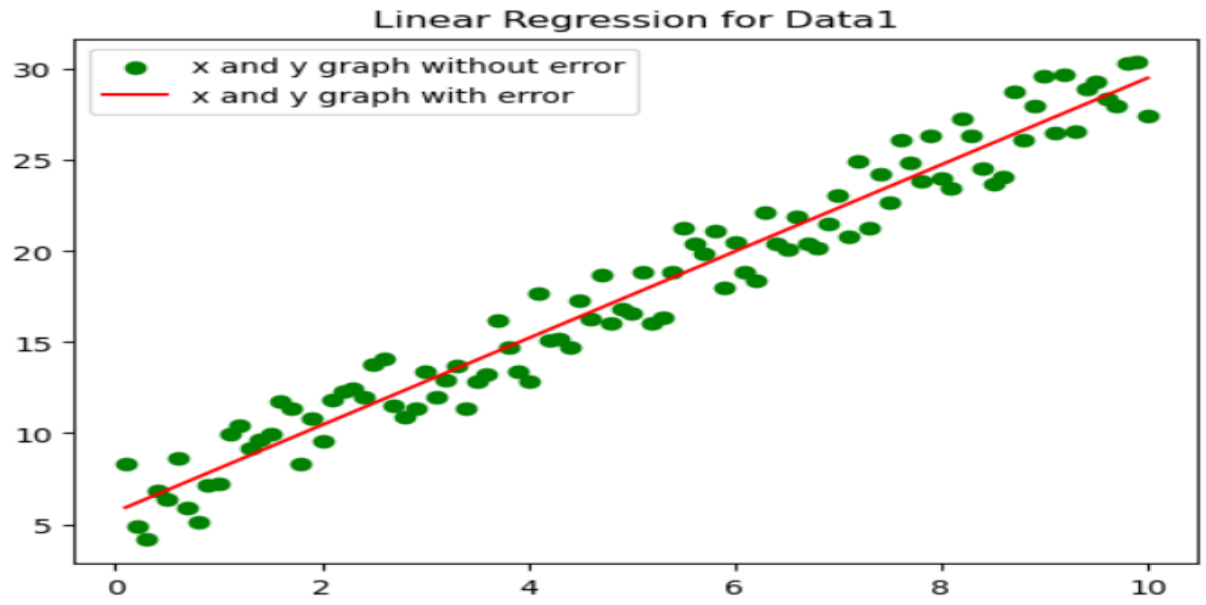
ML-Assignment 1

Linear Regression

Analysis of the datasets using Linear Regression. We have four datasets namely data1, data2, data3 and data4 . Analysis had been done for each of these data sets individually. In analysis we mainly focused on mean square error, mean absolute error, root mean square error and root square for the respective datasets. First we calculate these error by custom code and then we compared it with the scikit learn library.

Data Set 1

Output by	**MAE**	**MSE**	**RMSE**	**R-Square**	**Cost by GD**	**Cost by Numerical**
Numerical	1.280556	2.078525	1.4417091949	0.9998767	**[[5.68078713 2.38406007]]**	**[[5.68078713 2.38406007]]**
scikitlearn	1.280556	2.07525	1.441709145	0.95796		

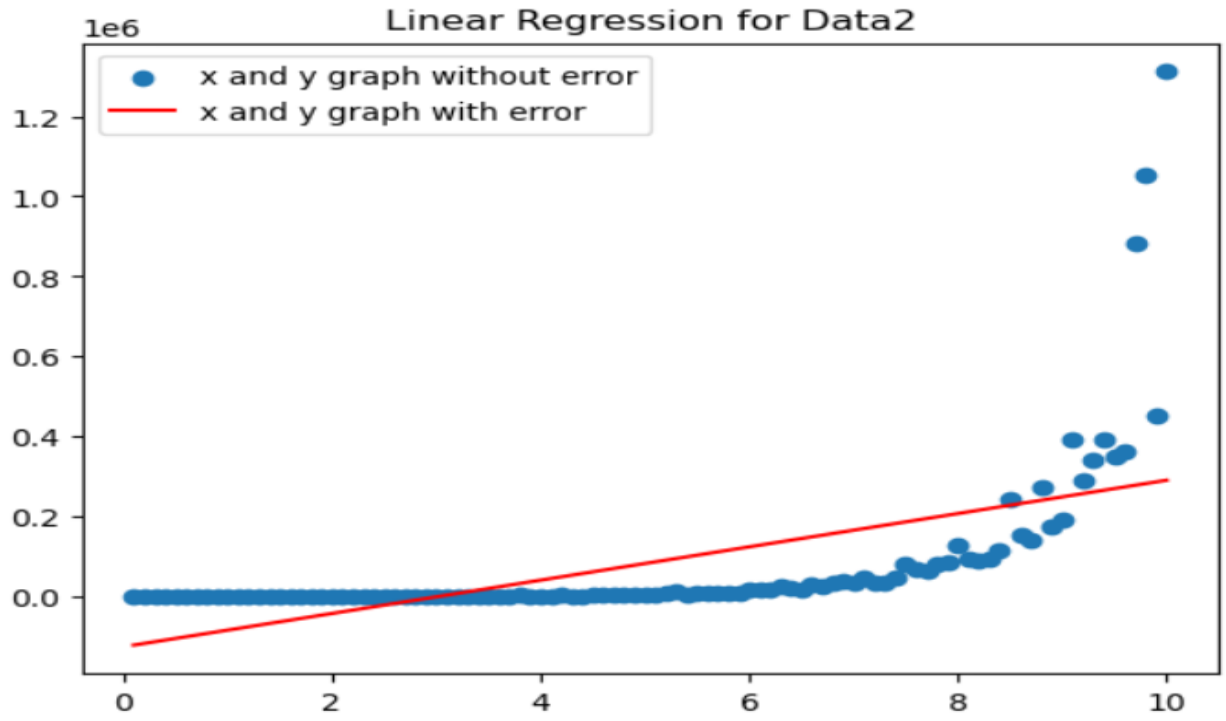


Analysis

Dataset 1: The dataset exhibits a notable linear trend when visualized on the x and y axes. The high R-squared value of 0.9579571905586357, nearing 1, suggests a robust linear relationship between the variables. Consequently, this dataset is well-suited for application with the standard Linear Regression algorithm.

Dataset 2 :

Output by	MAE	MSE	RMSE	R-Square	Cost by GD	Cost by Numerical
Numerical	99929.7832	27577785853.1640	166065.607074927	0.46225	**[[-125568.289921341603.63156151]]**	**[[-125568.2899213341603.63156151]]**
scikitlearn	99929.783297	27577785853.1634	166065.607074927	0.34339		

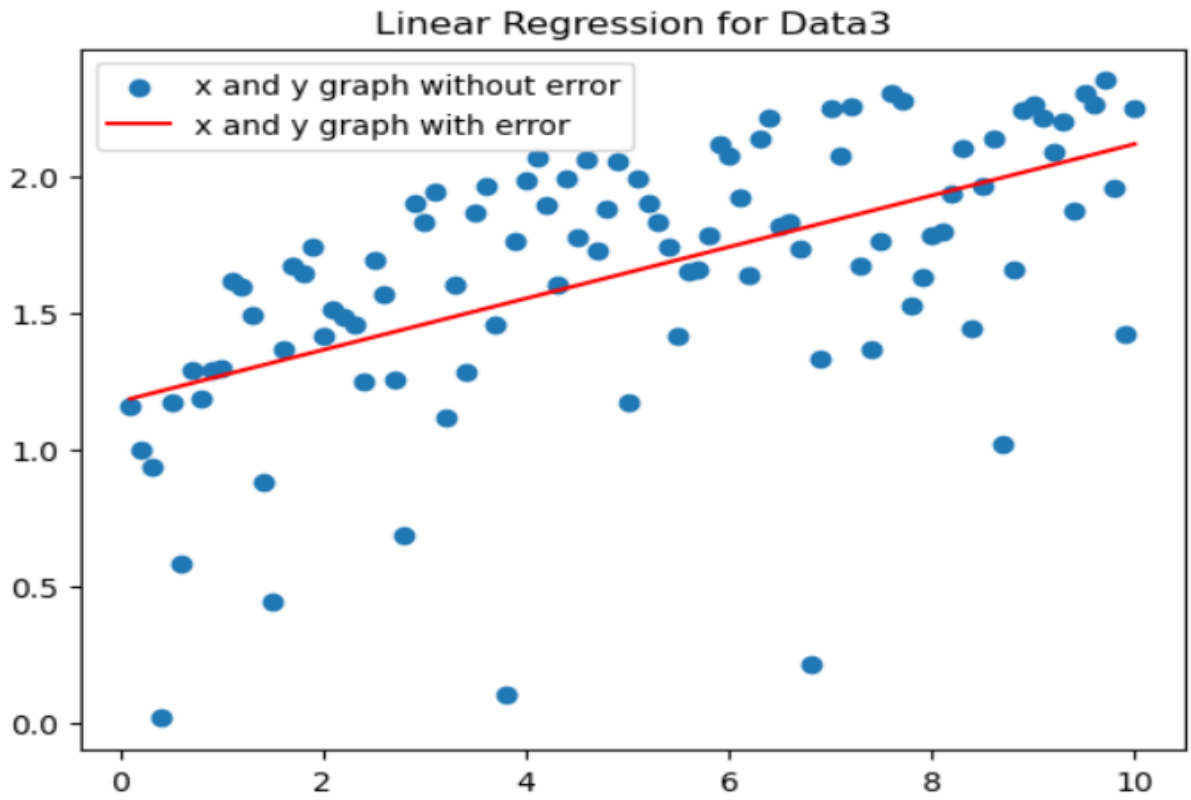


Analysis

Dataset2: At the outset, the dataset suggested an exponential trend among its data points. However, following a logarithmic transformation applied to the input features, the transformed data assumed a linear appearance. The R-squared value of 0.34339 Consequently, to render this dataset suitable for the standard Linear Regression algorithm, a nonlinear transformation of the input features is essential. The first plot illustrates the relationship between x and y before the transformation. If we convert this exponential into linear than this will best fit the line , this is because in that case our R-squared value turn out to be 0.9993234 which is closely to 1.

Dataset3:

Output by	MAE	MSE	RMSE	R-Square	Cost by GD	Cost by Numerical
Numerical	0.161730	0.294678		0.999999856	**[[1.17706208 0.09419021]]**	**[[1.17706208 0.09419021]]**
scikitlearn	0.161730	0.294678		0.31370		



Analysis

Dataset3: The widespread dispersion of points in the plane necessitates a nonlinear transformation of input features to establish an improved best-fit curve. In this dataset, the data points were widely scattered on the graph, posing challenges in identifying a clear pattern or relationship. With an R-squared value of 0.31369732267280803, considerably distant from 1, the evidence points to a weak linear connection between the variables. Consequently, the conventional Linear Regression algorithm may prove inadequate for this dataset. Exploring alternative regression methods or algorithms capable of handling non-linear relationships could be more suitable.

Dataset 4:

Output by	MAE	MSE	RMSE	R-Square	Cost by GD	Cost by Numerical
Numerical	5.155506	34.620481	5.883917133	0.99829	**[[13.23947579 6.13243433 2.3922683 7.74681094]]**	**[13.23947782 6.13243763 2.39226554 7.74681038]]**
scikitlearn	5.155506	34.6420481	5.883917113	0.98417		

Analysis:

Dataset4:

Consistent R-squared values near 1 from both sklearn and my analysis suggest a good fit for multivariate regression. The high R^2 indicates our model captures a large part of the data's variation. Lower MAE means our predicted curve closely follows the best-fit curve on average. Additionally, with an RMSE of 5.88 and an impressive R-Squared of 0.9983, our model exhibits strong performance, displaying minimal average errors and explaining nearly all the dataset's variability. In summary, our multivariate regression model demonstrates accuracy and effectiveness in representing the underlying patterns in the data.