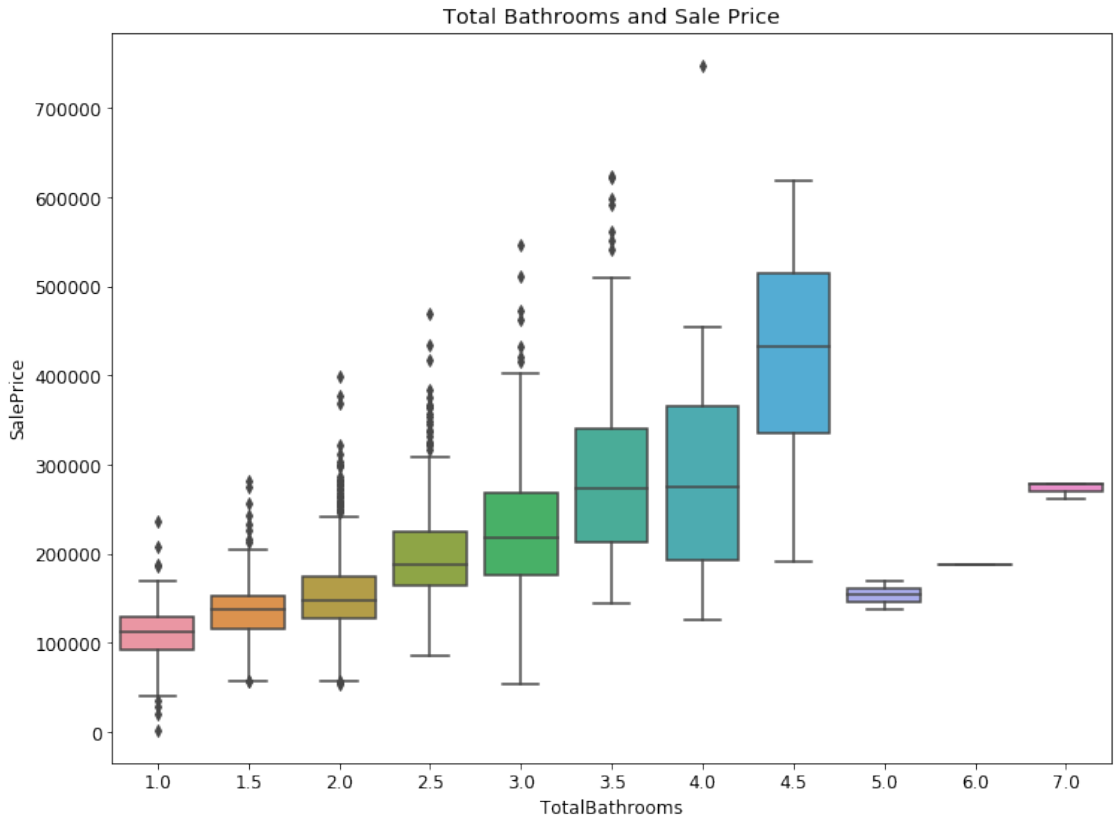
0.1 Question 2b

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

```
In [15]: sns.boxplot(training_data_with_bathrooms['TotalBathrooms'],training_data_with_bathrooms['SaleP
         plt.title('Total Bathrooms and Sale Price')
```

```
Out[15]: Text(0.5, 1.0, 'Total Bathrooms and Sale Price')
```



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.

First of all we can try to increase the complexity of our model by adding releveant features. Adding additional useful features to our data would reduce bias and improve its accuracy, which decreasing validation error up to a certain point. The second thing that we could try is cross validation. Implementing a k-fold cross validation allows us to repeat the splitting of our train and validation set and we can then pick the model with the lowest 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?

Based on the plot above it is difficult to observe a clear distinct relationship between the sale prices and their neighborhoods. It is difficult to discern the differences between the neighbors and sale prices as some neighbors with high counts had high prices while others with high counts had low prices.

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?

This is done intentionally because having the sixth category is redundant, removing it ensures full rank and invertibility.