Prosper Loan Data R Project

Vincent Man

11/07/2018

Prosper is a US based peer-to-peer online lending provider and it connects potential investors to help fund people in loans for those who are struggling to pay off their debts such as medical bills or have difficulties in their startup.

Borrowers can request for personal loans simply by creating their online profiles and borrower listings then investors will decide whether or not to accept their loan request or not. Rather than requesting loans from traditional banks, people are able to borrow from a big group of investors.

Prosper itself do not lend borrowers the money but earns its revenue through service fees.

In this finance data project, I will use R to perform EDA (exploratory data analysis) to analyse the Prosper dataset and uncover interesting insights by extracting important variables and detect outliers and anomalies. To do this I will use various techniques when employing EDA by plotting the raw data and simple statistics on one variable to multiple variables. By doing all of this I will be able to use the valuable information to help investors to pick the safest option when looking at potential borrowers.

```
getwd()
```

```
## [1] "/Users/vince/Desktop/Udacity/Vincent Man R Project"
```

```
setwd("~/Desktop/Udacity/Vincent Man R Project")

#Load all packages for this project

library(ggplot2)
library(gridExtra)
library(RColorBrewer)

# Load Prosper loan data
df <- read.csv("prosperLoanData.csv")
```

Look into the dataset dimensions

```
dim(df)
```

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

List all the Prosper variables

```
names(df)
```

```
##  [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"
```

Selecting my variables

Now I want to choose the variables that will be predominantly part of my analysis.

```
prosper <- subset(df, select=c("Term", "LoanStatus", "BorrowerRate",
               "ProsperScore", "ListingCategory..numeric.", "BorrowerState",
               "EstimatedReturn", "CreditScoreRangeUpper",
               "CreditScoreRangeLower", "CurrentDelinquencies",
               "IncomeRange", "LoanOriginalAmount",
               "LoanOriginationQuarter", "ProsperRating..Alpha.",
               "IsBorrowerHomeowner"))
```

Creating new variables

After looking at the dataset there are new variables that I need to create:

```
# Rather than counting the number of times a borrower has been a delinquent it wil
l be categorised as Delinquent Borrower (1) or not Delinquent Borrower (0).

prosper$DelinquentBorrower <- ifelse(prosper$CurrentDelinquencies == TRUE, 1, 0)

# There are two variables for the Credit Score (Lower range and Upper range) so I
want to combine them together by calculating the mean.

prosper$CreditScore <- (prosper$CreditScoreRangeUpper + prosper$CreditScoreRangeLo
wer) /2

# Convert ListingCategory..numeric into a factor variable with labels.

prosper$ListingCategory <- factor(prosper$ListingCategory..numeric.,
                            labels=c("Not available", "Debt consolidation",
                                     "Home improvement", "Business",
                                     "Personal loan", "Student use",
                                     "Auto", "Other", "Baby & Adoption Loans",
                                     "Boat", "Cosmetic Procedures",
                                     "Engagement Ring Financing",
                                     "Green Loans", "Household Expenses",
                                     "Large Purchases", "Medical/Dental",
                                     "Motorcycle", "RV", "Taxes",
                                     "Vacation", "Wedding Loans"))
```

Sorting existing variables

Factor variables that needs to be sorted into the correct order

```
prosper$LoanOriginationQuarter <- factor(prosper$LoanOriginationQuarter,
                                         levels = c("Q4 2005",
                                                    "Q1 2006","Q2 2006","Q3 2006",
"Q4 2006",
                                                    "Q1 2007","Q2 2007","Q3 2007",
"Q4 2007",
                                                    "Q1 2008","Q2 2008","Q3 2008",
"Q4 2008",
                                                    "Q2 2009","Q3 2009","Q4 2009",
                                                    "Q1 2010","Q2 2010","Q3 2010",
"Q4 2010",
                                                    "Q1 2011","Q2 2011","Q3 2011",
"Q4 2011",
                                                    "Q1 2012","Q2 2012","Q3 2012",
"Q4 2012",
                                                    "Q1 2013","Q2 2013","Q3 2013",
"Q4 2013",
                                                    "Q1 2014","Q2 2014","Q3 2014")
)

prosper$ProsperRating <- factor(prosper$ProsperRating..Alpha.,
                                levels= c("AA", "A", "B", "C", "D", "E", "HR", "NA
"))

prosper$ProsperScore <- factor(prosper$ProsperScore,
                               levels=c(11,10,9,8,7,6,5,4,3,2,1))


prosper$IncomeRange <- factor(prosper$IncomeRange,
                                levels= c("Not displayed", "Not employed",
                                          "$0", "$1-24,999", "$25,000-49,999",
                                          "$50,000-74,999", "$75,000-99,999",
                                          "$100,000+"))
```

Now let's take a look at the final dataset

```
str(prosper)
```

```
## 'data.frame':    113937 obs. of  19 variables:
##  $ Term                     : int  36 36 36 36 36 60 36 36 36 36 ...
##  $ LoanStatus               : Factor w/ 12 levels "Cancelled","Chargedoff",..:
3 4 3 4 4 4 4 4 4 4 ...
##  $ BorrowerRate             : num  0.158 0.092 0.275 0.0974 0.2085 ...
##  $ ProsperScore             : Factor w/ 11 levels "11","10","9",..: NA 5 NA 3 8
2 10 8 3 1 ...
##  $ ListingCategory..numeric.: int  0 2 0 16 2 1 1 2 7 7 ...
##  $ BorrowerState            : Factor w/ 52 levels "","AK","AL","AR",..: 7 7 12
12 25 34 18 6 16 16 ...
##  $ EstimatedReturn          : num  NA 0.0547 NA 0.06 0.0907 ...
##  $ CreditScoreRangeUpper    : int  659 699 499 819 699 759 699 719 839 839 ...
##  $ CreditScoreRangeLower    : int  640 680 480 800 680 740 680 700 820 820 ...
##  $ CurrentDelinquencies     : int  2 0 1 4 0 0 0 0 0 0 ...
##  $ IncomeRange              : Factor w/ 8 levels "Not displayed",..: 5 6 1 5 8
8 5 5 5 5 ...
##  $ LoanOriginalAmount       : int  9425 10000 3001 10000 15000 15000 3000 10000
10000 10000 ...
##  $ LoanOriginationQuarter   : Factor w/ 35 levels "Q4 2005","Q1 2006",..: 8 33
6 28 31 32 30 30 32 32 ...
##  $ ProsperRating..Alpha.    : Factor w/ 8 levels "","A","AA","B",..: 1 2 1 2 6
4 7 5 3 3 ...
##  $ IsBorrowerHomeowner      : Factor w/ 2 levels "False","True": 2 1 1 2 2 2 1
1 2 2 ...
##  $ DelinquentBorrower       : num  0 0 1 0 0 0 0 0 0 0 ...
##  $ CreditScore              : num  650 690 490 810 690 ...
##  $ ListingCategory          : Factor w/ 21 levels "Not available",..: 1 3 1 17
3 2 2 3 8 8 ...
##  $ ProsperRating            : Factor w/ 8 levels "AA","A","B","C",..: NA 2 NA 2
5 3 6 4 1 1 ...
```

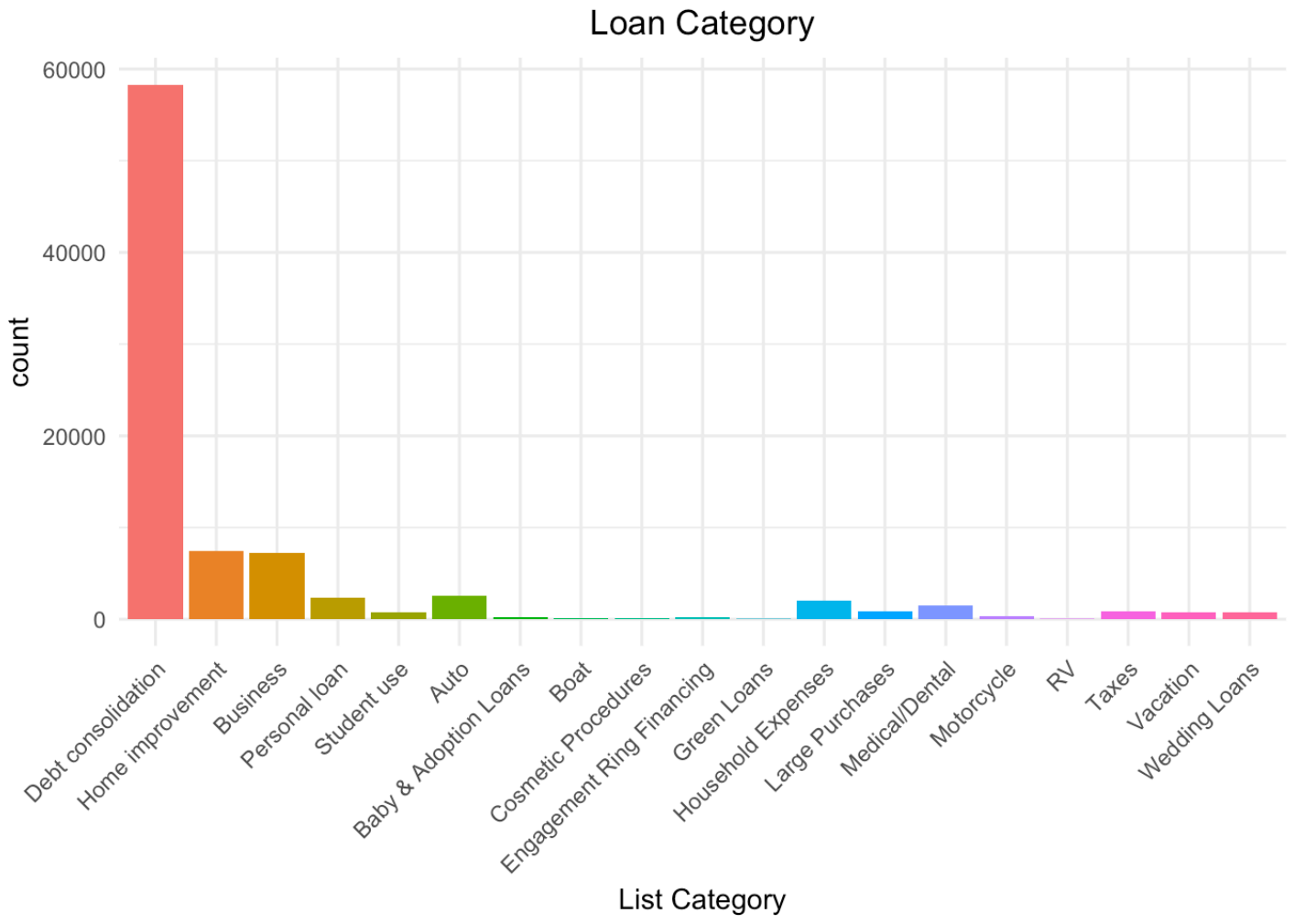
After investigating about the dataset, I want to further explore it with the following questions in mind:

- Who are the potential borrowers and what is their profile like?
- How is the Interest Rate created for new/current borrowers?
- How does Prosper measure risk for each borrower?
- How does Prosper estimate return for each borrower?

Univariate Plots Section

What is the main reason for borrowing?

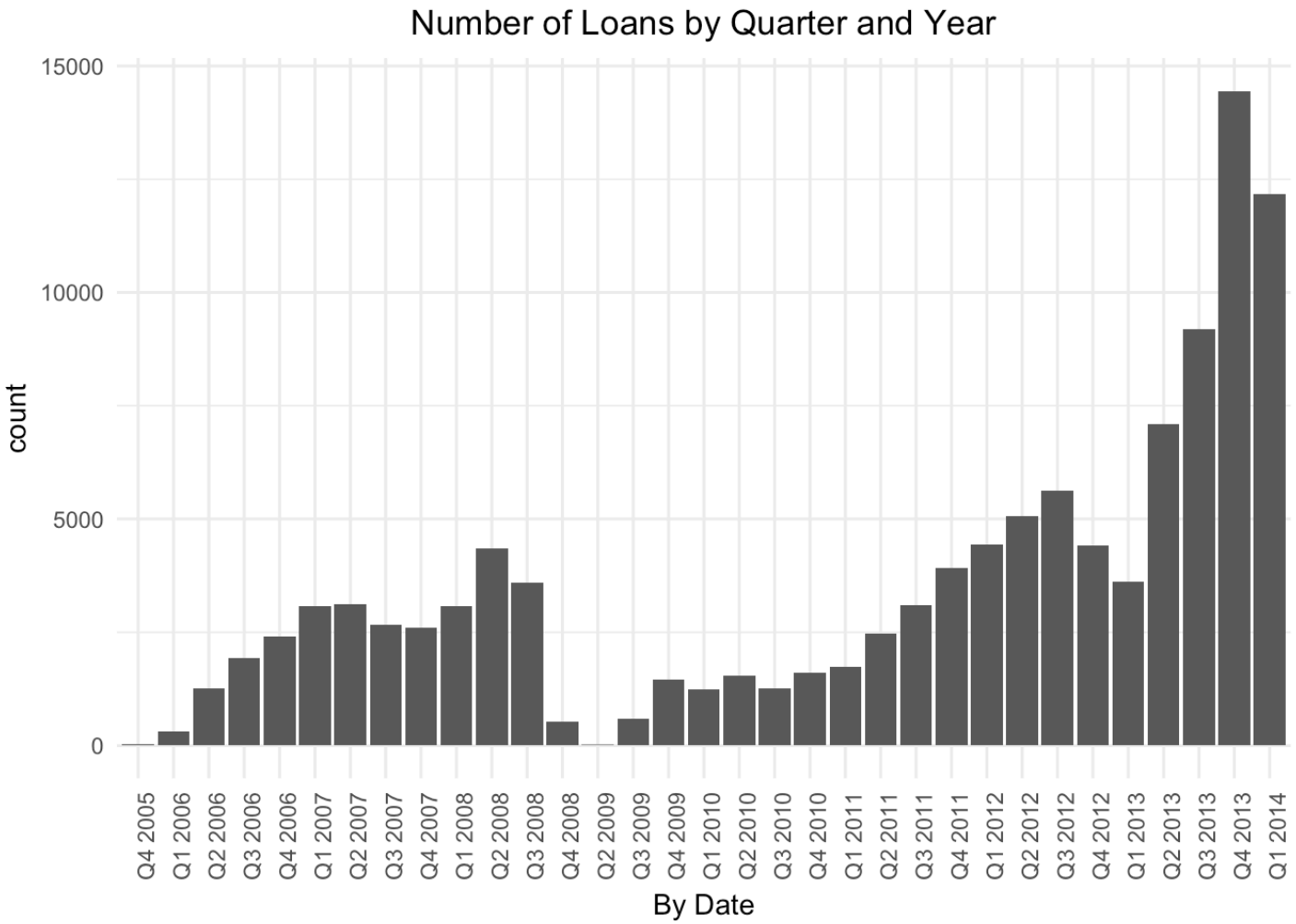
```
#I had to remove two categories by subsetting in order to make the analysis cleare
r
ggplot(data=subset(prosper, ListingCategory != "Not available" & ListingCategory !
= "Other"), aes(x=ListingCategory)) +
  geom_bar(aes(fill=ListingCategory)) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = .5)) +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1)) +
  guides(fill=FALSE) +
  labs(title="Loan Category",x="List Category")
```



Debt Consolidation is the biggest reason for borrowers applying for loans from Prosper. House Improvement and Business being the second and third largest reasons. It is a surprise to me that borrowers take Prosper loans to pay off other debts and loans.

What is the number of listings throughout the period?

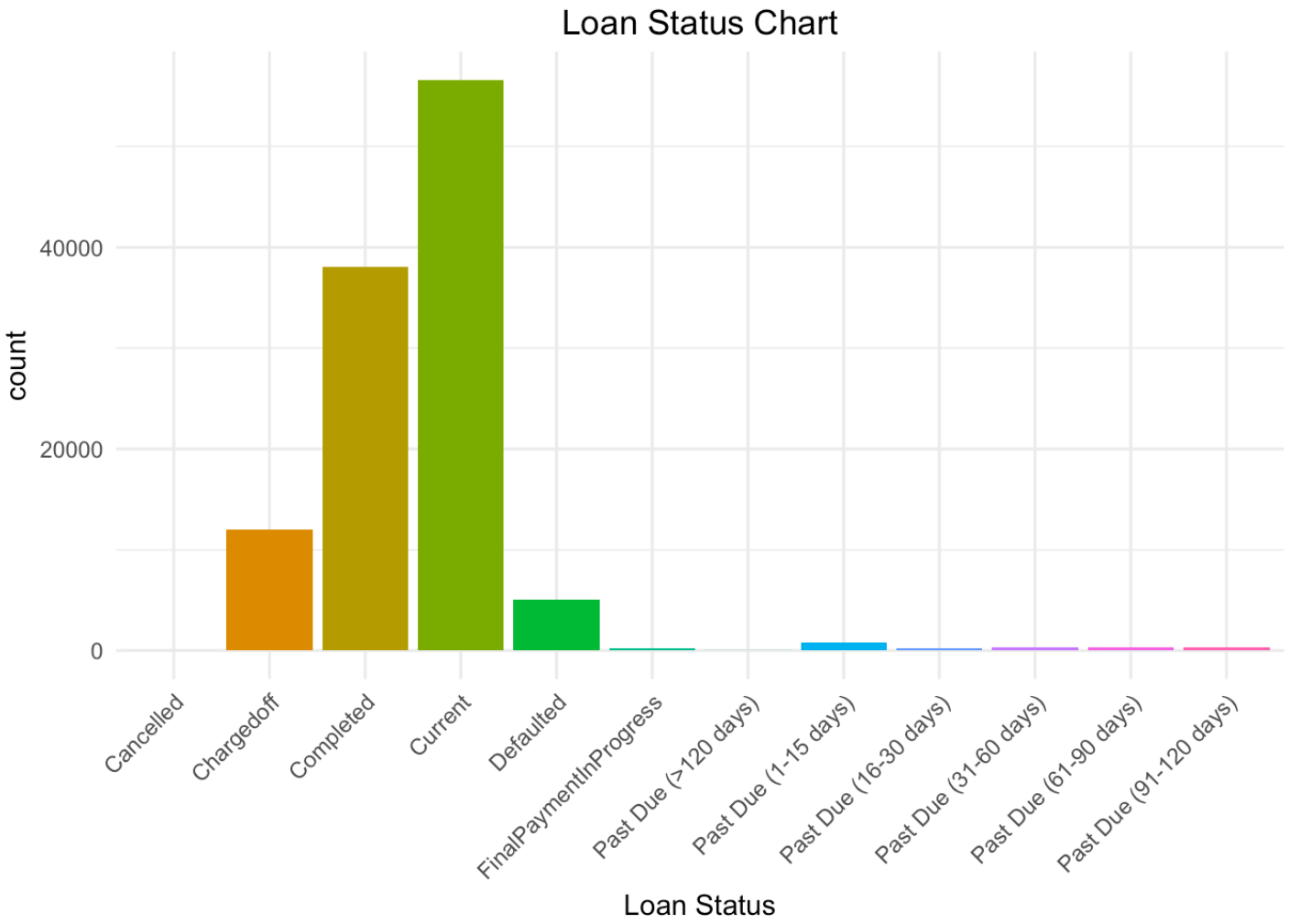
```
ggplot(data=prosper, aes(x= LoanOriginationQuarter)) +
  geom_bar() +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 1, hjust = 1)) +
  theme(plot.title = element_text(hjust = .5)) +
  guides(fill=FALSE) +
  labs(title="Number of Loans by Quarter and Year", x="By Date")
```



The number of borrowers fell dramatically from 2008 in the fourth quarter. This was due to the cease and desist order by SEC as Prosper was in violation of the Securities Act. Therefore, Prosper's peer-to-peer lending model had to be reviewed. Another reason could be Prosper's credit policy were not strict enough. After successfully obtaining SEC registration and the relaunch of Prosper's lending/investing website SEC in July 2009. The listings had started to grow gradually.

What is the Loan Status?

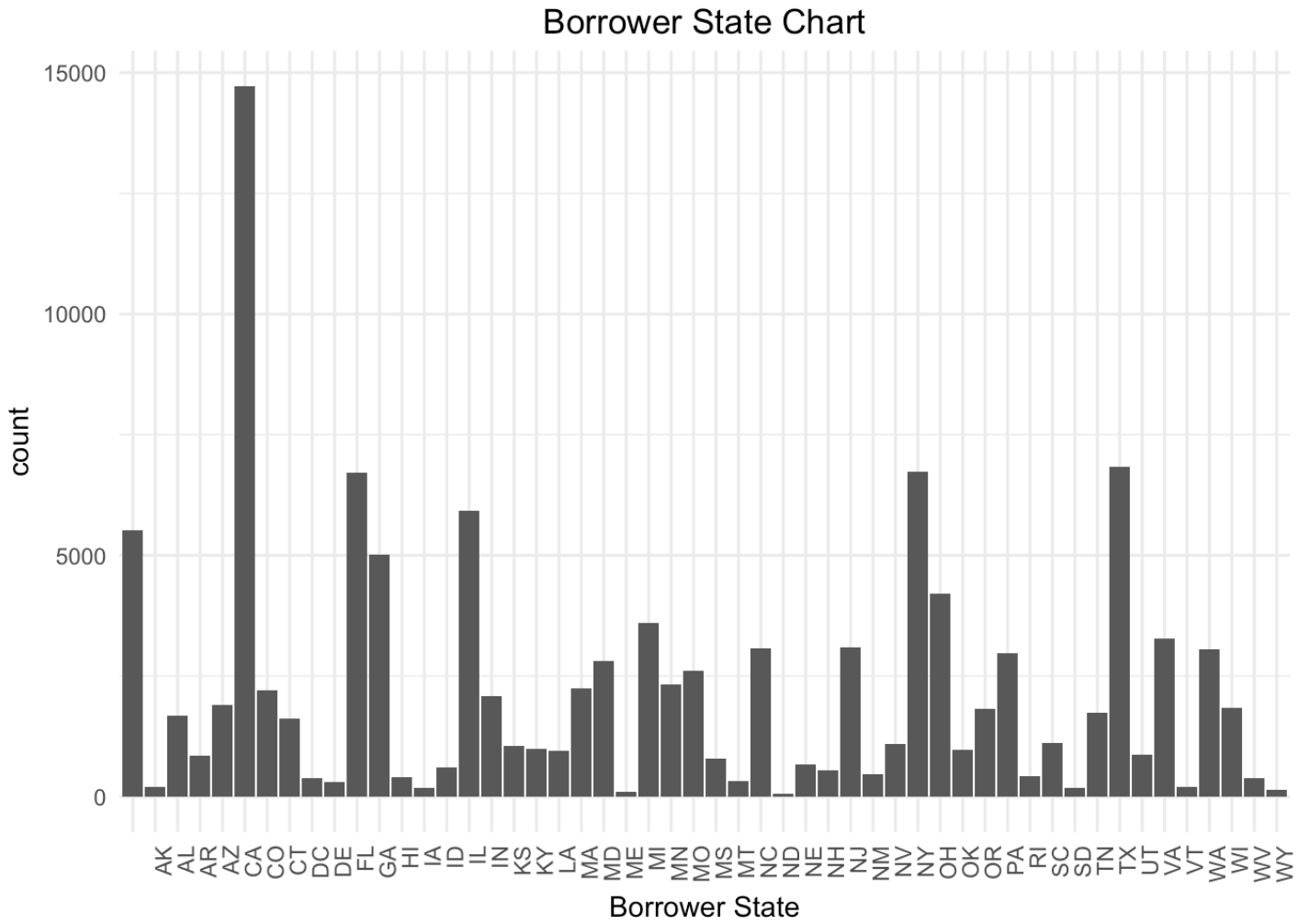
```
ggplot(data=prosper, aes(x= LoanStatus)) +
  geom_bar(aes(fill=LoanStatus)) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = .5)) +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1)) +
  guides(fill=FALSE) +
  labs(title="Loan Status Chart", x="Loan Status")
```



According to this plot, a vast majority of loans are mainly “Current” or “Completed”. However, there are a large number of borrowers who have a loan status of “Chargedoff” and “Defaulted” being the third and fourth largest. This means Prosper classified these borrowers as delinquents as they were unable to pay off their loan. Surprisingly none of the loans were cancelled.

What is the number of borrowers in each state?

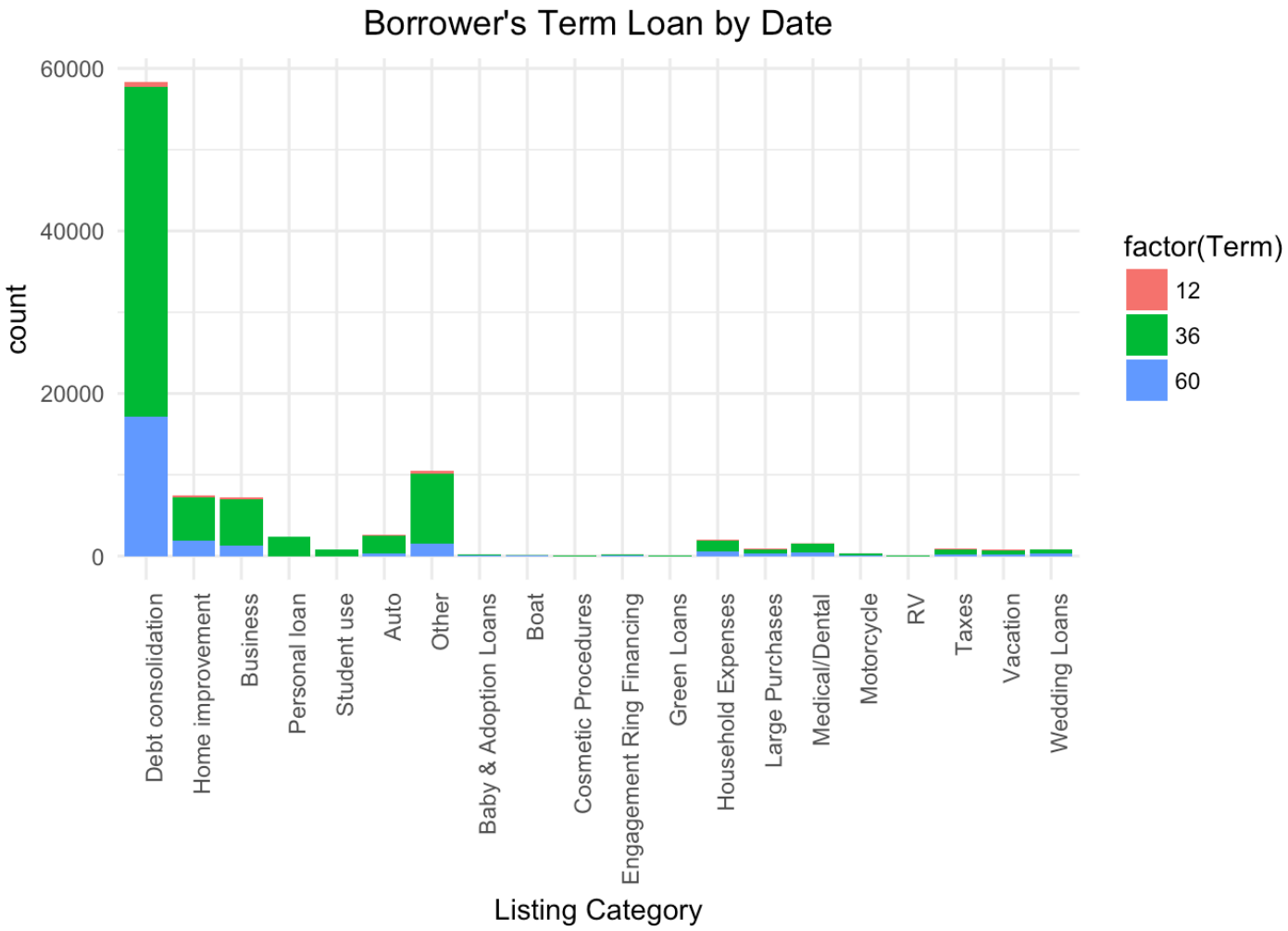
```
# Remove NAs in my analysis by subsetting.
ggplot(data=subset(prosper,!is.na(BorrowerState)), aes(x= BorrowerState)) +
  geom_bar() +
  theme_minimal() +
  theme(plot.title = element_text(hjust = .5)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 1, hjust = 1)) +
  guides(fill=FALSE) +
  labs(title="Borrower State Chart", x="Borrower State")
```



It can be seen that CA (California) has the most borrowers in the United States which is not a surprise as it is Prosper's original base. Moreover, FL (Florida), GA (Georgia), IL (IIIinois), NY (New York) and TX (Texas) contained a large number of borrowers with more than 5,000 in each state.

What is the number of Loans in Terms?

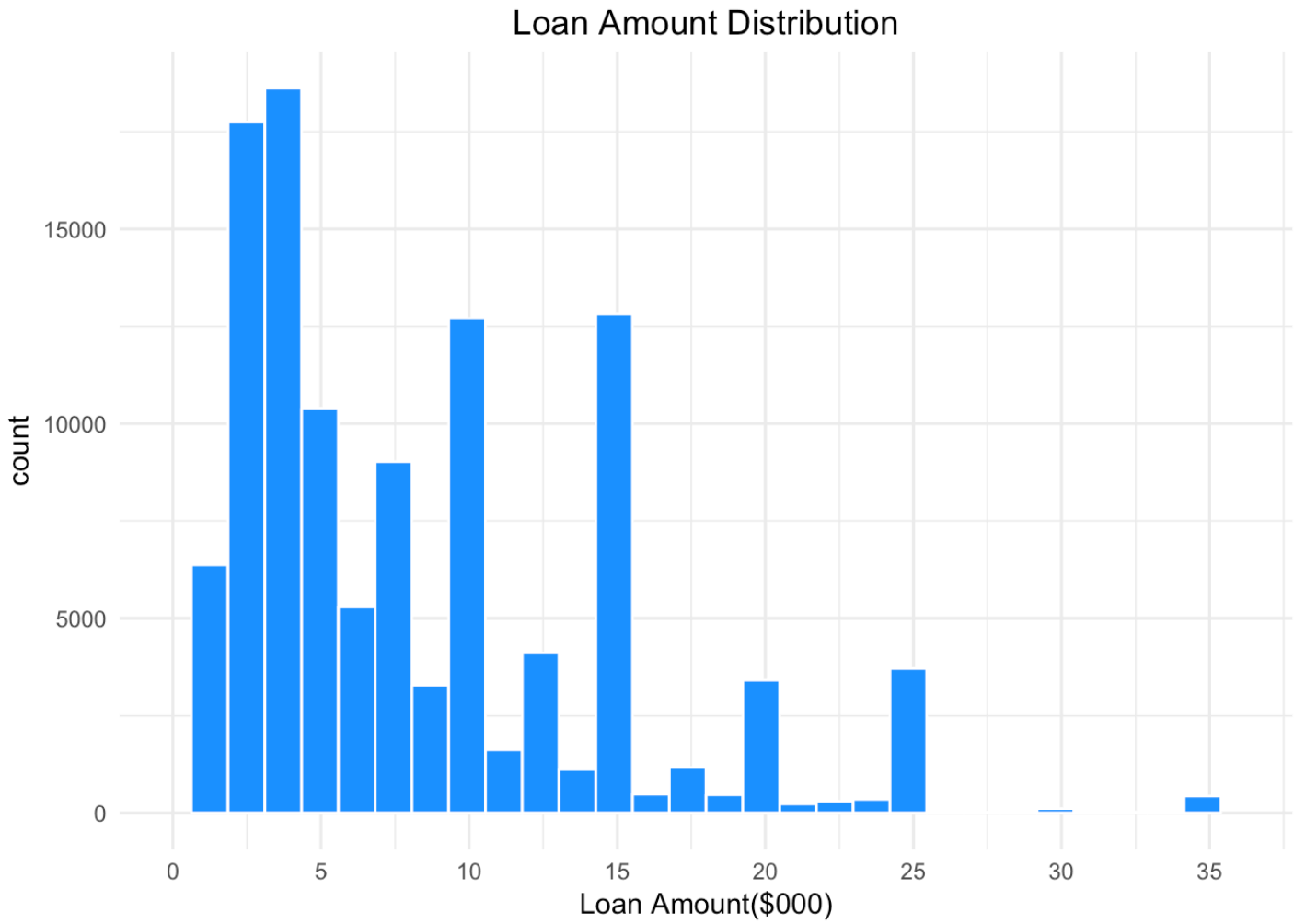
```
ggplot(data=subset(prosper, ListingCategory != "Not available"), aes(x= ListingCat
egory, fill=factor(Term))) +
  geom_bar() +
  theme_minimal() +
  theme(plot.title = element_text(hjust = .5)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 1, hjust = 1)) +
  labs(title="Borrower's Term Loan by Date", x="Listing Category")
```



Most borrowers chose 36 months to pay off their Prosper loans and 60 months being the second chosen term. A small number opted for 12 months.

How much do borrowers request for?

```
# I divided the Loan Amount by 1,000 in order to make it easier for the readers.
ggplot(data=prosper, aes(x=LoanOriginalAmount/1000)) +
  geom_histogram(color=I("white"), fill=I("dodgerblue")) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = .5)) +
  labs(title="Loan Amount Distribution", x="Loan Amount($000)") +
  scale_x_continuous(limits=c(0,36), breaks = seq(0,36,5))
```



The distribution for the loan is positively skewed and between $1,000 and $35,000 ($35,000 being an outlier). A majority of borrowers asked for loans below $15,000 and this signifies that most people are only using the loans to pay for small debts.

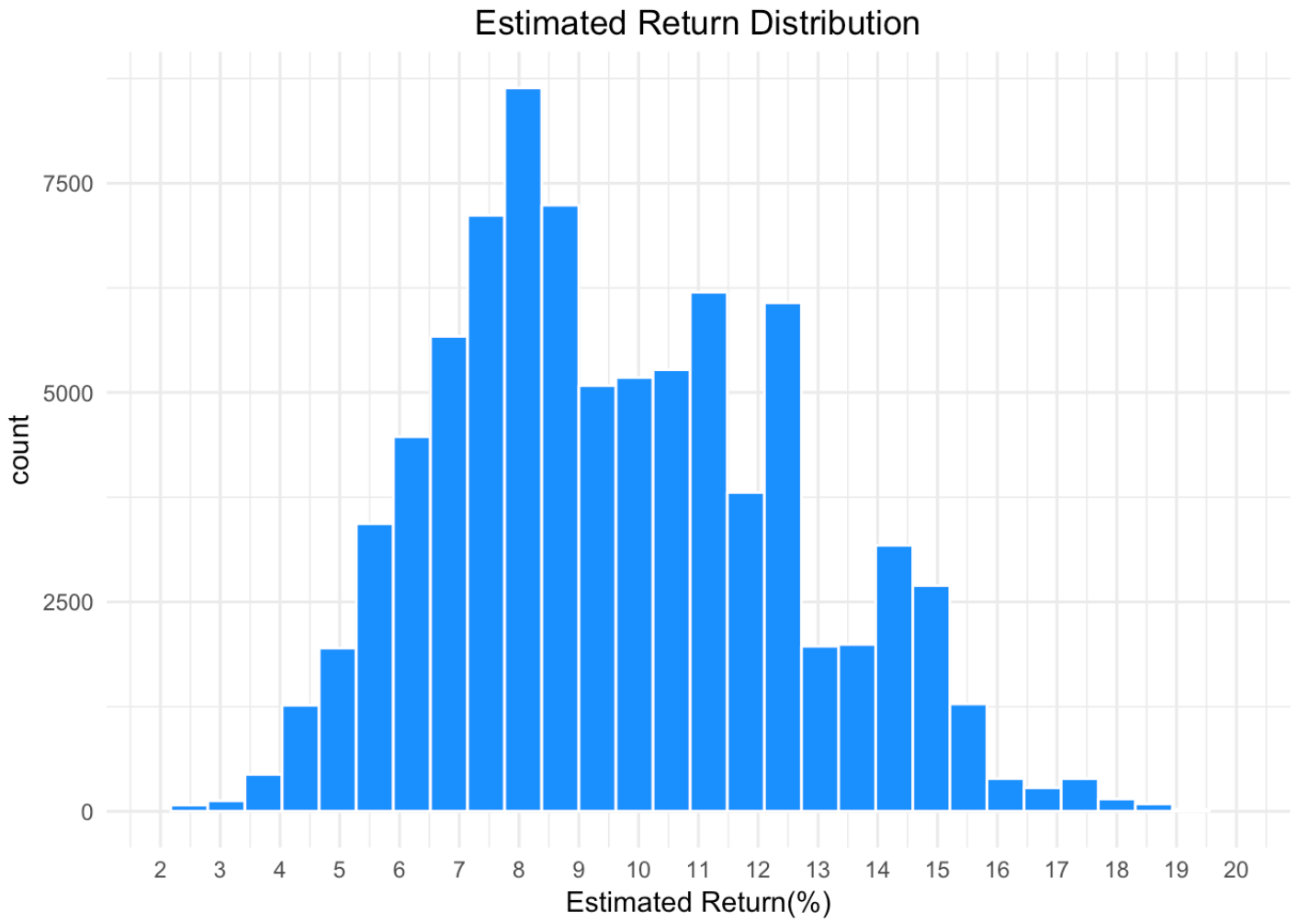
```
summary(prosper$LoanOriginalAmount)
```

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

On average people borrow $8,337. The minimum amount is $1,000 and maximum $35,000.

What is the distribution for Estimated Return?

```
# Multiplied Estimated Return by 100 to turn it into a percentage.
ggplot(data=prosper, aes(x= EstimatedReturn*100)) +
  geom_histogram(color=I("white"), fill=I("dodgerblue")) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = .5)) +
  labs(title="Estimated Return Distribution", x="Estimated Return(%)") +
  scale_x_continuous(limits=c(2,20), breaks = seq(2,20, 1))
```



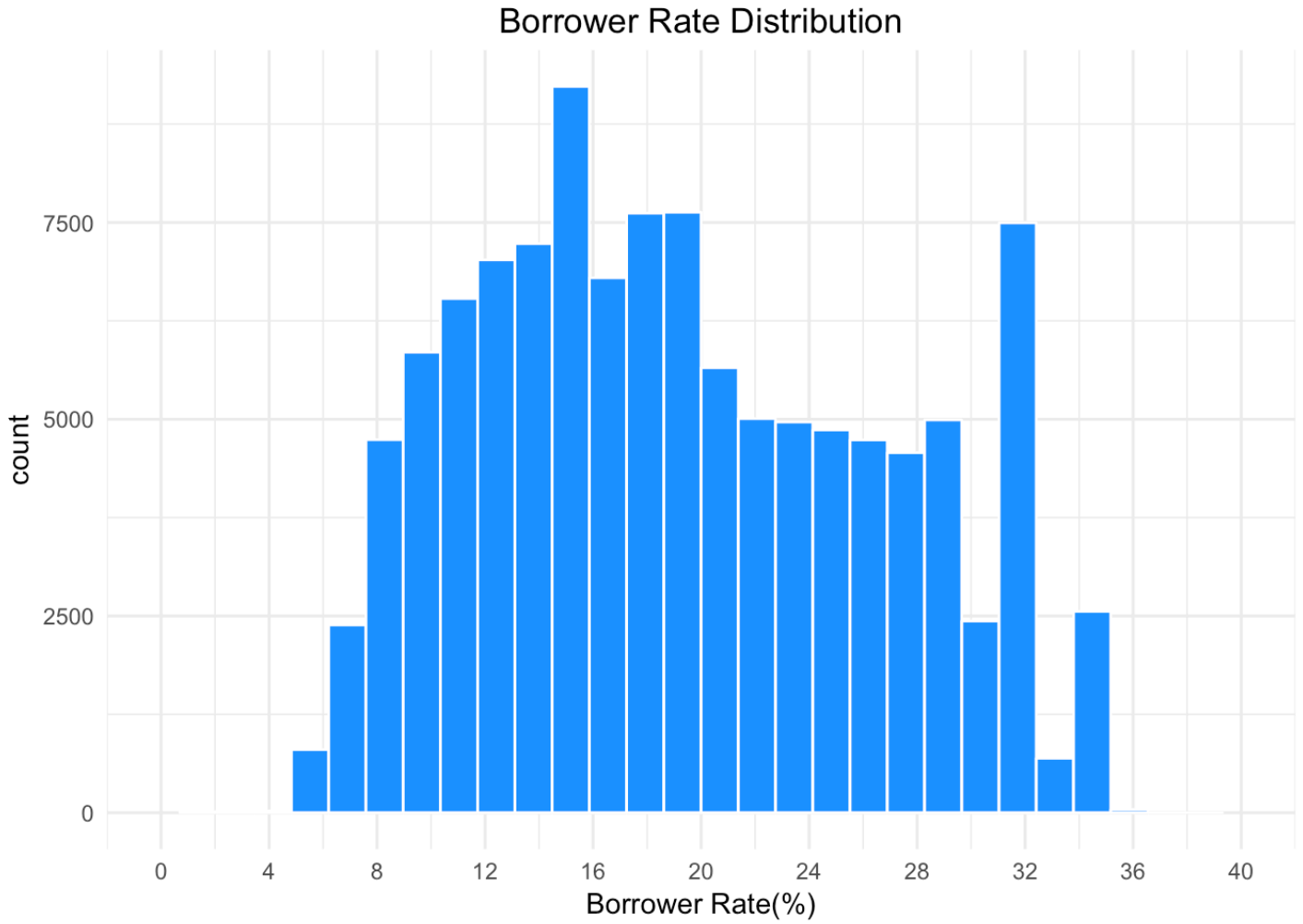

```
summary(prosper$EstimatedReturn)
```

```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  -0.183   0.074   0.092   0.096   0.117   0.284   29084
```

The chart for Estimated Return is normally distributed and it is between 2% to 19%. 2% being the lowest in making a return.

What is the distribution for Borrower Rate?

```
ggplot(data=prosper, aes(x= BorrowerRate*100)) +
  geom_histogram(color=I("white"), fill=I("dodgerblue")) +
  theme_minimal() +
  scale_x_continuous(limits=c(0,40), breaks=seq(0,40,4)) +
  theme(plot.title = element_text(hjust = .5)) +
  labs(title="Borrower Rate Distribution", x="Borrower Rate(%)")
```



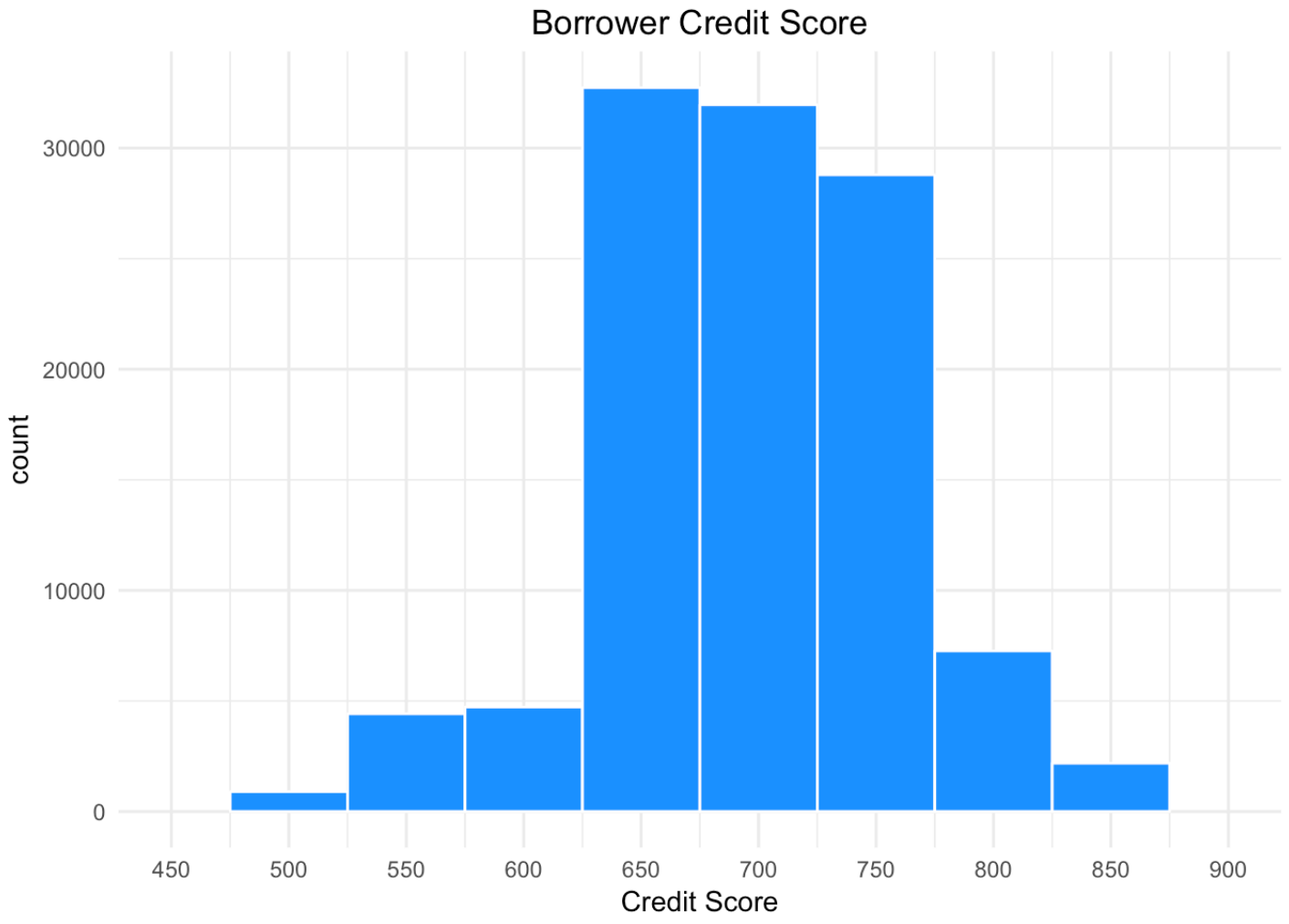

```
summary(prosper$BorrowerRate)
```

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

The Borrowers pay interest rates with mean of 19.28% to a maximum of 49.75% and with a spike at 16%. The interest rates seem very high in terms of borrowing, however, referencing back to the previous chart, one of the main reasons for the loan requests was “Debt Consolidation”. This means that borrowers are paying high interest mainly to pay off their bills as credit card companies may charge an even higher rate.

What is the distribution for Credit Score?

```
ggplot(data=prosper, aes(x=CreditScore)) +
  geom_histogram(binwidth=50,color=I("white"), fill=I("dodgerblue")) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = .5)) +
  labs(title="Borrower Credit Score", x="Credit Score") +
  scale_x_continuous(limits = c(450,900), breaks = seq(450,900,50))
```



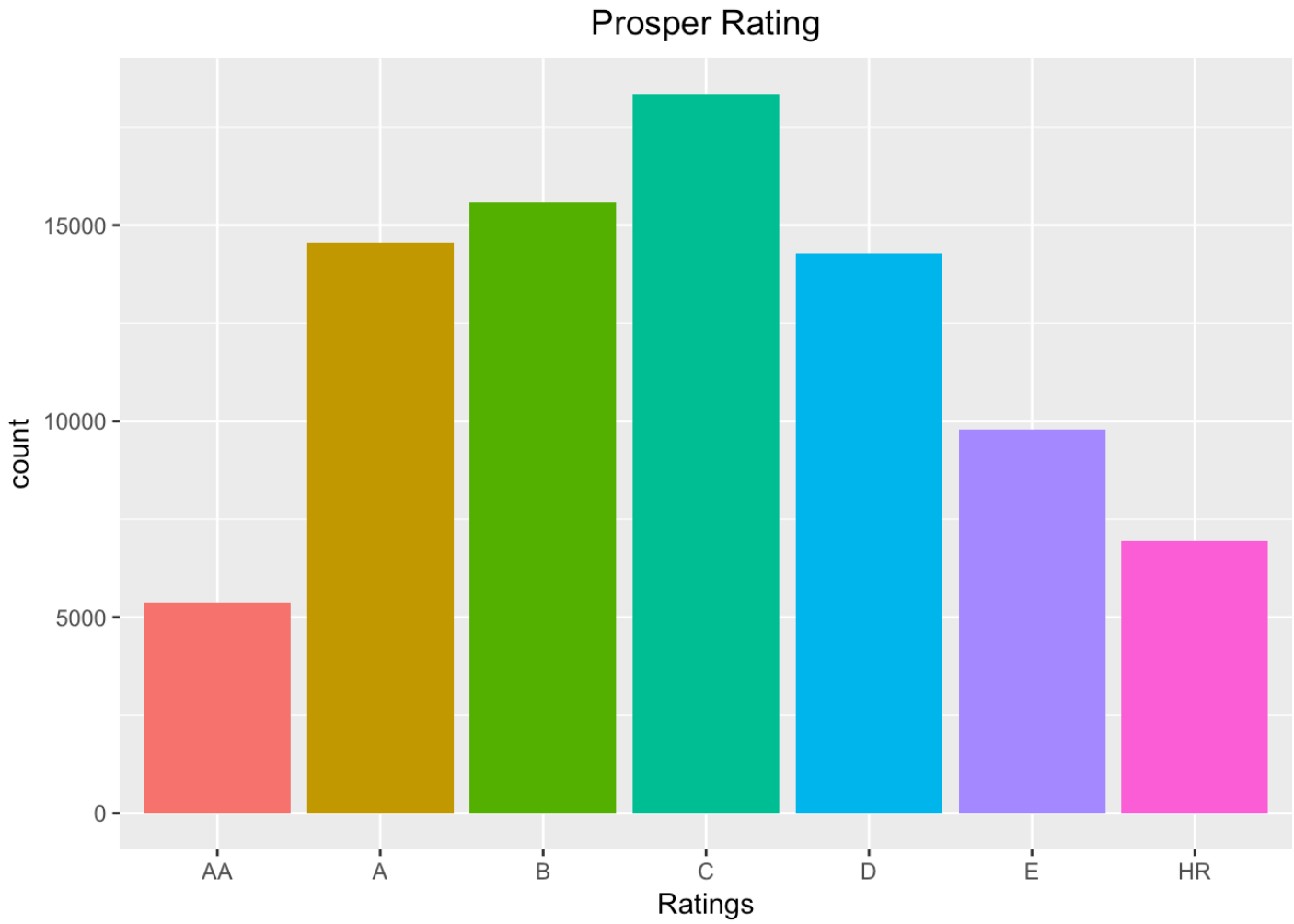

```
summary(prosper$CreditScore)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     9.5   669.5   689.5   695.1   729.5   889.5     591
```

From this distribution, a majority of borrowers have credit between 650 to 750. Although a number of borrowers have credit score below 600 and above 800.

What is the distibution for Prosper Rating?

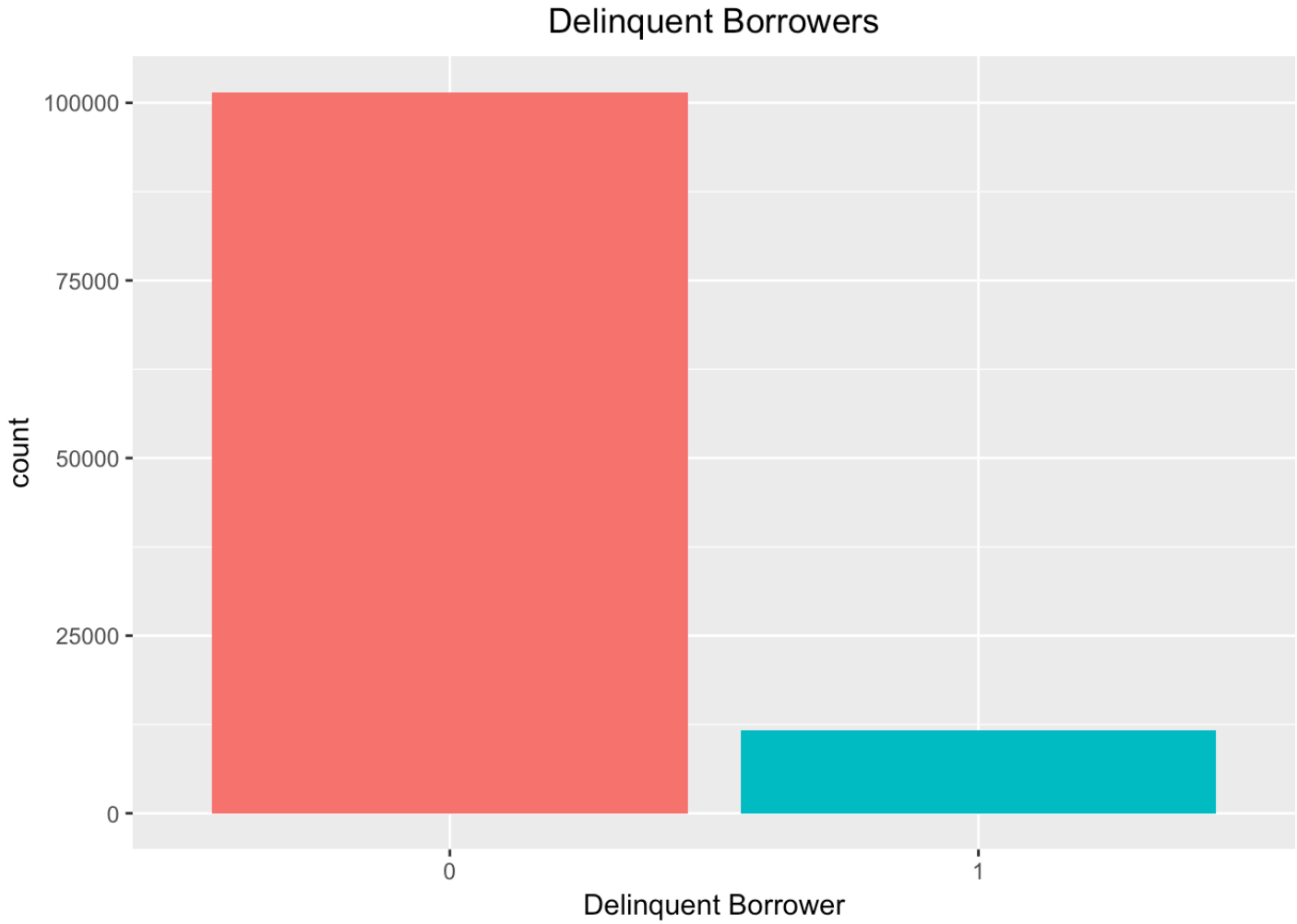
```
ggplot(data=subset(prosper,!is.na(ProsperRating)), aes(x=ProsperRating)) +
  geom_bar(aes(fill=factor(ProsperRating))) +
  theme(plot.title = element_text(hjust = .5)) +
  labs(title="Prosper Rating", x="Ratings") +
  guides(fill=FALSE)
```



The bar graph is normally distributed, and it shows that Prosper has 7 different ratings, AA being the lowest risk and HR the highest. Borrowers mostly had a Prosper rating of “C” and “AA” being the least. This is a proprietary rating system which is based upon the estimation of the borrower’s estimated loss rate which was introduced after the SEC registration

How many Delinquent Borrowers are there?

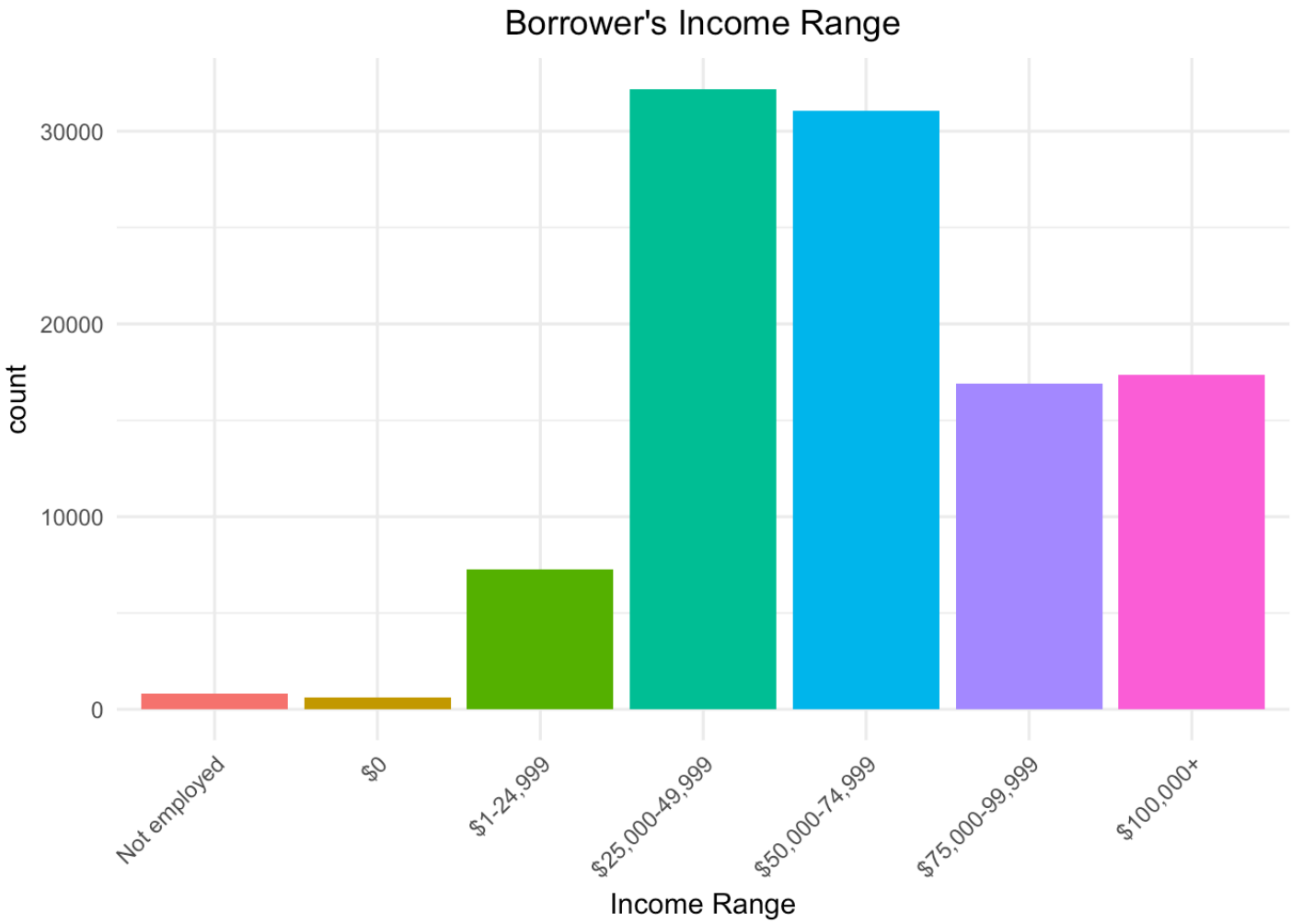
```
ggplot(data=subset(prosper,!is.na(DelinquentBorrower)), aes(x=DelinquentBorrower))
+
  geom_bar(aes(factor(DelinquentBorrower), fill=factor(DelinquentBorrower))) +
  theme(plot.title = element_text(hjust = .5)) +
  labs(title="Delinquent Borrowers", x="Delinquent Borrower", color="Delinquent Bo
rrower") +
  guides(fill=FALSE)
```



Here we can see that there are less delinquent borrowers than those who are not.

What are borrowers' income?

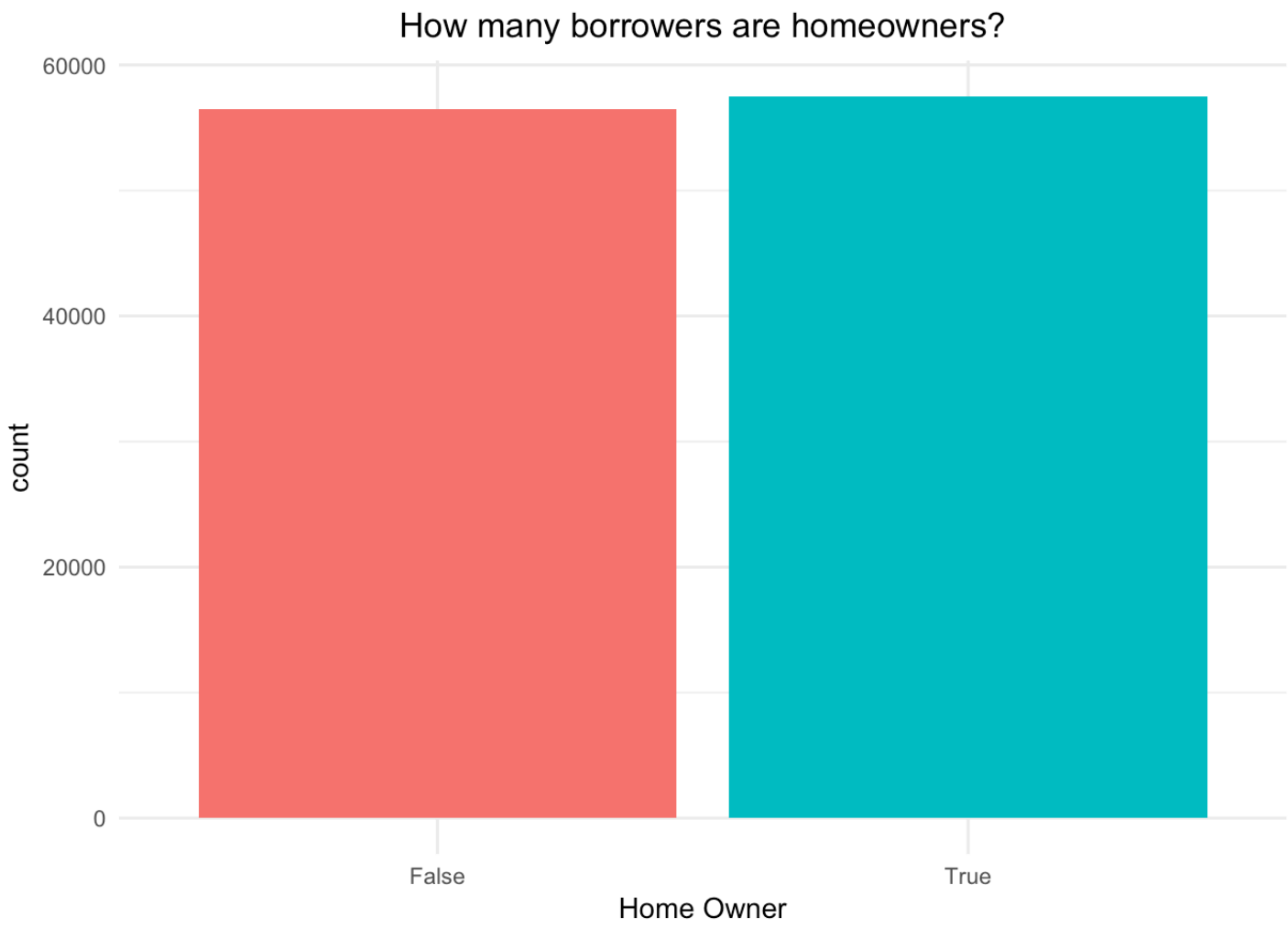
```
# Removed "Not displayed" category by subsetting to make analysis easier to unders
tand.
ggplot(data=subset(prosper, prosper$IncomeRange != "Not displayed"), aes(x= Income
Range)) +
  geom_bar(aes(fill=factor(IncomeRange))) +
  theme_minimal() +
  labs(title="Borrower's Income Range", x="Income Range")  +
  theme(plot.title = element_text(hjust = .5)) +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1)) +
  guides(fill=FALSE) +
  scale_x_discrete()
```



Here we see a distribution of borrower's income range. The vast majority of borrower's income is between $25,000 and $74,999. It can be seen that Prosper rarely approve loans to those who earn less than $25,000. Another surprising fact is that borrowers have an income of $0.

How many Borrowers are homeowners?

```
ggplot(data=prosper, aes(x=IsBorrowerHomeowner)) +
  geom_bar(aes(fill=IsBorrowerHomeowner)) +
  theme_minimal() +
  labs(title="How many borrowers are homeowners?", x="Home Owner")  +
  theme(plot.title = element_text(hjust = .5)) +
  guides(fill=FALSE)
```



Nearly half of the borrowers are homeowners.

Univariate Analysis

What is the structure of your dataset?

In this dataset, it contains 113,937 observations and 81 variables from the period of 2005 - 2014. However, I decided to only focus on 15 variables which are:

- Term
- LoanStatus
- BorrowerRate
- ProsperScore
- ListingCategory..numeric.
- BorrowerState
- EstimatedReturn
- CreditScoreRangeUpper
- CreditScoreRangeLower
- CurrentDelinquencies
- IncomeRange
- LoanOriginalAmount

- LoanOriginationQuarter
- ProsperRating..Alpha.
- IsBorrowerHomeowner

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

The main features of interest are Borrower Rate and Estimated Return

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

Term, Loan Status, ProsperScore, ListingCategory..numeric., CreditScoreRangeLower, CreditScoreRangeUpper, CurrentDelinquencies, IncomeRange, LoanOriginalAmount, LoanOriginationQuarter, ProsperRating..Alpha., IsBorrowerHomeowner.

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

I converted ListingCategory..numeric. from numeric variable into a factor variable using the category names that were shown in the spreadsheet. Moreover, I created a new categorical variable “DelinquentBorrower from CurrentDelinquencies whether or not there are any delinquents. Lastly, I created another variable to calculate the mean from the two variables (CreditScoreRangeLower, CreditScoreRangeUpper)

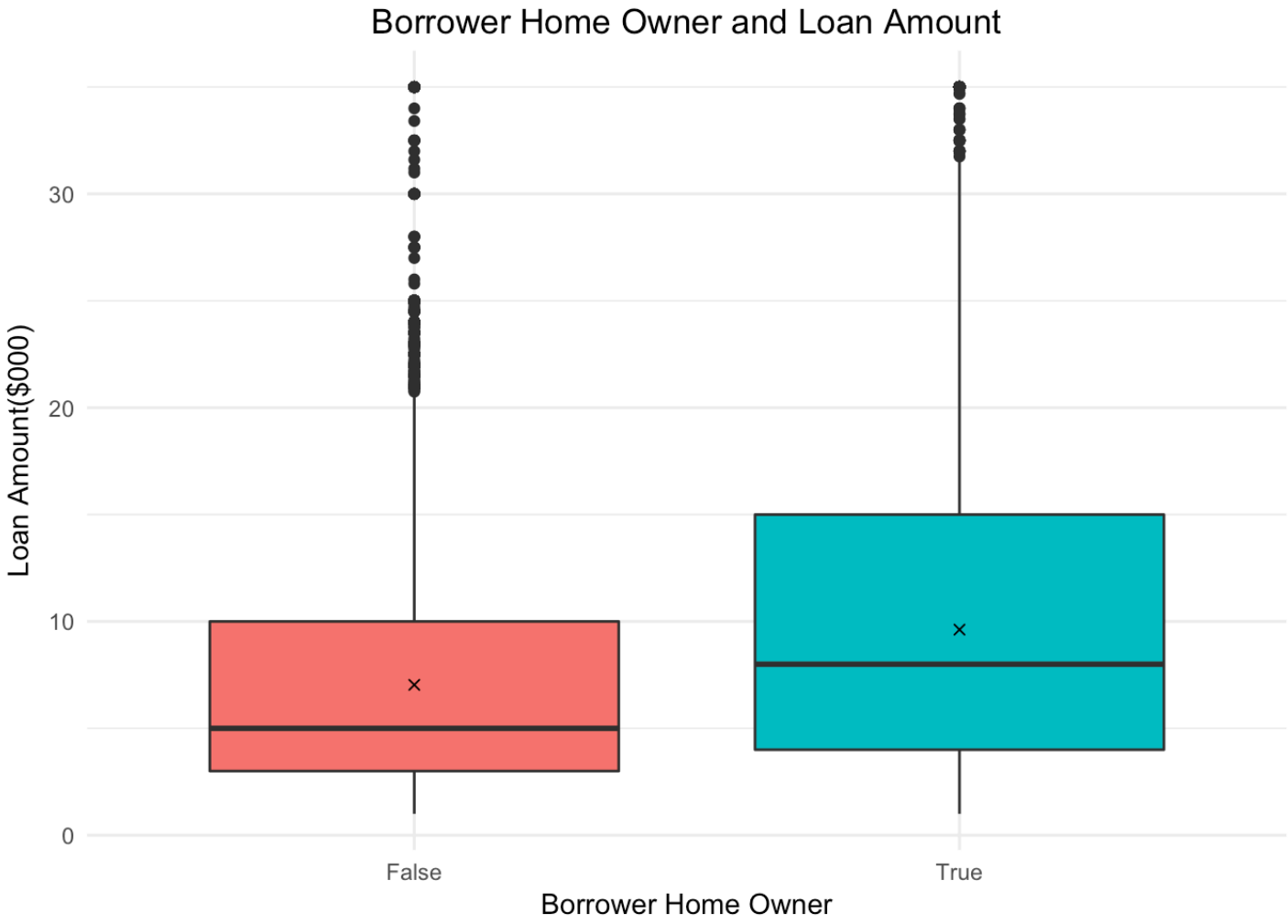
Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

There are a few factor variables which are not in the correct order so therefore, I had to sort them in order. For example, I had to sort LoanOriginationQuarter from the earliest year and quarter to the latest. I also had to sort CreditScore and ProsperRating from the lowest risk to the highest. In addition to Income Range, I sorted the variable from the lowest to the highest salary earned.

Bivariate Plots Section

What is the relationship between Borrower HomeOwner and Loan Amount?

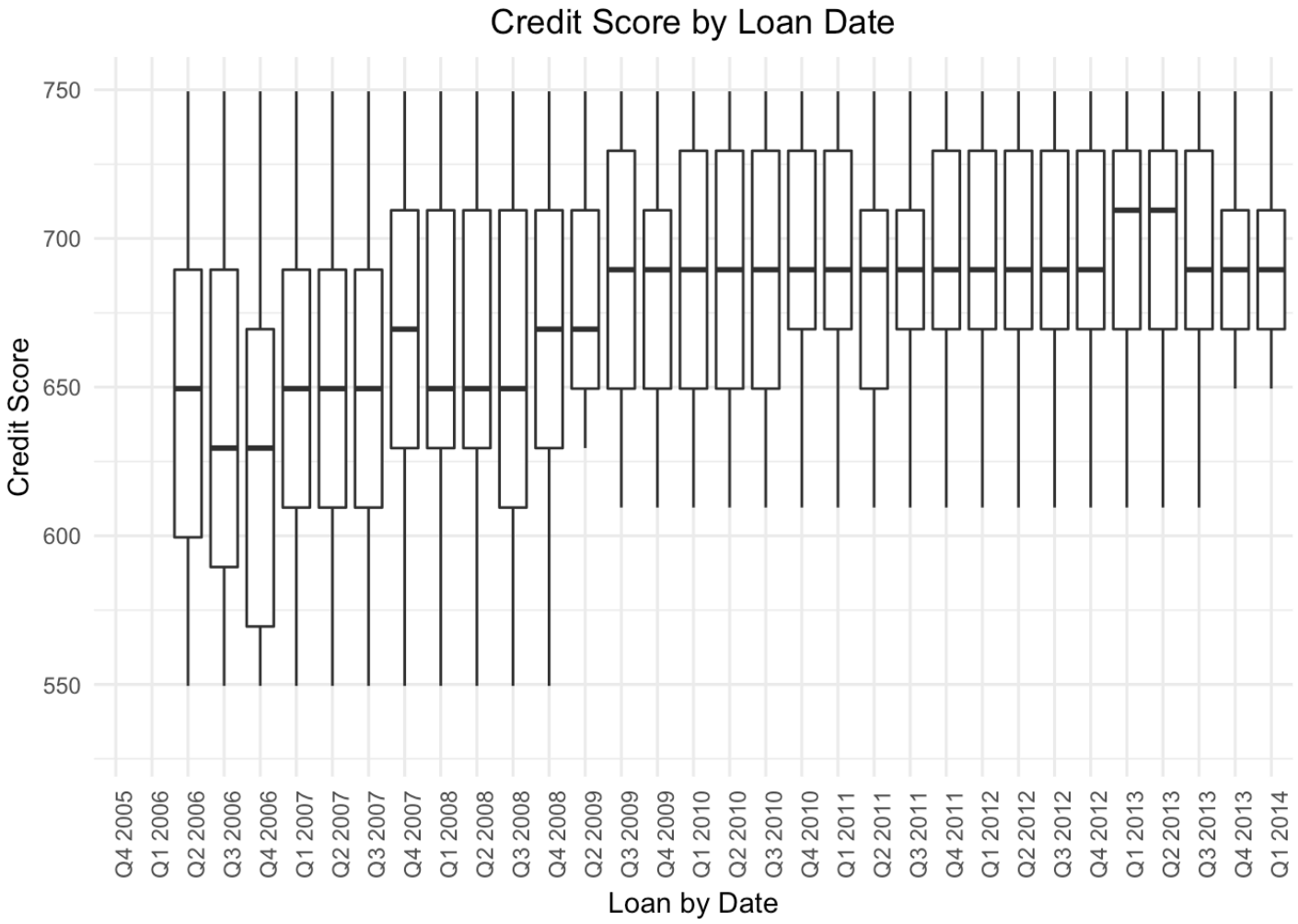
```
ggplot(data=prosper, aes(x=IsBorrowerHomeowner, y=LoanOriginalAmount/1000)) +
  geom_boxplot(aes(fill=IsBorrowerHomeowner)) +
  stat_summary(fun.y = mean, geom = "point", shape =4) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = .5)) +
  guides(fill=FALSE) +
  labs(title="Borrower Home Owner and Loan Amount",x="Borrower Home Owner", y="Loa
n Amount($000)")
```



The mean (labelled as 'x') for Loan Amount to those who are Homeowners is $10,000 and median is $8,000. Borrowers who owns a home tend to have a better chance of requesting larger sums of loan than those who do not.

What is the credit score throughout the quarter period?

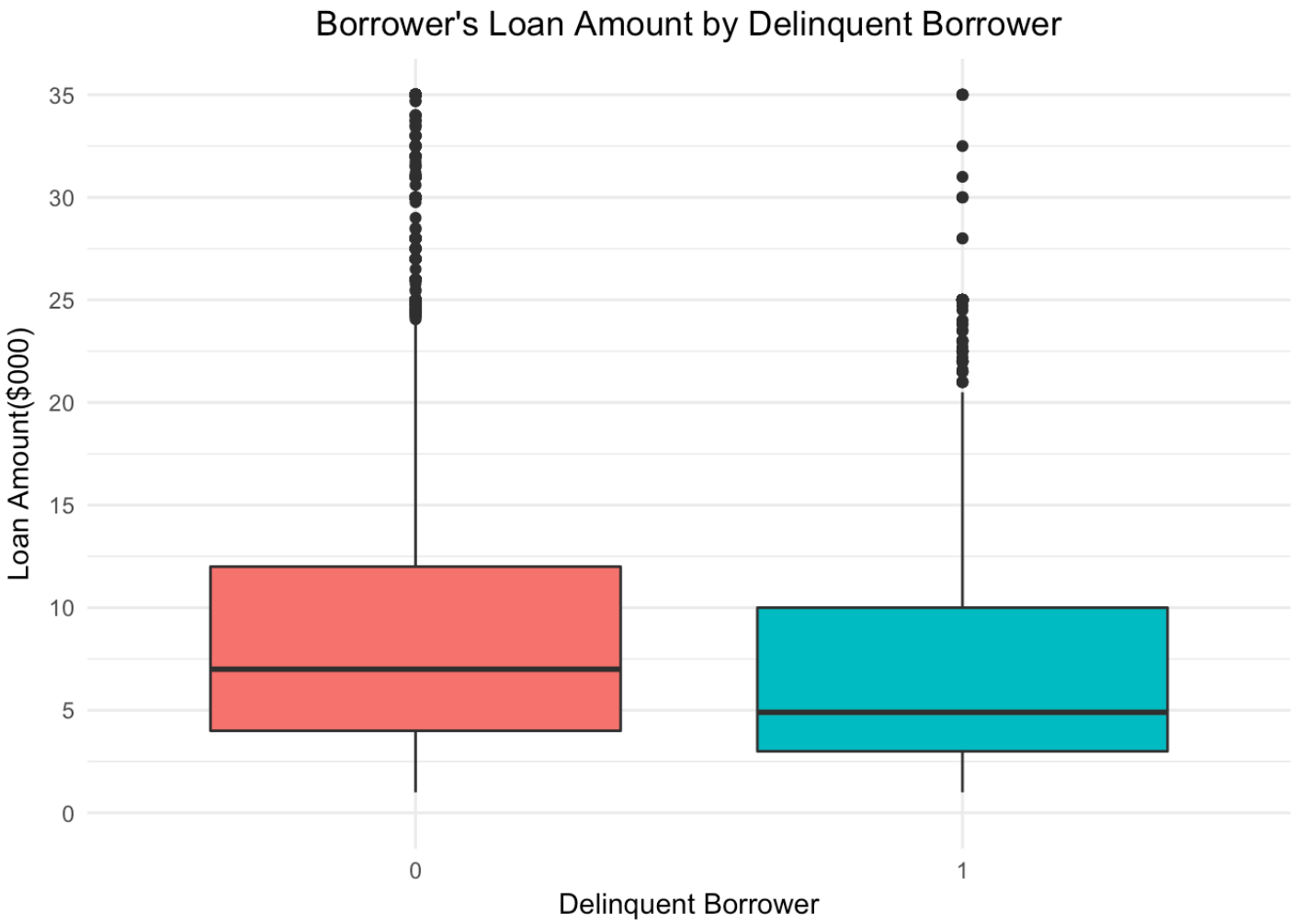
```
ggplot(data=prosper, aes(x=LoanOriginationQuarter, y=CreditScore)) +
  geom_boxplot() +
  theme_minimal() +
  theme(plot.title = element_text(hjust = .5)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 1, hjust = 1)) +
  guides(fill=FALSE) +
  labs(title="Credit Score by Loan Date",x="Loan by Date", y="Credit Score") +
  ylim(530,750)
```



This boxplot gives information about how Prosper changed its credit policies after obtaining their registration with SEC and the relaunch of their website in 2009. The credit score increased from the lowest limit of 520 (second quarter of 2006) to 630 (second quarter of 2009).

What is the relationship between Loan Amount and Delinquent Borrowers?

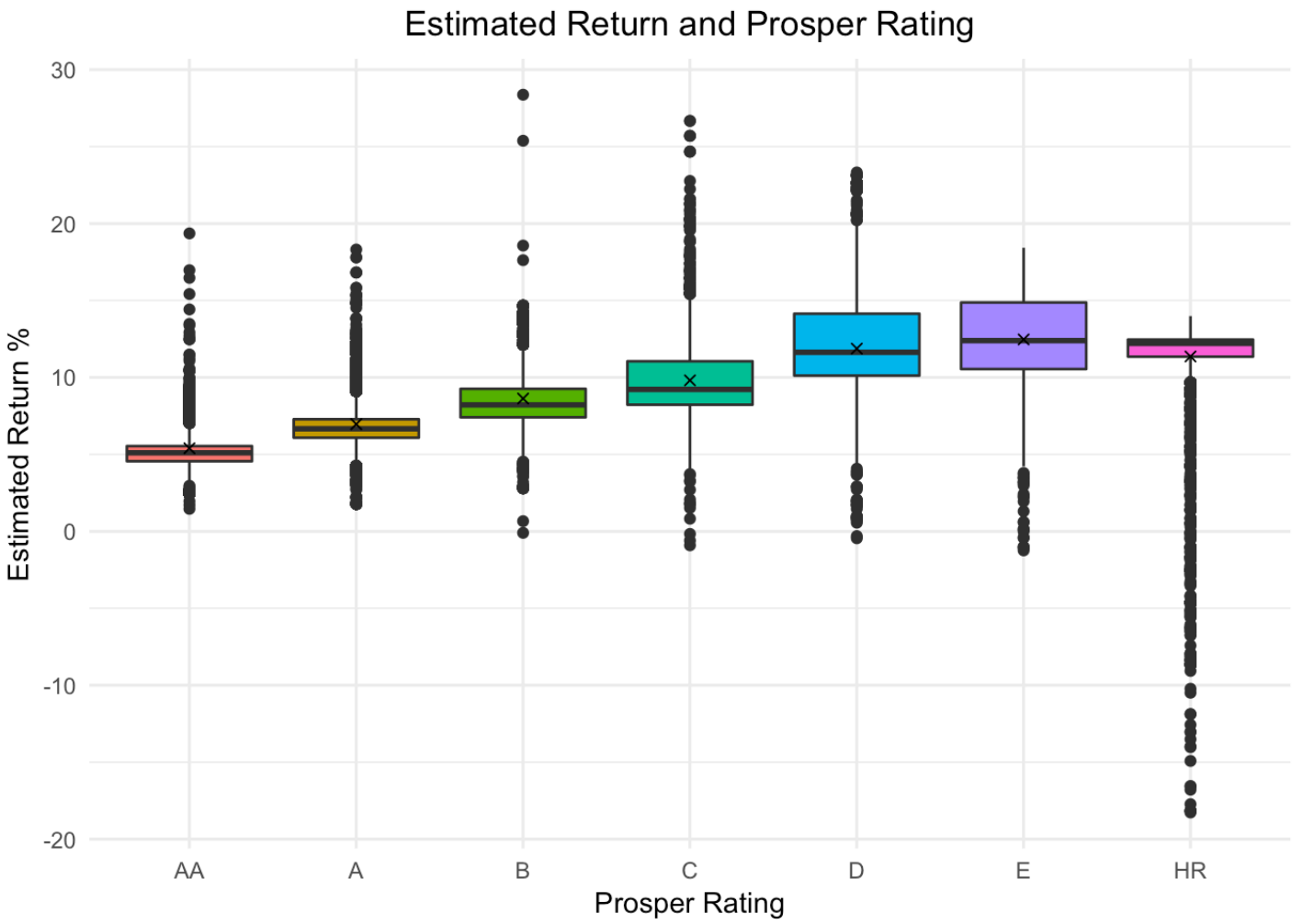
```
ggplot(data=subset(prosper,!is.na(DelinquentBorrower)), aes(x=DelinquentBorrower,
y=LoanOriginalAmount/1000)) +
  geom_boxplot(aes(factor(DelinquentBorrower), fill=factor(DelinquentBorrower))) +
  scale_y_continuous(limits = c(0,35), breaks = seq(0,35,5)) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = .5)) +
  guides(fill=FALSE) +
  labs(title="Borrower's Loan Amount by Delinquent Borrower", x="Delinquent Borrow
er", y="Loan Amount($000)")
```



We can see that the number of Delinquent borrowers is mainly those who borrowed less than $10,000 with the median at $5,000. Although there are outliers of people borrowing up to $35,000.

What is the relationship between Borrower Rate and Prosper Rating?

```
ggplot(data=subset(prosper, !is.na(ProsperRating)), aes(x=ProsperRating, y=Estimat
edReturn *100)) +
  geom_boxplot(aes(factor(ProsperRating), fill=factor(ProsperRating))) +
  stat_summary(fun.y = mean, geom = "point", shape =4) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = .5)) +
  guides(fill=FALSE) +
  labs(title="Estimated Return and Prosper Rating", x="Prosper Rating", y="Estimat
ed Return %")
```



For Prosper ratings of AA and A (lowest risk), the median is around 0.06 and there are outliers above where investors are able to gain returns with more than 10% and this is a good indicator for investment opportunities. For ratings E and HR, the median is around 13% which is higher than the other lower risk ratings. However, outliers for the rating of HR goes all of the way down to -18%.

What is the relationship between Credit Score Vs Prosper Rating and Prosper Score ?

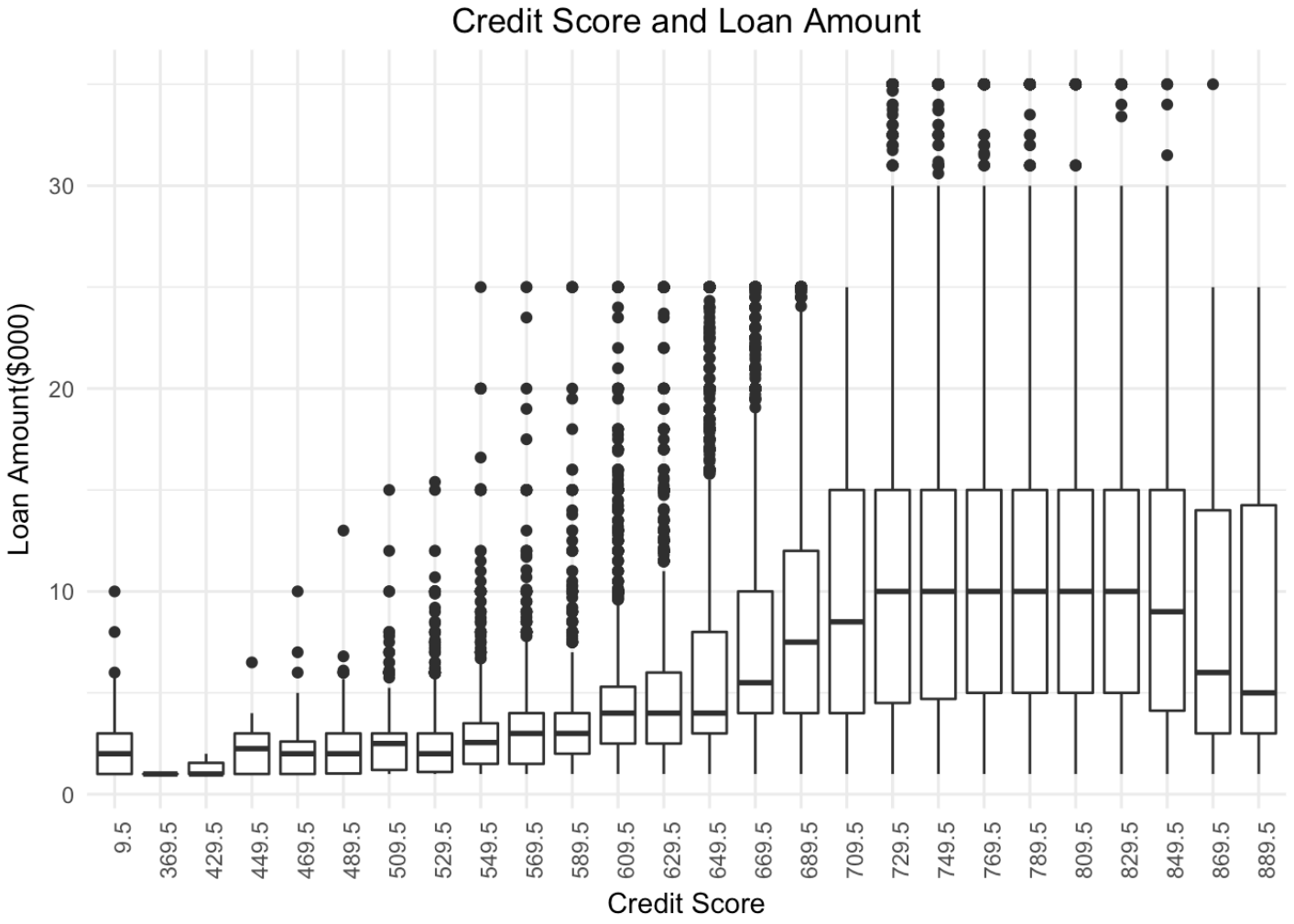
```
# I am using GridExtra to show multiple plots in one page to compare Credit scores with Prosper Rating and Prosper Score
p1<- ggplot(data=subset(prosper, !is.na(ProsperRating)), aes(x=ProsperRating, y=CreditScore)) +
  geom_boxplot(aes(factor(ProsperRating), fill=factor(ProsperRating))) +
  stat_summary(fun.y = mean, geom = "point", shape =4) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = .5)) +
  guides(fill=FALSE) +
  labs(title="Prosper Rating and Credit Score", x="Prosper Rating", y="Credit Score")

p2 <- ggplot(data=subset(prosper, !is.na(ProsperScore)), aes(x=ProsperScore, y=CreditScore)) +
  geom_boxplot(aes(factor(ProsperScore), fill=factor(ProsperScore))) +
  stat_summary(fun.y = mean, geom = "point", shape =4) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = .5)) +
  guides(fill=FALSE) +
  labs(title="Prosper Score and Credit Score", x="Prosper Score", y="Credit Score")

grid.arrange(p1, p2, nrow = 1)
```

The highest credit score is given “AA” with mean 780. Interestingly lowest rating “HR” with mean of 690 has a credit score higher than rating “E” with mean of 677.

With Prosper Score, the higher the credit score the higher Prosper score from 1 to 11. Prosper Score of 11 has a credit score with a mean of 730 (labelled as ‘x’) and for some reason Prosper Score of 10 has a higher mean (760) than Prosper Score 11.

What is the relationship between Credit Score and Loan Amount?

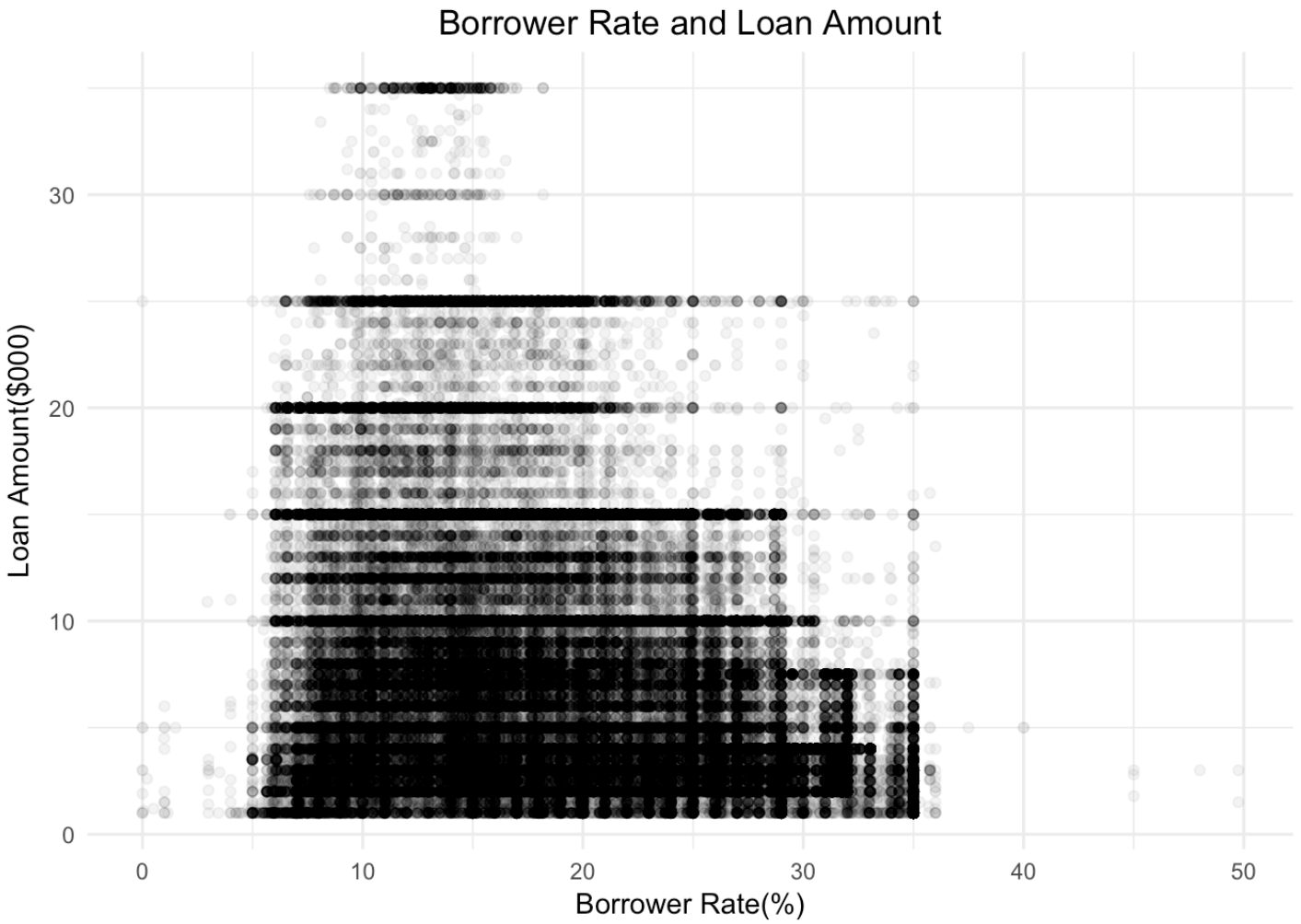
```
ggplot(data=subset(prosper, !is.na(CreditScore)), aes(x=CreditScore, y=LoanOrigina
lAmount/1000)) +
  geom_boxplot(aes(factor(CreditScore))) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = .5)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 1, hjust = 1)) +
  guides(fill=FALSE) +
  labs(title="Credit Score and Loan Amount", x="Credit Score", y="Loan Amount($000
)")
```



We can see that the higher loan amount was mostly approved based on their credit score so the higher the score the more you can borrow. Interestingly, there are outliers where borrowers with low credit scores were able to apply for larger sums especially those with a credit score of less than 600. This could be due to the fact that borrowers must have applied for loans with Prosper in the past, so therefore they have a good business relationship.

What is the relationship between Borrower Rate and Loan Amount?

```
ggplot(data=prosper, aes(x=BorrowerRate*100, y=LoanOriginalAmount/1000)) +
  geom_jitter(alpha=0.05, position = position_jitter(h=0)) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = .5)) +
  guides(fill=FALSE) +
  labs(title="Borrower Rate and Loan Amount", x="Borrower Rate(%)", y="Loan Amount
($000)")
```




```
cor(prosper$BorrowerRate, prosper$LoanOriginalAmount)
```

```
## [1] -0.3289599
```

The two variables have a negative relationship with a correlaton of -33%. This is because in the plot, it shows that interest rates are more correlated as loan amount increases which could be due to the borrower's good financial background.

What is the relationship between Borrower Rate and Estimated Return?

```
ggplot(data=prosper, aes(x=BorrowerRate*100, y=EstimatedReturn*100)) +
  geom_jitter(alpha=0.05) +
  scale_x_continuous(limits = c(5,38), breaks = seq(5,38,10)) +
  scale_y_continuous(limits=c(0,25), breaks = seq(0,25,10)) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = .5)) +
  labs(title="Estimated Return and Borrower Rate", x="Borrower Rate(%)", y="Estima
ted Return(%)") +
  geom_smooth(method="lm", color="red")
```

According to this scatter plot, we can determine that Estimated Return and Borrower Rate have a strong and positive correlation so the higher the interest rate the better the return for Prosper investors.

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?

- After the cease and desist order by SEC in 2008 due to Prosper's violation of the Securities Act for its Peer-to-peer model. The number of listings plummeted dramatically and so Prosper require borrowers to have a higher credit score. Prosper had been getting better growth after the relaunch of their website.
- Delinquent Borrowers tend to be those that take out smaller loans (less than $10,000).
- There is certainly a strong positive correlation between "Borrower Rate"" and "Estimated Return" so the higher the interest the higher the return. However, the relationship between "Borrower Rate" and "Loan Amount" is negative according to the plot, the higher the loan amount the lower the interest rate which could be due to the borrower's high credit score.

- Another strong relationship between Prosper Rating and Borrower Rate as the higher the risk the higher the interest rate, but in another plot the higher the loan amount the less risky which is due to the borrower's high credit score as mentioned above.

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

Borrowers with a Prosper Rating of "HR" tend to have higher credit scores than those with a Prosper Rating of "E".

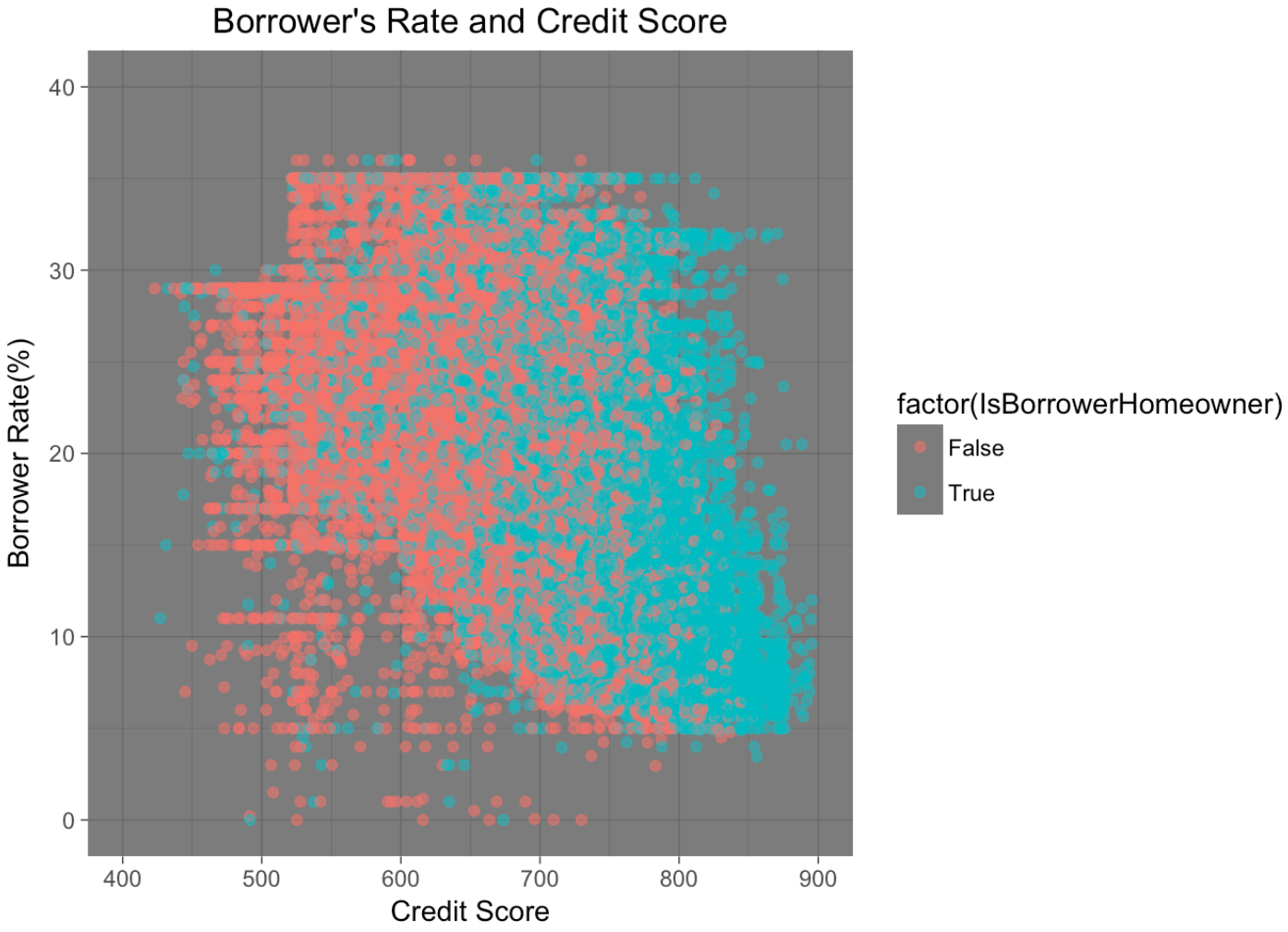
What was the strongest relationship you found?

The strongest relationship would be Estimated Return & Borrower Rate as well as a strong relationship between Loan Amount & Credit Score as Credit Score is a indicator for borrower's credit worthiness it determines how much you can borrow.

Multivariate Plots Section

What is the relationship between Borrower's Rate and Credit Score by Delinquent Borrowers?

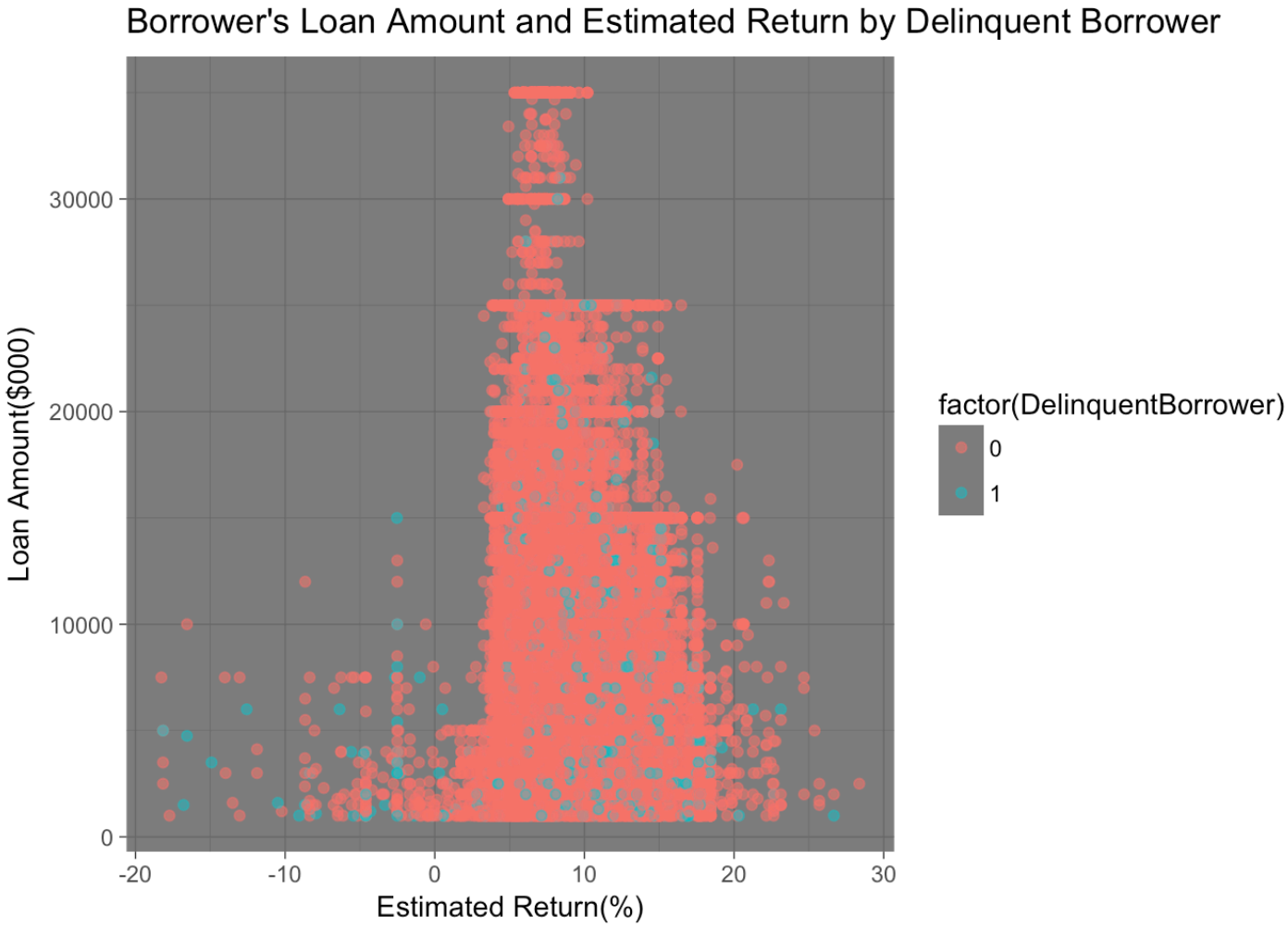
```
ggplot(data=subset(prosper,!is.na(IsBorrowerHomeowner)), aes(x=CreditScore, y=Borr
owerRate*100, color=factor(IsBorrowerHomeowner))) +
  geom_point(alpha=0.5, position = "jitter") +
  scale_x_continuous(limits=c(400,900)) +
  scale_y_continuous(limits=c(0,40)) +
  theme_minimal() +
  theme_dark() +
  theme(plot.title = element_text(hjust = .5)) +
  labs(title="Borrower's Rate and Credit Score", x="Credit Score", y="Borrower Rat
e(%)")
```



We can see that Homeowners were mainly those with higher credit scores. As the credit score increases, the interest rate decreases.

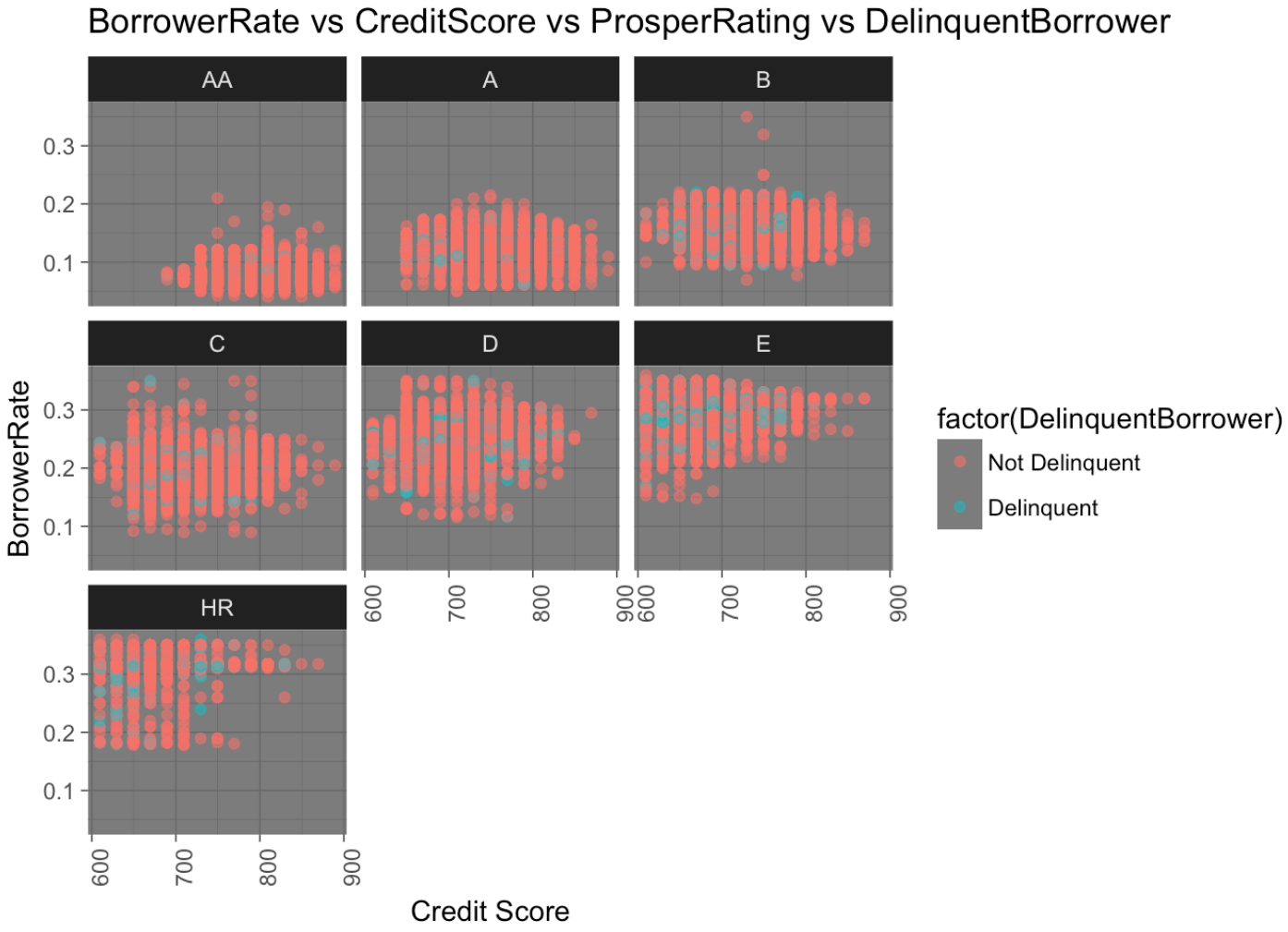
What is the relationship between Estimated Return and Loan Amount by Delinquent Borrower?

```
ggplot(data=subset(prosper,!is.na(DelinquentBorrower)), aes(x=EstimatedReturn*100,
y=LoanOriginalAmount, color=factor(DelinquentBorrower))) +
  geom_point(alpha=1/2) +
  scale_x_continuous() +
  theme_minimal() +
  theme_dark() +
  labs(title="Borrower's Loan Amount and Estimated Return by Delinquent Borrower",
x="Estimated Return(%)", y="Loan Amount($000)")
```



Most Delinquent Borrowers in green are mostly scattered below the loan amount less than $10,000 although there are outliers that are above as mentioned in the previous plot where the outliers go up to $35,000. Estimated Return for Delinquent Borrower is between -20% to 27%.

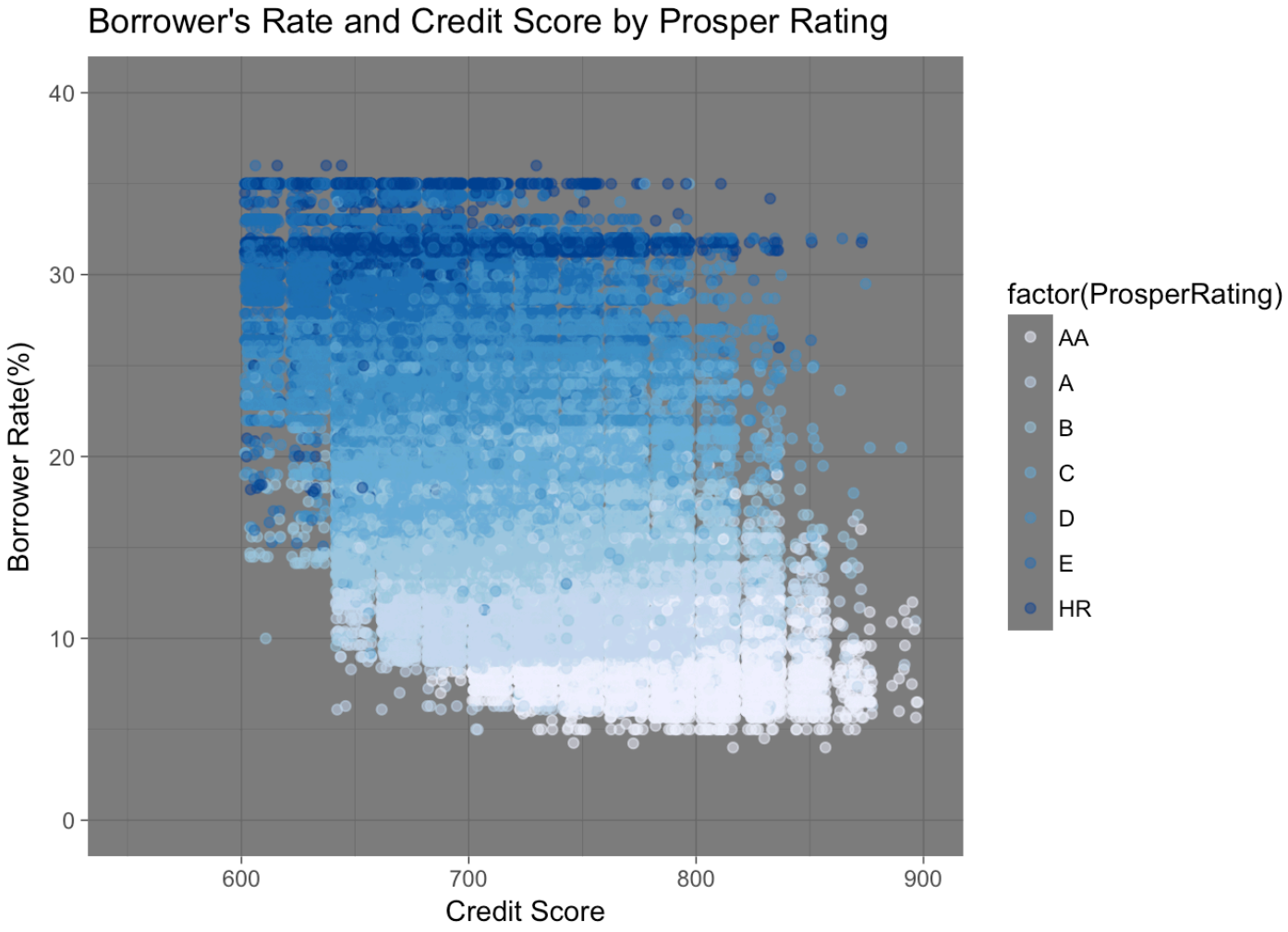
```
ggplot(data=subset(prosper, !is.na(DelinquentBorrower) & !is.na(ProsperRating)), a
es(x= CreditScore, y=BorrowerRate, color=factor(DelinquentBorrower))) +
  geom_point(alpha=1/2) +
  facet_wrap(~ProsperRating) +
  scale_color_discrete(labels= c("Not Delinquent", "Delinquent")) +
  theme_minimal() +
  theme_dark() +
  labs(title="BorrowerRate vs CreditScore vs ProsperRating vs DelinquentBorrower",
x="Credit Score", y="BorrowerRate") +
  theme(axis.text.x = element_text(angle = 90, vjust = 1, hjust = 1))
```



As the Prosper Rating goes down (from the highest rating of “AA” to the lowest rating of “HR”) the number of Delinquent Borrowers slowly increases as well as the interest rates and at the same time the credit score goes down.

What is the relationship between Borrower’s Rate and Credit Score by Prosper Rating?

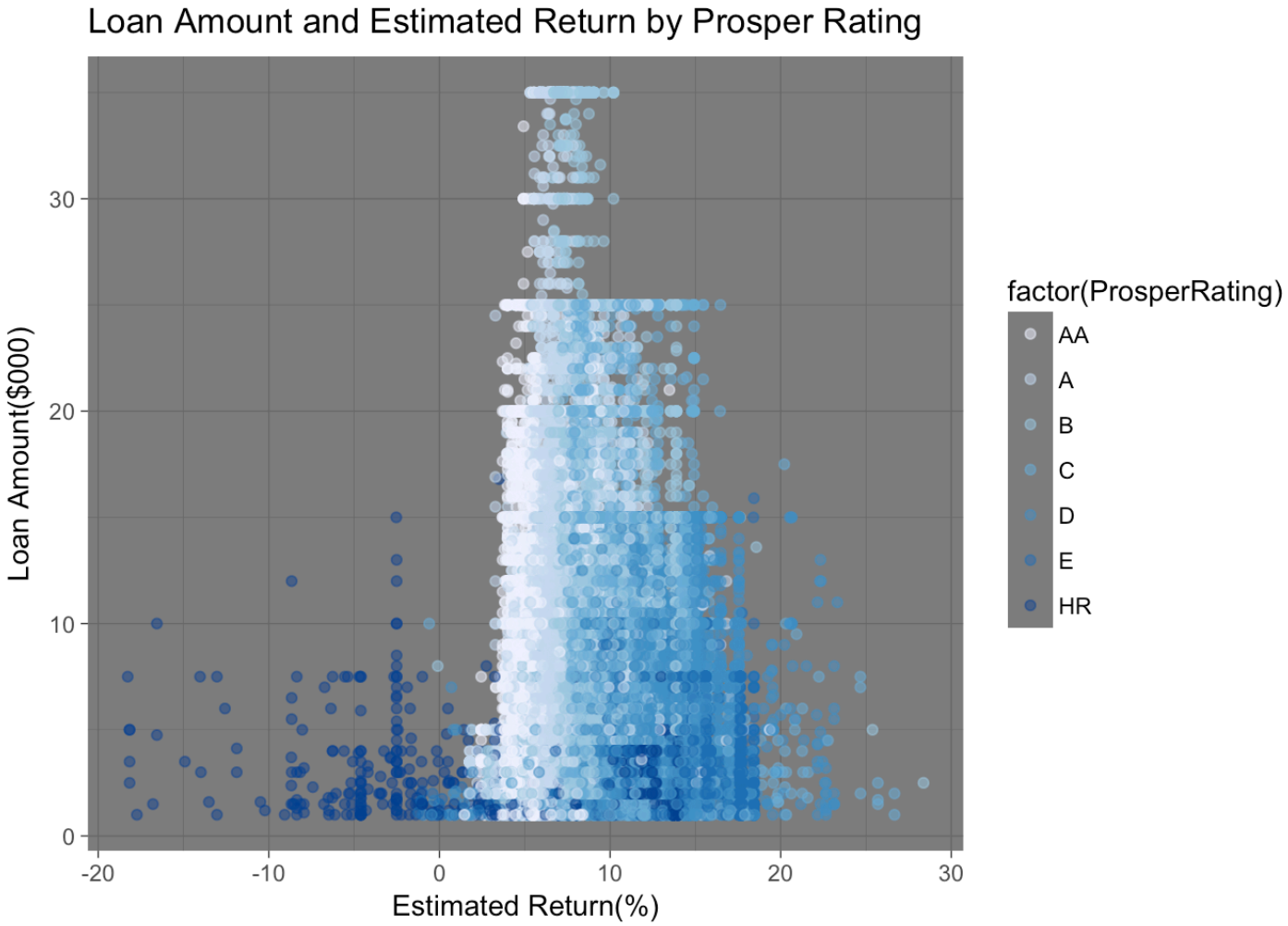
```
ggplot(data= subset(prosper,!is.na(ProsperRating)), aes(x=CreditScore, y=BorrowerR
ate*100, color=factor(ProsperRating))) +
  geom_point(alpha=1/2, position = "jitter") +
  scale_color_brewer(type = "seq", palette = 1, direction = 1) +
  scale_x_continuous(limits=c(550,900)) +
  scale_y_continuous(limits=c(0,40)) +
  theme_minimal() +
  theme_dark() +
  labs(title="Borrower's Rate and Credit Score by Prosper Rating", x="Credit Score
", y="Borrower Rate(%)")
```



Looking at this plot the lower the credit score and Prosper Rating, the higher the interest rate, therefore, the riskier it gets. It can be seen that borrowers with a credit score below 600 were mainly from before the cease order from SEC after questioning about their P2P model. Even though a small number of borrowers managed to get a low interest rate with a low credit score but Prosper have changed its credit policies and made the rules stricter.

What is the relationship between Estimated Return and Loan Amount by Prosper Rating?

```
ggplot(data=subset(prosper,!is.na(ProsperRating)), aes(x=EstimatedReturn*100, y=Lo
anOriginalAmount/1000, color=factor(ProsperRating))) +
  geom_point(alpha=1/2) +
  scale_color_brewer(type = "seq", palette = 1, direction = 1) +
  scale_x_continuous() +
  theme_minimal() +
  theme_dark() +
  labs(title="Loan Amount and Estimated Return by Prosper Rating", x="Estimated Re
turn(%)", y="Loan Amount($000)")
```



Here we can see that borrowers who applied for a lower amount of loan offer Prosper a small return and thereby higher risk. For instance, those who have a Prosper Rating of “HR”" and borrowed less than $15,000 the estimated return is less than 20% and the lowest being -20%. With high prosper rating borrowers can get larger loans and if it is lower than a “C” the loan will become less.

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?

When investigating on “Loan amount” and “Estimated Return”, as the loan amount increases the “Prosper Rating” becomes less risky(Prosper Rating AA being the least risky and HR the riskiest) and at the same time “Estimated Return”" increases this could be due to the borrower’s high credit score.

Were there any interesting or surprising interactions between features?

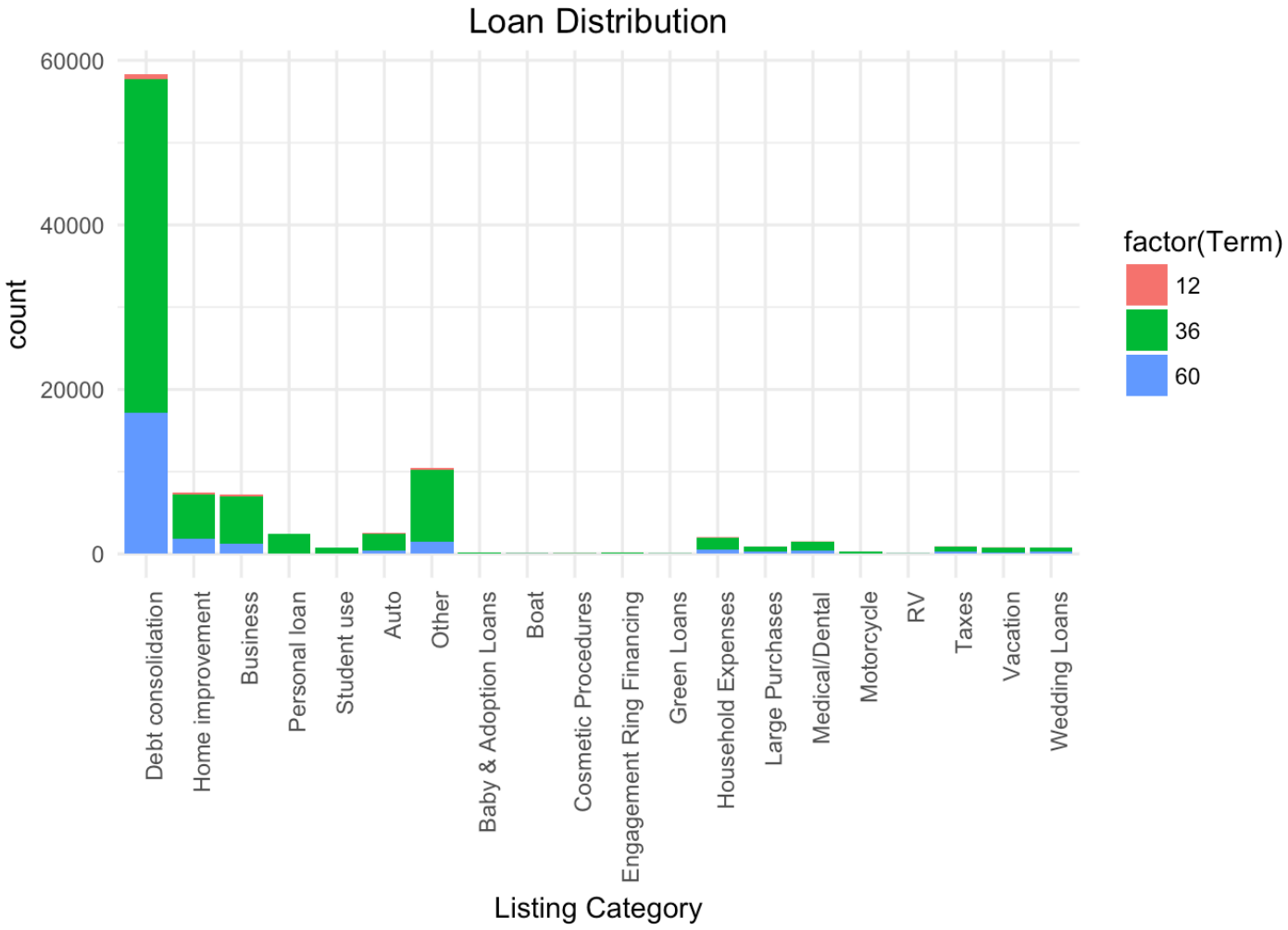
It is curious to say that when borrowing a high loan amount, the Prosper Rating should be going from AA (least risky) to HR (most risky) but in this plot it is showing the exact opposite a low amount of loan is the riskiest for investors.

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

No

Plot One

```
# # Final Plots and Summary
ggplot(data=subset(prosper, ListingCategory != "Not available"), aes(x= ListingCat
egory)) +
  geom_bar(aes(fill=factor(Term))) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = .5)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 1, hjust = 1)) +
  labs(title="Loan Distribution", x="Listing Category")
```

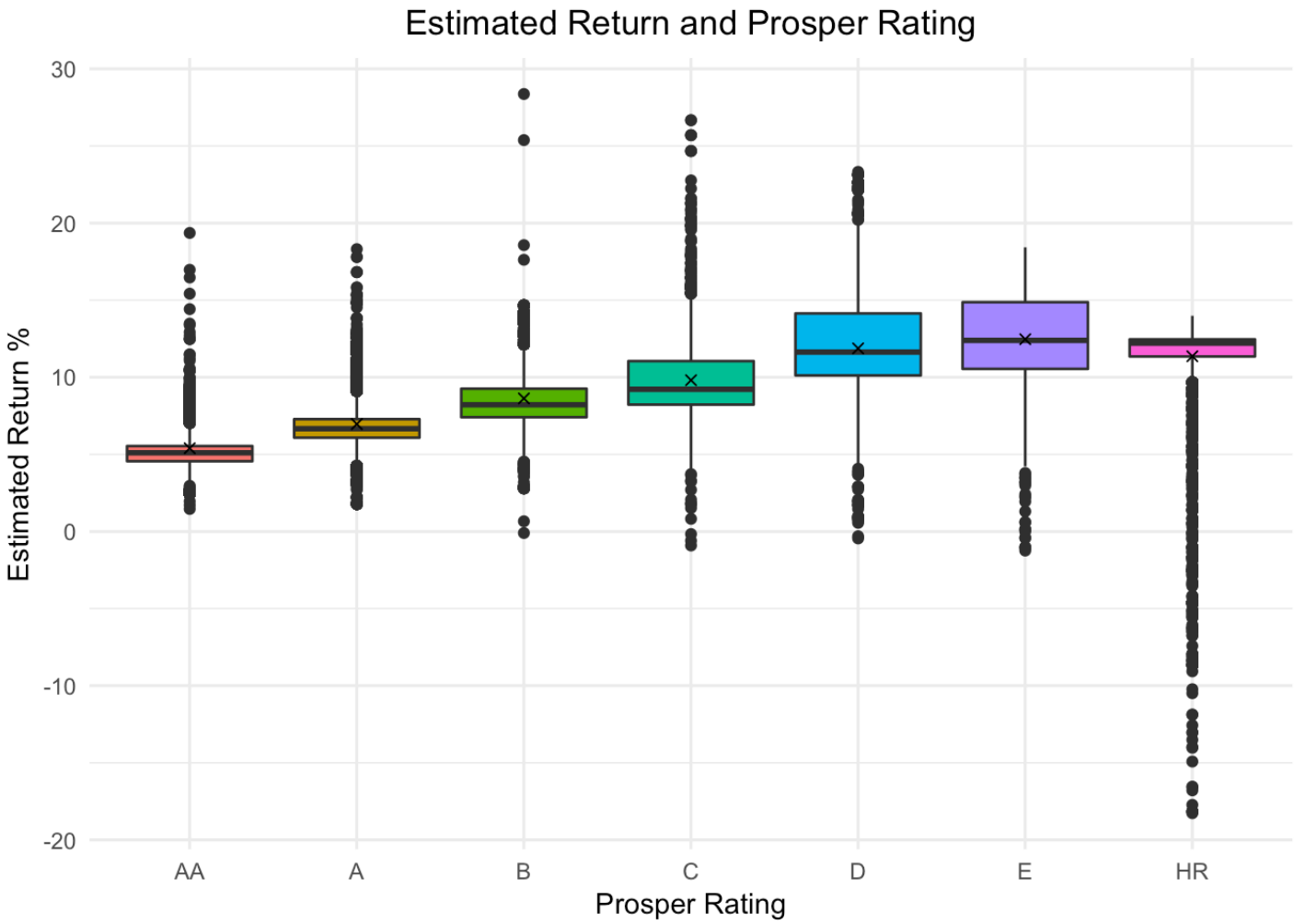


Description One

The first plot shows the vast majority of loans were mainly for Debt Consolidation, Home Improvement and Business purposes. Most borrowers opted to pay off the loans within 36 months which is their favourite choice out of the three terms. Borrowers may have used the loans to compensate other debts that they owed such as credit card bills.

Plot Two

```
 # What is the relationship between Borrower Rate and Prosper Rating?
ggplot(data=subset(prosper, !is.na(ProsperRating)), aes(x=ProsperRating, y=Estimat
edReturn *100)) +
  geom_boxplot(aes(factor(ProsperRating), fill=factor(ProsperRating))) +
  stat_summary(fun.y = mean, geom = "point", shape =4) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = .5)) +
  guides(fill=FALSE) +
  labs(title="Estimated Return and Prosper Rating", x="Prosper Rating", y="Estimat
ed Return %")
```

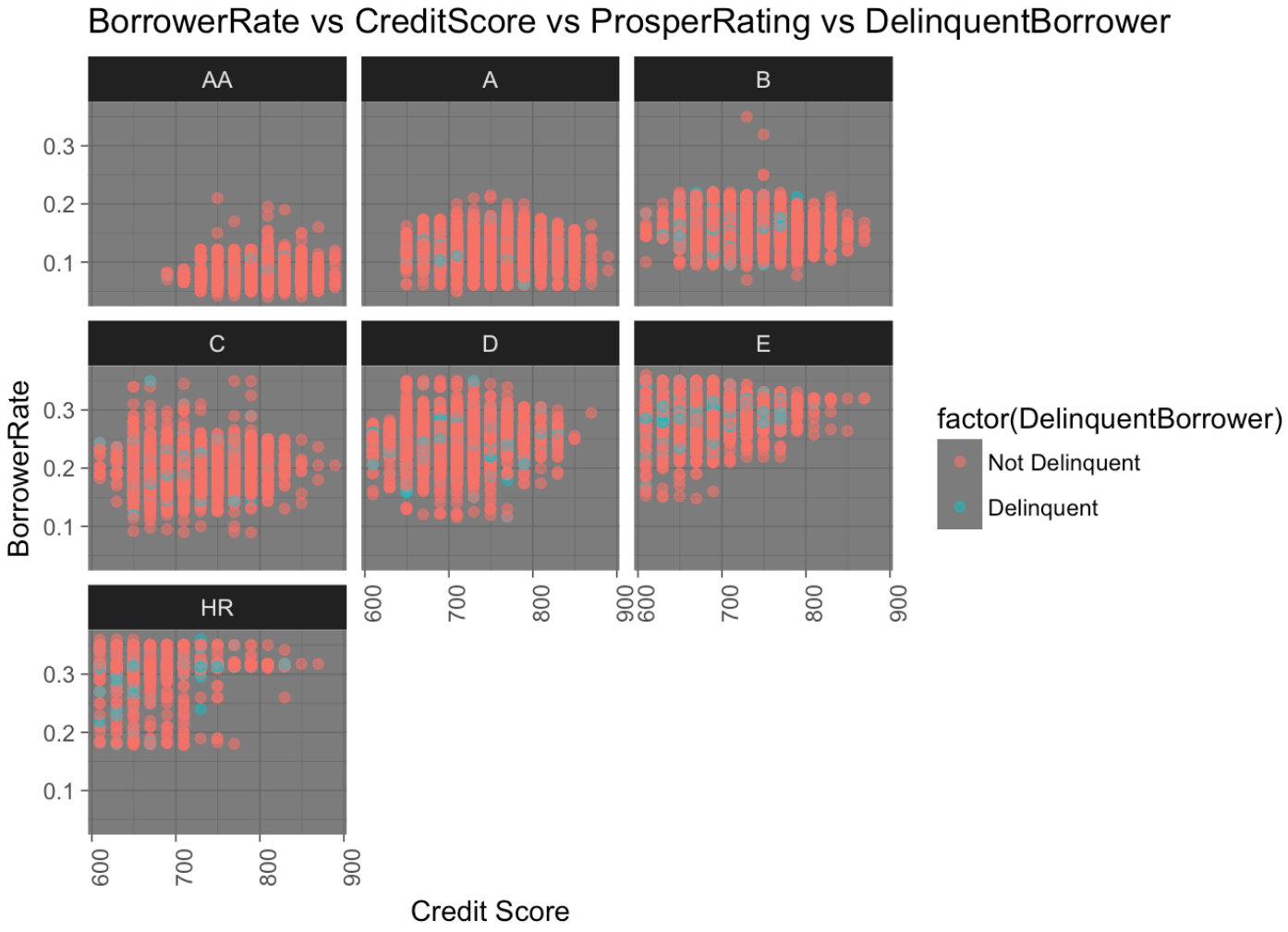


Description Two

The Estimated Return is determined once the loan is listed which will calculate how much risk there will be for investors (borrowers being defaulted or late monthly repayments etc). It is clear to say that as the Prosper Rating goes down (from AA to HR) the interest rates will increase significantly as the investment becomes riskier. Therefore, investors tend to charge borrowers less for those with a good Prosper Rating.

Plot Three

```
ggplot(data=subset(prosper, !is.na(DelinquentBorrower) & !is.na(ProsperRating)), a
es(x= CreditScore, y=BorrowerRate, color=factor(DelinquentBorrower))) +
  geom_point(alpha=1/2) +
  facet_wrap(~ProsperRating) +
  scale_color_brewer(type = "seq", palette = 1, direction = 1) +
  scale_color_discrete(labels= c("Not Delinquent", "Delinquent")) +
  theme_minimal() +
  theme_dark() +
  labs(title="BorrowerRate vs CreditScore vs ProsperRating vs DelinquentBorrower",
x="Credit Score", y="BorrowerRate") +
  theme(axis.text.x = element_text(angle = 90, vjust = 1, hjust = 1))
```



Description Three

For the good ratings (AA, A and B) delinquent borrowers in green fall more to the left where the credit scores are lower, and the interest rates are significantly less than the lower ratings. In other words, Delinquent borrowers have lower credit scores which could be due to their bad performance in the past for making late payments therefore, given higher interest rates. Borrowers who are in red tend to have higher credit scores thereby paying lower rates.

Reflection

The Prosper dataset is very large with nearly 114,000 observations and 81 variables as well as records dating from 2005 - 2014. It required a lot of planning to understand more about the company, its P2P lending model and the industry as a whole. So, choosing which variables to analyse was a bit of a challenge, but after spending some time studying the dataset I was able to find leads and thought of interesting questions as I continued to explore deeper on various different plots.

> After conducting my exploratory data analysis, I was able to find out interesting insights:

A majority of loans were requested for “Debt Consolidation” in order to pay off their other loans or debts.

Borrowers who own a home tend to get larger loans than those who do not.

The most popular choice for loan term is 36 months out of the 3 terms (12 and 60 months).

Delinquent Borrowers are mainly those who applied for loans less than $10,000 and within the low level of Prosper Ratings, Prosper Score and Credit Score but at the same time pay higher interest rates and better return.

Most Borrowers have a Prosper Rating of “C” and “AA” being the least.

Estimated Return tends to be higher as the Borrower’s rating lowers

Borrowers with a higher credit score were given a high rating of either “AA”, “A” and “B”, but surprisingly those who have a low rating of “HR” have credit scores higher than rating “E”.

Most borrowers earn between $25,000 - $74,999.

Based on all these findings I finally able to create the borrowers' profiles where I found a list of factors that would determine the borrower's rates, the estimated return, and indicate whether the investment is profitable or not. If borrowers have a low credit score, it is clear that they will be charged at a higher interest rate. However, depending on the borrowers' circumstances and many other internal and external factors, it might be a different story and they could pay at a lower rate. After this analysis, it had provided me a lot of valuable insights and I am now more confident in investing in borrowers' listings with a small percentage of risk as they offer more return and will not easily make a loss.
To go further with this analysis, additional data could immensely improve this dataset such as gathering more on borrower's gender and age. This will be helpful as I can understand more about the borrower whether this person is a male ih his late 40s or a female in her early 20s etc. Another idea for future work would be to apply a classification algorithm such as Logistic Regression or K Nearest Neighbors to make a predictive model and to see the investment is worth taking or not.

References

- Prosper: https://www.prosper.com (https://www.prosper.com)
- Kaggle: https://www.kaggle.com/jschnessl/prosperloans (https://www.kaggle.com/jschnessl/prosperloans)
- R Mark Down: https://www.rstudio.com/wp-content/uploads/2016/03/rmarkdown-cheatsheet-2.0.pdf (https://www.rstudio.com/wp-content/uploads/2016/03/rmarkdown-cheatsheet-2.0.pdf)
- Investopedia: https://www.investopedia.com/news/peertopeer-lender-prosper-sell-5-billion-loans/ (https://www.investopedia.com/news/peertopeer-lender-prosper-sell-5-billion-loans/)
- Stackoverflow: https://stackoverflow.com/jobs/companies/prosper-marketplace (https://stackoverflow.com/jobs/companies/prosper-marketplace)