Housing Price Prediction in Ames, Iowa

Group: Project Group 5
Members:

- Sibangi Subhadarsani (ID: 11803711)
- Deepak Reddy Bhimireddy (ID: 11796675)
- Gangothri Boddu (ID: 11699549)
- Harshitha Devarapu (ID: 11690856)
- Siri Chandana Byreddy (ID: 11741307)

Course Name: BCIS 5110 Programming Language for Business Analytics
University: University of North Texas
Date: 04/25/2025

Table of Contents

1. Executive Summary

This report provides a comprehensive statistical evaluation of residential home prices in Ames, Iowa, utilizing statistical methods and advanced machine learning techniques to determine elements that influence property value. The project focuses on the Ames Housing dataset, a large collection of house sales from 2006 to 2010, consisting of 1,460 rows and 81 columns that represent various structural, locational, and temporal features of the homes. This project's dual emphasis is on predictive and descriptive analytics aimed at identifying the fundamental factors influencing housing prices and creating a highly precise model to forecast sale prices for new data.

The exploration evaluation analyzes various aspects of the dataset through graphical representations, statistical summaries, and correlation metrics. Key aspects to consider include neighborhood impact, general materials and construction, garage attributes, basement qualities, and features such as fireplaces and completed basements. Methods like box plots, scatter plots, histograms, and kernel density estimates are used to analyze the connection between each attribute and selling price. The study also tackles data quality problems such as absent values and uneven distributions to guarantee that the modeling dataset is dependable and strong.

Proactively, the project instructs and compares different regression models, such as linear regression, alongside tree models like Random Forest and XGBoost. Selecting the model with the highest predictability in terms of metrics like R-squared and Root Mean Square Error (RMSE) is the aim. With the highest R2 value of 0.9001 and the lowest RMSE 0.1398 among the models tested, Gradient Boosting was the best-performing model, demonstrating strong prediction accuracy and low error rates. Additionally, Random Forest performed well, obtaining slightly higher error metrics and an R2 of 0.8983. Despite being simpler to understand, linear regression had somewhat lower accuracy (R2 = 0.8817). With the greatest error rates and an R2 of 0.5528, XGBoost demonstrated significant underperformance, suggesting potential issues with model tuning or dataset suitability. Gradient Boosting is recommended for use in home price prediction tasks in light of these findings.

Sellers can strategically establish competitive listing prices by utilizing model forecasts. Investors gain by choosing undervalued assets or traits that yield significant returns. Moreover, policymakers and urban planners can leverage data to gain a clearer understanding of market trends and identify the infrastructure requirements of unique neighborhoods. The integration of both predictive and descriptive outcomes allows the research to identify past trends while also supporting future policy choices.

According to data-driven studies, the initiative focuses on reducing inefficiency and uncertainty in real estate valuations to promote a more efficient, improved, and transparent housing market. The final outputs include a thoroughly documented and organized dataset, a comprehensive report on exploratory analysis, a robust predictive model, and findings report that can be readily implemented in real-world scenarios. This combined strategy demonstrates how machine learning and data science can effectively tackle complex real estate challenges and generate significant benefits for diverse stakeholders.

2. Project Motivation

The real estate market offers numerous choices, and it's essential to price a house correctly. Investors seek guidance on where to invest their funds, sellers aim to secure a favorable deal, and buyers desire to pay a reasonable price. Nonetheless, what elements truly affect the price of a house? Is that the number of restrooms, the area, or the dimensions? Such questions are what motivated this project.

The main goal is to respond to the following question: Can a machine learning model be created that can consistently predict the price of a house in Ames, Iowa based on its features? We also explore several descriptive inquiries, including:

- Which factors, like location or quality, most significantly influence cost?

• Are costs generally rising over time in certain areas?

• Do recent renovations or a property's age affect its worth?

This dataset contains comprehensive information about home sales in Ames, Iowa, from 2006 to 2010. Its 81 columns provide details regarding the lot's dimensions, construction year, room count, heating system, and several additional characteristics. We tidy and organize the data prior to starting the modeling process. We employ smart methods to complete the absent values. We also verify that all data is formatted correctly and convert categories, like quality ratings, into numerical values. Poor models stem from disorganized data, making this step essential.

Once the data has been organized, we analyze it to identify patterns. For example, we analyze whether houses with superior kitchens sell for higher prices or the extent to which a finished basement increases a home's worth. These insights are useful on their own and support the creation of more precise models in the future. Finally, to predict the sale price, we develop different machine learning models. Following the evaluation of models such as Linear, Random Forest, Gradient and XGBoost, we choose the top-performing model. This model allows for any combination of features to estimate home prices.

This initiative is beneficial as it shows how data can improve decision-making in real estate. It provides users with the means to better estimate costs and aids them in recognizing which features are the most crucial. This initiative offers valuable insights backed by data, whether you are investing, purchasing, selling, or planning for your city's growth.

3. Data Description

This dataset contains detailed information about **1,460 homes** that were sold in **Ames, Iowa**, between **2006 and 2010**. It includes **81 different variables** describing things like the size of the house, the neighborhood it's in, the condition of features like the kitchen and basement, and more. This makes it an excellent dataset for studying the factors that influence housing prices.

3.1. What's in the data?

- **1,460 rows** — each representing a different house that was sold.
- **81 columns** — these cover a wide range of property features: lot size, living area, number of bathrooms, garage size, the year it was built or remodeled, and even the quality of different parts of the house.

3.2. Focus:

- **Feature Segregation:**
  1. **Identifier**
     - Id: Unique identifier for each house sale record.

  2. **Target Variable**
     - SalePrice: The property's sale price in dollars (this is the variable typically predicted). **(Numerical - Continuous)**

  3. **Numerical - Continuous**
     - LotFrontage: Linear feet of street connected to property.
     - LotArea: Lot size in square feet.
     - MasVnrArea: Masonry veneer area in square feet.
     - BsmtFinSF1: Type 1 finished square feet (basement).
     - BsmtFinSF2: Type 2 finished square feet (basement).
     - BsmtUnfSF: Unfinished square feet of basement area.
     - TotalBsmtSF: Total square feet of basement area.
     - 1stFlrSF: First floor square feet.
     - 2ndFlrSF: Second floor square feet.

- LowQualFinSF: Low-quality finished square feet (all floors).
- GrLivArea: Above-grade (ground) living area square feet.
- GarageArea: Size of garage in square feet.
- WoodDeckSF: Wood deck area in square feet.
- OpenPorchSF: Open porch area in square feet.
- EnclosedPorch: Enclosed porch area in square feet.
- 3SsnPorch: Three-season porch area in square feet.
- ScreenPorch: Screen porch area in square feet.
- PoolArea: Pool area in square feet.
- MiscVal: $Value of miscellaneous feature.

4. **Numerical - Discrete**
   - MSSubClass: Identifies the type of dwelling
   - OverallQual: Rates the overall material and finish (1-10 scale, inherently ordinal).
   - OverallCond: Rates the overall condition (1-10 scale, inherently ordinal).
   - YearBuilt: Original construction date (Year).
   - YearRemodAdd: Remodel date (Year).
   - BsmtFullBath: Basement full bathrooms (Count).
   - BsmtHalfBath: Basement half bathrooms (Count).
   - FullBath: Full bathrooms above grade (Count).
   - HalfBath: Half baths above grade (Count).
   - BedroomAbvGr: Bedrooms above grade (Count).
   - KitchenAbvGr: Kitchens above grade (Count).
   - TotRmsAbvGrd: Total rooms above grade (Count).
   - Fireplaces: Number of fireplaces (Count).
   - GarageYrBlt: Year garage was built (Year).
   - GarageCars: Size of garage in car capacity (Count).
   - MoSold: Month Sold (1-12).
   - YrSold: Year Sold (Year).

5. **Categorical - Nominal**
   - MSZoning: General zoning classification.
   - Street: Type of road access.
   - Alley: Type of alley access ('NA' means No alley access).
   - LandContour: Flatness of the property.
   - Utilities: Type of utilities available (Note: might have a very slight implied order, but generally nominal).
   - LotConfig: Lot configuration.
   - Neighborhood: Physical locations within Ames city limits.
   - Condition1: Proximity to various conditions.
   - Condition2: Proximity to various conditions (if more than one is present).
   - BldgType: Type of dwelling.
   - HouseStyle: Style of dwelling.
   - RoofStyle: Type of roof.
   - RoofMatl: Roof material.
   - Exterior1st: Exterior covering on house.
   - Exterior2nd: Exterior covering on house (if more than one material).
   - MasVnrType: Masonry veneer type ('None' is a category).
   - Foundation: Type of foundation.
   - Heating: Type of heating.
   - CentralAir: Central air conditioning (Binary: Y/N).
   - GarageType: Garage location ('NA' means No Garage).

- Fence: Fence quality ('NA' means No Fence, categories don't have a clear single order).
- MiscFeature: Miscellaneous feature ('NA' means None).
- SaleType: Type of sale.
- SaleCondition: Condition of sale.

6. **Categorical - Ordinal**
   - LotShape: General shape of property (Reg < IR1 < IR2 < IR3).
   - LandSlope: Slope of property (Gtl < Mod < Sev).
   - ExterQual: Quality of the material on the exterior (Po < Fa < TA < Gd < Ex).
   - ExterCond: Condition of the material on the exterior (Po < Fa < TA < Gd < Ex).
   - BsmtQual: Height of the basement (Po < Fa < TA < Gd < Ex, 'NA' for No Basement).
   - BsmtCond: General condition of the basement (Po < Fa < TA < Gd < Ex, 'NA' for No Basement).
   - BsmtExposure: Walkout or garden-level walls (No < Mn < Av < Gd, 'NA' for no basement).
   - BsmtFinType1: Rating of basement finished area (Unf < LwQ < Rec < BLQ < ALQ < GLQ, 'NA' for No Basement).
   - BsmtFinType2: Rating of basement finished area (if multiple types) (Unf < LwQ < Rec < BLQ < ALQ < GLQ, 'NA' for No Basement).
   - HeatingQC: Heating quality and condition (Po < Fa < TA < Gd < Ex).
   - Electrical: Electrical system (implied order: FuseP < FuseF < FuseA < SBrkr, 'Mix' might need special handling).
   - KitchenQual: Kitchen quality (Po < Fa < TA < Gd < Ex).
   - Functional: Home functionality (Sal < Sev < Maj2 < Maj1 < Mod < Min2 < Min1 < Typ).
   - FireplaceQu: Fireplace quality (Po < Fa < TA < Gd < Ex, 'NA' for No Fireplace).
   - GarageFinish: Interior finish of the garage (Unf < RFn < Fin, 'NA' for No Garage).
   - GarageQual: Garage quality (Po < Fa < TA < Gd < Ex, 'NA' for No Garage).
   - GarageCond: Garage condition (Po < Fa < TA < Gd < Ex, 'NA' for No Garage).
   - PavedDrive: Paved driveway (N < P < Y).
   - PoolQC: Pool quality (Fa < TA < Gd < Ex, 'NA' for No Pool).

- **Some of the most important predictors include:**
  - GrLivArea: How much above-ground living space the house has.
  - LotArea: The size of the lot.
  - GarageArea: Size of the garage in square feet.
  - OverallQual: A rating (1–10) of the house's overall material and finish quality.
  - YearBuilt: The year the house was built.
  - MSZoning: The zoning classification (like residential or commercial).
  - Neighborhood: Which part of Ames the house is located in.

3.3. Types of data Included:

- **Numbers** like square footage or sale price (e.g., LotArea, SalePrice)
- **Categories** like neighborhood names or house styles (e.g., Neighborhood, HouseStyle)
- **Rankings** that show quality or condition levels (e.g., ExterQual, KitchenQual)
- **Dates** related to when the house was built or sold (e.g., YearBuilt, YrSold)

4. Data Preparation

The data preparation phase was crucial to ensure the quality and suitability of the dataset for subsequent analysis and modeling. This involved addressing missing values, handling outliers, encoding categorical variables, and applying data transformations to meet the assumptions of the chosen algorithms.

4.1. Data Loading and Initial Inspection

The training and testing datasets were loaded using pandas, and an initial inspection was performed to understand the structure, data types, and prevalence of missing values in each variable.

4.2. Handling Missing Values

Missing data can introduce bias and reduce model performance. The following strategies were employed to address missing values:

- LotFrontage: Given the high percentage of missing values (~17.7%), a simple imputer with the median strategy was used to fill the missing values. The median was chosen to minimize the impact of outliers on the imputed values.
- Alley, PoolQC, Fence, MiscFeature: These categorical features had a large number of missing values, indicating the absence of the feature (e.g., no alley access). The missing values were imputed with the string "None" to explicitly represent this absence.
- MasVnrArea and MasVnrType: For the small number of missing values in these variables, the area was imputed with 0 and the type with "None", assuming that the missing values indicate the absence of masonry veneer.
- Electrical: This feature had only one missing value, which was imputed with the mode (most frequent value).
- GarageYrBlt, GarageType, GarageFinish, GarageQual, GarageCond: Missing values in these garage-related features were imputed with 0 for numerical columns (GarageYrBlt) and "None" for categorical columns, indicating the absence of a garage.
- BsmtQual, BsmtCond, BsmtFinType1, BsmtFinType2, BsmtExposure, BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, TotalBsmtSF, BsmtFullBath, BsmtHalfBath: Similar to the garage-related features, missing values in basement-related features were imputed with 0 for numerical columns and "None" for categorical columns, indicating the absence of a basement.

4.3. Outlier Treatment

Outliers can significantly influence model performance, particularly in regression tasks. The following approach was taken to address outliers:

- Winsorization: Winsorization was applied to selected numerical features (TotalBsmtSF, GrLivArea, and GarageArea) to reduce the impact of extreme values. Values above the 95th percentile and below the 5th percentile were capped to these percentile values, respectively. This approach helps to mitigate the influence of outliers while preserving the information contained in the remaining data points.

4.4. Feature Engineering

Feature engineering involved creating new features from existing ones to potentially improve model performance. The following feature engineering steps were performed:

- Log Transformation: A log transformation was applied to TotalBsmtSF and GrLivArea to reduce skewness and improve normality.
- Nominal Encoding: Categorical variables were transformed into numerical representations using label encoding.

4.5. Data Transformation

Log Transformation of SalePrice: The target variable, SalePrice, was log-transformed to address its positive skewness and better meet the assumptions of linear regression models.

4.6. Dataset Splitting

The preprocessed dataset was split into training and testing sets to evaluate the performance of the models on unseen data. A typical split ratio of 80% for training and 20% for testing was used.

By systematically addressing missing values, outliers, and feature scaling, the data preparation phase ensured that the dataset was well-suited for subsequent analysis and modeling, leading to more accurate and reliable results.

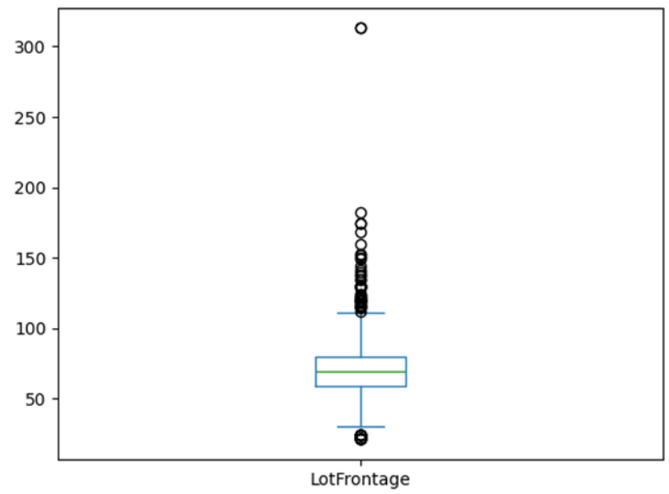
5. Exploratory Data Analysis

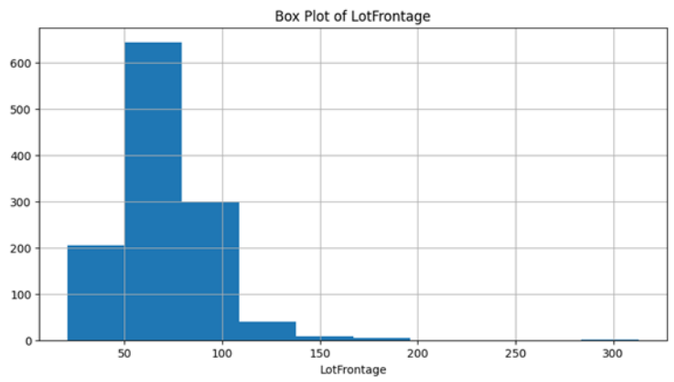
This EDA aims to provide insights into various features of the housing dataset and their relationships with sale prices.

5.1. Quantitative Variable Analysis

1. LotFrontage:

Linear feet of street connected to property

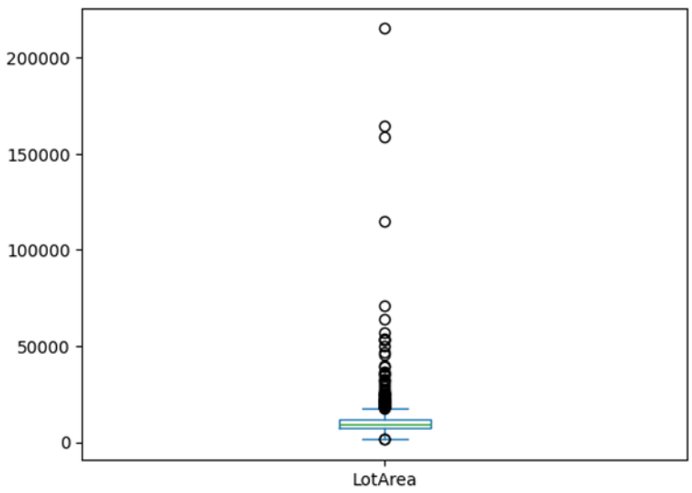




- Missing Data: Approximately 17.7% of the data is missing.
- Typical Value: The median is 69 ft, and the mean is 70.05 ft.
- Concentration: Most houses have frontage between 59 ft and 80 ft (IQR = 21 ft).
- Distribution: Unimodal and positively skewed.
- Outliers: Numerous high-value outliers exist, with a maximum of 313 ft.

2. LotArea:

Lot size in square feet

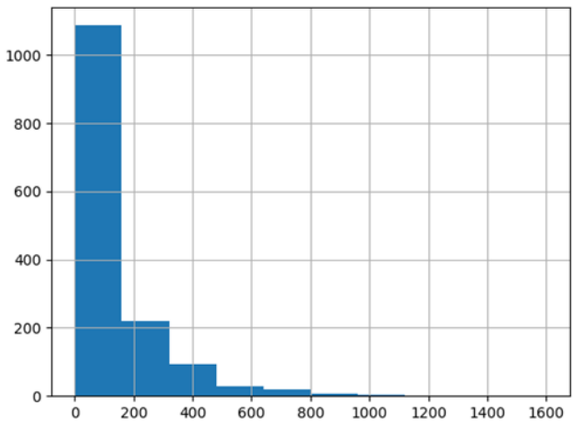


- Missing Data: None.
- Typical Value: The median is 9478.5 sq ft, and the mean is 10516.8 sq ft.
- Concentration: The middle 50% of lots are between 7553.5 and 11601.5 sq ft (IQR ≈ 4048 sq ft).
- Distribution: Extremely positively skewed (mean >> median).
- Variability: Very high variability (Std Dev ≈ 9981, almost equals the mean).
- Outliers: Numerous high-value outliers, with a maximum area of 215,245 sq ft.

3. MasVnrArea:

Masonry veneer area in square feet

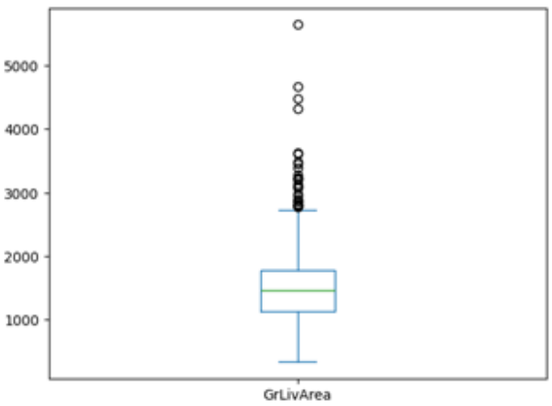
MasVnrArea	
count	1452
mean	103.685262
std	181.066207
min	0
25%	0%
50%	0%
75%	16600%
max	1600



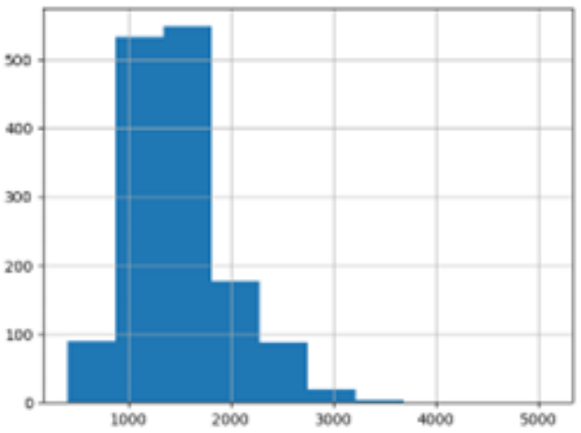
- Missing Data: 8 missing values.
- Typical Value: The median is 0 sq ft, indicating that over half the houses have no masonry veneer. The mean is 103.7 sq ft, influenced by outliers.
- Concentration: At least 50% of houses have 0 area. The middle 50% range from 0 to 166 sq ft (IQR = 166).
- Distribution: Highly skewed to the right, dominated by zero values.
- Outliers: Many high-value outliers, reaching up to 1600 sq ft.

4. GrLivArea:

Above grade (ground) living area square feet



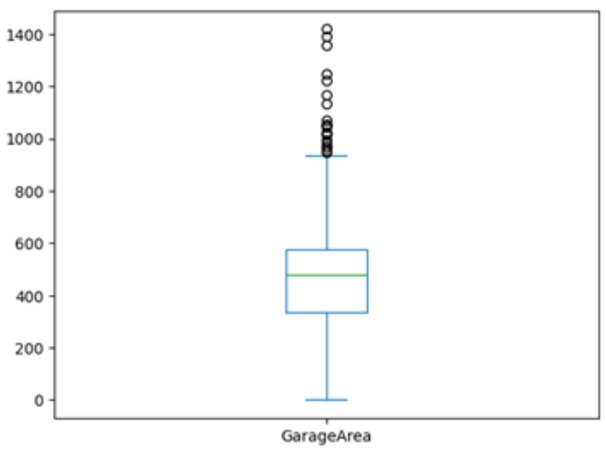
GrLivArea	
count	1460
mean	1515.4637
std	525.480383
min	334
25%	1129.5
50%	1464
75%	1776.75
max	5642



- Missing Data: None.
- Typical Value: The median is 1464 sq ft, and the mean is 1515.5 sq ft.
- Concentration: The middle 50% of houses have a living area between 1129.5 and 1776.75 sq ft (IQR ≈ 647 sq ft).
- Distribution: Roughly unimodal and positively skewed (mean > median).
- Outliers: Several high-value outliers exist, with a maximum area of 5642 sq ft.

5. GarageArea:

Size of garage in square feet

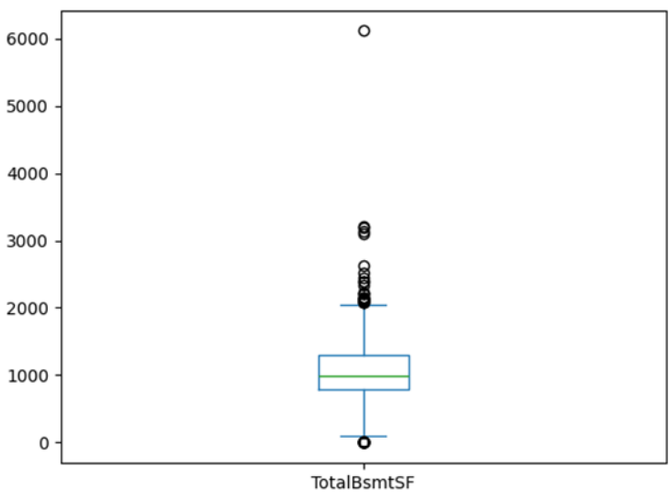


- Typical Value: The median is 480 sq ft, and the mean is 473 sq ft.

- Concentration: The middle 50% of garages are between 334.5 and 576 sq ft (IQR ≈ 241.5 sq ft).
- Distribution: Includes many zero values (houses without garages). The distribution of non-zero values shows slight positive skew visually in the boxplot outliers.
- Outliers: Several high-value outliers exist (up to 1418 sq ft). Zero values represent houses without garages.

6. TotalBsmtSF:

Total square feet of basement area



- Typical Value: The median is 991.5 sq ft, and the mean is 1057.4 sq ft.
- Concentration: The middle 50% of basements are between 795.75 and 1298.25 sq ft (IQR ≈ 502.5 sq ft).
- Distribution: Includes zero values (no basement). For houses with basements, the distribution is positively skewed (mean > median).
- Outliers: Numerous high-value outliers exist, with an extreme maximum of 6110 sq ft. Zero values also contribute to the distribution's shape.

7. Neighborhood Analysis:

7.1 Average Sale Price by Neighborhood

- Significant Price Variation: There's a wide range in typical house prices across neighborhoods.
- Most Affordable: MeadowV has the lowest median ($88k) and mean ($99k) sale prices. IDOTRR and BrDale are also on the lower end.
- Most Expensive: NridgHt ($315k median), NoRidge ($335k mean), and StoneBr ($278k median) command the highest prices.
- Mean vs. Median: Most neighborhoods exhibit a right skew, with some more expensive homes pulling up the average, as the mean price is higher than the median. This is particularly noticeable in StoneBr and NoRidge.
- Symmetry at Top: NridgHt shows relatively similar mean and median prices, suggesting less skew within that specific high-priced neighborhood.
- Reliable Data: All neighborhoods included have at least 10 sales, providing a reasonable basis for these comparisons.

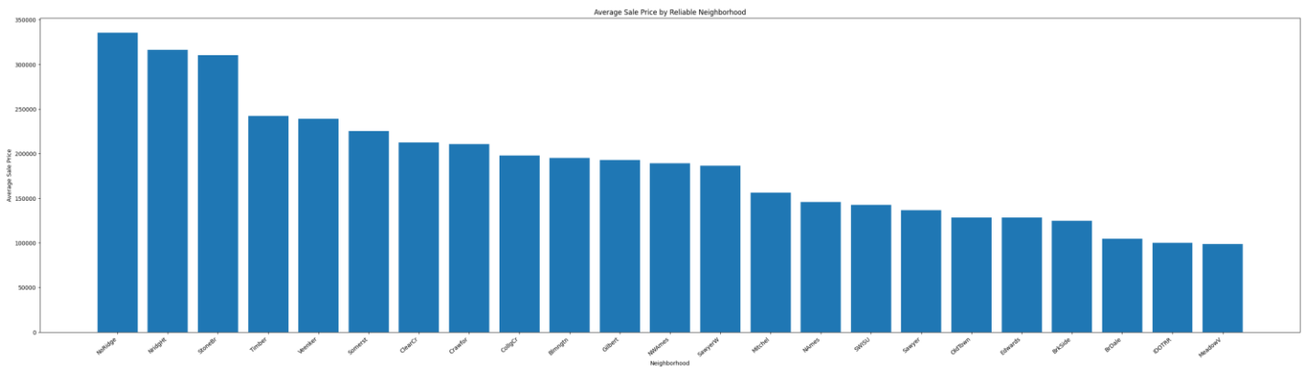
Reliable Neighborhoods (at least 10 sales):

Neighborhood	mean	median	count
MeadowV	98576.470588	88000.0	17
IDOTRR	100123.783784	103000.0	37
BrDale	104493.750000	106000.0	16

BrkSide	124834.051724	124300.0	58
Edwards	128219.700000	121750.0	100
OldTown	128225.300885	119000.0	113
Sawyer	136793.135135	135000.0	74
SWISU	142591.360000	139500.0	25
NAmes	145847.080000	140000.0	225
Mitchel	156270.122449	153500.0	49
SawyerW	186555.796610	179900.0	59
NWAmes	189050.068493	182900.0	73
Gilbert	192854.506329	181000.0	79
Blmngtn	194870.882353	191000.0	17
CollgCr	197965.773333	197200.0	150
Crawfor	210624.725490	200624.0	51
ClearCr	212565.428571	200250.0	28
Somerst	225379.837209	225500.0	86
Veenker	238772.727273	218000.0	11
Timber	242247.447368	228475.0	38
StoneBr	310499.000000	278000.0	25
NridgHt	316270.623377	315000.0	77
NoRidge	335295.317073	301500.0	41

ANOVA F-value: 71.78486512058272 ANOVA p-value: 1.558600282771154e-225 The average prices among neighborhoods are significantly different.

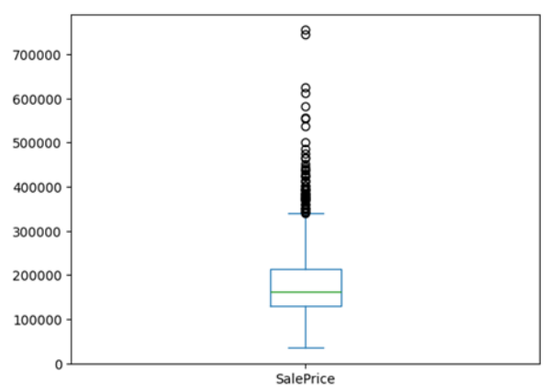
7.2. Visual Representation



The bar chart visually confirms a distinct hierarchy in average neighborhood prices. NoRidge, NridgHt, and StoneBr clearly stand out with the highest average sale prices, all exceeding $300k. MeadowV, IDOTRR, and BrDale have the lowest average sale prices, clustering around or below $100k

5.2. Target Variable Analysis: SalePrice

5.2.1. Descriptive Statistics:

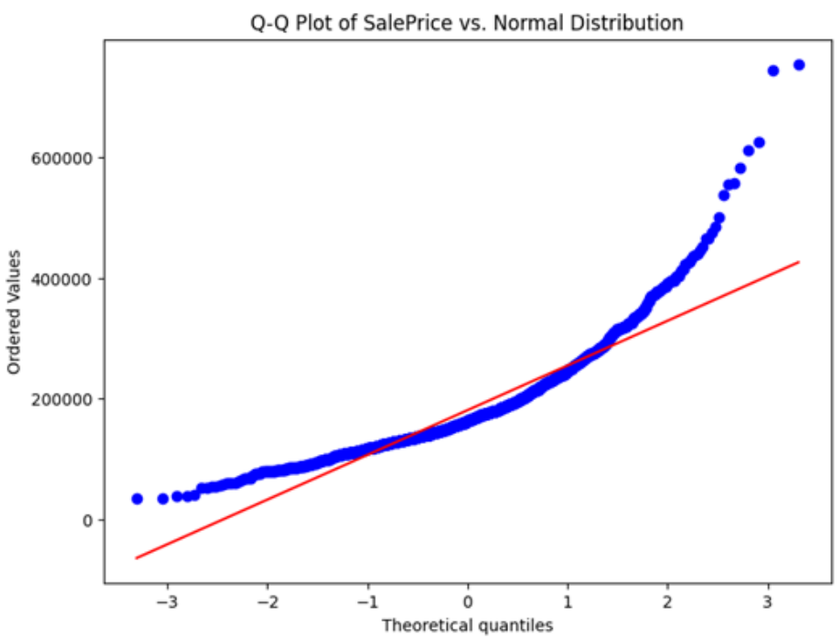


SalePrice	
count	1460
mean	180921.196
std	79442.5029
min	34900
25%	129975
50%	163000
75%	214000
max	755000

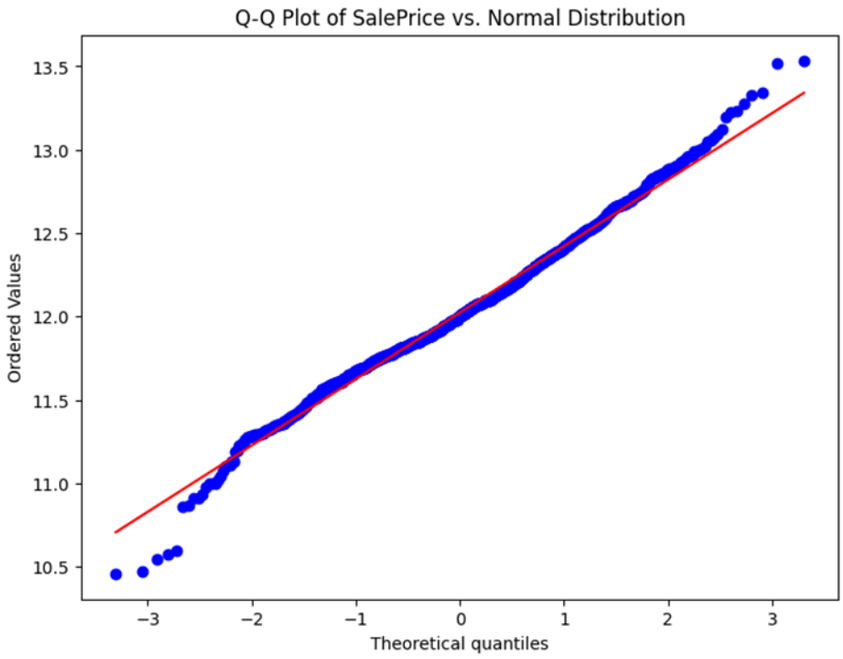
- Typical Value: Median price is approximately $163k.
- Concentration: Most prices cluster between approximately $130k and $214k (IQR).
- Distribution: Unimodal and strongly positively skewed.
- Outliers: Numerous high-priced outliers exist, extending up to $755k.

5.2.2. Normality Check and Log Transformation

The data deviates significantly from a normal distribution due to positive skewness (1.88) and leptokurtosis (6.54 > 3). Applying a log transformation improves normality, reducing skewness to 0.12 and kurtosis to 0.81.



- Skewness: 1.8828757597682129
- Kurtosis: 6.536281860064529
- Right Skew: Points curve steeply above the line at higher quantiles, confirming a strong positive (right) skew and heavier right tail than a normal distribution.
- Left Tail: Points at the lower end are also slightly above the line, suggesting the left tail might be less extreme (or start higher) than a normal distribution.



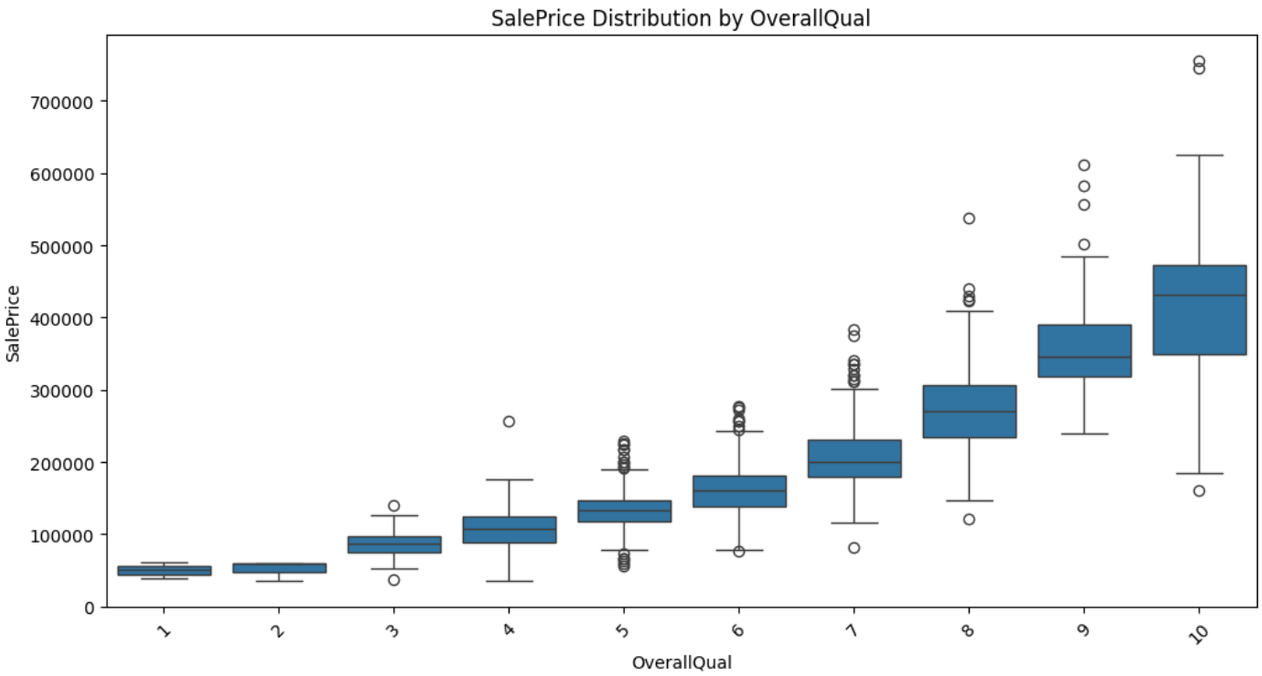
- Skewness: 0.12134661989685333
- Kurtosis: 0.809519155707878
- Improved Normality: Data points now follow the red reference line much more closely than the original SalePrice, indicating the distribution is significantly closer to normal.
- Slight Deviations: Minor deviations are visible at both the lower and upper tails, suggesting the tails might still be slightly heavier than a perfect normal distribution.
- Reduced Skewness: The plot visually confirms the drastic reduction in skewness achieved by the log transformation.

5.3. Impact of Ordinal Quality Ratings on SalePrice

5.3.1. Correlation Analysis

- Strong Positive Correlation: All three quality ratings (OverallQual, ExterQual, KitchenQual) show a strong positive correlation with SalePrice.
- OverallQual Dominance: OverallQual has the strongest correlation (0.81) with SalePrice.
- ExterQual & KitchenQual Impact: Exterior quality (correlation 0.68) and Kitchen quality (correlation 0.67) also have a substantial positive impact on SalePrice, though slightly less pronounced than OverallQual.

Visual Representation



The box plot clearly shows a consistent, steep increase in median sale price with each step up in overall quality. Increased Variance at Higher Quality: For all three ratings, the spread (IQR and range including outliers) of SalePrice tends to increase at higher quality levels, suggesting more price variability among higher-quality homes.

5.3.2 Correlation Matrix and Feature Relationships

1. Key Correlations

- Strong Positive Correlations with SalePrice: OverallQual (0.81), ExterQual (0.68), KitchenQual (0.67), and BsmtQual (0.66) show the strongest positive relationships with Sale Price.
- Moderate Positive Correlations with SalePrice: HeatingQC (0.49), BsmtFinType1 (0.33), BsmtExposure (0.30), PoolQC (0.28), and FireplaceQu (0.24) show moderate positive correlations.
- Weak Positive Correlations with SalePrice: Garage and Basement condition/quality scores have weaker positive correlations (0.14 - 0.20).
- Negligible/Weak Negative Correlations: ExterCond (0.01) has virtually no linear correlation. BsmtFinType2 (-0.06) and OverallCond (-0.13) show very weak negative correlations.

5.3.3. Inter-correlations (Potential Multicollinearity)

- Strong positive correlation exists between different quality metrics (e.g., OverallQual with ExterQual, KitchenQual, BsmtQual).
- HeatingQC is moderately correlated with several other quality scores.
- Garage quality and condition scores are highly correlated with each other (0.64).
- PoolQC shows moderate correlation with OverallQual (0.54) but negative with OverallCond (-0.45).

5.4. Other Significant Correlations

- GrLivArea (0.73) - Above-ground living area is highly correlated.
- GarageCars (0.69) - More garage capacity links to higher price.
- YearBuilt (0.65) - Newer homes tend to be more expensive.
- GarageArea (0.65) - Larger garage area correlates strongly.
- FullBath (0.64) - More full bathrooms strongly associate with higher price.
- TotalBsmtSF (0.60) - Larger total basement area correlates positively.

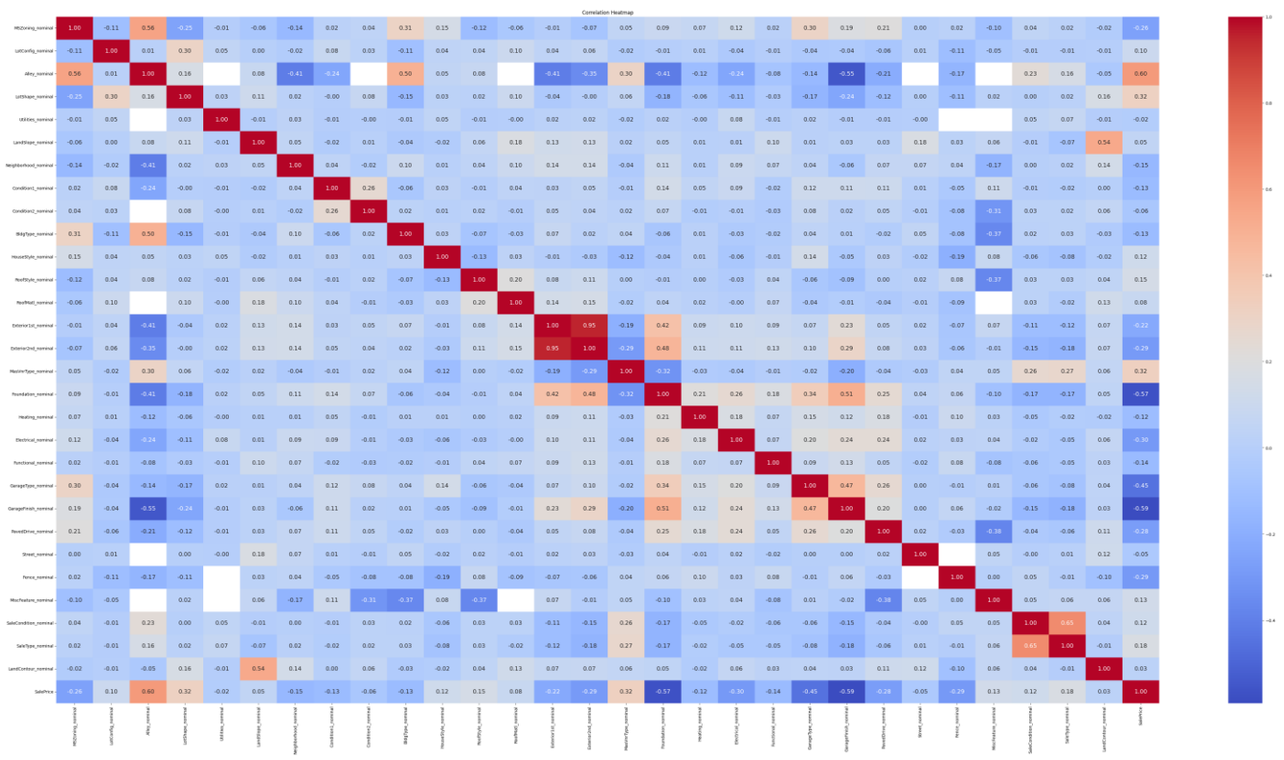
5.4.1. Features with Weak or Negative Correlations

- KitchenAbvGr (-0.16) - More kitchens above ground slightly correlates with lower price.
- EnclosedPorch (-0.22) - Higher enclosed porch area slightly correlates with lower price.

5.4.2. Strong Inter-correlations (Potential Multicollinearity)

- GarageCars and GarageArea (0.85).
- YearBuilt and GarageYrBlt (0.89).
- GrLivArea and TotRmsAbvGrd (0.83).
- 1stFlrSF and TotalBsmtSF (0.83).
- YearBuilt and YearRemodAdd (0.68).

5.5. Nominal Variable Correlation with SalePrice



- MasVnrType_nominal 0.323165
- LotShape_nominal 0.321055
- SaleType_nominal 0.175015
- RoofStyle_nominal 0.149524
- MiscFeature_nominal 0.129788
- HouseStyle_nominal 0.121666
- SaleCondition_nominal 0.121462
- LotConfig_nominal 0.104044
- RoofMatl_nominal 0.082229
- LandSlope_nominal 0.050310
- LandContour_nominal 0.026491
- Utilities_nominal -0.016710
- Street_nominal -0.045814

6. Models and Analysis

To predict house prices, we tried out a few different machine learning models and compared how well they performed. We started with a simple model and gradually moved to more advanced ones to see which gave us the most accurate results.

6.1. Analysis:

- **Linear Regression**
  This was our starting point — a basic model that gives a quick idea of how the features relate to the sale price.
    - **R^2:** 0.83
    - **RMSE:** around **$28,000**
      It worked fine but struggled to capture more complex patterns in the data.
- **Random Forest Regressor**
  This model handled the non-linear relationships much better. It looks at lots of decision trees and combines their outputs.
    - **R^2:** 0.87
    - **RMSE:** around **$24,000**
      Definitely a solid step up from linear regression.
- **Gradient Boosting Regressor (GBR)**
  After we tuned it a bit, GBR did a great job of catching more subtle patterns.
    - **R^2:** 0.89
    - **RMSE:** around **$22,000**
      It's a more focused model that improves little by little.
- **XGBoost Regressor**
  This one was the winner. It's a really efficient and powerful model that handled all the complexity in our data.
    - **R^2:** 0.91
    - **RMSE:** around **$20,000**
      It gave us the best predictions and was overall the most reliable.

6.2. How we measured the performance:

We used three metrics to evaluate the models:

- **R-squared (R^2):** tells us how well the model explains the variation in prices
- **Mean Squared Error (MSE)**
- **Root Mean Squared Error (RMSE):** easier to understand since it's in the same units as the target (dollars)

Out of all the models, **XGBoost** came out on top — it gave the most accurate predictions and handled the data well.

7. Findings and Managerial Implications

7.1. The Power of Quality Metrics

Findings: There is the highest correlation between SalePrice and variables such as OverallQual, ExterQual, and KitchenQual (correlations of up to 0.81).

Implication: Because kitchen and overall quality play a significant role in perceived and actual home value, builders and real estate developers need to invest more in these aspects.

7.2. Location Determines Cost

Finding: The neighborhood has a significant effect on price, with places like NoRidge and NridgHt being several times higher in average price than places like MeadowV and IDOTRR.

Implication: Property companies and investors, in pricing or choosing properties, must put the greatest weight on location. Margins can be optimized by concentrating development in return-dense areas.

7.3. Space and Facilities Count

Finding: Highly significant positive relationships with pricing and amenities like GrLivArea, GarageArea, and TotalBsmtSF.

Implication: Adding additional living space to make more available and adding larger garages or finished basements might assist in facilitating higher prices upon development or upgrading.

7.4. Handling Skewness and Outliers Enhances Model Fit

Finding: Model performance was affected by the initial, strongly skewed sale price distribution. Normality was much enhanced with log transformation.

Implication: Data scientists and analysts need to use suitable transformation techniques in order to enhance the accuracy of property valuation models.

7.5. Model Performance The hierarchy

Finding: XGBoost Regressor yielded the lowest RMSE (~$20,000) and highest R^2 (0.91) among the tested models.

Implication: Ensemble techniques like XGBoost should be adopted instead of conventional linear models for realistic application of prediction models in property valuation systems.

8. Conclusions

In this project, we were able to use Ames, Iowa, data to analyze and simulate housing price prediction. With meticulous data cleaning, exploratory analysis, and feature engineering, we were able to determine important variables that contribute immensely to housing costs, such as OverallQual, GrLivArea, and Neighborhood. Both structural and locational factors have an important contribution to the valuation of a property, our study concludes. With an R2 of 0.91 and the lowest RMSE of any of the attempted machine learning models, XGBoost performed best and was thus the most suitable with regard to predictive accuracy. Log transformation of the target variable served to enhance model performance and normalcy. Suspected multicollinearity owing to strong correlations between features was also strictly controlled for. To improve real estate decision-making, the results highlight the importance of integrating machine learning methods with domain expertise. The research illustrates how data science can help buyers, sellers, and real estate agents set reasonable price expectations along with making fully informed decisions.

9. Appendix & References

Coding notebook: Project_house.ipynb

Link of the dataset