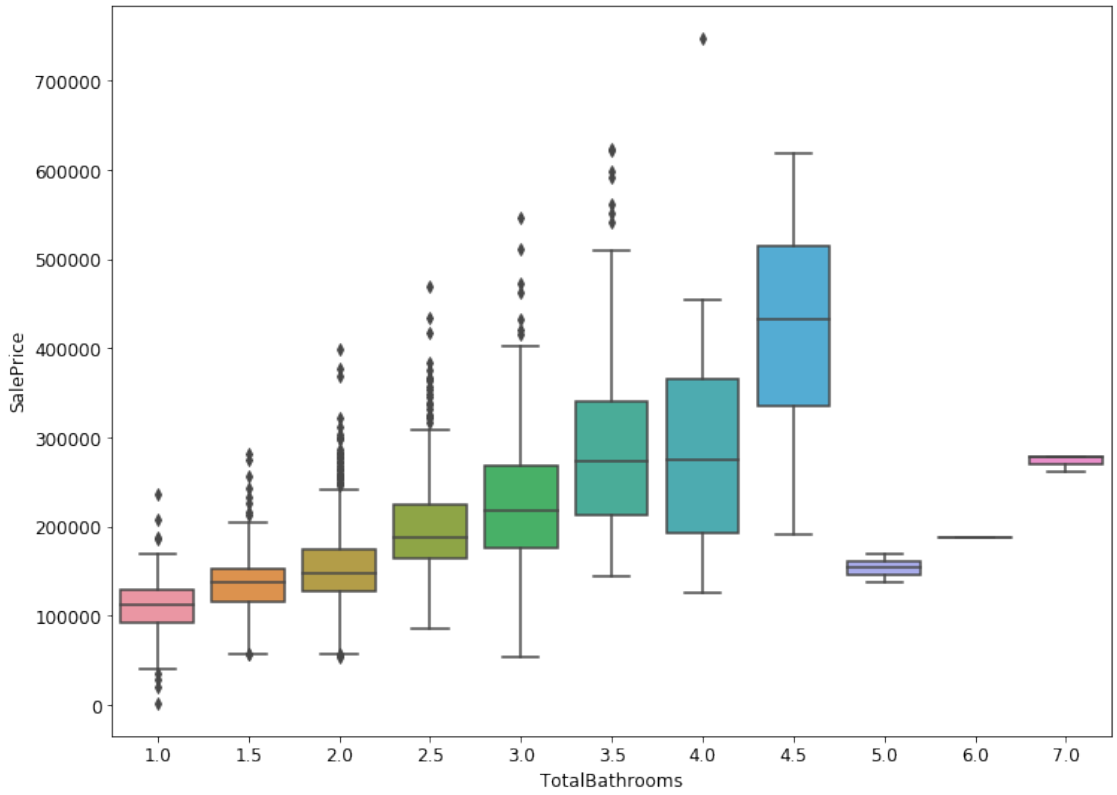
0.1 Question 2b

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

```
In [171]: sns.boxplot(data=training_data_with_bathrooms, x="TotalBathrooms", y="SalePrice");
```



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 can group neighborhoods into sections according to their median SalePrice. We could also turn columns such as Year_Built and Overall_Qual into categorical columns since they are nominal features and then use one_hot_encoding on the columns to ensure that the machine does not predict more recently built houses as more expensive.

Another thing we could do is regularize the columns which would ensure that the columns dealing with larger values such are Lot_Area do not have a greater effect on the prediction.

0.3 Question 6a

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

1. The distribution of houses across the neighborhoods is not even and neither is the distribution of the sale prices of the houses.
2. NAmes has the highest count but the distribution of the houses does not show a large interquartile range.
3. I would like to make the claim that the neighborhoods with a lower count have a higher median SalePrice but due to the lack of sufficient data on all neighborhood to make that claim.

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?

The dropping of one category has been done intentionally. This was done to ensure a full rank for calculating the ordinary least squares estimate of the coefficients. Dropping the sixth category does not mean we lose that information. If we are given 5 out the 6 columns, we can easily calculate the sixth column, But having the sixth column would mean our design matrix would not be full rank.