Data Science Engineering Project with R

Dataset: Car Price Prediction

Manuel Weihmann & Daniel Stepanovic

2025-10-19

Table of contents

1 Introduction

Introduction

This report looks at car prices and asks: do brands differ in price?

- Data handling: load the CSV and create a clean brand column
- Descriptive stats: quick numbers for price (min, median, mean, max)
- Visuals: a histogram of prices and boxplots by brand
- Simple model: test if price differs by brand
- Takeaways: what matters, what we found, and limits of the data

1.1 Libraries

```
library(tidyverse)
```

2 Data

2.1 Data source

We use the Car Price Prediction dataset (CarPrice_Assignment.csv). It contains 205 cars and 26 variables (price, brand/model, technical specs).

2.2 Data import

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

- 10 character variables
- 16 numeric variables

2.3 Data dictionary

```
dict <- data.frame(
  Variable = names(cars),
  Type     = sapply(cars, function(x) paste(class(x), collapse = ", "))
)

knitr::kable(dict)
```

	Variable	Type
car_ID	car_ID	numeric
symboling	symboling	numeric
CarName	CarName	character
fueltype	fueltype	character
aspiration	aspiration	character
doornumber	doornumber	character
carbody	carbody	character
drivewheel	drivewheel	character
enginelocation	enginelocation	character
wheelbase	wheelbase	numeric
carlength	carlength	numeric
carwidth	carwidth	numeric
carheight	carheight	numeric
curbweight	curbweight	numeric
enginetype	enginetype	character
cylindernumber	cylindernumber	character
enginesize	enginesize	numeric
fuelsystem	fuelsystem	character
boreratio	boreratio	numeric
stroke	stroke	numeric
compressionratio	compressionratio	numeric
horsepower	horsepower	numeric
peakrpm	peakrpm	numeric
citympg	citympg	numeric
highwaympg	highwaympg	numeric
price	price	numeric

3 Summary statistic tables

3.1 Numeric variables: mean, sd, min, Q1, median, Q3, max

```
nums <- names(cars)[sapply(cars, is.numeric)]

num_summary <- t(sapply(cars[nums], function(x) c(
  n      = sum(!is.na(x)),
  mean   = mean(x, na.rm = TRUE),
  sd     = sd(x, na.rm = TRUE),
  min    = min(x, na.rm = TRUE),
  q1     = as.numeric(quantile(x, 0.25, na.rm = TRUE)),
  median = median(x, na.rm = TRUE),
  q3     = as.numeric(quantile(x, 0.75, na.rm = TRUE)),
  max    = max(x, na.rm = TRUE)
)))
num_summary <- round(as.data.frame(num_summary), 2)
num_summary
```

```
# A tibble: 16 x 8
       n     mean       sd     min      q1   median       q3      max
   <dbl>    <dbl>   <dbl>   <dbl>   <dbl>    <dbl>    <dbl>    <dbl>
 1   205   103      59.3     1      52      103      154      205
 2   205     0.83    1.25   -2       0        1        2        3
 3   205    98.8     6.02   86.6    94.5     97      102.     121.
 4   205   174.     12.3   141.    166.     173.     183.     208.
 5   205    65.9     2.15   60.3    64.1     65.5     66.9     72.3
 6   205    53.7     2.44   47.8    52       54.1     55.5     59.8
 7   205  2556.    521.   1488    2145     2414     2935     4066
 8   205   127.     41.6    61      97      120      141      326
 9   205     3.33    0.27    2.54    3.15     3.31     3.58     3.94
10   205     3.26    0.31    2.07    3.11     3.29     3.41     4.17
11   205    10.1     3.97    7       8.6      9        9.4     23
12   205   104.     39.5    48      70       95      116      288
13   205  5125.    477.   4150    4800     5200     5500     6600
14   205    25.2     6.54   13      19       24       30       49
15   205    30.8     6.89   16      25       30       34       54
16   205 13277.   7989.   5118    7788    10295    16503    45400
```

The numeric table shows the main summary stats (mean, sd, quartiles) for each numeric variable (e.g., price, enginesize, horsepower).

3.2 Categorical variables: number of levels + top 5 most common

```
is_cat <- function(x) is.factor(x) || is.character(x)
cats <- names(cars)[sapply(cars, is_cat)]

cat_summary <- do.call(rbind, lapply(cats, function(v){
  tab <- sort(table(cars[[v]]), decreasing = TRUE)
  top <- head(paste(names(tab), tab, sep=": "), 5)
  data.frame(
    variable = v,
    n_levels = length(tab),
    top_levels = paste(top, collapse = " | "),
    row.names = NULL
  )
}))
cat_summary
```

```
# A tibble: 10 x 3
   variable       n_levels top_levels
   <chr>             <int> <chr>
 1 CarName             147 peugeot 504: 6 | toyota corolla: 6 | toyota corona: ~
 2 fueltype              2 gas: 185 | diesel: 20
 3 aspiration            2 std: 168 | turbo: 37
 4 doornumber            2 four: 115 | two: 90
 5 carbody               5 sedan: 96 | hatchback: 70 | wagon: 25 | hardtop: 8 |~
 6 drivewheel            3 fwd: 120 | rwd: 76 | 4wd: 9
 7 enginelocation        2 front: 202 | rear: 3
 8 enginetype            7 ohc: 148 | ohcf: 15 | ohcv: 13 | dohc: 12 | l: 12
 9 cylindernumber        7 four: 159 | six: 24 | five: 11 | eight: 5 | two: 4
10 fuelsystem            8 mpfi: 94 | 2bbl: 66 | idi: 20 | 1bbl: 11 | spdi: 9
```

The categorical table lists how many levels each label variable has and the top categories with counts.

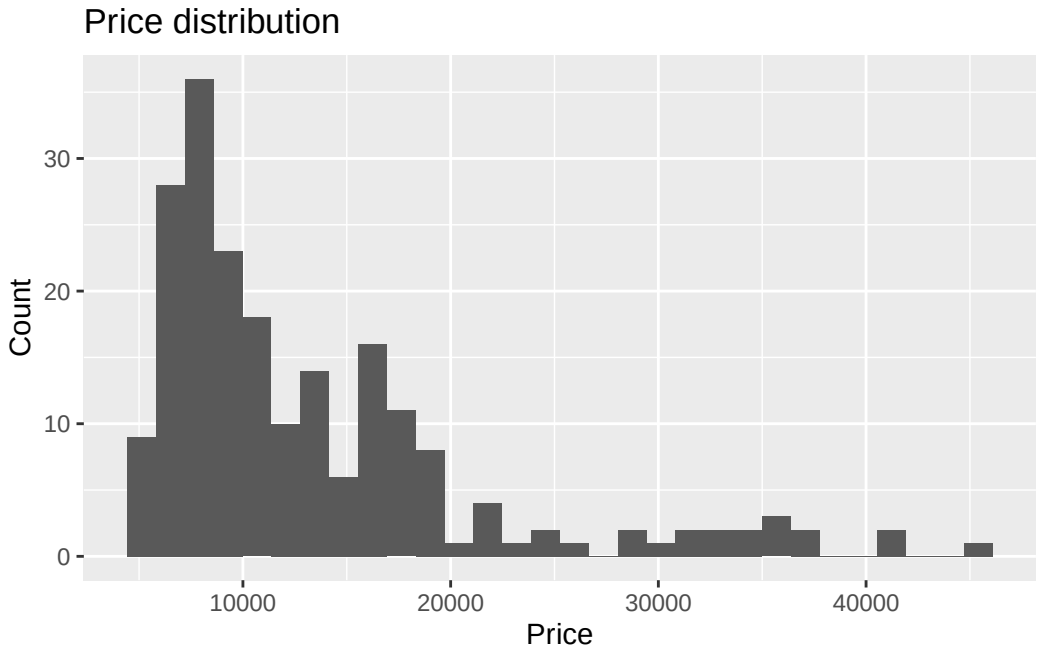
4 Data visualisations

Make a simple "brand" from the first word of CarName

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

4.1 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.

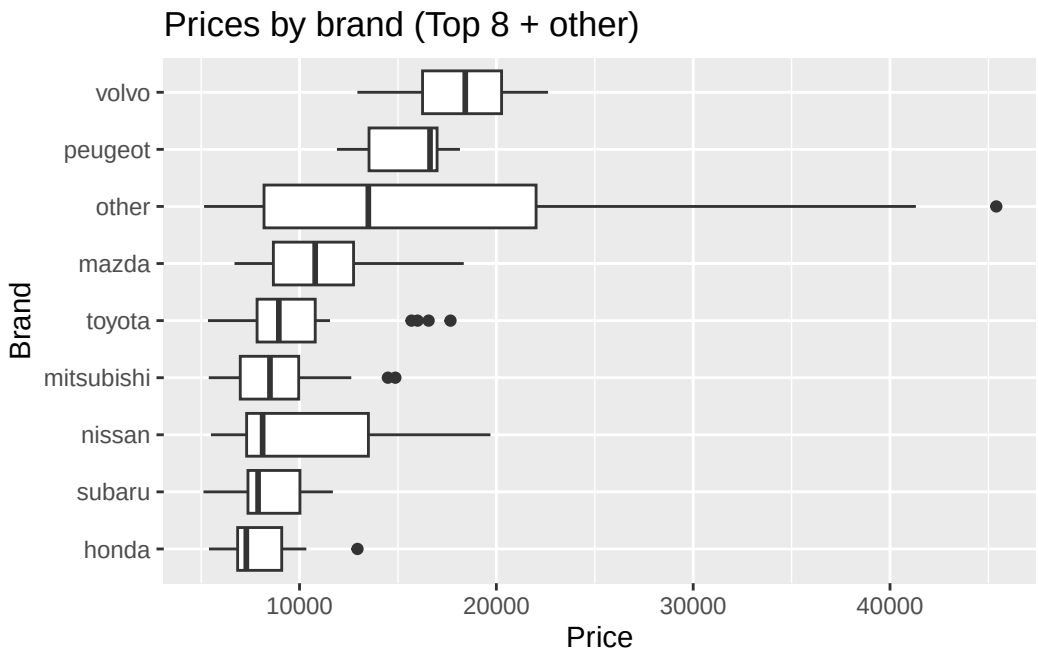
For plots/models, using log(price) will often give clearer patterns.

4.2 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")
```



4.3 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
```

5 Summary

- Brand matters: A simple ANOVA/linear model (price ~ brand) is significant.
- Effect size: Brand explains about 20% of price differences (R^2 0.21)—important, but not the whole story.
- Who looks pricier (vs. Honda baseline): Volvo, Peugeot, and "other" are notably higher; the other top brands are not clearly different in this simple model.
- Caveat: These gaps also reflect specifications (engine size, horsepower, weight). Without controls, we may overstate pure brand effects.

Brand matters for price (significant) and explains ~20% of the variation; Volvo/Peugeot/"other" are pricier than the baseline, but much of the price is driven by specs—brand premiums change once we control for engine size, horsepower, and weight.