Unlocking Housing Insights: An Exploratory Analysis of the Ames Housing Dataset.

Prepared by

Rebecca Stalley-Moores

Unlocking Housing Insights: An Exploratory Analysis of the Ames Housing Dataset.

Prepared by: **Rebecca Stalley-Moores**

Date: **17/06/2025**

Executive Summary

This analysis explores the factors influencing house sale prices using a comprehensive dataset of sold residential properties. Initial univariate analysis reveals that most homes tend to be modest in size with smaller garages, porches, and bathrooms being common, while room count features approximate a normal distribution. Key categorical variables such as `Neighborhood`, `House Style`, and `MS Subclass` display distinct groupings that may impact pricing, while ordinal quality indicators like `Overall Cond` and `Kitchen Qual` reflect expected ordered patterns with some inconsistencies.

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

Multivariate exploration using correlation matrices and grouped boxplots confirms that while many numerical features have overlapping distributions across categorical groups such as `Neighborhood` and `Garage Type`, `qual_living_area_interaction` consistently demonstrates clear median shifts and lower variability overlap. This underscores its value as a robust predictor of sale price.

Hypothesis testing focused on `qual_living_area_interaction` revealed significant differences in sale prices across low, medium, and high value groups, supported by both ANOVA and the non-parametric Kruskal-Wallis test. These results strongly validate the importance of this combined quality and area metric in explaining sale price variation.

Next steps include expanding hypothesis testing to other features such as `Neighborhood` and `Garage Type`, applying corrections for multiple comparisons to control Type I error, and employing advanced machine learning models like Random Forests and Gradient Boosting to better capture complex relationships.

Overall, this project provides a solid foundation for understanding key drivers of housing prices, with particular emphasis on engineered features that integrate quality and size, informing both future statistical analysis and predictive modelling efforts.

Data Summary

- The goal of this analysis 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.

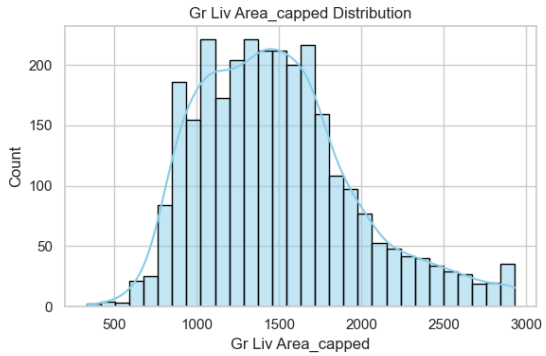
By retaining these features, the analysis remains sensitive to **non-normal, real-world patterns** that could otherwise be smoothed over if only symmetric or linear distributions were considered.

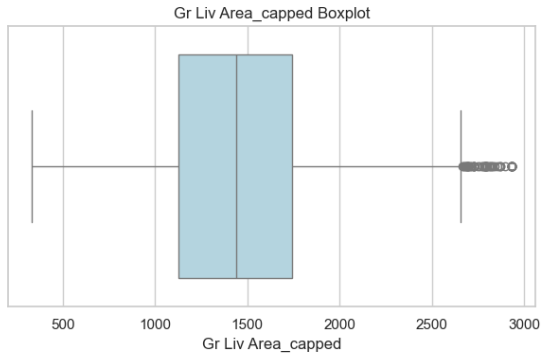
The final features selected for univariate exploration were:

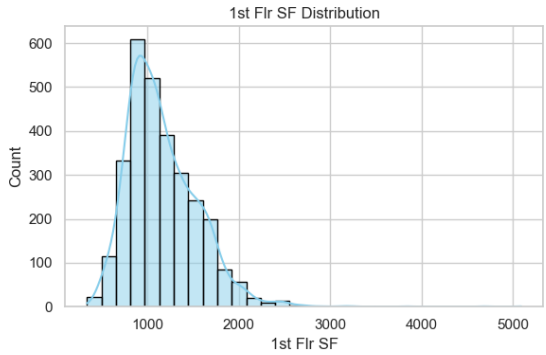
- Gr Liv Area_capped
- 1st Flr SF
- Total Bsmt SF
- Mas Vnr Area_capped
- Lot Area_capped
- Wood Deck SF_capped
- total_bsmt_finished
- qual_living_area_interaction
- Full Bath
- total_bathrooms
- TotRms AbvGrd
- Bedroom AbvGr
- Kitchen AbvGr
- avg_quality
- house_age
- remodel_age
- Yr Sold
- Garage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

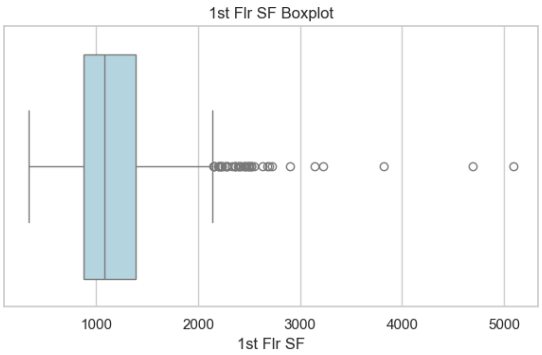
Each feature is grouped logically below, visualised using histograms and KDE and box plots to show the distribution, skewness and outliers

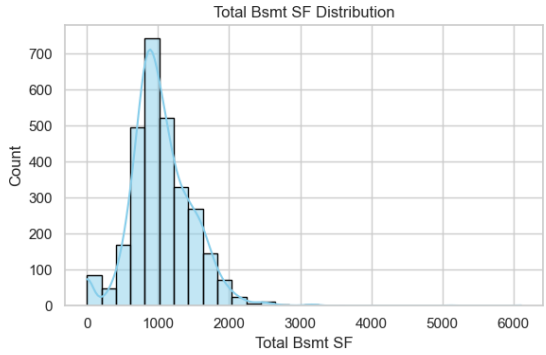
- **Size and Area Features**

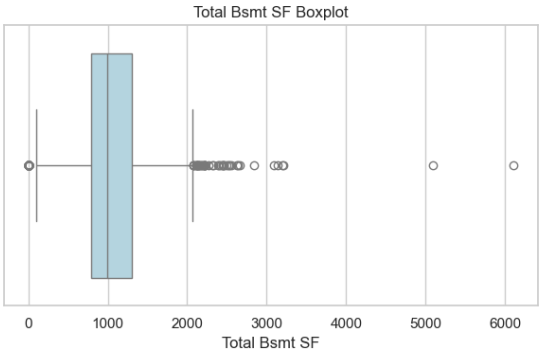


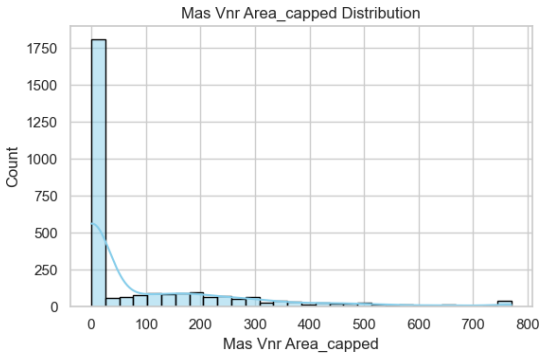


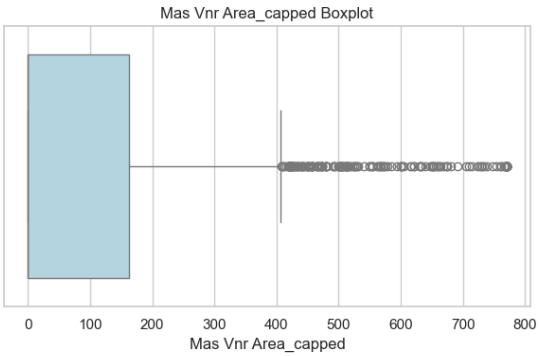


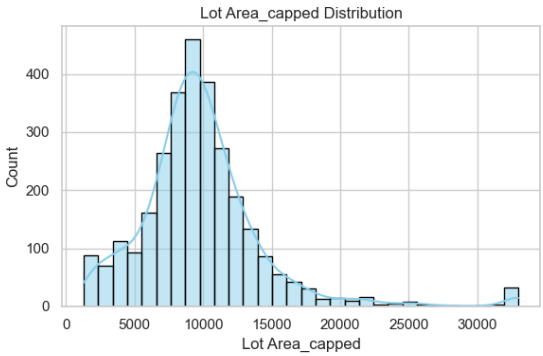


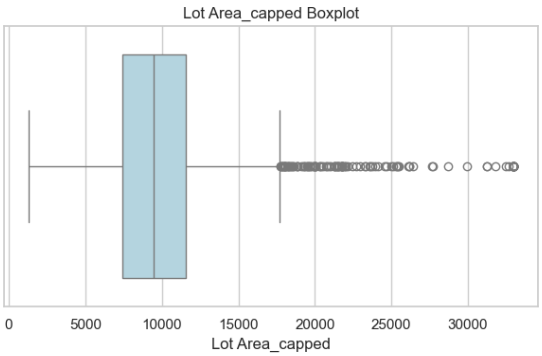


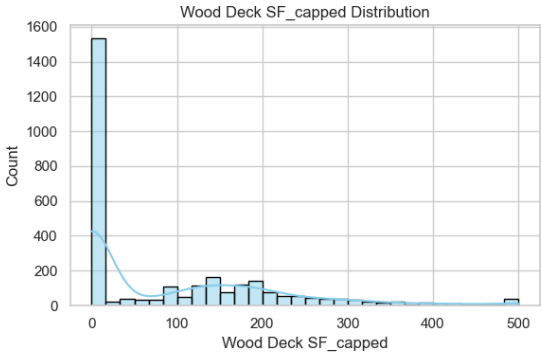


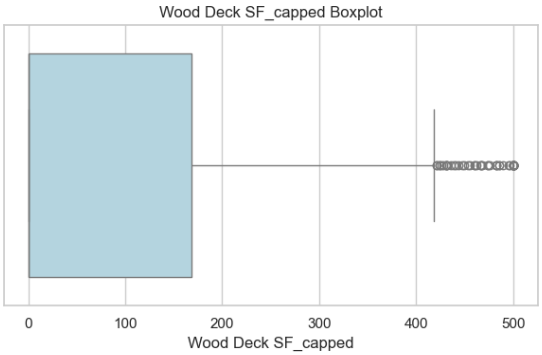

Mas Vnr Area_capped Distribution
Count
Mas Vnr Area_capped

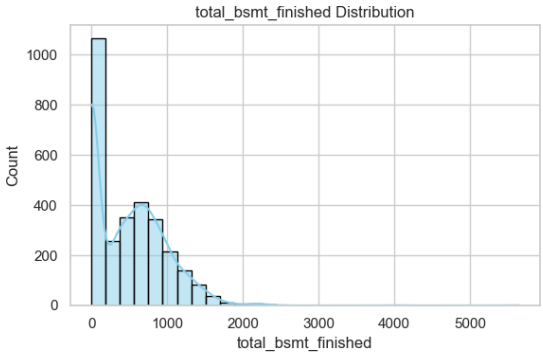

Mas Vnr Area_capped Boxplot
Mas Vnr Area_capped

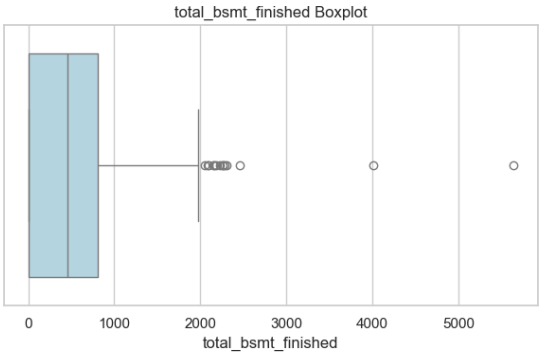

Lot Area_capped Distribution
Count
Lot Area_capped

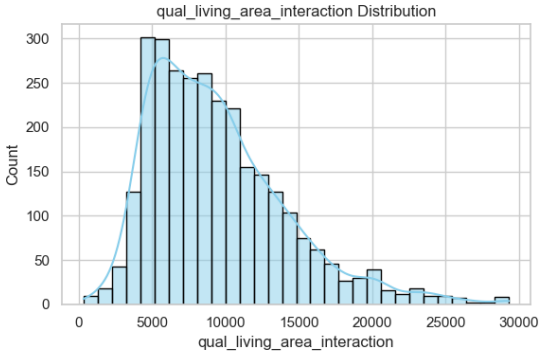

Lot Area_capped Boxplot
Lot Area_capped

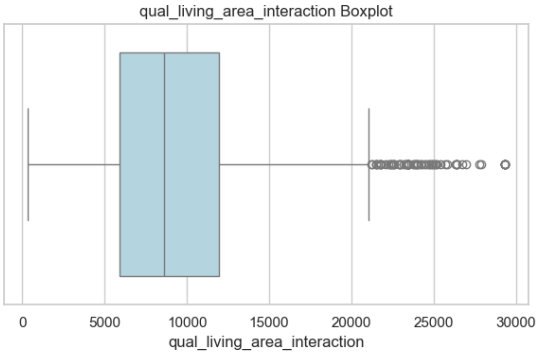

Wood Deck SF_capped Distribution
Count
Wood Deck SF_capped


Wood Deck SF_capped Boxplot
Wood Deck SF_capped

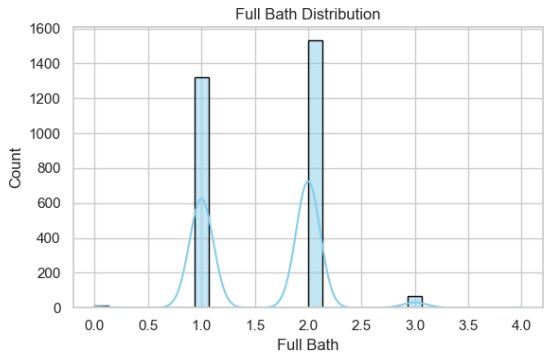

total_bsmt_finished Distribution
Count
total_bsmt_finished

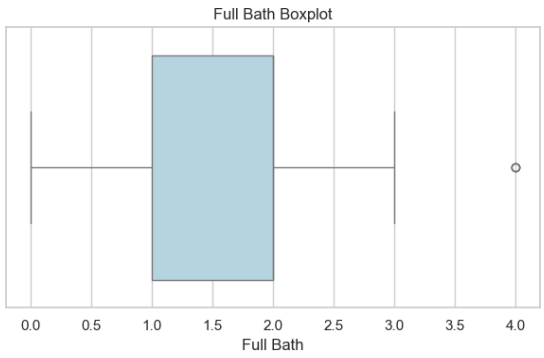

total_bsmt_finished Boxplot
total_bsmt_finished

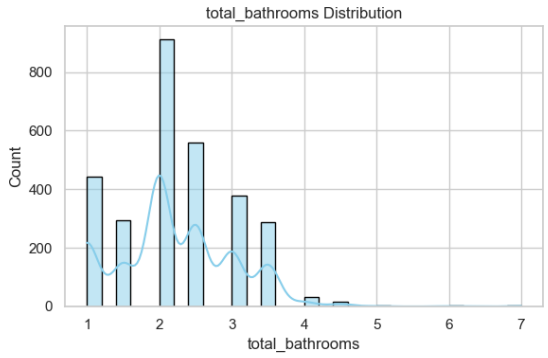

qual_living_area_interaction Distribution
Count
qual_living_area_interaction

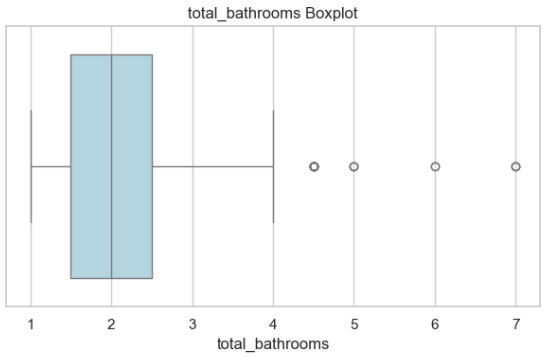

qual_living_area_interaction Boxplot
qual_living_area_interaction

- **Bathrooms and Plumping**

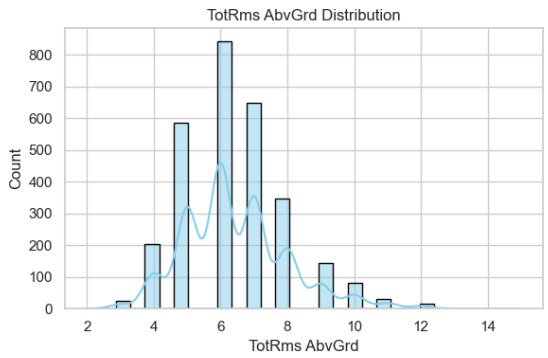


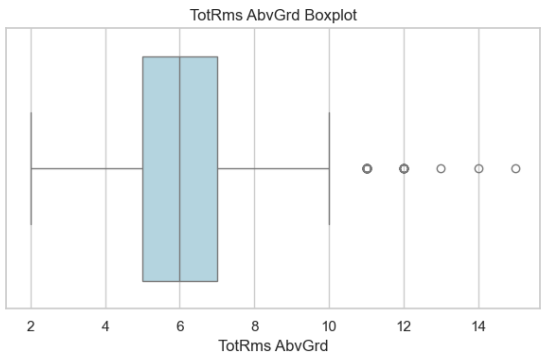


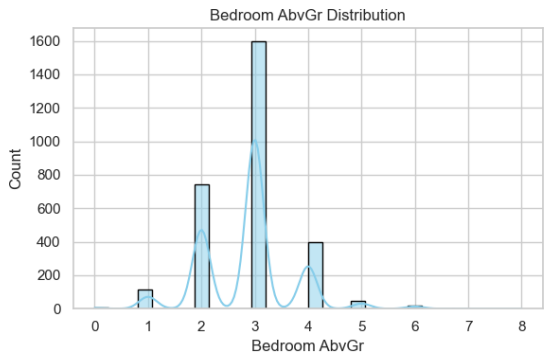


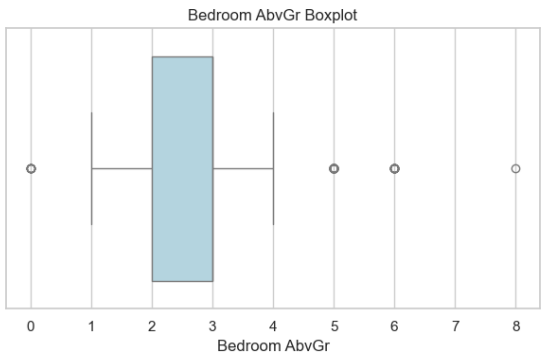


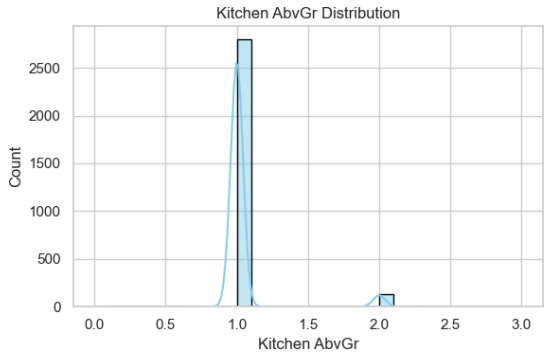
- **Rooms and Interior Count**

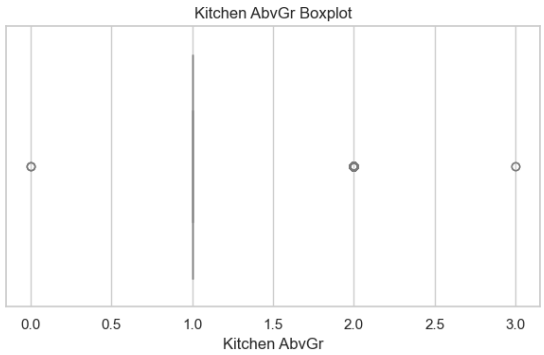




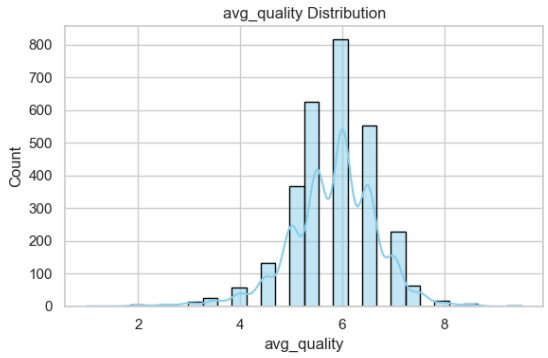


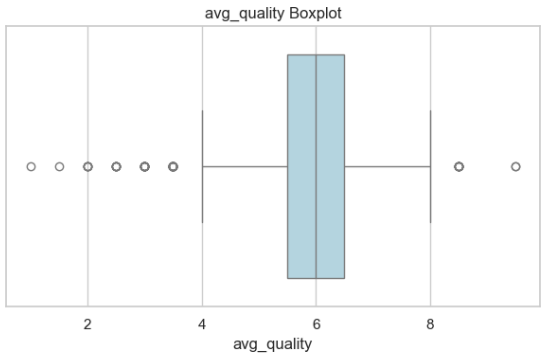


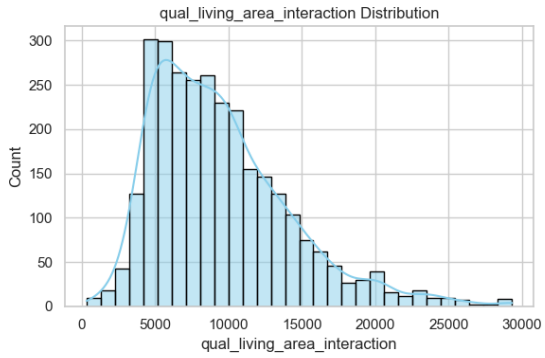


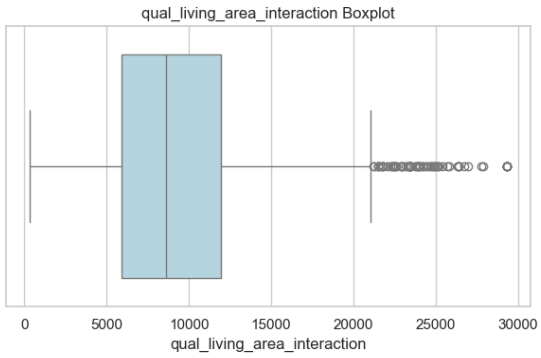


- **Quality / Composite Scores**

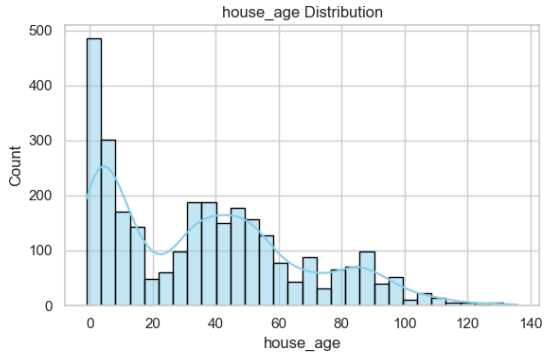


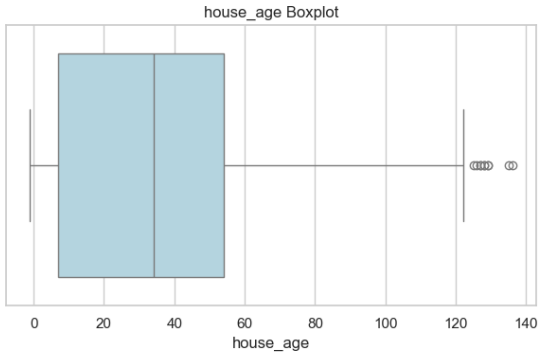


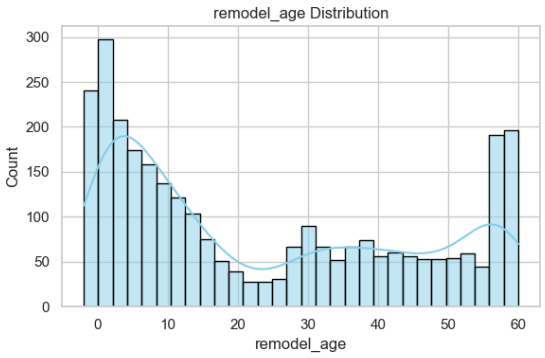


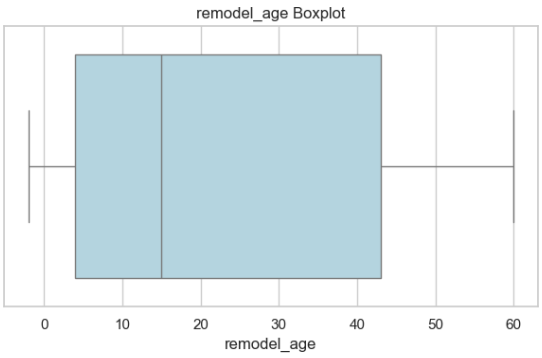


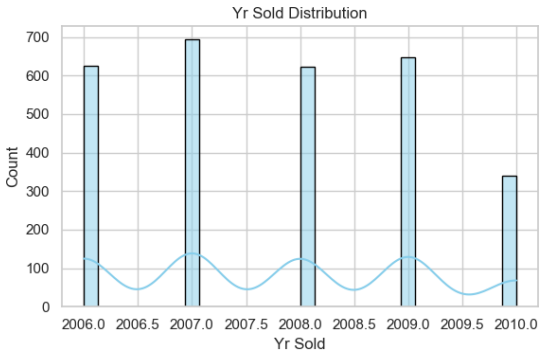
- **Age / Time-Based Features**

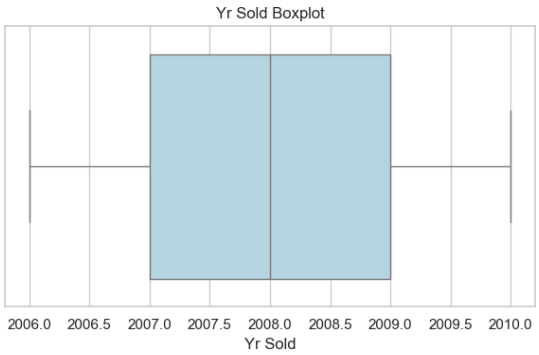




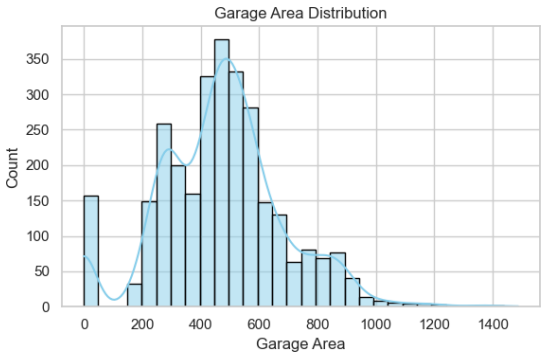


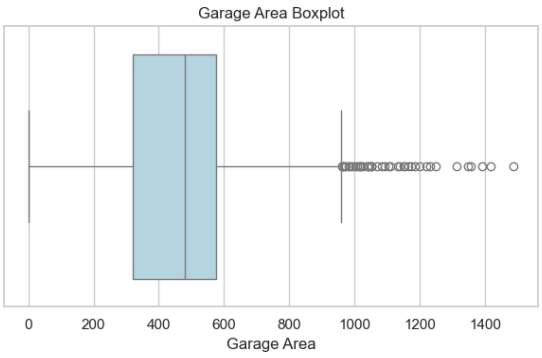


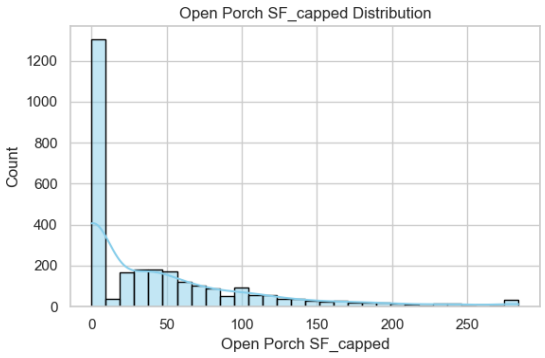


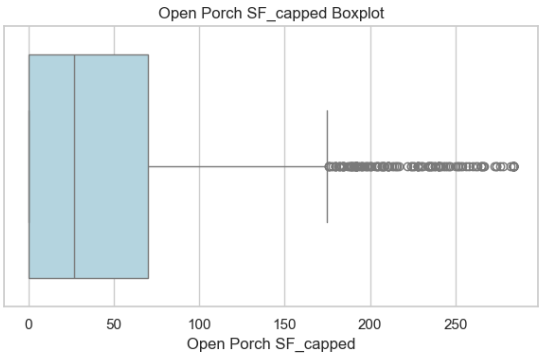


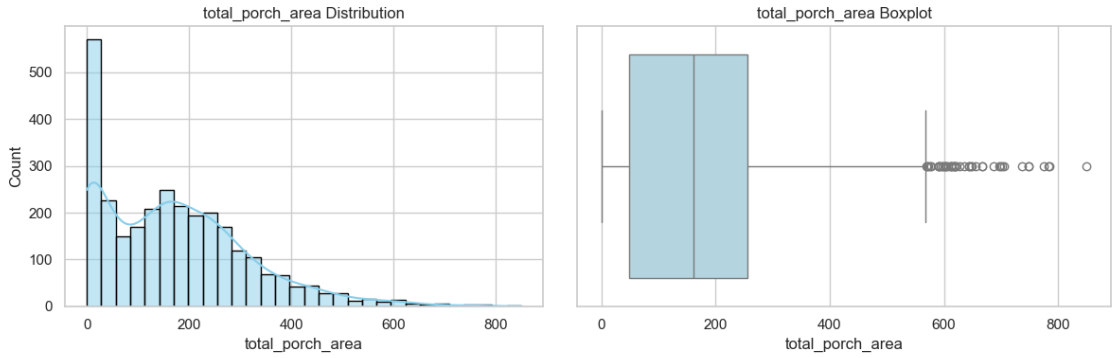
- **Garage, Porch and External Features**



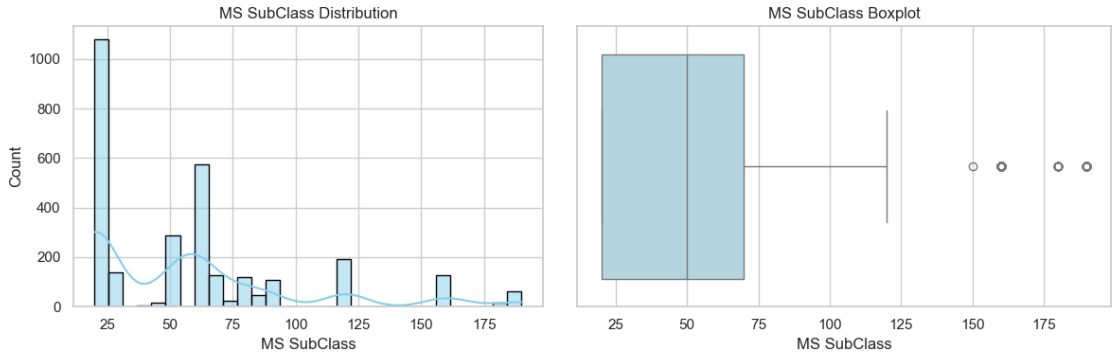


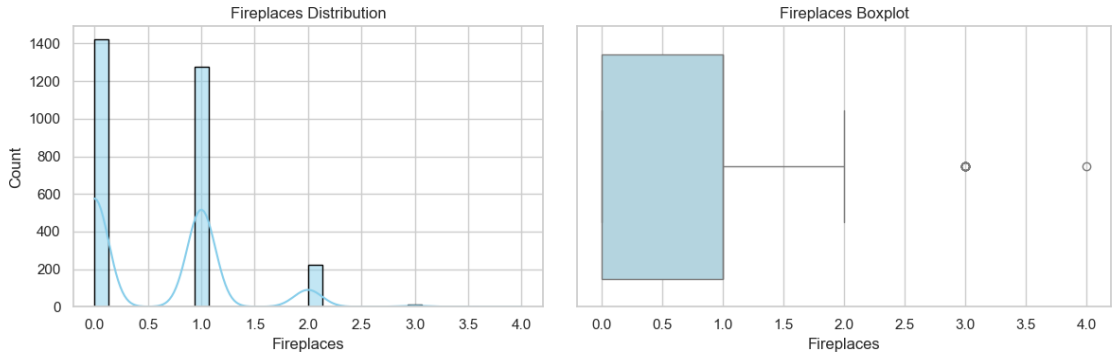






- **Structural / Other**





- **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:

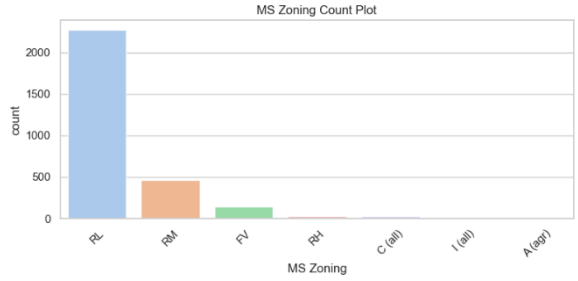
- Features with little to no variation across categories were dropped.
- The remaining features were ranked by significance using ANOVA p-values.
- The top 15 features were selected.
- An additional 5 domain-relevant features were added back for contextual importance.

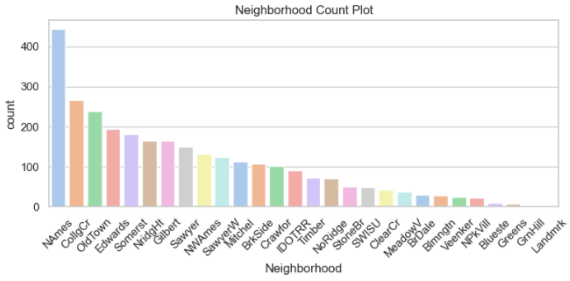
The final set of categorical features analysed is:

- `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`

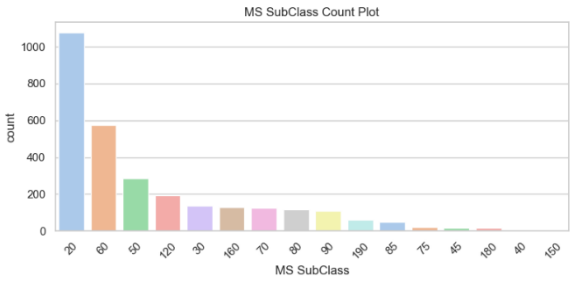
Each feature is grouped logically below, visualised using count plots to assess the distribution of categories within each feature.

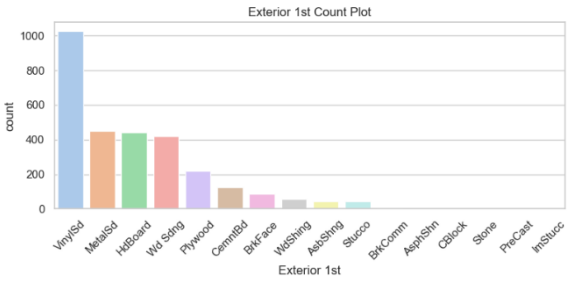
- **Location and Zoning**

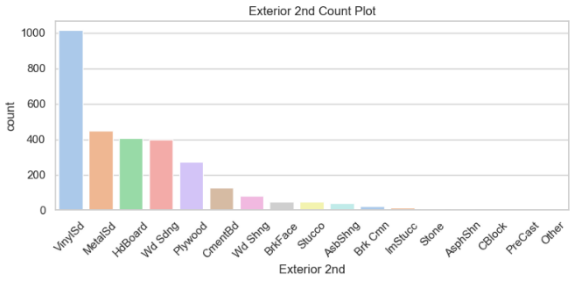


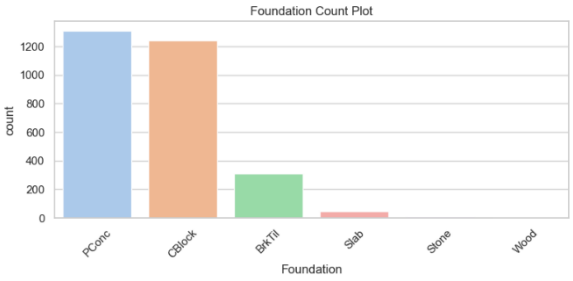


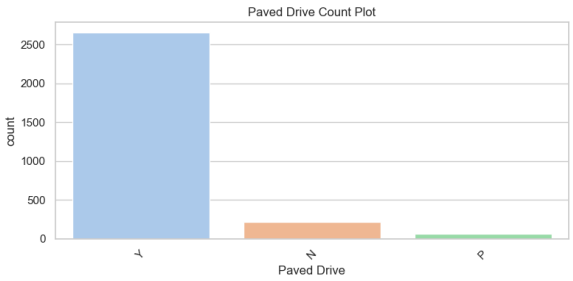
- **Exterior and Structural**

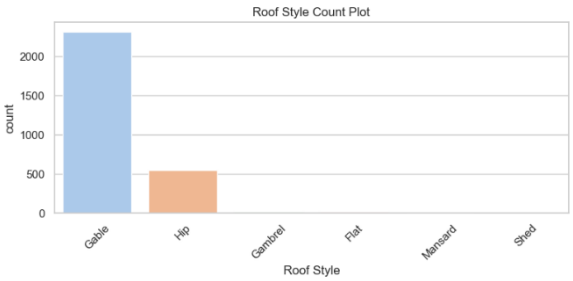




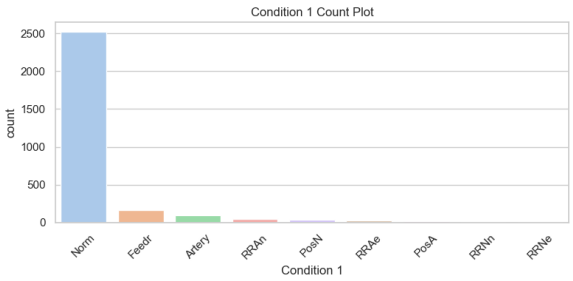


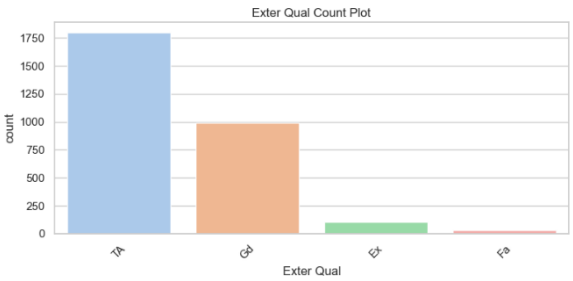


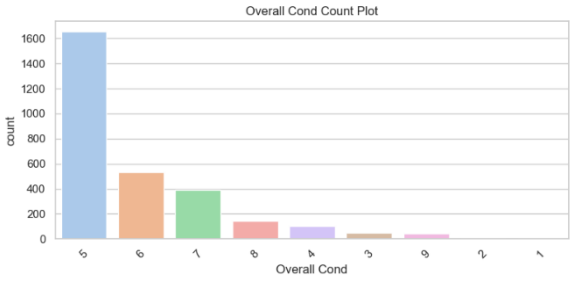




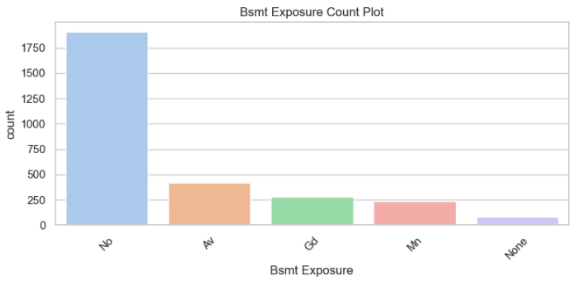
- **Quality and Condition Ratings**

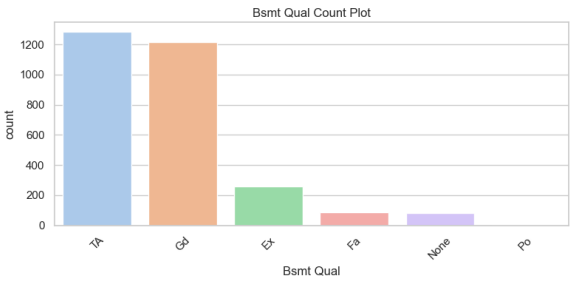


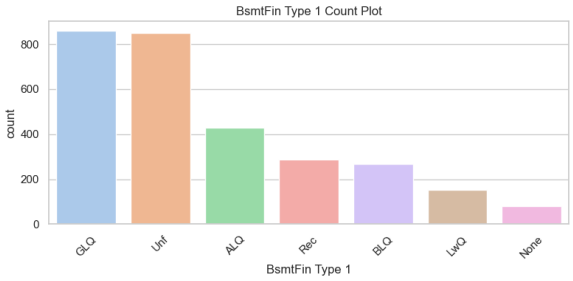




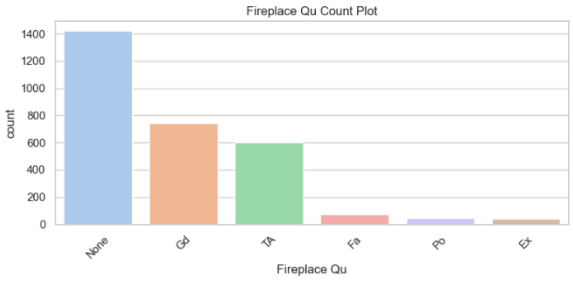
- **Basement Features**

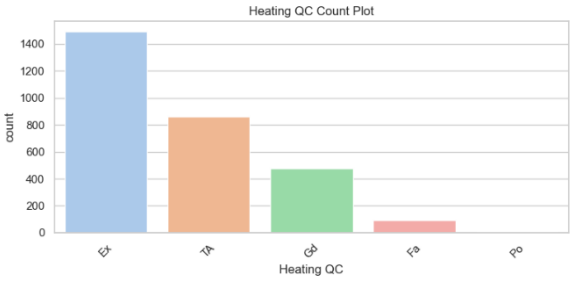




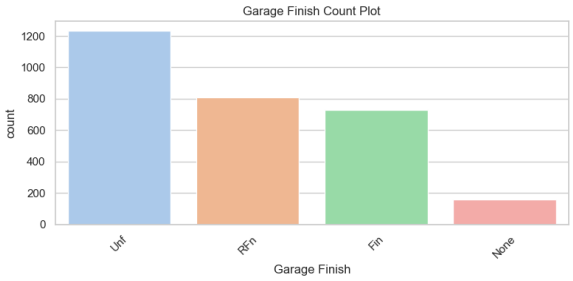


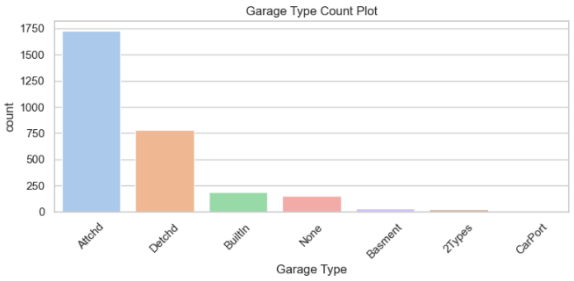
- **Heating and Fireplace**



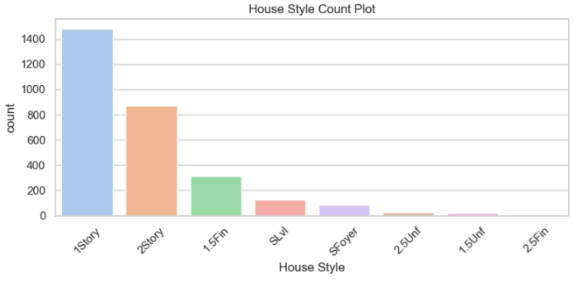


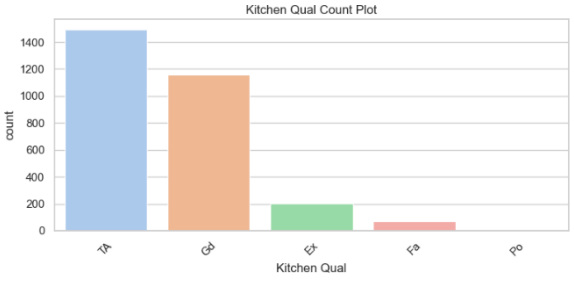
- **Garage Features**





- **Interior Quality and Style**





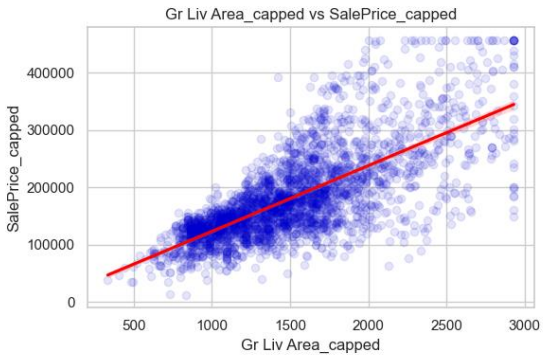
- Nominal categories like Neighborhood, Exterior 1st and Exterior 2nd, MS Subclass, and House Style exhibit clear, stepped differences. These well-separated differences between categories may be important when later considering their impact on the target variable.
- Ordinal categories such as Exter Qual, 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.

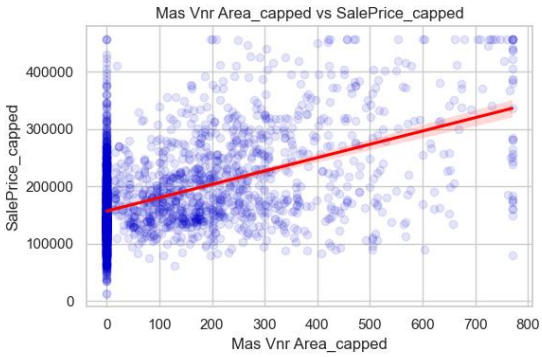
Bivariate Analysis

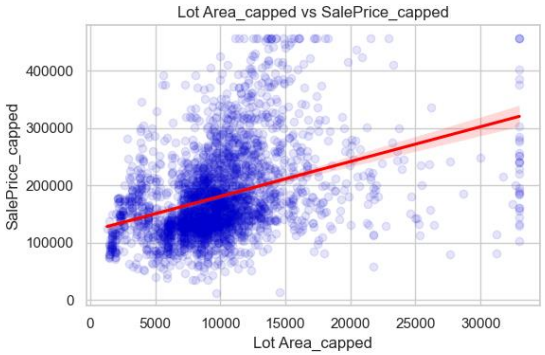
Numerical Versus Target

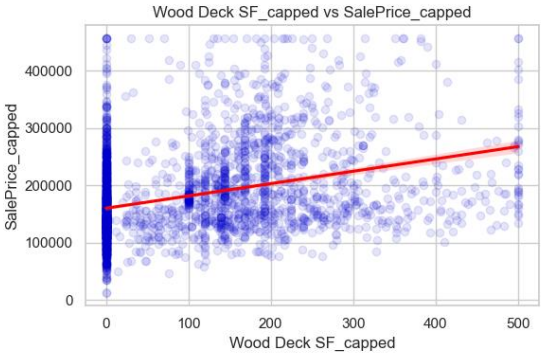
The following regplots show the scatter points with a best fit line for all selected key features against SalePrice_capped. As some features are ordinal in nature these are displayed as boxplots:

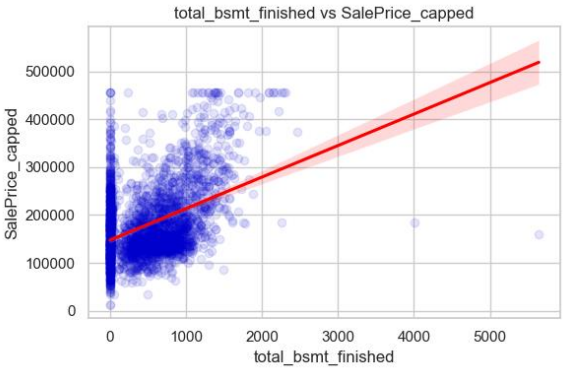
- **Size and Area Features**



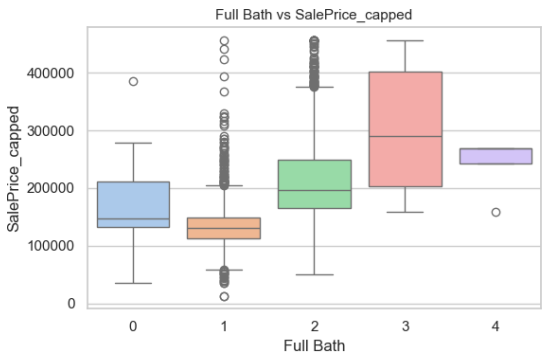


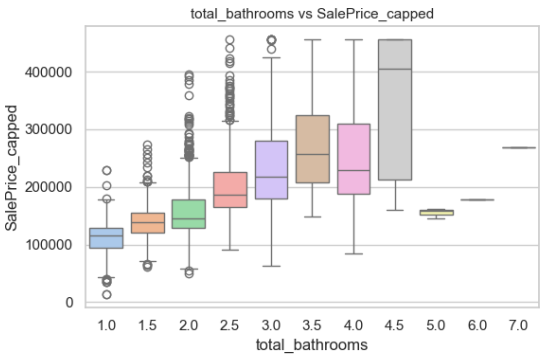




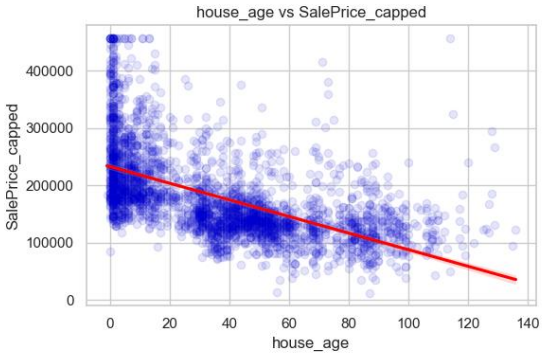


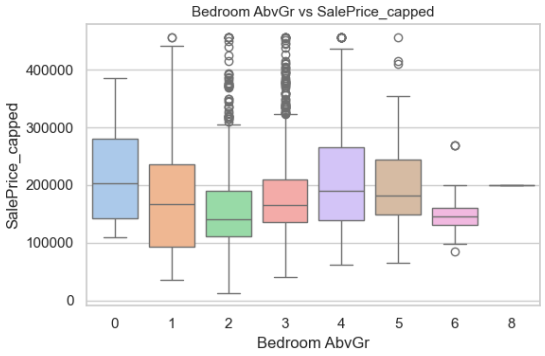
- **Bathrooms and Plumping**

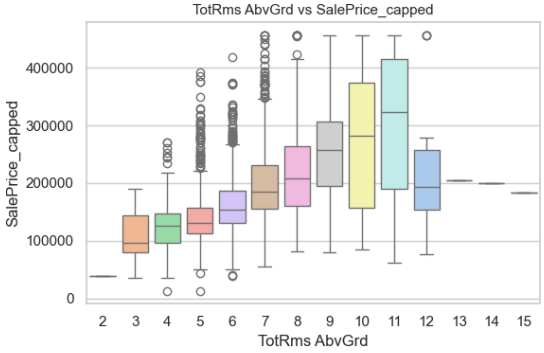


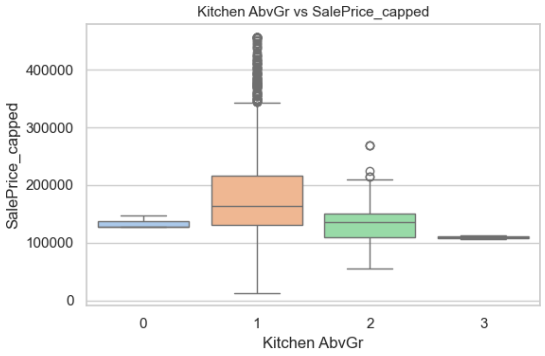


- **Rooms and Interior Count**

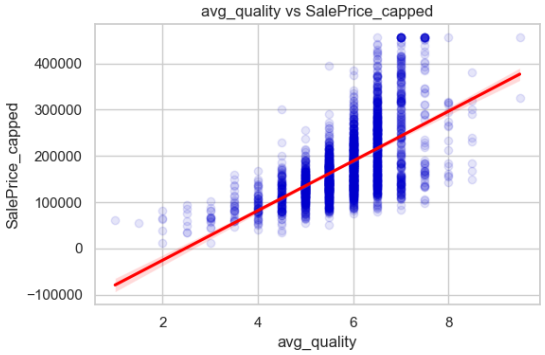




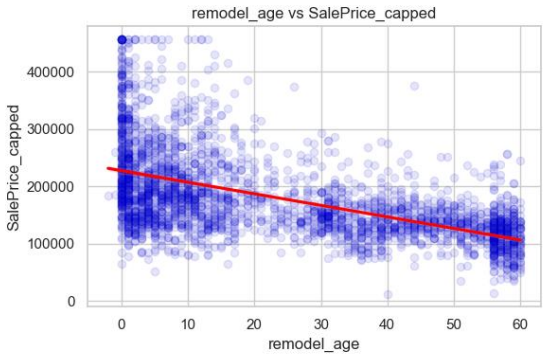




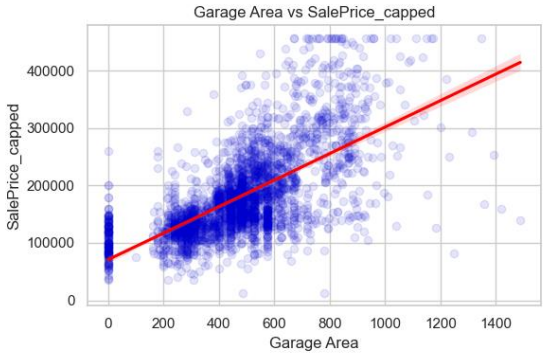
- **Quality / Composite Scores**

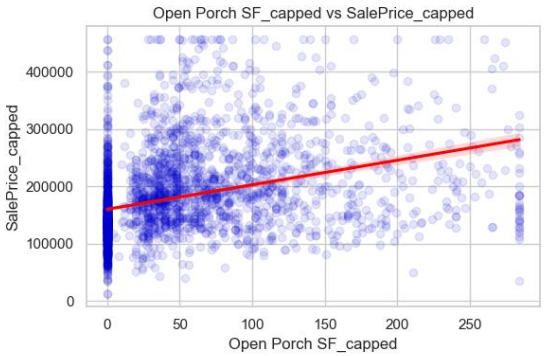


- **Age / Time-Based Features**

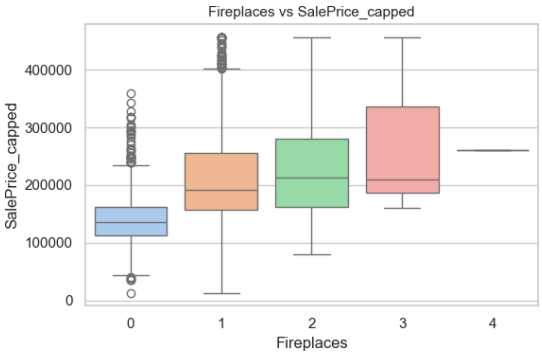


- **Garage, Porch and External Features**

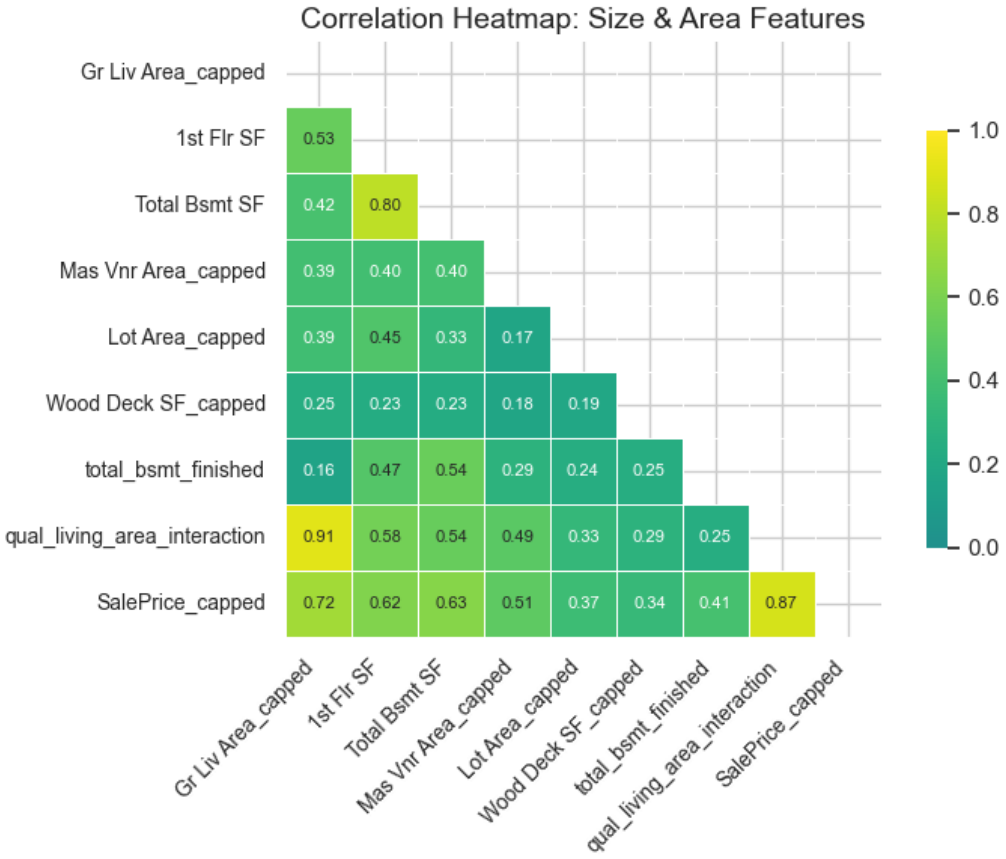


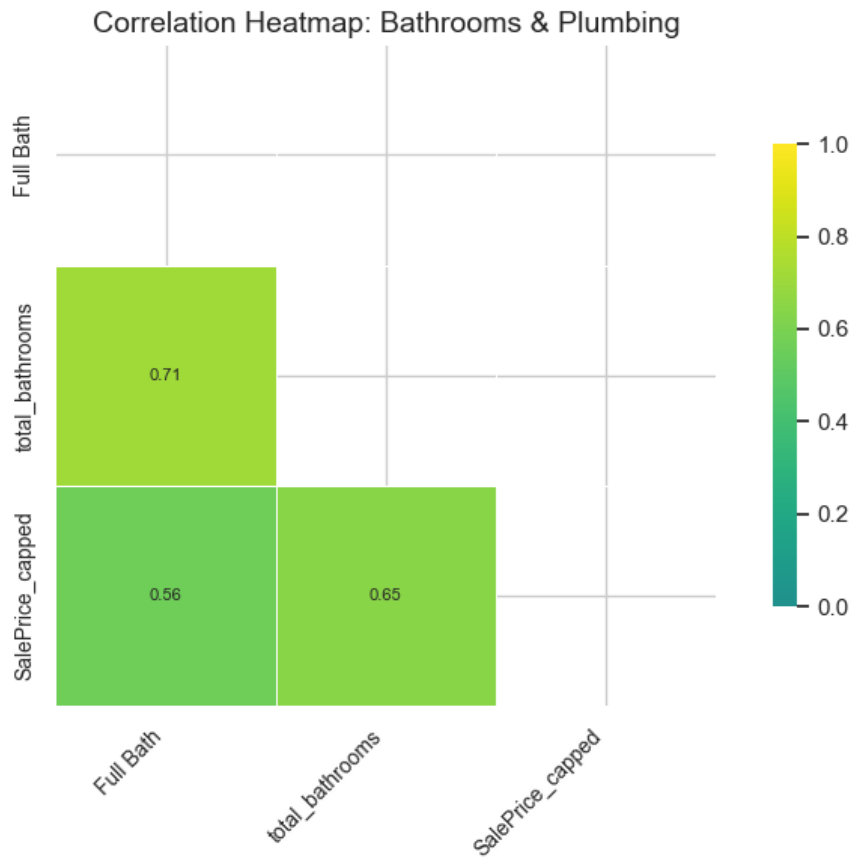


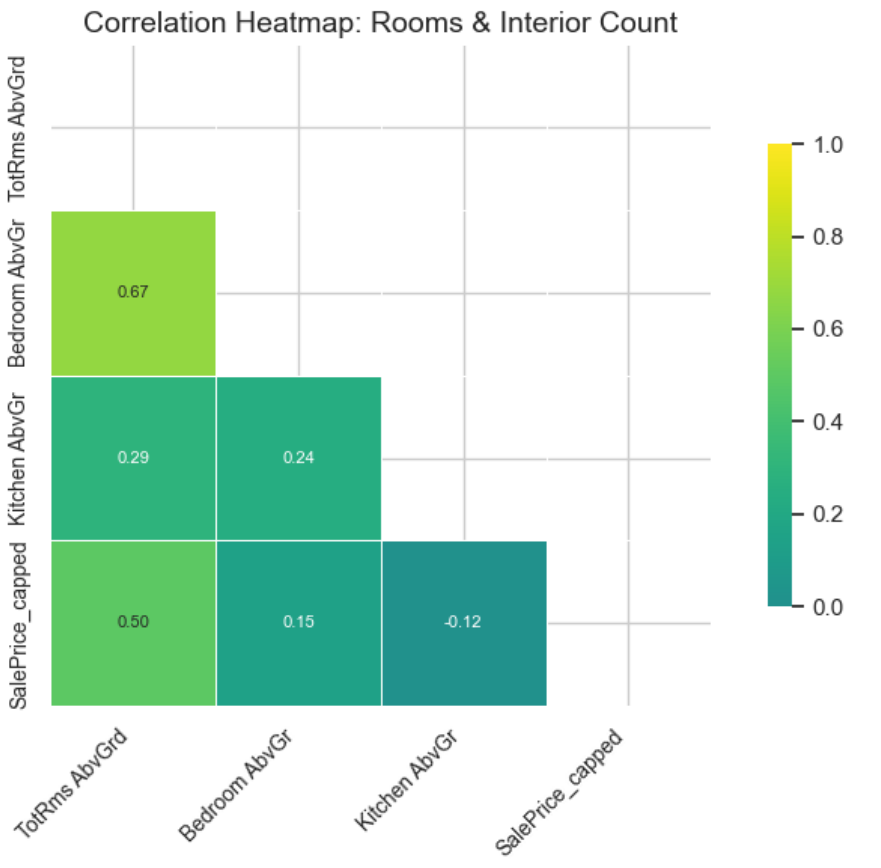
- **Structural / Other**



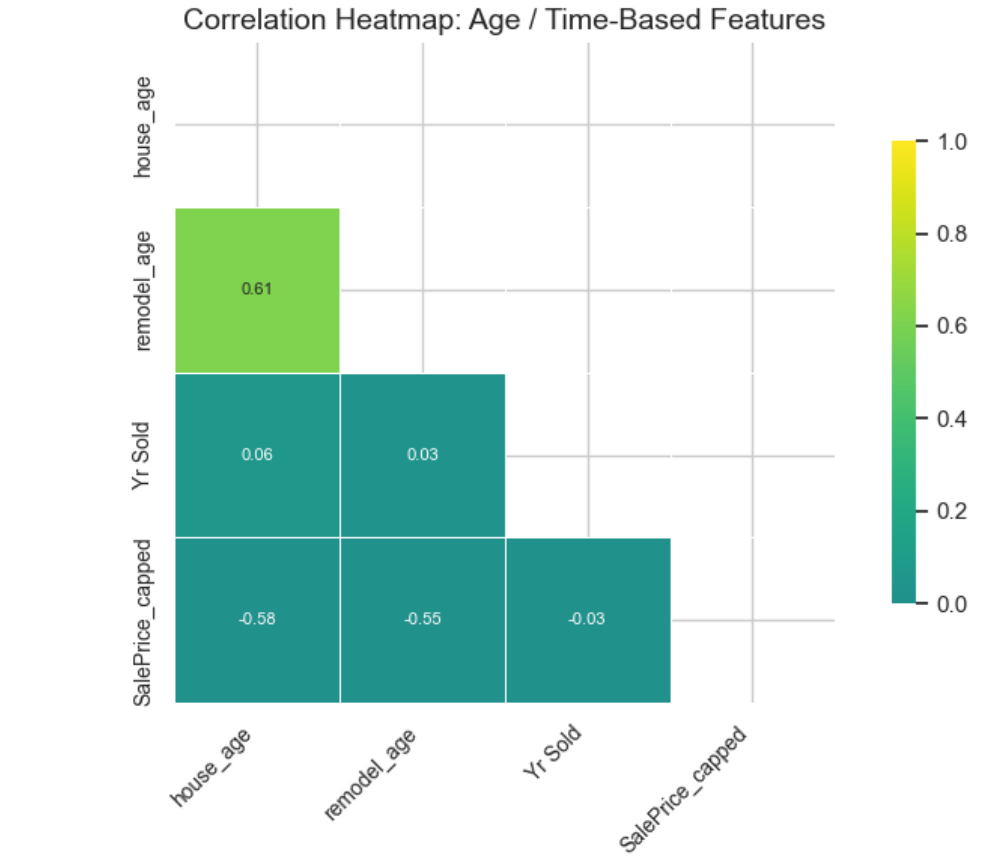
The following heatmaps show the strength of correlation between all key selected features as well as `SalePrice_capped`, grouped thematically.

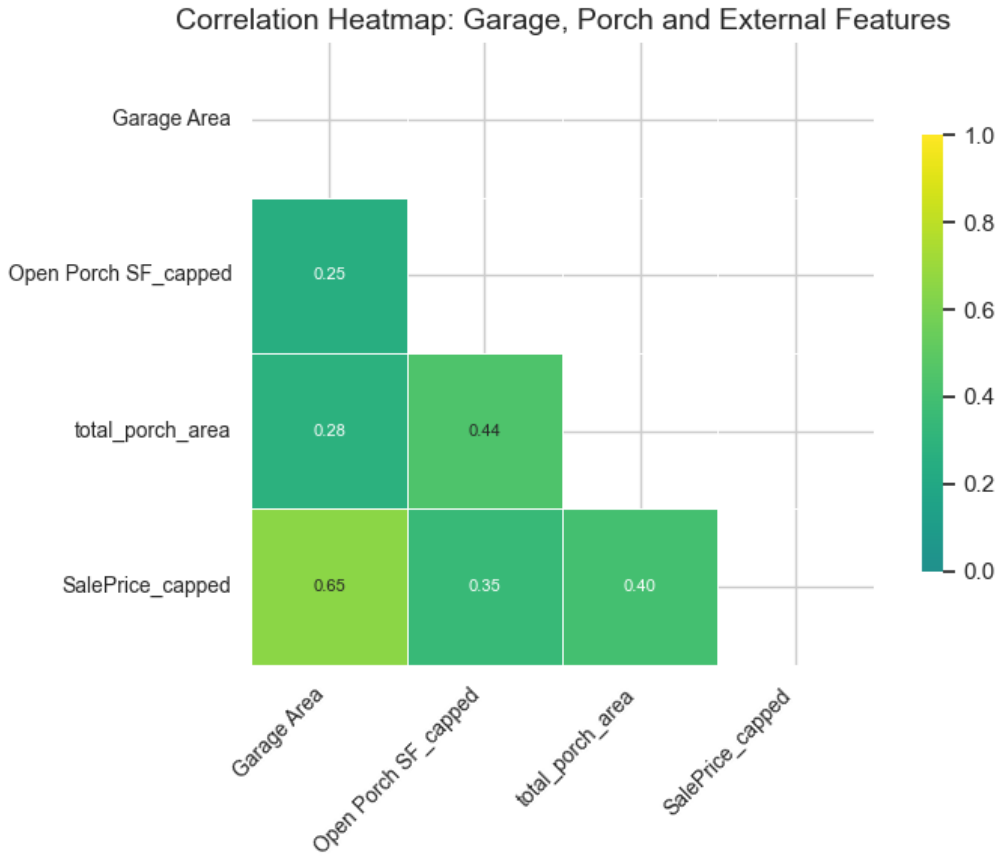


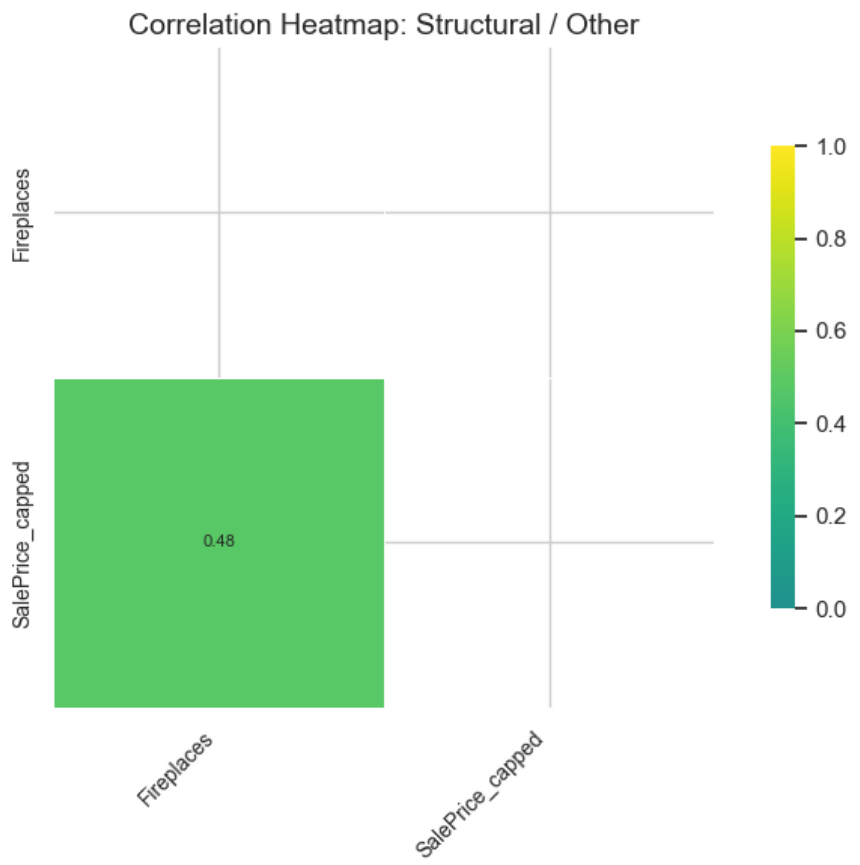



Correlation Heatmap: Rooms & Interior Count
TotRms AbvGrd
Bedroom AbvGr
Kitchen AbvGr
SalePrice_capped
0.67
0.29
0.24
0.50
0.15
-0.12
1.0
0.8
0.6
0.4
0.2
0.0


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


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`

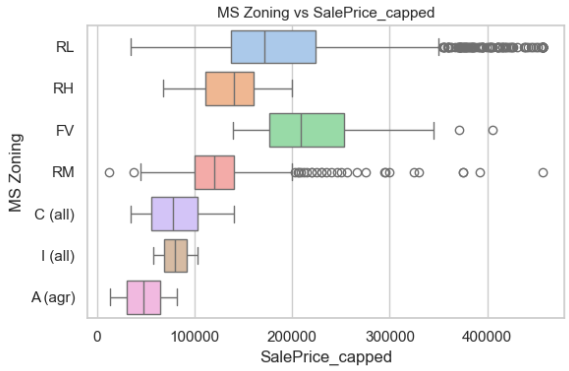
These correlations suggest some redundancy in the data, and a subset of these features could be removed during future model development.

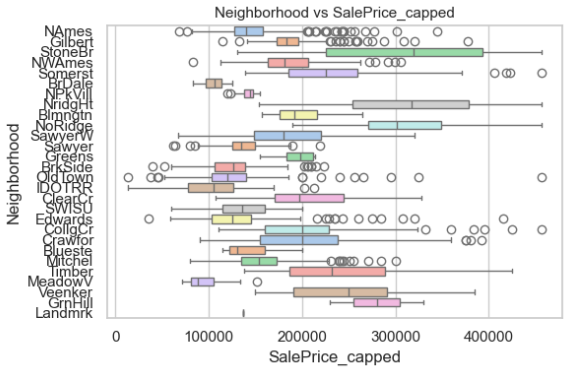
From the regplots, it's clear that `qual_living_area_interaction` has data points that closely follow the best-fit line. `Gr Liv Area_capped` follows the line well at lower square footage levels, but the relationship becomes weaker as square footage increases.

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.

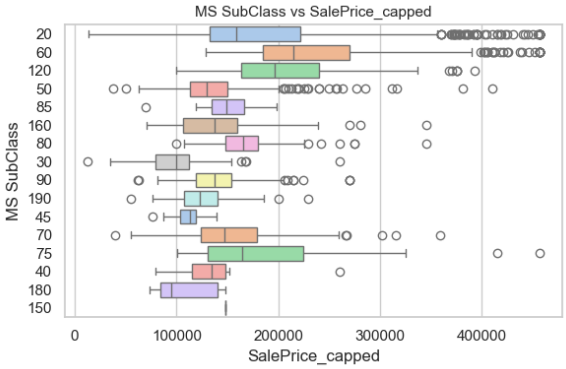
Categorical versus Target

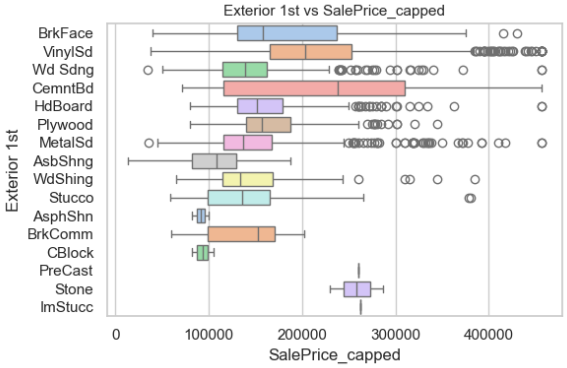
- **Location and Zoning**

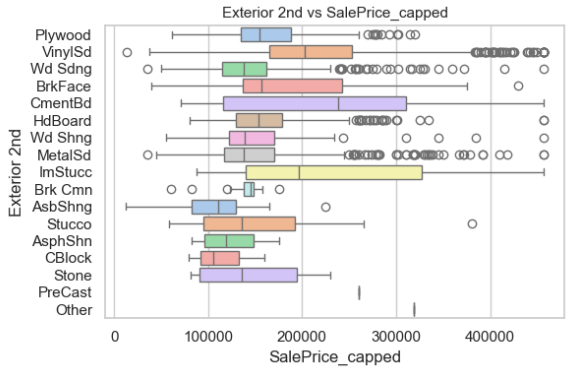


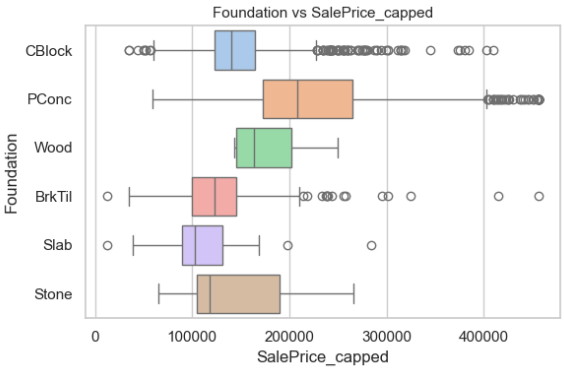


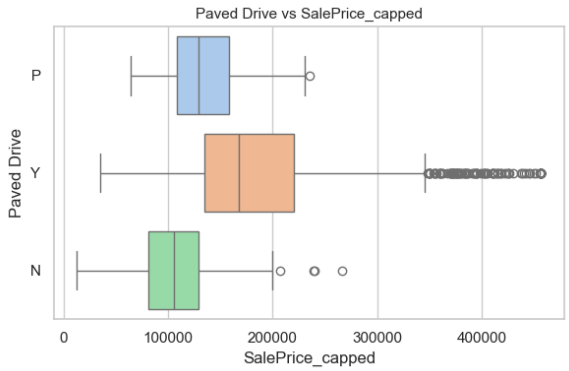
- **Exterior and Structural**

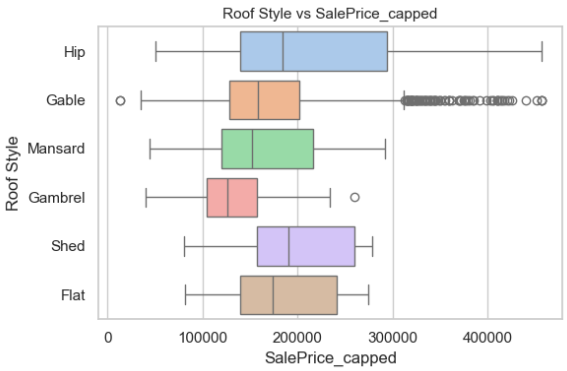




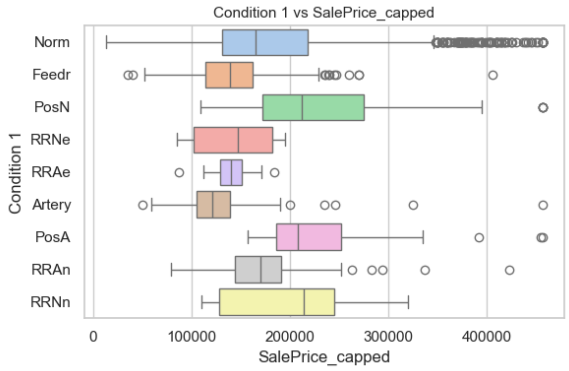


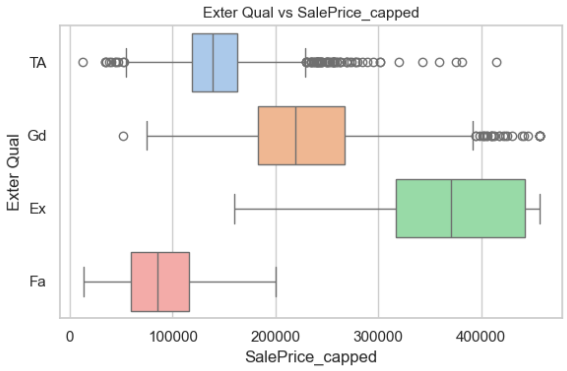


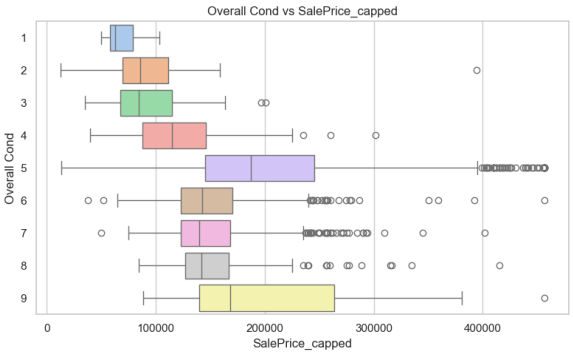




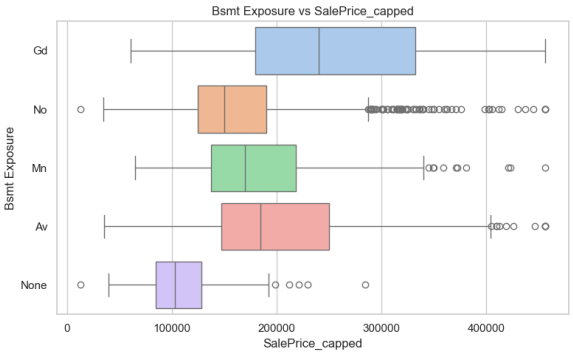
- **Quality and Condition Ratings**

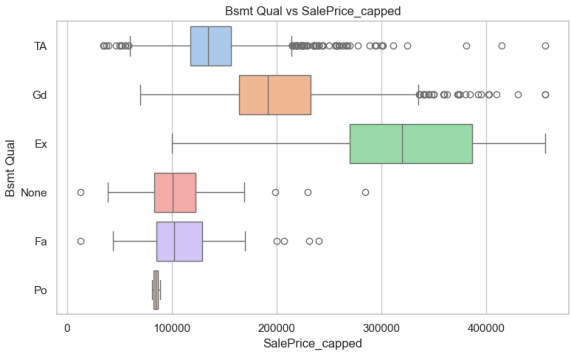


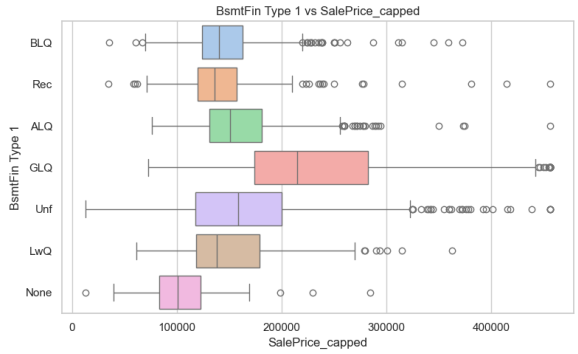




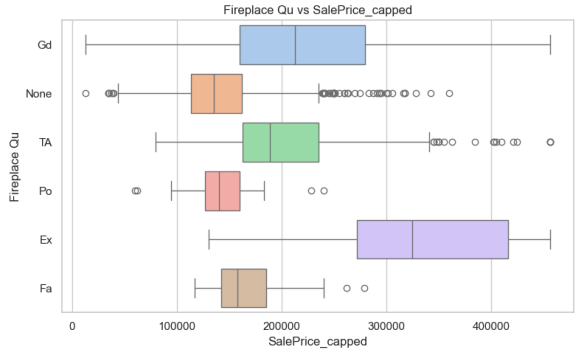
- **Basement Features**

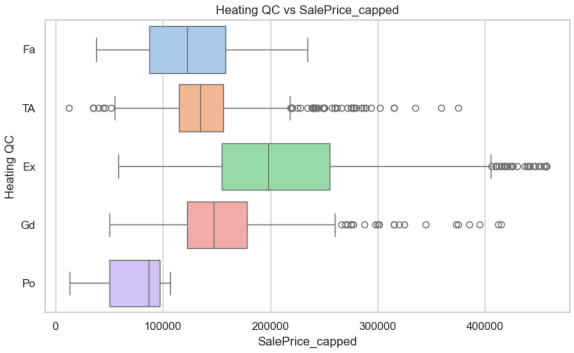




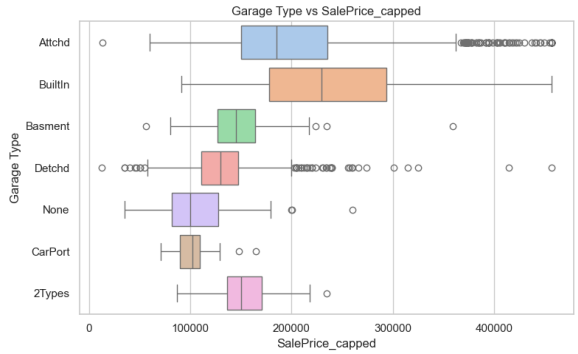


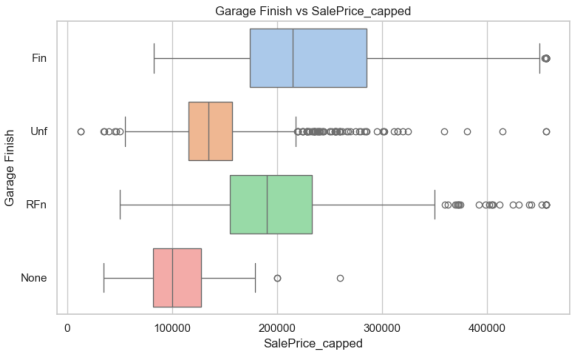
- **Heating and Fireplace**



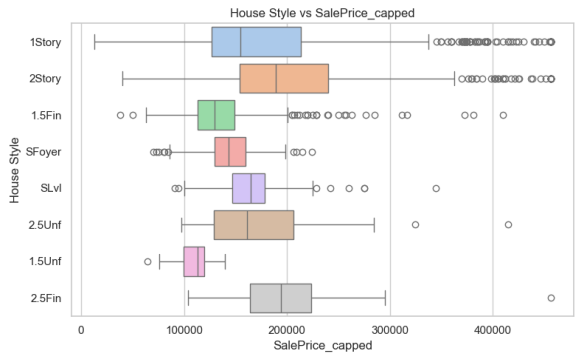


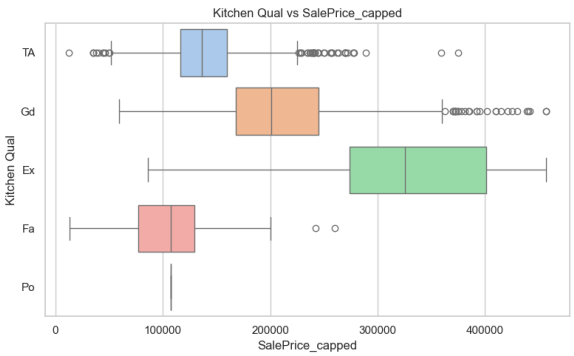
- **Garage Features**





- **Interior Quality and Style**

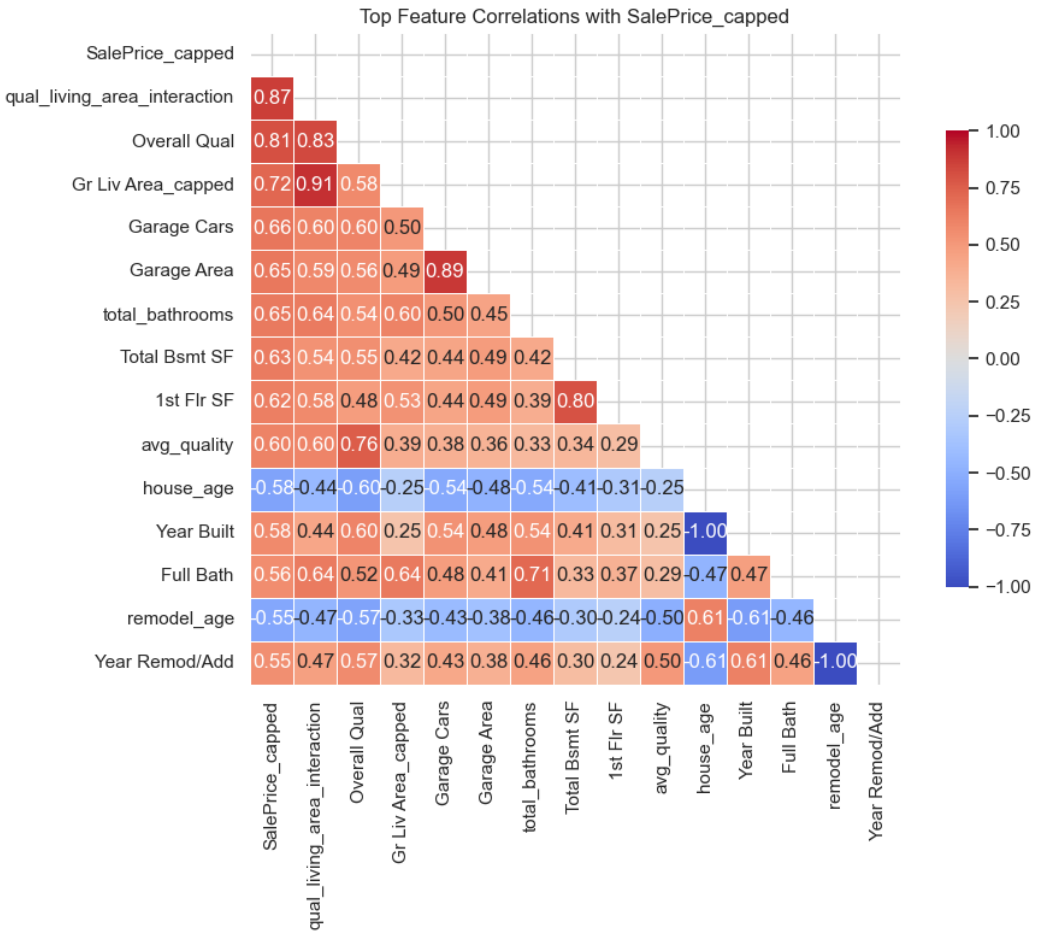




- Clear relationships and strong predictors:
    - **Neighborhood**: Despite some overlap, categories like *NoRidge* and *StoneBr* tend to have distinctly higher median sale prices, showing it's an important location factor.
    - **Overall Quality & Kitchen Quality**: Clear ordinal trends with sale 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:

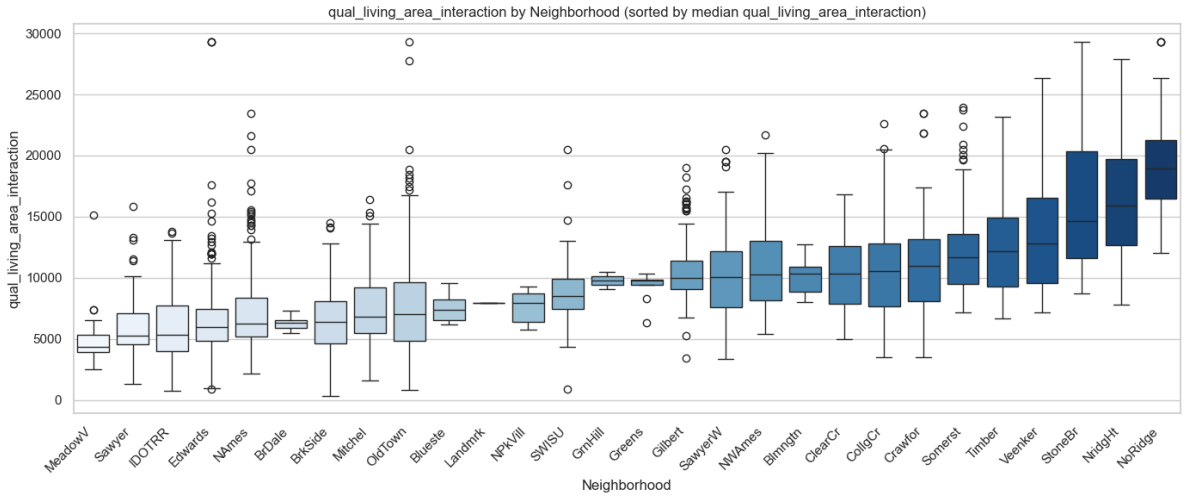
- **Neighborhood** – Most categories show clearly separated median sale prices, indicating that location is a strong driver of housing value.
- **Overall Qual** – Demonstrates a strong ordinal trend with price and other features, reinforcing its reliability as a quality proxy.
- **Garage Type** – Different 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.

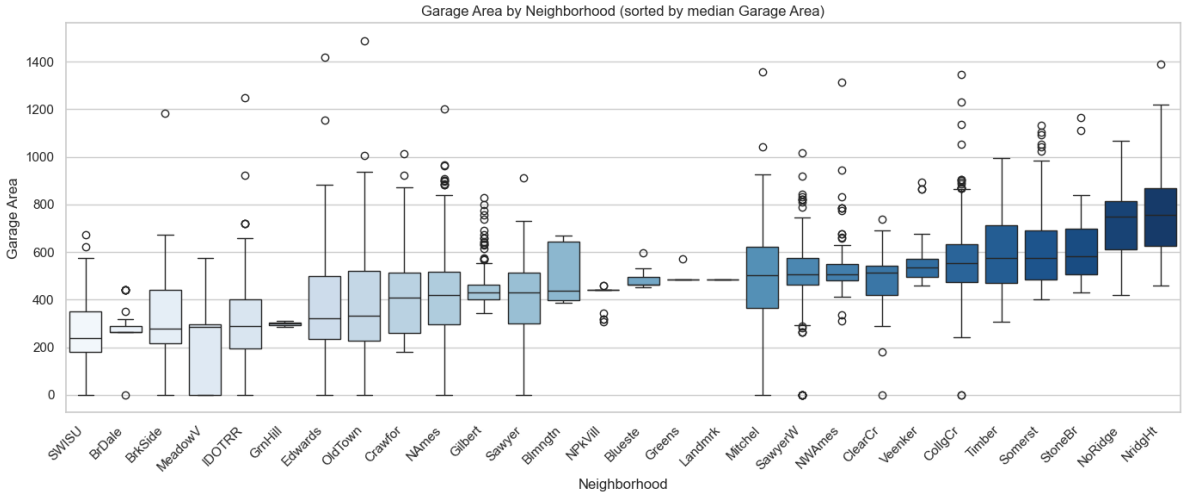
The numerical features selected were:

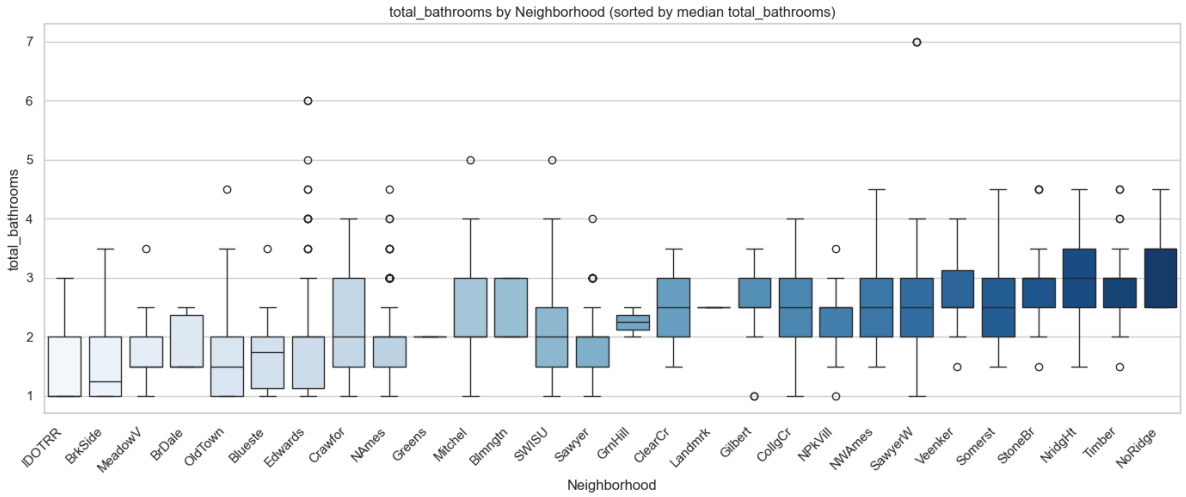
- **qual_living_area_interaction** - Strongest predictor; combines quality and size effects
- **Garage Area** - High correlation with price; reflects useful amenities
- **Total Bsmt SF** - Substantial living/functional space; relevant to house size
- **1st Flr SF** - Core structural metric; often strongly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

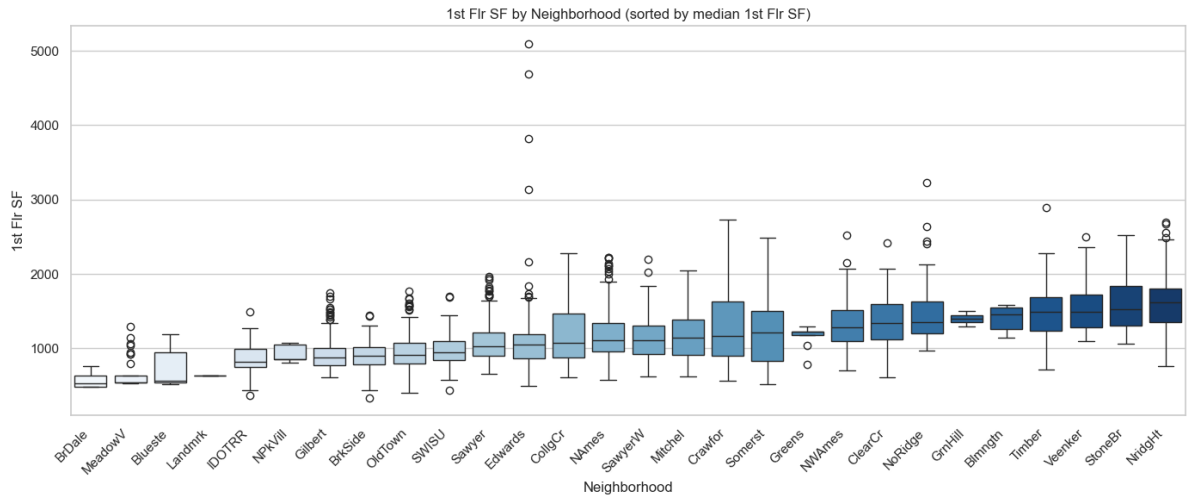
For each set of numerical features against a categorical feature the latter are ordered by median for ease of reading

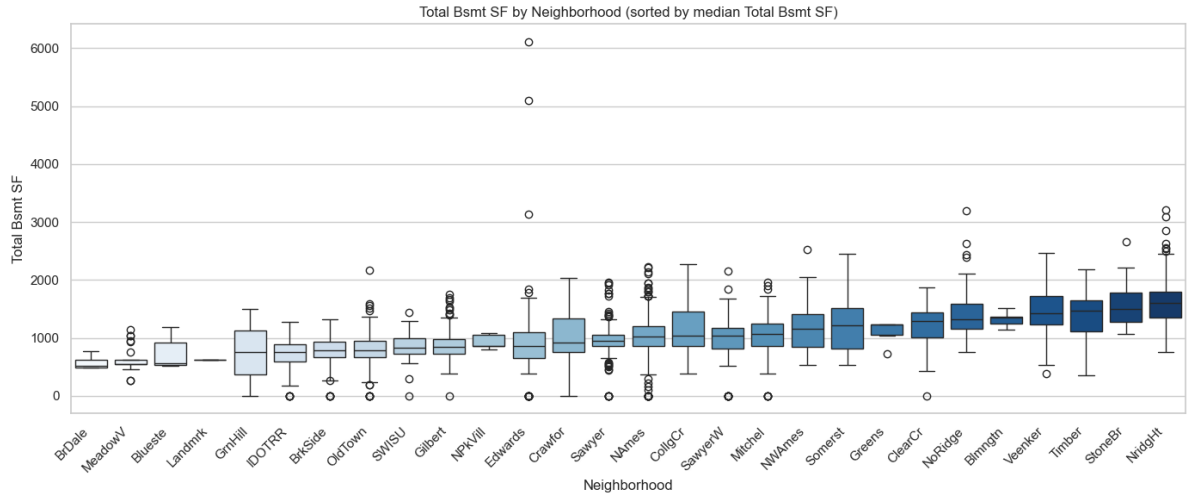
Numerical Features by Neighbourhood









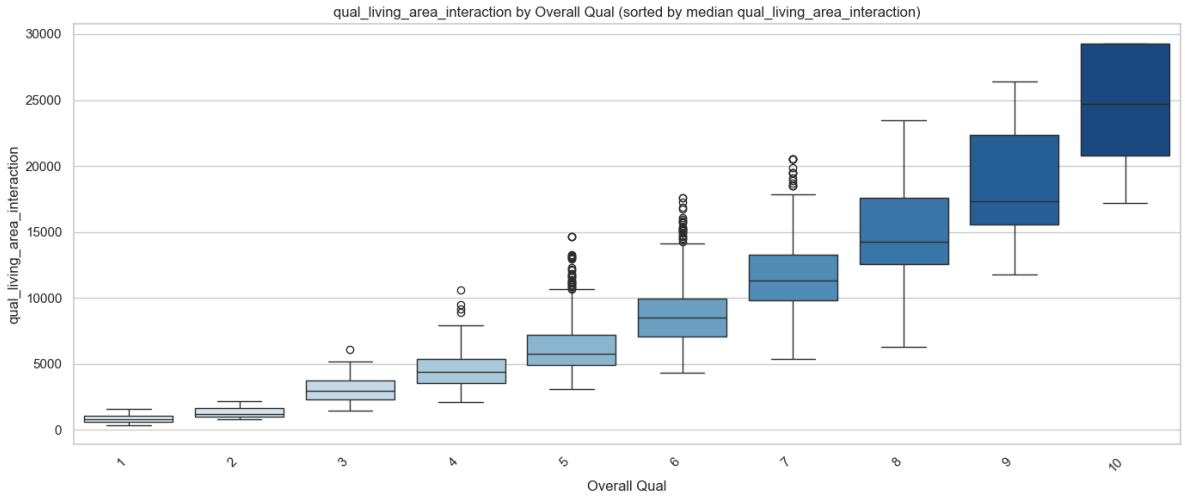


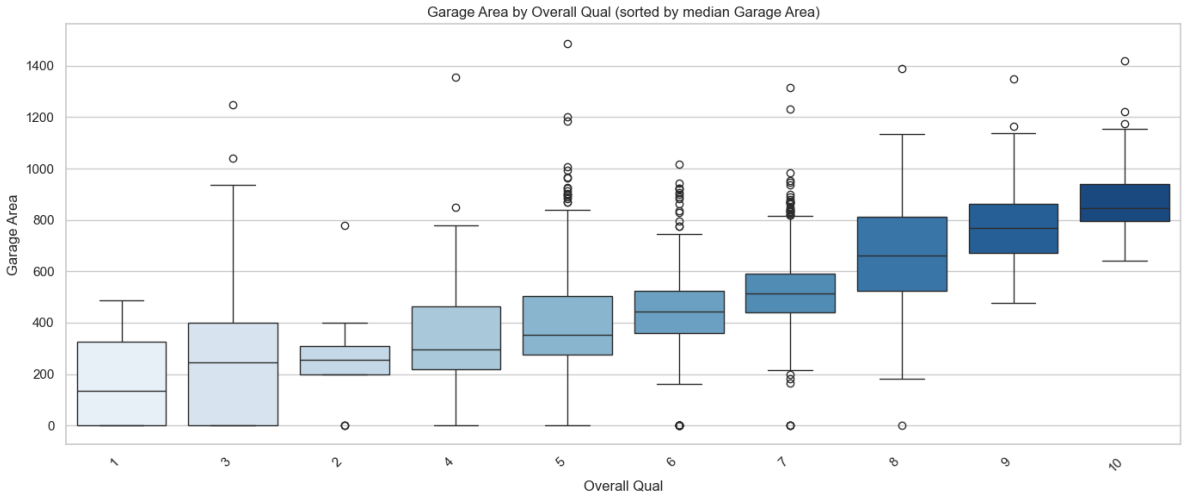
- **Median values** for most numerical features were somewhat to very homogeneous across categories. This homogeneity reduces for `Garage Area` and `qual_living_area_interaction`, where roughly 40% of the category medians are clearly distinct, offering more meaningful comparison.
- **Interquartile Range (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.

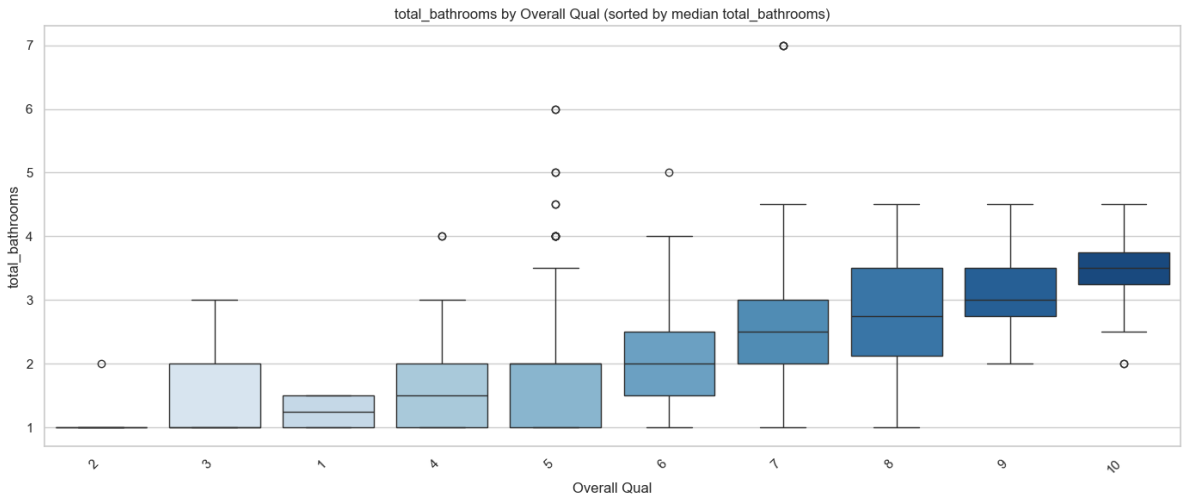
- The **Edwards** neighborhood stands out consistently for its widely dispersed outliers across all five numerical features, though not always numerous. This suggests a particularly high internal variability in home characteristics.

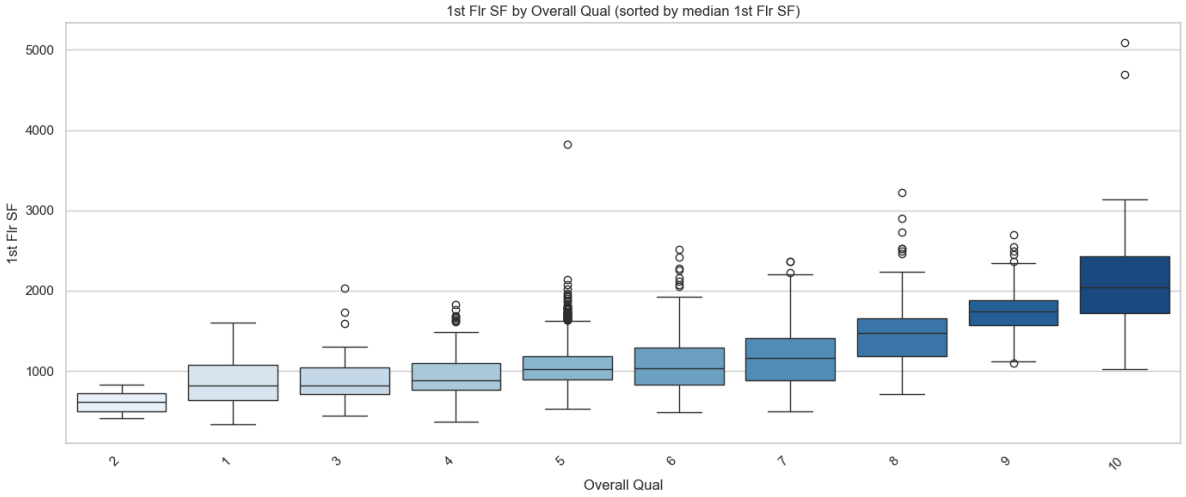
- **Overall,** these plots show that while some numerical features (like `Garage Area` or `qual_living_area_interaction`) vary somewhat across neighborhoods, most exhibit substantial internal variability and overlapping interquartile ranges. In many cases, the apparent differences are likely driven more 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.

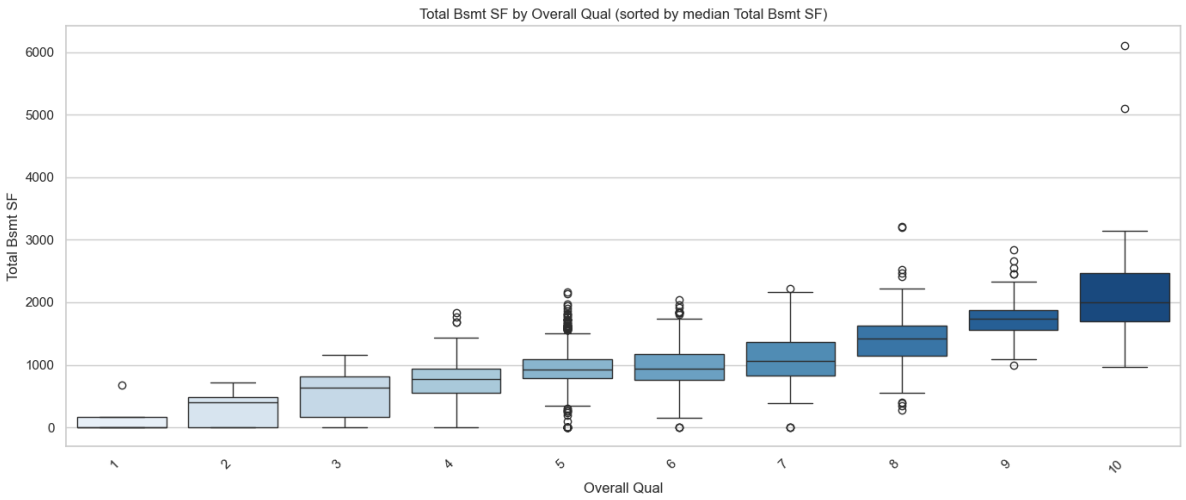
Numerical Features by Overall Qual









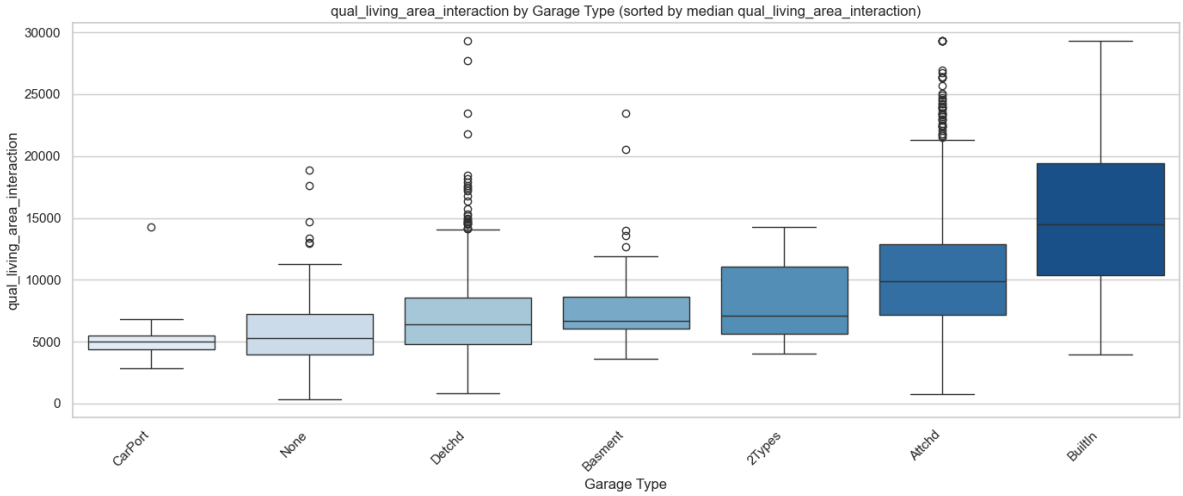


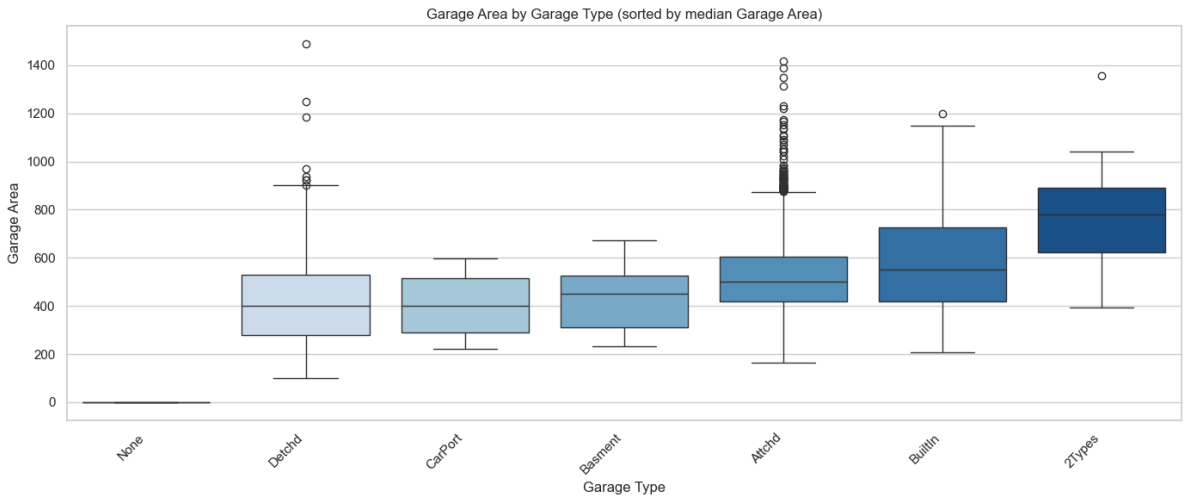
- **Median values** of `qual_living_area_interaction` across `Overall Qual` ordinal categories are clearly distinct, with differences increasing exponentially. This is expected because `qual_living_area_interaction` multiplies living space (`Gr Liv Area`) by `Overall Qual`, so the overall quality acts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

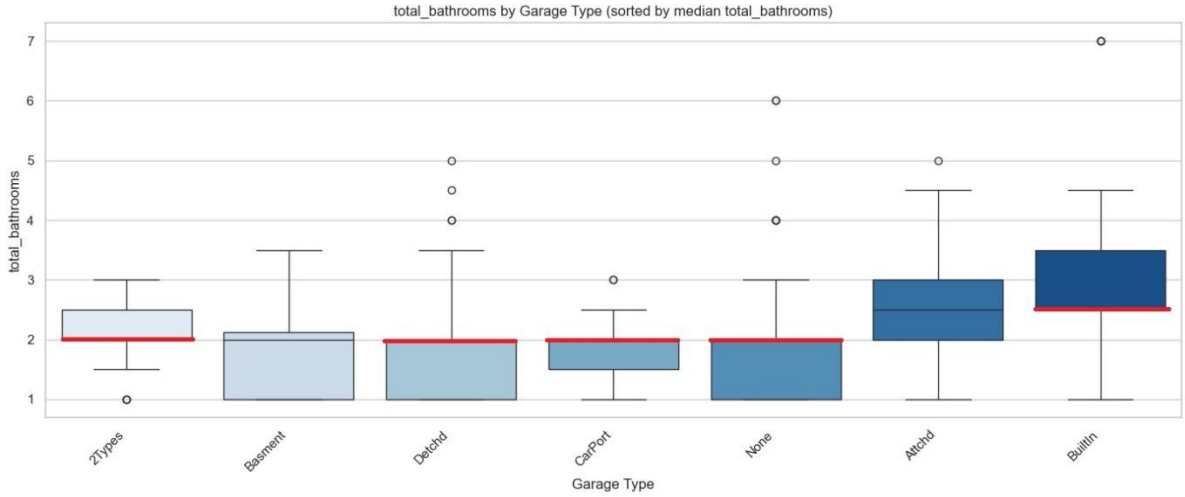
The median values of Garage Area, total_bathrooms, and 1st Flr SF are relatively homogeneous across the lower Overall Qual scores, but become more distinct and evenly spaced as Overall Qual increases. In contrast, Total Bsmt SF shows more consistent medians between Overall Qual scores 5 to 7, with greater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

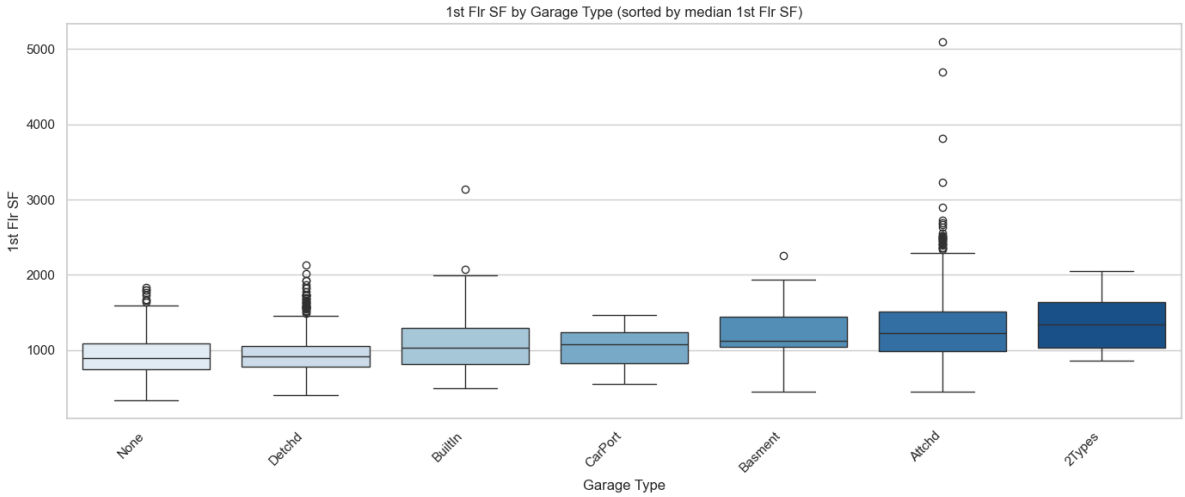
- **Interquartile range (IQR)** overlap is minimal between qual_living_area_interaction and Overall Qual scores — but this is to be expected, as qual_living_area_interaction is derived by multiplying Overall Qual with living area. For the remaining numerical features, there is considerable overlap across 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.

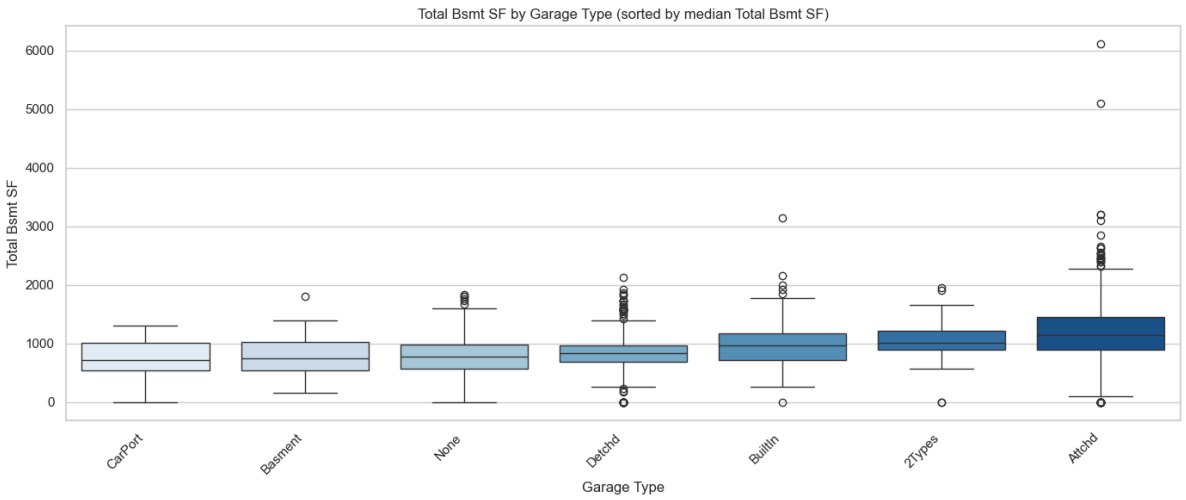
Numerical Features by Garage Type




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

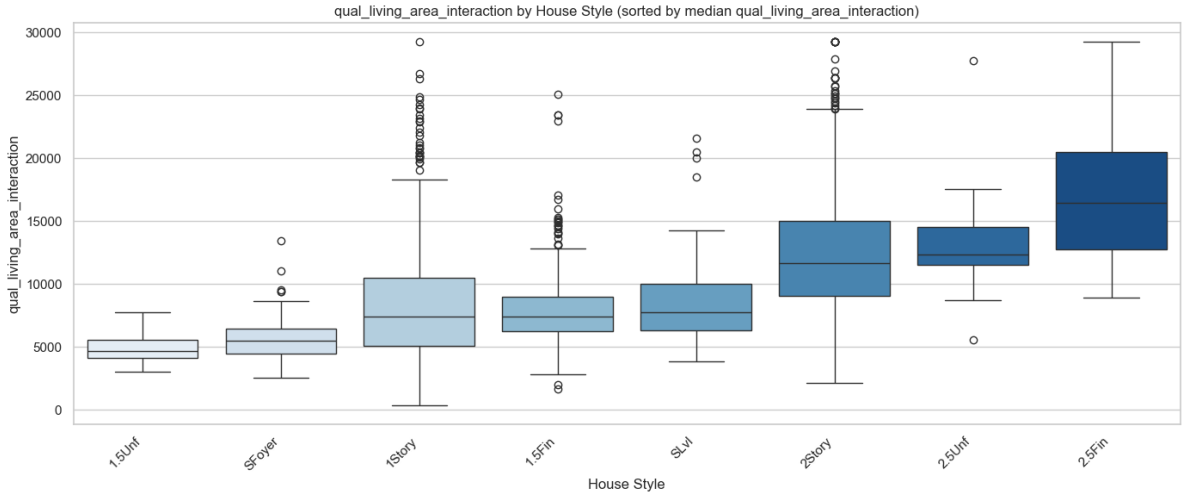

total_bathrooms by Garage Type (sorted by median total_bathrooms)
total_bathrooms
1
2
3
4
5
6
7
2Types
Basment
Detchd
CarPort
None
Attchd
BuiltIn
Garage Type

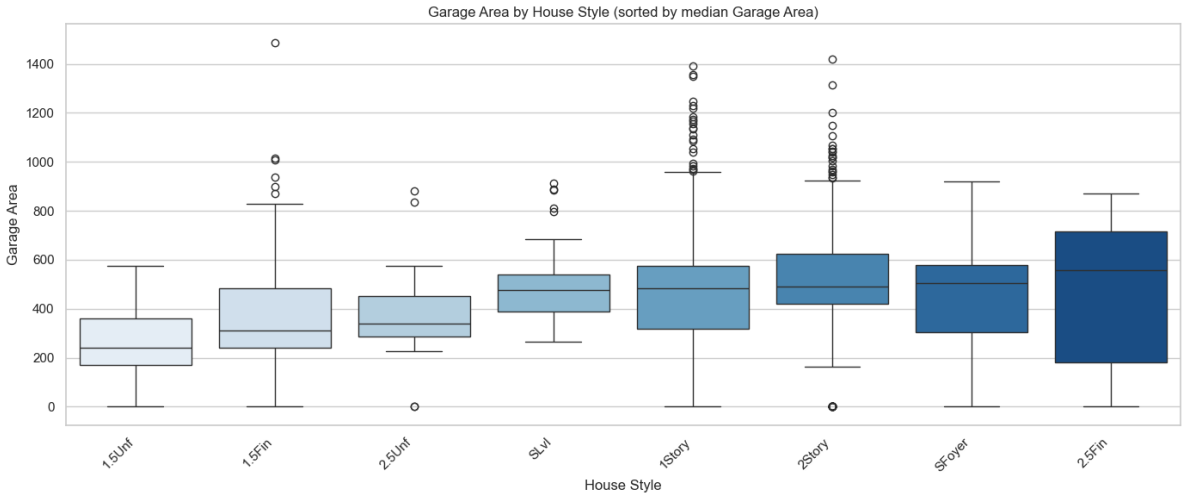

1st Flr SF by Garage Type (sorted by median 1st Flr SF)
1st Flr SF
1000
2000
3000
4000
5000
None
Detchd
BuiltIn
CarPort
Basment
Attchd
2Types
Garage Type

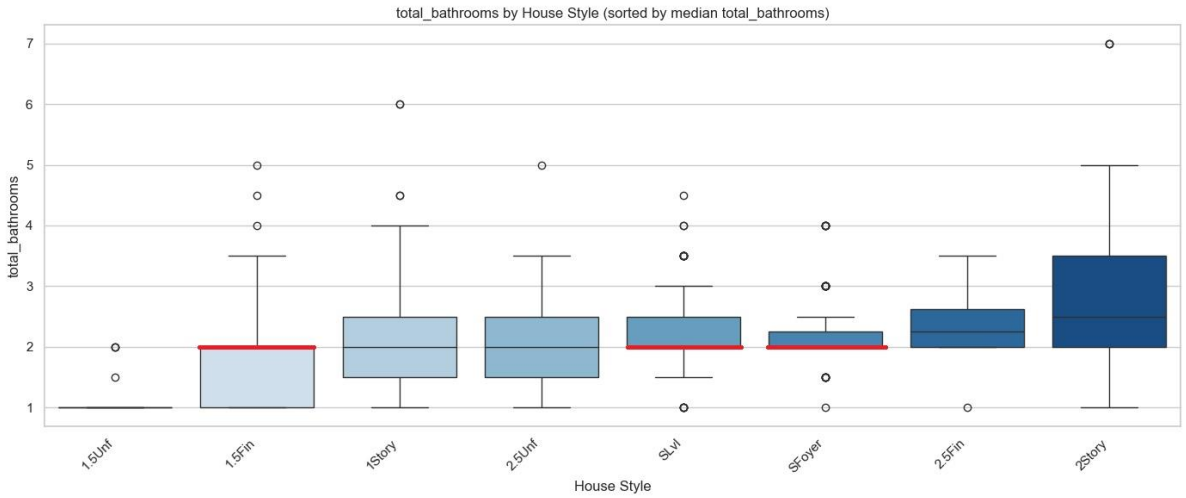


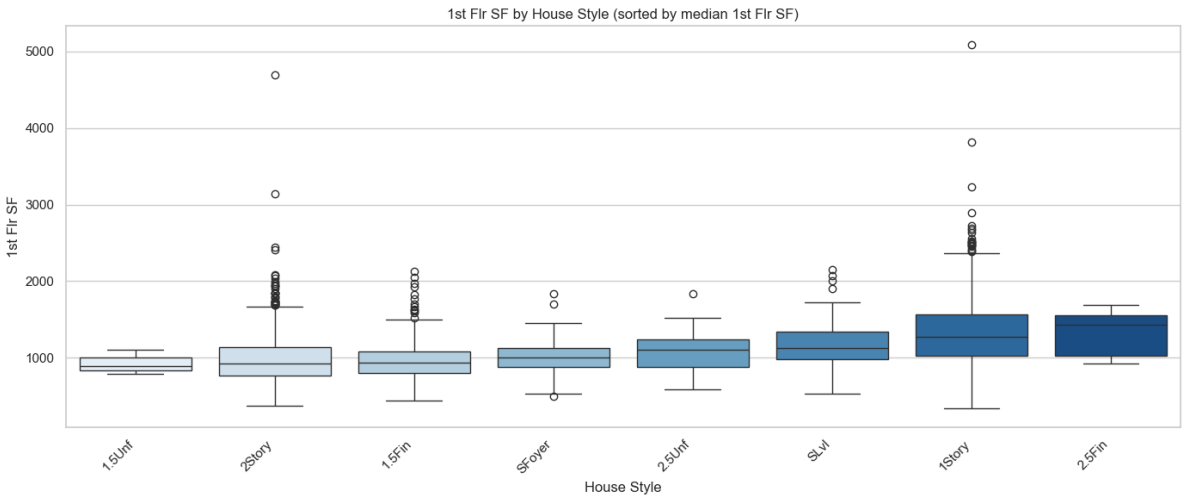
- For `1st Flr SF` and `Total Bsmt SF`, median values and interquartile ranges (IQRs) are fairly homogeneous across all Garage Types, with similar IQR overlaps, suggesting little to no relationship between these features and Garage Type.
- For `qual_living_area_interaction` and `total_bathrooms`, medians 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.

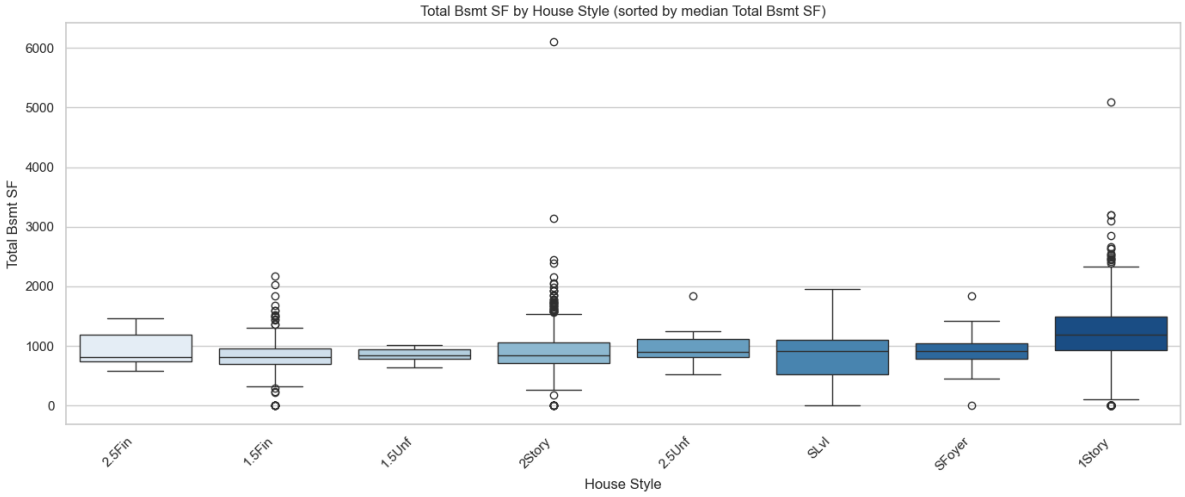
Numerical Features by House Style





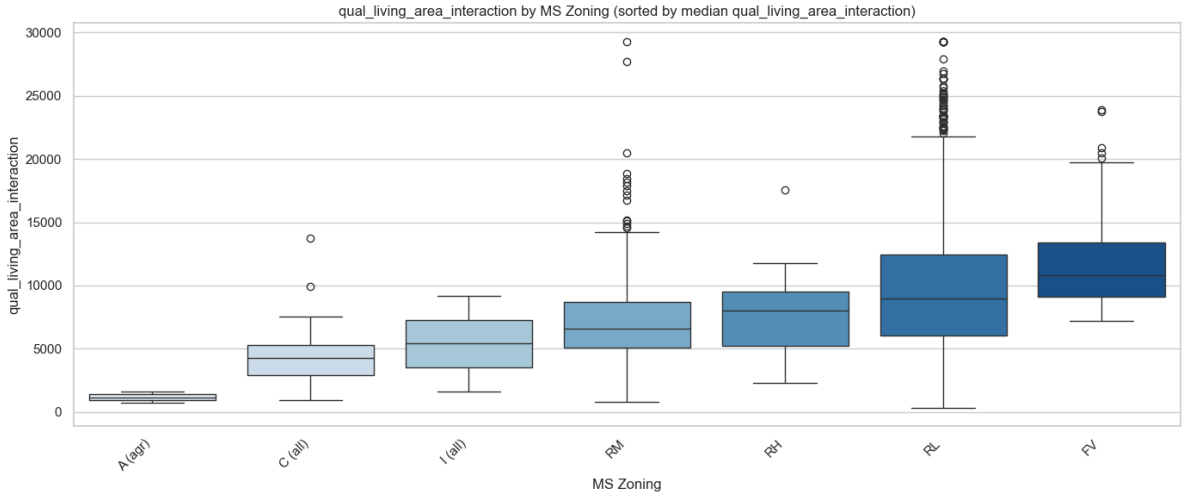


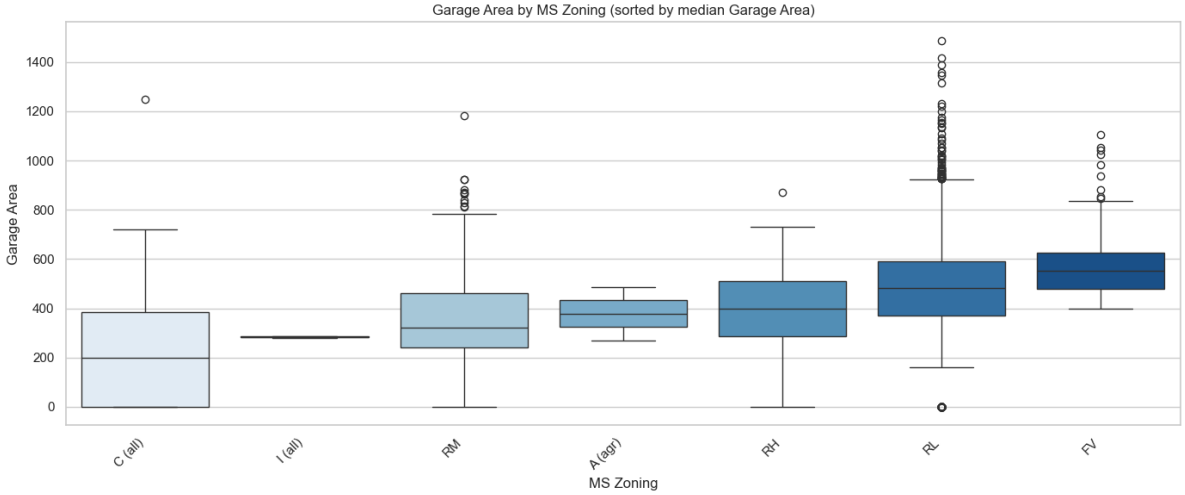


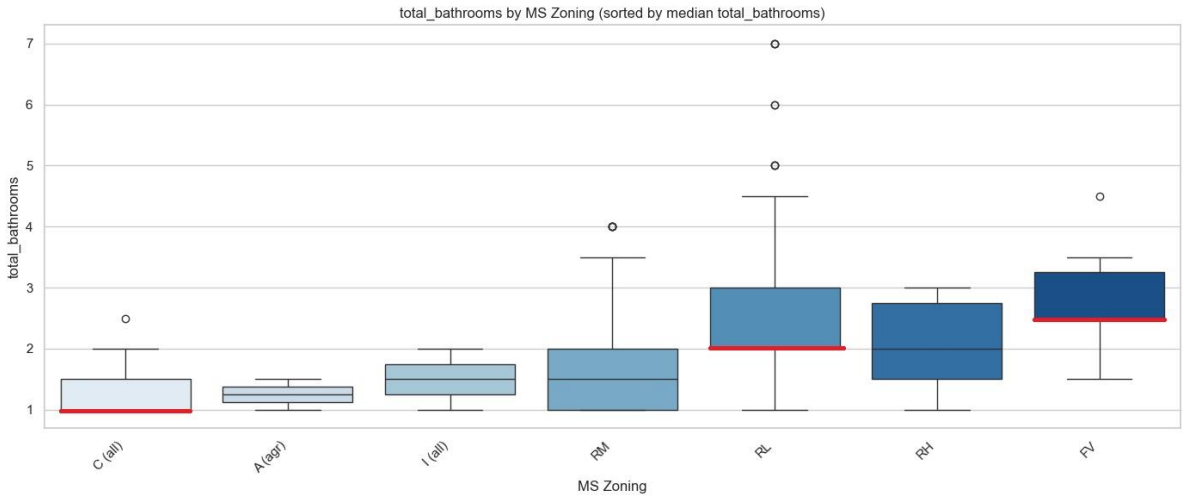


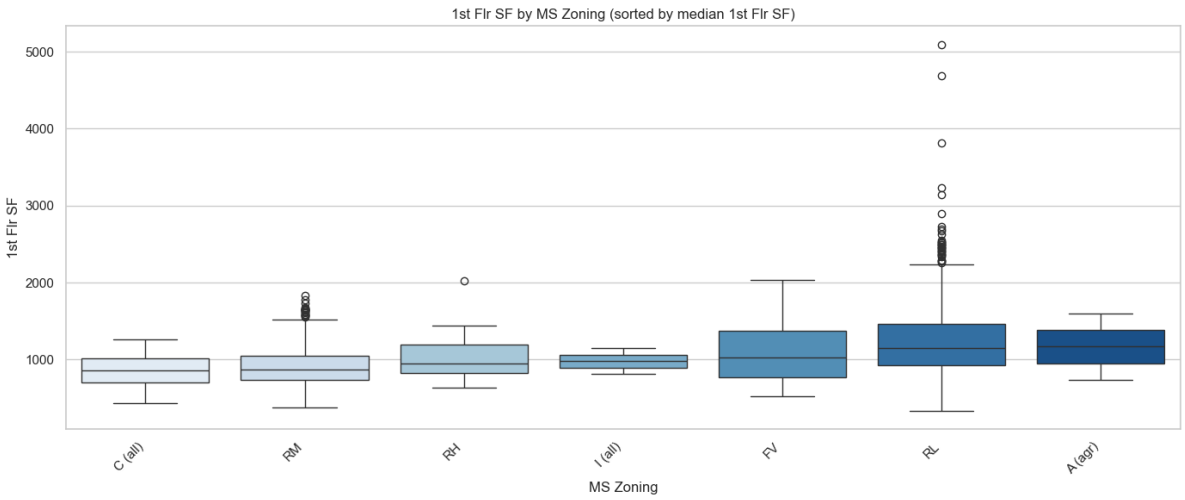
- For `qual_living_area_interaction`, median values tend to cluster into 3–4 overlapping groups with considerable IQR overlap and varied IQR widths.
- `1st Flr SF` and `Total Bsmt SF` show fairly homogeneous medians and IQRs across all House Styles, with substantial overlap, suggesting little relationship 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.

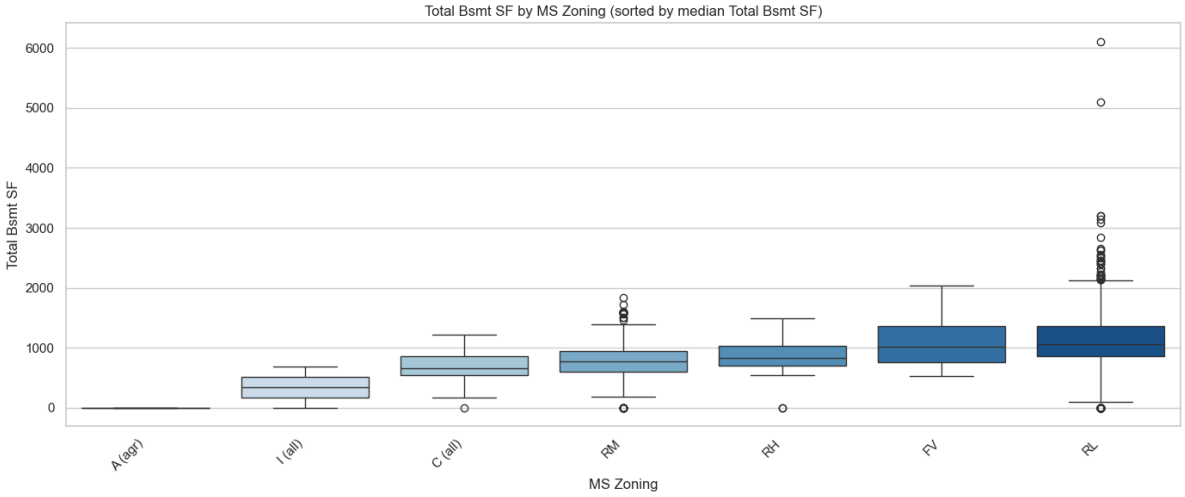
Numerical Features by MS Zoning











- 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.

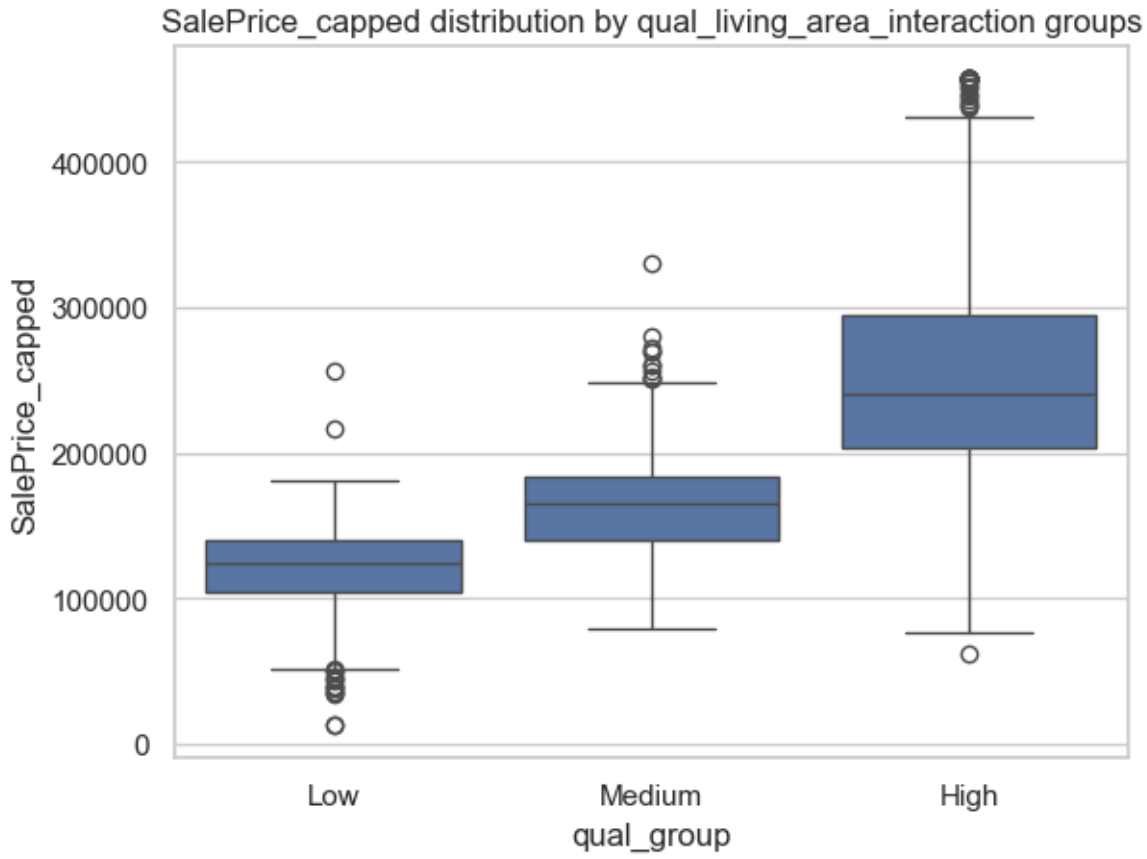
To investigate whether the engineered feature **qual_living_area_interaction** (which combines overall quality and living area) is significantly associated with sale price, the continuous variable was divided into three equal-sized groups (Low, Medium, High) using quantile-based binning. This

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.

Conclusion and Next Steps

Conclusion:
This exploratory analysis of the Ames Housing dataset reveals several important insights about factors influencing house sale prices. Univariate and bivariate analyses show that home size, quality, and engineered features such as `qual_living_area_interaction`—which combines overall quality and living area—are strongly associated with sale price. The multivariate analysis, supported by both correlation matrices and grouped boxplots, underscores that `qual_living_area_interaction` outperforms its individual components in explaining price variation. Location (`neighborhood`) and garage characteristics also play significant roles but exhibit more variability and overlap within categories. Hypothesis testing using ANOVA and Kruskal-Wallis confirmed a statistically significant difference in sale prices across different levels of `qual_living_area_interaction`, affirming its predictive strength.

Next Steps:
Building on these findings, future work could explore more advanced modelling techniques to capture nonlinearities and interactions more effectively. Potential directions include:

- **Further Hypothesis Testing:** Explore additional hypotheses, including those related to `neighborhood` effects and garage characteristics, as well as new ones inspired by domain knowledge or exploratory analyses. Instead of simple binning, future analyses could explore continuous modelling techniques such as regression analysis (linear, polynomial, or generalized additive models) or mixed-effects models. These methods better capture nonlinearities and interactions without arbitrarily grouping continuous variables. Adding confidence intervals to hypothesis tests and effect size estimates will improve the interpretability and practical relevance of the results, complementing significance testing and supporting better decision-making.

- **Statistical Methodology:** While a one-way ANOVA and Kruskal-Wallis provided initial insights into group differences, future analysis could incorporate post-hoc comparisons (e.g.

Tukey's HSD) and regression-based approaches to model continuous relationships more effectively.

- **Include Power Analysis and Effect Size:** To strengthen the statistical rigor, power analysis should be conducted before hypothesis testing to ensure sufficient sample size for detecting meaningful differences. Reporting effect sizes (e.g., eta-squared for ANOVA, Cohen's d for pairwise tests) will help quantify the practical significance of findings beyond just statistical significance.
- **Multiple Comparison Corrections:** Since multiple hypothesis tests may be conducted, it is crucial to apply statistical corrections for multiple comparisons—such as Bonferroni or Benjamini-Hochberg adjustments—to control the increased risk of Type I errors and ensure robust, reliable conclusions.
- **Feature Engineering:** Further refine engineered variables or create interaction terms to enhance predictive power, especially incorporating `neighborhood` and garage features alongside quality and size metrics.
- **Addressing Multicollinearity:** Given observed feature redundancies, dimensionality reduction techniques like Principal Component Analysis (PCA) or feature selection methods could improve model efficiency.
- **Machine Learning Models:** Implement regression-based models such as Random Forests, Gradient Boosting Machines, or XGBoost to leverage multiple features simultaneously and model complex relationships.
- **Model Validation:** Use cross-validation and out-of-sample testing to evaluate model generalizability and prevent overfitting.
- **Incorporate Temporal Factors:** Explore the impact of sale year, remodel age, and other time-based variables more deeply to understand market trends.
- **Outlier Analysis:** Investigate high-value outliers in certain neighborhoods and garage types to assess their influence and consider robust modelling approaches.

Overall, this analysis provides a strong foundation for developing predictive models that incorporate meaningful domain insights and maximize interpretability for housing price estimation.

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.