Data Analysis Final Report

The analysis of the house price dataset reveals several key insights into factors influencing sale prices.

1. Feature Correlations: The correlation heatmap indicates that `SalePrice` is most strongly correlated with `OverallQual` (Overall Quality), `GrLivArea` (Above Grade (Ground) Living Area Square Feet), `GarageCars` (Size of garage in car capacity), and `GarageArea` (Size of garage in square feet). Features such as `TotalBsmtSF` (Total square feet of basement), `1stFlrSF` (First Floor square feet), and `FullBath` (Full bathrooms above grade) also show a significant positive correlation with `SalePrice`. Conversely, features like `LowQualFinSF` (Low quality unfinished square feet) and `BsmtFinSF2` (Type 2 Finished Square Feet) exhibit a weak negative correlation, suggesting they have minimal positive impact on sale price.

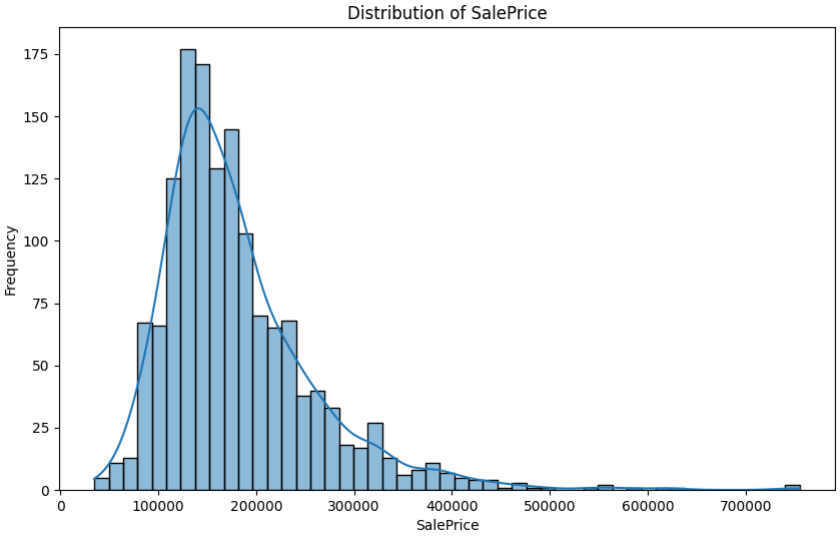
2. Distribution of Key Features: - **SalePrice:** The distribution of `SalePrice` is right-skewed, which is common for financial data. This indicates that while most houses fall within a certain price range, there is a tail of higher-priced properties. - **GrLivArea:** The distribution of `GrLivArea` shows a bimodal pattern, suggesting that there might be different typical sizes of living areas, possibly related to the house style (e.g., 1-story vs. 2-story). - **OverallQual:** The distribution of `OverallQual` is relatively normal, with a peak around quality rating 6 and 7, and fewer houses at the extreme low and high ends of the quality scale. This suggests that houses with higher overall quality tend to command higher prices.

3. Data Quality Observations: The initial data cleaning steps handled missing values by imputing them with the mean for numerical columns and the mode for categorical columns. Duplicate rows were removed. Subsequent encoding was applied to categorical features, using one-hot encoding for features with few unique values and label encoding for others. This ensures that the data is in a suitable format for modeling.

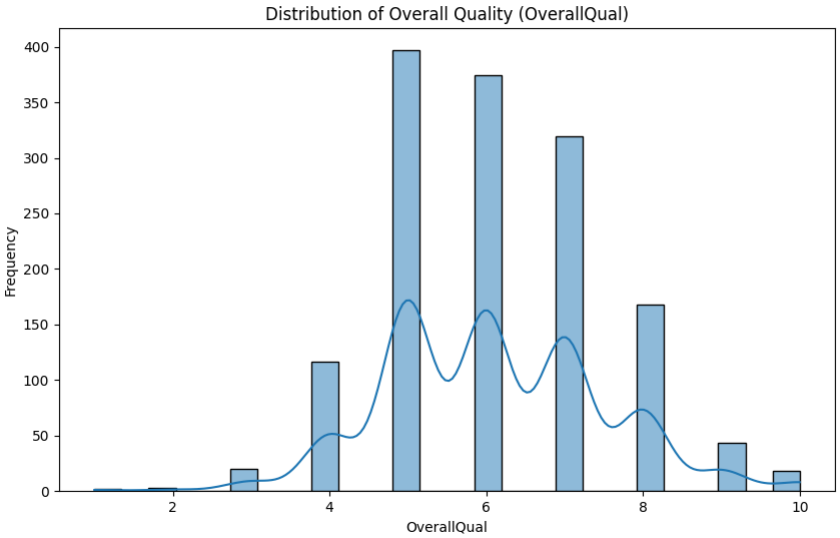
Conclusion: The most significant drivers of house prices in this dataset appear to be the overall quality of the house, its living area, and the presence and size of a garage and bathrooms. Understanding these relationships is crucial for accurate price prediction and for identifying properties with high market value.

Visual Evidence

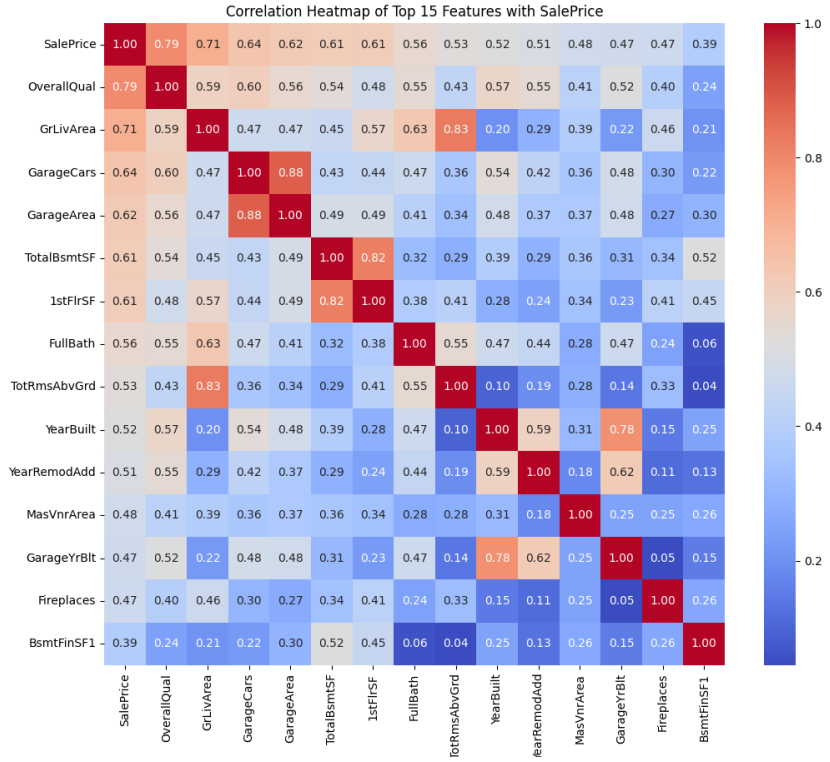
Visualization: plot_2_target_distribution.png



Visualization: plot_4_overallqual_distribution.png



Visualization: plot_1_correlation_heatmap_top15.png



Visualization: plot_3_grlivarea_distribution.png