Determinants of House Prices

Diego Deviez

2024-06-07

Background & Descriptive Statistics

Context and Research Question: The Ames Housing Prices data set contains many different variables tied to the sale of a house to help understand each individual residency. Some of the variables included privde information such as the type of dwelling, type of foundation, garage location, lot size, etc…

The goal is to use the Ames Housing Prices data to derive a model that can predict the sale price of a house. Answering this question is useful because it allows us to gain a deeper understanding of determinants for house prices. Specifically, answering this question can help determine predictors of house prices for real estate professionals, planners and developers, and financial institutions to aid them in understanding market trends, demands, opportunities, and minimize risk.

library(tidyverse)  
library(memisc)  
library(pander)  
library(Hmisc)  
library(writexl)  
  
hp <- read\_csv("R1housingprices.csv")

Data and Variables–Descriptive Statistics:

Dependent Variable: SalePrice (Quantitative|Continuous) Independent Variables: LotArea (Quantitative|Continuous), GrLivArea (Quantitative|Continuous) YearBuilt (Quantitative|Discrete), BldgType(Qualitative|Discrete), OverallCond(Quantitative|Discrete), KitchenQual (Qualitative|Discrete)

describe(hp[,c("SalePrice", "LotArea", "GrLivArea", "YearBuilt", "OverallCond")])

## hp[, c("SalePrice", "LotArea", "GrLivArea", "YearBuilt", "OverallCond")]   
##   
## 5 Variables 1460 Observations  
## --------------------------------------------------------------------------------  
## SalePrice   
## n missing distinct Info Mean Gmd .05 .10   
## 1460 0 663 1 180921 81086 88000 106475   
## .25 .50 .75 .90 .95   
## 129975 163000 214000 278000 326100   
##   
## lowest : 34900 35311 37900 39300 40000, highest: 582933 611657 625000 745000 755000  
## --------------------------------------------------------------------------------  
## LotArea   
## n missing distinct Info Mean Gmd .05 .10   
## 1460 0 1073 1 10517 5718 3312 5000   
## .25 .50 .75 .90 .95   
## 7554 9478 11602 14382 17401   
##   
## lowest : 1300 1477 1491 1526 1533, highest: 70761 115149 159000 164660 215245  
## --------------------------------------------------------------------------------  
## GrLivArea   
## n missing distinct Info Mean Gmd .05 .10   
## 1460 0 861 1 1515 563.1 848 912   
## .25 .50 .75 .90 .95   
## 1130 1464 1777 2158 2466   
##   
## lowest : 334 438 480 520 605, highest: 3627 4316 4476 4676 5642  
## --------------------------------------------------------------------------------  
## YearBuilt   
## n missing distinct Info Mean Gmd .05 .10   
## 1460 0 112 1 1971 33.88 1916 1925   
## .25 .50 .75 .90 .95   
## 1954 1973 2000 2006 2007   
##   
## lowest : 1872 1875 1880 1882 1885, highest: 2006 2007 2008 2009 2010  
## --------------------------------------------------------------------------------  
## OverallCond   
## n missing distinct Info Mean Gmd   
## 1460 0 9 0.814 5.575 1.111   
##   
## Value 1 2 3 4 5 6 7 8 9  
## Frequency 1 5 25 57 821 252 205 72 22  
## Proportion 0.001 0.003 0.017 0.039 0.562 0.173 0.140 0.049 0.015  
##   
## For the frequency table, variable is rounded to the nearest 0  
## --------------------------------------------------------------------------------

hp %>% count(BldgType)

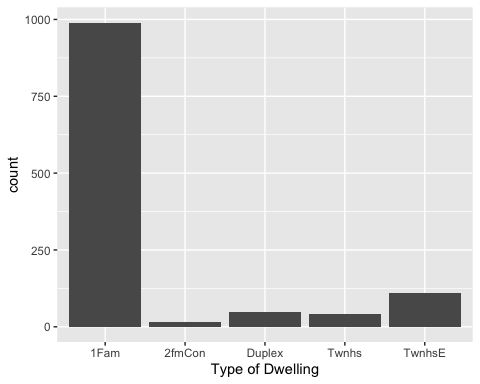
## # A tibble: 5 × 2  
## BldgType n  
## <chr> <int>  
## 1 1Fam 1220  
## 2 2fmCon 31  
## 3 Duplex 52  
## 4 Twnhs 43  
## 5 TwnhsE 114

Descriptive Visualizations: Removed outliers to generalized graphs and better understand trends.

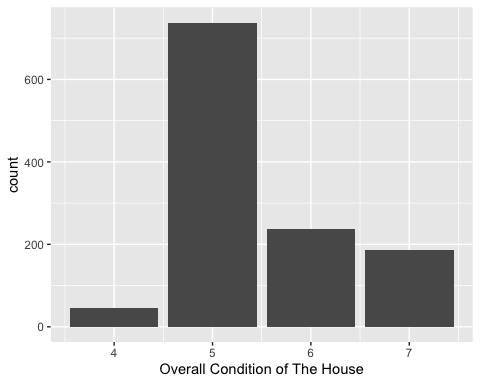
hpoutliers <- hp %>%   
 filter(LotArea < 17500, SalePrice < 350000, GrLivArea < 3000, YearBuilt > 1907, between(OverallCond,4,7))  
  
write\_xlsx(hpoutliers,"hpoutliers.xlsx")  
  
ggplot(data = hpoutliers) +   
 geom\_bar(mapping = aes(x = YearBuilt), binwidth = 0.6) +  
 labs(x = "Original Construction Date")



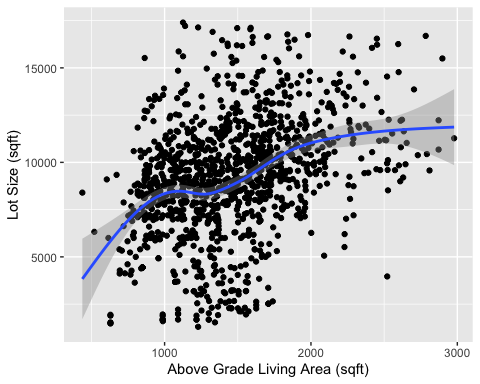
ggplot(data = hpoutliers) +   
 geom\_bar(mapping = aes(x = BldgType), binwidth = 0.5) +  
 labs(x = "Type of Dwelling")



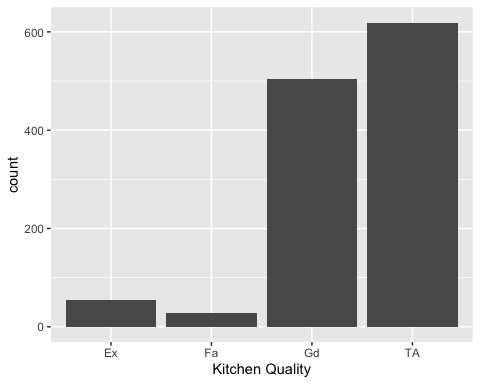
# 1Fam Single-family Detached   
# 2FmCon Two-family Conversion; originally built as one-family dwelling  
# Duplx Duplex  
# TwnhsE Townhouse End Unit  
# TwnhsI Townhouse Inside Unit  
  
ggplot(data = hpoutliers) +   
 geom\_bar(mapping = aes(x = OverallCond), binwidth = 0.5) +  
 labs(x = "Overall Condition of The House")



ggplot(data = hpoutliers) +   
 geom\_point(mapping = aes(x = GrLivArea, y = LotArea)) +  
 geom\_smooth(mapping = aes(x = GrLivArea, y = LotArea)) +  
 labs(x = "Above Grade Living Area (sqft)", y = "Lot Size (sqft)")



ggplot(data = hpoutliers) +   
 geom\_bar(mapping = aes(x = KitchenQual), binwidth = 0.6) +  
 labs(x = "Kitchen Quality")



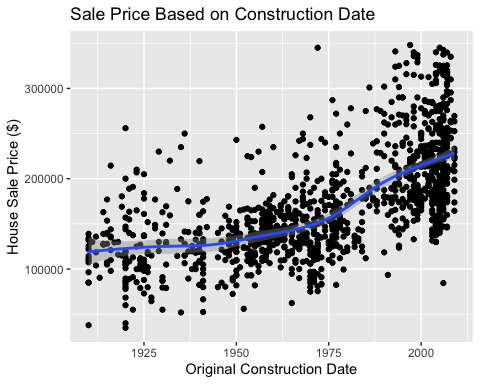
# Ex Excellent  
# Gd Good  
# TA Typical/Average  
# Fa Fair  
# Po Poor

Hypotheses: The selection of variables was based on general knowledge and theory. One of the things I feel justify the price of a house are property characteristics such as LotArea, larger lots are often priced higher; YearBuilt, older homes may be priced lower due to its high maintenance; BldgType, the type of home can impact the price; GrLivArea, larger living space can impact prices. Moreover, the homes overall condition (OverallCond) can also have a huge impact on price. I hypothesize that all of these variables will have a positive relationship with Sale Price and explain a large mount of the variance.

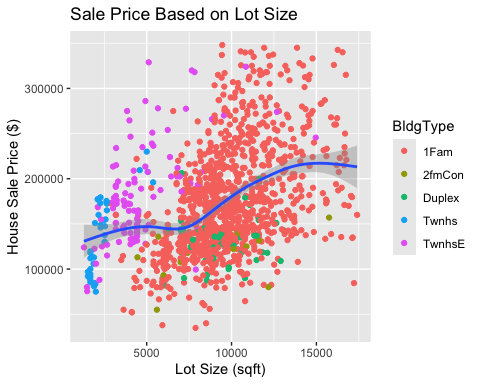
Model Results

Predictive Visualizations:

ggplot(data = hpoutliers, mapping = aes(x = YearBuilt, y = SalePrice)) +  
 geom\_point() +  
 geom\_smooth() +  
 scale\_y\_continuous(labels = function(y) format(y, scientific = FALSE)) +  
 labs(title = "Sale Price Based on Construction Date", x = "Original Construction Date", y = "House Sale Price ($)")



ggplot(data = hpoutliers, mapping = aes(x = LotArea, y = SalePrice)) +  
 geom\_point(mapping = aes(color = BldgType)) +  
 geom\_smooth() +  
 scale\_y\_continuous(labels = function(y) format(y, scientific = FALSE)) +  
 labs(title = "Sale Price Based on Lot Size",x = "Lot Size (sqft)", y = "House Sale Price ($)" )



# 1Fam Single-family Detached   
# 2FmCon Two-family Conversion; originally built as one-family dwelling  
# Duplx Duplex  
# TwnhsE Townhouse End Unit  
# TwnhsI Townhouse Inside Unit  
  
ggplot(data = hpoutliers, mapping = aes(x = GrLivArea, y = SalePrice)) +  
 geom\_point() +  
 geom\_smooth() +  
 facet\_wrap(~OverallCond, ncol = 3) +  
 scale\_y\_continuous(labels = function(y) format(y, scientific = FALSE)) +  
 labs(title = "Sale price Based on Above Grade Living Area per Overall Condition Grade of the House", x = "Above Grade Living Area (sqft)", y = "House Sale Price ($)")



Regression Model: After further analysis, I decided to remove BldgType from my regression model as it made little impact on variance explained.

model1 <- lm( formula = SalePrice ~ YearBuilt + OverallCond + BldgType, data = hpoutliers)  
model2 <- lm( formula = SalePrice ~ YearBuilt + LotArea + OverallCond, data = hpoutliers)  
model3 <- lm( formula = SalePrice ~ YearBuilt + OverallCond + LotArea + GrLivArea, data = hpoutliers)  
  
  
require(memisc)  
m123table <- mtable('Model1' = model1,  
 'Model 2' = model2,  
 'Model 3' = model3,  
 summary.stats = c('R-Squared','F','p','N'))  
pander(m123table)

* **Model1**:
  + **coef**: *-2854098*, *1519*, *5951*, *-5985*, *-34309*, *-59443*, *-29303*, *108186*, *52.47*, *1782*, *10769*, *6495*, *6919*, *4492*, *-26.38*, *28.96*, *3.339*, *-0.5558*, *-5.283*, *-8.591*, *-6.523*, *2.554e-121*, *3.12e-140*, *0.000868*, *0.5785*, *1.512e-07*, *2.639e-17*, *1.012e-10*, *-3066352*, *1417*, *2454*, *-27112*, *-47052*, *-73018*, *-38117*, *-2641844*, *1622*, *9448*, *15142*, *-21567*, *-45868* and *-20490*
  + **sumstat**:

Table continues below

| sigma | r.squared | adj.r.squared | F | numdf | dendf | p | logLik |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 43194 | 0.4677 | 0.4651 | 175.6 | 6 | 1199 | 2.59e-160 | -14580 |

| deviance | AIC | BIC | N |
| --- | --- | --- | --- |
| 2.237e+12 | 29176 | 29217 | 1206 |

* + **contrasts**:
    - **BldgType**: contr.treatment
  + **xlevels**:
    - **BldgType**: *1Fam*, *2fmCon*, *Duplex*, *Twnhs* and *TwnhsE*
  + **call**: lm(formula = SalePrice ~ YearBuilt + OverallCond + BldgType, data = hpoutliers)
* **Model 2**:
  + **coef**: *-2649848*, *1380*, *6.366*, *7064*, *97384*, *47.28*, *0.3731*, *1641*, *-27.21*, *29.19*, *17.06*, *4.305*, *2.002e-127*, *4.699e-142*, *1.264e-58*, *1.807e-05*, *-2840909*, *1288*, *5.634*, *3845*, *-2458786*, *1473*, *7.098* and *10283*
  + **sumstat**:

Table continues below

| sigma | r.squared | adj.r.squared | F | numdf | dendf | p |
| --- | --- | --- | --- | --- | --- | --- |
| 40729 | 0.5256 | 0.5244 | 443.9 | 3 | 1202 | 4.756e-194 |

| logLik | deviance | AIC | BIC | N |
| --- | --- | --- | --- | --- |
| -14511 | 1.994e+12 | 29031 | 29057 | 1206 |

* + **contrasts**:
  + **xlevels**:
  + **call**: lm(formula = SalePrice ~ YearBuilt + LotArea + OverallCond, data = hpoutliers)
* **Model 3**:
  + **coef**: *-2183601*, *1101*, *8952*, *2.882*, *72.93*, *71661*, *35.14*, *1186*, *0.2889*, *2.192*, *-30.47*, *31.34*, *7.551*, *9.974*, *33.27*, *1.498e-151*, *4.53e-158*, *8.512e-14*, *1.449e-22*, *1.428e-172*, *-2324195*, *1032*, *6626*, *2.315*, *68.63*, *-2043008*, *1170*, *11278*, *3.448* and *77.23*
  + **sumstat**:

Table continues below

| sigma | r.squared | adj.r.squared | F | numdf | dendf | p | logLik |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 29392 | 0.7531 | 0.7523 | 916 | 4 | 1201 | 0 | -14117 |

| deviance | AIC | BIC | N |
| --- | --- | --- | --- |
| 1.038e+12 | 28245 | 28276 | 1206 |

* + **contrasts**:
  + **xlevels**:
  + **call**: lm(formula = SalePrice ~ YearBuilt + OverallCond + LotArea + GrLivArea, data = hpoutliers)

summary(model3)

##   
## Call:  
## lm(formula = SalePrice ~ YearBuilt + OverallCond + LotArea +   
## GrLivArea, data = hpoutliers)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -118618 -17799 -1751 13592 126852   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.184e+06 7.166e+04 -30.471 < 2e-16 \*\*\*  
## YearBuilt 1.101e+03 3.514e+01 31.343 < 2e-16 \*\*\*  
## OverallCond 8.952e+03 1.186e+03 7.551 8.51e-14 \*\*\*  
## LotArea 2.881e+00 2.889e-01 9.974 < 2e-16 \*\*\*  
## GrLivArea 7.293e+01 2.192e+00 33.273 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 29390 on 1201 degrees of freedom  
## Multiple R-squared: 0.7531, Adjusted R-squared: 0.7523   
## F-statistic: 916 on 4 and 1201 DF, p-value: < 2.2e-16

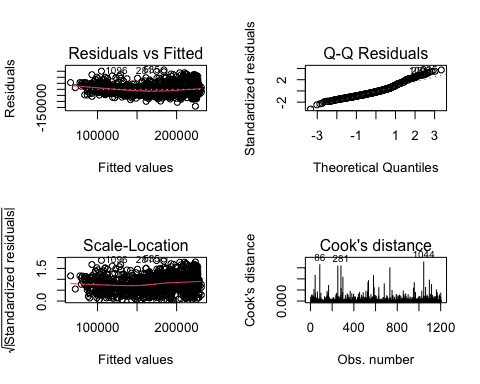
Regression Model3 Results:

The p-value for the model and the p-value for for the coefficients are both less than .05. We see the three stars. This model is highly significant.

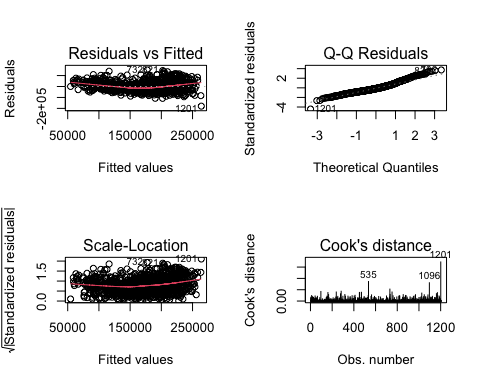
The coefficients are the following: YearBuilt (1,101), LotArea (2.881), GrLivArea (72.93), OverallCond (8,952). These numbers are all positive, meaning they all have a positive relationship with Sale Price. i.e. As the Year Built increases, the Sale Price also increases.

The R-squared value shows how much of the variance is explained by this model. The R-squared for this specific model explains 75.31% of the variance in Sale Price, making it a pretty effective model.

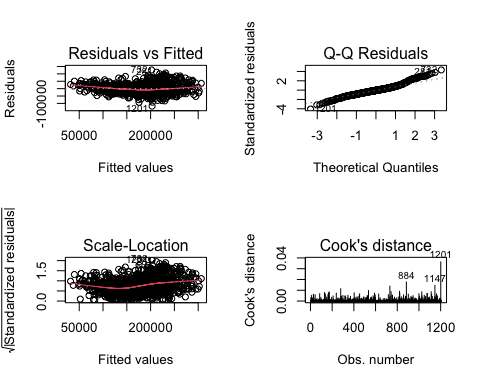
par(mfrow=c(2,2))  
plot(model1, which = c(1,2,3,4))



par(mfrow=c(2,2))  
plot(model2, which = c(1,2,3,4))



par(mfrow=c(2,2))  
plot(model3, which = c(1,2,3,4))

 Regression Model3 Robustness Tests:

1. Residuals vs Fitted Plot: The residuals are randomly scattered for the most part around a line that is somewhat horizontal.
2. Q-Q Residuals Plot: The are many points that lie perfectly along the reference line. However, there are significant deviations from the line towards both ends, suggesting departure from normality.
3. Scale-Location Plot: Not normally distributed, meaning the homoscedasticity of the regression model is not safe.
4. Cooks Distance: There are a couple of outliers observations that might require some investigation.

Overall, the data used to construct this model would need further investigation to improve these robustness tests.

Conclusions

Discussion: The original hypothesis, which stated that ” all of these variables will have a positive relationship with Sale Price and explain a large mount of the variance.”, was supported by the analysis. The regression model showed a significant and positive relationship between the independent variables(construction date, lot size in square feet, overall condition of the house, and above grade living area in square feet) and the dependent variable (sale price of the house). These different variables are responsible for explaining 75.31% of the sale price of the house. This is a significant effect that can effectively aid in understanding the sale price of a house.

The practical implications of this work are substantial. Real estate businesses can leverage these findings to develop pricing models that accurately reflect market values based on these key characteristics. For example, understanding that newer homes or those with larger living areas tend to fetch higher prices can inform renovation and marketing strategies, making properties more attractive to buyers.

Business Value: One way to use the findings to create value is to use the identified key variables to create a pricing model. Creating a an effective pricing model could lead to more competitive and accurate pricing, potentially increasing average sale prices by 5-10%. Another way to create value from these findings is to use them for strategic renovations. Focus on upgrading properties based on the most impactful variables, such as overall condition and living area, in order to increase the sale price. Such renovations could increase property values by 10-15%, providing a high return on investment.

Key Takeaways: Accurate Pricing: Using the identified key variables can help set more accurate property prices. Value-Adding Renovations: Strategic renovations can significantly boost property values and attractiveness.

Limitations and Future Research: The regression robustness tests leave soem room to be desired. Upon further analysis, the 4 graphs showed that the data used to construct the regression model could use further investigation. Another limitation would entail the data collection. Collecting data on more detailed property features and neighborhood specifics could enhance the model. For example, data on interior amenities would provide a more comprehensive view of property values.

Conclusion: The analysis done on the Ames Hosuing Pricing dataset confirms that key property characteristics chosen (construction date, lot size in square feet, overall condition of the house, and above grade living area in square feet) significantly influence house prices. By implementing pricing models and strategic renovations based on these findings, real estate businesses can enhance their competitiveness and profitability. While there are limitations in the current model and data, future research with more detailed data and more robust models can further refine these predictions and provide even greater value.