The goal of the task is to reproduce one table from an existing paper, plus bring in a new data source to generate an original insight. The following files are attached to these instructions:

1. Paper – Nagaraj_reimers_april21.pdf
2. Data – loans_merged.dta

We need you to do a few things:

1. Process the loans data and make it into a balanced sample between 2003 and 2011, such that there are a total of 88006 observations (one per book) in each calendar year. Each row is one loan event in the raw data. Impute 0 loans when a book has not been loaned in a given calendar year. Clean this data to keep only relevant columns and call it "master.dta"
2. Using this master.dta data, replicate Table 5 columns 1 and 2 ("main effect") in the main paper and spit out table_5.tex. We do not care about the formatting of your output to match – just the numbers. The variable "year_scanned" indicates the year when a book was scanned, and this variable is missing if it was not scanned. The "location" var has the physical location of the book.
3. Next, we'd like your help to understand whether the effect of digitization was different for different kinds of books. Using the book identifiers (e.g. book titles in the "title_display" var or OCLC numbers in the "_oclc" var) and any website/API of your choosing (e.g. ChatGPT, Google Books API, OpenLibrary), please retrieve any additional book-level information that you think might provide an interesting source of heterogeneity. Don't worry if you can't get info for all the books, just aim to have as good coverage as possible. Merge this new data into the master.dta and generate one additional result. Represent your key result using a figure neatly labeled for presentation and provide a short discussion interpreting your results.
4. Create a "lastname_firstname_output.tex" file that has both the table and figure above and a short description of your approach and results.

Please avoid asking questions unless extremely necessary (we cannot guarantee a quick response!) – part of the goal of this exercise is to judge your ability to work independently. In case of extremely necessary questions, please direct them to Cecil-Francis (cecil-francis.b@berkeley.edu). When you are done with this data task, please send us (data-innovation-lab@berkeley.edu) a zip file with your name (Last_First.zip) including all the following files:

1. The raw data that we sent you.
2. The new data that you collected. Name it "original_data"
3. Your Stata.do file (please comment on the code to make it easy to follow). Stata is preferred but we also accept R or Python. In either format, we should be able to run your codes on our computer and it should generate the final dataset and the files with the final tables/figures. Everything should be automatic (i.e. no need to copy-paste a number to excel or anything like that).
4. A short (say, 1-2 pages) written Latex document describing what you did, including how you went about retrieving your original data, and with the key figure and table (remember to attach all the tables and figures in this document). As before, we should be able to compile this Latex document and it should generate the PDF document on our computer using the figure table that were generated with the do-file. If you're using an online Latex editor like Overleaf, please include a link to your project in the email.
5. The final PDF output for the latex file above. Name it (Last_First.pdf).

We expect this task to take approximately 5-7 hours to complete. The goal is to assess the current status of your coding and computational skills related to empirical economics/management research, but a "perfect" score on this task is not a prerequisite to be considered for the job. With that in mind, please do this work yourself, without input from other people.