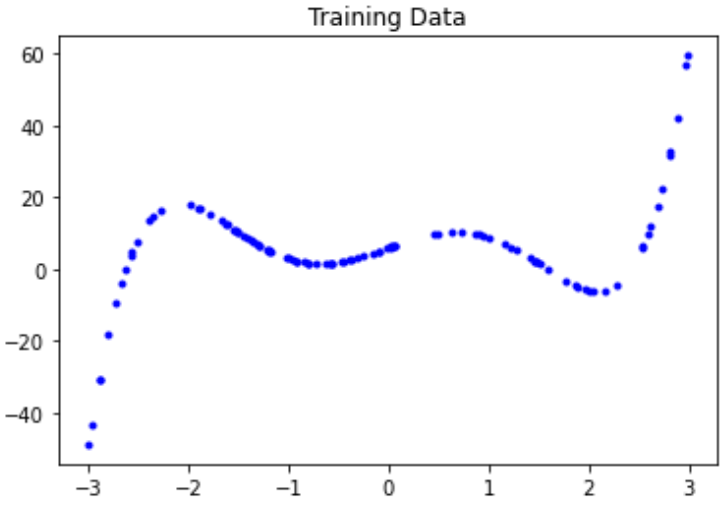
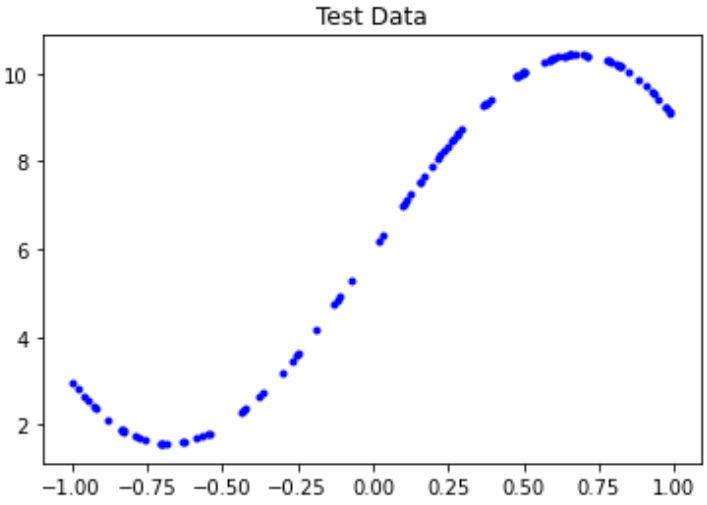
CS 5783 Assignment 1 – Report

S M Rakibur Rahman

Question 1:





a. Is the relationship linear?

From the visualization, it is clear that the relationship between variables is not linear.

b. Do you need feature engineering to add any non-linearity?

Yes, we do need to perform feature engineering.

i. If so, how can you engineer these features?

I can engineer these features using basis functions

ii. What are some functions that you can try?

From visualization, it looks like a polynomial basis function will be suitable. Certainly 1st order and 2nd order polynomial (Straight line) won't be enough. By analyzing the data, it appears that I will need polynomial of order 3 or more.

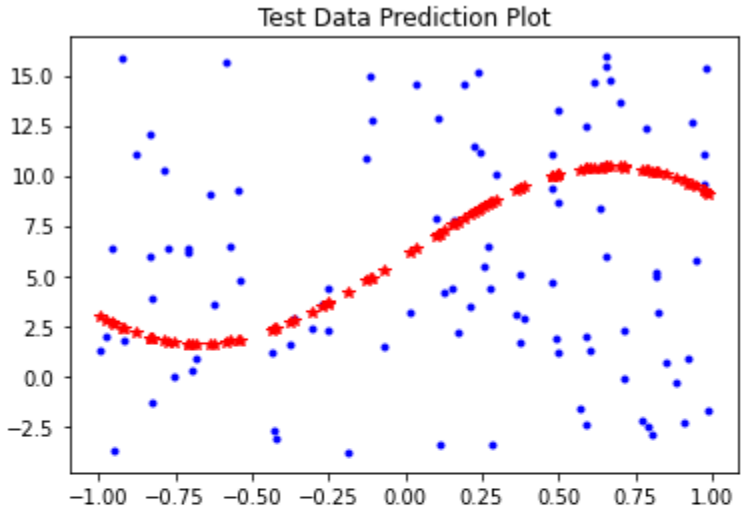
I have used normal equations.

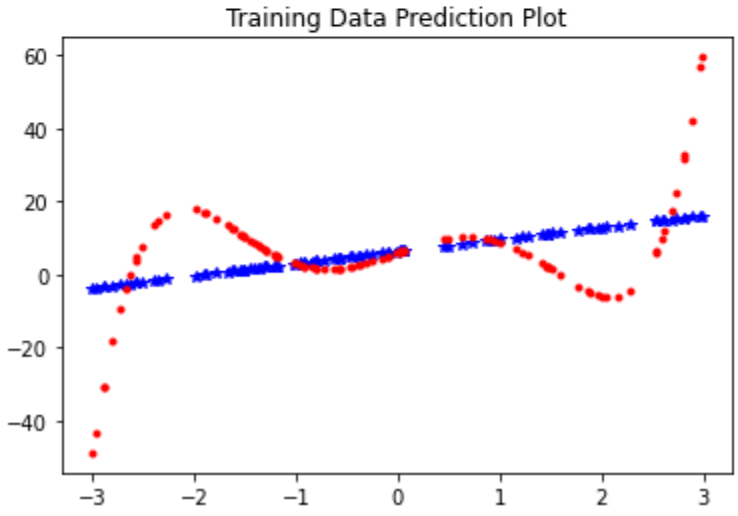
1. Plot each of them individually to verify!

Below are the plots

3-degree polynomial:

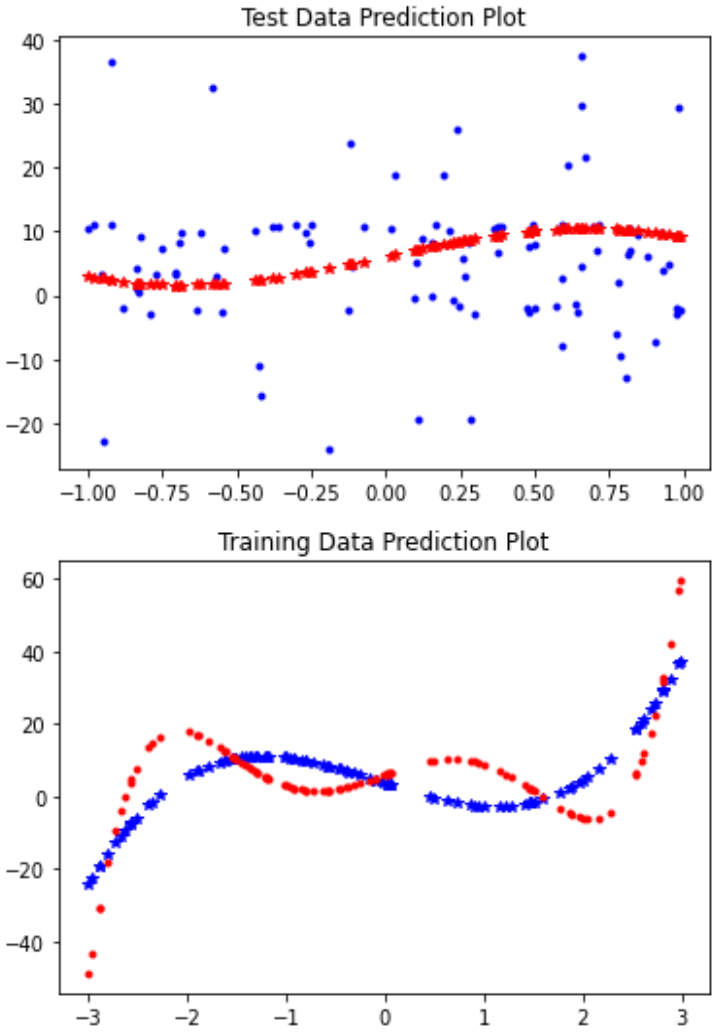
```
Average test data prediction error =  [41.70193992]
Average training data prediction error =  169.33131794384346
Text(0.5, 1.0, 'Training Data Prediction Plot')
```





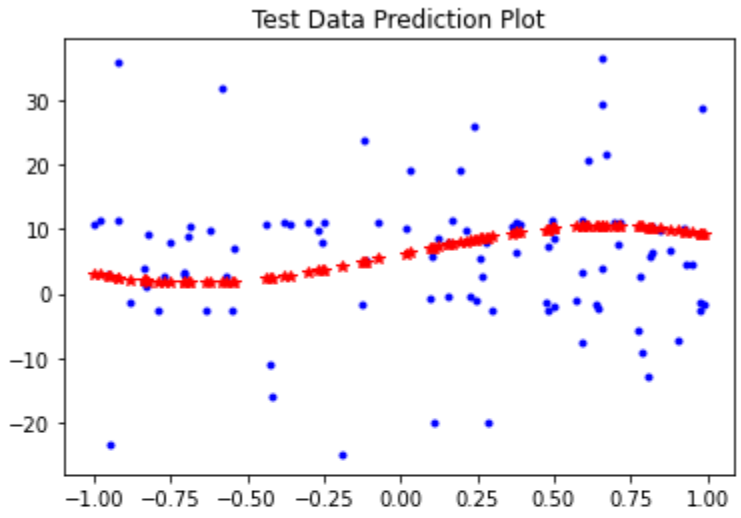
4 degree polynomial:

```
Average test data prediction error =  [131.41526738]
Average training data prediction error =  81.07774145765362
Text(0.5, 1.0, 'Training Data Prediction Plot')
```



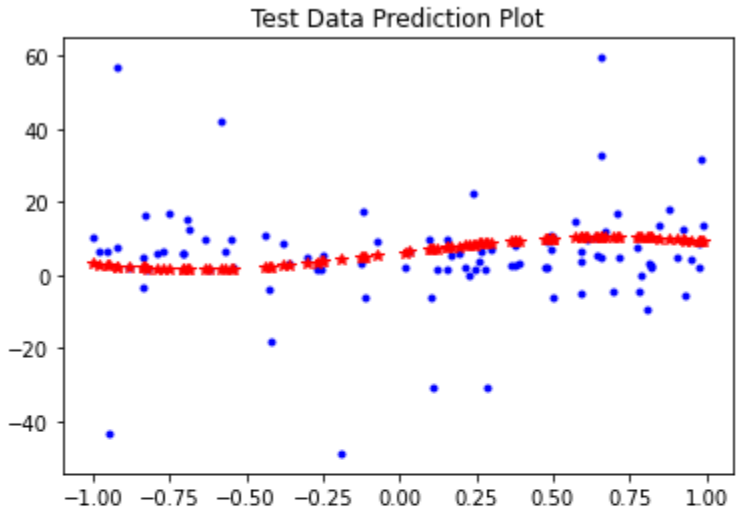
5 degree polynomial:

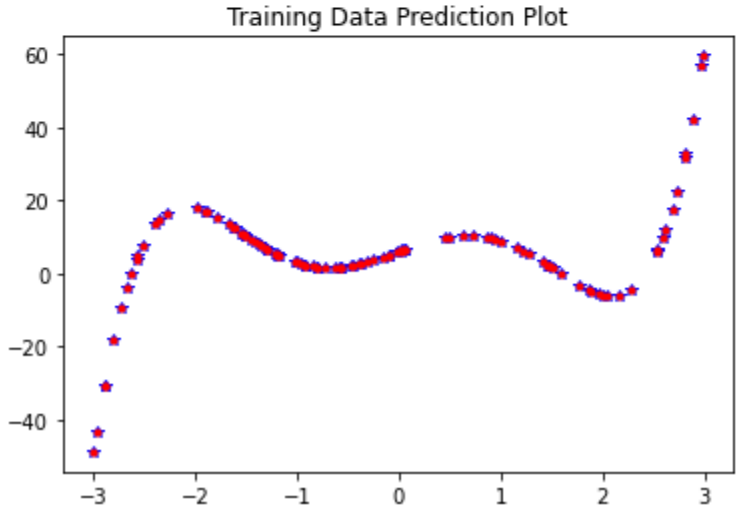
```
Average test data prediction error =  [131.15742946]
Average training data prediction error =  80.90354598712581
Text(0.5, 1.0, 'Training Data Prediction Plot')
```



6 degree polynomial:

```
Average test data prediction error =  [211.42515053]
Average training data prediction error =  7.461420903991277e-26
Text(0.5, 1.0, 'Training Data Prediction Plot')
```





Question 2:

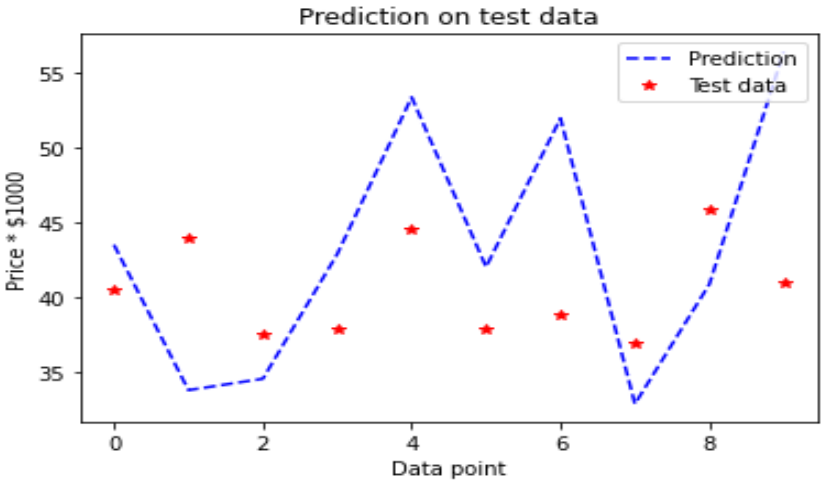
After loading the data from the csv file, I extracted the input x values (11 features excluding the house ID) and output y-values (prices). Then I split the data into training and test samples. First 18 samples are kept for training and remaining 10 samples are for testing. I used Normal Equations to calculate my model parameters theta.

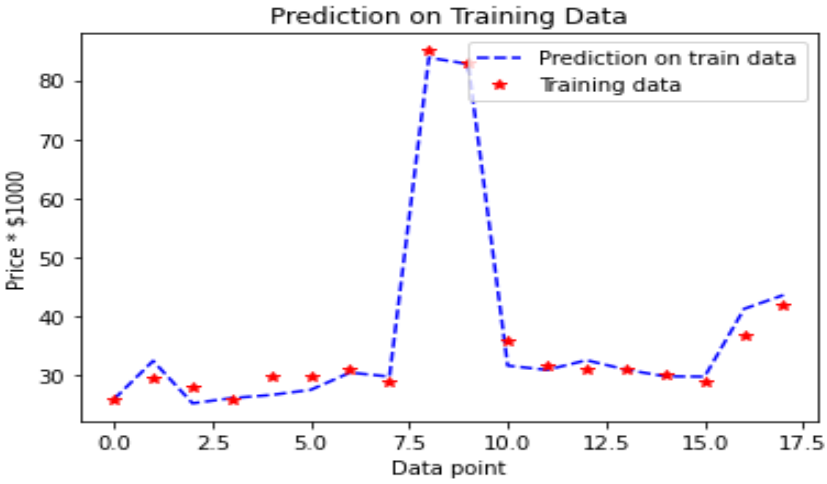
$$\Theta = (X^T X)^{-1} X^T y$$

In this case, theta is a vector of 11 parameters corresponding to 11 features. After calculating theta, I used the basis function equation to calculate my predictions.

$$y(x,\mathrm{w}) = \sum_{j=0}^{M-1} w_j \phi_j(x) = \mathrm{w}^T \phi(x)$$

, I got the following prediction plots.





Question 1: What is the average least squares error for the given data using your simple linear regression model?

Answer:

Average prediction error = 68.5098

Average training data prediction error = 4.2843

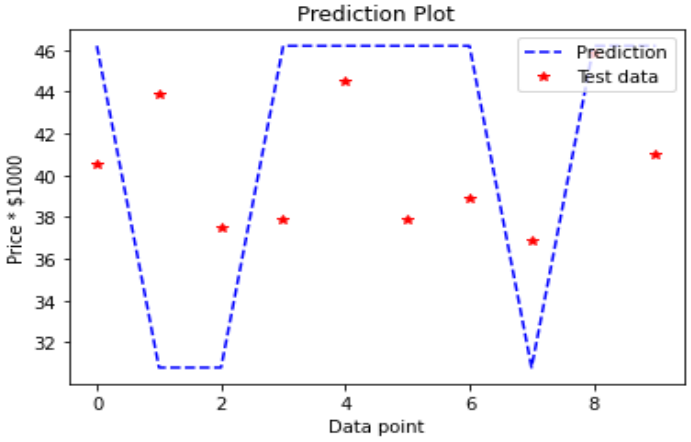
Question 2: Which factor has the most effect on the final value? How do you know this? Can you use only this feature to predict the price?

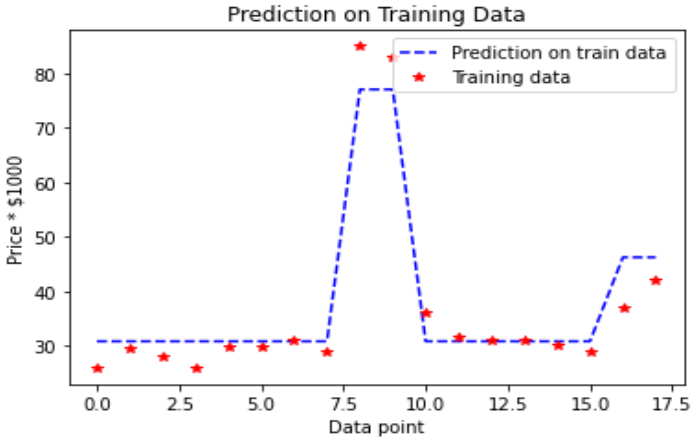
Answer: No of Bathroom has the most effect on the final pricing prediction. This can be understood from the value of the theta. The theta values as following,

Theta values are = [1.57, 25.74, 0.62, 0.21, 1.73, -5.73, 5.83, 0.12, -0.99, 2.93, 5.18]

The second theta value is for the no of bathrooms, which is the highest absolute value.

Using only this feature we can predict the house pricing; however, this doesn't lead to very good prediction. Especially if we look at the prediction plot. Following plots are for house prediction based on only the bathroom feature.





Average prediction error = 50.71701

Average training data prediction error = 16.4644

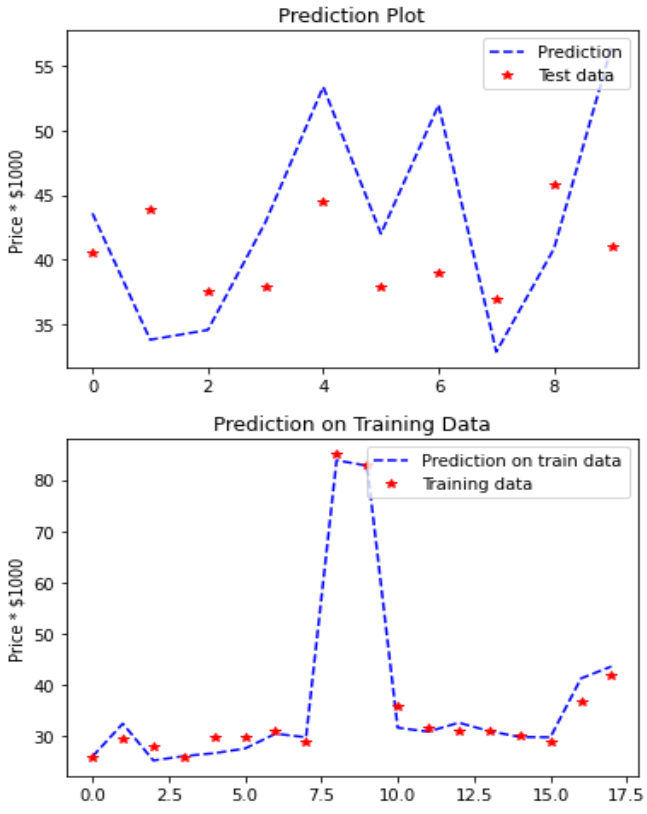
Question 3: Which factor has the least effect on the final value? How do you know this? What effect does removing this feature have on the performance?

Answer: Land Area feature has the least effect on the final pricing prediction. This is clear from the value of theta.

Theta values are = [1.57, 25.74, 0.62, 0.21, 1.73, -5.73, 5.83, 0.12, -0.99, 2.93, 5.18].

The fourth theta value is for the Land Area feature, which is the lowest absolute value.

Removing this feature doesn't have much effect on the prediction, as can be seen from the prediction plots. Following are the prediction plots after removing the land area feature. If we compare these with the original prediction plots, we can see that these plots are almost same as the original predication plots. The average prediction errors are also similar. Which verifies that the land area feature has the least effect.



Average prediction error = 69.5219

Average training data prediction error = 4.2851