R Exploratory Data Analysis

Code

Mubina Arastu

29 October, 2021

1 Introduction

Prosper is a peer-to-peer lending platform that aims to connect people who need money with those who have money to invest (https://WWW.prosper.com/ (https://WWW.prosper.com/)). In this Exploratory Data Analysis, I explored Prosper dataset containing loan information for over a 100,000 people between the years 2006 and 2013.

I first explored the complete dataset, which has 81 features, and then decided to pick variables that were of my interest.

I wanted my R exploratory analysis to give meaningful insights to the reader. In peer-to-peer lending, there are three main stakeholders: borrowers, lenders, and the organization itself. So, I decided to focus on Prosper as it has higher stake over others and designed this project to mainly explore how Prosper rating impact Borrower's APR, Monthly Income and Loan amounts.

The exploration is divided analytically into 3 segments: Univariate Plots, Bivariate Plots, and Multivariate Plots.This also includes Final three plots as well as a Reflection section at the end that summarizes my experience and thoughts working throughout this project.

1.1 Dataset Structure

Let’s take a look at the sheer size of this dataset:

```
## 'data.frame':    113937 obs. of  81 variables:
##  $ ListingKey                       : chr  "1021339766868145413AB3B" "102736024
99503308B223C1" "0EE9337825851032864889A" "0EF5356002482715299901A" ...
##  $ ListingNumber                    : int  193129 1209647 81716 658116 909464 1
074836 750899 768193 1023355 1023355 ...
##  $ ListingCreationDate              : chr  "09:29.3" "28:07.9" "00:47.1" "02:3
5.0" ...
##  $ CreditGrade                      : chr  "C" "" "HR" "" ...
##  $ Term                             : int  36 36 36 36 36 60 36 36 36 36 ...
##  $ LoanStatus                       : chr  "Completed" "Current" "Completed" "C
urrent" ...
##  $ ClosedDate                       : chr  "8/14/2009 0:00" "" "12/17/2009 0:0
0" "" ...
##  $ BorrowerAPR                      : num  0.165 0.12 0.283 0.125 0.246 ...
##  $ BorrowerRate                     : num  0.158 0.092 0.275 0.0974 0.2085 ...
##  $ 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.            : chr  "" "A" "" "A" ...
##  $ ProsperScore                     : int  NA 7 NA 9 4 10 2 4 9 11 ...
##  $ ListingCategory..numeric.        : int  0 2 0 16 2 1 1 2 7 7 ...
##  $ BorrowerState                    : chr  "CO" "CO" "GA" "GA" ...
##  $ Occupation                       : chr  "Other" "Professional" "Other" "Skil
led Labor" ...
##  $ EmploymentStatus                 : chr  "Self-employed" "Employed" "Not avai
lable" "Employed" ...
##  $ EmploymentStatusDuration         : int  2 44 NA 113 44 82 172 103 269 269
...
##  $ IsBorrowerHomeowner              : logi  TRUE FALSE FALSE TRUE TRUE TRUE ...
##  $ CurrentlyInGroup                 : logi  TRUE FALSE TRUE FALSE FALSE FALSE
...
##  $ GroupKey                         : chr  "" "" "783C3371218786870A73D20" ""
...
##  $ DateCreditPulled                 : chr  "41:46.8" "2/27/2014 8:28" "09:10.1"
"10/22/2012 11:02" ...
##  $ CreditScoreRangeLower            : int  640 680 480 800 680 740 680 700 820
820 ...
##  $ CreditScoreRangeUpper            : int  659 699 499 819 699 759 699 719 839
839 ...
##  $ FirstRecordedCreditLine          : chr  "10/11/2001 0:00" "3/18/1996 0:00"
"7/27/2002 0:00" "2/28/1983 0:00" ...
##  $ 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      : int  24 389 0 115 220 1410 214 101 219 21
```

```
9 ...
##  $ InquiriesLast6Months               : int  3 3 0 0 1 0 0 3 1 1 ...
##  $ TotalInquiries                     : int  3 5 1 1 9 2 0 16 6 6 ...
##  $ CurrentDelinquencies               : int  2 0 1 4 0 0 0 0 0 0 ...
##  $ AmountDelinquent                   : int  472 0 NA 10056 0 0 0 0 0 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             : int  0 3989 NA 1444 6193 62999 5812 1260
9906 9906 ...
##  $ BankcardUtilization                : num  0 0.21 NA 0.04 0.81 0.39 0.72 0.13
0.11 0.11 ...
##  $ AvailableBankcardCredit            : int  1500 10266 NA 30754 695 86509 1929 2
181 77696 77696 ...
##  $ TotalTrades                        : int  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            : int  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                        : chr  "$25,000-49,999" "$50,000-74,999" "N
ot displayed" "$25,000-49,999" ...
##  $ IncomeVerifiable                   : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
##  $ StatedMonthlyIncome                : num  3083 6125 2083 2875 9583 ...
##  $ LoanKey                            : chr  "E33A3400205839220442E84" "9E3B37071
505919926B1D82" "6954337960046817851BCB2" "A0393664465886295619C51" ...
##  $ 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 NA ...
##  $ 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 12325
7 88353 90051 121268 121268 ...
##  $ LoanOriginalAmount                 : int  9425 10000 3001 10000 15000 15000 30
00 10000 10000 10000 ...
##  $ LoanOriginationDate                : chr  "9/12/2007 0:00" "3/3/2014 0:00" "1/
17/2007 0:00" "11/1/2012 0:00" ...
##  $ LoanOriginationQuarter             : chr  "Q3 2007" "Q1 2014" "Q1 2007" "Q4 20
12" ...
##  $ MemberKey                          : chr  "1F3E3376408759268057EDA" "1D1337054
6739025387B2F4" "5F7033715035555618FA612" "9ADE356069835475068C6D2" ...
##  $ 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 ...
##  $ 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 ...
```

1.2 Dataset Summary

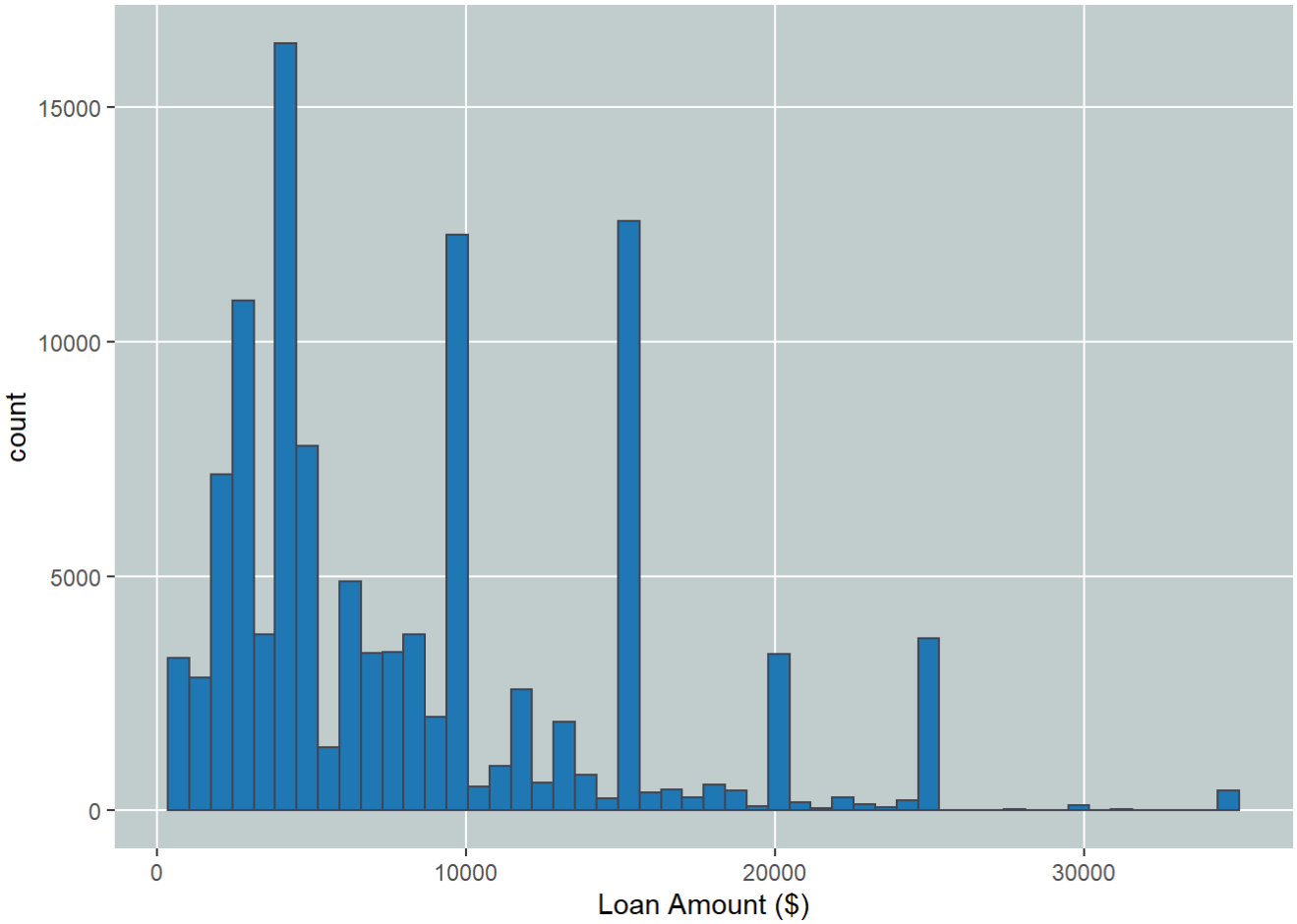
Code

```
##  BankCreditAvailable BorrowerState       BorrowerAPR      CreditGrade
##  Min.   :     0      Length:113937      Min.   : 0.653   C      : 5649
##  1st Qu.:   880      Class :character   1st Qu.:15.629   D      : 5153
##  Median :  4100      Mode  :character   Median :20.976   B      : 4389
##  Mean   : 11210                         Mean   :21.883   AA     : 3509
##  3rd Qu.: 13180                         3rd Qu.:28.381   HR     : 3508
##  Max.   :646285                         Max.   :51.229   (Other): 6604
##  NA's   :7544                           NA's   :25       NA's   :85125
##  CurrentCreditLines CurrentDelinquencies AvgCreditScore  EmploymentStatus
##  Min.   : 0.00      Min.   : 0.0000      Min.   :  9.5   Length:113937
##  1st Qu.: 7.00      1st Qu.: 0.0000      1st Qu.:669.5   Class :character
##  Median :10.00      Median : 0.0000      Median :689.5   Mode  :character
##  Mean   :10.32      Mean   : 0.5921      Mean   :695.1
##  3rd Qu.:13.00      3rd Qu.: 0.0000      3rd Qu.:729.5
##  Max.   :59.00      Max.   :83.0000      Max.   :889.5
##  NA's   :7604       NA's   :697          NA's   :591
##  IsBorrowerHomeowner InterestandFees     ServiceFees
##  Mode :logical       Min.   :   -2.35   Min.   :-664.87
##  FALSE:56459         1st Qu.:  274.87   1st Qu.: -73.18
##  TRUE :57478         Median :  700.84   Median : -34.44
##                      Mean   : 1077.54   Mean   : -54.73
##                      3rd Qu.: 1458.54   3rd Qu.: -13.92
##                      Max.   :15617.03   Max.   :  32.06
##
##            ListingCategory    LoanAmount     LoanStatus
##  Debt Consolidation:58308   Min.   : 1000   Length:113937
##  Not Available     :16965   1st Qu.: 4000   Class :character
##  Other             :10494   Median : 6500   Mode  :character
##  Home Improvement  : 7433   Mean   : 8337
##  Business          : 7189   3rd Qu.:12000
##  Auto              : 2572   Max.   :35000
##  (Other)           :10976
##  MonthlyLoanPayment  Occupation        ProsperRating   MonthlyIncome
##  Min.   :   0.0     Length:113937      C      :18345   Min.   :      0
##  1st Qu.: 131.6     Class :character   B      :15581   1st Qu.:   3200
##  Median : 217.7     Mode  :character   A      :14551   Median :   4667
##  Mean   : 272.5                        D      :14274   Mean   :   5608
##  3rd Qu.: 371.6                        E      : 9795   3rd Qu.:   6825
##  Max.   :2251.5                        (Other):12307   Max.   :1750003
##                                        NA's   :29084
##  Term
##  12: 1614
##  36:87778
##  60:24545
##
##
##
##
```

2 Univariate Plots Section

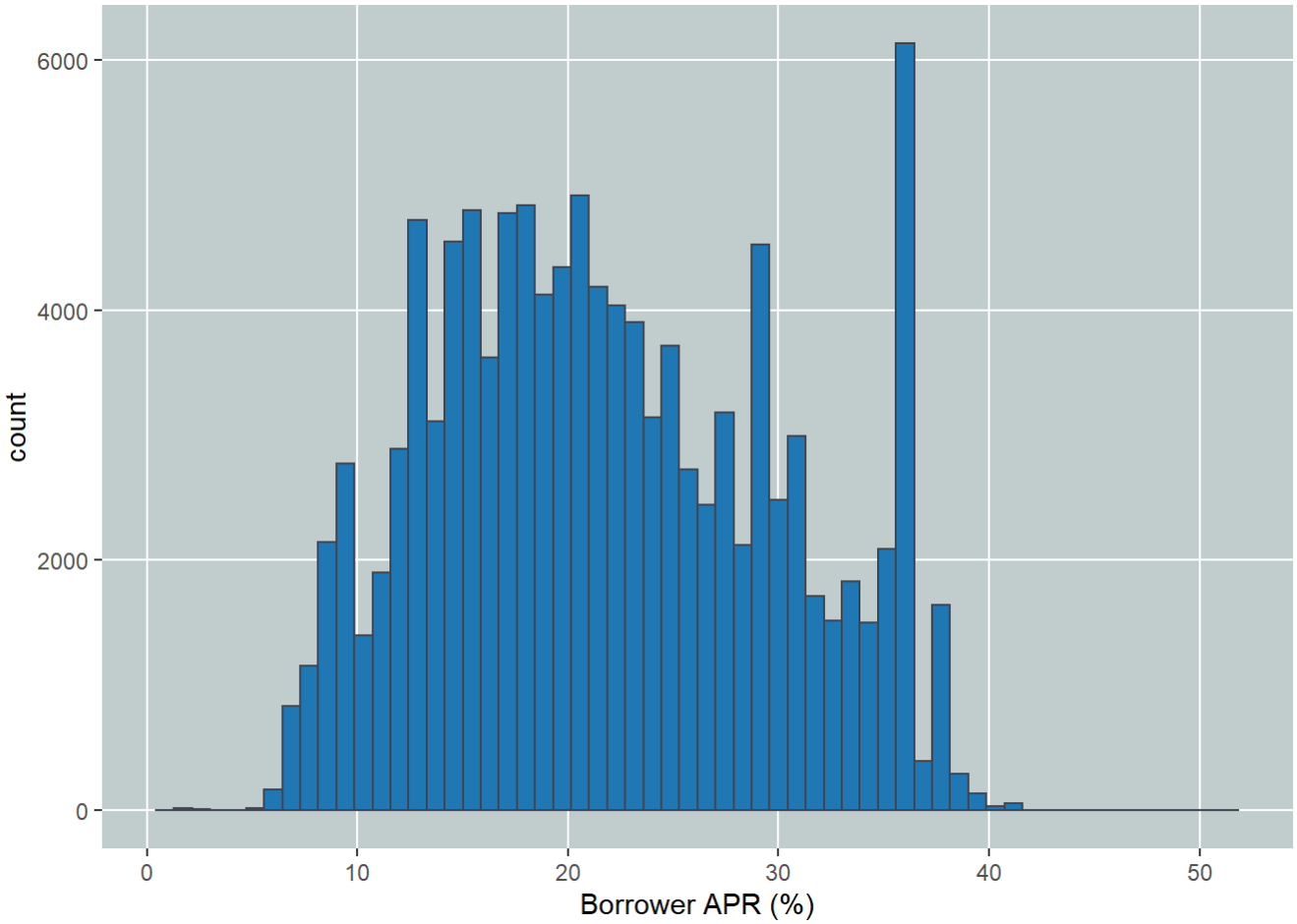
Let's draw Histogram & Bar graph to see Loan amounts, Borrower APR % and Terms. I chose Histogram to show accurate distribution of variables (loan amounts and Borrower APR) that have continuous measurements and also wanted to look for outliers.I selected Bar graph to show the comparison of discrete variable i.e Terms(In Months).

2.1 Loan Amounts - (Histogram)



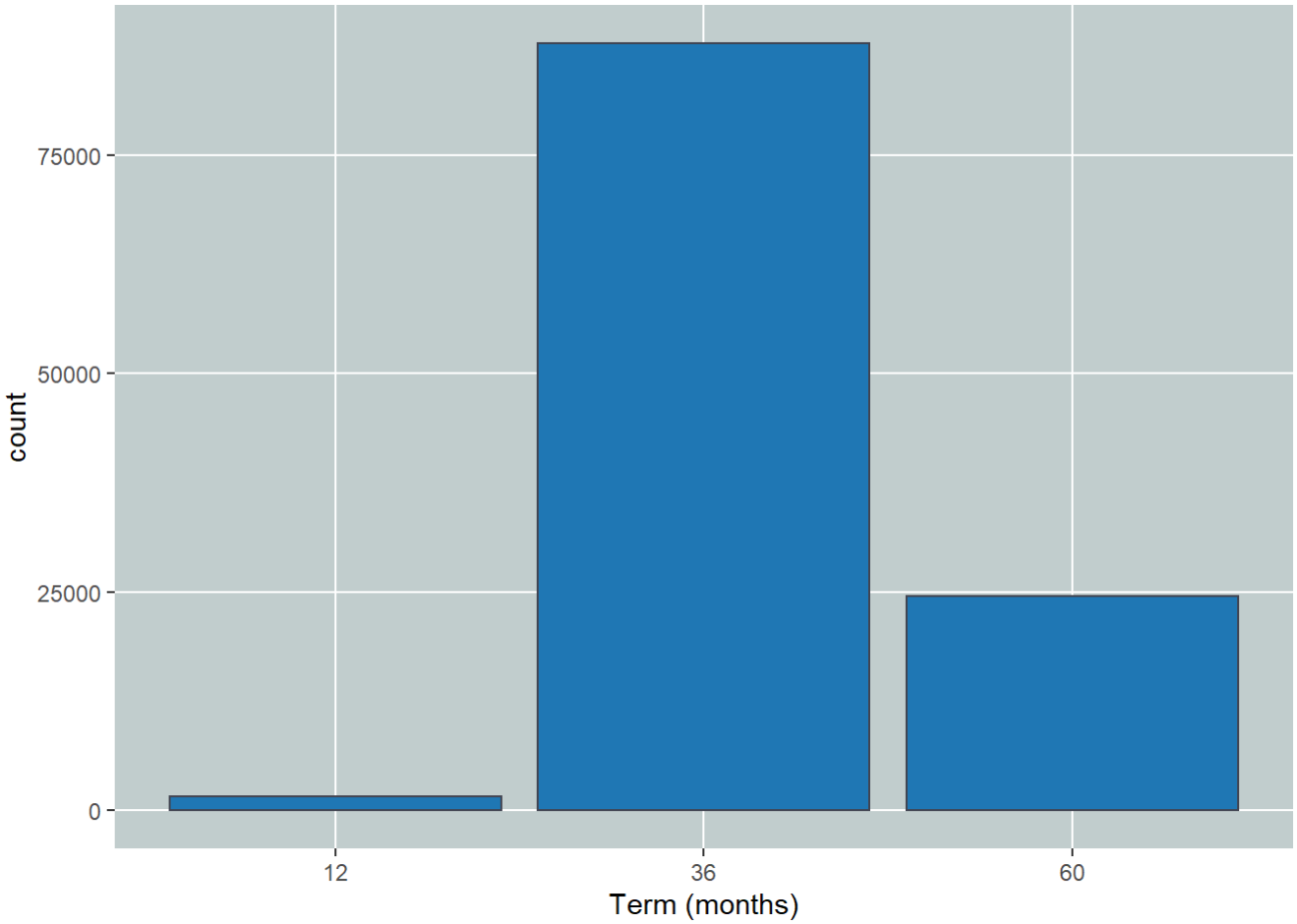
The loan plot shows that borrowers generally borrowed loans in round numbers (e.g. $7000,$10000,$15000,$20000).

2.2 Borrowers APR(%) - (Histogram)



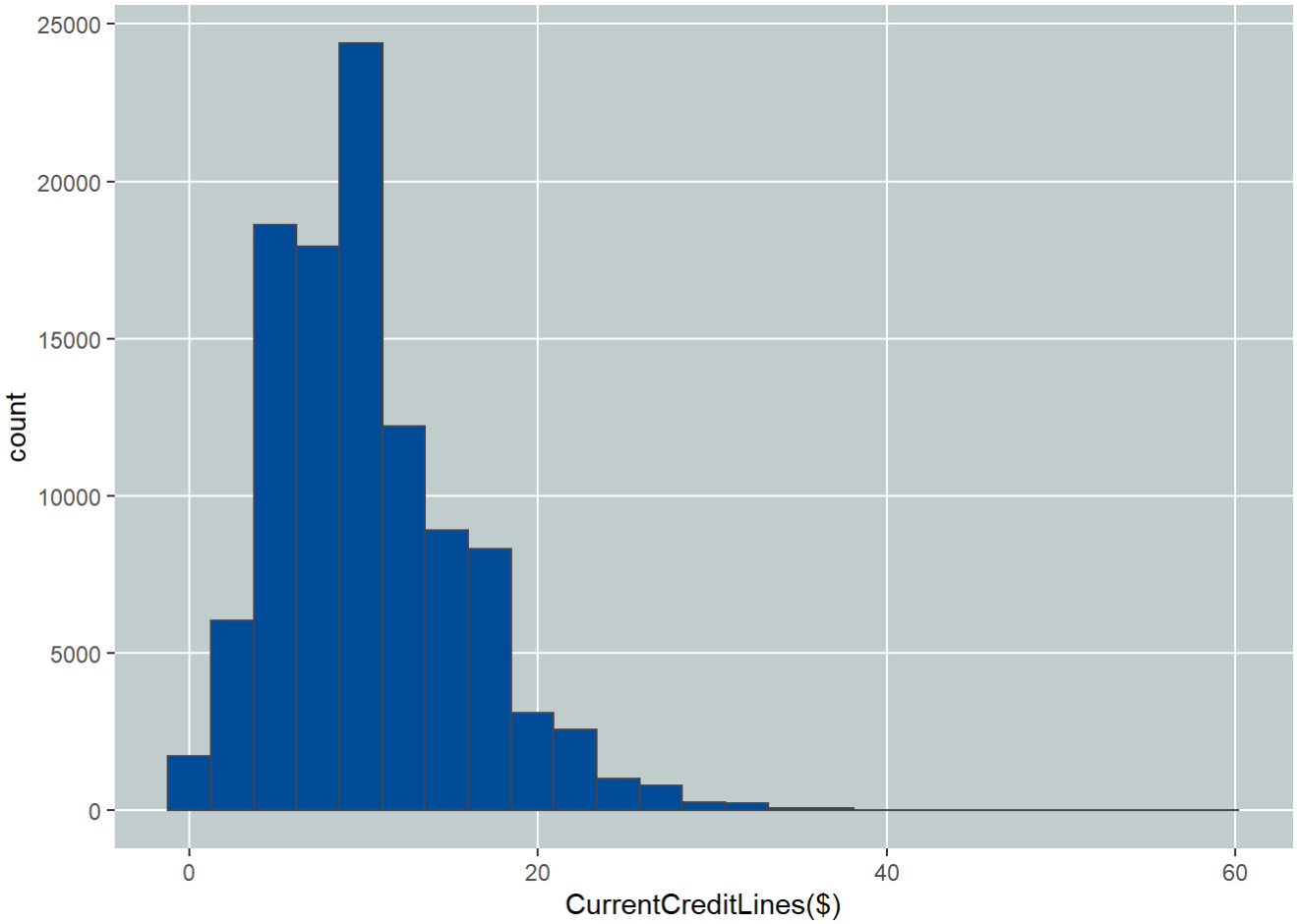
The Borrower APR shows a normal distribution centered around 21% with an anomalous mode around 36%. This shows that most likely the loan company sets specific APR tiers for borrowers in specific credit grade categories.

2.3 Term Trend - (Bar graph)



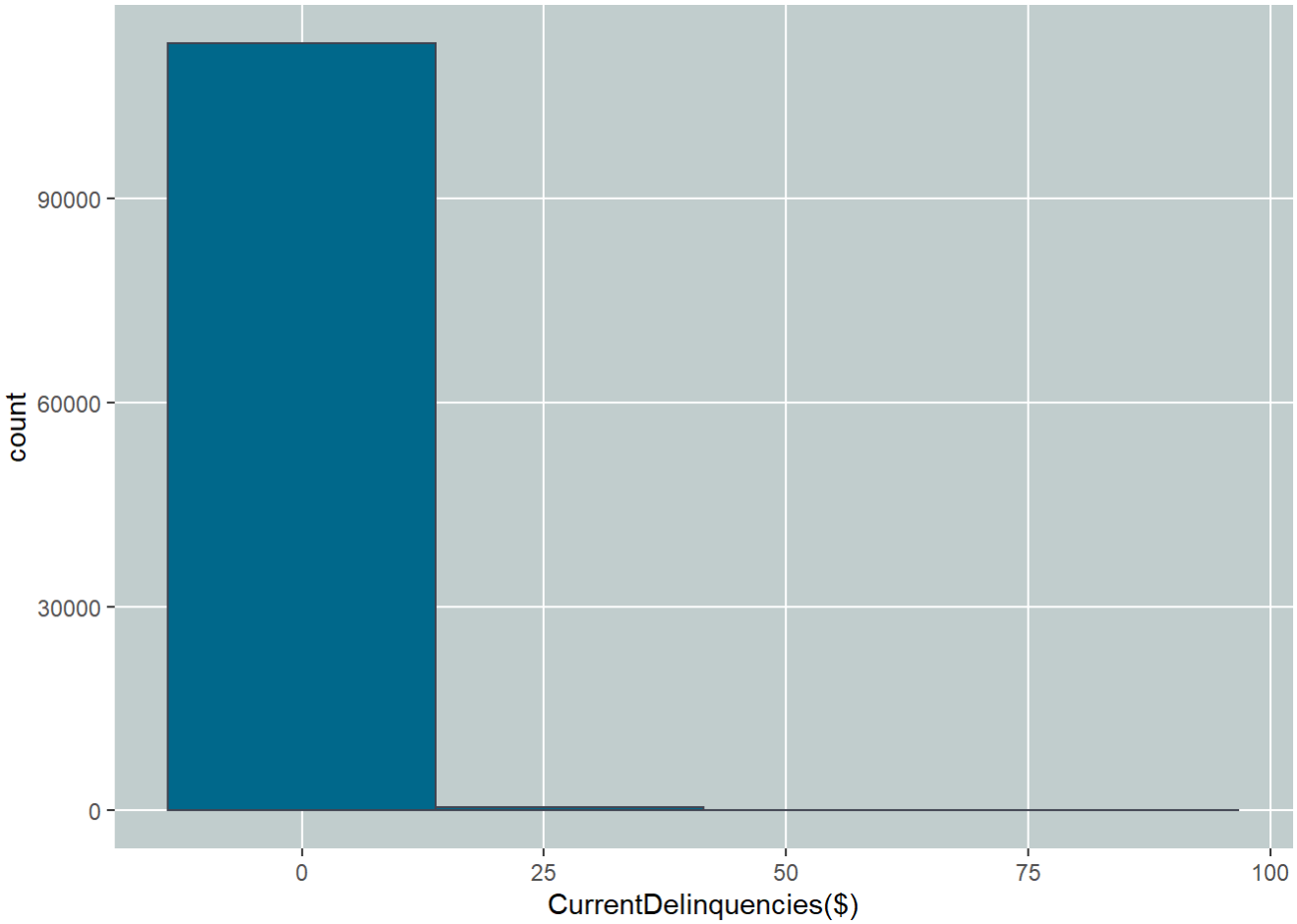
The above Term trend shows that loan amounts were borrowed mostly for 36 months term followed by 60 months term.

2.4 Current Credit Lines - (Histogram)



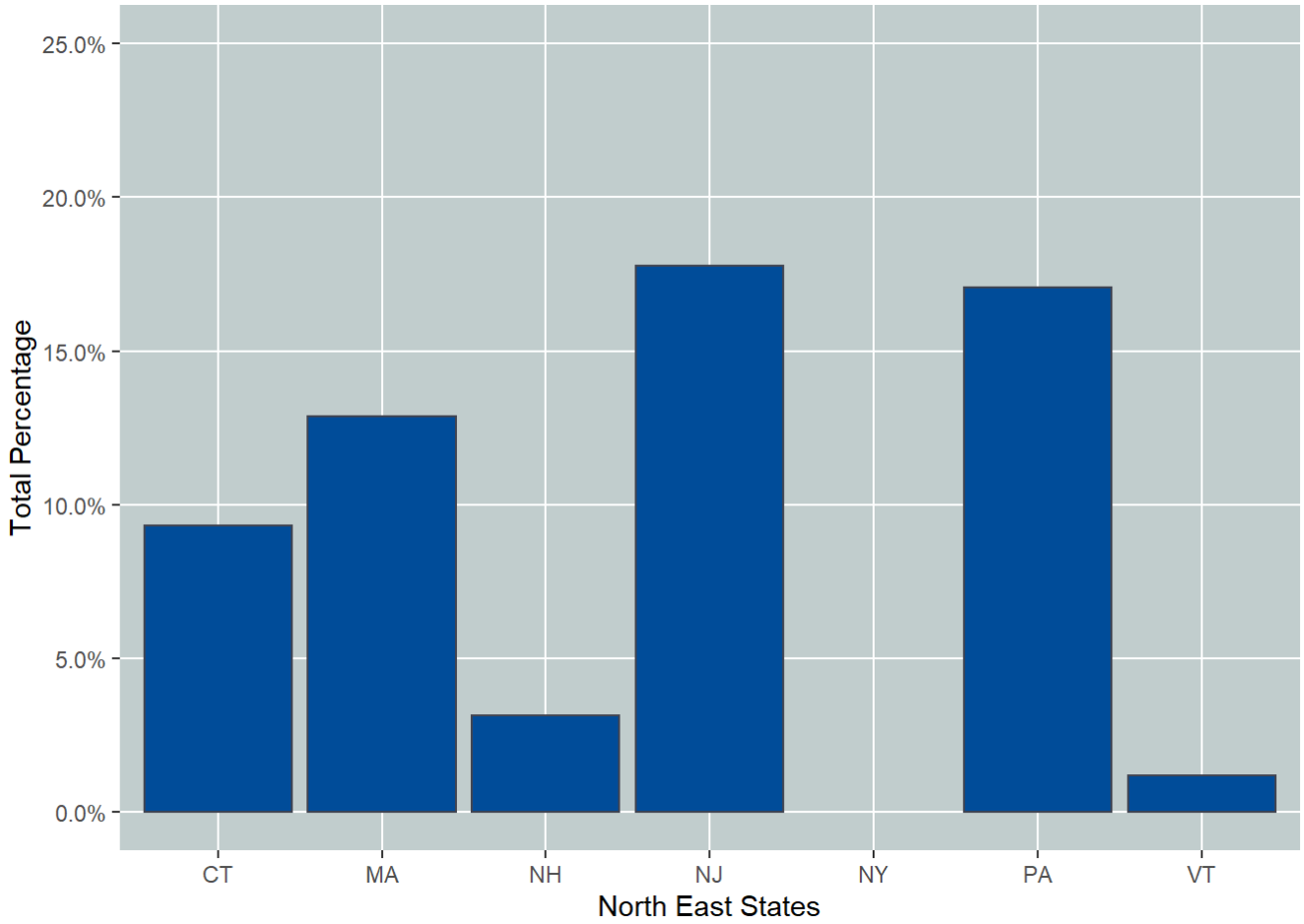
Initialy above histogram was designed to have '50' bin size which resulted in an outlier for (5-10) Credit lines exceeding 1500 counts. To show more accurate results, I adjusted the bin size to 25, now it clearly shows that current credit lines in the range (5-10) has seen a spike at around 25000 counts.

2.5 Current Delinquencies - (Histogram)



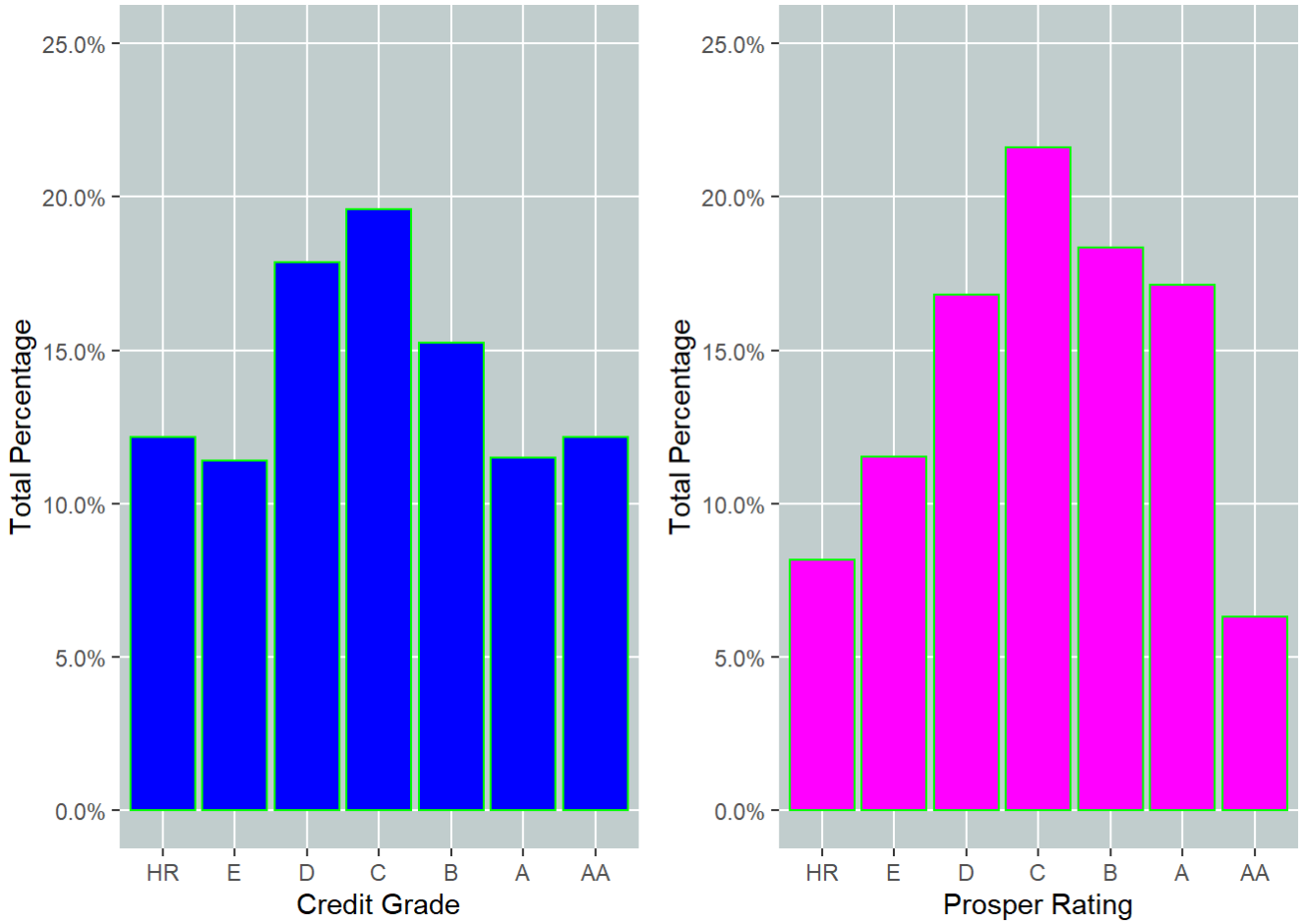
I have adjusted above histogram to show the results in 4 bins to depict the distribution of current delinquencies in three range group. It is clearly evident that borrowers mostly had current delinquencies in the range from (0-30).

2.6 North Eastern Borrower States - (Bar graph)



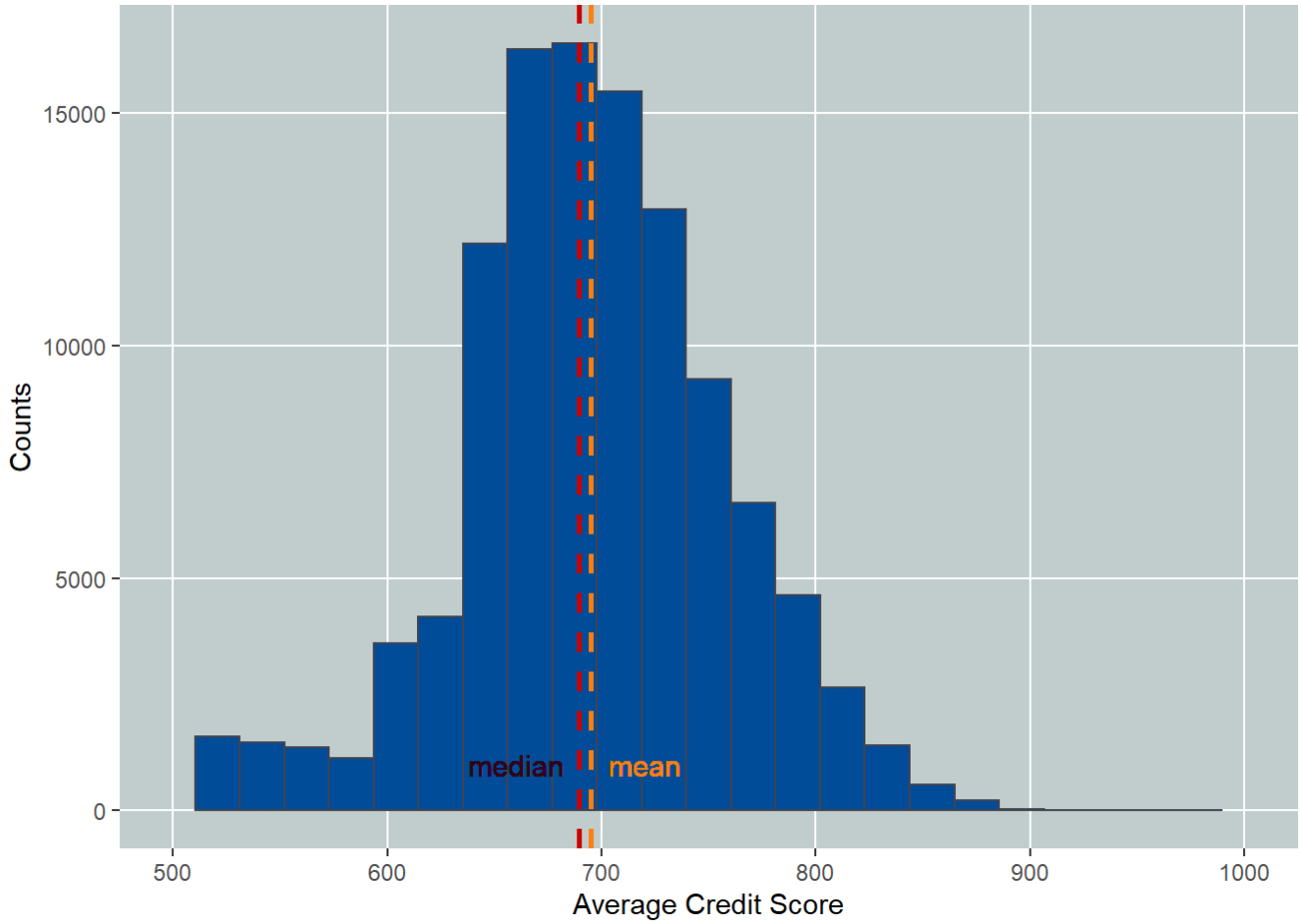
I have selected BAR graph here to draw accurate representation of categorical variable “states”. I have used filtered subset of Borrowers state to only focus on North eastern region. The graph shows that mostly borrowers are from New Jersey and Pennsylvania states.

2.7 Credit Grade & Prosper Rating - (Bar graph)



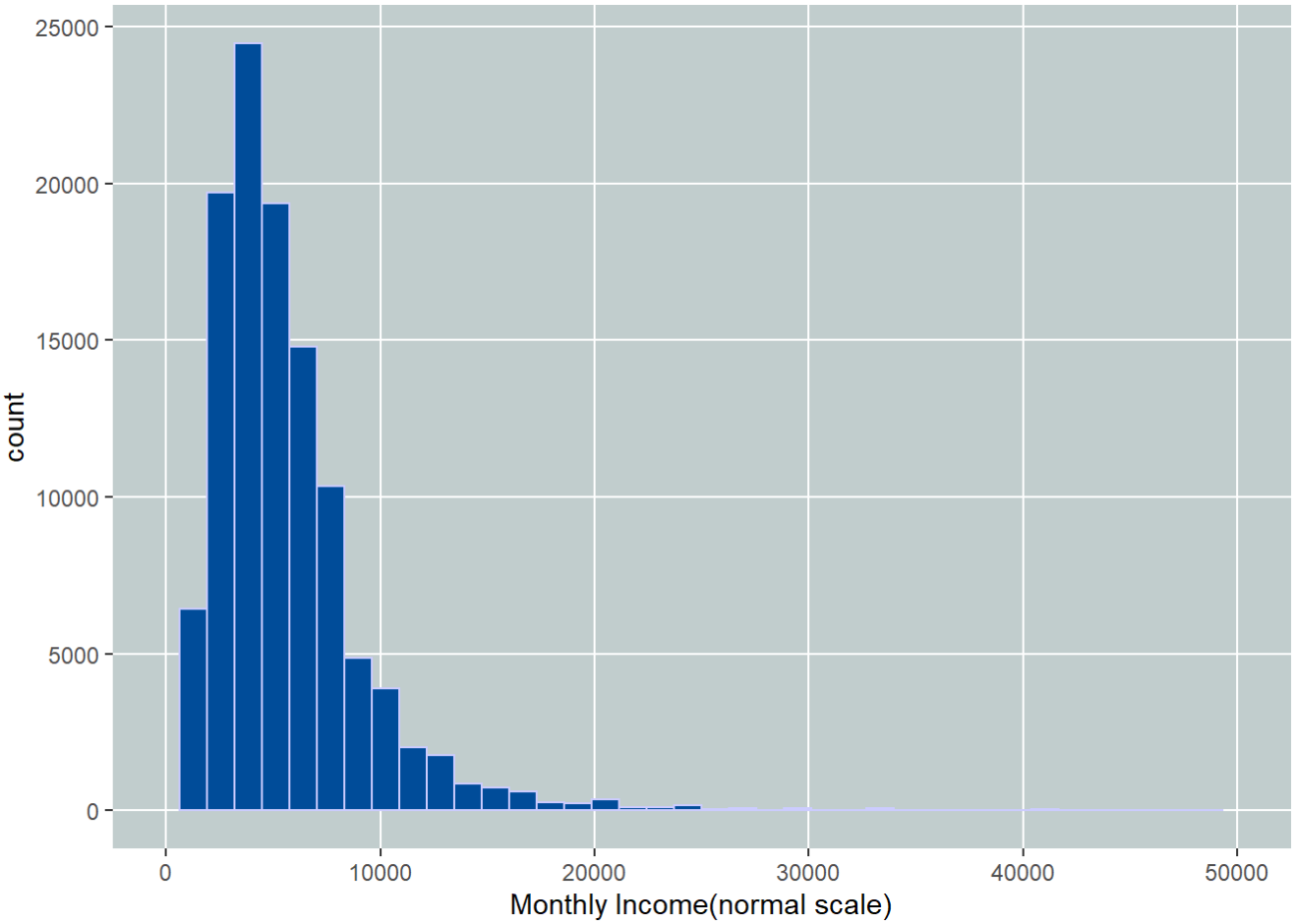
I chose bar graph to compare Credit grade with Prosper Rating. Above two plots reveals good comparison between credit grade and Prosper rating. It clearly signifies that Prosper gives positive 'AA' rating to very few borrowers. However, they also gave more 'E' than 'HR' ratings.

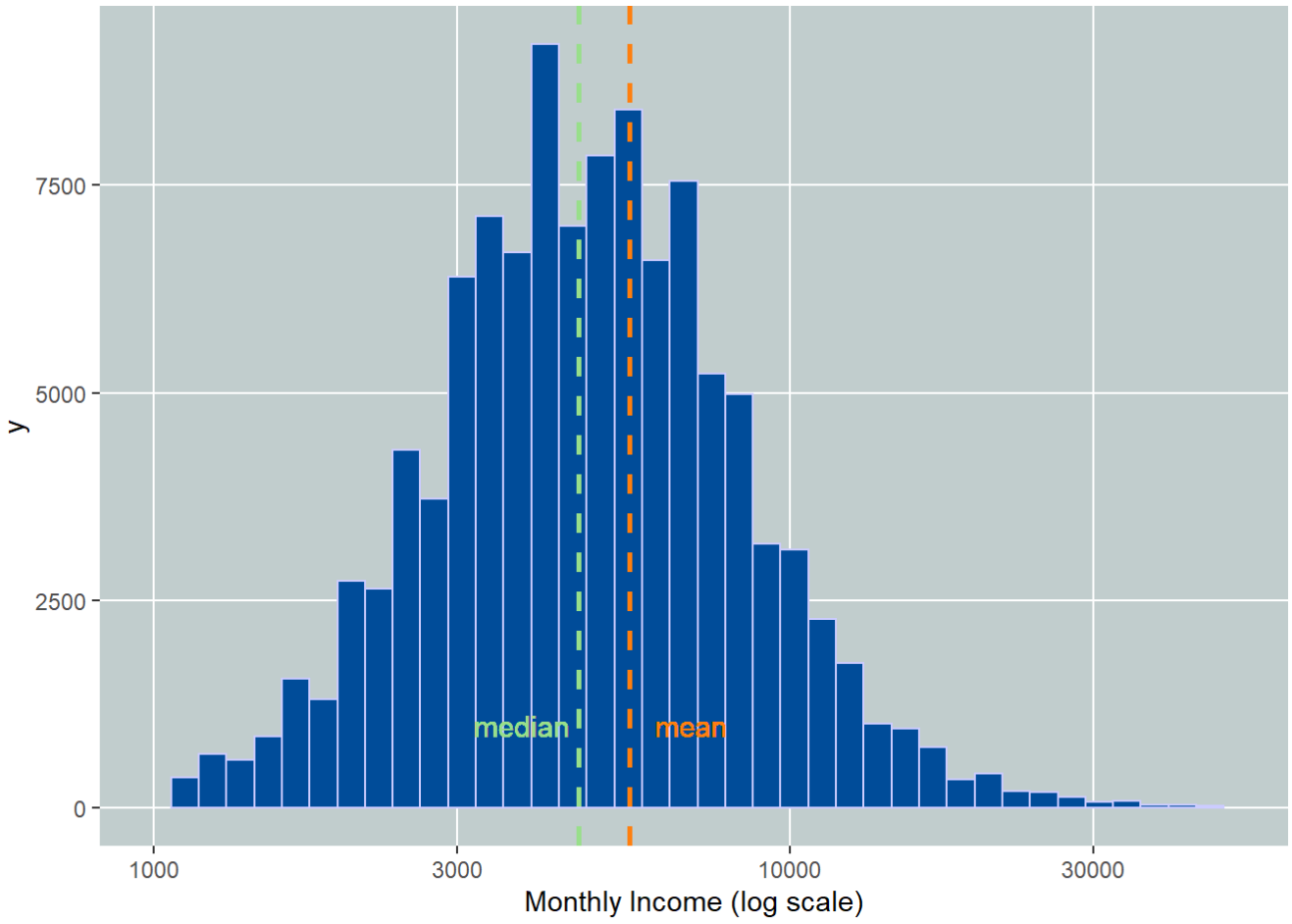
2.8 AvgCreditScore- (Histogram Normal scales distribution)



I chose normal scales histogram with 'geom lines' to show Mean and Median values along with the total Credit scores counts.The credit score is normally distributed with a median near 700.

2.9 Monthly Income - (Histogram Normal and Log scales distribution)

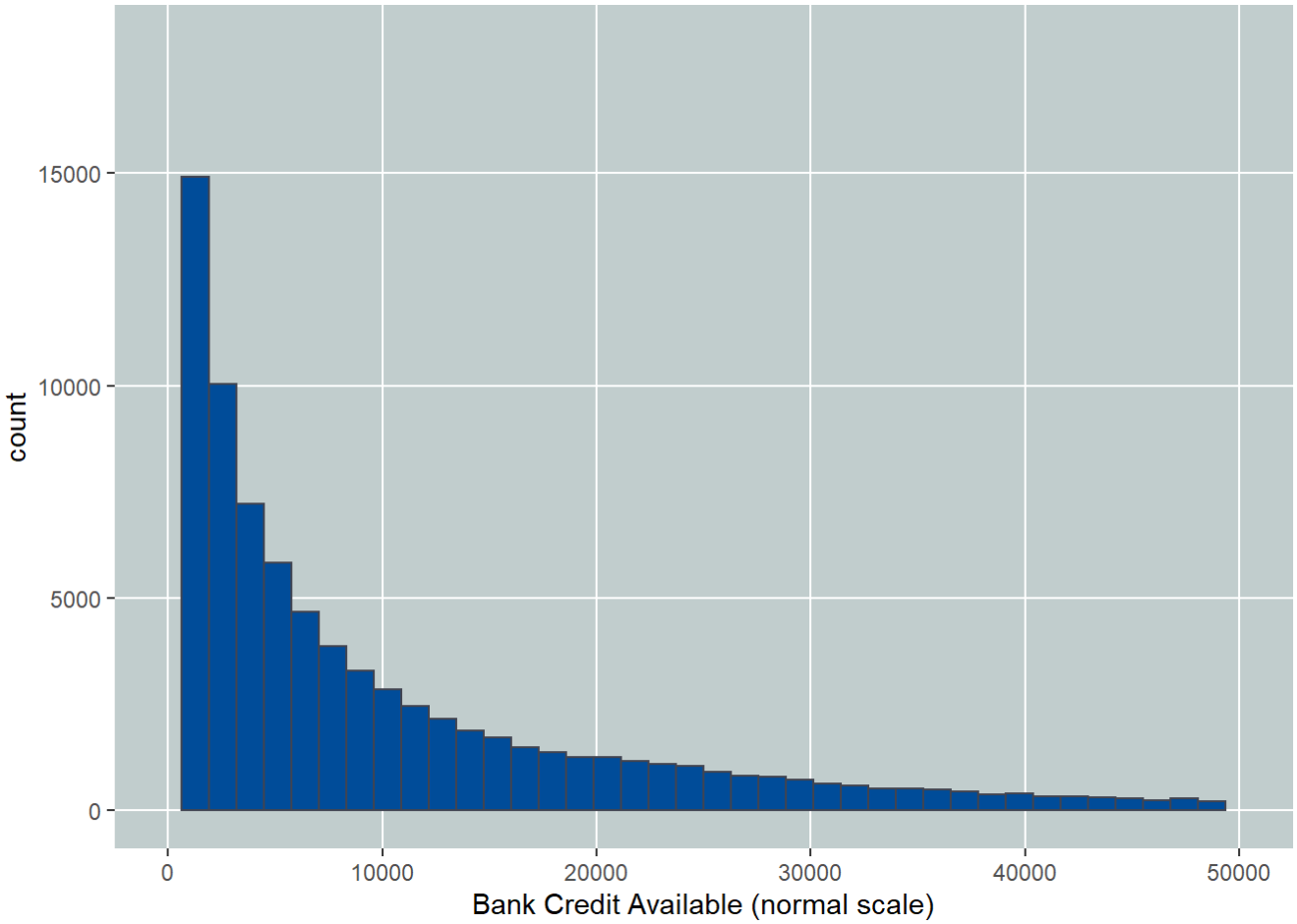


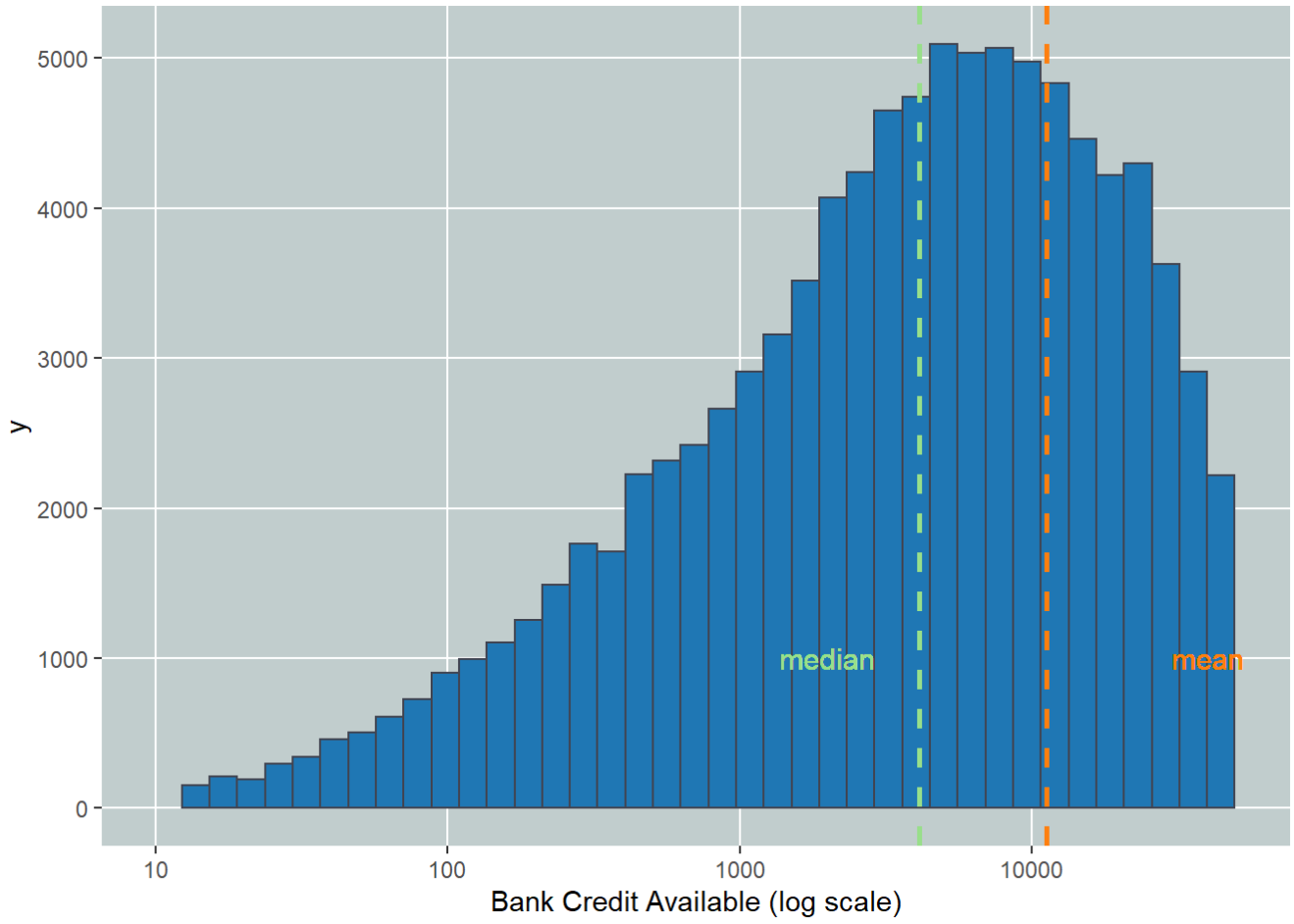


I chose normal scale since in a normal distribution 68% of the results fall within one standard deviation and

95% fall within two standard deviations.So I wanted to see where monthly Income falls.Interestingly, the histogram of all borrowers monthly income has a positive right-skewed long-tail distribution. I chose Log scale for monthly Income to show accurate monetary values counts along with Mean and Median values.On a log scale, the income data is centered around a median and mean of value of $4667 and $5608 respectively.

2.10 Bankcard Credit Available - (Histogram Normal and log scales distribution)



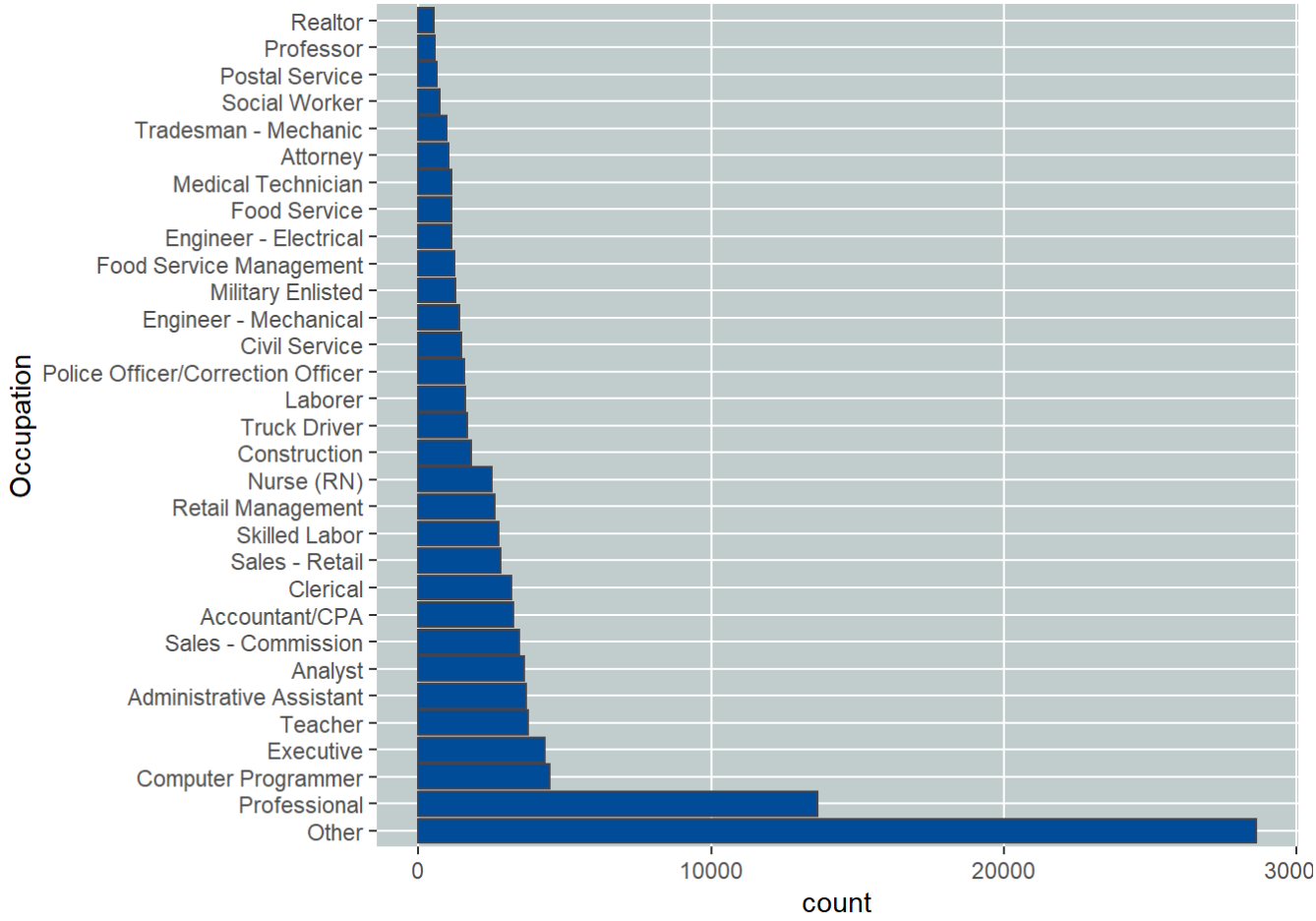


I chose normal scale initialy to see where BankCreditAvailable normally falls.Interestingly, the histogram of

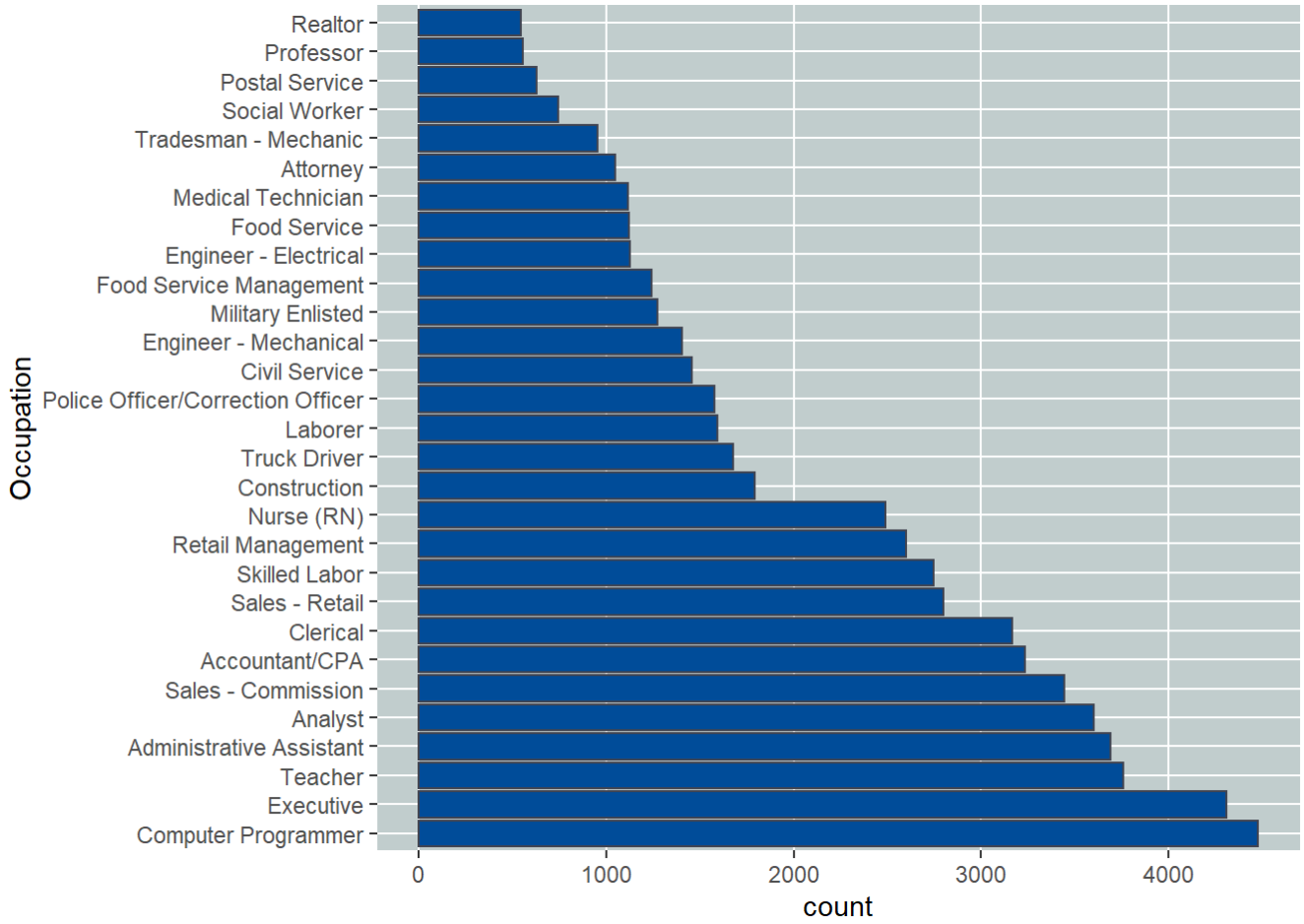
available bankcard credit data follows a left-skewed long-tail distribution trend. Additionally, I have used log scale and have adjusted the log limits to see clear mean and median values.Bank Credit available is left skewed with mean 10000 and median around 5000.

2.11 Occupation Categories by Least & Highest Counts

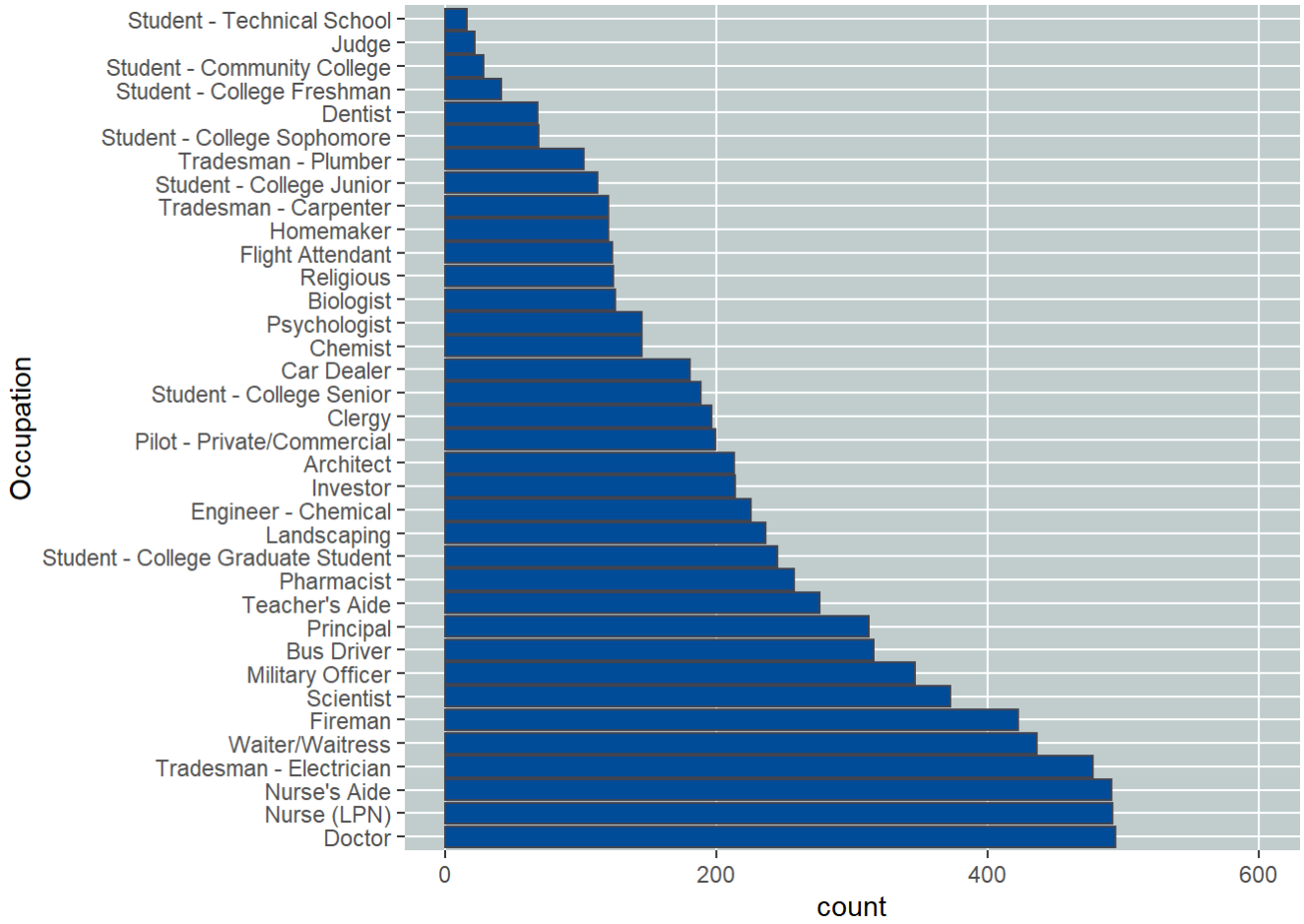
2.11.1 First Highest Range of Occupation Categories - (Histogram)



2.11.2 Second Highest Range of Occupation Categories - (Histogram)



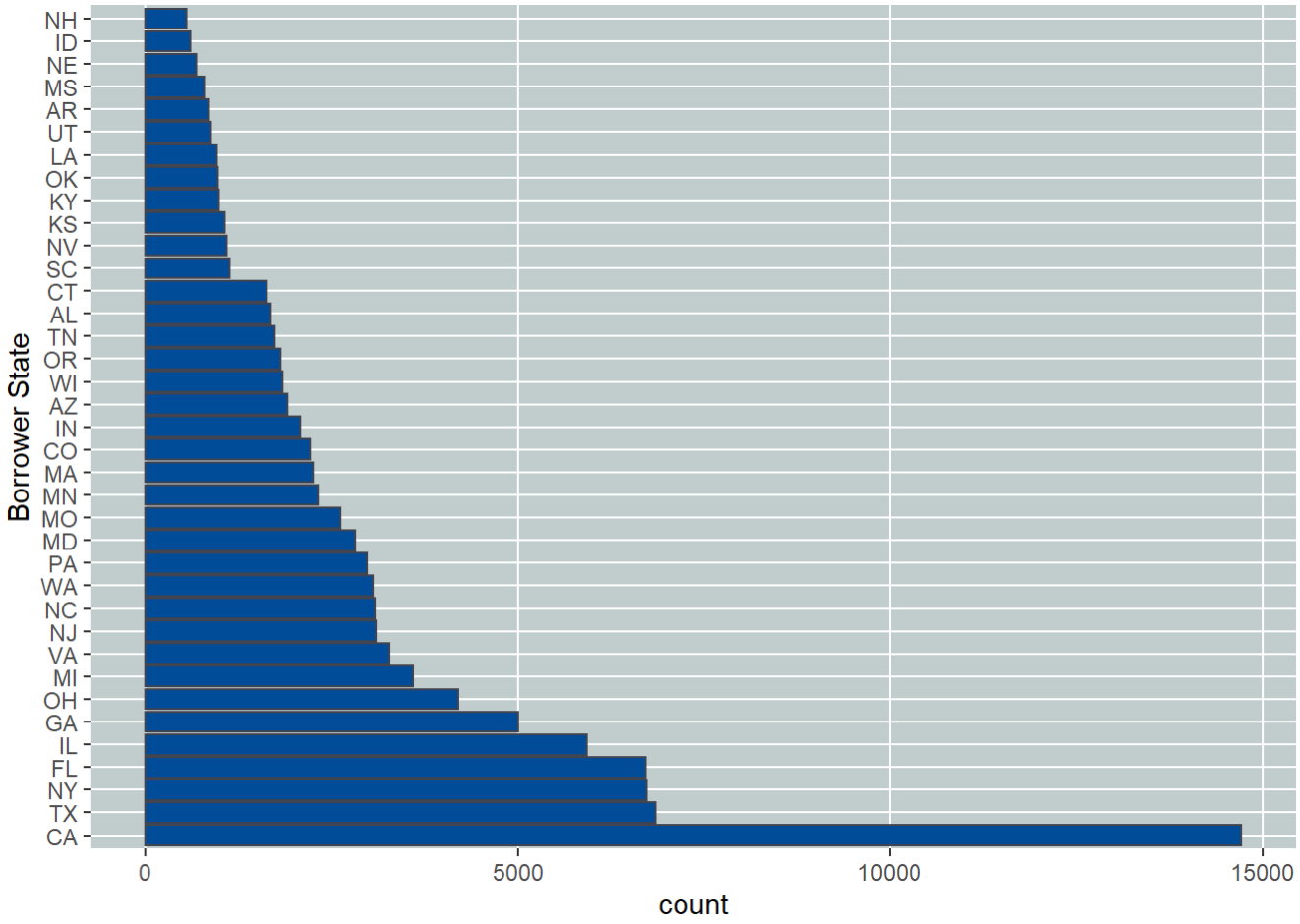
2.11.3 Lowest Range of Occupation Categories - (Histogram)

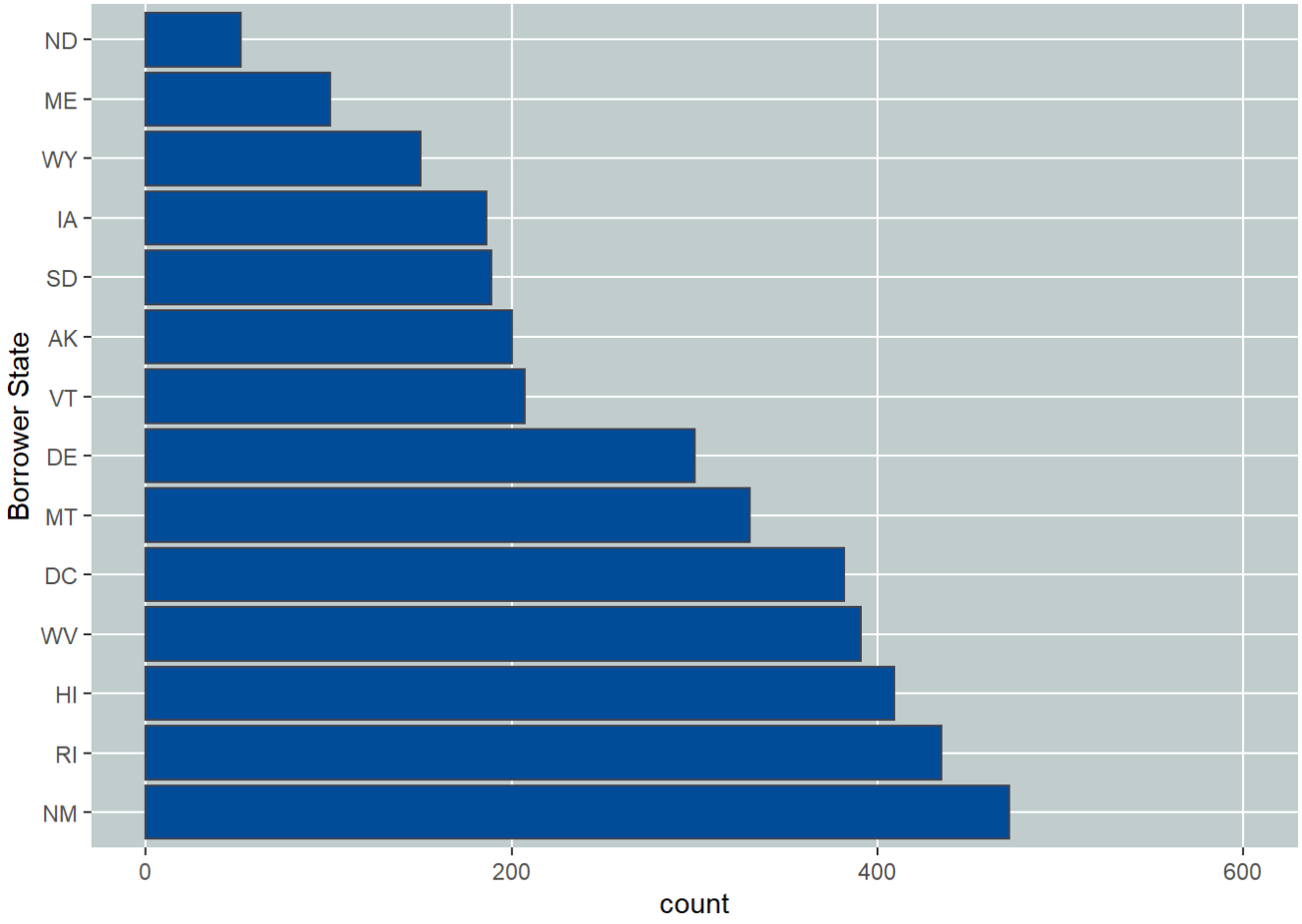


I chose histograms to show higher and smaller counts of occupation categories by subseting the data. Borrowers are mostly from 'Other' and 'Professional' Occupation categories.Second highest range of borrowers belong to 'Executive' and 'computer programmer' categories.The least common borrower occupation categories are of judge, dentist, and student.

2.12 Borrower States by Least & Highest Counts

2.12.1 Least & Highest counts by Borrower States- (Histogram)

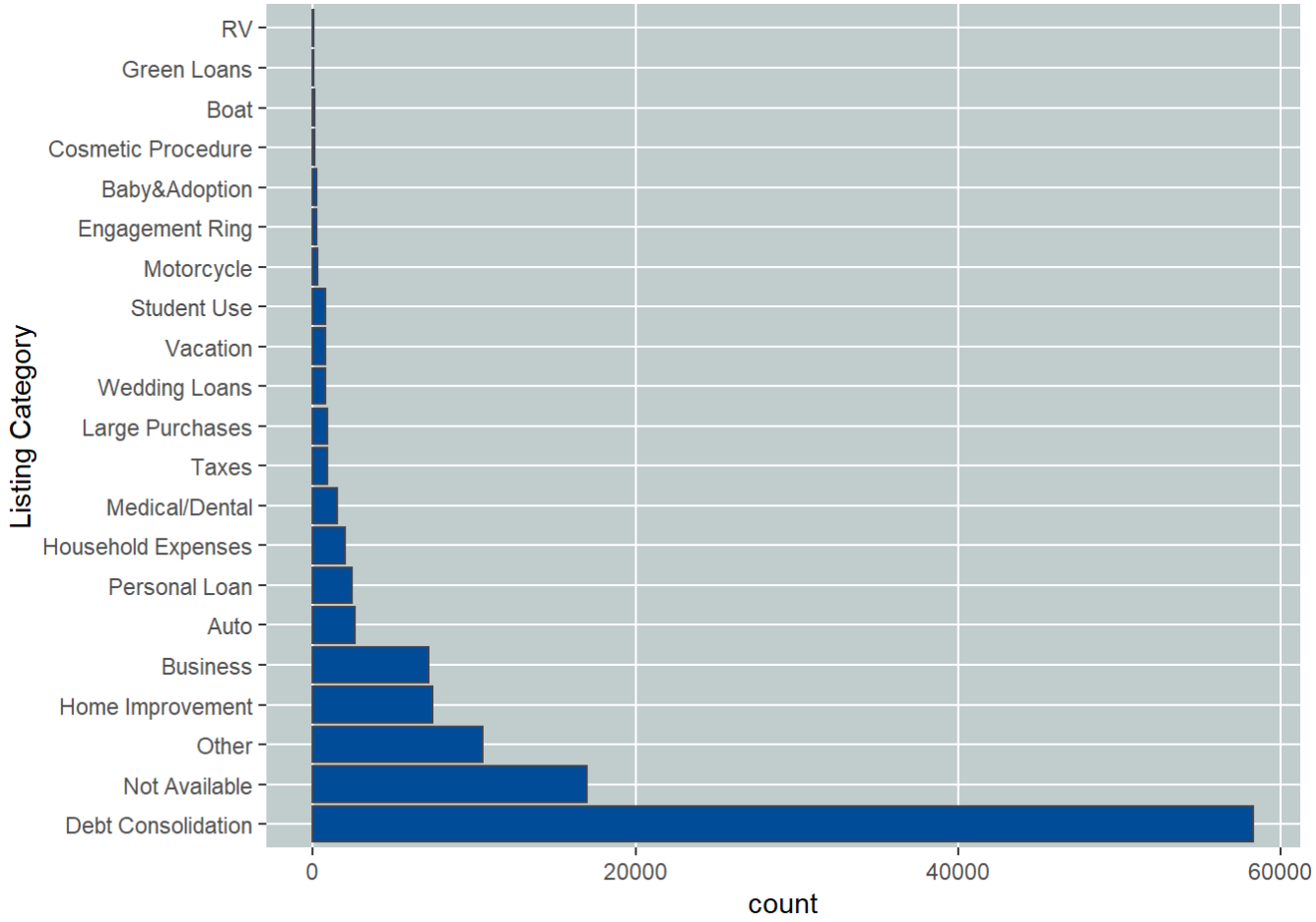




I chose histogram to visualize higher and smaller subsets of borrower states.From the above first graph it is

clearly evident that mostly Borrowers are from California state followed by Texas.Second graph shows that Borrowers are very least from North Dakota and Maine.

2.12.2 Listing Categories - (Re-Ordered Histogram)



I chose to show the distribution of Listing Category over normal scales and adjusted the graph to shows counts at x axis for better readability of listing categories at y axix.Above graph shows highest amount of loans were borrowed for 'Debt Consolidation' and 'other' categories.

3 Univariate Analysis

3.1 1. What is the structure of your dataset?

This data set contains 113,937 loans, with 81 features for each loan.I chose to filter out these features into 20 to see how borrowers data attributes like income, credit, state and occupation interact with the loan amount, term and APR.

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

Most interesting features of prosper loans dataset revolves around the loan term, monthly payment, and APR for a given loan amount. Another area of interest will be to look at the differences between the Credit Grade and the Prosper Ranking since they use the same ranking key (e.g. ‘C’, ‘B’, ‘A’, ‘AA’).

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

I would say investigation regarding borrower credit score and monthly income. Additionally, I would like to investigate the impact of occupation, listing category, available credit, and more.

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

The credit score for each loan in the original dataset was provided as a lower and upper credit bound. To make it more simple, I created a new variable ‘AvgCreditScore’ to take the average of the bound.

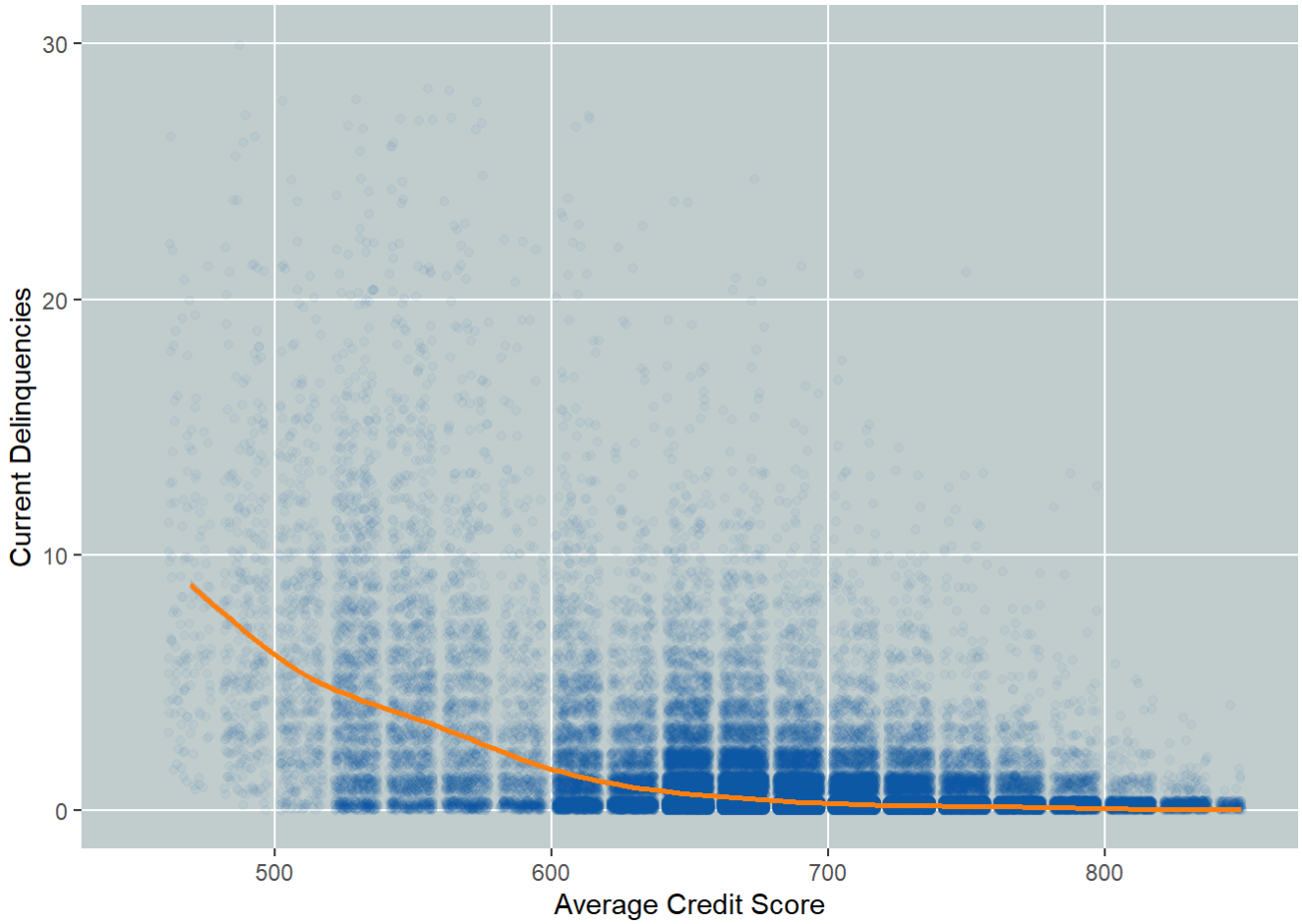
3.5 5. Of the features you investigated, were there any unusual distributions? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

The borrower APR has a normal distribution centered around 0.21%. However, anomalous peaks are seen in the histogram at 0.30% and 0.36%. The other interesting distribution in data was for the loan amount. The most frequent loan amounts occur at nicely rounded intervals such as $10000 and $15000. The Prosper Rating and credit grade factors were rearranged to put the “AA” at a higher ranking than the “A” rating. The “Listing Category” factors were renamed to the actual category name instead of a number key. The “Listing Category” and “Term” were converted from numeric to factor types for simplicity and ease of plotting.

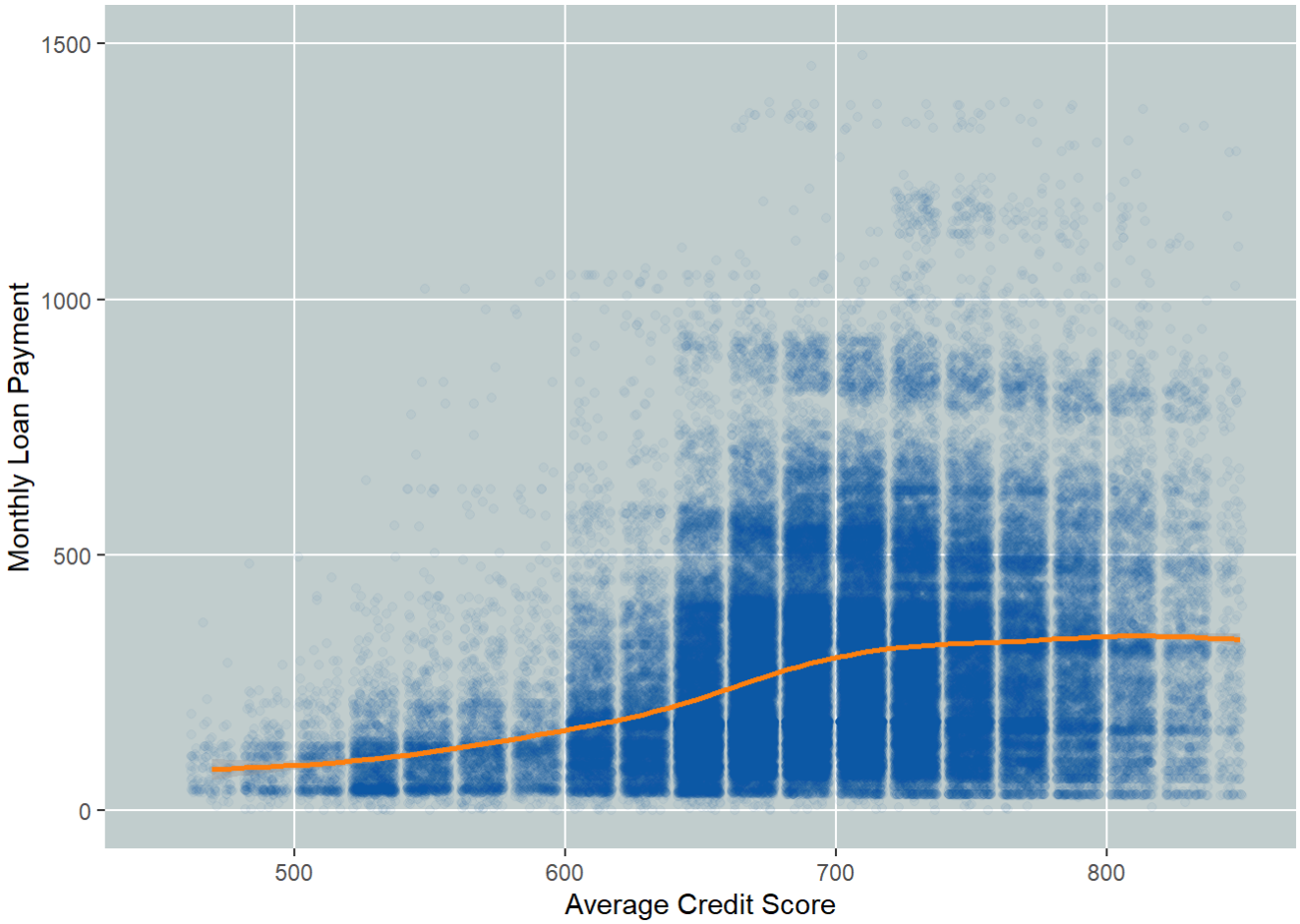
4 Bivariate Plots Section

4.1 Plot the comparison of Credit Scores with CurrentDelinquencies, MonthlyLoanPayment, CurrentCreditLines & BankCreditAvailable

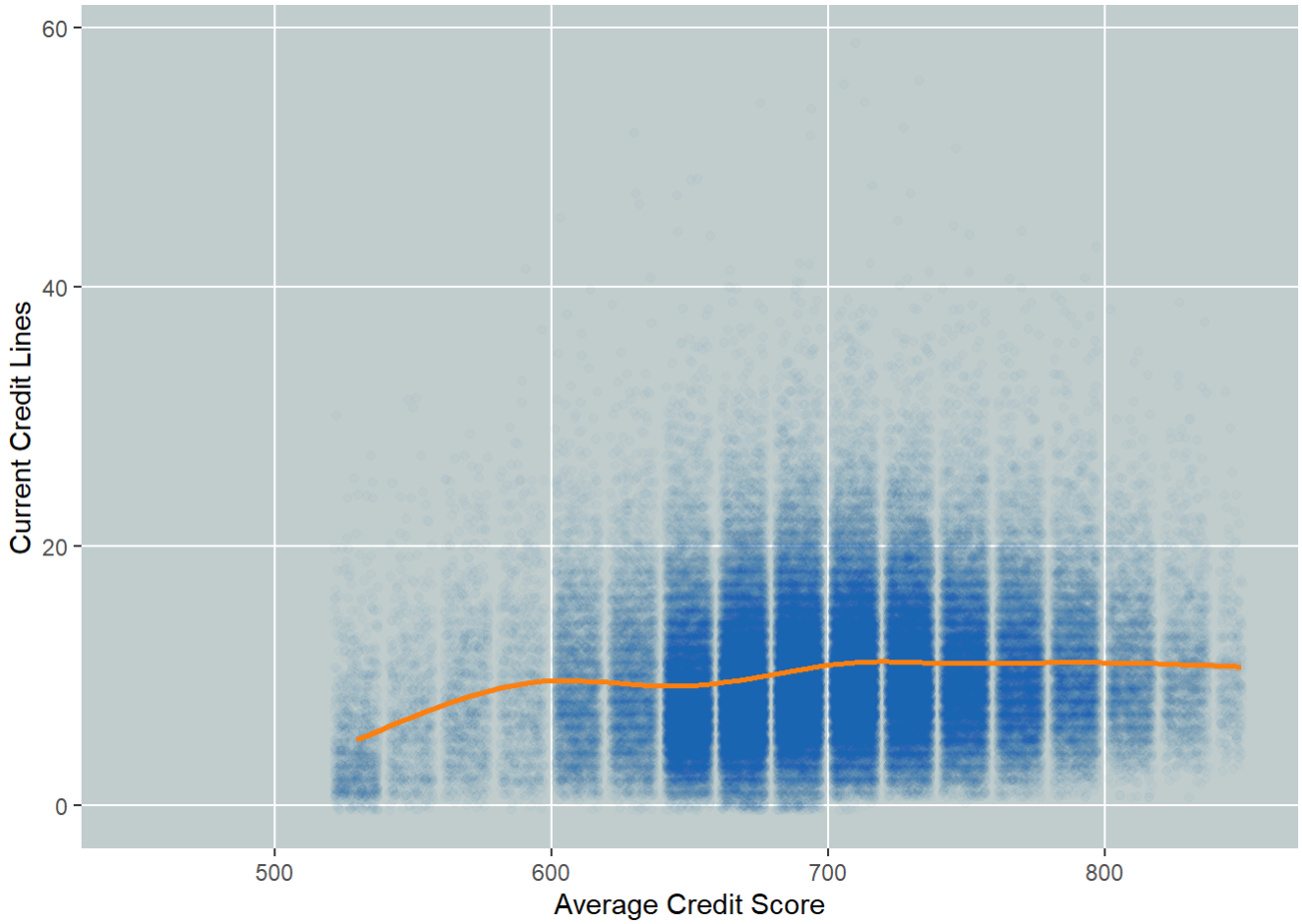
4.1.1 Average Credit Score vs Current Delinquencies - (Jitter Plot)



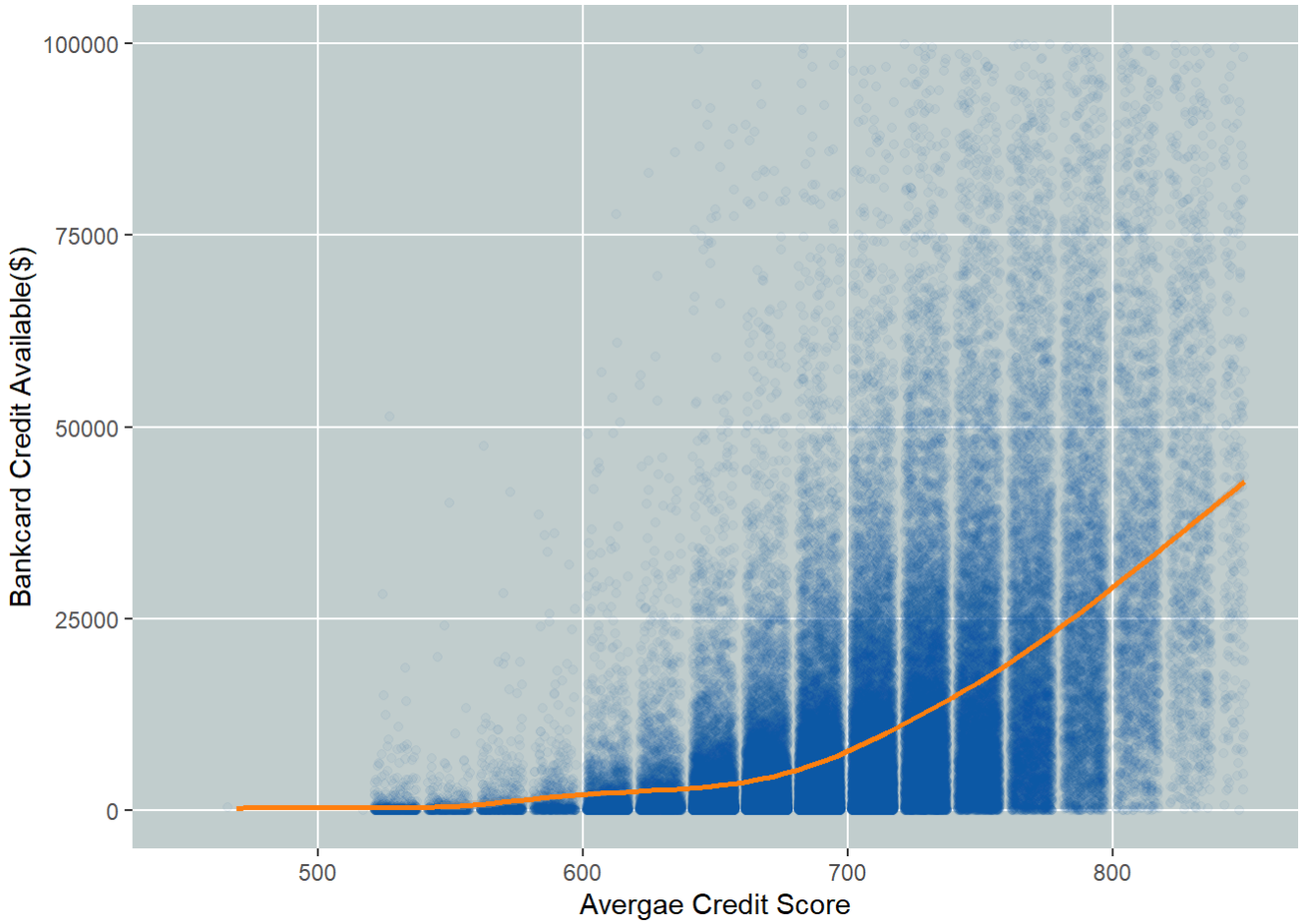
4.1.2 Average Credit Score vs Monthly Payment - (Jitter Plot)



4.1.3 Average Credit Score vs Current Credit Lines - (Jitter Plot)



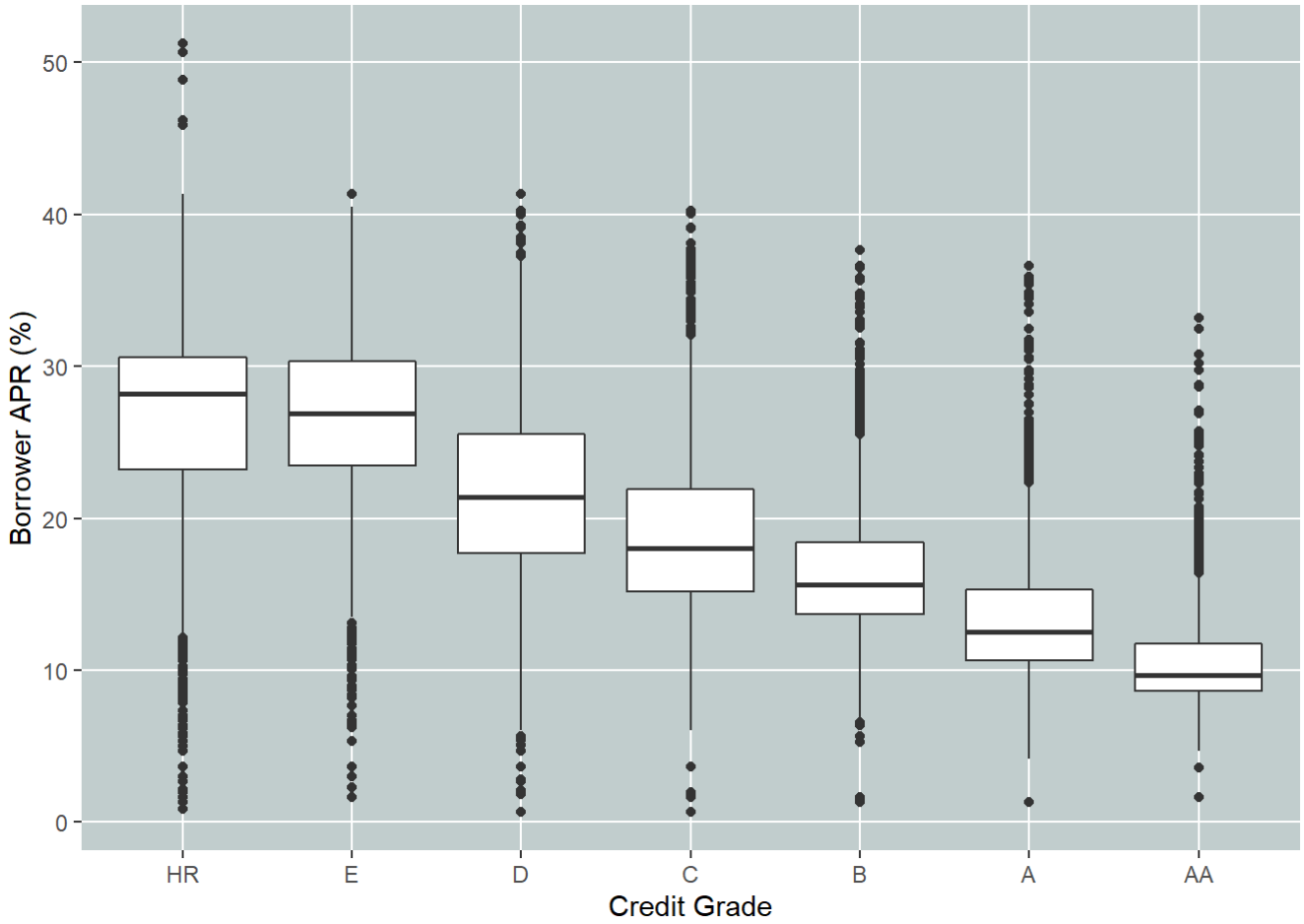
4.1.4 Average Credit Score vs Bank Credit Available - (Jitter Plot)



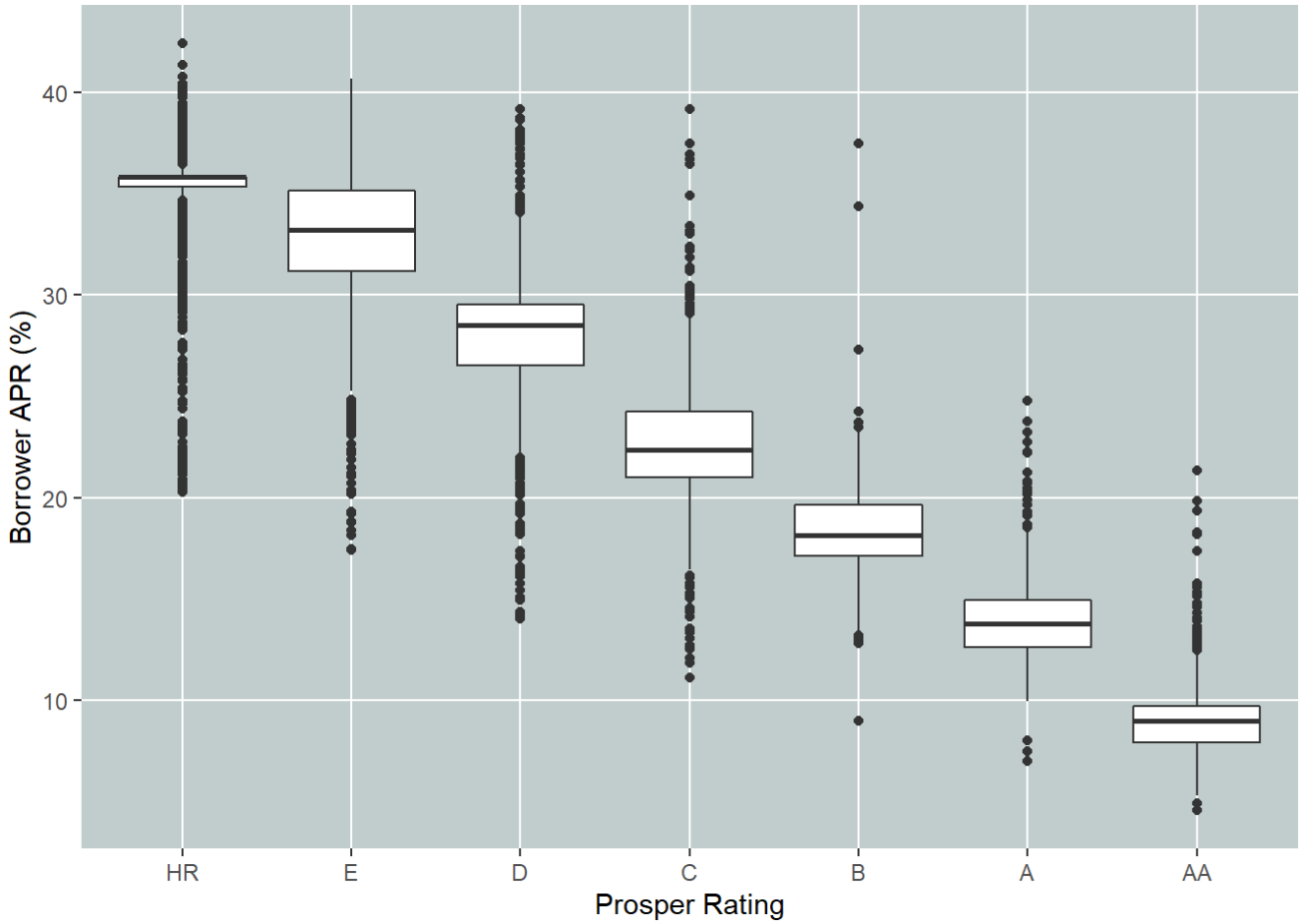
I have used “Jitter” plots as these are the best plots to prevent any sort of over plotting in statistical graphs.I have added random intervals to show accurate results. Credit delinquencies shows a weak negative correlation on the borrowers average credit score with some outliers.For example, a few borrowers with credit scores greater that 750 had more than 10 current delinquencies and at the same time very few borrowers with credit score than 600 had more than 25 current delinquencies. Borrowers with low credit (<600) scores tended to have lower monthly payments (and loan amounts) than those with higher credit scores (>600). Higher the credit score higher the payments. Credit lines did not show much correlation with credit score; the average was about 10 credit lines across the range. Bankcard credit available showed a clear positive correlation with credit score.

4.2 Plot the comparison of Credit Grade vs Prosper Rating based on Borrowers APR and Credit Score

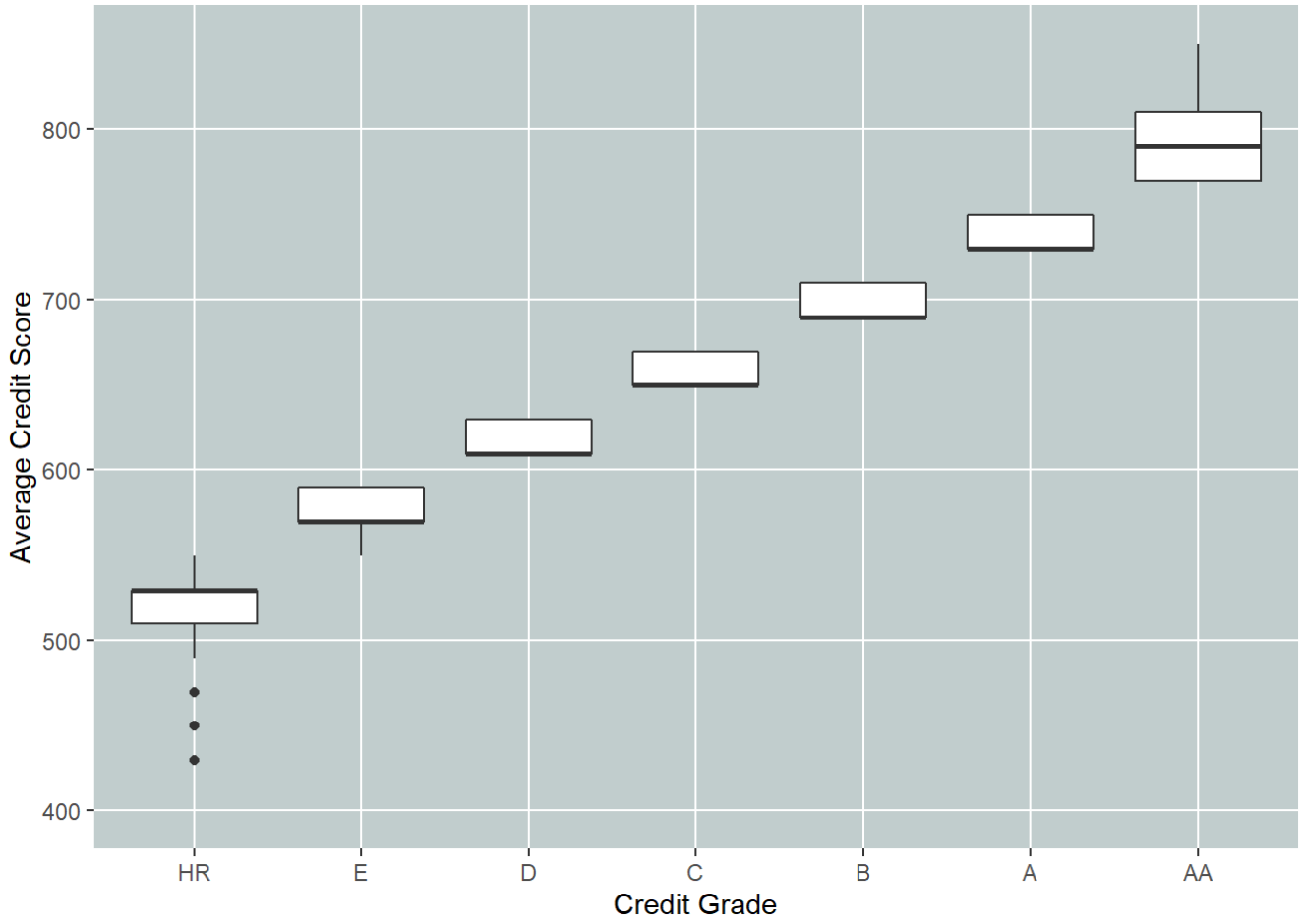
4.2.1 Borrowers APR by Credit Grade - (Box Plot)



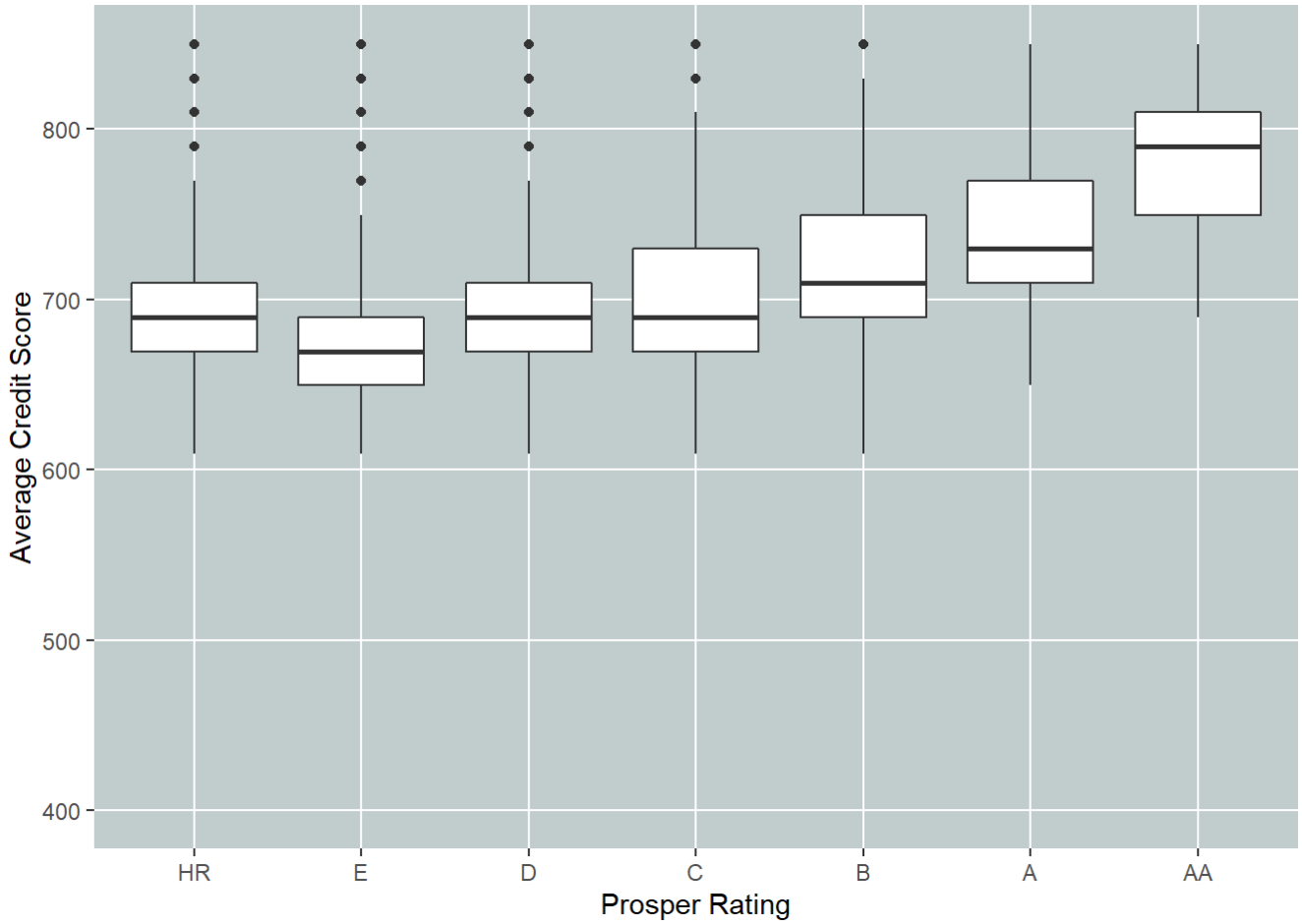
4.2.2 Borrowers APR by Prosper Rating - (Box Plot)



4.2.3 Credit Score by Credit Grade - (Box Plot)



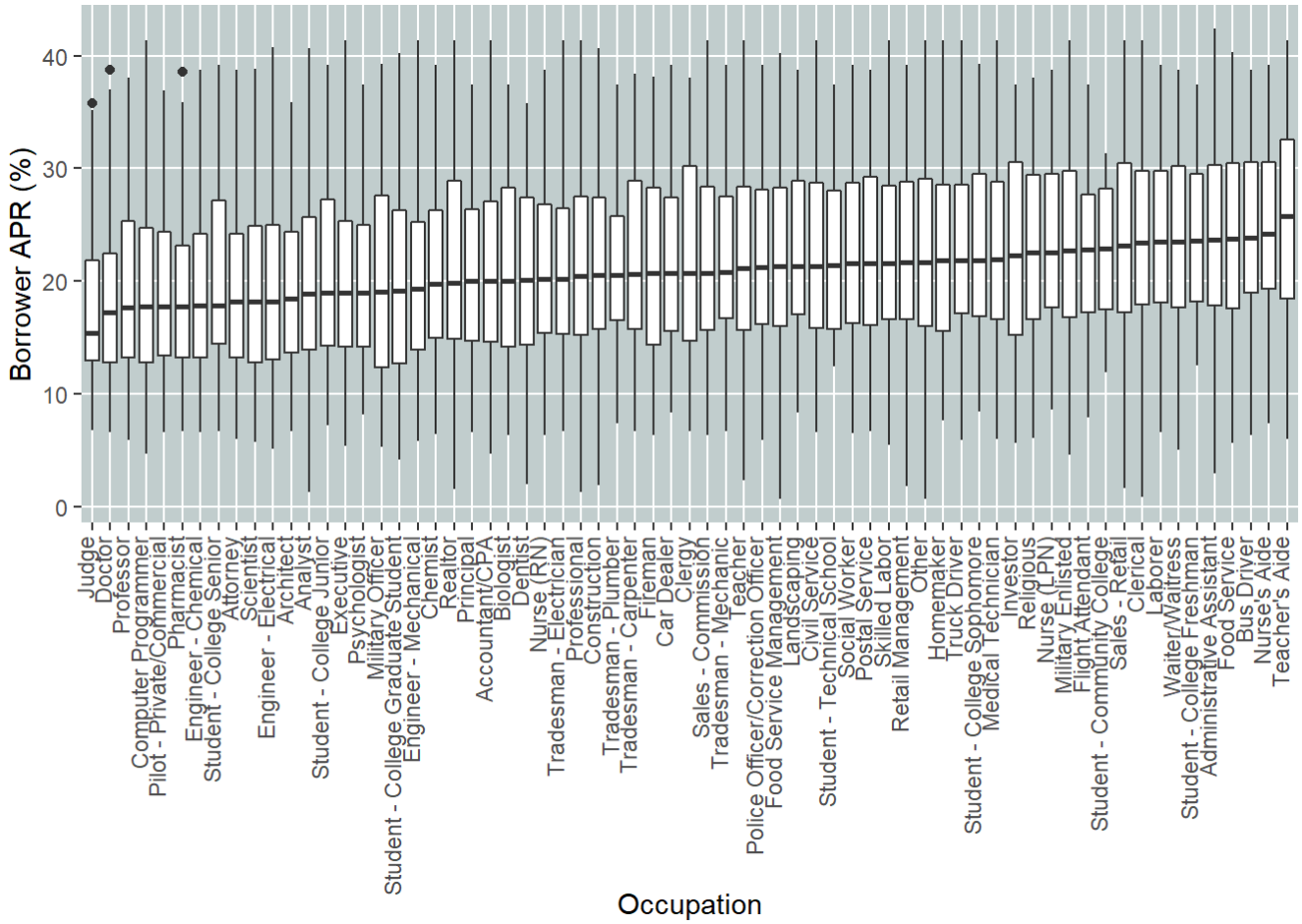
4.2.4 Credit Score by Prosper Rating - (Box plot)



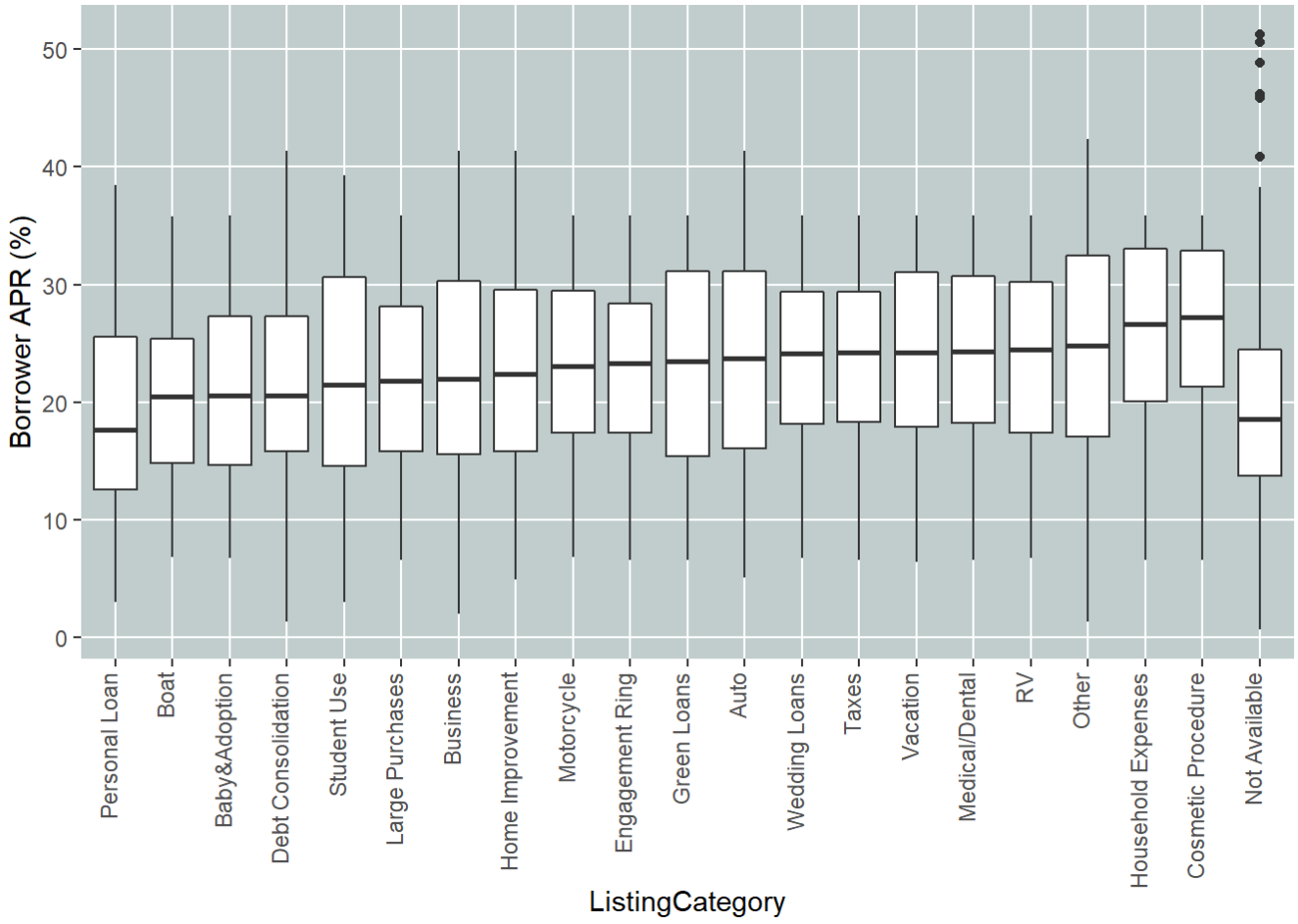
I chose box plots above to show the comparison of categorical variables “credit Grade” & “Prosper Rating”. The credit grade and prosper rating differed significantly with respect to credit score and borrower APR. Looking at borrower APR in terms of credit grade shows interesting histograms results.There seems to be a pattern where for each credit grade, there are several standard rates for APR. Most interestingly, the Prosper rating lines up in very clear stepping order with respect to APR. With respect to Credit score we can see that credit grade ‘HR’ has lower credit scores i.e <500. Higher the Prosper rating, higher credit scores can be seen.

4.3 Plot the comparison of Borrowers APR over Occupation, Listing Category, Monthly Income & Term

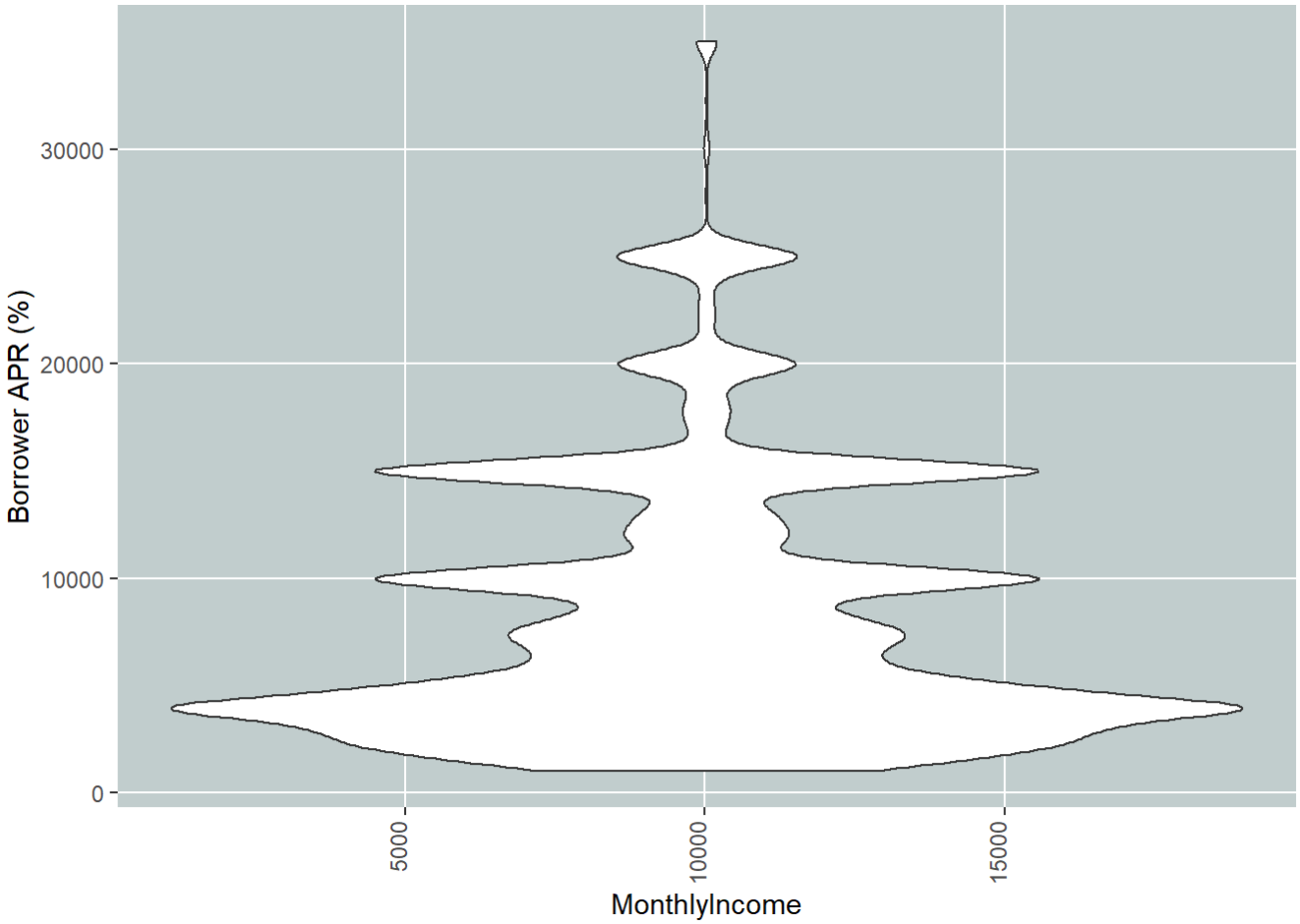
4.3.1 Borrower APR by Occupation - (Box Plot)



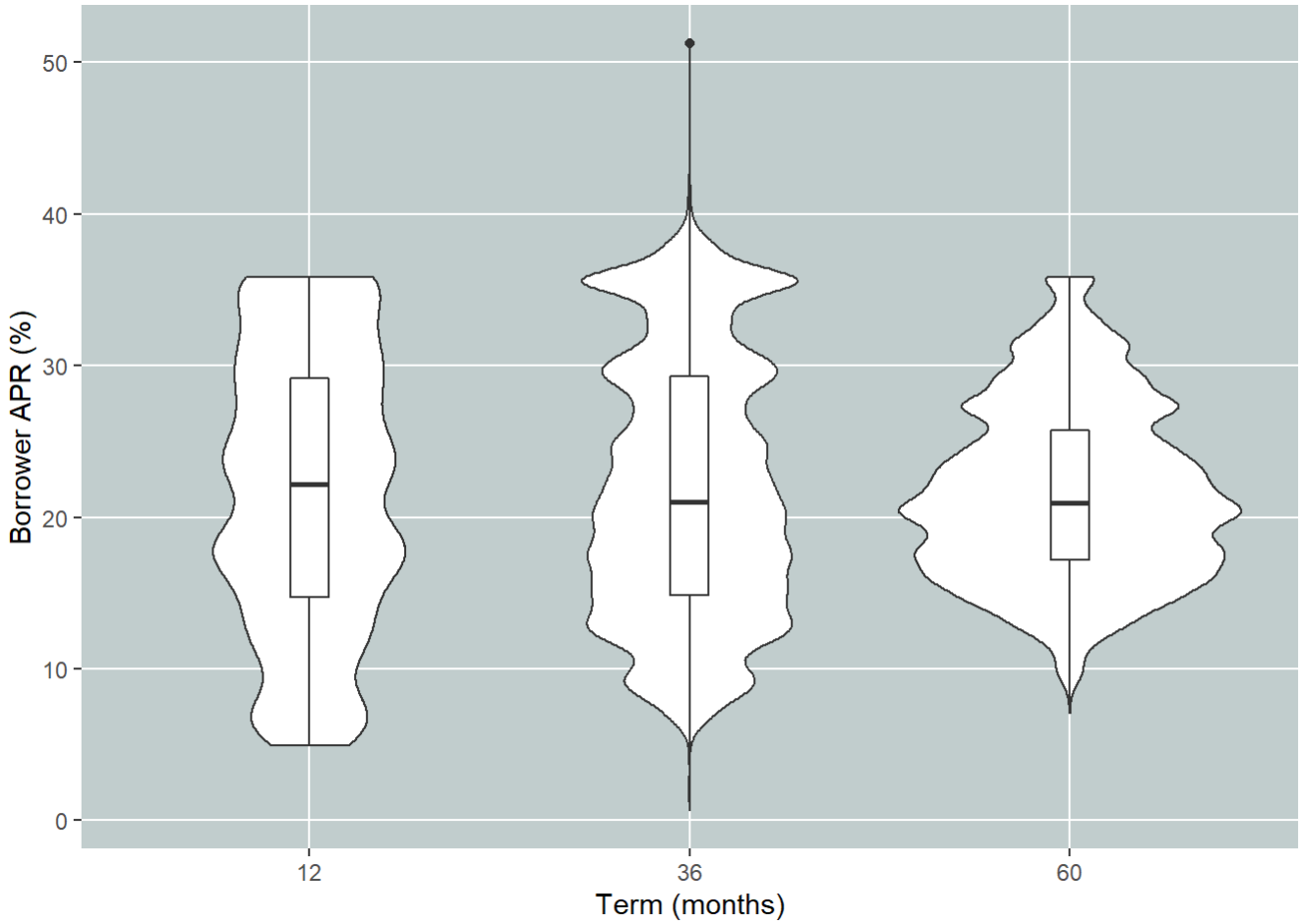
4.3.2 Borrower APR by Listing Category - (Box Plot)



4.3.3 Borrower APR by Monthly Income - (Violin Plot)



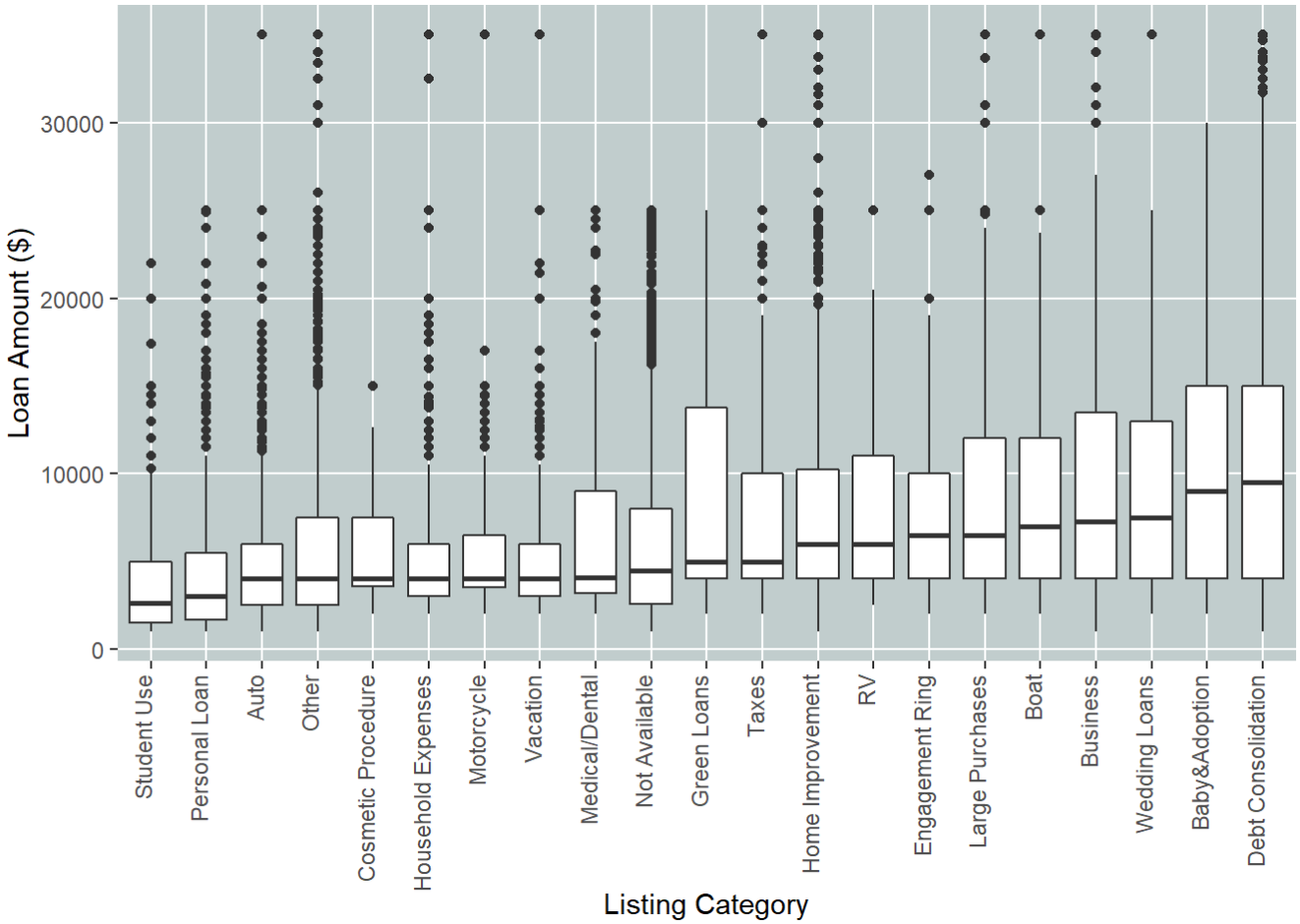
4.3.4 Borrower APR by Terms- (Violin & Box Plot)



I chose box plots above to show the Borrower APR percentage across different categorical variables like “Occupation” & “Listing Category”. I chose violin plots as they are used to show the kernel probability density of the data at different values. I have used here to show the Borrower APR percentage by Monthly Income and Terms and there probability density. The three occupations with the highest median APR are teacher’s aide,nurse’s aide and bus driver. Three occupations with the lowest median APR are are Judge, Doctor, and Professor. Monthly income, past $10k, didn’t have a huge factor on APR. Each loan term had a similar mean/median value in relation to the APR.

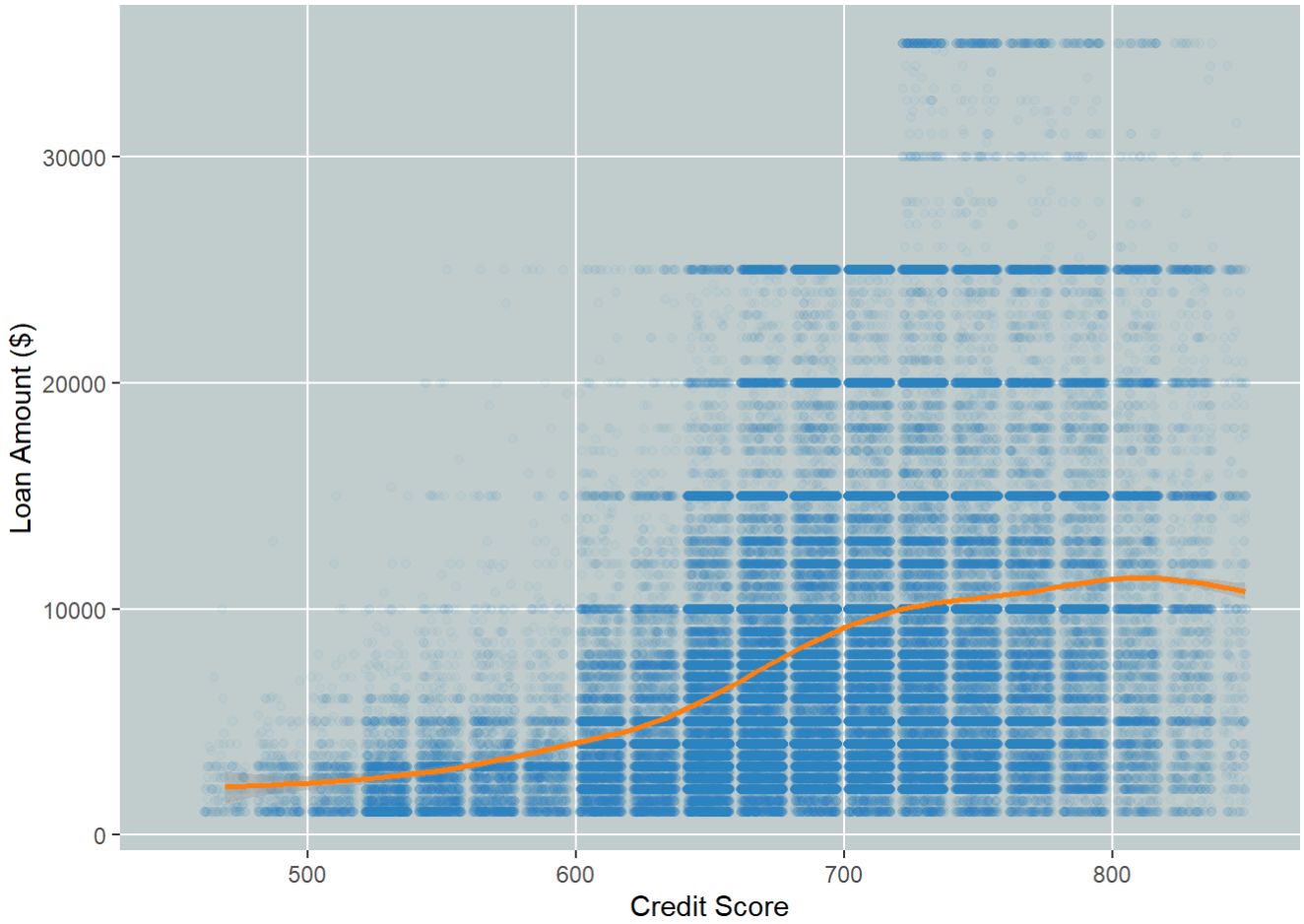
4.4 Plot the comparison of Loan amount over Listing Category, Monthly Income and Credit Score

4.4.1 Loan amount by Listing Category - (Box Plot)



4.4.2 Loan amount by Monthly Income - (Points)

4.4.3 Loan amount by Average Credit Score - (Jitter Plot)



I chose geom point to create the scatter plots and show distribution of Lon amount by monthly income I chose box plot with function median to show minimum, first quartile, median, third quartile, and maximum values. I chose Jitter to visually determine the degree and type of correlation between the Loan amount and average credit score. The median loan amounts for the baby/adoption and debt consolidation categories were higher than the other categories. For less than a monthly income of $10k, there was a weak correlation between monthly income and loan amount. Borrowers with higher credit scores (greater than 650) tended to take larger loans.

5 Bivariate Analysis

5.1 1.Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

Initial sets of plots focused on comparing credit score with other data set features.
First, I noticed that lower bankcard credit and high delinquencies had a negative effect on the credit score.Secondly, Credit lines did not show any significant relationship with credit score and Lastly, Credit grade had a direct relationship with credit score but not that great with the Prosper rate.
The next set of data set explorations focused on Borrower APR.

First, Credit grade had a direct relationship with APR and the Prosper rating was quite overlapped with APR.Secondly, In the listing category, median loan amounts for the baby/adoption and debt consolidation categories were higher than the 'other' categories.

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

It was noticed that each loan term had a similar distribution and median relationship with APR. I would say longer the loan term(months),higher the APR. The highest APR in terms of listing category was cosmetic procedure. I would say even automotive or medical should come in the top list due to its nature of urgent need.

5.3 3.What was the strongest relationship you found?

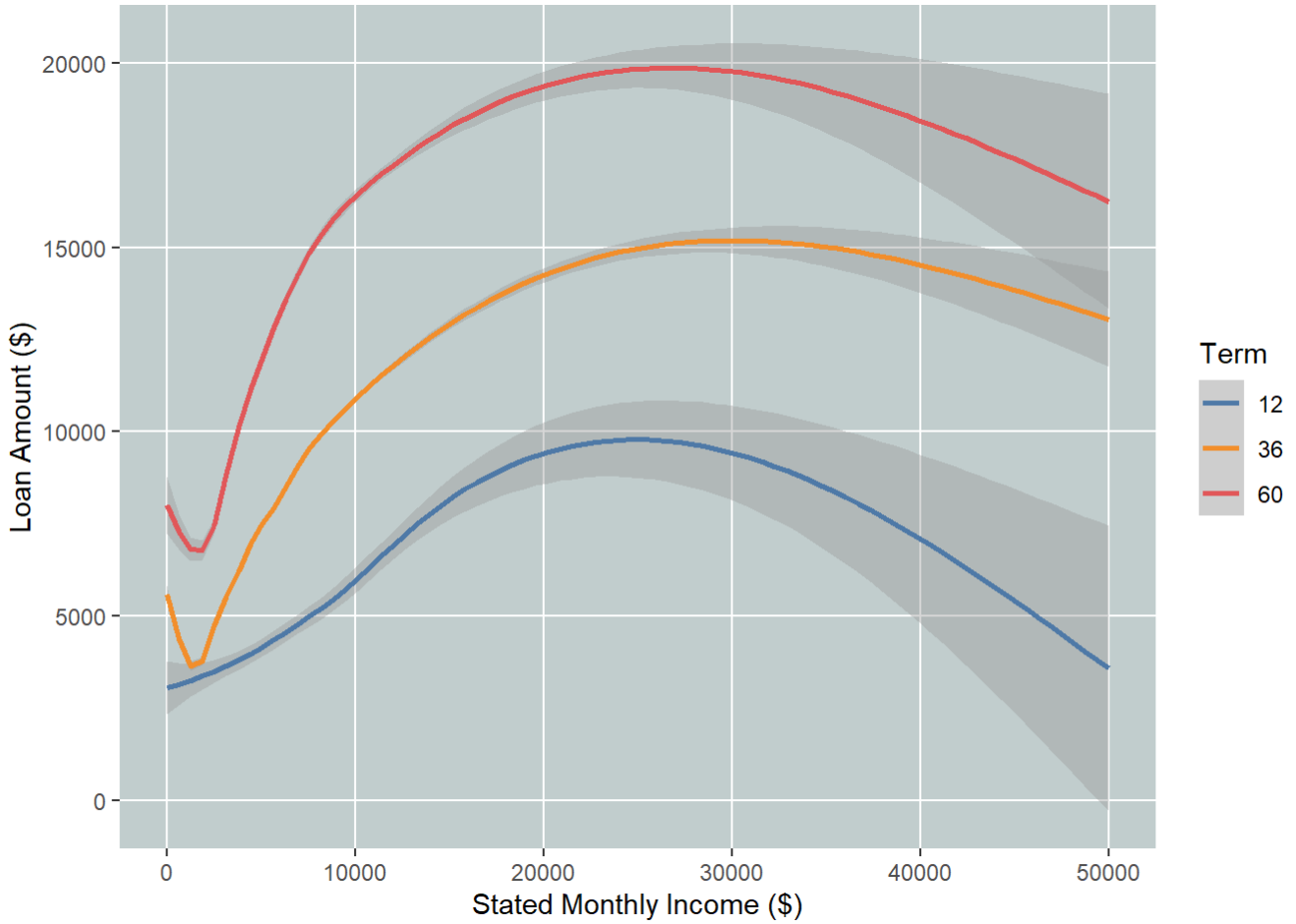
The strongest relationship was between the Credit grade and APR From a categorical perspective, the strongest relationship was between the Prosper rating and the APR.

6 Multivariate Plots Section

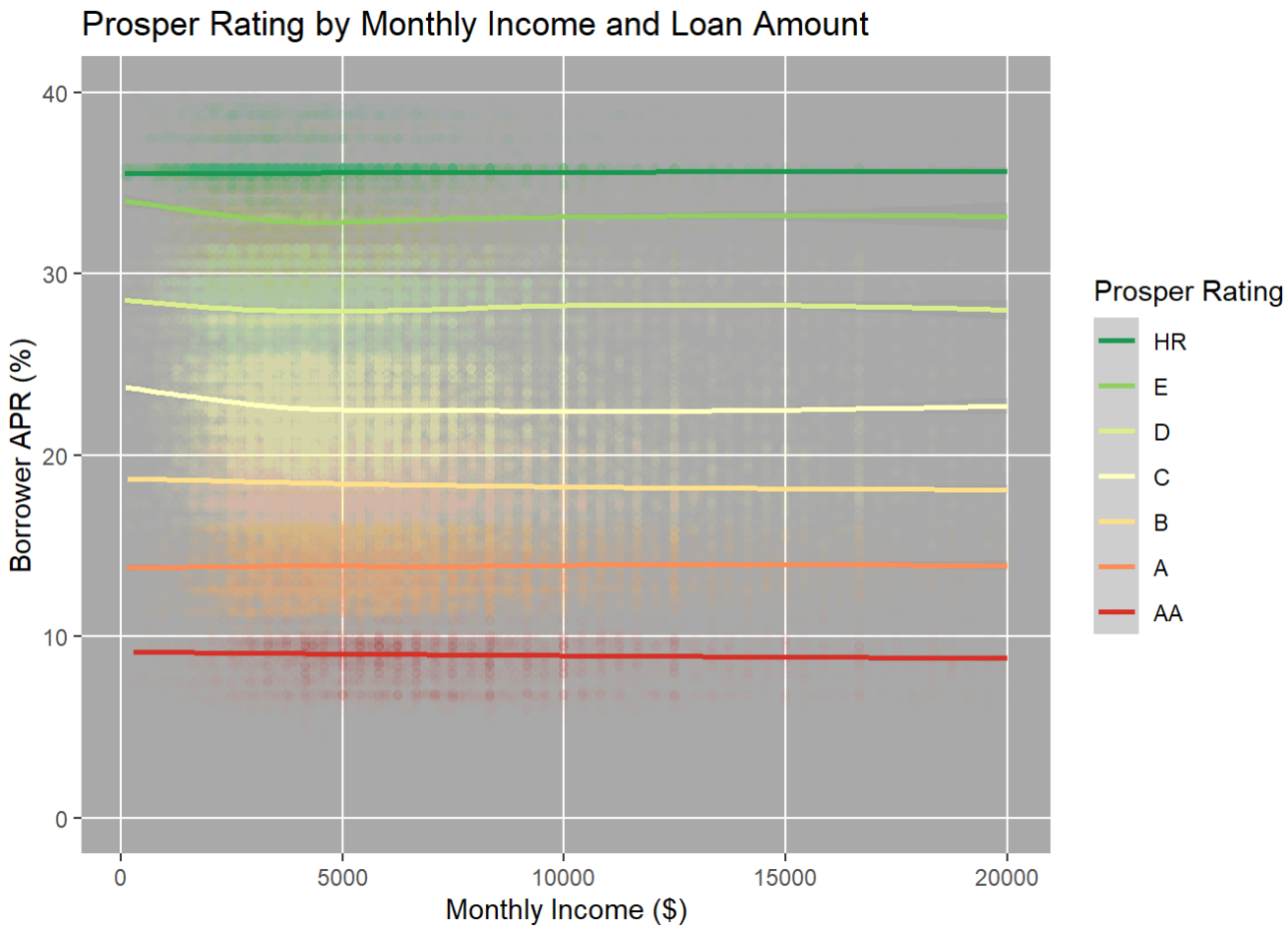
The most interesting relationships in the prior sections are related to the APR, monthly income, and Prosper ratings. In this Multivariate analysis series, I would like to explore 'Monthly Income' across the Term, Loan amount, APR, Prosper Rating & Credit Score.

6.1 Plot the comparison of Monthly Income over Loan amount, Term ,APR and Propser Rating

6.1.1 Term trend across Monthly Income and Loan Amount

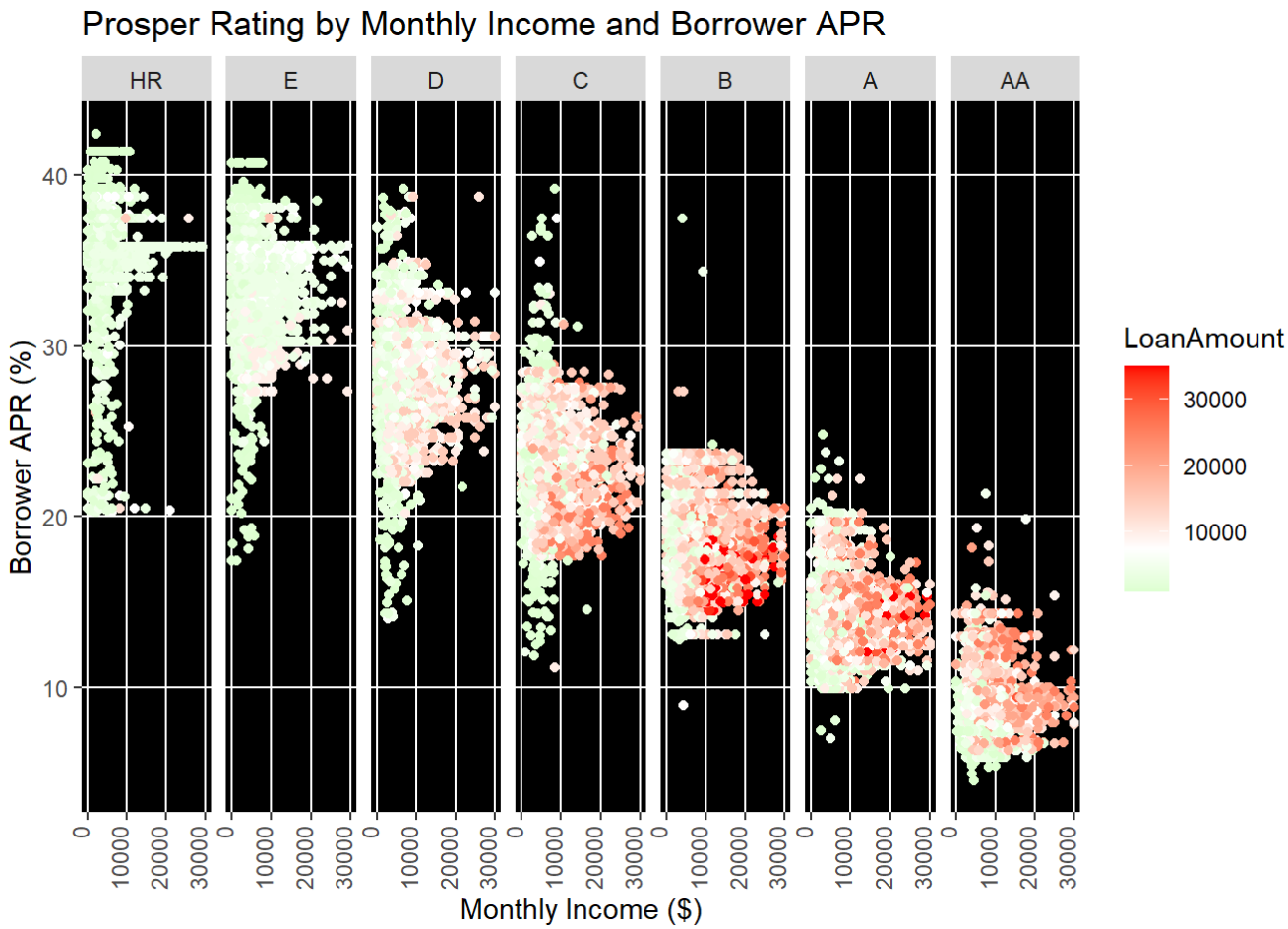


6.1.2 Prosper Rating by Monthly Income and Loan Amount


Prosper Rating by Monthly Income and Loan Amount
Borrower APR (%)
40
30
20
10
0
0
5000
10000
15000
20000
Monthly Income ($)
Prosper Rating
HR
E
D
C
B
A
AA

```
## <ggproto object: Class ScaleDiscrete, Scale, gg>
##     aesthetics: colour
##     axis_order: function
##     break_info: function
##     break_positions: function
##     breaks: waiver
##     call: call
##     clone: function
##     dimension: function
##     drop: TRUE
##     expand: waiver
##     get_breaks: function
##     get_breaks_minor: function
##     get_labels: function
##     get_limits: function
##     guide: legend
##     is_discrete: function
##     is_empty: function
##     labels: waiver
##     limits: NULL
##     make_sec_title: function
##     make_title: function
##     map: function
##     map_df: function
##     n.breaks.cache: NULL
##     na.translate: TRUE
##     na.value: NA
##     name: waiver
##     palette: function
##     palette.cache: NULL
##     position: left
##     range: <ggproto object: Class RangeDiscrete, Range, gg>
##         range: NULL
##         reset: function
##         train: function
##         super:  <ggproto object: Class RangeDiscrete, Range, gg>
##     rescale: function
##     reset: function
##     scale_name: tableau
##     train: function
##     train_df: function
##     transform: function
##     transform_df: function
##     super:  <ggproto object: Class ScaleDiscrete, Scale, gg>
```

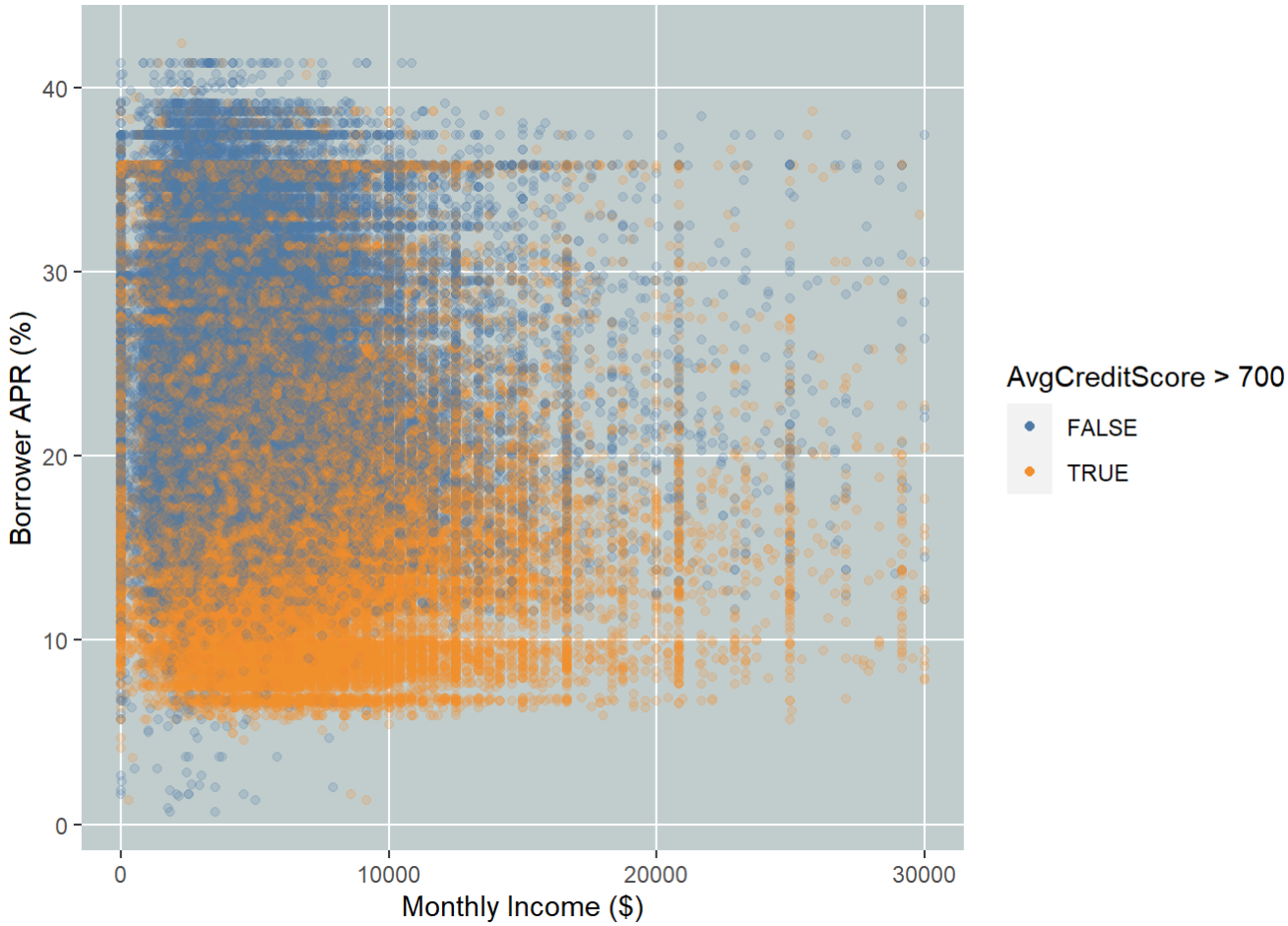
6.1.3 Prosper Rating by Monthly Income and Borrower APR



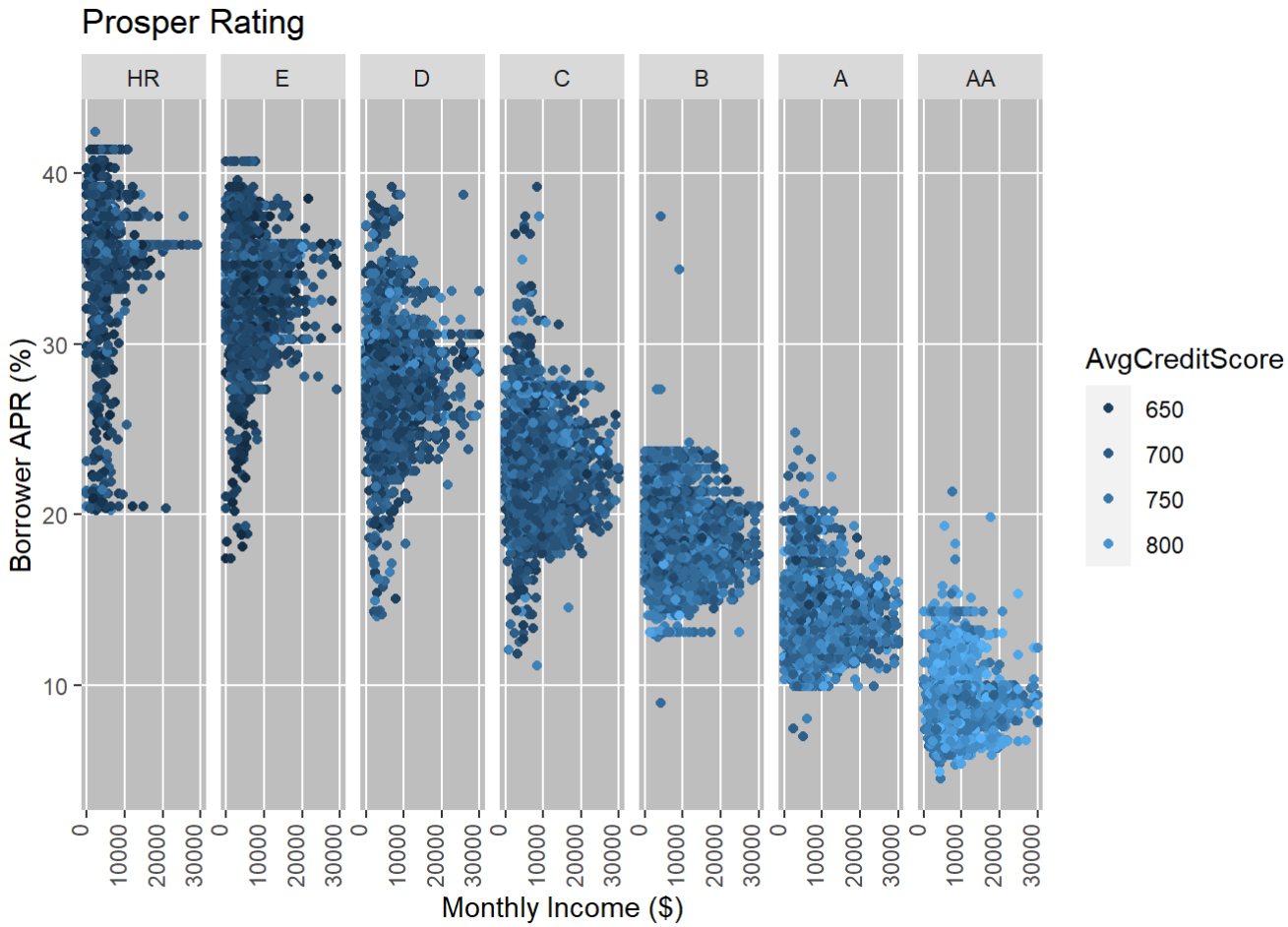
I chose geom_point to create scatter plots to show relationships between Stated Monthly Income with Borrower APR and Prosper Rating without Facet and relationships between Stated Monthly Income, Borrower APR and Loan amount with Facet as Prosper Rating. Comparing the loan amount, monthly income and term shows a very clear trend where larger loans are often associated with longer loan terms. In the bivariate plots section, clear relationship was evident between the Prosper rating and the APR.Interestingly it can be noticed here that monthly income does not play a large role in the Prosper/APR relationship.

6.2 Plot the comparison of Average credit score over Monthly Income and Borrower APR

6.2.1 Average credit score (>700) by Monthly Income and Borrower APR



6.2.2 Average credit score (<850) by Monthly Income and Borrower APR - Facet over Prosper Rating



6.2.3 Average credit score (>700) by Monthly Income and Borrower APR- Facet over Prosper Rating

I chose geom point to create scatter plots to show relationships between Stated Monthly Income with Borrower APR and Average Credit Score without Facet and relationships between Stated Monthly Income with Borrower APR, Average Credit Score and Prosper Rating with Facet as Prosper Rating.The median credit score for this data set was 700. It can be noticed here that Higher credit scores always have lower APR and higher Prosper ratings. Borrowers with larger loan amounts have higher Prosper ratings. Likewise, higher the bank card credit,better the APR and Prosper rating.

7 Multivariate Analysis

7.1 1. Talk about some of the relationships you observed in this part of the investigation. Were there features that strengthened each other in terms of looking at your feature(s) of interest?

The APR & Prosper rating feature were of the main interest in this section. Higher credit score, higher bankcard credit and higher monthly income all had clear and strong relationships with the borrowers Prosper rating and APR.

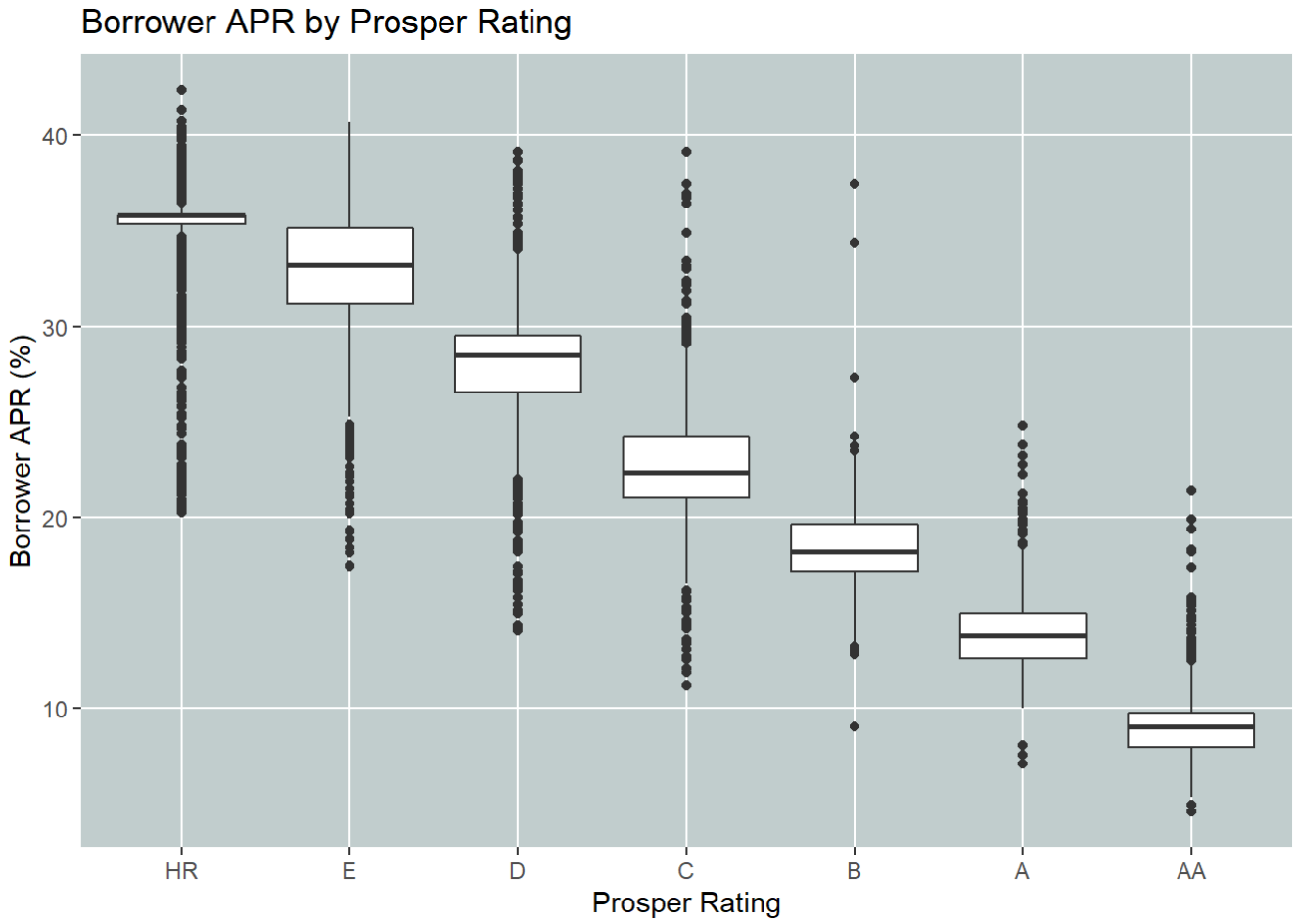
7.2 2. Were there any interesting or surprising interactions between features?

Home ownership status had little impact on the Prosper rating & APR. It is most likely that Prosper shows more interest in credit score, bankcard credit & monthly income of the borrower as opposed to the home ownership status.

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

8 Final Plots and Summary

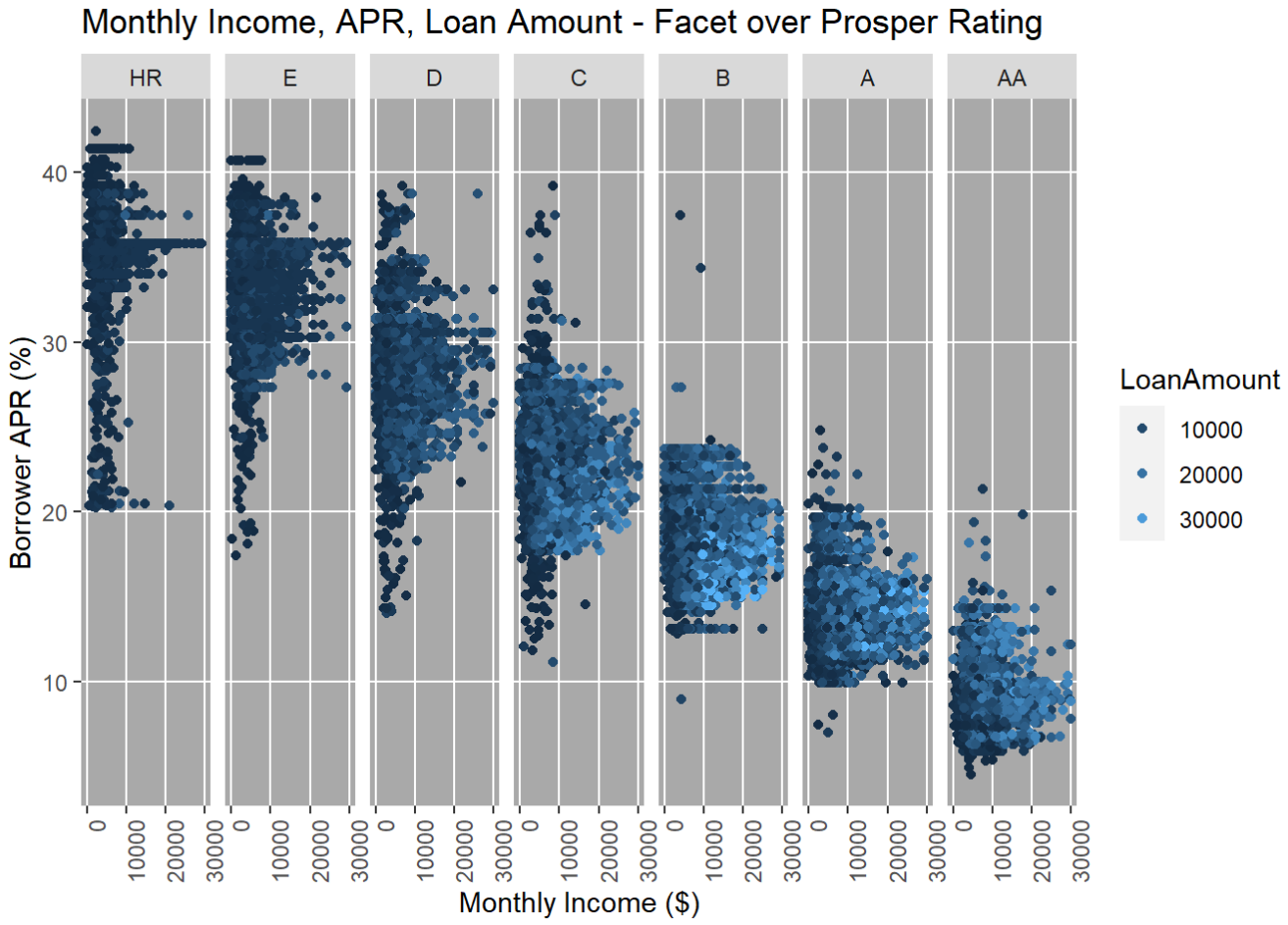
8.1 Plot One - Borrower APR by Prosper Rating



8.1.1 Description One

I chose Box plot for above graph to visualize APR data over categorical variable 'Prosper Rating'. Getting a loan with a lower APR is what most borrowers target. Once borrower is assigned a Prosper rating, it is clearly evident what APR range could be expected. "Higher the prosper rating lower the APR".

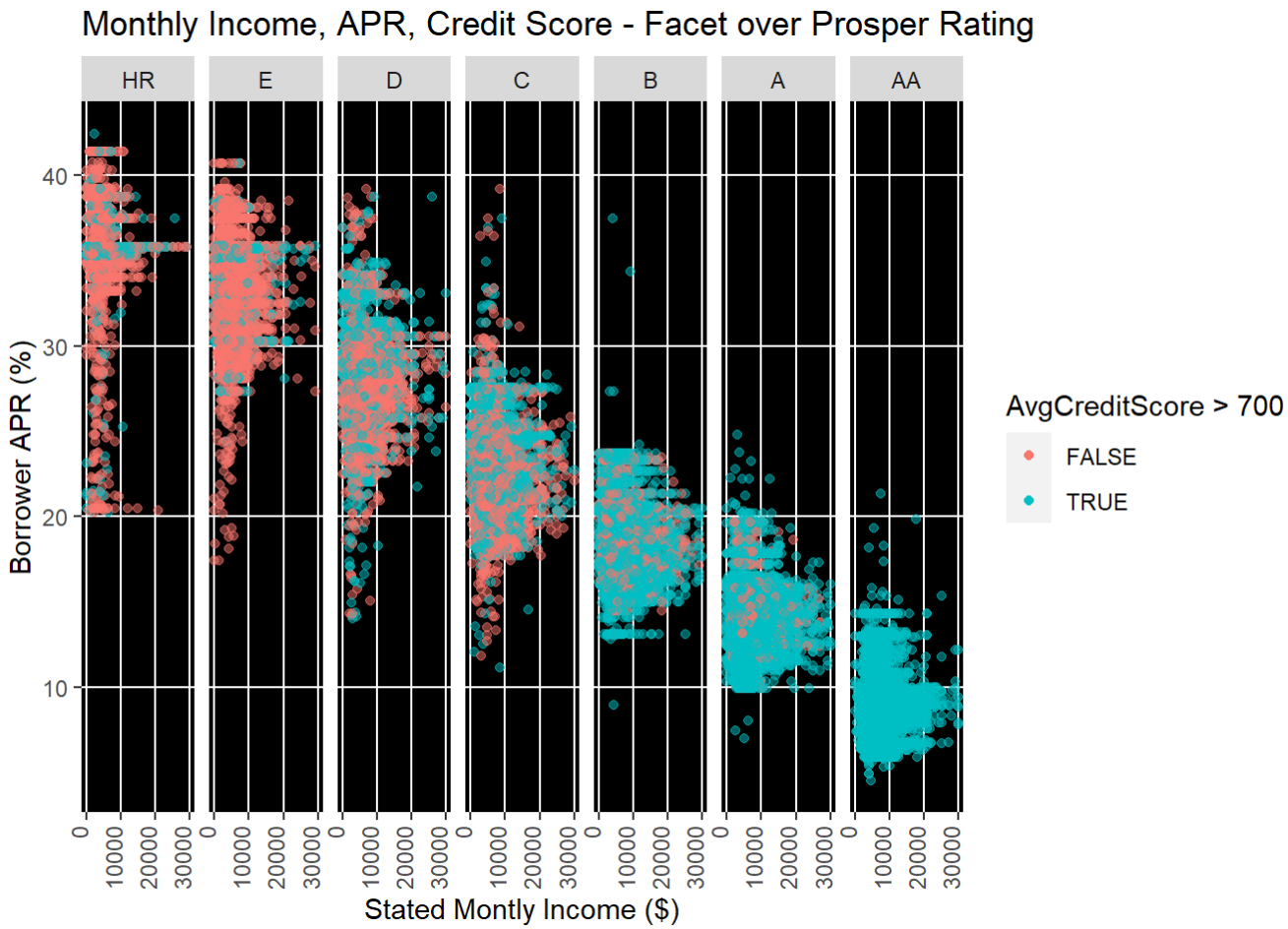
8.2 Plot Two - Monthly Income, APR, Loan Amount - Facet over Prosper Rating



8.2.1 Description Two

I chose geom_Jitter to see scattered relationships between Monthly Income, Loan amount and APR and no doubt Jitter plotted the distribution widely. It looks like if a borrower wants to take a larger loan amount with a lower APR, it helps in having higher monthly income and a good Prosper rating. In result, strong income can help offset a lower Prosper rating like “C” in getting large amount of loans.

8.3 Plot Three - Monthly Income, APR, Credit Score - Facet over Prosper Rating



8.3.1 Description Three

I chose geom_point to visualize scattered relationships between Credit score, Income & APR. Based on the above graph, it's clear that "Credit score does matter". There is a reason why so much emphasis is placed on credit score with regards to loans. A good credit score always plays a major role in getting a good APR.

9 Reflection

Since I work in the Insurance company, exploring Prosper data as part of this project was very interesting for me.I noticed Prosper data mostly focus on personal loan instead of homeowners. Personal Loan data gives good insights from occupation, income and listing category perspective. Most thought provoking and time consuming throughout this project was to come up with good design solutions, to choose the plots that fits best in describing relationships between different variables. I decided to facet with Prosper rating on Bi-variable and Multi-variable analysis. In future, would like to explore the data sets to get some additional insights on the relationships between investors, APR and loan origination date/quarter.