, 0.03), labels = percent) +
  xlab("Number of Recommendation") +
  geom_hline(yintercept = 0.18, color = "red", linetype = 2)
```




```
with(subset(loans, Recommendations < 4),
     by(BorrowerRate, Recommendations, summary))
```

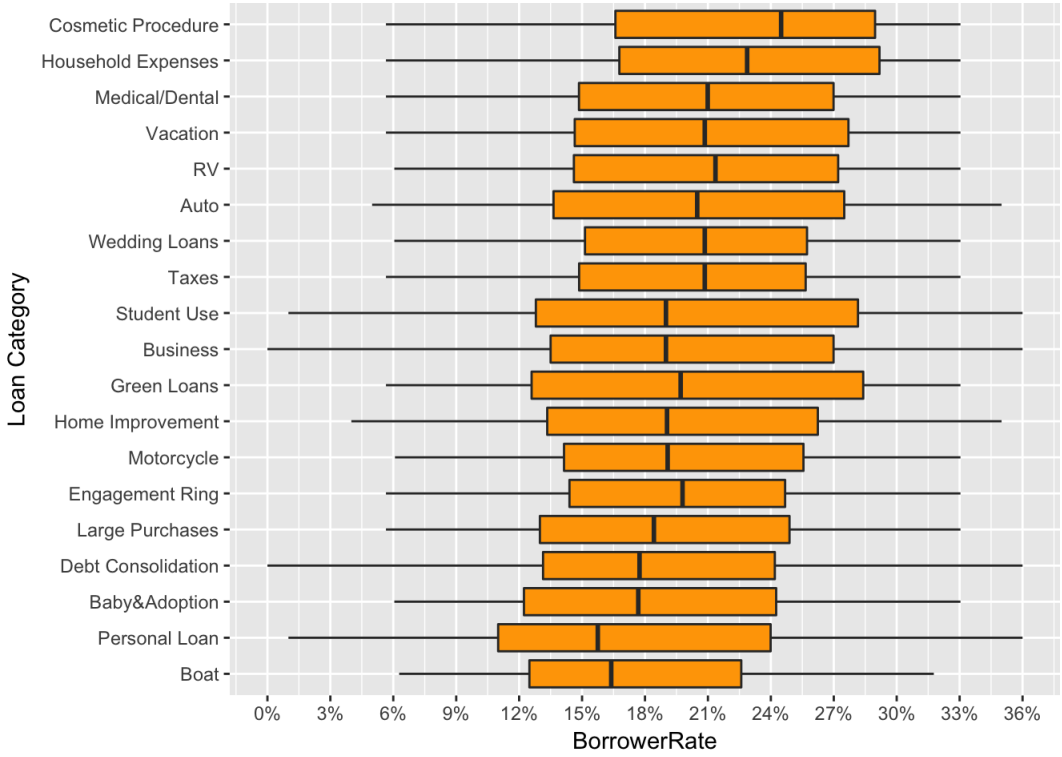
```
## Recommendations: 0
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.1349  0.1840  0.1932  0.2506  0.4975
## ------------------------------------------------------------
## Recommendations: 1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0001  0.1218  0.1700  0.1840  0.2400  0.3600
## ------------------------------------------------------------
## Recommendations: 2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0499  0.1200  0.1650  0.1785  0.2289  0.3500
## ------------------------------------------------------------
## Recommendations: 3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0425  0.1114  0.1840  0.1924  0.2577  0.3500
```

While more than 96% of loans did not record any recommendations by the borrower, I looked at this variable to see if the remaining 4% got a different interest rate. By looking at the counts of each number of recommendations in the second table, I decided to creat the boxplot only for recommendation levels between 0-3, since there were very few loans with more than 3 recommendations.

The boxplot does not indicate any clear correlation between interest rate and recommendations for the loan. The median interest rate for loans with 1 and 2 recommendations is slightly lower than loans with no recommendations. However, that median goes back up for loans with 3 recommendaations.

Interest Rates vs. Loan Category

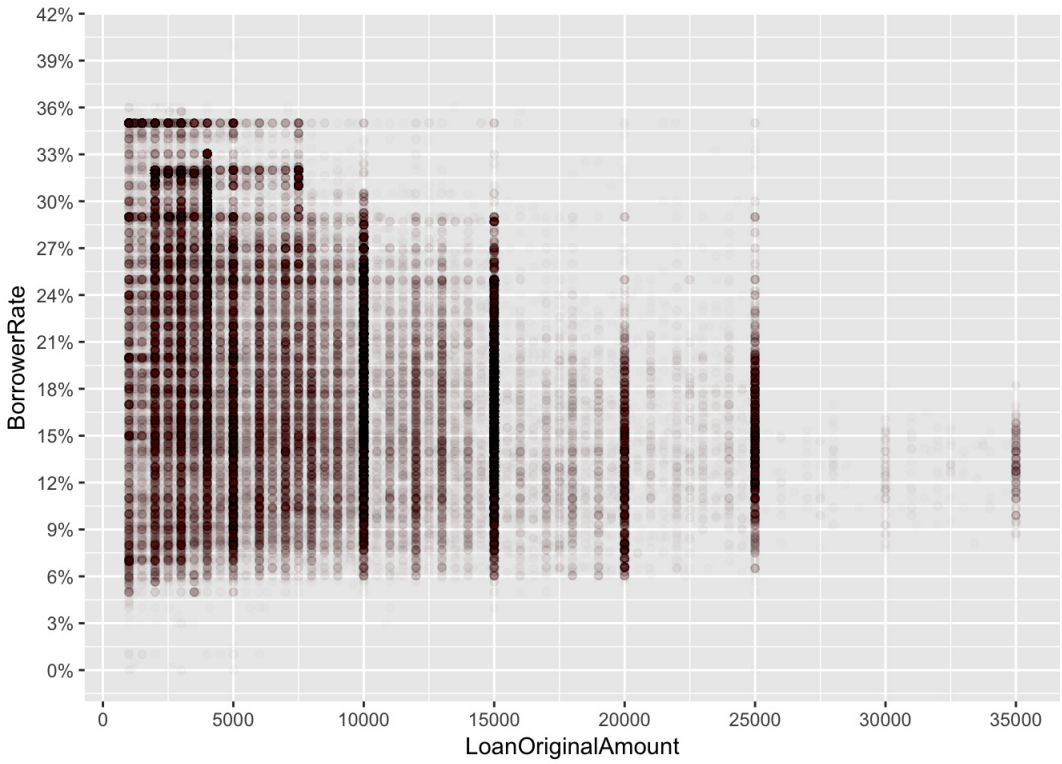
```
ggplot(aes(x=reorder(LoanCategory, BorrowerRate, fun=median), y=BorrowerRate),
       data = subset(loans, !LoanCategory %in% c("Other", "Not Available")))+
  geom_boxplot(fill = "orange") + coord_flip() +
  scale_y_continuous(breaks = seq(0,0.5, 0.03), labels = percent) +
  xlab("Loan Category")
```



It is interesting to see that “Cosmetic Procedure” (which has a very low count) is the category with the highest median interest rate on the loans. The Interquartile range for this category is 18%-29% with a median of 25%. On the other hand, we see that “Debt Consolidation” loans (which is the most frequent loan category by far) have a median of 18% which is lower than overall average interest rate and only higher than the median of three other categories. “Debt Consolidation” and Business" loans (which is also among the most frequent loan categories) have the highest spread of interest rate.

Interest Rates vs. Loan Amount

```
ggplot(aes(x=LoanOriginalAmount, y= BorrowerRate), data = loans)+
  geom_jitter(alpha = 0.01, color = I("black"), fill = I("darkred"), shape = 21)+
  scale_x_continuous(breaks = seq(0,35000, 5000)) +
  scale_y_continuous(breaks = seq(0, 0.5, 0.03),
                     labels = percent, limits = c(0,0.4))
```



It looks like there is a slight negative relationship between loan amount and loan interest rate. Loans with higher amounts have had lower interest rates and vice versa. We can see that almost all the loans with 35% interest rate were less than $7500 and on the other end, almost

all the $35000 loans had an interest rate between 9%-18%. We might attribute this to the fact that usually people with higher income apply for higher loan amounts and since people with higher income generally tend to have higher credit score, they get lower interest rates.

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 two most important relationships that were noticed were the exponential negative relationship between credit score and interest rate (lower credit score, higher interest rate) and also the impact of having delinquent accounts on lowering borrower's credit score.

We also saw that loan amount is to some extent correlated with annual income of a borrower. People with lower income tend to apply for lower amounts of loan. On the other hand, at high levels of loan amounts (above $25,000), we do not have any borrower with less than $100,000 in annual income.

There is a slight correlation between loan amount and interest rate in the negative direction. Loans between $20,000 - $35,000 rarely had an interest rate above 24%, while there were a lot of loans less than $7500 with an interest rate above 30%.

No strong correlation was found between credit score versus open credit lines and annual income, and also between interest rate versus loan term or number of recommendations.

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

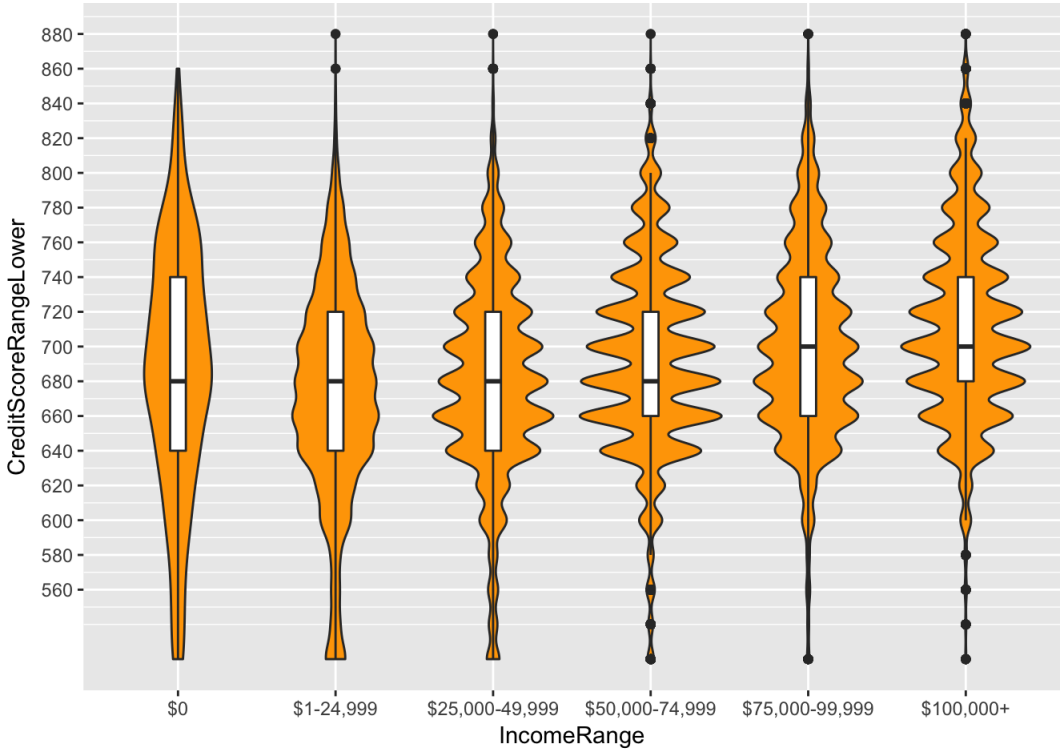
It was interesting to see how similar the states' trends are when it comes to median annual income versus median amount loan. While the states of North Dakota and Iowa are at the bottom of the list in both variables, New Jersey, Virginia, and Maryland are the states among the top with regard to both median income and median amount of loan. The states in between are also very close in these two variables.

What was the strongest relationship you found?

The strongest quantifiable relationship that I found was that of interest rate on loans versus the borrower's credit score. I noticed an exponential negative correlation between the two variables especially between credit score 600 - 880. Although the interest rates with the highest count in the dataset (31.7% and 35%) do not fit this exponential model.

Multivariate Plots Section

```
# adjusting the order of the factor variable IncomeRange
loans$IncomeRange <- ordered(loans$IncomeRange,
                             levels = c("$0", "$1-24,999", "$25,000-49,999",
                                        "$50,000-74,999", "$75,000-99,999",
                                        "$100,000+"))

ggplot(aes(y= CreditScoreRangeLower, x=IncomeRange),
       data = subset(loans, !is.na(IncomeRange) & !IncomeRange %in%
                       c("Not displayed", "Not employed")  &
                                                  CreditScoreRangeLower>0)) +
  geom_violin(fill = "orange") +
  geom_boxplot(width=0.1) + scale_y_continuous(breaks = seq(560, 880, 20))
```

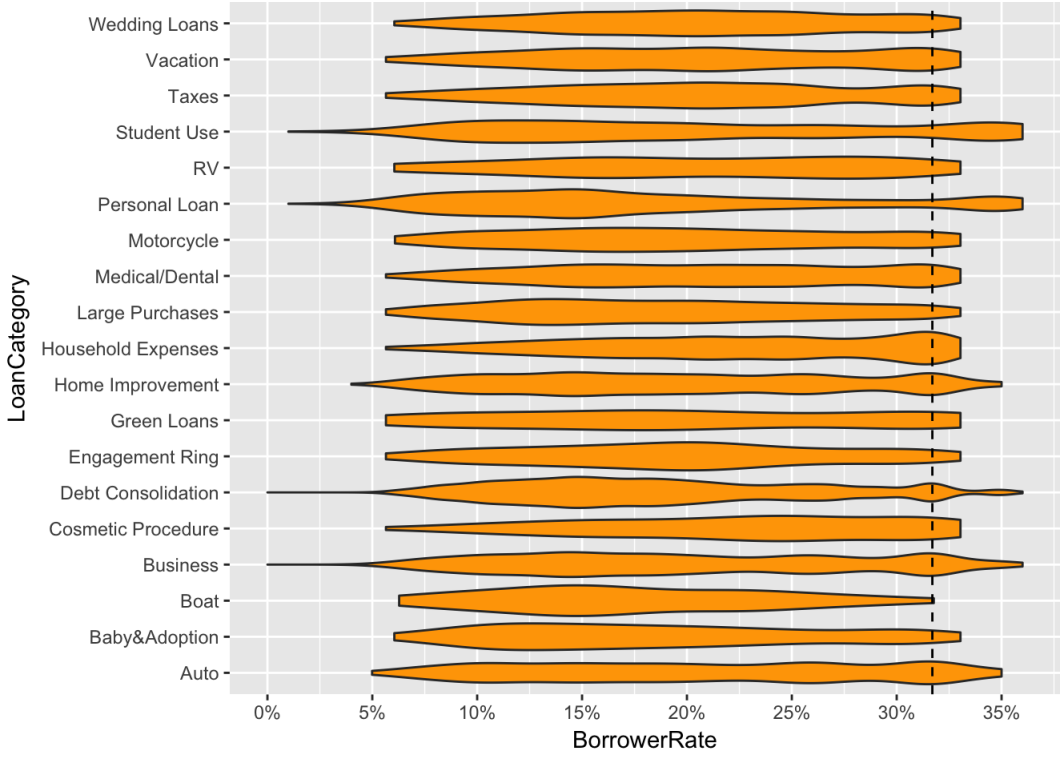


Looking at the violin chart of Credit Score against the 6 buckets of income range included in the data set reveals a clearer relationship between income and credit score. For credit scores above 750, we can see that “$100,000+” has a higher width compared to other buckets except for $0 bucket. (As mentioned before, credit score values are in intervals of 20 and so we see ups and downs on the violin chart at intervals of 20). Also, looking at the median credit score for each bucket, we can see that the median slightly goes up as the income range increases. The same goes for the credit scores less than 600 where we see a smaller width as income goes up.

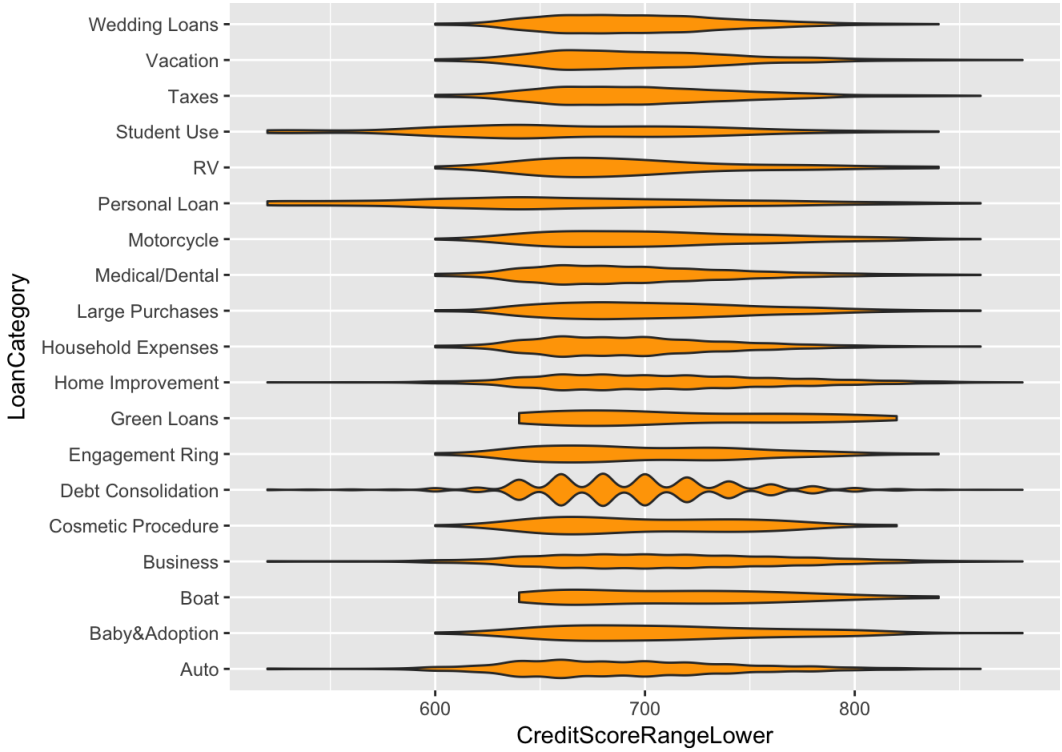
```
ggplot(aes(x=StatedMonthlyIncome*12,
           y=CreditScoreRangeLower),
       data = subset(loans, StatedMonthlyIncome <
                     quantile(StatedMonthlyIncome, 0.95) &
                     CreditScoreRangeLower > 0 &
                     !is.na(DoesBorrowerHaveDelinquency)))+
  geom_jitter(aes(color = DoesBorrowerHaveDelinquency), size = 0.3) +
  scale_x_continuous(breaks = seq(0,150000,25000), labels = dollar) +
  scale_y_continuous(breaks = seq(0, 900, 50)) +
  labs(x="Annual Income", y = "Credit Score")+
  scale_color_brewer(type = "qual", guide = guide_legend(title = "Delinquency"))
```

Here we can see the significant negative impact of having delinquency on credit score for every level of income. “FALSE” (green) represents the loans with no delinquent accounts for the borrower and “TRUE” (purple) represents the loans with delinquent accounts. The vast majority of delinquency belongs to people with income lower than $60,000 and we rarely see an income of above $125,000 with delinquent accounts.

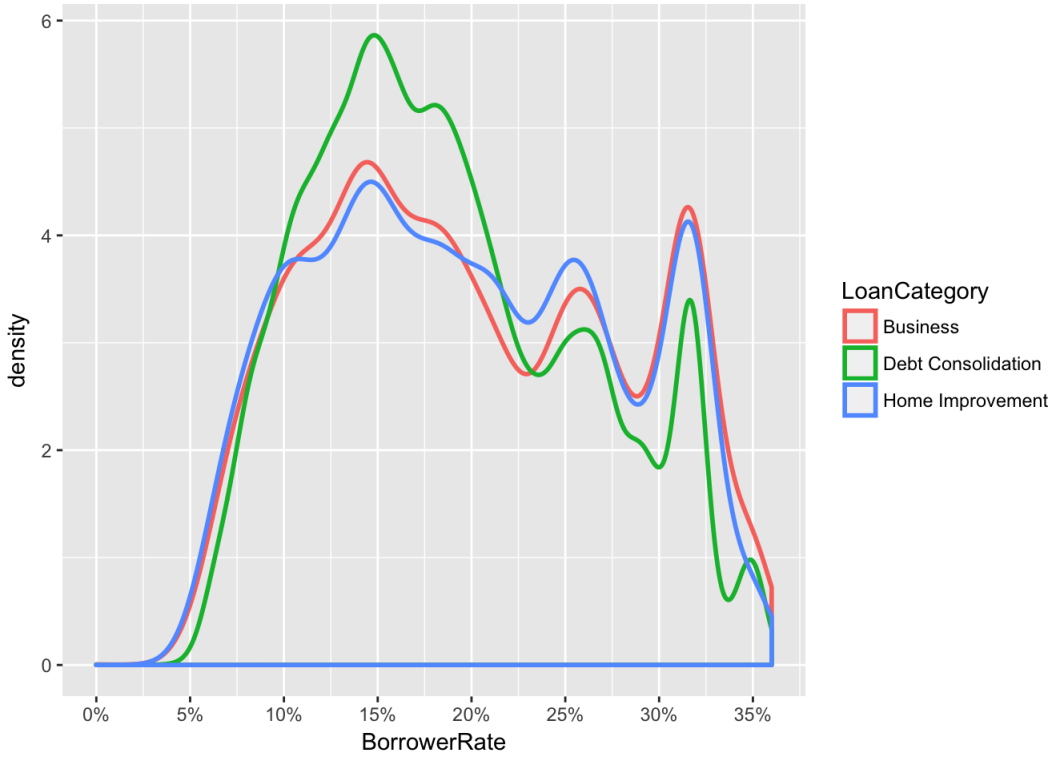
```
ggplot(aes(x= LoanCategory, y= BorrowerRate),
       data = subset(loans, !LoanCategory %in% c("Not Available", "Other"))) +
  geom_violin(fill = "orange") +
  scale_y_continuous(labels = percent, breaks = seq(0,0.4,0.05)) +
  geom_hline(yintercept = 0.317, linetype = 2) + coord_flip()
```




```
ggplot(aes(y= CreditScoreRangeLower, x= LoanCategory),
       data = subset(loans, !LoanCategory %in% c("Not Available", "Other"))) +
  geom_violin(fill = "orange") + coord_flip()
```



```
ggplot(aes(x = BorrowerRate), data = subset(loans, LoanCategory %in%
                                              c("Home Improvement", "Business",
                                                "Debt Consolidation"))) +
  geom_density(aes(color = LoanCategory), size = 1) +
  scale_x_continuous(labels = percent, breaks = seq(0, 0.4, 0.05))
```

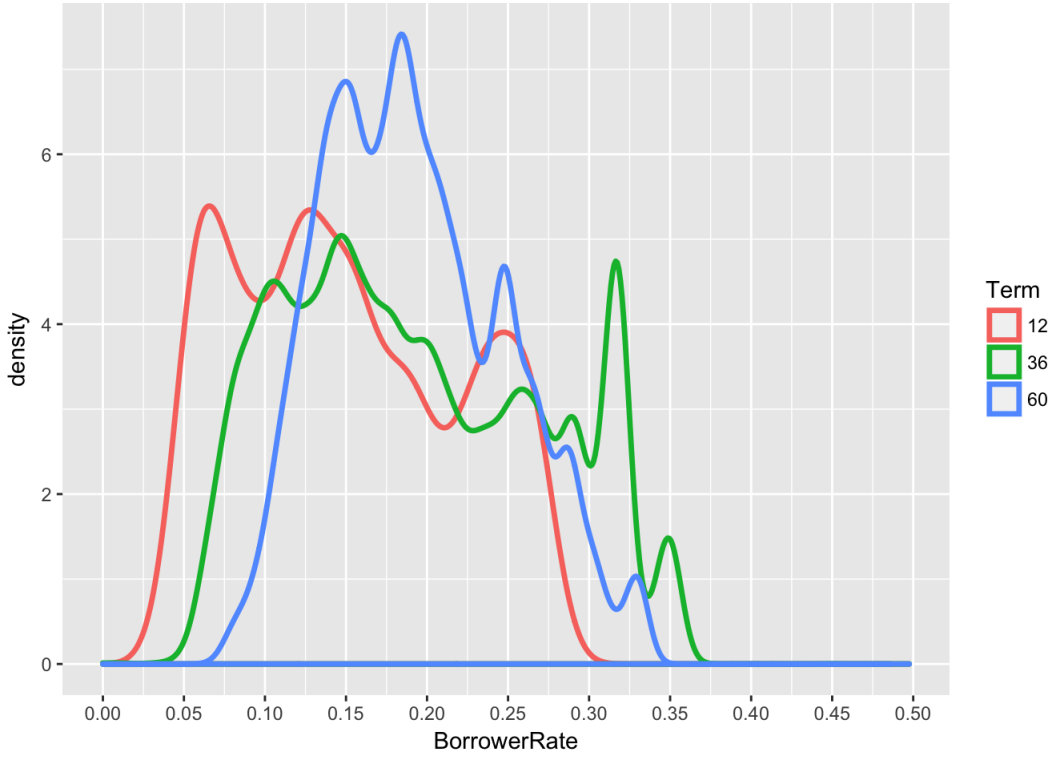


From the first violin chart, it looks like that the category with the highest concentration around the most frequent interest rate, 31.77% (dispalyed with a dashed line), is “Household Expenses”. “Student Use” is the category with the highest proporation of high interest rate (above 34%). The category with the lowest density around both 31.77% and above 34% interest is “Boat”.

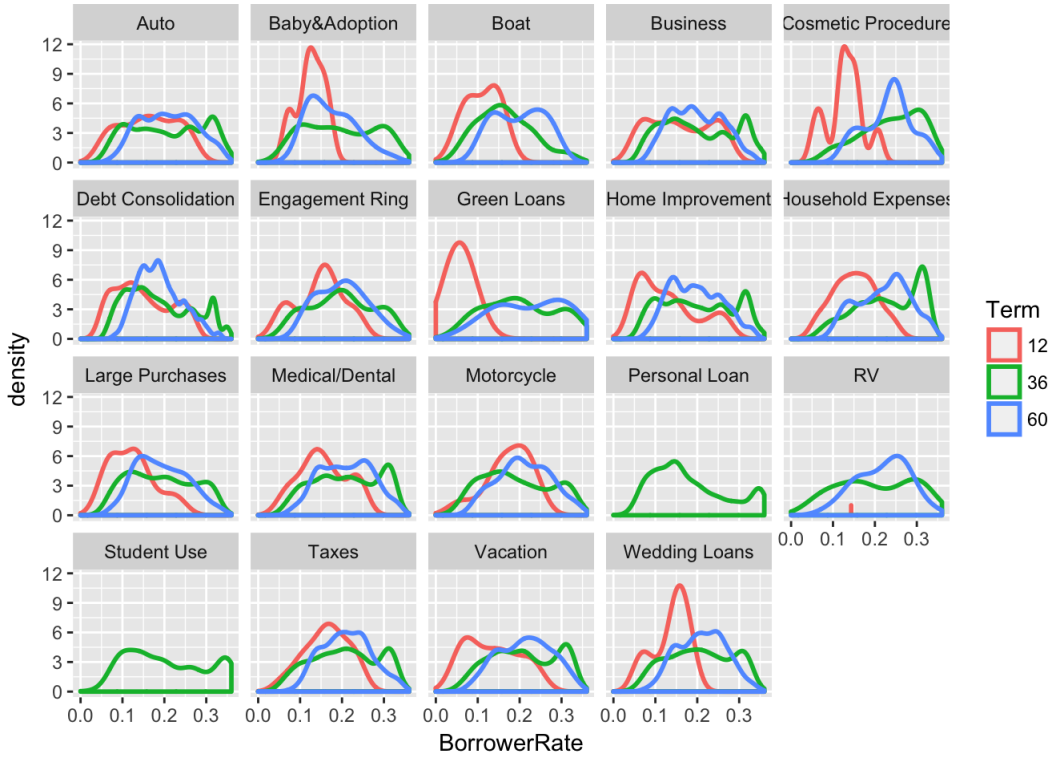
By looking at the violin chart for credit score of people who applied for “Student Use” loans and comparing that with violin chart of credit score for other loans, we can see that the gap in credit score can explain the gap in interest rate between “Student Use” loans and other loans to some degree (as “Student Use” is the one of the categories with higher density around credit scores lower than 600). However, credit score does not explain this for “Household Expenses” loans.

I was also interested specifically in “Debt Consolidation”, “Home Improvement”, and “Business” loans, which combined together, make up close to 70% of the loans. We can see from the third chart that all three categories have the highest density around 15%. Although for Business and Home Improvement the densities around 15% loans and 31.7% loans are very close.

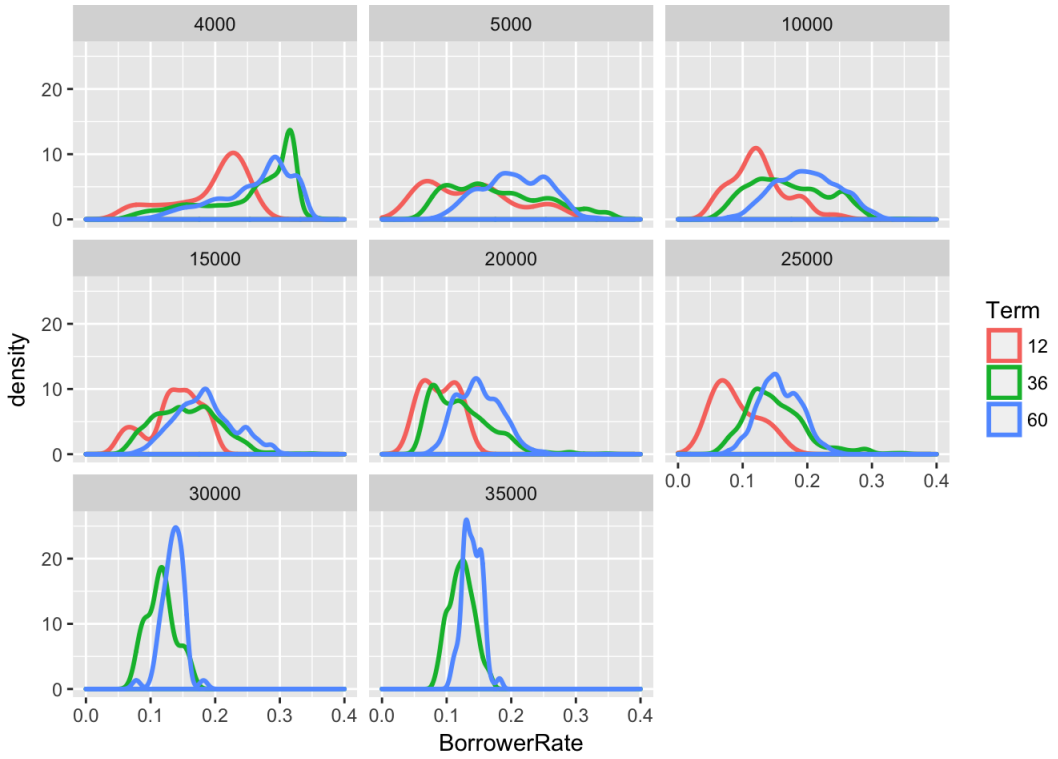
```
ggplot(aes(x= BorrowerRate), data = loans) +
  geom_density(aes(color = Term), size = 1.2) +
  scale_x_continuous(breaks = seq(0, 0.5, 0.05))
```




```
#Faceting by Loan Category
ggplot(aes(x= BorrowerRate), data = subset(loans, !LoanCategory %in%
                                           c("Not Available", "Other"))) +
  geom_density(aes(color = Term), size = 1) +
  scale_x_continuous(breaks = seq(0, 0.5, 0.1)) + facet_wrap(~LoanCategory)
```




```
#Faceting by Loan Amount
ggplot(aes(x= BorrowerRate), data = subset(loans, LoanOriginalAmount %in%
                                           c(4000, 5000, 10000, 15000, 20000,
                                             25000, 30000, 35000))) +
  geom_density(aes(color = Term), size = 1) +
  scale_x_continuous(breaks = seq(0, 0.5, 0.1)) + facet_wrap(~LoanOriginalAmount)
```

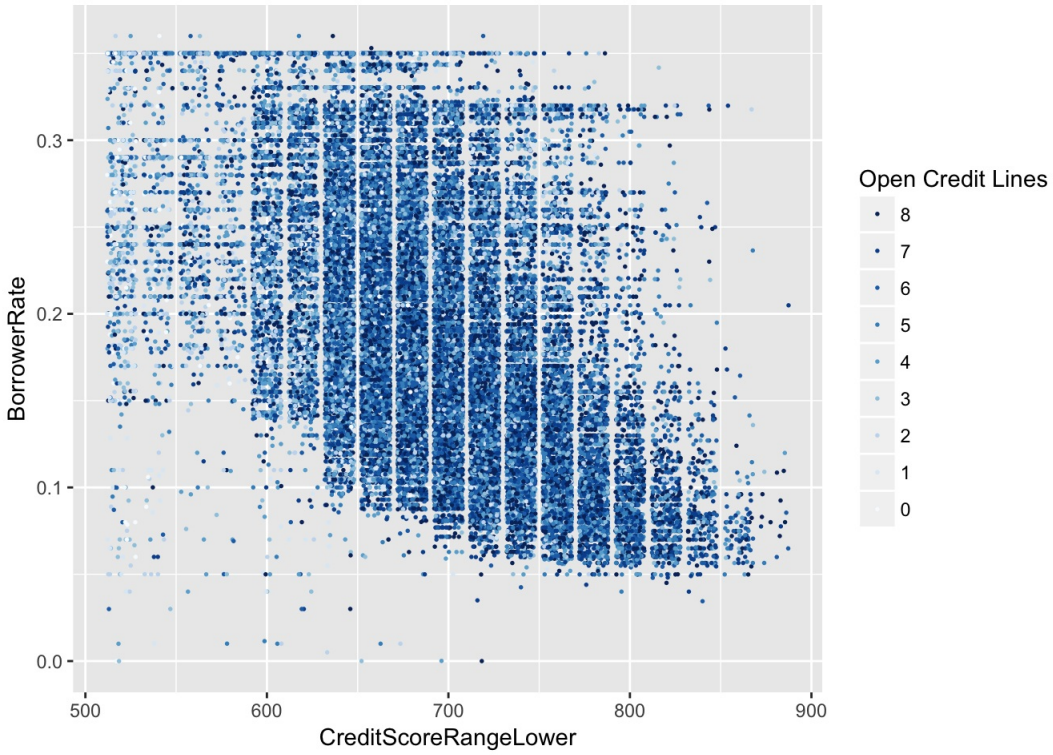


Based on the first density chart above, 36 month loans had a higher proportion of interest rates above 30% than 60 month and 12 month loans. On the other hand, 12 month loans had a higher density for interest rates lower than 10% followed by 36 month loans. The range in which we see a significantly higher density for 60 month loans is between 15%-20%.

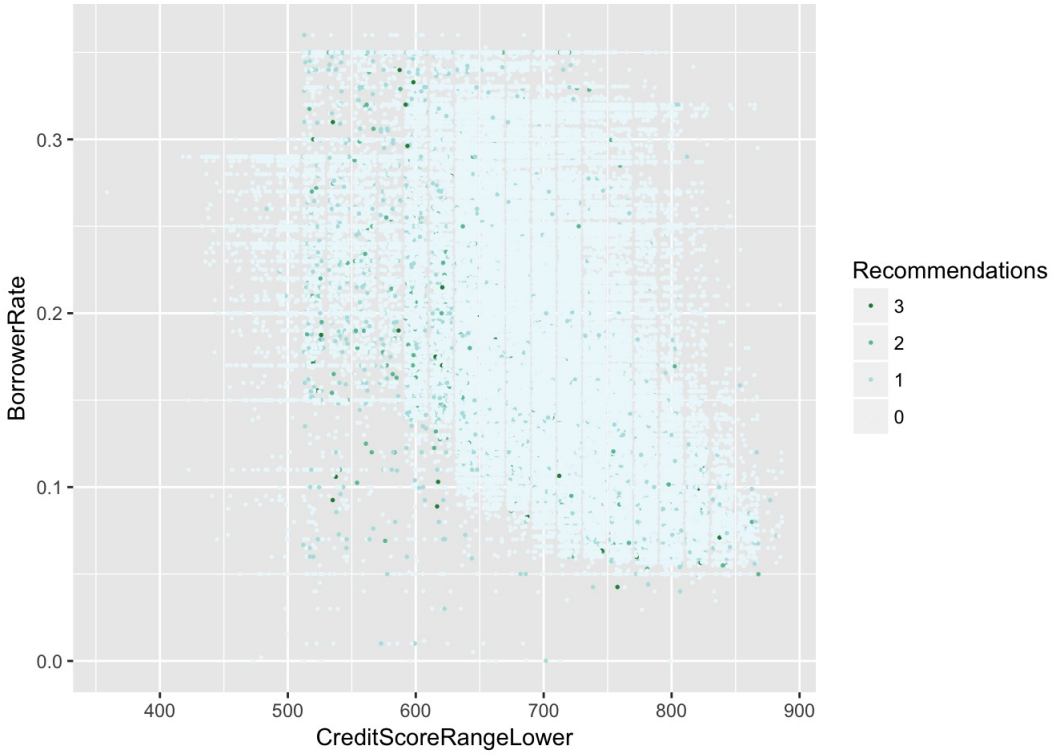
If we split up the first density chart based on loan category (second chart) we see that for all categories of loan (except for Geen Loans, which has the second lowest count), 36 month loans had a higher proportion of interest rates above 30% compared to 60 month and 12 month loans. On the other hand, 12 month loans had a higher proportion of interest rates below 10% in all categories except for Motorcycle. It is interesting to see that “Personal Loan” and “Student Use” loans only had 36 month terms.

For the third chart, I looked at the density chart for the most frequent loan amounts in the dataset. Although this is a continuous variable, the combined count of the selected 8 loan amounts accounts for about 50% of the data. Faceting the first density chart over loan amount shows that the main driver of the higher density of 36 moth loans for above 30% interest rate is $4000 loans (which has the highest count in the data). For interest rates of lower than 10%, 12 month loans almost always have a higher densit compared to 36 and 60 month loans (execpt for high amount loans of 30,000 and $35,000). The chart also reconfirms the results that higher amount loans had a higher proporation of lower interest rates compared to lower amount loans.

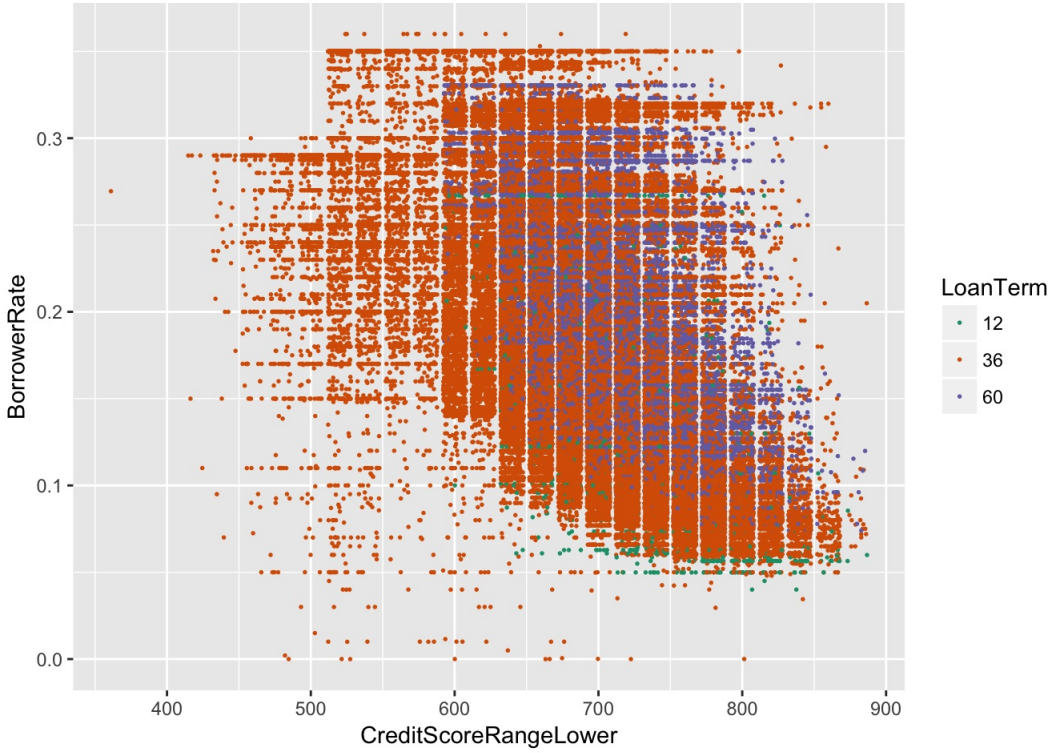
```
ggplot(aes(x=CreditScoreRangeLower, y=BorrowerRate),
       data = subset(loans, CreditScoreRangeLower >0 & OpenCreditLines < 9)) +
  geom_jitter(aes(color = factor(OpenCreditLines)), size = 0.3)+
  scale_color_brewer(type = "seq", guide =
                       guide_legend(reverse = T, title = "Open Credit Lines"))
```



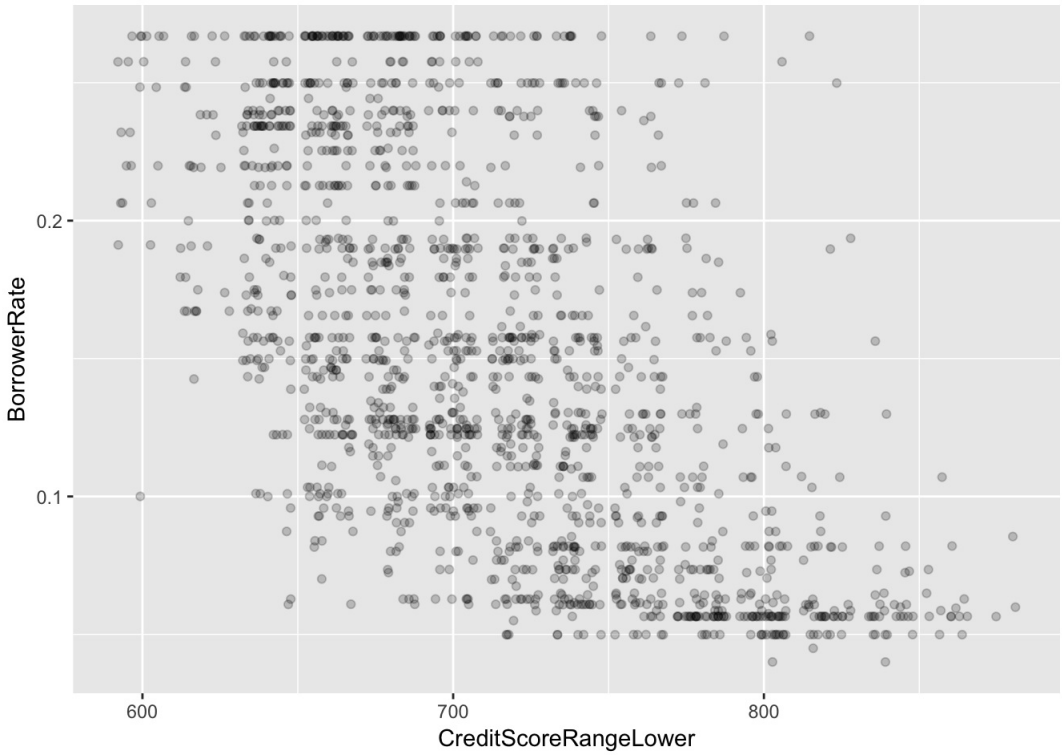

```
ggplot(aes(x=CreditScoreRangeLower, y=BorrowerRate),
       data = subset(loans, CreditScoreRangeLower >0 & Recommendations < 4)) +
  geom_jitter(aes(color = factor(Recommendations)), size = 0.3)+
  scale_color_brewer(type = "seq", palette = 2, guide =
                       guide_legend(reverse = T, title = "Recommendations"))
```




```
ggplot(aes(x=CreditScoreRangeLower, y=BorrowerRate),
       data = subset(loans, CreditScoreRangeLower >0)) +
  geom_jitter(aes(color = factor(Term)), size = 0.3)+
  scale_color_brewer(type = "qual", palette = 2, guide =
                       guide_legend(title = "LoanTerm"))
```




```
ggplot(aes(x=CreditScoreRangeLower, y=BorrowerRate),
       data = subset(loans, CreditScoreRangeLower >0 & Term== 12)) +
  geom_jitter(alpha = 0.2)
```



No clear relationship is seen between Open Credit Lines and interest rate if we account for constant credit score value. The only thing noticed here is that most of the lighter color points which represent less than 1 open credit line are concentrated on the left side of the chart, meaning around credit scores lower than 550.

The same goes for Recommendations vs interest rate where no clear trend is noticeable when holding credit score value constant (I looked at a subset of data where number of recommendation is less than 4, because the count for loans with recommendations above 4 was very low and I did not want to make the chart too busy). Interest rate does not vary much on the number of recommendations when holding credit score constant.

In the third chart, at first it looks like that most of 12 month loans have a lower interest rate for every level of credit score. However the last chart which is filtered to show only 12 month loans reveals that they are actually spread across the whole range of interest rates. So no real trend is noticed here either.

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?

The density chart of credit score based on annual income range showed a close relationship between these two variables, especially for credit scores above 800 and below 600. This relationship was not as clear in the scatter plot before.

Also, we were able to see that for every level of annual income, the credit score of people without delinquent accounts is significantly higher than credit score of people with delinquent accounts.

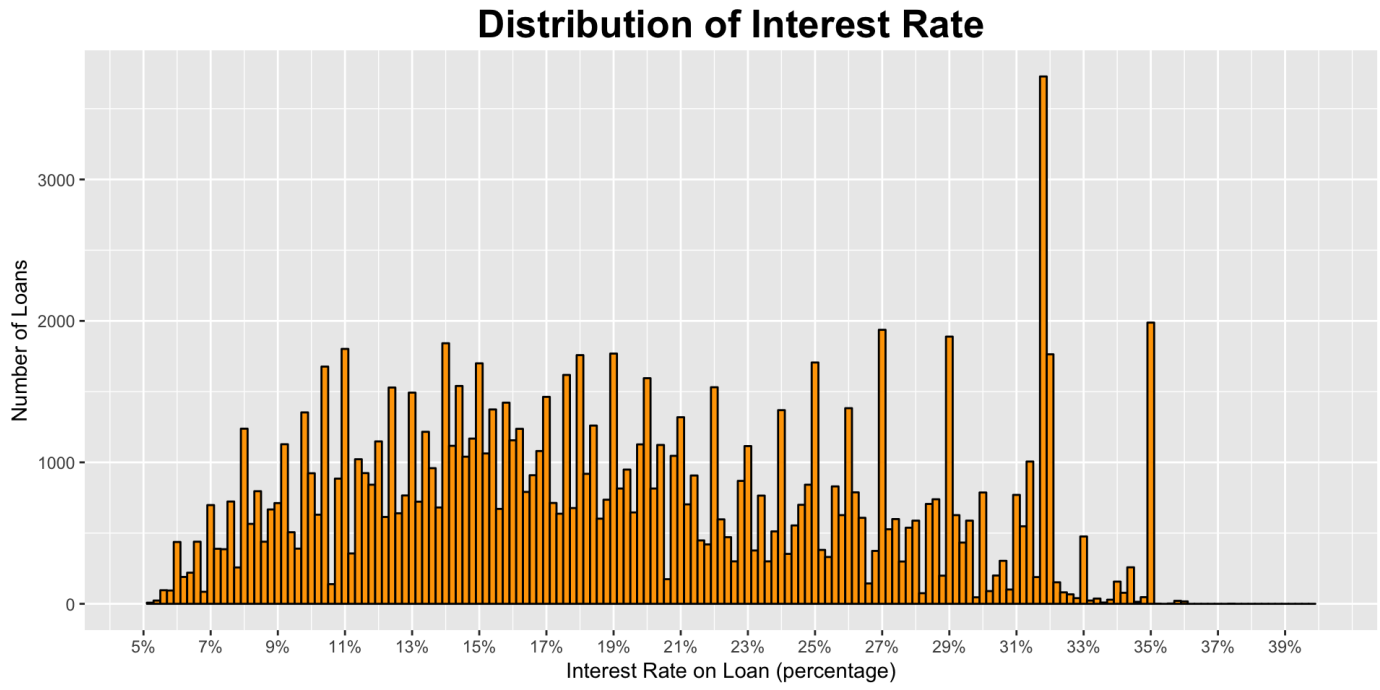
Were there any interesting or surprising interactions between features?

It was interesting to see that “Household Expenses” loans have, by far, the highest density around the most frequent interest rate, 31.7%. The next categories with high density around this interest rate have at least 50% lower density compared to “Household Expenses”. This is despite the fact that borrowers who applied for “Household Expenses” loans had a similar credit score density chart compared to borrowers of other categories.

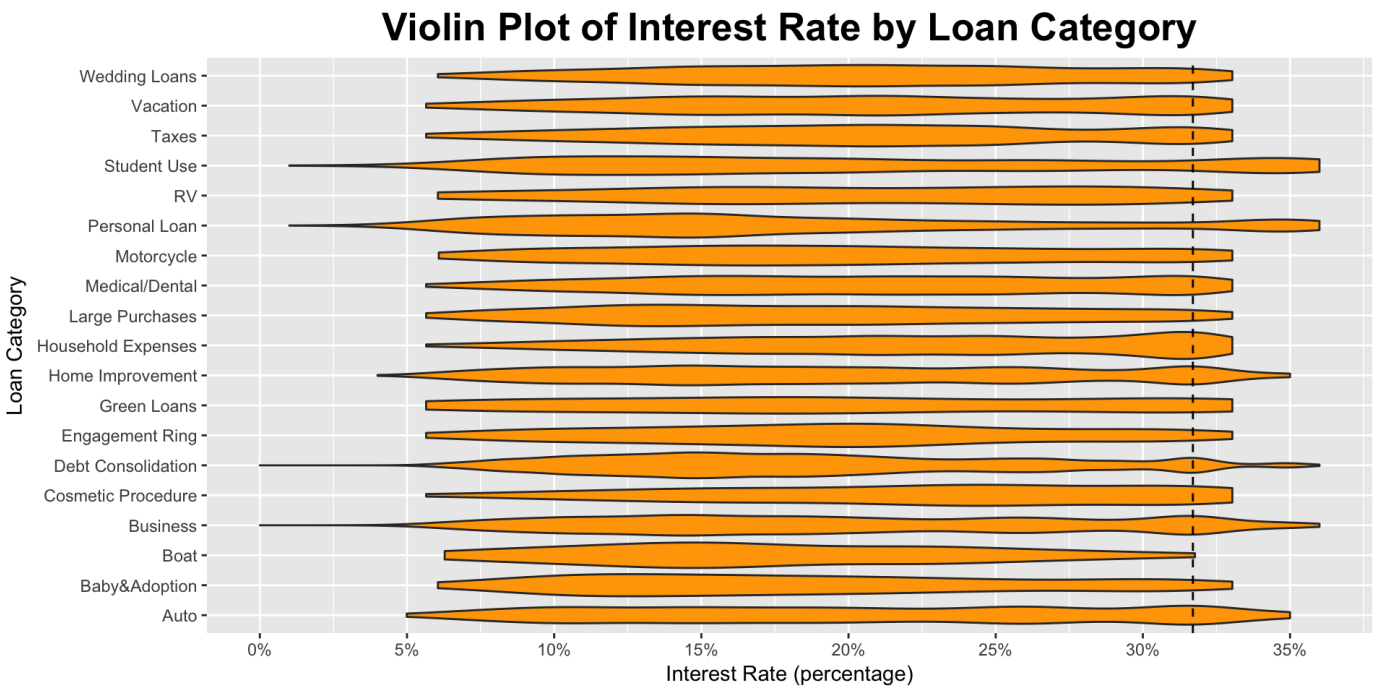
Final Plots and Summary

Plot One

```
ggplot(aes(x=BorrowerRate), data= loans)+
  geom_histogram(binwidth = 0.002, color = "black", fill = "orange")+
  scale_x_continuous(limits=c(0.05, 0.4),
                     breaks =seq(0.05, 0.4, 0.02), labels = percent) +
  labs(x="Interest Rate on Loan (percentage)", y="Number of Loans") +
  ggtitle("Distribution of Interest Rate")+
  theme(plot.title = element_text(face = "bold", size = 20, hjust = 0.5))
```




```
ggplot(aes(x= LoanCategory, y= BorrowerRate),
       data = subset(loans, !LoanCategory %in% c("Not Available", "Other"))) +
  geom_violin(fill = "orange") +
  scale_y_continuous(labels = percent, breaks = seq(0,0.4,0.05)) +
  geom_hline(yintercept = 0.317, linetype = 2) + coord_flip() +
  labs(y="Interest Rate (percentage)", x="Loan Category") +
  ggtitle("Violin Plot of Interest Rate by Loan Category")+
  theme(plot.title = element_text(face = "bold", size = 20, hjust = 0.5))
```

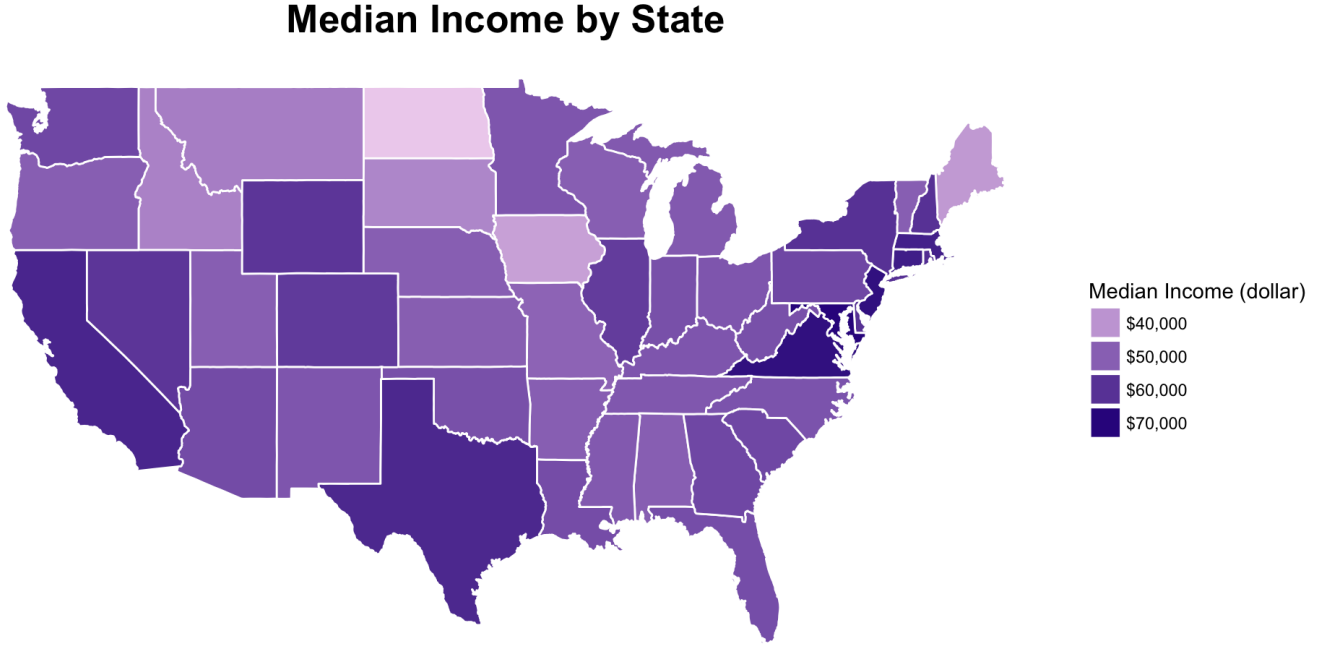


Description One

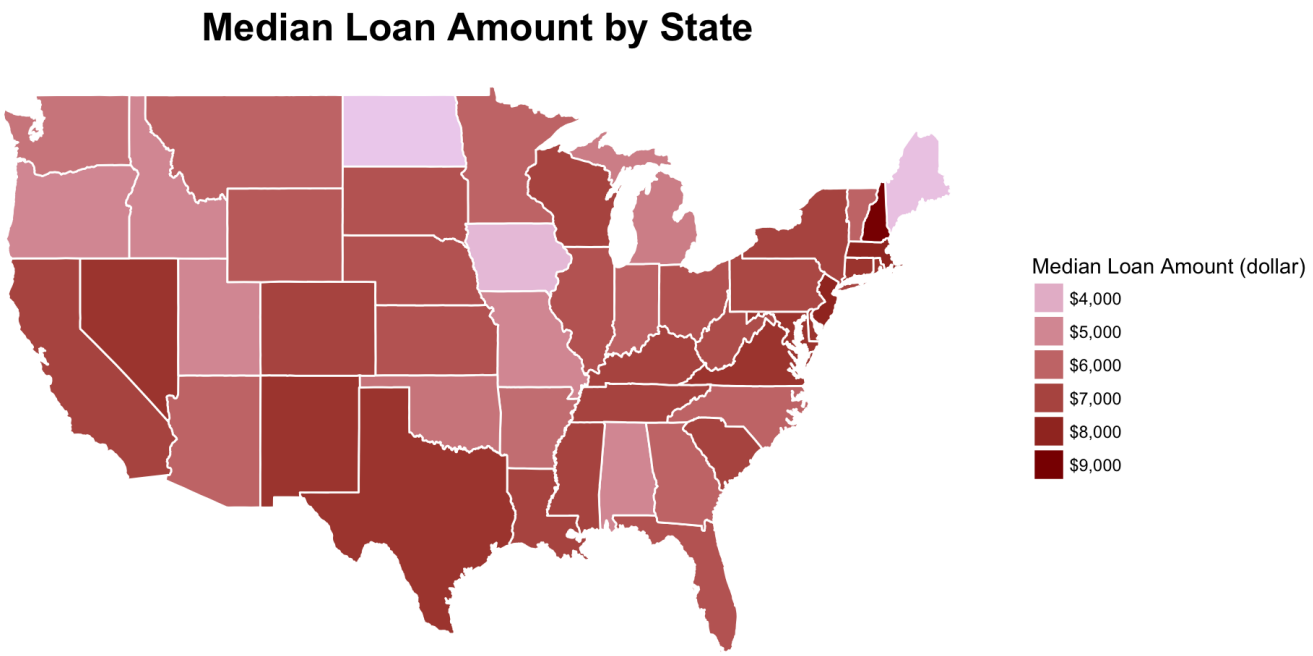
Setting the binwidth at 0.2% for interest rate distribution, it is interesting to see how the number of loans drastically goes up at interest rate around 31.7%. While all the other bins have a count of less than 2000 loans, there are close to 4000 loans with 31.7% rate in the dataset. The violin chart of interest rate based on loan category shows that “Household Expenses” is the category with the highest proportion of loans around 31.7% interest rate (displayed with a dashed line).

Plot Two

```
ggplot()+
  geom_map(data=all_states, map=all_states, aes(x=long, y=lat, map_id=region)) +
  geom_map(data=state_income_loan, map = all_states,
           aes(fill = median_income, map_id = BorrowerState), color="#ffffff") +
  labs(x= NULL, y = NULL) + theme(axis.text = element_blank()) +
  theme(panel.background = element_blank()) +
  theme(axis.ticks = element_blank()) +
  scale_fill_continuous(low = "thistle2", high = "navyblue",
                        labels = dollar,
                        guide = guide_legend(title = "Median Income (dollar)")) +
  ggtitle("Median Income by State")+
  theme(plot.title = element_text(face = "bold", size = 20, hjust = 0.5))
```



```
ggplot()+
  geom_map(data=all_states, map=all_states, aes(x=long, y=lat, map_id=region)) +
  geom_map(data=state_income_loan, map = all_states,
           aes(fill = median_loan, map_id = BorrowerState), color="#ffffff") +
  labs(x= NULL, y = NULL) + theme(axis.text = element_blank()) +
  theme(panel.background = element_blank()) +
  theme(axis.ticks = element_blank()) +
  scale_fill_continuous(low='thistle2', high='darkred',
                        labels = dollar, guide =
                          guide_legend(title = "Median Loan Amount (dollar)")) +
  ggtitle("Median Loan Amount by State")+
  theme(plot.title = element_text(face = "bold", size = 20, hjust = 0.5))
```

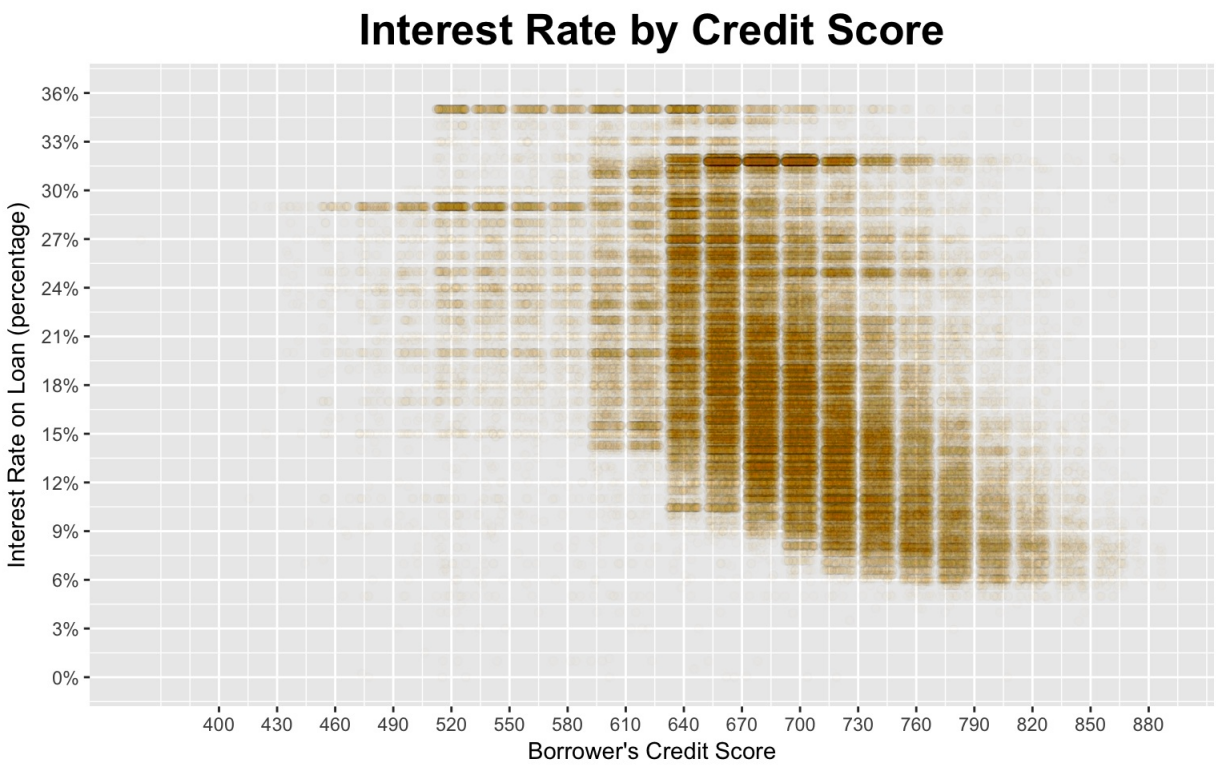


Description Two

The two maps above show how closely the order of median loan amount follows the order of median income stated by borrowers in each state. While DC and states like Maryland and Virginia are at the top of the list of both median income and median loan amount, North Dakota and Iowa are at the bottom, again both on median income and median loan amount.

Plot Three

```
ggplot(aes(x=CreditScoreRangeLower, y= BorrowerRate),
       data = subset(loans, CreditScoreRangeLower > 0))+
  geom_jitter(alpha = 0.01, color = I("black"), fill = I("#F79420"), shape = 21)+
  scale_x_continuous(breaks = seq(400,900, 30)) +
  scale_y_continuous(breaks = seq(0, 0.5, 0.03), labels = percent) +
  labs(x = "Borrower's Credit Score", y = "Interest Rate on Loan (percentage)") +
  ggtitle("Interest Rate by Credit Score")+
  theme(plot.title = element_text(face = "bold", size = 20, hjust = 0.5))
```



Description Three

We see a somewhat exponential relationship between interest rate and credit score in the range of 640-880. The vertical gaps in the data is due to the fact that credit score value is stated in intervals of 20 in the dataset. The direction of the relationship is negative. As credit score increases, interest rate and the level of variation in interest rate decrease.

The horizontal bands around the interest rates of 31% and 35% indicate a significant frequency of interest rates around these values which occur mostly for credit scores lower than 740.

Reflection

In this report, I conducted an analysis on the dataset containing almost 113,000 loans from 2005-2014. Each row in the dataset included 81 features of a loan, from which I picked 13 variables to explore. I looked at the distribution of the individual variables as well as some of the relationships between variables with a focus on interest rate and borrowers' credit score.

One of the struggles here was that most variables were either categorical or discrete numerical variables, so in many cases I was not able to find quantified correlations that are normally applied to continuous data. While I could find some patterns between credit score and variables such as delinquency, annual income, and open credit lines, finding variables with meaningful relationship with interest rate turned out to be a real challenge. The main variable impacting interest rate was credit score with an exponential relationship. It looked like that the relationship between interest rate versus other variables could somehow be explained indirectly through credit score. For example, the fact that lower amounts of loan tend to have higher interest rate could be explained based on the fact that lower amounts of loan tend to be applied for by people with lower income and people with lower income tend to have lower credit score. Or the fact that higher proportions of "Student Use" loans have high interest rates around 35% could be explained by the fact that lower proportions of borrowers who applied for "Student Use" loans had credit score above 700 compared to applicants of other categories of loan.

Since Prosper website indicates that there is about 11%-15% range in interest rate for borrowers within each Prosper credit rating, further investigation is required to find out what determines the exact interest rate within each rating. By finding more variables that have an impact on interest rate, we could build a model to predict the interest rate for each loan.