Loan Data Exploration:

By Dhriti Nirmal

DATASET:

This data set contains information on peer to peer loans facilitated by credit company Prosper. This document explores a dataset containing 113,937 loans with 81 variables on each loan, including loan amount, borrower rate (or interest rate), current loan status, borrower income, and many others.

Main findings from the exploratory data analysis:

My main feature of interest is BorrowerAPR (The Borrower's annual percentage rate) and the distribution of borrowers APR looks multimodal with most of the values are at the range of 0.05 and 0.4. The borrower APR is negatively correlated with the loan original amount, which means the more the loan amount, the lower the APR. At different sizes of the loan amount, the APR has a large range, but the range of APR decreases with the increase of loan amount. The Prosper rating also affects the borrower APR, which decreases with the better rating. Unemployed borrowers have greater APR and the employed borrowers have the least APR. Term doesn't really affect the relation between BorrowerAPR and original loan amount, which remains negatively correlated throughout all the terms. The loan amount increases with a better rating and the borrower APR decreases with better rating.Interestingly, the relationship between borrower APR and loan amount turns from negative to slightly positive when the Prosper ratings are increased from HR to A or better.

Outside of the main variables of interest, the distribution of stated monthly income is severely right screwed, with most of stated monthly income less than 20k. I only noticed that there are only 2219 people (0.026% of people) who make more than 15k and less than 0.003% of people make more than 30k. The majority of borrowers are employed. There are more 60 month loans on B and C ratings. There are only 36 months loans for HR rating borrowers. I noticed that employed, self-employed and those who belong to others category borrowers have more monthly income and loan amount than part-time, retired,full time and not employed borrowers. I also noticed that the current loans have the most amount being taken and charged off has the least. I noticed that except for the lowest ratings, defaulted credits tend to be larger than completed and most of the defaulted credits comes from individuals with low Prosper rating.

Key Insights for the Presentation:

My main variable of interest was the Borrower APR so I will be reflecting on the factors that affect this variable and its different visualizations depicting the relation between other variables and the APR.

Resources:

https://github.com/justinolgui/Communicate-Data-Findings/blob/master/Prosper_Loan_Data_Analysis.ipynb

https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.plot.pie.html

https://stackoverflow.com/questions/55729588/how-to-fix-the-truth-value-of-a-series-is-ambiguous-use-a-empty-a-bool-a-i

https://stackoverflow.com/questions/50319614/count-plot-with-stacked-bars-per-hue

https://stackoverflow.com/search?q=value+error+max+arg+is+an+empty+sequence

https://stackoverflow.com/search?q=move-legend-outside-figure

https://jakevdp.github.io/PythonDataScienceHandbook/04.06-customizing-legends.html

https://knowledge.udacity.com