Predicting Home Prices in Ames, Iowa: Machine Learning for Real Estate

To Fit or Not to Fit?

Background

The Ames Housing dataset is a widely used benchmark in the field of machine learning for real estate valuation. Compiled by Dr. Dean De Cock in 2011, it offers a rich resource for analyzing factors influencing single-family home prices in Ames, Iowa.

- Snapshot of the housing market during the mid-2000s, between 2006 and 2010.
- Reveals how location and size characteristics influence housing prices.
- Features like overall quality, number of bedrooms/bathrooms, and presence of amenities like fireplace or central air, provide insights into what buyers valued in Ames.

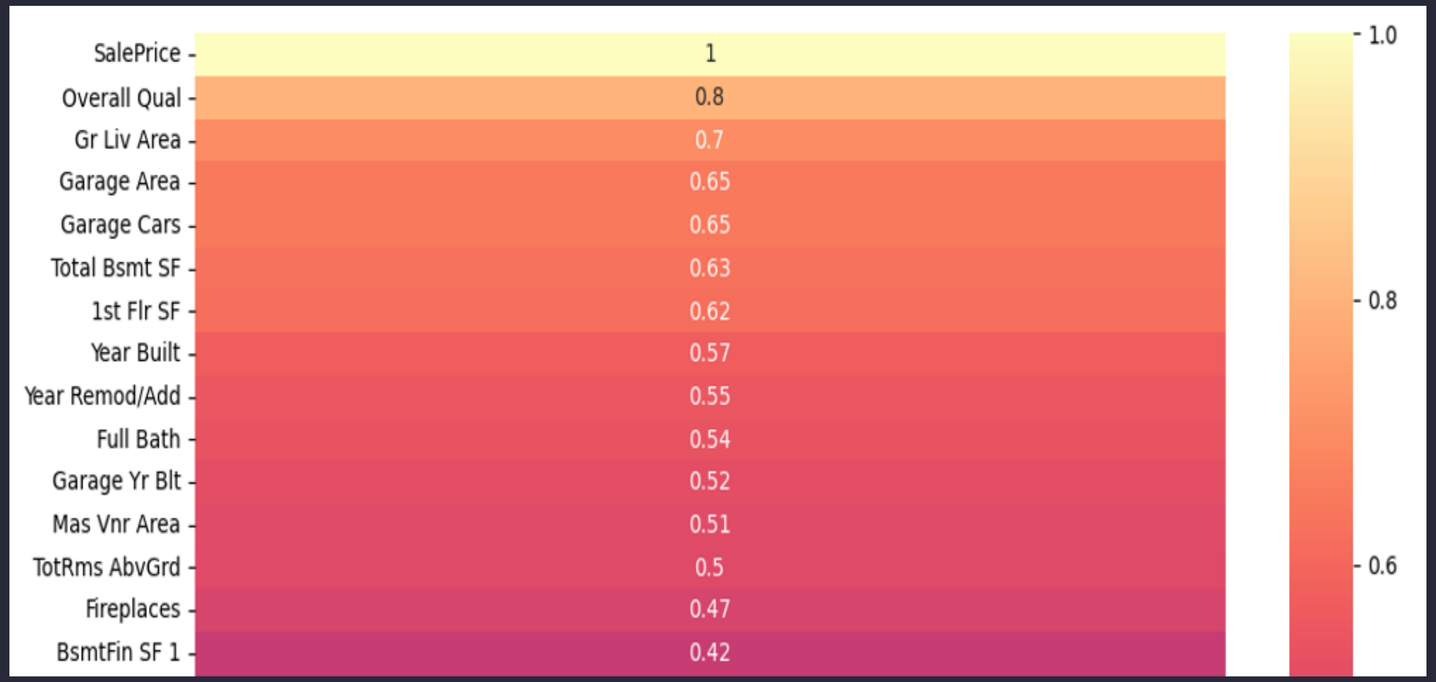
Problem Statement

Develop a machine learning model that accurately predicts the sale price of single-family homes in Ames, Iowa. Predictive modeling with regression is a valuable tool for real estate professionals and homeowners in the area.

Success Evaluation:

- Accuracy: a low Root Mean Squared Error (RMSE), indicates the model's ability to predict sale prices close to the true values.

Numerical Data

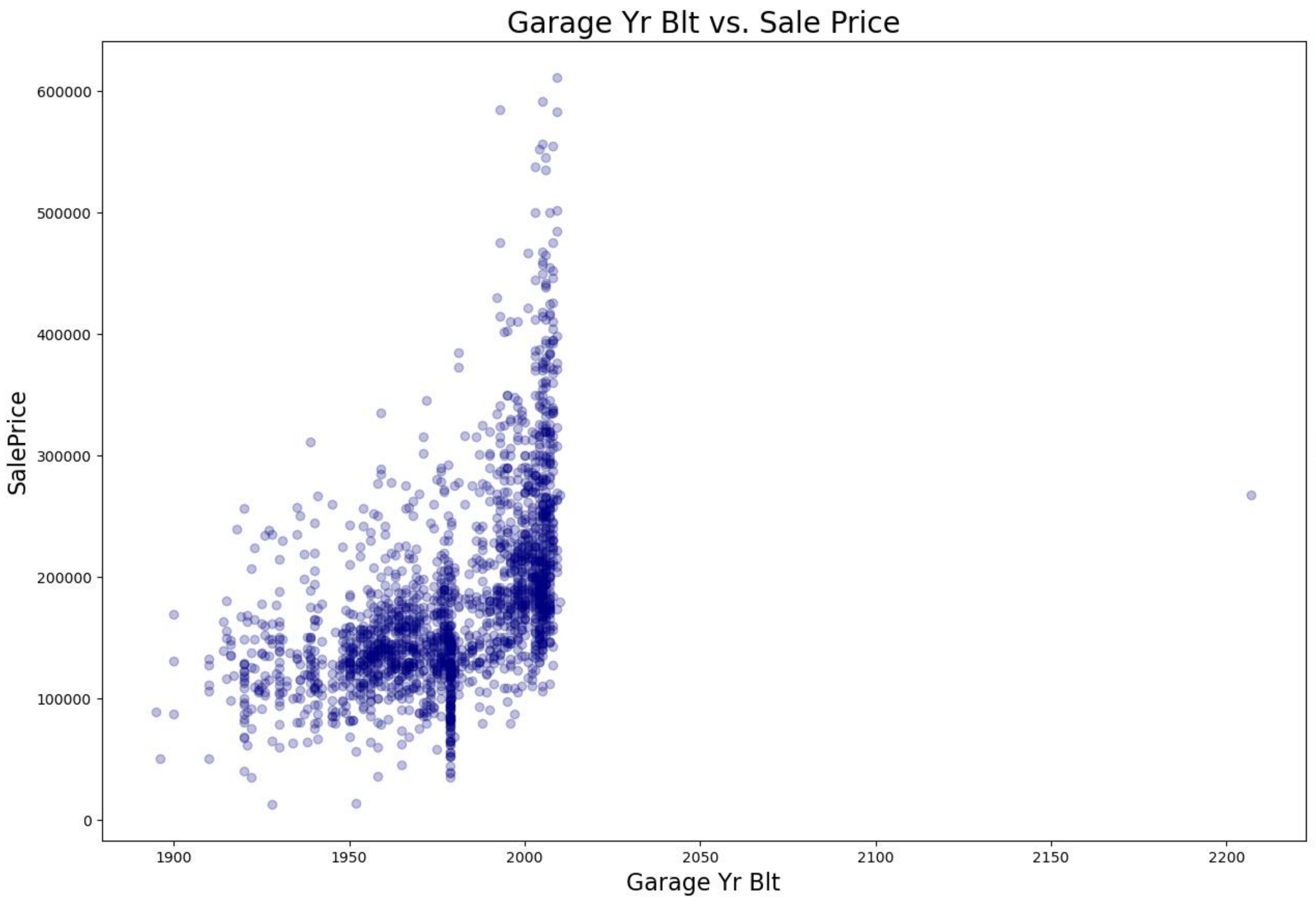


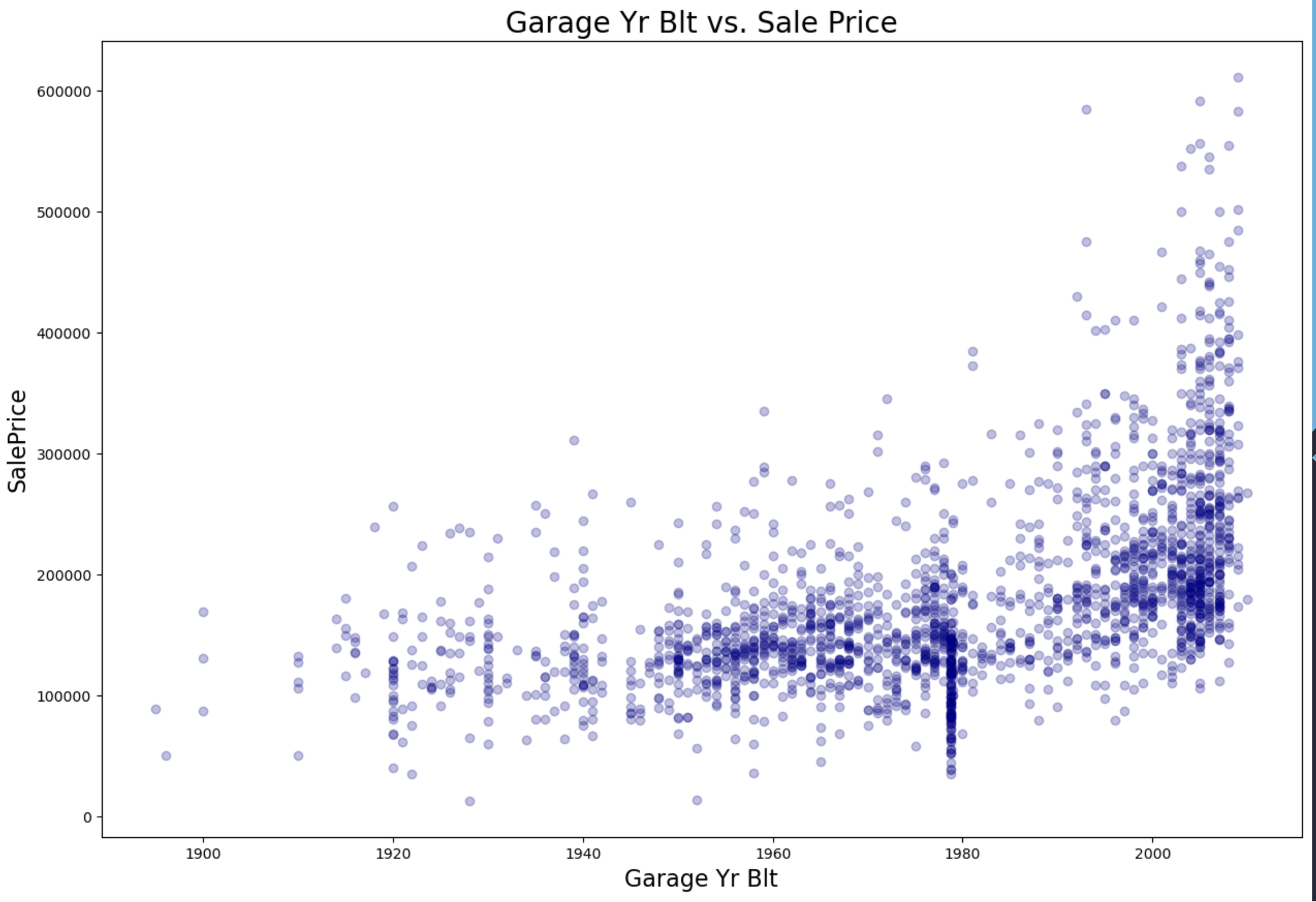
Correlation Coefficient:

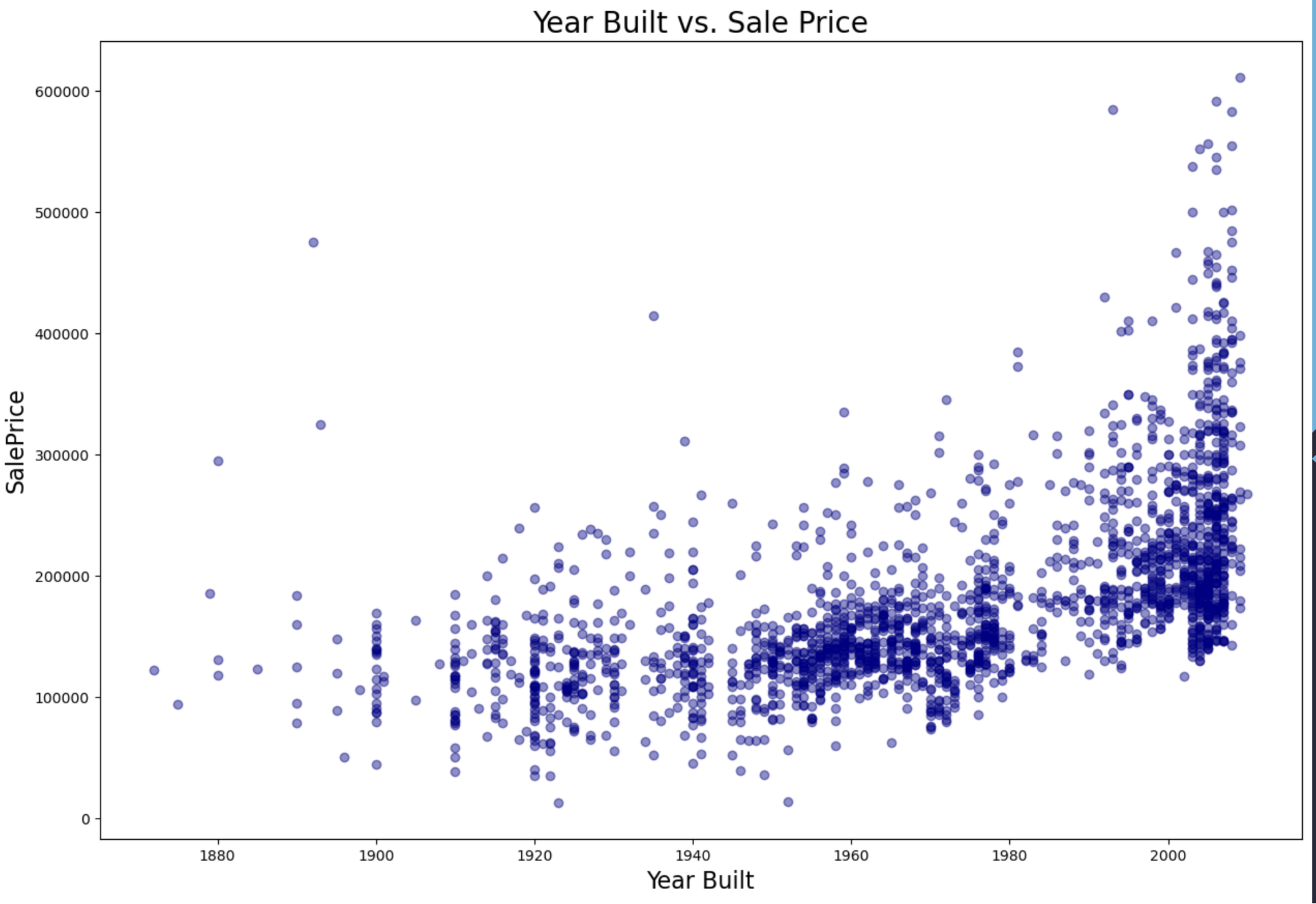
- Sorted by strength.
- Use the top numerical features for modeling.
- Plot graphs of features that are relevant in readable heatmap.

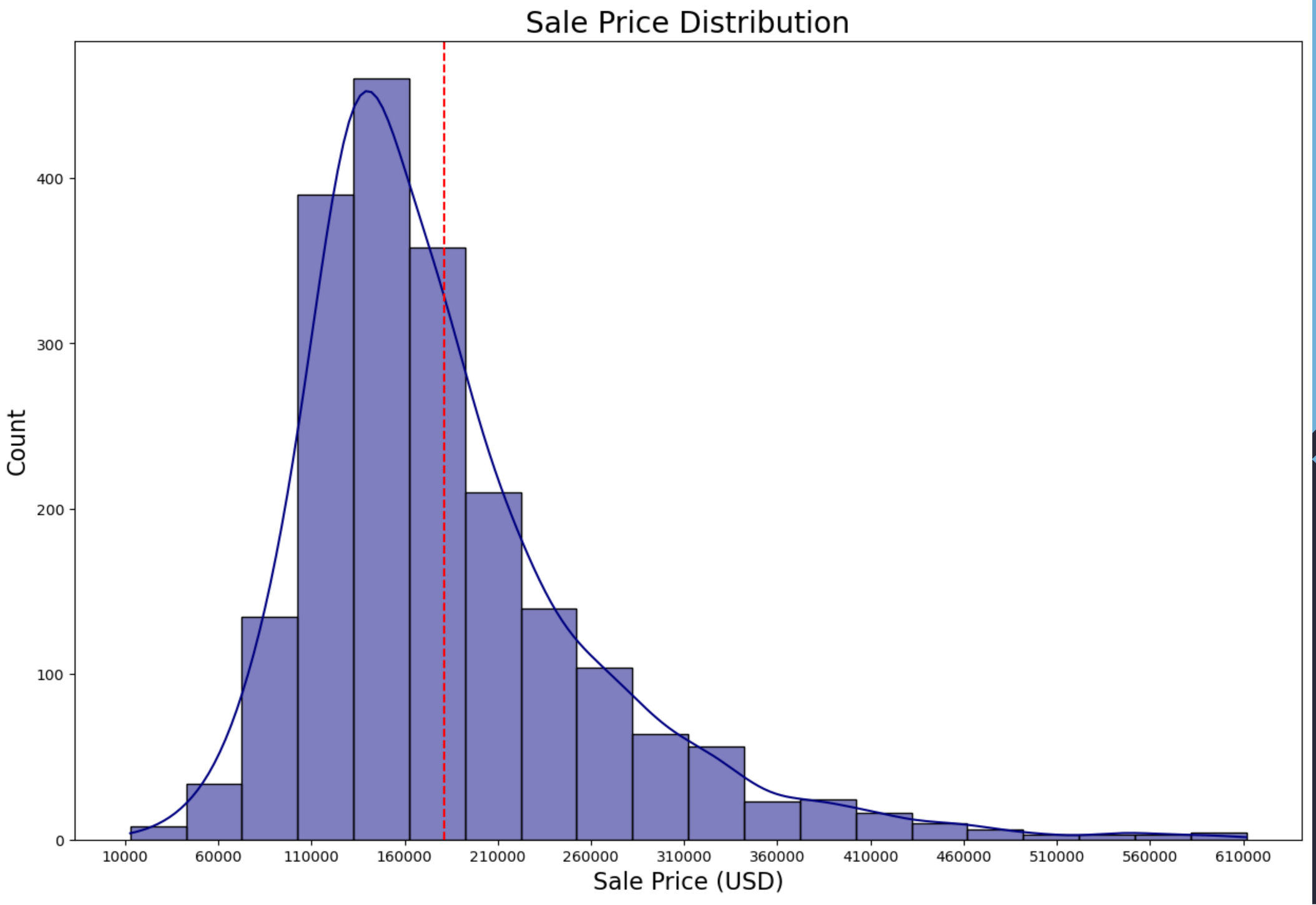
Models (Candidates):

- Multilinear Regression
- Ridge Regression
- Lasso Regression
- ElasticNet Regression


Garage Yr Blt vs. Sale Price
SalePrice
600000
500000
400000
300000
200000
100000
0
1900
1950
2000
2050
2100
2150
2200
Garage Yr Blt


Garage Yr Blt vs. Sale Price
SalePrice
600000
500000
400000
300000
200000
100000
0
1900
1920
1940
1960
1980
2000
Garage Yr Blt


Year Built vs. Sale Price
SalePrice
600000
500000
400000
300000
200000
100000
0
1880
1900
1920
1940
1960
1980
2000
Year Built


Sale Price Distribution
Count
0
100
200
300
400
10000
60000
110000
160000
210000
260000
310000
360000
410000
460000
510000
560000
610000
Sale Price (USD)


Sale Price Distribution: Box Plot
1
25000
75000
125000
175000
225000
275000
325000
375000
425000
475000
525000
575000
625000
Sale Price (USD)

Model Training

From the four candidates, Lasso Regression ends up as the best model at predicting Sale Price values. Let's look at relevant points on this modeling technique:

- Accuracy: Lasso minimizes the errors between predicted and actual prices. This ensures your model makes good predictions. (Think of it as real estate agent getting the ballpark price right.)

- **Feature Selection:** Unlike regular regression, Lasso adds "shrinkage" penalty. It can push some coefficients to zero, effectively removing them from the model. (Think of the agent focusing only on crucial aspects like size and location.)

- Trade-off: Lasso might eliminate a relevant factor if it's weakly correlated with other features. Compare Lasso with other models like Ridge Regression, which shrinks coefficients but doesn't eliminate them, can be valuable.

Model Evaluation: Test Metrics

Metric (ElasticNet)	Value
R - Squared	0.8504101451655979
Mean Absolute Error	20665.268367674693
Median Absolute Error	13833.485296685321
Mean Squared Error	888790440.2236749
Root Mean Squared Error	29812.588619971848

- R Squared (0.85): Imagine hitting a bullseye on a dartboard. An R-squared of 0.85 means the Lasso model is landing 85% of its predictions close to the center!
- Mean/Median Absolute Error ($20,665/$13,833): Think of this as the average difference between the predicted price and the actual price. A lower number means your predictions are typically within a few ten thousand dollars of the real value.
- Root Mean Squared Error ($29,813): For each house, calculate the difference between predicted price and actual price. Square those differences (account for negative errors), take average of those squared values, and finally take the square root. RMSE of $29,813 tells us that on average, the predictions are off by about $30,000 from the actual prices.

Model Evaluation: Test Metrics

Metric (Linear)	Value
R - Squared	0.8871324703612562
Mean Absolute Error	19059.089077633616
Median Absolute Error	14680.672247606039
Mean Squared Error	670604176.1029077
Root Mean Squared Error	25896.026260855306

Metric (Ridge)	Value
R - Squared	0.8872678653805913
Mean Absolute Error	19033.69638362567
Median Absolute Error	14572.504790880485
Mean Squared Error	669799724.4977359
Root Mean Squared Error	25880.489263105825

Metric (Lasso)	Value
R - Squared	0.8874283150793292
Mean Absolute Error	19039.356262231824
Median Absolute Error	14552.458647036576
Mean Squared Error	668846410.1266992
Root Mean Squared Error	25862.06507854118

Predictive Insights

	Coefficient
Overall Qual	19592.179437
Gr Liv Area	30637.903451
Garage Area	3613.118291
Garage Cars	3737.037427
Total Bsmt SF	17592.113083
Year Built	3678.452932
Year Remod/Add	13744.473126
Full Bath	-1370.370783
Half Bath	5507.341258
Garage Yr Blt	-644.201221
Mas Vnr Area	2766.668182
Bedroom AbvGr	-6064.639640
TotRms AbvGrd	2612.418961

Here are some interesting conclusions from the lasso coefficients listed in the DataFrame:

- holding all else constant, for every one car increase in Garage Cars, expect the Sale Price to increase by about $3,740.
- holding all else constant, for every one year increase in Year Built, expect the Sale Price to increase by about **$3,680**.
- holding all else constant, for every increase in value by 1, in Overall Quality, expect the Sale Price to increase by about **$19,590**.

Conclusion

Lasso Regression Model gives the lowest RMSE of 25862.065! Expect the Kaggle Score to be different on the leaderboards.

From the looks of it, this might generalize quite well on Kaggle's new test data.

Kaggle Challenge: (Late Submission)

- Public Score - **27447.39733**
- Private Score – **29633.40227**

Limitations:

- Did not use any dummy variables in modelling process.
- No function to find optimal alpha for ridge or lasso.

Citations

- De Cock, Dean. (2011). Ames Housing Dataset. www.kaggle.com/datasets/marcopale/housing
- Wickham, H., & Grothmann, C. (2016). *Tidy Modeling with R* www.tmwr.org/ames
- Tan, Alvin T., PhD. (2022). Cracking the Ames Housing Dataset With Linear Regression. www.towardsdatascience.com/wrangling-through-dataland-modeling-house-prices-in-ames-iowa-75b9b4086c96