The U.S. housing market is diverse, with significant regional variation in house prices influenced by a wide range of factors. As housing affordability continues to be a key issue in many parts of the country, understanding what drives house prices has become crucial for both buyers and sellers. This analysis seeks to identify the primary factors influencing house prices, focusing on basic features like **bedrooms**, **bathrooms**, and **home type**. We aim to determine the extent to which these factors predict house prices and explore whether regional variations play a significant role in price determination.

**Objective:**

The primary goal of this analysis is to create a model that predicts house prices based on commonly available features and assess how well these factors explain price variability across different states. We also aim to evaluate how the model performs in predicting prices for higher-value homes and identify areas where the model may struggle.

### **Source:**

### The data used in this analysis is sourced from Zillow, a leading real estate marketplace, and contains detailed information on housing properties across the United States. The dataset includes variables such as price, number of bedrooms, number of bathrooms, home type, and location (state and city), among others.

**Procedure:**

1. Data Cleaning: The dataset was preprocessed to remove entries with missing, zero, or infinite values for house prices. This ensured that only valid data points were used in the analysis.
2. Feature Selection: We selected bedrooms, bathrooms, and home type as the primary independent variables for the model. Although variables like living area and year built are important predictors of house prices, they were excluded from the model due to high collinearity with other features. For instance, living area tends to increase proportionally with the number of bedrooms and bathrooms, leading to redundancy in the data. Including these highly correlated features would distort the coefficients in the regression model, making it less reliable.
3. Modeling: A linear regression model was built using the least squares method. Dummy variables were created for home type to account for categorical data, and the model was trained on the selected features.
4. Evaluation: The model was evaluated using R-squared, mean squared error (MSE), and residual analysis to measure its performance and accuracy in predicting house prices.

### **Results:**

* **Initial Linear Regression**:
  + **R-squared: 0.26**: The model’s R-squared value indicates that only 26% of the variation in house prices is explained by the features (bedrooms, bathrooms, and home type). This low value suggests that while these features are somewhat predictive, 74% of the variability in house prices is driven by other factors not included in the model. As a result, the model may fail to capture important aspects of price variation, particularly for high-value homes.
  + **Mean Squared Error (MSE):** The model’s MSE was 167,497,164,786.96, indicating a large average squared difference between predicted and actual prices. The high MSE reflects the model's poor performance, especially for predicting high-end properties. Large errors occur more frequently in expensive homes, which likely have additional influential factors not captured by this simple model.
  + **Coefficients**:
    - **Bedrooms**: The coefficient for bedrooms suggests that, all else being equal, adding one more bedroom increases the predicted house price by $31,639.
    - **Bathrooms**: The coefficient for bathrooms indicates that each additional bathroom adds $217,777 to the predicted house price.
    - **Home Type**: Different home types (e.g., single-family vs. multi-family) were encoded as dummy variables, and the coefficients suggest how much a home’s type influences its price. For example, a single-family home increases the predicted price by $301,666, compared to the baseline category.
* **Logarithmic Transformation**:
  + To better handle skewed price data and high-value homes, we applied a **logarithmic transformation** to house prices.
  + **R-squared: 0.198**: The logarithmic model explained 19.8% of the variation in the transformed price data. While the log transformation helped to reduce skewness, it still did not improve the model significantly in terms of capturing variability in prices, especially for high-value homes.
  + **Root Mean Squared Error (RMSE)** for the log-transformed model was $415,272, indicating that the model still struggled to predict house prices accurately, particularly for homes in the higher price range.
* **Binning for Bedrooms and Bathrooms**:
  + By categorizing bedrooms and bathrooms into bins (e.g., homes with 1-2 bedrooms vs. 3-4 bedrooms), we captured non-linear relationships between these features and price.
  + **R-squared: 0.23**: This method yielded an R-squared value of 0.23, which, although similar to the initial linear model, allowed for more interpretable coefficients.
  + **Meaning of Binned Coefficients**:
    - For homes with 3-4 bedrooms, the predicted price increase was $88,229, while for homes with 5+ bedrooms, the predicted price was $102,859.
    - The price increase for homes with more bathrooms was larger, with homes containing 4+ bathrooms increasing the predicted price by over $149,570.
* **Median House Prices by State**
  + **Description:** We calculated the median house price for each state in the dataset to understand regional price differences.
  + **Analysis:** The median house prices across different states revealed significant regional variation. States like California (CA), Hawaii (HI), and the District of Columbia (DC) had the highest median house prices, exceeding $700,000, reflecting the impact of urbanization, market demand, and geographical desirability. In contrast, states such as Indiana (IN), Nebraska (NE), and Iowa (IA) had much lower median prices, hovering around $150,000. This highlights the stark differences between more densely populated, economically active regions and more rural states. The analysis shows the importance of considering location when analyzing house prices.

**Final Insights**:

* + Across all three approaches (normal linear regression, logarithmic transformation, and binning), the model consistently underperformed for higher-priced homes. The error distribution widened as house prices increased, suggesting that luxury homes have additional factors influencing their value—factors not captured in the model's features.
  + The median house price by state analysis highlighted significant regional differences in house prices, with states like California (CA), Hawaii (HI), and District of Columbia (DC) showing notably higher median prices than other states such as Indiana (IN) and Nebraska (NE). These differences emphasize the importance of considering geographic factors when building predictive models. Location, particularly at a more granular level than just state, plays a critical role in determining house prices. As a result, future models should incorporate location-specific variables (e.g., neighborhood-level data) to better capture regional housing market dynamics and improve overall model accuracy.

### **Criticism of Procedure:**

* **Limited Features**: By focusing solely on bedrooms, bathrooms, and home type, the model ignores other critical variables that influence house prices, such as location (beyond just the state), amenities, and market conditions. Living area and year built, which are significant predictors of house price, were excluded due to high collinearity with other features (e.g., number of bedrooms). While this decision avoided multicollinearity, it may have also reduced the model's overall explanatory power. Incorporating these features in a more sophisticated model will lead to better predictions.
* **Linear Model Assumptions**: The linear regression model assumes a linear relationship between features and price, which may not hold true for high-end or luxury homes. This assumption likely contributed to the model's poor performance for expensive properties.
* **Geographic Aggregation**: Grouping data at the state level may obscure important differences within states, particularly in regions with high variability in house prices (e.g., urban vs. rural areas). More granular location data would likely improve the model's accuracy.
* **Lack of Normalization**: The features used in the model (e.g., bedrooms, bathrooms, and price) were not normalized, meaning they were kept in their original scales. Bedrooms and bathrooms have small ranges, while house prices span several orders of magnitude. Without normalization, models such as linear regression may give disproportionate weight to features with larger scales (like price), leading to potential bias in the coefficients. Normalizing the features would ensure that all features contribute equally to the model, potentially improving model performance, particularly in gradient-based methods.

### **Conclusion:**

This project aimed to analyze the factors influencing house prices across the United States, focusing on commonly available features such as bedrooms, bathrooms, and home type. Through the use of linear regression models, we explored the extent to which these features predict house prices and identified areas where the model struggled—particularly for higher-end homes. While the model was able to explain some variation in house prices, with an R-squared of around 0.26, it became clear that additional factors, such as location, amenities, and property-specific features, are essential for more accurate predictions.

Our analysis of median state prices highlighted significant geographic disparities in housing costs, reinforcing the importance of location in price determination. Furthermore, the decision to exclude features like living area and year built due to collinearity limitations impacted the model's capacity to capture the full complexity of price variation. Finally, the lack of normalization and the reliance on linear assumptions limited the model's performance, especially in high-value housing markets.

In conclusion, while bedrooms, bathrooms, and home type provide a useful starting point for analyzing house prices, more advanced models incorporating non-linear relationships, granular location data, and additional housing characteristics are required to accurately predict house prices across diverse market segments. Future projects could aim to address these gaps, using richer datasets and more sophisticated techniques to capture the complexities of the U.S. housing market.