Exploratory Data Analysis of Real Estate Data

Key Insights and Findings

Presented by: Alham Hotaki

Date: 27.August.2024

Introduction

Objective

Provide an overview of the dataset and outline the goals of the analysis.

Dataset Overview

- **Entries:** 5,000 residential properties.
- **Columns:** 16 features, including `MLS`, `sold_price`, `bedrooms`, `bathrooms`, `lot_acres`, `taxes`, `year_built`, and more.

Analysis Goals

- Understand key factors influencing property prices.
- Prepare data for predictive modeling by addressing data quality issues.

Initial Observations

- Data Types: Several columns need type conversions.
- Missing Values: columns such as lot acres, square footage, garage, kitchen features and HOA should be handled.
- Potential Outliers: Columns like lot_acres show extreme values, suggesting potential outliers (e.g., maximum of 2154 acres).
- Categorical Features: The kitchen_features and floor_covering columns contain multiple values as comma-separated strings, which might need encoding.

Data Type Conversion Summary

Zipcodes: Converted from integers to strings to prevent incorrect numerical operations.

Year Built: Reformatted as year dates; zeros were replaced with NaN.

Fireplaces: Cleaned by replacing blank spaces with NaN and converted from strings to integers.

HOA Fees: Removed commas, replaced 'None' with NaN, and converted from strings to numeric values.

Data Validation Summary

To validate the data, we need to check for logical consistency across various columns. This involves ensuring that the values make sense in the context of the data and that there are no contradictions or anomalies.

Year Built: We checked that the years make sense—no future dates or unrealistic old years.

Bedrooms, Bathrooms, and Square Footage: We made sure the number of rooms and the home's size match up logically.

Lot Size: We verified that the lot sizes are realistic for homes, and fixed any extreme values.

Taxes: We checked that tax amounts are reasonable. e.g., no zero or negative taxes unless justified

Fireplaces: We made sure the number of fireplaces is realistic, with no negative numbers.

HOA Fees: We ensured that fees are appropriate. E.g., no negative values.

Handling Missing Values (1)

- MLS 0
- sold_price 0
- zipcode 0
- longitude 0
- latitude 0
- lot_acres 49
- taxes 0
- year_built 5
- bedrooms 0
- bathrooms 6
- sqrt_ft 56
- garage 7
- kitchen_features 33
- fireplaces 25
- floor_covering 1
- HOA 562

Strategies Employed

1. **Group-Based Imputation**: Used for columns like Lot Acres, Year Built, and Bathrooms to estimate missing values based on similar records.
2. **K-Nearest Neighbors (KNN) Imputation:** Applied for HOA Fees to predict missing values using related features.
3. **Simple Fill:** Replaced NaN values with appropriate placeholders, like 'No Features Listed' for Kitchen Features.

Handling Missing Values (2)

Group-Based Imputation Details

Lot Acres Imputation

- Group-Based Imputation by Zipcode, Bedrooms, and Sold Price
- Process: Calculated the mean Lot Acres within groups and used it to fill missing values.

Year Built Imputation

- Method: Group-Based Median Imputation by Zipcode and Sold Price
- Process: Replaced missing Year Built values with the median from similar properties.

Bathrooms Imputation

- Method: Group-Based Imputation by Bedrooms and Garage
- Process: Estimated missing Bathroom values based on the number of Bedrooms and Garage size.

Handling Missing Values (3)

Addressing Complex Missing Data

Square Footage

- Strategy: Introduced a new feature combining Bedrooms and Bathrooms to estimate missing Square Footage values.
- Process: Imputed missing values based on the average Square Footage of similar properties.

HOA Fees (Homeowners Association)

- Strategy: KNN Imputation
- Process: Used KNN to predict and impute missing HOA fees based on related property features like Sold Price, Year Built, and Lot Acres.

Fireplaces & Kitchen Features

- Fireplaces: Replaced missing values with 0, indicating no fireplaces.
- Kitchen Features: Filled missing values with 'No Features Listed', ensuring complete data for analysis.

Data Exploration

- Descriptive Statistics
- Distribution of Data
- Range of Data (detection of Outliers)
- Correlation Analysis
- Geospatial Analysis

Descriptive Statistics

General Observations

Sold Price: Ranges up to USD 5.3 million, indicating high-end properties.
Lot Acres: Maximum of 636.67 acres, far exceeding the 75th percentile, suggesting the presence of very large properties.
Taxes: Wide range from USD 0 to over USD 12 million, highlighting potential outliers or data entry errors.
Square Footage: Varies from 1,100 to 22,000 sq. ft., with a mean of 3,716 sq. ft., possibly including outliers.

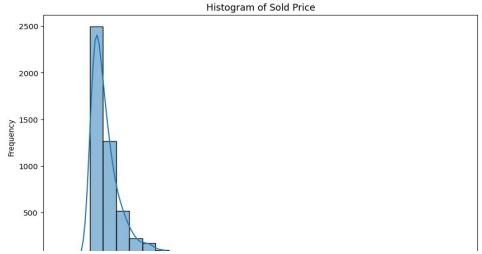
Key Insights

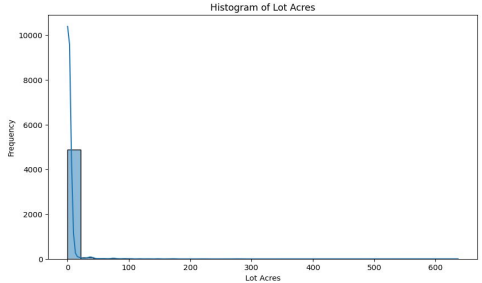
Year Built: Properties range from 1893 to 2019, with a median year of 1999, indicating a mix of older and newer homes.

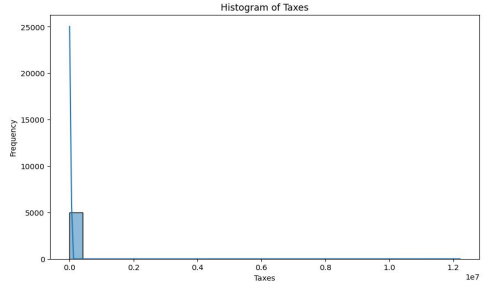
Bedrooms/Bathrooms: Median of 4 bedrooms and 4 bathrooms; maximum values suggest some luxury properties.

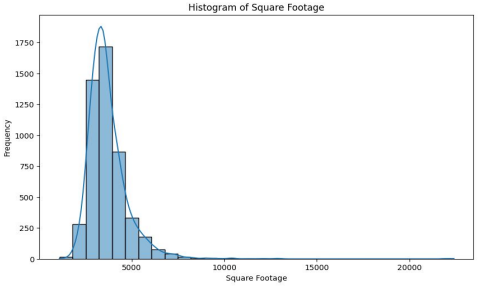
Fireplaces & Garages: Averages of 1.88 fireplaces and 2.81 garage spaces, typical for larger homes.

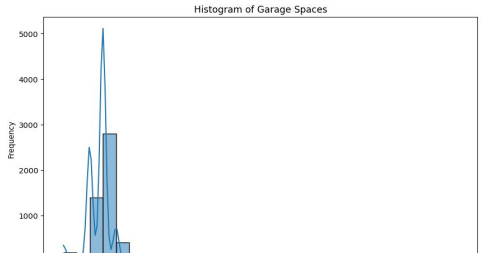
Distribution of Data

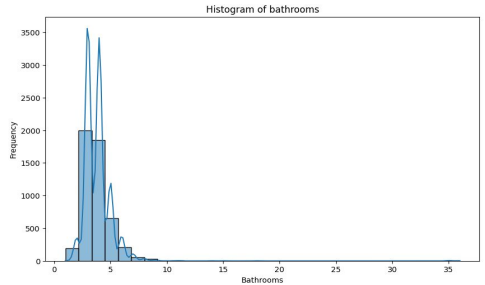


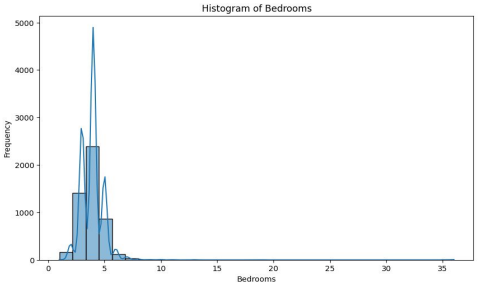


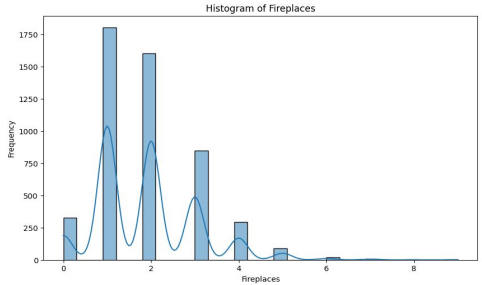


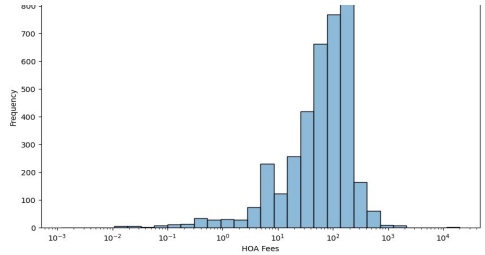












Distribution of Data

Sold Price: Right-skewed; most properties are under USD 1M, with a few high-end outliers.

Lot Acres: Majority have small lots; a few large properties up to 636 acres.

Taxes: Right-skewed; most properties have low taxes, with some extreme values.

Square Footage: Typically between 2,000-5,000 sq. ft.; some large, luxury properties.

Garage Spaces: Most have 1-3 spaces; a few outliers with up to 30.

Fireplaces: Commonly 1-2 fireplaces; multiple peaks indicate luxury homes.

HOA Fees: Right-skewed; most fees are low, with fewer high values.

Bedrooms: Centered around 3-4; few properties have more than 5.

Bathrooms: Mostly 2-4; more bathrooms indicate luxury properties.

Range of Data (detection of Outliers)

Why Outliers Matter?

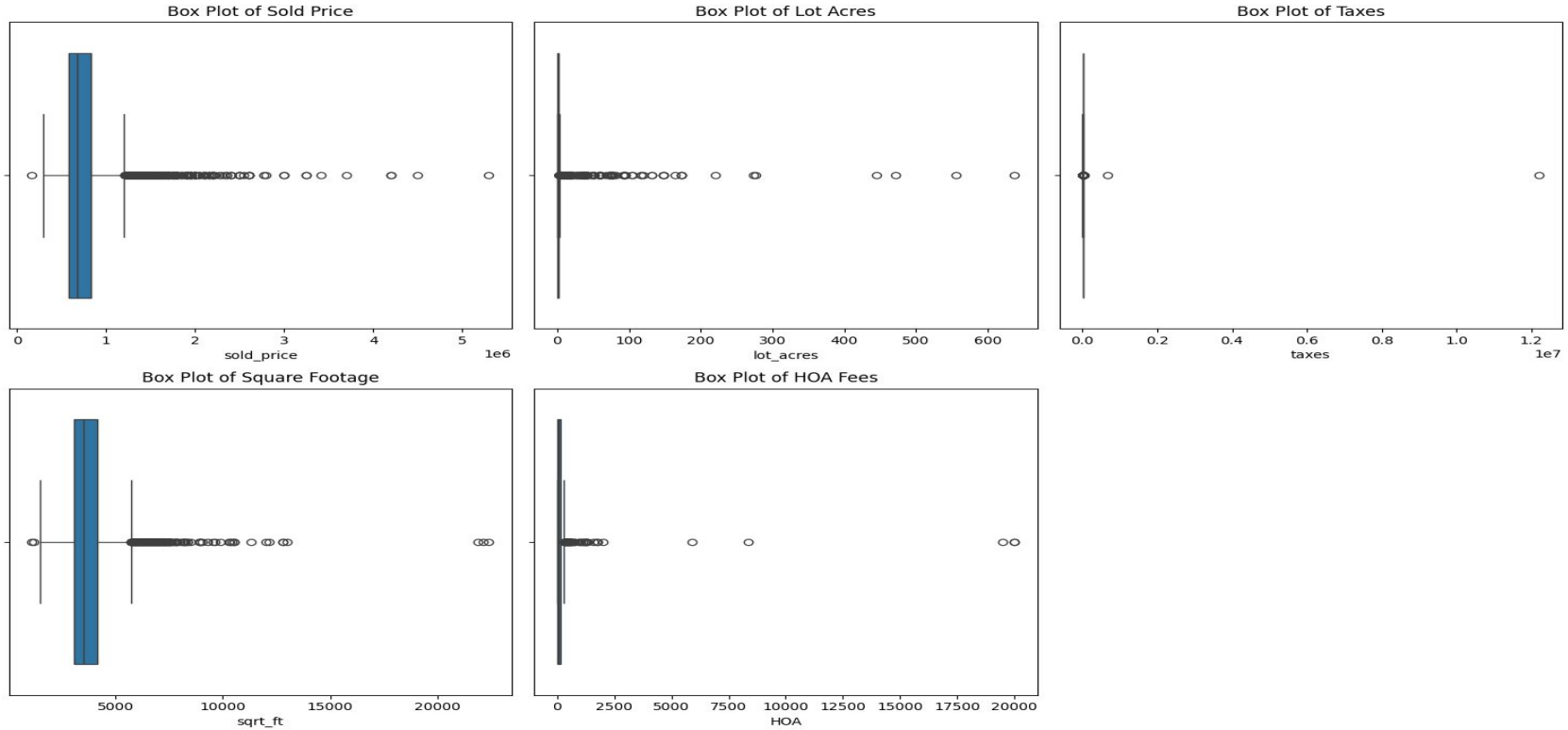
Impact: Outliers can skew analysis results and affect model performance.

Goal: Systematically identify and manage outliers to ensure robust analysis.

Outlier Handling Strategies:

1. **Remove Outliers:** For likely errors or extreme cases not representative of the data.
2. **Cap Outliers:** Replace outlier values with the nearest non-outlier value (e.g., at the 1st and 99th percentiles).
3. **Transform Data:** Use transformations like log to reduce the impact of outliers.

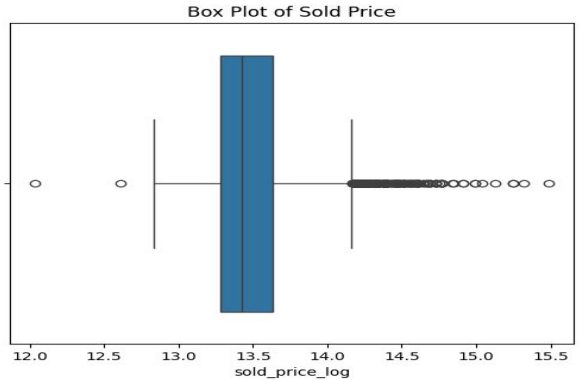
Range of Data (detection of Outliers)

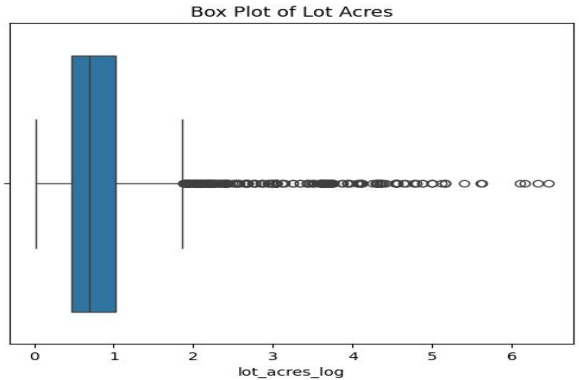


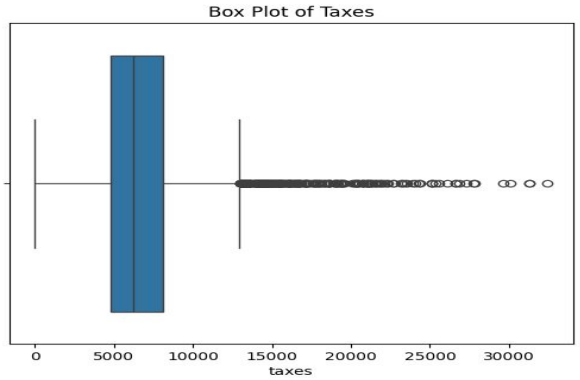
Range of Data (detection of Outliers)

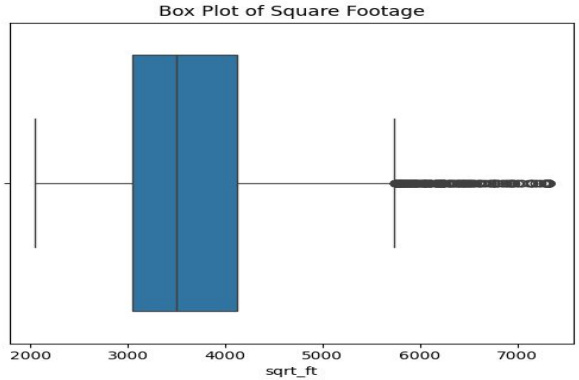
Sold Price: Significant outliers at the higher end, indicating luxury properties.
Strategy: Apply log transformation to reduce skewness.
Lot Acres: Few extreme outliers with large lot sizes.
Strategy: Log transformation and capping at the 99th percentile.
Taxes: Outliers with extremely high taxes, possibly errors.
Strategy: Manual review and correction of errors (e.g., misplaced decimal points).
Square Footage (sqrt_ft): Outliers in very large properties, potentially mansions or commercial buildings.
Strategy: Cap at a reasonable maximum (e.g., 99th percentile)
HOA Fees: Extreme outliers with very high fees.
Strategy: Cap at a certain threshold and manual review for accuracy.

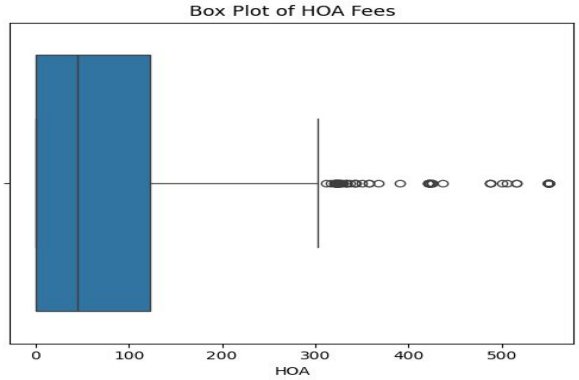
Range of Data (detection of Outliers)



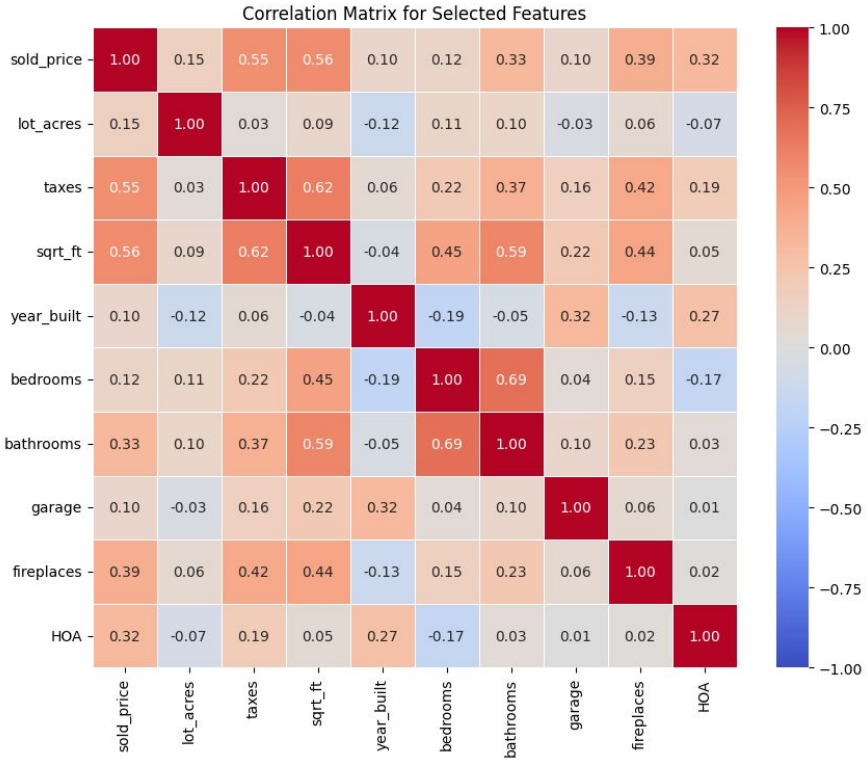


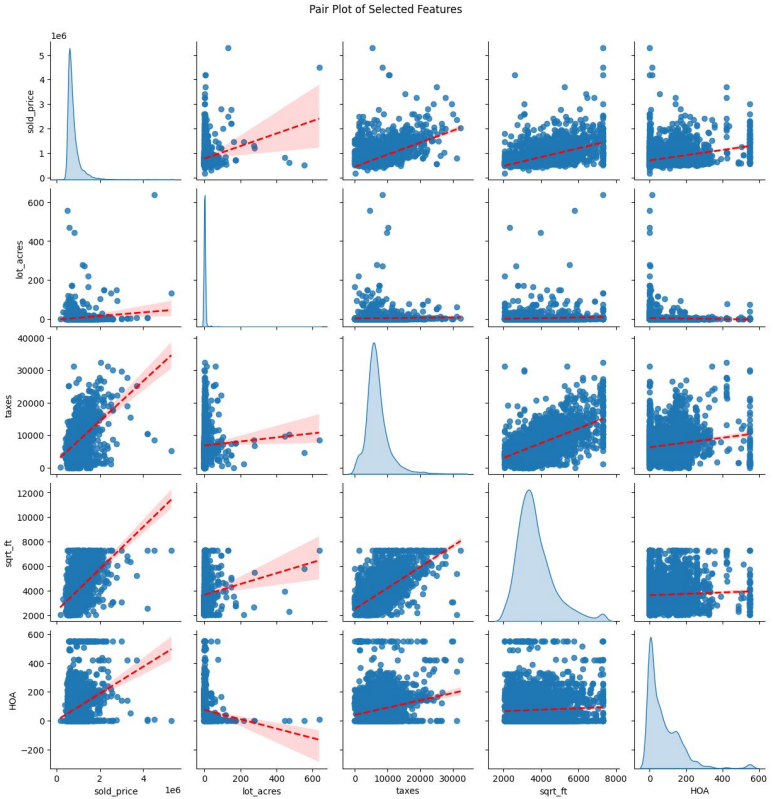




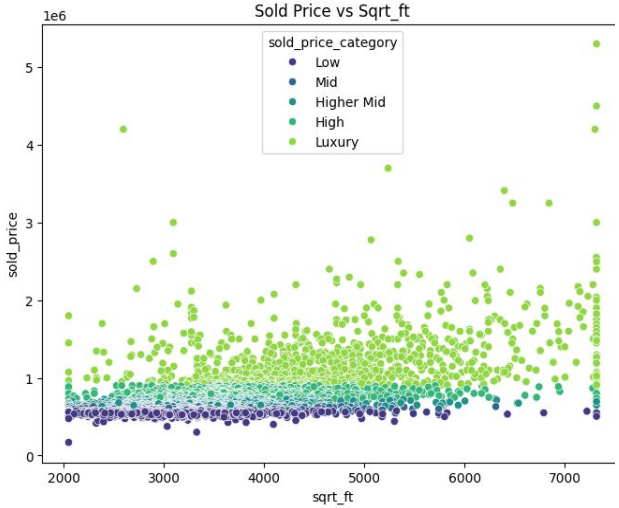


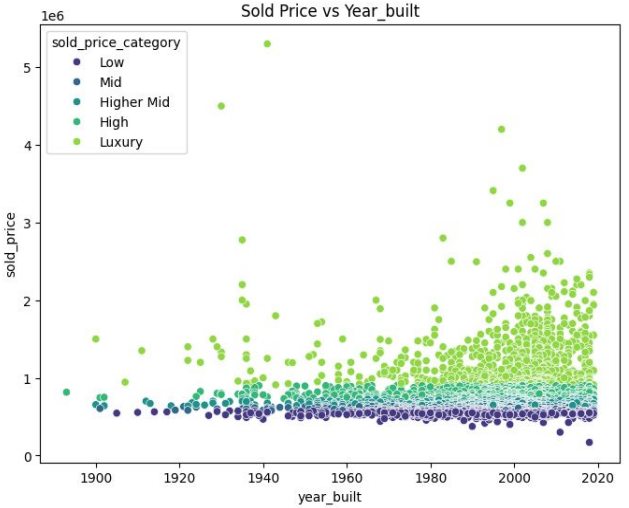
Correlation Analysis





Correlation Analysis





Correlation Analysis

Strong Predictors:

- Square Footage: Strong correlation (0.56) with higher sold prices.
- Taxes: Strong correlation (0.55); higher taxes often indicate more expensive homes.
- Fireplaces: Moderate correlation (0.39); more fireplaces usually mean higher prices.

Other Notable Correlations:

- Bathrooms: Moderately correlated (0.33) with sold price.
- HOA Fees: Moderate correlation (0.32); higher fees often link to higher property prices.

Modeling Strategy

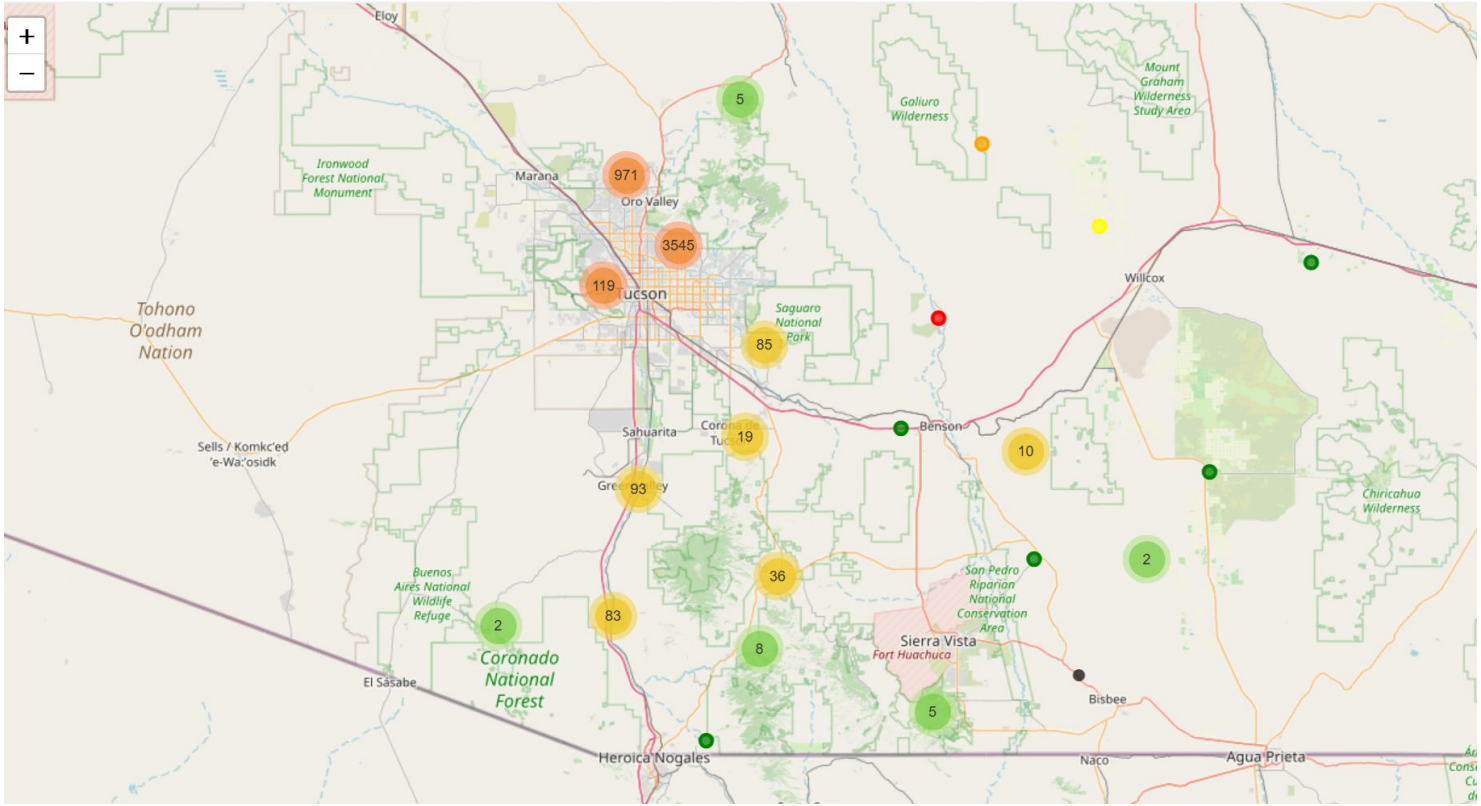
Handling Multicollinearity:

- Options: Remove one feature from highly correlated pairs, or apply PCA/regularization.
- New Composite Feature: Combined related features to simplify the model and reduce multicollinearity.

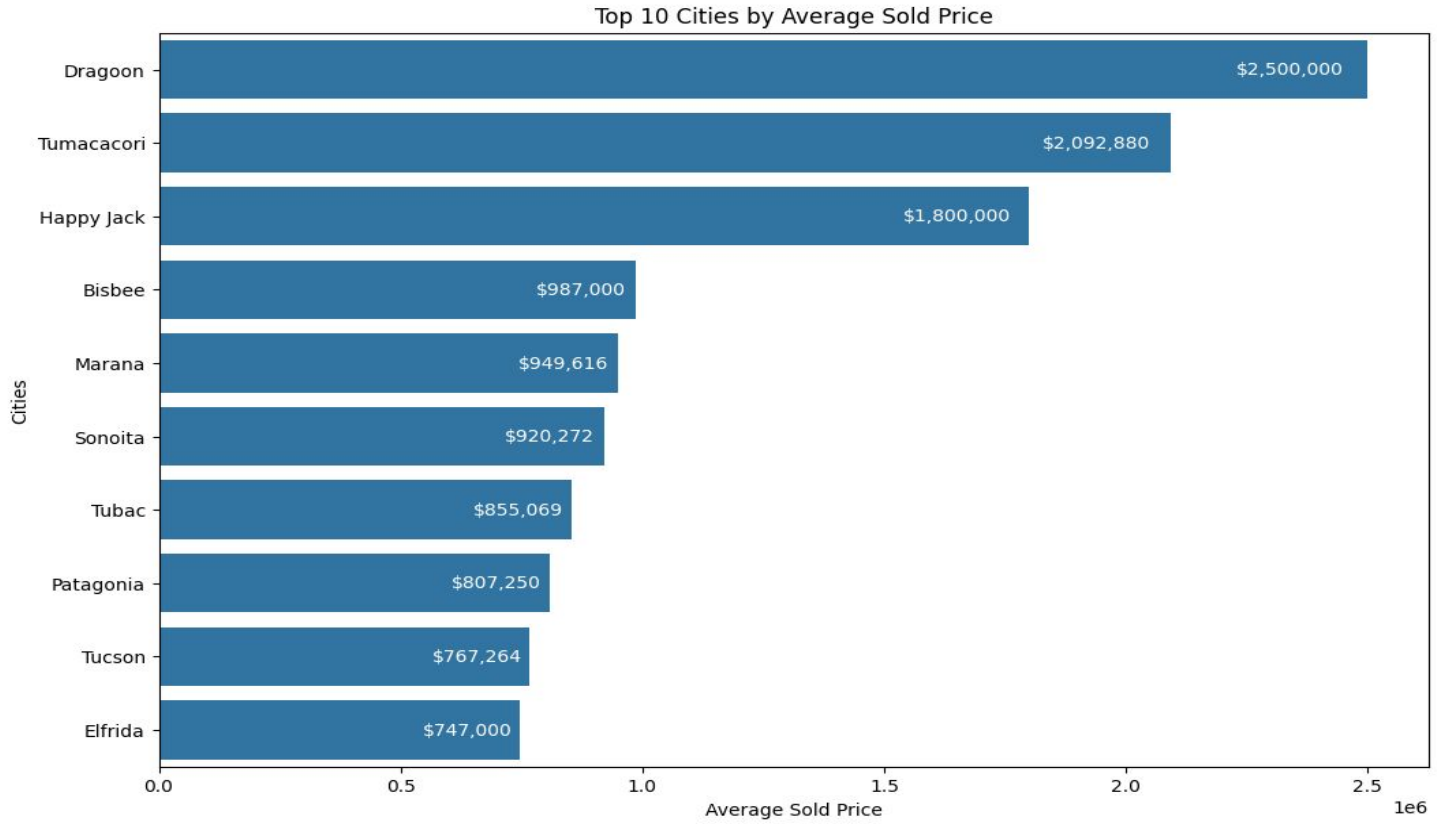
Focus for Prediction:

- Prioritize Square Footage, Taxes, Bathrooms, and Fireplaces in the model.
- Reevaluate or potentially exclude weaker features like Garage and Year Built based on model performance.

Geospatial Analysis



Geospatial Analysis



Geospatial Analysis

Data Integration:
Merged the main dataset with an additional dataset to map zip codes to city names, enhancing geographic analysis.

Interactive Map Creation:

- Used Folium to visualize property data centered on the dataset's average geographic coordinates.
- Employed marker clustering to manage and display numerous data points effectively.

Marker Representation:

- Properties are shown as color-coded markers based on pricing tiers.
- Interactive popups provide detailed information for each location.

Key Insights:

- Identified top cities for luxury homes.
- Observed geographic trends in property values across different regions.

Complementary Visualization:

- Bar plot highlighting the top 10 cities by average sold price.
- Provides a clear view of regions with the highest property values.

Thank You