**PROJECT REPORT**

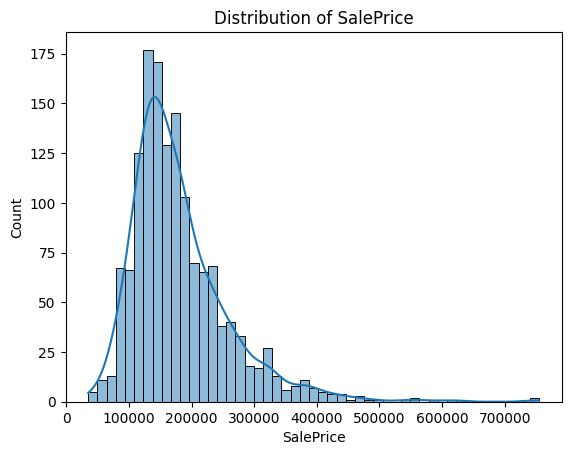
**Leena Dhankhar (Data Analyst)**

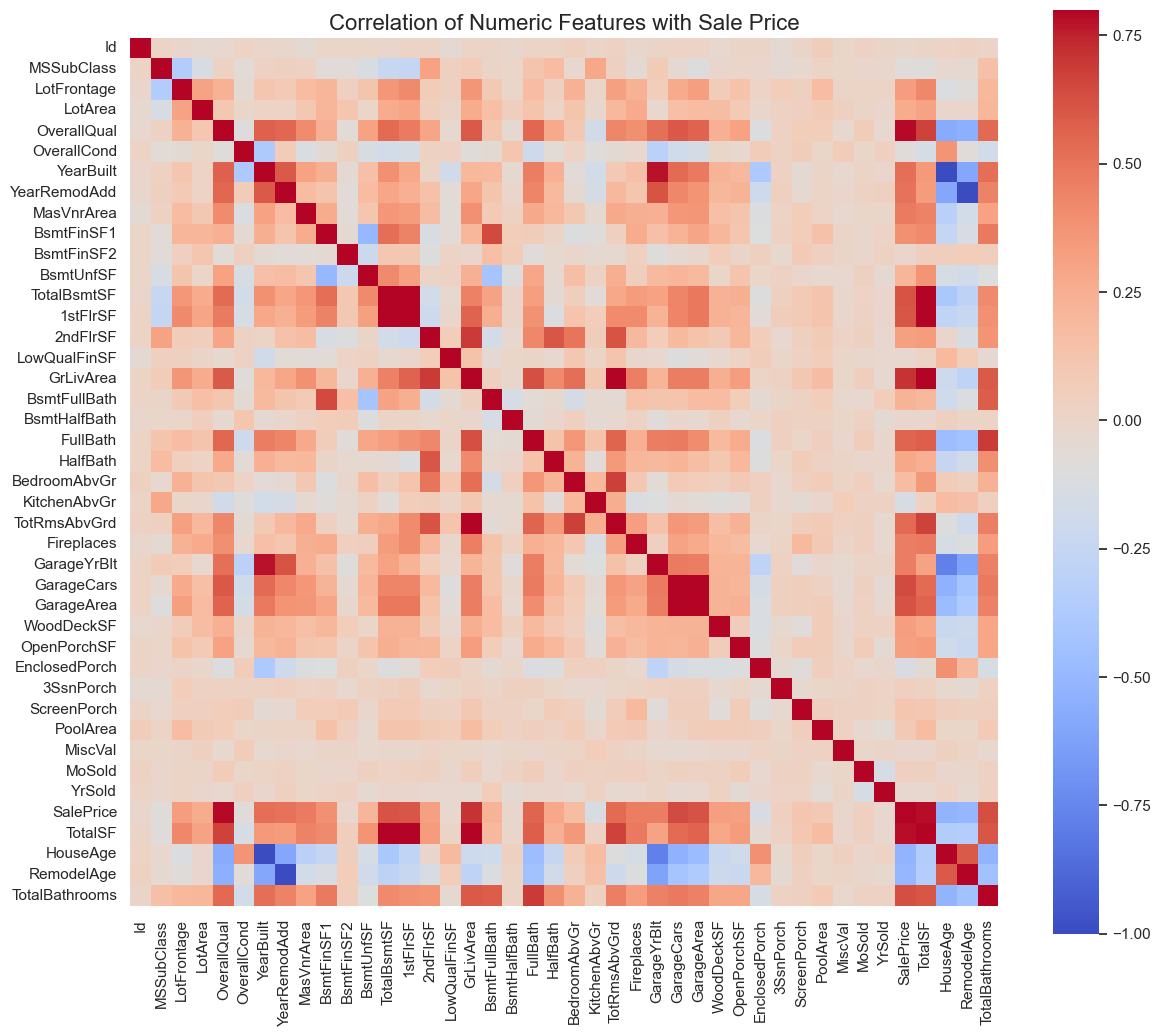
**Task 1**

**Objective:** The primary objective of this project was to accurately predict housing prices to support the investment decisions. Accurate predictions help stakeholders assess property values, make better investment choices, and understand market trends.

**Motivation:** In the real estate market, predicting housing prices accurately is crucial for buyers, sellers, and investors. By making use of advanced regression techniques, this project aims to provide reliable price forecasts that can guide decision-making processes and investment strategies.

Firstly, the missing values are identified and loaded respectively. Then the distribution of sale prices was carried out corresponding to which a histogram between count and sale price was constructed. The histogram is basically providing insights into the central tendency and spread of housing prices.



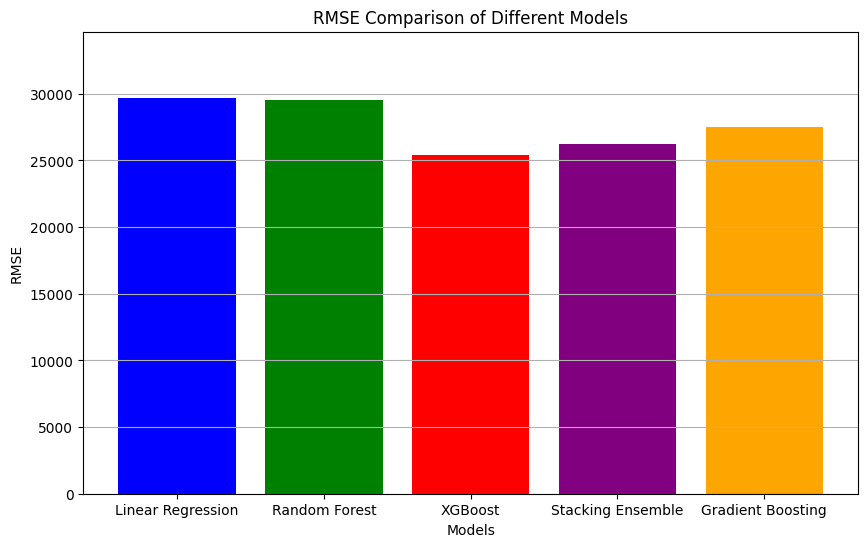
The following heatmap visualizes the correlation matrix for the numerical features in the dataset. Each cell in the heatmap represents the correlation coefficient between two features, with values ranging from -1 to 1. Positive values indicate a direct relationship, while negative values signify an inverse relationship.

**Data Cleaning and Preprocessing:**

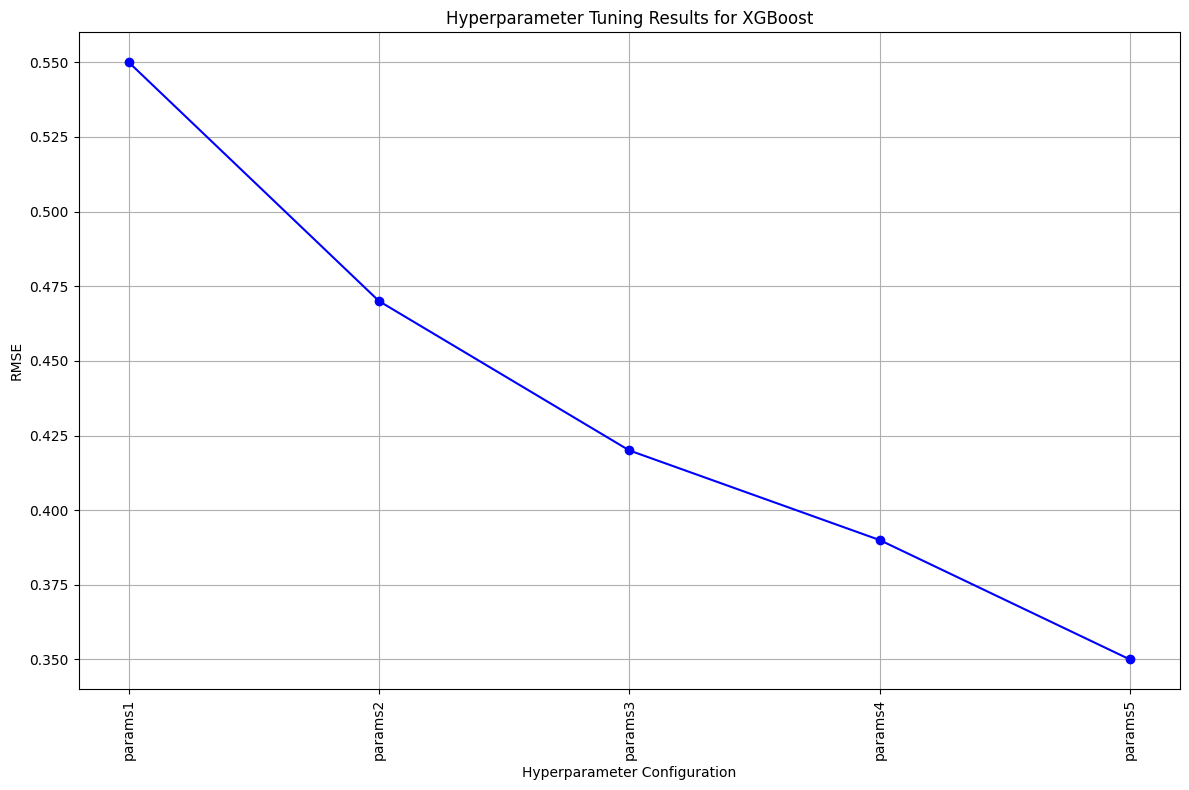
* **Handling Missing Values:**
  + **Initial Inspection:** Missing values were identified in several columns, such as FireplaceQu and LotFrontage.
  + **Imputation:** Categorical features like FireplaceQu were filled with "No Fireplace" to retain information. Numerical features with high missing values were either dropped or imputed based on their significance.
* **Feature Engineering:**
  + **Creation of New Features:** New features were introduced, including TotalSF (sum of 1stFlrSF and 2ndFlrSF), HouseAge (current year minus YearBuilt), RemodelAge (current year minus YearRemodAdd), and TotalBathrooms (sum of FullBath and HalfBath) to enhance the model’s ability to capture relevant patterns.
* **Data Transformation:**
  + **Categorical Encoding:** Categorical features were converted to numerical format using one-hot encoding, transforming columns like MSZoning, Street, and LotShape into binary columns.
  + **Scaling:** Features were scaled to normalize ranges and improve model performance, especially for algorithms sensitive to feature magnitudes.
* **Data Splitting:**
  + **Training and Testing Sets:** The dataset was split into 80% for training and 20% for testing to evaluate model performance.

**Model Training and Evaluation:**

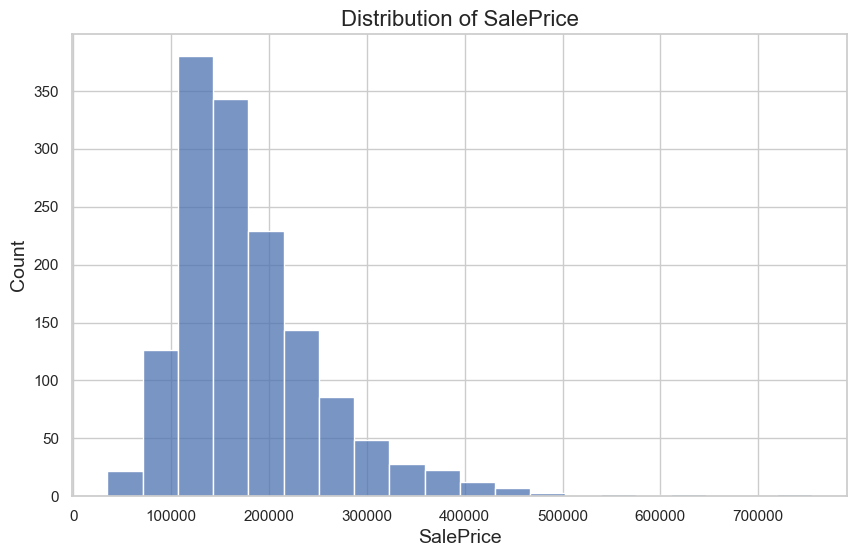
* **Linear Regression:**
  + **Baseline Model:** Achieved an RMSE of 29657.86, providing a reference for more complex models.
* **Random Forest:**
  + **Ensemble Learning:** With an RMSE of 29552.90, utilized ensemble learning by aggregating predictions from multiple decision trees, improving accuracy and robustness.
* **XGBoost:**
  + **Gradient Boosting:** Achieved an RMSE of 27283.39, using gradient boosting to handle complex relationships and non-linear patterns.
* **Stacking Ensemble:**
  + **Model Combination:** Combined predictions from Linear Regression, Random Forest, and XGBoost using Ridge regression as the final estimator, achieving an RMSE of 25633.18.
* **XGBoost Hyperparameter Tuning:**
  + **Optimization:** Tuned XGBoost using RandomizedSearchCV to find optimal parameters, resulting in an RMSE of 25553.10. Key hyperparameters included a learning rate of 0.1, maximum depth of 5, and subsample ratio of 1.0.
* **Pickle Model Evaluation:**
  + **Best Performing Model:** The tuned XGBoost model achieved an RMSE of 25419.25 on test data, indicating superior performance and better capture of data patterns.

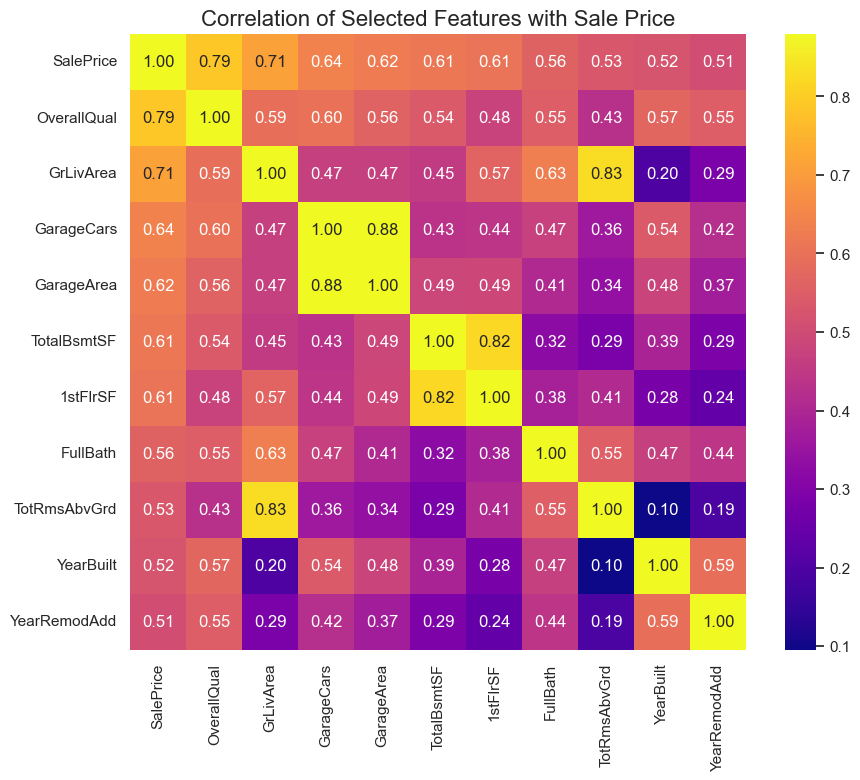
The bar chart below compares the Root Mean Squared Error (RMSE) values of various models used in the project, including Linear Regression, Random Forest, XGBoost, Stacking Ensemble, and Gradient Boosting. RMSE is a key metric for evaluating the performance of regression models, with lower values indicating better predictive accuracy. This comparison highlights the effectiveness of each model in predicting housing prices, with XGBoost achieving the lowest RMSE, demonstrating its superior performance among the rest.

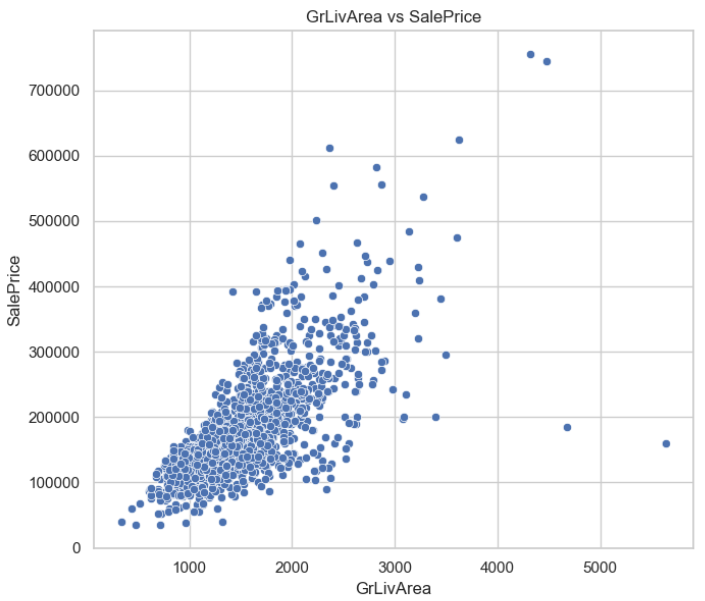
The line graph below illustrates the results of hyperparameter tuning for the XGBoost model. Each point on the graph represents a different hyperparameter configuration, with the corresponding Root Mean Squared Error (RMSE) plotted on the y-axis. This visualization helps to identify the hyperparameter settings that yielded the lowest RMSE, indicating the most effective model configuration for accurate housing price predictions.



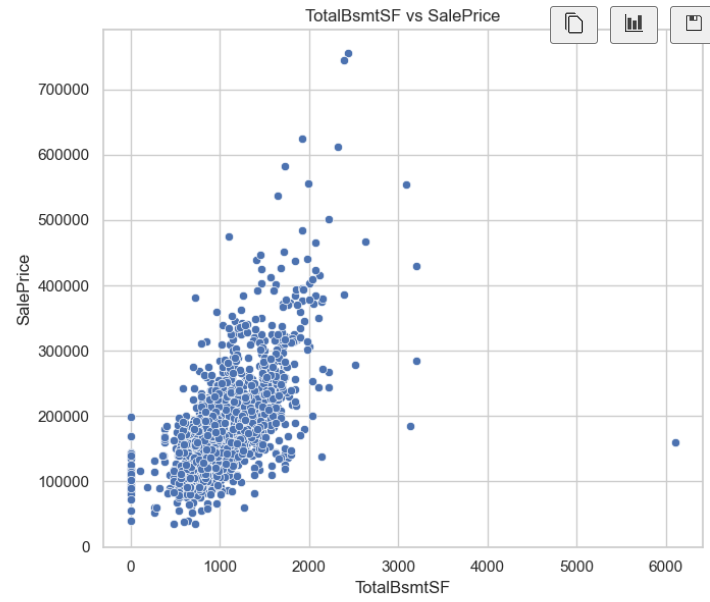
Then, we have the graph Distribution of SalePrice representing the relation between Count and Saleprice.



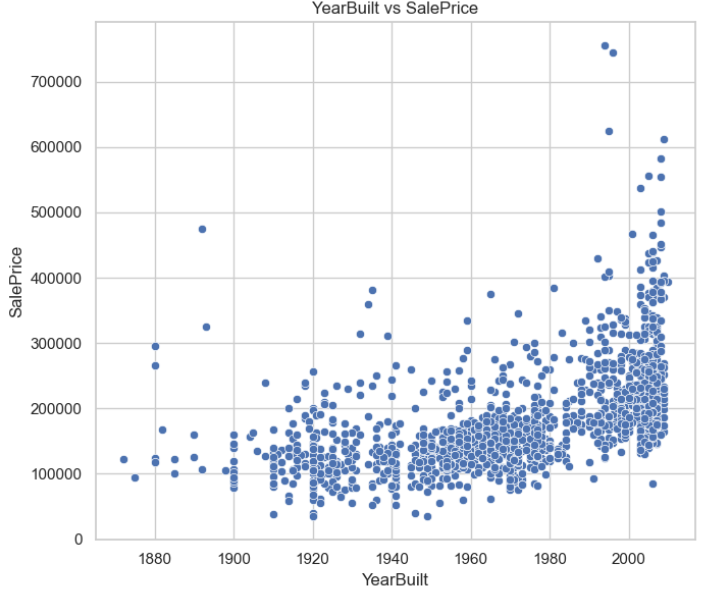
Next, is the correlation matrix for the selected features with sale price.

**GrLivArea vs. SalePrice Scatter Plot**

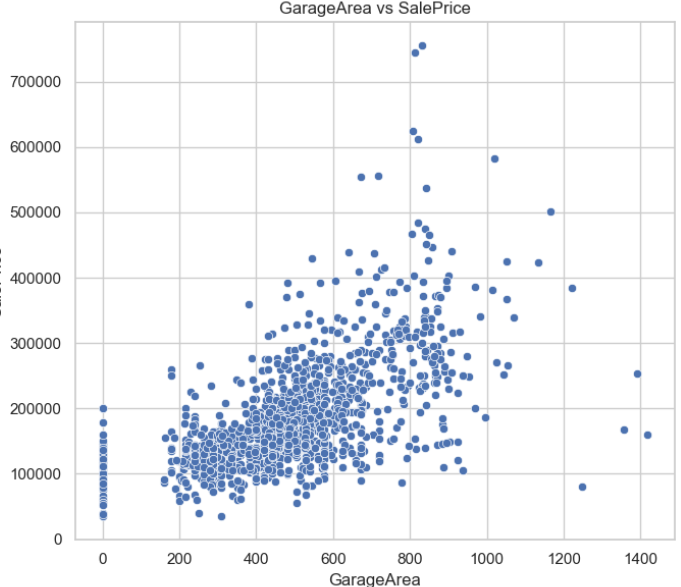
The scatter plot reveals that most houses have a Ground Living Area (**GrLivArea**) between 1000 to 3000 square feet and a **Sale** **Price** between 100,000 to 400,000 USD. There is a strong positive correlation between these variables, indicating that as **GrLivArea** increases, **Sale** **Price** typically rises. A few outliers exist, particularly for homes with **GrLivArea** above 4000 square feet, where prices don't increase proportionately, suggesting potential anomalies or unique factors.

**TotalBsmtSF vs SalePrice**

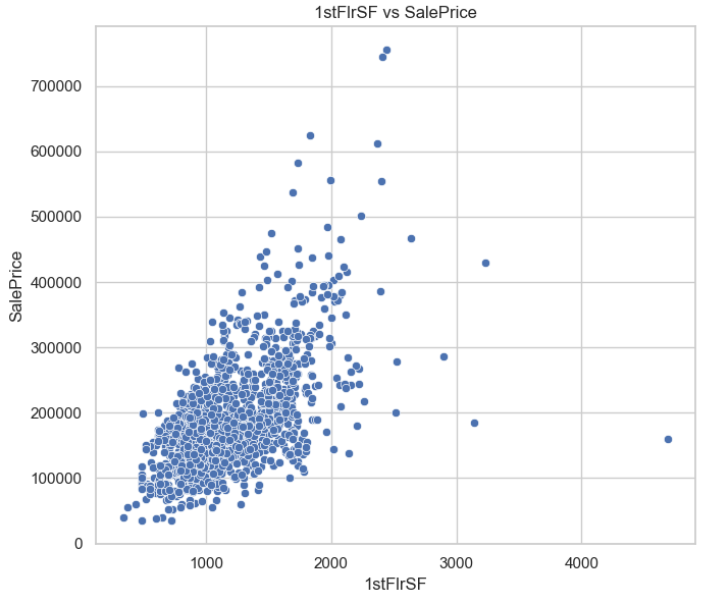
A few outliers with larger **basements** (over 3000 square feet) don't follow this trend, indicating potential anomalies. **TotalBsmtSF** is a significant predictor of house prices.

**YearBuilt vs. SalePrice Scatter Plot**

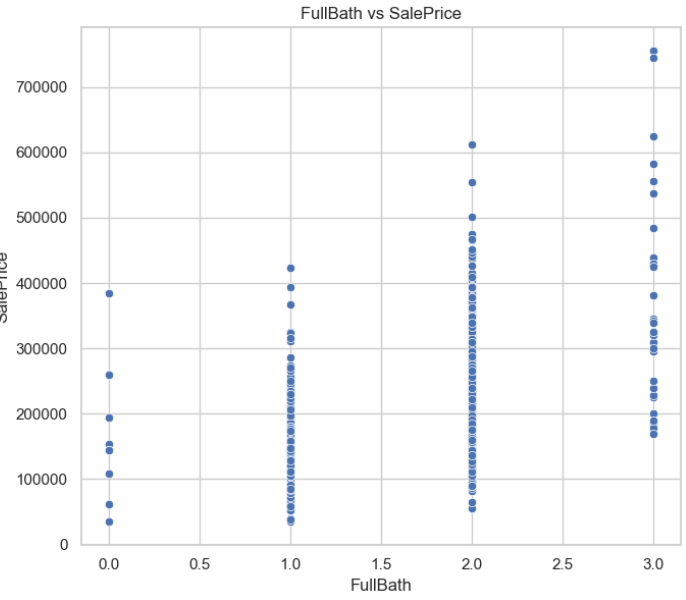
The scatter plot shows that houses **built** in more recent years generally have higher sale prices. Most houses built after 1960 show a noticeable upward trend in **Sale** **Price**, particularly after 2000, where prices increase sharply. Older houses, especially those built before 1940, tend to have lower sale prices, with few exceptions. The **Year** **Built** is a significant factor influencing house prices, with newer homes generally commanding higher prices.

**GarageArea vs. SalePrice Scatter Plot**

As the **garage** **area** increases, the **sale** **price** tends to rise as well. The majority of houses with **garage** **areas** between 200 and 600 **square** **feet** have **sale** **prices** ranging from $100,000 to $300,000. Larger **garage** **areas**, particularly those above 800 **square** **feet**, are associated with **sale** **prices** exceeding $500,000. Although there are some outliers, such as properties with **garage** **areas** over 1,000 square feet that have **sale** **prices** above $600,000, most data points are concentrated in the lower range, suggesting that most houses have **garage** **areas** under 600 square feet.

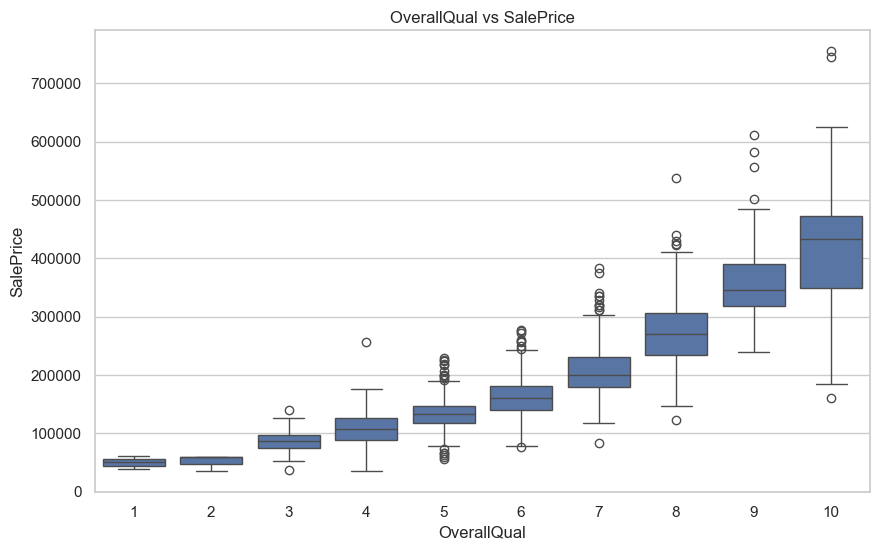
**1stFirSF vs SalePrice**

The scatter plot reveals a strong positive correlation between **1st** **Floor** **Square** **Footage** (1stFirSF) and **Sale** **Price**. Houses with a **1stFirSF** around 1,000 to 2,000 **square** **feet** typically have **sale** **prices** between $100,000 and $400,000. Larger homes with **1stFirSF** above 2,000 **square** **feet** can see **sale** **prices** exceeding $500,000, with some outliers even reaching over $700,000. The data is densely clustered below 2,000 **square** **feet**, indicating that most houses fall within this range.

**FullBath vs SalePrice**

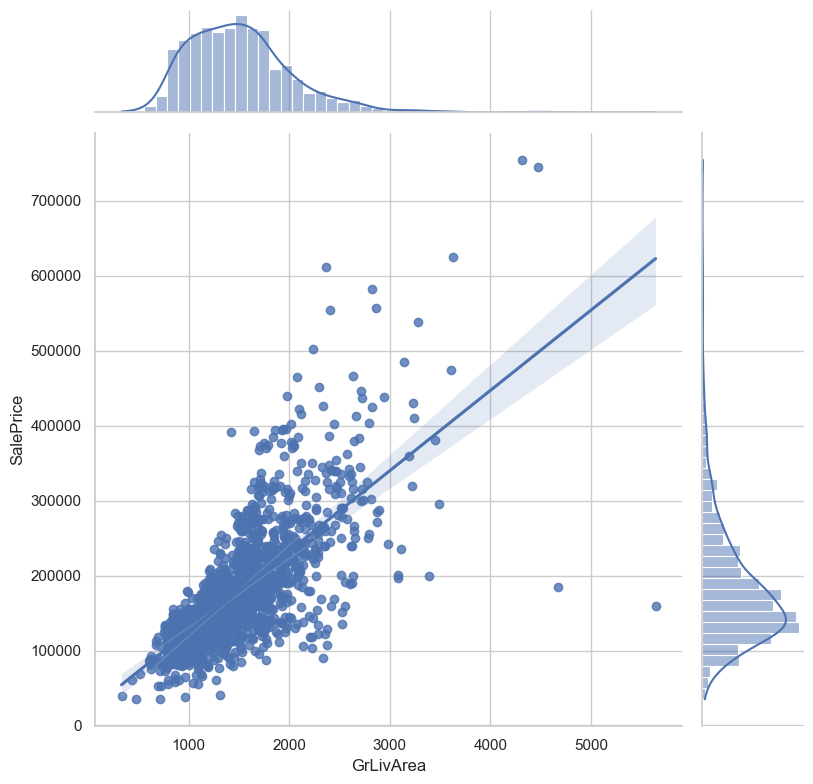
The scatter plot reveals a decent positive correlation between **1st** **Floor** **Square** **Footage** (1stFirSF) and **Sale** **Price**. Houses with a **1stFirSF** around 1,000 to 2,000 **square** **feet** typically have **sale** **prices** between $100,000 and $400,000. Larger homes with **1stFirSF** above 2,000 square feet can see **sale** **prices** exceeding $500,000, with some outliers even reaching over $700,000. The data is densely clustered below 2,000 **square** **feet**.

The following graph represents relationship b/w OverallQual vs SalePrice.

****

**Forecasted Sales Prices:**  
Based on the best-performing model, the forecasted sales prices for sample houses are:

* House 1: $139,935.80
* House 2: $333,368.47
* House 3: $103,031.56
* House 4: $162,947.81
* House 5: $312,906.88

***The graph is a scatter plot with marginal histograms, showing GrLivArea (above ground living area) on the x-axis and SalePrice on the y-axis. The curves on the top and right side of the graph are kernel density estimation (KDE) plots, providing a smoothed representation of the data distribution for GrLivArea and SalePrice*.**

*Some Key markers:*

1. *The scatter plot reveals a strong positive correlation between GrLivArea and SalePrice, indicating that larger living areas generally lead to higher sale prices.*
2. *The KDE plots show that both variables are right-skewed, with a higher concentration of homes with smaller living areas and lower sale prices, while fewer homes are larger and more expensive.*