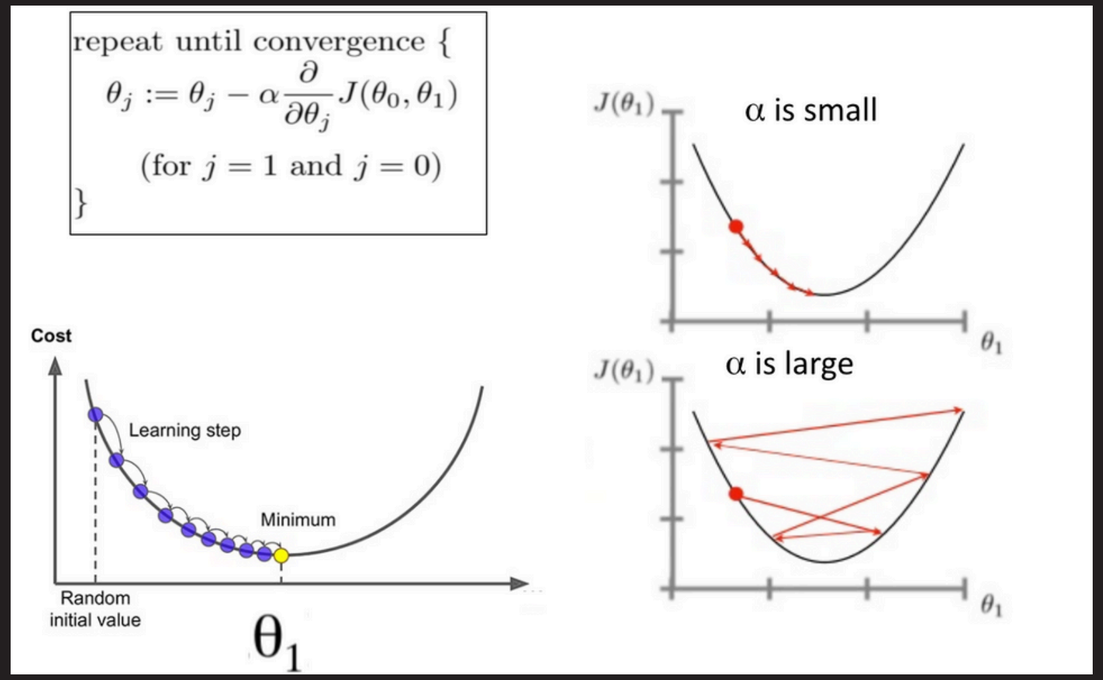
LINEAR REGRESSION WITH GRADIENT DESCENT ALGORITHM

200101005
YUSUF ÖMER TURSUN

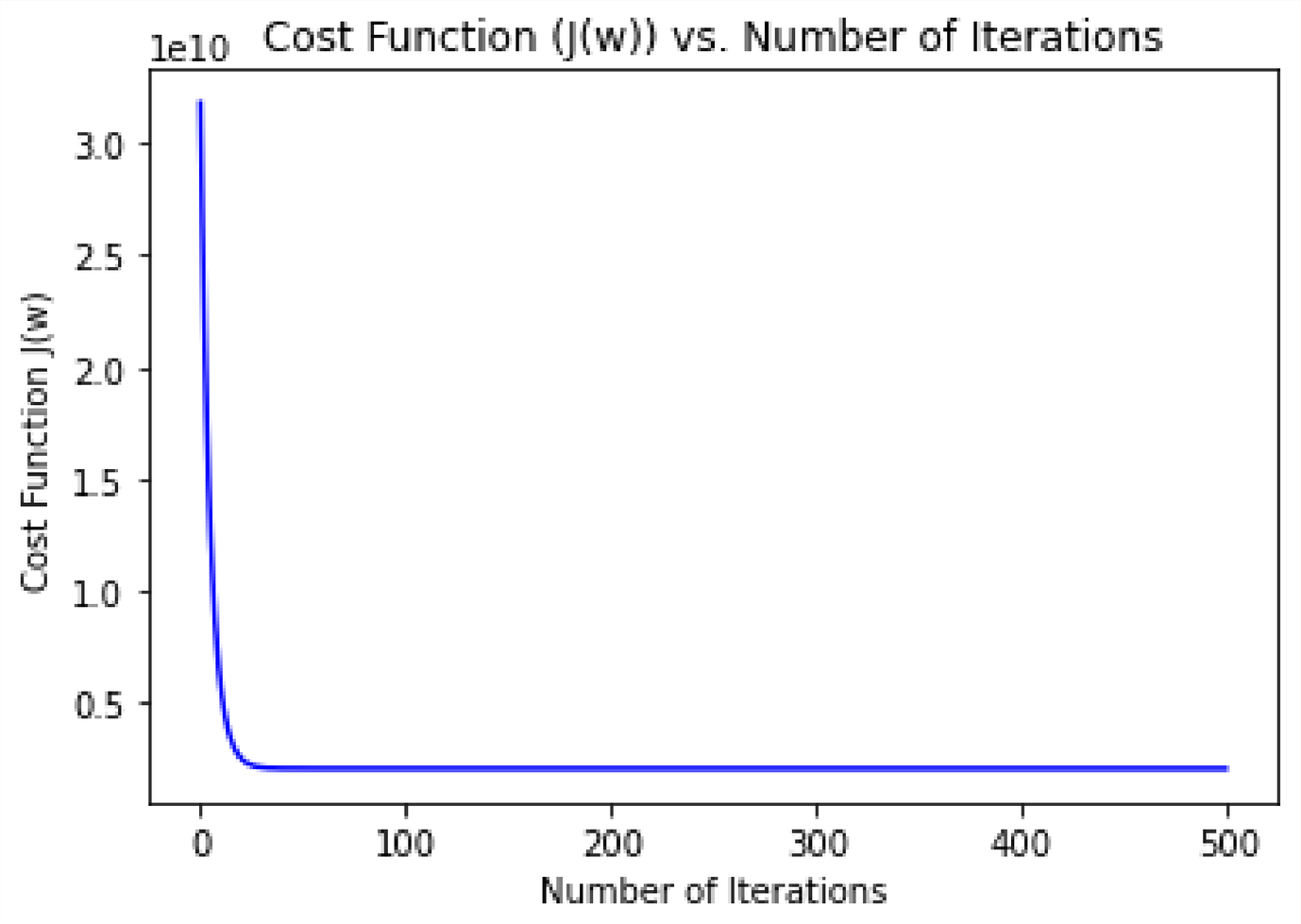


In this assignment we will try to implement a linear regression machine learning algorithm using the gradient descent algorithm.
We will try to find a linear relationship between house sizes and house prices.

RELATIONSHIP BETWEEN COST FUNCTION AND NUMBER OF ITERATIONS

As the number of iterations increases, the value of the cost function usually decreases. This means that the model is trained and the parameters are optimized. At the beginning of the training process, the cost can be high because the model's predictions are often poor. However, as the iterations progress, the model learns and its predictions improve.

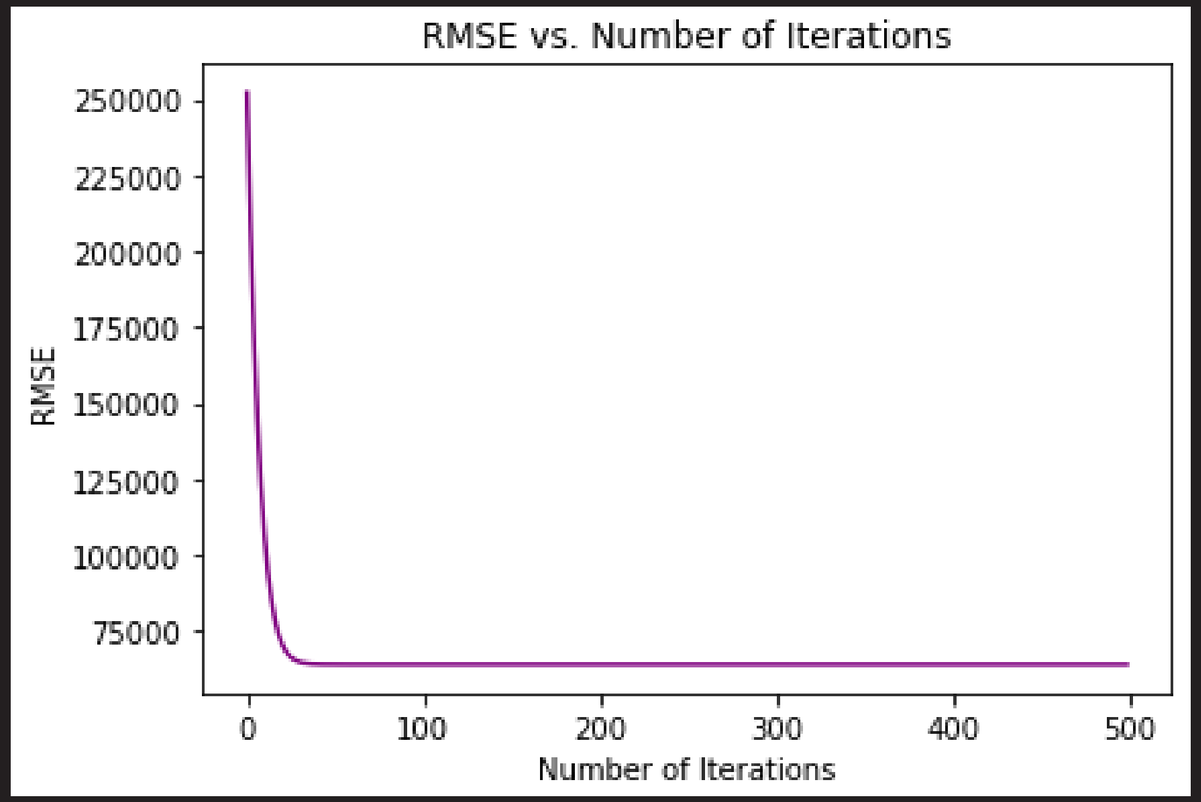
However, the cost function drops up to a certain point. Eventually, factors such as the learning rate or model complexity can have an impact. If the cost function does not decrease sufficiently or, conversely, increases, this may indicate that the model may be facing problems such as overfitting or underfitting.



RELATIONSHIP BETWEEN RMSE VALUES AND NUMBER OF ITERATIONS

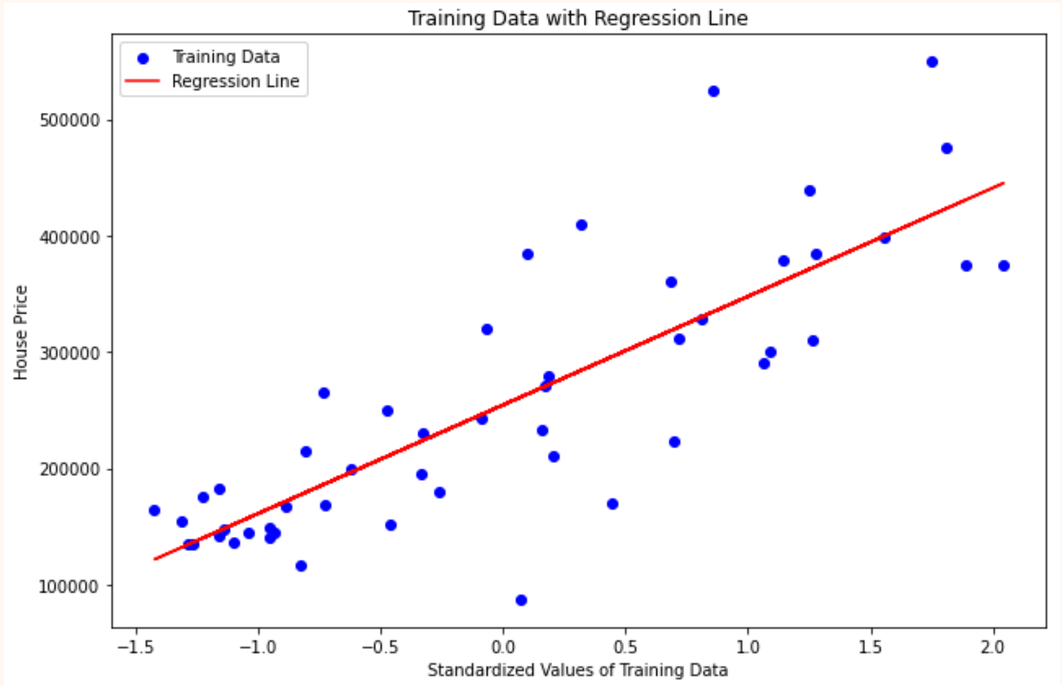
As the number of iterations increases, the RMSE value is generally expected to decrease. This indicates that the model starts to make better predictions during the learning process. At the beginning of the training process, the RMSE may be high because the model's predictions are usually bad. However, as the iterations progress, the model learns better and the RMSE value decreases.

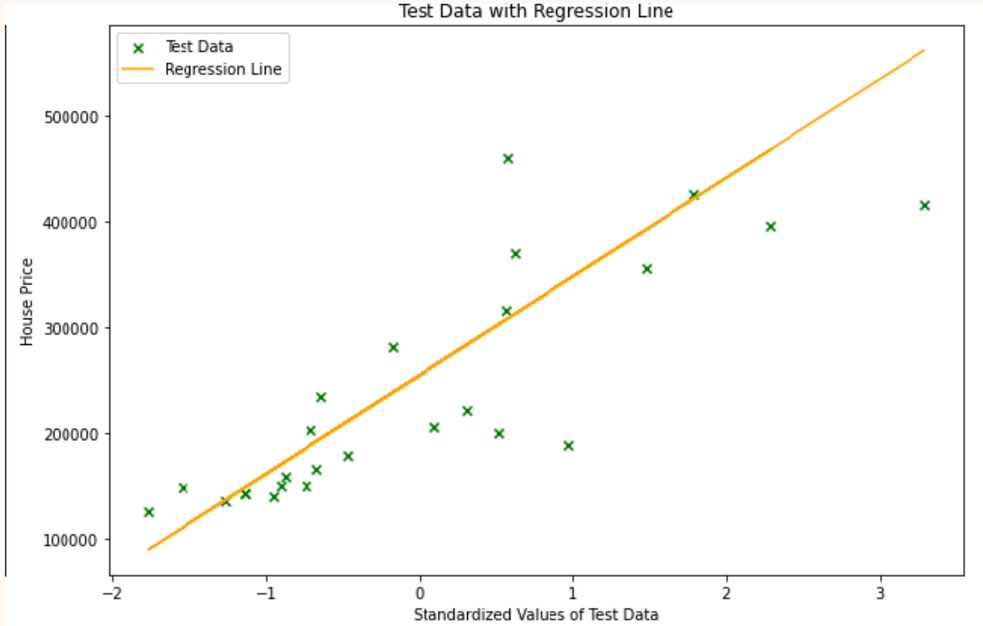
However, there may be cases where the RMSE value does not decrease or increases after a certain point. If the RMSE starts to increase after a certain point, this may be a sign of overfitting. As the model overfits the training data, its performance on general data may decrease.



REGRESSION LINES OF TRAINING AND TEST DATAS

The model is trained on training data. Therefore, the regression line usually fits the training data better. The closeness between the training data and the regression line indicates that the model provides a good fit on this data.
The distance between the distribution of the training data on the scatter plot and the regression line indicates the success of the model in the training process. If there is a lot of dispersion (width of the distribution) on the training data and it is located close to the regression line, this indicates that the model has learned the data well.

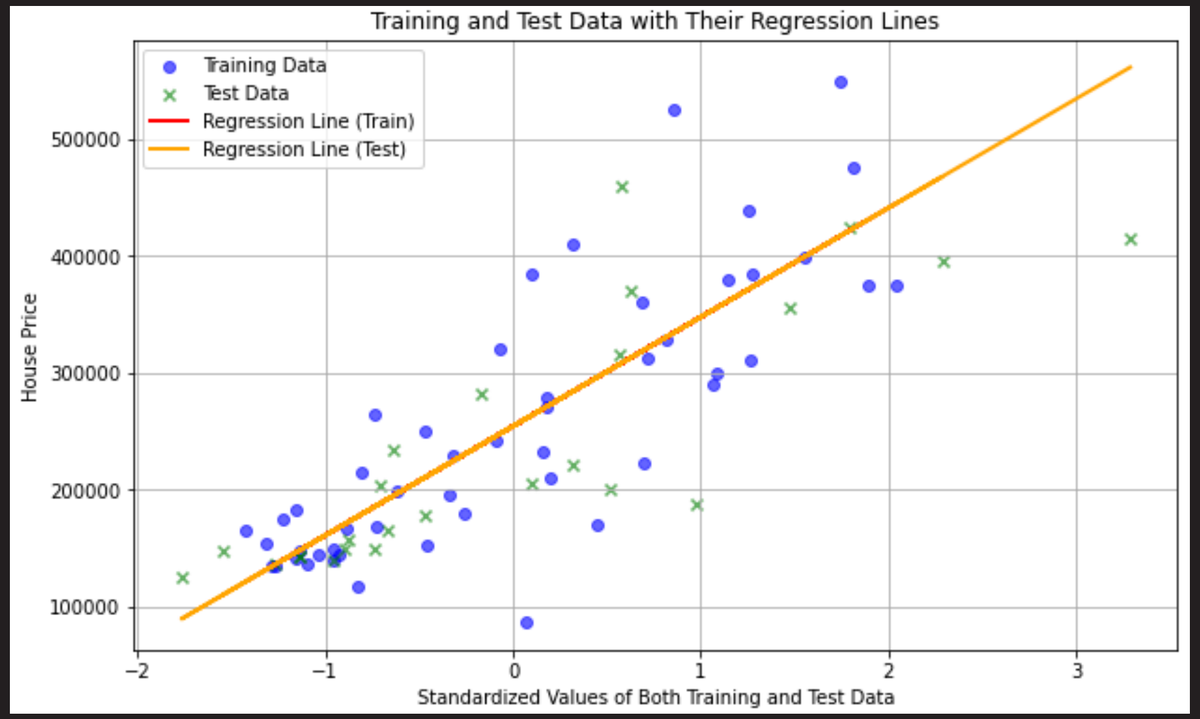




Test data consists of data that the model has not seen before. Therefore, the model's performance on this data indicates its generalization ability.
The distance between the regression line on the test data and the data points indicates the generalization ability of the model. If there is a large distance between the test data and the regression line, this indicates that the model does not perform well enough on this data and its generalization ability is low.

LINEAR SOLUTION LINE FOR TRAINING AND TEST DATA

Training Data: The training data, indicated by blue circles, represents the dataset on which the model was trained. This data was used in the model's learning process.
Test Data: The test data, denoted by green “x”, represents a separate dataset used to evaluate the overall performance of the model.



In general, the graph shows that the model performs well on the training data and is consistent with the test data. In this case, we can say that the model is a good predictor.

PARAMETERS' COMPARISON

Finally, I compared the w parameters obtained by applying the Gradient Descent algorithm with the parameters obtained with the Closed Form Solution, i.e. the Normal solution method.

w[0] - Gradient Descent = **254449.99999999988**

w[0] - Normal Solution = **254450.0**

w[1] - Gradient Descent = **93308.92010610425**

w[0] - Normal Solution = **93308.92010610428**