House Price Prediction Using Regression Models

Name: Mena Rossini R
Reg No: *21222040099*

---

1. Introduction and Objective

In this project, I used a dataset about house prices in Ames, Iowa, to build models that can predict the sale price of a house based on certain features. The goal was to explore different linear regression models and figure out which one does the best job of predicting prices accurately.

This kind of analysis can be helpful for real estate agents, property developers, or anyone trying to buy or sell a home, as it shows which features have the biggest impact on a home's price.

---

2. About the Dataset

The dataset comes from a Kaggle competition called "House Prices: Advanced Regression Techniques." It contains 1,460 rows and 80 features about homes that were sold in Ames, Iowa. Each row represents a different house, and the goal is to predict the `SalePrice`.

For this project, I chose a small set of important features to keep things simple and easier to interpret:

- GrLivArea – Above ground living area in square feet
- OverallQual – Overall material and finish quality (rated 1–10)
- GarageCars – Number of cars the garage can hold
- TotalBsmtSF – Total basement square footage
- YearBuilt – Year the house was built

I dropped any rows with missing values in these columns. For models that need it, I also standardized the features using a scaler.

---

3. Models Used and How They Performed

In this analysis, I used linear regression models like **Simple, Multiple, Ridge, and Lasso**. Other models like Random Forest, Gradient Boosting, and Neural Networks were not used because they are more complex and require more tuning. The focus was on linear models for simplicity and easier interpretation. Trying those advanced models could be a next step to improve accuracy.

🔹 Simple Linear Regression

- Used only one feature: `GrLivArea`
- Easy to understand but not very accurate

```
Simple Linear Regression (GrLivArea):
R²: 0.554
RMSE: 58471.76
```

🔹 Multiple Linear Regression

- Used all five feature.
- A noticeable improvement over the simple model

```
Multiple Linear Regression:
R²: 0.794
RMSE: 39763.30
```

🔹 Ridge Regression

- Like multiple linear regression, but adds regularization to reduce overfitting
- Best performance overall

```
Ridge Regression:
R²: 0.793
RMSE: 39807.84
```

🔹 Lasso Regression

- Another regularized model, also helps with feature selection
- Very similar to Ridge, slightly worse in this case

```
Lasso Regression:
R²: 0.794
RMSE: 39763.30
```

Model Performance Summary Table:

	Model	R^2 Score	RMSE
0	Simple Linear Regression	0.554263	58471.756526
1	Multiple Linear Regression	0.793865	39763.295266
2	Ridge Regression	0.793403	39807.843933
3	Lasso Regression	0.793865	39763.297972

---

4. Key Insights

After testing the models and looking at the results, I found that:

- `OverallQual` (the quality of the house) and `GrLivArea` (size of living space) had the biggest impact on price.
- Adding more features improved the prediction a lot compared to just using one feature.
- Ridge Regression performed the best, likely because it balances model complexity and accuracy well.
- Lasso was also effective and could be helpful if I had more features, since it helps with feature selection.

---

5. Limitations and Next Steps

While the model did pretty well, there are still some limitations:

- The model doesn't do as well with very expensive or unique houses (possibly due to outliers).
- I only used 5 features — there are many more in the dataset that might improve predictions.
- The model assumes linear relationships, which might not capture all patterns in the data.

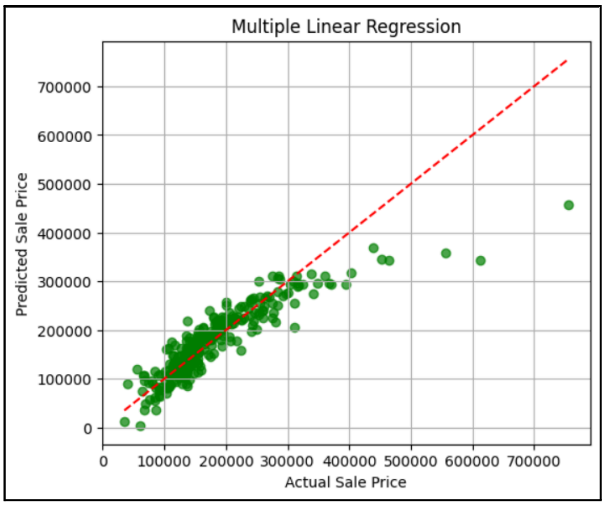
If I continue working on this, I'd like to:

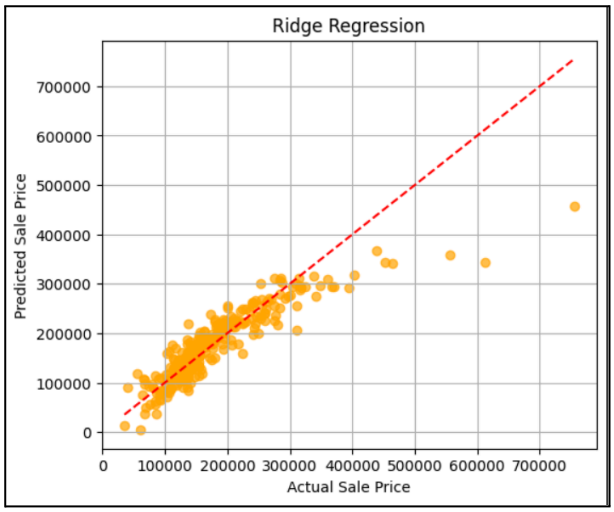
- Try models like Random Forest or XGBoost that can handle non-linear relationships better
- Use more features and possibly create new ones (like age of the house)
- Look more into the outliers and maybe remove or handle them differently

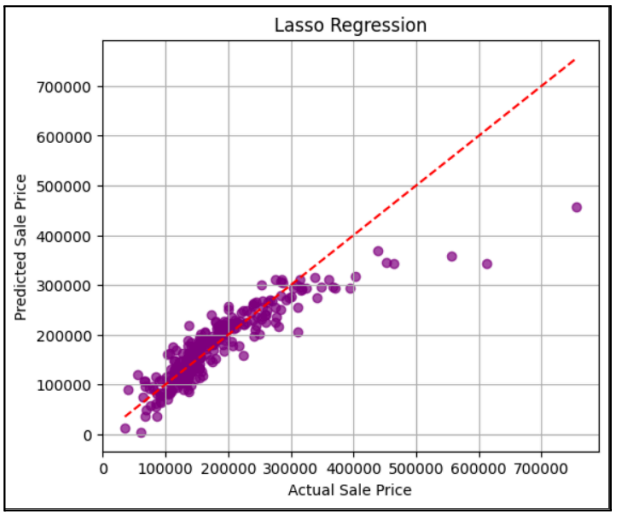
---

6. Visuals (Actual vs. Predicted)

I created scatter plots for each model to see how close the predicted prices were to the actual ones. The closer the dots are to the red diagonal line, the better the model performed.







Conclusion

Overall, this project helped me understand how different regression models work and how to compare them. Ridge Regression turned out to be the best model for this dataset, and the project gave me a better idea of what features influence house prices the most.

GitHub Link: [Regression_Hands_On]
DataSet Link: [Housing_DataSet]