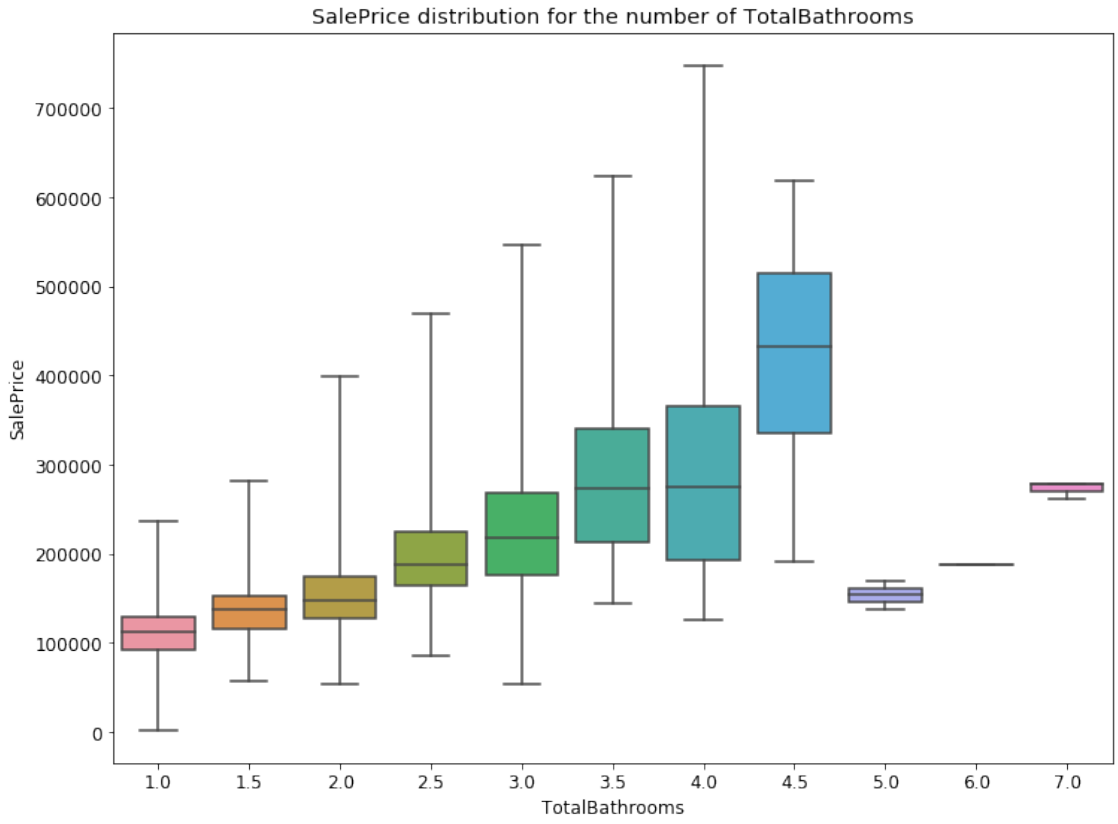
0.1 Question 2b

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

```
In [303]: sns.boxplot(x = 'TotalBathrooms', y = 'SalePrice', data = training_data, whis = 5);
          plt.title('SalePrice distribution for the number 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.

Firstly, we could take more, or a different combination of features that also influence SalePrice into account in X_train. This will give us a more realistic model that is more representative of reality, as there are many many factors that influence a house's sale price, like the zoning, Street (location), whether it has a pool or a fence, etc. Secondly, we could implement one-hot encoding into the model so that categorical/binary features can also be used to predict the SalePrice.

0.3 Question 6a

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

The neighborhood a house is in definitely has a significant impact on their sale price. It is clear that there is quite some variation in prices across neighborhoods. Moreover, theamount of data available is not uniformly distributed among neighborhoods. North Ames, for example,comprises almost 15% of the training data while Green Hill has a scant 2 observations in this data set.

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?

Average is pretty much the same as Fair, and as such isnt different enough of a variable to be taken into account.