Exploratory Data Analysis with Prosper Loan Dataset

Vishal Kandagatla

May 9, 2018

```
knitr::opts_chunk$set(echo = TRUE)
```

The Following is an analysis of data made on Prosper Company's Loan Prediction data set. Prosper is America's first marketplace lending platform. Get a personal loan at a low rate. Prosper personal loans require generally good credit; this peer-to-peer lender grades your loan so investors can decide whether to fund it.

Here are Some of the features of the loan offered by Prosper:

1. Low interest rate
2. Fixed term-3 or 5 years*
3. Single monthly payment
4. No hidden fees or prepayment penalties

We first read the dataset with the following command:

```
pld <- read.csv('prosperLoanData.csv', sep = ",")
```

To display the number of observations in the dataset:

```
dim(pld)
```

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

There are 82675 observations and 85 variables.

To display the List of variables:

```
install.packages('ggplot2', repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/Prajval/Documents/R/win-library/3.4'
## (as 'lib' is unspecified)
```

```
## package 'ggplot2' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\Prajval\AppData\Local\Temp\Rtmpw96SUE\downloaded_packages
```

```
library(ggplot2)

install.packages('ggthemes', repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/Prajval/Documents/R/win-library/3.4'
## (as 'lib' is unspecified)
```

```
## package 'ggthemes' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\Prajval\AppData\Local\Temp\Rtmpw96SUE\downloaded_packages
```

```
library(ggthemes)

names(pld)
```

```
##  [1] "ListingKey"
##  [2] "ListingNumber"
##  [3] "ListingCreationDate"
##  [4] "CreditGrade"
##  [5] "Term"
##  [6] "LoanStatus"
##  [7] "ClosedDate"
##  [8] "BorrowerAPR"
##  [9] "BorrowerRate"
## [10] "LenderYield"
## [11] "EstimatedEffectiveYield"
## [12] "EstimatedLoss"
## [13] "EstimatedReturn"
## [14] "ProsperRating..numeric."
## [15] "ProsperRating..Alpha."
## [16] "ProsperScore"
## [17] "ListingCategory..numeric."
## [18] "BorrowerState"
## [19] "Occupation"
## [20] "EmploymentStatus"
## [21] "EmploymentStatusDuration"
## [22] "IsBorrowerHomeowner"
## [23] "CurrentlyInGroup"
## [24] "GroupKey"
## [25] "DateCreditPulled"
## [26] "CreditScoreRangeLower"
## [27] "CreditScoreRangeUpper"
## [28] "FirstRecordedCreditLine"
## [29] "CurrentCreditLines"
## [30] "OpenCreditLines"
## [31] "TotalCreditLinespast7years"
## [32] "OpenRevolvingAccounts"
## [33] "OpenRevolvingMonthlyPayment"
## [34] "InquiriesLast6Months"
## [35] "TotalInquiries"
## [36] "CurrentDelinquencies"
## [37] "AmountDelinquent"
## [38] "DelinquenciesLast7Years"
## [39] "PublicRecordsLast10Years"
## [40] "PublicRecordsLast12Months"
## [41] "RevolvingCreditBalance"
## [42] "BankcardUtilization"
```

```
## [43] "AvailableBankcardCredit"
## [44] "TotalTrades"
## [45] "TradesNeverDelinquent..percentage."
## [46] "TradesOpenedLast6Months"
## [47] "DebtToIncomeRatio"
## [48] "IncomeRange"
## [49] "IncomeVerifiable"
## [50] "StatedMonthlyIncome"
## [51] "LoanKey"
## [52] "TotalProsperLoans"
## [53] "TotalProsperPaymentsBilled"
## [54] "OnTimeProsperPayments"
## [55] "ProsperPaymentsLessThanOneMonthLate"
## [56] "ProsperPaymentsOneMonthPlusLate"
## [57] "ProsperPrincipalBorrowed"
## [58] "ProsperPrincipalOutstanding"
## [59] "ScorexChangeAtTimeOfListing"
## [60] "LoanCurrentDaysDelinquent"
## [61] "LoanFirstDefaultedCycleNumber"
## [62] "LoanMonthsSinceOrigination"
## [63] "LoanNumber"
## [64] "LoanOriginalAmount"
## [65] "LoanOriginationDate"
## [66] "LoanOriginationQuarter"
## [67] "MemberKey"
## [68] "MonthlyLoanPayment"
## [69] "LP_CustomerPayments"
## [70] "LP_CustomerPrincipalPayments"
## [71] "LP_InterestandFees"
## [72] "LP_ServiceFees"
## [73] "LP_CollectionFees"
## [74] "LP_GrossPrincipalLoss"
## [75] "LP_NetPrincipalLoss"
## [76] "LP_NonPrincipalRecoverypayments"
## [77] "PercentFunded"
## [78] "Recommendations"
## [79] "InvestmentFromFriendsCount"
## [80] "InvestmentFromFriendsAmount"
## [81] "Investors"
```

The 85 variables are listed above.

Here is the summary of data obtained from the Prosper official websit:

After a borrower submits a loan application, Prosper first obtains a credit report from TransUnion to evaluate whether the applicant meets the underwriting criteria Prosper has established in conjunction with WebBank. As of December 23, 2016, applicants are subject to the following minimum eligibility criteria:

- Minimum FICO credit score of 640 (TransUnion FICO 08 score)
- Debt-to-income ratio below 50%
- Stated income greater than $0
- No bankruptcies filed within the last 12 months
- Fewer than seven credit bureau inquiries within the last 6 months
- Minimum of three open trades reported on their credit report

Repeat borrowers must meet several additional criteria:

- No previous loans on the Prosper platform which have been charged-off
- Must not have been declined for a loan through Prosper within the last four months due to
  - - delinquency or returned payments on a previous loan through Prosper

Summary of Loan Amounts

```
summary(pld$LoanOriginalAmount)
```

```
##    Min. 1st    Qu    Media     Mean 3rd Qu.     Max.
##    1000        .           n
##                    6500     8337 12000       35000
##                400
##                0
```

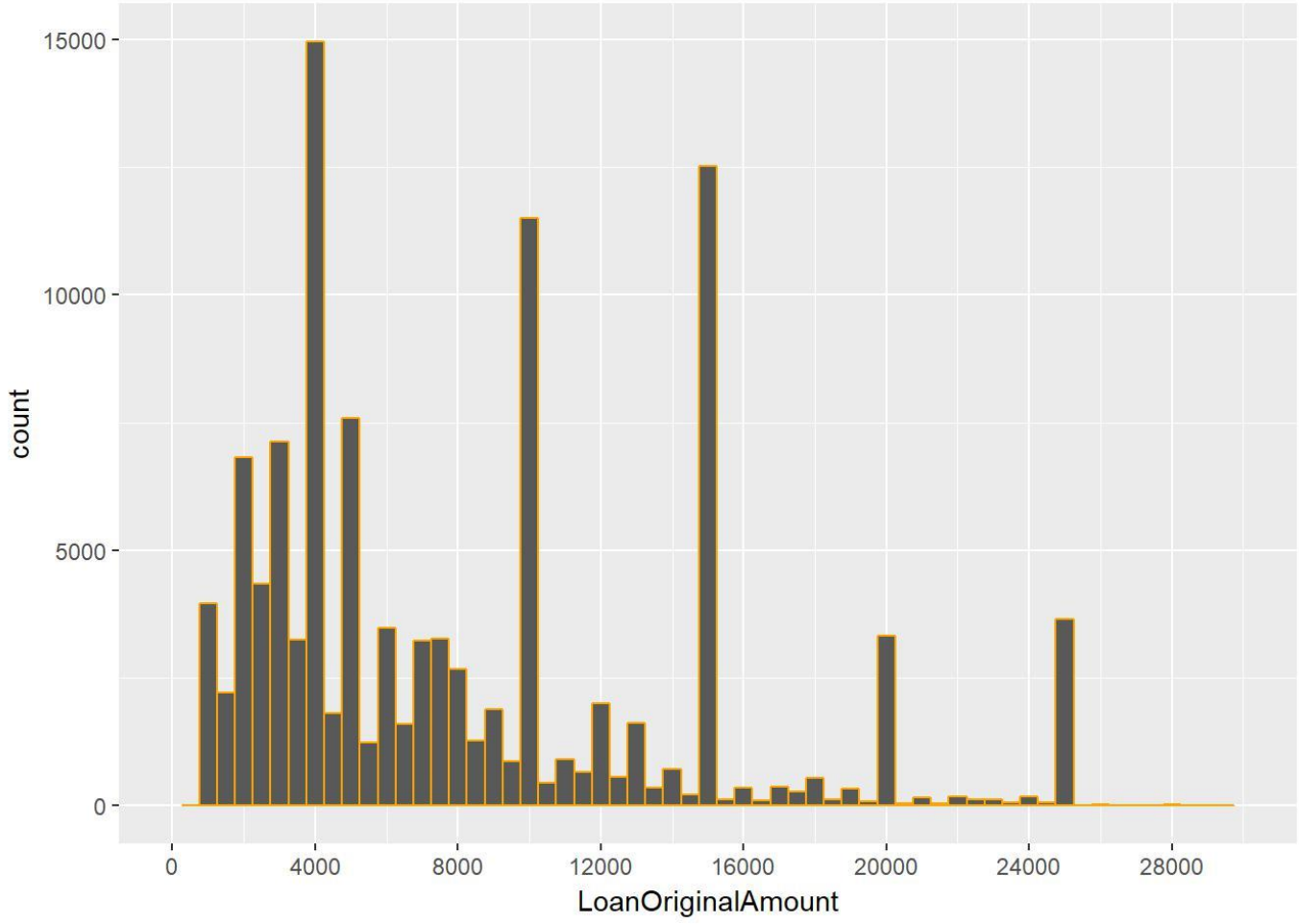
Min Loan amount is $1000 Max Loan amount is $35000

Prosper approves loans in the range $1000 to $35000.

Plotting Loan amount

```
ggplot(data = pld,aes(LoanOriginalAmount))+
  geom_histogram(binwidth=500,color=I('orange')) +
  scale_x_continuous(breaks=seq(0,35000,4000),lim=c(0,30000))
```

```
## Warning: Removed 503 rows containing non-finite values (stat_bin).
```



We see spikes at $4000 and $15000 mark. This is because of the fact that prosper approves loans below $4000 without any collateral. They need more documentation and proof of repayment for loans over $15,000 and hence less number of loans are approved.

Summary of monthly payments made by the borrowers:

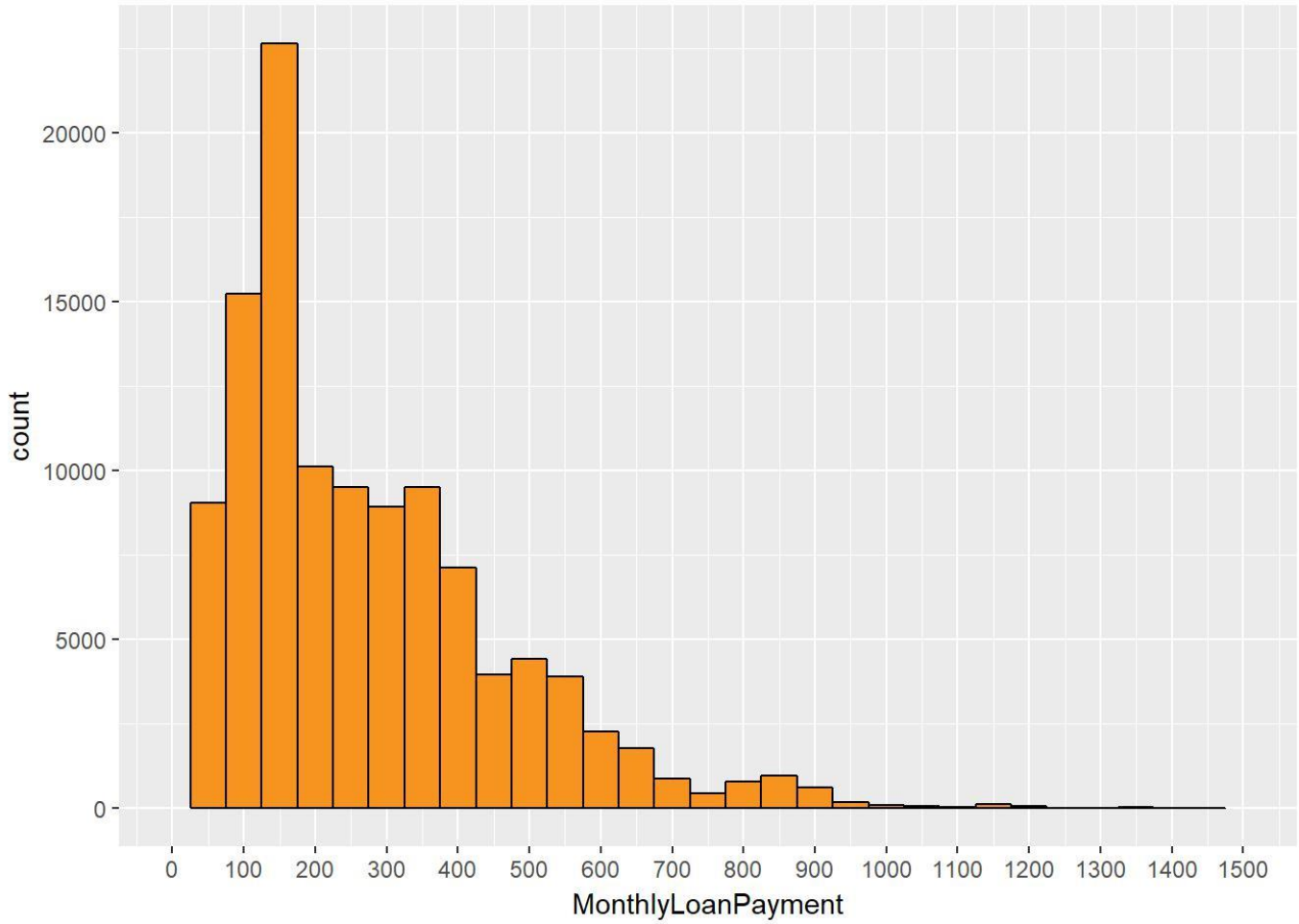
```
summary(pld$MonthlyLoanPayment)
```

```
## Min.    1st Qu.   Media    Mean 3rd Qu.    Max.
## 0.0       131.6       n   272.5 371.6     2251.5
                      217.7
```

Here is a plot of the monthly payments made by the borrowers:

```
ggplot(data=pld, aes(MonthlyLoanPayment))+
  geom_histogram(binwidth=50,color=I('black'),fill=I('#F79420'))+
  scale_x_continuous(limits=c(0,1500),breaks=seq(0,1500,100))
```

```
## Warning: Removed 22 rows containing non-finite values (stat_bin).
```



Most Loan Payments are made in the range $100 to

$200 Borrower Rate

```
summary(pld$BorrowerRate)
```

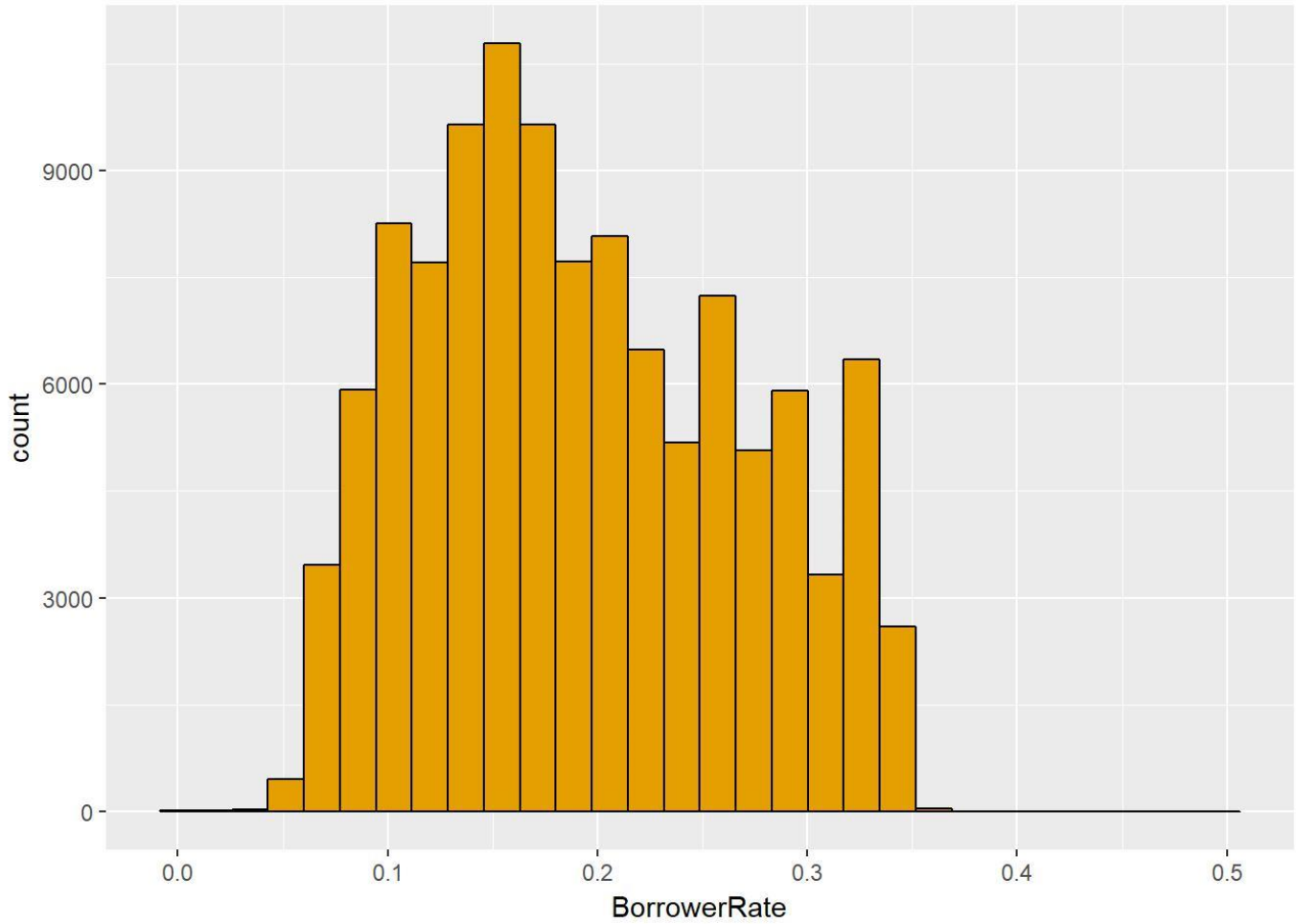
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.1340  0.1840  0.1928  0.2500  0.4975
```

Median borrower rate is 18.75%.

Here is a plot of the borrower rate:

```
ggplot(data=pld,aes(BorrowerRate)) +
  geom_histogram(color=I('black'),fill=I('#E69F00'))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Borrower rates typically lie between 0.1 and 0.3 That is 10% to 30%

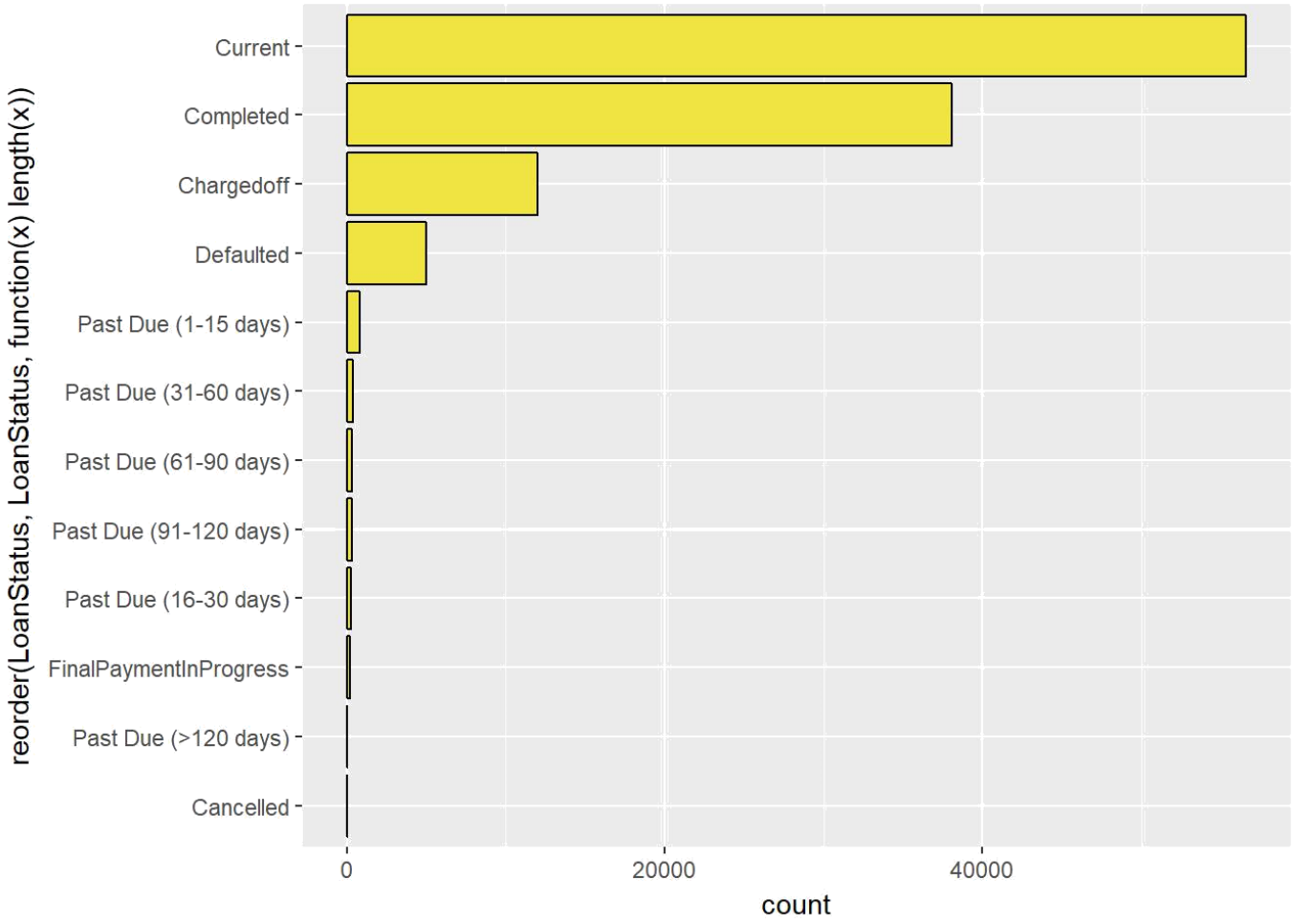
Summary of the Loan Status of borrowers:

```
summary(pld$LoanStatus)
```

```
##               Cancelled             Chargedoff              Completed 
##                       5                  11992                  38074 
##                 Current             Defaulted FinalPaymentInProgress 
##                   56576                   5018                    205 
##   Past Due (>120 days)   Past Due (1-15 days) Past Due (16-30 days) 
##                      16                    806                    265 
##  Past Due (31-60 days)  Past Due (61-90 days) Past Due (91-120 days) 
##                    363                    313                    304
```

We can plot this in the form a graph. The graph contains one variable:

```
ggplot(data=pld,aes(reorder(LoanStatus, LoanStatus, function(x) length(x)))) +
  geom_bar( color=I('black'),fill=I('#F0E442')) +
    coord_flip()
```



Number of Cutomers who have been defaulted:

```
nrow(pld[pld$LoanStatus=='Defaulted',])/nrow(pld)*100
```

```
## [1] 4.404188
```

1.11% of the Loans are Defaulted

Number of people who make late payments

```
nrow(subset(pld,
            LoanStatus=='Past Due (1-15 days)'  |
            LoanStatus=='Past Due (16-30 days)' |
            LoanStatus=='Past Due (31-60 days)' |
            LoanStatus=='Past Due (61-90 days)' |
            LoanStatus=='Past Due (91-120 days)'|
            LoanStatus=='Past Due (>120 days)'))/nrow(pld)*100
```

```
## [1] 1.81416
```

2.50% of people make late payments

Prosper scores of the borrowers:

```
summary(pld$ProsperScore)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00 4.00 6.00 5.95 8.00 11.00 29084
```

```
table(pld$ProsperScore)
```

```
 ##
 ##     1     2     3     4     5     6     7     8     9    10    11
##   992  5766  7642 12595  9813 12278 10597 12053  6911  4750  1456
```

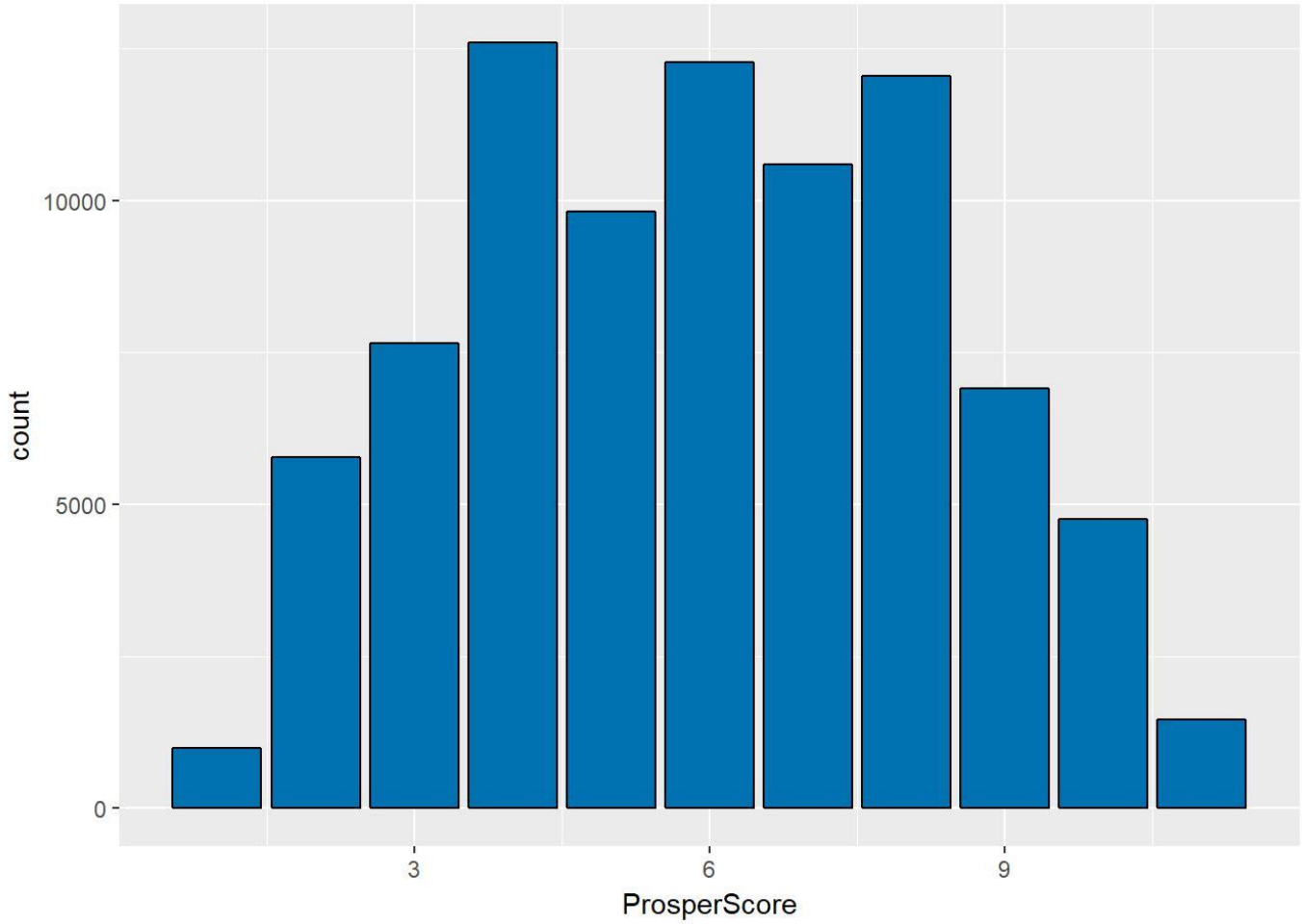
Scores lie in the range 1 - 11. Median Score is 6.

However, Prosper has a set a minimum score of 6.4 for Loan approval.

A plot of the customers Prosper Scores:

```
ggplot(data=pld,aes(ProsperScore))+
  geom_bar( color=I('black'),fill=I('#0072B2'))
```

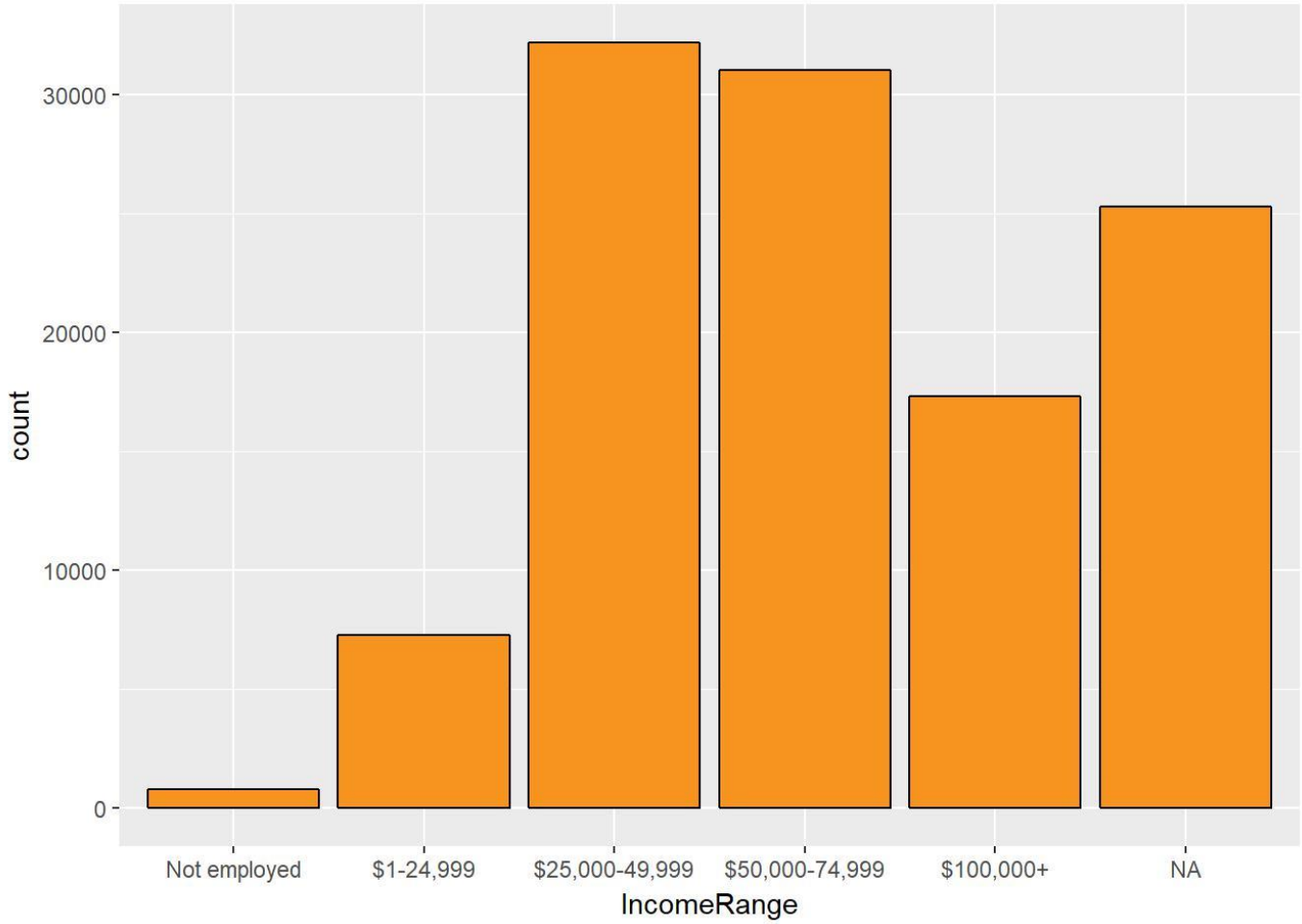
```
## Warning: Removed 29084 rows containing non-finite values (stat_count).
```



Here is a distribution of the borrower's income range:

```
pld$IncomeRange<-ordered(pld$IncomeRange,levels=c("Not employed",
                                                  "$1-24,999",
                                                  "$25,000-49,999",
                                                  "$50,000-74,999",
                                                  "75,000-99,999",
                                                  "$100,000+"))

ggplot(data=pld,aes(IncomeRange)) +
  geom_bar(color=I('black'),fill=I("#F79420"))
```



Most of the people who borrow a loan from Prosper have an income in the range $25000 - $75000. People who apply for loans having a income range over $100,000 are low in number.

Summary of the Debt to Income Ratio:

```
summary(pld$DebtToIncomeRatio)
```

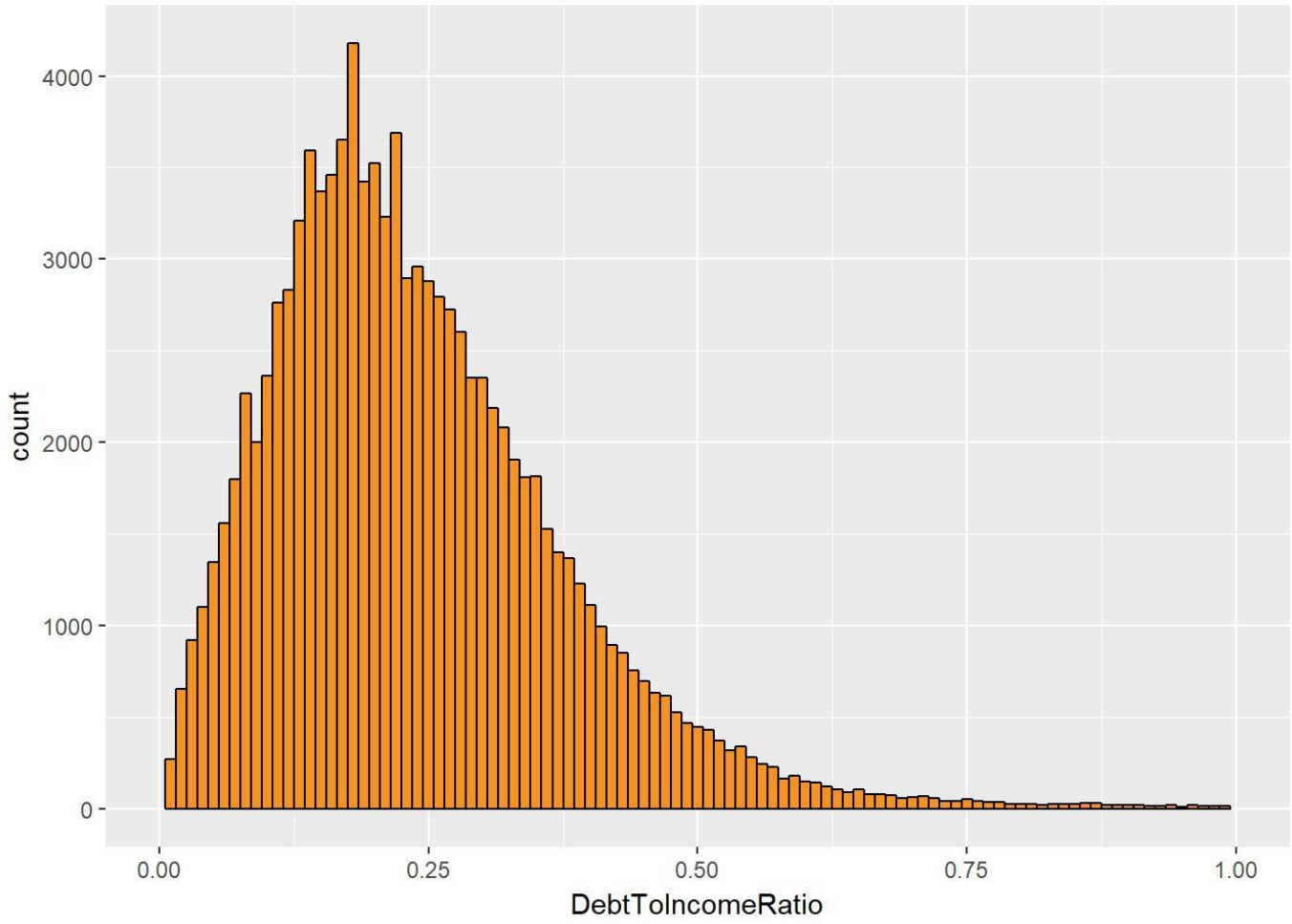
```
##  Min. 1st Qu.      Media      Mean 3rd Qu.      Max      NA's
##  0.000 0.140            n    0.276 0.320          .      8554
                       0.220                     10.01
                                                 0
```

Mean Debt to income ratio is 26%. That is borrowers, on an average spend 26% of their money and save the rest.

Debt to Income Ratio Plot

```
ggplot(data=pld,aes(DebtToIncomeRatio)) +
  geom_histogram(binwidth=0.01,color=I('black'),fill=I('#F79420')) +
  xlim(0,1)
```

```
## Warning: Removed 9353 rows containing non-finite values (stat_bin).
```



Most of the borrowers debt to income ratio lie in the range 0.15 to 0.30

Prosper grants loans to those who have a good debt to income ratio.

According to Prosper: - Maximum debt-to-income ratio: 50% (excluding mortgage) is set for the borrower. Credit Score Range Lower Summary

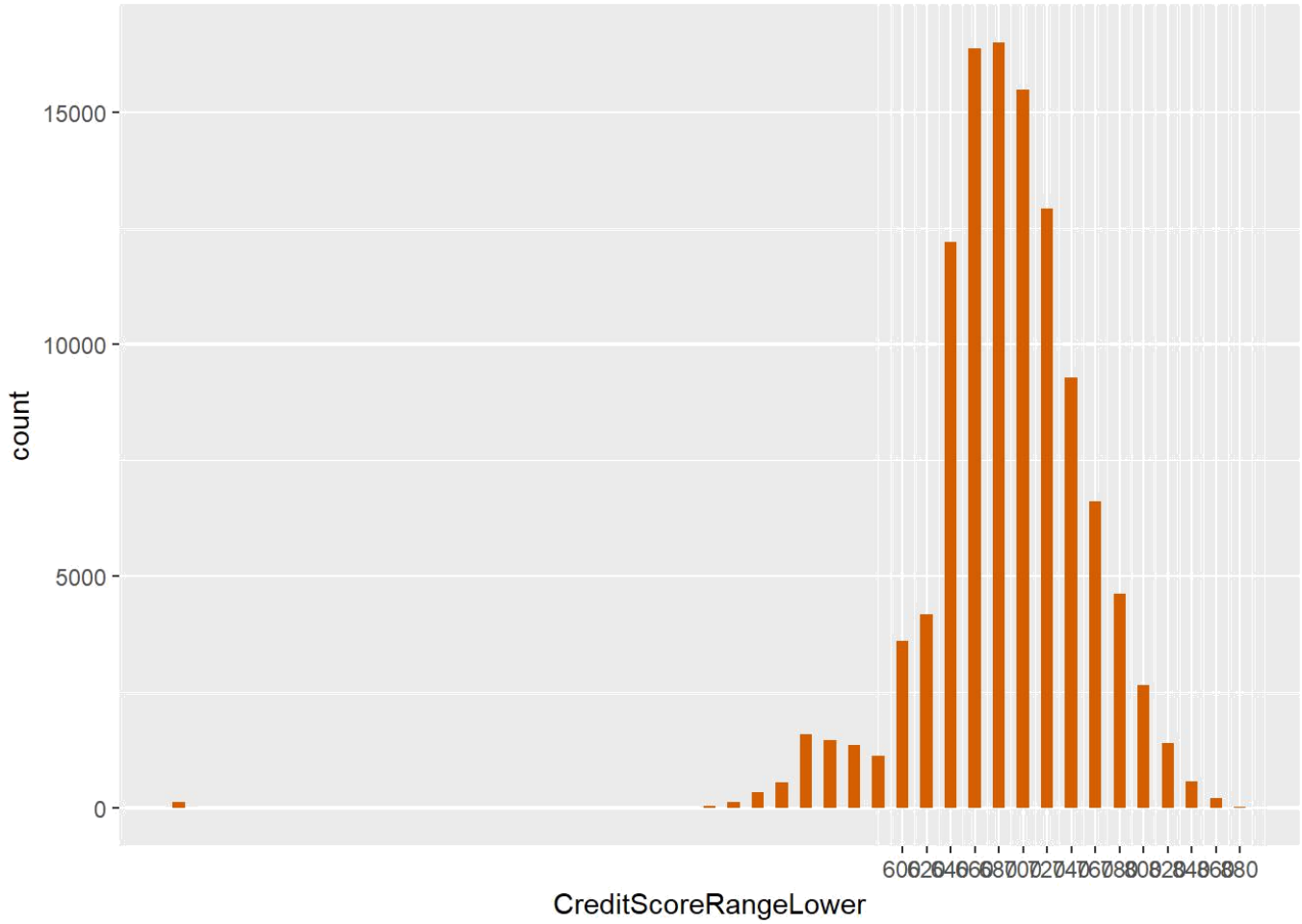
```
summary(pld$CreditScoreRangeLower)
```

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

Here is a plot of the Credit scores of the borrowers:

```
ggplot(data=pld,aes(CreditScoreRangeLower)) +
  geom_histogram(binwidth=10,fill=I("#D55E00") ) +
  scale_x_continuous(breaks=seq(600,880,20))
```

```
## Warning: Removed 591 rows containing non-finite values (stat_bin).
```



Credit score for most people lie in the range 660-720. Prosper generally grants loans for people having good credit score. They have set a minimum credit score requirement of 640 if the loan has to be approved. Else the borrower has to present additional documents to be able to get his loan sanctioned.

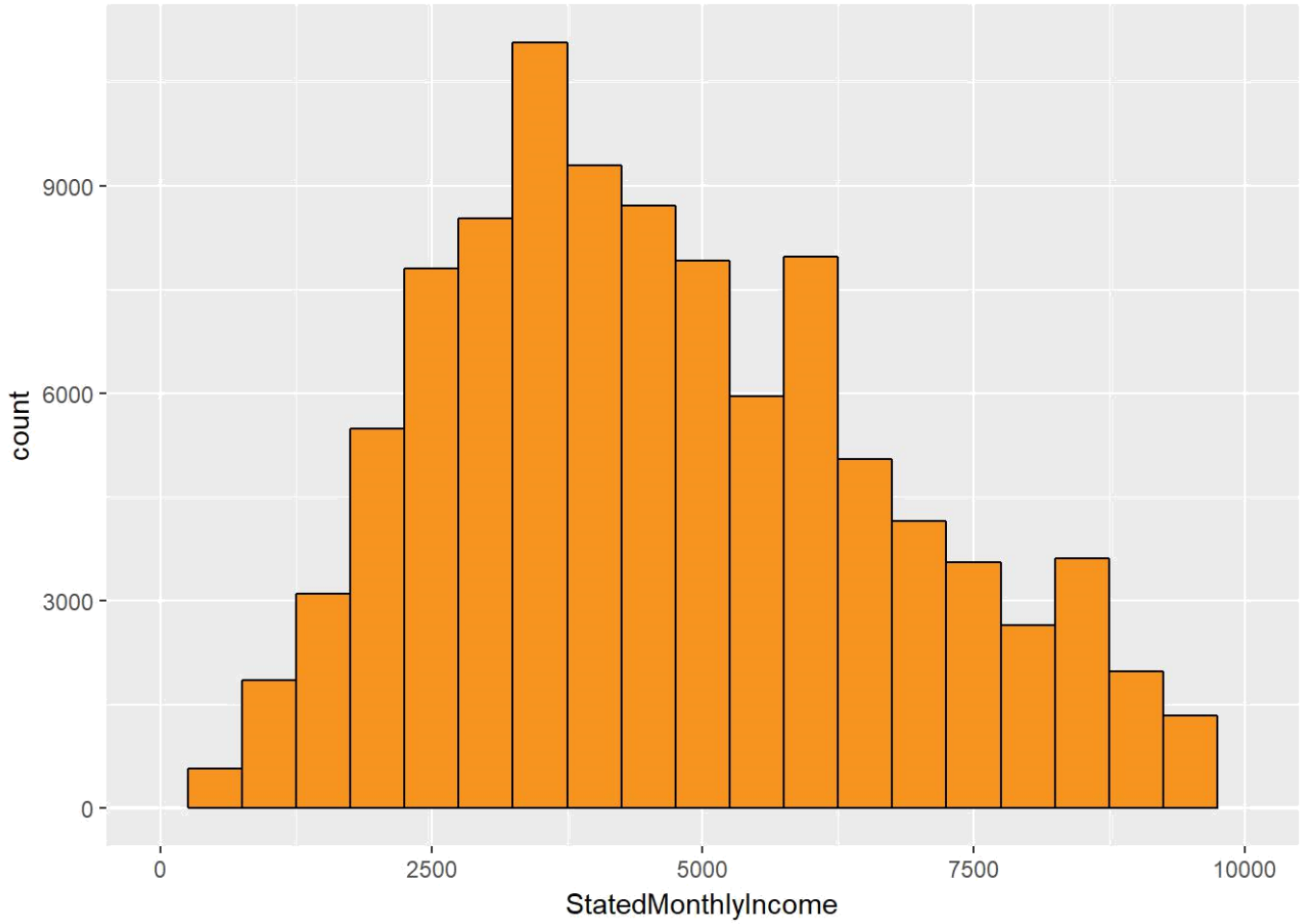
Here is the monthly income of the Prosper customers:

```
summary(pld$StatedMonthlyIncome)
```

```
## Min. 1st Qu.      Media     Mean 3rd Qu.    Max.
##                       n
## 0 3200             4667     5608    6825 1750003
```

```
  ggplot(data=pld,aes(StatedMonthlyIncome)) +
  geom_histogram(binwidth=500, color=I('black'), fill=I("#F79420") ) +
  xlim(0,10000)
```

```
## Warning: Removed 9780 rows containing non-finite values (stat_bin).
```



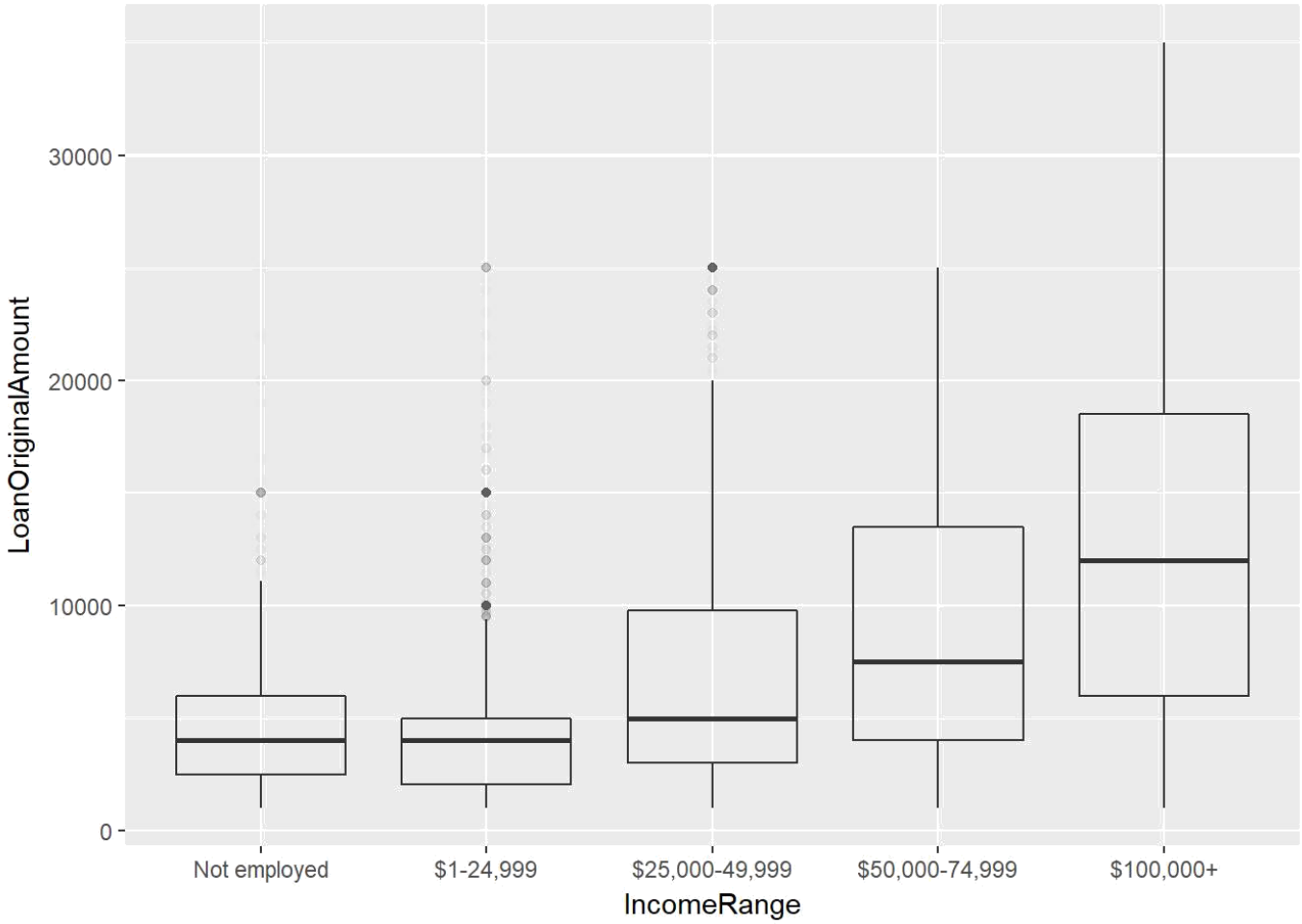
Monthly incomes typically lie in the range $2500 - $7000.

Bivariate Analysis:

This is a plot of income range vs Loan Amount of the borrowers.

```
ggplot(aes(x = IncomeRange, y = LoanOriginalAmount), data=subset(pld, !is.na(Incom
eRange))) +
  geom_boxplot(alpha=1/100) +
   geom_line(stat= 'summary', fun.y = mean, linetype = 2, color = 'blue')
```

```
## geom_path: Each group consists of only one observation. Do you need to
## adjust the group aesthetic?
```



We can make the following observations:

1. Loans above $20,000 are granted for people with income range more than $100,000 in most of the situations.
2. Prosper generally grants loans for people having income range more than $25000.
3. Loans in the range of $30,000 are only granted for people having a yearly income of $100,000 or more.

An analysis of the Debt to income ratio and Credit Score range:

We first make 2 columns that contain the mean and median values of the credit scores.

Here is a sample of our credit score data:

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
pld.fc_new <- pld %>%
  group_by(DebtToIncomeRatio) %>%
  summarise(CreditScoreRangeLower_mean = mean(CreditScoreRangeLower),
            CreditScoreRangeLower_median =
            median(CreditScoreRangeLower), n=n()) %>%
  arrange(DebtToIncomeRatio)

head(pld.fc_new)
```

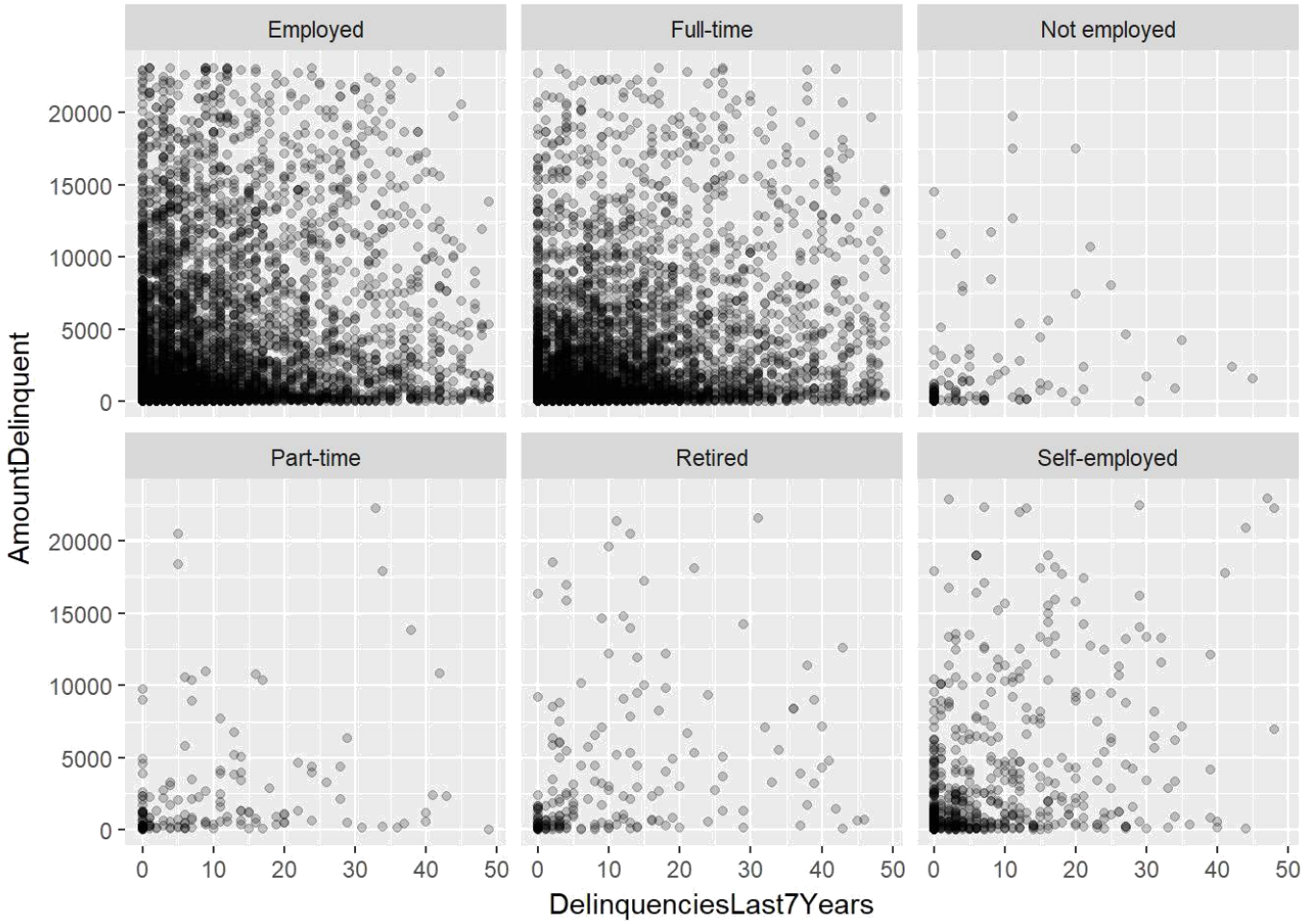
```
## # A tibble: 6 x 4
##   DebtToIncomeRatio CreditScoreRangeLower_mean
##               <dbl>                      <dbl>
## 1           0.00000                   666.3158
## 2           0.00044                         NA
## 3           0.00310                   520.0000
## 4           0.00611                         NA
## 5           0.00647                         NA
## 6           0.00677                         NA
## # ... with 2 more variables: CreditScoreRangeLower_median <dbl>, n <int>
```

A plot of Debt to income ratio vs Credit Score Range.

We make use of Facet wrap o depict the deliquencies of the customers based on their employment status.
The final plot we get is as shown below:

```
ggplot(aes(x = DelinquenciesLast7Years, y = AmountDelinquent),
       data = filter(pld, AmountDelinquent > 0 &
                           EmploymentStatus != "Other" )) +
  geom_point(alpha = 0.2) +
  xlim(0, quantile(pld$DelinquenciesLast7Years, 0.99, na.rm = T)) +
  ylim(0, quantile(pld$AmountDelinquent, 0.99, na.rm = T)) +
  facet_wrap(~EmploymentStatus)
```

```
## Warning: Removed 1373 rows containing missing values (geom_point).
```

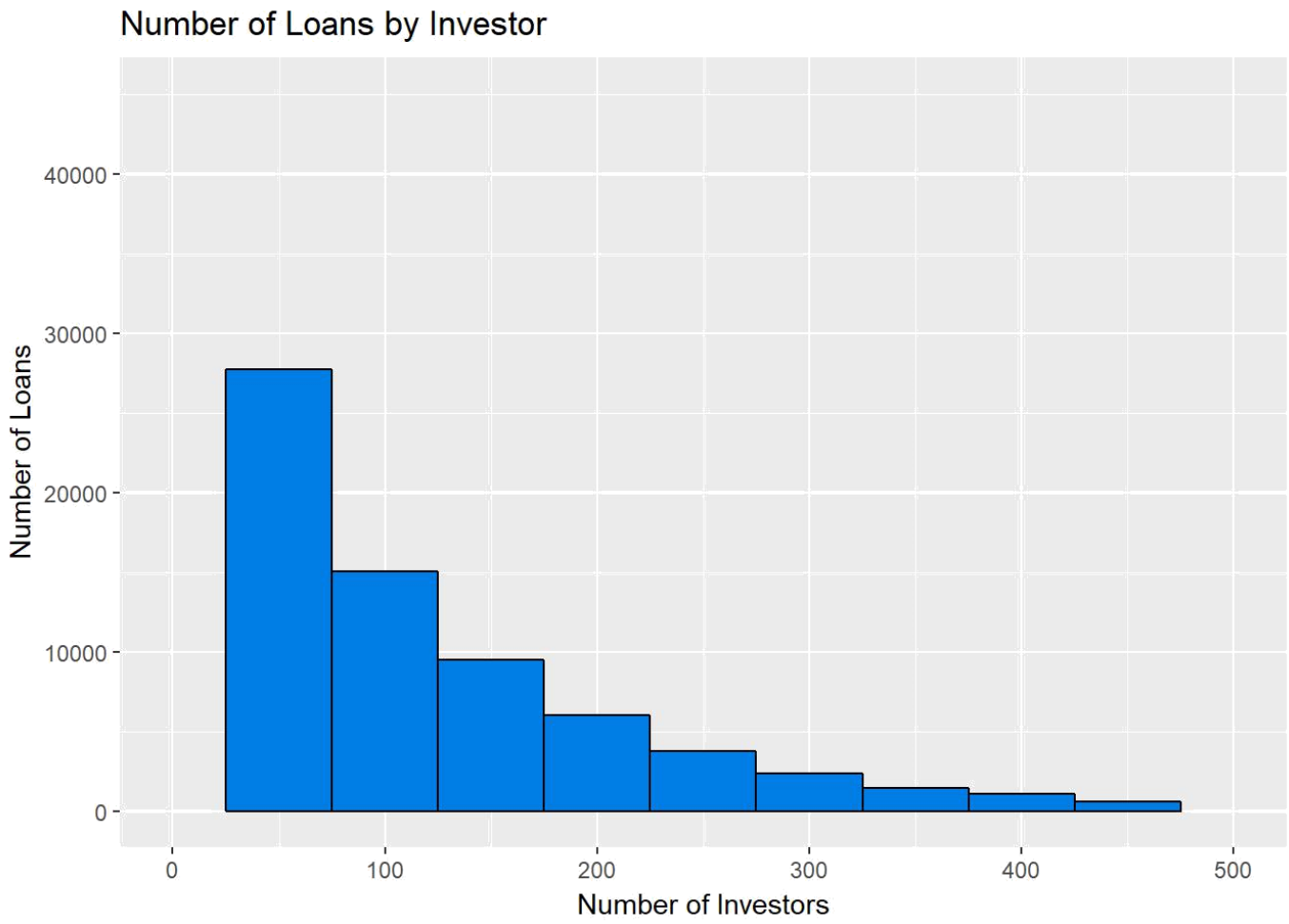


We can see more dots on the employed customers box. This is because Prosper offers loan to cadidates who are employed. People who are either un-employed or having part time jobs have low deliquencies and are thus rejected for loans most of the times.

People may invest in groups for loans. This is a rare situation in most of the banks. Let us have a look at the number of investors for loans through this graph:

```
ggplot(pld, aes(Investors)) +
  geom_histogram(color = 'black', fill = '#007EE5', binwidth = 50) +
  ggtitle('Number of Loans by Investor') + xlab('Number of
  Investors') +
  ylab('Number of Loans') +
  xlim(0,500)
```

```
## Warning: Removed 812 rows containing non-finite values (stat_bin).
```



Number of investors are below 150 for most loans.

Corelation:-

Let us have a look at the correlation between credit score given by the central authority and the prosper score given by the company:

```
cor.test(pld$CreditScoreRangeLower,pld$ProsperScore)
```

```
##
##  Pearson's product-moment correlation
##
## data:  pld$CreditScoreRangeLower and pld$ProsperScore
## t = 115.87, df = 84851, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3637793 0.3753979
## sample estimates:
##      cor
## 0.369603
```

The correlation between the two is 0.37 By this we can come to a conclusion that there is not much of a relation between the credit score and the prosper score. prosper score thus depends on different factors and is computed differently when compared to the credit score.

Correlation between Lender yield and Debt to income ratio:

```
cor.test(pld$LenderYield,pld$DebtToIncomeRatio)
```

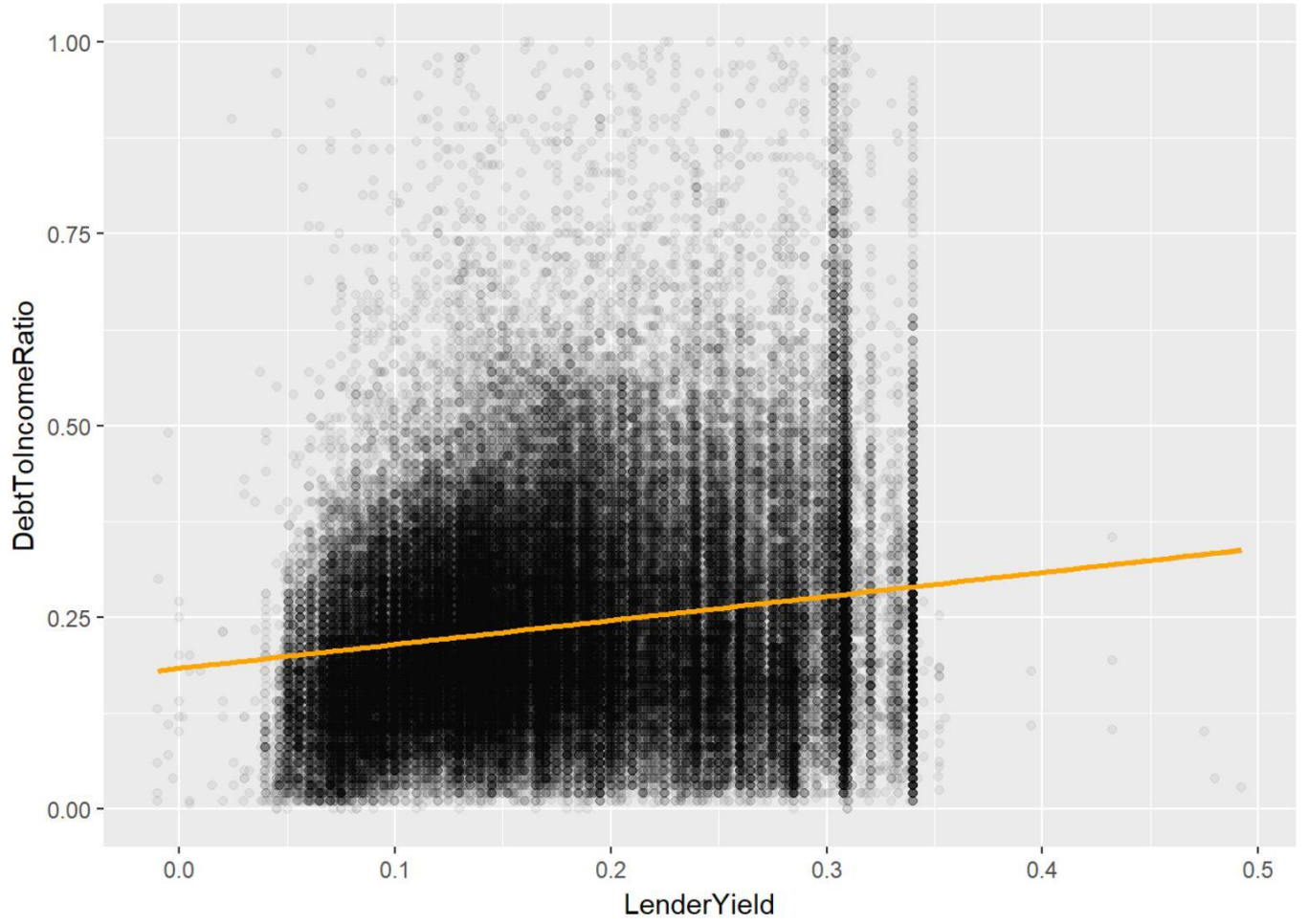
```
##
##  Pearson's product-moment correlation
##
## data:  pld$LenderYield and pld$DebtToIncomeRatio
## t = 20.147, df = 105380, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.05592580 0.06795465
## sample estimates:
##       cor
## 0.06194247
```

The correlation is just 0.12 Lender yeild is thus not quite related to debt to income ratio and it is evident from the following graph.

```
ggplot(data=pld,aes(x=LenderYield, y=DebtToIncomeRatio))+
  ylim(0, 1)+
  geom_point(alpha=0.05)+
  geom_smooth(method = 'lm', color= 'orange')
```

```
## Warning: Removed 9353 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 9353 rows containing missing values (geom_point).
```

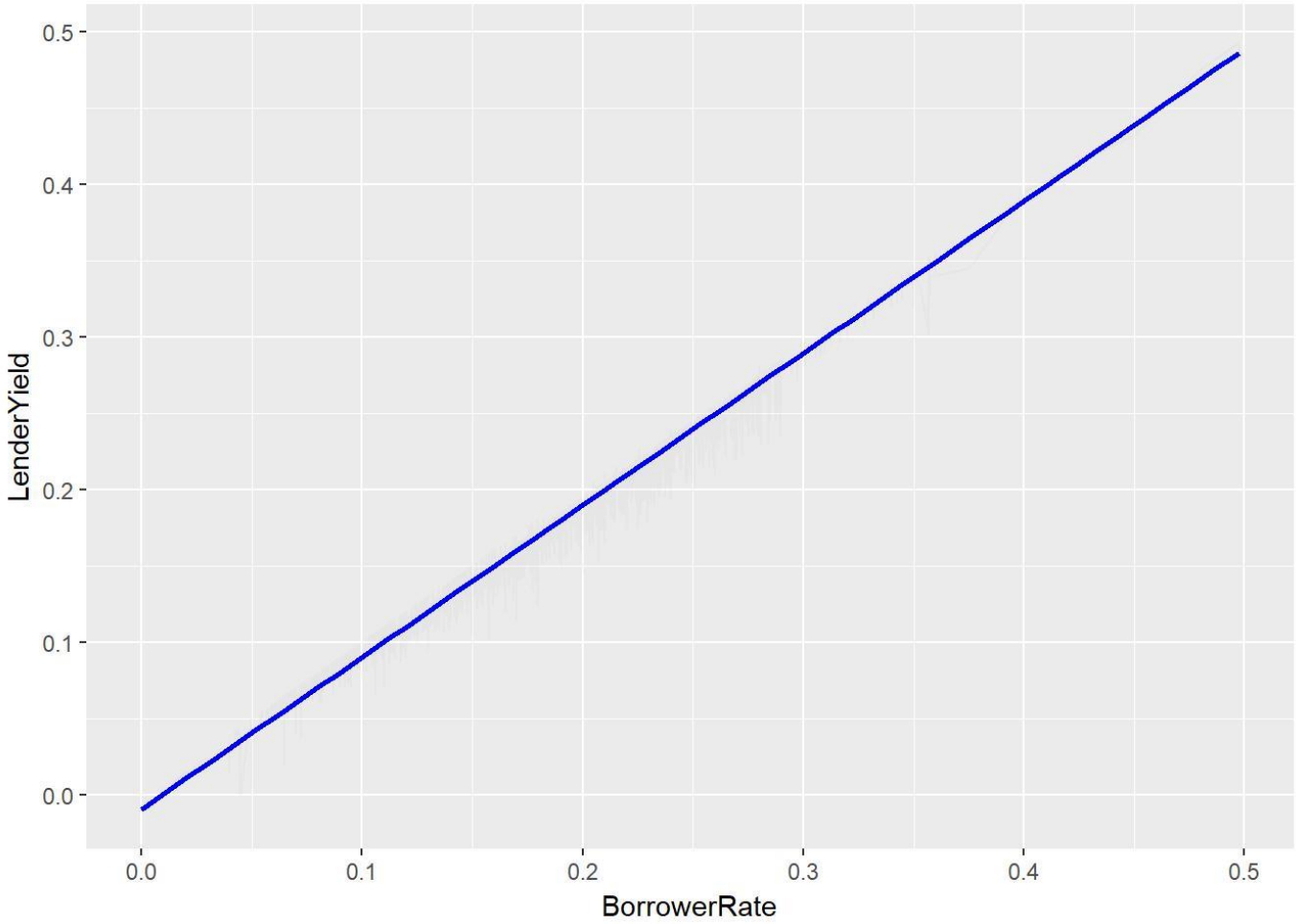



```
cor(pld$BorrowerRate,pld$LenderYield)
```

```
## [1] 0.9992113
```

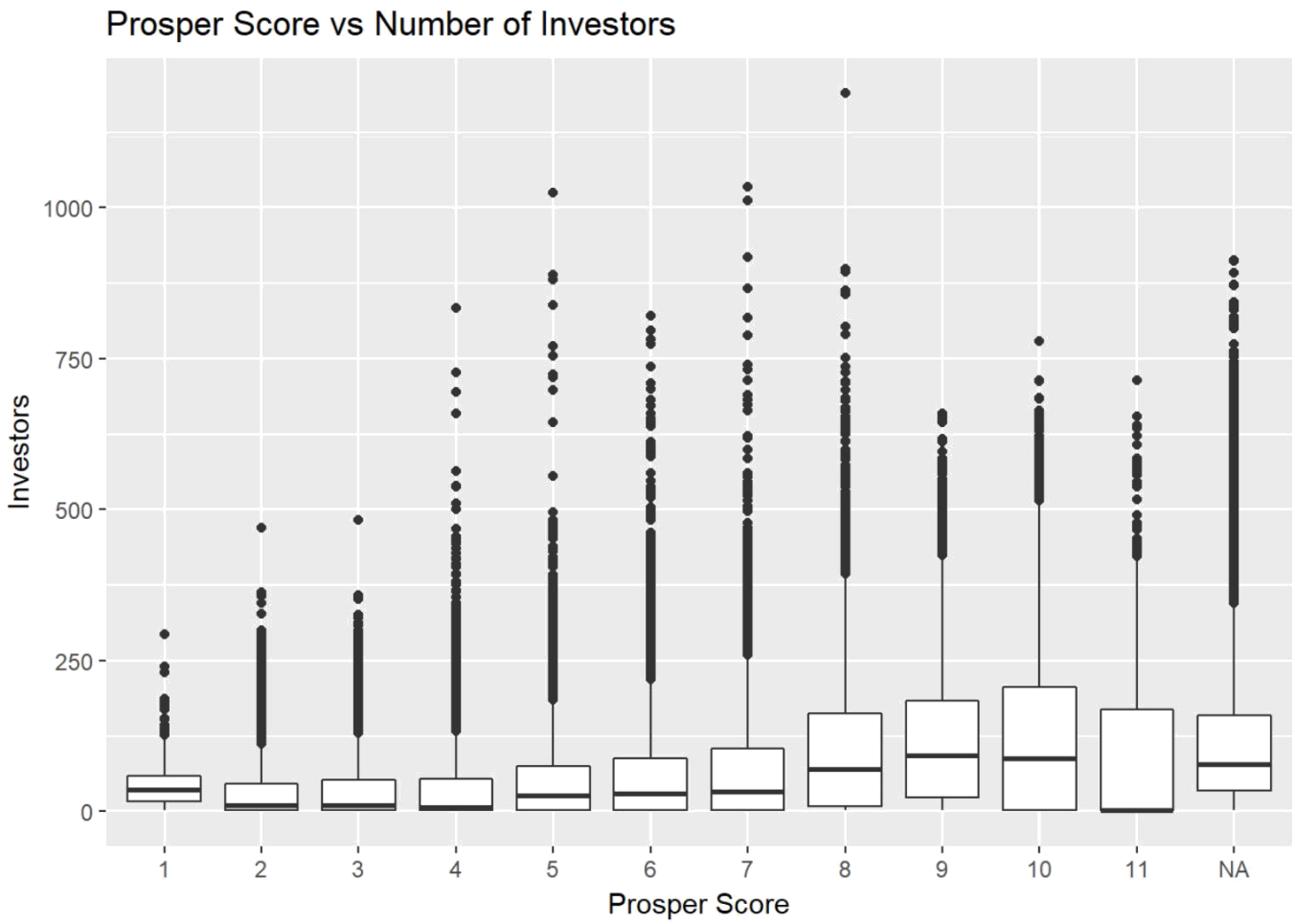
From this it is evident that the borrower rate is almost directly proportional to the Lender Yield. The graph depicts the same:

```
ggplot(data=pld,aes(x=BorrowerRate, y=LenderYield))+
  geom_line(alpha=1/50)+
  geom_smooth(method = 'lm', color= 'blue')
```



Prosper score vs Number of investors:

```
ggplot(aes(x = as.factor(ProsperScore), y = Investors),
       data = pld) +
  geom_boxplot() +
  ggtitle("Prosper Score vs Number of Investors")
  + xlab("Prosper Score")
```




```
  ylab("Investors")
```

```
 ## $y
 ## [1] "Investors"
##
 ## attr(,"class")
 ## [1] "labels"
```

The box plots also depict the median number of investors for the borrowers with certain prosper scores.

We can see from the plot that people having higher prosper scores have more number of investors for the loan.

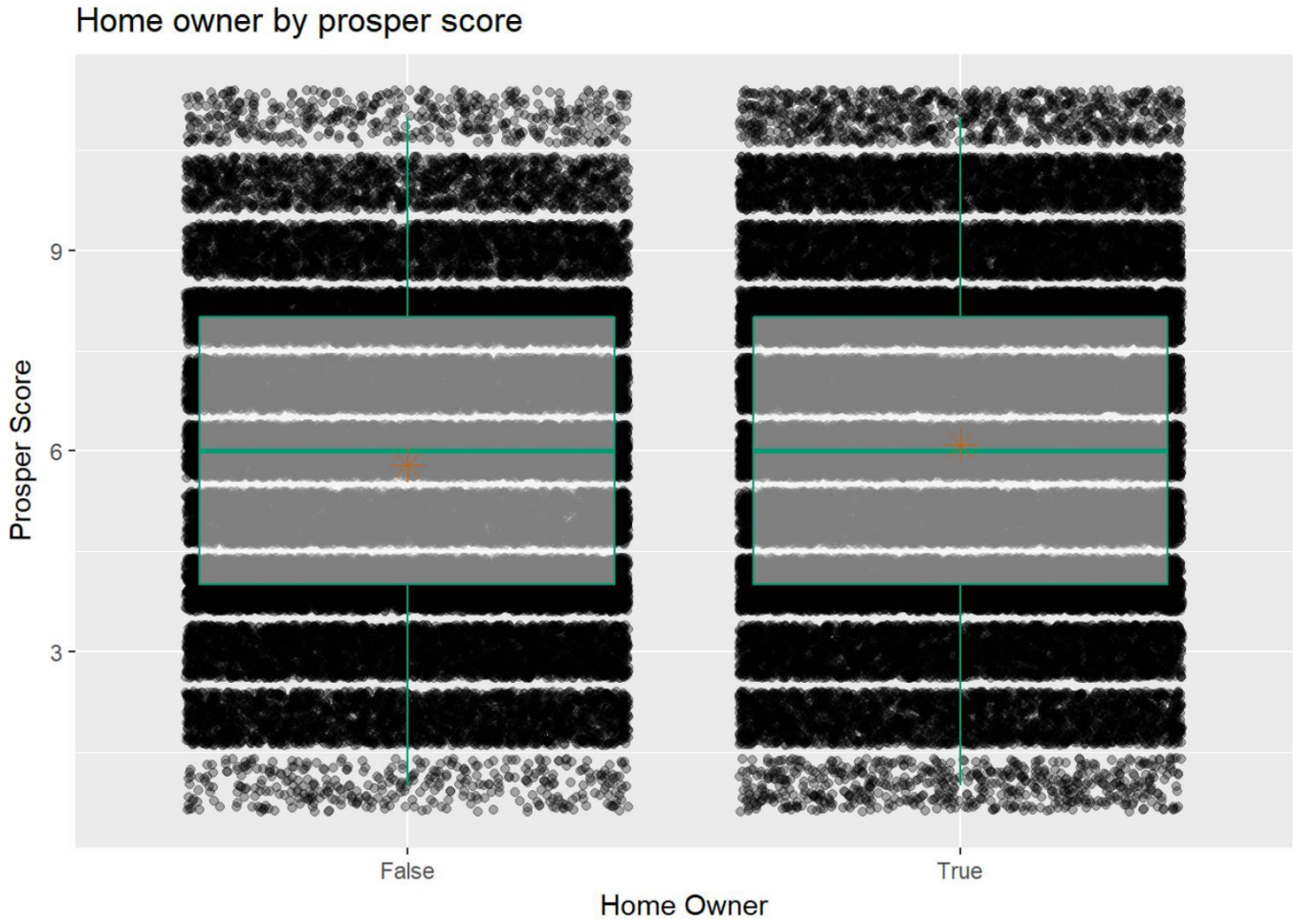
A Comparison of the prosper scores of the borrowers with and without own homes:

```
ggplot(aes(factor(IsBorrowerHomeowner),
           ProsperScore),
       data = pld) +
  geom_jitter( alpha = .3)  +
  geom_boxplot( alpha = .5,color = '#009E73')+
  stat_summary(fun.y = 'mean',
               geom = 'point',
               color = '#D55E00',
               shape = 8,
               size = 4)+
  ylab('Prosper Score') +
  xlab('Home Owner') +
  ggtitle('Home owner by prosper score')
```

```
## Warning: Removed 29084 rows containing non-finite values (stat_boxplot).
```

```
## Warning: Removed 29084 rows containing non-finite values (stat_summary).
```

```
## Warning: Removed 29084 rows containing missing values (geom_point).
```



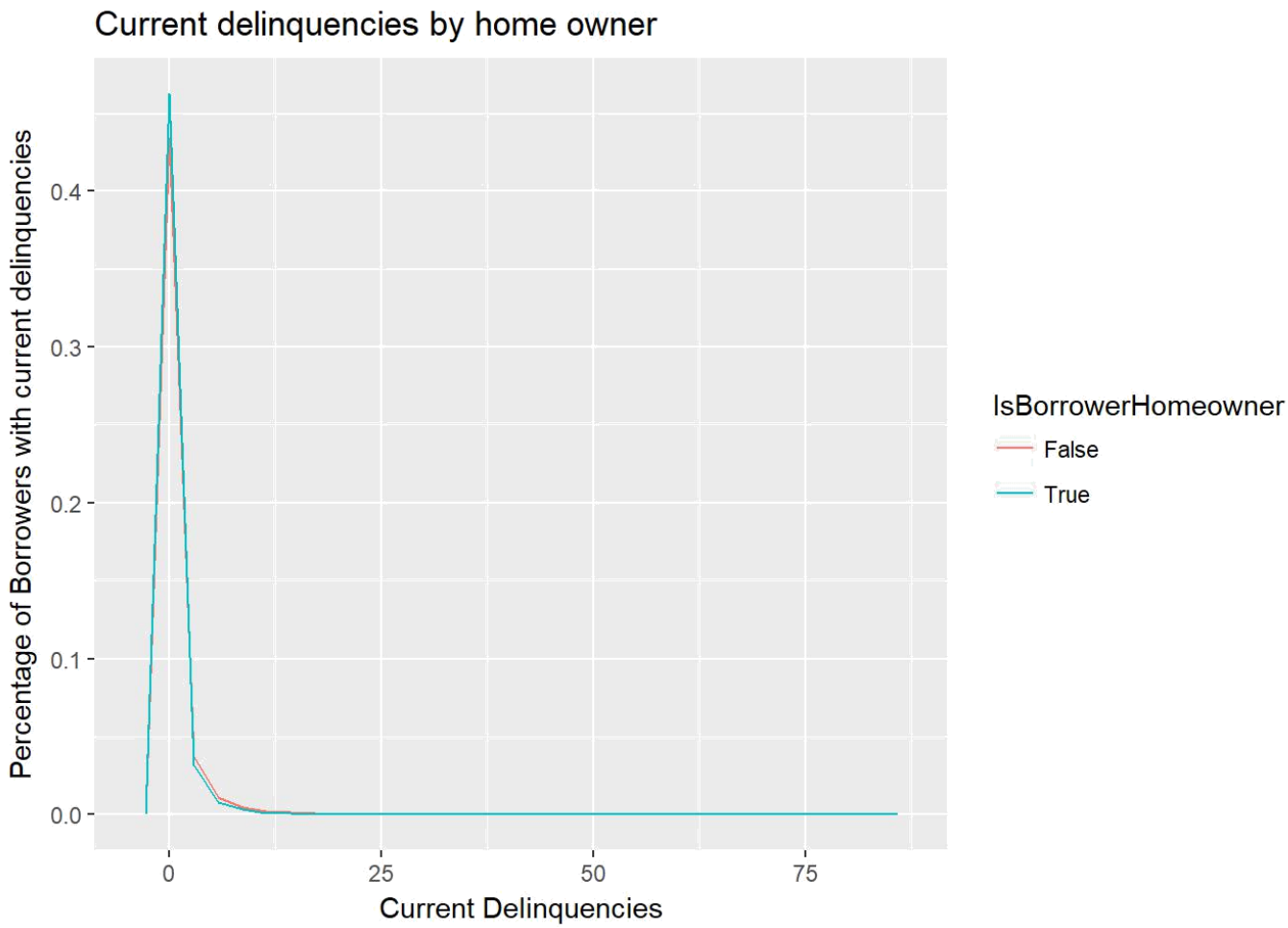
The graphs shows that there is not much difference in the median prosper scores of the borrowers with and without own homes. The customers with own homes have a slightly higher prosper score compared to the ones who dont.

Here is another plot comparing the delinquencies of the borrowers with and without their own homes:

```
ggplot(aes(x=CurrentDelinquencies, y=..count../sum(..count..)),
       data = subset(pld, !is.na(IsBorrowerHomeowner))) +
  geom_freqpoly( aes(color = IsBorrowerHomeowner))
  + xlab('Current Delinquencies') +
  ylab('Percentage of Borrowers with current delinquencies')
  + ggtitle('Current delinquencies by home owner')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

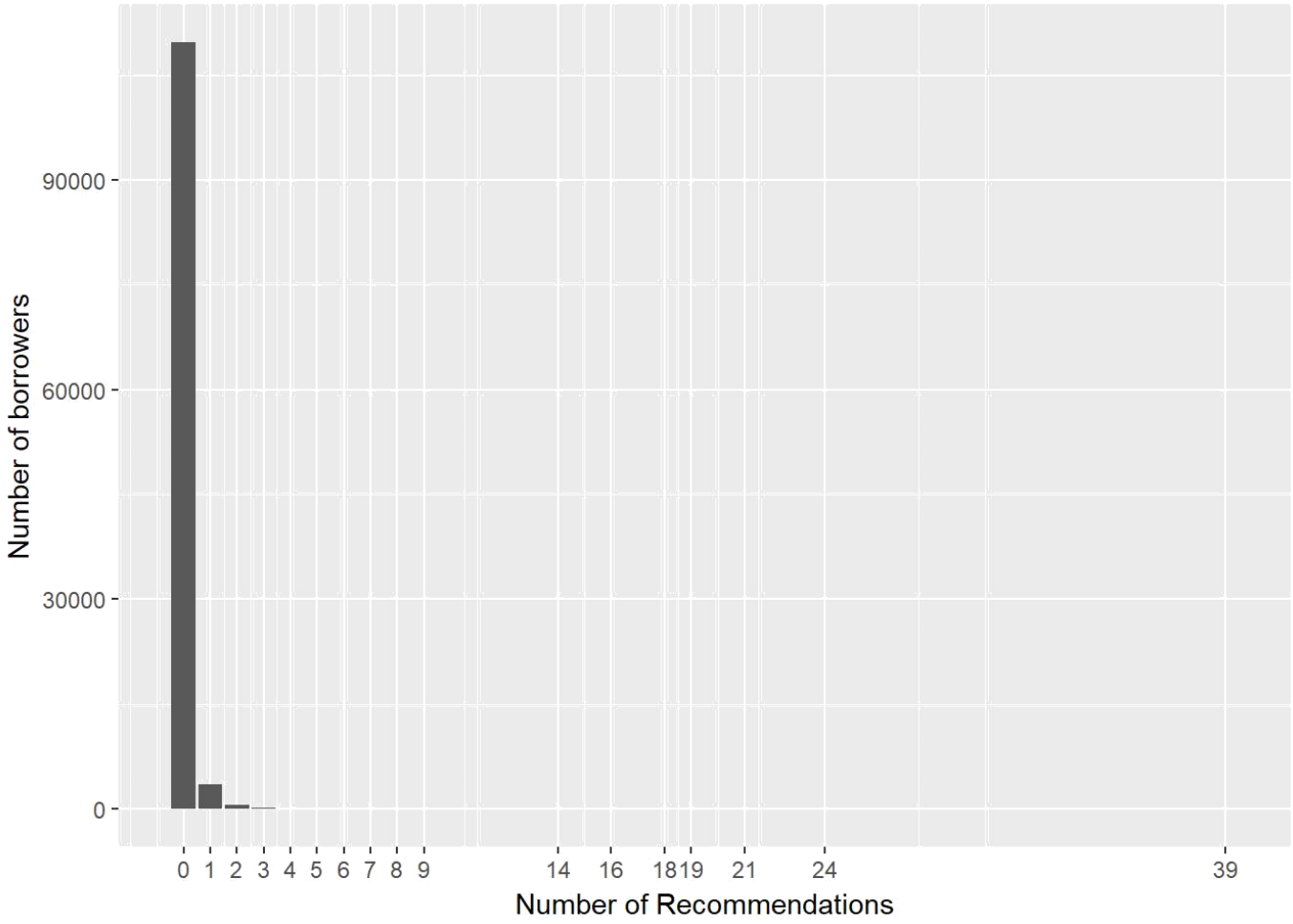
```
## Warning: Removed 697 rows containing non-finite values (stat_bin).
```



percentage of delinquencies remain almost the same for both the parties.

Number of Recommendations for the borrowers:

```
pld %>%
    group_by(Recommendations) %>%
    summarise(n = n()) %>%
    ggplot(aes(x = Recommendations, y = n)) +
    geom_bar(stat = 'identity', position="dodge") +
    ylab("Number of borrowers") +
    xlab("Number of Recommendations") +
    scale_x_continuous(breaks = unique(pld$Recommendations))
```



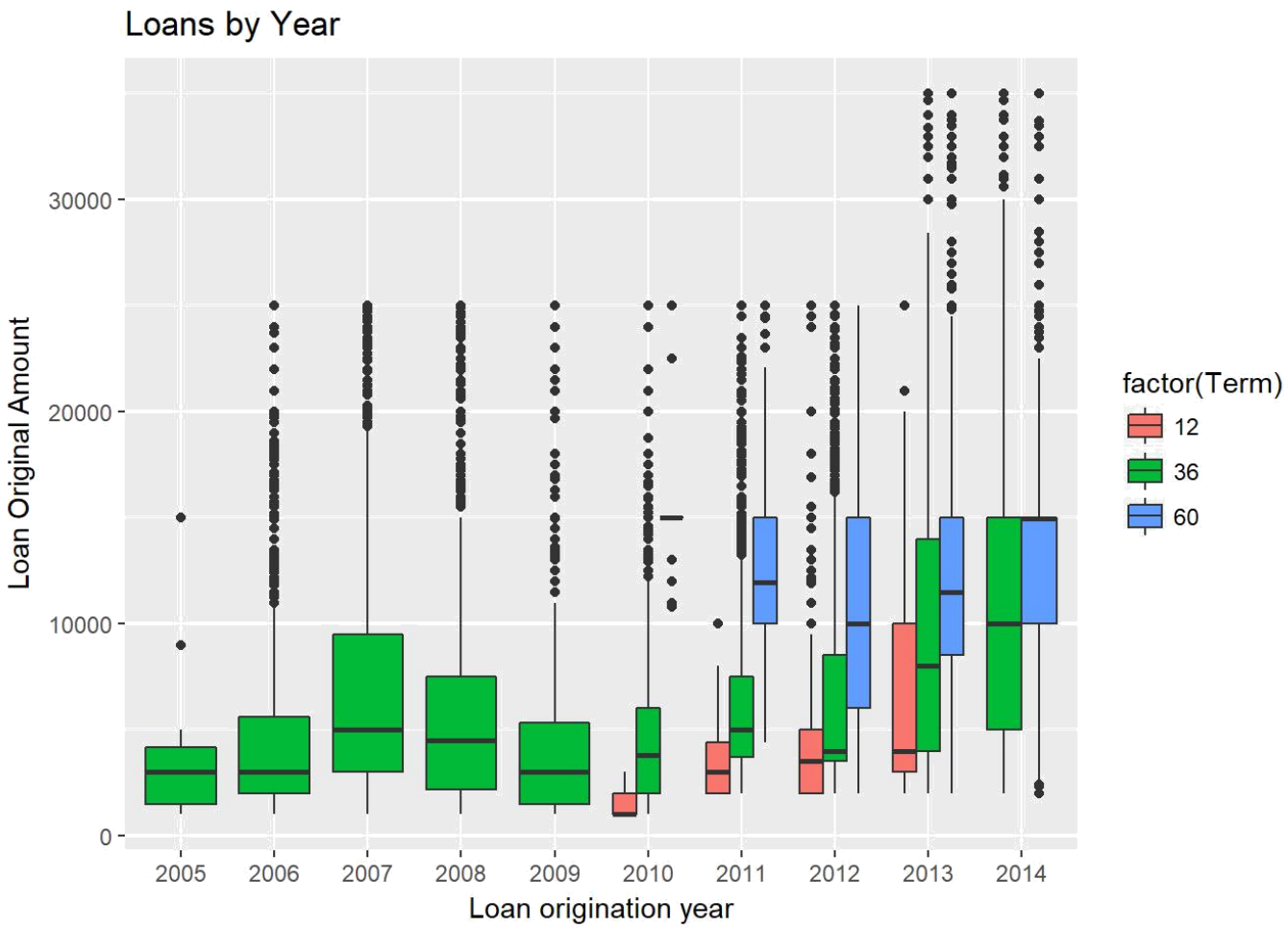
The Loans are granted for most people without any recommendations.

MultiVariate analysis:

We create boxbplot to show the variation in the loan amounts granted to the borrowers over years along with factor(term).

```
pld$LoanOriginationYear <-
  format(as.Date(pld$LoanOriginationDate, format="%Y-%m-%d"),"%Y")

ggplot(aes(x = LoanOriginationYear,
           y = LoanOriginalAmount,
           fill = factor(Term)),
       data = pld) +
  geom_boxplot()+
  ggtitle("Loans by Year") +
  labs(x = "Loan origination year", y = "Loan Original Amount")
```



The Loan amount granted have increase in the past few years. In the year 2014, the median loan granted has increased compared to the last 5 years. The factor or term required for repayment of the loan has also lengthened. The sizes of the boxes have also increased over the years, indicating that more number of people are granted with loans nowadays compared to 5 years ago.

We now take a look at the different types of people who apply for loan. We look at the background of the borrower and note down their occupations. We then make analysis based on the occupations of the borrowers:

```
pld$Group_By_Occupation <- factor(pld$Occupation)
levels(pld$Group_By_Occupation) <- list(
  Student=c("Student - College Graduate Student",
            "Student - College Senior",
            "Student - Community College",
            "Student - College Freshman",
            "Student - College Junior",
            "Student - College Sophomore",
            "Student - Technical School"),
  Medical=c("Doctor", "Nurse's Aide",
                    "Nurse (RN)",
                    "Nurse (LPN)",
                    "Dentist",
                    "Pharmacist",
                    "Medical Technician",
                    "Psychologist"),
  Sales=c("Sales - Commission",
          "Sales - Retail",
          "Car Dealer",
          "Realtor"),
```

```
  Service=c("Food Service Management",
            "Food Service",
            "Postal Service",
            "Social Worker",
            "Truck Driver",
            "Bus Driver",
            "Retail Management",
            "Waiter/Waitress",
            "Flight Attendant",
            "Clerical",
            "Religious",
            "Clergy"),
  Laborer=c("Construction",
            "Laborer",
            "Skilled Labor",
            "Landscaping",
            "Homemaker",
            "Fireman",
            "Executive",
            "Teacher's Aide",
            "Computer Programmer",
            "Administrative Assistant",
            "Professional",
            "Accountant/CPA",
            "Tradesman - Carpenter",
            "Tradesman - Mechanic",
            "Tradesman - Electrician",
            "Tradesman - Plumber",
            "Pilot - Private/Commercial"),
  HigherEduJobs=c("Architect",
                  "Biologist",
                  "Engineer - Electrical",
                  "Engineer - Mechanical",
                  "Engineer - Chemical",
                  "Judge", "Teacher",
                  "Scientist",
                  "Professor",
                  "Attorney", "Analyst", "Accountant/CPA"
  ),
  CivilService=c("Civil Service",
                  "Military Officer",
                  "Police Officer/Correction Officer",
                  "Military Enlisted"),
  Other=c("Other", "")
)
```

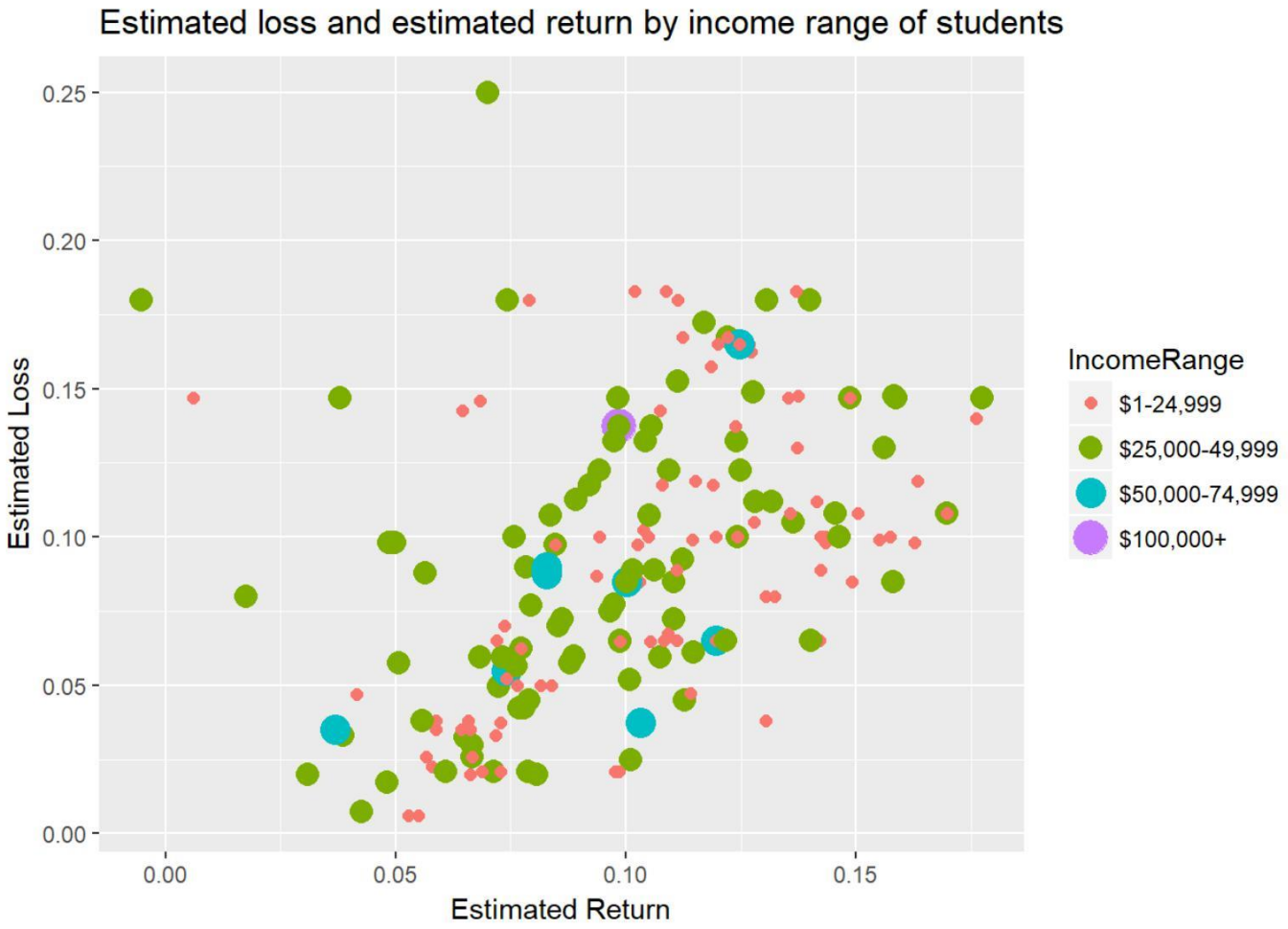
From this we take a sample of students as students are the ones who are loan applications have been on the rise.

We make a plot of Estimated return vs Estimated loss for the students keeping their income range in consideration:

```
ggplot(aes(EstimatedReturn, EstimatedLoss),
       data=subset(pld, Group_By_Occupation == 'Student'
                   & !is.na(Group_By_Occupation)
                   & !is.na(IncomeRange)
                   & IncomeRange != 'Not displayed'
                   & IncomeRange != 'Not employed'))+
  geom_point(aes(colour=IncomeRange, size=IncomeRange))+
  labs(x="Estimated Return", y = "Estimated Loss") +
  ggtitle('Estimated loss and estimated return by income range of students')
```

```
## Warning: Using size for a discrete variable is not advised.
```

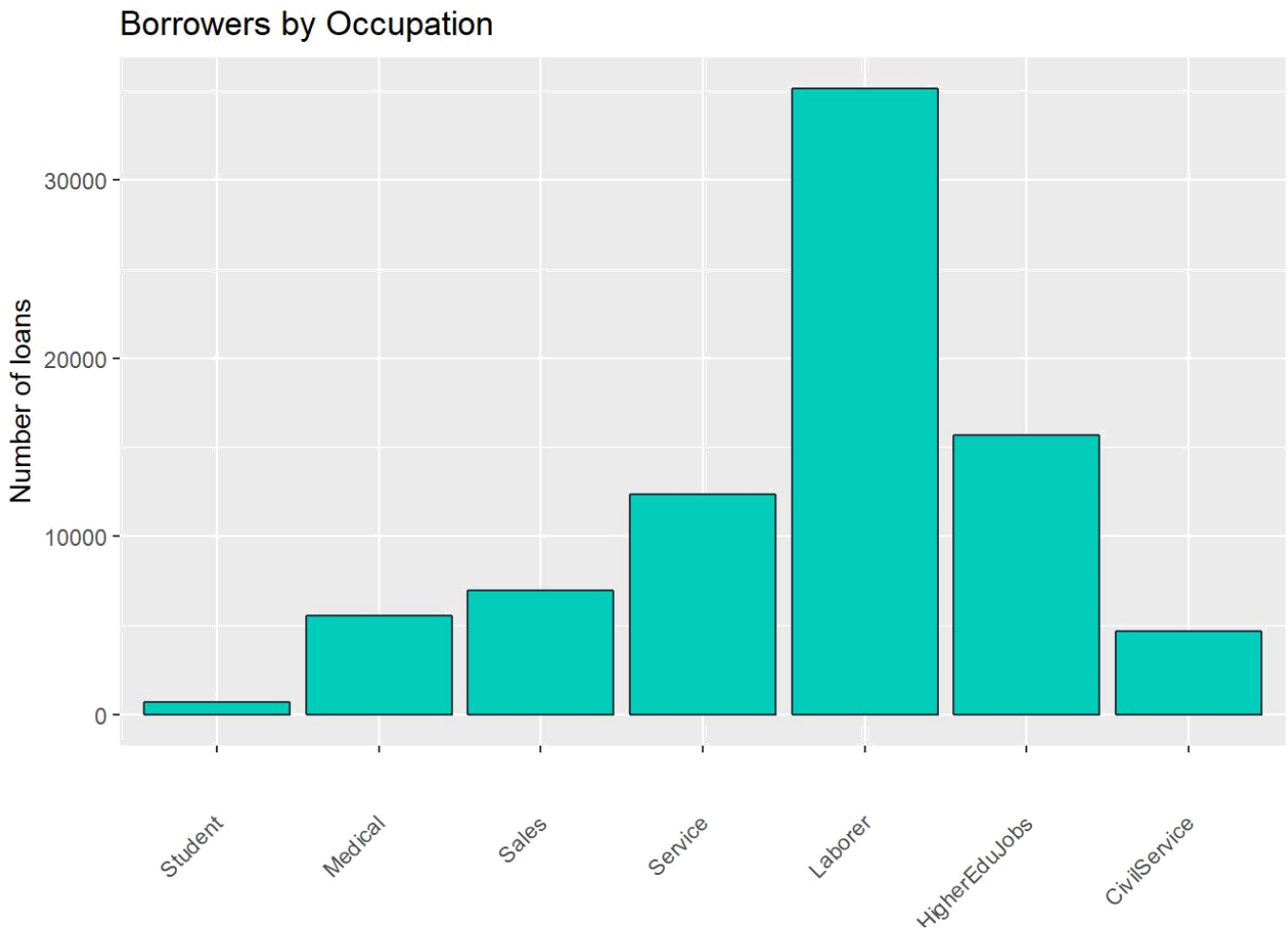
```
## Warning: Removed 276 rows containing missing values (geom_point).
```



It is evident that students who earn less have more Estimated loss due to the interest paid in the loans.
Prosper receives less return on the students who have an annual income range of over $50,000.

Now lets have a look at the occupation of the borrowers who take the most number of loans.

```
ggplot(data=subset(pld,
                   Group_By_Occupation != 'Other' &
                   !is.na(Group_By_Occupation)),
       x=Group_By_Occupation, aes(Group_By_Occupation)) +
  geom_bar(colour='#24323e', fill='#02ccba')+
  ggtitle("Borrowers by Occupation") + labs(x="", y =
  "Number of loans") +
  theme(axis.text.x=element_text(angle=45,hjust=1,vjust=0.5))
```

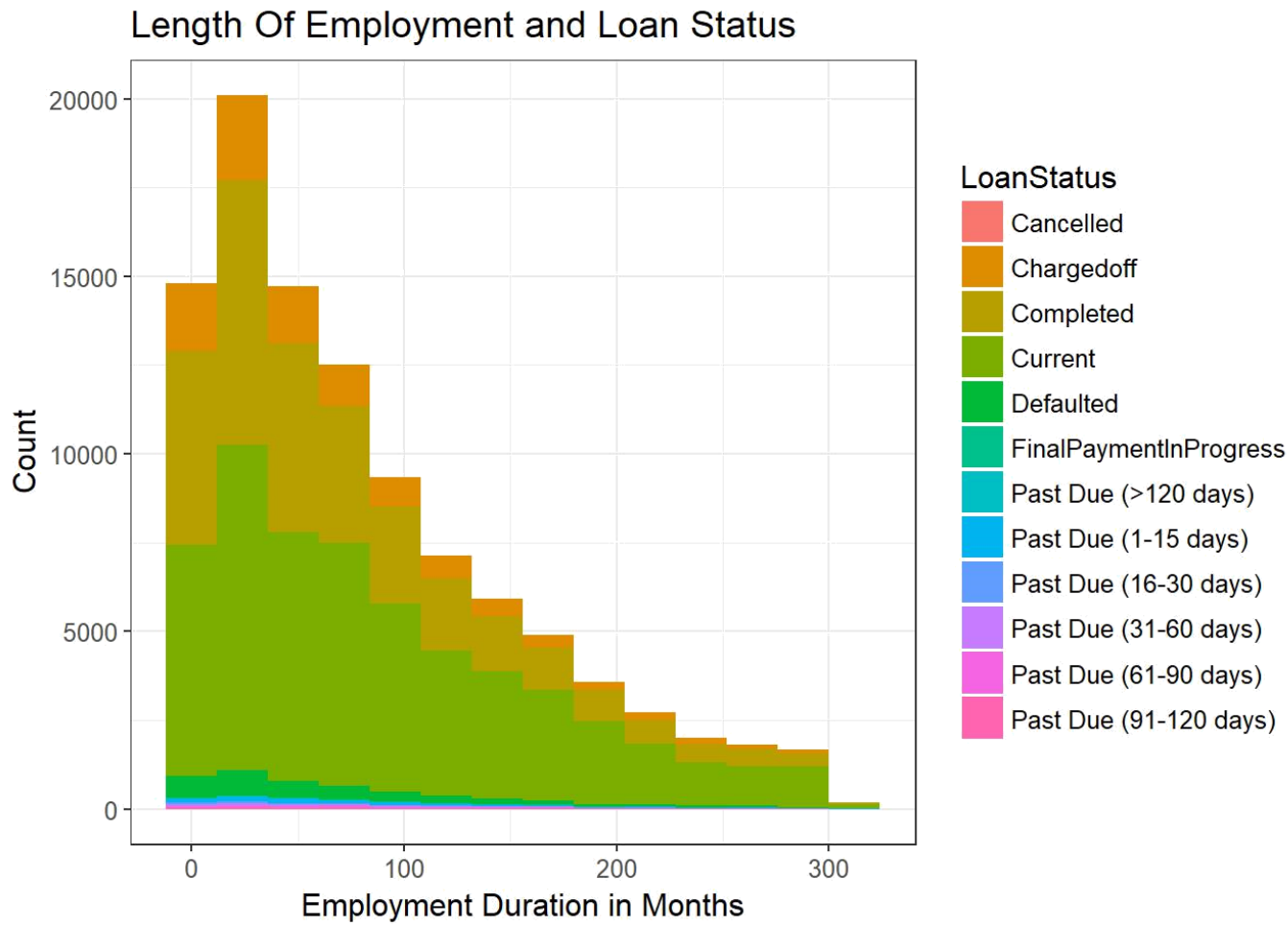


We see that most number of people who apply for loans are laborers. This is for them to make trades and make profit out of it. And it is surprising to see that Students borrow the least number of loans.

Let us see if there is any relation between the Employment duration and the Loan Status. Generally people who are employed longer have completed their loan terms. Let us see what the data has to say:

```
sub <- pld[,c("EmploymentStatusDuration", "LoanStatus")]
sub <- sub[-which(is.na(sub[,1]) | is.na(sub[,2])),]
sub <- sub[-which(sub[,1] %in% boxplot(sub[,1], plot=F)$out),]


ggplot() + geom_histogram(data=sub, aes(x=EmploymentStatusDuration,
                                        fill=LoanStatus),
                          binwidth=24) +
  xlab("Employment Duration in Months") +
  ylab("Count") +
  ggtitle("Length Of Employment and Loan Status")
  + theme_bw() + theme(text=element_text(size=12))
```



We can infer the following from the graph:

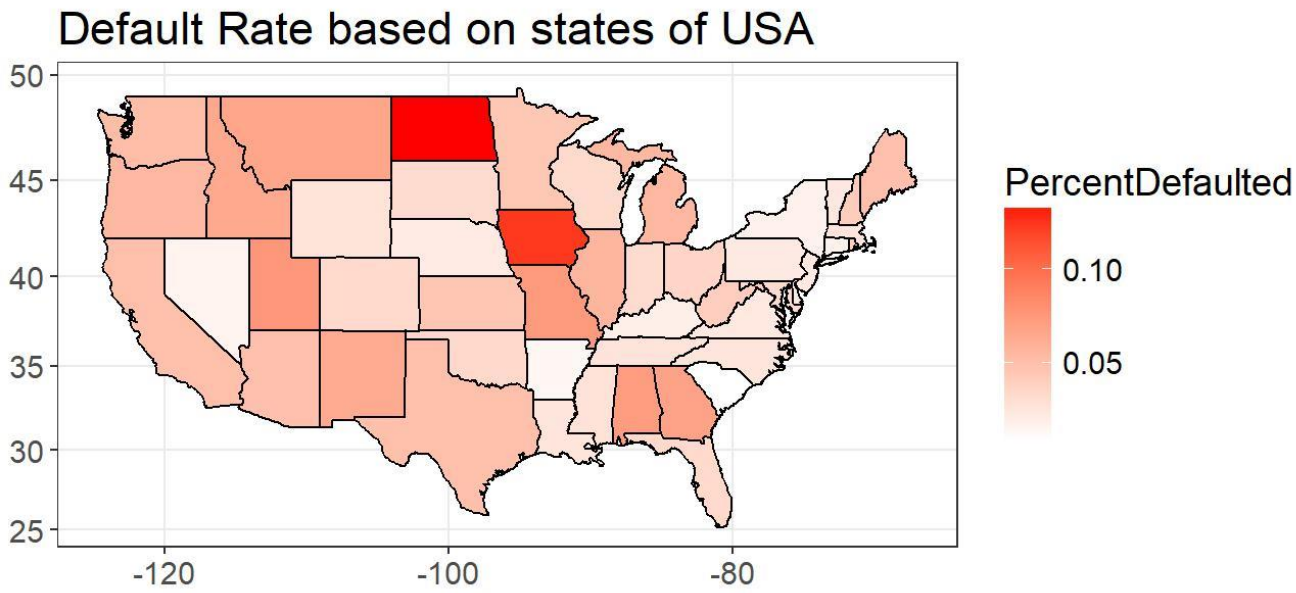
1. People apply for loans in the beginning stage after employment.
2. Around one quarter of the people pay back the loan amount in a short span of 100 months or less.
3. Many people take loans for long term periods and continue to repay their loans. The 3rd statement above is beneficial for the banks as long term repayments give more credit to the banks.

Here is the USA map showing the different states and the percentage of people defaulted in them.

```
library(maps)

library(mapproj)
states <- map_data("state")
status <- function(x) {
  l <- length(which(pld[which(pld$BorrowerState == x), "LoanStatus"] == "Defaulted
"))
  l / length(which(pld$BorrowerState == x))
}
s <- sapply(state.abb, status)
s <- data.frame("region"=tolower(state.name), "PercentDefaulted"=s)
map_df <- merge(states, s, by="region", all.x=T)
map_df <- map_df[order(map_df$order),]

ggplot(map_df, aes(x=long, y=lat, group=group)) +
  geom_polygon(aes(fill=PercentDefaulted)) +
  geom_path() +
  scale_fill_gradient(low="white", high="red",
                      na.value="white") +
  coord_map() + theme_bw() +
  ggtitle("Default Rate based on states of USA") +
  theme(text=element_text(size=14)) + xlab("") + ylab("")
```



From the map we can make following conclusions:

1. South Dakota and New mexico have more percentage of defaulters.
2. North Dakota and Iowa have the least number of defaulters.

It is important to understand the reason why the people borrow loans. Generally people buy loans to grow

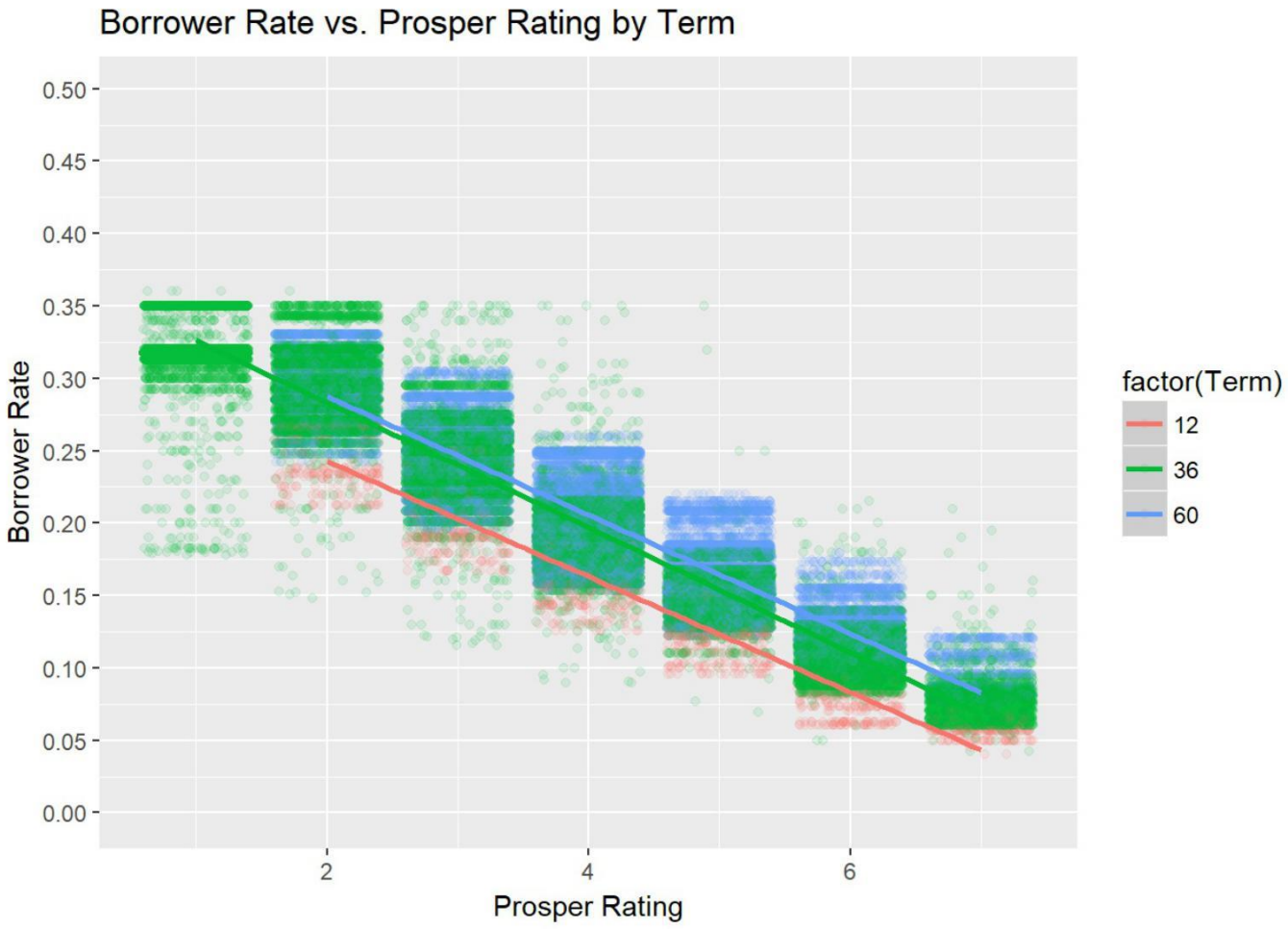
their business, buy a new car, build houses and clear their debts. Let us see what the prosper data set has to say:

Final Plot 1:

```
ggplot(data = pld,
       aes(x = ProsperRating..numeric.,
           y = BorrowerRate,
           color = factor(Term))) +
  geom_point(alpha = 0.1, position = position_jitter()) +
  geom_smooth(method = "lm") +
  labs(title = "Borrower Rate vs. Prosper Rating by Term",
       x = "Prosper Rating",
       y = "Borrower Rate") +
  scale_y_continuous(breaks = seq(0, .5, .05))
```

```
## Warning: Removed 29084 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 29084 rows containing missing values (geom_point).
```



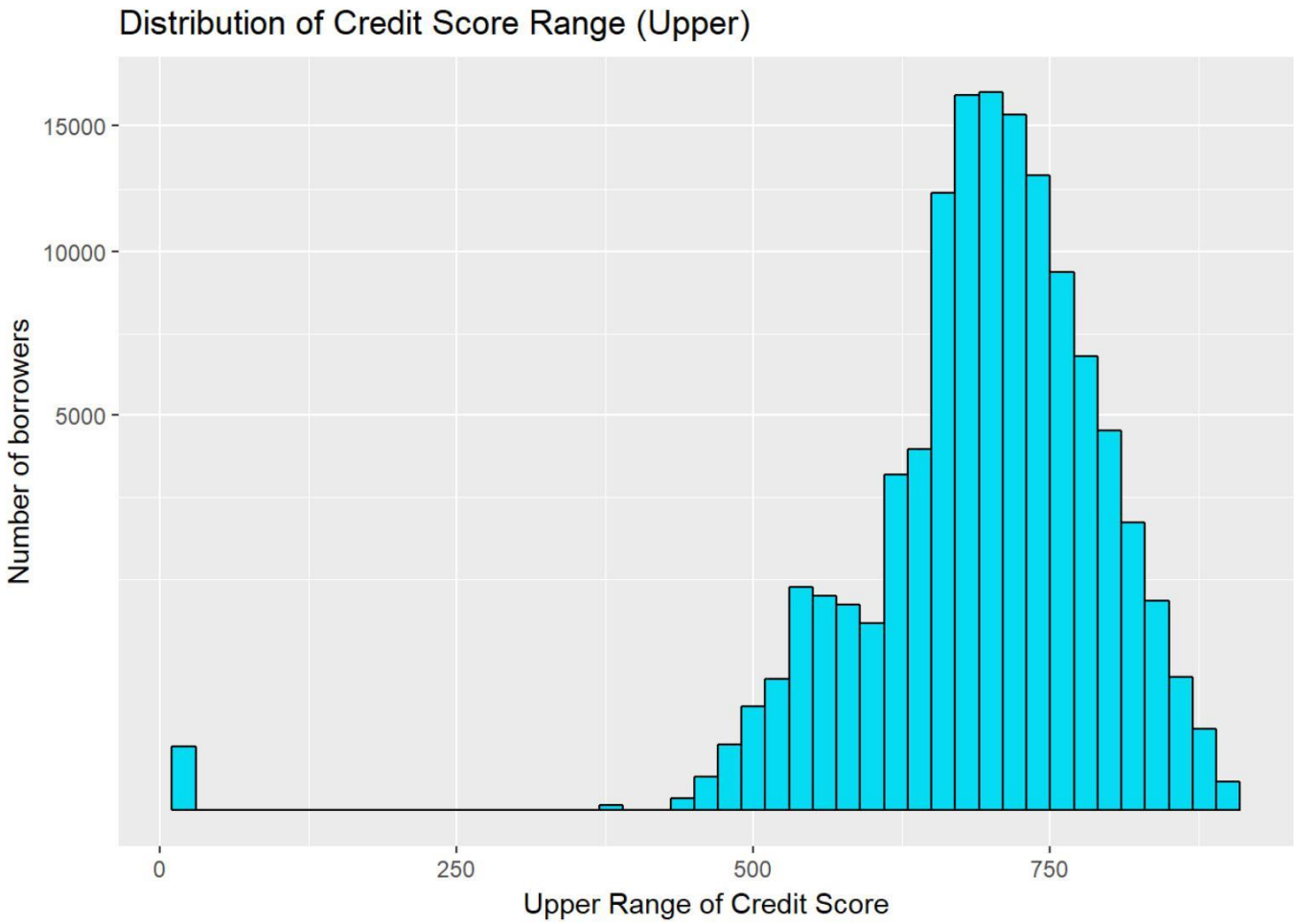
Summary 1:

This plot of the BorrowerRate vs. ProsperRating shows that a lower interest rate is expected for a borrower with a better credit rating. When grouping by the loan term in months, the interest rate tends to be lower for shorter terms.

Final Plot 2:

```
ggplot(pld, aes(x = CreditScoreRangeUpper)) +
  geom_histogram(aes(y = ..count..), binwidth = 20, fill = '#05DBF2',
        color='black', position="identity") +
  scale_y_sqrt() +
        labs(x= 'Upper Range of Credit Score', y ='Number of borrowers',
             title="Distribution of Credit Score Range (Upper)")
```

```
## Warning: Removed 591 rows containing non-finite values (stat_bin).
```




```
summary(pld$CreditScoreRangeLower)
```

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

```
summary(pld$CreditScoreRangeUpper)
```

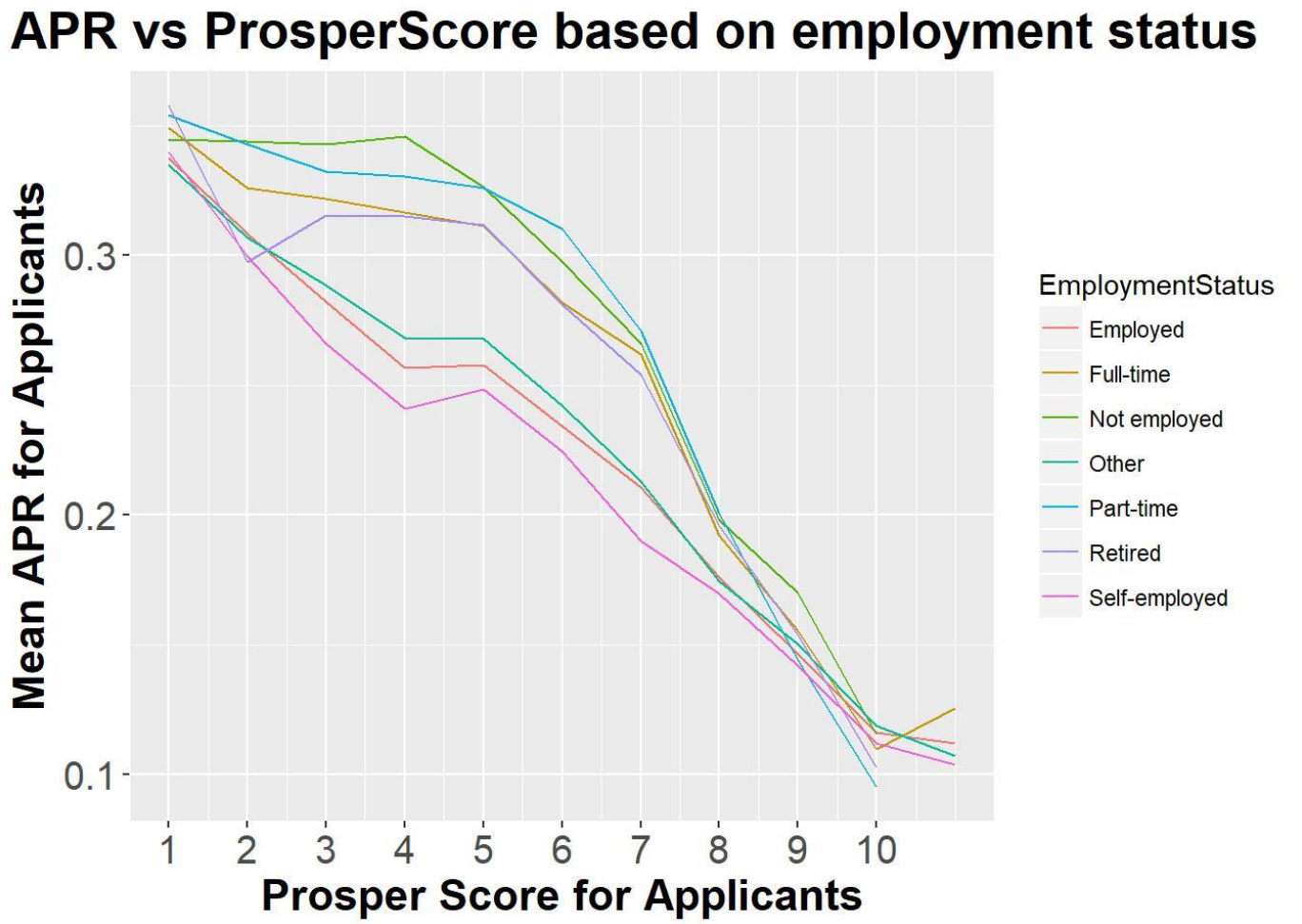
```
##    Min. 1st Qu.    Media     Mean 3rd Qu.     Max.     NA's
##    19.0   679.0        n    704.6   739.0    899.0      591
##                    699.0
```

Summary 2: Median Credit score range of a borrower is 680.0 to 699.0 and Mean ranges from 685.6 to 704.6. The max score is 880 to 899 and there are 133 borrowers with very low credit score (0-19).

Final Plot 3:

```
ggplot(data = pld, aes(x = ProsperScore, y = BorrowerAPR))+
  geom_line(aes(color = EmploymentStatus), stat = 'summary' , fun.y = mean)+
  scale_x_continuous(breaks = seq(1,10,1))+
  xlab("Prosper Score for Applicants")+
  ylab("Mean APR for Applicants")+
  ggtitle('Mean APR vs ProsperScore based on employment status')+
  theme(plot.title = element_text(face = 'bold', size = 20, hjust = 0.5),
        axis.title.x = element_text(face = 'bold', size = 17),
        axis.title.y = element_text(face = 'bold', size = 17),
        axis.text.x = element_text(size = 15),
        axis.text.y = element_text(size = 15))
```

```
## Warning: Removed 29084 rows containing non-finite values (stat_summary).
```



This third plot was chosen to show the effect of employment status on the APR by plotting the mean APR for each prosper score by the employment status. The plot shows that employed, self-employed applicants had a much lower APR than the not employed, part time applicants. This clearly shows the status did have a considerable effect on the APR. Although this difference can be seen till a score of 7 and the gap closed beyond that which can be attributed to the higher credit score, lower debt of the applicants despite their employment status.

Reflection:

The Prosper Loans dataset contains a wealth of information about 114K loans given between Nov 2005 and Mar 2014.

I began my investigation by going through each of the variables. I then categorized them into variables that are useful and should be taken into consideration and the ones that not be taken under consideration.

I continued exploring individual variables that I had noticed, like the loan amount, term, interest rates, credit scores, loan status, and loan purposes. At first I leaned towards the relationship between the interest rates and the credit score, but because I found it so evident I decided to give it another thought and I read again the definition of the variables. I thought backwards, like a borrower, and decided to explore what would give me a good or bad credit rating that would deserve me a lower interest rate.

Then I went on to decide the types of plots that would be apt for depicting the models in a clear and understandable way. I went through a few difficulties in plotting the models, especially when i took 3 variables into consideration. Another difficulty was that some of the variables had no corelation and it was difficult to express them in a graph and derive the results.

Summary:

In most of the cases people take loans for their own personal needs. In 10% of the case people borrow loans to buy automotives. 8% of the loans are borrowed to make payments. 6.2% of loans are taken for Home Improvement. Remaining loans are taken of other specific reasons mentioned in the graph.

Thus , an analysis has been made on the Prosper Loan Data Set. We saw on what basis the bank grants loans, the occupation of customers, the reason why they need loans and when the banks make a profit out of it.

Future Work:

We can create plots that instantly tells us whether a borrower is eligible for a particular loan amount or not based on his credit history and background. The plots can also describe the profit rates achieved and the most valuable customers thorugh the profits. Based on the plotting, banks can tune their interest rates for their borrowers depending on their occupation. All this contributes in growing the yeild of Prosper.