Explanatory Data Visualization - Prosper Loan Data

Background

Prosper Marketplace, Inc. is a San Francisco, California-based company in the peer-to-peer lending industry, which directly connects prospective borrowers to investors, cutting out the middle man, which is usually the banks. Prosper handles the servicing of the loan and collects and distributes borrower payments and interest back to the loan investors.

Summary

The purpose of this project is to analyze a dataset and create a story through visualizations. For this particular exercise a dataset containing information on 113,697 loans and 81 variables was analyzed. You can download the dataset here and access the definitions for all the variables here.

The goal of this exercise is to answer the following questions from the point of view of an investor:

1. What kind of loans are available on Prosper and what are the average returns and default rates for different categories of loans?
2. Are there particular regions/states in the US where investors should be more cautious of borrowers?
3. What variables (i.e. credit rating, occupation, income) should an investor consider when assessing a borrower's creditworthiness?

Design

As mentioned above, the basic structure of this exercise is to answer the what, where and who a loan investor should consider before deciding to invest.

The following variables were identified as integral to answering the questions above:

1. LoanStatus
2. EstimatedYield
3. EstimatedLoss
4. EstimatedReturn
5. ProsperRating
6. ListingCategory
7. BorrowerState
8. Occupation
9. IncomeRange
10. Employment Status

From the onset, the following types of loans were filtered out from the analysis as they were deemed to be either not descriptive enough or loans that would be inherently risky. For example, debt consolidation which is a form of debt refinancing that entails taking out one loan to pay off many others:

1. Not Available
2. Other

3. Debt Consolidation
4. Personal Loan

A few calculated fields were created to obtain the default rates:

1. Defaulted = IIF([Loan Status] = "Defaulted" OR [Loan Status] = "Chargedoff", 1, 0)
2. Default Rate = ROUND(AVG([Defaulted]), 3)

Default rates (i.e. % of defaulted loans), estimated yield, loss and return were used as the main parameters to assess the value and riskiness of a loan.

To view the initial version of the explanatory data visualizations, open the provided Tableau workbook entitled "Prosper Loans". To view the final version of the explanatory data visualization, please click here or open the provided workbook entitled "Proper Loans (Final)".

What types of loans?

The first bar chart on the top-left, "# of Records", is to provide an overall sense of what types of loans are most popular. The line chart below that is to demonstrate which loans generate the highest returns, but to also show that loans with higher returns in general have higher default rates. The scatter plot on the right further emphasizes that higher yield often means higher risk of losses and helps to compact a lot of information into one visualization and at the same time, helps to highlight student use loans as an outlier.

Assuming a risk-adverse investor, for the rest of the analysis, only loan categories with default rates < 10% were considered.
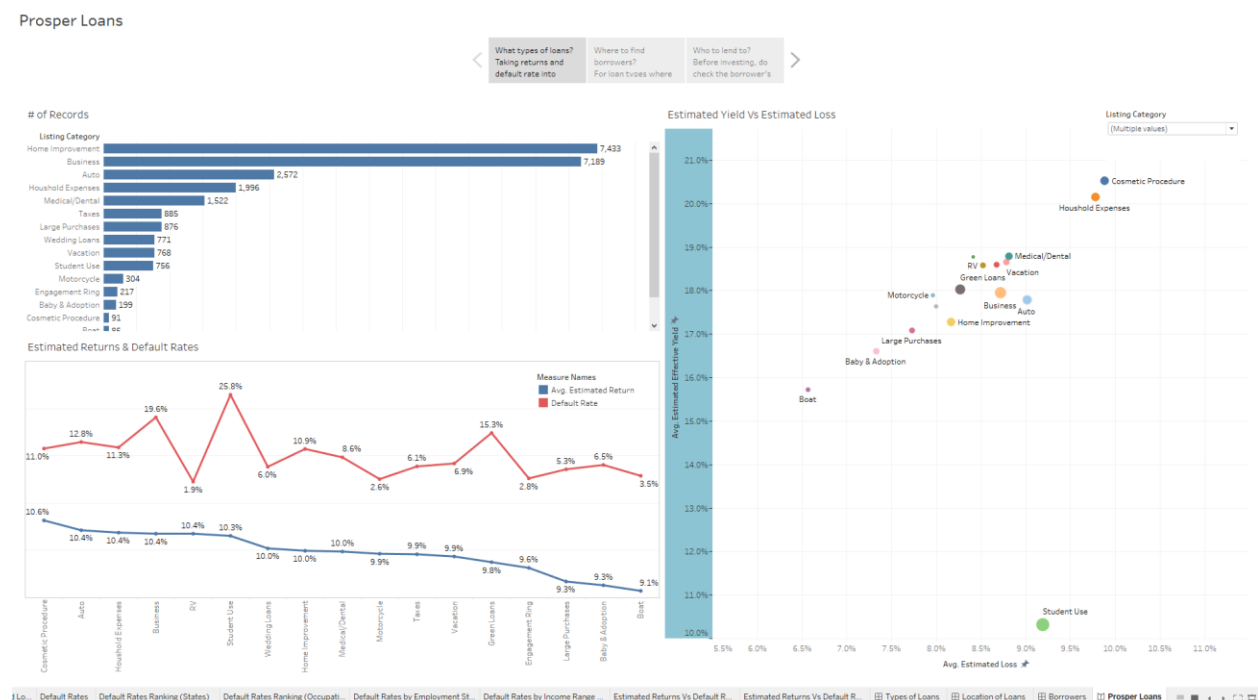
Where to find good borrowers?

Two plots, one geographical and one bar chart, were used to highlight in which states default rates were the lowest for loan categories where default rates were less than 10%. The rankings help investors to immediately identify that, statistically, they would have higher chances of finding good paymasters in Arizona, Illinois, Missouri, New Jersey and Indiana.
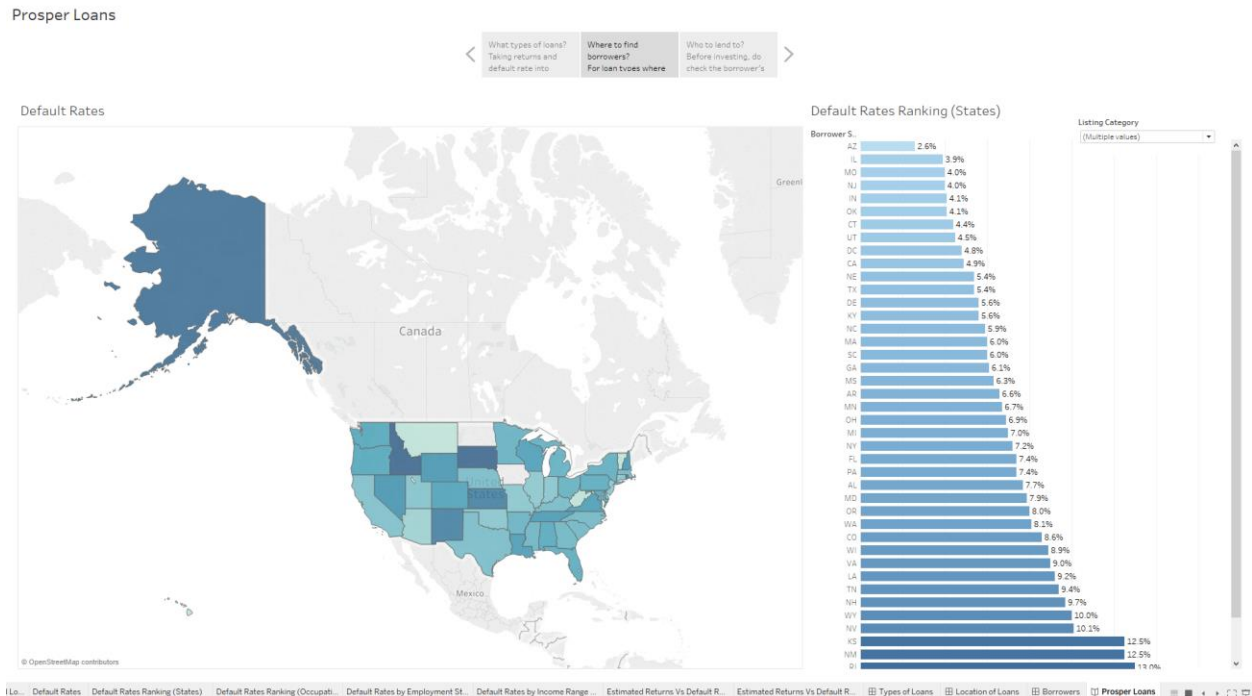
Who to lend to?

A few parameters were used to try and determine how a "good" borrower could be identified:

1. Employment status
2. Income verifiable
3. Occupation
4. Income range
5. Prosper rating

On the left side of the dashboard, the relationship between the five parameters above and riskiness were explored. At the top are the more generic parameters such as employment status and income verifiable, and as you proceed down the dashboard, the parameters become more detailed such as income range. Color coordinating the values in the tables helped to emphasize the level of riskiness.

On the right side of the dashboard, the relationship between estimated return and default rates are displayed based on income range and prosper rating. Both show the expected positive correlation between returns and risk, which is, as risk of defaulting increases so does the opportunity to earn more returns. Drawing a trendline helps to show that the relationship between risk and returns is different when considering income range and prosper ratings. The interesting thing to highlight is that there is a point of diminishing returns when considering prosper ratings. For borrowers rated D

and below, though the opportunity to earn increases, the incremental risk the investor bears increases a lot more.
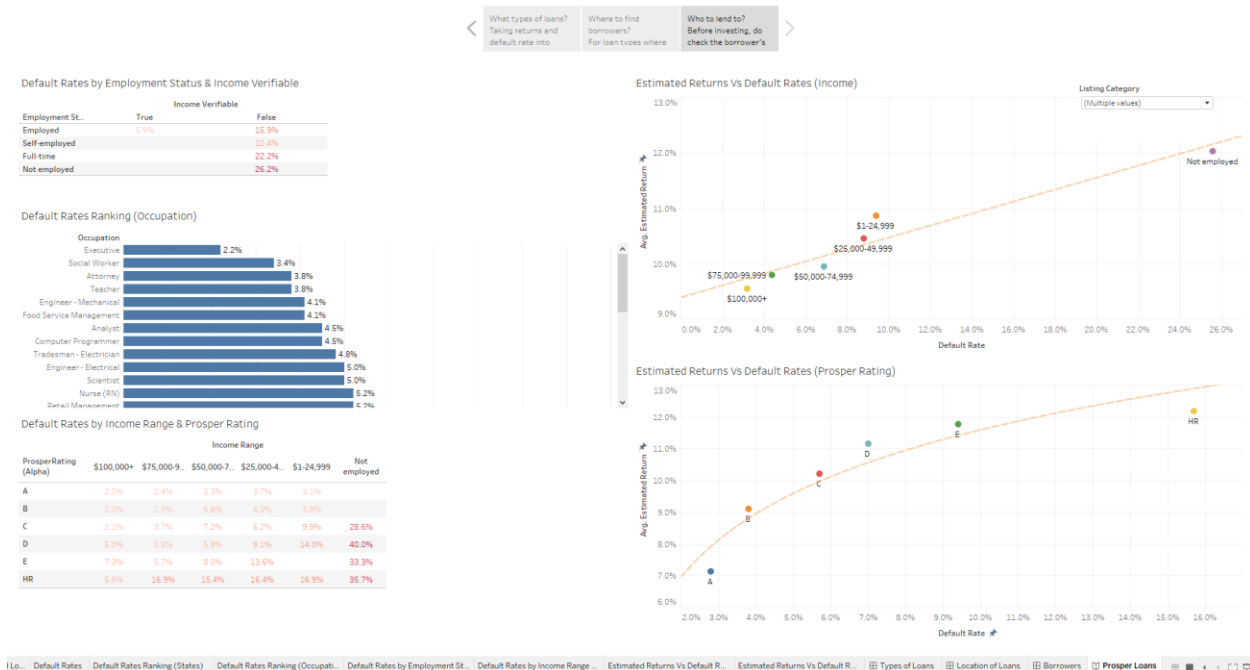
Feedback

I requested feedback on my data visualizations from friends who, should be highlighted, are not data analysts. Below are some of their feedback which I have used to improve my project:

1. Simplify plot titles and reduce size of titles
   - Reduced title font from 15 to 12
   - Removed unnecessary descriptions in the tile which are self-explanatory in the visualizations
2. Some plots are cluttered
   - Removed the axis labels and headers for plots where labels had already been added next to the line/bars themselves
3. Legend is required for the "Estimated Returns & Default Rates" by listing category to distinguish between estimated returns and default rates
   - Added legend
4. In the "Estimated Returns & Default Rates" by listing category, it doesn't serve a purpose for estimated returns to be a bar graph and default rates to be a line plot
   - Changed both to line plots
5. The "Estimated Yield Vs Loss per Category" is very informative as it clearly highlights that "Student Use" loans do not follow the same risk-reward relationship as the other types of loans. Student use loans have a significant lower yield for the losses that it tends to produce.
6. Change the "Estimated Returns Vs Default Rates (Prosper Rating)" to a scatter plot to better illustrate the relationship between estimated returns and default rates
   - Changed plot type from

7. Change the bar chart for the default rates by employment status and income verifiable to a table similar to default rates by income range and prosper rating.
   - Changed visualization from bar chart to table
8. The scatter plot of estimated return Vs default rates is very helpful to evaluate which type of ratings you would likely aim for based on your risk tolerance.
   - Added similar estimated returns Vs default rates plot for income range

Resources

1. [Wikipedia - Prosper Marketplace](#)
2. [Wikipedia - Debt Consolidation](#)