Predicting House Prices Made Simple

Using Data to Help Home Buyers and Sellers

Samuel Ndirangu Maina

March, 09 2025

Business Understanding

- This dataset contains detailed features of residential homes in Ames, Iowa, including physical characteristics, location details, and sale prices.
- The goal of this project is to analyze various factors influencing house prices in these locations and develop predictive models to estimate house prices based on these factors.

Objectives

Below are the objectives;

- Build an accurate predictive model for house prices.
- Identify key features influencing house prices.
- Provide actionable insights for real estate stakeholders.
- Compare different machine learning models for optimal performance.

Data Understanding.

- ❖ The dataset has a total of 81 variables with 1460 observations or homes, this can be seen from the 1460 rows and 81 columns.
- ❖ From the 81 variables SalePrice Distribution is the dependent variable with the other 80 variables been independent variables.
- ❖ From the housing dataset below are some of the statistical observations that can be made;

- ❖The average sale price is approximately $180,921.
- ❖The minimum sale price is $34,900.
- ❖The maximum is $755,000

Data Cleaning

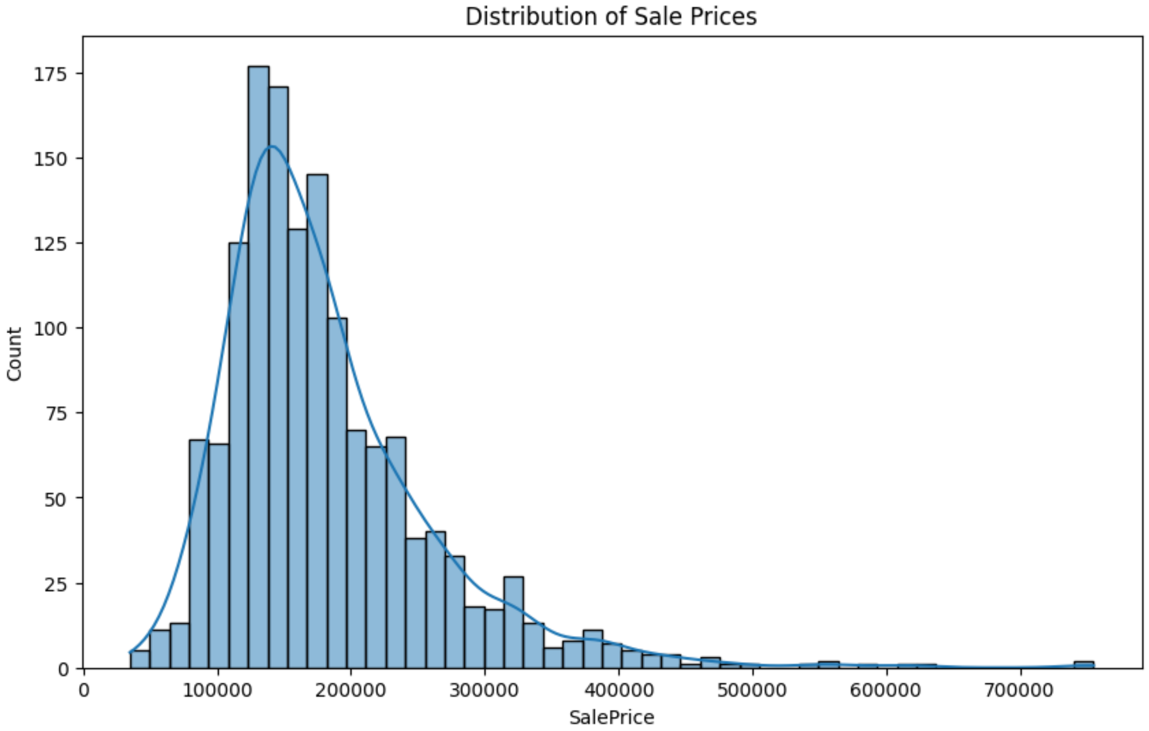
Below are the steps made towards cleaning the data;

- Fixing missing information such as; homes that didn’t list garage details.
- Removing duplicate entries to keep it fair.
- Through feature engineering we added new information like “Total Size” and “House Age” to make predictions better.

Explanatory Data Analysis.

A) Univariate Analysis;

The histogram depicts distribution of our dependent variable, SalePrice and from the observation made, most homes sell between $100,000 and $250,000.

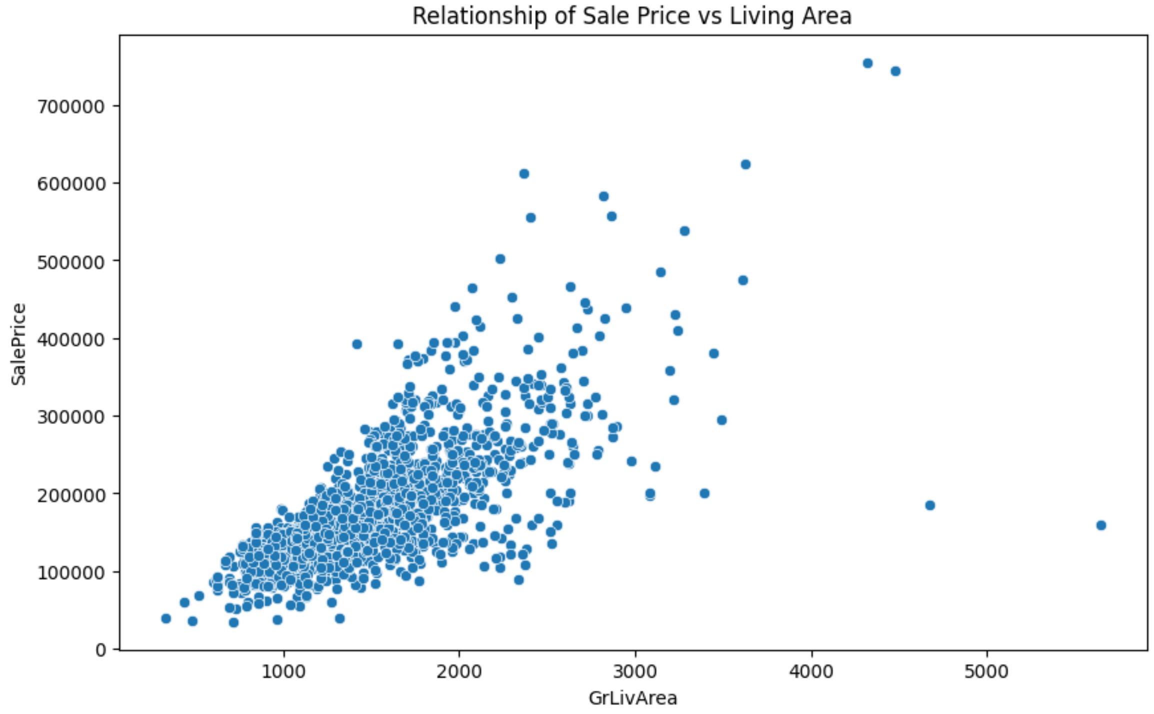


Explanatory Data Analysis. (Continuation…)

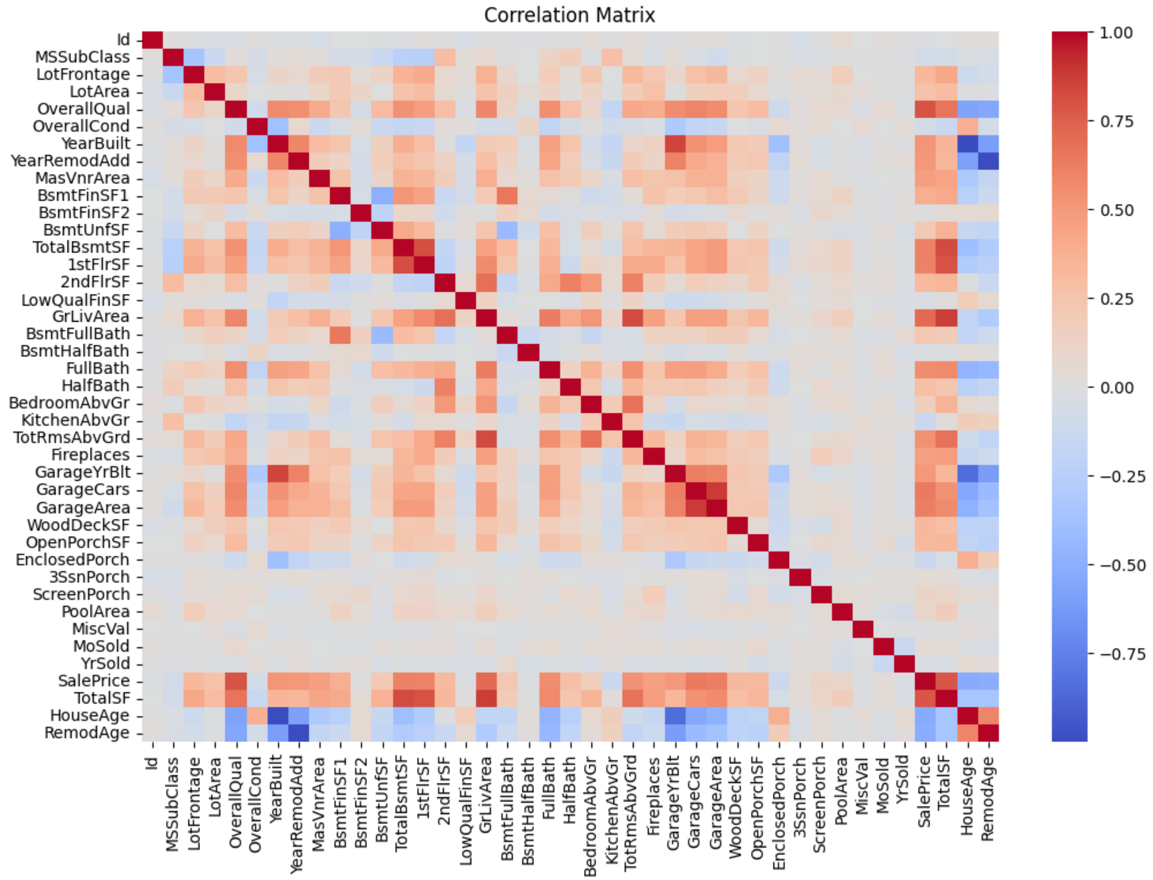
B) Bivariate Analysis

The Scatterplot depicts the relationship between living area and the price.

From the observation, homes with a bigger area costed more as would be expected.



Explanatory Data Analysis. (Continuation…)



C) Multivariate Analysis

The correlation matrix heatmap above provides a visual overview of the relationships between numerical features in the dataset.

Some of the Key Observations are as below;

Strong Positive Correlations: This can be identified by the strong positive correlations;

- Higher quality houses sell for more).
- More garage space correlates with more cars).
- Larger living area correlates with higher price).
- A larger total basement square footage correlates with a larger 1st floor square footage).

Data Preparation and Modeling

Through label encoding, categorical data was converted to numeric data.

Through Scaling we used 20% of our data to test it and 80% to train our model.

Upon trying different tools, the Random Forest emerged as the best option with the most accurate predictions.

Recommendation

- Focus on property size and quality for pricing decisions.
- Use the model for initial price estimates.
- Future improvements could include using advanced ensemble methods and feature selection techniques.

Thank You