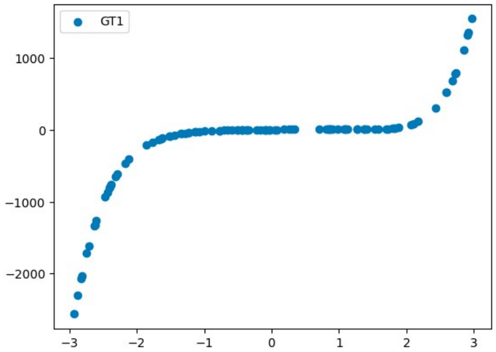
**Assignment 1**
**CS 5630/6630**

***SUBMIT THE CODE IN AN IPYNB FILE (using Google Collab). NO OTHER FORMAT IS ACCEPTED NOT EVEN .PY***

**[Question 1] [15 points]**

Suppose that you are conducting a scientific experiment where you are observing the effects of one variable (*x_train.npy* and *x_test.npy*) on the output (*y_train.npy* and *y_test.npy*). On visualizing the relationship between the variables, you see the following plot:



Your goal is to come up with a linear regression model that can take the training data (*x_train.npy* and *y_train.npy*) and model the relationship between the variables *x* and *y*. You should implement your own version of linear regression either using gradient descent or normal equations. **You SHOULD NOT use any pre-packaged library such as Sci-Kit Learn.**

Here are somethings to keep in mind for tackling this problem:

1. Try to plot this relationship on your own using matplotlib. You can also visualize the test data to see if it gives you any clues about the underlying relationship between the variables.
2. Use your knowledge gleaned from the previous step to answer the following questions:
    a. Is the relationship linear?
    b. Do you need feature engineering to add any non-linearity?
        i. If so, how can you engineer these features?
        ii. What are some functions that you can try?
            1. Plot each of them individually to verify!

You will need to write a short report detailing your thought process, the code you wrote in Python to implement the linear regression model and the equation that models the relationship between *x* and *y* that you found. You should provide evidence that corroborates your final statement such as plots, prediction errors, etc.

**[Question 2] [20 points]**

Imagine that you are a realtor in Auburn. You have data points (See excel file. Last column is the target variable.) that correspond to the recent sales of different houses in and around Auburn. Your goal is to help estimate the prices of houses that one can use to sell or buy listings. Can you use your knowledge of linear regression to find the best regression model? Use your implementation from Question 1 (without any basis functions) to answer the following questions:

1. What is the average least squares error for the given data using your simple linear regression model?
2. Which factor has the most effect on the final value? How do you know this? Can you use only this feature to predict the price?

3. Which factor has the least effect on the final value? How do you know this? What effect does removing this feature have on the performance?

**[Question 3] [15 points]**

Implement a locally weighted linear regression model for the data from Question 1. Refer to Slide 33 from Lecture 4 for reference. You should implement your own version of linear regression either using gradient descent or normal equations. **You SHOULD NOT use any pre-packaged library such as Sci-Kit Learn.**

Answer the following questions:

1. Do you need any basis functions when using the locally weighted approach?
2. What is the difference between this implementation and the one for Question 1?

**Submission Requirements:**

You will need to submit the following as a single IPYNB file:

1. Use the “text cells” on Colab to write your report.
2. Include a cell with README instructions at the top of your notebook to note on any dependencies that are required to run your code.

**Note**:

1. If your code does not run on Colab, you will not get any credit for the code segment. We will only grade what is in your report.
    a. This includes any syntax errors due to indentation, unnamed/unknown libraries that were not listed in the README file, etc.
2. Please submit code only in Python and in the IPython notebook format. You can write your answers as part of the notebook if you do not want a separate report file, but it must be comprehensive.
    a. Any code not in Python will not be graded at all.