Exploratory Analysis: Loan Data

Nikhil Haas

August 10, 2015

Introduction

A data set containing 113,937 rows of data and 81 variables was downloaded. Each row contains information on a loan, including loan amount, borrower rate (or interest rate), current loan status, borrower income, borrower employment status, borrower credit history, and the latest payment information. Data is from Prosper.com, a peer-to-peer lending marketplace. Prosper.com allows individuals to invest in other individuals and handles the loan servicing on behalf of the borrower and investor. Let's ask a couple questions or form a hypothesis about the relationships between some of the variables. These questions will be explored with several plots and analysis.

Questions

Several questions come to mind, considering that this data set could potentially contain information that 1) could illustrate macroeconomic trends, 2) demonstrate how the loan industry or investors might use various criteria to approve loans, and 3) show whether or not those criteria indeed give loan originators enough information make a prediction about the risk of loaning to a borrower.

- Macroeconomic relationships
    1. Can loan origination by state be a proxy for economic health?
    2. Is there a trend in delinquency rate coinciding with a recession?
    3. Are individuals with certain job titles more or less likely to apply for a loan?
- Inter-variable relationships
    4. What relationships (linear or otherwise) exist in the data?
    5. Do people with certain occupations have more income or delinquencies, or are they able borrow more than others?

6. Is credit score a good indicator of on-time payments?
7. Is number of past delinquincies an indicator of future delinquincies?

Macroeconomics Relationships

Geographic

We can begin by visualizing each loan origination by state. There were 871 duplicate values for *LoanKey*, which were removed prior to aggregating loans by state. Each *LoanKey* is a unique identifier for a given loan, so the duplicate entries might contain information about a loan from a different point in time.

In Fig. 1, we see that a large portion of loan borrower applications are from California. Texas, Florida, New York, and Illinois also stand out on the map. From a business standpoint, this might or might not give us much information, depending on how the company is marketing and where its user-base is. Interestingly, we see some parallels between the US states by GDP (Fig. 2) and the states from which many borrowers originate: the top states by GDP are also the top borrower states. These states are also likely the most populated, as well, which could be the reason why more loans come from those states. This kind of analysis gives us an over-arching view of the geospacial information contained in the data set that could help us to dive in deeper, much like a histogram gives us an overview of data before moving on to detailed statistical analysis. Further geospacial studies could help the business identify certain market opportunities.

We can also plot trends on the US map. The data contains loan origination dates, and by separating the data by year and calculating the percent difference in loans YoY, we have a new dimension (time) to incorporate. In Fig. 3 we can see which states are experiencing the most (and least) growth in loan originations over

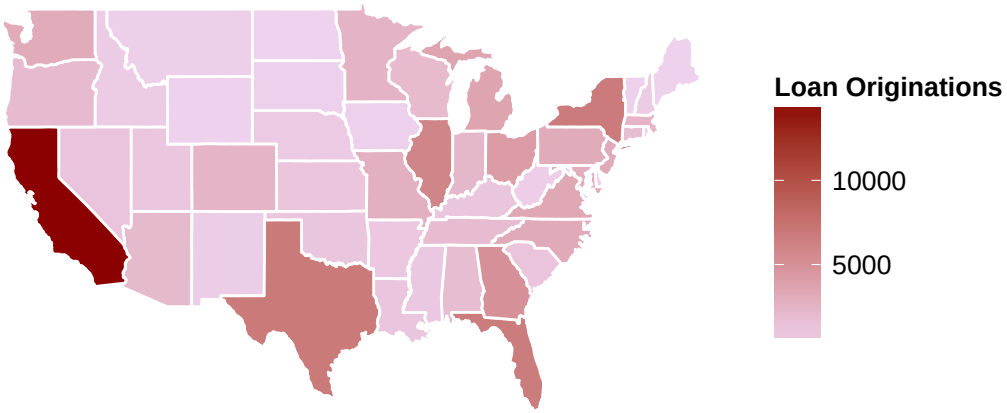

Figure 1: Loan originations by state.

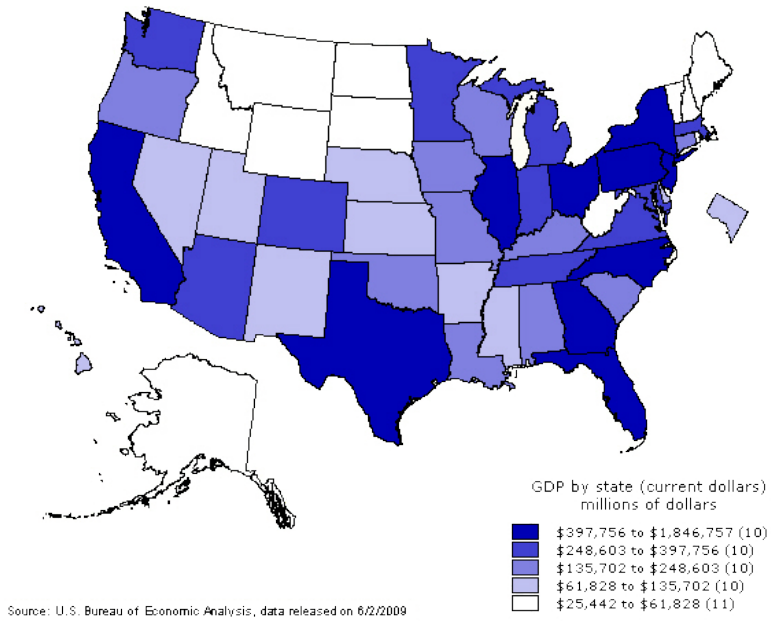

Figure 2: US states by GDP, 2008. Source: Wikipedia.org.

time. This might tell us about the sales history and trajectory of Propser.com and could be relevant for any marketing groups in the company. Fig. 3 could also be compared with trends in GDP by state to see if any states experienced negative growth when compared with change in GDP, which might indicate that Propser.com is losing market share in that state. We observe that no state experienced negative growth, so this should indicate that the business is growing overall, with some states bringing in more business at a faster rate than others.

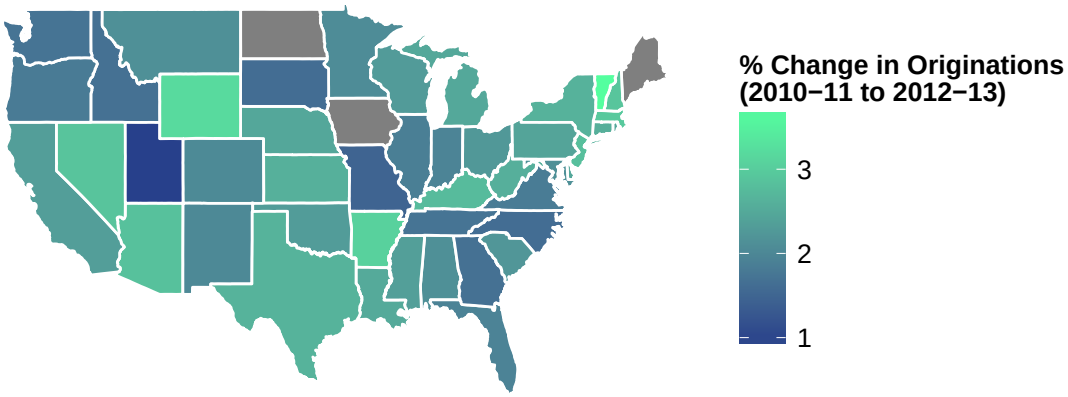

Figure 3: Trends in originations by state. Gray indicates no data available.

A lot of other interesting information was garnered from similar state maps, such as which states experienced the greatest default rates (not included). Geographic data could be explored *ad infinitum* - every statistical

measure we make could be broken down at the state level - but let's move on so we can explore other aspects of the data.

Occupation

There are approximately 63 occupations listed in the loans data. The occupations were consolidated into 43 occupations. For example, "Student - College" and "Student - High School" were consolidated into a single occupation "Student", In Fig. 4 we can see the loan originations by occupation. Certain occupations appear more than others. This doesn't necessarily mean that certain occupations are more likely to borrow or get approved to borrow; this is more likely a representation of the user base of Prosper.com.

The category of *Professional* is almost three times larger than most of the other categories. If the company were to try to use occupation to make predictions about or to better understand borrowers, Fig. 4 might suggest splitting up the *Professional* category on loan applications. This could potentially allow some hidden sub-categories of "Professional" occupations to surface. It could be the case that *Professional* has so many varied meanings that the category is useless as a predictor. Imagine if a third of Professionals identified themselves as "Technical Professionals" and another third identified as "Administrative Professionals". It could be the case that "Technical Professional" always defaulted on their loans and "Administrative Professional" never did, but since all those individuals are grouped into "Professional" that trend would never be exposed.

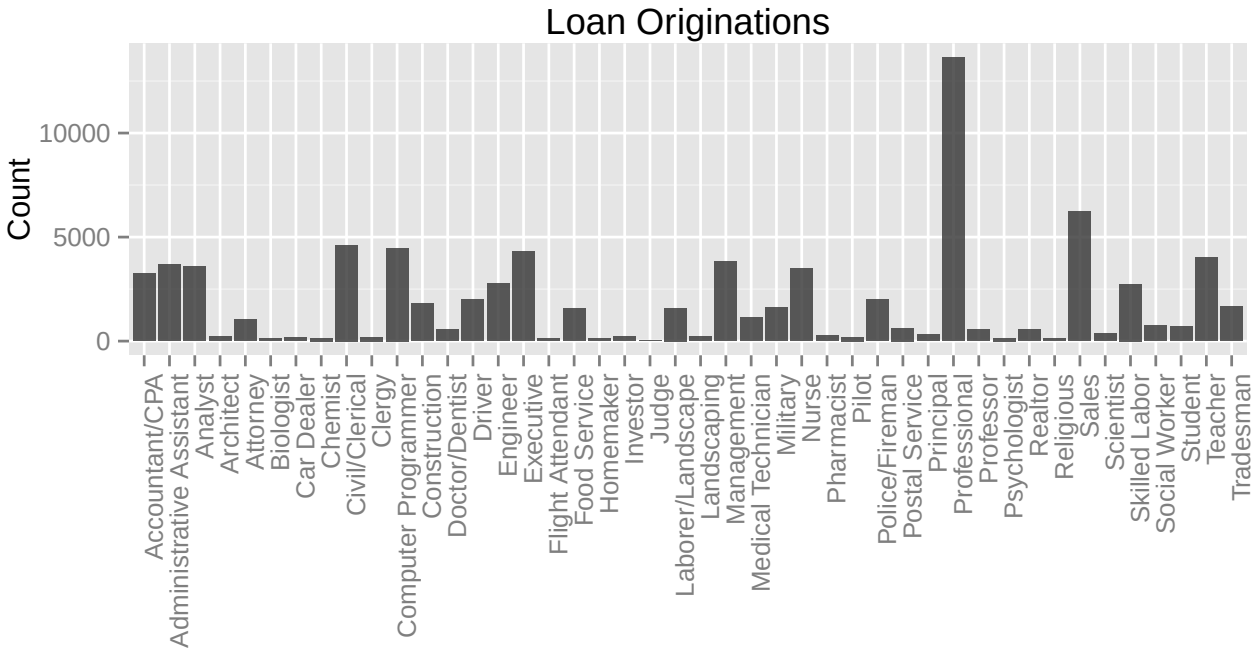

Figure 4: Loan data by Occupation

We might wonder if certain occupations experience more delinquencies than others, or if certain occupations are more likely to complete their loan payments. By calculating the frequencies of *LoanStatus* by occupation (Fig. 5), we can see that there are a high number of completed loans for students and a low number of completed loans for judges. This might seem unusual if at first glance one assumes that this indicates students are more reliable borrowers and judges are not (which would contradict our presumed understanding of those two populations). However, it is important to realize that these observed trends do not necessarilyy indicate causation. It is possible that Prosper.com first broke into the Student market and therefore that category of occupation shows a high percentage of completed loans; judges, on other hand, might only be recently exposed to Propser.com as a means to borrow money (e.g. now that Prosper.com is growing in popularity). If several judges recently opened loans, they might skew the proportion of loans that are current.

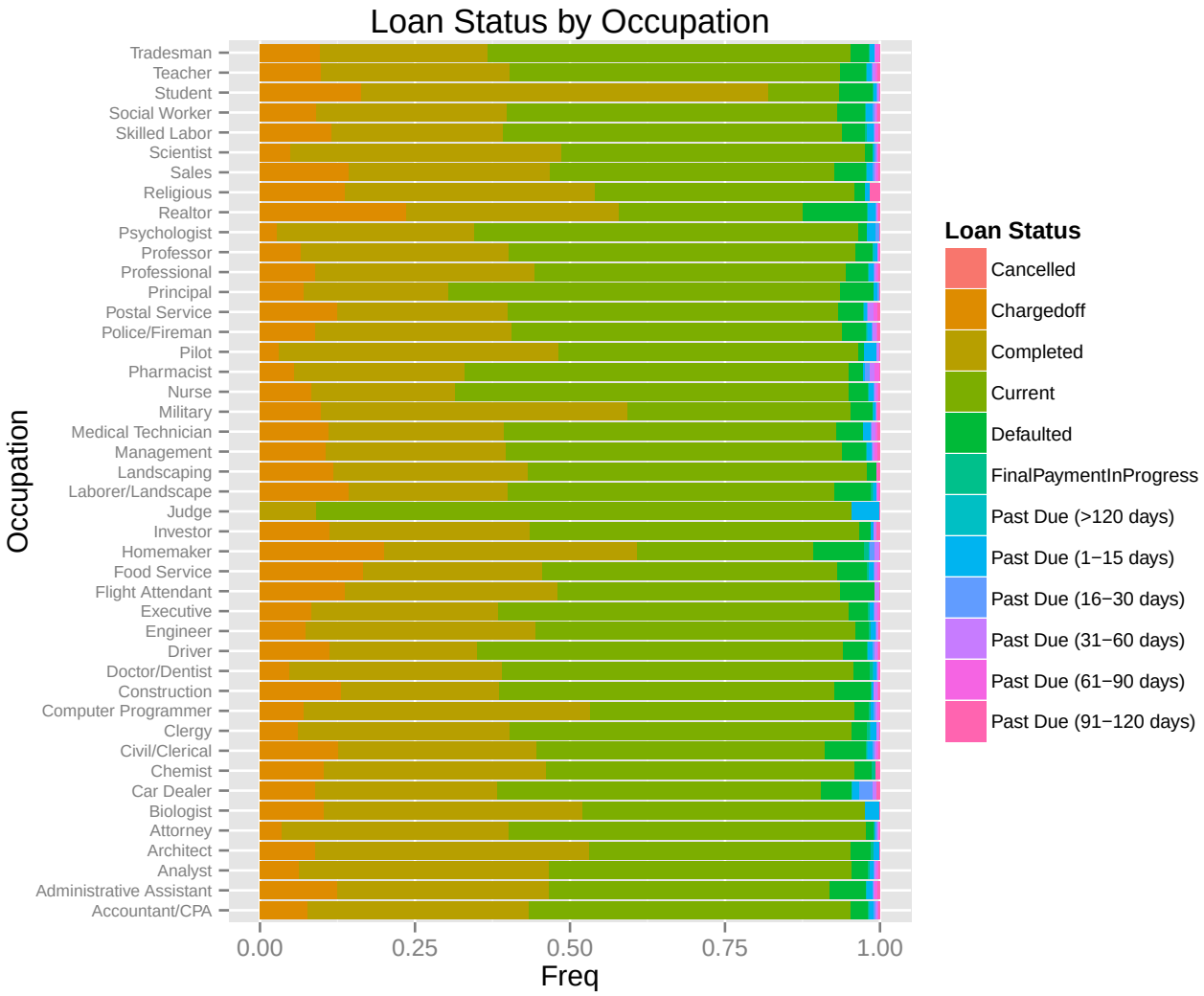

Figure 5: Loan status by occupation

Inter-variable Relationships

Relationships to Credit Score

We can pull out the numeric columns from the data and compare each column against others. Creditors often use credit score as one of the primary measures by which a borrower is rated, so let's explore relationships between various measure about a loan and the borrower's credit score. Here, we use the upper value of the borrower's credit score as reported by a credit agency, *CreditScoreRangeUpper*. As a creditor, we would be interested in exploring whether or not credit score is correlated with delinquencies. As observers of the loan industry, we can compare measures like *DelinqueniciesLast7Years*, *RevolvingCreditBalance*, *DebtToIncomeRatio* to see how a borrower's financial picture influences his or her credit score. As a borrower, we might be interested in seeing how our credit score changes the *BorrowerRate*. Furthermore, we can see whether or not there is a strong link between credit score and *ProsperScore*, the rating Prosper.com assigns to the borrower. In effect, we want to see what (if any) relationships exist in the data between credit score and other measures. The measures are scaled between 0 and 1 because we are comparing credit score with measures that have very different ranges - some are fractions representing a percent, others; such as monthly income, are orders of magnitude larger. We scale so we can plot on one axis.

In Fig. 6, which plots the relationships between credit score and these various measures, we can begin to understand some of these relationships. The discussion of these relationships will be short since the graph is fairly self-explanatory.

- **Credit score and delinquencies:** delinquencies drop at the upper and lower ranges of credit scores. Fewer delinquencies in the last seven years is correlated with higher credit scores.
- **Credit available and usage:** higher *AvailableBankcardCredit* appears to be strongly correlated with high credit scores, and *CurrentDelinquencies* inversely correlated with high credit scores.
- **Monthly income:** there do not appear to be any significant correlations between monthly income and credit score.
- **Borrower rate:** *BorrowerRate* looks be be capped for credit scores below approximately 550. Borrower rates are fairly evenly distributed across credit scores, which might indicate that there are other factors with significant influcene on *BorrowerRate* (such as *DebtToIncomeRatio*), or that *BorrowerRate* does not depend on credit score.
- **Prosper score:** scores assigned to a loan by Prosper seem to be evenly distributed, except very high credit scores generally receive a high *ProsperScore*.

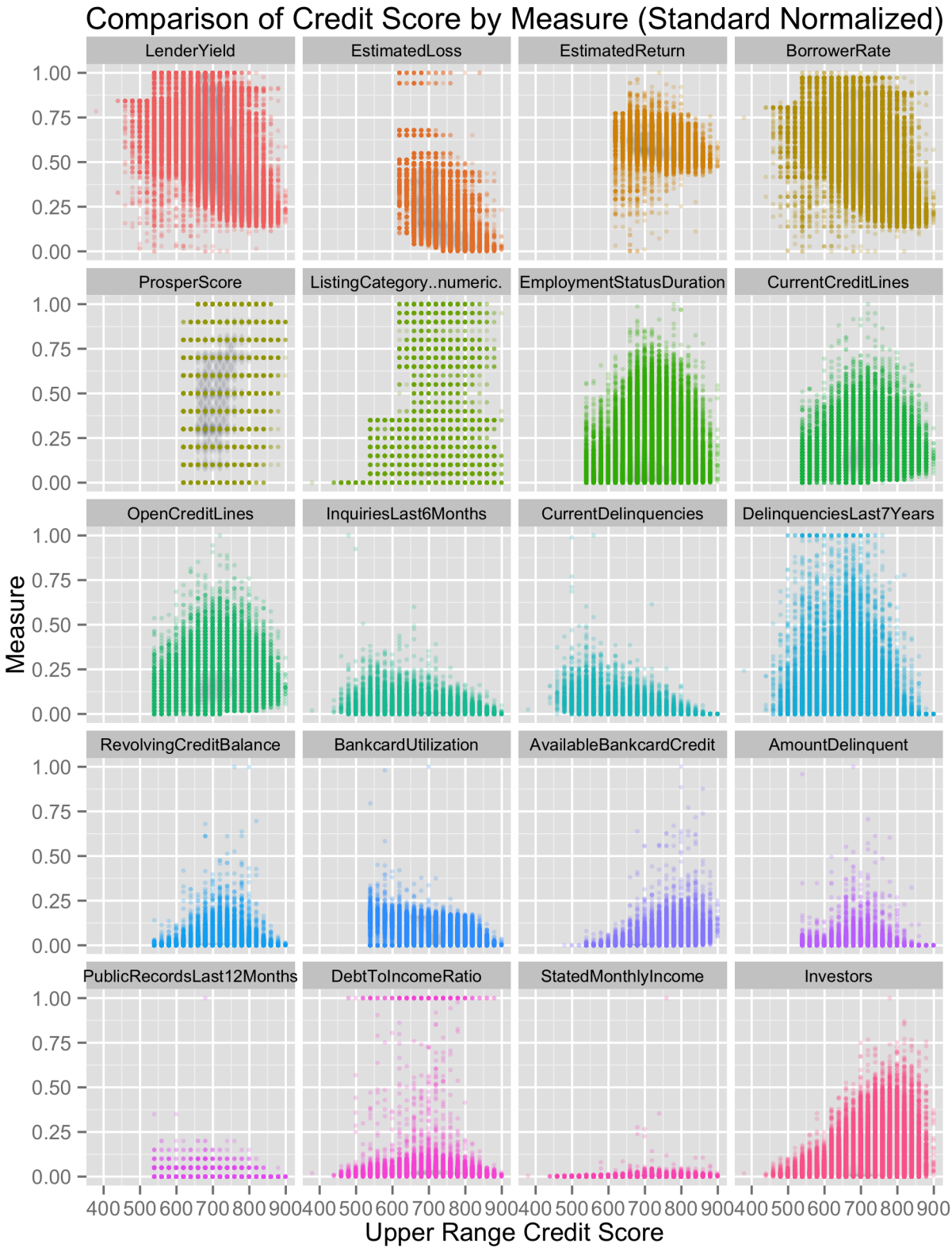

Figure 6: Scatter of various measurements by credit score with density in the background

This type of plot (a facet wrap) lets us easily compare multiple variables and multiple categories. We could increase the complexity, for example by making point size scale with the amount borrowed, to visualize even more data. However, the point of this graph is to identify correlations so we keep additional data off this plot to make it easier to identify patterns as credit score increases or decreases. Since the data is shown as a scatter plot, densities of the scatter have been placed behind the points so that we can recognize when trends in the scatter are due to variability in the data rather than a trend. We might also consider taking the second quantile (Q2) of data or removing outliers prior to plotting so they do not distract our eyes from possible trends.

Conclusion

There are a lot of plots and endless analysis that can be done on this data. With so many columns and a decent number of rows there are many visualizations that could be created with a specific hypothesis in mind. In this short report, I sought to explore general economic trends, get a basic idea of how credit score is impacted by a borrower's financial history, and see how credit score might influence his or her loans with Prosper.com.