Exploratory Data Analysis of Prosper Loan Distribution

Kristen Kellerman

==========

This analysis takes an exploratory look at loan data from Prosper.com between 2005 and 2014. The dataset includes more than 110,000 loan entries with 81 different variables per entry. We will explore 17 of those variables to discover the who(loan requstors), what(loan amount), why (loan type) and how (criteria for approval) of Prosper clients.

The loan variables that we will explore are:

CreditScoreRangeLower

CreditScoreRangeUpper

CurrentCreditLines

CurrentDelinquencies

DebtToIncomeRatio

EmploymentStatus

EstimatedReturn

IncomeRange

ListingCategory

LoanOriginalAmount

Occupation

OpenCreditLines

IsBorrowerHomeOwner

EstimatedLoss

PublicRecordsLast12Months

EffectiveYield

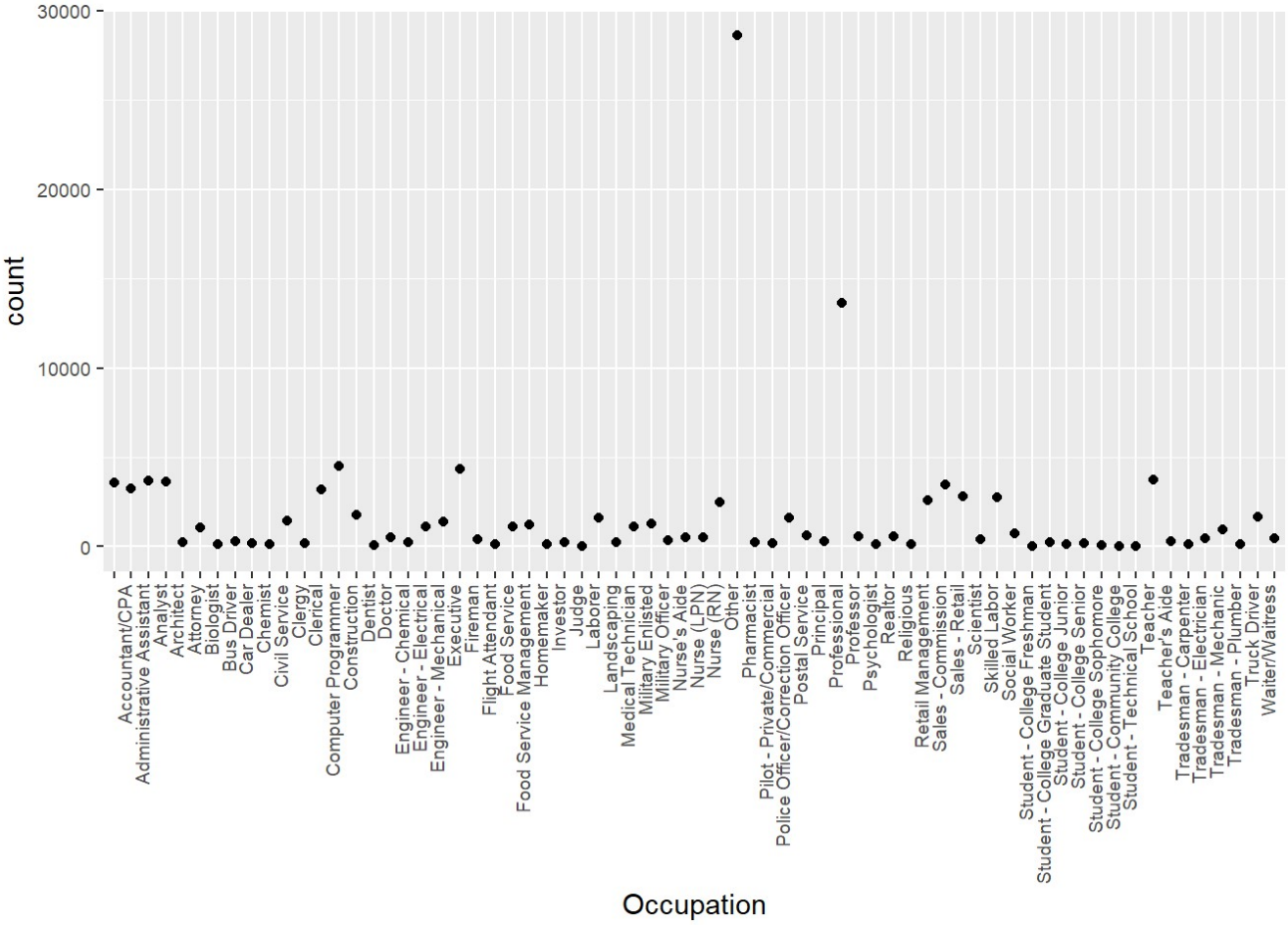
RevolvingCreditBalance

```
##  [1] "CreditScoreRangeLower"     "CreditScoreRangeUpper"
##  [3] "CurrentCreditLines"        "CurrentDelinquencies"
##  [5] "DebtToIncomeRatio"         "EmploymentStatus"
##  [7] "EmploymentStatusDuration"  "EstimatedReturn"
##  [9] "IncomeRange"               "ListingCategory..numeric."
## [11] "LoanOriginalAmount"        "Occupation"
## [13] "OpenCreditLines"           "IsBorrowerHomeowner"
## [15] "EstimatedLoss"             "LenderYield"
## [17] "EstimatedEffectiveYield"
```

```
## 'data.frame':    113937 obs. of  17 variables:
##  $ CreditScoreRangeLower    : int  640 680 480 800 680 740 680 700 820 820 ...
##  $ CreditScoreRangeUpper    : int  659 699 499 819 699 759 699 719 839 839 ...
##  $ CurrentCreditLines       : int  5 14 NA 5 19 21 10 6 17 17 ...
##  $ CurrentDelinquencies     : int  2 0 1 4 0 0 0 0 0 0 ...
##  $ DebtToIncomeRatio        : num  0.17 0.18 0.06 0.15 0.26 0.36 0.27 0.24 0.25 0.2
5 ...
##  $ EmploymentStatus         : chr  "Self-employed" "Employed" "Not available" "Empl
oyed" ...
##  $ EmploymentStatusDuration : int  2 44 NA 113 44 82 172 103 269 269 ...
##  $ EstimatedReturn          : num  NA 0.0547 NA 0.06 0.0907 ...
##  $ IncomeRange              : chr  "$25,000-49,999" "$50,000-74,999" "Not displayed
" "$25,000-49,999" ...
##  $ ListingCategory..numeric.: int  0 2 0 16 2 1 1 2 7 7 ...
##  $ LoanOriginalAmount       : int  9425 10000 3001 10000 15000 15000 3000 10000 100
00 10000 ...
##  $ Occupation               : chr  "Other" "Professional" "Other" "Skilled Labor" .
..
##  $ OpenCreditLines          : int  4 14 NA 5 19 17 7 6 16 16 ...
##  $ IsBorrowerHomeowner      : chr  "True" "False" "False" "True" ...
##  $ EstimatedLoss            : num  NA 0.0249 NA 0.0249 0.0925 ...
##  $ LenderYield              : num  0.138 0.082 0.24 0.0874 0.1985 ...
##  $ EstimatedEffectiveYield  : num  NA 0.0796 NA 0.0849 0.1832 ...
```

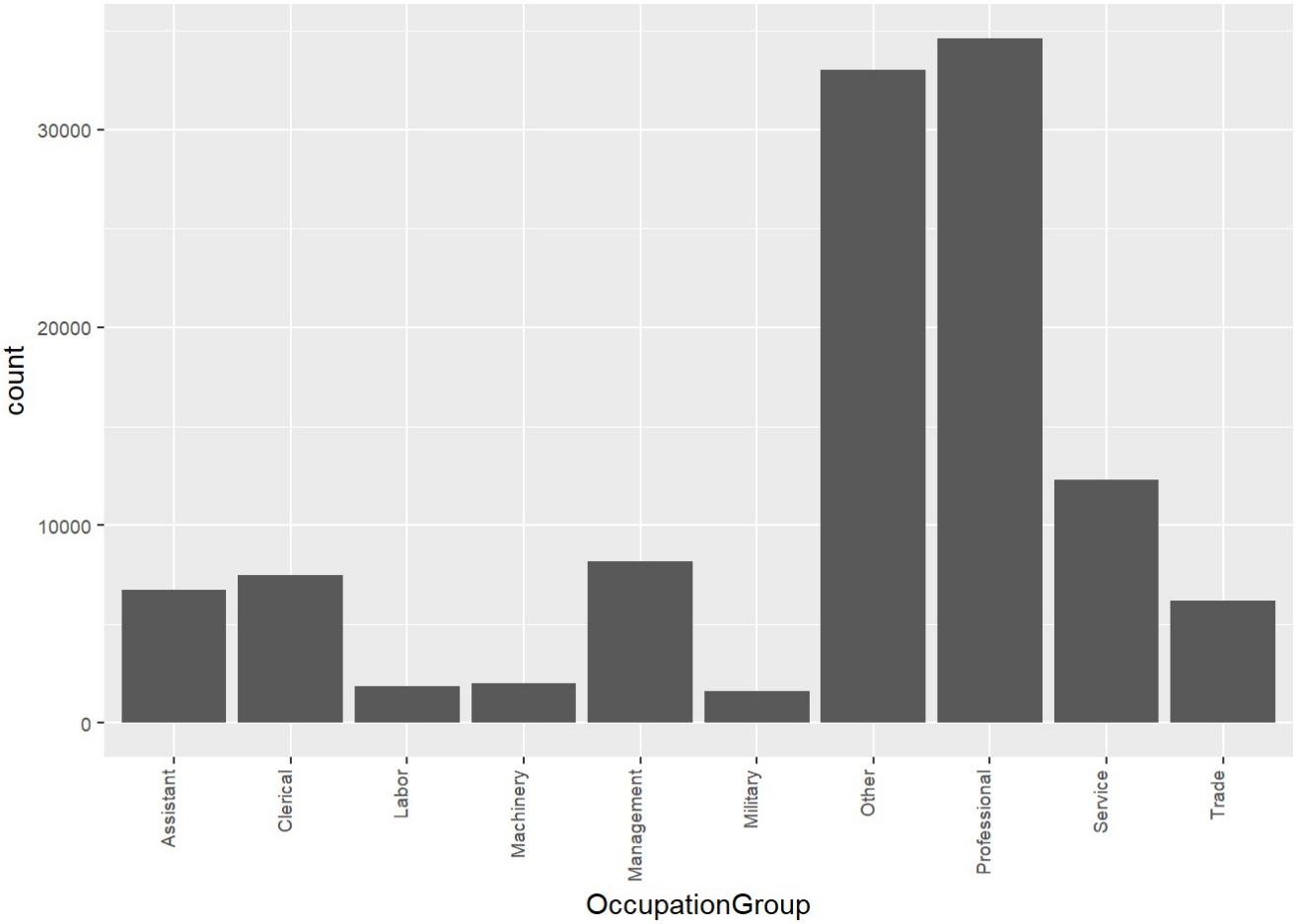
```
##  CreditScoreRangeLower CreditScoreRangeUpper CurrentCreditLines
##  Min.   :  0.0         Min.   : 19.0         Min.   : 0.00
##  1st Qu.:660.0         1st Qu.:679.0         1st Qu.: 7.00
##  Median :680.0         Median :699.0         Median :10.00
##  Mean   :685.6         Mean   :704.6         Mean   :10.32
##  3rd Qu.:720.0         3rd Qu.:739.0         3rd Qu.:13.00
##  Max.   :880.0         Max.   :899.0         Max.   :59.00
##  NA's   :591           NA's   :591           NA's   :7604
##  CurrentDelinquencies DebtToIncomeRatio EmploymentStatus
##  Min.   : 0.0000      Min.   : 0.000    Length:113937
##  1st Qu.: 0.0000      1st Qu.: 0.140    Class :character
##  Median : 0.0000      Median : 0.220    Mode  :character
##  Mean   : 0.5921      Mean   : 0.276
##  3rd Qu.: 0.0000      3rd Qu.: 0.320
##  Max.   :83.0000      Max.   :10.010
##  NA's   :697          NA's   :8554
##  EmploymentStatusDuration EstimatedReturn  IncomeRange
##  Min.   :  0.00           Min.   :-0.183   Length:113937
##  1st Qu.: 26.00           1st Qu.: 0.074   Class :character
##  Median : 67.00           Median : 0.092   Mode  :character
##  Mean   : 96.07           Mean   : 0.096
##  3rd Qu.:137.00           3rd Qu.: 0.117
##  Max.   :755.00           Max.   : 0.284
##  NA's   :7625             NA's   :29084
##  ListingCategory..numeric. LoanOriginalAmount  Occupation
##  Min.   : 0.000            Min.   : 1000      Length:113937
##  1st Qu.: 1.000            1st Qu.: 4000      Class :character
##  Median : 1.000            Median : 6500      Mode  :character
##  Mean   : 2.774            Mean   : 8337
##  3rd Qu.: 3.000            3rd Qu.:12000
##  Max.   :20.000            Max.   :35000
##
##  OpenCreditLines IsBorrowerHomeowner EstimatedLoss    LenderYield
##  Min.   : 0.00   Length:113937       Min.   :0.005   Min.   :-0.0100
##  1st Qu.: 6.00   Class :character    1st Qu.:0.042   1st Qu.: 0.1242
##  Median : 9.00   Mode  :character    Median :0.072   Median : 0.1730
##  Mean   : 9.26                       Mean   :0.080   Mean   : 0.1827
##  3rd Qu.:12.00                       3rd Qu.:0.112   3rd Qu.: 0.2400
##  Max.   :54.00                       Max.   :0.366   Max.   : 0.4925
##  NA's   :7604                        NA's   :29084
##  EstimatedEffectiveYield
##  Min.   :-0.183
##  1st Qu.: 0.116
##  Median : 0.162
##  Mean   : 0.169
##  3rd Qu.: 0.224
##  Max.   : 0.320
##  NA's   :29084
```

Occupation



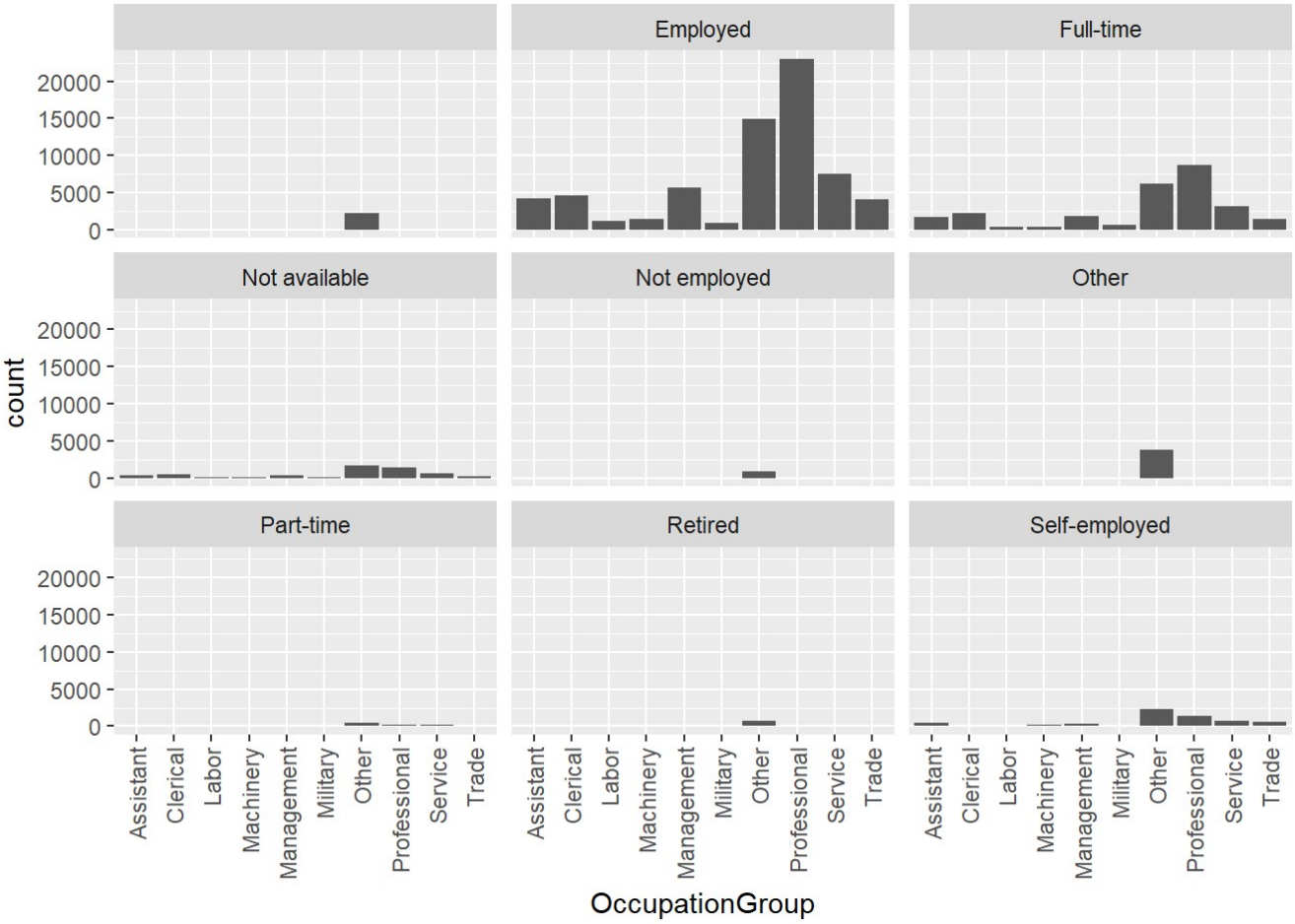
First we will take a look at the occupation of those who were given loans through Prosper. Almost 1/3 of professions are defined as “Other” and the second largest group is “Professional”.

We will re-organize the data to see if Occupation gives better insights of the Occupations likely to receive a loan from Prosper.



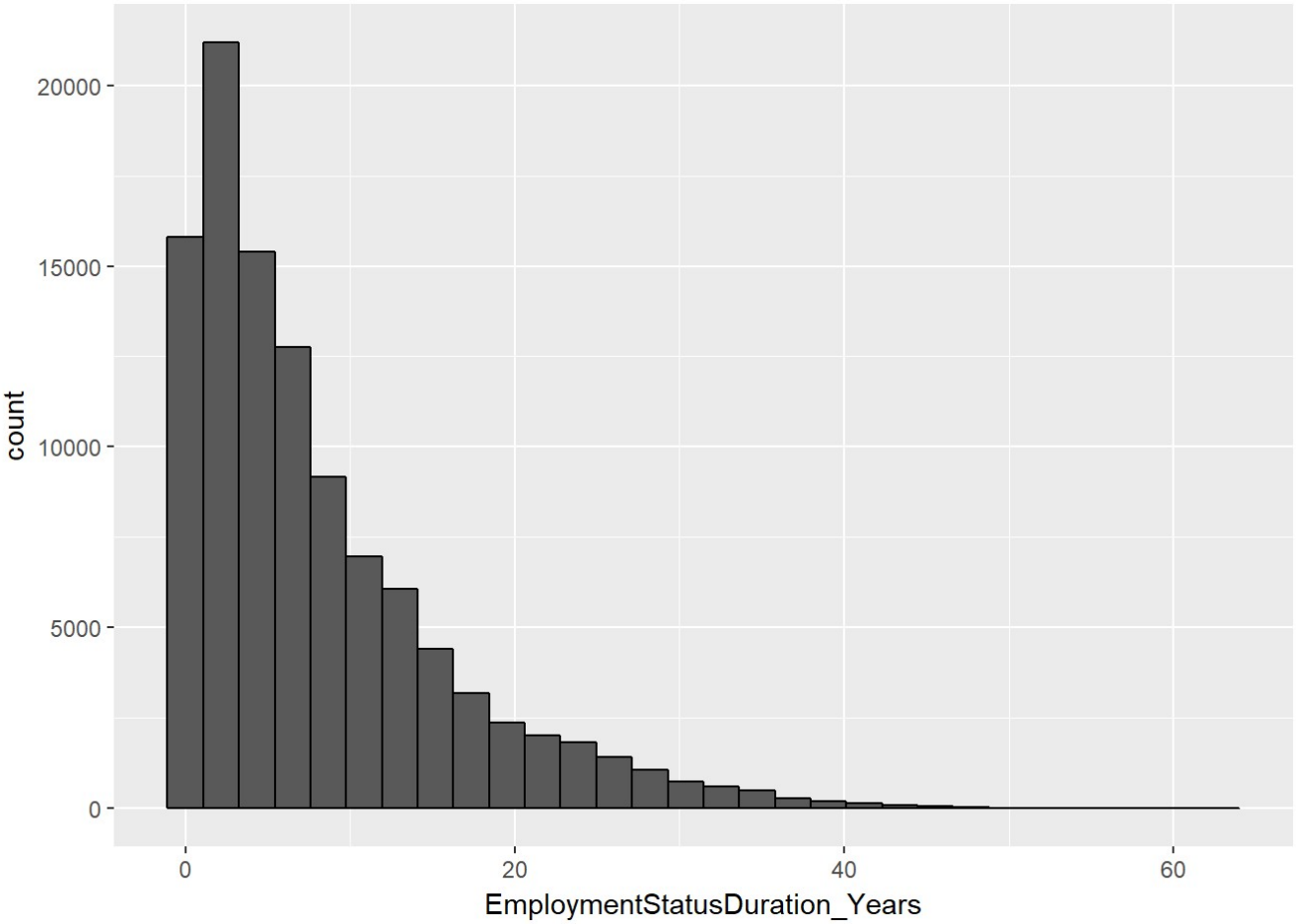
In order to better visualize the data by occupation we re-organize by categories as defined by the International Standard Classification of Occupations(ISCO).

While our two main groups are still Professional and Other, we get a much clearer picture of the break-down of loans by occupational group.

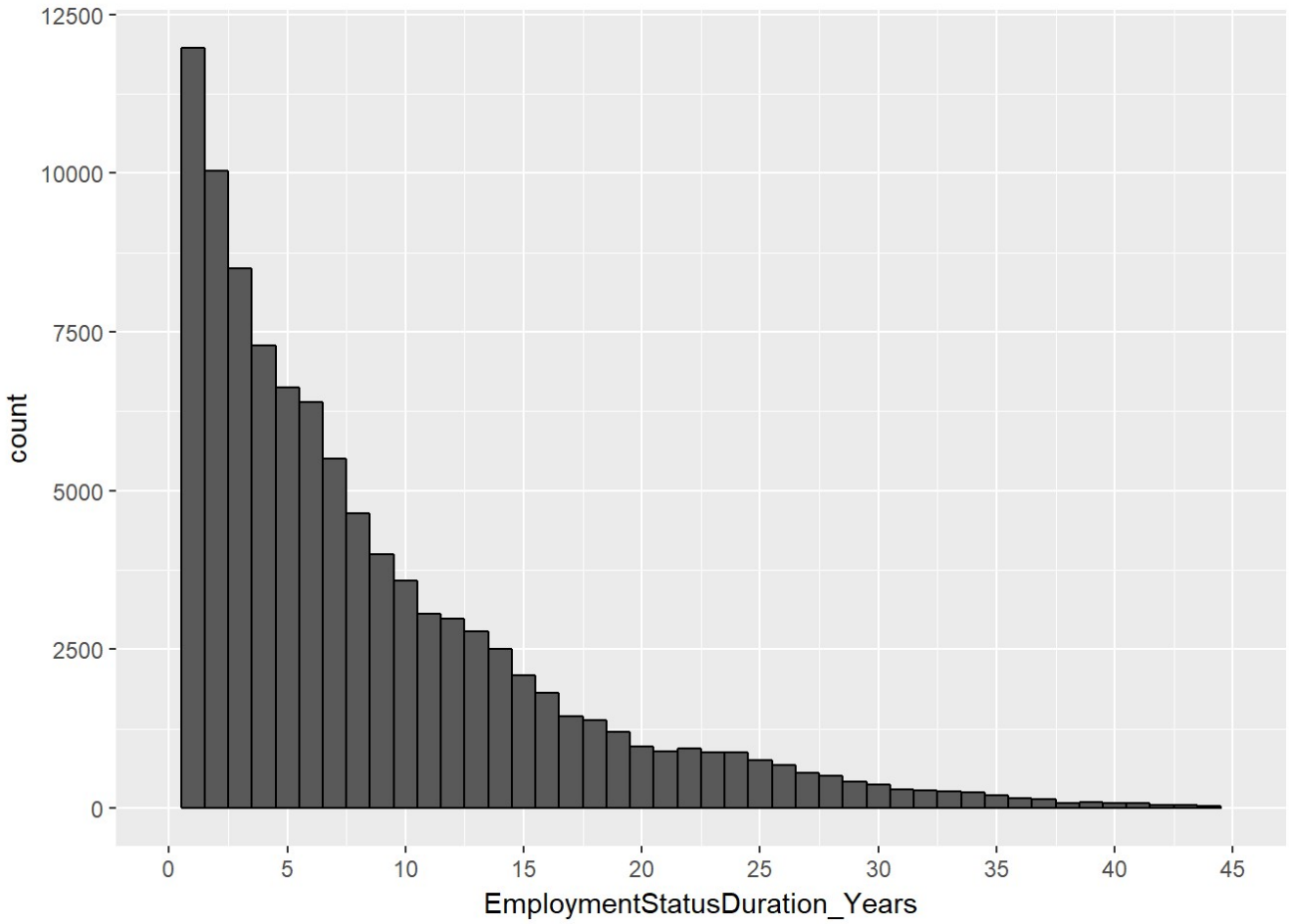



```
##                  Employed     Full-time Not available  Not employed
##          2255         67322         26355          5347           835
##         Other     Part-time       Retired Self-employed
##          3806          1088           795          6134
```

Evaluating employment status seemed to be another logical variable to investigate in determining who the loan requestor is however it is a dead end. The criteria used and the dispersion of values are not useful. It is interesting and perplexing to see that this value seems to have so little importance.



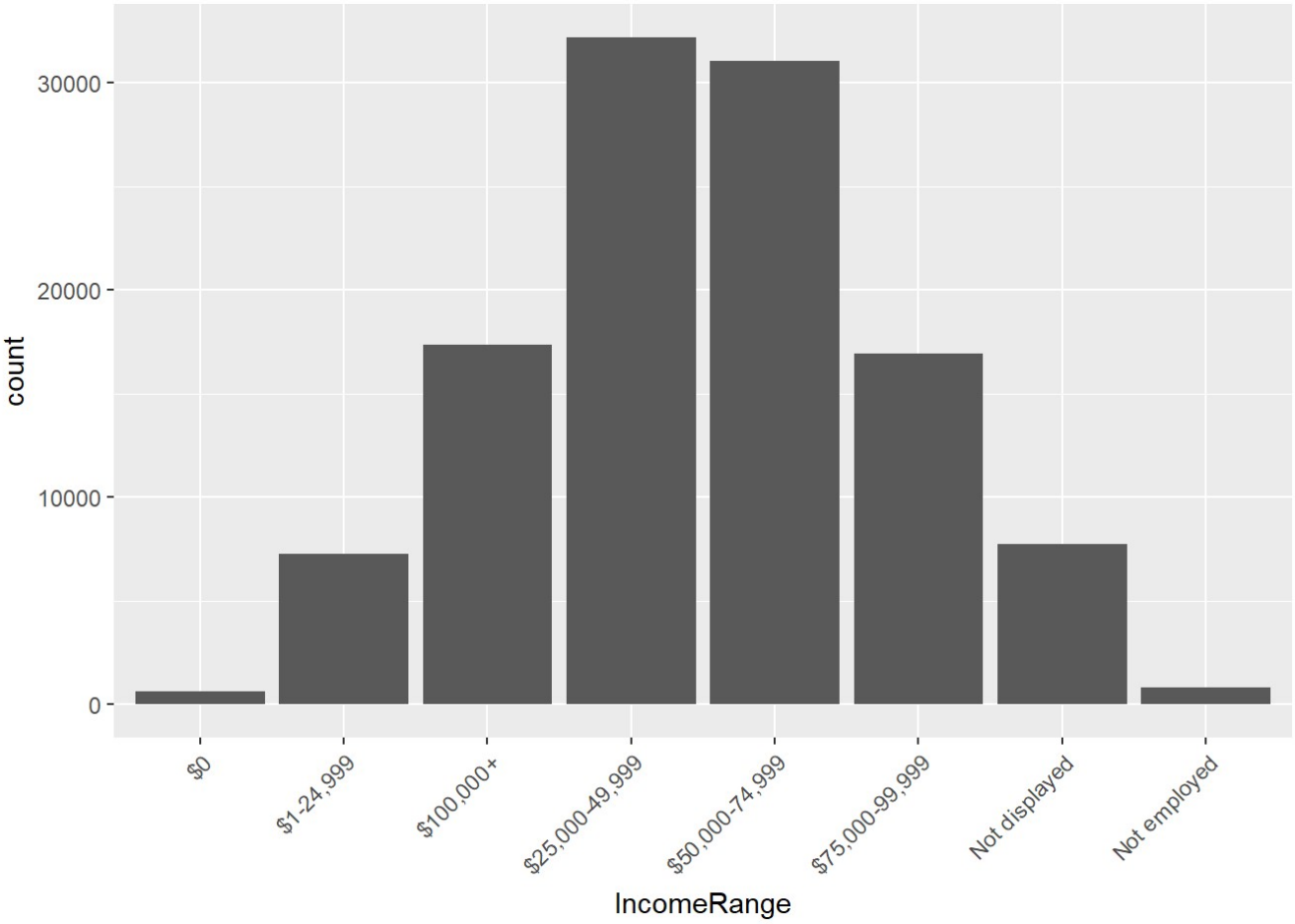
Employment status duration is in months so we will convert it to years for clarity. It looks like there may be considerable outliers here or there are a number of loan requestors are at least 80 years of age. If the latter is the case this would be quite surprising because Prosper's loan process is completely internet based.




```
summary(working_data$EmploymentStatusDuration_Years)
```

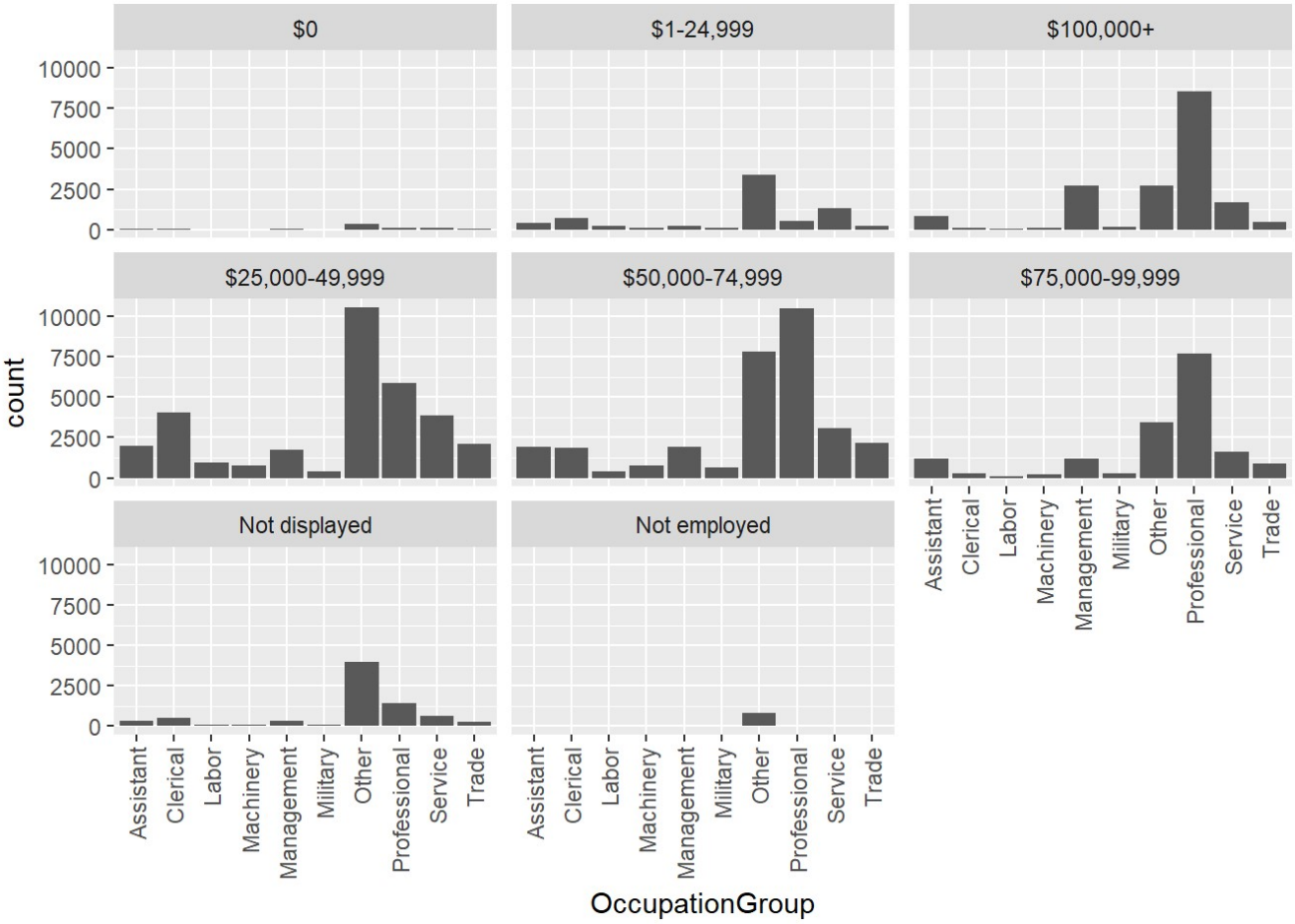
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   2.167   5.583   8.006  11.417  62.917    7625
```

The average time working is less than 10 years, with 75% of loan requestors spending 11 years in their job.



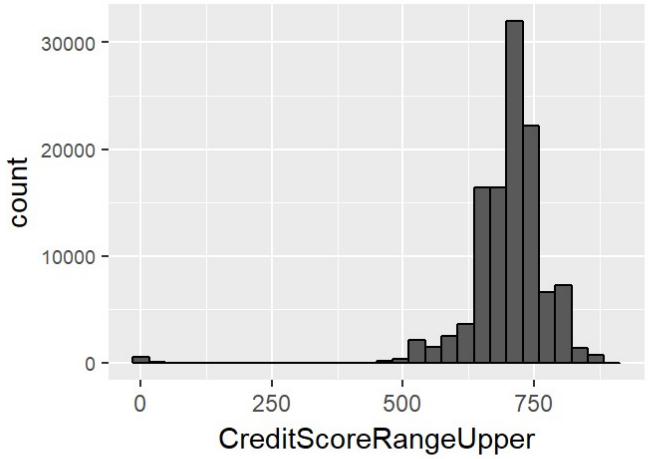

```
##             $0      $1-24,999      $100,000+ $25,000-49,999 $50,000-74,999
##            621           7274          17337          32192          31050
## $75,000-99,999  Not displayed   Not employed
##          16916           7741            806
```

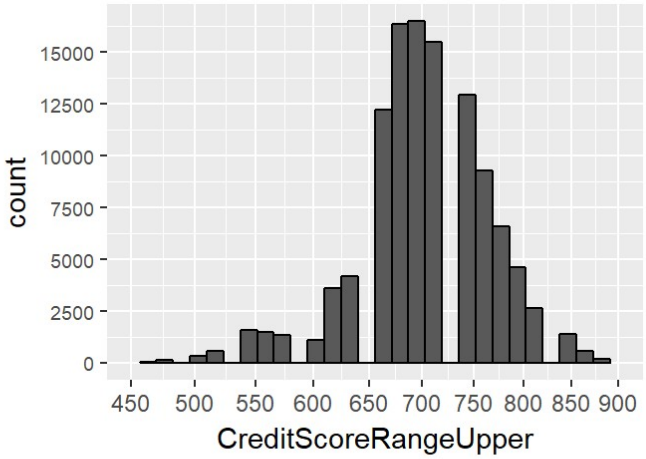
The choice was made to look at the Income Range from the dataset because it seemed less likely for a loan applicant to exaggerate income versus declaring a single income amount. This gives us the opportunity to examine other factors more thoroughly than interrogating the validity of income information.

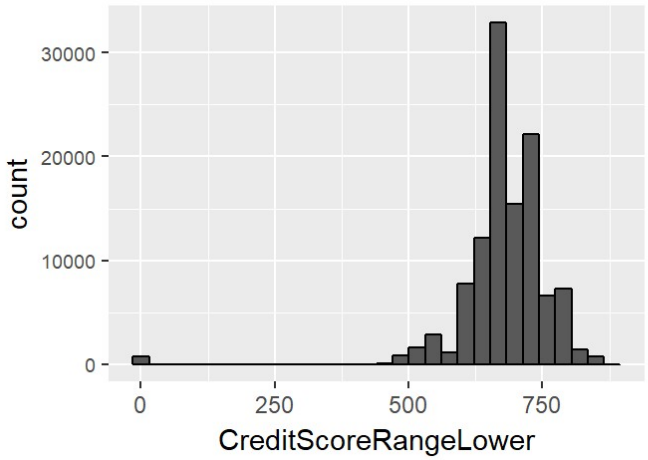


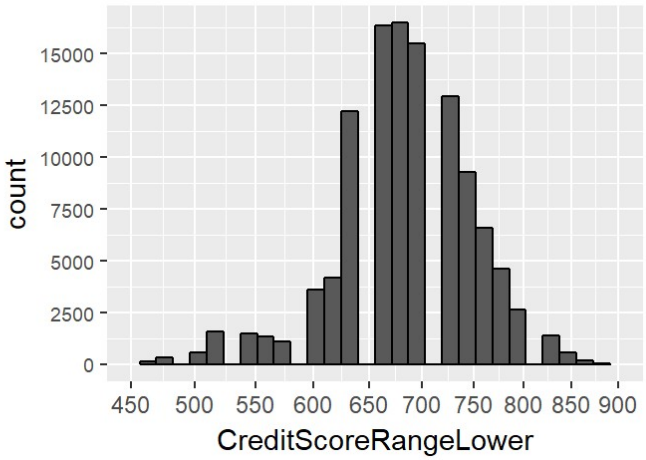
Credit

Prosper uses its own grading system for categorizing a loan requestor instead of traditional FICO score. The data available is not sufficient to derive how this scoring system equates to traditional methods. We decided to use the more traditional metric of credit score to shape the view of our loan requestor, and thus the variables CreditScoreRangeUpper and CreditScoreRangeLower are used.





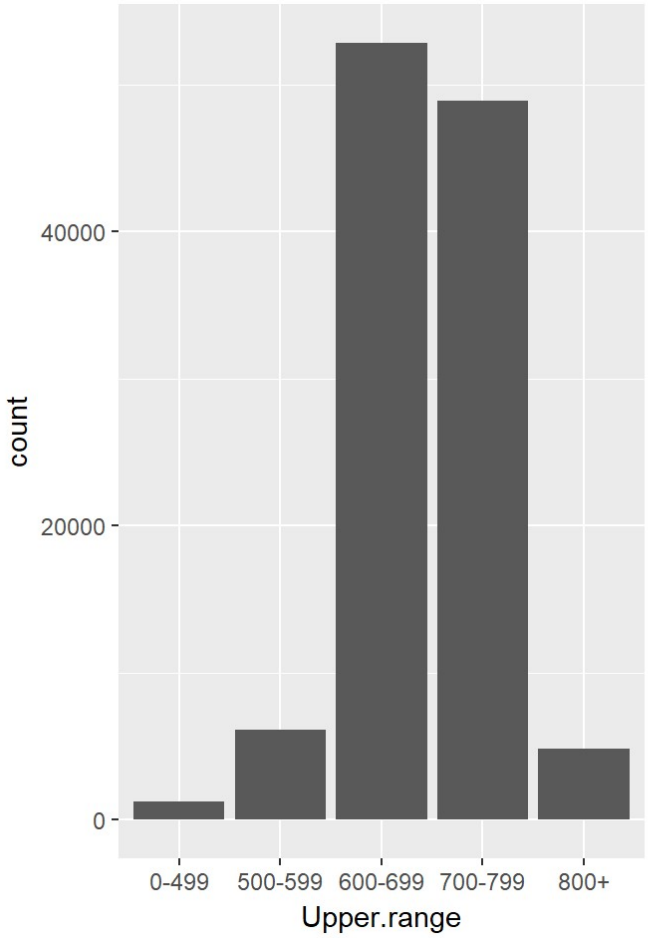


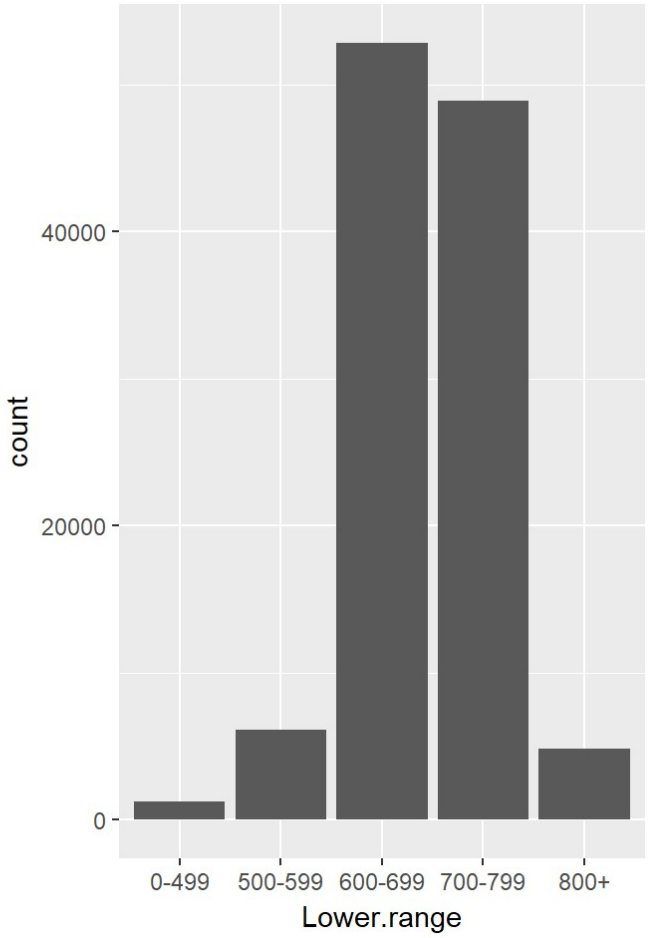



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   679.0   699.0   700.9   739.0   899.0
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0     660     680     682     720     880
```

Evaluating the upper and lowercredit score limit of the requestor is gives us an unexpected picture of the distribution of the upper credit scores. While it does reflect the quantile breakdown of the credit score it is worth noting that we can propose that in order to qualify for a Prosper Loan the requestor stands a significantly higher chance if the credit score is above 650.



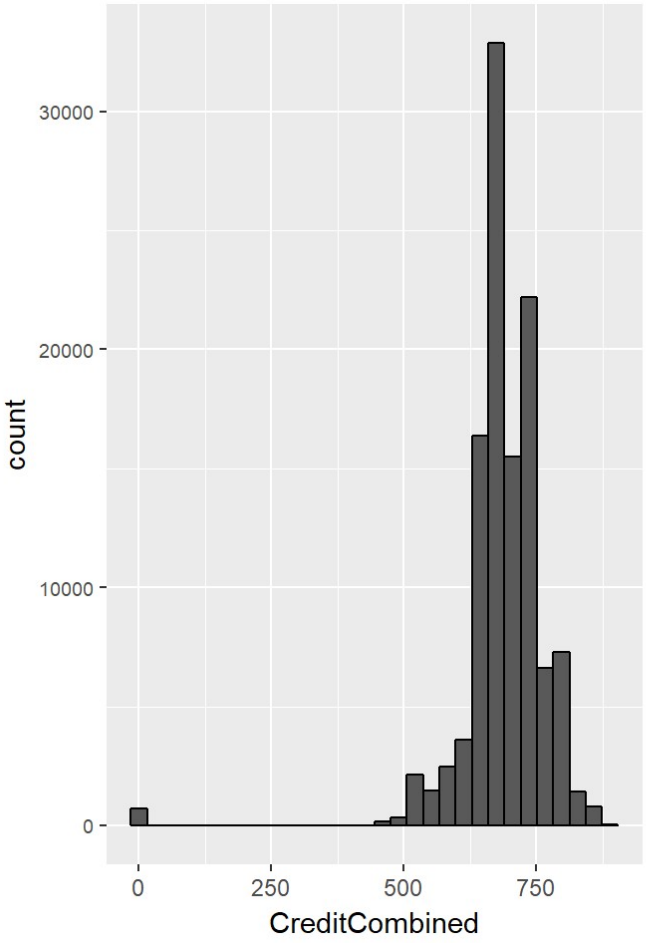


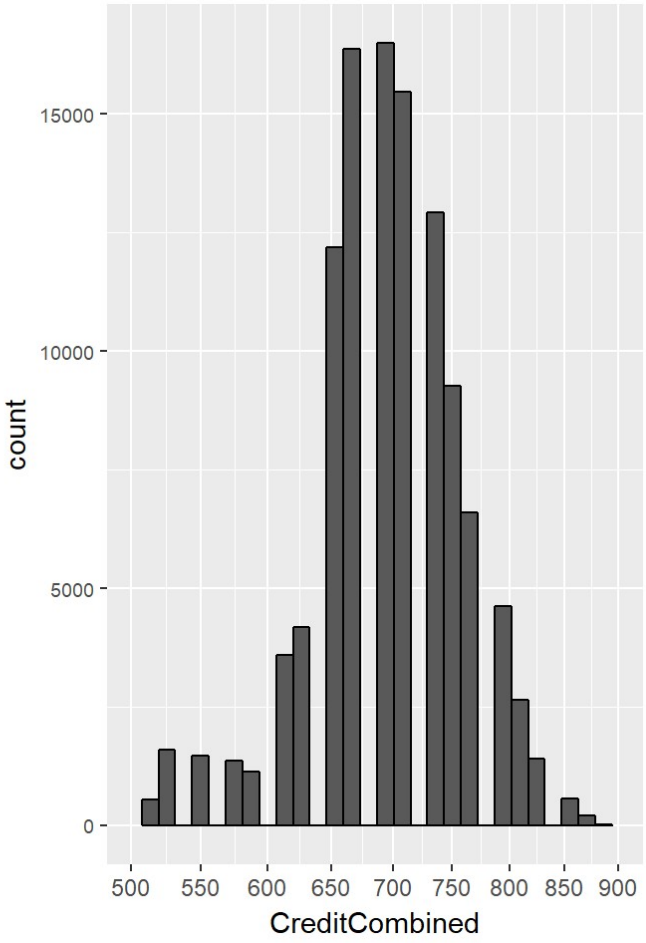
It is important to note that the values of 0 - 499 also contain missing data from the dataset. Specifically there were 451 values with no score.
The decision was made to include the data as a zero value because it had no impact on the overall values and did not affect the determination of who the loan requestor target is for a Prosper loan.

When we group upper credit scores it becomes more apparent that the majority of requestors have an upper credit score between 600 and 799.

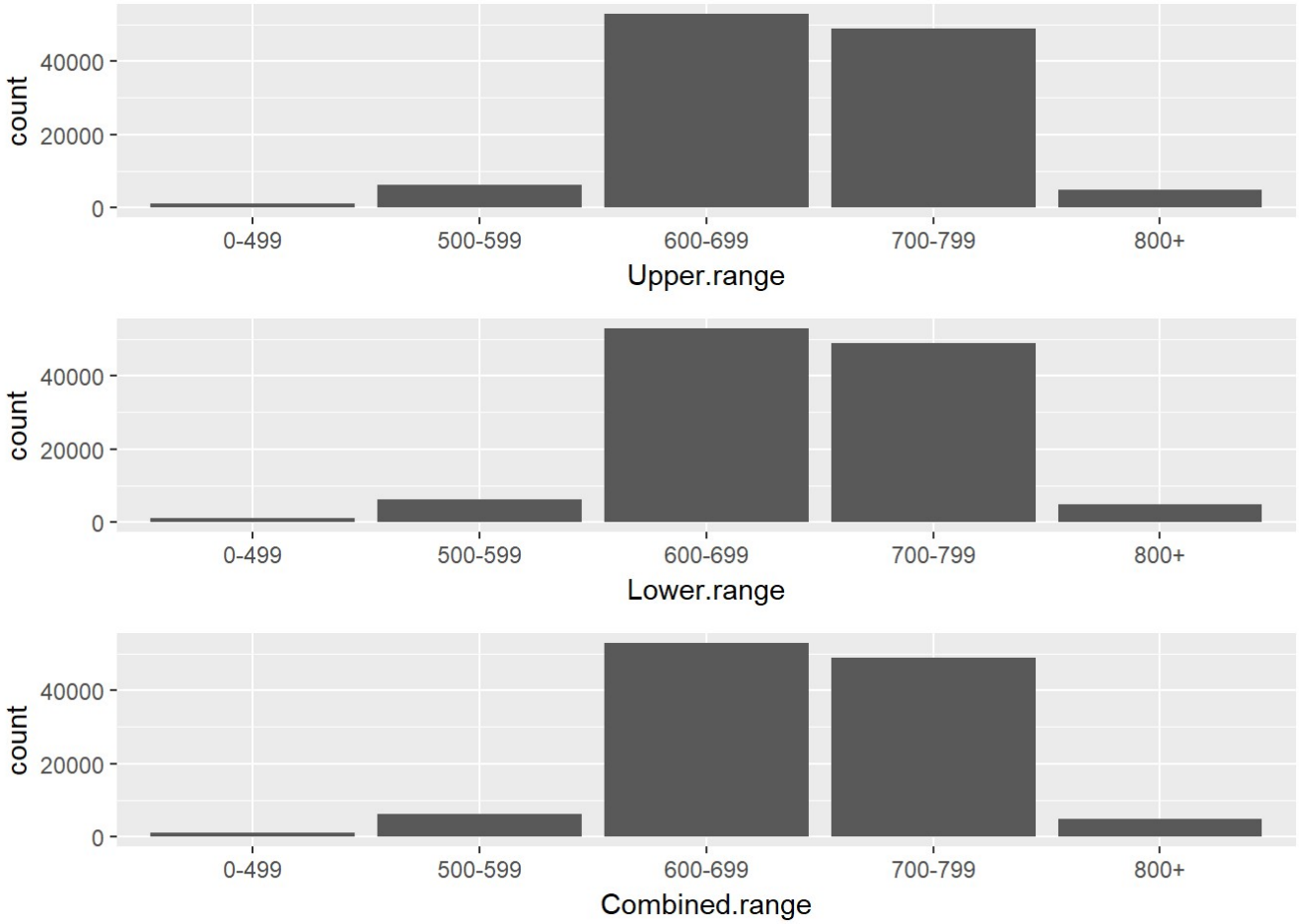
The lower credit score range is equally insightful with the majority of credit scores concentrated between 650 and 750. This also coincides with the upper credit score values indicating that there may not be a significant variance from upper to lower credit scores.

The results of categorization of the lower range returns the expected result. The lower range is not different from the upper range in value count. This tells us that either the requestor’s credit score was stable over time or that the initial request and second request were within a short time period.

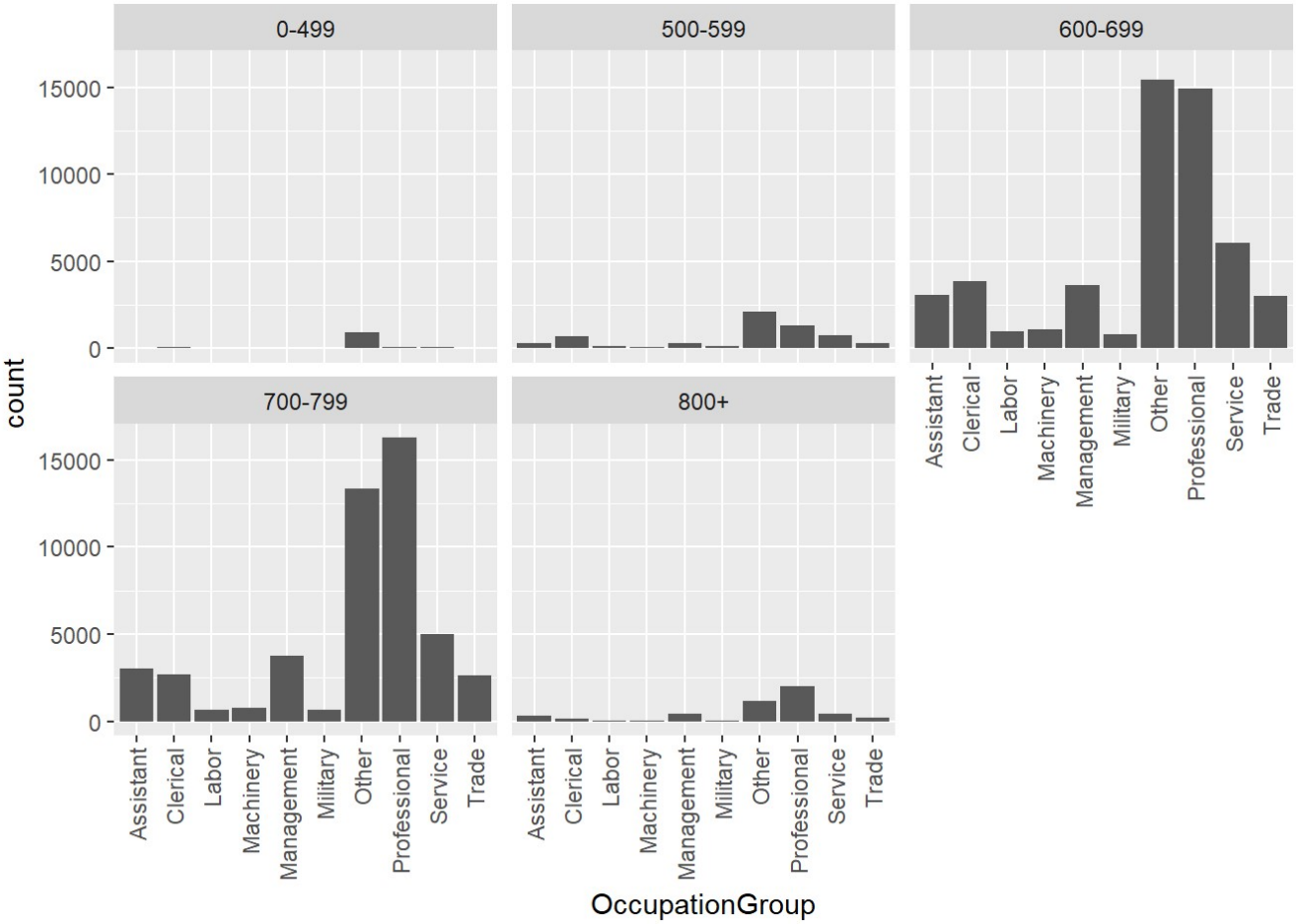





```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   669.5   689.5   691.5   729.5   889.5
```



The distribution across upper and lower when combined do not change.

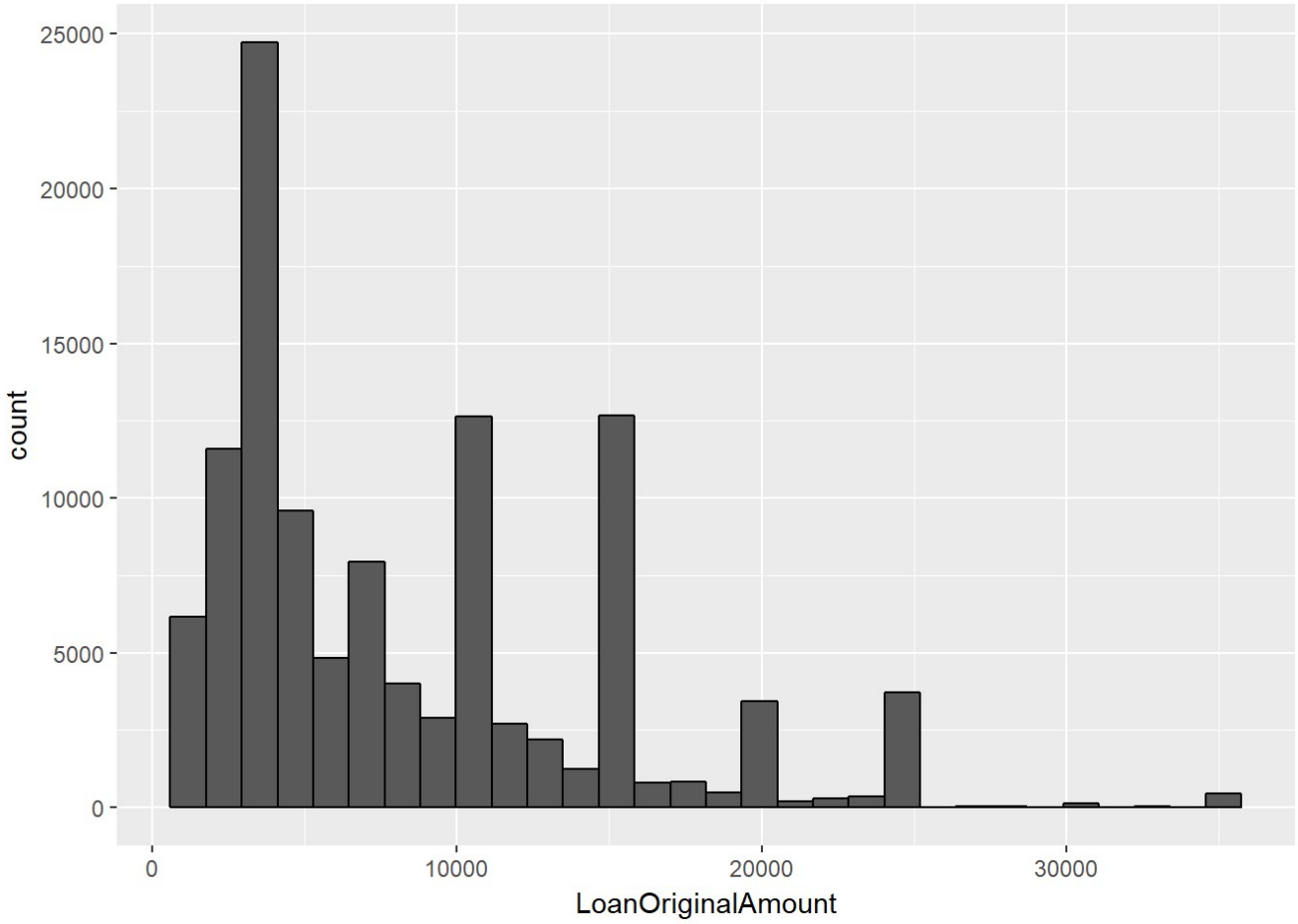


Evaluating by occupation supports the claim that in order to successful in the loan application process a requestor should have a score of at least 600.

```
## # A tibble: 10 x 4
##    Occupation   Lower Upper Combined
##    <fct>        <dbl> <dbl>    <dbl>
##  1 Assistant     690.  709.     699.
##  2 Clerical      669.  688.     678.
##  3 Labor         671.  690.     680.
##  4 Machinery     680.  699.     690.
##  5 Management    693.  712.     703.
##  6 Military      679.  698.     689.
##  7 Other         668.  686.     677.
##  8 Professional  695.  714.     705.
##  9 Service       680.  699.     689.
## 10 Trade         685.  704.     694.
```

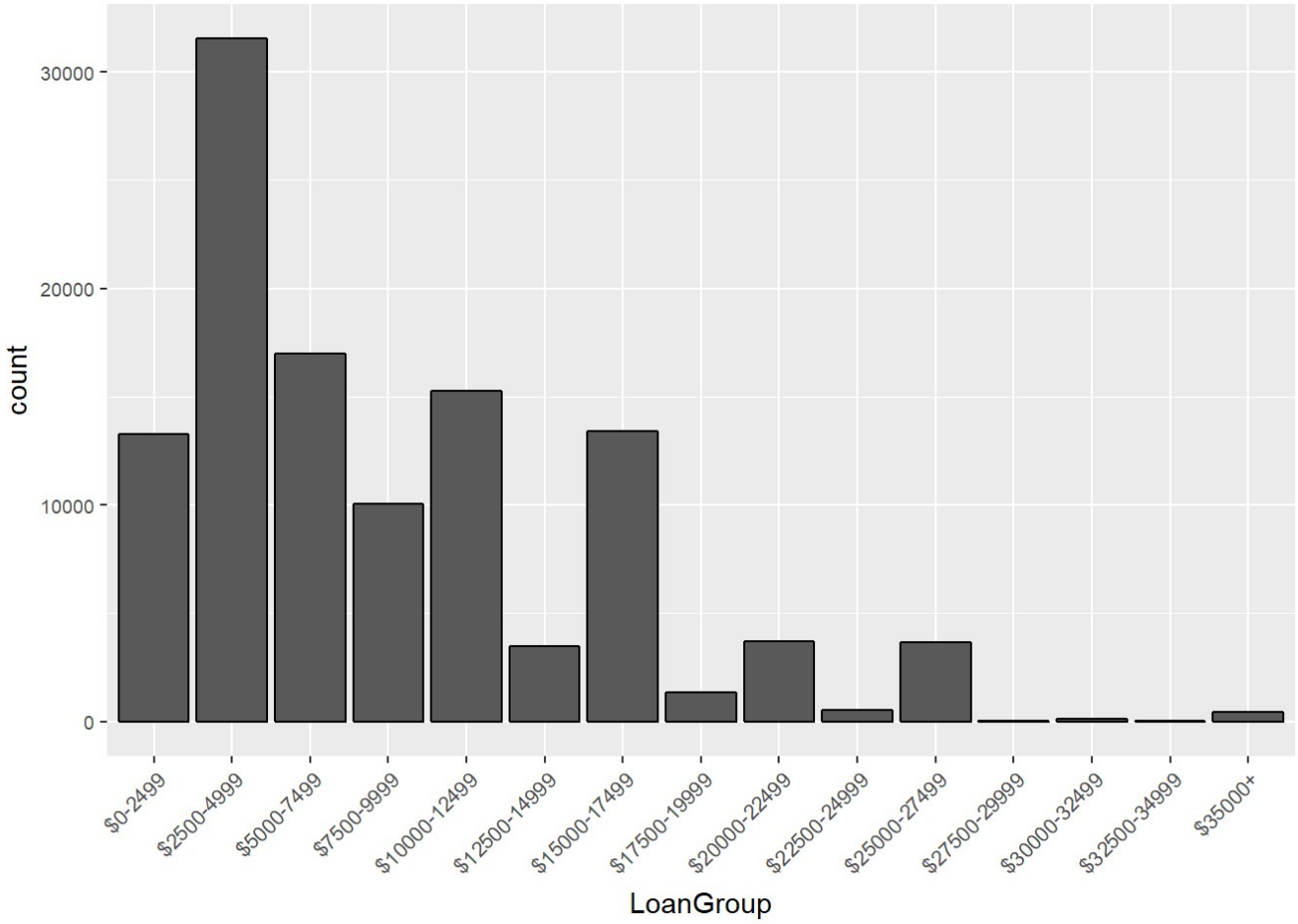
Once we average the lower and upper credit scores it is then that we begin to see the data shift slightly. The mean credit score across all groups drops slightly. Although there is a slight shift it solidifies the acceptable range for credit score when applying for a Prosper loan.

Loan Amount




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

We can see that 75% of the loans are between 1000 and 12,000 with a few loans reaching 35000. The loan amounts are relatively low which may be an indicator that these loans are intended for more personal use than a major purchase.




```
##
##       $0-2499    $2500-4999    $5000-7499    $7500-9999 $10000-12499
##         13288         31546         16972         10048         15258
## $12500-14999 $15000-17499 $17500-19999 $20000-22499 $22500-24999
##          3474         13405          1370          3712           554
## $25000-27499 $27500-29999 $30000-32499 $32500-34999       $35000+
##          3657            40           145            38           430
```

With the most popular loan range being between 2500 and 4999, and the majority of loans being less than 12,500, it appears that the primary loan market is personal loans with possibly a few major purchases.

Univariate Analysis

What is the structure of your dataset?

The original dataset contains 113,937 loan entries with 81 different variables. The dataset used for analysis contains 16 of those original variables with an additional 10 variables created to provide alternative insights into the data.

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

The motivation for this analysis is to discover who is the loan requestor, what is the purpose and amount of the loan, and what factors indicate a loan application will be approved. Because Prosper uses a proprietary system to determine these factors the followng features are the core of the analysis: Occupation, Income, Credit Score

and Employment Status. These factors were chosen because they are more traditional factors used in loan applications.

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

Other features that will assist in shaping the results of the analysis may be: Current Delinquencies, Debt to Income Ratio, Current Credit Lines, Estimated Return, Current Revolving Accounts, Public Records 12 months, Effective Estimated Yield and Estimated Loss. Many of these features are used in determining traditional credit scores but they may give insight into the range of loans a requestor qualifies for and the amount of the approved loan.

Did you create any new variables from existing variables in the dataset? Did you perform any operations on the data to tidy, adjust, or change the form of the data? If so, why did you do this?

There were a number of modifications made to the dataset, the first was to create the dataset that will be evaluated in the analysis.

The overall decision to create new variables instead of refactoring the original ones was made to provide consistency and not lose any other potential insights that the original values may provide.

The variable **OccupationGroup** was created to standardize the occupational observations of the loan requestor. This was broken down categorically in accordance with ISCO standards.

Both UpperCreditRange and LowerCreditRange were broken into categories **Upper.Range** and **Lower.Range** to give a more concise view of their ranges. Then those values were combined into **CreditCombined** and **CreditCombinedRange** which averages the upper and lower credit scores.
This was done because the difference between the upper and lower scores was marginal.

CategoryName was created from ListingCategory as the values are numeric to see the distribution of loans by name.

Finally, LoanOriginalAmount was categorized as **LoanGroup** to provide a more consistent approach to observing loan values.

Of the features you investigated, were there any unusual distributions?

UpperCreditRange and LowerCreditRange initially appear bimodal and when scaled it was very revealing in that it clearly shows Prosper’s preferred market from a credit score perspective.

LoanOriginalAmount has a multimodal distribution with peaks every 2500. While this may appear unusual it make sense that these amounts would be general cut off values given that Prosper Loans minumum and maximum values are currently 2000 - 40000.

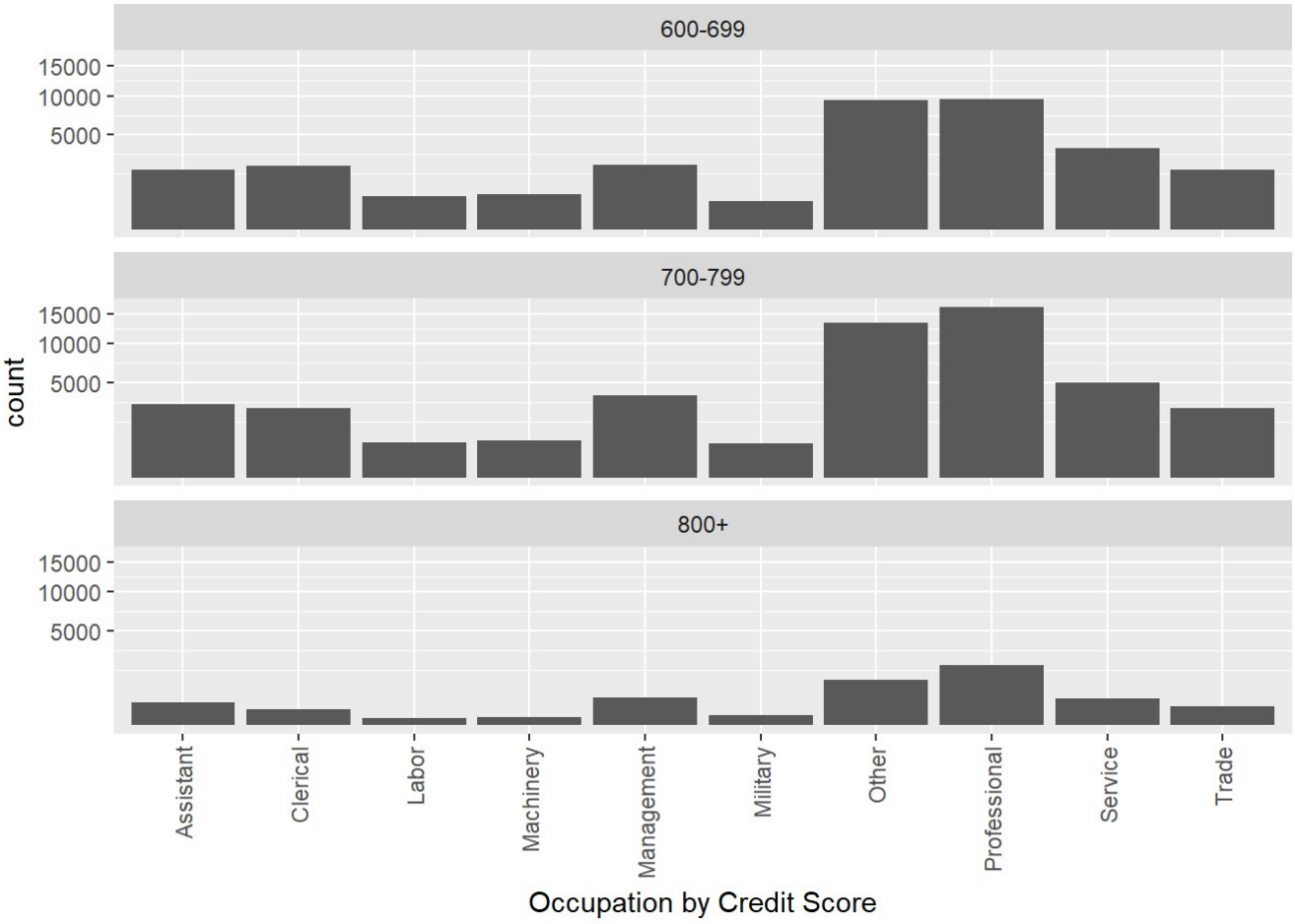
Reflections

With a cursory look at who Prosper is giving a loan, it appears that the majority of loan requestors are in one of these groups: Other, Professional, Service or Trade. They are employed and their income is more than $25,000 with a credit score of at least 650. The loan amount requested is less than $17,500 and is primarily in these catgories: Business, Debt Consolidation, Home Improvement, or Personal.

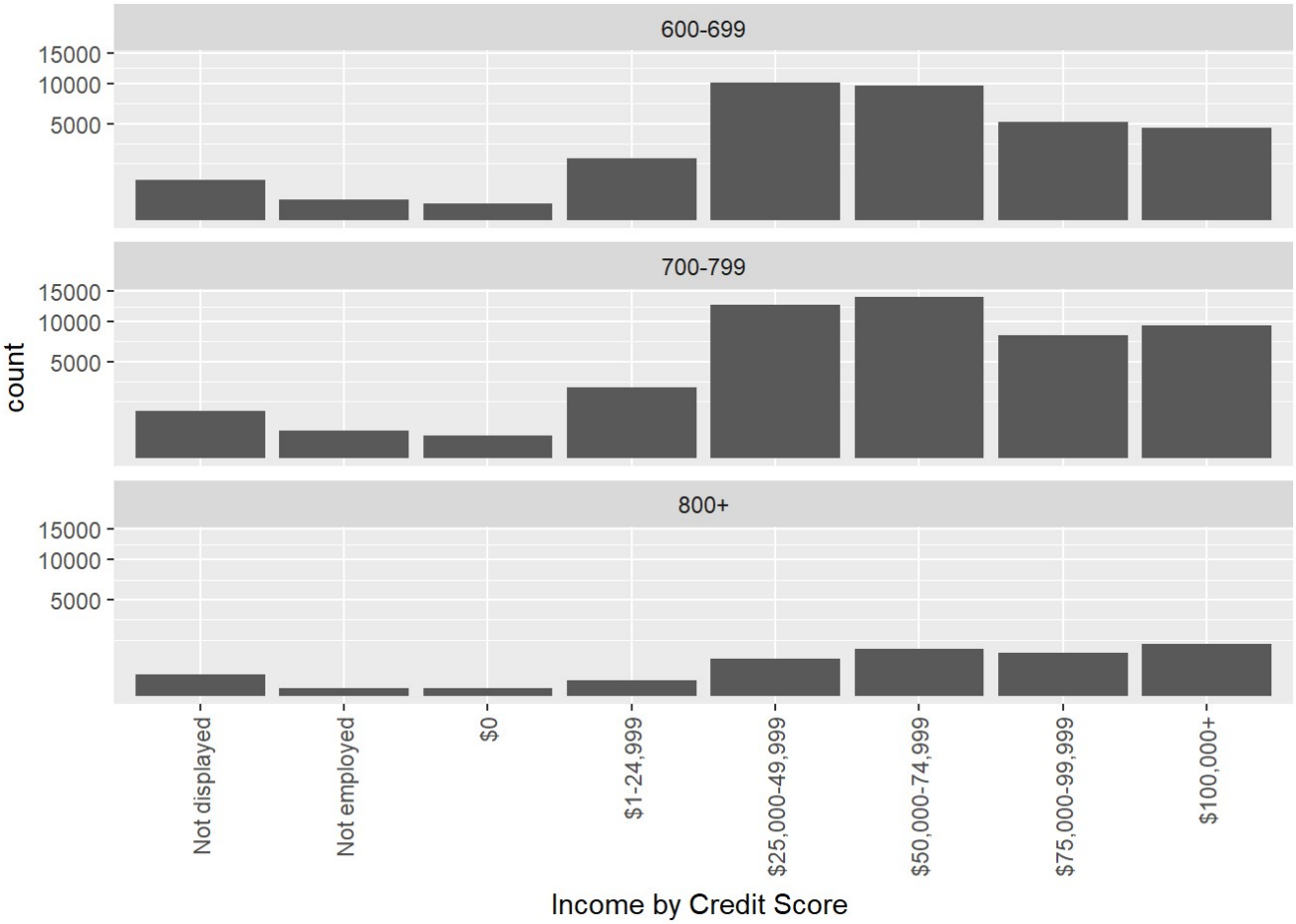
Bivariant

Loan Requestor

The there is a strong indication that minimum preferred credit score is 650, so the dataset has been narrowed to meet this minimum. This brings the total entries included from 113,937 to 86,608.

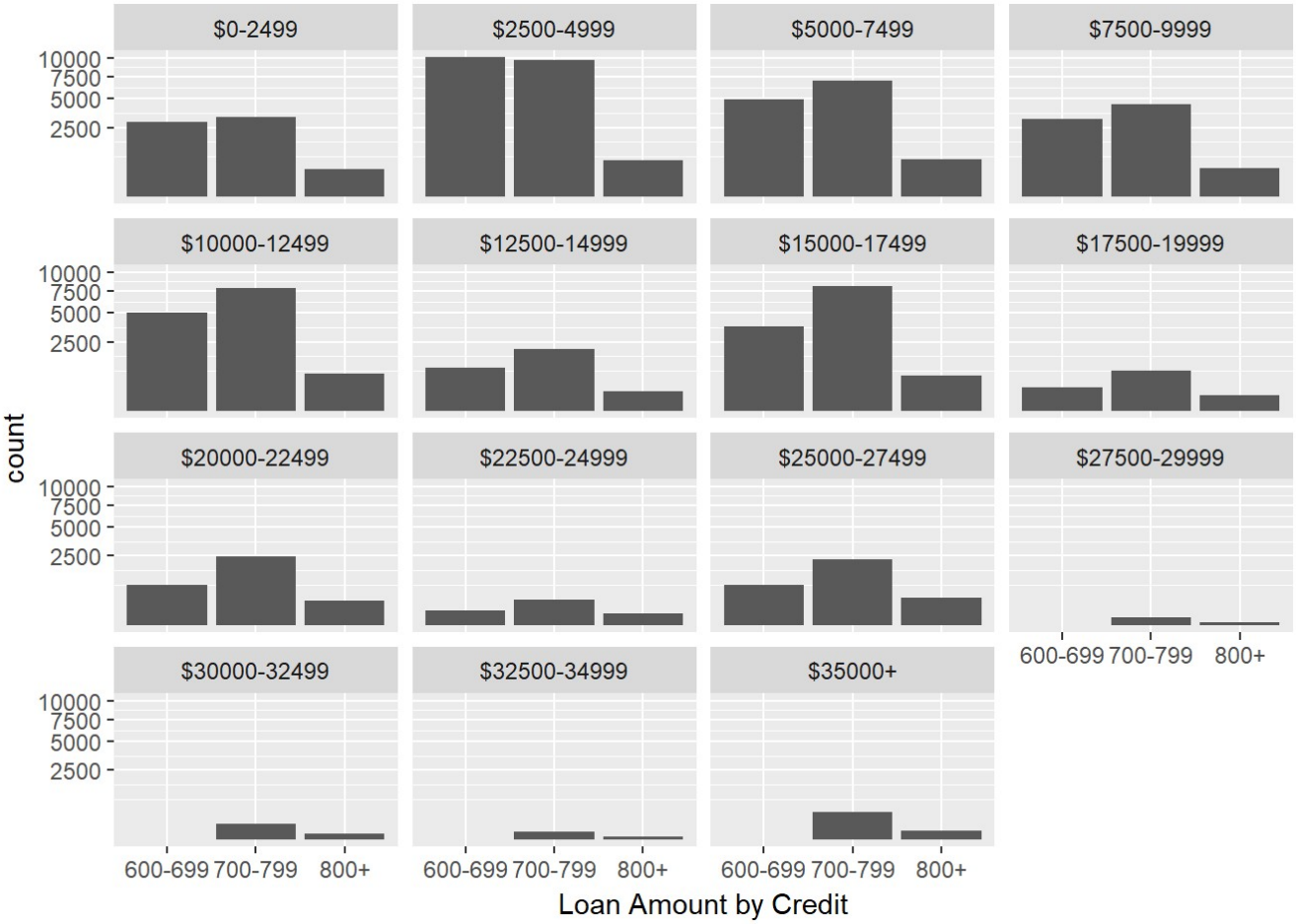


Labor, Machinery and Military have the least number of loan requestors in the range of 650 and above while all other occupations are well represented.

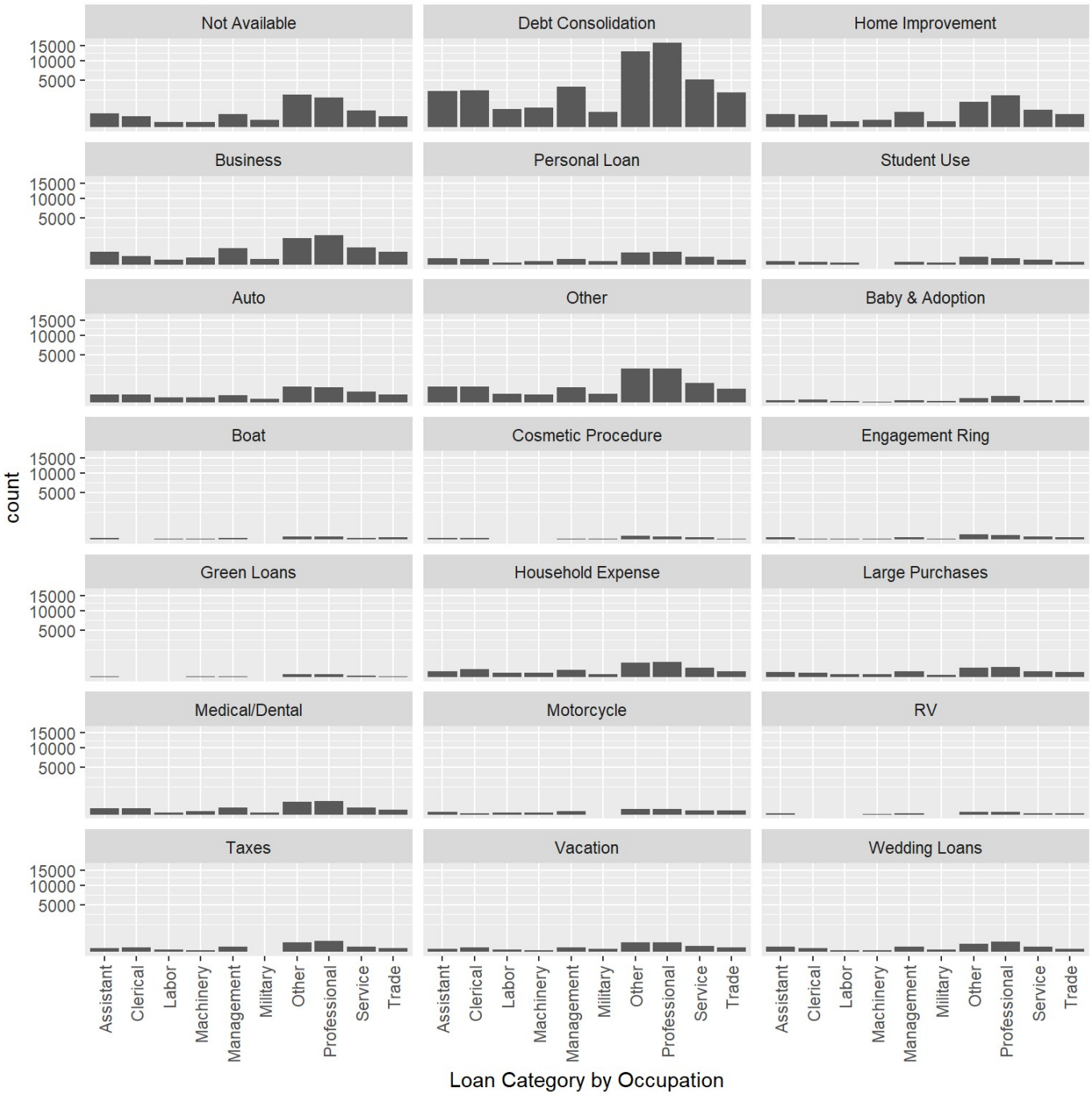


No surprise here to see that the income range is predominately 25,000 and above.

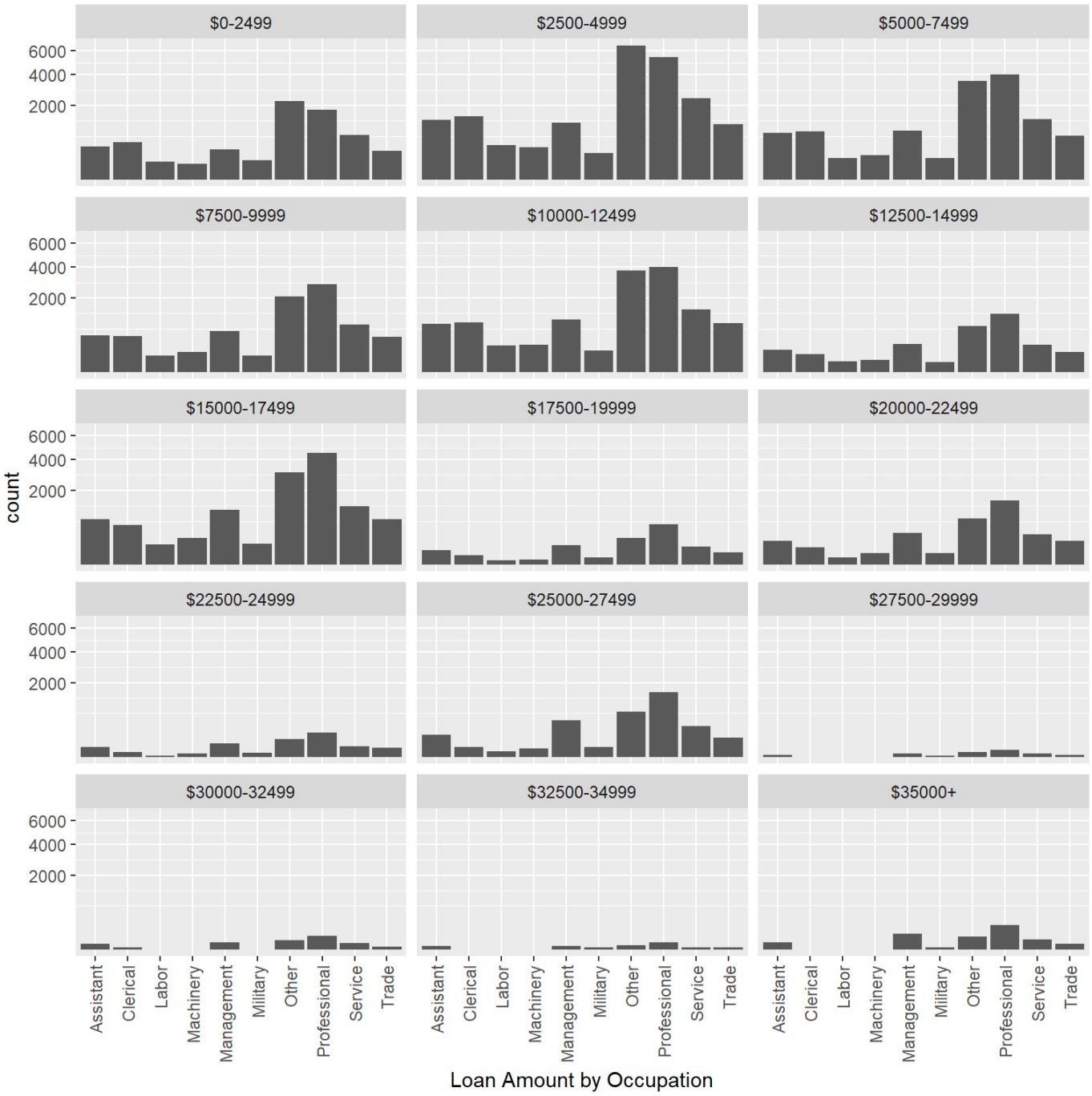
Loans, Requestor and Credit



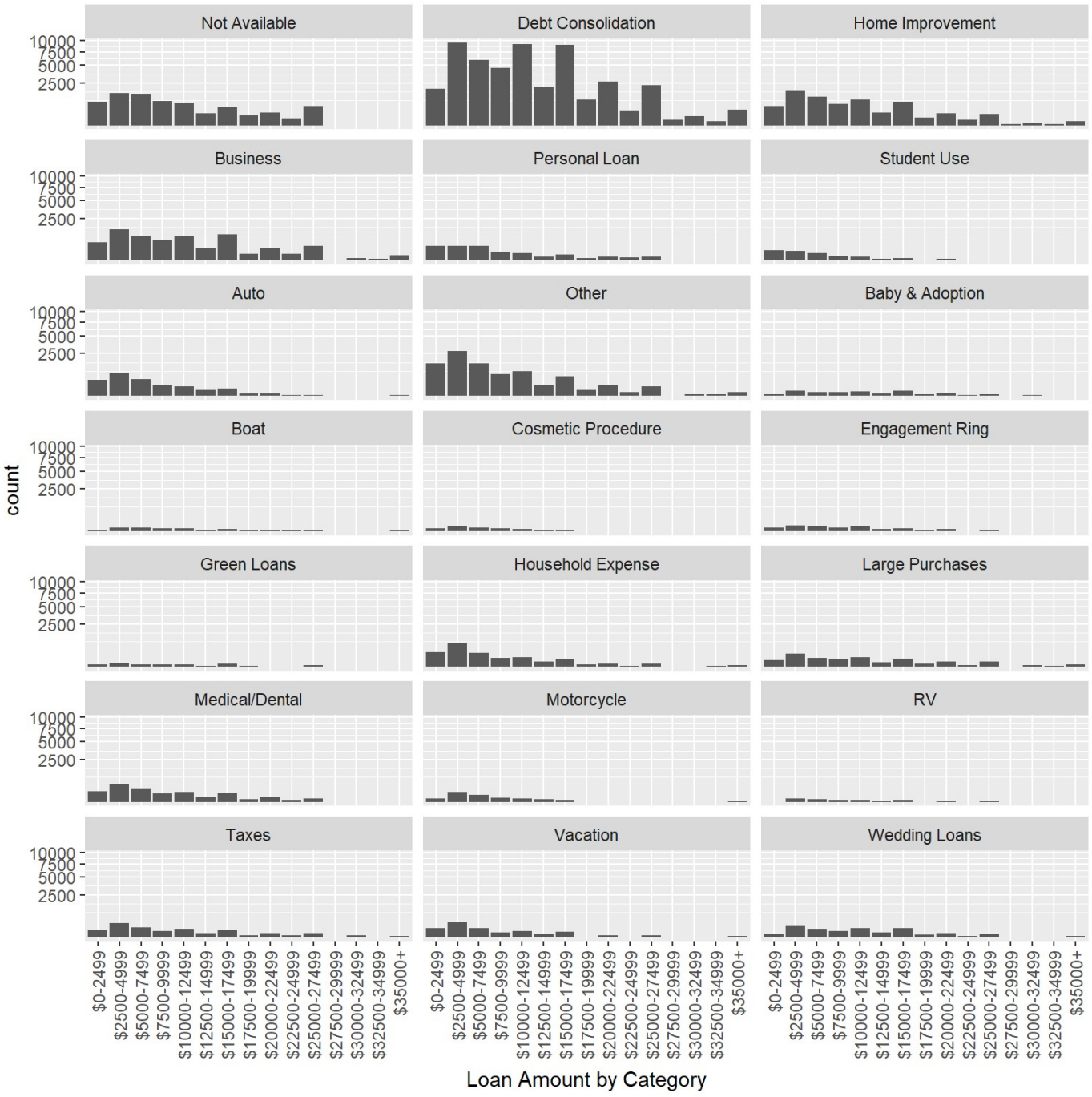
If the credit score is below 700, it appeears that the maximum acceptable loan is less than 27,500.



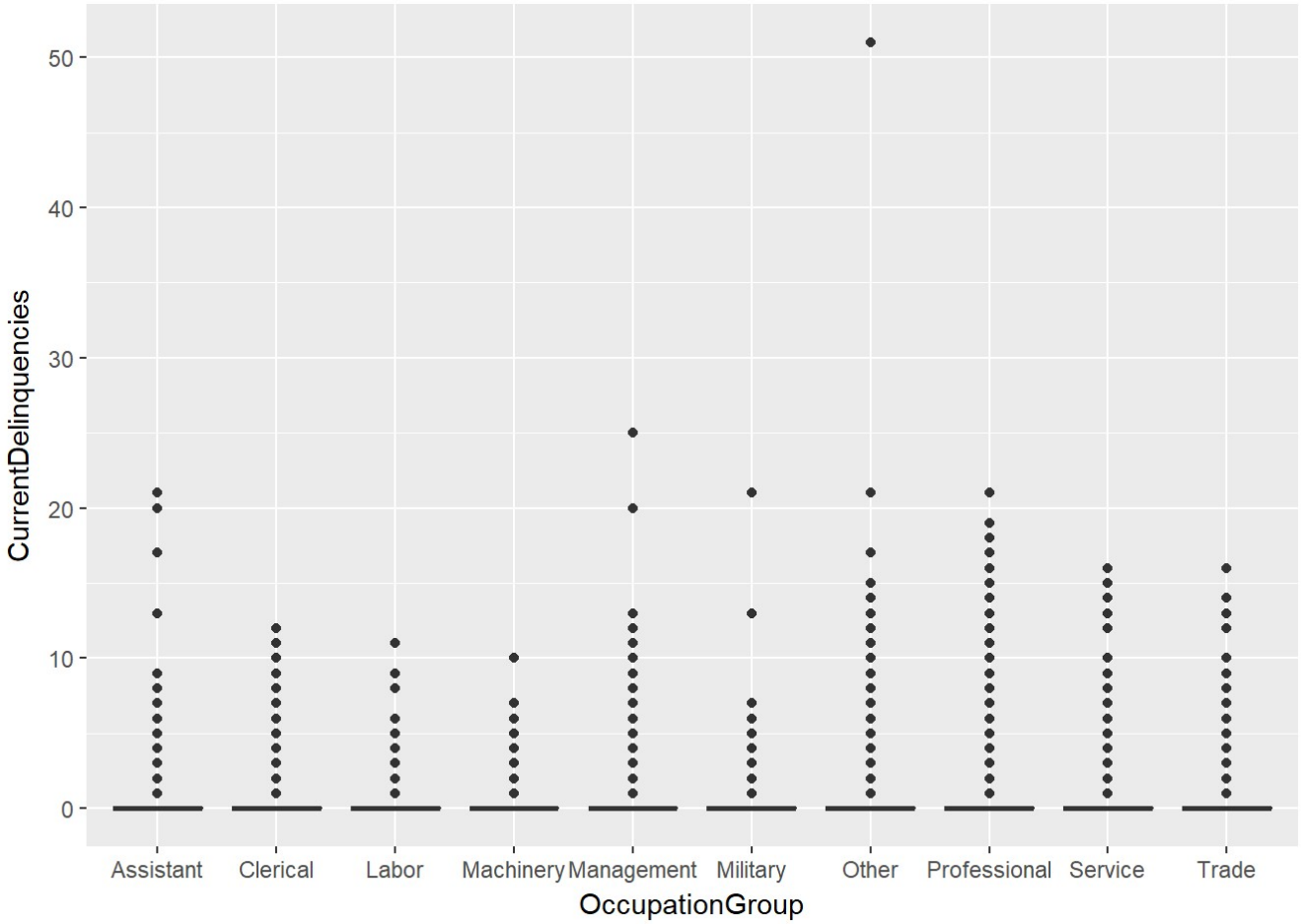
The data bears out that the majority of loans are for debt consolidation, instead of large purchases.



It seems that if your occupation is clerical, labor or machinery the likelyhood of requesting or receiving a loan greater than 27,000 is very unlikely. This is interesting because it may indicate that they are in the credit score range of less than 700.

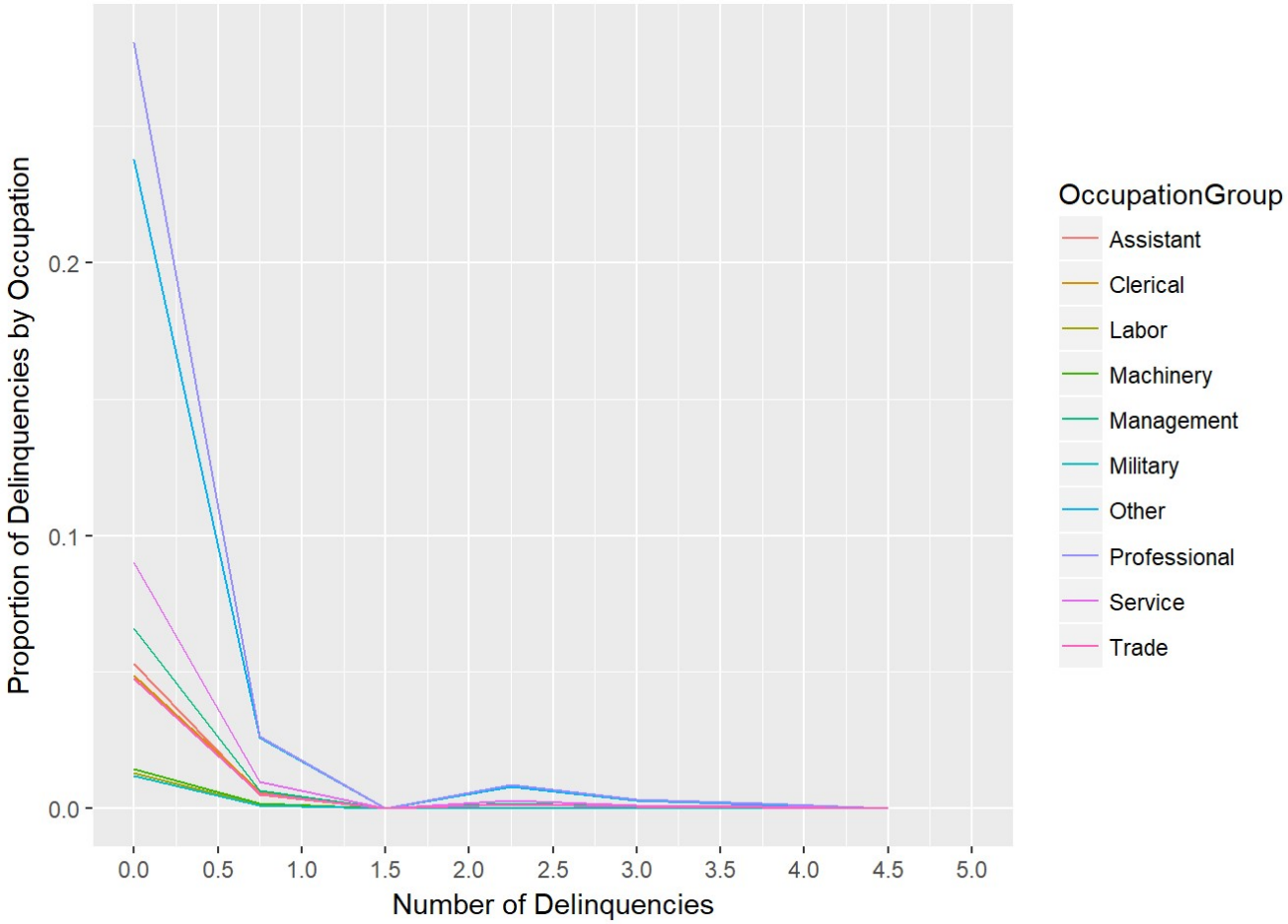


Here again, no matter how much money is borrow it is predominately for debt consolidation. This is very unusual and would indicate a high level of risk for those investing in loans.



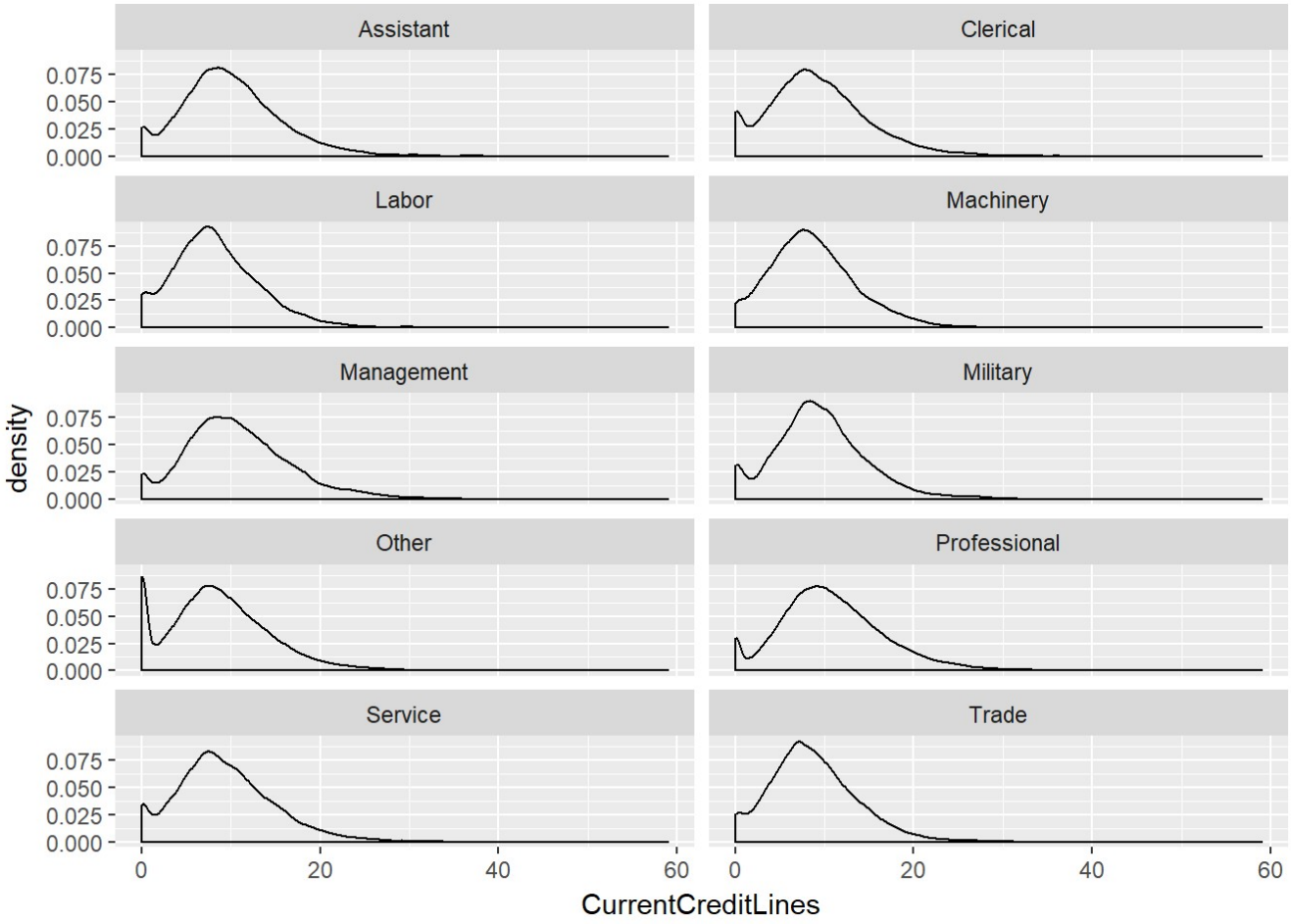
This is not what we expect to see when looking at delinquencies. It is safe to say that there are quite a number of outliers. We expect that the number of delinquencies be very low in order to qualify for a loan. It is very surprising that there are values in the 50 range.

Other Determining Factors



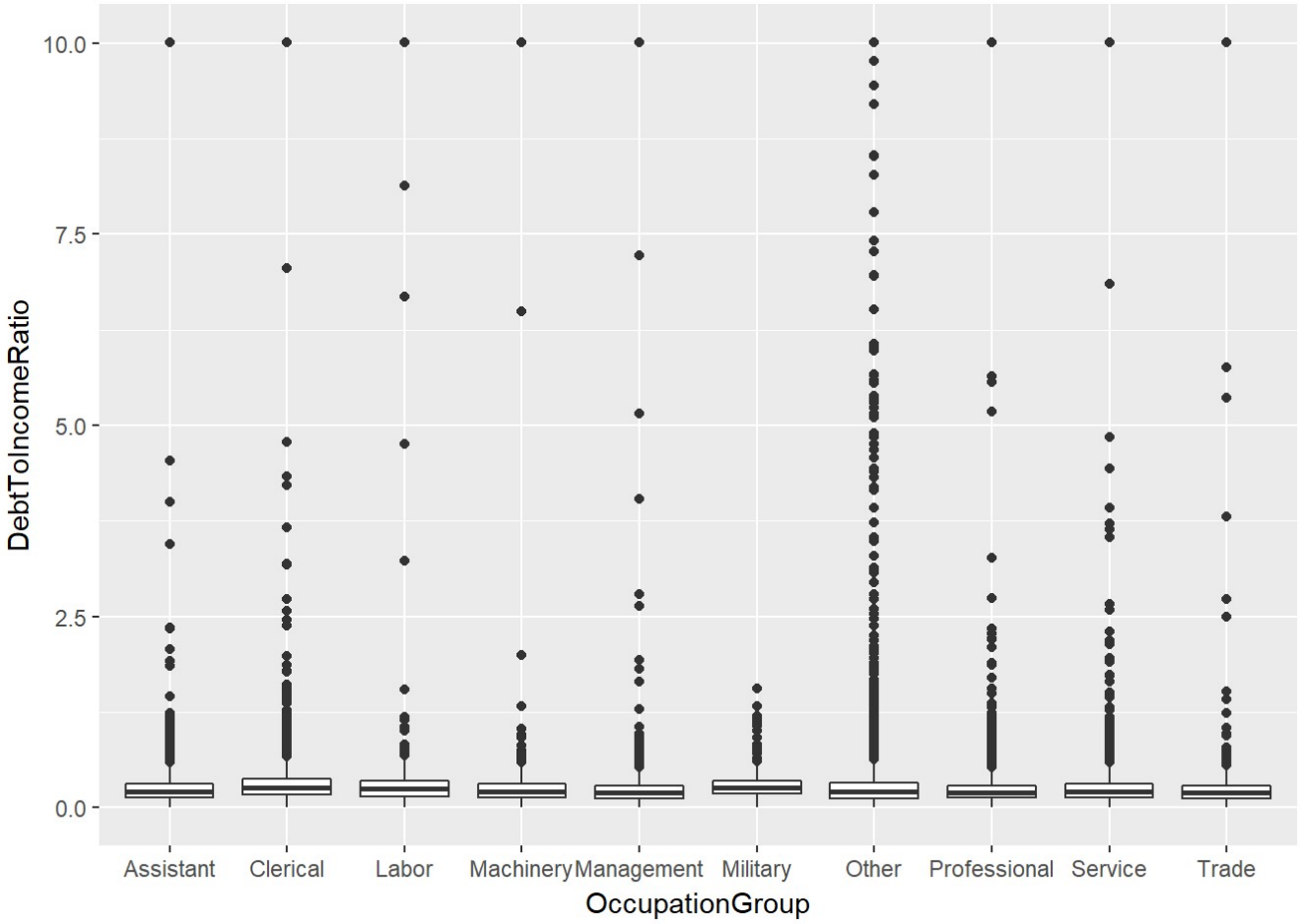

```
## # A tibble: 10 x 4
##    OccupationGroup mean_delinquencies median Count
##    <fct>                        <dbl>  <dbl> <int>
##  1 Assistant                    0.553      0  6735
##  2 Clerical                     0.851      0  7479
##  3 Labor                        0.629      0  1831
##  4 Machinery                    0.582      0  1991
##  5 Management                   0.454      0  8152
##  6 Military                     0.499      0  1618
##  7 Other                        0.667      0 33024
##  8 Professional                 0.501      0 34618
##  9 Service                      0.615      0 12303
## 10 Trade                        0.515      0  6186
```

Proportionally we can see that the majority of those who were granted a loan had less than 1 delinquency on their current credit report.



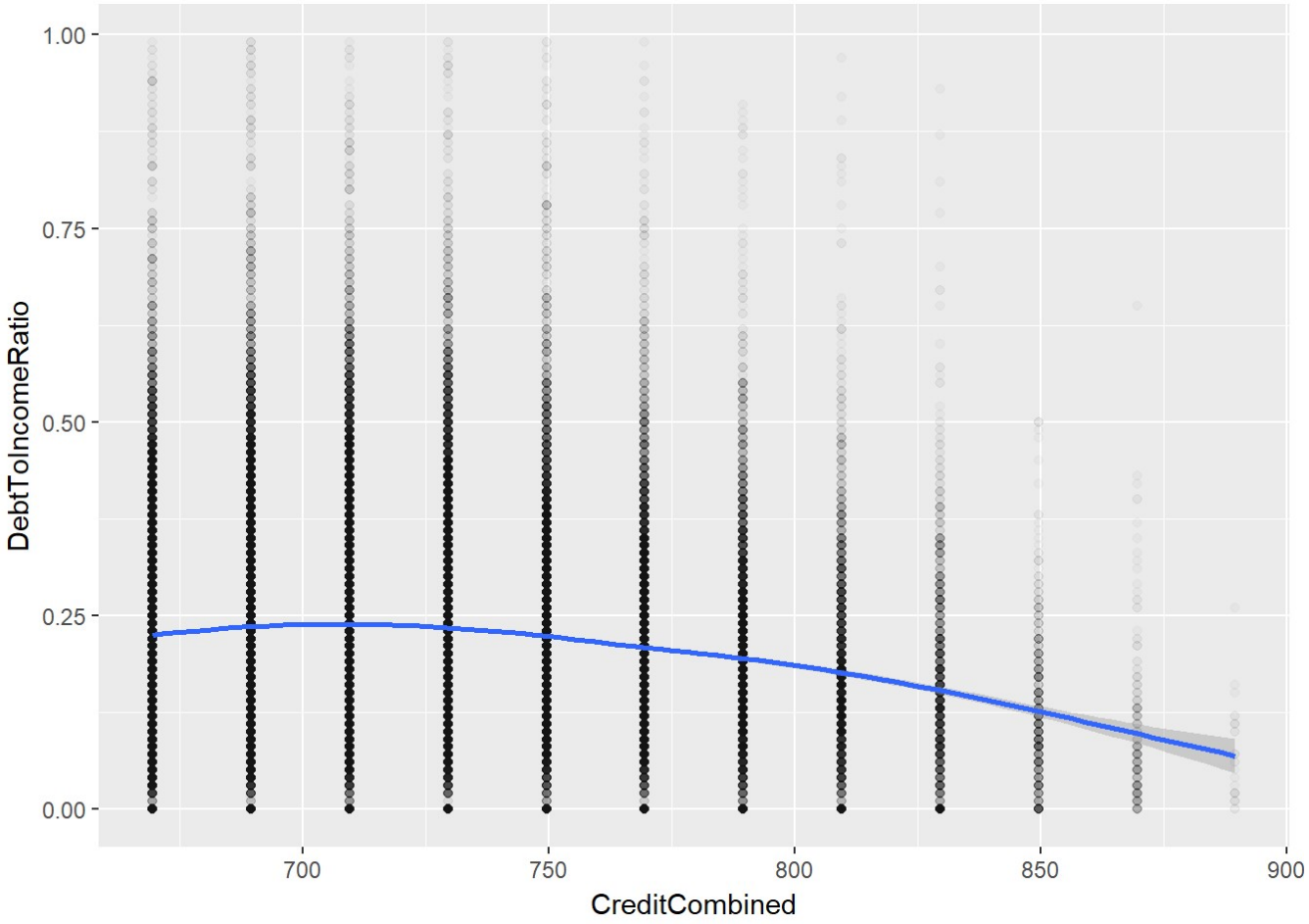

```
## # A tibble: 10 x 4
##    OccupationGroup  mean median Count
##    <fct>           <dbl>  <dbl> <int>
##  1 Assistant        9.93      9  6735
##  2 Clerical         9.15      9  7479
##  3 Labor            8.30      8  1831
##  4 Machinery        8.66      8  1991
##  5 Management      10.7      10  8152
##  6 Military         9.40      9  1618
##  7 Other            8.51      8 33024
##  8 Professional    10.9      10 34618
##  9 Service          9.35      9 12303
## 10 Trade            8.75      8  6186
```

The number of credit lines is fairly consistent across all occupations. It appears that the ideal number of credit lines is 10 or less.



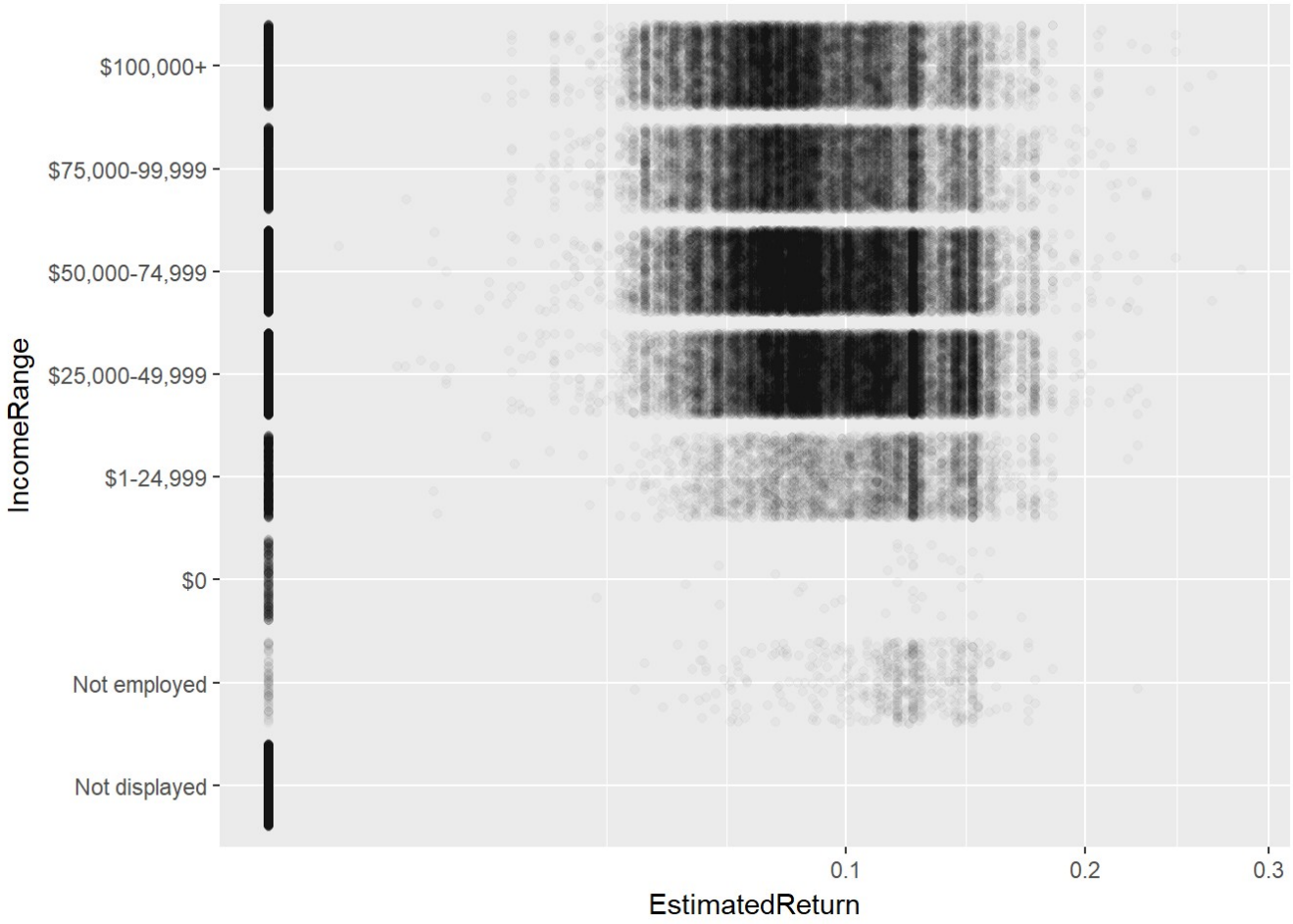

```
## # A tibble: 10 x 4
##    OccupationGroup mean_DTI median_DTI Count
##    <fct>              <dbl>      <dbl> <int>
##  1 Assistant          0.236       0.21  5304
##  2 Clerical           0.316       0.26  5061
##  3 Labor              0.275       0.24  1310
##  4 Machinery          0.262       0.21  1506
##  5 Management         0.225       0.2   6537
##  6 Military           0.273       0.26  1138
##  7 Other              0.282       0.21 23994
##  8 Professional       0.224       0.2  27850
##  9 Service            0.251       0.21  9096
## 10 Trade              0.229       0.2   4812
```

While there are quite a few outliers here as well, the average Debt to Income Ratio by occupation is less than 30%.




```
## # A tibble: 3 x 2
##   Combined.range mean_DTI
##   <fct>             <dbl>
## 1 600-699           0.256
## 2 700-799           0.255
## 3 800+              0.188
```

Also we speculate that the debt to income ratio should be no more than 30 percent in order to qualify for a Prosper loan. Although it appears the preferred range is much lower at approximately 25%


$100,000+
$75,000-99,999
$50,000-74,999
$25,000-49,999
$1-24,999
$0
Not employed
Not displayed
IncomeRange
0.1
0.2
0.3
EstimatedReturn

```
## OccupationGroup: Assistant
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.13040  0.05879  0.08172  0.07746  0.10740  0.24680
## ------------------------------------------------------
## OccupationGroup: Clerical
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.04610  0.06970  0.09060  0.08725  0.11660  0.22640
## ------------------------------------------------------
## OccupationGroup: Labor
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.10490  0.07060  0.08760  0.08539  0.11348  0.17610
## ------------------------------------------------------
## OccupationGroup: Machinery
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.08660  0.06943  0.08524  0.08439  0.11070  0.21580
## ------------------------------------------------------
## OccupationGroup: Management
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.01890  0.06081  0.08154  0.07896  0.10620  0.26670
## ------------------------------------------------------
## OccupationGroup: Military
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.06480  0.04825  0.08421  0.07601  0.11450  0.22160
## ------------------------------------------------------
## OccupationGroup: Other
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.17730  0.06027  0.08360  0.07914  0.11150  0.22650
## ------------------------------------------------------
## OccupationGroup: Professional
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.18270  0.05780  0.07923  0.07689  0.10530  0.28370
## ------------------------------------------------------
## OccupationGroup: Service
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.18160  0.06310  0.08370  0.08092  0.11116  0.23130
## ------------------------------------------------------
## OccupationGroup: Trade
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.02510  0.06499  0.08287  0.08066  0.10790  0.19480
```

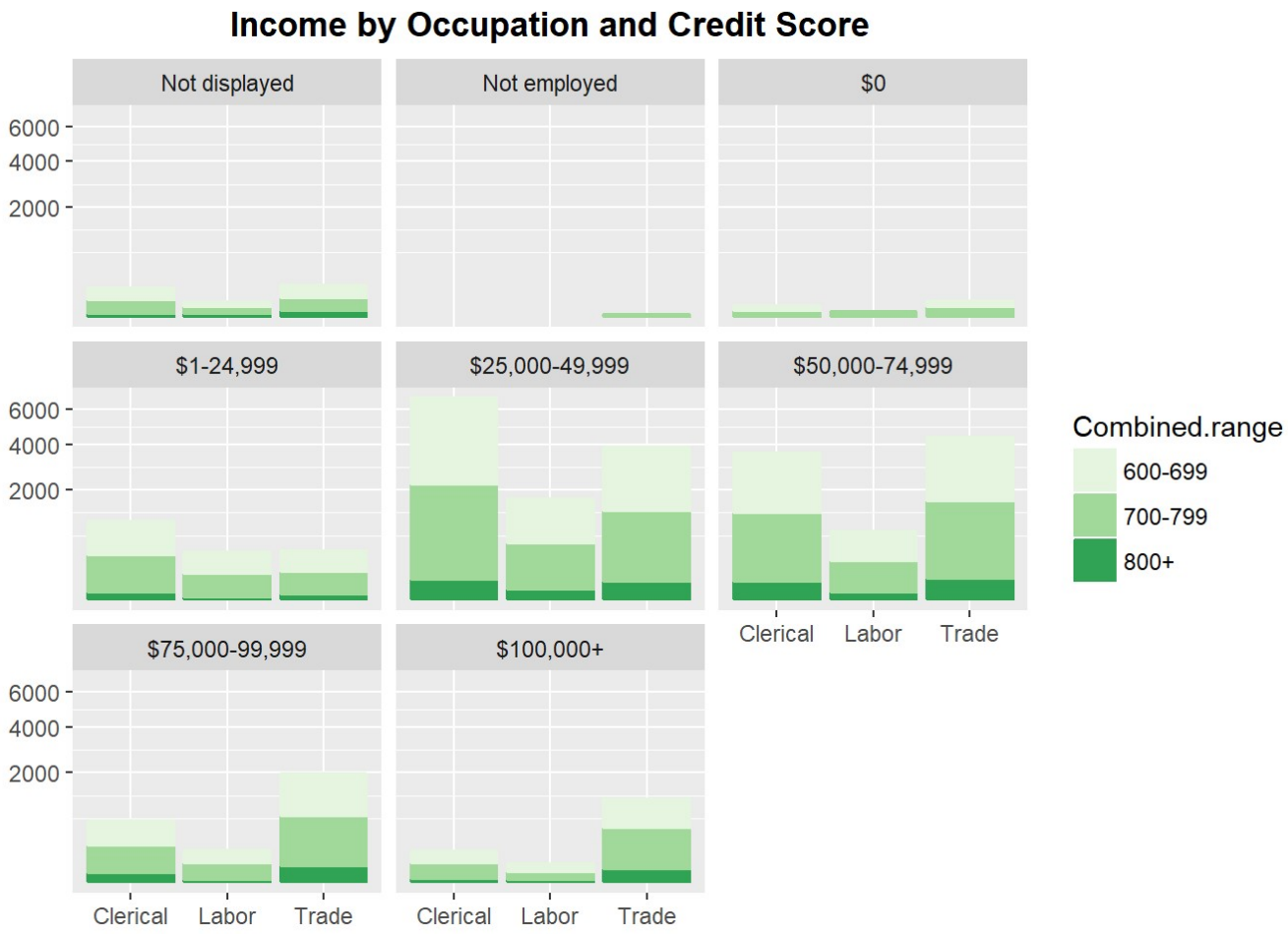
Clerical, labor and trade are the highest return rate among the occupational groups indicating that they are considered more of a risk than other groups.

```
## # A tibble: 8 x 3
##   IncomeRange      meanR  meanY
##   <fct>            <dbl>  <dbl>
## 1 Not displayed  0       0
## 2 Not employed   0.104   0.200
## 3 $0             0.00913 0.0162
## 4 $1-24,999      0.0847  0.156
## 5 $25,000-49,999 0.0847  0.151
## 6 $50,000-74,999 0.0816  0.140
## 7 $75,000-99,999 0.0800  0.135
## 8 $100,000+      0.0772  0.128
```

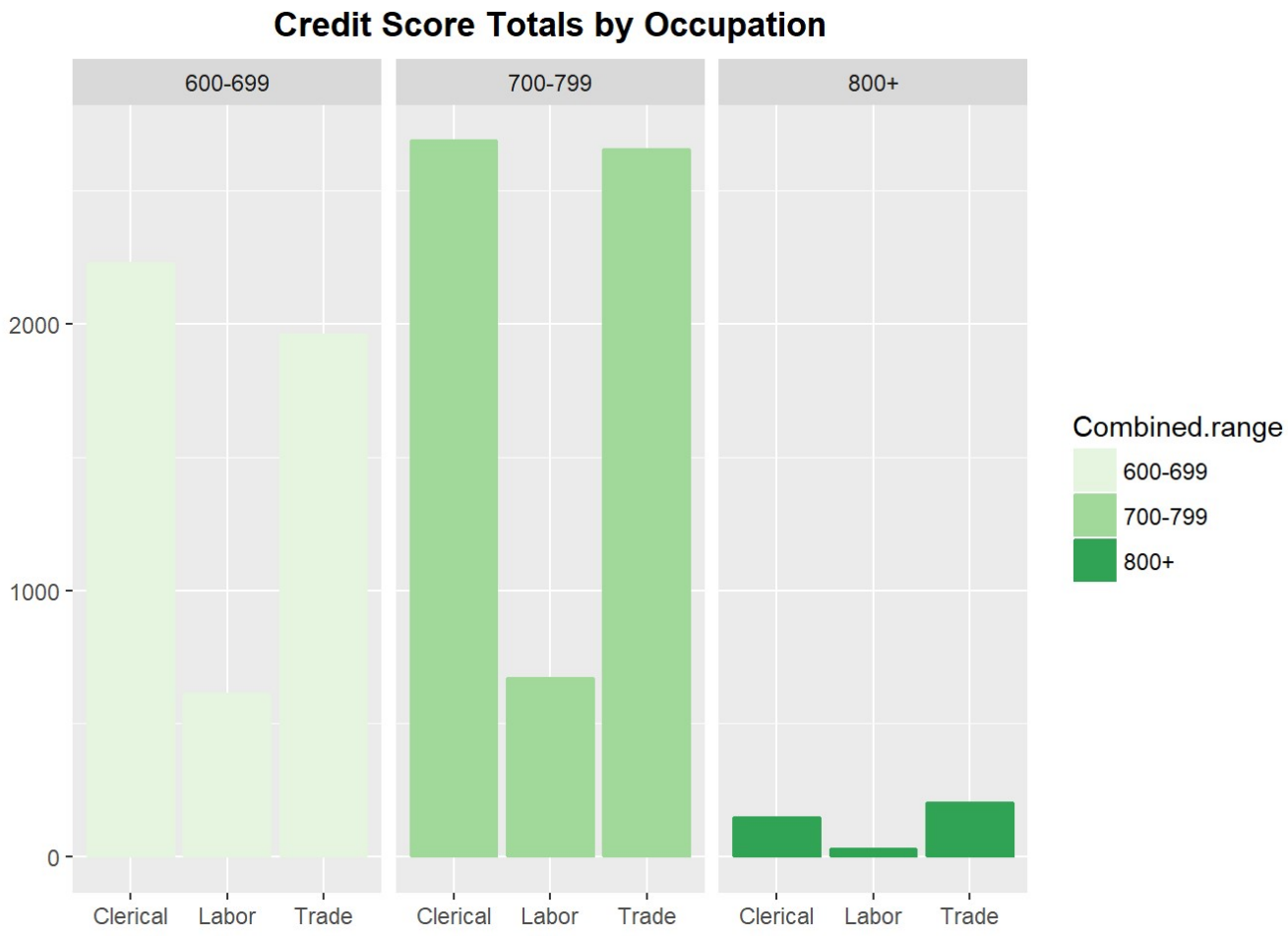
While there is no direct relationship between Income and Estimated Return, the range of estimated return appears in a risk versus reward fashion. Meaning, the lower the income the higher the estimated return. This is not particularly surprising because the data definition states that Estimated Return is in part based off of Estimated Effective Yield.

```
## # A tibble: 10 x 4
##    OccupationGroup  meanR meanEEY meanLY
##    <fct>            <dbl>   <dbl>  <dbl>
##  1 Assistant       0.0775   0.134  0.167
##  2 Clerical        0.0872   0.155  0.187
##  3 Labor           0.0854   0.156  0.186
##  4 Machinery       0.0844   0.151  0.177
##  5 Management      0.0790   0.135  0.163
##  6 Military        0.0760   0.133  0.176
##  7 Other           0.0791   0.140  0.175
##  8 Professional    0.0769   0.129  0.159
##  9 Service         0.0809   0.142  0.176
## 10 Trade           0.0807   0.142  0.171
```

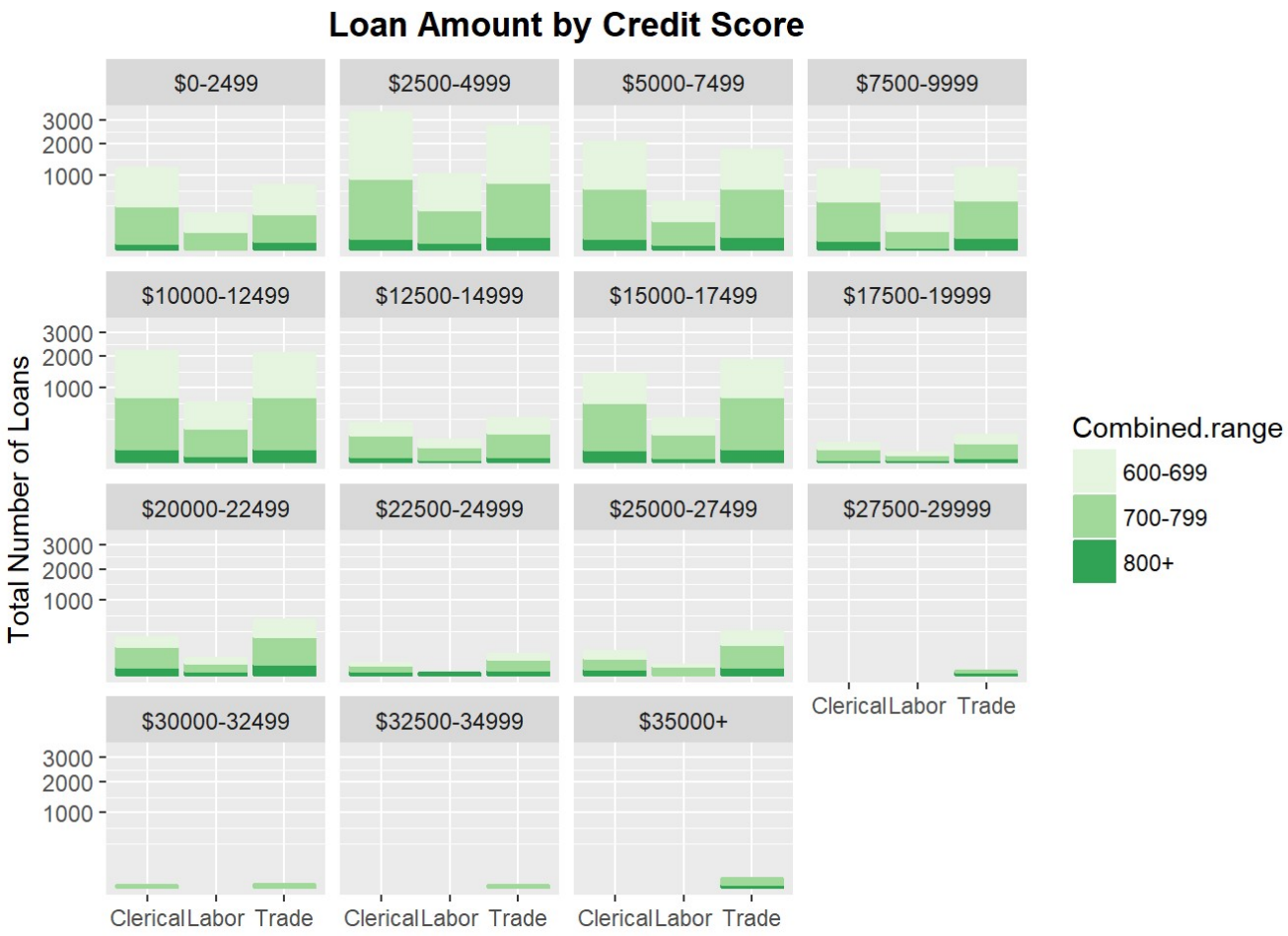
Again we see that clerical, labor and machinery have the highest estimated yield of all groups. This is worth investigating further.



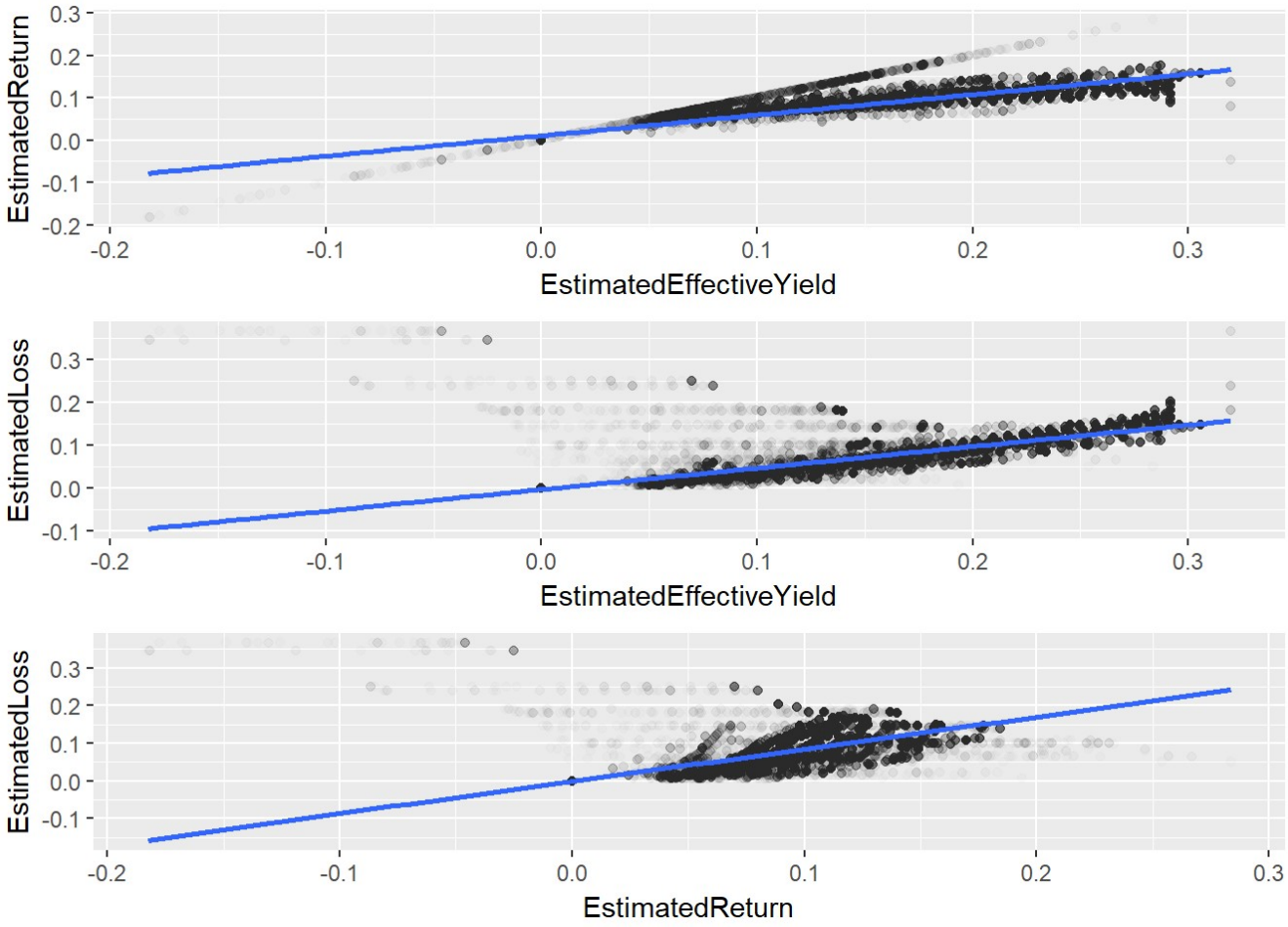
Income doesn't appear to be a factor in why these groups have lower loan amounts.



This may be the answer! I'm surprised that their credit scores are so low considering the income distribution is relative to other occupations.



While not completely definitive, it does indeed show that credit score is a factor for these groups lower loan amounts.



The relationship between estimated yield, loss and return is easy to see here. Prosper uses this to determine risk and reward for investors.

Bivariate Analysis

How did the feature(s) of interest vary with other features?

Occupation by EstimatedEffectiveYield varied significantly across three specific groups: Clerical, Labor, and Trade. These three groups had the highest estimated yield and estimated return of all other occupations. They also had the lowest loan amounts. When investigating further it was discovered they had proportionally lower credit scores and this was why those values were higher.

Did you observe any interesting relationships between the other features?

The relationship between EstimatedEffectiveYield, EstimatedLoss and EstimatedReturn were interesting in how closely related they are.

Due to the scope of interest the supporting variables relationship with the main features were of interest. The relationship between the combined credit score and debt to income ratio was interesting. The higher the credit score the lower the DTI. It's interesting because it was unexpected. The expected relationship is Income to DTI. It was also interesting to see that higher income and credit score were not related.

What was the strongest relationship you found?

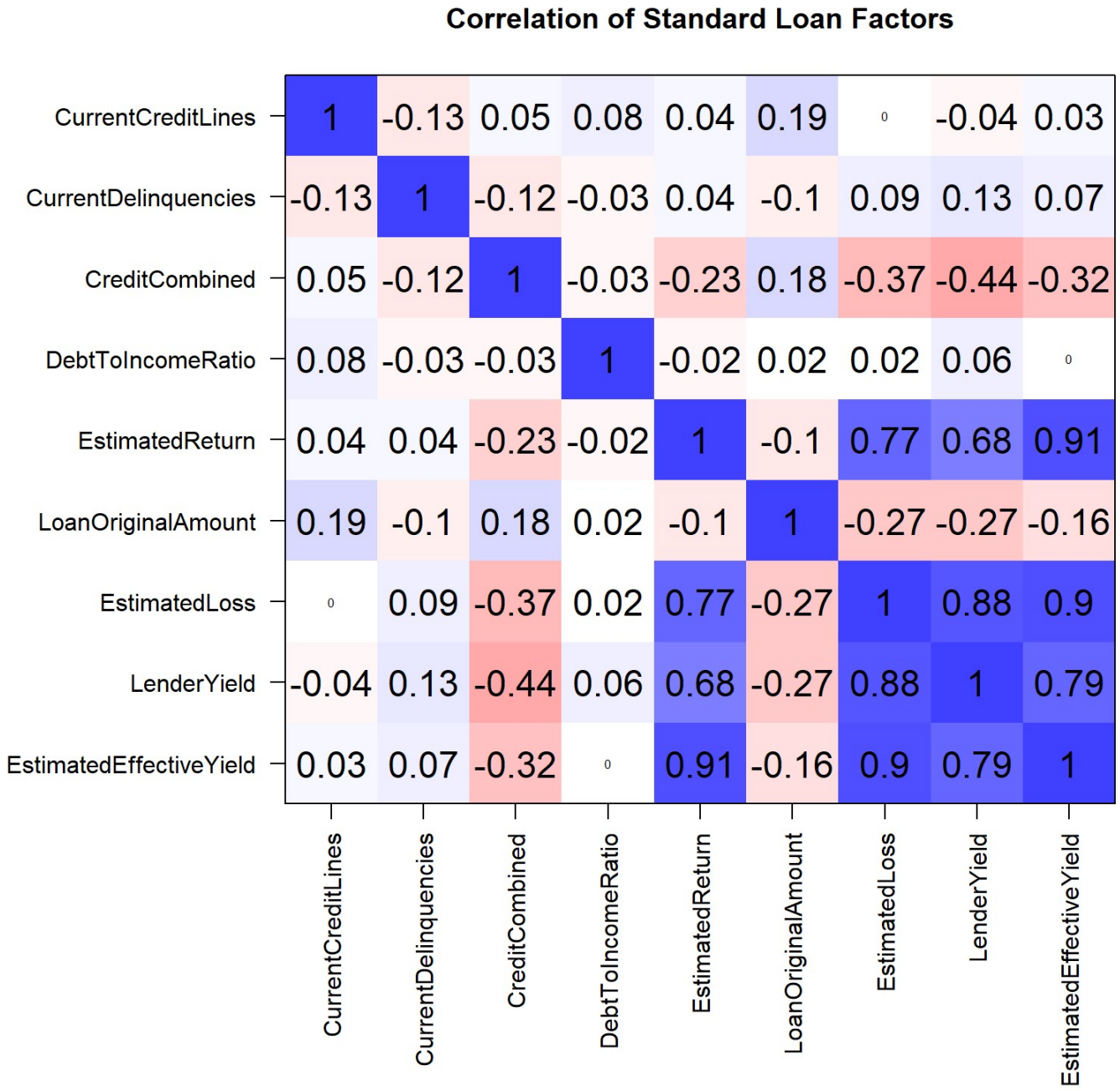
The strongest relationships were between EstimatedEffectiveYield, EstimatedLoss and EstimatedReturn. These values are used to determine how much money investors and Prosper will make over the life of the loan.

The relationship between credit score and the amount of the loan was the strongest of the observations.

Reflections

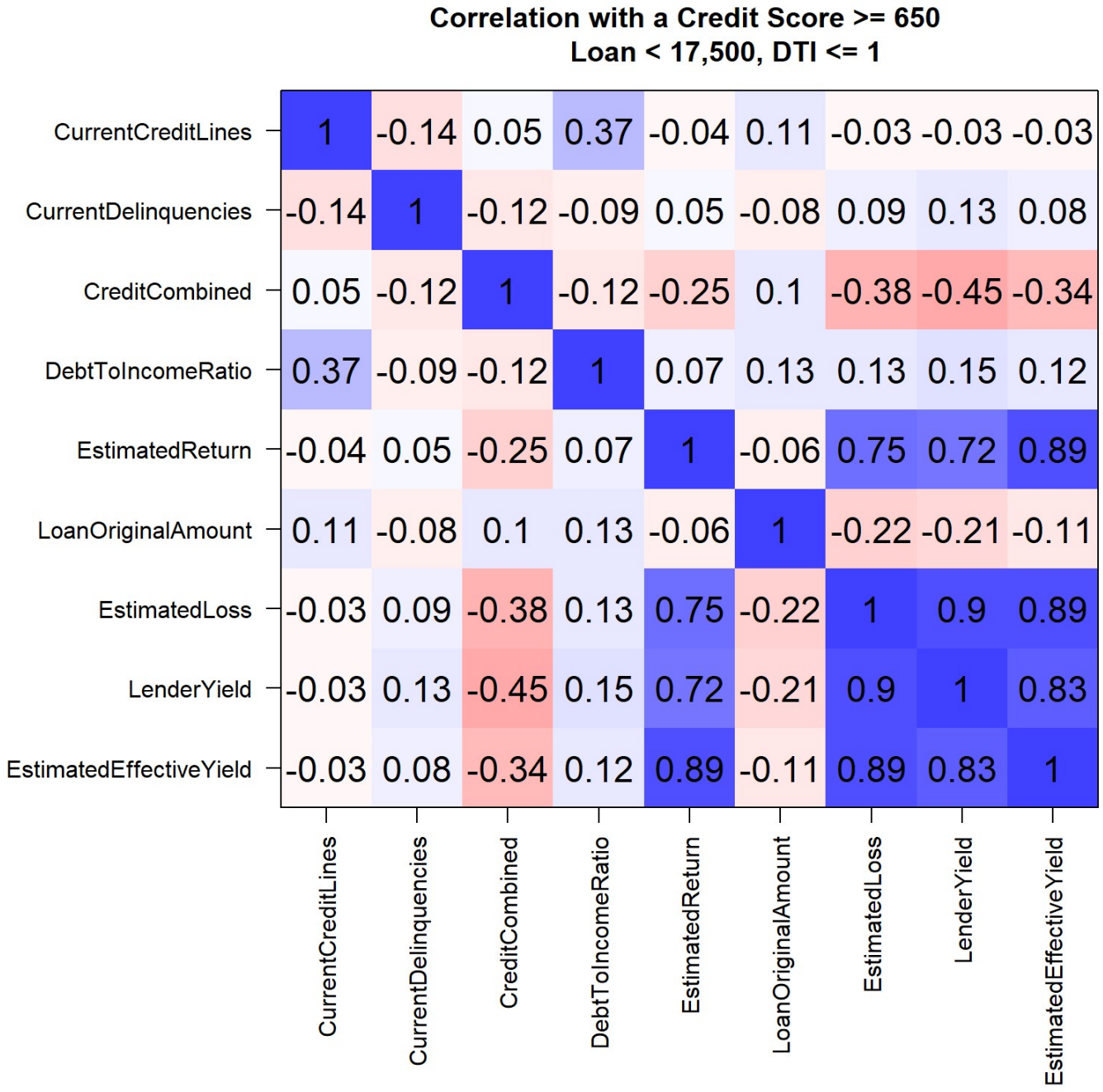
The criteria for receiving a loan with Prosper is now much more defined. The Debt to Income Ratio(DTI) should be no greater than **.30**, Income should be at least **$25,000**, Credit Score should be at least **650**, the category for the loan should be Debt Consolidation and the amount of the loan should be less than **$17,500**. We will try to narrow this down further and see if it holds true with multivariate plots.

Multivariate Plots Section



There are no surprising relationships here. Yield, Loss and Return are all directly related. This is used to

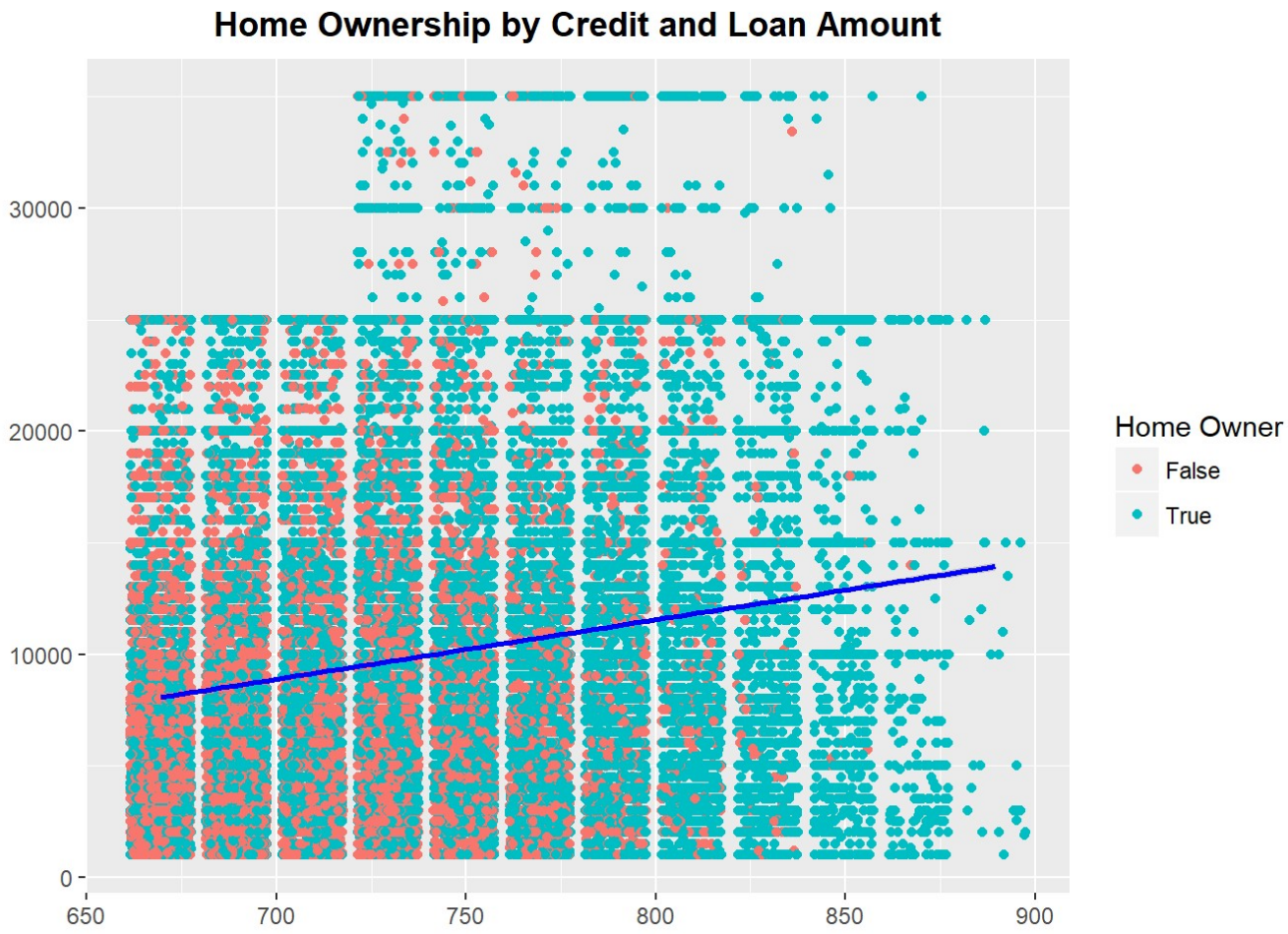
determine APR, Rate and profit for investors. The strong and moderate negative relationships between CreditCombined, Yield and Loss are significant.



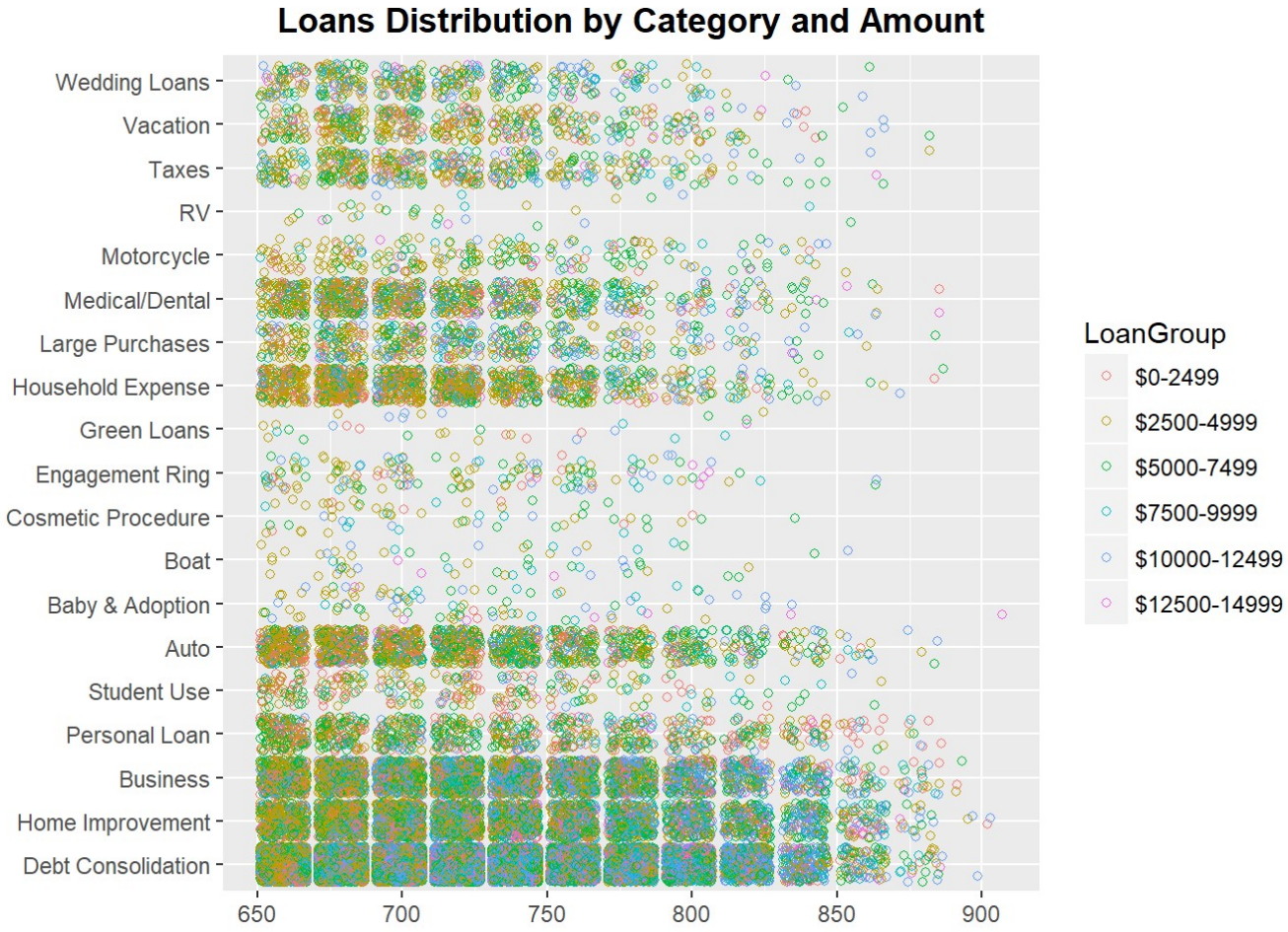
There are subtle differences between both of these plots. There is a difference of 0.01 - 0.27 between the plot without conditions and the plot that has conditions set at a credit score of at least 650, a loan amount less than **$17,500** and where the DTI is less than or equal to **1**.

That change in significance is very predictable because the largest variance in significance is between CurrentCreditLines and DebtToIncomeRatio. These two variables have a very limited scope of significance in that as CurrentCreditLines rises the significance of DTI falls. DTI shows the most significance when it is 1 or lower. It will show less significance as the value falls and equally so at a the value greater than 1. Putting it simply, the more Lines of credit you have the more debt you obtain. When you reach a debt of greater than your income the number no longer has significant value in regard to the number of credit lines.

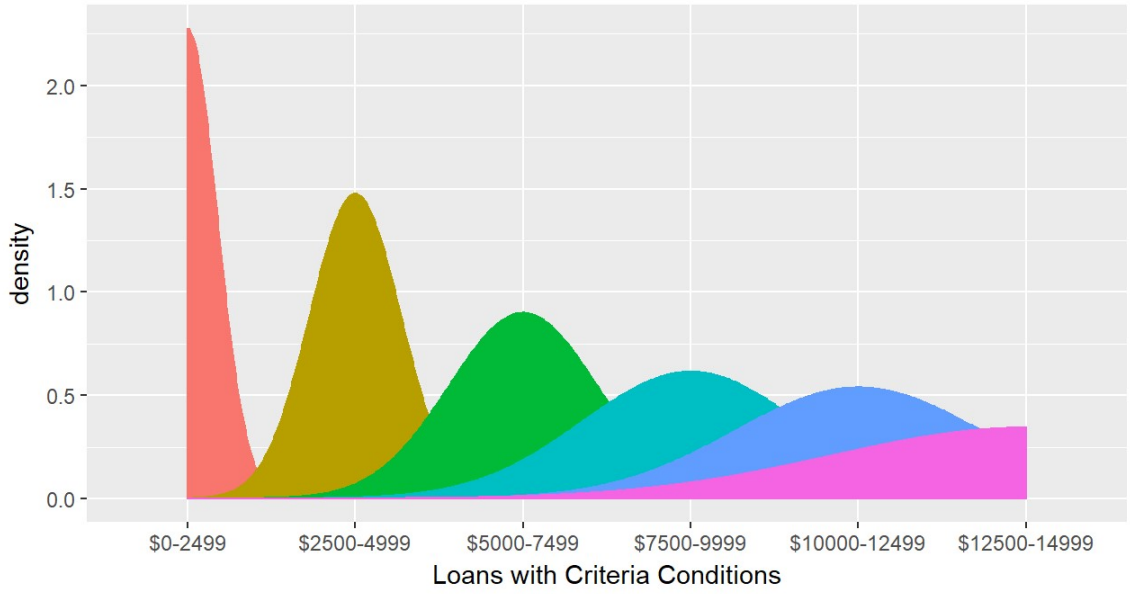
Loan requestors with a higher income tend to own homes and have higher credit score than those with lower incomes and credit scores.

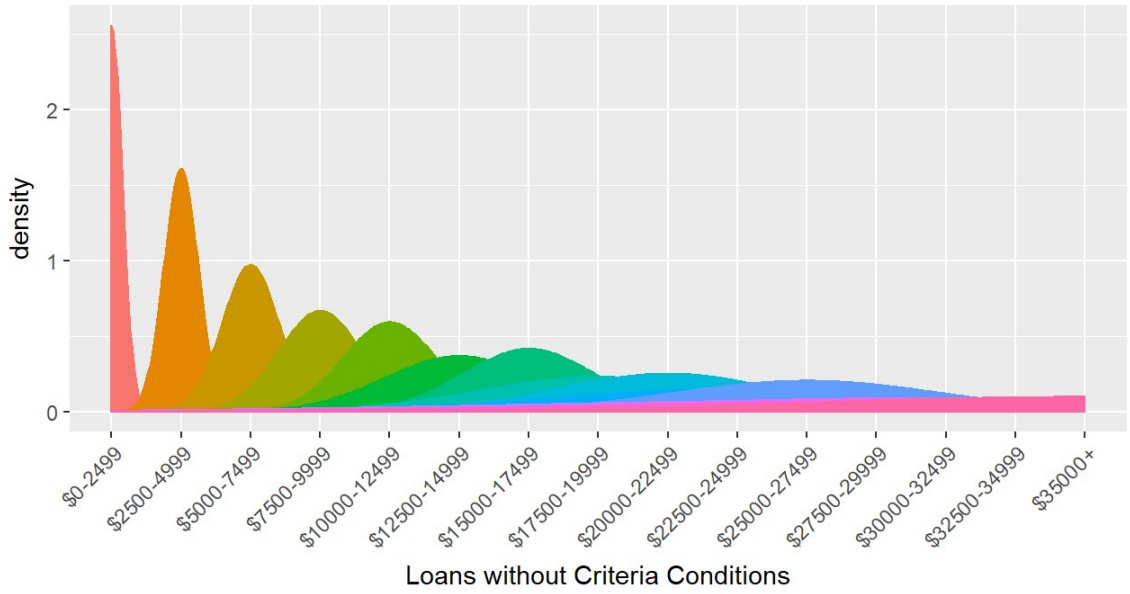


A common theme as shown above, is the relationship between loan amount and credit score.



Because of the non-linear relationship between our loan requestor and Prosper loan approval criteria we take an approach of narrowing the scope according to what we have learned about the requestor. Here we see the dispersion of loan for those whose credit score is greater than **650**, debt to income ratio is less than **.30**, income is greater than **$25,000** and loan request is less than **$15,000**.
Across distinctlyvdefined loan categories we see that the primary loan requests are for Business, Debt Consolidation, or Home Improvement.





The criteria set to restrict the loan amount to less than **$15,000** is very solid here. We see that once the loan amount reaches **$17,500** the plot starts to flatten out rapidly.

Multivariate Analysis

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

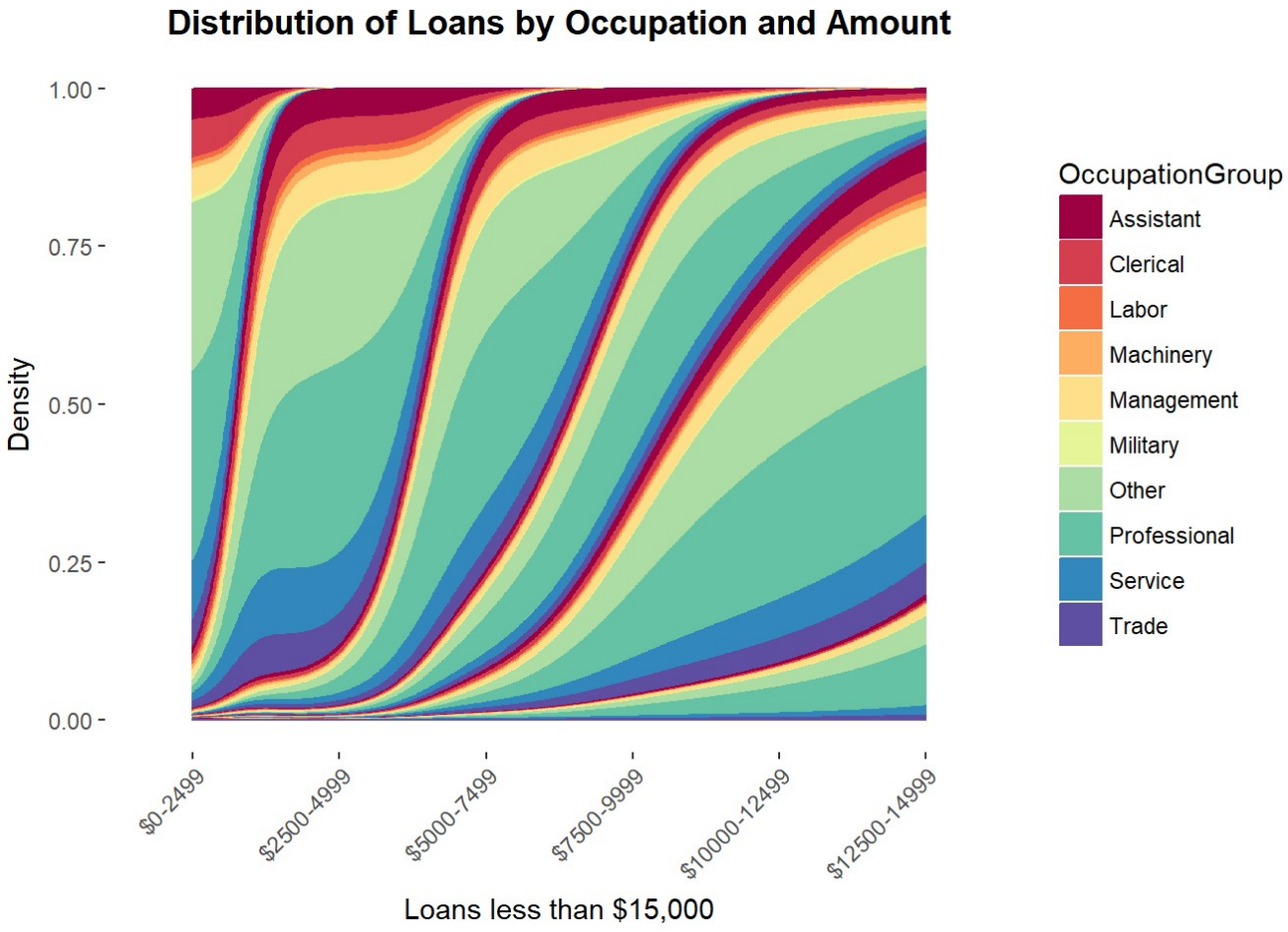
When the data was restricted to conditions proposed to be successful in applying for a loan with Prosper the results of the shift across all observations becomes much clearer. Furthermore, when we restrict the conditions closer to what is ideal, it is then that we can see a much more significant change in the data.

Were there any interesting or surprising interactions between features?

The relationship between DTI and Current credit lines was really interesting. When setting a conditional scope to observe correlation changes at first I was surprised to see such a dramatic change. I hadn't truly considered the impact on the relationship between the two variables.

Final Plots and Summary

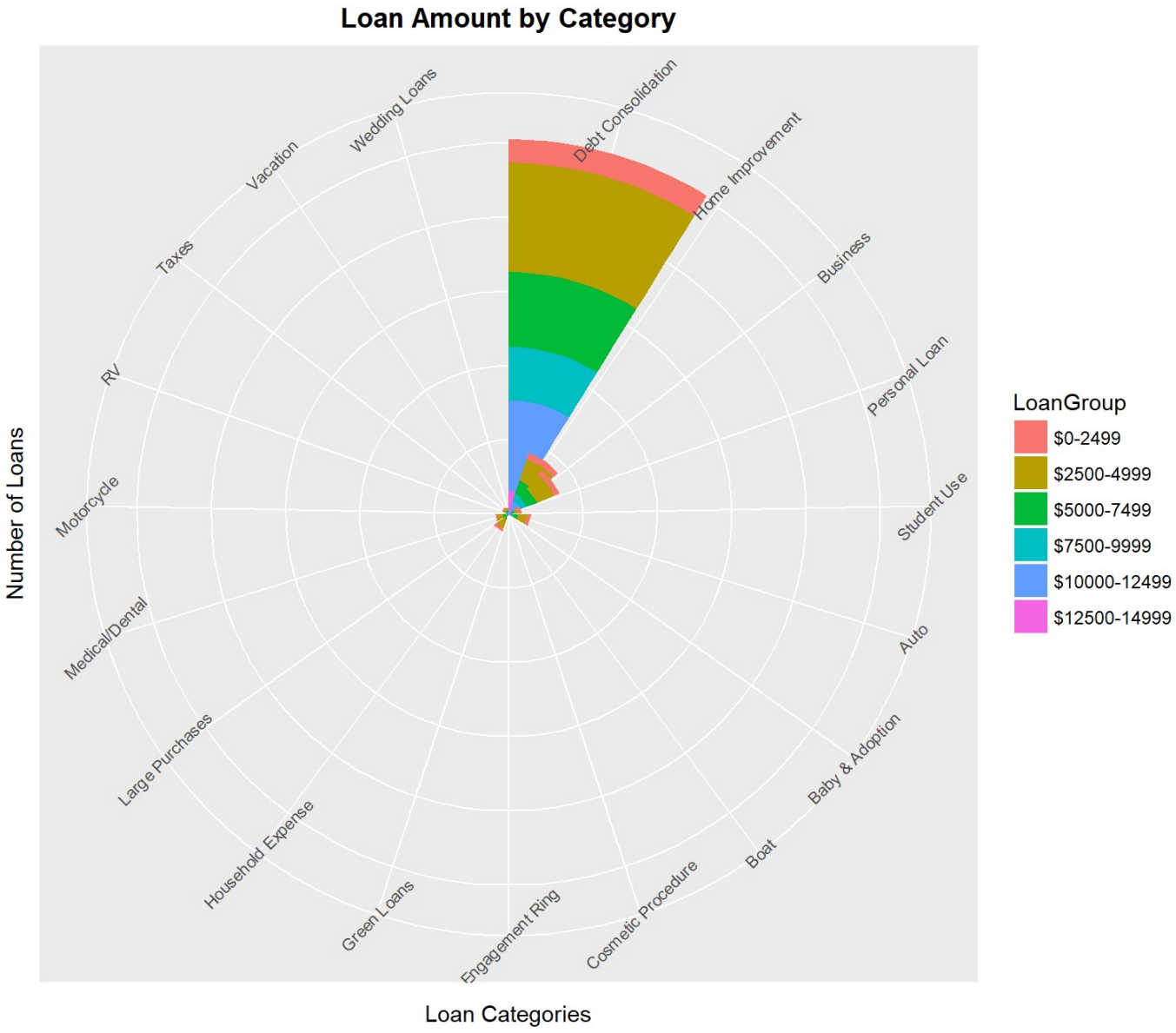
Plot One



Description One

I chose this plot because it answers the questions of who(loan requestor) and what(loan amount). It is very easy to see the relationship between occupations and loan amounts. All occupations are clearly represented which agrees with the assertion that occupation is not a factor in qualifying for a loan with Prosper. The division of volume by loan amount is clear as well. We can see as the loan amount increases the density drops off rapidly which agrees with the assertion that to be successful as a loan requestor the loan should be less than $15,000.

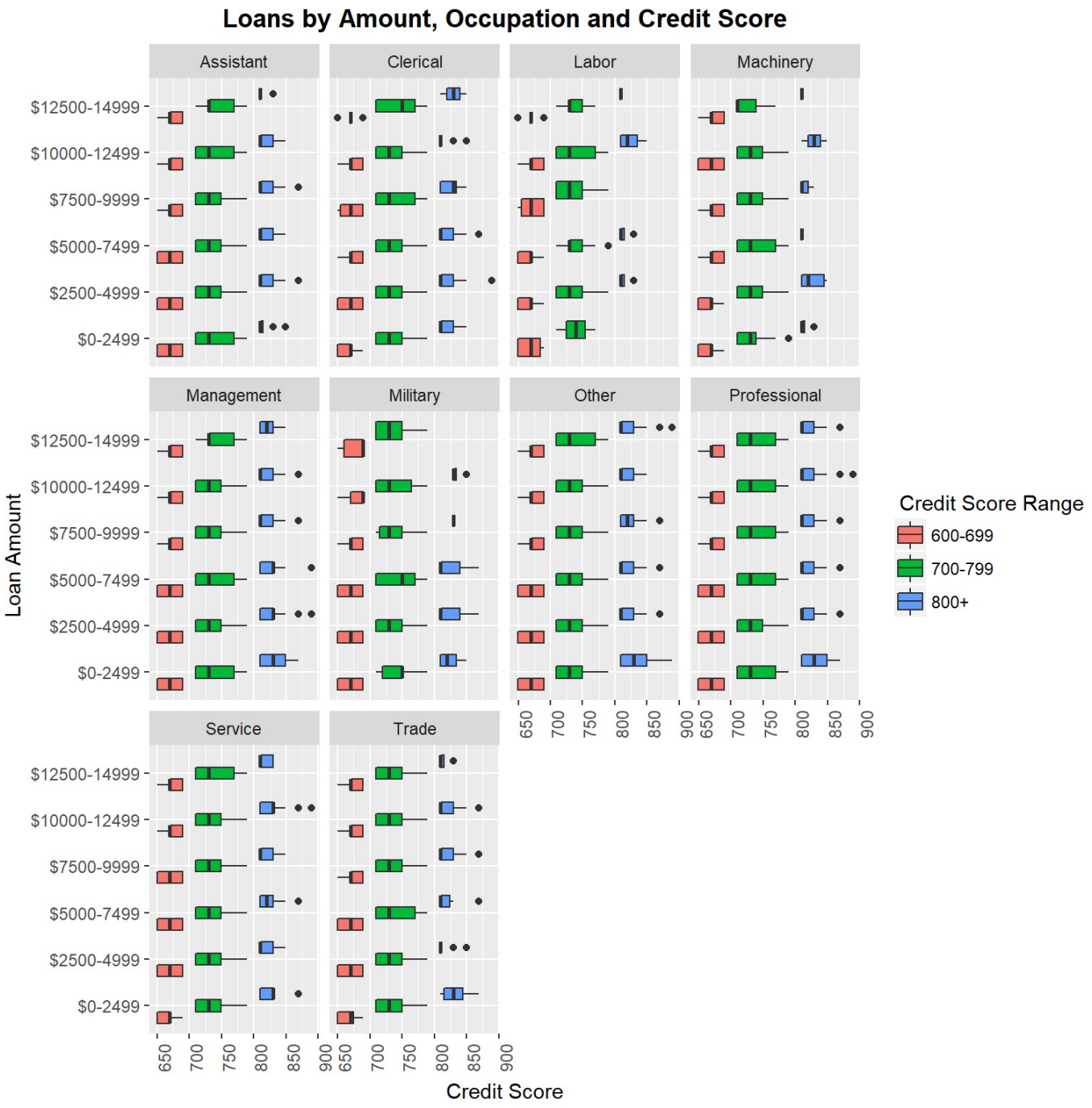
Plot Two



Description Two

This plot was chosen to answer the questions of what(loan amount), and why(loan category). We know the answer to the who of Prosper loans this plot shows that we will be more successful in the loan application process if the loan is for debt consolidation. However, if the loan is for home improvement or business and the amount requested is less than $7,500 we will be much more likely to be approved than for other loan categories.

Plot Three



Description Three

The final plot was chosen because it shows the how(criteria) of the Prosper loan requestor. A credit score of at least **650** and loan amounts lower than **$15,000**, DTI of less than **.30** and an income of at least **$25,000** are all represented here. While some occupations like Labor may have a harder time of qualifying for higher loan amounts with lower credit scores we can see that all the higher the credit score the more successful the request is across all occupations.

Reflection

The goal of this study was to answer the questions of who, what, why and how Prosper grants loans. The dataset used holds 113,937 entries across 81 variables. The total number of variables used across this study

both original and created totaled 26.

The choice of those specific variables was made to give the most meaning to the data by using a more traditional approach to the criteria of the loan process.

From there each was evaluated parallel to the goals of the study.
I was able to determine that while certain occupations have a higher volume of loans, all occupations were represented if they were within the parameters that were defined as preferred.

The final preferred criteria is:

- All Occupations
- Credit Score >= 650
- Current Delinquencies < 1
- Debt to Income Ratio < .30
- Income > $25,000
- Loan Amount < $15,000
- Loan Category of Debt Consolidation, Business or Home Improvement

There are so many more insights to be gained from this data set and I have just scratched the surface. Other things to explore could be how Prosper defines a good risk or how many of Propser's customers are repeat customers. It would also be interesting to take this data and develop a model to help investors determine the best loans to back and what loans to avoid.

Sources

- https://en.wikipedia.org/wiki/International_Standard_Classification_of_Occupations (https://en.wikipedia.org/wiki/International_Standard_Classification_of_Occupations)
- https://www.prosper.com/plp/loans/loan-types/ (https://www.prosper.com/plp/loans/loan-types/)
- http://r.789695.n4.nabble.com/Apostrophes-in-R-Commander-in-recode-td3704453.html (http://r.789695.n4.nabble.com/Apostrophes-in-R-Commander-in-recode-td3704453.html)
- https://www.stat.berkeley.edu/classes/s133/factors.html (https://www.stat.berkeley.edu/classes/s133/factors.html)
- https://www.cnbc.com/quotes/?symbol (https://www.cnbc.com/quotes/?symbol)=.SPX