Anirudh Rao
BE21B004

DA5401 – Data Analytics Lab

Assignment 2

Task 1

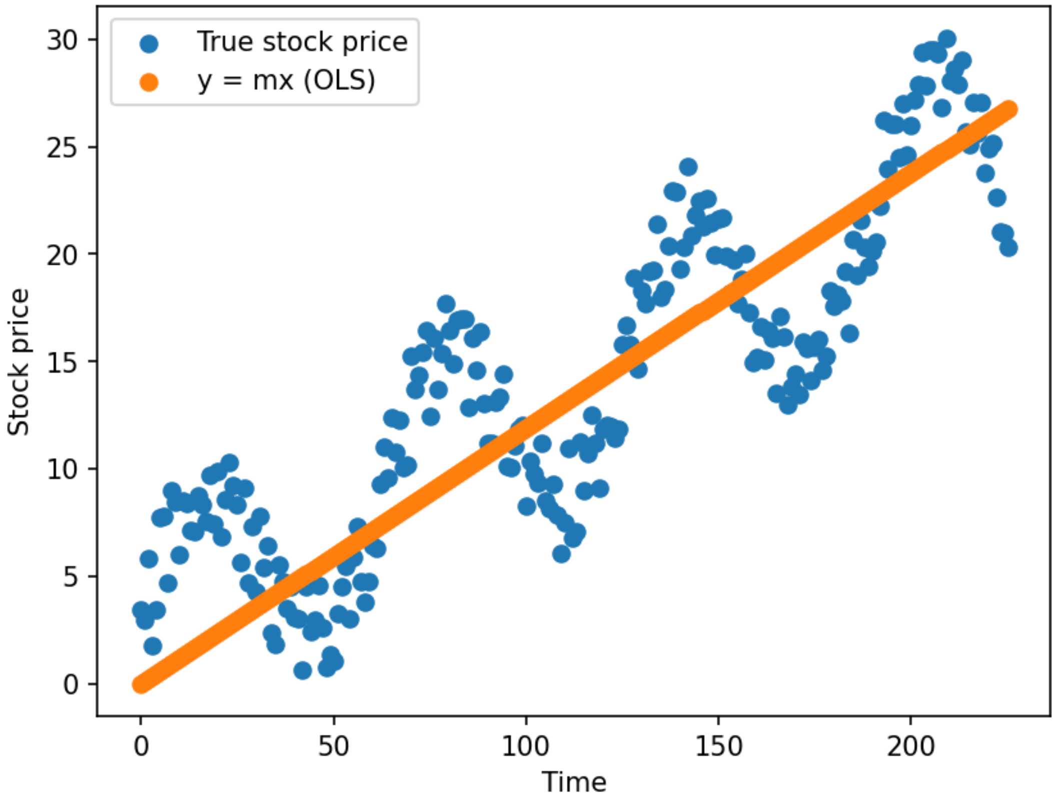
1. For the stock price data, the OLS closed form solution was used to fit a line with slope m and zero intercept. The indices of the data points, i.e., 0,1,2,…,225, was used as a proxy for the timepoints X. The values of the stock prices was taken as y. The OLS closed form solution is given by:

$$m_{\text{OLS}} = (X^T X)^{-1} X^T y$$

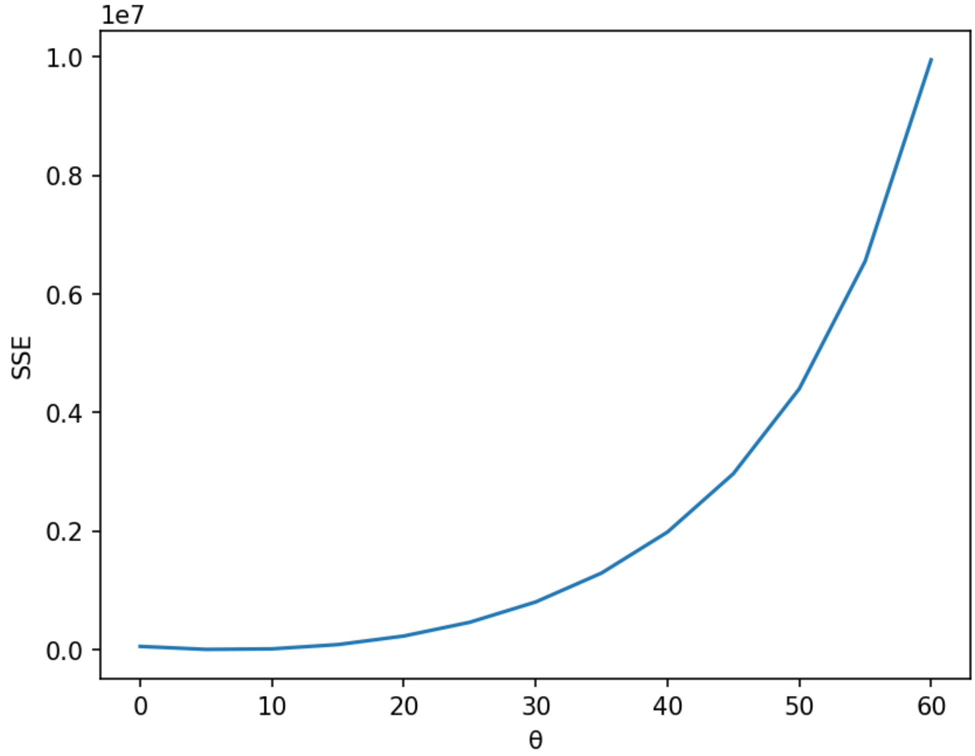
This was implemented using inbuilt Numpy functions and operators. From this, the slope was found to be $m_{\text{OLS}} = 0.11899412514961835$.

$$\boldsymbol{m_{\text{OLS}} \approx 0.119}$$

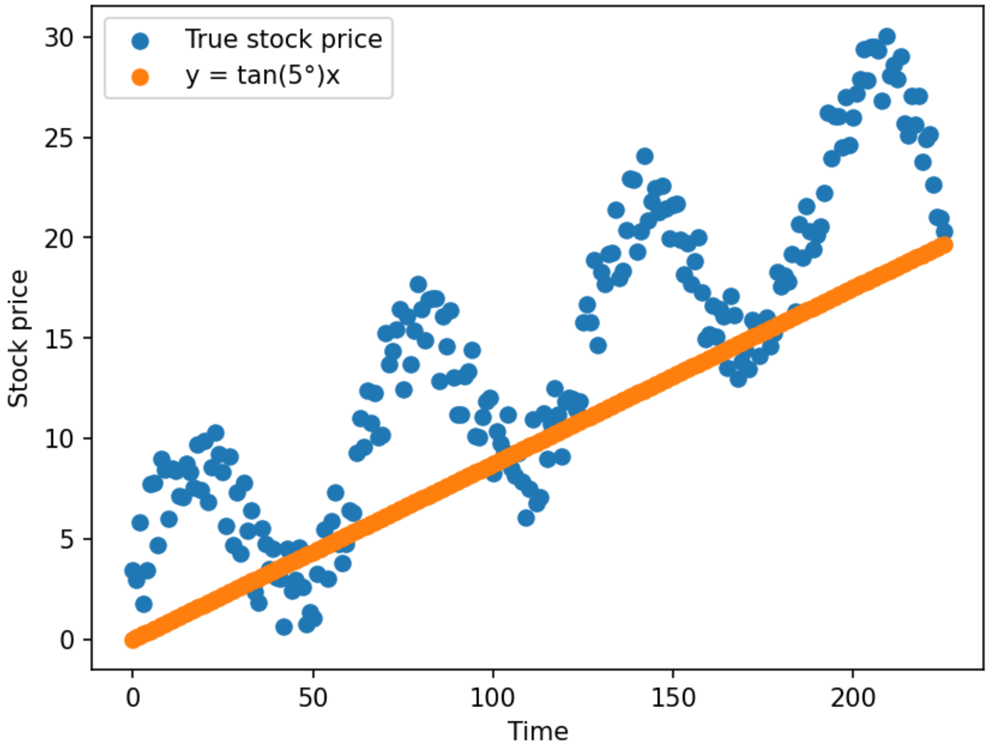
Plotting $y = mx$ gives us the linear fit for the stock price data. This has an SSE of 3850.335.



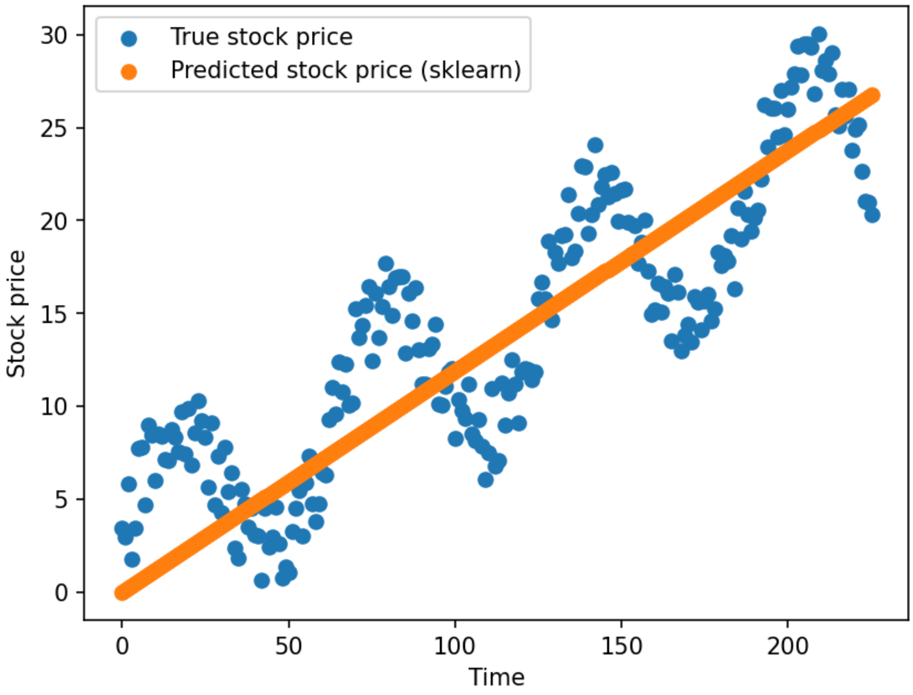
2. The alternative approach of performing a linear search for $m = \tan\theta$, with θ going from 0° to 60° in steps of 5°, was carried out. The plot of the variation of SSE with θ is shown below:



This revealed that the θ that minimised the SSE was $\boldsymbol{\theta = 5°}$, with an SSE of 7644.253. This value of θ is close to $\tan^{-1}(m_{\text{OLS}}) \approx 7°$.



3. Next, the same regression problem was solved using sklearn's `LinearRegression` class. This had an SSE of 3850.335, the same as the OLS solution.



4. The slopes of the regression lines from the three methods are:

$$m_{\text{OLS}} = 0.119$$

$$m_{\tan} = 0.087$$

$$m_{\text{sklearn}} = 0.119$$

The slopes from OLS and sklearn are the same, as expected. The slope obtained from the linear search is lower due to the step size of the linear search. If we used a step of 1, we would have identified $\theta = 7°$ as the θ that minimises the SSE, obtaining a slope similar to OLS and sklearn.

Task 2

1. The stock price data was used to construct separate training, evaluation, and test datasets for interpolation and extrapolation tasks. For interpolation, the original data was randomly split into the three sets with a 70%-15%-15% split. For extrapolation, the first 70% datapoints were used for training and the remaining 30% datapoints (from later timepoints) were split equally between evaluation and test datasets.

2. The stock price data resembles a sine curve with linearly increasing midline. This can be modelled as

$$y = A\sin(\omega x) + mx$$

From the first cycle, we can see that the stock price has a minimum of 0 and a maximum of 10. This means that the amplitude is $A = \frac{10-0}{2} = 5$.

We can also see that 4 cycles were completed in 226 time steps. Thus, the time period is

$$T = \frac{226}{4} = 56.5$$

The angular frequency is

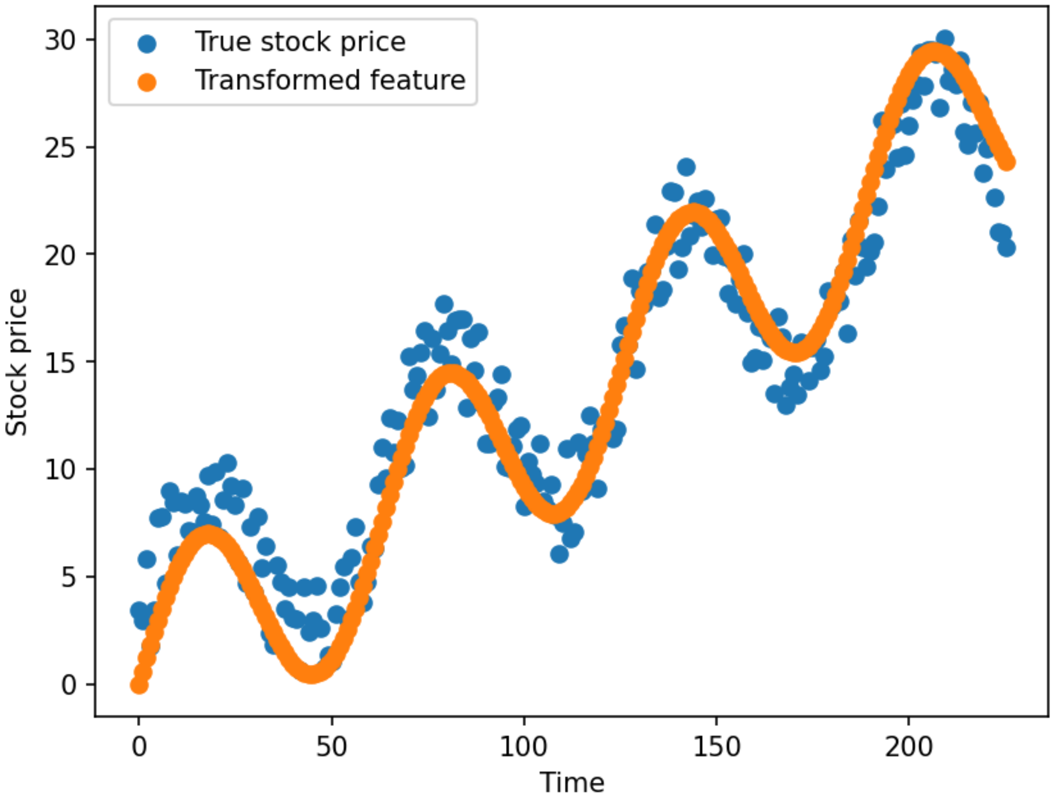
$$\omega = \frac{2\pi}{T} \approx 0.1$$

We already obtained the slope from the OLS solution as

$$m = 0.119$$

Thus, the feature transformation to be performed is

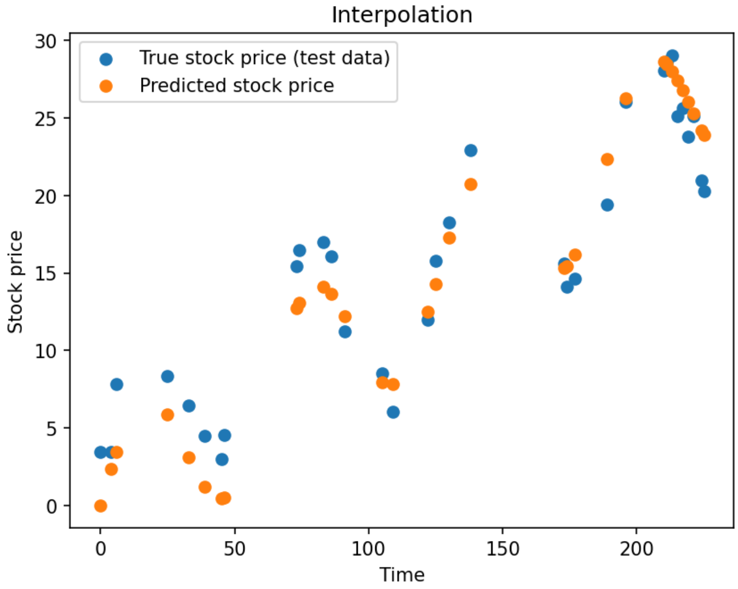
$$\boldsymbol{x \rightarrow 5\sin(0.1x) + 0.119x}$$



3. sklearn's `LinearRegression` class was used to fit the transformed feature against the stock price for interpolation. The performance metrics were obtained as:

$$\text{SSE(eval)} = 93.336$$

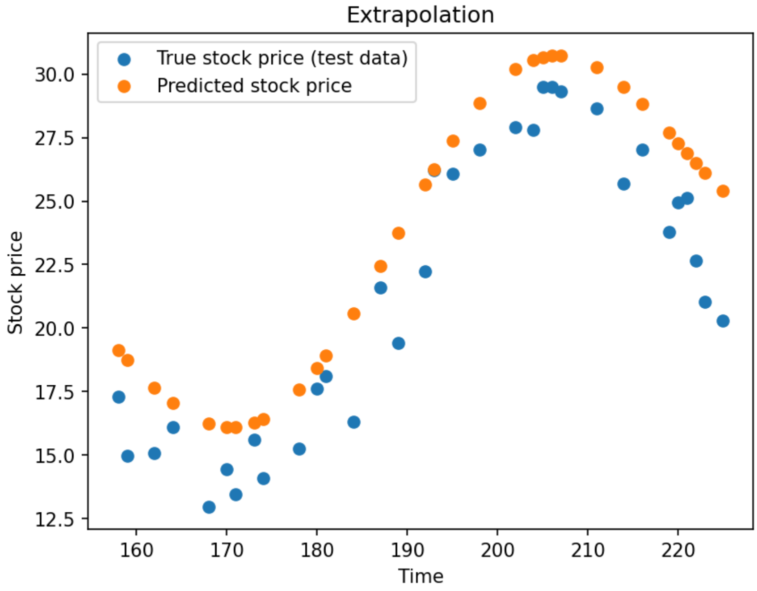
$$\text{SSE(test)} = 179.156$$



4. sklearn's `LinearRegression` class was used to fit the transformed feature against the stock price for extrapolation. The performance metrics were obtained as:

$$\text{SSE(eval)} = 237.834$$

$$\text{SSE(test)} = 241.053$$



Task 3

1. The spring position data was used to construct separate training, evaluation, and test datasets for interpolation and extrapolation tasks. For interpolation, the original data was randomly split into the three sets with a 70%-15%-15% split. For extrapolation, the first 70% datapoints were used for training and the remaining 30% datapoints (from later timepoints) were split equally between evaluation and test datasets.

2. The spring position data shows a damped oscillation. This can be modelled as

$$y = Ae^{-Bx}\sin(\omega x)$$

We can observe that the amplitude maximum (at $x = 0$) is 26.1.

We can also see that 4 cycles were completed in 226 time steps. Thus, the time period is

$$T = \frac{226}{4} = 56.5$$

The angular frequency is

$$\omega = \frac{2\pi}{T} \approx 0.1$$

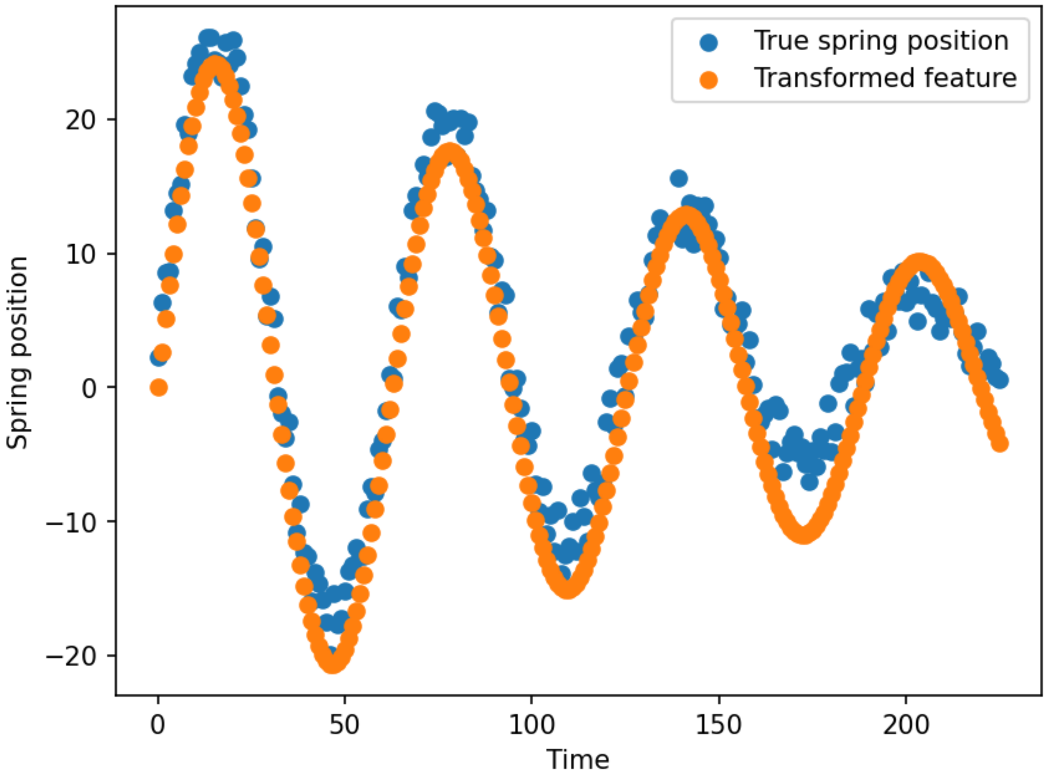
We observe that the amplitude maxima decreases from 26.1 to 20 in one time period. Mathematically,

$$\frac{20}{26.1} = e^{-BT}$$

$$B = \frac{-1}{T}\ln\left(\frac{20}{26.1}\right) = \frac{-1}{56.5}\ln\left(\frac{20}{26.1}\right) \approx 0.005$$

Thus, the final transformation is given by

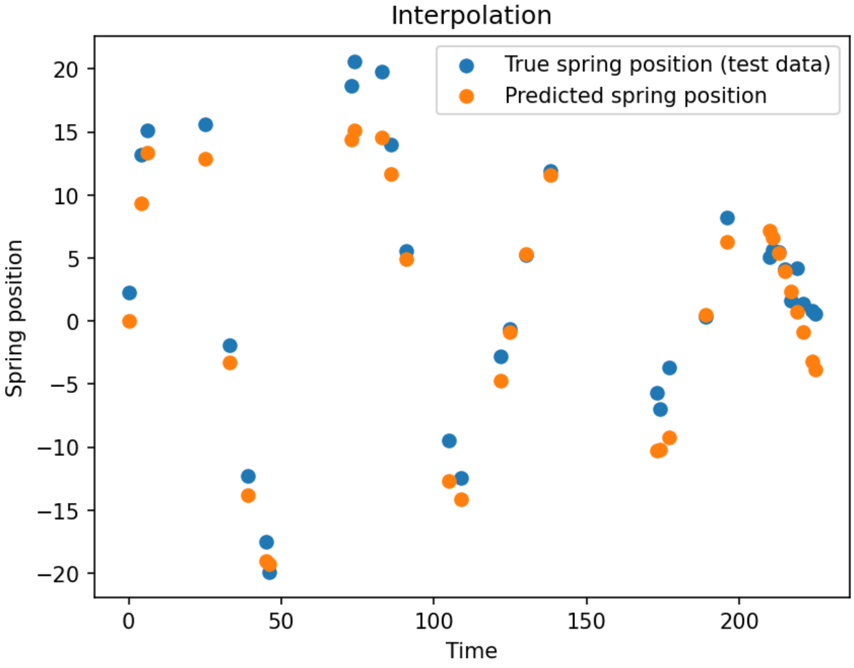
$$\boldsymbol{x \rightarrow 26.1e^{-0.005x}\sin(0.1x)}$$



3. sklearn's `LinearRegression` class was used to fit the transformed feature against the spring position for interpolation. The performance metrics were obtained as:

$$\text{SSE(eval)} = 364.791$$

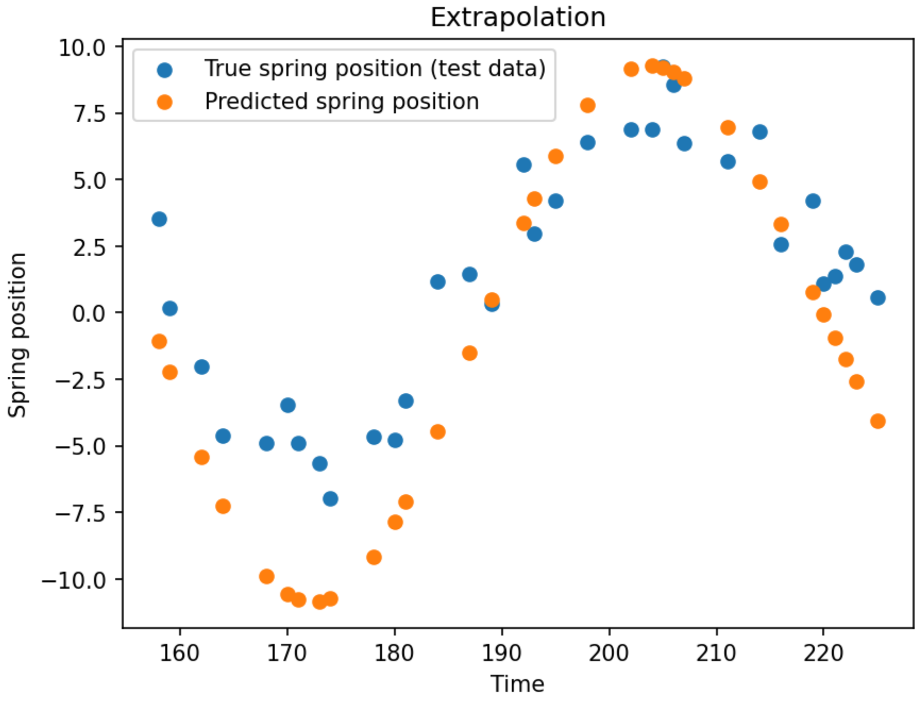
$$\text{SSE(test)} = 258.830$$



4. sklearn's `LinearRegression` class was used to fit the transformed feature against the spring position for extrapolation. The performance metrics were obtained as:

$$\text{SSE(eval)} = 566.629$$

$$\text{SSE(test)} = 390.936$$


Extrapolation
● True spring position (test data)
● Predicted spring position
Spring position
Time