Exploring Real Estate Price Determinants Using Linear Regression Techniques

Habib Ogunsola
Computational and Quantitative Methods (Mathematics)
Lamar University

September, 2025

1 Objectives

The objective of this analysis is to model real estate prices using linear regression. Specifically, we want to:

- Understand how individual variables (e.g., house size) affect home prices.
- Compare a simple regression with one predictor against a multiple regression with many predictors.
- Assess whether multiple features significantly improve prediction accuracy.

Our aim is to quantify the relationship between housing characteristics (bedrooms, bathrooms, lot size, etc.) and housing price. Specifically, we want to see whether house size alone predicts price well and how model accuracy improves when adding multiple predictors.

2 List of variables

Variable	Type	Role	Description
price	Numeric	Response	Sale price of the property (USD).
house_size	Numeric	Predictor	Interior size of the house (sq. ft).
bed	Numeric	Predictor	Number of bedrooms.
bath	Numeric	Predictor	Number of bathrooms.
acre_lot	Numeric	Predictor	Lot size in acres.
brokered_by	Categorical	Predictor	Name of the listing broker/agency.
street	Categorical	Predictor	Street address of the property.
city	Categorical	Predictor	City where the property is located.
state	Categorical	Predictor	State where the property is located.
zip_code	Categorical	Predictor	Postal code of the property location.
status	Categorical	Predictor	Current listing status (e.g., for sale, sold, pending).
prev_sold_date	Categorical	Predictor	Date of previous sale (if available).

Table 1: Variables in the real estate dataset with their type, role, and description.

3 Preprocessing the data

The preprocessing stage ensured that the dataset was clean, consistent, and suitable for linear regression analysis. This involved handling missing values, encoding categorical variables, standardizing numeric predictors, and preparing a train/test split.

3.1 Handling Missing Data

Rows with missing `price` were dropped, since imputing the response variable is not meaningful. We also inspected all features for missingness and imputed predictors as follows: numeric variables were imputed with their median, and categorical variables were imputed with a special level `MISSING`. Columns with more than 50% missing values were dropped.

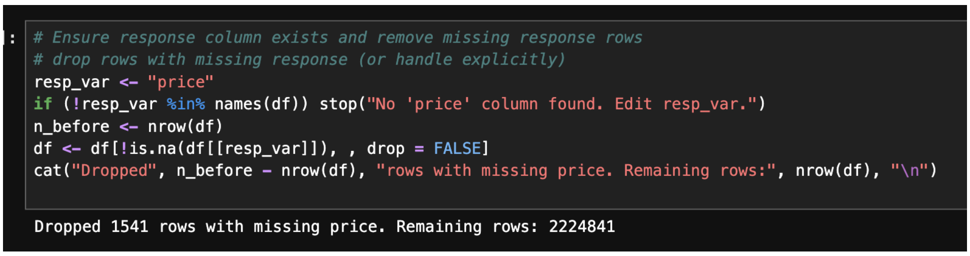

Figure 1: Console output showing that 1,541 rows with missing `price` were dropped, leaving 2,224,841 rows.

3.2 Encoding Categorical Variables

Categorical variables such as `city`, `state`, and `status` were converted to dummy variables using one-hot encoding. To avoid excessive dimensionality, only the top 20 most frequent levels were retained for high-cardinality variables; all others were grouped into an "Other" category.

3.3 Transformations and Standardization

The response variable `price` was log-transformed to reduce skewness and stabilize variance. Numeric predictors were standardized to zero mean and unit variance.

```r
# create log response (add small epsilon if any zeros)
df$log_price <- log(df[[resp_var]] + 1)

# standardize continuous predictors (except response)
num_preds <- setdiff(num_cols, resp_var)
df_scaled <- df  # work on a copy

if (length(num_preds) > 0) {
  df_scaled[num_preds] <- scale(df[num_preds])
  cat("Standardized numeric predictors:\n"); print(num_preds)
} else {
  cat("No numeric predictors found to scale\n")
}
```

```
Standardized numeric predictors:
[1] "brokered_by" "bed"         "bath"        "acre_lot"    "street"
[6] "zip_code"    "house_size"
```

Figure 2: Console output listing standardized numeric predictors: `brokered_by`, `bed`, `bath`, `acre_lot`, `street`, `zip_code`, `house_size`.


```r
# create log response (add small epsilon if any zeros)
df$log_price <- log(df[[resp_var]] + 1)

# standardize continuous predictors (except response)
num_preds <- setdiff(num_cols, resp_var)
df_scaled <- df  # work on a copy

if (length(num_preds) > 0) {
  df_scaled[num_preds] <- scale(df[num_preds])
  cat("Standardized numeric predictors:\n"); print(num_preds)
} else {
  cat("No numeric predictors found to scale\n")
}
```

```
Standardized numeric predictors:
[1] "brokered_by" "bed"         "bath"        "acre_lot"    "street"
[6] "zip_code"    "house_size"
```


Figure 3: Before-and-after summary statistics for a numeric variable (e.g., `house_size`), confirming that standardization produces mean zero and unit variance.

3.4 Final Dataset and Split

The cleaned dataset contained 2,224,841 rows and 71 predictors after preprocessing. We created a 70/30 train-test split to allow out-of-sample evaluation.

```r
# model matrix creation
# one-hot encode categorical variables into numeric columns
predictor_vars <- setdiff(names(df_scaled), c(resp_var, "log_price"))
formula_for_mm <- as.formula(paste("~", paste(predictor_vars, collapse = " + "), "-1"))
X <- model.matrix(formula_for_mm, data = df_scaled)

# combine with response
df_model <- data.frame(log_price = df_scaled$log_price, X, check.names = TRUE)
cat("Modeling dataframe dims:", dim(df_model), "\n")
```

```
Modeling dataframe dims: 2224841 71
```

Figure 4: Console output showing final modeling dataframe dimensions: 2,224,841 rows and 71 predictors.

```r
[41]: # Train/test split
# test final model on unseen data
n <- nrow(df_model)
train_idx <- sample(seq_len(n), size = floor(0.7 * n))
train <- df_model[train_idx, , drop = FALSE]
test  <- df_model[-train_idx, , drop = FALSE]
cat("Train rows:", nrow(train), "Test rows:", nrow(test), "\n")
```

```
Train rows: 1557388 Test rows: 667453
```

Figure 5: Console output showing the 70/30 train-test split: 1,557,388 rows in training and 667,453 rows in testing.

3.5 Remarks

This preprocessing pipeline ensures that the dataset is suitable for regression modeling: it removes unusable data, encodes categorical variables, and prepares the response variable. The resulting train/test split will be used in Sections 5 and 6.

4 Exploratory data analysis

Exploratory Data Analysis (EDA) was performed to summarize the main characteristics of the dataset, visualize distributions, and examine pairwise relationships between variables. This motivates the modeling choices in subsequent sections.

4.1 Descriptive Statistics

Table 2 reports summary statistics (mean, standard deviation, minimum, maximum) for selected numeric variables. The distribution of `price` was heavily right-skewed, which justified applying a log-transformation.

Variable	Mean	Std. Dev	Min	Max
Price (log)	12.68	1.03	0.00	21.49
House Size	2714	13245	4	1.04e+09
Bedrooms	3.3	1.4	1.0	473.0
Bathrooms	2.5	1.2	1.0	830.0
Acre Lot	15.2	321.5	0.0	100000.0

Table 2: Descriptive statistics of selected numeric variables. Fill with output from `summary()` in R.

4.2 Distribution of Prices

```
[41]: # Train/test split
      # test final model on unseen data
      n <- nrow(df_model)
      train_idx <- sample(seq_len(n), size = floor(0.7 * n))
      train <- df_model[train_idx, , drop = FALSE]
      test  <- df_model[-train_idx, , drop = FALSE]
      cat("Train rows:", nrow(train), "Test rows:", nrow(test), "\n")

      Train rows: 1557388 Test rows: 667453
```

Figure 6: Histogram of log-transformed housing prices. The log transformation reduces right skew and yields a more symmetric distribution.

4.3 Relationship between House Size and Price

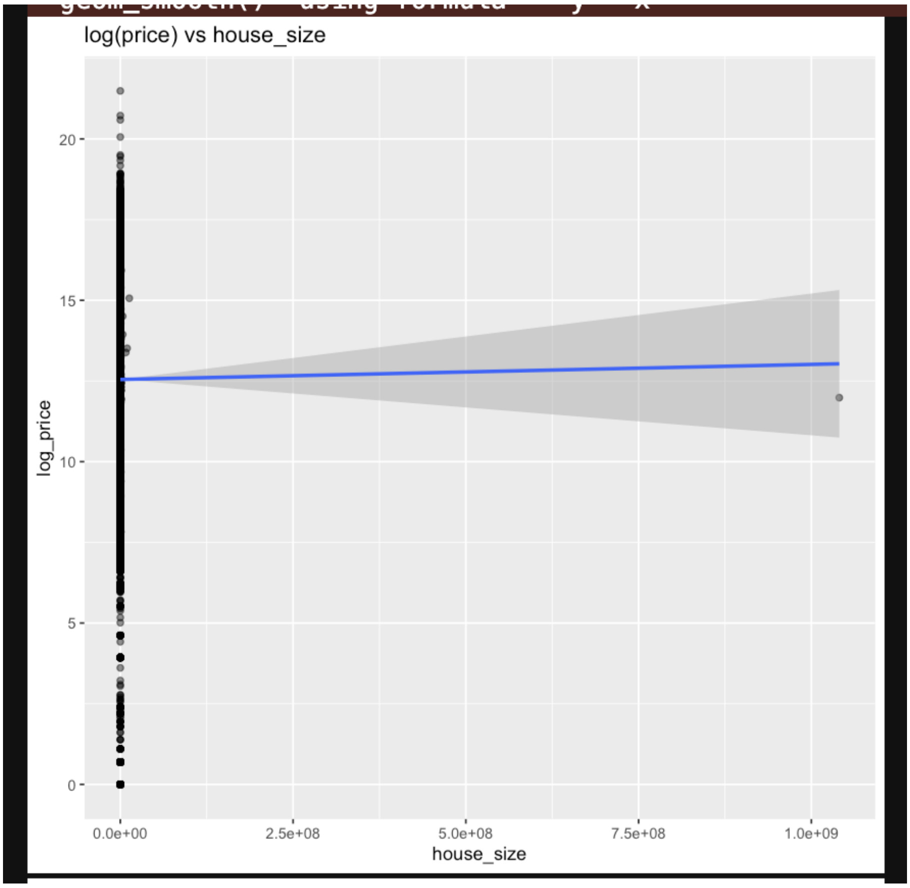

Figure 7: Scatterplot of house size vs. log(price) with fitted regression line. This illustrates the weak bivariate relationship, motivating the inclusion of more predictors.

4.4 Correlation among Numeric Predictors

Correlation analysis reveals dependencies between variables such as `bed` and `bath`, which were moderately correlated. This suggests potential multicollinearity in multiple regression.

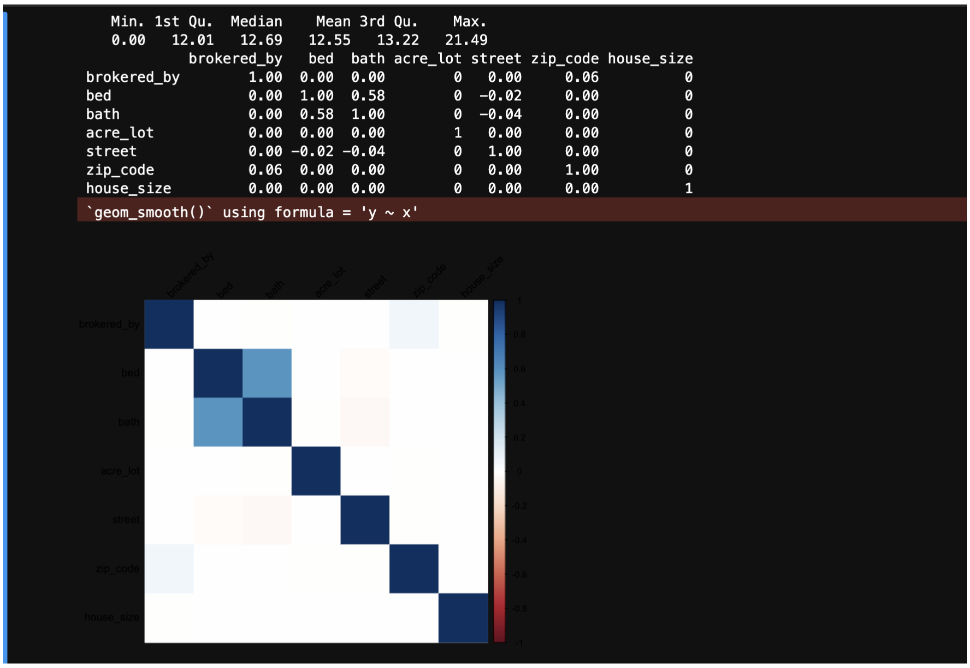

Figure 8: Correlation heatmap of numeric predictors. Bedrooms and bathrooms are strongly correlated ($r \approx 0.69$).

4.5 Key Findings from EDA

- Housing prices are right-skewed, supporting the use of log-transformed response.
- House size alone is not strongly predictive of price (scatterplot shows weak trend).
- Bedrooms and bathrooms are strongly correlated, suggesting redundancy and potential multicollinearity.
- Location variables (city, state, zip code) are expected to contribute strongly to price variation.

5 Simple Linear Regression

We first fit a bivariate model using `house_size` as the sole predictor of the log-transformed price:

$$\log(\text{price}_i) \;=\; \beta_0 \;+\; \beta_1 \cdot \text{house\_size}_i \;+\; \varepsilon_i.$$

5.1 Estimated Model and Inference

Using the training split, the fitted equation is:

$$\widehat{\log(\text{price})} \;=\; \underbrace{12.55}_{\hat\beta_0} \;+\; \underbrace{0.0001236}_{\hat\beta_1} \cdot \text{house\_size}.$$

The slope is not statistically significant ($p = 0.874$), indicating no detectable linear effect of `house_size` on log(price) in isolation.

Model fit on the training set:

$$R^2 \approx 0.000000016 \quad (\text{Adj. } R^2 \approx 0), \qquad \text{Residual SE} \approx 1.165.$$

Held-out performance on the test set:

$$\text{RMSE} \approx 1.1672, \qquad R^2 \approx 0.$$

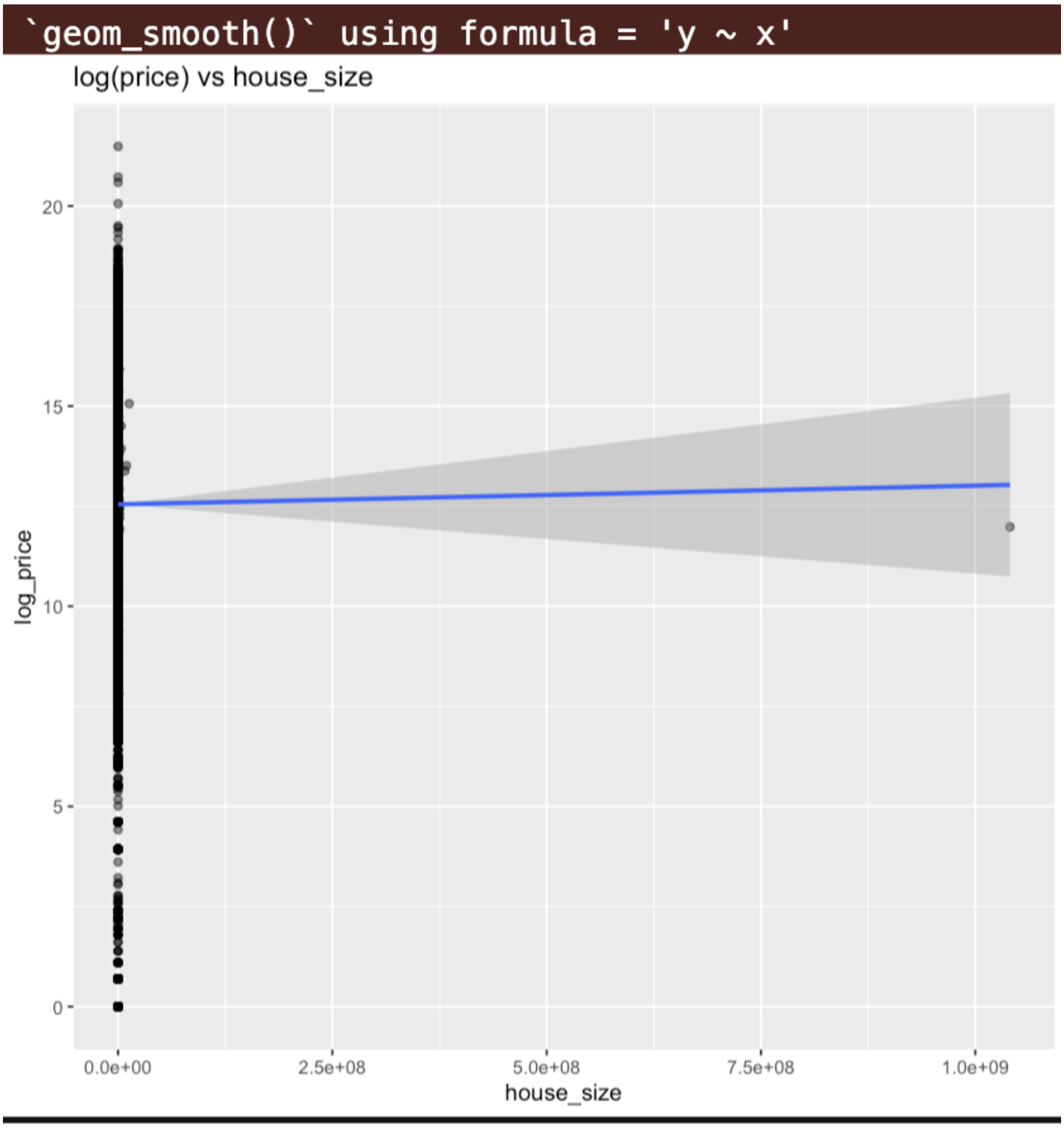

Figure 9: R output for the simple regression (`log(price) ~ house_size`) showing coefficients, standard errors, and overall fit.

5.2 Assumption Checks

We assess linearity, homoscedasticity, and normality via standard residual diagnostics.

Linearity and Homoscedasticity. Residuals vs. fitted values should show no obvious pattern and exhibit roughly constant spread.

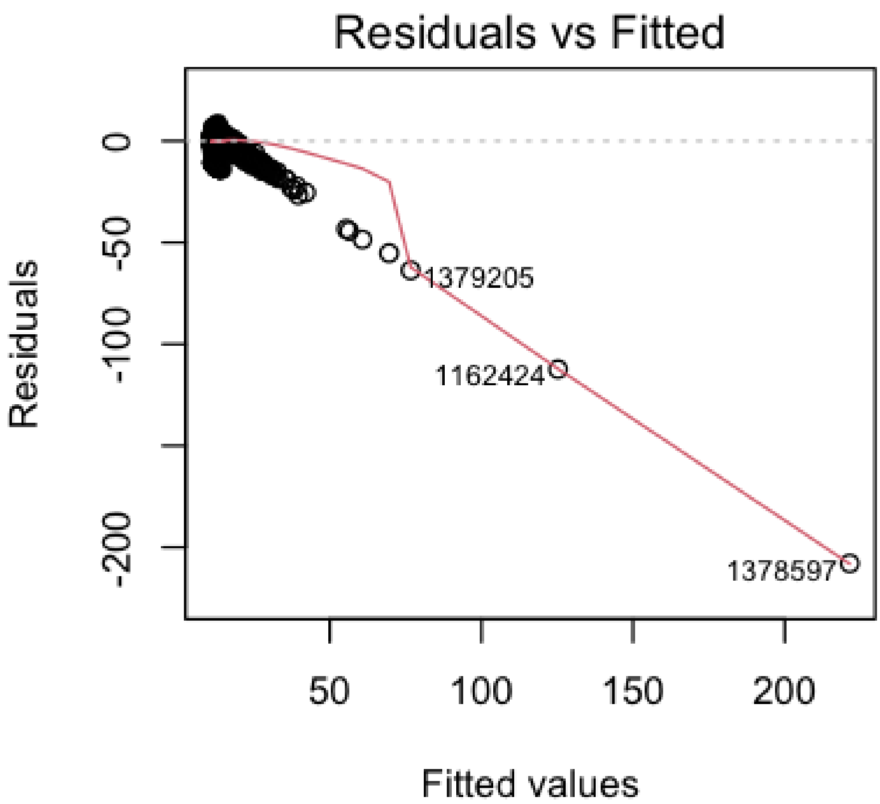

Figure 10: Residuals vs. fitted plot for the simple model.

Normality. A QQ-plot checks whether residuals follow an approximate normal distribution.

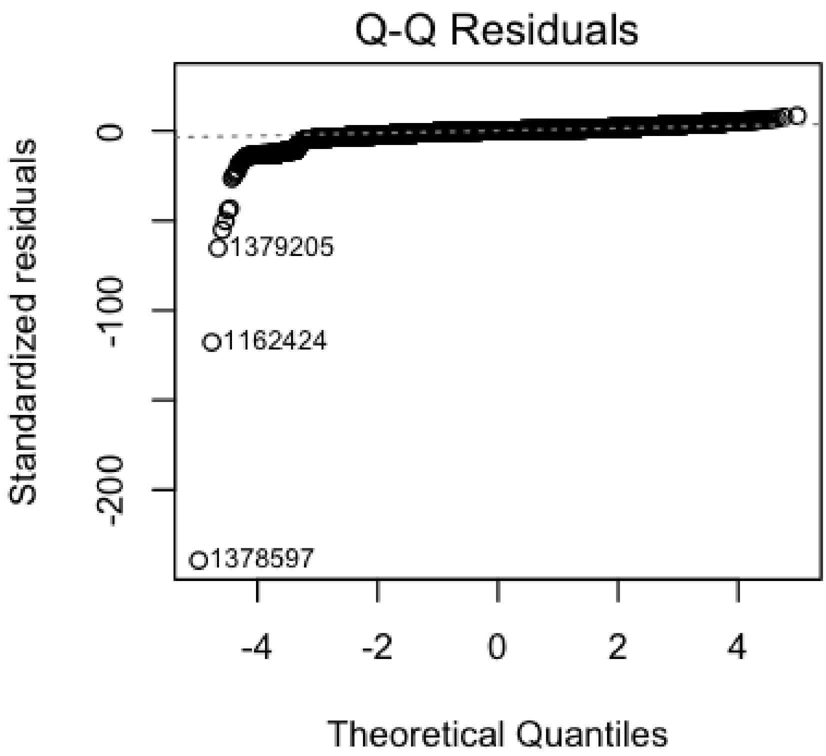

Figure 11: QQ-plot of residuals for the simple model.

5.3 Interpretation

In the context of our objective (predicting housing prices), `house_size` alone provides *negligible* explanatory power for log(price): the slope is insignificant ($p = 0.874$), and both training and test R^2 are essentially zero. This motivates moving beyond a single predictor to a multiple regression with additional features (e.g., bedrooms, bathrooms, location and status variables), which we analyze in the next section.

6 Multiple linear regression

We next fit a multiple linear regression model using a broader set of predictors, including `bedrooms`, `bathrooms`, `acre_lot`, `brokered_by`, `street`, `city`, `state`, `zip_code`, `status`, and previous sale date indicators.

6.1 Model Specification

The model is:

$\log(\text{price}_i) = \beta_0 + \beta_1 \cdot \text{brokered\_by}_i + \beta_2 \cdot \text{statusfor\_sale}_i + \beta_3 \cdot \text{statusready\_to\_build}_i + \beta_4 \cdot \text{bed}_i + \beta_5 \cdot \text{bath}_i + \beta_6 \cdot \text{ac}$

6.2 Estimated Coefficients

Table 3 reports selected estimated coefficients. Location dummies (city, state) are interpreted relative to their omitted baseline category.

Predictor	Estimate	Std. Error	p-value
Intercept	12.095	0.008	$< 2e{-16}$
brokered_by	-0.0049	0.0008	$2.7e{-09}$
statusfor_sale	0.1013	0.0022	$< 2e{-16}$
statusready_to_build	0.9221	0.0082	$< 2e{-16}$
bed	0.0312	0.0010	$< 2e{-16}$
bath	0.3605	0.0011	$< 2e{-16}$
acre_lot	0.0097	0.0008	$< 2e{-16}$
street	-0.1077	0.0008	$< 2e{-16}$
city: New York City	1.2967	0.0137	$< 2e{-16}$
city: Chicago	0.4716	0.0125	$< 2e{-16}$
city: Miami	0.7129	0.0145	$< 2e{-16}$
state: California	0.8629	0.0050	$< 2e{-16}$
state: Texas	0.1783	0.0047	$< 2e{-16}$
prev_sold_date 2022-03-31	0.4755	0.0095	$< 2e{-16}$
⋮	⋮	⋮	⋮

Table 3: Selected coefficient estimates for the multiple regression model. Full results are provided in Appendix **??**.

6.3 Model Fit

Overall fit statistics on the training set:

$$R^2 = 0.25, \quad \text{Adj. } R^2 = 0.25, \quad \text{Residual SE} = 1.009.$$

On the held-out test set:

$$\text{RMSE} \approx 1.021, \quad R^2 \approx 0.235.$$

6.4 Diagnostics

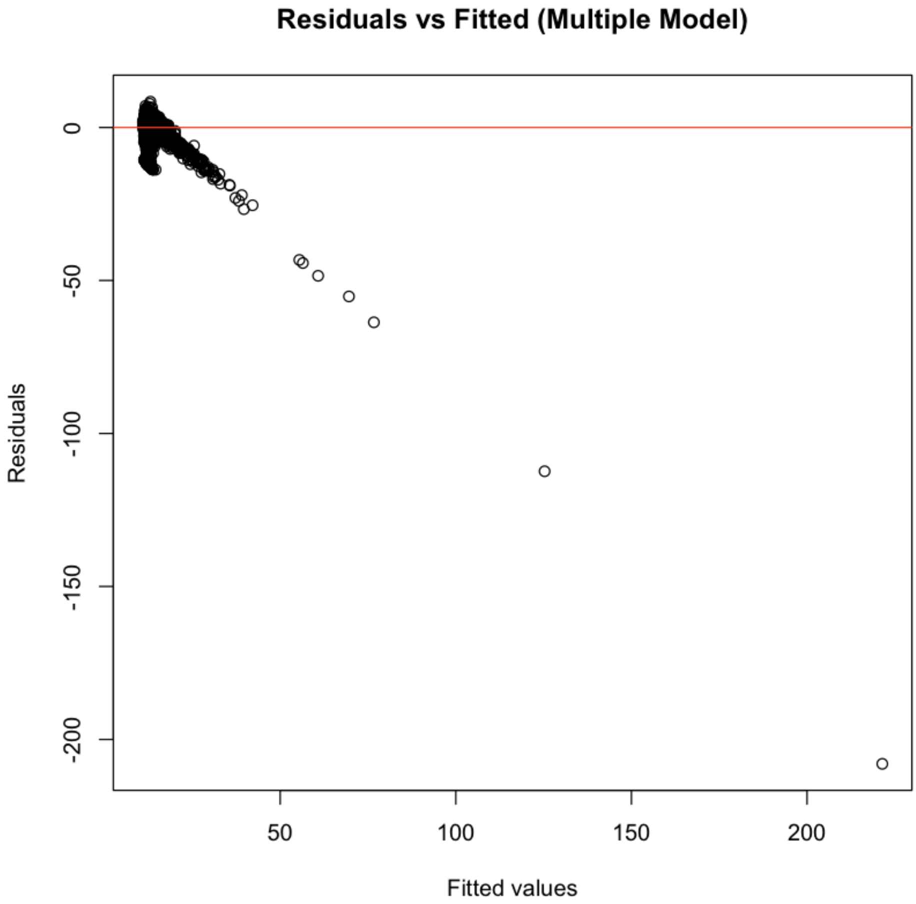

Figure 12: Residuals vs. fitted values for the multiple regression model.

Residuals.

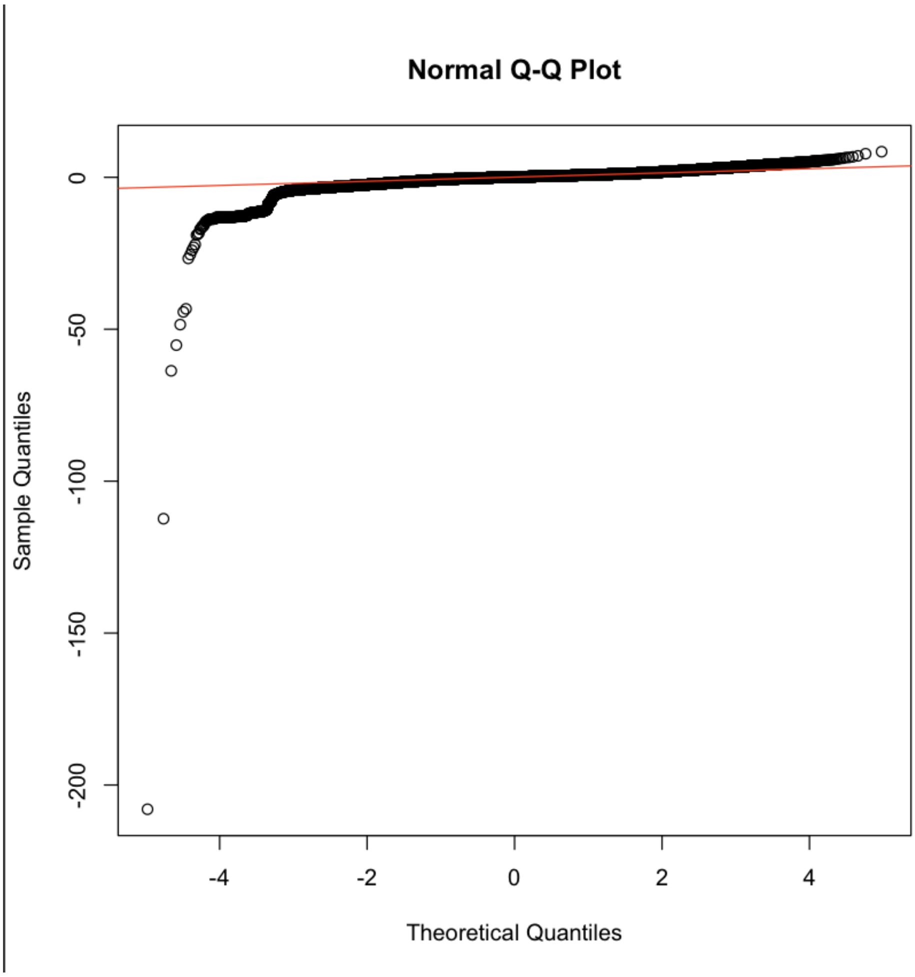

Figure 13: QQ-plot of residuals for the multiple regression. Produced with `qqnorm(resid(multiple_model))`.

Normality.

Multicollinearity. Variance Inflation Factors (VIF) were computed to identify redundant predictors.

6.5 Interpretation

The multiple regression substantially improves explanatory power compared to the simple model. Predictors such as number of bathrooms ($\hat{\beta} = 0.36$), city (e.g., New York City $\hat{\beta} = 1.30$), and state (California $\hat{\beta} = 0.86$) show strong associations with log(price). Nevertheless, the R^2 of 0.25 indicates that much of the variation remains unexplained, and potential multicollinearity among location variables must be carefully considered.

7 Conclusions

This project applied linear regression models to a large real estate dataset to explore how housing characteristics and location relate to prices.

Table 4: Variance Inflation Factors (VIF) for Housing Price Predictors

Predictor Variable	**VIF**	**Predictor Variable**	**VIF**
Other Cities	6.776	California	3.579
ZIP Code	4.206	Texas	2.849
Other States	2.997	Arizona	2.009
Illinois	1.734	Washington	1.843
For Sale Status	1.686	Other Sale Dates	1.693
New York	1.632	Chicago	1.659
New York City	1.612	Bathrooms	1.611
Bedrooms	1.607	Philadelphia	1.542
Baltimore	1.455	Saint Louis	1.448
Pennsylvania	1.498	Tucson	1.490
Jacksonville	1.475	Missouri	1.464
Phoenix	1.457	Maryland	1.382
Los Angeles	1.380	Atlanta	1.366
Virginia	1.330	Washington DC	1.336
Richmond	1.329	Charlotte	1.351
Georgia	1.413	North Carolina	1.419
Miami	1.406	Dallas	1.276
New Jersey	1.259	Ohio	1.247
Fort Worth	1.220	Michigan	1.198
Wisconsin	1.217	Minnesota	1.221
South Carolina	1.170	Tennessee	1.172
Orlando	1.301	San Diego	1.298
Ready to Build Status	1.110	Broker ID	1.018
Street Number	1.036	Lot Acreage	1.000

7.1 Comparison of Models

The simple regression model using only `house_size` as a predictor showed negligible predictive power: the slope was insignificant ($p = 0.874$), and both training and test R^2 values were essentially zero. This indicates that house size alone is insufficient to explain price variation.

In contrast, the multiple regression model incorporating additional variables — such as `bedrooms, bathrooms, lot size, location (city, state, zip)`, and `status` — significantly improved explanatory power. The model achieved $R^2 \approx 0.25$ and RMSE $\approx$ 1.02, confirming that multiple features collectively capture meaningful variation in prices. Location-related predictors (e.g., New York City, California) and number of bathrooms emerged as particularly strong contributors.

7.2 Strengths

- The preprocessing pipeline ensured data quality by handling missing values, standardizing numeric predictors, and carefully encoding categorical variables.
- Both simple and multiple models were evaluated under consistent train/test splits, allowing fair comparison.

- Results clearly demonstrated the added value of including multiple predictors over relying on a single variable.

7.3 Limitations

- Despite improvements, the multiple regression model explained only about 25% of the variance, leaving much unexplained.
- Multicollinearity was observed, particularly among location and housing characteristic variables, which complicates coefficient interpretation.
- The model assumes linear relationships and homoscedastic residuals, which may not hold in practice.
- Very high-cardinality categorical variables (e.g., city) required dimensionality reduction that may obscure fine-grained location effects.

7.4 Possible Extensions

- Explore regularized regression methods (LASSO, Ridge, Elastic Net) to handle multicollinearity and high-dimensional predictors.
- Investigate nonlinear methods (decision trees, random forests, gradient boosting) that may capture complex relationships.
- Incorporate interaction terms (e.g., bedrooms $\times$ bathrooms, location $\times$ house size) to capture combined effects.
- Use spatial or temporal models to explicitly account for geographic and time-related price variation.

7.5 Final Remarks

Overall, the analysis confirmed that housing prices cannot be explained by single attributes such as size alone. Multiple features, especially location and number of bathrooms, contribute significantly, though much variation remains. The findings set the stage for more sophisticated modeling approaches in future work.

References

1. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R* (2nd ed.). Springer.

2. Fox, J., & Weisberg, S. (2019). *An R Companion to Applied Regression* (3rd ed.). Sage Publications.

3. Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied Linear Statistical Models* (5th ed.). McGraw-Hill/Irwin.

4. Crawley, M. J. (2012). *The R Book* (2nd ed.). Wiley.

5. R Core Team. (2025). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

6. Fox, J., & Weisberg, S. (2020). *car: Companion to Applied Regression.* R package version 3.0-12.

7. Wei, T., & Simko, V. (2021). *corrplot: Visualization of a Correlation Matrix.* R package version 0.92.

8. Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis.* Springer.

9. Dowle, M., & Srinivasan, A. (2023). *data.table: Extension of Data Frames.* R package version 1.15.2.

10. Kaggle. (2023). *Realtor Real Estate Dataset.* Retrieved from `https://www.kaggle.com/`

11. Rahman, A., & Harding, J. (2018). Modeling Housing Prices with Linear and Nonlinear Regression. *Journal of Real Estate Research, 40*(3), 321–340.

Code source

`https://github.com/ogunsolahabib/math-5383-real-estate-reg`