Analysing Loan Data from Prosper

Flow of the Presentation

- Introduction
- Assessing Data
- Cleaning Data
- Univariate Exploration
- Bivariate Exploration
- Multivariate Exploration

Introduction

This dataset is financial dataset, and this is related to the loan, borrowers, lenders, interest rates and stuffs like that. Prosper or Prosper Marketplace Inc. is a San Francisco, California based company specializing in loans at low interest rates to the borrowers. In this dataset, we are using the data from the Prosper to analyse it and trying to find the pattern in the Prosper data. This may be tedious because of the sheer size of the dataset and the complicated nature of all the financial datasets. We are using Python libraries to plot some visualizations.

Our investigation will focus on analysing the factors that affect borrower's APR and What type of loan has been taken by what kind of borrower.

Assessing and Cleaning Data

1) What is the structure of Data?

```
- The data consist of a total 81 attributes and accounts for
113,937 entries. Each entry gives us idea about the borrower and
its background, and the details of the loan associated with
them.
```

2) What are the main features of your analysis?

```
- The above question is answered by keeping the attributes of
concern and removing the others. Done below.
```

3) Do all the entries have a prosper score associated with it?

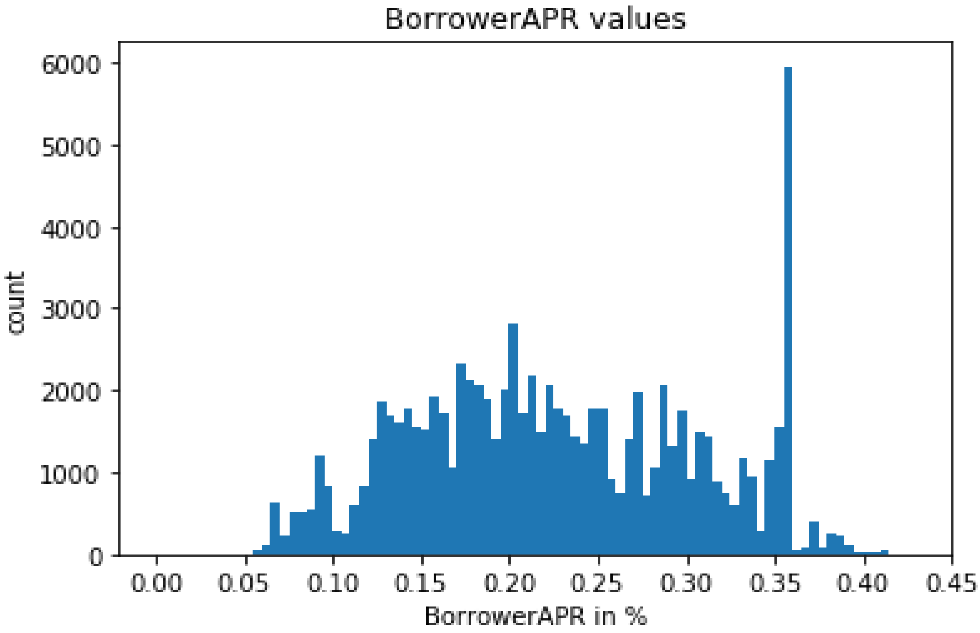
4) What feature(s) do I think will help most in the analysis of my interest?

```
- I assume that 'Borrower's APR', 'Prosper Rating', and
'Occupation' will play a major role in determining the result of
analysis. The answer to this question will be answered in detail
by visualization.
```

Univariate Exploration

- Which values of APR are most occurring?

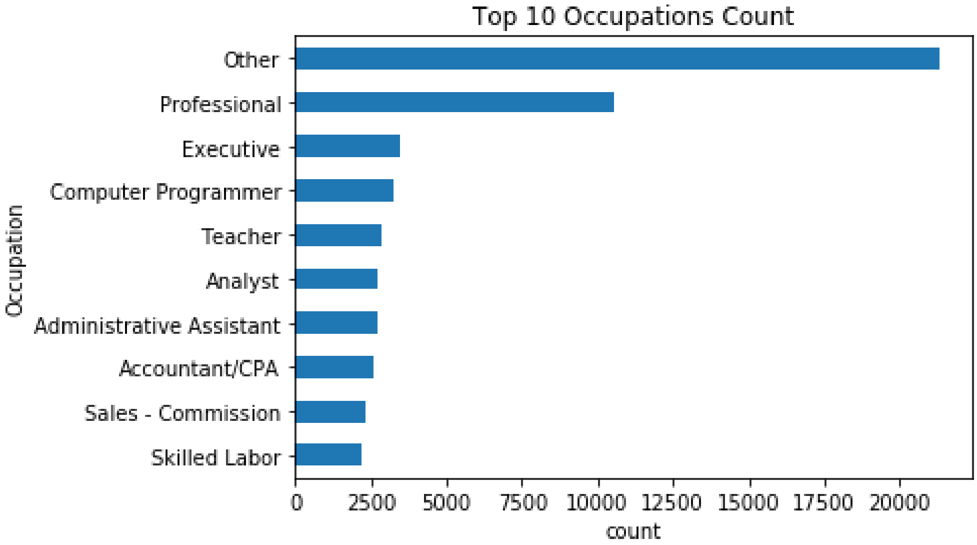
Let us plot different values of BorrowerAPR



BorrowerAPR distribution seems fairly normal with mean around ~0.19, with the exception of value ~0.35 which is surprisingly high and this needs to be probed.

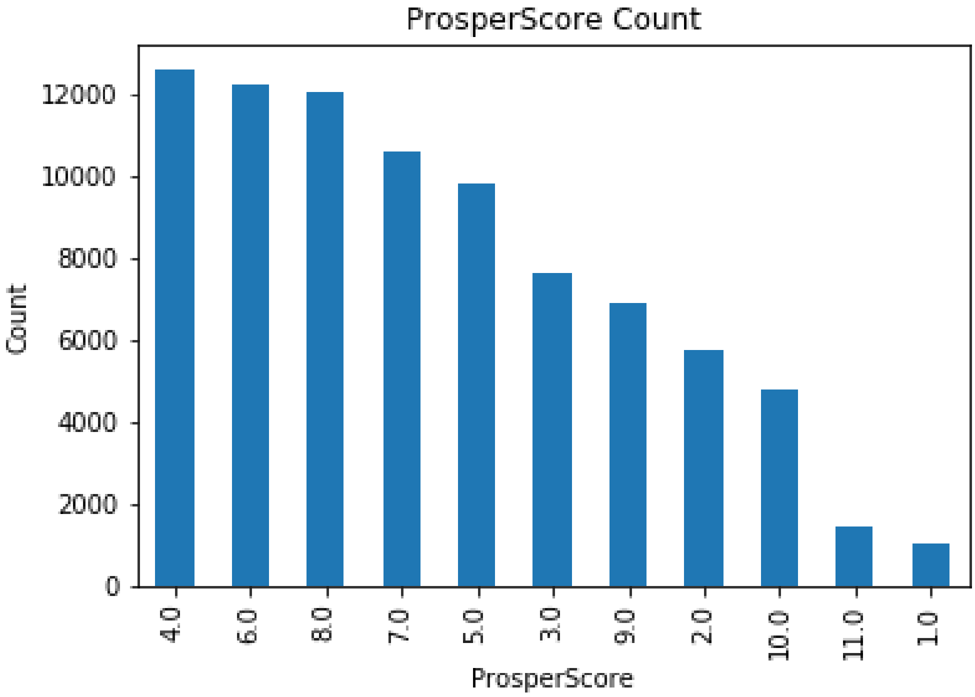
- What is the most pursued occupation in our dataset?

Let us see the types of occupation that our borrower is involved in.



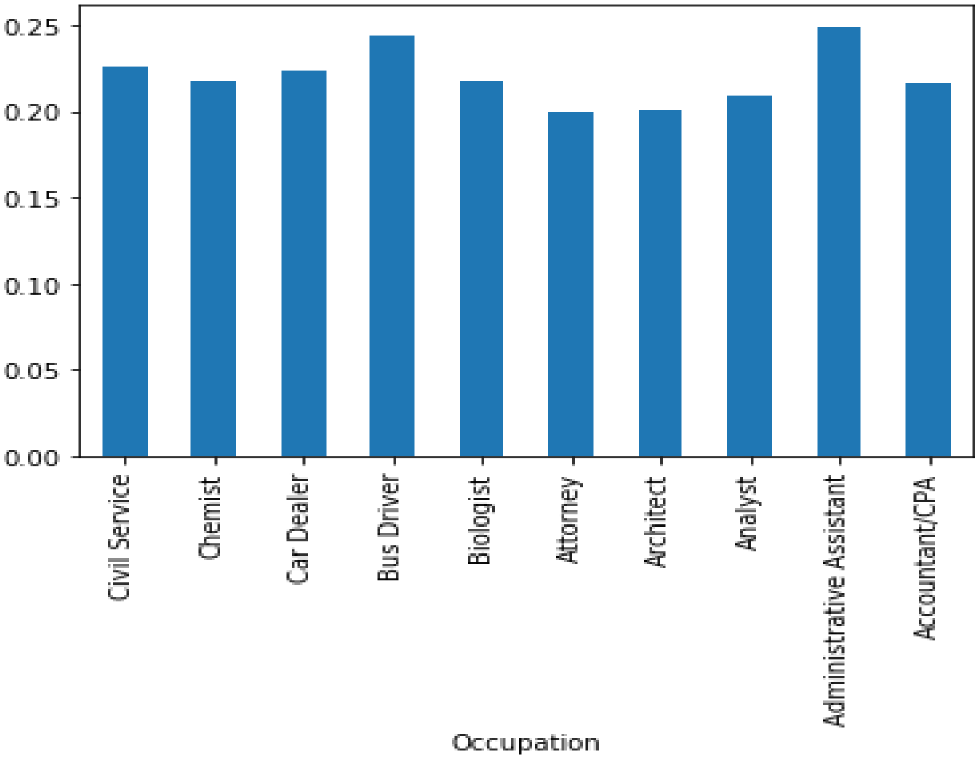
Occupations: We see a huge number of entries as 'Other' and 'Professional' we can deduce that these people are reluctant in disclosing their actual professions. The rest of occupation categories are evenly distributed.

- What Prosper scored occurs the most?



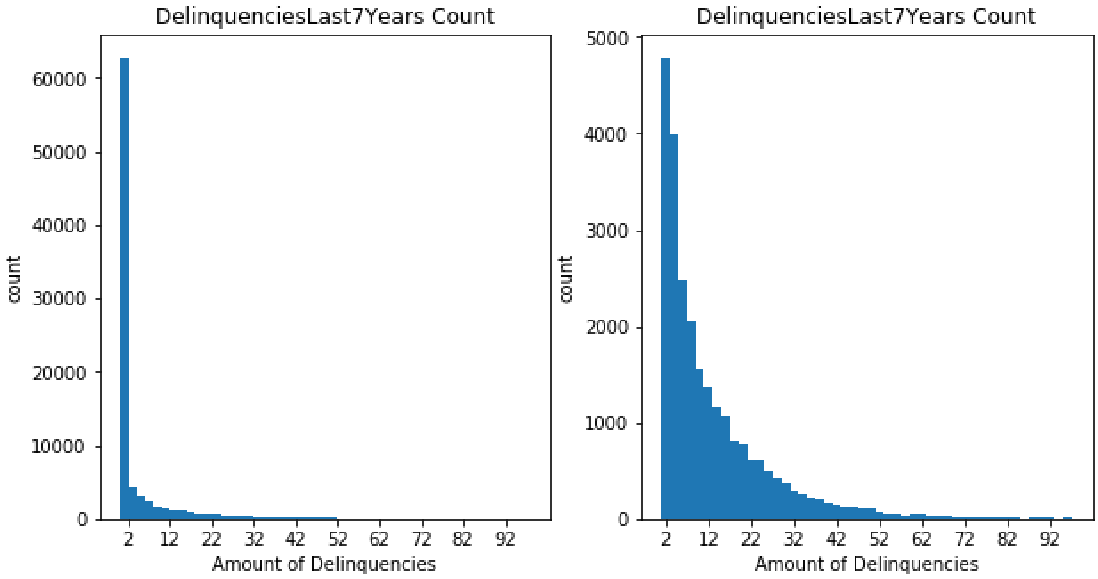
Prosper Score: Upon inspection we can say that more the money you borrow, the lesser Prosper Score you get.

- Let us bar plot for APR means of top 10 occupations.



We can observe that there is negligible difference between mean APR scores of top 10 occupations. Hence this exploration is not substantial. We can consider that Occupation might not be the most appropriate attribute in our analysis.

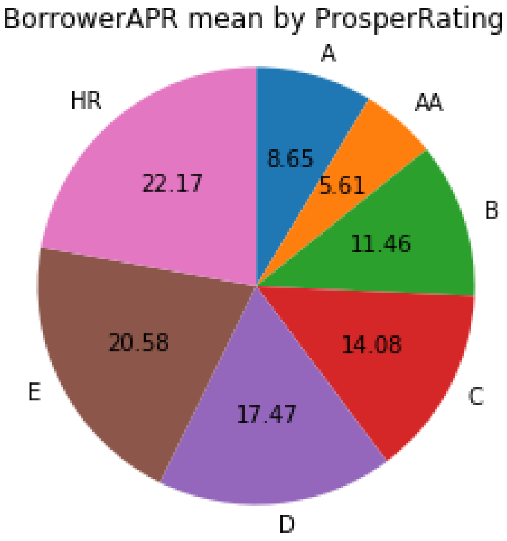
- Let us put some light on delinquencies from past.



DelinquenciesLast7Years: The first plot shows overall delinquencies and we can infer that most of the borrowers do not have any delinquency. But, as we remove those records where there is no delinquency, the spread becomes more distributed and the decrease is gradual.

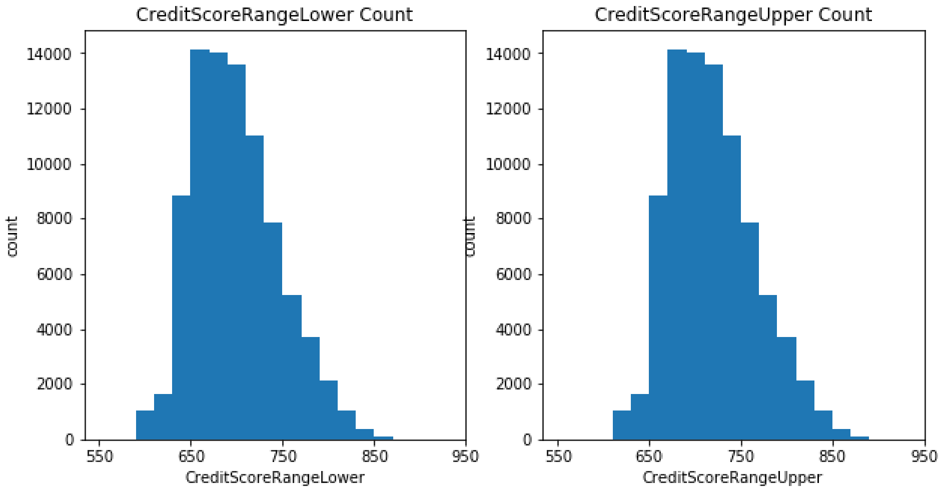
Bivariate Exploration

- Let us check for BorrowerAPR mean by ProsperRating mean.



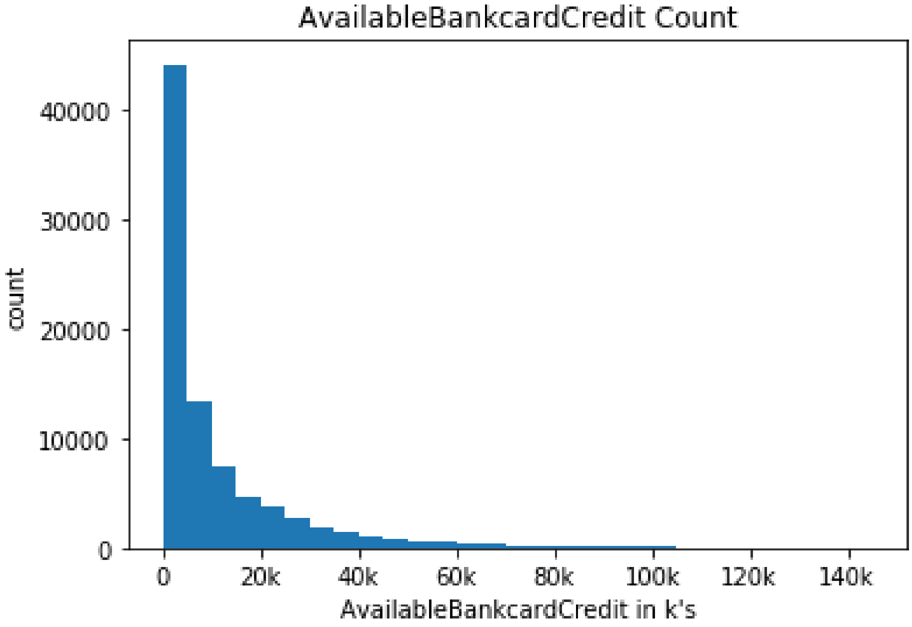
BorrowerAPR mean by ProsperRating : The ProsperRating categories have been labelled from highest to lowest manner (i.e. AA, A, B, C, D, E, HR). We can observe that the highest rating of AA has received lowest mean APR and vice-versa, i.e. lowest ratings have received highest mean APR. This is somewhat intuitive because better ratings induce more trust in customer and hence in case of any default also, the punishment in terms of APR is low.

- Comparing CreditScoreRangeLower & CreditScoreRangeUpper



CreditScoreRangeLower & CreditScoreRangeLower: The plots are quite similar and nothing substantial to differentiate.

-Let us see how the AvailableBankcardCredit is distributed.



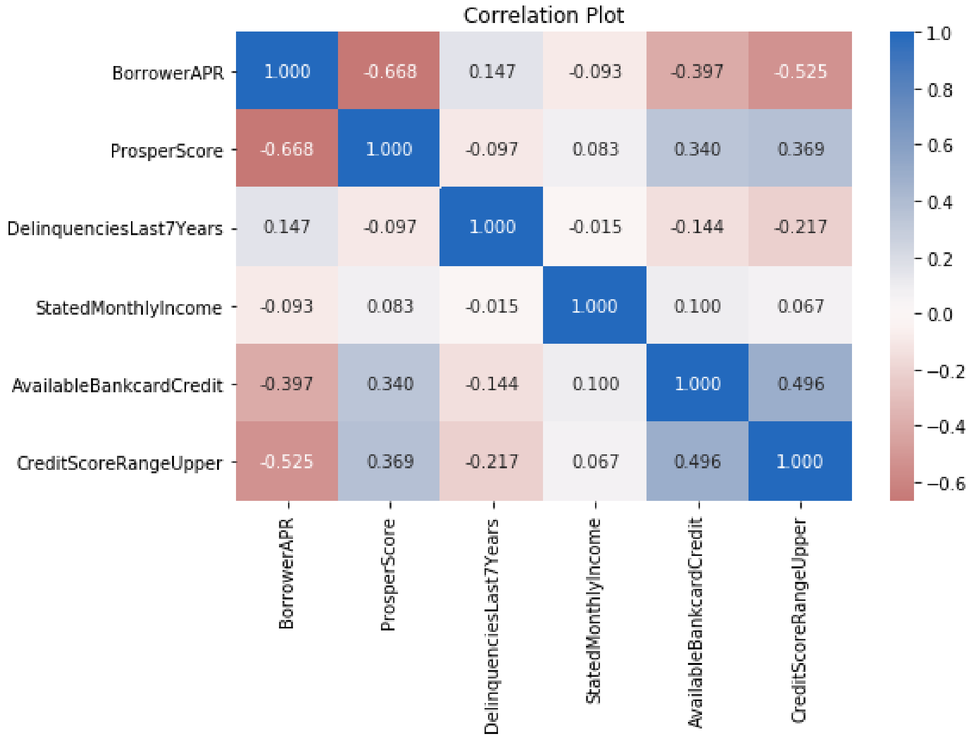
We can clearly see that the above plot now shows how the Available Bankcard Credit is distributed in more detail.

While performing this basic analysis, did you find any feature or attribute to be behaving unusually?

Yes, features like Occupation turned out to give no significant idea about how it is affecting Borrower's APR. In real, occupation can be generalized to give a sense of background and income which indirectly affects APR but from the above preliminary analysis, we cannot say so.

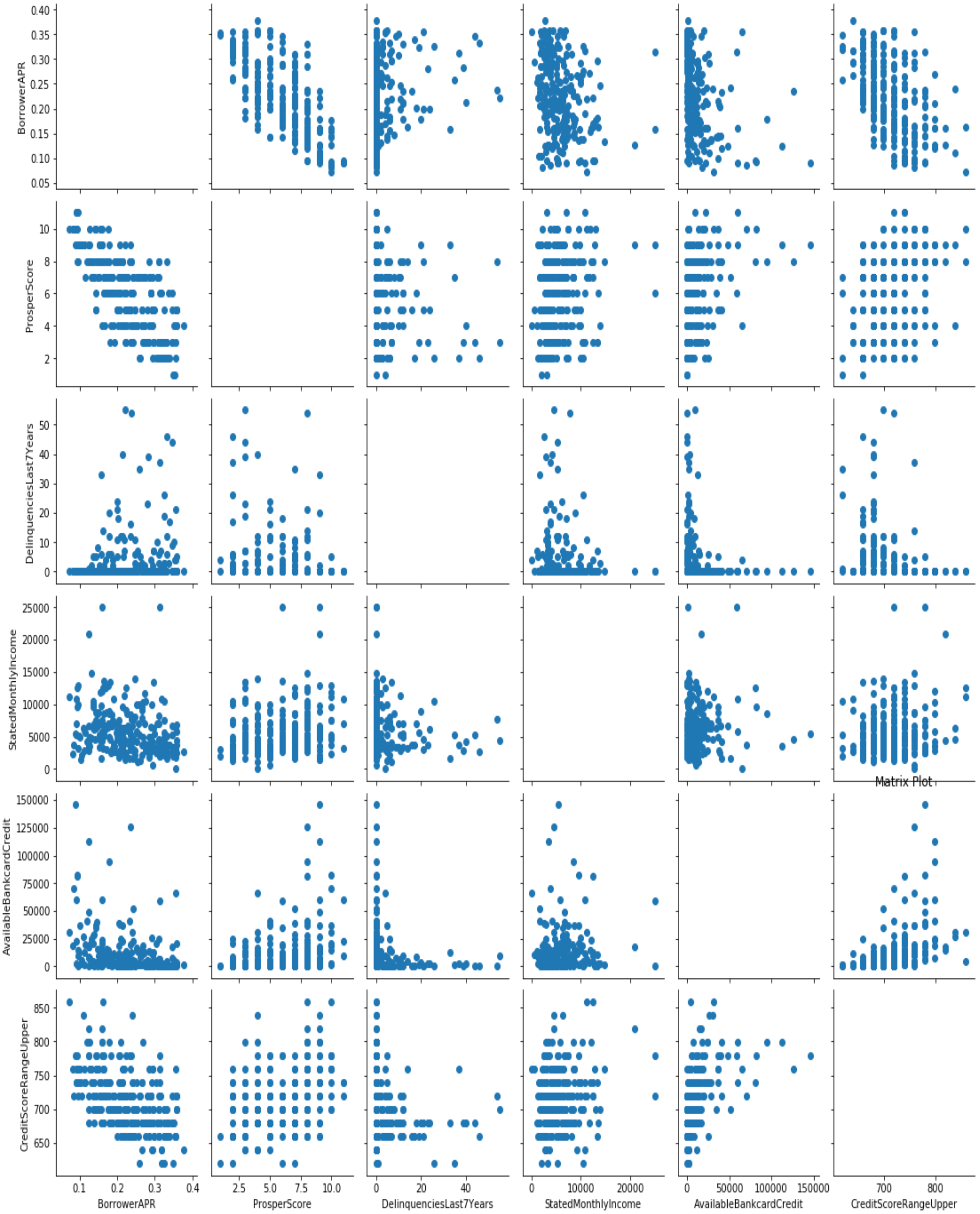
Multivariate Exploration

- Let us see how the attributes are influencing each other and also how are they correlated using the Correlation plot.



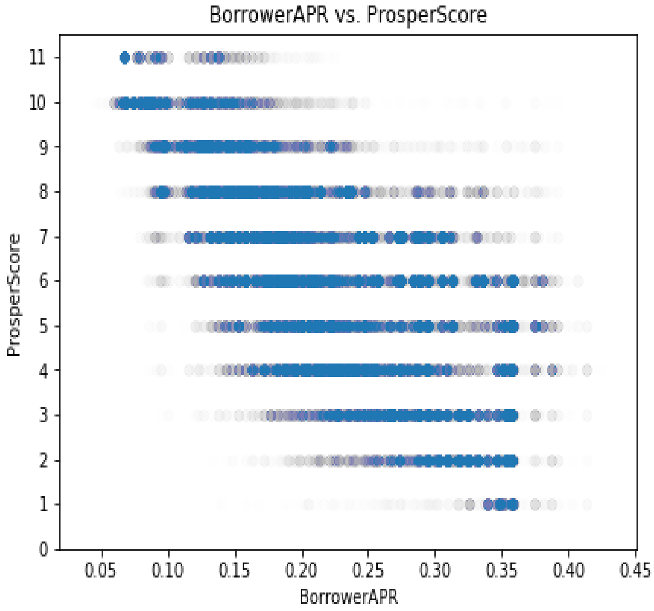
The correlations are mostly negative and that can be understood and is justified. The highest negative correlation is between 'ProsperScore' and 'BorrowerAPR', which should be true because higher the prosperscore, higher will be the trust in customer and his ability to repay the dues hence they are inversely proportional or negatively correlated. Same relation can be seen in Credit Score and BorrowerAPR also.

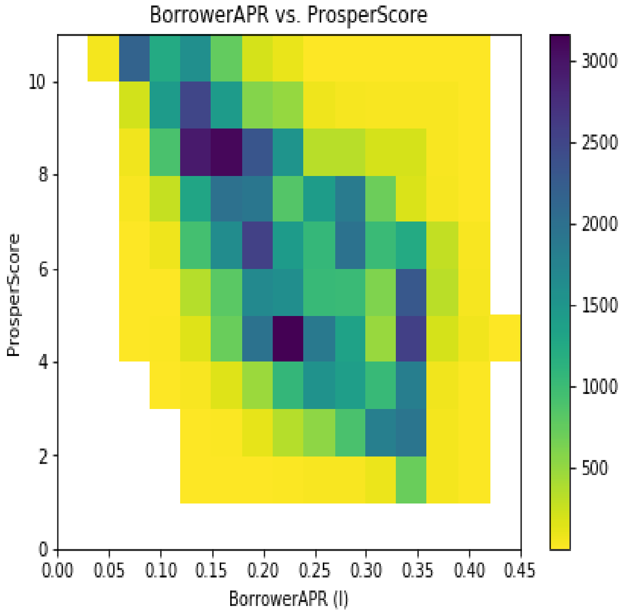
- Let us have a look at scatter plots for these variables



The Matrix plot also turns out to be like the correlation plot in determining which attributes have positive and which attributes have negative correlation. We can deduce that ProsperScore has most negative correlation with BorrowerAPR, we can try and analyze it better by explicitely plotting them in different and more informative manner.

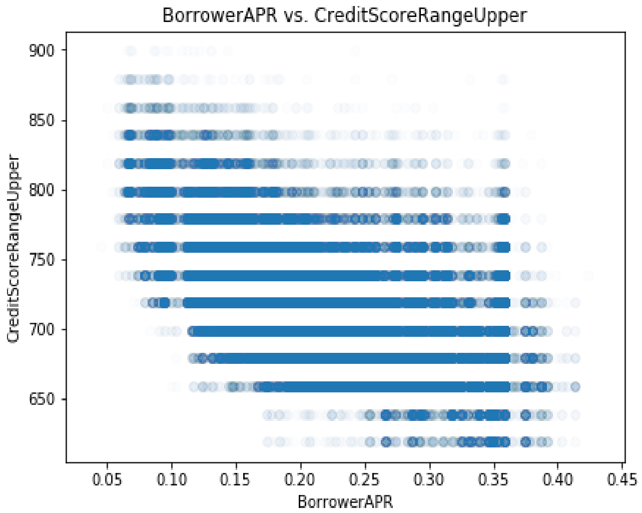
- Let us plot scatter and heat plot for comparing ProsperScore and BorrowerAPR.

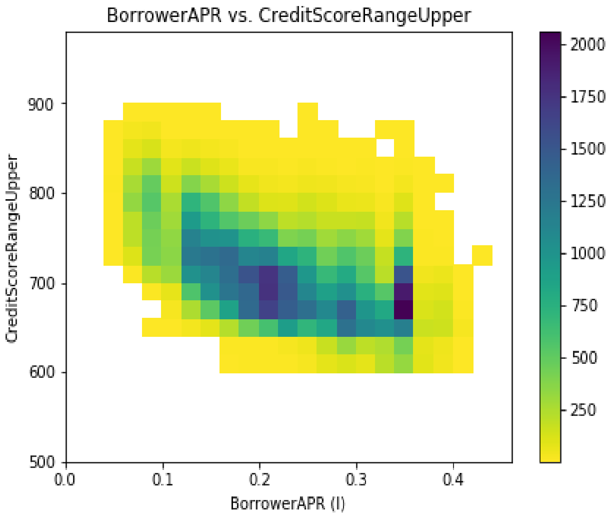




Our assumption is likely to be true about both variables being Negatively correlated.

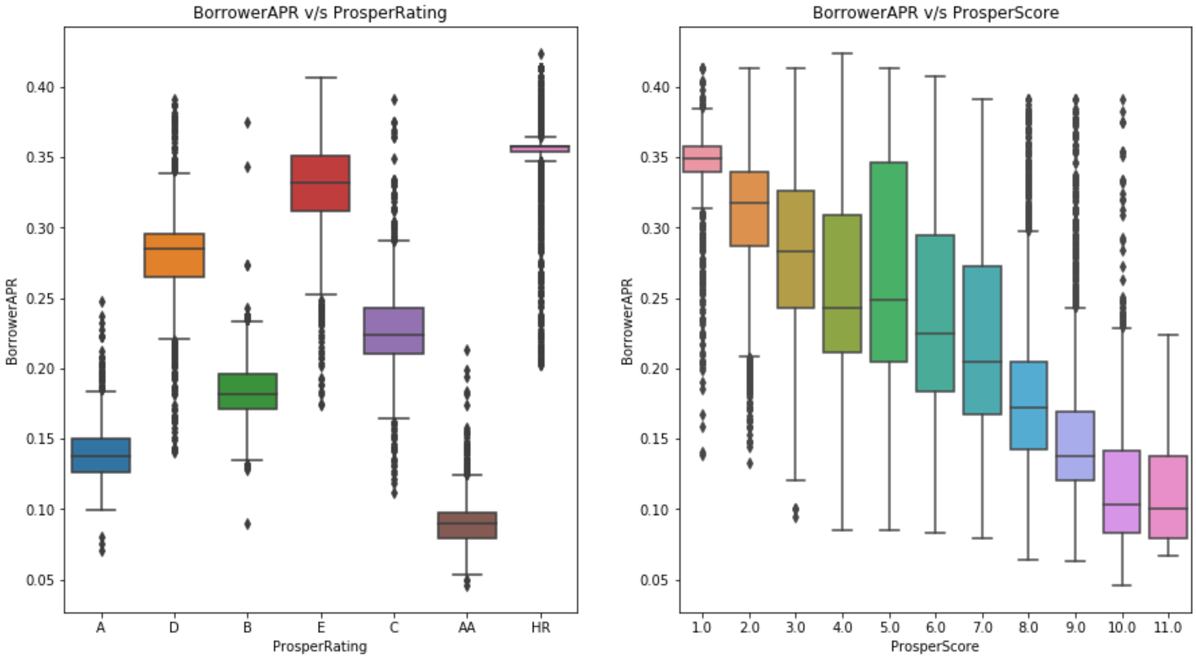
- Let us plot a similar scatter plot and heat map for other 2 variable i.e. BorrowerAPR and CreditScoreRangeUpper





The above plot makes sense because the higher the Credit Score, the lower will be the BorrowerAPR. The Credit Score also is positively correlated with the ProsperScore

- Let us try to plot relation between BorrowerAPR and ProsperScore & ProsperRating. The primary focus while doing this viz is to uncover some information that might not be visible using the above plotted plots.



The Violin plot for BorrowerAPR and ProsperRating & ProsperScore gives us the idea that ProsperRating cannot substantially help in seeing how it affects the BorrowerAPR. ProsperScore plot clearly shows a negative correlation with the BorrowerAPR.

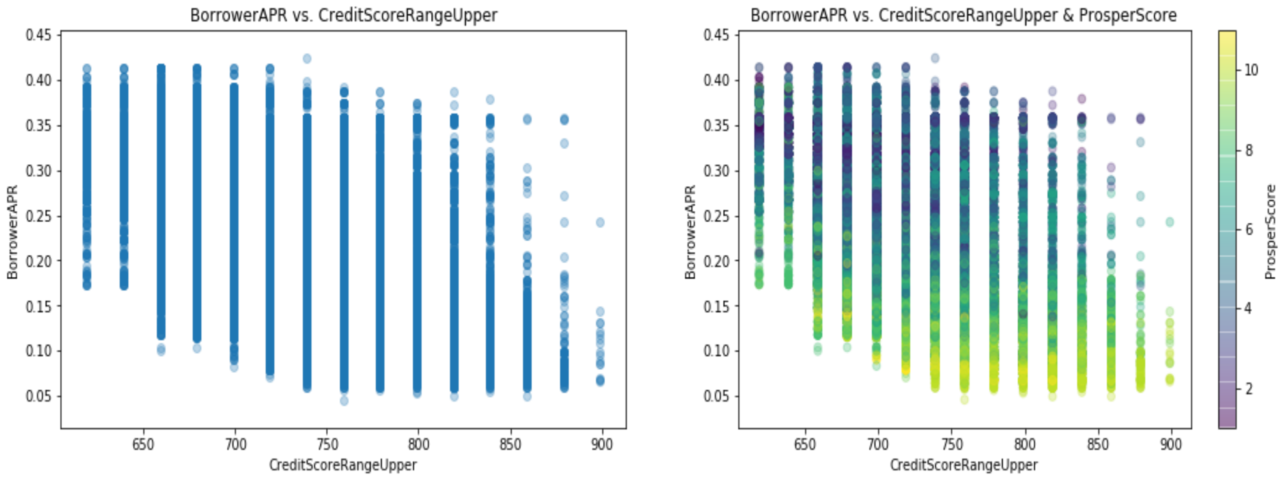
What are some of the relationships of our interest and what is their nature?

- The plots above have revealed 2 features / attributes having a relationship of our interest and should be probed for further insights.

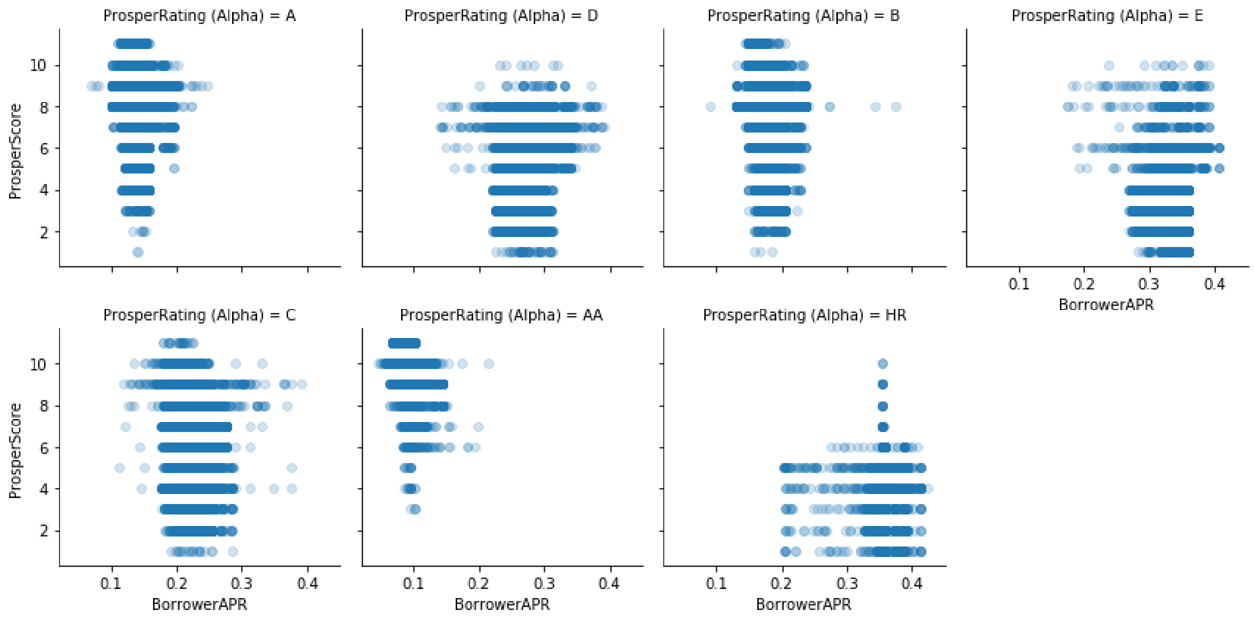
- We can say that ProsperScore and CreditScoreRangeUpper, both, have a negative correlation to the BorrowerAPR. A probable reason for such behavior are mentioned below each of the plots.

We will utilize these variables to try 2 kinds of plot and try and implement Multivariate Visualization

- Here we start by plotting two variables at first and then try to include third variable in the same plot to get a multivariate visualization
- The negative correlations (BorrowerAPR and ProsperScore & BorrowerAPR and CreditScoreRangeUpper) is visible from second plot itself.
- As the Credit Score Range goes higher, the BorrowerAPR can be seen decreasing.
- Similarly, with lower ProsperScore (i.e. bluish dots) the BorrowerAPR stays higher or on top of the plot unlike the yellow dots including lower BorrowerAPR and at the same time also some positive correlation with Credit Score.



Let's try and use FacetGrid() to do a similar multivariate visual



This plot shows us the correlation between BorrowerAPR and ProsperScore for each of the ProsperRating individually. We can deduce that the ProsperRating categories would not have any pattern of correlation with BorrowerAPR. But we can deduce that the people with lower ratings usually tend to have a higher APR and which can be true in real world too.