Team: Mannan Mishra, Suvass Ravala, Fuling Wang

GitHub Repository: https://github.com/UC-Berkeley-I-School/Project2_Mishra_Ravala_Wang

Primary dataset: link

Main question: What home features best improve Sale Price per Square Foot for Renovators?

As a home developer and renovator, home sale price is a critical factor determining business success. In the real estate industry, it is well understood that square footage influences a property's market value. Larger homes with many amenities tend to command higher prices. However, builders may not always have the luxury of working with expansive tracts of land required for the above developments.

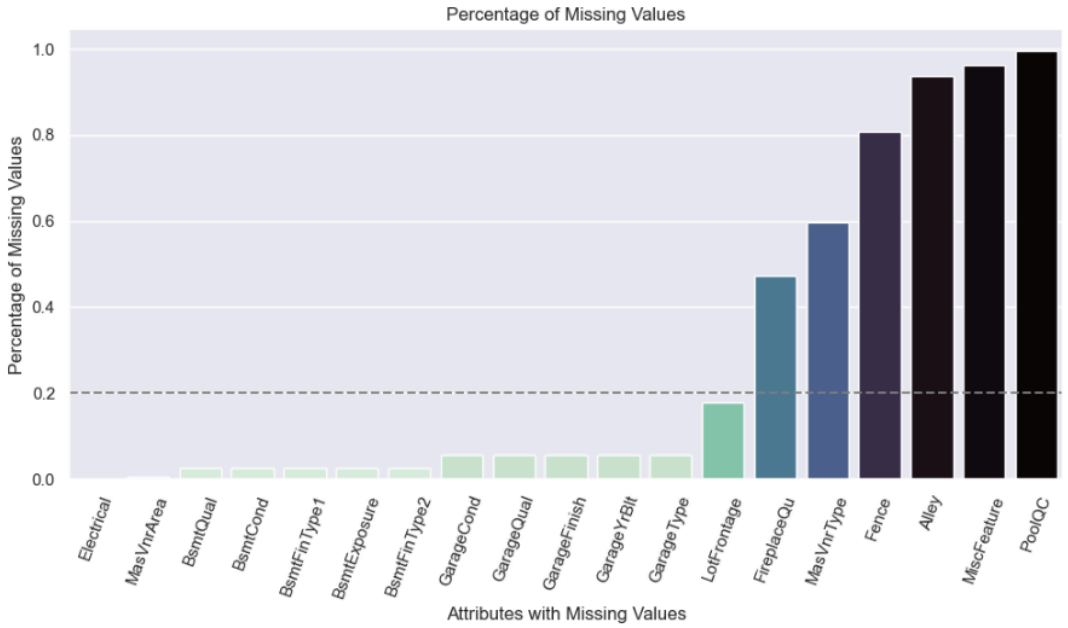
Therefore, our team recommends builders leverage price per square foot as a key metric to evaluate the success of their projects. Regardless of the type of home being developed, the primary objective should be to maximize the highest possible price per square foot.

In our analysis, our team explores optimizing the price per square foot to provide insights to builders. Specifically, we examine the housing landscape of Ames, Iowa, using a house prices dataset. The dataset includes 79 attributes, 36 quantitative and 43 qualitative, which we aim to transform into actionable insights that inform decision-making for home builders and renovators.

Data Cleaning - Ensuring data sanity

Let's understand the data we are working with. Identifying which of the 79 attributes are critical for our price per square foot analysis is essential to generate relevant insights. The house pricing dataset includes data on 1,460 homes in Ames, Iowa, encompassing attributes such as home sale price, Grlivarea (above ground living area), basement quality, etc…

Before continuing our analysis, we must first clean the dataset by identifying and addressing missing values (nulls). Null values in our dataset can compromise the accuracy of our insights so we need to calculate the percentage of nulls for each attribute to assess the severity of its missing data. To calculate the percentage of nulls, we find the count of null values in each column, then we divide that count by the total number of values for each column. Below are the results from this calculation:



As seen above, PoolQC, MiscFeature, Alley, Fence, and FireplaceQu have over 40% of their data as null values. Consequently, we have decided to remove these attributes from our dataset for the below reasons:

1. Data Integrity: Attributes with high percentages of null values may compromise the integrity of any analysis. Substituting missing values for these attributes would require assumptions that may not accurately represent the data distribution. For example, filling in null values for PoolQC based on the average of the other values will skew results. Furthermore, assigning PoolQC to houses without pools will misrepresent those values.
2. Lack of Value: Given the large percentage of nulls, these attributes represent a small subset of the population. This significant proportion of missing values makes these attributes unreliable for analysis, as the lack of data introduces high levels of uncertainty. This inherent rarity suggests that these attributes will have limited application in our findings.

By streamlining this dataset (removing sparsity), we aim to improve the clarity of our dataset.

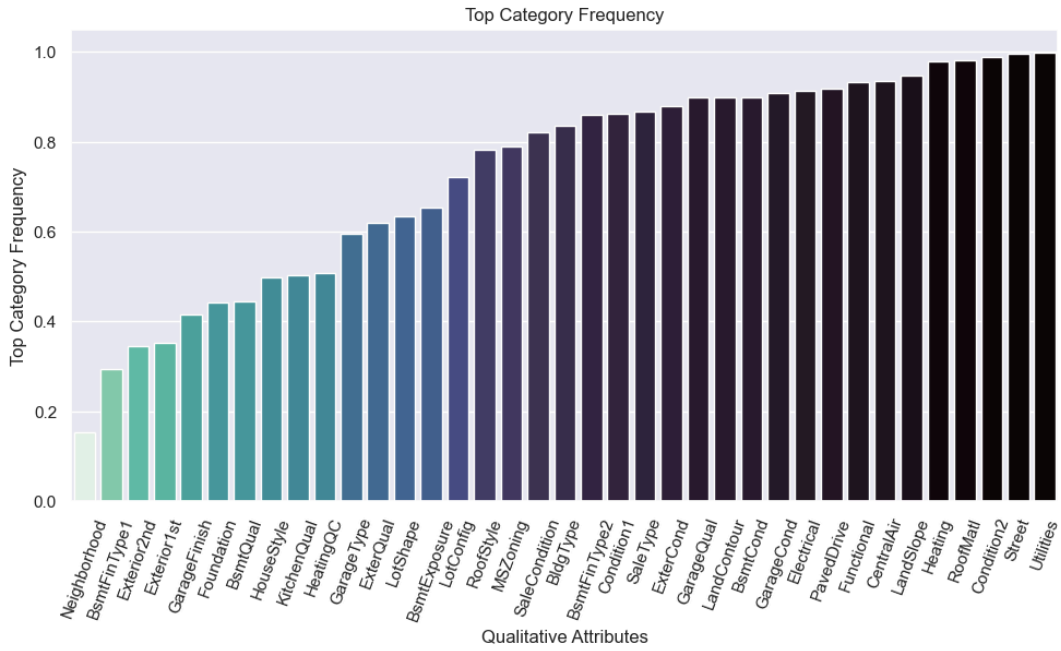
Feature Selection: Enhancing Model Interpretability
Qualitative Features: Distribution and Statistical Evaluation

Now that we cleaned our dataset, we want to focus our analysis on the most impactful attributes. While our dataset contains 79 attributes, conducting an in-depth analysis of each one is neither practical nor efficient. Rather, we can use statistical analysis to condense our search, identifying attributes most relevant to price per square foot.

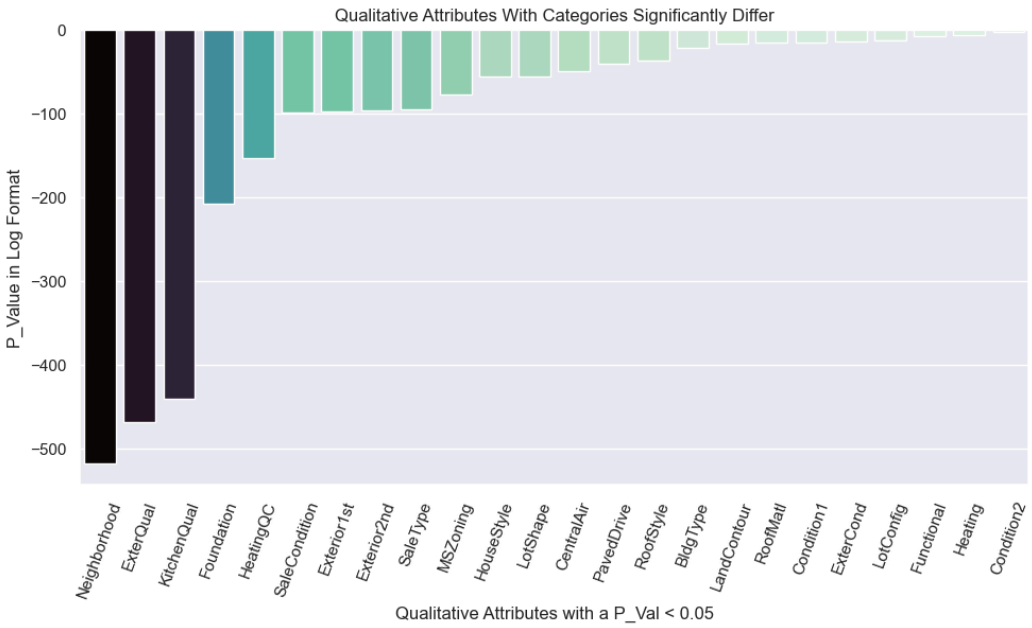
First we start by systematically assessing qualitative attributes to ensure their relevance. The focus is on their statistical and distribution properties in order to reveal desired patterns or undesired redundancies. We consider:

1. Distribution: Attributes with highly imbalanced distribution, where a single category dominates the results, are less likely to contribute meaningfully.
2. Relationship to Target Variable: To determine the utility of our qualitative attributes, we need to isolate attributes that show meaningful differences in relation to sale price per square foot. Statistical tests, such as one-way ANOVA, can help quantify whether the differences among groups are statistically significant, thereby justifying their inclusion in the model.
3. Multicollinearity and Redundancy: Categorical features that are highly correlated with one another may introduce multicollinearity (when independent variables exhibit strong linear relationship). These features need to be identified and addressed.

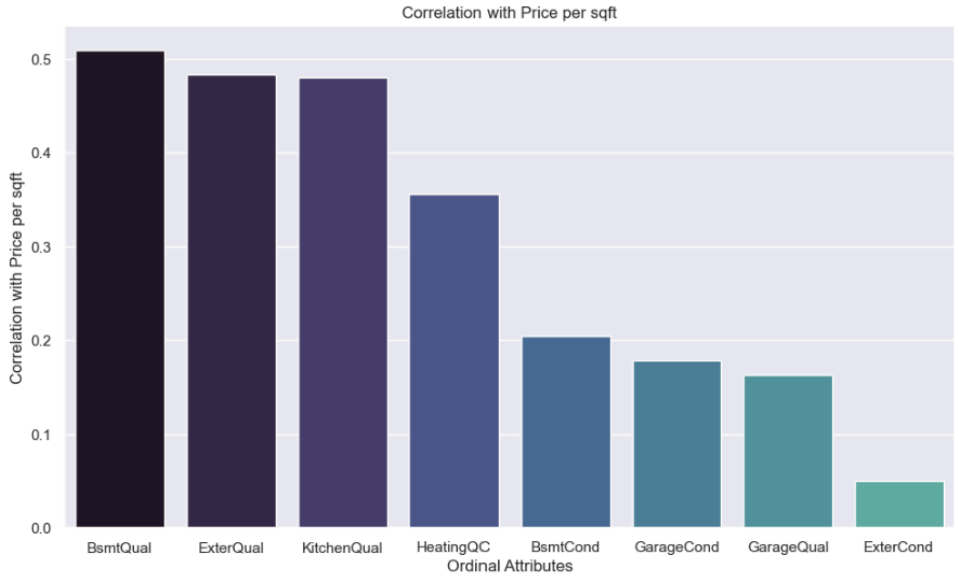
1. Visualizing the top-category frequencies revealed that several qualitative features are heavily concentrated in their dominant categories. These were further analyzed to determine whether their distribution adds meaningful variance to the model.



2. We then use ANOVA to identify whether the differences among groups were statistically significant, in other words, whether the average price per sqft varies significantly across categories. Features demonstrating significant group differences were prioritized for inclusion.



Quantifying Ordinal Features:



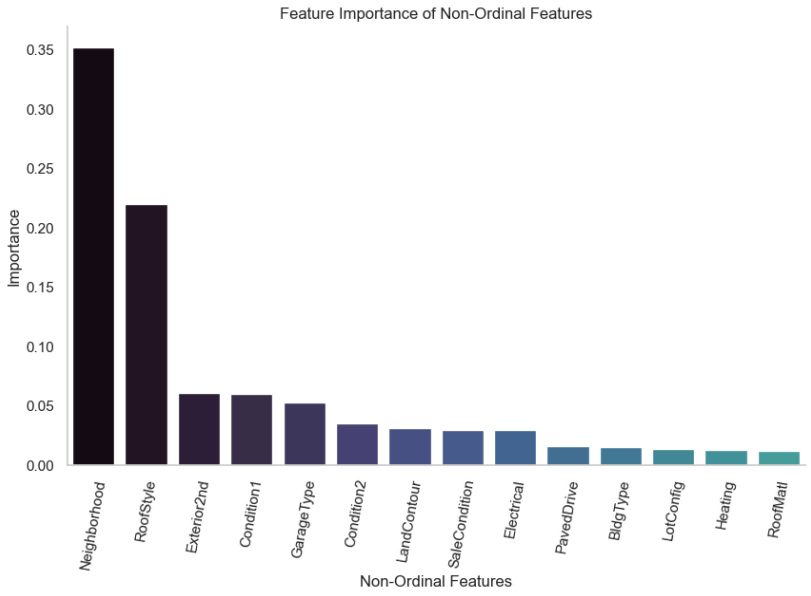
We then divide the qualitative attributes into two groups: ordinal vs non-ordinal. Ordinal attributes, often related to quality (e.g., **Kitchen Quality**, **Basement Quality**), were mapped to numerical values to facilitate correlation analysis with the target variable. This transformation provided a clear measure of how quality levels influence property prices.

Key Insights:

- **External Quality, Basement Quality**, **Kitchen Quality and Heating Quality** showed strong correlations with prices/sqft.
- **External Quality**, **Kitchen Quality and Heating Quality** were particularly notable due to their balanced categories and absence of missing data, further validating their reliability.

Quantifying non-Ordinal Features:

In addition to conducting ANOVA tests, we used a decision tree to quantify feature importance, providing insights into the most informative features to classify different ranges of price per sqft. See below graph for the results.

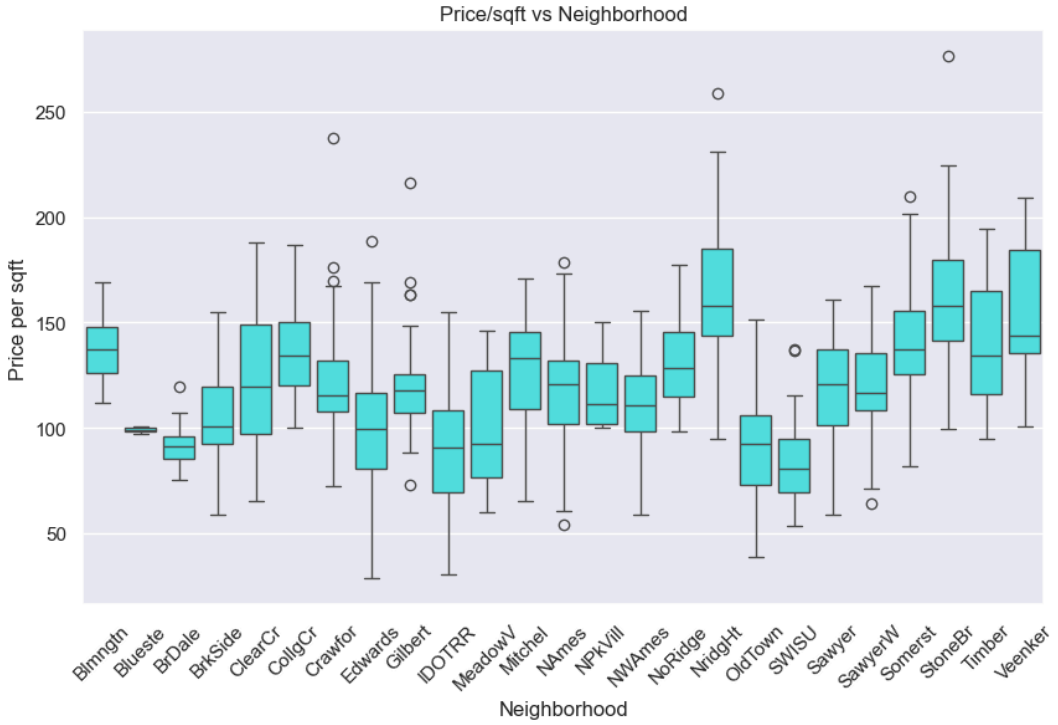


From the above analysis, we identified a set of 8 ordinal and 10 non-ordinal attributes that are important to investigate. The Ordinal features are BsmtQual, ExterQual, KitchenQual, HeatingQC, FireplaceQu, BsmtCond, GarageCond, GarageQual The Non-ordinal features are Neighborhood, RoofStyle, Exterior2nd, Condition1, GarageType, Condition2, LandContour, SaleCondition, Electrical, and PavedDrive.

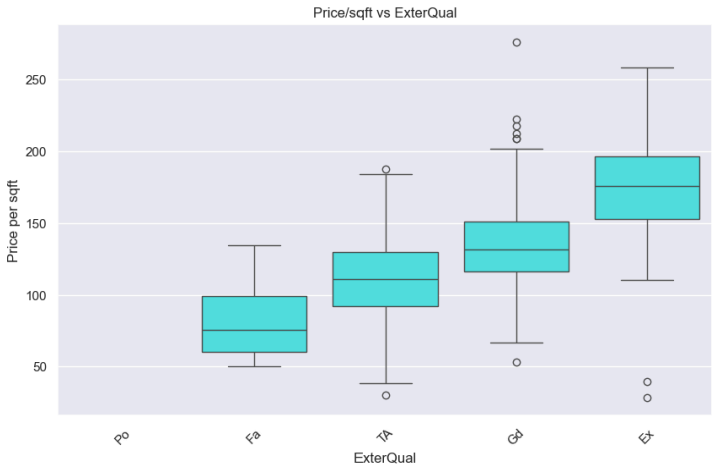
Note: The features (BsmtQual, FireplaceQu, BsmtCond, GarageCond, GarageQual, GarageType, Electrical) did not pass our ANOVA test (results were nan), indicating that they may be outliers or have imbalanced group distributions. Therefore we will exclude these attributes from further analysis. Excluding these problematic features leaves us with a refined list of **3 ordinal** and **8 non-ordinal** attributes.

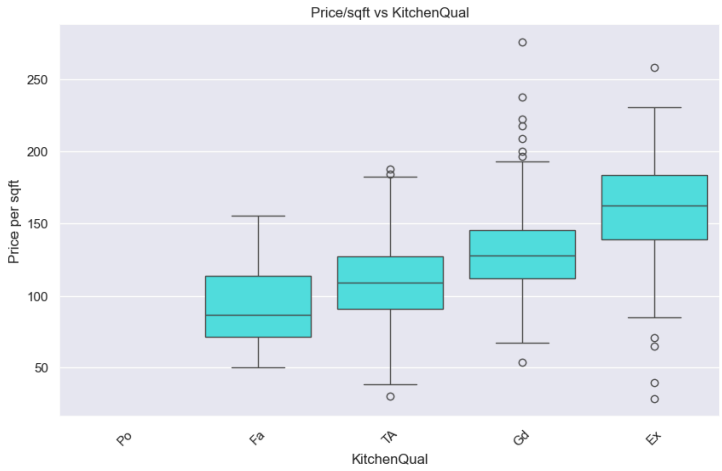
Results of Further Analysis:

From the above 11 attributes, below we will outline the attributes with the most defined outcomes.



The variation in sale price per square foot across different neighborhoods provides key insights into the role of location in determining property value. What we can see from the above histograms is that every neighborhood in Ames, Iowa has a different price per square foot distribution. For example, the StoneBr neighborhood has a far greater mean price per square foot than SWISU. The main insight from analyzing neighborhoods is that builders can distinguish which neighborhoods to prioritize luxury designs over cost-effective designs. Builders in high price per square foot neighborhoods will be more likely to recoup money on luxurious amenities over neighborhoods with low price per square foot. Therefore, builders should analyze the nearby neighborhoods and base their design philosophy on the location.

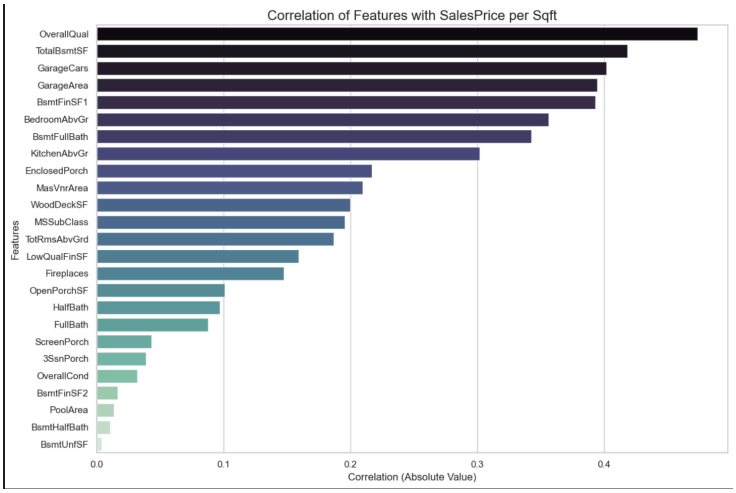




The two charts above signify the relationship between exterior quality, kitchen quality and price per square foot. The quality categories - Excellent(Ex), Good(Gd), Typical(TA), Fair(Fa), and Poor(Po) - show that higher quality exteriors and kitchen correlates with higher prices per square foot. Improving the exterior quality (removing mold, repainting, fixing the garden) and kitchen quality (new appliances and countertops) can significantly improve home value. Prioritizing these enhancements will be a cost-effective strategy to attract buyers and achieve premium pricing. Further analysis on quality is explained below in the quantitative section.

Quantitative Data:

Correlation:

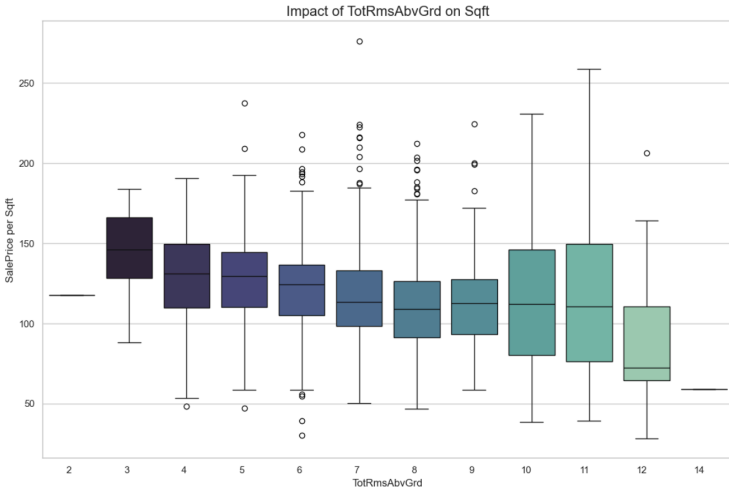


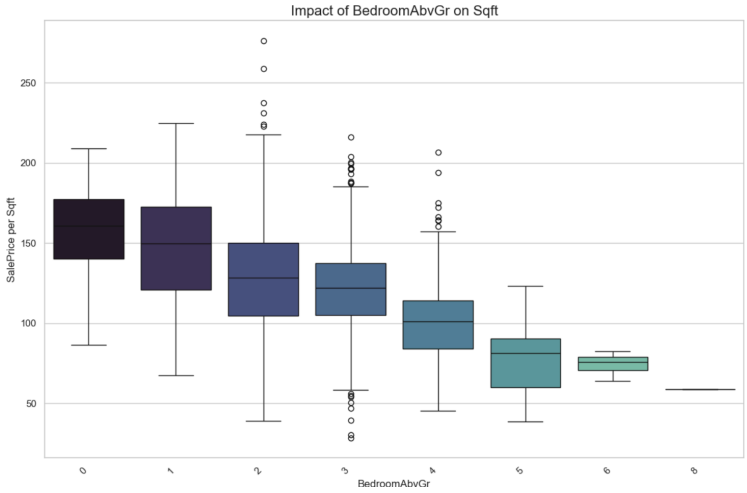
Now that we have explored qualitative data, we turn our focus to quantitative data. The first step in our analysis was to identify and exclude attributes that cannot be influenced by renovators, such as *Year Sold*. Additionally, we removed collinear variables, like *Living Area Square Feet*, to avoid redundancy and potential distortions in the analysis. Then we observed which variables had the highest correlation to identify potential variables that best signal higher

price per square foot. A strong candidate was Overall Quality, which relates to our qualitative insights above.

Below are highly correlated candidates with definitive insights:

Rooms:



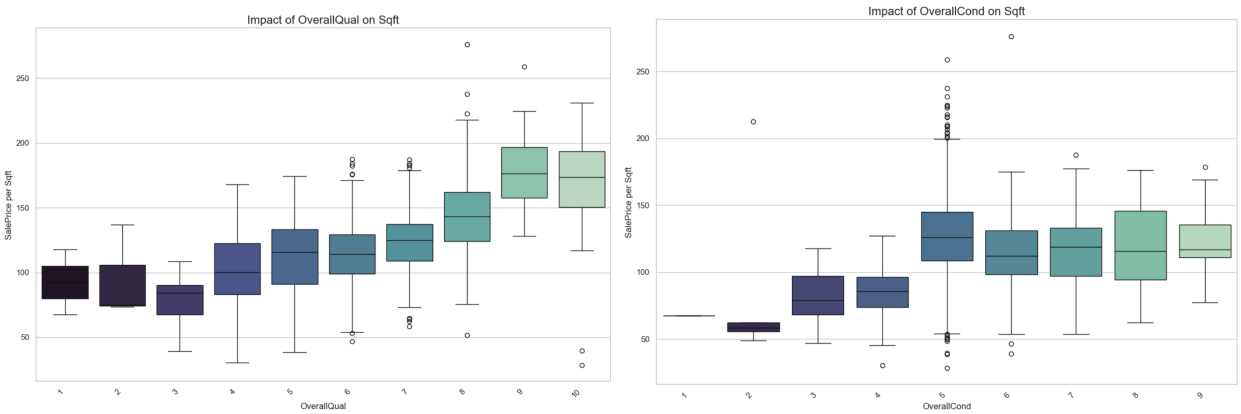


The above charts highlight the nuanced relation between room count, bedroom count, and house value. When we subset the data to isolate trends between rooms, bedrooms, and price we observe an interesting trend:

- Total number of rooms: Adding more rooms above ground did not show a clear trend in increasing or decreasing the price per square foot. This trend suggests that simply increasing the number of rooms without considering their function, room area, or quality does not necessarily add value to a home
- Bedrooms: However, adding more bedrooms above ground was associated with a decrease in price per square foot highlighting low value especially when a home has more than 4 bedrooms. This phenomenon is likely because expanding bedrooms without expanding square footage may indicate cramped rooms. The number of bedrooms alone do not contribute significantly to the functionality of a house, especially if they come at the expense of comfort or larger living spaces like kitchens.

These findings advise developers to be strategic when adding rooms to a house. Rather, developers should focus on optimizing the layout, improving existing spaces, or adding multifunctional rooms. These findings suggest that thoughtful design and comfort take precedence over quantity of rooms in driving price per square foot.

Condition and Quality:



The first chart reveals a strong positive correlation between Overall Quality and Price per Square Foot. As overall quality improves, price per square foot increases. This trend suggests that buyers place a premium on high quality construction and finishes. This finding implies that upgrades to enhance overall quality (modern kitchens, updated bathrooms, improved structural supports) are likely to yield substantial returns.

The relationship between Overall Condition and price per square foot is more nuanced. From conditions 1 to 5, price per square foot consistently increases. However, between condition levels 5 and 10, the mean price per square foot remains constant. This plateau suggests that once a home reaches a “good enough” condition, further improvements to condition may not significantly boost home value.

Renovators can obtain strong advice from these findings. While improving the condition of a property is beneficial, renovators should evaluate the cost-benefit of condition improvements beyond average. Rather, once the home has reached a “good enough” condition, improvements on overall quality (high end finishes, appliance upgrades, system upgrades) are more likely to maximize return on investment. Essentially, homes in average condition but with higher quality will command more premium than homes in excellent condition but lower quality.

ML Model:

We trained and tested a Machine Learning model as a fun side project to see whether an ML model agrees with our findings. From our ensemble machine learning model, which includes the models Linear Regression, Random forest, XGBoost, and Gradient Boosting, we saw that the 3 biggest factors that contributed to the price per square foot were Overall Quality, Garage Cars and Overall Condition. Overall Quality and Overall Condition were also major attributes analyzed in our exploratory data analysis above.

Final Takeaways:

The primary objective of this project was to derive insights for developers on features to maximize price per square foot. Based on our analysis, we recommend the following strategies:

1: Prioritize Neighborhood Research

The location of a property plays an important role in its value. As seen in our analysis, neighborhoods in Ames, Iowa vary significantly in their mean price per square foot. Developers should align their designs with neighborhood trends:

- High price per sqft neighborhoods: Prioritize luxury amenities and premium finishes. Buyers in these neighborhoods will be more likely to pay, enabling return on investment.
- Low price per sqft neighborhoods: focus on cost effective designs while still maintaining quality. This strategy will ensure better returns on investment without spending on unnecessary premium finishes.

2: Optimize Room and Bedroom Design

Simply increasing room count does not directly translate to higher price per square foot. Developers need to focus on the functionality and comfort of spaces rather than pure room count. Buyers prioritize utility and comfort over quantitative measures.

Adding more bedrooms, especially beyond four, can decrease price per square foot. For context, this finding relates to when bedrooms are added without increasing square footage. Recklessly adding more bedrooms means compromising the size of another space leading to cramped designs. Instead, developers should prioritize thoughtful layouts that maintain comfort and spaciousness.

3: Balance Condition and Quality Improvements

Both overall condition and overall quality influence price per square foot but their impacts differ:

- Overall Condition is essential, nobody wants to buy a home in poor condition. However, once a home reaches an average condition, further improvements yield diminishing returns.
- Improving overall quality will drive up price per square foot. High quality features like modern kitchens and well maintained exteriors drive up home value more effectively than condition alone.
- Developers should aim to fix the house to a decent condition, then allocate the remaining investments towards quality finishes, maximizing profitability.

Our exploratory data analysis highlights the importance of aligning home development efforts with market expectations to reap profitable results. By focusing on neighborhood research, thoughtful room layout, and high quality finishes, developers can maximize their profits within the constraints of limited square footage. These insights are versatile and can be applied to home construction of any scale, ensuring home developers enhance their price per square foot, and consequently property value.