Overview

You've come so far in your data science journey! Over the past few weeks you've covered some essential math for machine learning, data science research methods, and kept improving your R and Python programming skills.

In the third project you're going to be applying some of those data science research method skills. You can choose to do this project in either R or Python.

Concepts covered:

- Data Science methodology
- Statistical research methods
- Statistical programming in R or Python

The Dataset

The dataset we'll be contains housing price data from Boston, Massachusetts, USA; a description can be found here. This is a common data set and is available directly in R and in the *sklearn.datasets* module from scikit-learn in Python. A quick internet search will reveal directions for loading it. Alternatively, the data set can be downloaded directly from here.

Within the data set you'll find variables such as per capita crime rate, average number of rooms, property tax rate, and more. Note that the variables are not for individual households, but for groups of households known as "census tracts."

For this assignment, you should assume that the data is from a **sample** of census tracts in the Boston area, not all census tracts.

Requirements

For this project you should produce a Python or R notebook. There are two sets of specific questions as well as an open-ended experimental design. Your project will be evaluated on how well it answers the questions and on the quality of the experimental design:

1. Choose a variable other than CHAS and MEDV (the target, median home price).

    1. Compute the mean and standard deviation of the variable.
    2. Plot a histogram of the variable.
    3. What is the sample correlation between your chosen variable and median home price?
    4. Perform a regression, predicting MEDV from your chosen variable.

2. You have a theory that tracts that border the Charles River (CHAS) will have higher median price (MEDV or target) than those that do not.
    1. What is the null hypothesis?
    2. Calculate the *p*-value. Use the sample mean of the target as an estimate of the population mean.
    3. What is the 90% confidence interval for the target (price) of tracts that border the Charles River?
    4. Assume an effect size (Cohen's *d*) of 0.6. If you want 80% power, what group size is necessary?

3. Imagine you are the city planner of Boston and can add various new features to each census tract, such as a park. Be creative with your new "features" – we use the term loosely. You can assume that none of the tracts contained your features previously. Design an experiment to explore the effects of these features on the media house price in census tracts. You should include an explanation of the experimental design as well as a plan of analysis, which should include a discussion of group size and power. Be sure to apply the knowledge you learned in the Data Science Research Methods courses.

4.Submission

When you have completed the project, email your instructor a single notebook (R or Python). Include text blocks that describe how you answered the questions and your experimental design as well as any necessary code blocks.