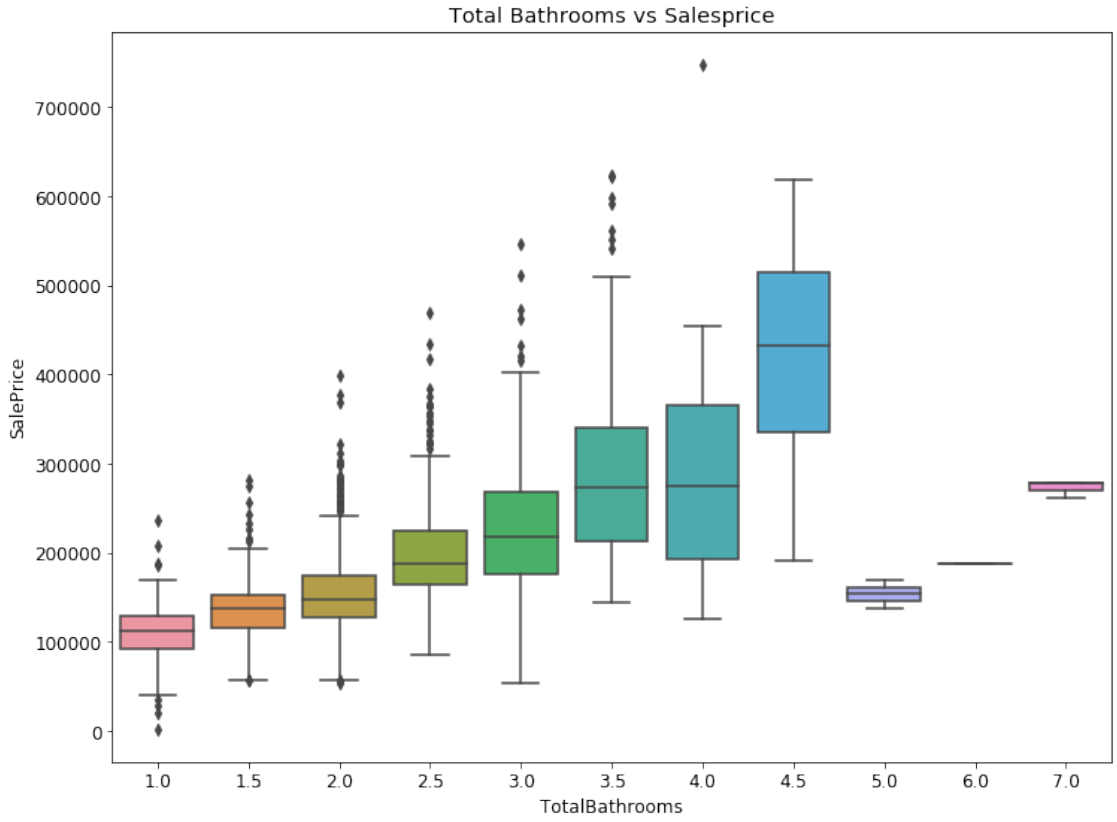
0.1 Question 2b

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)
         plt.title('Total Bathrooms vs Salesprice')
```

```
Out[15]: Text(0.5, 1.0, 'Total Bathrooms vs Salesprice')
```



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.

1. Increasing the size of training that we can have better learning about the relationship in the data.
2. Increasing the model complexity is also a good way, we can add more feature to improve its accuracy and lower the validation error. However, if the model is too complex, it will course overfit.
3. Using Cross-Validation.

0.3 Question 6a

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

I did not see much relationships between the sales price and the number of neighbours. Some houses with a high number of neighbours are below average, but some houses with a high number of neighbours are above average. The reverse is also true. Some housing prices with a low number of neighbours are below average, and some are above average. Therefore, there are no strong relationships between the sales price and the number of neighbours.

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?

It is done intentionally because 'Fireplace_Qu=Average' is not an indicator variable, therefore we can drop that to reduce redundancy and make our model is linear independent.