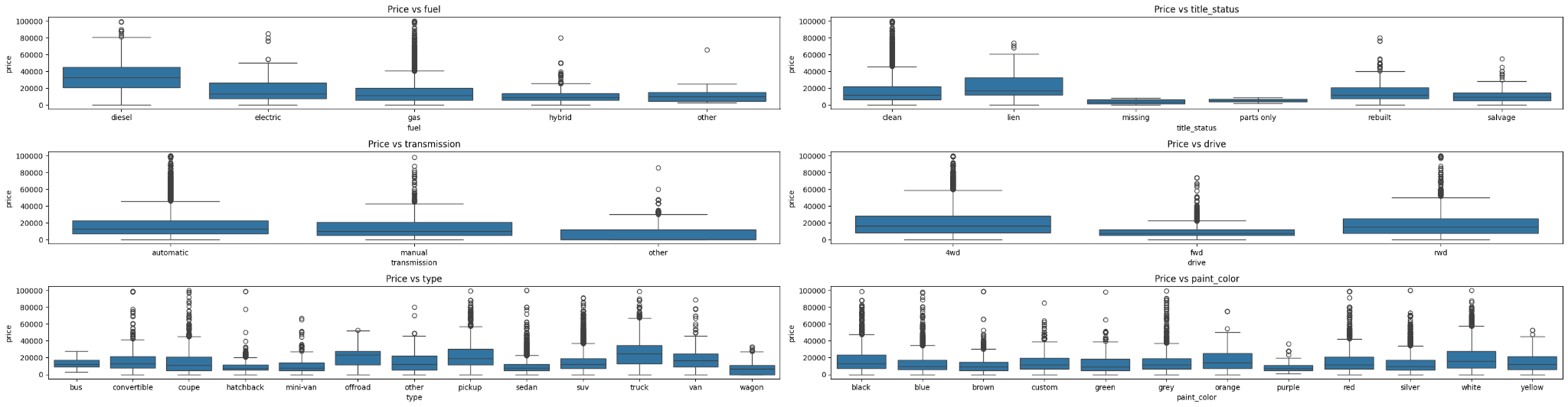


The histogram exhibits a right-skewed distribution, also known as a positively skewed distribution. This means that the majority of the car prices are concentrated towards the lower end, with a long tail extending towards the higher prices.

Key Observations:

* Peak: The highest bar in the histogram represents the most frequent price range, which appears to be between $0 and $10,000. This suggests that a large number of used cars are priced within this range.
* Tail: The long tail on the right side indicates the presence of some used cars with significantly higher prices compared to the majority. These could be luxury cars, rare models, or vehicles with special features.
* Spread: The overall spread of the histogram shows the range of used car prices in the dataset. The prices seem to range from near $0 to around $100,000, with a few outliers potentially exceeding that.

============================



These box plots provide valuable insights into how Price is influenced by various categorical factors.

Each box plot visualizes the distribution of Price for a specific category within a categorical variable.

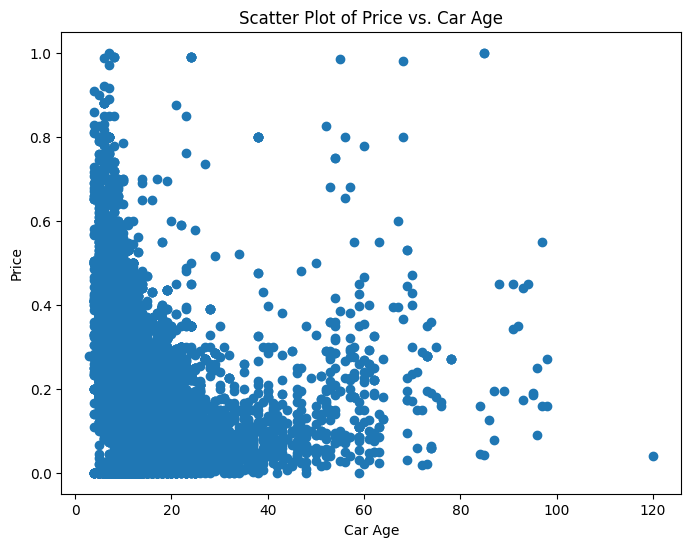
The box itself represents the interquartile range (IQR), containing the middle 50% of the Price data for that category.

The line inside the box is the median Price.

The whiskers extend to the most extreme data points within 1.5 times the IQR from the box's edges.

Individual points beyond the whiskers are considered potential outliers.

=============



Car age is the new feature created from the year

The scatter plot shows a general negative trend between Price and Car Age. This means that as Car Age increases, Price tends to decrease. This is a common observation in the used car market, as newer cars generally command higher prices.

We could calculate the correlation coefficient to quantify the strength and direction of the linear relationship between Price and Car Age.

===============================================

Findings 👍

LinearRegression Model

<class 'pandas.core.frame.DataFrame'>

Index: 67186 entries, 126 to 34850

Columns: 5280 entries, car\_age to cylinders\_other

dtypes: float32(1), float64(5279)

memory usage: 2.6 GB

Mean Squared Error: 664319962.6844393

R-squared: -3.9602328277333703

The linear regression model on the used car dataset seems to be performing poorly based on the evaluation metrics you provided:

* Mean Squared Error (MSE): 664319962.6844393 - This is a very high MSE, indicating that the model's predictions are far off from the actual prices.
* R-squared: -3.9602328277333703 - A negative R-squared is a clear sign that the model is not fitting the data well. It suggests that the model's predictions are worse than simply using the mean price as a prediction for all cars.

==================================

RadomRegression Model

Mean Squared Error: 95193894.02320598

R-squared: 0.28922220517709873

Interpretation of Metrics:

* MSE: The MSE measures the average squared difference between the predicted and actual car prices. A lower MSE indicates better accuracy. Compared to the Linear Regression model you previously mentioned (which had an MSE of 664319962.68), this MSE is significantly lower, suggesting that the Random Forest model is making more accurate predictions.
* R-squared: The R-squared value represents the proportion of variance in the target variable (car price) that is explained by the model. An R-squared of 0.29 means that the model explains about 29% of the variability in car prices. While this is not a very high value, it's a considerable improvement over the negative R-squared obtained with the Linear Regression model.

Analysis:

* Improved Performance: The Random Forest model is showing better performance than the Linear Regression model, as evidenced by the lower MSE and positive R-squared. This suggests that the Random Forest is better able to capture the complex relationships in the data.
* Room for Improvement: While the Random Forest model is an improvement, an R-squared of 0.29 indicates that there is still a significant portion of the price variability that the model is not capturing.

The Random Forest Regression model is a step in the right direction compared to the Linear Regression model. However, there may still be room for improvement.

=================