NAME: IBRAHIM AYODOTUN

Prosper Loan Dataset

Cleaning and Exploratory stage

INTRODUCTION:

This dataset contain information about loan from prosper (it is part of Udacity datasets). It is a dataset with 113,937rows and 81 columns. The aim off this analysis is to provide exploratory and explanatory data visualization. The dataset would be checked for some quality and tidiness issues with some exploratory data visualization. Some python inbuilt libraries like pandas, seaborn and matplotlib would also be used.

Step one: Quality and tidiness issues

A new Data Frame ("loan_2") would be created from the original dataset because not all 81 columns would be needed for this analysis. For the step all about checking the dataset for null values also some columns would be.

- A function was created that was used to fill null values with "not applicable" and "0" for categorical and integers respectively.
- Duplicated rows was also checked for which there was none.
- Because of reproducibility of code functions are used to convert columns to their correct datatype

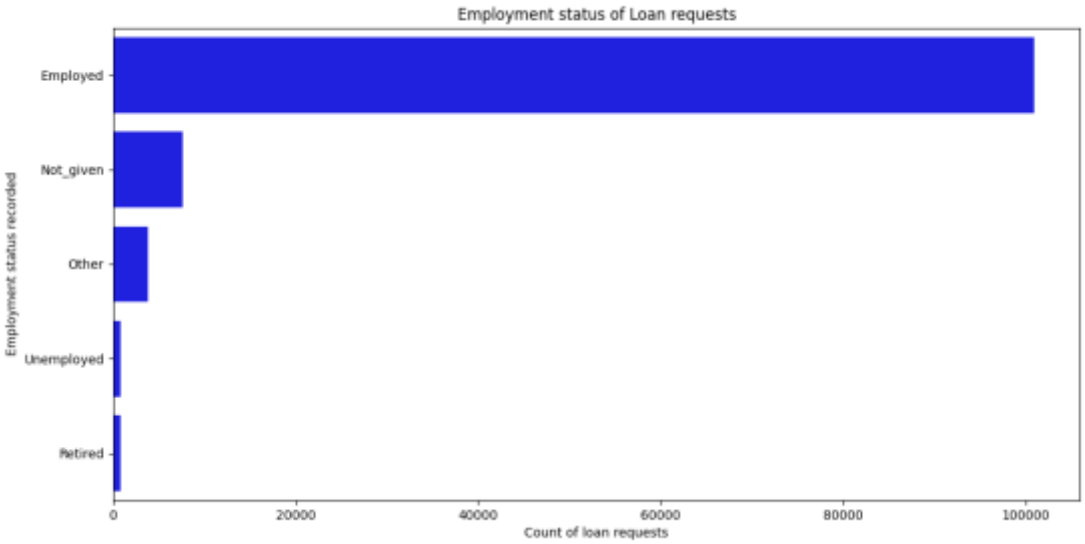
At the end of this step each column is a variable, each observation is a row and no duplicated values which makes the data tidy. The next step is to do some exploratory analysis with visualization.

Step two: Exploratory data analysis

Visualization would be produced to see some factors that influences loan approval and relationship between some variables and loan approval

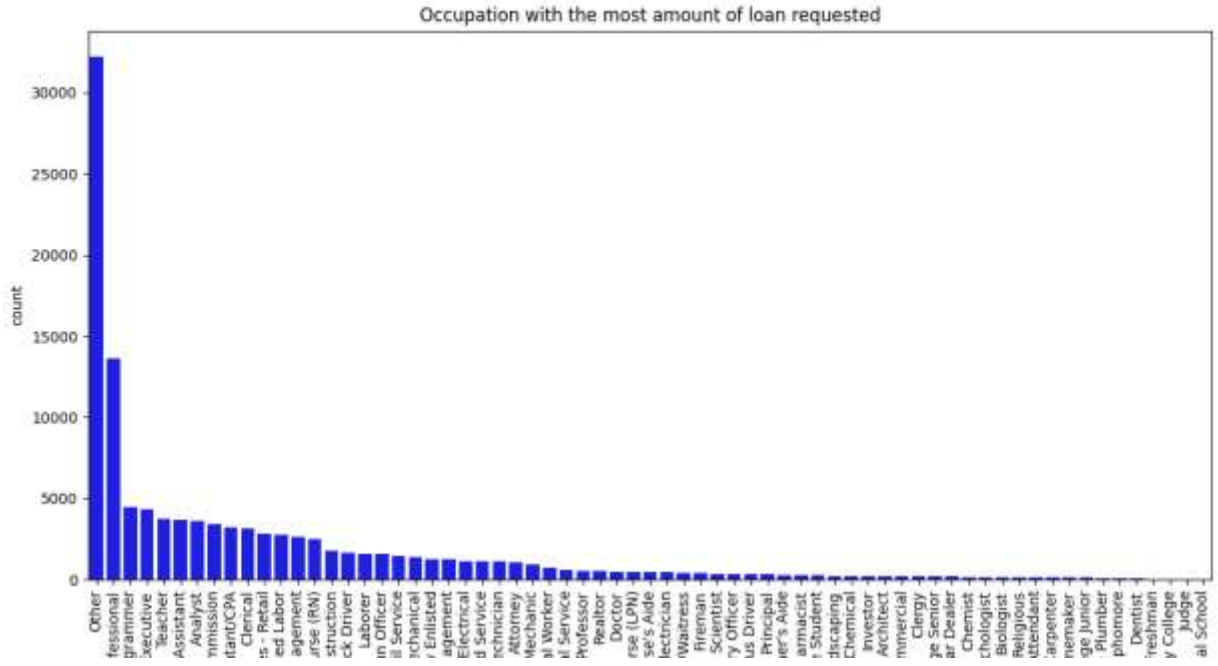
- Occupation of most people that collected Loan

From the chat below it shows most loan request are from employed individuals, this is either part time, full time or self-employed.

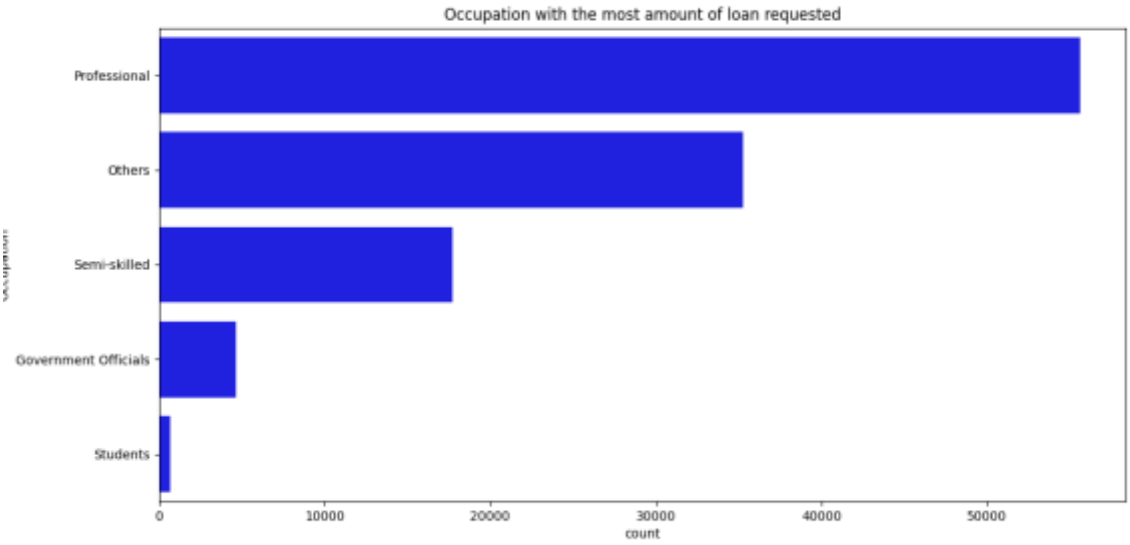


- Occupation of most people that collected Loan

  From the plot it shows occupation varies, “others” has the largest which either means their occupation was not on the list or they do not want to disclose. Student (Technical school) has the least number of loans requests.

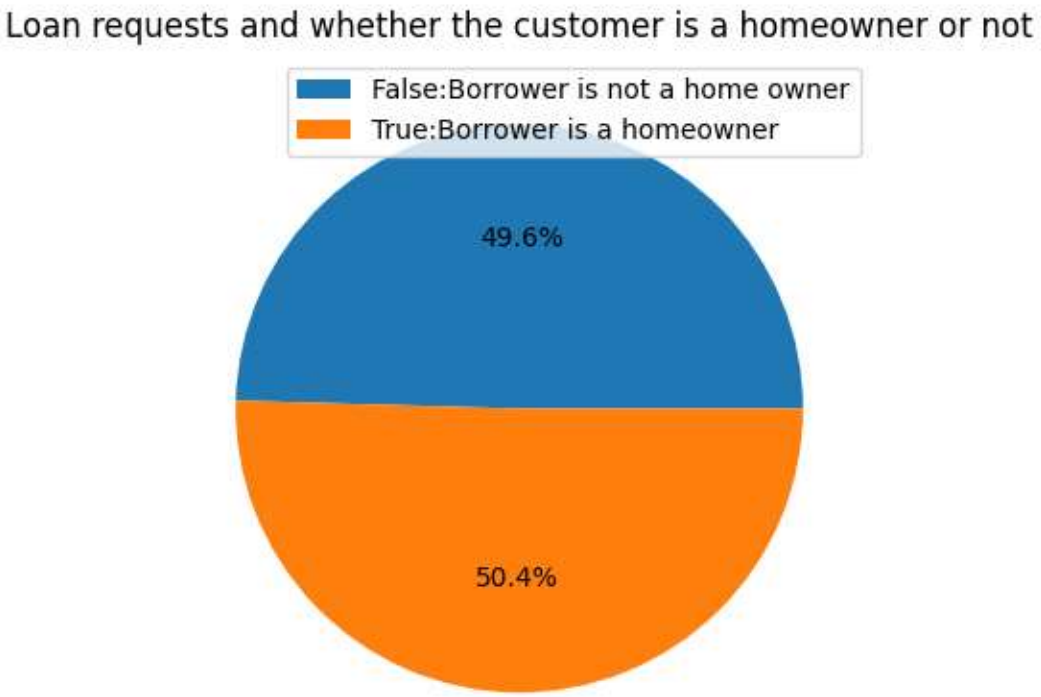


From the graph above it shows they are a lot of occupation this would be grouped into various categories because most of the occupations are just repetition, they can fit into a category. Professional consists of occupations like nurse, engineering, Judges. Others consists of occupations like car dealer, food services. Semi-skilled are administrative assistance, clergy men e.t.c



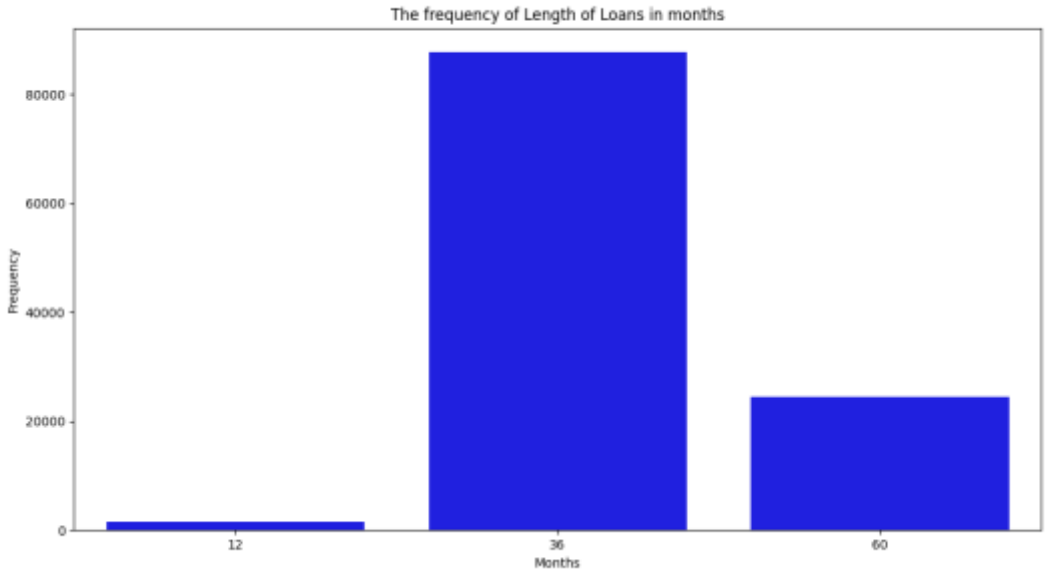
- “Homeowner” is another variable that would enable us know characteristics of borrowers

  The percentage of whether a customer is a homeowner or not is so close with homeowners less than 1% ahead, so it can be said that whether a customer is a homeowner or not do not lead to more loan requests.



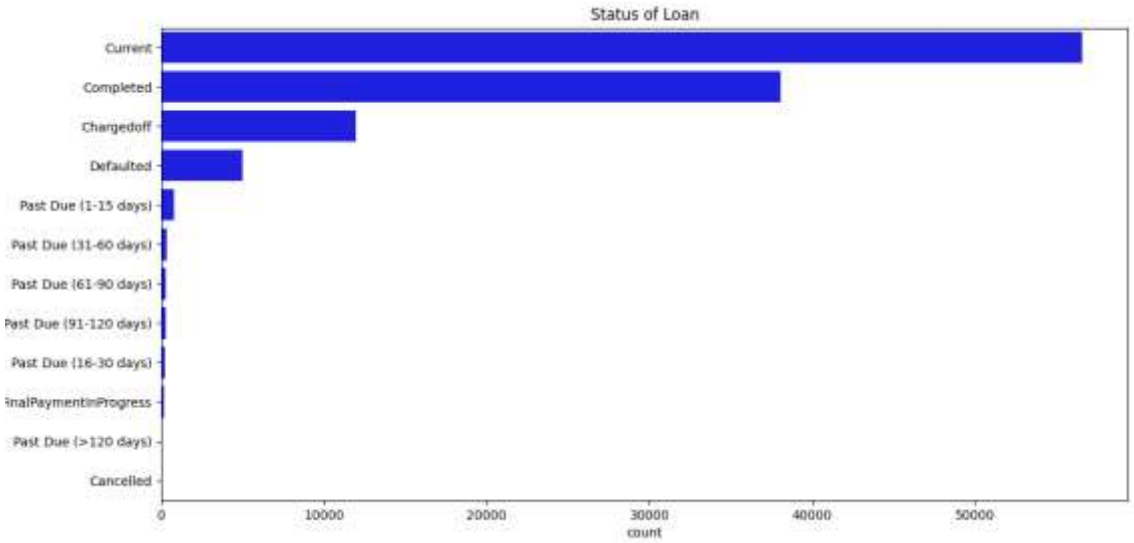
- Length of the loan

  The bar chart below shows the length of loan, most loan are for 36 months followed by 60 months and 12 months have the shortest occurrence of loans.



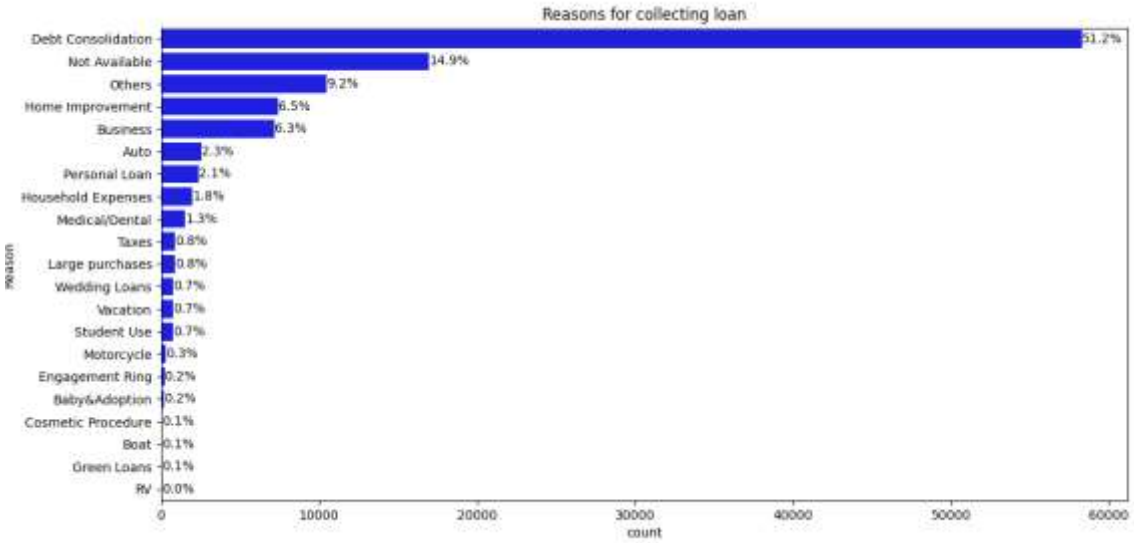
- Loan Status

  Checking for the status of loans, most of the loans are still current.

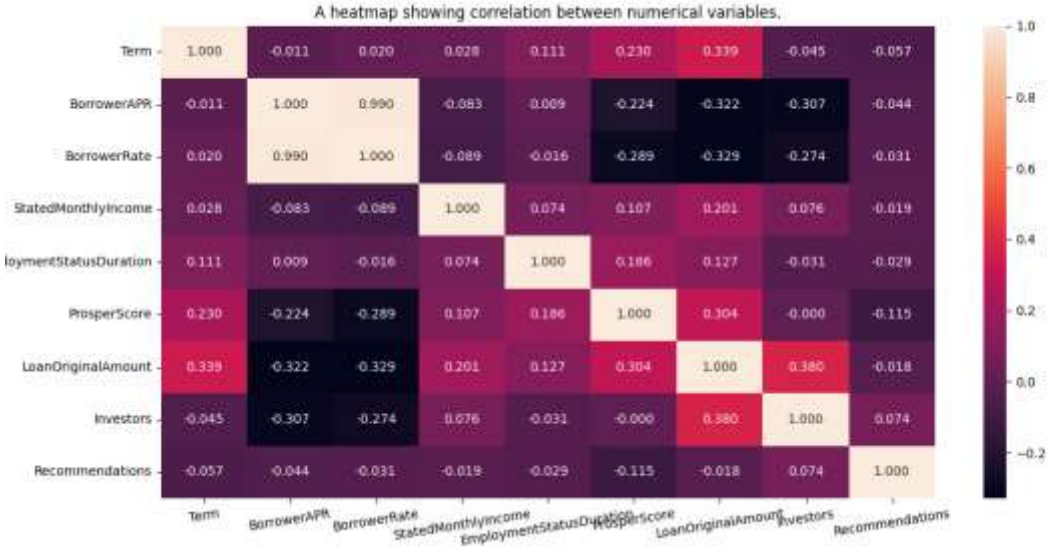


- Reason for collecting loan

  From the plot shown, over 50% of loans collected are used to settle debt.

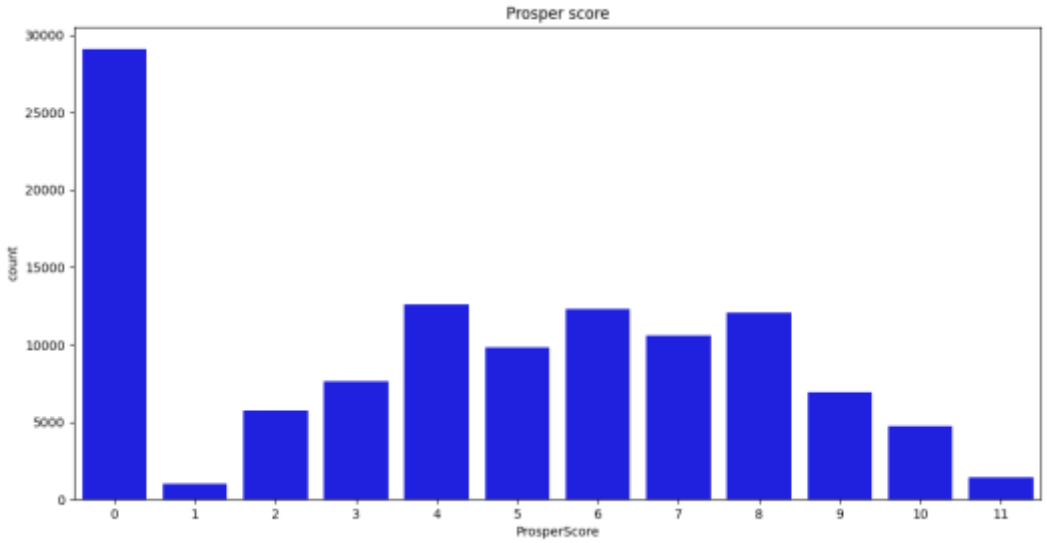


- The heatmap below shows the correlation between numerical variables. The correlation between borrower's APR and borrower's rate is high (0.990), it means they are highly correlated.



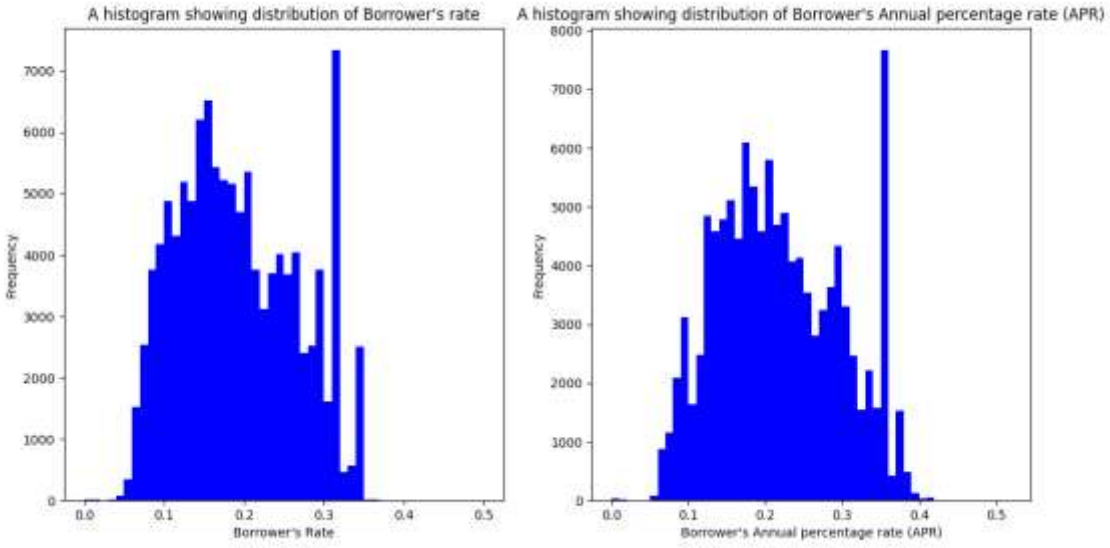
- Prosper risk score

  Prosper score is a rating the bank used for loan approvals, 0 is because no score was recorded before July 2009. Prosper score ranges from highest risk 1 through lowest risk 11. Most borrowers are scored from 4 through 8



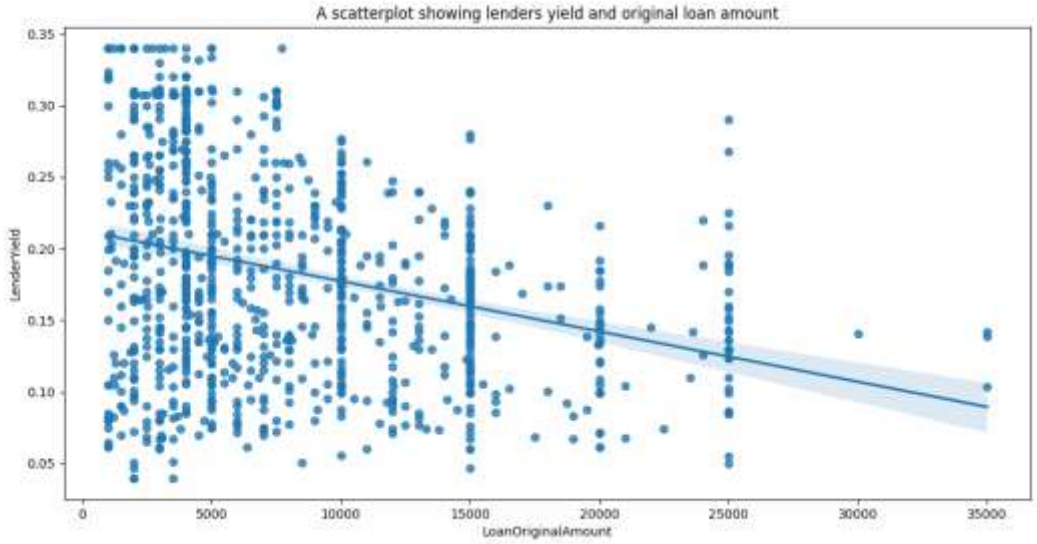
- Borrowers APR and Rate

  Borrowers APR is the annual percentage rate of the loan while borrower's rate is the annual interest rate of the loan. The distributions are multimodal, but APR would be used mostly because that is what the borrower would pay.



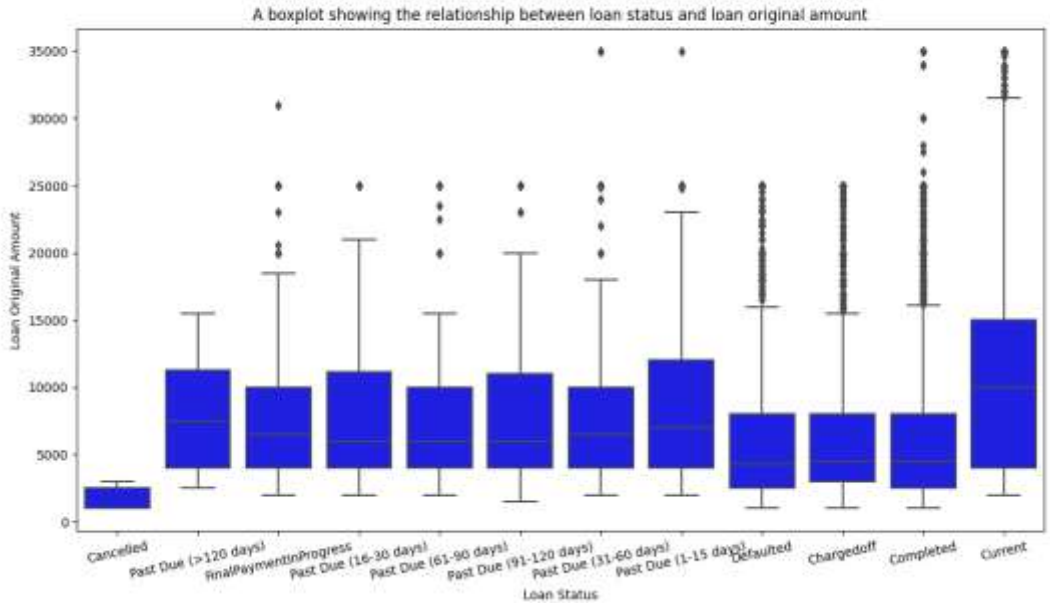
- Lenders yield and loan amount

  The scatter plot below shows the relationship between lender's yield and the original loan amount. A subset of the data (1000 points) was used for the plot because when I used all the point there was *overlapping* of points. There is a negative relationship between lenders yield and loan amount i.e as the loan amount increases the lenders yield drops



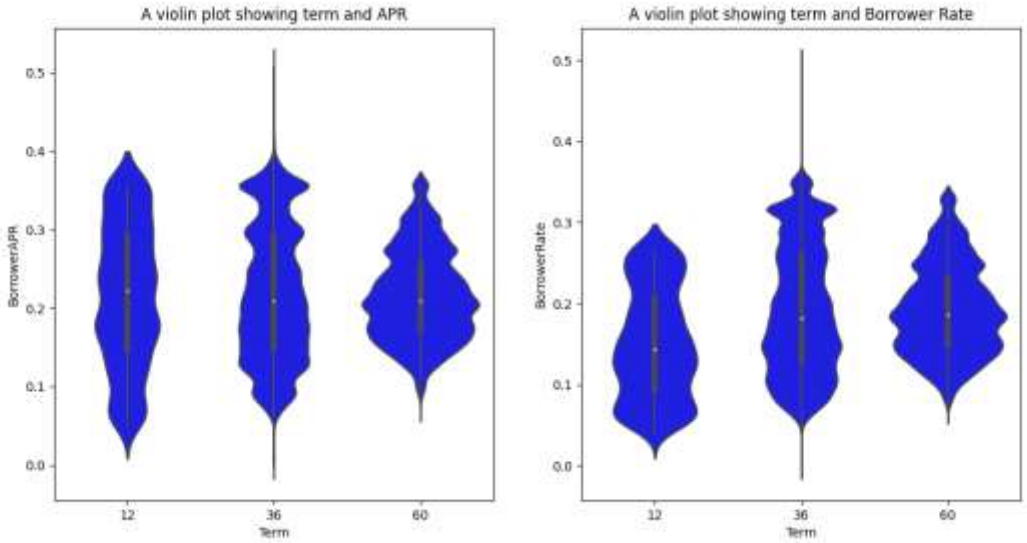
- Loan status and Loan Original amount

The box plot shown below shows the relationship between loan status and loan original amount and it appears most current loan have the most amount being taken and charged off have the least



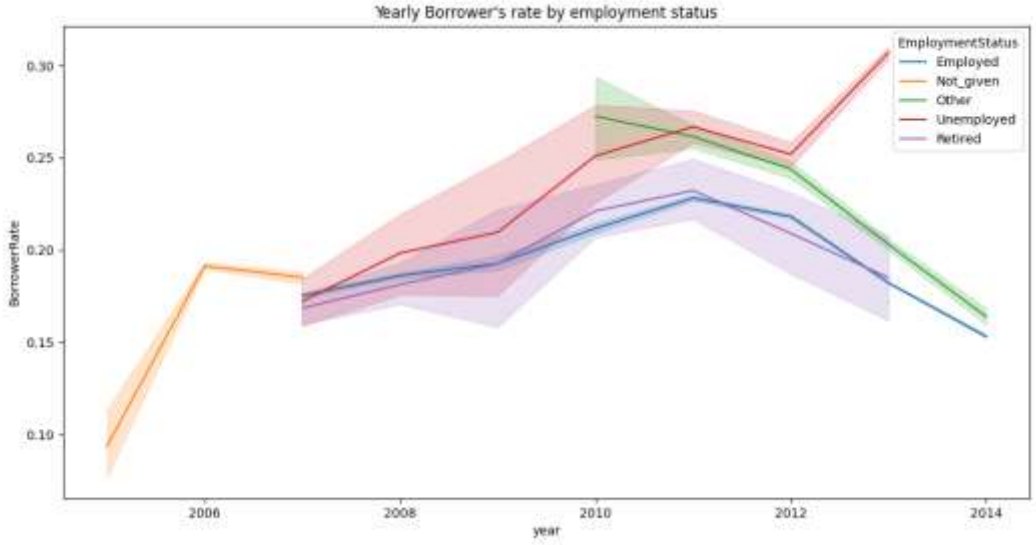
- Term in relation to borrower's APR and Rate

  The violin plot below shows the relationship between borrower's rate and APR. For longer loans the rate increases but there was a decrease in APR.



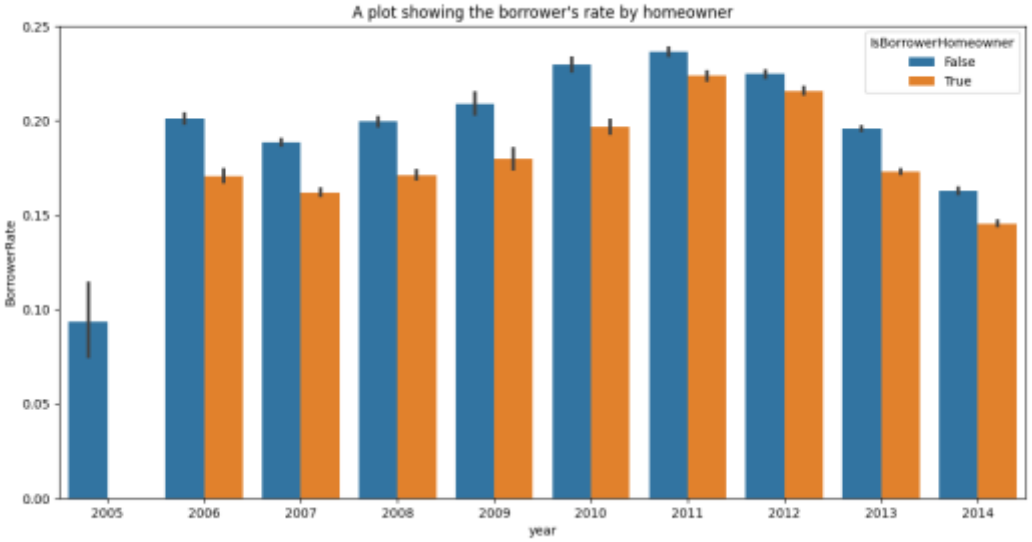
- Annual borrowers rate and employment status

  The line plot consists of three variables year, employment status and APR, from the plot the unemployed has the highest interest rate. Those employed has the lowest except in 2012 when retired persons had a lower borrowers rate.



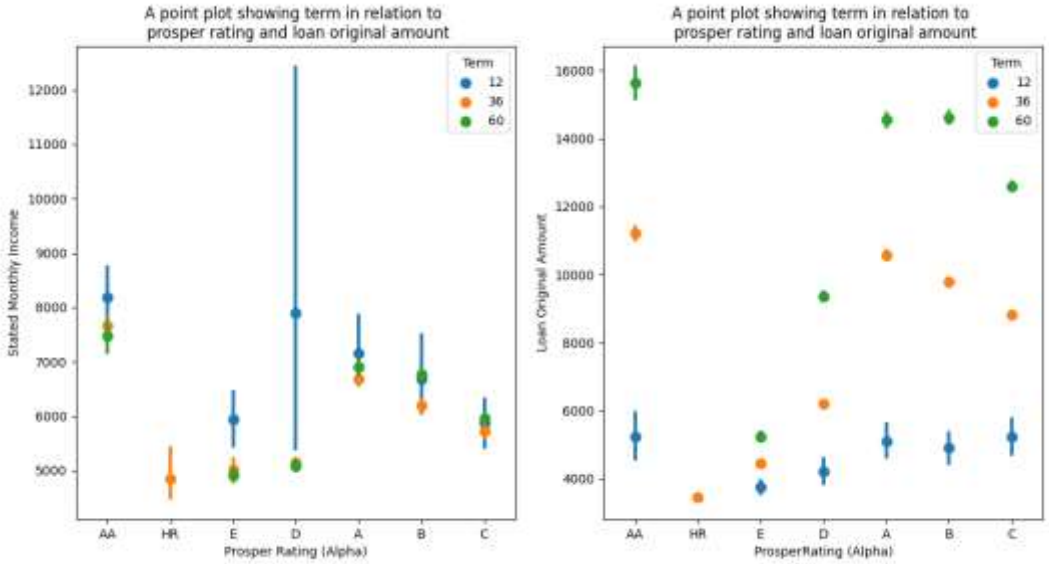
- Homeowners and borrowers rate

  The interest rate of people that are not homeowners are higher. So people who have houses have a lower interest rate.



- How the stated monthly income and loan original amount variables effect on Prosper Rating (Alpha)?

  From the point plot shown term and rating for the monthly income doesn't seem to have a relationship, but for loan amount the loan amount increases for all the term with a better rating.



- The file is saved, and explanatory analysis and conclusion would be in the other document