Paycheck Protection Program is an SBA-backed loan that helps businesses keep their workforce employed during the COVID-19 crisis. The loans are administered by private lenders, who then submit application information to the SBA.

The main objective of this project is to develop an understanding of why SBA removed certain loan applications. The SBA has not publicly offered a reason for the removal of these applications. We were provided with two datasets with over 500,000 records, including both accepted and removed records.

We first cleaned and merged the two datasets to describe them by generating summary statistics comparisons and visualizations to make hypotheses. Then, we performed statistical inference to verify the hypotheses based on gut feelings.

We chose a binary logistic regression model for statistical modeling to interpret the determinants of whether a loan was removed from the database. In the final step, we further utilized machine learning to perform a binary classification, using the LightGBM model we could accurately predict whether a loan will be removed from the dataset. We used Google Collab notebook for Python programming and Tableau for visualization throughout the process.

We found that overall, as of June 30, 2021, about 4% of loan records were removed from the database, with the majority of deleted records being updated between April and June. And the key determinants of whether a loan record will be removed from the SBA database or not are the loan status (esp. The record is in Exemption 4 status) and the length of the status generation cycle.