Notebook

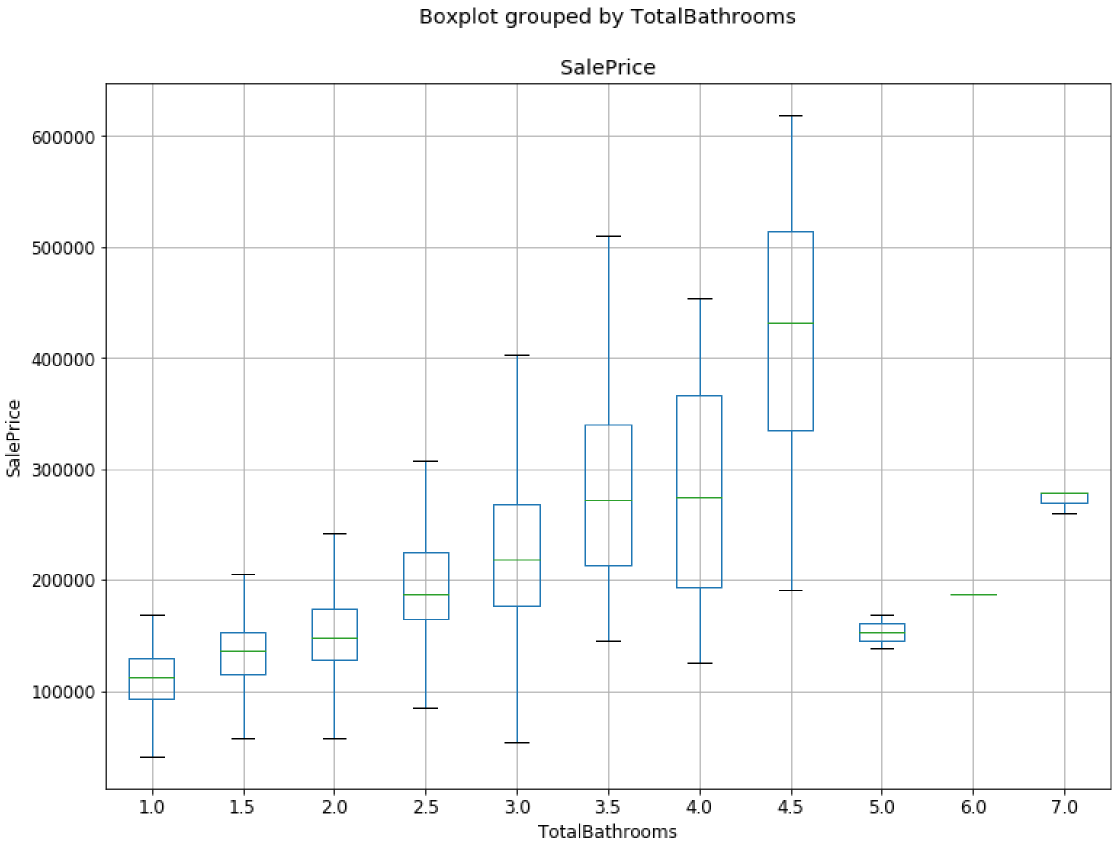
April 8, 2020

0.1 Question 5

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

```
In [17]: training_data.boxplot(column = 'SalePrice', by = 'TotalBathrooms', showfliers = False)
         plt.ylabel('SalePrice', Fontsize = 12)
```

```
Out[17]: Text(0, 0.5, 'SalePrice')
```



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.

First we can think of putting another variable. We may need to verify it through graphs, but I think it will be able to make more accurate predictions by putting in other variables that may affect the price. For example, variables like 'Year_Built' or 'Kitchen_Qual' can be expected to affect the price, and by putting these variables into the predict model, we would be able to make more accurate predictions.

Next, we would be able to make more accurate predictions by adjusting the slope values of the linear model. To optimize the slope value, we can use 'rsme' function. By applying multiple slope values and using the rsme function to get the slope value with the least error, we can make better predictions from the given data.