Ames Housing Data: Analysis

John Lawless

Problem Statement

- I am a data scientist for a large organization, who has taken on a realty company as a client.
- This company is interested in acquiring a data- driven marketing advantage in home buying.
- This company is not interested in “fixing up” homes, instead acting as a mediary for homes ready to sell, preferably for high prices.
- Current business model prioritizes:
    - Newer properties (both for value and expected lack of expense in remodelling)
    - Desirable locations (proficient neighborhoods, access to desired amenities, etc)
    - Highest overall inspector rankings.
- Their request to us is both to better understand marketing trends, and to offer them a method of accurate predicting selling prices of new homes as they come on the market.

The Data Problem

The key questions to answer, in addition to creating a predictive model:

1) What is the “average” selling price for all homes?
2) Do newer houses (those less than 50 years old) sell significantly higher than this average price?
3) Do newer houses sell significantly higher than older houses? *If they don’t, it may be evidence that some other factor is responsible other than age.*
4) If 2 and 3 are yes - what is the average selling price of these newer homes?
5) How does the overall ranking of properties affect selling prices of homes?
6) How does the location of a home (in terms of neighborhood) affect the price of a home?

Strategy for Approaching Data Problem

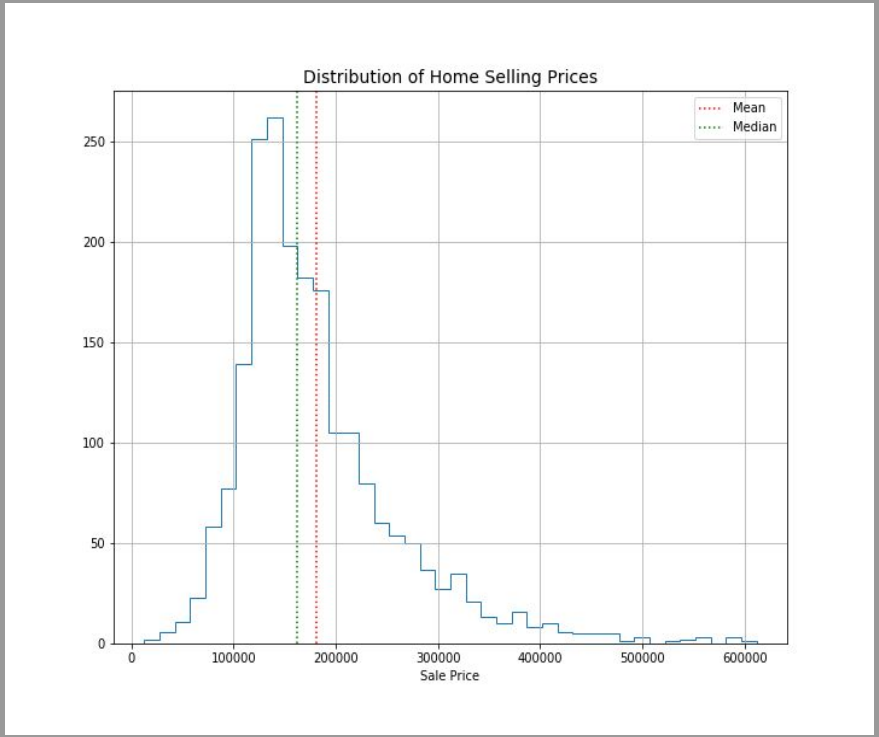
White box predictive models (such as linear regression) carry with them the benefit of interpretability of the data. My strategy, after exploratory analysis, is to build a simple linear model that focuses on explaining the effects (if any) key factors of interest to our client have on a home's selling price. After this, if needed, I will focus on a secondary model built for optimal predictive power and minimal error.

Analysis - what is the “average” selling price for all homes?

- I can infer with 95% confidence from this data that the true average selling price for all homes in this region is captured by the interval of $178,000 - $185,000.

However, as the histogram shows, the sales data is very right skewed, so I do not think this is a good number to utilize. A better representation is likely the median of the data:

- The same interval for the median selling price is $159,000 - $166,000, which I feel represents the central tendency of the data better.



Analysis - Newer Homes vs...

By hypothesis testing the various average selling prices of newer homes vs other types of homes, I can state that this data shows evidence that:

- Newer homes have a significantly different average selling price from homes in general (there is some difference).
- Newer homes do not have the same average as older homes.

There appears to be evidence that a home being newer does significantly change the average selling price.

With these factors considered, I can also state with high confidence that the true average selling price for newer homes is captured by the interval of $215,000 - $225,000.

It appears that our client's business model of focusing on new homes is not without merit!

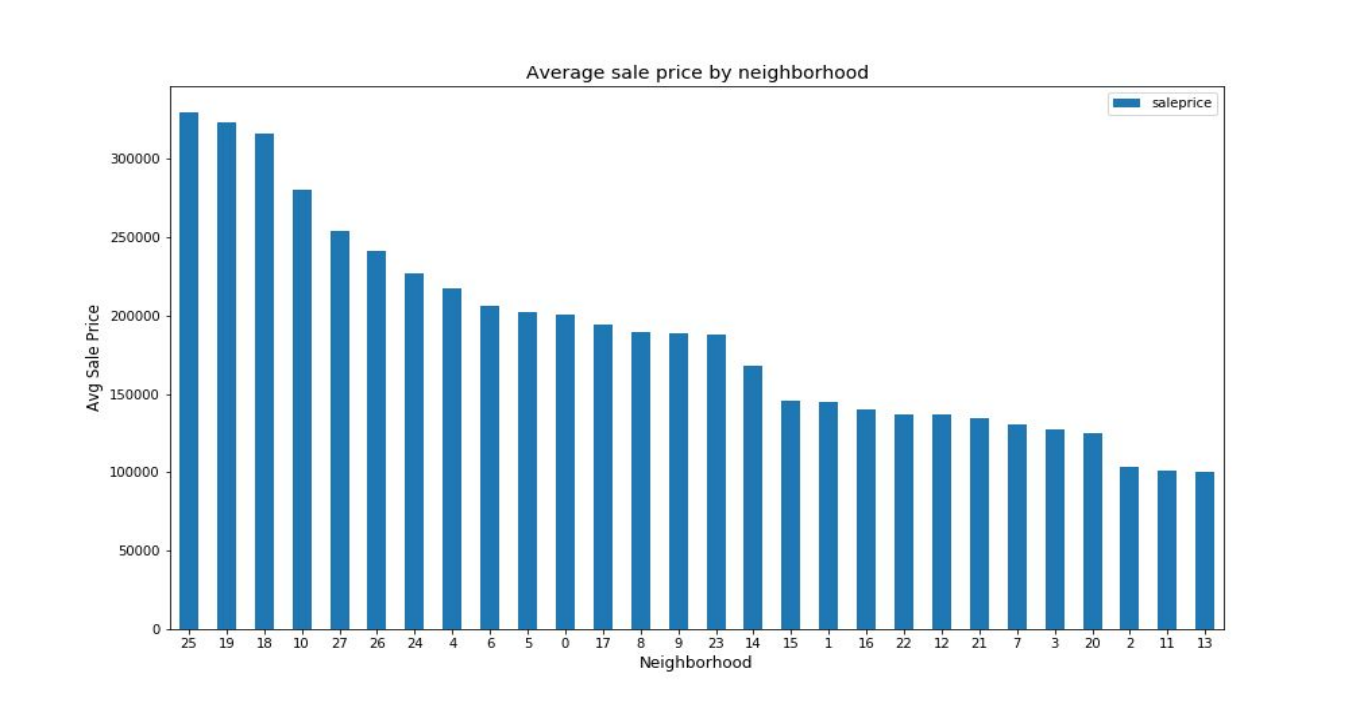
Analysis - Effect of overall ranking on home selling price

As this graph shows, there is a strong positive correlation between the overall home ranking and the average selling price.

It again seems that focusing on higher ranking homes is a good way to maximize selling price.

Analysis - Home Price by Neighborhood

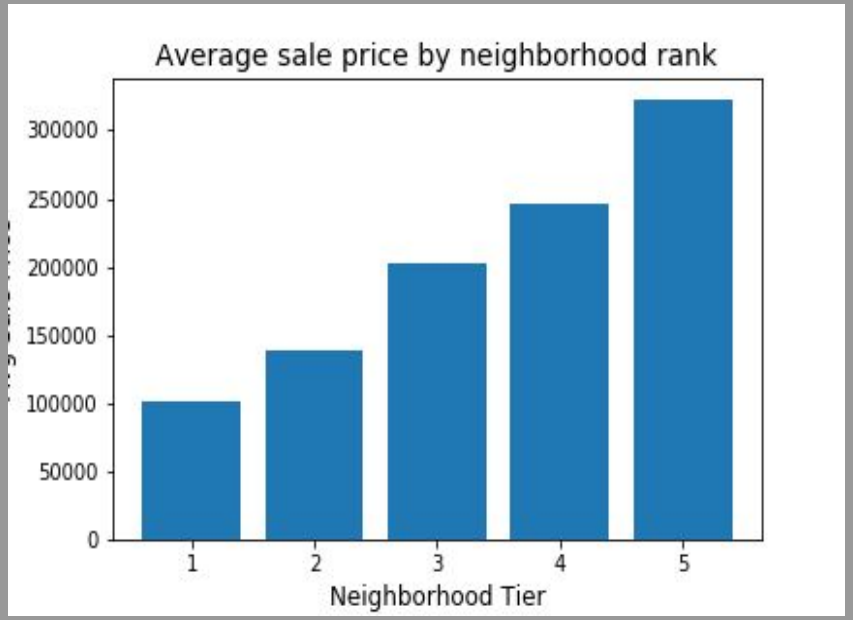
While there is a strong positive correlation here, this data is too complex to easily describe, as seen:



Analysis - Home Price by Neighborhood

To simplify this data, I organized these neighborhoods into ordinal tiers. This is both easier for our client to keep track of, and clearer to interpret from a model.

As can be seen, the same trend is preserved here, but each tier represents a collection of neighborhoods, so that we can analyze the effect they have on home selling price.



Analysis - Home Price by Age

While this trend is not perfectly linear, there is still a negative correlation between sale price and age of a home. In other words, as a home gets older, it generally decreases in potential selling price.

These key factors of our client's business model all do appear to be correlated to home selling prices in ways favorable to their business. But *how* do these factors affect home prices?

Results: How Client's Key Factors affect Home Selling Price (Interpretable Model)

By running a simple linear regression model on these key factors (and checking for valid conditions for inference), I can report the following:

- For every increase of one square foot of property size, we can expect to see an increase found between the interval of $1.47 - $2.10 in home selling price, holding all other factors constant (close to a 2:1 ratio).
- For every additional year of age a home has, we can expect to see a decrease of selling price between $183.37 - $484.69, with all other factors the same.
- A home being at the top overall rank ("Very Excellent") predicts a home selling for an increase captured by the interval $312,000 - $386,000, relative to that same home being ranked lowest ("Very Poor").
- A home being in the top tier of neighborhoods (Stone Brook, Northridge Heights, and Northridge), is expected to sell between $34,000 - $76,000 more, relative to that same home being in the lowest tier (Briardale, Iowa DOT and Railroad, and Meadow Village).

Optimizing Predictive Power of Linear Modelling

The interpretable model performed tolerably well, predicting roughly 77% of home selling price variance on unseen data. However, there was room to optimize.

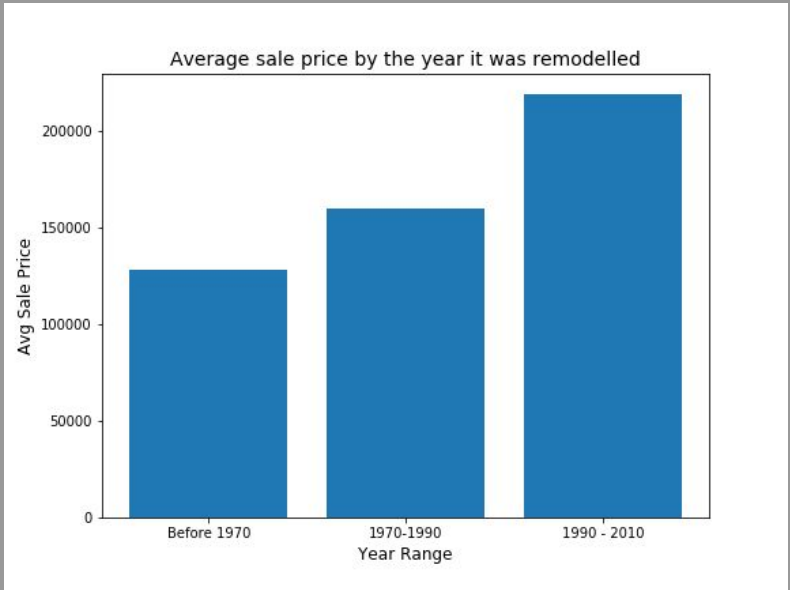
This was largely done by adding complexity to the model, giving the model more information to identify trends for when predicting the price of new homes on the market.

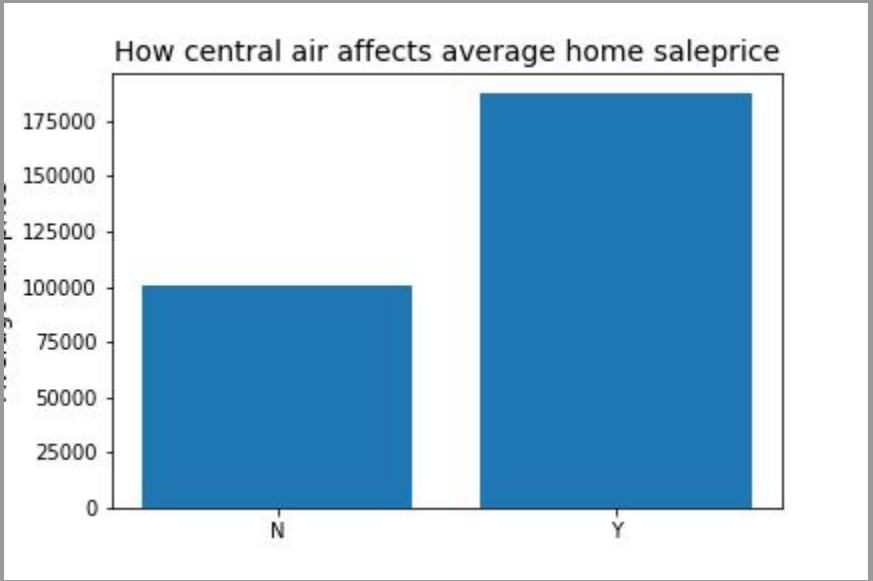
In addition to increasing complexity with existing features, new features were also engineered and interpreted to follow seen trends in the data.

Examples of many simple new feature questions: “Does this home have a pool?” “Does this home have a basement?” “Does it have a garage?” “Does it have a fireplace?”

Examples of New Features

There are far too many new features to explore them all, but here are illustrations of some new features added:

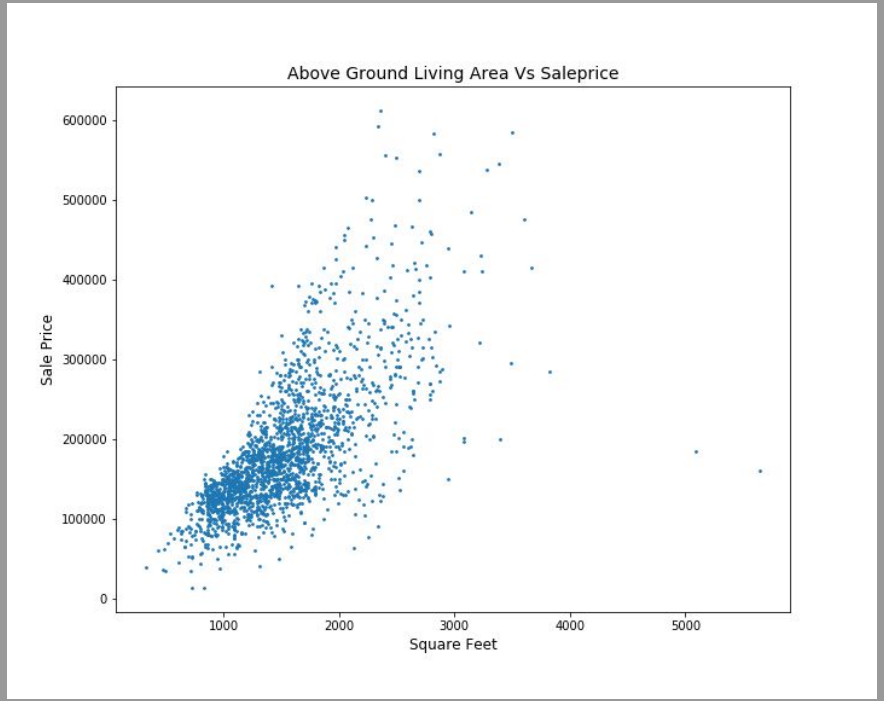




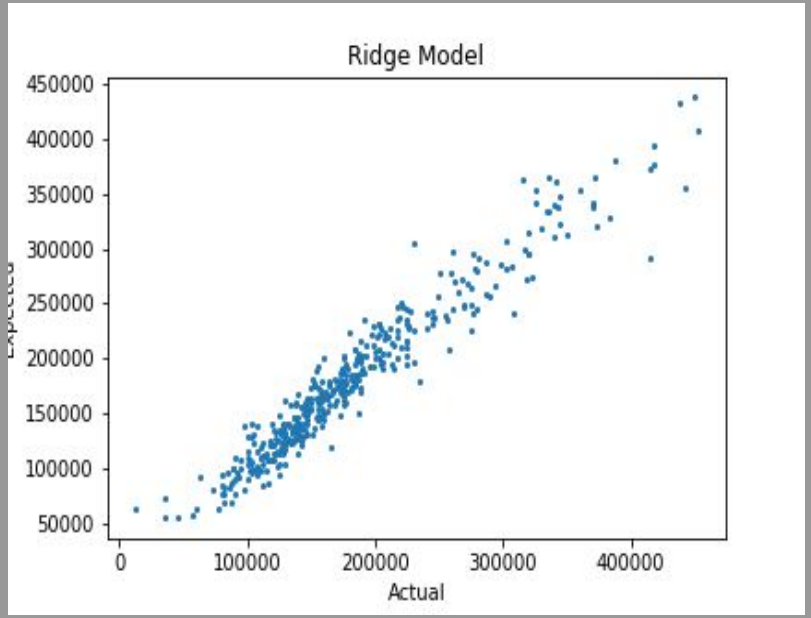
Example of Some Interpreted Features

The image here shows how above ground living area correlates to home sale price. As can be seen, it is not linear but has a strong positive curve. Many other features relate to sale price in similar correlations.

Adding exponential features to this model assisted it in following these trends appropriately, as well as logarithmic features to correct for skewed distributions.



Results of Optimized Model



The optimized model fit just under 90% of the variance in sale price of unseen data, indicating higher predictive power.

This image shows the actual sale price of data vs what the model predicted.

This model has a high predictive power and can be used for the benefit of our client, but be mindful that it will tend to underpredict very high selling price homes.

Recommendations from this study

- The priorities of our client's business model are effective choices for identifying high value homes, and utilizing the relationships shown in this report can help them capitalize on these factors.
- Consider including other factors as a priority according to this report, such as the presence of modern amenities such as central air, or considering older homes if they have been recently remodeled.
- Utilize the predictive model including in the deliverables to predict the selling price of new homes as they come on the market (but be cautious on the prediction of very high price homes!)

References Used in this Study

-https://www.opendoor.com/w/blog/factors-that-influence-home-value

- https://stackoverflow.com/questions/38420847/apply-standardscaler-to-parts-of-a-data-set

https://scikit-learn.org/stable/modules/generated/sklearn.compose.make_column_transformer.html

-http://jse.amstat.org/v19n3/decock/DataDocumentation.txt