R Notebook for Prosper Loan Data

PREPARING RSTUDIO AND THE DATA SET

This is Project 4 for the Udacity Data Analyst nanodegree. I am using R to exlore the Prosper Loan Dataset. This dataset included information about loans that Prosper sold. Prosper.com is a peer-to-peer lending marketplace. Borrowers make loan requests and investors contribute as little as $25 towards the loans of their choice. To begin, I installed the packages as instructed in the rubric.

```
library("ggplot2")
library("knitr")
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
```

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

Opening the Data Set

```
getwd()
```

```
## [1] "C:/Users/Nancy Olewnik/Documents"
```

```
pf <- read.csv('prosperLoanData.csv')
names(pf)
```

```
##  [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"
```

Running the data & summary files

```
data(pf)
summary(pf)
```

```
##                     ListingKey     ListingNumber
##  17A93590655669644DB4C06:     6   Min.   :      4
##  349D3587495831350F0F648:     4   1st Qu.: 400919
##  47C1359638497431975670B:     4   Median : 600554
##  8474358854651984137201C:     4   Mean   : 627886
##  DE8535960513435199406CE:     4   3rd Qu.: 892634
##  04C13599434217079754AEE:     3   Max.   :1255725
##  (Other)                :113912
##                    ListingCreationDate  CreditGrade         Term
##  2013-10-02 17:20:16.550000000:     6          :84984   Min.   :12.00
##  2013-08-28 20:31:41.107000000:     4   C      : 5649   1st Qu.:36.00
##  2013-09-08 09:27:44.853000000:     4   D      : 5153   Median :36.00
##  2013-12-06 05:43:13.830000000:     4   B      : 4389   Mean   :40.83
##  2013-12-06 11:44:58.283000000:     4   AA     : 3509   3rd Qu.:36.00
##  2013-08-21 07:25:22.360000000:     3   HR     : 3508   Max.   :60.00
##  (Other)                      :113912   (Other): 6745
##                 LoanStatus                  ClosedDate
##  Current               :56576                      :58848
##  Completed             :38074   2014-03-04 00:00:00:  105
##  Chargedoff            :11992   2014-02-19 00:00:00:  100
##  Defaulted             : 5018   2014-02-11 00:00:00:   92
##  Past Due (1-15 days) :  806   2012-10-30 00:00:00:   81
##  Past Due (31-60 days):  363   2013-02-26 00:00:00:   78
##  (Other)               : 1108   (Other)            :54633
##   BorrowerAPR       BorrowerRate     LenderYield
##  Min.   :0.00653   Min.   :0.0000   Min.   :-0.0100
##  1st Qu.:0.15629   1st Qu.:0.1340   1st Qu.: 0.1242
##  Median :0.20976   Median :0.1840   Median : 0.1730
##  Mean   :0.21883   Mean   :0.1928   Mean   : 0.1827
##  3rd Qu.:0.28381   3rd Qu.:0.2500   3rd Qu.: 0.2400
##  Max.   :0.51229   Max.   :0.4975   Max.   : 0.4925
##  NA's   :25
##  EstimatedEffectiveYield EstimatedLoss   EstimatedReturn
##  Min.   :-0.183          Min.   :0.005   Min.   :-0.183
##  1st Qu.: 0.116          1st Qu.:0.042   1st Qu.: 0.074
##  Median : 0.162          Median :0.072   Median : 0.092
##  Mean   : 0.169          Mean   :0.080   Mean   : 0.096
##  3rd Qu.: 0.224          3rd Qu.:0.112   3rd Qu.: 0.117
##  Max.   : 0.320          Max.   :0.366   Max.   : 0.284
##  NA's   :29084           NA's   :29084   NA's   :29084
##  ProsperRating..numeric. ProsperRating..Alpha.  ProsperScore
##  Min.   :1.000                  :29084         Min.   : 1.00
##  1st Qu.:3.000           C      :18345         1st Qu.: 4.00
##  Median :4.000           B      :15581         Median : 6.00
##  Mean   :4.072           A      :14551         Mean   : 5.95
##  3rd Qu.:5.000           D      :14274         3rd Qu.: 8.00
##  Max.   :7.000           E      : 9795         Max.   :11.00
##  NA's   :29084           (Other):12307         NA's   :29084
##  ListingCategory..numeric. BorrowerState
##  Min.   : 0.000            CA     :14717
##  1st Qu.: 1.000            TX     : 6842
##  Median : 1.000            NY     : 6729
##  Mean   : 2.774            FL     : 6720
##  3rd Qu.: 3.000            IL     : 5921
##  Max.   :20.000                   : 5515
##                            (Other):67493
##                    Occupation         EmploymentStatus
##  Other                     :28617   Employed     :67322
##  Professional              :13628   Full-time    :26355
##  Computer Programmer       : 4478   Self-employed: 6134
##  Executive                 : 4311   Not available: 5347
##  Teacher                   : 3759   Other        : 3806
##  Administrative Assistant: 3688                 : 2255
##  (Other)                   :55456   (Other)      : 2718
##  EmploymentStatusDuration IsBorrowerHomeowner CurrentlyInGroup
##  Min.   :  0.00           False:56459         False:101218
##  1st Qu.: 26.00           True :57478         True : 12719
##  Median : 67.00
##  Mean   : 96.07
##  3rd Qu.:137.00
##  Max.   :755.00
##  NA's   :7625
##                    GroupKey                 DateCreditPulled
##                        :100596   2013-12-23 09:38:12:     6
##  783C3371218786870A73D20:  1140   2013-11-21 09:09:41:     4
##  3D4D3366260257624AB272D:   916   2013-12-06 05:43:16:     4
##  6A3B336601725506917317E:   698   2014-01-14 20:17:49:     4
##  FEF83377364176536637E50:   611   2014-02-09 12:14:41:     4
##  C9643379247860156A00EC0:   342   2013-09-27 22:04:54:     3
##  (Other)                :  9634   (Other)            :113912
##  CreditScoreRangeLower CreditScoreRangeUpper
##  Min.   :  0.0         Min.   : 19.0
##  1st Qu.:660.0         1st Qu.:679.0
##  Median :680.0         Median :699.0
##  Mean   :685.6         Mean   :704.6
##  3rd Qu.:720.0         3rd Qu.:739.0
##  Max.   :880.0         Max.   :899.0
##  NA's   :591           NA's   :591
##         FirstRecordedCreditLine CurrentCreditLines OpenCreditLines
##                     :   697     Min.   : 0.00      Min.   : 0.00
##  1993-12-01 00:00:00:   185     1st Qu.: 7.00      1st Qu.: 6.00
##  1994-11-01 00:00:00:   178     Median :10.00      Median : 9.00
##  1995-11-01 00:00:00:   168     Mean   :10.32      Mean   : 9.26
##  1990-04-01 00:00:00:   161     3rd Qu.:13.00      3rd Qu.:12.00
##  1995-03-01 00:00:00:   159     Max.   :59.00      Max.   :54.00
##  (Other)            :112389     NA's   :7604       NA's   :7604
##  TotalCreditLinespast7years OpenRevolvingAccounts
##  Min.   :  2.00             Min.   : 0.00
##  1st Qu.: 17.00             1st Qu.: 4.00
##  Median : 25.00             Median : 6.00
##  Mean   : 26.75             Mean   : 6.97
##  3rd Qu.: 35.00             3rd Qu.: 9.00
##  Max.   :136.00             Max.   :51.00
##  NA's   :697
##  OpenRevolvingMonthlyPayment InquiriesLast6Months TotalInquiries
##  Min.   :    0.0             Min.   :  0.000      Min.   :  0.000
##  1st Qu.:  114.0             1st Qu.:  0.000      1st Qu.:  2.000
##  Median :  271.0             Median :  1.000      Median :  4.000
##  Mean   :  398.3             Mean   :  1.435      Mean   :  5.584
##  3rd Qu.:  525.0             3rd Qu.:  2.000      3rd Qu.:  7.000
##  Max.   :14985.0             Max.   :105.000      Max.   :379.000
##                              NA's   :697          NA's   :1159
##  CurrentDelinquencies AmountDelinquent   DelinquenciesLast7Years
##  Min.   : 0.0000      Min.   :     0.0   Min.   : 0.000
##  1st Qu.: 0.0000      1st Qu.:     0.0   1st Qu.: 0.000
##  Median : 0.0000      Median :     0.0   Median : 0.000
##  Mean   : 0.5921      Mean   :   984.5   Mean   : 4.155
##  3rd Qu.: 0.0000      3rd Qu.:     0.0   3rd Qu.: 3.000
##  Max.   :83.0000      Max.   :463881.0   Max.   :99.000
##  NA's   :697          NA's   :7622       NA's   :990
##  PublicRecordsLast10Years PublicRecordsLast12Months RevolvingCreditBalance
##  Min.   : 0.0000          Min.   : 0.000            Min.   :      0
##  1st Qu.: 0.0000          1st Qu.: 0.000            1st Qu.:   3121
##  Median : 0.0000          Median : 0.000            Median :   8549
##  Mean   : 0.3126          Mean   : 0.015            Mean   :  17599
##  3rd Qu.: 0.0000          3rd Qu.: 0.000            3rd Qu.:  19521
##  Max.   :38.0000          Max.   :20.000            Max.   :1435667
##  NA's   :697              NA's   :7604              NA's   :7604
##  BankcardUtilization AvailableBankcardCredit  TotalTrades
##  Min.   :0.000       Min.   :     0          Min.   :  0.00
##  1st Qu.:0.310       1st Qu.:   880          1st Qu.: 15.00
##  Median :0.600       Median :  4100          Median : 22.00
##  Mean   :0.561       Mean   : 11210          Mean   : 23.23
##  3rd Qu.:0.840       3rd Qu.: 13180          3rd Qu.: 30.00
##  Max.   :5.950       Max.   :646285          Max.   :126.00
##  NA's   :7604        NA's   :7544            NA's   :7544
##  TradesNeverDelinquent..percentage. TradesOpenedLast6Months
##  Min.   :0.000                      Min.   : 0.000
##  1st Qu.:0.820                      1st Qu.: 0.000
##  Median :0.940                      Median : 0.000
##  Mean   :0.886                      Mean   : 0.802
##  3rd Qu.:1.000                      3rd Qu.: 1.000
##  Max.   :1.000                      Max.   :20.000
##  NA's   :7544                       NA's   :7544
##  DebtToIncomeRatio         IncomeRange    IncomeVerifiable
##  Min.   : 0.000    $25,000-49,999:32192   False:  8669
##  1st Qu.: 0.140    $50,000-74,999:31050   True :105268
##  Median : 0.220    $100,000+     :17337
##  Mean   : 0.276    $75,000-99,999:16916
##  3rd Qu.: 0.320    Not displayed : 7741
##  Max.   :10.010    $1-24,999     : 7274
##  NA's   :8554      (Other)       : 1427
##  StatedMonthlyIncome                     LoanKey       TotalProsperLoans
##  Min.   :      0     CB1B37030986463208432A1:     6   Min.   :0.00
##  1st Qu.:   3200     2DEE3698211017519D7333F:     4   1st Qu.:1.00
##  Median :   4667     9F4B37043517554537C364C:     4   Median :1.00
##  Mean   :   5608     D895370150591392337ED6D:     4   Mean   :1.42
##  3rd Qu.:   6825     E6FB37073953690388BC56D:     4   3rd Qu.:2.00
##  Max.   :1750003     0D8F37036734373301ED419:     3   Max.   :8.00
##                      (Other)                :113912   NA's   :91852
##  TotalProsperPaymentsBilled OnTimeProsperPayments
##  Min.   :  0.00             Min.   :  0.00
##  1st Qu.:  9.00             1st Qu.:  9.00
##  Median : 16.00             Median : 15.00
##  Mean   : 22.93             Mean   : 22.27
##  3rd Qu.: 33.00             3rd Qu.: 32.00
##  Max.   :141.00             Max.   :141.00
##  NA's   :91852              NA's   :91852
##  ProsperPaymentsLessThanOneMonthLate ProsperPaymentsOneMonthPlusLate
##  Min.   : 0.00                       Min.   : 0.00
##  1st Qu.: 0.00                       1st Qu.: 0.00
##  Median : 0.00                       Median : 0.00
##  Mean   : 0.61                       Mean   : 0.05
##  3rd Qu.: 0.00                       3rd Qu.: 0.00
##  Max.   :42.00                       Max.   :21.00
##  NA's   :91852                       NA's   :91852
##  ProsperPrincipalBorrowed ProsperPrincipalOutstanding
##  Min.   :    0            Min.   :    0
##  1st Qu.: 3500            1st Qu.:    0
##  Median : 6000            Median : 1627
##  Mean   : 8472            Mean   : 2930
##  3rd Qu.:11000            3rd Qu.: 4127
##  Max.   :72499            Max.   :23451
##  NA's   :91852            NA's   :91852
##  ScorexChangeAtTimeOfListing LoanCurrentDaysDelinquent
##  Min.   :-209.00             Min.   :   0.0
##  1st Qu.: -35.00             1st Qu.:   0.0
##  Median :  -3.00             Median :   0.0
##  Mean   :  -3.22             Mean   : 152.8
##  3rd Qu.:  25.00             3rd Qu.:   0.0
##  Max.   : 286.00             Max.   :2704.0
##  NA's   :95009
##  LoanFirstDefaultedCycleNumber LoanMonthsSinceOrigination   LoanNumber
##  Min.   : 0.00                 Min.   :  0.0              Min.   :     1
##  1st Qu.: 9.00                 1st Qu.:  6.0              1st Qu.: 37332
##  Median :14.00                 Median : 21.0              Median : 68599  
##  Mean   :16.27                 Mean   : 31.9              Mean   : 69444  
##  3rd Qu.:22.00                 3rd Qu.: 65.0              3rd Qu.:101901  
##  Max.   :44.00                 Max.   :100.0              Max.   :136486  
##  NA's   :96985                                                            
##  LoanOriginalAmount          LoanOriginationDate LoanOriginationQuarter
##  Min.   : 1000      2014-01-22 00:00:00:   491   Q4 2013:14450         
##  1st Qu.: 4000      2013-11-13 00:00:00:   490   Q1 2014:12172         
##  Median : 6500      2014-02-19 00:00:00:   439   Q3 2013: 9180         
##  Mean   : 8337      2013-10-16 00:00:00:   434   Q2 2013: 7099         
##  3rd Qu.:12000      2014-01-28 00:00:00:   339   Q3 2012: 5632         
##  Max.   :35000      2013-09-24 00:00:00:   316   Q2 2012: 5061         
##                     (Other)            :111428   (Other):60343         
##                    MemberKey      MonthlyLoanPayment LP_CustomerPayments
##  63CA34120866140639431C9:     9   Min.   :   0.0     Min.   :   -2.35   
##  16083364744933457E57FB9:     8   1st Qu.: 131.6     1st Qu.: 1005.76   
##  3A2F3380477699707C81385:     8   Median : 217.7     Median : 2583.83   
##  4D9C3403302047712AD0CDD:     8   Mean   : 272.5     Mean   : 4183.08   
##  739C338135235294782AE75:     8   3rd Qu.: 371.6     3rd Qu.: 5548.40   
##  7E1733653050264822FAA3D:     8   Max.   :2251.5     Max.   :40702.39   
##  (Other)                :113888                                         
##  LP_CustomerPrincipalPayments LP_InterestandFees LP_ServiceFees   
##  Min.   :    0.0             Min.   :   -2.35   Min.   :-664.87  
##  1st Qu.:  500.9             1st Qu.:  274.87   1st Qu.: -73.18  
##  Median : 1587.5             Median :  700.84   Median : -34.44  
##  Mean   : 3105.5             Mean   : 1077.54   Mean   : -54.73  
##  3rd Qu.: 4000.0             3rd Qu.: 1458.54   3rd Qu.: -13.92  
##  Max.   :35000.0             Max.   :15617.03   Max.   :  32.06  
##                                                                  
##  LP_CollectionFees  LP_GrossPrincipalLoss LP_NetPrincipalLoss
##  Min.   :-9274.75   Min.   :  -94.2       Min.   : -954.5    
##  1st Qu.:    0.00   1st Qu.:    0.0       1st Qu.:    0.0    
##  Median :    0.00   Median :    0.0       Median :    0.0    
##  Mean   :  -14.24   Mean   :  700.4       Mean   :  681.4    
##  3rd Qu.:    0.00   3rd Qu.:    0.0       3rd Qu.:    0.0    
##  Max.   :    0.00   Max.   :25000.0       Max.   :25000.0    
##                                                              
##  LP_NonPrincipalRecoverypayments PercentFunded    Recommendations   
##  Min.   :    0.00                Min.   :0.7000   Min.   : 0.00000  
##  1st Qu.:    0.00                1st Qu.:1.0000   1st Qu.: 0.00000  
##  Median :    0.00                Median :1.0000   Median : 0.00000  
##  Mean   :   25.14                Mean   :0.9986   Mean   : 0.04803  
##  3rd Qu.:    0.00                3rd Qu.:1.0000   3rd Qu.: 0.00000  
##  Max.   :21117.90                Max.   :1.0125   Max.   :39.00000  
##                                                                     
##  InvestmentFromFriendsCount InvestmentFromFriendsAmount   Investors      
##  Min.   : 0.00000           Min.   :    0.00            Min.   :   1.00  
##  1st Qu.: 0.00000           1st Qu.:    0.00            1st Qu.:   2.00  
##  Median : 0.00000           Median :    0.00            Median :  44.00  
##  Mean   : 0.02346           Mean   :   16.55           Mean   :  80.48
##  3rd Qu.: 0.00000           3rd Qu.:    0.00           3rd Qu.: 115.00
##  Max.   :33.00000           Max.   :25000.00           Max.   :1189.00
##
```

Does my data set over 1,000 observations? Are there at least 8 different variables?

```
dim(pf)
```

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

113,937 observations with 81 variables

List out the description of variables and types

```
str(pf)
```

```
## 'data.frame':    113937 obs. of  81 variables:
##  $ ListingKey                         : Factor w/ 113066 levels "00003546482
094282EF90E5",..: 7180 7193 6647 6669 6686 6689 6699 6706 6687 6687 ...
##  $ ListingNumber                      : int  193129 1209647 81716 658116 909
464 1074836 750899 768193 1023355 1023355 ...
##  $ ListingCreationDate                : Factor w/ 113064 levels "2005-11-09
20:44:28.847000000",..: 14184 111894 6429 64760 85967 100310 72556 74019 97834
97834 ...
##  $ CreditGrade                        : Factor w/ 9 levels "","A","A
A","B",..: 5 1 8 1 1 1 1 1 1 1 ...
##  $ Term                               : int  36 36 36 36 36 60 36 36 36 3
6 ...
##  $ LoanStatus                         : Factor w/ 12 levels "Cancelled","Cha
rgedoff",..: 3 4 3 4 4 4 4 4 4 4 ...
##  $ ClosedDate                         : Factor w/ 2803 levels "","2005-11-2
5 00:00:00",..: 1138 1 1263 1 1 1 1 1 1 1 ...
##  $ BorrowerAPR                        : num  0.165 0.12 0.283 0.125 0.24
6 ...
##  $ BorrowerRate                       : num  0.158 0.092 0.275 0.0974 0.208
5 ...
##  $ LenderYield                        : num  0.138 0.082 0.24 0.0874 0.198
5 ...
##  $ EstimatedEffectiveYield            : num  NA 0.0796 NA 0.0849 0.1832 ...
##  $ EstimatedLoss                      : num  NA 0.0249 NA 0.0249 0.0925 ...
##  $ EstimatedReturn                    : num  NA 0.0547 NA 0.06 0.0907 ...
##  $ ProsperRating..numeric.            : int  NA 6 NA 6 3 5 2 4 7 7 ...
##  $ ProsperRating..Alpha.              : Factor w/ 8 levels "","A","A
A","B",..: 1 2 1 2 6 4 7 5 3 3 ...
##  $ ProsperScore                       : num  NA 7 NA 9 4 10 2 4 9 11 ...
##  $ ListingCategory..numeric.          : int  0 2 0 16 2 1 1 2 7 7 ...
##  $ BorrowerState                      : Factor w/ 52 levels "","AK","AL","A
R",..: 7 7 12 12 25 34 18 6 16 16 ...
##  $ Occupation                         : Factor w/ 68 levels "","Accountant/C
PA",..: 37 43 37 52 21 43 50 29 24 24 ...
##  $ EmploymentStatus                   : Factor w/ 9 levels "","Employe
d",..: 9 2 4 2 2 2 2 2 2 2 ...
##  $ EmploymentStatusDuration           : int  2 44 NA 113 44 82 172 103 269 2
69 ...
##  $ IsBorrowerHomeowner                : Factor w/ 2 levels "False","True":
2 1 1 2 2 2 1 1 2 2 ...
##  $ CurrentlyInGroup                   : Factor w/ 2 levels "False","True":
2 1 2 1 1 1 1 1 1 1 ...
##  $ GroupKey                           : Factor w/ 707 levels "","00343376901
312423168731",..: 1 1 335 1 1 1 1 1 1 1 ...
##  $ DateCreditPulled                   : Factor w/ 112992 levels "2005-11-09
00:30:04.487000000",..: 14347 111883 6446 64724 85857 100382 72500 73937 97888
97888 ...
##  $ CreditScoreRangeLower              : int  640 680 480 800 680 740 680 70
```

```
0 820 820 ...
##  $ CreditScoreRangeUpper              : int  659 699 499 819 699 759 699 71
9 839 839 ...
##  $ FirstRecordedCreditLine            : Factor w/ 11586 levels "","1947-08-2
4 00:00:00",..: 8639 6617 8927 2247 9498 497 8265 7685 5543 5543 ...
##  $ CurrentCreditLines                 : int  5 14 NA 5 19 21 10 6 17 17 ...
##  $ OpenCreditLines                    : int  4 14 NA 5 19 17 7 6 16 16 ...
##  $ TotalCreditLinespast7years         : int  12 29 3 29 49 49 20 10 32 3
2 ...
##  $ OpenRevolvingAccounts              : int  1 13 0 7 6 13 6 5 12 12 ...
##  $ OpenRevolvingMonthlyPayment        : num  24 389 0 115 220 1410 214 101 2
19 219 ...
##  $ InquiriesLast6Months               : int  3 3 0 0 1 0 0 3 1 1 ...
##  $ TotalInquiries                     : num  3 5 1 1 9 2 0 16 6 6 ...
##  $ CurrentDelinquencies               : int  2 0 1 4 0 0 0 0 0 0 ...
##  $ AmountDelinquent                   : num  472 0 NA 10056 0 ...
##  $ DelinquenciesLast7Years            : int  4 0 0 14 0 0 0 0 0 0 ...
##  $ PublicRecordsLast10Years           : int  0 1 0 0 0 0 0 1 0 0 ...
##  $ PublicRecordsLast12Months          : int  0 0 NA 0 0 0 0 0 0 0 ...
##  $ RevolvingCreditBalance             : num  0 3989 NA 1444 6193 ...
##  $ BankcardUtilization                : num  0 0.21 NA 0.04 0.81 0.39 0.72
0.13 0.11 0.11 ...
##  $ AvailableBankcardCredit            : num  1500 10266 NA 30754 695 ...
##  $ TotalTrades                        : num  11 29 NA 26 39 47 16 10 29 2
9 ...
##  $ TradesNeverDelinquent..percentage. : num  0.81 1 NA 0.76 0.95 1 0.68 0.8
1 1 ...
##  $ TradesOpenedLast6Months            : num  0 2 NA 0 2 0 0 0 1 1 ...
##  $ DebtToIncomeRatio                  : num  0.17 0.18 0.06 0.15 0.26 0.36
0.27 0.24 0.25 0.25 ...
##  $ IncomeRange                        : Factor w/ 8 levels "$0","$1-24,99
9",..: 4 5 7 4 3 3 4 4 4 4 ...
##  $ IncomeVerifiable                   : Factor w/ 2 levels "False","True":
2 2 2 2 2 2 2 2 2 2 ...
##  $ StatedMonthlyIncome                : num  3083 6125 2083 2875 9583 ...
##  $ LoanKey                            : Factor w/ 113066 levels "00003683605
746079487FF7",..: 100337 69837 46303 70776 71387 86505 91250 5425 908 908 ...
##  $ TotalProsperLoans                  : int  NA NA NA NA 1 NA NA NA NA N
A ...
##  $ TotalProsperPaymentsBilled         : int  NA NA NA NA 11 NA NA NA NA N
A ...
##  $ OnTimeProsperPayments              : int  NA NA NA NA 11 NA NA NA NA N
A ...
##  $ ProsperPaymentsLessThanOneMonthLate: int  NA NA NA NA 0 NA NA NA NA N
A ...
##  $ ProsperPaymentsOneMonthPlusLate    : int  NA NA NA NA 0 NA NA NA NA N
A ...
##  $ ProsperPrincipalBorrowed           : num  NA NA NA NA 11000 NA NA NA NA N
A ...
```

```
##  $ ProsperPrincipalOutstanding        : num  NA NA NA NA 9948 ...
##  $ ScorexChangeAtTimeOfListing        : int  NA NA NA NA NA NA NA NA NA N
A ...
##  $ LoanCurrentDaysDelinquent          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ LoanFirstDefaultedCycleNumber      : int  NA NA NA NA NA NA NA NA NA N
A ...
##  $ LoanMonthsSinceOrigination         : int  78 0 86 16 6 3 11 10 3 3 ...
##  $ LoanNumber                         : int  19141 134815 6466 77296 102670
123257 88353 90051 121268 121268 ...
##  $ LoanOriginalAmount                 : int  9425 10000 3001 10000 15000 150
00 3000 10000 10000 10000 ...
##  $ LoanOriginationDate                : Factor w/ 1873 levels "2005-11-15 0
0:00:00",..: 426 1866 260 1535 1757 1821 1649 1666 1813 1813 ...
##  $ LoanOriginationQuarter             : Factor w/ 33 levels "Q1 2006","Q1 20
07",..: 18 8 2 32 24 33 16 16 33 33 ...
##  $ MemberKey                          : Factor w/ 90831 levels "000033976974
13387CAF966",..: 11071 10302 33781 54939 19465 48037 60448 40951 26129 2612
9 ...
##  $ MonthlyLoanPayment                 : num  330 319 123 321 564 ...
##  $ LP_CustomerPayments                : num  11396 0 4187 5143 2820 ...
##  $ LP_CustomerPrincipalPayments       : num  9425 0 3001 4091 1563 ...
##  $ LP_InterestandFees                 : num  1971 0 1186 1052 1257 ...
##  $ LP_ServiceFees                     : num  -133.2 0 -24.2 -108 -60.3 ...
##  $ LP_CollectionFees                  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ LP_GrossPrincipalLoss              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ LP_NetPrincipalLoss                : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ LP_NonPrincipalRecoverypayments    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ PercentFunded                      : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ Recommendations                    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ InvestmentFromFriendsCount         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ InvestmentFromFriendsAmount        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Investors                          : int  258 1 41 158 20 1 1 1 1 1 ...
```

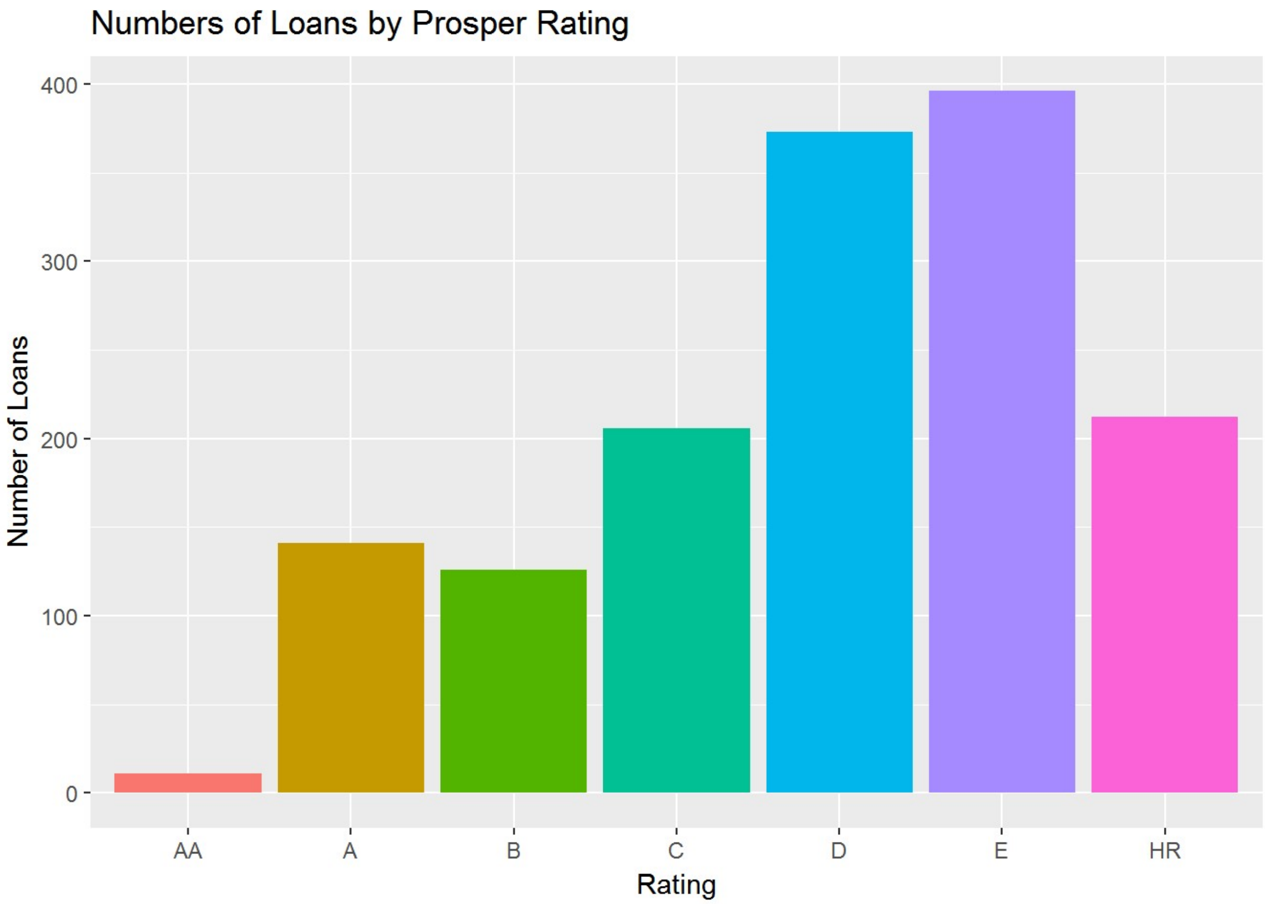
UNIVARIATE PLOT SECTIION

Factorizing rating for the key variable we’d investigate throughout the dataset

```
pf$ProsperRating.alpha = factor(pf$ProsperRating..Alpha.,
                                 levels = c("AA","A","B","C","D","E","HR","N
A"))
pf$ProsperRating <-factor(pf$ProsperRating..Alpha,
                        levels = c('AA', 'A', 'B', 'C', 'D', 'E', 'HR', 'NA'))
pf$ProsperScore = factor(pf$ProsperScore)
```

1U HISTOGRAM OF PROSPER RATING BY NUMBERS OF LOANS

```
ggplot(data = na.omit(pf), aes(ProsperRating.alpha)) +
  geom_bar(aes(fill = ProsperRating.alpha),stat="count") + guides(fill=FALSE) +
  ggtitle('Numbers of Loans by Prosper Rating') +
  xlab('Rating') +
  ylab('Number of Loans')
```



```
summary(pf$ProsperRating.alpha)
```

```
##     AA      A      B      C      D      E     HR     NA   NA's
##   5372  14551  15581  18345  14274   9795   6935      0  29084
```

Looks like “NA” and “C” rating loans account for the majority of the loans.

2U PROSPER RATING DISTRIBUTION

```
table(pf$ProsperRating..numeric., useNA = 'ifany')
```

```
##
##     1     2     3     4     5     6     7  <NA>
##  6935  9795 14274 18345 15581 14551  5372 29084
```

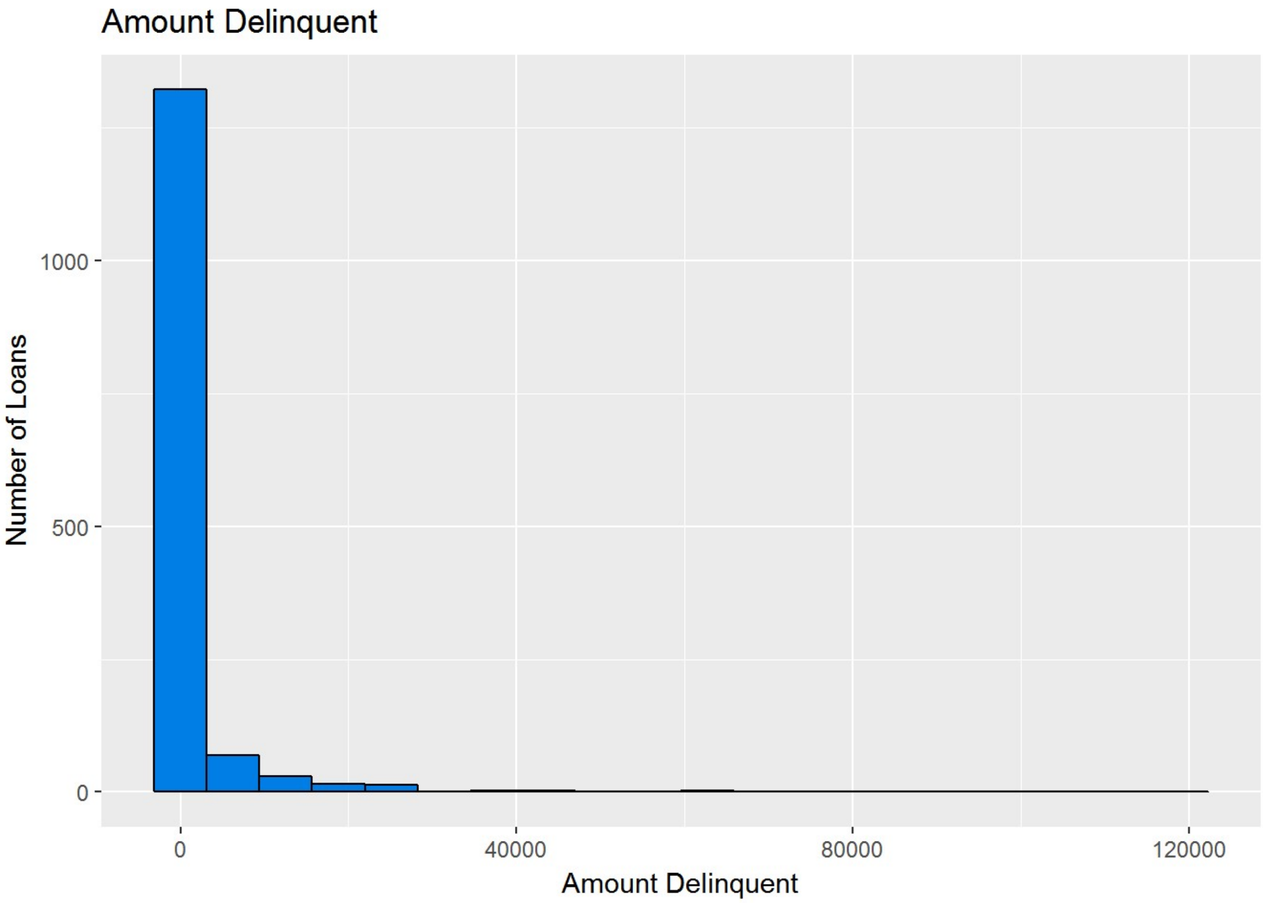
```
summary(pf$ProsperRating..numeric., useNA = 'ifany')
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   3.000   4.000   4.072   5.000   7.000   29084
```

The NA count of Prosper Rating and Prosper Score is similar (29,084). I’m curious how the Prosper Rating and Prosper Score varies.

3U AMOUNT DELINQUENT

```
ggplot(data = na.omit(pf), aes(AmountDelinquent)) +
  geom_histogram(aes(fill = AmountDelinquent), color = "black", fill = '#007EE
5',bins=20) +
  ggtitle('Amount Delinquent') +
  xlab('Amount Delinquent') +
  ylab('Number of Loans')
```



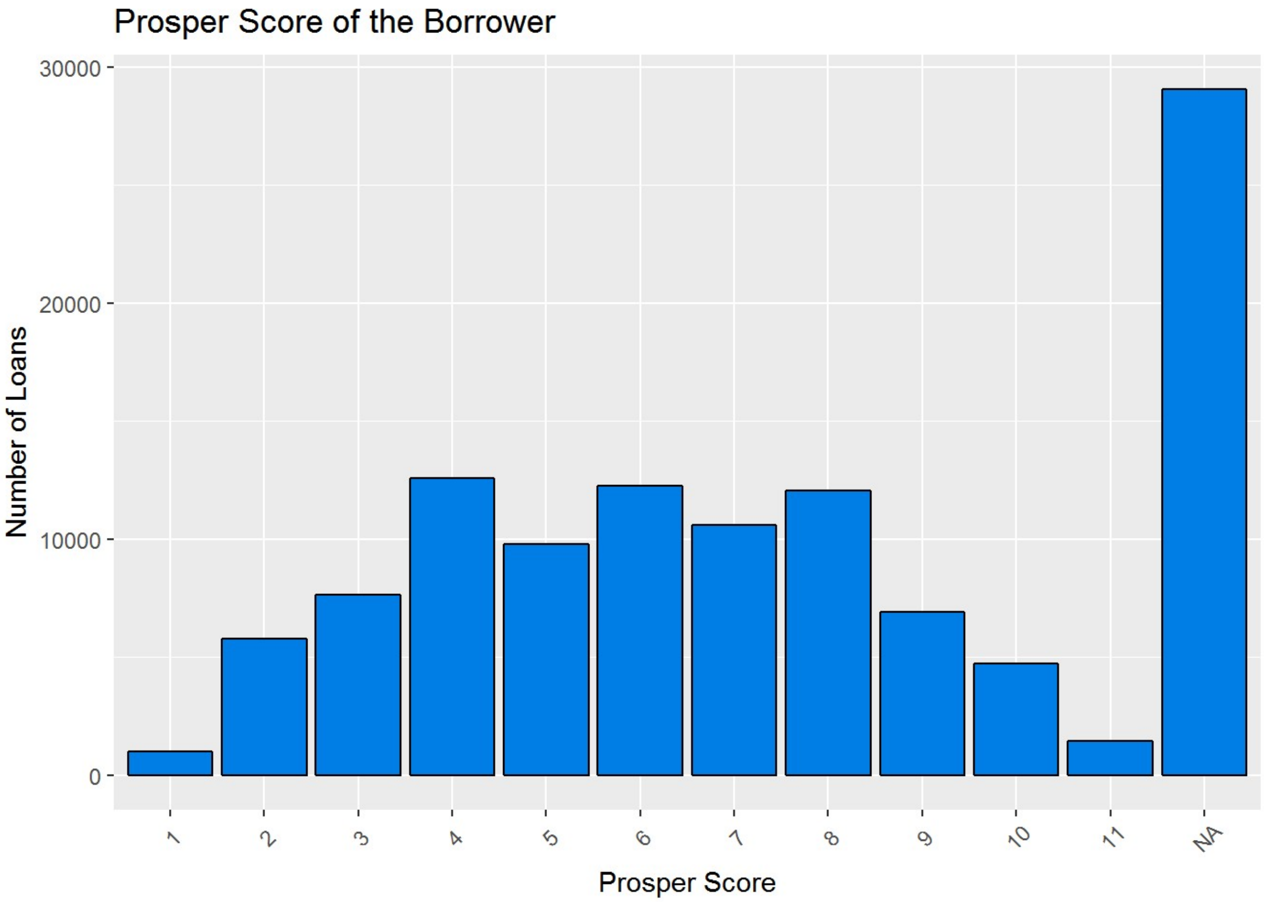
```
summary(pf$AmountDelinquent)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's 
##      0.0      0.0      0.0    984.5      0.0 463900.0     7622
```

This chart tells us that the mean amount deliquent is $985. The maximum in default is over $400,000. The bar chart shows the the most frequent deliquent amount is about $1,000.

4U SCORE DISTRIBUTION

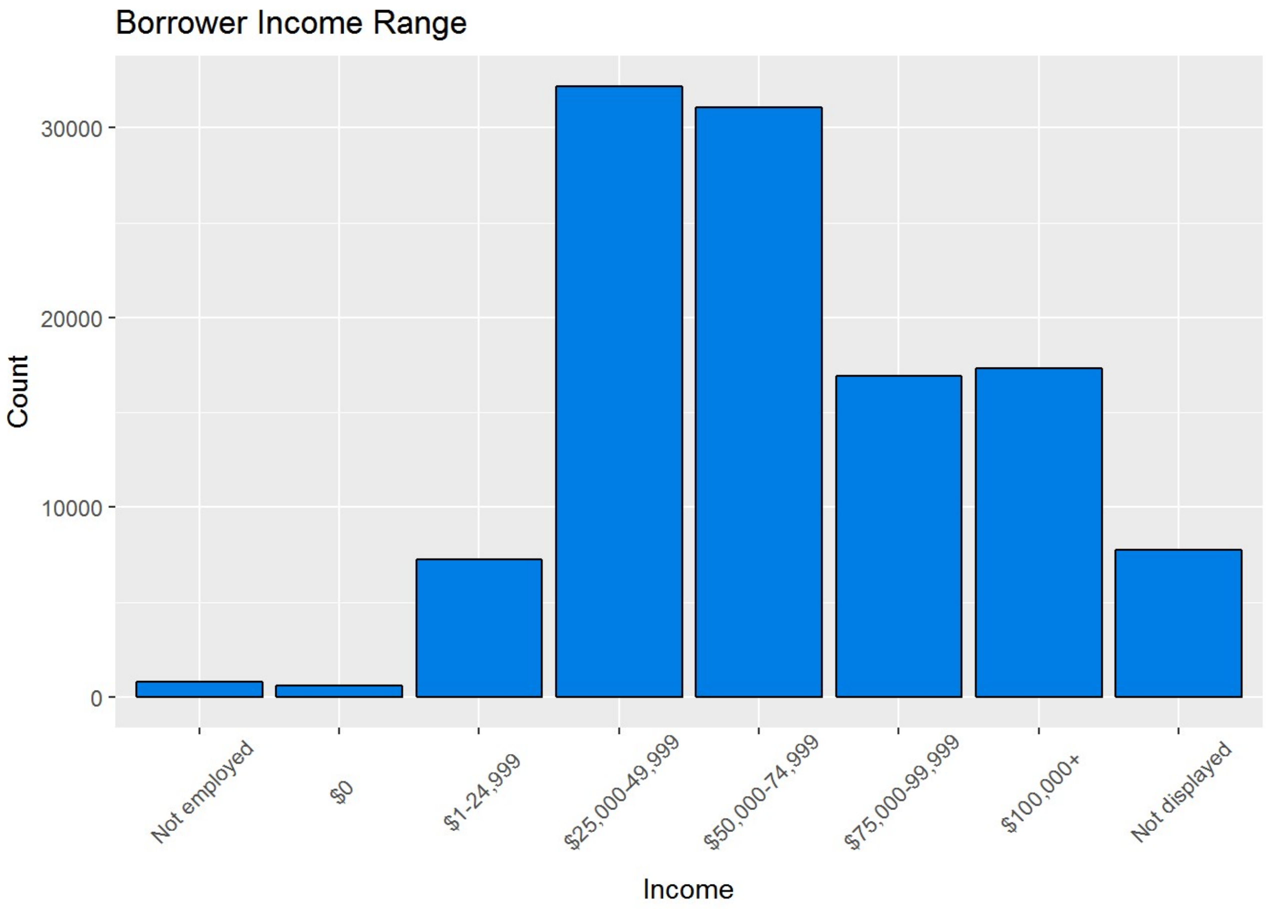
```
ggplot(data = pf, aes(ProsperScore)) +
  geom_bar(color="black", fill = '#007EE5') +
  ggtitle('Prosper Score of the Borrower') +
  xlab('Prosper Score') +
  theme(axis.text.x = element_text(angle = 45, vjust = 0.6)) +
  ylab('Number of Loans')
```



Again, the majority of the scores are “NA” and in the 4-8. category range. Why are there so many ProsperScores that are NA?

5U BORROWER INCOME RANGE

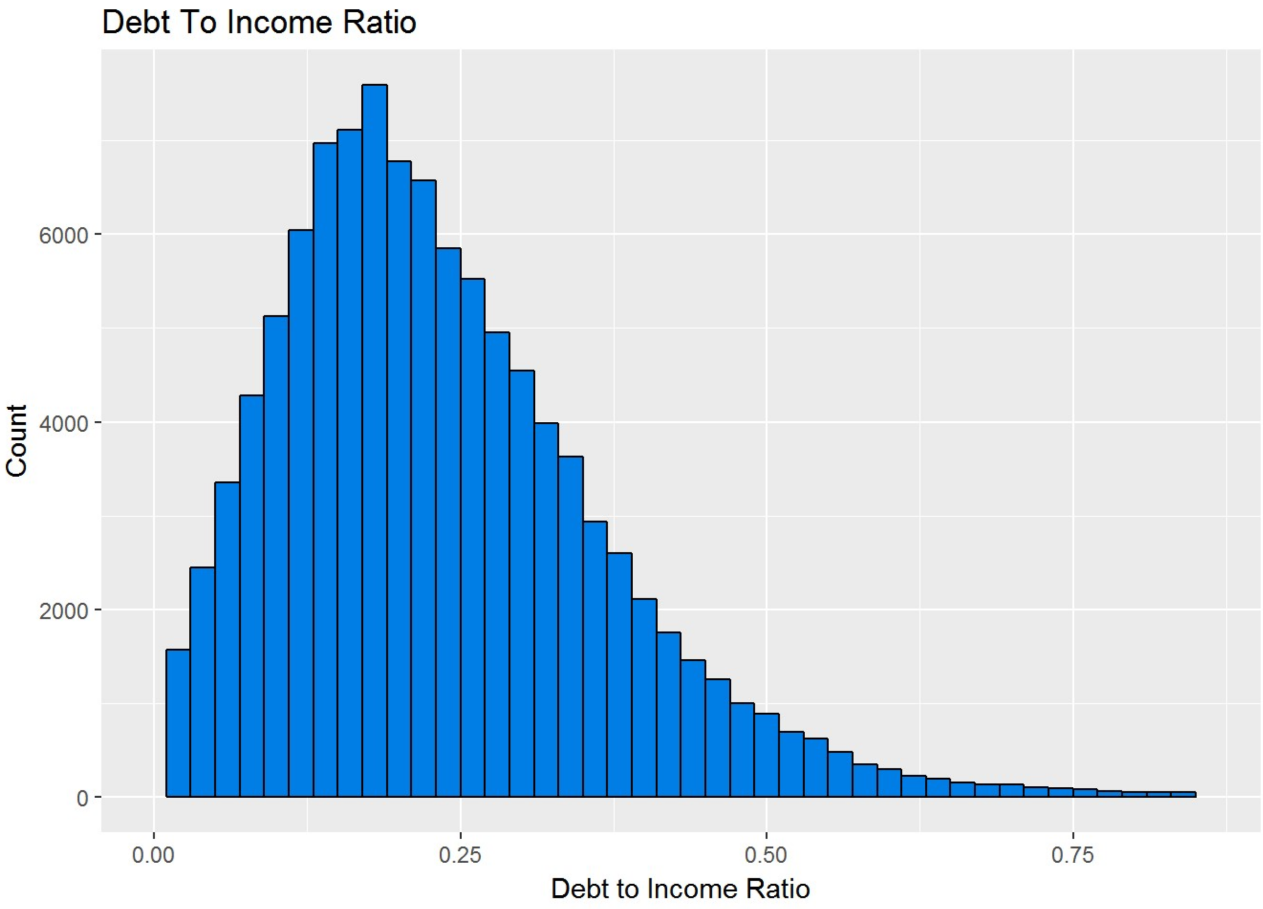
```
pf$IncomeRange = factor(pf$IncomeRange, levels=c("Not employed", "$0", "$1-24,9
99", "$25,000-49,999", "$50,000-74,999", "$75,000-99,999", "$100,000+", "Not di
splayed"))

ggplot(data = pf, aes(IncomeRange)) +
  geom_bar(color="black", fill = '#007EE5') +
  ggtitle('Borrower Income Range') +
  xlab('Income') +
  theme(axis.text.x = element_text(angle = 45, vjust = 0.6)) +
  ylab('Count')
```



The majority of borrowers are in the $25,000 - $75,000 range. I suspect this lower-middle class range needs loans for debt consolidations.

6U DEBT TO INCOME RATIO

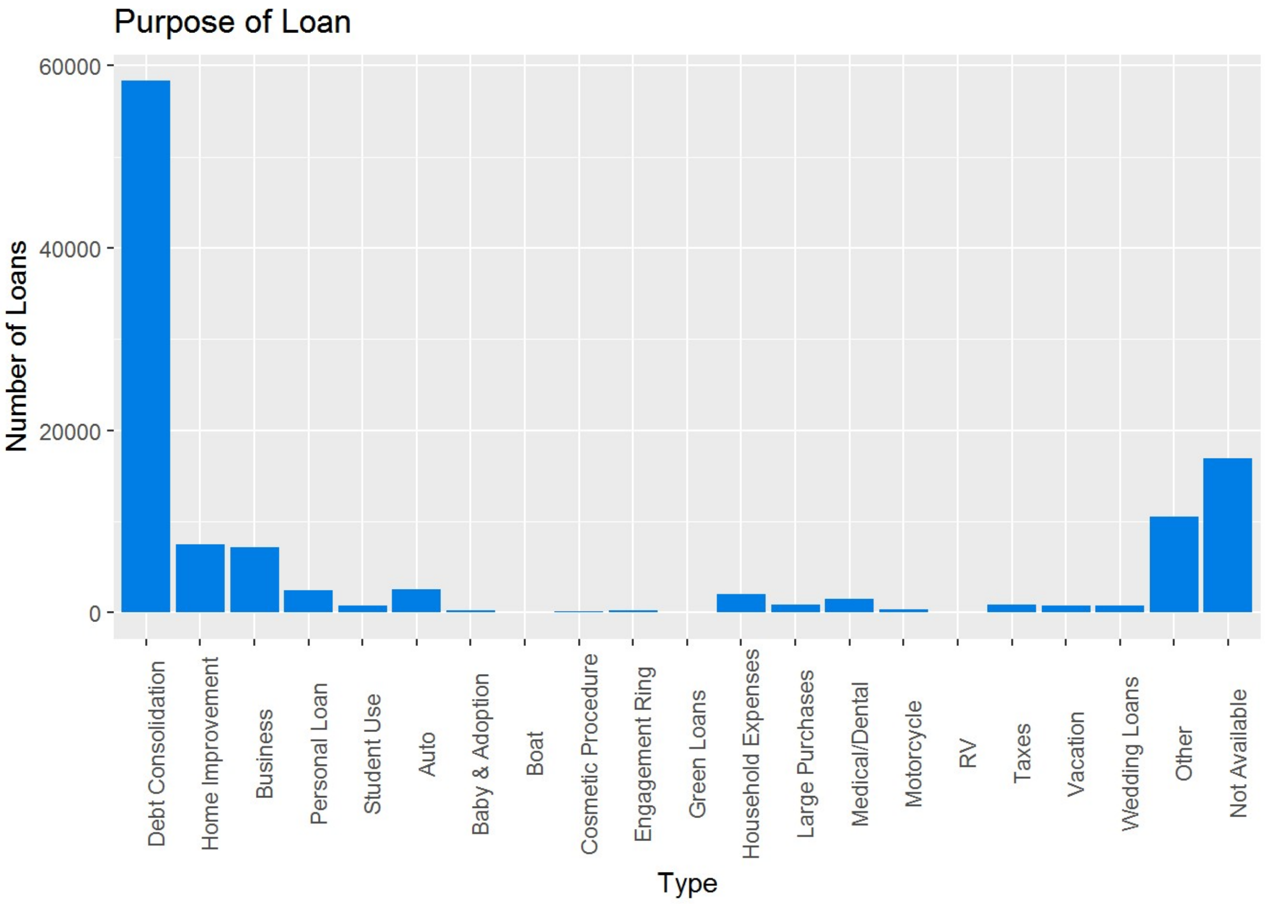
```
ggplot(data = pf, aes(x = DebtToIncomeRatio)) +
        geom_histogram(color = "black", fill = '#007EE5', binwidth = 0.02) +
        xlim(0, quantile(pf$DebtToIncomeRatio, prob = 0.99, na.rm=TRUE)) +
        ggtitle("Debt To Income Ratio") +
        xlab("Debt to Income Ratio") +
        ylab("Count")
```



The data is long-tailed right-skewed. It’s expected the majority of people in U.S have a credit history and the ratio should be low enough for a secured repayment.

7U BORROWER’S PURPOSE OF LOAN

```
x <- c('Debt Consolidation',
                                'Home Improvement','Business',
                                 'Personal Loan',
                                 'Student Use',
                                 'Auto',
                                 'Baby & Adoption',
                                 'Boat',
                                 'Cosmetic Procedure',
                                 'Engagement Ring',
                                 'Green Loans',
                                 'Household Expenses',
                                 'Large Purchases',
                                 'Medical/Dental',
                                 'Motorcycle', 'RV',
                                 'Taxes', 'Vacation',
                                 'Wedding Loans',
                                 'Other',
                                 'Not Available')

pf$ListingCategory <- factor(pf$ListingCategory..numeric., levels = c(1:6,8:20,
7,0), labels = x)

ggplot(data = pf, aes(x=ListingCategory)) +
  geom_bar(aes(y=..count..), size = 3, fill = '#007EE5', stat="count") +
  ggtitle('Purpose of Loan') +
  xlab('Type') +
  ylab('Number of Loans') +
  theme(axis.text.x = element_text(angle = 90))
```



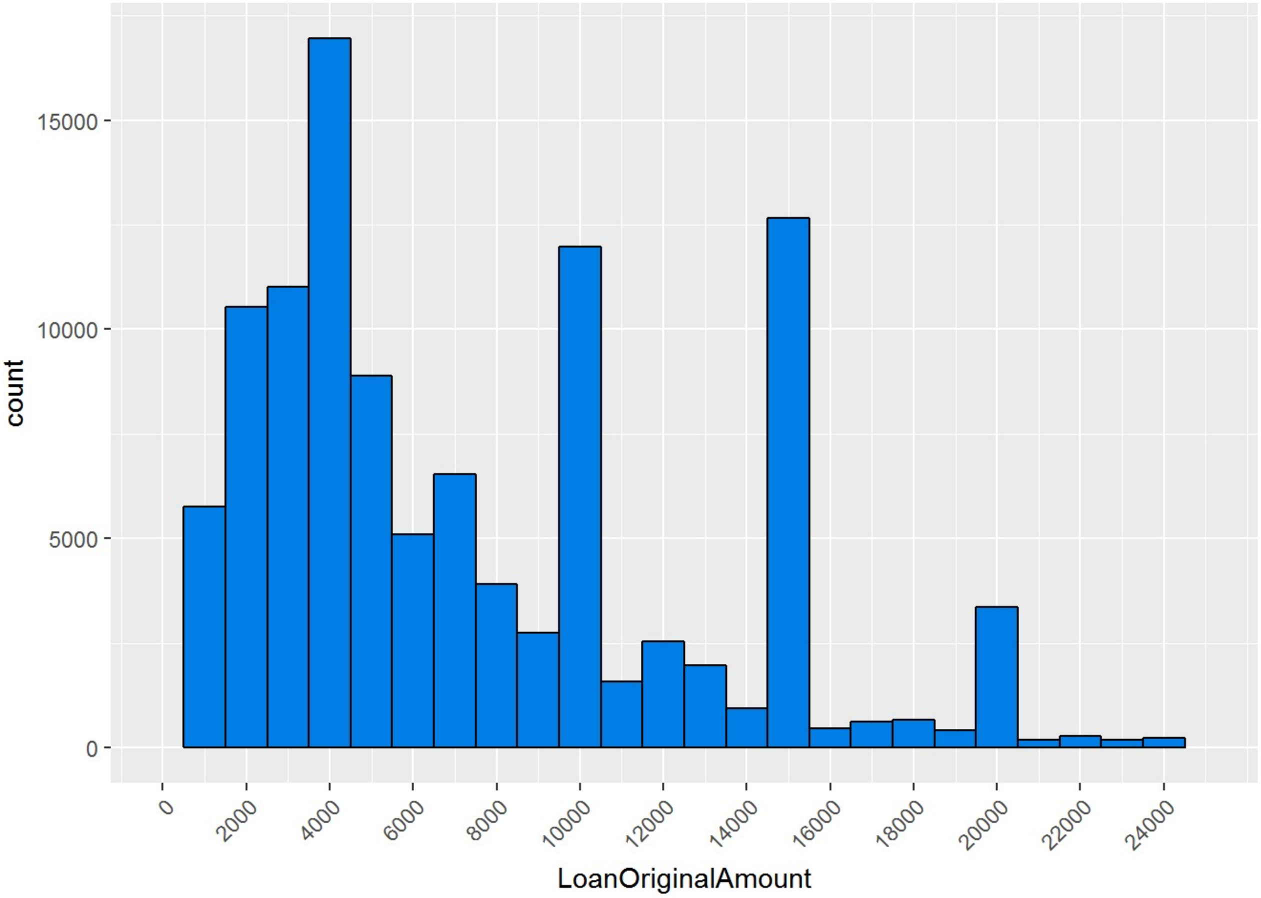
```
summary(pf$ListingCategory)
```

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

This chart tells us that not many people are willing to explain the purpose of the loan. I’m surprised that Prosper doesn’t require this field. It also looks like there is a high need, more than 50%, for loans for debt consolidation.

8U LOAN SPLIT BY AMOUNT

```
ggplot(pf, aes(LoanOriginalAmount)) +
                geom_histogram(color = "black", fill = '#007EE5', binwidth = 10
00) +
                scale_x_continuous(
                limits = c(0,quantile(pf$LoanOriginalAmount, 0.99,na.rm = TRU
E)),
                breaks = seq(0, quantile(pf$LoanOriginalAmount, 0.99, na.rm = T
RUE), 2000)) +
                theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



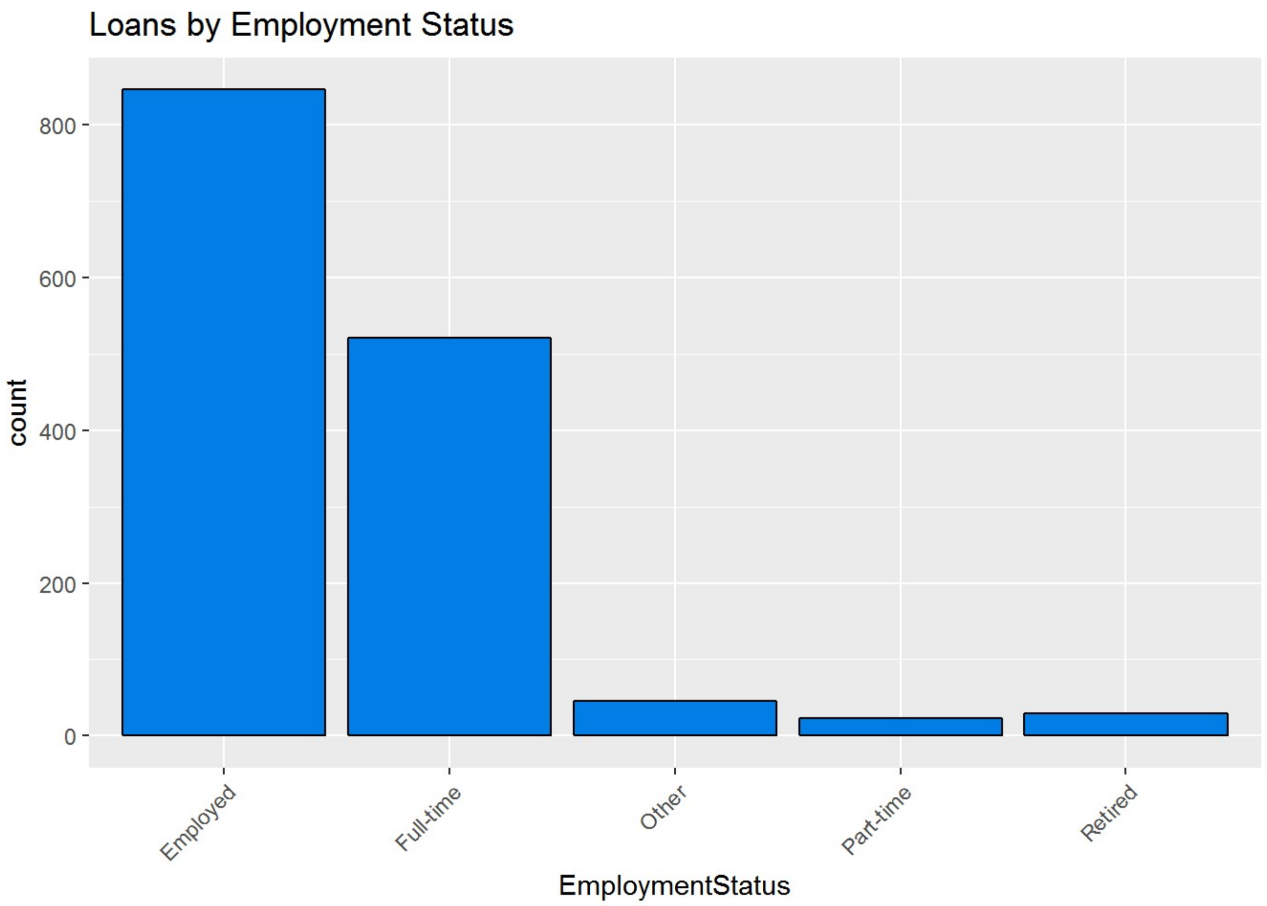
```
summary(pf$LoanOriginalAmount)
```

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

The minimum loan amount is $1,000. There appears to four main ranges wherre people borrow money ($5,000 - $10,000 - $15,000 - $20,000). Although this might be more than enough for them to cover their original need, people tend to check these rounded amount boxes.

9U EMPLOYMENT STATUS

```
ggplot(aes(x = EmploymentStatus), data = na.omit(pf)) +
                    geom_bar(color = "black", fill = '#007EE5') +
                    theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
                    ggtitle("Loans by Employment Status")
```



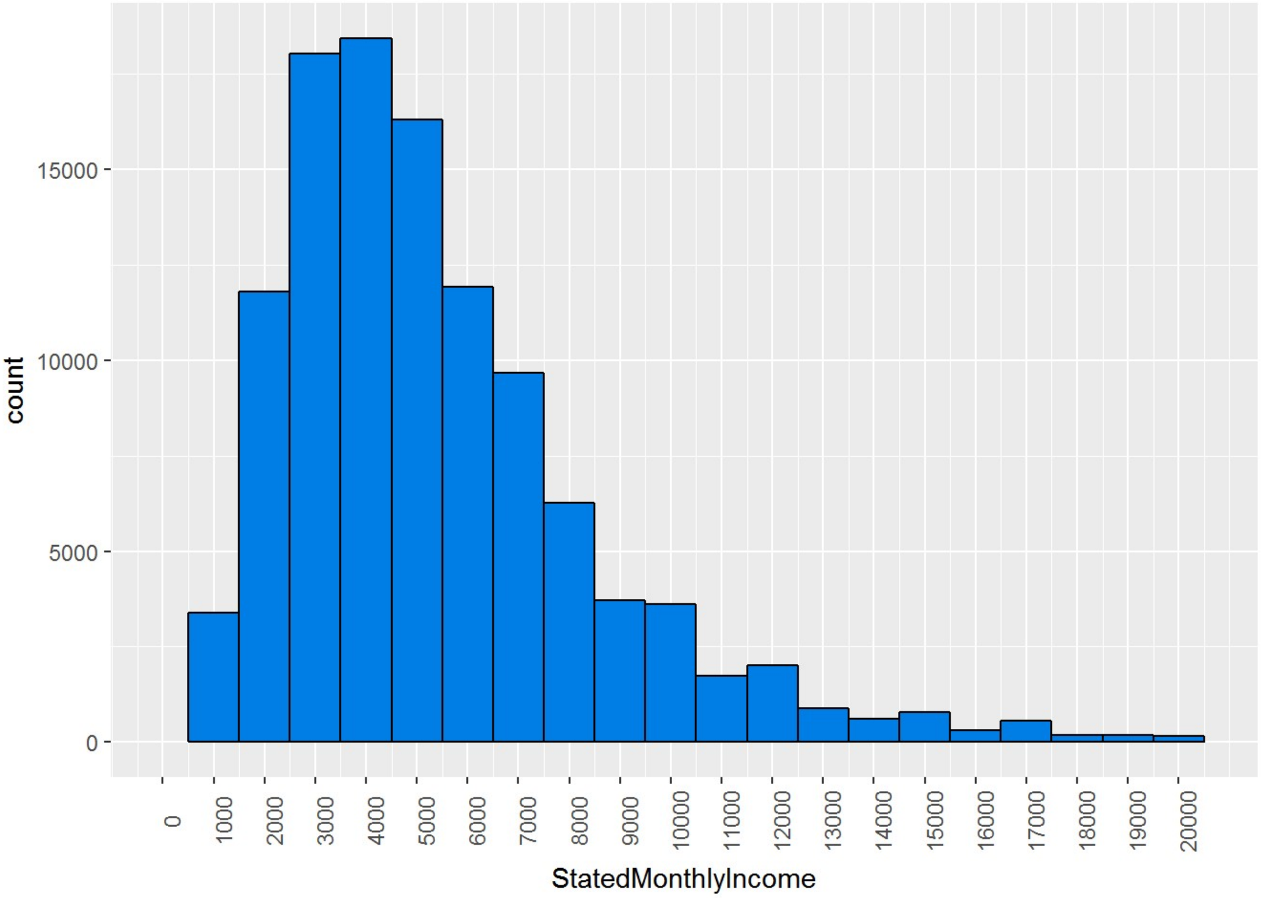
```
summary(pf$EmploymentStatus)
```

```
##                     Employed     Full-time Not available  Not employed 
##          2255         67322         26355          5347           835 
##         Other     Part-time       Retired Self-employed 
##          3806          1088           795          6134 
```

This chart shows that the majority is employed; however, this data could be skewed. Does the “employed” data include part-time or full-time?

10U STATED MONTHLY INCOME

```
ggplot(aes(x = StatedMonthlyIncome), data = pf) +
                  geom_histogram(color = "black", fill = '#007EE5', binwidth =
1000) +
                  scale_x_continuous(
                  limits = c(0, quantile(pf$StatedMonthlyIncome, 0.99,
                                                              na.rm = TRUE)),
                  breaks = seq(0, quantile(pf$StatedMonthlyIncome, 0.99,
                                                      na.rm = TRUE), 1000)) +
                  theme(axis.text.x = element_text(angle = 90))
```



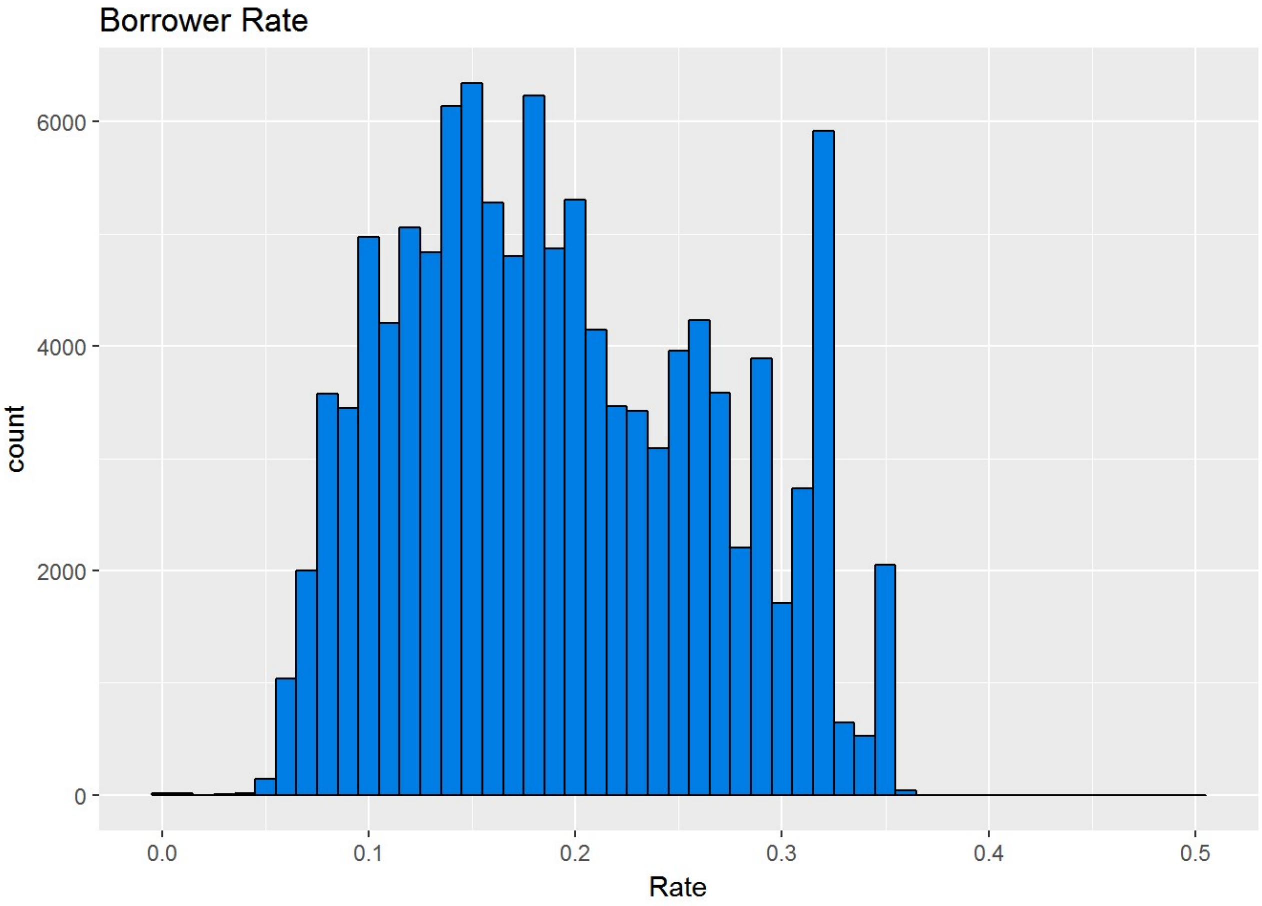
```
summary(pf$StatedMonthlyIncome)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    3200    4667    5608    6825 1750000
```

This chart tells us the most popular stated monthly income is $4,000 - $5,000.

11U BORROWER’S RATE

```
ggplot(data = pf, aes(x = BorrowerRate)) +
        geom_histogram(color = "black", fill = '#007EE5', binwidth = 0.01) +
        xlab("Rate") +
        ggtitle("Borrower Rate")
```



```
summary(pf$BorrowerRate)
```

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

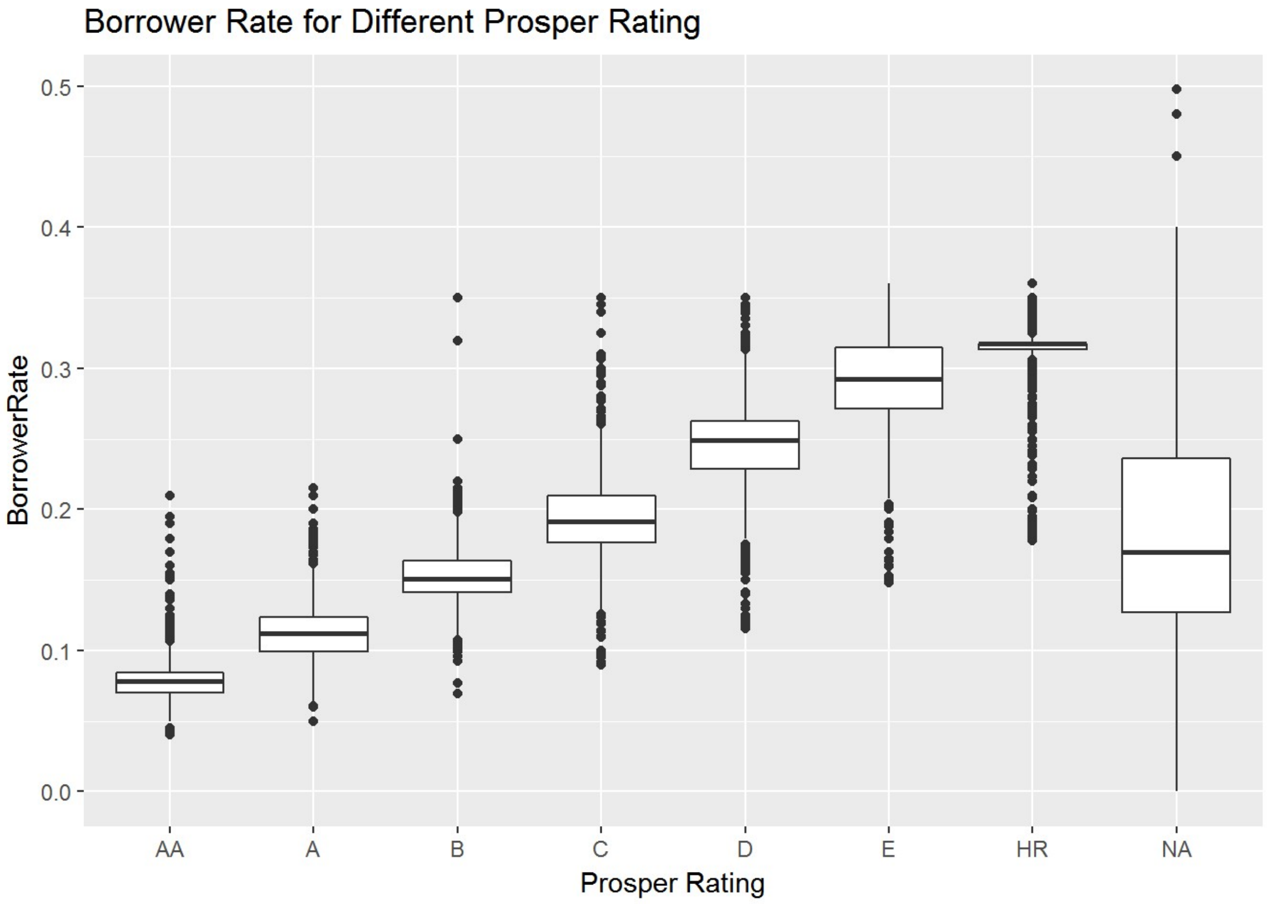
The most frequent rates are approximately 15%, 17% and 32%. This variation could be a factor of the amount or debt-to-income ratio.

BIVARIATE PLOT & ANAYLSIS SECTION

```
pf$ProsperRating.alpha = factor(pf$ProsperRating..Alpha.,
                                 levels = c("AA","A","B","C","D","E","HR","N
A"))
pf$ProsperRating <-factor(pf$ProsperRating..Alpha,
                        levels = c('AA', 'A', 'B', 'C', 'D', 'E', 'HR', 'NA'))
pf$ProsperScore = factor(pf$ProsperScore)
```

12B PROSPER DATA vs BORROWER RATE vs PROSPER RATE

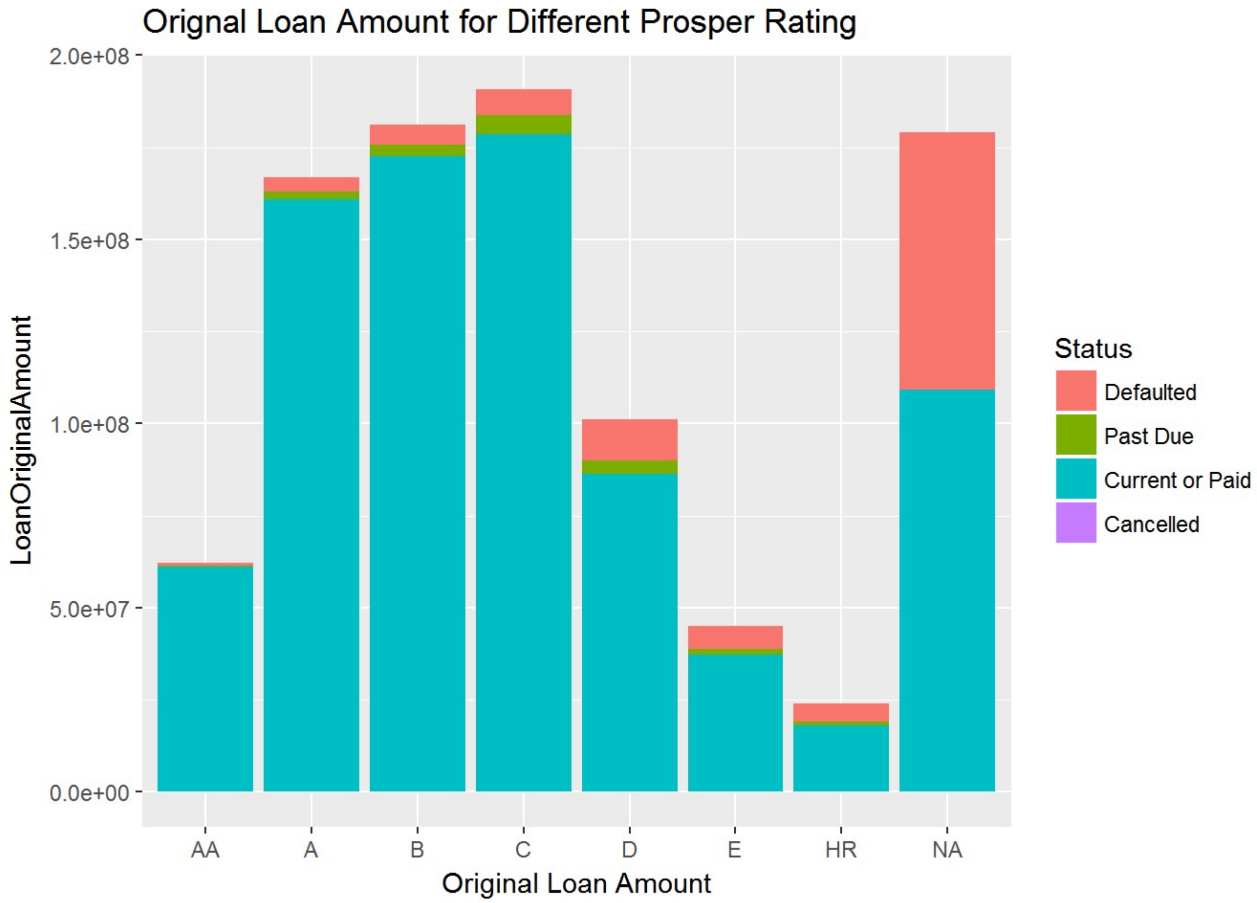
```
pf$ProsperRating.alpha <- factor(pf$ProsperRating.alpha)
ggplot(data = pf, aes(x = ProsperRating.alpha, y = BorrowerRate)) +
        geom_boxplot() +
        xlab("Prosper Rating") +
        ggtitle("Borrower Rate for Different Prosper Rating")
```



The better Prosper Rating means better rating. This shows that the better the Prosper Rate, the lower the prosper rating.

13B LOAN STATUS PER RATING

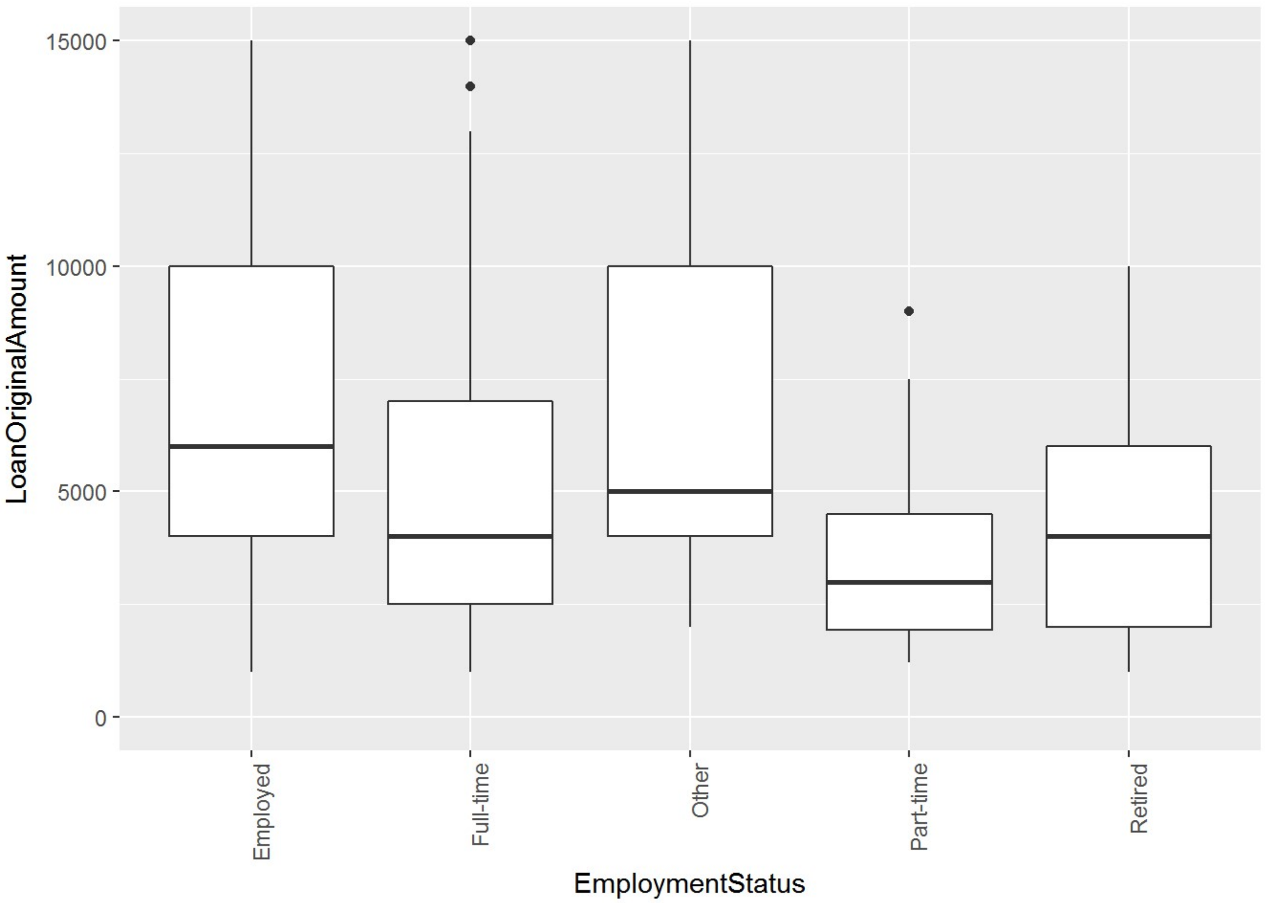
```
# create a new variable summarizing the result of each loan
pf <- pf %>% mutate(Status = ifelse(LoanStatus %in%
                     c("Chargedoff", "Defaulted"), 0,
                     ifelse(LoanStatus %in%
                     c("Completed", "Current", "FinalPaymentInProgress"), 2,
                     ifelse(LoanStatus %in%
                     "Cancelled",3,1))))

pf$Status <- factor(pf$Status, levels = 0:3,
                         labels = c("Defaulted",
                                    "Past Due",
                                    "Current or Paid",
                                    "Cancelled"))

ggplot(data = arrange(pf,Status), aes(x = ProsperRating.alpha,
                    y = LoanOriginalAmount, fill = Status)) +
                    geom_bar(stat = "identity") +
                    xlab("Prosper Rating") +
                    xlab("Original Loan Amount") +
                    ggtitle("Orignal Loan Amount for Different Prosper Rating")
```



This chart tells me that AA loans have the lowest default rate. The other loan categories have a varying loan default rate. Also, the NA loans have the largest default and he least amount of loans categorized as NA. This tells me that Prosper should require all the fields in order to avoid a high default amount.

14B BORROWER PROFILE - EMPLOYMENT STATUS ~ LOAN ORIGINAL AMOUNT

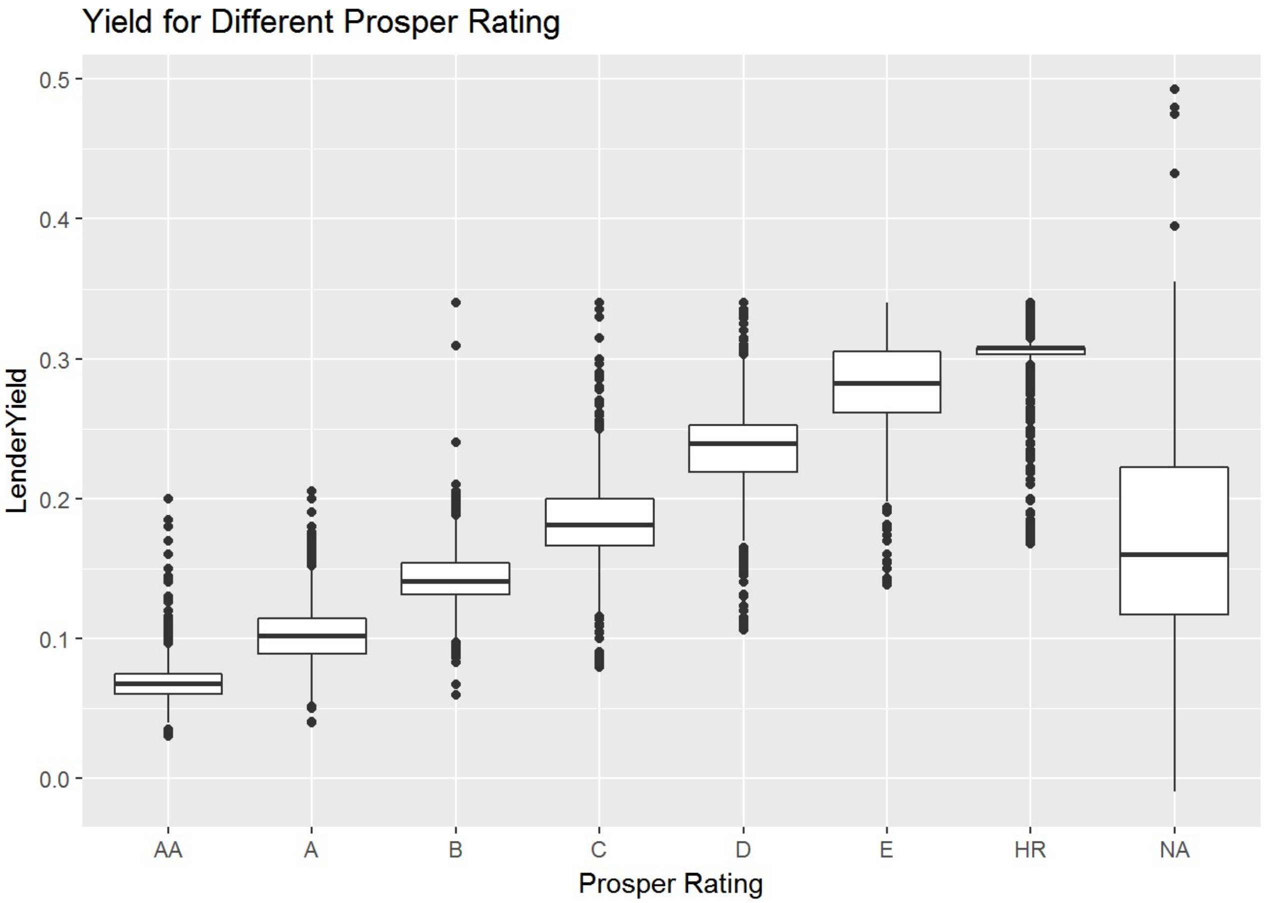
```
ggplot(aes(x = EmploymentStatus, y = LoanOriginalAmount), data = na.omit(pf)) +
                          geom_boxplot() +
                          scale_y_continuous(limits = c(0,15000)) +
                          theme(axis.text.x = element_text(angle = 90, hjust =
1))
```



This chart excludes monthly income over $9,000 and no income. Nothing significant stands out in this chart. This tells me that Prosper needs to clarify this data field. For example, you can be “Employed” and “Full-time”. I’m also curious what the “other” employment status means.

15B INVESTOR PROFILE - LENDER YIELD ~ PROSPER RATING

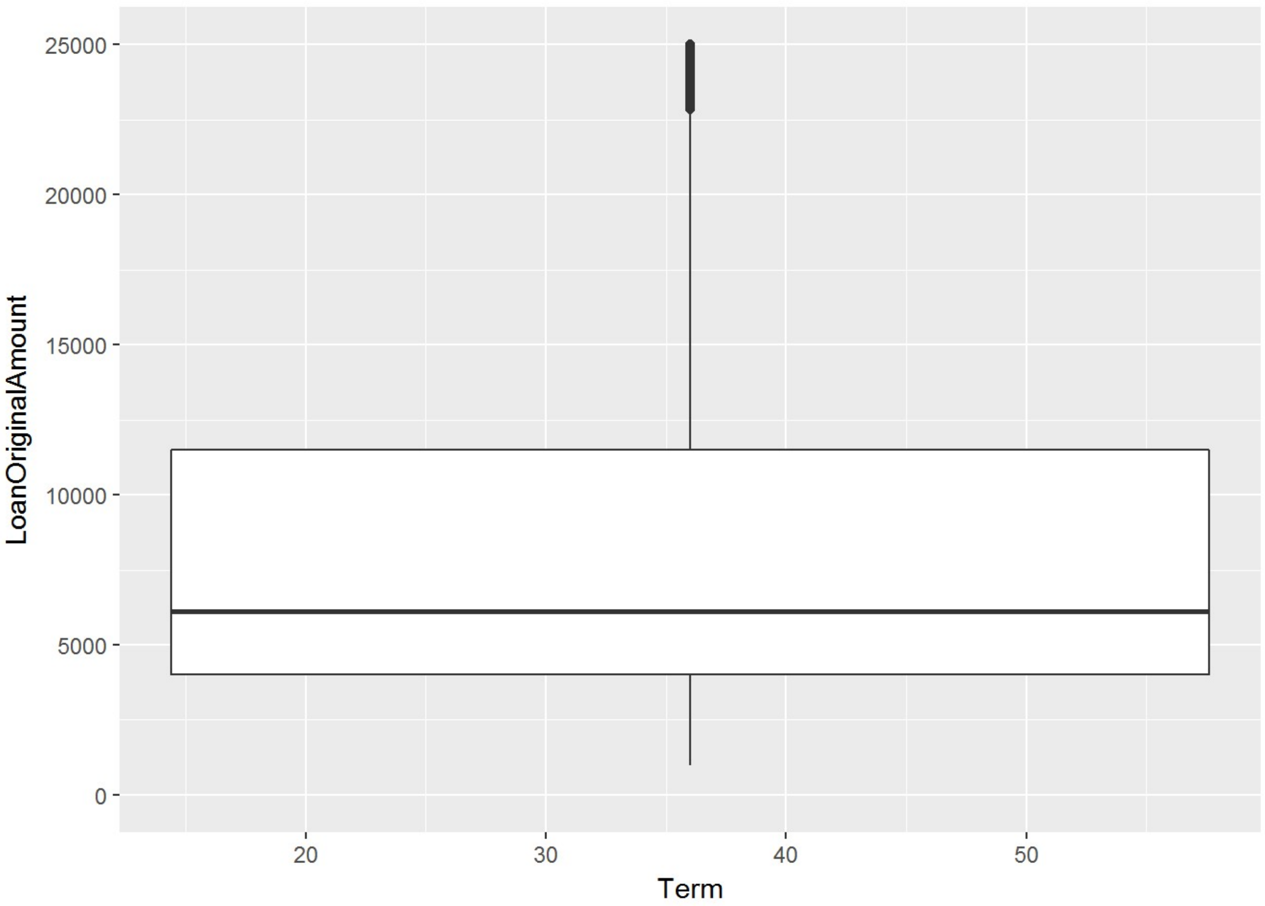
```
pf$ProsperRating.alpha = factor(pf$ProsperRating..Alpha.,
                                 levels = c("AA","A","B","C","D","E","HR","N
A"))
ggplot(data = pf, aes(x = ProsperRating.alpha, y = LenderYield)) +
        geom_boxplot() +
                               xlab("Prosper Rating") +
          ggtitle("Yield for Different Prosper Rating")
```



This chart doesn’t show many anything incredibly interesting. It shows that the worse the Prosper Rating, the higher the Lender Yield.

16B INVESTOR PROFILE - LOAN ORIGINAL AMOUNT ~ TERM

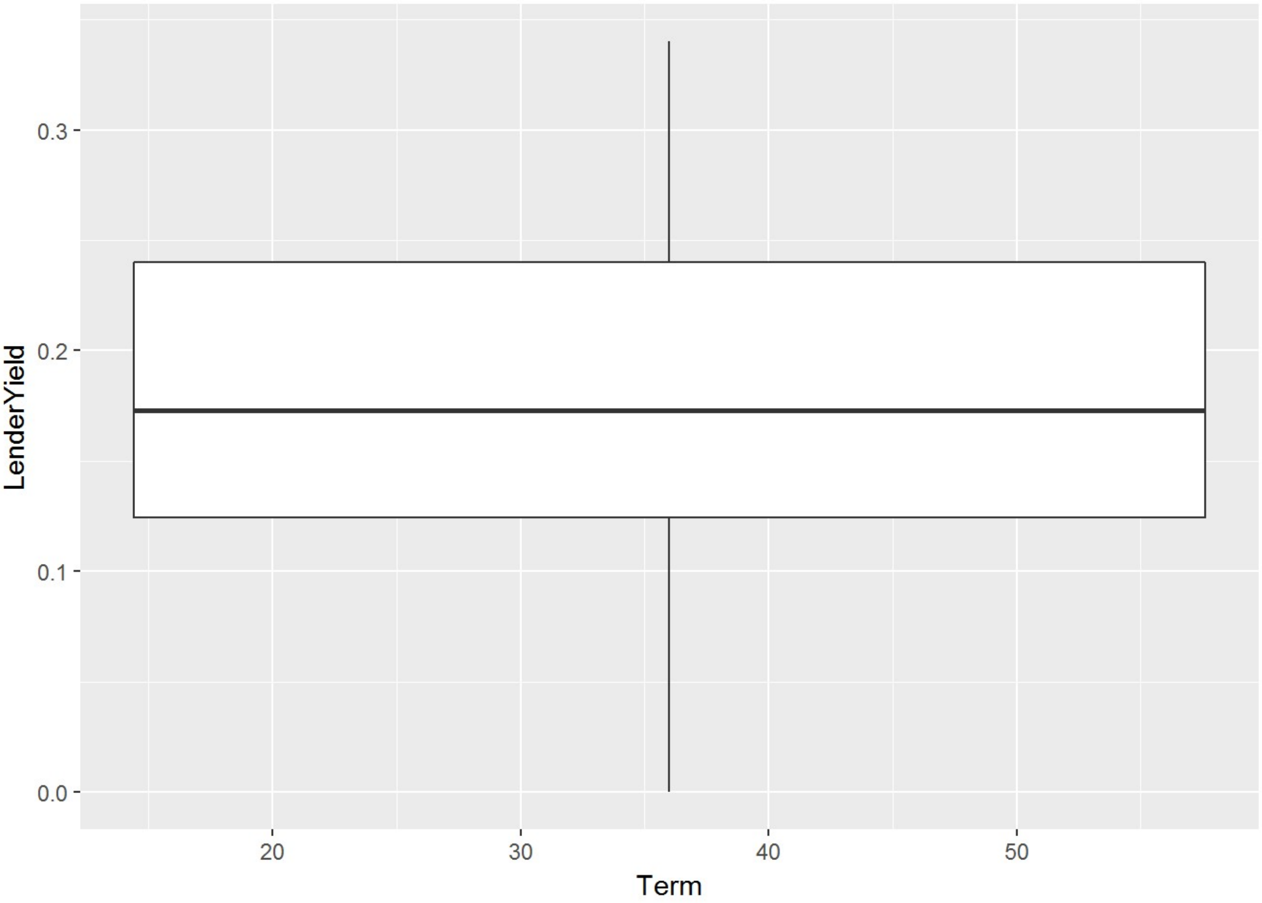
```
##26. Investor Profile - LoanOriginal Amount ~ Term
ggplot(aes(y = LoanOriginalAmount, x = Term), data = pf) +
                                             geom_boxplot() +
                                             scale_y_continuous(
    limits = c(0, quantile(pf$LoanOriginalAmount, 0.99, na.rm = TRUE)))
```



This chart shows the majority of loans 36-month term. The Loan original amount is significantly higher for 60 months term. This tells me that when people borrow more money, they spread out the loan terms.

17B INVESTOR PROFILE - LENDER YILED ~ TERM

```
ggplot(aes(y = LenderYield, x = Term), data = pf) +
                                             geom_boxplot() +
                                             scale_y_continuous(
    limits = c(0, quantile(pf$LenderYield, 0.99, na.rm = TRUE)))
```

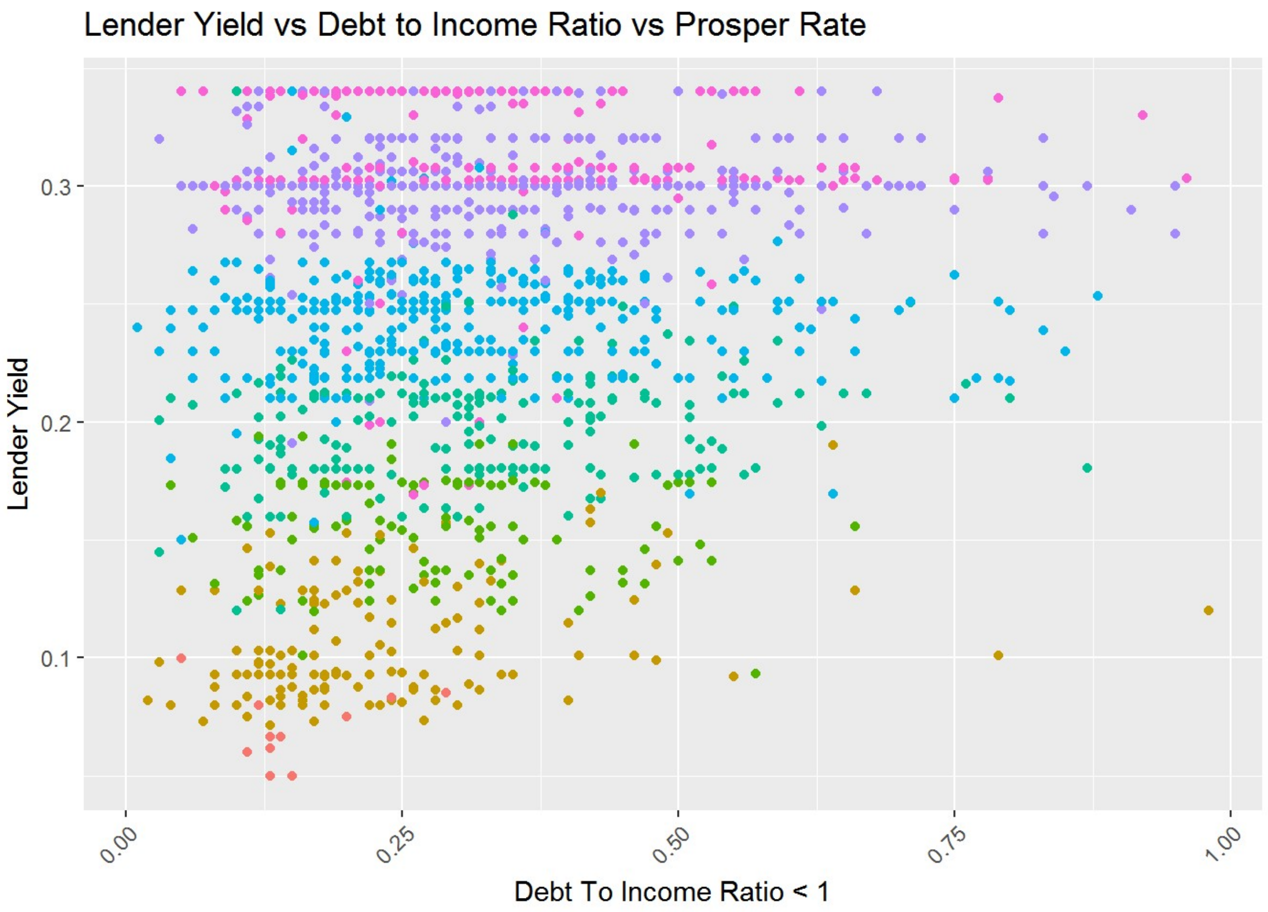


This chart doesn’t tell me anything new about the term, lender yield or prosper rating.

MULTIVARIATE PLOT & ANAYLSIS SECTION

18M DEBT TO INCOME RATIO - PROSPER RATING - LENDER YIELD

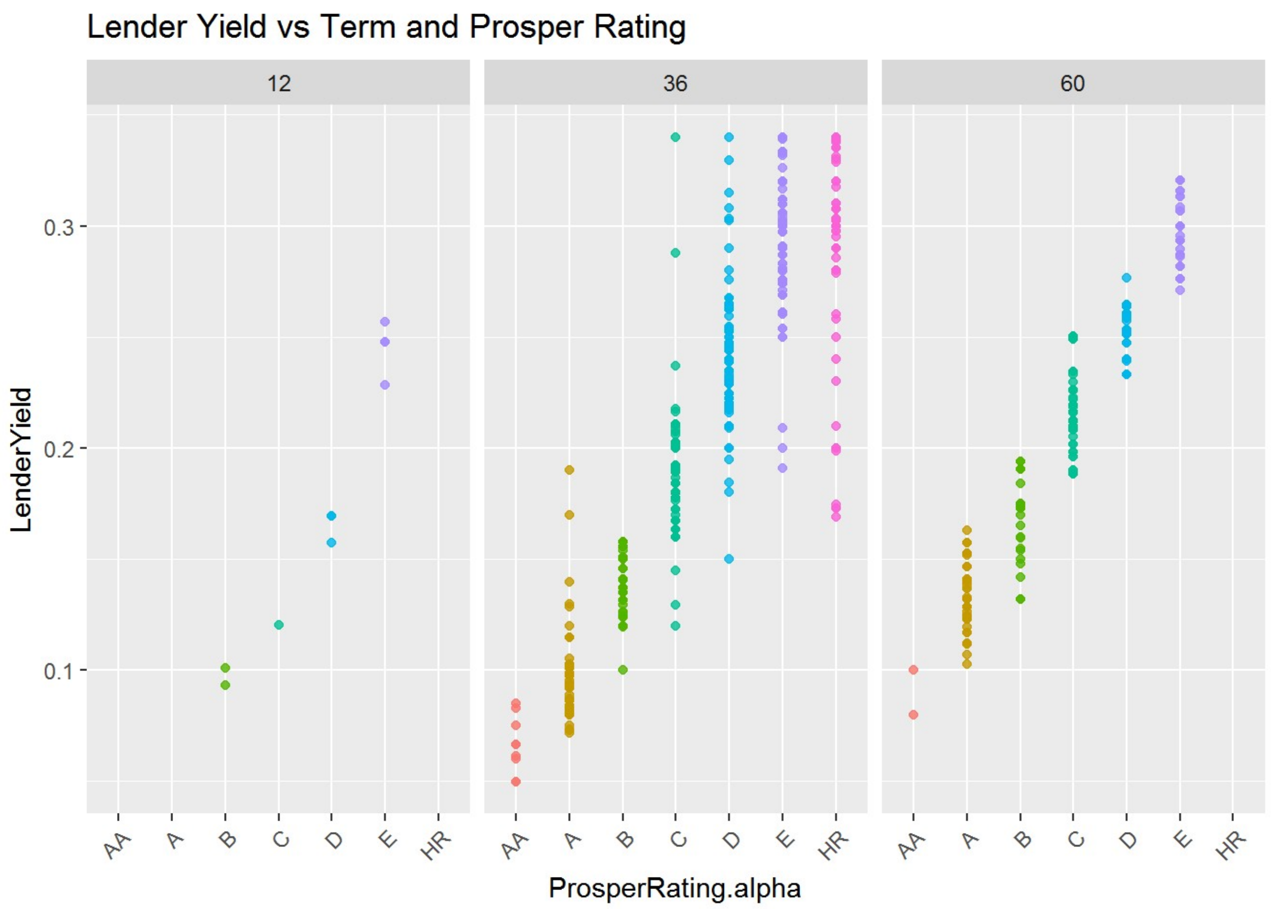
```
ggplot(aes(x= DebtToIncomeRatio, y=LenderYield, color=ProsperRating.alpha),
    data=na.omit(filter(pf, DebtToIncomeRatio < 1))) +
    geom_point(alpha = 1) +
    #scale_y_log10() +
    #facet_grid(.~ ProsperRating.alpha ) +
    theme(legend.position = "none",axis.text.x = element_text(angle = 45, hjus
t = 1))+
    ggtitle("Lender Yield vs Debt to Income Ratio vs Prosper Rate") +
    xlab ("Debt To Income Ratio < 1") +
    ylab ("Lender Yield") +
    scale_fill_discrete(name = "Prosper Rating")
```



This chart shows the coorelation of the Lender Yield, the Prosper Rating and the Debt-To-Income Ratio.

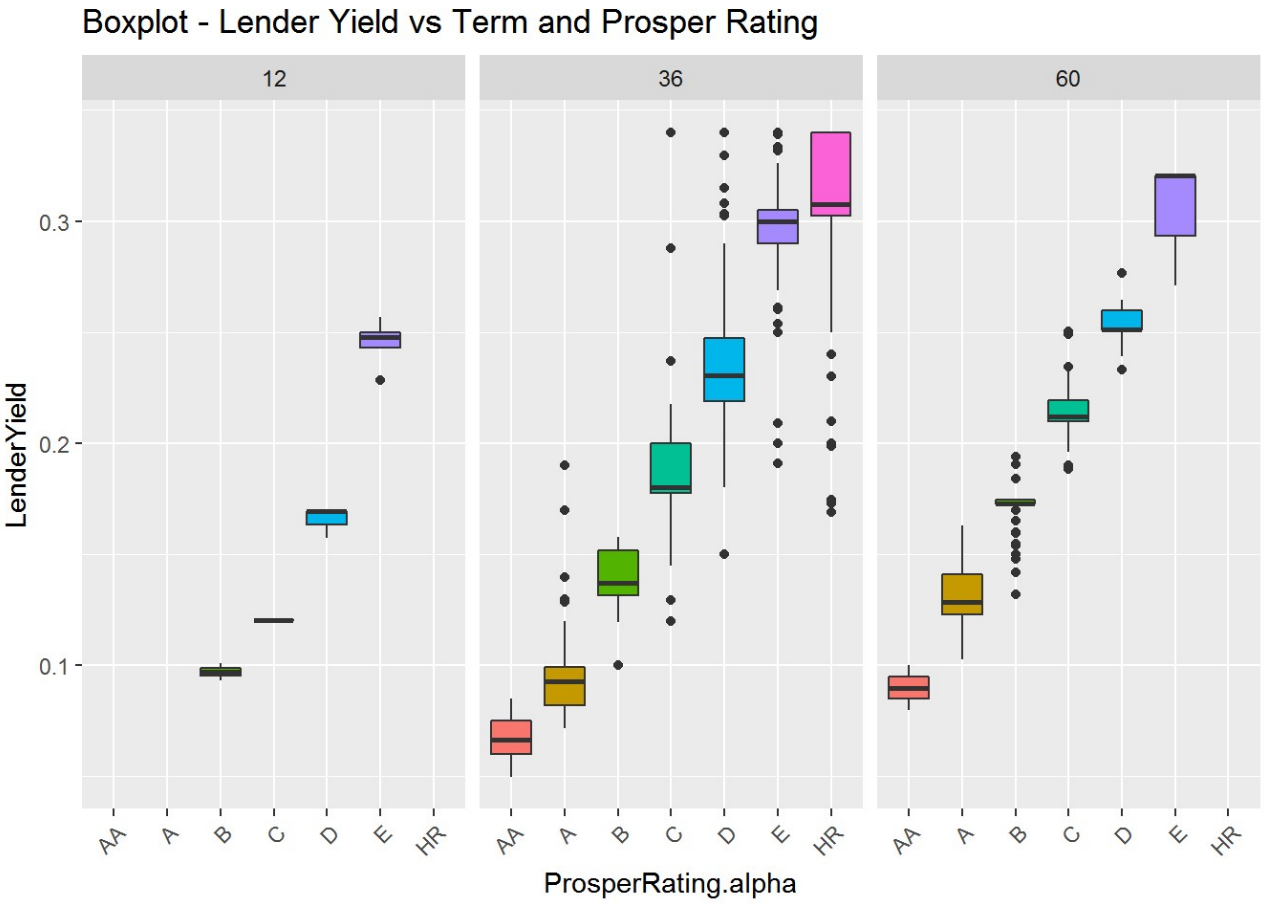
19M LENDER YIELD vs PROSPER RATE vs TERM

```
plot1 <- ggplot(aes(x= ProsperRating.alpha, y=LenderYield,
                                  color=ProsperRating.alpha),
       data=na.omit(filter(pf, DebtToIncomeRatio < 1))) +
    geom_point(alpha = 0.8) +
    facet_grid( .~ Term) +
    theme(legend.position = "none", axis.text.x = element_text(angle = 45, hjus
t = 1))+
    ggtitle("Lender Yield vs Term and Prosper Rating")
grid.arrange(plot1)
```



20M BOXPLOT - LENDER YIELD vs PROSPER RATE vs TERM

```
plot2 <- ggplot(aes(x= ProsperRating.alpha, y= LenderYield ),
       data=na.omit(filter(pf, DebtToIncomeRatio < 1))) +
    geom_boxplot(aes(fill = ProsperRating.alpha)) +
    facet_grid( .~ Term    ) +
    theme(legend.position = "none", axis.text.x = element_text(angle = 45, hjus
t = 1))+
    ggtitle("Boxplot - Lender Yield vs Term and Prosper Rating")
grid.arrange(plot2)
```



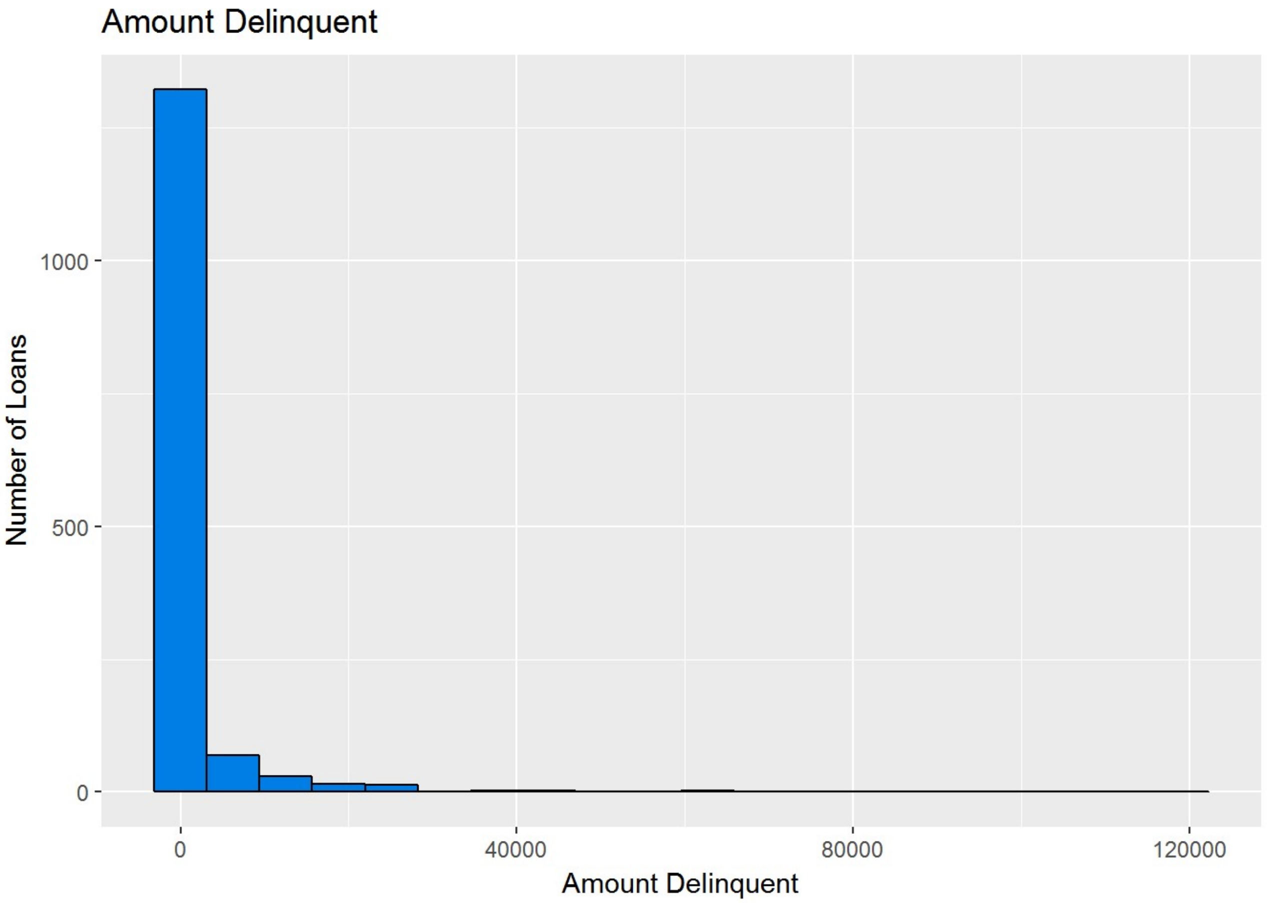
The chart looks at the term, lender yield and prosper rating. The majority of loans choose 36-month ter where the yield is higher.

FINAL PLOTS & SUMMARY

My favorite plots are 3U (Amount Delinquent), 4U (Prosper Score of Borrower), 7U (Borrower's purpose of loan) and 13B (Loan Status for Different Prosper Rating). These final charts tell me that Prosper needs to collect information about the purpose of the loan for all applicants. To remain profitable, Prosper Loan needs to find ways to less their amount of deliquent loans.

Final plot - 3U AMOUNT DELINQUENT

```
ggplot(data = na.omit(pf), aes(AmountDelinquent)) +
  geom_histogram(aes(fill = AmountDelinquent), color = "black", fill = '#007EE
5',bins=20) +
  ggtitle('Amount Delinquent') +
  xlab('Amount Delinquent') +
  ylab('Number of Loans')
```

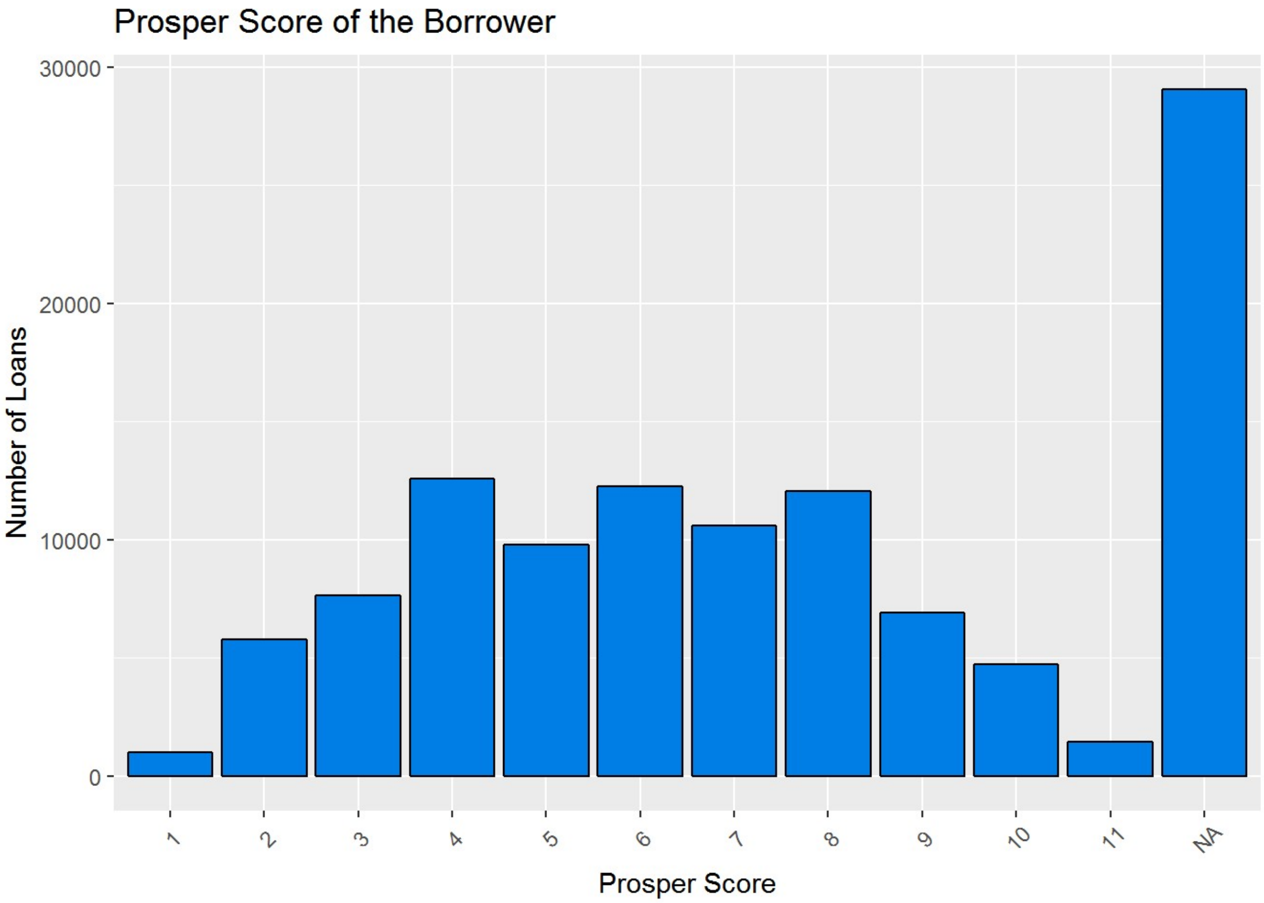


```
summary(pf$AmountDelinquent)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's 
##      0.0      0.0      0.0    984.5      0.0 463900.0     7622
```

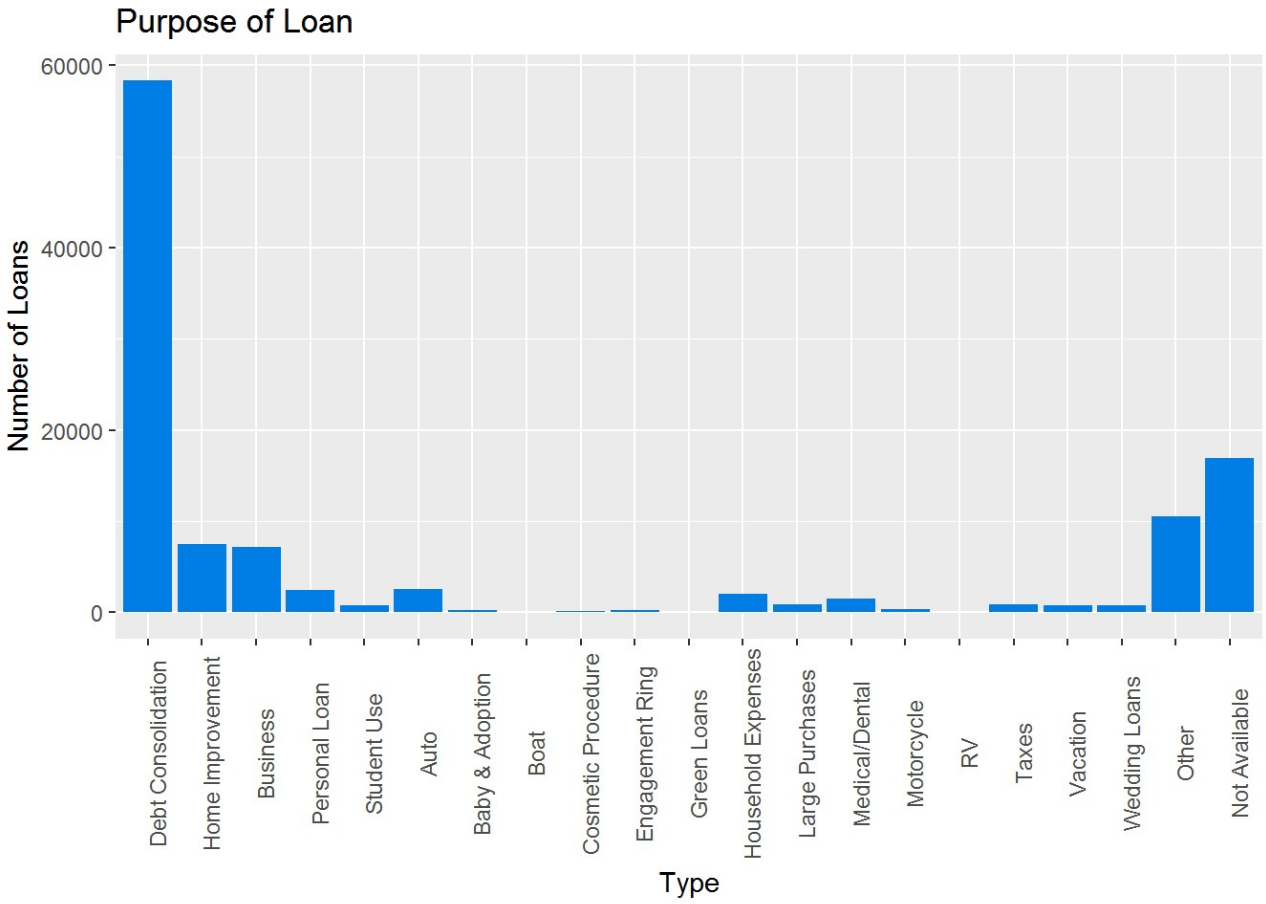
Final Plot - 4U SCORE DISTRIBUTION

```
ggplot(data = pf, aes(ProsperScore)) +
  geom_bar(color="black", fill = '#007EE5') +
  ggtitle('Prosper Score of the Borrower') +
  xlab('Prosper Score') +
  theme(axis.text.x = element_text(angle = 45, vjust = 0.6)) +
  ylab('Number of Loans')
```


Prosper Score of the Borrower
Number of Loans
30000
20000
10000
0
1
2
3
4
5
6
7
8
9
10
11
NA
Prosper Score

Final Plot - 7U BORROWER’S PURPOSE OF LOAN

```
x <- c('Debt Consolidation',
                                'Home Improvement','Business',
                                 'Personal Loan',
                                 'Student Use',
                                 'Auto',
                                 'Baby & Adoption',
                                 'Boat',
                                 'Cosmetic Procedure',
                                 'Engagement Ring',
                                 'Green Loans',
                                 'Household Expenses',
                                 'Large Purchases',
                                 'Medical/Dental',
                                 'Motorcycle', 'RV',
                                 'Taxes', 'Vacation',
                                 'Wedding Loans',
                                 'Other',
                                 'Not Available')

pf$ListingCategory <- factor(pf$ListingCategory..numeric., levels = c(1:6,8:20,
7,0), labels = x)

ggplot(data = pf, aes(x=ListingCategory)) +
  geom_bar(aes(y=..count..), size = 3, fill = '#007EE5', stat="count") +
  ggtitle('Purpose of Loan') +
  xlab('Type') +
  ylab('Number of Loans') +
  theme(axis.text.x = element_text(angle = 90))
```

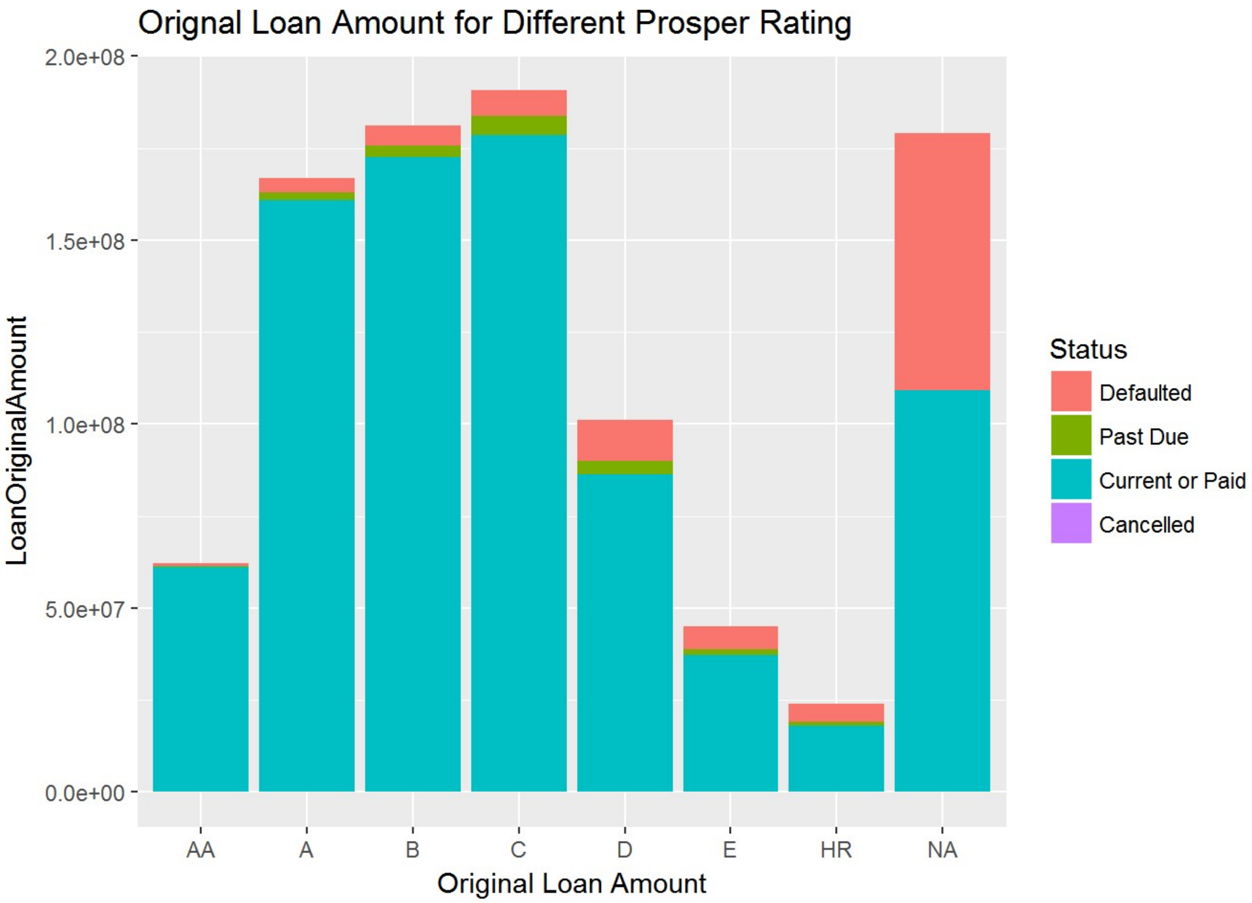


```
summary(pf$ListingCategory)
```

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

Final Plot - 13B LOAN STATUS PER RATING

```
# create a new variable summarizing the result of each loan
pf <- pf %>% mutate(Status = ifelse(LoanStatus %in%
                     c("Chargedoff", "Defaulted"), 0,
                     ifelse(LoanStatus %in%
                     c("Completed", "Current", "FinalPaymentInProgress"), 2,
                     ifelse(LoanStatus %in%
                     "Cancelled",3,1))))

pf$Status <- factor(pf$Status, levels = 0:3,
                         labels = c("Defaulted",
                                    "Past Due",
                                    "Current or Paid",
                                    "Cancelled"))

ggplot(data = arrange(pf,Status), aes(x = ProsperRating.alpha,
                    y = LoanOriginalAmount, fill = Status)) +
                    geom_bar(stat = "identity") +
                    xlab("Prosper Rating") +
                    xlab("Original Loan Amount") +
                    ggtitle("Orignal Loan Amount for Different Prosper Rating")
```



REFLECTION

1. What is the structure of your dataset?

The dataset has 113,937 observations and 81 variables. The dates ranges from 2005 through 2014. The types of variables are interger, numeric, date, and factor. The 88 variables could be split into two categories related to the borrower and investor.

2. What are the main features of interest in the dataset?

The dataset variables can be split into two for the borrower and lender. For the borrower, the variables of interest are Prosper Rating (numeric & alphabet) because it is an indicator of the quality of borrowers. Other variables of interest are debt-to-income ratio, verifiable income and credit grade. For the lender perspective, lender yield and estimated return are variables of interest.

3. What other features in the dataset do you think will help support your investigation into your features of interest?

I'm interested in comparing the ProsperScore to the Estimated Return/Loss. I'm curious to learn if their rating criteria has been modified throughout the years. There were approximately 28,000 loans that had NA for a ProsperScore. It would be helpful to investigate the criteria that makes up the ProsperScore.