FBN Data Science Take-Home Question (page 1 of 2)

Background

Agricultural data frequently contains spatial information -- typically the longitude and latitude at which each data point was collected. One specific example involves planting and harvest data from precision farming equipment. When a planter moves through a field in April, 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, when a harvester moves through that same field in October, it records thousands of measurements of its location and the crop's yield at that point.

Unfortunately, the planting and harvest data are almost never collected at exactly the same locations on a given field. Thus, if we want to estimate the relationship between harvest yield and the variables measured during planting, we must first 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).
    f. speed: continuous variable specifying the driving speed of the planter (miles per hour).
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).

Part A - Associate the Planter and Harvester Data

Write a function to associate the planting variables with the harvest points. The function should do the following:

1. Take two arguments: the planting filename and the harvest filename
2. Read in the files from your local file system
3. Return a data frame with the same number of rows as the harvest file, but with at least three extra columns, containing the values of variety, seeding rate, and seed spacing associated with each harvest point.

Please provide a brief written description of your plant-harvest point association algorithm, including some discussion of its efficiency.

Part B - Exploratory Data Analysis

Use the output of the function you wrote in Part A to do the following:

1. Perform an exploratory analysis of the combined plant-harvest data.
2. Quantify how the three planting variables (variety, seeding rate, and seed spacing) are associated with yield.
3. Conclude by telling us what else you might do if you had more time to analyze the data.

Your analysis should include some data visualization and statistical models. You should also provide brief written explanations with each step you take explaining what you are doing, why you are doing it, and what you found. In particular, whenever you present results in the form of a graph, summary, or table, briefly describe in writing what you think those results mean.

Your response to Part B of this take-home question is a very important part of your application to the Data Science team at FBN. We are looking for you to demonstrate thoughtfulness, creativity, and a solid understanding of your chosen statistical tools.

What to Submit to Us

1. The output of the function you wrote in Part A as a .csv file
2. A brief written description of your plant-harvest point association algorithm, including some discussion of its efficiency
3. Your written response to Part B, including any plots or relevant data summaries and all code necessary to recreate your analysis