HOUSE PRICES EDA REPORT

Exploratory Data Analysis

Kaggle House Prices Competition

Generated: 2025-11-09

EXECUTIVE SUMMARY

DATA OVERVIEW:

Samples: 1,460

Features: 81

Price Range: $34,900 – 755,000

DATA QUALITY:

Columns with Missing Values: 19

Highest Missing: PoolQC (99.52%)

KEY FINDINGS:

OverallQual: Correlation 0.791

GrLivArea: Correlation 0.709

GarageCars: Correlation 0.640

CONCLUSIONS:

- Target variable is right-skewed, log transform recommended

- Strong correlated features identified for modeling

- Missing values need to be handled

DATASET OVERVIEW

BASIC INFORMATION:

- Data Shape: (1460, 81)
- Numeric Features: 38
- Categorical Features: 43
- Memory Usage: 3.4 MB

TARGET STATISTICS:

- Mean: $180,921
- Median: $163,000
- Standard Deviation: $79,443
- Skewness: 1.883

MISSING VALUES ANALYSIS

SUMMARY:

- Columns with Missing Values: 19
- Total Missing Values: 7829
- Average Missing %: 28.2%

TOP MISSING COLUMNS:

1. PoolQC: 99.52%
2. MiscFeature: 96.3%
3. Alley: 93.77%
4. Fence: 80.75%
5. MasVnrType: 59.73%
6. FireplaceQu: 47.26%
7. LotFrontage: 17.74%
8. GarageYrBlt: 5.55%

TARGET VARIABLE ANALYSIS

STATISTICS:

- Skewness: 1.883
- Kurtosis: 6.536
- Coefficient of Variation: 0.439
- Recommendation: Log transformation

DISTRIBUTION NOTES:

- Distribution is right-skewed
- Few high-priced houses pull mean upward
- Log transformation will help

CORRELATION ANALYSIS

CORRELATION STRENGTH:

- Strong (>0.5): 11 features
- Moderate (0.3-0.5): 4 features
- Weak (<0.3): 0 features

TOP FEATURES:

1. OverallQual: 0.791

   (Very Strong)
2. GrLivArea: 0.709

   (Very Strong)
3. GarageCars: 0.640

   (Strong)
4. GarageArea: 0.623

   (Strong)
5. TotalBsmtSF: 0.614

   (Strong)

FEATURE ENGINEERING

NEW FEATURES:

- HouseAge
- TotalArea
- HasPool
- TotalBath
- log_SalePrice
- log_GrLivArea
- log_LotArea

CONCLUSIONS AND RECOMMENDATIONS

MAIN CONCLUSIONS:

- Good data quality with some missing values
- Target variable needs log transformation
- Strong features identified for modeling
- Feature engineering can improve performance

NEXT STEPS:

1. Advanced feature engineering
2. Handle categorical variables
3. Build predictive models
4. Model evaluation and tuning
5. Results interpretation

Report completed