UDAND PROJECT - Explore and Summarize Data

#DATASET - Prosper Loan Data

Introduction and Hypothesis

Prosper issues loans to individuals, which are described as "fractional loans", since multiple investors can fund one loan, which minimizes risk due to diversification. The most important question to investors (as always) is what constitutes a good loan to invest in? The null hypothesis can thus be generally summarized as the following:

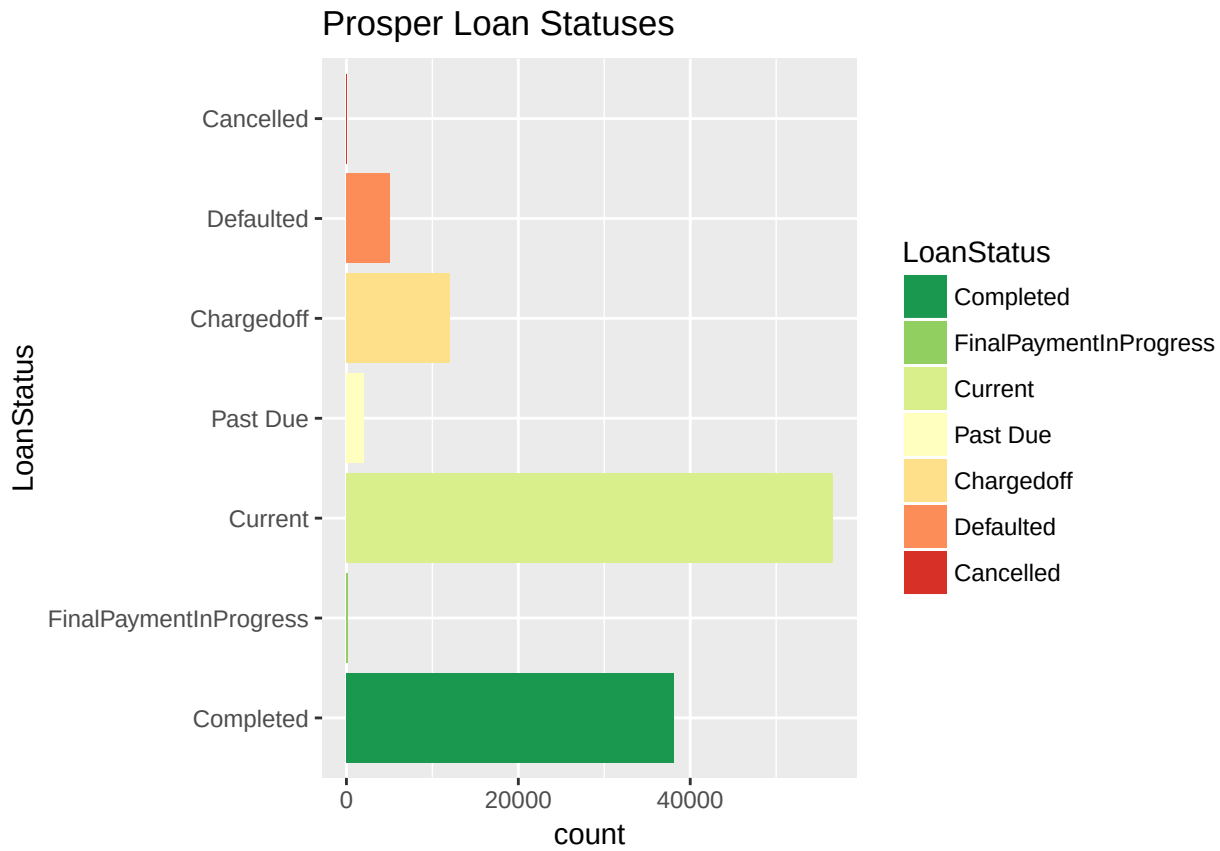
"*There is no relationship between a variable and a borrower's ability to pay off a loan.*"

This report will endeavour to detail and explain the most relevant variables that *do* have an effect on a borrower's ability to pay off a loan (i.e. the one investors should focus on).

For ease of use, we will create a new dataframe with just those relevant variables, and remove all other variables not investigated in this report.

Univariate Plots Section

Loan Status (LS)




```
##               Completed FinalPaymentInProgress                 Current
```

```
##                38074                  205                56576
##            Past Due            Chargedoff            Defaulted
##                2067                11992                 5018
##            Cancelled
##                    5
```

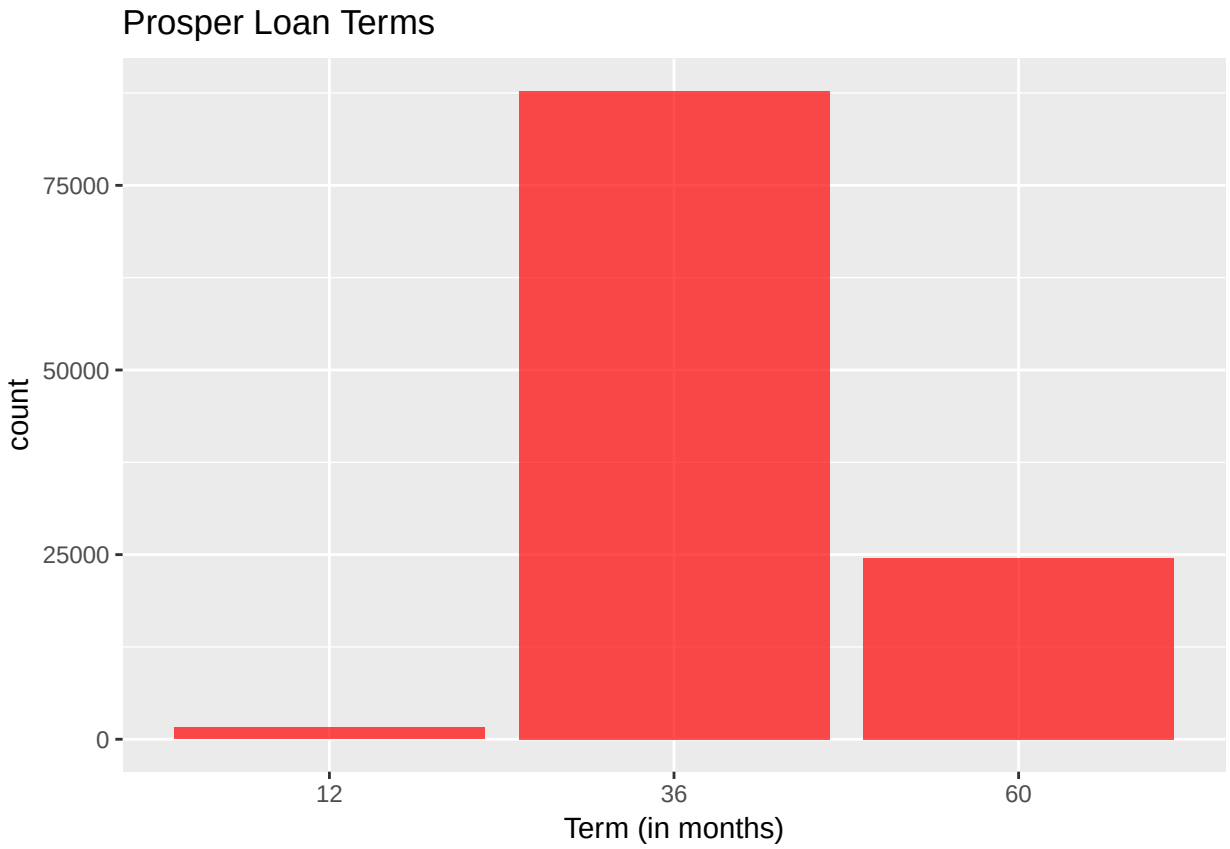
The most important variable, this variable directly states whether a borrower has been repaying a loan.

- **Current** loans have had all payments have been made, at least so far.
- **Completed** loans have been successfully paid off.
- **FinalPaymentInProgress** loans are almost completed.
- **Past Due** loans means one or more payments have not been made on the loan. The original data split Past Due by duration, but since there are a low amount of loans in this status, we will combine them for the purposes of this report.
- **ChargedOff** loans have missed four payments (i.e. loan payments not made for more than 120 days). These loans are now due in full, and the remaining balance is assumed to be lost.
- **Defaulted** loans are where the borrower is for certain reasons unable to repay the loan, due to delinquency, bankruptcy, death, etc.
- **Cancelled** loans seem self-explanatory, but the Prosper website does not mention a "Cancelled" Loan Status. I have assumed the loan application was made, but the borrower changed their mind before the loan was fully approved. Since there are only 5 occurrences, it does not make sense to specify a whole category for such a small group. *Thus, Cancelled loans will be ignored for the purposes of this project.*

To simplify terminology in this report, Complete, FinalPaymentInProgress, and Current loans will be referred to as "Good" loans, while Past Due, Chargedoff, and Defaulted loans will be referred to as "Bad" loans.

The amount of Good loans compared to Bad loans will be referred to as the Good-Bad Loan Ratio (GBR)

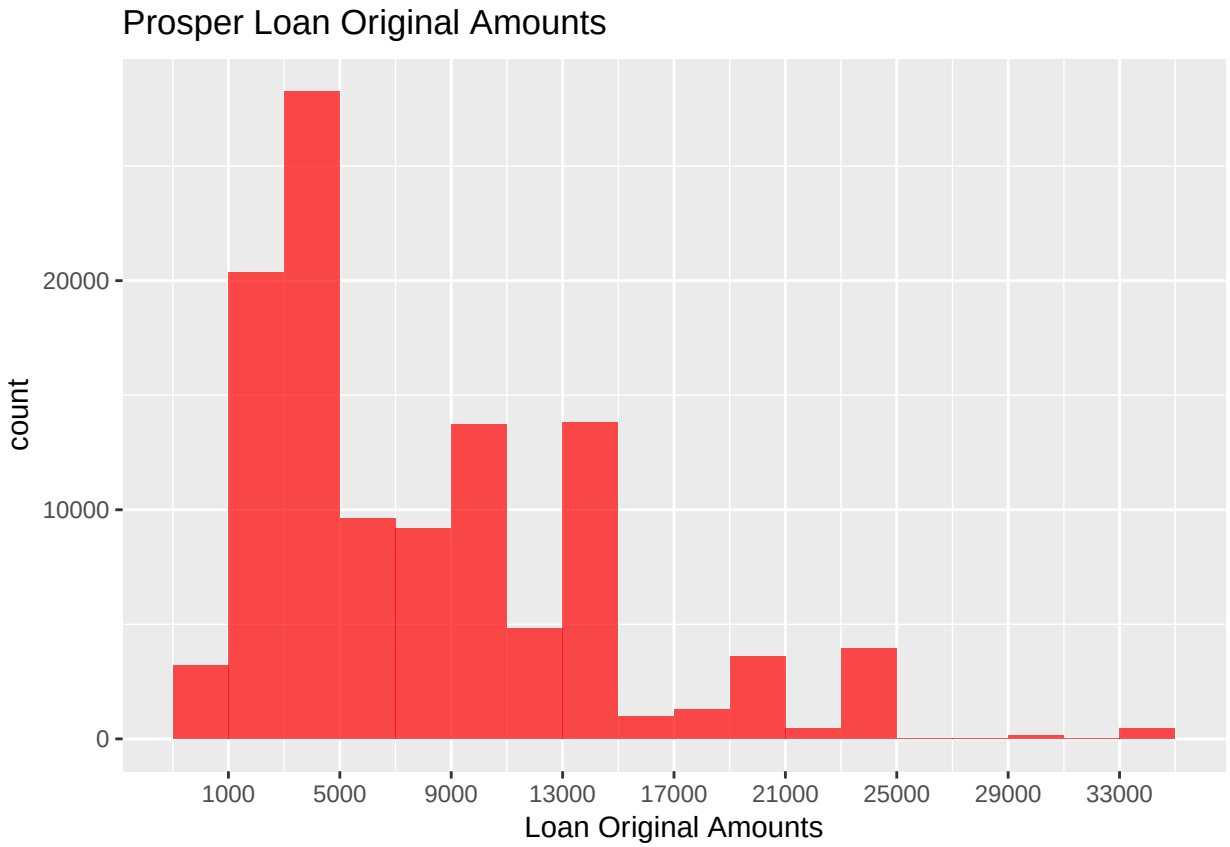
Term (T)




```
##    12    36    60
##  1614 87773 24545
```

This graph shows the length of time loans would be paid over. These are all clearly short term loans, with an overwhelming majority having a 36 month term, followed by 60 and 12 month terms.

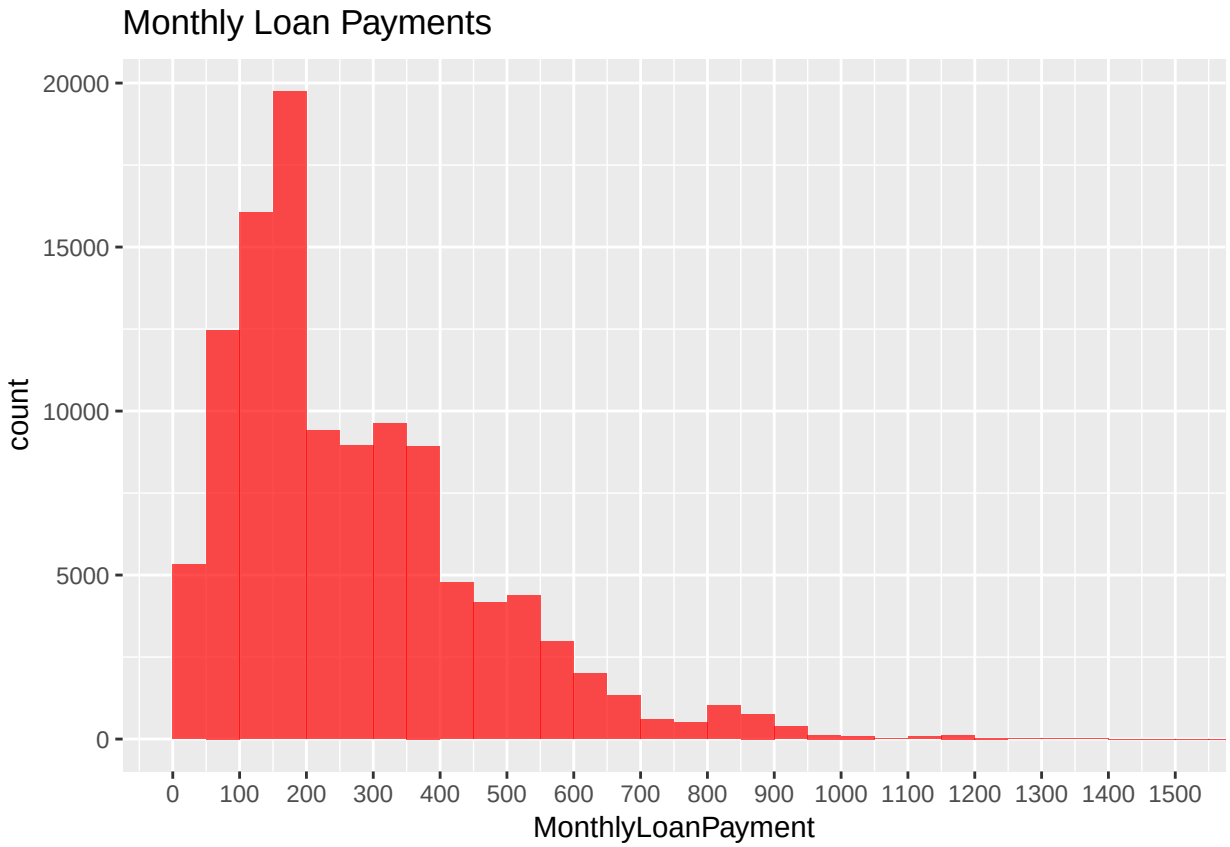
Loan Original Amount (LOA)




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

Here, we can see that the Loan original amounts are mostly ~$5k, but there are peaks at $10k and $15k as well, with smaller peaks at $20k and $25k. A higher LOA could mean a higher GBR.

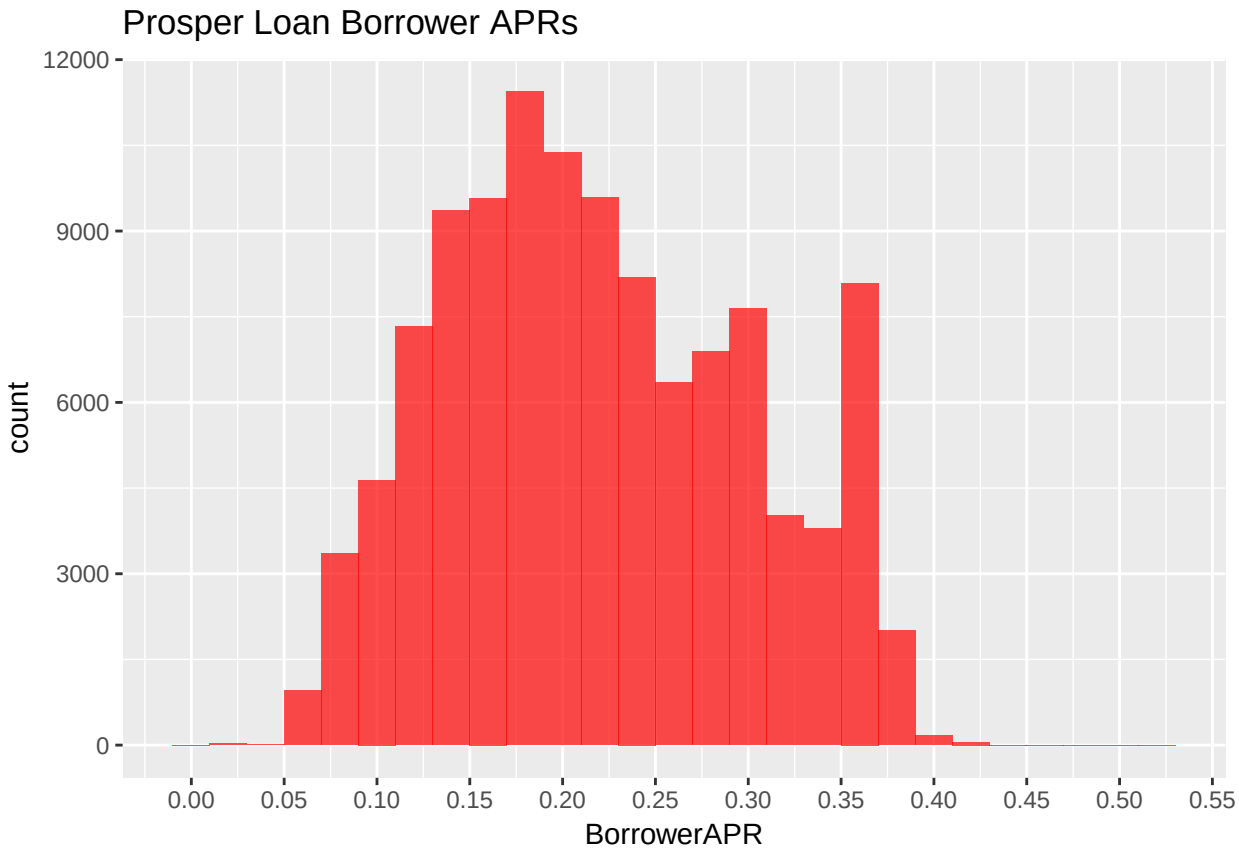
MonthlyLoanPayment (MLP)




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   131.6   217.7   272.5   371.6  2251.5
```

It could be initially argued that higher loan amounts are harder to pay off, and thus are more likely to not be paid. This graph is positively skewed, with a huge majority of loan payments between $100-200 per month. We can also see a steady amount of monthly loan payment between $200-$400. There are very few loans with $1k+ monthly payments.

BorrowerAPR (APR)

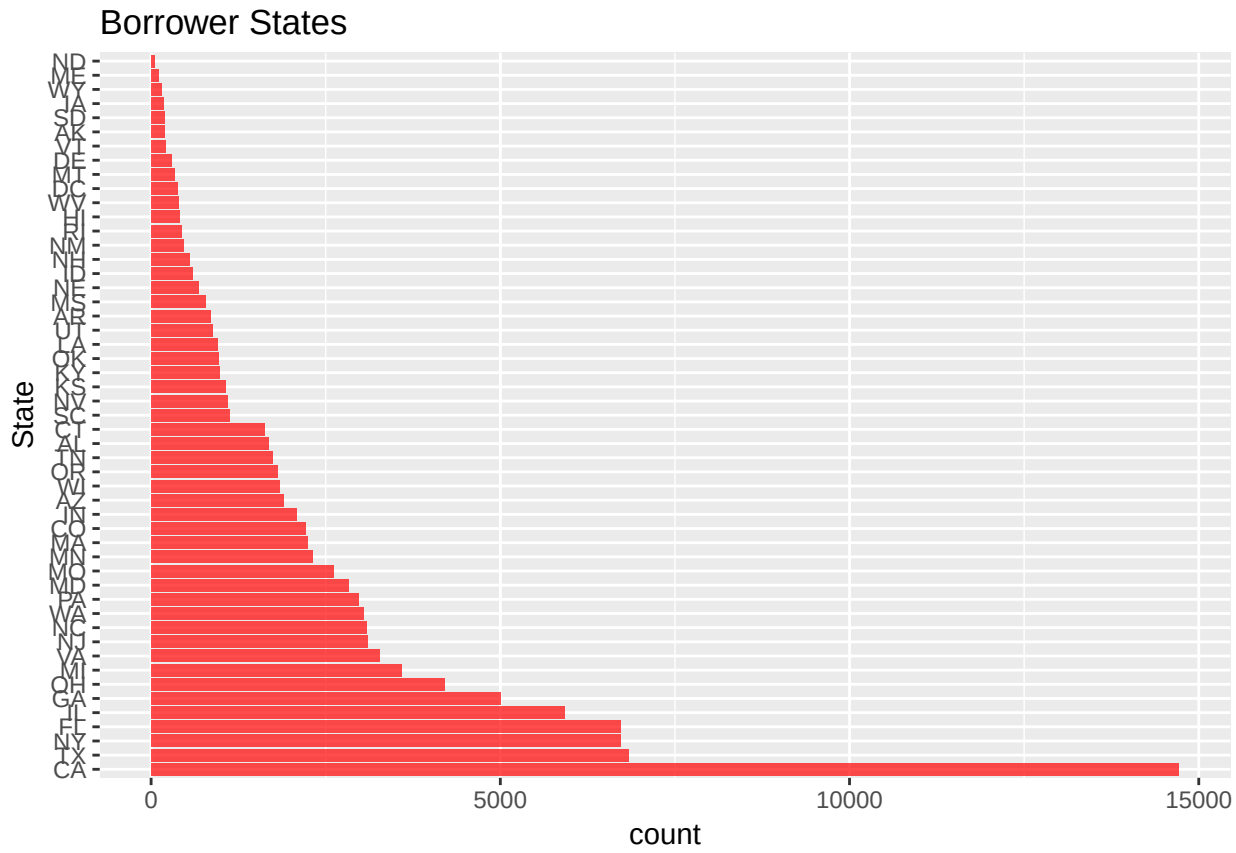



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## 0.00653 0.15629 0.20976 0.21883 0.28384 0.51229      25
```

There are numerous variables that follow the same pattern as BorrowerAPR (e.g. BorrowerRate, LenderYield, etc). However, BorrowerAPR was chosen for analysis as it accounts for more overall expenses *for the borrower*. Higher APR means higher expenses, which usually means higher difficulty of repayment.

The graph shows a fairly uniform distribution, with peaks between 0.15 and 0.20. There is an interesting peak just after 0.35 as well.

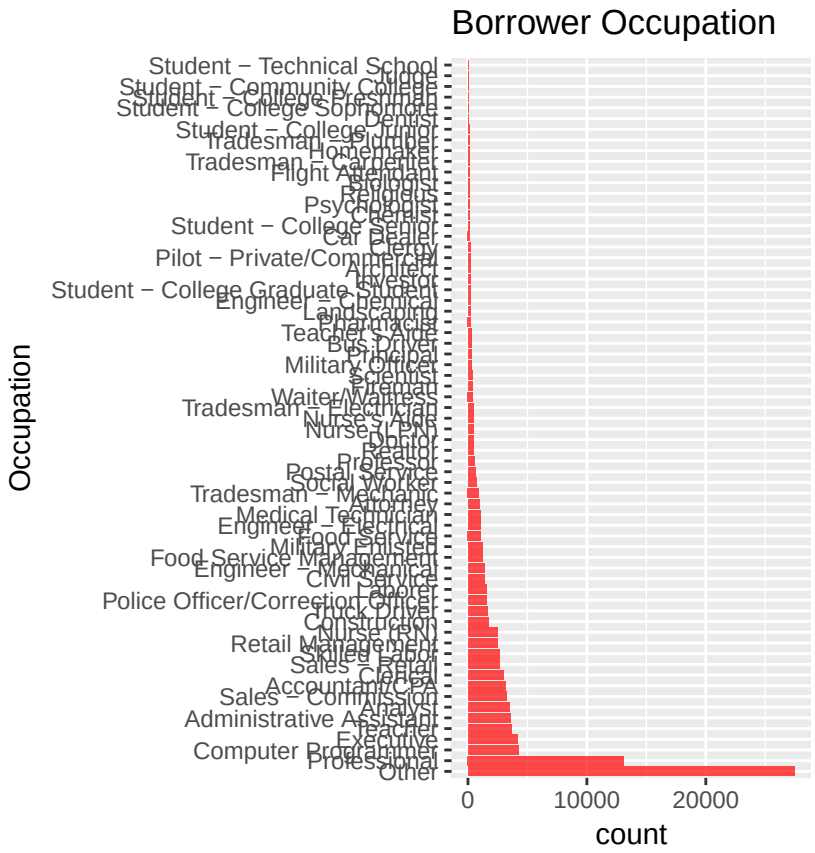
BorrowerState (BS)




```
##    AK    AL    AR    AZ    CA    CO    CT    DC    DE    FL    GA    HI
##   200  1679   855  1901 14717  2210  1627   382   300  6719  5008   409
##    IA    ID    IL    IN    KS    KY    LA    MA    MD    ME    MI    MN
##   186   599  5921  2078  1062   983   954  2242  2821   101  3593  2318
##    MO    MS    MT    NC    ND    NE    NH    NJ    NM    NV    NY    OH
##  2615   787   330  3083    52   674   551  3097   472  1090  6729  4197
##    OK    OR    PA    RI    SC    SD    TN    TX    UT    VA    VT    WA
##   971  1817  2972   435  1122   189  1737  6842   877  3278   207  3048
##    WI    WV    WY
##  1842   391   150
```

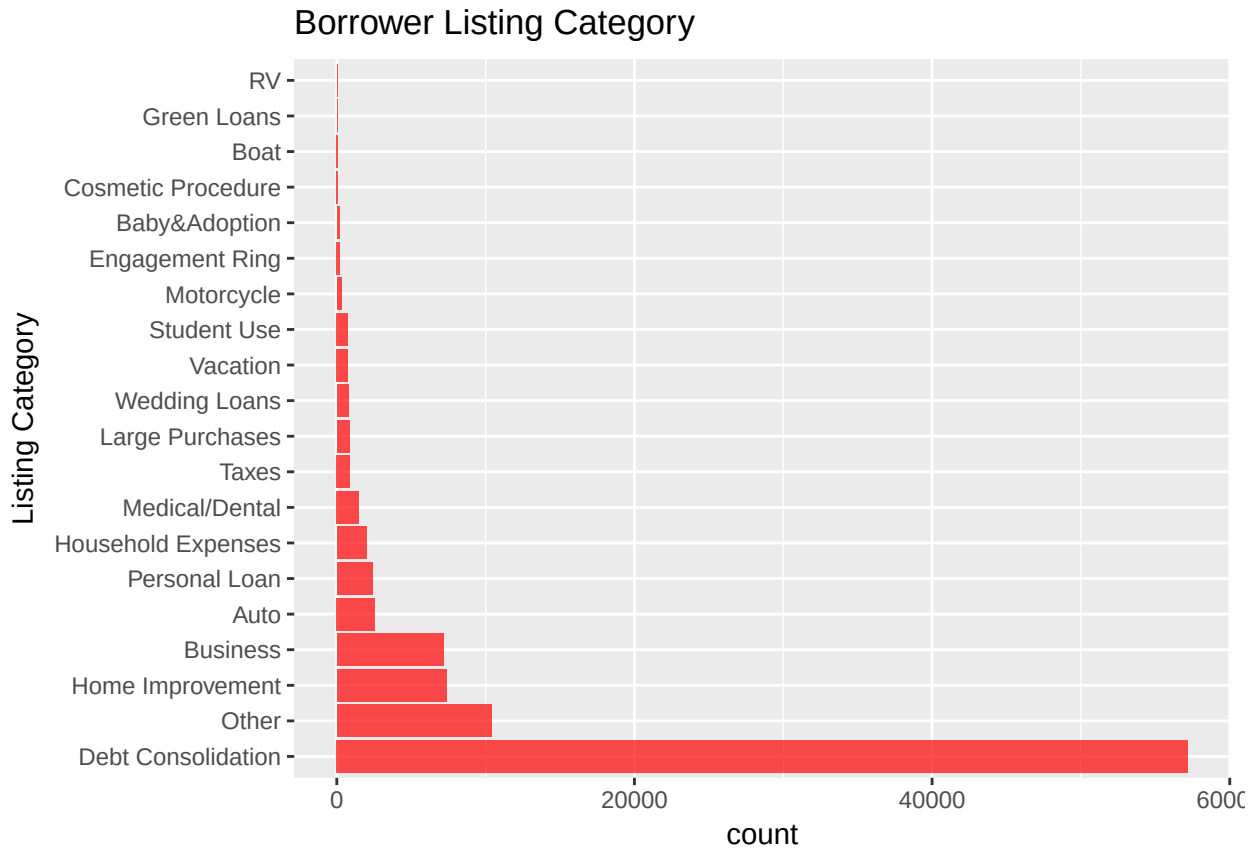
This graph shows the states borrowers live in. The overwhelming majority of loans were made to borrowers from CA, followed by NY, TX, and FL. Perhaps borrowers from different states have different GBRs?

Occupation (OCC)



This graph shows the occupations of the various borrowers. There are numerous categories of occupations, but there are two patterns that can be immediately seen: - The most popular Occupation labels are very vague (i.e. “Professional” & “Other”) - While there are many categories here, there are noticeably low number of loans made out to students of various education levels.

Listing Category (LC)




```
## Debt Consolidation   Home Improvement          Business
##              57184              7392              7177
##      Personal Loan       Student Use              Auto
##               2383               755              2559
##              Other     Baby&Adoption              Boat
##              10395               199                85
## Cosmetic Procedure    Engagement Ring       Green Loans
##                 91               217                58
## Household Expenses    Large Purchases    Medical/Dental
##               1975               865              1500
##         Motorcycle                RV             Taxes
##                304                51               880
##           Vacation     Wedding Loans
##                758               769
```

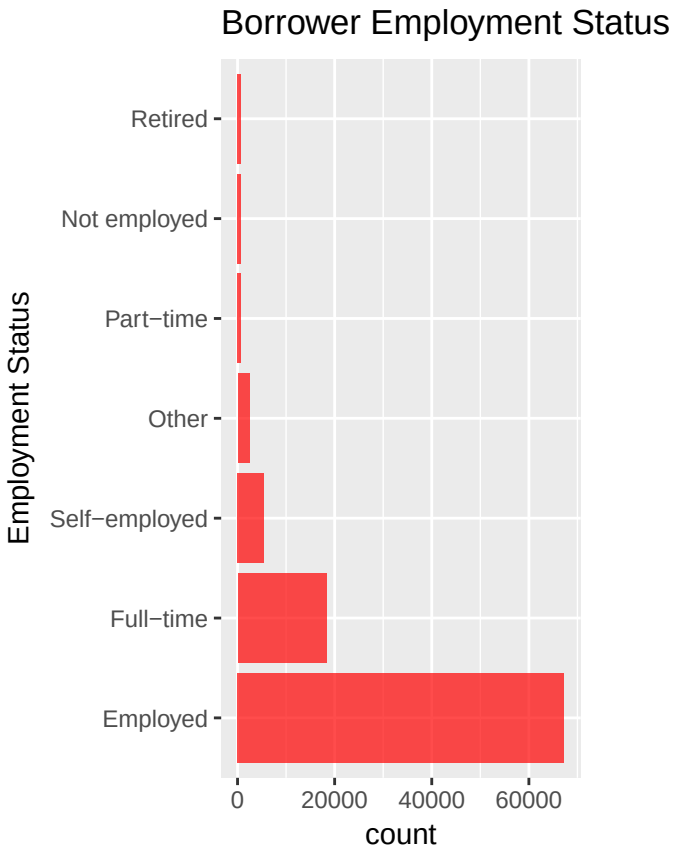
This graph shows the reasons for the loans issued. A huge majority of loans were made for debt consolidation purposes. Perhaps some listing categories are higher risk than others, and have higher rates of nonpayment?

Homeowner Status (HOS)

```
## False  True
## 45573 50024
```

There is an almost 50/50 ratio of homeowners to non-homeowners.

Employment Status (ES)

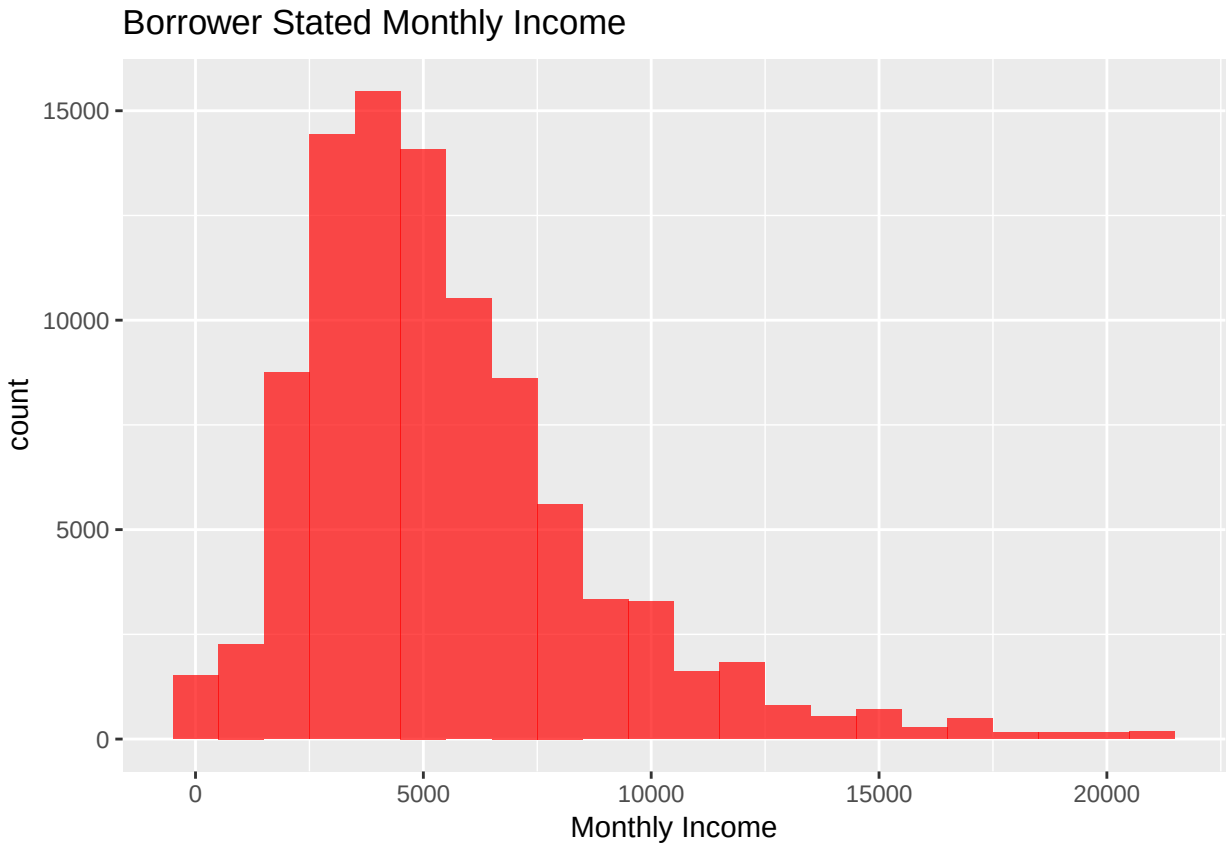



```
##                    Employed     Full-time Not available  Not employed
##             0         67306         18308             0           734
##         Other     Part-time       Retired Self-employed
##          2473           739           629          5408
```

This graph shows the various employment statuses of borrowers. These labels are not very accurate; some labels, including the most frequent occurence ("Employed") do not specify type of employment, whereas the other labels do (i.e. "Full-time", "Self-employed", etc.)

We could definitely surmise that having employment is a big factor in being able to pay off debt. However, a borrower can have still have income and not be employed.

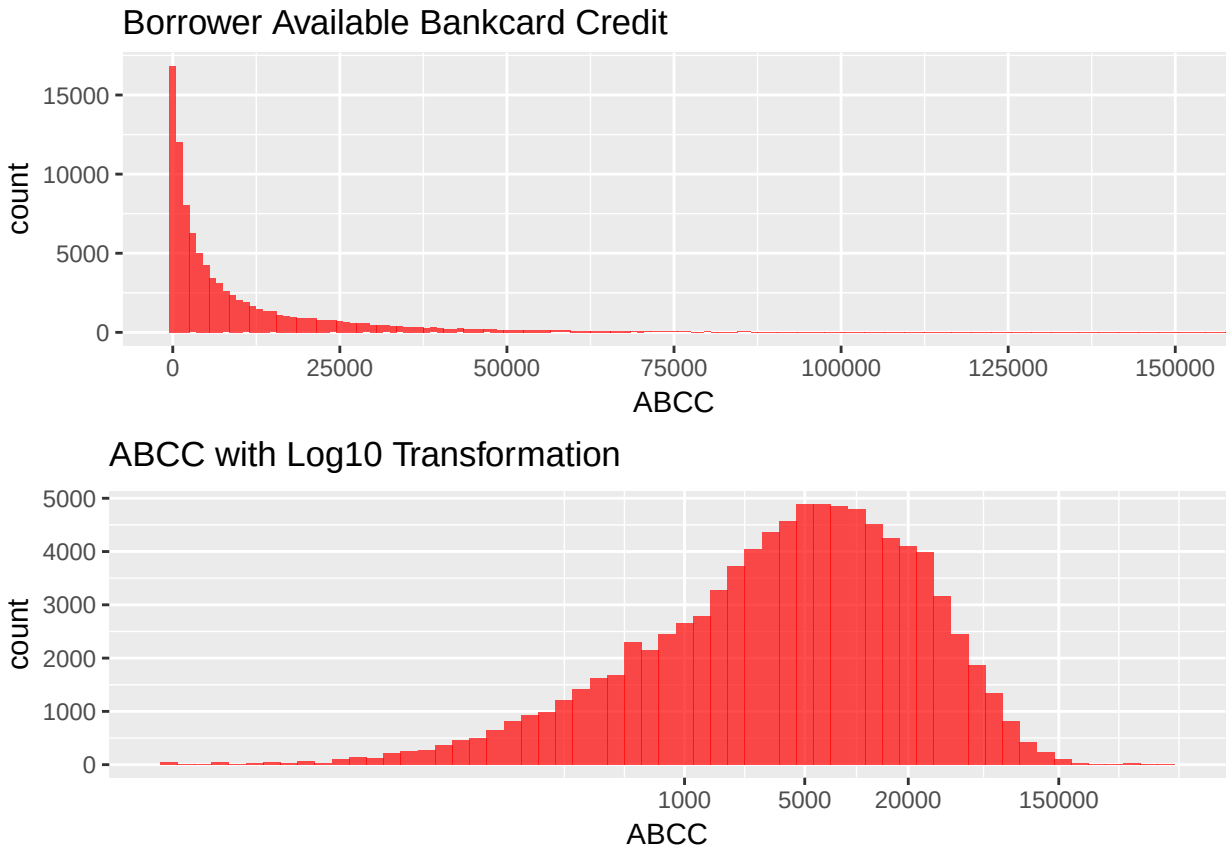
Stated Monthly Income (SMI)




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0    3333    4853    5534    7000   20833
```

This graph shows the monthly income of borrowers, which is vital to ascertaining whether a borrower can pay. Here, we can see that there is a positively skewed distribution, peaking around $4000. Only 99% of data is presented in order to eliminate any outliers (the maximum Stated Monthly Income is $175k, which must be a mistake).

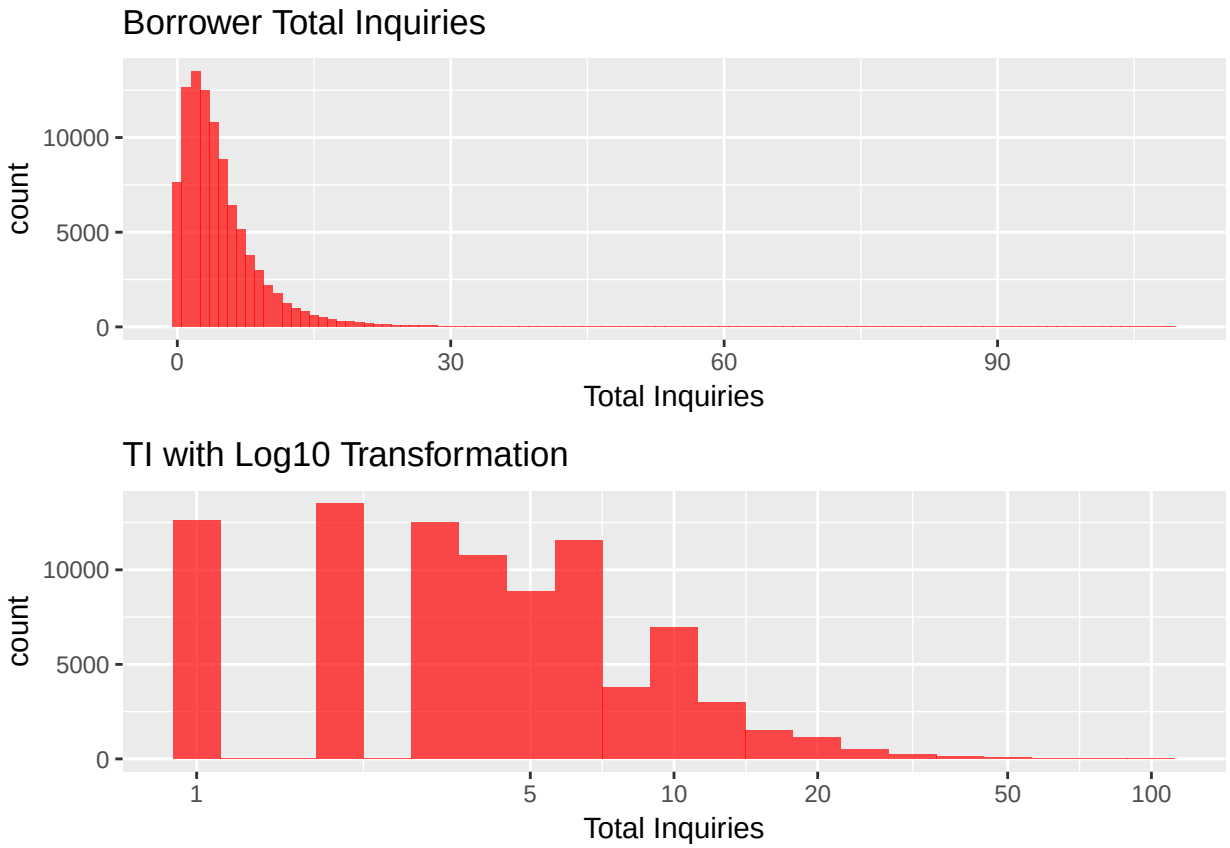
Available Bank Card Credit (ABCC)




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0     990    4318   11245   13496  572427
```

This graph shows the total available credit via bank card at the time a borrower's credit profile was pulled. This is also very positively skewed, with the majority of borrowers having a very low amount of bank card credit (between $0 and $25,000). This makes sense, as borrowing funds via credit card is an easier process than going through Prosper loans. The second plot shows a log10 transformation, which reveals a negatively skewed distribution, peaking around $5k-$10k.

Total Inquiries (TI)

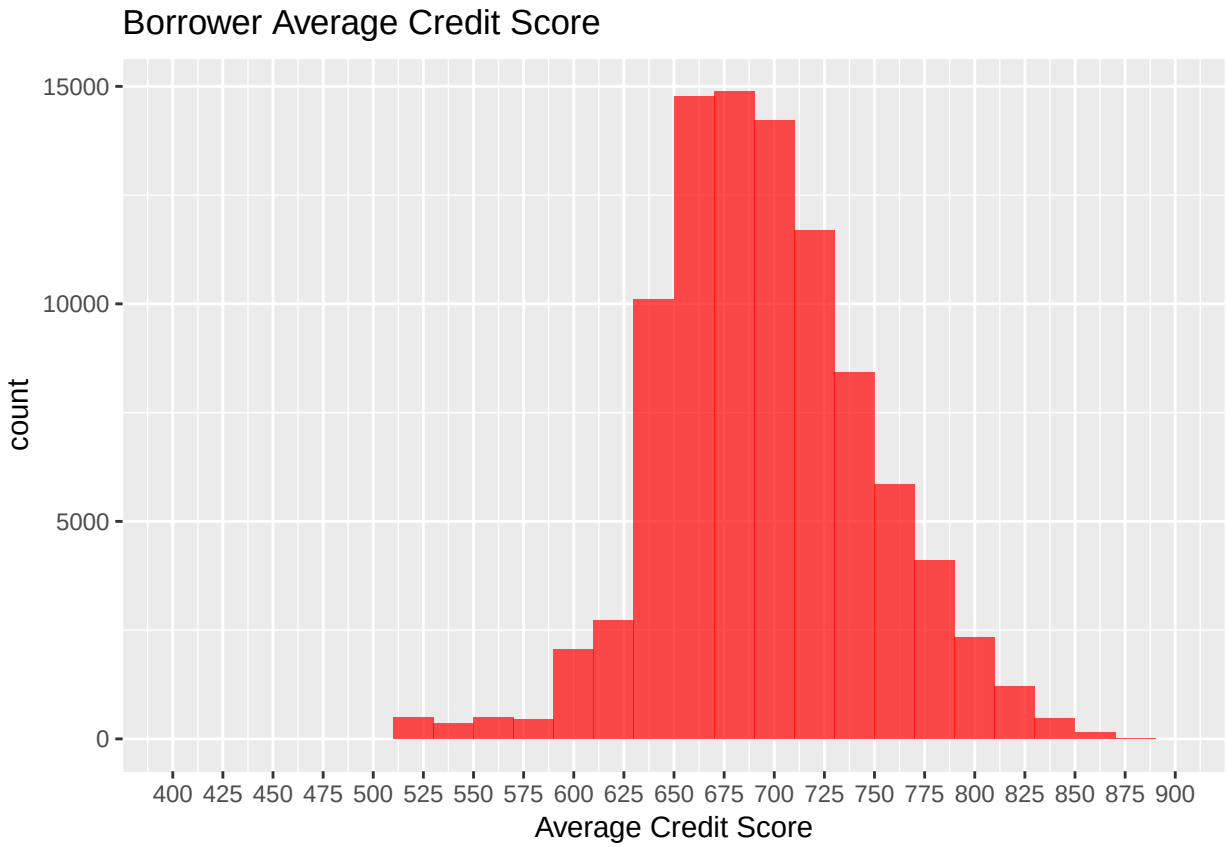



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.000   4.000   4.786   6.000 109.000
```

This graph shows the total number of inquiries at the time the credit profile was pulled. Inquiries are requests by businesses to check the score, usually for lending purposes. Higher numbers usually means higher risk.

This is very positively skewed, peaking at 2 inquiries. The log10 transformation shows majority of distribution between 1 and 6 inquiries, with a peak at 10 as well.

Average Credit Score (ACS)

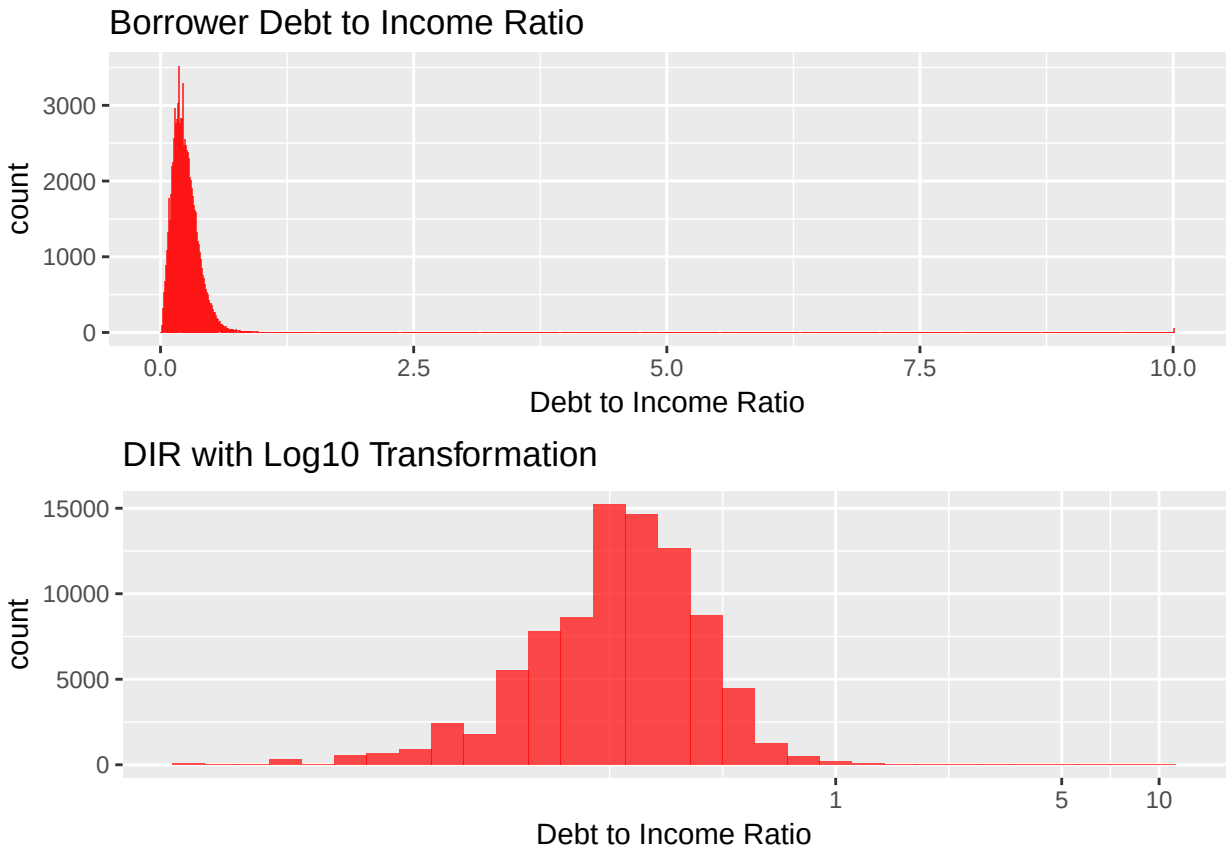



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   529.5   669.5   709.5   704.6   729.5   889.5
```

There are multiple various credit ratings used to label borrowers and their risk level. Credit ratings illustrate past borrowing/debt history, and thus can be potentially very relevant to a borrower's ability to pay a loan.

Based on our hypothesis, this report will *focus specifically on AverageCreditScore*, as this is a factor that is somewhat controllable by the borrower's actions, whereas the other variables are scores allocated to the borrowers based on ACS and other factors.

Debt to Income Ratio (DIR)




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   0.150   0.220   0.259   0.320  10.010    8103
```

This graph shows the debt to income ratio of the borrower at the time the credit profile was pulled. The log10 plot shows an overwhelming majority of data limited between 0 and 1, but a small amount also at 10.01.

Univariate Analysis

What is the structure of your dataset?

The Prosper Loan dataset contains 113,937 loans with 81 variables on each loan. Out of these 81 variables, this report focuses on 15, as shown above.

- Majority of Loan Statuses are Current
- Average LOA is $8337
- Average MLP is $272.50
- Majority of loan Terms is 36 months
- Average APR is 21.8%
- Majority of borrowers live in CA, followed by TX, FL, and NY
- Majority of borrowers are have occupation of “Other” and “Professional”
- Majority of loans used for purpose of debt consolidation
- There is a fairly even split between borrowers who do and do not own homes
- Majority of borrowers are employed
- Average monthly income is $5608

- Majority of borrowers have a very low amount of ABCC
- Mean number of total credit inquiries is around 5
- Average credit score is 695
- Average DIR is 0.276

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

The main feature of interest is LoanStatus, since it directly relevant to our null hypothesis. *LS is thus the response variable of this report.* The other variables will be analyzed in bivariate and multivariate analyses in combination with LoanStatus to see whether they are explanatory variables i.e. whether they correlate to LoanStatus in some way.

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

I believe the main secondary features should be related to borrower income (i.e. SMI), since incoming revenue is an indicator of cash flow, which is directly related to the payment of expenses, such as a loan.

However, there may be other features that aren't related to income, but describe the financial behavior of borrowers, and how capable they are of managing finances responsibly. ACS is one such variable, as is ABCC and TI.

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

ACS was created using the average of upper and lower credit scores for borrowers.

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?

DIR, TI, and ABCC all have very positively skewed relationships, and thus they had log10 transformations applied in order to reveal any interesting patterns. However, it must be noted that these transfomations remove any values with 0

LS had the "Cancelled" loans removed, as mentioned above.

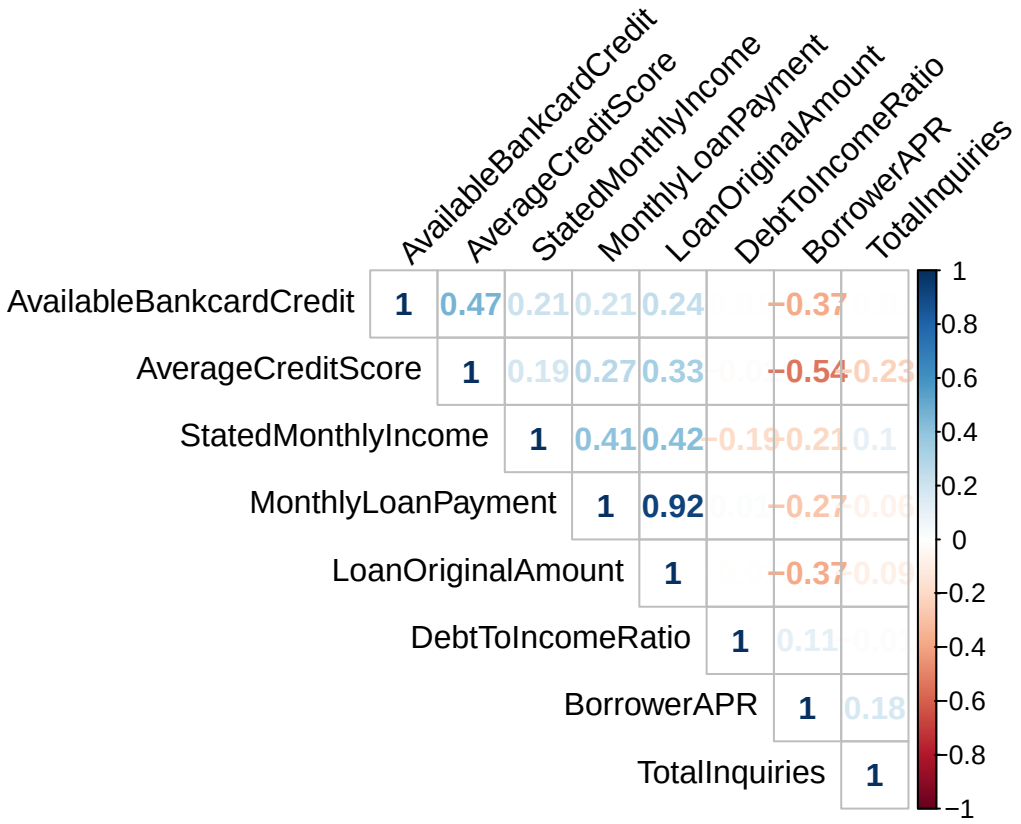
BS, OCC, ES, and LC all had either empty level names or levels named "Not Available". These were initially going to be kept, but further bivariate analysis showed that they interfered with correlation tests too much.

SMI also had numerous outliers, so the top 1% of data points were removed.

In all, 19,148 data points were removed.

Bivariate Plots Section

Correlation Matrix Between Quantitative Variables



This correlation matrix reveals a few interesting correlations between the quantitative secondary variables in this report.

LOA & MLP have an extremely strong positive relationship; thus we can see in the matrix that MLP exhibits slightly weaker correlations with other variables than that of LOA. This also means that *MLP can be virtually ignored as an independent variable*, due to its very strong correlation with LOA. *LOA/MLP are positively correlated with SMI, ACS, and ABCC*, as all these factors would be relevant in determining whether the borrower could pay off such a large loan.

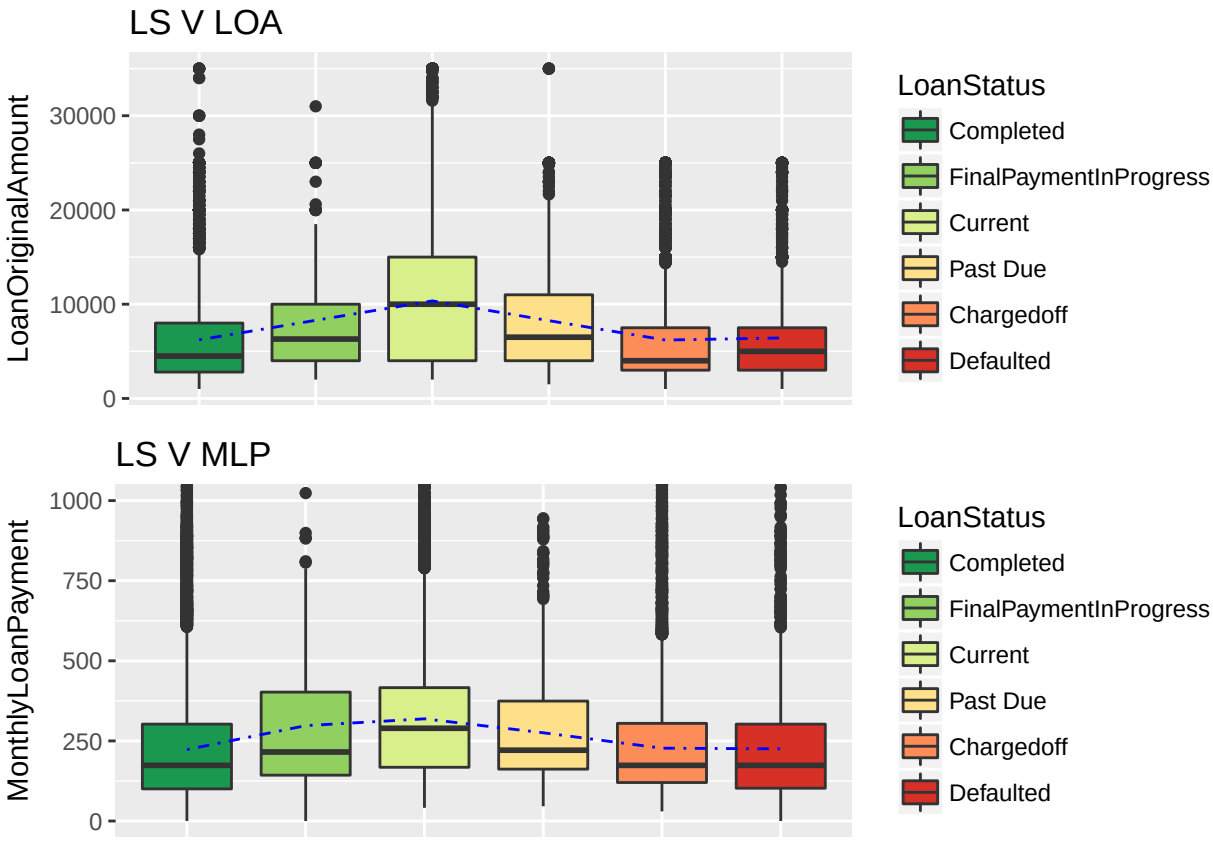
APR has negative correlations with many variables in varying degrees of strength. While correlation does not imply causation, we can see that the same variables positively correlated with LOA (SMI, ACS, ABCC) are also negatively correlated with APR.

ACS as mentioned above, has a fairly strong positive correlation with ABCC, and a fairly strong negative correlation with APR

SMI surprisingly does not have a strong correlation with the other quantitative variables (except LOA/MLP, as noted above). This shows that just

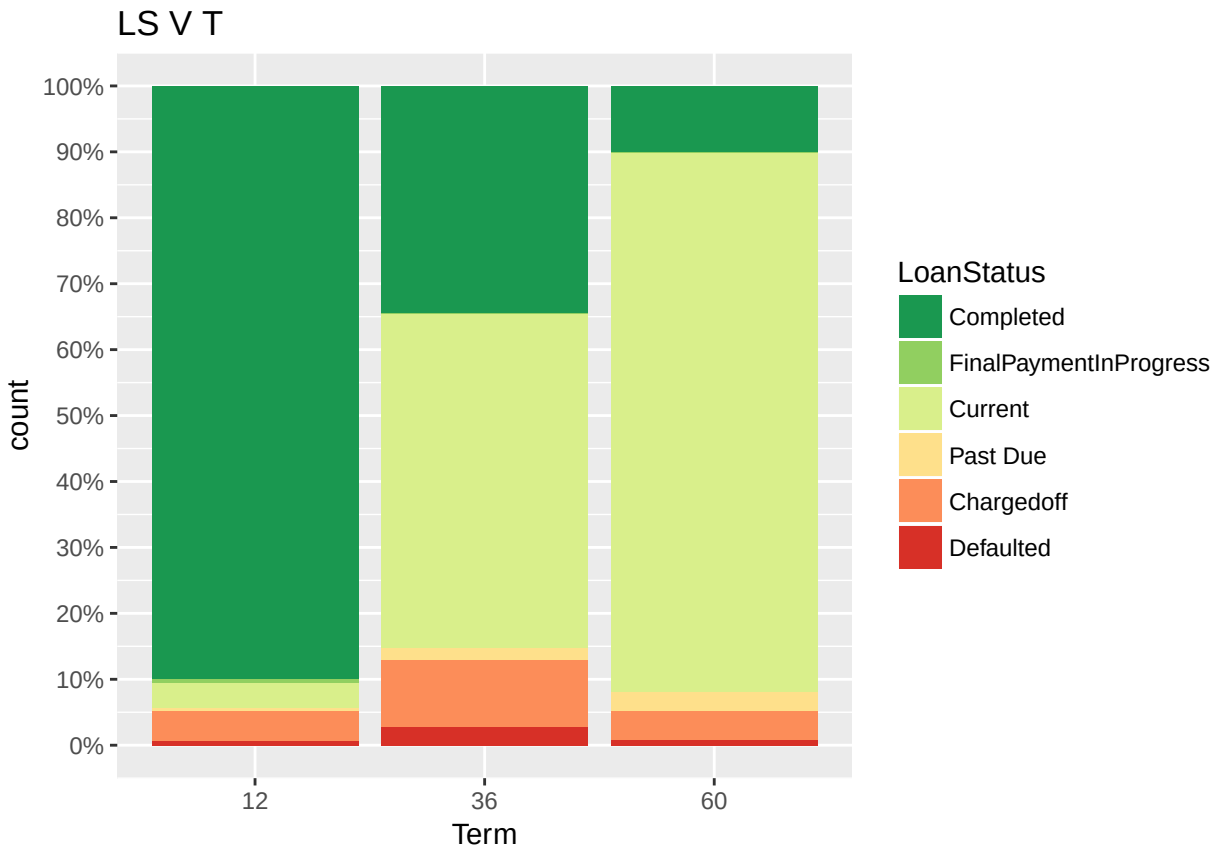
Now we will begin bivariate plotting with LoanStatus against our other supporting variables.

LS V LOA/MLP

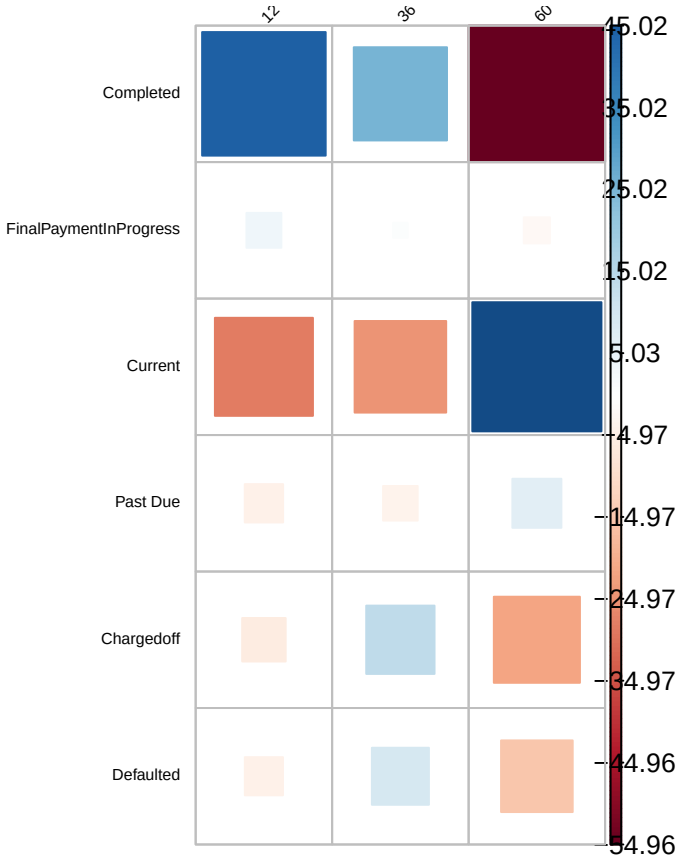


We surmised from the correlation matrix that LOA and MLP are almost identical in nature; these plots reinforces this. Mean LOA/MLP peaks for Current loans at 10k/300 respectively, but otherwise dip down in almost identical patterns for both Completed and Chargedoff/Defaulted Loans. Thus, LOA/MLP *is not* correlated with LS.

LS V T



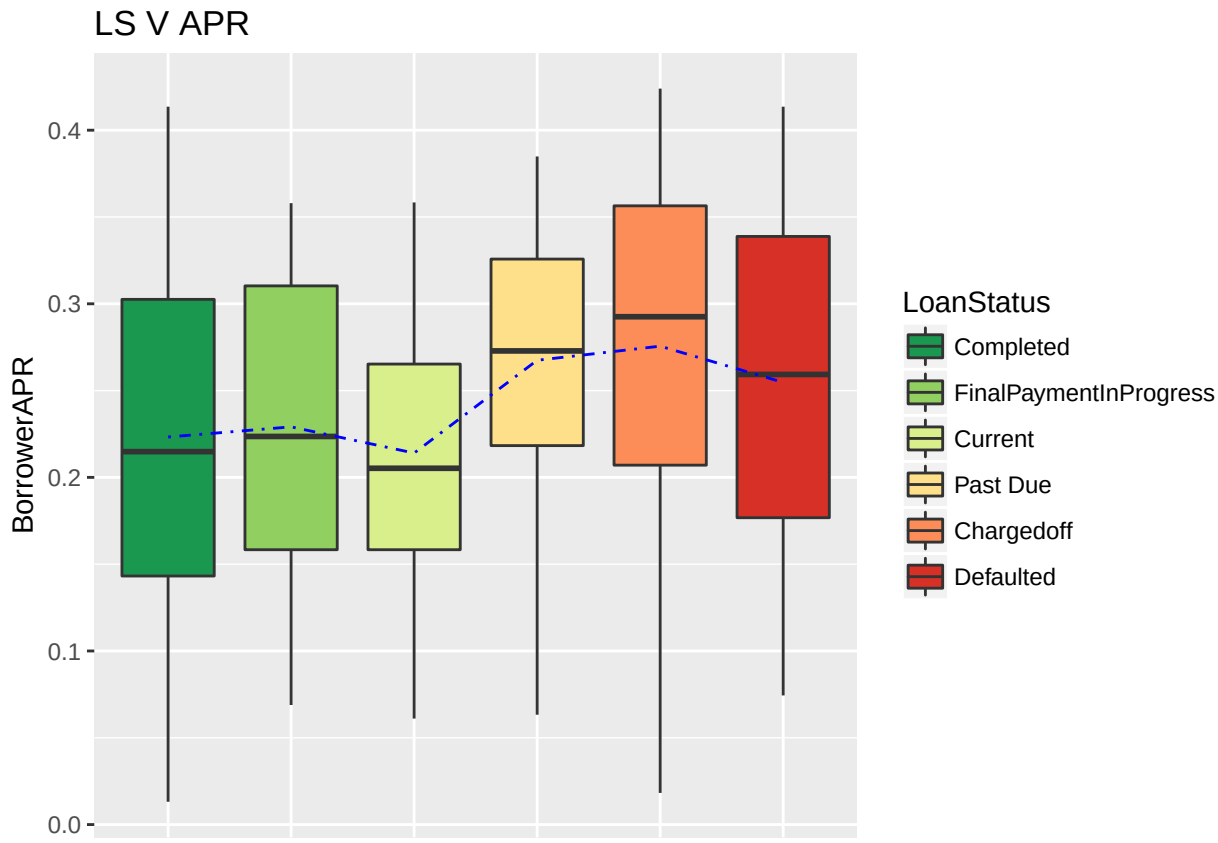
There is a clear difference in Loan Status between Terms. While the huge majority of loans have been made for 36 months, ~15% of those loans are Bad. By comparison, ~5% of 12 month loans are Bad. and ~8% of 60 month loans are also Bad.



The 12M and 60M Term, with higher GBRs, have negative correlations with Bad Loans, while 36M has a weak positive correlation with Bad Loans.

Interestingly, the Chi-square plot also implies that Current loans have strong correlations with 60 month terms, which means shorter term loans are not preferred anymore. However, it does not explain why 12M terms (which have high GBRs) are not used instead. This could be due to the fact that longer terms net more interest (good for investors), and also lower the monthly payments (good for borrowers). In any case, from these graphs, we argue that Term has a *strong positive correlation* with LS.

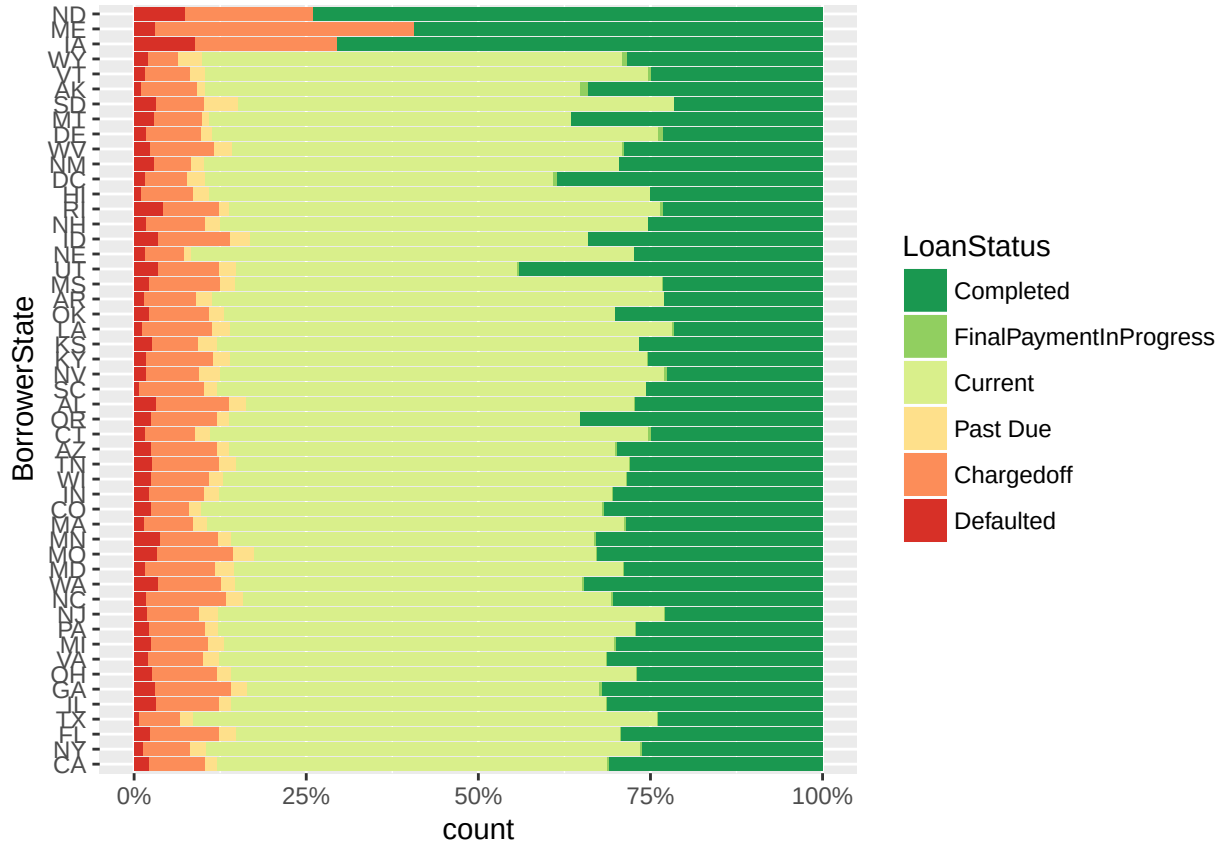
LS V APR



The mean APR mostly follows the median across LS, and while all those values are between 0.2 and 0.3, Good Loans clearly have lower overall mean/median APR (<0.25) than Bad Loans (>0.25).

Thus, we can assert that there is a clear *negative correlation between APR and a borrower's ability to pay.*

LS V BS

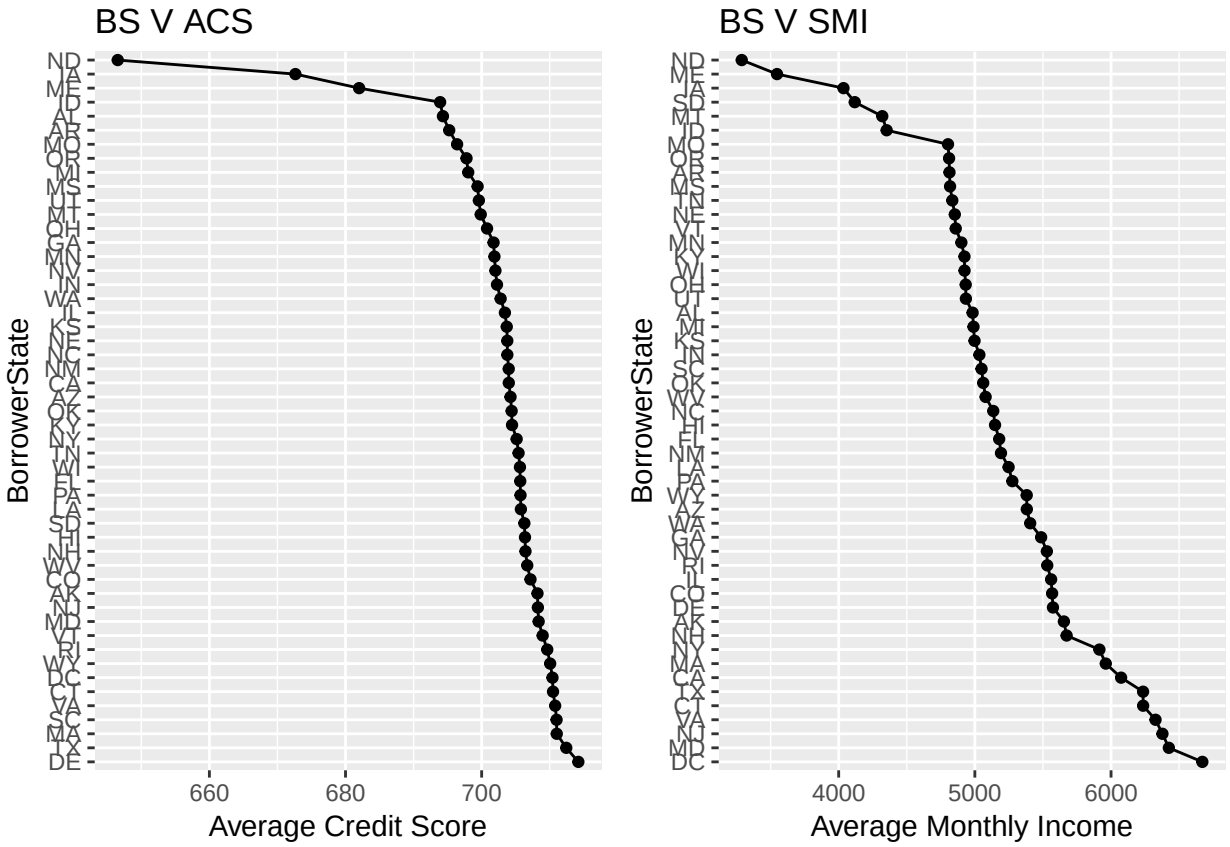


This graph shows Loans categorized by State, ordered descending by the number of loans, but split into proportions based on Loan Status.

It is telling that the states with the lowest GBRs (i.e. ND, ME, and IA have ratios under 75%) have no current loans. Prosper must no longer services borrowers in those states.

It could be argued that ND, ME, and IA had very few loans anyway, so Prosper simply decided to abandon outreach and service in those states. Looking at the data proves otherwise however. The first univariate BS analysis shows that WY, VT, AK (the next lowest occuring BorrowerStates) all have similarly low numbers of loans. Thus, the *frequency of loans does not factor* into why Prosper has ceased issuing loans to those "bad" states.

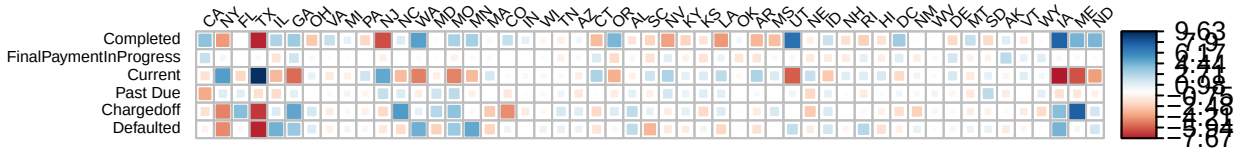
Thus, it must be the features of those states (e.g. sluggish economy, bad neighborhoods?) that somehow affect borrowers, since other states do not have a Bad Loan proportion exceeding 25%.



These first two graphs illustrate the average monthly income and average credit score of each state, ordered descending. We can clearly see that the states with the three lowest incomes (under $4k) and the three lowest ACS are the same "bad" states with no current loans: ND, ME, and IA. After those three though, the order differs.

Descending by SMI, the next three are SD, MD, ID Descending by ACS, the next three are ID, AL, AR

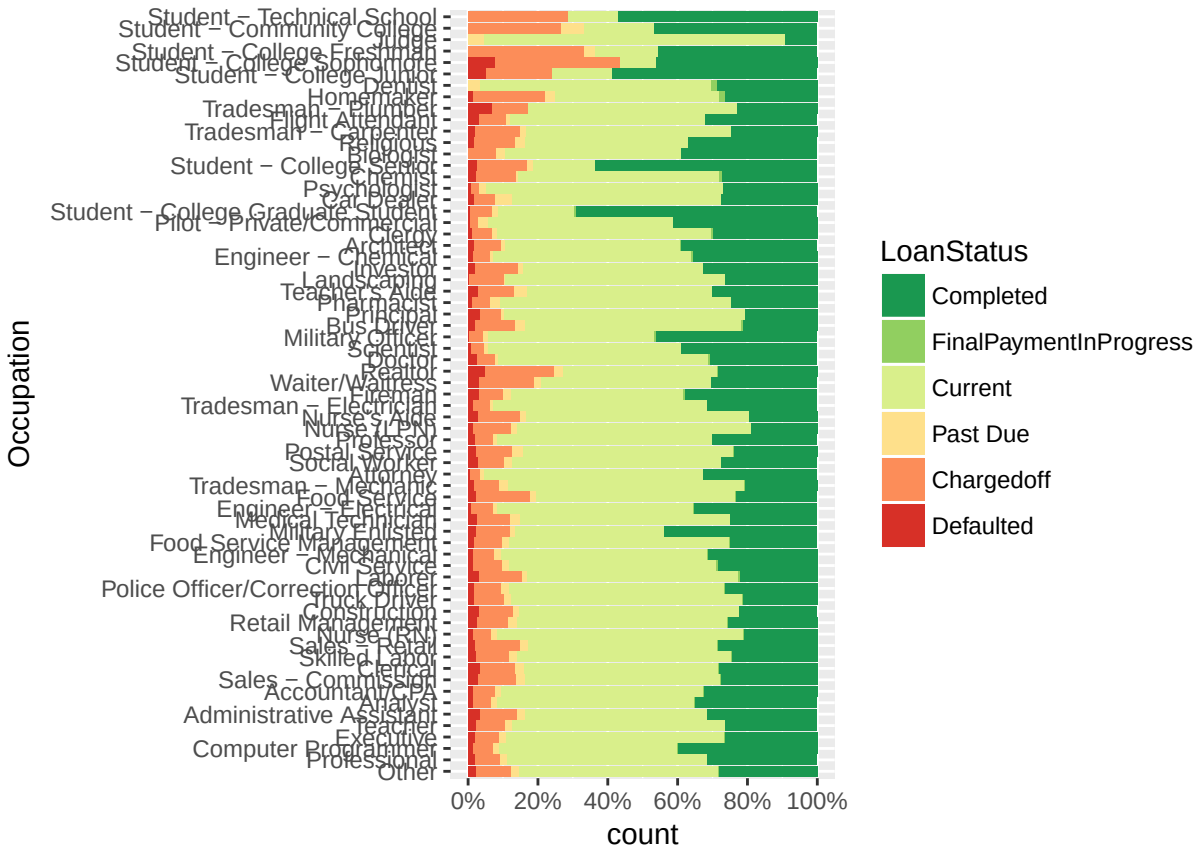
The main difference between the plots is that ACS is grouped closely together, with values no more than 70 points apart, whereas SMI can vary by thousands.



Here we look at the ChiSquare plot. It shows varying correlations between BS and LS. As expected, IA, ME, and ND have negative correlations with current loans, and positive correlations with Completed, Chargedoff, and Defaulted loans. Interestingly, we can also see this same pattern with other states, in particular UT, MO, WA, and GA. However, these patterns do not definitively establish that borrowers from these states have a higher chance of bad loans, only that there is a negative correlation with borrowers getting new loans. Some of those states (e.g. IA, ND, UT) also have stronger positive correlations with Completed status than Chargedoff or Defaulted status.

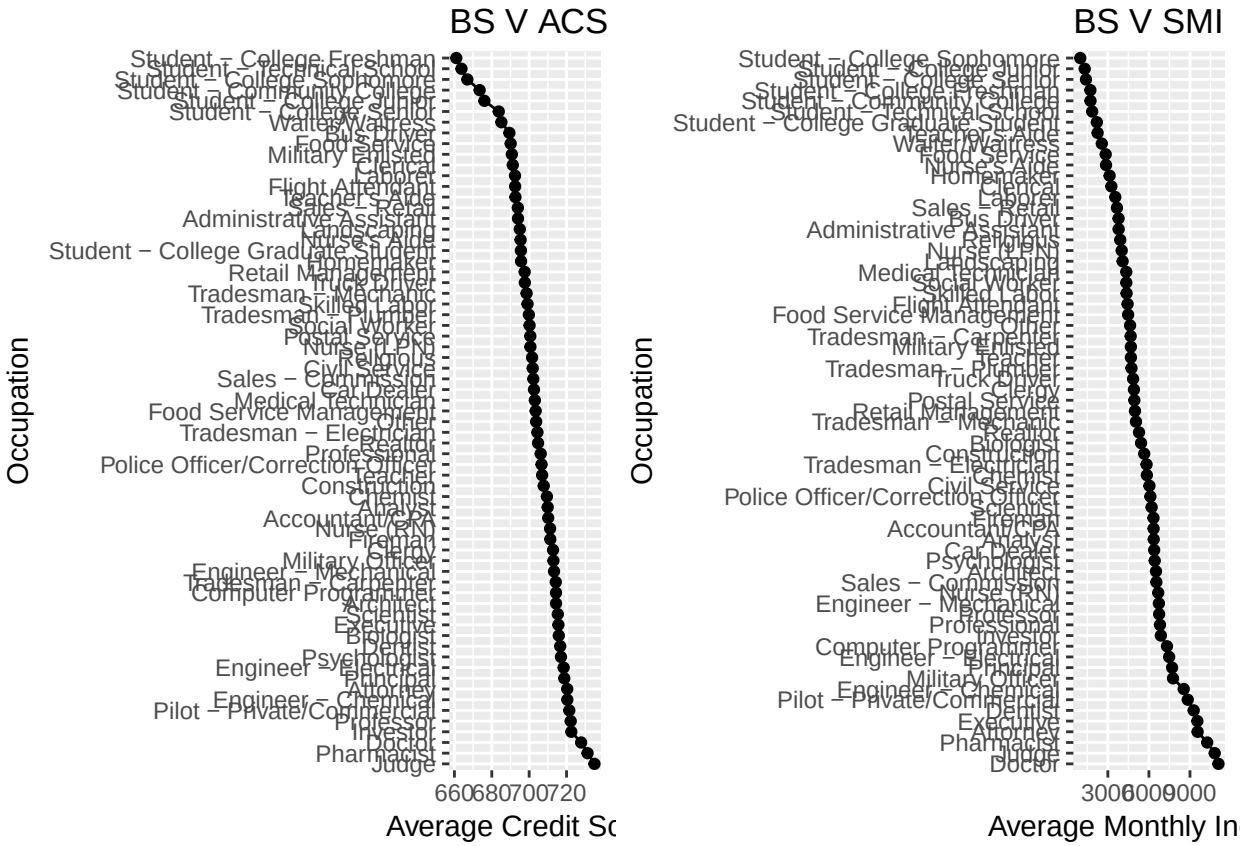
In any case, we can definitely establish that BS has a *strong correlation* with LS.

LS V OCC

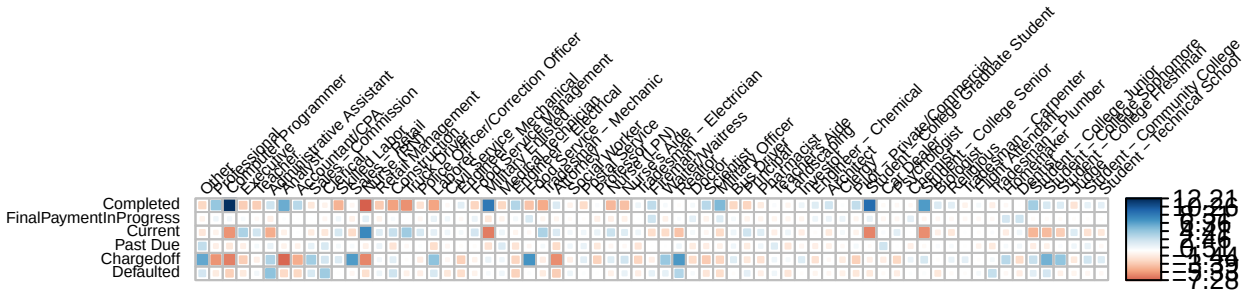


This graph shows Loans categorized by Occupation, ordered by the frequency of those occupations, but split into proportions based on Loan Status. The average proportion of Bad Loans appears to be around 20%.

In the univariate analysis, we saw a pattern of Students having low numbers of loans. Here, we can see another pattern: Students (whether of Technical Schools, Community College, College Freshman, Sophomore, Jr, Sr, or Graduates) have low proportions of Current loans. This pattern, combined with the lower GBRs (~30%, with the exception of Seniors and Graduate Students), implies that Prosper is ceasing loans to all Students.



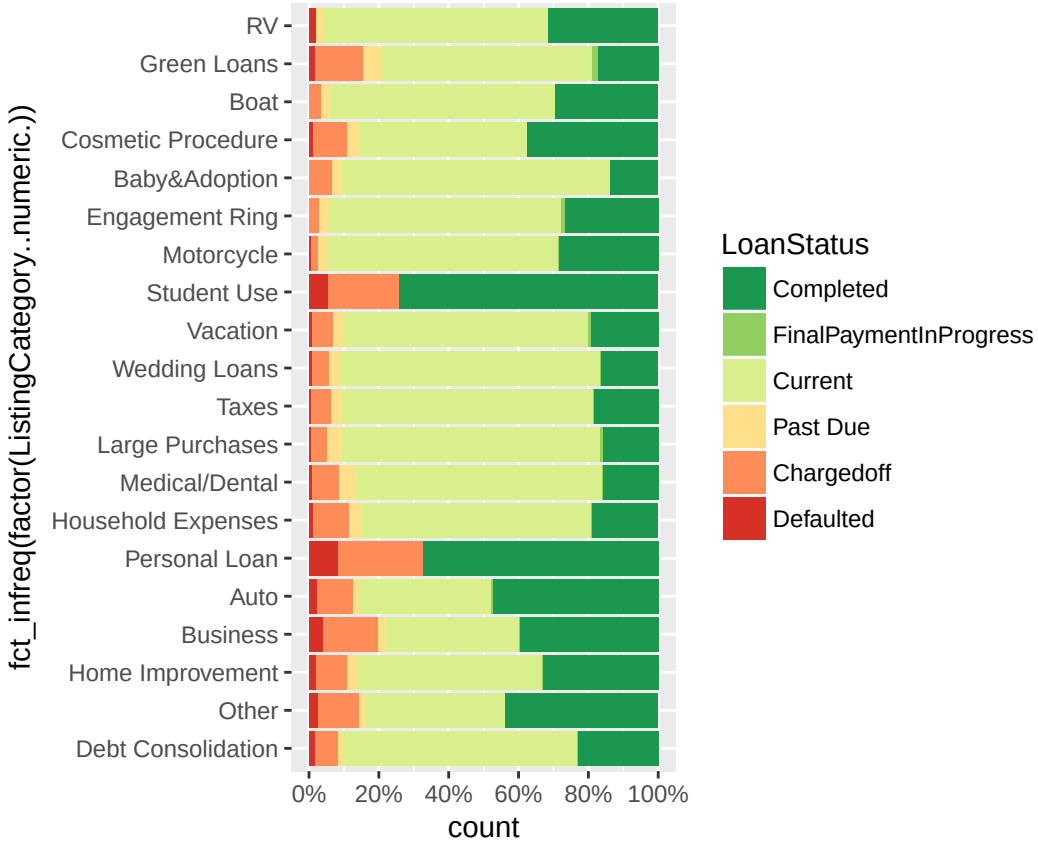
Orering occupations by average income further reinforces the importance of SMI and ACS in the context of occupation, as students clearly have the lowest amounts, followed by low-skilled jobs such as Teacher's Aide, Waiter/Waitress, Food Service, Nurse's Aide, Clerical, Laborers, etc.



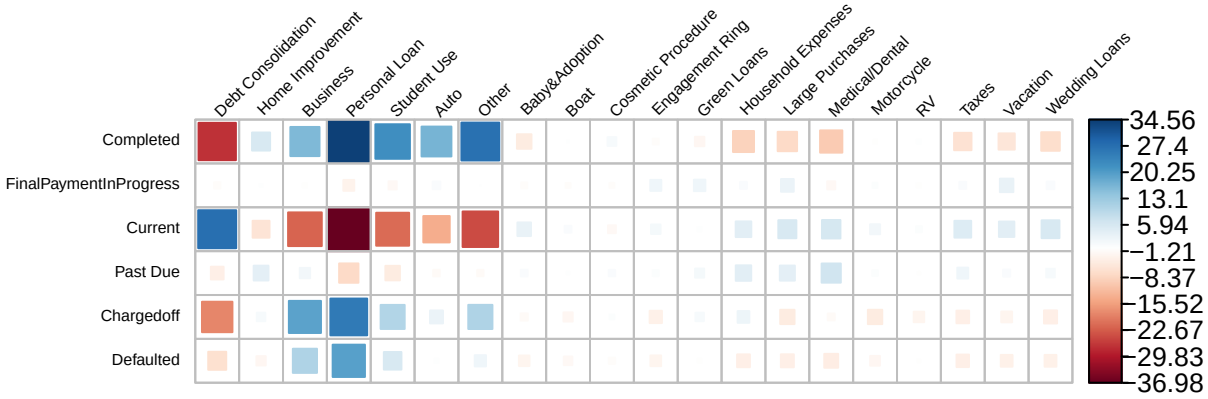
The Chi square plot shows a few more patterns: - *Professionals, Computer Programmers, Analysts, Miltary Enlisted/Officers, College Graduates, and College Seniors* actually have strong correlations with paying off loans successfully. This is visualized by blue positive shades in Good loan statuses, and red negative shades in Bad loan statuses. - On the other hand, occupations such as *Others, Sales - Commissions, Sales - Retail, Admin Assistants, Clerical, Laborer, Construction, Food Service, Realtors, and other Students* have strong positive correlations in Bad loan statuses, and vice versa.

We can definitely establish that OCC has a *strong correlation* with LS.

LS V LC



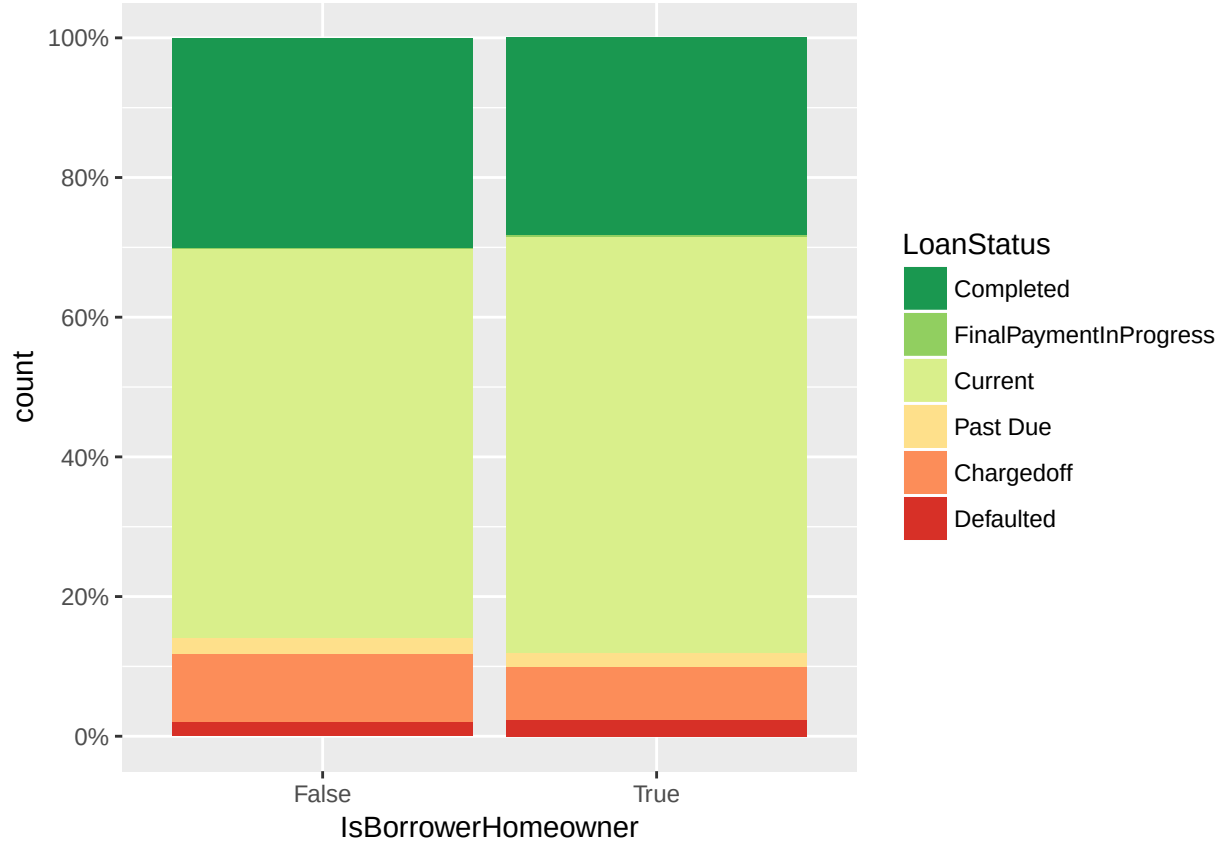
Personal and Student Loans are all either completed, charged off, or defaulted. These loans also have significantly lower GBRs (~60-65%). Comparing this with the occupation bivariate analysis above, it makes sense that loans for students are no longer permitted post-2009, since they had the highest levels of nonpayments. It also makes sense that Prosper requires a reason for loans. But what makes these listing categories more risky?



In the Chi Square plot, we can clearly see a very strong positive correlation between current Loans and Debt Consolidation, which means that many new borrowers are getting loans for this reason. In contrast, there are strong negative correlations for Current loans for LC values of Personal, Other, Business, Student, and Auto. This implies that these loans are being phased out. Personal and Business loans also have the strongest correlations to Chargedoff and Defaulted LoanStatuses. The other LC levels seem all follow the same pattern, with weak positive correlations towards Current loans. Note however that the LC values of Medical/Dental, Large Purchases, and Household Expenses also have weak positive correlations in Past Due!

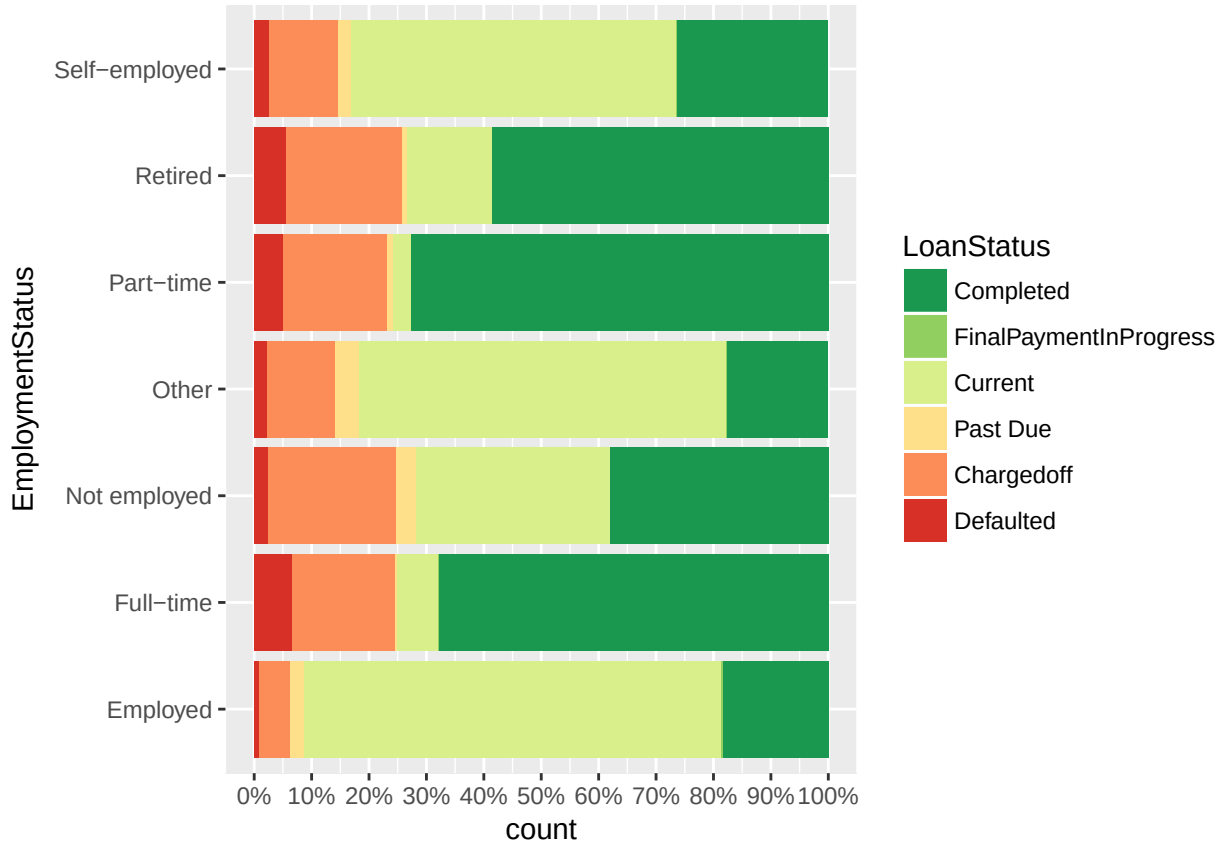
We can definitely assert that LC has a *strong correlation* with LS.

LS V HOS

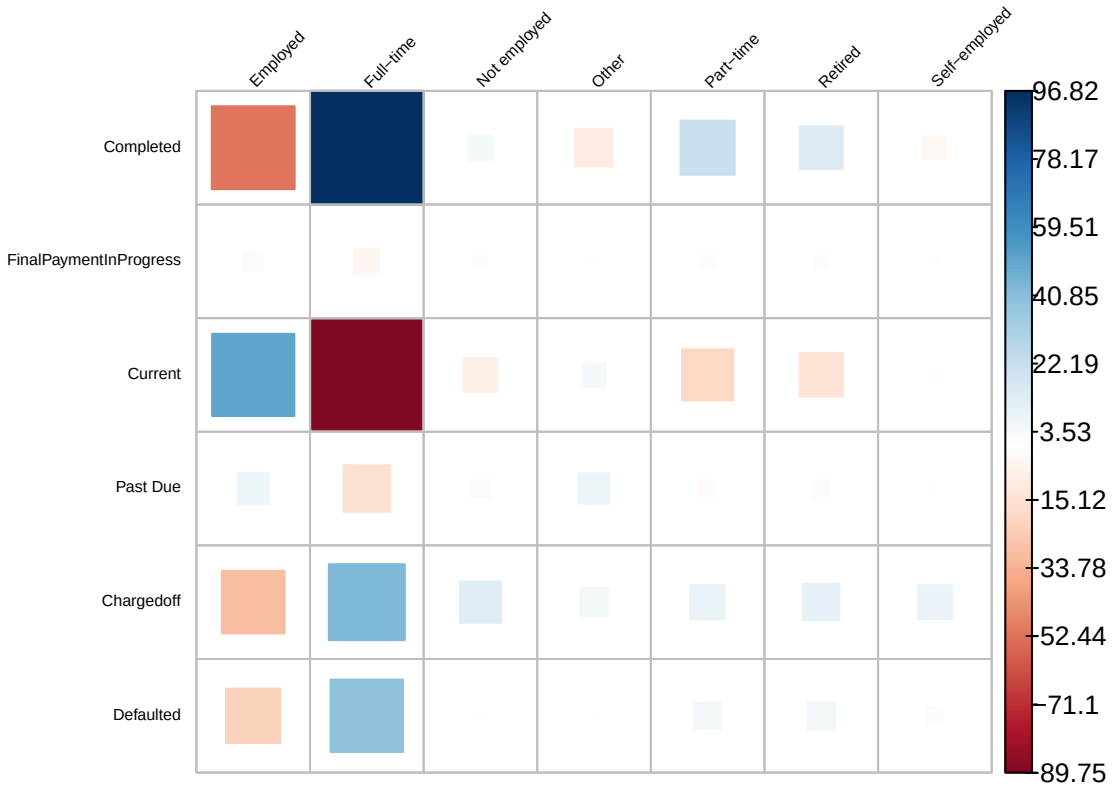


We already established that there is little difference in the amount of borrowers who are and are not homeowners; we can see that the difference in proportion of Good/Bad Loans is equally small. Non-homeowners have a *slightly* lower GBR, but the difference is surprisingly negligible. We surmise that homeowner status *does not* correlate with on a borrower's ability to pay off loans.

LS V ES



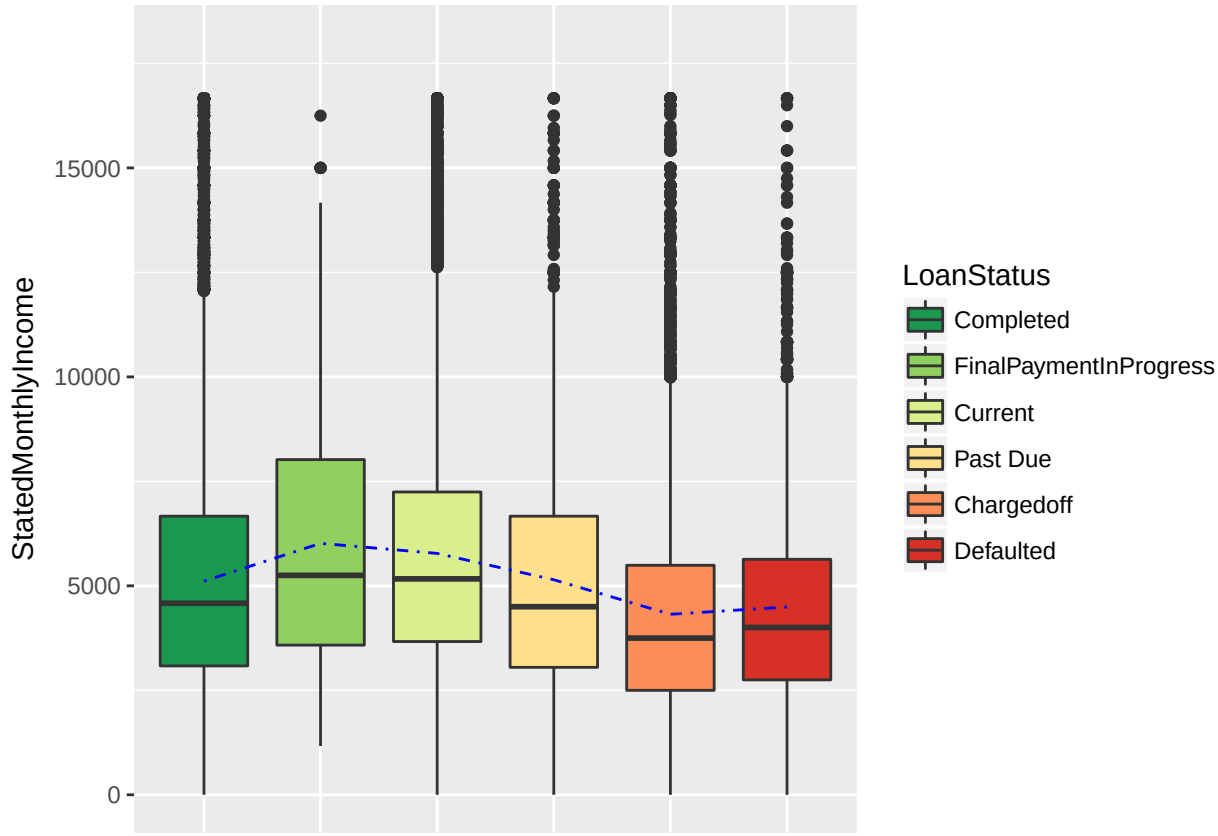
Here, we can see huge GBR differences across ES. Retired, Part-time,and Full-time ES levels have the lowest GBRs, and also low proportions of Current loans, which indicates that Prosper is no longer issuing loans to these kinds of borrowers. Borrowers with the Employed level on the other hand have the highest GBR, highest frequency, and highest Current loan proportion.



The Chi-square plot supports the above assertions. Full-time borrowers in particular have strong correlations with both Completed and Chargedoff/Defaulted LS. Employed is also the only level to have negative correlations with Bad loan statuses.

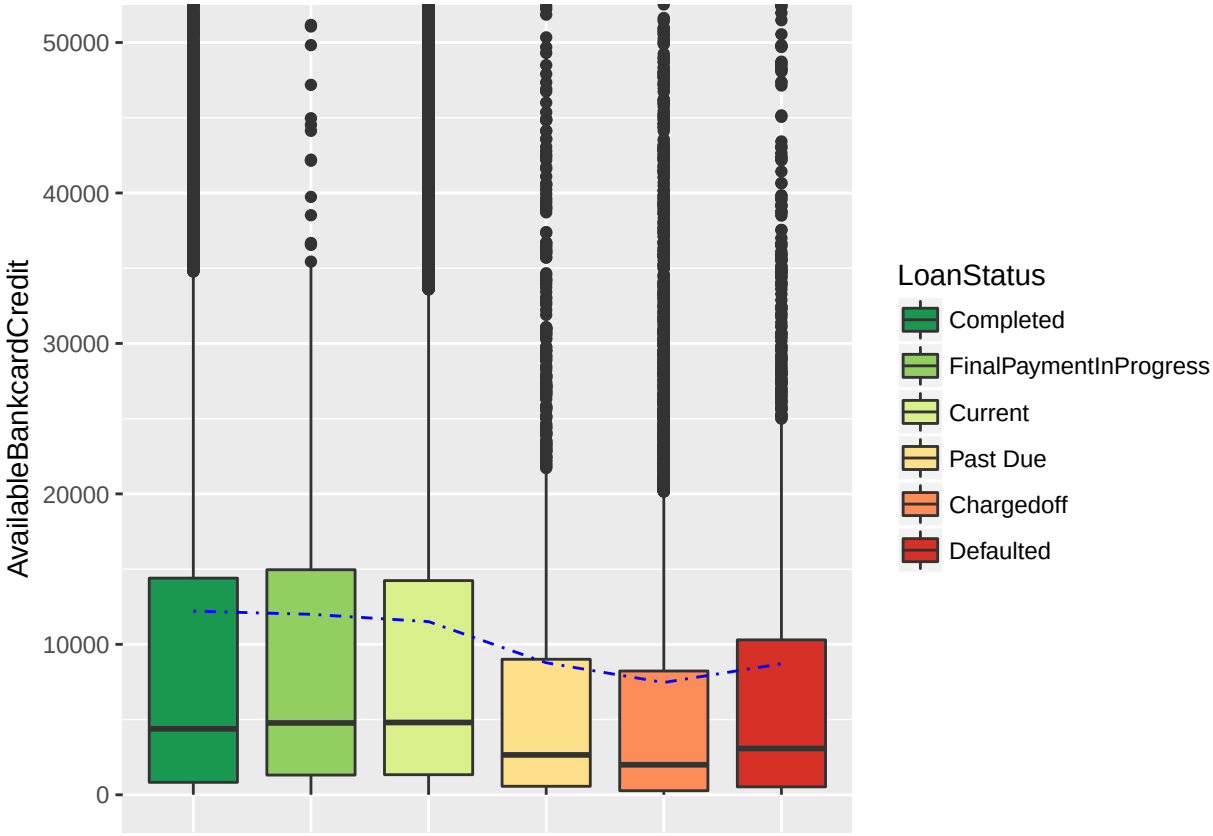
We can definitely assert that ES has a *strong correlation* with LS.

LS V SMI



SMI is distributed fairly similarly, though the mean lines are higher than median (most likely due to outliers). We can still see a clear pattern between Good Loans and Bad: Good Loans have higher SMI, and Bad loans have lower SMI. However, the difference is not as much as one would think; the medians and means across all LoanStatuses are actually quite close. Thus, we can assert that there is a *weak positive correlation* between SMI and LS.

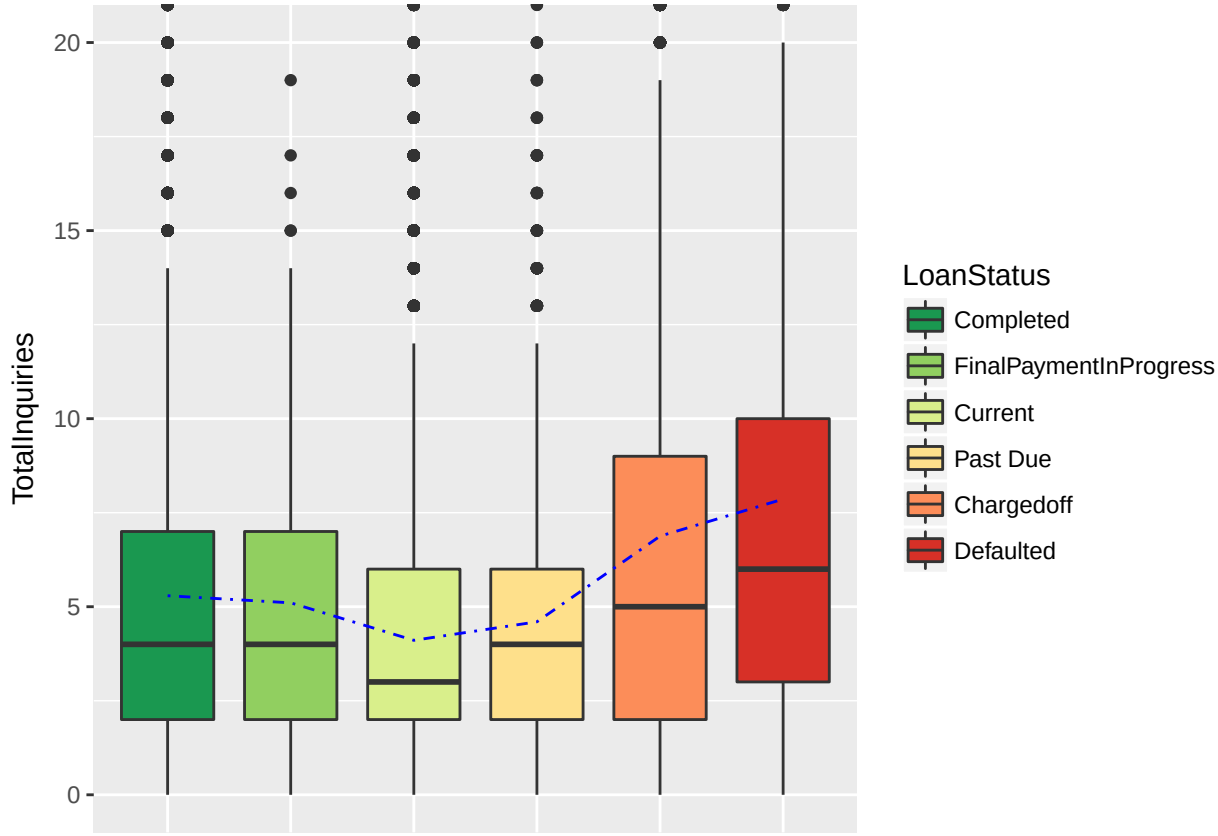
LS V ABCC



Here, we can see that ABCC by LS have similar 1st quantile levels. However, Good Loans clearly have higher medians (~5000 v ~2500), and have significantly higher 3rd quantiles (~14-15k v ~8-10k). The mean line also shows a clear dip between Good and Bad Loans.

Thus, we can assert that there is a *positive correlation* between ABCC and LS.

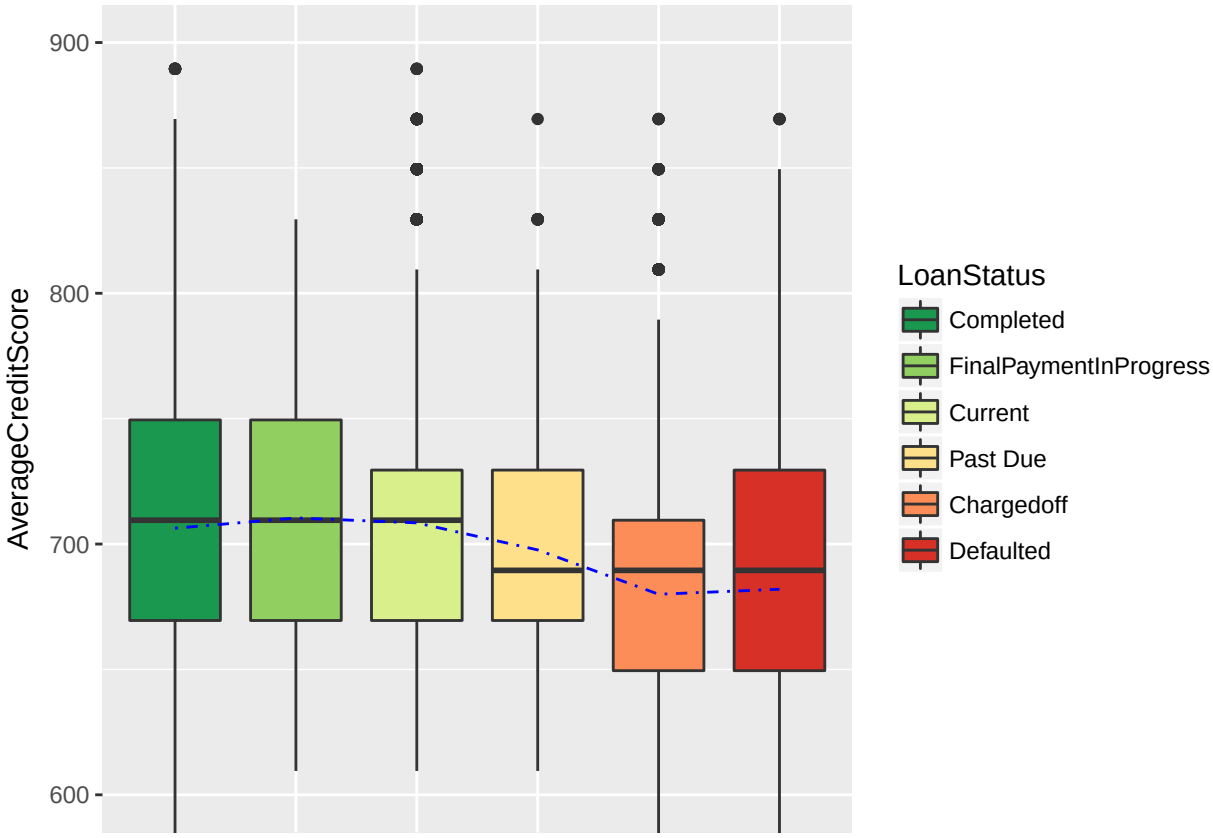
LS V TI



We can see a clear dramatic change in the average number of TotalInquiries between Good and Bad loans. Good loans are an average of 4-6 inquiries, but ChargedOff and Defaulted loans have much higher averages of around 9 and 11. The

Thus, we can assert that there is a *negative correlation* between TI and LS.

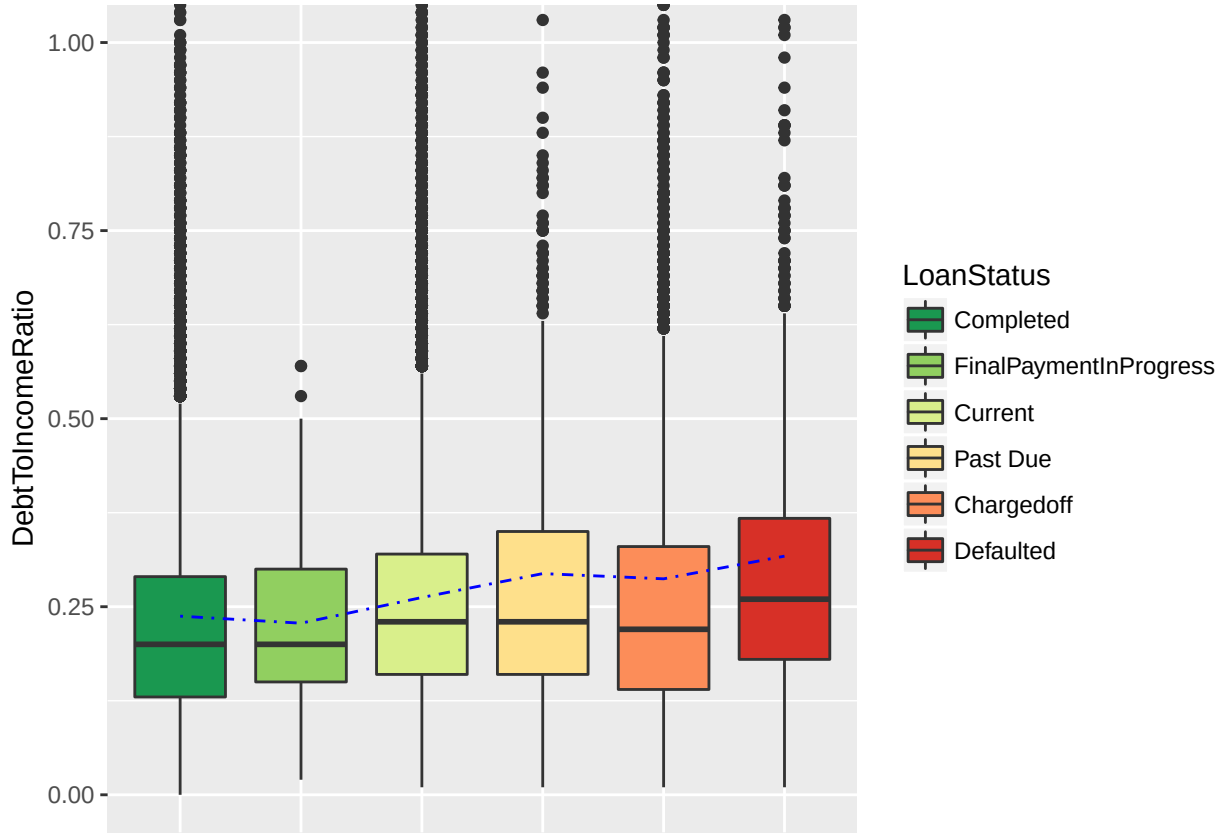
LS V ACS



Here, we can see a clear difference between ACS by LS. Bad loans have a clear lower median ACS <700, whereas Good loans have a median ACS >700. The mean line closely follows these medians. It is clear that there is a *positive correlation* with ACS and LS.

It is interesting to note that Final, Current, and PastDue loans do not have any borrowers with an ACS below 600. This implies that more recent Prosper loans only accept borrowers with an ACS of >600.

LS V DIR



Here we can also see a difference in average DIR between Good and Bad Loans. While they are all distributed quite similarly, more Bad loans have higher DIRs. It must be noted however that the difference is only in tenths of a percentage. Thus, we can assert that there is a *slight negative correlation* with DIR and LS.

Bivariate Analysis

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?

Our response variable is LS, and thus is the main focus of this report. Plotting LS against other explanatory variables, we were able to establish that strong correlations with the following:

LOA/MLP - Surprisingly, these appeared to have no correlation with LS.

T - 36 month terms definitely has the highest correlation to Bad Loans. However, 12 month and 60 month terms both have higher GBRs, despite being on opposite sides of the spectrum. I hypothesize that 12 month terms force borrowers to plan properly due to the short nature of the loan, and 60 month terms give borrowers more time and flexbility, hence both term amounts having high GBRs. It must be noted however, that 12M loans are mostly Completed, whereas 60M are mostly Current. Thus, they cannot be directly compared, as 60M loan data could change in the future.

OCC - students and borrowers with low-skilled jobs were correlated with Bad Loans. Interestingly, realtors and commissions-based salespeople are an anomaly, as they also have high rates of Bad loans, despite having higher ACS and SMI.

BS - borrowers from ND, ME, and IA were correlated with Bad Loans.

ES - borrowers with the "Employed" label had the highest GBR.

SMI - positive correlation, Good Loans are associated with higher SMI

ABCC - positive correlation, Good Loans are associated with higher ABCC

TI - negative correlation, Good Loans are associated with lower TI

ACS - positive correlation, Good Loans are associated with higher ACS

DIR - negative correlation, Good Loans are associated with lower DIR

Interestingly, home ownership has little effect on LoanStatus. I was theorizing that if a borrower already owned a home, they would be more responsible with debt. This does not seem to be the case.

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

The initial correlation matrix found correlations between secondary quantitative variables, notably MLP-LOA, ACS-ABCC, APR-ACS. These correlations will be analyzed across LS in the multivariate analysis section below.

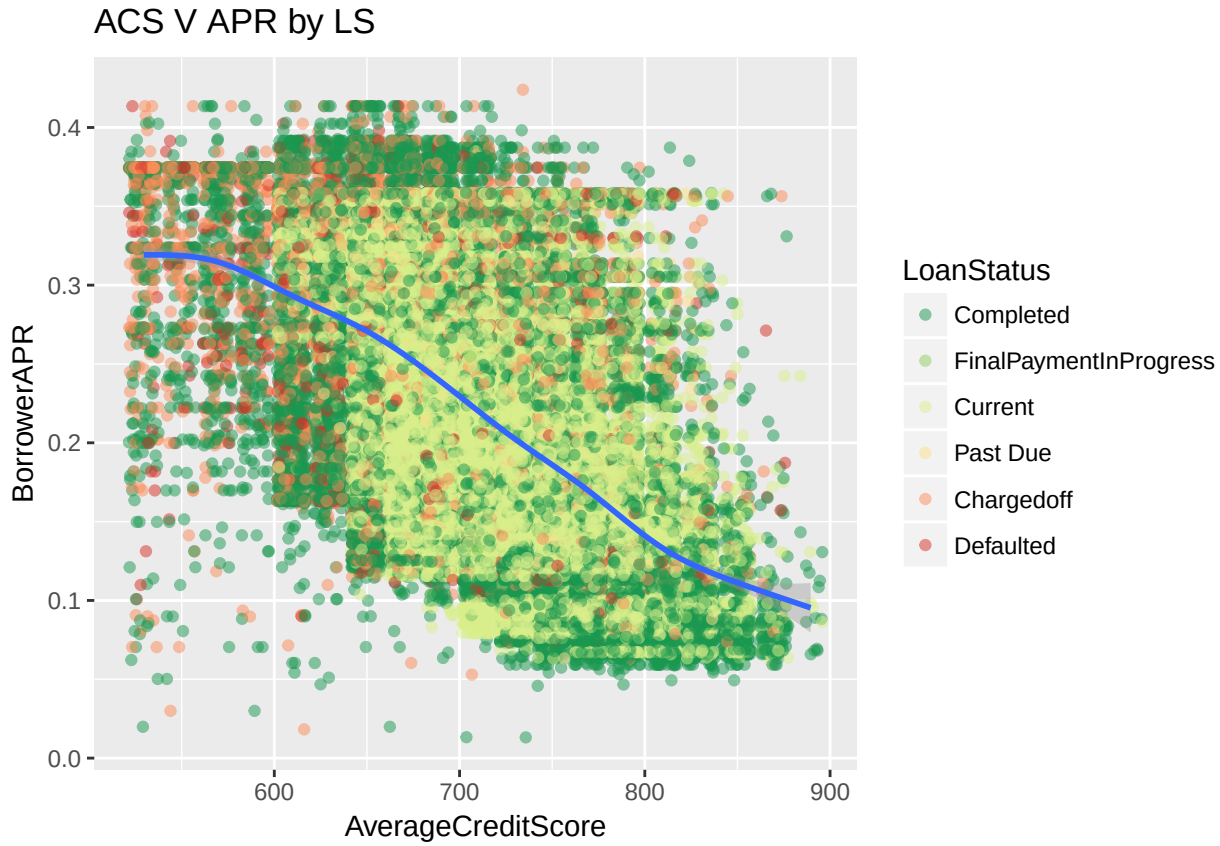
OCC and BS were also plotted across average SMI and ACS, and both OCC and BS were found to have strong correlations. Occupations and States with higher incomes and credit scores were far more likely to also have higher GBRs, and vice versa.

What was the strongest relationship you found?

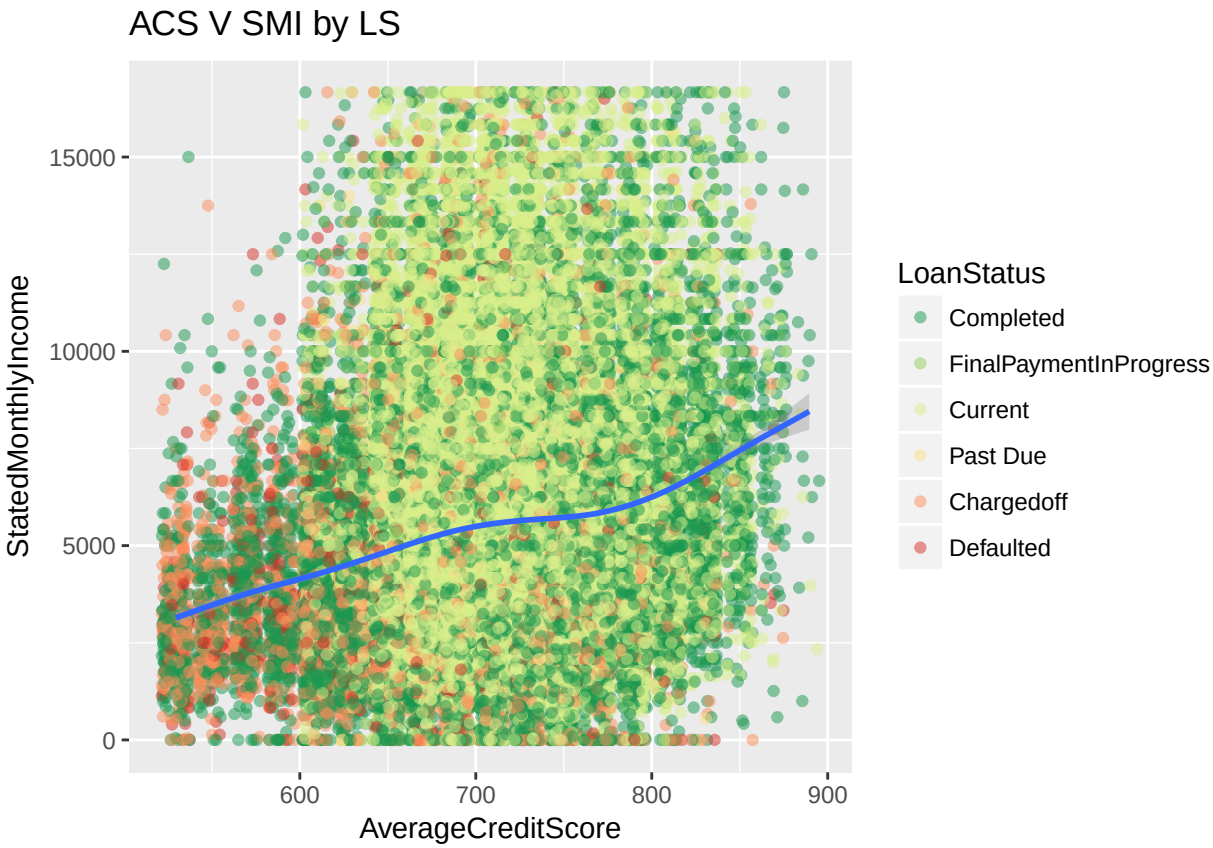
There are many strong relationships between LS and the other explanatory variables. However, I suggest that the strongest relationship was between *LS and ACS*. The qualitative variables do not alone indicate whether a loan will go bad (e.g. the BorrowerState does not directly affect LS, it was found to be SMI and APR); they are linked to other quantitative factors. And out of all the quantitative factors, ACS was chosen, for two reasons:

1. it is a number calculated directly by actions that measure the fiscal responsibility of a borrower
2. there are a minimal number of outliers in this plot compared to others, and yet we can see a clear difference in both mean and median between Good and Bad Loans.

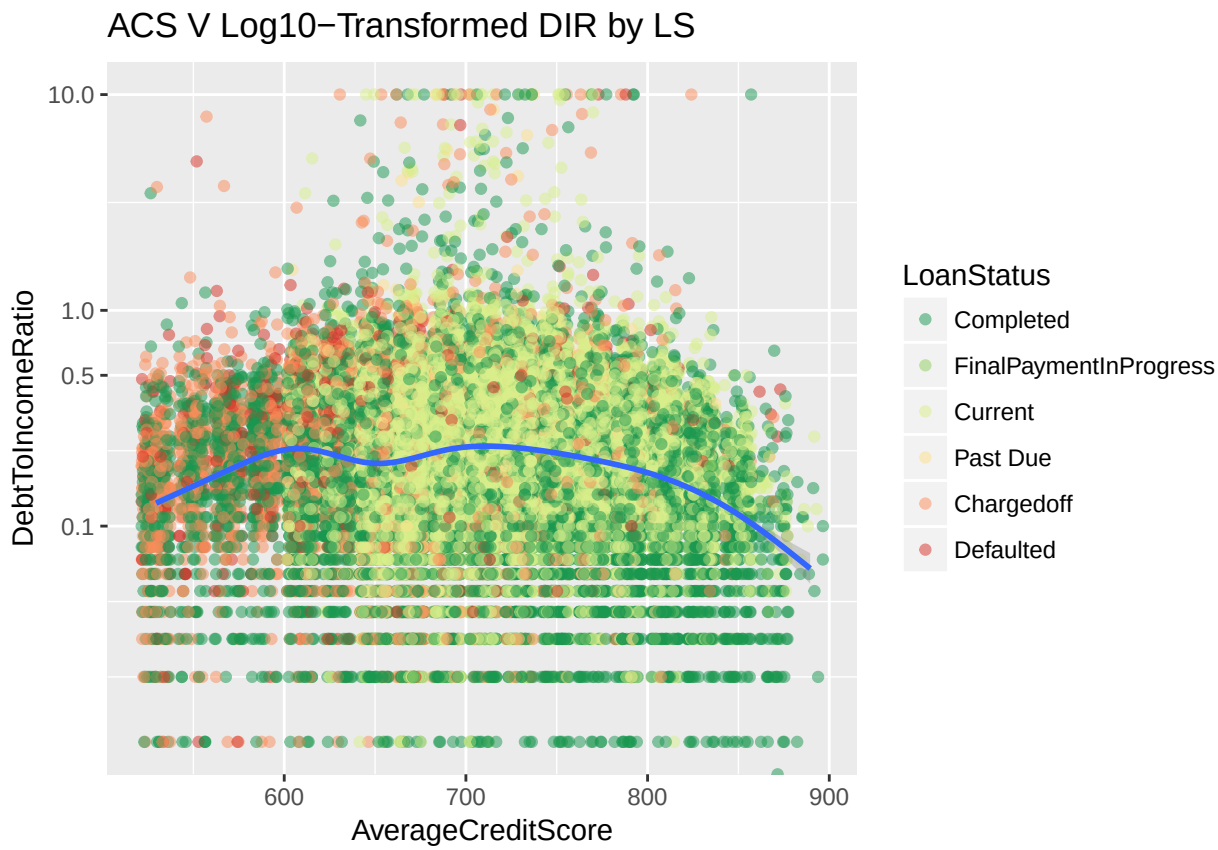
Multivariate Plots Section



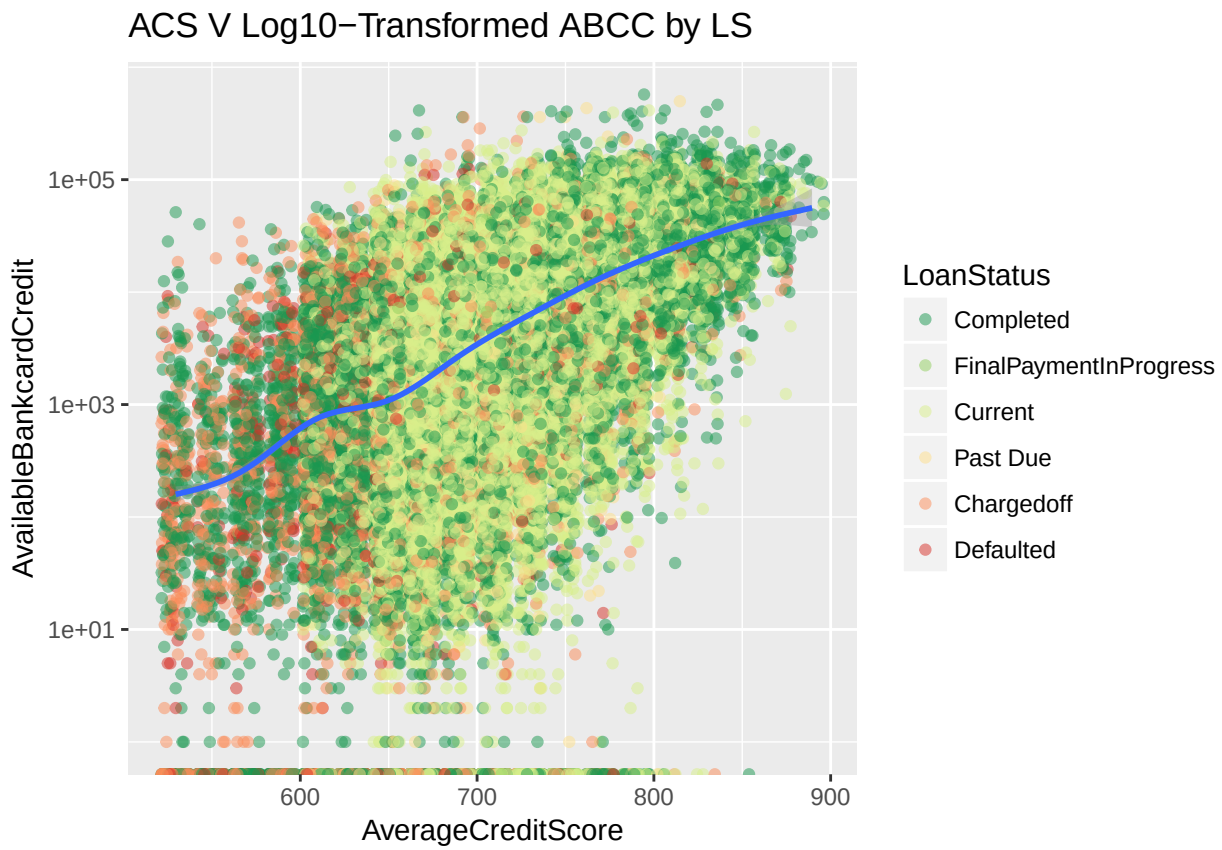
ACS and APR have the strongest negative correlation among all quantitative variables in this report, and thus it is plotted first. Here we can clearly see the negative mean line showing an overall decrease in APR as ACS increases. We can also see there are many Bad Loans clustered among APRs above 0.25 and under 650. However, while it is not extremely clear, we can still see Bad Loans dispersed among Good Loans in higher ACS values, as long as APR is high too.



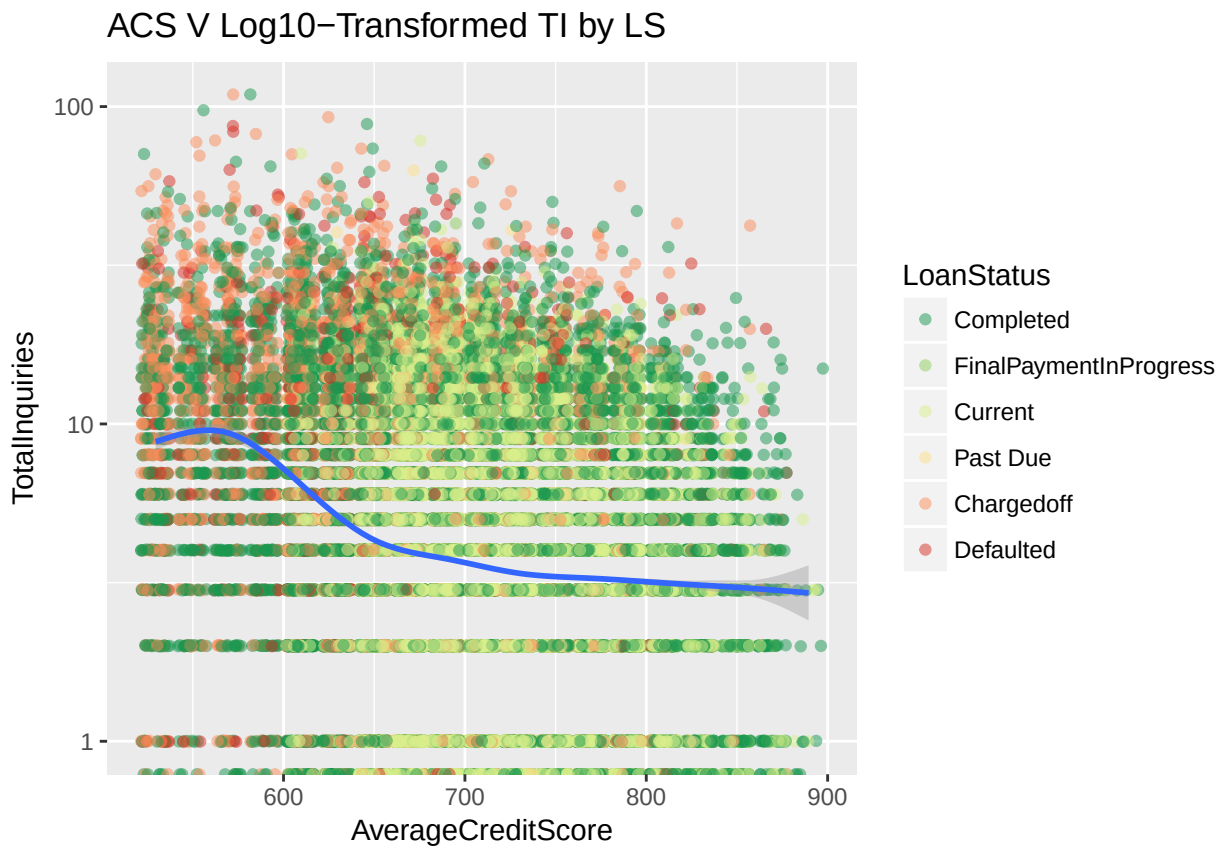
We see a similar positive correlation between SMI and ACS. Borrowers with low ACS values typically also have low SMI, and vice versa. However, the dispersion of Bad Loans seems to be distributed more widely than in the ACS-APR scatter plot around ACS values of 650, with Borrower SMI ranging across the whole income spectrum.



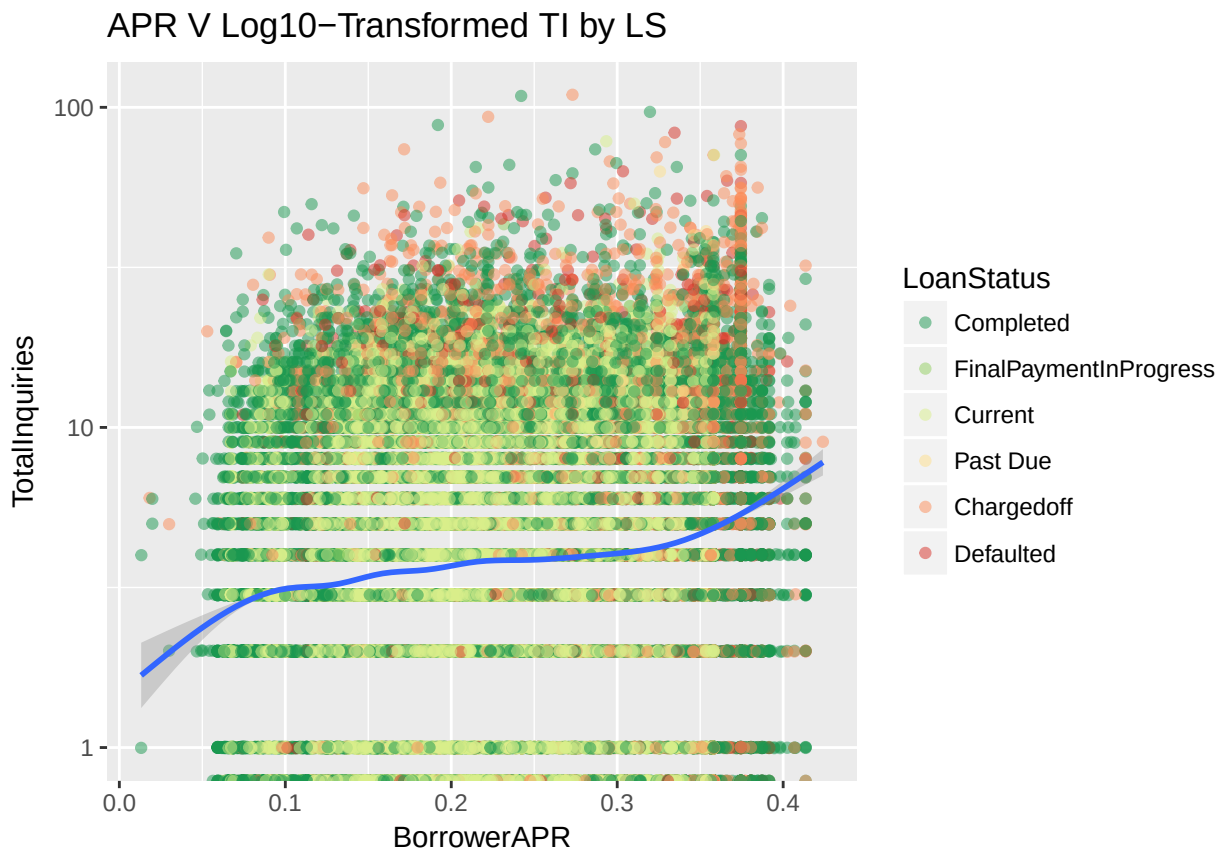
DIR only appears to matter up to 1.0. Beyond that, there are few borrowers, and there appears to be no discernable pattern in GBR. For data points with DIR between 0 and 1, the dispersion of Bad Loans again appears to be distributed fairly widely, with no distinct pattern (excepting the cluster of Bad Loans where ACS < 600). In fact, the mean line actually rises initially, then dips down at higher ACS levels.



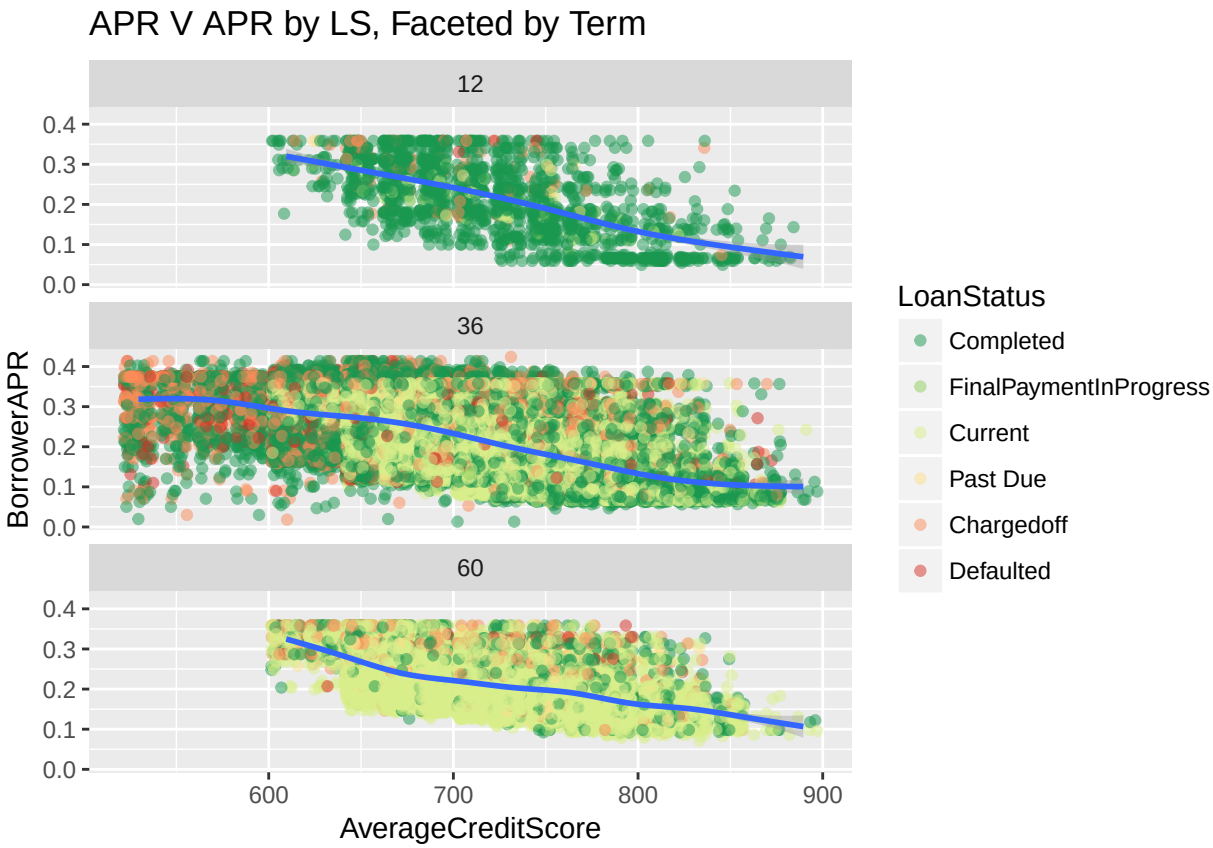
Here, we can see the strong positive correleation between ABCC and ACS. Unfortunately, there also appears to be a large spread of Bad loans among all ABCC levels.



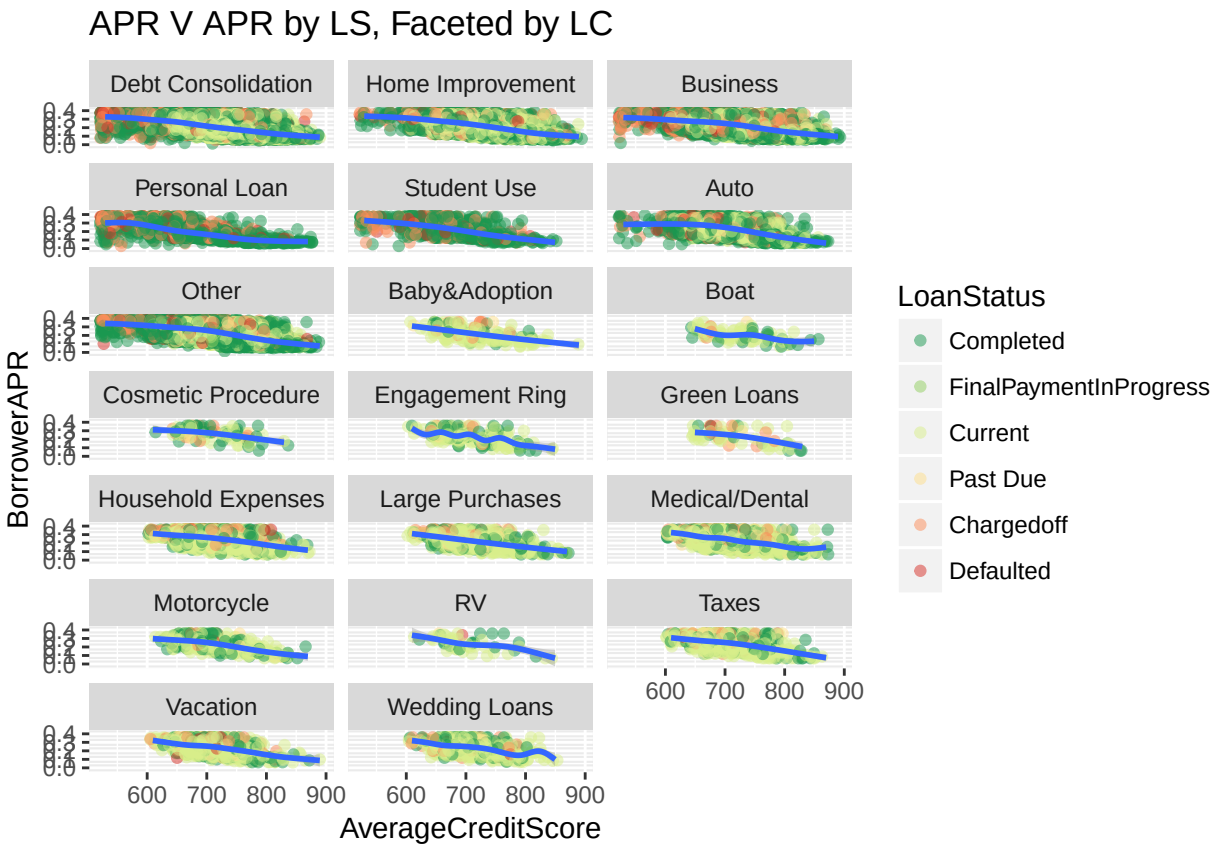
Here we can begin seeing big differences in LS. While there are Bad Loans for borrowers with ACS < 600, high TI values (> 10) also have high instances of Bad Loans, even when ACS is higher.



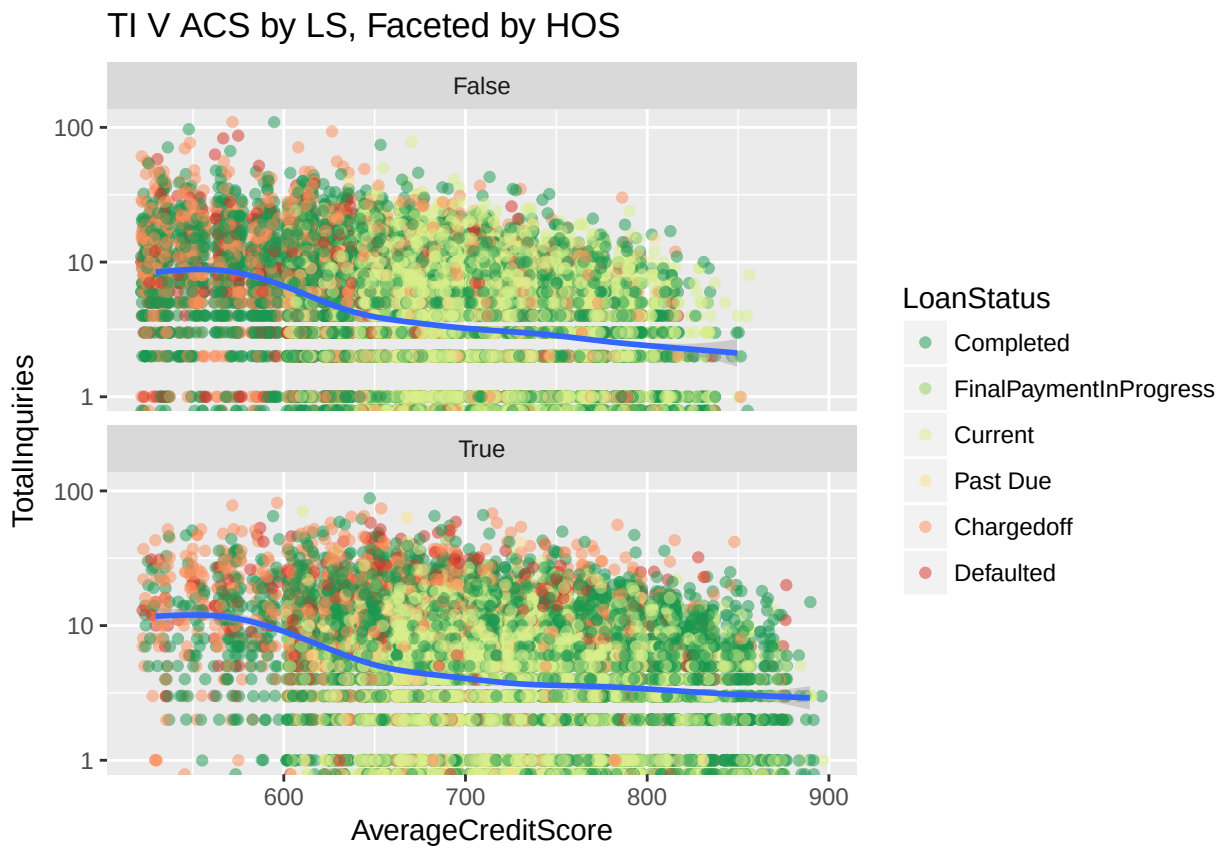
Plotting by APR with TI shows the same separation in LS. There are clearly more Bad Loans for borrowers at higher APR levels, but the instances of Bad Loans increases dramatically when TI is >10, even with low APR.



Faceting by Term also reveals differences in LS. 12M and 60M are mostly Good Loans (Completed and Current, respectively), with a few Bad Loans at higher APR levels. 36M Loans thus clearly hold the majority of Bad Loans, with the majority having higher APR levels.



Faceting by LC affirms the previous findings in the bivariate analysis. Business and Personal loans in particular have high amounts of Bad Loans at higher APR and ACS levels, hence the strong correlations in the Chi-square plot.



Here we can see another interesting pattern. We already have discovered that high TI has more Bad Loans, but faceting by HOS reveals that homeowners have more Bad Loans at higher ACS scores. Non-homeowners have the majority of Bad Loans isolated in lower ACS levels. With this plot, we can argue that homeowners actually have a *stronger negative correlation* with LS than non-homeowners, since high ACS does not preclude these borrowers from having Bad Loans.

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?

Almost all these scatter plots used ACS in the X axis, and thus we could see a consistent pattern of low GBRs at low ACS scores. Generally speaking, higher ACS would lead to higher GBR. We attempted to see if the Y axis variable would alter the pattern of high ACS borrowers in any meaningful manner.

Many variables that were found to have positive correlations with LS were not quite as distinctive once laid out on a scatterplot. ABCC, DIR, and SMI all had correlations with ACS, but unfortunately there were no clear patterns at higher ACS levels. All the Good and Bad Loans were spread across the spectrum.

The strongest pattern found was with TI. High numbers of inquiries seemed to constantly link with more Bad Loans, no matter the ACS or APR level.

LC and HOS were the only qualitative variables analyzed in this multivariate analysis, mainly because the other variables had far too many factors to plot cleanly. Furthermore, we already

Were there any interesting or surprising interactions between features?

There a few variables that, upon further analysis, did not give the results I was anticipating.

DIR was a surprise. Logically it would make sense that higher DIR would be linked to low ACS (as initially hypothesized in the bivariate analysis), thus creating a linear model. However, mean DIR actually rises at low ACS levels, then stablizes at 0.5, before dipping back down at high ACS levels. At first I thought that the log10 transformation removed too many variables, thus changing the correlation pattern, but removing the log10 showed the same pattern. I believe my incorrect hypothesis was due to a visual misrepresentation of the bivariate analysis, as all the LoanStatuses were distributed very similarly, with differences being only in tenths of a percent.

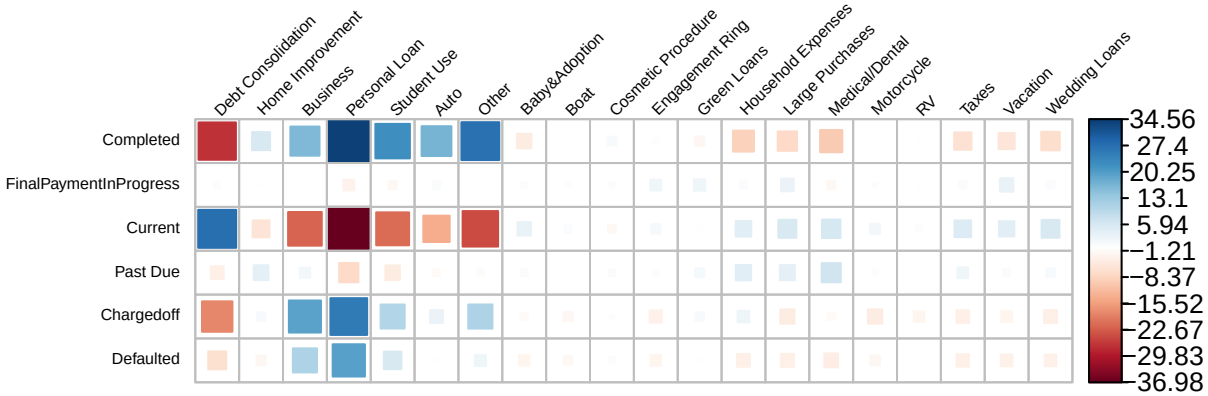
SMI and ABCC was also unexpected. We would normally assume that borrowers with higher income and more credit would typically have a lower amount of Bad Loans, but this assumption looks to be wrong. Higher income and bank credit does not mean a borrower will pay a loan off successfully.

HOS was also a surprise variable. At first, I expected it to have a big impact, but the bivariate analysis presented similar GBRs for borrowers, regardless of home ownership. However, cross referencing with ACS and TI shows that borrowers with homes actually had *Bad Loans at higher ACS and TI numbers*, whereas the majority of Bad Loans for non-homeowners were mostly limited to *low ACS* and high TI numbers. Perhaps homeowners in all credit score ranges would be more inclined to make more inquiries for whatever reason, maybe to purchase more homes, or participate in other investing activities.

Final Plots and Summary

Plot One

Chi-Square Plot - LoanStatus V ListingCategory

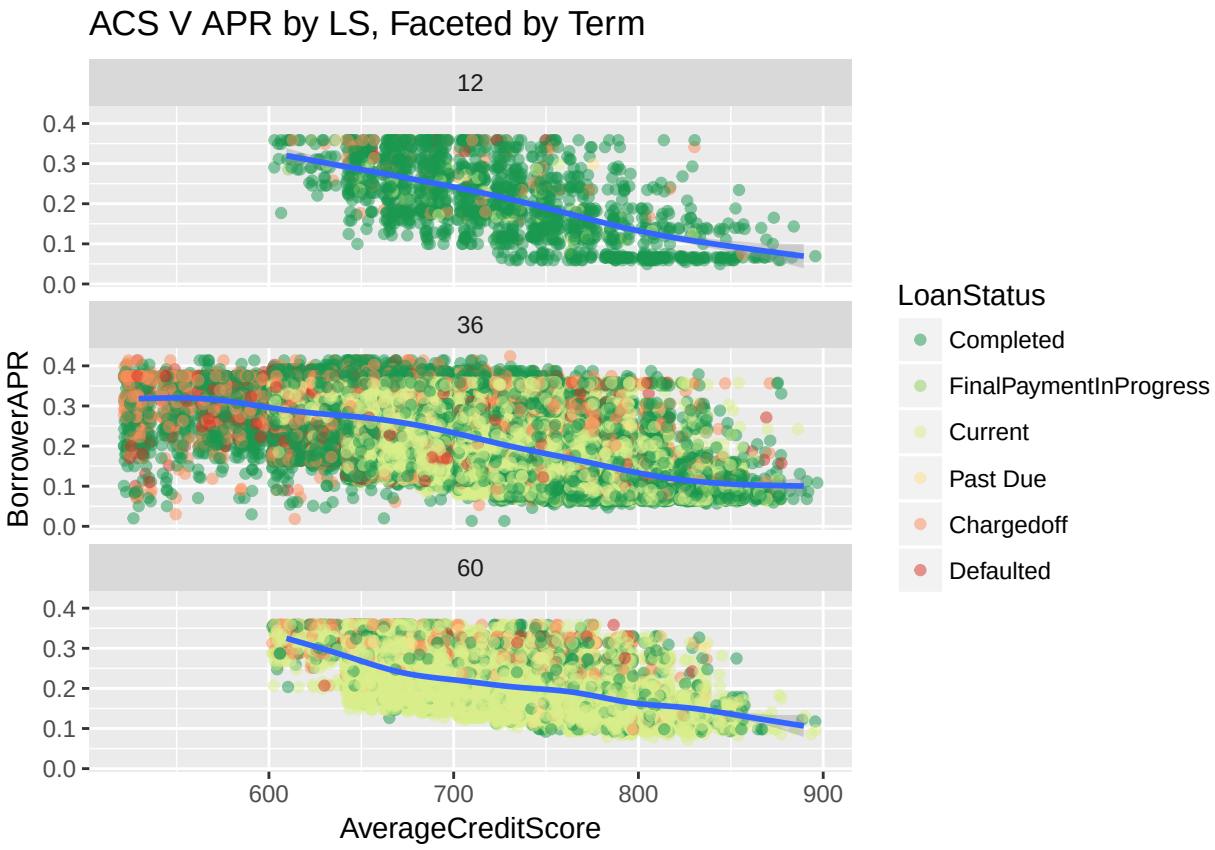


Description One

There were many Chi-square plots created in this report. They are simple, and clearly visualize relationships between variables. I chose the LC Chi-square plot over other qualititative variables specifically for the same reason I chose ACS as the primary quantitative variable: *it reflects the fiscal responsibility of a borrower.* This cannot be said for BS, ES, and OCC, the other qualitative variables.

Borrowers getting a Loan for Debt Reconciliation are interesting in paying down their loans and having a potentially lower interest rate. This implies that they are more responsible with their finances than someone getting a Personal Loan, which could cover anything from helping a friend to making an unnecessary purchase. We can see that "Other", Personal and Business Loans are more likely to be Bad Loans, due to their uncertain and riskier nature.

Plot Two



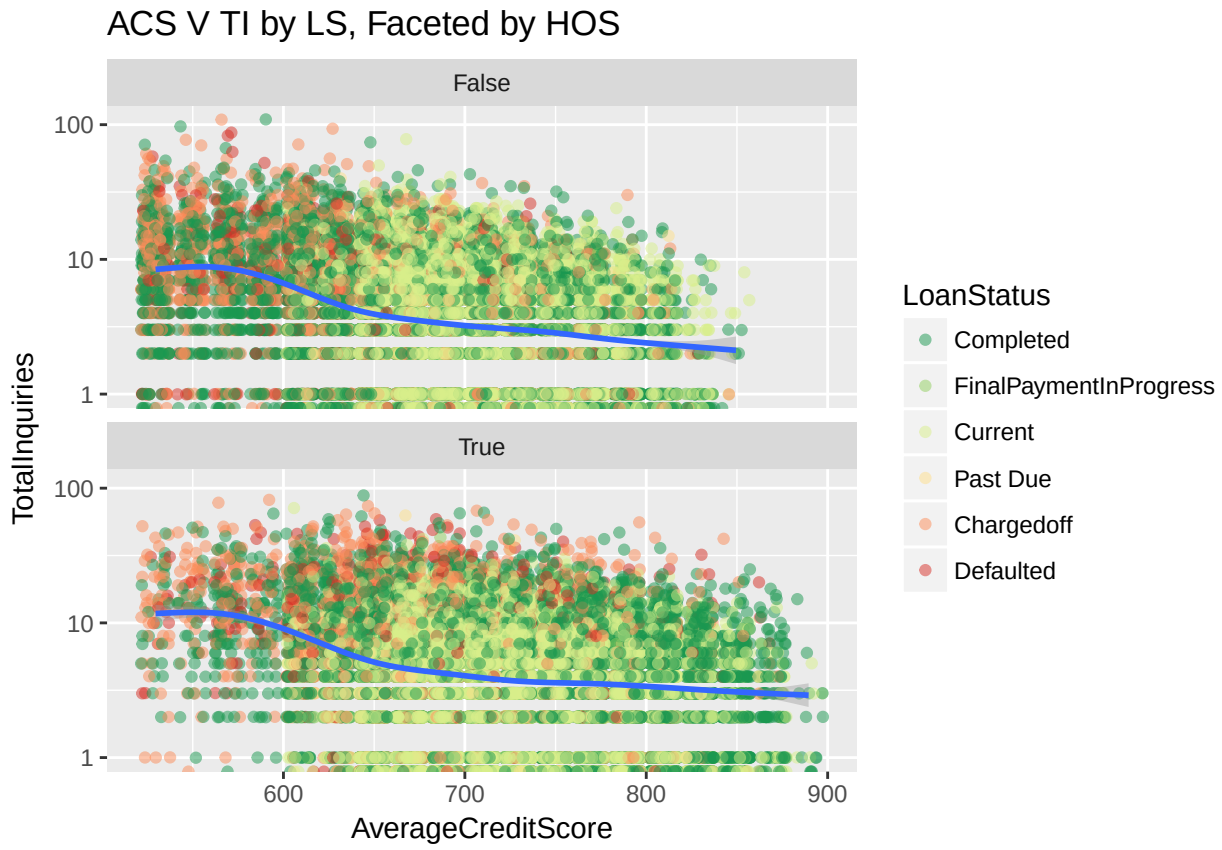
Description Two

In this plot we can clearly see three main conclusions gleaned from the univariate and bivariate analyses:

1. ACS's positive correlation with LS - the higher the credit score, the higher the amount of Good Loans (and consequently the lower the amount of Bad Loans)
2. APR's negative correlation with LS- higher APR leads to less Good Loans (and consequently more Bad Loans). Note that Bad Loans do exist at higher ACS levels when APR is also high.
3. Term's effect on LS

- 12M Loans are mostly Completed
- 60M Loans are mostly Current
- most Bad Loans have a 36M length

Plot Three



Description Three

In this plot we can clearly see three main conclusions gleaned from the univariate and bivariate analyses:

1. TI's negative correlation with LS - more inquiries generally means less Good Loans (and consequently more Bad Loans)
2. APR's negative correlation with LS- higher APR leads to less Good Loans (and consequently more Bad Loans). Note that Bad Loans do exist at higher ACS levels when APR is also high.
3. HOS's effect on LS:

- fairly equal number of borrowers who do and do not own homes
- Homeowners tend to have more Bad Loans with high TI values *despite* high ACS

Thus, we can conclude that *Homeowners with high TI numbers* have a *stronger negative correlation* with LS than Non-Homeowners.

Reflection

The null hypothesis was stated in the beginning of this report.

"There is no relationship between a variable and a borrower's ability to pay off a loan."

My report clearly illustrates that this is not the case. We actually have the opposite problem, where there are *infinitely many* relationships between variables and a borrower's ability to pay off a loan. The Prosper dataset was a very interesting project to tackle, with many possible avenues. I am happy with my choice of analyzing what factors would affect a borrower's ability to pay back debt, since this is an issue that 99% of people in the world would encounter in their life.

Isolating the variables analyzed in this report is a very important first step which I initially failed to take. In doing so, I was constantly distracted with other factors that should not have been a part of this project. I am still not sure if using LoanStatus as the response variable was a mistake, due to its qualitative nature. Perhaps using quantitative variables would provide more concrete or accurate conclusions.

On the other hand, the Chi-square plots appeared to very clearly define which factors were more relevant, and which were not. This was especially helpful where there were large numbers of qualitative factors (e.g. BS, LC). However, it is difficult to reorder chi-square plots, especially by correlation.

I feel that I correctly chose most variables, but one I should pursue next time is LoanCreationYear. Loans were not issued out equally, and Prosper clearly adjusted their policies as the company matured. Perhaps filtering by year would be an interesting direction to take the report, since what financially affected people 10 years ago would be somewhat different from the financial issues people face today.