A Story of Loan Data from Prosper

Summary

Background

Prosper.com was founded in 2006 as one of the online peer-to-peer (P2P) lenders, and it’s a large provider of personal loans.

Business Model

Prosper acts as the loan underwriter but doesn't lend the money itself. Alternatively, the company matches potential borrowers with investors who want to fund loans. Prosper makes money by charging fees for this service.

Dataset

This 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. Download link of dataset and data dictionary as below:

Download dataset

https://www.google.com/url?q=https://s3.amazonaws.com/udacity-hosted-downloads/ud651/prosperLoanData.csv&sa=D&ust=1551867497134000

Data dictionary (explains the variables in the data set)

https://docs.google.com/spreadsheets/d/1gDyi_L4UvIrLTEC6Wri5nbaMmkGmLQBk-Yx3z0XDEtI/edit#gid=0

Finding

- By using Tableau, we found that Propser Rating is a good indicator and according to "Risk-Return Tradeoff " theory, higher risk is associated with greater probability of higher return. However, loans with rating HR have the highest risk but they don't have the highest estimated return and loans with "prosper rate C & D " have the greatest principal loss than "prosper rate E & HR ".

- To examine the defaulters, we found that the higher the income, the lower default rate.
  Compare with other occupations, student group has higher default rate and borrow rate. Among student groups, College student has the highest no. of loans, while College Sophomore student has the highest default rate. In addition, “Prosper rating D” has the greatest default rate.
- Since 2009, loan amount and number of records of loan steadily increase and reach to maximum during 2014. Alternatively, prosper score is gradually decrease from 2009 to 2014. Default rate is the highest in CA.
- Prosper Score based on different category in which peronal loan, business and home improvement are contribute most of loan.

Design

Target audience

- Potential borrowers and investors.
- Individuals interested in Peer-to-Peer lending.

Visualization

- To find out any relationship between attributes and borrowers and investors for making better decisions.

Chart Types

- Map
  - Maps are probably the biggest subsection of visualization types. In this project, I used Choropleth map to visualize geographical (per state), color is the important part to this map. A color scale is assigned to numerical data (default rate), and the value for each region is used to color the region. It usually uses political boundaries as the regions. This method is very easy to understand.
- **Bar-chart**
  - In this project, I used both horizontal or vertical bar charts to show numerical comparisons across categories (e.g: income range, occupation, prosper rating). Bar charts are good for (1) summarizes a large data set in visual form, (2) displays relative numbers or proportions of multiple categories, (3) shows each data category in a frequency distribution, (4) allows a visual check of the accuracy and reasonable of calculations. However, it requires additional expiation.
- **Stacked bar**
  - In this project, I also employed simple stacked bar graphs, which placed each value for the segment after the previous one. The total value of the bar is all the segment values added together. In this case, ideal for comparing the total amounts of default rates across each prosper rating. For example, to show how a larger category (e.g: prosper rating) is divided into smaller categories (e.g: income range) and what the relationship of each part has on the total amount (e.g: default rate).
- **Line plot**
  - In this project, I also applied line graph plots, which used continuous data as points and then joins them with a line. Multiple data sets (e.g: different borrower state) can be graphed together, such that we can compare them easily. However, it can only use with continuous data.
- **Scatter plot**
  - I used scatter plot to show a trend in the data relationship between (1) loan original amount and month of loan date. (2) Default rate and median borrower rate).The advantage of plot will retain exact data values and sample size, it can also show minimum/maximum and outliers. But despite that is hard to visualize results in large dataset and data on both axes should be continuous.

- **Area chart**
    - Area charts are commonly used to showcase data that depicts a time-series relationship. After reviewer feedback, I changed line plot into area plot. Area charts are primarily used when the magnitude of the trend is to be communicated. The filling between the line segments and the axis line facilitates understanding the magnitude, something which cannot be done using the line chart.

Encoding

- **Color marks**
    - I used color to differentiate different categories (e.g: listing category, income range).
- **Dimensions**
    - In scatter plots, I also used dimensions to differentiate (1) average default rate, (2) count of listing category.
- **Filters**
    - I used interactive filters and shelf filters.
- **Calculated fields**
    - I used calculated fields as they can be used to create more robust visualizations.

Feedback

After I completed my first sketch, I shared with my peer classmates. Their feedbacks as follow:

No.	Feedback	Action	Lesson learnt
1	In story TWO & THREE too much colors in these two bar charts.	I realized the colors add no meaning or value and usually distract people. I change the bar chart into single color.	Use different colors only when they correspond to differences of meaning in the data.
2	In story FOUR, feel difficulty to identify which states are contribute the most default rate.	I changed the map color to purple palette with 5 steps.	The success of choropleth maps, depends upon understanding how we manipulate colors to systematically encode our geographic data using the three dimensions of color: hue, lightness, and saturation. Dimensions of color have powerful effects as visual variables, causing the eye to see groups, order, or some combination thereof. Choosing colors carefully based on the nature of the data we're mapping will ensure that the map's message is made clear.
3	In story FOUR, there are too many lines in a line chart.	I changed the line chart into area plot and showed Top 4 Borrower State where they contribute the most loan original amount and number of records.	Area charts are commonly used to showcase data that depicts a time-series relationship. This area chart gives us a quick comparison of the trend in the States over the period of a year.

Initial version (Before feedback)
https://public.tableau.com/views/ProsperLoanDatav01/ProsperLoanStoryv1?:embed=y&:display_count=yes&publish=yes

Final version (After feedback)

https://public.tableau.com/views/ProsperLoanDatav02/ProsperLoanStoryv2?:embed=y&:display_count=yes&publish=yes

Resources

- Prosper: https://www.prosper.com
- Tableau Tutorial https://www.tutorialspoint.com/tableau/
- Data Visualization 101 https://cdn2.hubspot.net/hub/53/file-863940581-pdf/Data_Visualization_101_How_to_Design_Charts_and_Graphs.pdf