Explaining House Selling Price Variation Using Linear Regression

Logan Bradley-Trietsch

Purdue University

Work Product - Insight2Profit

March 13, 2020

What Is This Project?

What Is This Project?

- Build a multiple linear regression model to explain house price variation based on past selling data

What Is This Project?

- Build a multiple linear regression model to explain house price variation based on past selling data
- Analyze which variables contribute most to a house's selling price

What Is This Project?

- Build a multiple linear regression model to explain house price variation based on past selling data

- Analyze which variables contribute most to a house's selling price

- Demonstrate potential to make predictions about house prices

Why Is This Important?

Why Is This Important?

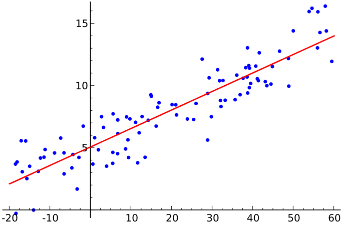
- Useful for homeowners/home buyers/realtors to understand the market and bargain

Why Is This Important?

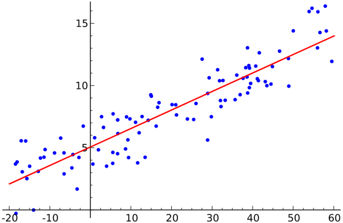
- Useful for homeowners/home buyers/realtors to understand the market and bargain
- Accurate home value estimates are very profitable
  - Ad revenue
  - Real estate bots that buy/sell properties

Overview of Linear Regression

Overview of Linear Regression



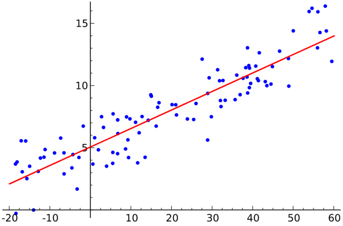
Overview of Linear Regression



Basic model: $y = \beta_1 x + \beta_0$

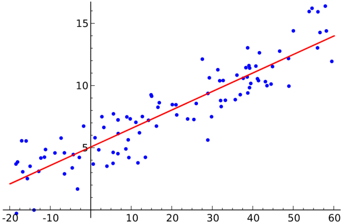
Overview of Linear Regression

Basic model: $y = \beta_1 x + \beta_0$



$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

Overview of Linear Regression



Basic model: $y = \beta_1 x + \beta_0$

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

To find β_0, take the partial with respect to β_0, set equal to 0, solve for β_0:

$$\frac{\partial}{\partial \beta_0} \left[RSS(\beta_0, \beta_1) \right] = 0$$

Verify Assumptions!

Verify Assumptions!

1. Residuals have an average of 0

Verify Assumptions!

1. Residuals have an average of 0
2. Variance is constant

Verify Assumptions!

1. Residuals have an average of 0
2. Variance is constant
3. There is no relationship between residuals

Verify Assumptions!

1. Residuals have an average of 0
2. Variance is constant
3. There is no relationship between residuals

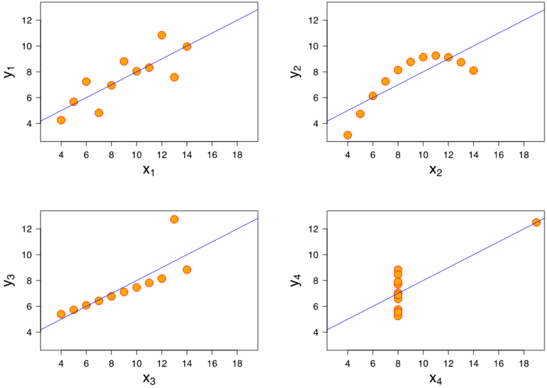


Figure: Different data, **same** model!

What Does the Dataset Look Like?

What Does the Dataset Look Like?

- 1460 houses in Ames, Iowa

What Does the Dataset Look Like?

- 1460 houses in Ames, Iowa
- 79 Explanatory Variables

What Does the Dataset Look Like?

- 1460 houses in Ames, Iowa
- 79 Explanatory Variables
  - Continuous (`LotArea`, `1stFlrSF`)

What Does the Dataset Look Like?

- 1460 houses in Ames, Iowa
- 79 Explanatory Variables
  - Continuous (`LotArea`, `1stFlrSF`)
  - Discrete (`FullBath`, `Kitchen`)

What Does the Dataset Look Like?

- 1460 houses in Ames, Iowa
- 79 Explanatory Variables
  - Continuous (`LotArea`, `1stFlrSF`)
  - Discrete (`FullBath`, `Kitchen`)
  - Factors (`MSZoning`, `Neighborhood`)

What Does the Dataset Look Like?

- 1460 houses in Ames, Iowa
- 79 Explanatory Variables
  - Continuous (`LotArea`, `1stFlrSF`)
  - Discrete (`FullBath`, `Kitchen`)
  - Factors (`MSZoning`, `Neighborhood`)
- Response: Sale price of a home (`SalePrice`)

What Does the Dataset Look Like?

- 1460 houses in Ames, Iowa
- 79 Explanatory Variables
  - Continuous (`LotArea`, `1stFlrSF`)
  - Discrete (`FullBath`, `Kitchen`)
  - Factors (`MSZoning`, `Neighborhood`)
- Response: Sale price of a home (`SalePrice`)
- Lots of missing values and redundant variables (`Utilities`)

Data cleaning

Data cleaning

- Remove redundant variables

Data cleaning

- Remove redundant variables
- Used logarithm transformations on the response (Sale Price) and other continuous predictors in order to ensure constant variance

Data cleaning

- Remove redundant variables
- Used logarithm transformations on the response (Sale Price) and other continuous predictors in order to ensure constant variance

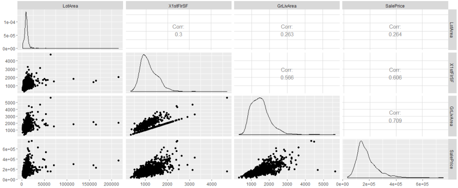


Figure: **Before** transformation

Data cleaning

- Remove redundant variables
- Used logarithm transformations on the response (Sale Price) and other continuous predictors in order to ensure constant variance

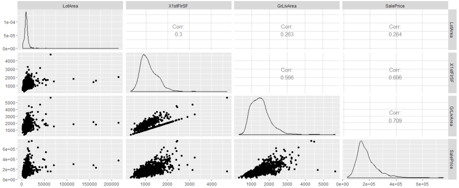


Figure: **Before** transformation

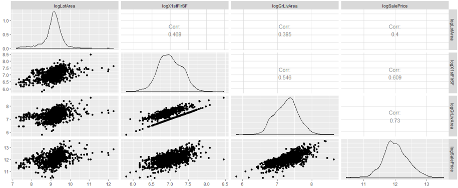


Figure: **After** transformation

Variable Selection

Variable Selection

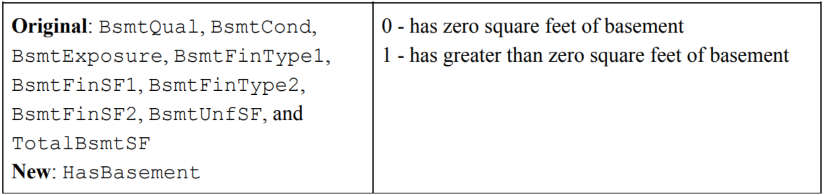
- Out of 79 variables, how did we choose which ones to include in the model?

Variable Selection

- Out of 79 variables, how did we choose which ones to include in the model?
- Created an indicator variable for multicollinear variables

Variable Selection

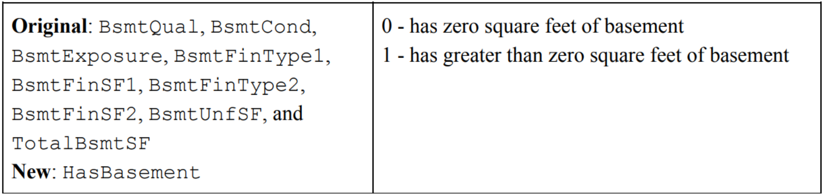
- Out of 79 variables, how did we choose which ones to include in the model?
- Created an indicator variable for multicollinear variables



Original: BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinSF1, BsmtFinType2, BsmtFinSF2, BsmtUnfSF, and TotalBsmtSF **New**: HasBasement	0 - has zero square feet of basement 1 - has greater than zero square feet of basement

Variable Selection

- Out of 79 variables, how did we choose which ones to include in the model?
- Created an indicator variable for multicollinear variables

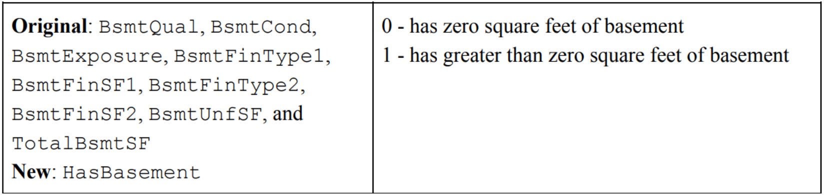


Original: BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinSF1, BsmtFinType2, BsmtFinSF2, BsmtUnfSF, and TotalBsmtSF **New**: HasBasement	0 - has zero square feet of basement 1 - has greater than zero square feet of basement

- Selected the rest of the variables using Bayesian information criterion

Variable Selection

- Out of 79 variables, how did we choose which ones to include in the model?
- Created an indicator variable for multicollinear variables



Original: BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinSF1, BsmtFinType2, BsmtFinSF2, BsmtUnfSF, and TotalBsmtSF **New**: HasBasement	0 - has zero square feet of basement 1 - has greater than zero square feet of basement

- Selected the rest of the variables using Bayesian information criterion
- Analyzed our model using ANalysis Of VAriance (ANOVA) to remove one redundant variable, Roof Material

Verifying Assumptions - Residuals

Verifying Assumptions - Residuals

Remember that we want:

Verifying Assumptions - Residuals

Remember that we want:

1. Average of the residuals is 0

Verifying Assumptions - Residuals

Remember that we want:

1. Average of the residuals is 0
2. Constant variance

Verifying Assumptions - Residuals

Remember that we want:

1. Average of the residuals is 0
2. Constant variance
3. No pattern in residuals

Verifying Assumptions - Residuals

Remember that we want:

1. Average of the residuals is 0
2. Constant variance
3. No pattern in residuals

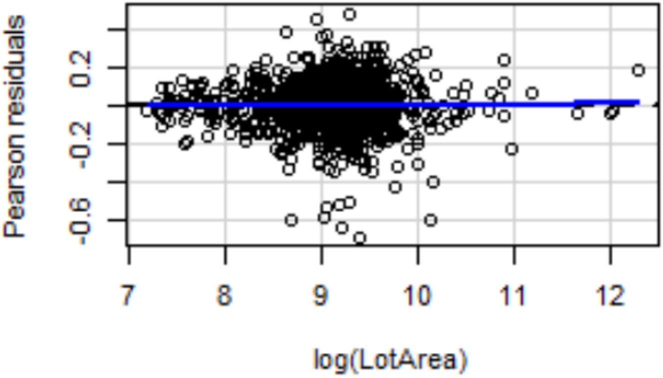


Figure: Average of the residuals is 0; constant variance; no pattern

Verifying Assumptions - Residuals

Verifying Assumptions - Residuals

We have a problem!

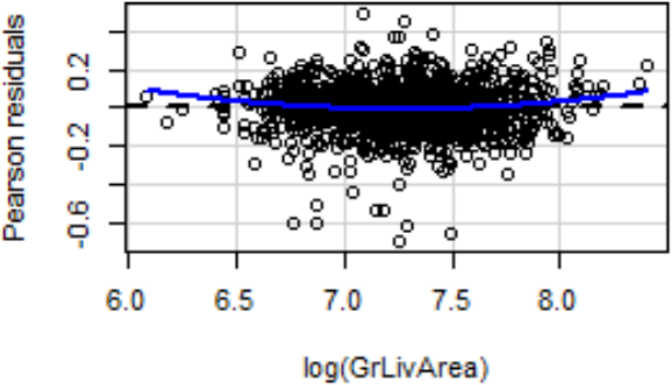


Figure: Average of the residuals is **not** 0, **nonconstant** variance

Verifying Assumptions - Residuals

We have a problem!

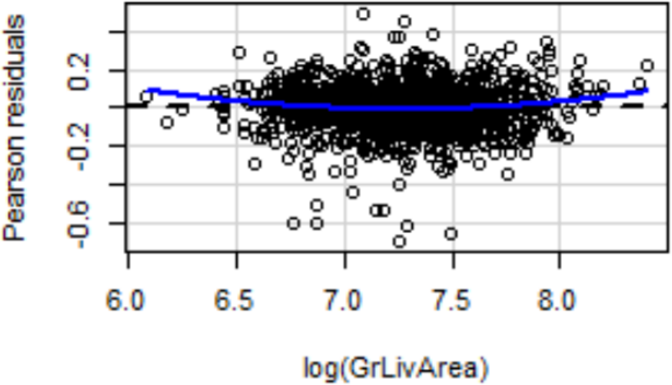


Figure: Average of the residuals is **not** 0, **nonconstant** variance

- We used methods to correct for this nonconstant variance

Effect Size: Above Grade Square Feet

Effect Size: Above Grade Square Feet

- So, what variables affect a house's price?!

Effect Size: Above Grade Square Feet

- So, what variables affect a house's price?!

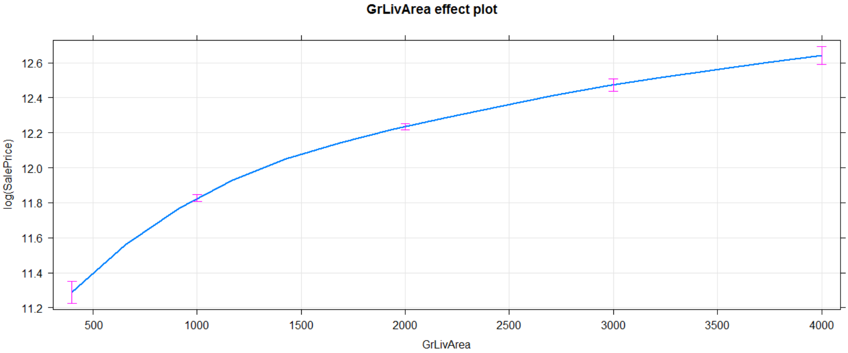


Figure: More above grade sq. ft. ⇒ higher selling price

Effect Size: Above Grade Square Feet

- So, what variables affect a house's price?!

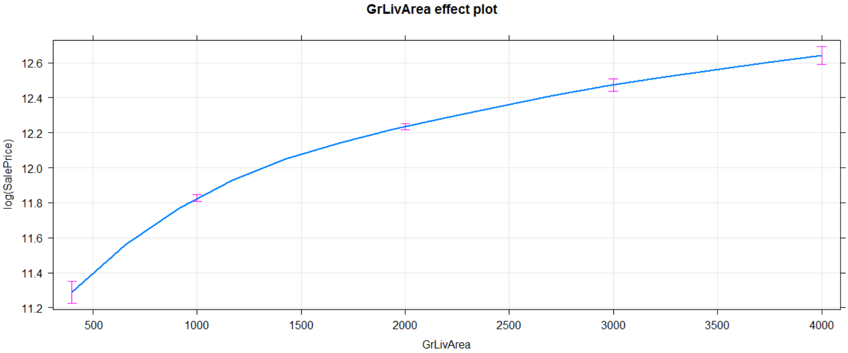


Figure: More above grade sq. ft. ⇒ higher selling price

- Linear relationship of living area vs sale price (remember we are in log scale)

Effect Size: Lot Area

Effect Size: Lot Area

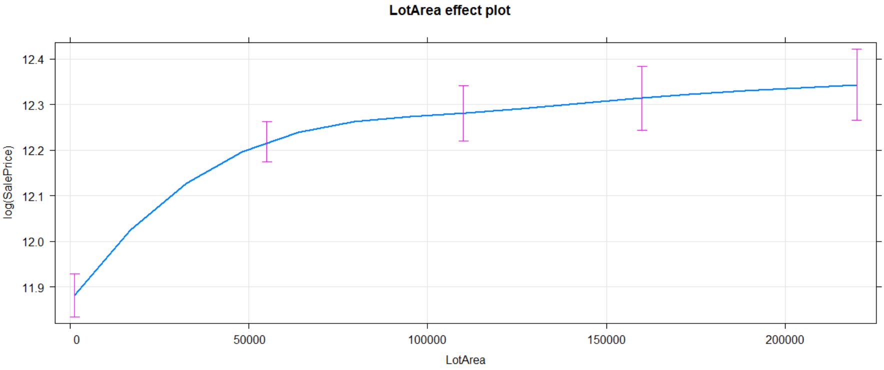


Figure: More lot area ⇒ higher selling price

Effect Size: Lot Area

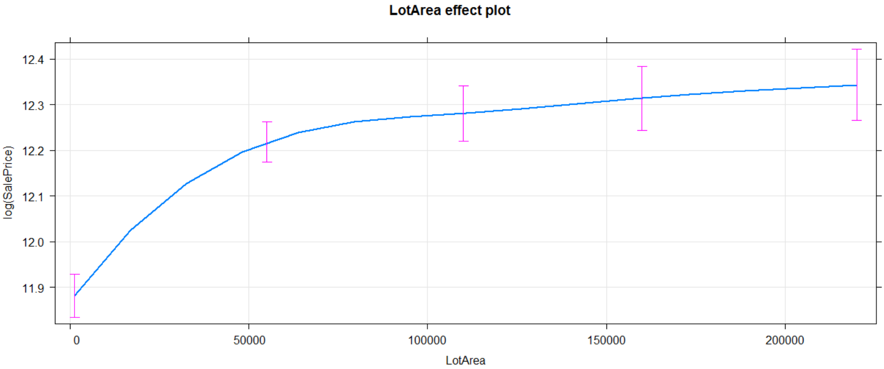


Figure: More lot area ⇒ higher selling price

- Again, approx. linear relationship between lot area and sale price (log scale)

Effect Size: Year of Remodel

Effect Size: Year of Remodel

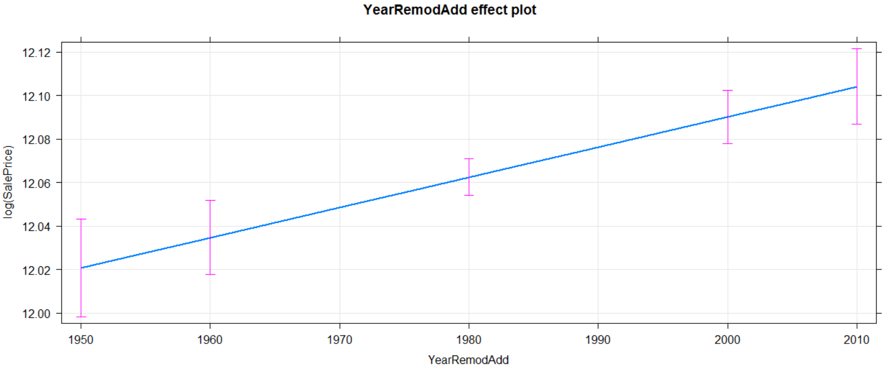


Figure: More recent remodel ⇒ higher selling price

Effect Size: Year of Remodel

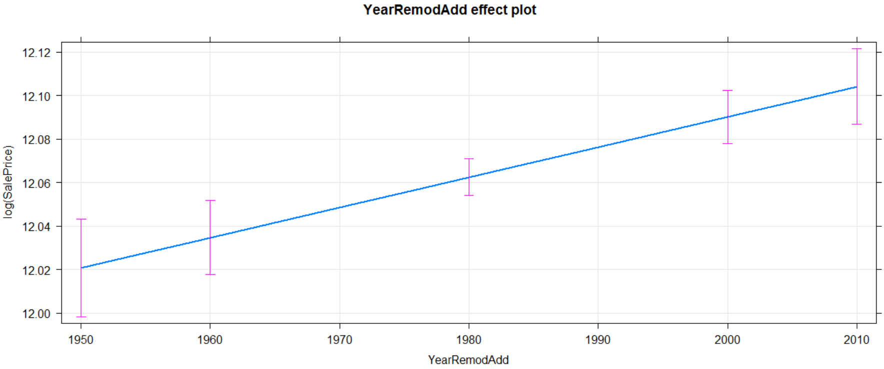


Figure: More recent remodel ⇒ higher selling price

- High variation on very old or very new remodels

Effect Size: Fireplaces

Effect Size: Fireplaces

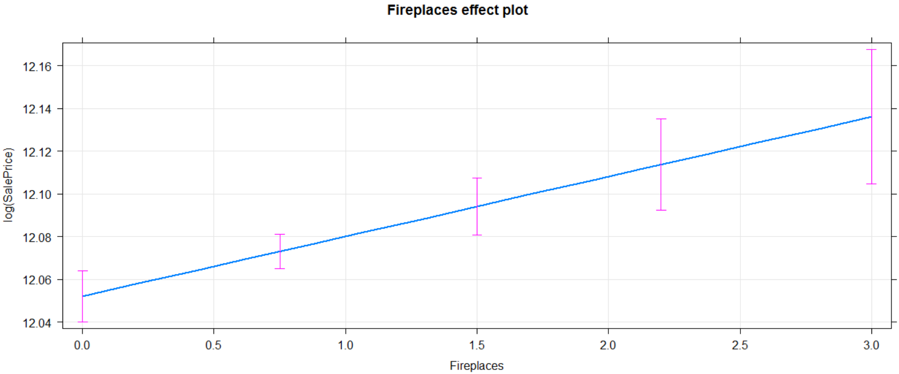


Figure: More fireplaces ⇒ higher selling price

Effect Size: Fireplaces

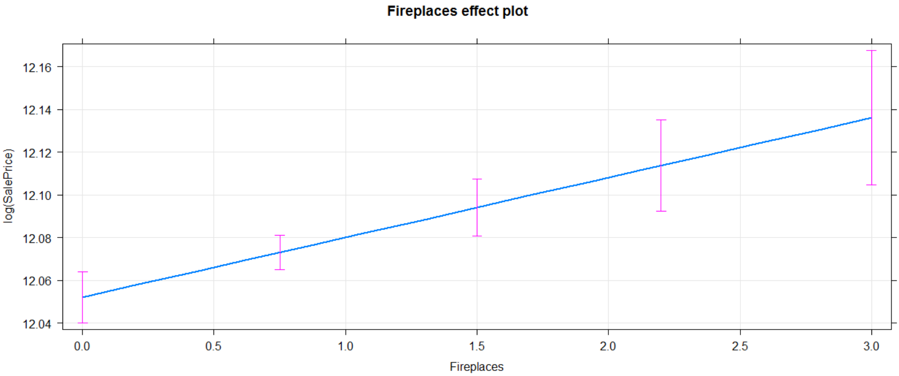


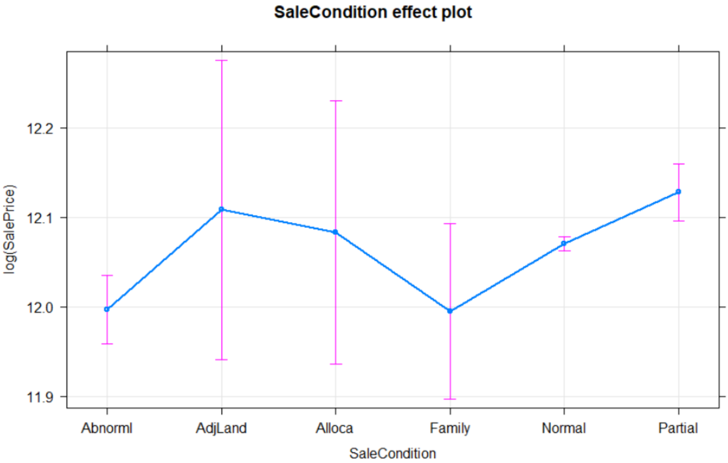
Figure: More fireplaces ⇒ higher selling price

- Increasing variation in sale price as number of fireplaces increases

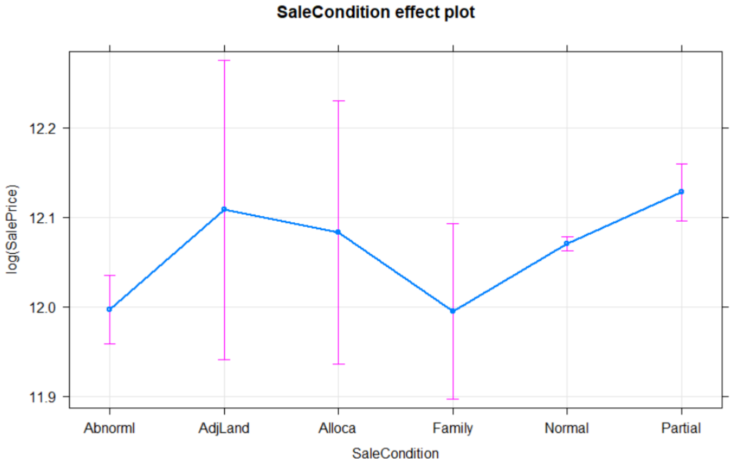
Effect Size: Sale Condition

- Abnorml = trade, foreclosure, short sale
- AdjLand = Adjoining Land Purchase
- Alloca = Allocation - two linked properties with separate deeds, typically condo with a garage unit
- Family = sale between family members
- Normal = normal sale
- Partial = home was not completed when last assessed (associated with new homes)

Effect Size: Sale Condition

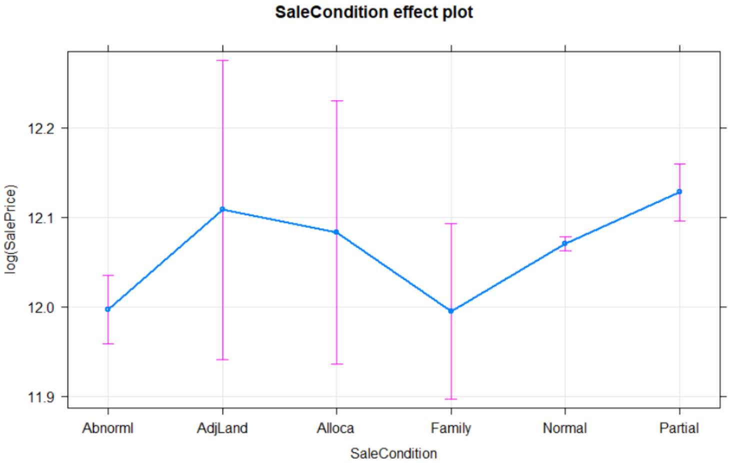


Effect Size: Sale Condition



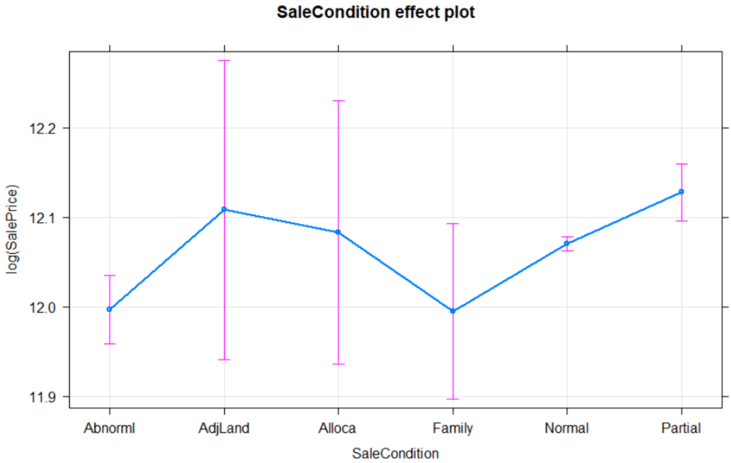
- Family lower than Normal

Effect Size: Sale Condition



- Family lower than Normal
- Normal has very little variation

Effect Size: Sale Condition



- Family lower than Normal
- Normal has very little variation
- Very high variation on allocations and adjoining land purchases!

Effect Size: Overall Quality

Effect Size: Overall Quality

- Rates overall material and finish of the house

Effect Size: Overall Quality

- Rates overall material and finish of the house

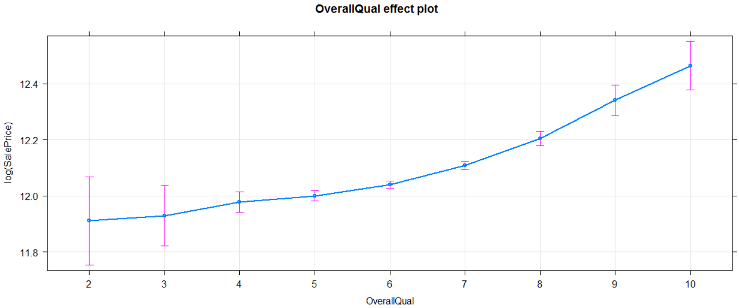


Figure: Higher overall quality ⇒ higher sale price

Effect Size: Overall Quality

- Rates overall material and finish of the house

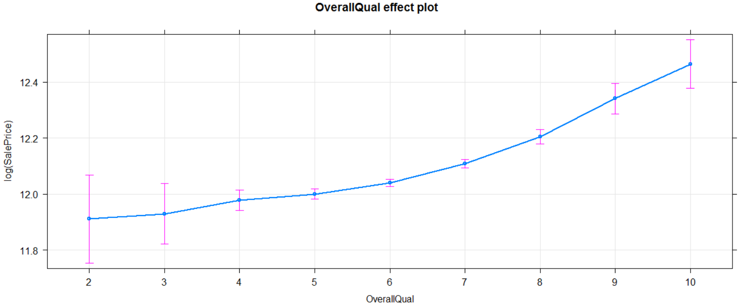


Figure: Higher overall quality ⇒ higher sale price

- Subjective measure surprisingly shows clear relationship?!

Effect Size: Overall Quality

- Rates overall material and finish of the house

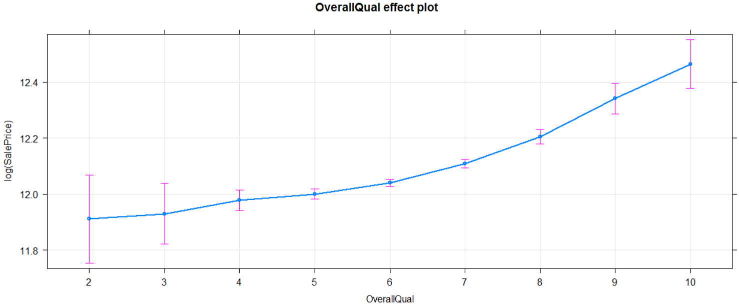


Figure: Higher overall quality ⇒ higher sale price

- Subjective measure surprisingly shows clear relationship?!
- Again, higher variation in sale price in extremes

Effect Size: Overall Condition

- Rates the overall condition of the house

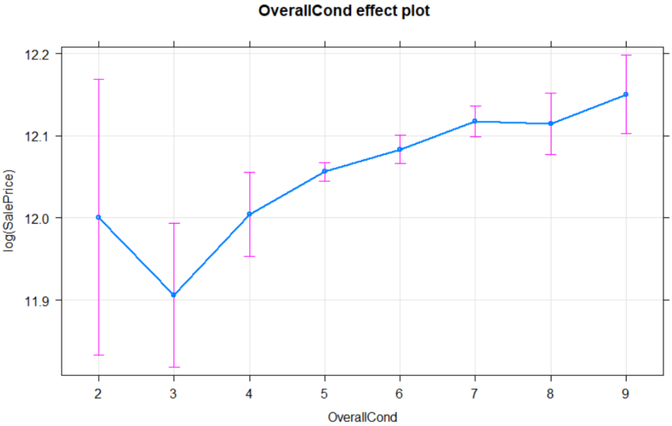


Figure: Higher overall condition ⇒ higher sale price

Effect Size: Overall Condition

- Rates the overall condition of the house

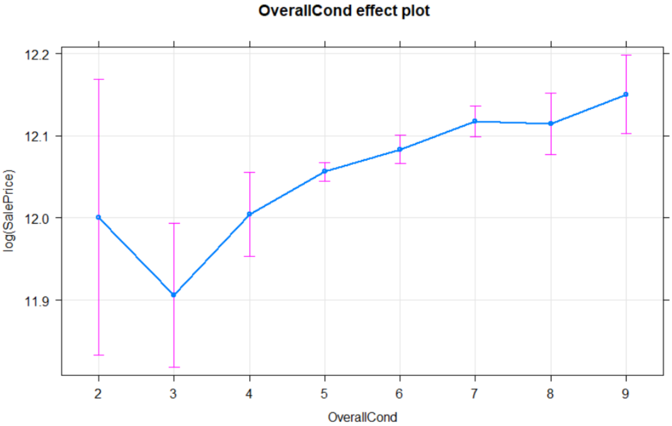


Figure: Higher overall condition ⇒ higher sale price

- This is the weird relationship expected in a subjective measure

Effect Size: Central Air

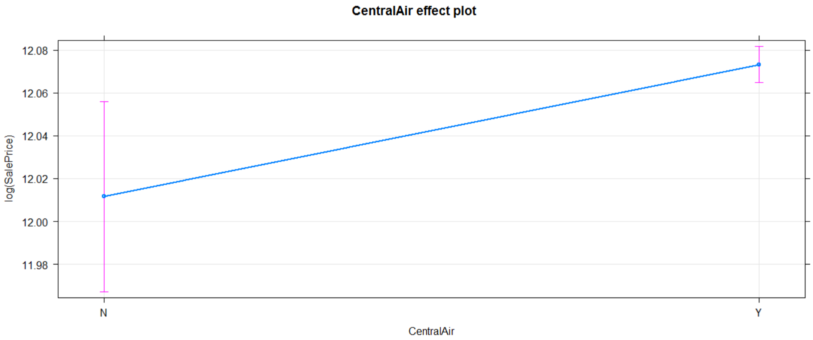


Figure: Presence of central air ⇒ higher sale price

Effect Size: Central Air

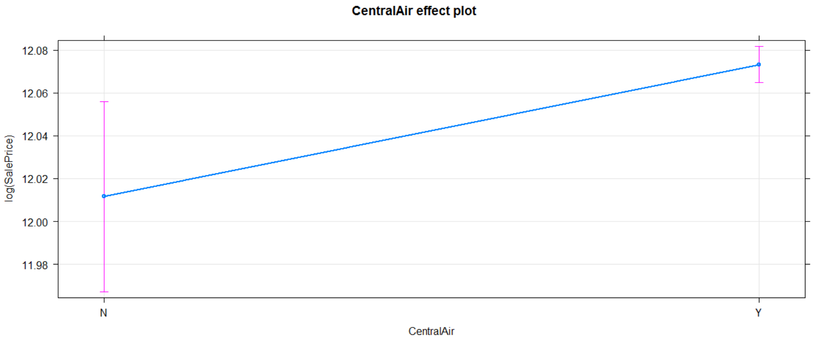


Figure: Presence of central air ⇒ higher sale price

- More variation in "No" than in "Yes"

Effect Size: Has Basement

- We created this variable to simplify the model

Effect Size: Has Basement

- We created this variable to simplify the model

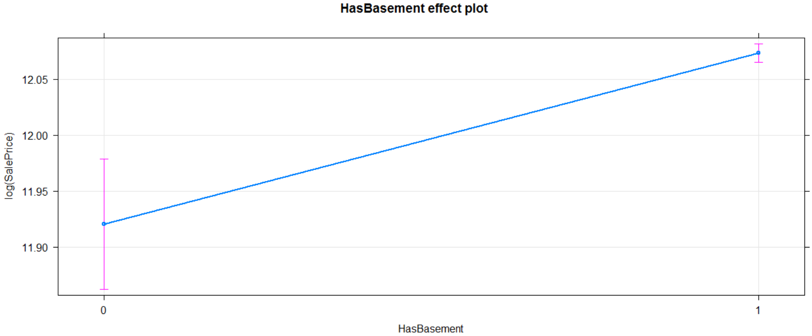


Figure: Presence of basement ⇒ higher sale price

Effect Size: Has Basement

- We created this variable to simplify the model

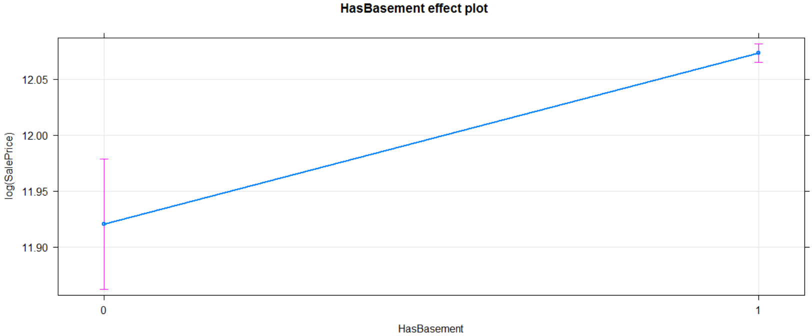


Figure: Presence of basement ⇒ higher sale price

- Again, more variation in "No" than in "Yes"

Effect Size: Has Garage

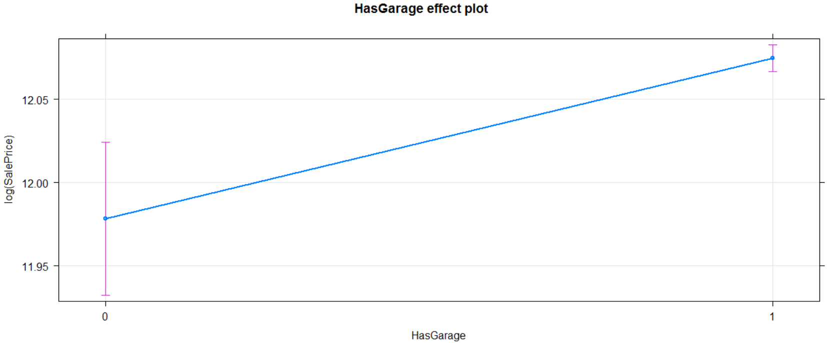


Figure: Presence of garage ⇒ higher sale price

Effect Size: Has Garage

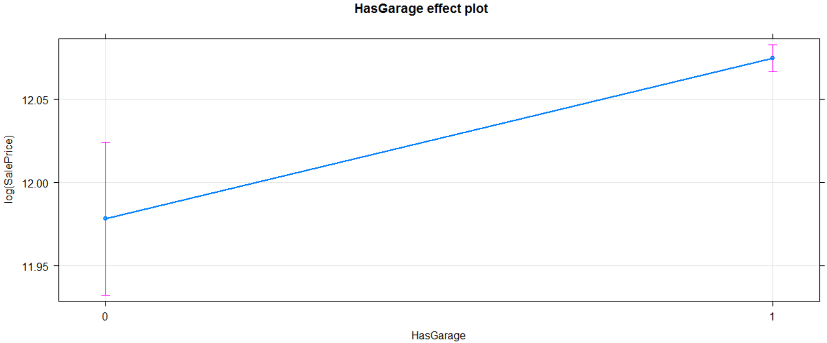


Figure: Presence of garage ⇒ higher sale price

- Again, more variation in "No" than in "Yes"

Unexpected Relationships: Has 2nd Floor

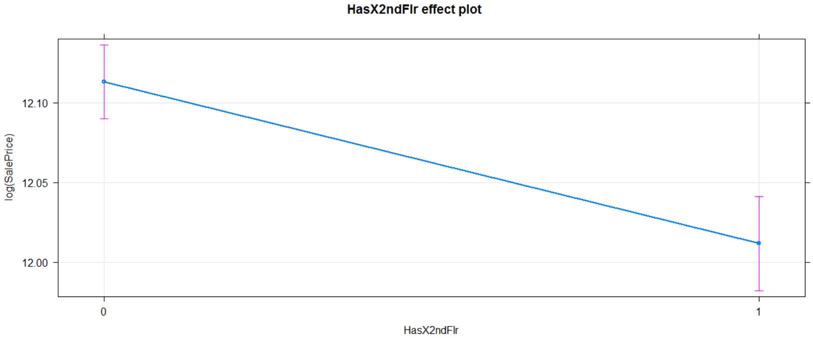


Figure: Presence of 2nd floor ⇒ lower sale price?

Unexpected Relationships: Has 2nd Floor

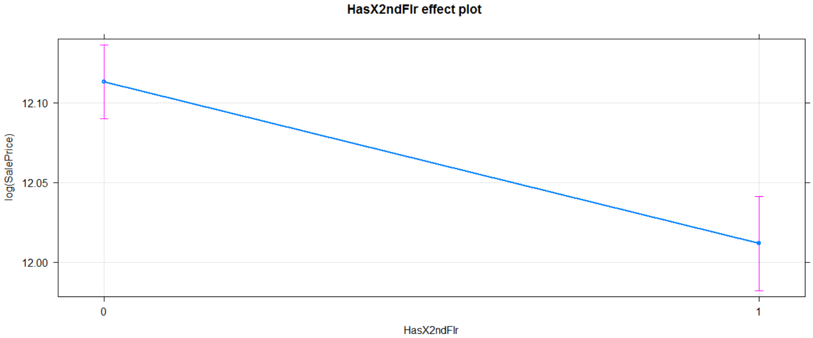


Figure: Presence of 2nd floor ⇒ lower sale price?

- We attribute this unexpected relationship to other variables that already encode this information (e.g., Above ground living area)

Unexpected Relationships: Has 2nd Floor

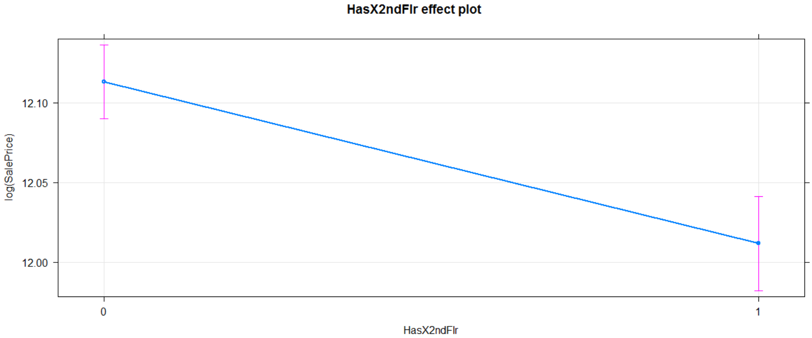


Figure: Presence of 2nd floor ⇒ lower sale price?

- We attribute this unexpected relationship to other variables that already encode this information (e.g., Above ground living area)
- 2nd floor homes have a higher average sale price

Unexpected Relationships: Kitchens Above Grade

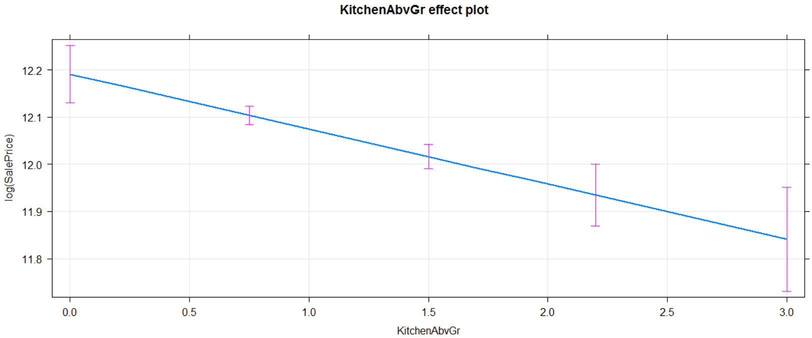


Figure: More kitchens ⇒ lower sale price?

Unexpected Relationships: Kitchens Above Grade

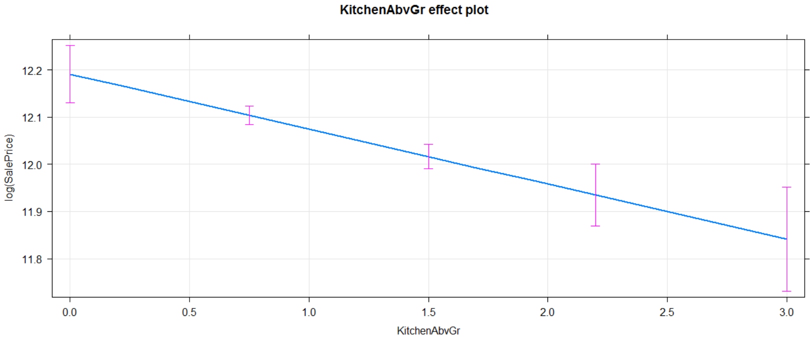


Figure: More kitchens ⇒ lower sale price?

- We attribute this unexpected relationship to other variables that already encode this information (e.g., Kitchen Quality)

Effect Size: Kitchens Quality

- Ex - excellent
- Gd - good
- TA - typical/average
- Fa - fair

Effect Size: Kitchens Quality

- Ex - excellent
- Gd - good
- TA - typical/average
- Fa - fair

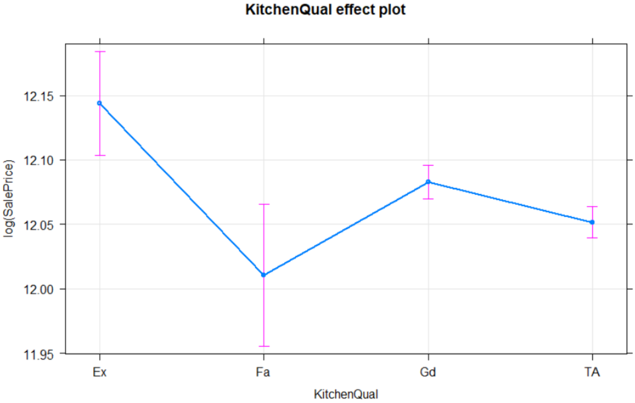


Figure: Higher quality kitchens ⇒ higher sale price

Predictive Potential

Predictive Potential

- R^2 is .91

Predictive Potential

- R^2 is .91
- 91% of the variation in sale price is accounted for by the explanatory variables

Predictive Potential

- R^2 is .91
- 91% of the variation in sale price is accounted for by the explanatory variables
- There is high predictive potential for Ames, Iowa from 2006-2010

Predictive Potential

- R^2 is .91
- 91% of the variation in sale price is accounted for by the explanatory variables
- There is high predictive potential for Ames, Iowa from 2006-2010

Limitations of This Project

Limitations of This Project

- Data was collected from 2006-2010

Limitations of This Project

- Data was collected from 2006-2010
- Only 1,460 observations!

Limitations of This Project

- Data was collected from 2006-2010
- Only 1,460 observations!
- Only in Ames, Iowa

Limitations of This Project

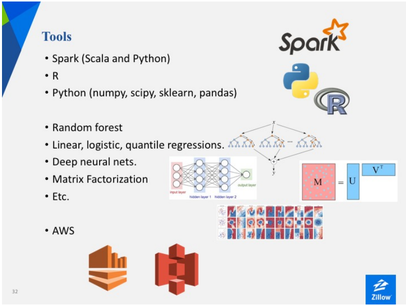
- Data was collected from 2006-2010
- Only 1,460 observations!
- Only in Ames, Iowa
- Potentially misleading results or self-fulfilling predictions

Limitations of This Project

- Data was collected from 2006-2010
- Only 1,460 observations!
- Only in Ames, Iowa
- Potentially misleading results or self-fulfilling predictions
- Correlation $\neq$ Causation

Limitations of This Project

- Data was collected from 2006-2010
- Only 1,460 observations!
- Only in Ames, Iowa
- Potentially misleading results or self-fulfilling predictions
- Correlation $\neq$ Causation
- I'm not Zillow!
  - There are more accurate (and complex) methods



Tools

- Spark (Scala and Python)
- R
- Python (numpy, scipy, sklearn, pandas)

- Random forest
- Linear, logistic, quantile regressions.
- Deep neural nets.
- Matrix Factorization
- Etc.

- AWS

Conclusion

Conclusion

- Goal: What factors affect house prices? How do these factors affect house prices?

Conclusion

- Goal: What factors affect house prices? How do these factors affect house prices?
- Verified that our model is used legitimately and all assumptions are addressed

Conclusion

- Goal: What factors affect house prices? How do these factors affect house prices?
- Verified that our model is used legitimately and all assumptions are addressed
- Certain factors have clear relationships (Living Area, Central Air)

Conclusion

- Goal: What factors affect house prices? How do these factors affect house prices?
- Verified that our model is used legitimately and all assumptions are addressed
- Certain factors have clear relationships (Living Area, Central Air)
- Others are quite complicated (Number of Kitchens, Kitchen Quality)

Conclusion

- Goal: What factors affect house prices? How do these factors affect house prices?
- Verified that our model is used legitimately and all assumptions are addressed
- Certain factors have clear relationships (Living Area, Central Air)
- Others are quite complicated (Number of Kitchens, Kitchen Quality)
- Our model shows potential to predict sale price on similar homes, but other methods are better at prediction

Conclusion

- Goal: What factors affect house prices? How do these factors affect house prices?
- Verified that our model is used legitimately and all assumptions are addressed
- Certain factors have clear relationships (Living Area, Central Air)
- Others are quite complicated (Number of Kitchens, Kitchen Quality)
- Our model shows potential to predict sale price on similar homes, but other methods are better at prediction

Questions?