Prosper Data

Alissa McBain

6/2/2018

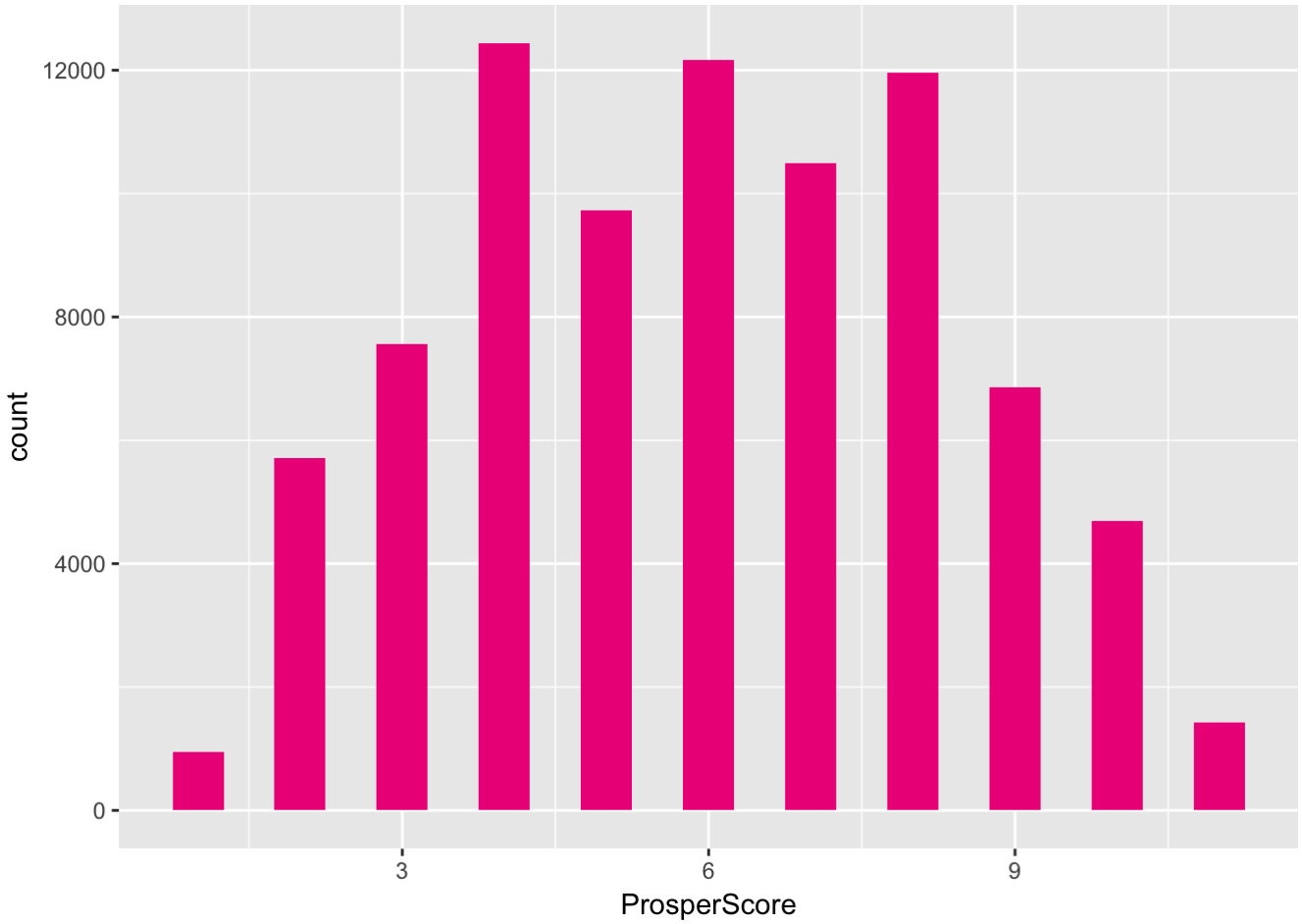
Why Prosper?

I chose to use the Prosper data set. Although it is large, the data is most interesting to me. I have a background in finance, and I love personal finance. I enjoy helping people understand their finances, set budgets and improve their credit scores, since these steps have helped me out in life so so much. I personally am an investor at Prosper, so I thought it would be interesting to try to predict the score Prosper gives a potential loan ('ProsperScore'), or try to predict a loan's APR.

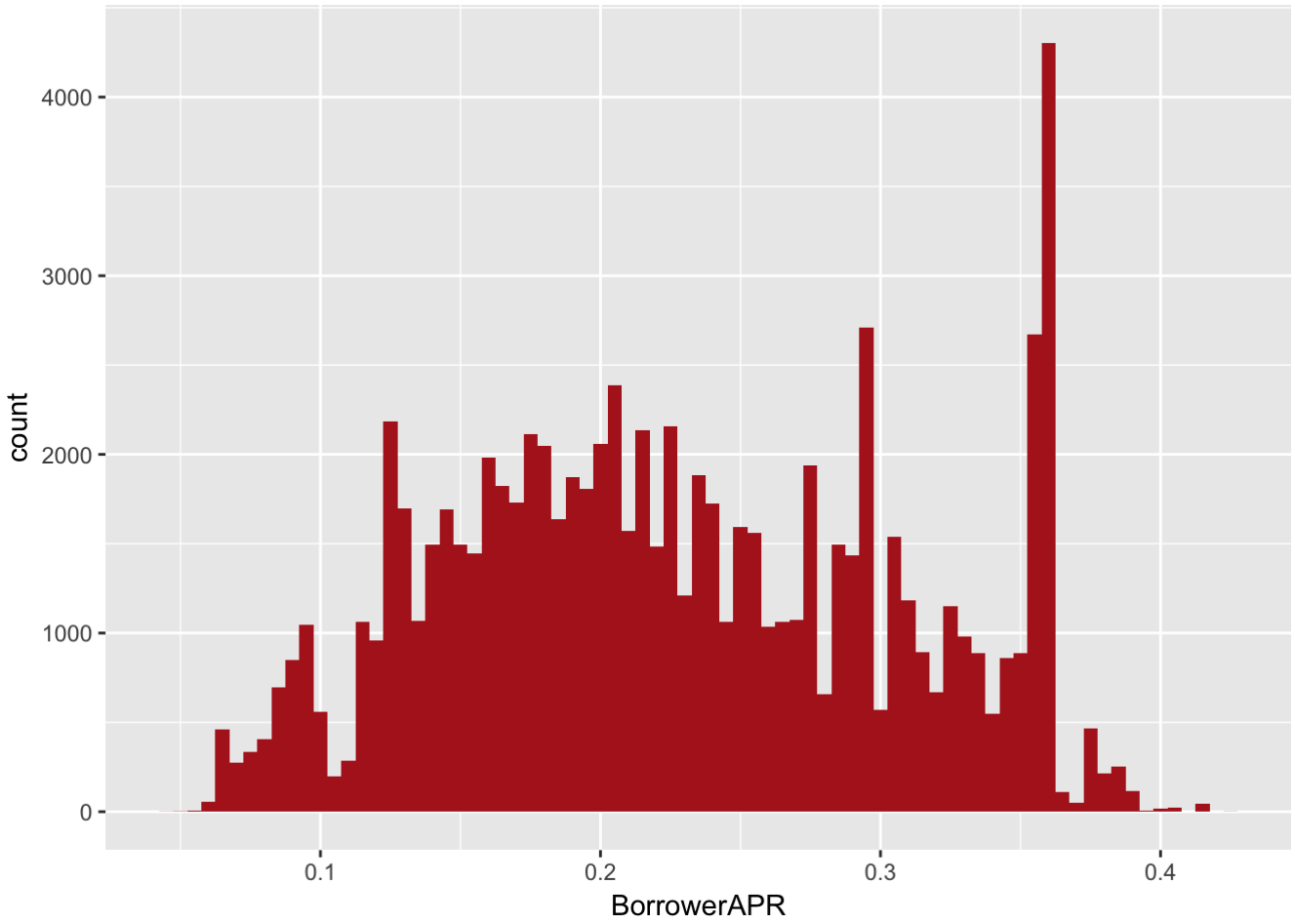
Univariate Plots and Analysis

```
##   BorrowerAPR       ProsperScore         EmploymentStatus
##  Min.   :0.04583   Min.   : 1.000   Employed     :66586
##  1st Qu.:0.16361   1st Qu.: 4.000   Full-time    : 7926
##  Median :0.21945   Median : 6.000   Self-employed: 4456
##  Mean   :0.22694   Mean   : 5.953   Other        : 3742
##  3rd Qu.:0.29254   3rd Qu.: 8.000   Not employed :  649
##  Max.   :0.42395   Max.   :11.000   Retired      :  367
##                                     (Other)      :  256
##  IsBorrowerHomeowner OpenRevolvingMonthlyPayment RevolvingCreditBalance
##  False:39560          Min.   :    0.0            Min.   :     0
##  True :44422          1st Qu.:  155.0            1st Qu.:  3804
##                       Median :  310.0            Median :  9310
##                       Mean   :  430.4            Mean   : 17936
##                       3rd Qu.:  563.0            3rd Qu.: 20335
##                       Max.   :13765.0            Max.   :999165
##
##  BankcardUtilization DebtToIncomeRatio StatedMonthlyIncome
##  Min.   :0.000       Min.   : 0.000    Min.   :      0
##  1st Qu.:0.330       1st Qu.: 0.150    1st Qu.:   3427
##  Median :0.600       Median : 0.220    Median :   5000
##  Mean   :0.564       Mean   : 0.259    Mean   :   5931
##  3rd Qu.:0.830       3rd Qu.: 0.320    3rd Qu.:   7083
##  Max.   :2.500       Max.   :10.010    Max.   :1750003
##                      NA's   :7214
##  LoanOriginalAmount  CreditScore
##  Min.   : 1000      Min.   :609.5
##  1st Qu.: 4000      1st Qu.:669.5
##  Median : 7500      Median :709.5
##  Mean   : 9061      Mean   :709.0
##  3rd Qu.:13500      3rd Qu.:729.5
##  Max.   :35000      Max.   :889.5
##
```

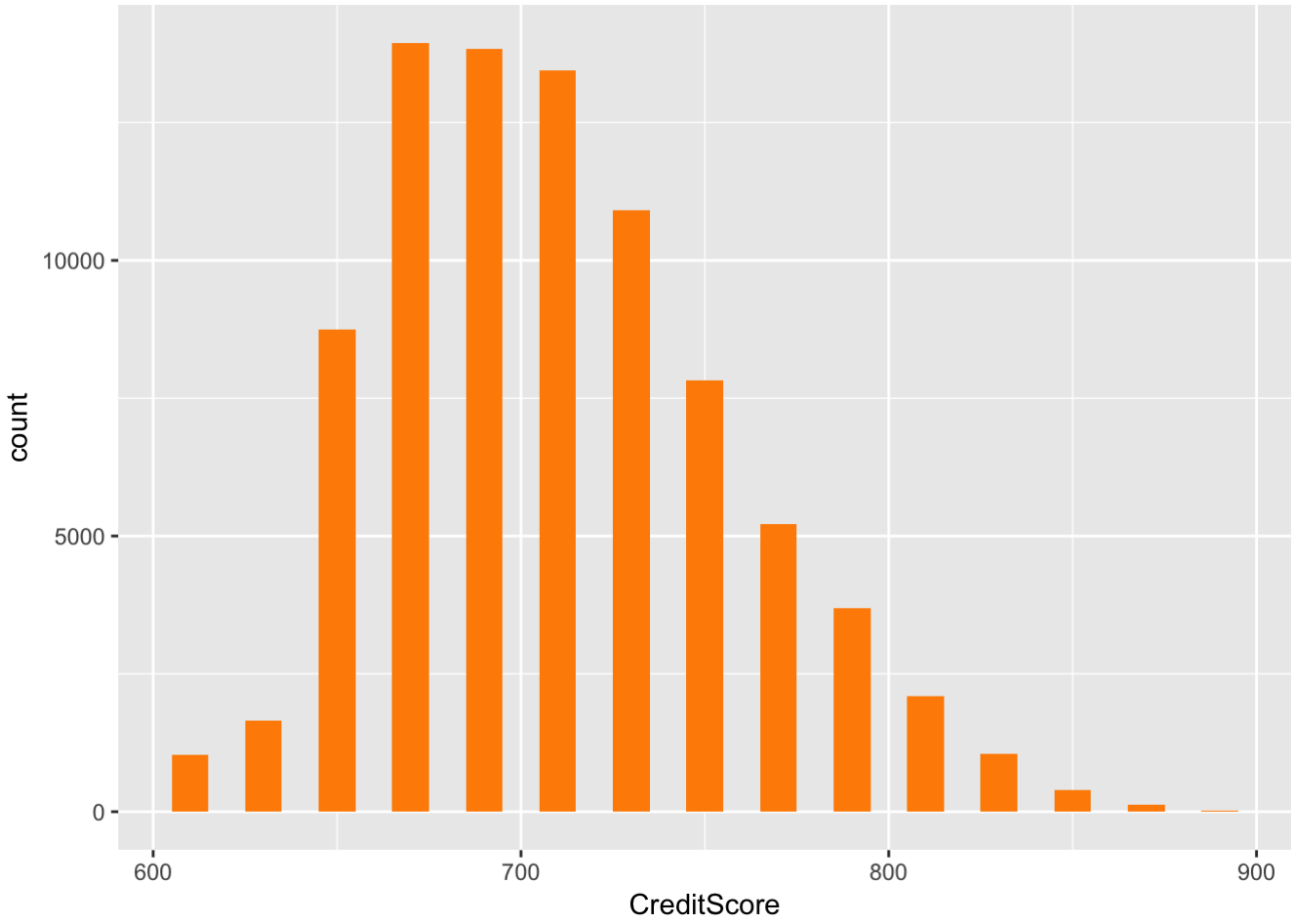
My data set consists of eleven variables, two of which I will be trying to predict (ProsperScore or BorrowerAPR). The other nine variables are facts about a customer's credit and lifestyle. One variable, EmploymentStatus, is factored with nine levels. Another variable, IsBorrowerHomeowner is also factored, but with only a True or False value. The remaining seven variables are all continuous variables. Five of these variables (OpenRevolvingMonthlyPayment, RevolvingCreditBalance, BankcardUtilization, DebtToIncomeRatio and StatedMonthlyIncome) have very large outliers, so I'm going to keep that in mind as I move on.



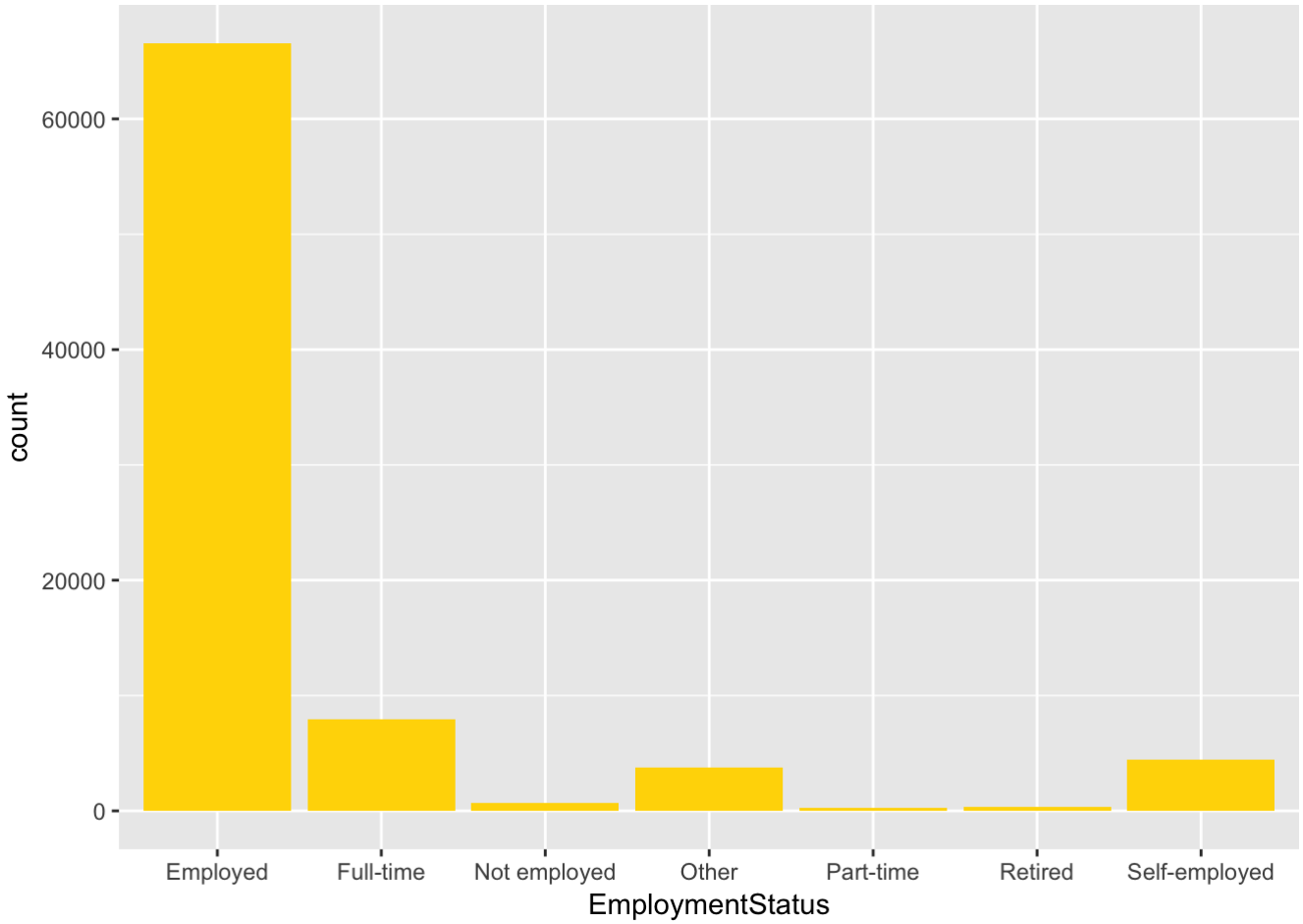
This histogram of ProsperScore shows that it is fairly normal distributed, with discrete values.



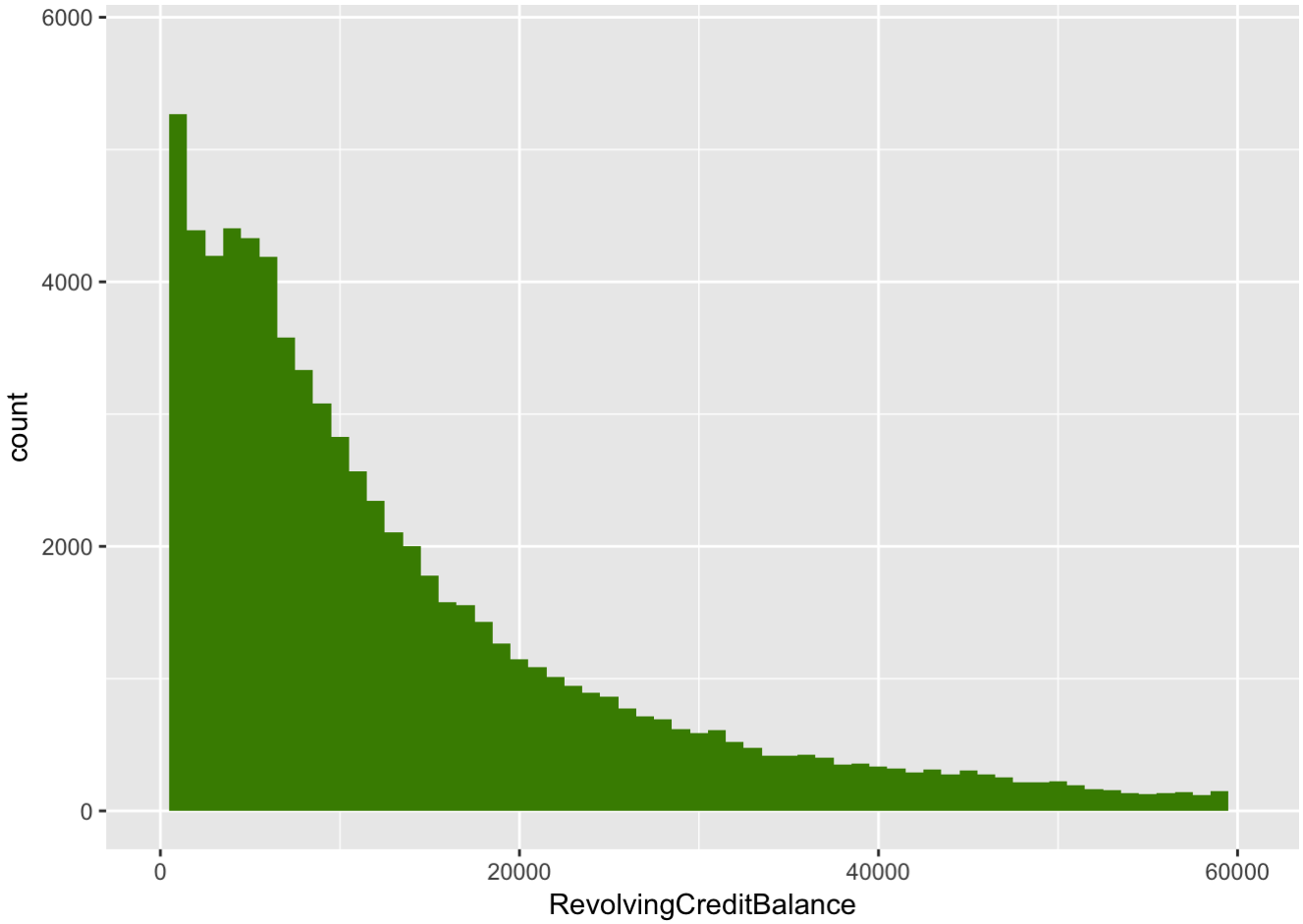
The histogram for BorrowerAPR is also fairly normally distributed, with a large spike around 0.36/0.37, so that is a bit curious. There is also a spike at 0.29.



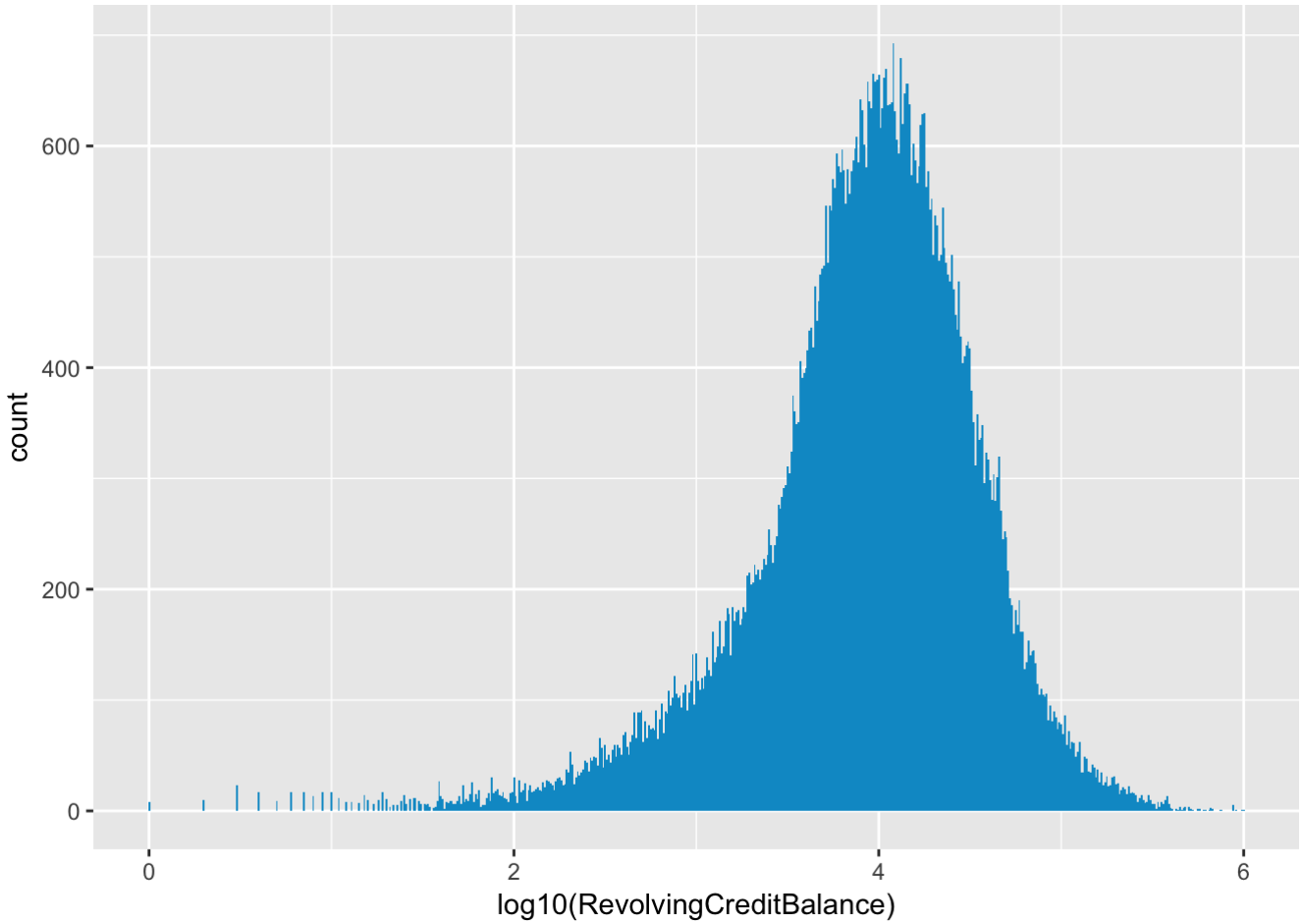
This histogram of CreditScore is normal, with a right skewness, indicating that there may be outliers on the larger side of CreditScore that are pulling the mean. Most CreditScores are between 650 and 750.



Most borrowers in the data set are Employed.

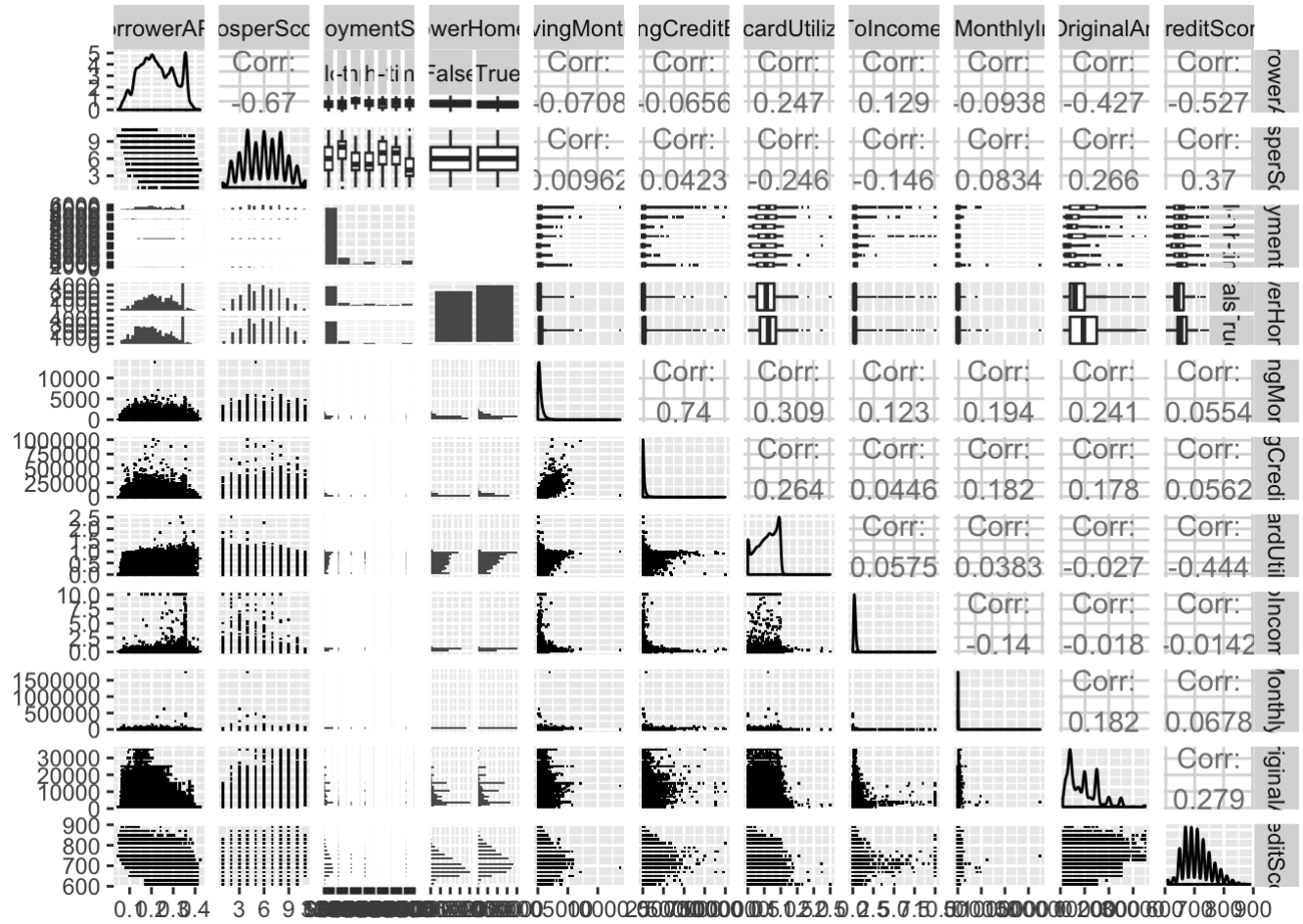


RevolvingCreditBalance was one of the variables with larger outliers that I noticed from the Summary table (looking at the 3rd Quartile value vs. the Max, they are very spread apart). Limiting the x axis to the 95th quantile cuts off many of those large outliers, bringing the max down from 999,165 to about 60,000. Much cleaner and easier to read.



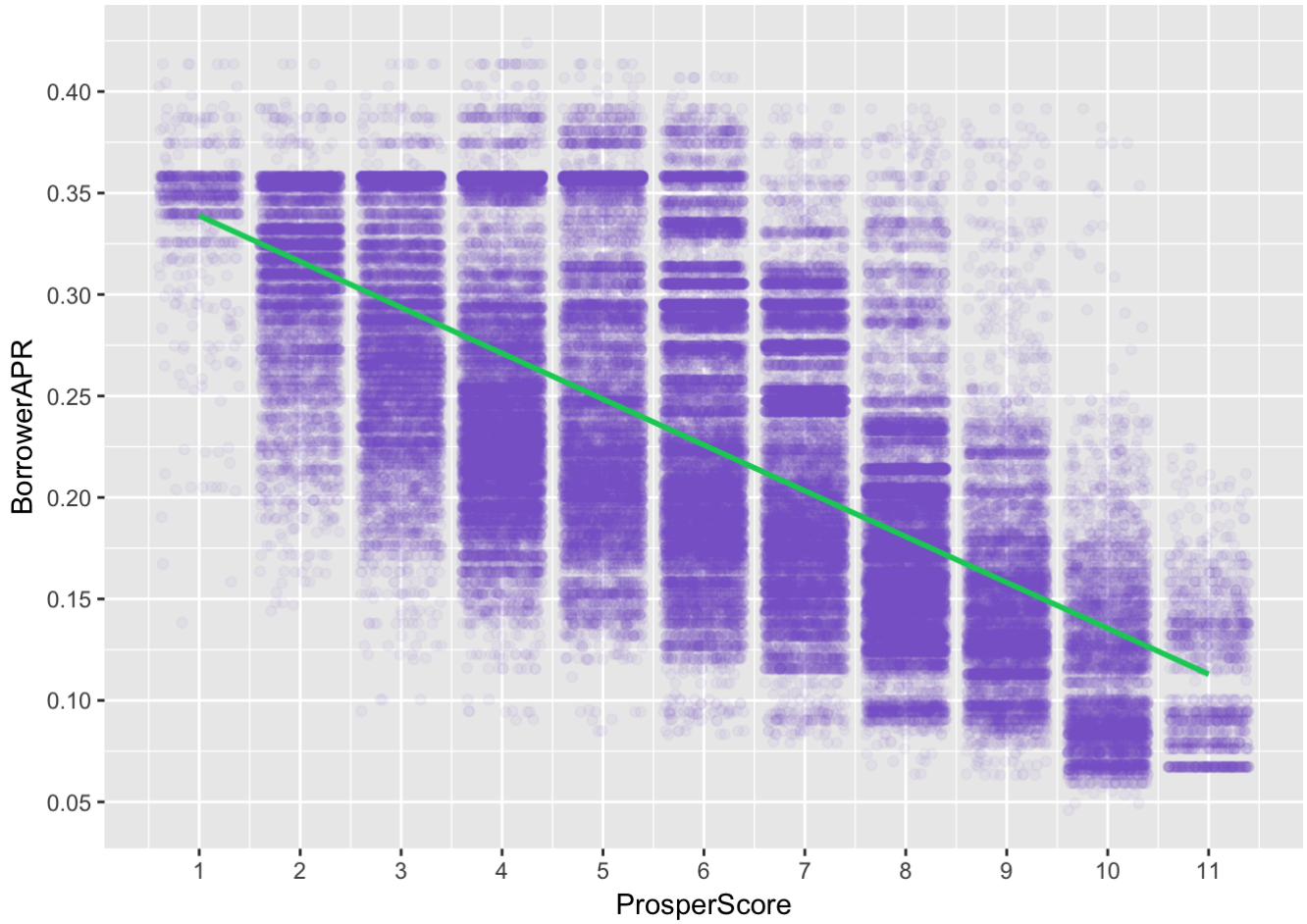
Also, taking the log10 of RevolvingCreditBalance made the data much more normally distributed (slightly left-skewed), so I'm going to keep that in mind for future analysis as well.

Bivariate Plots and Analysis

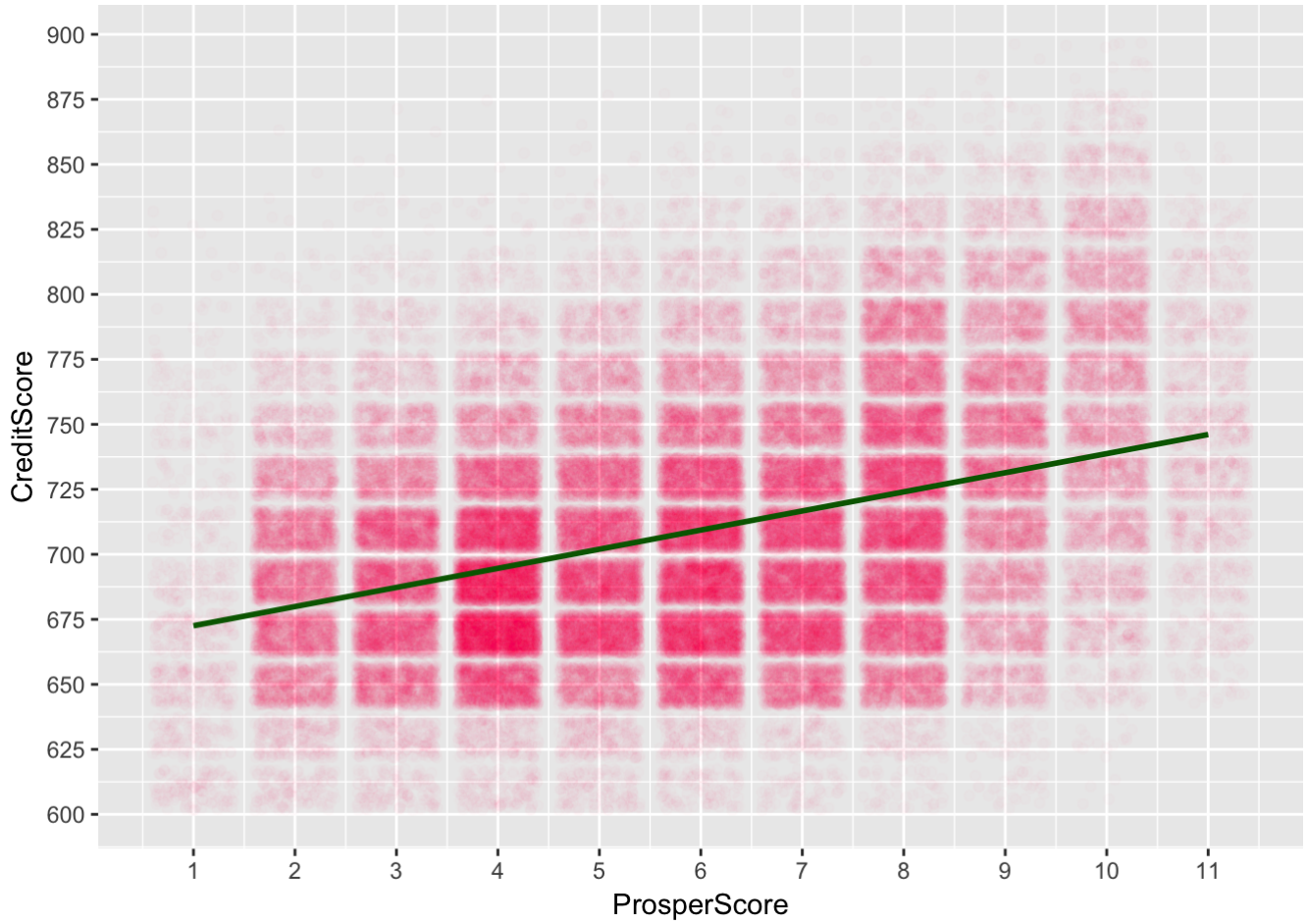


Observing this GGPairs plots, and keeping in mind I'd like to predict ProsperScore or BorrowerAPR, what do I want to explore more? Well, looking at the correlation coefficients, the variable most correlated with ProsperScore is BorrowerAPR. This makes sense because both of these variables are the outcome of the analysis that Prosper does on an individual. Next most correlated with ProsperScore is CreditScore.

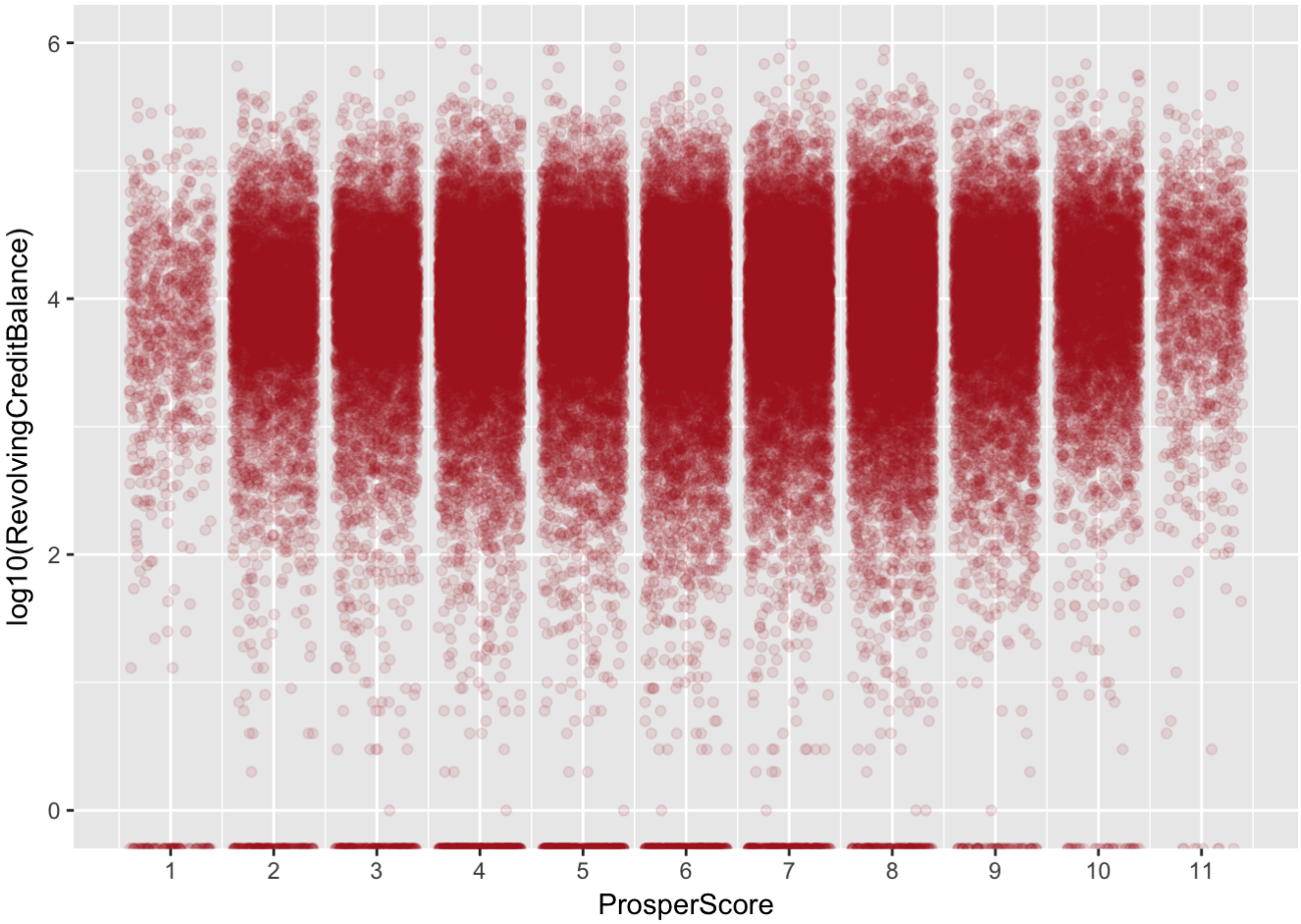
The strongest correlation of all my variables pairs RevolvingCreditBalance and OpenRevolvingMonthlyPayment. This is understandable because the higher the credit balance is, the higher the monthly payment should be (perhaps a lower monthly payment for someone with a higher CreditScore? <- something to explore).



The first relationship I looked at was ProsperScore vs BorrowerAPR. They were indeed correlated, although not as strongly as I would have assumed. There is quite a lot of data at the 0.36 APR line (36%), which pulls the line of best fit higher than I would expect.

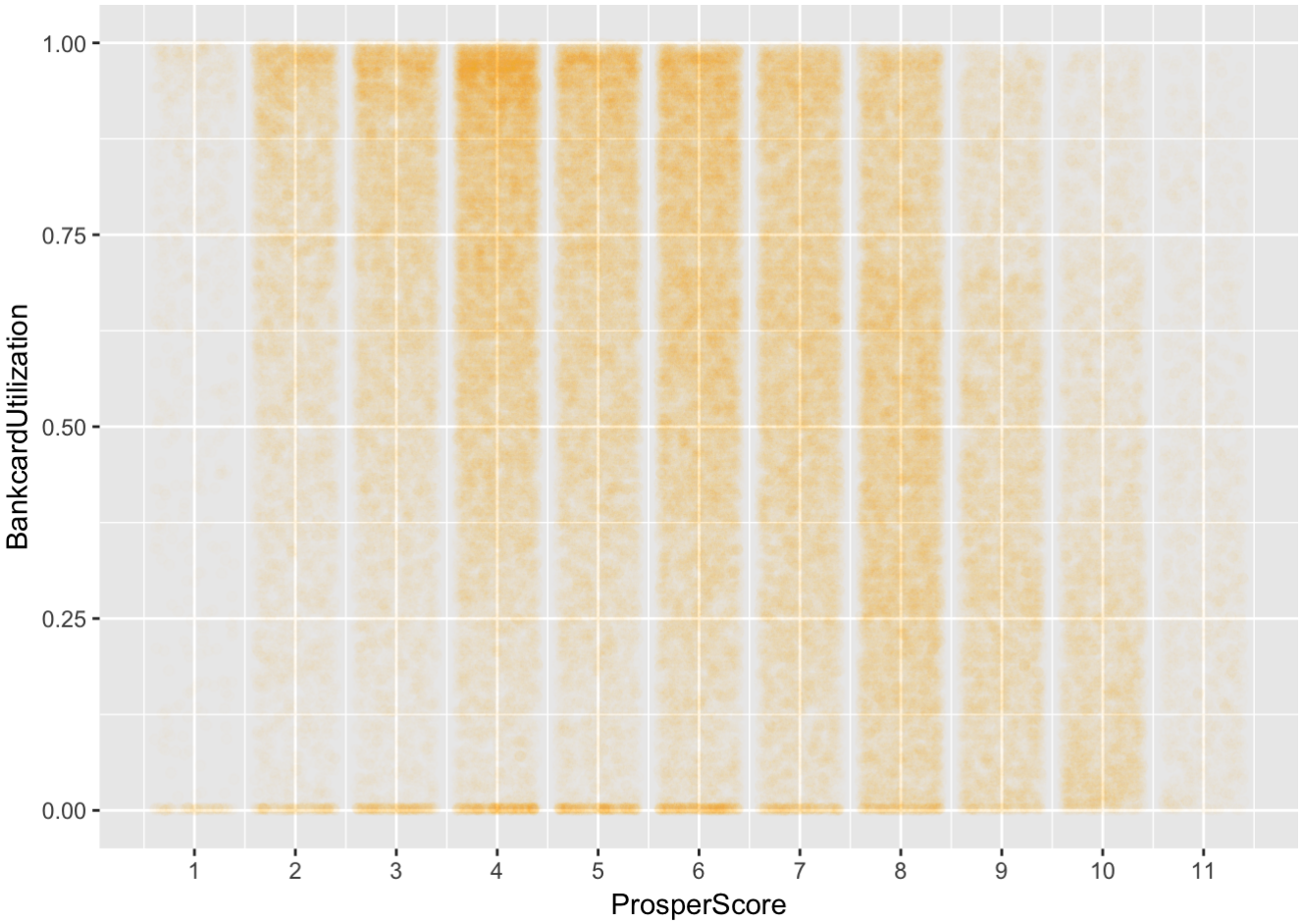


Second, I looked at ProsperScore vs CreditScore. Again, these two variables were correlated, and when I get to making a model to predict ProsperScore, I'm sure CreditScore will be included. The variance of the data was larger than I was expecting, as I can many borrowers with credit scores in the range of 640 to 800 all received ProsperScores of 8, for example.



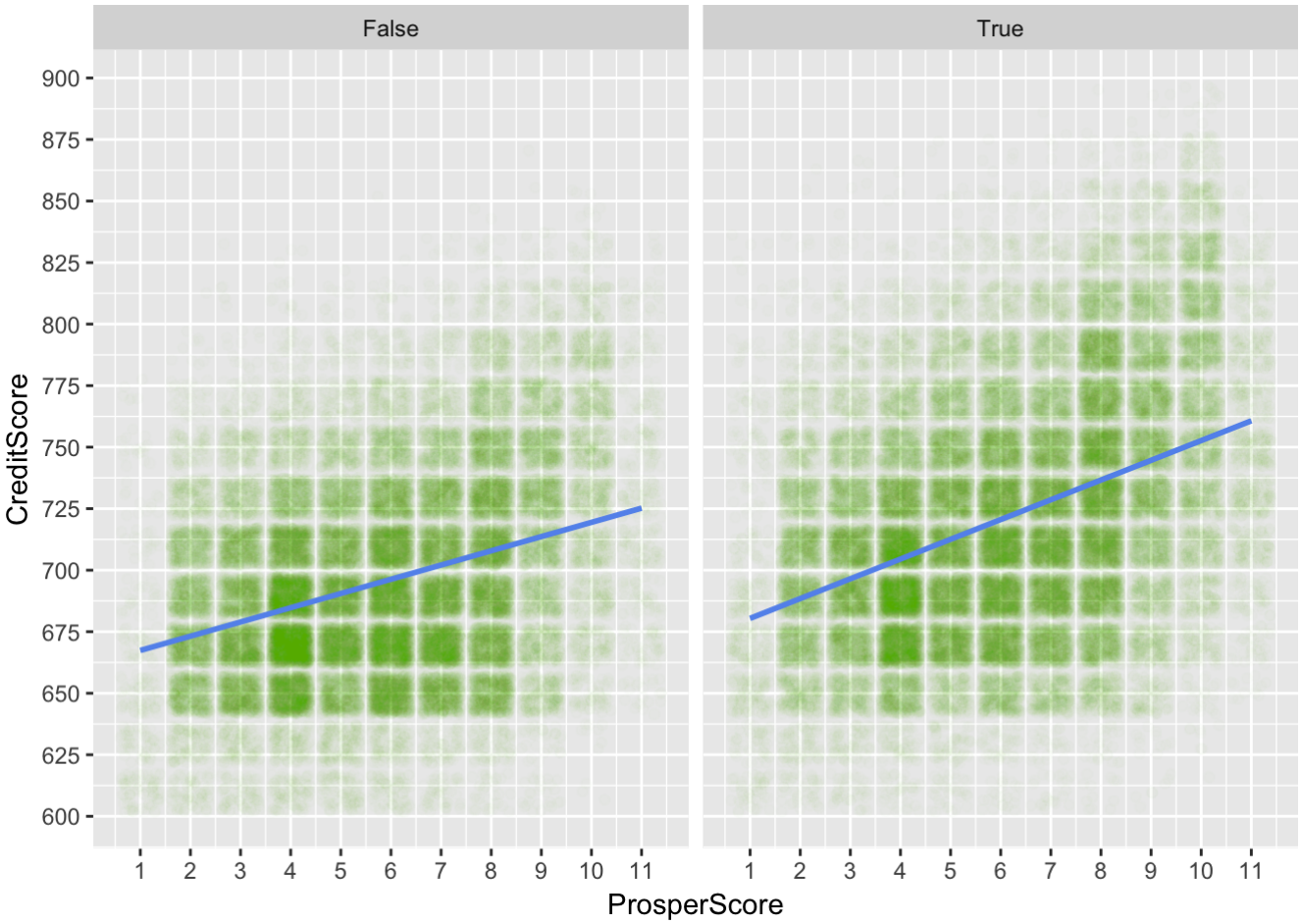
Next I looked at ProsperScore vs log(RevolvingCreditBalance), and didn't find much correlation. Unfortunately this particular analysis didn't add any information to my investigation.

The two variables with the highest correlation on my GGPairs plots was RevolvingCreditBalance vs OpenRevolvingMonthlyPayment. Both of these variables have large outliers, so I limited the plot to only the 95th percentile for each variable. The plot shows a strong positive correlation, which makes sense when I think about what each variable represents. The higher a borrower's credit balance is, the higher (I would assume) their monthly payments on that balance would be.

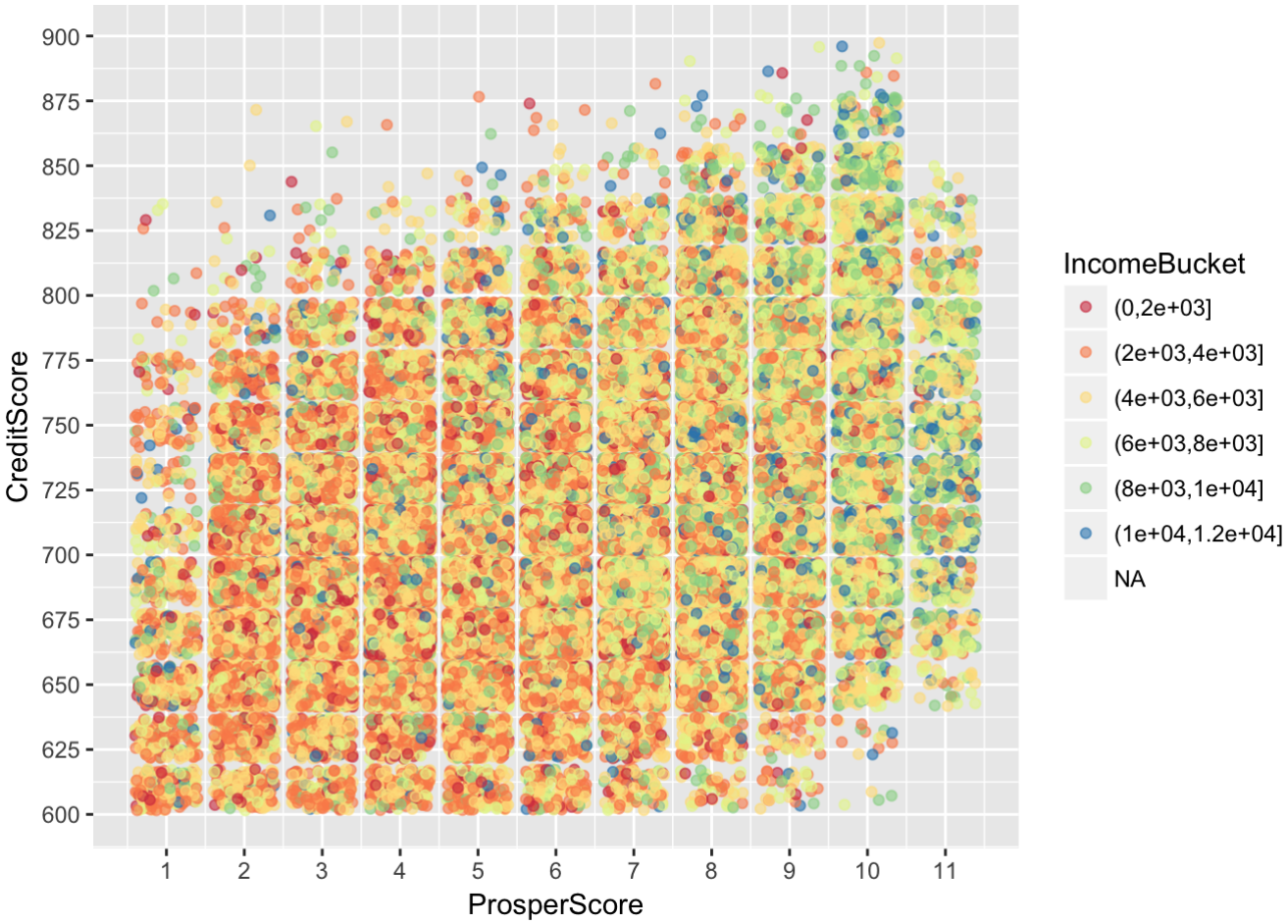


Lastly, expects say that 30% of a person's credit score is determined by the amount of outstanding debt that they currently have, so I was curious to see if BankcardUtilization has an impact on a person's ProsperScore. I limited y to 1 because I'm not sure that it's possible to have over 100% utilization, and if so, those would be outliers anyways. Unfortunately, I didn't see a strong correlation, but there was a faint negative trend to the data. As a borrower's utilization went up, their ProsperScore went down, in general.

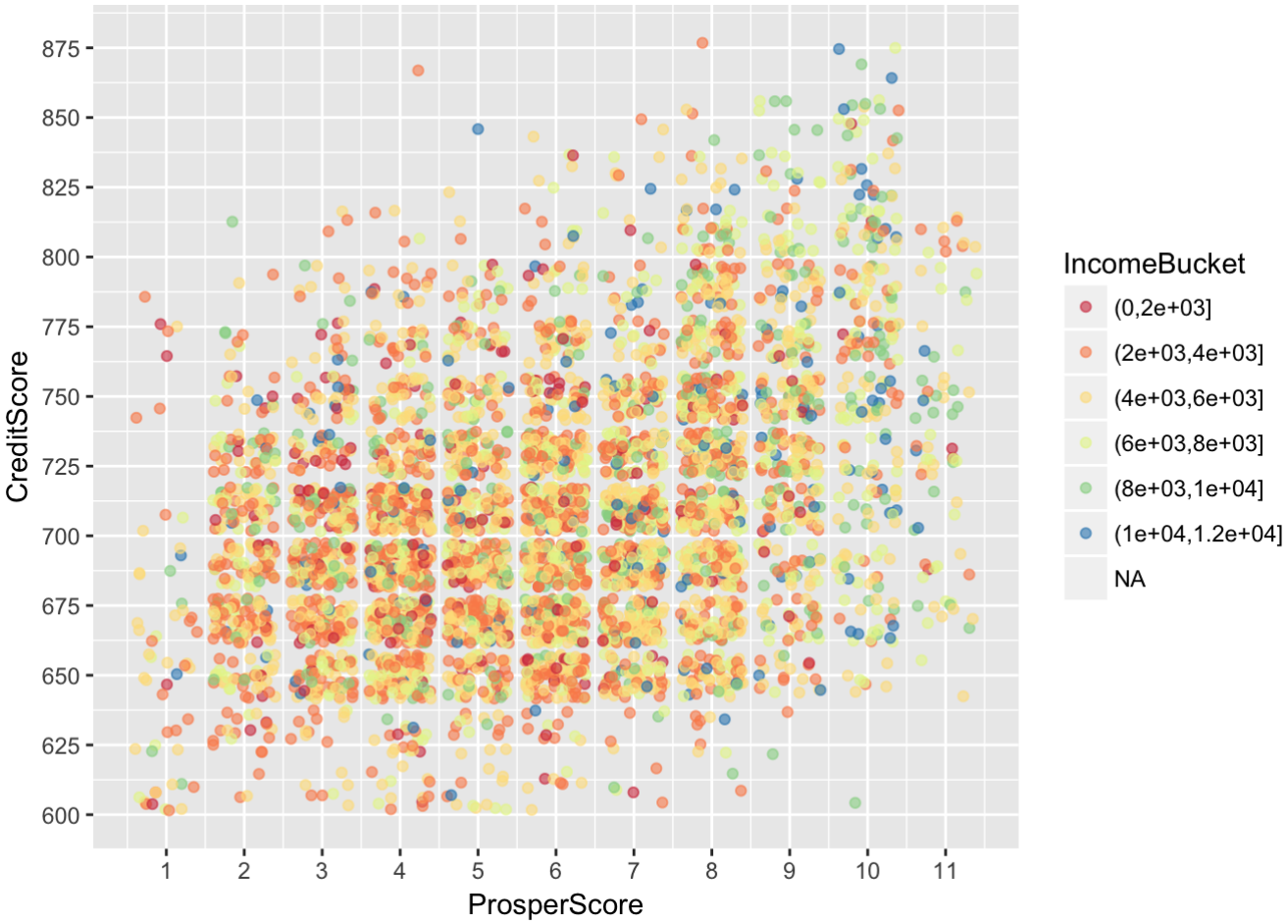
Multivariate Plots and Analysis



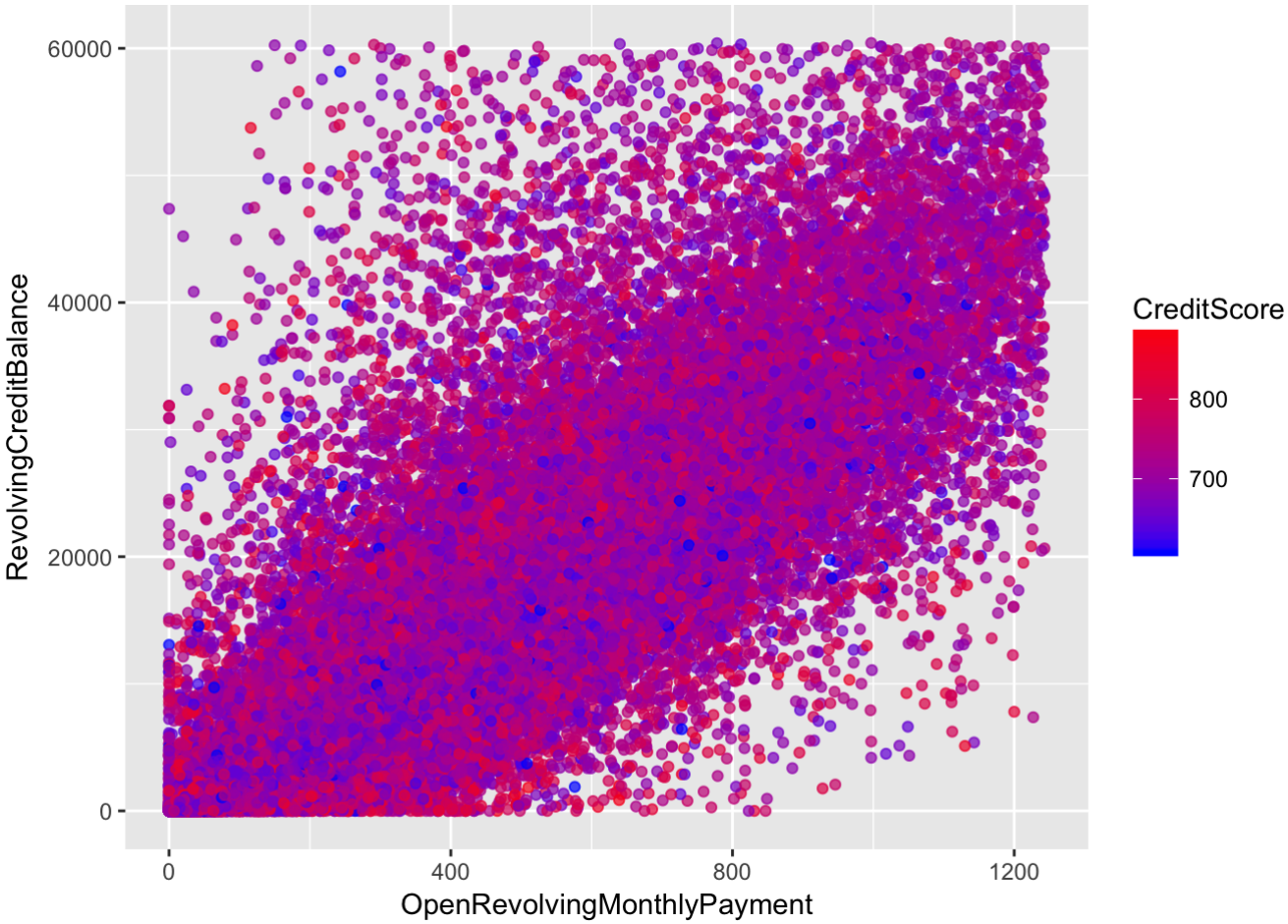
The first thing I looked at was ProsperScore vs CreditScore, faceted by whether or not the borrower was a homeowner. I noticed that the line of best fit was slightly more steep for homeowners, meaning that the correlation between ProsperScore and CreditScore was slightly stronger for them. Also, I noticed that, in general, homeowners had a wider range of ProsperScores, filling up more of the higher ProsperScores than non-homeowners.



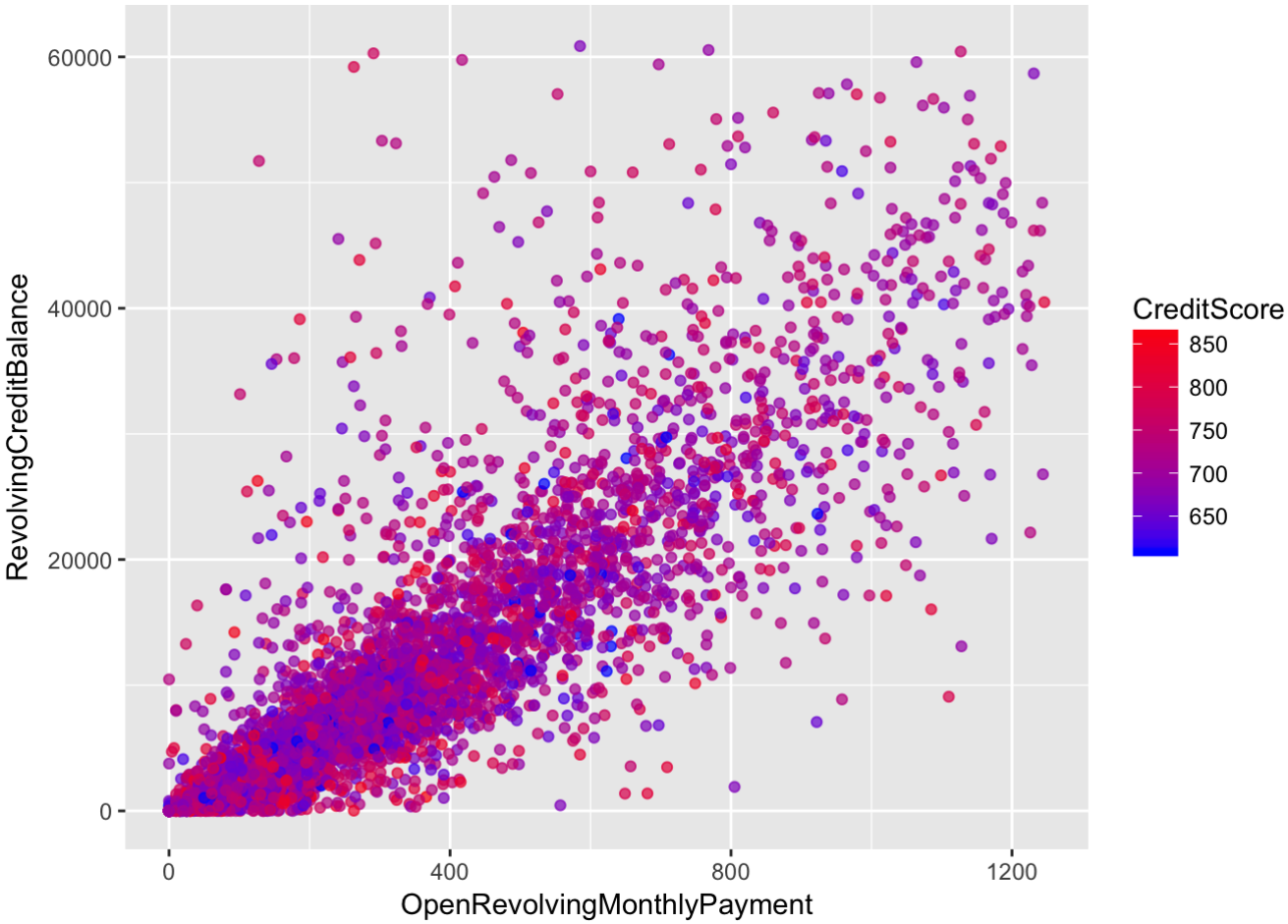
Next, this plot is a funfetti plot, if I've ever seen one. Again, I looked at ProsperScore vs CreditScore, this time coloring the points by a borrower's income range (which I bucketed, to make the plot easier to interpret). There is a clear correlation with income range; I can detect a slight correlation of higher incomes with higher Prosper and Credit Scores, and lower incomes with lower Prosper and Credit Scores (the plot is more green/blue in the top right, and slightly more red/orange in the bottom left). But, this plot is very overplotted, even using 60% transparency, so I'd like to take a look at just a sample of the data.



To get a more clear picture, I took a random sample of 5,000 data points and created a new data frame from those. Then I replotted the same graph with my random sample, to see if I could gather any trends from this smaller data set. It seems about the same, with a lower income slightly correlated to a lower Credit and Prosper Score, and vice verse.



Lastly, since my strongest correlation was between OpenRevolvingMonthlyPayment and RevolvingCreditBalance, I wanted to see if CreditScore was a factor in the correlation. I plotted them, with CreditScore going from blue to red, and as you can see, the plot is VERY overplotted and very hard to interpret because of this.



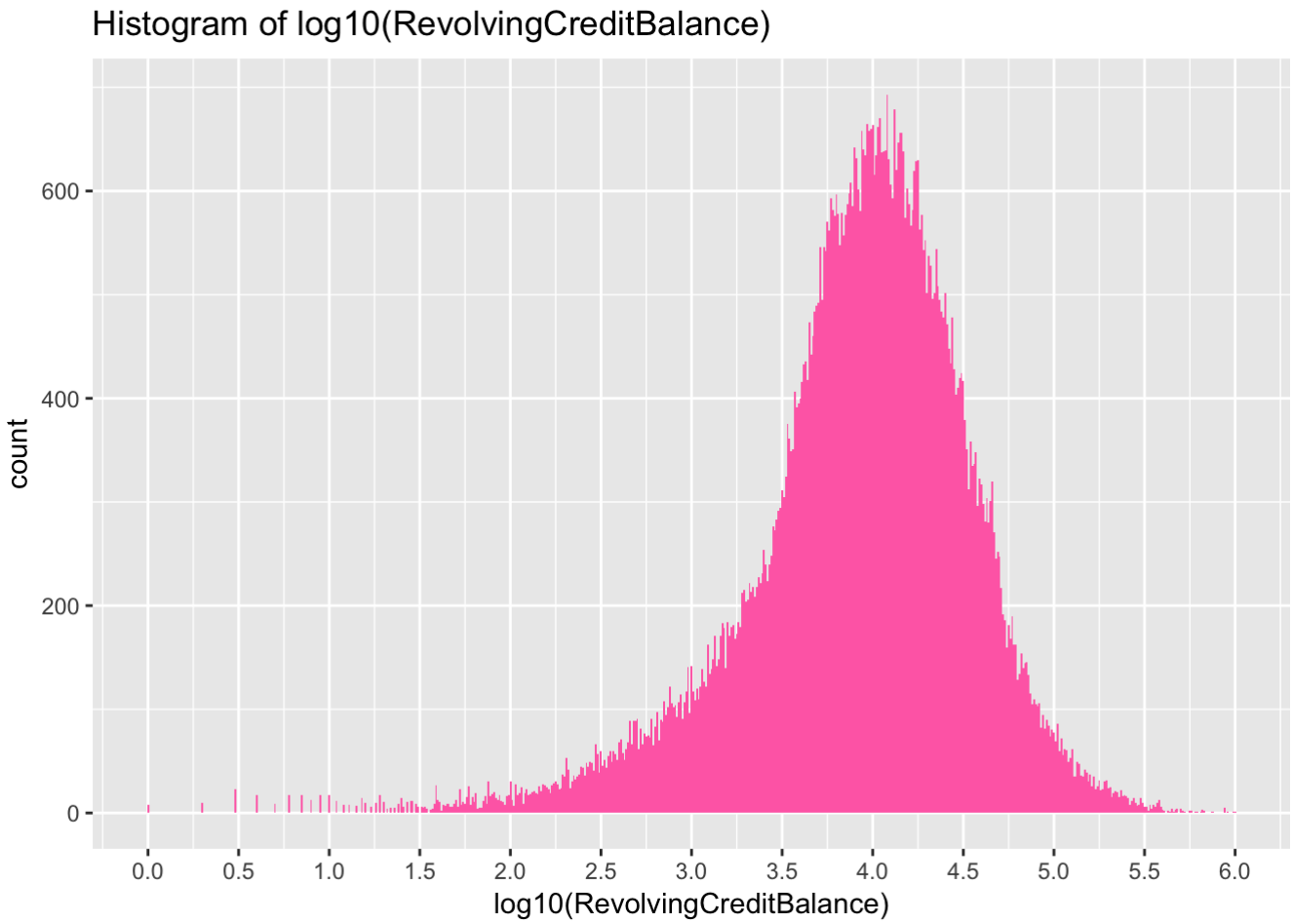
So, I used the sample data frame I made earlier and plotted that instead. This plot is much easier to read, and I can see that the correlation between OpenRevolvingMonthlyPayment and RevolvingCreditBalance is much stronger with lower values, and becomes much more dispersed with higher values of each variables. Unfortunately, there is no discernible pattern to CreditScore on this data.

```
##
## Calls:
## m1: lm(formula = ProsperScore ~ CreditScore, data = dfsubset)
## m2: lm(formula = ProsperScore ~ CreditScore + log10(DebtToIncomeRatio +
##     0.01), data = dfsubset)
## m3: lm(formula = ProsperScore ~ CreditScore + log10(DebtToIncomeRatio +
##     0.01) + IsBorrowerHomeowner, data = dfsubset)
## m4: lm(formula = ProsperScore ~ CreditScore + log10(DebtToIncomeRatio +
##     0.01) + IsBorrowerHomeowner + EmploymentStatus, data = dfsubset)
## m5: lm(formula = ProsperScore ~ CreditScore + log10(DebtToIncomeRatio +
##     0.01) + IsBorrowerHomeowner + EmploymentStatus + log10(BankcardUtilization +
##     0.01), data = dfsubset)
##
## ============================================================================================
=======================================
##                                                    m1              m2              m3
            m4              m5
## --------------------------------------------------------------------------------------------
---------------------------------------
##   (Intercept)                                    -7.246***       -8.816***       -9.0
95***       -9.165***      -10.006***
##                                                  (0.115)         (0.114)         (0.1
16)         (0.115)         (0.122)
##   CreditScore                                     0.019***        0.019***        0.0
19***        0.019***        0.021***
##                                                  (0.000)         (0.000)         (0.0
00)         (0.000)         (0.000)
##   log10(DebtToIncomeRatio + 0.01)                                -2.566***       -2.5
52***       -2.476***       -2.627***
##                                                                  (0.031)         (0.0
31)         (0.030)         (0.031)
##   IsBorrowerHomeowner: True/False                                                -0.1
83***       -0.200***       -0.263***
##                                                                                  (0.0
16)         (0.016)         (0.016)
##   EmploymentStatus: Full-time/Employed
             0.996***        1.025***
##
            (0.025)         (0.025)
##   EmploymentStatus: Not employed/Employed
             3.719           3.742
##
            (2.053)         (2.048)
##   EmploymentStatus: Other/Employed
            -0.500***       -0.479***
##
            (0.036)         (0.036)
##   EmploymentStatus: Part-time/Employed
             1.316***        1.404***
##
            (0.146)         (0.146)
##   EmploymentStatus: Retired/Employed
             0.494***        0.563***
```

```
##
            (0.115)          (0.115)
##   EmploymentStatus: Self-employed/Employed
            -1.839***        -1.841***
##
            (0.317)          (0.316)
##   log10(BankcardUtilization + 0.01)
                              0.354***
##
                             (0.017)
## ------------------------------------------------------------------------------------
-------------------------------------
##   R-squared                                          0.137           0.222           0.2
23           0.243           0.247
##   adj. R-squared                                     0.137           0.222           0.2
23           0.243           0.247
##   sigma                                              2.205           2.082           2.0
80           2.053           2.048
##   F                                                13324.565       10927.496        7342.9
91        2733.551        2514.873
##   p                                                  0.000           0.000           0.0
00           0.000           0.000
##   Log-likelihood                                  -185586.103     -165213.415     -165145.6
47     -164158.035     -163951.410
##   Deviance                                         408458.057      332665.313      332078.5
07      323643.182      321905.658
##   AIC                                              371178.206      330434.829      330301.2
94      328338.070      327926.820
##   BIC                                              371206.221      330471.823      330347.5
37      328439.804      328037.802
##   N                                                 83982           76768           76768
         76768           76768
## ====================================================================================
=====================================
```

The last step I took was to build a model with the five mostly likely predictors of ProsperScore. Even with these five variables, the R squared value is only 0.247, so I would not feel confident saying this is a strong model to predict a borrower's ProsperScore.

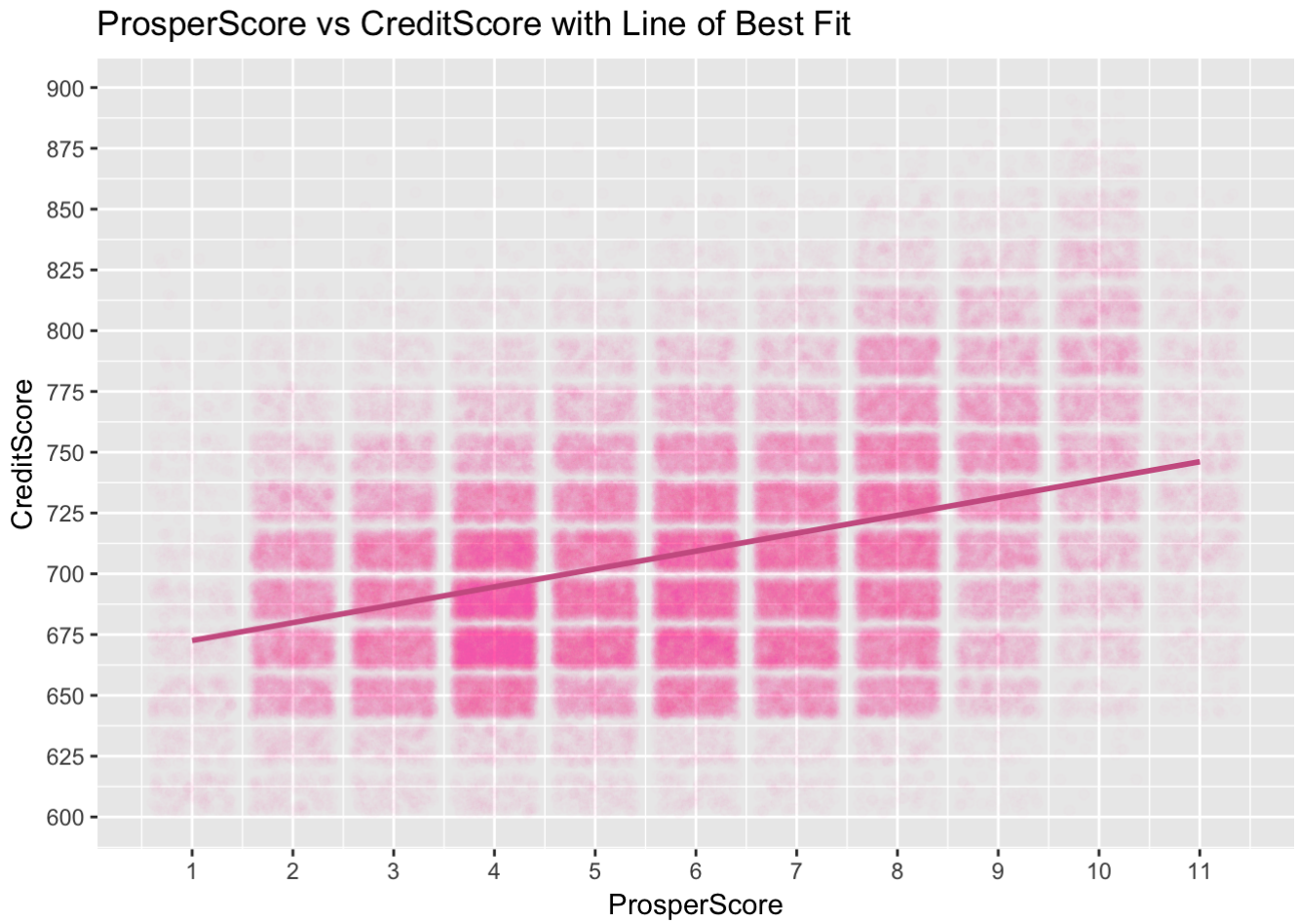
Plot One



Description One

I chose this plot as my first graphic because I found several variables that had large outliers (including RevolvingCreditBalance) and those outliers pulled the data and made them very right skewed. By taking the log 10 of these variables, it made their distributions much easier to work with.

Plot Two



Description Two

I chose this plot because I used this relationship as a basis for predicting ProsperScore. Though the relationship is not perfectly linear, CreditScore was the variable with the highest correlation to ProsperScore.

Plot Three

Description Three

I chose this plot as my third most interesting plot. I changed the colors from a funfetti, to a sliding scale from lavender to red. Though there isn't a STRONG correlation between ProsperScore and CreditScore, with this plot I can definitely tell there is a correlation between those two variables and Income. The top right of the plot is much more red, and the bottom left of the plot is much more lavender.

Reflection

In this exploration, I didn't find the strong relationships that I was hoping to find, but was pleasantly surprised to find some correlations. I struggled in the beginning, narrowing down this large data set; deciding what to keep and what to toss was difficult. Perhaps in the future, a better model could be made to predict ProsperScore with more of the variables that I discarded. Of course, Prosper has their algorithm, which uses some, if not all, of these variables.