# Datathon: LendingClub

*Presented by Correlation One*

## Problem Statement

Welcome to the 2022 Datathon! This document explains the topic of the Datathon, important details about the datasets you'll be using, and guidance on how to submit your results.

### Background

LendingClub was originally founded in 2006 as a peer-to-peer lending Facebook application. After receiving over $10 million in Series A funding in late 2007, LendingClub blossomed into a full-scale peer-to-peer lending platform that allows borrowers to create unsecured personal loans for a variety of reasons including home refinancing and debt consolidation.

LendingClub screens potential borrowers and *services* the loans once their application is approved. In essence, LendingClub acts as an origination vehicle for the loan, but they **do not fund the loan**. Ultimately, **investors make the decision on whether or not to fund the loan**, based on the loan specifications and debtor's credit history.

This service offers debtors a number of unique advantages when searching for suitable credit. First, debtors can receive their rates on loans without initiating a hard credit inquiry, thus preserving their credit score. Additionally, each loan has a fixed rate, no hidden fees, and no obligation prior to the transmission. LendingClub also services debt investors by providing access to un-securitized personal debt, while affording investors the ability to shop for loans that fit their risk profiles.

While generally credited with pioneering the business model of peer-to-peer lending, LendingClub also generated some controversy about the discriminatory practices that some investors might have engaged on the platform. The argument has been made both in favor of LendingClub's business model for increasing access to credit and against it for betraying investors' trust. The period between 2014 and 2018, over which the provided data was collected, also encompasses many important changes for LendingClub, including its IPO in 2014 and the resignation of its founder and CEO in 2016.

### Your Task

You are asked to pose your own question and answer it using the available datasets, as well as any supplementary datasets that you find to aid your analysis. Both the creativity of your

question, and the quality of your analysis are of paramount importance. **You need not be comprehensive; depth of insight is more important than breadth of question poses.**

Submissions may be predictive, using machine learning to classify or predict patterns. Submissions may also be illuminating by way of data visualization or sound statistical testing.

While the problem of default prediction might come to mind most naturally, we suggest selecting topics that have received less attention, unless you feel confident that your approach to default prediction will help you stand out. Consider exploring one of the sample questions below, or using them as an inspiration for your exploratory work. Creativity in formulating your question is encouraged; **however, it should not be at the expense of analytical depth, precision, and rigor, which are far more important.**

Sample Question 1: Assuming that the demographic information can be inferred from a borrower's zip code or home state, can we identify a bias (whether positive or negative) in loan rejections towards a particular group of borrowers?

Sample Question 2: Can the proportion of borrowers, entering the hardship plan, be better explained by the inter-temporal macroeconomic changes, their demographic characteristics, or purely their ex-ante creditworthiness?

Sample Question 3: Was there a noticeable change to LendingClub's overall business model or discriminatory practices after the resignation of its founder and CEO in 2016, following some public controversy?

## Datasets

The provided datasets are stored in the "Datathon Materials" folder on Google Drive. Your team should only use the data / datasets that are relevant to your choses question / topic.

The datasets consist of two tables: one table with all accepted loan applications from 2014 to 2018, and one table with all rejected loan applications from 2014 to 2018.

The accepted loan applications table includes a wide array of demographic and financial data pursuant to both the debtor, and any secondary applicants. Information is also included on the funding for the loan, and the loan terms.

The rejected loan applications table includes a handful of data elements detailing the applicant's demographic profile with some additional indicators on financial risk and credit worthiness.

**Lending_Club_Accepted_2014_2018.csv**
This dataset represents all accepted loan applications from 2014 to 2018.
*2,029,952 rows & 141 columns. Size: 1.5GB*

**Lending_Club_Rejected_2014_2018.csv**
This dataset represents all rejected loan applications from 2014 to 2018.
*26,132,308 rows & 9 columns. Size: 1.68GB*

## Additional Datasets

Participants are welcome to scour the internet for their own custom datasets to supplement their analysis. All additional data used should be public and reputable. Additionally, any supplementary datasets should not exceed 1 GB unzipped (consult Correlation One's R&D team if you believe your idea is worthy of an exception).

**For example, we would encourage you to explore the United States Census data to enrich the data sets provided by the demographic characteristics of the borrowers at the zip-code level.**

## Other Materials

We will provide you with the schema for each of the data tables in another packet.

## Submissions: Content

Submissions should have two components:

1. Report – this should have two main sections:
    a. Non-Technical Executive Summary – What is the question that your team set out to answer? What were your key findings, and what are their significance? You must communicate your insights clearly – summary statistics and visualizations are encouraged to help explain your thought process
    b. Technical Exposition – What was your methodology / approach towards answering the questions? Describe your data manipulation and exploration process, as well as your analytical and/or modeling steps. Again, the use of visualizations is highly encouraged when appropriate.
2. Code – please include all relevant code that was used to generate your results. **Although your code will not be graded, you MUST include it, otherwise your entire submission will be discarded.**

Additional information (e.g. roadblocks encountered, caveats, future research areas, and unsuccessful analysis pathways) may be placed in an appendix.

Judges will be evaluating your technical report without your team there to explain it; therefore, **your submission must "speak for itself".** Please ensure that your main findings are clear and that any visualizations are functionally labeled.

## Submissions: Evaluations

The competition will have multiple rounds of evaluation. Your Report will judged as follows:

- **Technical Executive Summary**

    - *Insightfulness of Conclusions.* What is the question that your team set out to answer, and how did you choose that question? Are your conclusions precise and nuanced, as opposed to over-generalizations?
- **Technical Expositions**
    - *Wrangling & Cleaning Process.* Did you conduct proper quality control and handle common error types? How did you transform the datasets to better use them together? What sorts of feature engineering did you perform? Please describe your process in detail within your Report.
    - *Investigative Depth.* How did you conduct your exploratory data analysis (EDA) process? What other hypothesis tests and ad-hoc studies did you perform, and how did you interpret the results of those tests and analyses? What patterns did you notice, and how did you use these to make subsequent decisions?
    - *Analytics & Modeling Rigor.* What assumptions and choices did you make, and how did you justify them? How did you perform feature selection? If you build models, how did you analyze their performance, and what shortcomings do they exhibit? If you constructed visualizations and/or conducted statistical tests, what was the motivation behind the particular models you build, and what did you tell you?

## Submissions: Formats

Reports can be produced using any tool you prefer (Python Notebook, Shiny Application, Microsoft Office, etc.); however, **your report MUST be in a universally accessible and readable format (HTML, PDF, PPT, Web link)**. It must not require dedicated software to open. For example, if your report is a Python Notebook, it should be exported to HTML. If you create a Shiny App, it should be published at an accessible Web link.

**Please include the source file used to generate your report.** For example, if you submit a PDF with math-type, equations, or symbols please include your LaTeX source file.

Code should be submitted in a single zipped collection of files separate from your report.

Your team will be sent a Google Form at the beginning of the competition; you will use this form to submit and upload your content. **Submissions MUST be received by Sunday, July 24th 5pm ET. Any submissions received after that time will NOT be evaluated by the judges.**

## Tips and Recommendations

Since this is a virtual event and you will have almost a week to work on your submissions, you should start thinking early about what problem you want to solve. The outcome of this Datathon, and your overall success, will largely be a product of how well you planned prior to the event, and the insightfulness of the problem that you chose to solve.

For data engineering, exploration, and modeling, we highly recommend that you install Jupyter Notebook: http://jupyter.org/install.html. Jupyter Notebook is an interactive, real-time

development environment that eliminates many pain points of the standard "terminal & text editor" environment, and is compatible with both Python and R.

We also recommend that your team stick to tools and techniques that you have previously used. Learning new skills is certainly valuable, but it can consume a large portion of your available time, leaving less time for completing the task at hand.

We've compiled 3 additional commonalities of successful teams and 3 pitfalls that successful teams will actively avoid. Of course, these may not apply to every team, so we recommend that you and your team apply any tips accordingly.

| Tips for Success | Try to Avoid |
|---|---|
| **1.** Focus on hypothesis testing when brainstorming your research question | **1.** Do not try to exhaust all different models you know just to yield an ideal cross validation accuracy |
| **2.** Spend at least 3 hours on your report to ensure strong communications through both visualizations and writing | **2.** Do not violate assumptions of statistical models. Sometimes, specific models require specific features so it is best to make sure those conditions are sufficiently met |
| **3.** Engage in proper causal analysis. Just because your model passes standard cross-validation checks it does not demonstrate (or even suggest) causality | **3.** Do not pick research statements and blindly stick to it trying to get it to work. Often times, further data exploration will show that it's not true or worthwhile |

## Ask for Help

Correlation One's R&D team is here to help. Let us know about your struggles as early on as you can and we may be able to offer advice on how to best move forward.