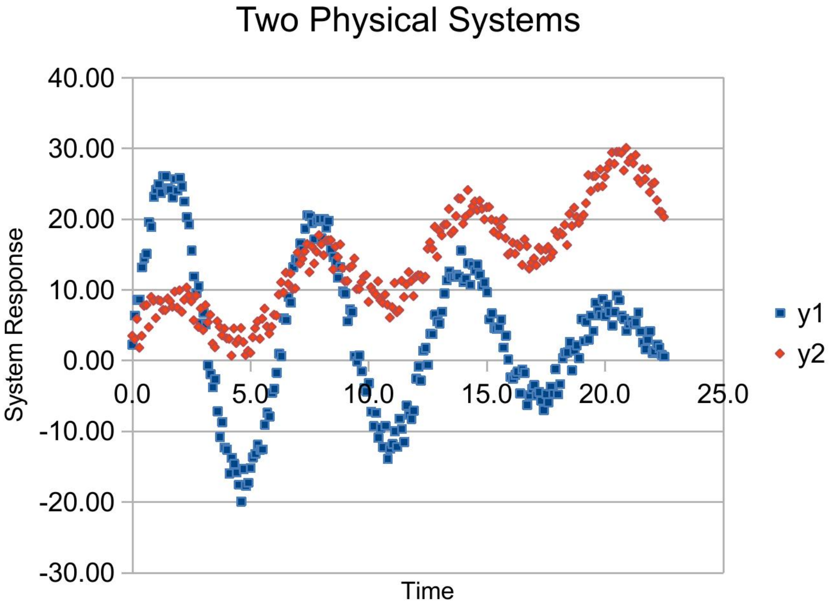
DA5401 Assignment #2

Let's rollup our sleeves to get into the first data science task, where we try to model two cyber physical tasks. The first physical system in a suspended spring with an unknown load at the end. The second physical system is a stock price tracker. Our goal here is a) to model the harmonic oscillation of the spring and b) to forecast the stock price at a future timeframe. The supplied dataset consists of two columns representing both physical systems (call them 'y1' and 'y2'). Assume that the data points are sampled at constant time intervals (call it 'x'); the scale does not matter. The visualization of the data points, assuming the same time scale, will look like below.



Task 1 [20 points]

Let's consider the 'y2' dataset against time 'x'. Fit a linear model y2 = mx that minimizes the SSE. Note that there is not need for the 'c', as you observe that the trend line (linear model) is passing through '0' (origin).

1. Implement the OLS closed form solution using numpy's matrix operators to find the value of 'm' that minimizes SSE.
2. Implement a linear search (the single parameter search version of grid search) for m = tan θ, where θ in [0, 60] in steps of 5 degrees and measure the SSE at every choice of θ. Create a plot that shows SSE vs θ. Report the θ, that minimizes SSE.
3. Implement the solution using sklearn's LinearRegression class.
4. Compare the estimated 'm' values through the above three methods and justify the differences if there are.

Task 2 [15 points]

You will notice that the linear model is an ok fit for the y2. What should be the mathematical model of stock price dataset? If you notice the periodicity in the data, you should factor that in your mathematical model using an appropriate function that's periodic. The challenge here is; the trend of the magnitude is also increasing, which you confirmed in your previous task. So, the math model should consider both properties.

1. Split your data into Train, Eval & Test.
    - ***Interpolation***: When you randomly split the data into train, eval and test; your test and evaluation data points may be inside the data range (time range). When you can predict those points correctly, you are essentially recovering missing data in the regression line. This is also called the interpolation problem.
    - ***Extrapolation***: In this scenario, the test and eval points should be outside the time range of the training data. If your model is a good fit, and when you predict the data point outside the range, you are essentially extrapolating the regression line. This is also called the "Forecasting" task.
2. Implement the regression model (OLS or LinearRegression or equivalent) using appropriate feature transformation so that the SSE is lower than that of Task 1.
3. Train the regression model for interpolation and evaluate the SSE.
4. Train the regression model for extrapolation and evaluate the SSE.

Task 3 [15 points]

Having finished the mathematical models of stock price prediction, let's switch to the mathematical model for approximating the oscillation of a loaded spring. Here you notice that the oscillations have a constant time period, but the height of the wave is decreasing until cessation. You should repeat Task 2 on the spring system dataset.

Please note that you are building only the linear models despite the datasets look non-linear.