2019-02-27

Making Houses More Appealing to Buyers - Milestone project

1. Problem statement

The housing market is the target of several studies and the reasons are simple to understand: it affects an important part of the society and it involves large amounts of money. A house in the market may be just another asset of an investor or a family's entire life worth of savings.

Regardless of the profile, every owner wants to make a good deal when selling their properties. The most significant actions an owner can take to get a better evaluation of their houses are performing house upgrades. But there are so many possible enhancements available, wouldn't it be great if we could identify which services would impact the selling price the most?

As far as this study is concerned, the first part would be to identify which house features are immutable, like location, type of dwelling, etc. and which features can be affected by house projects. The immutable features will be used in the future as a way to classify a house, therefore they will be called "classifying variables". The second group are the ones that the owners can act upon with house projects, therefore they are the "affectable variables".

In the second part of this study, Data Science techniques will be used to identify which "affectable variables" have the most positive impact on the selling price. For this study, this will be the final deliverable.

2. Description of the dataset

The dataset used for this study is the Ames Housing Dataset, which presents 79 explanatory variables describing several aspects of residential homes in Ames, Iowa, United States. The dataset can be found following the link below:

https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview

Once this study is finalized, with minor adjustments, the algorithms generated may be applied for different regions.

To treat this dataset, the following operations were necessary:

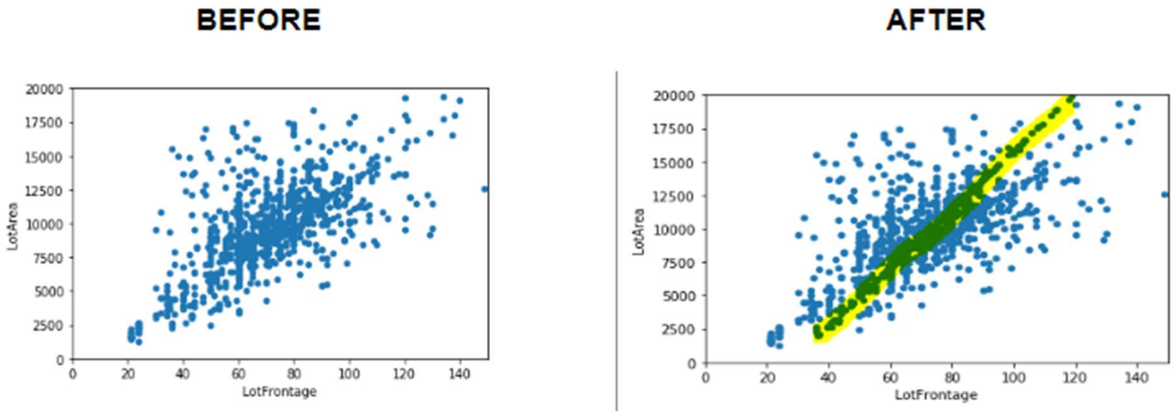
- Checked for errors – mistyped information, for example
- Cleaned or filled missing information
- Checked for consistency among columns
- Treated outliers

All those steps are detailed in the Data Wrangling Report.

3. Initial findings

a. Association between variables LotFrontage and LotArea

When treating the missing information for column LotFrontage, instead of just deleting the NaNs, we identified a linear relation between LotFrontage X LotArea. Here are two scatter plot of the two variables, one before filling the missing values, the other after:


BEFORE

AFTER

This association saved us from losing valuable data. In addition, we confirmed a high correlation between the LotArea and the price of the house, as expected.

b. Roof Style and Roof Material

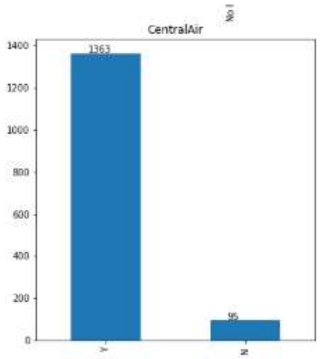
Roof styles are predomintantly "Gable" with material "standard Shingle". As we have very few observations of the other types of Roof and types of material, we cannot draw statistically relevant conclusions. As a result, this variable will not be among those we will suggest to be improved.

c. Pool Quality

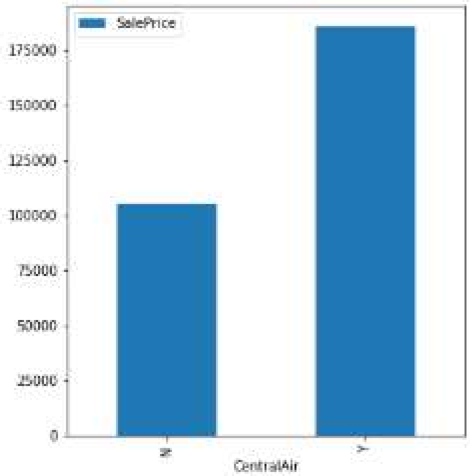
Not enough houses with pool to draw significant conclusions. Improving the quality of the pool will not be suggested to owners.

d. Central Air

Most of the houses on this sample have central air,

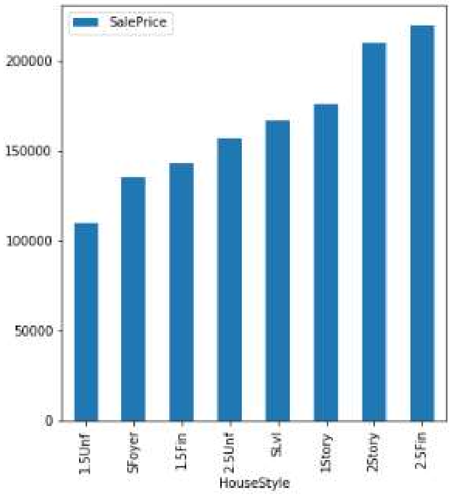


which means that it is possible to find houses with several combinations of the other variables that have that feature. In other words, a buyer will either look for another house or make a very low offer. We understand it is not a simple service but, comparing a house with central air with a similar one without central air, we notice an increment of $80,000 on average.



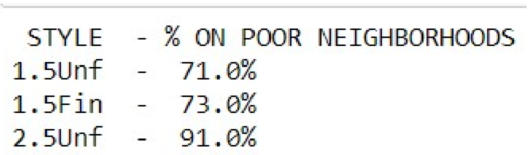
e. Price based on number of stories

We were not expecting the average selling price of a 1 story house being more expensive than a one and a half story (finished or unfinished) and more expensive than a two stories house with 2nd level unfinished.



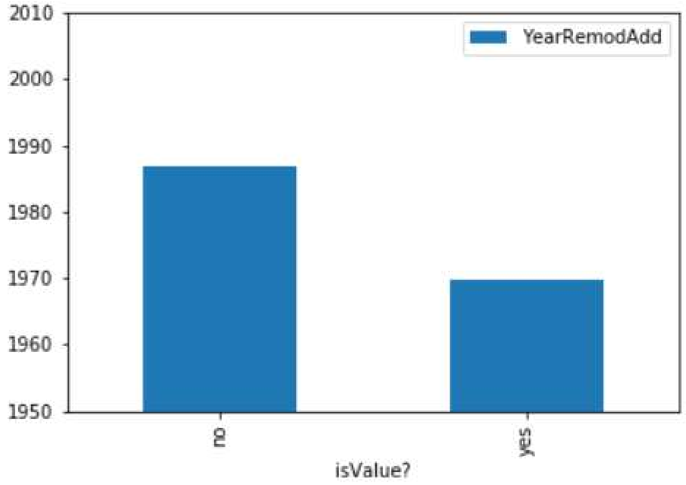
We investigated this and found that:

- 1.5 and 2 stories houses with 2nd level unfinished are located in "poor" neighborhoods



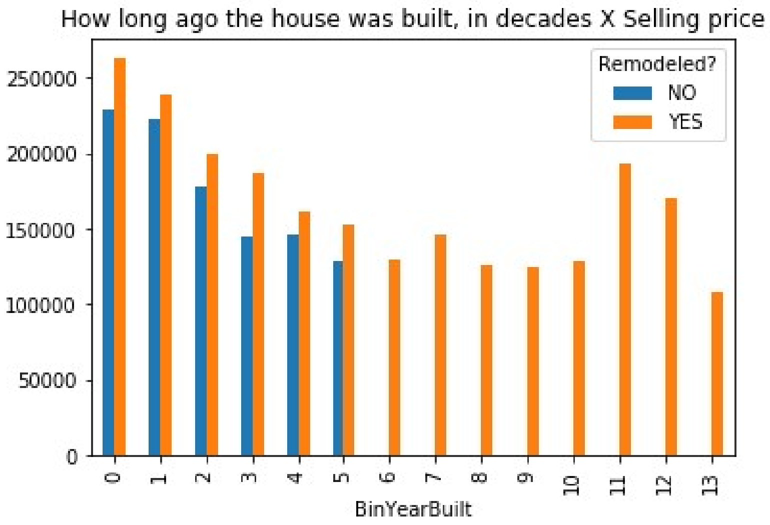
```
 STYLE   - % ON POOR NEIGHBORHOODS
1.5Unf   -  71.0%
1.5Fin   -  73.0%
2.5Unf   -  91.0%
```

- 1.5 and 2 stories houses with 2nd level unfinished are older



f. Difference in price between remodelled houses X non-remodelled houses

We were able to prove that the difference in price of remodelled houses is indeed more expensive than non-remodelled houses, which is the cornerstone of our project. The following plot helps us see that:



Another interesting fact this chart shows is that every house built more then 60 years ago have undergone a house improvement.

Regarding the hypothesis test, check the file "Applying inferential statistics to Capstone Dataset". Here is a summary of what was done:`

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):

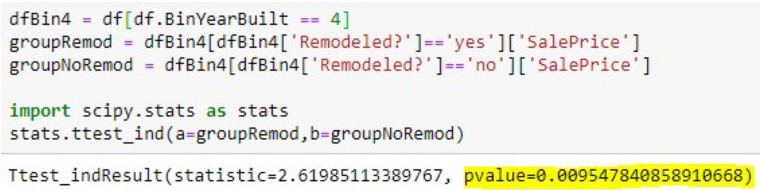
For houses built between 40 to 49 years ago:

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,



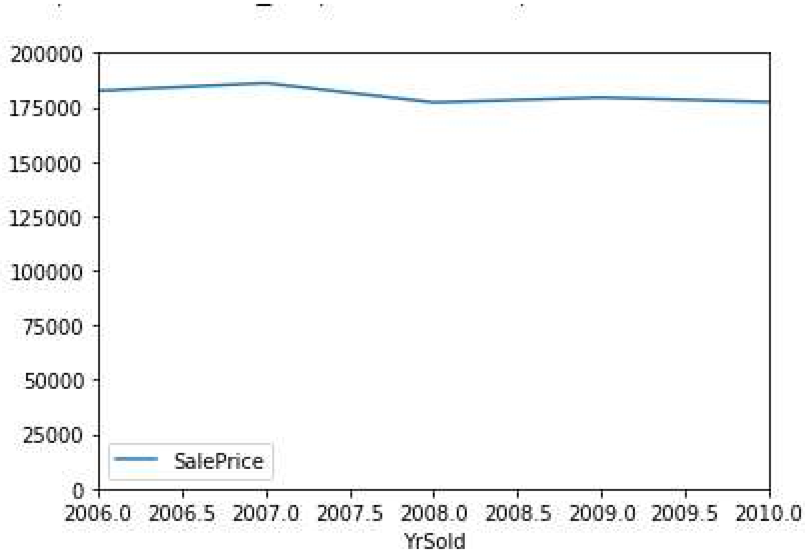
```
dfBin4 = df[df.BinYearBuilt == 4]
groupRemod = dfBin4[dfBin4['Remodeled?']=='yes']['SalePrice']
groupNoRemod = dfBin4[dfBin4['Remodeled?']=='no']['SalePrice']

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

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

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.

g. Selling prices do not increase along the years

We do not need to take into consideration the inflation over the years, as the average selling price does not change significantly



h. Basement improvement

The jump on the average selling price when the basement finished area goes from any category to Good Living Quarters is enormous. It is almost $100,000.00. Which means: if you're going to improve your basement, make it very good because it is worth it. Otherwise, your basement may fall into an average quality and you'll lose your investment. The plot is the same we used for item e.