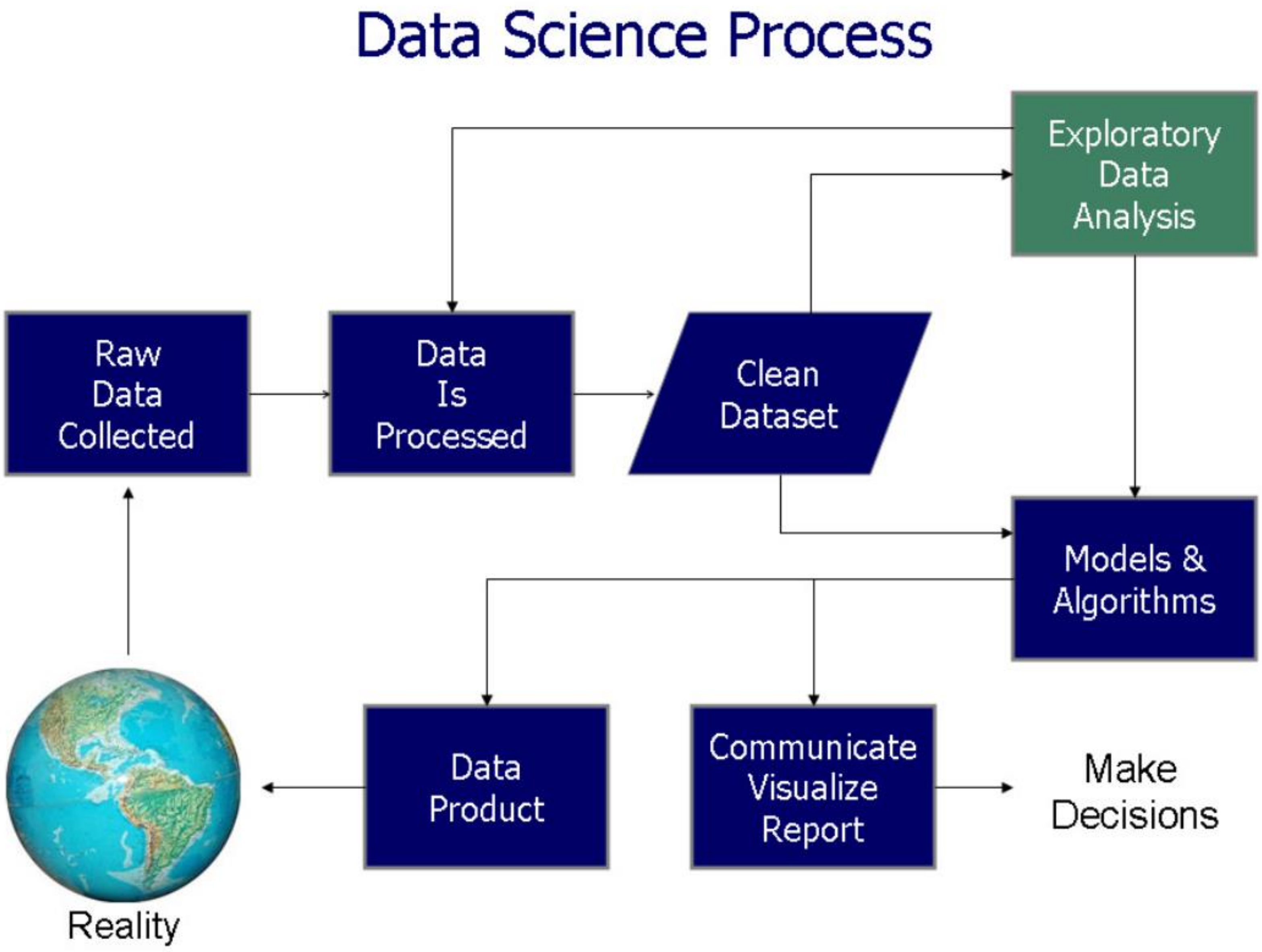
Exploratory Data Analysis (EDA) & Data Cleaning for House Pricing Dataset

PREPARED BY VAISHALI PATELIA

Outline

- Introduction
- Purpose of the Project
- Dataset Overview
- Python packages for EDA
- Data Cleaning
- Data Visualization
- Conclusion
- Recommendation
- References
- Appendices



Source: https://datasilk.com/data-analysis/

Purpose of the Project

Goal: Clean and prepare a housing dataset for the modeling team.

Tasks Performed:

1. Data Loading & Initial Analysis
2. Handling Missing Values
3. Data Type Conversion
4. Outlier Detection & Removal
5. Data Visualization & Insights

Dataset Overview

- Total Rows & Columns: Displayed from df.shape (5000, 16)
- Feature Summary: Key columns (e.g., sold_price, sqrt_ft, bedrooms, lot_acres)
- Missing Values: Shown using df.info() (lot_acres, bathrooms, sqrt_ft, garage, HOA)
- Data Types: Numeric & Categorical columns (bathrooms, sqrt_ft, garage, fireplaces, HOA)

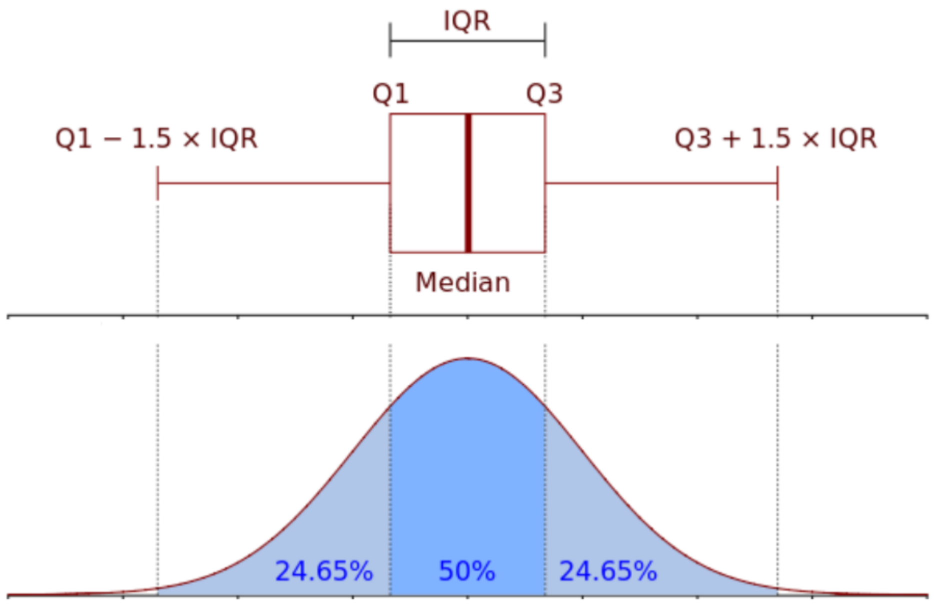
```
df.head()
```

	MLS	sold_price	zipcode	longitude	latitude	lot_acres	taxes	year_built	bedrooms	bathrooms	sqrt_ft	garage	kitchen_features	fireplaces	floor_cove
0	21530491	5300000.0	85637	-110.378200	31.356362	2154.00	5272.00	1941	13	10.0	10500.0	0.0	Dishwasher, Freezer, Refrigerator, Oven	6.0	Mexicar V
1	21529082	4200000.0	85646	-111.045371	31.594213	1707.00	10422.36	1997	2	2.0	7300.0	0.0	Dishwasher, Garbage Disposal	5.0	Natural S (
2	3054672	4200000.0	85646	-111.040707	31.594844	1707.00	10482.00	1997	2	3.0	NaN	NaN	Dishwasher, Garbage Disposal, Refrigerator	5.0	Natural S Other:
3	21919321	4500000.0	85646	-111.035925	31.645878	636.67	8418.58	1930	7	5.0	9019.0	4.0	Dishwasher, Double Sink, Pantry: Butler, Refri...	4.0	Ceramic Lami V
4	21306357	3411450.0	85750	-110.813768	32.285162	3.21	15393.00	1995	4	6.0	6396.0	3.0	Dishwasher, Garbage Disposal, Refrigerator, Mi...	5.0	Ca Con

Python packages for EDA

Data Cleaning

- **Handled Missing Values** : Used median imputation for numerical data & 0 for categorical values
- **Outlier Detection & Removal** : Applied Interquartile Range (IQR) method
- **Fixed Data Types** : Converted numerical values stored as strings
- **Removed Duplicates** : Ensured data consistency
- **Standardized Column Names** : For easy reference in modeling

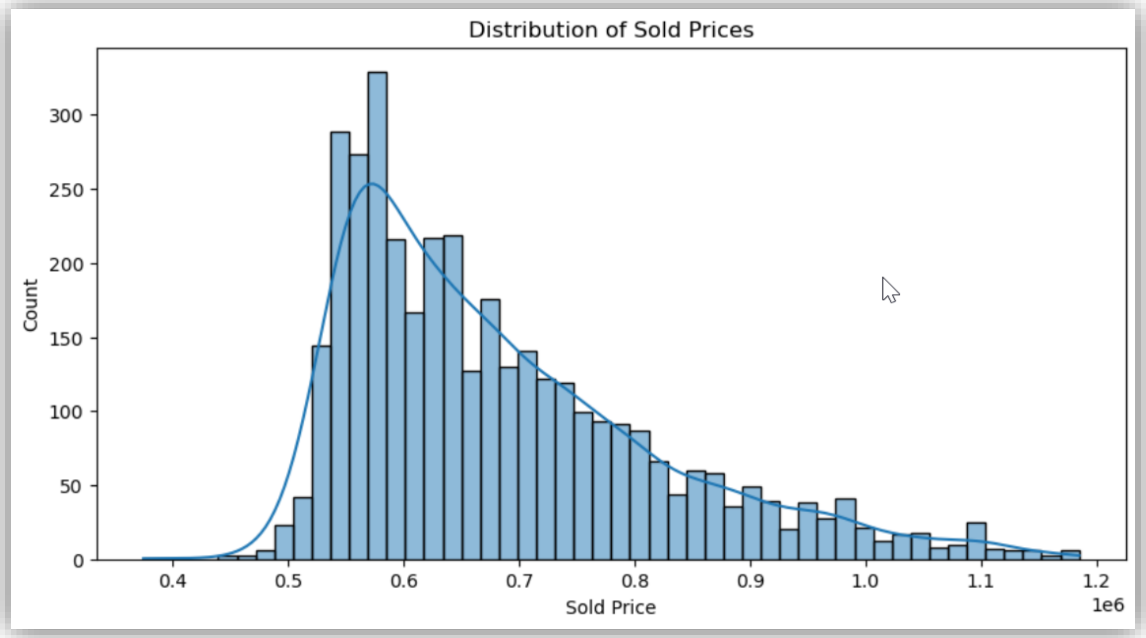


Data Visualization

- Helps detect patterns, trends, and anomalies
- Makes data easier to interpret and communicate

Visuals Used in this Project:

- Histograms & Boxplots : Understanding distributions & outliers
- Correlation Heatmap : Finding relationships between variables
- Scatterplots : Checking trends between house size & price
- Bar Chart **:** Compare categorical data distributions (e.g., house type, location)
- Pair Plot : Shows relationships between multiple numerical variables

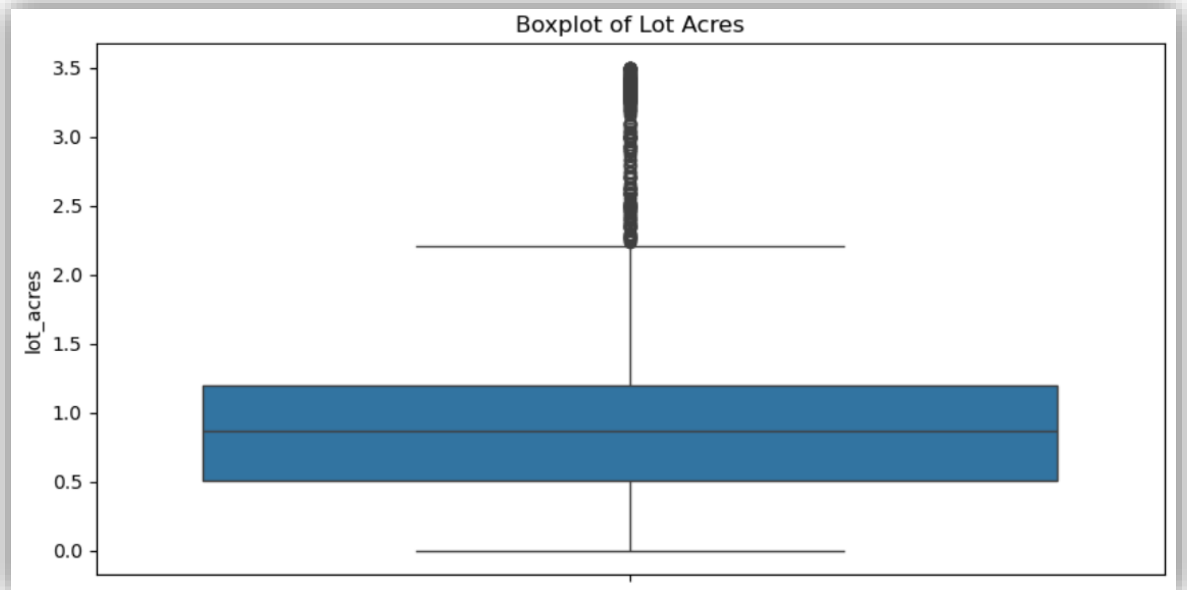


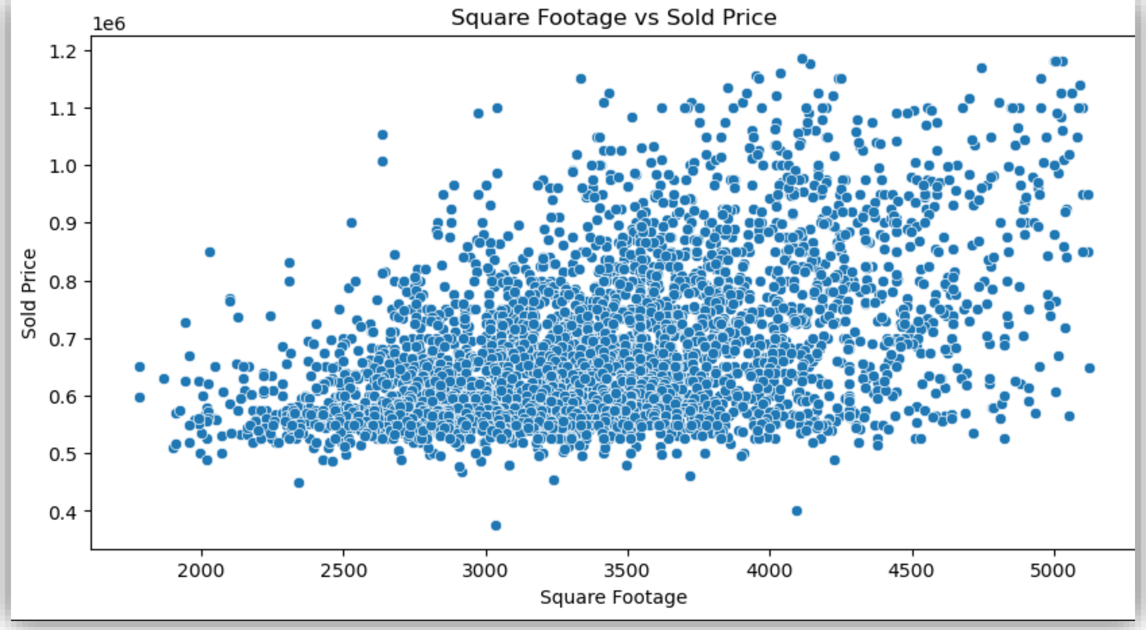
Histogram of Sold Prices

House prices were right-skewed, meaning some high-value properties affected the average.

Boxplot of Lot Acres

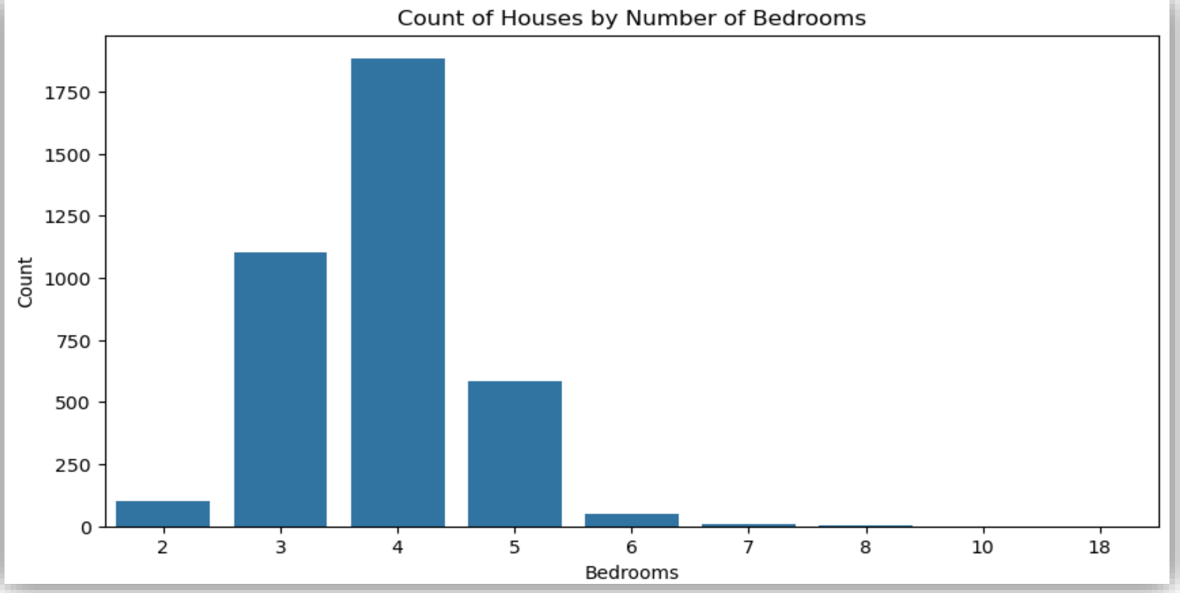
Shows data distribution & outliers using quartiles.





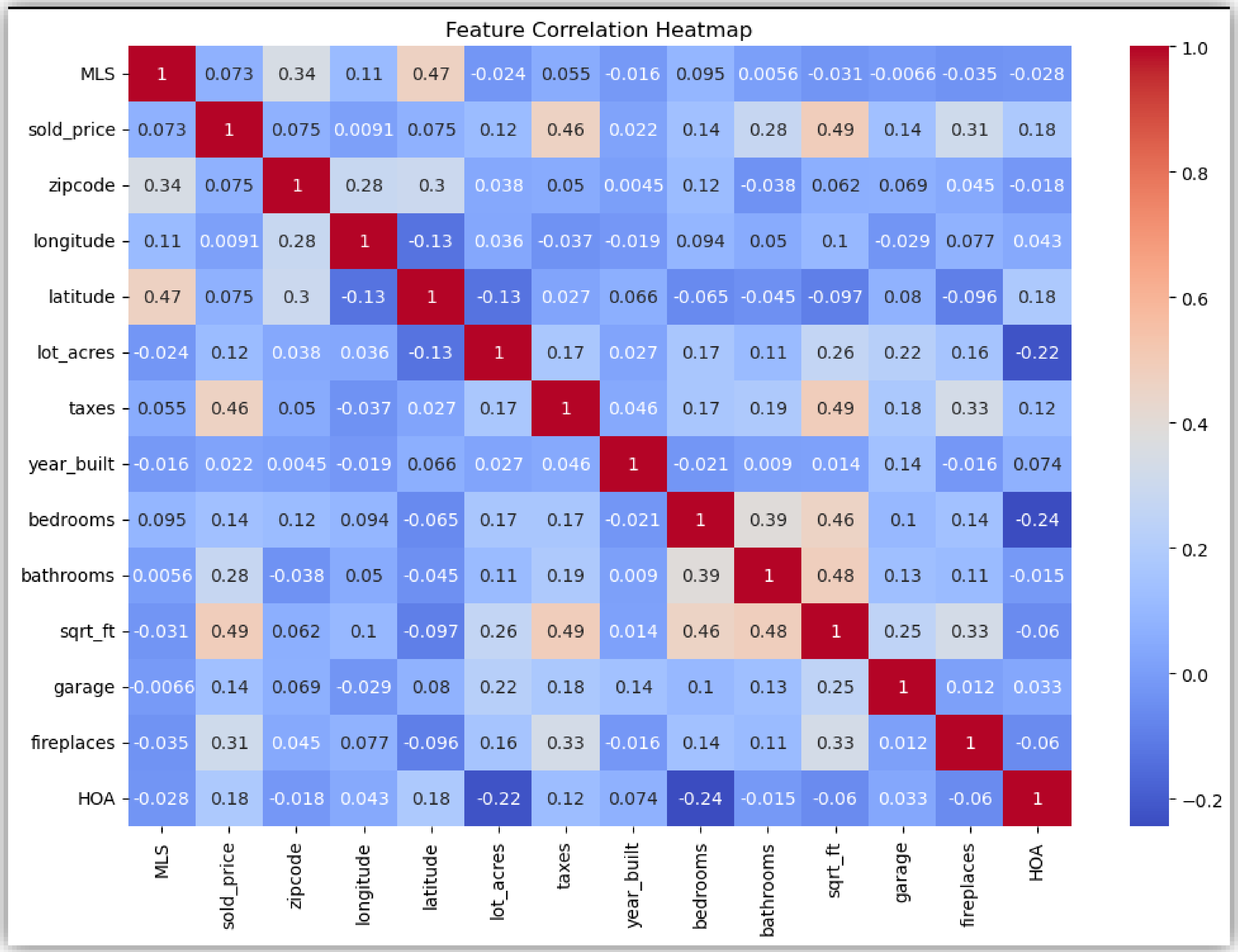
Scatter plot of Square Footage vs Price

Used to visualize the relationship between house size (square footage) and sold price.
This helps identify whether larger homes tend to have higher prices.



Count Plot of Number of Bedrooms

Used to visualize the distribution of houses based on the number of bedrooms.
This helps identify the most common house types in the dataset.



Correlation Heatmap

Helps identify feature relationships to select relevant predictors for modeling.

Square footage, number of rooms, and overall quality had the highest correlation with house prices.

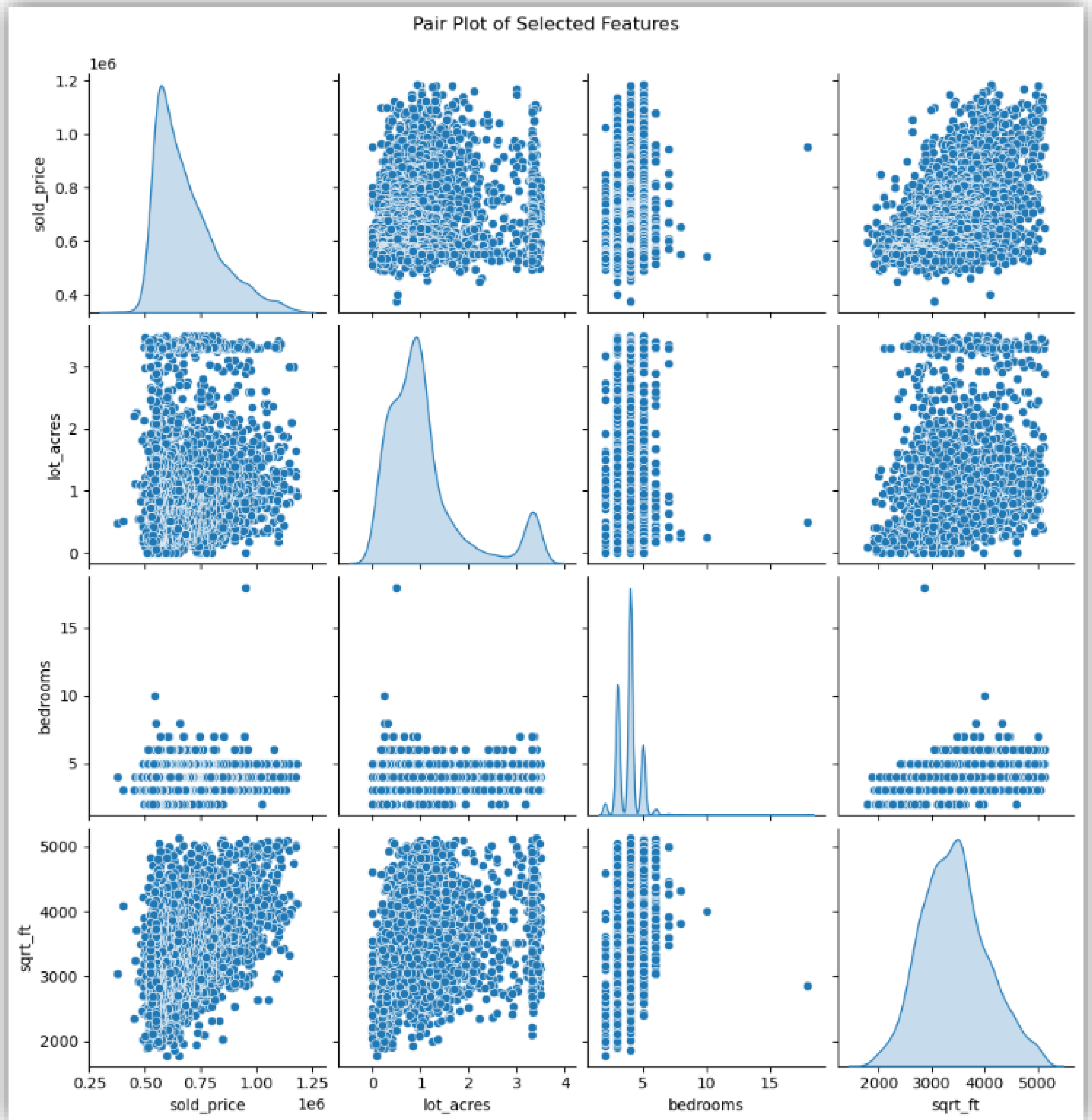
Lot size had weak correlation, indicating it might not be a strong predictor.

PairPlot of Selected Features

House price vs. square footage showed a strong positive correlation (larger homes tend to be more expensive).

Year built vs. price suggested that newer homes generally have higher prices.

Outliers were visible in some features like lot size.



Conclusion

- Missing values handled using median imputation
- Outliers removed using IQR method to improve model accuracy
- Right-skewed price distribution, suggesting need for log transformation
- Strong correlation between house size, number of rooms, and price
- Categorical features (e.g., location, house type) significantly impact price

Recommendation

- Feature Engineering for model improvement
- Train machine learning models on cleaned data
- Fine-tune models for better predictions

References

- Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques.* Elsevier.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.
- Hawkins, D. M. (1980). *Identification of Outliers.* Springer.
- García, S., Luengo, J., & Herrera, F. (2015). *Data Preprocessing in Data Mining*. Springer.
- Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow.* O'Reilly Media.
- Pedregosa, F., et al. (2011). *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research.
- Seabold, S., & Perktold, J. (2010). *Statsmodels: Econometric and Statistical Modeling with Python*.

Appendices

What is EDA?

EDA is the **process of analyzing and summarizing datasets** to uncover patterns, detect anomalies, and gain insights before building models.

Key Objectives:

- Understand data structure and distributions
- Identify missing values and inconsistencies
- Detect and handle outliers
- Explore correlations between variables
- Generate visualizations to interpret trends

Techniques Used in EDA:

- **Descriptive Statistics:** Mean, median, standard deviation
- **Data Visualization:** Histograms, Boxplots, Scatterplots, Heatmaps
- **Outlier Detection:** Interquartile Range (IQR), Z-score

Handled Missing Values

Mean

The average of all values in a dataset

Best for normally distributed data

Median

The middle value when data is ordered

Better for skewed data or when outliers are present

Standard Deviation (SD)

Measure of spread or dispersion of data from the mean

Lower SD = data clustered closer to the mean

Higher SD = data more spread out

Quantifies variability and consistency in data

Outlier Detection

Interquartile Range (IQR)

- IQR is the range between the first quartile (Q1) and third quartile (Q3)
- IQR = Q3 - Q1
- Outlier identification:
  - Lower bound: Q1 - 1.5 * IQR
  - Upper bound: Q3 + 1.5 * IQR
  - Any data point outside these bounds is considered an outlier

Z-score

- Measures how many standard deviations a data point is from the mean
- $Z = (X - \mu) / \sigma$
  Where X is the data point, μ is the mean, and σ is the standard deviation
- Outlier identification:
  - Typically, $|Z| > 3$ is considered an outlier
  - This threshold can be adjusted based on the specific needs of the analysis


Thank You