Data Science Engineering Project with R

CarPrice_Assignment.csv

Manuel Weihmann & Daniel Stepanovic

2025-10-19

Table of contents

1 Macht die Marke einen Unterschied beim Preis?

1.1 1) Setup

```
library(tidyverse)
```

1.2 2) Load and clean the data

```
cars <- readr::read_csv("data/CarPrice_Assignment.csv")
```

2 Make a simple “brand” from the first word of CarName

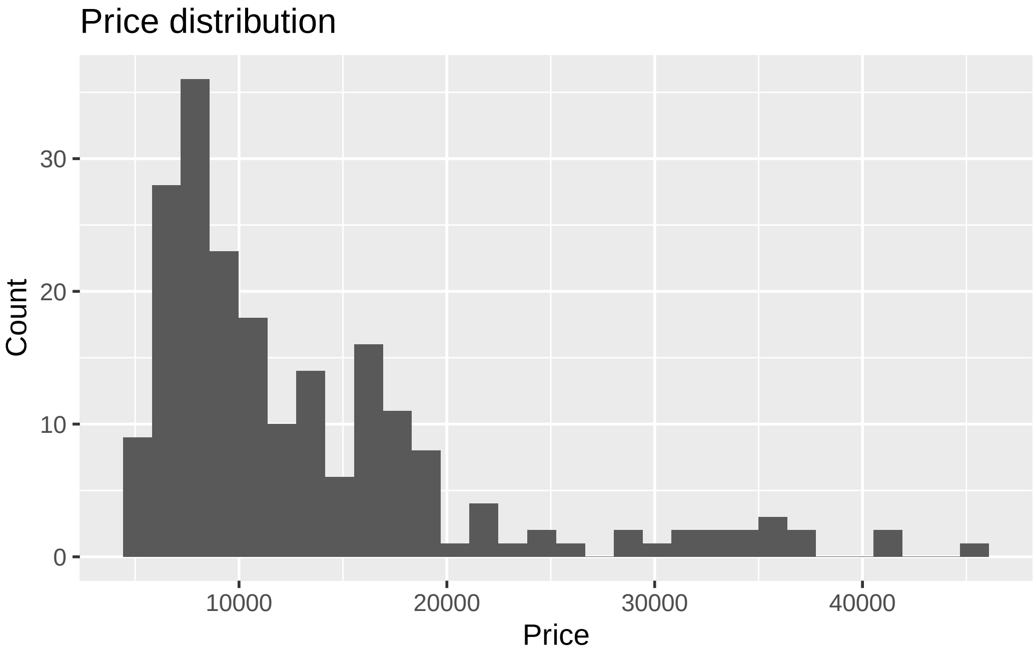
```
cars <- cars |>
  mutate(brand = tolower(sub(" .*", "", CarName)))
```

3 2) Quick look at price

```
# Basic summary
summary(cars$price)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   5118    7788   10295   13277   16503   45400 
```

```
# Price histogram (shape of prices)
ggplot(cars, aes(price)) +
  geom_histogram(bins = 30) +
  labs(title = "Price distribution", x = "Price", y = "Count")
```



Range: from 5,118 to 45,400 → very wide spread.

Middle (median): 10,295 → half the cars cost 10,295 and half 10,295.

Average (mean): 13,277, which is higher than the median → suggests a right-skewed distribution (some expensive cars pull the average up).

Middle 50% (IQR): Q3 − Q1 = 16,503 − 7,788 = 8,715 → typical prices in the middle half span about 8.7k.

Possible high outliers: A common cutoff is Q3 + 1.5×IQR 29,576. Since the max is 45,400, there are likely high-price outliers.

What we can conclude

Prices are skewed to the right with some very expensive cars.

The median (10,295) is a better “typical price” than the mean here.

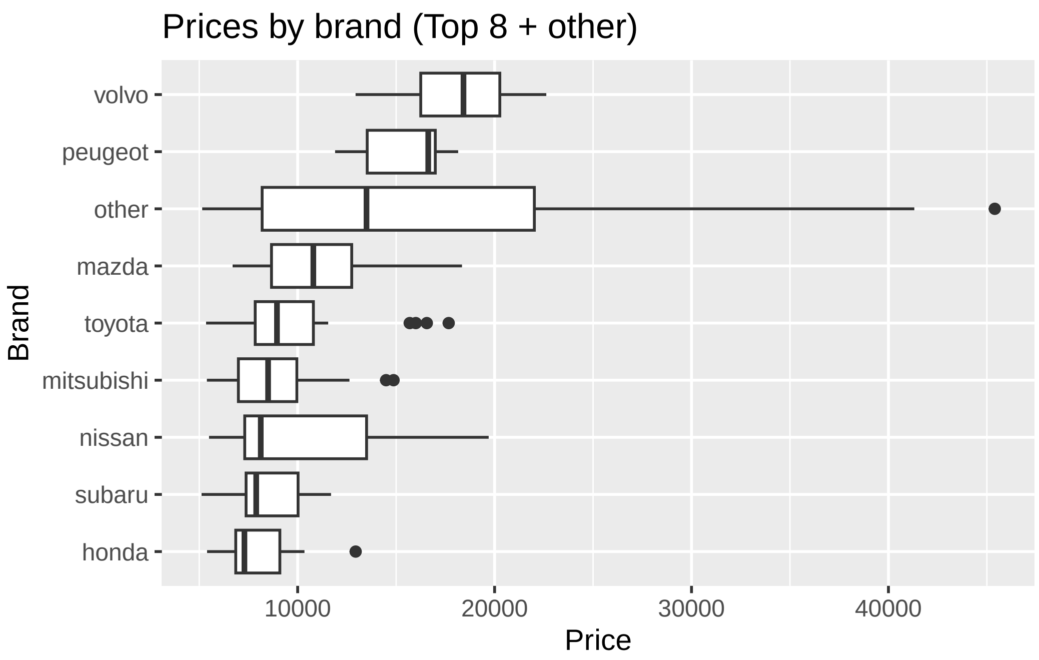
For plots/models, using log(price) will often give clearer patterns.

4 3) Do brands look different?

```
# Keep the 8 most common brands (everything else = "other") so the plot is clean
top_brands <- cars |>
  count(brand, sort = TRUE) |>
  slice_head(n = 8) |>
  pull(brand)

cars_small <- cars |>
  mutate(brand_simple = if_else(brand %in% top_brands, brand, "other"))

# Boxplot: prices by brand
ggplot(cars_small, aes(x = reorder(brand_simple, price, FUN = median), y = price)) +
  geom_boxplot() +
  coord_flip() +
  labs(title = "Prices by brand (Top 8 + other)", x = "Brand", y = "Price")
```



5 4) One-line check: brand effect

```r
# Super simple test: price ~ brand (no other variables)
fit <- lm(price ~ brand_simple, data = cars_small)
summary(fit)
```

```
Call:
lm(formula = price ~ brand_simple, data = cars_small)

Residuals:
     Min       1Q   Median       3Q      Max 
-11710.1  -3423.2   -938.2   1852.4  28538.9 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)              8184.7     2006.0   4.080 6.55e-05 ***
brand_simplemazda        3135.9     2740.7   1.144  0.25393    
brand_simplemitsubishi   1055.1     2836.9   0.372  0.71036    
brand_simplenissan       2231.0     2632.5   0.847  0.39777    
```

```
brand_simpleother        8676.4     2161.0   4.015 8.46e-05 ***
brand_simplepeugeot      7304.4     2963.0   2.465  0.01455 *
brand_simplesubaru        356.6     2895.4   0.123  0.90212
brand_simpletoyota       1512.0     2389.9   0.633  0.52770
brand_simplevolvo        9878.5     2963.0   3.334  0.00102 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7233 on 196 degrees of freedom
Multiple R-squared:  0.2125,    Adjusted R-squared:  0.1804
F-statistic: 6.611 on 8 and 196 DF,  p-value: 1.22e-07
```

Abschlussfazit

Ja, Marke spielt eine Rolle: Einfache ANOVA/Regression mit price ~ brand ist gesamt signifikant.

Größe des Effekts: Marke erklärt ca. 20 % der Preisunterschiede (R^2 0.21) – also relevant, aber nicht alles.

Wer teurer wirkt (vs. Referenz Honda): Volvo, Peugeot und die Gruppe „other“ liegen deutlich höher; die übrigen Top-Marken sind im einfachen Modell nicht klar verschieden.

Aber: Die Unterschiede spiegeln auch Ausstattung wider (Motor, PS, Gewicht). Ohne Kontrolle überschätzen wir „reine“ Markeneffekte.

“Brand matters for price (significant), explaining ~20% of variation; Volvo/Peugeot/‘other’ are pricier than the baseline, but much of the price is still driven by specs—so brand premiums shrink or shift once we control for engine size, horsepower, etc.”