Exercise 03

This should be completed **individually**.

Section 1: Analysis

A friend of yours owns an frozen drink shop. On hot days, she seems extra happy. She says that she sees a line extending around the block and knows that means more sales. However, even before your friend started the business, she always seemed to love summer and thrived in the heat, so you think her jubilant attitude might be intrinsic. She makes a bet with you that sales really are higher on hotter days. She gets data on the sales numbers (in dollars) and the daily temperatures (in degrees Fahrenheit).

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 drinks.xlsx. Your analysis should have the following elements:

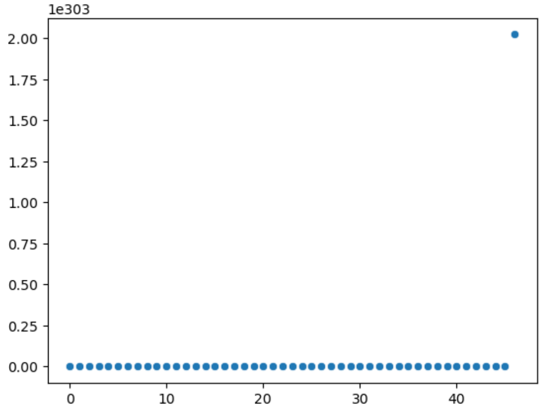
7. An explanation of why the analysis is being conducted and what the hypothesis is
8. Descriptive information about the data, including summary statistics (such as number of observations, measures of central tendency, & measures of dispersion) and plots of the data distributions
9. Descriptive information about the relationships between the two variables, including correlation and scatterplots
10. 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. Choose a method to handle rows with missing data.
11. 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

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.

1. Cite the source you used
2. 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 1000 iterations so that you can see what is happening. Produce 3 graphs of iterations (x) and loss (y) for 3 learning rates (0.1, 0.01, 0.001)

Here is an example scatterplot of iterations vs loss



3. 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.
4. Rerun your section 1 model using the standardized inputs and 2 different learning rates (0.01, 0.001). Compare your results using gradient descent and using the module you used in Section 1 for the standardized inputs.
5. Plot the loss (i.e. cost) function over the iterations for both learning rates on the same graph