[50 points]

In this coding assignment you are to implement a linear regression algorithm and use it with a real world dataset. The dataset contains a collection of real estate listings in San Luis Obispo county.

https://wiki.csc.calpoly.edu/datasets/wiki/Houses

To make it simpler for this assignment, the dataset has been slightly modified. It contains the following fields:

1. MLS: Multiple listing service number for the house (unique ID).
2. Price: the most recent listing price of the house (in dollars).
3. Bedrooms: number of bedrooms.
4. Bathrooms: number of bathrooms.
5. Size: size of the house in square feet.
6. Price/SQ.ft: price of the house per square foot.

Write a program to predict housing price vs size using batch gradient descent. Some MATLAB commands are suggested below but you don't have to use these commands. Type "doc <command>" at the MATLAB prompt for more information.

1. Read "housing_price_data.dat"
    - MATLAB: dlmread
2. Take the price and size columns only (columns 2 and 5). Sort the data then scatter plot it (price vs size). Label the plot
    - MATLAB: sort, scatter, title
3. Remove the outliers, i.e., consider only entries 25 to 600
    - By now you should have 2 vectors (matrix of size 576x2), the first vector for the size of the house and another vector for the price of the house.
4. Normalize the size-of-the-house vector, i.e., divide by (max-min). This is to ensure the scale is compatible with the bias, which is 1.
5. Initialize the weight vector randomly
6. Perform gradient descent until convergence
    - You can define convergence by some measure, such as the difference between the MSE of the previous iteration and the MSE of the current iteration being less than some small number
7. Plot the line defined by the optimal weight vector over the same scatter plot of the training examples
    - MATLAB: figure, hold off, plot
    - "hold off" causes MATLAB to plot over the current figure
8. Print the number of iterations required to achieve convergence for different learning rates.
9. Plot error vs iterations for a given learning rate

Hints:

```
x=dlmread('housing_price_data.dat');
F=sort([x(:,5) x(:,2)]);

figure; scatter(F(:,1), F(:,2));
title('Scatter plot of housing prices');
xlabel('Size in square Feet');
ylabel('Price');

%% Linear Regression/Gradient Descent
%% Your code ...

% Display result
figure; hold off;
scatter(x(:,2)*normalize,y, 5, '.'); hold;
plot(x(:,2)*normalize,yhat,'r');
   % yhat is the predicted price of the house
   % normalize is the normalizing factor (max-min)
xlabel('Size in square Feet'); ylabel('Price');

figure; plot(err_arr);
   % err_arr is an array of the error for each iteration
xlabel('Iterations'); ylabel('Error');

fprintf('Algorithm converges after %d iterations, learning rate=%5.3f \n', itr,
nu);
```