Tarunvir Singh

What factors best predict engine size?

Introduction

In this project, I see how different car features relate to engine size using the Automobile dataset from the UCI Machine Learning Repository. The purpose of the analysis is to identify which variables are most effective at predicting engine size and to compare several linear regression models. After reviewing the dataset, I listed a variety of possible numerical predictors such as horsepower, wheelbase, curb weight, fuel mileage, and price. I then saw each predictor using scatter plots and selected the variables that showed the strongest relationships with engine size. The models will be assessed using R^2 and AIC, where higher R^2 and lower AIC indicate better predictive performance. I chose engine size as the response variable because it is a central element of a vehicle's performance and design, and many car characteristics such as horsepower, fuel efficiency, and price. They are strongly affected by the size of the engine. This makes engine size a meaningful outcome for analyzing how different vehicle features relate to one another.

Methods Section

Below is a snippet of the first six rows of the imported Automobile dataset (import85 file).

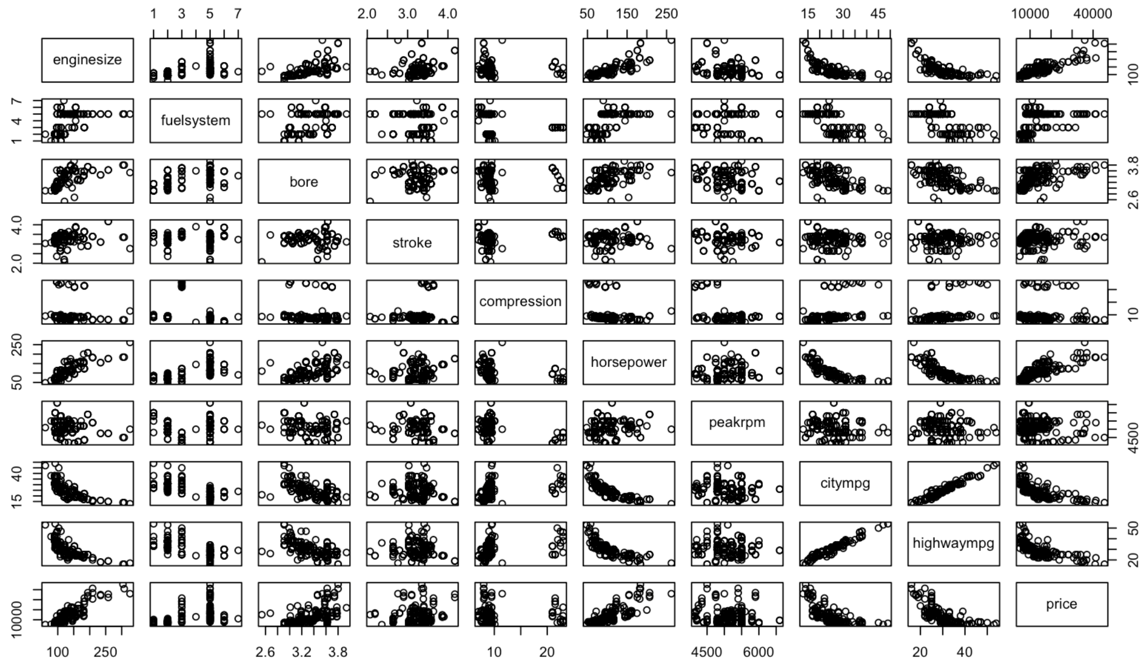
	symboling	make	fuel	aspiration	numdoors	bodystyle	drivewheels	enginelocation	wheelbase	length
1	3	alfa-romero	gas	std	two	convertible	rwd	front	88.6	168.8
2	3	alfa-romero	gas	std	two	convertible	rwd	front	88.6	168.8
3	1	alfa-romero	gas	std	two	hatchback	rwd	front	94.5	171.2
4	2	audi	gas	std	four	sedan	fwd	front	99.8	176.6
5	2	audi	gas	std	four	sedan	4wd	front	99.4	176.6
6	2	audi	gas	std	two	sedan	fwd	front	99.8	177.3

	width	height	curbweight	enginetype	numcylinders	enginesize	fuelsystem	bore	stroke	compression	horsepower
1	64.1	48.8	2548	dohc	four	130	mpfi	3.47	2.68	9.0	111
2	64.1	48.8	2548	dohc	four	130	mpfi	3.47	2.68	9.0	111
3	65.5	52.4	2823	ohcv	six	152	mpfi	2.68	3.47	9.0	154
4	66.2	54.3	2337	ohc	four	109	mpfi	3.19	3.40	10.0	102
5	66.4	54.3	2824	ohc	five	136	mpfi	3.19	3.40	8.0	115
6	66.3	53.1	2507	ohc	five	136	mpfi	3.19	3.40	8.5	110

	peakrpm	citympg	highwaympg	price
1	5000	21	27	13495
2	5000	21	27	16500
3	5000	19	26	16500
4	5500	24	30	13950
5	5500	18	22	17450
6	5500	19	25	15250

I selected engine size as the response variable and examined a wide range of possible predictors, including:

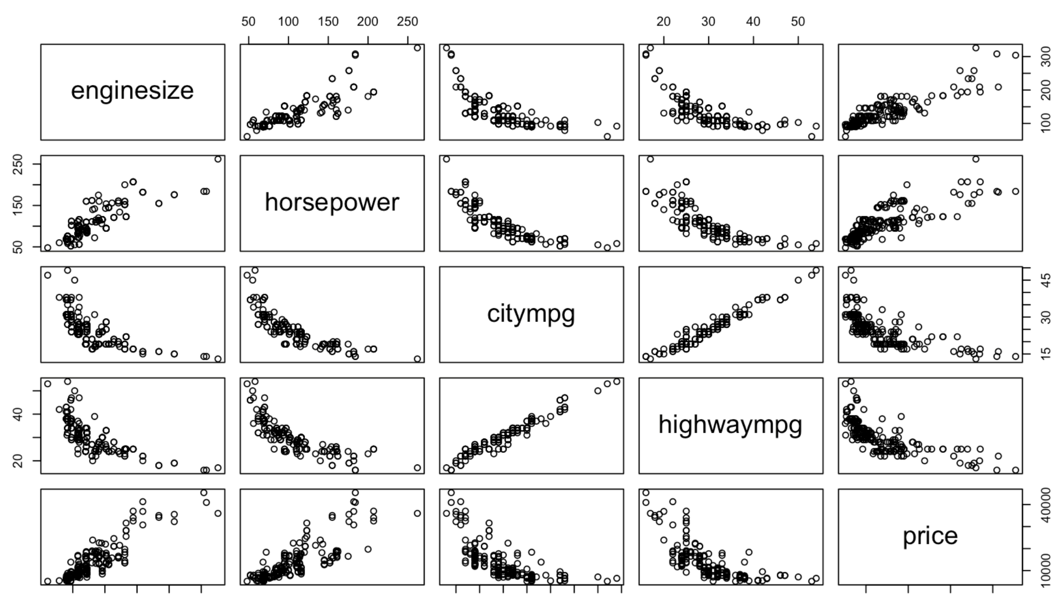
- horsepower
- fuelsystem
- bore
- stroke
- Compression
- peakrpm
- citympg
- highwaympg
- price



To narrow these down, I created scatter plots to visually check the strength, direction, and shape of each relationship with engine size. Scatter plots are useful because they show whether the relationship looks linear, curved, weak, or too scattered to model.

After reviewing every graph, I found that horsepower, city mpg, highway mpg, and price were the strongest candidates. These variables showed clear patterns, either positive linear trends (horsepower, price) or inverse trends (city and highway mpg). In contrast, variables like bore, stroke, compression, peakrpm, wheelbase, and curb weight showed relationships that were either too weak or too dispersed to be useful in a linear model.

Here are the models for them in Scatter Plots



Model 1 - Single Predictor

I created one-predictor regression models for the four strongest candidates: horsepower, city mpg, highway mpg, and price. I compared their performance using Adjusted R^2 and AIC.

Model	**Predictor**	**Adjusted R^2**	**AIC**
m1.hp	Horsepower	0.7131	1749.688
m1.ct	Citympg	0.5106	1852.726
m1.hw	Highwaympg	0.5416	1840.130
m1.pr	Price	0.7888	1690.528

Among the four, price stood out clearly. It had the highest Adjusted R^2 and the lowest AIC, making it the strongest single predictor of engine size.

For Model 1, I used price as the single predictor

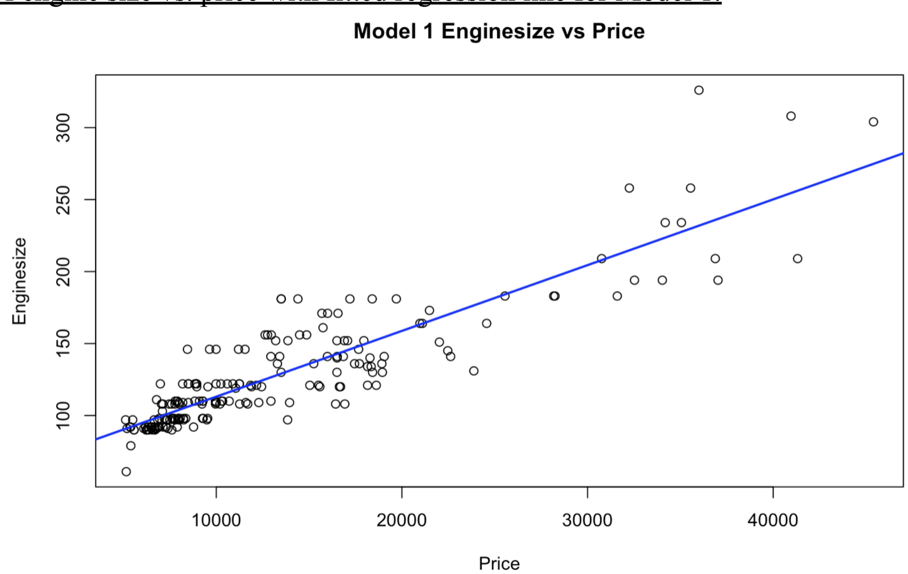
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.742e+01  2.650e+00   25.44   <2e-16 ***
price       4.570e-03  1.705e-04   26.80   <2e-16 ***
```

The R gave me

Intercept	67.42	Predicted engine size when price = 0 (not meaningful, but needed for the line)
price	0.00457	Engine size increases by 0.00457 units for each $1 increase in price

Equation: enginesize = 67.42 + 0.00457(price)

Scatterplot of engine size vs. price with fitted regression line for Model 1.



Model 2 - Multiple Predictors

Next, I built several multi-predictor models to see whether adding variables would improve predictive performance.

Model	**Predictor**	**Adjusted R^2**	**AIC**
m2.pr.hp	Price + horsepower	0.8329	1646.379
m2.pr.hw	Price + Highwaympg	0.808	1673.186
m2.pr.hp.hw.ct	Price + Horsepower + Highwaympg + Citympg	0.84	1639.909
m2.pr.hp.sq	price + price ^2 + horsepower + horsepower^2	0.8417	1637.826

The best-performing model was m2.pr.hp.sq, which included price, horsepower, and their squared terms. This model had the highest Adjusted R^2 (0.8417) and the lowest AIC (1637.826)

Full Equation from R

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.763e+01  9.146e+00   6.301 2.04e-09 ***
price            7.245e-04  8.071e-04   0.898  0.37051
I(price^2)       5.147e-08  1.716e-08   2.998  0.00308 **
horsepower       4.754e-01  1.991e-01   2.388  0.01792 *
I(horsepower^2) -6.302e-05  7.323e-04  -0.086  0.93152
```

Enginesize = 57.63 + 0.0007245 (price) + 0.00000005147 (price^2) + 0.4754 (horsepower) − 0.00006302 (horsepower^2)

Interpretation of coefficients:

- **Horsepower** has a statistically significant positive effect ($p = 0.0179$), meaning vehicles with higher horsepower generally have larger engines.
- **Price²** is also significant ($p = 0.00308$), suggesting a curved relationship where more expensive cars tend to have much larger engines.
- **Horsepower²** was not significant, but allowing curvature still helped model fit.
- **Price** was not significant alone because the curved effect explains more of the relationship.

Model Comparison and Conclusion

Model 1 uses price alone to predict engine size ($R^2 = 0.7888$, AIC = 1690.528), revealing a clear linear relationship. Model 2 incorporates price, horsepower, and their quadratic terms, achieving superior performance ($R^2 = 0.8417$, AIC = 1637.826). The quadratic terms capture nonlinear patterns where engine size increases more rapidly at higher price and horsepower levels, explaining why Model 2 outperforms the simpler linear approach.

This analysis identified price and horsepower as the strongest predictors of engine size. The best model demonstrates that more expensive, higher-horsepower vehicles have disproportionately larger engines due to accelerating growth patterns. While Model 1 provides interpretability through its simplicity, Model 2 offers significantly better predictive accuracy by accounting for the curved relationships in the data. This project highlights the importance of testing various model specifications, including polynomial terms, to identify which factors matter most and how they interact with the response variable.