Housing Prices:
Multiple Regression Prediction

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

UNC 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:

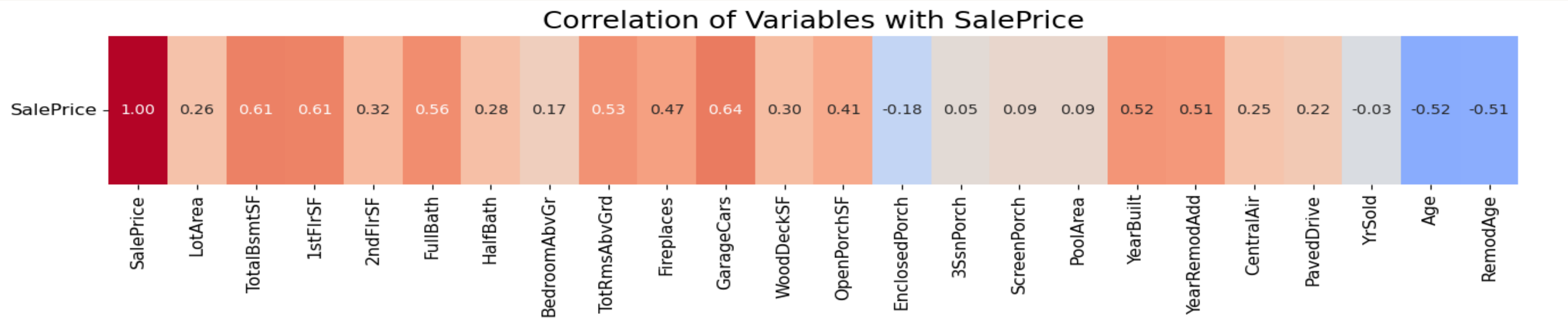
- **Optimizing Predictors:** Rigorously transforming and filtering features to manage.
- **Model Building:** Employing Standardized Linear Regression on selected features to establish a predictive relationship between property characteristics and SalePrice.
- **Validating Investment Value:** Using hypothesis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

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

Data Preparation and Exploration

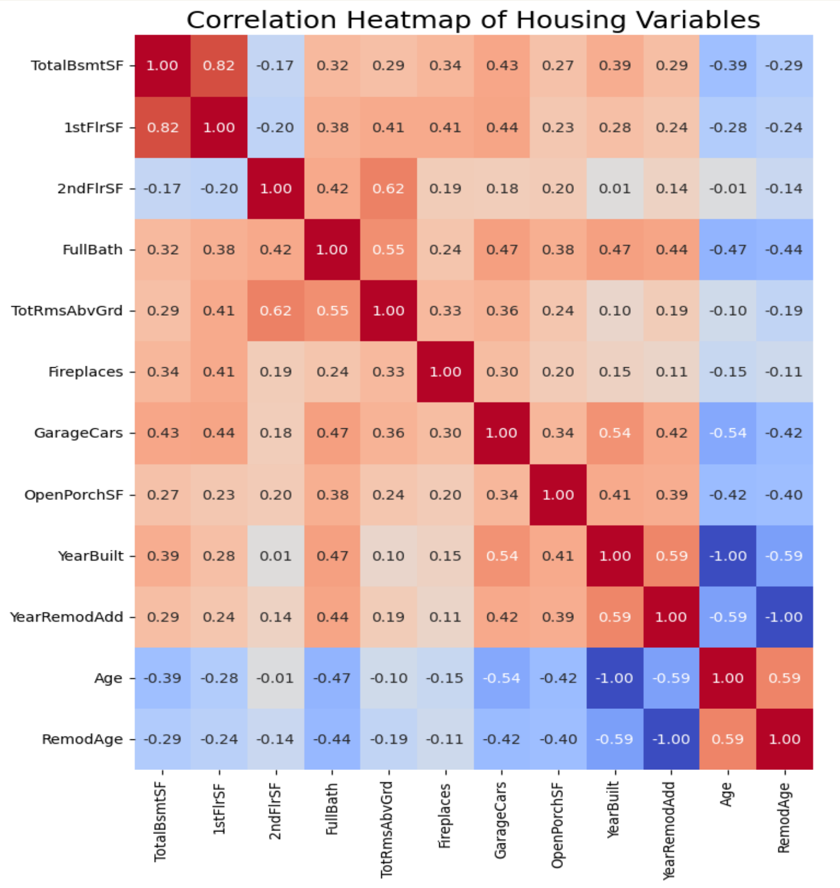
Our initial exploration focused on understanding the structure of the dataset and identifying variables strongly related to the target variable, SalePrice.

- **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. |corr| > 0.3).



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 many zero values 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 correlation matrix to the right). We strategically removed highly dependent features to mitigate multicollinearity.



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 including **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.

Baseline Model	Coefficient Values
1stFlrSF	98.612153
2ndFlrSF	65.724110
FullBath	-3375.551315
Fireplaces	12636.826823
GarageCars	20309.040344
OpenPorchSF	7649.813905
Age	-662.831892

R^2 scores: Train, 0.710; Test, 0.761

Normalized Model	Coefficient Values
1stFlrSF	37806.903032
2ndFlrSF	28789.907886
FullBath	-1468.700543
Fireplaces	8015.583393
GarageCars	15692.821506
OpenPorchSF	3342.047712
Age	-20142.956304

R^2 scores: Train, 0.713; Test, 0.738

Hypothesis Testing

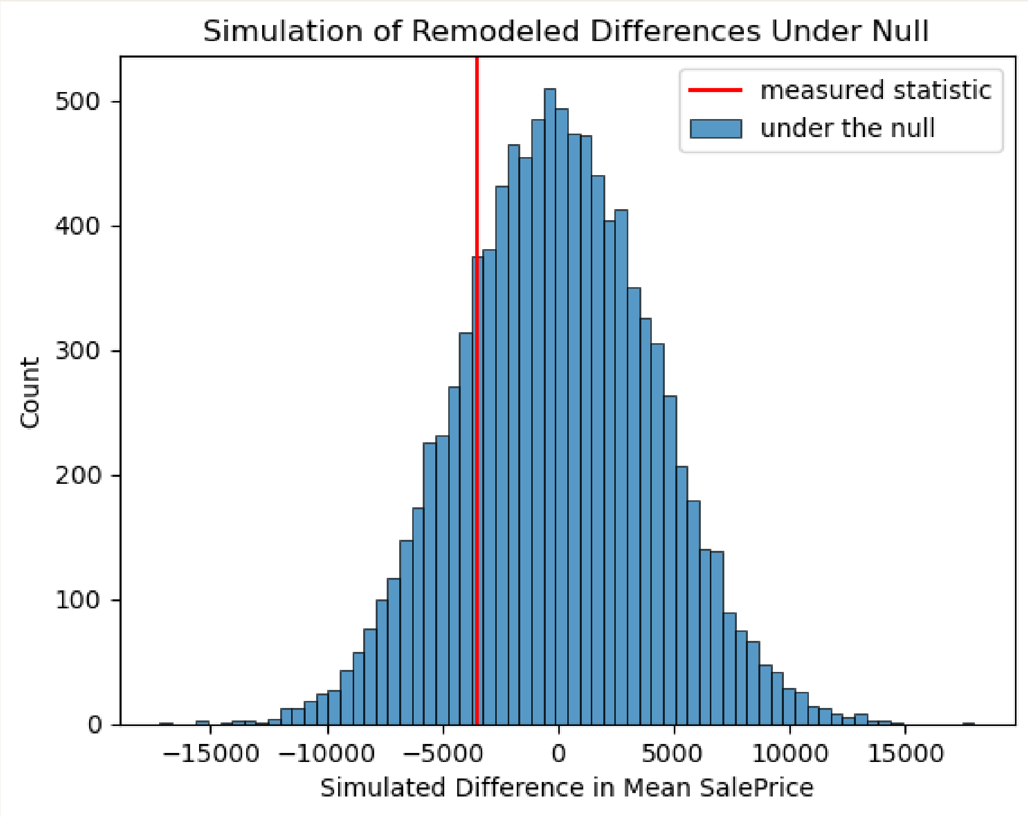
To move beyond correlation and prove investment value, we performed two rigorous statistical testing.

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), allowing 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 graph to the right).
- **Conclusion:** This test confirms that a remodel leads to a statistically significant price increase.



Conclusion

In this project, we built a multiple linear regression model that predicts SalePrice fairly well ($R^2 \approx 0.71$ train, 0.74 test). We also found that house size and age are the strongest drivers of price. Our tests suggest CentralAir is linked to higher prices, and 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. As a real estate consultant, we recommend prioritizing usable living space and high-impact features such as fireplaces and garages.

Link to Code