




Ames Housing Market Predictions



Boldly predicting what no man has predicted
before.



Who? What? Why?

- I am a Data Scientist currently enrolled in General Assembly Data Science Immersive.
- I've been given sample data of housing prices in Ames, Iowa from 2006 - 2010.
- I am looking to predict future housing prices using a various regression models to help better understand the housing market trends in this area.

Baseline

The baseline for our model is the sale price average of \$181469.70
The Kaggle submission score for this baseline model is 81146.36592

The objective for this study is to beat the baseline score. Which will more accurately predict housing prices on unseen data.

Cleaning

- Although many missing values were found, `SimpleImputer()` will be able to handle any missing values.
- All columns were lower cased, and all spaces between column names were replaced with an underscore(_).

Help reduce error when calling columns.

- Half and full bath were converted into `total_baths`

(1 bath + 2 half baths = 2 total baths)

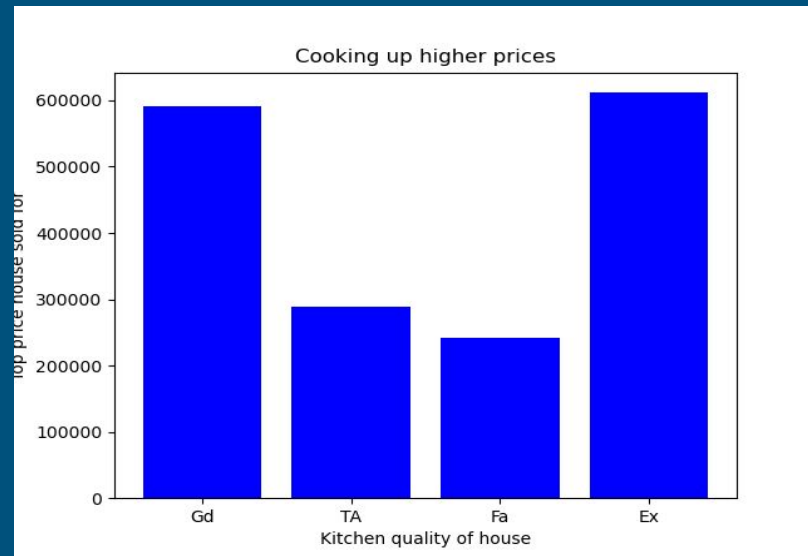
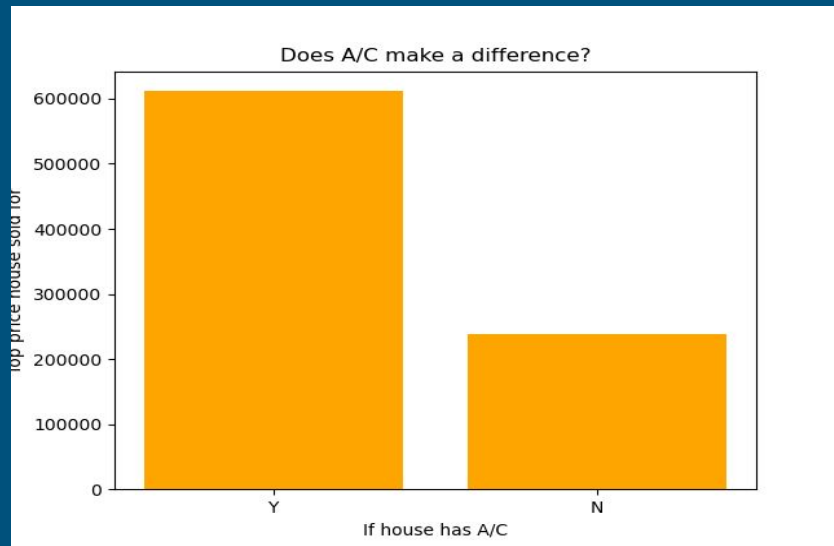
- Total square footage was calculated by adding up all above ground square footage and finished basement square footage.

$$\text{Total_sf} = (\text{1st fl} + \text{2nd fl}) + (\text{bsmnt sf} - \text{unfinished bsmt sf})$$

EDA

- Mean price/sq_foot is 93 dollars. Which is above the national average of 83 dollars/sf
(<https://www.statista.com/statistics/682549/average-price-per-square-foot-in-new-single-family-houses-usa/>)
- 93% of houses have central air. Thoses house sold for an average of \$87,000 more than those without.
- Houses with a kitchen quality of excellent compared to good sold for an of \$124,000 more.
- Houses with a basement quality of excellent compared to good sold for an average of \$126,000 more.
- The sweet spot for bathrooms is 4.5 with the average selling price of \$337,000, almost \$50,000 than any other bathroom count.

Graphical EDA



Gd = Good

TA = Average

FA = fair

EX = Excellent

Finding our variables

	overall_qual	overall_cond	Total Bath	fireplaces	bedroom_abvgr	garage_cars	totrms_abvgrd	wood_deck_sf	price_per_sq	total_sq	saleprice
overall_qual	1.000000	-0.082770	0.542808	0.388920	0.053373	0.587423	0.382025	0.257081	0.500294	0.536865	0.800207
overall_cond	-0.082770	1.000000	-0.188893	-0.006463	-0.009908	-0.168513	-0.093576	0.011034	-0.002689	-0.089139	-0.097019
Total Bath	0.542808	-0.188893	1.000000	0.325330	0.202630	0.496635	0.422829	0.291446	0.030428	0.689654	0.629500
fireplaces	0.388920	-0.006463	0.325330	1.000000	0.079194	0.310687	0.311765	0.238887	0.076391	0.495485	0.471093
bedroom_abvgr	0.053373	-0.009908	0.202630	0.079194	1.000000	0.085918	0.655439	0.034758	-0.123316	0.245211	0.137067
garage_cars	0.587423	-0.168513	0.496635	0.310687	0.085918	1.000000	0.368034	0.240721	0.341219	0.481051	0.648220
totrms_abvgrd	0.382025	-0.093576	0.422829	0.311765	0.655439	0.368034	1.000000	0.182835	0.043461	0.548259	0.504014
wood_deck_sf	0.257081	0.011034	0.291446	0.238887	0.034758	0.240721	0.182835	1.000000	0.063288	0.330321	0.326490
price_per_sq	0.500294	-0.002689	0.030428	0.076391	-0.123316	0.341219	0.043461	0.063288	1.000000	-0.187825	0.464890
total_sq	0.536865	-0.089139	0.689654	0.495485	0.245211	0.481051	0.548259	0.330321	-0.187825	1.000000	0.718685
saleprice	0.800207	-0.097019	0.629500	0.471093	0.137067	0.648220	0.504014	0.326490	0.464890	0.718685	1.000000

Pre Processing

Pre Processors:

`SimpleImputer()` - Replaced all empty values with the mean.

`PolynomialFeatures()` - Squaring all target columns.

`OneHotEncoded()` - Making binary columns for categorical information.

`StandardScaler()` - Scaling each column into computer only understandable data.

Problems:

Keeping track of each variable name for every processor became a difficult task.

Solutions:

Wrapping all processors into a Column Transformer simplified all pre processors into one variable.

Regularization and R2 scores

Regularization is the process of making an overfit model closer to the training score.

As my image to the right indicates, my model was not overfit prior to the Ridge and Lasso regression tests. Which only made my models test score a very small percentage better.

```
===== lr =====  
0.9386194675823075  
0.8906042029215874  
===== Ridge =====  
0.9314080680978707  
0.9045809306107356  
===== Lasso =====  
0.938527954700138  
0.892651846272057
```

Findings

Baseline rmse = 81146.36592

My rmse = 38869.25083

With the intention of creating a model to beat our baseline score, we would call this a success

Inferring my model is capable of predicting future housing prices on unseen data.

Sources

<https://www.statista.com/statistics/682549/average-price-per-square-foot-in-new-single-family-houses-usa/>