Lending Club Data Analysis

Xin XU

Problem:

Lending club is a peer-to-peer company. It offers loan trading on a secondary market. Lending club is the world's largest peer-to-peer lending platform. It enables borrowers to create unsecured personal loans between $1000 and $40000. The standard load period is three years. Investors can search and browse the loan listings on Lending Club website and select loans that they want to invest in based on the information supplied about the borrower, amount of loan, loan grade, and loan purpose. Investors make money from interest. Lending Club makes money by charging borrowers an origination fee and investors a service fee.
(https://en.wikipedia.org/wiki/Lending_Club)

Reason:

Peer-to-peer(P2P) company is quite popular in the world recent years. In general, the interest is much higher than the bank. More and more people try to invest some money on P2P. However, it's always more risky to invest money by P2P website or company than in the bank. I'm very interested in the true situation in the P2P company. And I want to do some data related work in finance industry after I graduate. So I want to do some data and finance related project.

Data:

I will use the loan data from the Lending Club. The data will include the current loan status, latest payment information, borrower credit scores, number of finance inquiries, address, etc. The data file is from Lending Club official website:
https://www.lendingclub.com/info/download-data.action
This data file can also be found in Kaggle

Problem definition:

Potential Questions:
What's the trend of the Lending Club amount of the clients? i.e. More and more people use this platform? Or less and less people use it?
What kind of people use this platform most? For example, which profession use Lending Club most.
What kind of people have bad record in the end? If possible, can we make a decision whether to lend money to someone?
Techniques:
For many questions like first two, they shouldn't be very difficult to answer if we can clean the data. After we got our cleaned data, we can easily summarize the data, and give some good data visualization. By the statistics and visualization, we can answer these kind of questions easily.

For the last questions, we may need to use logistics regression or decision tree to predict whether a client would be potential 'bad' customer and then refused his/her request.

Preliminary Data Manipulation:
See Jupyter notebook(Including data assessing, data cleaning, statistics summary, and data visualization)