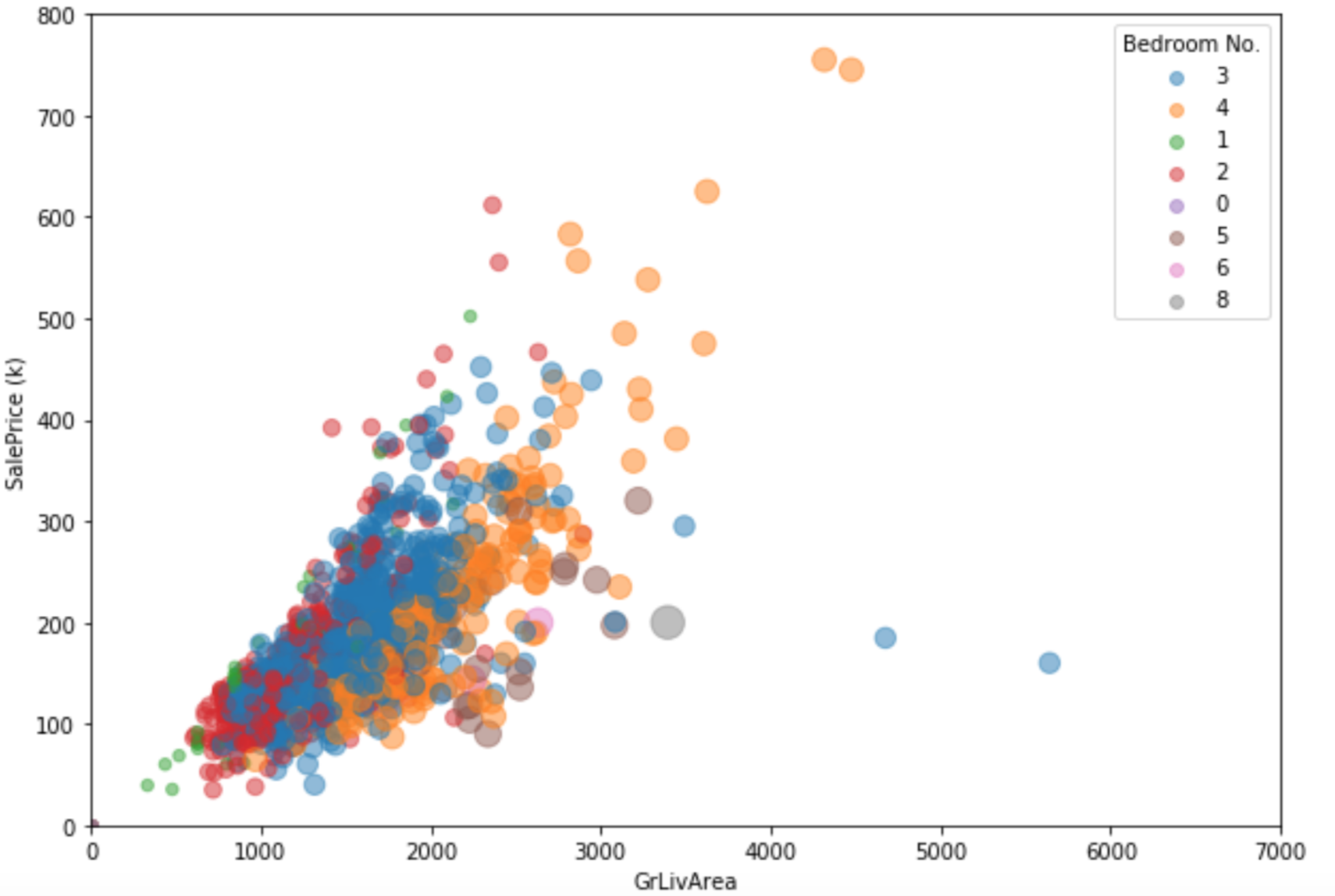
Capstone Project 1: Story Telling

## Capstone project 1: Housing prices advanced regression

### Data  from <https://www.kaggle.com/c/house-prices-advanced-regression-techniques>

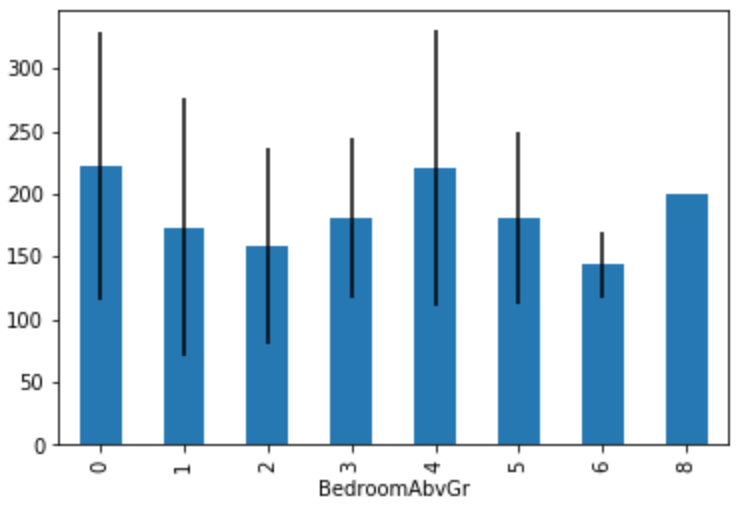
### Questions:

#### How is bedroom number affects Sale Price?

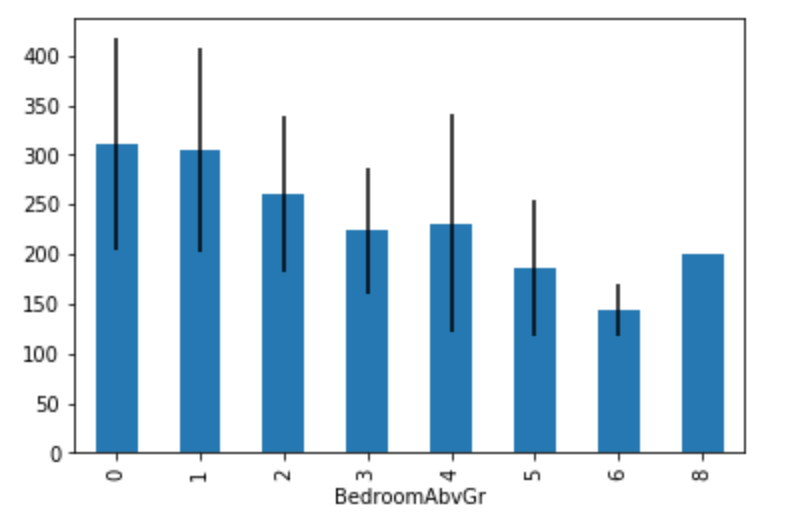


It looks like at at similar living area, the less the bedroom number is, the higher the price is, which is surprising.

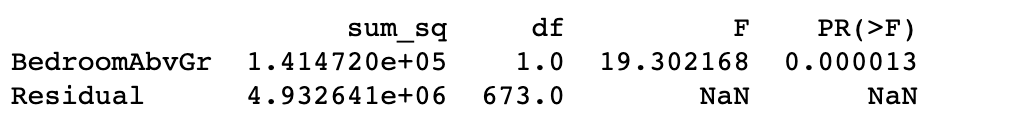
The mean Sale Price grouped with bedroom number for whole data sets:



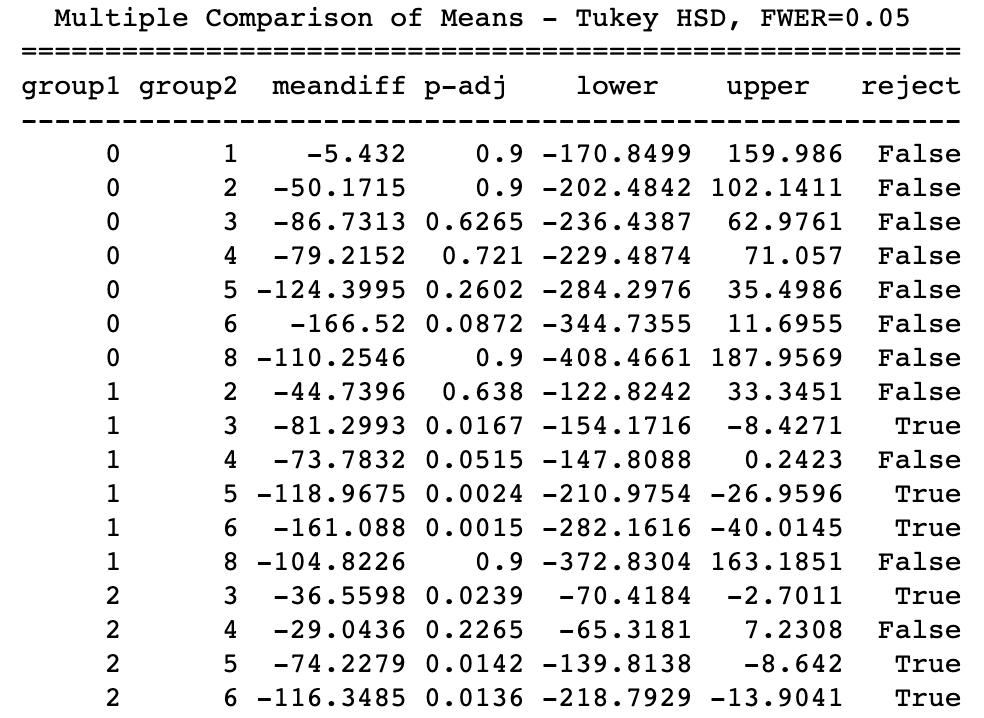
The mean Sale Price grouped with bedroom number for living area bigger than 1500:



It looks like when living area is bigger than 1500, bed room number is negatively correlated with Sale Price. Mean number of Sale Price at different bedroom groups are significantly different according to anova analysis.



And with tukey hsd post pairwise comparison, we saw the difference between each group of bedroom numbers.



Is that True?

My hypothesis is if bedroom number significantly affects SalePrice, then if we have a data set that only has randomized bedroom number and the other attributes are the same, then we still could expect to see a negative correlation between SalePrice and Living area.

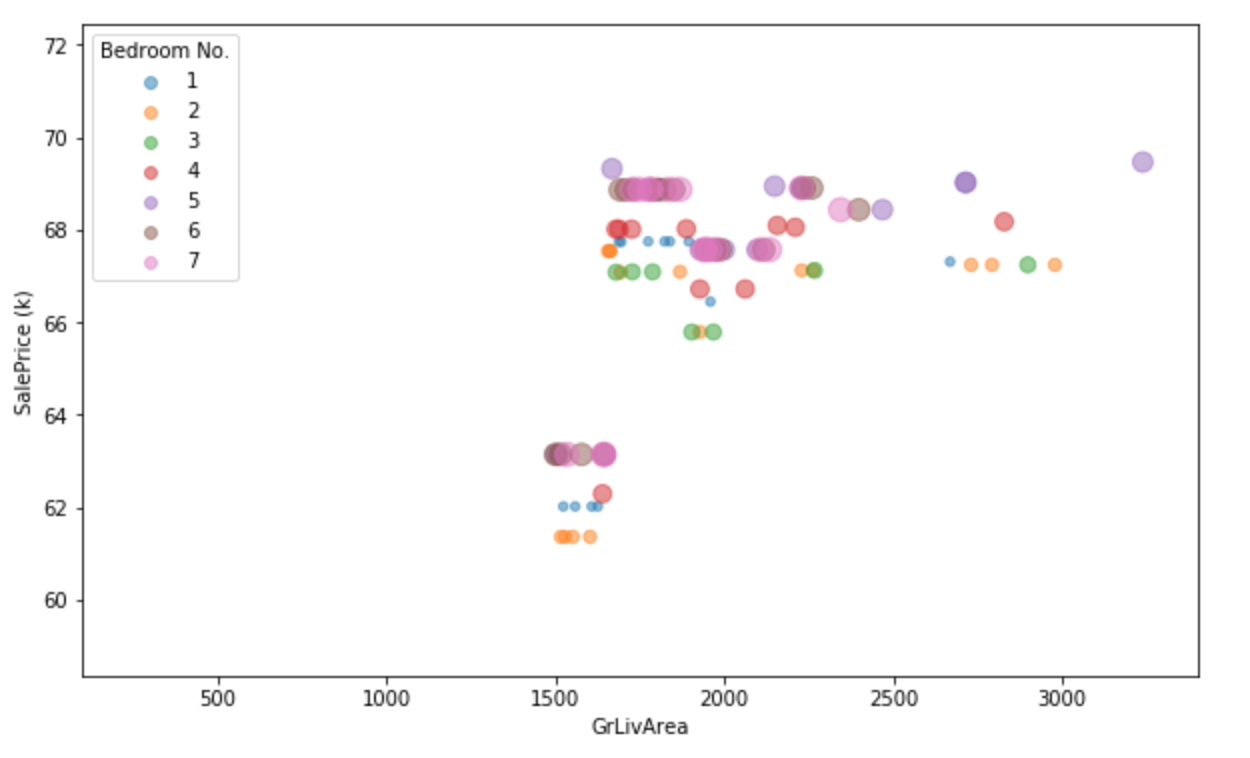
To test this hypothesis, we used simulation to randomly pick a house, and assign everything else as the same with this house, but assign the bedroom number randomly from 0 to 5 (let’s add some boundaries, to make sure bedroom number is reasonable by making it less that ‘living area/300’, which means a house with less than 1000sqrt could maximum expect 3 bedrooms instead of more, to make the bedroom number more realistic). Also, we keep living area as the same as the original data set, and use xgboost optimized model to predict the SalePrice using the new simulated data set.

## Experimenting:

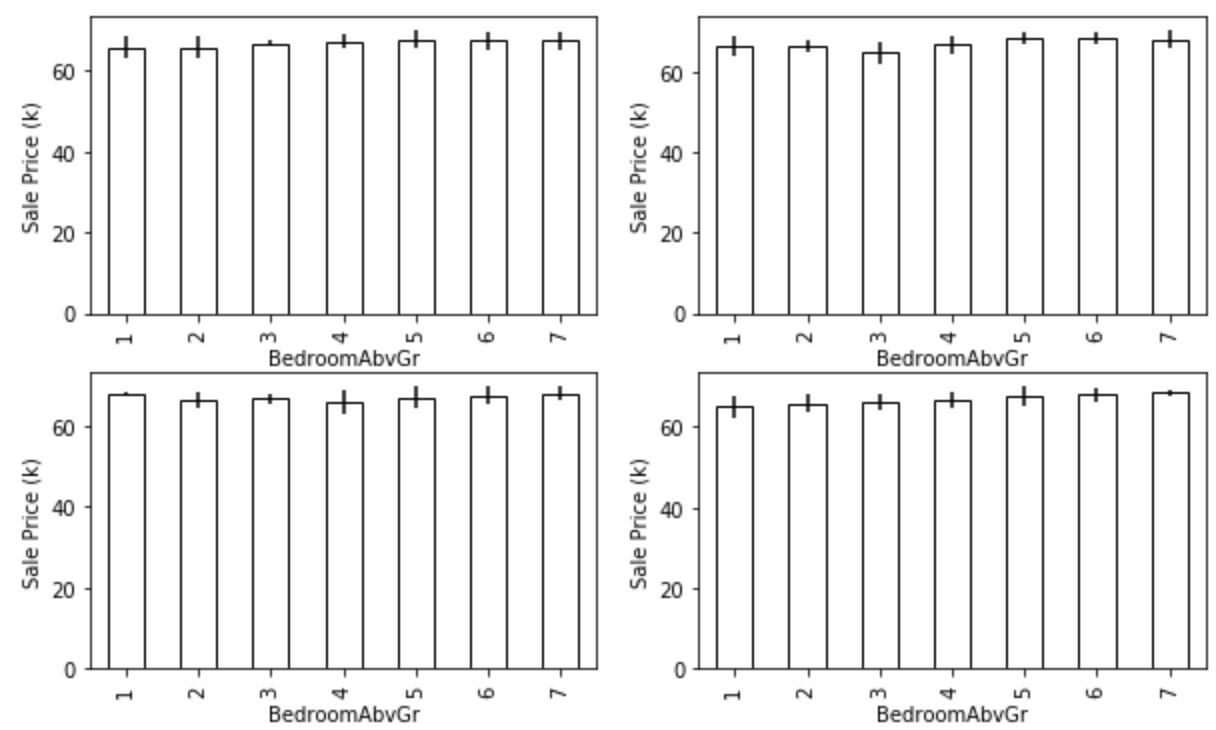
First, We choose the best model after grid search for best parameters for Xgboost model, and fit the training data with the best model.

Second, we do a simulation:

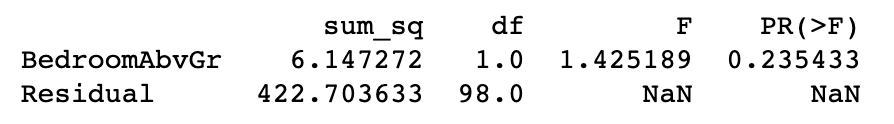
One example of the 4 simulations, others are very similar.



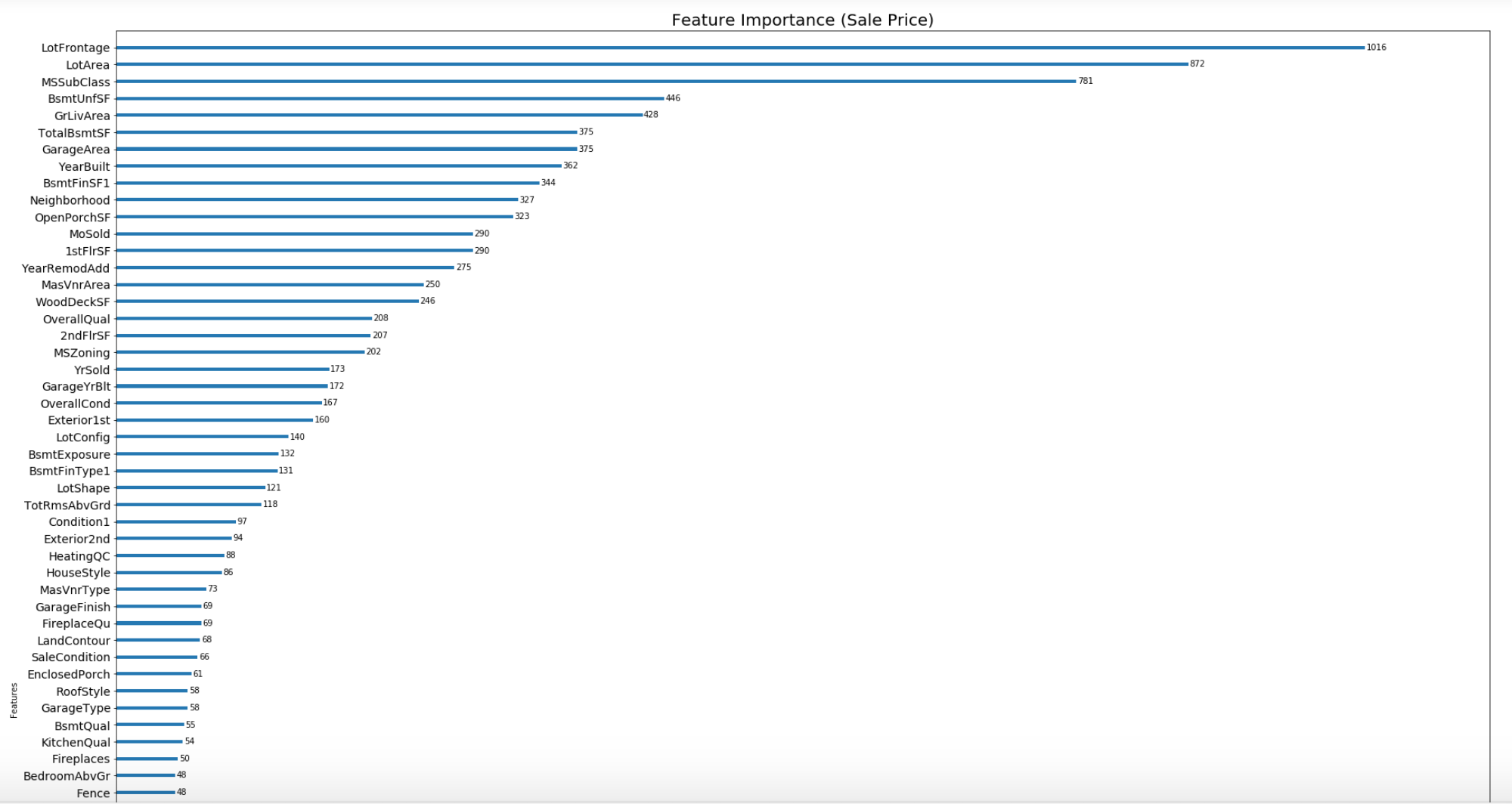
With four different simulations, we could not observe the negative correlation as indicated in the original data set.



ANOVA analysis also showed no significant difference between bedrooms (PR>0.05):



In fact, in the feature importance analysis, bedroom number is ranked 44 in all 79 features, which is not that important.



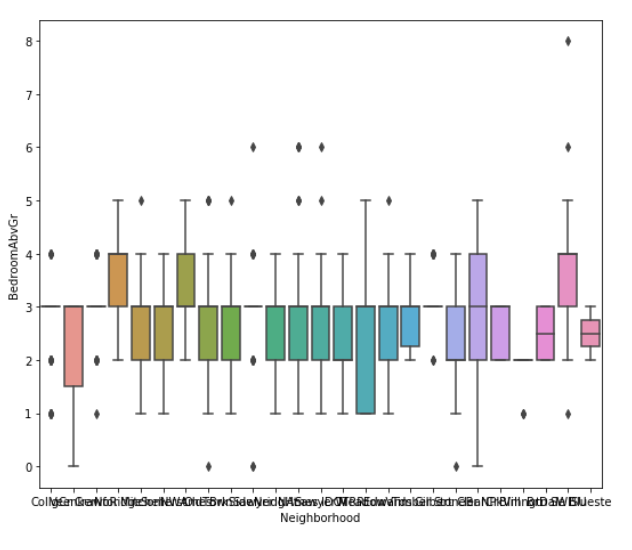
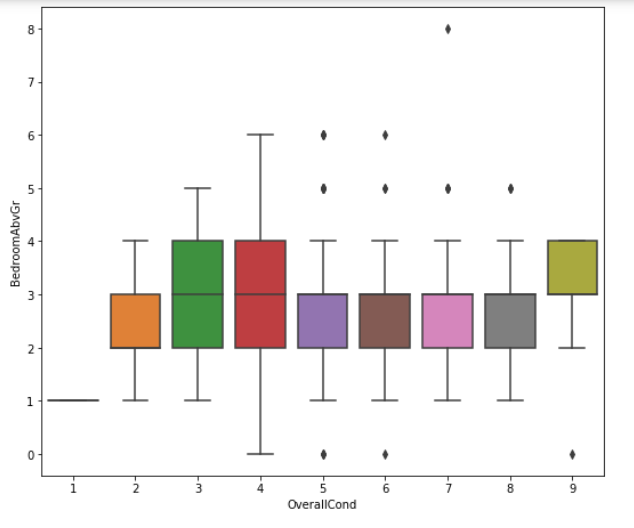
Then what is the reason for us to observe the negative correlation between bedroom number and sale price?

#### Hypothesis one:

There are one or more features that are more important is negatively correlated with Bedroom number. For example: overall quality or neighborhood.

I plotted all relationships between bedroom number and other features, no obvious correlation could be found.

Examples:

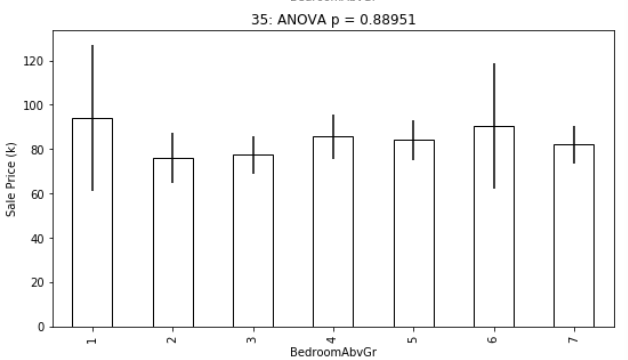


#### Hypothesis two:

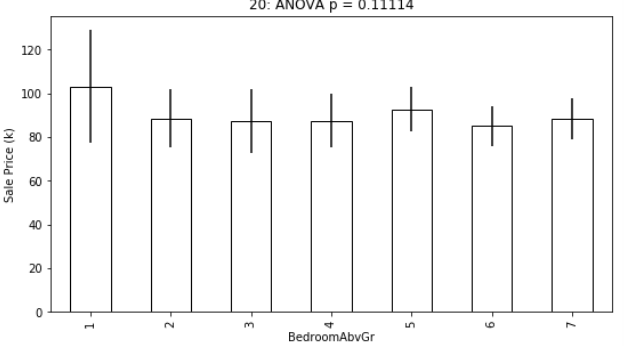
It is a multi-feature combination outcome.

When I added more features to the simulation, I started to see some negative tendency as indicated in the original data. Although the p value is not as significant as in the original data.

Example one: features added from rank 34 to rank 34+35 = 69. Rank by importance in feature importance analysis (rank number the bigger the more important).



Example two: features added from rank 20 + 34 = 54 to rank 79 (rank number the bigger the more important).



## Conclusion:

The ‘bedroom number’ is negatively associated with ‘Sale Price’ in the data but not causally correlated.

## Insights:

### Even though we saw a negative correlation in the original data sets between bedroom number and Sale Price, after simulation experimenting, we found, this correlation is not a cause but a random association, and bedroom number is actually one of the least considered features to affect house prices.

### 

### This study is one example to show that we should be careful to avoid miss interpreting association with causation.