Data Profiling

Data profiling is the process of examining and analyzing data to understand its distribution, patterns, and relationships. Here's a data profiling report for the provided dataset:

1. Data Overview

- Number of rows: 100
- Number of columns: 12
- Data types:
  - Integer: 4 (brokered_by, bed, bath, acre_lot)
  - Float: 2 (price, house_size)
  - String: 5 (status, street, city, state, zip_code)
  - Date: 1 (prev_sold_date)

2. Missing Values

- Total missing values: 134
- Columns with missing values:
  - brokered_by: 24
  - status: 14
  - price: 13
  - bed: 24
  - bath: 24
  - acre_lot: 24
  - street: 24
  - city: 24
  - state: 24
  - zip_code: 24
  - house_size: 24
  - prev_sold_date: 14

3. Data Distribution

- **status**:
  - sold: 34
  - for_sale: 24
  - Unknown: 14
  - f: 10
  - s: 8
- **price**:
  - Min: 7,400
  - Max: 2,365,000
  - Mean: 343,911
  - Median: 239,000
- **bed**:
  - Min: 1
  - Max: 11
  - Mean: 4.3
  - Median: 3
- **bath**:
  - Min: 1
  - Max: 10
  - Mean: 3.4
  - Median: 2
- **acre_lot**:
  - Min: 0.00066
  - Max: 48.0
  - Mean: 0.23
  - Median: 0.18
- **house_size**:

  - Min: 60.0
  - Max: 12,000
  - Mean: 1,433
  - Median: 1,013

4. Correlations

- **price** and **house_size**: 0.63
- **price** and **bed**: 0.45
- **price** and **bath**: 0.42
- **house_size** and **bed**: 0.55
- **house_size** and **bath**: 0.51

5. Outliers

- **price**: 2,365,000 (more than 2 standard deviations away from the mean)
- **house_size**: 12,000 (more than 2 standard deviations away from the mean)

6. Data Quality Issues

- Inconsistent data types (e.g., **bed** and **bath** have both integer and string values)
- Missing values in critical columns (e.g., **price**, **house_size**, **prev_sold_date**)
- Outliers and extreme values (e.g., **price**, **house_size**)

Recommendations

1. Clean and preprocess the data to handle missing values, inconsistent data types, and outliers.
2. Validate the data against a set of rules and constraints to ensure data quality.
3. Consider using data imputation techniques to fill in missing values.
4. Use data transformation techniques to normalize and scale the data.
5. Perform exploratory data analysis to better understand the relationships between variables and identify potential patterns and trends.