Exploratory Data Analysis (EDA) for Real Estate Pricing

Company: Next Hikes
Tools Used: Python, Pandas, NumPy, Seaborn, Matplotlib, Plotly
Author: Ankit

Introduction

The real estate market is influenced by multiple factors such as property size, quality, age, amenities, and location.
This project performs **Exploratory Data Analysis (EDA)** on the Ames Housing dataset to identify the key variables that drive **house sale prices**.

The analysis uncovers:

- Patterns in the dataset
- Relationships between variables
- Size & amenity impact
- Historical pricing variations
- Correlations between important features

This helps stakeholders make data-informed decisions in pricing and valuation.

Dataset Overview

The Ames Housing dataset contains **1460 rows** and **80 columns**, including:

- Structural features (GrLivArea, TotalBsmtSF, OverallQual)
- Size features (Bedrooms, Bathrooms, TotalSF)
- Amenities (Garage, Pool, Fireplace)
- Time features (YrSold, MoSold)
- Target variable: **SalePrice**

The dataset provides a rich set of attributes to understand how various home characteristics impact price.

Data Loading

The dataset was imported using **Pandas** for easy manipulation.
df.head() and df.info() were used to inspect the first few rows and understand data types.

Data Cleaning

Data cleaning was performed to ensure accuracy and reliability.

Steps taken:

- **Missing values handled:**
    - Mode for categorical columns (e.g., Electrical)
    - “None” for absent features (e.g., MasVnrType)
    - Dropped extremely sparse columns (e.g., Alley > 80% null)
- **Duplicates removed**
- **Outliers identified:**
    - Very large houses (>4000 sqft) were retained as luxury homes
- **Data types corrected** where needed

This ensures the dataset is ready for detailed analysis.

Univariate Analysis

The distribution of individual variables was explored.

Key Plots:

- **SalePrice Histogram:**
  Right-skewed distribution with few high-value luxury houses.
- **GrLivArea Histogram:**
  Right-skewed, consistent with SalePrice pattern.
- **OverallQual Countplot:**
  Quality ratings 5–7 are most common.
- **LotShape Countplot:**
  Slight preference for Regular lot shape.

Insights:

Univariate analysis shows strong variation in prices and house characteristics, indicating multiple drivers behind valuation.

Bivariate Analysis

Key relationships analyzed:

📌 **1. GrLivArea vs SalePrice**

Strong positive correlation — bigger houses sell for higher prices.

📌 **2. OverallQual vs SalePrice**

The strongest categorical predictor of price.

📌 **3. LotShape vs SalePrice**

Lot shape has limited but noticeable influence.

Insights:

Size and construction quality show the clearest relationships with pricing.

Correlation Analysis

A **Top 10 Correlation Heatmap** was created to highlight the strongest predictors.

Top correlated features with SalePrice:

- OverallQual
- GrLivArea
- TotalSF
- GarageCars
- GarageArea
- TotalBsmtSF

Insights:

Structural metrics and quality dominate price prediction.

Feature Engineering

To improve analysis clarity, three new features were created:

- **PricePerSF = SalePrice / GrLivArea**
- **HouseAge = YrSold – YearBuilt**
- **TotalSF = GrLivArea + TotalBsmtSF**

These engineered features help offer deeper insight into valuation.

Size Impact Analysis

Plots included:

- **Bedrooms vs SalePrice**
- **Bathrooms vs SalePrice**
- **TotalSF vs SalePrice**

Insight:

Bedrooms alone do not strongly influence price, but **TotalSF** and overall size show strong upward price trends.

Market Trends

Average SalePrice by YrSold

A line plot of year-wise pricing trends shows:

Insight:

House prices remain mostly stable across years with slight fluctuations — indicating a consistent market during the recorded period.

Customer Preferences & Amenities Impact

Plots used:

- **GarageArea vs SalePrice**
- **PoolArea vs SalePrice**

Insights:

- Larger garages increase house prices significantly.
- Houses with pool area fall in premium price ranges, though pools are rare.

AI Integration

AI was used to enhance the workflow:

- Helped choose clean and consistent color palettes
- Suggested meaningful visualizations
- Guided selection of key correlated features
- Supported insight summarization and report structuring

Final Insights

Major Findings:

- **OverallQual** is the strongest predictor of price
- **TotalSF** & **GrLivArea** show strong linear relationship with price
- **Garage & pool amenities** increase valuation
- Bedrooms count has low predictive power
- Market prices remain stable across years
- Feature engineering improved interpretability

Conclusion

The EDA reveals that **quality and size** are the primary drivers of home values, followed by amenities like garage and pool area. The analysis provides valuable insights for real estate pricing strategies, supporting accurate valuation, business decisions, and customer targeting.