# From Insights to Predictions: Regression Modelling of Ames Housing Prices.

Prepared by

**Rebecca Stalley-Moores**

# From Insights to Predictions: Regression Modelling of Ames Housing Prices.

---

Prepared by: **Rebecca Stalley-Moores**

Date: **20/08/2025**

---

## Executive Summary

This analysis **investigated** the factors influencing house sale prices using a comprehensive dataset of sold residential properties. During the **exploratory data analysis (EDA)**, univariate summaries **showed** that most homes were modest in size, with smaller garages, porches, and bathrooms being common, while room count features approximated a normal distribution. Key categorical variables such as Neighborhood, House Style, and MS Subclass **displayed** distinct groupings potentially relevant for pricing, and ordinal quality indicators like Overall Cond and Kitchen Qual **reflected** expected ordered patterns.

Bivariate analysis **highlighted** strong positive correlations between sale price and engineered features combining living area and overall quality, particularly the **qual_living_area_interaction** variable, which **outperformed** its individual components in capturing price variation. Other size-related features, including garage and basement areas, **showed** moderate to strong relationships with price. Several features **exhibited** high collinearity, indicating potential redundancy for future modelling.

The **iterative modelling workflow refined** these insights:

- **Full OLS and Outlier-Removed Models:** Initial regression **identified** influential predictors but **was** affected by multicollinearity and extreme observations. Removing four extreme observations improved coefficient stability, providing a cleaner basis for subsequent models, though it permanently excluded these potentially influential points..
- **Yeo-Johnson Transformation: Stabilized** variance and **improved** model robustness while retaining all data.
- **Stepwise BIC Selection: Produced** a parsimonious feature set with good predictive accuracy and interpretability, emphasizing predictors like **qual_living_area_interaction** and select neighborhood dummies.
- **Ridge Regularization: Was selected** as the preferred model, achieving predictive performance comparable to the BIC model (Test $R^2 \approx 0.912$, RMSE ≈ 0.210) while drastically **reducing** multicollinearity (VIFs < 10). Ridge **retained** all predictors, **downweighted** over-represented neighborhood effects, and **emphasized** stable structural features such as **BsmtFin SF 1**, **Lot Area_capped**, **total_bathrooms**, **remodel_age**, and **house_age**. Coefficient signs **remained** consistent with previous models, confirming directionality of effects, while shrinkage **improved** overall stability.

Initial hypothesis testing during EDA supported the importance of qual_living_area_interaction, showing significant differences in sale prices across low, medium, and high groups. Neighborhood

and garage characteristics **remained** relevant, though Ridge **demonstrated** that penalization **reduced** over-reliance on extreme or collinear categorical effects, emphasizing features with broader generalizability.

**Next Steps:**

- Apply the train/test split immediately after EDA to ensure unbiased evaluation of model generalization.
- Explore additional interactions and nonlinear transformations for temporal and categorical features.
- Evaluate nonlinear and machine learning models (e.g., Random Forest, Gradient Boosting) to capture complex patterns beyond linear effects.
- Conduct robustness checks for outliers and consider robust regression techniques if necessary.
- Perform post-hoc analyses including effect sizes, confidence intervals, and cross-validation to strengthen interpretability and stability.
- Incorporate domain insights from Ridge and BIC models to guide feature selection, hypothesis testing, and business-relevant interpretations.

Overall, the project **demonstrated** that engineered features combining quality and size were central to explaining sale price variation. Iterative modelling, culminating in Ridge regularization, **provided** a robust, interpretable, and generalizable framework for predictive analysis, balancing accuracy with structural understanding of Ames property values.

# Objectives

Building upon a comprehensive exploratory data analysis of the Ames Housing dataset, this analysis aims to **develop and evaluate linear regression models to accurately predict residential property sale prices**.

## Primary Objectives:

1. **Build predictive models** using the key features identified in the EDA, particularly the engineered qual_living_area_interaction variable that showed the strongest correlation (r=0.87) with sale prices
2. **Compare multiple regression approaches** to determine the optimal modelling strategy, including:
    - Simple linear regression using our best single predictor
    - Multiple linear regression incorporating highly correlated features
    - Regularized regression methods to address the multicollinearity identified
    - Polynomial regression to capture any non-linear relationships detected
3. **Quantify predictive performance** using appropriate metrics ($R^2$, RMSE, MAE) and validate model generalizability through proper train-test evaluation
4. **Identify the most important predictive features** and provide actionable insights for real estate stakeholders about what drives property values in Ames, Iowa

## Success Criteria:

- Achieve strong predictive accuracy while maintaining model interpretability
- Address the multicollinearity issues identified in the correlation analysis
- Handle the right-skewed distributions and outliers appropriately
- Provide reliable predictions that could inform pricing decisions

---

# Data Summary

- The goal of the analysis portion is to explore the factors influencing house prices.
- The dataset is sourced directly from the Ames Housing dataset on the Kaggle platform.
- The original dataset contains 2,930 rows and 82 columns, representing residential property sales in Ames, Iowa.
- The variables include a mix of numeric and categorical data types, such as:

    - Structural features (e.g., OverallQual, YearBuilt, GrLivArea)
    - Location details (e.g., Neighborhood, LotConfig)
    - Sale-related info (e.g., SalePrice, SaleType)

- In its original format there are 10 float columns, 28 integer columns and 43 object columns.
- A typical variable example:

    - OverallQual: Rates the overall material and finish of the house (1 = Poor, 10 = Excellent)

- The target variable for analysis is `SalePrice`, representing the sale price of each home.

---

# Data Cleaning and Feature Engineering

## Data Cleaning

- Null values in numeric features were carefully analysed to determine whether they indicated absence of a feature (e.g., no garage) or were genuinely missing data. Nulls indicating absence were replaced with zeros, while genuine missing data were imputed using median values.
- A similar approach was taken for null categorical values; where the null was due to feature absence, it was replaced with the category 'None'. For the Electrical feature, the mode was used as the nulls appeared to be genuine missing data, since most houses are expected to have electricity.
- Data types were verified and appeared correct—only three features with discrete numeric counts were converted from floats to integers.
- No duplicated rows were found.
- Checks for categorical variations found consistent categories across variables.
- Many features exhibited substantial skewness and outliers.

    - Log and Yeo-Johnson transformations were attempted but did not improve skewness and reduced interpretability.
    - Consequently, some features were capped based on the following decision-making process:

| Feature | Q3 (75%) | Max | Outlier Factor | What it is | Decision |
|---|---|---|---|---|---|
| Lot Area | 11,555 | 215,245 | 18.6× Q3 | Land size (sq ft) | Cap at 99th percentile |
| Mas Vnr Area | 162.75 | 1,600 | 9.8× Q3 | Masonry veneer area | Cap at 99% or investigate |
| Wood Deck SF | 168 | 1,424 | 8.5× Q3 | Wood deck square footage | Cap at 99% |
| Open Porch SF | 70 | 742 | 10.6× Q3 | Open porch square footage | Cap |
| Enclosed Porch | 0 | 1,012 | ∞ | Closed porch square footage | Usually 0, so anything >200 is extreme . Cap or binarize |
| 3Ssn Porch | 0 | 508 | ∞ | 3-season porch | Cap at 95–99% |
| Screen Porch | 0 | 576 | ∞ | Screened porch | Cap at 95–99% |
| Pool Area | 0 | 800 | ∞ | Size of pool | Most homes have no pool . Bin into 0 / >0 or cap |
| Misc Val | 0 | 17,000 | ∞ | Miscellaneous (e.g., sheds, tennis courts) | Bin or cap at 99% |
| SalePrice | 213,500 | 755,000 | 3.5× Q3 | Target: sale price | Cap cautiously or not at all |
| Gr Liv Area | 1742.75 | 5642.00 | 3.28 x Q3 | Square footage of living space | Cap at 95–99% |

- Logical inconsistencies were checked, for example:
    - Rows where `Garage Cars` > 0 but `Garage Area` and `Garage Yr Blt` were zero (2 rows) were removed.
    - 521 inconsistencies were found where `Gr Liv Area` was less than the sum of `1st Flr SF, 2nd Flr SF,` and `Low Qual Fin SF`. Due to the high count and lack of clear correction, a flag column was created to mark these rows for modelling consideration.
- No unusual string lengths or blank strings were found.

## Data Wrangling and Feature Engineering

- New features were created to enhance predictive power and insight, including:
    - **house_age:** Years since construction (`Yr Sold - Year Built`)
    - **remodel_age:** Years since last remodel (`Yr Sold - Year Remod/Add`)
    - **years_to_remodel**: Time from build to remodel (`Year Remod/Add - Year Built`)
    - **total_bathrooms:** Sum of all full and half bathrooms (basement and above grade), with half baths weighted as 0.5
    - **total_porch_area:** Sum of porch areas (`Wood Deck SF, Open Porch SF, Enclosed Porch, 3Ssn Porch, Screen Porch`) to capture outdoor living space

    - **total_bsmt_finished**: Sum of basement finished areas (`BsmtFin SF 1` and `BsmtFin SF 2`)
    - **living_area_ratio**: Ratio of above-ground living area (`Gr Liv Area`) to `Lot Area`, representing housing density
    - **avg_quality:** Combined rating from `Overall Qual` and `Overall Cond` (mean)
    - **season_sold:** Categorization of `Mo Sold` into seasons (Winter, Spring, Summer, Fall)
    - **qual_living_area_interaction:** Interaction between `Overall Qual` and `Gr Liv Area`
- These transformations were saved into a new DataFrame to preserve the original data.

---

# Data Exploration Plan

The objective of the exploratory data analysis (EDA) is to develop an understanding of the key factors that influence house prices in the Ames Housing dataset. This will inform hypothesis generation and future predictive modelling efforts.

## Analytical Approach

- **Univariate Analysis**
  Examine the distribution of individual variables to understand their range, skewness, central tendency, and presence of outliers.

    - *Numerical features*: Histograms and boxplots (particularly for features that were capped or transformed).
    - *Categorical features*: Count plots to assess frequency distributions and category imbalances.

- **Bivariate Analysis**
  Investigate the relationships between the target variable (`SalePrice_capped`) and other features to suggest relationships and identify potential predictors.

    - *Numerical vs Target*: Scatter plots and correlation heatmaps.
    - *Categorical vs Target*: Boxplots grouped by category to examine how sale prices vary by feature.

- **Multivariate Analysis**
  Explore interactions between multiple variables to uncover deeper patterns and potential collinearity. This will include:

    - Correlation matrix to identify strongly related features.
    - Grouped boxplots for categorical and numerical combinations.
    - Outlier and anomaly detection to flag extreme values or inconsistencies that could affect interpretation and modelling.

## Exploration Goals

The analysis will focus on answering the following questions:

- **Which features show the strongest correlation with sale price?**
  *Identifying informative predictors for modelling.*

- **How do categorical variables (e.g., `Neighborhood`, `House Style`, Exterior) influence sale prices?**
  *Understanding categorical effects that numerical variables may not capture.*
- **How can engineered features improve predictive power beyond raw variables?**
  *Investigating the value of composite features (e.g., `qual_living_area_interaction`) that integrate multiple aspects like quality and size to explain sale price variation more effectively than individual variables.*

These insights will form the foundation for developing hypotheses, selecting impactful variables, and preparing the dataset for subsequent predictive modelling.

---

# EDA and Discussion

## Univariate Analysis

### Numerical Features

The dataset contains numerous numerical variables but plotting and analysing all would lead to excessive complexity and dilute interpretability. Therefore, a carefully selected subset of numerical features was chosen for exploration based on:

- **Correlation with SalePrice_capped**:

Features strongly correlated with the target variable were:

- `qual_living_area_interaction` `0.870268`
- `Gr Liv Area_capped` `0.721902`
- `Garage Cars` `0.662413`
- `Garage Area` `0.652544`
- `total_bathrooms` `0.645520`
- `Total Bsmt SF` `0.634097`
- `1st Flr SF` `0.622991`
- `avg_quality` `0.602766`
- `house_age` `-0.576052`
- `Year Built` `0.575599`
- `Full Bath` `0.555079`
- `remodel_age` `-0.552184`
- `Year Remod/Add` `0.550198`
- `Mas Vnr Area_capped` `0.506096`
- `TotRms AbvGrd` `0.497449`
- `Fireplaces` `0.481518`
- `BsmtFin SF 1` `0.430806`
- `total_bsmt_finished` `0.412349`

Features with medium correlation with the target variable were:

- `total_porch_area` `0.400291`
- `Lot Area_capped` `0.367188`
- `Open Porch SF_capped` `0.349387`
- `Lot Frontage` `0.340024`
- `Wood Deck SF_capped` `0.337733`

- **Removing Overlapping Features**

To reduce redundancy and improve clarity, overlapping features were reviewed and one from each pair was retained:

  - **Garage Area** was kept instead of `Garage Cars` as it provides a continuous measure of garage size.
  - **house_age** was kept instead of `Year Built` for a more intuitive view of property age.
  - **remodel_age** was kept instead of `Year Remod/Add` to maintain consistency with other age-based features.

This helped streamline analysis, clarify insights and reduce the risk of multicollinearity in future modelling.

- **Domain Relevance**:

The following features may not show a strong individual correlation with `SalePrice_capped`, but they represent important buyer considerations in the real estate market. As such, they may interact meaningfully with other variables and add valuable contextual insight to the exploration.

| **Feature** | **Reason for Inclusion** |
|---|---|
| Bedroom AbvGr | Commonly considered by buyers; affects liveability perception. |
| Kitchen AbvGr | Number of kitchens could reflect multi-family or large homes |
| TotRms AbvGrd | Overall room count impacts functionality and space. |
| MS SubClass | Proxy for building type (e.g., 1-story vs. 2-story), important for valuation. |
| Yr Sold | Useful to assess temporal market trends or inflation impact. |

- **Distribution Diversity**

In addition to correlation and domain knowledge, features were selected for analysis based on their statistical distribution. Features showing interesting patterns, such as strong skewness, long tails, or multimodal distributions, were retained as they may signal unique subgroups or outlier behaviour that could be informative. For example:

- **Mas Vnr Area** and **Lot Area** exhibit strong right skew due to a large number of properties having minimal veneer and overall lot area and a small set of homes with significant stone/brickwork and lot area, which may signal luxury elements.
- **Fireplaces** and **TotRms AbvGrd** show natural clustering into discrete groups (e.g., 1, 2, or 3+ fireplaces), potentially aligning with property types or buyer preferences.
- **Open Porch SF** and **Wood Deck SF** also show heavy-tailed distributions, capturing variation in outdoor space amenities.

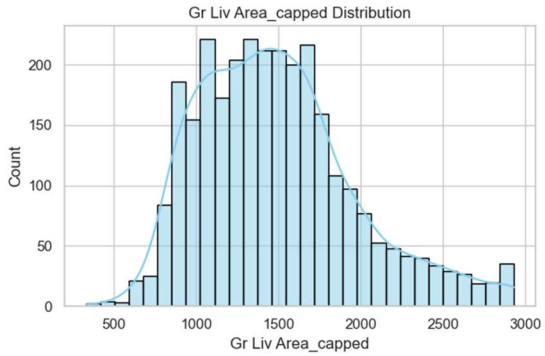
By retaining these features, the analysis remains sensitive to **non-normal, real-world patterns** that could otherwise be smoothed over if only symmetric or linear distributions were considered.

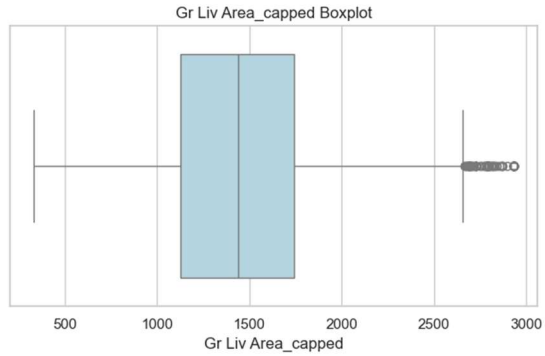
The final features selected for univariate exploration were:

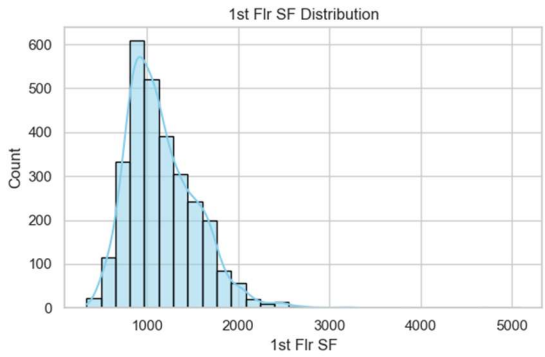
- Gr Liv Area_capped
- 1st Flr SF
- Total Bsmt SF
- Mas Vnr Area_capped
- Lot Area_capped
- Wood Deck SF_capped
- total_bsmt_finished
- qual_living_area_interaction
- Full Bath
- total_bathrooms
- TotRms AbvGrd
- Bedroom AbvGr
- Kitchen AbvGr
- avg_quality
- house_age
- remodel_age
- Yr Sold
- Garage Area
- Open Porch SF_capped
- total_porch_area
- MS SubClass
- Fireplaces

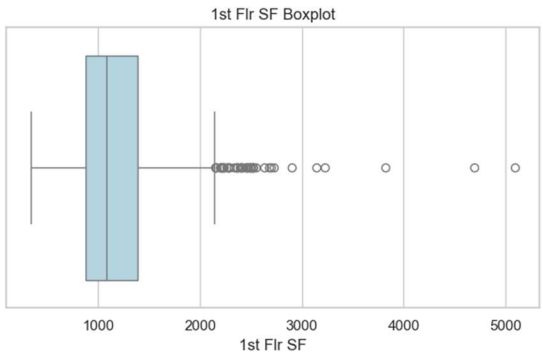
Each feature is grouped logically below, visualised using histograms and KDE and box plots to show the distribution, skewness and outliers

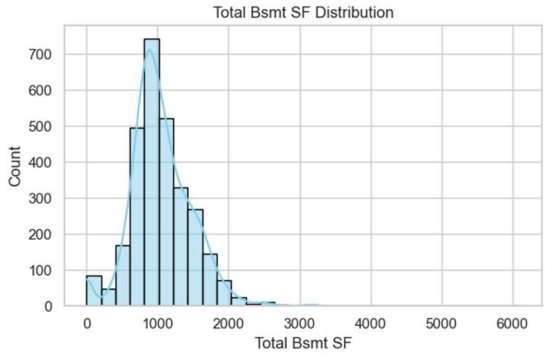
- **Size and Area Features**

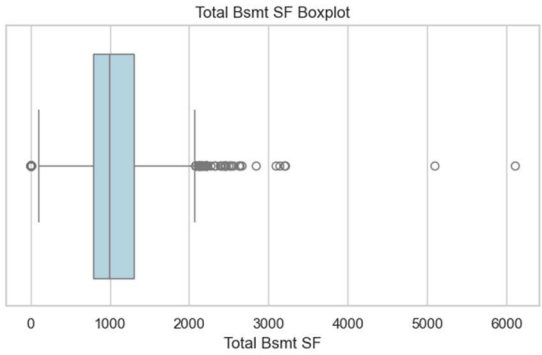


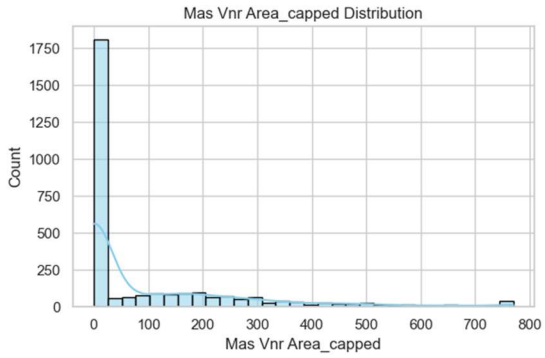



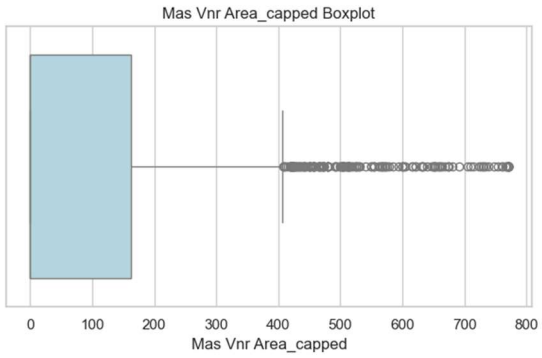



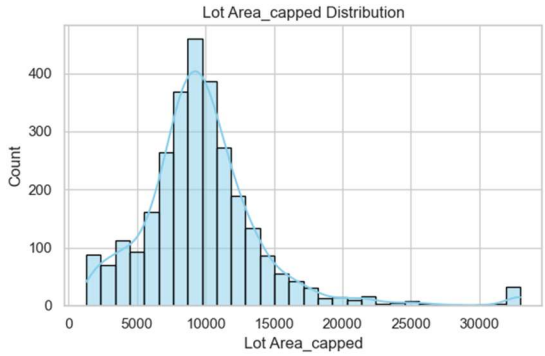



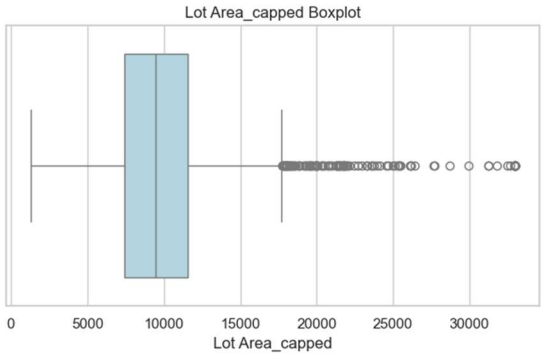



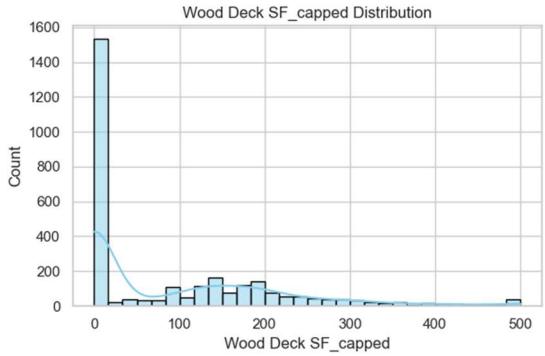



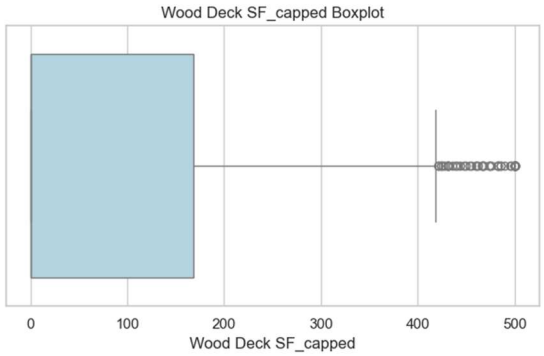

Mas Vnr Area_capped Distribution
Count
Mas Vnr Area_capped

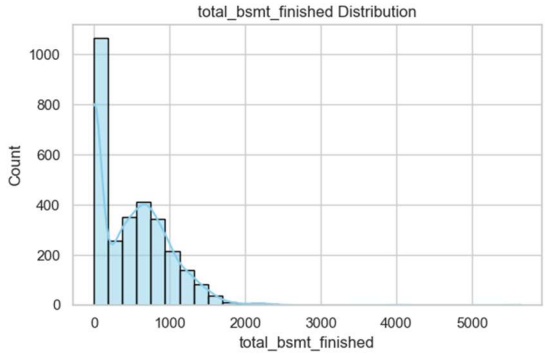

Mas Vnr Area_capped Boxplot
Mas Vnr Area_capped

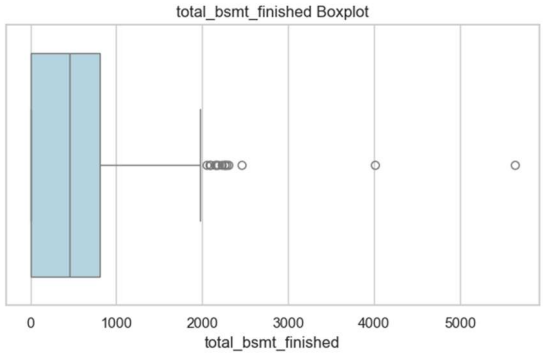

Lot Area_capped Distribution
Count
Lot Area_capped

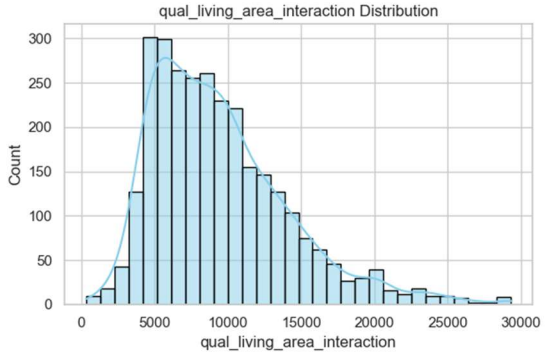

Lot Area_capped Boxplot
Lot Area_capped

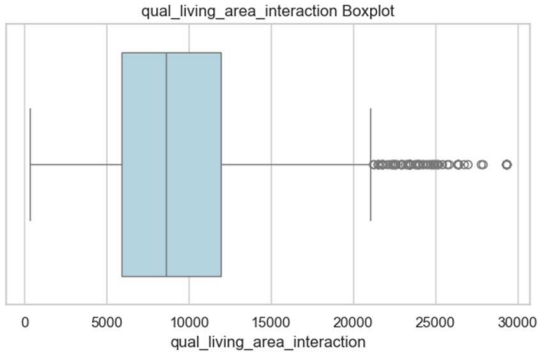

Wood Deck SF_capped Distribution
Count
Wood Deck SF_capped


Wood Deck SF_capped Boxplot
Wood Deck SF_capped

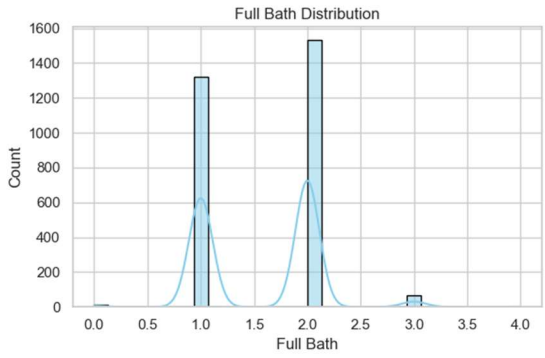

total_bsmt_finished Distribution
Count
total_bsmt_finished

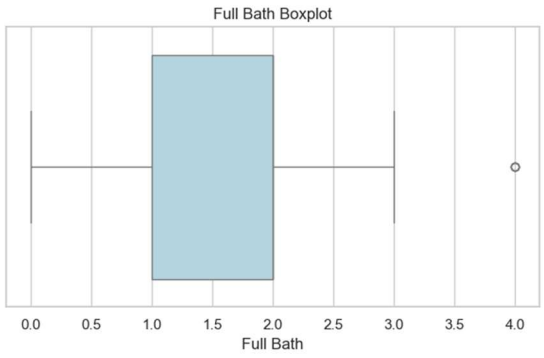

total_bsmt_finished Boxplot
total_bsmt_finished

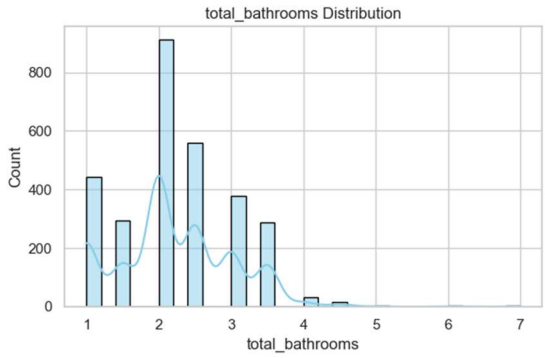

qual_living_area_interaction Distribution
Count
qual_living_area_interaction

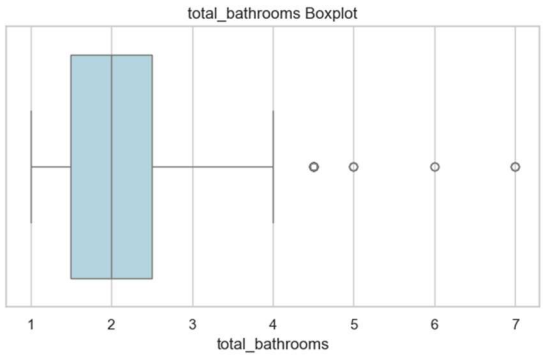

qual_living_area_interaction Boxplot
qual_living_area_interaction

- **Bathrooms and Plumping**

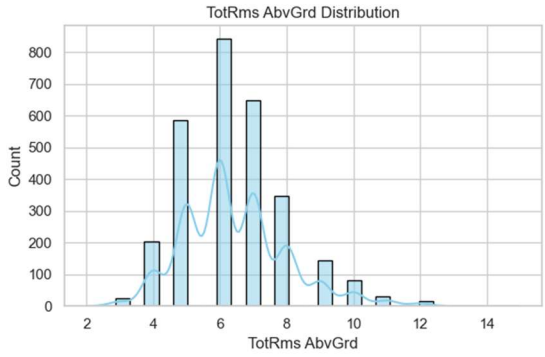


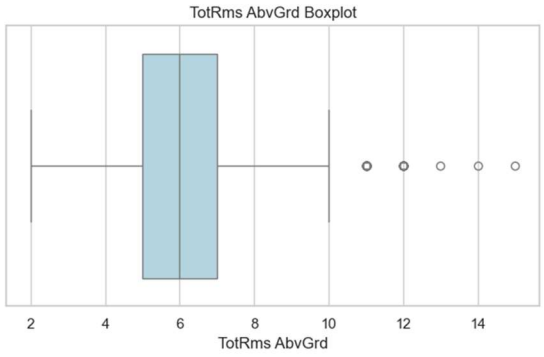



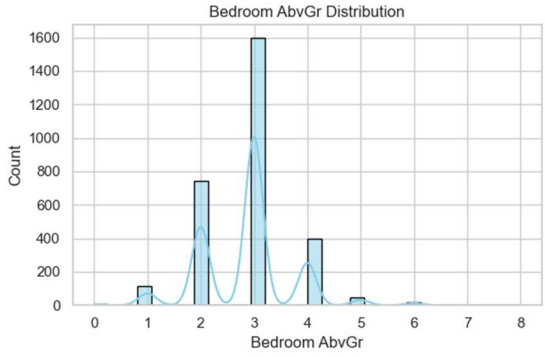



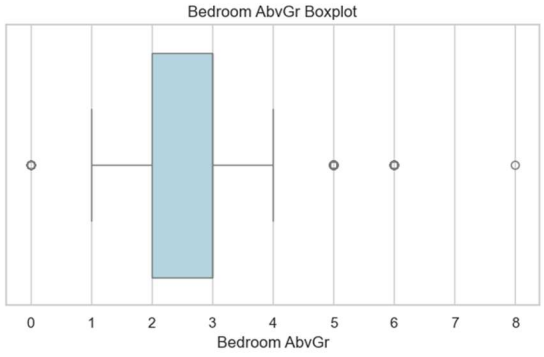



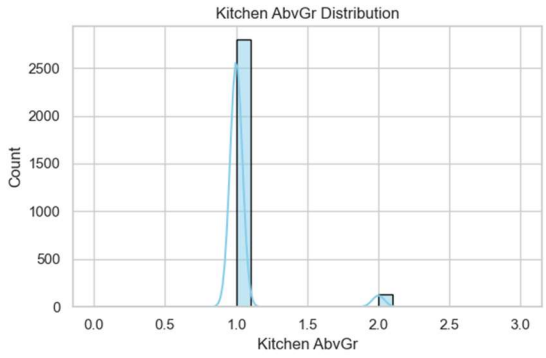

- **Rooms and Interior Count**

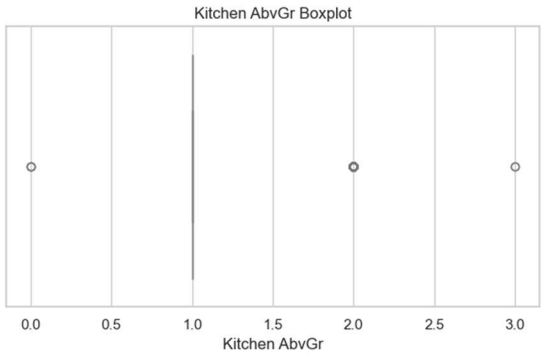


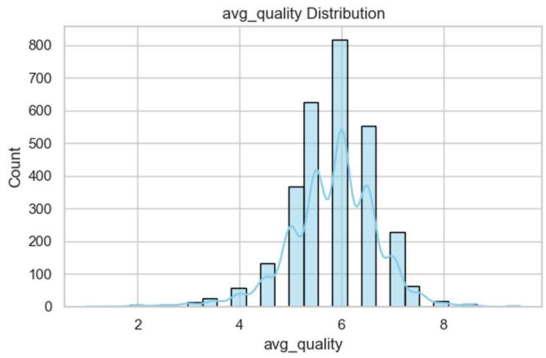



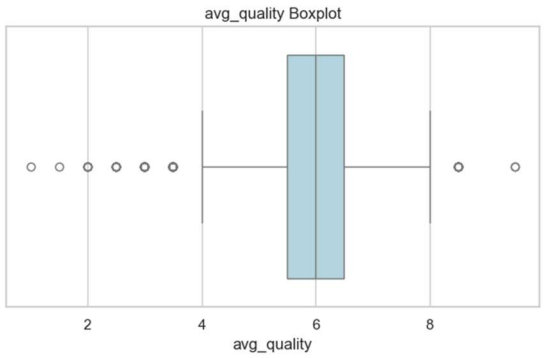



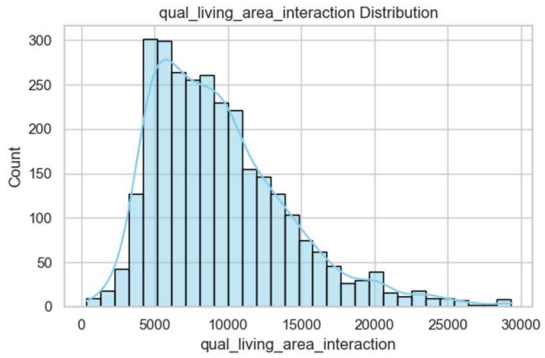


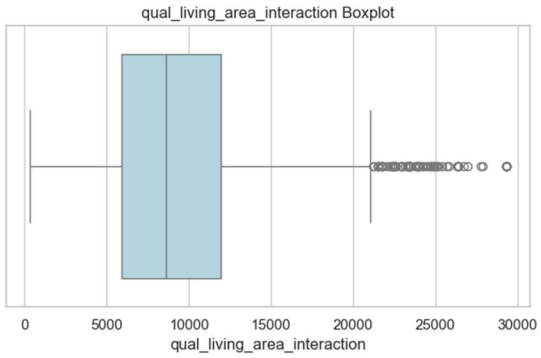




- **Quality / Composite Scores**

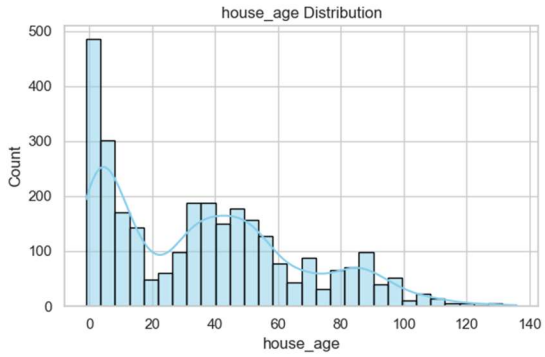


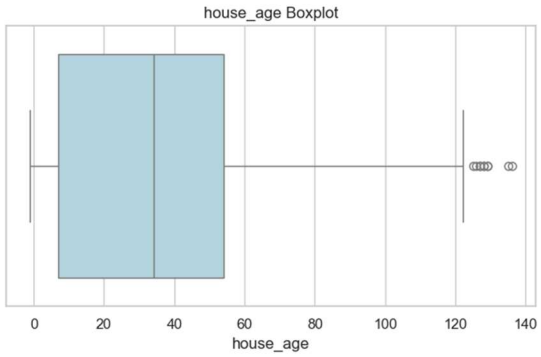



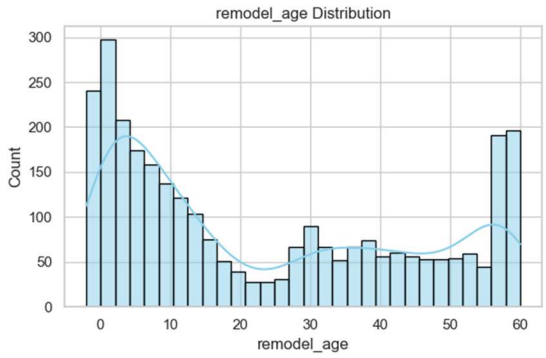



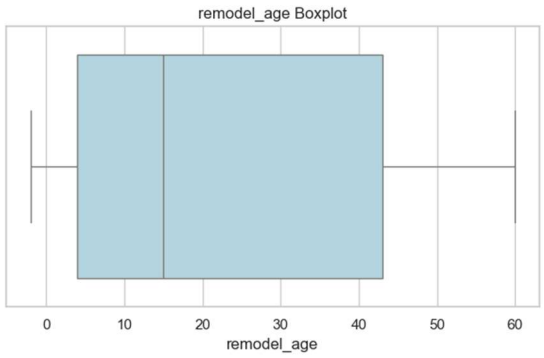



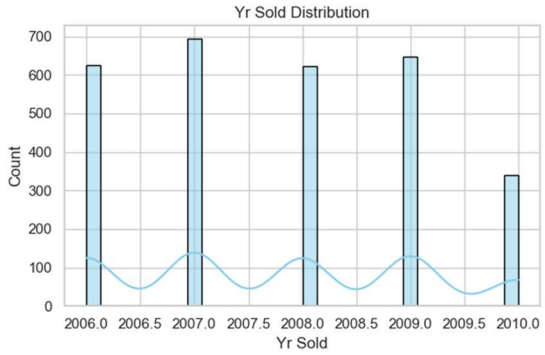

- **Age / Time-Based Features**

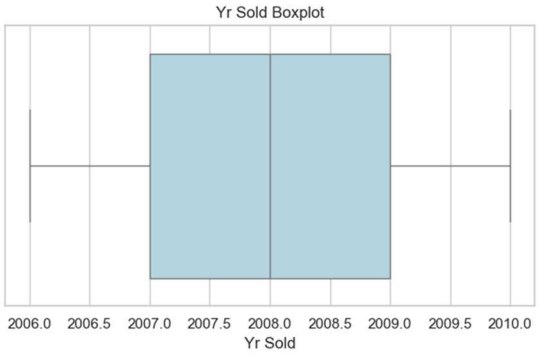


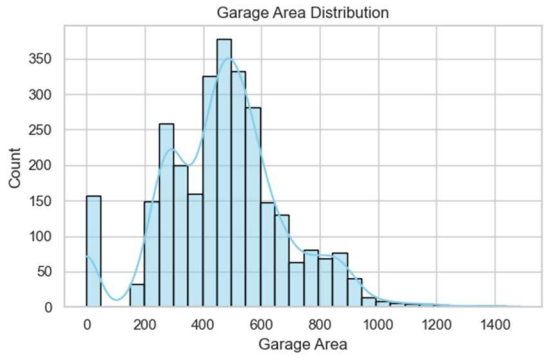


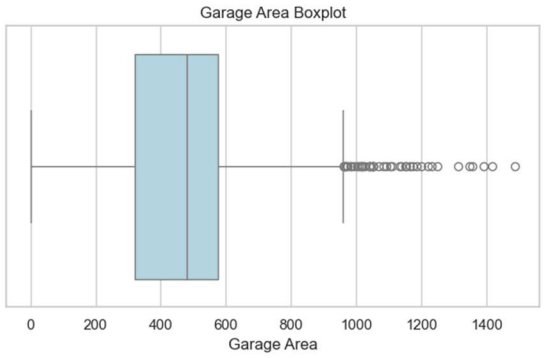



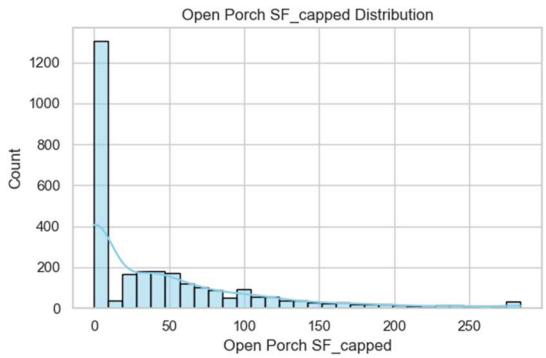



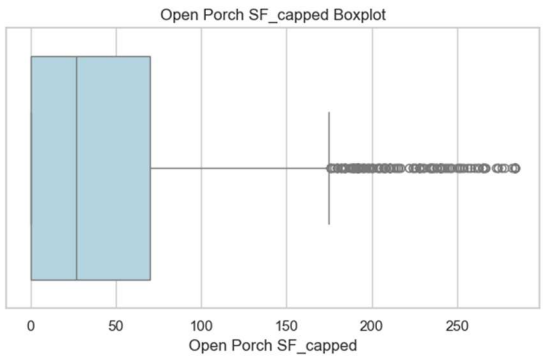



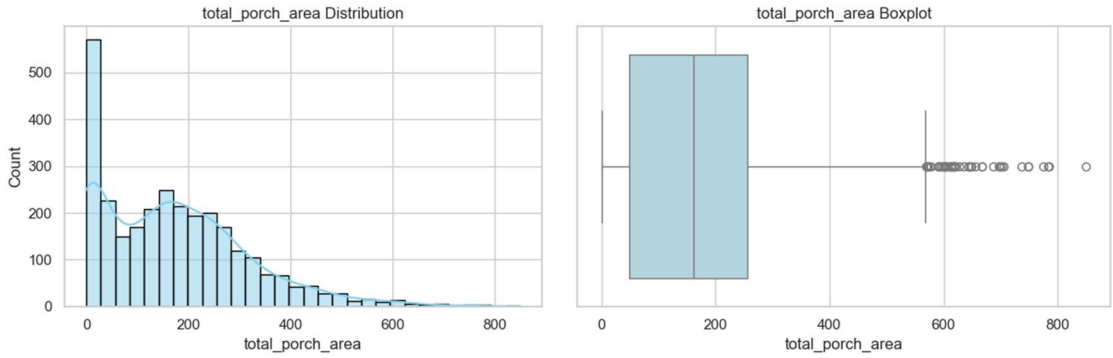

- **Garage, Porch and External Features**



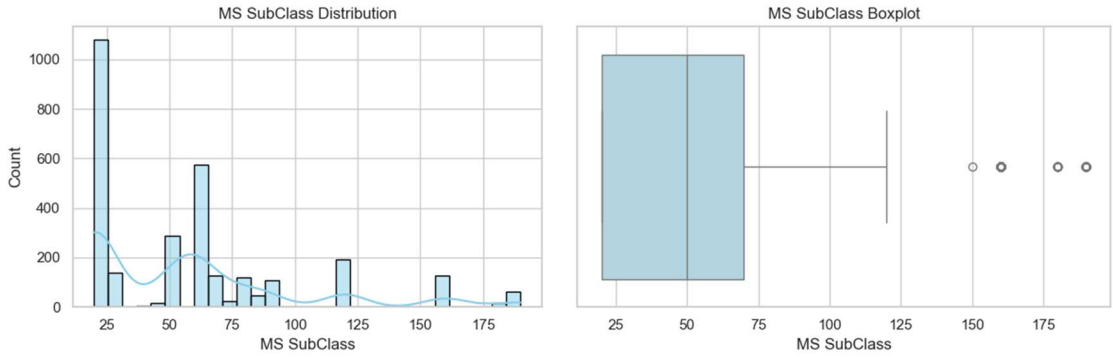



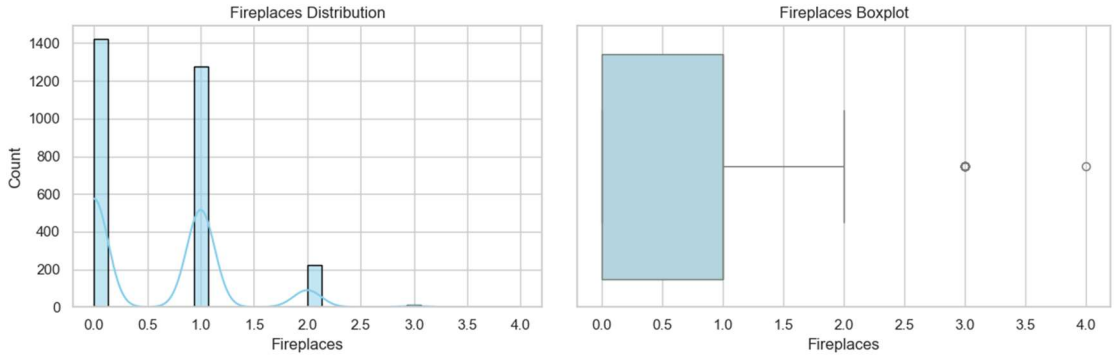






- **Structural / Other**




- **Size and area-related features** such as `Gr Liv Area_capped`, `Lot Area_capped`, `Total Bsmt SF`, `1st Flr SF`, and `total_bsmt_finished` show strong right-skewness, with a wide range and numerous high-end outliers.
- **Both bathroom and plumbing-related features** display multimodal distributions with only a limited number of distinct values. `Total_bathrooms`, however provides the most informative distribution, showing moderate right skew and high-end outliers.
- **Room count features** such as `TotRms AbvGrd` and `Bedroom AbvGr` are closer to normal distributions with only slight right skew and a few high-value outliers.
- **avg_quality** is the only feature that appears to have a slight left skew, indicating more homes with above-average quality.
- **Time-based features** show varying patterns:
  - `Yr Sold` is relatively uniform across years but shows a drop in 2010.
  - `remodel_age` exhibits a **U-shaped distribution**, suggesting that many houses were either never remodelled or remodelled recently.
  - `house_age` is right-skewed, with most properties being newer but some much older outliers.
- **Garage, porch, and exterior features** are all heavily right-skewed with numerous zero values and high-end outliers.

- Many features such as `Mas Vnr Area_capped`, `Wood Deck SF_capped`, `total_bsmt_finished`, `Half Bath`, `Bsmt Full Bath`, `house_age`, `Open Porch SF_capped`, `total_porch_area`, and `Fireplaces` exhibit **zero-inflated distributions**, where a substantial portion of the data is concentrated at zero.
- Features like `Gr Liv Area_capped`, `Lot Area_capped`, and `qual_living_area_interaction` show **wide ranges**, reflecting diverse property sizes, while features such as Full Bath, and Kitchen AbvGr are more **tightly clustered**, with fewer distinct values.

**Categorical Features**

Given the large number of categorical features in the dataset, plotting all of them would result in an overwhelming amount of visual information and reduce interpretability. Therefore, a curated subset of features was selected for analysis. The selection process was:

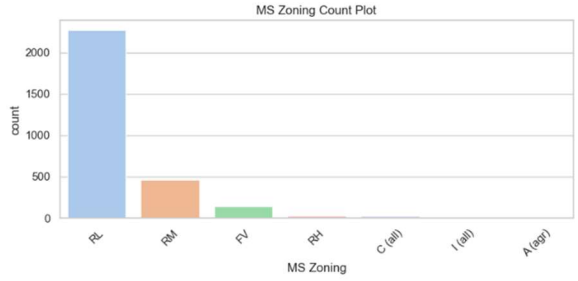
- Features with little to no variation across categories were dropped.
- The remaining features were ranked by significance using ANOVA p-values.
- The top 15 features were selected.
- An additional 5 domain-relevant features were added back for contextual importance.

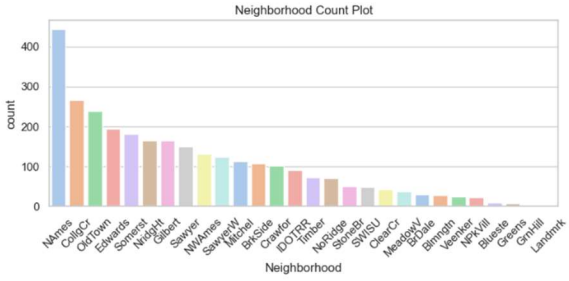
The final set of categorical features analysed is:

- `Neighborhood`
- `MS SubClass`
- `Exter Qual`
- `Bsmt Qual`
- `Kitchen Qual`
- `Garage Finish`
- `Fireplace Qu`
- `Foundation`
- `Garage Type`
- `BsmtFin Type 1`
- `Heating QC`
- `Bsmt Exposure`
- `Exterior 1st`
- `Exterior 2nd`
- `Overall Cond`
- `MS Zoning`
- `House Style`
- `Condition 1`
- `Roof Style`
- `Paved Drive`

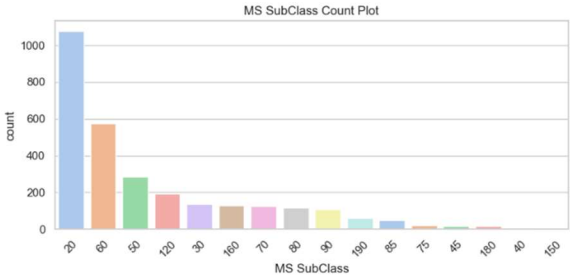
Each feature is grouped logically below, visualised using count plots to assess the distribution of categories within each feature.

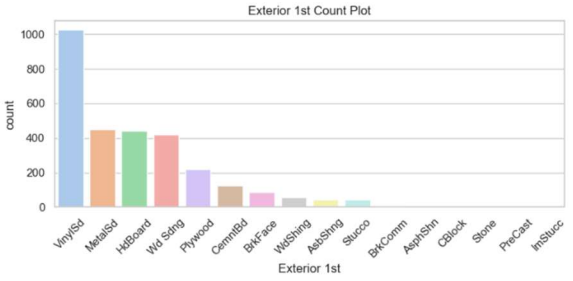
- **Location and Zoning**

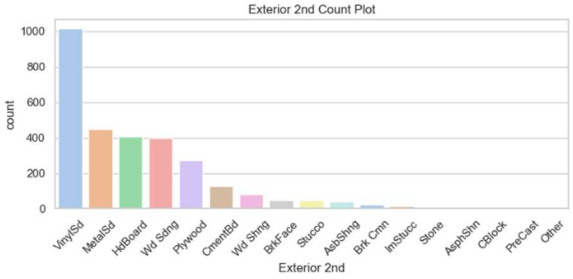


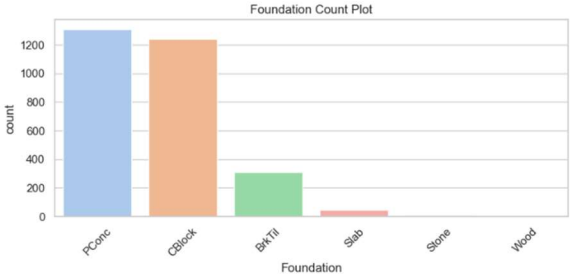



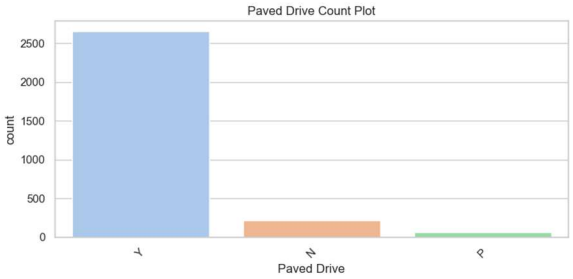
- **Exterior and Structural**

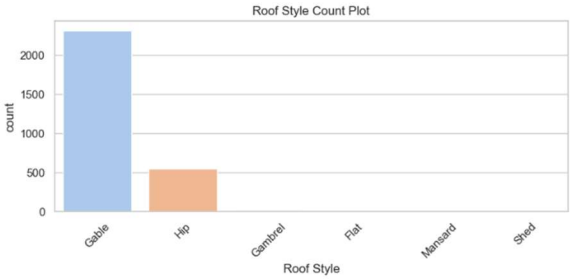


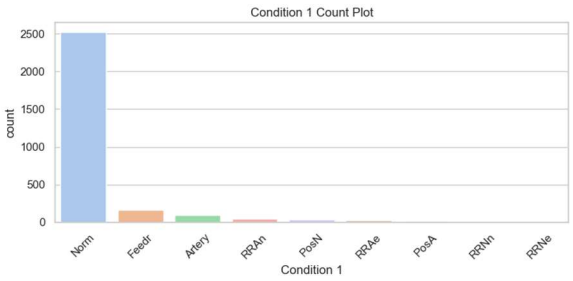



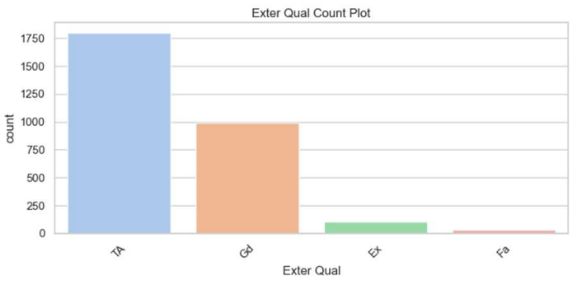



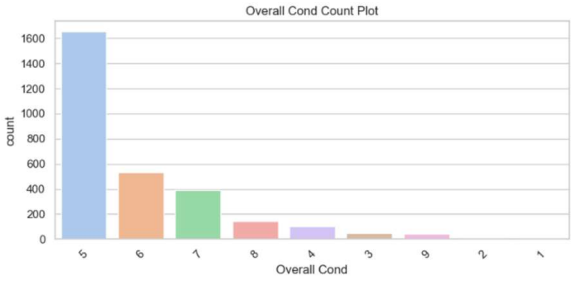



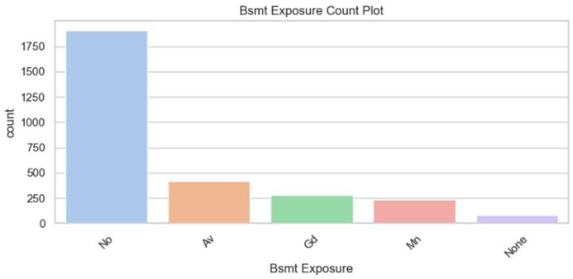

- **Quality and Condition Ratings**

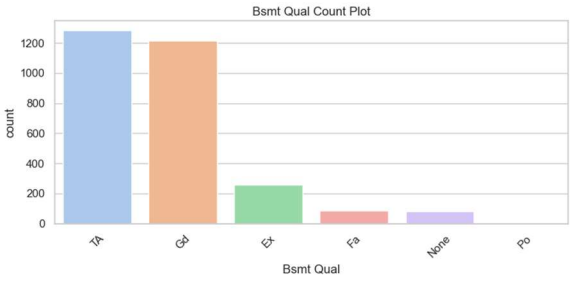


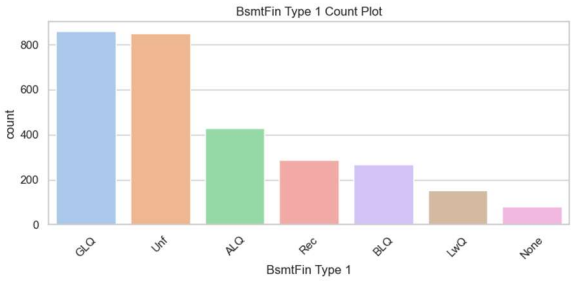



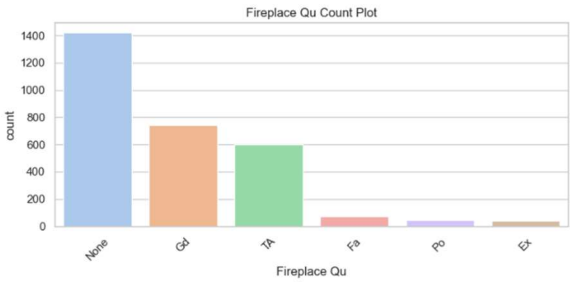
- **Basement Features**

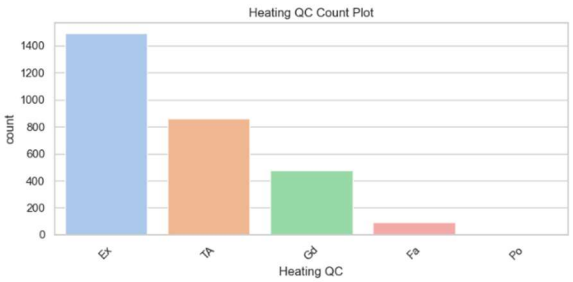


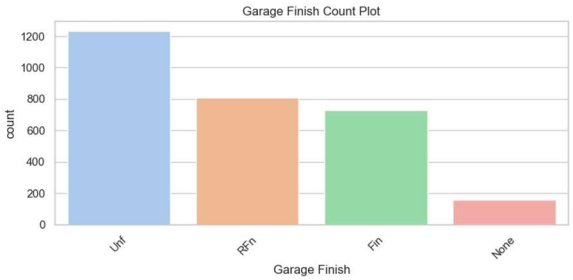



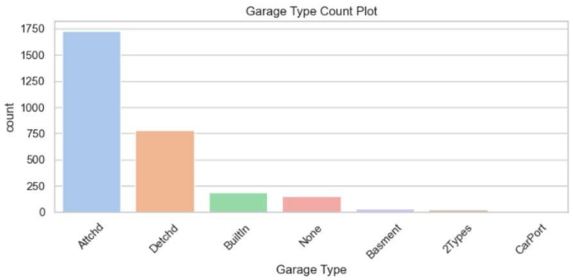

- **Heating and Fireplace**



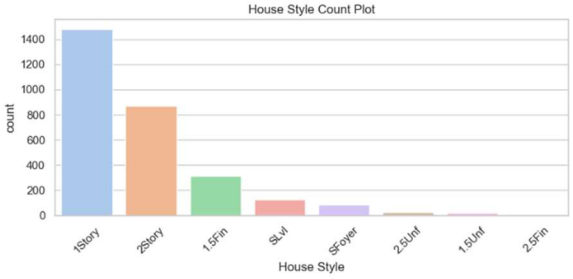




- **Garage Features**




- **Interior Quality and Style**



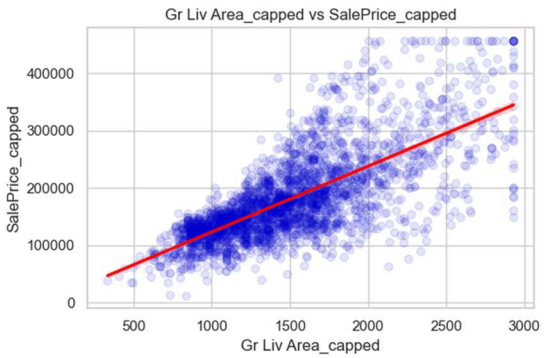

- Nominal categories like Neighborhood, Exterior 1st and Exterior 2nd, MS Subclass, and House Style exhibit clear, stepped differences. These well-separated differences between categories may be important when later considering their impact on the target variable.
- Ordinal categories such as Exter Qual, Overall Cond, Fireplace Qu, Heating QC, and Bsmt Fin Type 1 show more uneven category separations. These differences might still be important when considering their impact on the target, as the categories represent meaningful ordered levels of quality or condition.
- Some features display dominant categories for example MS Zoning mainly shows *RL,* Sale Condition is dominated by *Normal*, Overall Condition mostly has the value *5*, Paved Drive is primarily *Y* (Yes), Bsmt Exposure is largely *No,* Exterior Condition mainly shows *TA* (Typical/Average), Exterior 1st and Exterior 2nd are mostly *VinylSd* (Vinyl Siding), Condition 1 is predominantly *Norm* (Normal) and Roof Style is mainly *Gable*.
- Some features have missing or rare categories, indicating limited observations for certain classes:
    - Sale Condition, Garage Type, Condition 1, MS Zoning, Exterior 1st, Exterior 2nd, Foundation, and Roof Style each have some categories with very few data points or missing entries.
    - Since Exterior 1st and Exterior 2nd are highly similar in category distribution and material types, one could consider dropping Exterior 2nd to reduce redundancy.

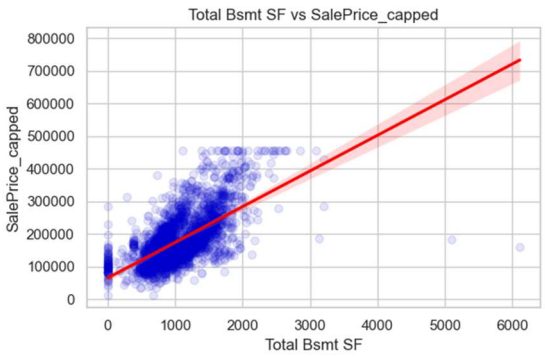
## Bivariate Analysis

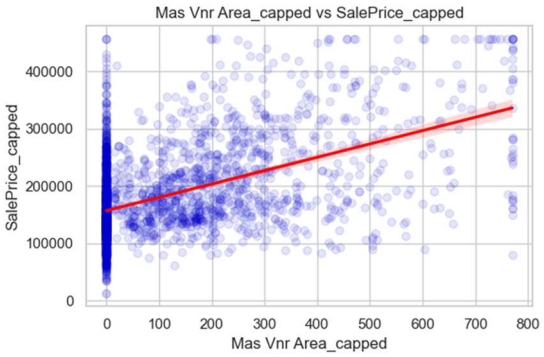
**Numerical Versus Target**

The following regplots show the scatter points with a best fit line for all selected key features against SalePrice_capped. As some features are ordinal in nature these are displayed as boxplots:

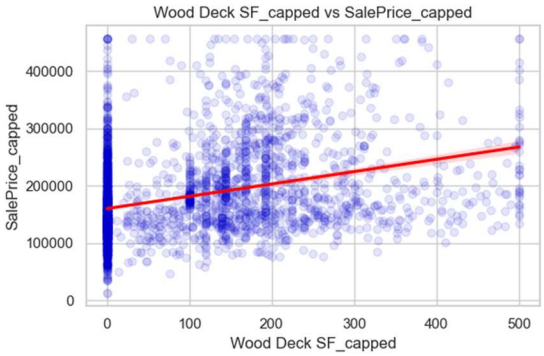
- **Size and Area Features**

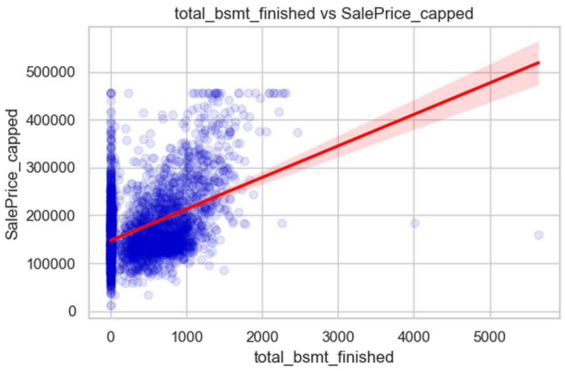


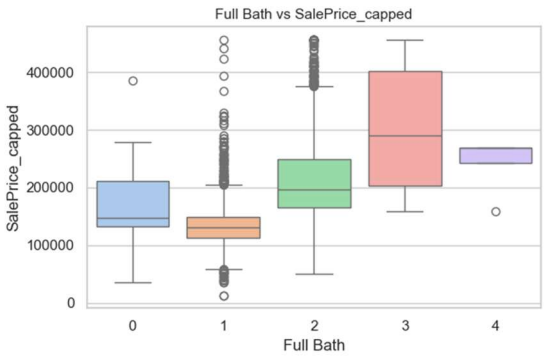


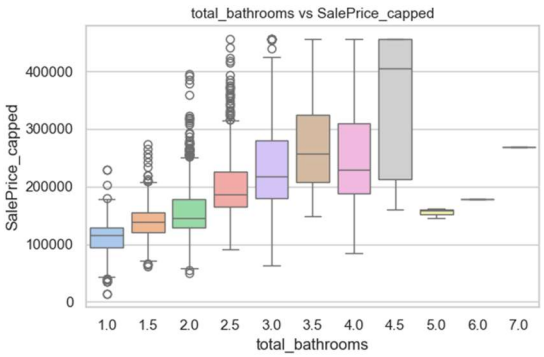


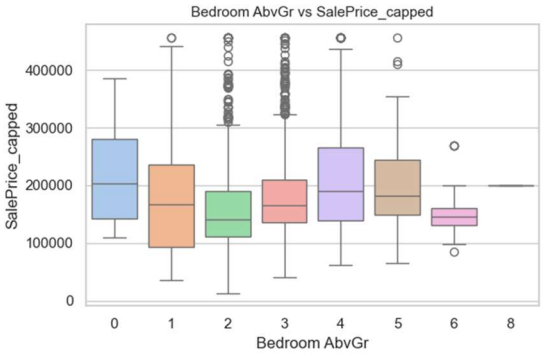


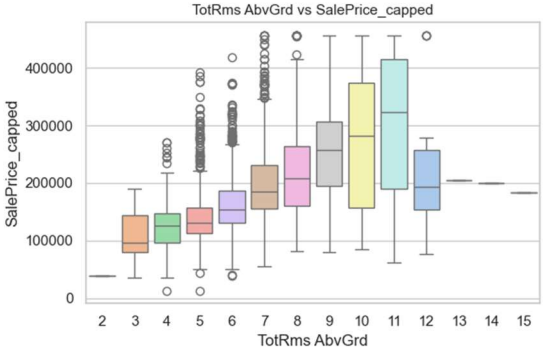
- **Bathrooms and Plumping**

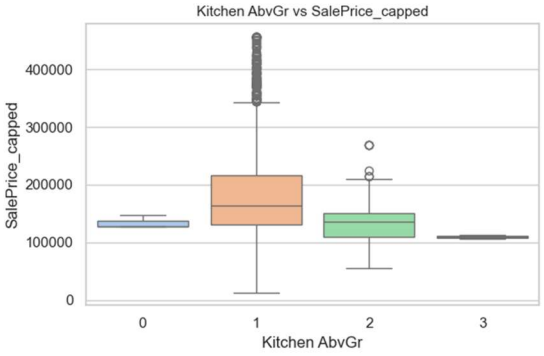


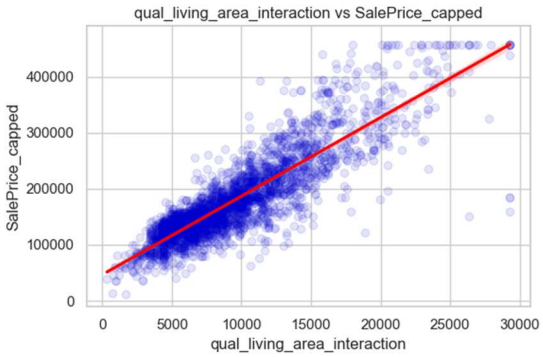
- **Rooms and Interior Count**

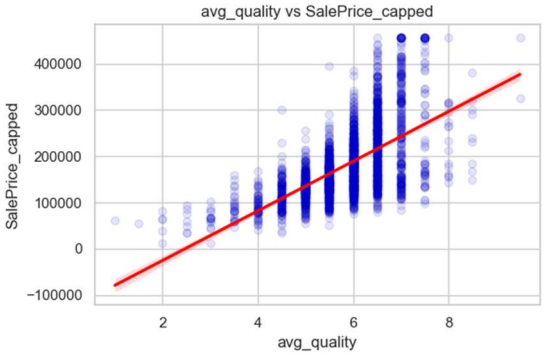



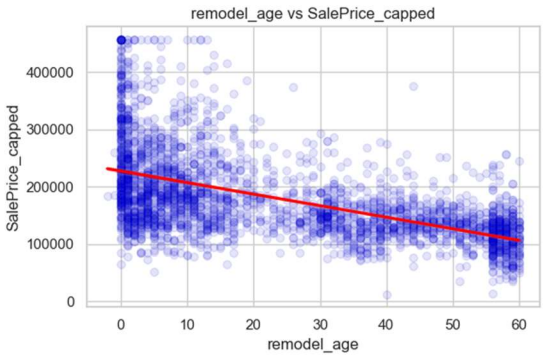



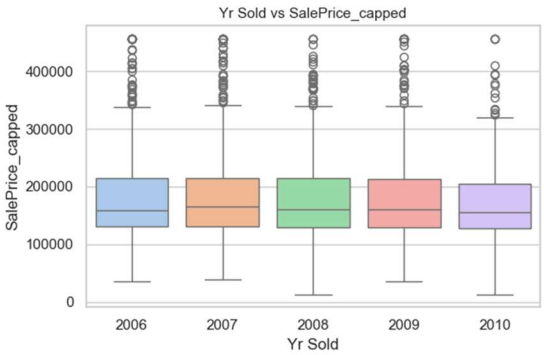

- **Quality / Composite Scores**



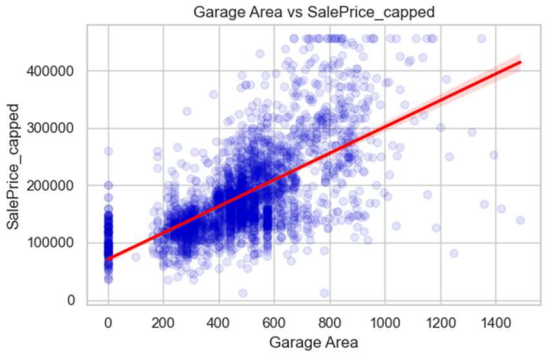



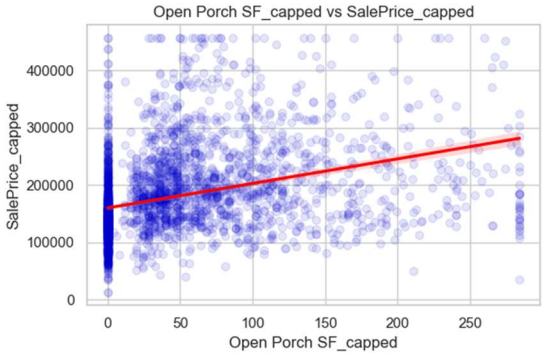

- **Age / Time-Based Features**



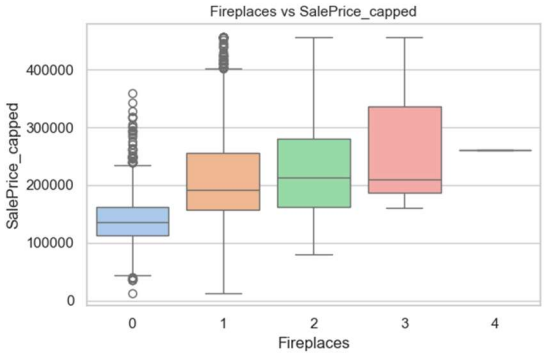
- **Garage, Porch and External Features**



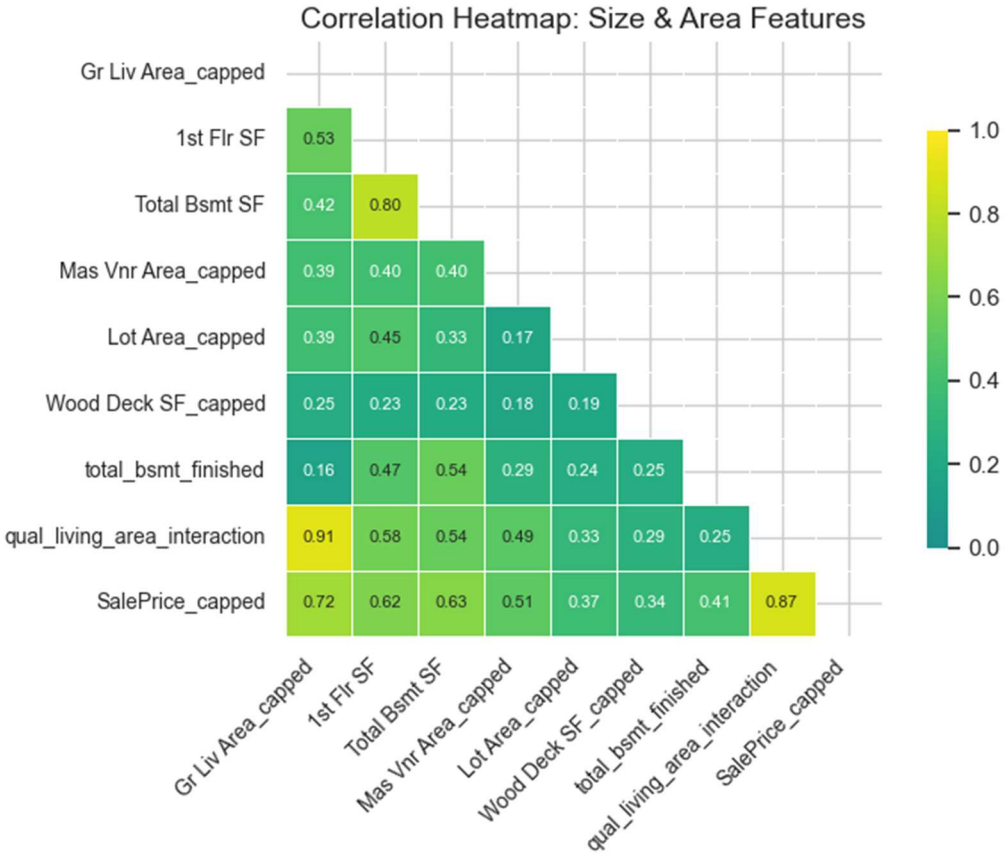



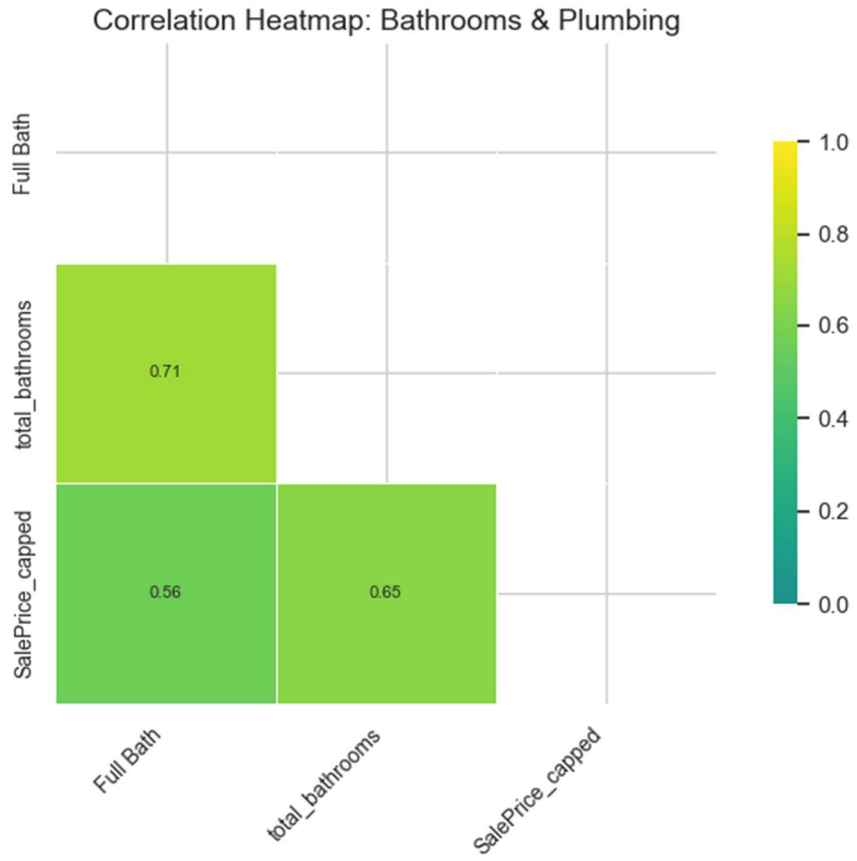


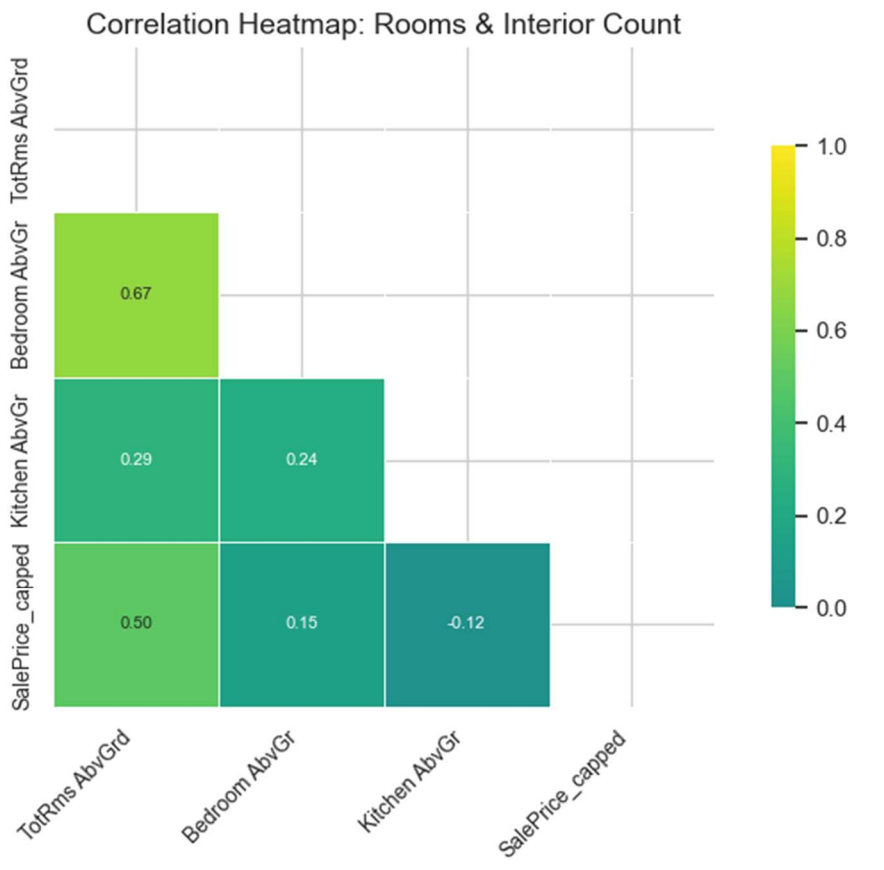
- **Structural / Other**



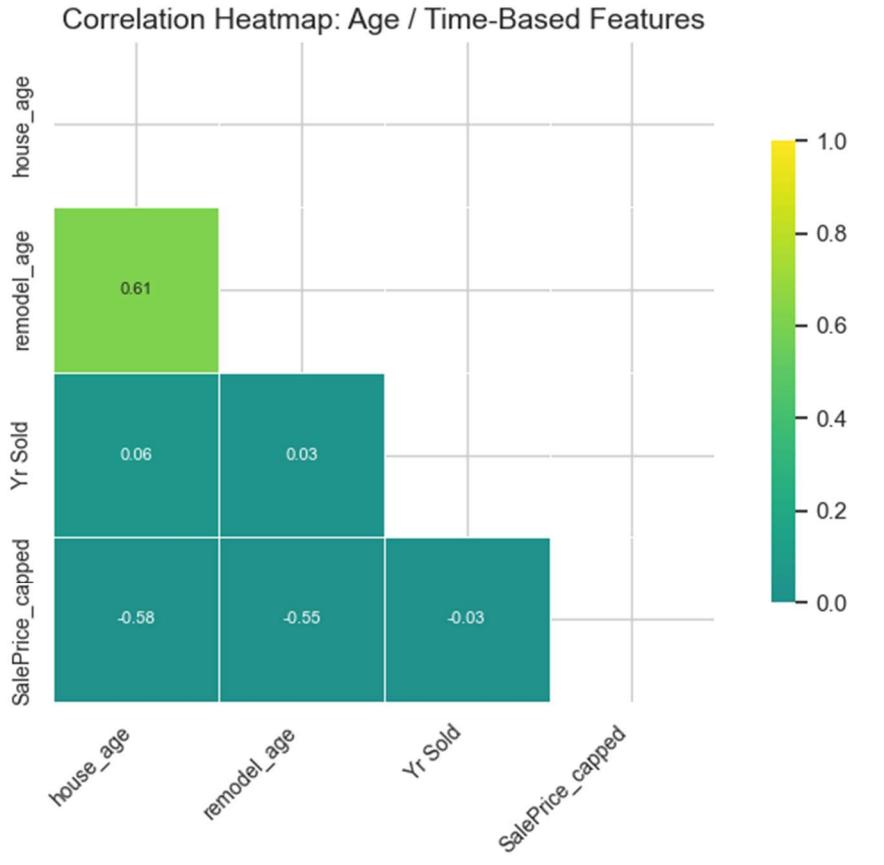
The following heatmaps show the strength of correlation between all key selected features as well as `SalePrice_capped`, grouped thematically.

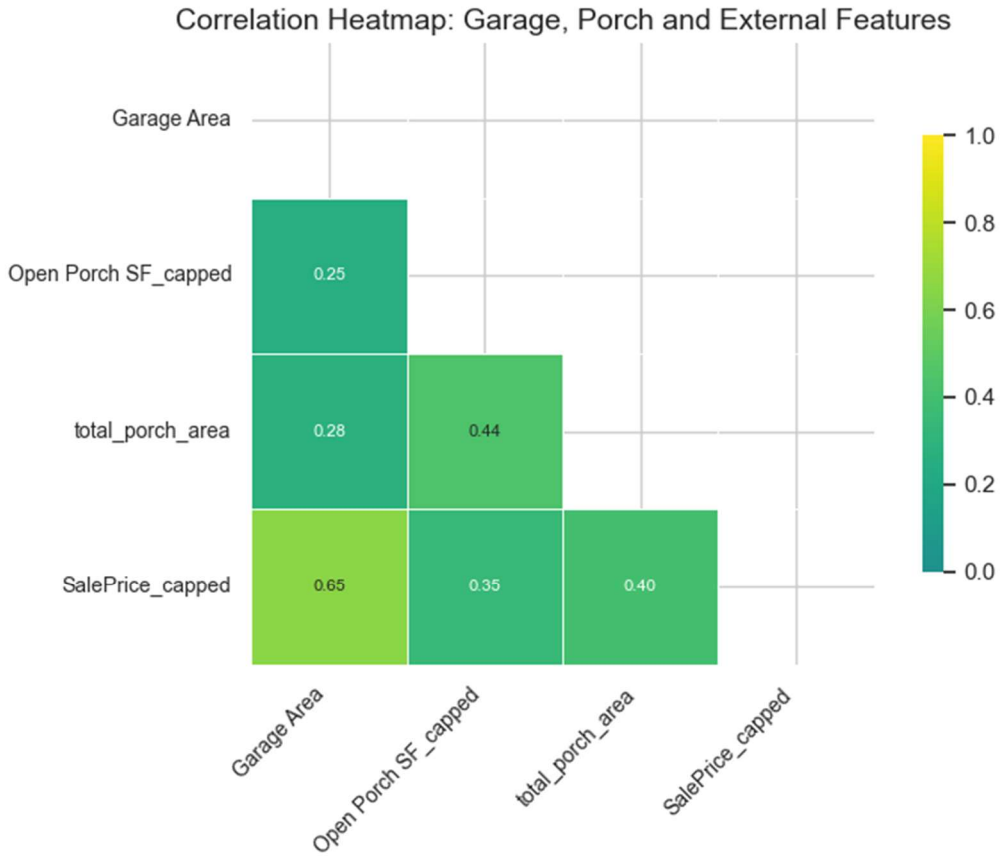


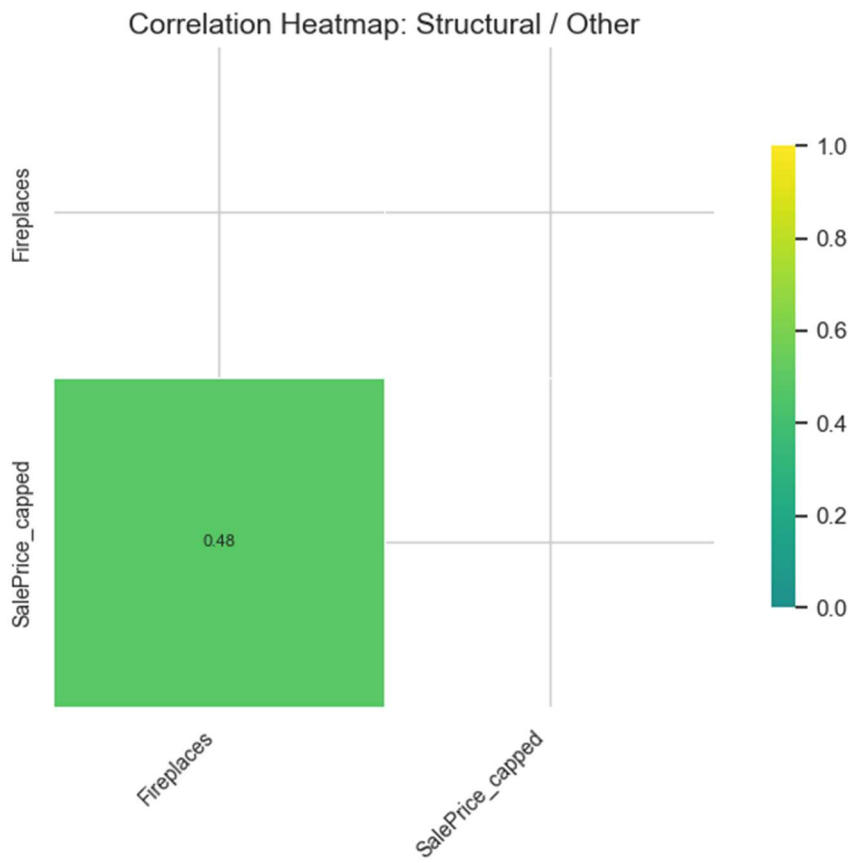




Correlation Heatmap: Rooms & Interior Count
TotRms AbvGrd
Bedroom AbvGr
Kitchen AbvGr
SalePrice_capped
0.67
0.29
0.24
0.50
0.15
-0.12
1.0
0.8
0.6
0.4
0.2
0.0
TotRms AbvGrd
Bedroom AbvGr
Kitchen AbvGr
SalePrice_capped


Correlation Heatmap: Quality / Composite Scores
avg_quality
qual_living_area_interaction
SalePrice_capped
0.60
0.60
0.87
1.0
0.8
0.6
0.4
0.2
0.0
avg_quality
qual_living_area_interaction
SalePrice_capped


Correlation Heatmap: Age / Time-Based Features
house_age
remodel_age
Yr Sold
SalePrice_capped
0.61
0.06
0.03
-0.58
-0.55
-0.03
house_age
remodel_age
Yr Sold
SalePrice_capped
1.0
0.8
0.6
0.4
0.2
0.0


Correlation Heatmap: Garage, Porch and External Features
Garage Area
Open Porch SF_capped
total_porch_area
SalePrice_capped
0.25
0.28
0.44
0.65
0.35
0.40
Garage Area
Open Porch SF_capped
total_porch_area
SalePrice_capped
1.0
0.8
0.6
0.4
0.2
0.0




From the heatmaps we can see that quality and composite scores (e.g. `avg_quality` and `qual_living_area_interaction`) show strong correlation with `SalePrice_capped`, as do `Gr Liv Area_capped` and `Garage Area`. Other features such as `1st Flr SF`, `Total Bsmt SF`, `total_bathrooms`, and `Full Bath` also show moderately strong correlation.

Surprisingly, rooms and interior counts (e.g. `TotRms AbvGrd`, `Bedroom AbvGr`, `Kitchen AbvGr`) and age/time-based features (e.g. `house_age`, `remodel_age`, `Yr Sold`) show relatively weak correlation with `SalePrice_capped`. `Fireplaces` also exhibits low correlation.

In terms of collinearity, several features appear highly correlated with each other:

- `Total Bsmt SF` with `1st Flr SF`
- `qual_living_area_interaction` with both `Gr Liv Area_capped` and `avg_quality`
- `total_bathrooms` with `Full Bath`
- `Bedroom AbvGr` with `TotRms AbvGrd`
- `remodel_age` with `house_age`

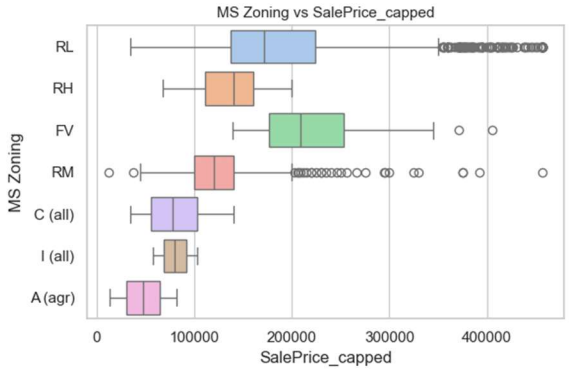
These correlations suggest some redundancy in the data, and a subset of these features could be removed during future model development.

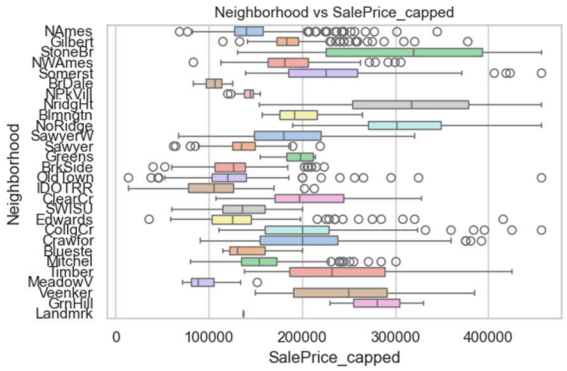
From the regplots, it's clear that `qual_living_area_interaction` has data points that closely follow the best-fit line. `Gr Liv Area_capped` follows the line well at lower square footage levels, but the relationship becomes weaker as square footage increases.

For `1st Flr SF`, `Total Bsmt SF`, and `total_bsmt_finished`, the data points are concentrated at lower square footage values, with relatively few observations in the higher range.

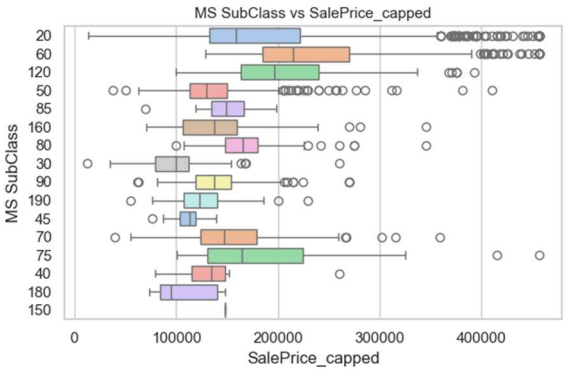
## Categorical versus Target

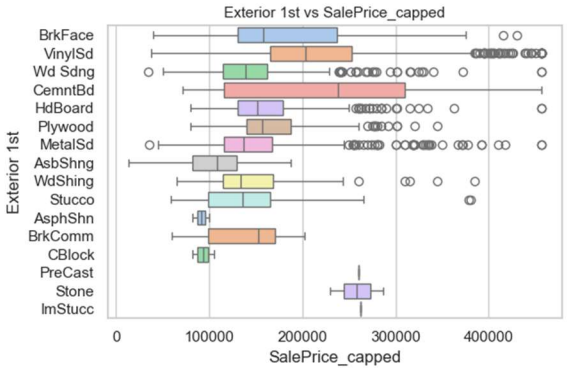
- **Location and Zoning**

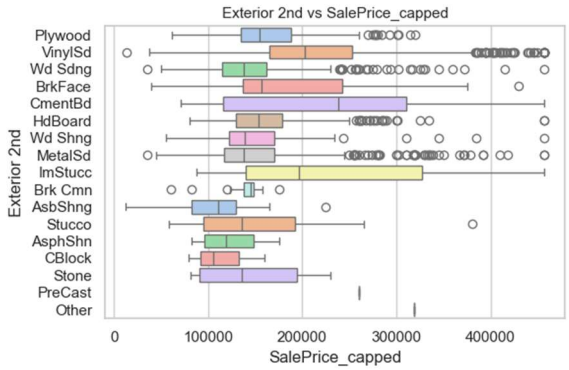


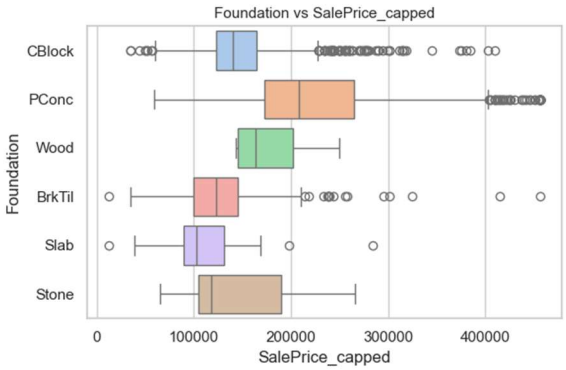



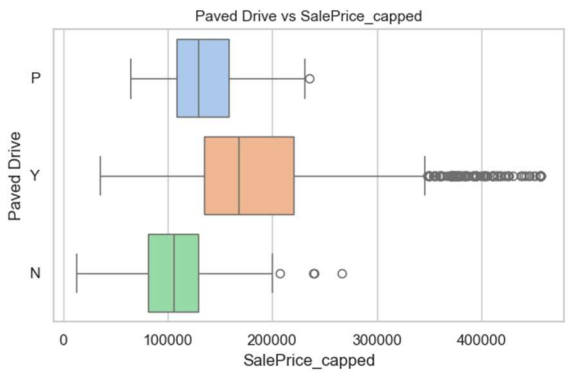

- **Exterior and Structural**

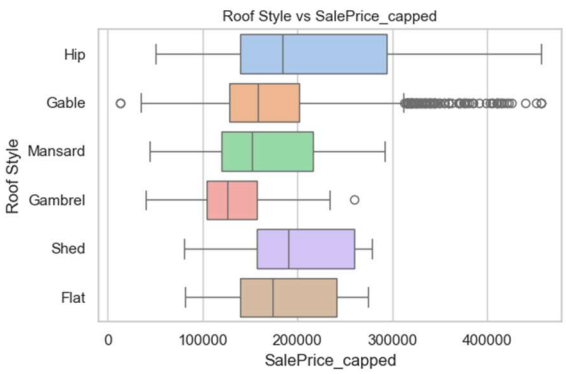


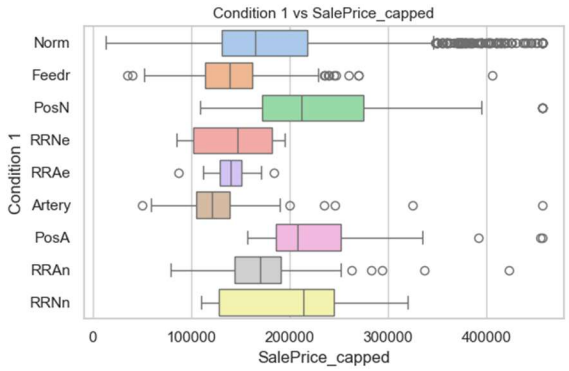



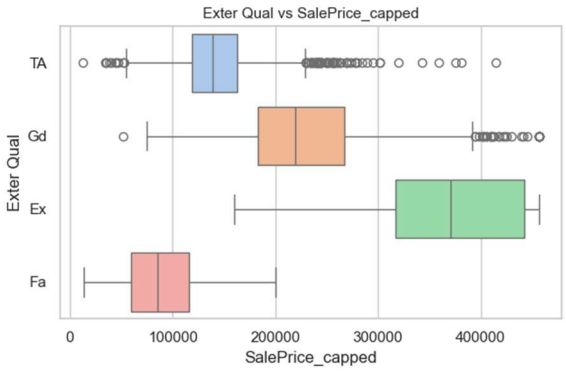



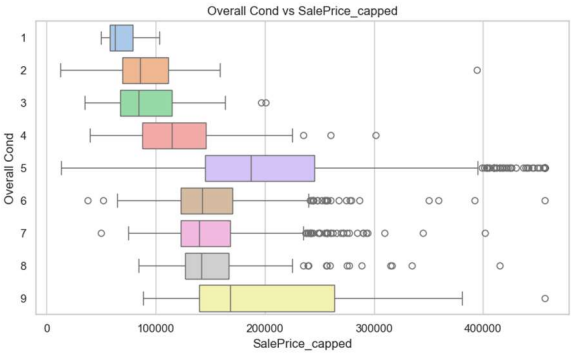


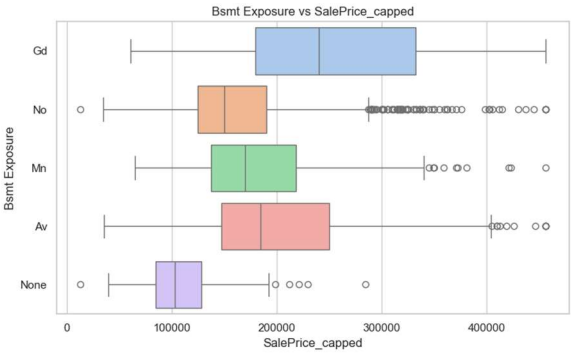

- **Quality and Condition Ratings**

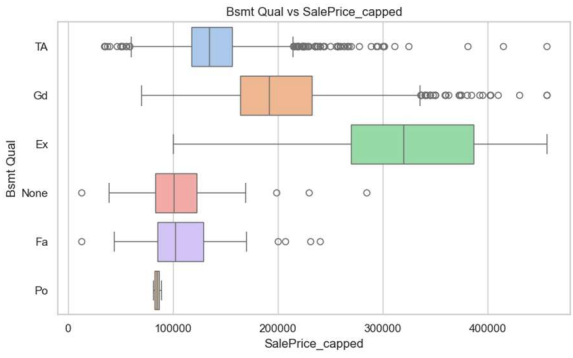


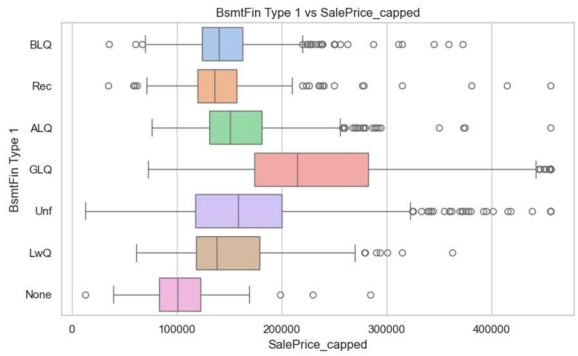



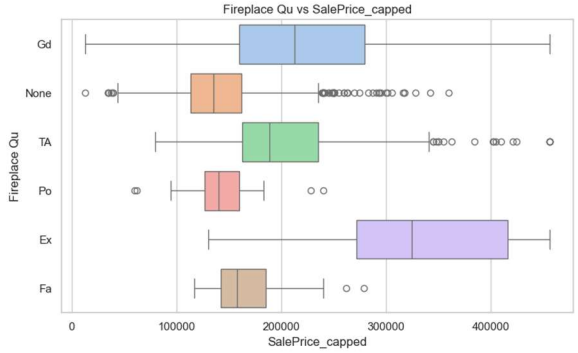

- **Basement Features**

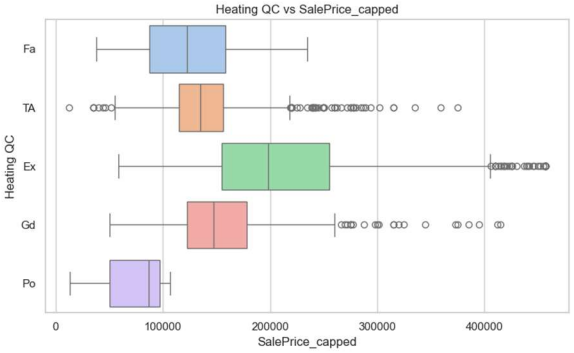


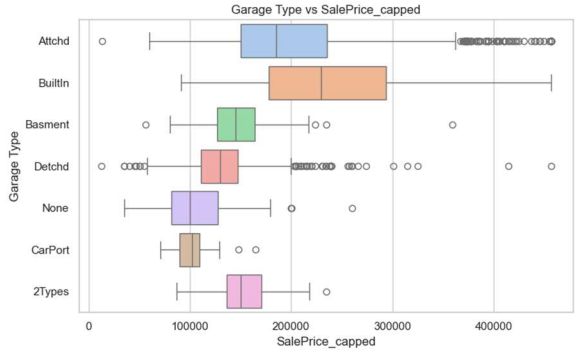


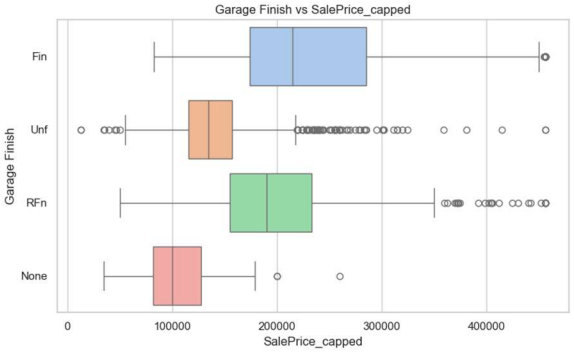

- **Heating and Fireplace**



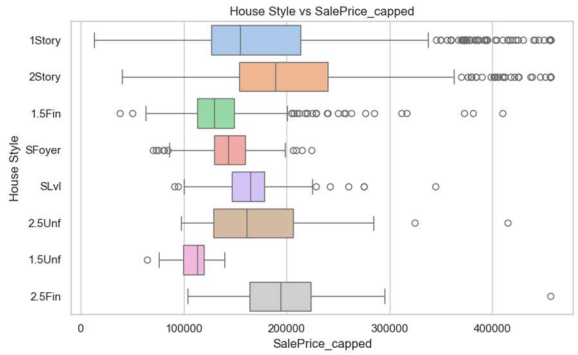



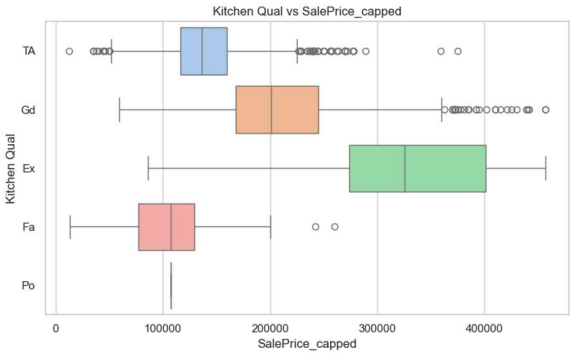

- **Garage Features**




- **Interior Quality and Style**

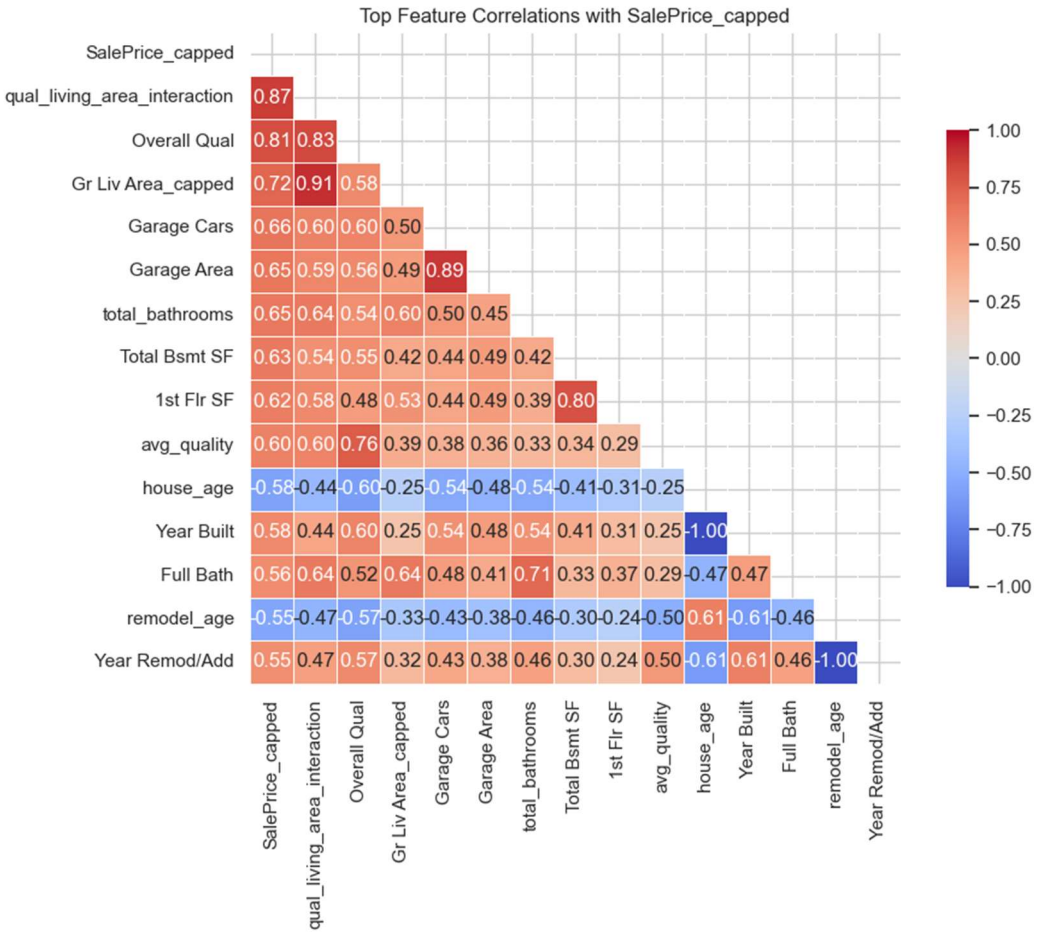


- Clear relationships and strong predictors:
    - **Neighborhood**: Despite some overlap, categories like *NoRidge* and *StoneBr* tend to have distinctly higher median sale prices, showing it's an important location factor.
    - **Overall Quality & Kitchen Quality**: Clear ordinal trends with sale price can be seen, where higher quality categories correspond to higher median prices.
    - **Garage Type and Garage Finish**: Show meaningful variation in sale price, with better-finished garages associated with higher prices.
    - **MS Zoning**: While RL and RM overlap a lot, there is a clear difference between residential and commercial/industrial categories.
    - **House Style:** Categories like *1Story* and *2Story* often show different price medians, indicating impact on value.
    - **Basement Finish Type:** *GLQ* (Good Living Quarters) and other higher-quality basement finishes correspond with higher sale prices.
- **IQR Overlap:** Features such as `Neighborhood`, `MS SubClass`, `Exterior 1st`, `Exterior 2nd`, `Roof Style`, `Condition 1`, `Overall Cond`, `BsmtFin Type 1`, and `House Style` show considerable overlap in interquartile ranges, suggesting they may be weaker discriminators of sale price.
- **Outlier Patterns**: High-value outliers often cluster in specific categories, reflecting internal variability:
    - `MS Zoning`: Outliers mostly in *RL* and *RM*.
    - `Neighborhood`: Most show outliers, except a few like *StoneBr* and *NoRidge*.
    - `MS SubClass`: Categories *20*, *50*, and *60* show the most outliers.
    - `Exterior 1st` / `Exterior 2nd`: Outliers common in types like *BrkFace* and *CBlock*.
    - `Paved Drive`: '*Y*' (Yes) category has many outliers.
    - `Roof Style` / `Condition 1`: Outliers especially in *Gable* and *Norm*.
    - Quality ratings (`Exter Qual`, `Bsmt Qual`, `Kitchen Qual`, `Heating QC`): Mid-to-high quality categories tend to show wider price ranges.
    - `Garage Type` / `Garage Finish`, `House Style`: Common types show substantial spread.
- **Spread & Median Patterns**:
    - **Wide spreads**: Found in *RL/RM* (`MS Zoning`), *StoneBr/NridgHt* (`Neighborhood`), *20/60* (`MS SubClass`), *PConc* (`Foundation`), *GLQ* (`BsmtFin Type 1`), *Ex* (`Exter Qual` & `Heating QC`), and *1Story/2Story* (`House Style`).
    - **Tight spreads**: Seen in rarer categories like *Landmrk* (`Neighborhood`), *150* (`MS SubClass`), and *Po* (`Kitchen Qual`).
    - **Similar medians**: Noted between related categories such as *Po* & *Fa* (`Kitchen Qual`), *85* & *70* (`MS SubClass`), *2Story* & *2.5Fin* (`House Style`).
    - **Distinct medians**: Most features show clearly separated medians, enhancing interpretability.
- **Ordinal Trends**: Most ordinal features display clear progression across levels. However:
    - `Kitchen Qual`: '*Po*' (Poor) and '*Fa*' (Fair) share the same median.
    - `Overall Cond`: Levels *2* & *3* and *6, 7, 8* have similar medians.
    - `Exter Qual`: The *Po* (Poor) category is absent.

## Multivariate Analysis

**Correlation matrix**




- The correlation matrix confirms expected positive relationships between house size features (like `Gr Liv Area_capped` and `TotRms AbvGrd`) and the target variable (`SalePrice_capped`).
- Several features exhibit very strong inter-correlations (e.g., `Gr Liv Area_capped` and `qual_living_area_interaction` at 0.91), suggesting overlapping information among these predictors.
- Moderate correlations among some age-related variables (`house_age` and `remodel_age`) hint at underlying temporal patterns worth further exploration.
- Features with weaker correlations to sale price (e.g., `Fireplaces`) may require more complex modelling or could be less influential in predicting price.
- The `qual_living_area_interaction` feature appears well justified, as it draws from the moderately correlated components `Gr Liv Area_capped` and `Overall_Qual` (0.58), both of which are also strongly correlated with the target (0.72 and 0.81, respectively).
- An interaction between `Garage Cars` (or `Garage Area`) and `total_bathrooms` could offer added value, as these variables show moderate correlations both with each other and with

the target. Together, they may capture a dimension of overall practicality or luxury not reflected in the individual features alone.

- Overall, the matrix provides a useful overview of variable relationships that will inform feature selection and further analysis.

**Grouped Boxplots (categorical vs numerical)**

Grouped boxplots were used to explore how key categorical variables relate to important numerical predictors — specifically those found to be highly correlated with sale price in earlier analysis. While sale price itself is not plotted here (having been covered in the bivariate section), these visualisations provide insight into how categorical features may influence or align with other predictive metrics such as living area, garage space, and basement size.

The categorical features selected were:

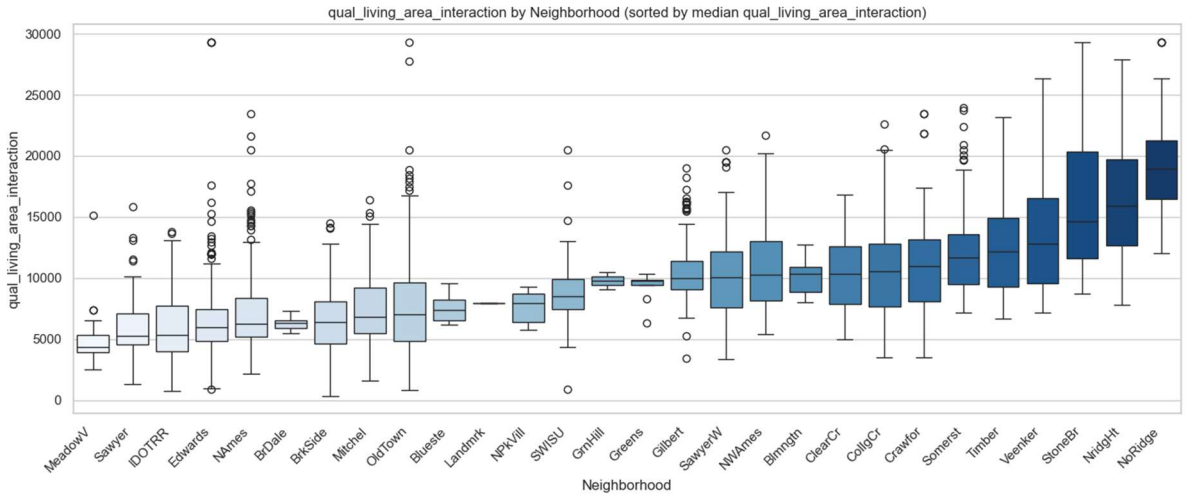
- **Neighborhood** – Most categories show clearly separated median sale prices, indicating that location is a strong driver of housing value.
- **Overall Qual** – Demonstrates a strong ordinal trend with price and other features, reinforcing its reliability as a quality proxy.
- **Garage Type** – Different garage configurations are associated with noticeably different price ranges, suggesting an impact on perceived value.
- **House Style** – Most categories show distinct medians, highlighting architectural style as an influential characteristic.
- **MS Zoning** – Despite being a nominal variable, zoning types show distinct price medians, reflecting zoning's influence on residential desirability and development constraints.

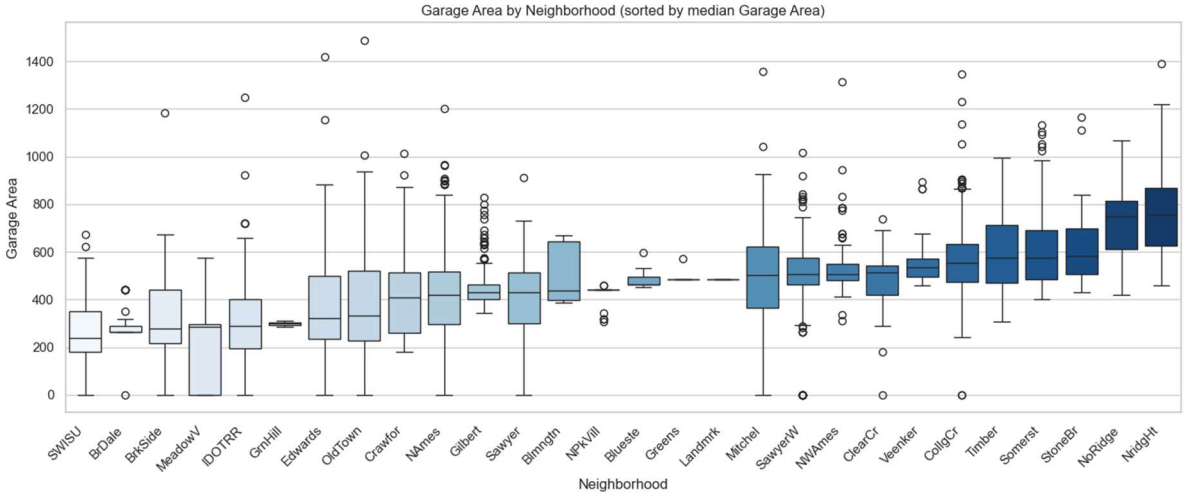
The numerical features selected were:

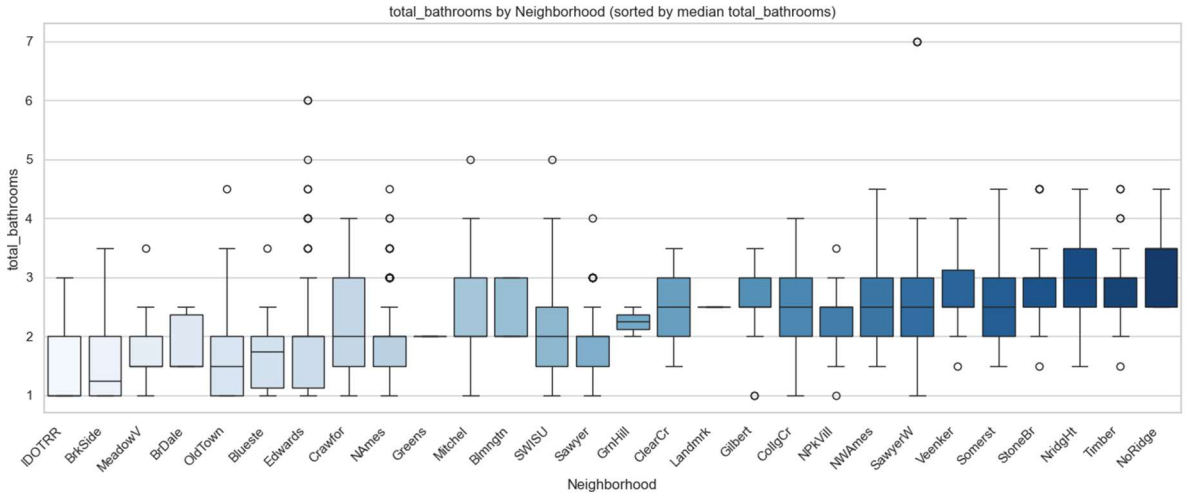
- **qual_living_area_interaction** - Strongest predictor; combines quality and size effects
- **Garage Area** - High correlation with price; reflects useful amenities
- **Total Bsmt SF** - Substantial living/functional space; relevant to house size
- **1st Flr SF** - Core structural metric; often strongly related to layout and value
- **total_bathrooms** - Practical utility feature; combines full and half baths for better insight

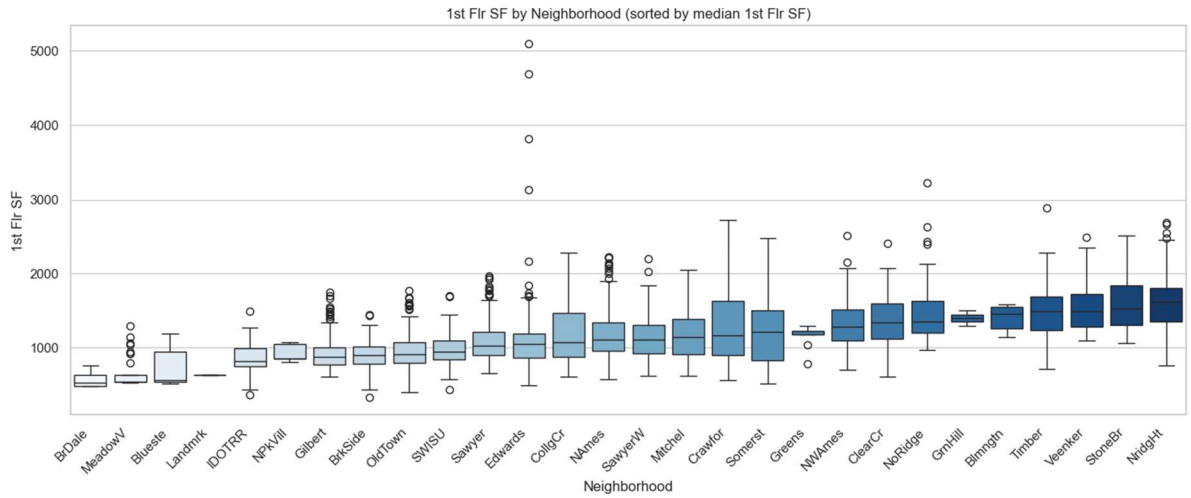
For each set of numerical features against a categorical feature the latter are ordered by median for ease of reading

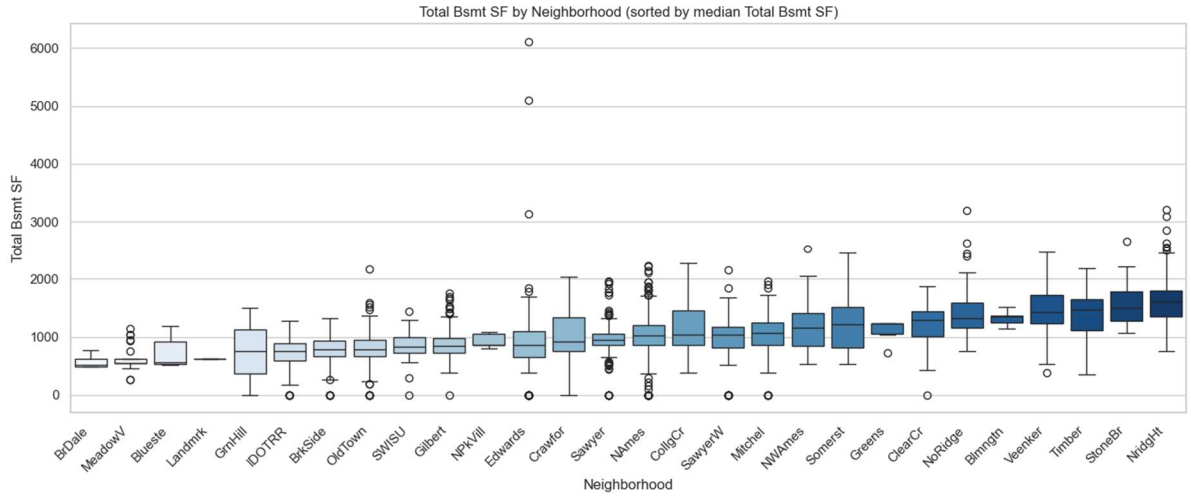
## Numerical Features by Neighbourhood



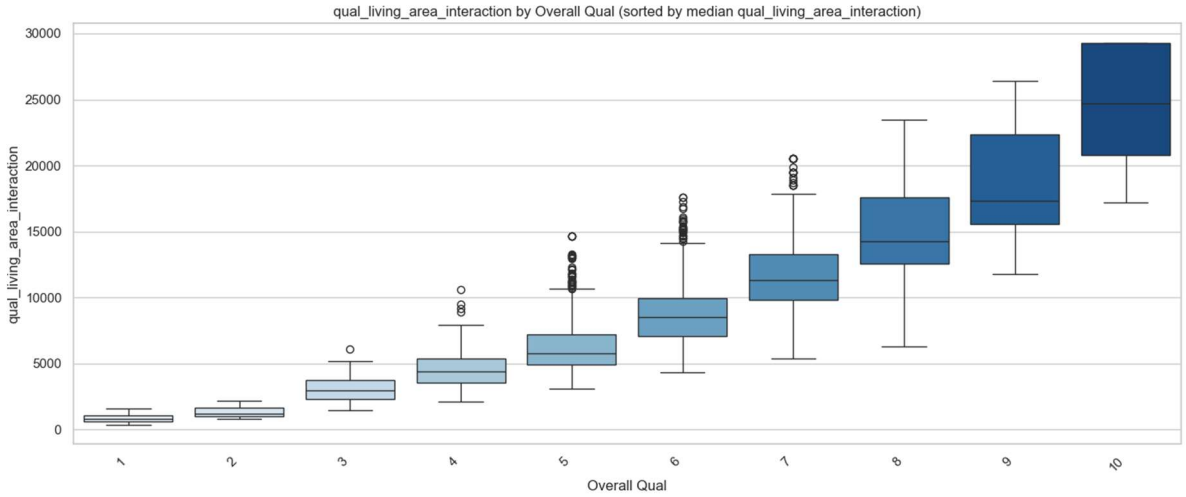


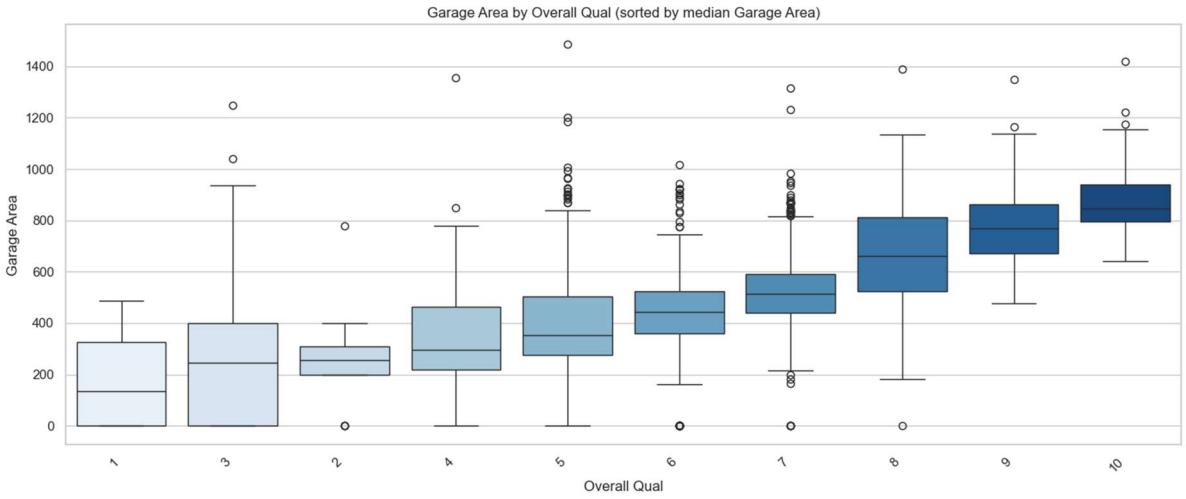



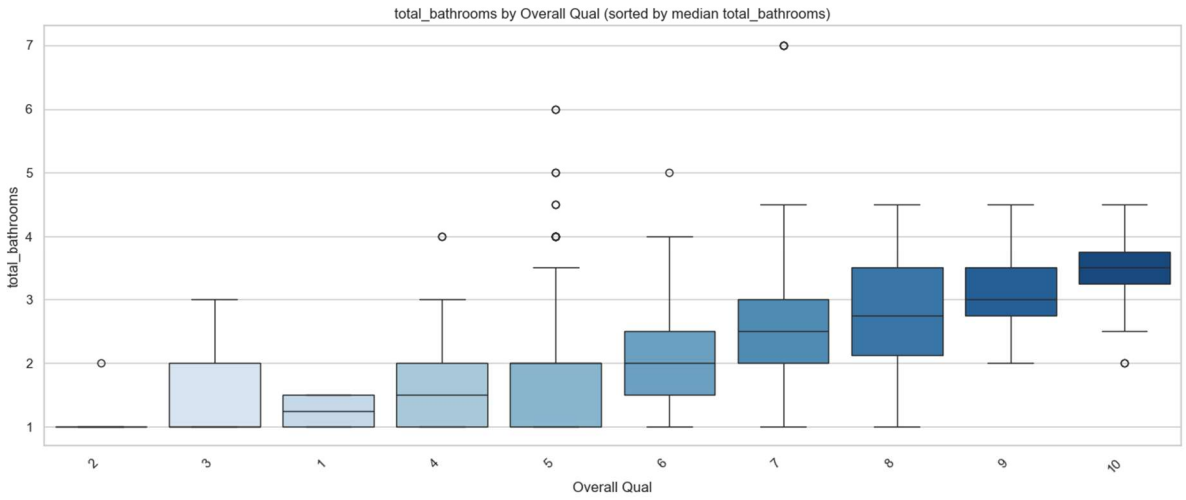

- **Median values** for most numerical features were somewhat to very homogeneous across categories. This homogeneity reduces for `Garage Area` and `qual_living_area_interaction`, where roughly 40% of the category medians are clearly distinct, offering more meaningful comparison.
- **Interquartile Range (IQR) Overlap** is common across most categorical groupings, suggesting substantial within-group variation in numerical features. This overlap becomes slightly less prominent for `total_bathrooms`, `Garage Area`, and `qual_living_area_interaction`, but remains notable overall. This implies that `Neighborhood` alone does not explain much of the variation in these numerical attributes.
- **Outlier Patterns**:
    - **Total Bsmt SF and 1st Flr SF**: Roughly half of the neighborhoods exhibit outliers, though these tend to lie just above the upper IQR limit.
    - **total_bathrooms:** Outliers appear in about a third of neighborhoods, and while fewer in number, they are more widely dispersed.
    - **qual_living_area_interaction and Garage Area:** Outliers are more common (seen in ~two-thirds of neighborhoods), more numerous, and vary from dense clusters to widely scattered points, some close to the IQR and others far above it.

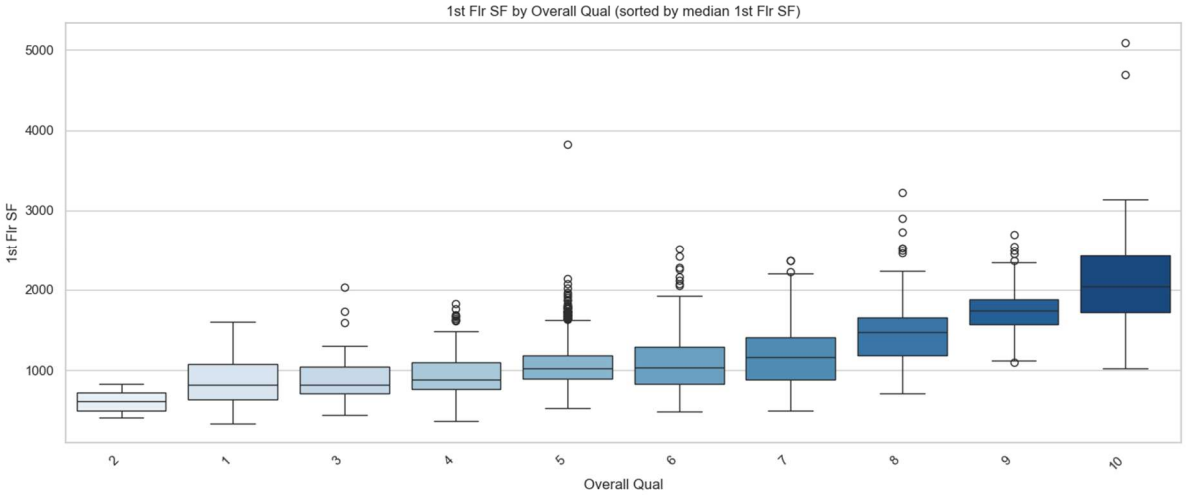
    - The **Edwards** neighborhood stands out consistently for its widely dispersed outliers across all five numerical features, though not always numerous. This suggests a particularly high internal variability in home characteristics.
- **Overall,** these plots show that while some numerical features (like `Garage Area` or `qual_living_area_interaction`) vary somewhat across neighborhoods, most exhibit substantial internal variability and overlapping interquartile ranges. In many cases, the apparent differences are likely driven more by outliers than by consistent shifts in median values. This suggests that Neighborhood, on its own, may not reliably explain variation in these numerical features, and its predictive strength may be better realized when used in combination with other variables or as part of interaction terms.

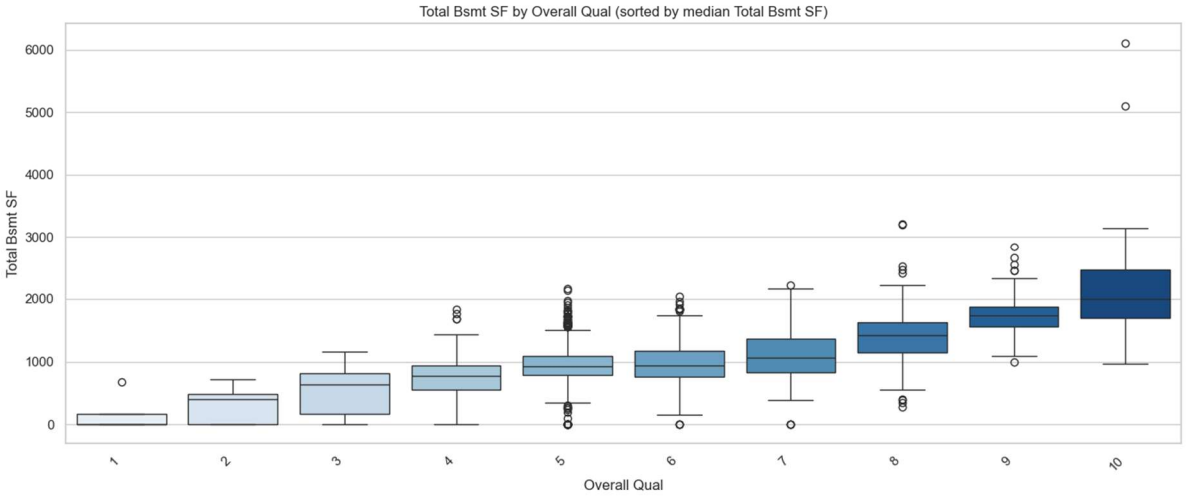
**Numerical Features by Overall Qual**





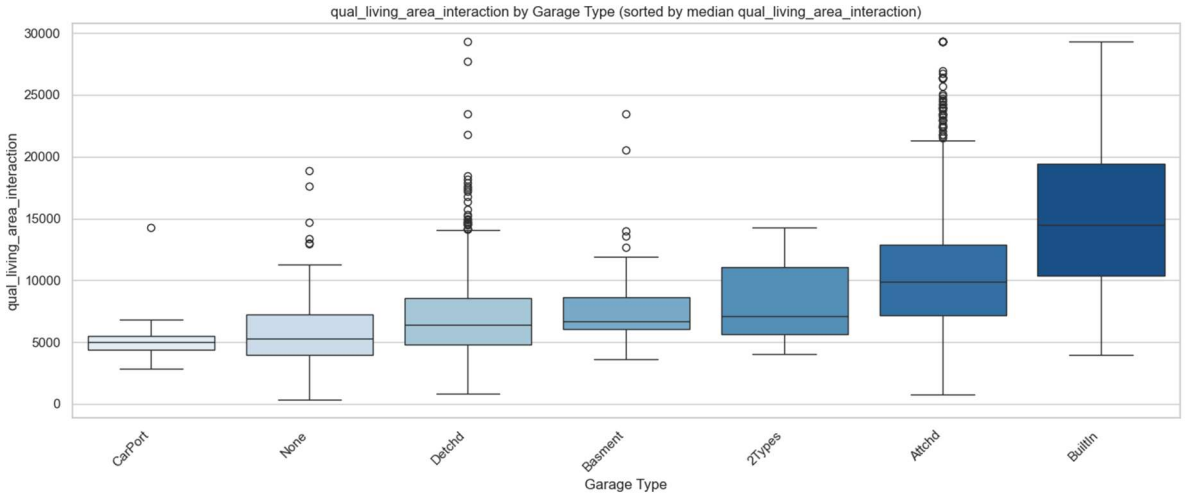



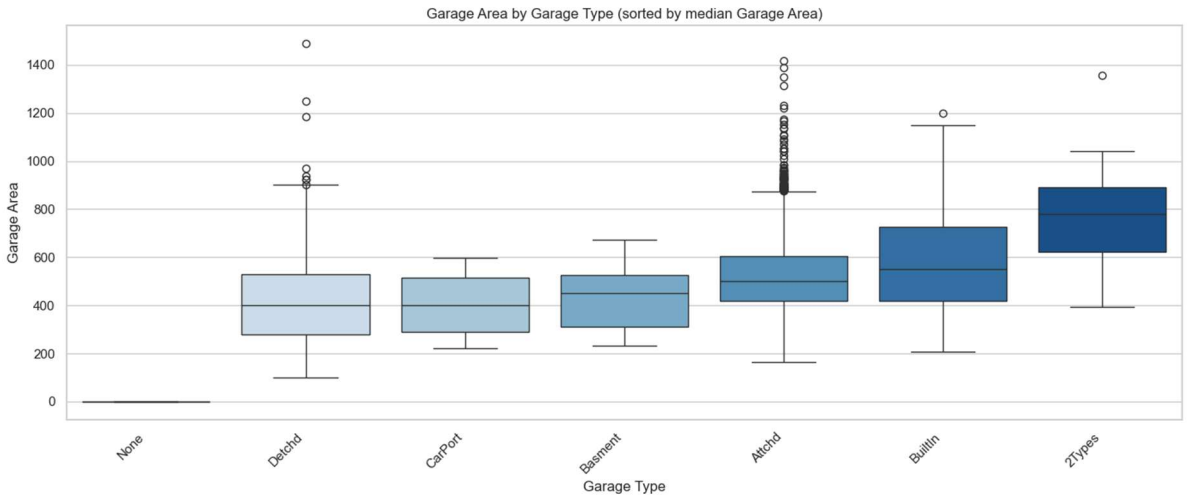

- **Median values** of `qual_living_area_interaction` across `Overall Qual` ordinal categories are clearly distinct, with differences increasing exponentially. This is expected because `qual_living_area_interaction` multiplies living space (`Gr Liv Area`) by `Overall Qual`, so the overall quality acts as a multiplier, amplifying the effect of living area on this feature.

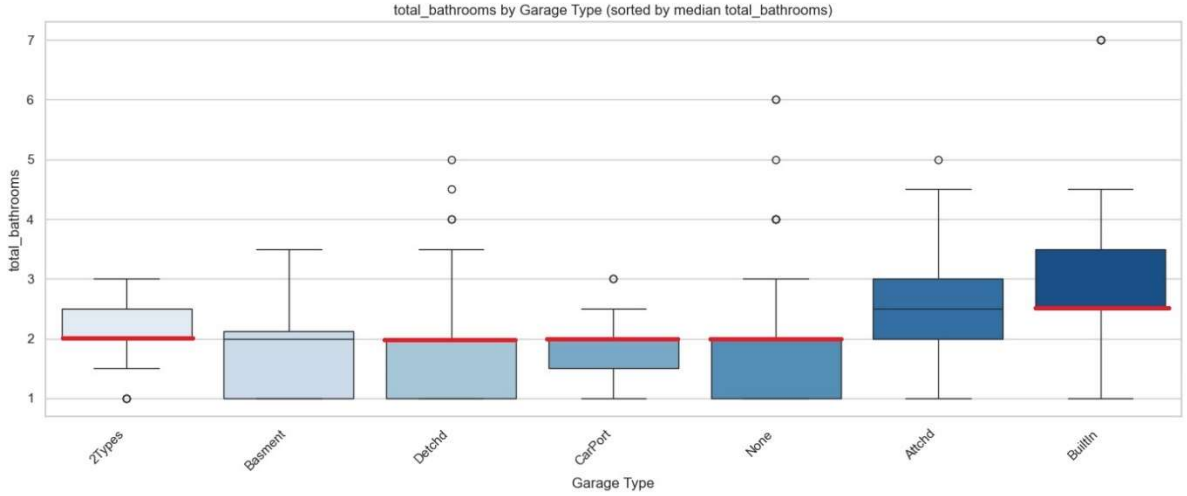
The median values of `Garage Area`, `total_bathrooms`, and `1st Flr SF` are relatively homogeneous across the lower `Overall Qual` scores, but become more distinct and evenly spaced as `Overall Qual` increases. In contrast, `Total Bsmt SF` shows more consistent medians between `Overall Qual` scores 5 to 7, with greater variation at both the lower and higher ends of the quality scale.

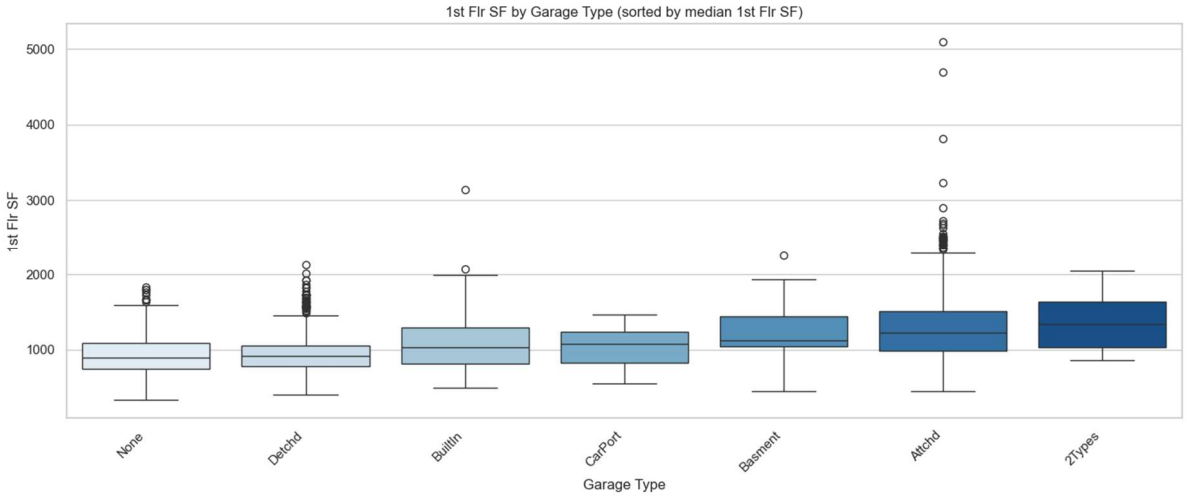
- **Interquartile range (IQR)** overlap is minimal between `qual_living_area_interaction` and `Overall Qual` scores — but this is to be expected, as `qual_living_area_interaction` is derived by multiplying `Overall Qual` with living area. For the remaining numerical features, there is considerable overlap across `Overall Qual` scores, though this tends to decrease slightly at the higher quality levels. Some variation in IQR is observed throughout.
- **Outlier Patterns**: All numerical features apart from `total_bathrooms` show dense clusters of outliers above the IQR, particularly around Overall Qual score 5, and to a lesser extent, score 6. `total_bathrooms` exhibits the fewest outliers overall, with those present being sparsely distributed.
- **Overall,** these plots show that `Overall Qual` captures meaningful variation in several numerical features, particularly at higher quality levels where medians separate more clearly and interquartile ranges narrow. At lower quality levels, many features remain relatively homogeneous with overlapping distributions, indicating weaker differentiation. Engineered features like `qual_living_area_interaction` exhibit predictable patterns by design, while others—such as `Garage Area` and `Total Bsmt SF`—show more nuanced, nonlinear behaviour. Outliers clustered around mid-level quality scores may obscure some trends. This suggests that, while `Overall Qual` clearly relates to these numerical features at higher quality levels, incorporating additional variables or interaction terms may be necessary to fully capture variation across the entire quality spectrum.

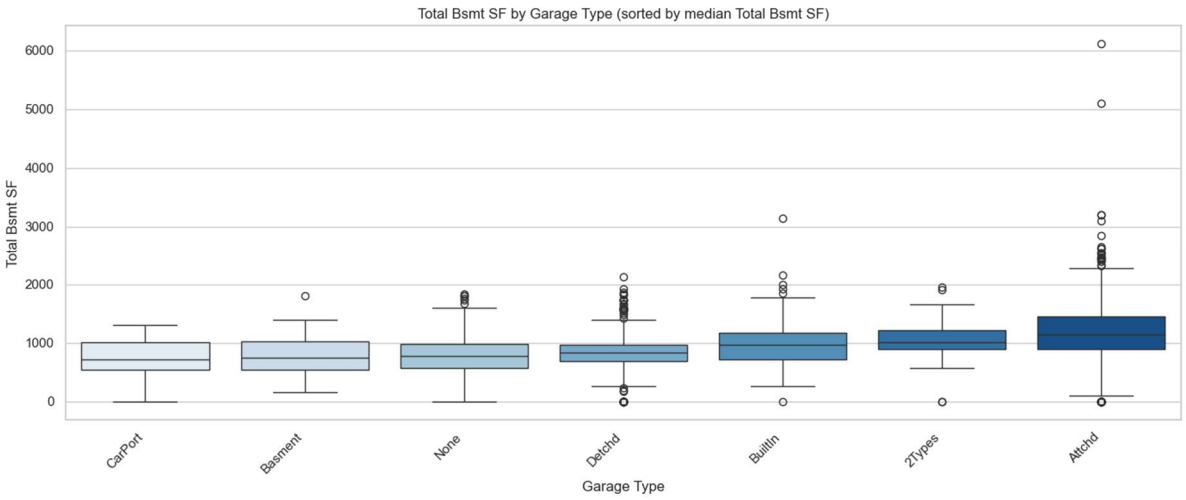
**Numerical Features by Garage Type**




Garage Area by Garage Type (sorted by median Garage Area)
Garage Area
0
200
400
600
800
1000
1200
1400
None
Detchd
CarPort
Basment
Attchd
BuiltIn
2Types
Garage Type

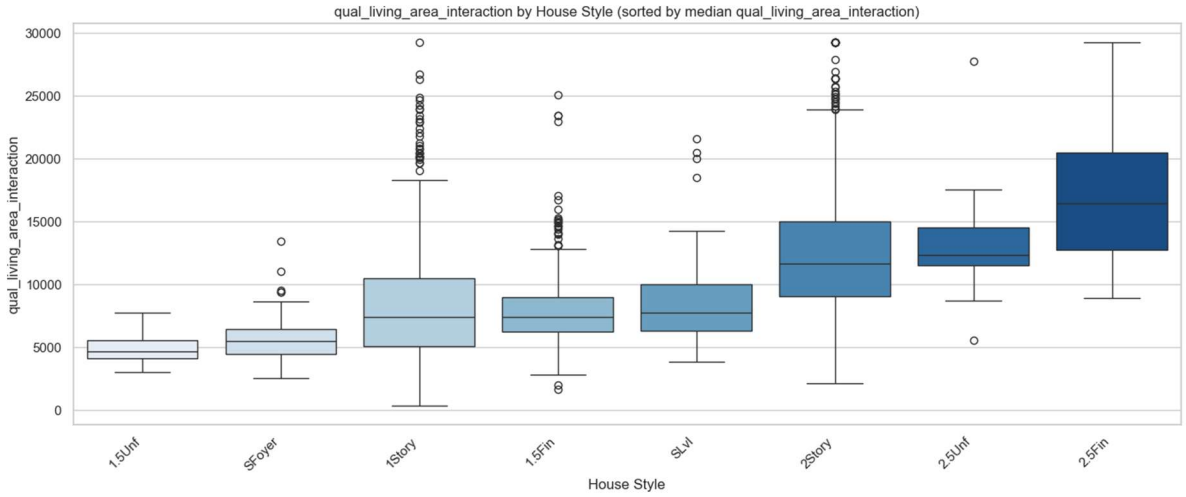

total_bathrooms by Garage Type (sorted by median total_bathrooms)
total_bathrooms
1
2
3
4
5
6
7
2Types
Basment
Detchd
CarPort
None
Attchd
BuiltIn
Garage Type

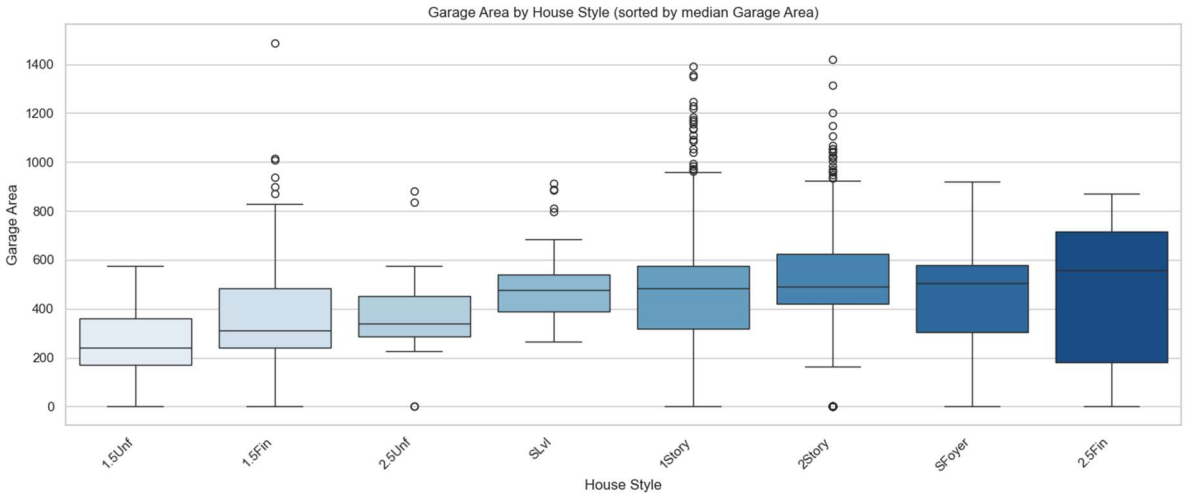

1st Flr SF by Garage Type (sorted by median 1st Flr SF)
1st Flr SF
1000
2000
3000
4000
5000
None
Detchd
BuiltIn
CarPort
Basment
Attchd
2Types
Garage Type

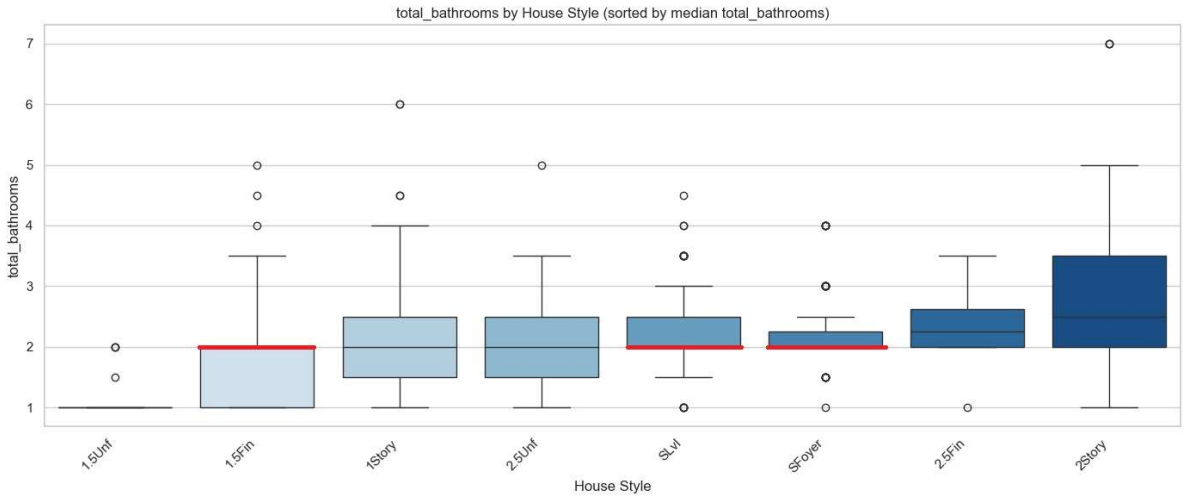


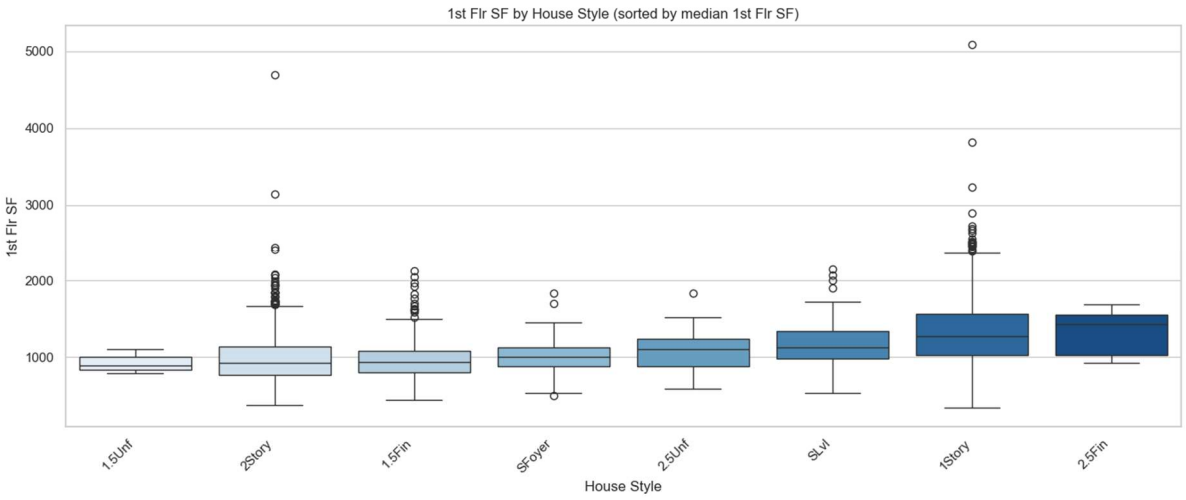

- For `1st Flr SF` and `Total Bsmt SF`, median values and interquartile ranges (IQRs) are fairly homogeneous across all Garage Types, with similar IQR overlaps, suggesting little to no relationship between these features and Garage Type.
- For `qual_living_area_interaction` and `total_bathrooms`, medians are generally homogeneous as well, except for the *Attached* and *BuiltIn* garage types, which show higher values of `qual_living_area_interaction`. This pattern is supported by the IQRs and their overlaps.
- A similar pattern of homogeneity is observed for `Garage Area` across most Garage Types, except for the *None* category, which is clearly associated with zero garage area, and the *2Types* category, which tends to correspond to higher garage areas.
- `total_bathrooms` exhibits very few outliers across all `Garage Type` categories, indicating relatively consistent values within each group.
- For the remaining numerical features, properties with Detached garages show the highest number of outliers, followed by those with Attached garages. Nearly all these outliers lie above the upper bound of the interquartile range (IQR), suggesting some homes in these categories have unusually large values.
- Overall, most numerical features show limited variation across Garage Types, with median values and IQRs often overlapping. However, some features—such as `qual_living_area_interaction`, `total_bathrooms`, and `Garage Area`—do show noticeable associations, particularly higher values for *Attached* and *BuiltIn* garages, and distinctly zero `Garage Area` for the *None* category. Outliers are generally few for `total_bathrooms` but more prevalent in Detached and Attached garages for other features. These patterns suggest a mix of stability and variation that should be considered in further analysis or modelling.

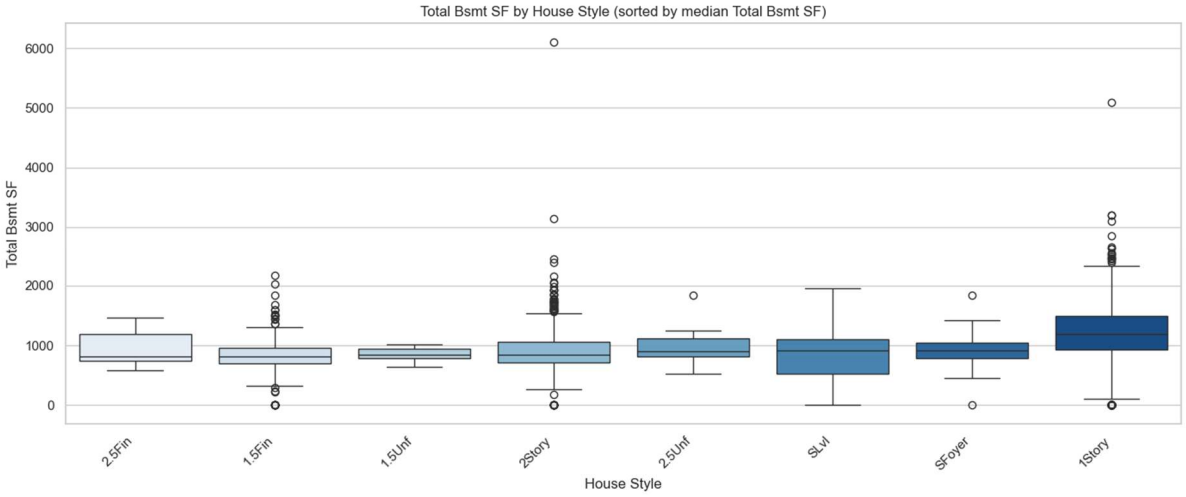
## Numerical Features by House Style



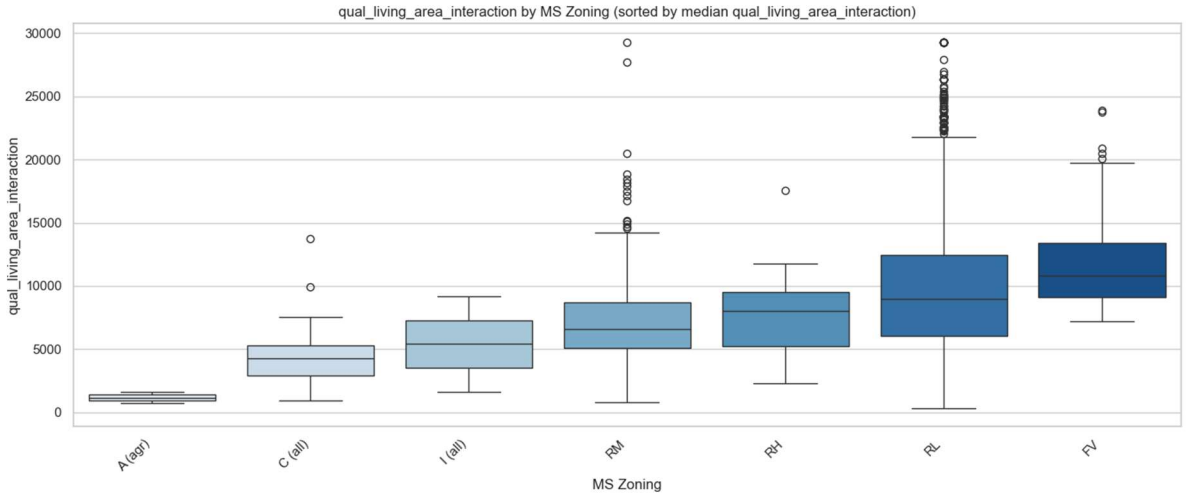



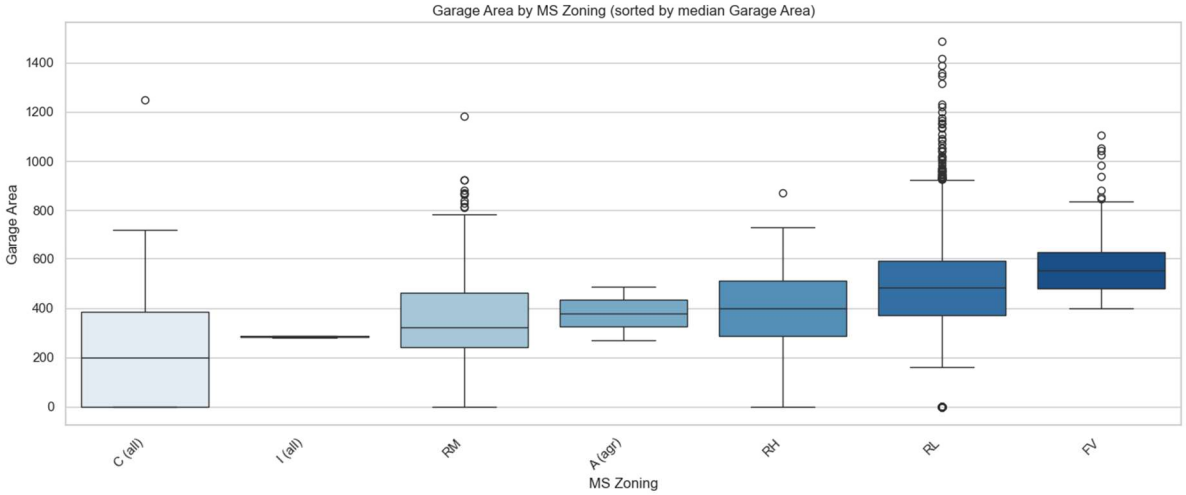


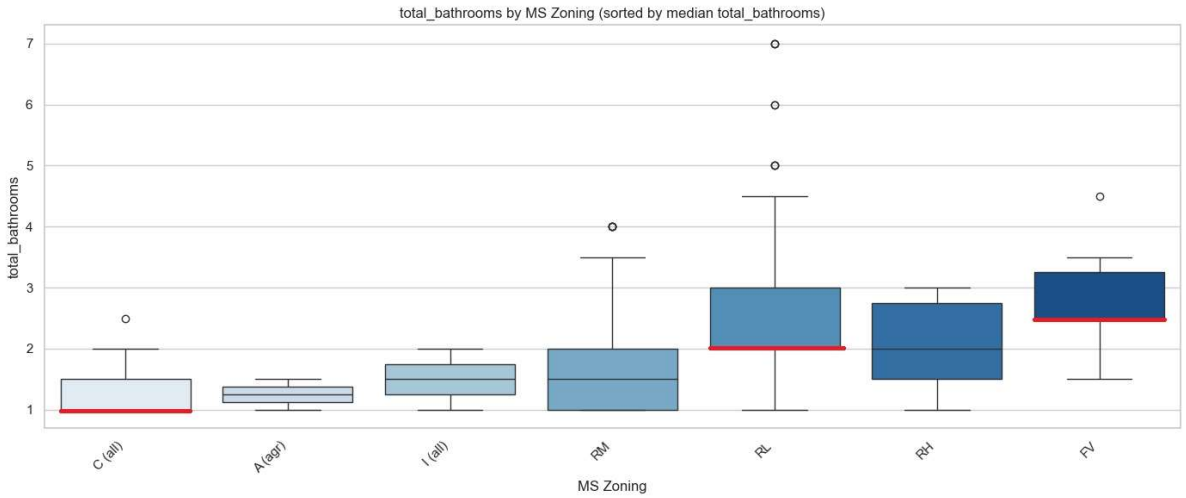



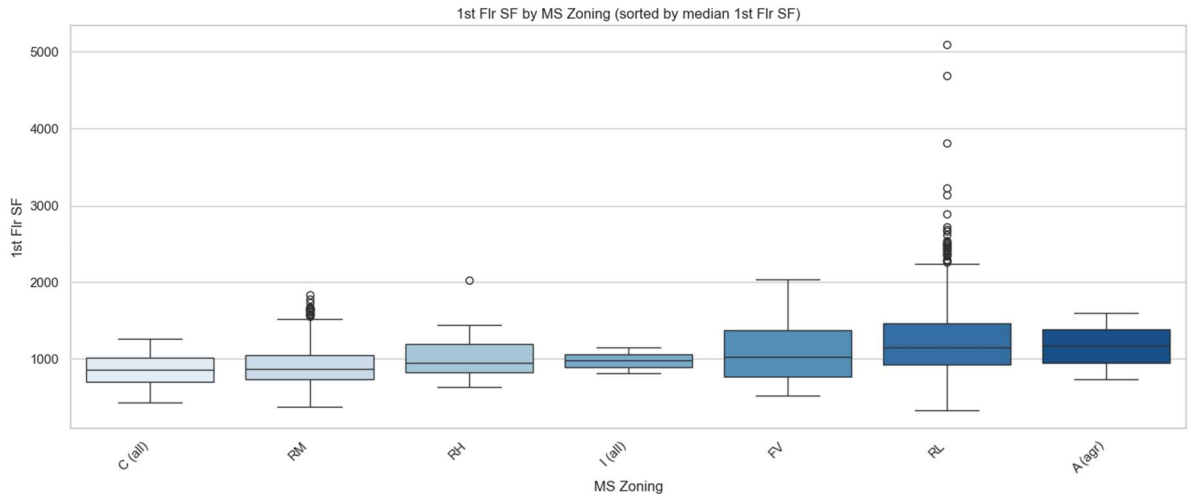

- For `qual_living_area_interaction`, median values tend to cluster into 3–4 overlapping groups with considerable IQR overlap and varied IQR widths.
- `1st Flr SF` and `Total Bsmt SF` show fairly homogeneous medians and IQRs across all House Styles, with substantial overlap, suggesting little relationship with `House Style`.
- `Garage Area` medians also show little variation, with extensive IQR overlap; however, the IQR sizes themselves are generally small.
- `Total_bathrooms` displays homogeneity across most categories except for *1.5Unf*, which has notably lower medians, and *2.5Fin* and *2Story*, which have higher medians. Although the IQR sizes vary considerably, they mostly overlap.
- With the exception of `total_bathrooms`, which has fewer and more evenly dispersed outliers, *2Story*, *1.5Fin*, and *1Story* house styles exhibit the most outliers, nearly all occurring above the upper interquartile range.
- Overall, numerical features show considerable overlap across House Styles, with medians and IQRs largely similar. However, some features—like `qual_living_area_interaction` and `total_bathrooms`—exhibit noticeable differences, particularly in categories such as *1.5Unf, 2.5Fin*, and *2Story*. Outliers are relatively rare for `total_bathrooms` but occur more frequently in *2Story, 1.5Fin*, and *1Story* styles, mostly above the upper IQR. These observations suggest mostly stable distributions with some meaningful variation that could inform further analysis or modelling.

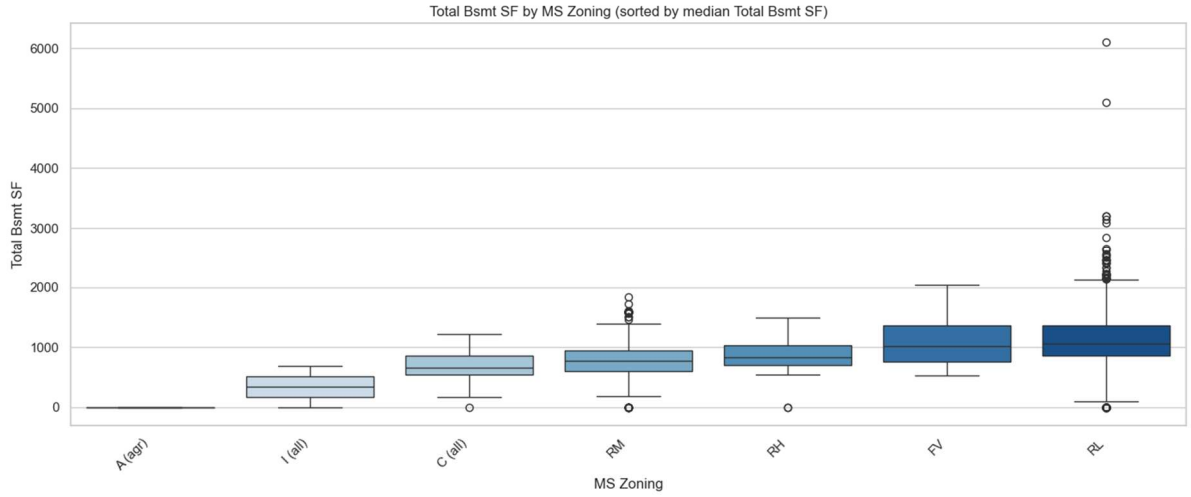
## Numerical Features by MS Zoning






- For `qual_living_area_interaction`, the medians are all distinct; however, the IQRs are mostly similar and overlap considerably, except for the *A(agr*) category.
- For `Garage Area`, the medians are also distinct, although some are less separated compared to `qual_living_area_interaction`. The IQR sizes vary significantly, with considerable overlap between categories.
- For `total_bathrooms`, the medians overlap somewhat, and the categories' IQRs also overlap despite some variation in IQR size.
- `1st Flr SF` shows fairly homogeneous medians and IQRs across all `MS Zoning` categories, with substantial overlap, suggesting little relationship with `MS Zoning`.
- `Total Bsmt SF` categories generally have homogeneous medians and overlapping IQRs, except for *A(agr),* which has nearly zero IQR and zero square footage.
- With the exception of `total_bathrooms`, which has very few outliers overall, *RL* followed by RM have the largest number of outliers, nearly all above the upper IQR
- Overall, numerical features show substantial overlap in medians and IQRs across `MS Zoning` categories, suggesting limited differentiation. Exceptions are seen in `qual_living_area_interaction` and `Garage Area`, where medians differ noticeably across categories despite overlapping IQRs. For these numerical features the *A(agr)* category is a clear outlier, showing a distinct median value and very narrow IQRs with little overlap. `Total_bathrooms` shows overlapping medians with few outliers, while *RL* and *RM* zones

exhibit the most outliers, primarily above the upper IQR. These patterns suggest generally stable distributions with some variation warranting further exploration.

*A full set of visualisations is included in the notebook appendix for reference. This ensures that, while not all plots are shown in the main report, the full exploratory context is preserved.*

---

## Key Insights

The univariate analysis reveals that most homes in the dataset are relatively modest in size, with smaller garages, porches, and bathrooms being more common. Room count features tend to follow a normal distribution, while size-related features often exhibit right skewness, with a smaller number of high-end properties extending the distribution tail. In categorical features, nominal variables like `Neighborhood`, `MS Subclass`, `House Style`, and exterior features show well-defined groupings that could influence sale price, while ordinal features like `Overall Cond` and `Kitchen Qual` reflect ordered quality levels with varying degrees of separation. Some features are dominated by a single category (e.g. `MS Zoning` is mostly *RL*), and high redundancy exists between `Exterior 1st` and `Exterior 2nd`, suggesting one could be dropped to reduce noise.

Bivariate analysis highlights strong correlations between sale price and quality-related features, such as `qual_living_area_interaction`, `avg_quality`, and `Gr Liv Area_capped`, while area-based features like `Garage Area` and `Total Bsmt SF` also show moderate to strong correlations. In contrast, room counts (e.g., `Bedroom AbvGr`, `TotRms AbvGrd`) and time-based features (e.g., `house_age`) are weakly correlated with price. Some features exhibit high collinearity, such as `1st Flr SF` with `Total Bsmt SF` and `Gr Liv Area_capped` with `qual_living_area_interaction`, indicating redundant information that could be streamlined in future modelling. Regplots confirm strong linear relationships for some features, especially engineered variables like `qual_living_area_interaction`, while others show diminishing strength at higher values.

The categorical features analysis shows that variables like `Neighborhood`, `House Style`, `Garage Type`, and `BsmtFinType1` influence sale price, with premium categories (e.g., *NoRidge, StoneBr, GLQ basements*) consistently associated with higher median prices. However, many categories across features like `MS Sub Class`, `Exterior1st`, `Exterior2nd`, and `Roof Style` exhibit overlapping interquartile ranges, reducing their discriminative power. Certain ordinal variables follow expected price gradients, though inconsistencies exist—for example, `Kitchen Qual`'s *Poor* and *Fair* categories share the same median, and some `Overall Cond` levels are indistinguishable in pricing. Outliers are more prevalent in common residential zones and well-known neighborhoods, often inflating the apparent price spread within those groups.

The multivariate analysis, incorporating both grouped boxplots and the correlation matrix, revealed key patterns in how numerical and categorical features relate to each other. The correlation matrix highlighted strong positive associations between sale price and features like `qual_living_area_interaction`, `Gr Liv Area_capped`, `Garage Area`, and `avg_quality`, with

qual_living_area_interaction showing one of the highest correlation coefficients. It also exposed redundancy between several features—for instance, Gr Liv Area_capped and qual_living_area_interaction, and Total Bsmt SF with 1st Flr SF—indicating that some variables might offer overlapping information.

The grouped boxplots further supported these findings by showing that while many numerical features (such as 1st Flr SF and Total Bsmt SF) had homogeneous medians and overlapping interquartile ranges across categories like Neighborhood, Garage Type, and MS Zoning, qual_living_area_interaction consistently stood out. It displayed distinct median shifts and reduced IQR overlap across multiple groupings, especially by Overall Qual, where its values increased predictably and meaningfully. These patterns suggest that qual_living_area_interaction performs especially well in capturing meaningful variation in house characteristics and pricing, outperforming its component features when examined in multivariate contexts.

---

## Hypotheses

The following hypotheses are proposed for further investigation:

- **Hypothesis 1:**
  There is a positive relationship between the engineered feature qual_living_area_interaction and sale price. Homes with higher values of this feature, which combines overall quality and living area, will generally have higher sale prices. This hypothesis builds on observed strong correlations and the clear pattern seen in regression plots, suggesting that the interaction of size and quality is a key driver of home value.
- **Hypothesis 2:**
  Neighborhood plays a significant role in determining sale price, with certain neighborhoods consistently commanding higher median sale prices than others. For example, neighborhoods such as *NoRidge* and *StoneBr* show notably higher medians compared to others. This hypothesis recognizes the importance of location as a major factor in real estate pricing, reflecting local demand, amenities, and community prestige.
- **Hypothesis 3:**
  The type and quality of a home's garage are associated with sale price. Homes featuring attached or built-in garages tend to have higher sale prices compared to those with detached garages or no garage at all. This hypothesis is supported by observed variations in median prices and outlier distributions across garage categories, indicating that garage characteristics contribute to overall property value.

## Hypothesis Testing

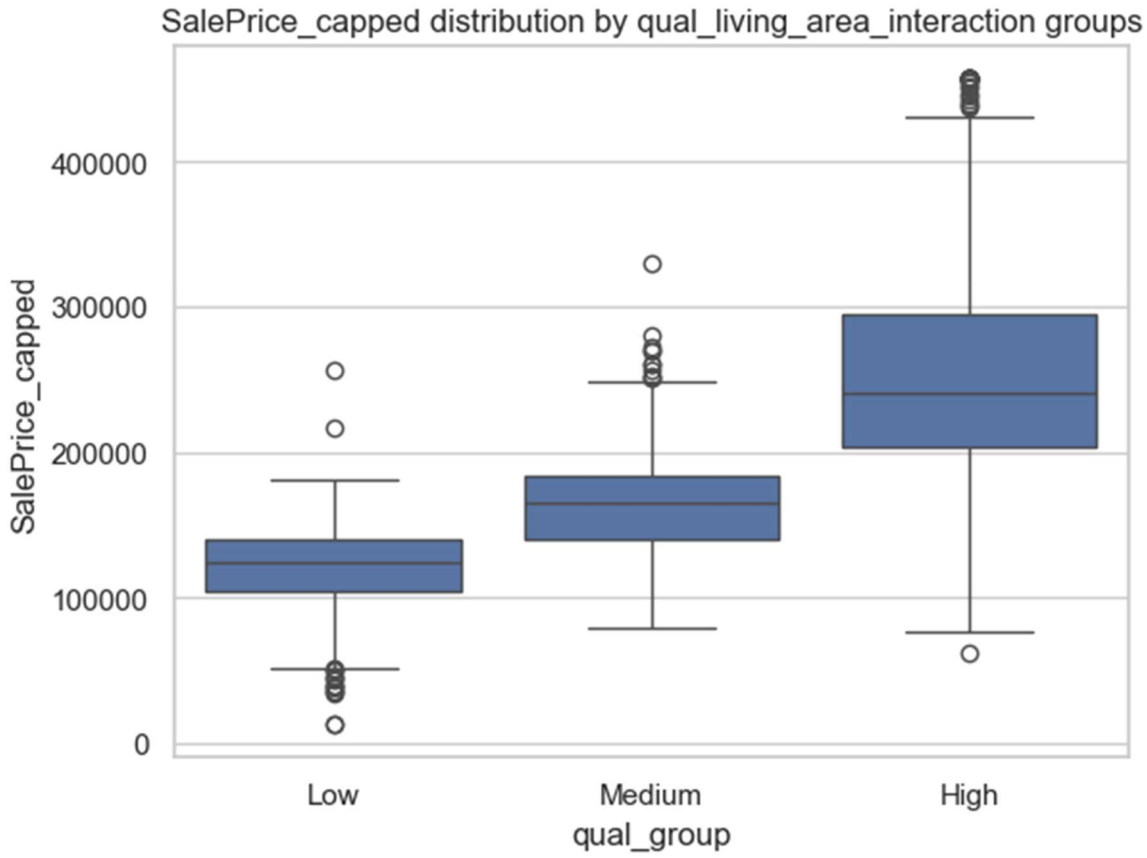
Hypothesis 1:
**There is a positive relationship between the engineered feature qual_living_area_interaction and sale price.**

To investigate whether the engineered feature **`qual_living_area_interaction`** (which combines overall quality and living area) is significantly associated with sale price, the continuous variable was divided into three equal-sized groups (Low, Medium, High) using quantile-based binning. This grouping allowed comparison of average sale prices across different levels of **`qual_living_area_interaction.`**

**Hypotheses:**

- **Null hypothesis ($H_0$):** There is no difference in mean sale price among the three **`qual_living_area_interaction`** groups.
- **Alternative hypothesis ($H_1$):** At least one group's mean sale price differs significantly from the others.

Using these groups, boxplots were generated to visualize the distribution of capped sale prices for each category, revealing an apparent increase in median sale price from Low to High groups.




An initial **one-way ANOVA** was conducted to assess whether the **mean** sale price differs significantly across the three groups. The test yielded a highly significant **F-statistic of 1823.76** and a **p-value < 0.0001**, providing strong evidence to **reject the null hypothesis** that group means are equal.

*The probability of making a Type I error (i.e., wrongly rejecting the null hypothesis when it is true) is extremely low, given the very small p-value. Similarly, the large test statistic and clear group differences reduce the likelihood of a Type II error (i.e., failing to detect a true difference).*

However, the distribution of sale prices within each group is known to be **right-skewed**, violating the ANOVA assumption of normally distributed residuals. As this can erroneously inflate the risk of Type I errors in the ANOVA test, a **non-parametric Kruskal–Wallis H-test** was conducted as a robustness check. This test does **not require normality** and instead compares the **median ranks** across groups.

**Kruskal–Wallis Test:**

The Kruskal–Wallis test produced a **statistic of 1882.61** and a **p-value < 0.0001**, again providing strong evidence to **reject the null hypothesis**.

**Conclusion:**

Both the ANOVA and the Kruskal–Wallis test confirm that **sale prices differ significantly across levels of `qual_living_area_interaction`**. This supports the hypothesis that **homes with higher values of this engineered feature** — which multiplies living area by overall quality — **tend to sell for more**. It reinforces the importance of this feature in capturing variation in housing value, and suggests it may be a particularly useful predictor in future modelling efforts.

Having confirmed through statistical testing that an engineered feature can significantly differentiate sale prices, developing comprehensive predictive regression models is a logical next step. This modelling phase will leverage the relationships and feature insights discovered throughout the exploratory analysis to build reliable tools for predicting housing prices and quantifying the relative importance of different property characteristics

---

# Regression Modelling Analysis

## Preprocessing Workflow Before Modelling

The modelling workflow began with a carefully prepared dataset derived from the exploratory data analysis (EDA). The key preprocessing steps ensured the data was machine-learning ready and that relationships identified in the EDA were properly captured.

- **Starting Dataset**: Included cleaned variables, engineered features such as `qual_living_area_interaction`, and capped continuous variables like Gr_Liv_Area_capped.
- **Categorical Encoding**:
    - Ordinal categorical variables (e.g., Overall_Qual, Kitchen_Qual) were encoded as integers reflecting their natural rank.
    - Nominal categorical variables (e.g., Neighborhood, Garage_Type) were transformed via one-hot encoding to avoid imposing any order. This added 173 columns to the dataframe
- **Polynomial Features:** To capture nonlinear relationships with the target (SalePrice_capped), numeric features were evaluated for potential benefit from polynomial transformations. For each feature, correlations with the target were compared for the original, squared, and cubic terms. Features showing improved correlation in their squared or cubic form were selected

for polynomial expansion. Polynomial features up to degree 3, including interaction terms, were generated separately for continuous and ordinal variables. This process added 627 new features to the dataset, which were then combined with the original data prior to model fitting.

- **Feature Scaling**: Numeric features were scaled using RobustScaler(), particularly for regularized regression models (Ridge, Lasso, ElasticNet) to ensure comparability of coefficients.
- **Feature Assessment**: Following encoding, transformation, and polynomial expansion, correlations with the target (SalePrice_capped) and multicollinearity indicators (Variance Inflation Factor, VIF) were reviewed to guide feature selection and ensure model stability. Boolean features were converted to numeric (0/1) for correlation assessment. VIF-based pruning reduced extreme multicollinearity, resulting in 230 features. A subsequent correlation filter removed features with very weak correlation to the target, reducing the dataset to 100 predictors. This approach provided a balance between retaining informative features for linear regression and leaving additional weak or correlated features to be handled later through regularized regression techniques (Ridge, Lasso, ElasticNet).
- **Note on Preprocessing Order and Potential Data Leakage:**
  In this analysis, polynomial feature generation and feature scaling were applied to the full dataset prior to performing the train-test split. Ideally, these transformations should be fitted on the training data only and then applied to the test set to avoid data leakage. Applying them beforehand means that information from the test set could have influenced the derived features, potentially inflating model performance. Consequently, the predictive results reported here are likely to be slightly more optimistic than they would be if the model were trained on fully unseen data. Future work should implement transformations post-split to ensure a fully unbiased evaluation.

## Model Evaluation and Diagnostics

- **Metrics:** For Model 1 and Model 2, performance was assessed using $R^2$, RMSE, and MAE. Model 1 used an 80/20 train–test split, while Model 2 employed 5-fold cross-validation due to its higher dimensionality. This approach provides robust evaluation of predictive performance and reduces sensitivity to any single train–test partition.
- **Diagnostics:**
  - Residual plots to check for heteroscedasticity and linearity violations (*Residuals vs Fitted scatter plots*).
  - Cook's distance and leverage statistics to identify influential observations (*Cook's Distance plots and Influence Plots*).
  - Normality tests on residuals (*QQ plots, Shapiro-Wilk test, and Jarque-Bera test*).
- **Implementation:** Detailed diagnostics were applied to Model 1 to understand baseline feature behavior and assumptions. For Model 2, diagnostics were deferred until a final feature selection and regularization step, at which point they will be applied comprehensively.
- **Rationale:** Systematic diagnostics ensure that assumptions of linear regression are monitored, potential data or model issues are flagged, and interpretation of coefficients is reliable in later stages.

- **Expected Outcome:** Identification of opportunities for improvements such as outlier treatment, variable transformations, regularization, or model simplification in the final model evaluation.

## Initial Model 1: Simple Linear Regression

- **Approach:** Five separate simple linear regression models were run, each with one of the top correlated features as the predictor of sale price. Features tested (individually):
    - qual_living_area_interaction (corr. 0.87)
    - Overall Qual Exter Qual Fireplace Qu (corr. 0.76)
    - Overall Qual^2 BsmtFin Type 1 (corr. 0.68)
    - total_bathrooms (corr. 0.65)
    - Total Bsmt SF (corr. 0.63)
- **Rationale:** These features were selected as they are the five strongest individual predictors of sale price. Running them as individual regressions would a clear view of their standalone explanatory power.
- **Hypothesis:** The qual_living_area_interaction feature was expected to produce the best-performing simple model, followed by quality and size-related predictors.
- **Expected Outcome:** These models were expected to provide a benchmark for how much variance in sale price can be explained by single features before moving to more complex models with multiple predictors and polynomial terms.
- **Evaluation and Diagnostics**: Performance metrics ($R^2$, RMSE, MAE) were calculated for both the training and test sets to assess predictive power and generalization. Overall, qual_living_area_interaction showed the strongest fit (Train $R^2$=0.746, Test $R^2$=0.796), followed by the other features in descending order of correlation. RMSE and MAE were generally low, with small differences between training and test metrics, though some results may be slightly optimistic due to data leakage from applying polynomial transformations and scaling before the train-test split.

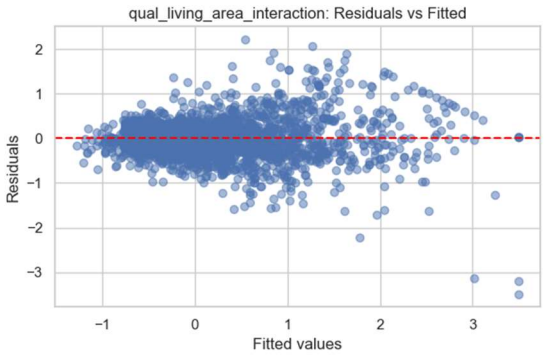



**Figure 1:** Residuals vs fitted values for qual_living_area_interaction showing mostly elliptical scatter with minor outliers, indicating the linearity assumption is roughly met.

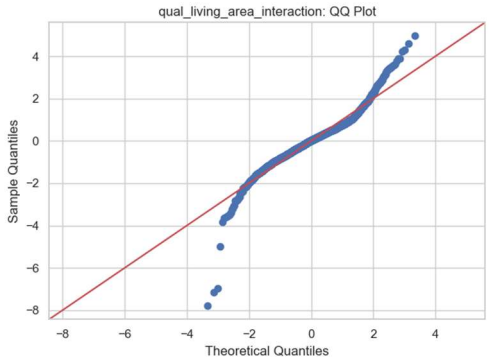



**Figure 2:** Figure 2: QQ plot of residuals for qual_living_area_interaction. Most residuals align with the theoretical quantiles, with slight deviations at the extremes. Normality tests (Shapiro-

Wilk W=0.943, p<0.001; Jarque-Bera JB=3153, p<0.001) confirmed departures from normality, though predictive performance remained meaningful.

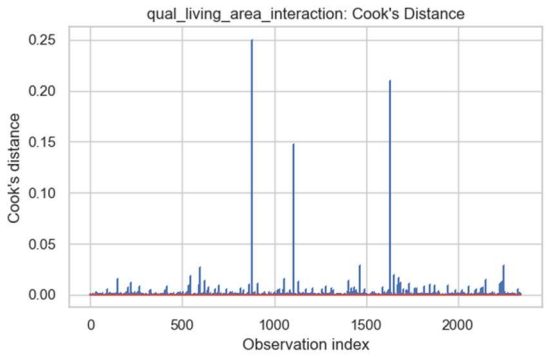



**Figure 3:** Cook's distance for qual_living_area_interaction. A few observations have high influence, warranting caution in inference but not invalidating predictive insights.

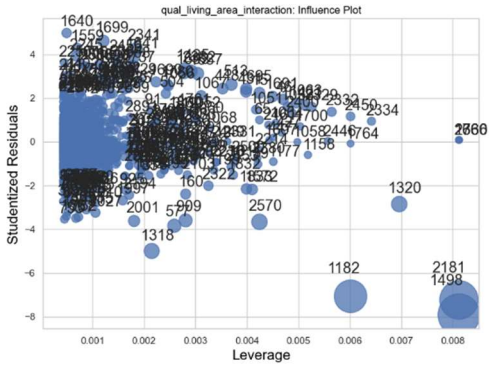



**Figure 4:** Influence Plot — Influence plot for qual_living_area_interaction. Shows the distribution of leverage and residuals, highlighting a few high-leverage points.

Across the other features, diagnostics showed similar issues: residuals were roughly random with occasional clustering, QQ plots deviated more strongly at the extremes, and Cook's distance highlighted a handful of influential observations (particularly in Total Bsmt SF and total_bathrooms). All residual normality tests rejected the null hypothesis, consistent with the visual deviations

Despite these assumption violations, Model 1 establishes clear baseline performance for the top individual predictors. These results highlight the need to expand into multiple regression with polynomial terms to better capture complex relationships, while also addressing issues such as influential outliers and residual non-normality.

| Feature | Train $R^2$ | Test $R^2$ | Train RMSE | Test RMSE | Train MAE | Test MAE |
|---|---|---|---|---|---|---|
| **qual_living_area_interaction** | 0.746 | 0.796 | 0.449 | 0.437 | 0.319 | 0.321 |
| **Overall Qual Exter Qual Fireplace Qu** | 0.557 | 0.63 | 0.592 | 0.589 | 0.447 | 0.44 |
| **Overall Qual^2 BsmtFin Type 1** | 0.464 | 0.429 | 0.651 | 0.732 | 0.487 | 0.531 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **total_bathrooms** | 0.417 | 0.415 | 0.679 | 0.741 | 0.494 | 0.535 |
| **Total Bsmt SF** | 0.389 | 0.447 | 0.695 | 0.72 | 0.539 | 0.539 |

**Table 1:** Performance Metrics for Model 1 Features. Training and test performance metrics ($R^2$, RMSE, MAE) for all five top features in Model 1. Differences between training and test metrics are generally small, though results may be slightly optimistic due to pre-split transformations.

## Initial Model 2: Multiple Linear Regression

- **Features:** Built from the post-assessment feature set (~100 predictors after VIF pruning and correlation filtering). Polynomial transformations (squares and cubes) were selectively applied to numeric variables showing curvature or nonlinear residual patterns in EDA. All features are scaled.
- **Rationale**: The expanded feature set aims to capture a wider range of signal than Model 1, balancing complexity with pruning to limit excessive collinearity.
- **Hypothesis:** A richer set of predictors, including nonlinear terms, should provide a more flexible representation of the housing price function and yield stronger predictive performance than Model 1.
- **Scope of this stage:** At this point, no further adjustments are made to address heteroscedasticity, skewness, bias/variance trade-offs, or outliers. These issues will be revisited after assessing Model 2's fit and diagnostic behaviour.
- **Expected Outcome:** Improved predictive accuracy relative to Model 1, albeit with possible risks of higher variance and interpretability challenges due to the enlarged feature space.
- **Evaluation:** Performance was assessed using 5-fold cross-validation, calculating $R^2$, RMSE, and MAE on both training and validation folds. This provides a robust estimate of model generalization and reduces sensitivity to a single train-test split.

| Fold | Train $R^2$ | Test $R^2$ | Train RMSE | Test RMSE | Train MAE | Test MAE |
|---|---|---|---|---|---|---|
| 0 | 0.9065 | 0.9109 | 0.2719 | 0.2892 | 0.1812 | 0.1821 |
| 1 | 0.9057 | 0.9158 | 0.2805 | 0.2537 | 0.1842 | 0.1876 |
| 2 | 0.9168 | 0.8626 | 0.2633 | 0.3242 | 0.1758 | 0.1930 |
| 3 | 0.9112 | 0.8925 | 0.2696 | 0.2982 | 0.1790 | 0.1959 |
| 4 | 0.9081 | 0.9030 | 0.2752 | 0.2792 | 0.1841 | 0.1895 |

**Table 1:** Cross-validation results per fold for Model 2

| Metric | Value |
|---|---|
| Train $R^2$ | 0.9097 |
| Test $R^2$ | 0.8970 |
| Train RMSE | 0.2721 |
| Test RMSE | 0.2889 |
| Train MAE | 0.1808 |
| Test MAE | 0.1896 |

**Table 2:** Average performance across folds for Model 2 - summarises the average performance across all folds, highlighting overall predictive ability and variability.

- **Observations:**

- Training performance is consistently high across folds, indicating the model fits the data well.
- Validation (test) performance shows more variability, particularly in fold 2, suggesting sensitivity to data partitioning and mild overfitting.
- Despite fold-to-fold variation, the model demonstrates meaningful predictive power and establishes a strong baseline for multi-feature regression.

**Visual Analysis of Model 2 Performance Across Folds**

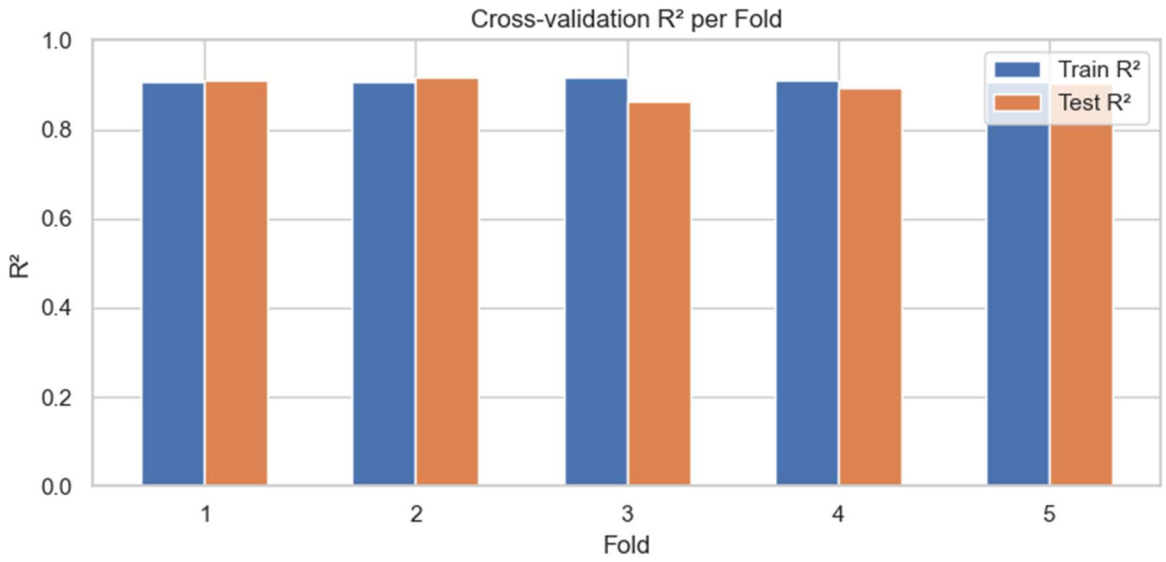



**Figure 1:** Train and validation $R^2$ scores for each fold.

- Training $R^2$ is consistent (~0.91), showing stable fit across training splits.
- Test $R^2$ varies more (0.86–0.92), with fold 3 showing the largest drop, indicating some subsets are harder to predict.
- High training $R^2$ with variable test $R^2$ hints at mild overfitting.

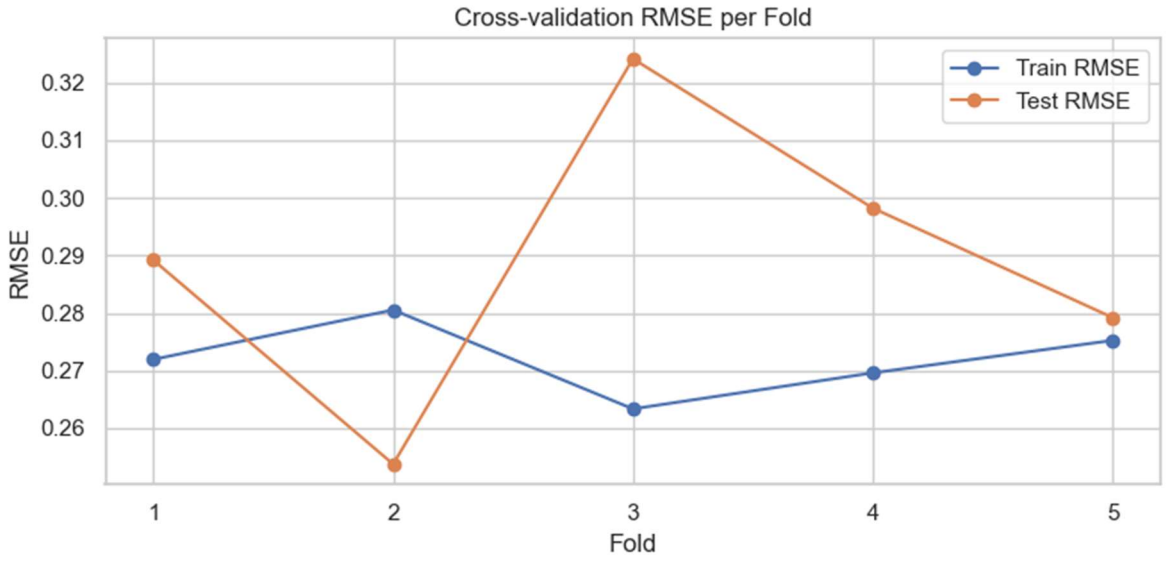



**Figure 2:** Train and validation RMSE across folds.

- Training RMSE is fairly consistent across folds, indicating a stable fit on the training data, with only minor variations between folds.
- Test RMSE closely matches training RMSE only in fold 5; in the other folds, it is either notably higher, as in fold 3, or substantially lower, as in fold 2.
- This suggests the model generalizes reasonably well for certain folds, but shows variability in performance across other test subsets.

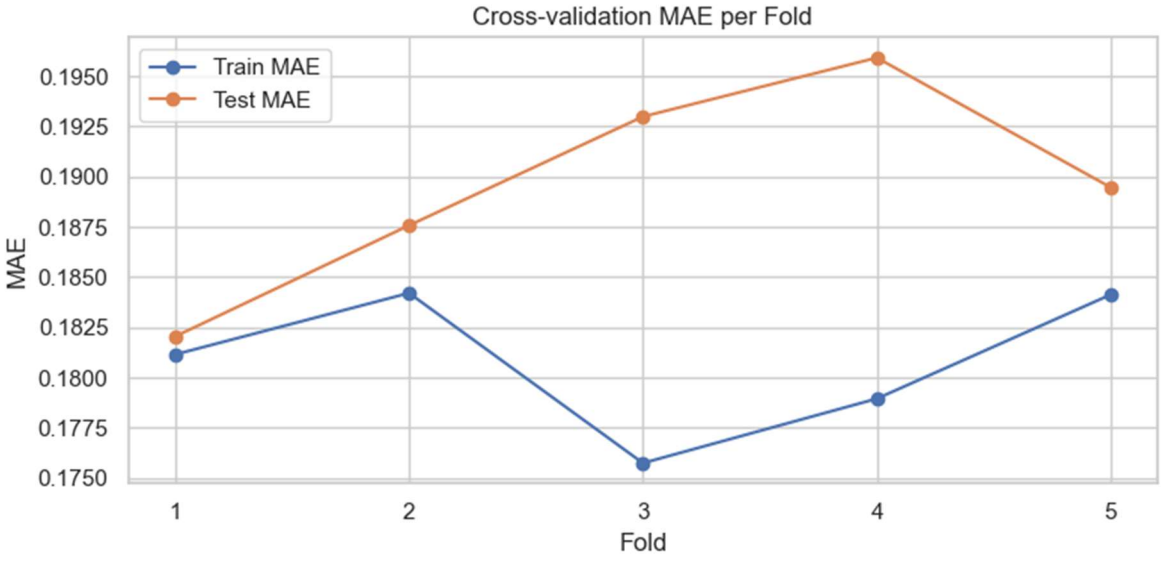



**Figure 3:** Train and validation MAE across folds.

- Training MAE shows variability across folds.
- Test MAE is very close to test MAE in fold 1 but higher than training MAE in every fold, with fold 4 showing the largest gap.

- This indicates mild overfitting, with predictive accuracy varying across different test subsets.

**Feature Importance Analysis**

To better understand the drivers of Model 2's predictions, the top features were assessed based on their regression coefficients:

| Feature | Coefficient |
|---|---|
| qual_living_area_interaction | 0.619 |
| Neighborhood_NridgHt | 0.434 |
| Exterior 2nd_CmentBd | 0.377 |
| Neighborhood_StoneBr | 0.377 |
| Neighborhood_NoRidge | 0.343 |
| Neighborhood_Somerst | 0.314 |
| Exterior 1st_CemntBd | -0.308 |
| MS SubClass_190 | -0.285 |
| Sale Condition_Partial | 0.271 |
| Condition 2_PosA | 0.263 |

**Interpretation:**

- Positive coefficients indicate features that increase predicted sale price, while negative coefficients reduce it.
- The engineered qual_living_area_interaction feature is the strongest predictor, confirming its importance from the EDA.
- Neighborhood indicators are consistently positive, showing that properties in desirable neighborhoods drive higher sale prices.
- Certain exterior materials and building types have negative coefficients, suggesting they slightly reduce predicted sale price relative to baseline categories.
- Overall, the coefficients highlight which features contribute most to the model's predictions, helping interpret the model and guiding potential feature engineering for future iterations.

**Summary of Observations**

- Training performance is consistently high across folds, indicating a strong fit on the training data.
- Validation (test) performance shows more variability, particularly in fold 3, suggesting sensitivity to data partitioning and mild overfitting.
- Feature importance analysis highlights key predictors, such as **qual_living_area_interaction**, **Neighborhood_NridgHt**, and **Exterior 2nd_CmentBd**, which are driving model predictions.
- Despite fold-to-fold variation, the model demonstrates meaningful predictive power and establishes a strong baseline for multi-feature regression.

## Strengths and Weaknesses of Models 1 and 2

**Model 1: Five Separate Simple Linear Regressions**

**Strengths**

- Clear interpretability: Each model focuses on a single feature, making coefficients and relationships easy to understand.
- Strong baseline performance: The qual_living_area_interaction predictor alone explains a substantial proportion of variance (Train $R^2$=0.746, Test $R^2$=0.796).
- Low overfitting risk: Training and test metrics are generally close, suggesting the simple models generalize reasonably well.
- Diagnostic clarity: Residuals show roughly elliptical scatter, allowing easy visual assessment of linearity and outliers.

**Weaknesses**

- Limited predictive scope: Single-feature models cannot capture interactions or combined effects beyond the polynomial/interaction term already included.
- Minor data leakage: Polynomial transformations applied before train-test split may slightly inflate performance metrics.
- Residual issues: QQ plots show departures from normality at the extremes, and a few observations have high leverage, indicating potential sensitivity to outliers.
- Predictive power constrained: Other top features explain less variance individually (Test $R^2$ 0.429–0.63), highlighting the need for multivariate modelling.

---

**Model 2: Multivariate Regression (~100 features including polynomials and interactions)**

**Strengths**

- High predictive power: Average Train $R^2$=0.910, Test $R^2$=0.897, showing strong overall model fit.
- Captures complex relationships: Polynomials and interactions allow modelling of non-linear effects and combined feature influence.
- Stable training metrics: RMSE and MAE are consistent across folds, suggesting good fit on training data.
- Scalable baseline: Provides a solid foundation for iterative refinement with more features, while multicollinearity has been reduced through preprocessing.

**Weaknesses**

- Mild overfitting: Test metrics are more variable across folds, particularly fold 3 (Test $R^2$ drop, higher RMSE/MAE), indicating some sensitivity to data partitioning.

- Reduced interpretability: Large feature set and polynomial terms make it harder to communicate model insights.
- Residual patterns: Although no funnel or strong curvature was observed, residuals are not perfectly randomly distributed, suggesting minor deviations from assumptions may persist.
- Scope for refinement: Outlier handling, variance stabilisation, and further feature engineering could improve predictive consistency across folds.

## Iterative Refinement

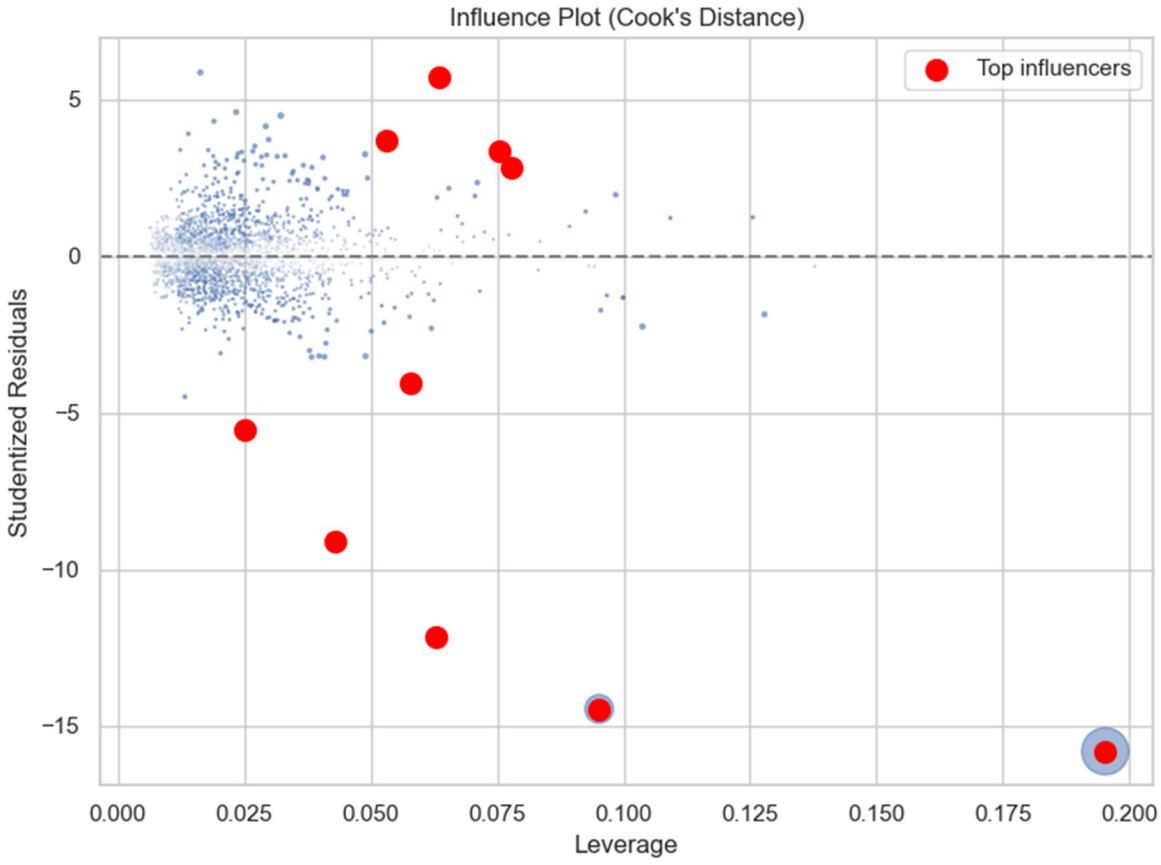
The strengths and weaknesses observed across Models 1 and 2 highlight areas where the models perform well and aspects that require further adjustment. While Model 2 achieves high predictive accuracy and captures complex relationships, mild overfitting, residual deviations, and fold-to-fold variability indicate room for improvement. To enhance robustness, interpretability, and generalization, iterative refinements are proposed, beginning with outlier handling.

### 1. Outlier Treatment

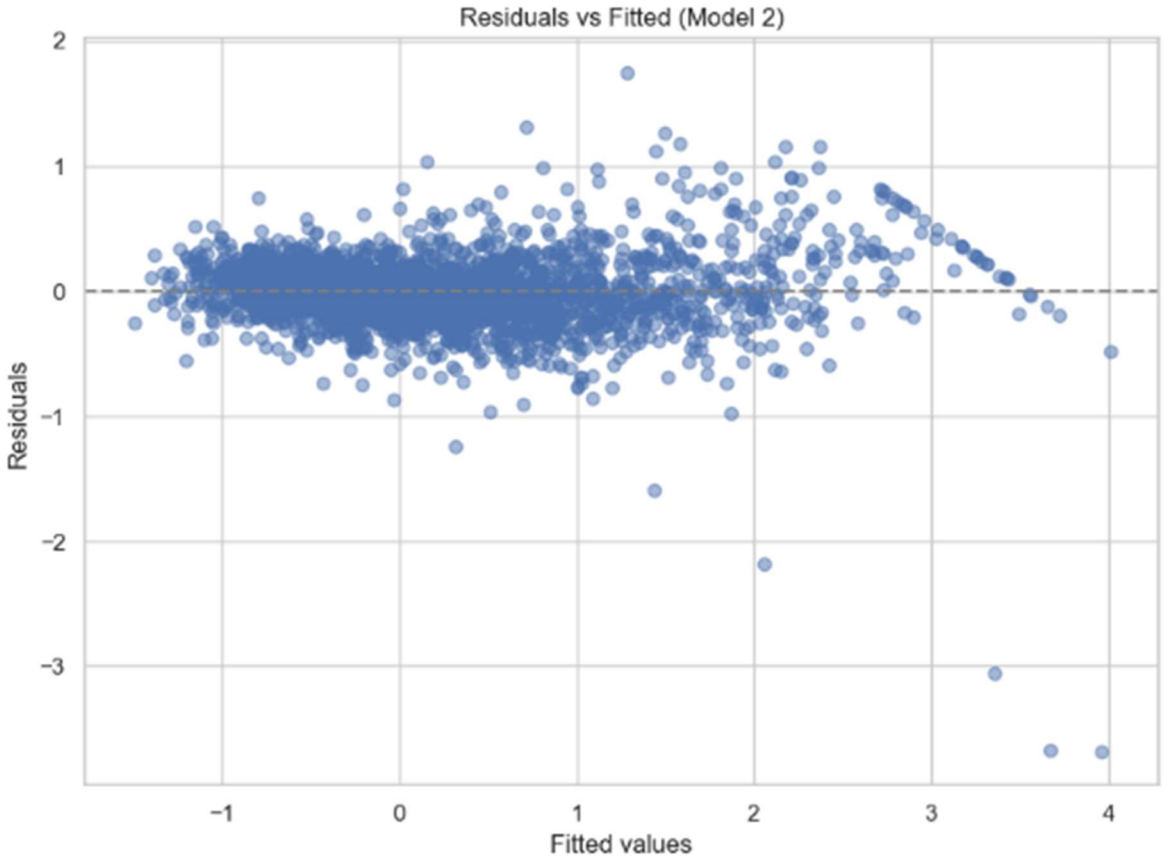
**Objective:** To determine whether removing extreme or influential points improves the predictive accuracy of the linear regression model on SalePrice_capped.

#### 1.1. Baseline Diagnostics

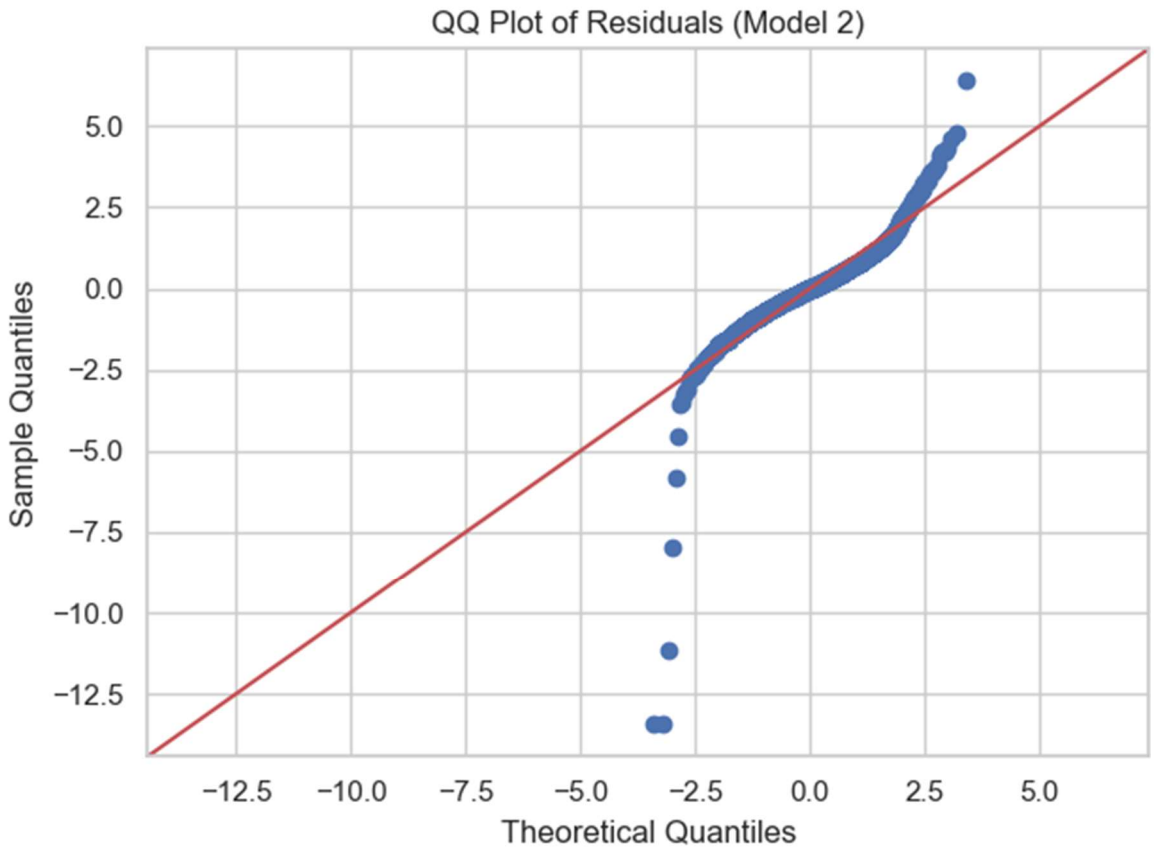
- **Cook's Distance:** Several observations exceed the 4/n threshold, with one dominant point in the bottom-right of the influence plot.




- **Residuals vs Fitted:** The cluster is elliptical but denser on the left, indicating mild heteroscedasticity. At higher fitted values, ~5 outliers fall between 1–4 (fitted) and -1.5 to -3.8 (residuals), while ~10 points fall between 3–4 (fitted) and -0.5 to 0.5 (residuals). This pattern is typical of skewed housing price targets.




- **QQ Plot:** Most points align with the 45° line at upper quantiles, though a slight right skew is evident. Two extreme lower-quantile points are observed at (-3, -13).




- **Normality:** The Shapiro–Wilk test yielded $p \approx 0$, indicating non-normal residuals. Given the large dataset, minor deviations are not expected to harm predictive accuracy.

- **Multicollinearity (VIF):**

| Feature | VIF |
|---|---|
| house_age | 106.346 |
| years_to_remodel | 84.171 |
| remodel_age | 56.866 |
| total_bathrooms | 50.768 |
| total_porch_area | 41.139 |
| Wood Deck SF_capped | 29.441 |
| Full Bath | 26.134 |
| Bsmt Full Bath | 19.463 |
| qual_living_area_interaction | 11.922 |
| Garage Cars Mo Sold avg_quality | 9.325 |

Several features exhibit very high multicollinearity. *house_age*, *years_to_remodel*, and *remodel_age* are near-perfectly collinear, while *total_bathrooms*, *total_porch_area*, and *Wood Deck SF_capped* also display inflated VIFs. This indicates that coefficients may be unstable and regularization will be needed in future iterations.

- **Diagnostics Summary:** Model 2 already achieves high accuracy (Test $R^2 \approx 0.897$). Outlier removal may offer incremental improvements, but wholesale removal of hundreds of points risks damaging model structure.

### 1.2. Identification of Influential Points

Cook's distance flagged 156 points above the 4/n threshold. The largest values included indices 1498, 2180, 2181, and 1182. A case-by-case review compared their features against dataset medians and IQRs.

| Index | Action | Rationale |
|---|---|---|
| 1498 | Remove | Extremely large, top quality, but very low sale price → likely data error |
| 2180 | Remove | Very large, high quality, but sale price unusually low |
| 2181 | Remove | Similar to 2180; large, high quality, but suspiciously low price |
| 1182 | Remove | Large, high quality, but very small basement and unusually low sale price |
| 1945 | Monitor | Large home, low quality, but price above median → inconsistent |
| 2737 | Monitor | Large home, moderate quality, high condition, price above median → unusual but plausible |
| 2570 | Keep | Large home, decent quality, slightly unusual features but within range |
| 1782 | Keep | Smaller home, decent quality, features consistent with dataset |
| 91 | Keep | Medium-large, good quality, price above median → reasonable |
| 2666 | Optional | Large, top-quality, very old home with high price → unusual historic case |

### 1.3. Outlier Sensitivity Analysis

- **Approach**:
  - Tested all combinations of removing the four strongest candidates (1498, 2180, 2181, 1182).
  - Evaluated metrics ($R^2$, RMSE, MAE) via 5-fold CV.
  - Compared results against the original full dataset..
- **Results**

| Removed Indices | Test $R^2$ | Δ Test $R^2$ | Test RMSE | Δ Test RMSE | Test MAE | Δ Test MAE |
|---|---|---|---|---|---|---|
| (1498, 2181, 2180, 1182) | 0.925 | 0.028 | 0.248 | -0.041 | 0.178 | -0.011 |
| (1498, 2181, 2180) | 0.921 | 0.025 | 0.253 | -0.036 | 0.180 | -0.010 |
| (1498, 2180, 1182) | 0.917 | 0.020 | 0.259 | -0.030 | 0.180 | -0.010 |
| (1498, 2180) | 0.914 | 0.017 | 0.263 | -0.026 | 0.181 | -0.009 |
| (1498, 2181, 1182) | 0.912 | 0.015 | 0.266 | -0.023 | 0.181 | -0.008 |
| (1498, 2181) | 0.909 | 0.012 | 0.270 | -0.019 | 0.182 | -0.007 |
| (2181, 2180, 1182) | 0.909 | 0.012 | 0.271 | -0.018 | 0.184 | -0.005 |
| (1498, 1182) | 0.907 | 0.010 | 0.275 | -0.014 | 0.183 | -0.006 |
| (2180, 1182) | 0.904 | 0.007 | 0.278 | -0.011 | 0.186 | -0.004 |
| (2181, 2180) | 0.904 | 0.007 | 0.276 | -0.013 | 0.185 | -0.005 |
| (1498,) | 0.903 | 0.006 | 0.280 | -0.008 | 0.185 | -0.004 |
| (2180,) | 0.902 | 0.005 | 0.281 | -0.008 | 0.186 | -0.003 |
| (2181,) | 0.901 | 0.004 | 0.284 | -0.005 | 0.188 | -0.002 |
| (2181, 1182) | 0.901 | 0.004 | 0.281 | -0.007 | 0.188 | -0.002 |
| (1182,) | 0.898 | 0.001 | 0.287 | -0.001 | 0.190 | 0.001 |

| | | | | | | |
|---|---|---|---|---|---|---|
| () | 0.897 | 0.000 | 0.289 | 0.000 | 0.190 | 0.000 |

- **Key observations:**

    - Removing all four indices produced the largest improvements in $R^2$, RMSE, and MAE
    - Smaller subsets also improved performance, though less substantially.
    - The model is sensitive to extreme outliers, but not dominated by any single case.

### 1.4. Interpretation

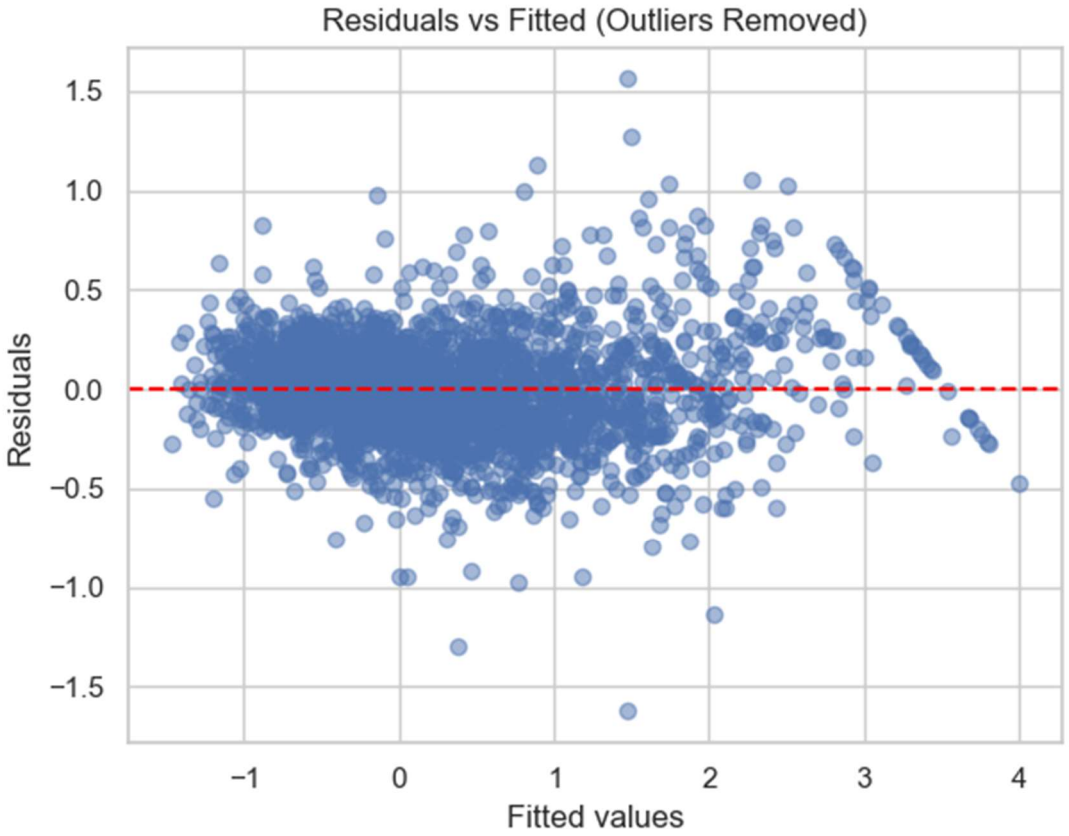
- The baseline model (Model 2) was already highly predictive (Test $R^2 \approx 0.897$).
- Removing the four most extreme outliers increased Test $R^2$ to 0.925 and lowered RMSE and MAE, confirming their disproportionate influence.
- Even partial removal improved stability, highlighting sensitivity to extremes.

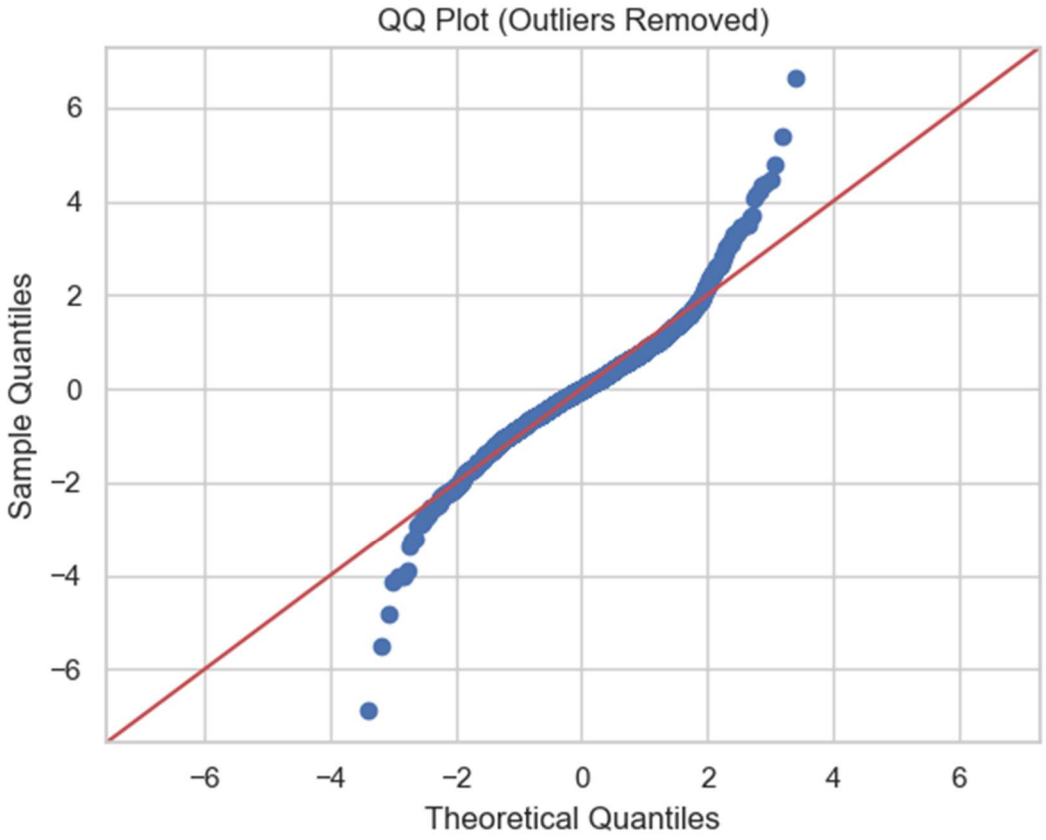
- **Recommendation:** For maximum predictive accuracy, exclude indices 1498, 2180, 2181, and 1182. For realism and completeness, retaining them is defensible, though with slightly weaker metrics.
- **Conclusion:** The regression model is generally robust, but targeted removal of extreme or erroneous observations can yield tangible performance gains.

### 1.5. Evaluation and Assumption Validation after Outlier Treatment

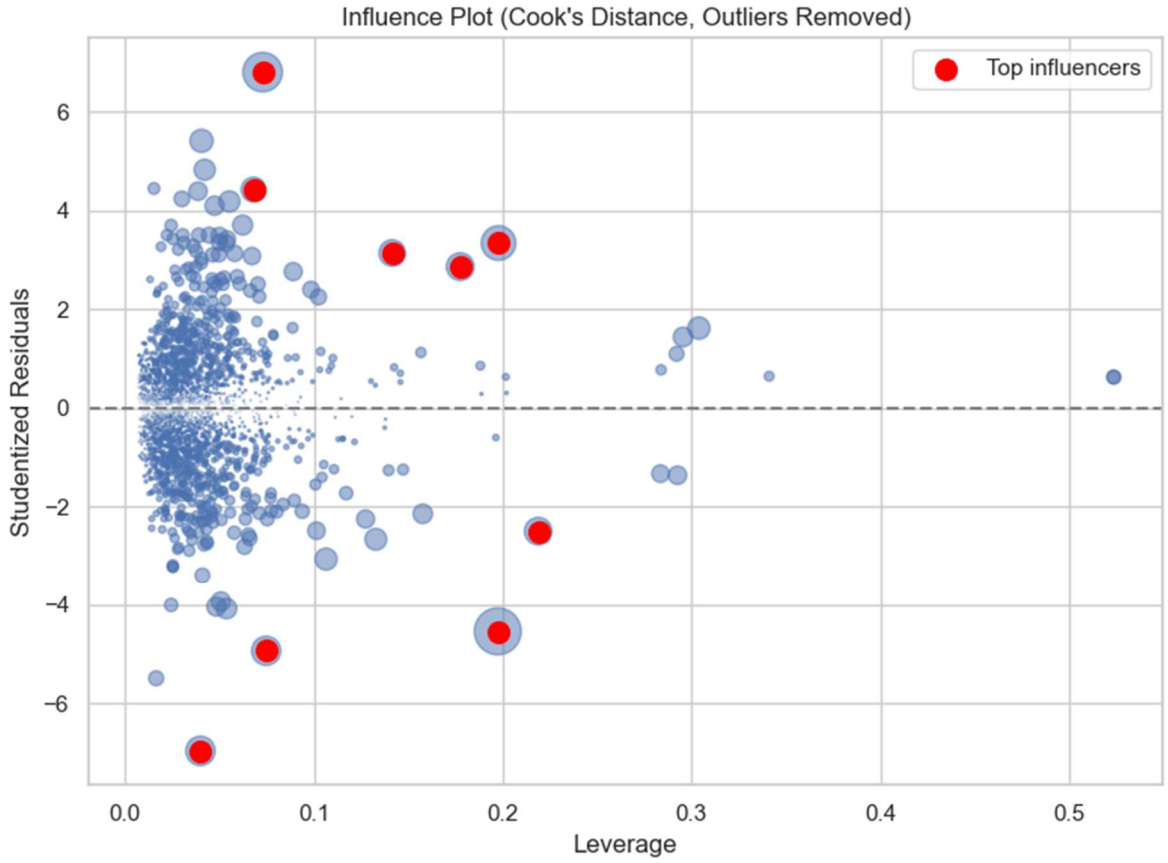
After the Outlier Treatment iteration was completed, the following checks and metrics were recorded:

- **Residual Diagnostics**
    - Residuals vs fitted plot shows reduced noise and improved stability across most of the dataset.
    - QQ plots indicate improved normality, with deviations limited to the extremes.
    - Shapiro–Wilk test ($p \approx 0$) still rejects perfect normality, but deviations are minor and acceptable given sample size.
    - Cook's distance shows that the previously dominant outlier is gone; remaining points are less influential.


Residuals vs Fitted (Outliers Removed)
Residuals
Fitted values
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1
0
1
2
3
4


QQ Plot (Outliers Removed)
Sample Quantiles
Theoretical Quantiles
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6

| Index | Cook's Distance | Target |
|---|---|---|
| 1356 | NaN | 0.000000 |
| 782 | 0.049983 | -0.202381 |
| 2737 | 0.035529 | 3.035714 |
| 241 | 0.027192 | 1.762262 |
| 1782 | 0.019679 | -0.154762 |
| 2570 | 0.019375 | 0.892857 |
| 1945 | 0.017491 | 1.685714 |
| 1027 | 0.017269 | 1.577381 |
| 2666 | 0.016047 | 3.531743 |
| 1425 | 0.014083 | 3.531743 |




- **Variance Inflation Factors (post removal):**

| Feature | VIF |
|---|---|
| Bldg Type_Duplex | inf |
| MS SubClass_90 | inf |
| house_age | 847.4285 |
| years_to_remodel | 661.6787 |
| remodel_age | 444.5772 |
| Garage Finish_None | 236.0888 |
| Garage Type_None | 234.8742 |
| total_bathrooms | 52.65361 |

| | |
|---|---|
| total_porch_area | 42.75427 |
| Sale Type_New | 39.75365 |

Several predictors now exhibit inflated VIFs due to structural shifts after removal. This will be addressed in later iterations (e.g., stepwise selection, regularization)

- **Performance Metrics**

The table below shows that after removing the 4 influential observations, the model shows clear improvement in both training and test performance. Train and test $R^2$ increased, while RMSE and MAE decreased, indicating a better fit and more accurate predictions. This suggests that the removal of extreme points stabilized the model without overfitting.

| Metric | Original Model 2 | Post-Removal Model |
|---|---|---|
| Train $R^2$ | 0.9097 | 0.9330 |
| Test $R^2$ | 0.8970 | 0.9249 |
| Train RMSE | 0.2721 | 0.2345 |
| Test RMSE | 0.2889 | 0.2475 |
| Train MAE | 0.1808 | 0.1703 |
| Test MAE | 0.1896 | 0.1784 |

### 1.6. Feature Importance Analysis

To better understand the drivers of Model 2's predictions, the top features were assessed based on their regression coefficients:

| Rank | Model 2 (All Data) | Coefficient | Outlier-Removed Model | Coefficient |
|---|---|---|---|---|
| 1 | qual_living_area_interaction | 0.619 | qual_living_area_interaction | 0.553 |
| 2 | Neighborhood_NridgHt | 0.434 | Exterior 2nd_CmentBd | 0.427 |
| 3 | Exterior 2nd_CmentBd | 0.377 | Exterior 1st_CemntBd | -0.323 |
| 4 | Neighborhood_StoneBr | 0.377 | MS SubClass_190 | -0.308 |
| 5 | Neighborhood_NoRidge | 0.343 | Neighborhood_NridgHt | 0.284 |
| 6 | Neighborhood_Somerst | 0.314 | Neighborhood_StoneBr | 0.273 |
| 7 | Exterior 1st_CemntBd | -0.308 | Neighborhood_Somerst | 0.221 |
| 8 | MS SubClass_190 | -0.285 | Neighborhood_NoRidge | 0.218 |
| 9 | Sale Condition_Partial | 0.271 | BsmtFin SF 1 | 0.217 |
| 10 | Condition 2_PosA | 0.263 | Foundation_Slab | 0.203 |

- Qual_living_area_interaction is still the dominant feature, but slightly less influential now that outliers are removed. This suggests that extreme homes (likely very large/small + quality) were previously inflating the importance of this interaction.
- Neighbourhood effects are still positive and important, but coefficients have shrunk; removing outliers has reduced the "premium" those neighbourhoods seemed to add — extreme prices in those areas were likely driving higher coefficients.
- Exterior material is remarkably stable which suggests exterior finish effects are robust to outliers

- Outlier removal lets more “core structural features” (basement finish, foundation) rise in importance, while some “edge case” sale conditions drop out of the top 10. This suggests the model is now less driven by rare transaction types and more by general house characteristics.
- **Summary:** The outlier-removed model is **more balanced and generalisable**, focusing less on unusual transactions and extreme properties, and more on consistent structural features.

### 1.7. Outlier Treatment Summary

- **Residuals & Normality:** Outlier removal improved residual stability, with deviations largely confined to the extremes and QQ plots showing closer alignment to normality.
- **Influence & Leverage**: Removing four highly influential cases eliminated distortion in coefficients and residual structure, reducing extreme leverage.
- **Predictive Performance:** Test $R^2$, RMSE, and MAE all improved after outlier removal, indicating stronger generalisation without added overfitting.
- **Parsimony & Robustness:** Multicollinearity persists, but feature importance is now more balanced and less sensitive to rare or extreme cases.
- **Feature Importance:** Core predictors remain consistent, while structural features gain prominence and rare transaction-specific effects diminish.
- **Overall Decision:** Outlier removal yields more stable, accurate, and generalisable models, though at the cost of excluding some legitimate extreme cases.

## 2. Feature Refinement and Response Transformation

- **Motivation:**
    - Residual analysis indicated no missing non-linear patterns (so no need for additional polynomial or interaction terms).
    - However, a slight funnel shape was observed in the residuals vs fitted plot, suggesting mild heteroscedasticity

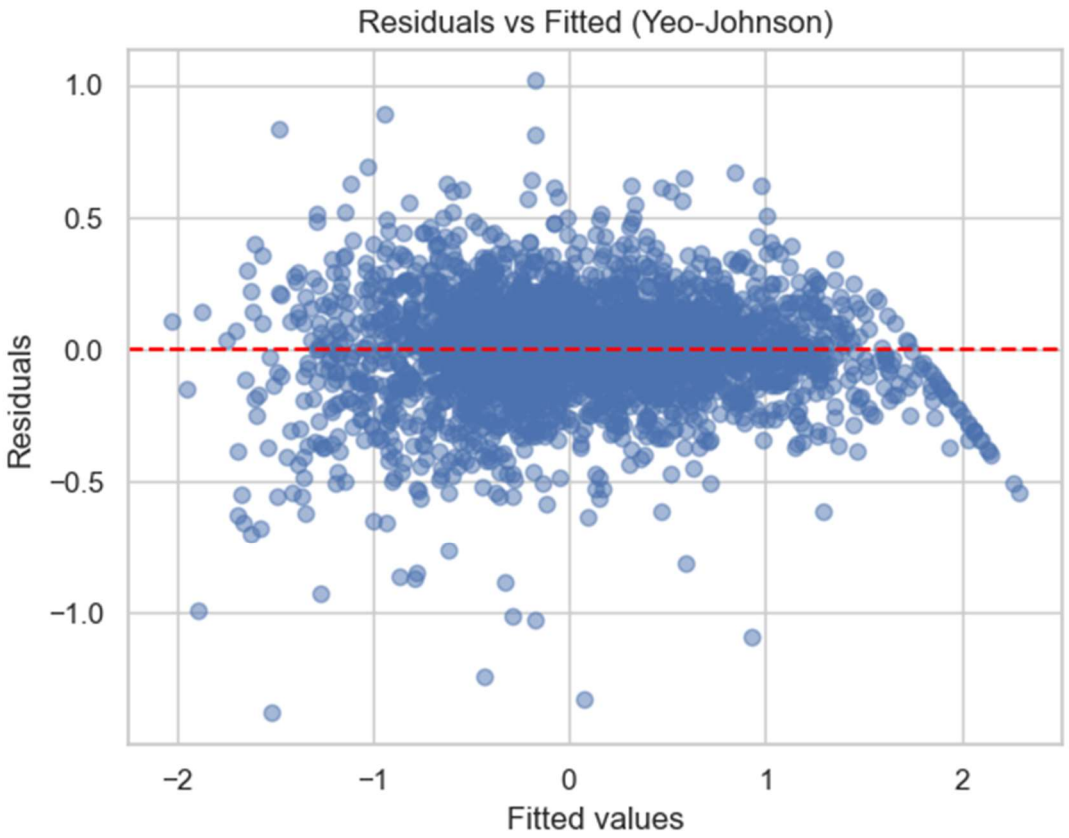
### 2.1. Transformation Method

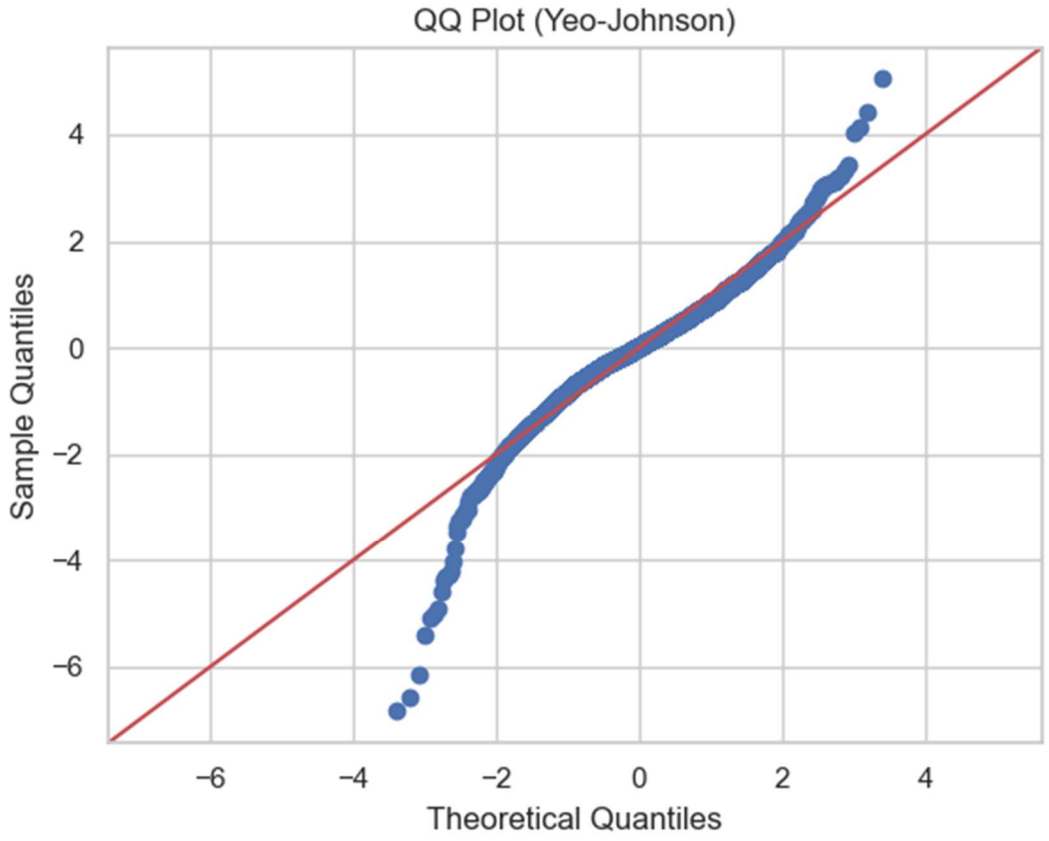
- **Objective:** To determine whether applying a Yeo-Johnson transformation to the SalePrice_capped response variable improves variance stability, satisfies linear regression assumptions, and maintains predictive accuracy.
- **Model Refit:** Linear regression was refitted on transformed y
- **Diagnostics Re-assessed:** Residual plots, QQ plots, Cook’s distance, Shapiro–Wilk test, and VIF were recomputed. Cross-validation (5-fold) was applied to measure predictive performance

### 2.2. Evaluation and Assumption Validation

- **Residual Diagnostics:**
- **Residuals vs Fitted**: Funnel shape largely eliminated. Residuals evenly scattered around zero; range reduced to approximately -1 to 1.
- **QQ Plot:** More points aligned on the 45° line; normality improved, particularly in the upper quantiles.

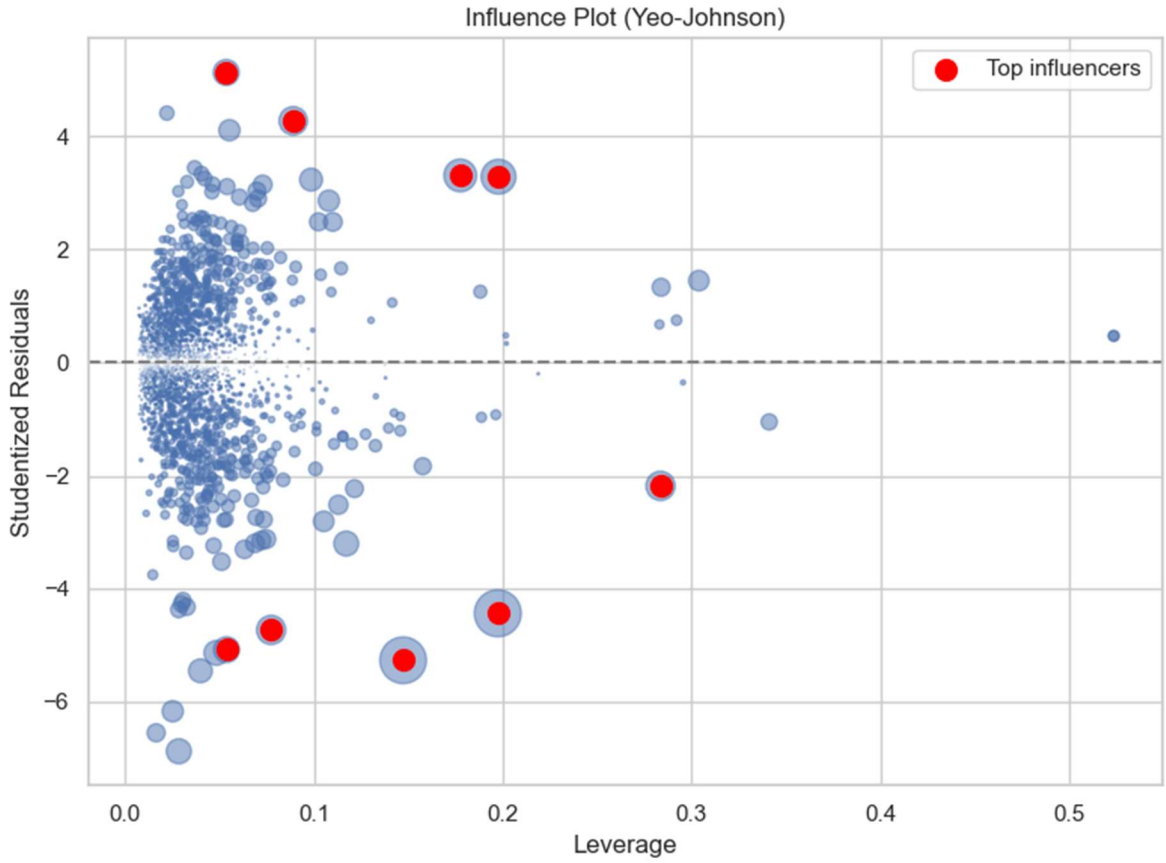
- **Cook's Distance:** Threshold = 0.0014; 236 influential points identified. Largest leverage = 0.22; studentized residuals ranged from -7 to 7, showing a wider range than raw residuals due to scaling by each point's standard deviation and leverage. Influence plot similar to previous iteration, with points concentrated at lower leverage values.
- **Shapiro–Wilk Test: Statistic** = 0.9915, p = 1.2e-08. Residuals are closer to normal compared with untransformed model.






- Cook's distance threshold: 0.0014
- Number of influential points: 236
- Top 10 influential points:

| Index | Cook's Distance | Target |
|---|---|---|
| 1356 | NaN | 0.000000 |
| 782 | 0.047715 | -0.202381 |
| 1553 | 0.047029 | -1.748810 |
| 241 | 0.026186 | 1.762262 |
| 1945 | 0.023245 | 1.685714 |
| 432 | 0.018526 | 3.531743 |
| 1555 | 0.018483 | -1.428571 |
| 2881 | 0.017534 | -0.535714 |
| 1407 | 0.014584 | 1.190476 |
| 2298 | 0.014414 | -0.952381 |




- **Variance Inflation Factors (post transformation):** No change; top multicollinear features remain:

| Feature | VIF |
|---|---|
| Bldg Type_Duplex | inf |
| MS SubClass_90 | inf |
| house_age | 847.4285 |
| years_to_remodel | 661.6787 |
| remodel_age | 444.5772 |
| Garage Finish_None | 236.0888 |
| Garage Type_None | 234.8742 |
| total_bathrooms | 52.6536 |
| total_porch_area | 42.7543 |
| Sale Type_New | 39.7537 |

- **Performance Metrics**
  - Predictive performance remains strong; Test $R^2$ slightly decreased compared to post-outlier removal but still exceeds original Model 2.
  - RMSE and MAE decreased in absolute terms, indicating tighter residuals and improved overall fit.
  - Transformation successfully addressed heteroscedasticity and improved normality of residuals.

| Metric | Post-Removal | Yeo-Johnson |
|---|---|---|
| Train $R^2$ | 0.9330 | 0.9209 |
| Test $R^2$ | 0.9249 | 0.9110 |
| Train RMSE | 0.2345 | 0.2006 |
| Test RMSE | 0.2475 | 0.2123 |
| Train MAE | 0.1703 | 0.1445 |
| Test MAE | 0.1784 | 0.1522 |

### 2.3. Feature Importance Analysis

| Rank | Model 2 (All Data) | Coefficient | Outlier-Removed Model | Coefficient | Yeo–Johnson Model | Coefficient |
|---|---|---|---|---|---|---|
| 1 | qual_living_area_interaction | 0.619 | qual_living_area_interaction | 0.553 | qual_living_area_interaction | 0.360 |
| 2 | Neighborhood_NridgHt | 0.434 | Exterior 2nd_CmentBd | 0.427 | MS SubClass_160 | -0.337 |
| 3 | Exterior 2nd_CmentBd | 0.377 | Exterior 1st_CemntBd | -0.323 | Neighborhood_MeadowV | -0.299 |
| 4 | Neighborhood_StoneBr | 0.377 | MS SubClass_190 | -0.308 | MS Zoning_C (all) | -0.254 |
| 5 | Neighborhood_NoRidge | 0.343 | Neighborhood_NridgHt | 0.284 | Exterior 2nd_CmentBd | 0.209 |
| 6 | Neighborhood_Somerst | 0.314 | Neighborhood_StoneBr | 0.273 | Exterior 1st_CemntBd | -0.205 |
| 7 | Exterior 1st_CemntBd | -0.308 | Neighborhood_Somerst | 0.221 | Sale Condition_Partial | 0.201 |
| 8 | MS SubClass_190 | -0.285 | Neighborhood_NoRidge | 0.218 | Central Air_Y | 0.170 |
| 9 | Sale Condition_Partial | 0.271 | BsmtFin SF 1 | 0.217 | Neighborhood_Somerst | 0.149 |
| 10 | Condition 2_PosA | 0.263 | Foundation_Slab | 0.203 | MS SubClass_60 | -0.144 |

- **Shift in Top Features:** The Yeo-Johnson transformation with outlier removal has slightly altered the top 10 features. Some structural and zoning variables (e.g., MS SubClass_160, Neighborhood_MeadowV, MS Zoning_C) now appear more prominently, while some previously strong predictors like MS SubClass_190, Neighborhood_NridgHt, and BsmtFin SF 1 have dropped out of the top 10.
- **Magnitude Changes:** Coefficients in the Yeo-Johnson model are generally smaller, reflecting the rescaling effect of the transformation on the response.
- **Interpretation:** The transformation improved variable distributions and residual homoscedasticity but also slightly shifted the relative importance of predictors. This indicates that the transformation can influence the model's sensitivity to certain features, particularly categorical structural variables.
- **General Across Models:** Despite these changes, core predictors like qual_living_area_interaction and major exterior or structural features remain consistently important across all iterations, highlighting their robust influence on SalePrice_capped

### 2.4. Feature Refinement and Response Transformation Refinement Summary

- Applying the Yeo-Johnson transformation stabilized variance across fitted values and produced a more symmetric residual distribution.
- Model assumptions are better satisfied, improving confidence in coefficient estimates and hypothesis testing.
- Although multicollinearity persists, predictive accuracy remains high; regularization will be addressed in subsequent iterations.

- Slight compromise on $R^2$ is justified by substantial gains in homoscedasticity and residual stability.
- Importantly, the Yeo–Johnson transformation combined with outlier removal slightly shifted the ranking of the top predictive features: some structural and zoning variables (e.g., MS SubClass_160, Neighborhood_MeadowV, MS Zoning_C) now feature more prominently, while previously dominant predictors such as MS SubClass_190 and Neighborhood_NridgHt dropped slightly in relative importance.
- Overall, core drivers like qual_living_area_interaction and major exterior or structural features remain consistently important, confirming that the model continues to capture the key determinants of SalePrice_capped while benefiting from improved residual symmetry and stabilized variance.
- **Recommendation:** Retain the Yeo-Johnson transformed model for subsequent analysis. This balances predictive performance and adherence to linear regression assumptions, ensuring reliable predictions while maintaining interpretability.

## 3. Stepwise Feature Selection

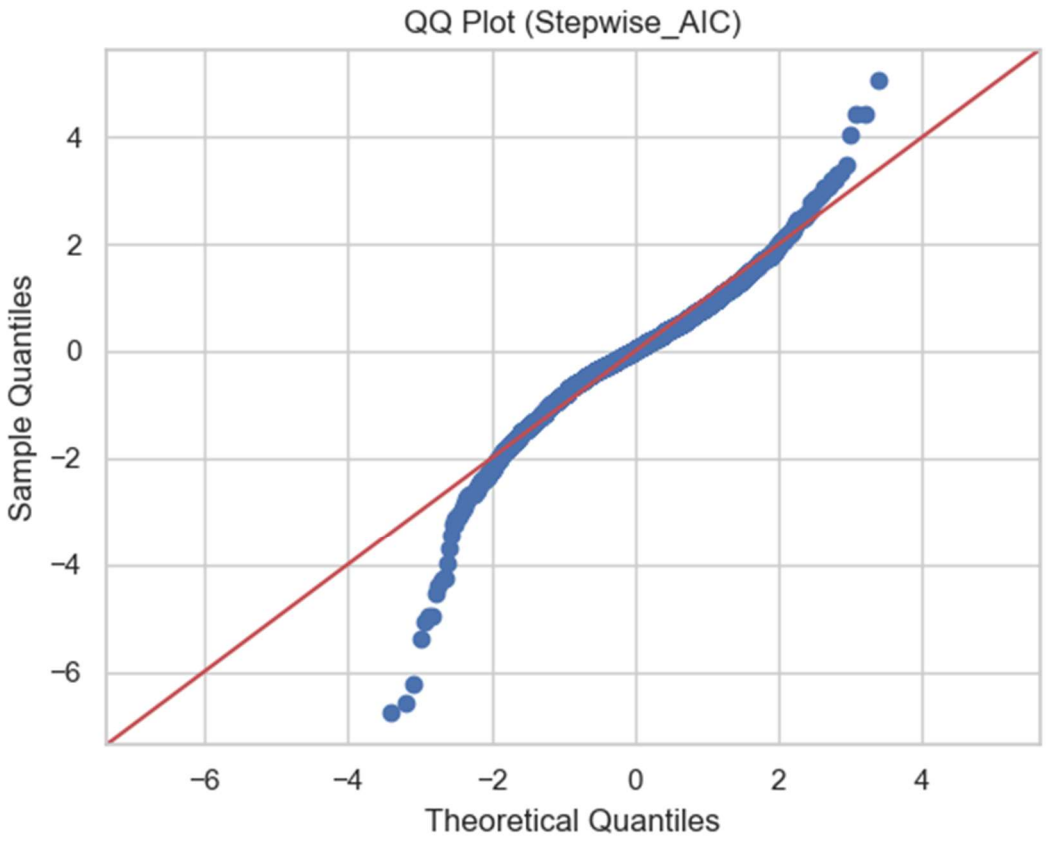
- **Purpose:** With ~100 predictors, stepwise selection provides a systematic way to simplify the model while retaining predictive power. It allows identification of the most relevant features, flagging of potential multicollinearity issues, and evaluation of the influence of outliers before moving to regularization.
- **Method**: Stepwise regression iteratively adds or removes predictors based on an information criterion, such as AIC or BIC. Forward selection starts with no predictors and adds the most informative feature at each step. Backward elimination starts with all predictors and removes the least informative one iteratively. Bidirectional stepwise regression combines both approaches, allowing features to be added or removed at each iteration.
- **Objective:** The stepwise models serve as a **diagnostic check**. They reveal which pre-existing features (including polynomial and interaction terms) are most influential, highlight potential multicollinearity and outliers, and provide a baseline understanding of the model's behavior before applying regularization methods such as Ridge, Lasso, or Elastic Net

### 3.1. AIC Evaluation and Assumption Validation

- **Residual Diagnostics:**
    - **Residuals vs Fitted**: The residuals are less dense than in the Yeo Johnson model and centred around 0. Residual values range from approximately -2 to 1, and fitted values range from -2 to 2. Most residuals are concentrated near 0, with a spread that increases slightly on the left side. Overall, the model fit is reasonable, and the residuals show only mild deviation from normality.
    - **QQ Plot:** Sample quantiles are very similar to the Yeo-Johnson model, indicating residuals are approximately normal aside from a few extreme low-end points.
    - **Cook's Distance**: Threshold = 0.0014; 233 influential points identified. The highest Cook's D is 0.049, well below 1, indicating no single point dominates the regression. Influential points include extreme low and high target values, but with moderate leverage, so their effect on coefficients is limited. Most points are far below the threshold, showing the model is robust to outliers. Compared with Yeo-Johnson, the

Cook's D distribution is slightly tighter, reflecting a more even influence across observations.

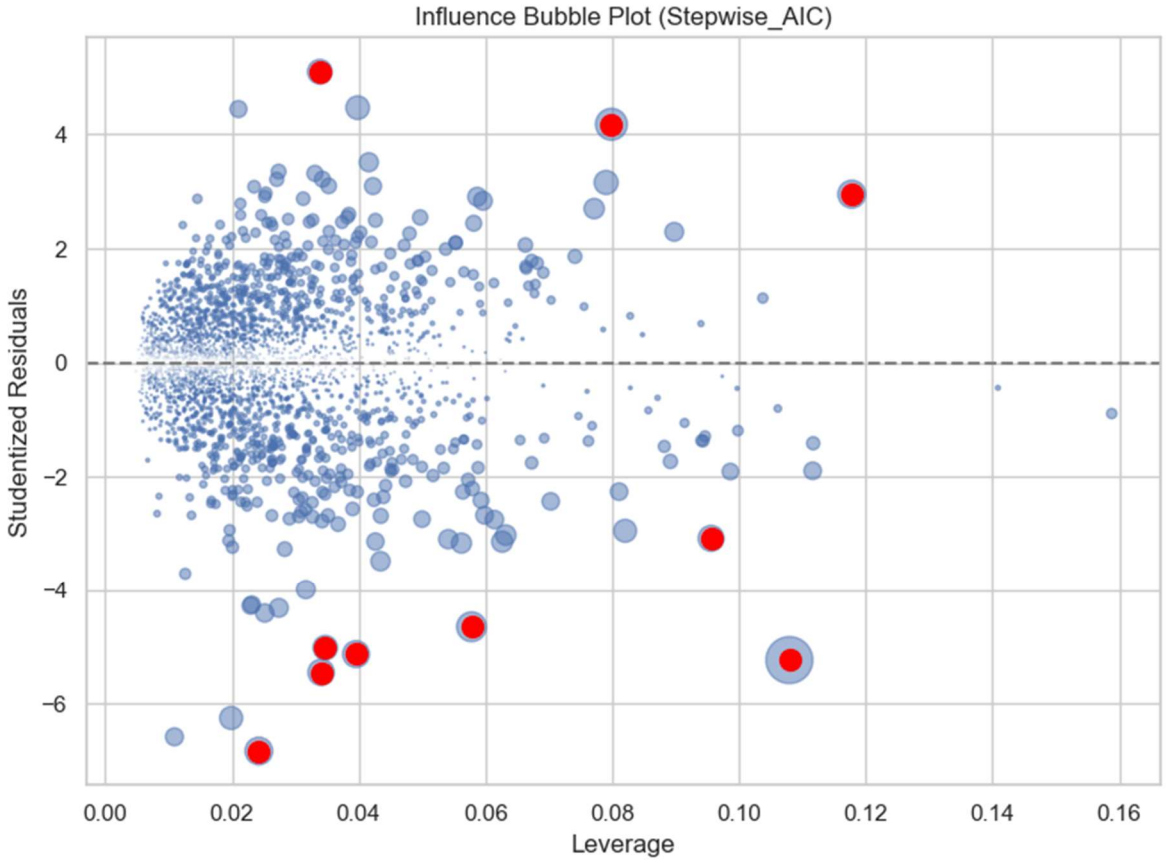
- **Shapiro–Wilk Test:** 0.9575, p = 1.97e-28, indicating residuals are closer to normality than in the original model but slightly deviated due to low-end outliers.
- **Multicollinearity (VIF) -** Drastic improvement. Highest VIF is 41.89 (MS Zoning_RL), followed by 20.68 (Central Air_Y) and 17.96 (Sale Type_WD). Remaining predictors have moderate correlation.




- Cook's distance threshold: 0.0014
- Number of influential points: 233
- Top 10 influential points (Stepwise_AIC):

| Index | Cook's Distance | Target (Residualized) |
|---|---|---|
| 1553 | 0.048974 | -2.888 |
| 2881 | 0.022485 | -0.648 |
| 1555 | 0.019626 | -2.198 |
| 1945 | 0.017331 | 1.088 |
| 181 | 0.017013 | -2.897 |
| 125 | 0.016021 | -1.202 |
| 1782 | 0.015464 | -0.164 |
| 2730 | 0.014906 | -0.146 |

| | | |
|---|---|---|
| 1407 | 0.013480 | 0.846 |
| 2298 | 0.013384 | -1.301 |




| Feature | VIF |
|---|---|
| MS Zoning_RL | 41.886406 |
| Central Air_Y | 20.683111 |
| Sale Type_WD | 17.959372 |
| Sale Condition_Normal | 11.227537 |
| MS Zoning_RM | 11.048847 |
| qual_living_area_interaction | 11.008145 |
| Garage Cars Mo Sold avg_quality | 10.706381 |
| Foundation_PConc | 10.464856 |
| House Style_2Story | 10.250826 |
| Garage Type_Attchd | 9.228414 |

- **Performance Metrics**

| Metric | Yeo-Johnson | AIC |
|---|---|---|
| Train $R^2$ | 0.9209 | 0.9197 |
| Test $R^2$ | 0.911 | 0.9137 |
| Train RMSE | 0.2006 | 0.2021 |
| Test RMSE | 0.2123 | 0.2090 |
| Train MAE | 0.1445 | 0.1455 |

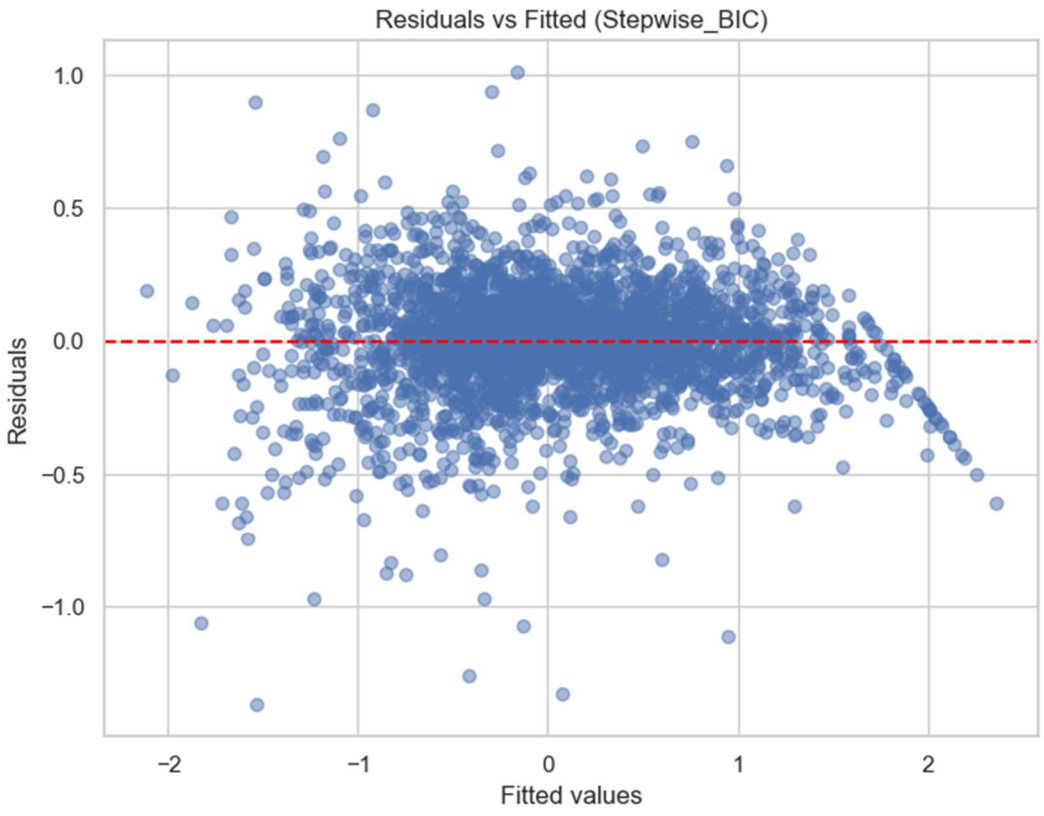
| | | |
|---|---|---|
| Test MAE | 0.1522 | 0.1501 |

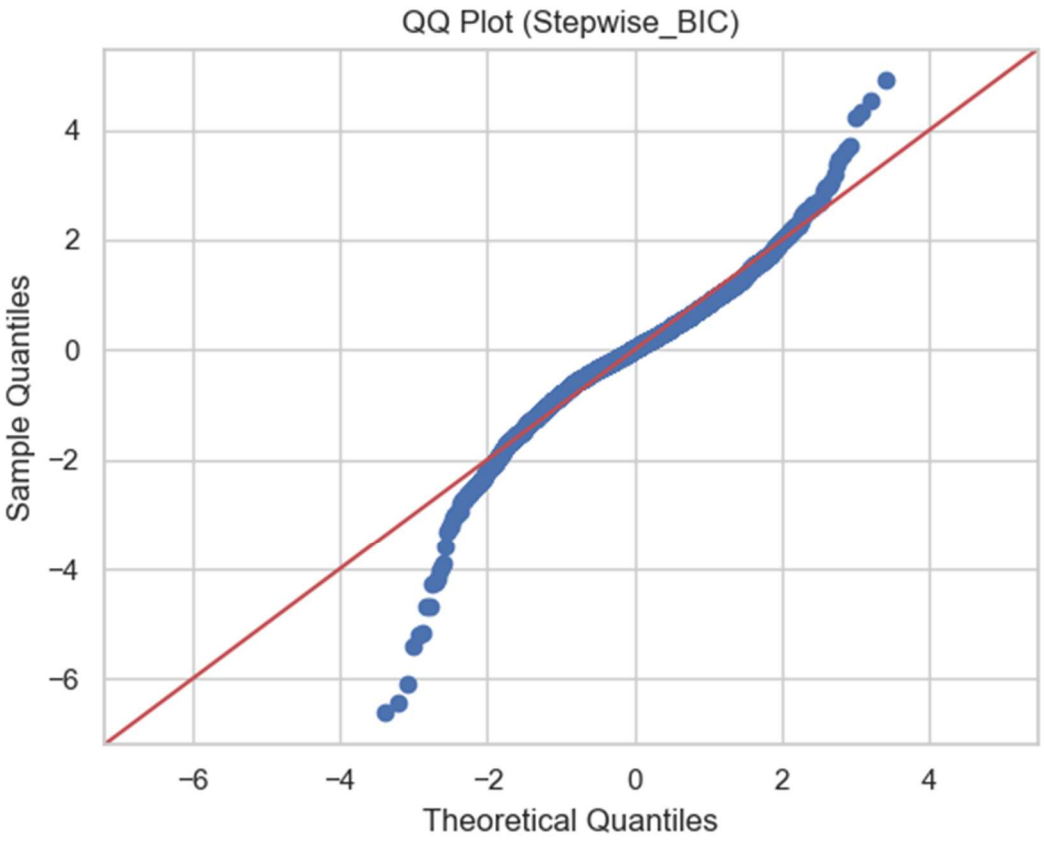
- **Model fit is very similar:** Both Yeo-Johnson and Stepwise AIC have comparable Train and Test $R^2$ values (~0.92 for training, ~0.91–0.913 for testing), indicating strong predictive performance.
- **Error metrics are close:** Train and Test RMSE/MAE are nearly identical, with Stepwise AIC showing slightly lower Test RMSE (0.209 vs 0.212) and Test MAE (0.150 vs 0.152), suggesting marginally better generalization.
- **Stepwise AIC slightly more balanced:** Overall, Stepwise AIC achieves nearly the same fit as Yeo-Johnson while keeping training and testing errors very consistent, indicating it handles outliers well without overfitting.

- **Summary:**
  - **Residuals:** Centered around 0, less dense than Yeo-Johnson, mild deviation from normality, ranges approx. -2 to 1 (residuals) and -2 to 2 (fitted).
  - **Influence:** Cook's D threshold 0.0014; 233 points influential but all moderate leverage (max 0.049), slightly tighter than Yeo-Johnson.
  - **Multicollinearity:** Drastically improved; highest VIFs: MS Zoning_RL 41.89, Central Air_Y 20.68, Sale Type_WD 17.96.
  - **Performance:** Train $R^2$ ~0.92, Test $R^2$ ~0.914; slightly lower Test RMSE (0.209) and MAE (0.150) than Yeo-Johnson, indicating good generalization.

### 3.2. BIC Evaluation and Assumption Validation

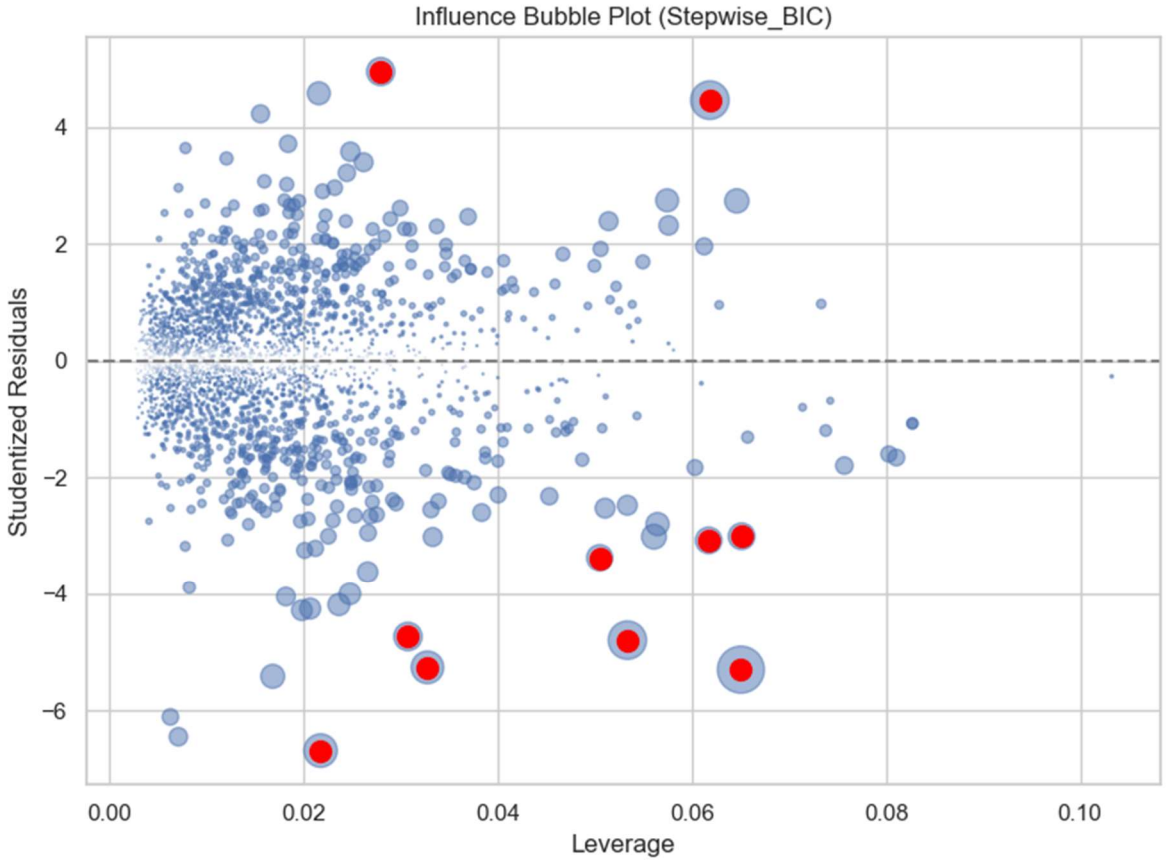
- **Residual Diagnostics:**
  - **Residuals vs Fitted**: Similar to Stepwise AIC. Residuals range roughly from -2 to 1, fitted values from -2 to 2, with residuals centered around 0 and slightly looser spread than Yeo-Johnson. A few low-end outliers exist, but the overall pattern shows balanced dispersion without extreme leverage effects.
  - **QQ Plot:** Matches Stepwise AIC and Yeo-Johnson plots, confirming approximate normality
  - **Cook's Distance:** Threshold = 0.0014; 235 influential points identified. Highest Cook's D is 0.046, slightly lower than AIC, again indicating no extreme leverage. Top influencers show a mix of extreme low and high target values, similar to AIC, but the spread is a bit tighter, with fewer moderate Cook's D values above 0.02. Most points remain far below the threshold, confirming robustness.
  - **Shapiro–Wilk Test:** 0.9581, p = 3.08e-28. Residuals slightly deviate from normality but remain comparable to AIC.
  - **Multicollinearity (VIF) -** Identical to Stepwise AIC. Top VIF remains 41.89 (MS Zoning_RL) and other predictors show moderate correlation


Residuals vs Fitted (Stepwise_BIC)
Residuals
Fitted values
1.0
0.5
0.0
−0.5
−1.0
−2
−1
0
1
2


QQ Plot (Stepwise_BIC)
Sample Quantiles
Theoretical Quantiles
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4

- Cook's distance threshold: 0.0014
- Number of influential points: 235
- Top 10 influential points (Stepwise_BIC):

| Index | Cook's Distance | Target |
|---|---|---|
| 1553 | 0.046212 | -2.888197 |
| 2881 | 0.031060 | -0.648166 |
| 1555 | 0.030724 | -2.197789 |
| 181 | 0.023305 | -2.896581 |
| 125 | 0.022161 | -1.201991 |
| 2298 | 0.016845 | -1.300925 |
| 1407 | 0.016638 | 0.846297 |
| 423 | 0.014982 | 1.753460 |
| 2730 | 0.014818 | -0.145732 |
| 2880 | 0.014402 | -2.313157 |




| Feature | VIF |
|---|---|
| MS Zoning_RL | 41.886406 |
| Central Air_Y | 20.683111 |
| Sale Type_WD | 17.959372 |
| Sale Condition_Normal | 11.227537 |
| MS Zoning_RM | 11.048847 |
| qual_living_area_interaction | 11.008145 |

| | |
|---|---|
| Garage Cars Mo Sold avg_quality | 10.706381 |
| Foundation_PConc | 10.464856 |
| House Style_2Story | 10.250826 |
| Garage Type_Attchd | 9.228414 |

- **Performance Metrics**

| Metric | Yeo-Johnson | AIC | BIC |
|---|---|---|---|
| Train $R^2$ | 0.9209 | 0.9197 | 0.9165 |
| Test $R^2$ | 0.9110 | 0.9137 | 0.9126 |
| Train RMSE | 0.2006 | 0.2021 | 0.2060 |
| Test RMSE | 0.2123 | 0.2090 | 0.2102 |
| Train MAE | 0.1445 | 0.1455 | 0.1486 |
| Test MAE | 0.1522 | 0.1501 | 0.1514 |

  - BIC yields a slightly more conservative model, with marginally lower Test $R^2$ (0.9126) compared to Yeo-Johnson (0.911) and AIC (0.9137). This reflects BIC's stronger penalty for additional predictors, favoring simpler, more parsimonious models while maintaining most of the predictive performance..
  - **Higher training errors:** Train RMSE (0.206) and Train MAE (0.149) are slightly higher than the other models, reflecting its more restrained fit on the training data.
  - **Comparable generalization:** Test RMSE (0.210) and Test MAE (0.151) remain similar to the other models, suggesting BIC still generalizes well despite being simpler.
  - **Balanced and robust:** Overall, BIC achieves nearly the same predictive performance while using fewer variables, making it more parsimonious and less sensitive to extreme points.
- **Summary**
  - **Residuals & Normality:** Residuals range -2 to 1, fitted values -2 to 2, centred around 0 with a slightly looser spread than Yeo-Johnson; QQ plot and Shapiro-Wilk (0.9581, p=3.08e-28) indicate approximate normality.
  - **Influential Points:** Cook's D threshold = 0.0014; 235 points identified, highest 0.046, showing no extreme leverage and robust influence distribution.
  - **Multicollinearity:** VIFs identical to AIC, top predictor MS Zoning_RL = 41.89; remaining predictors show moderate correlation.
  - **Performance & Fit:** BIC is slightly more conservative, with marginally higher training errors (Train RMSE 0.206, MAE 0.149) but similar Test metrics ($R^2$ 0.9126, RMSE 0.210, MAE 0.151), achieving nearly the same predictive performance with fewer variables.

### 3.3. Feature Importance Analysis

| Rank | Model 2 (All Data) | Coefficient | Outlier-Removed Model | Coefficient | Yeo–Johnson Model | Coefficient | Stepwise BIC | Coefficient |
|---|---|---|---|---|---|---|---|---|

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1 | qual_living_area_interaction | 0.619 | qual_living_area_interaction | 0.553 | qual_living_area_interaction | 0.360 | qual_living_area_interaction | 0.388 |
| 2 | Neighborhood_NridgHt | 0.434 | Exterior 2nd_CmentBd | 0.427 | MS SubClass_160 | -0.337 | Neighborhood_MeadowV | -0.316 |
| 3 | Exterior 2nd_CmentBd | 0.377 | Exterior 1st_CemntBd | -0.323 | Neighborhood_MeadowV | -0.299 | MS Zoning_C (all) | -0.315 |
| 4 | Neighborhood_StoneBr | 0.377 | MS SubClass_190 | -0.308 | MS Zoning_C (all) | -0.254 | MS SubClass_160 | -0.304 |
| 5 | Neighborhood_NoRidge | 0.343 | Neighborhood_NridgHt | 0.284 | Exterior 2nd_CmentBd | 0.209 | Sale Condition_Partial | 0.235 |
| 6 | Neighborhood_Somerst | 0.314 | Neighborhood_StoneBr | 0.273 | Exterior 1st_CemntBd | -0.205 | Central Air_Y | 0.181 |
| 7 | Exterior 1st_CemntBd | -0.308 | Neighborhood_Somerst | 0.221 | Sale Condition_Partial | 0.201 | Sale Condition_Normal | 0.144 |
| 8 | MS SubClass_190 | -0.285 | Neighborhood_NoRidge | 0.218 | Central Air_Y | 0.170 | Garage Type_None | -0.139 |
| 9 | Sale Condition_Partial | 0.271 | BsmtFin SF 1 | 0.217 | Neighborhood_Somerst | 0.149 | MS SubClass_30 | -0.138 |
| 10 | Condition 2_PosA | 0.263 | Foundation_Slab | 0.203 | MS SubClass_60 | -0.144 | Kitchen AbvGr | -0.137 |

- **Dominant Predictors:** qual_living_area_interaction consistently ranks first across all iterations, confirming its central role in explaining SalePrice_capped.
- **Neighborhood Effects:** Neighborhood-related features remain important, though their relative influence shifts slightly across models. In the Stepwise BIC model, Neighborhood_MeadowV and MS SubClass_160 enter the top 10, reflecting refined importance after collinearity reduction and feature selection.
- **Structural Features:** Core structural attributes (e.g., exterior material, basement finish, foundation, kitchen size, central air) rise in prominence in the outlier-removed, Yeo–Johnson, and Stepwise BIC models, highlighting stable, generalizable drivers of price beyond rare or extreme cases.
- **Stepwise BIC Refinement:** The BIC-selected model emphasizes parsimony, favouring a leaner set of highly predictive and interpretable features. Transactional and property-type features (e.g., Sale Condition_Partial, MS Zoning_C) enter the top 10, demonstrating the model's ability to prioritize informative predictors once less relevant or collinear variables are removed.
- **Coefficient Stability:** Across all iterations, feature signs remain largely consistent. Coefficients are moderated by response transformation and outlier handling, improving interpretability and model robustness. Stepwise BIC selection further stabilizes the set of influential predictors.
- **Summary:** The evolution of feature importance illustrates how model refinements—outlier treatment, response transformation, and stepwise selection—shift the focus from rare or extreme observations toward consistent structural, neighbourhood, and property-type characteristics, supporting both predictive accuracy and interpretability.

### 3.4. Overall Stepwise Summary

- **Residuals & Normality:** Both Stepwise and the Yeo Johnson models show approximately normal residuals in QQ plots. Stepwise AIC and BIC have slightly looser residual spreads than Yeo-Johnson but remain well-centred around 0.
- **Influence & Leverage:** BIC has fewer moderate leverage points than AIC. Both stepwise models are slightly more robust to low-end outliers compared with Yeo-Johnson, reducing extreme influence from individual points.
- **Predictive Performance:** Train and Test $R^2$ are comparable across all three models (AIC, BIC and Yeo Johnson: ~0.92 for training, ~0.91–0.913 for testing). Stepwise models have slightly lower training errors than Yeo-Johnson, with BIC being marginally more conservative.
- **Parsimony & Robustness:** BIC achieves nearly the same predictive performance as the other models while using fewer variables. Its more restrained fit makes it less sensitive to extreme points, reflected in a tighter influence distribution and consistent error metrics. Feature importance analysis confirms that the BIC model retains the most predictive and interpretable structural and neighbourhood features, emphasizing stable drivers of SalePrice_capped while eliminating less informative or collinear variables.
- **Overall Decision:** While all three models perform similarly, the Stepwise BIC model stands out for its balance of predictive accuracy, reduced leverage influence, and parsimony. Therefore the BIC-selected model is retained and used as the basis for the next model iteration.

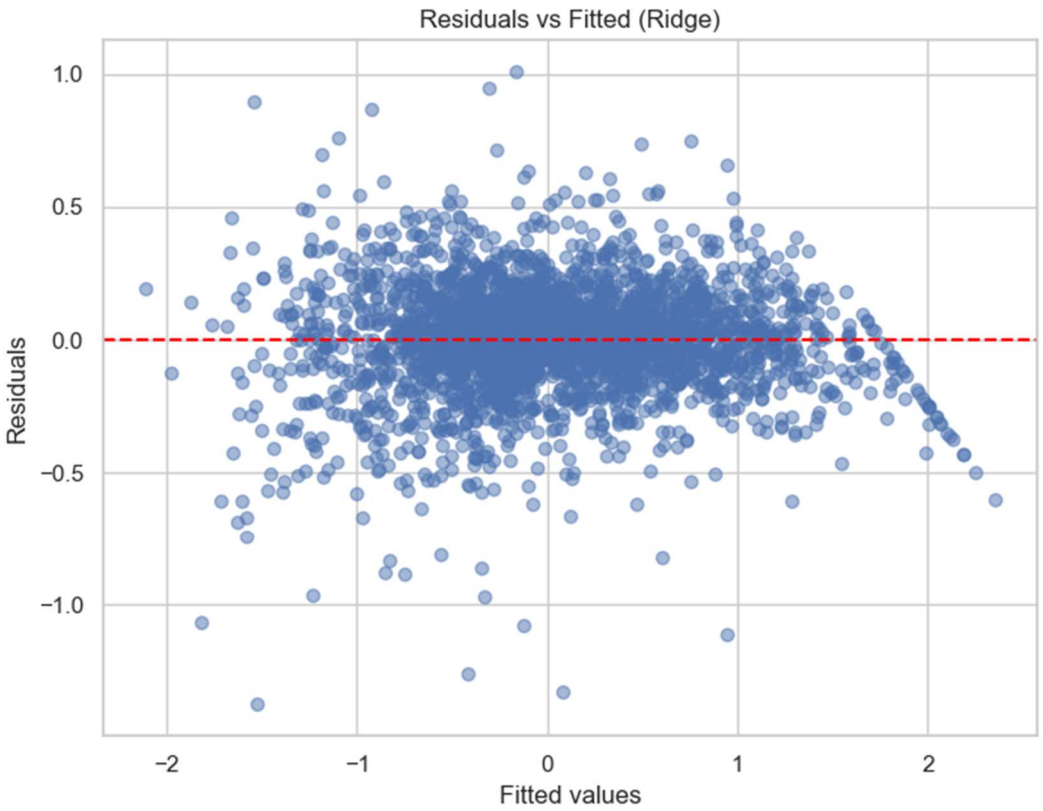
## 4. Regularization (Ridge / Lasso / Elastic Net)

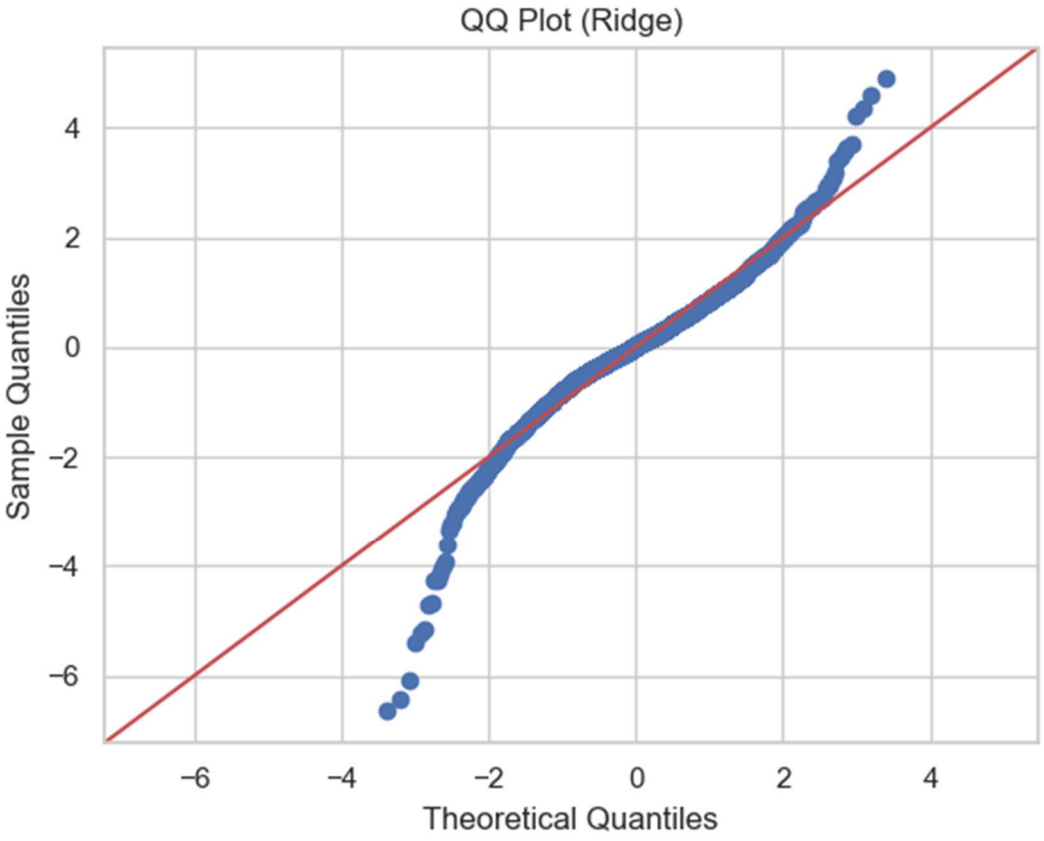
- **Purpose / Relevance:** Regularization techniques, including Ridge, Lasso, and ElasticNet, are employed to address multicollinearity and reduce the risk of overfitting when modeling with a relatively large set of correlated predictors. By introducing a penalty on coefficient sizes, these methods constrain the model, helping to stabilize predictions and improve generalizability to unseen data. Lasso and ElasticNet, in particular, can shrink some coefficients to zero, providing a form of automatic feature selection and promoting a more parsimonious model. This makes regularization highly relevant for the BIC-selected feature set, where a smaller number of influential variables already exists but correlated features could still distort traditional OLS estimates.
- **Methods**:
    - **Ridge Regression (L2 penalty):** Useful for reducing variance when predictors are highly correlated, shrinks coefficients but does not perform feature selection.
    - **Lasso Regression (L1 penalty):** Shrinks some coefficients to exactly zero, providing automatic feature selection.
    - **Elastic Net Regression (L1 + L2 penalty):** Balances Ridge and Lasso effects, particularly useful for correlated predictors.
- **Implementation Strategy:**
    - Cross-validated GridSearch used to tune hyperparameters (alpha for Ridge/Lasso, alpha and l1_ratio for ElasticNet)
    - Models fitted on Yeo-Johnson-preprocessed features, ensuring regularization addressed multicollinearity and outlier influence..

- Evaluation included performance metrics ($R^2$, RMSE, MAE) and diagnostics (residual plots, QQ plots, Cook's distance, VIFs)

### 4.1. Regularisation Evaluation and Assumption Validation

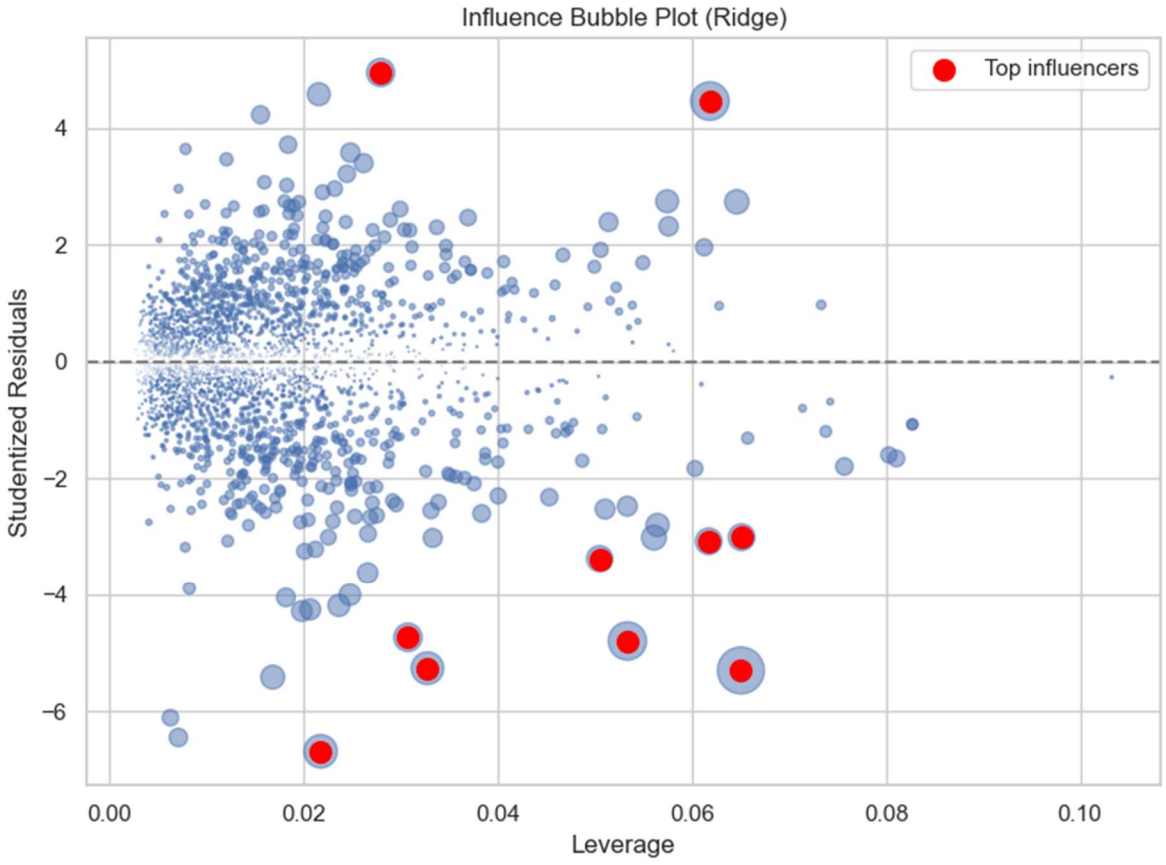
- **Residual Diagnostics**
  - **Residual vs Fitted:** Nearly identical to the Stepwise BIC models. Regularisation has stabilized coefficients without altering predictive accuracy
  - **QQ Plots** – Identical to the Stepwise BIC model, indicating normality is unchanged.
  - **Influential Points (Cook's Distance):** All three models show the same 235 influential points as Stepwise BIC; Regularisation does not address influence points.
  - **Shapiro–Wilk Normality Test:** Very close to Stepwise BIC baseline (~0.957)
  - **Variance Inflation Factors (VIFs):** Significantly reduced, with top VIF at 9.31 and most others under 6, showing effective control of multicollinearity.






- Cook's distance threshold: 0.0014
- Number of influential points: 235
- Top 10 influential points (Ridge):

| index | Cook's Distance | target |
|---|---|---|
| 1553 | 0.046212 | -2.888197 |
| 2881 | 0.031060 | -0.648166 |
| 1555 | 0.030724 | -2.197789 |
| 181 | 0.023305 | -2.896581 |
| 125 | 0.022161 | -1.201991 |
| 2298 | 0.016845 | -1.300925 |
| 1407 | 0.016638 | 0.846297 |
| 423 | 0.014982 | 1.753460 |
| 2730 | 0.014818 | -0.145732 |
| 2880 | 0.014402 | -2.313157 |




| Feature | VIF |
|---|---|
| qual_living_area_interaction | 9.311305 |
| Total Bsmt SF | 6.340397 |
| MS SubClass_60 | 5.943589 |
| House Style_2Story | 5.643816 |
| Garage Cars Bsmt Unf SF avg_quality | 5.604685 |
| BsmtFin SF 1 | 4.996992 |
| living_area_ratio avg_quality^2 | 4.767918 |
| house_age | 4.612149 |
| Overall Qual Exter Qual Fireplace Qu | 4.025758 |
| total_porch_area | 3.669426 |

- **Performance Metrics**

| Metric | Stepwise | Ridge | Lasso | ElasticNet |
|---|---|---|---|---|
| Train $R^2$ | 0.9165 | 0.9165 | 0.9164 | 0.9165 |
| Test $R^2$ | 0.9126 | 0.9126 | 0.9124 | 0.9126 |
| Train RMSE | 0.2061 | 0.2061 | 0.2062 | 0.2061 |
| Test RMSE | 0.2102 | 0.2102 | 0.2105 | 0.2103 |

| | | | | |
|---|---|---|---|---|
| Train MAE | 0.1486 | 0.1486 | 0.1485 | 0.1486 |
| Test MAE | 0.1514 | 0.1514 | 0.1514 | 0.1514 |

### 4.2. Feature Importance Analysis

| Rank | Model 2 (All Data) | Coefficient | Outlier-Removed Model | Coefficient | Yeo–Johnson Model | Coefficient | Stepwise BIC | Coefficient | Ridge Regularization | Coefficient |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | qual_living_area_interaction | 0.619 | qual_living_area_interaction | 0.553 | qual_living_area_interaction | 0.360 | qual_living_area_interaction | 0.388 | qual_living_area_interaction | 0.293 |
| 2 | Neighborhood_NridgHt | 0.434 | Exterior 2nd_CmentBd | 0.427 | MS SubClass_160 | -0.337 | Neighborhood_MeadowV | -0.316 | house_age | -0.084 |
| 3 | Exterior 2nd_CmentBd | 0.377 | Exterior 1st_CemntBd | -0.323 | Neighborhood_MeadowV | -0.299 | MS Zoning_C (all) | -0.315 | BsmtFin SF 1 | 0.074 |
| 4 | Neighborhood_StoneBr | 0.377 | MS SubClass_190 | -0.308 | MS Zoning_C (all) | -0.254 | MS SubClass_160 | -0.304 | Lot Area_capped | 0.071 |
| 5 | Neighborhood_NoRidge | 0.343 | Neighborhood_NridgHt | 0.284 | Exterior 2nd_CmentBd | 0.209 | Sale Condition_Partial | 0.235 | remodel_age | -0.070 |
| 6 | Neighborhood_Somerst | 0.314 | Neighborhood_StoneBr | 0.273 | Exterior 1st_CemntBd | -0.205 | Central Air_Y | 0.181 | Sale Condition_Partial | 0.064 |
| 7 | Exterior 1st_CemntBd | -0.308 | Neighborhood_Somerst | 0.221 | Sale Condition_Partial | 0.201 | Sale Condition_Normal | 0.144 | MS SubClass_160 | -0.062 |
| 8 | MS SubClass_190 | -0.285 | Neighborhood_NoRidge | 0.218 | Central Air_Y | 0.170 | Garage Type_None | -0.139 | Sale Condition_Normal | 0.054 |
| 9 | Sale Condition_Partial | 0.271 | BsmtFin SF 1 | 0.217 | Neighborhood_Somerst | 0.149 | MS SubClass_30 | -0.138 | living_area_ratio avg_quality^2 | 0.049 |
| 10 | Condition 2_PosA | 0.263 | Foundation_Slab | 0.203 | MS SubClass_60 | -0.144 | Kitchen AbvGr | -0.137 | total_bathrooms | 0.048 |

- **Dominant Predictors:** qual_living_area_interaction consistently ranks as the most influential predictor across all models, including the Ridge regularization. While its coefficient magnitude is smaller under Ridge (reflecting shrinkage), its dominance confirms that overall living area adjusted for quality remains the key driver of capped sale prices.
- **Neighborhood Effects:** Neighbourhood-related predictors (e.g., NridgHt, StoneBr, NoRidge, Somerst, MeadowV) are highly influential in the unrestricted and outlier-removed models, but their importance diminishes in the Ridge model. This suggests that while location matters, penalization reduces over-reliance on neighborhood dummies that may capture extreme or collinear effects.

- **Structural Features:** Ridge regularization elevates more stable structural features such as BsmtFin SF 1, Lot Area_capped, total_bathrooms, and remodel_age. These are intuitive, interpretable drivers of value that generalize better across the dataset compared with rare exterior or zoning categories, which were prominent earlier.
- **Ridge Regularization Refinement:** Ridge regularization stabilizes the feature set by shrinking coefficients toward zero, reducing the impact of collinear predictors while retaining all variables. This shifts importance away from niche categorical effects and toward robust, generalizable structural and transactional features (e.g., house_age, BsmtFin SF 1, total_bathrooms), ensuring interpretability without sacrificing predictive performance.
- **Coefficient Stability:** Coefficient magnitudes decline under Ridge due to penalization, but sign consistency is preserved across most iterations (e.g., qual_living_area_interaction positive, MS SubClass_160 negative). Regularization mitigates coefficient inflation from collinear variables and tempers extreme swings observed in earlier models.
- **Summary:** The evolution of feature importance shows a clear trajectory: early models were influenced by neighborhood and exterior categories, outlier removal and transformation improved robustness, BIC refined the set for interpretability, and Ridge regularization confirmed the central role of living area, while redistributing importance toward generalizable structural features. Together, these steps yield a final model that is both predictive and stable, balancing parsimony, interpretability, and robustness.

### 4.3. Overall Regularisation Summary

- **Predictive Performance:** Ridge, Lasso, and ElasticNet achieve nearly identical performance to the Stepwise BIC model, with Test $R^2 \approx 0.912$ and Test RMSE ≈ 0.210, showing robust generalizability.
- **Diagnostics:** Residuals, QQ plots, Shapiro-Wilk tests (~0.957), and Cook's distance all confirm stability similar to Stepwise BIC.
- **Multicollinearity:** Regularisation reduces VIFs dramatically (top VIF ≈ 9.31 vs 41+ in Stepwise), mitigating multicollinearity while retaining key predictors.
- **Interpretation:** Coefficient shrinkage simplifies interpretation; Lasso and ElasticNet may set minor predictors near zero, emphasizing the most important relationships.
- **Model Selection Implication:** Ridge regression is preferred. While Lasso and ElasticNet could induce additional sparsity, the BIC feature set is already parsimonious, so zeroing coefficients risks losing important predictors. Ridge retains all features, reduces multicollinearity, and maintains predictive accuracy, offering a stable and interpretable model aligned with project objectives of building a strong predictive model and identifying key drivers of Ames property values.

## Final Model Selection and Validation

- **Criteria:** Candidate models were evaluated not only on predictive accuracy ($R^2$, RMSE) and coefficient interpretability but also against the objectives of this project: preserving key features identified in EDA (e.g., **qual_living_area_interaction**), addressing multicollinearity, and handling outliers. Feature importance was considered, acknowledging that rankings vary across iterations due to transformations, outlier adjustments, and penalisation.

- **Selection:** Ridge regression was selected as the final model. Built on the BIC feature set, Yeo-Johnson transformation, and the outlier-adjusted dataset, Ridge delivered predictive performance on par with the best prior models (Test $R^2 \approx 0.912$, RMSE ≈ 0.210), while drastically reducing variance inflation factors and providing superior coefficient stability compared to Lasso and ElasticNet. By shrinking but retaining all predictors, Ridge preserves explanatory breadth and avoids losing relevant features, even though feature importance varies across iterations. Compared to stepwise BIC, Yeo-Johnson OLS, and outlier-removed models, Ridge provides the most robust combination of predictive accuracy, interpretability, and feature retention.
- **Validation:** For an unbiased performance estimate, the full pipeline should be rerun with the train/test split applied prior to feature engineering (polynomial expansions, scaling, transformations). This ensures that model refinement choices, including Ridge shrinkage, are not influenced by the full dataset.
- **Outcome:** The Ridge model meets the project objectives by delivering strong predictive performance while retaining all important features identified in EDA. The Ridge model offers a stable and interpretable framework for predicting Ames property values while retaining all plausible drivers. Feature importance should be interpreted as relative and context-dependent, given variation across iterations, but Ridge provides the most reliable assessment of key predictors. Although not yet deployable, this model establishes a clear refinement plan for improving validation and reassessing generalisation.

## Final Model Interpretation

The Ridge model demonstrates strong predictive performance, achieving a Test $R^2 \approx 0.912$ and RMSE ≈ 0.210, on par with the stepwise BIC model. In addition to maintaining accuracy, Ridge drastically reduces multicollinearity, with VIFs brought down to acceptable levels (under 10), ensuring more reliable coefficient estimates. Although not yet deployable, this specification provides valuable insights into drivers of Ames property values. Coefficient signs and magnitudes indicate the direction and relative influence of predictors, while Ridge shrinkage stabilises estimates and mitigates collinearity, preserving all features in the model.

**Feature Importance Insights:**

- **Dominant Predictor: qual_living_area_interaction** remains the top driver across all models, reflecting its critical role identified during EDA.
- **Structural Features:** Ridge highlights generalizable structural predictors such as BsmtFin SF 1, Lot Area_capped, total_bathrooms, and remodel_age, consistent with their practical relevance to house value.
- **Neighborhood Effects**: Previously dominant neighborhood dummies are downweighted, showing that Ridge penalization shifts importance toward robust structural and transactional predictors.
- These findings confirm that the model satisfies the objectives: it preserves EDA-identified key features, mitigates multicollinearity, and emphasizes interpretable, generalizable drivers of Ames house prices.

---

# Conclusion and Next Steps

**Conclusion:**
This project integrated EDA and iterative regression modelling to identify drivers of Ames housing prices and evaluate predictive models against the stated objectives. EDA highlighted **overall home size, quality, and qual_living_area_interaction** as primary predictors, with neighborhood and garage characteristics showing additional influence. Iterative modelling confirmed:

1. **Full OLS model on all data** (Original Model 2)— established baseline performance but suffered from multicollinearity and outlier influence.
2. **Outlier-removed model** — improved coefficient stability and reduced extreme residuals, but removed a few influential points.
3. **Yeo-Johnson transformation** — further stabilised variance and improved model robustness while retaining all data.
4. **Stepwise BIC selection** — provided a parsimonious feature set with good predictive accuracy and interpretability.
5. **Ridge regularization** — emerged as the best-performing specification, matching the predictive strength of the BIC model (Test $R^2 \approx 0.912$, RMSE ≈ 0.210) while drastically reducing multicollinearity (VIFs < 10). Ridge retained all features, downweighted over-represented neighbourhood effects, and emphasized generalizable structural and transactional features such as **BsmtFin SF 1**, **Lot Area_capped**, **total_bathrooms**, and **remodel_age**.

Diagnostics confirmed that Ridge produced more stable coefficients, mitigated fold-to-fold variability, and preserved explanatory breadth, aligning with the project objectives of predictive accuracy, interpretability, and key feature retention.

**Next Steps:**
Building on the iterative results and Ridge insights, the following actions are recommended:

- **Improved Data Partitioning:** Apply the train/test split immediately after EDA and prior to polynomial expansion or scaling to ensure unbiased performance estimates.
- **Advanced Feature Engineering:** Explore additional interactions and nonlinear transformations, especially involving temporal factors (sale year, remodel age) and key categorical variables (neighborhood, garage type).
- **Nonlinear and Machine Learning Models:** Consider regression-based and tree-based approaches (Random Forest, Gradient Boosting, XGBoost) to capture complex patterns beyond linear effects.
- **Robustness Checks:** Investigate residual outliers and assess alternative robust regression techniques if high-value observations disproportionately influence the model.

- **Post-hoc Statistical Analysis:** Include effect size calculations, confidence intervals, and detailed comparison of key predictors to strengthen interpretability.
- **Validation and Generalisation**: Conduct repeated cross-validation and out-of-sample testing to ensure stability and prevent overfitting.
- **Domain-Informed Refinement**: Use insights from Ridge and previous iterations to guide future feature selection, hypothesis testing, and business-relevant interpretations.

This iterative approach establishes a strong foundation for predictive modelling, highlighting key drivers while providing a roadmap to more robust, generalizable models for Ames property valuation.

---

## References

To enhance reproducibility, the complete Jupyter notebook, data files, and environment specifications are publicly available at [https://github.com/rebeccastalleymoores/ames-housing-eda/blob/main/Unlocking%20Housing%20Insights.ipynb]. This allows replication of the analysis, visualizations, and hypothesis testing presented here.