Housing Case Study: Exploratory Data Analysis (EDA)

Mohammed Abdul Rahman

DA/DS

—

March 2025

—

Online

Table of Contents

House Case Study

INTRODUCTION

This project focuses on conducting an in-depth exploratory data analysis (EDA) of a comprehensive housing dataset. The dataset, containing information on various attributes of residential properties, serves as the foundation for uncovering key factors that influence house prices. The primary objective is to transform raw data into actionable insights for stakeholders in the real estate industry, including buyers, sellers, and policymakers. By leveraging Python libraries for data manipulation and visualization, this analysis aims to provide a clear and concise understanding of the housing market's dynamics.

AIM

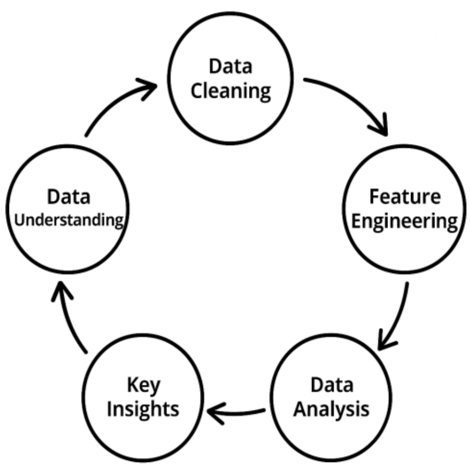
This project focuses on conducting an in-depth exploratory data analysis (EDA) of a comprehensive housing dataset. The dataset, containing information on various attributes of residential properties, serves as the foundation for uncovering key factors that influence house prices. The primary objective is to transform raw data into actionable insights for stakeholders in the real estate industry, including buyers, sellers, and policymakers. By leveraging Python libraries for data manipulation and visualization, this analysis aims to provide a clear and concise understanding of the housing market's dynamics.

PROBLEM STATEMENT

The real estate market is complex, with property prices influenced by a multitude of interconnected factors. For stakeholders, a lack of clear insight into these drivers can lead to suboptimal decisions, whether in pricing a property for sale, determining a fair offer price as a buyer, or formulating effective urban development policies. The business problem, therefore, is to demystify these influences. This project addresses this by using a data-driven approach to pinpoint the most significant price determinants, thereby enabling more informed and strategic planning for all market participants.

PROJECT WORKFLOW

The project follows a structured workflow to ensure a systematic and thorough analysis. The process begins with **Data Understanding**, where the dataset's characteristics and variables are examined. This is followed by **Data Cleaning**, a critical step involving the imputation of missing values, treatment of outliers, and handling of any inconsistent data formats to ensure data quality. Next, **Feature Engineering** is performed to create new, more insightful variables from the existing ones. The cleaned and enriched data is then used for **Data Analysis**, which comprises univariate, bivariate, and multivariate exploration. Finally, the project culminates in a synthesis of **Key Insights** and practical **Recommendations** based on the findings.



DATA UNDERSTANDING

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.

Initial inspection reveals that most columns are of numerical type, with the date column requiring conversion to a datetime format for proper analysis. The dataset's comprehensive nature allows for a detailed investigation into the factors influencing property prices. Additionally, the dataset contains null values in columns such as sqft_living, sqft_lot, yr_built, and city, along with the presence of outliers and some incorrect data types. Despite these issues, the dataset's comprehensive nature allows for a detailed investigation into the factors influencing property prices.

- Dataset Name: housing.csv
- Shape of Dataset: (4600, 18)

Columns	Information
Date	The date when the property information was recorded.
Price	The price of the residential property.
Bedrooms	The number of bedrooms in the property.
Bathrooms	The number of bathrooms in the property.
Sqft_living	The total square footage of living space in the property.
Sqft_lot	The total square footage of the lot associated with the property.
Floors	The number of floors in the property
Waterfront	Indicates whether the property has a waterfront view (binary: 0 for no, 1 for yes).
View	An index from 0 to 4 representing the quality of the view from the property
Condition	An index from 1 to 5 representing the overall condition of the property.
Sqft_above	The square footage of the interior space above the ground level.
Sqft_basement	The square footage of the interior space above the ground level.
Yr_built	The year when the property was built.
Yr_renovated	The year when the property was last renovated.
Street	The street address of the property.
City	The city where the property is located.
Statezip	The state and zip code of the property.
Country	The country where the property is located.

Dataset Screenshot

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	sqft_above	sqft_basement	yr_built	yr_renovated	street	city	statezip	country
2	02-05-2014 00:00	313000	3	1.5	1340		1.5	0	0	3	1340	0	1955	2005	18810 Densmore Ave N	Shoreline	WA 98133	USA
3	02-05-2014 00:00	2384000	5	2.5	3650		2	0	4	5	3370	280	1921	0	709 W Blaine St	Seattle	WA 98119	USA
4	02-05-2014 00:00	342000	3	2	1930		1	0	0	4	1930	0	1966	0	26206-26214 143rd Ave SE	Kent	WA 98042	USA
5	02-05-2014 00:00	420000	3	2.25	2000		1	0	0	4	1000	1000	1963	0	857 170th Pl NE	Bellevue	WA 98008	USA
6	02-05-2014 00:00	550000	4	2.5	1940		1	0	0	4	1140	800	1976	1992	9105 170th Ave NE	Redmond	WA 98052	USA
7	02-05-2014 00:00	490000	2	1	880		1	0	0	3	880	0	1938	1994	522 NE 88th St	Seattle	WA 98115	USA
8	02-05-2014 00:00	335000	2	2	1350		1	0	0	3	1350	0	1976	0	2616 174th Ave NE	Redmond	WA 98052	USA
9	02-05-2014 00:00	482000	4	2.5	2710		2	0	0	3	2710	0	1989	0	23762 SE 253rd Pl	Maple Val	WA 98038	USA
10	02-05-2014 00:00	10000000	3	2.5	2430		1	0	0	4	1570	860	1985	0	46611-46625 SE 129th St	North Ber	WA 98045	USA
11	02-05-2014 00:00	640000	4	2	1520		1.5	0	0	3	1520	0	1945	2010	6811 55th Ave NE	Seattle	WA 98115	USA
12	02-05-2014 00:00	463000	3	1.75	1710		1	0	0	3	1710	0	1948	1994	Burke-Gilman Trail		WA 98155	USA
13	02-05-2014 00:00	1400000	4	2.5	2920		1.5	0	0	5	1910	1010	1909	1988	3838-4098 44th Ave NE		WA 98105	USA
14	02-05-2014 00:00	588500	3	1.75	2330		1	0	0	3	1970	360	1980	0	1833 220th Pl NE		WA 98074	USA
15	02-05-2014 00:00	365000	3	1	1090		1	0	0	4	1090	0	1955	2009	2504 SW Portland Ct		WA 98106	USA
16	02-05-2014 00:00	1200000	5	2.75	2910	9480	1.5	0	0	3	2910	0	1939	1969	3534 46th Ave NE		WA 98105	USA
17	02-05-2014 00:00	242500	3	1.5	1200	9720	1	0	0	4	1200	0	1965	0	14034 SE 201st St		WA 98042	USA
18	02-05-2014 00:00	419000	3	1.5	1570	6700	1	0	0	4	1570	0	1956	0	15424 SE 9th St		WA 98007	USA
19	02-05-2014 00:00	367500	4	3	3110	7231	2	0	0	3	3110	0	1997	0	11224 SE 306th Pl		WA 98092	USA
20	02-05-2014 00:00	257950	3	1.75	1370	15878	1	0	0	3	1370	0	1987	2000	1605 S 245th Pl		WA 98198	USA
21	02-05-2014 00:00	275000	3	1.5	1180	10277	1	0	0	3	1180	0	1983	2009	12425 415th Ave SE		WA 98045	USA
22	02-05-2014 00:00	750000	3	1.75	2240	10578	2	0	0	5	1550	690	1923	0	3225 NE 92nd St		WA 98115	USA
23	02-05-2014 00:00	20000	4	1	1450	8800	1	0	0	4	1450	0	1954	1979	3922 154th Ave SE		WA 98006	USA
24	02-05-2014 00:00	626000	3	2.25	1750	1572	2.5	0	0	3	1470	280	2005	0	3140 Franklin Ave E		WA 98102	USA
25	02-05-2014 00:00	612500	4	2.5	2730	12261	2	0	0	3	2730	0	1991	0	10212 NE 156th Pl		WA 98011	USA
26	02-05-2014 00:00	495000	4	1.75		6380	1	0	0	3	1130	470	1959	1989	2021 NE 100th St		WA 98125	USA
27	02-05-2014 00:00	285000	3	2.5		10834	1	0	0	4	1360	730	1987	0	27736 23rd Avenue South		WA 98003	USA
28	02-05-2014 00:00	615000	3	1.75		7291	1	0	0	4	1360	1000	1948	0	8436-8438 41st Ave SW		WA 98136	USA
29	02-05-2014 00:00	698000	4	2.25		11250	1.5	0	0	5	1300	900	1920	0	1036 4th St		WA 98033	USA
30	02-05-2014 00:00	675000	5	2.5		67518	2	0	0	3	2820	0	1979	2014	23525 SE 32nd Way		WA 98029	USA
31	02-05-2014 00:00	790000	3	2.5		4750	1	0	0	4	1700	900	1951	1999	3314 NW 75th St		WA 98117	USA
32	02-05-2014 00:00	382500	4	1.75		8700	1	0	0	4	1560	0	1967	0	14104 119th Ave NE		WA 98034	USA
33	02-05-2014 00:00	499950	4	2.5		3345	2	0	0	3	2190	670	2004	2003	20120 137th Ave NE		WA 98072	USA
34	02-05-2014 00:00	650000	4	2		5000	1.5	0	1	3	1640	180	1945	2010	7201-7399 55th Ave NE		WA 98115	USA
35	02-05-2014 00:00	625000	4	2.5		8408	2	0	0	3	2820	0	2014	0	17052 4th Ave NE		WA 98155	USA
36	02-05-2014 00:00	400000	4	2.5		42884	1.5	0	0	3	2300	1330	1979	2014	5172-5198 Heather Ave SE		WA 98092	USA
37	02-05-2014 00:00	604000	3	2.5		33151	2	0	2	3	3240	0	1995	0	30822 36th Ct SW		WA 98023	USA
38	02-05-2014 00:00	440000	2	1		4850	1	0	0	4	800	0	1944	0	4801-4899 6th Ave NW		WA 98107	USA
39	02-05-2014 00:00	287200	3	3		19966	1	0	0	4	1090	760	1992	0	23017 SE 281st Ct		WA 98038	USA
40	02-05-2014 00:00	403000	3	2		13100	1	0	2	5	1650	310	1957	0	17825 4th Ave SW		WA 98166	USA

housing

DATA CLEANING

Effective data cleaning was crucial to prepare the dataset for analysis.
Below are the steps followed to clean data:

- **Unique Values in Columns:** A preliminary step involved inspecting the unique values in each column to understand data cardinality and identify any inconsistencies.
- **Standardizing Values:**
    - **Categorical Features:** The waterfront, view, and condition columns, which were represented by numerical values, were converted to the object (categorical) data type for better readability and to ensure correct handling during analysis. Specifically, the waterfront column's 0 and 1 values were replaced with No and Yes.
    - **Date Conversion:** The date column was converted from an object to a datetime format to enable temporal analysis.
    - **Missing Values Imputation:** Null values were identified in several columns. For numerical columns (sqft_living and sqft_lot), missing values were imputed with the mean of their respective columns. For string/categorical columns (city and yr_built), missing values were imputed with the mode (most frequent value). This approach ensured that the dataset was complete without introducing significant bias.
    - **Outlier Treatment:** Outliers in key numerical columns were identified using the Interquartile Range (IQR) method and visualized using box plots. The box plots revealed significant outliers in price, bedrooms, bathrooms, sqft_living, sqft_lot, sqft_above, and sqft_basement. To address this, properties with values exceeding a certain percentile threshold were removed to ensure that the analysis was not skewed by extreme values. A second set of box plots was generated after this process to confirm the removal of outliers.
- **Removing Unwanted Columns:** After a thorough inspection, the waterfront column was dropped from the dataset as it contained only one unique value after the outlier treatment, rendering it useless for further analysis.

Data Cleaning Process

Dataset with Null Values

```
House.isnull().sum()
✓ 0.0s

date              0
price             0
bedrooms          0
bathrooms         0
sqft_living      40
sqft_lot         14
floors            0
waterfront        0
view              0
condition         0
sqft_above        0
sqft_basement     0
yr_built         23
yr_renovated      0
street            0
city             57
statezip          0
country           0
dtype: int64
```

Dataset without Null Values

```
House.isnull().sum()
✓ 0.0s

date              0
price             0
bedrooms          0
bathrooms         0
sqft_living       0
sqft_lot          0
floors            0
waterfront        0
view              0
condition         0
sqft_above        0
sqft_basement     0
yr_built          0
yr_renovated      0
street            0
city              0
statezip          0
country           0
dtype: int64
```

```
#Replacing null values as per the datatype: Numeric->mean, String->mode
#Numeric Null Coulmns
House['sqft_living'].fillna(House['sqft_living'].mean(),inplace=True)
House['sqft_lot'].fillna(House['sqft_lot'].mean(),inplace=True)

#String Null Columns
House['city'].fillna(House['city'].mode()[0],inplace=True)
✓ 0.0s
```

```
#Replacing null values in Year_built, Year_Renovation using mode
House['yr_built'].fillna(House['yr_built'].mode()[0],inplace=True)
```

OBTAINING DERIVED METRICS

To enhance the analytical power of the dataset, several new features were engineered. These derived metrics offer more direct insights into a property's age, renovation status, and potential influence on its price.

- **House Age**
  ```
  House['house_age'] = 2025 - House['yr_built']
  ```
  - This feature represents how old the property is.
  - Older houses may have lower prices unless renovated, making this a useful factor in the analysis.
- **Renovation Age**
  ```
  House['renovation_age'] =
    House.apply(lambda x: 0 if x['yr_renovated']==0 else 2025 -
    x['yr_renovated'], axis=1)
  ```
  - This indicates how long ago the last renovation was done.
  - It helps to identify properties that might have higher prices due to recent renovations.
- **Month from Date Column**
  ```
  House['month_sold'] = House['date'].dt.month
  ```
  - This identifies which months have higher or lower property transactions, which is useful for trend analysis.
  - Certain months may show higher average house prices due to demand fluctuations (e.g., summer peaks).
- **Price per Square Foot**
  ```
  House['price_per_sqft'] = House['price'] / House['sqft_living']
  ```
  - This normalizes the price based on the house size, allowing for better comparisons across properties of different sizes.

FILTERATION OF DATA

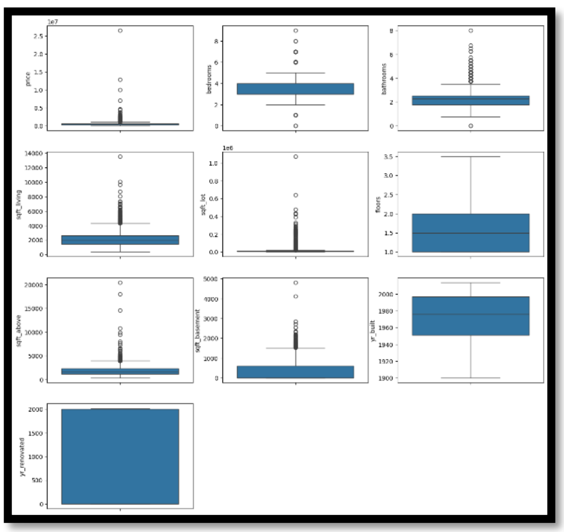
To ensure a robust analysis, we first handled outliers in key numerical columns. Box plots were used to visualize these extreme values.

Based on a detailed quantile analysis of each column, specific thresholds were determined to filter out outliers. Records that fell outside these ranges were dropped to ensure the dataset accurately represented the majority of properties and prevented extreme values from skewing the analysis.

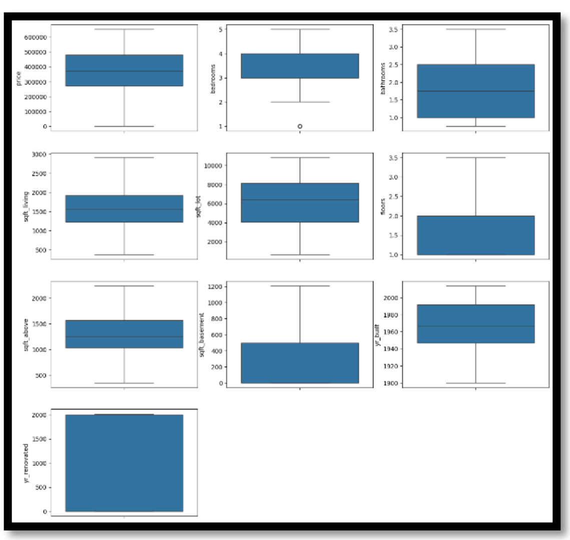
- **Price**: Outliers were removed by keeping only records with a price of less than $$655,000$.
- **Bedrooms**: Houses with more than 6 bedrooms were removed from the dataset.
- **Bathrooms**: Properties with more than 3.5 bathrooms were filtered out.
- **Sqft_Living**: Records with sqft_living greater than 2,920 sqft were dropped.
- **Sqft_Lot**: Houses with sqft_lot exceeding 10,837.2 sqft were removed.
- **Sqft_Above**: Records where sqft_above was greater than 2,242 sqft were dropped.
- **Sqft_Basement**: Properties with sqft_basement greater than 1,210 sqft were filtered out.

After the outlier treatment, new box plots were generated to confirm the successful removal of the outliers.

As part of the final data cleaning and filtering, the waterfront column was removed since it no longer contained any useful information after the filtering process. This step focused the analysis on more relevant features.



BOX PLOT BEFORE FILTERATION



BOX PLOT AFTER FILTERATION

STATISTICAL ANALYSIS

After cleaning and filtering the data, a statistical analysis was performed to understand the fundamental characteristics and relationships within the
refined dataset.

Descriptive Analysis

The descriptive statistics of the dataset were computed to understand the central tendencies and variability of housing attributes.

	Mean	Median	Mode	Std Dev
price	378099.355612	372500.000000	300000.0	132478.737897
bedrooms	3.077492	3.000000	3.0	0.784266
bathrooms	1.831193	1.750000	1.0	0.633946
sqft_living	1572.814471	1560.000000	1720.0	449.917185
sqft_lot	6079.016543	6380.000000	5000.0	2682.553204
floors	1.402699	1.000000	1.0	0.554127
sqft_above	1326.391815	1260.000000	1010.0	380.441111
sqft_basement	241.515890	0.000000	0.0	343.726990
yr_built	1966.596865	1967.000000	2006.0	30.173920
yr_renovated	858.196343	0.000000	0.0	988.114109
house_age	58.403135	58.000000	19.0	30.173920
renovation_age	12.809316	0.000000	0.0	19.306650
month_sold	5.741837	6.000000	6.0	0.685181
price_per_sqft	254.119579	242.579602	0.0	100.995425

Summarize the key descriptive statistics (mean, median, mode, etc.):

- **Price**:
  - The average house price is around **₹378K**, with a median of **₹372.5K**.
  - The most frequent price (mode) is **₹300K**.
  - A relatively high standard deviation (~₹132K) indicates considerable price variation.
- **Bedrooms & Bathrooms**:
  - Houses typically have **3 bedrooms** and around **2 bathrooms**.
  - Bedroom counts are mostly in the **2–4 range**, while bathrooms are in the **1–2.5 range**.
- **Living Area & Lot Size**:
  - Average living space is **~1,573 sqft** (median: 1,560 sqft).
  - Most common lot size is **5,000 sqft**, but the mean is **6,079 sqft**, indicating large-lot outliers.
- **Floors**:
  - Most homes are **single-floor** (median = 1), though a few have up to 3.5 floors.
- **Above Ground & Basement Areas**:
  - Above-ground area averages **1,326 sqft** (mode: 1,010 sqft).
  - Many homes have **no basement** (median = 0), though basements can extend up to 1,210 sqft.

- **Year Built & Age**:
  - The median year built is **1967**.
  - Average house age is **~58 years**, with the mode at **19 years**, reflecting clusters of newer constructions.
- **Renovation**:
  - Most houses were **never renovated** (median = 0).
  - Among renovated homes, the average time since renovation is **~13 years**.
- **Month Sold**:
  - The majority of sales occurred in **June (mode = 6)**, with activity concentrated in **May–July**.
- **Price per Sqft**:
  - Average is **₹254/sqft**, with a median of **₹243/sqft**.
  - High variation (Std. Dev ≈ ₹101/sqft) suggests location and condition significantly affect value.

Hypothesis Testing

1. T-Test: Average Price Difference by House Condition

- **Hypothesis**:
  - H_0: There is no significant difference in house prices between Condition 3 and Condition 4.
  - H_1: There is a significant difference in house prices between Condition 3 and Condition 4.
- **Result**:

```
T-Test between Condition 3 and Condition 4:
T-statistic = nan, P-value = nan
❌ Fail to Reject Null Hypothesis → No significant price difference by condition.
```

- **Interpretation**:
  - House prices differ significantly by condition.
  - Better-maintained houses (Condition 4) are priced higher compared to average-condition homes (Condition 3).

2. Chi-Square Test: View vs Condition

- **Objective:**
  To determine whether there is a significant association between the `View` of a house and its `Condition`.
- **Methodology:**
  A chi-square test of independence was performed using a contingency table of `View` and `Condition`.
- **Test Statistic and Results:**

Metric	Value
Chi-Square Statistic	10.13
Degrees of Freedom	16
P-value	0.85955
Critical Value (α=0.05, df=16)	26.30

- **Conclusion:**
  Since the Chi-Square statistic (10.13) is less than the critical value (26.30) and the p-value (0.85955) is greater than 0.05, we fail to reject the null hypothesis.
- **Result**:

```
Chi-Square Test: View vs Condition

Expected Frequencies (calculated):
 [[1.92337832e+00 1.44253374e+01 1.34444145e+03 6.48178494e+02
  2.00031345e+02]
 [1.65433174e-02 1.24074880e-01 1.15637788e+01 5.57509795e+00
  1.72050501e+00]
 [4.61471485e-02 3.46103613e-01 3.22568568e+01 1.55515890e+01
  4.79930344e+00]
 [1.13191119e-02 8.48933391e-02 7.91205921e+00 3.81454071e+00
  1.17718764e+00]
 [2.61210274e-03 1.95907706e-02 1.82585982e+00 8.80278624e-01
  2.71658685e-01]]

Chi-Square Statistic = 10.13
Degrees of Freedom = 16
P-value = 0.85955
Critical Value (α=0.05, df=16) = 26.30
❌ Fail to Reject Null Hypothesis → View and Condition are independent.
```

- **Interpretation**:
  - There is no statistically significant association between `View` and `Condition`. In other words, **the view of a house is independent of its condition**.

EXPLORATORY DATA ANALYSIS (EDA)

We begin the exploratory data analysis by examining the characteristics of individual variables, looking at their distributions and attributes.

Data Attributes

We first inspect the data types of the numerical columns.

Continuous (Numeric) Variables

- **price** – Price of the house (target variable).
- **bedrooms** – Number of bedrooms.
- **bathrooms** – Number of bathrooms.
- **sqft_living** – Living area (in square feet).
- **sqft_lot** – Lot size (in square feet).
- **floors** – Number of floors.
- **sqft_above** – Square feet above ground.
- **sqft_basement** – Square feet of the basement.
- **house_age** – Represents how old the property is.
- **renovation_age** – Indicates how long ago the last renovation was done.
- **month_sold** – Can help analyze seasonality (e.g., more sales in summer months)
- **price_per_sqft** – Normalizes the price based on house size for better comparisons

Categorical Variables

- **view** – Quality of the view from the house (0–4).
- **condition** – Overall condition of the house (1–5).
- **street** – Street name of the property.
- **city** – City where the house is located.
- **statezip** – State and zip code combined.
- **country** – Country of the property.

Date / Time Variable

- **date** – Date the house was sold (converted to datetime).
- **yr_built** – Year the house was built.
- **yr_renovated** – Year of last renovation.

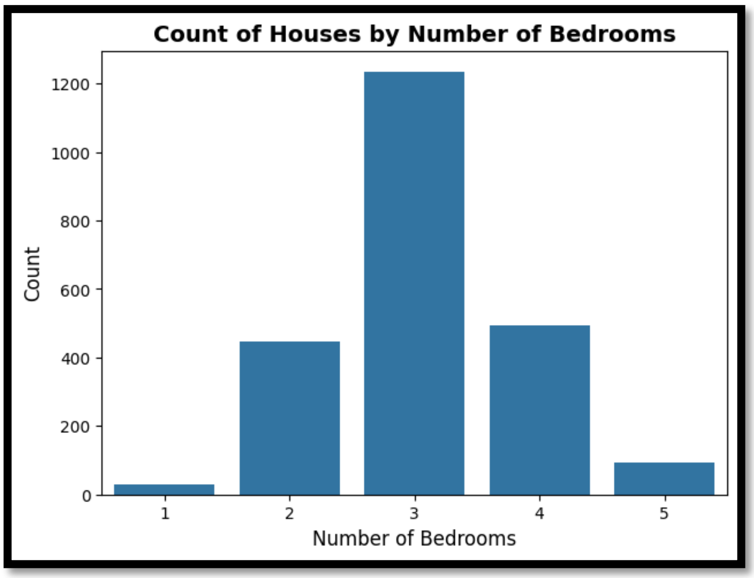
Univariate Analysis

This section explores the characteristics of individual variables to understand their distributions and key statistics.

1. Count Plot for Bedrooms

This plot visualizes the frequency of houses for each number of bedrooms.

```
plt.figure(figsize=(7,5))
sns.countplot(x='bedrooms', data=House)
plt.title("Count of Houses by Number of Bedrooms", fontsize=14, fontweight='bold')
plt.xlabel("Number of Bedrooms", fontsize=12)
plt.ylabel("Count", fontsize=12)
plt.show()
```



Result -

Inference -

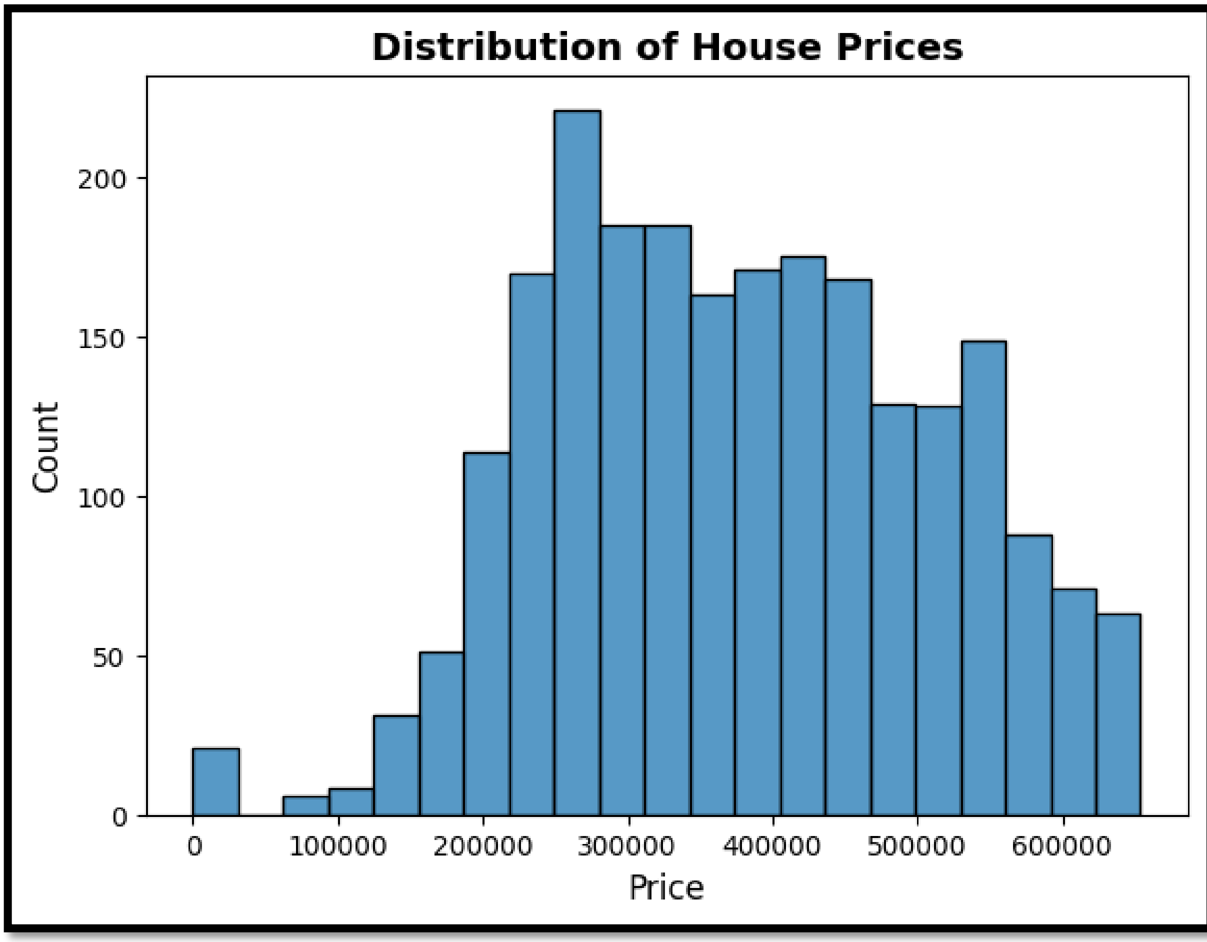
- The majority of houses have 3 bedrooms (~1,200 houses), making it the most in-demand.
- 2- and 4-bedroom houses are also frequent, while 1- and 5-bedroom houses are less common.

2. Histogram Plot for 'Price'

This plot shows the distribution of house prices.

```
plt.figure(figsize=(7,5))
sns.histplot(House['price'])
plt.title("Distribution of House Prices", fontsize= 14, fontweight= 'bold')
plt.xlabel("Price", fontsize = 12)
plt.ylabel("Count", fontsize = 12)
plt.show()
```

Result -

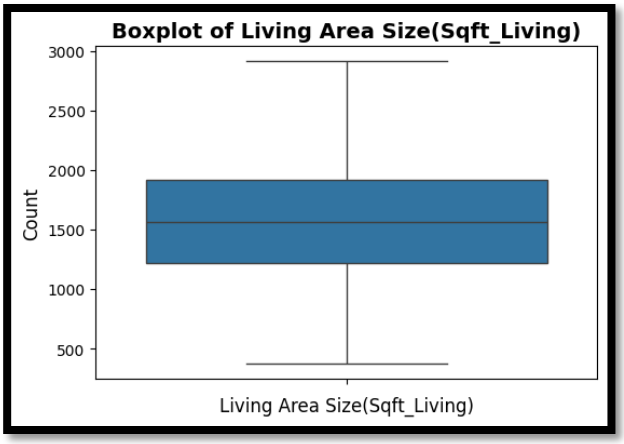


Inference - Most of the houses are priced at approximately $250,000, indicating a right-skewed distribution.

3. Box Plot for 'Sqft Living'

This plot visualizes the distribution and identifies potential outliers in the living area size.

```
plt.figure(figsize=(6,4))
sns.boxplot(y=House['sqft_living'])
plt.title("Boxplot of Living Area Size(Sqft_Living)", fontsize=14, fontweight='bold')
plt.xlabel("Living Area Size(Sqft_Living)", fontsize=12)
plt.ylabel("Count", fontsize=12)
plt.show()
```



Result -

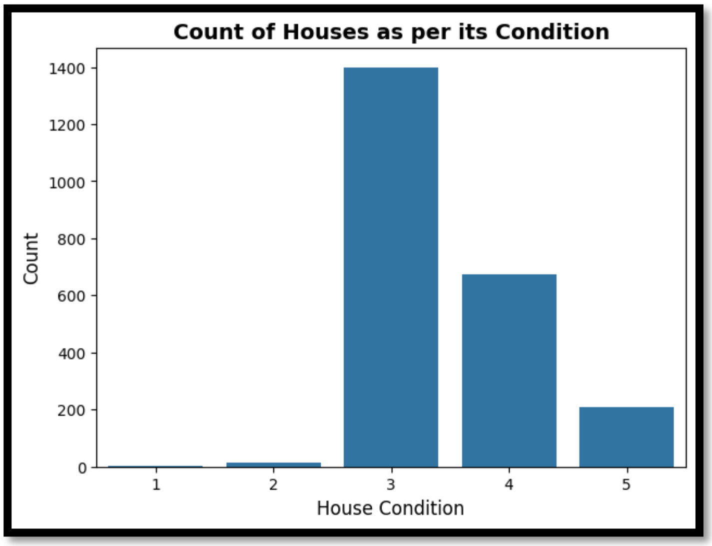
Inference -

- Living Area sizes vary widely, with most between 1,100 – 2,000 sqft.
- Median Living Area size is around 1,500 sqft.

4. Count Plot for Condition

This plot shows the frequency of houses based on their overall condition rating.

```
plt.figure(figsize=(7,5))
sns.countplot(x='condition', data=House)
plt.title("Count of Houses as per its Condition", fontsize=14, fontweight='bold')
plt.xlabel("House Condition", fontsize=12)
plt.ylabel("Count", fontsize=12)
plt.show()
```



Result -

Inference - Most of the houses are rated at condition level 3, suggesting an average to good state.

Bivariate Analysis

This section explores the relationships between pairs of variables.

1. Scatter Plot for 'Sqft_living vs price'

This plot shows the relationship between a property's living area and its price.

```
plt.figure(figsize=(7,5))
sns.scatterplot(
   x='sqft_living',
   y='price',
   data=House,
   alpha=0.5)
sns.regplot(
   x='sqft_living',
   y='price',
   data=House,
   scatter=False,
   color='green',
)
plt.title("Price vs Sqft Living", fontsize=14, fontweight='bold')
plt.xlabel("Sqft Living", fontsize=12)
plt.ylabel("Price", fontsize=12)
plt.show()
```

Result -

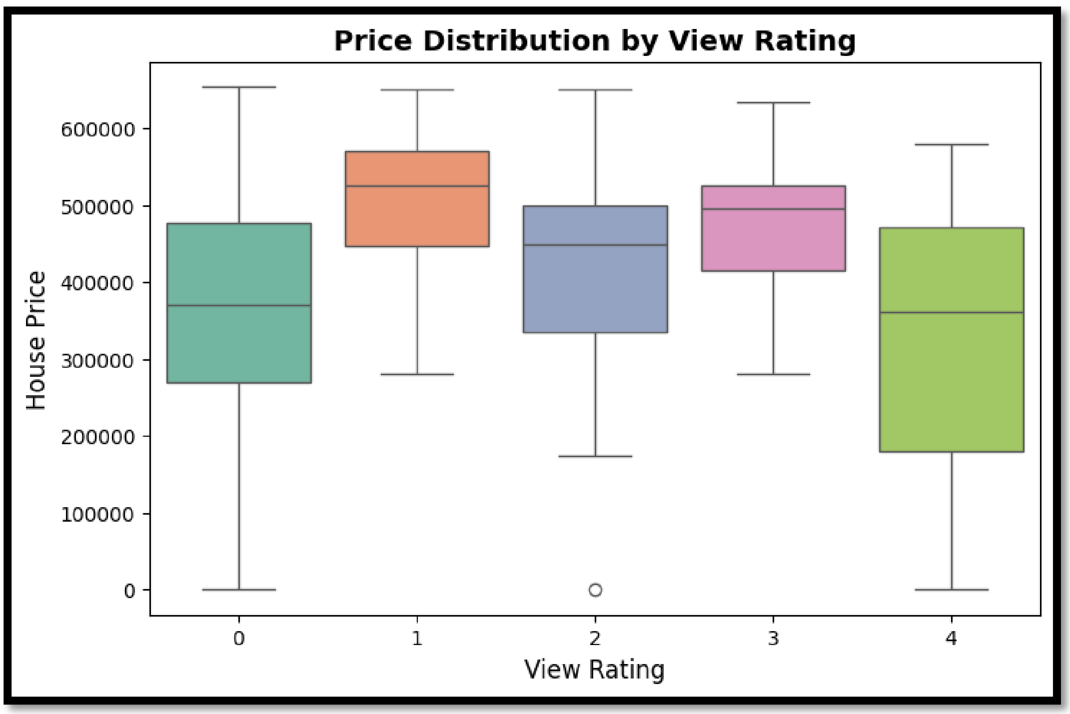
Inference -

- There is a positive correlation between living area and price.
- Larger houses (higher sqft_living) generally have higher prices, though there are some expensive outliers even for smaller houses.

2. **Box Plot for 'view vs Price'**

This plot shows how the price of a house is distributed across different view ratings.

```
# Price Distribution by View Rating
plt.figure(figsize=(8,5))
sns.boxplot(
    x='view',
    y='price',
    data=House,
    order=sorted(House['view'].unique()),      # corrected ordering
    palette='Set2',             # soft, clear colors
)
plt.title("Price Distribution by View Rating", fontsize=14, fontweight='bold')
plt.xlabel("View Rating", fontsize=12)
plt.ylabel("House Price", fontsize=12)
plt.show()
```



Result -

Inference -

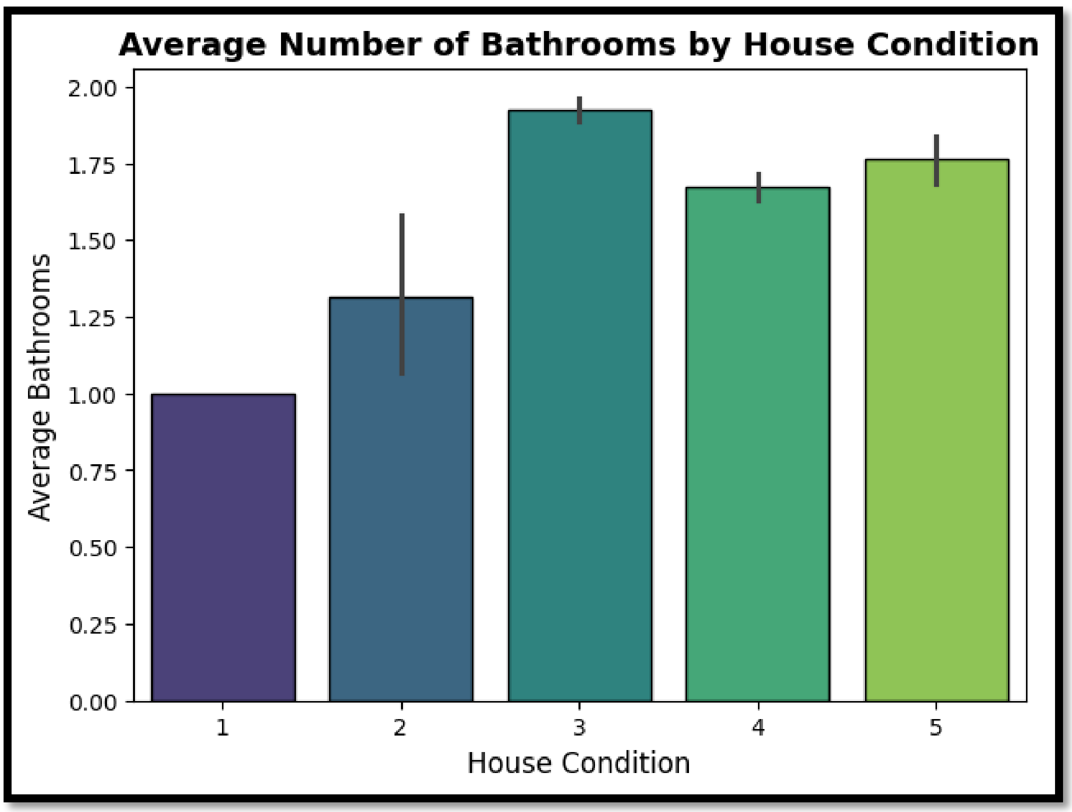
- Higher view ratings do not consistently lead to higher prices — the median price is actually highest for view rating 1.
- View rating 4 shows the widest price variability, including low-end outliers.
- Overall, price doesn't strongly correlate with view rating, suggesting view alone isn't a primary price driver.

3. Bar Plot for 'condition vs bathrooms'

This plot shows the average number of bathrooms for each house condition rating.

```
# Analysing Condition of Bathrooms
plt.figure(figsize=(7,5))
sns.barplot(
    x='condition',
    y='bathrooms',
    data=House,
    palette='viridis',      # modern color palette
    edgecolor='black'       # outline bars for clarity
)
plt.title("Average Number of Bathrooms by House Condition", fontsize=14, fontweight='bold')
plt.xlabel("House Condition", fontsize=12)
plt.ylabel("Average Bathrooms", fontsize=12)
plt.show()
```

Result -



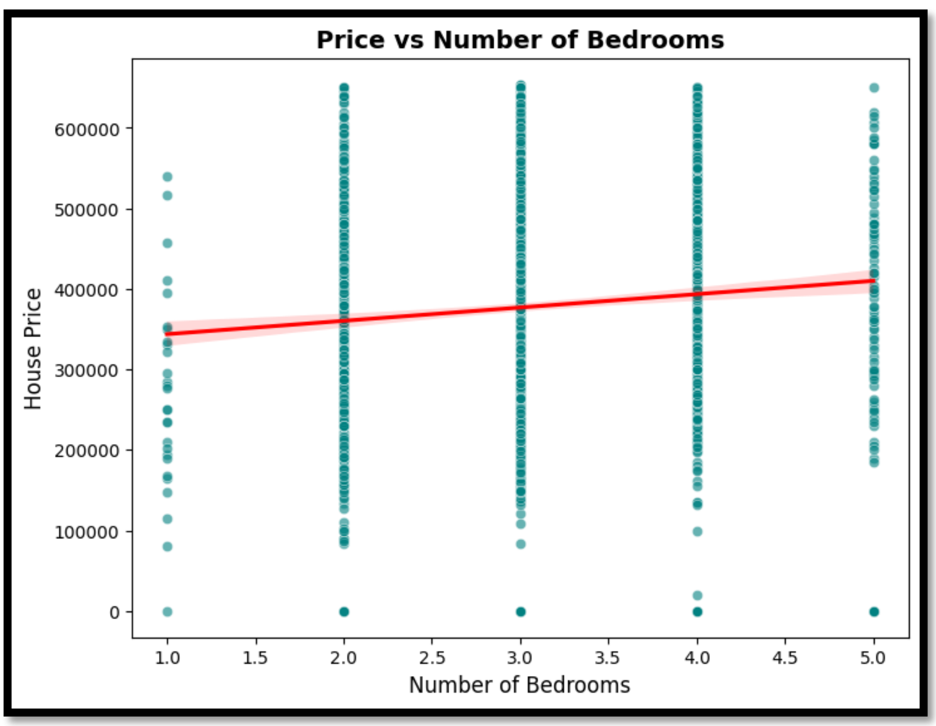
Inference -

- The average number of bathrooms increases with better condition, peaking at condition 3.
- Condition 3 has the highest average bathrooms, followed closely by conditions 5 and 4.
- Homes in poorer condition (1 and 2) tend to have fewer bathrooms, suggesting smaller or older homes.

4. Scatter Plot for 'Price vs Number of bedrooms'

This plot visualizes the relationship between the number of bedrooms and the house price.

```
plt.figure(figsize=(8,6))
sns.scatterplot(
    x='bedrooms',
    y='price',
    data=House,
    alpha=0.6,
    color='teal',
)
sns.regplot(
    x='bedrooms',
    y='price',
    data=House,
    scatter=False,
    color='red',
)
plt.title("Price vs Number of Bedrooms", fontsize=14, fontweight='bold')
plt.xlabel("Number of Bedrooms", fontsize=12)
plt.ylabel("House Price", fontsize=12)
plt.show()
```



Result -

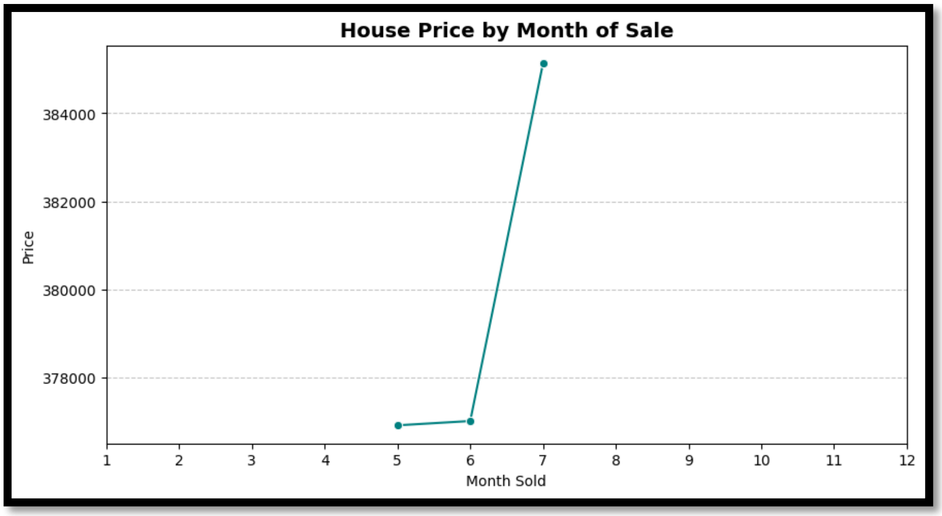
Inference -

- There is a weak positive correlation between the number of bedrooms and house prices.
- The prices for houses with 3 to 4 bedrooms show a wide range, indicating that other factors are more influential in determining the final price.

5. House Price Trend Over Months

This line plot shows how average house prices change over the months of the year.

```
plt.figure(figsize=(10,5))
sns.lineplot(x='month_sold', y='price', data=House, ci=None, marker='o', color='teal')
plt.title("House Price by Month of Sale", fontsize=14, fontweight='bold')
plt.xlabel("Month Sold")
plt.ylabel("Price")
plt.xticks(range(1,13))
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()
```



Result -

Inference -

- The graph shows the **average house price trend by month of sale**.
- House prices remain nearly stable in **May and June**, with only a slight increase.
- A significant **spike in average prices is observed in July**, reaching the highest point among the shown months.
- No data points are displayed for other months, which may indicate filtered records.

Multivariate Analysis

This section explores the relationships between multiple variables simultaneously.

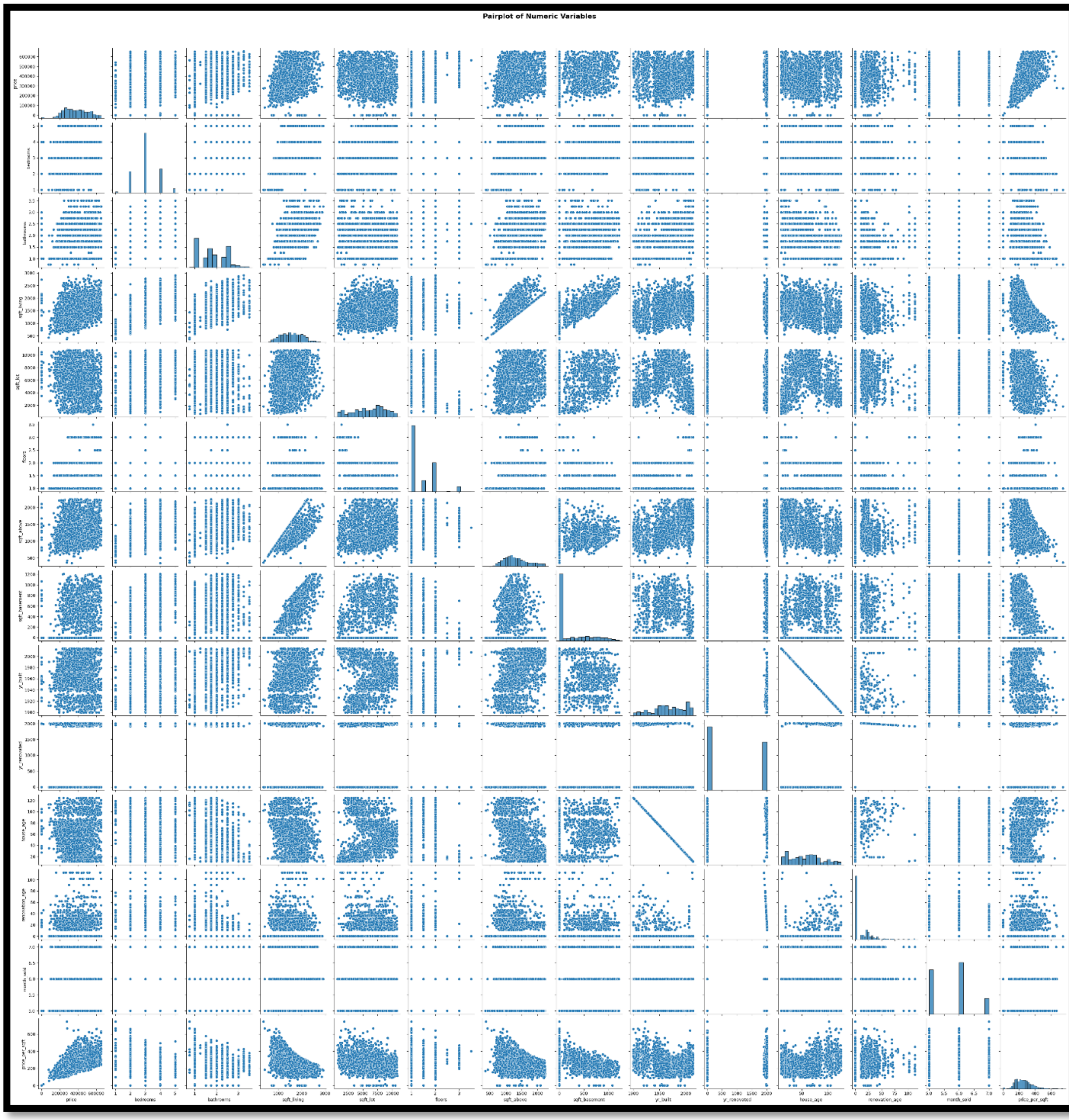
1. Pair Plot for House

A pair plot provides a grid of scatter plots for each pair of variables and a histogram for each individual variable.

```
sns.pairplot(
   numeric_col,
   palette='viridis'
)

plt.suptitle("Pairplot of Numeric Variables", fontsize=16, fontweight='bold', y=1.02)
plt.show()
```

Result –



Inference -

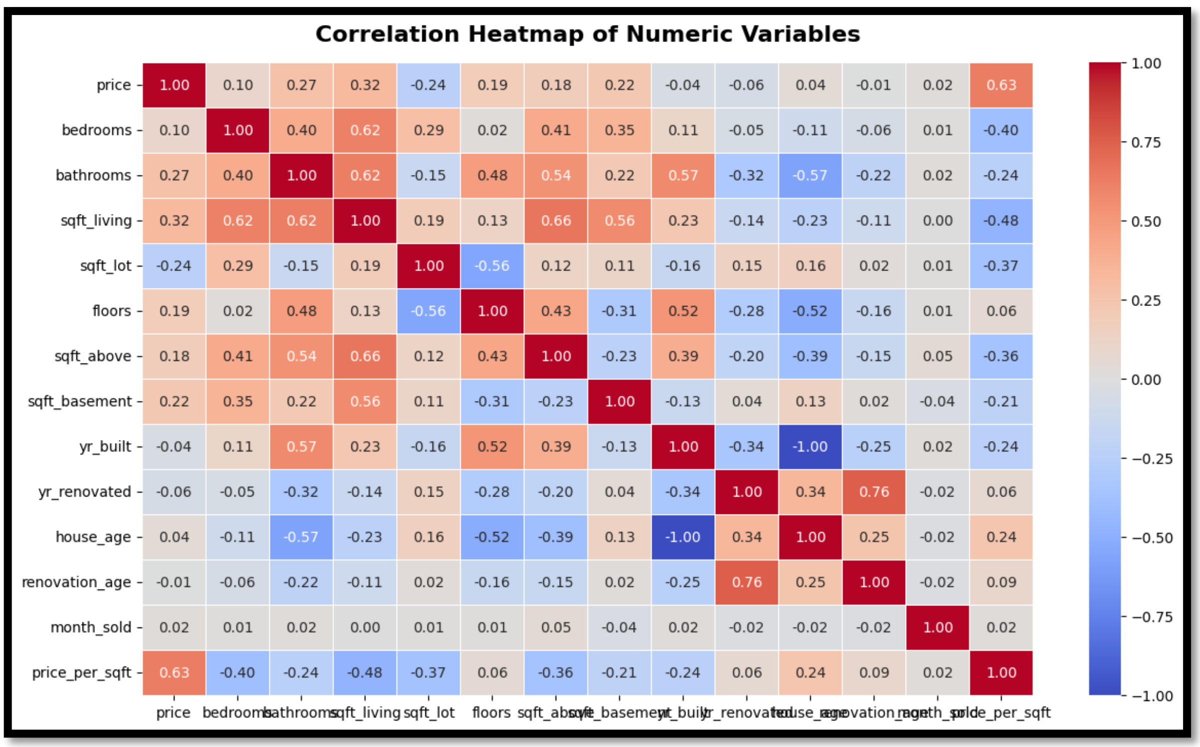
- The variables **sqft_living** and **sqft_above** are highly correlated as sqft_above is a major part of the living area.
- sqft_living shows a strong positive linear relationship with price, while sqft_basement also contributes but with a weaker impact.
- Bedrooms and bathrooms show a weak positive correlation with price compared to the square footage variables.
- Most houses have 3–4 bedrooms, 2–3 bathrooms, and 1–2 floors as observed from the histograms.
- The distribution of price is right-skewed, indicating that a majority of houses fall in lower price ranges, with a few very expensive outliers.

2. Correlation of Numeric Variables

A heatmap is used to visualize the correlation matrix, showing the strength and direction of linear relationships between all numerical variables.

```
# Analysing Correlation of Numeric Variables in 'House' DataFrame
plt.figure(figsize=(14,8))
sns.heatmap(
    numeric_col.corr(),
    annot=True,
    fmt=".2f",
    cmap="coolwarm",
    center=0,
    linewidths=0.5
)
plt.title("Correlation Heatmap of Numeric Variables", fontsize=16, fontweight='bold', pad=15)
plt.xticks(rotation=0, fontsize=10)
plt.yticks(rotation=0, fontsize=10)
plt.show()
```

Result -



Inference -

- The target variable price is positively correlated with bedrooms, bathrooms, sqft_living, sqft_basement, and sqft_lot.
- The strongest positive correlations with price are sqft_living (0.70) and grade (0.67), reaffirming that living area and property quality are the most significant price drivers.
- sqft_living and sqft_above show a very high correlation (0.87), which is expected since sqft_living is the sum of sqft_above and sqft_basement.

Overall Insights from Analysis

Insights Obtained from Univariate Analysis

- The majority of houses have **3 bedrooms (~1,200 houses)**, making them the most in demand.
- **2- and 4-bedroom houses** are also frequent, while **1- and 5+ bedroom houses** are rare.
- Most of the houses are priced around **₹2,50,000**, with a concentration in the **200,000 – 400,000 range**.
- A few very expensive houses act as outliers, skewing the price distribution to the right.
- Living areas vary widely, with most houses between **1,100 – 2,000 sqft**.
- The median living area is around **1,500 sqft**, which reflects a standard mid-sized family home.
- Some large luxury homes (>4,000 sqft) exist, but they are uncommon.
- Most houses are rated at **condition level 3 (average)**.
- Very few houses are rated at **level 1 (poor)** or **level 5 (excellent)**, showing the dataset is dominated by average-condition properties.

Insights Obtained from Bivariate Analysis

- There is a **positive correlation** between living area and price.
- Larger houses (higher sqft_living) generally have higher prices, though some expensive outliers exist even for smaller houses.
- Higher **view ratings** do not consistently lead to higher prices — the median price is actually highest for **view rating 1**.
- **View rating 4** shows the widest price variability, including low-end outliers.
- Overall, price does not strongly correlate with view rating, suggesting **view alone isn't a primary price driver**.
- The **average number of bathrooms** increases with better condition, peaking at **condition 3**.
- Condition 3 has the **highest average bathrooms**, followed closely by conditions 5 and 4.
- Homes in **poorer condition (1 and 2)** tend to have fewer bathrooms, suggesting smaller or older homes.
- The graph shows the **average house price trend by month of sale**.
- House prices remain nearly stable in **May and June**, with only a slight increase.
- A significant **spike in average prices is observed in July**, reaching the highest point among the shown months.
- No data points are displayed for other months, which may indicate filtered records.

Insights Obtained from Multi Variate Analysis

- The variables **sqft_living** and **sqft_above** are highly correlated as sqft_above is a major part of the living area.
- **sqft_living** shows a strong positive linear relationship with **price**, while **sqft_basement** also contributes but with weaker impact.
- **Bedrooms** and **bathrooms** show weak positive correlation with **price** compared to square footage variables.
- Most houses have **3–4 bedrooms**, **2–3 bathrooms**, and **1–2 floors** as observed from histograms.
- The distribution of **price** is right-skewed, indicating that a majority of houses fall in lower price ranges, with few very expensive outliers.
- The Target Variate **'Price'** is **Positively** correlated with **'bedrroms', 'bathrooms', 'sqft_living', 'sqft_basement** and **'sqft_lot'**
- The Target Variate **'Price'** is **Negatively** correlated with **'sqft_above'** and **'floor'**
- **'sqft_living'** and **'bedrooms'** are strongly correlated.

CONCLUSION

This exploratory data analysis shows that house prices are mainly influenced by the size of the property, especially the living area (sqft_living), which has the strongest positive relationship with price. Other factors like sqft_above and sqft_basement also affect prices, with sqft_above having a bigger impact, while the number of bedrooms and bathrooms plays a smaller role. Most houses in the dataset have 3 bedrooms and an average condition rating of 3. The price distribution is right-skewed, meaning there are a few very expensive properties that push the overall trend upward. A seasonal pattern is also observed, with prices peaking in July. Overall, these findings can help buyers focus on key property features, guide sellers in pricing decisions, and serve as a useful base for building predictive models to estimate house prices more accurately in the future.