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

1. Introduction

Accurately understanding 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 investors interested in single family houses in Ames, Iowa investigating 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? We hypothesized that these six factors would have a positive correlation on sale price; however, we plan to study more variables 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. Data and Methodology

This analysis uses a dataset from Kaggle (DeCock 2011) that contains property records collected through the Ames Assessor's Office between 2006 and 2010. Each row represents a single home sale, with 80 variables describing a variety of features that could influence price. These variables include characteristics such as location, lot area, and quality, which we assessed in the construction of our descriptive model.

The original data with 1460 observations went through a number of data cleaning and wrangling steps prior to the construction of the model. First, we filtered 817 observations from the data to only look at sales from 2006 and 2007 to avoid the edge case of the 2008 housing crisis. We further filtered by single family homes to avoid miscellaneous property types such as duplexes which were not well represented by the dataset, removing 105 observations, leaving us with 538 samples. From here, we verified that all of the variables of interest were not missing any values. Some categorical variables, like neighborhood, included a large number of different categories which we grouped into broader ones to simplify the model and avoid overfitting.

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 resulting heatmap, we selected variables that had a high level of correlation with Sale price like the Above ground living area, overall quality, overall condition, the year house was built, the year remodelling/additions were added, total basement area, number of full baths, kitchen quality, and the 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.

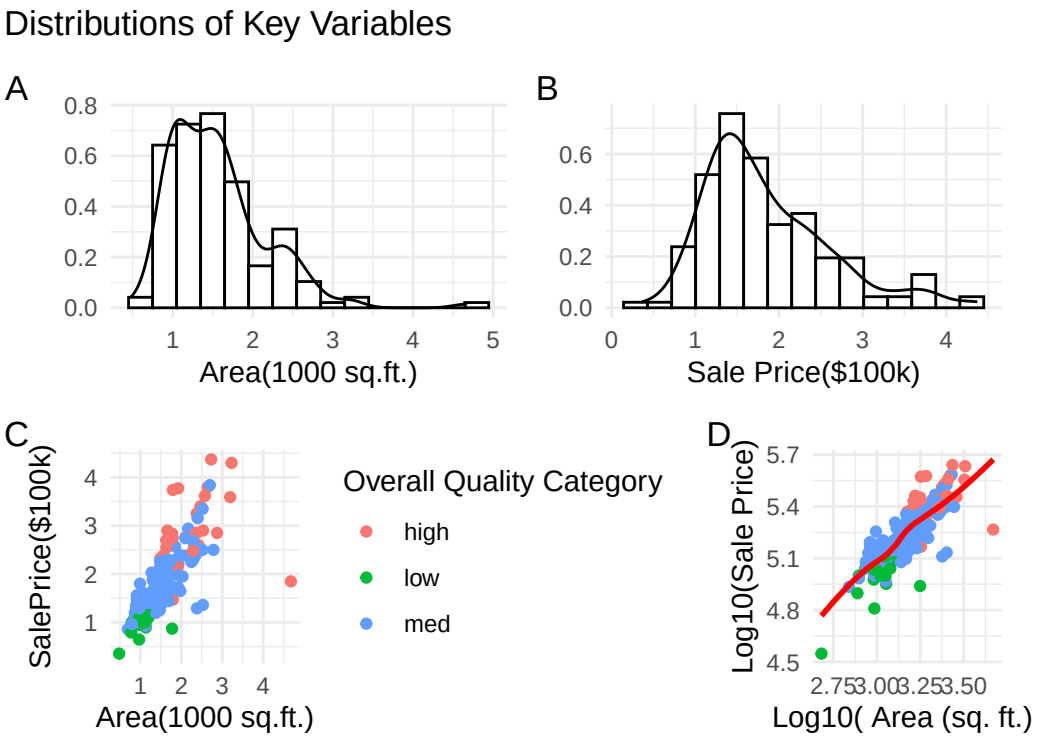

Figure 1: In panel (A) we report the distribution of Total Living Area above ground, which has a long-right tail. In panel (B) we report the distribution of Single Family homes Sale Price, which presents an almost classic normal distribution with a slight skew to the right. In panel (C) we present the 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.

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 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.

3. Model Specifications and Assumptions

We explain the variation in house sale prices using a sequence of linear regression models operating on metric within our dataset. Our dependent variable is SalePrice, measured in U.S. dollars, transformed by a log function in order to stabilize variance and improve the normality of errors. We report robust standard errors due to heteroskedasticity concerns in the data. Our final model includes the independent variables: GrLivArea (living area above ground), which is transformed by a log function to represent the nonlinear and diminishing returns it has on SalePrice; TotalBsmtSF (Total Basement Square Footage), which is transformed by a log function to represent the nonlinear and diminishing returns it has on SalePrice; YearBuilt; OverallQual (Overall House Quality), which is an ordinal variable measured on a scale of 1 to 10 which we have split into categories of "low", "med", and "high", with "low" being the basis category; and OverallCond (Overall House Condition) is also an ordinal variable split into those same categories.

% 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 - 17:12:58

While the sample size (n=377) supports the Central Limit Theorem, we closely examined the key assumptions of linear regression. In order to verify if the data is **IID**, we examined the residuals and applied log transformations to sale price and square footage to reduce the impact of large variations in house size and price. This transformation reduced clustering and the influence of extreme values, as seen in Figure 4. 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 thus satisfying the assumption of **No Perfect Collinearity** . The log-log model, created from the above-mentioned transformation, better satisfies the **linear conditional expectation** assumption, as it corrects for the curved residual patterns and improves fit across the range of sale prices.

We also attempted to address the assumption of **Constant Error Variance** with the log transformations in the final model, the p-value from the Breusch-Pagan test increased from $2.3160859 \times 10^{-11}$ to 0.0046931, indicating some improvement, but still with heteroskedasticity present. Despite this, Figure 2 shows that the residual distribution appears to be nearly normal

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*** (0.08)	0.54*** (0.08)
OverallQualCategory		−0.25*** (0.04)	−0.15*** (0.03)
log(TotalBsmtSF)		−0.11*** (0.02)	−0.07*** (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*** (0.05)	3.28*** (0.27)	−0.04 (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

with a slight skew to the left. Further, the Q-Q plots from Figure 4 show reduction in the heavy tails for the residuals and the skewness improved from -0.811701 to -0.0812877 , indicating near symmetry. While still leptokurtic, the kurtosis reduced from 8.1475096 to 1.1508506 due to fewer extreme values. Based on these values, we can conclude that the residual distribution for the final model can be accepted as having **Normally Distributed Errors**.

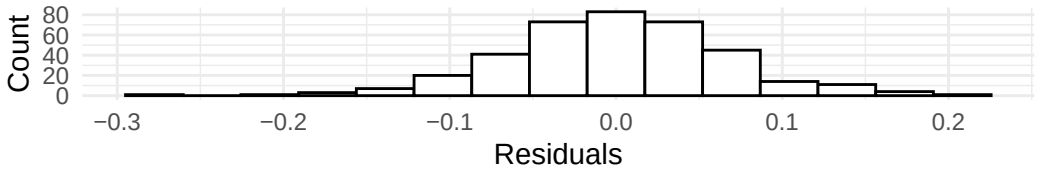

Figure 2: Residuals from Final Linear Model predicting Sale Price

4. Model Results and Interpretation

```
      RMSE   Rsquared        MAE
0.06607040 0.86493041 0.05056557
```

The stargazer chart reports three models predicting log(SalePrice). Column 3 is our preferred "Advanced Model," which explains 82.3% of the variance of log(SalePrice) within the training

set and results in an adjusted R-squared value of 0.815. We transformed the above-ground living area with a log in the final model to account for the diminishing returns with price that this feature is associated with. Every variable other than medium overall cond was statistically significant with p-values under 0.05. Since the Studentized Breusch-Pagan test revealed heteroskedasticity, we used robust standard errors in our final model. This ensures more accurate significance results, despite the violation of the linear regression assumption of constant variance. When we apply the Advanced model to the test set (70% of the data), it achieves a Root Mean Square Error value of 0.066 and R-squared value of approximately 0.865.

The results of the model largely match our intuition. Based on its coefficient of 0.54, a 1% increase in GrLivArea corresponds to a 0.54% increase in sale price, holding all else equal. Similarly, year built and basement size were statistically significant and positively correlated with sale price, indicating that newer homes and larger basements tend to sell for a higher price. The categorical variables for quality and condition show that while high quality was associated with higher prices, medium condition was not significantly distinguishable.

5. Discussion

Our final descriptive model explains 86.5% of the variation within the test set for the log transform of sale prices by using variables such as the log of above-ground living area, basement size, year built, overall quality, and condition. The log-log form for above-ground living area captures the diminishing returns observed by the fact that larger homes increase non-proportionally in value, with all else held equal. While overall quality was a statistically significant predictor if the condition was high, it was not significant with other values, indicating that only homes in the best condition stand out in terms of price.

The coefficients were practically meaningful and aligned with expectations. Further improvements could involve additions such as collecting data on household income and grouping neighborhoods by their average income to better capture socioeconomic effects. Overall, the model provides a foundation for understanding the drivers of sale price in Ames and supports the improvement of the underwriting process for investment properties through data-driven decision making.

Appendix

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

2. **A List of Model Specifications Tried.** Specifications of Attempted Models:

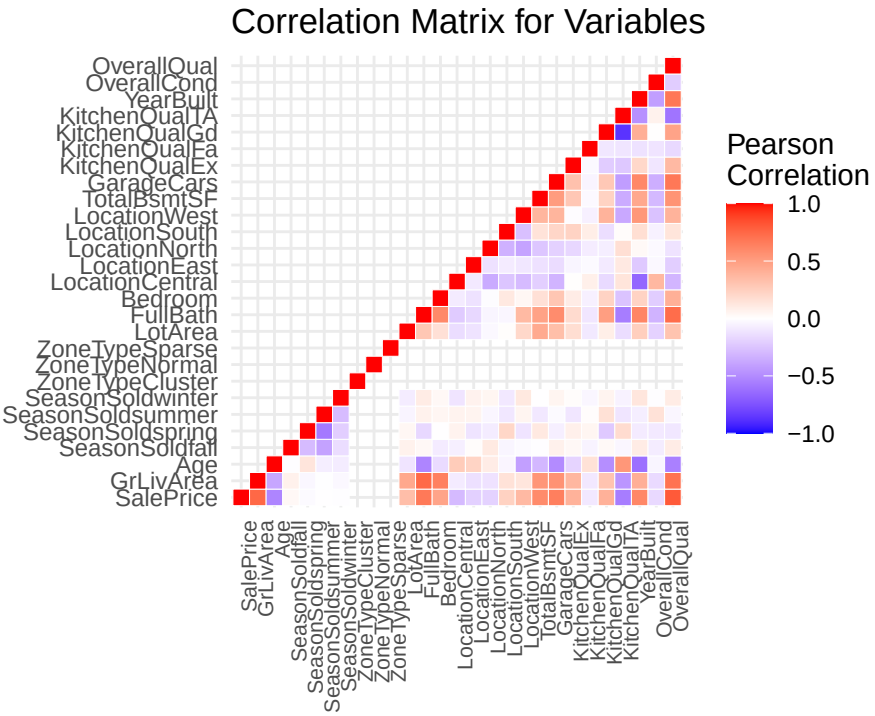

Figure 3: Correlation Matrix

1. SalePrice ~ GrLivArea Initial model, showed positive relationship between GrLivArea and SalePrice
2. SalePrice ~ GrLivArea + LotArea We discovered that LotArea is not a significant predictor of SalePrice
3. SalePrice ~ GrLivArea + FullBath + SeasonSold We discovered Seasonsold is not a significant predictor of SalePrice
4. SalePrice ~ sqrt(GrLivArea) + TotalBsmtSF + YearBuilt + GarageCars + KitchenQual We discovered GarageCars is not a significant predictor of SalePrice
5. log10(SalePrice) ~ sqrt(GrLivArea) + sqrt(TotalBsmtSF) + YearBuilt + KitchenQual We discovered that KitchenQual in its baseline state was not a statistically significant predictor of the log of SalePrice.
6. log10(SalePrice) ~ GrLivArea We made an alternate version of our initial model with the new method of predicting the log of SalePrice, this model has considerably better fit than the original model without the transformation.
7. log10(SalePrice) ~ sqrt(GrLivArea)+sqrt(TotalBsmtSF) + YearBuilt + OverallQual + OverallCond Initial testing on linearity of OverallQual and OverallCond, they work nearly like a metric variable, however OverallQual has collinearity concerns if it is not converted into categories.

8. log10(SalePrice) ~ log10(GrLivArea) + TotalBsmtSFLogged + YearBuilt + OverallQualCategory + OverallCondCategory We grouped values of OverallQual and OverallCond into categories to better reflect them as ordinal variables. This considerably reduces collinearity concerns as evidenced by VIF testing.

3. Residuals-vs-Fitted-values Plot

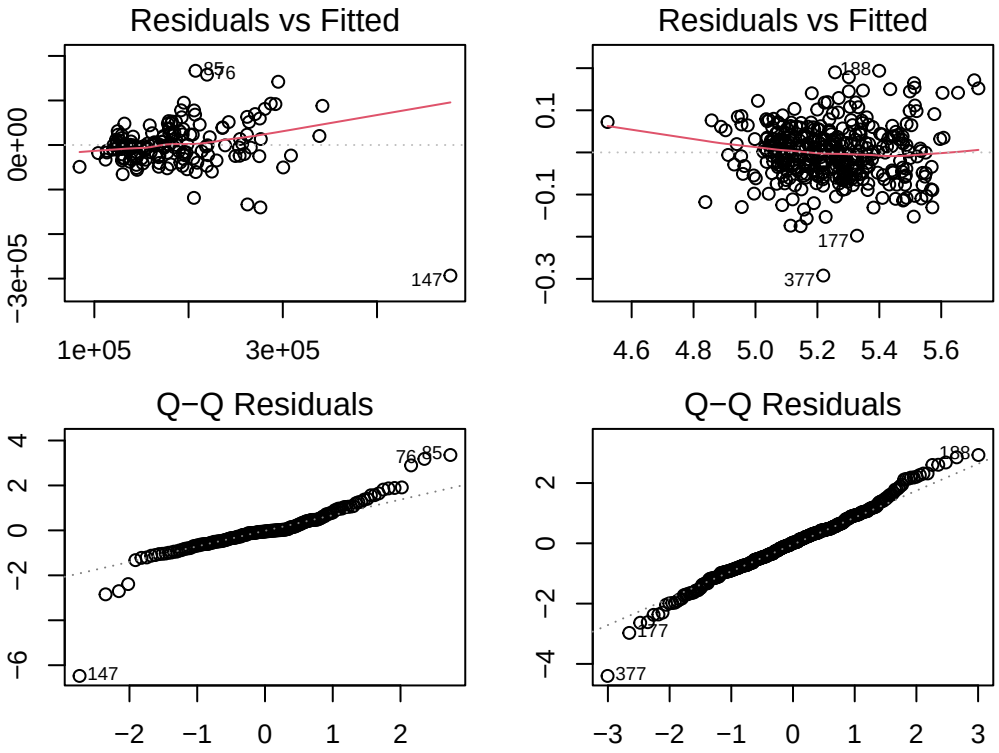

Figure 4: Comparison of Inital(left) and Final(right) Model Residuals and QQ plot.

References

DeCock, Dean. 2011. "Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project." *Journal of Statistics Education* 19 (3): 87–115.