Relatório Desafio Precificação

Rodolfo Viana

06-12-2015

Relatório Desafio Precificação

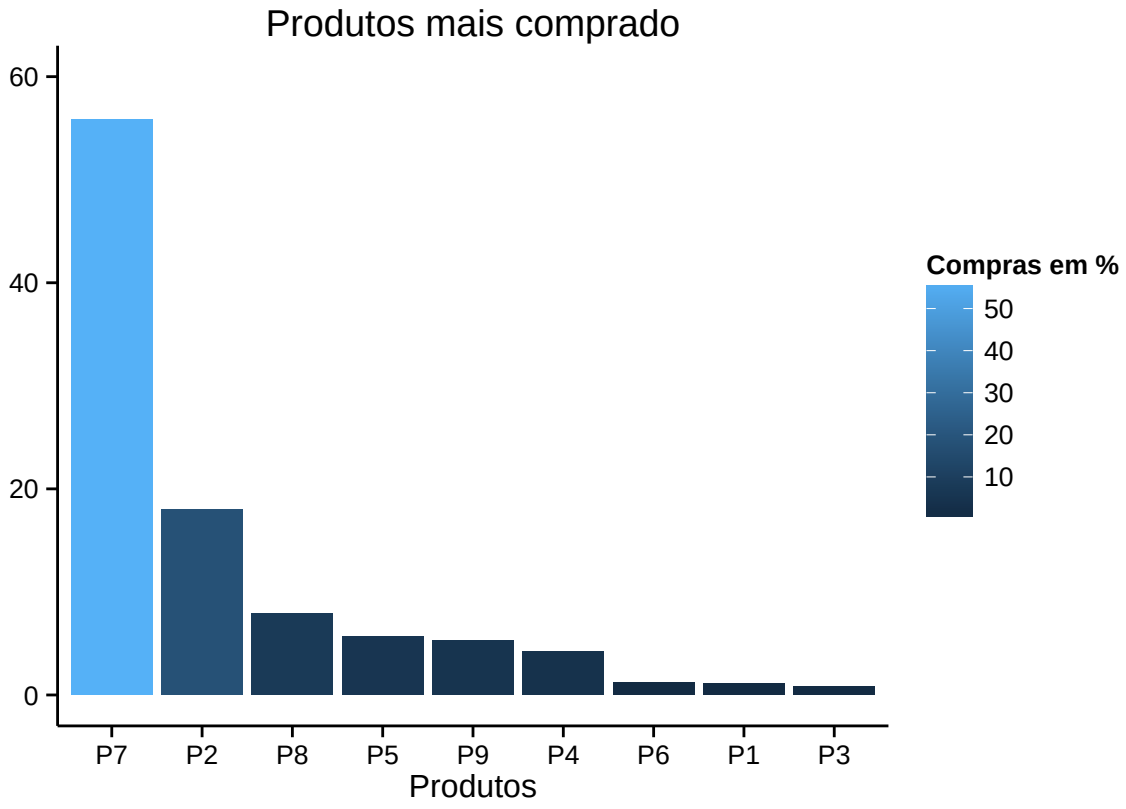
O problema a seguir trata do desafio de precificação, onde se buscar meios para otimizar a relação demanda e preço. Os dados utilizados são uma pequena amostra de um comércio eletrônico, onde temos basicamente dois arquivos "sales.csv" e "comp_prices.csv com 351.091 e 50.114 linhas, respectivamente.

Arquivo sales.csv

O arquivo sales.csv contém informações transacionais, onde cada linha representa uma venda. Possui as colunas:

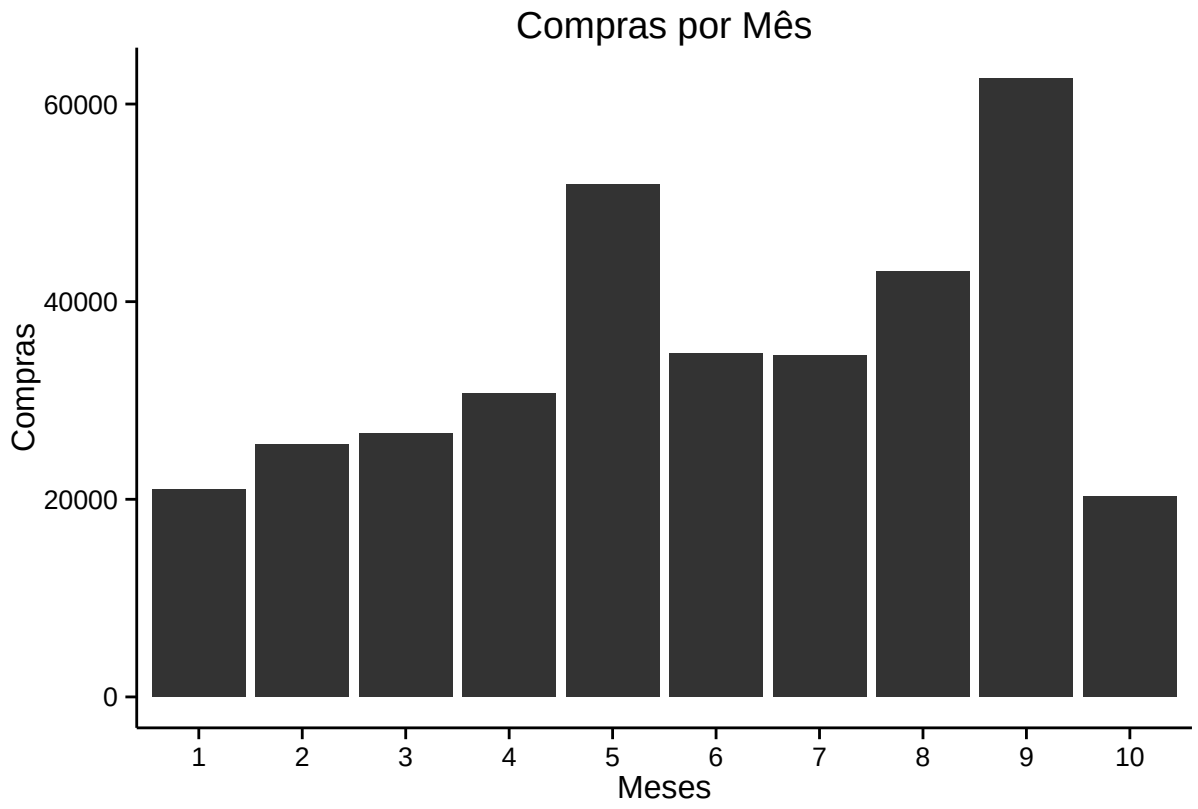
- PROD_ID: ID de produto. Com os dados para nove produtos, P1 a P9
- DATE_ORDER: Data da venda, no formato YYYY-MM-DD
- QTY_ORDER: Quantidade vendida
- PRICE: Preço da venda

Vamos primeiro analisar a quantidade de produtos vendidos.



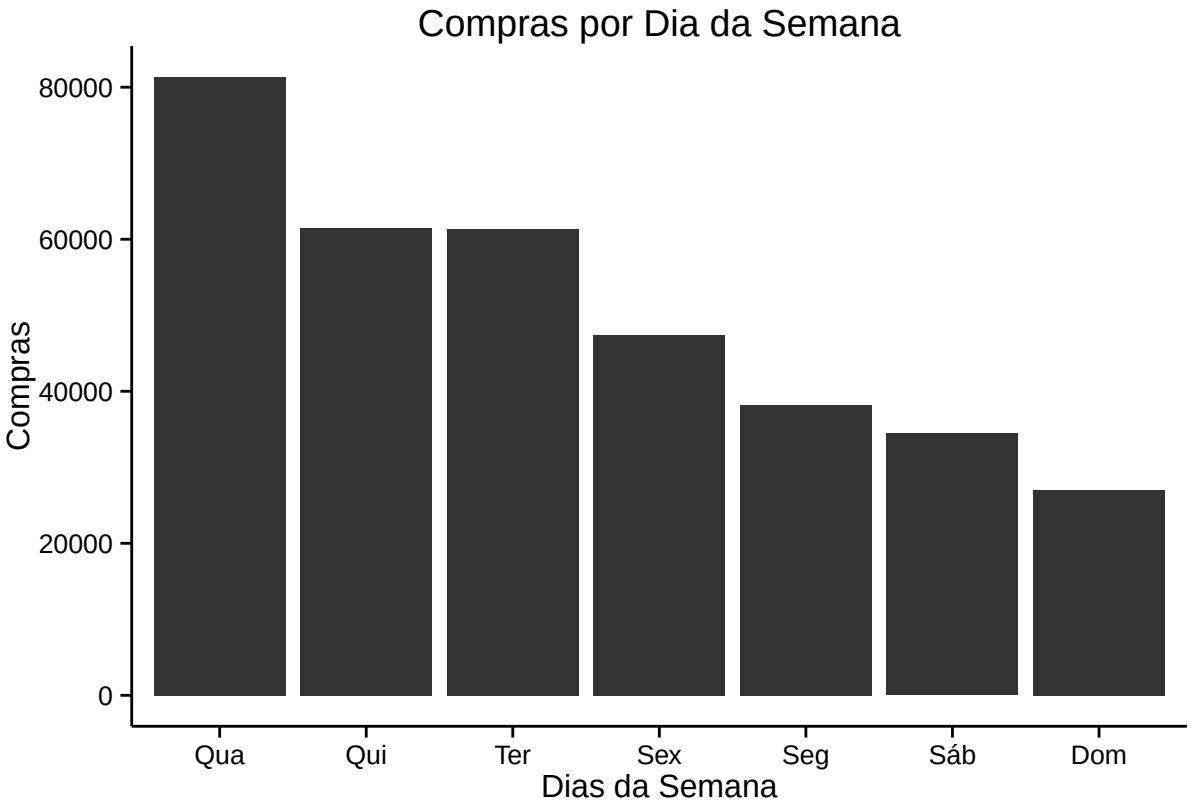
É possível notar que o produto P7 foi o produto mais comprado, com 55% das compras. O segundo colocado é o P2, que não chega nem a 20% das compras.

Podemos observar como é a distribuição das compras ao longo do ano.



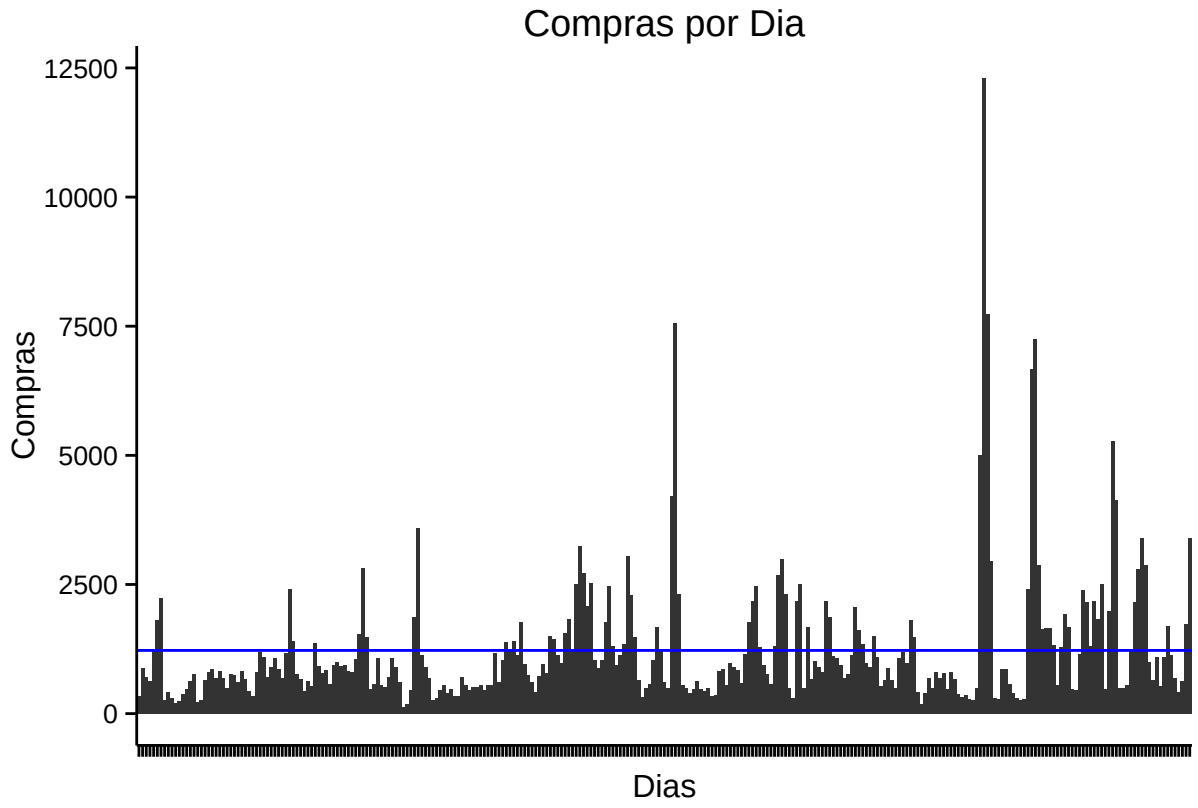
O mês com maior volume de compras é o mês de Setembro (9). Interessante notar que o mês de Maio (5) ficou em segundo lugar, sendo esse o mês do dia das mães.

É interessante investigar também em qual dia da semana que acontece um maior volume de compras.



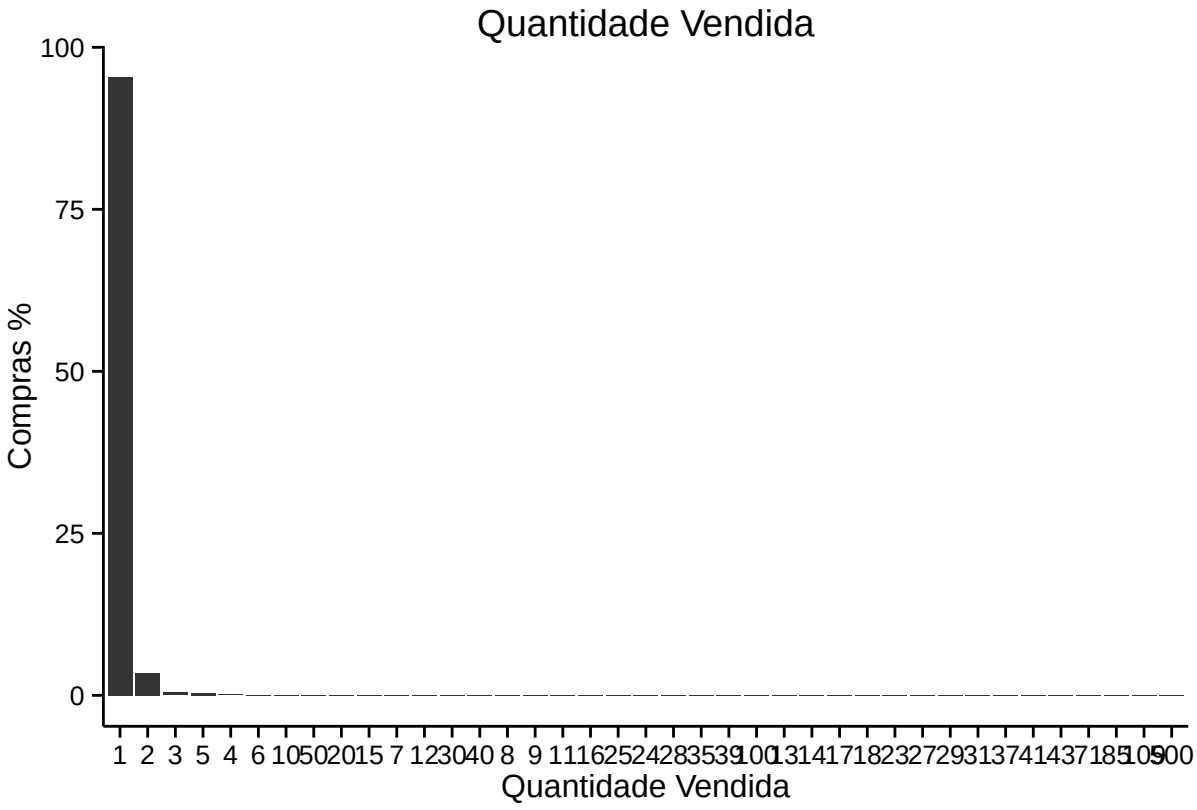
Para essa amostra de dados, o dia que acontece o maior volume de compras é a quarta-feira.

Podemos verificar também se o volume de compras ao longo dos dias é uniforme ou não.

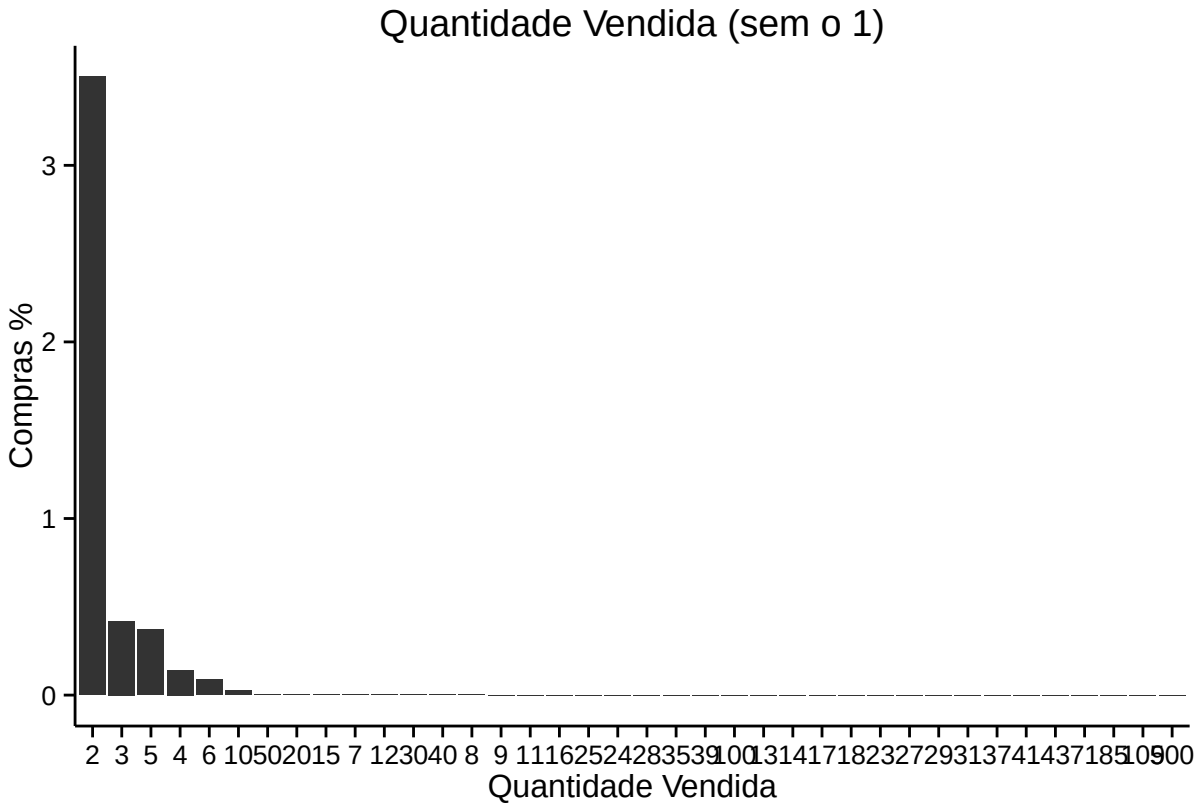


A linha azul representa a média total de compras. Podemos observar que alguns dias superam e muito a média de compras. Os dias 19 e 20 Agosto foram os dias em que se teve o maior volume de compras

Analisamos também a quantidade vendida.

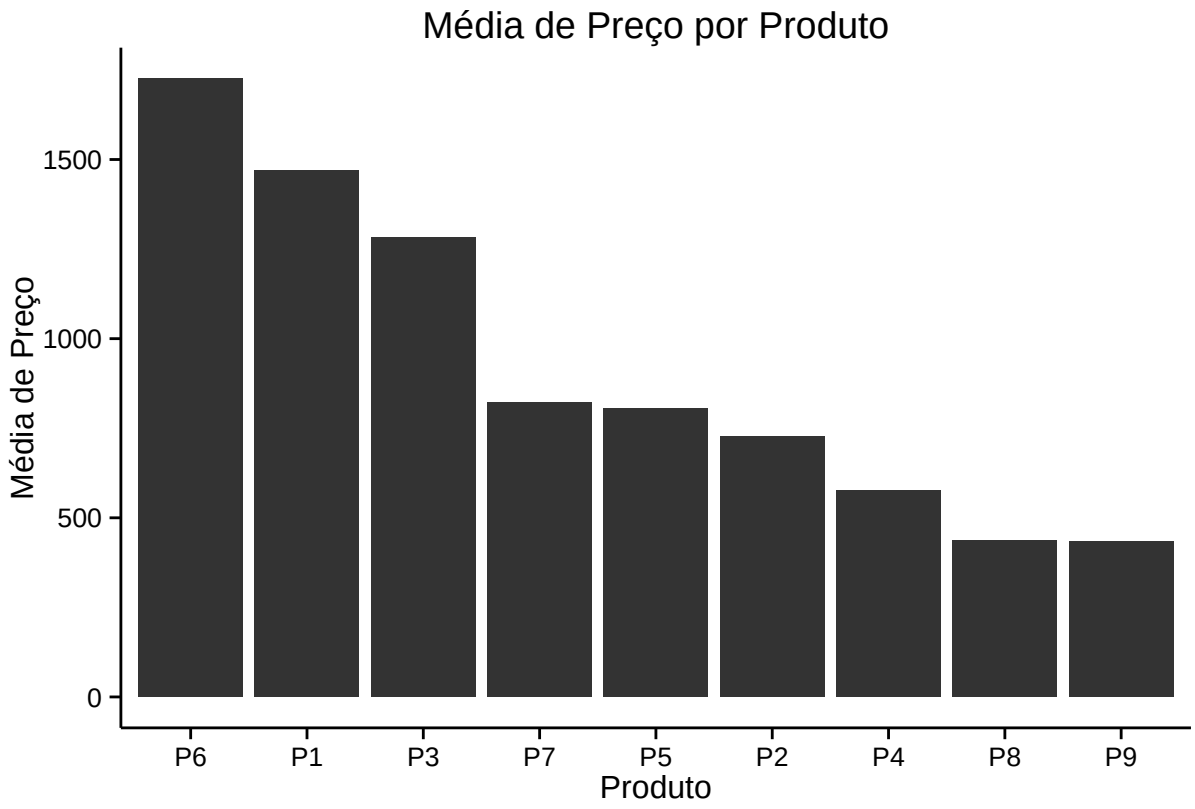


Em 95% das compras apenas 1 produto foi comprado.



2 produtos foram comprados em menos de 4% das compras. Mostrando que o padrão de compra desse comércio eletrônico é majoritariamente de 1 produto.

Podemos também analisar a média de preços por produto.



É possível observar que o produto P6 possui maior média de preço.

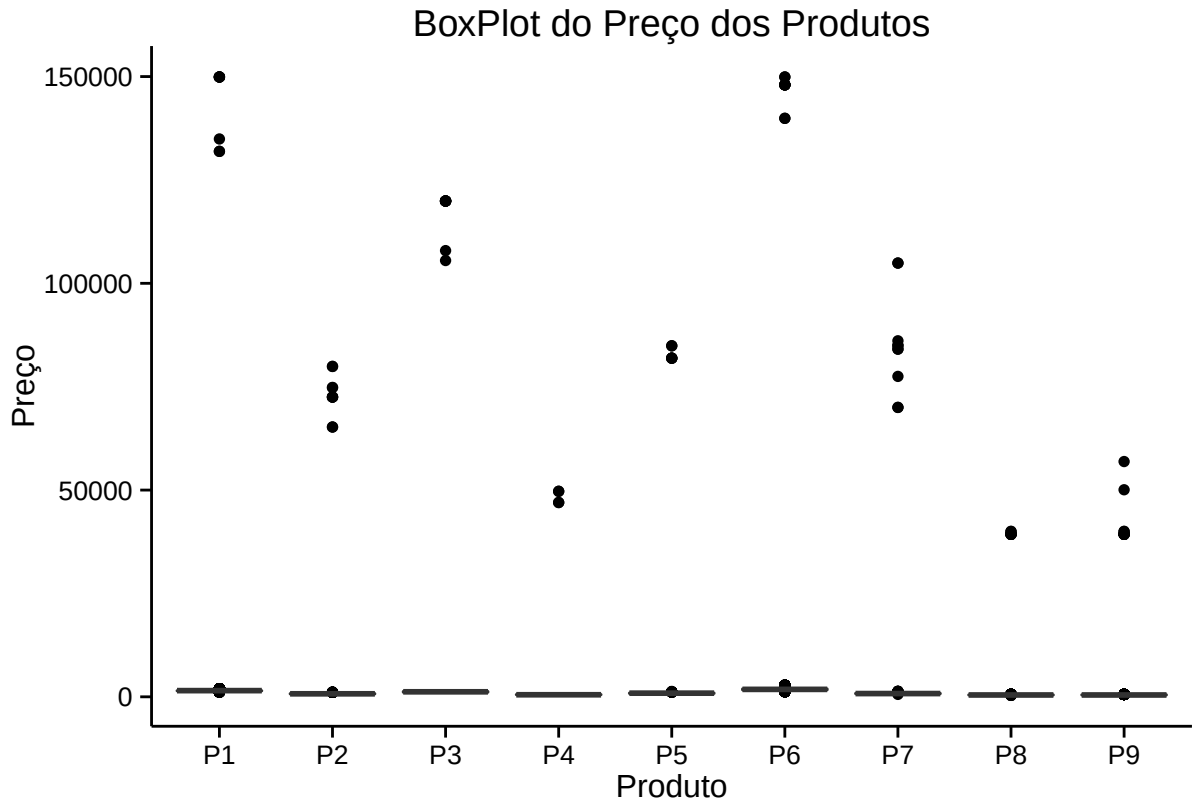
Arquivo comp_prices.csv

O arquivo comp_prices.csv contém dados de monitoramento dos preços dos concorrentes. É possível encontrar os dados de seis concorrentes, C1 a C6, que são monitorados duas vezes ao dia.

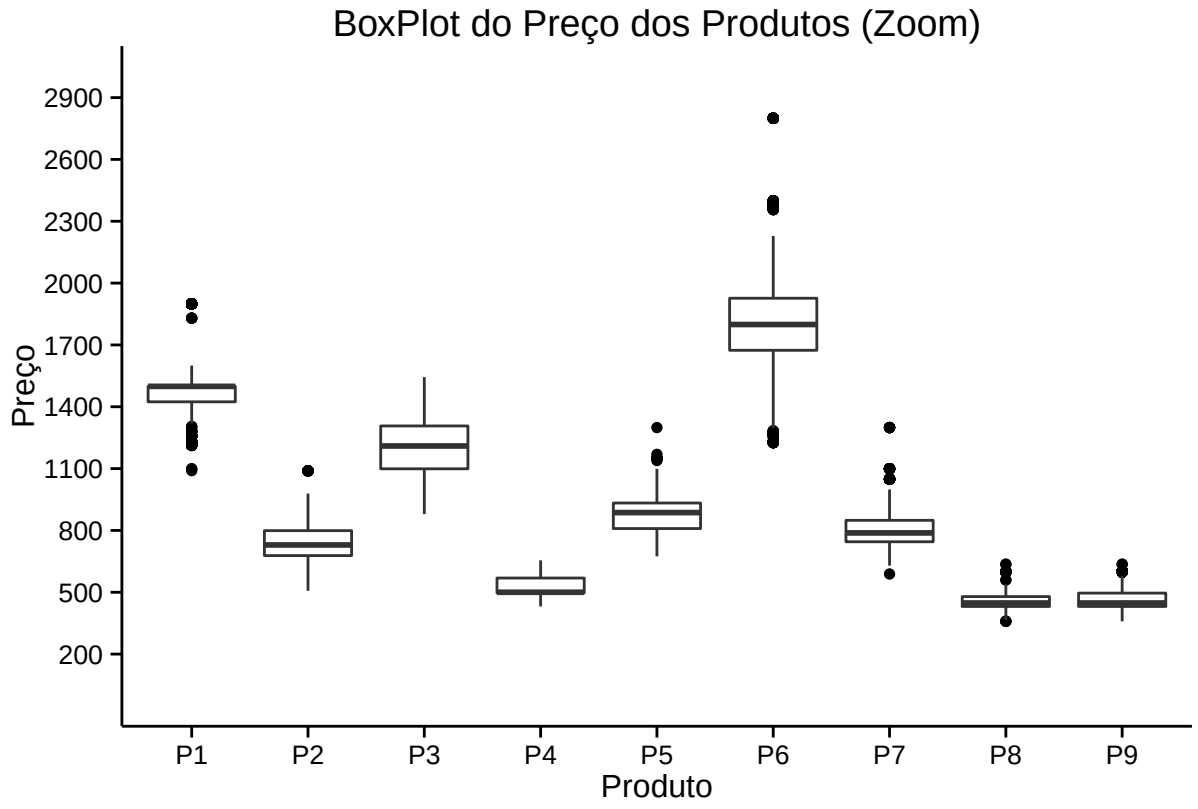
Possui as colunas:

- PROD_ID: ID de produto
- DATE_EXTRACTION: Data e hora da extração dos preços dos concorrentes, no formato YYYY-MM-DD HH:MM:SS
- COMPETITOR: ID do concorrente (C1 a C6)
- COMPETITOR_PRICE: Preço do concorrente, por produto
- PAY_TYPE: Tipo de pagamento

Vamos analisar a distribuição do valor dos preços por produto.

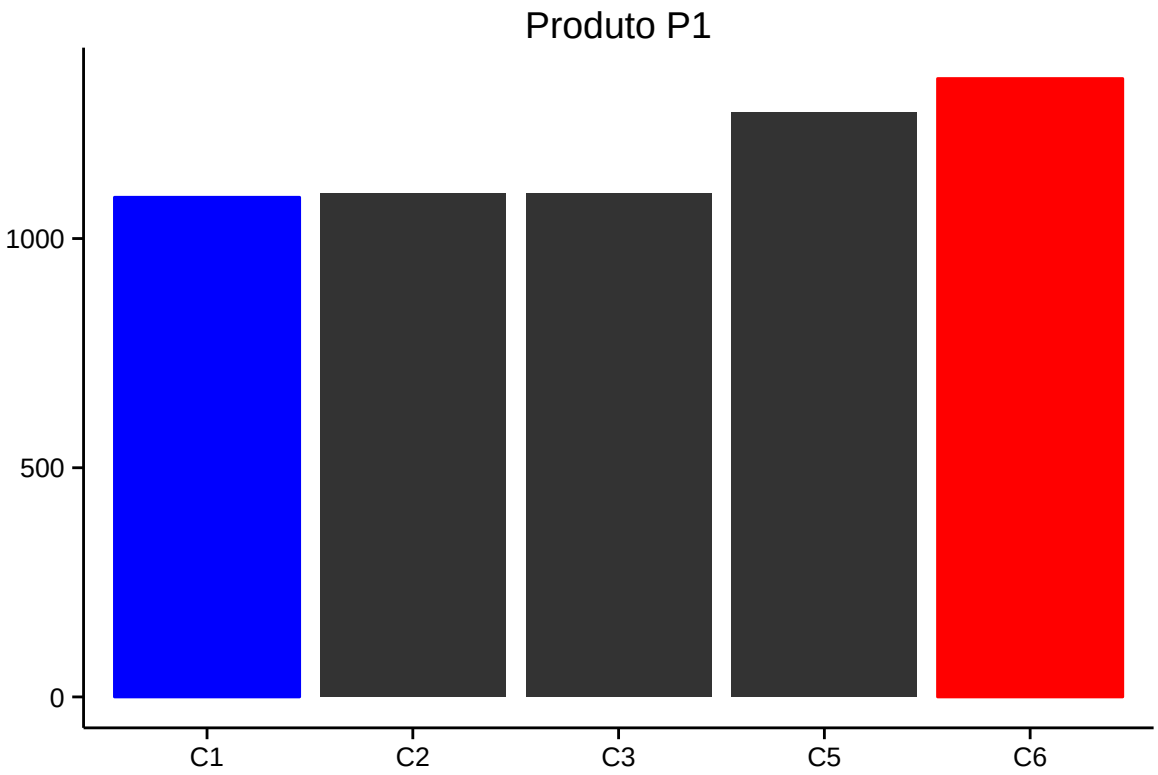


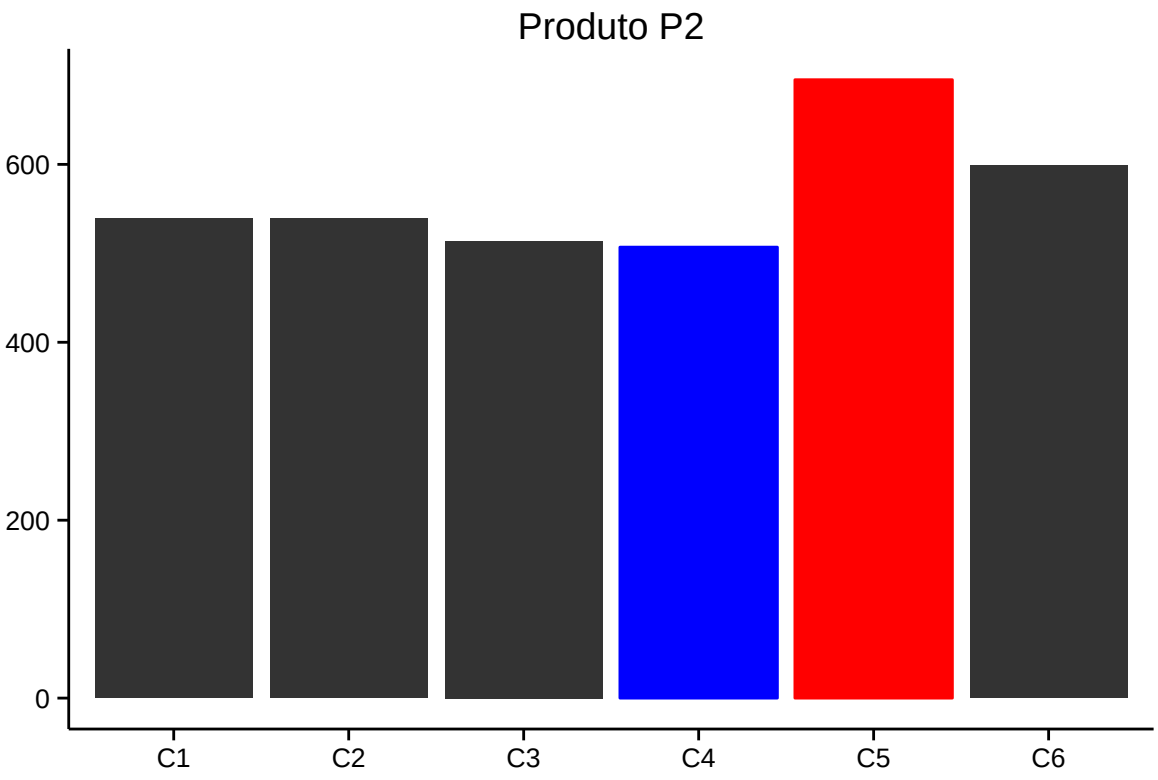
É possível perceber que todos os produtos possuem outlier. Isso pode às vezes significar erro de digitação, erro na hora de coletar os dados.

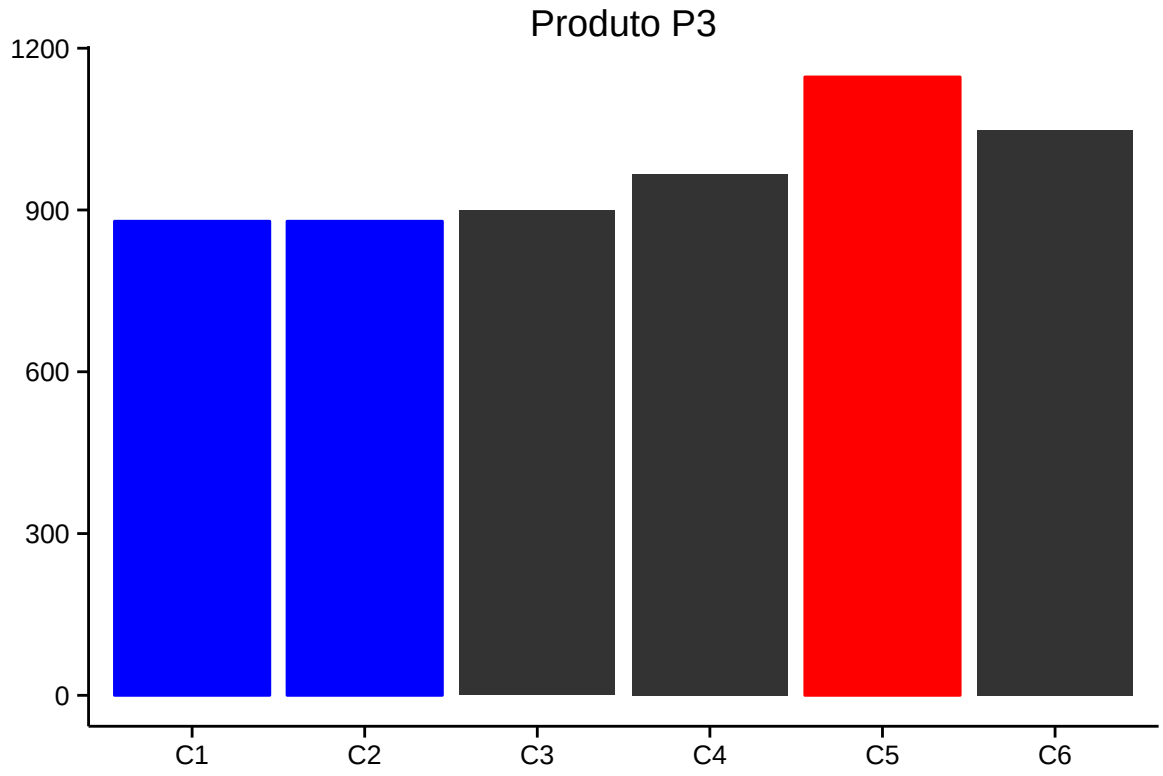


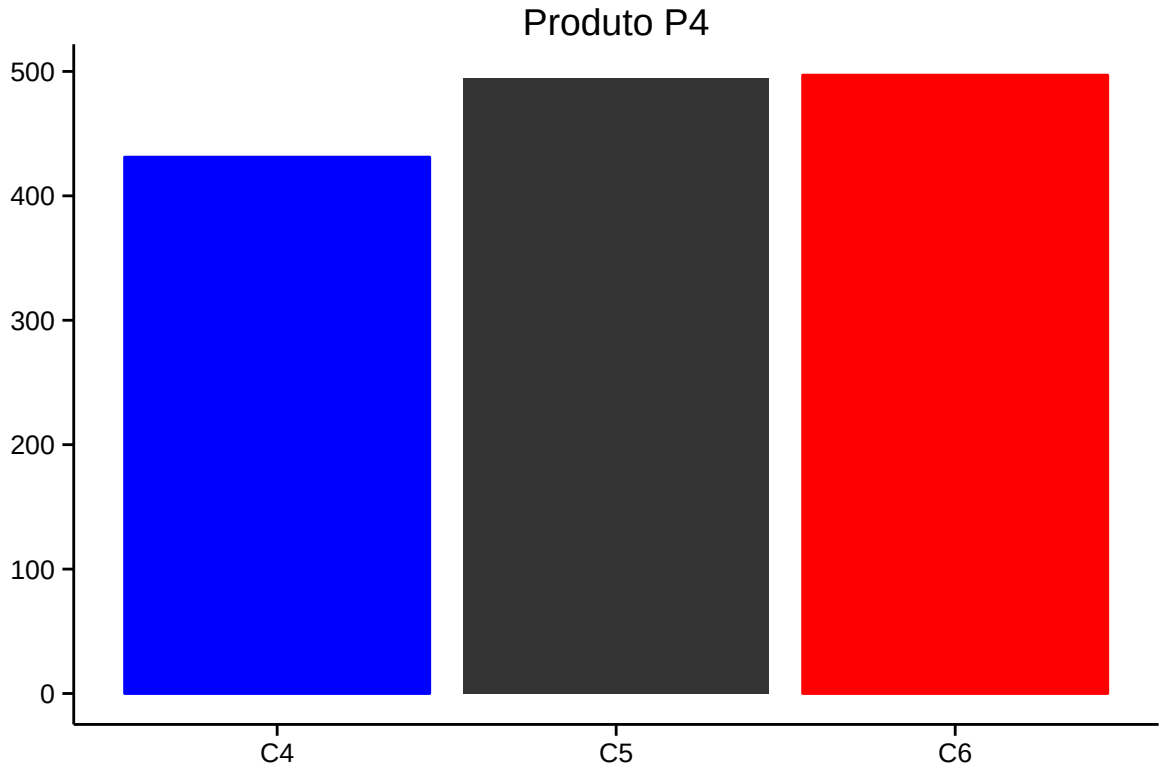
Podemos perceber que o P6 é o produto que possui a média mais alta. Já o produto P8 possui a média mais baixa.

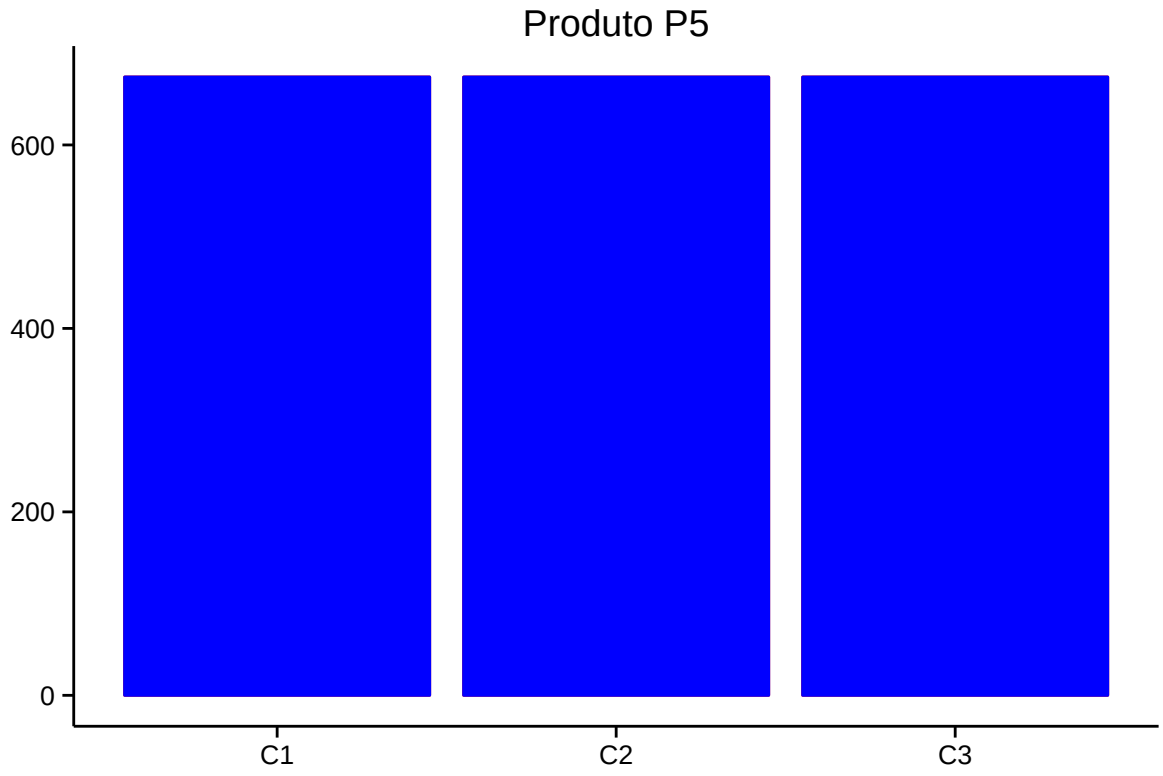
Interessante notar qual o competidor que possui o maior e menor preço para cada produto.

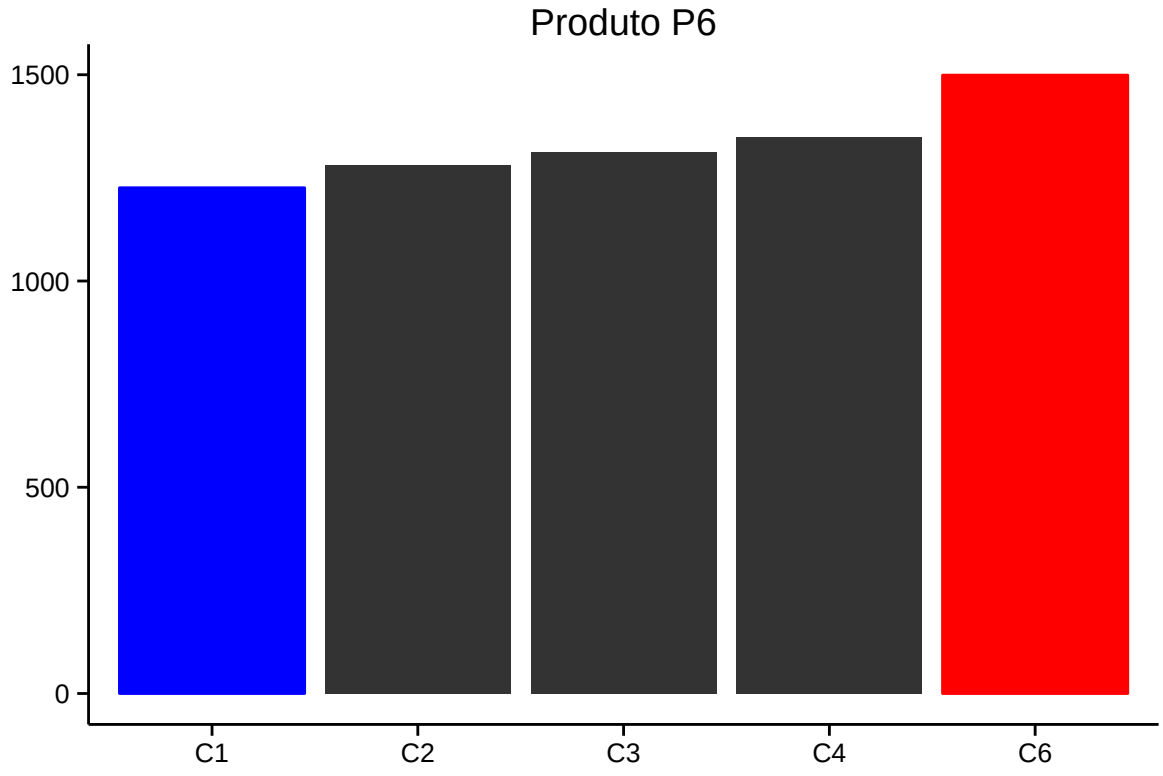


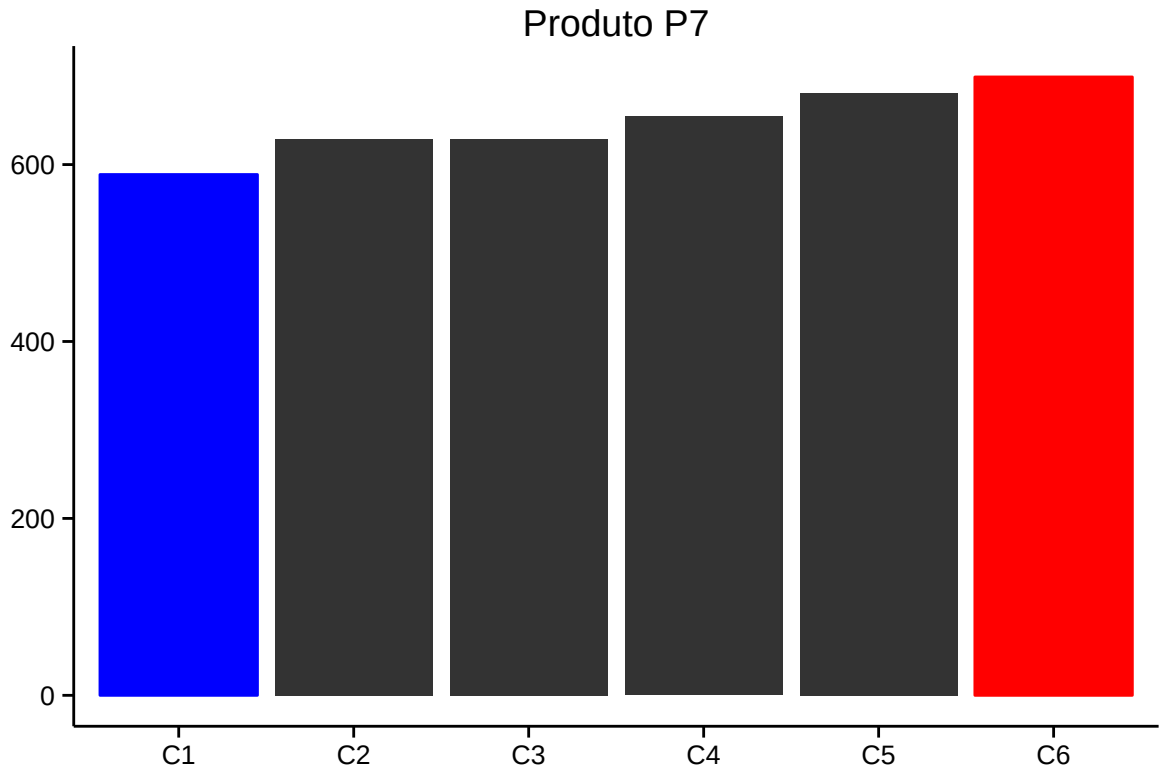


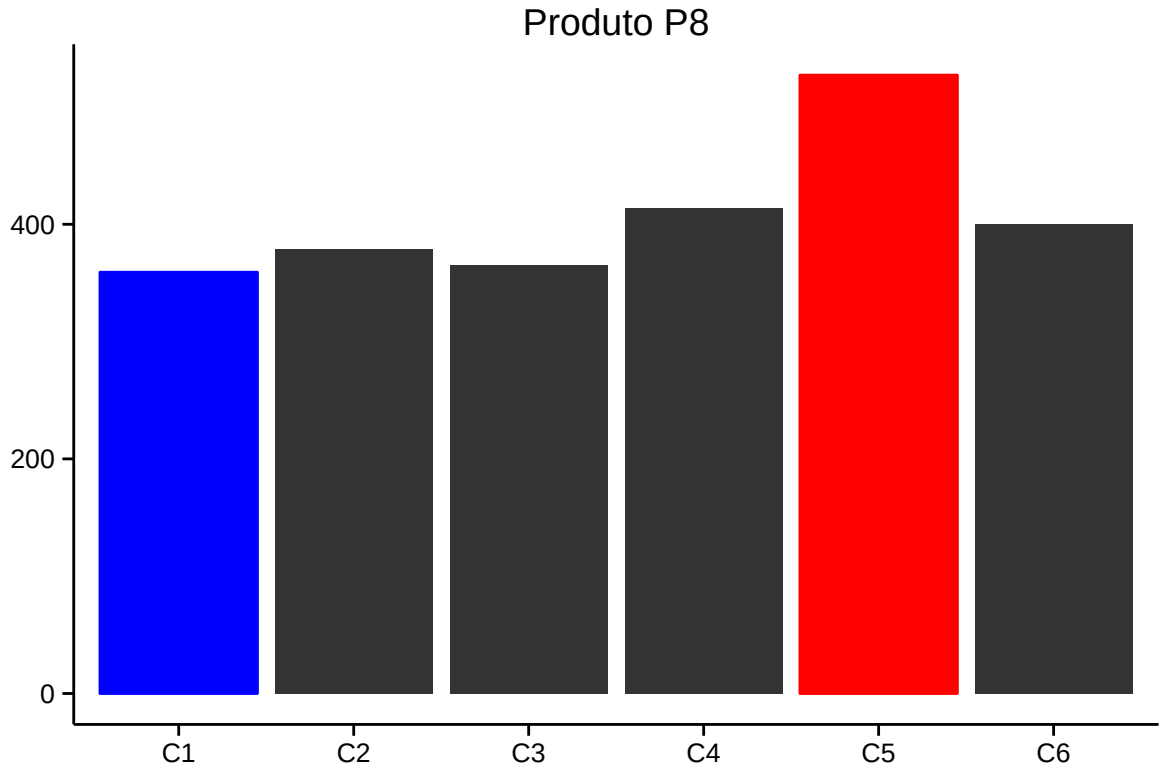

Produto P3
1200
900
600
300
0
C1
C2
C3
C4
C5
C6

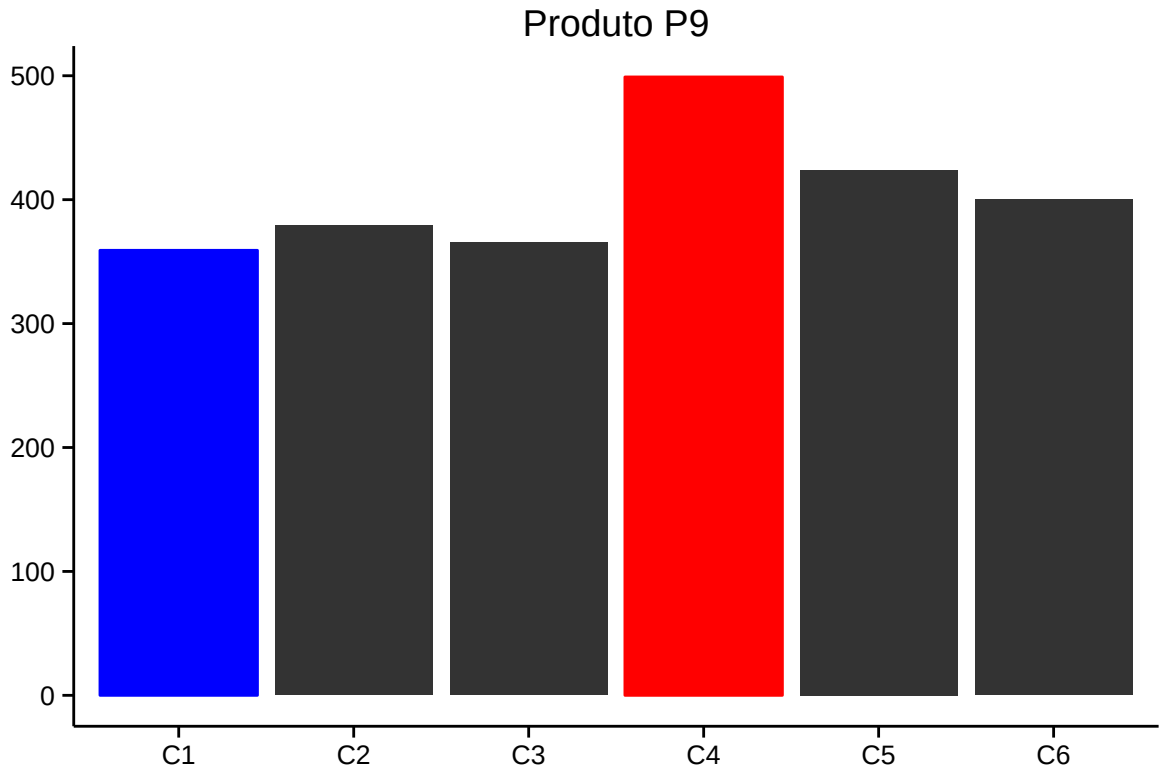

Produto P4
500
400
300
200
100
0
C4
C5
C6


Produto P5
600
400
200
0
C1
C2
C3


Produto P6
1500
1000
500
0
C1
C2
C3
C4
C6

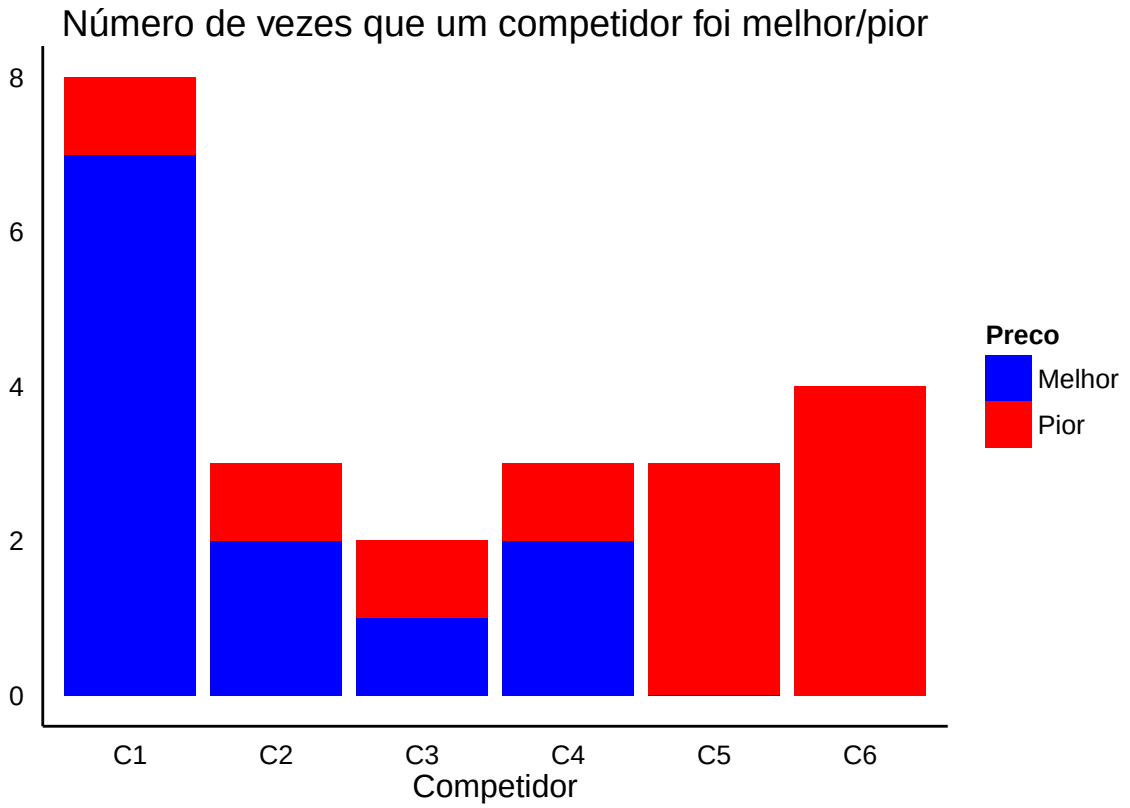

Produto P7
600
400
200
0
C1
C2
C3
C4
C5
C6


Produto P8
400
200
0
C1
C2
C3
C4
C5
C6



É importante notar que para o P5 todos os 3 concorrentes possuem o mesmo valor mínimo. A coluna em azul significa que aquele concorrente possui valor mínimo, já a coluna em vermelho significa que aquele concorrente possui valor máximo em comparação com os outros concorrentes.

Qual é então o concorrente mais importante?



Se a gente considerar que um concorrente é importante quando aplica os melhores preços, então podemos concluir que o concorrente C1 é o melhor concorrente e o concorrente C6 é o pior corrente.

Pré-processamento

Foram feitos vários pre-processamentos com o objetivo de criar um data frame ideal para a solução do problema. Para mais detalhe de implementação olhar o código em pre-processamento.Rmd. Como resultado foi criado um data frame com as seguintes colunas:

- "PROD_ID"
- "dia"
- "qty"
- "preco"
- "date"
- "hora"
- "mes"
- "semana"
- "dia_da_semana_num"
- "C1"

- “C2”
- “C3”
- “C4”
- “C5”
- “C6”

O modelo

Para a criação do modelo utilizei a biblioteca h2o, por se tratar de um open-source software para big-data analysis. O h2o é bastante rápido e flexível, podendo assim ser possível carregar uma grande quantidade de dados. Faz parte de uma comunidade que vem crescendo cada dia mais.

```
##
## H2O is not running yet, starting it now...
##
## Note:  In case of errors look at the following log files:
##     /tmp/RtmpWTbAeT/h2o_rodolfo_started_from_r.out
##     /tmp/RtmpWTbAeT/h2o_rodolfo_started_from_r.err
##
##
## .............Successfully connected to http://127.0.0.1:54321/
##
## R is connected to the H2O cluster:
##     H2O cluster uptime:         18 seconds 220 milliseconds
##     H2O cluster version:        3.0.1.7
##     H2O cluster name:           H2O_started_from_R_rodolfo_mzp350
##     H2O cluster total nodes:    1
##     H2O cluster total memory:   0.85 GB
##     H2O cluster total cores:    4
##     H2O cluster allowed cores:  4
##     H2O cluster healthy:        TRUE
```

Inicialmente vamos dividir o dataset entre o dataset de treino e validação. Essa divisão é importante por evita o overfitting, que ocorre quando um modelo estatístico super se adapta ao conjunto treinado, dessa forma quando o modelo recebe um valor pelo o qual ele não foi treinado, ele vai gerar uma predição muito ruim. É importante essa divisão entre treino e validação para verificar em qual ponto o modelo começa a sofrer overfitting.

```
##
  |
  |                                                                      |   0%
  |
  |======================================================================| 100%

##
  |
```

```
  |                                                                      |   0%
  |
  |======================================================================| 100%
```

A minha estratégia é primeiro achar o melhor modelo que encontre a quantidade diária que será vendida para cada produto, e só depois encontrar o melhor modelo que encontre o preço diário desse produto.

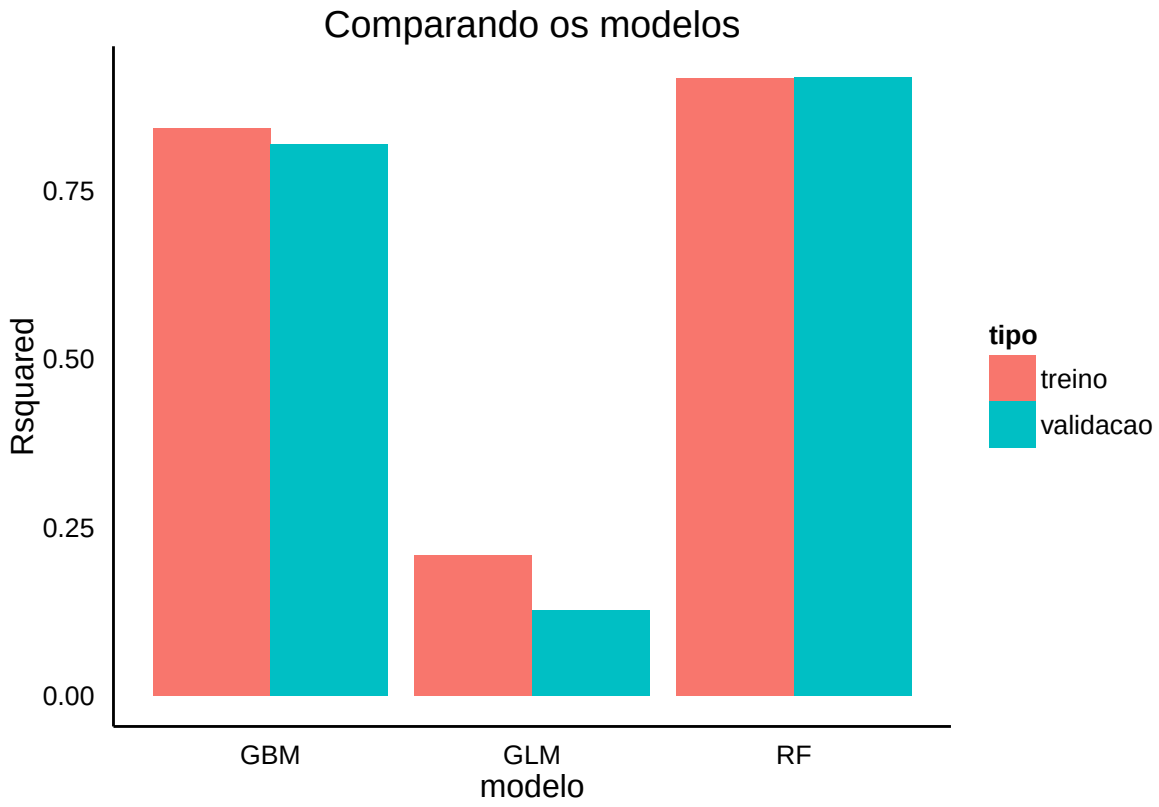
Vamos inicialmente trabalhar com os modelos GBM, random florest e GLM. O ideal seria inicialmente rodar todos os modelos com um grande número de árvores, grande profundidade e uma taxa de aprendizado pequena por interação, porém isso leva um tempo grande na minha máquina atual (com apenas 4GB)

```
##
  |
  |                                                                      |   0%
  |
  |=                                                                     |   2%
  |
  |===                                                                   |   4%
  |
  |=====                                                                 |   8%
  |
  |==========                                                            |  16%
  |
  |================                                                      |  24%
  |
  |==================                                                    |  28%
  |
  |=====================                                                 |  32%
  |
  |============================                                          |  42%
  |
  |==================================                                    |  52%
  |
  |====================================                                  |  56%
  |
  |========================================                              |  62%
  |
  |============================================                          |  68%
  |
  |================================================                      |  76%
  |
  |====================================================                  |  84%
  |
  |======================================================                |  86%
  |
  |==========================================================            |  92%
  |
  |================================================================      |  96%
  |
  |======================================================================| 100%

##
  |
  |                                                                      |   0%
```

```
|
|=                                                                |   2%
|
|===                                                              |   4%
|
|====                                                             |   6%
|
|======                                                           |  10%
|
|========                                                         |  12%
|
|==========                                                       |  16%
|
|============                                                     |  18%
|
|==============                                                   |  22%
|
|=================                                                |  26%
|
|===================                                              |  28%
|
|=====================                                            |  32%
|
|========================                                         |  36%
|
|==========================                                       |  40%
|
|============================                                     |  42%
|
|==============================                                   |  46%
|
|==================================                               |  52%
|
|====================================                             |  56%
|
|======================================                           |  58%
|
|========================================                         |  62%
|
|==========================================                       |  66%
|
|============================================                     |  70%
|
|==============================================                   |  74%
|
|================================================                 |  78%
|
|==================================================               |  82%
|
|====================================================             |  86%
|
|======================================================           |  92%
|
|========================================================         |  94%
```

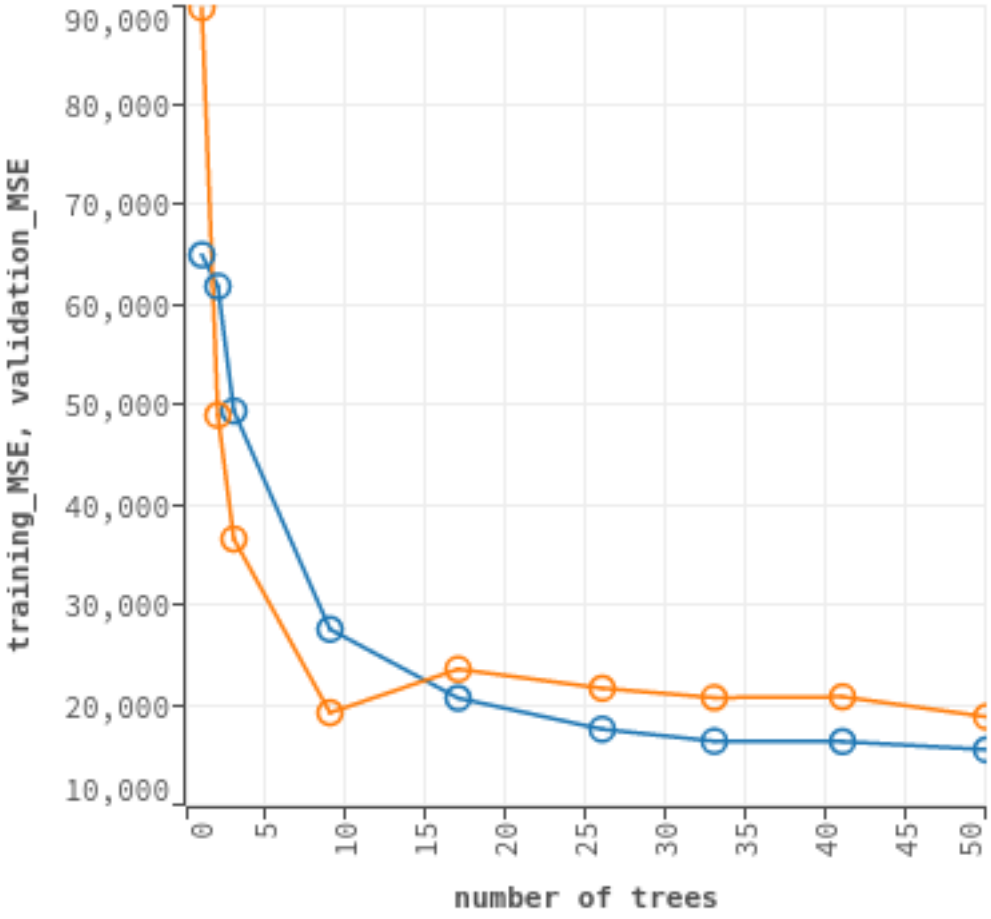
```
  |
  |======================================================================| 100%

##
  |
  |                                                                      |   0%
  |
  |=========                                                             |  14%
  |
  |======================================================================| 100%
```

Para verificar qual dos 3 modelos é o melhor, utilizamos a métrica Rsquared, onde o valor do Rsquared (entre 0 e 1) é o percentual de variância explicada pelo o modelo. Quanto maior foi o Rsquared melhor é o modelo.



É possível notar que o Random Florest teve um melhor resultado do que os outros modelos, obtendo assim um Rsquared maior, então optamos por escolher o Random Florest para o caso da predição da quantidade de vendas de produtos.

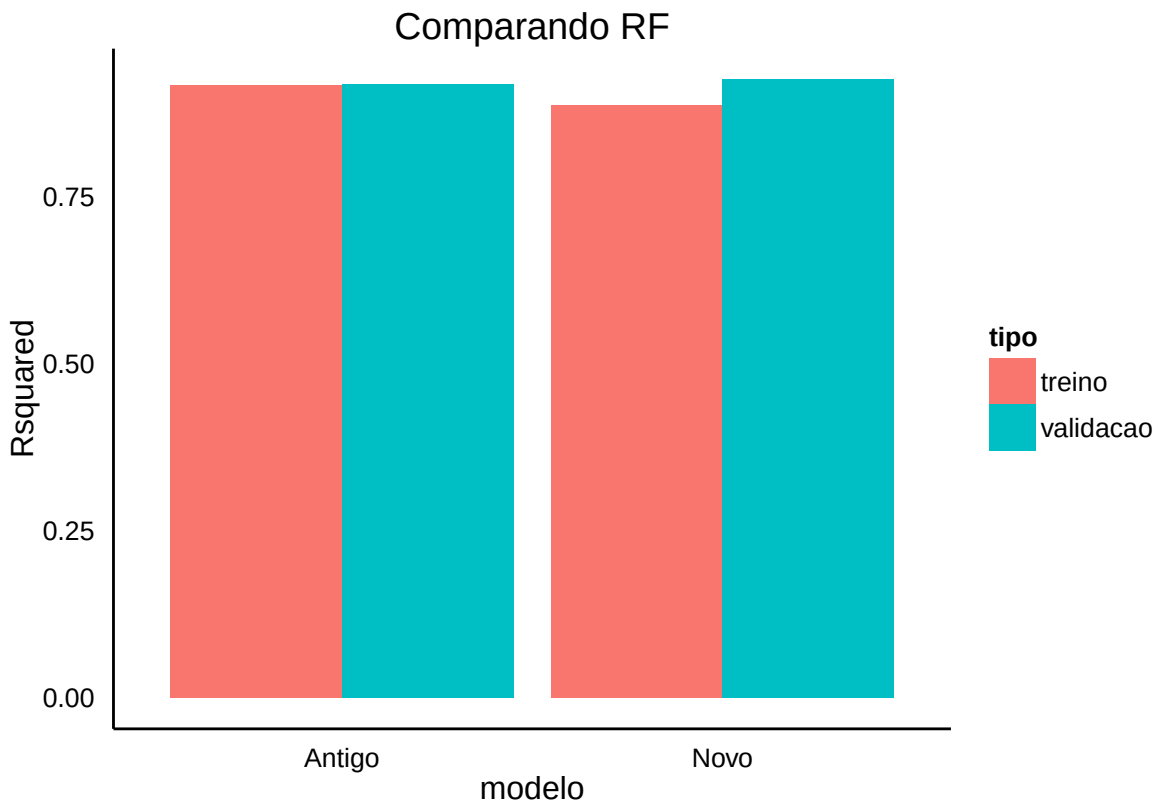
Como o Random Florest foi o escolhido, é interessante observar como se deu o treinamento ao longo das criações das árvores. Para evitar o overfitting dividimos os dados de treino em treino e validação. Dessa forma podemos observar o exato momento em que o modelo passa a sofrer o overfitting.



É possível notar que depois da árvore 17, o modelo passa a sofrer overfitting. Por esse motivo criamos um novo modelo, dessa vez parando o treinamento na árvore 17. A linha azul significa a evolução do treino e a linha laranja significa validação.

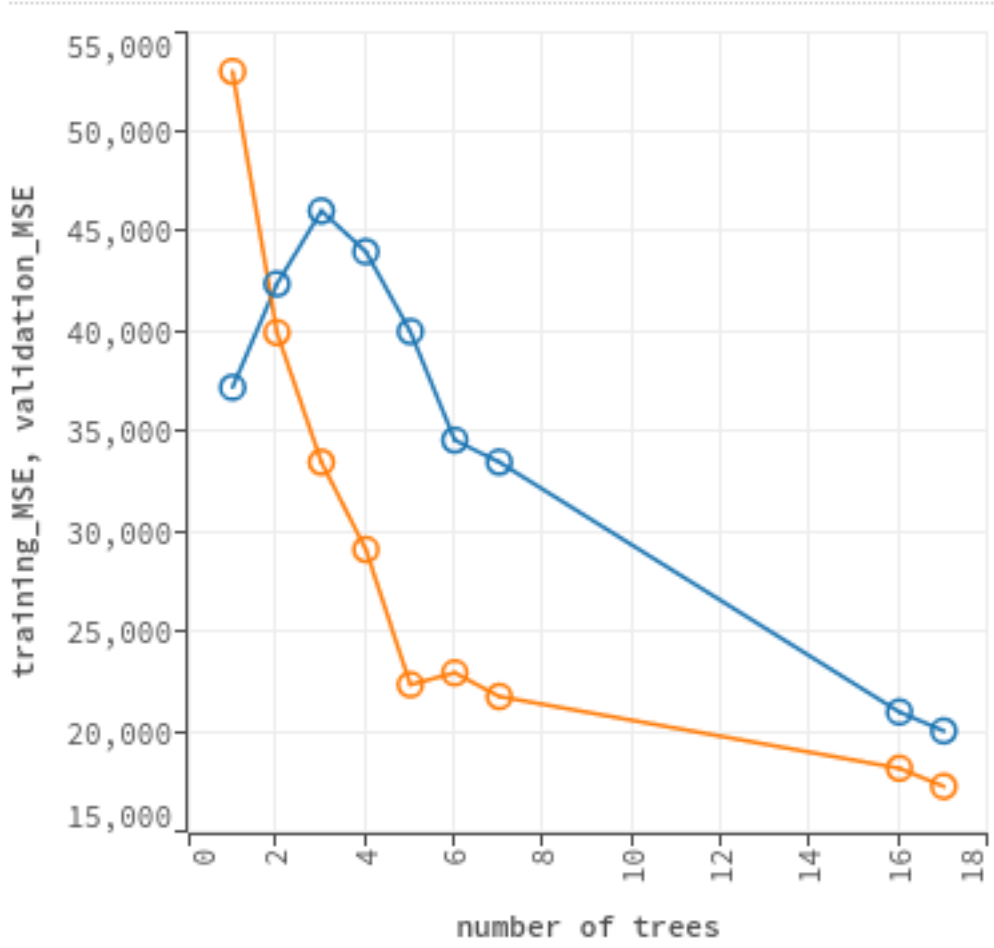
```
##
  |
  |                                                                      |   0%
  |
  |====                                                                  |   6%
  |
  |===========                                                           |  18%
  |
  |===================                                                   |  29%
  |
  |=======================                                               |  35%
  |
  |==============================                                        |  47%
  |
  |=======================================                               |  65%
  |
  |================================================                      |  76%
  |
  |==================================================                    |  82%
```

```
|
|=============================================================     |  94%
|
|=================================================================| 100%
```

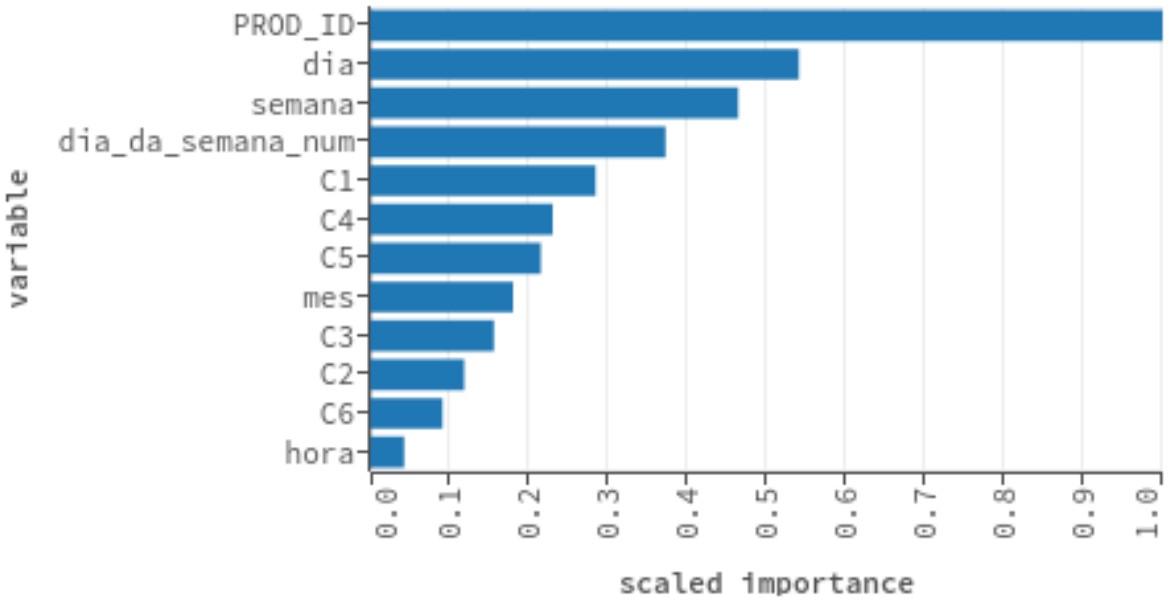


É possível notar que o valor de Rsquared do treino sofreu uma ligeira queda, porém o valor do Rsquared da validação sofreu um aumento. Dessa forma a gente evita o overfitting

Evolução do treinamento



É interessante notar também a importância das variáveis para a criação dos modelos.



Notamos que a variável **PROD_ID, dia e semana** foram as variáveis com maior importância para a criação do modelo. Por esse motivo é interessante criar novas variáveis que podem agregar valor a essas variáveis para melhorar o modelo.

Uma boa alternativa é criar uma variável que mostre a quantidade de compras que o item teve no dia anterior. Outra boa alternativa é criar uma variável que mostre a média de preço e a quantidade de compras do item na última semana.

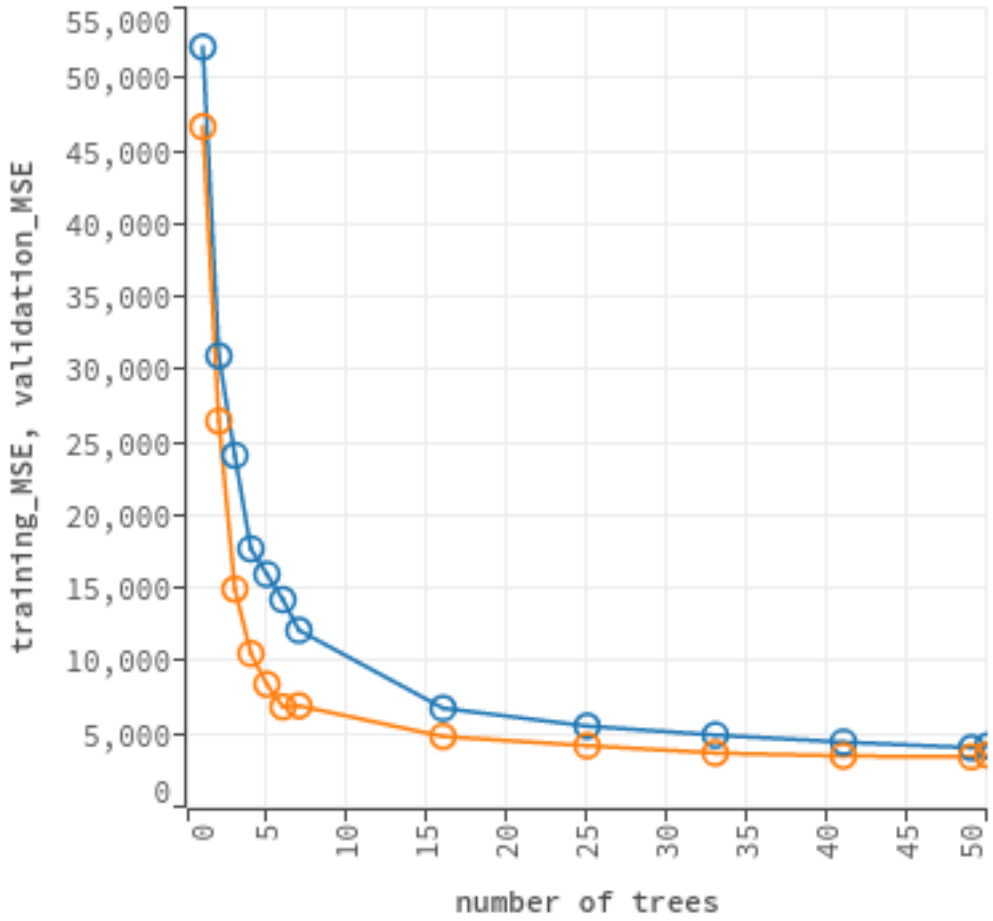
Foi feito então um novo pré-processamento, onde foram adicionadas as três novas colunas **qty_semana, media_preco_semana, qty_anterior**.

```
##
  |
  |                                                                      |   0%
  |
  |======================================================================| 100%

##
  |
  |                                                                      |   0%
  |
  |======================================================================| 100%
```

Um novo modelo será criado, seguindo os mesmo passos anteriores: * Rodar RF com 50 árvores e 30 de profundidade * Encontrar o ponto onde o modelo começa a ficar especialista (overfitthing) * Rodar RF com o número de árvores otimizado

```
##
  |
  |                                                                      |   0%
  |
  |=                                                                     |   2%
  |
  |====                                                                  |   6%
  |
  |======                                                                |  10%
  |
  |=========                                                             |  14%
  |
  |============                                                          |  18%
  |
  |==============                                                        |  22%
  |
  |=================                                                     |  26%
  |
  |=====================                                                 |  32%
  |
  |======================                                                |  34%
  |
  |========================                                              |  38%
  |
  |============================                                          |  42%
  |
  |==============================                                        |  48%
```

```
|
|================================                                  |  50%
|
|=================================                                 |  54%
|
|====================================                              |  58%
|
|=====================================                             |  62%
|
|========================================                          |  66%
|
|==========================================                        |  70%
|
|============================================                      |  74%
|
|==============================================                    |  78%
|
|================================================                  |  80%
|
|==================================================                |  84%
|
|====================================================              |  90%
|
|======================================================            |  94%
|
|========================================================          |  96%
|
|==================================================================| 100%
```

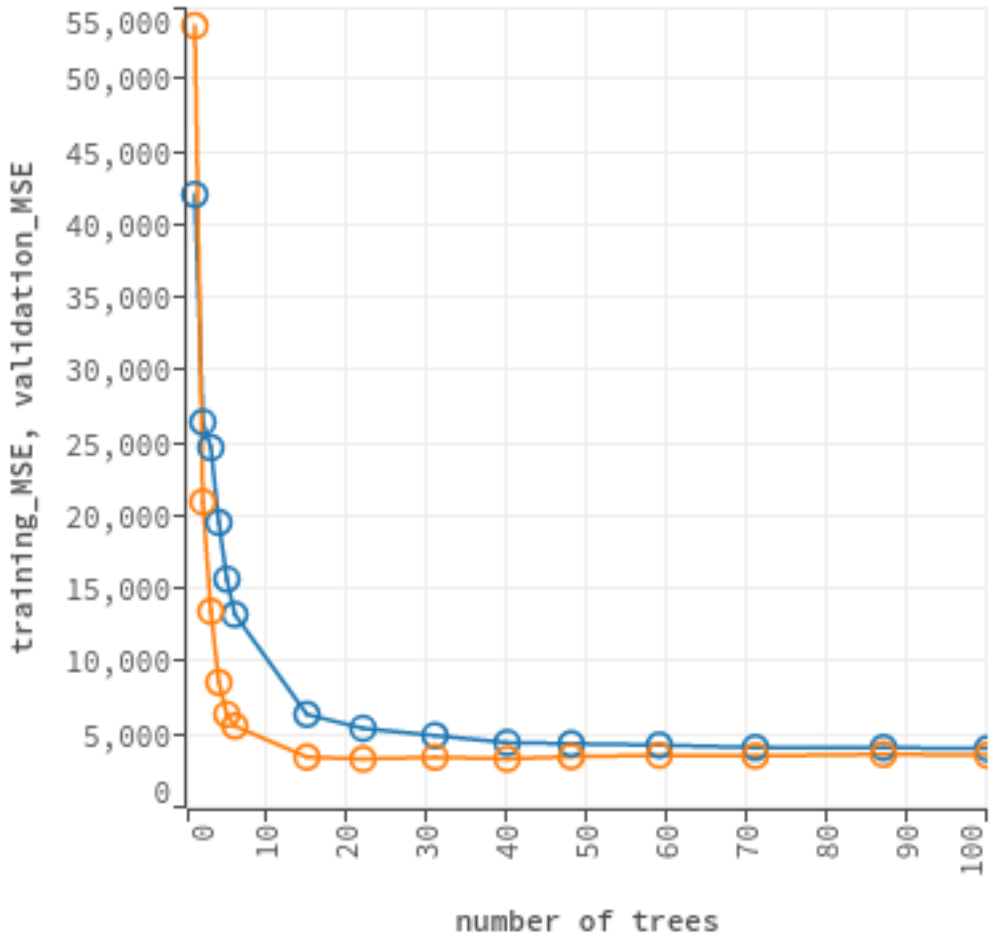


Para esse novo modelo não foi possível notar o ponto onde o modelo começa a ficar especialista. Por esse motivo foi criado um novo modelo, dessa vez utilizando 100 árvores e 30 de profundidade.

```
##
  |
  |                                                                      |   0%
  |
  |=                                                                     |   1%
  |
  |==                                                                    |   3%
  |
  |===                                                                   |   5%
  |
  |====                                                                  |   6%
  |
  |======                                                                |   9%
  |
  |=======                                                               |  11%
  |
  |========                                                              |  13%
  |
  |=========                                                             |  14%
  |
  |==========                                                            |  16%
```

```
|
|============                                                   |  18%
|
|=============                                                  |  20%
|
|==============                                                 |  21%
|
|================                                               |  24%
|
|=================                                              |  25%
|
|==================                                             |  28%
|
|====================                                           |  29%
|
|=====================                                          |  31%
|
|======================                                         |  33%
|
|========================                                       |  35%
|
|========================                                       |  37%
|
|=========================                                      |  39%
|
|==========================                                     |  41%
|
|============================                                   |  43%
|
|=============================                                  |  46%
|
|==============================                                 |  48%
|
|================================                               |  50%
|
|=================================                              |  53%
|
|==================================                             |  54%
|
|====================================                           |  56%
|
|=====================================                          |  58%
|
|======================================                         |  61%
|
|=======================================                        |  63%
|
|========================================                       |  65%
|
|=========================================                      |  66%
|
|==========================================                     |  68%
|
|============================================                   |  71%
```

```
|
|=============================================                    |  73%
|
|==============================================                   |  75%
|
|===============================================                  |  77%
|
|================================================                 |  78%
|
|=================================================                |  80%
|
|==================================================               |  82%
|
|===================================================              |  85%
|
|====================================================             |  87%
|
|=====================================================            |  89%
|
|======================================================           |  92%
|
|=======================================================          |  94%
|
|========================================================         |  96%
|
|========================================================= |  98%
|
|==========================================================| 100%
```

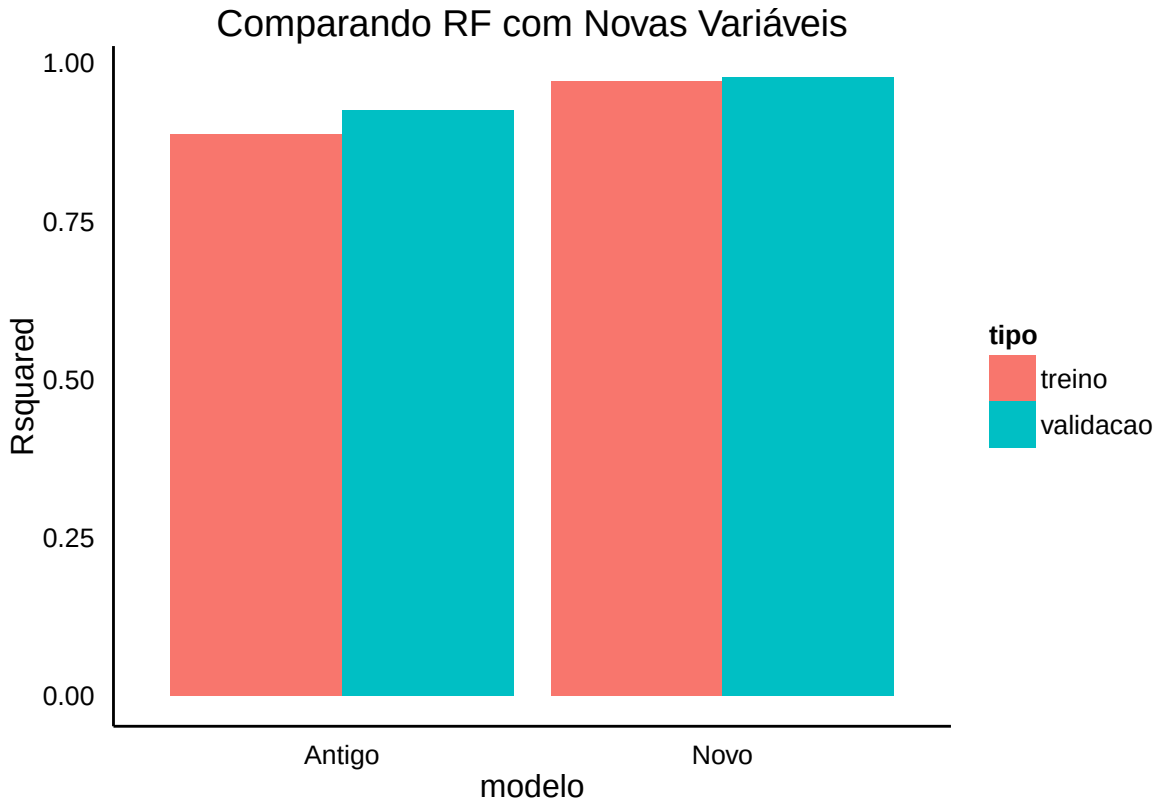


É possível notar que depois da árvore 60, tanto o treinamento quando a validação se estabilizam. Por esse motivo, o número de árvores ideal para esse conjunto de dados é 60 (com mais árvores só estaríamos "desperdiçando" processamento para ter pouca melhora no modelo).

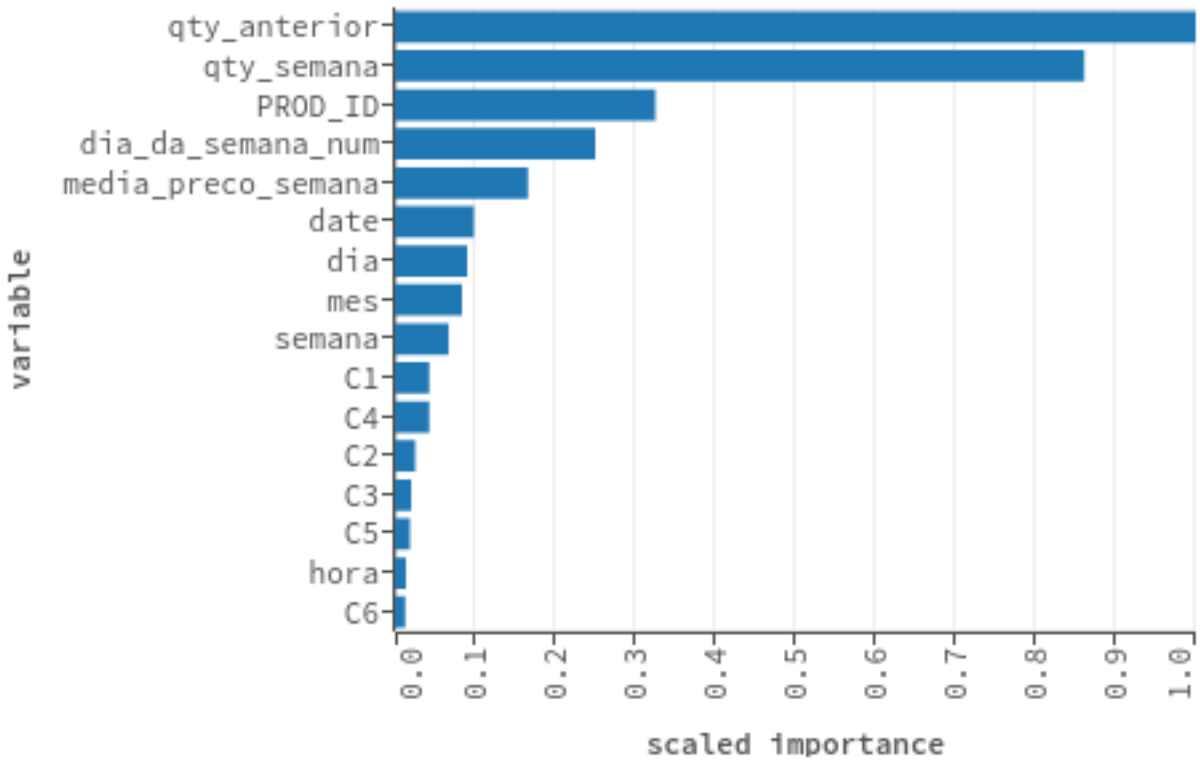
```
##
  |
  |                                                                      |   0%
  |
  |=                                                                     |   2%
  |
  |===                                                                   |   5%
  |
  |======                                                                |  10%
  |
  |=========                                                             |  13%
  |
  |===========                                                           |  17%
  |
  |=============                                                         |  20%
  |
  |===============                                                       |  23%
  |
  |=================                                                     |  27%
```

```
|
|====================                                              |  30%
|
|======================                                            |  33%
|
|=========================                                         |  38%
|
|=========================                                         |  40%
|
|============================                                      |  43%
|
|==============================                                    |  47%
|
|==================================                                |  52%
|
|====================================                              |  55%
|
|======================================                            |  57%
|
|========================================                          |  62%
|
|==========================================                        |  65%
|
|============================================                      |  68%
|
|==============================================                    |  72%
|
|================================================                  |  75%
|
|==================================================                |  78%
|
|====================================================              |  82%
|
|======================================================            |  85%
|
|========================================================          |  88%
|
|============================================================      |  93%
|
|==============================================================    |  97%
|
|================================================================  |  98%
|
|==================================================================| 100%
```

Podemos agora compara e verificar se existiu algum ganho entre o melhor modelo com as novas variáveis



É possível notar que o valor do Rsquared subiu em relação ao modelo anterior. Devemos agora investigar quais foram as variáveis que mais contribuíram para essa mudança.



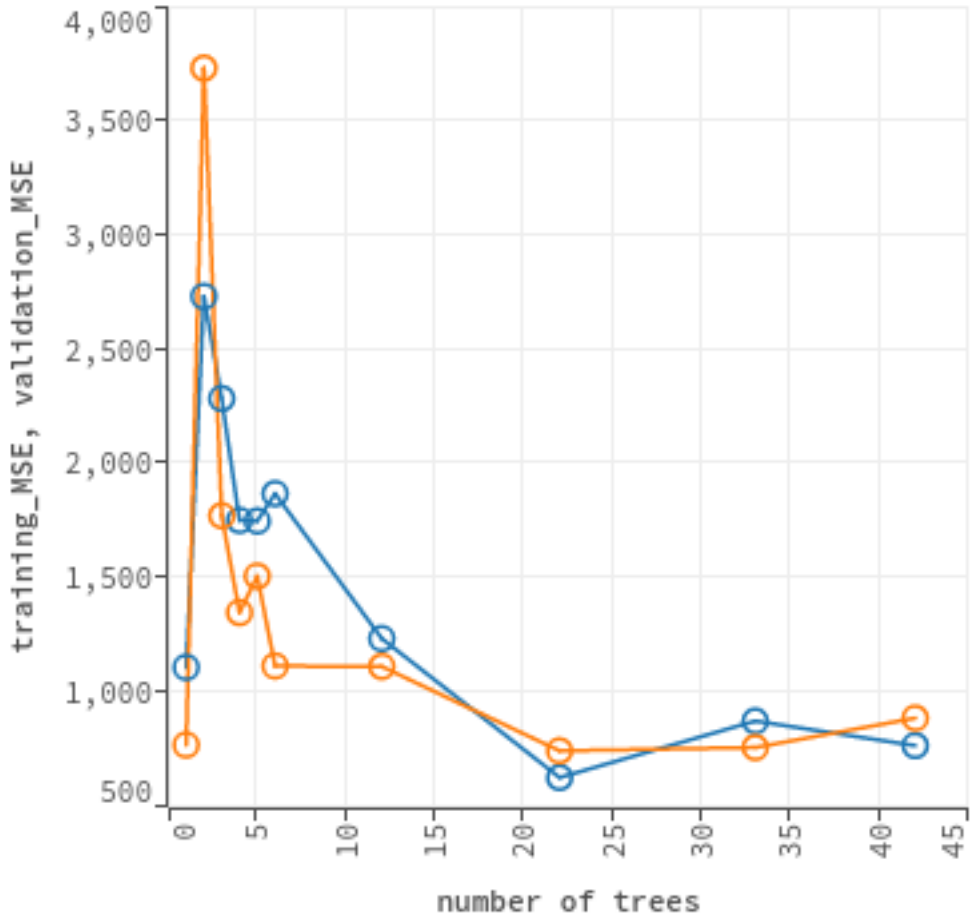
Notamos que as três novas variáveis adicionadas se encontram no Top 5 das variáveis que mais contribuíram para a criação do modelo. Podemos afirmar então, que as novas variáveis influenciaram positivamente o modelo.

Modelo para prever o preço de um certo produto por dia

Vamos agora repetir os passos para criar um novo modelo, para prever o preço de um certo produto por dia.

```
##
  |
  |                                                                      |   0%
  |
  |=                                                                     |   2%
  |
  |====                                                                  |   6%
  |
  |======                                                                |  10%
  |
  |=========                                                             |  14%
  |
  |============                                                          |  18%
  |
  |==============                                                        |  22%
  |
  |==================                                                    |  26%
  |
  |======================                                                |  32%
  |
  |========================                                              |  36%
  |
  |============================                                          |  42%
  |
  |================================                                      |  48%
  |
  |==================================                                    |  52%
  |
  |====================================                                  |  56%
  |
  |========================================                              |  62%
  |
  |============================================                          |  66%
  |
  |==============================================                        |  68%
  |
  |================================================                      |  72%
  |
  |====================================================                  |  78%
  |
  |========================================================              |  84%
  |
  |==========================================================            |  86%
```

```
  |
  |=============================================================      |  92%
  |
  |============================================================== |  98%
  |
  |================================================================| 100%
```

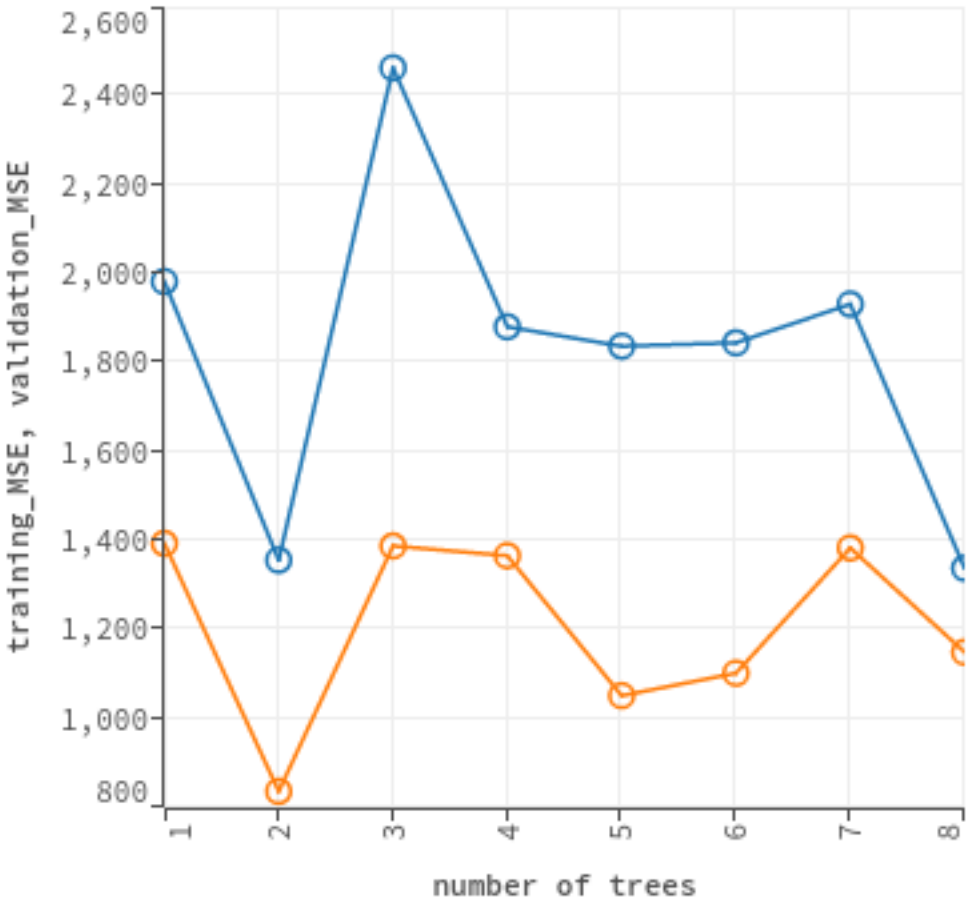


É possível notar que o após 17 árvores o modelo se torna especialista. Por esse motivo foi criado um novo modelo com apenas 17 árvores.

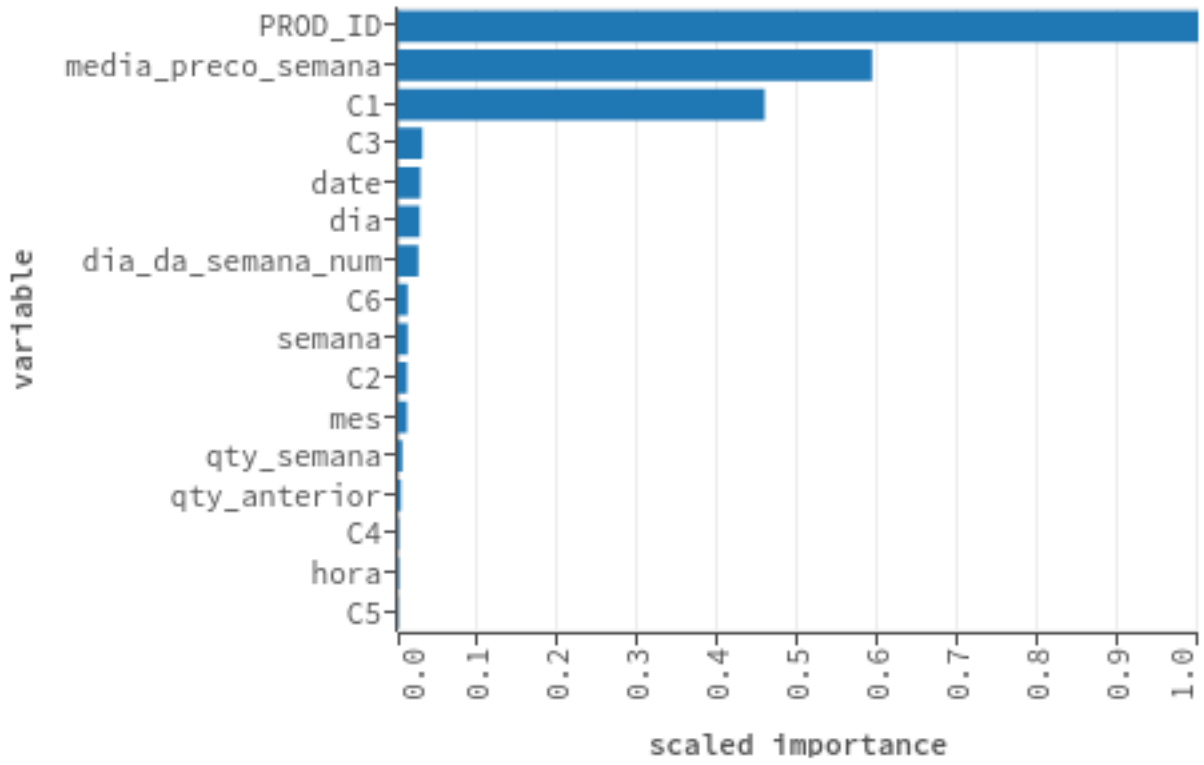
```
##
  |
  |                                                                 |   0%
  |
  |========                                                         |  12%
  |
  |===============                                                  |  24%
  |
  |=======================                                          |  35%
  |
  |==============================                                   |  47%
  |
  |=================================                                |  53%
```

```
|
|======================================                           |  65%
|
|================================================                 |  82%
|
|============================================================     |  94%
|
|=================================================================| 100%
```

Agora o modelo não sofre mais overfitting



É importante observar quais foram as variáveis quem mais contribuíram para a criação do modelo.



É interessante notar que o concorrente C1 se encontra no Top 3 das variáveis que mais contribuíram para a criação do modelo. Esse mesmo concorrente foi destacado como sendo o concorrente que possui os melhores preços antes mesmo da criação de qualquer modelo

Com isso temos dois modelos finais: * Um modelo para prever a quantidade diária que será vendida de cada produto * Um modelo para prever o preço diário de cada produto

---

Prevendo o preço e a quantidade

Podemos agora utilizar os dados de teste (que foi separado antes do modelo, e que é totalmente disjunto dos dados de treino) para prever qual seria o resultado em um mundo real.

```
##
  |
  |                                                                      |   0%
  |
  |======================================================================| 100%

##
  |
  |                                                                      |   0%
  |
  |======================================================================| 100%
```

```
##
  |
  |                                                                      |   0%
  |
  |======================================================================| 100%
```

Podemos agora simular qual seria o real valor Rsquared no mundo real. Para o modelo que prever a quantidade de compras, temos um Rsquared de:

```
## [1] 0.9768916
```

Para o modelo que prever a quantidade o preço, temos um Rsquared de:

```
## [1] 0.9964762
```

É possível notar que o Rsquared ficou acima de 0.97 para os dois casos. Esse valor no mundo real é considerado bastante alto. Cheguei nesse valor tão alto por se tratar de um conjunto de dados no qual não condiz com o de um mundo real (ter apenas 9 produtos, 6 concorrentes, etc). Todos esses fatores influenciaram para se ter um modelo com uma precisão tão alta.

Mesmo com um valor alto, ainda é possível melhorar sem causar overfitting. Se eu tivesse mais tempo, iria criar um novo modelo retirando alguns outlier, pois acredito que eles influênciam negativamente o modelo. Também iria retirar as variáveis que pouco contribuem para o modelo, deixando ele assim mais rápido.

Além disso, iria investigar e criar novas variáveis para melhorar o modelo. Algumas variáveis que tem potencial para melhorar o modelo:

- Vendabilidade - Proporção que o produto foi comprado em relação aos outros produtos

Acredito que ser importante para o modelo saber quais são os produtos com maior potencial de compras

- Top_meses - Proporção de vendas de um mês em relação aos outros meses.

Acredito que ser importante para o modelo saber quais são os meses que mais possuem compras (Ex: Dia das mãe, pais, etc)

- Dia_Mes - Mostra se o dia se encontra no começo, meio ou fim do mês

Acredito que algumas pessoas preferem comprar no começo do mês (quando recebem o salário) outras devem preferir comprar no fim do mês

Também acho importante utilizar outra biblioteca além do h2o, como por exemplo, o caret, que eu já usei aqui