Assignment 2

Q1. Generate a dataset using numpy random as I had shown in the 1st class of this week. Train a linear regression model using |x-xhat|^3 as your loss function and a polynomial regression model using |x-xhat|^7 as your loss function. (Note that you will need to derive the gradient descent algorithms for these functions yourselves). You are allowed to use only numpy, pandas and Matplotlib. Then train a linear regression model using the sklearn library on the same dataset. At last, plot the dataset and curves obtained from all models in the same figure.

Q2. Dataset: Air quality of an Italian city
(https://archive.ics.uci.edu/ml/datasets/Air+quality)
The dataset contains 9358 instances of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multi Sensor Device. The device was located on the field in a significantly polluted area, at road level, within an Italian city. Data were recorded from March 2004 to February 2005 (one year) representing the longest freely available recordings of on field deployed air quality chemical sensor devices responses. Ground Truth hourly averaged concentrations for CO, Non Metanic Hydrocarbons, Benzene, Total Nitrogen Oxides (NOx) and Nitrogen Dioxide (NO2) and were provided by a co-located reference certified analyzer. Missing values are tagged with -200 values.

Your objective is to predict the Relative Humidity of a given point of time based on all other attributes
affecting the change in RH.

(i) Perform the data pre-processing steps on the dataset as explained in the class. Handle missing values, get insights from correlation matrix and deal with outliers.

(ii) Split the dataset into a 85:15 ratio into training and test dataset using the sklearn library.

(iii) Train a linear regression model from scratch using only numpy, pandas and matplotlib and train a linear regression model using the sklearn library on the training dataset.

(iv) Calculate the r2 score and mean squared error using the test dataset. Compare the results obtained and plot your results.