Lab 2: Description Using Models

A descriptive statistical analysis of housing prices in Ames, Iowa

Ben Robbins Micah Collins Shikha Sharma Scott Stossel

April 18, 2025

Does the total livable square footage, neighborhood, season of sale, number of bedrooms/baths, finished basement, and recent remodeling/upgrades correlate to the sale price in the real estate market of Ames, Iowa?

1. Introduction

Being able to accurately understand the relationship between various aspects of a house and the sale price is a crucial first step for real estate investors to understand the underwriting process. As small time investors interested in single family in Ames, Iowa who have heard of certain factors affecting price, we wanted to answer the question: Does the total livable square footage, neighborhood, season of sale, number of bedrooms/baths, finished basement, and recent remodeling/upgrades correlate to the sale price in the real estate market of Ames, Iowa? While in the research question we called out six factors that we hypothesized would have a positive correlation on sale price, we plan to study more variables that are contained within the dataset. Understanding these relationships help us improve our underwriting process and lay the foundation for further research to eventually make predictive and causal models.

2. Description of the Data Source

This analysis uses a dataset from Kaggle that contains property records collected through the Ames Assessor's Office between 2006 and 2010.1 Each row represents a single home sale, with 79 variables describing a variety of features about the property.2 These variables include characteristics such as location, lot area, and quality, which we assessed in the construction of our descriptive model.

The data went through a number of data cleaning and wrangling steps prior to the construction of the model. First, we filtered the data to only look at sales from 2006 and 2007. Since the goal was to better understand housing features and their relationship with price, we decided to avoid the dramatic effect the 2008 recession had on housing prices. We further filtered by

single family homes to avoid miscellaneous property types such as duplexes which were not well represented by the dataset. From here, we verified that all of the variables of interest were not missing any values. We condensed certain categorical variables into larger categories to avoid

3. Data Wrangling

The original data contained 1460 records of sales from 2006-2010, but to avoid impact of the housing crisis from 2008, we decided to use data from 2006-2007, which removed 817 observations. Further, we decided to focus specifically on single family homes, which further reduced our data by 105 observations. To separate the steps of data filtering and wrangling logic we split the steps of filtering the data and deleting rows with nulls in a data filtering file and the steps of deriving new columns and fixing data types into a data wrangling file. We added four new columns to aid in our regression analysis. To analyse effect of seasonality, we added a column "SeasonSold" based on month of sale. "ZoneType" was added to categorize by housing density. "Age" was added to using the year house was built. "NeighborhoodGroup" was added to categorize the neighborhoods by geographical location. Then the data is split into two sets, 30% for exploratory and 70% for confirmatory analysis.

4. Operationalization

In order to obtain a better understanding of the relation between the variables in the data, we generated a correlation matrix to help us identify which variables to select for our analysis. From the heatmap generated using the correlation matrix, we selected variables that had a high level of correlation with Sale price like the Above ground living area, the overall material and finish of the house, the overall condition of the house, the year house was built, the year remodelling/additions were added, total basement area, number of full baths, kitchen quality, car capacity of garage. We also made note of the variables that had a high level of correlation with each other suggesting possible collinearity. This included lot size, total above ground area, number of full baths.

5. Data Visualization

To gain an understanding of the distributions of the variables, we explored univariate and bivariate distributions. As depicted in Figure 1, panel A shows a more pronounced right skewed distribution for the total above ground living area than the skew shown by Sale Price in panel B. This contributes to the higher variance in our model residuals for the higher price range of larger homes. Panel C shows the joint distribution of both these variables and the Overall Quality of the houses which shows a linear relationship for lower size houses but the relationship appears to change to non linear for larger size homes. Finally, panel D shows the final model predictions overlayed on the actual sale price. As expected, there is high accuracy

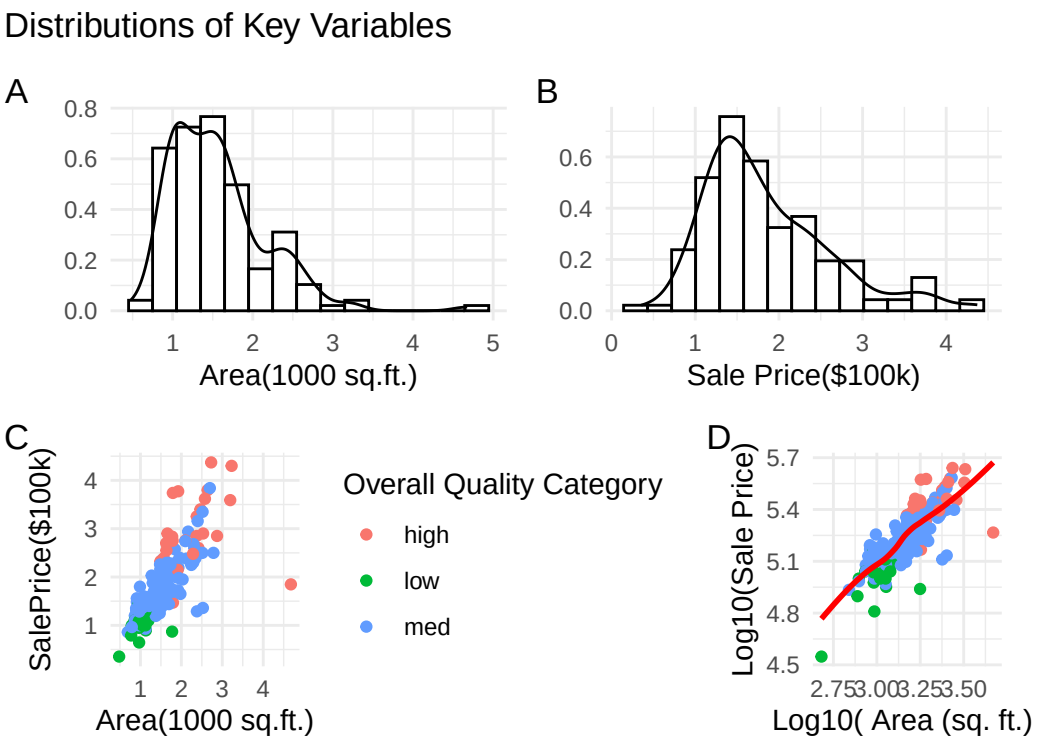

Figure 1: In panel (A) we report the distribution of Total Living Area above ground, which has a long-right tail. In panel (B) we report the distribution of Single Family homes Sale Price, which presents an almost classic normal distribution with a slight skew to the right. In panel (C) we present the joint distribution of the two, together with Overall Quality of the house noting there is a non linear relationship. Finally, panel (D) overlays the final model predictions in red over the joint distribution of the log10(actual sale price) and square root of house area.

for lower house prices, but for higher house prices, the model is unable to capture some of the non-linearity of the actual sale price. The effect of the data transformation is also visible by comparing panel C and panel D.

6. Model Specification

We explain the variation in house sale prices using a sequence of linear regression models operating on both metric and ordinal variables within our housing dataset. Our dependent variable is SalePrice, measured in U.S. dollars. We eventually model log base 10 of SalePrice in order to stabilize variance and improve the normality of errors.

After using our exploratory data to test a number of different models we found the best model

had 5 statistically significant features: sqrt(Living area), overall quality, sqrt(Total Basement area), overall condition, and year built. We also performed a log transformation on the sale price since it has a heavy tail. Likewise we performed the square root transformation on the areas due to heavy tails and there being some 0 values which can't be handled by logs. Overall quality and overall condition were treated as metric variables as opposed to ordinal due to them being measured by professionals with a strict set of rules that we believe make the distances between the numbers approximately equal.

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Fri, Apr 18, 2025 - 12:38:14

7. Model Assumptions

IID Although due to the size of the dataset 377 CLT assumptions will apply, there were some concerns we had to address. Temporal clustering was an issue especially due to the financial crisis of 2008. Generalization of the observations from this data will have to be constrained to other similar suburban university towns. In order to examine if the data is IID, we examined the residuals of the linear model. To reduce the impact due to large variations in the house size and the sale price, we used the log of the sale price and square root of the house price. This helped reduced the clustering in the residuals as can be seen by Figure 4 in the Appendix.

Perfect Collinearity As we can observe from Figure 3, there is some collinearity between variables like lot area and house size and number of full baths. We chose only one of these variables in our model to avoid issues related to near perfect collinearity. In addition we ran a VIF test and the highest value 1.8099738 is well below the allowed value of 4.

Linear Conditional Expectation Overall, the assumption of a linear conditional expectation is supported by the log-transformed model, but there are serious concerns within the original model in terms of possible non-linear trends as well as highly influential outliers. As can be seen from Figure 4 the original model overpredicts for the lower sale price and underpredicts for the higher sale price. The final model has better overal prediction though it still under predicts for very low prices and very high prices and slightly overpredicts for most of the data.

Constant Error Variance The studentized Breush-Pagan test for the initial model gives a very low p-value of $2.3160859 \times 10^{-11}$ which indicates heteroskedasticity in the original data. Due to the transformations performed for the final model, the p-value 0.0046931 from the test is high ($p>0.05$), we cannot reject the null hypothesis of constant error variance (homoskedasticity).

Normally Distributed Errors The residual distribution appears to nearly normal as visible from Figure 2. Further, the Q-Q plots from Figure 4 show reduction in the heavy tails for the residuals. Comparing the skewness and kurtosis for the model residuals, we observe the kurtosis reduced from 8.1475096 to 1.1508506, indicating Leptokurtic residuals. The skewness

Table 1: Comparison of initial, intermediate and Final Regression Models.

	Dependent variable:		
	Log10(SalePrice)		
	Initial Model	Intermediate Model	Advanced Model
	(1)	(2)	(3)
GrLivArea	0.0002^{***}		
	(0.0000)		
log(GrLivArea)		0.65^{***}	0.54^{***}
		(0.08)	(0.08)
OverallQualCategory		-0.25^{***}	-0.15^{***}
		(0.04)	(0.03)
log(TotalBsmtSF)		-0.11^{***}	-0.07^{***}
		(0.02)	(0.02)
YearBuilt			0.03^{***}
			(0.01)
OverallCondCategory			0.002^{***}
			(0.0003)
Intercept			-0.12^{**}
			(0.05)
OverallCondCategorymed			-0.05^{**}
			(0.02)
Constant	4.91^{***}	3.28^{***}	-0.04
	(0.05)	(0.27)	(0.49)
Observations	161	161	161
R^2	0.55	0.74	0.82
Adjusted R^2	0.54	0.74	0.82
Note:			*p<0.1; **p<0.05; ***p<0.01

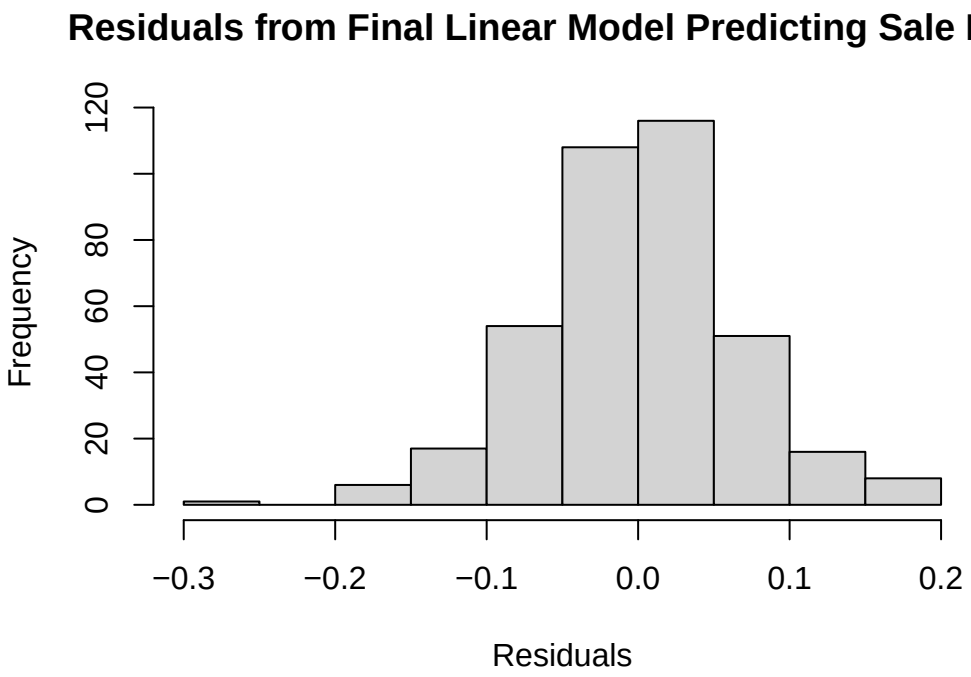

Figure 2: Residuals from Final Linear Model predicting Sale Price

reduced from -0.811701 to -0.0812877. Based on these values, we can conclude that the residual distribution for the final model can be accepted as having a fairly normal distribution.

8. Model Results and Interpretation

A report that scores in the top level will correctly interpret statistical significance, clearly interpret practical significance, and comment on the broader implications of the results. It may want to include statistical tests besides the standard t-tests for regression coefficients. When discussing practical significance, comment on both the direction and magnitude of your coefficients, placing them in context so the reader can understand if they are important. To help the reader understand your fitted model, you may want to describe hypothetical datapoints (e.g. a hypothetical person with 1 cat is predicted to spend $2400 on pet care. That rises to $3200 for a hypothetical person with 2 cats…).

Appendix

1. **Data Source** Anna Montoya and DataCanary. House Prices - Advanced Regression Techniques, 2016, Kaggle

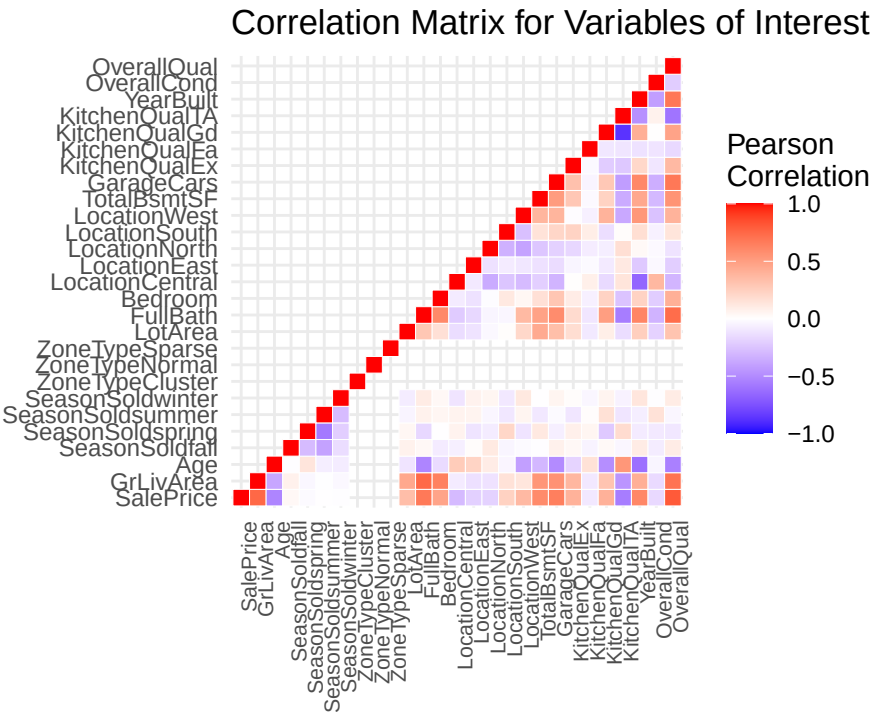

Figure 3: Correlation Matrix

2. **A List of Model Specifications Tried.** We are interested in seeing how you arrived at your final model. In just a sentence, please provide a reason or something that you learned from each specification.

3. Residuals-vs-Fitted-values Plot

Model Summary

```
Call:
lm(formula = log10(SalePrice) ~ log10(GrLivArea) + OverallQualCategory +
    TotalBsmtSFLogged + YearBuilt + OverallCondCategory, data = confirm_data)
```

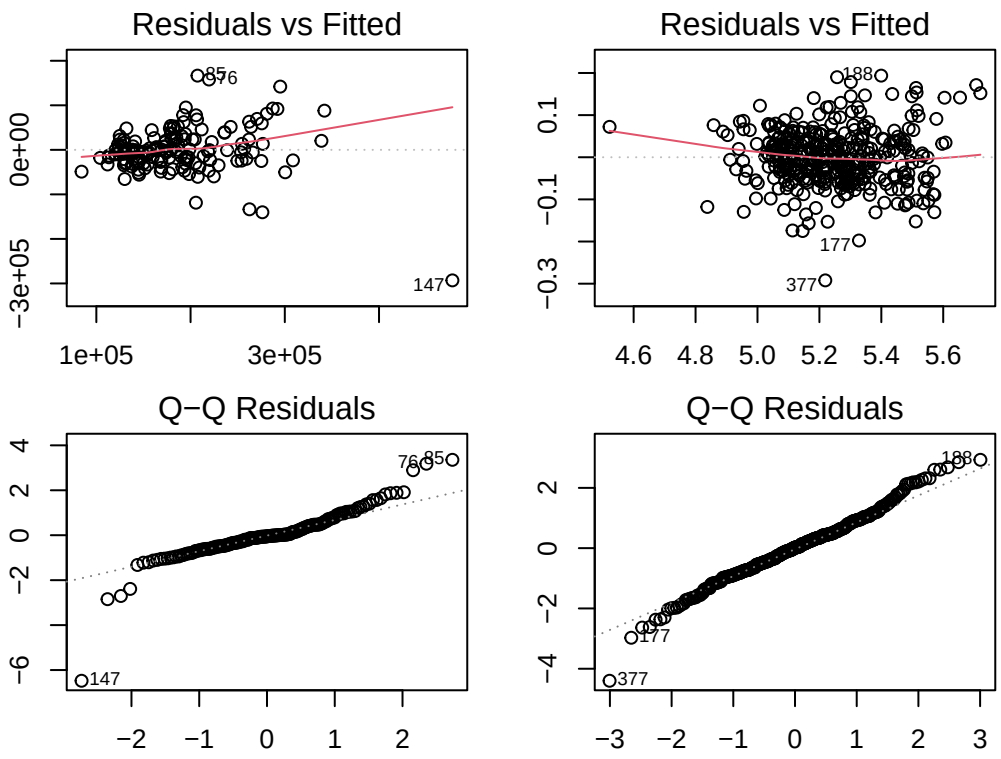

Figure 4: Comparison of Inital(left) and Final(right) Model Residuals and QQ plot.

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.292184 -0.042467  0.001402  0.037546  0.193714

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)              -0.1447465  0.2713563  -0.533   0.5941
log10(GrLivArea)          0.6612428  0.0281844  23.461  < 2e-16 ***
OverallQualCategorylow   -0.1327365  0.0187269  -7.088 6.96e-12 ***
OverallQualCategorymed   -0.1121440  0.0109506 -10.241  < 2e-16 ***
TotalBsmtSFLogged         0.0175272  0.0037004   4.737 3.11e-06 ***
YearBuilt                 0.0016818  0.0001324  12.703  < 2e-16 ***
OverallCondCategorylow   -0.1417861  0.0210633  -6.731 6.44e-11 ***
OverallCondCategorymed   -0.0423983  0.0166521  -2.546   0.0113 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06678 on 369 degrees of freedom
```

```
Multiple R-squared:  0.8649,    Adjusted R-squared:  0.8624
F-statistic: 337.6 on 7 and 369 DF,  p-value: < 2.2e-16
```