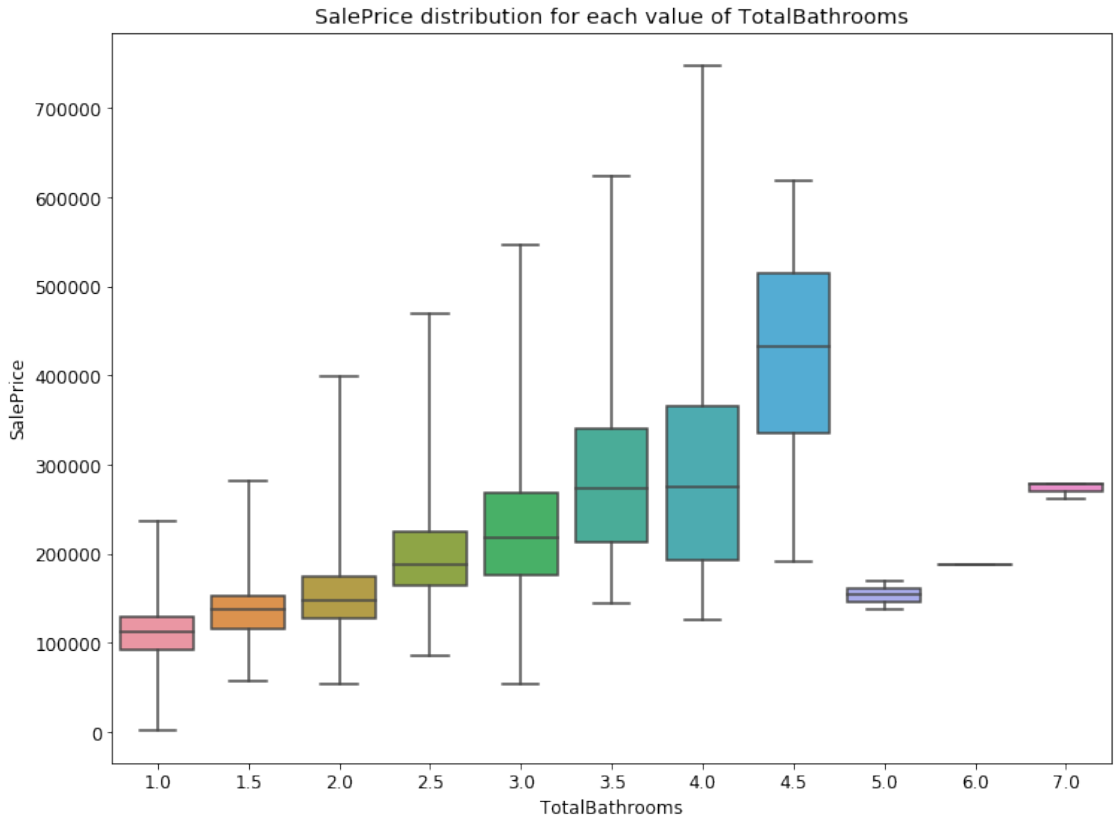
0.1 Question 2b

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

```
In [15]: sns.boxplot(x = 'TotalBathrooms', y = 'SalePrice', data = training_data_with_bathrooms, whis =
         plt.title('SalePrice distribution for each value of TotalBathrooms')
```

```
Out[15]: Text(0.5, 1.0, 'SalePrice distribution for each value of TotalBathrooms')
```



0.2 Question 5d

What changes could you make to your linear model to improve its accuracy and lower the validation 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.

We could add more features that are likely to correlate with expensive houses, or encode the neighborhoods as features, since houses from the same neighborhood might be more likely to be closer in price. We could also use regularization to improve the accuracy and lower the validation error.

0.3 Question 6a

Based on the plot above, what can be said about the relationship between the houses' sale prices and their neighborhoods?

There is quite some variation in prices across neighborhoods it seems. The amount of data available is not the same among neighborhoods. For example, North Ameshas 299 observations in the training data while Green Hill has just 2 observations.

0.4 Question 8a

Although the fireplace quality variable that we explored in Question 2 has six categories, only five of these categories' indicator variables are included in our model. Is this a mistake, or is it done intentionally? Why?

If we want to calculate the least square estimate of our coefficients, then the design matrix has to be full rank. If each of the fireplace quality variables's six categories were represented in the model, this would not be the case