Descriptive analysis

Karine Almeida

2023-10-12

Business Understanding

A fictitious car buying and selling company is having difficulties reselling used cars in its catalogue. With the aim of pricing your catalog more competi- tively and thus recovering the poor performance in this sector, I will analyze the data to answer business questions and create a predictive model that prices the customer's cars so that they are as close to market values. **In this notebook you will have access to a descriptive analysis of the data, insights and answers to some business questions.**

Descriptive analysis

Viewing the database

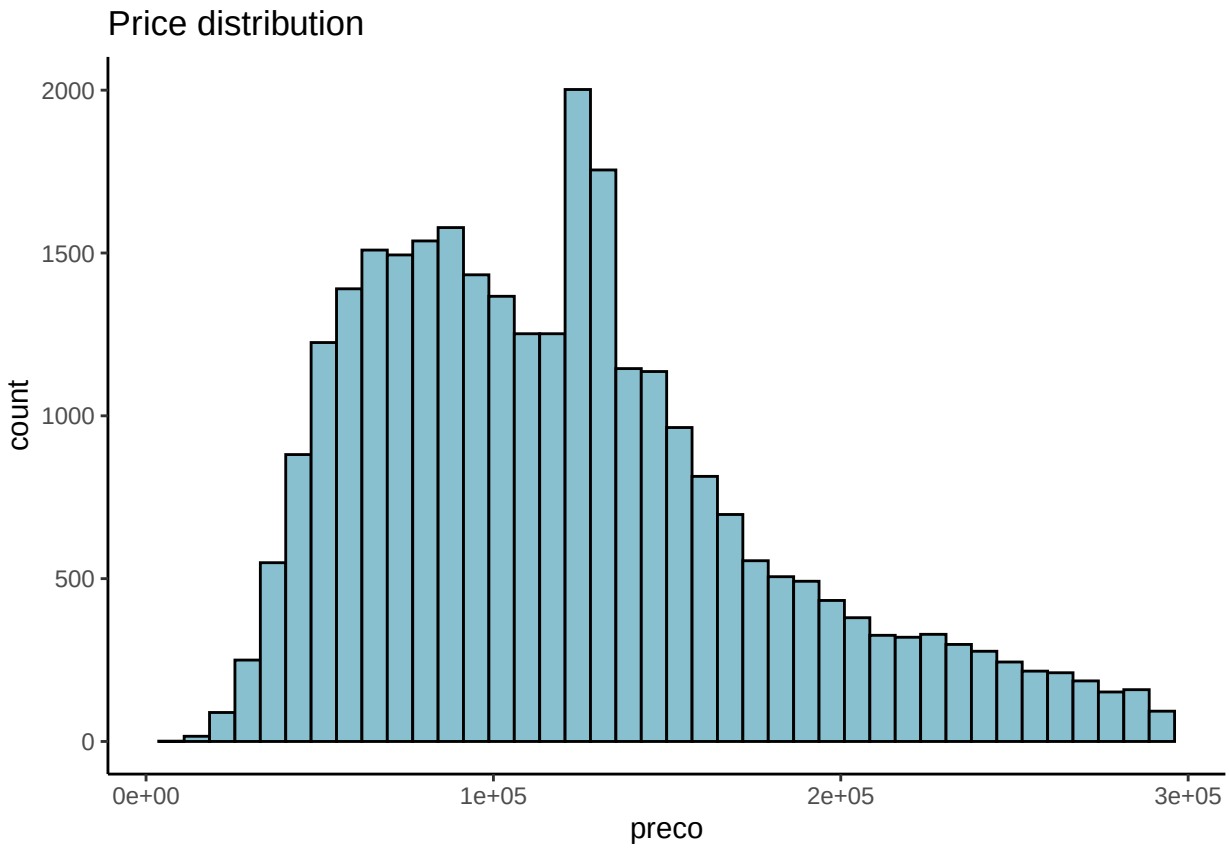
```
cars_train <- read_csv("../data/cars_train_clean1.csv", show_col_types = FALSE)
glimpse(cars_train)
```

```
## Rows: 29,513
## Columns: 23
## $ num_fotos          <dbl> 8, 8, 16, 14, 8, 13, 14, 15, 8, 15, 8, 8, 16, 8, 8,~
## $ marca              <chr> "NISSAN", "JEEP", "KIA", "VOLKSWAGEN", "SSANGYONG",~
## $ modelo             <chr> "KICKS", "COMPASS", "SORENTO", "AMAROK", "KORANDO",~
## $ versao             <chr> "1.6 16V FLEXSTART SL 4P XTRONIC", "2.0 16V FLEX LI~
## $ ano_de_fabricacao  <dbl> 2017, 2017, 2018, 2013, 2013, 2017, 2019, 2016, 201~
## $ ano_modelo         <dbl> 2017, 2017, 2019, 2015, 2015, 2018, 2019, 2017, 201~
## $ hodometro          <dbl> 67772.00, 62979.00, 44070.00, 85357.00, 71491.00, 8~
## $ cambio             <chr> "CVT", "Automática", "Automática", "Automática", "A~
## $ num_portas         <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, ~
## $ tipo               <chr> "Sedã", "Sedã", "Sedã", "Picape", "Utilitário espor~
## $ blindado           <chr> "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", "~
## $ cor                <chr> "Branco", "Branco", "Preto", "Branco", "Preto", "Br~
## $ tipo_vendedor      <chr> "PF", "PF", "PJ", "PJ", "PF", "PJ", "PJ", "PJ", "PF~
## $ cidade_vendedor    <chr> "Rio de Janeiro", "Belo Horizonte", "Santos", "Soro~
## $ estado_vendedor    <chr> "São Paulo (SP)", "Minas Gerais (MG)", "São Paulo (~
## $ anunciante         <chr> "Pessoa Física", "Pessoa Física", "Concessionária",~
## $ entrega_delivery   <lgl> FALSE, FALSE, TRUE, TRUE, FALSE, TRUE, TRUE, FALSE,~
## $ troca              <lgl> FALSE, FALSE, FALSE, TRUE, FALSE, TRUE, TRUE, FALSE~
## $ dono_aceita_troca  <chr> "não aceita troca", "Aceita troca", "Aceita troca",~
## $ veiculo_único_dono <chr> "mais de um dono", "mais de um dono", "mais de um d~
## $ ipva_pago          <chr> "IPVA pago", "IPVA pago", "ipva não pago", "IPVA pa~
## $ veiculo_licenciado <chr> "Licenciado", "não licenciado", "não licenciado", "~
## $ preco              <dbl> 74732.59, 81965.33, 162824.81, 123681.36, 82419.76,~
```

First of all it is important to know the characteristics of our target

Price distribution

```
ggplot(cars_train, aes(preco)) +
  geom_histogram(bins=40, col="black", fill="#88C0D0FF")+
  ggtitle('Price distribution') +
  theme_classic()
```



The target has a distribution tending to be asymmetric to the right (positive asymmetry).

Price descriptive statistics

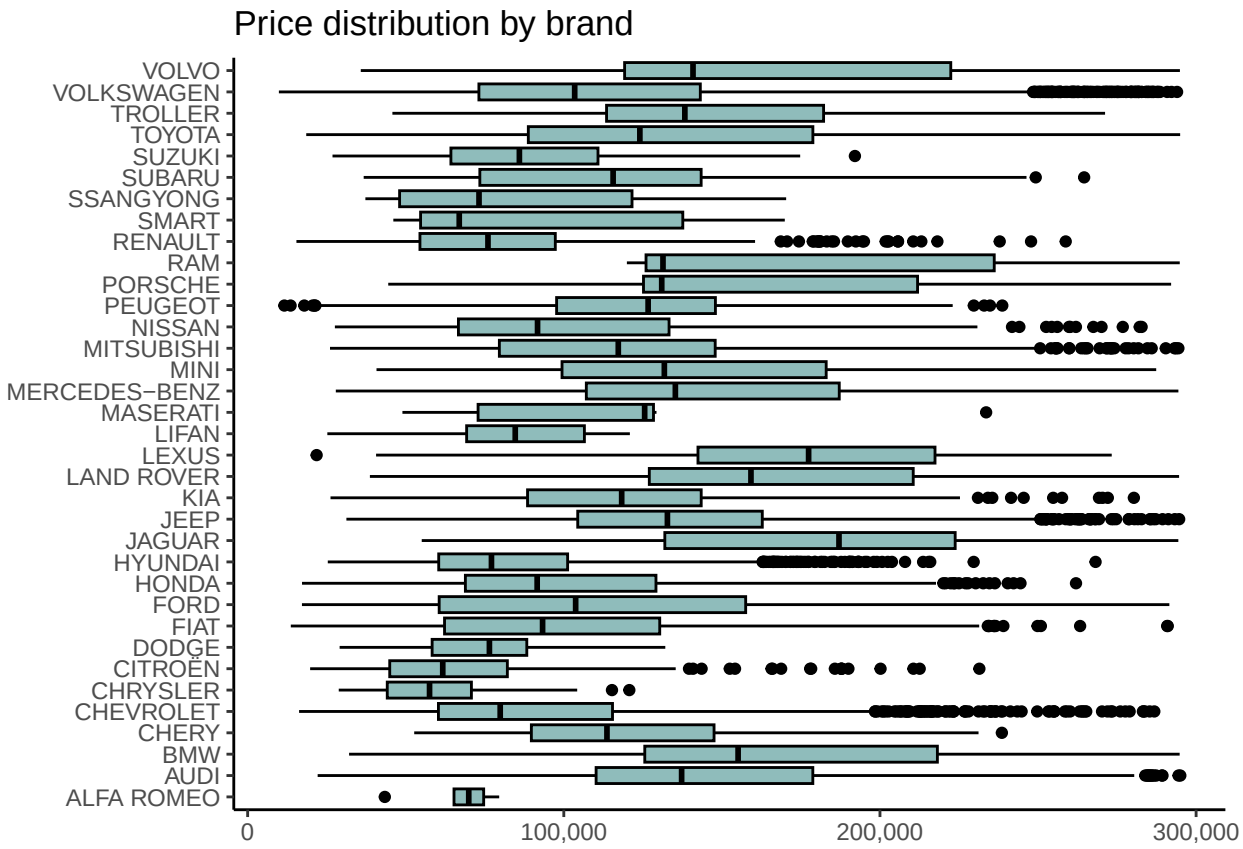
```
summary(cars_train$preco)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9870   76631  114435  121209  151654  295002
```

The price distribution reveals that 75% of cars sold have a value below 151,584 reais. The maximum value of a car sold was 295,000 reais and the minimum 9,870 reais.

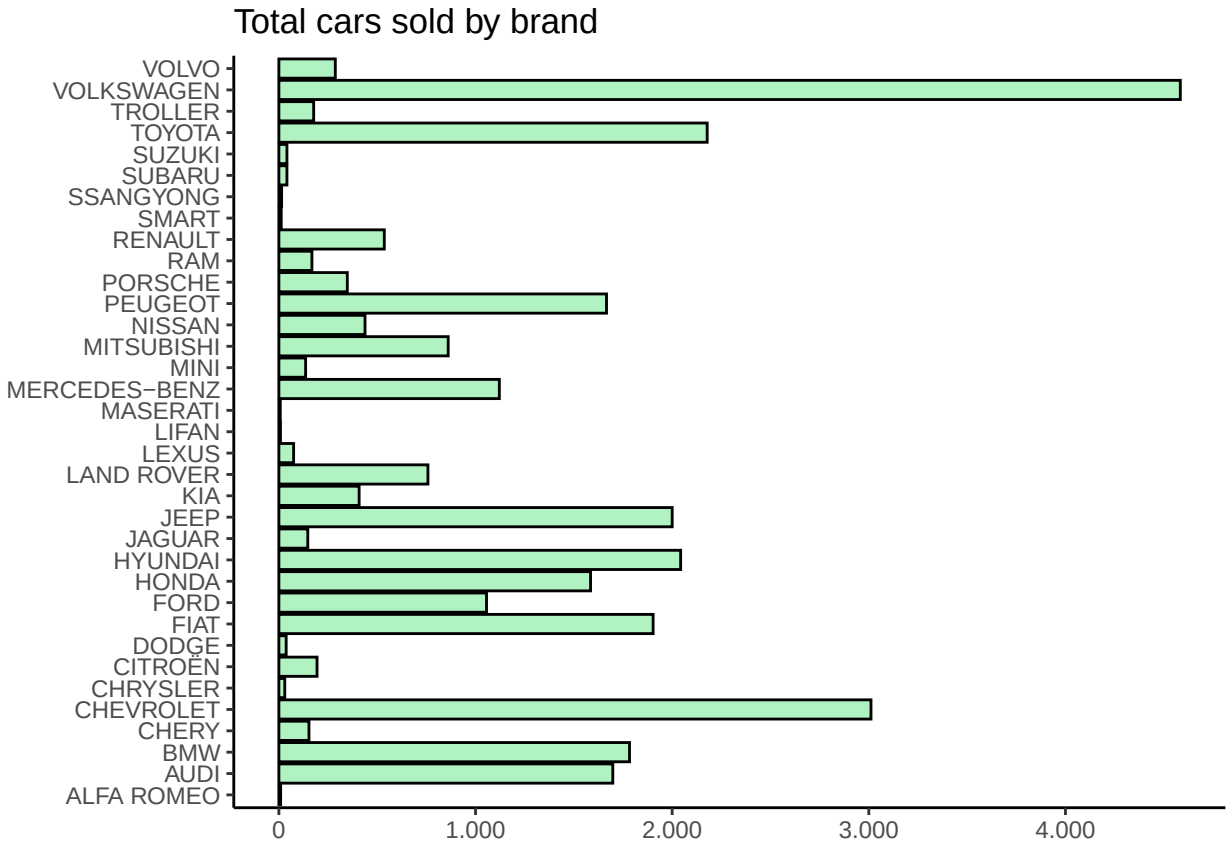
Relationship between price and brand

```
ggplot(cars_train, aes(x=preco, y=marca))+
  geom_boxplot(col= "black", fill= "#8FBCBBFF")+
  scale_x_continuous(labels = comma_format(decimal.mark = "."))+
  labs(y=NULL, x= NULL, title = 'Price distribution by brand') +
  theme_classic()
```



Which brands sell the most cars?

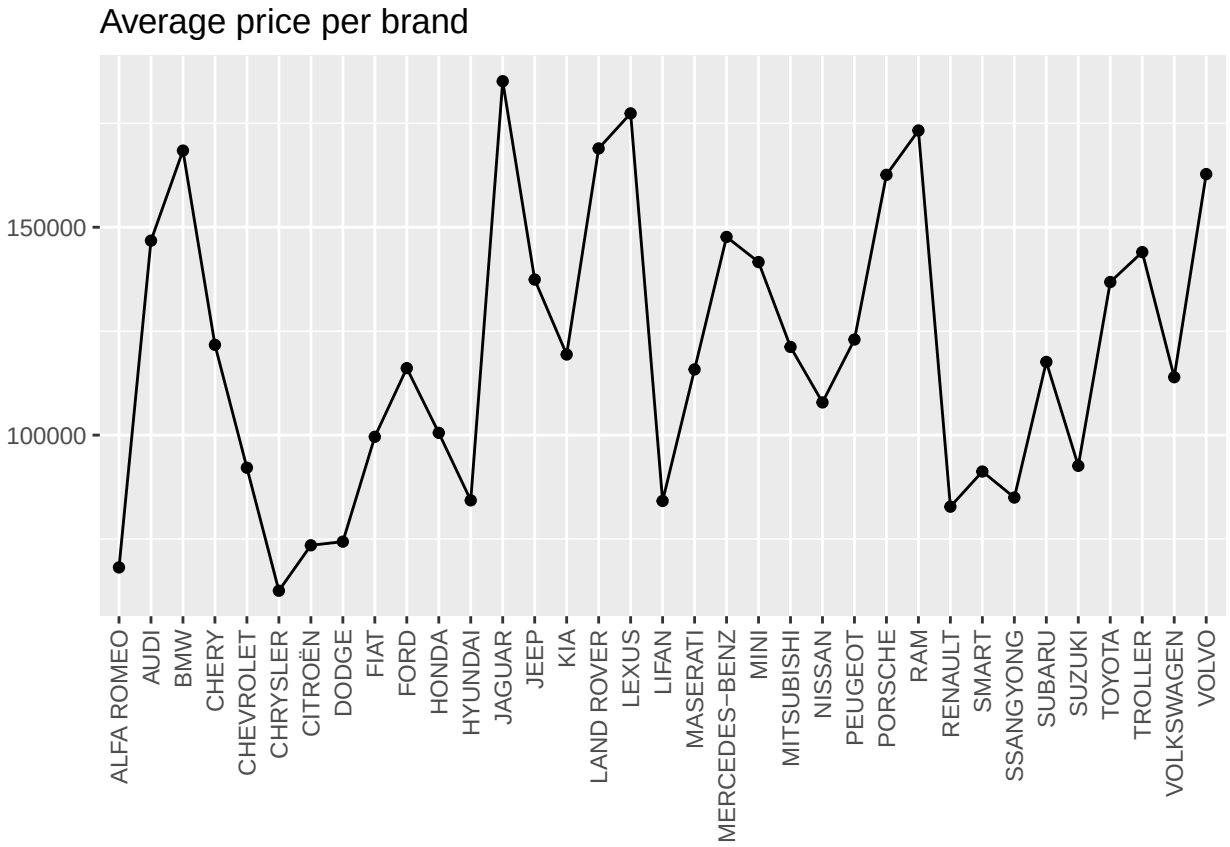
```
cars_train %>%
  group_by(marca) %>%
  summarise(total = n()) %>%
  ggplot(aes(x=total, y= marca))+
  geom_bar(stat = 'identity', fill="#b0f2c2", col="black", position = "dodge")+
  scale_x_continuous(labels = comma_format(big.mark = "."))+
  labs(y=NULL, x= NULL, title = "Total cars sold by brand")+
  theme(axis.text.y = element_blank())+
  theme_classic()
```



The 3 brands that sell the most are, respectively, **VOLKSWAGEN, CHEVROLET and TOYOTA**.

Average price per brand

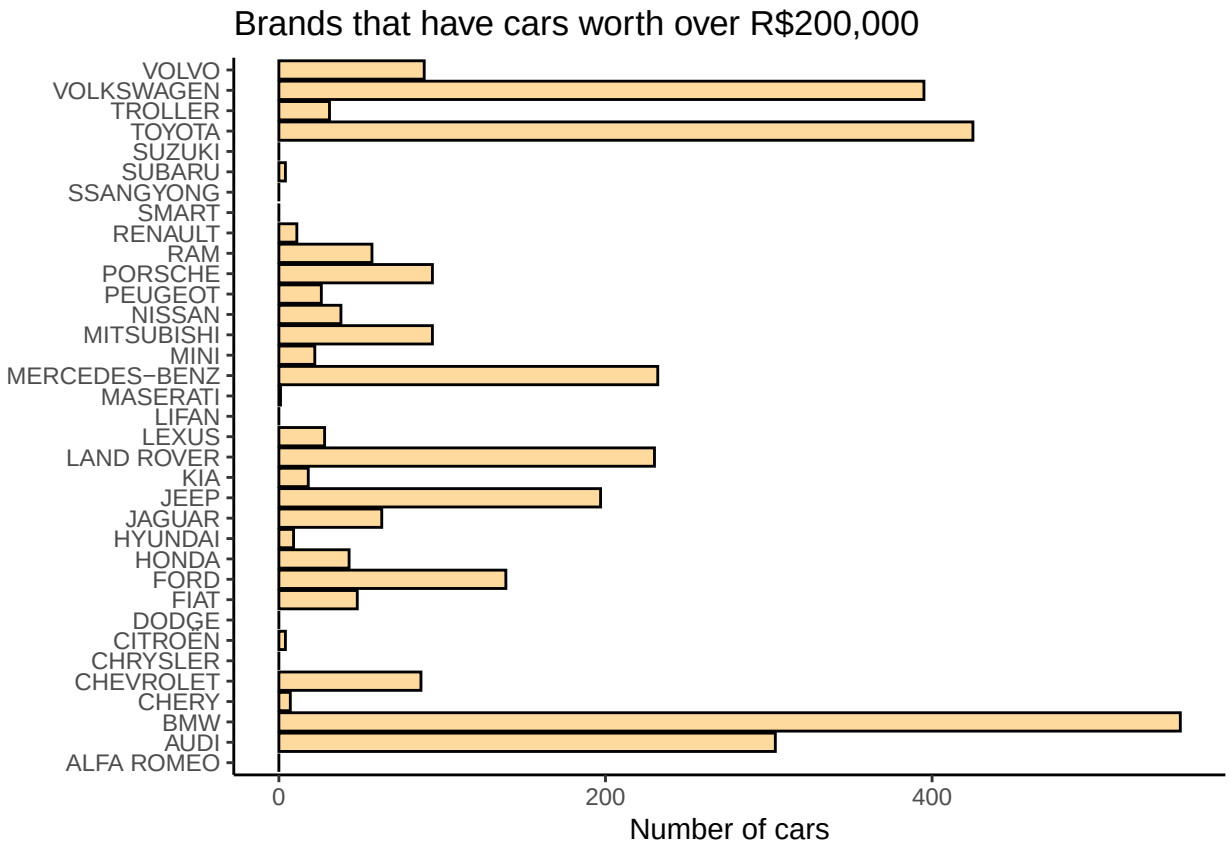
```
cars_train %>%
  group_by(marca) %>%
  summarise(mediaPreco = mean(preco)) %>%
  ggplot(aes(x = marca, y = mediaPreco, group=1))+
  geom_line()+
  geom_point()+
  labs(y=NULL, x= NULL, title = 'Average price per brand')+
  guides(x =  guide_axis(angle = 90))+
  theme()
```



Cars with the highest average price are JAGUAR, LEXUS, RAM and BMW

Which brands have cars worth more than R$200,000 and how many cars do they have?

```
cars_train %>%
  group_by(marca) %>%
  summarise(Preco = sum(preco > 200000))  %>%
  ggplot(aes(x=Preco, y= marca))+
  geom_bar(stat = 'identity', fill="#ffda9e", col="black", position = "dodge")+
  labs(y=NULL, x= 'Number of cars', title = 'Brands that have cars worth over R$200,000')+
  theme_classic()
```



Among the 40 registered brands, which ones are responsible for 50% of sales?

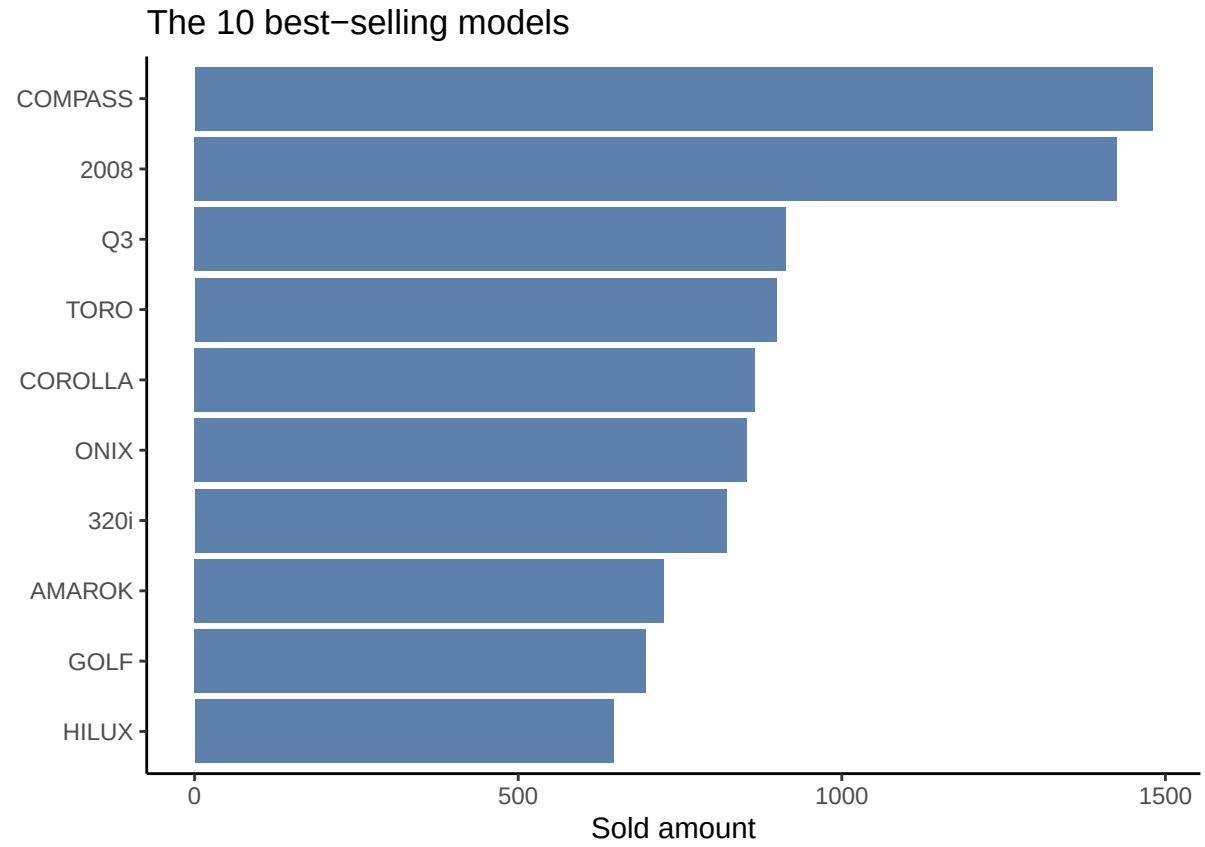
```
cars_train %>%
  group_by(marca) %>%
  summarise(percent= (n()/nrow(cars_train)*100)) %>%
  slice_max(percent, n=6) %>%
  arrange(desc(percent)) %>%
  kable(col.names = c('Brand', 'Percentage'))
```

Brand	Percentage
VOLKSWAGEN	15.532138
CHEVROLET	10.202284
TOYOTA	7.379799
HYUNDAI	6.922373
JEEP	6.776675
FIAT	6.448006

Among all the models listed, which are the 10 best sellers?

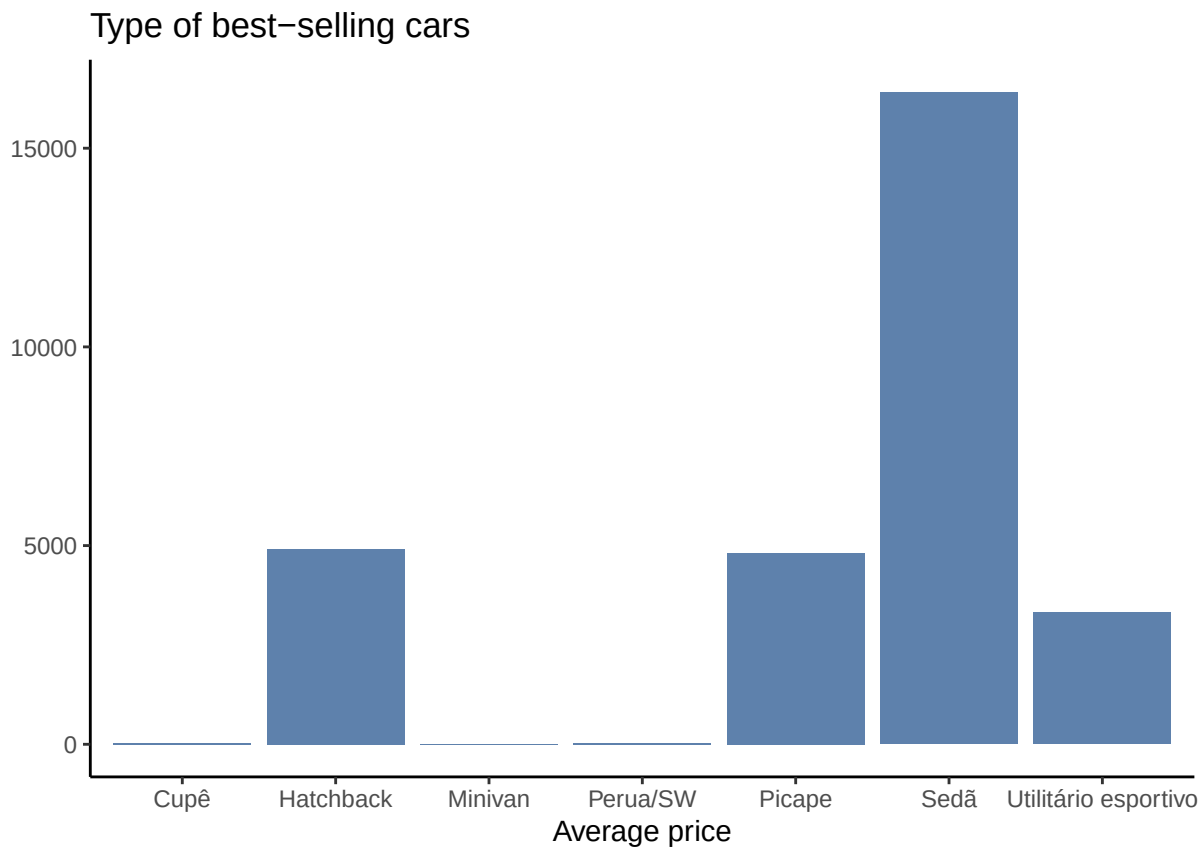
```
cars_train %>%
  group_by(modelo) %>%
  summarise(qtd_vendida = n()) %>%
  slice_max(qtd_vendida, n=10) %>%
  mutate(modelo = reorder(modelo, qtd_vendida)) %>%
```

```
  ggplot(aes(y=modelo, x=qtd_vendida)) +
  geom_col(just = 0.5, fill="#5E81ACFF") +
  labs(y="", x="Sold amount", title= "The 10 best-selling models") +
  theme_classic()
```



What types of cars sell the most?

```
cars_train %>%
  group_by(tipo) %>%
  summarise(qtd_vendida = n()) %>%
  ggplot(aes(x=tipo, y=qtd_vendida)) +
  geom_col(fill="#5E81ACFF")+
  labs(y="", x="Average price", title= "Type of best-selling cars") +
  theme_classic()
```



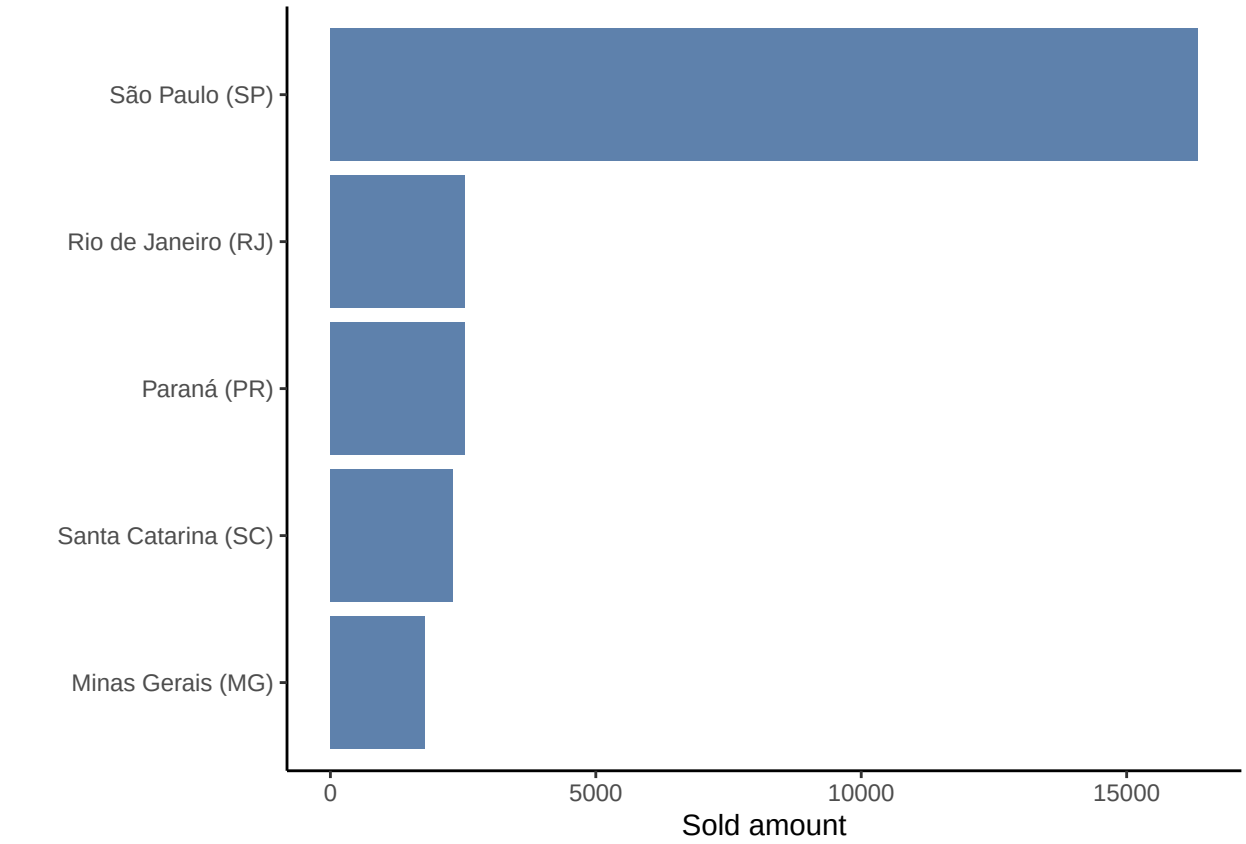
What are your favorite car versions? Let's make a top 10.

```
cars_train %>%
  group_by(versao) %>%
  summarise(total = n()) %>%
  slice_max(total, n=10) %>%
  mutate(versao = reorder(versao, total)) %>%
  kable(col.names = c('Version', 'Sold amount'))
```

Version	Sold amount
1.6 16V FLEX ALLURE PACK 4P AUTOMÁTICO	1346
2.0 16V FLEX LIMITED AUTOMÁTICO	1032
2.0 TFSI AMBIENTE QUATTRO 4P GASOLINA S TRONIC	757
1.0 200 TSI COMFORTLINE AUTOMÁTICO	465
2.0 HIGHLINE 4X4 CD 16V TURBO INTERCOOLER DIESEL 4P AUTOMÁTICO	429
1.8 16V EVO FLEX FREEDOM AT6	406
1.4 MPFI LS CS 8V FLEX 2P MANUAL	352
2.0 SPORT 16V TURBO ACTIVE FLEX 4P AUTOMÁTICO	341
2.0 XEI 16V FLEX 4P AUTOMÁTICO	324
2.0 TSI GTI 16V TURBO GASOLINA 4P AUTOMÁTICO	306

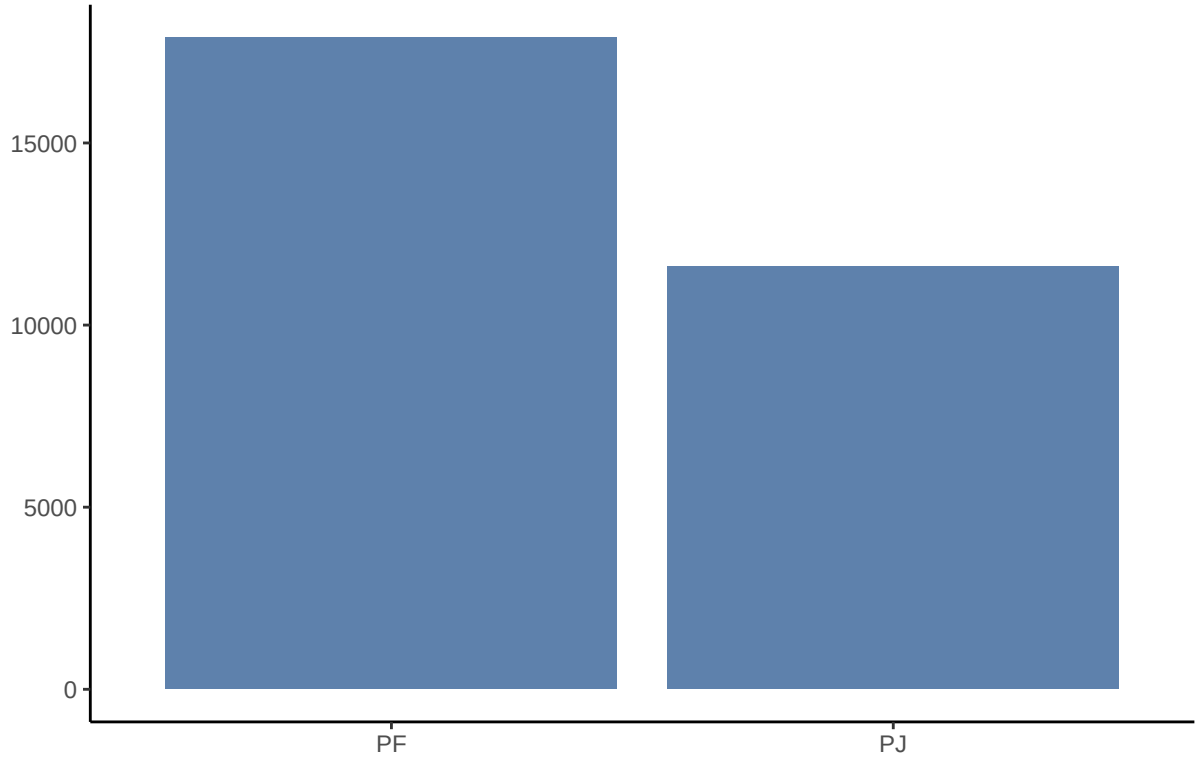
What are the 5 states with the highest number of sales?

```r
cars_train %>%
  group_by(estado_vendedor) %>%
  summarise(qtd_vendas = n()) %>%
  slice_max(qtd_vendas, n=5) %>%
  mutate(estado_vendedor= reorder(estado_vendedor, qtd_vendas)) %>%
  ggplot(aes(y=estado_vendedor, x=qtd_vendas)) +
  geom_col(just = 0.5, fill="#5E81ACFF") +
  labs(y="", x="Sold amount") +
  theme_classic()
```



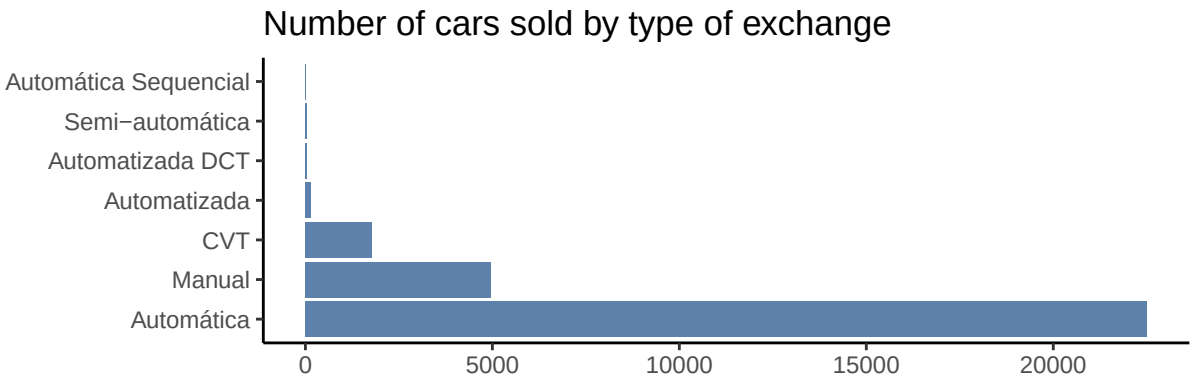
Number of cars sold by PF and PJ

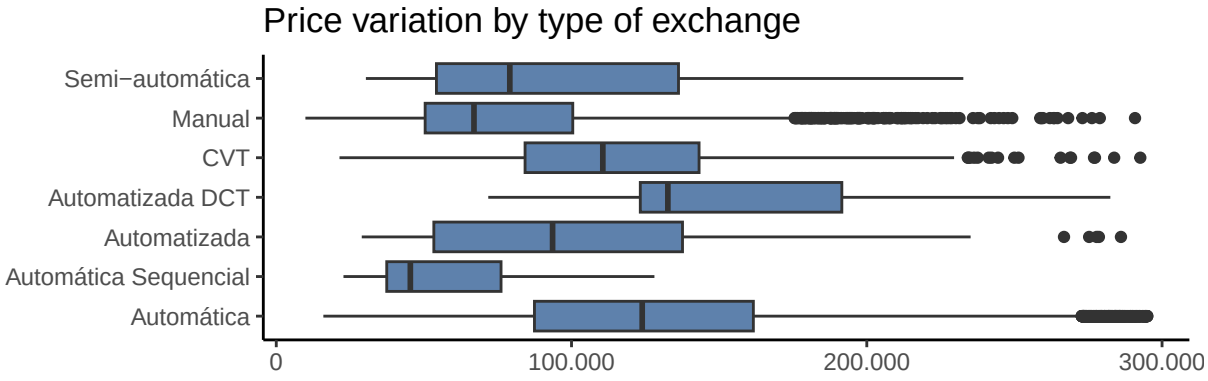
```r
cars_train %>%
  group_by(tipo_vendedor) %>%
  summarise(total = n()) %>%
  ggplot(aes(y=total, x=tipo_vendedor)) +
  geom_col(just = 0.5, fill="#5E81ACFF") +
  labs(y="", x="", title= "") +
  theme_classic()
```



Exchange vs price

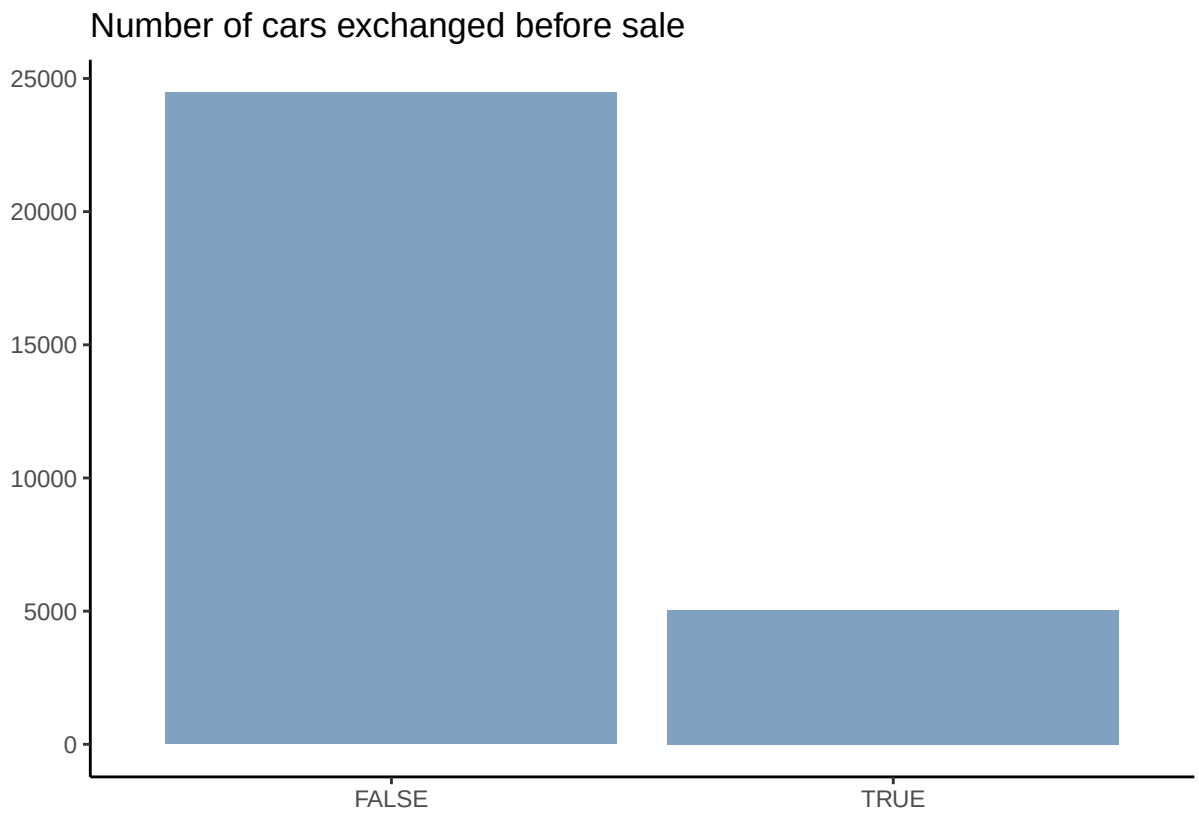
```
g7 <- cars_train %>%
  group_by(cambio) %>%
  summarise(total = n()) %>%
  mutate(cambio = reorder(cambio, -total)) %>%
  ggplot(aes(x=total, y=cambio)) +
  geom_col(just = 0.5, fill="#5E81ACFF") +
  labs(y="", x="", title= "Number of cars sold by type of exchange") +
  theme_classic()

g8 <- cars_train %>%
  ggplot(aes(x=preco, y=cambio)) +
  geom_boxplot(show.legend = FALSE, fill="#5E81ACFF") +
  labs(y="", x="", title= "Price variation by type of exchange") +
  scale_x_continuous(labels = comma_format(big.mark = ".")) +
  theme_classic()

grid.arrange(g7, g8, ncol=1)
```





Car exchange

```
cars_train %>%
  ggplot(aes(x=troca)) +
  geom_bar(fill= "#81A1C1FF") +
  labs(title="Number of cars exchanged before sale", x="", y="") +
  theme_classic()
```

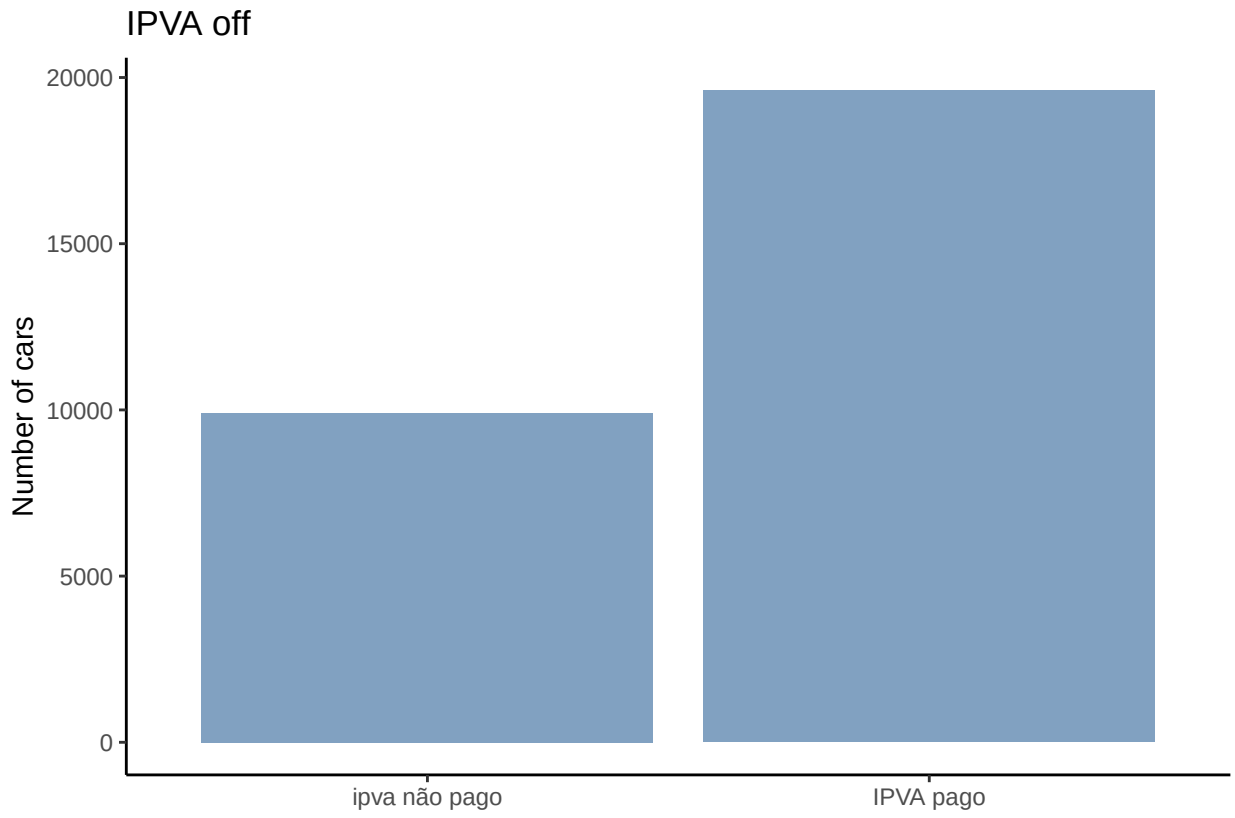



```
print(paste(sum(cars_train$troca=='FALSE'),
            "cars were never exchanged, therefore, more than 80% of cars sold always belonged to the sa
```

```
## [1] "24480 cars were never exchanged, therefore, more than 80% of cars sold always belonged to the s
```

What percentage of cars have IPVA paid off?

```
cars_train %>%
  ggplot(aes(x=ipva_pago)) +
  geom_bar(fill= "#81A1C1FF") +
  labs(title="IPVA off", x="", y="Number of cars") +
  theme_classic()
```

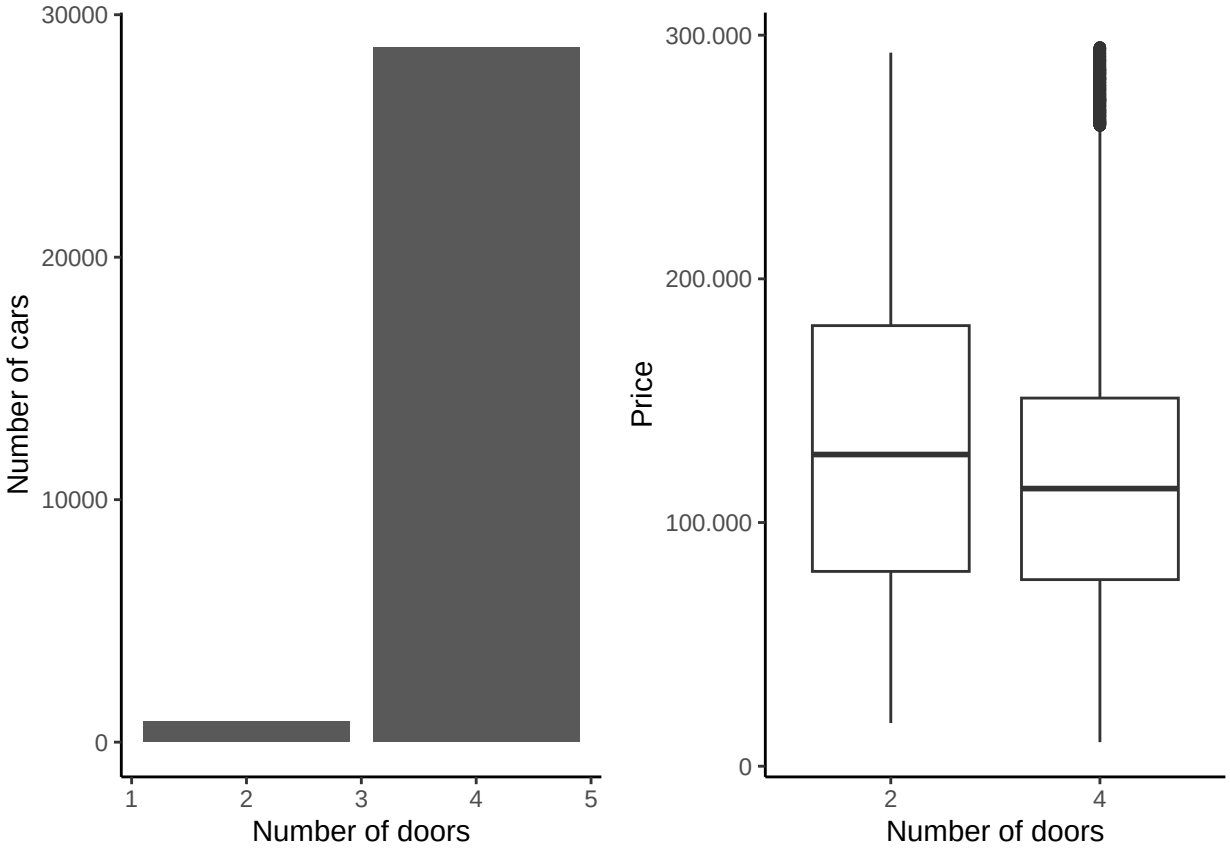



```
print(paste(sum(cars_train$ipva_pago=='IPVA pago'), "cars have paid IPVA, which represents approximatel
    round((sum(cars_train$ipva_pago=='IPVA pago')*100/nrow(cars_train))),"% of registered cars."))
```

```
## [1] "19614 cars have paid IPVA, which represents approximately 66 % of registered cars."
```

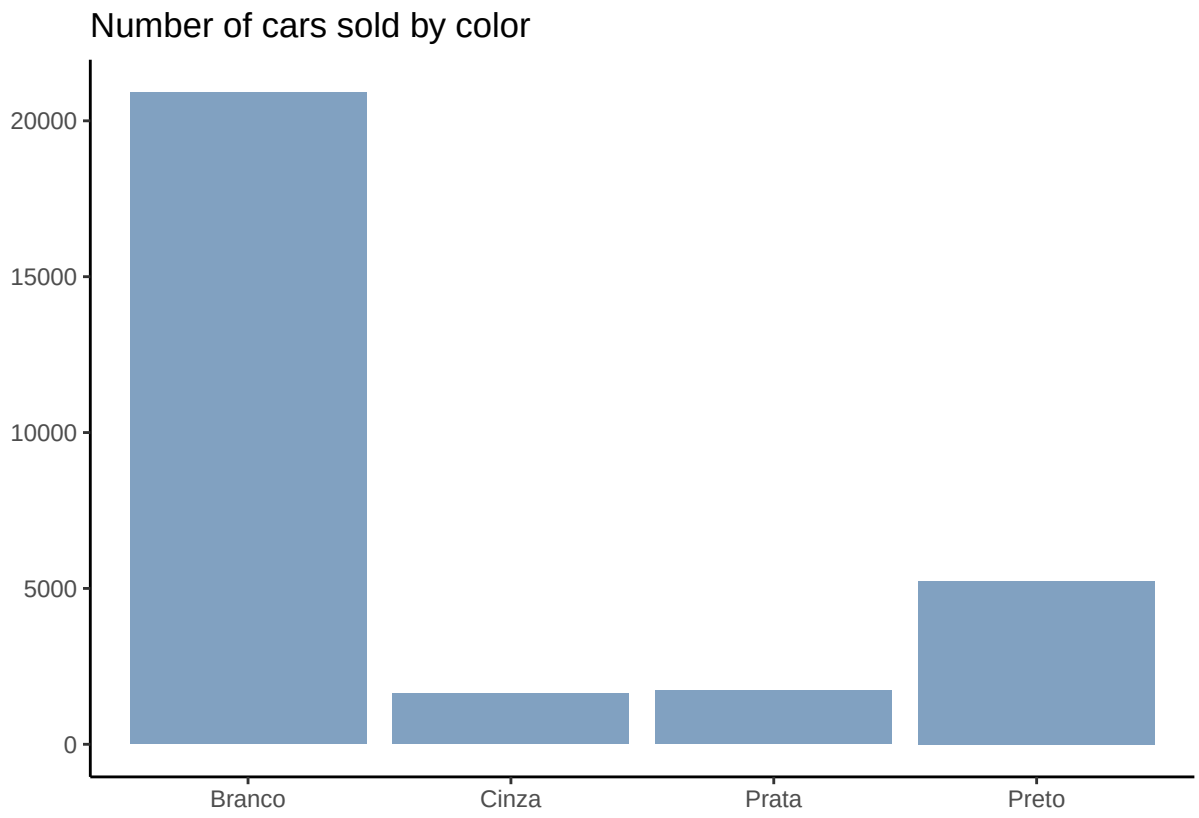
Relationship between the number of doors and the number of cars sold and the price

```
cars_train$num_portas[cars_train$num_portas == 3] <- 4

g9 <- cars_train %>%
  ggplot(aes(x=num_portas)) +
  geom_bar() +
  labs(x="Number of doors", y="Number of cars") +
  theme_classic()

g10 <- cars_train %>%
  ggplot(aes(x=as.factor(num_portas), y=preco))+
  geom_boxplot(show.legend = FALSE)+
  labs(x="Number of doors", y="Price") +
  scale_y_continuous(labels = comma_format(big.mark = "."))+
  theme_classic()

grid.arrange(g9, g10, ncol=2)
```



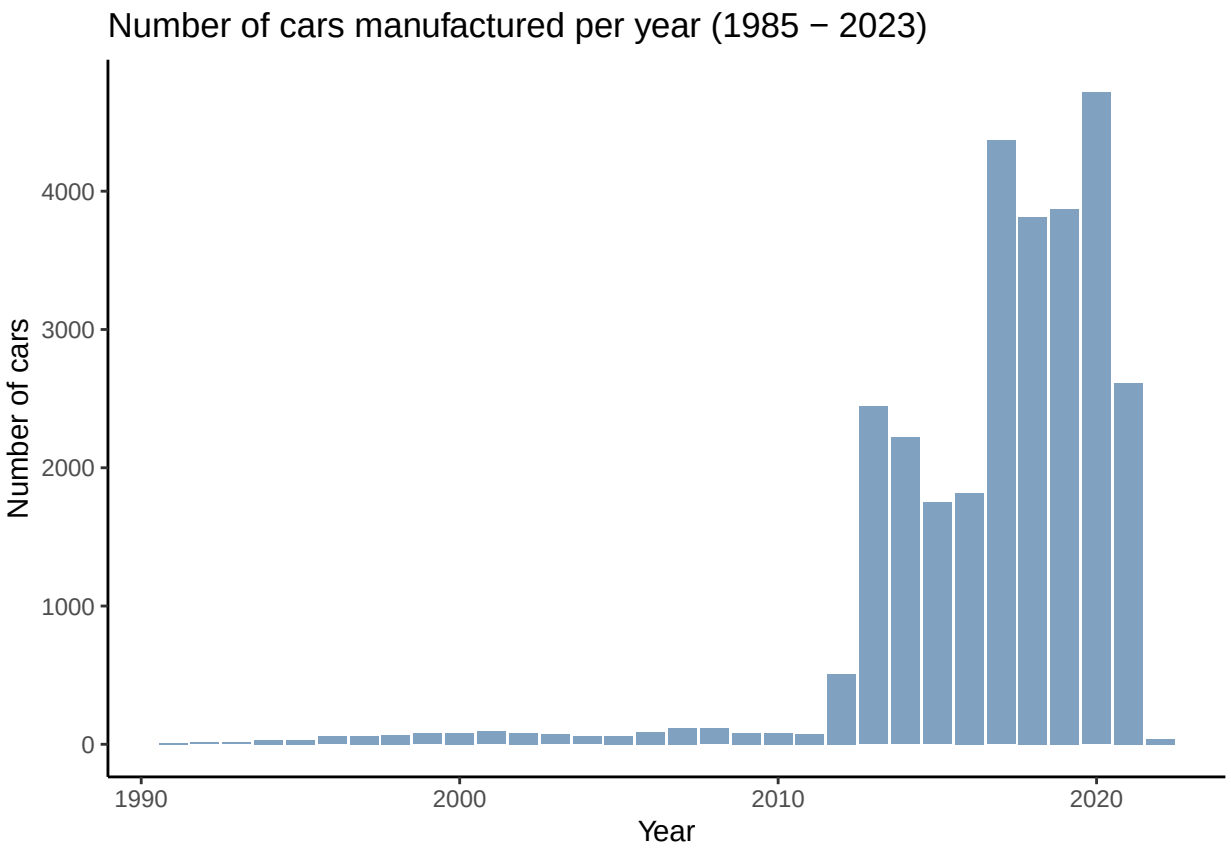
Black and white are the best-selling colors. Watch below:

```
cars_train %>%
  ggplot(aes(x=cor)) +
  geom_bar(fill= "#81A1C1FF") +
  labs(x="", y="", title="Number of cars sold by color") +
  theme_classic()
```



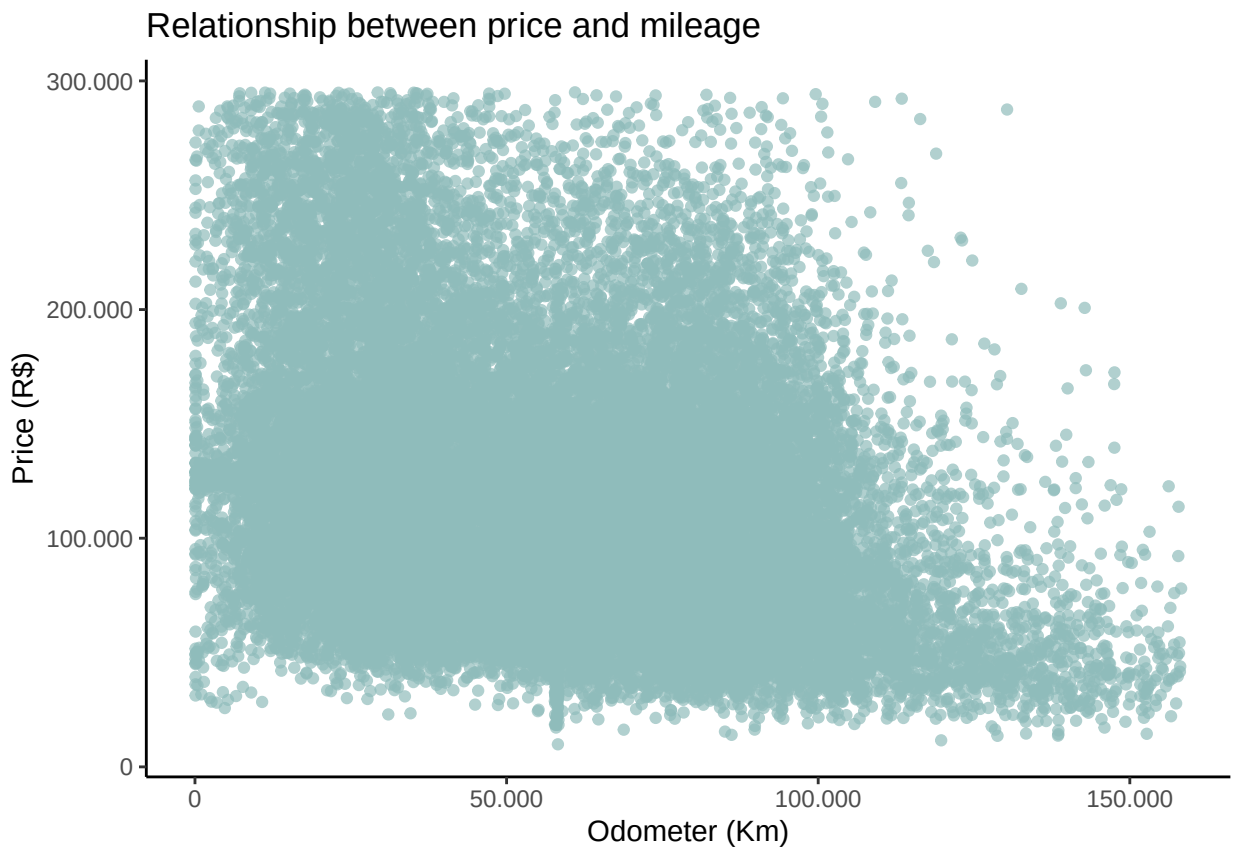
The best-selling cars were manufactured in the last 10 years

```
cars_train %>%
  ggplot(aes(x=ano_de_fabricacao)) +
  geom_bar(fill= "#81A1C1FF") +
  labs(x="Year", y="Number of cars", title = "Number of cars manufactured per year (1985 - 2023)") +
  theme_classic()
```



Relationship between price and mileage

```
cars_train %>%
  ggplot(aes(x= hodometro, y = preco)) +
  geom_point(alpha=0.7, color= "#8FBCBBFF", fill="black") +
  scale_y_continuous(labels = comma_format(big.mark = "."))+
  scale_x_continuous(labels = comma_format(big.mark = "."))+
  labs(x="Odometer (Km)", y="Price (R$)", title="Relationship between price and mileage") +
  theme_classic()
```



Although the data shows a slight trend that the higher the mileage, the lower the price of the car, it is observed from the distribution of the data that the majority of prices are not influenced by mileage.

What is the best state registered in the database to sell a popular brand car and why?

To facilitate the analysis, I will only filter popular brands, so we can analyze the data more specifically. For this I will create a new object.

```
marca_popular <-
  cars_train[cars_train$marca == c("NISSAN","CHERY","CITROËN","VOLKSWAGEN","PEUGEOT", "HONDA",
```

Now I will select the 5 states with the highest number of popular car sales.

```
marca_popular %>%
  group_by(estado_vendedor) %>%
  summarise(qtd_vendas = n()) %>%
  slice_max(qtd_vendas, n=5) %>%
  mutate(estado_vendedor= reorder(estado_vendedor, -qtd_vendas)) %>%
  kable()
```

estado_vendedor	qtd_vendas
São Paulo (SP)	933
Rio de Janeiro (RJ)	161
Santa Catarina (SC)	129
Paraná (PR)	127

estado_vendedor	qtd_vendas
Rio Grande do Sul (RS)	98

São Paulo is the state with the highest sales of popular cars, which may indicate turnover and greater consumption of this product in relation to other states.

Let's look at the average price and variation in the price of these cars in each state.

```
marca_popular %>%
  group_by(estado_vendedor) %>%
  summarise(avg_preco = mean(preco),
            desvio_padrao = sd(preco)) %>%
  kable()
```

estado_vendedor	avg_preco	desvio_padrao
Acre (AC)	97821.98	58232.94
Alagoas (AL)	119079.69	59875.26
Amazonas (AM)	111625.94	32596.03
Bahia (BA)	103739.41	60536.01
Ceará (CE)	141138.51	86341.41
Espírito Santo (ES)	105613.26	48961.01
Goiás (GO)	116494.84	48519.24
Mato Grosso (MT)	179038.95	60423.63
Mato Grosso do Sul (MS)	40480.62	NA
Minas Gerais (MG)	111967.20	54965.84
Paraná (PR)	117864.53	52111.68
Paraíba (PB)	90557.32	51455.10
Pará (PA)	110208.58	43157.31
Pernambuco (PE)	93443.73	38737.13
Rio Grande do Norte (RN)	139463.59	36485.38
Rio Grande do Sul (RS)	116516.23	52318.16
Rio de Janeiro (RJ)	107892.86	55319.18
Santa Catarina (SC)	105282.51	55544.94
Sergipe (SE)	97894.39	70117.65
São Paulo (SP)	104336.03	52345.24
Tocantins (TO)	67021.22	NA

São Paulo does not have the lowest average price in relation to other states, and the standard deviation reveals that the variation in the value of these cars is approximately 59 thousand reais in relation to the average. Even though it does not have the most favorable values compared to other states, it is important to consider that:SP has the most populous city in Brazil.

As it is a city with a large territorial area, many people, in search of greater comfort, buy pre-owned popular cars, as they are the most affordable.

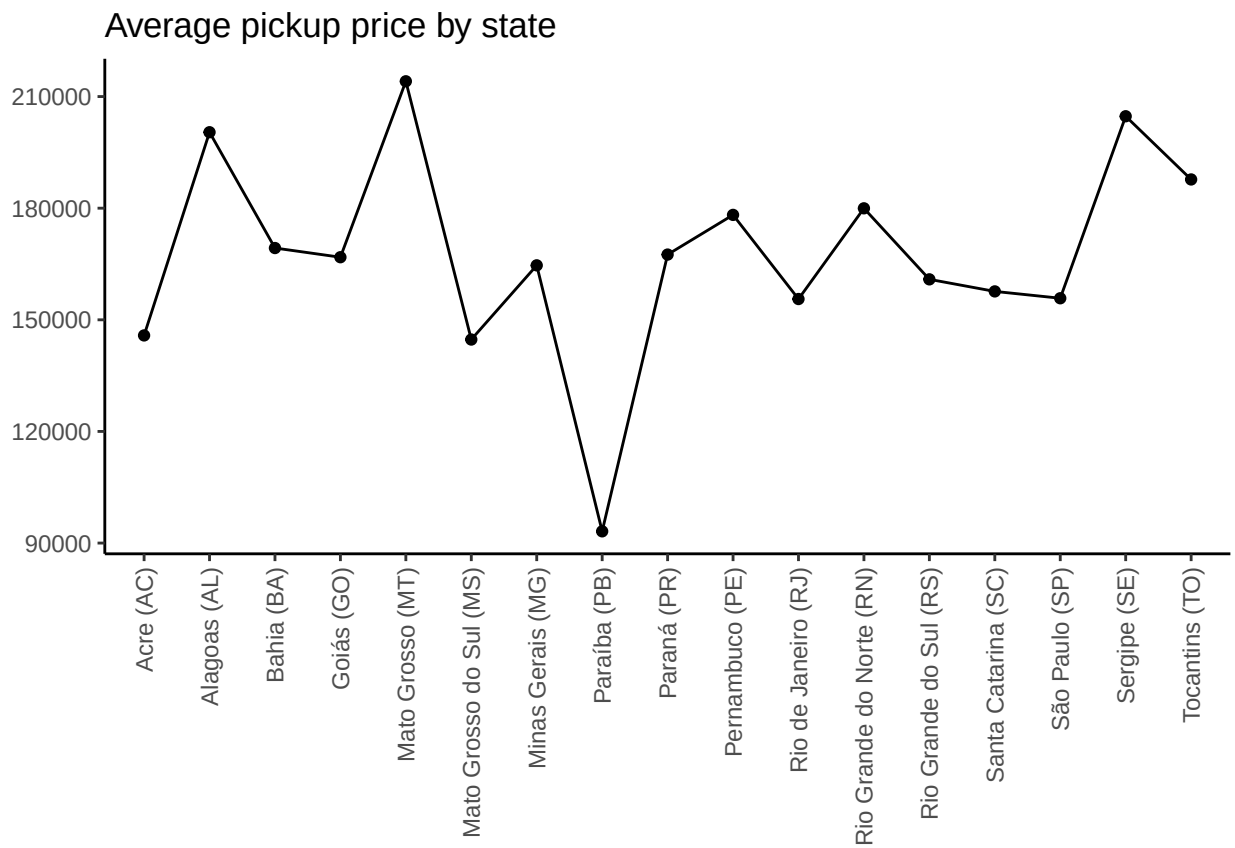
With this, I conclude that São Paulo is the best state to sell popular cars.
What is the best state to buy a pickup truck with an automatic transmission and why?

Firstly, I will create an object that filters the type of car and exchange rate, so that my variables bring data specific to this demand.

```
picape <- cars_train[cars_train$tipo == "Picape" & cars_train$cambio == "Automática", ]
```

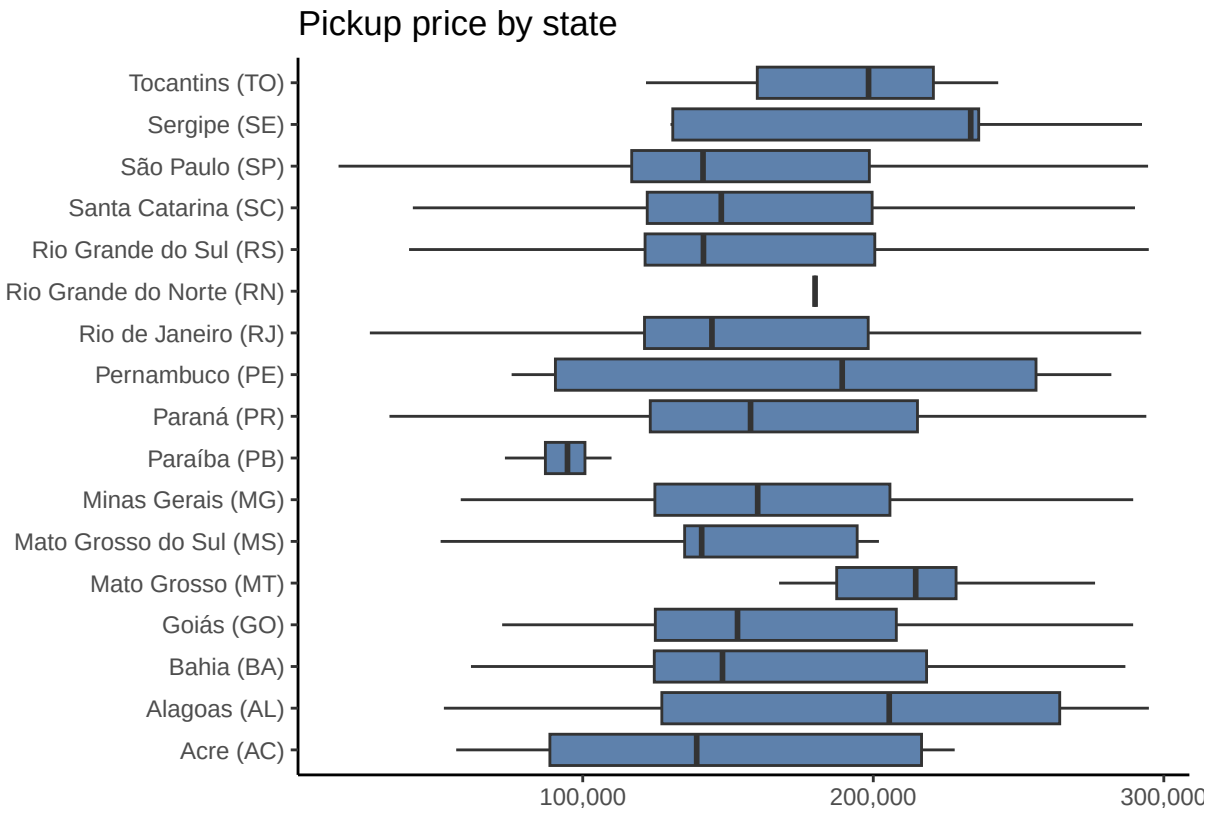
Now I will calculate the average price of an automatic pickup truck by state

```
#Gráfico 1
picape %>%
  group_by(estado_vendedor) %>%
  summarise(avg_preco = mean(preco)) %>%
  ggplot(aes(y=avg_preco, x=estado_vendedor, group=1)) +
  geom_line()+
  geom_point()+
  labs(y=NULL, x= NULL, title = 'Average pickup price by state')+
  guides(x =  guide_axis(angle = 90))+
  theme_classic()
```



Let's see how the price of this type of car varies within the states?

```
picape %>%
  group_by(estado_vendedor) %>%
  ggplot(aes(x=preco, y=estado_vendedor)) +
  geom_boxplot(fill="#5E81ACFF") +
  labs(y="", x="", title= "Pickup price by state") +
  scale_x_continuous(labels = comma_format(decimal.mark = "."))+
  theme_classic()
```



Paraíba seems to be the most financially suitable option. Although it is in the northeast of the country and the buyer's demand may be in the south, south-east or north and the buyer has to pay the shipping, it is still more advantageous, as the average price difference in relation to other states is at least 100 thousand reais (excluding the Acre).