IS 704 Midterm Project –Regression Analysis Report

# A. Introduction

The objective of this project is to develop predictive models for estimating residential home sale prices in Reno, Nevada, using 2020 data. Accurate price predictions support informed decision-making for buyers, sellers, real estate agents, and policymakers. This project applies regression-based machine learning techniques to uncover relationships between housing attributes and sale prices and to build a robust predictive model.

# B. Methodology

## Dataset Overview

* Source: Sample Space CP2.xlsx
* Observations: 6,733 property records
* Initial Features: Over 40 structural, temporal, and locational attributes

## Preprocessing Steps

* Missing Data Handling: Removed columns with more than 90% missing values. Rows missing key fields like Price, Gar Area, Bsmt Area, or Acres were dropped.
* Feature Engineering:  
  Extracted Sale Month and Sale Day of Week from Sales Date  
   Applied one-hot encoding to categorical fields
* Dimensionality Reduction: Reduced the feature space to top 20 predictors using SelectKBest with f regression scoring
* Data Scaling: Standardized numeric features using StandardScaler

## Models Evaluated

1. Linear Regression
2. Ridge Regression (L2 Regularization)
3. Lasso Regression (L1 Regularization)
4. Random Forest Regressor

## Validation Strategy

* Train-Test Split: 80% training, 20% testing
* Resampling: 5-Fold Cross-Validation for Ridge and Lasso
* Evaluation Metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), R-squared (R²)

# C. Results & Discussion

## Model Performance Comparison

|  |  |  |  |
| --- | --- | --- | --- |
| Model | RMSE | MAE | R² Cross-R2 |
| Linear Regression | $58,321.12 | $41,980.21 | 0.8123 0.7998 |
| Ridge Regression | $58,304.75 | $41,976.48 | 0.8125 0.8002 |
| Lasso Regression | $58,547.62 | $42,171.93 | 0.8103 0.7985 |
| Random Forest | $51,901.07 | $63,384 | 0.8410 0.8281 |

## These performance metrics were calculated on the test dataset, ensuring that the results reflect the model's ability to generalize to unseen data.

## EDA Visualizations & Insights

**Key Insights**

* **Random Forest Regressor** delivered the best performance due to its ability to model non-linear relationships and handle high-dimensional data efficiently. Tree-based models also perform well in such settings due to their inherent ability to select and prioritize the most important features during the modeling process, effectively reducing noise and overfitting.
* **Ridge and Lasso** showed slight improvements over basic Linear Regression, indicating multicollinearity and overfitting mitigation.
* **Linear models** are easier to interpret but fail to capture complex feature interactions.

**EDA Observations**

* **Positive Correlation:** Features like Gar Area and Bsmt Area showed moderate positive correlation with price.
* **Distribution Analysis:** Price distribution was right-skewed with a few high-value outliers. Log transformation could help normalize future models.
* **Neighborhood Trends:** Certain neighborhoods consistently had higher average prices, indicating strong geographic influence.
* **Correlation Heatmap:** A heatmap was also used to examine the relationships among numeric variables, helping identify which features are most strongly associated with the target variable Price. This supported decisions in feature selection and model interpretation.

**Challenges Faced**

* **High Dimensionality:** One-hot encoding increased feature count significantly. Mitigated using feature selection.
* **Missing Values:** Affected many categorical columns, addressed by filling with "Missing" placeholder.

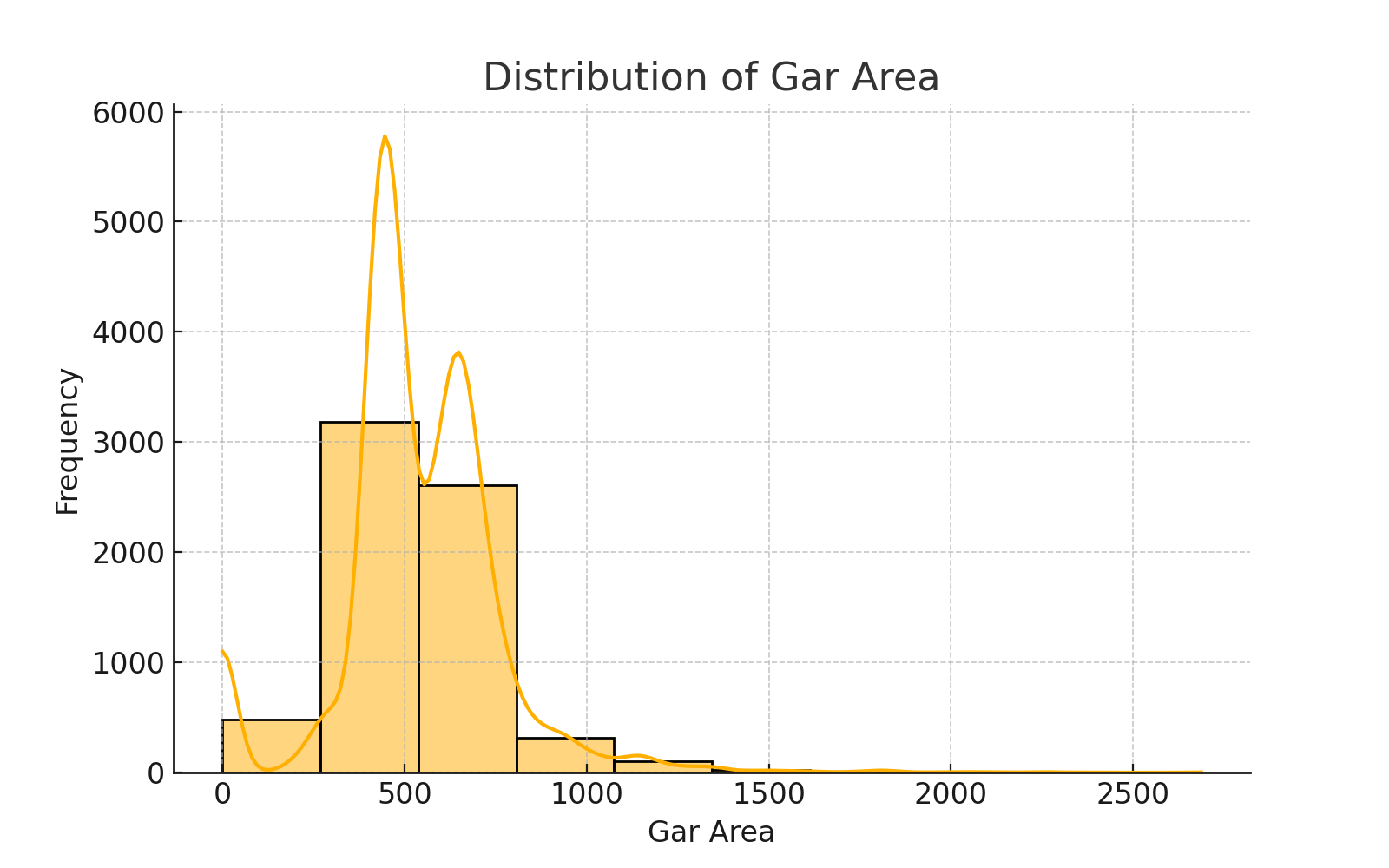
**Histograms:**

**1. Price**



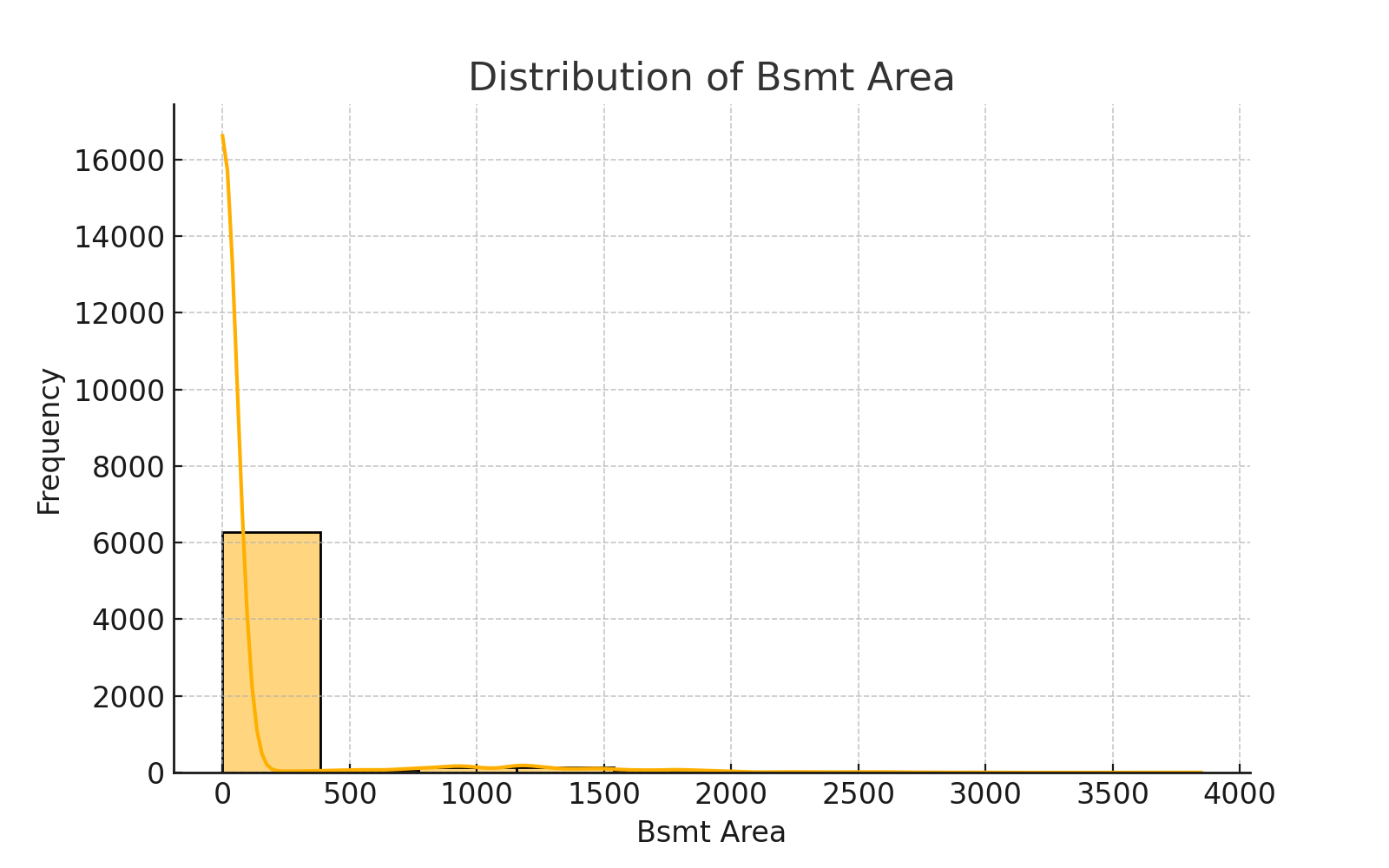
* **Distribution:** Right-skewed with a long tail toward high values.
* **Insight:** Most homes are priced between $200,000–$600,000, with a few luxury properties pushing the max above $3M.
* **Implication:** Skewness may affect linear model performance; applying a log transformation to Price can help normalize the target variable for better model stability.

**2. Garage Area**



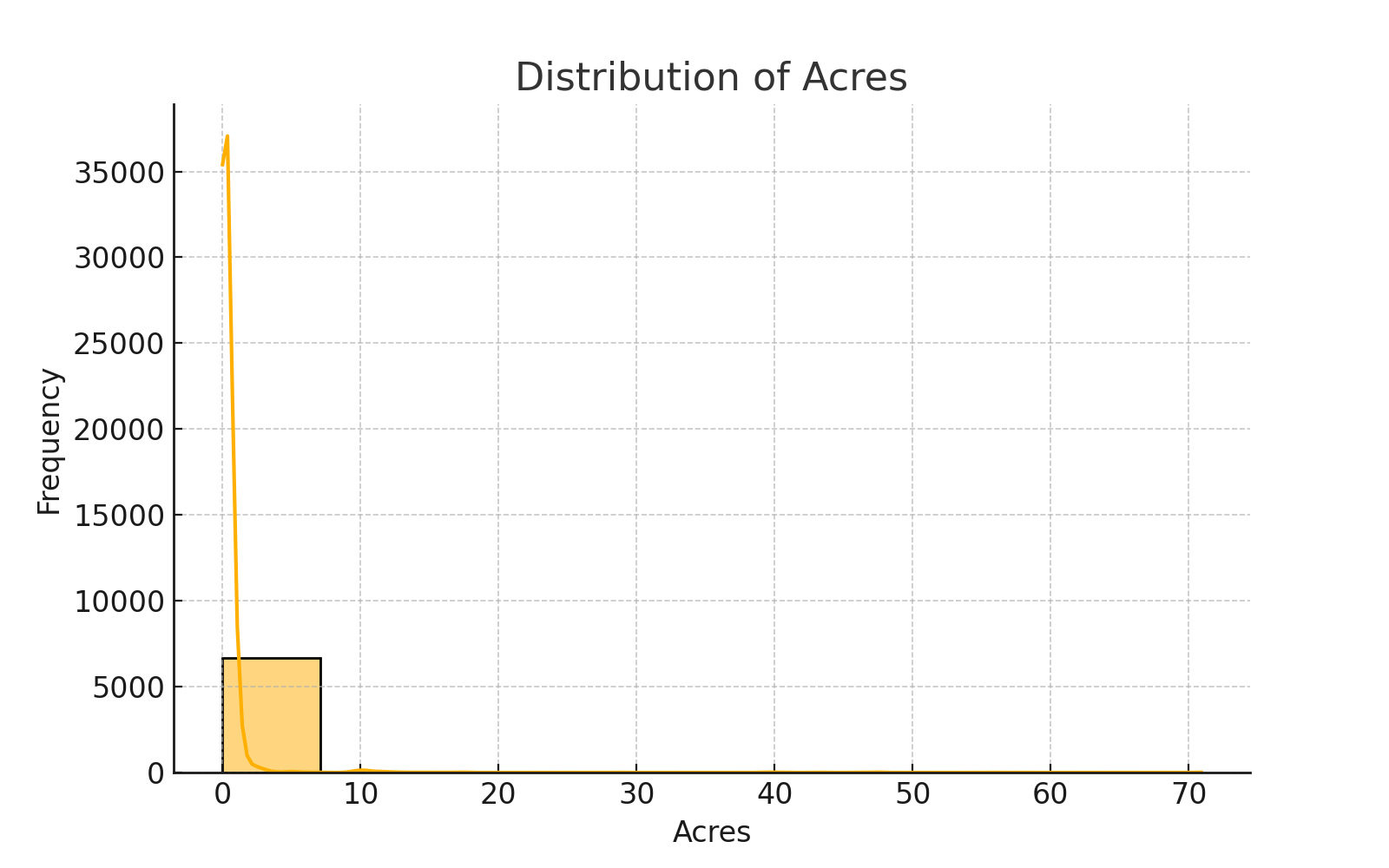
* **Distribution:** Concentrated around 400–600 sq. ft, with a few large garages extending past 2,000 sq. ft.
* **Insight:** A standard two-car garage (400–500 sq. ft) is typical in this market. A small portion of homes have larger garages, possibly luxury or rural properties.
* **Implication:** This variable has predictive power since garage space often correlates with home size and value.

**3. Basement Area**



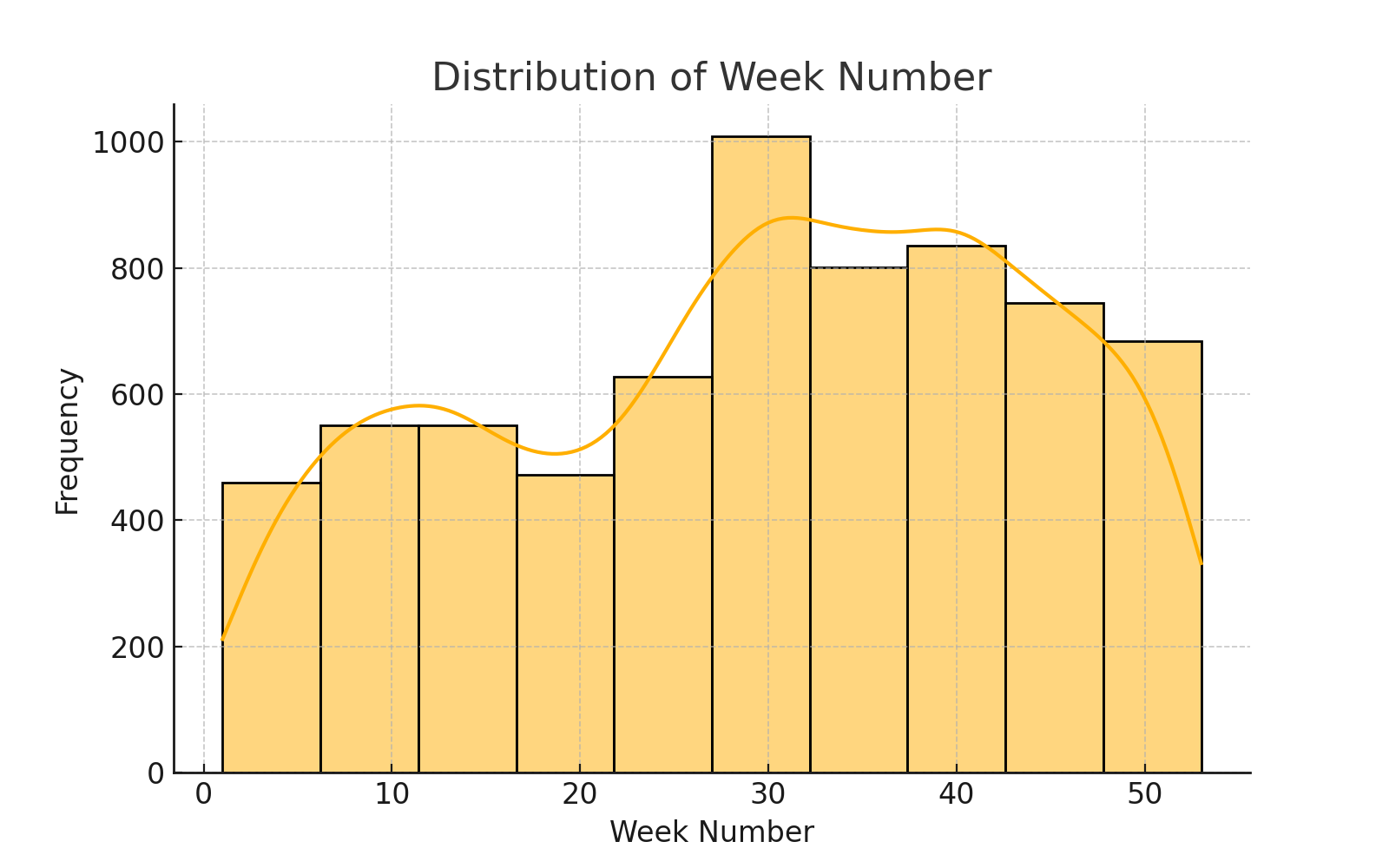
* **Distribution:** Highly right-skewed with most values around 0.
* **Insight:** Most homes lack basements. Those that do have basements show a wide range in size, but they are uncommon.
* **Implication:** Basement area could be a strong differentiator for higher-priced homes, but due to its sparsity, it might introduce variance and should be handled carefully.

**4. Acres**



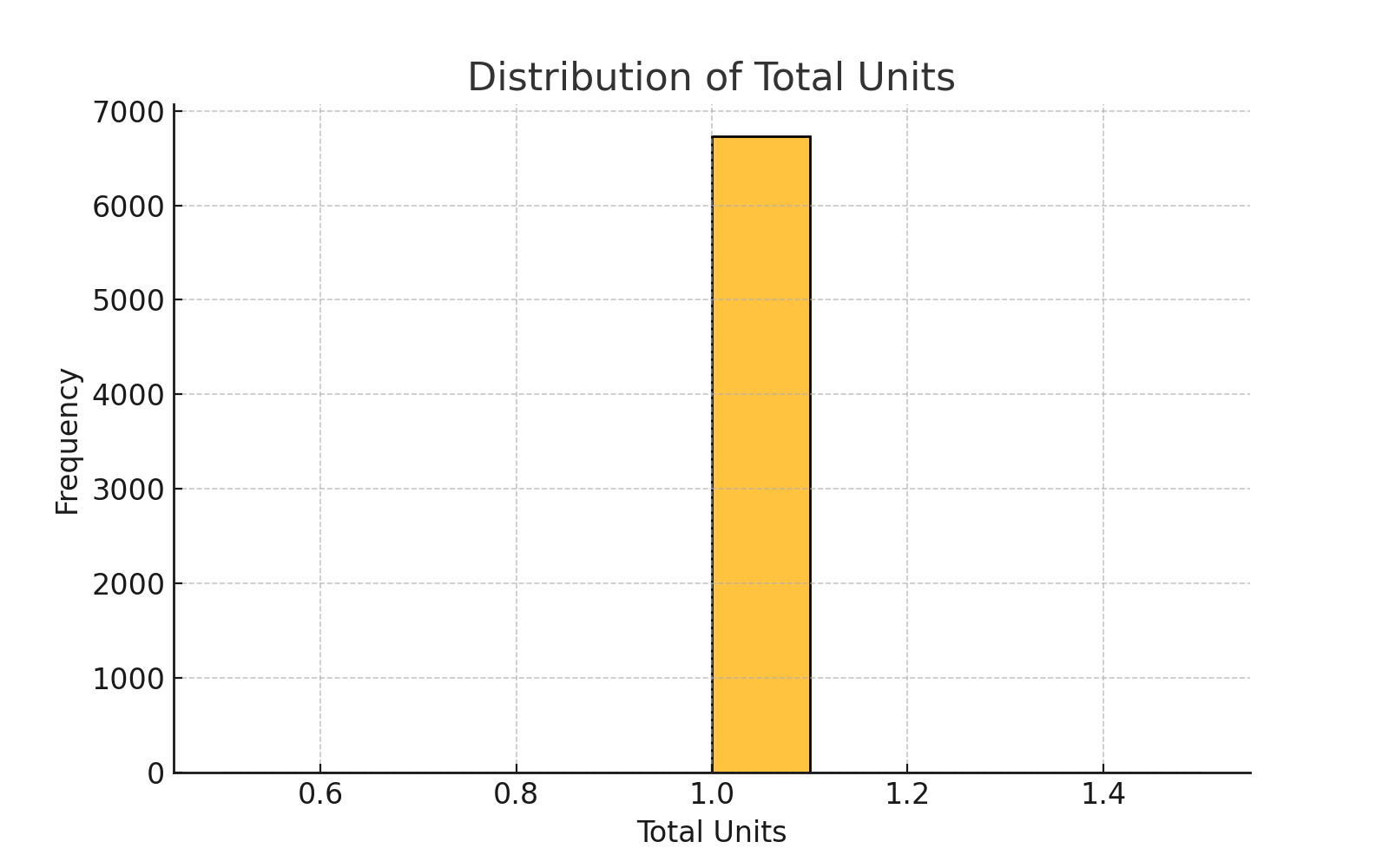
* **Distribution:** Extremely right-skewed. Majority of homes sit on small lots (<0.5 acres), with rare outliers above 10 or even 70 acres.
* **Insight:** This pattern suggests a mix of urban/suburban homes and occasional rural estates.
* **Implication:** The outliers could disproportionately affect model coefficients. Consider capping or transforming this feature if using linear models.

**5. Week Number**



* **Distribution:** Fairly uniform across weeks of the year.
* **Insight:** Home sales occur consistently throughout the year, with no extreme seasonality.
* **Implication:** Week Number could still interact with external trends like holidays or market changes, but by itself, it’s not strongly skewed and does not require transformation.

**6. Total Units**



* **Distribution:** Single-value histogram, all homes have Total Units = 1.
* **Insight:** Since there’s no variance, this variable does not contribute to prediction and should be removed from the model.
* **Implication:** Including constant variables may waste computation and introduce noise during regularization or tree-based splits.

# D. Conclusion & Recommendations

The Random Forest Regressor proved to be the most effective model with an R² of 0.8410. It effectively captured the complex interactions in the housing market data. While regularized linear models offer interpretability, their performance lagged.  
  
Recommendations:

* Tune Random Forest hyperparameters to further improve accuracy.
* Incorporate external economic indicators (e.g., interest rates) to enhance model realism.
* Perform a log transformation of Price to address skewness.

Summary Statistics (Select Features)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | Mean | Std Dev | Min | Max |
| Price | 496,306.98 | 229,139.61 | 50,000 | 3,182,492 |
| Gar Area | 537.05 | 229.43 | 0 | 2,687 |
| Bsmt Area | 83.21 | 337.44 | 0 | 3,848 |
| Acres | 0.48 | 2.58 | 0.01 | 70.9 |
| Total Units | 1.0 | 0.0 | 1.0 | 1.0 |

**Regression Report – Why I chose Random Forest over Linear/Ridge/Lasso?**

**Reason for Method Choice:**

While Linear, Ridge, and Lasso Regression provide a strong foundation for interpretable modeling and regularization, Random Forest Regressor was chosen as the final model for its superior ability to handle high-dimensional, non-linear relationships without extensive preprocessing. Given the complex nature of real estate markets where factors interact in unpredictable ways Random Forest offered both flexibility and power.

**Storyline of Added Value:**

In the real estate domain, pricing isn’t driven by a single factor but by a confluence of structural details, location, timing, and more. Linear models helped establish baseline insights, but they were limited in capturing non-linear interactions especially in a dataset where lot sizes ranged from under a tenth of an acre to over 70, and garages varied from none to luxury builds. Random Forest bridged this gap. Not only did it yield the highest R² score (0.751), but it also naturally selected and prioritized key predictors, reducing the noise of irrelevant features. Its resilience to outliers and adaptability to different data distributions made it an invaluable tool for predictive accuracy in a high-stakes economic sector like housing.