MTCARS 2.0

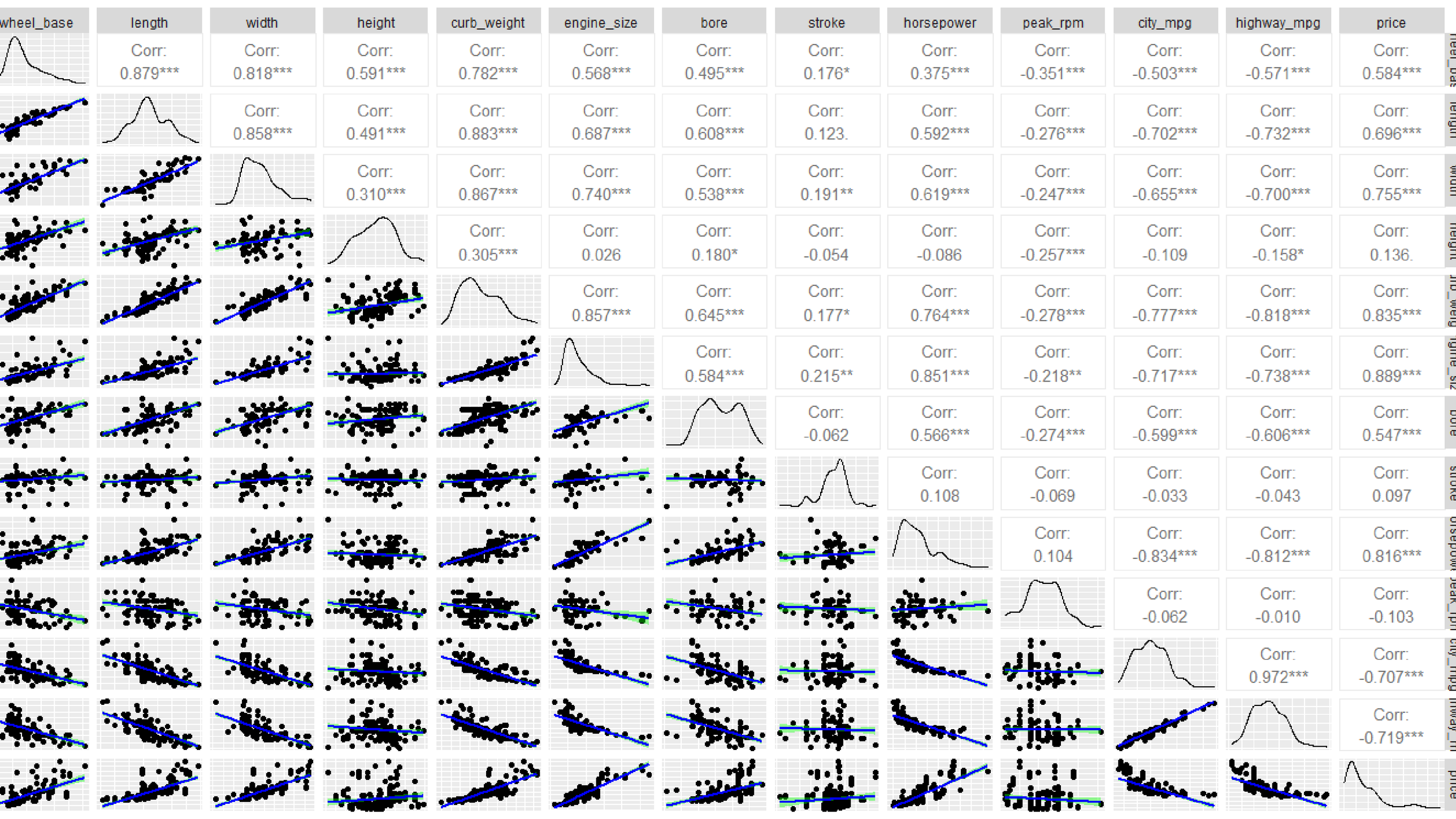
GABRIEL, MARLA,
MOISÉS

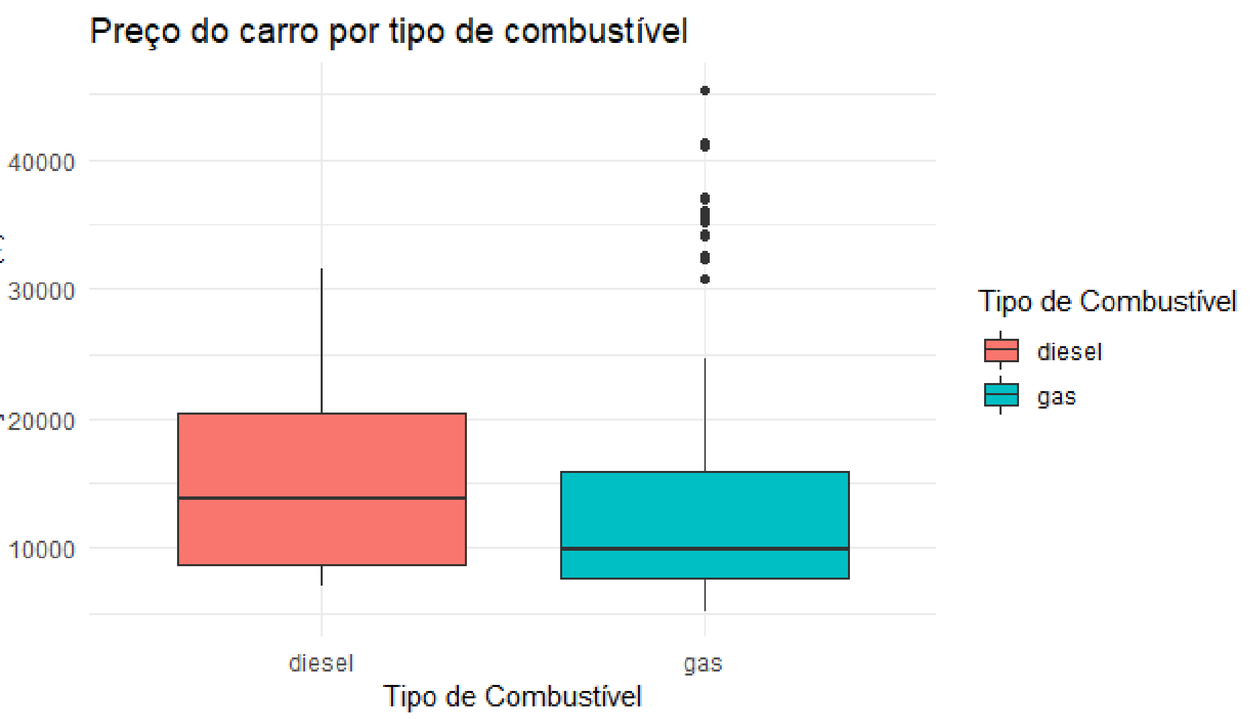
Tratamento de dados

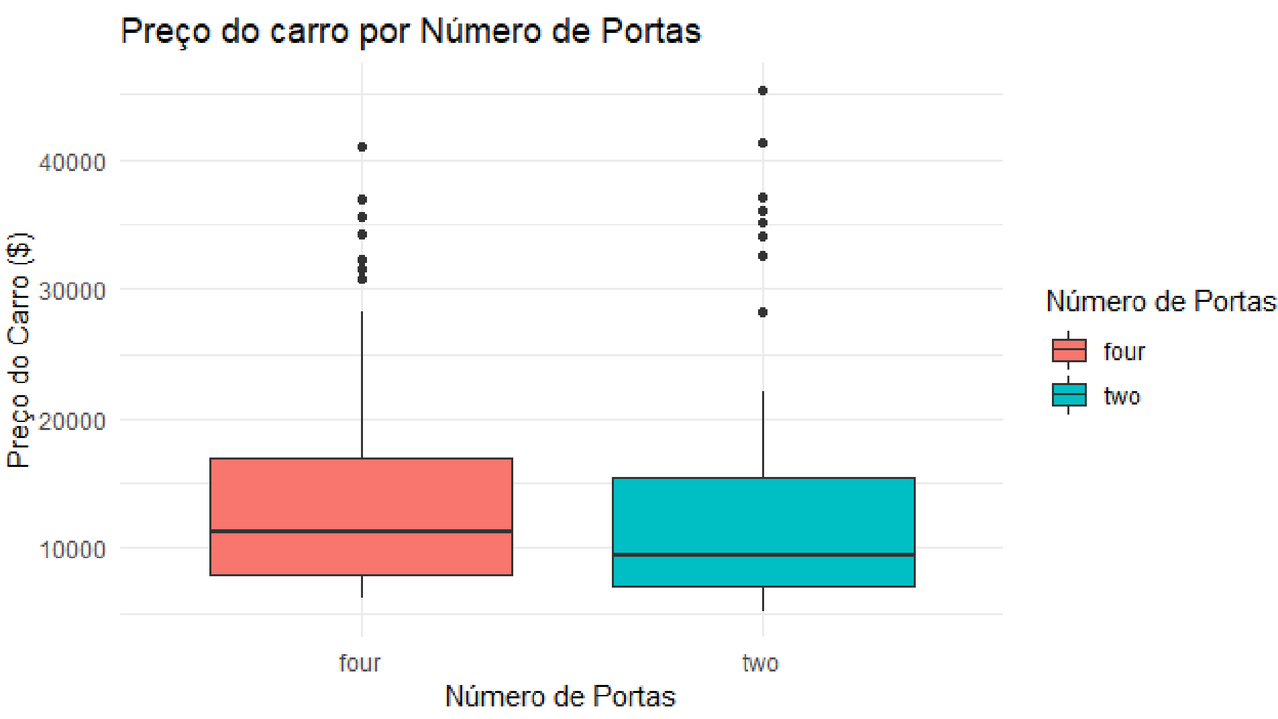
Resumo geral dos dados;
Variáveis sem nome;
Tipos/classes das variáveis;
Valores faltantes;
Categorização: Comṕress_ratio, num_cylinders.

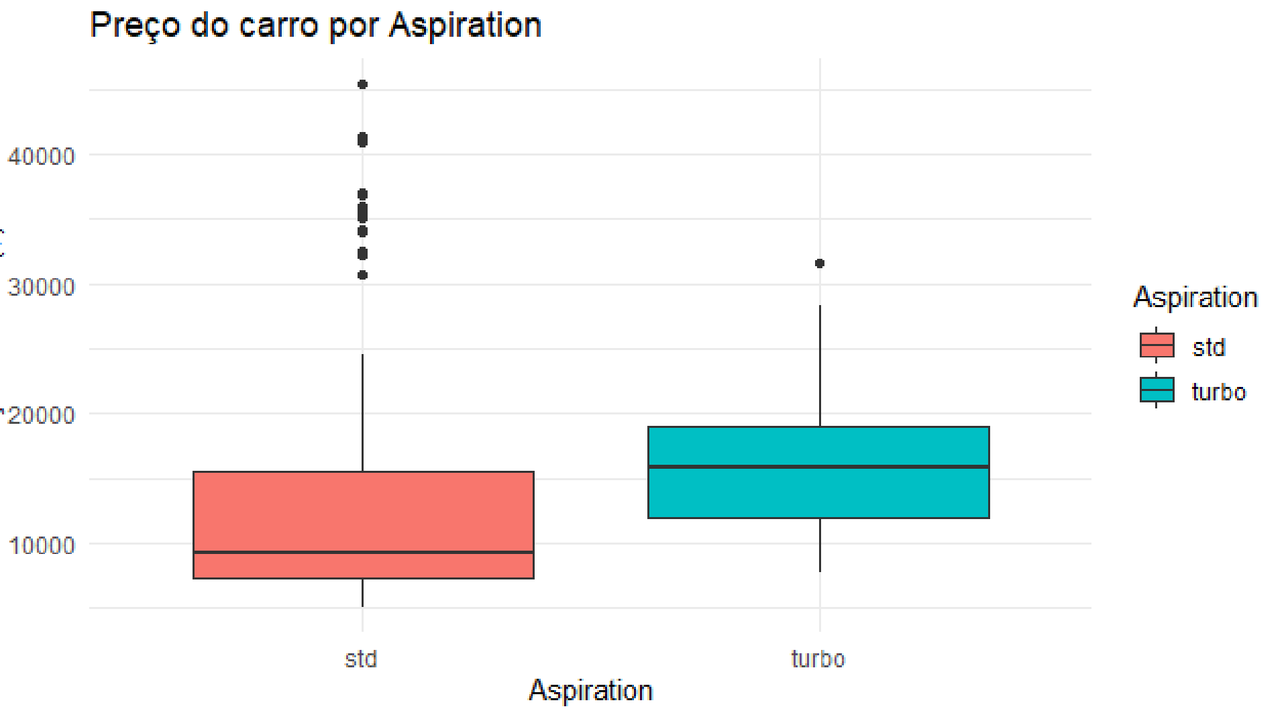
Exploratória

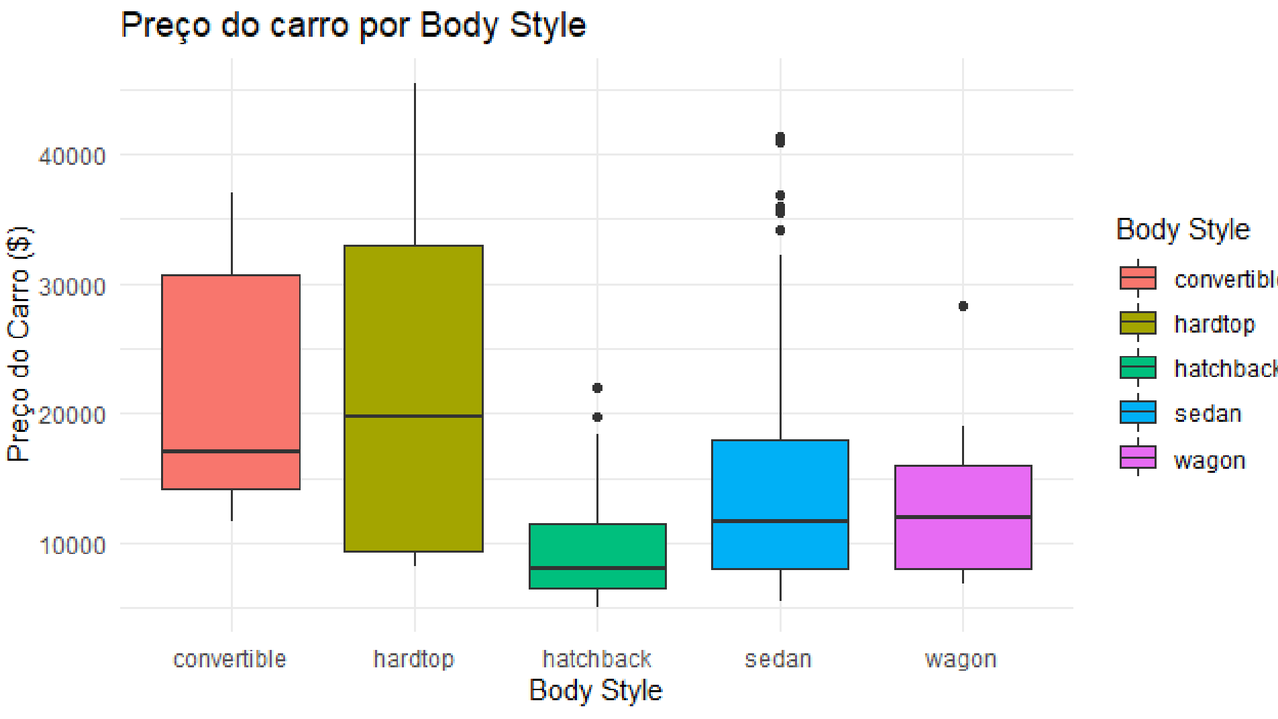
- Cálculo de descritivas;
- Gráficos bi-variados com o preço;
- Verificação de correlação entre variáveis explicativas;
- Testes de hipótese para correlação;
- Correlação de Crammer para variáveis categóricas.

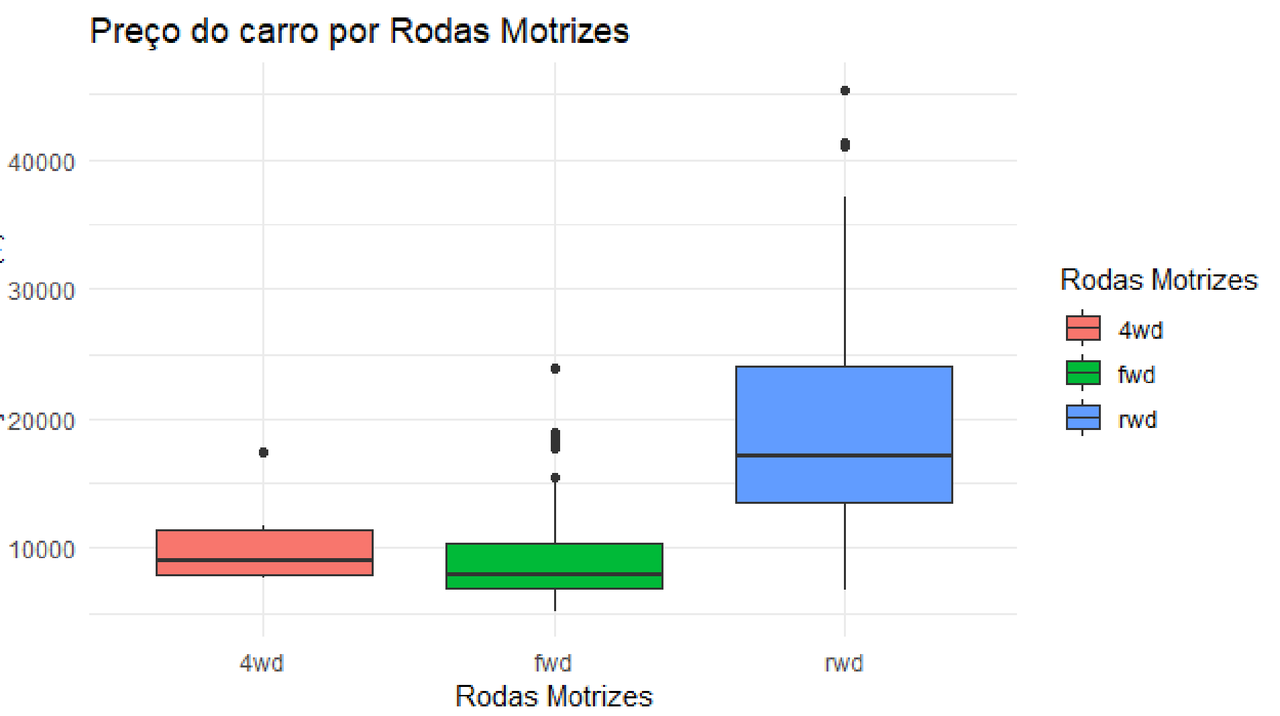

wheel_base
length
width
height
curb_weight
engine_size
bore
stroke
horsepower
peak_rpm
city_mpg
highway_mpg
price
Corr: 0.879***
Corr: 0.818***
Corr: 0.591***
Corr: 0.782***
Corr: 0.568***
Corr: 0.495***
Corr: 0.176*
Corr: 0.375***
Corr: -0.351***
Corr: -0.503***
Corr: -0.571***
Corr: 0.584***
Corr: 0.858***
Corr: 0.491***
Corr: 0.883***
Corr: 0.687***
Corr: 0.608***
Corr: 0.123.
Corr: 0.592***
Corr: -0.276***
Corr: -0.702***
Corr: -0.732***
Corr: 0.696***
Corr: 0.310***
Corr: 0.867***
Corr: 0.740***
Corr: 0.538***
Corr: 0.191**
Corr: 0.619***
Corr: -0.247***
Corr: -0.655***
Corr: -0.700***
Corr: 0.755***
Corr: 0.305***
Corr: 0.026
Corr: 0.180*
Corr: -0.054
Corr: -0.086
Corr: -0.257***
Corr: -0.109
Corr: -0.158*
Corr: 0.136.
Corr: 0.857***
Corr: 0.645***
Corr: 0.177*
Corr: 0.764***
Corr: -0.278***
Corr: -0.777***
Corr: -0.818***
Corr: 0.835***
Corr: 0.584***
Corr: 0.215**
Corr: 0.851***
Corr: -0.218**
Corr: -0.717***
Corr: -0.738***
Corr: 0.889***
Corr: -0.062
Corr: 0.566***
Corr: -0.274***
Corr: -0.599***
Corr: -0.606***
Corr: 0.547***
Corr: 0.108
Corr: -0.069
Corr: -0.033
Corr: -0.043
Corr: 0.097
Corr: 0.104
Corr: -0.834***
Corr: -0.812***
Corr: 0.816***
Corr: -0.062
Corr: -0.010
Corr: -0.103
Corr: 0.972***
Corr: -0.707***
Corr: -0.719***

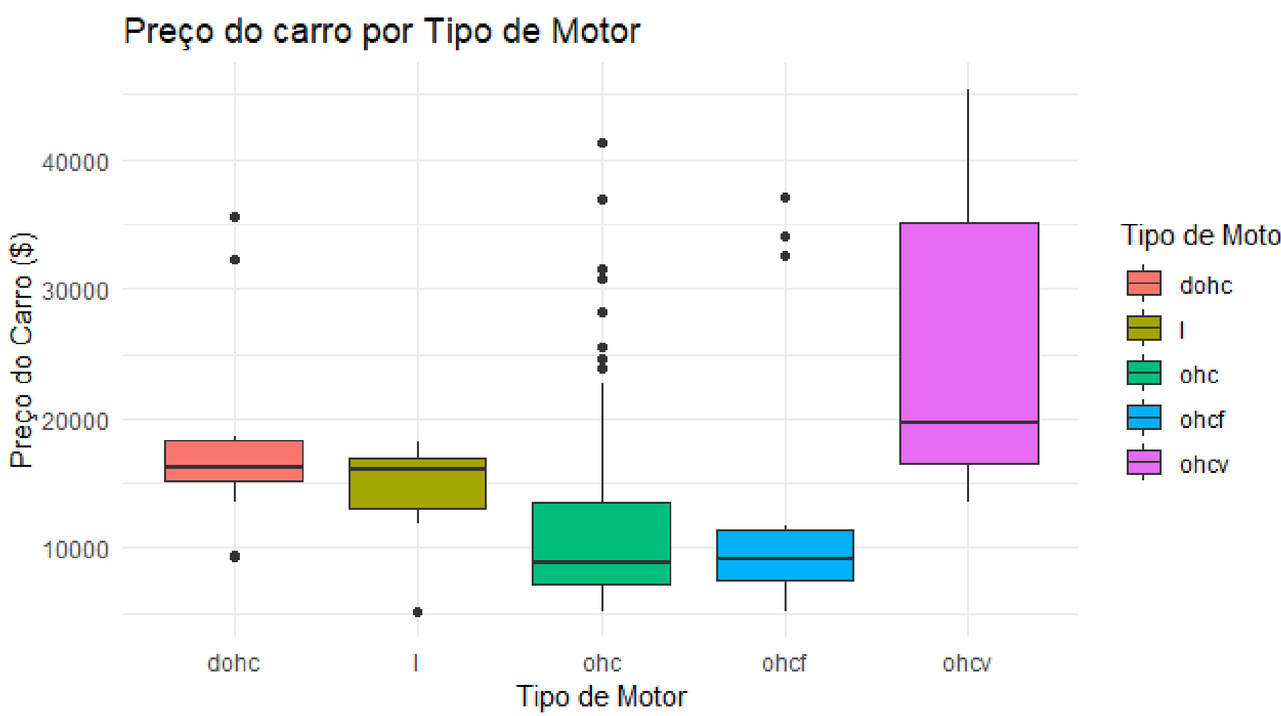

Preço do carro por tipo de combustível
40000
30000
20000
10000
diesel
gas
Tipo de Combustível
Tipo de Combustível
diesel
gas

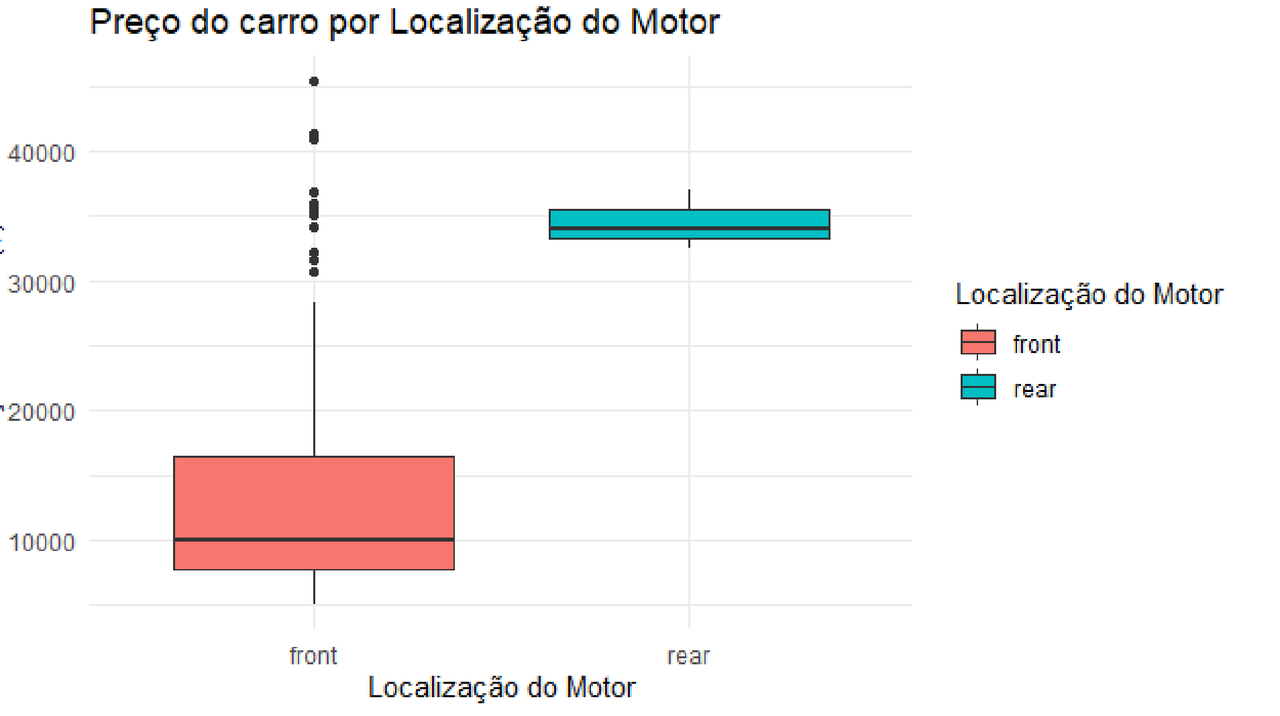

Preço do carro por Número de Portas
Preço do Carro ($)
40000
30000
20000
10000
four
two
Número de Portas
Número de Portas
four
two

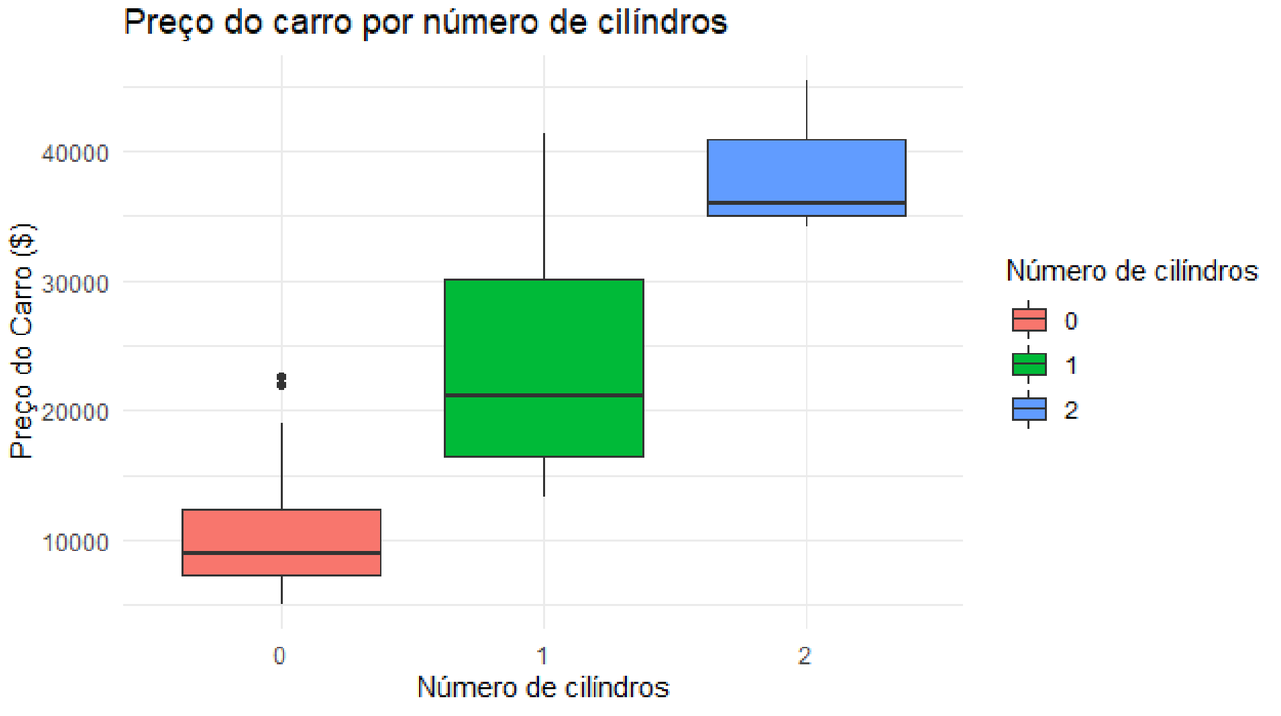

Preço do carro por Aspiration
40000
30000
20000
10000
std
turbo
Aspiration
Aspiration
std
turbo

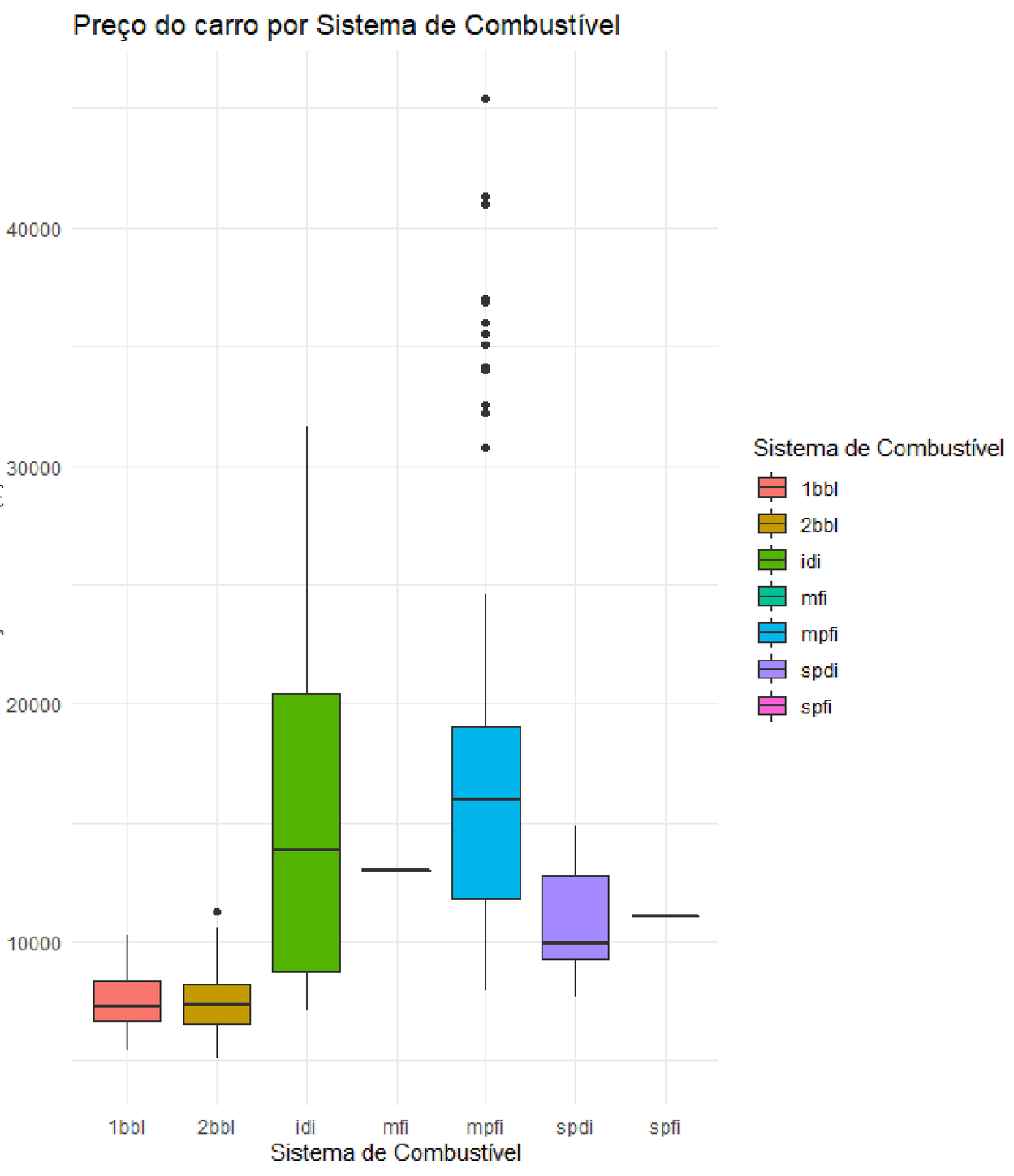

Preço do carro por Body Style
Preço do Carro ($)
40000
30000
20000
10000
convertible
hardtop
hatchback
sedan
wagon
Body Style
Body Style
convertible
hardtop
hatchback
sedan
wagon

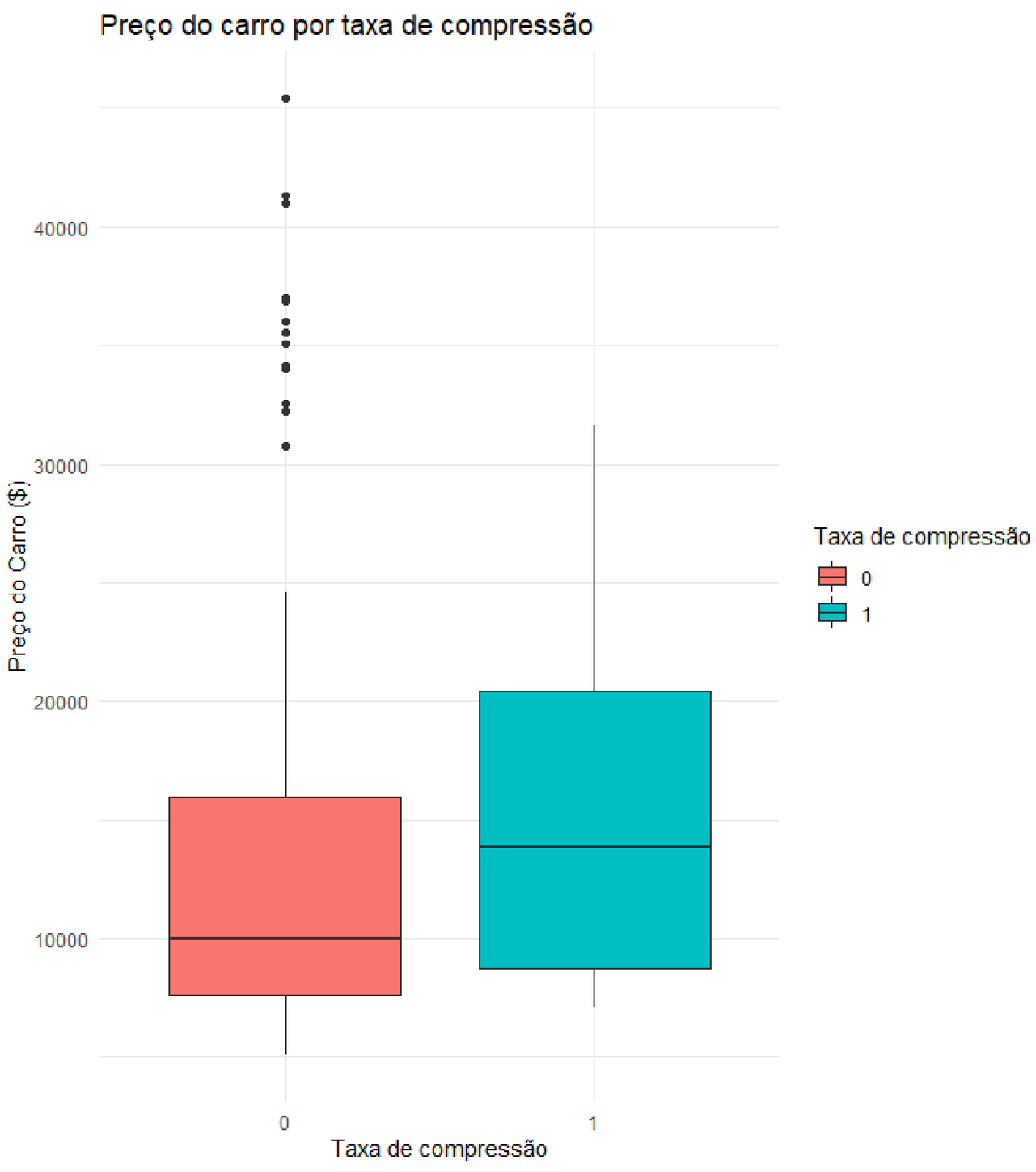

Preço do carro por Rodas Motrizes
40000
30000
20000
10000
4wd
fwd
rwd
Rodas Motrizes
Rodas Motrizes
4wd
fwd
rwd

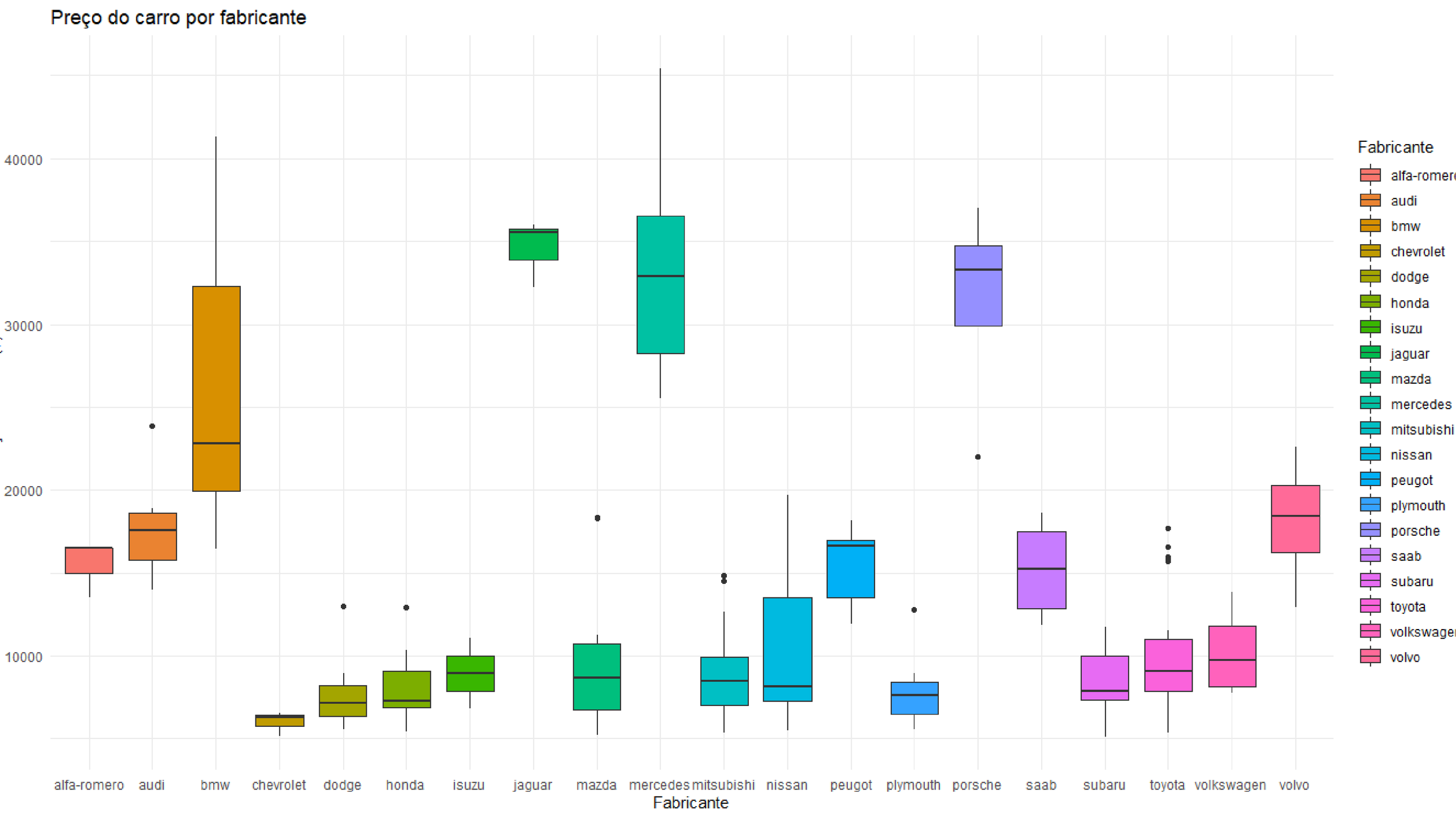

Preço do carro por Tipo de Motor
Preço do Carro ($)
40000
30000
20000
10000
dohc
l
ohc
ohcf
ohcv
Tipo de Motor
Tipo de Moto
dohc
l
ohc
ohcf
ohcv


Preço do carro por Localização do Motor
40000
30000
20000
10000
front
rear
Localização do Motor
Localização do Motor
front
rear


Preço do carro por número de cilíndros
Preço do Carro ($)
40000
30000
20000
10000
0
1
2
Número de cilíndros
Número de cilíndros
0
1
2


Preço do carro por Sistema de Combustível
40000
30000
20000
10000
1bbl
2bbl
idi
mfi
mpfi
spdi
spfi
Sistema de Combustível
Sistema de Combustível
1bbl
2bbl
idi
mfi
mpfi
spdi
spfi


Preço do carro por taxa de compressão
40000
30000
20000
10000
Preço do Carro ($)
0
1
Taxa de compressão
Taxa de compressão
0
1


Preço do carro por fabricante
40000
30000
20000
10000
alfa-romero
audi
bmw
chevrolet
dodge
honda
isuzu
jaguar
mazda
mercedes
mitsubishi
nissan
peugot
plymouth
porsche
saab
subaru
toyota
volkswagen
volvo
Fabricante
Fabricante
alfa-romero
audi
bmw
chevrolet
dodge
honda
isuzu
jaguar
mazda
mercedes
mitsubishi
nissan
peugot
plymouth
porsche
saab
subaru
toyota
volkswagen
volvo

Modelos simples

price ~ length + length^2;
price ~ width + width^2;
orice ~ curb_weigh + curb_weight^2;
price ~ engine_size;
price ~ horsepower;
price ~ city_mpg + city_mpg^2 + city_mpg^3 + city_mpg^4;
price ~ highway_mpg + highway_mpg^2 + highway_mpg^3.

Modelos múltiplos

Ajuste de dois modelos múltiplos cheios para seleção de variáveis.
Consideraram-se: length, width, curb_weight, engine_size, horsepower, city_mpg, highway_mpg, compression_ratio, num_cylinders, engine_type, engine_location, drive_wheels, body_style, aspiration, make para o modelo 1.

Modelos múltiplos

MODELO 2

Igual ao modelo 1, sem a variável make (muito correlacionada com as outras segundo correlação de Crammer)

STEPWISE

Seleção de variáveis utilizando AIC

Melhores modelos

NORMALIDADE

Os dois modelos tiveram alguns problemas com normalidade, por conta de uns poucos pontos.

HOMOCEDASTICIDADE

Ambos pareciam homocedásticos e passaram no Goldfeld.Quandt. No entanto, falharam no Breush-Pagan.

MEDIDAS DE INFLUÊNCIA

Alguns pontos são extremamente influentes. Número razoável de pontos que talvez não tenham sido bem ajustados.

Melhores modelos

VARIAÇÕES

Os pontos influentes mudam a depender do modelo.

COMPARAÇÕES

Os Modelo 1 (após o STEPWISE) era o que tinha o menor AIC.

MULTICOLINEARIDADE

Uma imagem vale mais que mil palavras:

Modelos múltiplos

```
> aux2[5]
[[1]]
[[1]]$VIFs
                                          GVIF Df
poly(width, degree = 2, raw = F)      3.693301e+01  2
poly(curb_weight, degree = 2, raw = F) 2.697714e+02  2
poly(engine_size, degree = 1, raw = F) 2.040108e+01  1
poly(city_mpg, degree = 4, raw = F)    6.152620e+01  4
engine_type                            3.837155e+03  4
drive_wheels                           9.817384e+00  2
body_style                             6.002099e+00  4
aspiration                             1.779449e+00  1
make                                   1.492038e+06 19
```

MODELO 1

Parece que make é realmente bem correlacionada com as outras;

```
[[1]]
[[1]]$VIFs
                                          GVIF Df
poly(length, degree = 2, raw = F)      41.142117  2
poly(width, degree = 2, raw = F)       17.628598  2
poly(curb_weight, degree = 2, raw = F)   118.236216  2
poly(horsepower, degree = 1, raw = F)     14.379579  1
poly(city_mpg, degree = 4, raw = F)    14024.077264  4
poly(highway_mpg, degree = 3, raw = F)  9929.354454  3
compression_ratio                          3.284065  1
num_cylinders                             17.411708  2
engine_type                               13.793931  4
engine_location                            2.868308  1
drive_wheels                               4.820219  2
```

MODELO 2

O modelo 2 selecionou algumas variáveis muito correlacionadas entre si, sendo que algumas delas foram descartadas pelo Modelo 1.

Considerações

A FAZER

Verificar relações entre variáveis categóricas e quantitativas explicativas;
Possivelmente adicionar termos de interação;

MAIS UMA MEDIDA

Utilizar outras medidas para seleçaõ de variáveis, como BIC e o R^2, verificar performance de modelos sem algumas variáveis que sabemos causar multicolinearidade;

POR ENQUANTO É SÓ.

Obrigado pela atenção!
Dúvidas ou sugestões?