Udacity Data Analyst Nanodegree – Project 8

Introduction

For the final project in the Udacity Data Analyst Nanodegree, the students are asked to create a Tableau Story from the data. I chose one of Udacity's curated data sets i.e. the Prosper Loan Data. The dictionary can be found here.

Link to the story : Prosper Data Story

(mistakenly over wrote the thing so I only have one version)

Summary

Prosper is a peer to peer lending company that offers personal loans at low rates. These loans are unsecured, which means you do not have to put up any collateral (like a house or car) that could get taken away if you can't make payments. Each loan is typically funded by multiple people all over the United States.

However, on November 24, 2008, the SEC found Prosper to be in violation of the Securities Act of 1933. As a result, the SEC imposed a cease and desist order on Prosper. In July 2009, Prosper reopened their website for lending ("investing") and borrowing after having obtained SEC registration for its loans ("notes").

The given data set consists of data from 2006-2014 and 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 where various aspects of the data have been explored to present a story to the viewers.

The story follows the downfall of November 2008 and its impact on Prosper in terms of borrowing and lending money. Moreover, the economic climate at the time with the peak of the Global Financial Crisis and its aftermath on loan taking is also explored. Prosper seems to have had a hard time coping up with the Crisis that it underwent, both internally and externally.

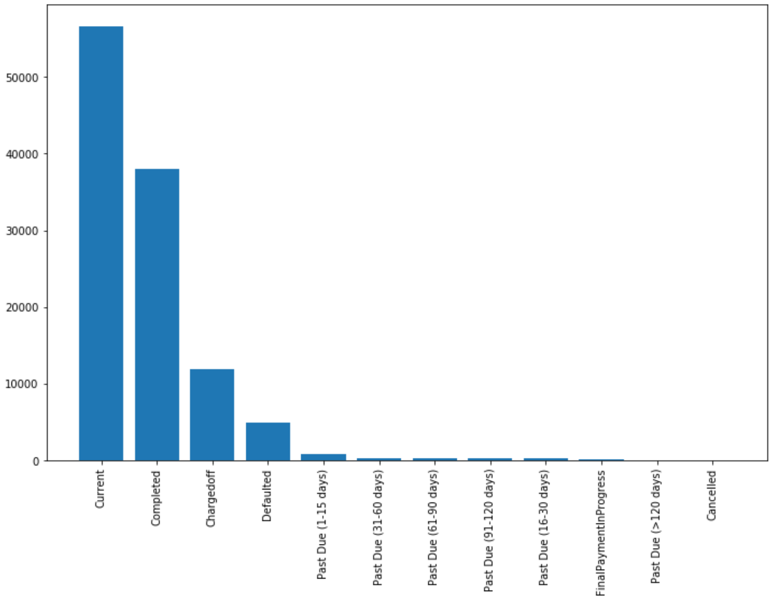
Data Exploration

The initial data was explored using Python and Jupyter notebooks. To get a sense of the data I plotted a simple visual to check how many what the loan status was and it seemed like most of the loans were still ongoing i.e. current. The second largest category was of completed loans. These were the two categories where I would have most of my analysis on.

Then I cleaned the data:

- I changed all date and time columns to only date columns for better time analysis.
- I dropped all but one primary column i.e. Member Key. All other columns with unique values only were dropped to decrease the size of the data set.
- I removed about 500 columns which had null values where there should have been no null values.

- For columns such as Occupation and Employment Status which had categories of ‘Other’ and ‘Not Available’, I filled all the Nulls with those values.
- Changed State abbreviations to State names from the data dictionary found online here.
- Then I split the data in two, pre-crisis data and post-crisis data. This splitting is then used in Tableau for analysis.



The data set was then downloaded as two separate datasets, and these were used for analysis and visualization. The background behind splitting the data set was threefold

1. Due to Prosper and its investigations in 2008-2009, there is an automatic split in the data. From 2006 to 2009 and then from August 2009 to 2014. There is also a split in the data where till July 2009, Prosper is using Credit Ratings to give loans and after July 2009, and implementing stricter credit guidelines for borrowers, it uses Prosper Rating as a measure.
2. From 2007-2009, the US economy had taken a downturn with the Great Financial Crisis. Recovery had started after 2009 so it seemed logical to split the data to show the Borrowers and Lenders landscape separately.
3. Smaller data sets meant fast analysis and I was going to draw a comparision.

Design

I wanted to have a clear design mechanism where the use of color was going to be purposeful:

- Each page in the story was going explain a single point only.
- Since I noticed the major divide in the data early on and I decide to explore that. I had to build an over arching point that would also zoom in when needed.
- Time is continuous, so a line graph would be the most appropriate for visualizing time related elements.
- To highlight the change before and after the crisis, there needs to be a comparison drawn. That means the same chart before the crisis and after the crisis.
- States are going to mapped by a country heat map.
- Categorical data can be plotted via color or different shapes.
- Text must be intentional to fill in the background. Any information that can be obtained by looking at the visual should not be put in words.

Page by Page Design

Page 1:

- It is important to show a drill down of the main overall time line side by side to the yearly time line. This design is important to show how there was a closure between 2008 and 2009, and you cannot see that from a yearly timeline.
- The interaction will highlight the need to drill down and look closely at the data.
- Loans are only Past Due from 2010. This is to be shown to highlight the economic climate of that period, although I don't need to show any other detail, so a simple dot chart will do the trick.

Page2:

- Showing the heatmap with the Debt to Income Ratio before an after the Crisis highlights the economic playground in the US.
- Use of different color in both maps is done to highlight the different scales. Debt to Income Ratio has risen after the crisis
- The scatter plot is the best at showing comparison between 2 or more dimensions.
- Instead of using color in the scatter plot, I used 2 different shapes, for 4 comparisons. This ensures that a reader sees 4 points but can draw comparisons. The filled and unfilled shapes are also used to make comparisons because it points to similar measures but at different points in time i.e. before and after the Crisis.
- The interaction can draw multiple measures together to highlight how each state measured before and after the Crisis. The only way to draw attention to the interaction was to add a tool tip.

Page3:

- The color was to highlight the different Prosper Ratings. Only with the color encoding, do we see that there is a connection with the Estimated Loss and the Prosper Rating.
- The income range was also encoded with the different sizes. The larger the shape, the larger the income range which highlights the category of the range without extra text.
- A scatter plot is the perfect plot as 2 measures are plotted together.

Page4:

- Since I am highlighting factors over the same timeline, line charts are the way to go.
- Color encoding is done because Prosper Rating is categorical.

Feedback

I added my initial design on the Slack workspace for all my peers to review. Following are the reviews I got:

Leah – "I love the 'hover to find out' interaction. Hovering to 'find out' though gives me the facts but doesn't explain 'why'? The colors of the last tab are attention getting. I suggest putting a title clear title above each graph to explain what we are looking at (the y-axis says, but a nice little title on top would be easier to read). Good Job!"

Ahmed – “You need to add some sort of description on Page 2. There’s no indication that you have to click on the world map to find the data associated with it. Bring the chart on the right to the bottom so it occupies more space and can be seen by the eyes. Good Job otherwise.”

Design Changes

- Update point 2 of the story and bring the right most chart to the bottom. Also add a tooltip to help readers in finding the interaction
- Update point 1 to add a tooltip and a pointer to work with hover interaction.
- Updated point 3 of the story to remove shape encoding because there were multiple encodings for the same range.
- Added titles to all the graphs to ensure readability.