Used Car Price Analysis

Shuai Yuan

2025-11-21

Introduction

This report analyzes a used car data containing information on model year, mileage, horsepower, fuel type, color, transmission type, accident history, brand characteristics, and price.

The analysis includes:

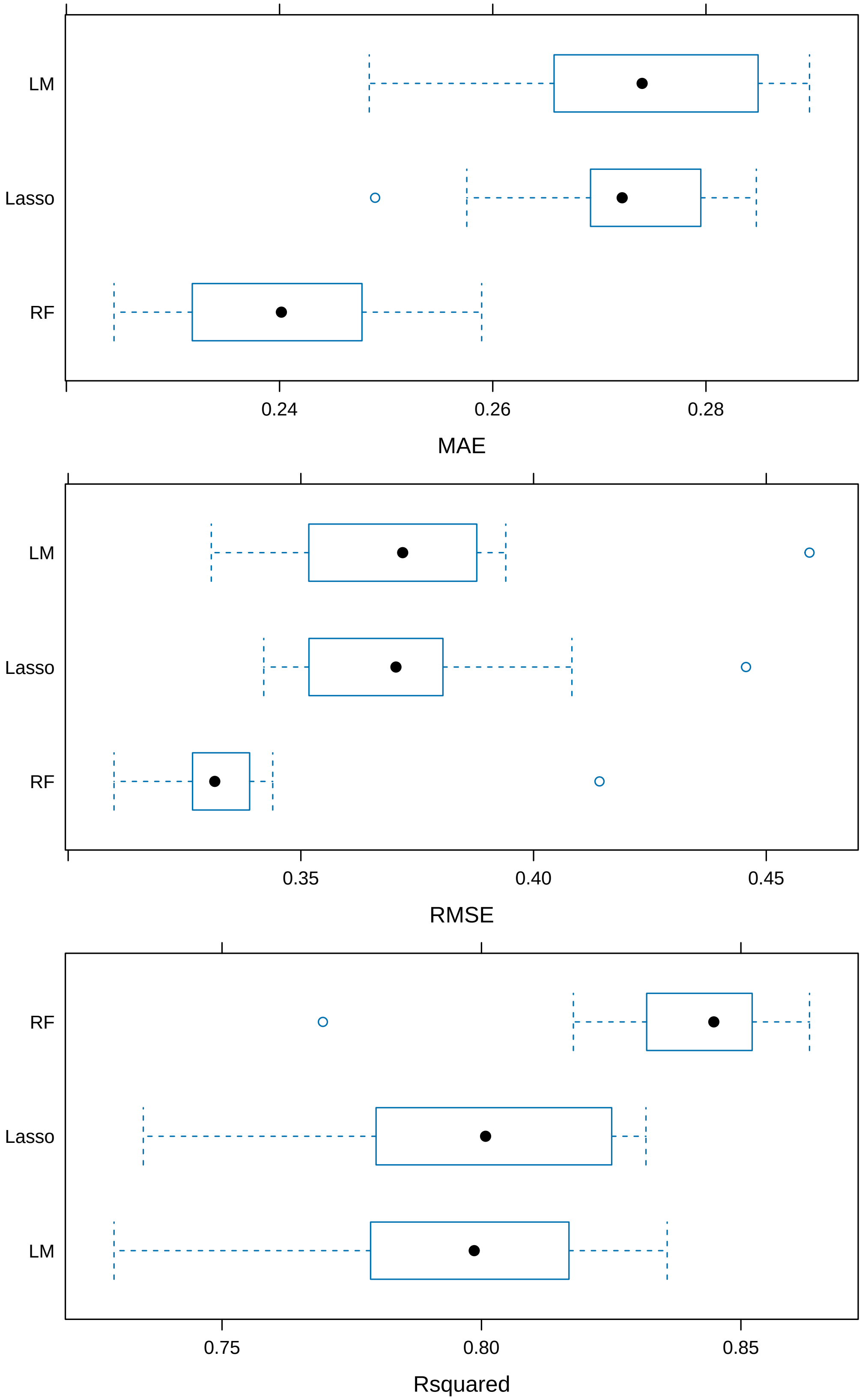
1. Data Processing
2. Predictive modeling (Linear Regression, Lasso, Random Forest)
3. Correlation analysis and clustering

Data Processing

The raw data contains 4009 observations and 12 variables, including vehicle brand, model year, mileage, engine description, multiple color fields, transmission type, accident history, and advertised price. Because several variables contained dozens of messy or extremely sparse categories, I applied targeted cleaning steps. Specifically, I extracted numeric mileage and horsepower, collapsed fuel types into three groups (Gasoline, Diesel, Hybrid), simplified transmission into Automatic/Manual/CVT, and merged exterior and interior colors into a small set of major categories, grouping rare levels into "Other." Accident history was converted to a binary indicator, and brand was compressed into a four-tier price-based category. After cleaning and removing incomplete rows, I obtained 2895 complete cases with the following analysis variables: brand_tier, model_year, mileage, fuel_type, HP, transmission, ext_color, int_color, accident, log_price.

Predictive modeling

I fitted three predictive models for log(price) – a linear regression, a lasso regression, and a random forest—using 10-fold cross-validation via the *caret* package. Across the resamples, the random forest achieved the best predictive accuracy, with the lowest average MAE (~0.24) and RMSE (~0.34) and the highest R^2 (around 0.84), while the linear and lasso models performed similarly with MAE around 0.27, RMSE around 0.38, and R^2 around 0.80.

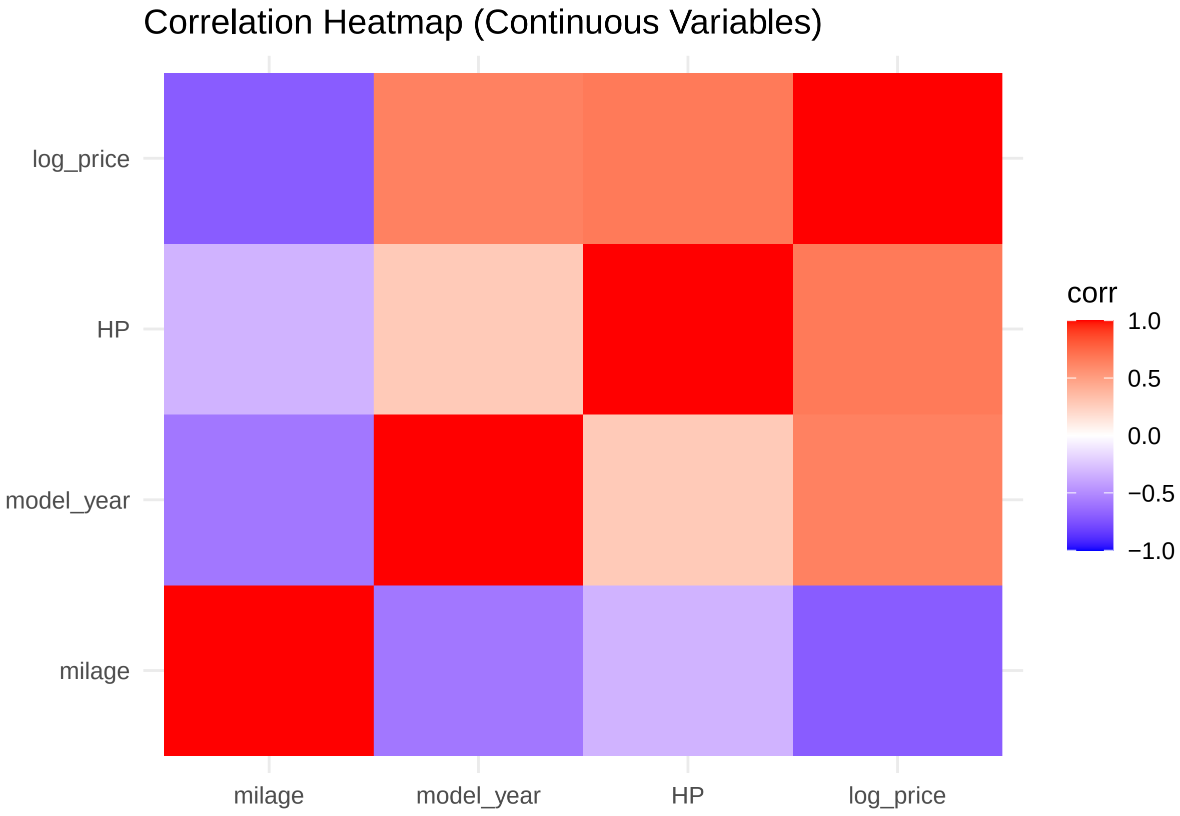

LM
Lasso
RF
0.24
0.26
0.28
MAE
LM
Lasso
RF
0.35
0.40
0.45
RMSE
RF
Lasso
LM
0.75
0.80
0.85
Rsquared

Correlation Analysis and Clustering

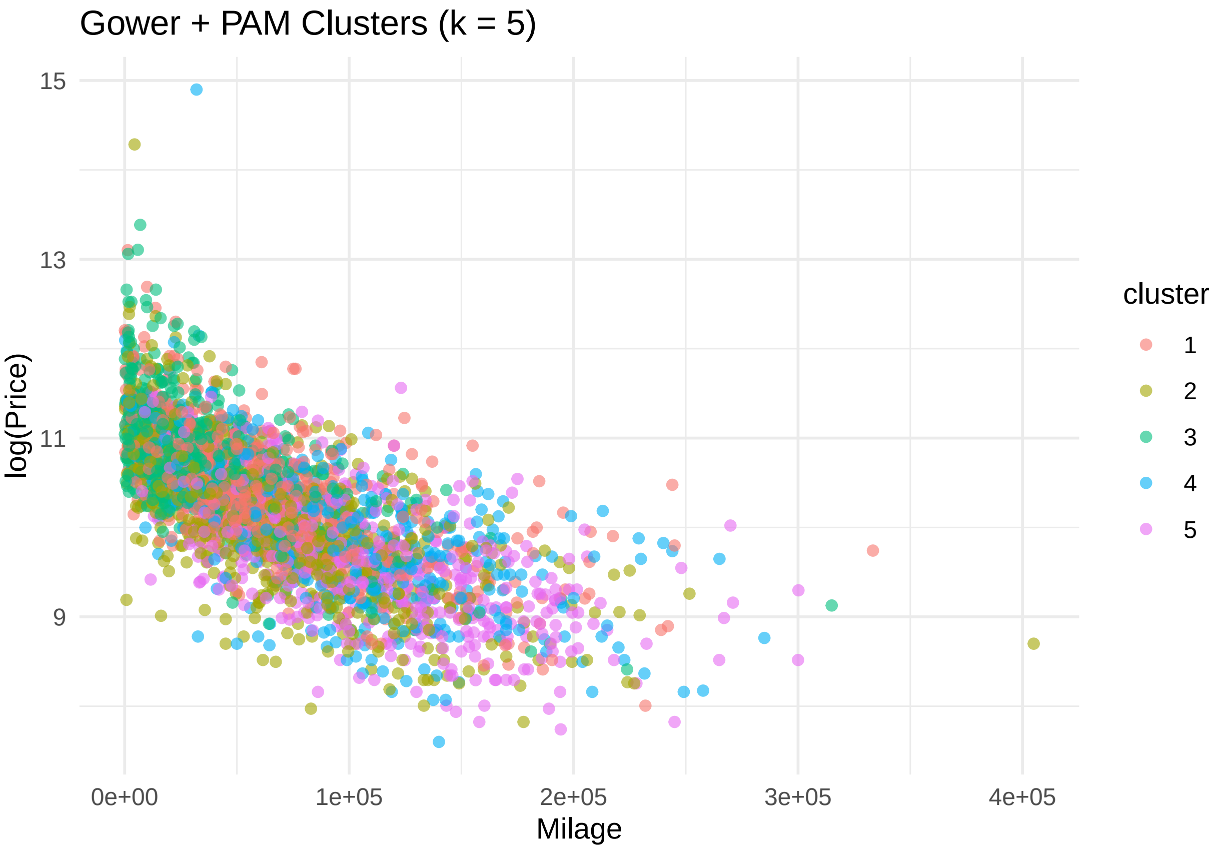
I first examined relationships among the four continuous predictors—mileage, model year, horsepower, and log-price—using a correlation heatmap. The results show clear intuitive patterns: log-price is strongly positively correlated with model year and horsepower, and strongly negatively correlated with mileage. These relationships confirm that newer, more powerful, and lower-mileage cars tend to have higher prices.

To evaluate whether the full data (including categorical features) exhibits natural group structure, I computed Gower distance, which handles mixed data types, and performed PAM (Partitioning Around Medoids) clustering for k = 2 through k = 10. Average silhouette widths across all k values remained very low (0.12–0.14), indicating weak or nonexistent cluster structure in the data. Visualizing the resulting clusters across mileage, horsepower, and model year confirmed that the clusters largely overlap and do not correspond to meaningful interpretable groups.

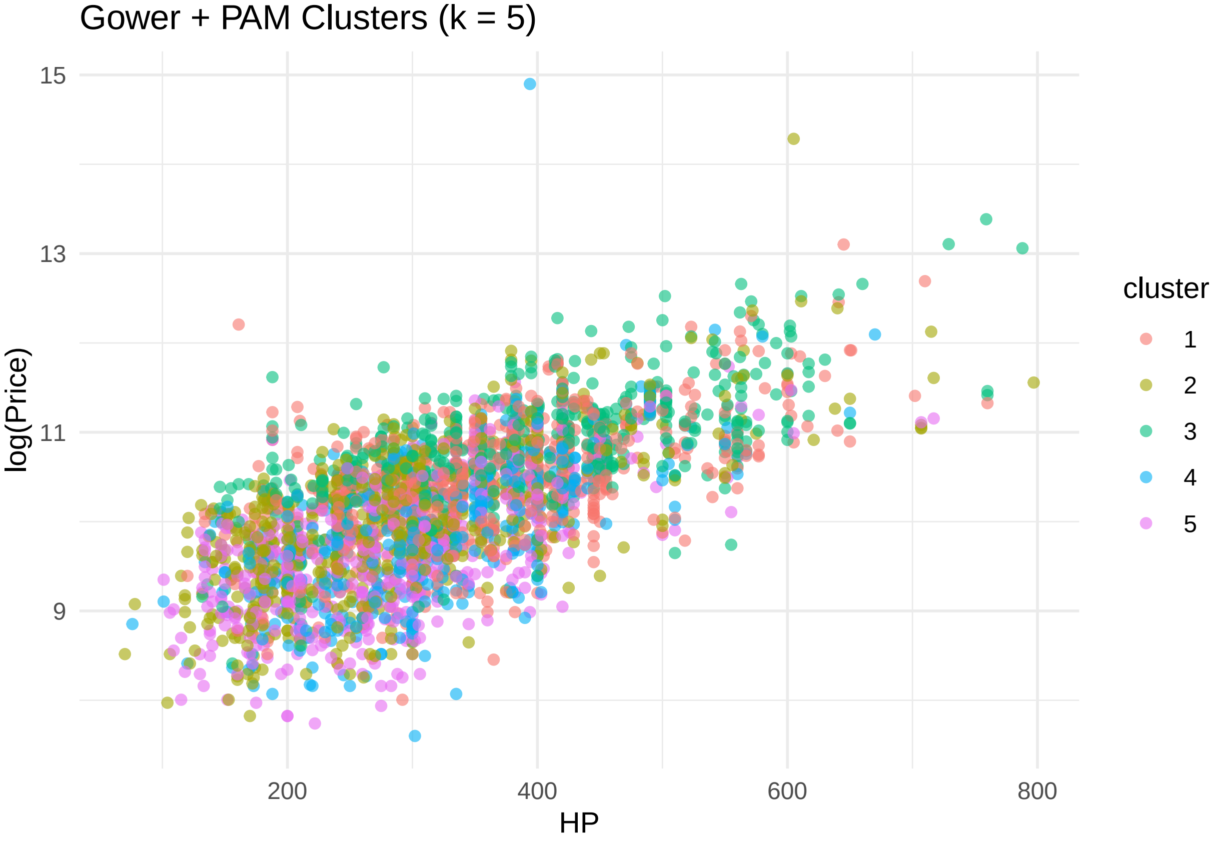
Overall, both the silhouette diagnostics and the cluster visualizations suggest that this data does not naturally cluster, and applying unsupervised clustering provides little additional insight beyond the supervised modeling of price.



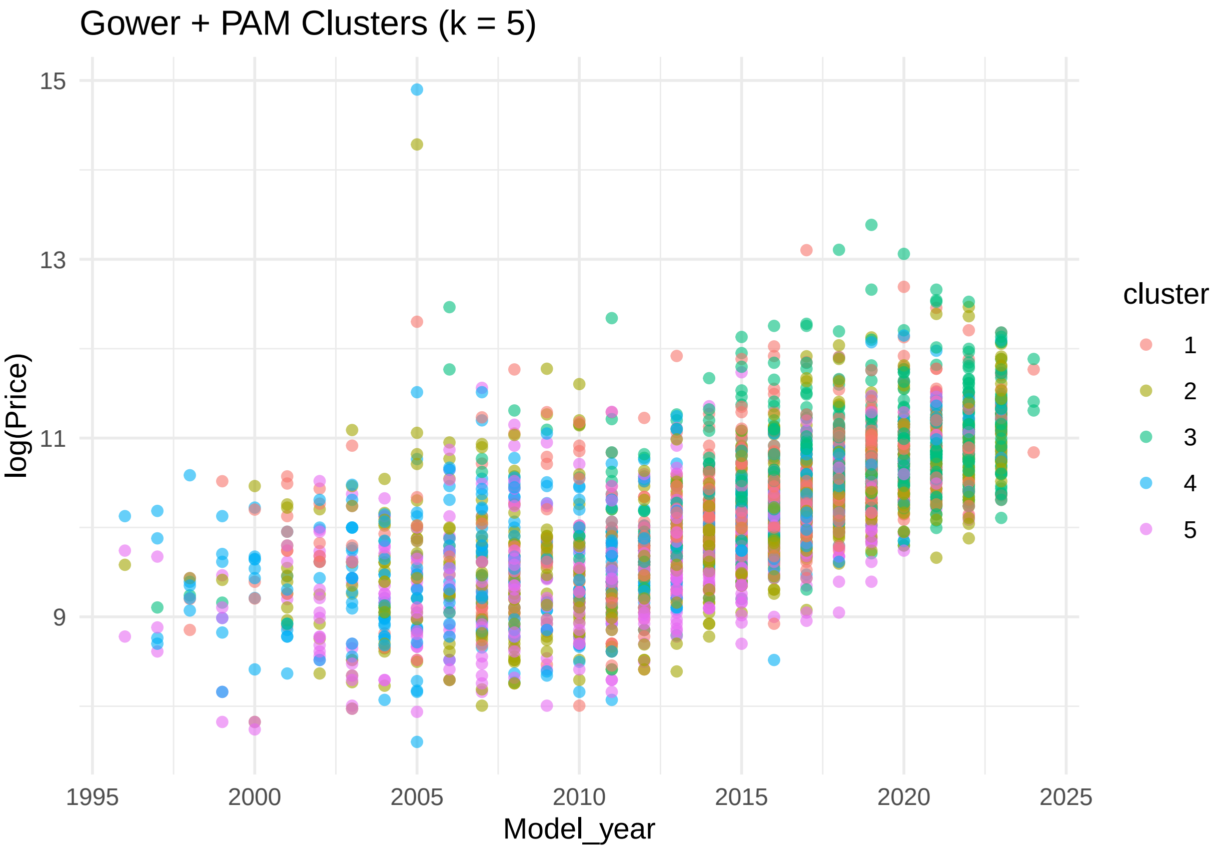
```
##
##   1   2   3   4   5
## 661 671 571 446 546
```



Gower + PAM Clusters (k = 5)



Gower + PAM Clusters (k = 5)


Gower + PAM Clusters (k = 5)
log(Price)
15
13
11
9
Model_year
1995
2000
2005
2010
2015
2020
2025
cluster
1
2
3
4
5