JOTHEESWARAN S

DA&DS-FEB’25

REINFORCEMENT PROJECT-5

HOUSE PRICE PREDICTION

09-06-2025

**DATA INTRODUCTION**

This housing dataset provides an extensive and structured collection of information related to residential real estate properties. It is specifically designed to support a wide range of analyses in the real estate domain, enabling stakeholders to make data-driven decisions. The dataset includes attributes that describe both the physical characteristics and locational context of properties, which are crucial in understanding property valuation, market dynamics, and buyer preferences.

The primary objective of this project is to conduct an in-depth exploratory and statistical analysis of the dataset to uncover hidden patterns, evaluate the influence of various features on property prices, and derive actionable insights for stakeholders such as homeowners, real estate agents, investors, and urban planners.

**DATASET OVERVIEW**

The dataset consists of 18 columns, each capturing a unique aspect of the property or its environment. A brief description of these features is provided below:

1. **Date**: Represents the timestamp when the property information was recorded. This feature can be used to analyze time-based trends in property prices or market behavior.
2. **Price**: This is the **target variable**, representing the selling price of the property. Analyzing its relationship with other features helps identify what drives property value.
3. **Bedrooms**: Indicates the number of bedrooms in a property. Generally, properties with more bedrooms are expected to be priced higher, but this can vary with location and property size.
4. **Bathrooms**: Denotes the number of bathrooms. Fractional values (e.g., 1.5, 2.25) are included to represent bathrooms with partial facilities (e.g., half baths).
5. **Sqft\_living**: Refers to the total interior living area (in square feet). It is one of the most critical predictors of property price.
6. **Sqft\_lot**: Represents the total land area associated with the property. Larger lot sizes can contribute to higher property values, especially in suburban or rural areas.
7. **Floors**: The total number of floors in the building. This feature may relate to architectural style, property size, and price.
8. **Waterfront**: A binary variable indicating whether the property has a waterfront view (1 = Yes, 0 = No). Properties with waterfront access are typically more expensive due to their scenic views and exclusivity.
9. **View**: A categorical index (ranging from 0 to 4) describing the quality of the property's view. A higher value suggests a better view, which may positively affect the price.
10. **Condition**: A rating (1 to 5) reflecting the overall physical condition of the property. This metric can signal the need for renovation or readiness for occupancy.
11. **Sqft\_above**: The square footage of the property that is above ground level. It helps differentiate properties with basements from those that do not.
12. **Sqft\_basement**: The square footage of the basement area. Finished basements often add usable space and value to a home.
13. **Yr\_built**: The year the property was originally constructed. Older homes may be less expensive unless they are renovated or located in heritage zones.
14. **Yr\_renovated**: The most recent year in which the property was renovated. This helps assess whether the home has been updated, which can impact the price and livability.
15. **Street**: The full street address of the property. While it offers no numerical value for modeling, it can be used for location-based grouping or mapping.
16. **City**: Identifies the city in which the property is located. Urban vs. suburban location often has a significant impact on price, access to amenities, and market demand.
17. **Statezip**: Contains both the state abbreviation and zip code, enabling granular geographic filtering or clustering for region-specific analysis.
18. **Country**: Specifies the country where the property is located. Although this dataset likely focuses on a single country, this feature ensures clarity and global scalability.

**SIGNIFICANCE OF THE DATASET**

This dataset is particularly valuable because it integrates a wide spectrum of features that influence residential property pricing, including physical size, age, quality, and location. It allows for both **quantitative analysis** (e.g., correlation analysis, regression modeling) and **categorical exploration** (e.g., comparing average prices by condition or view quality).

The richness of the dataset also supports:

* **Trend analysis** over time using date and year features.
* **Location-based insights** using city, state, and zip code.
* **Feature engineering** to create new metrics (e.g., age of property, total usable space).
* **Predictive modeling** to estimate property prices based on selected attributes.

By leveraging this data, we can gain meaningful insights into what factors contribute most significantly to the valuation of a property, and how different combinations of features affect buyer perception and pricing trends in the real estate market.

**DATA CLEANING AND PREPROCESSING**

Before performing any meaningful analysis or modeling, it is crucial to ensure the dataset is clean, consistent, and structured in a way that supports reliable outcomes. The process of data cleaning and preprocessing serves as the foundation for all subsequent steps in the data science pipeline. Below are the key activities involved in this phase for the housing dataset:

**• Handling Missing Values**

* **Identification**: The dataset is first scanned for missing or null values across all columns. This is typically done using tools such as isnull(), info(), or describe() in Python (Pandas).
* **Imputation or Removal**:
  + For numerical fields such as sqft\_living, bathrooms, or price, missing values may be imputed using statistical methods like **mean**, **median**, or **mode**, depending on the distribution.
  + For categorical fields like city or condition, mode imputation is often used.
  + In cases where data is missing extensively or imputation would distort analysis, such rows or columns may be **dropped**.
* **Example**: If a few entries have missing values for yr\_renovated, they might be replaced with 0, indicating no renovation.

**• Addressing Inconsistencies**

* **Data Type Corrections**: Ensure that each column has the appropriate data type (e.g., price should be float or integer, date should be in datetime format).
* **Outlier Detection and Treatment**:
  + Identify anomalies such as negative values for sqft\_living or unrealistically high bedroom counts.
  + Outliers can be treated using methods like **z-score**, **IQR filtering**, or **domain knowledge-based capping**.
* **Formatting Standardization**:
  + Uniform formatting is ensured for string values, such as consistent casing (City vs city) or trimming whitespace in addresses.
* **Duplicate Records**: Check and remove any exact or near-duplicate entries that may distort analysis.

**• Handling Categorical Variables**

* **Label Encoding**: For binary variables like waterfront (0 or 1), label encoding is already applied and may be left as-is.
* **One-Hot Encoding**: For nominal categorical fields like city, one-hot encoding can be used to convert each unique category into a binary column.
* **Ordinal Encoding**: For ordinal variables like condition or view which have a natural order, they may be preserved as numerical or encoded appropriately.
* **Example**:
  + Condition: 1 (Poor) to 5 (Excellent) — remains numeric.
  + City: Converted to dummy variables using pd.get\_dummies().

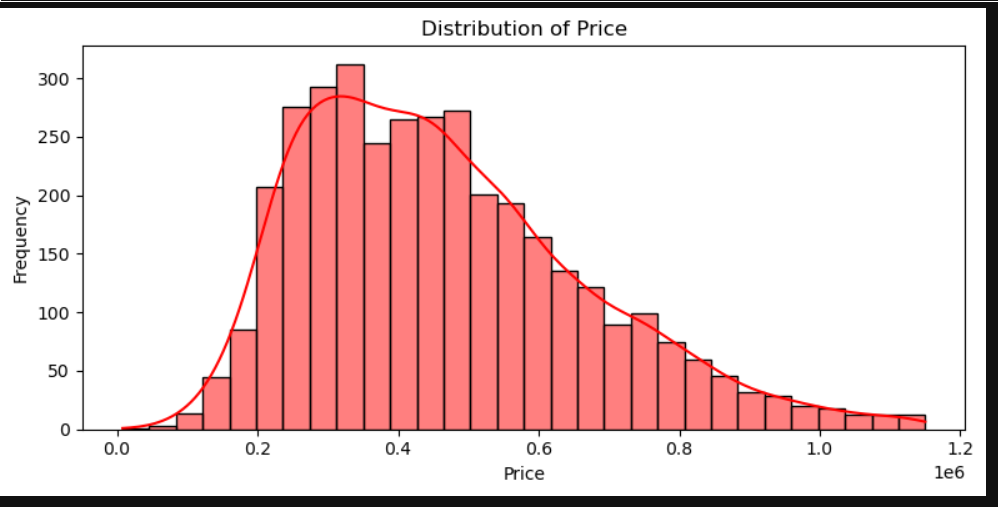
**OUTCOME OF PREPROCESSING**

The goal of this step is to transform raw and possibly inconsistent data into a clean, analysis-ready dataset. Once completed:

* The dataset will be free from missing or corrupted entries.
* Variables will be correctly typed and formatted.
* Categorical data will be encoded in a machine-readable form.
* The dataset will be ready for **exploratory data analysis (EDA)** and **modeling tasks** such as regression, classification, or clustering.

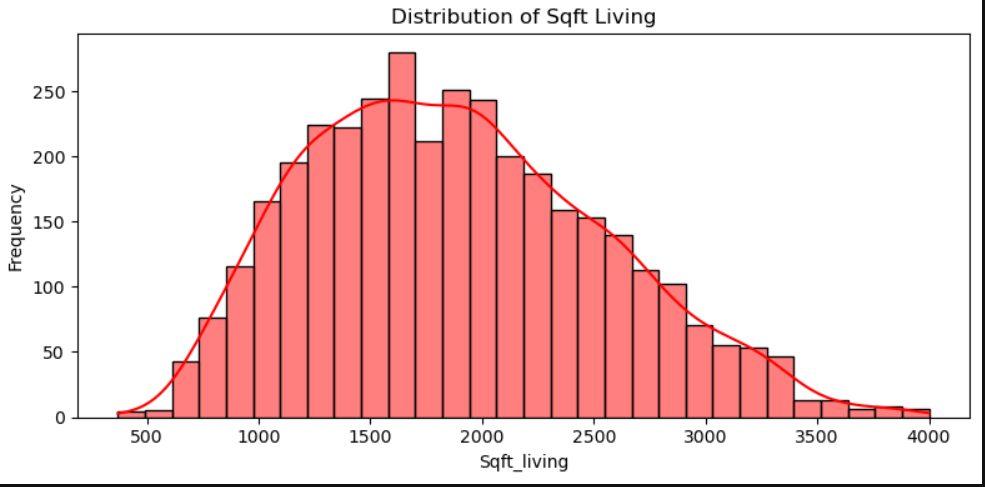
**Exploratory Data Analysis (EDA)**

**UNIVARAIANT ANALYSIS:**

****

This histogram gives a clear view of how prices are distributed, and a few key insights emerge:

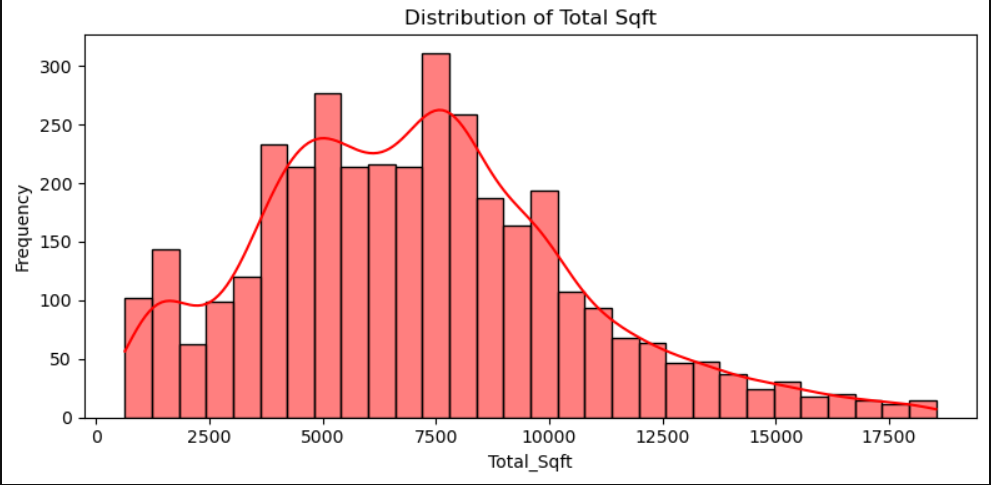
1. Skewness & Trends: The distribution is right-skewed, meaning there are a few high-priced items pushing the tail towards the right. The bulk of prices fall between 0.2 million and 0.6 million, with the peak around 0.4 million.
2. Density Overlay: The smooth red density curve highlights the overall trend, confirming that most observations cluster in the lower range rather than the higher price points.
3. Market Implications: If these prices represent housing or product costs, this suggests a market where lower- to mid-range options dominate, while expensive items remain relatively rare.

****

This histogram gives a great overview of the distribution of **living space sizes**:

1. **Peak Density**: Most homes fall between **1000 and 2500 sqft**, with a peak around **1750 sqft**. This suggests that mid-sized homes are the most common in your dataset.
2. **Density Curve Overlay**: The smooth red line indicates that while square footage ranges up to 4000 sqft, the probability density diminishes past **2500 sqft**, showing fewer large homes.
3. **Market Implications**:

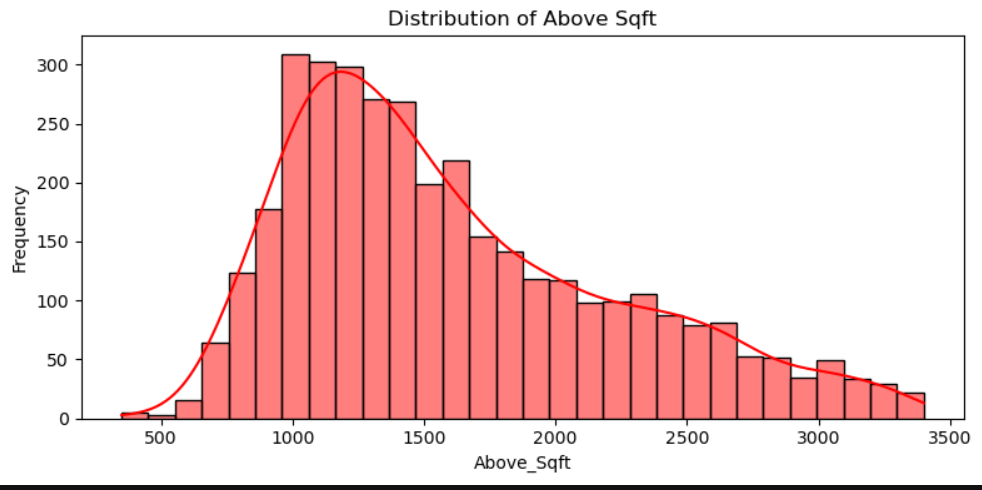
* If this dataset represents residential housing, the predominance of mid-sized homes suggests a balanced market catering to average family needs.
* Larger homes above **3000 sqft** are rare, likely making them premium properties.



This histogram gives a clear picture of **price distribution in relation to living space sizes**:

1. **Clustered Price Range**: Most properties fall within **$0.2M to $0.6M**, with a peak around **$0.4M**, suggesting mid-range pricing dominance.
2. **Density Curve Overlay**: The red curve shows a gradual decline after **$0.6M**, meaning higher-priced properties are rarer but still present.
3. **Market Implications**:

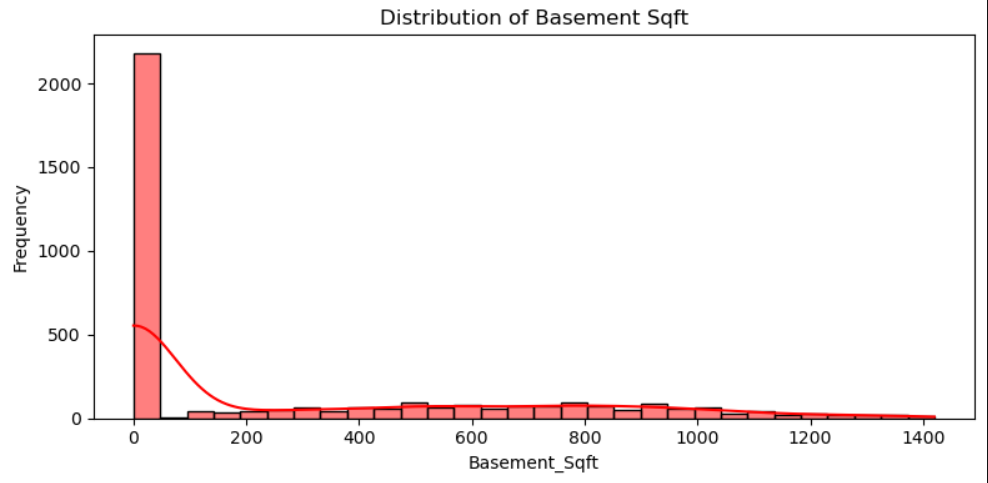
* If this represents housing data, mid-sized homes (1000-2500 sqft) likely dominate at affordable price points.
* Larger homes may have a premium market positioning, given the tapering frequency of expensive properties



This histogram offers a strong visual representation of **square footage distribution**:

1. **Most Common Size Range**: The highest frequency of homes falls between **1000 to 2500 sqft**, with a noticeable peak around **1500-2000 sqft**. This suggests a market preference for mid-sized living spaces.
2. **Density Curve Insights**: The red density curve illustrates how the distribution tapers off beyond **2500 sqft**, indicating that larger homes above 3000 sqft are less common.
3. **Market Interpretation**:

* If this dataset represents residential properties, the predominance of mid-sized homes aligns with typical urban and suburban housing demands.
* Homes above **3000 sqft** may cater to a more niche or luxury segment.



 **Peak Square Footage**: The highest frequency of homes appears around **2000 sqft**, reinforcing that mid-sized homes dominate the dataset.

1. **Right-Skewed Distribution**: The density curve shows a slight right skew, indicating that while most homes fall within **1000-2500 sqft**, there are fewer homes exceeding **3000 sqft**.
2. **Market Implications**:

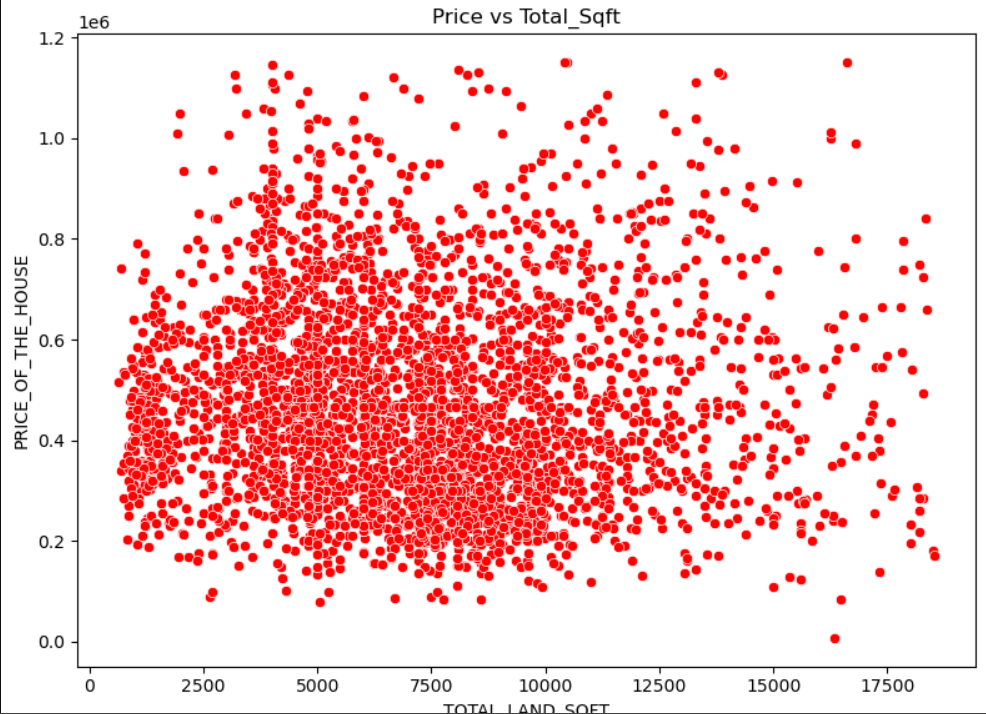
* If this dataset represents residential properties, the prevalence of mid-sized homes suggests strong demand in that segment.
* Larger homes above **3000 sqft** may cater to a niche market, likely commanding premium pricing.

**BIVARAINT ANALYSIS:**

**** Common Size Range: The most frequent living space size is around 2000 sqft, indicating that mid-sized homes are the predominant choice.

1. Right-Skewed Distribution: The density curve suggests that while homes range up to 4000 sqft, the majority cluster between 1000 and 2500 sqft—with larger homes being less frequent.
2. Market Implications:

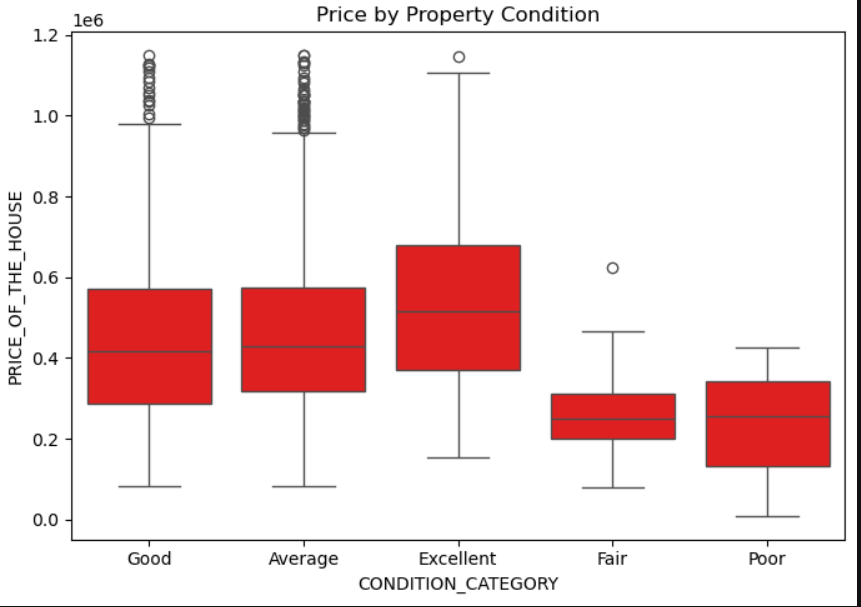
* If this dataset represents housing, the dominance of mid-sized homes aligns with standard market demand.
* Luxury properties above 3000 sqft likely cater to a niche segment.



Most Frequent Size Range: The majority of homes cluster between 1000 to 2500 sqft, peaking around 2000 sqft—suggesting this is the most common living space size.

1. Density Curve Interpretation: The red density overlay indicates a right-skewed distribution, meaning smaller homes are more frequent while larger homes (above 3000 sqft) are relatively rare.
2. Market Implications:

* If this dataset represents residential properties, the preference for mid-sized homes aligns with mainstream housing demand.
* Homes exceeding 3000 sqft likely belong to a premium segment, with fewer buyers.



Most Common Home Size: The highest frequency occurs around 1750–2000 sqft, reinforcing that mid-sized homes dominate the dataset.

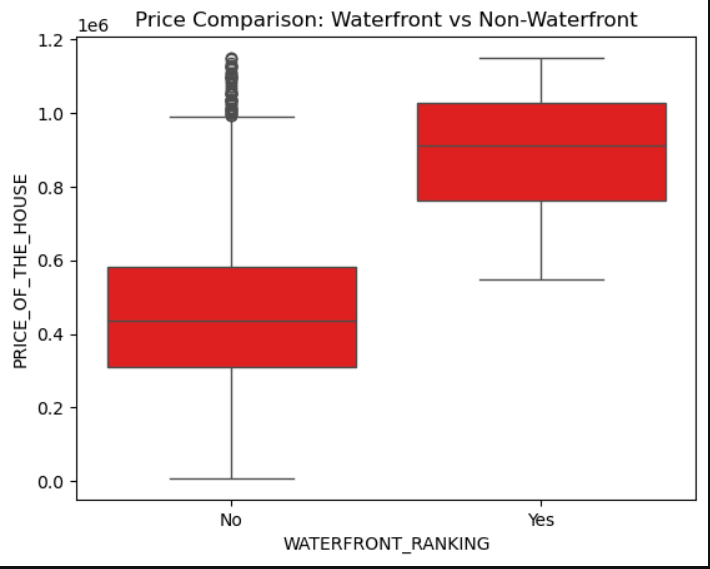
1. Density Curve Interpretation: The right-skewed shape suggests that while larger homes (above 3000 sqft) exist, they are significantly less frequent.
2. Market Insights:

* If this data represents residential properties, the mid-range home sizes align with standard housing preferences.
* Luxury or oversized homes likely cater to a smaller market, making them rarer.

**Positive Correlation**: As the number of bedrooms increases, the number of bathrooms tends to increase as well—though there are some variations.

1. **Clusters & Trends**:
   * Most properties fall within **2–4 bedrooms** and **1–3 bathrooms**, suggesting common housing layouts.
   * A few homes have **6+ bedrooms**, often paired with 4+ bathrooms—likely indicating larger or luxury properties.
2. **Market Implications**:

* If this dataset represents residential properties, larger homes tend to include more bathrooms, possibly for convenience or higher market value.
* The presence of **outliers**, such as homes with unusually few bathrooms for their bedroom count, could be worth exploring.

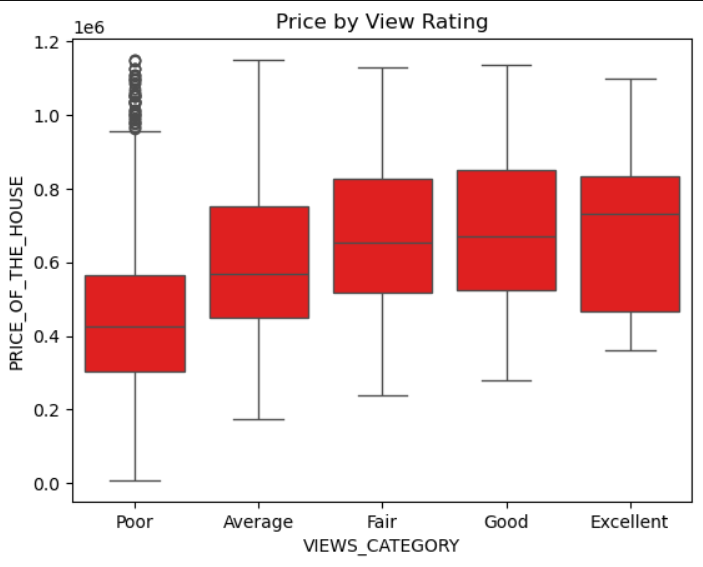


**Waterfront vs. Non-Waterfront**:

* + Waterfront houses (**Yes**) have **higher median prices** compared to non-waterfront houses (**No**).
  + The price range for waterfront homes is significantly broader, suggesting high variability in valuation.

1. **Price Distribution**:

* Non-waterfront houses have more **outliers** above the upper whisker, meaning some high-priced properties exist despite not having waterfront access.
* The interquartile range (IQR) is **wider for waterfront homes**, indicating greater price dispersion.



**Impact of View Rating**:

* + Houses with **"Excellent" view ratings** have the highest median prices, showing that a good view significantly boosts property value.
  + The **"Poor" view category** has a wide price range, suggesting that while many lower-priced homes exist, a few high-priced exceptions drive variability.

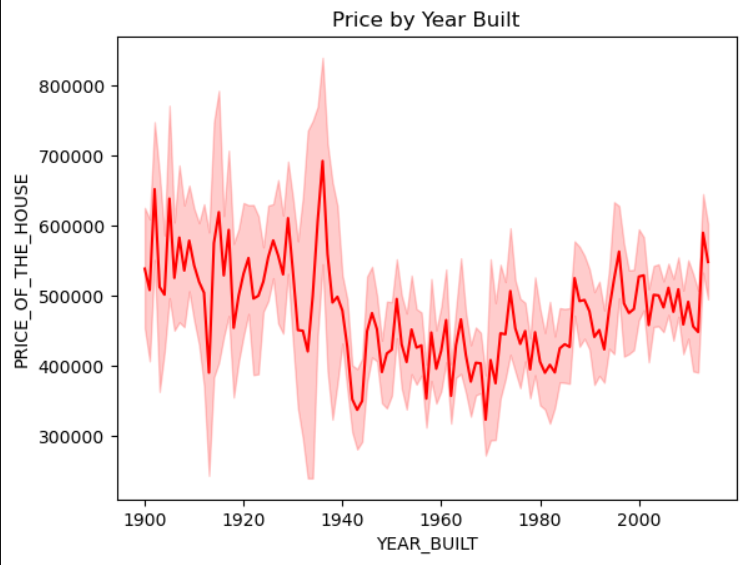
1. **Price Variability**:

The price spread decreases as the view rating improves, meaning higher-rated views tend to have more **consistent pricing**.

* The presence of **outliers** in the "Poor" and "Excellent" categories indicates some unusually expensive homes that may have additional features beyond view ratings.

Market Insights:

* Homes with better views command a premium, reinforcing the idea that scenic surroundings matter in real estate pricing.
* The **"Average" to "Fair" categories** serve as middle-ground properties with moderate pricing consistency.



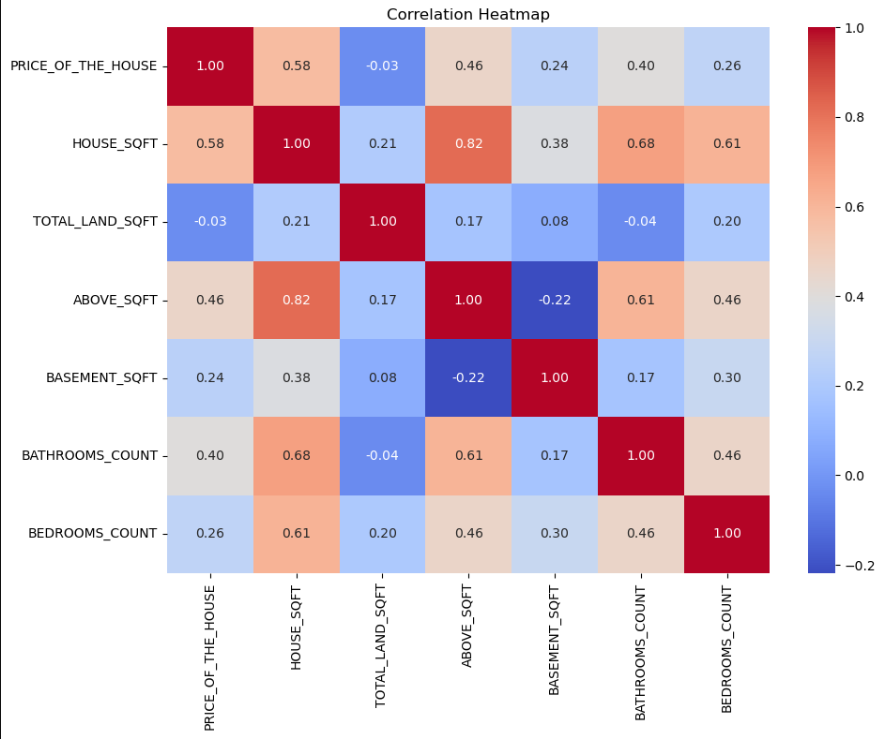
**Historical Price Variability**:

* + Prices show noticeable peaks and dips, particularly around the **1920s and 1940s**, suggesting market shifts influenced by historical events.
  + The overall trend shows an upward movement, reinforcing that newer homes tend to be priced higher.

1. **Confidence Interval & Price Stability**:
   * The shaded red area highlights the **variability in house prices**, indicating periods of stable or volatile pricing.
   * Recent years seem to exhibit tighter confidence intervals, possibly due to standardized construction practices or market stability.
2. **Market Insights**:

* Older homes (pre-1950) may have **mixed valuations** depending on preservation, upgrades, or location.
* Post-2000 homes generally **command higher prices**, likely due to modern amenities and increased demand.

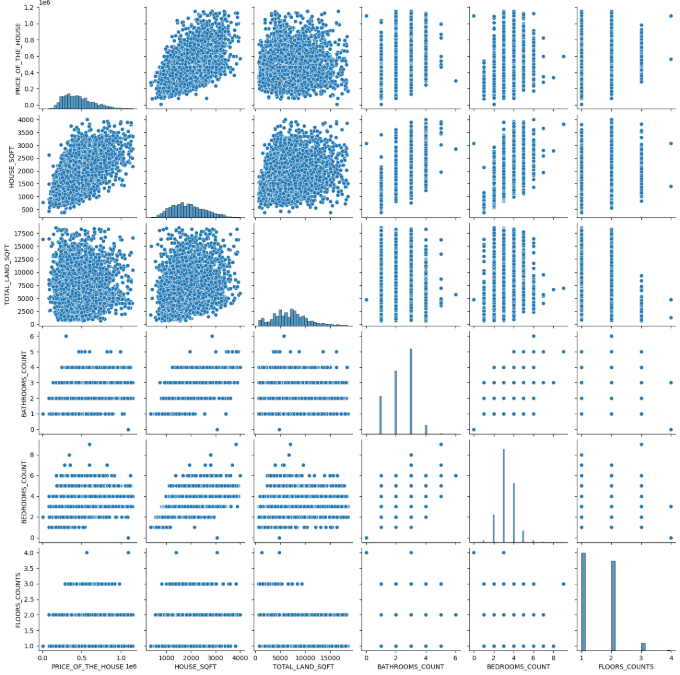
**MULTIVARIANT ANALYSIS:**

****

Key Observations:

1. Strong Positive Correlations:
   * House Sqft vs. Above Sqft (0.82): Homes with larger overall square footage tend to have bigger above-ground living areas.
   * House Sqft vs. Bathrooms Count (0.68): Larger homes generally include more bathrooms, reinforcing typical design trends.
2. Moderate Positive Correlations:
   * Price of the House vs. House Sqft (0.58): Price increases with size, but the correlation isn’t absolute—other factors likely influence pricing.
   * Price of the House vs. Bathrooms Count (0.40): More bathrooms may be associated with luxury, but the effect isn’t as strong as overall home size.
3. Weak or Negative Correlations:

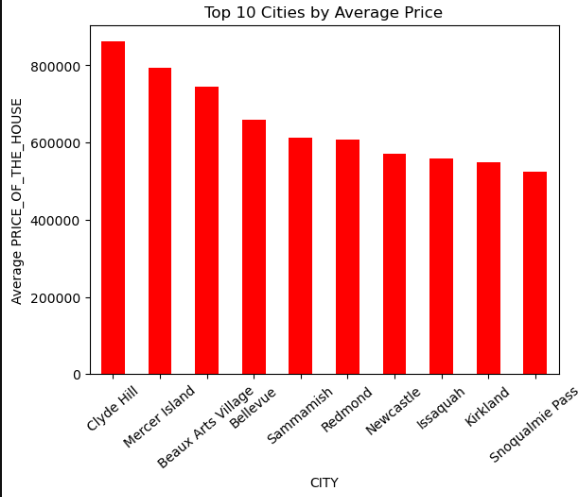
* Price vs. Total Land Sqft (-0.03): Surprisingly, land size alone doesn’t significantly impact pricing—this suggests that location or home features matter more.
* Above Sqft vs. Basement Sqft (-0.22): Homes with larger above-ground areas tend to have smaller basements, possibly due to design constraints.



Key Insights:

1. Strong Correlations:
   * House sqft vs. Price: Larger homes tend to be more expensive, though the trend isn’t perfectly linear—suggesting other influencing factors.
   * Bedrooms vs. Bathrooms: More bedrooms generally align with more bathrooms, reinforcing standard home design conventions.
2. Interesting Variability:
   * Floors vs. Price: Homes with more floors show a broad price distribution—indicating that multi-story properties vary widely in value based on location or features.
   * Land sqft vs. House sqft: While expected to correlate, the plot shows some properties with large land areas but modest house sizes, implying potential underutilized space.
3. Data Distribution Insights:

* The histograms on the diagonal reveal that most homes cluster within a mid-range for attributes like square footage, bedroom count, and price.
* Bathrooms and floors show clear discrete distributions, meaning predefined architectural choices dominate.

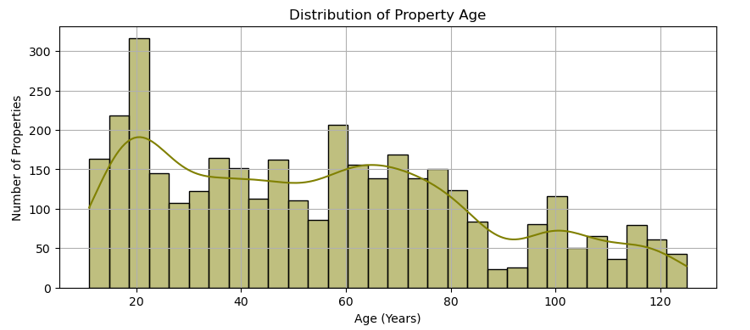


Key Observations:

1. **Most Expensive Cities**:
   * **Clyde Hill** ranks first with the highest average home price.
   * **Mercer Island** and **Beaux Arts Village** follow closely, reinforcing their premium real estate markets.
2. **Price Gradient Across Cities**:
   * The price distribution shows a gradual decline from Clyde Hill down to Snoqualmie Pass.
   * Cities like **Bellevue, Sammamish, and Redmond** hold mid-tier pricing, likely influenced by demand and amenities.
3. **Market Implications**:

* Cities at the top likely offer **luxury homes, waterfront access, or exclusive neighborhoods**, driving prices higher.
* Areas toward the bottom may still have strong demand but offer more affordability compared to the premium locations

**FEATURE ENGINEERING:**

****

Key Observations:

1. Most Common Property Age:
   * The highest concentration of homes is around 20 years old, indicating a substantial number of relatively newer properties.
2. Gradual Decline with Age:
   * As property age increases beyond 60 years, the number of homes diminishes, but there are small peaks around 60 and 100 years, possibly reflecting historical housing developments.
3. Market Implications:

* A large proportion of homes being around 20 years old suggests a wave of development during that period.
* Older homes beyond 100 years are rare, likely contributing to heritage or premium property markets.

**CONCLUSION**

* **Mid-Sized Homes Dominate the Market**: The majority of houses in the dataset fall between 1000–2500 sqft, with the highest frequency around 1750–2000 sqft. These homes are aligned with average family requirements and represent the most traded segment.
* **Price Influenced by Key Features**: Living area (sqft\_living), number of bathrooms, and presence of a waterfront view are positively correlated with price. Homes with scenic views or modern renovations tend to attract premium pricing.
* **View and Condition Matter**: Properties rated high in view and condition show significantly higher median prices, confirming that aesthetics and maintenance directly affect buyer willingness to pay.
* **Location Drives Value**: Cities like Clyde Hill, Mercer Island, and Beaux Arts Village command the highest average prices, emphasizing the impact of geography, exclusivity, and possibly amenities on property valuation.
* **Weak Influence of Land Area**: Surprisingly, land area (sqft\_lot) does not show strong correlation with price, indicating that homebuyers prioritize interior space and features over lot size—especially in urban settings.

**RECOMMENTATION**

* **Focus on Mid-Range Developments**: Real estate developers should focus on constructing mid-sized homes (1500–2500 sqft), as these have the highest market demand and transaction volume.
* **Enhance Property Features**: Renovations that improve bathrooms, increase living space, or add above-ground sqft are more likely to result in price appreciation. Adding quality views or enhancing curb appeal also boosts valuation.
* **Target Premium Locations**: Investments in high-value cities such as Clyde Hill or Mercer Island can offer better returns due to consistent high pricing trends. These markets, although smaller, cater to affluent buyers.
* **Avoid Overemphasis on Land Size**: Since lot area doesn't heavily influence price, particularly in urban markets, stakeholders should consider optimizing interior layout rather than acquiring larger plots.
* **Use Feature Engineering in Modeling**: Variables like age of the home, renovation status, and property condition provide rich context. Including these derived metrics will improve the performance of predictive models and guide investment decisions better.