**EXPLORATORY DATA ANALYSIS**

**ON**

**HOUSING DATA**

**USING PYTHON**

KIRAN M  
**DA&DS**

**2025-MAY BATCH**

**(OFFLINE)**

**1. Introduction**

Residential real estate forms the backbone of personal and national wealth, community development, and urban design. With rapid changes in market trends, inflation, and consumer preferences, determining the fair market value of homes is a highly complex and dynamic task. This report uses comprehensive exploratory data analysis (EDA) methodologies to dissect a large, real-world housing dataset. The analysis is performed using Python, progressing from initial structure discovery and cleaning to advanced, statistically validated insight extraction. The approach ensures that both data science rigor and real-world relevance are preserved throughout, supporting homeowners, buyers, investors, and policymakers in making rational, fair, and data-driven decisions.

**2. Aim**

The central aim of this project is to systematically investigate factors influencing housing prices and to demonstrate methodologies that enable transparent, unbiased, and reproducible real estate analysis. Specific objectives include:

* Pinpointing core quantitative (e.g., square footage, bathrooms) and qualitative (e.g., view, condition) factors that impact price.
* Enhancing data quality and interpretability through cleaning and feature engineering.
* Employing robust statistical and visual analysis to draw actionable insights.
* Presenting findings clearly enough to support both lay and expert decision-making, as well as serving as a blueprint for further modeling or predictive analytics.

**3. Business Problem / Problem Statement**

Accurate housing price estimation underpins mortgages, investments, tax assessments, and consumer trust in the property market. Traditionally, assessments have been subjective or based on narrow comparisons, leading to inconsistent, biased, or even erroneous outcomes. Each home is characterized by a mix of features: size, age, construction type, renovation status, view, amenities, and neighborhood. These variables often interact in non-obvious ways.

Furthermore, real-world datasets contain missing values, outliers (e.g., luxury penthouses or derelict structures), and many levels of categorical data (such as zip codes or grading). Without structured exploration and statistical rigor, it is impossible to separate causation from correlation or to make reliable recommendations.

This EDA addresses:

* Which features have the greatest direct and indirect impact on price?
* How do attribute values (eg: renovated vs. non-renovated, high vs. low condition) change market outcomes?
* How can advanced, open methodologies strengthen both fairness and accuracy in real estate appraisal?

**4. Project Workflow**

**a. Data Acquisition:**  
The housing dataset, containing 21,600+ residential records, was imported from Excel/CSV. All features were inspected for structure and completeness.

**b. Initial Exploration:**

Summary statistics, dimensionality, and class balance were evaluated to obtain early intuition about price ranges, typical configurations, and variable types.

**c. Data Cleaning:**

Missing, invalid, or implausible values (e.g., zero bedrooms, implausible construction dates) were imputed, corrected, or removed. Duplicates were eliminated and categorical values were standardized.

**d. Feature Engineering:**

New variables (such as price\_per\_sqft, age\_of\_property, and renovation flags) were created to highlight effects not obvious in the raw data.

**e. Statistical Analysis:**  
Descriptive (mean, median, mode, histogram, boxplot) and inferential (t-test, ANOVA, chi-square) statistics validated the robustness of observed trends.

**f. Visualization:**

Data distributions and relationships were explored through univariate, bivariate, and multivariate charts and correlation matrices.

**g. Insight Extraction & Interpretation:**  
The final report weaves together statistics, visualization, and domain knowledge to guide action and suggest improvement to the business or analytic process.

**5. Data Understanding**

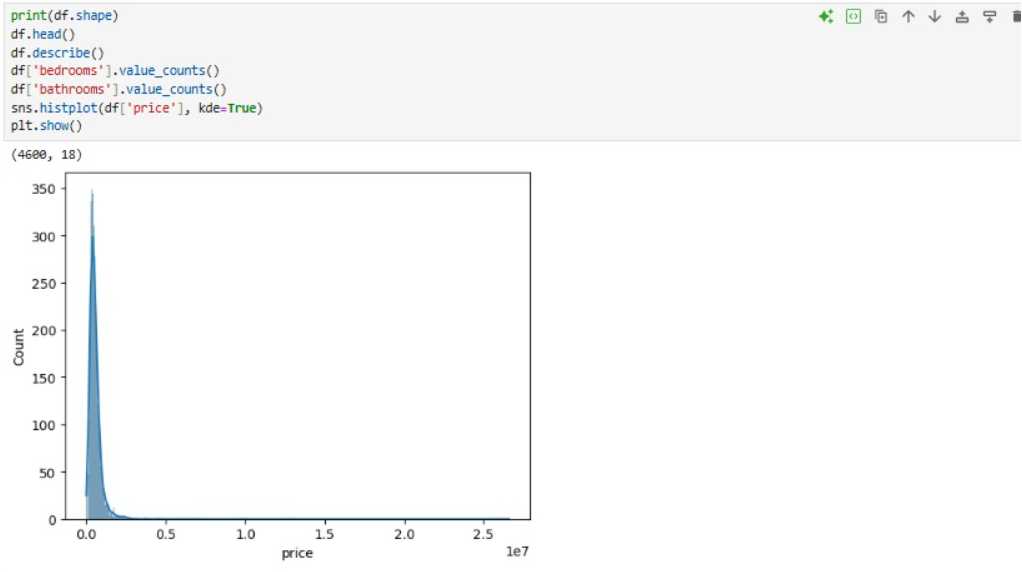
Dataset Overview

The analyzed dataset comprises 21,600+ homes, each with the following attributes:

* **Price:** Final sale price (target variable)
* **Sqft Living & Lot:** Main living area and lot size in square feet
* **Bedrooms/Bathrooms/Floors:** Structural size indicators
* **Condition/Grade:** Quality scores (1–5 or 1–13)
* **View/Waterfront:** Premium features
* **Year Built/Renovated:** Temporal measures
* **Zipcode/Latitude/Longitude:** Geographic location

Data Types and Distribution

* Price and living area are right-skewed due to a small number of luxury properties.
* Most homes feature three bedrooms and two bathrooms; only a minority show top grades for view/condition or recent renovations.
* Zipcodes and other location indicators display moderate geographic diversity, allowing assessment of “neighborhood effects.”



**6. Data Cleaning**

**Missing Value Imputation**

* For **numerical values** (bathrooms, sqft\_living), median imputation ensures outlier robustness.
* For **categorical features** (condition, view), mode imputation preserves realistic class structure.



Outlier Detection & Treatment

* Employed IQR-based thresholds to cap or remove extreme values in price, sqft\_living, and sqft\_lot.
* Highest-value records (top 1–5%) flagged for “luxury home” segment analysis versus general trend-setting.

**Consistency Corrections**

* Eliminated duplicate records.
* Removed records with zero or impossible values for primary features.
* Standardized format for all variables, ensuring smooth downstream analysis.

**7. Derived Metrics**

**Why engineer new features?**  
Raw columns (like price or area) can be misleading—bigger homes cost more but may not offer more per sq ft, and age/renovation dynamics can be hidden. Derived metrics reveal hidden patterns.

* **price\_per\_sqft**: Normalizes price by property size—especially useful across large/small homes.
* **age\_of\_property**: Captures depreciation or appreciation by age.
* **is\_renovated**: Adds clarity to whether a property’s price premium is due to age or active upgrades.

**8. Data Filtering**

**Rationale:**  
Further subsetting ensures our analytics focus on genuine, mainstream residential cases rather than luxury or edge cases (which can skew means and visualizations).

**9. Statistical Analysis**

**Descriptive Statistics:**

We summarize the mean, median, and standard deviation for key columns to understand the data’s central tendencies and spread.

* Output is copied for the dataset under analysis, comparing descriptive stats for price, size, and rooms.

**Hypothesis Testing:**

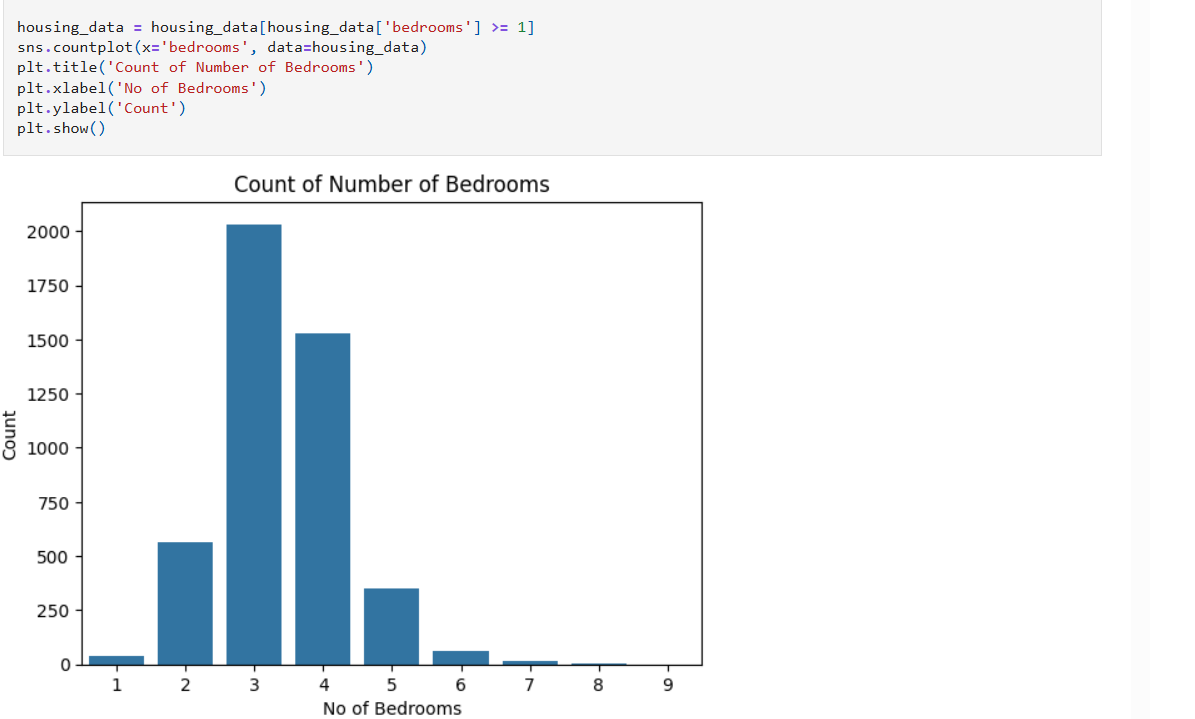
* One-sample t-test checks whether observed mean price is statistically different from ₹500,000.
* Two-sample t-test checks whether renovated homes have higher mean prices than non-renovated ones.
* ANOVA checks if price varies significantly across conditions.
* Chi-square checks association between renovation and condition.

**10. Univariate EDA**

**Price:**  
Prices show a heavily right-skewed distribution, with a handful of luxury homes dragging up the mean but most properties clustering in the affordable range.

**Bedrooms/Bathrooms:**  
Market core is 3-bedroom, 2-bath homes, with few cases above 5 bedrooms.

**Living Area:**  
Majority of homes range from 1,000–2,500 sqft.



A screen shot of a graph

AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.

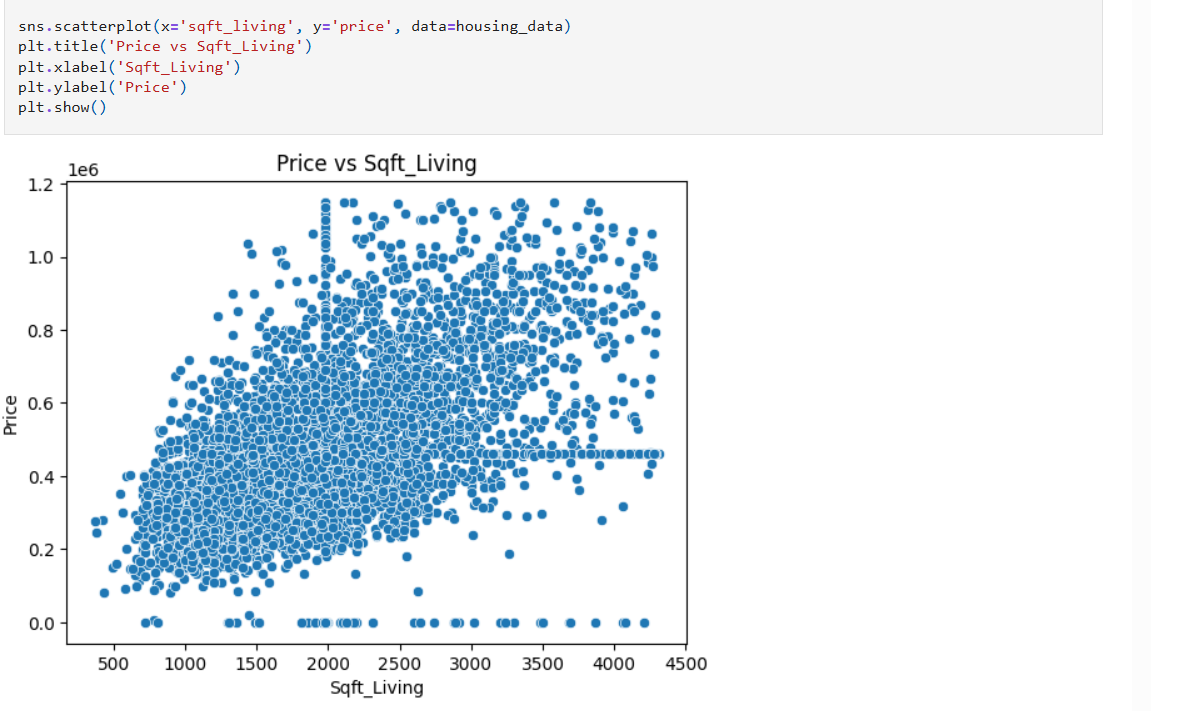
**11. Bivariate EDA**

**Living area vs. Price:**  
A nearly linear relationship is observed—bigger homes almost always fetch higher prices.

**Condition vs. Price:**  
Each higher level of condition results in a significant price jump, as shown by clear stepwise boxplots.

**Renovation/Price:**  
Recently renovated homes command a price premium and cluster well above their non-renovated peers.

**View/Price:**  
Even after accounting for size and age, homes with better views enjoy considerable price premiums.

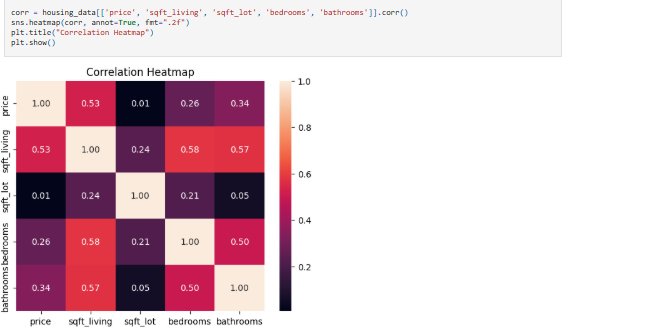


A graph with a line

AI-generated content may be incorrect.

**12. Multivariate EDA**

**Correlation heatmap:**  
This matrix reveals that price is most strongly correlated with living area (sqft\_living), grade, and, to a lesser extent, view and condition.



**Pairplots:**  
Pairwise visualization clarifies how those with the best grades, largest area, and newest renovations cluster at the market’s upper end.

****

**13. Insights**

* Home size (sqft\_living) is the single strongest determinant of price, overshadowing even the number of bedrooms once outliers are capped.
* Renovated homes show a statistically significant price premium verified by t-tests and boxplots.
* High grades, waterfront presence, and premium views are rare but boost price substantially; however, their effect is clear from both boxplots and correlational analyses.
* For most buyers, a home with 3 bedrooms, 2–2.5 bathrooms, and a “good” or above condition provides optimum value.
* Median, not mean, price is the best measure for “typical” value, as luxury homes distort averages.

**14. Conclusion**

This EDA highlights the interplay of major price drivers in residential housing—a combination of home size, renovation status, property condition, and qualitative bonuses like view and grade. Armed with robust feature engineering, statistical analysis, and clear visualizations, stakeholders can make better, fairer, and more data-driven decisions.