R Notebook for Prosper Loan Data

Code

A. PREPARING RSTUDIO AND THE DATA SET

A1. Installing the packages as instructed in the rubric.

Hide

```
library("ggplot2")
library("knitr")
library("dplyr")
library(gridExtra)
```

A2. Opening the Data Set

Hide

```
getwd()
```

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

Hide

```
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"
```

A3. Running the data & summary files

Hide

```
data(pf)
```

```
data set pf not found
```

Hide

```
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
2013-10-02 17:20:16.550000000:     6          :84984
2013-08-28 20:31:41.107000000:     4   C      : 5649
2013-09-08 09:27:44.853000000:     4   D      : 5153
2013-12-06 05:43:13.830000000:     4   B      : 4389
2013-12-06 11:44:58.283000000:     4   AA     : 3509
2013-08-21 07:25:22.360000000:     3   HR     : 3508
(Other)                       :113912   (Other): 6745
     Term                        LoanStatus
Min.   :12.00   Current              :56576
1st Qu.:36.00   Completed            :38074
Median :36.00   Chargedoff           :11992
Mean   :40.83   Defaulted            : 5018
3rd Qu.:36.00   Past Due (1-15 days) :  806
Max.   :60.00   Past Due (31-60 days):  363
                (Other)              : 1108
               ClosedDate     BorrowerAPR       BorrowerRate
                   :58848   Min.   :0.00653   Min.   :0.0000
2014-03-04 00:00:00:  105   1st Qu.:0.15629   1st Qu.:0.1340
2014-02-19 00:00:00:  100   Median :0.20976   Median :0.1840
2014-02-11 00:00:00:   92   Mean   :0.21883   Mean   :0.1928
2012-10-30 00:00:00:   81   3rd Qu.:0.28381   3rd Qu.:0.2500
2013-02-26 00:00:00:   78   Max.   :0.51229   Max.   :0.4975
(Other)            :54633   NA's   :25
 LenderYield      EstimatedEffectiveYield EstimatedLoss
Min.   :-0.0100   Min.   :-0.183          Min.   :0.005
1st Qu.: 0.1242   1st Qu.: 0.116          1st Qu.:0.042
Median : 0.1730   Median : 0.162          Median :0.072
Mean   : 0.1827   Mean   : 0.169          Mean   :0.080
3rd Qu.: 0.2400   3rd Qu.: 0.224          3rd Qu.:0.112
Max.   : 0.4925   Max.   : 0.320          Max.   :0.366
                  NA's   :29084           NA's   :29084
EstimatedReturn  ProsperRating..numeric. ProsperRating..Alpha.
Min.   :-0.183   Min.   :1.000                  :29084
1st Qu.: 0.074   1st Qu.:3.000          C       :18345
Median : 0.092   Median :4.000          B       :15581
Mean   : 0.096   Mean   :4.072          A       :14551
3rd Qu.: 0.117   3rd Qu.:5.000          D       :14274
Max.   : 0.284   Max.   :7.000          E       : 9795
NA's   :29084    NA's   :29084          (Other):12307
```

```
 ProsperScore   ListingCategory..numeric. BorrowerState
Min.   : 1.00   Min.   : 0.000            CA     :14717
1st Qu.: 4.00   1st Qu.: 1.000            TX     : 6842
Median : 6.00   Median : 1.000            NY     : 6729
Mean   : 5.95   Mean   : 2.774            FL     : 6720
3rd Qu.: 8.00   3rd Qu.: 3.000            IL     : 5921
Max.   :11.00   Max.   :20.000                   : 5515
NA's   :29084                             (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
                      :   697     Min.   : 0.00
1993-12-01 00:00:00:   185     1st Qu.: 7.00
1994-11-01 00:00:00:   178     Median :10.00
1995-11-01 00:00:00:   168     Mean   :10.32
1990-04-01 00:00:00:   161     3rd Qu.:13.00
1995-03-01 00:00:00:   159     Max.   :59.00
(Other)               :112389     NA's   :7604
OpenCreditLines TotalCreditLinespast7years
```

```
Min.   : 0.00   Min.   :  2.00
1st Qu.: 6.00   1st Qu.: 17.00
Median : 9.00   Median : 25.00
Mean   : 9.26   Mean   : 26.75
3rd Qu.:12.00   3rd Qu.: 35.00
Max.   :54.00   Max.   :136.00
NA's   :7604    NA's   :697
OpenRevolvingAccounts OpenRevolvingMonthlyPayment
Min.   : 0.00         Min.   :    0.0
1st Qu.: 4.00         1st Qu.:  114.0
Median : 6.00         Median :  271.0
Mean   : 6.97         Mean   :  398.3
3rd Qu.: 9.00         3rd Qu.:  525.0
Max.   :51.00         Max.   :14985.0

InquiriesLast6Months TotalInquiries    CurrentDelinquencies
Min.   :  0.000      Min.   :  0.000   Min.   : 0.0000
1st Qu.:  0.000      1st Qu.:  2.000   1st Qu.: 0.0000
Median :  1.000      Median :  4.000   Median : 0.0000
Mean   :  1.435      Mean   :  5.584   Mean   : 0.5921
3rd Qu.:  2.000      3rd Qu.:  7.000   3rd Qu.: 0.0000
Max.   :105.000      Max.   :379.000   Max.   :83.0000
NA's   :697          NA's   :1159      NA's   :697
AmountDelinquent   DelinquenciesLast7Years
Min.   :     0.0   Min.   : 0.000
1st Qu.:     0.0   1st Qu.: 0.000
Median :     0.0   Median : 0.000
Mean   :   984.5   Mean   : 4.155
3rd Qu.:     0.0   3rd Qu.: 3.000
Max.   :463881.0   Max.   :99.000
NA's   :7622       NA's   :990
PublicRecordsLast10Years PublicRecordsLast12Months
Min.   : 0.0000          Min.   : 0.000
1st Qu.: 0.0000          1st Qu.: 0.000
Median : 0.0000          Median : 0.000
Mean   : 0.3126          Mean   : 0.015
3rd Qu.: 0.0000          3rd Qu.: 0.000
Max.   :38.0000          Max.   :20.000
NA's   :697              NA's   :7604
RevolvingCreditBalance BankcardUtilization
Min.   :      0        Min.   :0.000
1st Qu.:   3121        1st Qu.:0.310
Median :   8549        Median :0.600
Mean   :  17599        Mean   :0.561
3rd Qu.:  19521        3rd Qu.:0.840
Max.   :1435667        Max.   :5.950
NA's   :7604           NA's   :7604
AvailableBankcardCredit  TotalTrades
Min.   :     0          Min.   :  0.00
```

```
 1st Qu.:   880          1st Qu.: 15.00
 Median :  4100          Median : 22.00
 Mean   : 11210          Mean   : 23.23
 3rd Qu.: 13180          3rd Qu.: 30.00
 Max.   :646285          Max.   :126.00
 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
 Min.   :      0     CB1B37030986463208432A1:     6
 1st Qu.:   3200     2DEE3698211017519D7333F:     4
 Median :   4667     9F4B37043517554537C364C:     4
 Mean   :   5608     D895370150591392337ED6D:     4
 3rd Qu.:   6825     E6FB37073953690388BC56D:     4
 Max.   :1750003     0D8F37036734373301ED419:     3
                     (Other)               :113912
 TotalProsperLoans TotalProsperPaymentsBilled
 Min.   :0.00      Min.   :  0.00
 1st Qu.:1.00      1st Qu.:  9.00
 Median :1.00      Median : 16.00
 Mean   :1.42      Mean   : 22.93
 3rd Qu.:2.00      3rd Qu.: 33.00
 Max.   :8.00      Max.   :141.00
 NA's   :91852     NA's   :91852
 OnTimeProsperPayments ProsperPaymentsLessThanOneMonthLate
 Min.   :  0.00        Min.   : 0.00
 1st Qu.:  9.00        1st Qu.: 0.00
 Median : 15.00        Median : 0.00
 Mean   : 22.27        Mean   : 0.61
 3rd Qu.: 32.00        3rd Qu.: 0.00
 Max.   :141.00        Max.   :42.00
 NA's   :91852         NA's   :91852
 ProsperPaymentsOneMonthPlusLate ProsperPrincipalBorrowed
 Min.   : 0.00                   Min.   :    0
 1st Qu.: 0.00                   1st Qu.: 3500
```

```
Median : 0.00                   Median : 6000
Mean   : 0.05                   Mean   : 8472
3rd Qu.: 0.00                   3rd Qu.:11000
Max.   :21.00                   Max.   :72499
NA's   :91852                   NA's   :91852
ProsperPrincipalOutstanding ScorexChangeAtTimeOfListing
Min.   :    0               Min.   :-209.00
1st Qu.:    0               1st Qu.: -35.00
Median : 1627               Median :  -3.00
Mean   : 2930               Mean   :  -3.22
3rd Qu.: 4127               3rd Qu.:  25.00
Max.   :23451               Max.   : 286.00
NA's   :91852               NA's   :95009
LoanCurrentDaysDelinquent LoanFirstDefaultedCycleNumber
Min.   :   0.0            Min.   : 0.00
1st Qu.:   0.0            1st Qu.: 9.00
Median :   0.0            Median :14.00
Mean   : 152.8            Mean   :16.27
3rd Qu.:   0.0            3rd Qu.:22.00
Max.   :2704.0            Max.   :44.00
                          NA's   :96985
LoanMonthsSinceOrigination   LoanNumber     LoanOriginalAmount
Min.   :  0.0              Min.   :     1   Min.   : 1000
1st Qu.:  6.0              1st Qu.: 37332   1st Qu.: 4000
Median : 21.0              Median : 68599   Median : 6500
Mean   : 31.9              Mean   : 69444   Mean   : 8337
3rd Qu.: 65.0              3rd Qu.:101901   3rd Qu.:12000
Max.   :100.0              Max.   :136486   Max.   :35000

         LoanOriginationDate LoanOriginationQuarter
2014-01-22 00:00:00:   491   Q4 2013:14450
2013-11-13 00:00:00:   490   Q1 2014:12172
2014-02-19 00:00:00:   439   Q3 2013: 9180
2013-10-16 00:00:00:   434   Q2 2013: 7099
2014-01-28 00:00:00:   339   Q3 2012: 5632
2013-09-24 00:00:00:   316   Q2 2012: 5061
(Other)             :111428   (Other):60343
                  MemberKey      MonthlyLoanPayment
63CA34120866140639431C9:     9   Min.   :   0.0
160833647449334457E57FB9:     8   1st Qu.: 131.6
3A2F3380477699707C81385:     8   Median : 217.7
4D9C34033020477712AD0CDD:     8   Mean   : 272.5
739C338135235294782AE75:     8   3rd Qu.: 371.6
7E1733653050264822FAA3D:     8   Max.   :2251.5
(Other)                :113888
LP_CustomerPayments LP_CustomerPrincipalPayments
Min.   :   -2.35    Min.   :    0.0
1st Qu.: 1005.76    1st Qu.:  500.9
Median : 2583.83    Median : 1587.5
```

```
Mean   : 4183.08    Mean   : 3105.5
3rd Qu.: 5548.40    3rd Qu.: 4000.0
Max.   :40702.39    Max.   :35000.0

LP_InterestandFees LP_ServiceFees    LP_CollectionFees
Min.   :   -2.35   Min.   :-664.87   Min.   :-9274.75
1st Qu.:  274.87   1st Qu.: -73.18   1st Qu.:    0.00
Median :  700.84   Median : -34.44   Median :    0.00
Mean   : 1077.54   Mean   : -54.73   Mean   :  -14.24
3rd Qu.: 1458.54   3rd Qu.: -13.92   3rd Qu.:    0.00
Max.   :15617.03   Max.   :  32.06   Max.   :    0.00

LP_GrossPrincipalLoss LP_NetPrincipalLoss
Min.   :  -94.2       Min.   : -954.5
1st Qu.:    0.0       1st Qu.:    0.0
Median :    0.0       Median :    0.0
Mean   :  700.4       Mean   :  681.4
3rd Qu.:    0.0       3rd Qu.:    0.0
Max.   :25000.0       Max.   :25000.0

LP_NonPrincipalRecoverypayments PercentFunded
Min.   :    0.00                Min.   :0.7000
1st Qu.:    0.00                1st Qu.:1.0000
Median :    0.00                Median :1.0000
Mean   :   25.14                Mean   :0.9986
3rd Qu.:    0.00                3rd Qu.:1.0000
Max.   :21117.90                Max.   :1.0125

Recommendations    InvestmentFromFriendsCount
Min.   : 0.00000   Min.   : 0.00000
1st Qu.: 0.00000   1st Qu.: 0.00000
Median : 0.00000   Median : 0.00000
Mean   : 0.04803   Mean   : 0.02346
3rd Qu.: 0.00000   3rd Qu.: 0.00000
Max.   :39.00000   Max.   :33.00000

InvestmentFromFriendsAmount   Investors
Min.   :    0.00            Min.   :   1.00
1st Qu.:    0.00            1st Qu.:   2.00
Median :    0.00            Median :  44.00
Mean   :   16.55            Mean   :  80.48
3rd Qu.:    0.00            3rd Qu.: 115.00
Max.   :25000.00            Max.   :1189.00
```

A4. Does my data set over 1,000 observations? Are there at least 8 different variables?

Hide

```
dim(pf)
```

```
[1] 113937     81
```

113,937 observations with 81 variables

A5. Does my data set contain at leasT one categorical variable?

Hide

```
lapply(pf,class)
```

```
$ListingKey
[1] "factor"

$ListingNumber
[1] "integer"

$ListingCreationDate
[1] "factor"

$CreditGrade
[1] "factor"

$Term
[1] "integer"

$LoanStatus
[1] "factor"

$ClosedDate
[1] "factor"

$BorrowerAPR
[1] "numeric"

$BorrowerRate
[1] "numeric"

$LenderYield
[1] "numeric"

$EstimatedEffectiveYield
[1] "numeric"

$EstimatedLoss
[1] "numeric"

$EstimatedReturn
[1] "numeric"

$ProsperRating..numeric.
[1] "integer"

$ProsperRating..Alpha.
[1] "factor"

$ProsperScore
[1] "numeric"
```

```
$ListingCategory..numeric.
[1] "integer"

$BorrowerState
[1] "factor"

$Occupation
[1] "factor"

$EmploymentStatus
[1] "factor"

$EmploymentStatusDuration
[1] "integer"

$IsBorrowerHomeowner
[1] "factor"

$CurrentlyInGroup
[1] "factor"

$GroupKey
[1] "factor"

$DateCreditPulled
[1] "factor"

$CreditScoreRangeLower
[1] "integer"

$CreditScoreRangeUpper
[1] "integer"

$FirstRecordedCreditLine
[1] "factor"

$CurrentCreditLines
[1] "integer"

$OpenCreditLines
[1] "integer"

$TotalCreditLinespast7years
[1] "integer"

$OpenRevolvingAccounts
[1] "integer"

$OpenRevolvingMonthlyPayment
```

```
[1] "numeric"

$InquiriesLast6Months
[1] "integer"

$TotalInquiries
[1] "numeric"

$CurrentDelinquencies
[1] "integer"

$AmountDelinquent
[1] "numeric"

$DelinquenciesLast7Years
[1] "integer"

$PublicRecordsLast10Years
[1] "integer"

$PublicRecordsLast12Months
[1] "integer"

$RevolvingCreditBalance
[1] "numeric"

$BankcardUtilization
[1] "numeric"

$AvailableBankcardCredit
[1] "numeric"

$TotalTrades
[1] "numeric"

$TradesNeverDelinquent..percentage.
[1] "numeric"

$TradesOpenedLast6Months
[1] "numeric"

$DebtToIncomeRatio
[1] "numeric"

$IncomeRange
[1] "factor"

$IncomeVerifiable
[1] "factor"
```

```
$StatedMonthlyIncome
[1] "numeric"

$LoanKey
[1] "factor"

$TotalProsperLoans
[1] "integer"

$TotalProsperPaymentsBilled
[1] "integer"

$OnTimeProsperPayments
[1] "integer"

$ProsperPaymentsLessThanOneMonthLate
[1] "integer"

$ProsperPaymentsOneMonthPlusLate
[1] "integer"

$ProsperPrincipalBorrowed
[1] "numeric"

$ProsperPrincipalOutstanding
[1] "numeric"

$ScorexChangeAtTimeOfListing
[1] "integer"

$LoanCurrentDaysDelinquent
[1] "integer"

$LoanFirstDefaultedCycleNumber
[1] "integer"

$LoanMonthsSinceOrigination
[1] "integer"

$LoanNumber
[1] "integer"

$LoanOriginalAmount
[1] "integer"

$LoanOriginationDate
[1] "factor"
```

```
$LoanOriginationQuarter
[1] "factor"

$MemberKey
[1] "factor"

$MonthlyLoanPayment
[1] "numeric"

$LP_CustomerPayments
[1] "numeric"

$LP_CustomerPrincipalPayments
[1] "numeric"

$LP_InterestandFees
[1] "numeric"

$LP_ServiceFees
[1] "numeric"

$LP_CollectionFees
[1] "numeric"

$LP_GrossPrincipalLoss
[1] "numeric"

$LP_NetPrincipalLoss
[1] "numeric"

$LP_NonPrincipalRecoverypayments
[1] "numeric"

$PercentFunded
[1] "numeric"

$Recommendations
[1] "integer"

$InvestmentFromFriendsCount
[1] "integer"

$InvestmentFromFriendsAmount
[1] "numeric"

$Investors
[1] "integer"
```

A6. List out the description of variables and types

Hide

```
str(pf)
```

```
'data.frame':   113937 obs. of  81 variables:
 $ ListingKey                          : Factor w/ 113066 levels "00003546482094
282EF90E5",..: 7180 7193 6647 6669 6686 6689 6699 6706 6687 6687 ...
 $ ListingNumber                       : int  193129 1209647 81716 658116 90946
4 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 978
34 ...
 $ CreditGrade                         : Factor w/ 9 levels "","A","AA","B",..:
5 1 8 1 1 1 1 1 1 1 ...
 $ Term                                : int  36 36 36 36 36 60 36 36 36 36 ...
 $ LoanStatus                          : Factor w/ 12 levels "Cancelled","Charge
doff",..: 3 4 3 4 4 4 4 4 4 4 ...
 $ ClosedDate                          : Factor w/ 2803 levels "","2005-11-25 0
0:00:00",..: 1138 1 1263 1 1 1 1 1 1 1 ...
 $ BorrowerAPR                         : num  0.165 0.12 0.283 0.125 0.246 ...
 $ BorrowerRate                        : num  0.158 0.092 0.275 0.0974 0.208
5 ...
 $ LenderYield                         : num  0.138 0.082 0.24 0.0874 0.1985 ...
 $ 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","AA","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/CP
A",..: 37 43 37 52 21 43 50 29 24 24 ...
 $ EmploymentStatus                    : Factor w/ 9 levels "","Employed",..: 9
2 4 2 2 2 2 2 2 2 ...
 $ EmploymentStatusDuration            : int  2 44 NA 113 44 82 172 103 269 26
9 ...
 $ 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 "","00343376901312
423168731",..: 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 978
88 ...
 $ CreditScoreRangeLower               : int  640 680 480 800 680 740 680 700 82
0 820 ...
 $ CreditScoreRangeUpper               : int  659 699 499 819 699 759 699 719 83
9 839 ...
```

```
 $ FirstRecordedCreditLine            : Factor w/ 11586 levels "","1947-08-24 0
0:00:00",..: 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 32 ...
 $ OpenRevolvingAccounts              : int  1 13 0 7 6 13 6 5 12 12 ...
 $ OpenRevolvingMonthlyPayment        : num  24 389 0 115 220 1410 214 101 219
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 0.1
3 0.11 0.11 ...
 $ AvailableBankcardCredit            : num  1500 10266 NA 30754 695 ...
 $ TotalTrades                        : num  11 29 NA 26 39 47 16 10 29 29 ...
 $ 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 0.2
7 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 "00003683605746
079487FF7",..: 100337 69837 46303 70776 71387 86505 91250 5425 908 908 ...
 $ TotalProsperLoans                  : int  NA NA NA NA 1 NA NA NA NA NA ...
 $ TotalProsperPaymentsBilled         : int  NA NA NA NA 11 NA NA NA NA NA ...
 $ OnTimeProsperPayments              : int  NA NA NA NA 11 NA NA NA NA NA ...
 $ ProsperPaymentsLessThanOneMonthLate: int  NA NA NA NA 0 NA NA NA NA NA ...
 $ ProsperPaymentsOneMonthPlusLate    : int  NA NA NA NA 0 NA NA NA NA NA ...
 $ 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 NA ...
 $ LoanCurrentDaysDelinquent          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ LoanFirstDefaultedCycleNumber      : int  NA NA NA NA NA NA NA NA NA NA ...
 $ LoanMonthsSinceOrigination         : int  78 0 86 16 6 3 11 10 3 3 ...
 $ LoanNumber                         : int  19141 134815 6466 77296 102670 123
257 88353 90051 121268 121268 ...
 $ LoanOriginalAmount                 : int  9425 10000 3001 10000 15000 15000
3000 10000 10000 10000 ...
 $ LoanOriginationDate                : Factor w/ 1873 levels "2005-11-15 00:0
```

```
0:00",..: 426 1866 260 1535 1757 1821 1649 1666 1813 1813 ...
 $ LoanOriginationQuarter             : Factor w/ 33 levels "Q1 2006","Q1 200
7",..: 18 8 2 32 24 33 16 16 33 33 ...
 $ MemberKey                          : Factor w/ 90831 levels "000033976974133
87CAF966",..: 11071 10302 33781 54939 19465 48037 60448 40951 26129 26129 ...
 $ 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 ...
```

B. UNIVARIATE PLOT SECTIION

B0. Factorizing rating for the key variable we'd investigate throughout the dataset

Hide

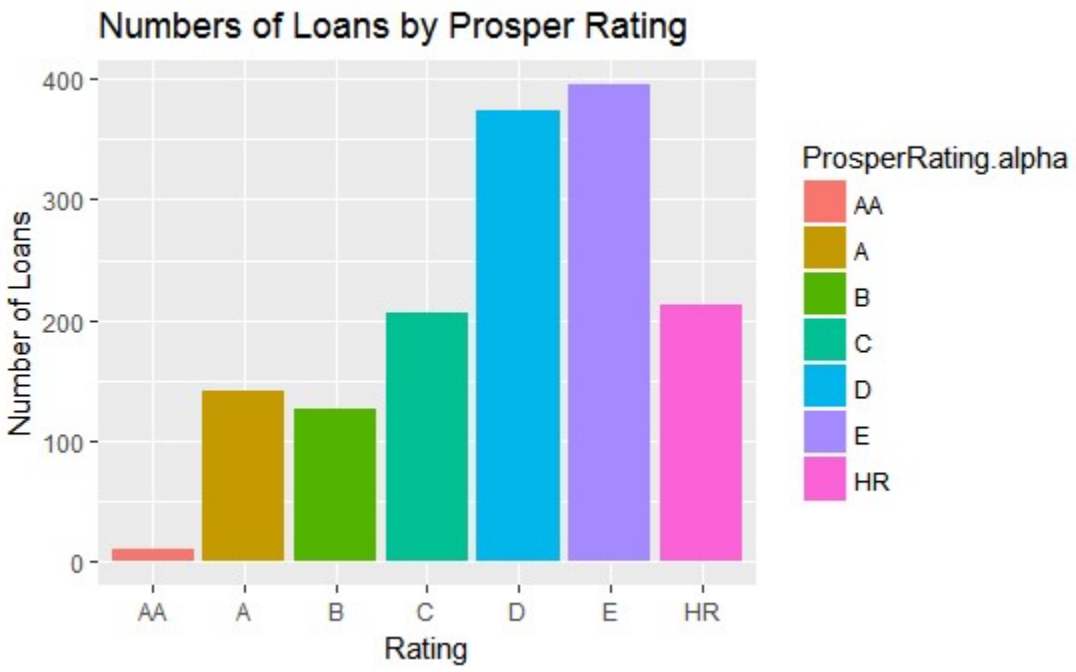
```
pf$ProsperRating.alpha = factor(pf$ProsperRating..Alpha.,
                                levels = c("AA","A","B","C","D","E","HR","N
A"))
pf$ProsperRating <-factor(pf$ProsperRating..Alpha,
                        levels = c('AA', 'A', 'B', 'C', 'D', 'E', 'HR', 'NA'))
pf$ProsperScore = factor(pf$ProsperScore)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

B1. HISTOGRAM OF PROSPER RATING BY NUMBERS OF LOANS

Hide

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

```
Ignoring unknown parameters: binwidth, bins, pad
```



Hide

```
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.

B1 - Part 2. PROSPER RATING DISTRIBUTION

Hide

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

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

Hide

```
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
```

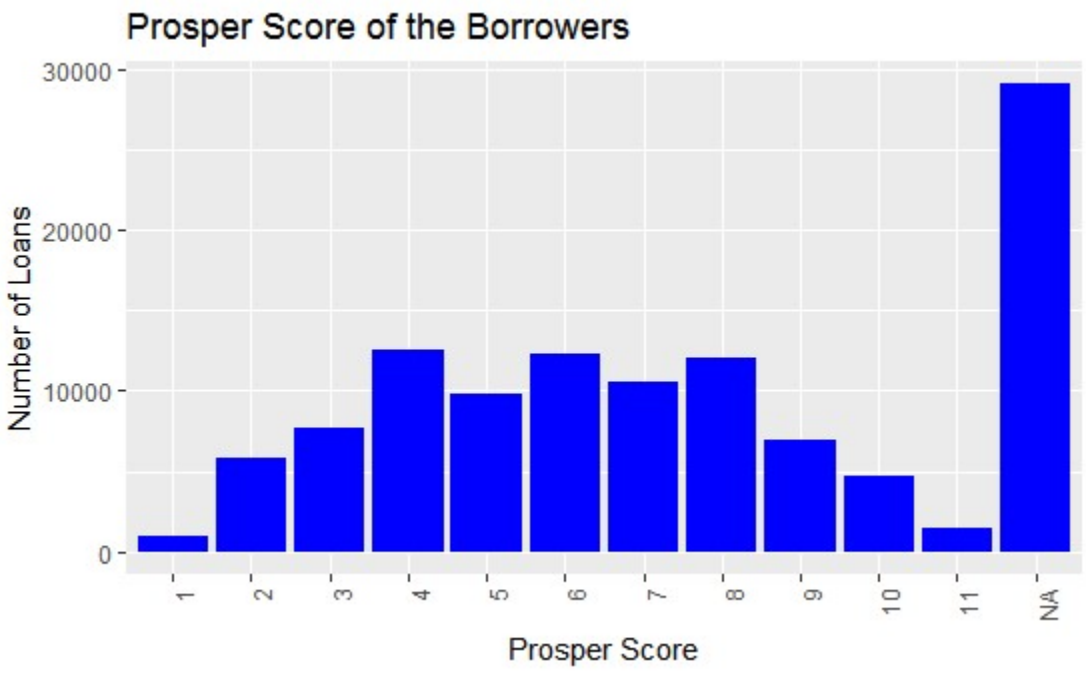
The NA count of Prosper Rating and Prosper Score is similar (29,084). I'm curious how the Prosper Rating and Prosper Score varies.

B2. PROSPER SCORE DISTRIBUTION

Hide

```
ggplot(pf, aes(x=ProsperScore)) +
                geom_histogram(aes(y=..count.., vjust=-0.9,   hjust=0.5), binwi
dth=500, size = 3, fill="blue", stat="count") +
  ggtitle('Prosper Score of the Borrowers') +
  xlab('Prosper Score') +
  ylab('Number of Loans') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

```
Ignoring unknown parameters: binwidth, bins, padIgnoring unknown aesthetics: vj
ust, hjust
```



Hide

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

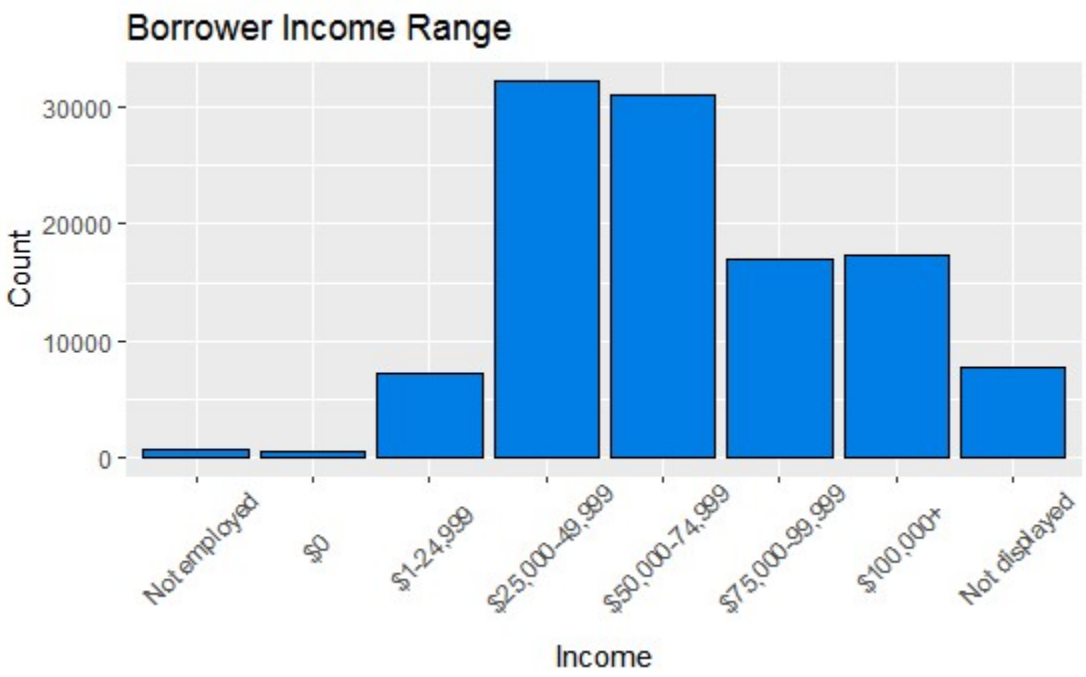
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   1.000   1.000   2.774   3.000  20.000
```

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

B3. BORROWER INCOME RANGE

Hide

```
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')
```

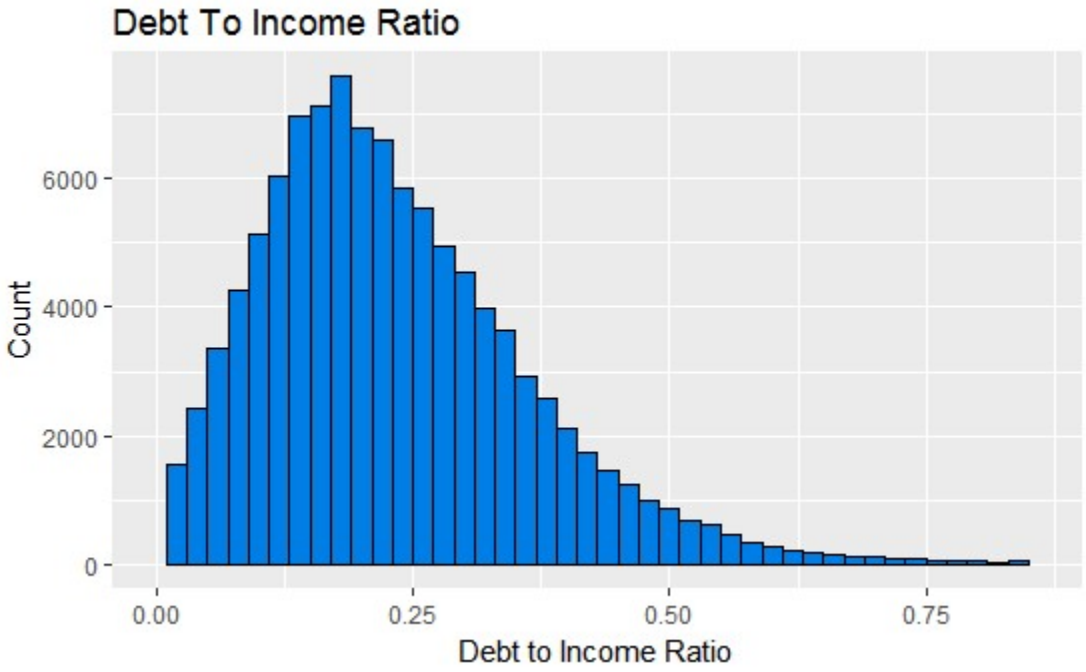


The majority of borrowers are in the $25,000 - $75,000 range. I suspect this lower-middle class range needs loans for debt consolidations.

B4. DEBT TO INCOME RATIO

Hide

```
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")
```



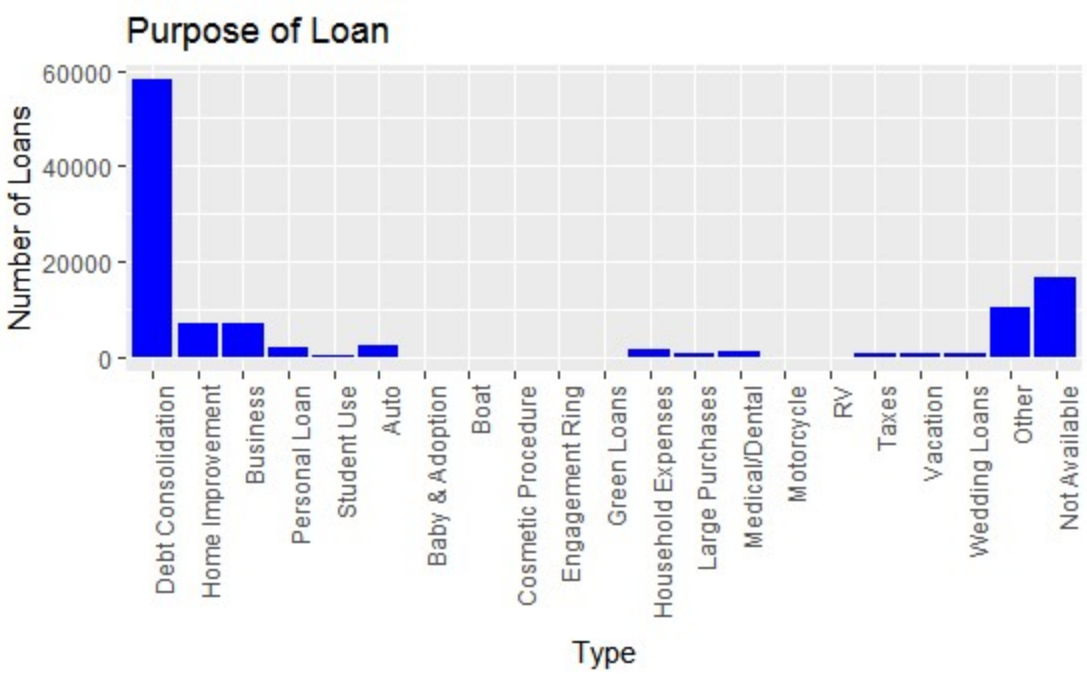
The data is long-tailed right-skewed. It's expected the majority of people in U.S have a credit history and the ratio should be low enough for a secured repayment.

B5. BORROWER'S PURPOSE OF LOAN

Hide

```
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(pf, aes(x=ListingCategory)) +
                geom_histogram(aes(y=..count.., vjust=-0.9,   hjust=0.5), binwi
dth=500, size = 3, fill="blue",stat="count") +
  ggtitle('Purpose of Loan') +
  xlab('Type') +
  ylab('Number of Loans') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

```
Ignoring unknown parameters: binwidth, bins, padIgnoring unknown aesthetics: vj
ust, hjust
```



Hide

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

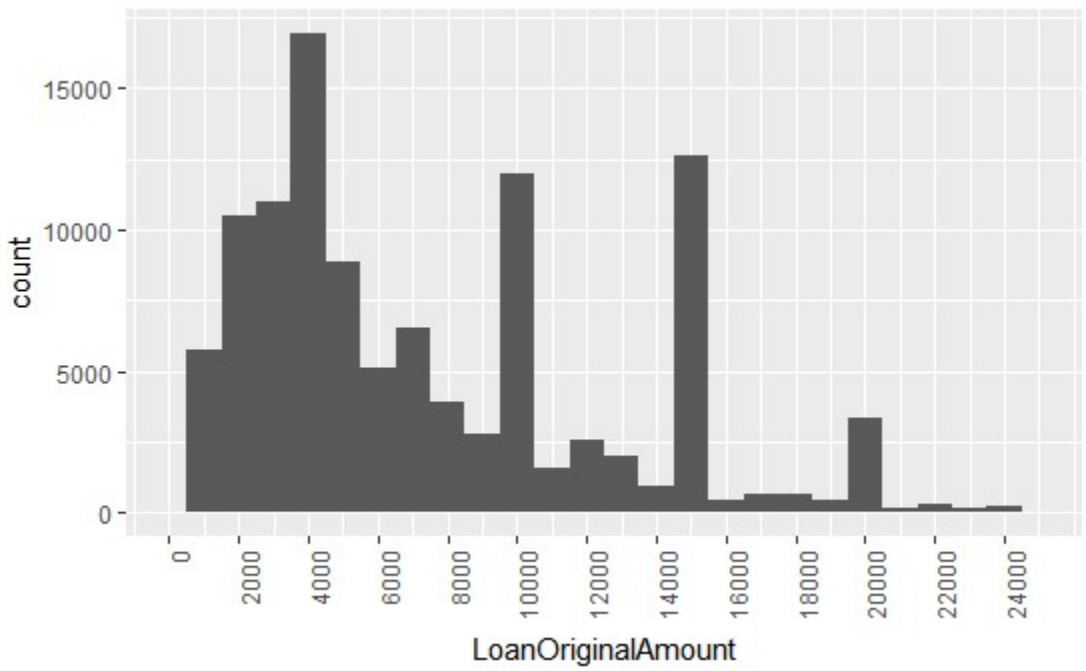
```
Debt Consolidation   Home Improvement           Business
             58308               7433               7189
     Personal Loan        Student Use               Auto
              2395                756               2572
   Baby & Adoption              Boat Cosmetic Procedure
               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.

B7. LOAN SPLIT BY AMOUNT

Hide

```
ggplot(pf, aes(LoanOriginalAmount)) +
                geom_histogram(binwidth = 1000) +
                scale_x_continuous(
                limits = c(0,quantile(pf$LoanOriginalAmount, 0.99,
                                                        na.rm = TRUE)),
                breaks = seq(0, quantile(pf$LoanOriginalAmount, 0.99,
                                                    na.rm = TRUE), 2000)) +
                theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Hide

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

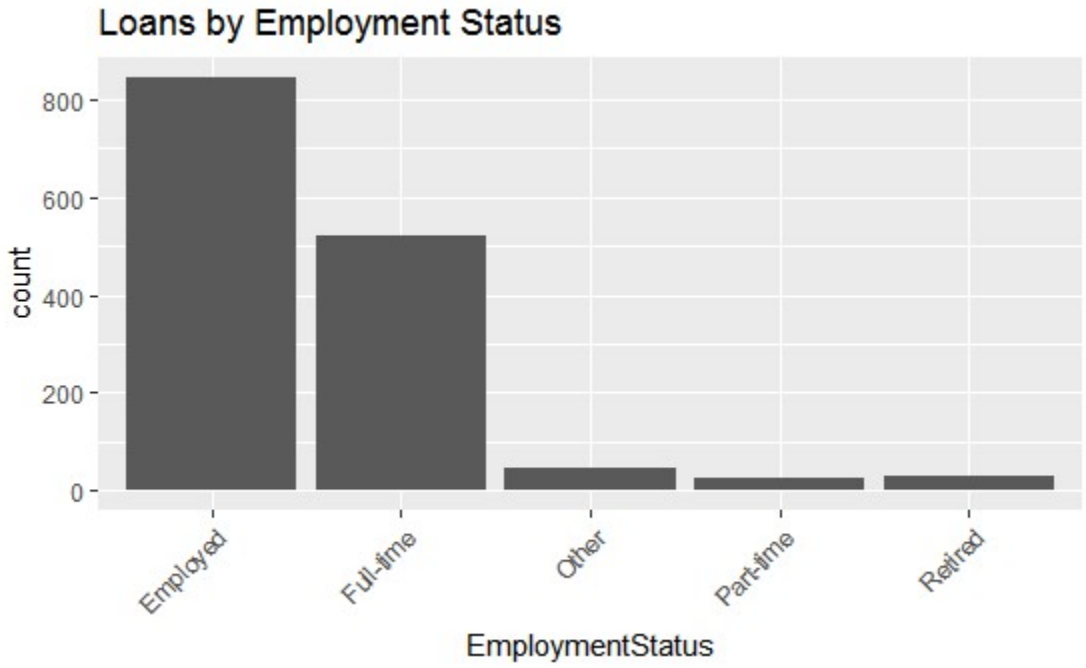
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1000    4000    6500    8337   12000   35000
```

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.

B8. EMPLOYMENT STATUS

Hide

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



Hide

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

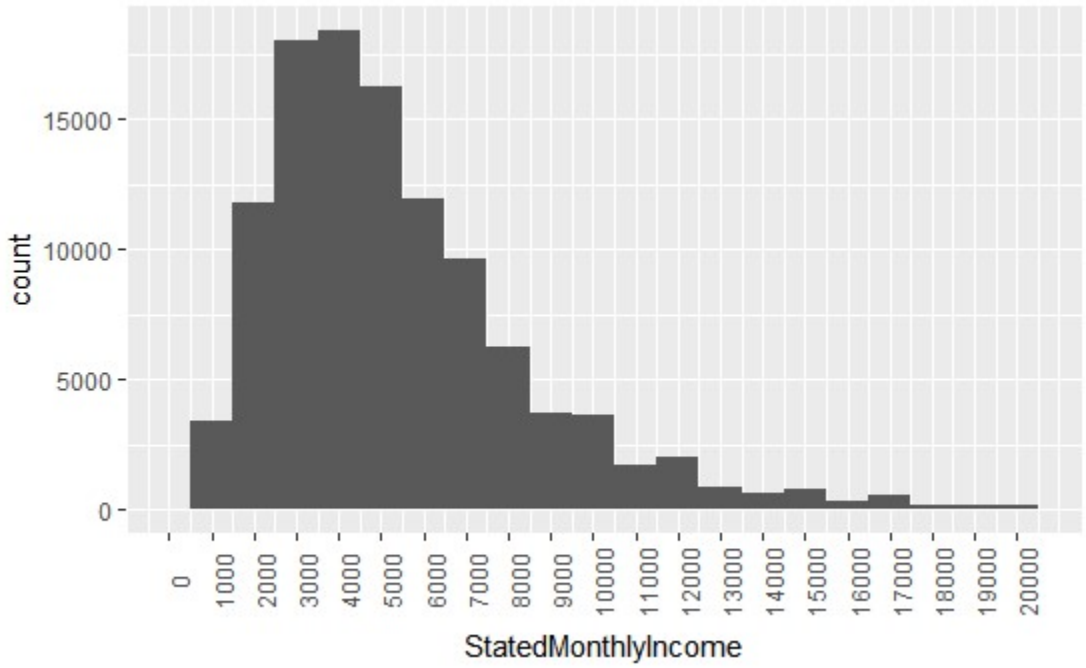
```
                  Employed     Full-time Not available
        2255         67322         26355          5347
 Not employed         Other     Part-time       Retired
         835          3806          1088           795
Self-employed
         6134
```

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

B9. STATED MONTHLY INCOME

Hide

```
ggplot(aes(x = StatedMonthlyIncome), data = pf) +
                  geom_histogram(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))
```



Hide

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

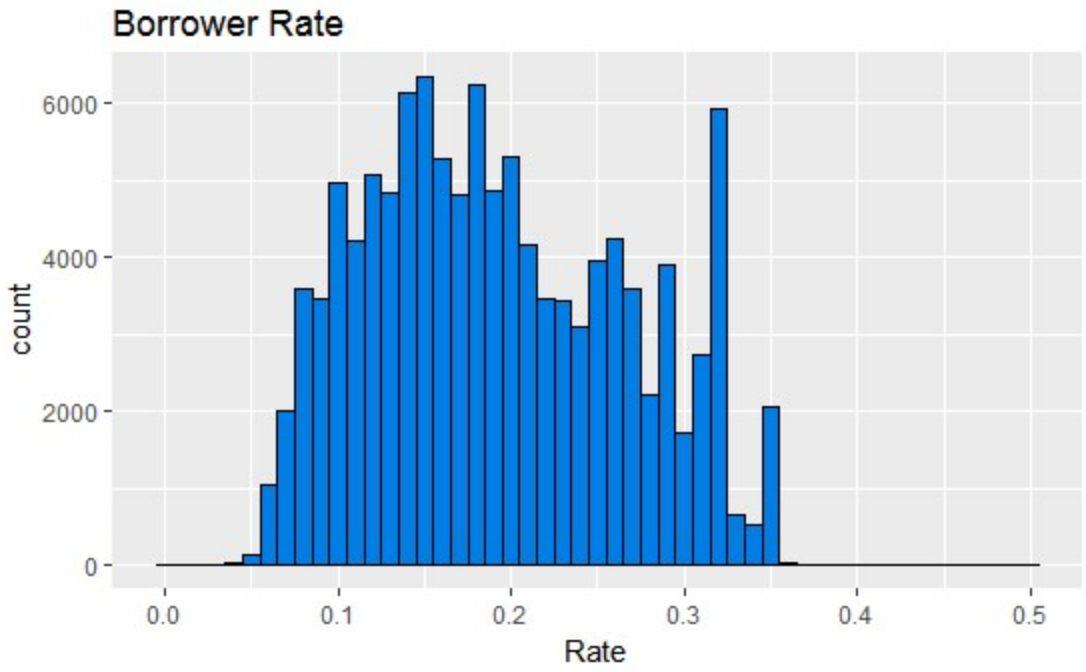
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0    3200    4667    5608    6825 1750000
```

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

B10. BORROWER'S RATE

Hide

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



Hide

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

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

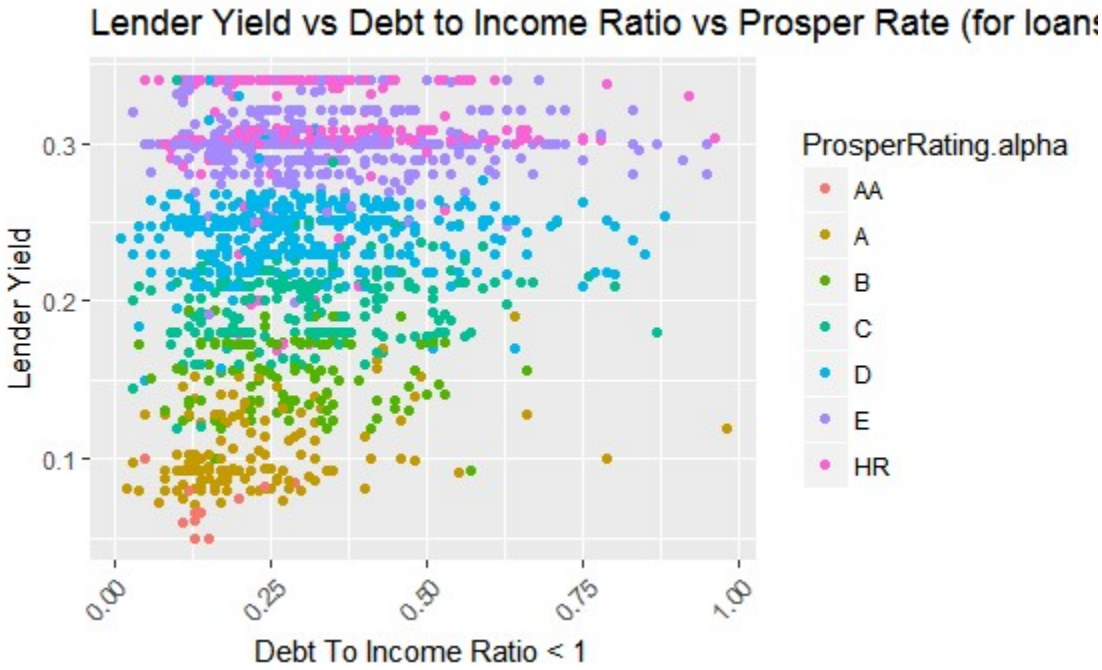
The most frequent rates are approximately 15%, 17% and 32%. This variation could be a factor of the amount or debt-to-income ratio.

C. MULTIVARIATE PLOT & ANAYLSIS SECTION

C1. DEBT TO INCOME RATIO - PROSPER RATING - LENDER YIELD

Hide

```
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(axis.text.x = element_text(angle = 45, hjust = 1))+
    ggtitle("Lender Yield vs Debt to Income Ratio vs Prosper Rate (for loans wi
th rating") +
    xlab ("Debt To Income Ratio < 1") +
    ylab ("Lender Yield") +
    scale_fill_discrete(name = "Prosper Rating")
```

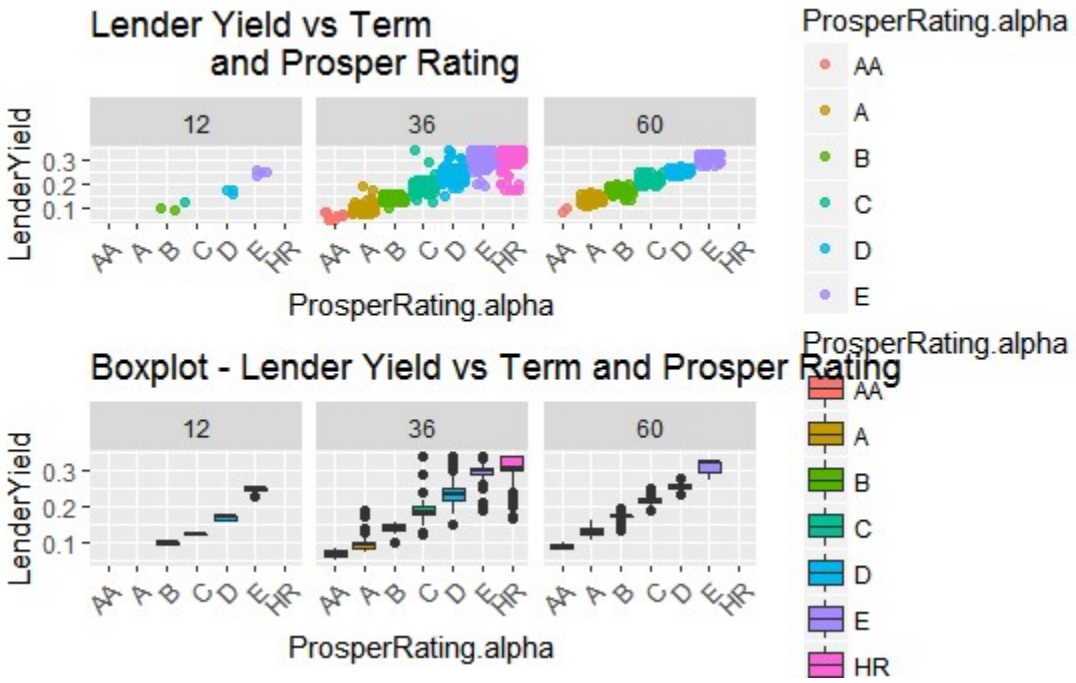


This chart shows the coorelation of the Lender Yield, the Prosper Rating and the Debt-To-Income Ratio.

C2. LENDER YIELD vs PROSPER RATE vs TERM

Hide

```
plot1 <- ggplot(aes(x= ProsperRating.alpha, y=LenderYield,
                                    color=ProsperRating.alpha),
       data=na.omit(filter(pf, DebtToIncomeRatio < 1))) +
    geom_point(alpha = 0.8, position = "jitter") +
    facet_grid( .~ Term  ) +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))+
    ggtitle("Lender Yield vs Term
            and Prosper Rating")
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(axis.text.x = element_text(angle = 45, hjust = 1))+
    ggtitle("Boxplot - Lender Yield vs Term and Prosper Rating")
grid.arrange(plot1, plot2, ncol=1, nrow =2)
```



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

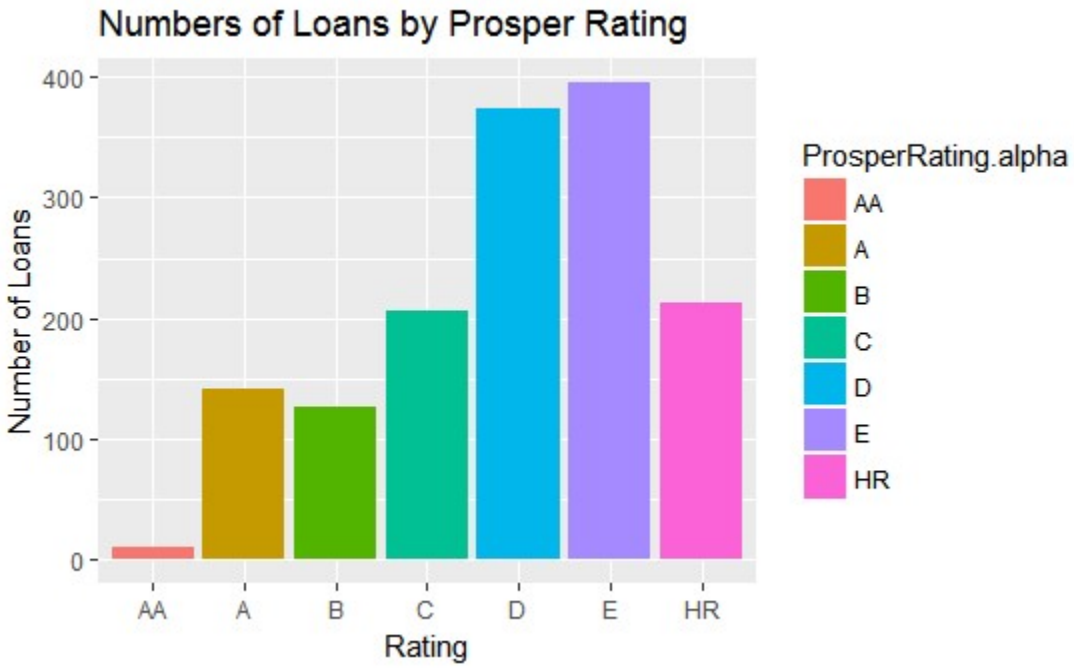
D. FINAL PLOTS & SUMMARY

PROSPER RATING

Hide

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

```
Ignoring unknown parameters: binwidth, bins, pad
```



Hide

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

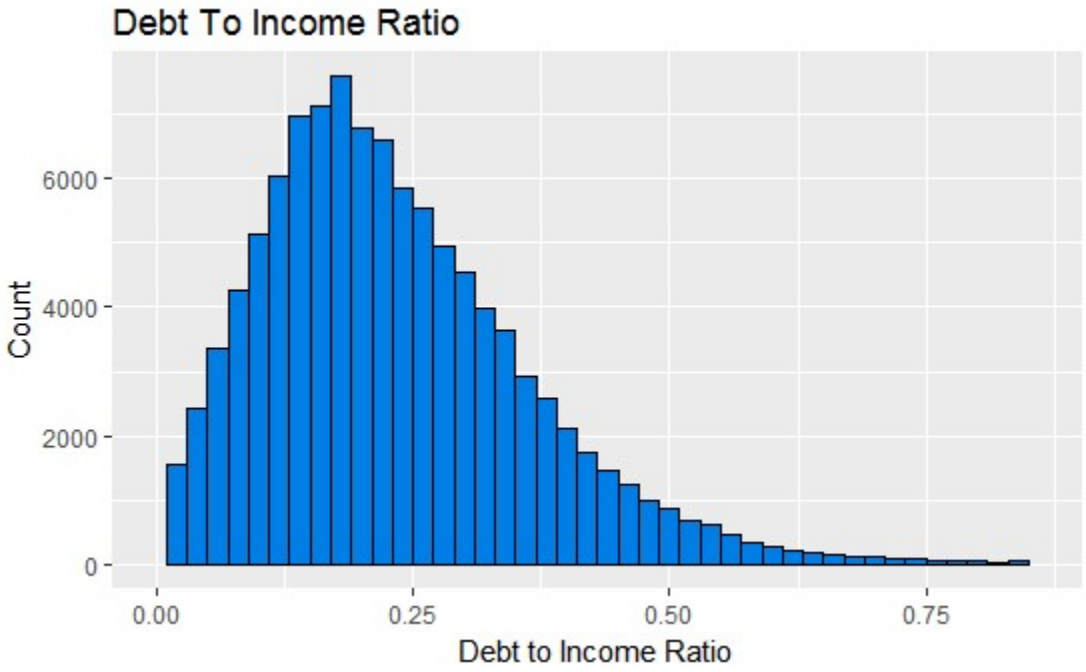
```
   AA     A     B     C     D     E    HR    NA  NA's
 5372 14551 15581 18345 14274  9795  6935     0 29084
```

I chose this graph as a final graph because it's important to see to the breakdown of Prosper Rating amount the loans. The most popular Prosper Ratings are D and E.

DEBT TO INCOME RATIO

Hide

```
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")
```

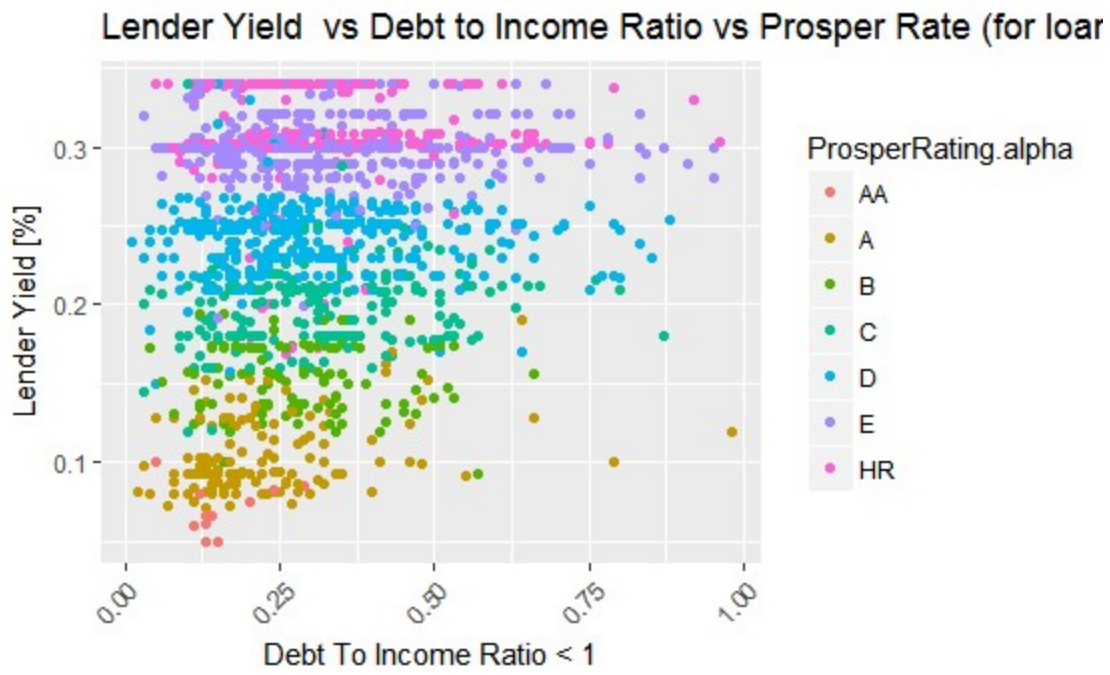


Similar to the graph above, I chose the Debt to Income Ratio graph as a final chart because it is important to see that the debt-to-income ratio for most borrowers is less than 0.25.

DEBT TO INCOME RATIO, LENDER YIELD AND PROSPER RATING

Hide

```
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(axis.text.x = element_text(angle = 45, hjust = 1))+
  ggtitle("Lender Yield  vs Debt to Income Ratio vs Prosper Rate (for loans wit
h rating)") +
  xlab ("Debt To Income Ratio < 1") +
  ylab ("Lender Yield [%]") +
  scale_fill_discrete(name = "Prosper Rating")
```



I chose to show this chart as one of the final three plots because I believe this shows the relationships between the Lender Yield and Prosper Rating. This shows that the higher the risk, the lower the rating and the better lender yield. A high Prosper Rating rating woul have a good debt-to-income ratio, which creates the upward triangle shape.

E. 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.