Exercise 04

This should be completed **individually**.

Section 1: Analysis

A friend of yours works at the local frozen custard place. On hot days, she says her work is harder. It feels like she doesn't get a break. She says it is because they have too many customers. However, even before your friend started working there, she always seemed tired and cranky when the weather was hot. You think it is because she doesn't tolerate the heat well. She makes a bet with you that there really are more customers on hot days. She gets data on the sales numbers and the daily temperatures.

1. What is the outcome?
2. What is the main effect/predictor she wants to understand the impact of?
3. What is the hypothesis?

Use the data she collected to conduct an analysis, test the hypothesis, and report results. The dataset is frzn_custard.xlsx. Your analysis should have the following elements:

1. An explanation of why the analysis is being conducted and what the hypotheses are
2. Descriptive information about the data, including summary statistics (such as number of observations, measures of central tendency, & measures of dispersion) and plots of the data distributions
3. Descriptive information about the relationships between the two variables, including correlation and scatterplots
4. A regression analysis to test the hypothesis. If you have trouble getting the regression analysis to work, look closely at the data. Your friend wasn't always able to get sales data for each day. Rows with missing data will need to be removed.
5. A description of the results of the analysis. Included in this description should be an interpretation of the coefficients, description of the goodness of fit, and a discussion of whether the results are statistically significant.

Section 2: Gradient Descent

1. Write a program in Python that uses gradient descent to find the regression coefficients (βs) for the frozen custard data in Section 1. There are many examples of how to do this on the internet. Most of these examples use the `np.dot` rather than `np.matmul`. Make sure your program uses `np.matmul` instead.
    a. Cite the source you used
    b. This problem is an example of how gradient descent can fail with raw data. There are more sophisticated modifications to gradient descent that address those failures. When you try this problem, it will not converge. Try it anyway and graph the loss function over iterations so that you can see what is happening.
    c. Standardize your X and Y variables. To do this, subtract the mean from each value and divide it by the standard deviation. Please note that the input array for X needs to have a column of 1's in it. Do not standardize the 1's.

d. Rerun your section 1 model using the standardized inputs. Compare your results using gradient descent and using the module you used in Section 1 for the standardized inputs.
e. Plot the loss (i.e. cost) function over the iterations