Raw House Data Analysis

•••

By Ram Sunder on March 24, 2025

Table of Content

1. Overview
2. Libraries used for EDA
3. Understanding Colums
4. Potential Issues to Address
5. Final Classification
6. Challenges(Future Warning)
7. Data Visualization
8. Summary
9. Question and Answers

Overview

Initial Observations from the Dataset

- Total Records: 5,000 rows
- Columns: 16
- Data Types:
  - Numerical: MLS, sold_price, zipcode, longitude, latitude, lot_acres, taxes, year_built, bedrooms, fireplaces,bathrooms,garage,sqrt_ft
  - Categorical (Objects): kitchen_features, floor_covering, HOA
- Missing Values:
  - lot_acres (10 missing)
  - fireplaces (25 missing)
  - sqrt_ft, garage, and HOA have None values (Need conversion to NaN)

Technology used in EDA

- Google Collab
- Python

- Libraries
  - Pandas: Read, clean, and manipulate data
  - NumPy: Handle numerical operations
  - Matplotlib & Seaborn: Create plots

Understanding Dataset Columns

Column Name	Description
MLS	Listing ID
sold_price	Sale price of the house
zipcode	Location identifier
longitude/latitude	Geographic location
lot_acres	Land area in acres
taxes	Annual property tax
year_built	Year house was built
bedrooms	Number of bedrooms
bathrooms	Number of bathrooms

Column Name	Description
sqrt_ft	Total square footage
garage	Number of garage spaces
kitchen_features	Kitchen specifications
fireplaces	Number of fireplaces
floor_covering	Type of flooring
HOA	Homeowners Association fee

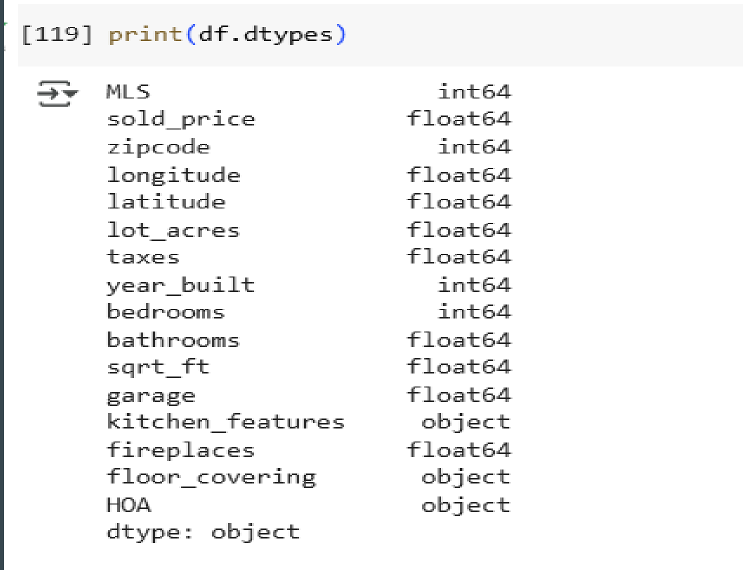
Potential Issues to Address

1. Bathrooms, Square Footage (sqrt_ft), and Garage columns are objects
   - Convert them to appropriate numerical types.
2. Missing Values Handling
   - lot_acres and fireplaces can be imputed or removed.
3. HOA contains None, which might indicate missing values
   - Need to check whether None means "No HOA" or a missing value.
4. Kitchen Features and Floor Covering are categorical
   - Need to encode or analyze further.

Potential Issues to Address

Missing Values Summary:

- lot_acres → 10 missing
- bathrooms → 6 missing
- sqrt_ft → 56 missing
- garage → 7 missing
- fireplaces → 25 missing
- HOA → 580 missing (possibly because many properties don't have HOA fees)



```
[119] print(df.dtypes)

MLS                     int64
sold_price            float64
zipcode                 int64
longitude             float64
latitude              float64
lot_acres             float64
taxes                 float64
year_built              int64
bedrooms                int64
bathrooms             float64
sqrt_ft               float64
garage                float64
kitchen_features       object
fireplaces            float64
floor_covering         object
HOA                    object
dtype: object
```

Potential Issues to Address

Handle missing values:

- lot_acres, bathrooms, sqrt_ft, garage, and fireplaces → Fill with median or mode.
- HOA → Treat missing values as zero (if None means no HOA).

Final Classification

Numerical (Continuous & Discrete)

- sold_price, longitude, latitude, lot_acres, taxes, year_built, bedrooms, fireplaces, bathrooms, sqrt_ft, garage, HOA

Categorical

- kitchen_features, floor_covering

Drop / Ignore

- MLS (just an ID, not useful for predictions)

Pandas/Seaborn Future Warning Fix

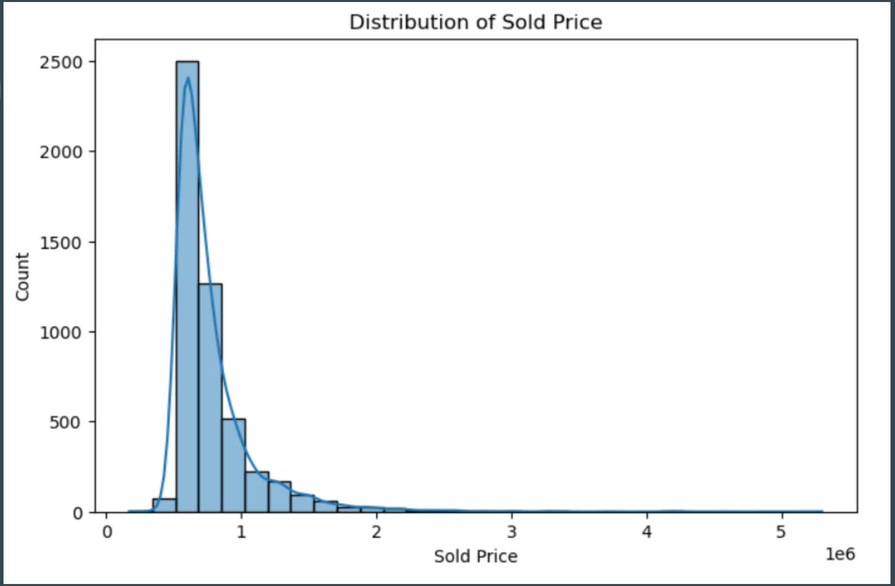
Instead of: df['lot_acres'].fillna(df['lot_acres'].median(), inplace=True)

Use: df['lot_acres'] = df['lot_acres'].fillna(df['lot_acres'].median())

Instead of: sns.kdeplot(df['sold_price'], shade= True, color= 'green')

Use: sns.kdeplot(df['sold_price'], shade= True, color= 'green')

Histogram (Frequency Plot)



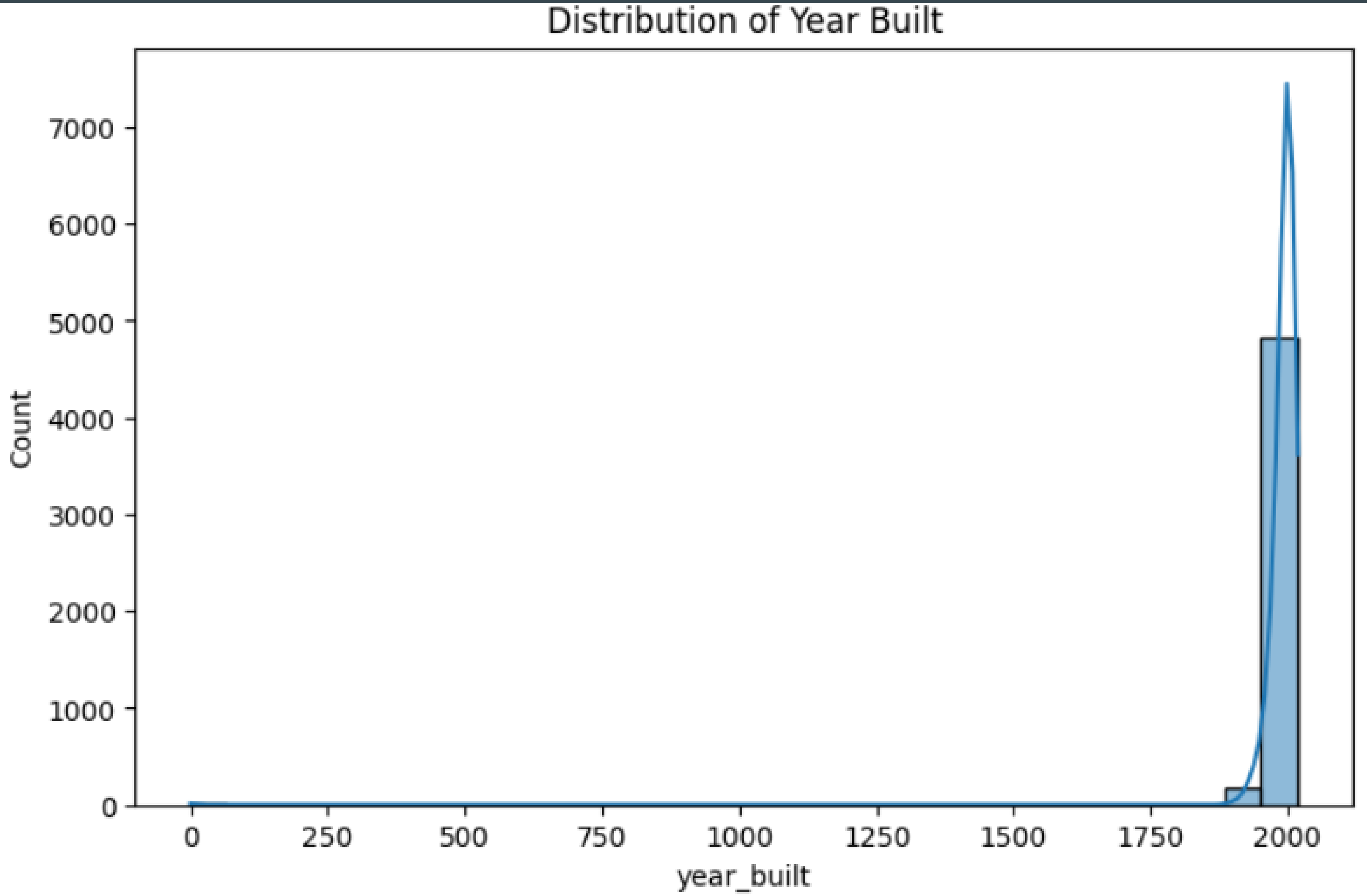
Used for: Checking the distribution of a numerical variable.

- The X-axis: Value of the variable (e.g., sold_price).
- The Y-axis: Count (frequency) of how many times a value appears.

Explanation

- sns.histplot() → Creates a histogram.
- bins=30 → Divides the data into 30 bins (smaller bins mean lower counts).
- kde=True → Adds a smooth curve (Kernel Density Estimation) to show distribution shape.

📌 Example:
If sold_price values range from 100K to 1M, this plot shows how many houses fall in each price range.


Distribution of Year Built
7000
6000
5000
4000
3000
2000
1000
0
Count
0
250
500
750
1000
1250
1500
1750
2000
year_built

Distribution Plot

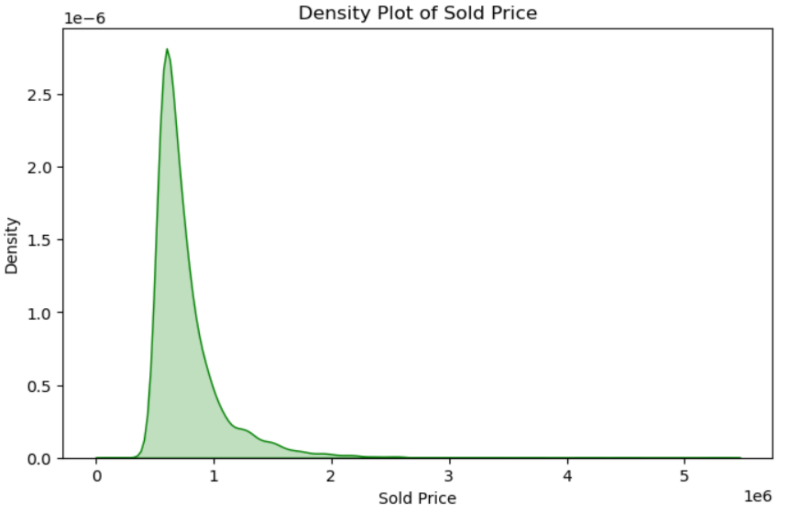
Used for: Checking if the data follows a normal distribution.

- Similar to a histogram but focuses more on smooth curves.

Explanation

- sns.kdeplot() → Creates a density plot.
- shade=True → Fills the area under the curve.
- This helps check if the data is skewed (left/right) or normally distributed.

Example:
If most house prices are between 200K-500K, the curve will peak there.



Pair Plot

Used for: Checking relationships between multiple numerical variables.

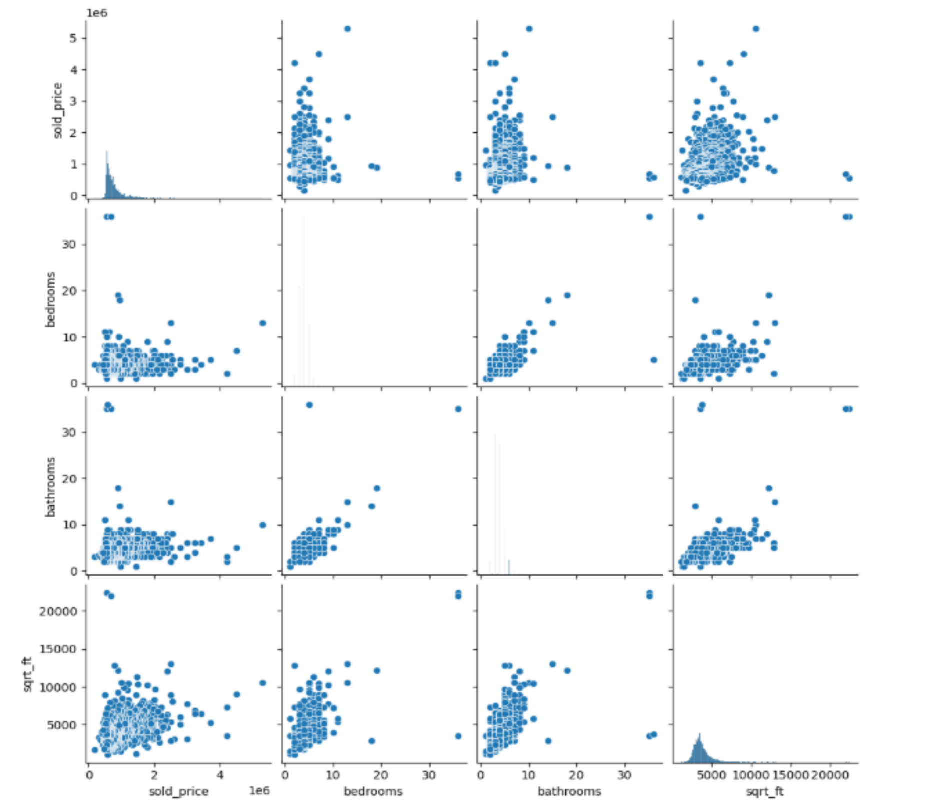
- It creates scatter plots + histograms for every combination.

Explanation

- sns.pairplot(df[columns]) → Plots all numerical columns.
- Helps detect correlations (e.g., larger houses have higher prices?).

Example:

- Sold Price vs Bedrooms → Do more bedrooms increase price?
- Sold Price vs sqrt_ft → Bigger houses → Higher price?



Box and Whisker Plot (Boxplot)

Used for: Identifying outliers in numerical data.

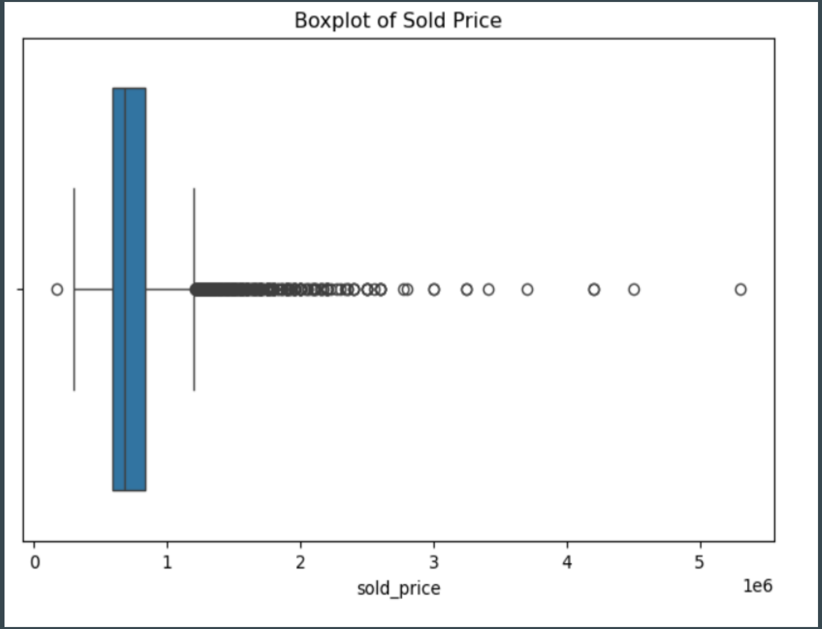
- Shows median, quartiles, and extreme values.

Explanation

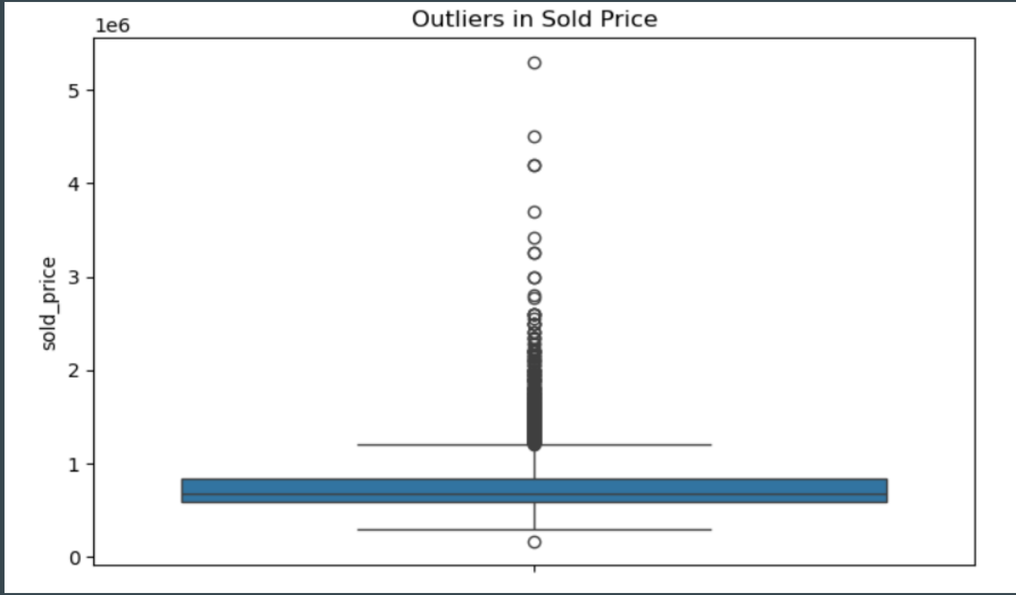
- sns.boxplot() → Creates a box plot.
- The box shows Q1 (25%), median (50%), and Q3 (75%).
- Outliers appear as individual dots beyond whiskers.

Example:
If most houses are under 500K, but some are 5M+,

those 5M+ values will appear as outliers.



Outlier Detection with Boxplot



Used for: Identifying extreme values in specific columns.

Explanation

- sns.boxplot(y=df['sold_price']) → Detects extreme price points.
- Outliers = Dots beyond whiskers (data points far from median).

Example:

- If 90% of house prices are below 1M, but some are 10M, those 10M houses are outliers

Correlation Matrix (Heatmap)

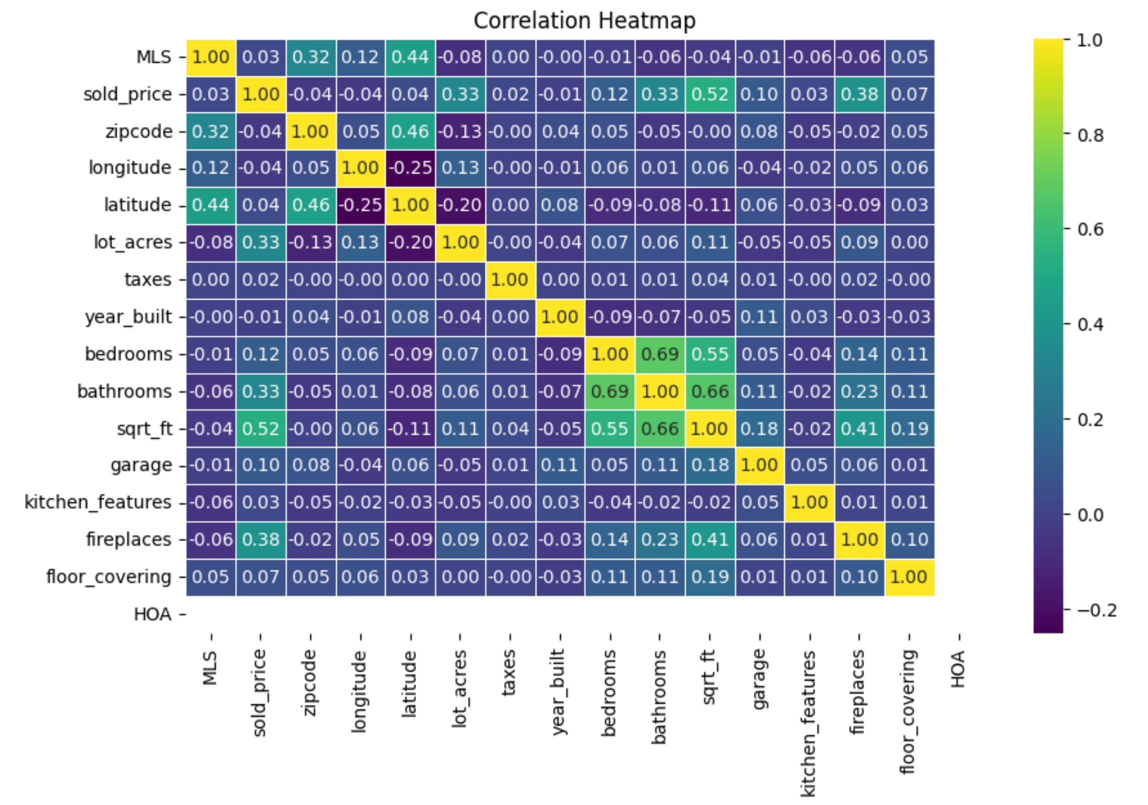
- Used for: Finding relationships between numerical columns.
- Shows which variables affect each other.

Explanation

- df.corr() → Calculates correlation between numerical columns.
- sns.heatmap() → Displays it as a color-coded matrix.
- Closer to +1 → Strong positive correlation (e.g., more sqft → higher price).
- Closer to -1 → Strong negative correlation.
- 0 means no relation.

Example:

- Sold Price & Square Feet = 0.85 → Bigger houses cost more.
- Sold Price & Bedrooms = 0.40 → Weak relation (more bedrooms don’t always mean higher price).



Correlation Heatmap

	MLS	sold_price	zipcode	longitude	latitude	lot_acres	taxes	year_built	bedrooms	bathrooms	sqrt_ft	garage	kitchen_features	fireplaces	floor_covering	HOA
MLS	1.00	0.03	0.32	0.12	0.44	-0.08	0.00	-0.00	-0.01	-0.06	-0.04	-0.01	-0.06	-0.06	0.05	
sold_price	0.03	1.00	-0.04	-0.04	0.04	0.33	0.02	-0.01	0.12	0.33	0.52	0.10	0.03	0.38	0.07	
zipcode	0.32	-0.04	1.00	0.05	0.46	-0.13	-0.00	0.04	0.05	-0.05	-0.00	0.08	-0.05	-0.02	0.05	
longitude	0.12	-0.04	0.05	1.00	-0.25	0.13	-0.00	-0.01	0.06	0.01	0.06	-0.04	-0.02	0.05	0.06	
latitude	0.44	0.04	0.46	-0.25	1.00	-0.20	0.00	0.08	-0.09	-0.08	-0.11	0.06	-0.03	-0.09	0.03	
lot_acres	-0.08	0.33	-0.13	0.13	-0.20	1.00	-0.00	-0.04	0.07	0.06	0.11	-0.05	-0.05	0.09	0.00	
taxes	0.00	0.02	-0.00	-0.00	0.00	-0.00	1.00	0.00	0.01	0.01	0.04	0.01	-0.00	0.02	-0.00	
year_built	-0.00	-0.01	0.04	-0.01	0.08	-0.04	0.00	1.00	-0.09	-0.07	-0.05	0.11	0.03	-0.03	-0.03	
bedrooms	-0.01	0.12	0.05	0.06	-0.09	0.07	0.01	-0.09	1.00	0.69	0.55	0.05	-0.04	0.14	0.11	
bathrooms	-0.06	0.33	-0.05	0.01	-0.08	0.06	0.01	-0.07	0.69	1.00	0.66	0.11	-0.02	0.23	0.11	
sqrt_ft	-0.04	0.52	-0.00	0.06	-0.11	0.11	0.04	-0.05	0.55	0.66	1.00	0.18	-0.02	0.41	0.19	
garage	-0.01	0.10	0.08	-0.04	0.06	-0.05	0.01	0.11	0.05	0.11	0.18	1.00	0.05	0.06	0.01	
kitchen_features	-0.06	0.03	-0.05	-0.02	-0.03	-0.05	-0.00	0.03	-0.04	-0.02	-0.02	0.05	1.00	0.01	0.01	
fireplaces	-0.06	0.38	-0.02	0.05	-0.09	0.09	0.02	-0.03	0.14	0.23	0.41	0.06	0.01	1.00	0.10	
floor_covering	0.05	0.07	0.05	0.06	0.03	0.00	-0.00	-0.03	0.11	0.11	0.19	0.01	0.01	0.10	1.00	
HOA																

Summary of Exploratory Data Analysis (EDA)

Key Insights & Steps

✓ Understanding Data: Identified numerical & categorical columns
✓ Handling Missing Values: Used Mean/Median/Mode for imputation
✓ Data Distribution: Checked skewness & outliers using histograms & boxplots
✓ Correlation Analysis: Found relationships using heatmaps & pair plots
✓ Outlier Detection: Used boxplots method to detect extreme values

Visualizations Used

✓ Histogram , Boxplot, Pairplot,Heatmap

Conclusion

✓ The dataset has been successfully cleaned and saved as a CSV file, ensuring it is ready for further data analysis and insights.

Question and Answer