Housing Prices in Ames, Iowa

Final Project Report: Team Blue

MTH 361: Applied Statistics

April 28, 2023

Alec Uyematsu

Caden Deutscher

Devam Patel

Table of Contents

[Introduction 3](#_Toc133776674)

[Background Research/Motivation 3](#_Toc133776675)

[Research Questions 3](#_Toc133776676)

[Hypothesis Statements 4](#_Toc133776677)

[Dataset Description 4](#_Toc133776678)

[Data Exploration 5](#_Toc133776679)

[Statistical Method/Point Estimates 7](#_Toc133776680)

[Sampling: 7](#_Toc133776681)

[Simple Linear Regression: 7](#_Toc133776682)

[Performing Simple Linear Regression: 10](#_Toc133776683)

[Simple Linear Regression Interpretation: 10](#_Toc133776684)

[Multiple Linear Regression: 11](#_Toc133776685)

[Performing Multiple Linear Regression: 13](#_Toc133776686)

[Multiple Linear Regression Interpretation: 13](#_Toc133776687)

[Results and Discussions 14](#_Toc133776688)

[Conclusion 14](#_Toc133776689)

[Appendix A: R Code for Diagrams 16](#_Toc133776690)

[Works Cited 19](#_Toc133776691)

# **Introduction**

## Background Research/Motivation

## 

The group of students developing this research project are all senior students at the University of Portland. This means that we have had the realization that life is getting closer and closer to hitting us. This includes making all the difficult decisions that come with that, such as, deciding what house to buy. With this information, we have set out to arm ourselves with some statistically backed rationale for whether we should purchase an older/renovated homes or not.

Doing a little bit of research, it is not hard to see that typically home value is influenced by age. It seems almost common sense that older homes will see more general wear and tear and ultimately a higher degradation of their condition; however, it is also clear that is not always the case. Some homes despite their age contain novelties that are worth being protected. Could those older novelties result in better care despite the fact that there is a higher age? Additionally, it is important to still recognize that things age no matter what. There is a lifetime for the roof, ceiling, wiring…etc. No matter age catches up. It will be interesting to see if our data from Iowa demonstrates these factors.

Additionally, there is a lack of consensus around the internet on the viability and worth of old homes. Some people believe that old homes have character (a good thing to some), and are built better. Others disagree and think the opposite considering a character a bad thing. How will these pieces of information play out in our discoveries? Only time will tell. Will the relatively subjective conception of condition lead to older homes showing a less negative condition in comparison to newer homes than might be expected, or will we see the expected fall off in condition for older homes. Despite the fact that there are few to no formal articles on this subject it is clear that the internet is divided in-favor of younger homes. Maybe this project can provided some light on this topic.

There are also additional factors outside of simply age that affect the overall condition of a home. The period in which homes went through can significantly increase/decrease the speed at which a home age. In other words, a home may not age at a linear pace. The owners who owned the home affect its aging. The materials used in its construction affect the aging rate, and region and the weather that the home experienced can affect its longevity.

## Research Questions

**The Question:** What is the relationship between the age of a home and its overall condition? This question can be examined with a few different variables.

**Additional variables to include:**

* *Neighborhood: Neighborhood – used to sort data*
* *Year Built: Year.Built*
* *The last year of remodel: Year.RemodAdd*
* *Overall Condition: Overall.Cond*

If(Year.Built == Year.RemodAdd) then there is no remodel

## Hypothesis Statements

We hypothesized that as the Year.Built and Year.RemodAdd variables get closer to present date, the overall condition of the house will improve.

Part 1:

H0\_YearBuild: Year built does not, on average, affect the general condition of the house.

H0\_YearBuild\_Symbol: beta\_yearBuilt = 0

HA\_YearBuilt: New houses, on average, have a better general condition than older houses.

HA\_YearBuild\_Symbol: beta\_yearBuilt > 0

Part 2

H0\_RemodAdd: Year remolded/additions and year built does not, on average, affect the general condition of the house.

H0\_RemodAdd\_Symbol: : beta\_yearBuilt = 0 and beta\_remodelDate = 0

HA\_RemodAdd: New remodels/additions to the house and more recently built houses, on average, have a better general condition than older houses.

HA\_RemodAdd\_Symbol: beta\_yearBuilt > 0 and beta\_remodelDate > 0

## Dataset Description

The Dataset used in this project was created by Dean De Cock from Truman State University. It was created from data passed to him by the Ames, Iowa Assessor’s office. The dataset was created from all the residential homes sold from 2006 until 2010 in Ames. Originally, the dataset contained 3970 observations and 113 different variables, but the dataset was shrunk in order to only provide the variables information that could be understood without additional information or knowledge. The dataset created by Dean De Cock contains 2930 observations and 80 variables.

# **Data Exploration**

What is the purpose of these variables in our research? There are two aspects to our research. One, we will be doing a general dive into the relationship on age vs condition. Specifically, we will be looking at the age from the year a house was built and comparing that to how its condition and the condition of other houses are affected. This is straight forward and involves the Year Built being the explanation variable and the overall condition being the response variable. The second observation we will be doing is looking at how a remodel disrupts the association of age on condition. This means that we will have two explanation variables year built and last year of remodel which will have a response variable of overall condition. Together, these two aspects of our research will give us a full grasp of how the age of a building affects its overall condition.

Calendar

Description automatically generated with low confidence

**Figure 1:** Year Built vs. Overall Condition

Table

Description automatically generated

**Figure 2:** Year of Remodel vs. Overall Condition

Chart, scatter chart

Description automatically generated

**Figure 3:** Year Built and Year of Remodel vs Overall Condition

### Statistical Method/Point Estimates

### Sampling:

One potential option is to use the entire sample provided in the data. This is okay if the data can be compared without regard for variables like initial home price, the neighborhood where the house is, or other variables that could create bias in how the house is treated over the years.

The other option is to split the data into different sampling distributions based on those variables pointed out above. We could take a stratified sampling of the neighborhoods, or homes of different starting costs and run the statistical methods on those samples. This is something that could eliminate bias of care.

Ultimately for the purpose of this project we used the whole data set for our sample, but this section was designed to demonstrate some of the thoughts that we grappled with when designing our tests.

### Simple Linear Regression:

Checking OLS Conditions

1st Condition:

**Independent Random Samples**: For this specific data, we know that it is derived from the purchases of homes from 2006-2010. Because of this we will assume that the data is random and independent.

2nd Condition:

**Residuals Normally Distributed and Homoscedastic:** According to the figure below, the residuals are normally distributed which satisfies this condition.

Chart, histogram

Description automatically generated

**Figure 4:** Residual Histogram

3rd Condition:

**No Pattern in Plotting Residuals:** Based on the figure below the residuals do have a pattern. This does not mean that we are unable to perform a linear regression. It does mean that a more complexed test of the data might make more sense for this specific study, but for this course we will stick to a linear regression model, although it is important to note that a different model might be more suited for the data.

Chart, box and whisker chart

Description automatically generated

**Figure 5:** Residual Plot

*A picture containing chart

Description automatically generated*

**Figure 6**: Scatter Plot with Linear Regression

*(Comparing age with condition)*

### Performing Simple Linear Regression:

**Call:  
lm(formula = Overall.Cond ~ Year.Built, data = ames\_df)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-5.5302 -0.3891 -0.1072 0.4320 3.6624   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 32.2802434 1.2446698 25.93 <2e-16 \*\*\*  
Year.Built -0.0135527 0.0006313 -21.47 <2e-16 \*\*\*  
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
  
Residual standard error: 1.033 on 2928 degrees of freedom  
Multiple R-squared: 0.136, Adjusted R-squared: 0.1357   
F-statistic: 460.9 on 1 and 2928 DF, p-value: < 2.2e-16**

### Simple Linear Regression Interpretation:

Condition = -0.013x + 32.280, where x is age of the home

The slope of -0.013 means that for each year added to the age of the home, the overall condition of the home roughly decreases by 0.013. Another way to say this is that for each year that a house’s age moves away from the current date in gains 0.013 in condition. This means that according to the linear regression newer homes have worse condition.

The intercept of 32.280 means that if the home was very old and built in 0 A.D, the predicted condition of the home would be 32.280 condition units. This does not have meaning, to us because there is no reasonable way a house would date back that far.

This above code performs a simple linear regression analysis on the Ames dataset using the lm() function in R. The lm() function fits a linear model to the data and in this case, the predictor variable is Year.Built and the response variable is Overall.Cond. The summary() function is then used to obtain a summary of the linear regression model, which includes information such as the coefficients, standard errors, t-values, p-values, and R-squared value. The results of the summary show the relationship between the predictor and response variable. The point estimate is -0.0135527 for the year built. This means that for every one-unit increase in the year built (moving towards todays date), the overall condition of the house decreases by 0.0135527. The P value is 2.2e-16, which is very small which would tell us to reject the null hypothesis.

### Multiple Linear Regression:

Checking OLS Conditions

1st Condition:

**Independent Random Samples:** Once again, for this specific data, we know that it is derived from the purchases of homes from 2006-2010.. Because of this we will assume that the data is random and independent.

2nd Condition:

**Residuals Normally Distributed and Homoscedastic:** The residuals for this test are very roughly normally distributed, but not really. This means that a different model might give better results, however for this class we are going to stick with the multiple linear regression. This does not satisfy this condition.

Chart, histogram

Description automatically generated

**Figure 7:** Multiple Linear Regression Residual Histogram

3rd Condition:

**No Pattern in Plotting Residuals:** There is a pattern of the residuals like in the linear model this means this condition is not stratified however like above we are going to continue with this test but recognize that another model might be better suited.

Chart

Description automatically generated

**Figure 8:** Multiple Linear Regression Residual Patterns

Table

Description automatically generated

**Figure 9:** Scatter Plot with Multiple Linear Regression

### Performing Multiple Linear Regression:

**Call:  
lm(formula = Overall.Cond ~ +Year.Remod.Add + Year.Built, data = ames\_df)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-5.4337 -0.2825 -0.2597 0.7121 3.2965   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 5.4417553 1.6935130 3.213 0.00133 \*\*   
Year.Remod.Add 0.0232966 0.0010746 21.679 < 2e-16 \*\*\*  
Year.Built -0.0233876 0.0007412 -31.555 < 2e-16 \*\*\*  
---  
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
  
Residual standard error: 0.9594 on 2927 degrees of freedom  
Multiple R-squared: 0.2555, Adjusted R-squared: 0.255   
F-statistic: 502.3 on 2 and 2927 DF, p-value: < 2.2e-16**

### Multiple Linear Regression Interpretation:

Overall.Cond = 5.4417553 + 0.0232966 \* Year.Remod.Add - 0.0233876 \* Year.Built

Where:

* Intercept (5.4417553): This is the estimated overall condition when both the year remodeled and year built are 0 A.D.
* Year.Remod.Add coefficient (0.0232966): This represents the average increase in overall condition for each additional year since the house was remodeled or added. In this case, a one-year increase in the year remodeled or added is associated with an increase of 0.0232966 units in overall condition. This means that the more recent the remodel the better the condition of the house.
* Year.Built coefficient (-0.0233876): This represents the average decrease in overall condition for each year added to the house year.built date. In this case, a one-year increase in the year built is associated with a decrease of 0.0233876 units in overall condition. This means that the more recent the house was build the worse its condition is.

The code above is trying to predict overall condition based on year remodeled and year built. Our Point Estimates are 0.0232966 for year remodeled and - 0.0233876 for year built. This means for every one unit increase in year remodeled that the overall condition increases by .0232966 and for every one unit increase in year built the overall condition decreases by .0232966 units.

The Multiple R-squared tells us that 25.55% of the overall condition can be explained by year built and year remodeled. Since our pval is very small (less than 2.2e-16) we know that our results are replicable. It also tells us we should reject the null hypothesis.

# Results and Discussions

After having completed the simple linear regression and the multiple linear regression on our data, we have recognized that the null hypothesis does not hold, based on the evidence presented to use in the tests. There is clearly an effect that age and remodeling date have on the condition of the house, but it is not at all what we expected. The more recent the remodeling date is, the better the condition of the house is. This is what we expected, but the fact that newer houses have worse conditions is hard to believe. This means that the tests support the rejection of our null hypothesis, but the adoption of the alternative hypothesis we originally came up with.

One reason we may not have achieved the expected results, is we did not fully understand the dataset we were working with. Now, the result makes more sense to us. If we consider the fact that all the data comes from a dataset founded around the homes bought from 2006-2010, it is easy to recognize why the dataset may be skewed. 2008 was a recession. This means that people may not have been buying newer homes because they might have been too expensive, and if they were bought they may not have been fully done because many construction companies did not finish building houses that they started building. This means that ultimately during this time frame we could have seen only the cheaper house selling and they may be more run down or in a poor condition. May be that would be another good question for another research project.

# Conclusion

One critical thing we recognized over the course of our research, is that the linear regression model and the multiple linear regression both are tests that do not feel natural or fitting for the data we are analyzing. A different or more complicated research model might be more appropriate for the analysis of this data, but for the purpose of this project we will discuss the results of the tests we performed. Something else to keep in mind is that the rounding of conditional units maybe the reason why the linear and multiple linear regression models are not very fitting for these specific variables of the data set.

In the linear regressions above the value to pay attention to are the point estimate, the slope. In the simple linear regression, we can see that the slope -0.0135527 for associating year built with overall condition. What this tells us is that there is a slight association with a decrease in a buildings overall condition when the year\_built increases.

In the second example, the multiple linear regression we have two values to look at. We have 0.0232966, which tells us that there is a slight positive association with a newer remodel age create a better overall house condition, while the second value -0.0233876 tells us there is a slight negative association between overall condition and increases in year built. So far, these simple tests confirm that older houses and houses with more recent renovations seem to have a better overall condition.

Finally, for both the linear regression and the multiple linear regression the p values are very low. This tells us that we need to reject the null hypothesis. Using these research models, we have successfully answered our research question “What is the relationship between the age of a home and its overall condition?” However, we must remember to be skeptical of these results as the model that was used is not the most appropriate for the data, and our result does not match the expected alternative hypothesis.

# 

# **Appendix A:** R Code for Diagrams

**Figure 1:** Year Built vs. Overall Condition

ggplot(ames, aes(x = Year.Built, y = Overall.Cond)) +  
  geom\_point() +  
  labs(title = "Year Built vs. Overall Condition",  
       x = "Year Built",  
       y = "Overall Condition") +  
  theme\_minimal()

**Figure 2:** Year of Remodel vs. Overall Condition

ggplot(ames, aes(x = Year.Remod.Add, y = Overall.Cond)) +  
  geom\_point() +  
  labs(title = "Year of Remodel/Addition vs. Overall Condition",  
       x = "Year of Remodel/Addition",  
       y = "Overall Condition") +  
  theme\_minimal()

**Figure 3:** Year Built and Year of Remodel vs Overall Condition

install.packages("plotly")  
library(plotly)

plot\_ly(data = ames,  
        x = ~Year.Built,  
        y = ~Year.Remod.Add,  
        z = ~Overall.Cond,  
        type = "scatter3d",  
        mode = "markers",  
        marker = list(size = 3)) %>%  
  layout(title = "Year Built, Year of Remodel/Addition, and Overall Condition",  
         scene = list(xaxis = list(title = "Year Built"),  
                      yaxis = list(title = "Year of Remodel/Addition"),  
                      zaxis = list(title = "Overall Condition")))

**Figure 4:** Residual Histogram and Residual Plot

#dataset

data(ames)

# regression model

model <- lm(formula = Overall.Cond ~ Year.Built, data = ames)

#residuals

res <- resid(model)

# residual vs. fitted plot

plot(ames$Overall.Cond, res)

#add a horizontal line at 0

abline(0,0)

#Graph

ggplot(data.frame(residuals = res), aes(x = residuals)) +

geom\_histogram()

**Figure 5**: Scatter Plot with Linear Regression

#Dataset

data(ames)

#Graph

ggplot(ames, aes(x = Year.Built, y = Overall.Cond)) + geom\_point() + geom\_smooth(method = "lm", se = FALSE, color = "blue") + labs(title = "Year Built vs. Overall Condition", x = "Year Built", y = "Overall Condition") + theme\_minimal()

**Figure 6:** Multiple Linear Regression Residual Histogram and Multiple Linear Regression Residual Patterns

#dataset

data(ames)

# regression model

model <- lm(formula = Overall.Cond ~ +Year.Remod.Add + Year.Built, data = ames)

#residuals

res <- resid(model)

# residual vs. fitted plot

plot(ames$Overall.Cond, res)

#add a horizontal line at 0

abline(0,0)

ggplot(data.frame(residuals = res), aes(x = residuals)) +

geom\_histogram()

**Figure 7**: Scatter Plot with Multiple Linear Regression

#Dataset

data(ames)

#Graph

ggplot(ames, aes(x = Year.Remod.Add, y = Overall.Cond)) + geom\_point() + geom\_smooth(method = "lm", se = FALSE, color = "blue") + labs(title = "Year of Remodel/Addition vs. Overall Condition", x = "Year of Remodel/Addition", y = "Overall Condition") + theme\_minimal()

# **Works Cited**

De Cock, Dean. "Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project." Journal of Statistics Education, vol. 19, no. 3, 2011, pp. 1-20, <https://www.tandfonline.com/doi/abs/10.1080/10691898.2011.11889627>.

“Internachi's Standard Estimated Life Expectancy Chart for Homes.” *InterNACHI® - InternationalAssociationofCertifiedHomeInspectors*, https://www.nachi.org/life-expectancy.htm.

Vercelletto, Niko.

“5 Reasons Why Buying an Old House Is a Great Idea.” *Real Estate News & InsightsRealtor.com®*,16Aug.2022, https://www.realtor.com/advice/buy/why-buy-old-home-instead-of-new-one/#:~:text=Old%20homes%20have%20better%2Dquality%20construction&text=Even%20the%20walls%20are%20likely,better%20sound%20barrier%20and%20insulation.

“Old House vs. New House: What to Consider.” *Old House Vs. New House: What To Consider | Rocket Mortgage*, https://www.rocketmortgage.com/learn/old-house-vs-new-house.