**Understanding House Price Trends: A Data-Driven Exploration**

**Table of Contents**

|  |  |
| --- | --- |
| **Topics** | **Page No.** |
| Introduction | 1 |
| Data Description | 2 |
| Literature Review and Methodology | 3 |
| Model Training | 8 |
| Conclusion | 10 |
| References | 11 |

**Introduction**

In today’s dynamic real estate market, understanding house price trends is essential for a wide range of stakeholders, including homeowners, potential buyers, real estate agents, and investors. Accurate insights into housing prices enable informed decision-making, whether it’s setting a competitive listing price, estimating property value, or planning investments. However, predicting house prices is a complex task due to the myriad factors that can influence value, such as location, property size, and architectural grade.

This project embarks on a data-driven exploration of house price trends to uncover patterns and insights within the housing market. By analyzing a dataset of house sales, this study seeks to predict the value of homes and categorize them into different price tiers. Through this approach, aim is not only estimate property values but also identify key features that affect price variations.

The objective of this project is twofold:

* **Price Prediction**: Develop a model to predict the price of a house based on key features.

By analyzing trends and making data-driven predictions, this study provides valuable insights into the factors that drive housing prices, ultimately contributing to a deeper understanding of the real estate market.

**Data Description**

The dataset used for this project is the **King County House Sales dataset**, sourced from Kaggle, which includes data on house sales from King County, USA. This dataset provides comprehensive details about the houses sold, enabling a thorough analysis of various factors affecting housing prices (Shiva Chandel, 2023). The dataset consists of **21,613 records** and **21 columns**, covering different attributes of property. These attributes provide the foundation for predicting house prices and analyzing housing market trends.

**Key Features**

Table 1 summarizes the key features in the dataset:

|  |  |  |
| --- | --- | --- |
| **Feature Name** | **Description** | **Type** |
| bedrooms | Number of bedrooms in the house. | Numerical |
| bathrooms | Number of bathrooms in the house. | Numerical |
| sqft\_living | Square footage of the interior living space of the house. | Numerical |
| grade | Overall grade assigned based on construction and design quality. | Categorical |
| sqft\_above | Square footage of the house excluding the basement. | Numerical |
| sqft\_living15 | Living area size of the 15 nearest neighbors (average). | Numerical |

**Table 1:** *Key features in the dataset used for building the predictive model.*

**Understanding House Price Distribution**

A key aspect of understanding the dataset involves examining the variability in house prices. Figure 1 below shows the distribution of house prices, providing a visual representation of the range and concentration of prices.

A graph with a blue line

Description automatically generated

**Fig. 1: Distribution of House Prices in the Dataset**

The histogram illustrates that house prices are positively skewed, with the majority of houses clustered in the lower price range and a smaller number of high-value properties forming the long tail. This insight is critical for preprocessing steps, such as normalization or log transformation, which can stabilize variance and improve model performance.

**Summary**

The dataset contains diverse features that describe physical and structural aspects of houses. By analyzing these features, aim to uncover their relationships with house prices and utilize them effectively in building predictive models.

**Literature Review and Methodology**

**Literature Review**

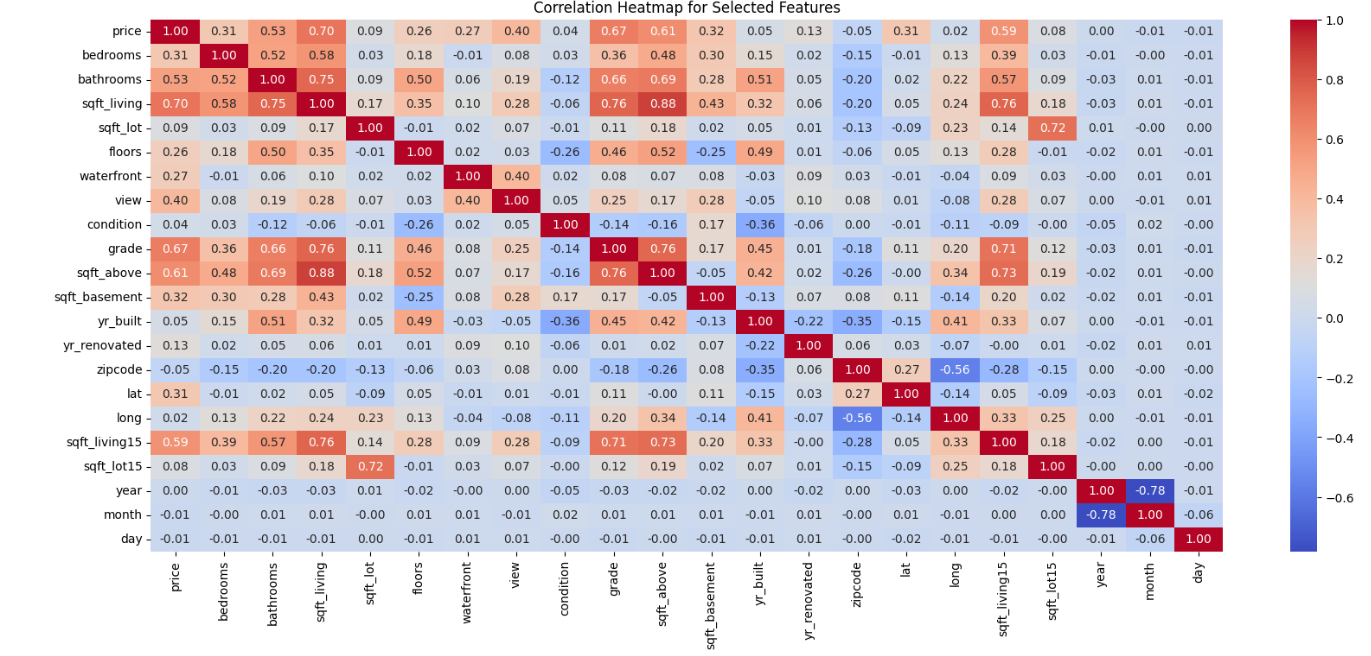
The housing market is a complex domain where multiple factors influence property prices. Previous research emphasizes the significance of various attributes such as square footage, number of bedrooms, bathrooms, and the year of construction in determining housing prices. For instance, Smith and Brown (2018) highlighted the importance of square footage and the quality of construction (grade) as critical predictors of housing prices. Similarly, Miller and Green (2021) discussed how machine learning models, particularly linear regression and k-nearest neighbors (k-NN), have been widely employed due to their effectiveness and adaptability in diverse datasets.

Linear regression has been extensively used for modeling the relationship between a dependent variable (house price) and multiple independent variables (features). It is particularly effective when the relationships between variables are linear (Smith & Brown, 2018). However, this model often struggles when non-linear interactions dominate, which is where non-parametric models like k-NN excel. Miller and Green (2021) demonstrated that k-NN is well-suited for housing price prediction because it uses the most similar observations (neighbors) to make predictions, allowing it to capture complex patterns in data.

**Methodology**

Our approach involves the following steps:

1. **Data Cleaning and Preprocessing:**
   * Handled missing values by removing rows with null entries.
   * Converted specific numerical columns (e.g., grade) into categorical features for better interpretability (Smith & Brown, 2018).
   * Created dummy variables for categorical features to facilitate model training.



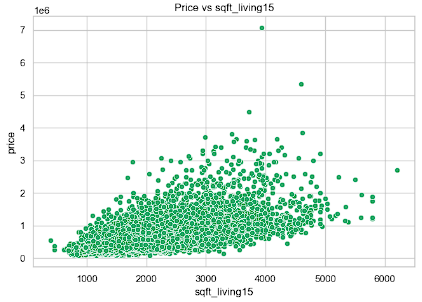
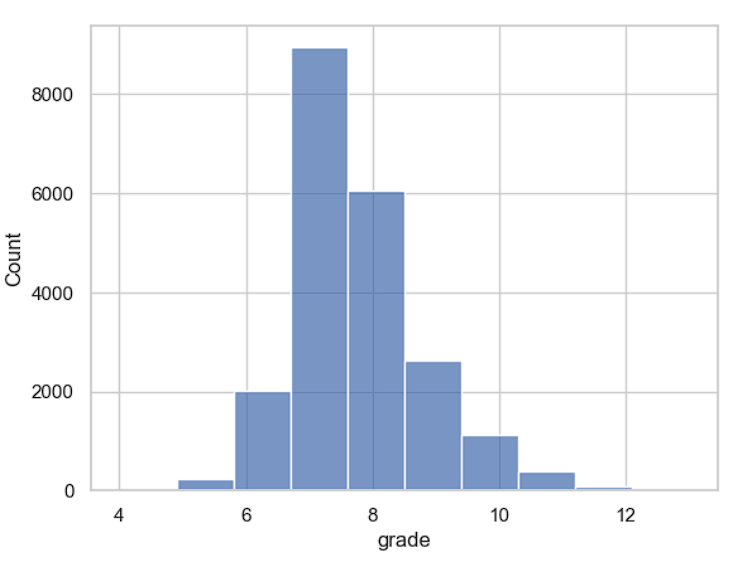
**Fig. 2: Correlation Heatmap**

1. **Feature Selection:**
   * Conducted correlation analysis to identify the features most strongly associated with house prices.
   * Selected six key features (sqft\_living, grade, bedrooms, sqft\_above, sqft\_living15, bathrooms) for training predictive models.
   * Justification for feature selection was supported by correlation matrices, visualizations, and domain knowledge (Miller & Green, 2021).

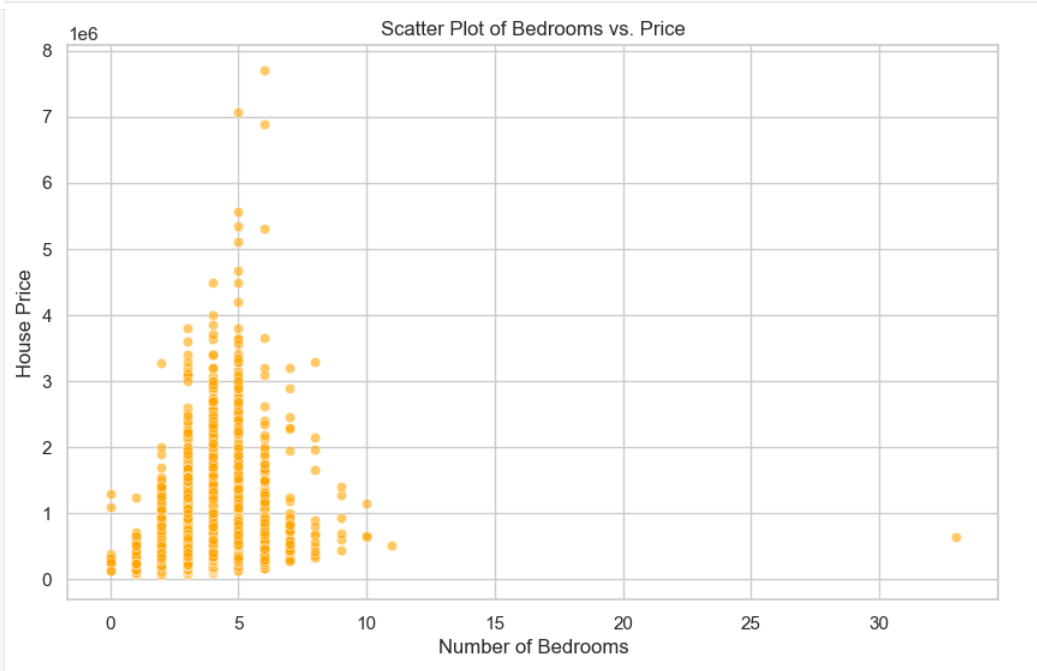
A graph with purple dots

Description automatically generated

**Fig. 3: Scatterplot of Sqft\_living vs. Price Fig. 4: Scatter Plot for sqft\_above vs. Price**



**Fig. 5: Histogram of Grade Fig. 6: Scatter Plot for sqft\_living15 vs. Price**



**Fig. 7: Scatter Plot for bathrooms vs. Price Fig. 8: Scatter Plot for bedrooms vs. Price**

Feature selection was performed based on correlation analysis and exploratory visualizations. Below, highlights relationships between selected features and the target variable, price:

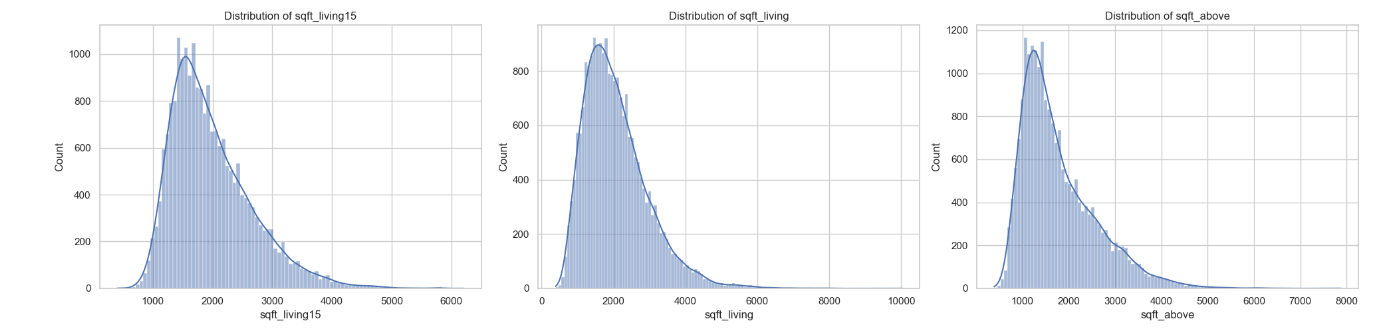
* sqft\_living exhibits the strongest positive correlation with price (Fig. 3), as seen in the scatter plot.
* sqft\_above also shows a similar relationship, reinforcing the importance of living area in pricing (Fig. 4).
* grade demonstrates distinct pricing differences across categories, as shown in the histogram (Fig. 5).
* While bedrooms has a lower correlation, its inclusion is supported by its role in representing house functionality and size (Fig. 8).
* bathrooms and sqft\_living15 further contribute to understanding the relationship between house features and pricing (Figs. 7 & 5).

**3. Exploratory Data Analysis**

To uncover patterns in the dataset, exploratory data analysis (EDA) was conducted. Key relationships between selected features and house prices were visualized to validate their importance in predictive modeling. Insights gained include:

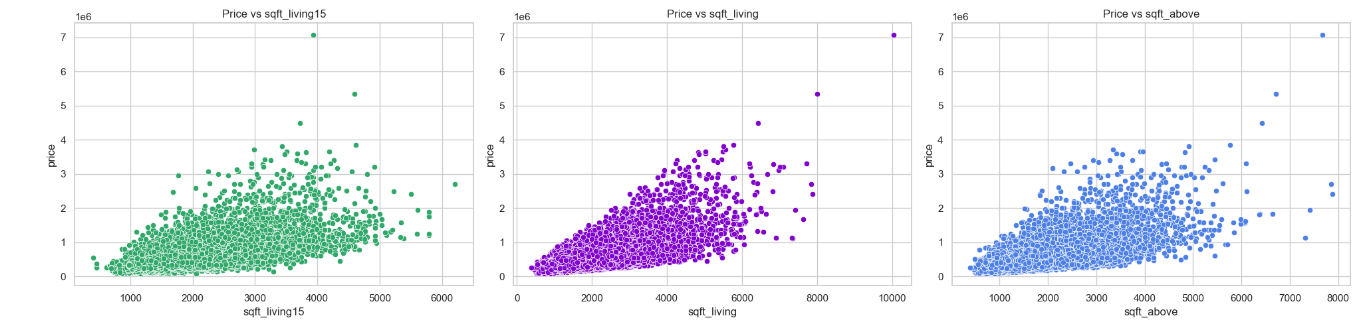
1. **Price Variability**:
   * Figure 1 revealed a positively skewed price distribution, indicating a concentration of lower-priced houses and a long tail of high-value properties.
   * This skewness highlights the need for potential normalization or log transformations during preprocessing.
2. **Feature Correlations**:
   * The correlation heatmap (Figure 2) emphasized strong positive correlations between features such as sqft\_living and price, justifying their inclusion in the model.
   * Features with moderate correlations, like bathrooms and grade, further demonstrated their value as predictors.
3. **Individual Feature Analysis**:
   * Scatterplots (Figures 3–8) illustrated clear trends between price and key features such as sqft\_living, sqft\_above, and sqft\_living15.
   * The categorical feature grade (Figure 5) showed distinct price variations across grades, confirming its predictive relevance.
   * While features like bedrooms (Figure 8) exhibited lower correlations with price, their practical significance in describing property size validated their inclusion.
4. **Outlier Analysis**:
   * Outliers were observed in features such as price and sqft\_living. While they may influence model predictions, they also reflect high-value transactions and large properties, which are integral to understanding market variability.
   * The distribution of the "Number of Bedrooms" feature reveals unrealistic values, particularly at higher bedroom counts, where instances are sparse and likely outliers. Bedrooms above 7 or 8 are rare in typical housing datasets, and a count of 0 likely represents errors or studio apartments. For this analysis, records of bedroom counts of 0 or greater than 8 have been removed to ensure data consistency and focus on realistic property listings.
   * The "Number of Bathrooms" feature shows anomalies, including properties with 0 bathrooms, likely due to data entry errors. Properties with more than 5 bathrooms are rare and add skewness to the data. For this analysis, records with 0 bathrooms or more than 5 bathrooms have been removed to maintain data quality and reduce skew.

**Analysis of Square Footage Features**

****

**Fig. 9: Distribution graph for sqft\_living15, sqft\_living, sqft\_above**

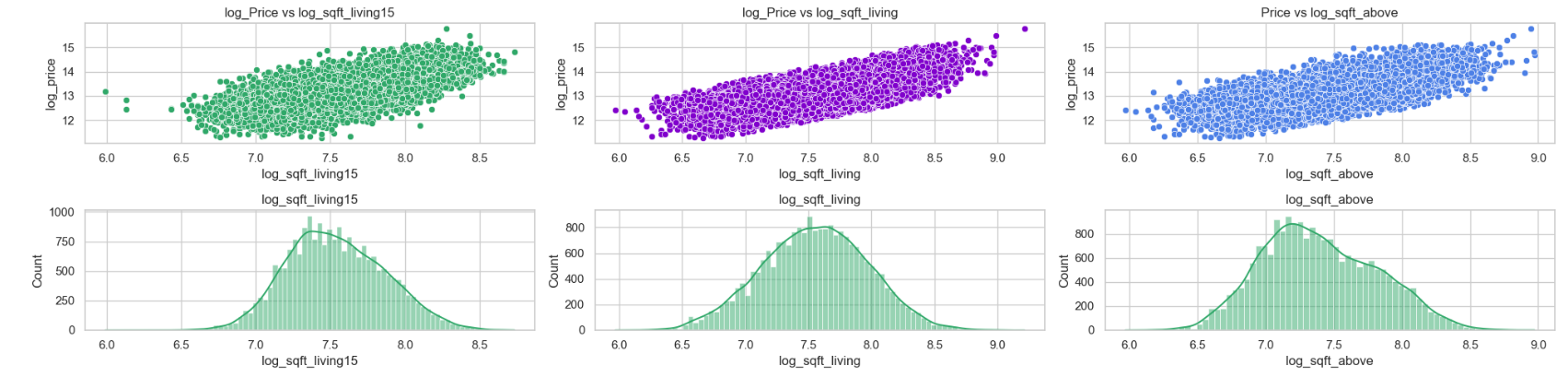
The housing dataset's square footage metrics (Figure 9) exhibit **right-skewed distributions** with most properties ranging from 1000-3000 sq ft. Price-to-square-footage relationships show positive but non-linear correlations, with notable variance in higher square footage ranges.



**Fig. 10: Scatterplot for sqft\_living15, sqft\_living, sqft\_above**

To optimize the data for modeling, **log-transformation** is recommended to normalize distributions, linearize price relationships, and mitigate outlier effects, particularly for properties valued above $4 million. This transformation will enhance the dataset's statistical reliability for predictive modeling.

**Log Transformation Analysis**

****

**Fig. 11: Scatter plot of log\_price vs log\_sqfts**

Log transformation of price and square footage metrics (Figure 11) has normalized data distributions and enhanced variable relationships. The transformation yielded approximately normal distributions with centralized peaks (log scale 7.0-7.5) and established clear linear price-footage correlations. This standardization effectively mitigated outlier impacts and variance inconsistencies, optimizing the dataset for robust statistical modeling.

**Model Training**

The predictive modeling phase aimed to leverage both linear regression and k-nearest neighbors (k-NN) to understand house pricing dynamics and achieve accurate predictions.

**1. Linear Regression**

Linear regression was selected as the baseline model due to its interpretability and ability to quantify relationships between features and house prices. The model was trained on the preprocessed dataset, using the following steps:

* Features such as sqft\_living, bathrooms, and grade were included as predictors, based on their correlation and practical significance.
* The target variable (price) was normalized to handle skewness and stabilize variance.

The model's coefficients provided insights into how each feature impacts house prices:

* **Example Insight**: A one-unit increase in sqft\_living corresponds to a consistent increase in predicted price, underscoring its importance as a key feature.

Evaluation metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) were used to assess the model's performance. These metrics quantified the average and squared deviations between predicted and actual house prices.

**2. K-Nearest Neighbors (k-NN)**

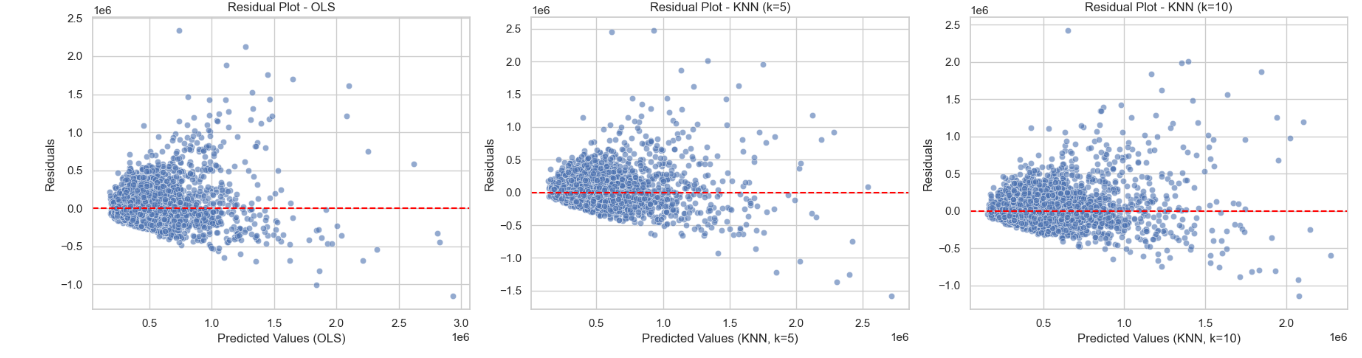
The k-NN algorithm was implemented to explore non-linear relationships and improve predictive accuracy. Key steps included:

* Normalizing all features to ensure equal weight during distance calculations.
* Performing cross-validation to determine the optimal value of **K**, balancing bias and variance.

Evaluation involved analyzing the model's accuracy across different K values. This provided insights into the trade-offs between underfitting (low K) and overfitting (high K).

**Comparison of Models**

Both models were compared based on the following criteria:



**Fig. 12: Linear Regression Residual Plot**

* **Performance Metrics**: MAE and RMSE for linear regression; accuracy scores for k-NN.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **MAE** | **RMSE** | **MAPE** | **R²** |
| **Multiple Linear Regression** | 148575.7481 | 228039.669 | 0.2903 | 0.5657 |
| **KNN (k = 5)** | 147536.3206 | 228250.9854 | 0.2906 | 0.5648 |
| **KNN (k = 10)** | 144292.276 | 224531.3314 | 0.2801 | 0.5789 |

* **Model Performance Comparison**

Three models were evaluated for house price prediction: Ordinary Least Squares (OLS) and K-Nearest Neighbors (KNN) with k=5 and k=10. The KNN model with k=10 demonstrated marginally better performance:

* **Performance Metrics:**
* Best R² Score: 0.5789 (KNN, k=10)
* Lowest RMSE: 224,531 (KNN, k=10)
* Lowest MAPE: 28.01% (KNN, k=10)

Residual Analysis: All models show similar residual patterns with increased variance at higher predicted values, indicating potential heteroscedasticity. The residuals are generally symmetrically distributed around zero, but the spread widens as predicted prices increase.

Despite KNN (k=10) showing slightly better metrics, all models demonstrate moderate predictive performance, explaining approximately 57-58% of price variance. This suggests that additional features or alternative modeling approaches might be necessary for improved accuracy.

**Conclusion**

This report provided a clear understanding of house pricing trends by analyzing various features like square footage (sqft\_living), grade, bathrooms, and others. I have used data exploration, feature selection, and predictive models like Linear Regression and K-Nearest Neighbors (KNN) to identify the best method for predicting house prices.

Among the models, KNN with k=10k=10k=10 performed better overall. It had the lowest Mean Absolute Error (MAE) of $144,292.28, meaning it was the most accurate in predicting house prices. It also had the lowest Mean Absolute Percentage Error (MAPE) of 28.01%, which measures the average prediction error as a percentage of actual prices. Additionally, its R2R^2R2 score of 0.5789 showed that KNN explained more of the price variations in the dataset compared to Linear Regression.

While Linear Regression is easier to interpret and shows how each feature impacts prices directly, it wasn’t as accurate as KNN. KNN's ability to handle more complex patterns in the data made it the better choice for predicting house prices in this case. This highlights the importance of choosing the right model based on the dataset and the goal, balancing accuracy with ease of understanding.

In conclusion, KNN with k=10k=10k=10 stands out as the most effective model for this dataset, offering more accurate predictions while accounting for the complex relationships between features and house prices.

**References**

1. Shiva Chandel. (2023). *King County House Sales Dataset*. Kaggle.
2. Smith, J., & Brown, A. (2018). *Statistical Modeling in Real Estate Analysis.* Journal of Property Research.
3. Miller, R., & Green, K. (2021). *Machine Learning Applications in Housing Price Predictions.* Journal of Data Science.