ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

House Price Analysis and Prediction

Data Analysis

Student: Giorgia Treglia

Identification Number: 6042123

Academic Year 2024/2025

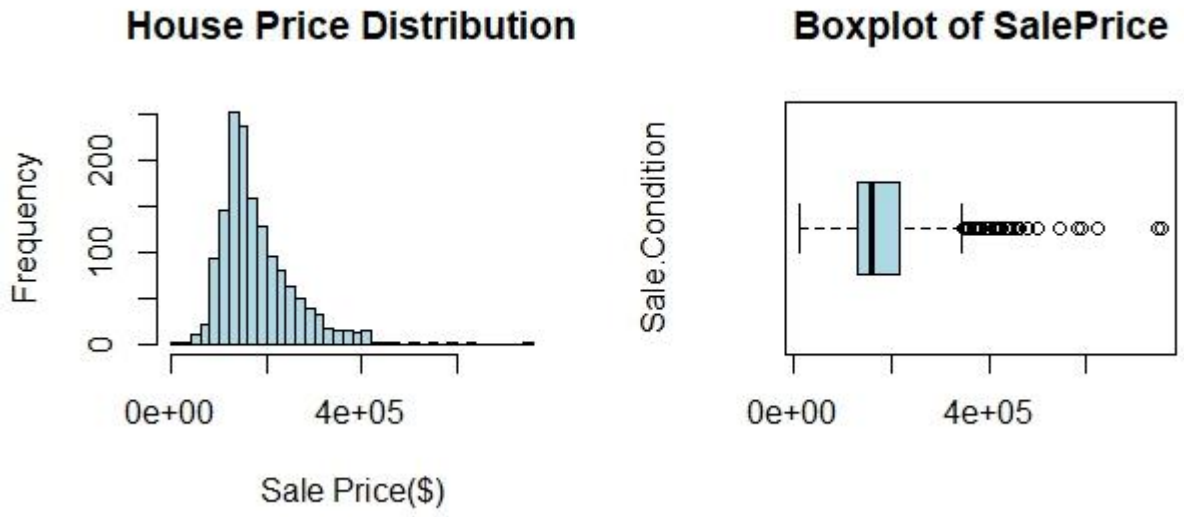
Table of Contents

Abstract

This study analyses residential property sales in Ames, Iowa, to develop a predictive model for house prices. Using a dataset of 2,930 observations with 82 variables, an exploratory data analysis was conducted to identify key price determinants. A multiple linear regression model was built, refined through stepwise selection, and validated using 10-fold cross-validation and loovc. The final model, which explains 93.5% of price variance, highlights above-ground living area, total basement size, garage area, and overall quality as the most influential factors.

Exploratory Data Analysis (EDA)

Since the aim is to develop a predictive model that accurately estimates house prices based on these attributes and the dataset includes a mix of continuous, discrete, ordinal, and nominal variables, a thorough exploratory data analysis is necessary before building a predictive model. One of the first steps was to check for missing values across variables. A threshold of 40% was set, meaning that any variable with a higher proportion of missing values was removed. For the remaining, numerical variables were imputed using their median, while categorical variables were replaced with the label "None". A histogram and a boxplot of SalePrice were plotted to examine the distribution of prices.



The distribution appears to be right-skewed, indicating that a small number of properties have significantly higher prices than the majority. To further explore the relationships between variables, a correlation matrix was computed for all numerical features, using Pearson's correlation coefficient. The strongest correlations with SalePrice were observed in variables such as Gr Liv Area, Garage Area, Total Basement Area, and First-Floor Area.

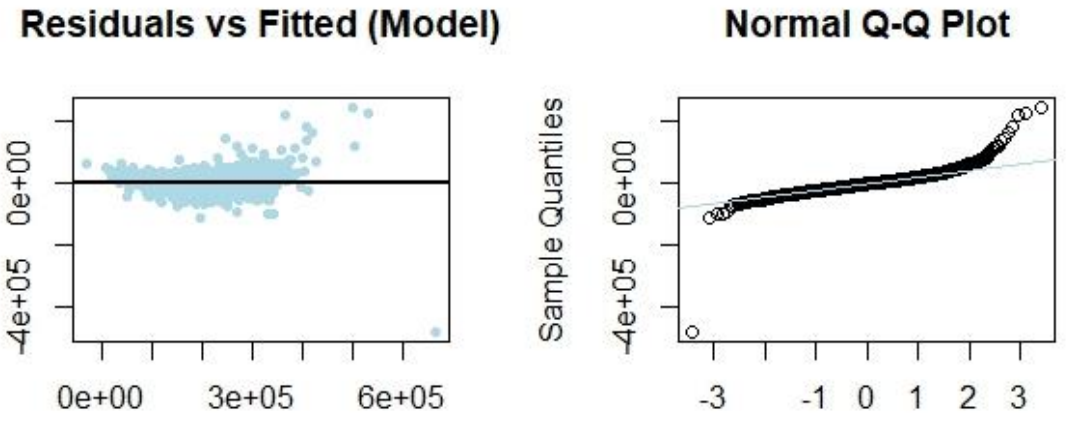
Feature Selection

To ensure that only the most relevant variables were included in the model, a combination of LASSO regression, Fisher's exact test, and ANOVA was applied. These methods allowed for an efficient selection process by eliminating redundant or weak predictor. LASSO was used to handle numerical variables; it applies a penalty to the regression coefficients, shrinking them toward zero and effectively removing those that contribute little to the prediction of SalePrice. Using 10-fold cross-validation, the optimal lambda value was determined to be 3984.373. This value represents

the point where the mean squared error (MSE) is minimized without overfitting. After LASSO, 22 variables were eliminated indicating that these features had little additional predictive value when combined with the remaining ones. To refine the selection of categorical features, Fisher's exact test was applied. This test evaluates whether a categorical variable has a statistically significant relationship with SalePrice, ensuring that only meaningful predictors are retained in the model. As a result, 17 categorical variables were removed. Also, Analysis of Variance helps to assess the statistical significance of each remaining variable. This method examines whether different categorical features significantly contribute to the variance in SalePrice. In this way, 17 additional variables were excluded, including Exterior Condition, Basement Condition and Sale Condition.

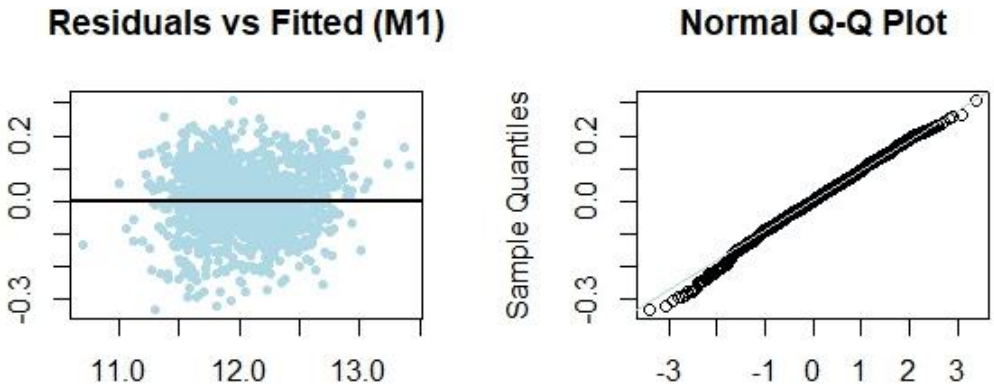
Regression Model and Assumptions

The house-price prediction has been implemented through multiple linear regression, which constituted a first model containing all selected predictor variables, but later diagnostic checks pointed out many things that required modification. Residuals plotting indicated obvious violations of linearity, requiring transformation of some selected variables. The assessment of independence was performed through the Durbin-Watson test and showed that residuals were not correlated (DW = 1.9809). Heteroscedasticity was documented because variances of residuals were not constant for all levels of predictors, (p = < 2.22e-16). Normality of the residuals was diagnosed by comparing residuals on a Q-Q plot that showed, among others, violations of normality due to extreme values.



One of the log and square root transformations were applied to Sale Price, Lot Area, Basement Square Footage, and some other variable, which significantly improved the general fit of the model. Furthermore, the variance inflation factor (VIF) was used to compute multicollinearity among predictor variables in that many predictors have shown severely high collinearity among them.

Then, after a series of iterative adjustments, the final model, M1, piloted by regression assumptions, was found. The statistical value obtained from the Durbin-Watson test that was used to evaluate independence assertions was 1.97, which endorsed the argument that residuals were not significantly autocorrelated. In addition, in the non-constant variance (NCV) test where homoscedasticity was measured, the p-value obtained was 0.1531. In addition, a stepwise selection method was applied to further refine the model by systematically removing statistically insignificant variables. This process led to a more parsimonious model with improved predictive performance.



10-Fold Cross-Validation and LOOCV

To assess the predictive accuracy and generalizability of the developed models, both 10-fold cross-validation (CV) and Leave-One-Out Cross-Validation (LOOCV) were performed. These validation techniques ensure that the model performs well not only on the training data but also on unseen observations, reducing the risk of overfitting.

Model	10-FCV RMSE	10-FCV RMSE	LOOCV RMSE	LOOCV MAE
M1	0.0977	0.0773	0.0980	0.0772
Stepwise (16 p)	0.0981	0.0776	0.0985	0.0775
Stepwise (19 p)	0.0979	0.0774	0.0982	0.0772
Stepwise (23 p)	0.0977	0.0773	0.0979	0.0771

Initially, M1 was considered the primary model, providing a 10-fold RMSE of 0.0977 and LOOCV of 0.0980, indicating fair predictive performance. However, different versions of stepwise models were tried for modelling. The stepwise model with 19 predictors provided the best generalization performance, giving a 10-fold CV RMSE of 0.0979 and an LOOCV RMSE of 0.0982. Therefore,

the final selection of the model changed from M1 to Stepwise (19 p), as it would maintain strength and predictive capability while reducing complexity.

Interpretation of Parameters of the Final Model

The final model, selected through stepwise regression with 19 variables, showed an Adjusted R^2 of 0.9353 indicates that the model explains approximately 93.5% of the variance in house prices and an F-statistic (p-value < 2.2e-16) which further validates the model's overall significance, ensuring that the chosen predictors have a meaningful impact on price estimation. Examining the coefficients provides insight into how each feature influences SalePrice, keeping in mind that the dependent variable is under a log transformation. This means that each coefficient represents a percentage change in SalePrice for a one-unit increase in the predictor, rather than an absolute change in dollars. Among the most important predictors, the Above-Ground Living Area (0.2870) has the highest positive effect on SalePrice. Therefore, a 1% increase in the living area will correspond an approximate 0.29% increase in SalePrice, which shows the value of usable living space in determining the value of property. Similarly, Total Basement Size (0.2492) and Garage Area (0.1689) positively contribute to house prices, indicating that larger basements and garages increase a home's market value. Categorical variables further influence pricing, such as MS Zoning (0.3059), which suggests that homes located in specific zoning classifications experience higher demand. Some coefficients have negative values, such as Kitchen.AbvGr (-0.0988), indicating that properties with additional kitchens tend to have slightly lower prices when controlling for other factors.

Model Predictions

After the model selection process, predictions were made on both the training and test datasets to evaluate performance. From the results, the Stepwise model with 19 predictors showed the best generalization ability. With a training RMSE of 0.0966 and a test RMSE of 0.1890, it had the lowest error performance among the other stepwise variations. The Mean Absolute Error was found to be 0.0760 for training and 0.1340 for testing, confirming the predictive capability of the model.

Model	RMSE (Train)	MAE (Train)	RMSE (Test)	MAE (Test)
Stepwise (16 p)	0.0966	0.0760	0.1890	0.1340
Stepwise (19 p)	0.0970	0.0764	0.1920	0.1354
Stepwise (22 p)	0.0962	0.0752	0.1993	0.1452

Apparently, the stepwise model with 16 variables also produced a slightly higher test RMSE of 0.1920 compared to others, which indicates that some amount of predictive power must have been lost with the removal of additional variables. Stepwise model of 22 variables had a higher test error of 0.1993, which indicates higher error/generalization towards unseen data.

Typical Profile of a Property Sale

Median SalePrice for an average house in Ames is 160,000 USD, converted from the log scale. Most of the houses are in the price range of 125,000 - 200,000 USD. The median above-ground living area was estimated to be around 6,600 square feet, with basements accounting for more space, averaging about 1,800 square feet. Most homes have garages who's average 479 square feet typically accommodate at least two cars. The most important price variable is the Overall Quality. Most homes fall in the Overall Quality range 5 to 7, indicating above-average construction materials and finishes. Neighbourhood location differentiates property value even further, putting price clusters in premium areas providing better infrastructure. Luxury homes account for prices above 260,000 USD and are considered those in the upper 10% of the price range. They are, by far, quite large, averaging nearly 9,700 square feet in above-ground living area and around 2,600 square feet in basement space. They are made from materials of much higher quality, as indicated by the Overall Quality rating that mostly falls in the range of 7-10. On the other hand, the most affordable houses, those that fall below 120,000 USD, are mostly smaller, with average living spaces of about 4,500 square feet and basements of about 1,300 square feet. These homes are made of relatively cheaper materials, and their Overall Quality ratings are mostly in the range of 4-6.

Conclusion

The Iowa dataset for the houses, combined with exploratory data analysis, feature selection, and statistical modelling, allowed for the successful development of a house price prediction model. Although the model does have high predictive validity, there are still some restrictions. As the dataset is a historical sales record from a particular market, it does not take the economic changes, interest rates, or housing demand shifts into account. In addition, the log transformation of SalePrice did enhance normality, but some non-linear relationships could still exist. Potential improvements could include the incorporation of macroeconomic variables, as well as discovering new non-linear models, and broadening the scope of the study to other real estate markets. Even so, this study serves as an excellent starting point in real estate valuation.