Applying inferential statistics to Capstone Dataset

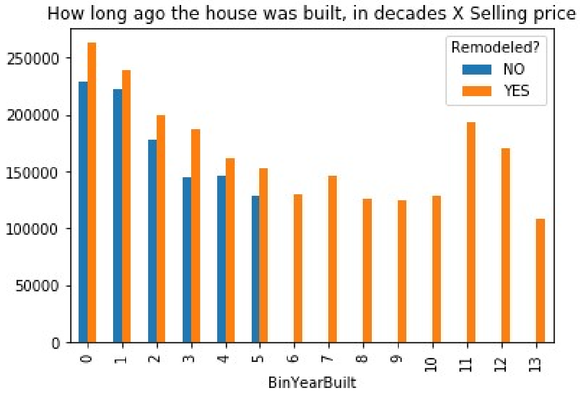
As stated on previous reports, the goal of this project is to make recommendations to owners about home improvements they can make to their properties so that its value increase, enabling them to sell their houses for a price that would compensate their additional investment.

Inferential statistics will allow us to better estimate the likelihood of substantially increasing the price of a house by improving it. To do so, consider the following variables:

YearsSinceBuilt: using year when house was sold as reference, how long ago was it built? (in years)

YearsSinceLastRemod: using year when house was sold as reference, how long ago was it last remodelled? (in years).

We have two groups of interest: houses that have been remodelled and houses that haven't. The following plot, produced on the last report, shows the comparison between the average selling price of these two groups, categorized by the YearsSinceBuilt (in decades):



Considering that the smallest difference between the prices was shown for houses in bin 4 (built between 40 - 49 years before selling), if we can prove that the difference in the mean price for houses that were remodelled and houses that were not is statistically significant, the other houses, that fit other bins, will also be significant.

Let's find the p-value associated with the following hypothesis test (1-sided):

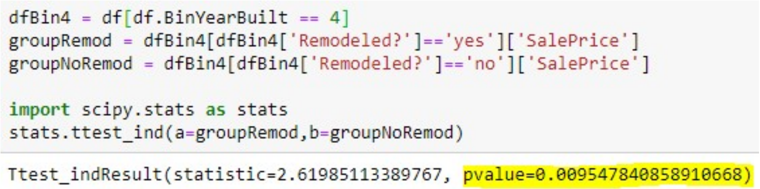
For houses built between 40 to 49 years ago:

H0: Mean(remodelled) – Mean(not-remodelled) = 0

H1: Mean(remodelled) – Mean(not-remodelled) > 0

We will estimate the mean and the standard deviation from the sample and use t-statistic.

Using python code (see below), we notice that the p-value is less then 1%, confirming that we should discard our null hypotheses and accept that remodelled houses tend to be sold for more.



```
dfBin4 = df[df.BinYearBuilt == 4]
groupRemod = dfBin4[dfBin4['Remodeled?']=='yes']['SalePrice']
groupNoRemod = dfBin4[dfBin4['Remodeled?']=='no']['SalePrice']

import scipy.stats as stats
stats.ttest_ind(a=groupRemod,b=groupNoRemod)
```

```
Ttest_indResult(statistic=2.61985113389767, pvalue=0.009547840858910668)
```

For improved results, we should check on the average price of the house improvements the house have experienced, as well as the increase on the selling price. Unfortunately, we don’ t have data on the cost of home improvements, but our findings are promising nonetheless.