Problem Statement

Organization Name: U.S. Small Business Administration (SBA)

Dataset Name: PPP Removed Applications

Difficulty: ⚙ ⚙ ⚙

Level 3: Participants with some data analysis background.

The problem statement is open-ended about what the final product may look like. The dataset may contain many variables of interest. Analyses from different angles by various techniques are encouraged.

Background

Enacted by Congress in 2020 to respond to the economic impact of the COVID-19 pandemic, the Paycheck Protection Program provided nearly $800 billion in loans to small businesses in order to retain payrolls. The Small Business Administration has oversight over the PPP program, although the loans are administered by private lenders, who then submit application information to the government. The loans could be fully forgiven if certain conditions were met.

The SBA has periodically released data on the more than 11.5 million approved applications, but it also has removed applications that had been previously present in the dataset.

The SBA has not offered publicly a reason for the removal of these applications, although some theories seem reasonable to consider:

1. Applicants may have withdrawn their applications.
2. Lenders may have sought additional information to confirm application details and not received it.
3. Lenders may have sought additional information to confirm application details and decided to cancel the loan application.
4. Lenders may have determined that applications were fraudulent.

Your task: through data analysis, develop an understanding of why these loans might have been removed.

For this challenge, we are providing two datasets, both describing loans to businesses in Georgia.

- Loans to Georgia businesses that were removed from the PPP database (ppp-removed-ga)
- All loans to Georgia businesses that remain in the PPP database (ppp_applicants_ga_full)

Questions

1. What are some defining characteristics of the removed loans?

   This could include demographic information about the borrower, the industries of borrowers, the size of borrower businesses, the lenders, the amount of undisbursed funds or something else.

2. How do the characteristics of loans that were removed from the PPP data compare to loans that were not removed from the PPP data?

3. Is it possible to accurately predict whether or not a loan was removed from the data, using fields in the data set and/or additional information you incorporate?

Data Considerations

- A data dictionary is included in the repo.
- The raw data includes a field with the race of the borrower, but most loans do not include it. We have excluded it from the data we are providing to you. If you wish to ask questions related to race or other demographic information, you might incorporate U.S. Census data at the level of County, ZIP Code or (if you geocode the addresses) Census Tract.

Additional Datasets

- Datasets from other sources, such as the Census, American Community Survey, or other sources may also be considered.