You've collected your data, cleaned it up, wrangled it into shape and explored it. Now it's time to perform some in-depth data analysis using machine learning. This step depends on you and your mentor, but here are some suggestions to get you going.

1. How do you frame your main question as a machine learning problem? Is it a supervised or unsupervised problem? If it is supervised, is it a regression or a classification?

Original Research question:

What is the literacy rate for the households? Is literacy related to other variables? If so, how? The households include African Americans living in Philadelphia during 1847. They include male and female family members of various ages. Many of the residents were not born in the state of Pennsylvania. The residents perform a variety of occupations and also demonstrate difference in their ability to read and/or write (literacy).

Research Question with Machine Learning in Mind:

Is a family literate given family size, male occupation, female occupation, and whether the family members were born slaves? The method is supervised learning because of the format of the data. Based on the variable of prediction, we frame the issue as logistic regression.

2. What are the main features (also called independent variables or predictors) that you'll use?

My independent variables are family size, male occupation, female occupation, and whether the family members were born slaves. The output variable is born slave (1 indicates yes, and 0 indicates no).

3. Which machine learning technique will you use?

Multiple logistic regression seems to fit best because born slave is a binary variable. We consider the problem of predicting born slave using multiple predictors.

4. How will you evaluate the success of your machine learning technique? What metric will you use?

For this project, a training dataset will be created with 80% of the data from the original dataset, and the testing data set includes 20% of the data.

i) This logistic regression is used to test, make predictions of the probability of literacy within these households.

ii) The metrics used to evaluate the model will be the confusion matrix in order to determine specificity and sensitivity. These will prove critical for identifying the ROC curve.