Prosper Loan (exploration)

- Data-frame Explanation

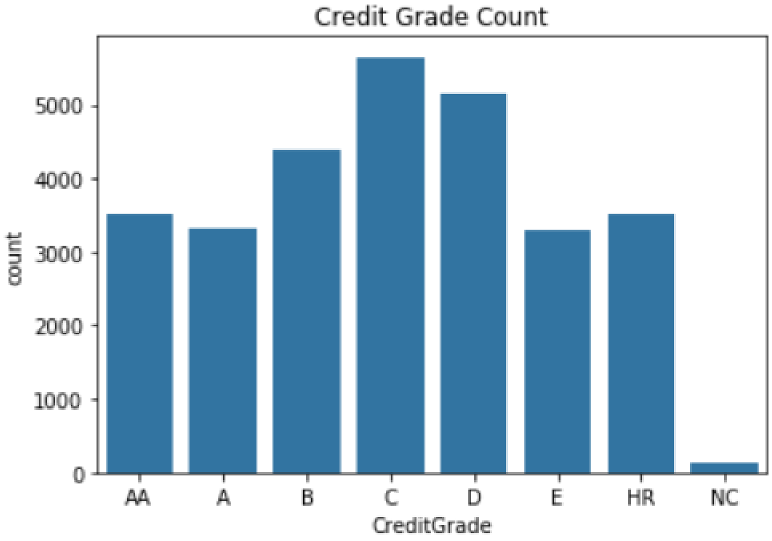
Prosper loan data is gathered from 2007 to 2014. The original dataset had 81 columns and 11397 rows. During the data cleaning process, four data-frame were created:

df_clean(‘prosper_data_ready.csv’) data-frame contains cleaned the whole dataset, loan_unclosed(‘unclosesd_loan.csv’) data-frame contains unclosed loan status data which is no fully paid on the loan, loan_closed(‘closed_loan.csv’) data-frame contains closed loan status data which is paid off for the loan, and prosper_date(‘date_data.csv’) contains data with timestamp index for time series analysis.

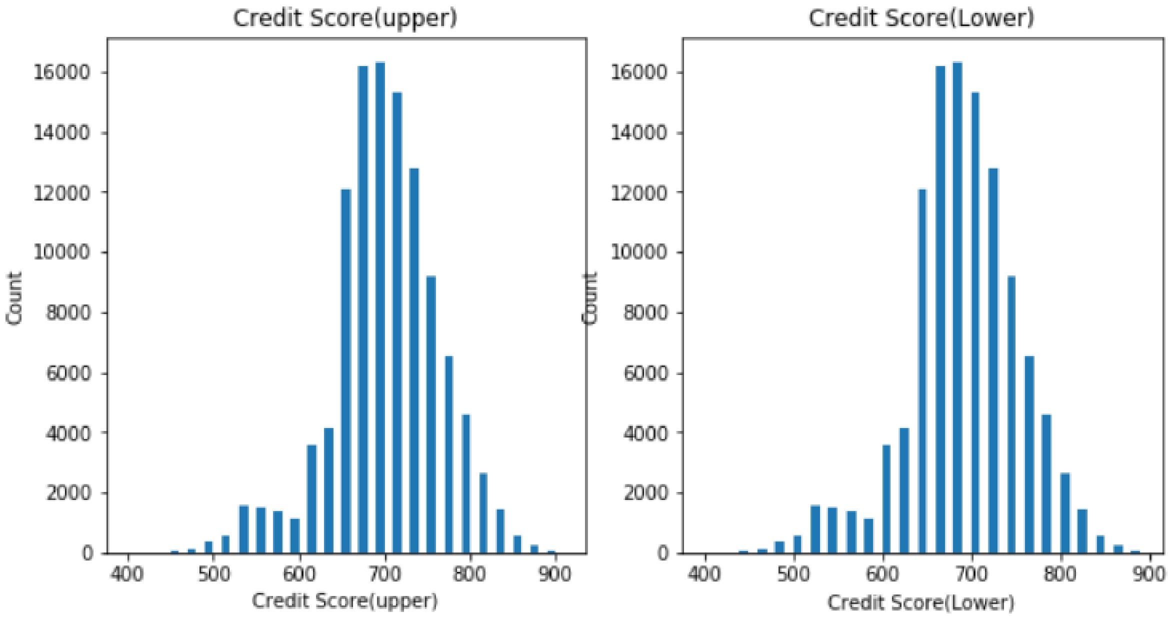
Prosper Loan dataset was explored in univariate, bivariate, and multivariate exploration. To explore the dataset, a lot of different graph method was used, such as count plot, heat map, scatter plot, etc.

- Univariate Exploration

Univariate exploration analyzes one variable at a time, count plot, and histogram was used for the exploration.

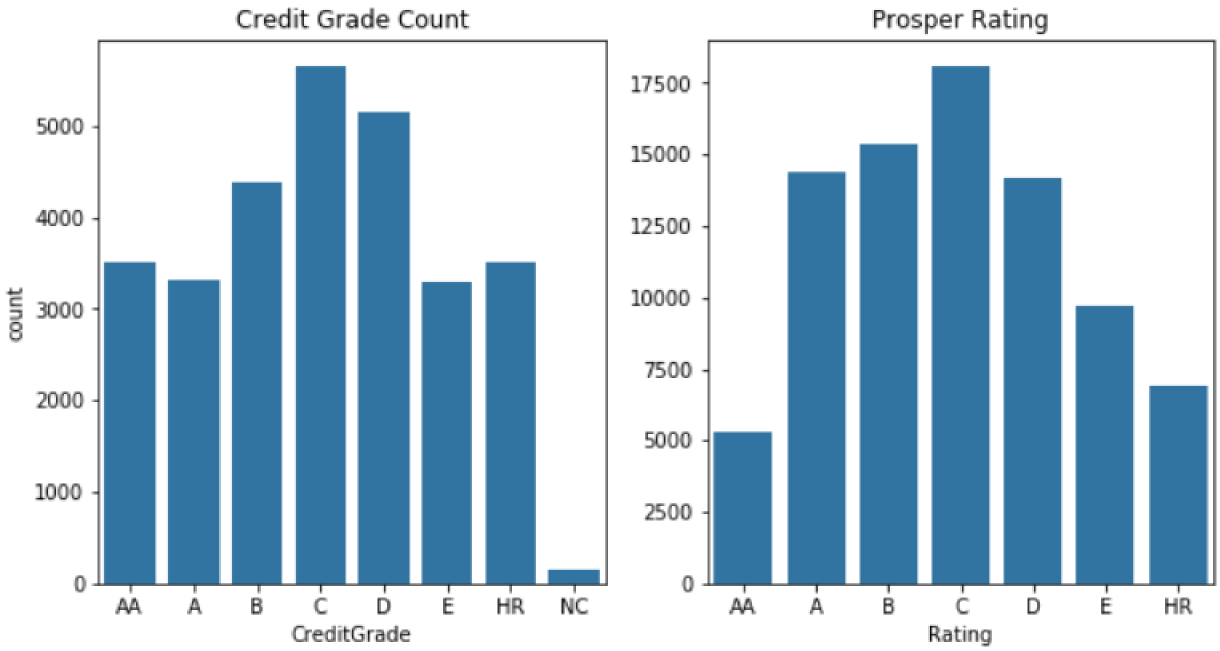


Credit grade data were plotted on count plot, credit grade C was the highest proportion of the dataset, D was the second, and B was the third. NC(No Credit) grade had the lowest count of all.

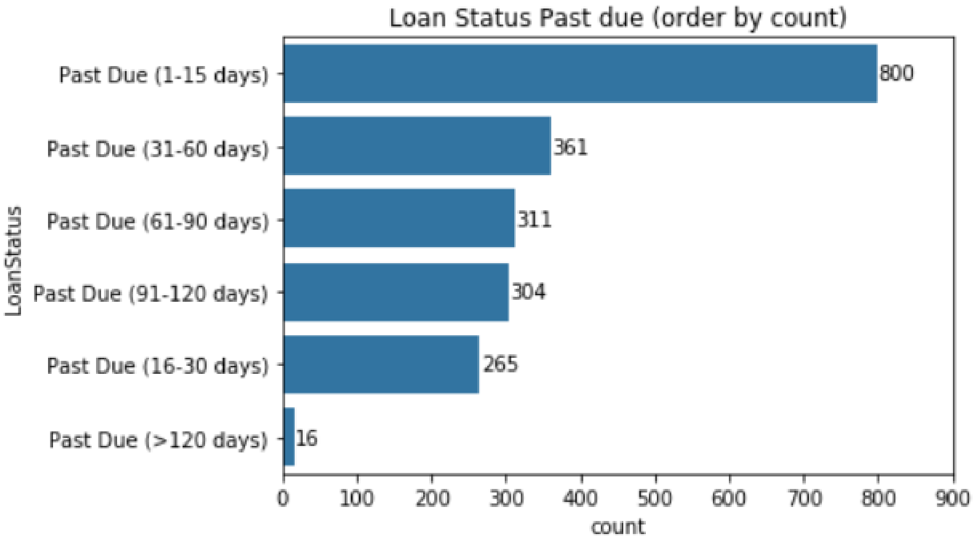


Credit Score range data was analyzed to see if there is any big difference between a lower score and high score. The above two histogram claims that there is only a very small difference in the credit score range.

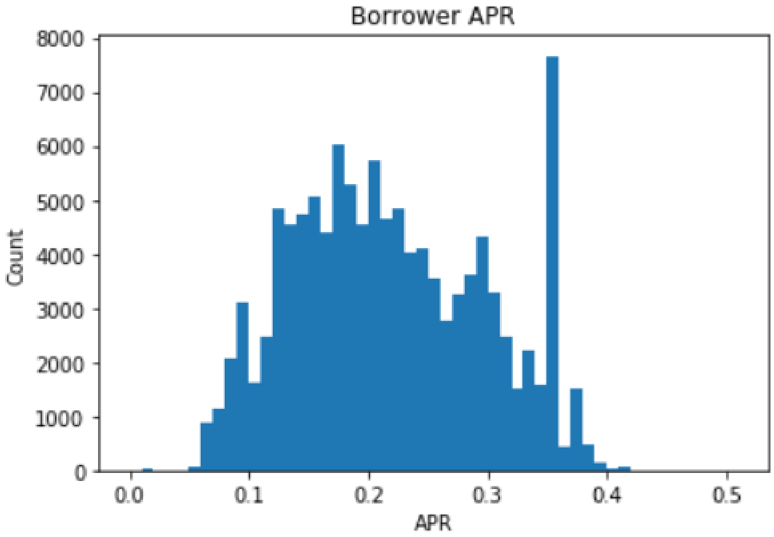
Moreover, when compare the credit score graph with credit grade graph, they have a similar shape when converting credit grade shape from right to left, credit score data and credit grade data match each other. It seems that a credit score around 700 is grade C in credit grade data.



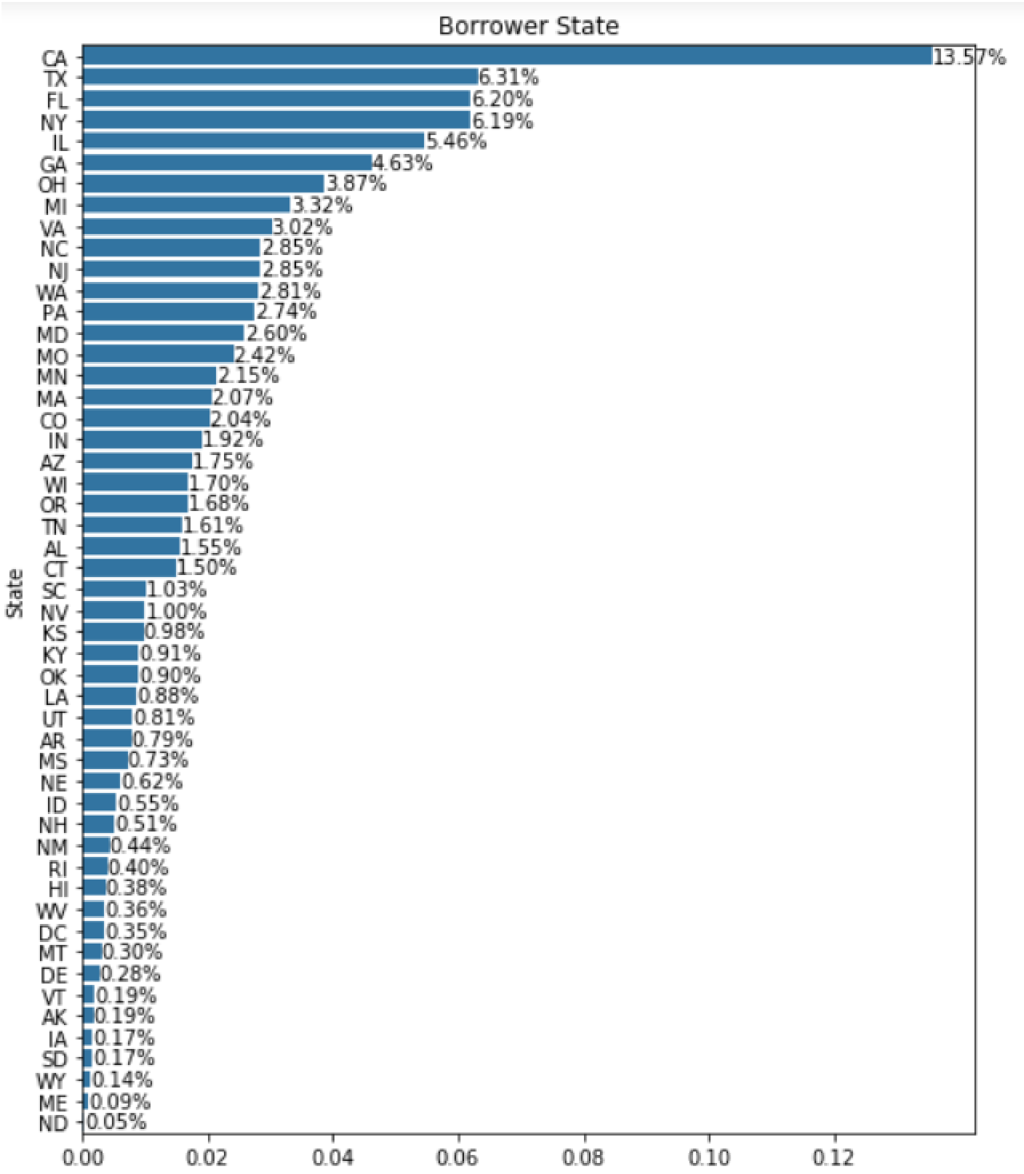
Prosper rating grade was plotted to see if there is a difference between regular credit grade system. Prosper loan company rated borrowers without credit information on their data. Prosper rating have a higher count since prosper have fewer null values; their highest count is also graded C. Prosper rating seems to have similar grading quota as a regular credit grading system.



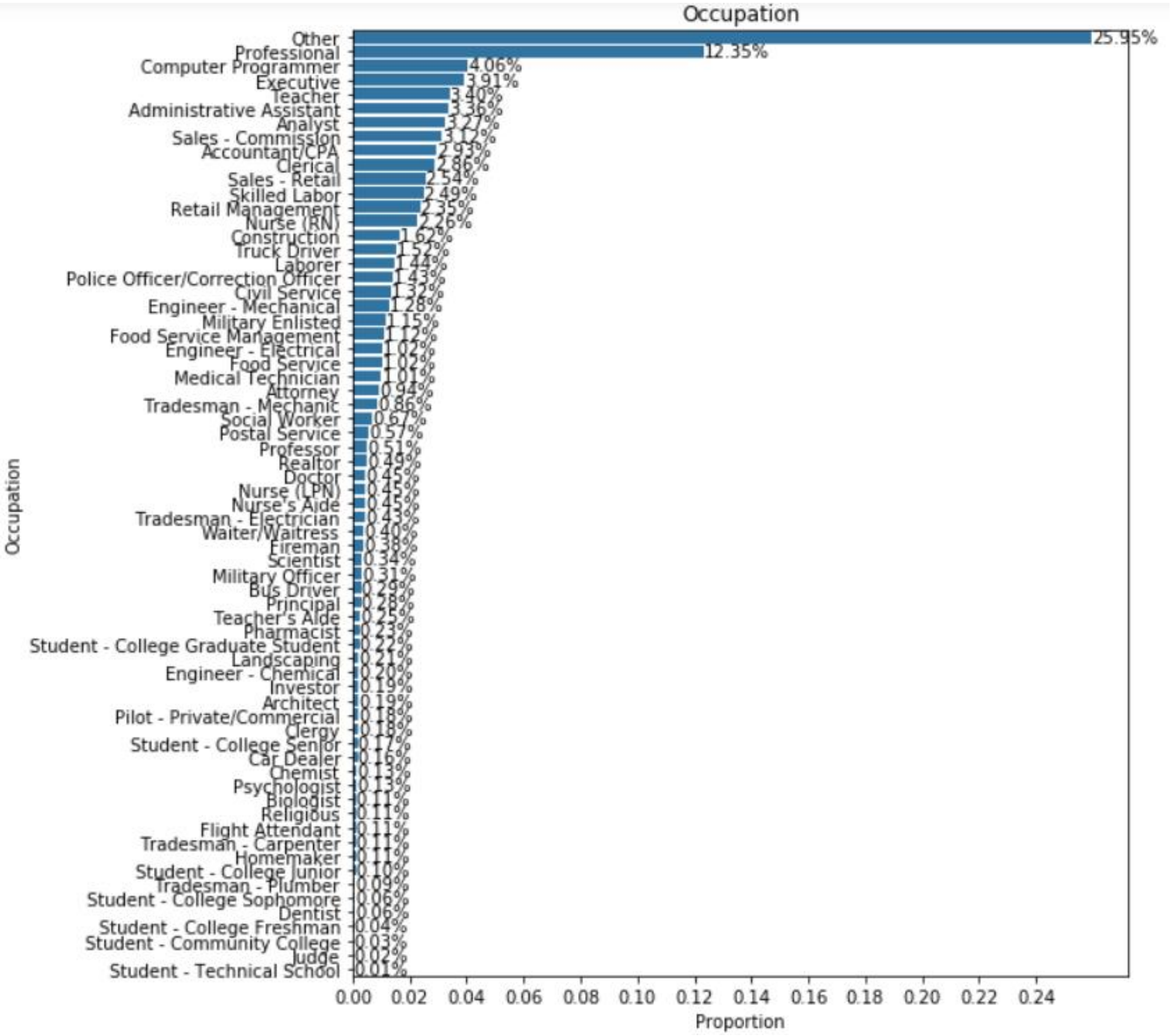
On Loan Status data, with all paid or current data in the graph, the count plot did not show all past due data because of the huge gap in the frequency. Only past due data was extracted to see past due patterns of loan borrowers. On past due count plot around 1- 15 days had the highest proportion of all past due. Even though past due count has slightly increased in day 31-60, past due frequency has decreased over time.



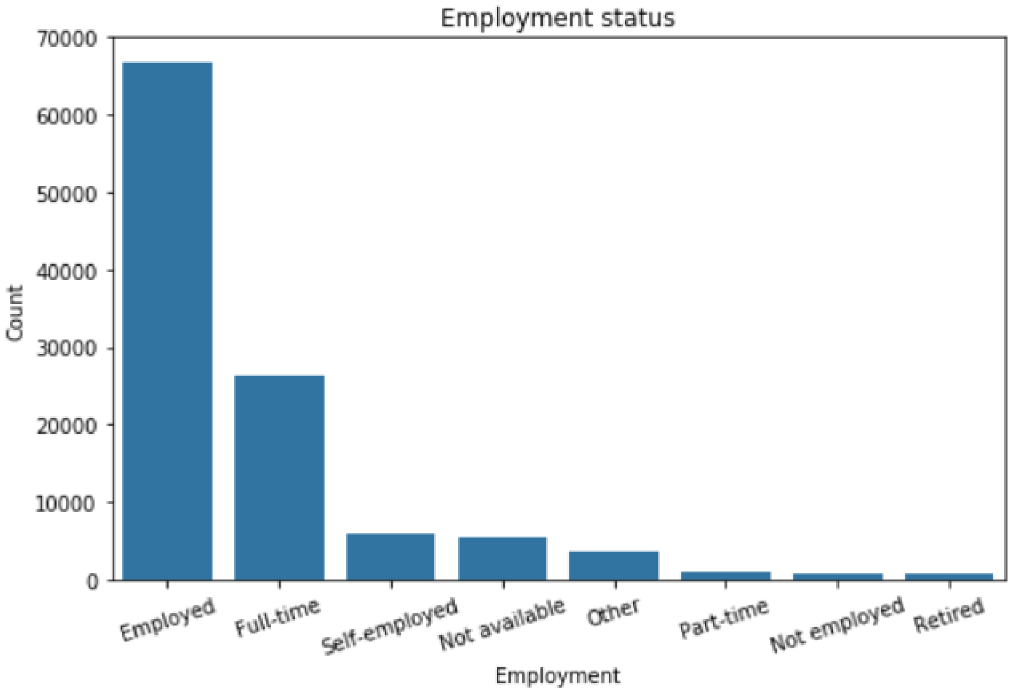
Borrower APR was plotted on count plot; there is a high spike between 0.3 and 0.4 APR. Eliminating the highest count on the graph, the plot will shape close to normal distribution.



Borrower State count-plot were plotted to see where borrowers are from. California had the highest count of 51 states recording 13.57% of all borrowers; the second was Texas with 6.31% proportion. Adding top 10 states proportion, the ten total proportion was around 55%.



Occupation count-plot was plotted to see which occupation is the borrower of prosper loan. The highest proportion of occupation was "Other" which indicates that it is not in the prosper loan occupation category. Top 5 occupation on the dataset is almost 50% of the borrowers.

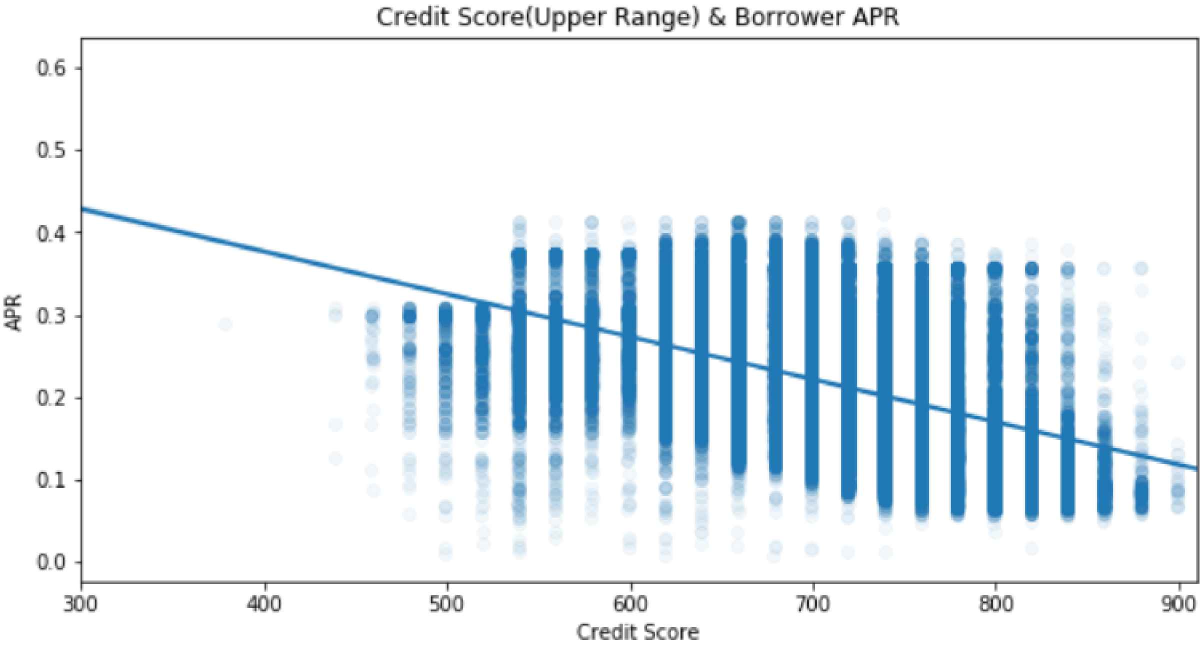


Employment status was plotted in count plot to see if borrowers are employed or not.

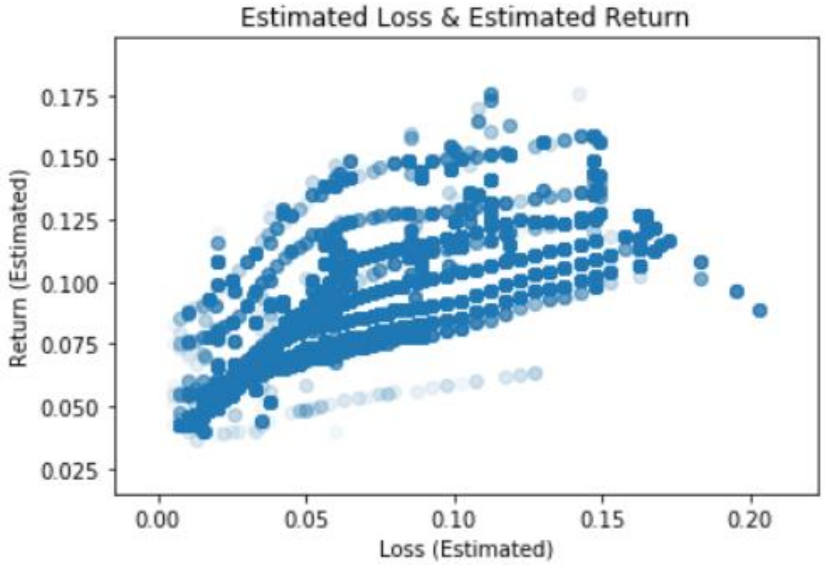
The graph above indicates that most of the borrowers are employed. Through this employment status, there is a very low chance that ‘Other’ on occupation is unemployed.

- **Bivariate Exploration**

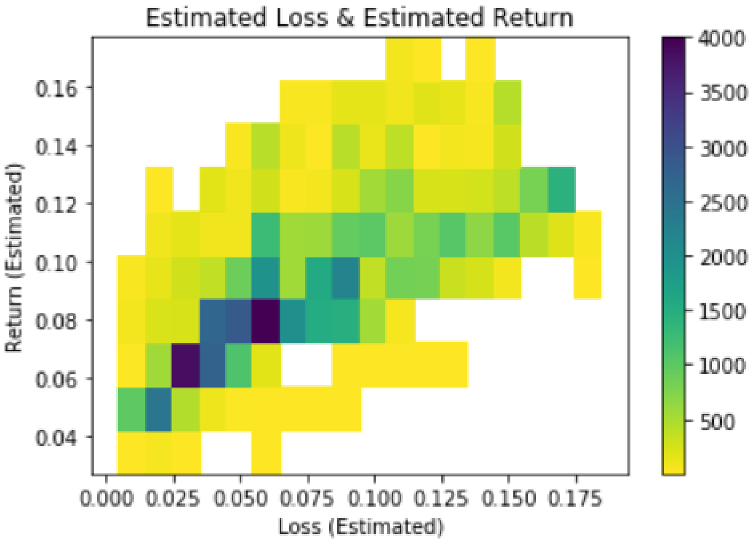
Bivariate Exploration analyzes two variables at a time. Scatter plots, heat map, box plot, and time series were used to see the relationship between two variables.



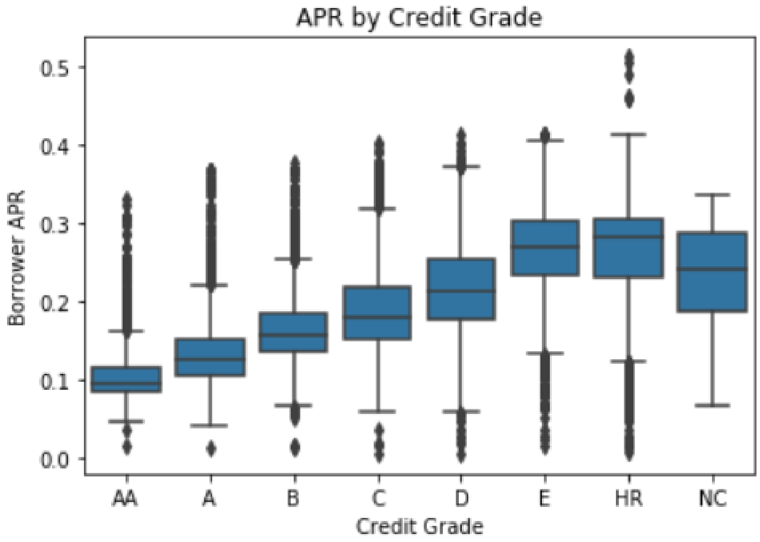
Credit Score and Borrower APR data were plotted on scatter plot to see the relationship. X-axes limit was held from 300 to 900 to eliminate outliers on 0 credit scores. When plotted without alpha rate, the plot points were spread out wide alpha rate 1/20 was held to see where plots are more likely to distributed. APR and Credit score have a negative relationship, as a credit score increase, APR decrease.



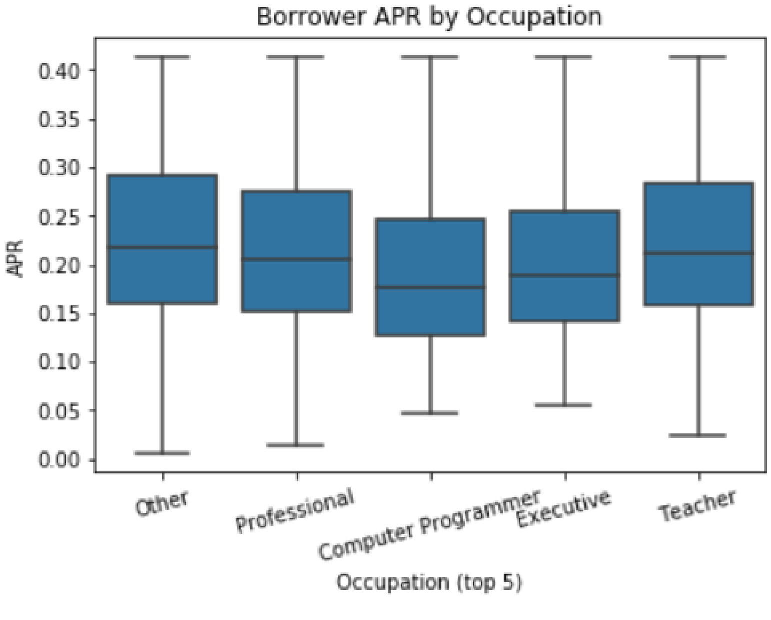
Estimated Loss and Estimated Return data were plotted on the scatter plot to see the relationship. Through the above graph, patterns can be founded; there are multiple lines with a positive relationship between estimated loss and return.



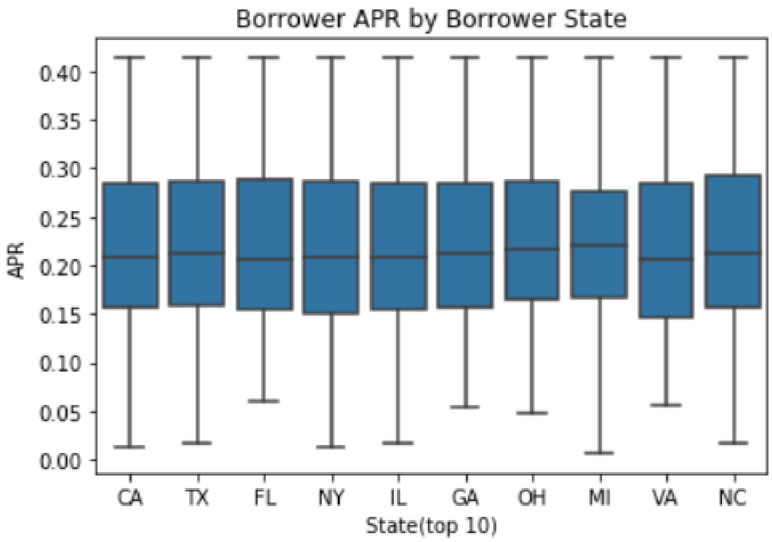
Estimated Loss and Estimated Return were plotted on the heat map graph. It is clear that estimated loss and return have a positive relationship, but from estimated loss point 0.15 estimated return decreased. There is a small concave upward relationship between estimated return and loss variables.



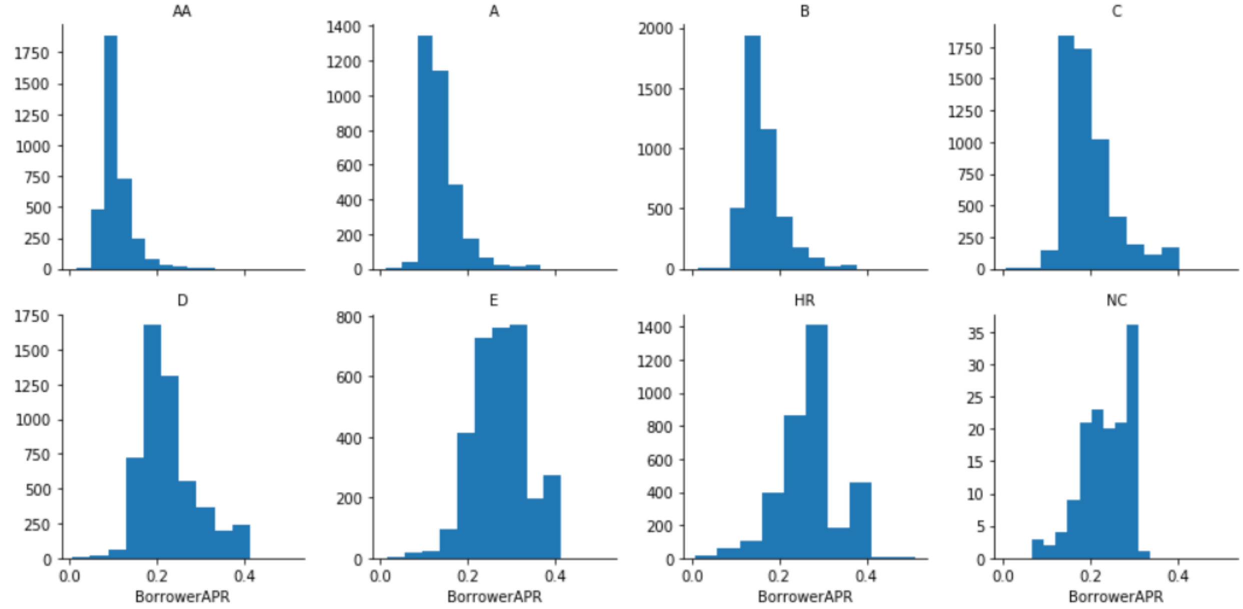
Credit Grade and Borrower APR variables were plotted on a box plot. As credit grade decreased borrower APR increased. The plot indicates that NC(no credit) has a lower APR than grade ‘E’ and ‘HR.’



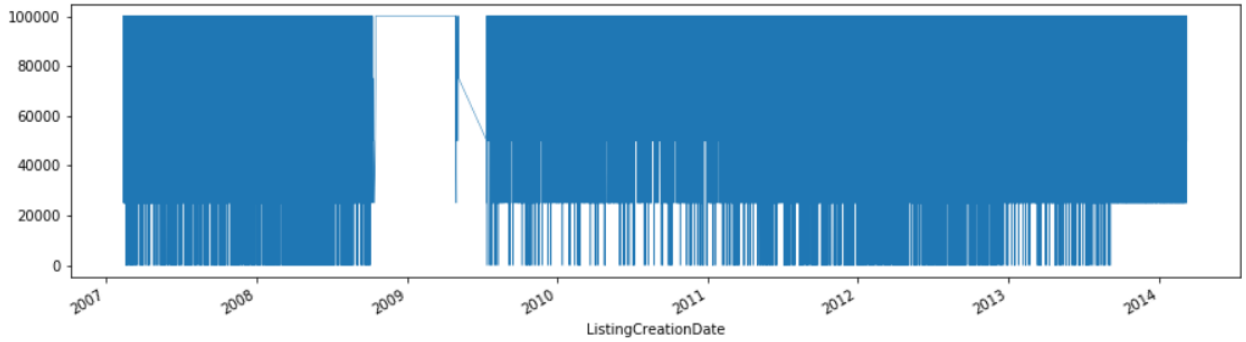
Top 5 count occupation and Borrower APR were plotted on a box plot. According to the graph, Executive occupation had the highest minimum APR, and computer programmer is the second. However, the computer programmer had the lowest median APR.

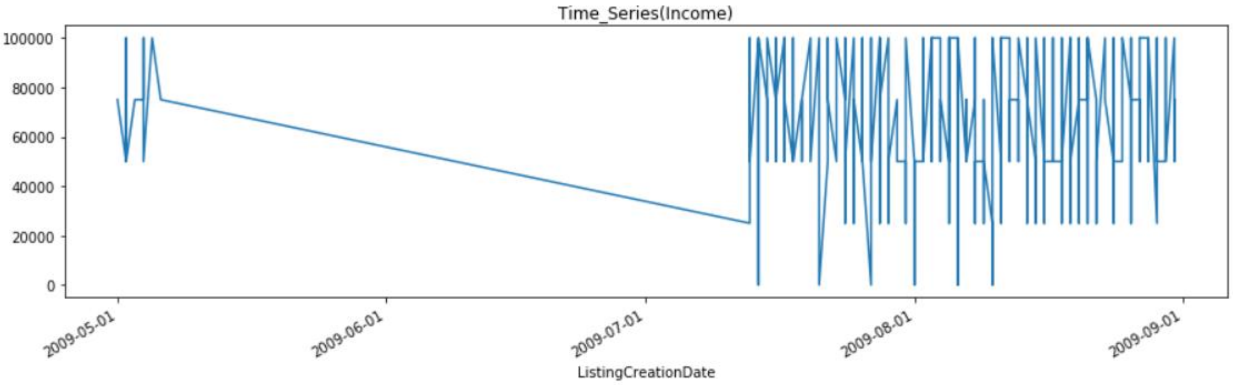


Top 10 count state and Borrower APR were plotted on the box plot. The states seem to have similar median APR, but Florida, Georgia, Ohio, and Virginia’s minimum APR were higher than the other six states.



Borrower APR was plotted on each credit grade histogram. When credit score was high histogram shaped left skew but as credit score decreased histogram peak moved toward the right. Right skew indicates that a high proportion of borrowers have lower APR and left skew means a high proportion of borrower’s APR is high.

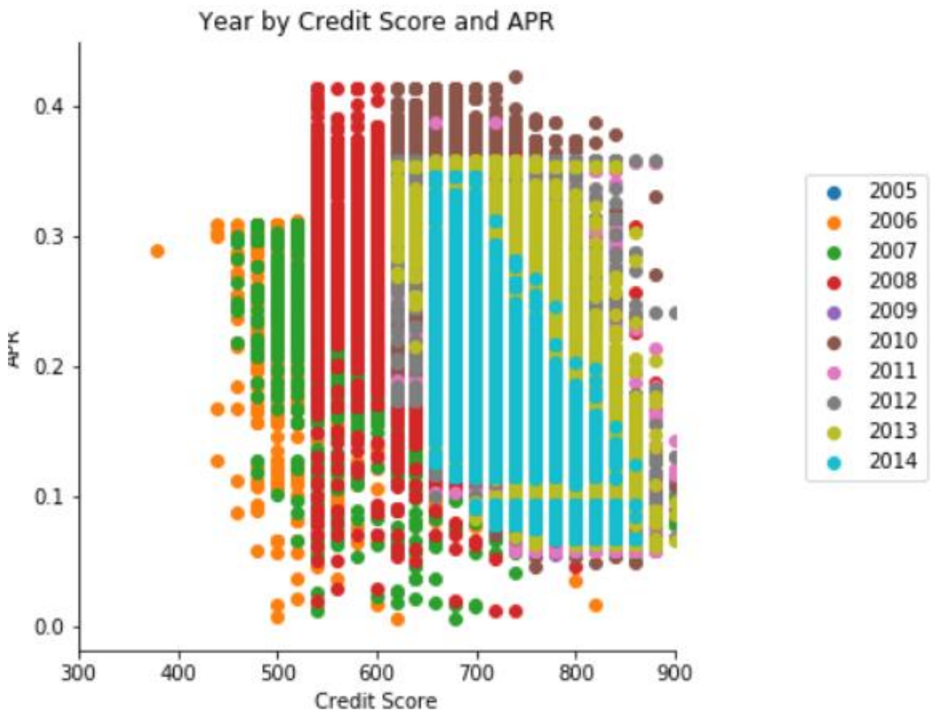




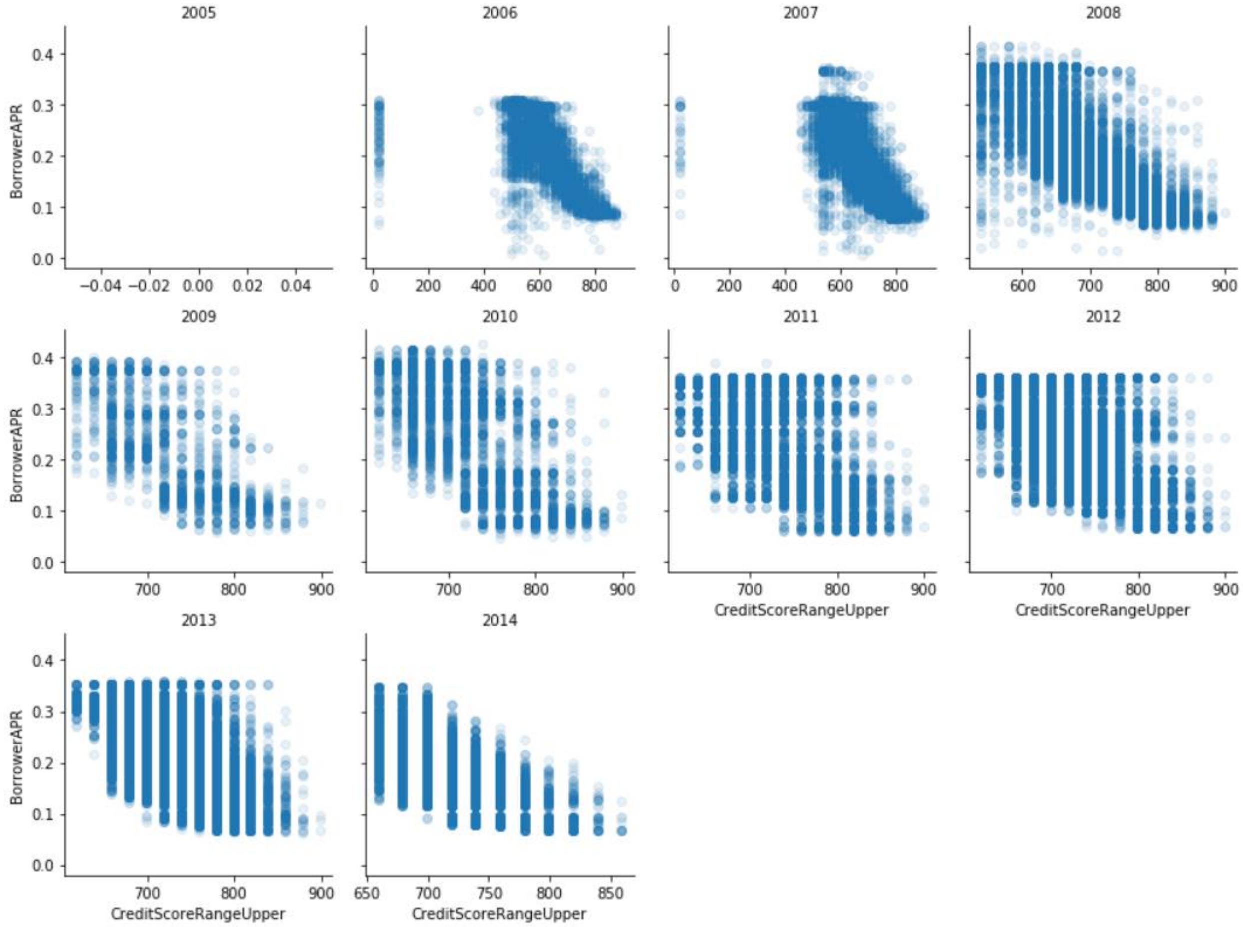
Income data were plotted with Listing creation date data. Even though the graph is not accurate; the graph shows data input gap between from 2009 May to 2009 July. There must be a data management problem or company problem between the time.

- Multivariate Exploration

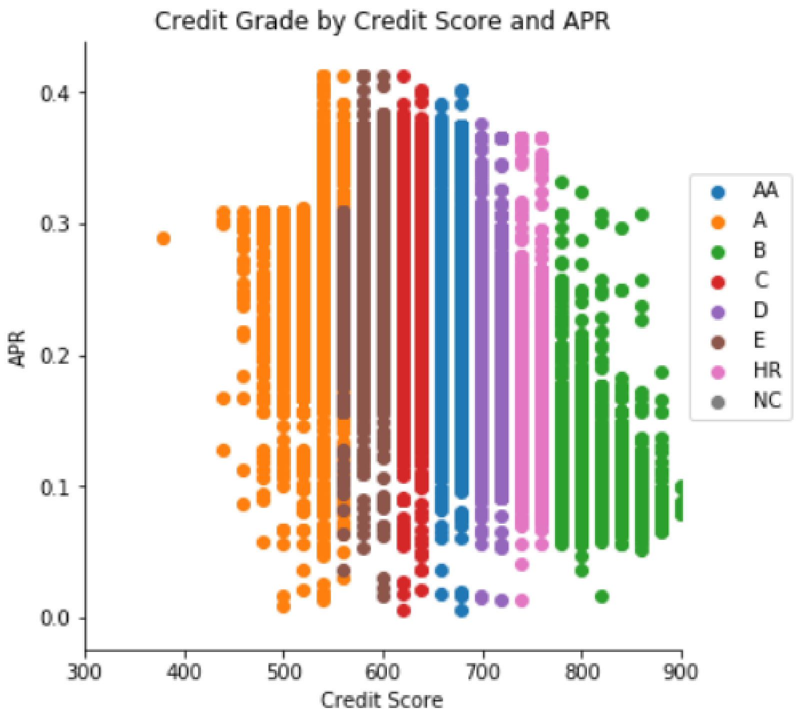
Multivariate exploration analyzes three or more variables at a time. Scatter plots were mainly used with Facet-grid to analyze data.



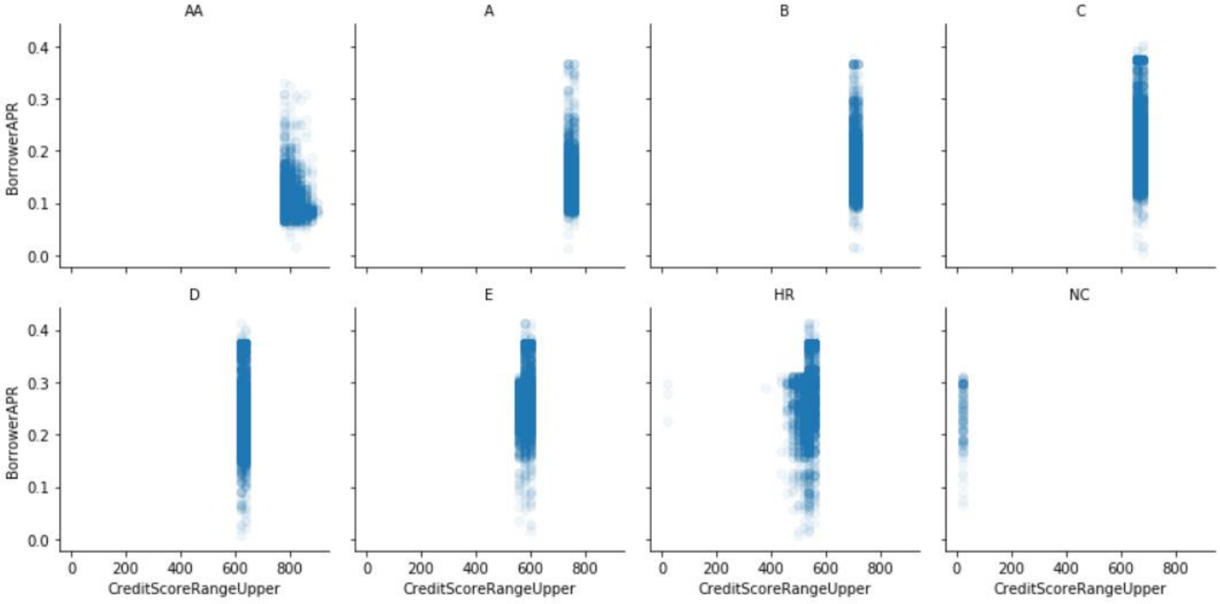
Borrower APR and Credit score were plotted by year on scatter plot. The graph clearly shows that plots are grouped by year. Facet grid was used to analyze further.

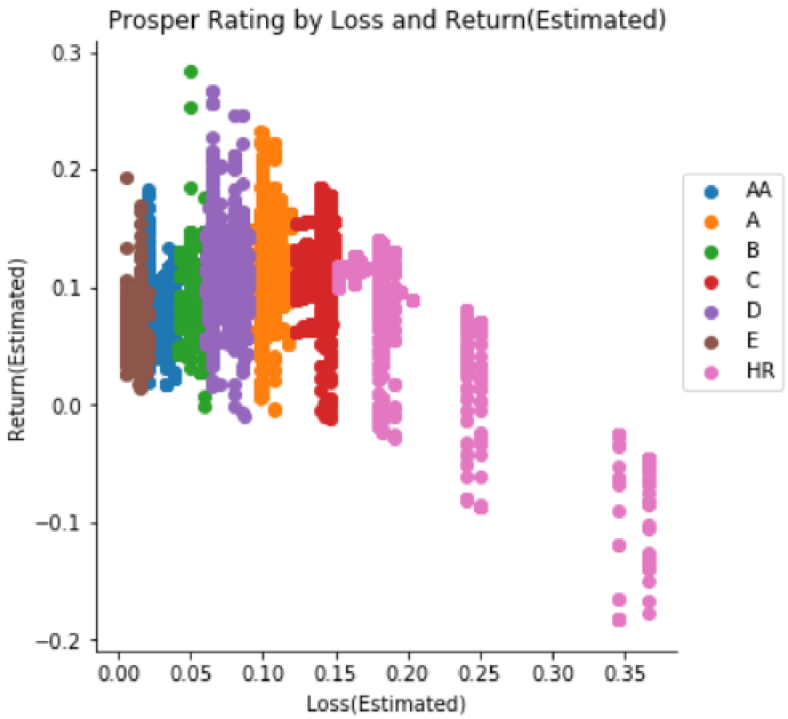


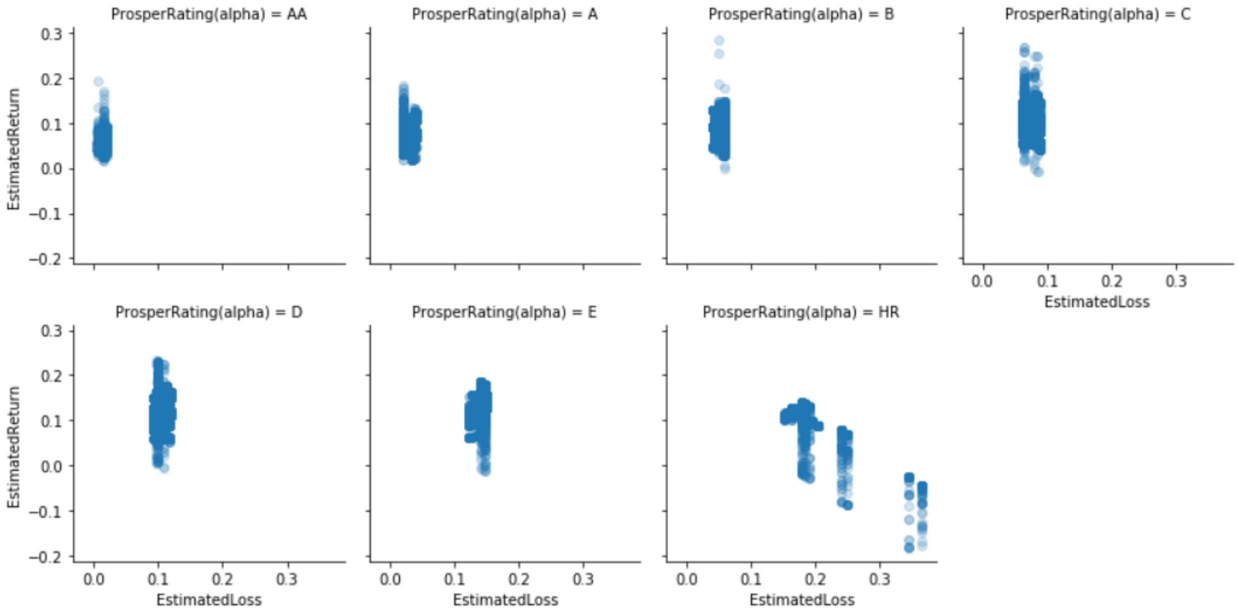
Borrower APR and Credit score were plotted by year on scatter plot using Facetgrid. 2009 data did not have information about APR and credit score. The time before 2012, there are 0 credit score data; however, after 2012, the graphs do not have any 0 credit data. Over time the scatter plots have formed more organized on the negative relationship between APR and credit score.



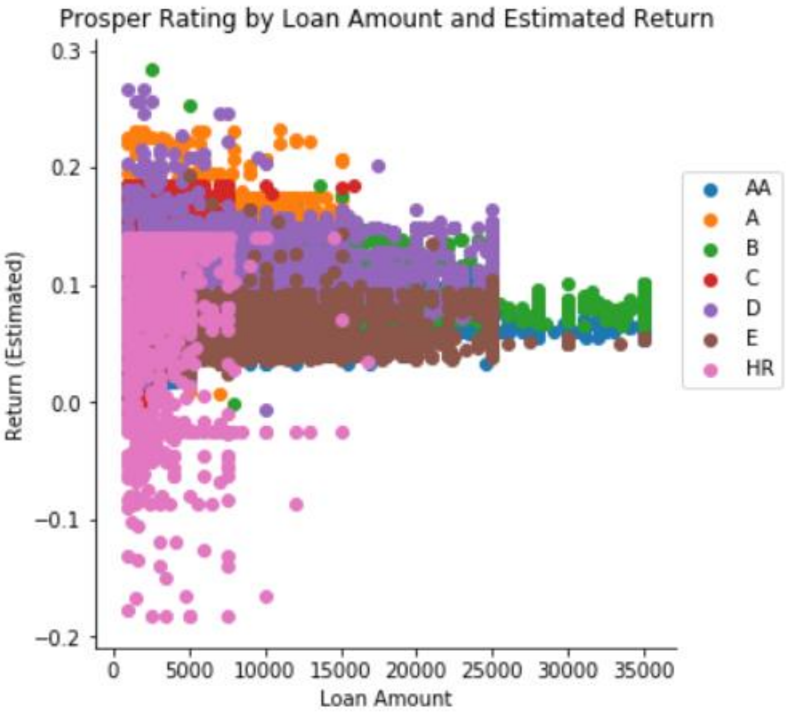
Credit score and Borrower APR were plotted on scatter plot by credit grade. A high credit score indicates high credit grade. As credit score increase APR decreased.

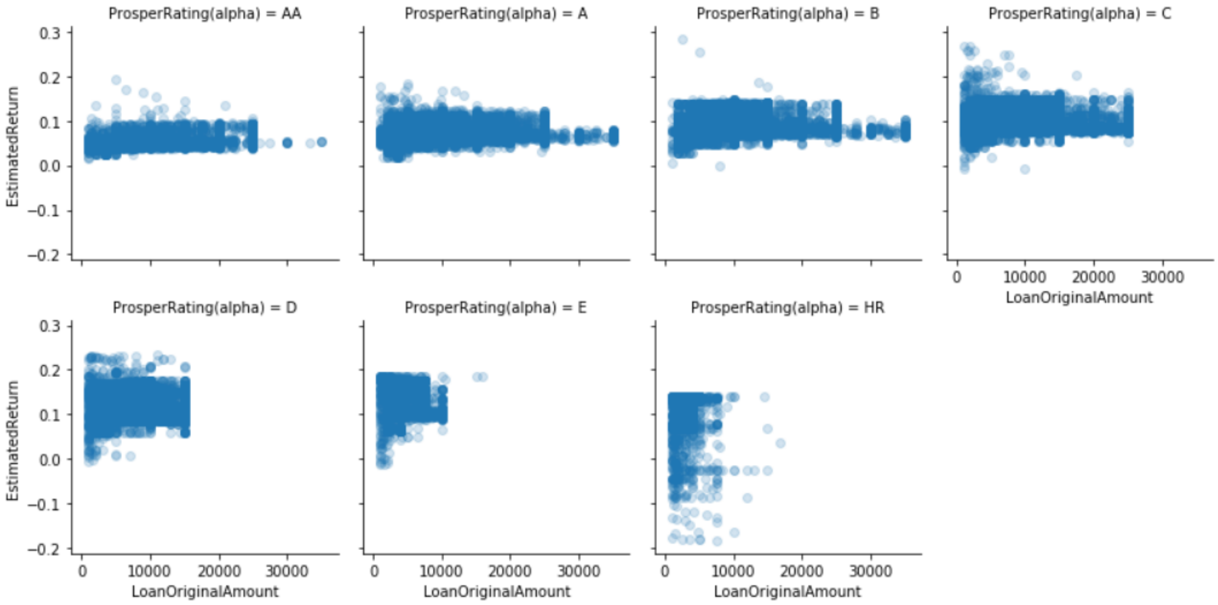






Estimated Loss and Estimated Return were plotted in scatter plot by prospering rating grade. Prosper rating divides estimated loss, and returns scatter plot. Prosper rating grade did not have a big difference on a return until rating became ‘HR,’ as rating grade reached ‘HR’ estimated return has decreased drastically. However, Estimated loss increased on each proper rating down a step.





Estimated Return and Original Loan Amount were plotted on scatter plot by prosper rating. The scatter plot was formed on horizontal shape until rating reached ‘D’ and the plot became vertical when rating reached ‘HR.’

Further analyses will be processed in the future…