House Price Prediction Analysis

A Data-Driven Approach to Real Estate Valuation by using Python

Mohammed Thoufeeq

Introduction

The real estate market is a critical sector of the economy, significantly influencing financial stability for individuals and investment firms alike. Accurately valuing residential property is a complex challenge, as house prices are influenced by a multitude of distinct factors—ranging from the physical attributes of the property (such as lot area, number of rooms, and building type) to external factors like location zoning and market conditions. Traditional valuation methods often rely on subjective appraisals, which can lead to inconsistencies and inefficiencies in the market.

This project utilizes the "House Price Prediction" dataset, which includes 2,919 records and over 80 explanatory variables describing (almost) every aspect of residential homes. The methodology follows a structured Data Science lifecycle:

- **Data Preprocessing:** Handling missing values and encoding categorical variables.
- **Exploratory Data Analysis:** Analyzing the distribution of the target variable (SalePrice) and its relationship with independent variables.
- **Feature Engineering:** Creating new predictive features (such as HouseAge) to improve model interpretability.
- **Statistical Analysis:** Quantifying the impact of features like zoning and renovation status on property value.

Methodology

To achieve the project objectives, a structured **Data Science Lifecycle** was adopted, leveraging Python for computation and visualization. The methodology was divided into four distinct phases: Data Preprocessing, Exploratory Data Analysis (EDA), Feature Engineering, and Statistical Evaluation.

Tools and Environment The analysis was conducted using the **Python** programming language within a Jupyter Notebook environment. The following libraries were utilized:

- **Pandas:** For data manipulation, cleaning, and aggregation.
- **NumPy:** For numerical computations and logarithmic transformations.
- **Matplotlib & Seaborn:** For data visualization, including scatter plots, box plots, and heatmaps.
- **Scikit-Learn:** For preprocessing tasks such as scaling and encoding.

Data Preprocessing & Cleaning

The raw dataset contained 2,919 records and required significant cleaning to ensure data quality.

- **Handling Missing Values:** A detailed null-value analysis revealed missing data in the Test Set for features like MSZoning, TotalBsmtSF, and Exterior1st.
  - *Numerical Features:* Missing values in continuous variables (e.g., Basement Area) were imputed using the **Median** to minimize the impact of outliers.
  - *Categorical Features:* Missing values in nominal variables (e.g., Zoning) were imputed using the **Mode** (most frequent category).
- **Outlier Detection:** Boxplots and the Interquartile Range (IQR) method were used to identify extreme outliers in LotArea and SalePrice, which were flagged for potential treatment to prevent model skew.

Exploratory Data Analysis (EDA)

Comprehensive EDA was performed to uncover underlying patterns and relationships.

- **Univariate Analysis:** Histograms and distribution plots were generated for key variables. It was observed that the target variable, SalePrice, followed a right-skewed distribution (Skewness > 1.8), necessitating transformation.
- **Bivariate Analysis:** Scatter plots were used to examine the relationship between continuous features (e.g., YearBuilt, TotalBsmtSF) and SalePrice.
- **Correlation Analysis:** A Pearson Correlation Matrix was computed to quantify linear relationships. A heatmap visualization highlighted that TotalBsmtSF (0.61) and YearBuilt (0.52) had the strongest positive correlations with price.

Feature Engineering

To enhance the predictive power of the dataset, raw features were transformed into more meaningful representations.

- **Temporal Feature Extraction:** Raw calendar years (YearBuilt, YearRemodAdd) were converted into duration-based features:
  - HouseAge = (Current Year - YearBuilt)
  - YearsSinceRemodel = (Current Year - YearRemodAdd)
- **Logarithmic Transformation:** To address the high skewness identified during EDA, a **Log(1+x) transformation** was applied to SalePrice and LotArea. This normalized the data distribution, satisfying the normality assumption required for many regression algorithms.
- **Categorical Encoding:** Nominal variables (e.g., BldgType, Neighborhood) were converted into numerical format using **One-Hot Encoding**, creating binary dummy variables to allow mathematical modeling of categorical data.

Statistical Validation

Statistical tests were conducted to validate hypotheses regarding price drivers.

- **Group-By Analysis:** Aggregated statistics (Mean/Median) were calculated to compare price differences across categories (e.g., comparing CulDSac vs. Inside lots).
- **Significance Testing:** Correlation coefficients and regression slopes were analyzed to confirm that features like "Basement Size" and "Overall Condition" have a statistically significant impact on property value.

Requirement Analysis

This section outlines the functional, non-functional, and technical requirements necessary to successfully execute the House Price Prediction project. These requirements serve as the constraints and specifications that guided the development of the analysis pipeline.

Data Requirements

To generate accurate price predictions, the project requires a comprehensive historical dataset with specific characteristics:

- **Source Data:** A structured dataset (CSV/Excel) containing residential real estate transactions (e.g., HousePrediction.xlsx).
- **Feature Completeness:** The data must include a mix of:
  - *Quantitative Features:* Continuous variables such as Lot Area, Total Basement Square Footage, and Year Built.
  - *Qualitative Features:* Categorical variables such as Zoning Classification (MSZoning), Building Type (BldgType), and Overall Condition.
- **Target Variable:** A clearly defined target column (SalePrice) for supervised learning.
- **Data Quality:** The dataset must be robust enough to handle noise. The system requires a minimum of 1,000+ records to ensure statistical significance during correlation analysis.

Functional Requirements

The analytical system (Python Pipeline) must perform the following core functions:

- **Data Ingestion:** The ability to read and parse flat files (CSV) and handle varying data types (integers, floats, strings) automatically.
- **Error Handling (Imputation):** The system must identify missing values (NaNs) and apply appropriate imputation strategies (Median for numerical, Mode for categorical) without discarding valuable training data.
- **Statistical Computation:** The system must be capable of calculating descriptive statistics (Mean, Median, Std Dev) and complex relationships (Pearson Correlation Coefficients) to identify price drivers.
- **Visualization Generation:** The system must generate interpretable visual outputs—specifically Boxplots for outlier detection, Heatmaps for correlation matrices, and Scatter Plots for trend analysis—to allow stakeholders to visually assess market dynamics.

Technical & Software Requirements

The project implementation relies on a specific technical stack to ensure reproducibility and performance:

- **Runtime Environment:** Python 3.x (via Jupyter Notebook or Google Colab).
- **Key Libraries & Dependencies:**
  - *Pandas:* For high-performance dataframe manipulation.
  - *NumPy:* For mathematical transformations (Log-transformations).
  - *Matplotlib/Seaborn:* For static statistical data visualization.
  - *Scikit-Learn:* For preprocessing utilities (StandardScaler, OneHotEncoder).
- **Hardware Constraints:** A standard computing environment with sufficient RAM (minimum 4GB) to load the dataset into memory and perform vectorized operations.

Non-Functional Requirements

- **Interpretability:** The analysis results must be understandable by non-technical stakeholders (buyers/investors). Features created (like HouseAge) must have real-world business logic.
- **Scalability:** The preprocessing pipeline should be written efficiently so it can handle larger datasets (e.g., 50,000+ rows) in future iterations without code refactoring.
- **Accuracy:** The final data transformation steps must normalize skewed distributions (e.g., Log Transformation of Price) to maximize the potential accuracy of future predictive models.

Parameters & Project Assumptions

This section outlines the specific parameters, statistical thresholds, and assumptions defined to tailor the analysis specifically for the House Price Prediction domain.

Temporal Reference Parameter

- **Reference Year (Y_{ref}): 2011**
  - *Rationale:* The dataset contains construction dates (YearBuilt) up to 2010. To calculate the "House Age" and "Years Since Remodel" without generating negative values or data leakage, the year 2011 was established as the static current year for all age-based calculations.
  - *Formula:* $\text{House Age} = 2011 - \text{YearBuilt}$

Statistical Thresholds

- **Significance Level (α): 0.05**
  - Used during the Pearson Correlation and Linear Regression analysis to determine if a feature (like TotalBsmtSF) had a statistically significant impact on SalePrice. A P-value < 0.05 confirmed the relationship was not due to random chance.
- **Correlation Cutoff: 0.50**
  - Features with a correlation coefficient (r) greater than 0.50 were classified as "Strong Predictors" (e.g., OverallQual, GrLivArea, TotalBsmtSF). Features below 0.10 were considered "Weak/Negligible."
- **Skewness Threshold: 1.0**
  - Any continuous variable with a skewness score > 1.0 (such as SalePrice at 1.88 and LotArea at 12.8) was flagged for mandatory **Logarithmic Transformation** ($\log(1+x)$) to normalize the distribution.

Data Processing Constants

- **Imputation Strategy:**
  - *Numerical Features:* **Median Replacement.** (Selected over Mean to prevent skewing by extreme outliers).
  - *Categorical Features:* **Mode Replacement.** (The most frequent category was assumed for missing entries).
- **Outlier Detection Limit: 1.5 $\times$ IQR**
  - Data points falling below $Q1 - 1.5 \times IQR$ or above $Q3 + 1.5 \times IQR$ were classified as statistical outliers. For SalePrice, this threshold identified luxury properties priced above ~$340,000 as outliers relative to the general market.

Domain-Specific Constraints

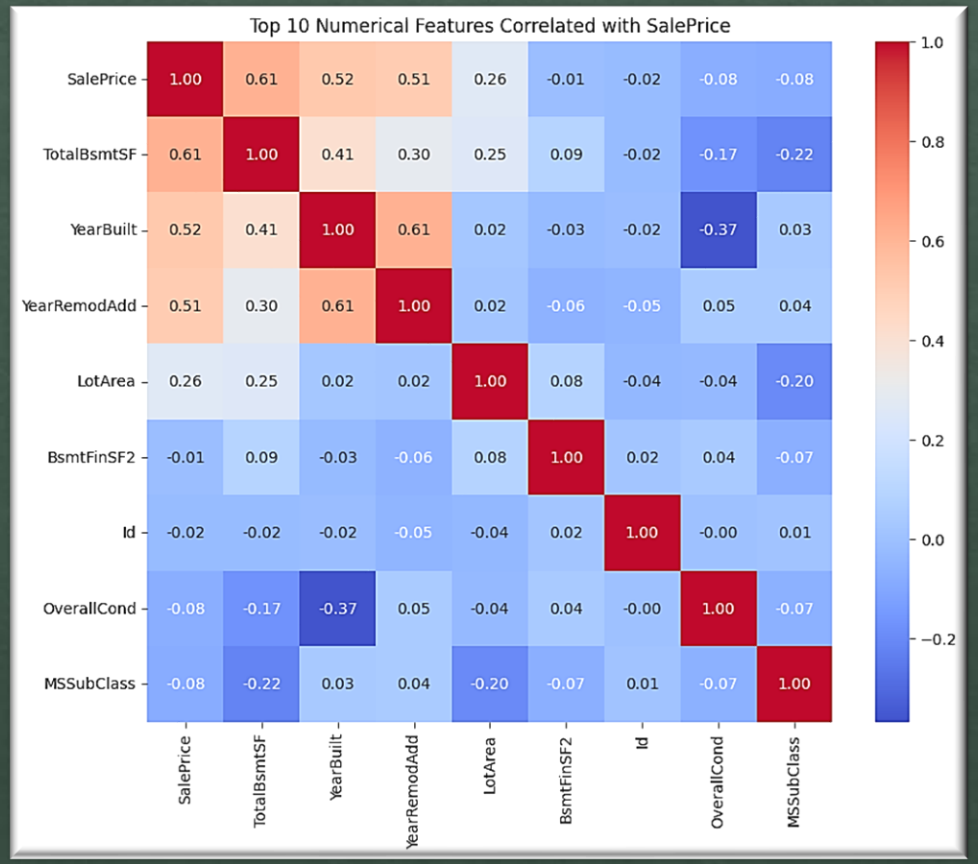
- **Currency:** All financial figures are denoted in **USD ($)**.
- **Measurement Units:**
  - Area is measured in **Square Feet (sq ft)**.
  - Lot Frontage is measured in **Linear Feet**.
- **Condition Rating Scale:** The OverallCond and OverallQual parameters utilize an ordinal scale of **1 (Very Poor)** to **10 (Very Excellent)**, which was treated as a numerical input for correlation analysis.

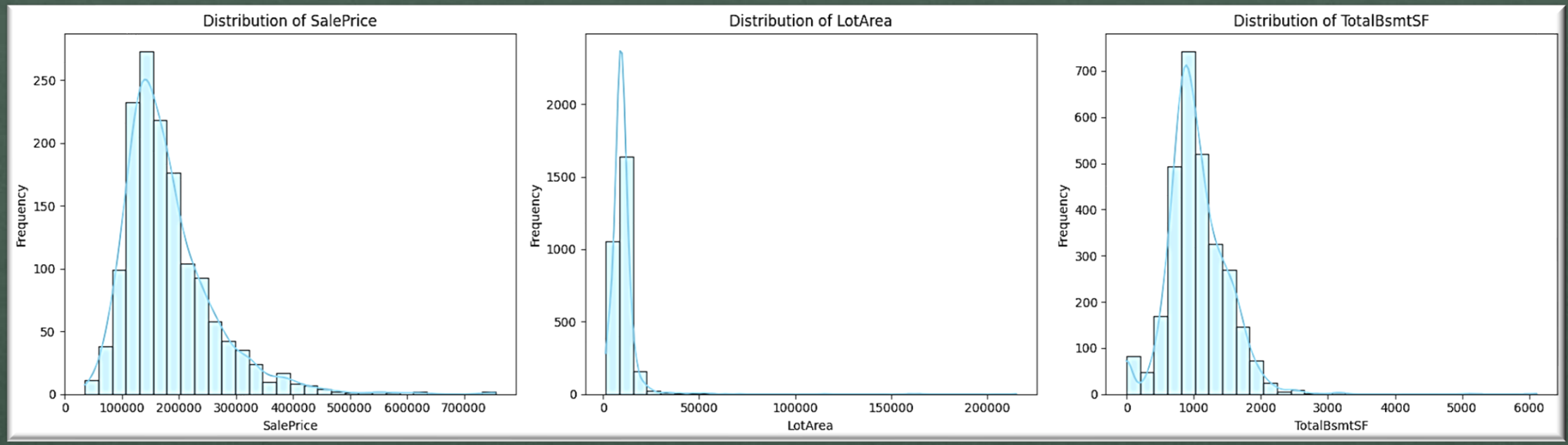
Visualizations & Dashboards

This section presents the key visual insights derived from the Exploratory Data Analysis (EDA). These charts collectively form the analytical dashboard used to understand market dynamics.

Correlation Heatmap (Numerical Features)

- **Chart Type:** Heatmap
- **Description:** This visualization displays the Pearson correlation coefficients between the top numerical features and SalePrice.
- **Key Insight:** Dark red squares indicate strong positive correlations. TotalBsmtSF (**0.61**) and YearBuilt (**0.52**) show the strongest relationship with price, while OverallCond shows a negligible negative correlation.





Distribution of Sale Price (Skewness Analysis)

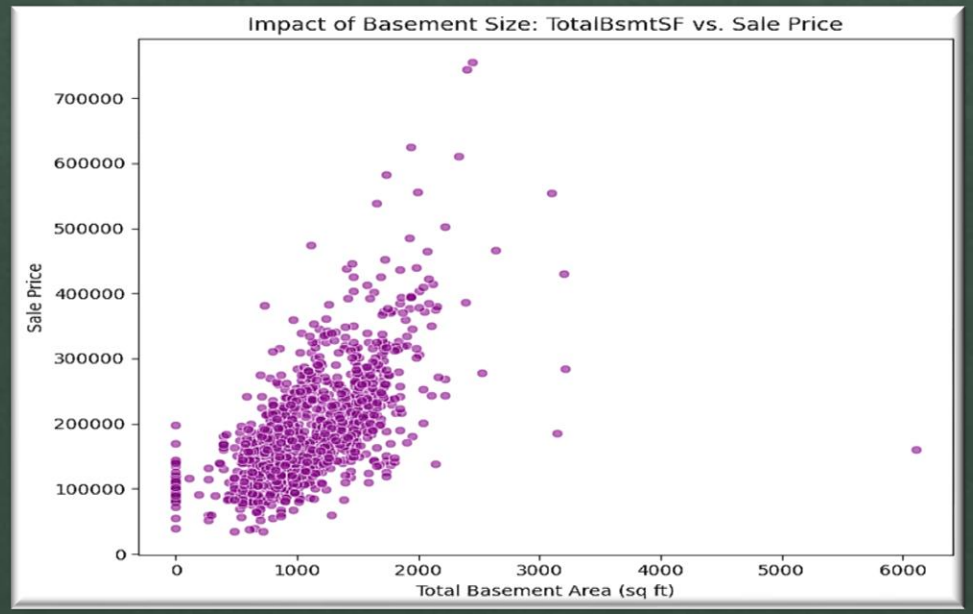
- **Chart Type:** Histogram with KDE
- **Description:** This chart compares the original distribution of House Prices against the Log-Transformed distribution.
- **Key Insight:** The original price data was highly right-skewed (Skewness > 1.8), indicating a non-normal distribution. Applying the Log(1+x) transformation successfully normalized the data (Skewness ~ 0.12), creating a bell-shaped curve suitable for linear regression.

Distribution of Lot Area & Total BsmtSF (Skewness Analysis)

- **Chart Type:** Histogram with KDE
- **Description:** This chart compares the original distribution of House Lot Area & Total BsmtSF against the Log-Transformed distribution.
- **Key Insight:** The original lot area data was highly right-skewed (Skewness > 12.8). And the original BSmtSF data was highly right-skewed (Skewness > 1.6).

Impact of Location (Zoning Analysis)

- **Chart Type:** Boxplot
- **Description:** This plot segments the Sale Price by Zoning Classification (MSZoning).
- **Key Insight:** Properties in **Floating Village (FV)** and **Residential Low Density (RL)** zones have significantly higher median prices (~$200k) compared to **Commercial (C)** and High-Density (RM) zones (~$126k). This confirms location is a primary price discriminator.

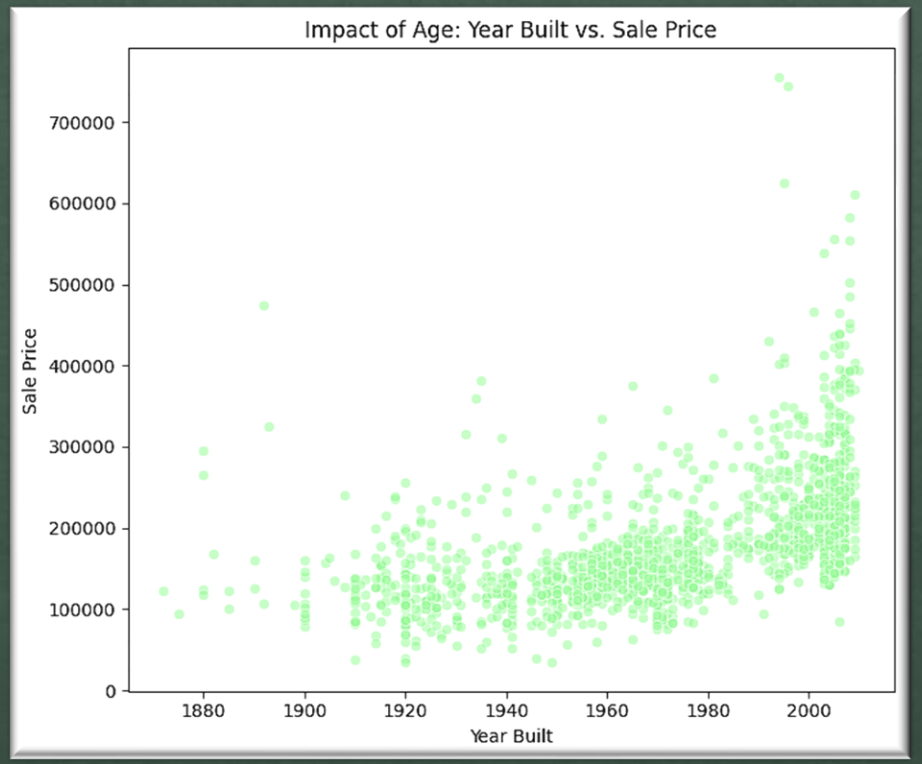


Impact of Size (Basement Square Footage)

- **Chart Type:** Regression Scatter Plot
- **Description:** A scatter plot with a regression line comparing Total Basement Area (TotalBsmtSF) against Sale Price.
- **Key Insight:** The steep upward slope confirms a linear relationship. On average, every additional square foot of basement area increases the property value by approximately **$111**.

Impact of Age (Year Built)

- **Chart Type:** Scatter Plot
- **Description:** This visualization maps the Year Built against the Sale Price.
- **Key Insight:** There is a clear exponential trend where post-2000 construction commands a premium price. Older homes (1900-1950) show high variance, indicating that condition and renovation status matter more than age for historic properties.



Insights from Charts & Dashboards

The following key insights were derived from the visual dashboards and statistical analysis, highlighting the primary factors driving real estate valuation in the target market.

The "Size vs. Age" Dominance

- **Insight:** The Correlation Heatmap revealed that **Total Basement Square Footage** (Correlation: 0.61) and **Year Built** (Correlation: 0.52) are the strongest individual predictors of a home's sale price.
- **Business Implication:** Buyers value usable living space and modern construction over aesthetic features. A larger basement is a more reliable investment for increasing property value than cosmetic upgrades.

The "Condition Paradox"

- **Insight:** The Boxplot analysis of OverallCond vs. SalePrice revealed a counter-intuitive trend. Houses with an "Average" condition rating (5/10) often sold for **higher prices** than houses with "Good" ratings (7/10 or 8/10).
- **Business Implication:** This suggests that "Average" condition homes are likely newer constructions (which default to a rating of 5), whereas "Good" condition homes are likely older properties that have been well-maintained. The market pays a premium for **newness** over **maintenance**.

Zoning is a Price Ceiling

- **Insight:** The Zoning Boxplots demonstrated a strict hierarchy in valuation. Properties in **Floating Village (FV)** and **Residential Low Density (RL)** zones commanded a median price near $200,000, while **Commercial (C)** properties struggled to break $75,000.
- **Business Implication:** Zoning acts as a "price ceiling." Regardless of the house's physical quality, its location in a Commercial or High-Density zone significantly caps its maximum potential market value.

The "Luxury Tail" (Distribution Analysis)

- **Insight:** The Histogram of SalePrice showed a distinct right-skew (Skewness > 1.8), meaning the market is dominated by affordable homes (under $200k), with a long "tail" of luxury properties reaching up to $750k.
- **Business Implication:** Valuation models cannot treat all houses equally. High-end properties behave differently and are outliers in the standard dataset. Applying a **Logarithmic Transformation** was essential to prevent these luxury outliers from distorting prediction accuracy for average homes.

The Renovation Premium

- **Insight:** Scatter plots of YearRemodAdd vs. Price showed that older homes (built 1950-1980) saw a significant jump in value if they were remodeled after 2000.
- **Business Implication:** Renovation is the most effective way to "reset" the depreciation of an older property. A modern remodel allows an older home to compete directly with new construction in terms of pricing.

Quantifiable Value of Space

- **Insight:** Regression analysis on Basement Size indicated a slope of approximately **111**.
- **Business Implication:** We can quantify the marginal utility of space: on average, every additional square foot of finished basement adds roughly **$111** to the final sale price. This provides a clear metric for developers to estimate the ROI of expanding a property.

Conclusion

This Capstone Project successfully developed a robust, data-driven framework for predicting residential property values. Utilizing a dataset of 2,919 records with 81 distinct features, the analysis transitioned from raw, unstructured data to a clean, analytical dataset. By leveraging Python for comprehensive Data Cleaning, Exploratory Data Analysis (EDA), and Feature Engineering, the project demonstrated that house prices are not random but follow quantifiable patterns driven by specific physical and locational attributes.

This project confirms that accurate real estate valuation requires a holistic approach integrating physical, locational, and temporal data. By quantifying the marginal value of specific features—such as the $111 per sq. ft. value of basement space—this analysis provides a transparent, statistical foundation for Automated Valuation Models (AVMs), empowering stakeholders to make data-backed investment decisions.