**A Story About Student Loan Data**

**Author: Mason Krug Udacity Nanodegree Project**

**Story Version 1, Story Version 2, GitHub**

**Summary**

The Prosper loan data set contains 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. The dataset and data-dictionary can be found on below links. Discoveries were found by comparing categorical and continuous variables against states. For examples, which states have the highest return on investment and how that compares to the average return on investment. Additionally, homeownership is explored alongside credit rating and average return on investment.

* [Prosper Loan Dataset](https://s3.amazonaws.com/udacity-hosted-downloads/ud651/prosperLoanData.csv)
* [Loan Data Variable Definitions](https://docs.google.com/spreadsheets/d/1gDyi_L4UvIrLTEC6Wri5nbaMmkGmLQBk-Yx3z0XDEtI/edit#gid=0)

**Design Decisions**

I’m very interested in how data can vary by regions. People have a specific amount of variance based on genetics. However, differences greatly manifest themselves when geographic regions are taken into account. Culture, resources distribution, socioeconomic factors, any many other variables can be very interesting to look at when latitude and longitude are explored. Coastal states, such as California, have wildly different economies than Nebraska. This is why I chose maps as the primary way to explore relationships. Additionally, I chose scatter plots to explore relationships. I chose them because they can be used to explore many variables at once.

There is an obvious relationship between a credit rating and the gratuity a lending company would have when deciding loan rates. I wanted to explore how strong – or weak – this relationship is. I chose to utilize bar graphs to demonstrate this relationship mostly. They offer a quick visual way to see trends.

**Project Feedback**

My coworker viewed Tableau Story I had created and made a few recommendations. His first comment was to remove null values from the dataset. While some data was lost in this decision, I have to agree. Having a clean data set can allow me to make stronger conclusions about my findings. With a sample of this size, there is not much data loss overall.

Second, he made the resolution to focus on the contiguous United States. The reason behind this change was twofold: the maps in my project can be presented in a much cleaner fashion by aiding readability on smaller screens such as mobile devices. The second reason for this change was because Hawaii and Alaska are somewhat of outliers from the data set. They are vastly different from the contiguous United States in a number of ways. Their populations are small relative to their size and their geographic difference from the contiguous United States allows me to reduce anomalies in my data visualization.

**References**

2014 Population Statistics - census.gov/quickfacts/fact/table/US/PST040219#PST040219