Problem Identification

1. Problem statement formation
    a. As a representative of the bank, should I grant a loan to a particular small business (Company X)? Why or why not? By assessing a loan's risk, specifically in California.
2. Context
    a. The dataset is from the U.S. Small Business Administration (SBA). The U.S. SBA was founded in 1953 on the principle of promoting and assisting small enterprises in the U.S. credit market (SBA Overview and History, US Small Business Administration (2015)). Small businesses have been a primary source of job creation in the United States; therefore, fostering small business formation and growth has social benefits by creating job opportunities and reducing unemployment. There have been many success stories of start-ups receiving SBA loan guarantees such as FedEx and Apple Computer. However, there have also been stories of small businesses and/or start-ups that have defaulted on their SBA-guaranteed loans.
3. Criteria for success
    a. Identifying Risk Factors
4. Scope of solution space
    a. The assessment is accomplished by estimating the loan's default probability through analyzing this historical dataset and then classifying the loan into one of two categories: (a) higher risk—likely to default on the loan or (b) lower risk—likely to pay off the loan in full. The process of making this determination requires students to conceptually understand the statistical concepts and how to apply them.
5. Constraints
    a. Location by State
    b. Industry
    c. Gross disbursement
    d. New versus established businesses
    e. Economy may impact default rates
    f. Whether a loan is backed by real estate
    g. The portion which is the percentage of the loan that is guaranteed by SBA
6. Stakeholders
    a. The bank
7. Data sources
    a. "National SBA" dataset (named SBAnational.csv) from the U.S. SBA which includes historical data from 1987 through 2014 (899,164 observa- tions)
    b. "SBA Case" dataset (named SBAcase.csv) which is used in the assignment described in this paper (2102 observations). The "SBA Case" dataset is a subset of the "National SBA."
    c. https://www.tandfonline.com/doi/full/10.1080/10691898.2018.1434342?scroll=top&needAccess=true
    d. https://www.kaggle.com/mirbektoktogaraev/should-this-loan-be-approved-or-denied

Here are some questions to consider to help you get started:

1. What is the problem you want to solve?
    a. California-Based Case Study: I'm a loan officer for Bank of America, and have received two loan applications from two small businesses: Carmichael Realty (a commercial real estate agency) and SV Consulting (a real estate consulting firm). As a loan officer, that needs to determine if they should grant or deny these two loan applications and provide an explanation as to "why or why not." To make this decision, I will need to assess the loan's risk by calculating the estimated probability of default using logistic regression. I will then want to classify these loans as either: "higher risk—more likely to default" or "lower risk— more likely to pay in full" when making a decision.
2. Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis?
    a. Inform a bank whether they should grant such a loan because of the high risk of default.
3. What data are you using? How will you acquire the data?
    a. Data acquired from the "National SBA" dataset (named SBAnational.csv) from the U.S. SBA which includes historical data from 1987 through 2014 (899,164 observations)
4. Briefly outline how you'll solve this problem. Your approach may change later, but this is a good first step to get you thinking about a method and solution.
    a. Clean Data
        i. Overview the missing Data
            1. Define what is worth keep or taking out
        ii. Work through the stages in model building and validation
        iii. Build a logistic regression model to estimate the default probability of the various loan applications.
            1. Apply logistic regression to classify a loan based on predicted risk of default
        iv. Calculate the misclassification rate using different levels of cutoff probability.
        v. Use the final logistic regression model to generate the estimated probability of default rate for each of the loans in the "test" data sample.
        vi. Classify the loans in the testing data as either "higher risk" or "lower risk" using the decision rules
        vii. Validate the final model by applying it to the other half of the data
        viii. Final logistic regression model generated to determine the estimated probability of default of a specific loan
        ix. Make a scenario-based decision informed by data analyses (i.e., whether to fund the loan).