

Capstone Project - Housing Price Prediction

A cartoon of a small town

AI-generated content may be incorrect.

April 4, 2025

SHeng Miao

Table of Contents

[**1.** **Introduction** 2](#_Toc195106174)

[1.1 Opening Sentence 2](#_Toc195106175)

[1.2 Context for the Problem 2](#_Toc195106176)

[**2.** **Dataset** 2](#_Toc195106177)

[2.1 Data Source 2](#_Toc195106178)

[2.2 Dataset Description 2](#_Toc195106179)

[**3.** **Exploratory Data Analysis (EDA)** 3](#_Toc195106180)

[3.1 Key Decisions 3](#_Toc195106181)

[3.2 Univariate Analysis – SalePrice distribution 4](#_Toc195106182)

[3.3 Bivariate Analysis - Key Feature Relationships with SalePrice 4](#_Toc195106183)

[3.4 Correlation Matrix 10](#_Toc195106184)

[**4.** **Feature engineering** 12](#_Toc195106185)

[**5.** **Modeling** 13](#_Toc195106186)

[5.1 Modeling and optimization 13](#_Toc195106187)

[5.2 Model performance Summary 14](#_Toc195106188)

[5.3 Prediction of Testing dataset 16](#_Toc195106189)

[5.4 Feature Importance under the best model 17](#_Toc195106190)

[5.5 SHAP dependent plots of top 10 most important features 19](#_Toc195106191)

[**6.** **Conclusion** 29](#_Toc195106192)

[6.1 Living Space Is the Most Powerful Predictor 29](#_Toc195106193)

[6.2 Quality Matters More Than Condition 29](#_Toc195106194)

[6.3 Garage Capacity Shows Value Saturation 30](#_Toc195106195)

[6.4 First Floor Area Shows Nonlinear Influence 30](#_Toc195106196)

[6.5 YearBuilt – Age of Home Matters, Especially for Small Homes 31](#_Toc195106197)

[6.6 YearRemodAdd – Remodeling Improves Value, But Only Up to a Point 31](#_Toc195106198)

[6.7 Interacting Effects Are Crucial 32](#_Toc195106199)

## **Introduction**

### 1.1 Opening Sentence

"Driven by a long-standing curiosity about what makes a home a great deal—or how to sell one at the best price—I set out to explore the key factors that influence housing values. In this project, I aim to develop a robust, data-driven model for predicting house prices using advanced machine learning techniques. The goal is not only to achieve accurate predictions but also to uncover meaningful insights that can empower real estate professionals, homebuyers, and sellers to make more informed decisions in a competitive market.

### 1.2 Context for the Problem

The "Housing Prices Competition for Kaggle Learn Users" dataset from Kaggle (<https://www.kaggle.com/competitions/home-data-for-ml-course/overview>) was chosen as my Springboard Capstone project. This dataset includes 79 explanatory variables describing nearly every aspect of residential homes in Ames, Iowa.

A circular chart with different colored circles

AI-generated content may be incorrect.

This pie chart illustrates the proportion of the 79 explanatory variables grouped into major feature categories. The largest category, General Property Info, accounts for 17.7% of the features and includes attributes such as zoning, lot configuration, and neighborhood. Other key categories include Foundation & Basement, Roof & Exterior, Garage, and Interior Space, each reflecting specific structural or functional aspects of the home. Smaller categories, such as Fireplaces, Building Type & Style, and Overall Quality & Condition, contribute fewer variables but may still have significant predictive power. Grouping features this way provides a clearer overview of the dataset’s structure and the breadth of information used for modeling.

## **Dataset**

### 2.1 Data Source

Kaggle’s "Housing Prices Competition for Kaggle Learn Users."

### 2.2 Dataset Description

The dataset is divided into two parts: a training set and a testing set, each containing 1,460 records of residential home sales in Ames, Iowa. The training set includes 79 explanatory variables along with the target variable SalePrice, while the testing set contains 78 explanatory variables—identical in structure except for the absence of the SalePrice column.

## **Exploratory Data Analysis (EDA)**

### 3.1 Key Decisions

* Features with more than 50% missing values--PoolQC, Alley, MiscFeature, Fence, MasVnrType, and FireplaceQu--were removed from both the training and testing datasets.
  + For features with less than 5% missing values, the entire records containing those missing values were removed. Similarly, five garage-related features—GarageType, GarageYrBlt, GarageFinish, GarageQual, and GarageCond—had slightly higher missing rates (under 6%) but shared the same missing pattern across records; therefore, those complete records were also removed.
  + Rare categories, defined as those representing less than 1% of the training data (fewer than 13 observations), were either merged with the most similar category or grouped into a new “Other” category. To maintain consistency, rare category consolidation was based on the training set and then applied to the testing set.
  + LotFrontage, which had approximately 18% missing values, was imputed using predictions based on its linear relationship with 1stFlrSF.
  + Four categorical features—Street, Utilities, Heating, and Condition2--and four numerical features—PoolArea, 3SsnPorch, LowQualFinSF, and BsmtHalfBath--were removed due to highly imbalanced distributions (dominant category > 99% of records) and limited predictive relevance (p-Value > 0.1).
  + Among the 34 categorical features, 10 with a clear ordinal relationship were converted using ordinal encoding, while CentralAir, a binary feature, was encoded as 0/1. The remaining categorical featues were transformed using one-hot encoding. As a result, training dataset contains 1,338 records and 151 features, while the testing dataset includes 1,459 records and 150 features.

### 3.2 Univariate Analysis – SalePrice distribution

A graph of a number of blue bars

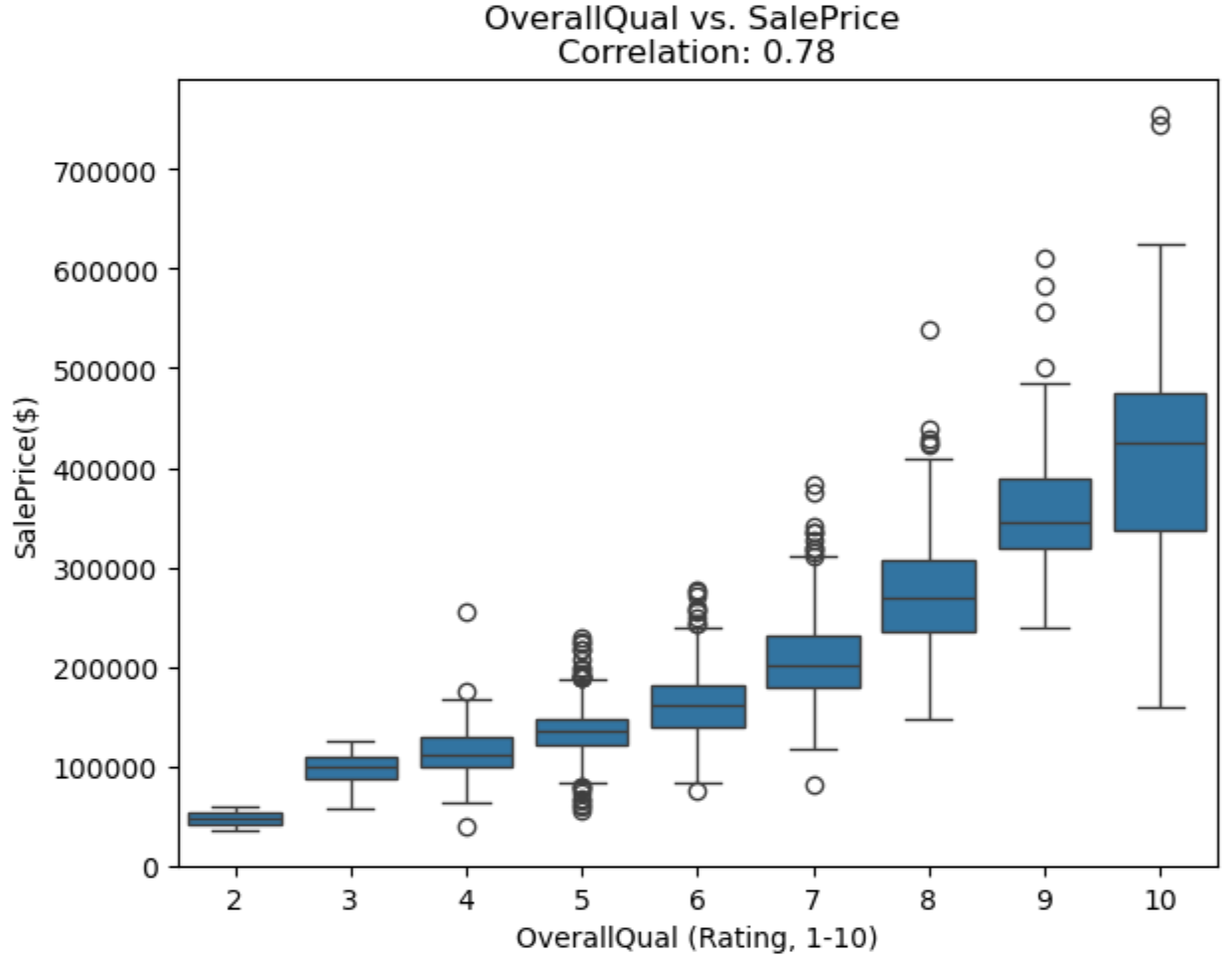
AI-generated content may be incorrect.

This histogram visualizes the frequency distribution of house sale prices. The x-axis represents SalePrice in dollars, ranging from 0 to 800,000, while the y-axis shows the frequency (number of occurrences). The distribution is right-skewed, with a peak around 100,000–200,000, indicating that most homes in the dataset sell within this price range. As the price increases beyond 300,000, the frequency drops sharply, with very few homes selling above 500,000, suggesting that higher-priced homes are less common in this dataset.

### 3.3 Bivariate Analysis - Key Feature Relationships with SalePrice

I selected the following scatter plots of the most interesting and highly correlated features (based on correlation coefficients) for deeper analysis. These charts compare features to the target feature, SalePrice, and include the most important features from the correlation analysis.

#### 3.3.1 OverallQual vs. SalePrice (Correlation: 0.78)



This box plot visualizes the relationship between OverallQual (overall material and finish quality) and SalePrice in training dataset. The x-axis represents OverallQual, a rating of the overall material and finish quality of the house, ranging from 1 (very poor) to 10 (very excellent). The y-axis represents SalePrice, the sale price of houses, in dollars. Each box represents the distribution of SalePrice for a given OverallQual rating: the central line in each box indicates the median SalePrice, the box itself spans the interquartile range (IQR, from the 25th to 75th percentile), and the whiskers extend to the minimum and maximum values within 1.5 times the IQR. Outliers beyond this range are plotted as individual points. The correlation coefficient between OverallQual and SalePrice is 0.78, as noted in the title, indicating a strong positive correlation. The plot shows that higher OverallQual ratings are associated with higher median sale prices, with a clear upward trend: houses rated below 3 have median sale prices below 100,000, while those rated above 7 have median sale prices above 200,000, with some outliers at higher ratings reaching up to 700,000.

* + *Story and Interest:* This box plot shows a strong positive correlation (0.78) between OverallQual and SalePrice. Higher quality homes consistently command higher prices, making this relationship visually compelling.
  + *Takeaway:* As expected, higher-quality homes sell for more due to their durability, aesthetics, and desirability, reflecting a key determinant of real estate value.
  + *Why We’re Seeing This:* The strong correlation aligns with expectations, as quality encapsulates a home’s appeal.

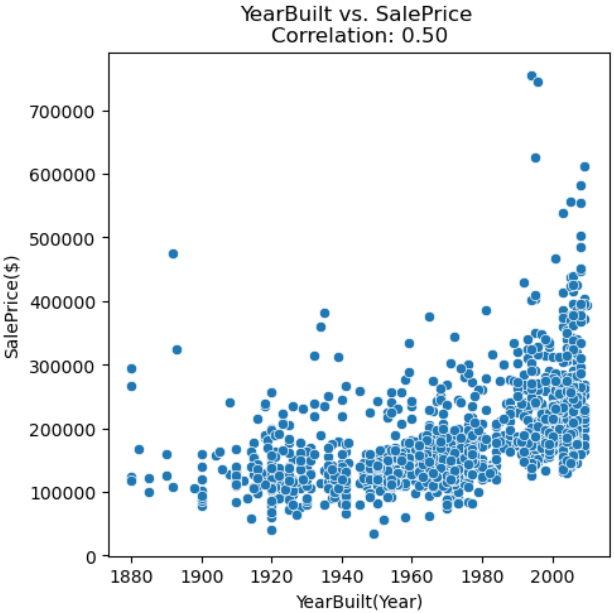
#### 3.3.2 GrLivArea vs. SalePrice (Correlation: 0.71)



This scatter plot illustrates the relationship between GrLivArea (Above grade (ground) living area square feet) and SalePrice in training dataset. The x-axis represents GrLivArea, measured in square feet, ranging from approximately 0 to 5,500. The y-axis represents SalePrice, the sale price of houses, in dollars. Each blue dot represents an individual house, with its position indicating its above ground living area and corresponding SalePrice.

* + *Story and Interest:* This plot shows a moderately strong positive correlation (0.71) between GrLivArea and SalePrice, with a general upward trend. However, scatter at higher living areas suggests other factors influence price beyond size.
  + *Takeaway:* Larger homes sell for more, as expected, due to their increased living space.
  + *Why We’re Seeing This:* Living area drives value due to functionality, but variability reflects differences in location, architectural style, or other conditions not captured here.

#### 3.3.3 YearBuilt vs. SalePrice (Correlation: 0.50)



This scatter plot visualizes the relationship between YearBuilt (the year a house was built) and SalePrice in a housing dataset. The x-axis represents YearBuilt, ranging from approximately 1870 to 2010. The y-axis represents SalePrice, the sale price of houses, in dollars. Each blue dot represents an individual house, with its position indicating the year it was built and its corresponding SalePrice.

* + *Story and Interest:* This scatter plot shows a weaker positive correlation (0.50) between YearBuilt and SalePrice, characterized by substantial variability and a nonlinear upward trend. Before 1970, there is no clear upward pattern; from 1970 to 1990, prices show a gradual increase, followed by a more accelerated rise in prices after 1990, accompanied by greater variation.
  + *Takeaway:* Surprisingly, age alone isn’t a strong driver of price, suggesting other factors overshadow YearBuilt.
  + *Why We’re Seeing This:*
  + **Pre-1970: Flat Trend, Low Variation**

*Homes built before 1970 often share similar construction standards, materials, and layouts, which limits variation in perceived value. Many of these homes may not have been significantly renovated, and thus their sale prices today reflect age-related depreciation and outdated features. The housing market was also more stable and regulated during this time, limiting price volatility.*

* + ***1970–1990: Gradual Increase***

*During this period, building standards, design preferences, and materials began to evolve — improving energy efficiency, layout functionality, and home aesthetics. The U.S. experienced moderate economic growth, which translated into slightly rising home prices for newer builds, though still constrained by relatively conservative construction trends. Many homes from this era still resemble older homes in structure but benefit from incremental improvements.*

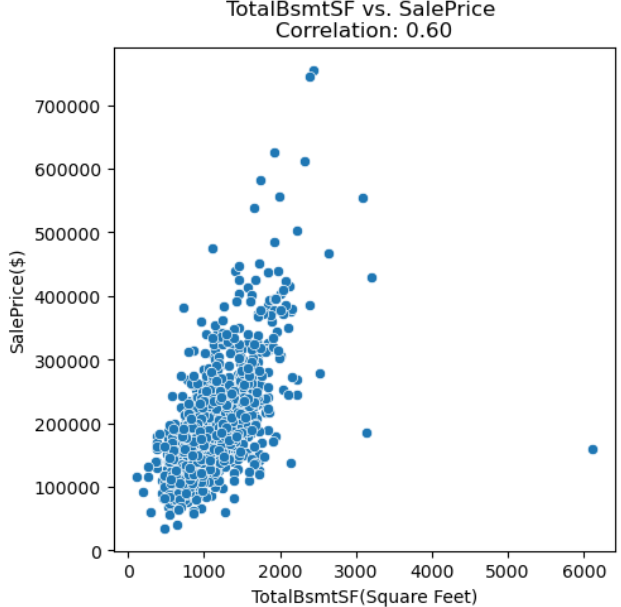
* + ***Post-1990: Accelerated Price Growth and Higher Variation***

*After 1990, the housing industry experienced rapid modernization:Open floor plans, attached garages, larger kitchens, energy-efficient materials, and higher ceilings became standard. There was a shift toward larger, higher-end construction in many regions.*

* + ***Large variation***

*The wide spread of points in the scatter plot — especially in recent construction years — reflects the influence of multiple interacting factors beyond just YearBuilt, such as remodeling and renovations, quality, etc.*

#### 3.3.4 TotalBsmtSF vs. SalePrice (Correlation: 0.60)



This scatter plot illustrates the relationship between TotalBsmtSF (Total Basement Square Footage) and SalePrice in a housing dataset. The x-axis represents TotalBsmtSF, measured in square feet. The y-axis represents SalePrice, the sale price of houses, in dollars. Each blue dot represents an individual house, with its position indicating its TotalBsmtSF and corresponding SalePrice.

* + *Story and Interest:* This plot shows a positive correlation (0.60) between TotalBsmtSF and SalePrice, with an upward trend but some variability.
  + *Takeaway:* As expected, larger basements increase sale prices due to added utility, though variability suggests other factors like finish matter.
  + *Why We’re Seeing This:* Basements provide storage or potential living space, but variability could stem from finish quality, location risks (e.g., flooding), or buyer preferences for above-ground space. This also may be due to a correlation with overall square footage of the house.

*Note:* These four features—OverallQual, GrLivArea, YearBuilt, and TotalBsmtSF—have the highest correlations with SalePrice of 0.78, 0.71, 0.50, and 0.60, respectively. They reflect key buyer priorities (quality, size, age) and offer nuanced insights into price drivers.

### 3.4 Correlation Matrix

#### 3.4.1 Correlation Matrix Heatmap of Features with High Correlations (>0.8)

A screenshot of a data visualization

AI-generated content may be incorrect.

This heatmap visualizes the pairwise correlation coefficients between filtered features, where filtering was applied to retain only features that have at least one strong correlation (absolute value greater than 0.8 but less than 1) with another feature. Specifically, the correlation matrix was scanned to identify feature pairs meeting this threshold, and only those features involved in at least one such high-correlation relationship were included in the heatmap. Both the x-axis and y-axis list the features, including SalePrice (the target feature), GrLivArea (Ground Living Area), OverallQual (Overall Quality), TotalBsmtSF (Total Basement Square Footage), 1stFlrSF (First Floor Square Footage), 2ndFlrSF (Second Floor Square Footage), GarageCars (Number of Cars the Garage Can Hold), GarageArea (Garage Area), YearBuilt (Year the House Was Built), and others related to house characteristics and zoning. The color of each cell represents the correlation coefficient between the corresponding pair of features, with the color gradient on the right indicating the scale: dark blue represents a strong positive correlation (close to 1.0), white represents no correlation (close to 0.0), and dark red represents a strong negative correlation (close to -1.0). The diagonal cells (where a feature correlates with itself) are blacked out, as they always have a correlation of 1.0. The heatmap highlights strong positive correlations, such as between GrLivArea and SalePrice, and between GarageCars and GarageArea, indicating that these features tend to increase together, while negative correlations, such as between certain exterior materials and SalePrice, suggest an inverse relationship.

*Comments on Multicollinearity:* After filtering for correlations above 0.8 (highlighted in red), I observed strong correlations between certain one-hot encoded categories, such as GarageType\_Detchd and GarageType\_Attchd. These high correlations can introduce multicollinearity in linear models. To mitigate this, it may be more effective to drop the most common category—rather than the first—during one-hot encoding, ensuring more balanced and interpretable feature representation.

#### 3.4.2 Features with the Strongest Correlation to SalePrice

A graph of blue bars with white text

AI-generated content may be incorrect.

This bar plot displays the correlation coefficients between top correlated features with the target feature SalePrice, ranked from left to right in descending order. The x-axis lists the features, including GrLivArea (Ground Living Area), OverallQual (Overall Quality), OverallCond (Overall Condition), GarageCars (Number of Cars the Garage Can Hold), GarageArea (Garage Area), TotalBsmtSF (Total Basement Square Footage), FullBath (Number of Full Bathrooms), TotRmsAbvGrd (Total Rooms Above Ground), YearBuilt (Year the House Was Built), and YearRemodAdd (Year of Remodeling or Addition). The y-axis represents the correlation coefficient of the respective feature with SalePrice.

*Comments on Target Correlations:* The features most strongly correlated with SalePrice include OverallQual (0.78) and GrLivArea (0.71), highlighting that overall quality and above-ground living area are key drivers of home value. This aligns with real estate intuition and suggests that these features are also strong predictors in the model.

## **Feature engineering**

After converting all categorical features to numerical form, I obtained a total of 150 independent numerical features, of which 111 are binary. The remaining 39 continuous or multi-level features require scaling. The appropriate scaling method for each feature was selected based on its distribution and statistical properties. The logic behind the scaling method selection is outlined below.

A screenshot of a computer

AI-generated content may be incorrect.

Excluding the year-related features (YearBuilt, YrSold, YearRemodAdd), which will be adjusted by subtracting a reference year, and the month feature (MoSold), which will be transformed using sine-cosine encoding, all remaining features will be automatically assigned to appropriate scalers. This assignment is handled by a custom function that categorizes features based on statistical thresholds and predefined ordinal classifications. The resulting feature groupings are shown below:

A screenshot of a computer

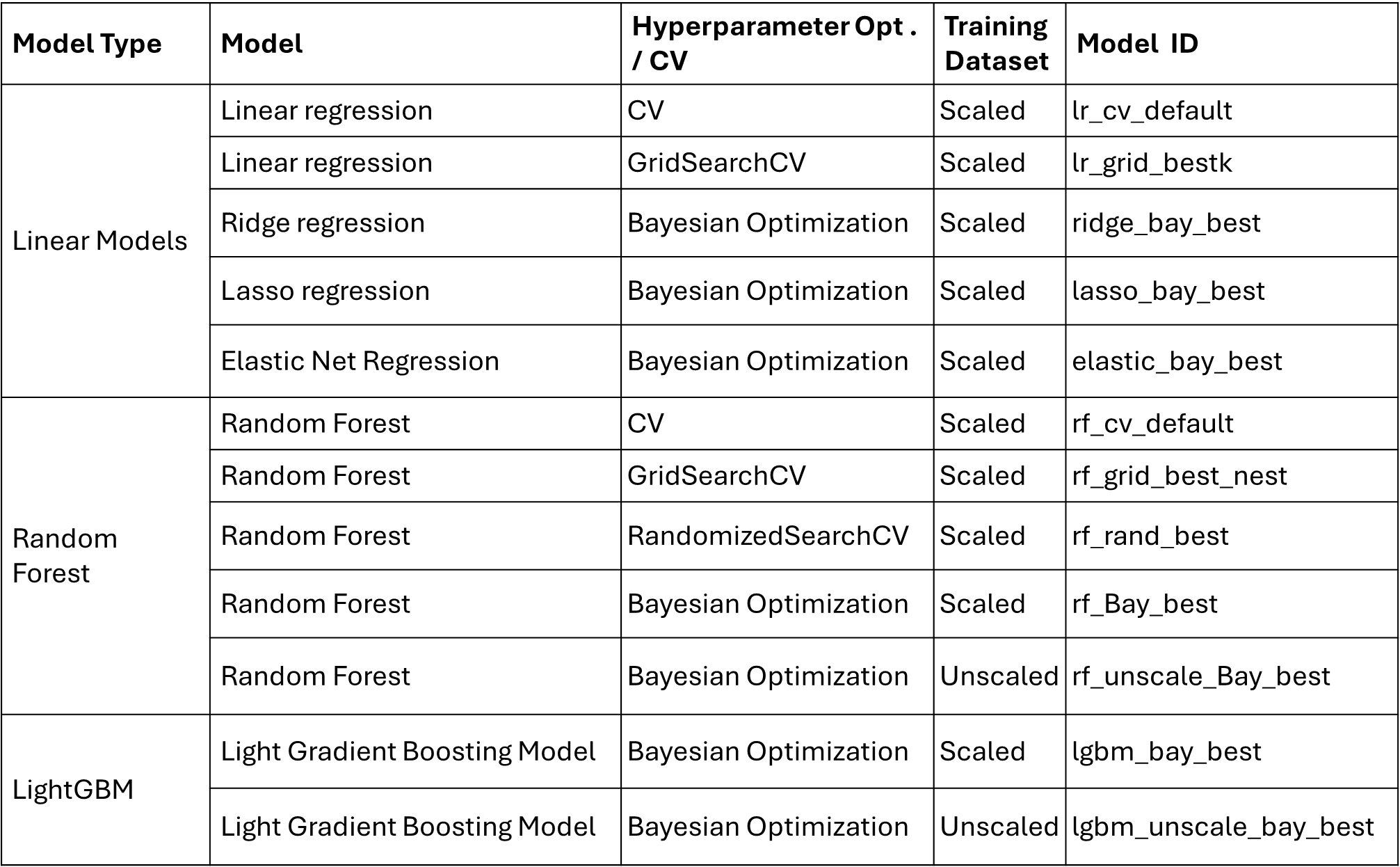
AI-generated content may be incorrect.

After inspecting the distribution of each feature, the automatic classification appeared reasonable. Subsequently, each feature was scaled using the method assigned to its respective group.

## **Modeling**

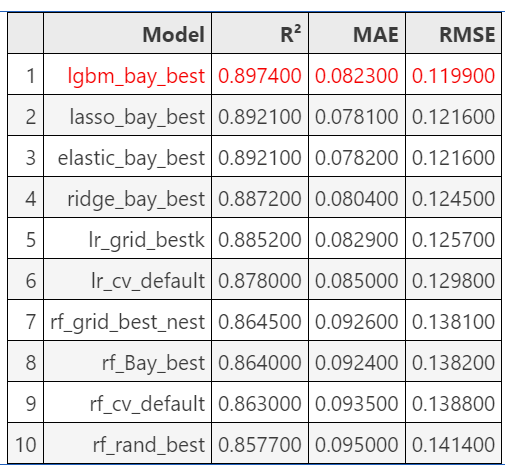
### 5.1 Modeling and optimization

To predict SalePrice, I implemented models from three major categories: Linear Models, Random Forest, and Light Gradient Boosting Machine (LightGBM). Within each category, multiple model variants were developed using a combination of different modeling techniques, hyperparameter optimization strategies (including Grid Search, Randomized Search, and Bayesian Optimization), cross-validation schemes, and training data configurations (scaled vs. unscaled). This process resulted in a total of 12 distinct model variants.



### 5.2 Model performance Summary

The performance and predictive accuracy of each model were evaluated using three key metrics: R² (coefficient of determination), MAE (Mean Absolute Error), and RMSE (Root Mean Squared Error). Based on R² scores, the top three models were lgbm\_bay\_best, lasso\_bay\_best, and elastic\_bay\_best. Among them, lgbm\_bay\_best (highlighted in red) was selected as the best-performing model due to its highest R², lowest RMSE, and fourth-lowest MAE. This model was subsequently used for the downstream analysis.



A graph of a bar graph

AI-generated content may be incorrect.A graph of a graph with text

AI-generated content may be incorrect.

A graph of a bar graph

AI-generated content may be incorrect.

These three bar plots compare the performance of 12 different model variants across three evaluation metrics: R² (coefficient of determination), MAE (mean absolute error), and RMSE (root mean squared error). Across all three metrics, the LightGBM model with Bayesian optimization (lgbm\_bay\_best) consistently outperforms the others, achieving the highest R² and the lowest MAE and RMSE, indicating superior predictive accuracy and model fit. Linear models such as lasso\_bay\_best, elastic\_bay\_best, and ridge\_bay\_best also perform competitively, particularly in terms of low error metrics. In contrast, the Random Forest models—especially the default and randomized versions—show the lowest R² and highest error values, suggesting relatively weaker performance in this regression task. Overall, the plots highlight the effectiveness of LightGBM combined with Bayesian optimization, and underscore the benefit of tuning and scaling strategies across all model categories.

### 5.3 Prediction of Testing dataset

In a Kaggle competition, the prediction on the testing dataset generated by the best-performing model—LightGBM with Bayesian optimization (lgbm\_bay\_best)—ranked 392nd out of 7,003 submissions, placing in the top 5%. To further improve this ranking, several enhancements are proposed:

* + Explore top-performing public notebooks for modeling strategies and feature engineering ideas
  + Experiment with advanced ensembling techniques (e.g., combining LightGBM, XGBoost, and CatBoost)
  + Implement cross-validation with out-of-fold predictions to better control overfitting
  + Conduct deeper feature engineering, which is often a key factor distinguishing top-tier solutions

### 5.4 Feature Importance under the best model

To interpret the predictions of the chosen best model-- LightGBM with Bayesian optimization (lgbm\_bay\_best) and understand the impact of each input feature, SHAP (SHapley Additive exPlanations) values offer a powerful and theoretically grounded approach. Based on cooperative game theory, SHAP assigns each feature a contribution value for individual predictions, enabling a more nuanced and consistent measure of feature importance. Unlike traditional feature importance methods that may rely on model-specific heuristics, SHAP values provide a unified framework that is both model-agnostic and locally accurate. In this analysis, SHAP was used to evaluate and visualize the contribution of each feature to the predicted target feature SalePrice, offering clearer insights into the model’s decision-making process.

A screen shot of a graph

AI-generated content may be incorrect.

This SHAP summary plot illustrates the relative importance and impact of features on the model's house price predictions. Each dot represents a SHAP value for an individual data point, showing how much a feature contributed to increasing or decreasing the predicted value. Features are ranked by their overall importance (mean absolute SHAP value), with GrLivArea, OverallQual, and OverallCond being the most influential. The color gradient indicates the feature's actual value—red for high and blue for low. For example, larger GrLivArea values (red) tend to increase predictions (positive SHAP values), while smaller values (blue) lower them. This plot reveals that key drivers of house price predictions are related to living space and overall quality, and also highlights non-linear and feature-specific effects that would be missed by traditional linear models.

A bar graph with red and white text

AI-generated content may be incorrect.

This SHAP feature impact bar plot shows the average contribution of each feature to the model’s predictions, measured by the mean absolute SHAP value. It highlights the top 9 most impactful features individually, while the remaining 141 features are grouped into a single aggregated bar. GrLivArea (above-ground living area) and OverallQual (overall material and finish quality) stand out as the most influential features, each contributing an average of +0.08 to the model's output. Other features like OverallCond, TotalBsmtSF, and ExterQual follow, each contributing around +0.02. The aggregated impact of the remaining features, although individually less influential, adds up to a significant portion of the model’s explanation (+0.25), suggesting a broad set of minor contributors. This plot confirms that while a few features dominate model predictions, many smaller ones still play a meaningful cumulative role.

### 5.5 SHAP dependent plots of top 10 most important features

#### 5.5.1 GrLivArea (Above grade (ground) living area square feet)

A graph of a graph showing a line of purple and blue dots

AI-generated content may be incorrect.

This SHAP dependence plot shows how the feature GrLivArea (above-ground living area in square feet) influences the model’s prediction of house prices, with OverallQual (overall quality) encoded by color. Each point represents a house, with the x-axis showing the actual GrLivArea and the y-axis indicating the SHAP value—i.e., the feature’s contribution to the predicted price. There is a strong positive relationship: as GrLivArea increases, its SHAP value increases, indicating a greater positive impact on predicted prices. The color gradient reveals an interaction effect—houses with both large living areas and high OverallQual (shown in red) contribute more strongly to higher predicted values. **The dashed vertical line at GrLivArea = 1425 marks a threshold where the impact shifts from negative to positive.** This plot demonstrates that larger, high-quality homes drive up predictions and emphasizes the compound influence of size and quality in determining house prices.

#### 5.5.2 OverallQual (Rates the overall material and finish of the house)

A graph of different colored dots

AI-generated content may be incorrect.

This SHAP dependence plot illustrates how OverallQual (a categorical measure of overall material and finish quality, ranging from 1 to 10) affects the model's predicted house prices, with GrLivArea (above-ground living area) shown as a color gradient. The x-axis shows the OverallQual values, and the y-axis represents the corresponding SHAP values—indicating how much each quality level contributes to the prediction. The plot shows a clear positive, stepwise relationship: as OverallQual increases, so does its SHAP value, meaning higher-quality homes contribute more positively to predicted prices. The color shading adds an interaction layer, showing that **higher GrLivArea (red points) tends to enhance the positive effect of high OverallQual.** This suggests a synergistic relationship between size and quality—large, high-quality homes exert the greatest upward influence on price predictions.

#### 5.5.3 OverallCond (Rates the overall condition of the house)

A graph of a graph of a number of objects

AI-generated content may be incorrect.

This SHAP dependence plot shows how OverallCond (overall condition of the house) impacts the model’s predicted house prices, with OverallQual represented by color. The x-axis displays OverallCond values (typically ranging from 1 to 9), while the y-axis indicates the SHAP value—how much this feature contributes to each individual prediction. The plot reveals a more subtle and nonlinear relationship. Homes with very low condition scores (2–3) have notably negative SHAP values, strongly reducing predicted prices, albeit with considerable variation. Most homes cluster around a condition score of 5–6, where SHAP values are closer to zero or slightly positive, suggesting a nearly neutral effect on predictions. However, beyond a condition rating of 6, the SHAP values flatten, indicating that **further improvements in condition beyond 6 have only a minimal positive effect on model predictions**. The color gradient shows that homes with high OverallQual (red) appear across all condition levels, but the effect of condition is more pronounced when quality is not high. This suggests that while poor condition significantly lowers predicted prices, excellent condition doesn't proportionally increase them beyond a certain threshold.

#### 5.5.4 TotalBsmtSF (Total square feet of basement area)

A graph of a graph showing a number of numbers and a line

AI-generated content may be incorrect.

This SHAP dependence plot illustrates how TotalBsmtSF (total basement square footage) influences the model’s predicted house prices, with GrLivArea (above-ground living area) shown by the color gradient. There is a clear positive relationship: as basement size increases, the SHAP value also increases, meaning larger basements contribute more positively to predicted price. The vertical dashed line at 988 sq ft marks the threshold where basement size begins to have a consistently positive impact. Notably, when the basement area falls below approximately 700 sq ft, it has the strongest negative impact, accompanied by high variability, indicating that small basements are heavily penalized in the model. This negative effect is particularly strong in homes with smaller GrLivArea (indicated by blue points), suggesting **that limited total living space—both above and below ground—substantially lowers the model's predicted value**. Conversely, the highest SHAP values are concentrated in homes with both large basements and large above-ground areas (in red), highlighting a compounding effect of total usable space. This pattern shows that the model prioritizes overall living area—both above and below ground—as a key driver of house value.

#### 5.5.5 ExterQual (Evaluates the quality of the material on the exterior)

A graph of different colored dots

AI-generated content may be incorrect.

This SHAP dependence plot illustrates how ExterQual (exterior quality rating) affects the model’s predicted house prices, with OverallQual represented by color. The x-axis shows the exterior quality levels (typically rated from 1 to 5), while the y-axis displays the SHAP values, indicating the contribution of each ExterQual level to the prediction. The plot reveals a clear stepwise pattern: houses with low ExterQual ratings (around 3) tend to have negative SHAP values, reducing the predicted price, whereas higher ratings (4 and 5) have positive SHAP values and increase predicted prices. **This positive impact is more pronounced in higher-quality homes (seen in red).** However, **upgrading exterior material quality from level 4 to 5 does not yield a substantial additional increase in predicted price.** Additionally, the color gradient shows that higher exterior quality often coincides with higher overall quality (red), further reinforcing their combined positive influence on the model’s output.

#### 5.5.6 YearBuilt (Original construction date)

A graph showing the value of a company

AI-generated content may be incorrect.

This SHAP dependence plot illustrates how YearBuilt (the actual construction year of a house) influences predicted house prices, with GrLivArea (above-ground living area) represented by color. The x-axis shows the build year, while the y-axis displays the SHAP value, indicating how much the construction year contributes to the model’s output. The plot is segmented into four time periods for clearer interpretation. Before 1900, homes over a century old show the strongest negative SHAP values with large variation, despite often having larger living areas—suggesting that extreme age significantly lowers predicted value, possibly due to outdated design or maintenance concerns. From 1900 to 1958, there's a gradual increase in SHAP values, with considerable spread; notably, homes with smaller living areas face more negative impact. Between 1958 and 1985, the SHAP values stabilize at a slightly negative level, again with more pronounced penalties for smaller homes**. After 1985, the trend becomes sharply, and YearBuilt has a strong positive influence on predicted prices—particularly for smaller homes, which benefit most from being recently built.** This pattern highlights how the model recognizes not only the chronological value of newer construction but also how it offsets size limitations in modern homes.

#### 5.5.7 YearRemodAdd (Remodel date)

A graph of a graph showing the number of years

AI-generated content may be incorrect.

This SHAP dependence plot illustrates how YearRemodAdd (year a house was remodeled) impacts the model’s predicted house prices, with YearBuilt shown by color. The x-axis represents the remodeling year, and the y-axis shows the SHAP value, reflecting how much that year contributes to the prediction. The data is divided into three distinct phases by vertical lines at 1970 and 1992. Before 1970, YearRemodAdd has a consistently negative impact with large variation, and the SHAP values show a very slow increasing trend, indicating that during this period, home prices were not very sensitive to remodeling year. Between 1970 and 1992, there is a sharp and steady increase in SHAP values, showing that **more recent remodels in this window significantly boosted predicted home values**—likely due to major improvements in design standards or materials. However, after 1992, the SHAP values plateau, indicating that further remodeling beyond this point does not yield significantly higher predictions. **This positive impact is more pronounced in newly-built homes (seen in red).** **This pattern implies diminishing returns on remodel timing after the early 1990s,** as modern updates beyond that point may be considered comparable in quality and thus no longer differentiating in the model’s view.

#### 5.5.8 BsmtFinSF1(Type 1 finished square feet)

A graph of a graph showing a number of points

AI-generated content may be incorrect.

This SHAP dependence plot illustrates how BsmtFinSF1 (finished square footage of the basement’s primary area) impacts the model’s prediction of house prices, with TotalBsmtSF (total basement size) encoded by color. The x-axis shows the actual BsmtFinSF1 values, while the y-axis represents the SHAP value, indicating the contribution of this feature to each prediction. There is a clear positive relationship up to around 2000 sq ft, where increasing finished basement space contributes increasingly to higher predicted prices. A vertical dashed line marks a threshold at approximately 226 square feet, suggesting that **below this value, finished basement area contributes little or even negatively to price predictions. Beyond this point, the SHAP values rise steeply, especially for homes with larger total basements (red points),** indicating that finishing more basement space significantly boosts model predictions. The points beyond 2000 sq ft appear to level off, this is due to only two outliers, and not a consistent pattern.

#### 5.5.9 1stFlrSF (First Floor square feet)

A graph with numbers and lines

AI-generated content may be incorrect.

This SHAP dependence plot illustrates how 1stFlrSF (first floor square footage) influences the predicted house price, with TotalBsmtSF (total basement area) represented by the color gradient. The two vertical dashed lines at approximately 670 and 1078 square feet divide the data into three distinct regions. In the first zone (below 670 sq ft), 1stFlrSF has the strongest negative impact, as shown by the sharply negative SHAP values and a wide spread, indicating high sensitivity and variation in how extremely small floor areas reduce predicted price. In the middle zone (670–1078 sq ft), SHAP values are relatively stable and hover around a slightly negative level, suggesting that moderate first floor sizes have limited impact on the model’s predictions. In the third zone (above 1078 sq ft), SHAP values begin to rise rapidly, showing that larger first floor sizes strongly increase predicted house prices. **This nonlinear pattern reveals that only when first floor area exceeds a certain threshold does it become a significant value driver in the model.**

#### 5.5.10 GarageCars (Size of garage in car capacity)

A graph of different colored dots

AI-generated content may be incorrect.

This SHAP dependence plot illustrates how the number of garage spaces (GarageCars) influences the model’s predicted house prices, with OverallQual (overall quality) represented by the color gradient. The x-axis shows the number of garage spaces (ranging from 1 to 4), and the y-axis displays the SHAP values, which indicate the contribution of each value to the prediction. The plot reveals a stepwise pattern: homes with 1 garage space generally have negative SHAP values, reducing the predicted price, while 2-car garages cluster around zero, suggesting a neutral impact. A sharp jump occurs at 3 garage spaces, where SHAP values become strongly positive, indicating a significant boost to the predicted price. **This positive impact is more pronounced in higher-quality homes (seen in red).** However, this group also exhibits a large spread, suggesting that the added value of a 3-car garage varies significantly depending on other features such as quality. Although data for 4-car garages is sparse, **increasing from 3 to 4 garage spaces does not provide additional benefit,** as the SHAP values remain similar or even slightly lower. This implies the model captures a saturation point where additional garage capacity no longer meaningfully increases home value. **Overall, the model treats 3-car garages as a key value-add feature, while fewer garage spaces may lower a home's predicted price.**

## **Conclusion**

This analysis leverages SHAP (SHapley Additive exPlanations) values to interpret the influence of individual features on the chosen best model-- LightGBM with Bayesian optimization (lgbm\_bay\_best). SHAP enables us to quantify and visualize how specific attributes of a property—such as size, quality, construction year, and remodeling—contribute to its predicted value.

The findings not only confirm several well-established real estate intuitions but also uncover subtle patterns and threshold behaviors that might otherwise remain hidden in traditional linear models.

### 6.1 Living Space Is the Most Powerful Predictor

* Key Features: GrLivArea, TotalBsmtSF, and BsmtFinSF1
* Insight: Larger above-ground living area and basement size consistently increase predicted house prices.
* Threshold Behavior:
  + Basements below ~700 sq ft show the strongest negative SHAP values, particularly in smaller homes.
  + A clear inflection occurs at ~988 sq ft (TotalBsmtSF) and ~226 sq ft (BsmtFinSF1), after which SHAP values begin to shift from negative to positive.
* Why It Matters: This aligns with the market's emphasis on total usable living space. The model captures a compounding effect—homes with both large basements and above-ground areas yield the greatest value.

### 6.2 Quality Matters More Than Condition

* Key Features: OverallQual, OverallCond
* Insight:
  + OverallQual shows a strong, stepwise positive relationship with predicted price.
  + OverallCond, however, has limited predictive power beyond average scores (~5), except for low scores (2–3), which are penalized heavily.
* Interpretation:
  + Expected for OverallQual—buyers and models alike value superior craftsmanship and finish.
  + Unexpected for OverallCond—suggests that condition ratings may be redundant or less discriminative in the presence of other features like remodeling and quality.

### 6.3 Garage Capacity Shows Value Saturation

* Key Feature: GarageCars
* Insight:
  + 1-car garages decrease predicted price.
  + 2-car garages are neutral.
  + 3-car garages provide the largest boost, though with significant variation. The positive impact is more pronounced in higher OverallQual.

*The positive impact is more pronounced in homes with higher OverallQual — indicating that garage size adds more value when paired with high-quality construction.*

* + Increasing to 4 cars shows no additional value.
* Why This Is Interesting:
  + Suggests an optimal garage size from a value perspective—bigger isn't always better.
  + The high spread in 3-car garage SHAP values highlights interaction with other features like home quality (OverallQual) and size, meaning garage impact is context-dependent.

### 6.4 First Floor Area Shows Nonlinear Influence

* Key Feature: 1stFlrSF
* Three Behavioral Zones:
  + <670 sq ft: Strongly negative SHAP values with wide variation.
  + 670–1078 sq ft: Stable, mildly negative effect.
  + >1078 sq ft: Shift from negative to positive, sharp increase in positive SHAP values.
* Significance:
  + Identifies a tipping point where first floor size shifts from being a liability to a major asset.
  + Provides actionable insight for homeowners and developers aiming to optimize layout space.

### 

### 6.5 YearBuilt – Age of Home Matters, Especially for Small Homes

* Insight:
  + Pre-1900: Strongest negative SHAP values, despite many having large living areas.
  + 1900–1958: Gradual improvement; small homes penalized more.
  + 1958–1985: Stable, mild negative effect.
  + Post-1985: Sharp positive increase, particularly for smaller homes.
* Unexpected but Insightful:
  + Smaller homes benefit more from newer construction.
  + Suggests the model recognizes that modern layout, materials, and efficiency in newer small homes can outweigh the size advantage of larger homes.

### 6.6 YearRemodAdd – Remodeling Improves Value, But Only Up to a Point

* Phases Identified:
  + Before 1970: Negative SHAP values with high variability — remodeling had little impact.
  + 1970–1992: Strong upward trend — remodel year increasingly boosts predicted price.
  + Post-1992: Plateau — no additional value from newer remodels. The positive remodeling impact is more pronounced in newly-built homes, suggesting a reinforcing effect between remodeling and new construction.
* Why It’s Surprising:
  + One might expect more recent remodels to continuously raise value.
  + Possible Explanation: Renovations after the early 1990s have become more standardized, offering diminishing marginal improvements that the model no longer distinguishes.
  + Additionally, in new homes, remodeling may enhance value more effectively due to alignment with modern design standards or upgraded finishing practices.

### 6.7 Interacting Effects Are Crucial

* Smaller homes benefit disproportionately from recent construction.
* High GrLivArea amplifies the effect of quality, while high OverallQual amplifies the effect of ExterQual and garage size.
* The SHAP interaction patterns reveal that contextual value (e.g., new + small) often matters more than individual features alone.