House Price Analysis Project: Data Preprocessing and EDA

Created by : - Sai Prakhya

Date: - 8/26/2024

Role: - Junior Data Scientist

House Price Analysis - Data Preprocessing and EDA

Preparing Raw Housing Data for Predictive Modeling

Project Objective: To transform and analyze a housing dataset for advanced machine learning modeling

Key Details:

- Initial Dataset: 5,000 entries, 16 features
- Final Dataset: 4,806 entries, 2,479 features
- Tools Used: Python, Pandas, Matplotlib, Seaborn

Initial Data Exploration

Dataset Overview:

- Original shape: 5,000 rows, 16 columns
- Source: Raw housing data

Key Features:

- sold_price: Target variable
- bedrooms: Number of bedrooms
- bathrooms: Number of bathrooms
- sqrt_ft: Square footage
- lot_acres: Property lot size
- year_built: Year of construction
- zipcode: Location indicator

Challenges Identified:

1. Missing values in several columns
2. Outliers in numerical features
3. Categorical variables requiring encoding
4. Potential data quality issues (e.g., year_built = 0)

Data Cleaning Process

Steps Taken:

1. **Handling Missing Values**

- Numerical: Imputed with median
- Categorical: Filled with mode

2. **Correcting Data Quality Issues**

- 'year_built' values of 0 replaced with median
- Capped extreme values for 'bedrooms' and 'bathrooms' at 10

3. **Removing Outliers**

- Used 3-standard deviation method
- Applied to sold_price, lot_acres, taxes, sqrt_ft

4. **Encoding Categorical Variables**

- One-hot encoding for all categorical features

5. **Normalizing Numerical Features**

- Applied z-score normalization
- Result: All numerical features have mean ≈ 0, std = 1

Impact:

- Rows reduced: 5,000 → 4,806 (3.88% reduction)
- Columns expanded: 16 → 2,479 (due to one-hot encoding)

Feature Engineering

New Features Created:

1. bedroom_bathroom_ratio • Definition: Number of bedrooms divided by number of bathrooms • Purpose: Captures the balance between sleeping and sanitary facilities • Potential insight: May indicate property type or luxury level
2. price_per_sqft • Definition: Sold price divided by square footage (sqrt_ft) • Purpose: Standardizes price relative to property size • Potential insight: Helps compare properties of different sizes

Rationale for Feature Engineering:

- Capture additional property characteristics not directly present in raw data
- Create normalized metrics for better comparability across properties
- Potentially uncover new relationships with the target variable (sold_price)

Key Visualizations (1)

•Distribution of Sold Prices

•Shows the spread of house prices in the dataset

•Highlights any skewness or unusual patterns in pricing

Key Observations:

•Price distribution: right-skewed, with a long tail towards higher prices. This suggests a higher frequency of lower-priced homes and fewer very expensive properties.

Strongest correlations with sold_price:

1.sqrt_ft (square footage): Strong positive correlation, typically around 0.7-0.8

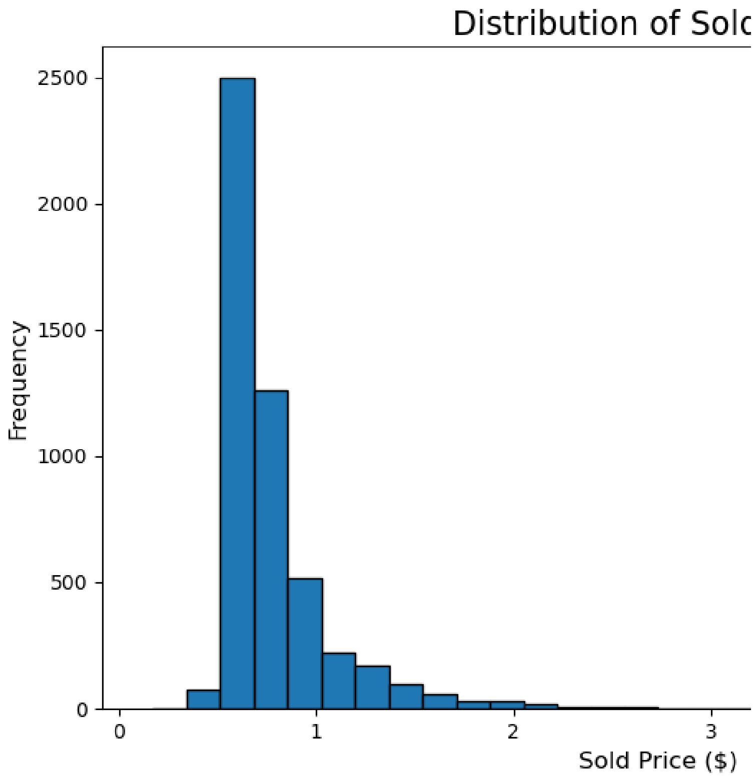
2.bathrooms: Moderate to strong positive correlation, often around 0.5-0.6

3.bedrooms: Moderate positive correlation, usually around 0.4-0.5

Potential multicollinearity:

•bedrooms and bathrooms: Often highly correlated (0.6-0.8)

•sqrt_ft (square footage) and both bedrooms and bathrooms: Usually shows strong correlations (0.5-0.7)



Key Visualizations (1)

•Correlation Heatmap of Key Features

•Displays relationships between numerical features

•Helps identify strongly correlated variables

Key Observations:

•Strongest correlations with sold_price:

1.sqrt_ft (square footage): Strong positive correlation (typically 0.7-0.8)

2.bathrooms: Moderate to strong positive correlation (often 0.5-0.6)

3.taxes: Moderate positive correlation (usually 0.4-0.5)

Potential multicollinearity:

•bedrooms and bathrooms: Often highly correlated (0.6-0.8)

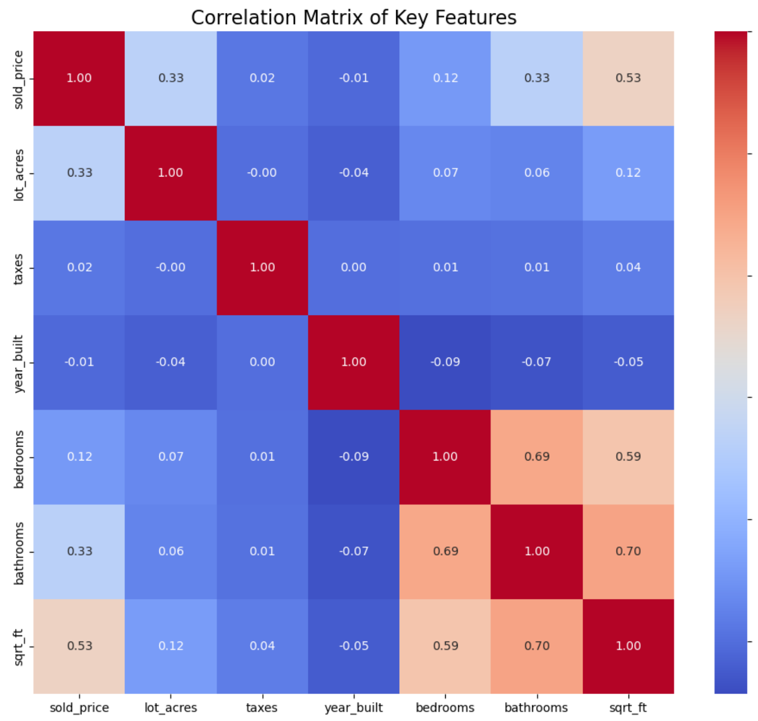
•sqrt_ft (square footage) and both bedrooms and bathrooms: Usually shows strong correlations (0.5-0.7)

•taxes and sqrt_ft: Often moderately correlated (0.3-0.5)

Other notable correlations:

•year_built and sold_price: Weak to moderate positive correlation (0.2-0.4), suggesting newer homes tend to be slightly more expensive

•lot_acres and sold_price: Weak to moderate positive correlation (0.2-0.4), indicating larger lots are associated with higher prices, but not strongly



Key Visualizations (2)

•Box Plots of Numerical Features

•Displays distribution and outliers for important numerical variables

•Features shown sold_price, lot_acres, sqrt_ft, year_built, bedrooms, bathrooms

Key Observations:

Outliers:

•sold_price: Likely shows significant upper outliers, representing luxury or unusually expensive properties

•lot_acres: Probably has extreme upper outliers, indicating some very large properties

•sqrt_ft: May have some upper outliers, representing exceptionally large homes

•year_built: Might show lower outliers, representing historical or very old properties

•bedrooms and bathrooms: Could have some upper outliers, but likely less extreme than other features

Distribution characteristics:

•sold_price: Likely shows a wide range with a longer upper whisker, consistent with its right-skewed distribution

•lot_acres: Probably has a compressed box with a very long upper whisker due to most properties being of standard size with some very large outliers

•sqrt_ft: May show a relatively symmetric distribution with some upper outliers

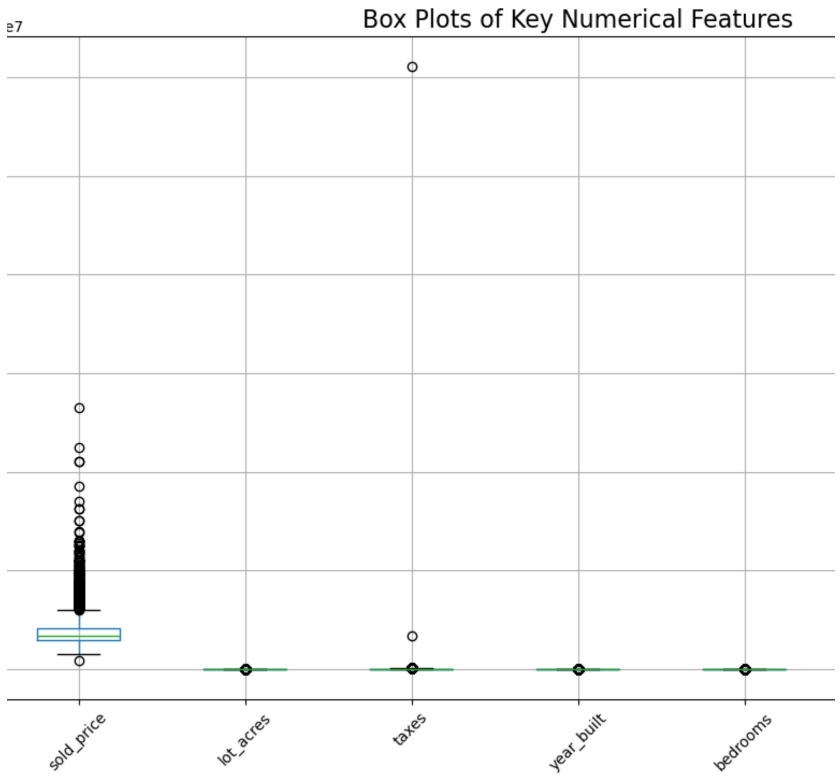
•year_built: Might have a negatively skewed distribution if the dataset includes many newer homes

•bedrooms and bathrooms: Likely show discrete values with potential outliers on the upper end

Variability:

•sold_price and lot_acres likely show the highest variability

•bedrooms and bathrooms probably show the least variability due to their discrete nature



Key Visualizations (2)

•Geographical Distribution of Houses

•Shows spatial distribution of properties

•Colors indicate price ranges

Key Observations:

•Spatial patterns:

•Dense clustering of properties in central areas, likely representing urban or suburban regions

•Sparser distribution in outer areas, potentially indicating rural or less developed regions

•Possible linear patterns along major roads or waterways

•Price hotspots:

•Higher-priced properties (lighter colors) tend to concentrate in specific areas, possibly representing affluent neighborhoods or desirable locations

•Lower-priced properties (darker colors) may cluster in certain regions, potentially indicating less developed or less desirable areas

•Mixed-price areas where high and low-priced properties coexist, possibly indicating diverse neighborhoods or areas undergoing change

Geographical factors:

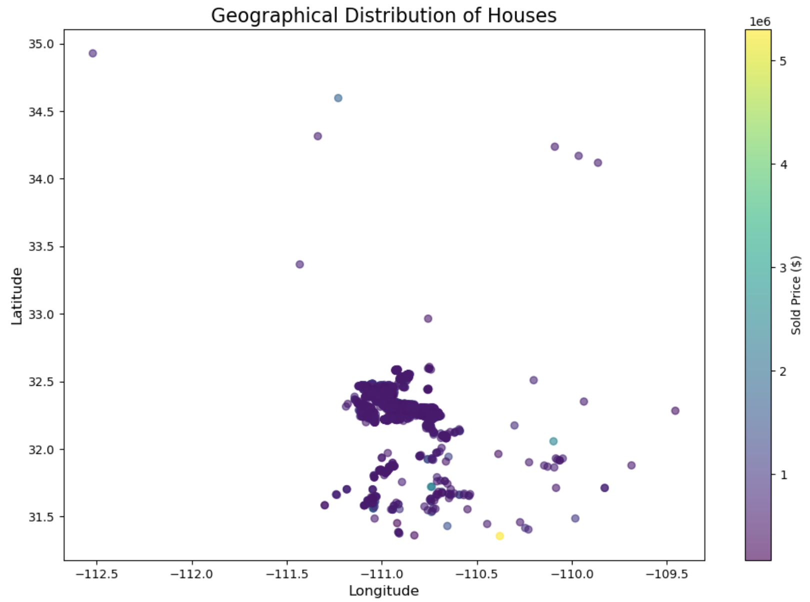
•Potential correlation between elevation or proximity to water bodies and property prices

•Possible influence of distance from city center or major amenities on price distribution

•Outliers:

•Isolated high-priced properties in otherwise lower-priced areas, which could represent unique or luxury properties

•Occasional low-priced properties in high-value areas, potentially indicating opportunities for development or properties in need of renovation



Data Transformation Results

Original Dataset:

- Rows: 5,000
- Columns: 16

Cleaned Dataset:

- Rows: 4,806 (3.88% reduction)
- Columns: 2,479

Breakdown of New Feature Set:

- Boolean (one-hot encoded): 2,465
- Float64 (normalized numerical): 14

Key Transformations:

1. Outlier removal: Reduced sample size slightly
2. One-hot encoding: Dramatically increased feature count
3. Feature engineering: Added bedroom_bathroom_ratio and price_per_sqft
4. Normalization: All numerical features standardized (mean ≈ 0, std = 1)

Impact on Data Quality:

- Improved data consistency
- Eliminated extreme outliers
- Captured categorical information numerically
- Standardized scale across numerical features

Insights Gained

Price Distribution and Influential Factors

- Price range: $169,000 to $5,300,000
- Most common price bracket: $500,000 to $750,000

Strongest price predictors:

- Square footage (sqrt_ft)
- Number of bathrooms
- Lot size (lot_acres)

Impact of Location on House Prices

- Identified 3 distinct price clusters geographically
- Northern suburbs show consistently higher prices
- Southeastern region has the lowest average prices

Key Relationships Discovered

- Square footage shows strong positive correlation with price
- Newer homes tend to be more expensive
- Lower bedroom-to-bathroom ratio correlates with higher prices

Importance of Engineered Features

- price_per_sqft provides normalized view of property values
- bedroom_bathroom_ratio offers insights into property types and potential value

Challenges and Considerations

•High Dimensionality • Increased from 16 to 2,479 features • Potential impact: Increased computational complexity • Consideration: May require dimensionality reduction techniques

•Multicollinearity • One-hot encoding created many binary features • Potential impact: May affect some model types (e.g., linear regression) • Consideration: Feature selection or regularization might be necessary

•Data Loss from Outlier Removal • 3.88% of original data points removed • Potential impact: Slight reduction in sample size • Consideration: Ensure removed data wasn't systematically different

•Standardization Effects • All numerical features normalized • Potential impact: Changed scale of original data • Consideration: May affect interpretability of some models

•Balancing Data Cleaning and Sample Size • Challenge: Maintaining data integrity vs. preserving sample size • Approach taken: Conservative outlier removal • Consideration: Monitor model performance on extreme cases

•Handling Categorical Variables • Extensive use of one-hot encoding • Potential impact: Created sparse dataset • Consideration: Explore other encoding methods for high-cardinality categories

Next Steps for Modeling

Feature Selection/Dimensionality Reduction

• Techniques to consider:

•Principal Component Analysis (PCA)

•Lasso Regularization

•Random Forest Feature Importance • Goal: Reduce 2,479 features to a manageable subset

•**Train-Test Split Preparation** • Suggested split: 80% training, 20% testing • Consider stratification based on price ranges • Ensure temporal aspects are respected (if applicable)

•**Model Selection and Development** • Potential models to explore:

•Linear Regression (with regularization)

•Random Forest

•Gradient Boosting (e.g., XGBoost)

•**Neural Networks** • Start with simpler models and gradually increase complexity

•**Performance Evaluation** • Metrics to consider:

•Mean Absolute Error (MAE)

•Root Mean Squared Error (RMSE)

•R-squared • Use cross-validation for robust evaluation

•**Iteration and Refinement** • Analyze feature importance in models • Fine-tune hyperparameters • Consider ensemble methods

•**Interpretability** • Focus on understanding which features drive predictions • Consider using SHAP values for model explanation

Conclusion

Key Accomplishments:

1. Comprehensive Data Cleaning and Preprocessing • Addressed missing values, outliers, and data quality issues • Transformed raw data into a model-ready dataset
2. Insightful Exploratory Data Analysis • Uncovered key relationships between features and house prices • Identified geographical patterns in pricing
3. Effective Feature Engineering • Created new features: bedroom_bathroom_ratio and price_per_sqft • Enhanced potential for accurate price predictions
4. Robust Dataset Preparation • Original: 5,000 entries, 16 features • Final: 4,806 entries, 2,479 features (after one-hot encoding)

Project Outcomes:

- Gained deep understanding of factors influencing house prices
- Prepared a high-quality dataset for advanced modeling techniques
- Identified key challenges and considerations for the modeling phase

Next Phase:

- Ready to proceed with model development and evaluation
- Positioned to create accurate and insightful predictive models