How to House Hunt: A Novel Analysis of Physical Attributes Influencing Sale Price in Ames, Iowa 2006-2010

House prices have been growing rapidly every year. According to Zillow, the typical American home price increased by 13.2% compared to last year (Adamczyk, 2021). There are many factors that determine the value of a residential property, which are used by real estate appraisers in complex algorithms to pair a property with an implicit price.

Although the housing market – which is defined as the supply and demand of homes – is one of the most critical determinants in sale price ("6 Factors That Influence a Home's Value", 2017), the lone customer has no significant influence over this variable. One person cannot sway the overall national economy. Instead, I am interested in investigating the internal and external physical attributes of a home that may influence sale price, which can be more directly understood and modified by the prospective homeowner.

Previous studies on the housing market have cited numerous factors that have a statistically significant impact on determining sale price. Thomas Thibodeau (2003) studied house prices in Dallas between 1991-1993 through examining hedonic house price equations, which are frequently used to estimate the market value of a subject property or to mark residential property values to market. He found that a property's size, age, and price of neighboring houses were the most important determinants in a specific unit's price. Other literature has identified current and expected interest rates to be the most impactful determinants of consumer purchasing behavior towards buying houses (Dua, 2008). This study assessed data based on 1984-2005 survey research from the University of Michigan.

However, due to the impact of the national housing market crash in late 2007 (Mortgage, n.d.), I am interested in examining an updated dataset that contains data from before and after the catastrophic event. The housing crisis undoubtably altered the American real estate landscape, so the factors influencing pricing – and thus, customer decision making – are likely also transformed. The goal of my study was to distill this information into a modern, simplified set of variables for customers to examine during their house hunt.

In this study, I used 2006-2010 housing data from Ames, Iowa to investigate the factors that are most statistically significant in predicting sale price. These factors included location, style, material finish and quality, square footage, garage size, and other variables. I aimed to make the complex real estate market more accessible to the average prospective homeowner by identifying these most influential qualitative and quantitative factors contributing to a dwelling's valuation. This would enable consumers to more effectively assess whether the home is over- or underpriced and thus whether it is worthy of an offer. I hypothesized that house style, building class, overall quality, home functionality, year built, garage area, and size were the most important factors in determining house price.

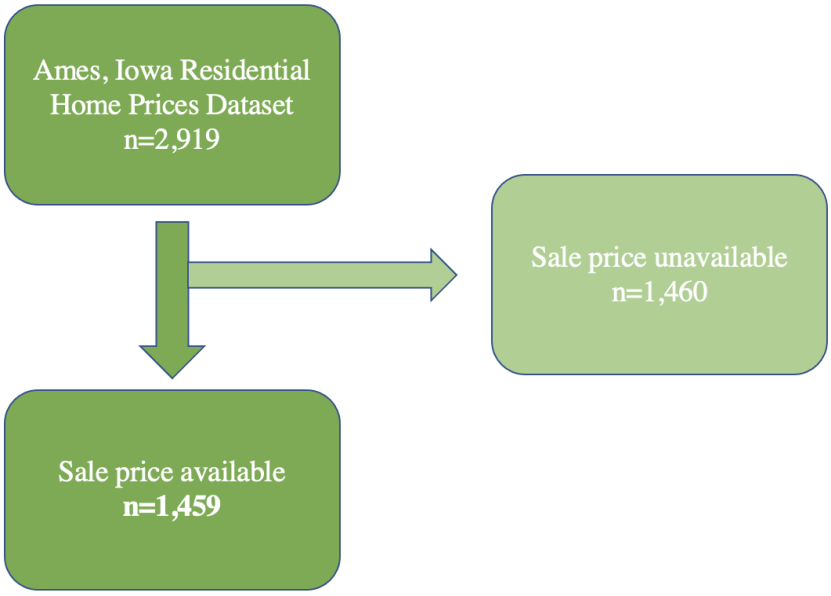
Methods

The use of data and measurements are discussed below.

Data

I collected the dataset from kaggle.com (De Cock, 2011). The data consisted of 2,919 residential home property sales that occurred in Ames, Iowa between 2006 and 2010 as well as the physical attributes of the home. My outcome variable was the sale price in dollars of each home. The exposure variables included house style, house zone, overall quality and condition, home functionality, age, garage area, and house size (see Table 1). For my analysis, I used these statistics only for observations with available sale prices, which reduced my sample size from 2,919 to 1,459 homes (see Figure 1).

Figure 1
Data reduction diagram



Measures

Previous research studies have identified a variety of factors influencing the sale price of a house, including interest rates (Dua, 2008) and the housing market ("6 Factors That Influence a Home's Value", 2017). But instead of these abstract concepts, I wanted to investigate the physical attributes of a home that have the largest impact on sale price, which would make purchasing a home more convenient for customers. Below, I highlight the variables that I found most important from my dataset. All of the variables besides "Sale Price" were used as predictors. For each predictor, I did not exclude any sales from my total of 1,459 observations to avoid losing any information.

Sale Price. Sale price was my continuous outcome variable which indicated the property sale price in American dollars. There were no sale prices of $0, so I did not exclude any observations. Prices ranged from $34,900 to $755,000.

House Style. House style was defined as the size of the house in terms of the floor plan. I used five dichotomous variables to examine house style, using one-story homes as the reference

variable: one for two-story buildings, one for one and one-half stories with second level finished and unfinished, one for split foyer and split-level homes, and a final variable for two and one-half story buildings with second level finished and unfinished.

MS Zoning. The zoning variable indicated the general zoning classification of the property. I transformed the eight different zones into two dummy variables: "residential" or "non-residential," with "residential" as the reference variable. The "residential" zones included floating village, high density, low-density, low-density park, and medium density residences. The "non-residential" zones included agriculture, commercial, and industrial properties.

Building Type. This variable complemented the "House Style" variable in describing the type of property being analyzed. I transformed this variable into four dichotomous variables with "Single family detached" homes as the reference variable: one for "Two family conversion," one for "Duplex," and one for "Townhouse End Unit" and "Townhouse Inside Unit."

Overall Rank. Overall Rank summed the dataset variables "OverallQual," which described the material and finish of the house, and "OverallCond," which described the overall condition of the house. Each of these variables was given a quantitative score from 1 to 10: "Very Poor" to "Very Excellent." I summed these variables to provide a more comprehensive and concise score for the condition and quality of the house that ranged from 2 to 20.

Total Floor Square Footage. For the total square footage variable, I summed the square footages of the first floor ("1stFlrSF") and second floor ("2ndFlrSF") to obtain a total square footage of the unit. This variable ranged in size from 334 square feet to 5,642 square feet.

Functional. This variable referred to the home's functionality rating based on the severity of building damages. This would provide the buyer with context on how much work needed to be done to make the unit livable. I created five dichotomous variables with "typical functionality" (no damages) as the reference variable: one for minor deductions 1 and 2, one for moderate deductions, one for major deductions 1 and 2, and one for severe damage and salvage only.

Age. This variable was the age of the building derived from the original year of construction ("YearBuilt"). This age variable ranged from 11 to 149 years. Many past studies found this to be an extremely important confounder when determining sale price. Since I investigated this dataset in 2021, I decided to subtract the year built from the current year to calculate the age of the house as variable "Age."

Garage Area. This continuous variable represented the square footage of the property's garage. Garage area was quantified in square feet with measurements ranging from 100 to 1,488.

Table 1

Descriptive Statistics: Outcome and Explanatory Variables of Ames, Iowa Homes in 2006-2010

Characteristic	Total Sample
	N = 1459
Sale Price	
Mean	179k (dollar)
St. Deviation	16.5k (dollar)
MS Zoning	
Residential Low Density	1114 (76.3%)
Residential Medium Density	242 (16.6%)
Other	103 (7%)
BldgType	
Single-family	1205 (82.6%)
Townhouse End	113 (7.7%)
Other	141 (9.7%)
HouseStyle	
One Story	745 (50%)
Two Story	427 (29.3 %)
Other	287 (19.7%)
OverallRank*	
Mean	11.6
St. Deviation	1.73
Age*	
Mean	49.6
St. Deviation	30.4
TotalFlrSF*	
Mean	1.48k (square feet)
St. Deviation	485 (square feet)
Functional	
Typical Function	1357 (93.0 %)
Minor Deduction2	36 (2.5%)
Other	66 (4.5 %)
GarageArea	
Mean	473 (square feet)
St. Deviation	217 (square feet)

Note: OverallRank was derived from the sum of OverallCond and OverallQual. TotalFlrSF was derived from the sum of 1stFlrSF and 2ndFlrSF. Age was derived from the subtraction of BuiltYear from 2021. Since Age is a quantitative variable, the table represents the mean age (49.3 years) and standard deviation (30.4) of the participants. The median of respondents was 48 years.

Analysis

The primary hypothesis, that there would be significant impacts of my confounders on my outcome variable, "Sale Price," was tested using an OLS regression model to account for both categorical and continuous variables. I first used R statistical software (version 3.5.1) to clean my dataset and create my dichotomous variables. Next, I checked assumptions for multicollinearity, normal distribution, and homoscedasticity before conducting my analysis.

Then, I used a multiple linear regression model to predict the estimator coefficients for each of my independent variables and used adjusted R-square values to inspect the proportion of variance for a dependent variable that is explained by the independent variable's estimators. I used a significance level of 0.05 to identify statistically significant predictors.

Results

To determine the key predictors of sale price, I ran a multiple linear regression with all chosen independent variables. In this analysis, the dependent variable was the sale price in dollars for each residence. My predictor variables were garage area (in square feet), age, overall rank, total floor square footage, house style, zone type, building type, and functionality. To satisfy the normality assumptions for multiple linear regression, I took the log of sale price before conducting my analysis. In addition, I investigated the scatterplot of residuals to ensure that the assumption of homoscedasticity was satisfied, and that there was no correlation between the independent variables and the residuals. I was wary of multicollinearity because two predictors – total floor square footage (first floor and second floor area) and overall rank (quality rating and condition rating) – were calculated by summing existing variables. To combat multicollinearity, I calculated the variance inflation factors for each of the eight predictors and found that they were all less than 10. Furthermore, I did not identify any outliers in my dataset. My residuals were normally distributed, as were my quantitative variables (see Table 1) except for "Age." Taking the log transformation of this predictor still did not satisfy the normality assumption, so this is a limitation of my study that should be addressed in future research.

My results showed that the linear combination of independent variables accounted for 83.52% of the variance in sale price (R^2 =0.8352, F(16, 1443) = 463, p<.001). From these variables, all eight predictors were significant with at least p<0.05 (see Table 2).

Table 2

Multiple Linear Regression Results: Predicting Sale Price from Eight Predictor Variables

Predictor Variable	b	SE	t	p
(Intercept)	10.960	0.071	154.609	p<.001***
GarageArea	0.000	0.000	9.617	p<.001***
Age	-0.194	0.008	-22.923	p<.001***
OverallRank	0.065	0.003	21.019	p<.001***
TotalFlrSF	0.000	0.000	31.742	p<.001***
HouseStyle				
OneStory	NA	NA	NA	NA
OneHalfStory	-0.079	0.016	-5.074	p<.001***
TwoStory	-0.133	0.012	-11.450	p<.001***
TwoHalfStory	-0.147	0.040	-3.630	p<.001***
Split	-0.016	0.017	-0.901	0.368
MSZoning				
Nonresidential	NA	NA	NA	NA
Residential	0.383	0.053	7.256	p<.001***
BldgType				
Single	NA	NA	NA	NA
Two	-0.021	0.030	-0.694	0.488
Duplex	-0.122	0.024	-5.055	p<.001***
Townhouse	-0.035	0.017	-2.105	0.035*
Functionality				
Typical	NA	NA	NA	NA
MinorDed	-0.038	0.021	-1.765	0.078
ModDed	-0.112	0.043	-2.603	0.009**
MajorDed	-0.147	0.038	-3.830	p<.001***
Severe	-0.454	0.163	-2.792	0.0053**

Note: N=1,443. Model R^2=0.837, F(16, 1443) = 463, p<.001, Adjusted R^2=0.8352, b=unstandardized multiple regression coefficient, SE=standard error, t=t-test statistic, p=probability value, *p<.05, **p<.01,***p<.001

I found that residential properties on average cost 46.67% more than nonresidential properties (b=0.383, p<.001) when all other variables are held constant. As age increases by one year, the sale price decreases by 21.4% (b= -0.194, p<.001). The effect of house style was statistically significant (p<.001) for three out of six housing categories. By using single story houses as a baseline and holding other variables constant, I determined that two-story houses are associated with a decrease in sale price by 14.2% (b= -0.133, p<.001), while one and one-half story houses are associated with a decrease in sale price by only 8.22% (b= -0.079, p<.001). The estimated effect of functionality on sale price increases as the severity of deductions increases compared to a baseline of typical functionality. Having a “severe” classification decreases sale price by 57.46% (b= -454, p<.001) compared to 11.85% (b= -0.038, p=0.009) for “moderate”

deductions. The effect of functionality for minor deductions, the effect of building type for two family homes, and the effect of house style for split houses did not have a statistically significant impact on sale price.

Complete results of this regression are presented in Table 2.

Discussion

My findings support my hypothesis: a majority of the variables I identified as most likely to be influential had statistically significant effects on sale price. As expected, as the age of the home increases, the sale price decreases, and as quality and condition ratings increase, the sale price increases. This was also consistent with Thibodeau's (2003) findings that a property's size and age were the most important determinants in a specific unit's price. Furthermore, as the severity of deductions increases, the house will require more renovations, which would decrease the price significantly. Although the effect of garage area and total floor square footage were significant and positively associated with sale price, it is surprising that the estimated effect is so small. This may be because a $1ft^2$ increase is not large enough of a difference to make an impact on the sale price, and in fact, that increase may not even be practically significant. Another surprising finding was that buying a home with more than one story would be less expensive than one with a single story, which may be due to a higher proportion of costly amenities, such as kitchens and baths. Future studies should include the ratio of kitchen and bathroom square footage to total home square footage to address this relationship.

Nationwide housing prices have skyrocketed in 2021 by 17.2% compared to 2020 (Forbes, 2021). Therefore, cost effective and evidence-based purchasing strategies are a necessity for any prospective homeowner. With the findings of this study in mind, homeowners should look most closely at these key predictors of potential houses and use them as points of comparison. Distilling the most impactful determinants of pricing to eight variables will make it much easier to make decisions such as buying a two story vs. one story house or how large of a garage a customer's budget can reasonably accommodate.

My study does have limitations. Similar to past studies, the data that I analyzed is siloed to a specific time period and place. In my case, the data from Ames, Iowa between 2006 and 2010 may not be representative of real estate properties across the country. When considering my results, it is important to note that my confounders may affect sale prices differently in bigger cities, other parts of the country, or in different time periods. Furthermore, my dataset did not include interest rates or the condition of the housing market, which other studies have identified to be significant determinants of sale price (Chan, 2016). One suggestion for future studies is to look at the same confounders decade by decade for 50 or 60 years. This would allow us to contextualize the real estate market and highlight trends that may be occurring over time.

References

6 factors that influence a home's value. (2017, August 7). Inman. https://www.inman.com/2017/08/07/6-factors-that-influence-a-homes-value/

Adamczyk, A. (2021, June 16). *The typical home price is up a record 13.2% compared to last year, according to Zillow*. CNBC. https://www.cnbc.com/2021/06/16/typical-us-home-price-up-record-13point2percent-compared-to-last-year.html

Campisi, N. (2021, August 9). *Will The Housing Market Cool Off Soon? Here's What Experts Predict*. Forbes Advisor. https://www.forbes.com/advisor/mortgages/when-will-the-housing-market-cool-off/

Chan, S., Dastrup, S., & Ellen, I. G. (2016). Do Homeowners Mark to Market? A Comparison of Self-Reported and Estimated Market Home Values During the Housing Boom and Bust. *Real Estate Economics*, *44*(3), 627–657. https://doi.org/10.1111/1540-6229.12103

Dua, P. (2008). Analysis of Consumers' Perceptions of Buying Conditions for Houses. *The Journal of Real Estate Finance and Economics*, *37*(4), 335–350. https://doi.org/10.1007/s11146-007-9084-0

Edelman, E. R., van Kuijk, S. M. J., Hamaekers, A. E. W., de Korte, M. J. M., van Merode, G. G., & Buhre, W. F. F. A. (2017). Improving the Prediction of Total Surgical Procedure Time Using Linear Regression Modeling. *Frontiers in Medicine*, *4*, 85. https://doi.org/10.3389/fmed.2017.00085

Mortgage, A. H. (n.d.). *When Did the Housing Bubble Burst?* Retrieved October 29, 2021, from https://www.amerifirst.com/amerifirst-blog/bid/83428/when-did-the-housing-bubble-burst

Suriyadeepan, R. (2020, December 30). Exploratory Data Analysis of Iowa Housing Price Prediction Problem. *Analytics Vidhya*. https://medium.com/analytics-vidhya/exploratory-data-analysis-of-iowa-housing-price-prediction-problem-3d50a016797a

Thibodeau, T. G. (2003). Marking Single–Family Property Values to Market. *Real Estate Economics*, *31*(1), 1–22. https://doi.org/10.1111/j.1080-8620.2003.00055.x