Due data: 10/2/2024, end of the day. **Please submit an .ipynb file via Canvas.**

Instructions:

1) The .ipynb file shall include not only the **source code**, but also necessary **plots/figures** and **discussions** which include your *observations*, *thoughts* and *insights*.
2) Please avoid using a single big block of code for everything then plotting all figures altogether. Instead, use a small bock of code for each sub-task which is followed by its plots and discussions. This will make your homework more readable.
3) Please follow common software engineering practices, e.g., by including sufficient **comments** to functions, important statements, etc.

Programming Problem:

Write a program to find the coefficients for a linear regression model for the dataset provided (data2.txt). Assume a linear model: $y = w_0 + w_1 * x$. You need to

1) Plot the data (i.e., x-axis for the 1st column, y-axis for the 2nd column),

and use Python to implement the following methods to find the coefficients:

2) Normal equation, and

3) Gradient Descent using **batch** AND **stochastic** modes respectively:
   a) Split dataset into 80% for training and 20% for testing.
   b) Plot MSE vs. iteration of **each mode** for **both training set and testing set** (i.e., batch – training and testing; stochastic – training and testing). Compare (with discussions) batch and stochastic modes (with discussion) in terms of accuracy (of testing set) and speed of convergence (You need to determine an appropriate termination condition, e.g., when cost function is less than a threshold, and/or after a given number of iterations.)
   c) Plot MSE of the testing set vs. learning rate (using 0.001, 0.002, 0.003, 0.004, 0.005, 0.006, 0.007, 0.008, 0.009, 0.01) and determine the best learning rate.

Please implement the algorithms by yourself and **do NOT use the fit() function** of the library.