# More Than a White Picket Fence: How Home Features Influence Sale Price

Yvonne Renard, Alex Teboul, Sara Elkasevic

## Why is it important to understand how Sale Price is affected by the features of a home?

*...In the midst of growing economic uncertainty and concerns over the housing market, a better understanding of home value can help homeowners and investors alike.*

## 2 Benefits of this Research:

1. More informed homeowners can better price their properties to sell.
2. More informed buyers can identify value in properties to ensure they're getting their money's worth compared to other buyers.

# Introduction - Dataset

- Our goal was to discover the underlying factors driving Sale Prices in the Ames Housing Dataset and determine the extent to which we could predict prices based on the features of a home.

- Ames Housing Dataset
  - 2930 observations of homes in Ames, Iowa
  - 79 explanatory variables of home features
  - *Sale Price*


# Literature Review

i. Performance of Multiple Linear Regression and Non-Linear Neural Networks and Fuzzy Logic Techniques in Modeling House Prices

by: Siti Amari and Gurudeo Anand Tularam

- Uses: Linear Stepwise Multivariate Regression, Neural Networks, and Adaptive Neuro-Fuzzy
- Neural networks are non-linear data-driven methods

ii. Forecasting House Prices in OECD Economies

by: N.Kundan Kishor and Hardik A. Marfatia

- Uses: Forecast Combination Methods
- Allows for macroeconomic changes over time and across countries

iii. Property Renovations and Their Impact on House Price Index Construction

by: A. N. Bogin and W. M. Doerner

- Uses: Ordinary Least Squares Estimator
- Introduces renovation control and fixed effects for HPI estimations

# Methods / Research Questions

1. Principal Component Analysis
2. Common Factor Analysis
3. Random Forest
4. Linear Multivariate Regression

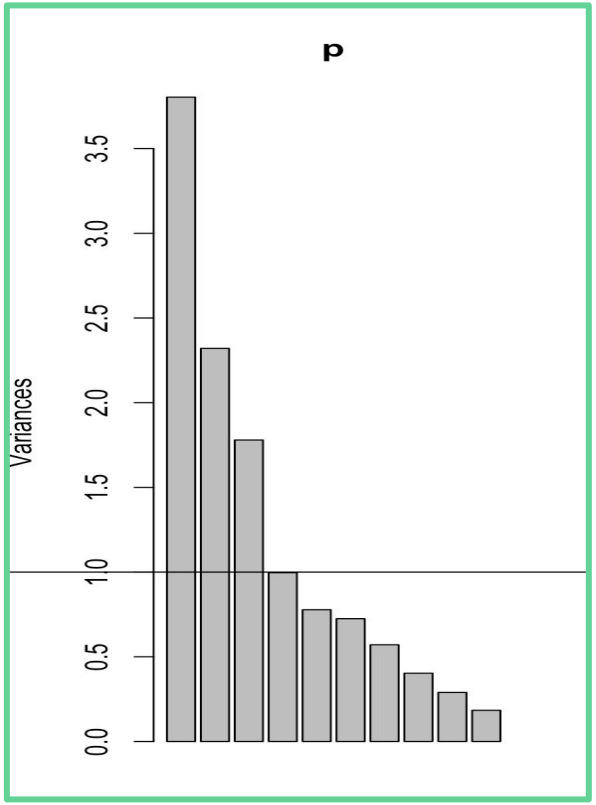
What are the most important features of (and their marginal effects) on home prices in Ames, Iowa?

Summary Statistics

| | min. | max. | median | mean |
|---|---|---|---|---|
| Sale Price ($) | 12,789 | 755,000 | 160,000 | 180,796 |
| Lot Area (sq ft) | 1,300 | 215,245 | 9,436 | 10,148 |
| Bedrooms | 0 | 8 | 3 | - |
| Full Bathrooms | 0 | 4 | 2 | - |
| Fireplaces | 0 | 4 | 1 | - |
| Overall Quality | 1 | 10 | 6 | - |
| Garage Cars | 0 | 5 | 2 | - |

# Principal Component Analysis

- PCA analysis was done on numerical variables using a varimax rotation
- Three components were the most optimal and explained the most variance
- Bartlett's Test of Sphericity ➜ p-value < 2.22e-16
- KMO Sampling Adequacy ➜ 0.83



# Common Factor Analysis

**What we did:**

1. Ran CFA with varimax rotation on our numeric variables to better understand factors at play in our dataset:
   a. Factor 1: Quality, Age, Size, and Quantity
   b. Factor 2: Basement
   c. Factor 3: Fireplaces
2. Conclusions:
   a. 58.3% Cumulative Variance
   b. Newer, Larger homes ➔ Higher Quality Ratings
   c. Basement variables could be consolidated.

| | Factor1 | Factor2 | Factor3 |
|---|---|---|---|
| overallqual | 0.803 | | |
| yearbuilt | 0.750 | | |
| yearremodadd | 0.725 | | |
| fullbath | 0.596 | | |
| garagecars | 0.674 | | |
| garagearea | 0.639 | | |
| exterqual | 0.824 | | |
| bsmtqual | 0.673 | | |
| heatingqc | 0.586 | | |
| kitchenqual | 0.763 | | |
| bsmtfinsf1 | | 0.813 | |
| bsmtunfsf | | -0.695 | |
| bsmtfullbath | | 0.692 | |
| bsmtfintype1 | | 0.756 | |
| fireplaces | | | 0.936 |
| fireplacequ | | | 0.868 |
| x1stflrsf | 0.459 | | |

| | Factor1 | Factor2 | Factor3 |
|---|---|---|---|
| SS loadings | 5.518 | 2.346 | 2.047 |
| Proportion Var | 0.325 | 0.138 | 0.120 |
| Cumulative Var | 0.325 | 0.463 | 0.583 |

# PCA Results

1. Component 1
   a. Year Built ➔ 0.831
   b. Year Remodeled ➔ 0.812
   c. Garage Year Built ➔ 0.858
2. Component 2
   a. Finished Square footage of Basement ➔ 0.840
3. Component 3
   a. Lot Frontage ➔ 0.753
   b. Lot Area ➔ 0.710
4. Cumulative Variance of 62%

```
Loadings:
             RC1    RC2    RC3
overallqual   0.764
yearbuilt     0.831
yearremodadd  0.812
fullbath      0.645
garageyrblt   0.858
bsmtqual      0.720
heatingqc     0.671
kitchenqual   0.744
bsmtfinsf1           0.840
bsmtfullbath         0.824
bsmtfintype1         0.837
lotfrontage                 0.753
lotarea                     0.710
x1stflrsf                   0.705
masvnrarea                  0.429

                 RC1   RC2   RC3
SS loadings    4.850 2.345 2.080
Proportion Var 0.323 0.156 0.139
Cumulative Var 0.323 0.480 0.618
```

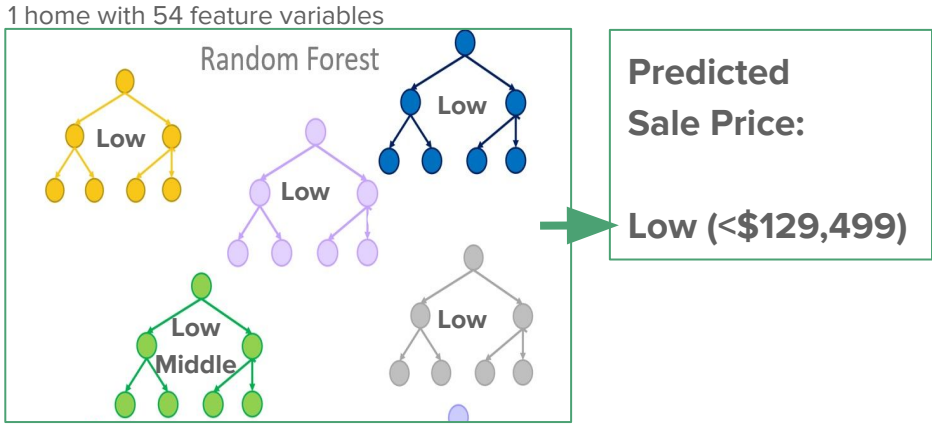
# Random Forest – Sale Price Classification

**What we did:**

1. We split Sale Price into 4 equally sized groups:
   a. **Low** (<$129,499)
   b. **Low Middle** ($129,500 - $160,000)
   c. **High Middle** ($160,001 - $213,500)
   d. **High** (>$213,501)
2. Subset of 54 numeric home feature variables
   ➔ Train/Test Split (70/30)
3. Train Random Forest classifier
4. Test RF classifier & Report Results

**How it works:**

- 500 Decision Trees **predict Low**, **Low Middle**, **High Middle**, or **High** for each home based on the 54 variables. Trees **split on Impurity**. We want homogenous nodes. **RF Majority Vote**.



# Random Forest Results

**5 Important Variables in the RF Model:**

1. Overall Quality
2. 1st floor square footage
3. Year Built
4. Lot Area
5. Year Remodeled

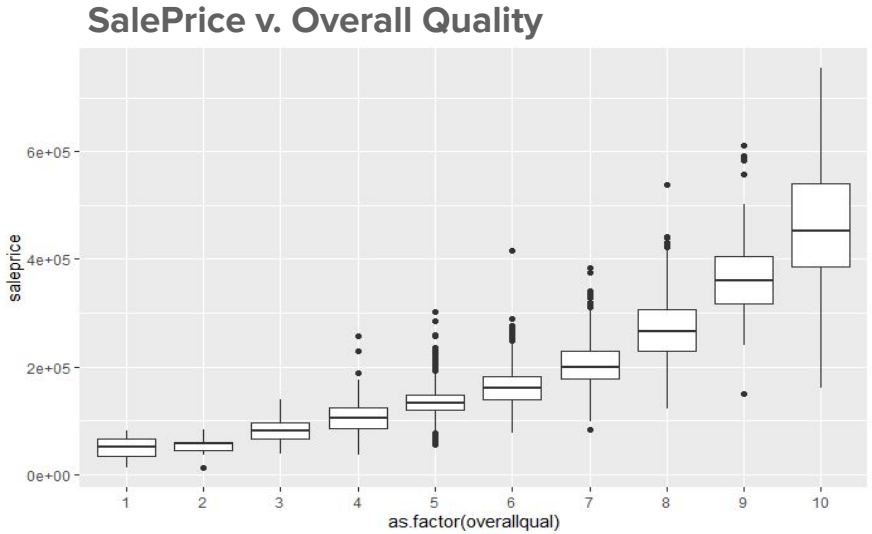

**Training Set:**

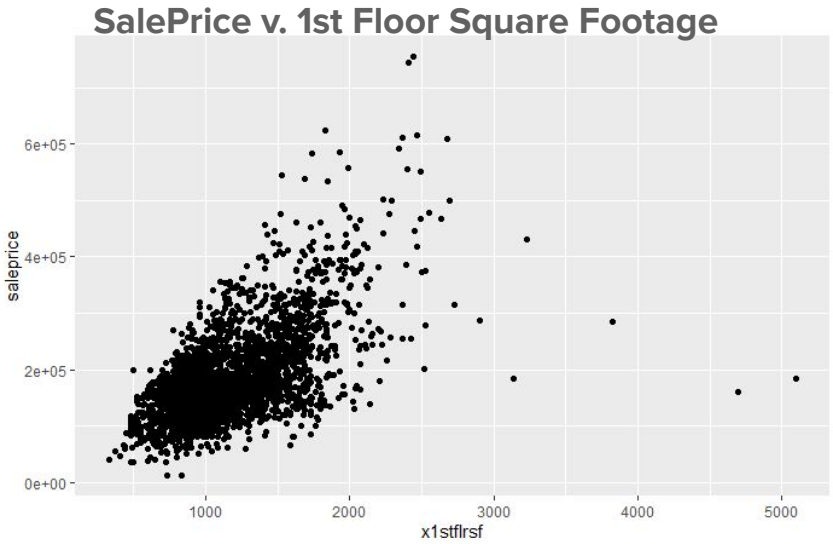
```
Confusion matrix:
             high low middle high middle low class.error
high          436   0          72          8   0.1550388
low             0 415           4         86   0.1782178
middle high    48   1         382         93   0.2709924
middle low      3  94          63        346   0.3162055
```

**Testing Set:** **92.7% Accuracy**

```
Confusion Matrix and Statistics


housing_prediction high low middle high middle low
       high          193   0          5          0
       low             0 213          1          8
       middle high     8   0        211         10
       middle low      0  13         19        198

Overall Statistics

               Accuracy : 0.9272
                 95% CI : (0.908, 0.9435)
```

# Important 2 Variables in Random Forest



# Backward Stepwise Multivariate Linear Regression

$$\widehat{Sale\ Price} = Type\ of\ Home + Lot\ Area + Overall\ Quality + Overall\ Condition + Year\ Built + Masonry\ Type + Sq\ Ft\ of\ Finished\ Basement + Rating\ of\ Basement + Sq\ Ft\ of\ Unfinished\ Basement + Full\ Bath\ in\ Basement + Full\ Bathrooms + Bedrooms\ Above\ Ground + Kitchen\ Above\ Ground + Total\ Rooms\ Above\ Ground + Garage\ Cars + Garage\ Area + Wood\ Deck\ Sq\ Ft + Open\ Porch\ Sq\ Ft + Screen\ Porch\ Sq\ Ft + Pool\ Area + Year\ Sold + Exterior\ Quality + Basement\ Quality + Basement\ Condition + Basement\ Exposure + Rating\ of\ Second\ Basement + Heating\ Quality + Kitchen\ Quality + Home\ Functionality + Fireplace\ Quality + Garage\ Quality + Pool\ Quality + \varepsilon$$

Assumptions:

- House Sale price is linearly related with independent variables
- No severe multicollinearity exists
- There are no influential outliers
- Errors are homoscedastic and not autocorrelated

# Results

Negative Coefficients:

- Bedrooms above ground
- Kitchen above ground
- Open porch (sq ft)
- Pool area
- Year sold
- Basement condition
- Garage quality

Adjusted R-Squared: 85.39%

| | | |
|---|---|---|
| Overall Quality | 11,680*** | (786.3) |
| Year Built | 120.4*** | (34.39) |
| Kitchen Quality | 9,409*** | (1,346) |
| Total Rooms Above Ground | 1,977** | (756.9) |
| Full Bathroom | 3,254* | (1,582) |
| Rating of 2nd Finished Basement | 31.47*** | (5.228) |
| Basement Full Bathroom | 6,767*** | (1,499) |
| Garage Size (cars) | 6,471*** | (1,840) |
| Screen Porch (sq ft) | 53.59*** | (10.38) |
| Intercept | 1,358,000 | (872,400) |

Significant codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’

# Conclusions

1. The condition of the home and its newness is significant in determining how high a house is priced.
2. Size is an important feature in determining price as well. The more square footage there is, the higher the price will be.
3. A limitation in this analysis is that the data was collected between 2006 and 2010, and thus data may have been affected by the financial crisis of the time.
4. Further research can include more features related to distances to the city center for example or more macroeconomic data such as percent of population that are family households. This would help build a more detailed analysis.

# References


- Amri S., & Tularam G.A. (2012) Performance of Multiple Linear Regression and Nonlinear Neural Networks and Fuzzy Logic Techniques in Modelling House Prices. *Journal of Mathematics and Statistics 2012*, 8 (4), 419-434.

- Bogin, A.N., & Doerner, W.M. (*2019*) Property Renovations and Their Impact on House Price Index Construction. *Journal of Real Estate Research*: 2019, 41( 2), 249-283.

- Kishor, K.N., & Marfatia, H.A. (2016) Forecasting House Prices in OECD Economies. Journal of Forecasting, 37(2), 170-190.


Thank you!

# Q&A