

# Housing Prices: Multiple Regression Prediction

Melissa Chen, Dylan Whitmire,  
Vanessa Convers, Neha Jumaní  
Data 110: Professor Harlin Lee



SCHOOL OF DATA SCIENCE  
AND SOCIETY

## Objective

The project aims to build a reliable **multiple linear regression** model for predicting house sale prices. This is achieved by:

- **Optimizing Predictors:** Rigorously transforming and filtering features to manage.
- **Model Building:** Employing Standardized Linear Regression on selected features to establish a predictive relationship between property characteristics and SalePrice.
- **Validating Investment Value:** Using hypothesis testing to formally prove the statistically significant, independent price impact of key features.

The final product is a predictive tool using data-driven rules for smart real estate investment choices.

## Data Preparation and Exploration

Our initial exploration focused on understanding the structure of the dataset and identifying variables strongly related to the target variable, SalePrice.

- **Initial Feature Assessment:** We identified and noted non-ordinal variables for removal since they were unsuitable for direct inclusion in a standard linear regression model.
- **Feature Filtering via Correlation:** We calculated the correlation between all original features and SalePrice (see graph below). We filter the dataset by keeping only the predictors that demonstrated a clear relationship (i.e.  $|\text{corr}| > 0.3$ ).

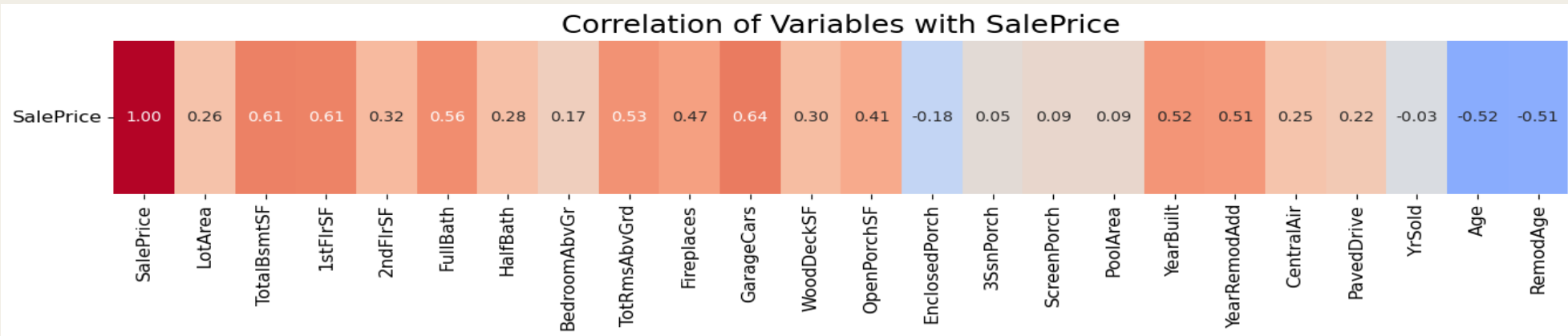


Fig 1.1

This stage ensured our data was clean, numerical, and met the assumptions for a stable multiple regression model.

- **Cleaning:** We removed columns with poor data quality or non-ordinal types.
- **Encoding:** Binary categorical variables and luxury features with class imbalance were mapped into binary.
- **Age Features:** We engineered new age-related predictors from the year columns for better capture.
- **Multicollinearity Mitigation:** A pairwise correlation matrix was used to identify highly correlated predictors (see fig. 1.2). We strategically removed highly dependent features to mitigate multicollinearity.

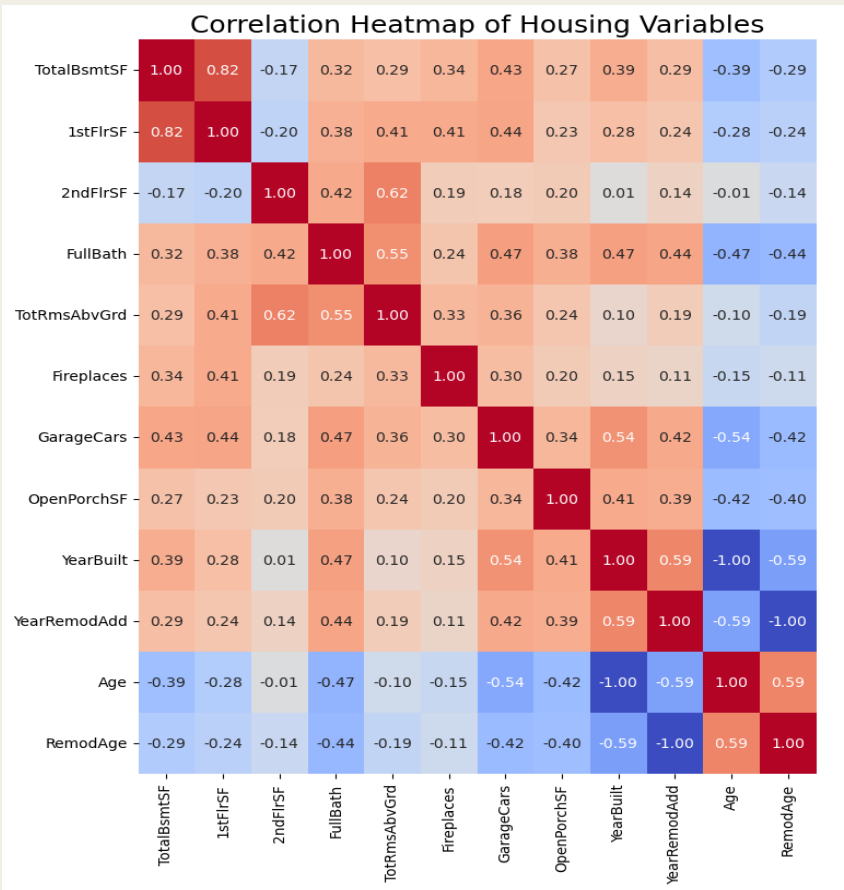


Fig. 1.2

## Regression Model Building

We split the dataset into training (90%) and testing (10%) subsets and developed a core linear model using two approaches:

- **Target Variable:** The dependent variable is **SalePrice**.
- **Predictor Variables:** Our final model uses a refined set of features: **1stFlrSF**, **2ndFlrSF**, **FullBath**, **Fireplaces**, **GarageCars**, **OpenPorchSF**, and **Age**.
- **Baseline Model Construction:** An initial linear regression model was built using the cleaned, non-standardized features.
- **Standardized Model Development:** To enhance model stability and enable the direct comparison of feature importance, all predictors were standardized using Z-scores.

## Regression Model Evaluation

**Performance Assessment:** Both the baseline and standardized models achieved strong predictive power, with the standardized model yielding a high  $R^2$  score

**Overfit Assessment:** Both the train and test scores are similar in value, so the model is neither overfitting or underfitting.

**Feature Importance:** By analyzing the magnitude of standardized coefficients, we identified **1stFlrSF**, **2ndFlrSF**, and **Age** as the most influential factors driving SalePrice.

**Multicollinearity Diagnostics:** A Variance Inflation Factor (VIF) analysis confirmed severe multicollinearity remained among core features. This instability resulted in an abnormal negative coefficient on FullBath in the multiple regression, despite its positive correlation in simple, univariate plots. This is a clear example of the suppression effect.

Fig. 2.1

Features	Baseline Coefficients	Normalized Coefficients	VIF
1stFlrSF	98.612153	37806.903032	18.7859
2ndFlrSF	65.724110	28789.907886	15.7457
FullBath	-3375.551315	-1468.700543	10.1598
Fireplaces	12636.826823	8015.583393	2.9817
GarageCars	20309.040344	15692.821506	2.9033
OpenPorchSF	7649.813905	3342.047712	2.5383
Age	-662.831892	20142.956304	2.2587

**Baseline  $R^2$  scores:** Train, 0.710; Test, 0.761

**Normalized  $R^2$  scores:** Train, 0.713; Test, 0.738

## Hypothesis Testing

To move beyond correlation and prove investment value, we performed two rigorous statistical testing.

**Central Air Test (T-Test):** We formally tested the impact of **CentralAir** (binary indicator) on SalePrice.

- **Null Hypothesis:** Central Air has no impact on price
- **Alternative Hypothesis:** Central Air has a significant impact on price
- **Finding:** The test resulted in an extremely small P-value ( $1.81e-22$ ), leading us to reject the Null Hypothesis.
- **Conclusion:** Central Air is a statistically significant, positive value-adding feature, validating its importance as a predictor.

**Remodeling Status Test (Permutation Test):** We performed a 10,000-iteration permutation test to determine if **IsRemodeled** (binary indicator) leads to a statistically significant price increase.

- **Methodology:** This involved randomly shuffling the SalePrice values across the "remodeled" and "original" groups 10,000 times to simulate the Null Hypothesis (no difference). We compared the observed price difference to this simulated null distribution (see fig. 3.1).
- **Conclusion:** This test confirms that a remodel leads to a statistically significant price increase.

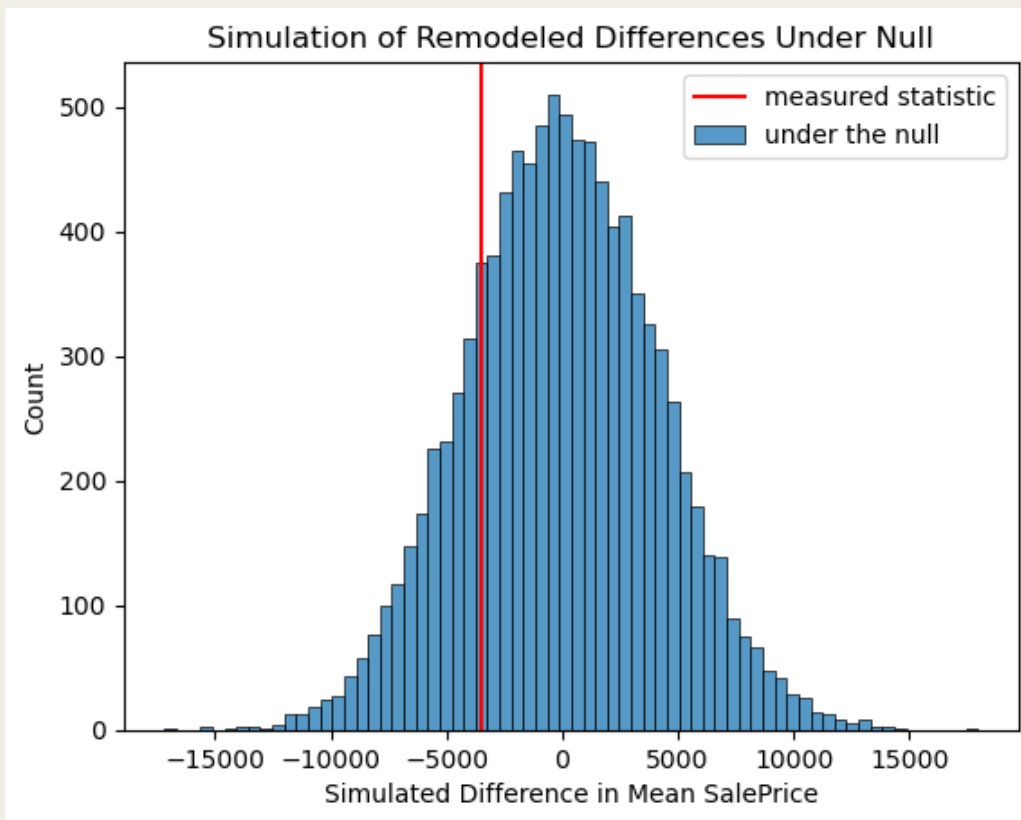


Fig. 3.1

## Conclusion

In this project, we built a multiple linear regression model that predicts SalePrice well ( $R^2 \approx 0.71$  train, 0.74 test). We found that house size and age are the strongest drivers of price. Our tests suggest CentralAir is linked to higher prices, while remodeling has no reliable effect in our simulation. We discovered multicollinearity (especially among size-related features) that helps explain FullBath's unexpected negative coefficient. This motivates a follow-up analysis. We recommend prioritizing usable living space and high-impact features such as fireplaces and garages when predicting real estate prices. We advise ensuring the property is younger while paying less attention to the remodeling status.

Link to Code

