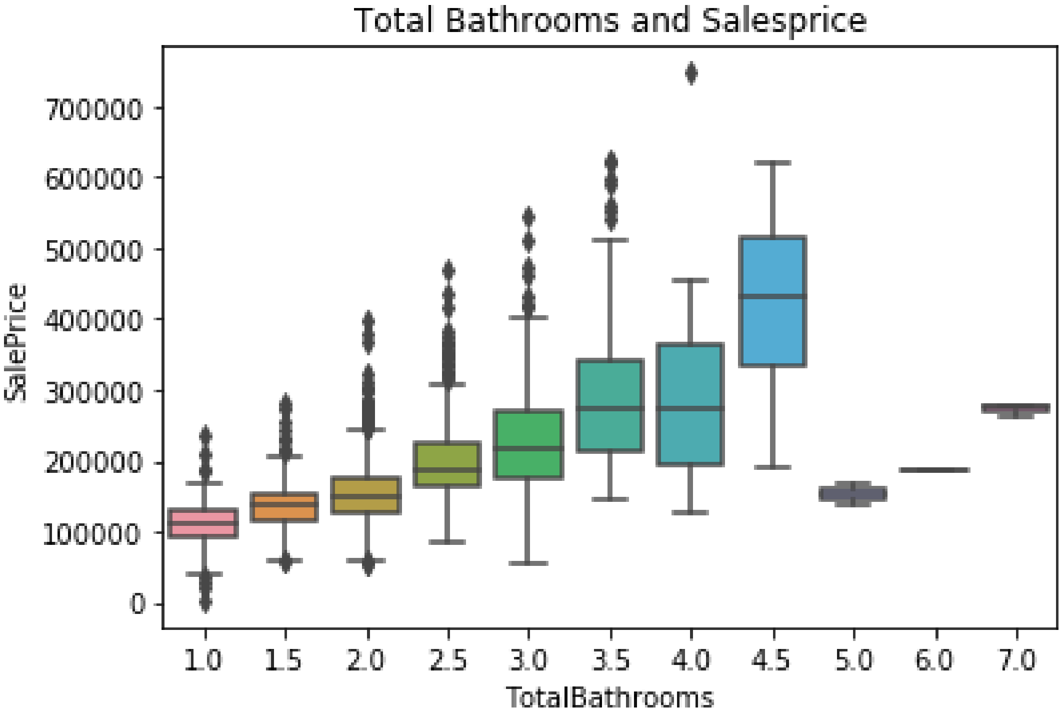
## 0.1 Question 2b

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

```
Out[15]: Text(0.5, 1.0, 'Total Bathrooms and 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.

I would try increasing the model complexity by adding a relevant feature. This could potentially improve my model accuracy because it reduces bias and increases model variance. This is because more parameters would mean more possible combinations of parameters and thus increased variance. The catch is that test error might incerase if the increase in model variance surpasses the decrease in bias. This is when regularization (penalizing larger weighted features) would be useful in decreasing variance.

I would also try cross validation, spliting the training data into partitions and using one split as the validation set. Repeating this a number of times to find the validation error or the average of errors and selecting the model with the lowest error. This process helps prevent overfitting and help with the variance.

## 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 not a clear relationship between the house sale price and their neighborhoods. Neighborhoods with high counts have mean saleprices that are both high and low and that's also the case for neighborhoods with low counts. Meaning, there is a lot of variation in prices across the different neighborhoods. Additionally, the count is different for every neighborhood, some neighborhoods have a much higher count than others. It's hard to establish a clear relationship based off of that.

## 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 the variable that is dropped is redundant and thus would make the columns linearly dependent. Dropping one of the indicator variables would ensure full rank and linear independence for invertability.