Notebook

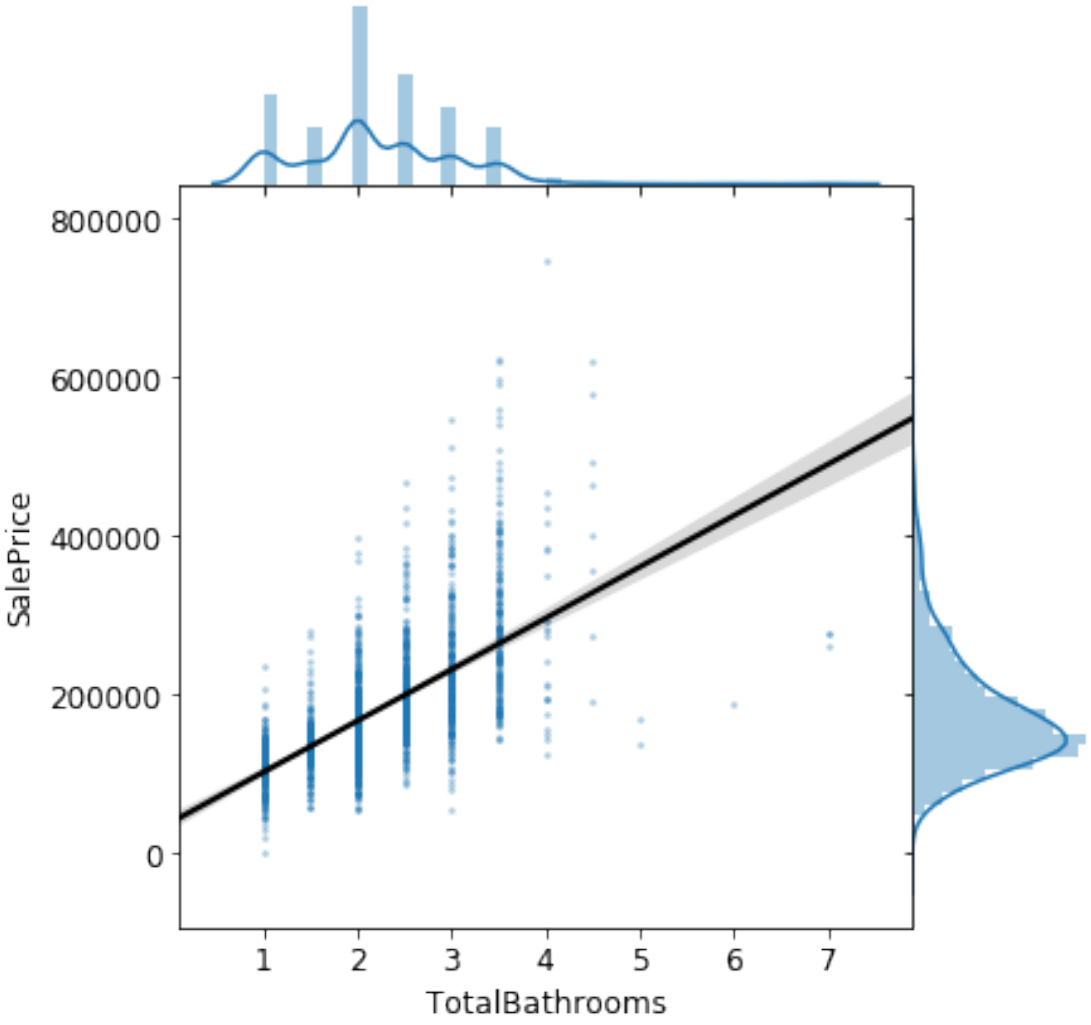
July 25, 2019

0.1 Question 5

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

```
In [76]: sns.jointplot(
             x='TotalBathrooms',
             y='SalePrice',
             data=training_data,
             stat_func=None,
             kind="reg",
             ratio=4,
             space=0,
             scatter_kws={
                 's': 3,
                 'alpha': 0.25
             },
             line_kws={
                 'color': 'black'
             }
         )
```

```
Out[76]: <seaborn.axisgrid.JointGrid at 0x7fee61ea3208>
```


800000
600000
400000
200000
0
SalePrice
1
2
3
4
5
6
7
TotalBathrooms

Ideally, we would see a horizontal line of points at 0 (perfect prediction!). The next best thing would be a homogenous set of points centered at 0.

But alas, our simple model is probably too simple. The most expensive homes are systematically more expensive than our prediction.

0.2 Question 8d

What changes could you make to your linear model to improve its accuracy and lower the test 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. Add more features, see how they can improve the accuracy, involve the features that contribute more on the accuracy improvement.
2. Increase the order of polynomial regression model may help, since if the model is too simple, we could make it more complex but not too complex. We can train several polynomial regression models with different orders, draw a graph of test error and pick the model with lowest test error.