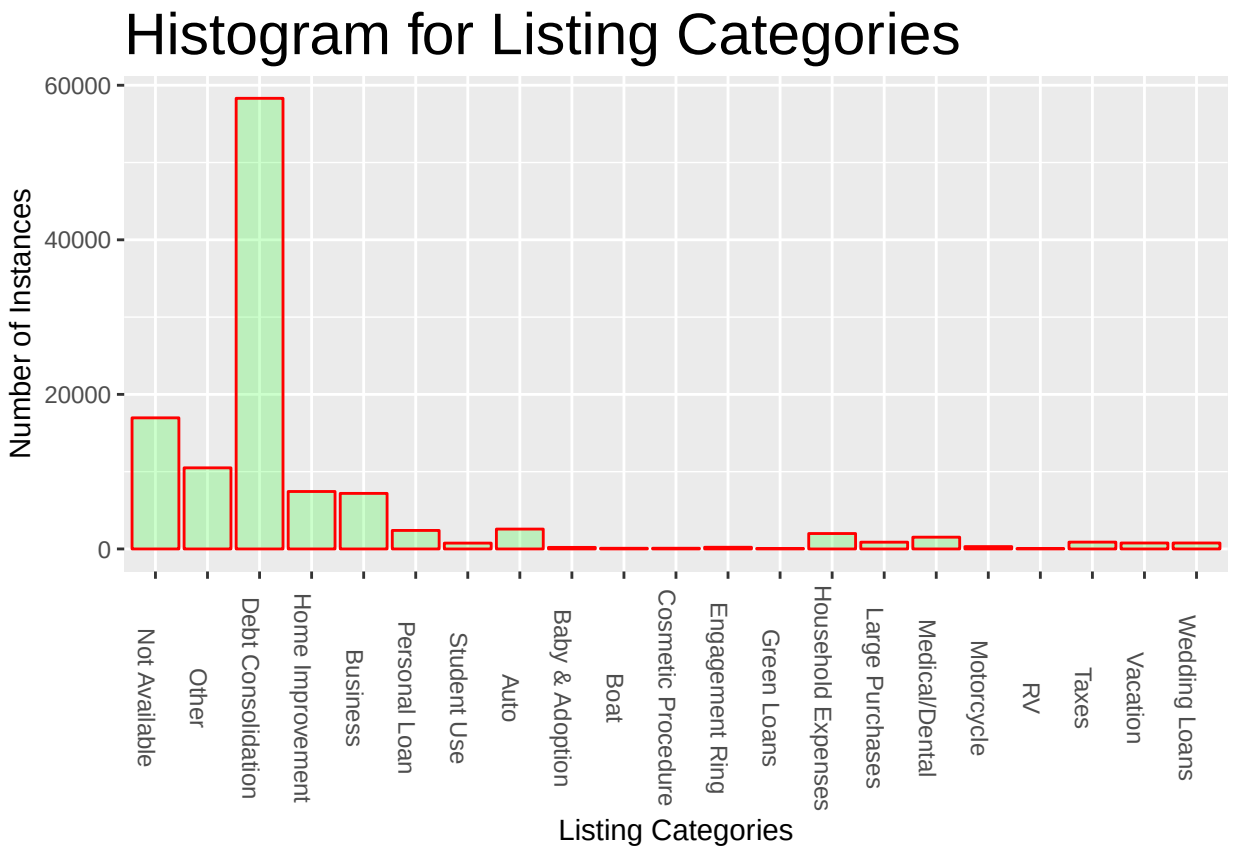
PROSPER LOAN DATA ANALYSIS BY AKASH DUTTA

Prosper is a San Francisco based company where people can invest in personal loans or request to borrow money. What is interesting here is that it has a peer-to-peer lending process i.e. the company itself does not loan out the money but rather connects the borrower to the lender. This is a innovative approach and benefits its customers when compared to loan processes in various traditional banking institutions.

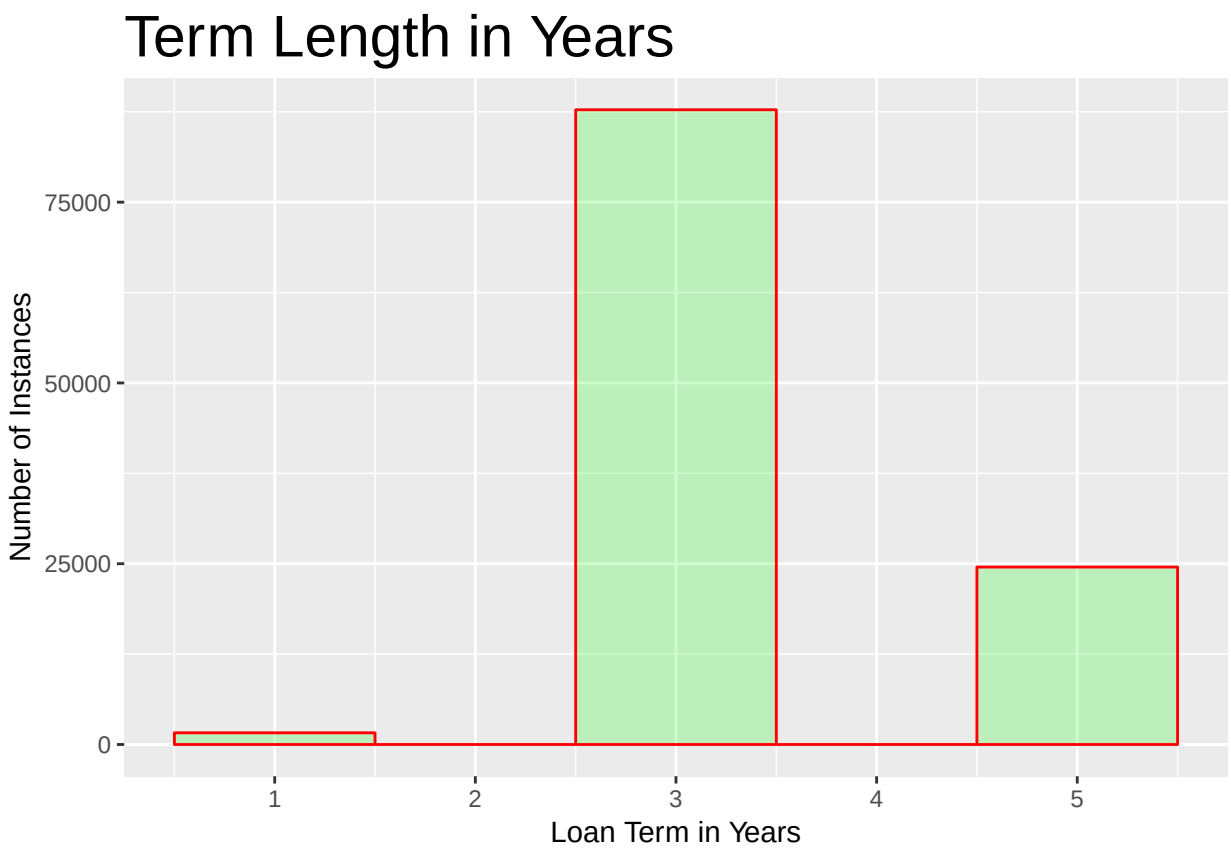
The dataset we have here is immense. It encompasses all the various data points considered when a loan is processed. I will attempt to deconstruct this vast .csv file to convey understanding of the data in a much more lucid way than scrolling through the many instances of loans provided in this dataset.

Do not complain about your data. DO MUNGING

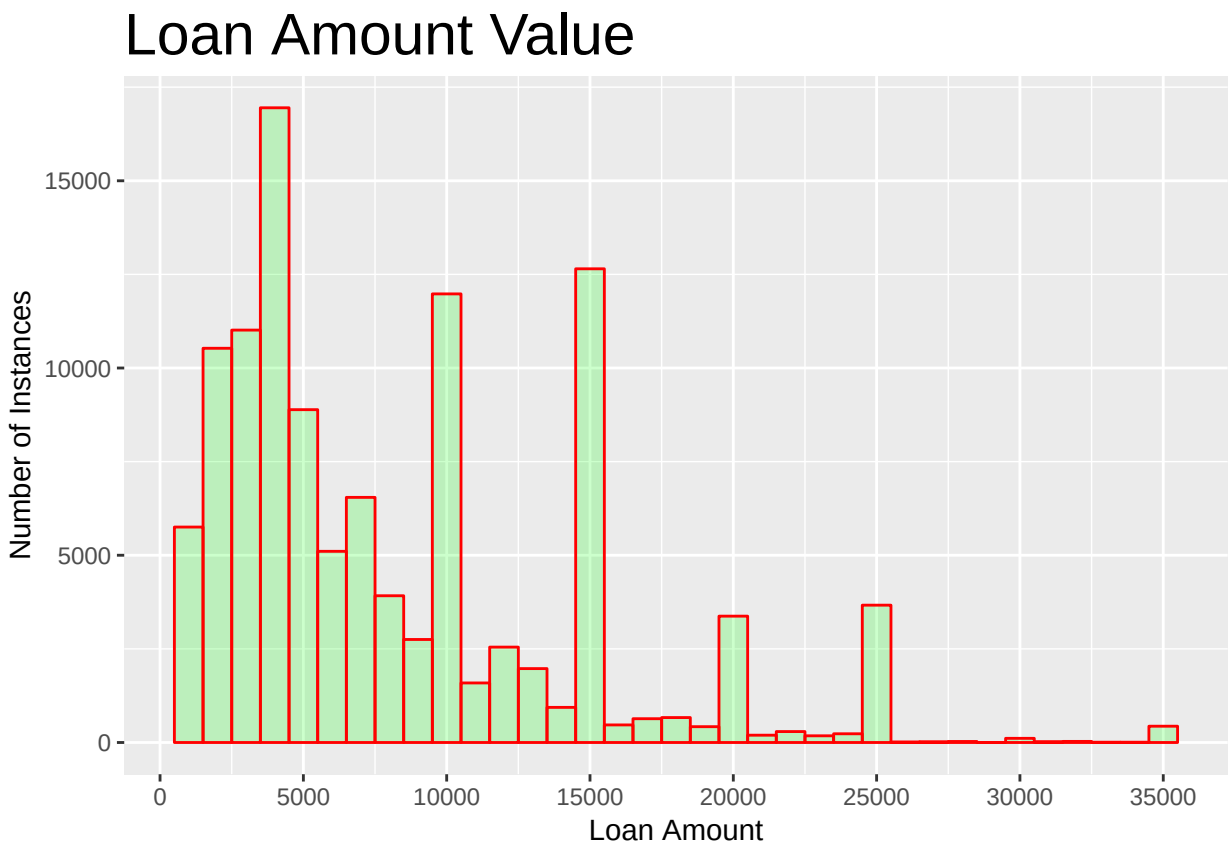
VAMOS!!!! Let us explore our data set



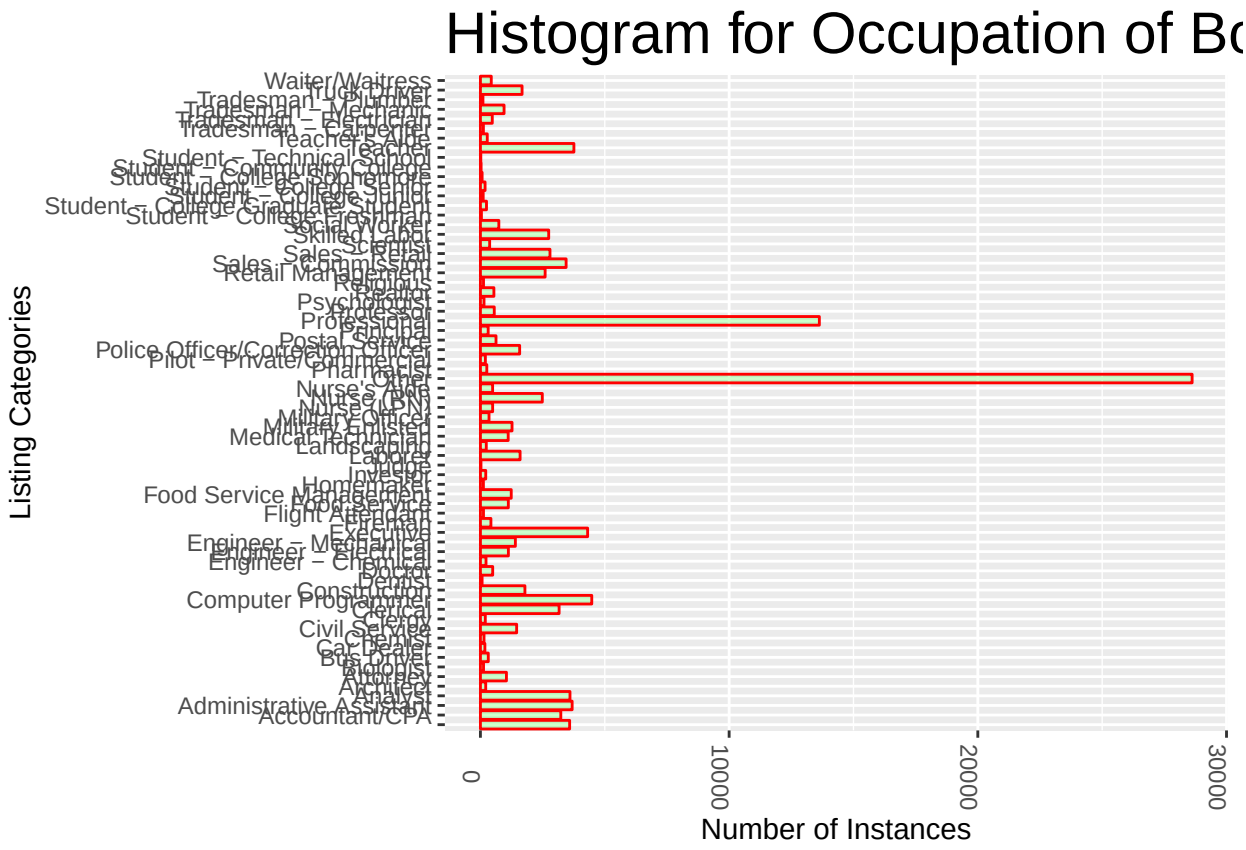
Why do People take Loans? Not surprisingly most instances belong to Debt consolidation and they lead the chart by a huge margin. For those who are unaware Debt Consolidation is to take a loan to pay another loan and yes it really exists. Business and Personal loans are other notable shareholders in the spread of Loans by Listing Categories

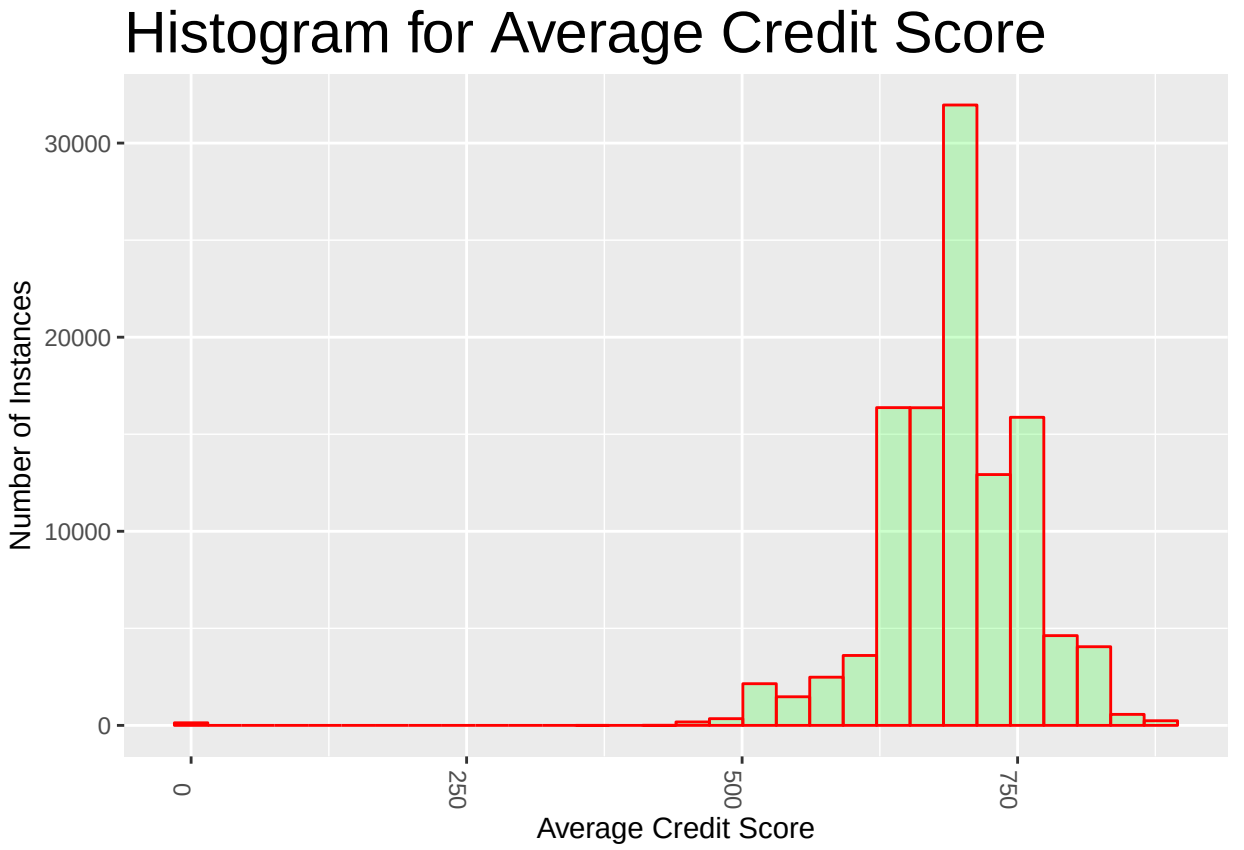


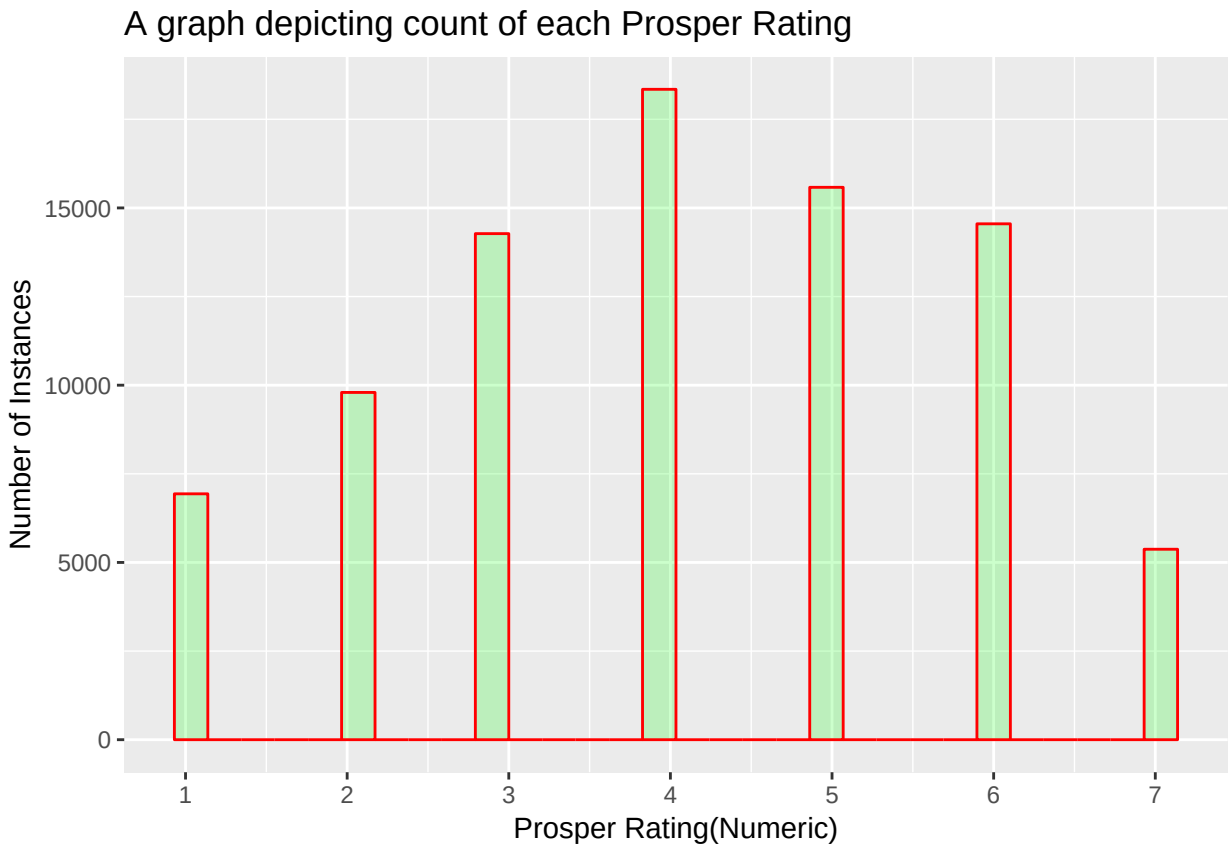
What is the most common time period for Loans? This graph was predictable. Most people take loans for a 3-year time period followed by 5-year time period. Very few loans are granted for a 1-year period. Reasons for this may be the low ROI from the investors perspective.

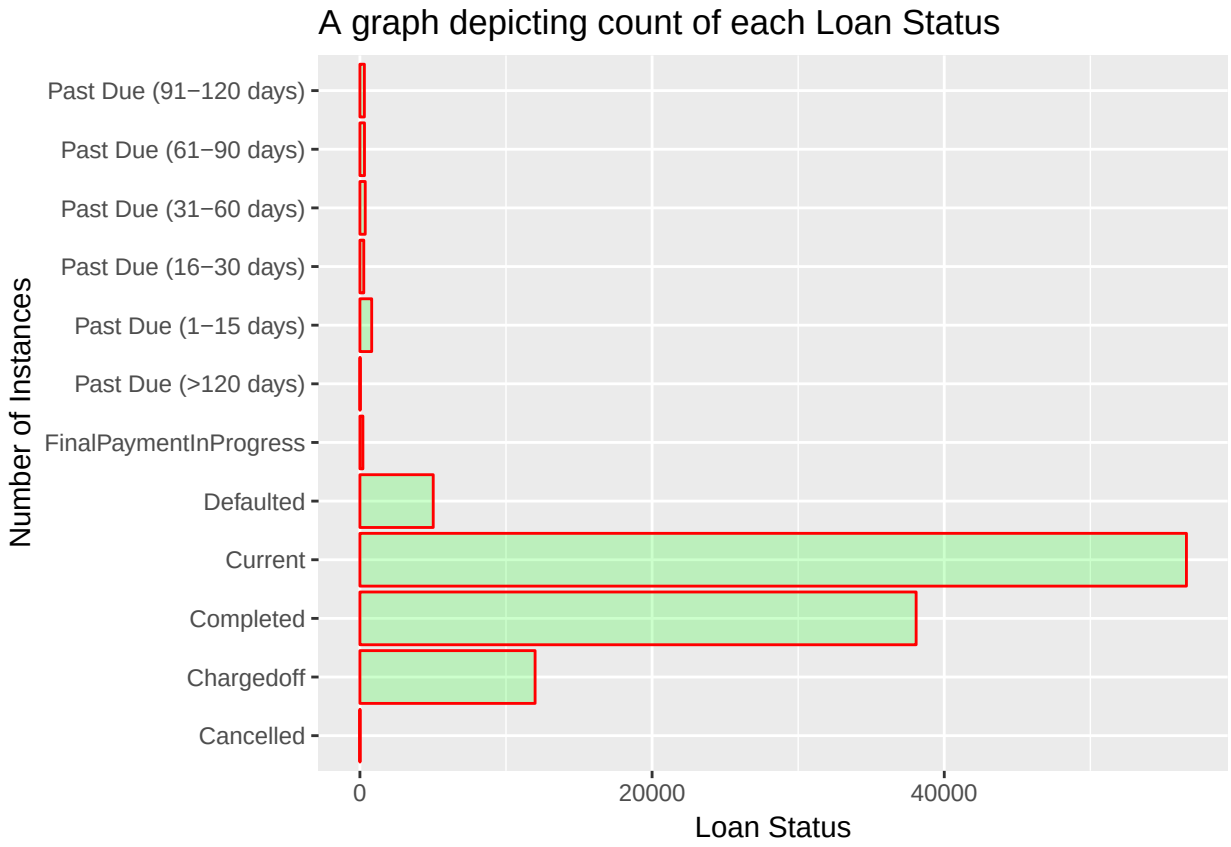


How much Loan do people take on an average? Prosper, being a peer-to-peer loan company, has a rather low median of Loan amount. We can see that the mode Loan value is $5000. The graph forms nice peaks on round values which is again an expected occurence.


Histogram for Occupation of Bo
Listing Categories
Waiter/Waitress
Police Officer/Correction Officer
Pilot – Private/Commercial
Food Service Management
Computer Programmer
Civil Service
Car Dealer
Bus Driver
Administrative Assistant
Accountant/CPA
0
10000
20000
30000
Number of Instances

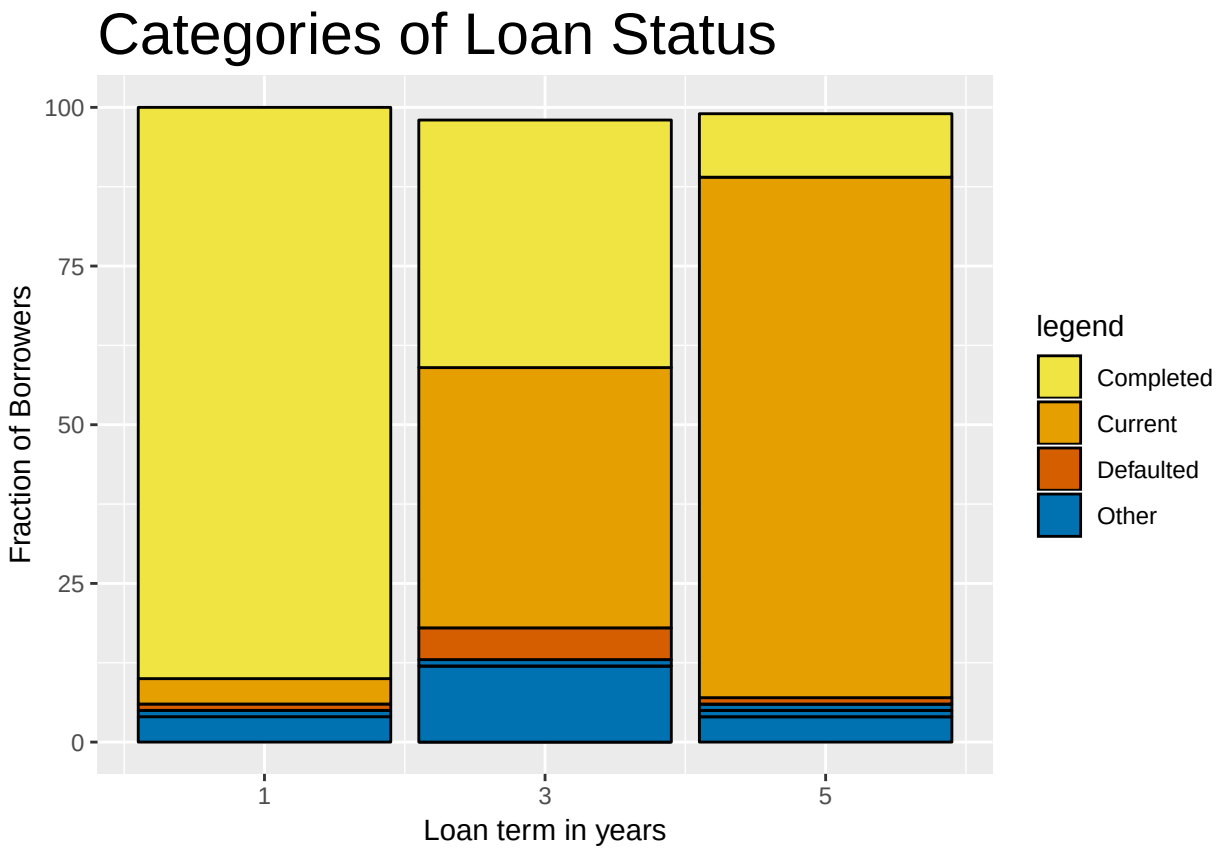

Histogram for Average Credit Score
Number of Instances
30000
20000
10000
0
0
250
500
750
Average Credit Score


A graph depicting count of each Prosper Rating
Number of Instances
15000
10000
5000
0
1
2
3
4
5
6
7
Prosper Rating(Numeric)

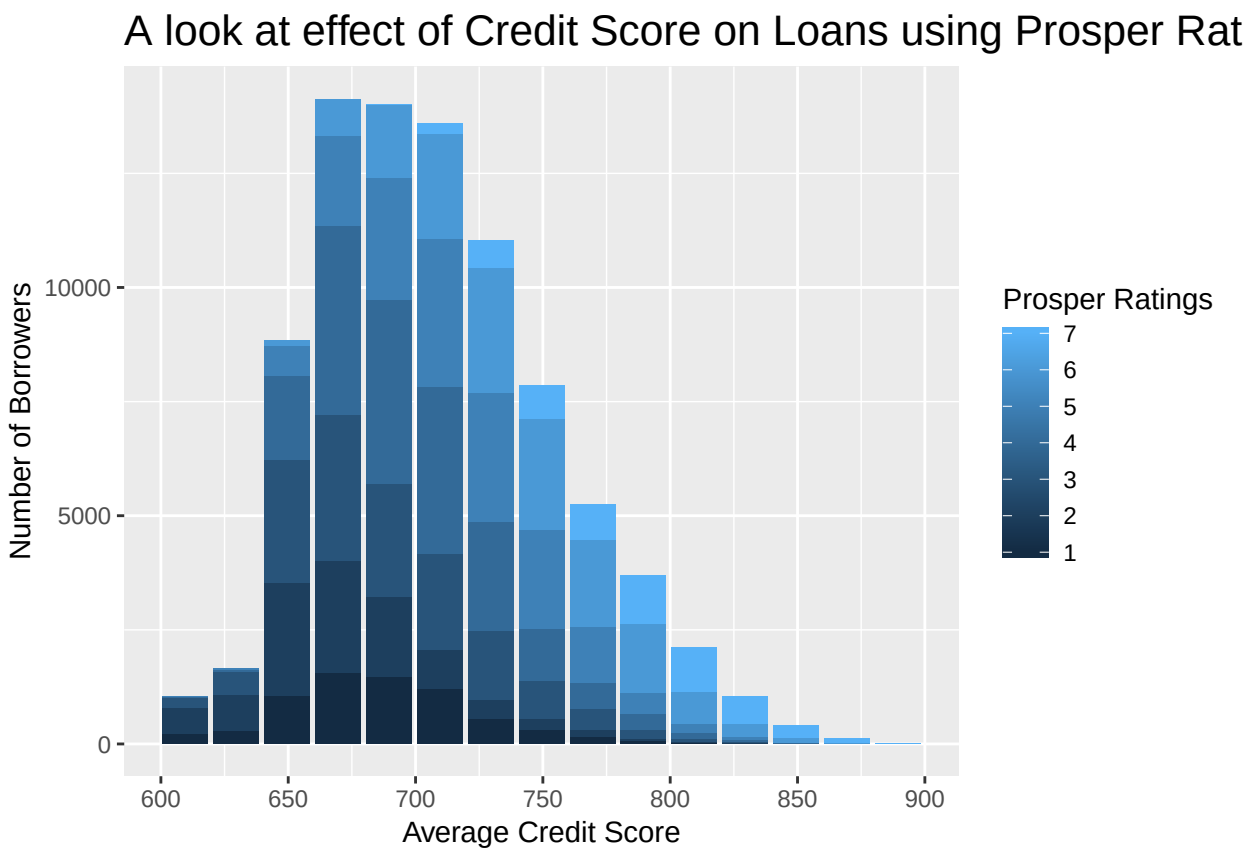

A graph depicting count of each Loan Status
Number of Instances
Past Due (91−120 days)
Past Due (61−90 days)
Past Due (31−60 days)
Past Due (16−30 days)
Past Due (1−15 days)
Past Due (>120 days)
FinalPaymentInProgress
Defaulted
Current
Completed
Chargedoff
Cancelled
0
20000
40000
Loan Status

People engaged in which occupation take loans? People are probably unwilling to list out their occupation and that is why we see two major peaks at “Professional” and “Other”. Notable peaks in the rest are Teachers, Computer Programmer and Executives.

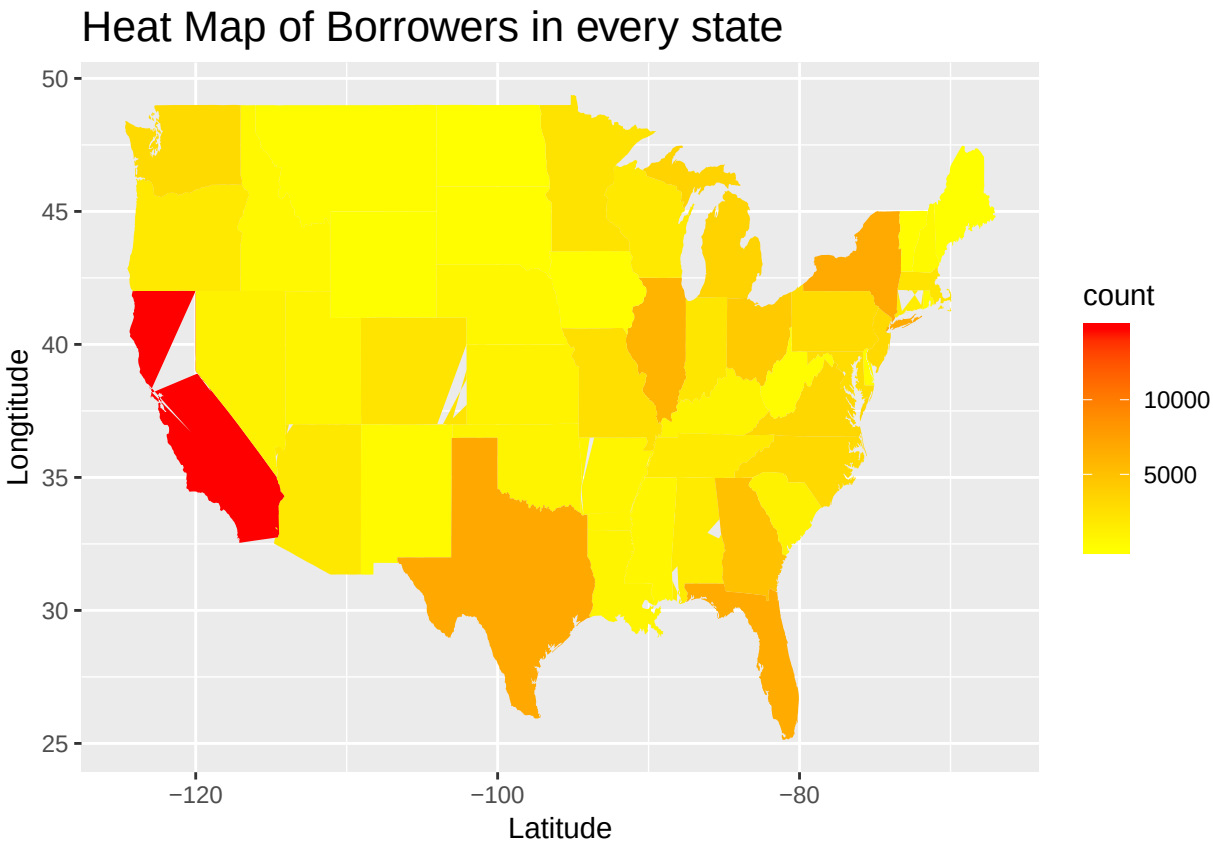
Let us now expand our approach.



In this graph we look at what fraction of the borrowers are completing their loans compared to the fraction of them defaulting. We observe that almost 9/10 loans are completed when the Term period is 1 year. But if we recall from earlier analysis we realise that this only accounts for a small percentage of total loans. As the time period increases we observe that a majority of loans are still ongoing especially if it is 5 years.

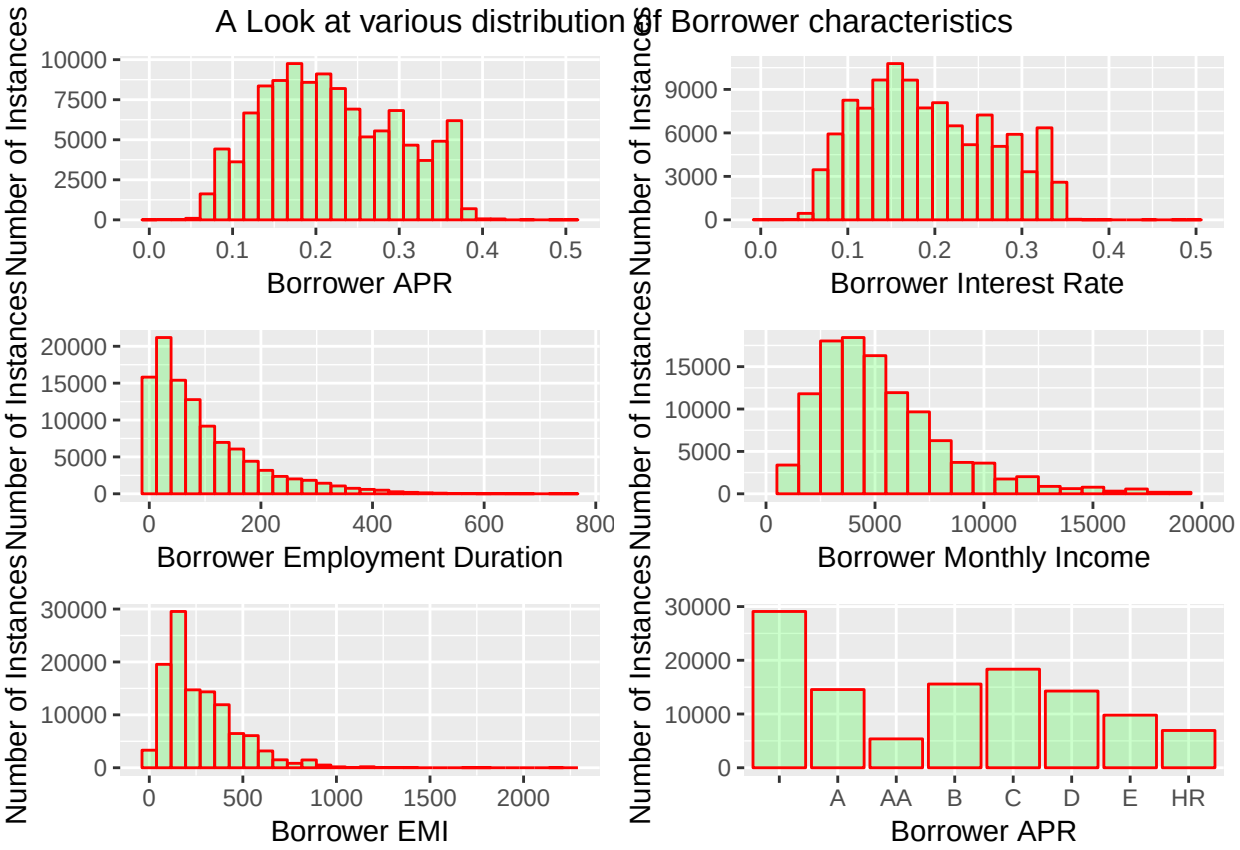


In this graph we explore the effect of Credit Score and Loans granted while also keeping tabs on the Prosper Rating assigned at the time of the Listing creation. The grey bars are Listings for which a Prosper Score was not available in the data. The Prosper ratings are synonomous to Credit Grade with 7 being the highest grade and 1 being the lowest. We observe that loans were mostly granted to people whose credit score was greater than about 640. We see that as the credit score increases the fraction of Prosper Rating being 7 also increases which seems appropriate.



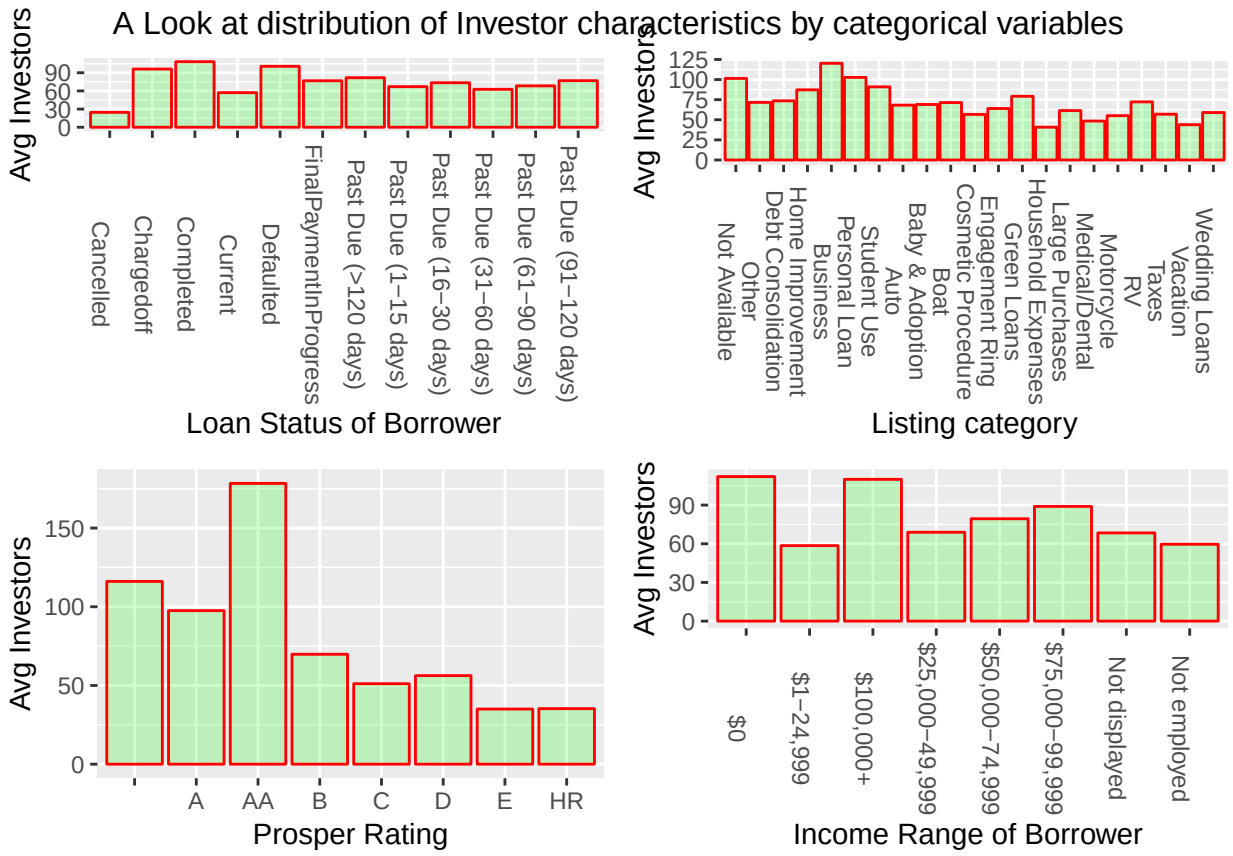
This graph was created to give a fresh perspective and also involve maps into the scene. Unsurprisingly, Prosper being a California based company has most of its loans in California. Other states which have a significant amount of Loans dispersed are Texas, New York, Florida and Illinois.

This graph is a pretty straighforward attempt to find a relationship between people whose income cannot be verified and their Prosper Ratings. We see that people having the top grade are rarely people whose income cannot be verified while almost one out of five people with below average credit score cannot verify their income source. This can be a datapoint for future analysis and risk mitigation.



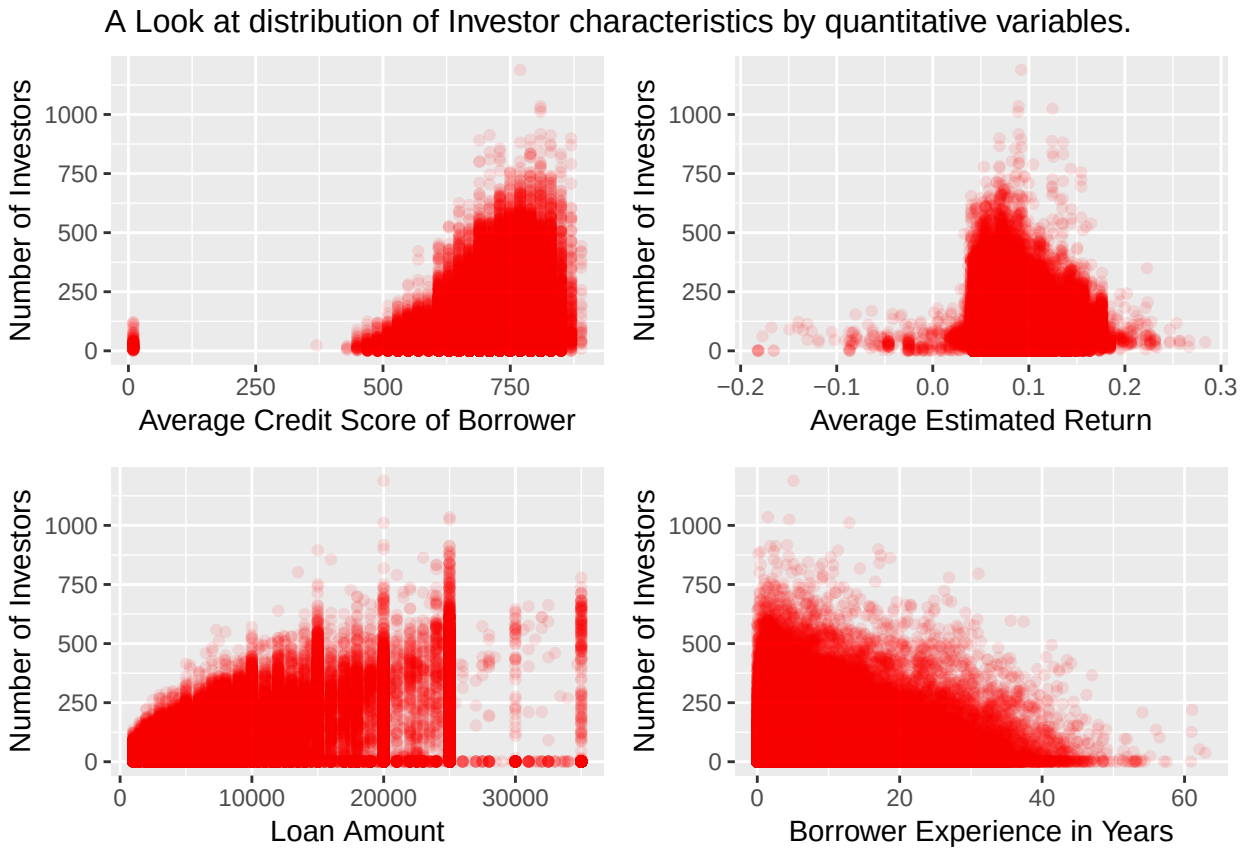
In this set of graphs we look at various borrower characteristics to get a better idea about the nature of our borrowers. Some interesting takeways from this set of graphs are:

1. APR for the loans can vary from as low as 10% to as high as 37% and almost similar values for the interest rate.
2. As the borrower experience duration increases the number of loans go down. Did people realise that Loans can be a burden or do people with years of experience require loans less than people fresh into their professional careers. Correlation does not mean causation.
3. The average yearly income of people in USA is $45k whereas in places where Prosper prospers like California, New York and Texas the average monthly income is about $55k. In our dataset though it seems that the people who take Loans earn less than the average income.
4. The equated monthly installment(EMI) of the borrowers follows the Loan amount taken and if compared to the graph analysed at the beginning we will see a similar shape.
5. A 'C' credit rating is deemed average by FICO and though Prosper seems to have a slightly different system we can see that Loans are taken by people whose ratings are around average.



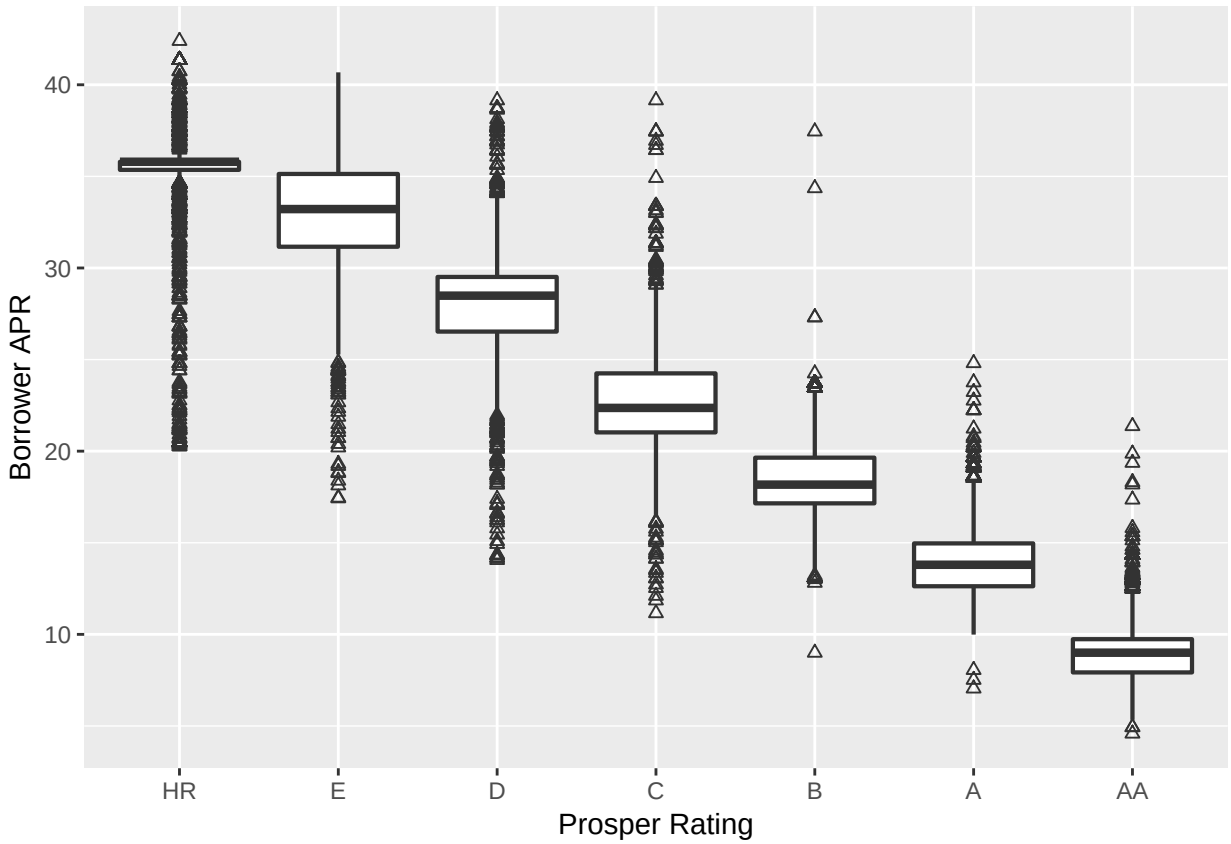
In the previous section we looked at various Borrower characteristics, now let us view the data from an Investor's perspective. Some takeaways are:

1. The average number of investors are almost same in completed Loans and Defauted Loans. Also there are significant bars in the Past Due status.
2. In the Listing categories graph focus on the shortest graphs and we find that investrs are probbably unilling to invest if the reason behind the loan is a vacation or any large purchases.
3. Unsurprisingly, as the Prosper rating reduces the average number of investors also decreases.
4. We can see that as the Income range increases the average number of Investors also increases.

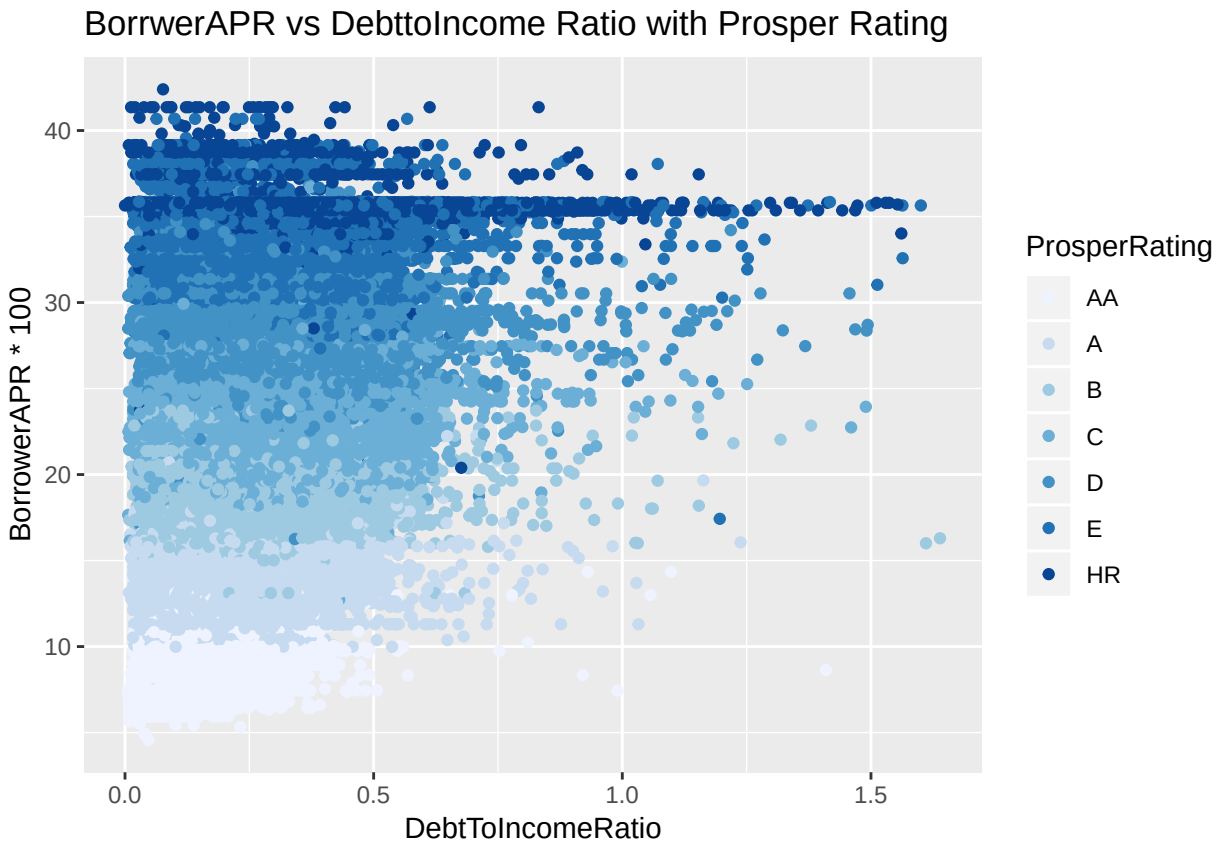


This set of graphs give us a view of some of the quantitative variables:

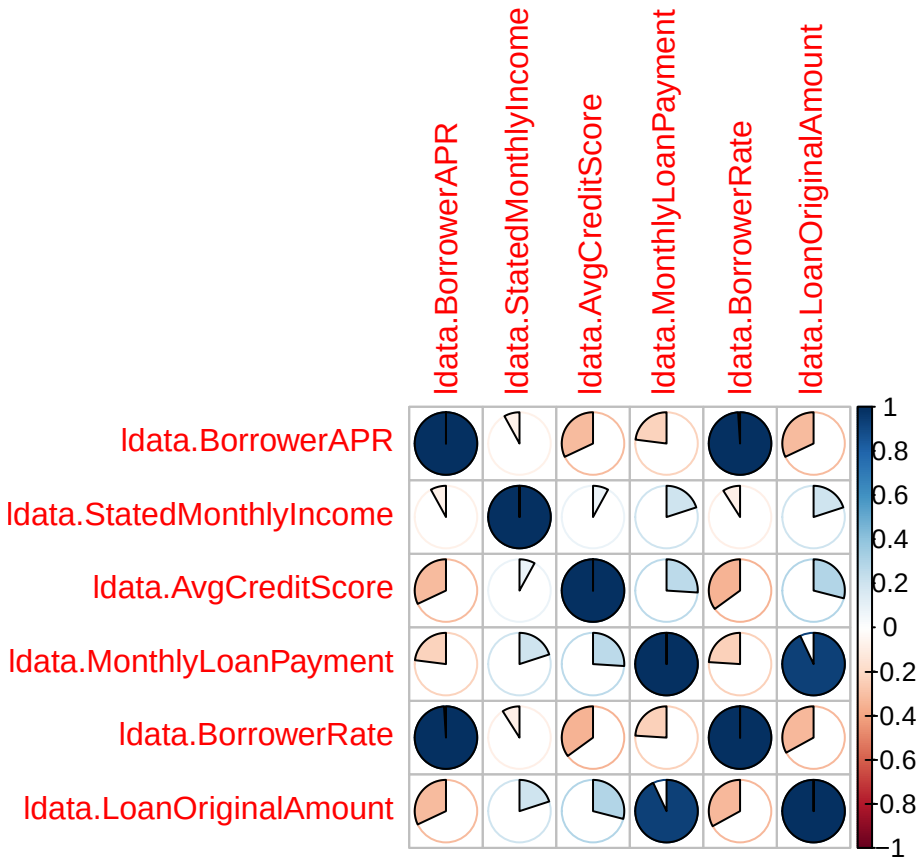
1. We see the a better average credit score will ensure more investors and that is expected. It is also interesting to note that when credit scores are unavailable loans are still disbursed albeit the number of investors are significantly low.
2. Tha major chunk of investors hover around the estimated return percentage of about 5-10% and the fact that the number of investors does not increase with increase of estimated return might be an indication of investors sticking to a safe value rather than charging more.
3. The shape of this graph is expected as when the value of the Loan increases it is expected that the number of investors will go up. There are well defined peaks at every 5000$ and that is an expected outcome too.
4. Borrower Experience in Years has surprised me the most. Some people have(or claim to have) than 50 years of experience. Earlier we have seen that the number of loans decreases with increase in experience and that is reflected here as well.

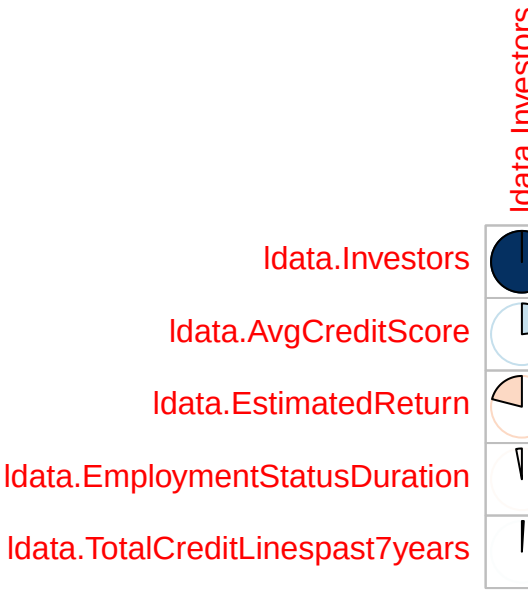


This graph gives us an insight to the relationship between Borrower APR and the prosper Rating.



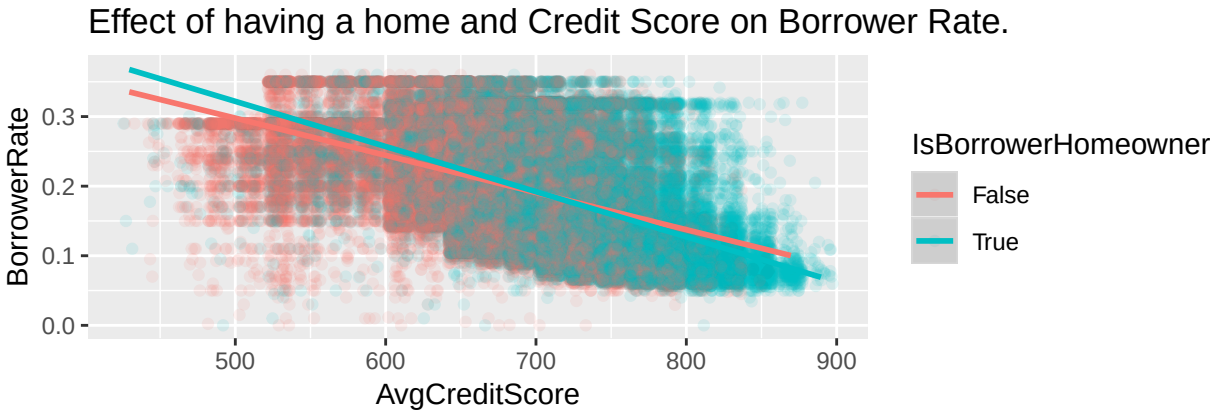
This graph helps us analyze the effect of Debt-to-Income ratio on Borrower APR for various Prosper Ratings

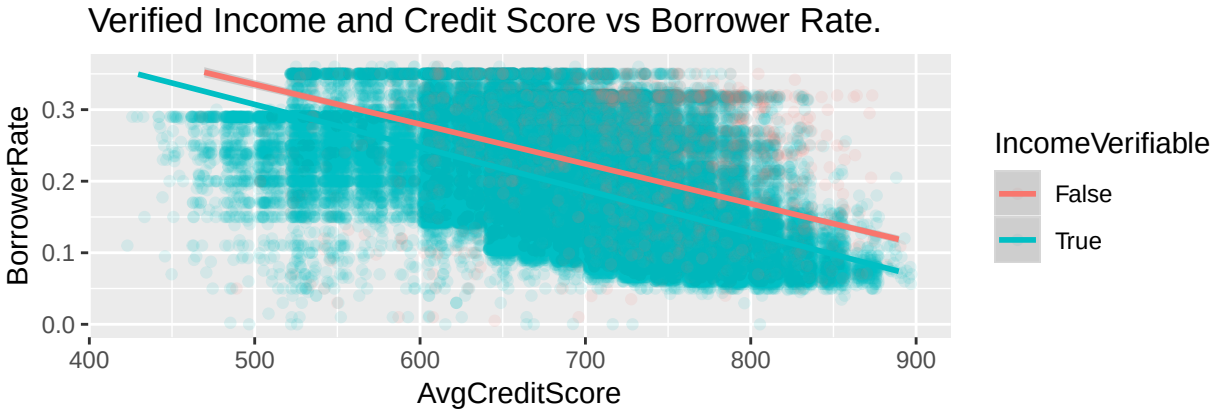




In the above two graphs we observe correlation between various Borrower and Investor characteristics.

In the Investors graph we find weak relationships between Average Credit Score, Estimated Return and Total Credit Lines in the past 7 years with the nymber of Investors. In the Borrowers graph we find a weak negative correlation of Borrower APR with the Average Credit Score. Also we can see a weak positive correlation with Average Credit Score and these values are not anything out of the ordinary.



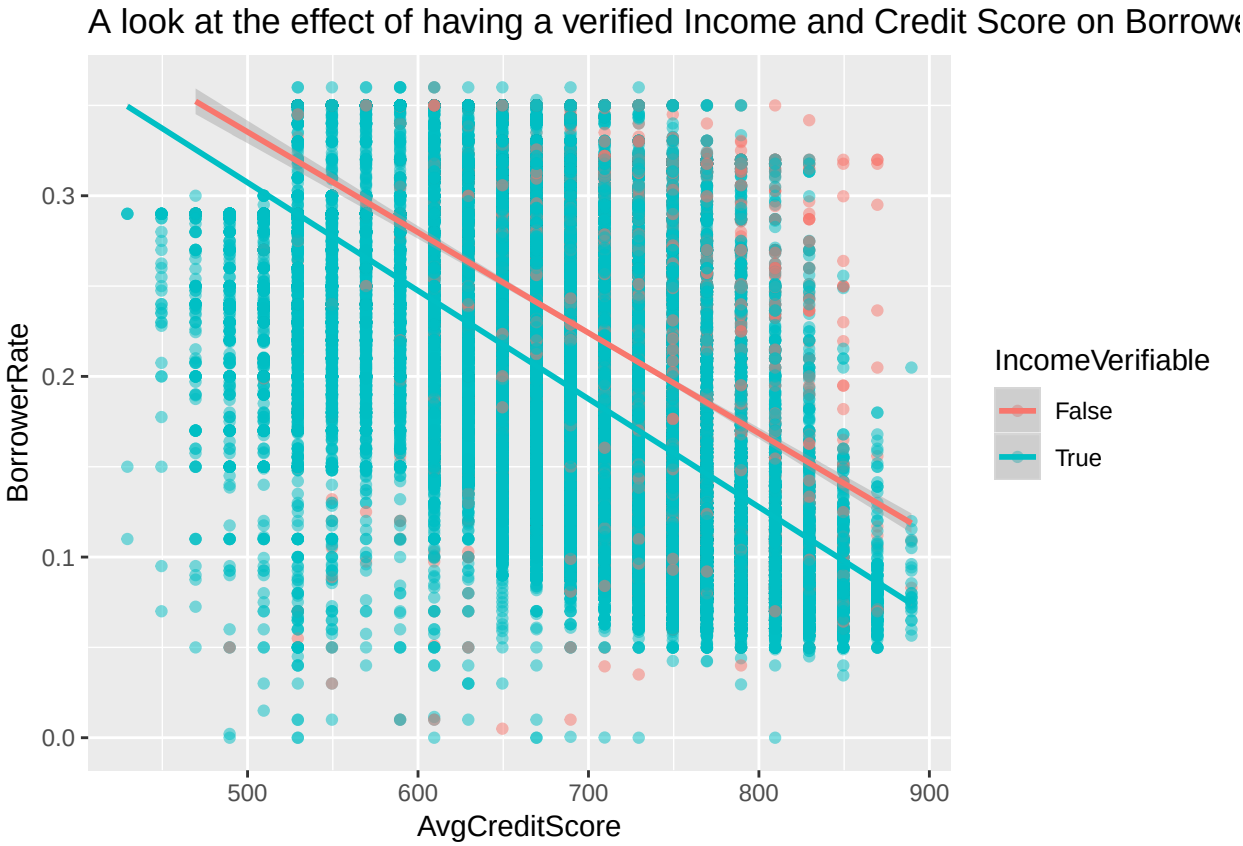


The above two graphs are meant to provide insight into the comparisons of three variables together.

In the first graph we see that if the Borrower owns an home and has a credit score till 720 his rate seems higher. After 720 this theme seems to change. This might be due to the number of loans decrease after a certain cedit score as we have seen earlier. In the second graph we can see that if the income source of the borrower cannot be verified then he always has a higher rate than individuals whose income can be verified. Although both show a decreasing trend as the credit score increases. Interestingly a lot of people seem to not have a home and yet get a loan granted whereas that is not the case with income verification.

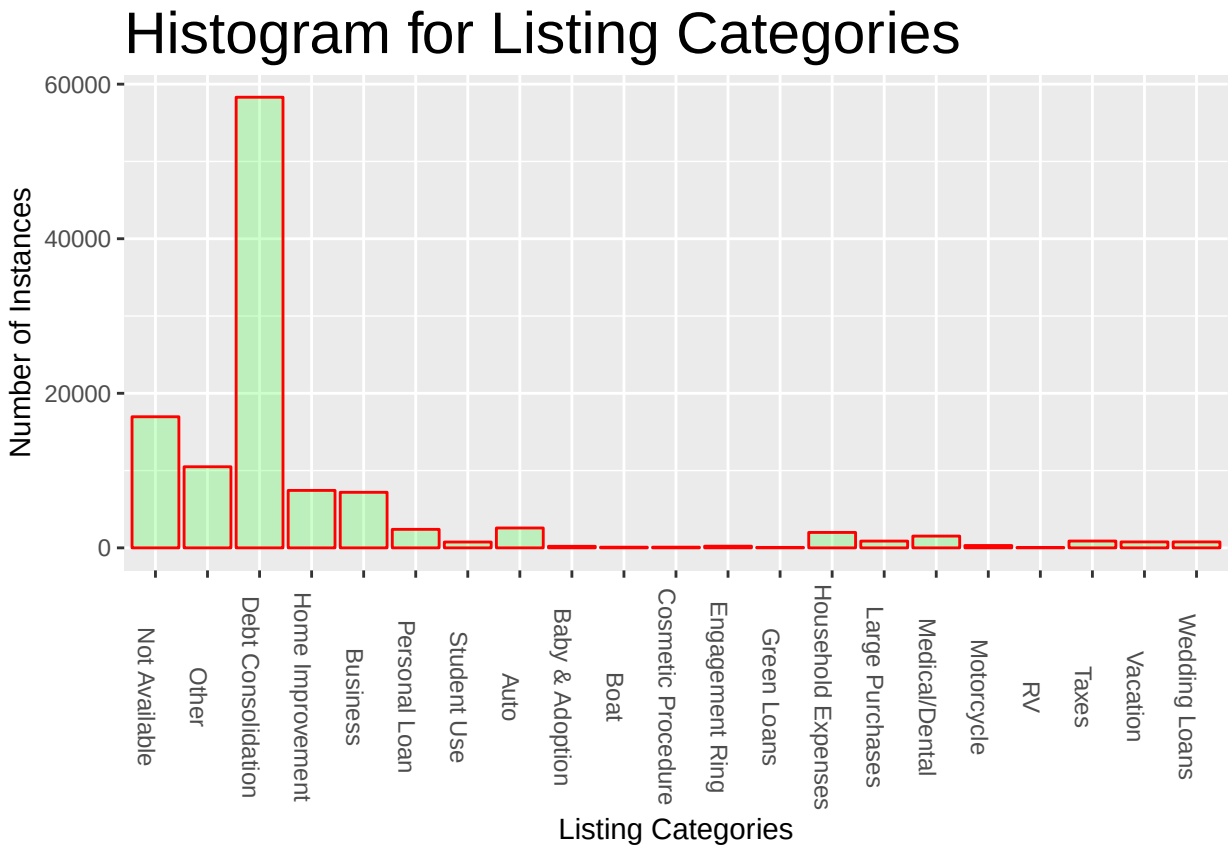
Final Plots and Summary

Plot 1



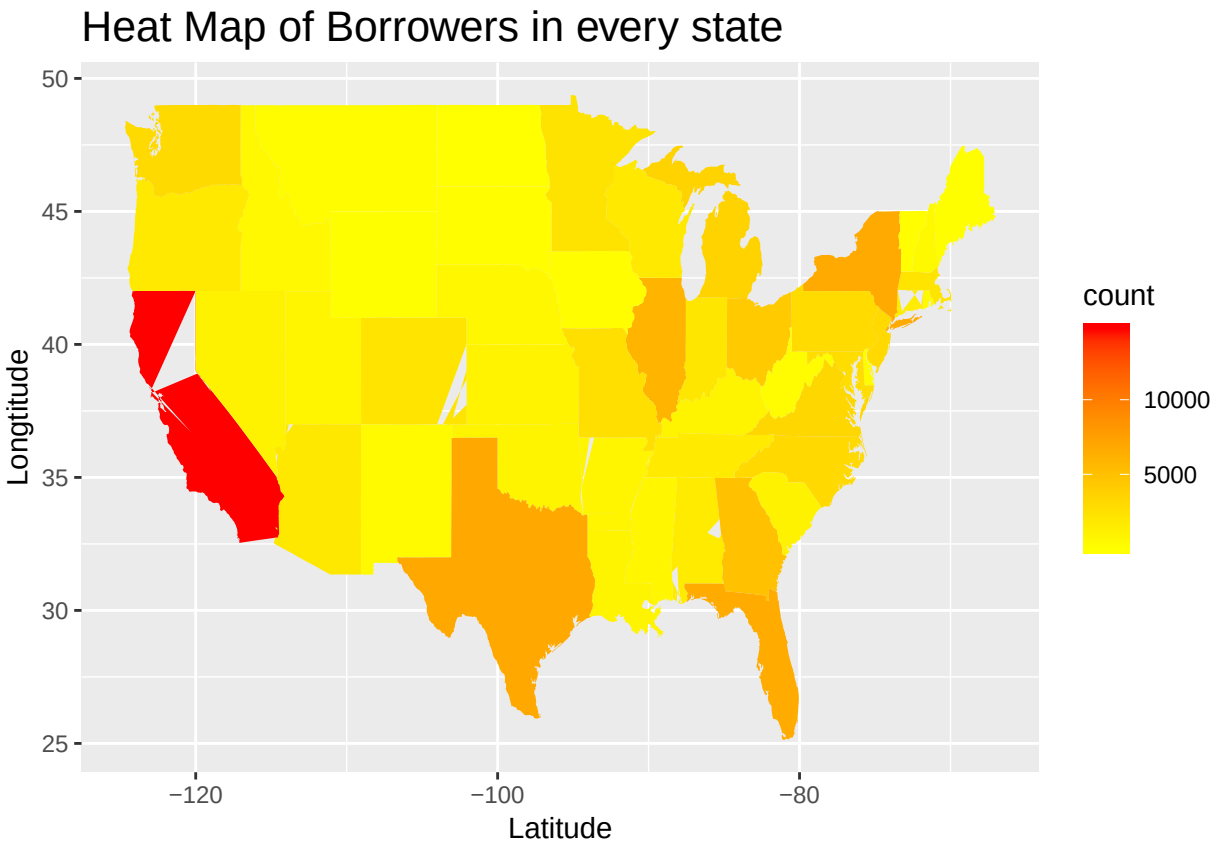
In my opinion this plot gave us a lot of insight of how general trends work. If one can have his income verified and even if the have a low credit score they can find investors to get loans. It is tough to find investors when income isn't verified and also the Interest rate remains significantly higher. This plot also shows us the advantage of having a higher credit score which generally would lead to lower rates of Interest.

Plot 2



The fact that people take loans to repay other loans might seem dumbfounding to non-finance people but it is true. Home improvement, Business and Auto seems to be okay with investors as they must include something tangible and specific to invest their money in.

Plot 3



I am fascinated by map plots and this map gives us many details apart from the obvious. The states of New York, Texas and Illinois have a higher than national average income whereas as Florida is just about the same. California being the headquarters of Prosper is obviously the dense zone. It might be interesting to see some investor and Borrower characteristics of these speific states.

Reflection

This dataset is immense and provides the scope of insightful analysis. I tried to cover as many variables as possible but there remains a lot of opportunities to explore this dataset further.

Understanding the meaning and effect of certain variables was the chief challenge in this dataset. To establish the apt combination of variables which yield meaningful analysis forms the crux of the problem here. Also some variables like Rate, APR etc. made for some confusion in deciding which variable would yield the best analysis.

The most interesting parts that I have discovered are the ones where I worked with multiple variables and their effect on each other. There is a lot of scope here and I would like to explore this domain more. Cleaning of data and unexpected results are part of any real-life dataset and this was a major part of my learning.

A variety of visualizations have made reduced the dataset to various components which can be further analyzed. Apart from this I have succesfully found that most of my assumptions where close to the real world values.

As part of future developments I would like to try out more combinations to figure out more advanced prediction models which might help in predicting a potential loan defaulter. Another direction might be to figure out more combinations and to find a better correlation between them which might help in production too.