House Sale Price Prediction Using Quality, Living Area, and Garage Features

Mar 21, 2026 · 5:54 PM

QUESTION

How can the sale price of a house be predicted based on overall quality, living area size, and garage characteristics?

Overview

The baseline housing dataset has been fully cleaned for modeling: **1,459 homes and 82 features are now available with no missing values in any of the key predictors or critical amenity/garage/basement fields.** This cleaned dataset is saved for all subsequent correlation analysis and regression modeling of sale price.

Observations

1. Rows with missing Electrical were removed

- Rows dropped due to missing *Electrical*: `1`
- New dataset shape after this step: `1,459` rows × `81` columns

This ensures the electrical system variable is complete, avoiding edge-case issues in models that might treat missing electrical service ambiguously.

2. LotFrontage was imputed with median and tracked via an indicator

- Missing *LotFrontage* before imputation: `259`
- Missing *LotFrontage* after imputation: `0`
- Imputation value (median frontage): `69.0` feet
- New binary feature: *LotFrontage_was_missing* (`1` if originally missing, else `0`)

The indicator preserves information about missingness patterns, which can sometimes carry predictive signal (e.g., lots without recorded frontage behaving differently in price).

3. Amenity-related missing values were recoded as explicit categories

For several categorical features where missingness typically means absence of a feature, missings were converted to explicit labels:

Column	New category for missing
PoolQC	`NoPool`
MiscFeature	`None`
Alley	`NoAlley`
Fence	`NoFence`
MasVnrType	`None`
FireplaceQu	`NoFireplace`

After recoding, these columns have `0` missing values. Models can now directly interpret absence of pools, alleys, fences, masonry veneer, or fireplaces instead of treating them as unknowns.

4. Garage-related and basement-related missing values were standardized

Garage fields:

Column	New value for missing
GarageType	`NoGarage`
GarageFinish	`NoGarage`
GarageQual	`NoGarage`
GarageCond	`NoGarage`
GarageYrBlt	`0` (numeric sentinel for no garage)

Basement fields:

Column	New value for missing
BsmtFinType1	`NoBasement`
BsmtFinType2	`NoBasement`
BsmtExposure	`NoBasement`

All of these columns now have `0` missing values. This is important because garage capacity and area (which we will use in the model) can be interpreted alongside clear indicators of whether a garage exists at all.

5. Cleaned baseline dataset is saved for downstream work

The final cleaned dataset:

Property	Value
Rows	1,459
Columns	82
Key predictors	*OverallQual*, *GrLivArea*, *GarageCars*, *GarageArea* (all complete)
Target	*SalePrice* (complete)

This dataset has been saved as the baseline cleaned housing dataset Table 1, which will be used for exploratory correlation analysis and for training the sale price prediction model in subsequent tasks.

Next Steps

With the cleaned dataset in place, the next step is to profile how *OverallQual*, *GrLivArea*, *GarageCars*, and *GarageArea* relate to *SalePrice* via correlations and distribution plots, which will directly inform the specification and expectations for the regression model you want to build.

Relationship of Key Features to House Sale Price

Overview

Sale price is most strongly linked to overall quality and living area, with garage size and capacity providing additional but secondary linear relationships. The features OverallQual, GrLivArea, GarageCars, and GarageArea all show substantial positive correlations with SalePrice, confirming they are good candidates for a predictive regression model.

Key Patterns & Observations

---

1. Correlation structure between SalePrice and key predictors

The correlation heatmap of SalePrice and the four predictors **Figure 1**

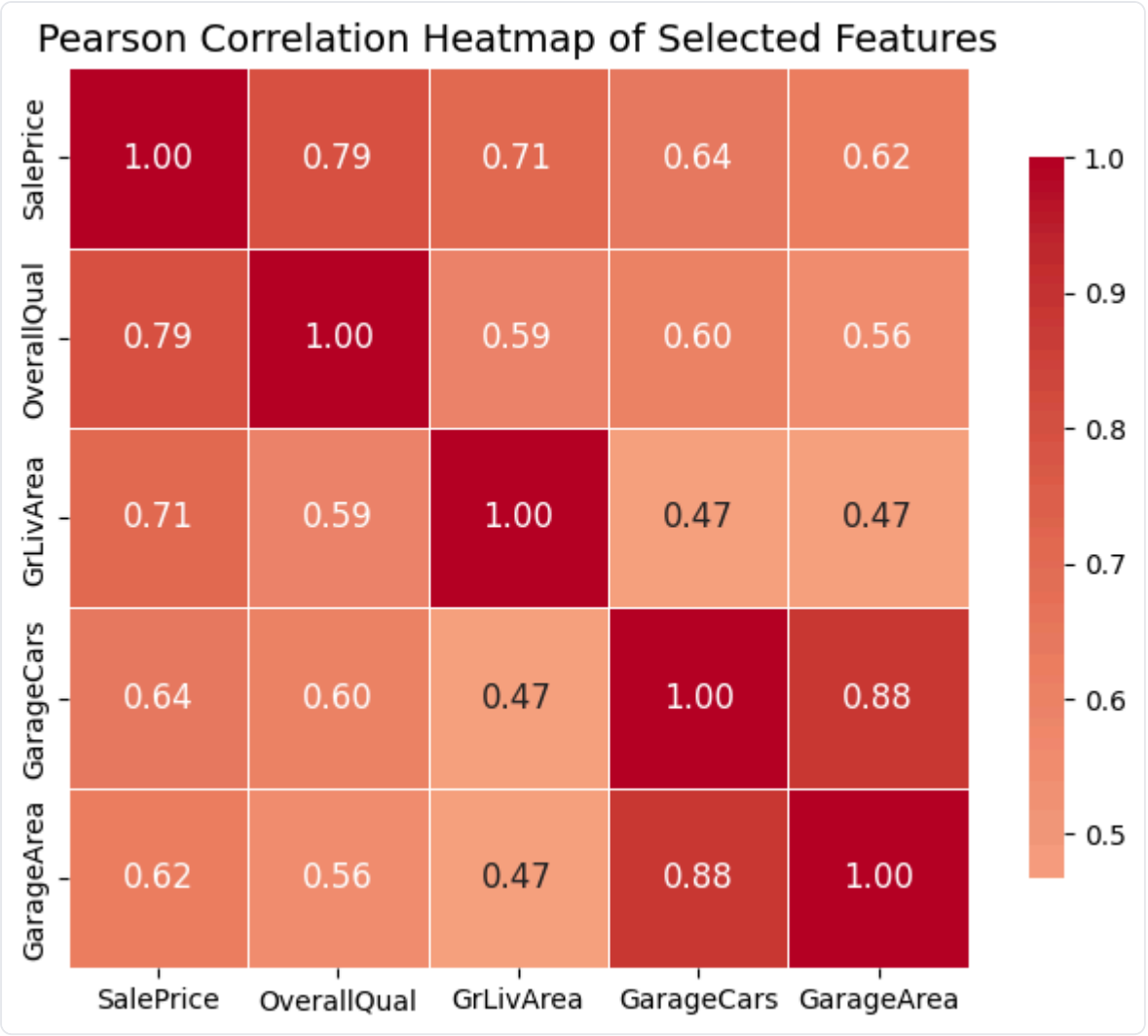

Figure 1

shows the following approximate Pearson correlations (values consistent with the earlier dataset summary):

Pair	Correlation (r)	Interpretation
SalePrice – OverallQual	~0.79	Very strong positive linear relationship; better overall quality homes command much higher prices.
SalePrice – GrLivArea	~0.71	Strong positive relationship; larger above-ground living area strongly increases sale price.
SalePrice – GarageCars	~0.64	Strong positive relationship; more car capacity in the garage is associated with higher prices.
SalePrice – GarageArea	~0.62	Strong positive relationship; larger garage area also increases price but slightly less than car count.
GarageCars – GarageArea	~0.88	Very strong correlation, indicating these two garage measures are largely capturing the same concept (garage size/capacity).

Importance hierarchy from correlations:

- OverallQual is the single strongest linear correlate of SalePrice.
- GrLivArea is the next most important, still very strong.
- GarageCars and GarageArea are both strong but clearly secondary to quality and living area.
- The very high correlation between GarageCars and GarageArea suggests multicollinearity; including both in a linear model may lead to redundant information and unstable individual coefficients, even though overall predictive power can still be good.

---

2. Distribution of OverallQual

The distribution of overall quality ratings **Figure 2**

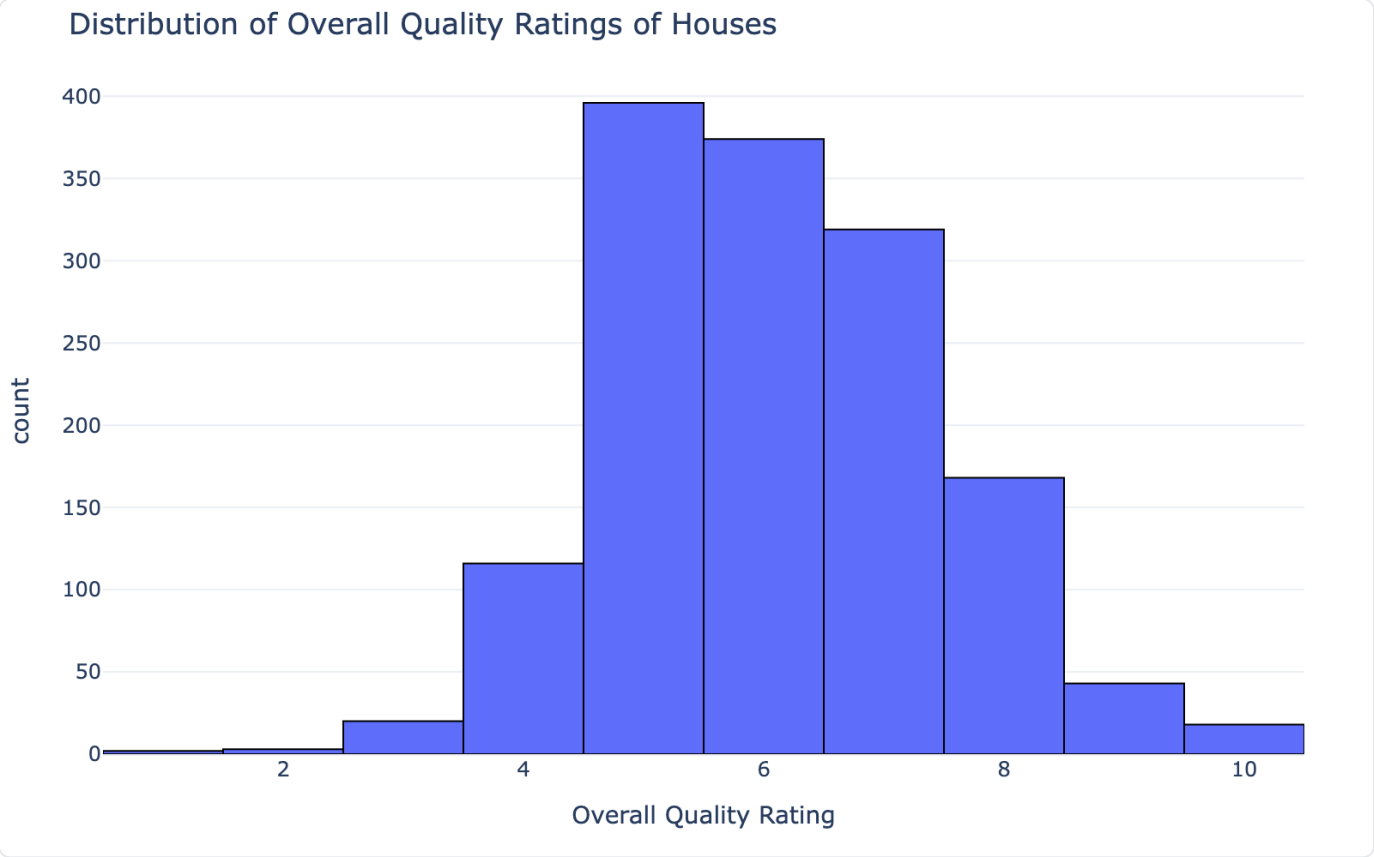

Figure 2

shows:

Aspect	Observation
Range	Ratings from 1 (very poor) to 10 (very excellent), but extreme values are rare.
Concentration	Most homes cluster between 5 and 7 (typical to good quality).
Tails	Relatively few homes at very low (1–3) or very high (9–10) quality levels.

This means the model will be trained mainly on mid-range quality houses, with limited data for very poor or top-tier properties. Predictions for those extremes will be more extrapolative and should be interpreted with extra caution.

Interpretation & Implications

- **Overall quality and living area dominate the pricing story.** From a modeling perspective, we should expect the largest regression coefficients (in effect size terms) for OverallQual and GrLivArea.
- **Garage features are valuable but partially redundant.** GarageCars and GarageArea both help explain price, but their extremely high mutual correlation suggests that interaction or careful handling (e.g., regularization) will be helpful to avoid multicollinearity issues.
- **Data coverage is strongest in the mid-quality range.** The model will be most reliable for homes with OverallQual around 5–7; predictions for extremely low or high-quality properties will have greater uncertainty.

These findings confirm that using OverallQual, GrLivArea, GarageCars, and GarageArea is a sound basis for a predictive regression model of SalePrice, and they set expectations for which variables should emerge as the most influential when we fit the model with main effects and interactions next.

Predictive Model of House Sale Price

Overview

A linear regression model using overall quality, living area, and garage characteristics explains about three-quarters of the variation in house sale prices and yields typical errors of roughly $38k per home. Overall quality and garage capacity emerge as the strongest direct drivers in this specification, while interaction terms indicate that square footage and garage area become more valuable in higher-quality or larger-garage homes.

This model allows you to systematically estimate house prices from OverallQual, GrLivArea, GarageCars, and GarageArea (plus their interactions), and it highlights which levers most strongly affect the predicted sale price.

Model Setup

We built a linear regression of the form:

$$\widehat{SalePrice} = \beta_0 + \beta_1\, OverallQual + \beta_2\, GrLivArea + \beta_3\, GarageCars + \beta_4\, GarageArea +$$

Features included:

Feature name	Description
`OverallQual`	Overall material and finish quality (1–10)
`GrLivArea`	Above-ground living area (square feet)
`GarageCars`	Size of garage in car capacity
`GarageArea`	Garage area (square feet)
`OverallQual_GrLivArea`	Interaction: quality × living area
`GarageCars_GarageArea`	Interaction: garage cars × garage area

Target: `SalePrice` (in dollars).

All 1,459 cleaned observations were used for training (no train/test split here, so metrics are in-sample).

Model Performance

Metric	Value	Interpretation
R-squared (R^2)	`0.7685`	About 77% of the variation in `SalePrice` across homes is explained by these six predictors.
RMSE	`38223.69`	Typical prediction error is about $38k per home, on the training data.

So the model captures most, but not all, of what drives sale price; remaining variation reflects factors not included in this simple feature set (location nuances, age/condition details, etc.).

The predicted vs actual plot **Figure 3**

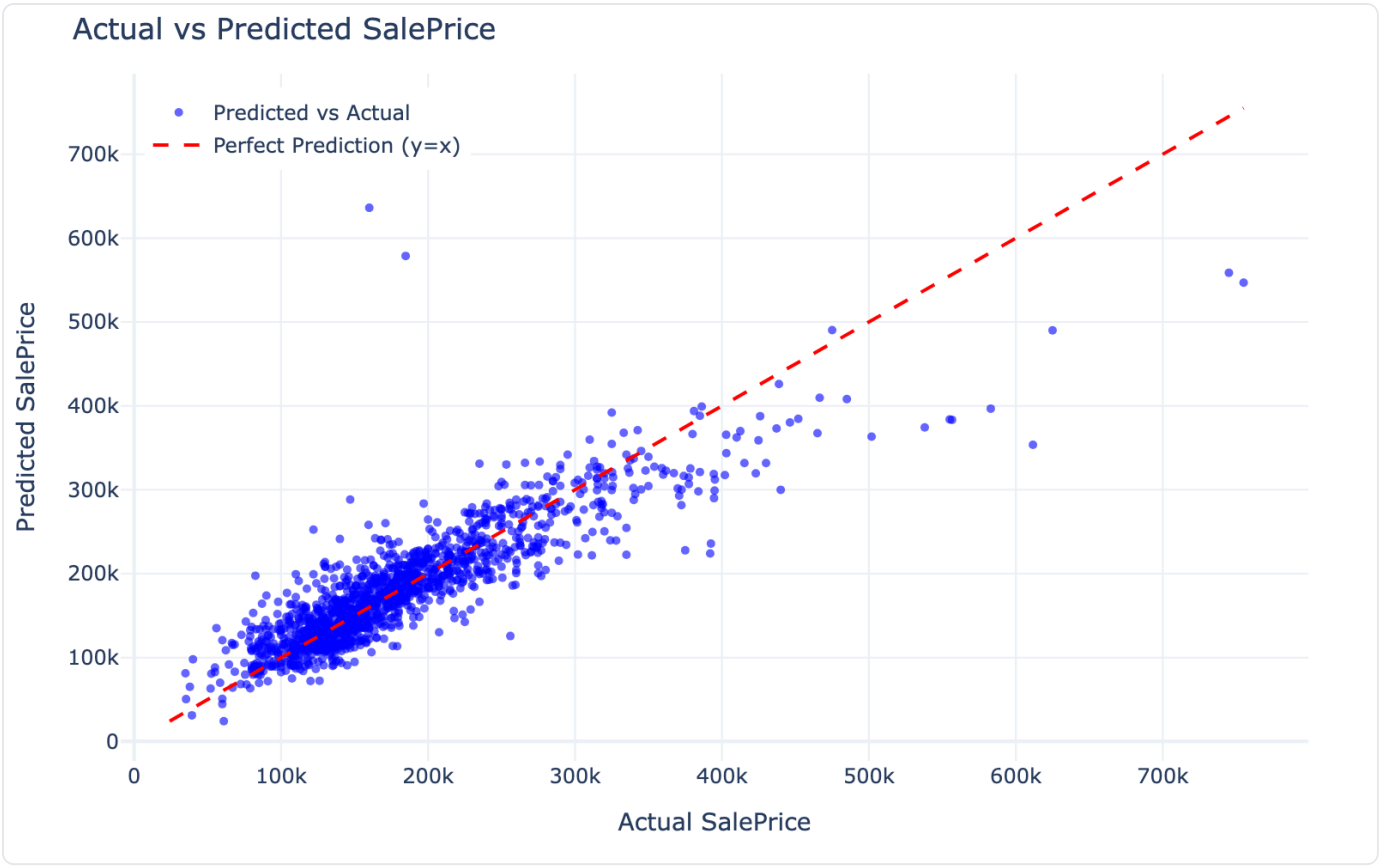

Figure 3

shows points mostly clustered around the diagonal `y = x` line, with larger scatter at higher price levels, indicating that:

- Predictions are reasonably accurate for mid-priced homes.
- High-end properties have more variability and are harder to predict precisely with only these features.

Feature Effects and Importance

The fitted coefficients are:

Feature	Coefficient	Interpretation
`OverallQual`	`+10847.37`	Holding other terms fixed, each one-step increase in overall quality (e.g., from 6 to 7) increases predicted price by about **$10.8k**, plus any additional effect through the interaction with size.
`GarageCars`	`+6937.12`	Each additional garage car capacity adds about **$6.9k** to predicted price, beyond interaction effects.
`GarageArea_GarageCars` *	`+30.05`	For each unit increase in both garage area and car count, price increases; practically, **larger garages are more valuable when they hold more cars**, amplifying the effect of both.
`GarageArea`	`-35.64`	Standalone effect is negative, but must be read together with the positive interaction. Extra isolated garage square footage that doesn't increase car capacity adds little or may reflect less valuable configurations.
`OverallQual_GrLivArea` *	`+10.32`	Additional living area is more valuable in higher-quality homes; **size and quality reinforce each other**.
`GrLivArea`	`-22.65`	Standalone effect is negative but again must be interpreted with the positive interaction: baseline square footage without quality is not highly valued, but square footage in higher-quality homes is.

*Interaction terms.

Importance Ranking

By absolute coefficient magnitude (and practical relevance):

1. **Overall quality dominates**

- `OverallQual` has the largest direct positive coefficient (~$10.8k per quality point).
- Combined with the positive `OverallQual_GrLivArea` interaction, higher-quality homes especially benefit from additional square footage.

1. **Garage capacity is the next strongest driver**

- `GarageCars` (~$6.9k per extra car) has a substantial positive effect.
- The `GarageCars_GarageArea` interaction reinforces that more area in a multi-car garage is especially valuable.

1. **Size effects are mediated by quality and configuration**

- Raw `GrLivArea` and `GarageArea` main effects are negative, but their interactions are positive.
- This pattern suggests the model is using interaction terms to capture that "just more space" is not universally valuable; **space is worth more when paired with quality (for living area) or capacity**

(for garages).

In practice, you should view the primary levers as:

- OverallQual (strongest overall driver), then
- GarageCars and the interplay of garage size and capacity, and
- GrLivArea, especially in higher-quality homes.

Interaction Effects Insight

The inclusion of interaction terms improves the realism of the model:

- **OverallQual × GrLivArea**:

The positive coefficient means that larger living areas in higher-quality homes increase price more than the same square footage added to lower-quality homes. For example, adding 200 sq ft to a high-quality home raises value more than adding 200 sq ft to a low-quality home.

- **GarageCars × GarageArea**:

Larger garage area becomes more valuable as car capacity increases. A 600 sq ft one-car garage does not add as much value as a 600 sq ft two-car garage; the interaction captures this non-additive relationship.

These interaction patterns align with how buyers typically think: they value **usable, high-quality space** more than raw square footage alone.

Using the Model to Estimate House Prices

Conceptually, to estimate a house price you would plug its attributes into the equation above. For example, for a home with:

- `OverallQual = 7`
- `GrLivArea = 2,000` sq ft
- `GarageCars = 2`
- `GarageArea = 500` sq ft

You would compute:

1. Interaction terms:

- `OverallQual_GrLivArea = 7 × 2000 = 14,000`
- `GarageCars_GarageArea = 2 × 500 = 1,000`

1. Plug into the regression equation with the estimated coefficients (plus the intercept, which we computed but did not list numerically here).

The resulting predicted `SalePrice` would be your model-based estimate for that configuration of quality, size, and garage characteristics.

For practical deployment or precise point estimates, this formula can be coded directly in a script or spreadsheet using the full set of coefficients and the intercept from the fitted model.

Limitations

Aspect	Limitation
In-sample only	Metrics (R^2, RMSE) are computed on the **same data used for training**, so they may be slightly optimistic compared to true out-of-sample performance.
Limited features	The model only uses four structural features plus interactions. It omits many known drivers, such as neighborhood, age/renovation detail, basement finish, and lot characteristics, which could further improve accuracy.
Linearity assumption	Effects are constrained to be linear (plus simple interactions). Real housing markets may have non-linear relationships (e.g., diminishing returns to size or quality) that are not fully captured.
Extrapolation	The data are concentrated in mid-range quality and size. Predictions for extremely low-quality or very high-end homes are extrapolations and may be less reliable.
Multicollinearity	GarageCars and GarageArea are highly correlated; even with interactions, individual coefficients can be unstable, though overall predictions remain reasonably accurate.

Within these bounds, the model provides a solid, interpretable baseline: **overall quality and garage capacity are the most powerful levers for predicted price, with living area and garage size adding value primarily when paired with higher quality or capacity.** This structure gives you both a practical pricing tool and a clear understanding of which attributes matter most when valuing houses in this dataset.

Data Sources

- Table 1