Problem and Goal:

Our goal is to develop a predictive model to estimate used car prices using features from our historical vehicle sales data. We will start by examining a sample of the dataset to identify any data issues and the attributes that describe car features and prices.

Following this analysis, our focus will shift to feature selection. This step involves identifying the most relevant features that significantly influence car prices. With these selected features, we will proceed to build a predictive model designed to utilize this information effectively.

Finally, we will use the insights generated by the model to understand which features used car buyers value the most. By following this comprehensive approach, we aim to uncover patterns and relationships that can inform strategic decisions for the used car dealership, ensuring they meet market demands and optimize pricing strategies.

Findings Summary:

I analyzed the data to understand what consumers value in used cars. Key steps in data preparation included dropping non-essential fields like ID, VIN, region, and state to streamline the dataset. Adding regional information back could be an interesting future exercise.

I also created a new car\_age field, which essentially calculates the age of the car based on its year. I made sure to remove rows that were missing all feature columns to keep the data clean and accurate. Interestingly, I found that using the median for price imputation rather than the mean increased the Mean Squared Error (MSE), so I went with the median. My strategy for handling missing data involved using the "most frequent" values for categorical features and the "mean" for numeric features. Outliers were removed to improve the model’s accuracy, and I adopted a logarithmic scale for better visualization of the results.

For modeling and feature selection, I compared the Mean Squared Error (MSE), R², and cross-validation scores across Ridge, Linear, and Lasso Regression models. Ridge and Linear Regression performed equally well, so I selected Ridge Regression to better help deal with features with very low coefficients. Assessing the impact of making car\_age and odometer polynomial or interactive features showed no significant improvement in the MSE score. I used CrossGridCSV for model validation and employed Random Forest to double-check feature importance.

The MSE values for both training and test data were 0.35, indicating consistent performance on both known and unseen data. The R² score of 0.65 shows that approximately 65% of the variance in price can be explained by the model's features, which is strong. The consistency between the test MSE and cross-validation mean MSE further supports the model’s reliability and generalizability.

The major factors negatively impacting car prices were Car Age and Odometer. This aligns with common expectations, as older cars and those with higher mileage generally sell for less. Other features like Model, Title Status, Cylinders, and Fuel had much smaller coefficients, indicating they have less influence on the price. Among these, the Model feature had the highest positive coefficient, but its impact was minimal compared to Car Age and Odometer.

Moving forward, focus on car age and odometer readings in your marketing and valuation strategies, as these are the most significant factors influencing car prices. Highlighting attributes like the car model and title status can also add value, even if their impact is smaller. Consider adding regional data back for deeper analysis, such as reviewing if dealership placement influences prices. Lastly, continuously update the model with new data to stay aligned with market trends and consumer preferences, and revisit when new data is available.

**For details, please see FinalCarPrice.ipynb and be sure to load data\vehicles.csv for code execution.**