To-do list - not including final write/clean ups for final deliverable

\*Update/change this file as needed

* ~~Get a smaller sample of the data - Victor~~
* Figure out necessary clean/transform steps and add them to Colab file - Victor
* Create test set (Santi)
* Create pre-processing pipeline (Santi)
* Add feature selection (if we are not sort on time maybe test 2 methods?) (Santi)
* Select 3-4 predictive model (in the guidelines it says we should try all models seen in class unless they don’t apply for some reason)

Models:

* Logistic Regression and evaluate on train set (CV)
  + Fine tune model
* Decision Tree and evaluate on train set (CV)
  + Fine tune model
* Random Forest and evaluate on train set (CV)
  + Fine tune model
* \* SVC evaluate on train set (CV) (leave for last)
  + Fine tune model
* KNN evaluate on train set(CV)
  + Fine tune model
* Compare model performance and pick the best one
* Test final model predictions on test set

Notes:

-loan\_status: need to discuss how to treat the different values

-we get 87 features before feature selection, how do we want to clean/change those?

-how to deal with zip codes (800+ zips)?

-how to deal with earliest credit line (600+ date values)?