

Name: Mazen Fraihat

Course: STA 2260.05-1 – Probability & Statistics for CS Engineers

Instructor: Kevin Bailey

Date: 11/06/2025

Predicting Horsepower from Automobile Characteristics

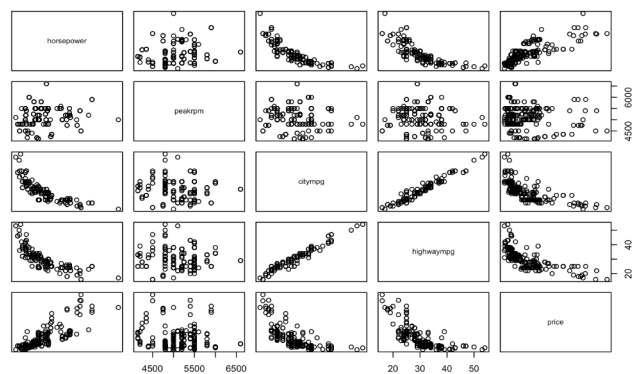
Introduction

The goal of this project is to model and predict automobile horsepower using numerical variables from the *imports85* dataset. The analysis aims to identify which features best explain horsepower variation and to compare single- and multiple-predictor models using the R^2 and Akaike Information Criterion (AIC). These metrics evaluate how much variability is explained and penalize overfitting respectively.

The dataset was obtained from the UCI Machine Learning Repository and contains automobile attributes such as engine size, price, horsepower, and more. The response variable for this study is horsepower, while potential predictors include price, engine size, highway MPG, and peak RPM.

horsepower	peakrpm	citympg	highwaympg	price
111	5000	21	27	13495
111	5000	21	27	16500
154	5000	19	26	16500
102	5500	24	30	13950
115	5500	18	22	17450
110	5500	19	25	15250
110	5500	19	25	17710
110	5500	19	25	18920
140	5500	17	20	23875
101	5800	23	29	16430
101	5800	23	29	16925
121	4250	21	28	20970
121	4250	21	28	21105
121	4250	20	25	24565
182	5400	16	22	30760
182	5400	16	22	41315
182	5400	15	20	36880
48	5100	47	53	5151
70	5400	38	43	6295

Data



Relationship

Model 1 – Single Predictor

Goal

To determine which single variable best predicts horsepower using simple linear regression.

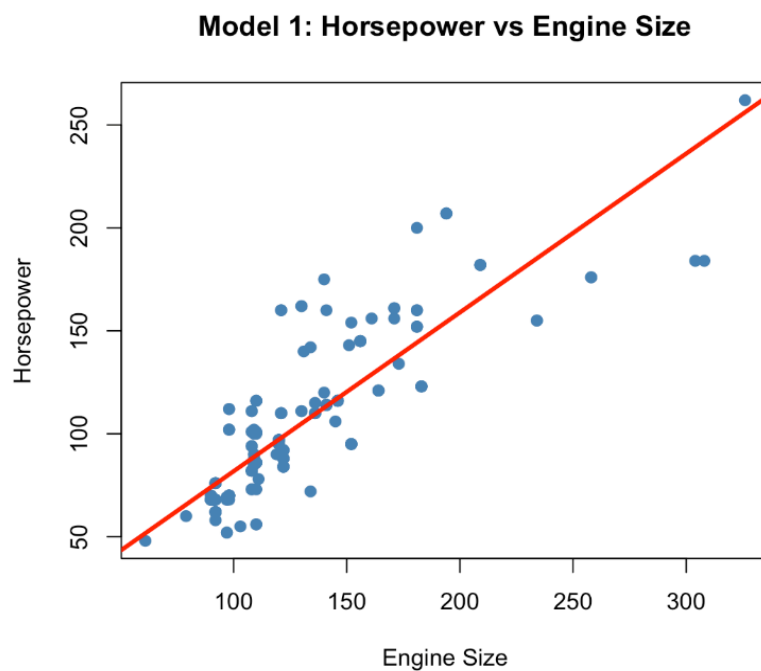
Four predictors were considered: price, highway MPG, peak RPM, and engine size.

From the results in Table 1, engine size was found to be the best single predictor of horsepower.

It had the highest R^2 value (0.7131) and the lowest AIC (1714.432), meaning it explains more variation in horsepower than the other variables.

Model	Predictor	R^2	AIC
m1.price	Price	0.6583	1748.155
m1.highwaympg	Highway MPG	0.6577	1748.501
m1.rpm	Peak RPM	0.005097	1954.416
m1.enginesize	Engine Size	0.7131	1714.432

The scatterplot below also shows a clear positive relationship—cars with larger engines produce more horsepower. The Figure shows the relationship between horsepower and engine size and the red line represents the fitted regression line from the Model 1 equation



Model 2 – Multiple Predictors

We tested two multiple linear regression models to see if adding a second variable improves horsepower prediction.

Model	Predictor	R ²	AIC
m2.enginesize.price	Enginesize + price	0.7296	1704.006
m1.m2.enginesize.bore	Enginesize + bore	0.7216	1709.586
m2.enginesize.bore.price.wheelbase	Engine size + bore + price + wheelbase	0.7694	1675.196

Among all models tested, the engine size + bore + price + wheelbase combination achieved the highest R² (0.7694) and the lowest AIC (1675.196), indicating the best overall fit.

This means that including bore, price, and wheelbase alongside engine size allows the model to explain roughly 77% of the variance in horsepower.

The improvement in both R² and AIC demonstrates that adding these predictors significantly enhances model performance without signs of overfitting.

Comparing Model 1 with Model 2

Compared to Model 1, which used only engine size (R² = 0.7131, AIC = 1714.432), the updated Model 2 shows a substantial improvement with R² = 0.7694 and AIC = 1675.196.

By including bore, price, and wheelbase alongside engine size, the model explains a greater portion of the variation in horsepower and achieves a stronger overall fit.

This enhanced model captures more of the underlying relationships between vehicle characteristics and horsepower while still maintaining simplicity and avoiding overfitting, making it more reliable for future predictions.

Conclusion

In this project, we found that engine size is the strongest single predictor of horsepower, explaining about 71% of the variation in the data.

Expanding the model to include bore, price, and wheelbase further improved performance, increasing R^2 to 0.7694 and reducing the AIC to 1675.196.

Overall, this multiple-predictor model provides a stronger and more accurate fit while remaining straightforward to interpret.

It effectively captures how multiple automobile characteristics influence horsepower, making it one of the most reliable models for predicting engine performance.