Data Cleaning and EDA for House Sales Dataset

Prepared by Manoj

Date – 27-08-2024

Index -

Overview -

Project Background -

- Analyzing a real estate dataset with 5,000 entries to understand the factors that influence house prices.

Dataset -

- Contains key features like `sold_price’ , `bedrooms’ , `bathrooms’ , `square_footage’ & `lot_size

Importance -

- Insights from this analysis will support accurate property pricing and informed decision-making in the real estate market.

Objective:

Primary Goal:

- Prepare a clean, structured dataset ready for predictive modeling by addressing missing values, converting data types, and managing outliers.

Specific Tasks:

- **Data Cleaning:** Ensure completeness and accuracy by handling missing values and converting data types.
- **EDA:** Uncover patterns and relationships through visualizations to inform the modeling process.

Outcome:

- Deliver a cleaned dataset and documented insights, setting the foundation for the next phase of model development.

Dataset Overview -

- **Size:** 5,000 entries with 16 features, including sold_price, bedrooms, bathrooms, and square footage.

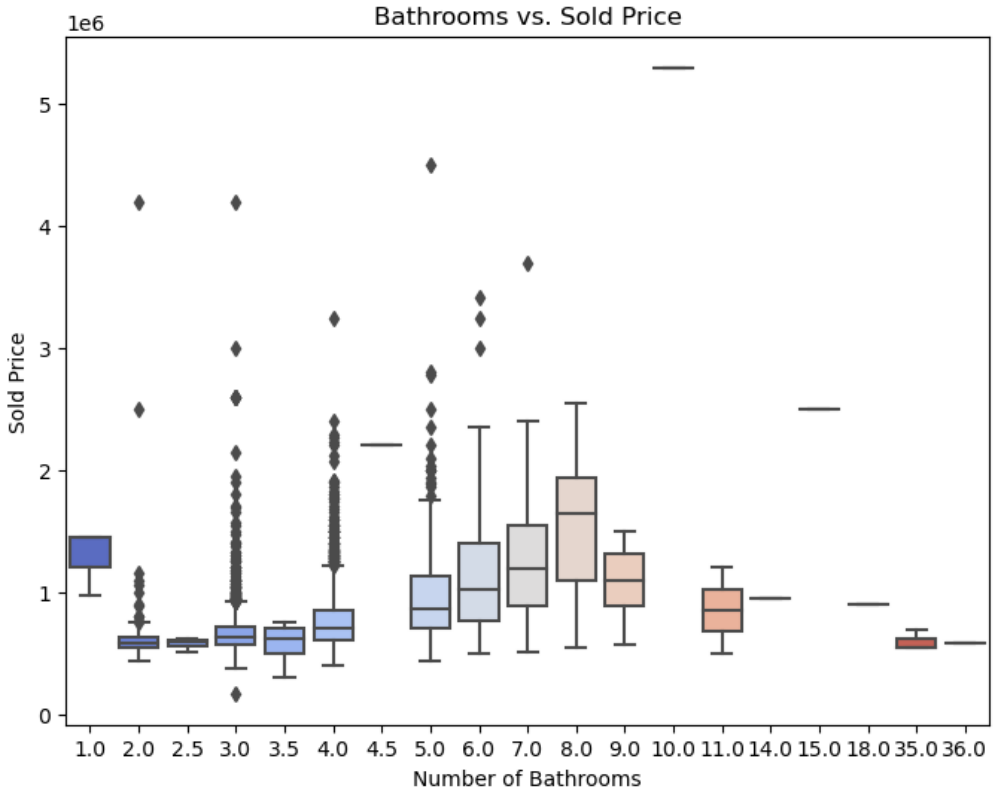
- **Key Attributes:**
  - **Sold Price:** Target variable for predicting property values.
  - **Lot Size, Square Footage:** Indicators of property size.
  - **Garage, HOA Fees:** Features affecting market value

Data Cleaning Steps -

- **Missing Values :** Filled gaps in lot_acres, bathrooms, and square footage using median values.
- **Data Types:** Converted columns like bathrooms, garage, and HOA to numeric formats for consistency.
- **Outliers:** Identified and managed extreme values in key features like sold_price to avoid skewed results.

Key Visuals -

1. **Box Plot: Bathrooms vs. Sold Price**

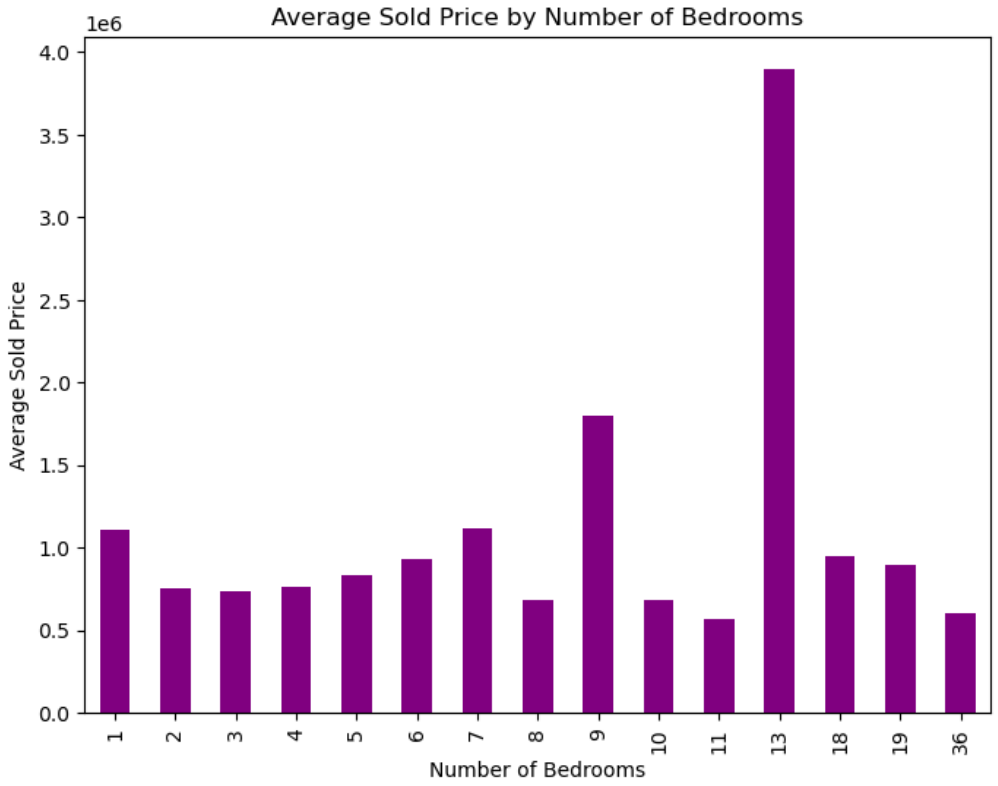


Insights:

- The box plot suggests that the number of bathrooms is positively correlated with the sold price, especially as the number of bathrooms increases from 5 to 9.
- However, there are significant outliers in almost every category, indicating that other factors besides the number of bathrooms also play a substantial role in determining the sold price.
- Properties with very high numbers of bathrooms (15, 18, 36) seem to be rare and may fall into niche markets

Key Visuals -

2. Bar Graph: Average Sold Price by Number of Bedrooms



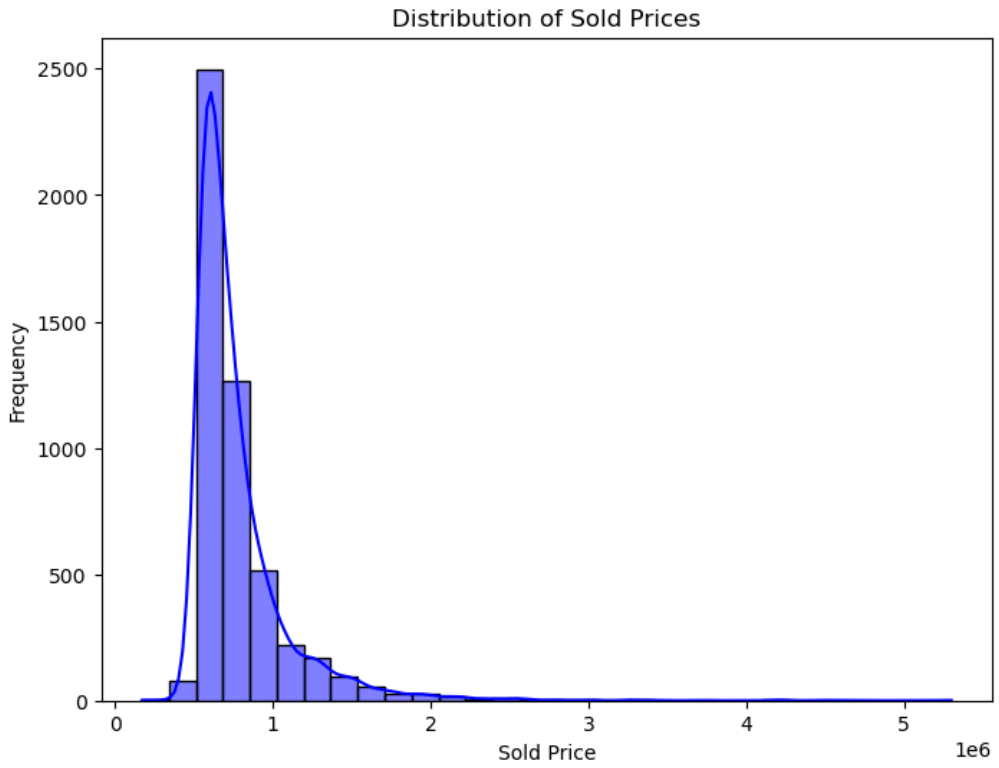
Insights:

The graph indicates that while most properties with typical bedroom counts (2-7) tend to have similar average sold prices, properties with a very high number of bedrooms (especially 13) command significantly higher prices, suggesting they fall into a different category of luxury or specialized real estate.

The spike at 13 bedrooms could be an outlier or due to a small number of high-value transactions.

Key Visuals -

3. Histogram: Distribution of Sold Prices



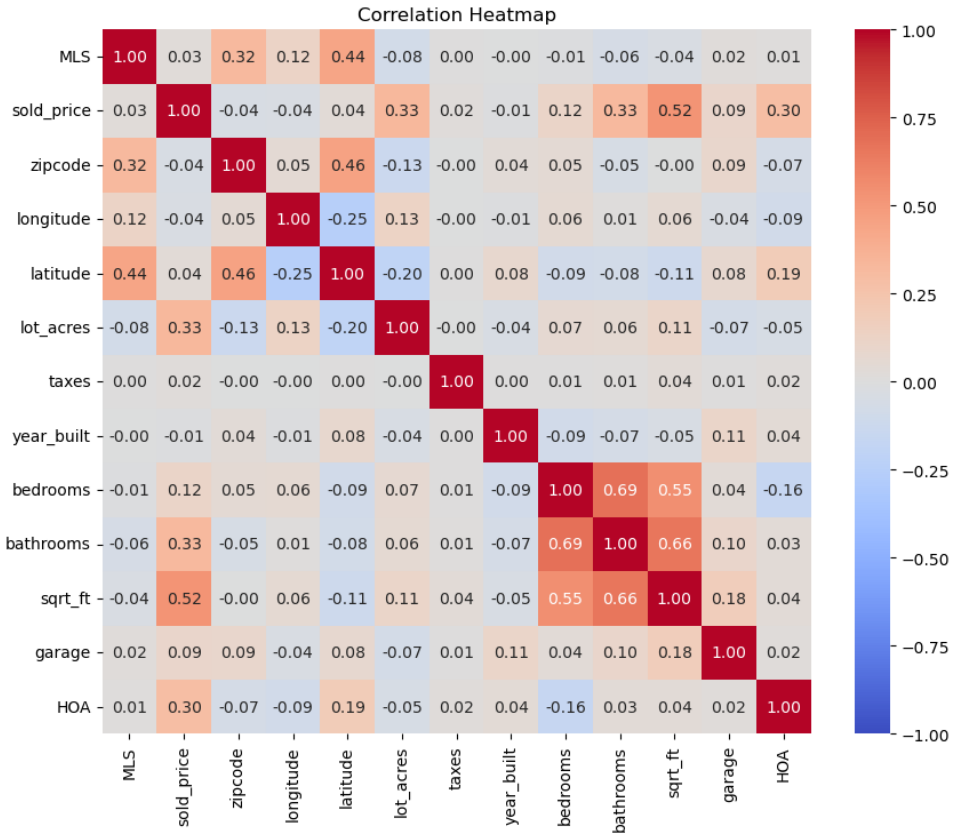
Insights:

- The distribution is **right-skewed**: Most sales are clustered at lower price ranges, with a long tail extending to the right, indicating that there are some sales at much higher prices, but these are less frequent.
- The peak (mode) is at a lower price range, with the frequency rapidly decreasing as prices increase.

This kind of distribution is typical in datasets where most items are sold at a lower price, but a few outliers exist at much higher prices.

Key Visuals -

4. Heatmap: Correlation Between Features



Insight:

The heatmap provides a clear visual representation of how variables in this dataset relate to each other. It can help identify key factors that influence the sold price, such as square footage, lot size, and the number of bathrooms.

Negative Correlation

- MLS, longitude, and latitude have little to no correlation with most other variables.
- Negative correlations, though weak, include latitude with lot_acres and longitude with sqft_ft.

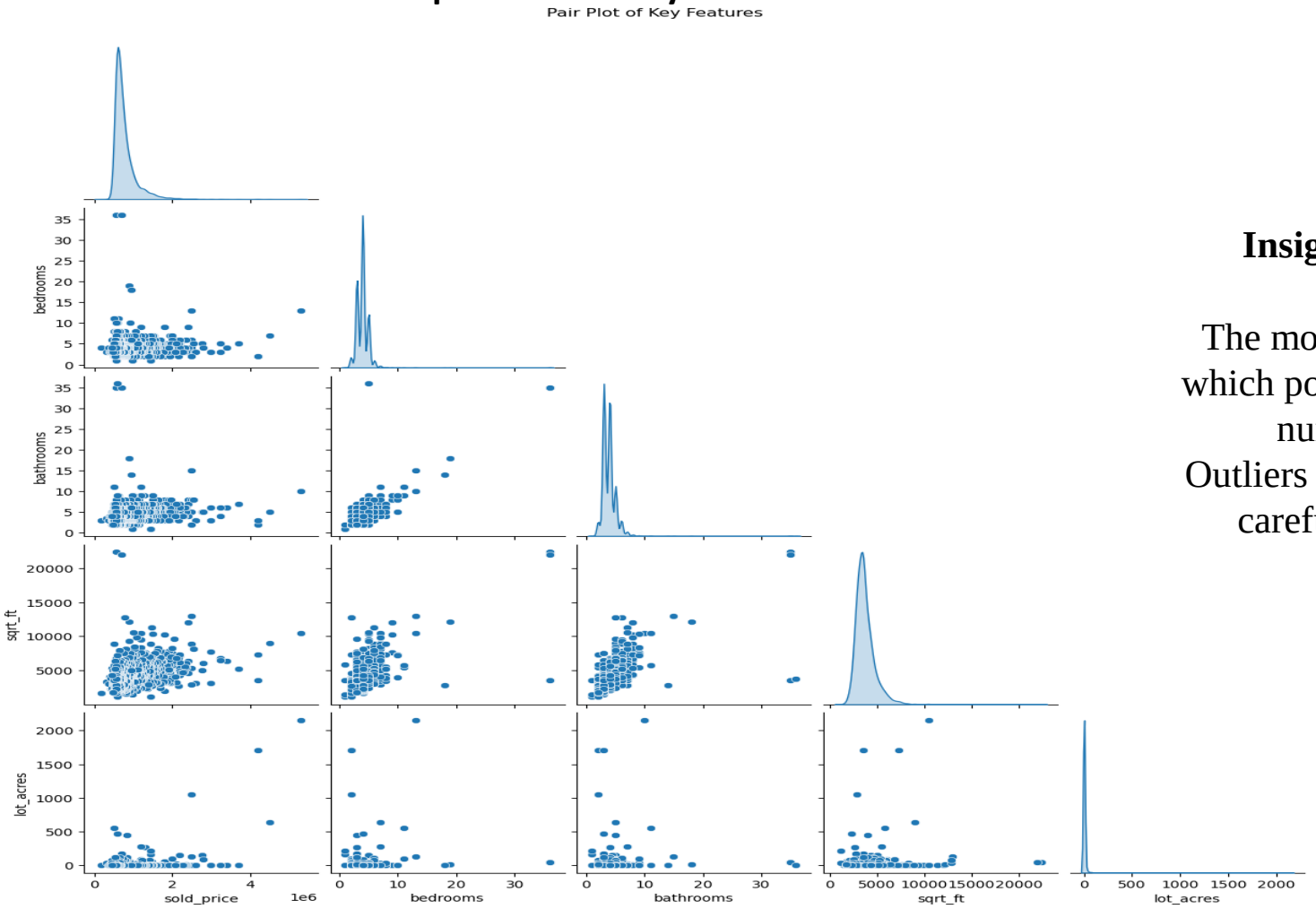
Positive Correlation

- The sold_price is positively correlated with several variables:
  - sqft_ft (0.52): Strong positive correlation.
  - bathrooms (0.33) and lot_acres (0.33): Moderate positive correlations.

These indicate that properties with more square footage, more bathrooms, and larger lots tend to have higher sold prices

Key Visuals -

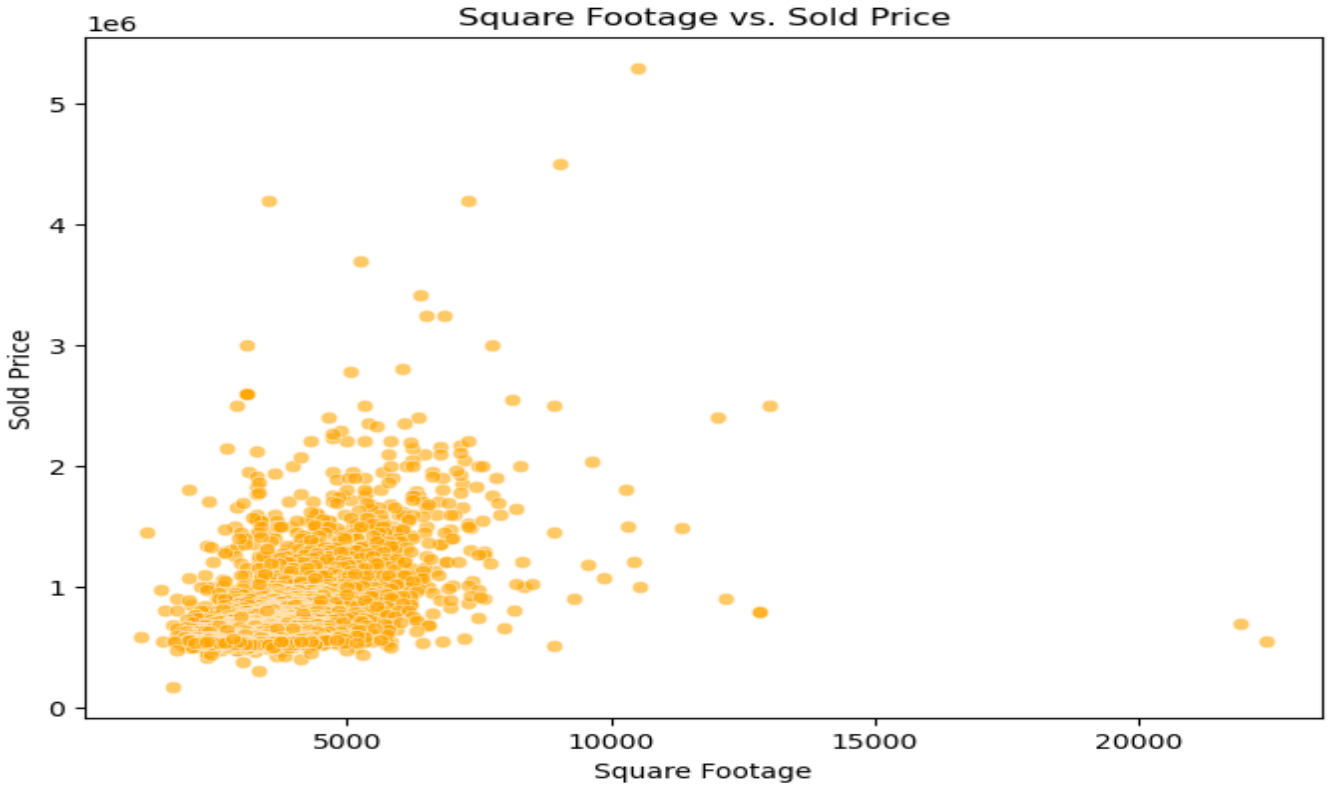
5. Pair Plot: Relationships Between Key Features



Insight: The pair plot is useful for visually assessing the relationships between the key features.
The most notable relationships in this plot involve sold_price, which positively correlates with features like square footage, the number of bathrooms, and the number of bedrooms.
Outliers and skewness are also visible, highlighting the need for careful interpretation when analyzing these relationships.

Key Visuals -

6. Scatter Plot: Square Footage vs. Sold Price



Insight: The scatter plot suggests a positive correlation between square footage and sold price, but with a significant amount of variation that could be due to additional factors not depicted in the graph.
The outliers indicate that very large or expensive properties do not necessarily follow the same pricing patterns as more typical properties.

EDA Summary:

1. **Sold Price Distribution**:
   - The sold_price distribution is right-change, indicating a few high-priced properties driving up the average.
2. **Lot Acres Distribution**:
   - The lot_acers distribution is highly change with a few large outliers. Most properties have small lot sizes.
3. **Bedroom Count**:
   - The majority of homes have 3-4 bedrooms, which aligns with typical residential properties.
4. **Square Footage vs. Sold Price**:
   - There is a positive correlation between sqrt_ft (square footage) and sold_price, as expected. Larger homes generally have higher selling prices.
5. **Lot Acres vs. Sold Price**:
   - Most properties with larger lot sizes tend to have higher prices, although the relationship is not as strong as square footage.
6. **Bathrooms vs. Sold Price**:
   - Homes with more bathrooms generally have higher selling prices, but there are some variations, possibly due to other factors like location or home age.

Conclusion

- The task has done are Data Cleaning and EDA.
- Missing values in critical features like lot_acres , bathrooms , and square footage were identified and filled
- Converted non-numeric data (e.g., garage , HOA) to numeric formats.

- Outliers in features like either transformed or capped.

- Standardized units of measurement (e.g., converting square footage consistently) and ensured consistency across categorical features.

- The Visualization effectively convey the most important aspects of the data, enabling a deeper understanding of the relationships, distributions, and potential issues that could impact the modelling process.

- Now the Data is ready for the modelling phase

.

Thank you