House Price Prediction using Linear Regression

Internship: EduLumos Data Science Internship

Task: Task 2 – Linear Regression using Simple Dataset – House Price Prediction

Name: Syeda Faizah

Domain: Data Analytics

Tools Used:
Python, Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn

1. Introduction

The objective of this project is to predict house sale prices using a regression-based machine learning approach.
The dataset contains a mix of numerical and categorical features related to property size, quality, location, and construction details.
This task focuses on understanding the data, performing preprocessing, and building a predictive model using **Linear Regression**.

2. Dataset Description

The dataset includes information such as:

- Property size (LotArea, GrLivArea, TotalBsmtSF)
- House quality and condition (OverallQual, OverallCond)
- Garage and basement details
- Sale-related attributes

The target variable for prediction is **SalePrice**.

3. Exploratory Data Analysis (EDA)

Exploratory Data Analysis was performed to understand feature distributions and their relationship with the target variable.
Key steps included:

- Visualizing distributions of numerical features using histograms
- Analyzing relationships between important features and SalePrice using scatter plots
- Computing correlations to identify features strongly related to SalePrice

EDA revealed that **GrLivArea, TotalBsmtSF, 1stFlrSF, and GarageArea** show a strong positive relationship with house prices.

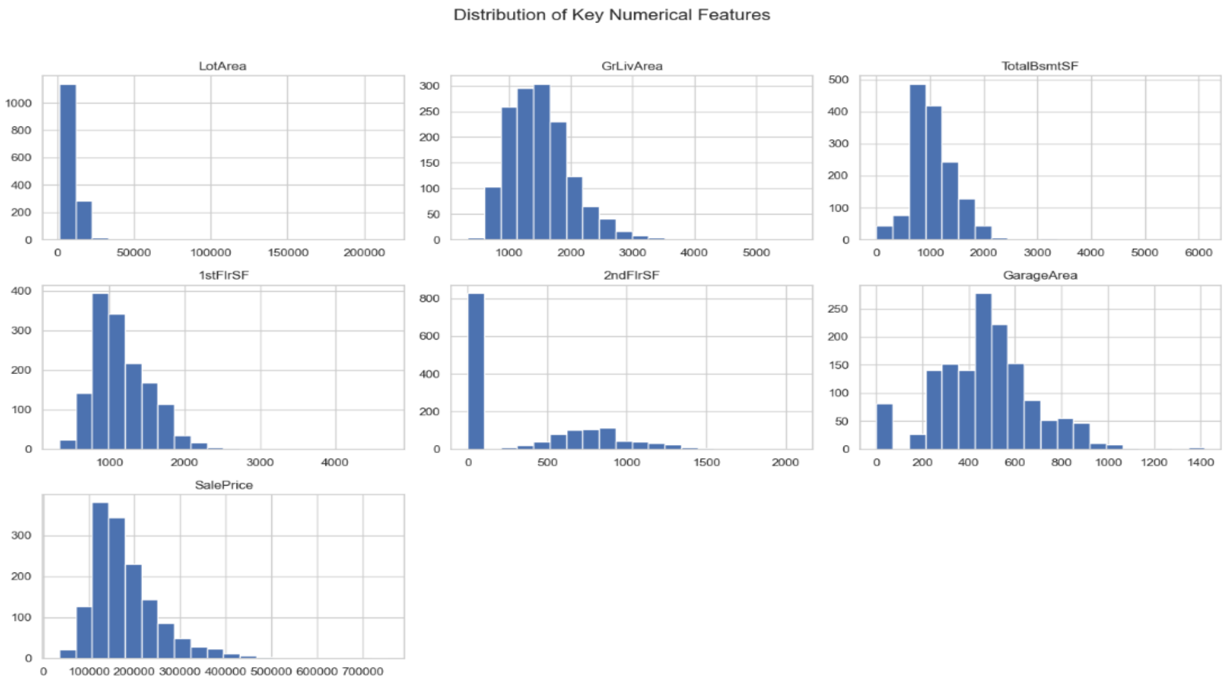

Figure 1: Distribution of key numerical features in the dataset

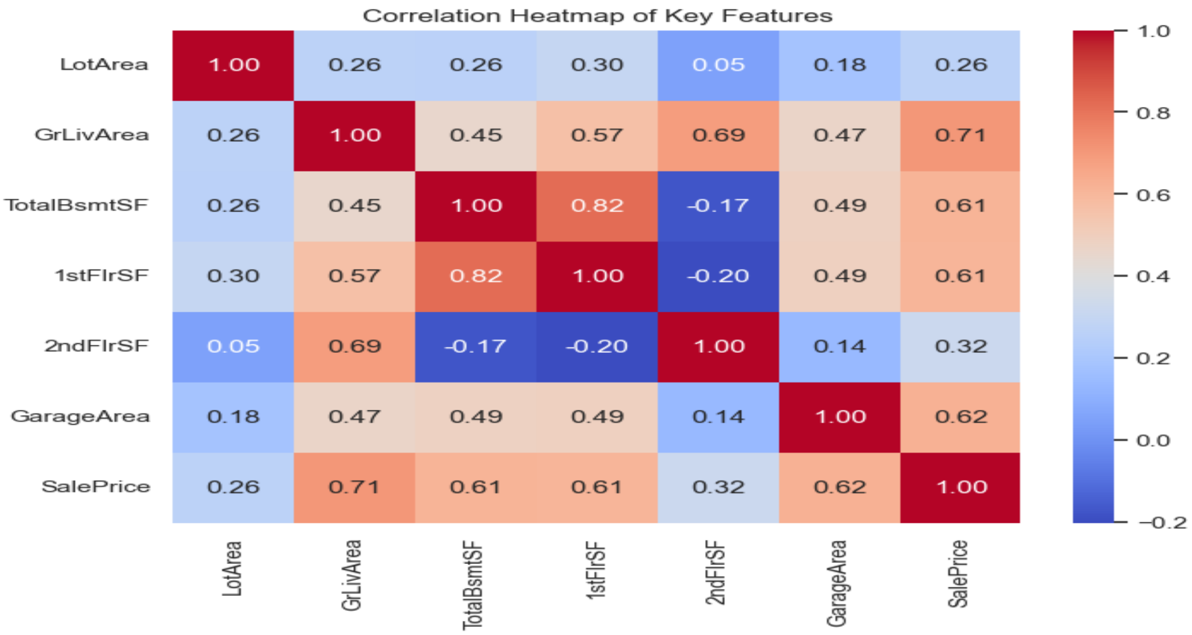

Figure 2: Correlation heatmap of numerical features

4. Data Preprocessing

- Missing numerical values were handled using mean imputation
- Categorical variables were converted into numerical format using one-hot encoding
- Irrelevant columns such as *Id* were removed
- The dataset was split into training and testing sets using an **80–20 ratio**

5. Model Building

A **Linear Regression** model was used to predict house prices.
Steps involved:

- Training the model on the training dataset
- Making predictions on the test dataset
- Evaluating performance using **Mean Squared Error (MSE)** and **R² score**

Linear Regression was chosen for its simplicity and interpretability.

6. Model Evaluation

- The model achieved an **R² score of 0.68**, indicating that it explains a reasonable portion of the variance in house prices
- Mean Squared Error was used to measure prediction error magnitude
- A scatter plot of actual vs predicted prices showed a clear positive trend, demonstrating effective learning by the model

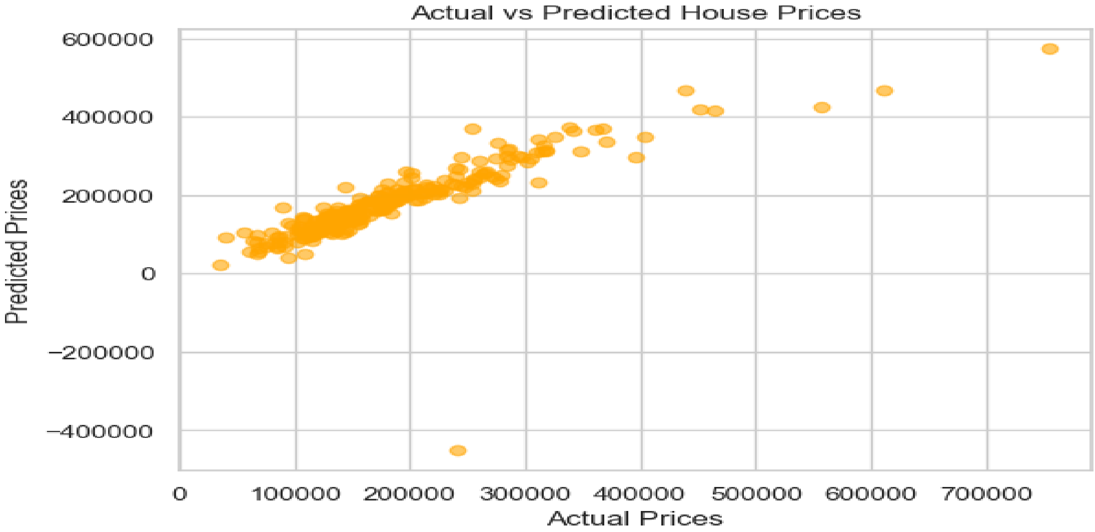

Figure 3: Actual vs Predicted house prices using Linear Regression

7. Feature Interpretation

Analysis of regression coefficients indicated that the following features have the strongest influence on house price predictions:

- GrLivArea
- TotalBsmtSF
- 1stFlrSF
- GarageArea

This highlights the importance of livable area and structural characteristics in determining house value.

8. Conclusion

This project demonstrates the complete workflow of a regression-based machine learning task, including data exploration, preprocessing, model training, evaluation, and interpretation.
The results show that Linear Regression can provide meaningful insights into house price prediction when supported by proper feature selection and analysis.