Python for Data Analysis and Scientific Computing - Project Proposal

**Project Members-**

Fiona Tang: [fionatang@berkeley.edu](mailto:fionatang@berkeley.edu)

Zhicheng (Jason) Xue: [emailxjason@gmail.com](mailto:emailxjason@gmail.com)

**Data source-**

<https://www.kaggle.com/wendykan/lending-club-loan-data/data>

**Dataset Description-**

LendingClub is the world’s largest peer-to-peer lending company, and it was the first peer-to-peer lender to register its offerings as securities with the Securities and Exchange Commission. The dataset includes information on all loans issued from 2007 to 2015. The data consists of information on the borrowers (“ability-to-pay” metrics such as credit score, length of employment, finance inquiries) as well as the loans themselves (loan status, loan amount, loan term, interest rate). The dataset includes ~890,000 observations and 75 variables.

**Why we chose this project-**

We chose this project because the correlations between borrowing habits and loan characteristics is an interesting topic. For instance, do people default on their loans because they take out too much, because they come from lower-income backgrounds, or both? Do we see different borrowing habits in different geographical regions across the nation? These are some of the questions we wish to explore. The project also has powerful predictive implications; although studying past trends/patterns can offer insight into how borrowers will manage their loans in the future.

**Goals (what we want to solve) & Project Plan (how we intend to solve it)-**

Our goal is to perform descriptive and diagnostic data analysis, to get a better understanding of the relationship between borrower traits and loan characteristics. We will analyze loans with “bad” loan status, and whether it seems to correlate with factors such as income level and loan amount. We will use scipy.stats to perform hypothesis testing, perhaps a 2-sample t-test, to compare two groups, borrowers with bad loan status versus borrowers with good loan status, and test whether differences between these groups is statistically significant.

**Final deliverable-**

The outcome will be data exploration, visualization, and statistical analysis of loan data using numpy, matplotlib, and scipy.