TEAM# IC22030

Team Members: Ling Fang, Yu-Tung Chang, Upasana Mohapatra

To predict the removal of loans from the public dataset, we're working on building a **classification model** to predict whether loans will be removed, across various industries and regions using three datasets - **removed loans**, the original **public dataset** including all loans, and a third dataset **combining the two datasets** provided.

Comparing the average values for loan amount in the full dataset and the combined dataset **(~$44K)** with the average loan amount from the removed dataset**($20K)**, we see a huge difference in loan borrowing. We then estimated the **size of businesses** in the datasets by looking into the jobs retained per service industry(source: NAICS data). Most of the jobs retained in the **removed loan** dataset were from miscellaneous service industries(**mostly B2C**) such as beauty salons and barbershops that are **smaller**. For the **full dataset**, most of the jobs retained were in **bigger service** industries(**B2B, B2C**) like Accommodation and Food Services, Retail Trade, etc.

Furthermore, we are creating **flags** for **lenders** with the maximum number of approved and removed loans, and **cities** that have the maximum number of removed loans in Georgia to build a smarter predictive model for our investigation. Some of the more prominent cities from this analysis are – '**Atlanta', 'Decatur', 'Fairburn'** while some of the lenders approving the maximum number of removed loans are **- 'Capital Plus Financial LLC', 'Prestamos CDFI LLC'**. We are now looking into the **average incomes and covid cases** in these cities for the given time period and feed that information in our model.