Optimizing Sale Price in Ames, Iowa

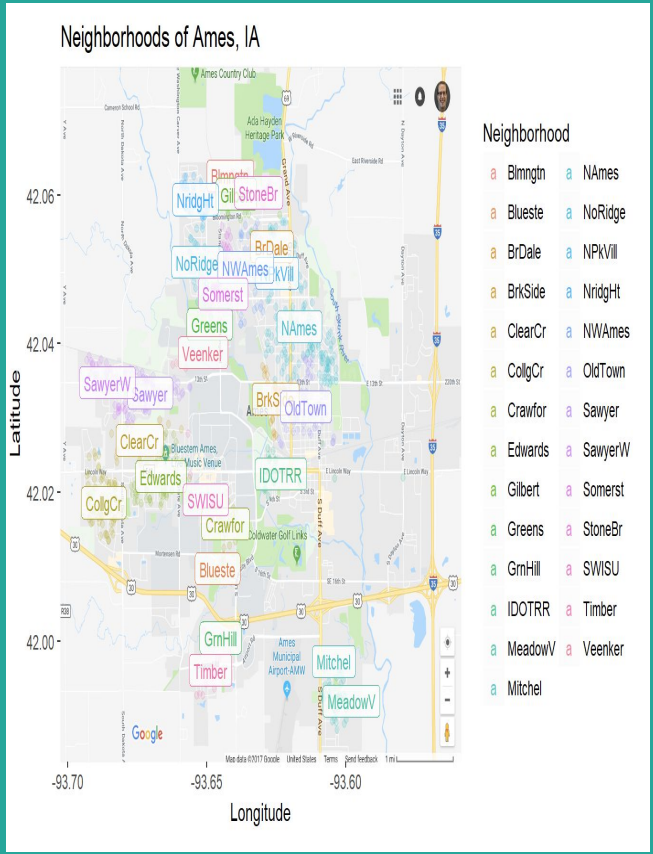
—

By Rose Dennis

Overview

- Problem Statement
- The Data and its Caveats
- The Impact of Neighborhood
- Kitchen Quality
- Modeling
- Conclusions

How can we maximize the sale price of a home in Ames, Iowa? What features are most important to potential clients?



*Full credit to BEH statistical consulting for this image

The Data and its Caveats

27 Neighborhoods
2,051 homes
80 original features

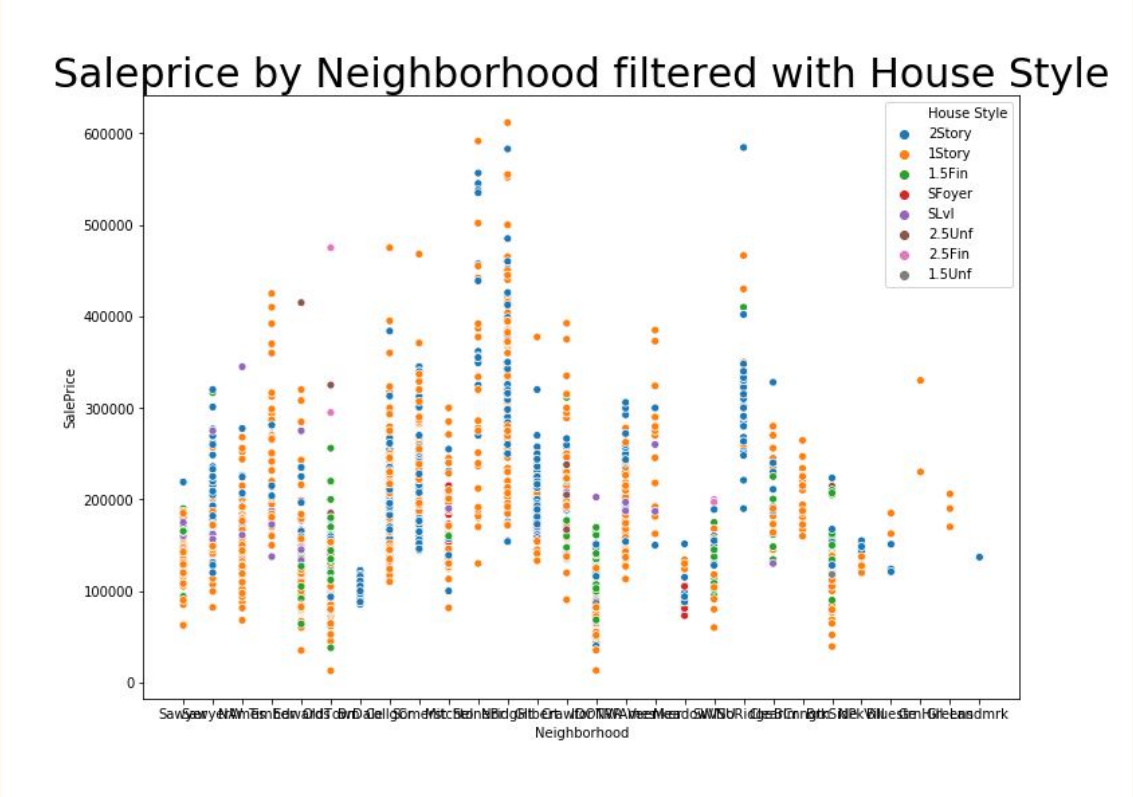
- Unbalanced Classes (ie. Neighborhood)
- Outliers
- Missing Values
- Ambiguous feature names (ie. Overall Quality)
- Multiple Linear Regression Assumptions

The Impact of Neighborhood

North Ridge Heights

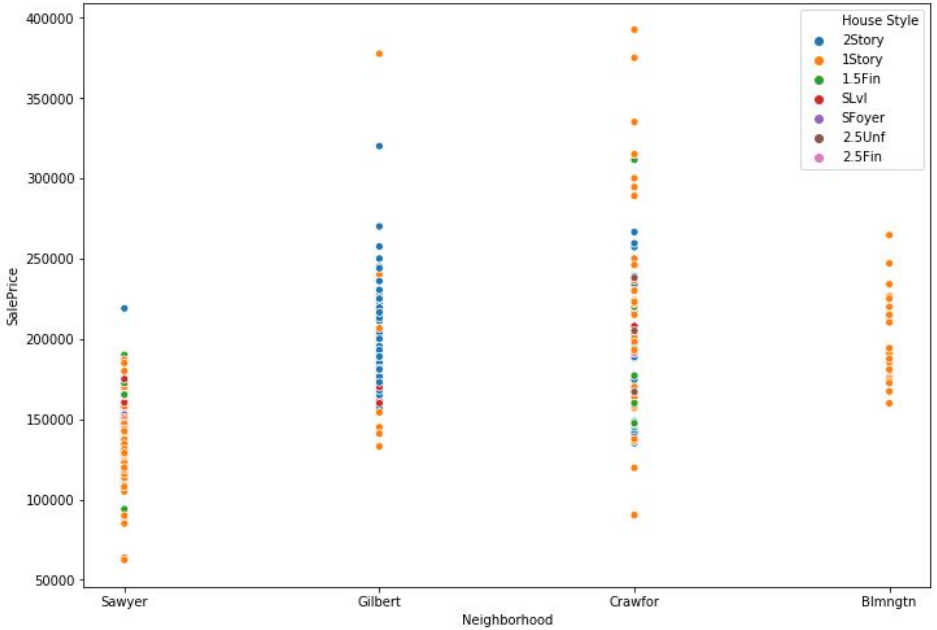
Gilbert

The Impact of Neighborhood-Diving into the Data



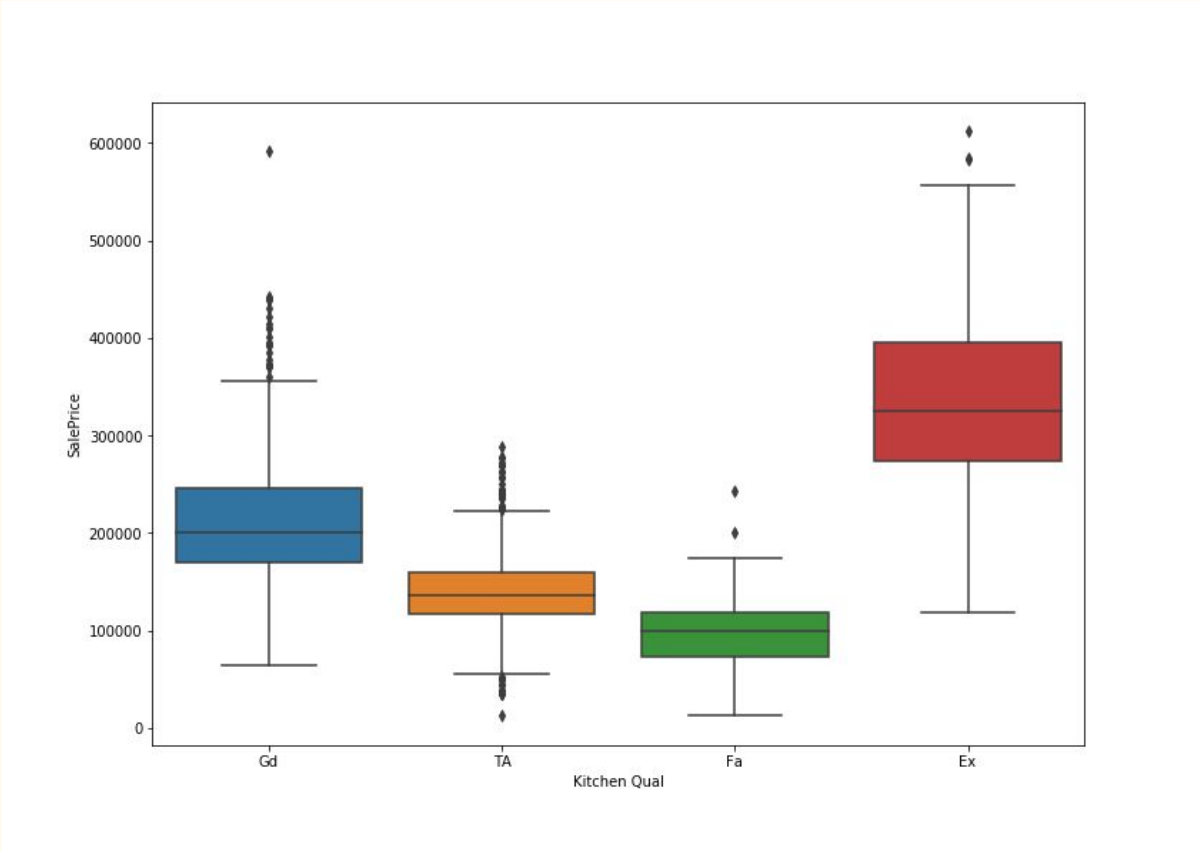
The Impact of Neighborhood-Diving into the Data Random Sample of 5 Neighborhoods

- Gauge different mean/median
- Different number of potential outliers for each
- Types of houses differ
- Mix/Max Sale price



Kitchen Quality

The median price for each kitchen quality is clearly different. We will put this in our model. We can also see outliers.



Modeling-Features

- TotRms AbvGrd
- Have Pool
- **Neighborhood**
- Utilities
- Bldg Type
- Exter Qual
- Heating
- Central Air
- **Kitchen Qual**
- Functional
- Garage Qual
- Overall Qual
- Gr Liv Area
- Garage Area
- 1st Flr SF
- Year Built
- Year Remod/Add
- Full Bath
- Lot Frontage
- Mas Vnr Area
- Total Bsmt SF

Linear Regression Model Formula

$$Y_i = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon_i$$

Interpretation for our model. Y_i represents our predicted values for Sale Price. The intercept implies if we had 0 features, this is how much the house would cost. Each beta represents for every one unit increase in X, there's a respective beta i increase in Y.

H_0: $\beta_0 = \beta_1 = \beta_2 = \ldots \beta_p = 0$
H_A: At least one $\beta_i \neq 0$

Our Model- The Math

$Y_i = 5.2004 + 0.0741$(Overall Qual) $+ 0.0003$(Garage Qual) $+ 0.0013$(Year Built)... - 0.0101(TotRms AbvGrd) $+ 0.4575$(Neighborhood_GrnHill) $+ 0.0935$(Neighborhood_Greens) $+$...

R^2 Score $= 0.887$

Remember! Our model uses the log(SalePrice) so we have to interpret slightly differently.

Our Model- Interpreting the coefficients

Numerical Feature
Overall Quality: $\beta = 0.0741$
Interpretation: For one unit increase in the overall quality, the sale price increases by about 7%.
Categorical Feature
Green Hill: $\beta = 0.4575$
Highest Beta coefficient while also having a significant p-value.
Interpretation: All else being held equal, the sale price of a house in Green Hill is about 45% more expensive than the next highest neighborhood, Greens.

Possible Next Steps

- Interaction term between Neighborhood and Total Rooms above ground
- Fine tuning the model
- Unbalanced Classes- up/down sampling or weighting

Conclusion

As expected- Neighborhood matters!

Renovating? Look at the kitchen quality first.

Let’s sell some houses!

Outside Resources:

http://www.tomrandallrealestateteam.com/Newsletter

https://rstudio-pubs-static.s3.amazonaws.com/337439_24918eaefe724411be93e41ede48b256.html