Notebook

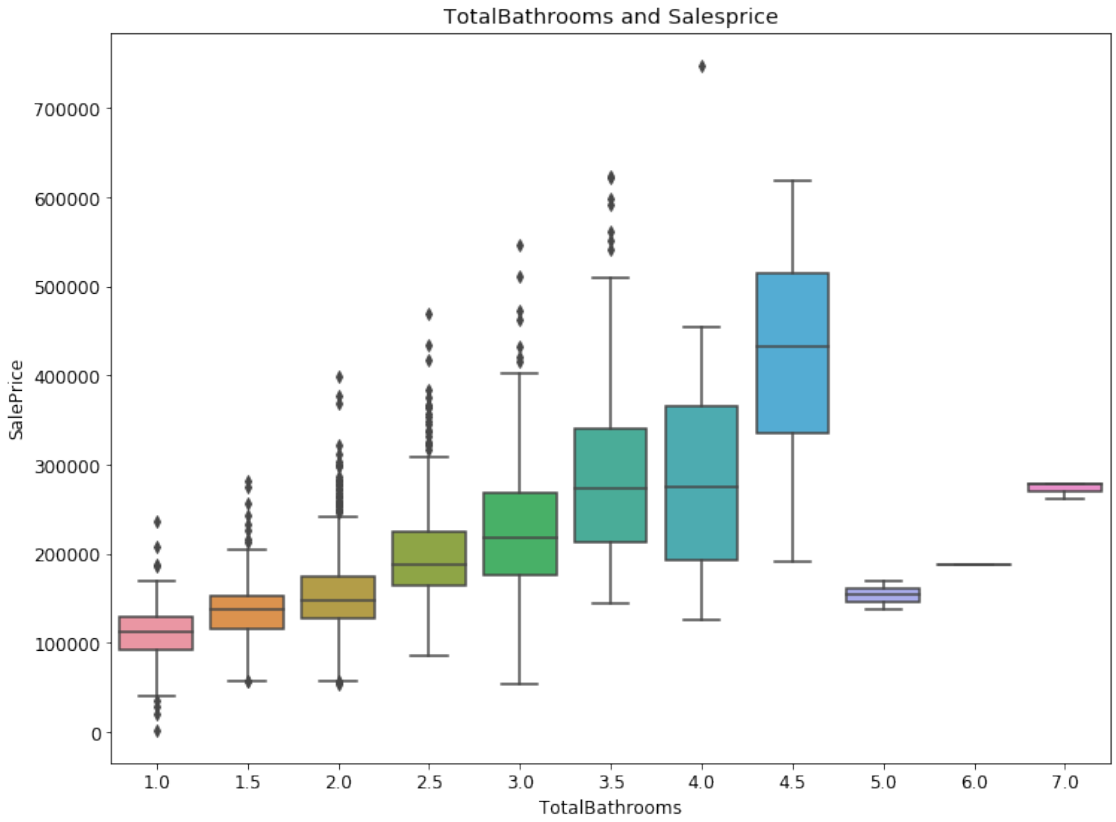
July 26, 2019

0.1 Question 5

Create a visualization that clearly and succintly shows that `TotalBathrooms` is associated with `SalePrice`. Your visualization should avoid overplotting.

```
In [235]: sns.boxplot(x="TotalBathrooms", y="SalePrice", data=training_data)
          plt.title('TotalBathrooms and Salesprice')
```

```
Out[235]: Text(0.5, 1.0, 'TotalBathrooms and Salesprice')
```



Ideally, we would see a horizontal line of points at 0 (perfect prediction!). The next best thing would be a homogenous set of points centered at 0.

But alas, our simple model is probably too simple. The most expensive homes are systematically more expensive than our prediction.

0.2 Question 8d

What changes could you make to your linear model to improve its accuracy and lower the test error? Suggest at least two things you could try in the cell below, and carefully explain how each change could potentially improve your model's accuracy.

1. Increase the model complexity by adding a useful feature while guaranteeing that it decreases bias more than it increases variance. Adding a useful feature to the data reduces bias and increases model variance, since models with many parameters have many possible combinations of parameters and therefore have higher variance than models with few parameters. However, as complexity of the model goes up, the test error would first decrease then increase as the increased model variance outweighs the decreased model bias. Therefore, we need to strike a balance between model bias and variance.

2. Cross validation We can implement k-fold cross validation on our training data. We can split the training data into K equal sized partitions, using K-1 splits to train, last split as validation set. We would repeat this for K times and come up with average of K errors, the validation error. Finally we can pick the model with the lowest validation error. The repeated estimates can mitigate the variance of splits and help preventing overfitting of the training data. This can help improve the model's accuracy in predicting in our test data.

3. Regularization If we add more useful features into the model, we can use regularizaition to penalize the large weighted features, in order to decrease the variance of the model.