PREDICTIVE MODELING FOR HOUSE PRICES USING ADVANCED REGRESSION TECHNIQUES

1. PROJECT OVERVIEW

- **PROBLEM STATEMENT:** Given a dataset containing various residential property features, the challenge is to develop a predictive model that accurately estimates house sale prices, aiding buyers, sellers, and real estate professionals in making informed decisions.

- **OBJECTIVE:** To build and evaluate a predictive model for house prices using advanced regression techniques, ensuring high accuracy and reliability in estimating property values based on key features.
- **Important Feature Highlight:** This project focuses more on 6 unique features which MSSUbclass, MSZoning, Utilities Type, Lot Frontage, YearBuilt and Neighborhood in relation to the target feature which is the Sales Price.

- **Stakeholders:** People that are likely to benefit from this project are buyers, sellers and real estate professionals
- **Model Framework**: Explain the model you are using and how it fits into your analysis.
- **Assumptions**:
    - **Influence of Year Built**: The year the house was built is one of the key features that greatly influences the price of the house

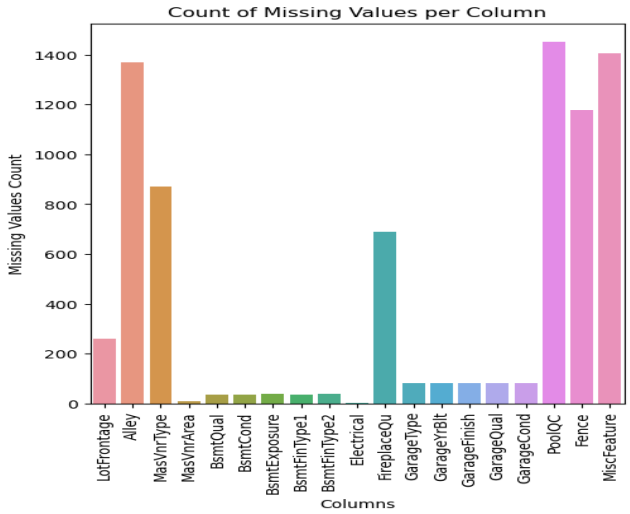
2. DATA COLLECTION

- **DATA SOURCE:** The dataset was gotten from a Kaggle competition.

- **DATA DESCRIPTION:** The dataset folder contains two dataset which contains the train and test dataset. The train dataset has 81 columns and 1,460 rows including the target variable which is the sales price. The test dataset is an unlabeled data without the target variable, it has 80 columns and 1,459 rows. The aim is to predict the unlabeled data after building the model.

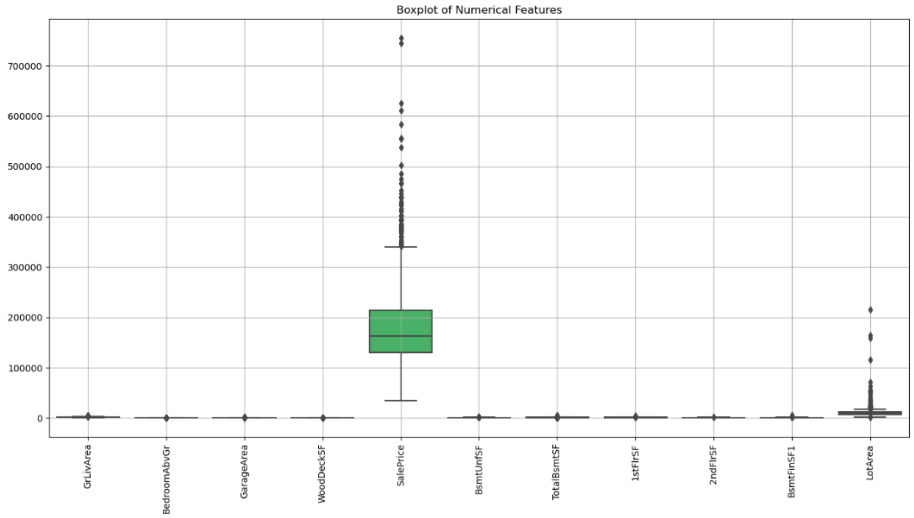
3. DATA CLEANING AND EXPLORATORY DATA ANALYSIS

- **MISSING VALUES:** There are 19 columns with missing values in both the train and test dataset, with PoolQC being the column with the highest number of missing values.



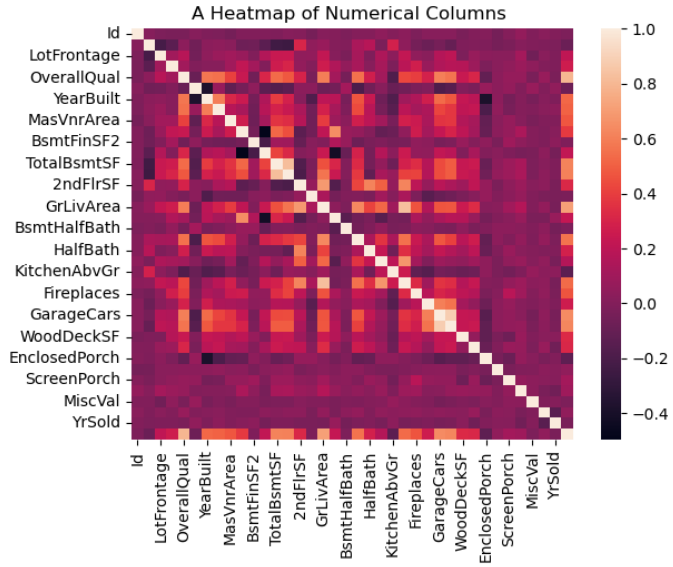
The missing values might be due to unavailability or lack of presence of such column in that particular house. The categorical columns with missing values were filled with NIL while numerical columns were filled with 0 in both datasets. No duplicate was found in both the train and test dataset. The datatypes were all formatted correctly.

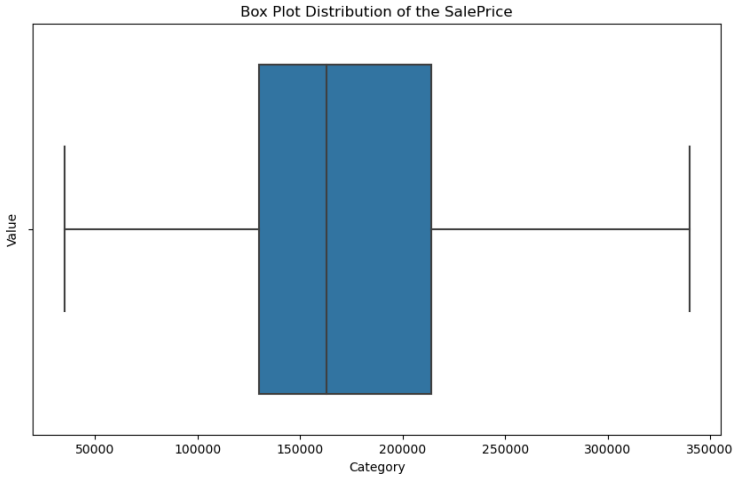
- **OUTLIERS:** The outliers in the dataset we plotted using the boxplot, which showed the whiskers representing the outliers in our dataset.



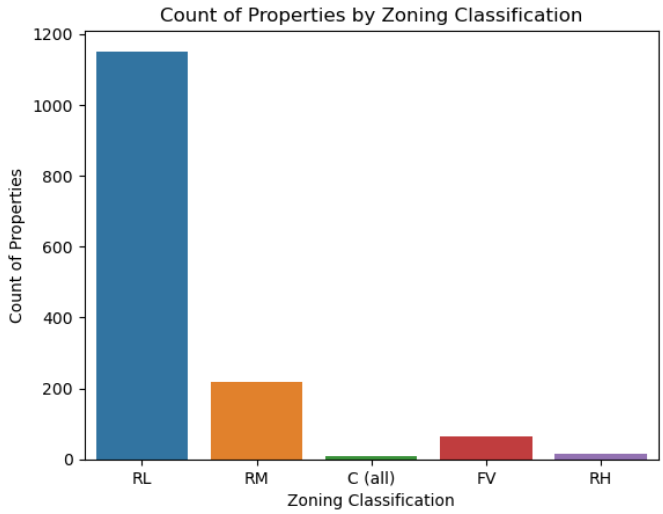
The outliers were handled using the Interquartile Range (IQR) and filled using the clip method which set the outliers to a fixed boundary level either the minimum or maximum accepted values in the data distribution. For example, if there is an outlier that is below the minimum value, it is filled with the minimum values, same goes for maximum values. The idea behind handling outliers through methods like clipping is to mitigate the impact that extreme values can have on statistical analysis and machine learning models. Outliers can skew the results and lead to misleading conclusions.

- **EXPLORATORY DATA ANALYSIS:** The variables exhibit a mix of correlations. Some are strongly correlated, indicating they might provide similar information, while others are weakly correlated or uncorrelated, suggesting they contribute distinct information to the dataset. This diversity in correlations can be valuable for building a robust predictive model, as it combines both redundant and unique features.





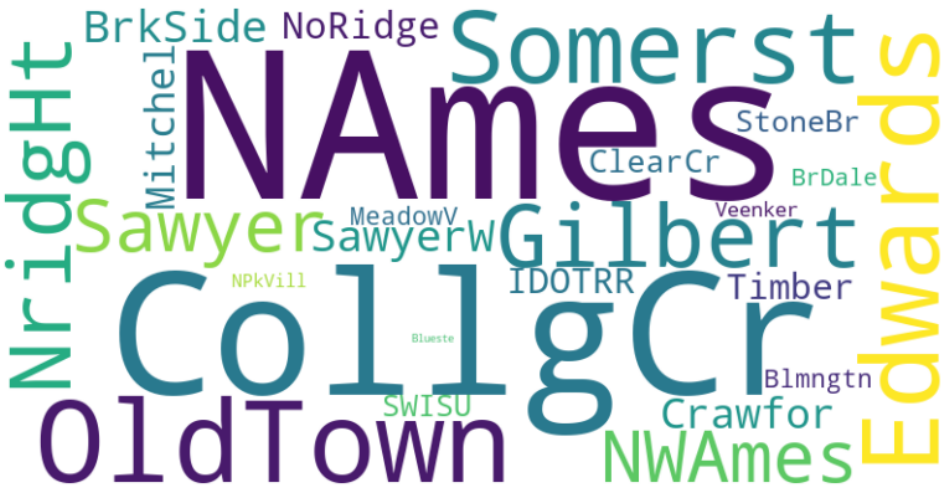
The boxplot shows the distribution of the sales price with the minimum price estimated to be $34,000, the median been $163,000 and the maximum estimated to be around $340,000.



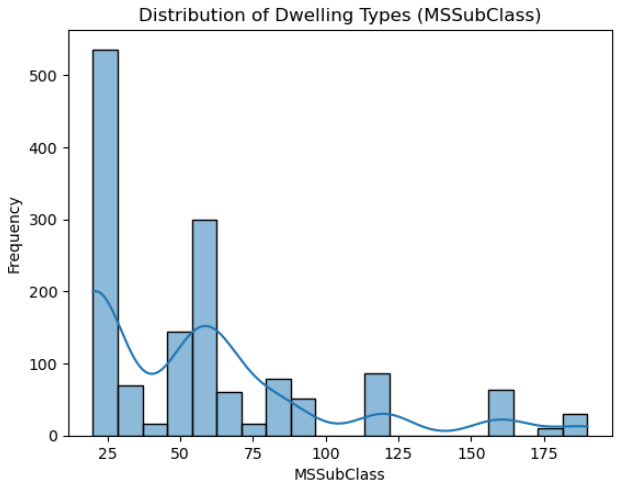
MSZoning also known as Municipal Zoning Identifies the general zoning classification of each property. The meaning of each classification can be found below:

a. C: Commercial,
b. FV: Floating Village Residential,
c. RH: Residential High Density,
d. RL: Residential Low Density,
e. RM: Residential Medium Density

The Residential Low Density (RL) has high number of houses in this zone than any other type of zoning. Properties zoned for residential use typically have different market values compared to those zoned for commercial use. Residential zones often cater to family living, while commercial zones can include businesses, factories which might increase property value due to higher potential rental income. Zoning laws may dictate the density of housing that can be built. For example, areas zoned for single-family homes may have higher property values compared to those zoned for high-density apartments or multi-family units. Zoning can influence the availability of amenities and infrastructural development.



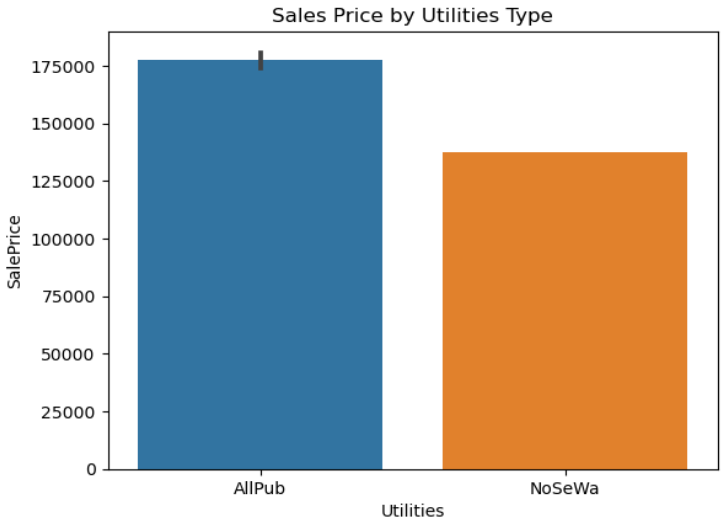
The word cloud shows the frequency of each Neighborhood based on how bold they are. **The neighborhood called NAmes, Edwards, OldTown, CollgCr and Gilbert appears to be the most frequent base on their font sizes.**



The chart you see is a histogram showing the distribution of different dwelling types in the dataset, classified by a code called "MSSubClass." Each bar represents the number of houses in each category, and the height of the bar shows how many houses fall into that category.
Here's a breakdown of what this chart tells us:
MSSubClass Codes: These codes represent different types of houses. For example:

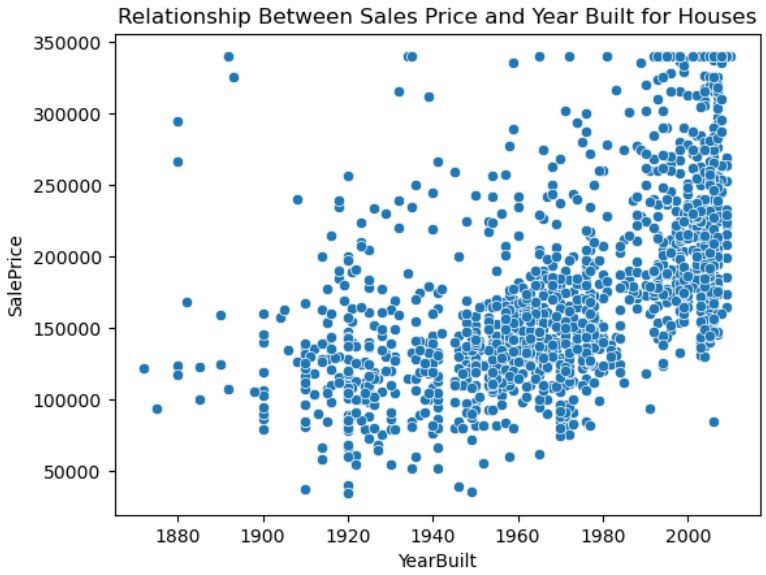
- 20: 1-story houses built before 1946.
- 60: 2-story houses built after 1945.
- 120: 1-story PUD (Planned Unit Development).

Frequency: The y-axis shows the number of houses in each category. The higher the bar, the more houses there are of that type. **The tallest bar is at MSSubClass code 20, indicating that the most common type of house in this dataset is the 1-story houses built before 1946.** There is also a significant number of houses in the MSSubClass code 60, which are 2-story houses built after 1945. In summary, this chart shows that in this dataset, older 1-story houses are the most common, followed by 2-story houses built after 1945. Other types of houses are less common.

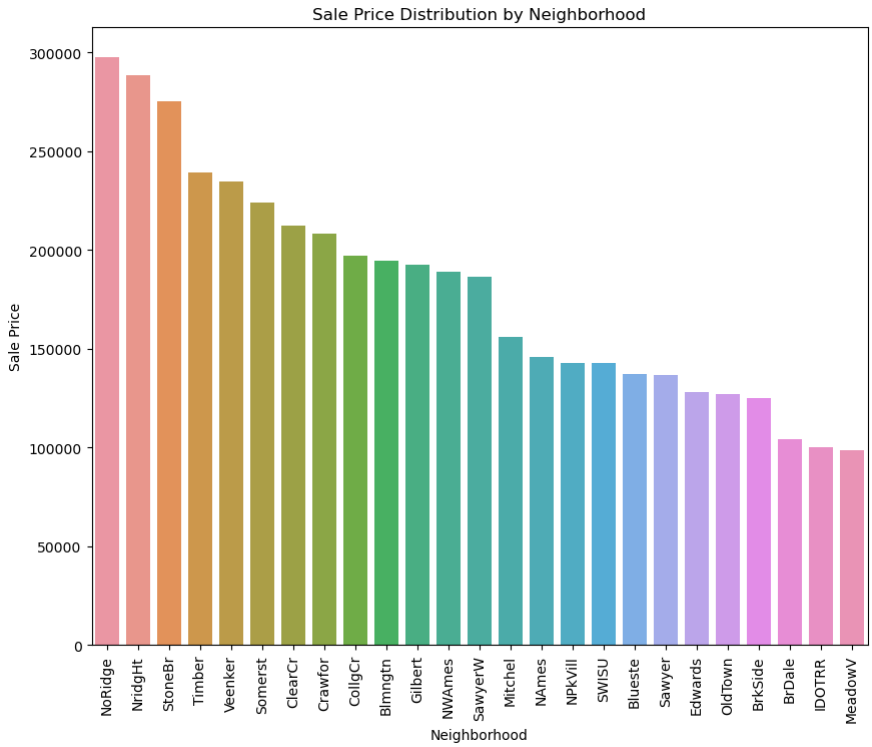


AllPub: This stands for "All Public utilities." It means that the property has access to all essential public utilities, which generally include: Public water supply, Public sewer system, Electricity,Gas, Telephone, Internet
NoSeWa: This is shorthand for "No Sewer or Water." It indicates that the property does not have access to public sewer and water services. Instead, the property might need:

A septic system for sewage disposal
A private well for water supply
Properties with "NoSeWa" utilities are typically found in more rural or undeveloped areas where public utility infrastructure is not available. Owners of such properties must arrange for private solutions for water and sewage. **This shows the prices of properties in relation to the Utilities Type. Properties with AllPub has higher price compared to properties with NoSewa.**



The variability in sale prices increases for houses built in recent years. This suggests that while newer houses can fetch higher prices, there is also a wide range of prices within each time period.
The variability is lower for older houses (before 1940), indicating a more uniform pricing in that period.



The 3 Neighborhood with the highest house price are NoRidge, Nridght and StoneBr. The type of neighborhood can also affect the prices of house. The more developed a neighborhood is, the higher the prices of houses.

4. FEATURE ENGINEERING AND FEATURE SELECTION

- **Feature Engineering:** 5 Nine features were created from existing columns, which are house age, remodel age, total baths, total square footage and lot frontage ratio.
- **Preprocessing:** Features were encoded using one-hot encoding, label encoding, and standard scaling. Ordered features were encoded using the label encoder, while non-ordered features were encoded using one-hot

encoding to avoid introducing false orderliness. Numerical features were scaled using StandardScaler to ensure that all features contribute equally to the model. A pipeline and column transformer were used to streamline these steps.

- **Feature Selection:** After preprocessing, the number of features increased to 211. To reduce dimensionality, Principal Component Analysis (PCA) was used, setting the number of components to 5.

5. MODEL TRAINING AND EVALUATION

The train data was split into train and test dataset. 5 machine learning algorithm was used to train the model which are Random Forest, Gradient Boosting, SVR, Decision Tree and Linear Regression. **JUSTIFICATION FOR EACH MODEL is given below:**

- **Linear Regression** provides a baseline and is highly interpretable. It assumes a linear relationship between the independent and dependent variables and provides a baseline performance. It's easy to implement and understand, making it a good starting point for any regression problem.
- **Decision Trees** offer flexibility in capturing non-linear relationships. Decision Trees can handle both linear and non-linear relationships by splitting the data based on feature values. They are useful for capturing non-linear patterns and interactions between features. While they can overfit the data, they provide a basis for more complex ensemble methods.
- **Random Forest** improve predictive performance by addressing the limitations of individual trees and leveraging ensemble techniques. Random Forest is an ensemble method that combines multiple decision trees to improve performance and robustness. It mitigates the risk of overfitting present in individual decision trees and can capture complex non-linear relationships.
- **Gradient Boosting** is another ensemble method that builds models sequentially, where each new model corrects errors made by the previous ones. It can handle complex relationships and interactions between features and often achieves higher accuracy than simpler models. It's effective in capturing both linear and non-linear patterns in the data.
- **SVR** handles non-linearity effectively and works well with high-dimensional data. SVR is useful for regression tasks where you want to find a function that fits the data within a certain margin of tolerance. It can handle non-linear relationships by using kernel functions to transform the feature space. SVR is particularly beneficial for regression tasks with complex and high-dimensional data.

Hyperparameter Tuning:

The metrics used for evaluating the models are Root mean squared, Mean Absolute error and R2 score. Before tuning, Random Forest happens be the best performing model with the lowest **RMSE**: 23,682.70, **MAE**: 16,976.52 and highest **R2 Score: 0.885.**

Hyperparameter was carried on each model using the RandomizedSearchCV. The reason for my choice of RandomizedSearchCV is that, **it requires less processing time.** The tuning was done on 3 most important parameters of each model. It selected the best parameters and its values from the list of values given.

The best performing model after hyperparameter tuning is Random Forest with the lowest **RMSE**: 23,426.85, **MAE**: 16,883.95, and highest **R2 score:** 0.887 after tuning.

Choosing the Best Performing Model

Based on the performance metrics (RMSE, MAE, and R2), the Random Forest model is the best among the options. The reasons for the assumption are:

1. **RMSE (Root Mean Squared Error):** Lower RMSE indicates better predictive accuracy. The Random Forest model has the lowest RMSE (23,426.85), suggesting it makes smaller errors on average.

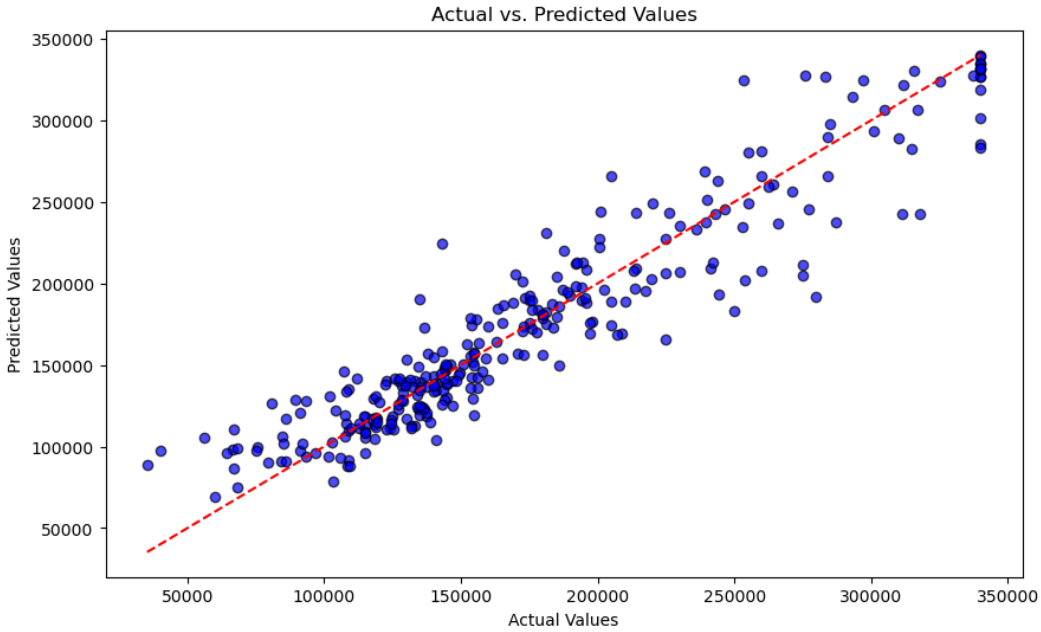
2. **MAE (Mean Absolute Error)**: Like RMSE, a lower MAE is better. The Random Forest model also has the lowest MAE (16,883.95), indicating it predicts closer to the actual values.

3. R**2 (Coefficient of Determination):** R2 measures the proportion of variance explained by the model. A higher R2 value is better. The Random Forest model has the highest R2 (0.887), indicating it explains 89% variance in the data.

Given these metrics, the Random Forest model outperforms the other models in all key areas, making it the best choice.

INTERPRETATION OF THE BEST PERFORMING MODEL FEATURES

1. **HouseAge**
    - Older houses might be less expensive due to wear and tear, outdated designs, or the need for more repairs and maintenance.
    - Conversely, historic or well-maintained older houses in desirable locations might be more expensive due to their charm and character.
    - Newer houses tend to be more expensive as they are built with modern materials and designs, and are likely to require fewer immediate repairs.
2. **YearRemodAdd**
    - The year when the house was last remodeled or had additions.
    - Houses with recent renovations might be more expensive because they incorporate modern amenities and designs, making them more appealing to buyers.
    - Extensive remodeling can significantly increase the value of a house by updating essential systems like plumbing, electrical, and HVAC.
    - The degree and quality of remodeling also play a crucial role. High-quality renovations can substantially boost a house's price.
3. **BsmtFinType1**
    - Type of basement finish (e.g., unfinished, partially finished, fully finished).
    - Fully finished basements add usable living space, which can increase the overall value of the house.
    - Basements that are finished to a high standard (e.g., with living areas, bathrooms, and high-end finishes) can significantly increase the house's price.
    - Partially finished or unfinished basements offer less additional value but still provide potential for future improvements.
4. **TotalSF (Total Square Footage)**
    - The total square footage of the house, including all living spaces.
    - Generally, larger homes are more expensive because they offer more living space.
    - The total square footage is a direct measure of the house's size, and larger homes can accommodate more amenities and features, which adds to their value.
    - However, the value also depends on how the space is utilized. Efficient use of space can sometimes be more valuable than sheer size.
5. **2ndFlrSF (Second Floor Square Footage)**
    - The square footage of the second floor of the house.
    - Houses with a significant amount of second-floor space can be more appealing to families who need multiple bedrooms or additional living areas.
    - The presence of a second floor can also indicate a more complex and potentially more valuable architectural design.
    - The functionality and design of the second floor (e.g., master suites, additional bathrooms) can further influence the overall house price.



Actual vs. Predicted Plot: Shows how well the predicted values match the actual values. The red dashed line represents a perfect prediction.

The Sales Price was predicted for the unlabeled data and added to the dataset after which it was converted to a csv file.

6. **TOOLS USED:**
    - **Data Cleaning and EDA:** Pandas, NumPy, Matplotlib, Seaborn
    - **Feature Engineering and Preprocessing:** Scikit-learn (LabelEncoder, OneHotEncoder, StandardScaler, PCA, Pipeline, ColumnTransformer)
    - **Model Training and Evaluation:** Scikit-learn (Linear Regression, Decision Tree, Random Forest, Gradient Boosting, SVR, RandomizedSearchCV)

7. **CONCLUSION:**

KEY FINDSINGS

- The neighborhood, utilities type, year the house was built influences the price of the house.
- Random Forest outperformed other models based on lower RMSE, MAE, and higher R^2 score.
- The model provides accurate price predictions, useful for stakeholders in the real estate market