Statistical Methods Project
Results & Conclusions

Your Team

March 2025

Study Design Recap

- Ten pre-registered hypotheses covering **numeric–numeric**, **numeric–categorical**, and **categorical–categorical** relationships.
- $\alpha = 0.05$ throughout.
- Effect-sizes reported: Pearson/Spearman r, Cramér's V, Welch–d ($\approx$ SMD), partial η^2 and pseudo-R^2 where relevant.

Key Correlations (Tests 1–4)

Pair	**r** / ρ	p	
Engine Size $\leftrightarrow$ Price	0.87	$< 2 \cdot 10^{-16}$	[0
Horsepower $\leftrightarrow$ Price	0.81	$< 2 \cdot 10^{-16}$	[0
City MPG $\leftrightarrow$ Highway MPG	0.97	$< 2 \cdot 10^{-16}$	[0
Fuel-Efficiency $\leftrightarrow$ Engine Size (Spearman)	-0.78 (Z = –10.43)	$< 10^{-4}$	

Take-away. Vehicle performance attributes (engine size, horsepower) almost perfectly predict price. Engine downsizing *does* trade off fuel economy.

Group Differences (Tests 5–7)

Contrast	Test	p	Effect	C
Price (gas vs diesel)	Welch t	0.12	$d = 0.36$ (small)	N
Horsepower (std vs turbo)	Welch t	$1.1 \cdot 10^{-4}$	$d = -1.05$	T
Norm. Losses $\sim$ Symboling	One-way ANOVA	$3.3 \cdot 10^{-13}$	$\eta^2 = 0.46$	R
Price $\sim$ Symboling	Kruskal–Wallis	$4.9 \cdot 10^{-11}$	$\eta^2_H \approx 0.35$	S

Take-away. Aspiration and risk rating (symboling) produce the largest mean shifts; fuel type alone does not.

Categorical Associations (Tests 8–9)

Table	χ^2 (MC)	p_{MC}	V	Strength
Fuel Type $\times$ Aspiration	33.0	$1.0 \cdot 10^{-4}$	0.40	**Moderate–Strong**
Body Style $\times$ Drive Wheels	26.6	0.0042	0.26	**Moderate**

Interpretation.

- Diesels overwhelmingly use *standard* aspiration, whereas gas vehicles split between turbo/standard.
- Rear-wheel drive is concentrated in coupes and convertibles; 4-wheel drive almost exclusive to hatchbacks.

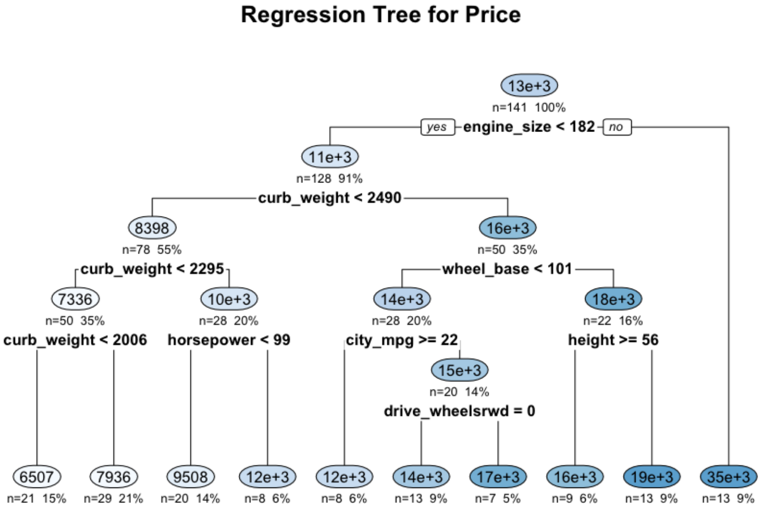
Model Line-up & Metrics

Model	RMSE	MAE	R^2 (test)
Multiple Linear Regression	2 103	1 668	0.920
Ridge (10-fold CV)	2 053	—	0.925
Lasso (10-fold CV)	**1 965**	—	**0.931**
CART (10-fold CV, cp $=$ 0.0008)	2 368	—	0.898

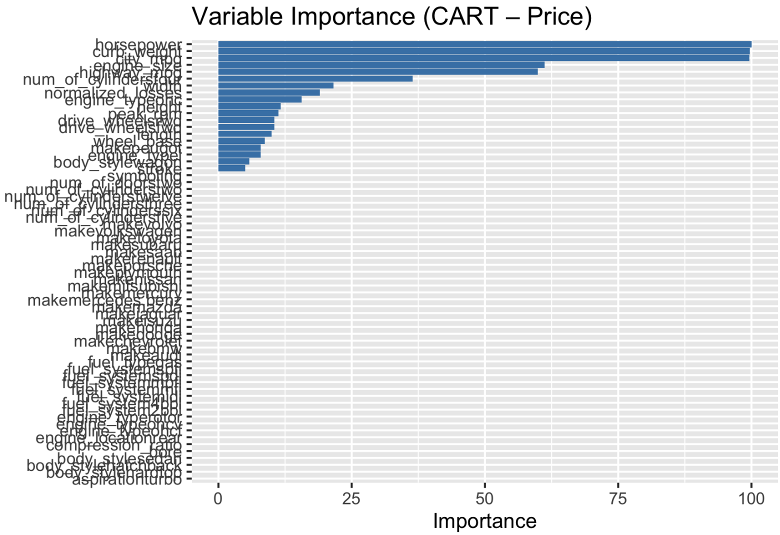
Observations.

- Regularization nudges linear RMSE down by 6–7 %.
- Tree sacrifices a little accuracy for interpretability and non-linearity.
- All models explain $\geq 90\%$ of test-set variance.

Regression Tree (CART)



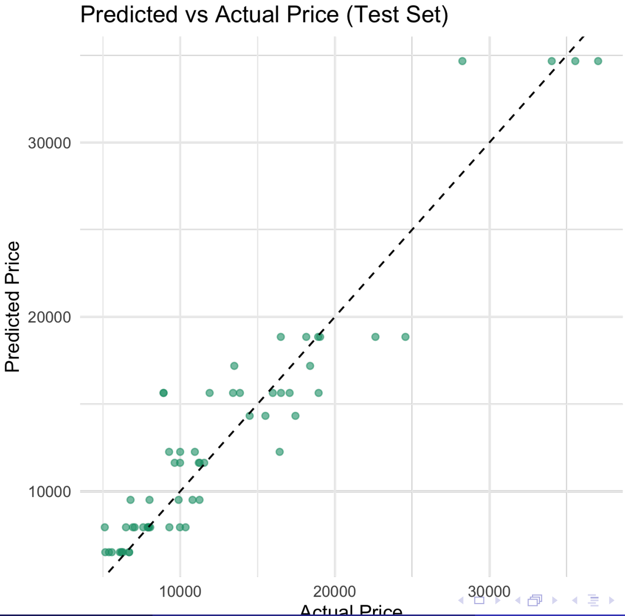
Variable Importance (CART)



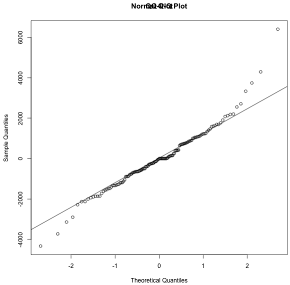
Tree Insights

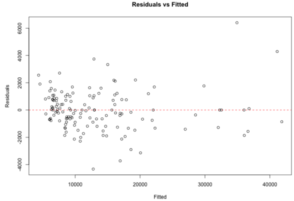
- First split on **engine_size ¡ 182** cc: distinguishes mainstream vs. luxury/performance cluster.
- In the “small-engine” branch, **curb_weight** and **horsepower** refine price bands.
- For large engines, wheel_base and height separate premium sedans from GT-style coupes ($\approx$ \$35 k leaf).

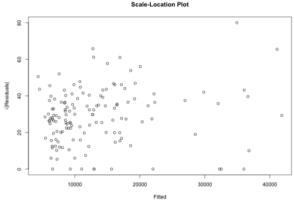
Predicted vs Actual Price (Linear Model)



Regression Diagnostics (Linear)







- **QQ-Plot**: residuals close to normal except mild tails.
- **Residuals–Fitted**: no funnel shape⇒ homoscedasticity acceptable.
- **Scale–Location**: variance fairly constant across fitted.

Key Findings

- **Pricing drivers.** Engine size, horsepower, curb weight, and wheel base dominate both correlation and predictive importance.
- **Risk & cost.** Higher symboling scores *and* higher normalized losses cluster around more expensive, powerful vehicles— significant at $p < 10^{-10}$.
- **Fuel economy trade-off.** Spearman $\rho = -0.78$ confirms large engines penalise average MPG.
- **Best predictive model.** Lasso regression (10-fold CV) achieved the lowest RMSE (1 965) and highest R^2 (0.93). Tree remains most interpretable.

Future Work

- Gradient-boosted trees (XGBoost, LightGBM) could improve RMSE while retaining some interpretability via SHAP.
- Explore non-linear interactions (e.g. engine_size $\times$ fuel_type) with GAMs or polynomial terms.
- Integrate external data – MSRP inflation adjustments or safety-rating scores – to refine the price model.

Thank you!

Questions?