**EXPLORATORY DATA ANALYSIS USING PYTHON**

**HOUSING DATASET ANALYSIS**

**R. SARANYA**

**DA&DS**

**MAY -2025 BATCH**

**(OFFLINE)**

**1.INTRODUCTION:**

The housing dataset provides comprehensive information on various attributes associated with residential properties, including price, number of bedrooms and bathrooms, square footage, location details, and other relevant features. The objective of this project is to conduct an in-depth analysis of the dataset to derive valuable insights for stakeholders in the real estate industry.

**2.AIM:**

The aim of this analysis is to perform a comprehensive analysis of residential housing data to uncover key insights into property pricing and influential features. By exploring variables such as price, location, size, condition, and amenities, the project seeks to identify patterns, correlations, and trends that impact real estate value. The objective also includes data cleaning, feature engineering, and visualization techniques to enhance understanding. Ultimately, this analysis will support better decision-making for real estate investors, developers, and buyers through data-driven insights and geospatial mapping of housing prices.

**3.PROBLEM STATEMENT:**

**4.PROJECT WORKFLOW:**

The housing data analysis follows a systematic workflow consisting of key steps that guide the project from raw data to actionable insights. The table below summarizes each step and its purpose:

| **Step** | **Description** |
| --- | --- |
| Data Collection | Load and inspect the housing dataset for structure and completeness. |
| Data Understanding | Explore the attributes, data types, and distributions to build context. |
| Data Cleaning | Handle missing values, outliers, and inconsistencies to ensure data quality. |
| Feature Engineering | Generate derived metrics such as price per square foot, property age, etc. |
| Data Filtering | Select relevant records for focused analysis based on criteria. |
| Statistical Analysis | Perform descriptive analysis and apply statistical tests where needed. |
| Exploratory Data Analysis | Visualize and interpret patterns using univariate, bivariate, and multivariate techniques. |
| Geospatial Mapping | Map average prices and metrics across locations for spatial insights. |
| Insight Generation & Report | Summarize key findings and visualizations for stakeholder understanding. |

**5.DATA UNDERSTANDING:**

1. Date: The date when the property information was recorded.

2. Price: The price of the residential property.

3. Bedrooms: The number of bedrooms in the property.

4. Bathrooms: The number of bathrooms in the property.

5. Sqft\_living: The total square footage of living space in the property.

6. Sqft\_lot: The total square footage of the lot or land area associated with the property.

7. Floors: The number of floors in the property.

8. Waterfront: Indicates whether the property has a waterfront view (binary: 0 for no, 1

for yes).

9. View: An index from 0 to 4 representing the quality of the view from the property.

10. Condition: An index from 1 to 5 representing the overall condition of the property.

11. Sqft\_above: The square footage of the interior space above the ground level.

12. Sqft\_basement: The square footage of the basement space in the property.

13. Yr\_built: The year when the property was built.

14. Yr\_renovated: The year when the property was last renovated.

15. Street: The street address of the property.

16. City: The city where the property is located.

17. Statezip: The state and zip code of the property.

18. Country: The country where the property is located.

**6. DATA CLEANING**

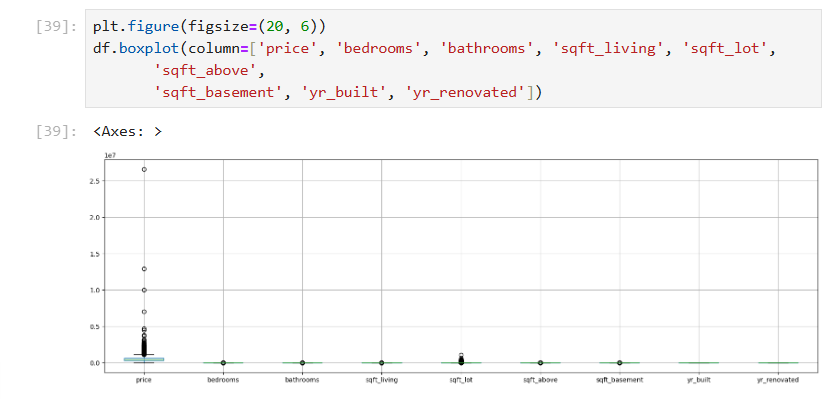
Data cleaning is a critical step in any data analysis project as it directly impacts the reliability and accuracy of the results. The housing dataset initially contained several imperfections, including missing values, outliers, and inconsistencies that needed to be addressed before proceeding with analysis.

**• Missing Values Imputation**

The dataset had missing entries in key features such as sqft\_lot and yr\_renovated. For numerical columns with few missing values, mean or median imputation was applied based on distribution characteristics. In some cases, such as renovation year, missing values were replaced with 0 to denote “never renovated” as a logical default.

**•Outlier Treatment**

Outliers are a datapoint, that differs from rest of data. For example, properties with excessively large square footage or unusually high prices were reviewed and excluded if deemed unrealistic.



**• Handling Inconsistent Values**

 Inconsistencies such as incorrect data formats (e.g., string-formatted dates), duplicated rows, or impossible values (e.g., zero bedrooms in a house) were identified and corrected. Duplicate rows were dropped to maintain dataset integrity. Categorical values such as city and statezip were standardized to ensure uniformity in grouping and analysis.

**7. OBTAINING DERIVED METRICS**

**Age of the Property:** A new column named age was generated by subtracting the year\_built value from the current year. This metric provides a clearer understanding of how old each property is, which can influence factors such as price, condition, and market demand. Instead of analyzing raw construction years, using property age allows for better grouping and comparison across listings.



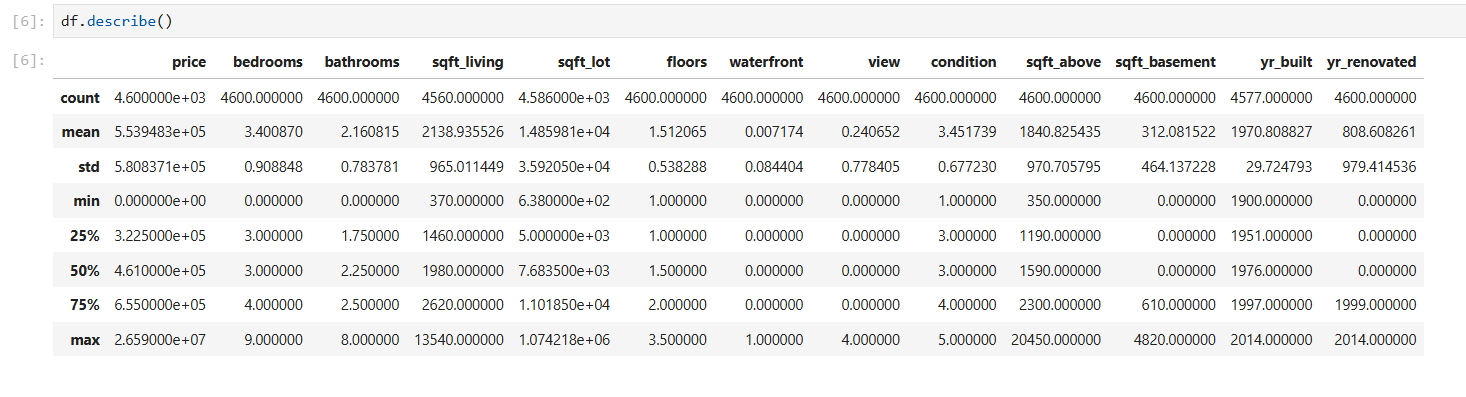
**8.FILTERING DATA FOR ANALYSIS**

**Date Components Extraction:** From the original date column, three separate columns — day, month, and year — were extracted. This transformation supports time-series analysis and enables trend evaluations across different time periods. For instance, it becomes easier to identify seasonal patterns in property sales or changes in pricing over specific months or years.

****

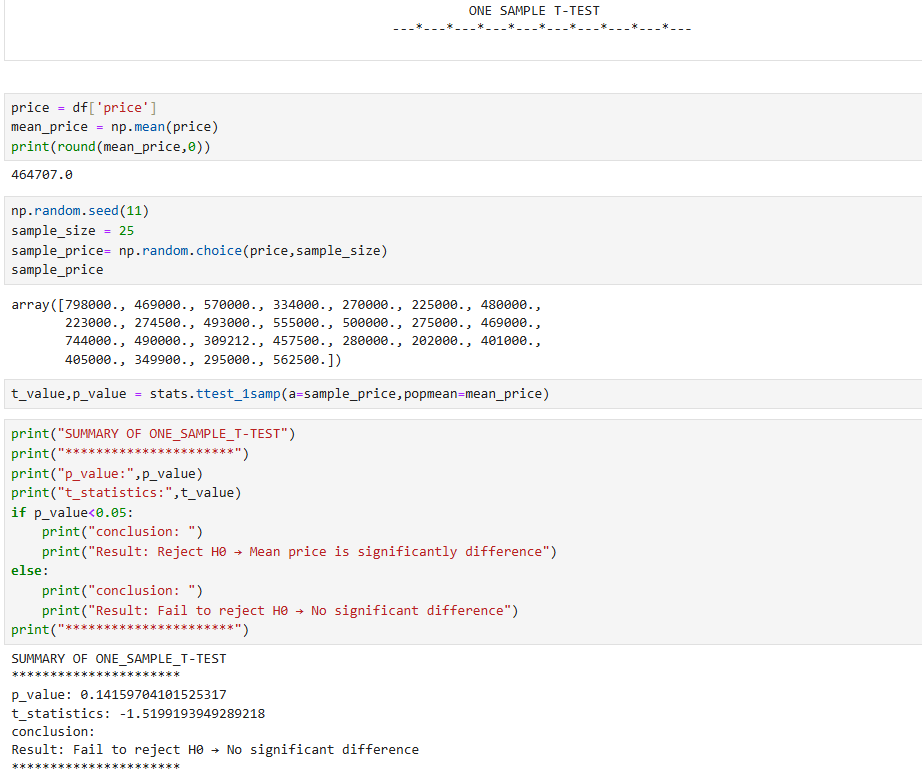
**9.STATISTICAL ANALYSIS**

**• Descriptive Analysis:**

* The average house price is approximately $553,948, with prices ranging from $0 to $26,590,000, indicating potential outliers or extreme luxury properties.
* The mean number of bedrooms is 3.4, while the median is 3, suggesting a slightly right-skewed distribution.
* For bathrooms, the average is 2.16, with most homes having between 1.75 and 2.5 bathrooms (IQR).
* The year-built ranges from 1900 to 2014, with a mean around 1970, indicating a mix of both old and new constructions.
* Living area (sqft\_living) has a wide range, with an average of 2,138 sqft, while lot size varies significantly, suggesting heterogeneity in property types.
* **** Categorical-like features such as waterfront and view have low average values, indicating most homes do not have these features.
* **Test Statistics and Hypothesis Testing**
* **One sample T-test:**

Hypotheses are

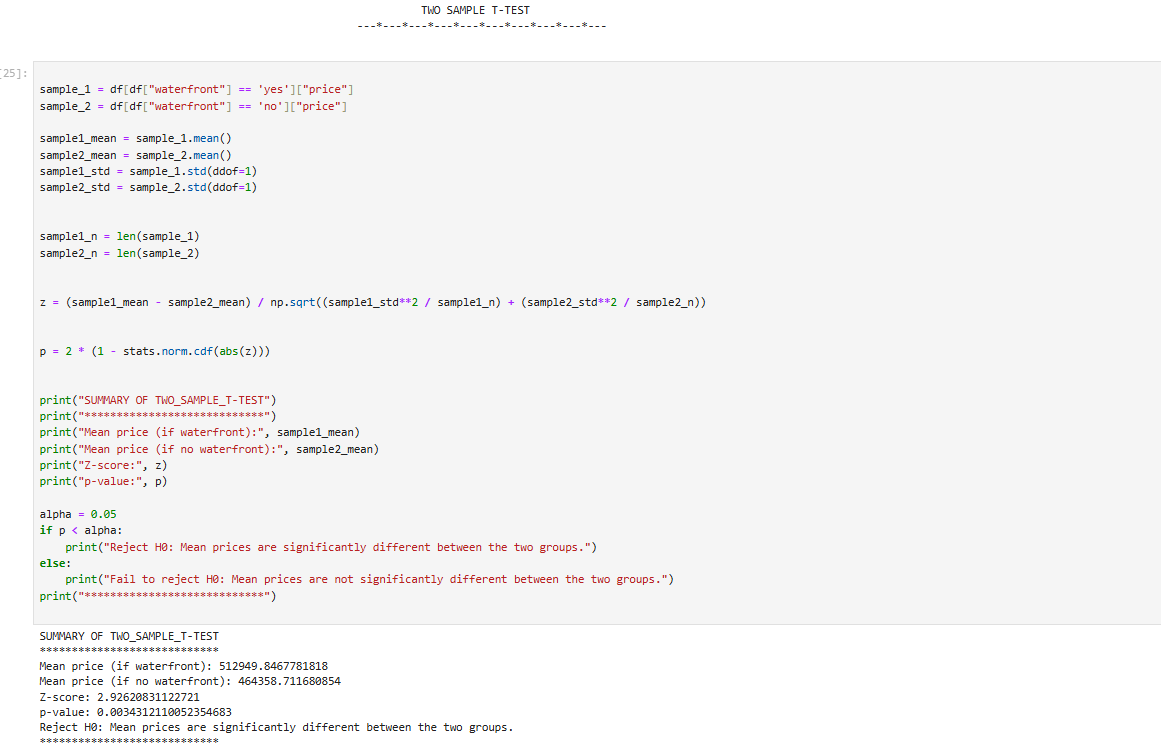
* Null Hypothesis (H₀): The mean price of the sample is equal to the population mean price.
* Alternative Hypothesis (H₁): The mean price of the sample is significantly different from the population mean price.

****

* **Two sample T-test:**

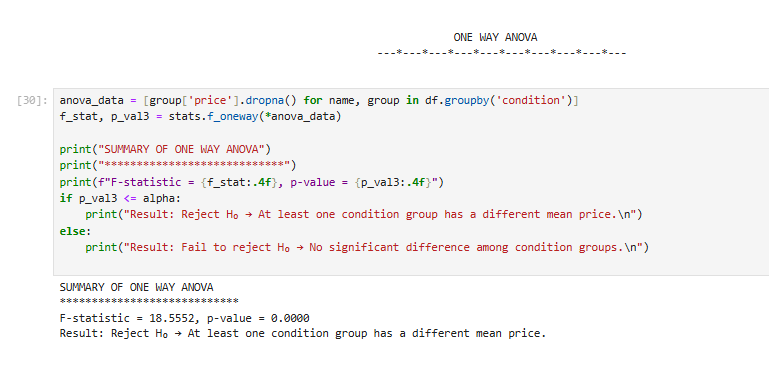
A two-sample t-test was conducted to determine whether there is a significant difference in the average house prices between waterfront and non-waterfront properties.

* Null Hypothesis (H₀): The mean house price is the same for waterfront and non-waterfront properties.
* Alternative Hypothesis (H₁): The mean house price differs between the two property types.

****

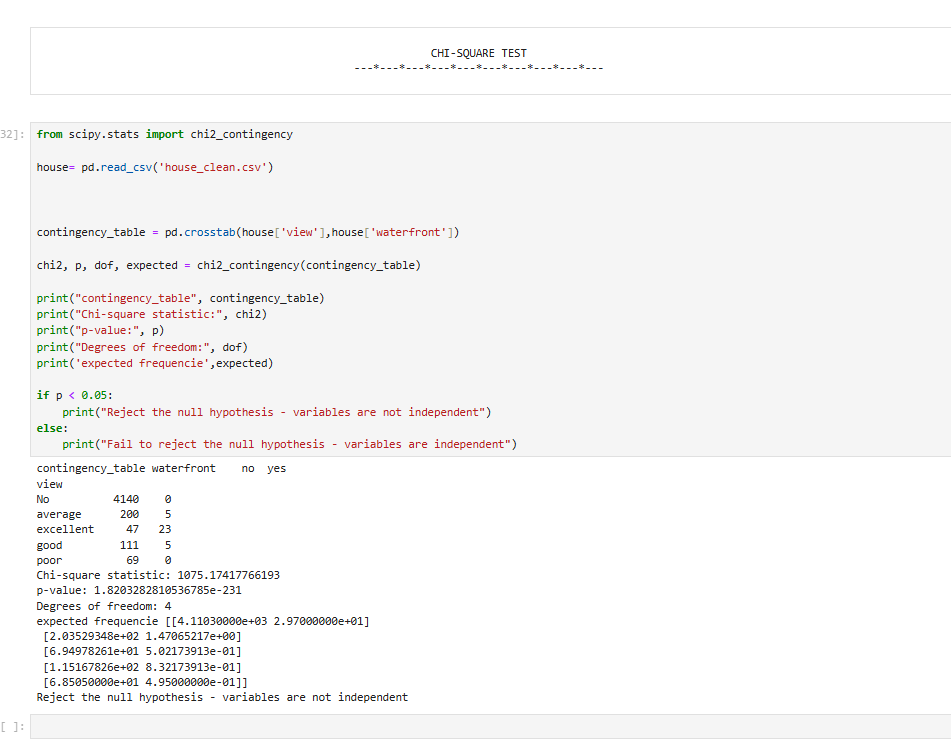
* **One way ANOVA**

A One-Way ANOVA (Analysis of Variance) was conducted to determine whether there are significant differences in mean house prices across different property condition levels.

* Null Hypothesis (H₀): All condition groups have the same mean price.
* ****Alternative Hypothesis (H₁): At least one condition group has a different mean price.
* **Chi-square test**

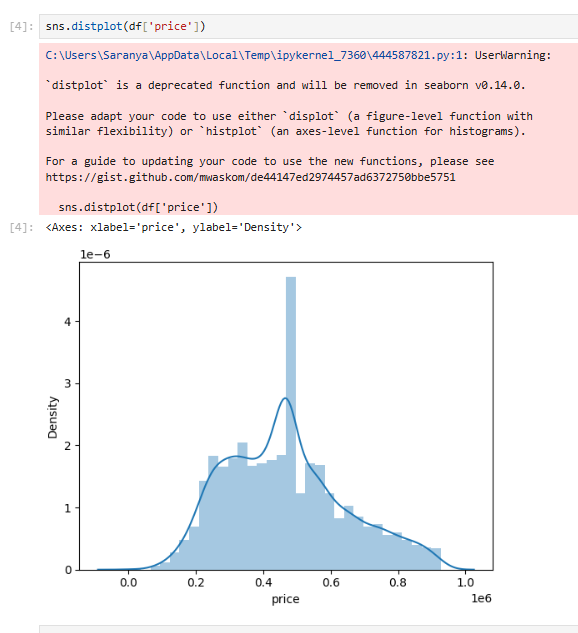
A Chi-Square Test of Independence was conducted to assess whether there is a significant association between the ‘view’ and ‘waterfront’ features of houses.

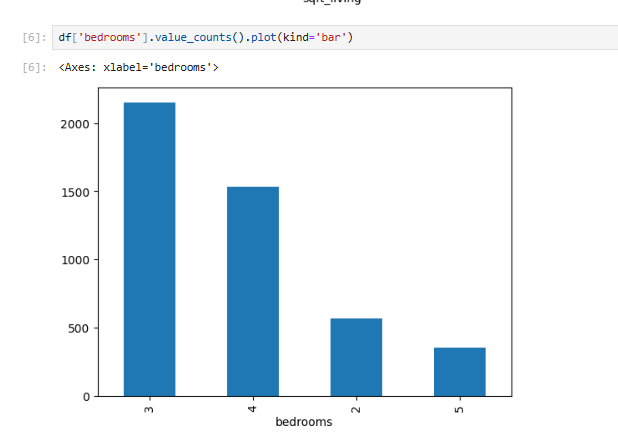
* Null Hypothesis (H₀): The variables view and waterfront are independent.
* Alternative Hypothesis (H₁): The variables view and waterfront are not independent.

****

**10.EXPLORATORY DATA ANALYSIS**

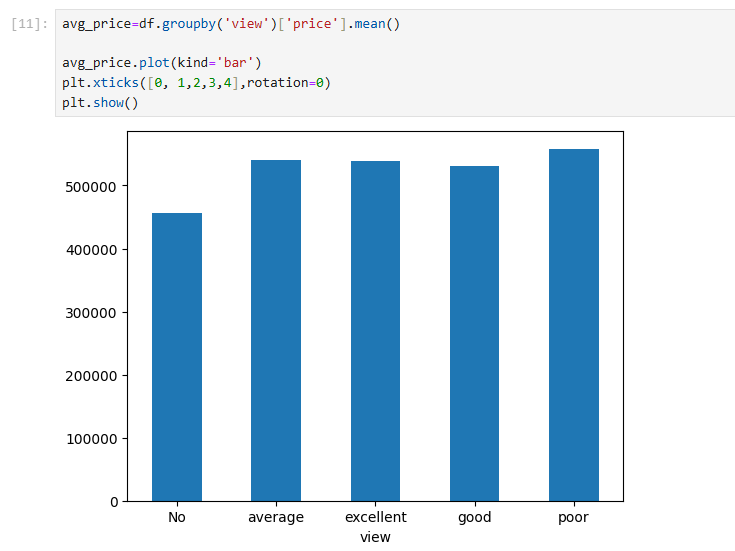
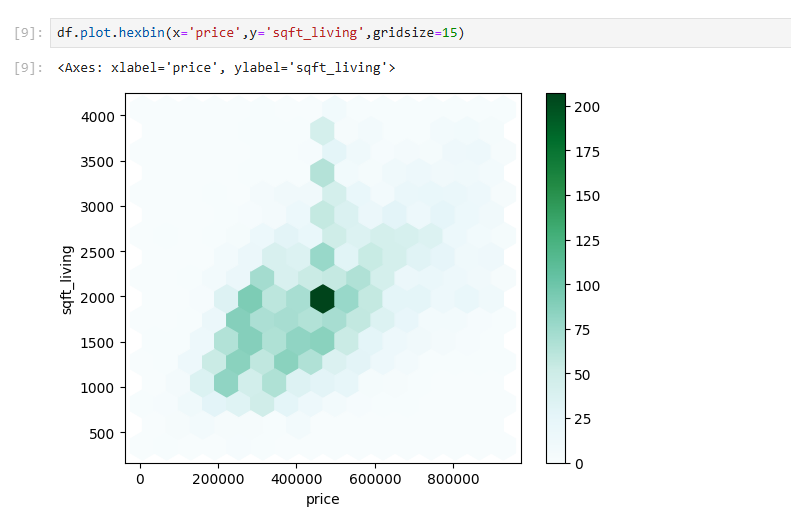
* **Univariate Analysis:**
* Univariate analysis was performed to examine the distribution of individual variables in the dataset. A density plot was created for the price column, showing a moderately right-skewed distribution. This visualization helped in identifying potential outliers and understanding the central price tendency across the dataset.
* Additionally, a bar chart was plotted for the bedrooms feature, where it was found that most houses have 3 or 4 bedrooms. This frequency distribution aids in understanding the common housing structure and will assist in feature engineering during model development.

****



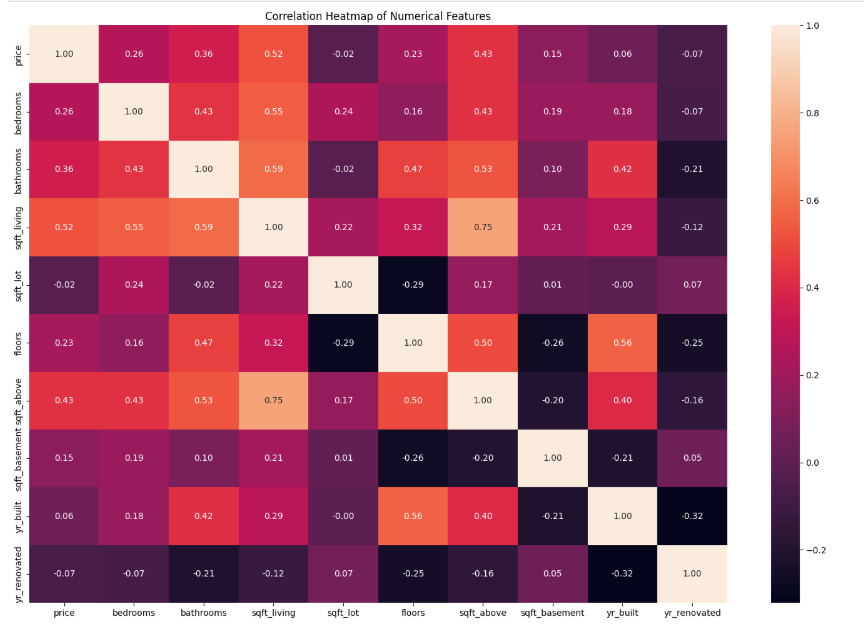
* **Bivariate Analysis:**

The hexagonal plot shows a strong positive correlation between price and square footage, with highest density around $400K-500K for 2,000-2,500 sqft properties. Properties range up to 4,000 sqft and $1M, confirming larger homes command higher prices.



* **Multivariate Analysis:**

Multivariate analysis (MVA) is a set of statistical methods used to analyze data sets with multiple variables. It helps in understanding the relationships and structures within the data, particularly when variables are correlated. MVA is crucial for exploring complex data, identifying patterns, and making informed decisions in various fields.

 A heatmap is a data visualization technique that represents data values as colors in a matrix format. It uses a color scale to indicate the magnitude or intensity of data points, allowing for quick identification of patterns, trends, and areas of high or low concentration within a dataset.

**10.OVERALL INSIGHTS FROM ANALYSIS**

* Larger homes (higher sqft, more bedrooms) command higher prices.
* Prices differ based on city. May be high population in those regions.
* Clusters of high prices in cities like **Seattle (WA), San Francisco (CA), or New York (NY)** suggest strong demand, luxury markets, or economic growth.
* **waterfront properties** (e.g., Miami, Malibu) priced higher than inland homes.
* in south USA, house average price is very low. May be in those regions economic and populations differs compare to north regions.
* Renovated homes are **~9.7% more expensive** on average than unrenovated homes. This shows that renovations generally add value, but the premium is moderate.

**11.CONCLUSION:**

The exploratory data analysis (EDA) of the housing dataset provided valuable insights into the factors influencing residential property prices. Through systematic data cleaning, feature engineering, and statistical analysis, we uncovered key patterns and relationships that can guide real estate stakeholders in decision-making.