Story5_Data608

Mubashira Qari

2025-03-24

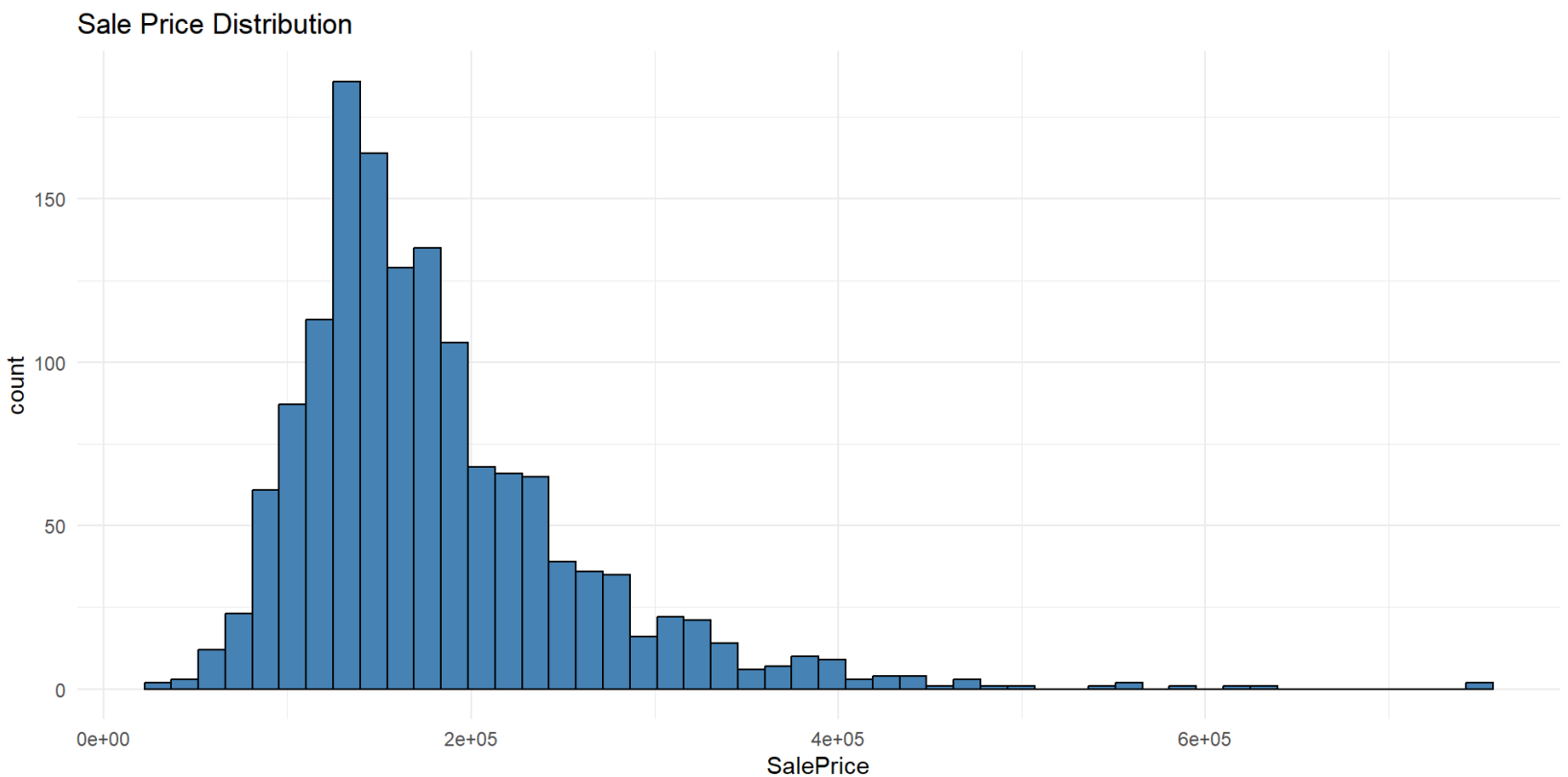
Defining Business Question

This week we will learn to visualize models, and the next week we will learn about dimensionality reduction techniques. A good dataset to practice these tools on is the Ames housing price dataset, which is a sort of “model organism” for machine learning practice. The emphasis of this assignment (described in the pdf) is on building a model or models of the factors determining housing prices and using visualizations to explain their meaning and implications.

Exploratory Data Analysis:

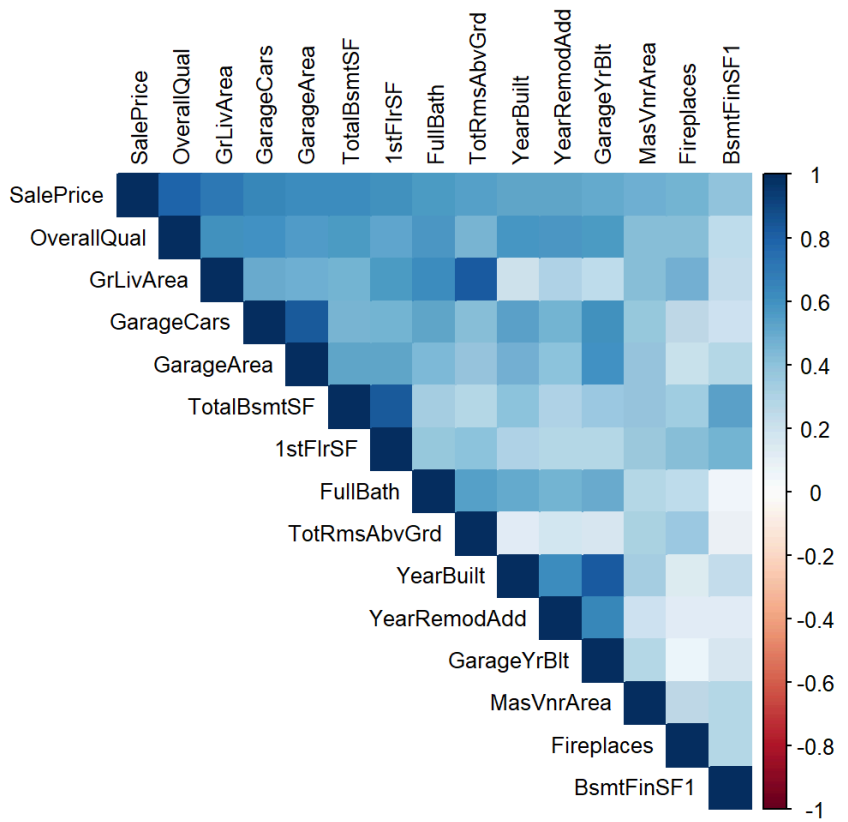
Histogram:

- Visualizing the distribution of housing sale prices to spot skewness or outliers.



Correlation Analysis:

- Finding top numerical features most correlated with SalePrice and plots them in a heatmap for insight.



Feature Engineering:

- Createing new variables:
- log-transformed price, total bathrooms, total porch area, and house age to enhance modeling.

Visual EDA:

- Showing relationships (e.g., living area vs. sale price)
- And correlations between important features using scatter and pair plots.


Living Area vs Sale Price
SalePrice
6e+05
4e+05
2e+05
0e+00
1000
2000
3000
4000
5000
GrLivArea

Train/Test Split:

- Spliting the data into training (80%) and testing (20%) for unbiased model evaluation.

Linear Model:

- Training a basic regression model with key features.

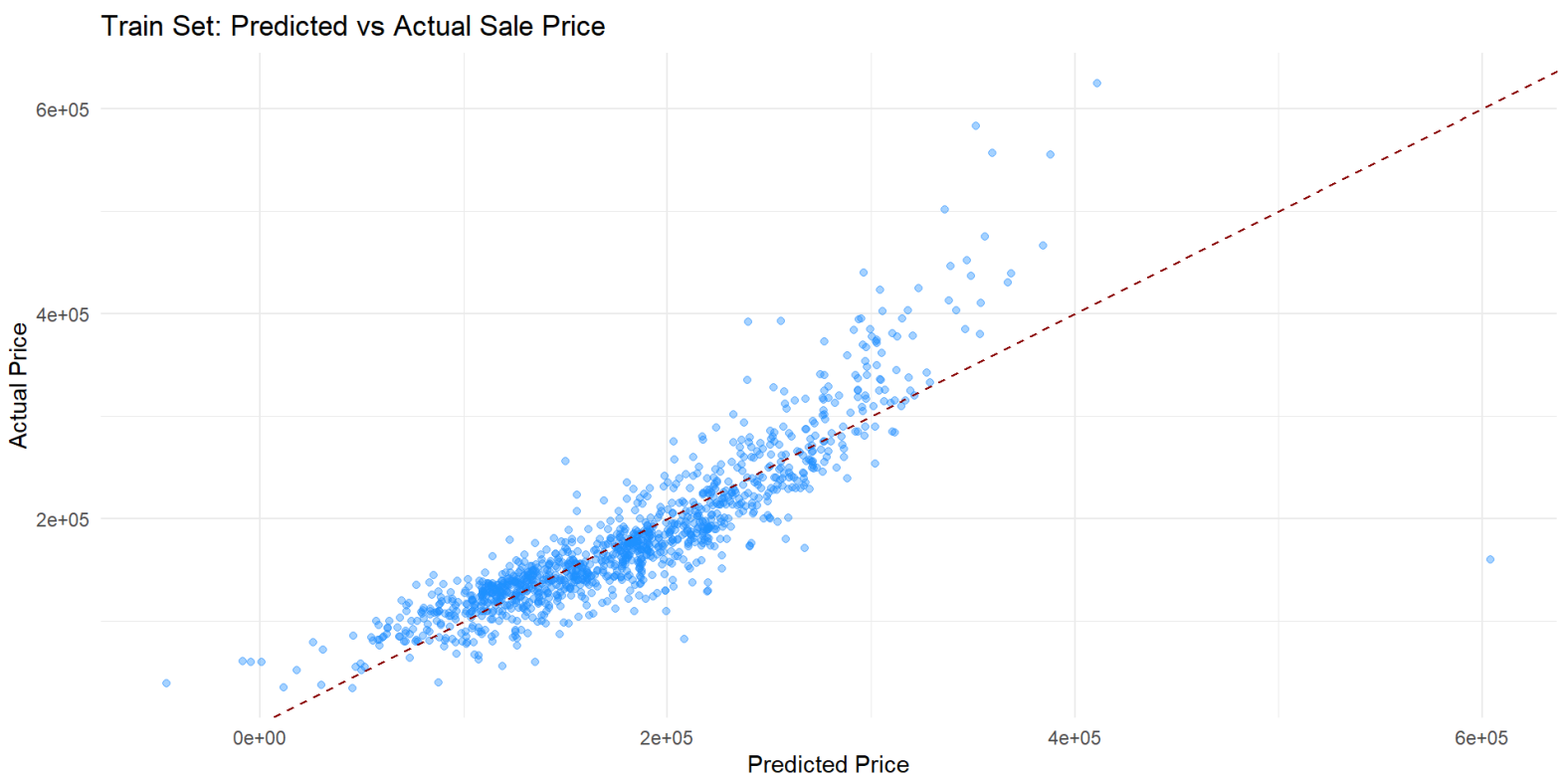
```
Call:
lm(formula = SalePrice ~ GrLivArea + OverallQual + GarageCars +
    TotalBsmtSF + YearBuilt, data = train_df)

Residuals:
    Min      1Q  Median      3Q     Max
-444139  -19365   -2431   16247  231503

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.027e+05  8.769e+04  -8.013 2.70e-15 ***
GrLivArea    4.721e+01  2.778e+00  16.996  < 2e-16 ***
OverallQual  2.074e+04  1.186e+03  17.481  < 2e-16 ***
GarageCars   1.618e+04  1.880e+03   8.609  < 2e-16 ***
```

Visualizing Results:

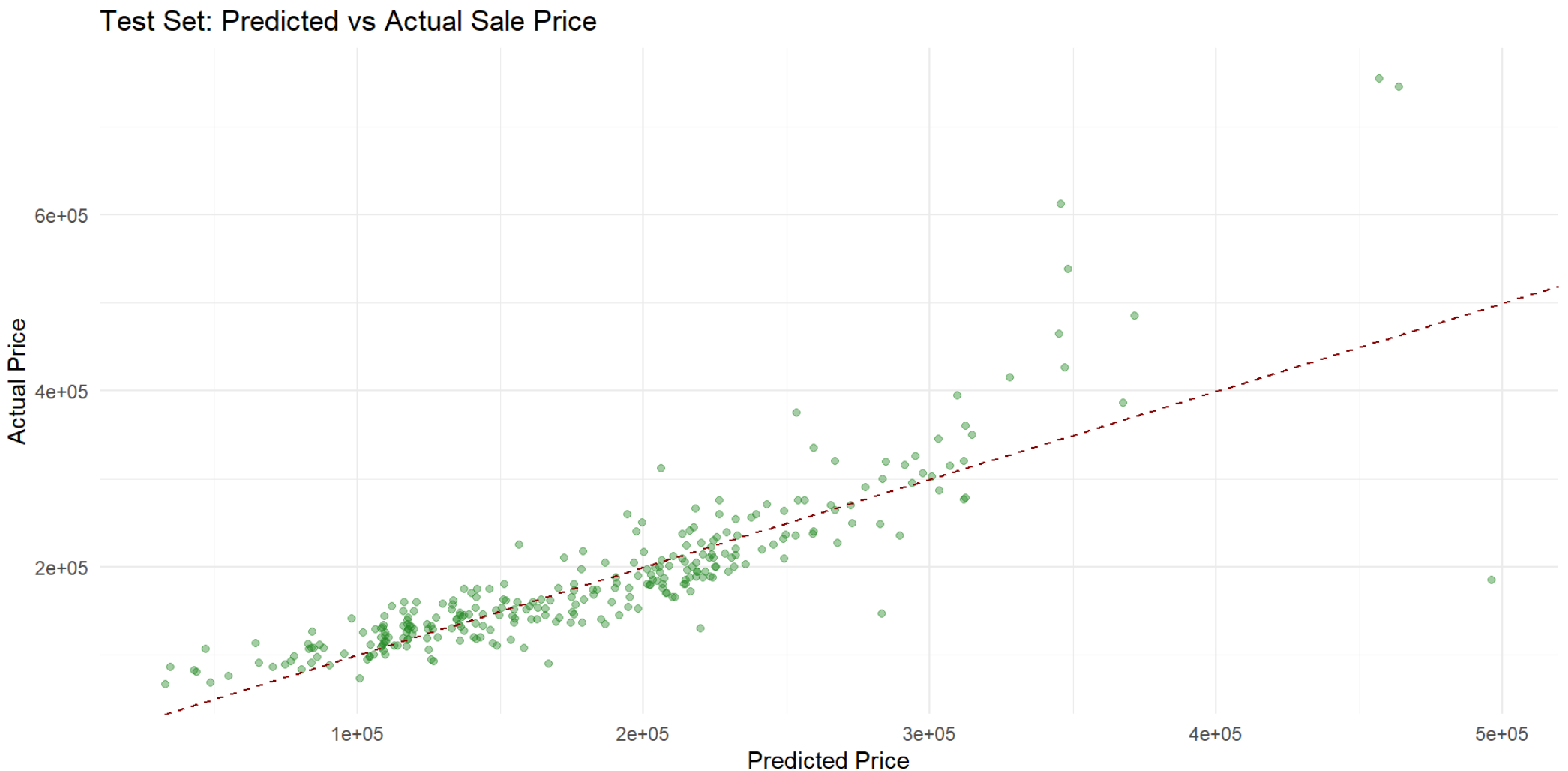
- Predicted vs Actual Sale Prices (Train & Test Sets)
- Means close your model's predictions are to reality



Interpretation:

- Most points cluster tightly along the diagonal line suggest model is accurately predicting many house prices.
- As the predicted price increases (especially > $350k), the points start to spread out more vertically, also called heteroscedasticity.
- This means your model becomes less reliable for more expensive homes
- A few houses are way above or below the red line meaning large errors for a small number of homes.
- These could be homes with unusual characteristics (e.g., poor condition, luxury materials, unique locations)
- Current predictors don’t explain those price differences

Visualization (Test)

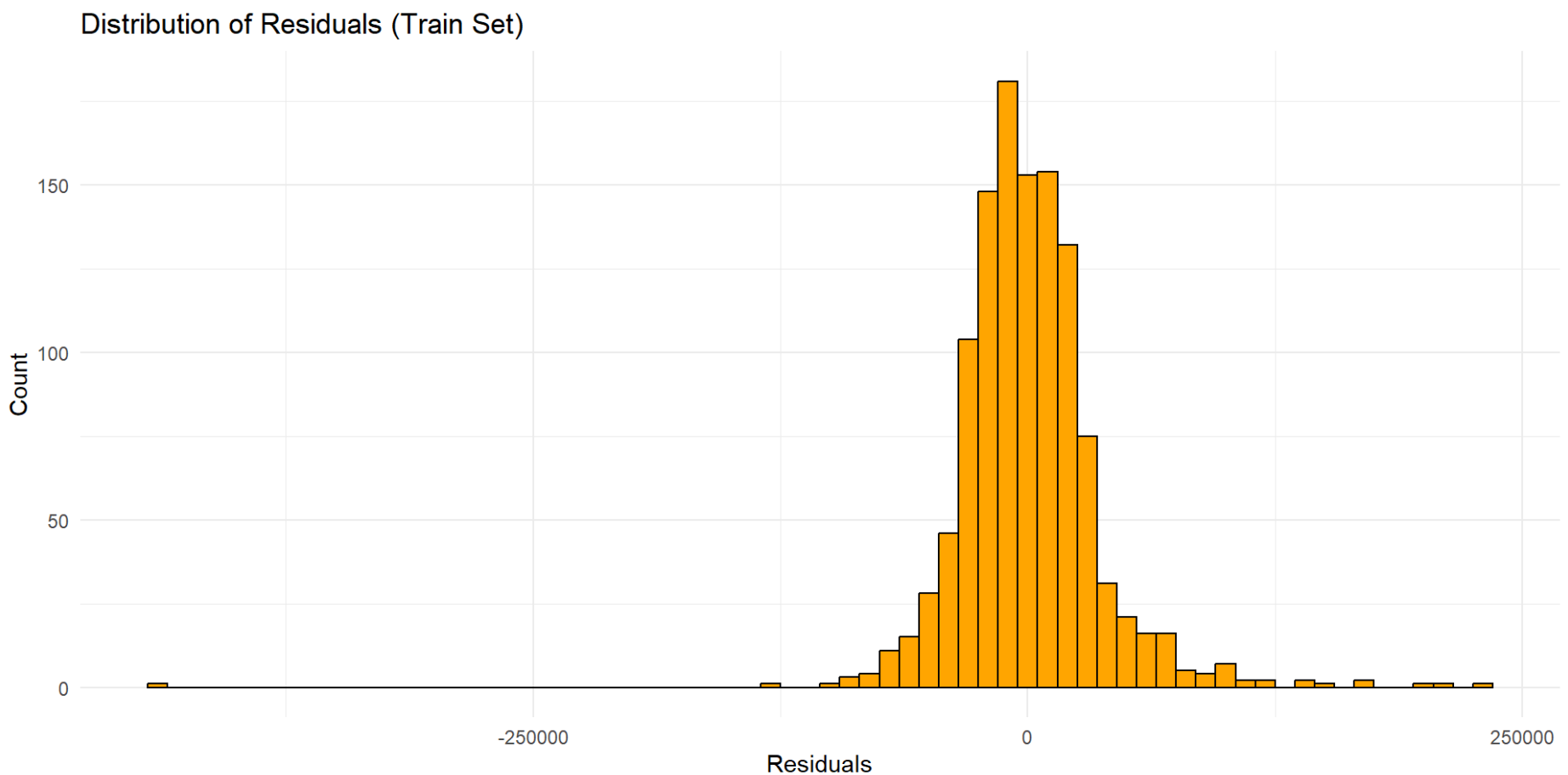


Interpretation:

- Model performs fairly well on average-priced homes and generalizes decently for typical properties in Ames.
- However, it's less reliable for expensive or atypical homes, likely due to:
- Missing features (e.g., neighborhood, condition, amenities)
- Linear model limitations (not capturing complex interactions or non-linear effects)
- This reduced accuracy on the test set especially at higher prices
- Suggests that your model is underfitting for complex cases.

Residuals Distribution

- Whether your errors are centered and symmetric

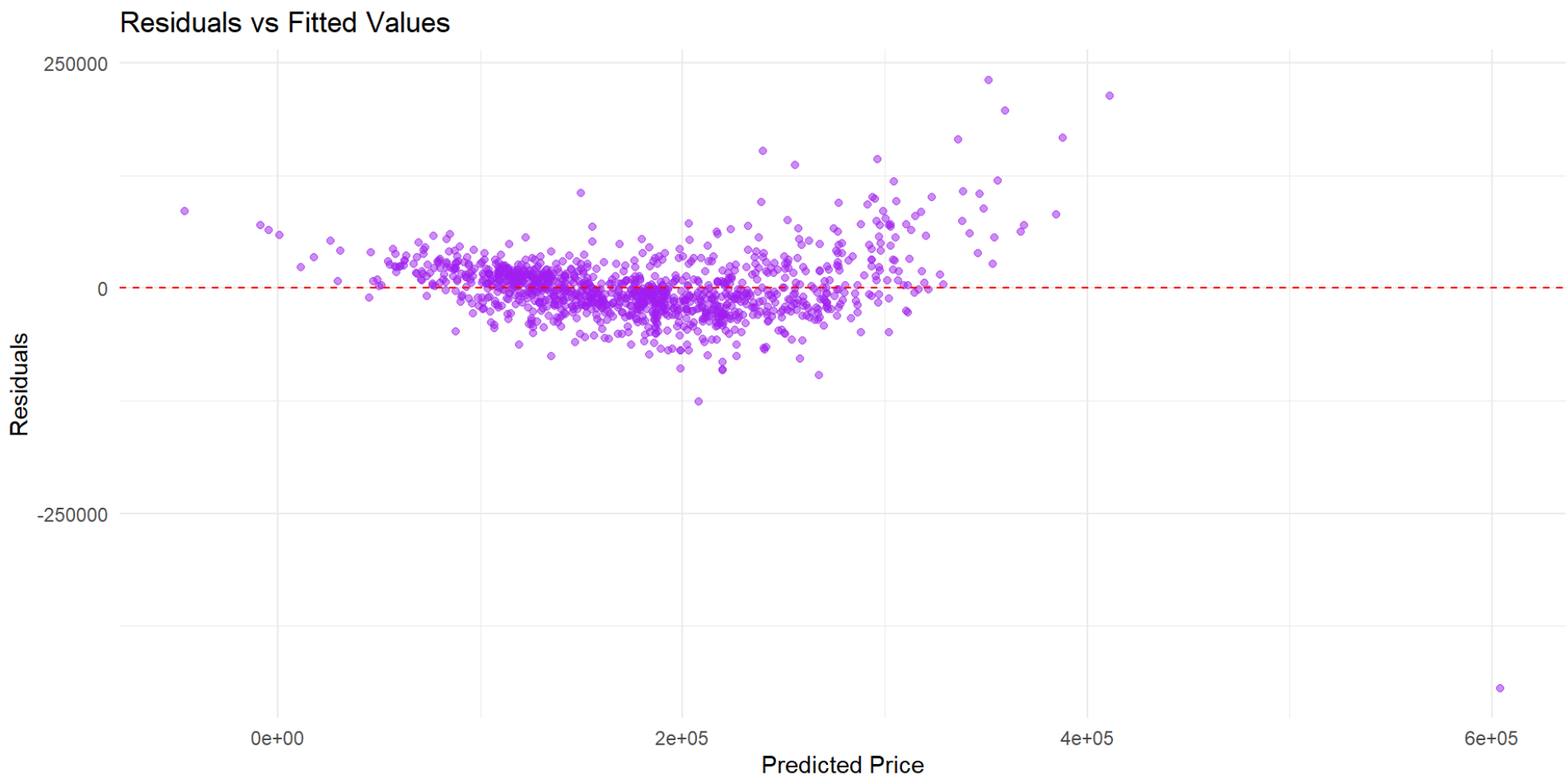


Interpretation:

- The model meets a key assumption of linear regression:
- Residuals are approximately normally distributed.
- However, the slight skew and long tails suggest:
- A few homes are underpredicted by large margins (positive residuals).
- There may be outliers or missing predictors influencing these extreme cases.
- Overall, this supports that the model is well-calibrated for typical homes, but could be improved for edge cases.

3. Residuals vs Fitted Values

– If error patterns suggest model issues (e.g. non-linearity)

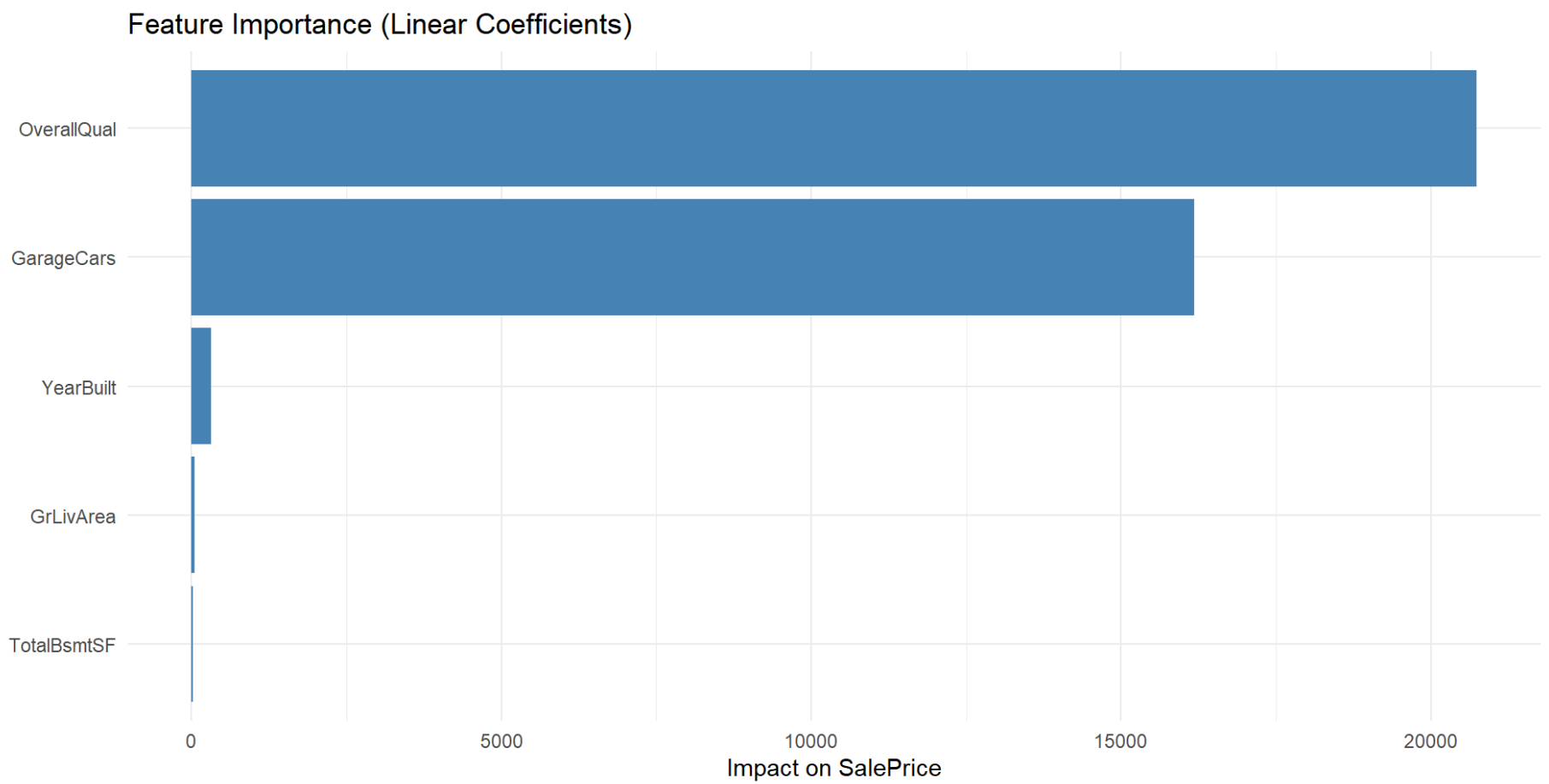


Interpretation

- Most residuals are clustered tightly around zero.
- Wider spread of residuals as predicted prices increase suggests heteroscedasticity.
- Indicates increasing variance in errors — especially for expensive homes.
- Linear model is less reliable for high-priced homes.

4. Feature Importance via Coefficients

- Which features drive price most significantly



Interpretation:

- Build quality and garage space are the strongest linear predictors of home value in Ames and these are the top priorities for buyers.
- Surprisingly, living area and basement size, though important intuitively, contribute less per unit in the linear model — possibly due to:
- Their smaller coefficient scale (e.g., price per sq ft is lower)

Model Performance on Test Set

- Performance check on test data:
- How well your model generalizes beyond training data

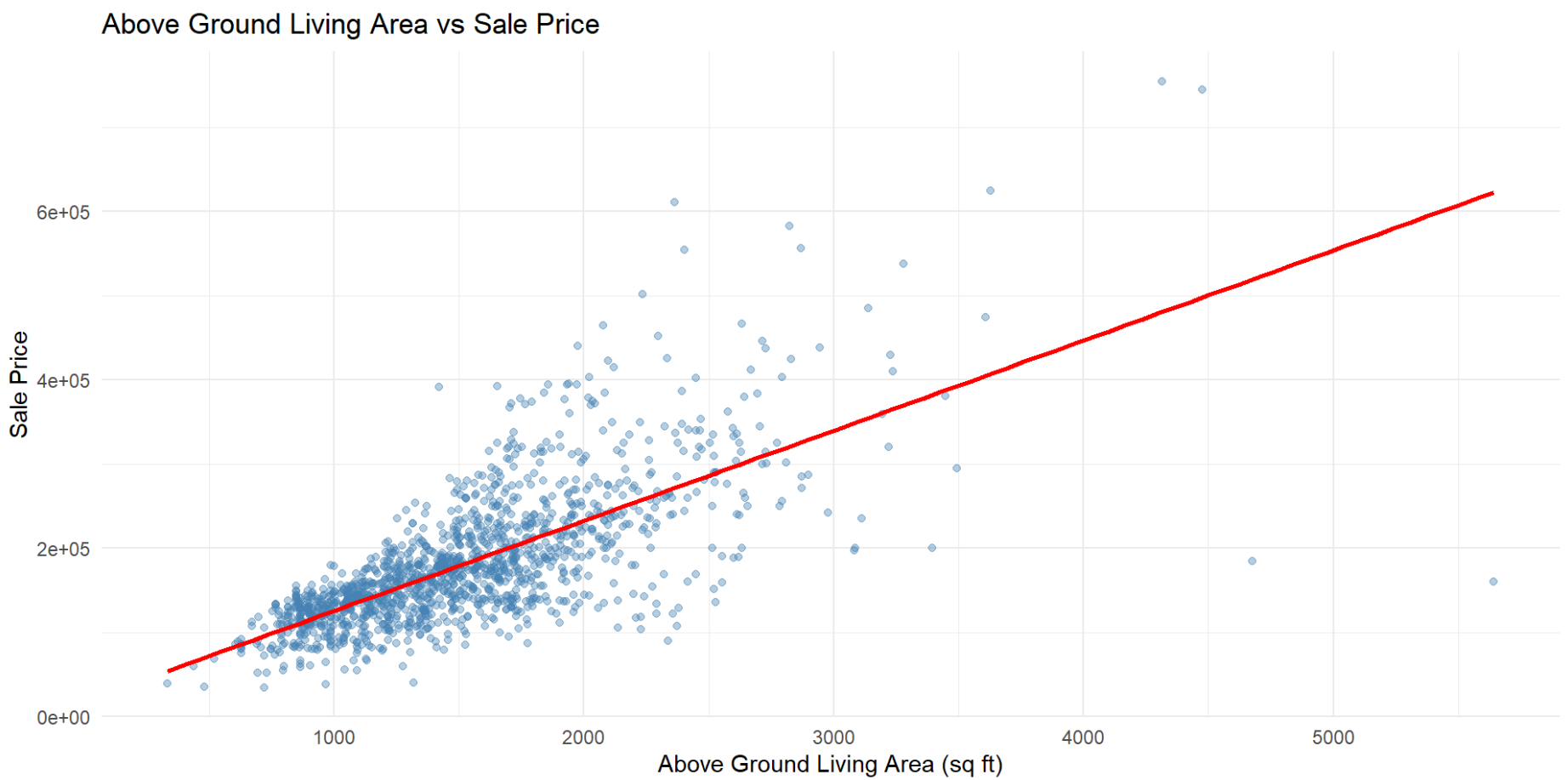
```
Test RMSE: 47295.59
Test R-squared: 0.7309
```

Interpretation:

- RMSE (Root Mean Squared Error) measures the average prediction error in dollars
- On average, model's predictions are off by about $47,296.
- R-squared explains how much variation in sale prices your model accounts for.
- Model explains 73.1% of the variability in home prices on unseen test data.
- Considering the typical home prices in Ames range around $150,000–$300,000, this is a moderate error, roughly 15–30% of a typical sale price.

key visualizations

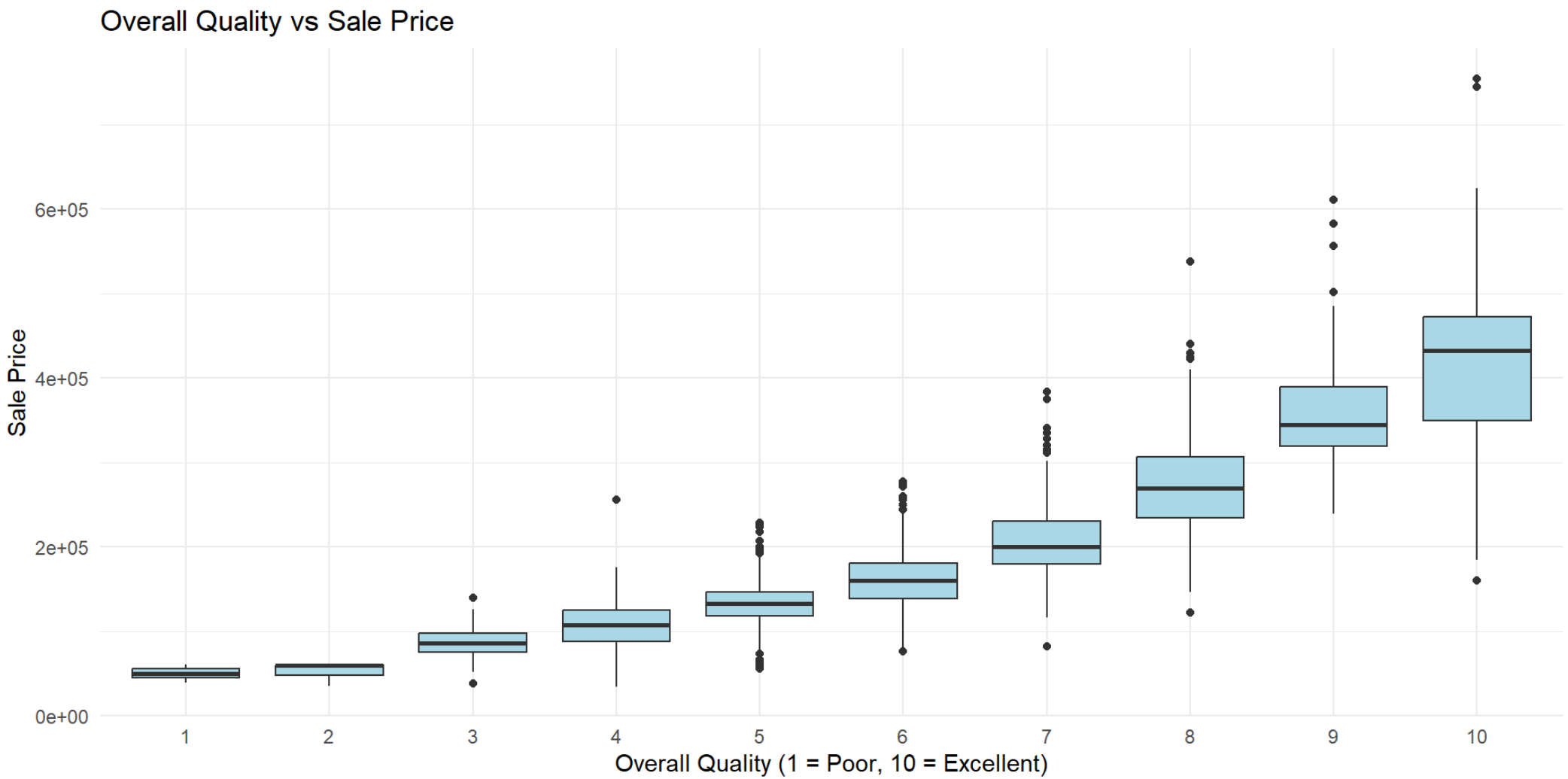
- GrLivArea vs. SalePrice



Interpretation:

- There’s a clear positive linear relationship: larger homes generally cost more.
- Some potential outliers exist (very large homes at lower prices).

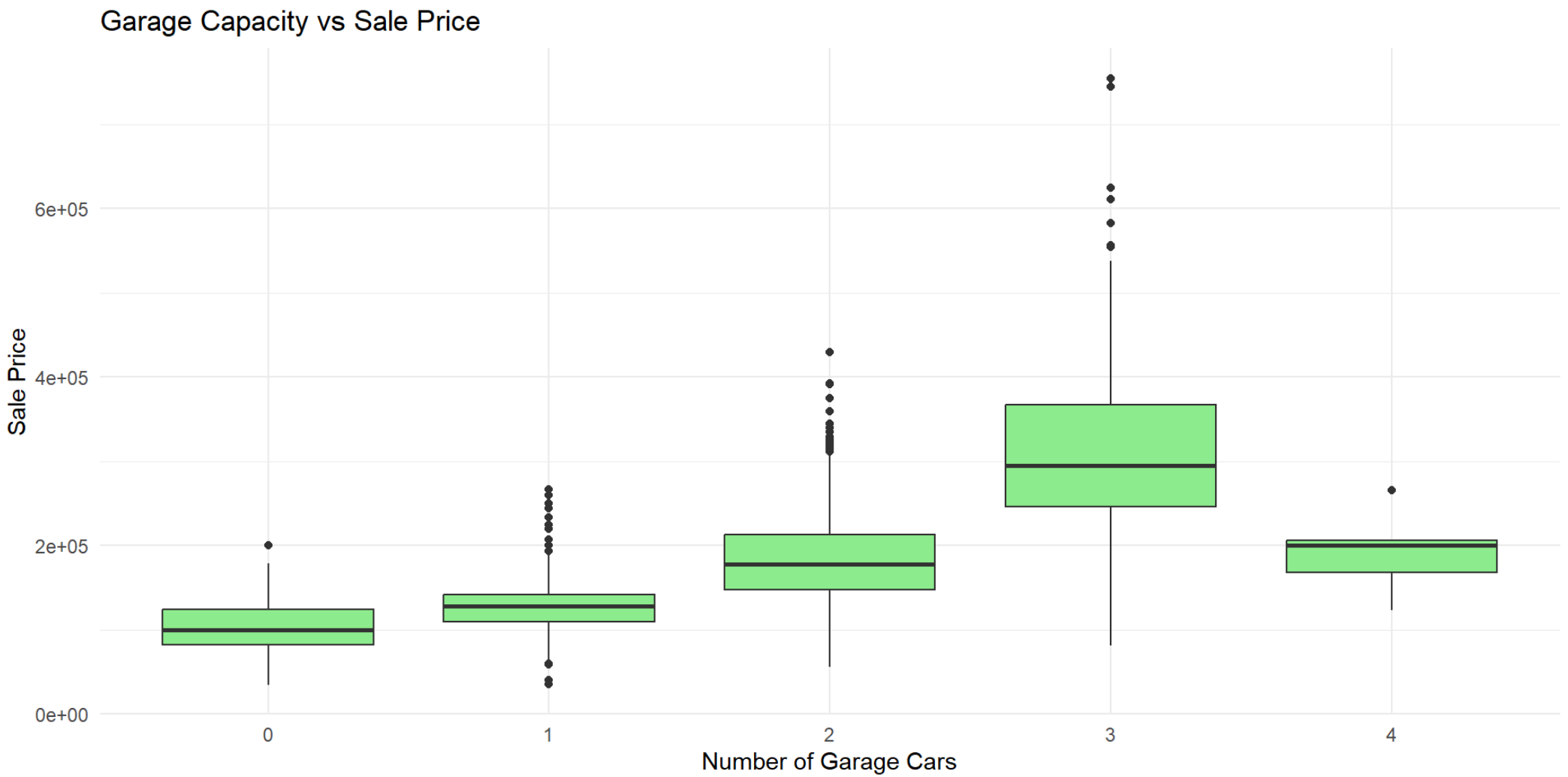
Overall Quality vs. SalePrice



Interpretation:

- Sale price rises sharply with better overall quality.
- A house rated 8 or above commands a premium price.
- The relationship is non-linear and very categorical-sensitive.

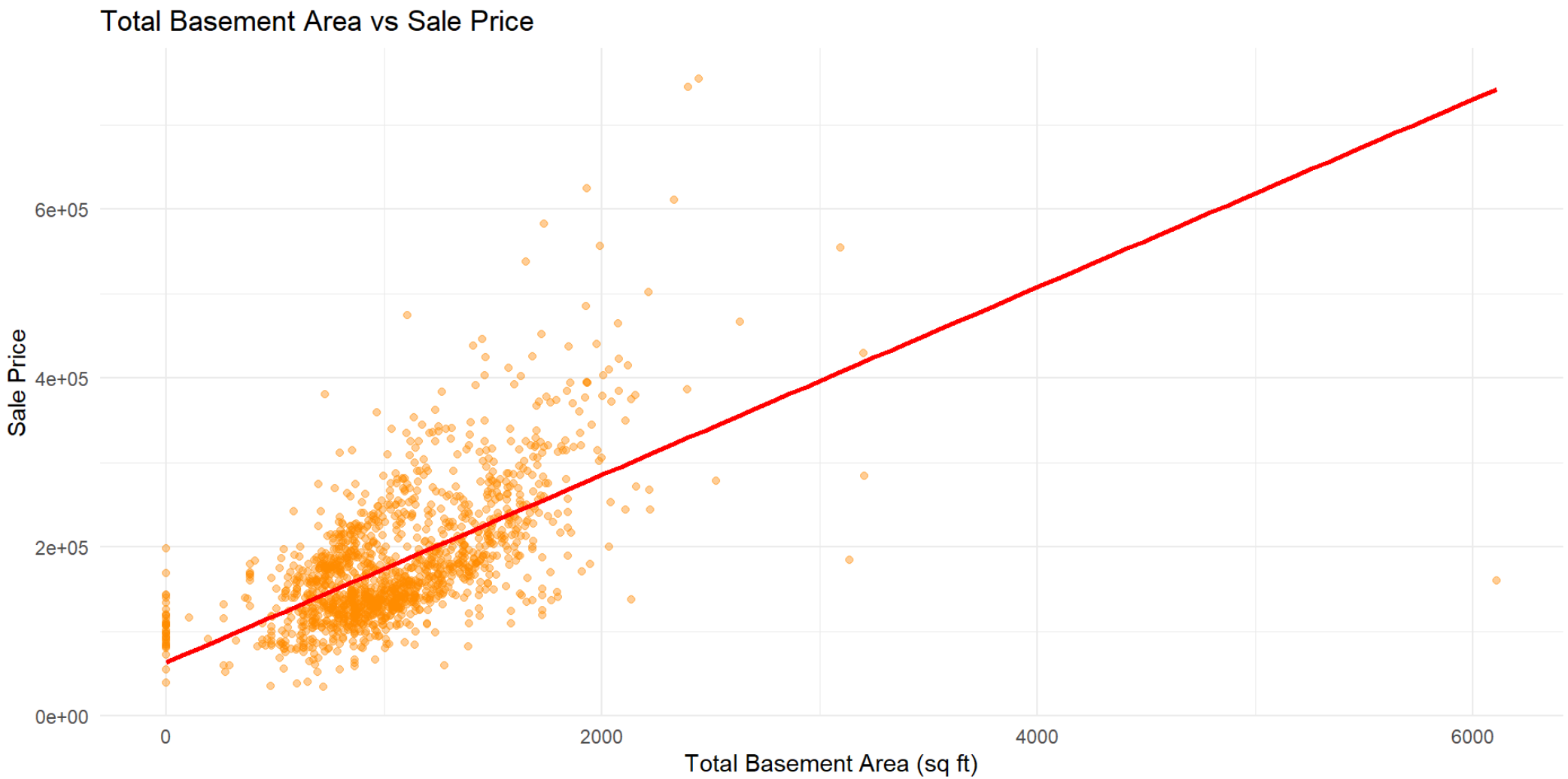
3. Garage Capacity vs. SalePrice



Interpretation:

- Houses with 2+ car garages fetch significantly higher prices.
- Garage capacity is both a space and status factor for buyers.

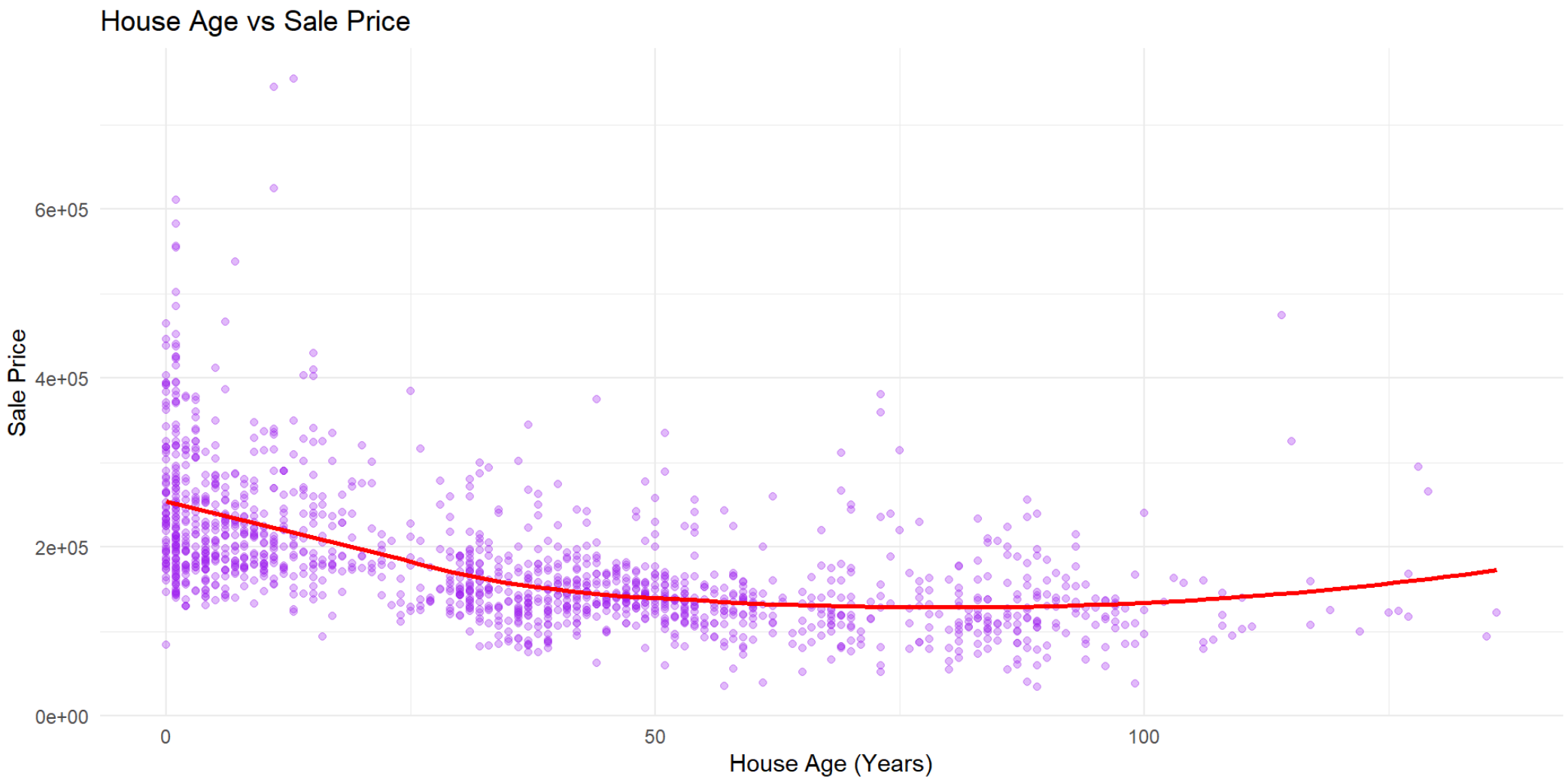
4. Total Basement Area vs. SalePrice



Interpretation:

- Larger basements correlate with higher prices.
- Relationship is positive but diminishing after 2000 sq ft.

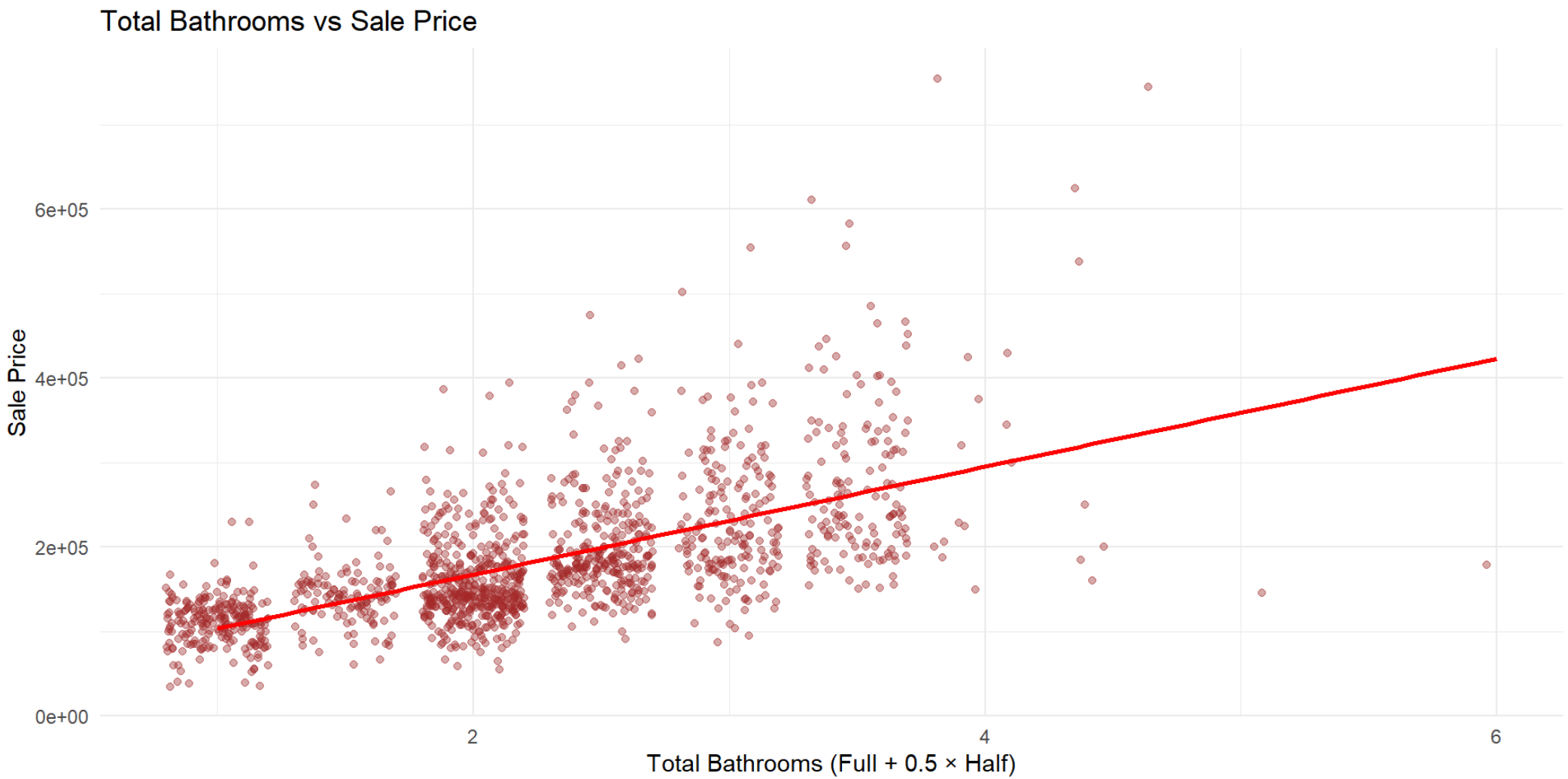
5. House Age vs. SalePrice



Interpretation:

- Newer homes (lower age) are valued higher.
- There’s a non-linear drop-off: value decreases sharply for older homes up to ~30 years, then levels off.

6. Total Bathrooms vs. SalePrice



Interpretation:

- More bathrooms = higher sale price, but effect plateaus after ~3.5.
- Suggests marginal benefit decreases for higher counts.