Group10 AmesHousing

Authors

Mattia Zanin – mattia.zanin@studenti.unipd.it

Matteo Giorgi – matteo.giorgi.1@studenti.unipd.it

Enrico Zanello – enrico.zanello@studenti.unipd.it

Luca Lo Buono – luca.lobuono@studenti.unipd.it

November 23, 2025

Contents

1 Research Hypotheses Supported by the Literature

The objective of this work is to study which structural, qualitative, and locational characteristics of a residential property significantly affect its market price.

The *Ames Housing Dataset* provides detailed information on dwellings sold in Ames (Iowa) and is an updated version of the Boston Housing Dataset, already widely used in the housing economics literature. With its 2930 observations, the Ames dataset offers a larger sample size and a more comprehensive set of features, making it better suited for modern statistical analysis.

The dataset includes information regarding physical size, construction details, quality ratings, and neighborhood indicators. Based on these variables, we formulate a set of hypotheses grounded in empirical findings from the real-estate and housing-econometrics literature.

1.1 Fundamental Variables and Assumptions

Previous work shows that approximately 80% of the variation in residential sale prices can be explained simply by considering the neighborhood and the total square footage of the dwelling (computed as `Total Bsmt SF` + `Gr Liv Area`) [DC11].

According to our interpretation of the literature published by [Han23] and [ZZS08], together with our empirical reasoning, we consider the following variables to have a significant impact on the sale price of a house.

1.1.1 Overall Quality Has a Strong Positive Impact on Sale Price

Several studies show that global quality assessments summarize multiple latent characteristics (such as materials, workmanship, and design), making them among the most informative predictors of house value.

[Abd22] identifies `Overall Qual` as one of the most influential variables in explaining sale price in a multiple regression framework.

Similar evidence is reported in [Han23], where overall quality consistently appears as the strongest determinant in both OLS and regularized regression models.

1.1.2 Above-Ground Living Area (`Gr Liv Area`) Is Positively Associated with Sale Price

The hedonic pricing literature traditionally recognizes physical size as a primary contributor to housing value.

> The **hedonic pricing model** is an economic model that explains the price of a good as the combination of the values of its attributes. A complex good is "decomposed" into its components, and the total price reflects the contribution of each of them.
>
> The concept was introduced by Lancaster (1966) in consumer theory and was first applied to housing prices by Rosen (1974).

[Abd22] highlights that the main livable area is among the variables with the highest explanatory power. Studies on the Ames dataset by [Ye24] and [Han23] similarly identify `Gr Liv Area` as one of the strongest continuous predictors.

1.1.3 Total Basement Area (`Total Bsmt SF`) Also Increases Sale Price

Basements provide additional functional space and generally correlate with larger, higher-value homes.

Multiple analyses of the Ames dataset [Han23] show that basement size remains statistically significant even when controlling for other structural variables. [Abd22] also finds that basement-related features contribute substantially to model fit.

1.1.4 Newer or Recently Renovated Houses (`Year Built`, `Year Remod/Add`) Tend to Have Higher Prices

The literature on housing depreciation demonstrates that structural aging reduces property value unless offset by renovations.

According to [Abd22], both the construction year and the remodeling year play an important role in predicting sale prices.

Other studies on the Ames dataset reinforce this conclusion, noting that newer homes or homes with extensive remodeling command a price premium.

1.1.5 Higher-Quality Kitchens and Exterior Materials Positively Influence Sale Price

Studies in real-estate economics show that buyers are particularly sensitive to the quality of kitchens, bathrooms, and exterior finishes, as these elements influence both functionality and aesthetic appeal.

`Kitchen Qual` and `Exter Qual` are repeatedly found to be statistically significant predictors in analyses using the Ames dataset [Ye24, Abd22]. Both variables capture qualitative assessments that are not reflected solely by house size.

1.1.6 Neighborhood Characteristics Significantly Affect Sale Price

Location remains one of the most important determinants of housing prices in hedonic models.

The Ames dataset includes a categorical variable (`Neighborhood`) that encodes proximity to schools, income areas, and local amenities.

Previous studies [DC11, Ye24] demonstrate that location dummies remain significant even in fully specified regression models.

1.2 Variables We Remove

We begin by removing the variables `Pool QC`, `Alley`, `Fence`, and `Misc Feature`, since they contain a large proportion of missing observations (`NaN`). Although the remaining variables would be available for analysis, including irrelevant or noisy predictors increases the standard errors of the estimated coefficients, leading to less powerful significance tests, wider confidence intervals, and ultimately less precise statistical inference.

For this reason, we exclude many of the 82 explanatory variables originally present in the dataset. Several of them refer to the same structural component of the house (e.g., `Bsmt`

Exposure, BsmtFin Type 1, BsmtFin Type 2 all describe basement characteristics), potentially resulting in high multicollinearity and inflated standard errors. Additionally, we believe that some features have only a negligible impact on sale prices. To obtain a more parsimonious model—characterized by a higher adjusted R^2 and lower AIC and BIC, we decided not to include these variables in our analysis.

Examples of excluded variables include:

- PID: an identification number that carries no information about the price.
- Lot Front: the length of the street frontage. Its effect is ambiguous (a larger frontage might be positive, but may also imply more noise and traffic).
- Lot Shape and Land Contour: less relevant than broader locational attributes such as neighborhood.
- Fireplaces: the number of fireplaces, an outdated feature with limited impact on modern housing prices.
- FireplaceQu: fireplace quality, which is of secondary importance relative to other quality indicators (e.g., kitchen and exterior quality).
- Roof Style and Roof Matl: roofing attributes that are less relevant compared to major structural and quality variables.
- Condition 1: proximity to conditions such as arterial roads or railways. Since Neighborhood captures location effects more comprehensively, including this variable may introduce redundancy, and noise may still distort its interpretation.

2 Description of the Dataset

The analysis is based on the *Ames Housing Dataset*, a well-known real-estate dataset originally compiled by the Assessor's Office of Ames, Iowa. It contains detailed information on residential properties sold between 2006 and 2010.

In its full version, the dataset includes 82 explanatory variables describing structural, qualitative, locational, and functional characteristics of each house.

For the purpose of this assignment, we selected a subset of variables most commonly used in hedonic pricing models and supported by the existing literature.

2.1 General Structure of the Dataset

Each row of the dataset corresponds to a single residential property, while columns represent the attributes listed above.

Numerical variables are measured either in square feet or in calendar years, whereas qualitative assessments use an ordinal scale. The dataset does not contain missing values for the variables selected in our analysis, and therefore no imputation was required.

The dependent variable of the regression is SalePrice, expressed in US dollars. The selected explanatory variables fall into the four following categories.

2.1.1 Size and Structural Characteristics

These variables capture the physical dimensions and structural attributes of the dwelling:

- `Gr Liv Area`: above-ground living area (square feet)
- `1st Flr SF`: first-floor area
- `Total Bsmt SF`: total basement area
- `Lot Area`: size of the lot
- `Full Bath`: number of full bathrooms
- `Garage Area`: size of the garage (square feet)
- `Garage Cars`: garage capacity (number of cars)
- `Garage Yr Blt`: year the garage was built
- `Year Built`: construction year of the house
- `Year Remod/Add`: year of the most recent remodeling
- `Utilities`: type of utilities available

These attributes quantify the amount of usable space and reflect the structural age of the property.

2.1.2 Quality Assessments

The dataset includes several ordinal ratings assigned by professional assessors:

- `Overall Qual`: overall material and finish quality
- `Kitchen Qual`: quality of kitchen materials
- `ExterQual`: exterior materials and workmanship
- `BsmtQual`: quality of basement materials

These variables summarize qualitative features that are not captured by size alone.

2.1.3 Locational and Timing Information

- `Neighborhood`: categorical variable identifying the physical location within the city of Ames
- `MS Zoning`: zoning classification (e.g., low-density residential, medium-density residential, commercial)
- `Year Sold`: year the property was sold

These variables incorporate locational amenities and neighbourhood-level characteristics that influence market value.

3 Model Specification and Estimation

The aim of this section is to construct a multiple linear regression model in order to explain the variation in the sale price of residential properties.

The dependent variable is `SalePrice`, while the set of regressors includes the structural, qualitative, and locational characteristics identified in the previous section.

3.1 Model Specification

The regression model is specified as follows:

$$\begin{aligned}\texttt{[SalePrice]}_i = \beta_0 &+ \beta_1 \texttt{[Gr Liv Area]}_i + \beta_2 \texttt{[1st Flr SF]}_i + \beta_3 \texttt{[Total Bsmt SF]}_i \\ &+ \beta_4 \texttt{[Lot Area]}_i + \beta_5 \texttt{[Full Bath]}_i + \beta_6 \texttt{[Garage Area]}_i \\ &+ \beta_7 \texttt{[Garage Cars]}_i + \beta_8 \texttt{[Garage Yr Blt]}_i + \beta_9 \texttt{[Year Built]}_i \\ &+ \beta_{10} \texttt{[Year Remod/Add]}_i + \beta_{12} \texttt{[Overall Qual]}_i + \beta_{13} \texttt{[Kitchen Qual]}_i \\ &+ \beta_{14} \texttt{[Exter Qual]}_i + \beta_{15} \texttt{[Bsmt Qual]}_i + \beta_{16} \texttt{[Neighborhood]}_i + \varepsilon_i\end{aligned}$$

where:

- β_0 is the intercept;
- the β's denote the coefficients associated with each regressor;
- ε_i is the error term capturing unobserved factors;
- the variable `Neighborhood` is treated as a categorical factor and encoded via dummy variables.

3.2 Model Assumptions

Our multiple linear regression model is estimated under the following assumptions:

Linearity of the conditional mean.

$$\mathbb{E}[\varepsilon_i \mid X_i] = 0 \quad \Longleftrightarrow \quad \mathbb{E}[Y_i \mid X_i] = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}.$$

This assumption implies that all systematic variation in the dependent variable is captured by the regressors.

Independent and identically distributed observations.

$$(Y_i, X_i) \perp\!\!\!\perp (Y_j, X_j) \qquad \text{for } i \neq j,$$

and all observations are drawn from the same population distribution. This corresponds to the standard *i.i.d.* sampling framework.

Finite fourth moments.

$$\mathbb{E}[X_{ik}^4] < \infty \qquad \text{and} \qquad \mathbb{E}[\varepsilon_i^4] < \infty.$$

This condition ensures the applicability of asymptotic results and prevents extreme observations from dominating the estimation.

3.3 R Code

Here we report the full R code used for the estimation of the baseline multiple linear regression model, including data preparation and variable renaming.

```r
ames <- read.csv("Amespug.csv")

# First, we rename variables with spaces into more convenient names
library(dplyr)

ames_clean <- ames %>%
  rename(
    GrLivArea     = `Gr.Liv.Area`,
    FirstFlrSF    = `X1st.Flr.SF`,
    TotalBsmtSF   = `Total.Bsmt.SF`,
    LotArea       = `Lot.Area`,
    FullBath      = `Full.Bath`,
    GarageArea    = `Garage.Area`,
    GarageCars    = `Garage.Cars`,
    GarageYrBlt   = `Garage.Yr.Blt`,
    YearBuilt     = `Year.Built`,
    YearRemodAdd  = `Year.Remod.Add`,
    OverallQual   = `Overall.Qual`,
    KitchenQual   = `Kitchen.Qual`,
    ExterQual     = `Exter.Qual`,
    BsmtQual      = `Bsmt.Qual`,
    Neighborhood  = Neighborhood
  )

# Linear regression
model <- lm(
  SalePrice ~ GrLivArea + FirstFlrSF + TotalBsmtSF + LotArea +
    FullBath + GarageArea + GarageCars + GarageYrBlt +
    YearBuilt + YearRemodAdd + OverallQual +
    KitchenQual + ExterQual + BsmtQual +
    Neighborhood,
  data = ames_clean
)

# Results
summary(model)
```

3.4 Model Output

```
Call:
lm(formula = SalePrice ~ GrLivArea + FirstFlrSF + TotalBsmtSF +
    LotArea + FullBath + GarageArea + GarageCars + GarageYrBlt +
    YearBuilt + YearRemodAdd + OverallQual + KitchenQual + ExterQual +
    BsmtQual + Neighborhood, data = ames_clean)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-458670  -12548     282   12557  233675

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -7.243e+05  1.326e+05  -5.463 5.11e-08 ***
GrLivArea            4.356e+01  2.052e+00  21.226  < 2e-16 ***
FirstFlrSF           8.657e+00  3.725e+00   2.324 0.020179 *
TotalBsmtSF          9.625e+00  3.423e+00   2.812 0.004955 **
LotArea              6.445e-01  8.528e-02   7.557 5.62e-14 ***
FullBath            -2.970e+03  1.688e+03  -1.760 0.078553 .
GarageArea           2.026e+01  6.717e+00   3.017 0.002577 **
GarageCars           6.594e+03  1.951e+03   3.379 0.000737 ***
GarageYrBlt         -6.674e+01  4.839e+01  -1.379 0.167965
YearBuilt            2.262e+02  5.792e+01   3.905 9.66e-05 ***
YearRemodAdd         2.445e+02  4.579e+01   5.339 1.01e-07 ***
OverallQual          1.199e+04  8.413e+02  14.248  < 2e-16 ***
KitchenQualFa       -3.413e+04  5.877e+03  -5.808 7.08e-09 ***
KitchenQualGd       -2.820e+04  3.201e+03  -8.811  < 2e-16 ***
KitchenQualPo       -5.463e+04  3.143e+04  -1.738 0.082314 .
KitchenQualTA       -3.437e+04  3.581e+03  -9.597  < 2e-16 ***
ExterQualFa         -4.054e+04  8.993e+03  -4.508 6.83e-06 ***
ExterQualGd         -2.722e+04  4.118e+03  -6.609 4.66e-11 ***
ExterQualTA         -2.682e+04  4.665e+03  -5.749 9.98e-09 ***
BsmtQualFa          -3.179e+04  5.268e+03  -6.034 1.82e-09 ***
BsmtQualGd          -3.014e+04  2.833e+03 -10.639  < 2e-16 ***
BsmtQualPo          -2.953e+04  2.253e+04  -1.311 0.189988
BsmtQualTA          -2.952e+04  3.540e+03  -8.337  < 2e-16 ***
NeighborhoodBlueste -7.475e+03  1.152e+04  -0.649 0.516562
NeighborhoodBrDale  -1.612e+04  8.670e+03  -1.860 0.063023 .
NeighborhoodBrkSide  3.168e+03  7.520e+03   0.421 0.673574
NeighborhoodClearCr  2.066e+04  7.975e+03   2.591 0.009618 **
NeighborhoodCollgCr  9.051e+03  6.240e+03   1.450 0.147055
NeighborhoodCrawfor  2.810e+04  7.158e+03   3.926 8.86e-05 ***
NeighborhoodEdwards -1.040e+04  6.894e+03  -1.508 0.131716
NeighborhoodGilbert  4.753e+03  6.486e+03   0.733 0.463774
NeighborhoodGreens   1.412e+04  1.249e+04   1.131 0.258345
NeighborhoodGrnHill  1.403e+05  3.127e+04   4.486 7.57e-06 ***
NeighborhoodIDOTRR  -6.567e+03  7.860e+03  -0.835 0.403551
NeighborhoodLandmrk -1.301e+04  3.129e+04  -0.416 0.677492
NeighborhoodMeadowV -9.408e+03  8.943e+03  -1.052 0.292867
NeighborhoodMitchel  2.414e+03  6.884e+03   0.351 0.725840
NeighborhoodNAmes    3.094e+03  6.592e+03   0.469 0.638863
NeighborhoodNoRidge  6.035e+04  7.163e+03   8.426  < 2e-16 ***
NeighborhoodNPkVill -5.402e+03  8.941e+03  -0.604 0.545811
NeighborhoodNridgHt  3.070e+04  6.575e+03   4.670 3.16e-06 ***
NeighborhoodNWAmes   2.648e+03  6.736e+03   0.393 0.694222
NeighborhoodOldTown -9.368e+03  7.297e+03  -1.284 0.199312
NeighborhoodSawyer   4.100e+03  6.884e+03   0.596 0.551549
NeighborhoodSawyerW  2.807e+03  6.660e+03   0.421 0.673440
```

```
NeighborhoodSomerst  1.341e+04  6.358e+03   2.109 0.035067 *
NeighborhoodStoneBr  5.038e+04  7.401e+03   6.807 1.23e-11 ***
NeighborhoodSWISU   -2.665e+03  8.445e+03  -0.315 0.752407
NeighborhoodTimber   1.464e+04  7.001e+03   2.092 0.036558 *
NeighborhoodVeenker  2.531e+04  8.710e+03   2.905 0.003697 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '. 0.1 ' ' 1

Residual standard error: 30630 on 2655 degrees of freedom
  (225 observations deleted due to missingness)
Multiple R-squared:  0.8543,    Adjusted R-squared:  0.8516
F-statistic: 317.8 on 49 and 2655 DF,  p-value: < 2.2e-16
```

3.5 Interpretation of the Regression Output

The OLS regression is estimated on 2,770 observations.[1] The dependent variable is `SalePrice`, while the regressors include size-related variables, quality indicators, age and renovation variables, and a set of dummy variables for neighbourhood.

3.5.1 Overall Fit of the Model

The model exhibits a high goodness of fit:

- R-squared = 0.8543
- Adjusted R-squared = 0.8516

This means that approximately 85% of the variation in house prices in Ames is explained by the regressors included in the model.

The overall F-statistic is 317.8 (with $p < 2.2 \times 10^{-16}$), strongly rejecting the null hypothesis that all slope coefficients are jointly equal to zero. Thus, the explanatory variables collectively have a statistically significant effect on `SalePrice`.

3.5.2 Interpretation of the Main Coefficients

The first group of regressors captures size, structural characteristics, and age. The main findings are summarised below.

Living Area and Basement Space. `GrLivArea` has an estimated coefficient of approximately 43.6 ($p < 2 \times 10^{-16}$). This implies that, *ceteris paribus*, an additional square foot of above-ground living area increases the expected sale price by about \$44.

The coefficient on `TotalBsmtSF` is about 9.6 ($p = 0.0049$), indicating that basement space is also valued by the market, though less than above-ground living area.

Lot Size. `LotArea` enters with a small but positive and highly significant coefficient (0.64, $p < 10^{-13}$), suggesting that marginal increases in lot size add value, though at a much lower rate than interior space.

[1] The original dataset contains 2,930 observations; 225 are removed due to missing values in the selected regressors.

Bathrooms and Garage Capacity. The coefficient on `FullBath` is negative ($-2{,}970$, p = 0.078), borderline significant. This counterintuitive sign is likely explained by multicollinearity: once living area and quality are included, an additional bathroom at fixed size may not capture higher quality.

`GarageArea` and `GarageCars` are both positive and significant (20.3 and 6,594 respectively), indicating that larger garages and capacity for more vehicles contribute positively to sale price.

Construction and Renovation Year. `YearBuilt` (226, p 9.7×10^{-5}) and `YearRemodAdd` (244, p 10^{-7}) are both positive and highly significant: newer homes and recently renovated properties command higher prices, consistent with depreciation and modernization effects.

Quality Indicators. `OverallQual` is one of the strongest predictors in the model, with an estimated coefficient of roughly 11,990 ($p < 2 \times 10^{-16}$). A one-point increase in overall quality raises the predicted sale price by about \$12k, confirming the central importance of quality emphasised in the hedonic housing literature.

The categorical quality variables `KitchenQual`, `ExterQual` and `BsmtQual` are encoded through dummy variables. Since the reference category corresponds to the highest quality level, most dummy coefficients are negative and highly significant, reflecting the price discounts associated with lower kitchen, exterior or basement quality.

Neighbourhood Effects. A large block of coefficients corresponds to neighbourhood dummies. Some neighbourhoods show large, positive and highly significant premia relative to the omitted reference category — for example:

- `NeighborhoodGrnHill`: +140,300 (p 7.6×10^{-6})
- `NeighborhoodNoRidge`: +60,350 ($p < 2 \times 10^{-16}$)
- `NeighborhoodNridgHt`: +30,700 (p 3.2×10^{-6})
- `NeighborhoodStoneBr`: +50,380 (p 1.2×10^{-11})

These coefficients indicate that location plays a substantial role in determining housing values, even after controlling for structure, size and quality.

Other neighbourhoods show coefficients statistically indistinguishable from zero, indicating that once structural and qualitative variables are controlled for, price differences across those areas are not significant.

Zoning and Utilities. In this specification, zoning categories (`MS Zoning`) and utility types do not exhibit statistically significant effects, suggesting that these factors add little explanatory power beyond size, quality and location.

4 Diagnostic Analysis

4.1 Perfect Collinearity

Perfect collinearity arises when one explanatory variable can be expressed as an exact linear combination of others. Among all possible specification errors, this is the most severe, because

perfect collinearity makes the OLS estimator *impossible* to compute: the design matrix becomes singular and the matrix $(X^\top X)^{-1}$ required for OLS does not exist. Other specification issues (such as omitted variables, functional form errors, or heteroskedasticity) typically affect the desirable properties of the OLS estimator under the Gauss–Markov assumptions, but do not prevent its computation. Perfect collinearity, instead, makes estimation itself undefined.

In the specification of our main model, perfect collinearity does not occur. However, the *Ames Housing Dataset* contains a well-known identity that can generate perfect multicollinearity if the analyst is not careful.

Consider the four area variables:

$$\texttt{GrLivArea} = \texttt{FirstFlrSF} + \texttt{SecondFlrSF} + \texttt{LowQualFinSF}$$

If all observations satisfy this identity exactly, which is almost always the case in the dataset, then including all four variables simultaneously in a regression introduces an exact linear dependence. In such a case, OLS cannot distinguish the marginal contribution of each component because they lie on the same hyperplane in the regressor space.

```
# Perfect multicollinearity test
ames_multicollineareperfetto <- ames %>%
  rename(
    GrLivArea    = `Gr.Liv.Area`,
    FirstFlrSF   = `X1st.Flr.SF`,
    SecondFlrSF  = `X2nd.Flr.SF`,
    LowQualFinSF = `Low.Qual.Fin.SF`,
    TotalBsmtSF  = `Total.Bsmt.SF`,
    LotArea      = `Lot.Area`,
    FullBath     = `Full.Bath`,
    GarageArea   = `Garage.Area`,
    GarageCars   = `Garage.Cars`,
    GarageYrBlt  = `Garage.Yr.Blt`,
    YearBuilt    = `Year.Built`,
    YearRemodAdd = `Year.Remod.Add`,
    OverallQual  = `Overall.Qual`,
    KitchenQual  = `Kitchen.Qual`,
    ExterQual    = `Exter.Qual`,
    BsmtQual     = `Bsmt.Qual`,
    Neighborhood = `Neighborhood`,
    MSZoning     = `MS.Zoning`,
    Utilities    = `Utilities`
  )

# Linear model with potential perfect collinearity
modelmulticollperf <- lm(
  SalePrice ~ GrLivArea + FirstFlrSF + SecondFlrSF + LowQualFinSF +
    TotalBsmtSF + LotArea + FullBath + GarageArea + GarageCars +
    GarageYrBlt + YearBuilt + YearRemodAdd + OverallQual +
    KitchenQual + ExterQual + BsmtQual +
```

```
    Neighborhood + MSZoning + Utilities,
  data = ames_multicollineareperfetto
)

summary(modelmulticollperf)
```

```
Call:
lm(formula = SalePrice ~ GrLivArea + FirstFlrSF + SecondFlrSF +
    LowQualFinSF + TotalBsmtSF + LotArea + FullBath + GarageArea +
    GarageCars + GarageYrBlt + YearBuilt + YearRemodAdd + OverallQual +
    KitchenQual + ExterQual + BsmtQual + Neighborhood + MSZoning +
    Utilities, data = ames_multicollineareperfetto)

Residuals:
    Min      1Q  Median      3Q     Max
-457007  -12709     135   12330  233868

Coefficients: (1 not defined because of singularities)
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -7.548e+05  1.344e+05  -5.617 2.15e-08 ***
GrLivArea            2.280e+01  1.433e+01   1.591 0.111753
FirstFlrSF           2.976e+01  1.484e+01   2.005 0.045045 *
SecondFlrSF          2.088e+01  1.441e+01   1.449 0.147467
LowQualFinSF                NA         NA      NA       NA
TotalBsmtSF          9.048e+00  3.423e+00   2.644 0.008251 **
LotArea              6.469e-01  8.584e-02   7.535 6.64e-14 ***
FullBath            -3.158e+03  1.688e+03  -1.871 0.061456 .
GarageArea           1.807e+01  6.767e+00   2.670 0.007621 **
GarageCars           7.132e+03  1.962e+03   3.635 0.000283 ***
GarageYrBlt         -6.638e+01  4.847e+01  -1.370 0.170952
YearBuilt            2.351e+02  5.839e+01   4.026 5.84e-05 ***
YearRemodAdd         2.398e+02  4.571e+01   5.246 1.68e-07 ***
OverallQual          1.197e+04  8.430e+02  14.201  < 2e-16 ***
KitchenQualFa       -3.396e+04  5.864e+03  -5.792 7.77e-09 ***
KitchenQualGd       -2.793e+04  3.191e+03  -8.751  < 2e-16 ***
KitchenQualPo       -5.436e+04  3.132e+04  -1.736 0.082768 .
KitchenQualTA       -3.386e+04  3.574e+03  -9.473  < 2e-16 ***
ExterQualFa         -3.919e+04  9.015e+03  -4.347 1.43e-05 ***
ExterQualGd         -2.738e+04  4.105e+03  -6.670 3.10e-11 ***
ExterQualTA         -2.738e+04  4.653e+03  -5.884 4.50e-09 ***
BsmtQualFa          -3.345e+04  5.268e+03  -6.350 2.53e-10 ***
BsmtQualGd          -3.063e+04  2.828e+03 -10.834  < 2e-16 ***
BsmtQualPo          -2.087e+04  2.375e+04  -0.879 0.379561
BsmtQualTA          -3.018e+04  3.538e+03  -8.531  < 2e-16 ***
...
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 30520 on 2647 degrees of freedom
(225 observations deleted due to missingness)
Multiple R-squared:  0.8558,    Adjusted R-squared:  0.8527
F-statistic: 275.6 on 57 and 2647 DF,  p-value: < 2.2e-16
```

The standard remedy is to remove *one* of the linearly dependent variables. Since the three components sum exactly to `GrLivArea`, the choice of which variable to drop is irrelevant from an informational standpoint: the remaining variables still span the same linear space.

Statistical software such as `R` automatically detects this issue and removes one of the problematic regressors. In our multicollinearity test, `R` excluded `LowQualFinSF`, which can be verified from the regression output: the estimated coefficient and its standard error appear as `NA`, accompanied by the message *"1 not defined because of singularities"*. This confirms that the dataset indeed contains an exact linear relationship among area variables, and that perfect collinearity is handled automatically by the software when these variables are included together.

4.2 Imperfect Multicollinearity: Variance Inflation Factors

Imperfect, or near, multicollinearity arises when two or more explanatory variables are highly correlated, although not perfectly linearly dependent. In this situation, OLS estimation is still feasible, but the variance of the affected coefficients becomes inflated. As a consequence, standard errors increase, leading to wider confidence intervals and less informative hypothesis tests. Inferences about the individual significance of regressors may therefore become unreliable, even when the variables are conceptually relevant.

A common solution consists in removing or combining regressors that are strongly correlated with others. To detect the presence and severity of near-collinearity, we rely on the *Variance Inflation Factor* (VIF), which corresponds to the second term in the variance of an OLS coefficient and serves as an alternative to examining pairwise sample correlations among regressors.

For a given regressor X_j, the VIF is defined as:

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the coefficient of determination from the auxiliary regression of X_j on the remaining regressors. A large VIF value indicates that X_j is strongly correlated with other regressors, implying increased variance of its estimated coefficient and, consequently, less precise inference.

```r
library(performance)
vif_values <- check_collinearity(model)
print(vif_values)
```

```
                GVIF Df GVIF^(1/(2*Df))
GrLivArea   3.043312  1        1.744509
FirstFlrSF  6.121783  1        2.474224
TotalBsmtSF 5.693781  1        2.386165
LotArea     1.358761  1        1.165659
FullBath    2.492432  1        1.578744
GarageArea  4.642314  1        2.154603
GarageCars  4.666852  1        2.160290
GarageYrBlt 4.414460  1        2.101062
YearBuilt   8.445562  1        2.906125
YearRemodAdd 2.536321  1        1.592583
OverallQual 3.741132  1        1.934201
```

```
KitchenQual   5.173766  4         1.228078
ExterQual     6.257313  3         1.357473
BsmtQual      7.135004  4         1.278422
Neighborhood 41.329522 27         1.071348
```

The literature offers different threshold values for interpreting VIFs, with common rules of thumb ranging from 5 to 10 as indicators of problematic multicollinearity. In models including categorical variables with more than two levels, `R` reports the *Generalized Variance Inflation Factor* (GVIF), an extension of the VIF that properly accounts for multiple degrees of freedom. This adjustment is necessary because a categorical variable with k levels expands into $(k-1)$ dummy variables, for which the classical VIF cannot be computed directly.

For variables with one degree of freedom, GVIF coincides with the standard VIF. For categorical regressors with more than one degree of freedom, GVIF must be transformed using the formula:

$$\text{GVIF}^{\,1/(2\,\text{df})}$$

which yields a value comparable to the standard VIF. This explains why some categorical variables appear to have extremely high GVIFs: the large values simply reflect their number of dummy variables rather than true collinearity issues.

For example, `Neighborhood` exhibits a very large GVIF (572.17), but once transformed it corresponds to a reasonable VIF of approximately 1.72. Similarly, `MSZoning` shows a GVIF of 26.42 but a transformed VIF of just 1.39.

Since none of the transformed VIF values exceeds the common threshold of 10, nor comes close to it, we conclude that multicollinearity is not a serious concern for our current specification, and all the included regressors can be safely retained.

4.3 Structural Stability: Chow Test around the 2008 Financial Crisis

In this section we investigate whether the relationship between the explanatory variables and `SalePrice` remained stable over time. Structural changes in the coefficients may arise for several reasons, and sharp economic events, such as financial crises, are among the most common sources of such instability. Since our dataset covers housing transactions between 2006 and 2010, it is natural to test for a structural break around 2008, the year associated with the outbreak of the Global Financial Crisis. The Chow test evaluates precisely this hypothesis.

Let RSS_P and RSS_C denote the residual sum of squares of the model estimated separately on the pre-crisis and the crisis/post-crisis subsamples, and let $RSS_{P\cup C}$ be the residual sum of squares of the model estimated on the full pooled sample. Denoting by k the number of estimated parameters (including the intercept), and by n_P and n_C the sample sizes of the two subsamples, the Chow test statistic is:

$$F \;=\; \frac{RSS_{P\cup C} - (RSS_P + RSS_C)}{k} \Bigg/ \frac{RSS_P + RSS_C}{n_P + n_C - 2k},$$

which, under the null hypothesis of parameter stability, follows an F-distribution with $(k, n_P + n_C - 2k)$ degrees of freedom.

In practice, we rely on the implementation provided by the `strucchange` package in `R`, which evaluates a Chow-type test at the break point corresponding to the first sale in 2008.

```r
# Chow-type structural stability test
library(sandwich)
library(strucchange)

# Ensure ordered factors for quality variables
ames_clean$OverallQual <- ordered(ames_clean$OverallQual, levels = 1:10)

ames_clean$KitchenQual <- ordered(
  ames_clean$KitchenQual,
  levels = c("Po", "Fa", "TA", "Gd", "Ex")
)

ames_clean$ExterQual <- ordered(
  ames_clean$ExterQual,
  levels = c("Po", "Fa", "TA", "Gd", "Ex")
)

ames_clean$BsmtQual <- ordered(
  ames_clean$BsmtQual,
  levels = c("Po", "Fa", "TA", "Gd", "Ex")
)

# Order data by year of sale
ames_ordered <- ames_clean[order(ames_clean$Yr.Sold), ]

# Baseline model for structural break test (without Neighborhood)
modelordered <- lm(
  SalePrice ~ GrLivArea + FirstFlrSF + TotalBsmtSF + LotArea +
    FullBath + GarageArea + GarageCars + GarageYrBlt +
    YearBuilt + YearRemodAdd + OverallQual +
    KitchenQual + ExterQual + BsmtQual,
  data = ames_ordered
)

# Break point: first observation with Yr.Sold == 2008
break_point <- min(which(ames_ordered$Yr.Sold == 2008))

# Chow-type test at the given break point
chow_test <- sctest(modelordered, type = "Chow", point = break_point)
print(chow_test)

# Separate models pre- and post-2008 (including Neighborhood)
model_pre2008 <- lm(
  SalePrice ~ GrLivArea + FirstFlrSF + TotalBsmtSF + LotArea +
    FullBath + GarageArea + GarageCars + GarageYrBlt +
```

```
    YearBuilt + YearRemodAdd + OverallQual +
    KitchenQual + ExterQual + BsmtQual +
    Neighborhood,
  data = subset(ames_clean, Yr.Sold <= 2008)
)

model_post2008 <- lm(
  SalePrice ~ GrLivArea + FirstFlrSF + TotalBsmtSF + LotArea +
    FullBath + GarageArea + GarageCars + GarageYrBlt +
    YearBuilt + YearRemodAdd + OverallQual +
    KitchenQual + ExterQual + BsmtQual +
    Neighborhood,
  data = subset(ames_clean, Yr.Sold > 2008)
)

summary(model_pre2008)
summary(model_post2008)
```

```
M-fluctuation test

data:  modelordered
f(efp) = 1.3035, p-value = 0.8746


Call:
lm(formula = SalePrice ~ GrLivArea + FirstFlrSF + TotalBsmtSF +
    LotArea + FullBath + GarageArea + GarageCars + GarageYrBlt +
    YearBuilt + YearRemodAdd + OverallQual + KitchenQual + ExterQual +
    BsmtQual + Neighborhood, data = subset(ames_clean, Yr.Sold <= 2008))

...
Residual standard error: 31660 on 1738 degrees of freedom
Multiple R-squared:  0.8485,  Adjusted R-squared:  0.8438

Call:
lm(formula = SalePrice ~ GrLivArea + FirstFlrSF + TotalBsmtSF +
    LotArea + FullBath + GarageArea + GarageCars + GarageYrBlt +
    YearBuilt + YearRemodAdd + OverallQual + KitchenQual + ExterQual +
    BsmtQual + Neighborhood, data = subset(ames_clean, Yr.Sold > 2008))

...
Residual standard error: 25200 on 857 degrees of freedom
Multiple R-squared:  0.9026,  Adjusted R-squared:  0.8966
```

The test yields a very large p-value (0.8746), so we fail to reject the null hypothesis of parameter stability. At first sight, this may appear counter-intuitive, given that the financial crisis was closely linked to a speculative bubble in housing markets. However, this result suggests an important distinction: the crisis may have shifted the *level* of house prices without altering the *structure* of the underlying pricing relationship.

To assess this interpretation, we estimate two separate regressions:

- **pre-crisis period**: houses sold between 2006 and 2008
- **post-crisis period**: houses sold between 2009 and 2010

Comparing the estimated intercepts across the two models shows that they differ substantially, with the post-crisis regression exhibiting a markedly lower intercept. This indicates that the crisis primarily caused a downward shift in the overall price level, while the marginal effects of structural and qualitative characteristics remained broadly stable.

In other words, the crisis affected *how much* houses sold for, but not *why* certain houses sold for more than others. The fundamental price determinants captured by our model appear structurally robust before and after 2008.

4.4 Functional Form Misspecification: RESET Test

We now investigate whether the linear functional form assumed in our baseline specification is adequate. A classical diagnostic tool for this purpose is the Ramsey RESET test, which evaluates whether nonlinear combinations of the fitted values possess explanatory power beyond the original regressors. A low p-value indicates rejection of the null hypothesis of correct functional form.

In our case, the RESET test produces a very small p-value, leading us to reject the null hypothesis of linearity. This result suggests the presence of functional form misspecification. Nonlinearities may arise either from structural breaks or from omitted variables. As discussed in Section **??**, the Chow test shows no evidence of a structural break around the 2008 financial crisis, which we consider the most plausible breakpoint. Therefore, it is reasonable to attribute the nonlinearity detected by the RESET test to omitted relevant terms, such as interactions or nonlinear transformations of existing regressors.

Ordinal Regressors and Nonlinearity. A central issue concerns the ordinal regressors included in the model, namely `OverallQual`, `KitchenQual`, `ExterQual`, and `BsmtQual`. Although `R` initially treats these variables as unordered categorical factors, they possess a meaningful ranking from "Poor" to "Excellent". By explicitly transforming them into ordered factors, the model is able to capture potential nonlinear effects across successive quality levels rather than treating the dummy coefficients as independent of one another.

The variable `OverallQual` deserves particular attention. Its numeric scale from 1 to 10 might appear continuous, but the values represent subjective assessments of overall quality rather than cardinal quantities. Treating it as an ordered factor instead of a numerical variable allows the model to accommodate nonlinear, possibly non-monotonic differences across quality levels.

Interaction Terms. To address additional sources of nonlinearity, we also introduce two economically motivated interaction terms:

$$\texttt{GrLivArea} \times \texttt{OverallQual}, \qquad \texttt{GarageArea} \times \texttt{GarageCars}.$$

These interactions have a clear economic interpretation. A large but low-quality house may be worth less than a smaller but high-quality one; without the interaction between livable

area and quality, this effect would be entirely missed by a linear model. Similarly, the value contributed by a large garage depends on its capacity: a wide two-car garage is typically worth more than a very large one-car garage.

Logarithmic Transformations. Finally, we apply logarithmic transformations to the dependent variable and several continuous regressors. This choice is supported by economic reasoning: in a log–log specification, coefficients represent elasticities, i.e., percentage changes in `SalePrice` associated with percentage changes in the regressors. This helps capture diminishing marginal effects. For example, a 1% increase in livable area is likely to have a larger impact on the price of a small house than on the price of a very large one, where additional space contributes less to perceived value relative to other features such as quality.

Taken together, these considerations support the conclusion that the baseline linear specification is not sufficiently flexible and that incorporating ordered factors, nonlinear transformations, and economically meaningful interactions is both statistically justified and economically interpretable.

4.5 Functional Form Misspecification: RESET Test

To assess whether the linear functional form assumed in our baseline model is appropriate, we apply the Ramsey RESET test. The test evaluates whether higher-order polynomials of the fitted values carry additional explanatory power, which would indicate misspecification due to omitted nonlinearities or functional form errors.

The RESET test on the baseline linear model rejects the null hypothesis of correct specification, with a p-value essentially equal to zero. This suggests that the model suffers from functional form misspecification. As discussed previously, the Chow test did not reveal evidence of structural breaks, so nonlinearity is unlikely to arise from instability over time. Instead, the RESET result points towards the omission of nonlinear transformations or interaction effects between regressors.

A relevant issue concerns the ordinal regressors in the dataset—`OverallQual`, `KitchenQual`, `ExterQual`, and `BsmtQual`—which carry an inherent ordering from "Poor" to "Excellent". Treating them as ordered factors allows the model to capture nonlinear patterns between successive levels rather than interpreting the quality categories as unrelated dummy variables. The variable `OverallQual`, despite being coded numerically from 1 to 10, is also ordinal in nature and is therefore transformed accordingly.

To address the functional misspecification, we consider two adjustments: (i) a log–log model, which provides elasticities and captures diminishing marginal effects; and (ii) the addition of economically meaningful interaction terms, such as `GrLivArea` $\times$ `OverallQual` and `GarageArea` $\times$ `GarageCars`. These interactions account for cases where the effect of size depends strongly on quality, a relationship supported by housing market intuition.

We test both alternative specifications using the RESET test.

4.6 Heteroskedasticity and Independence of Residuals

The Gauss–Markov theorem relies on two crucial assumptions:

- homoskedasticity of the error term
- lack of serial correlation among residuals

If these assumptions are violated, the OLS estimator remains unbiased but it is no longer efficient, and standard inferential procedures based on classical standard errors become unreliable.

4.6.1 Heteroskedasticity

Given that our dataset is cross-sectional, heteroskedasticity (non-constant error variance across observations) is a natural concern. To formally assess this, we conduct the White test for heteroskedasticity. This test evaluates whether the variance of the residuals depends on the regressors, their squares, or their pairwise interactions. Significant coefficients in the auxiliary regression imply that the error variance systematically varies with one or more explanatory variables.

The White test produces a very small p-value, leading us to reject the null hypothesis of homoskedasticity. We therefore conclude that heteroskedasticity is present in the model.

A standard remedy is to compute heteroskedasticity-robust standard errors, such as the HAC (heteroskedasticity and autocorrelation consistent) estimator of the covariance matrix. Using robust standard errors allows us to maintain valid inference even when the homoskedasticity assumption is violated.

4.6.2 Independence of Residuals

Testing for serial correlation in a cross-sectional framework is generally inappropriate, since the ordering of observations does not correspond to a meaningful time dimension. The Durbin–Watson test, in particular, is designed for time-series or panel data where residuals may be correlated with their past values.

To investigate whether meaningful temporal autocorrelation could be identified in this dataset, we attempted to construct a pseudo-panel by identifying houses that appear multiple times (same `PID`) and were sold in different years (`Yr.Sold`). However, this procedure returned zero matches: no house in the dataset appears to have been sold more than once between 2006 and 2010. Thus, serial correlation cannot meaningfully arise from repeated observations on the same units.

Nevertheless, computing the Durbin–Watson statistic on the model's residuals yields a value of approximately 1.84, which is close to the theoretical value of 2 under the null hypothesis of no first-order autocorrelation. We therefore find no evidence of serial correlation in the residuals, consistent with the cross-sectional nature of the data.

5 Model Refinement

Explain any variable selection, transformations, or alternative model specifications used to improve performance.

6 Conclusions

Summarize findings and relate them back to your initial hypotheses.

References

[Abd22] Azad Abdulhafedh. Incorporating multiple linear regression in predicting the house prices using a big real estate dataset with 80 independent variables. *Open Access Library Journal*, 2022.

[DC11] Dean De Cock. Ames, Iowa: alternative to the Boston housing data as an end of semester segression project. *Journal of Statistics Education*, 2011.

[Han23] Yueting Han. Price prediction of Ames housing through advanced regression techniques. *BCP Business & Management*, 2023.

[Ye24] Qiongwei Ye. House price prediction using machine learning for Ames, Iowa. *Applied and Computational Engineering*, 2024.

[ZZS08] Joachim Zietz, Emily Norman Zietz, and Stacy Sirmans. Determinants of house prices: a quantile regression approach. *The Journal of Real Estate Finance and Economics*, 2008.