MobAuto - Prova Técnica

Daniel David de Oliveira

04 de Março de 2021

Prova Técnica

Contexto da MobAuto

Atuamos no setor de mobilidade automobilística e temos por objetivo facilitar a dinâmica do mercado, tornando o processo de vendas mais fácil e lucrativo aos vendedores e também auxiliando as pessoas na escolha do carro a ser comprado.

Problema

O desafio consiste em fornecer informações a quatro concessionárias para entender melhor o comportamento sobre o mercado de carros. Elas desejam auxílio para compor seu estoque, escolhendo quais produtos devem trazer para a sua loja, tais como marca, ano e cor, por exemplo.

Cada uma está em um Estado diferente, sendo eles: São Paulo, Distrito Federal, Bahia e Rio Grande do Sul.

Em um primeiro momento, ambas concessionárias pediram pelas seguintes informações:

1) A minha região (estado) possui algum comportamento que se difere do Brasil? Nos quesitos:

a) A cor de um carro parece importar no seu preço?

b) Quais são as marcas e modelos mais presentes no mercado?

2) Olhando para os modelos mais predominantes (maior presença) na minha região:

a) Algum deles apresenta uma maior desvalorização? Ou uma maior valorização (caso ocorra)? Comparar preços ao longo do tempo (outubro, novembro e dezembro).

b) Existe uma desvalorização do carro baseado na quilometragem dele? Comparar preços mediante à quilometragem.

Bibliotecas a serem utilizadas

Resposta

```r
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(lattice)
library(gplots)
```

```
##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess
```

```r
library(dunn.test)

options(scipen = 10000000) #Tirar a notação científica dos gráficos
```

Lendo o banco de dados e nomeando as variáveis

```r
setwd("C:\\Users\\danie\\OneDrive\\Documentos\\R\\Cases\\MobAuto")

dados = read.csv("car_data_intern.csv",
                 sep = ",", na.strings="NA",
                 stringsAsFactors=T)

#criando um data.frame espelho
dds <- dados


#Nomeando as variáveis
colnames(dds) = c('data','marca','modelo',
                  'ano','preco',
                  'km','estado',
                  'cor')
```

Verificando se há dados faltantes

```r
    #Usei no R, e tirei do .pdf
#dds[!complete.cases(dds),]
```

```r
summary(dds$preco)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
##         0     34900     44205     54622     61990 121212120       120
```

A variável Preço possui NAs nas quais substutuirei pela mediana, pois são 120 NAs, corresponde menos de 1% das observações do banco de dados.

Substituindo

```r
#Substituindo os NAs pela mediana
dds[is.na(dds$preco),]$preco = median(dds$preco, na.rm = T)
```

1) a) A cor de um carro parece importar no seu preço?

Estou filtrando as cores de carros que mais vende, e isto nos proporciona 95% das observações.

```r
preco <- dds

#Corresponde a 95% do banco de dados
preco <- preco[preco$cor %in% c('Branco', 'Cinza',
                                'Prata','Preto',
                                'Vermelho'),]
```

Respondendo a pergunta, se a cor do carro influência em seu preço.

```r
#H0: A cor do carro não importa em seu preço
#Ha: A cor do carro importa em seu preço

kruskal.test(preco$preco, preco$cor)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  preco$preco and preco$cor
## Kruskal-Wallis chi-squared = 5410, df = 4, p-value <
## 0.00000000000000022
```

Resposta: Sim, a cor do carro é um fator importante em seu preço.

Sofisticando a pergunta, será que o preço do carro da cor vermelha possui diferença significativa em relação ao preço do carro branco?

```r
dunn.test(preco$preco, preco$cor, method="holm")
```

```
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 5409.9596, df = 4, p-value = 0
##
##
##                            Comparison of x by group
##                                      (Holm)
## Col Mean-|
## Row Mean |     Branco      Cinza      Prata      Preto
## ---------+-----------------------------------------------
##    Cinza |  -3.090377
##          |    0.0010*
##          |
##    Prata |   55.97575   44.82371
##          |    0.0000*    0.0000*
##          |
##    Preto |   28.43519   24.95951  -22.13386
##          |    0.0000*    0.0000*    0.0000*
##          |
## Vermelho |   51.09203   47.48888   18.50039   31.68452
##          |    0.0000*    0.0000*    0.0000*    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha/2
```

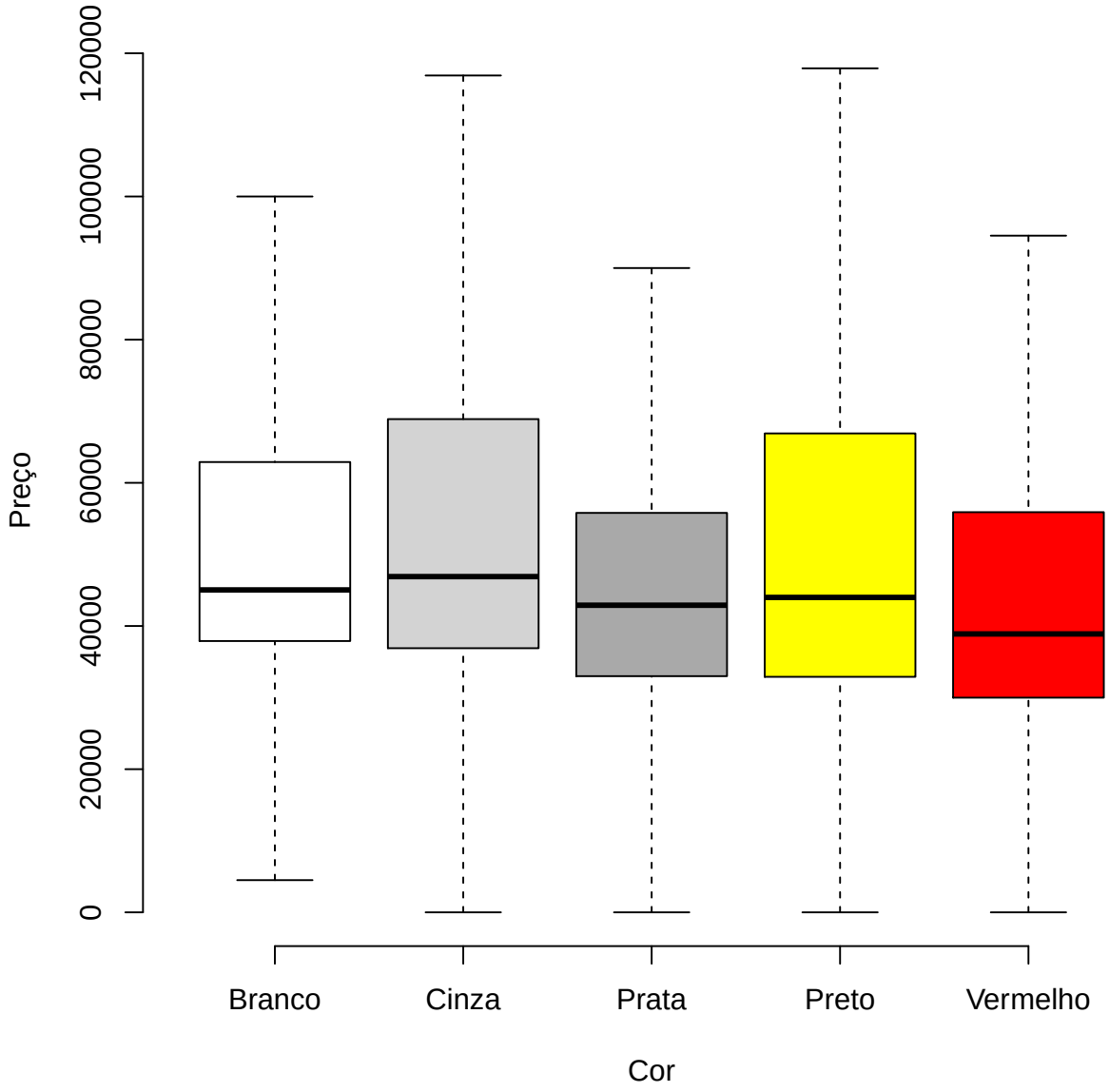
Sim, a cor do carro é um fator significativo na composição do preço. Todas combinações par a par, por exemplo, o preço do carro prata é diferente do preço do carro preto.

Vamos buscar um meio de explicitar essa conclusão em imagens.

Primeira idéia é buscar gráficos boxplot lado a lado e comparar-los.

Plotei sem os outlines, pois levando-os em conta, o *dashboard* ficaria com difícil interpretação.

```r
#removendo fatores não usados
preco$cor = factor(preco$cor)

#boxplot da variável preço e cor
boxplot(preco ~ cor,
        data = preco, xlab = "Cor",
        ylab = "Preço",
        frame = FALSE, col = c('white','light gray',
                               'dark gray', 'yellow',
                               'red'),outline = F)
```



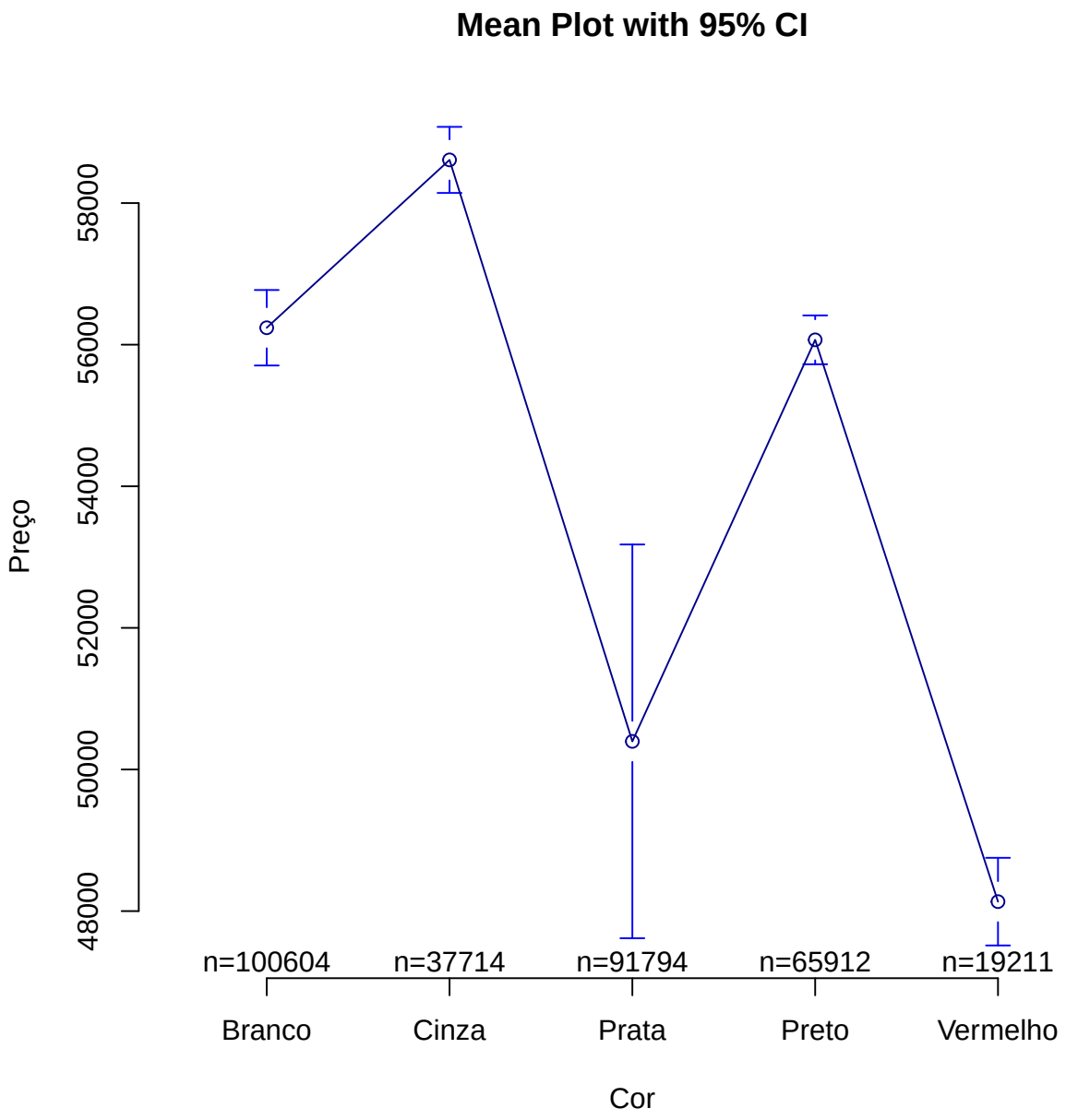
Perceba que os boxplots possuem tamanhos diferentes validando a conclusão dada no teste de kruskal-Wallis.

Outro meio de verificar se o preço do carro é influenciável pela cor é usando a função plotmeans do pacote gplots.

```
library(gplots)

plotmeans(preco ~ cor, data = preco, frame = FALSE, xlab = "Cor",
          ylab = "Preço",main="Mean Plot with 95% CI",
          col = 'dark blue')
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "frame" não é um
## parâmetro gráfico

## Warning in axis(1, at = 1:length(means), labels = legends, ...): "frame" não é
## um parâmetro gráfico
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "frame" não é um
## parâmetro gráfico
```



Há distâncias significativas das médias dos preços em relação a cor, portanto temos fortes evidências que a cor do carro influencia em seu preço.

b) Quais são as marcas e modelos mais presentes no mercado?

A primeira pergunta que vem a mente é: Quais as marcas que mais vende nacionalmente? E os modelos?

Para responder essas perguntas verificarei as marcas mais vendidas e modelos, respectivamente.

Logo em seguida, iremos comparar com as regiões correspondentes as concessionárias.

Quais são as marcas mais vendidas nacionalmente?

O processo de criação dos top10 das marcar e modelos vai ser repetido nesse processo: fazer um

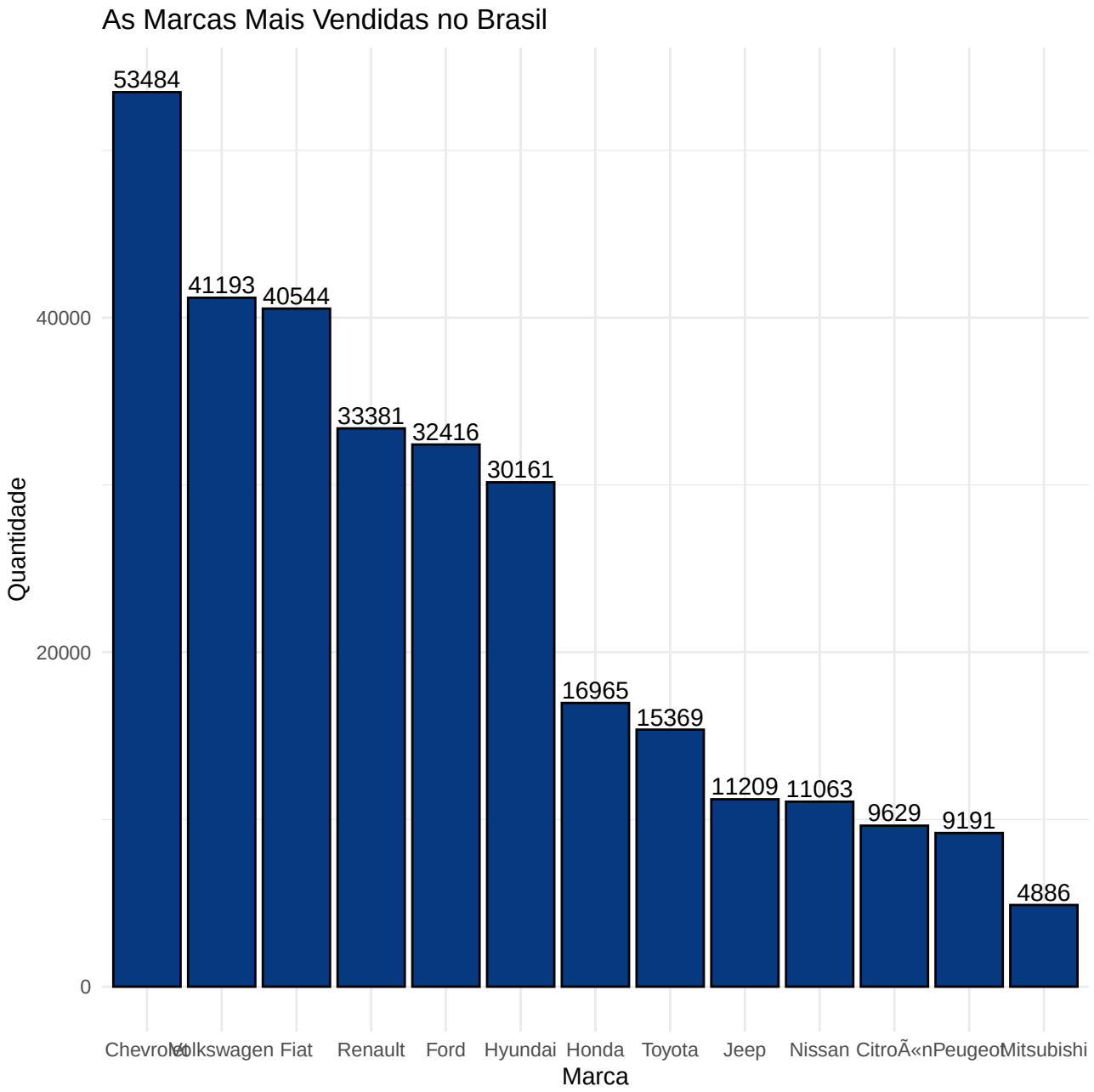
gráfico simples vendo em ordem decrescente das marcas/modelos mais vendidas, depois, vamos filtrar e mostrar um gráfico mais agradável visualmente.

Gráfico simples

```
#Verificando se há NAs na variável
dds[is.na(dds$marca),]
```

```
## [1] data   marca  modelo ano    preco  km     estado cor
## <0 rows> (or 0-length row.names)
```

```
count_data <- dds %>%
  count(marca)

      #plotando
ggplot(count_data[count_data$n > 4000,],
       aes(x = reorder(marca,-n), y = n )) +

  geom_bar(stat = 'identity',
           fill = '#073980', color = 'black') +

  geom_text(aes(label = n), vjust = -.25) +
  labs(x = 'Marca', y = 'Quantidade',
       title = 'As Marcas Mais Vendidas no Brasil') +

   theme_minimal()
```



Vimos as marcas que mais vende nacionalmente. A quantidade está explicitado acima das barras. Buscaremos visualisar as 10 maiores marcas e modelos. Com intuíto de não poluir os *dashboards*. Apoiado no gráfico acima, criaremos um data.frame chamado top10.

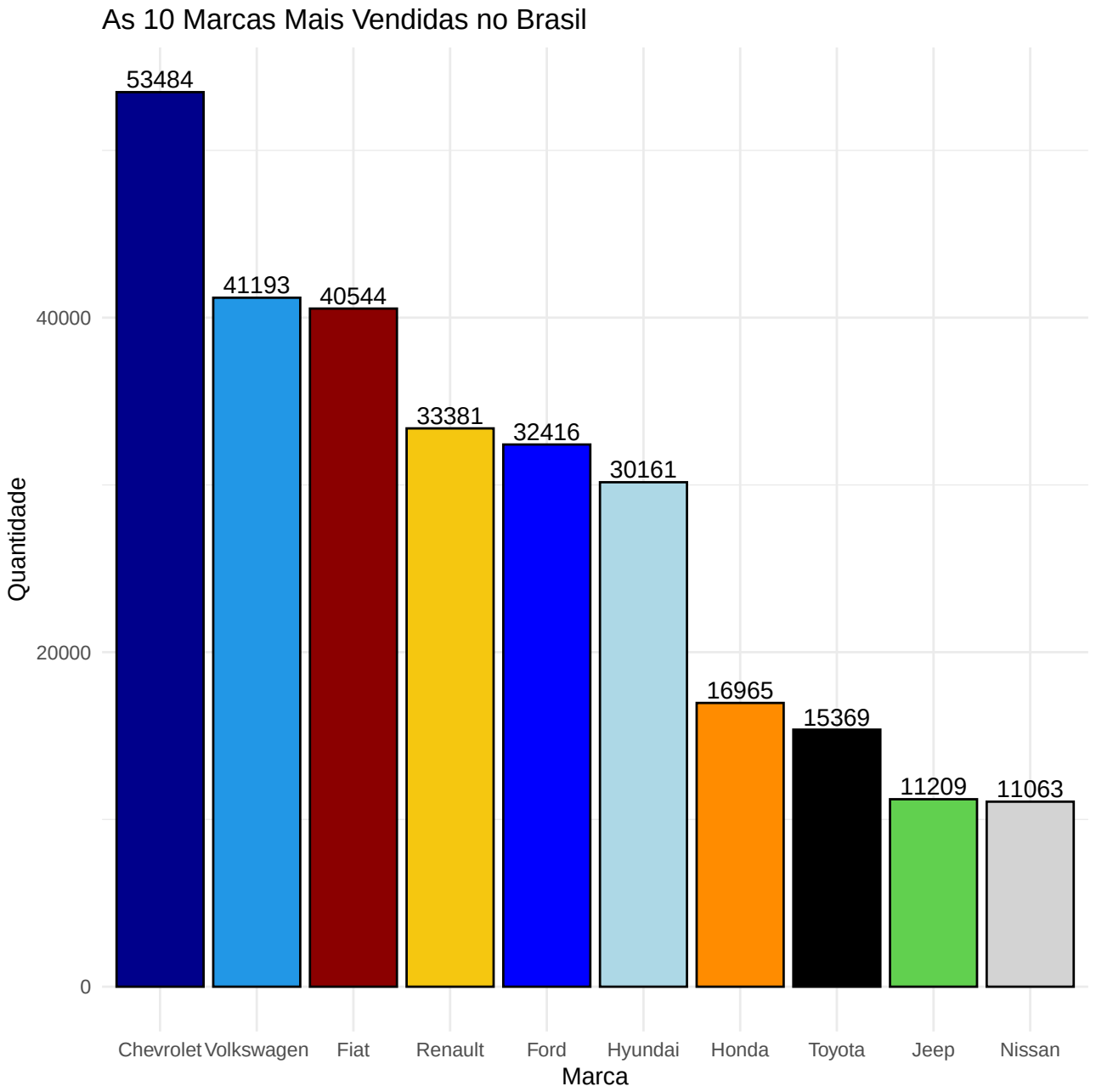
Gráfico filtrado

```
#### Selecionando as 10 Marcas mais Vendidas do Brasil ####

top10 <- dds[dds$marca %in% c('Chevrolet', 'Fiat',
                              'Ford', 'Honda',
                              'Hyundai', 'Jeep',
                              'Nissan', 'Renault',
                              'Toyota', 'Volkswagen'),]
```

Para as demais regiões apresentaremos somente os gráficos filtrados.

As marcas mais presente no mercado nacional são

```r
count_data <- top10 %>%
  count(marca)

      #plotando
ggplot(count_data, aes(x = reorder(marca,-n), y = n )) +

  geom_bar(stat = 'identity',
           fill = c('dark blue','dark red','blue',
                    'dark orange','light blue', '100403',
                    'light gray','7f935b','black',
                    '073980'), color = 'black') +

  geom_text(aes(label = n), vjust = -.25) +
  labs(x = 'Marca', y = 'Quantidade',
       title = 'As 10 Marcas Mais Vendidas no Brasil') +

   theme_minimal()
```

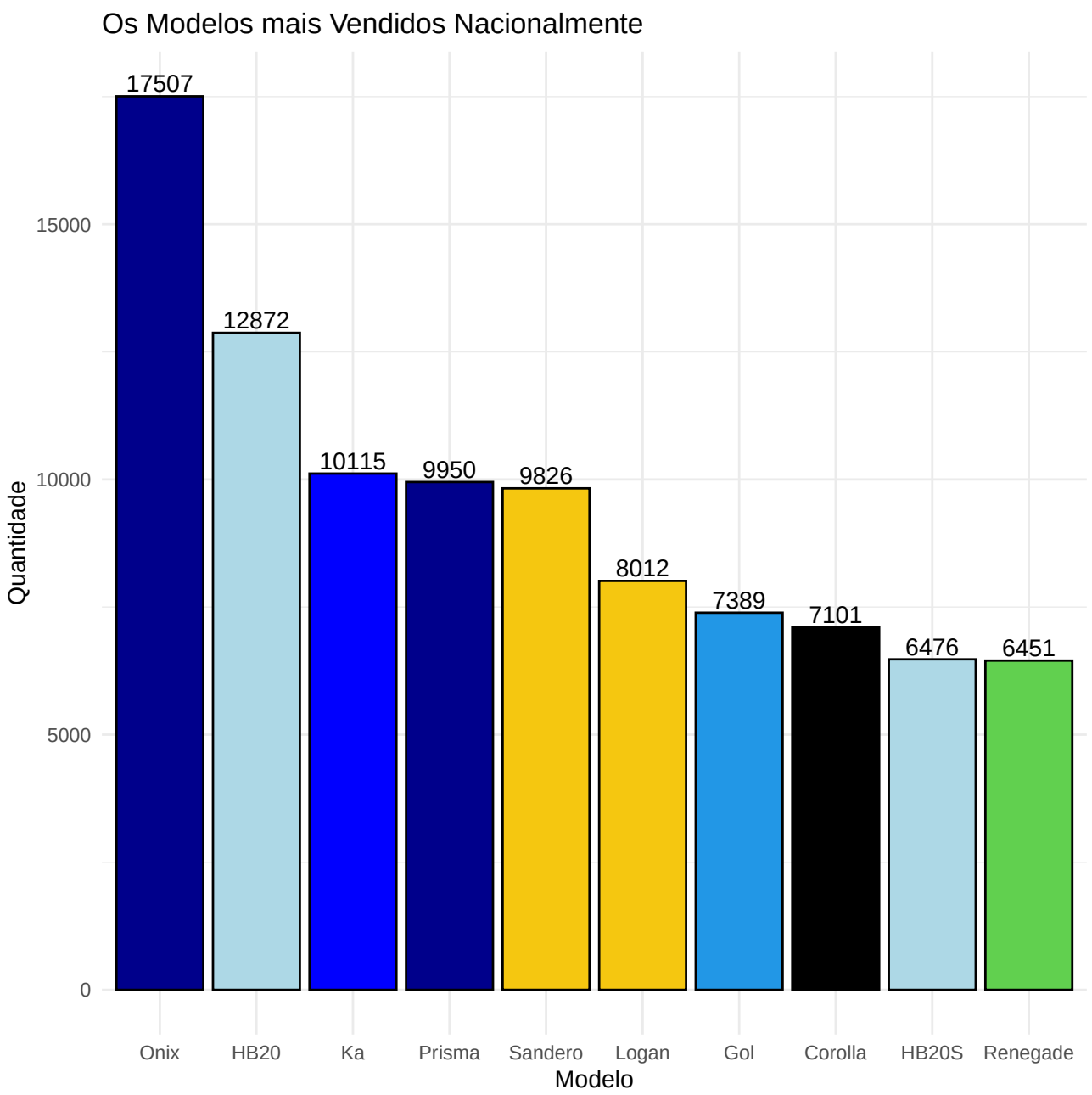


Chevrolet, Volkswagen, Fiat, Renault, Ford, Hyundai, Honda, Toyota, Jeep e Nissan, respectivamente.

Os modelos mais vendidos nacionalmente são

```
#### Nacional -  Modelo ####
count_data <- dds %>%
  count(modelo)

ggplot(count_data[count_data$n > 6000,],
       aes(x = reorder(modelo,-n), y = n )) +

  geom_bar(stat = 'identity',
           fill = c('black', '073980',
                    'light blue','light blue',
                    'blue', '7f935b',
```

```
                    'dark blue','dark blue',
                    '100403','7f935b'),color = 'black') +

  geom_text(aes(label = n), vjust = -.25) +
  labs(x = 'Modelo', y = 'Quantidade',
       title = 'Os Modelos mais Vendidos Nacionalmente') +

   theme_minimal()
```



Onix, HB20, Ka, Prisma, Sandero, Logan, Gol, Corolla, HB20S, Renegade.

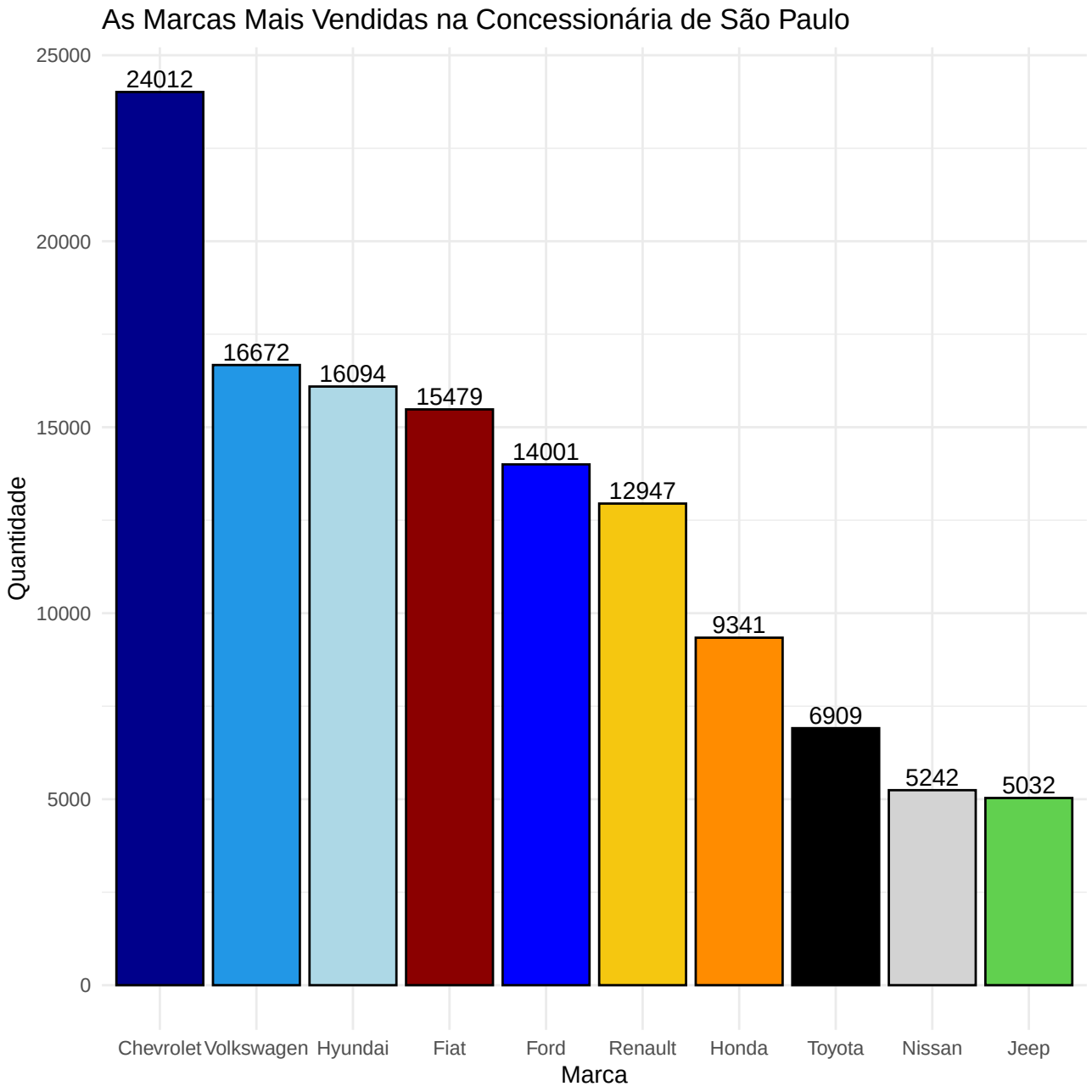
Perceba que mantivemos as cores das marcas em cada modelo, podemos ver que Onix e Prisma são modelos da Chevrolet, enquanto que Sandero e Logan são da Renault.

Partiremos para a segunda parte. Vereficar quais modelos e marcas mais presentes nas regiões de São Paulo, Distrito Federal, Bahia e Rio Grande do Sul.

São Paulo

As marcas mais presentes em São Paulo são

```
#### São Paulo - Marca ####
sp <- dds[dds$estado == 'SP',]



count_data <- sp %>%
  count(marca)

    #Plotando
ggplot(count_data[count_data$n > 5000,],
       aes(x = reorder(marca,-n), y = n )) +

  geom_bar(stat = 'identity',
           fill = c('dark blue','dark red',
                    'blue','dark orange',
                    'light blue', '100403',
                    'light gray','7f935b',
                    'black','073980'),color = 'black') +

  geom_text(aes(label = n), vjust = -.25) +
  labs(x = 'Marca', y = 'Quantidade',
       title = 'As Marcas Mais Vendidas na Concessionária de São Paulo') +

  theme_minimal()
```

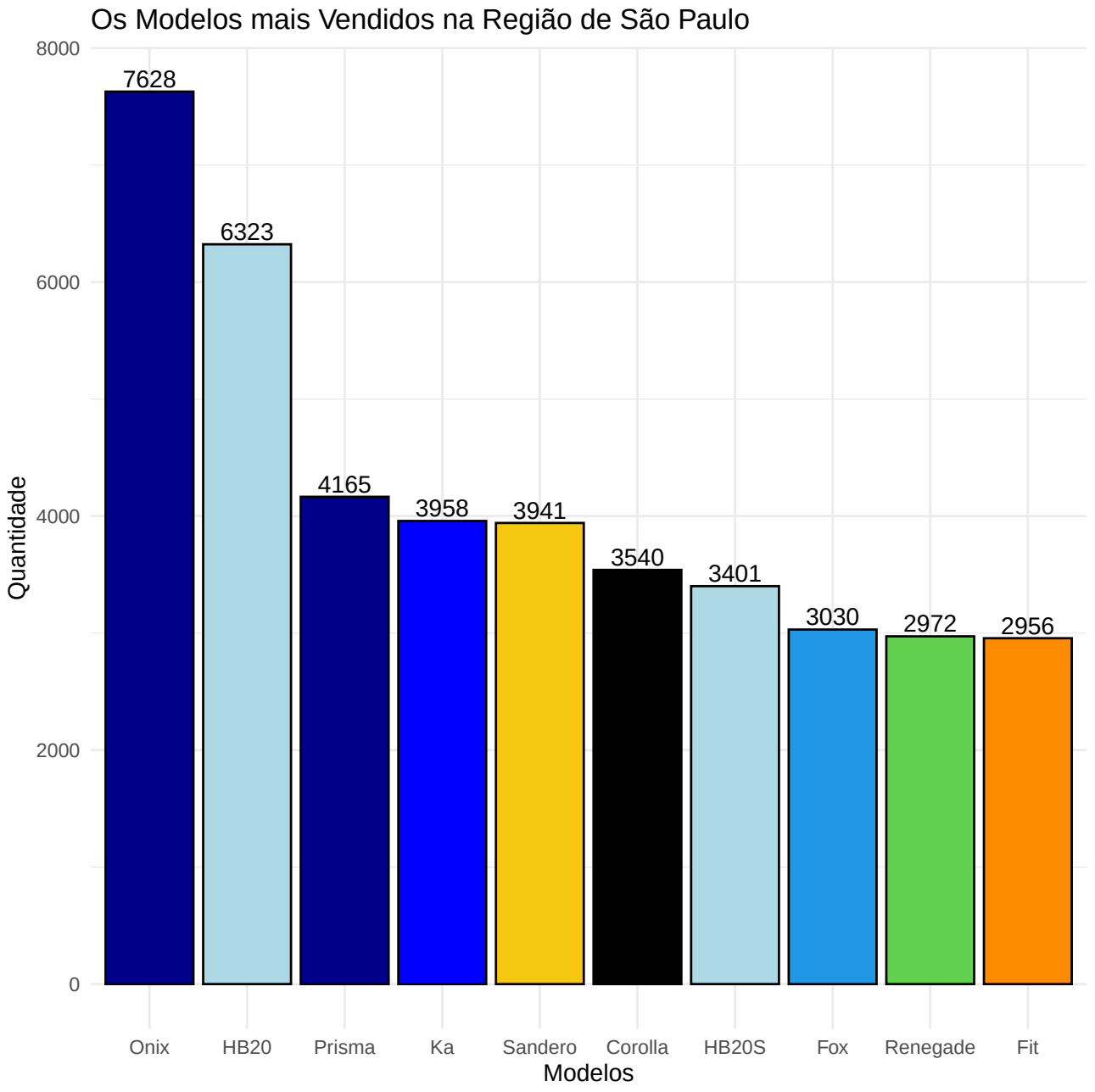


Volkswagen, Hyundai, Fiat, Ford, Renault, Honda, Toyota, Nissan e Jeep

Os Modelos mais vendidos em São Paulo são

```
sp <- dds[dds$estado == 'SP',]


count_data <- sp %>%
  count(modelo)


#plotando
ggplot(count_data[count_data$n > 2800,],
       aes(x = reorder(modelo,-n), y = n )) +
```

```
geom_bar(stat = 'identity',
         fill = c('black', 'dark orange',
                  '073980', 'light blue',
                  'light blue', 'blue',
                  'dark blue', 'dark blue',
                  '100403', '7f935b'),


         color = 'black') +

geom_text(aes(label = n), vjust = -.25) +
labs(x = 'Modelos', y = 'Quantidade',
     title = 'Os Modelos mais Vendidos na Região de São Paulo') +

theme_minimal()
```

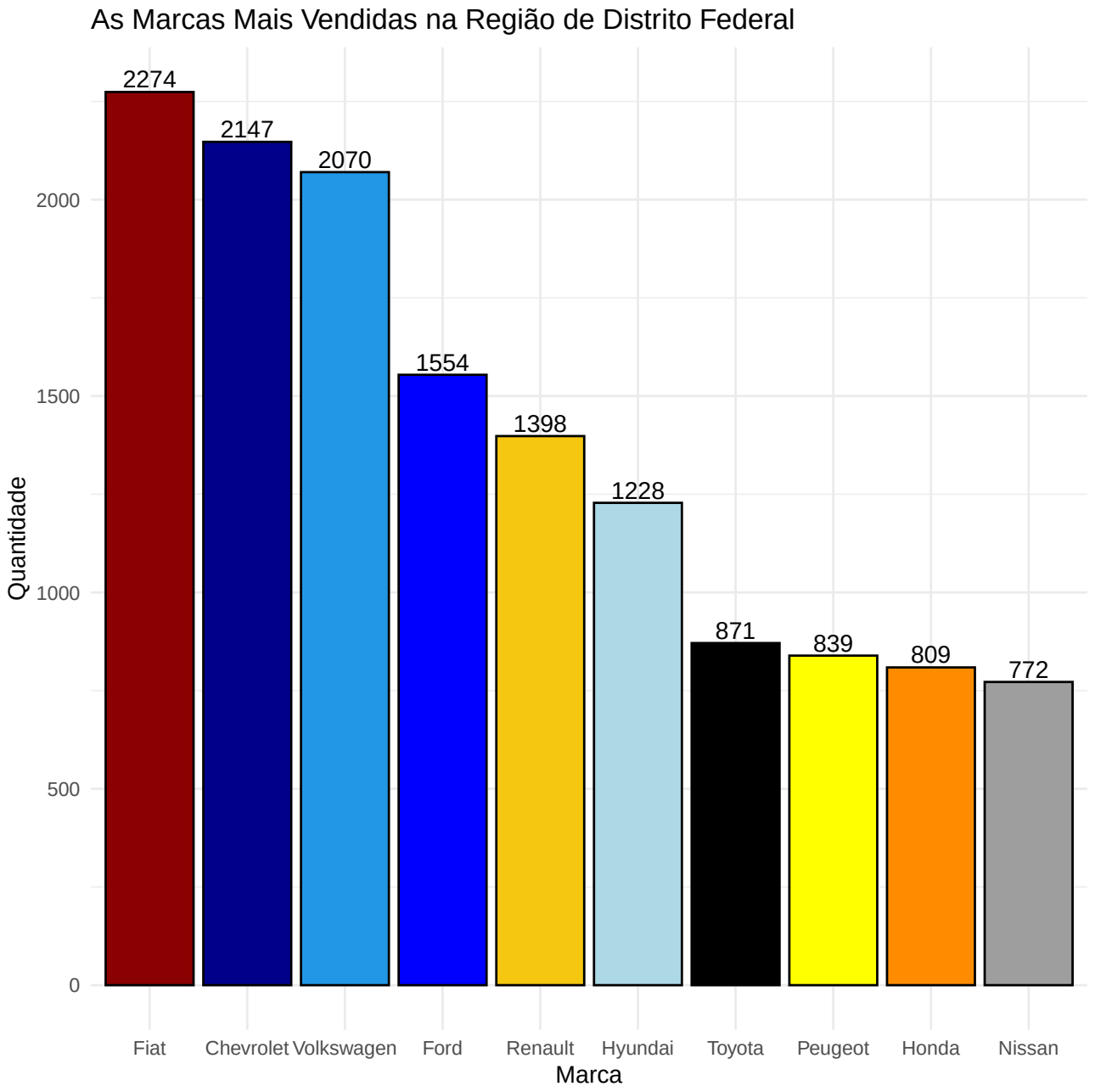


Onix, HB20, Prisma, Ka, Sandero, Corolla, HB20S, Fox, Renegade e Fit, respectivamete.

Distrito Federal

As marcas mais presentes em Distrito Federal são

```
#### Distrito Federal - Marca ####

df <- dds[dds$estado == 'DF',]


count_data <- df %>%
  count(marca)

    #Plotando
ggplot(count_data[count_data$n > 600,],
       aes(x = reorder(marca,-n), y = n )) +

  geom_bar(stat = 'identity',
           fill = c('dark blue','dark red',
                    'blue','dark orange',
                    'light blue','98856d',
                    'yellow','7f935b',
                    'black','073980'),color = 'black') +

  geom_text(aes(label = n), vjust = -.25) +
  labs(x = 'Marca', y = 'Quantidade', title = 'As Marcas Mais Vendidas na Região de Dist
  theme_minimal()
```

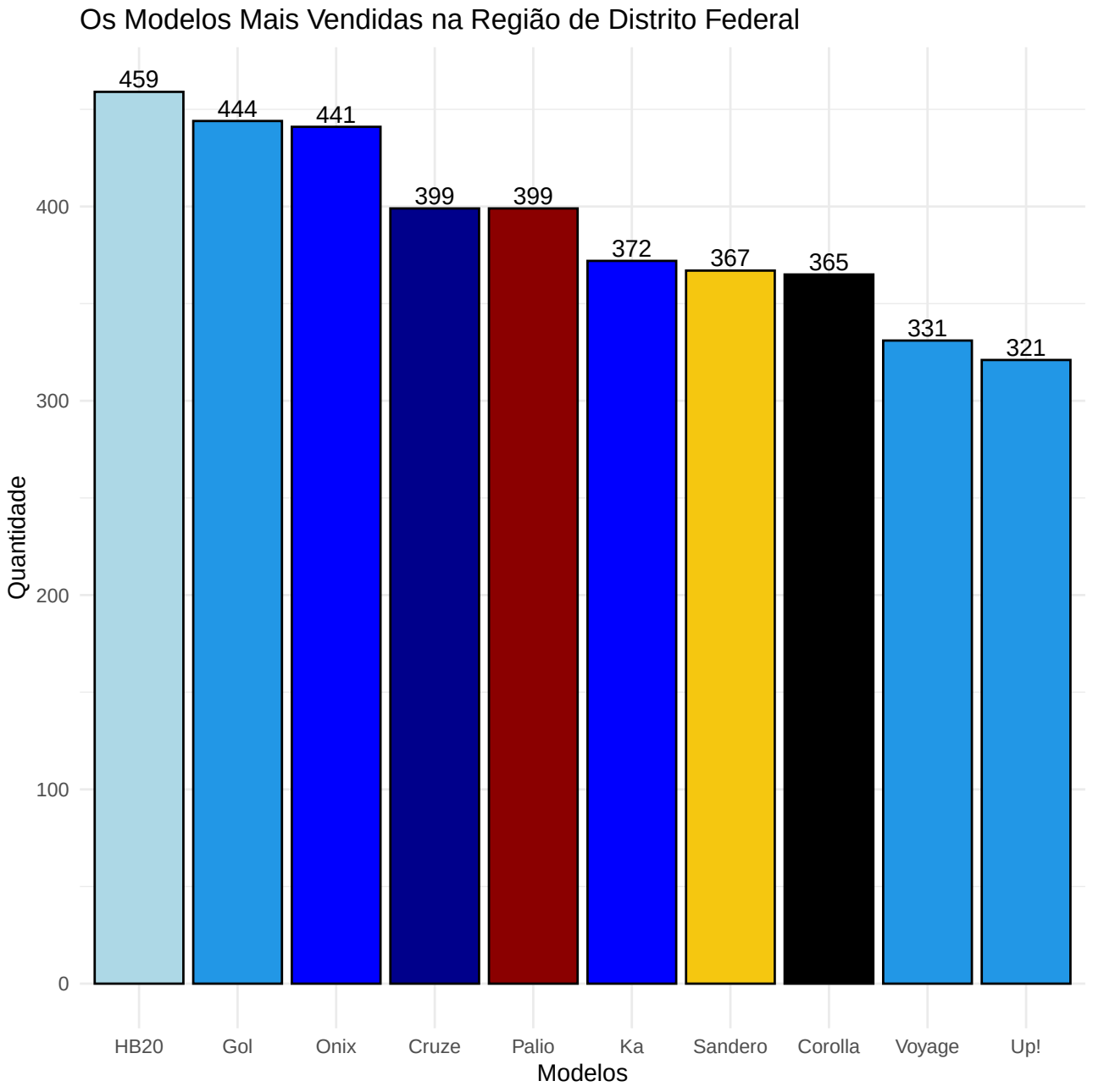


Fiat, Chevrolet, Volkswagen, Ford, Renault, Hyundai, Toyota, Pegeout, Honda e Nissan, respectivamente.

Os modelos mais vendidos no Distrito Federal são

```
#### Distrito Federal - Modelo ####
df <- dds[dds$estado == 'DF',]


count_data <- df %>%
  count(modelo)

ggplot(count_data[count_data$n > 310,],
       aes(x = reorder(modelo,-n), y = n )) +

  geom_bar(stat = 'identity',
```

```
          fill = c('black', 'dark blue',
                   '073980', 'light blue',
                   'blue', 'blue',
                   'dark red', '7f935b',
                   '073980', '073980'),
          color = 'black') +

geom_text(aes(label = n), vjust = -.25) +
labs(x = 'Modelos', y = 'Quantidade',
     title = 'Os Modelos Mais Vendidas na Região de Distrito Federal') +

theme_minimal()
```

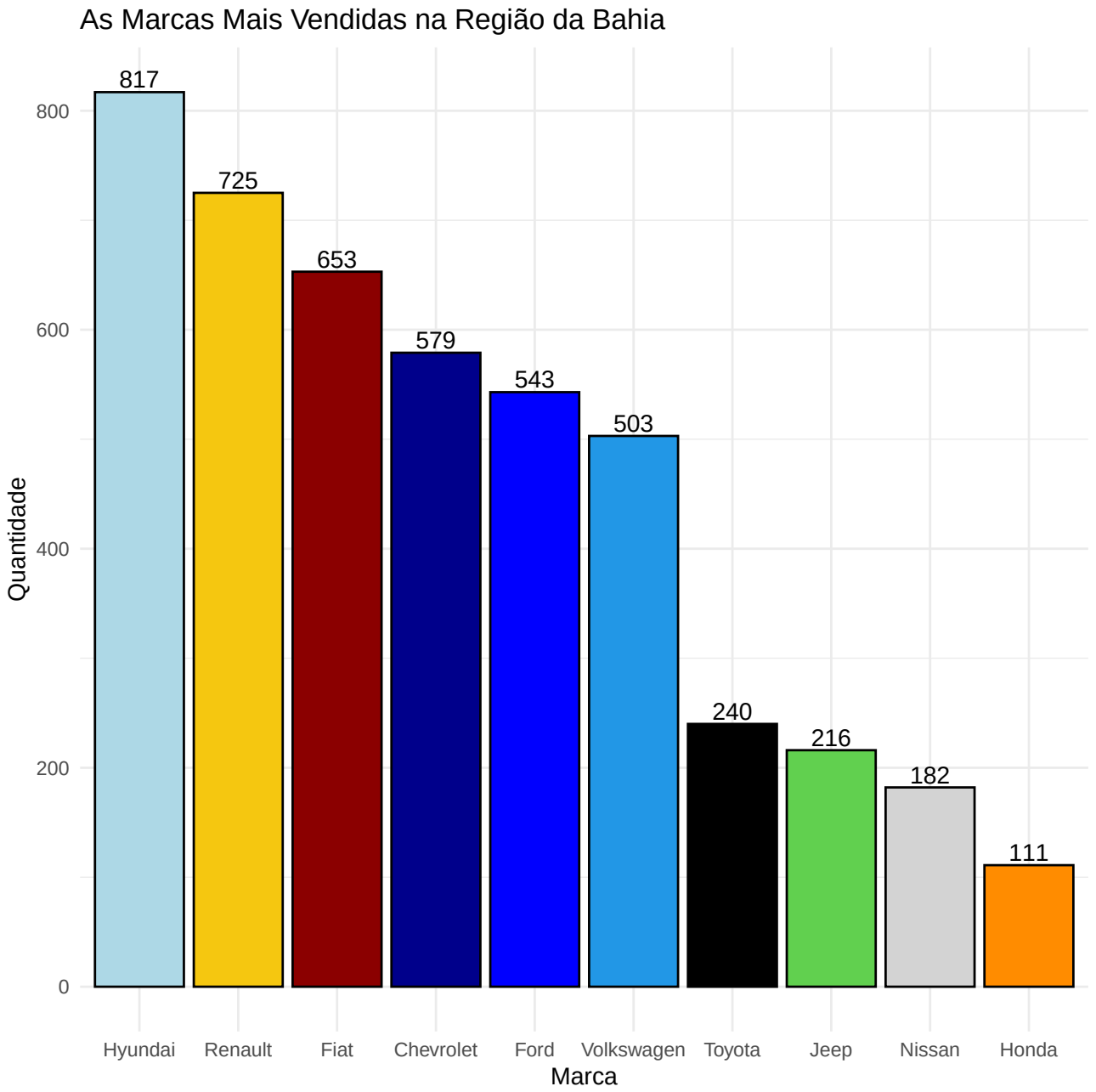


HB20, Gol, Onix, Cruze, Palio, Ka, Sandero, Corolla, Voyage e UP!, respectivamente.

Bahia

As marcas mais vendidos na Bahia são

```
#### Bahia - Marca ####
ba <- dds[dds$estado == 'BA',]


count_data <- ba %>%
  count(marca)

ggplot(count_data[count_data$n > 100,],
       aes(x = reorder(marca,-n), y = n )) +

  geom_bar(stat = 'identity',
           fill = c('dark blue','dark red',
                    'blue','dark orange',
                    'light blue', '100403',
                    'light gray','7f935b',
                    'black','073980'), color = 'black') +


  geom_text(aes(label = n), vjust = -.25) +

  labs(x = 'Marca', y = 'Quantidade',
       title = 'As Marcas Mais Vendidas na Região da Bahia') +

  theme_minimal()
```

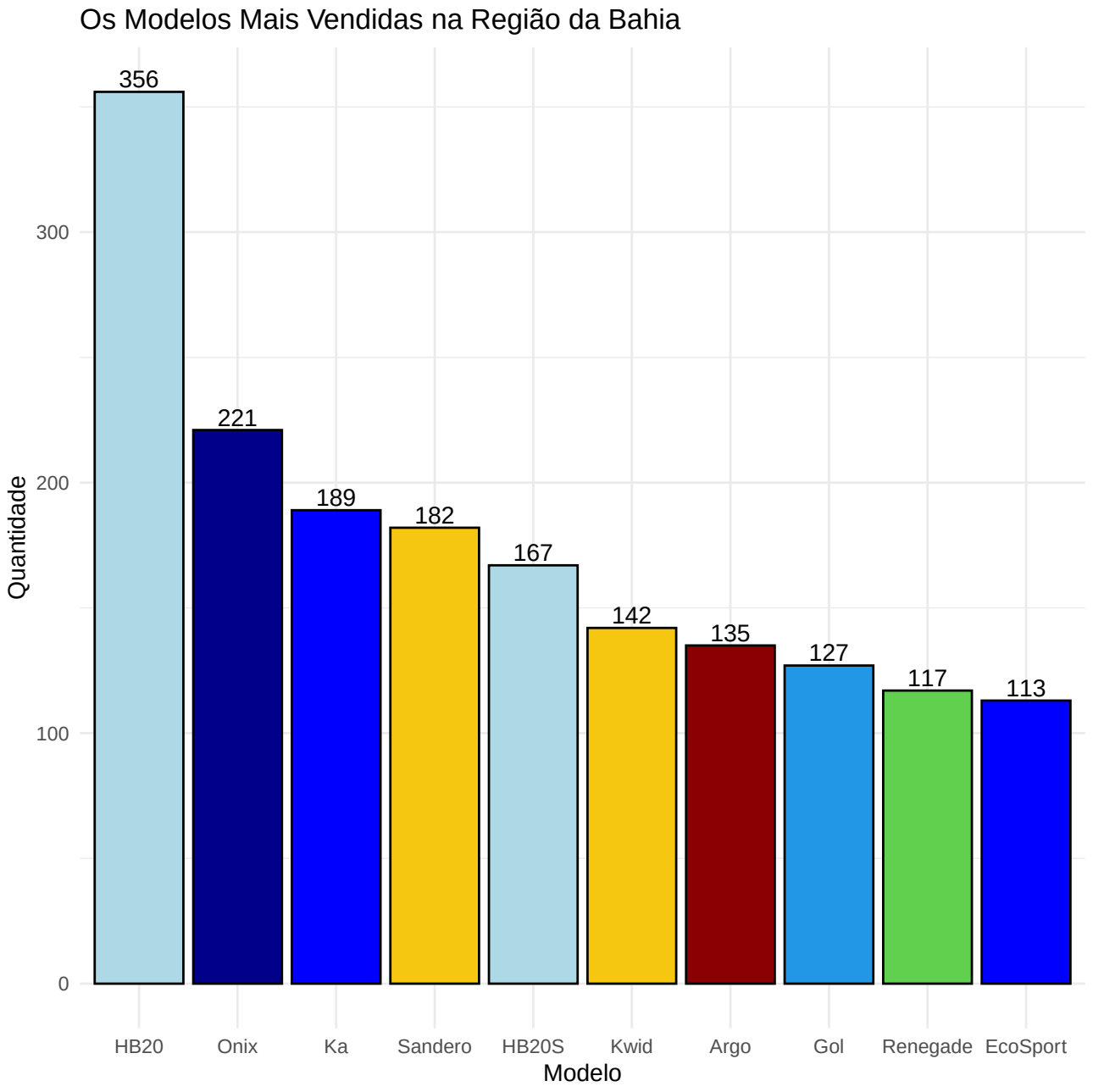


Hyundai, Renault, Fiat, Chevrolet, Ford, Volkswagen, Toyota, Jeep, Nissan e Honda, respectivamente.

O modelo mais vendidos são

```
#### Bahia- Modelo ####
ba <- dds[dds$estado == 'BA',]


count_data <- ba %>%
  count(modelo)

ggplot(count_data[count_data$n > 110,],
       aes(x = reorder(modelo,-n), y = n )) +

  geom_bar(stat = 'identity',
```

```
          fill = c('dark red', 'blue',
                   '073980', 'light blue',
                   'light blue', 'blue',
                   '7f935b', 'dark blue',
                   '100403', '7f935b'),
          color = 'black') +

geom_text(aes(label = n), vjust = -.25) +
labs(x = 'Modelo', y = 'Quantidade',
     title = 'Os Modelos Mais Vendidas na Região da Bahia') +

theme_minimal()
```

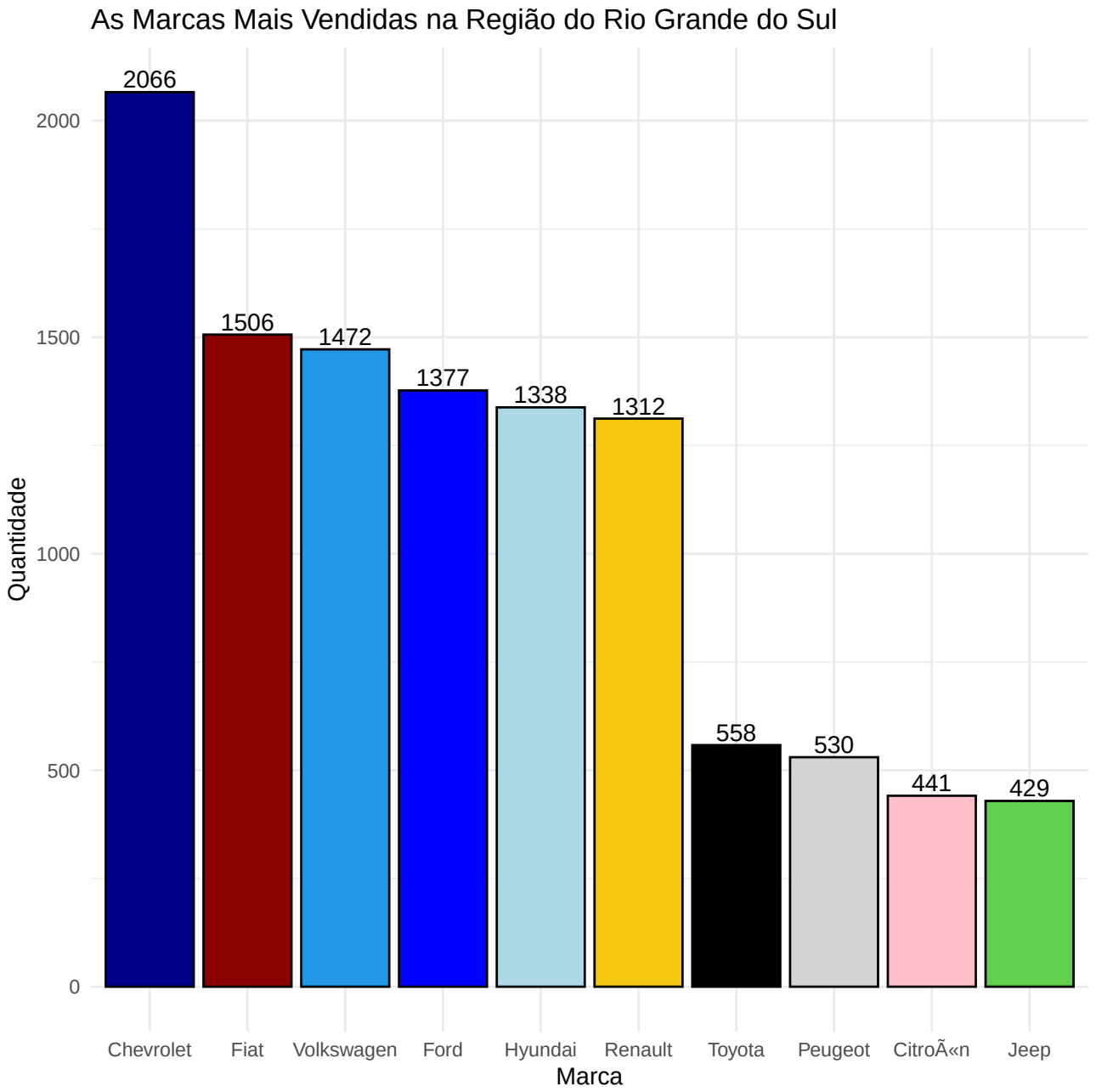


HB20, Onix, Ka, Sandero, HB20S, Kwid, Argo, Gol, Renegade e EcoSport, respectivamente.

Rio Grande do Sul

As marcas mais vendidos no Rio Grande do Sul são

```
#### Rio Grande do Sul - Marca ####


rs <- dds[dds$estado == 'RS',]


count_data <- rs %>%
  count(marca)

      #Plotando

ggplot(count_data[count_data$n > 360,],
       aes(x = reorder(marca,-n), y = n )) +

  geom_bar(stat = 'identity',
           fill = c('dark blue','pink',
                    'dark red','blue',
                    'light blue','100403',
                    'light gray','7f935b',
                    'black','073980'),color = 'black')+

  geom_text(aes(label = n), vjust = -.25) +
  labs(x = 'Marca', y = 'Quantidade',
       title = 'As Marcas Mais Vendidas na Região do Rio Grande do Sul') +

  theme_minimal()
```

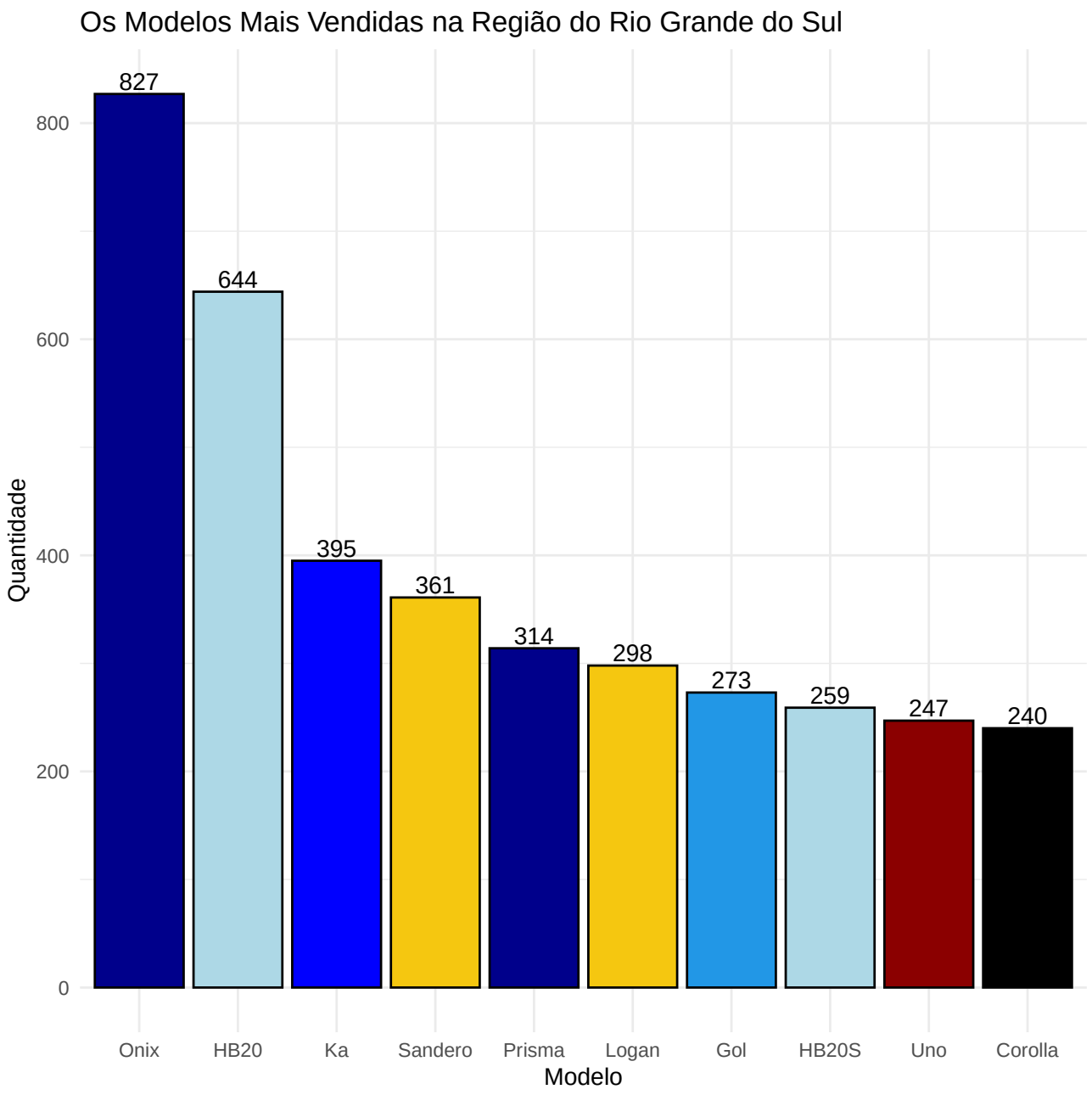


Chevrolet, Fiat, Volkswagen, Ford, Hyundai, Renault, Toyota, Pegeout, Citroen e Jeep, respectivamente.

Os modelos mais presentes no Rio Grande do Sul são

```
#### Rio Grande do Sul - Modelo ####
rs <- dds[dds$estado == 'RS',]


count_data <- rs %>%
  count(modelo)

ggplot(count_data[count_data$n > 220,],
       aes(x = reorder(modelo,-n), y = n )) +

  geom_bar(stat = 'identity',
```

```
          fill = c('black', '073980',
                   'light blue', 'light blue',
                   'blue', '7f935b',
                   'dark blue', 'dark blue',
                   '7f935b', 'dark red'),
          color = 'black') +

geom_text(aes(label = n), vjust = -.25) +
labs(x = 'Modelo', y = 'Quantidade',
     title = 'Os Modelos Mais Vendidas na Região do Rio Grande do Sul') +

theme_minimal()
```



Onix, HB20, Ka, Sandero, Prisma, Logan, Gol, HB20S, Uno e Corolla, respectivamente.

A seguir responderei o cliente por estados nas quais possuí concessionária. Na ordem de SP, DF, BA, DF.

2) Olhando para os modelos mais predominantes (maior presença) na minha região:

a) Algum deles apresenta uma maior desvalorização? Ou uma maior valorização?

b) Existe uma desvalorização do carro baseadona quilometragem dele?

```
#### 2A ####

#verificando a variável data
#summary(dds$data)

dds$data <- as.character(dds$data)

#Modificando a Variável data nos dias de venda para o mês que foi vendido

dds[dds$data %in% c('2020-10-01','2020-10-02','2020-10-03',
                    '2020-10-04','2020-10-05','2020-10-06',
                    '2020-10-07','2020-10-08','2020-10-09',
                    '2020-10-10','2020-10-11','2020-10-12',
                    '2020-10-13','2020-10-14','2020-10-15',
                    '2020-10-16','2020-10-17','2020-10-18',
                    '2020-10-19','2020-10-20','2020-10-21',
                    '2020-10-22','2020-10-23','2020-10-24',
                    '2020-10-25','2020-10-26','2020-10-27',
                    '2020-10-28','2020-10-29', '2020-10-30',
                    '2020-10-31') ,]$data = 'Outubro'


dds[dds$data %in% c('2020-11-01','2020-11-02','2020-11-03',
                    '2020-11-04','2020-11-05','2020-11-06',
                    '2020-11-07','2020-11-08','2020-11-09',
                    '2020-11-10','2020-11-11','2020-11-12',
                    '2020-11-13','2020-11-14','2020-11-15',
                    '2020-11-16','2020-11-17','2020-11-18',
                    '2020-11-19','2020-11-20','2020-11-21',
                    '2020-11-22','2020-11-23','2020-11-24',
                    '2020-11-25','2020-11-26','2020-11-27',
                    '2020-11-28','2020-11-29','2020-11-30'
                    ) ,]$data = 'Novembro'


dds[dds$data %in% c('2020-12-01','2020-12-02','2020-12-03',
                    '2020-12-04','2020-12-05','2020-12-06',
                    '2020-12-07','2020-12-08','2020-12-09',
                    '2020-12-10','2020-12-11','2020-12-12',
                    '2020-12-13','2020-12-14','2020-12-15',
                    '2020-12-16','2020-12-17','2020-12-18',
                    '2020-12-19','2020-12-20','2020-12-21',
                    '2020-12-22','2020-12-23','2020-12-24',
                    '2020-12-25','2020-12-26','2020-12-27',
                    '2020-12-28','2020-12-29','2020-12-30',
                    '2020-12-31') ,]$data = 'Dezembro'
```

São Paulo

Selecionando os modelos mais vendidos no Estado de SP.

```
#### 2B #####
sp <- dds[dds$estado == 'SP' &
            dds$modelo %in% c('Onix','HB20','Prisma',
                              'Ka', 'Sandero'),]
```

O sumário foi usado para filtrar as variável preço no passo seguinte, pois filtraremos um valor um pouco acima do terceiro quartil para evitarmos um gráfico de dispersão pouco informativo.

```
summary(sp$km)
```

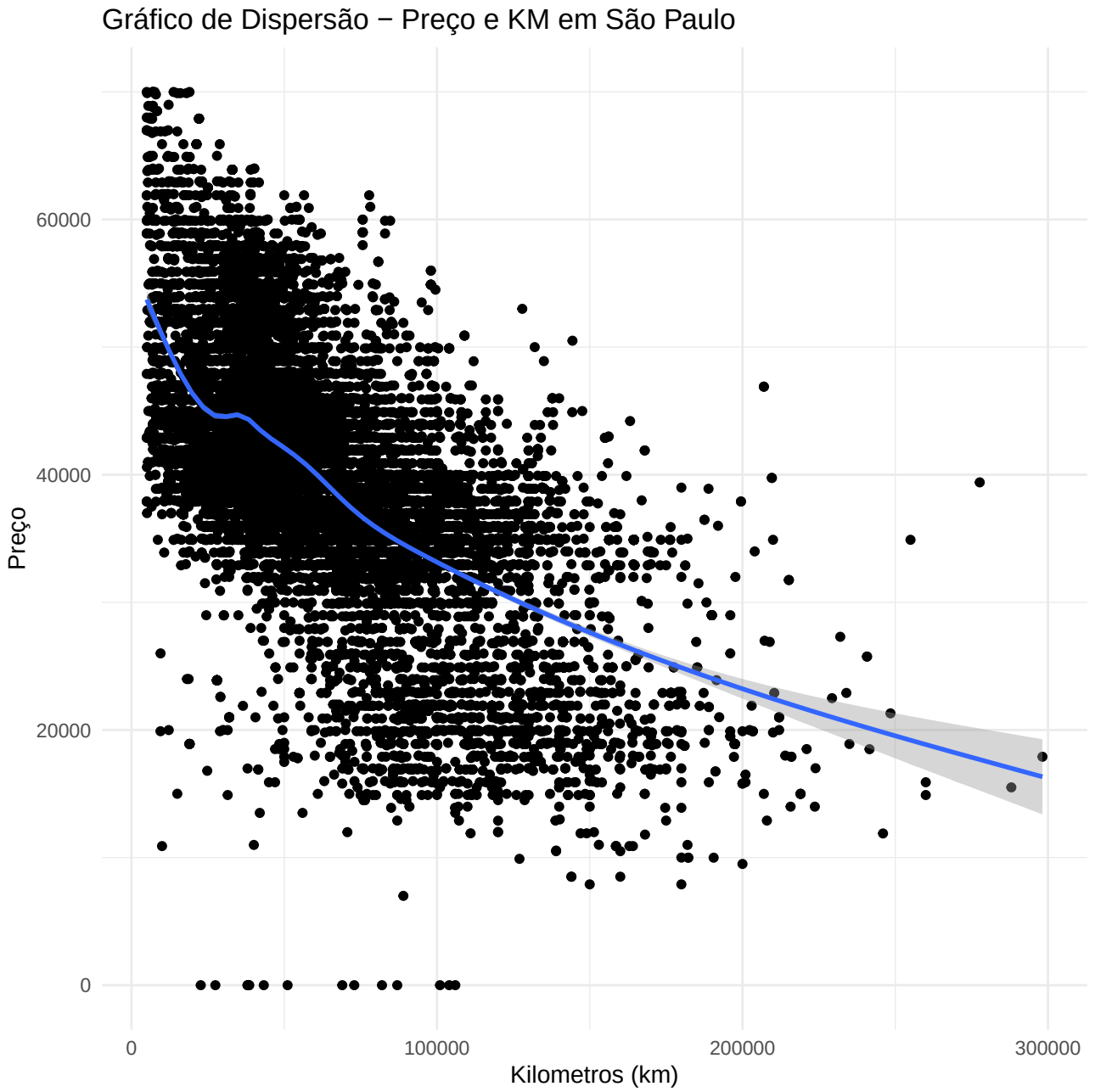
```
##                Min.             1st Qu.              Median                Mean
##                5000               37900               45864      42710402177508
##             3rd Qu.                Max.
##               70489 1111111111111111168
```

```
summary(sp$preco)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##        0    36990    41500    42690    44900 46990000
```

```
sp %>%
  filter(preco < 70000 & km < 300000) %>%
  ggplot(aes(x = km, y = preco))+
  labs(x = 'Kilometros (km)', y =
         'Preço', title = 'Gráfico de Dispersão - Preço e KM em São Paulo')+
  geom_point() +
  geom_smooth() +
  theme_minimal()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



Graficamente vemos que há uma tendência de quanto maior a quilometragem do carro menor o valor de mercado ele terá, e pelo comportamento do gráfico usaremos correlação de Spearman.

```
cor <- cor.test(sp$km, sp$preco, method="spearman")
```

```
## Warning in cor.test.default(sp$km, sp$preco, method = "spearman"): Cannot
## compute exact p-value with ties
```

```
cor
```

```
##
##  Spearman's rank correlation rho
##
## data:  sp$km and sp$preco
## S = 4686672891902, p-value < 0.00000000000000022
```

```
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.5971452
```

Temos evidências de que há correlação entre as variáveis e pela saída do R temos que a correlação corresponde a conclusão dada no gráfico de dispersão, ou seja, quanto mais rodado o carro menor o seu valor de venda.

Os modelos mais vendidos de SP apresenta alguma desvalorizaçao?

```
kruskal.test(sp$preco, sp$data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  sp$preco and sp$data
## Kruskal-Wallis chi-squared = 372.38, df = 2, p-value <
## 0.00000000000000022
```

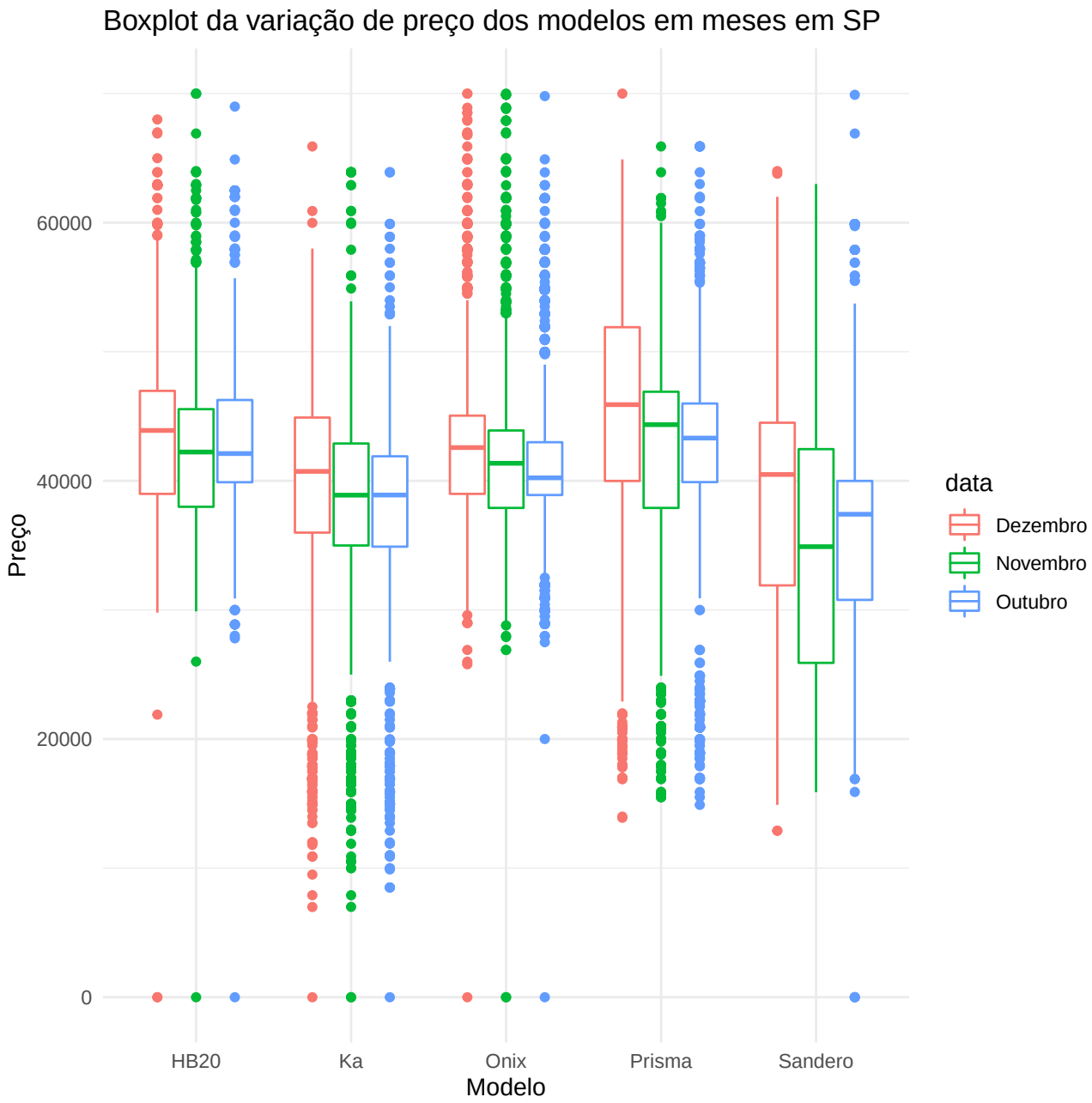
Há diferença dos preços de venda em relação aos meses.

```
dunn.test(sp$preco, sp$data,method="holm")
```

```
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 372.3825, df = 2, p-value = 0
##
##
##                             Comparison of x by group
##                                      (Holm)
## Col Mean-|
## Row Mean |   Dezembro   Novembro
## ---------+----------------------
## Novembro |   15.51727
##          |    0.0000*
##          |
##  Outubro |   18.14495   1.810489
##          |    0.0000*     0.0351
##
## alpha = 0.05
## Reject Ho if p <= alpha/2
```

Não existe diferença de preços entre os meses de Novembro e Outubro. Há diferença do mes de Dezembro aos demais.

```r
sp %>%
  filter(preco < 70000) %>%
  ggplot(aes(x = modelo,y = preco, color = data))+
  labs(x = 'Modelo', y = 'Preço', title = 'Boxplot da variação de preço dos modelos em m
  geom_boxplot() +
  theme_minimal()
```



Ford Ka, Onix e Prisma aparentam ter uma desvalorização ao longo dos meses - de outubro à dezembro. HB20 desvaloriza de outubro a novembro. E o Renault Sandero desvaloriza de outubro a novembro e em dezembro valoriza mas não chega no mesmo patamar de valor comparada em Outubro.

Distrito Federal

Selecionando os carros mais vendidos no Distrito Federal.

```r
df <- dds[dds$estado == 'DF' & dds$modelo %in% c('HB20','Gol','Onix',
                                                 'Cruze','Palio'),]
```

```r
summary(df$preco)
```

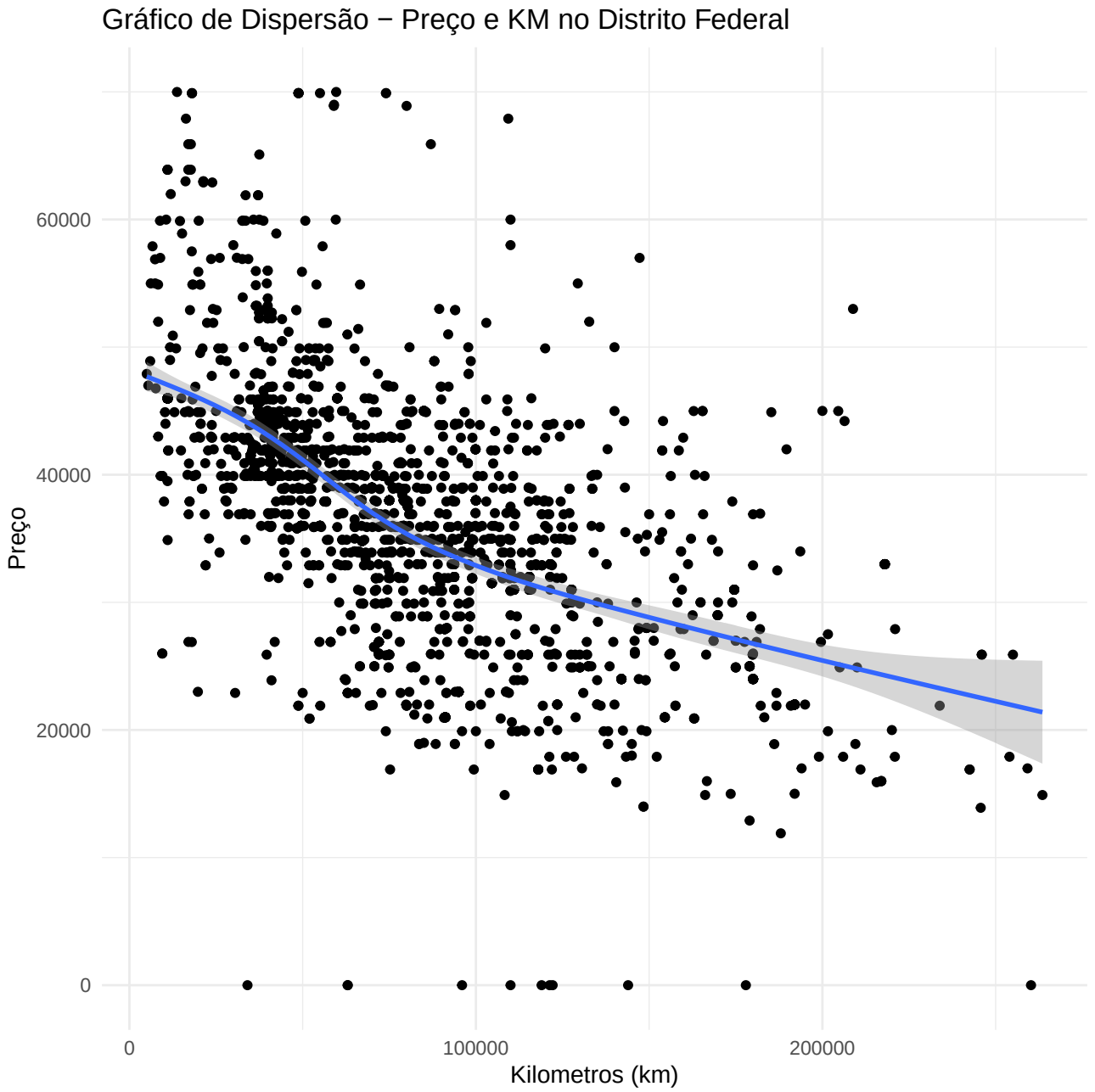
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0   31923   38900   39416   45900   95900
```

```r
summary(df$km)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5000   38000   70000   75918  102387  696029
```

```r
df %>%
  filter(preco < 70000 & km < 300000) %>%
  ggplot(aes(x = km, y = preco))+
  labs(x = 'Kilometros (km)', y =
         'Preço', title = 'Gráfico de Dispersão - Preço e KM no Distrito Federal')+
  geom_point() +
  geom_smooth() +
  theme_minimal()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



Graficamente vemos que há uma tendência de quanto maior a quilometragem do carro menor o valor de mercado ele terá, e pelo comportamento do gráfico usaremos correlação de Spearman.

```
cor <- cor.test(df$km, df$preco, method="spearman")
```

```
## Warning in cor.test.default(df$km, df$preco, method = "spearman"): Cannot
## compute exact p-value with ties
```

```
cor
```

```
##
##  Spearman's rank correlation rho
##
## data:  df$km and df$preco
## S = 2802719660, p-value < 0.00000000000000022
```

```
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## -0.711089
```

Temos evidências de que há correlação entre as variáveis e pela saída do R temos que a correlação corresponde a conclusão dada no gráfico de dispersão, ou seja, quanto mais rodado o carro menor o seu valor de venda.

Os modelos mais vendidos de DF apresenta alguma desvalorizaçao?

```
kruskal.test(df$preco, df$data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  df$preco and df$data
## Kruskal-Wallis chi-squared = 7.4354, df = 2, p-value = 0.02429
```

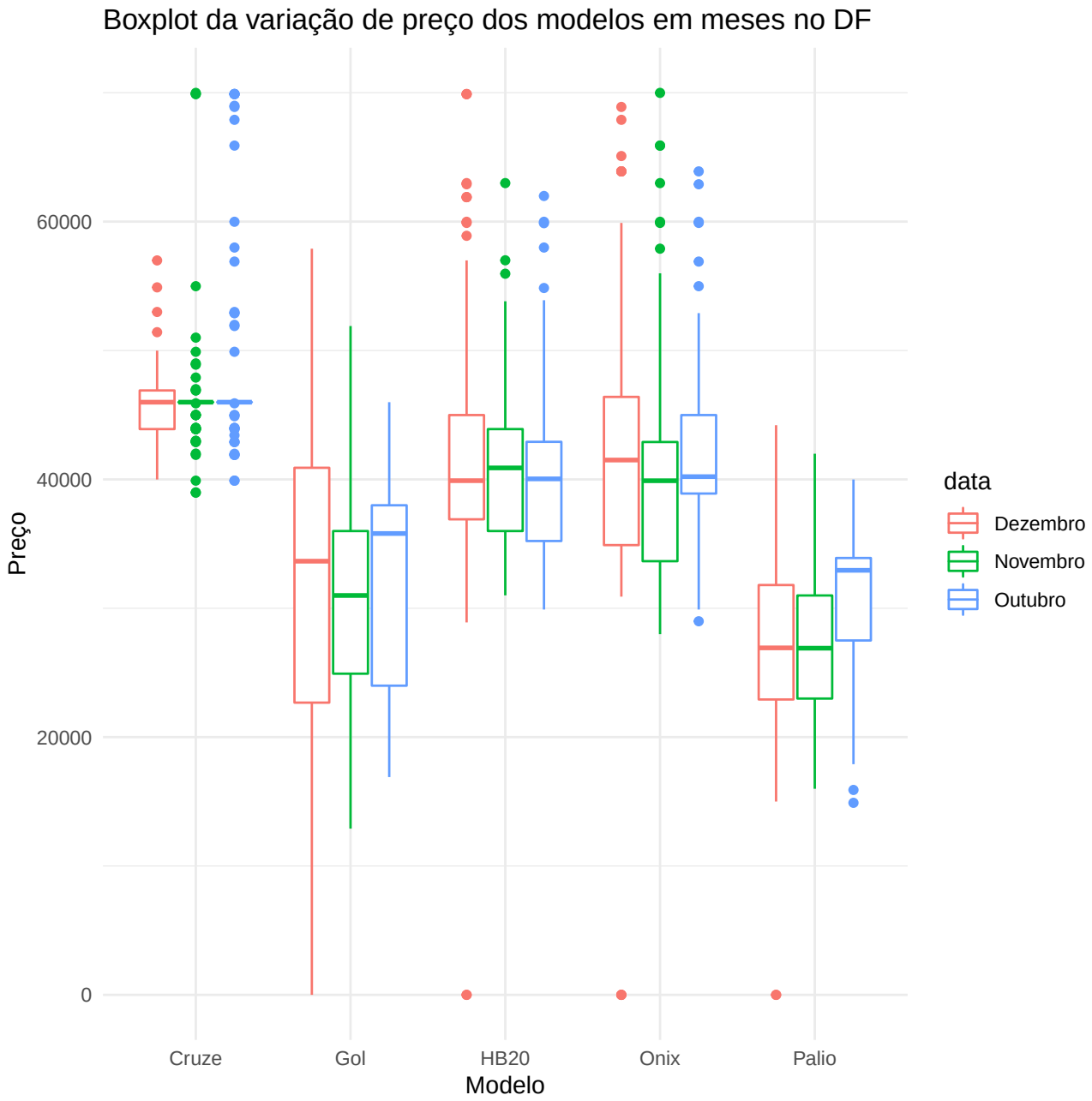
Há diferença dos preços de venda em relação aos meses.

```
dunn.test(df$preco, df$data,method="holm")
```

```
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 7.4354, df = 2, p-value = 0.02
##
##
##                             Comparison of x by group
##                                      (Holm)
## Col Mean-|
## Row Mean |   Dezembro   Novembro
## ---------+----------------------
## Novembro |  -2.608514
##          |    0.0136*
##          |
##  Outubro |  -2.066990   0.463233
##          |     0.0387     0.3216
##
## alpha = 0.05
## Reject Ho if p <= alpha/2
```

Não existe diferença de preços entre os meses de Novembro e Outubro. Há diferença do mes de Dezembro aos demais.

```
df %>%
  filter(preco < 70000) %>%
  ggplot(aes(x = modelo,y = preco, color = data))+
  labs(x = 'Modelo', y = 'Preço', title = 'Boxplot da variação de preço dos modelos em m
  geom_boxplot() +
  theme_minimal()
```



Fiat Palio obtém valorização do preço no mês de Dezembro. VW Gol sua média de preço obtem o menor valor no mês de NOvembro e em Dezembro há uma valorização em seu preço.

Bahia

```
ba <- dds[dds$estado == 'BA' & dds$modelo %in% c('HB20','Onix','Ka',
                                                 'Sandero','HB20S'),]
```

```
summary(ba$km)
```

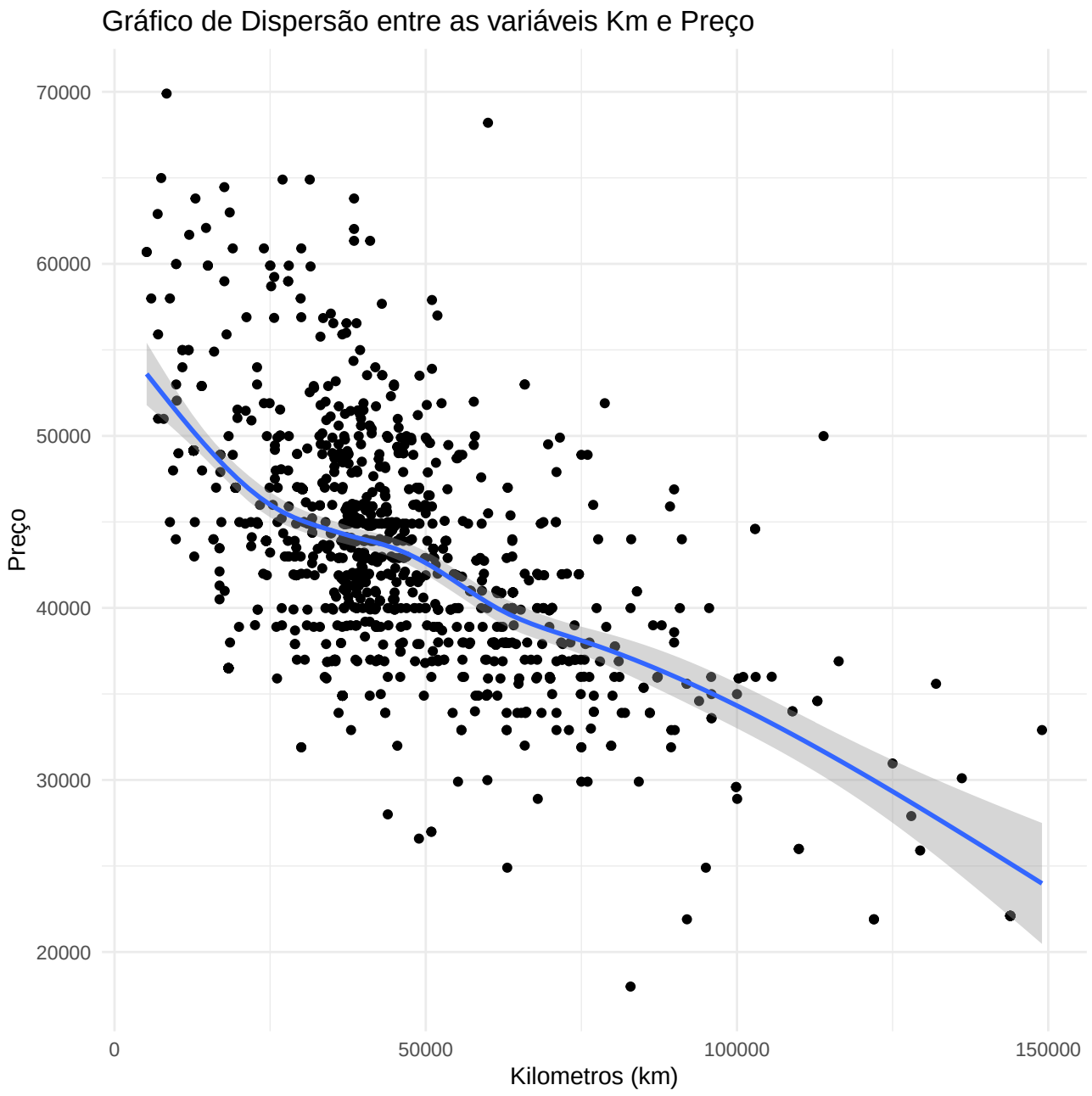
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5166   30458   41082   44801   55781  149000
```

```r
summary(ba$preco)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17990   38900   43210   43317   47550   72990
```

```r
ba %>%
  filter(preco < 70000 & km < 300000) %>%
  ggplot(aes(x = km, y = preco))+
  labs(x = 'Kilometros (km)', y =
         'Preço', title = 'Gráfico de Dispersão entre as variáveis Km e Preço')+
  geom_point() +
  geom_smooth() +
  theme_minimal()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



Graficamente vemos que há uma tendência de quanto maior a quilometragem do carro menor o valor de mercado ele terá, e pelo comportamento do gráfico usaremos correlação de Spearman.

```
cor <- cor.test(ba$km, ba$preco, method="spearman")
```

```
## Warning in cor.test.default(ba$km, ba$preco, method = "spearman"): Cannot
## compute exact p-value with ties
```

```
cor
```

```
##
##  Spearman's rank correlation rho
##
## data:  ba$km and ba$preco
## S = 356614004, p-value < 0.00000000000000022
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## -0.5435667
```

Temos evidências de que há correlação entre as variáveis e pela saída do R temos que a correlação corresponde a conclusão dada no gráfico de dispersão, ou seja, quanto mais rodado o carro menor o seu valor de venda.

Os modelos mais vendidos de BA apresenta alguma desvalorizaçao?

```
kruskal.test(ba$preco, ba$data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  ba$preco and ba$data
## Kruskal-Wallis chi-squared = 3.0676, df = 2, p-value = 0.2157
```

Não há diferença dos preços de venda em relação aos meses. Portanto não faz sentido analisar os gráficos boxplots buscando alguma variação dos preços ao longo dos meses.

Rio Grande do Sul

```
rs <- dds[dds$estado == 'RS' & dds$modelo %in% c('Onix','HB20','Ka',
                                                 'Sandero','Prisma'),]
```

```
summary(rs$preco)
```

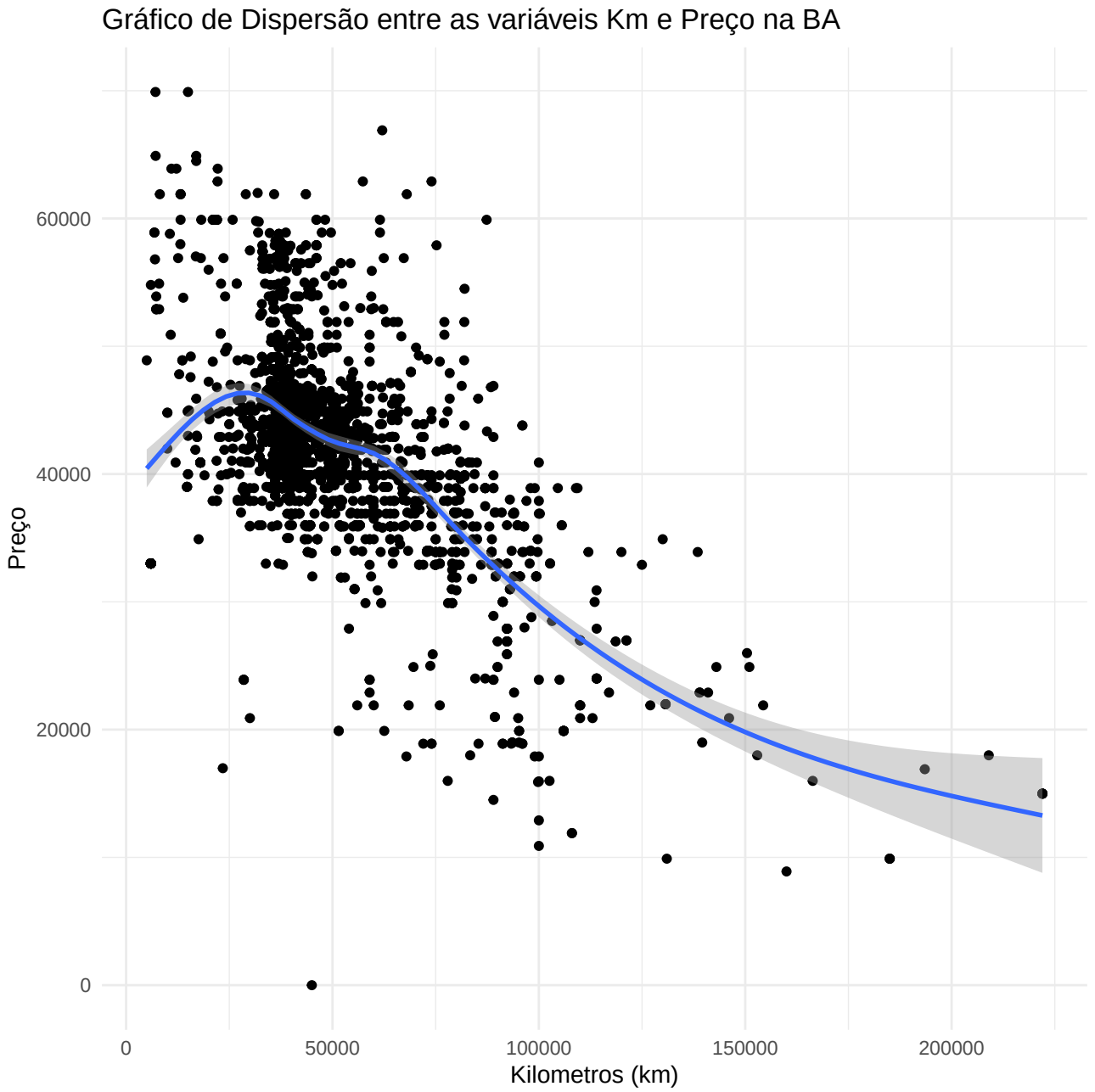
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0   38900   42500   42157   45900   79900
```

```r
summary(rs$km)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5000   37185   43087   57174   56169  999999
```

```r
rs %>%

  filter(preco < 70000 & km < 300000) %>%

  ggplot(aes(x = km, y = preco))+
  labs(x = 'Kilometros (km)', y =
         'Preço', title = 'Gráfico de Dispersão entre as variáveis Km e Preço na BA')+
  geom_point() +
  geom_smooth() +
  theme_minimal()
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



Graficamente vemos que há uma tendência de quanto maior a quilometragem do carro menor o valor de mercado ele terá, e pelo comportamento do gráfico usaremos correlação de Spearman.

```
cor <- cor.test(rs$km, rs$preco, method="spearman")
```

```
## Warning in cor.test.default(rs$km, rs$preco, method = "spearman"): Cannot
## compute exact p-value with ties
```

```
cor
```

```
##
##  Spearman's rank correlation rho
##
## data:  rs$km and rs$preco
## S = 3941371182, p-value < 0.00000000000000022
```

```
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.4414005
```

Temos evidências de que há correlação entre as variáveis e pela saída do R temos que a correlação corresponde a conclusão dada no gráfico de dispersão, ou seja, quanto mais rodado o carro menor o seu valor de venda.

Os modelos mais vendidos de RS apresenta alguma desvalorizaçao?

```
kruskal.test(rs$preco, rs$data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  rs$preco and rs$data
## Kruskal-Wallis chi-squared = 70.457, df = 2, p-value =
## 0.0000000000000005018
```

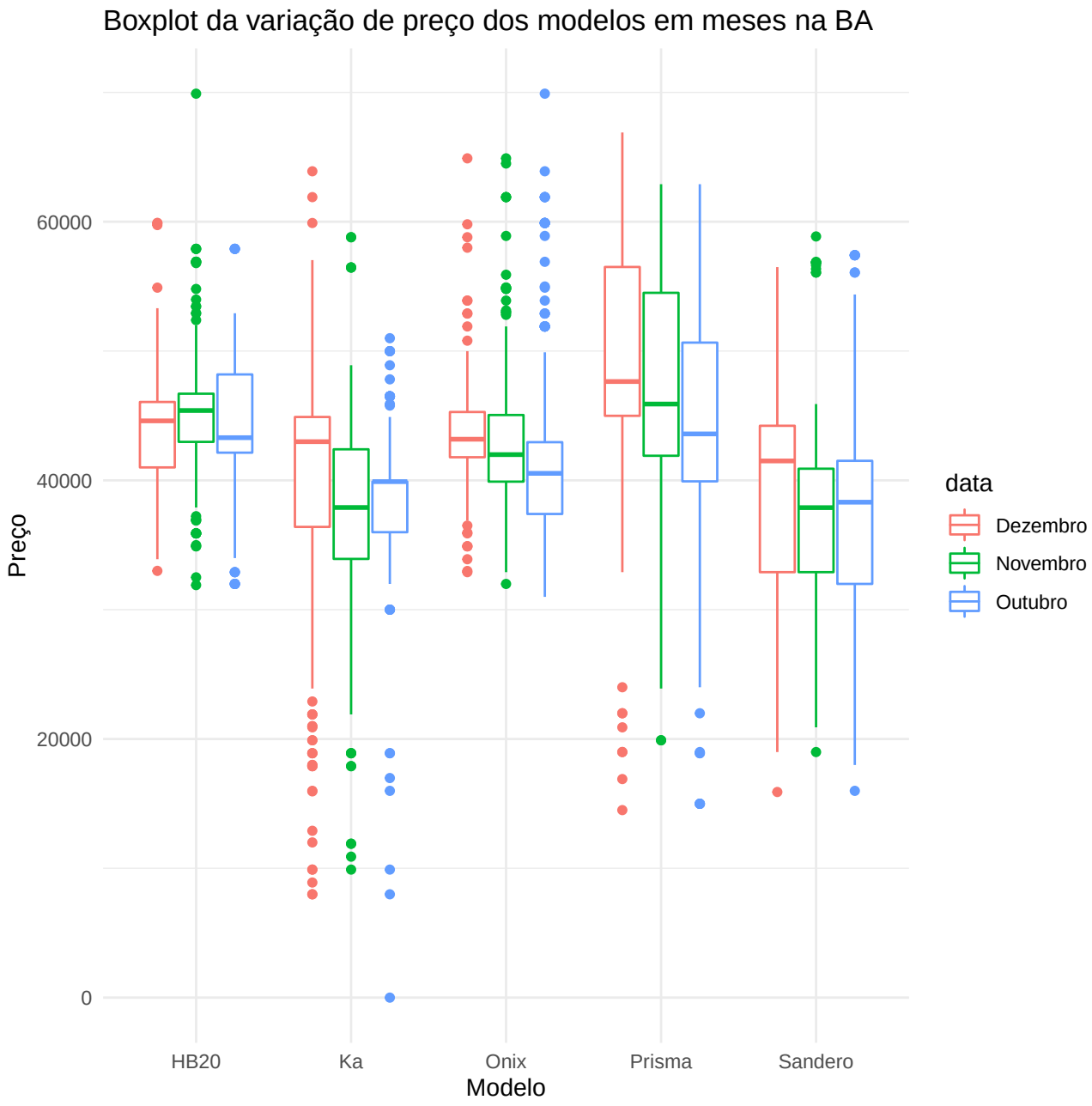
Há diferença dos preços de venda em relação aos meses.

```
dunn.test(rs$preco, rs$data,method="holm")
```

```
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 70.4565, df = 2, p-value = 0
##
##
##                            Comparison of x by group
##                                     (Holm)
## Col Mean-|
## Row Mean |   Dezembro   Novembro
## ---------+----------------------
## Novembro |   2.011144
##          |    0.0222*
##          |
##  Outubro |   8.025544   5.552099
##          |    0.0000*    0.0000*
##
## alpha = 0.05
## Reject Ho if p <= alpha/2
```

Há diferença de preços nos 3 meses.

```
rs %>%
  filter(preco < 70000) %>%
  ggplot(aes(x = modelo,y = preco, color = data))+
  labs(x = 'Modelo', y = 'Preço', title = 'Boxplot da variação de preço dos modelos em m
  geom_boxplot() +
  theme_minimal()
```



Há uma desvalorização ao longo dos meses os modelos Prisma e Onix.