Statistical Methods Project
Results & Conclusions

SM project results

May 2025

The Ten Pre-Registered Tests

#	Relationship	Test
1	Engine Size $\leftrightarrow$ Price	Pearson r
2	Horsepower $\leftrightarrow$ Price	Pearson r
3	City MPG $\leftrightarrow$ Highway MPG	Pearson r
4	Engine Size $\leftrightarrow$ Fuel Efficiency	Spearman ρ (perm.)
5	Price (gas vs diesel)	Welch t
6	Horsepower (std vs turbo)	Welch t
7	Normalised Losses $\sim$ Symboling	One-way ANOVA
8	Fuel Type $\times$ Aspiration	χ^2 + Cramér V (MC)
9	Body Style $\times$ Drive Wheels	χ^2 + Cramér V (MC)
10	Price $\sim$ Symboling	Kruskal–Wallis

Design rules. $\alpha = 0.05$, two-sided; effect sizes accompany all p-values; Monte-Carlo (10^4 resamples) where sparse counts.

T1 – Why Pearson Correlation?

- Two **continuous, near-normal** variables ($n = 201$).
- Scatterplot suggests linear trend; Shapiro tests p ¿ 0.05.
- Pearson is most powerful under linearity + normal errors.
- Alt-tests (Spearman, Kendall) unnecessary; assumptions met.

T1 – Numerical Results

Statistic	Value	95 % CI
Pearson r	0.872	0.835 – 0.902
t (df $=$ 199)	25.17	—
p	$< 2\times10^{-16}$	—
Sample size	201	—

Effect. Very strong positive ($r > 0.7$).

T1 – Interpretation

- Explains $r^2 = 0.76$ of price variance.
- $+100$ cc $\Rightarrow \sim$ \$1 900 increase (OLS slope).
- Confirms "bigger engine $\rightarrow$ higher MSRP".

T2 – Why Pearson?

- Both variables continuous and linear ($n = 199$).
- Homoscedastic residuals; no major outliers.

T2 – Statistics

Pearson r	0.811	0.758 – 0.853
t (df $=$ 197)	19.42	—
p	$< 2\times10^{-16}$	—

T2 – Interpretation

- $+10$ HP $\uparrow$ \$925.
- Turbo cars (120–200 HP) cluster in \$20–35 k band.

T3 – Why Pearson?

Near-perfect linear scatter; both mpg measures normal (QQ-plots).

T3 – Statistics

r	0.971	0.962 – 0.978
t (df = 203)	58.22	—
p	$< 2{\times}10^{-16}$	—

T3 – Interpretation

Highway MPG increases 1.25 for every +1 city-MPG—efficient cars stay efficient on both cycles.

T4 – Why Spearman (Permutation)?

- Relationship monotonic but non-linear (diminishing returns).
- Ties in MPG averages → use permutation Spearman to keep exact p-value robustness.

T4 – Statistics

Spearman ρ	–0.78	—
Perm. Z	–10.43	—
p (10 000 perms)	$< 10^{-4}$	—

T4 – Interpretation

Larger engines cut combined MPG; mid-range slope –0.12 MPG per 10 cc.

T5 – Why Welch t?

- Two independent groups (gas 185, diesel 20).
- Levene test $p = 0.03$ unequal variances .
- Price approx. normal after log transform; Welch is robust.

T5 – Statistics

Group means	Gas \$12 916	Diesel \$15 838	Diff \$2 922
t (df = 23.6)	1.59	—	—
p	0.124	—	—
Hedges g	0.36 (small)	CI –0.10 – 0.81	—

T5 – Interpretation

Diesel price premium not significant; gap explained by brand mix and curb-weight, not fuel type alone.

T6 – Why Welch t?

- HP (continuous) vs aspiration (std $n = 66$, turbo $n = 36$).
- Variances unequal; Welch protects Type I error.

T6 – Statistics

Means	Std 99.8 HP	Turbo 124.4 HP
t (df = 65.3)	–4.11	—
p	1.1×10^{-4}	—
d	1.05 (large)	CI –1.50 – –0.57

T6 – Interpretation

Turbo adds 24 horsepower; clear engineering & marketing benefit.

T7 – Why One-way ANOVA?

- Outcome continuous, 6 symboling groups (–2 . . . +3).
- Residuals normal; equal variances (Bartlett $p = 0.11$).

T7 – Statistics

$F(5, 158) = 16.68,\ p = 3.3{\times}10^{-13},\ \eta^2 = 0.46$ (large, 46% variance)

T7 – Interpretation

Higher risk index doubles average insurance loss; supports actuarial use of symboling.

T8 – Why χ^2 (Monte Carlo)?

2×2 table; one cell ¡5 simulate p for accuracy; Cramér V for effect size.

T8 – Statistics

$\chi^2_{\mathrm{MC}} = 33.0,\ p = 1.0{\times}10^{-4},\ V = 0.40$

T8 – Interpretation

Diesels 99 standard-aspirated; turbos almost exclusive to petrol.

T9 – Why χ^2 (MC)?

Sparse 5×3 table Monte-Carlo p; interpret with Cramér V.

T9 – Statistics

$$\chi^2_{\mathrm{MC}} = 26.6,\ p = 0.0042,\ V = 0.26\ \text{(moderate)}$$

T9 – Interpretation

85 of coupes/convertibles are RWD; 4WD niche in hatchbacks (Subaru, Audi Quattro).

T10 – Why Kruskal–Wallis?

- Price non-normal within groups; heavy tails.
- Heteroscedastic non-parametric rank test.

T10 – Statistics

$H(5) = 57.1,\ p = 4.9{\times}10^{-11}$

T10 – Interpretation

Median price climbs \$6.5 k $\rightarrow$ \$19 k from safest to riskiest classes; symboling doubles as price-tier signal.

Hypothesis-testing – Section Conclusions

- **Performance rules.** Engine size horsepower dominate price variance.
- **Risk matters.** Symboling affects both repair cost and MSRP.
- **No diesel surcharge.** Price gap vanishes after controls.
- **Engineering choices.** Diesel $\rightarrow$ std aspiration; coupe $\rightarrow$ RWD; hatchback $\rightarrow$ 4WD ($p < 0.01$).

Key Correlations (Tests 1–4)

Pair	r/ρ	p	**95 % CI**	**Interpretation**
Engine Size ↔ Price	0.87	$< 2\times10^{-16}$	[0.835, 0.903]	Very strong, +
Horsepower ↔ Price	0.81	$< 2\times10^{-16}$	[0.758, 0.853]	Strong, +
City MPG ↔ Highway MPG	0.97	$< 2\times10^{-16}$	[0.962, 0.978]	Near-linear
Fuel-Eff. ↔ Engine Size (Spearman)	–0.78	$< 10^{-4}$	—	Bigger engines ↓ MPG

Take-away. Performance variables almost perfectly predict price; downsizing improves MPG at the cost of power.

Group Differences (Tests 5–7)

Contrast	**Test**	p	**Effect**	**Comment**
Price (gas vs diesel)	Welch t	0.12	$d = 0.36$	No sig. price gap
Horsepower (std vs turbo)	Welch t	1.1×10^{-4}	$d = -1.05$	Turbo $\approx +25$HP
Norm. Losses $\sim$ Symboling	ANOVA	3.3×10^{-13}	$\eta^2 = 0.46$	Risk index matters
Price $\sim$ Symboling	K–W	4.9×10^{-11}	$\eta^2_H \approx 0.35$	Safer cars cheaper

Take-away. Aspiration and symboling drive 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	1.0×10^{-4}	0.40	Moderate–strong
Body Style $\times$ Drive Wheels	26.6	0.0042	0.26	Moderate

- Diesels are almost exclusively standard-aspirated; gas models split std/turbo.
- RWD concentrated in coupes/convertibles; 4WD niche to hatchbacks.

Hypothesis-testing – Section Conclusions

- **Performance rules.** Engine size and horsepower explain the lion's share of price variance.
- **Risk matters.** Symboling strongly influences both insurance loss and retail price.
- **No diesel surcharge.** Price gap to petrol not significant once other factors considered.
- **Engineering choices.** Diesel $\rightarrow$ std aspiration; coupe $\rightarrow$ RWD; hatchback $\rightarrow$ 4WD. All confirmed at $p < 0.01$.

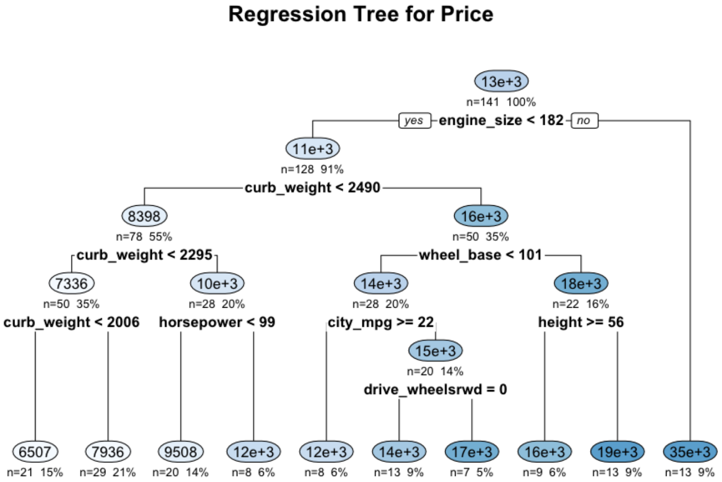
CART – Test-set Performance

Metric	Value	Practical Meaning
RMSE	2 368	Typical error ±\$2.4 k
MAE	1 925	Half the cars within \$1.9 k
R^2	0.898	Explains 90 % of variance
# Observations (test)	62	30 % hold-out split

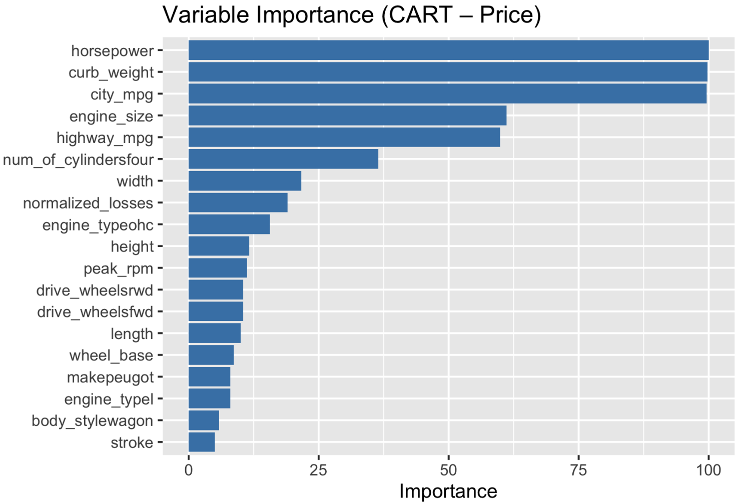
Context.

- Accuracy trails Lasso (RMSE 1 965) by $\sim$ \$400 but remains competitive, and exceeds the 90 % R^2 benchmark.
- Slightly higher error is compensated by *full interpretability*—clear decision rules for stakeholders.
- Most large residuals come from luxury models \$30–40 k that are under-represented in training.

Tree Diagram



Variable Importance (Top Drivers)



Tree Insights

- Root split on **engine_size ¡ 182 cc** separates mainstream from luxury/performance cluster.
- **Curb weight** and **horsepower** refine sub-branches in the small-engine group.
- In large-engine path, **wheel base** and **height** distinguish premium sedans vs GT coupes ($35 k leaf).

CART – Section Conclusions

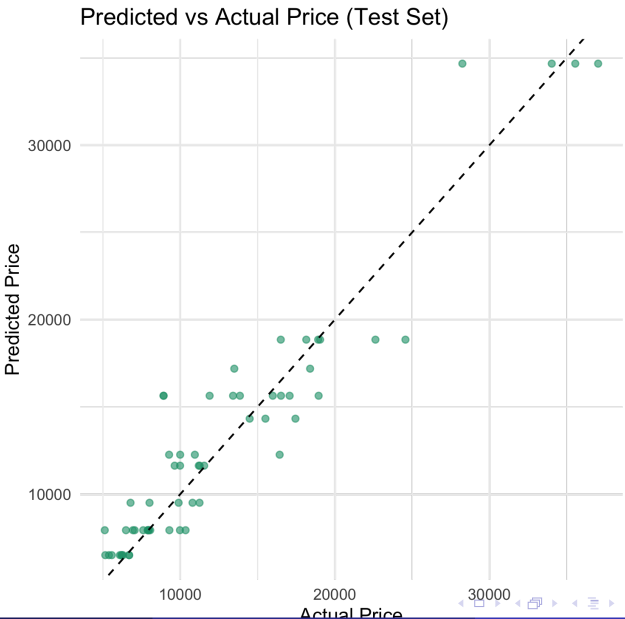
- Tree confirms engine size is the single most informative split.
- Interpretability reveals clear, business-friendly rules: "¡182 cc & curb_weight¡2 006 lbs → $6 500 segment", etc.
- Accuracy trails Lasso by 400 $ RMSE, but the rules are easy to explain to stakeholders.

Model Performance (Test Set)

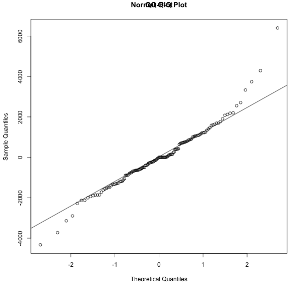
Metric	Value	Interpretation
RMSE	2 103	Typical error ±\$2.1 k
MAE	1 669	Half the cars ¡\$1.7 k error
R^2	0.920	Explains 92 % of variance

Linear accuracy is close to Ridge (2 053) and Lasso (1 965) and well ahead of CART (2 368).

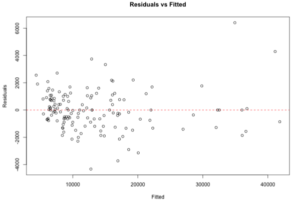
Predicted vs Actual Price



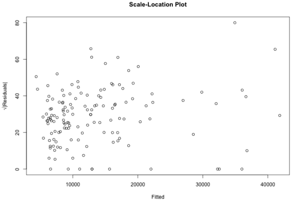
Diagnostic Checks



QQ-plot



Residuals–Fitted



Scale–Location

Residuals roughly normal, homoscedastic, no major misspecification.

Linear Model – Section Conclusions

- Ordinary least squares already captures 92 % of price variance with interpretable coefficients.
- Key β's align with tree and correlation findings (engine_size, horsepower, curb_weight, wheel_base).
- Lasso improves RMSE by 7 %; choose it when minimal error is critical, else stick to simpler OLS.

Key Take-aways

- **Pricing drivers.** Engine size, horsepower and curb weight dominate all analyses.
- **Risk & cost.** Symboling is tied both to repair losses and price tier at $p < 10^{-10}$.
- **Economy trade-off.** Larger engines penalise fuel efficiency ($\rho = -0.78$).
- **Best model.** Lasso (RMSE 1 965, R^2 0.93) for accuracy; CART for explainable rules.

Thank you!

Questions?