It is commonplace for agricultural data to contain spatial information -- typically the longitude and latitude at which each data point was collected. However, one of the challenges this introduces is that different datasets often do not contain data that were collected at the same locations. A specific example involves planting and harvest data from precision farming equipment. As a planter moves through a field, it records thousands of measurements of its location along with various data about the planting operation (such as the variety of seed planted, the seeding rate, and the spacing between seeds). Similarly, as a harvester moves through a field, it records thousands of measurements of its location and the crop's yield at that point. A question of agricultural interest is to estimate the relationship between yield and the variables measured during planting. However, the harvest data points do not necessarily exactly overlap the planting data points, so before any analysis can be conducted to probe these relationships, it's necessary to determine which planting points are associated with which harvest points.

The planting and harvest files for one corn field are attached (both files are from the same year). Both tables contain a row for each data point collected by the machine. The columns in each file are as follows:

1. planting_sample_data.csv
    a. long: longitude where data point was collected.
    b. lat: latitude where data point was collected.
    c. variety: the seed variety planted at that location.
    d. seeding_rate: continuous variable specifying the number of seeds planted per acre (in thousands).
    e. seed_spacing: continuous variable specifying the distance between seeds (inches).
2. harvest_sample_data.csv
    a. long: longitude where data point was collected
    b. lat: latitude where data point was collected
    c. yield: continuous variable specifying the yield of the crop (in bushels/acre).

Your task is twofold:

1. Write a function that takes two arguments: the planting filename and the harvest filename. The function should read in the files from your local filesystem, and then determine the values of the planting variables (variety, seeding rate, and seed spacing) that should be associated with each harvest point. Please design, describe, and implement an algorithm that performs this association. The function should return a data frame containing the same number of rows as the harvest file, but with three extra columns, containing the values of variety, seeding rate, and seed spacing associated with each harvest point.
2. Use the output of the function you wrote in step 1 to perform some exploratory data analysis on the data provided. The goal is to quantify how the three planting variables (variety, seeding rate, and seed spacing) are associated with yield. This exploratory

analysis should involve some data visualization and some implementation of statistical models.

Please provide the output of the function in step 1 as a .csv file, any plots or relevant data summaries from step 2, a written summary of your findings (describe your plant-harvest point association algorithm, its efficiency, and your data analysis/modeling approach from step 2), and all code necessary to recreate this analysis.

Feel free to contact Matt Meisner (608-234-8161, mmeisner@farmersbusinessnetwork.com) with any questions!