Exploratory Data Analysis Report

House Prices Dataset

Data Analysis Project

November 12, 2025

Abstract

This report presents an exploratory data analysis of the Kaggle House Prices dataset containing 1,460 residential homes in Ames, Iowa with 79 explanatory variables. Through systematic data cleaning and visualization, we identified that above-ground living area, overall quality, and year built are strong predictors of sale price, while lot area shows surprisingly weak correlation.

1 Introduction

The Kaggle House Prices dataset provides detailed information about residential properties in Ames, Iowa, with 79 variables ranging from basic metrics like square footage to nuanced features such as material quality and architectural styles. This analysis aims to identify which features most strongly relate to house prices, understand pricing patterns, and uncover surprising relationships in the data.

The dataset contains 1,460 houses with 38 numerical and 43 categorical features. The target variable `SalePrice` ranges from \$34,900 to \$755,000, with a mean of \$180,921 and median of \$163,000. The difference between mean and median indicates a right-skewed distribution, typical of real estate markets where a few luxury properties pull the average upward.

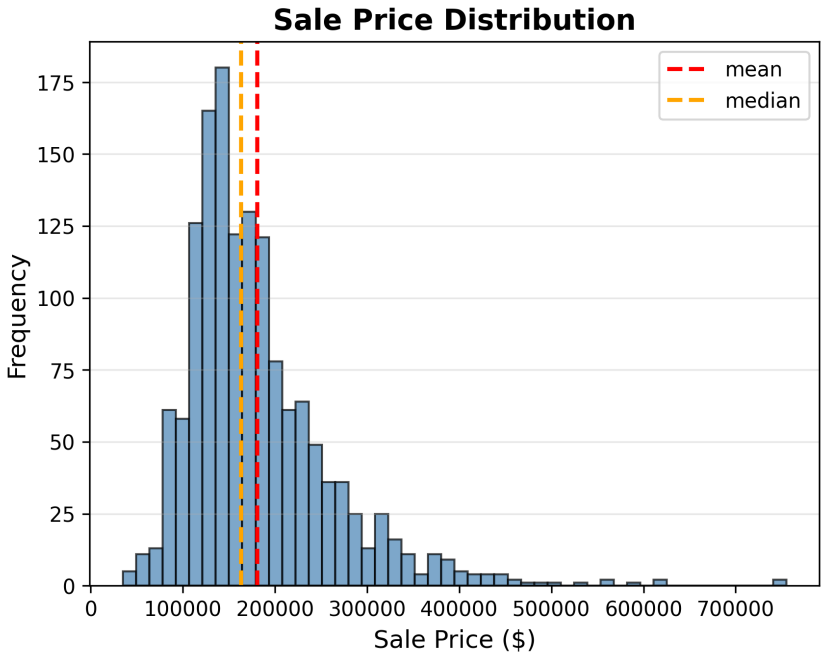

Figure 1: Sale Price Distribution showing right-skewed pattern with mean and median lines

2 Data Cleaning Methodology

2.1 Addressing Missing Values

Missing value analysis revealed that 19 columns contained gaps, with some showing extreme missingness. Notably, `PoolQC` was missing 99.52% of values, `MiscFeature` 96.30%, `Alley` 93.77%, and `Fence` 80.75%. Rather than viewing this as a data quality problem, we recognized that missingness often represents genuine feature absence—houses without pools naturally lack pool quality ratings.

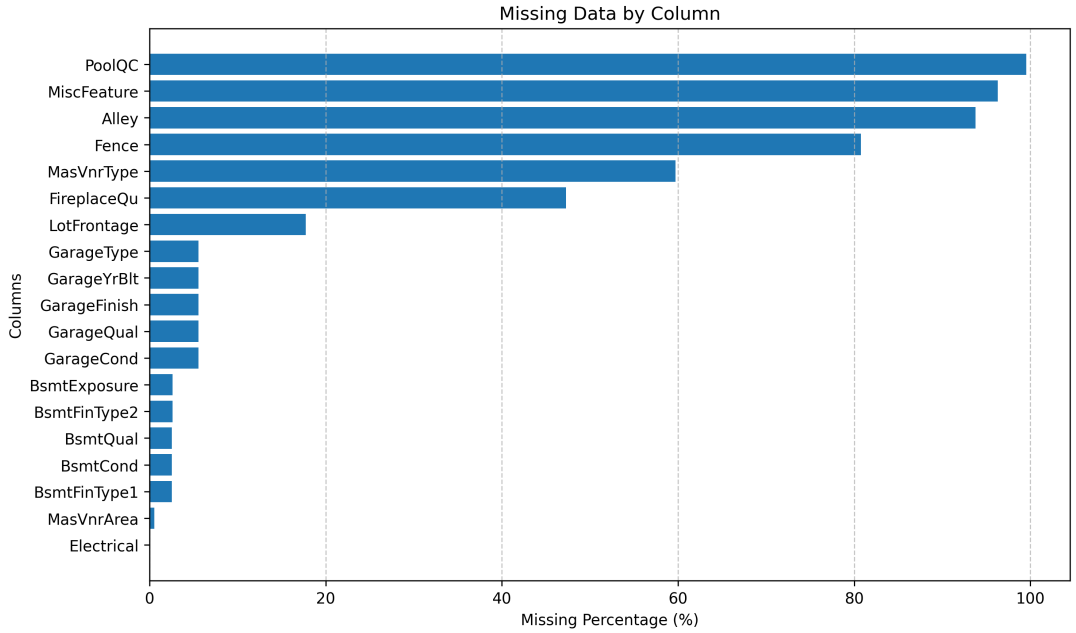

Figure 2: Missing data percentages by column

We strategically removed six columns with extreme missingness: `PoolQC`, `PoolArea`, `MiscFeature`, `MiscVal`, `Alley`, and `Fence`. These features apply to so few properties that retaining them would provide minimal predictive value while potentially introducing noise into models.

2.2 Intelligent Imputation

For `MasVnrType` (masonry veneer type), we filled missing values with "None" rather than using statistical imputation. This explicit labeling respects the semantic meaning—missing veneer type indicates no veneer at all. This approach enables tree-based models to learn that "having versus not having veneer" is a meaningful distinction.

2.3 Encoding Quality Features

Many features use ordinal quality ratings: Ex (Excellent), Gd (Good), TA (Typical), Fa (Fair), and Po (Poor). We converted these text ratings to numerical form using the mapping: None=0, Po=1, Fa=2, TA=3, Gd=4, Ex=5. This preserves the inherent ordering while making features mathematically usable. Before encoding, we filled missing values with "None" to represent absent features. We excluded `BsmtExposure` from this encoding as it uses a different rating system. This transformation affected features like `KitchenQual`, `FireplaceQu`, and various basement quality measures, converting them from categorical strings to ordinal integers suitable for modeling.

	ExterQual	ExterCond	BsmtQual	BsmtCond	HeatingQC	KitchenQual	FireplaceQu	GarageQual	GarageCond
0	Gd	TA	Gd	TA	Ex	Gd	NaN	TA	TA
1	TA	TA	Gd	TA	Ex	TA	TA	TA	TA
2	Gd	TA	Gd	TA	Ex	Gd	TA	TA	TA
3	TA	TA	TA	Gd	Gd	Gd	Gd	TA	TA
4	Gd	TA	Gd	TA	Ex	Gd	TA	TA	TA
5	TA	TA	Gd	TA	Ex	TA	NaN	TA	TA
6	Gd	TA	Ex	TA	Ex	Gd	Gd	TA	TA
7	TA	TA	Gd	TA	Ex	TA	TA	TA	TA
8	TA	TA	TA	TA	Gd	TA	TA	Fa	TA
9	TA	TA	TA	TA	Ex	TA	TA	Gd	TA

Quality features before encoding

	ExterQual	ExterCond	BsmtQual	BsmtCond	HeatingQC	KitchenQual	FireplaceQu	GarageQual	GarageCond
0	4	3	4	3	5	4	0	3	3
1	3	3	4	3	5	3	3	3	3
2	4	3	4	3	5	4	3	3	3
3	3	3	3	4	4	4	4	3	3
4	4	3	4	3	5	4	3	3	3
5	3	3	4	3	5	3	0	3	3
6	4	3	5	3	5	4	4	3	3
7	3	3	4	3	5	3	3	3	3
8	3	3	3	3	4	3	3	2	3
9	3	3	3	3	5	3	3	4	3

Quality features after encoding

3 Exploratory Findings

3.1 Distribution Analysis

Histograms of key features revealed important patterns. `GrLivArea` (above-ground living area) clusters between 1,000-2,000 square feet with a right skew toward larger homes, mirroring the sale price distribution. `LotArea` shows extreme right skewness with most lots under 15,000 square feet but a few extending beyond 100,000 square feet—these could be outliers or rural properties. `OverallQual` shows a roughly normal distribution centered around quality level 5-6, indicating most homes are average to slightly above-average quality.

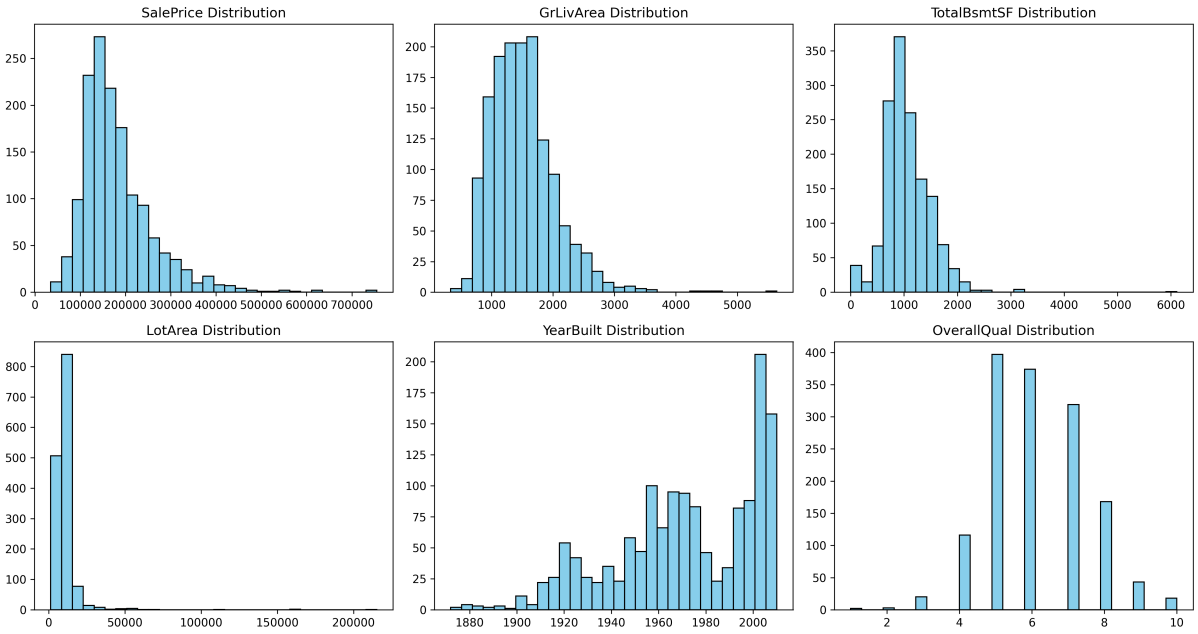

Figure 3: Distribution histograms for key features

3.2 Relationship with Sale Price

Scatter plots revealed varying relationship strengths with `SalePrice`. `GrLivArea` demonstrates the strongest positive linear relationship—as living area increases, price increases proportionally. However, we identified outliers: extremely large houses (over 4,000 square feet) selling for surprisingly low prices, potentially representing distressed sales or data anomalies.

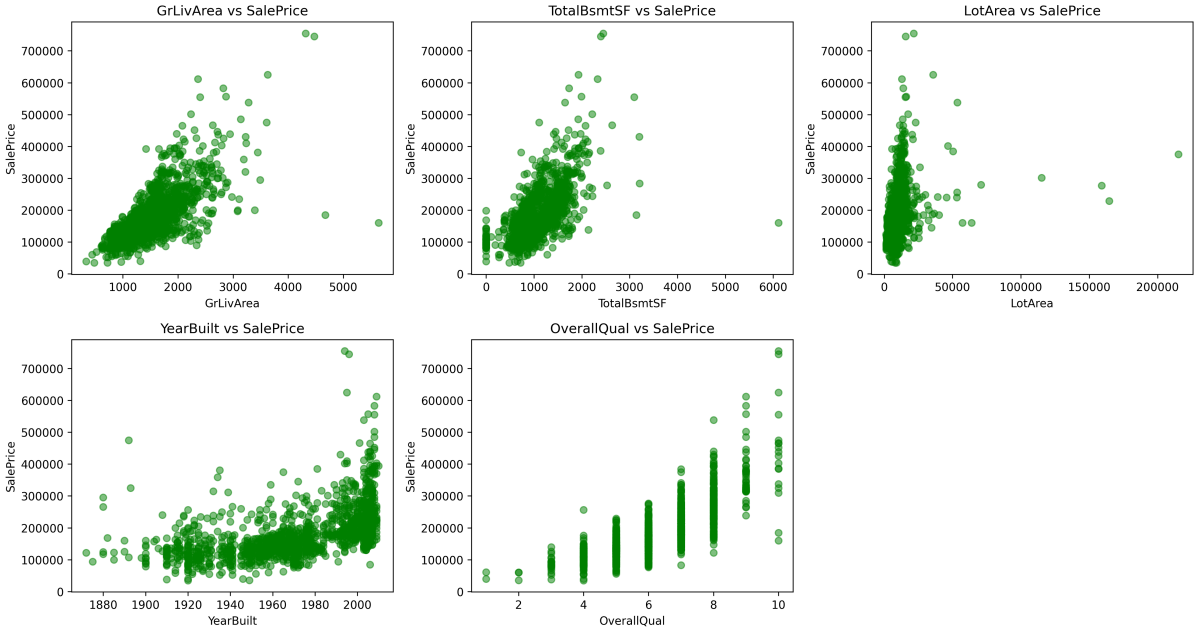

Figure 4: Scatter plots showing relationships between key features and Sale Price

`OverallQual` shows perhaps the clearest relationship, with distinct price separation between quality levels. Homes rated 8 or above consistently command premium prices, while those rated 4 or below cluster in lower ranges. This suggests subjective quality assessments effectively

capture multiple value dimensions beyond raw metrics.

`YearBuilt` shows moderate correlation, with newer homes generally selling for more, though well-maintained older homes can still command high prices due to location or architectural charm. `TotalBsmtSF` demonstrates positive but weaker correlation than living area, likely because basement space is valued less than above-ground space.

Most surprisingly, `LotArea` shows weak, inconsistent relationship with sale price. Despite conventional wisdom that land drives value, many expensive homes sit on modest lots while properties with very large lots sell for moderate prices. This suggests buyers in Ames prioritize home quality and size over land area—possibly reflecting abundant land availability in Iowa or zoning constraints that limit lot size variability.

4 Key Insights

Strong Predictors: `GrLivArea`, `OverallQual`, and `YearBuilt` emerged as robust price predictors. Living area shows consistent positive correlation across the price spectrum, while quality ratings effectively capture holistic value including materials, craftsmanship, and finish.

Surprising Findings: Lot area's weak predictive power contradicts conventional real estate wisdom. In this market, house characteristics—size, quality, condition—matter far more than land area. The presence of outliers (very large homes with low prices) warrants investigation for distressed sales or unique circumstances.

Distribution Characteristics: The right-skewed sale price distribution suggests median is more representative than mean, and predictive models should consider log transformation to normalize the target variable.

Categorical Features: Encoded quality features like `KitchenQual` and `FireplaceQu` show clear value relationships. Homes with excellent features command premiums, while those with absent or poor-quality features cluster lower, validating our encoding strategy.

5 Conclusion

This analysis identified key drivers of residential property values in Ames, Iowa through systematic data cleaning and visualization. Above-ground living area, overall quality ratings, and construction year are the strongest predictors, while lot size proved surprisingly weak. Our methodology of explicitly encoding absent features and converting ordinal ratings to numerical scales prepared the dataset for robust predictive modeling. These insights will inform feature selection, transformation strategies, and model evaluation in subsequent analysis phases, ultimately enabling accurate house price predictions.