Desafio | Cientista de dados - Oncase

Candidato: Tharso Rossiter

Análise descritiva dos dados (EDA)

Descrição do desafio

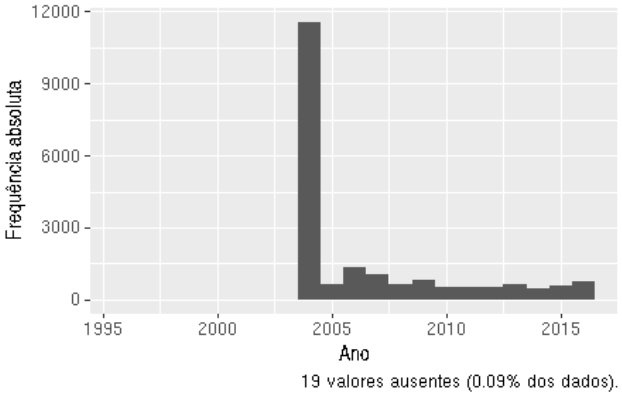
Nesse desafio, recebemos um arquivo JSON com dados de receitas. Entre as variáveis nesse conjunto de dados, estão:

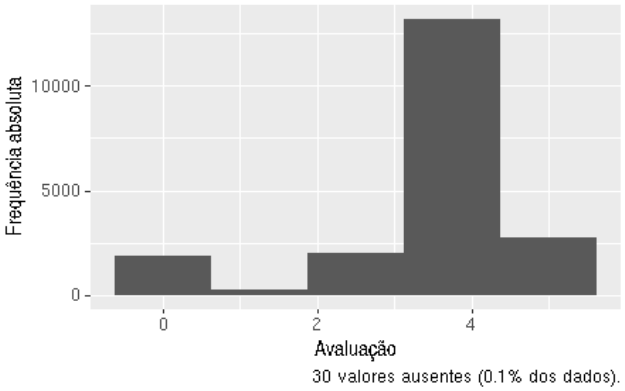
- **Variáveis de texto**: o título da receita, *tags* de categorias, descrição dos ingredientes e passo a passo da receita.
- **Variáveis numéricas**: quantidade de gordura, proteína, calorias e sódio em cada receita, assim como a avaliação da receita, numa escala de 0 a 5. As unidades de medida não foram informadas.
- A **data** na qual a receita foi publicada.

Análise exploratória de dados

Avaliações e ano de publicação

Abaixo, plotamos os gráficos referentes às avaliações e ano de publicação das receitas:



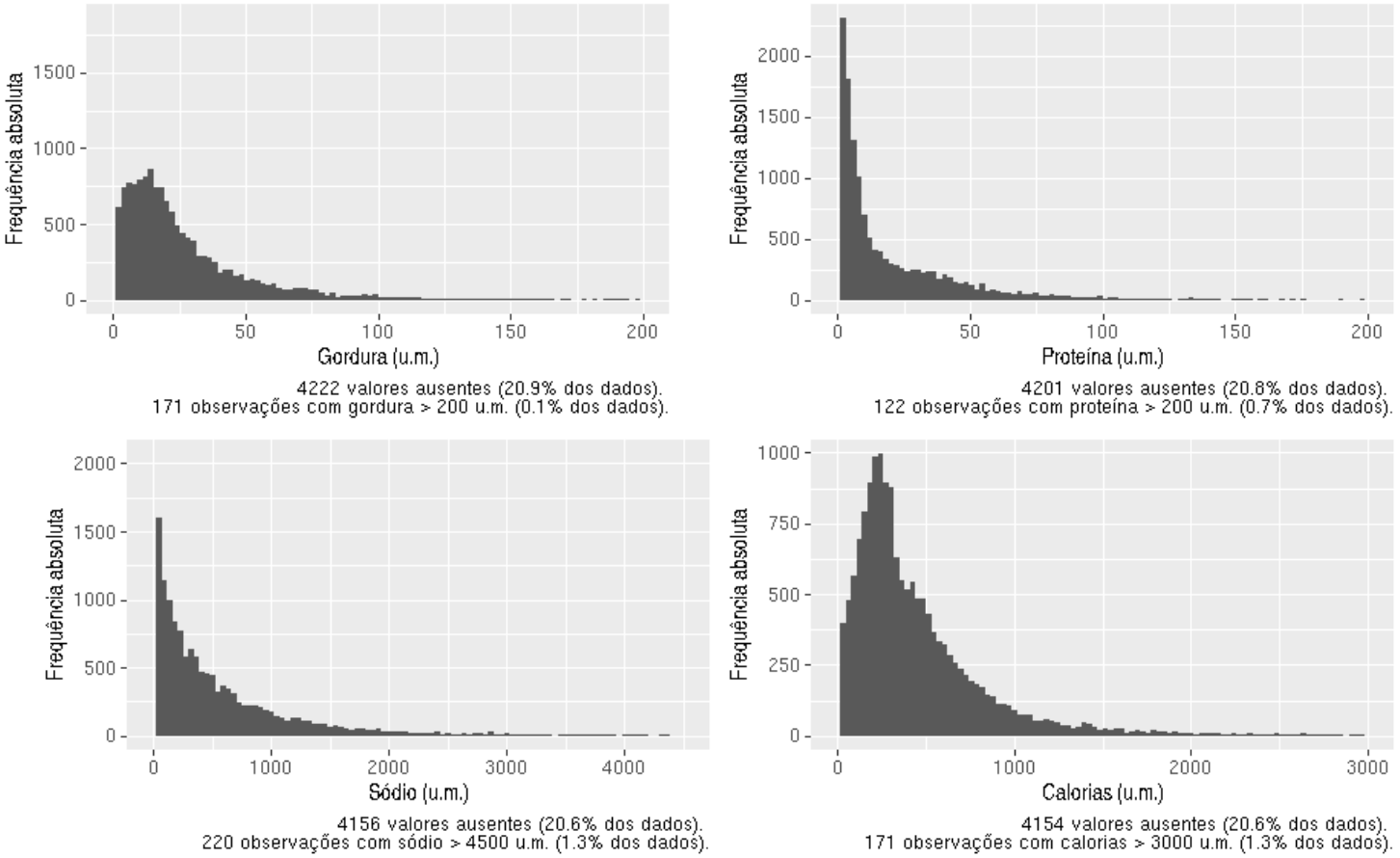


Alguns *insights* extraídos desses gráficos, além de análises posteriores são:

1. A primeira receita foi publicada em 1996. A maior parte das receitas (57% do total) foi publicada em 2004.
2. A maioria das receitas (33% do total) tem avaliação entre 3 e 4 estrelas.

Valor nutricional

Abaixo, plotamos gráficos referentes ao valor nutricional das receitas.



Alguns *insights* extraídos desses gráficos são:

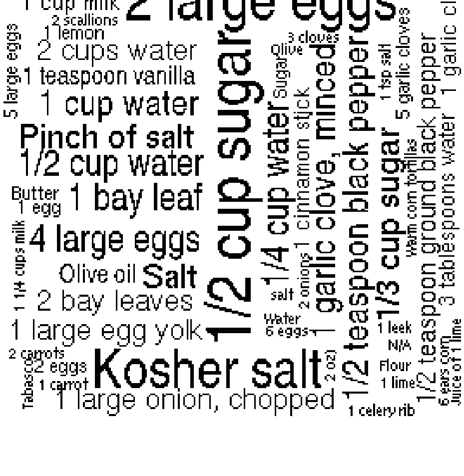
1. A maioria das receitas tem menor valor nutricional para todas as variáveis. Uma hipótese possível é que isso ocorre porque a maioria das receitas é elaborada para poucas porções. Outra hipótese, não-excludente com a primeira, é de que boa parte das receitas são dietéticas.

Análise de texto

Abaixo, plotamos nuvens de palavras para o título, categorias e ingredientes das receitas (nessa ordem):


cream
cheese
grilled
cake
with
pie
fresh
sweet
mustard
crust
dried
ice
bell
blue
mint
lime
herb
oil
torte
crab
tofu
kale
mini
bok
bay
fig
date
casserole
pot
best
rouille
rich
jelly
gin

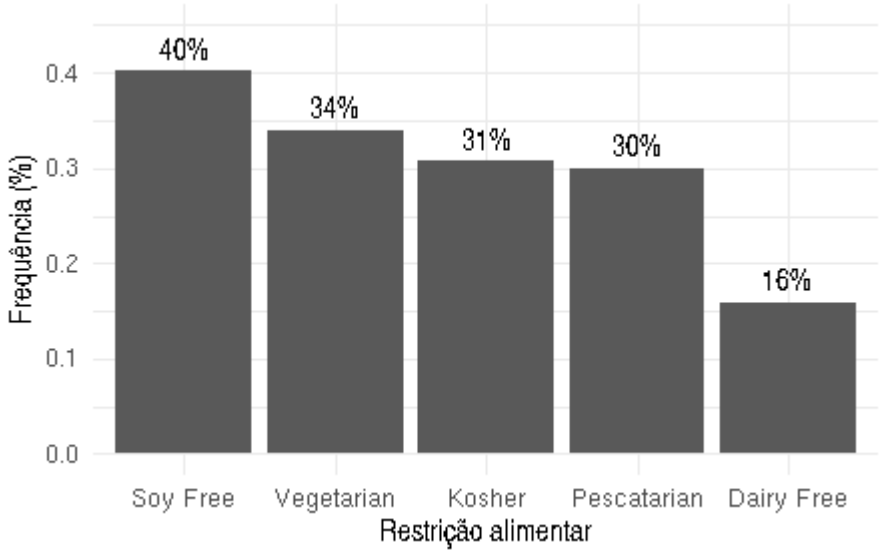

Soy Free
Vegetarian
Pescatarian
Gourmet
Kosher
Bake
Summer
Winter
Fruit
Almond
Egg
Port
Pork
Fall
Ginger
Potato
Walnut
Fish
Cake
Chile
Dinner
Dessert
Dairy
Free
Lemon
Herb
Party
Yogurt
Corn
Mint
Duck
Wok
Quince
Rice
Pie
Tea
Boil
Nut
Chill
Rum
Side
Chicken
Beef
Self
Dip
Semolina
Cashew
Brine
Sake
Prune
Meatball


2 large eggs
1/2 cup sugar
Kosher salt
Coarse kosher salt
Milk
1/2 cup dry red wine
lime wedges
1 cup milk
2 scallions
1 lemon
2 cups water
1 teaspoon vanilla
1 cup water
Pinch of salt
1/2 cup water
Butter
1 egg
1 bay leaf
4 large eggs
Olive oil
Salt
2 bay leaves
1 large egg yolk
2 carrots
2 eggs
1 carrot
Tabasco
1 large onion, chopped
5 large eggs
1 1/4 cups milk
1/4 cup water
1 cinnamon stick
1 garlic clove, minced
1 teaspoon black pepper
1/3 cup sugar
1/2 teaspoon black pepper
1/2 teaspoon ground black pepper
3 tablespoons water
1 garlic clove
5 garlic cloves
3 cloves
Olive
Sugar
1 tsp salt
2 onions
salt
Water
6 eggs
2 oz
1 leek
N/A
Flour
1 lime
1 celery rib
6 ears corn
Juice of 1 lime

Alguns *insights* desses gráficos são:

1. Há uma grande diversidade de receitas no conjunto de dados, já que as palavras frequentes nos títulos se referem a ingredientes presentes em pratos principais, sobremesas e *drinks*.
2. A maioria das receitas aparenta ser feita com ingredientes naturais ou minimamente, já que não há presença de nomes de marcas na nuvem de palavras dos títulos.
3. Boa parte das receitas é feita para pessoas com restrições alimentares. Isso porque a nuvem de palavras das categorias inclui diversos termos frequentes relacionados a restrições alimentares, como: "*soy free*" (livre de soja), "*vegetarian*" (vegetariano), "*pescatarian*" (pescetariano), "*dairy free*" (livre de laticínios) e "*kosher*" (comidas *kosher* são aquelas que obedecem as restrições alimentares contidas na Torá, o livro sagrado da religião judaica).

Com base no *insight* 3, investigamos a frequência de termos ligados a restrições alimentares nas categorias das receitas. Esses foram os resultados encontrados:

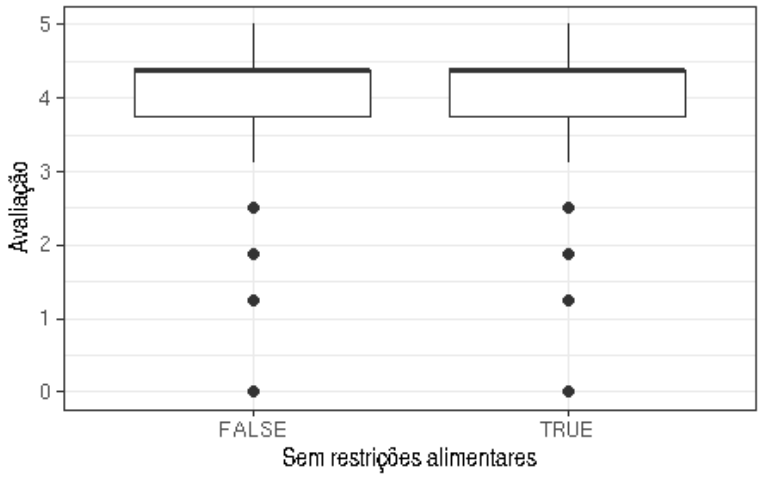


Avaliação das receitas

Vamos investigar quais variáveis podem afetar a avaliação das receitas.

Restrições alimentares

Primeiramente, analisamos se a avaliação da receita é afetada pelo fato dela obedecer às restrições alimentares mencionadas acima. O resultado está ilustrado no *boxplot* abaixo:

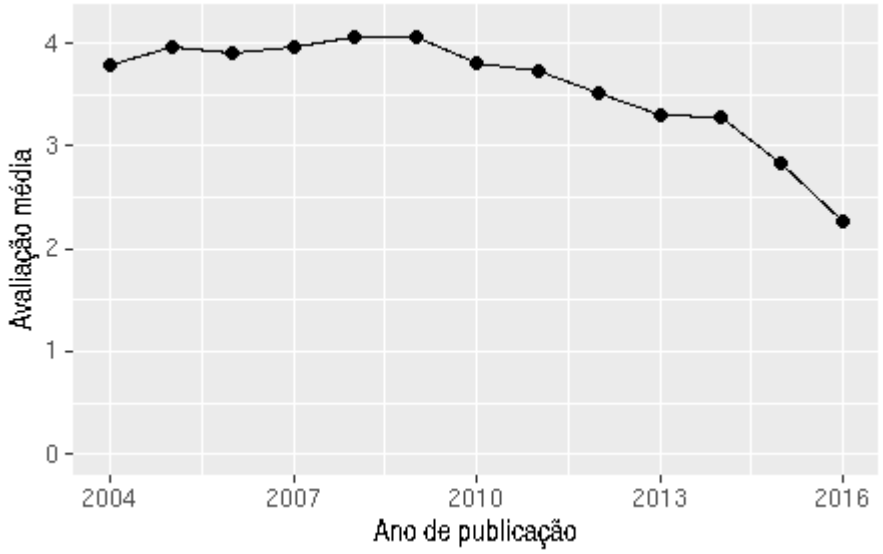


Os *boxplots* são muito parecidos. Logo, não há evidências de que restrições alimentares afetam a avaliação da receita, seja positivamente ou negativamente.

Fizemos a mesma análise para cada restrição alimentar, porém os resultados foram semelhantes. Escolhemos omitir esses boxplots, mas os códigos para gerá-los estão disponíveis no script no GitHub.

Ano de publicação

Analisamos também a média das avaliações das receitas a cada ano. Omitimos os anos antes de 2004, pois o número de receitas publicadas nestes anos é muito pequeno.



Nota-se que as receitas publicadas nos primeiros anos têm avaliações maiores que aquelas publicadas nos anos posteriores.

Palavras associadas a receitas de avaliação máxima

Abaixo, plotamos uma nuvem de palavras das palavras frequentes nos títulos das receitas com avaliação 5 (nota máxima).

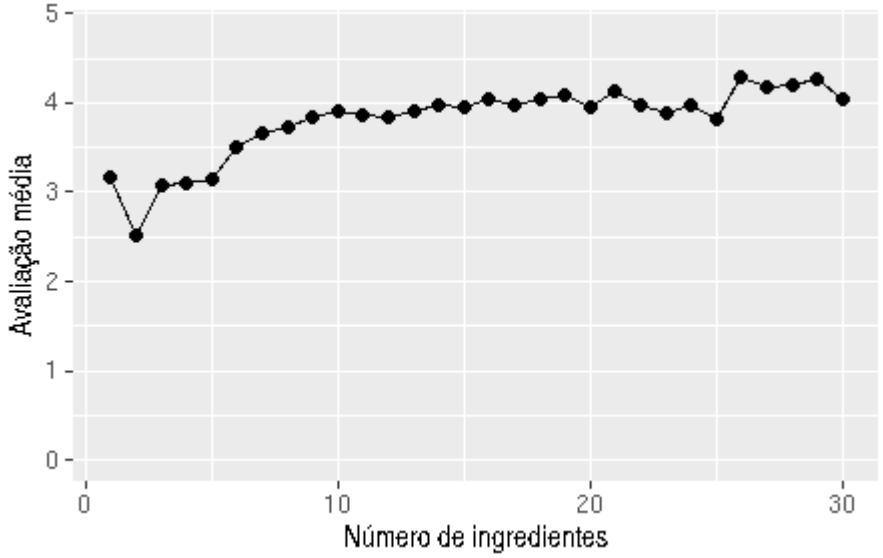
Nota-se que há uma grande frequência de:

- saladas (*salad*);
- comidas grelhadas (*grilled*);
- tortas (*pie*);
- cremes (*cream*);
- aioli (*alioli*), um molho de alho de origem francesa;
- receitas com manteiga (*butter*);
- e receitas com ingredientes naturais : milho, feijões, sangue (*corn*, *beans*, *blood*)

Número de ingredientes

Investigamos se o número de ingredientes em uma receita tem influência na sua avaliação. No gráfico abaixo, plotamos o número de ingredientes pela média da nota das receitas.

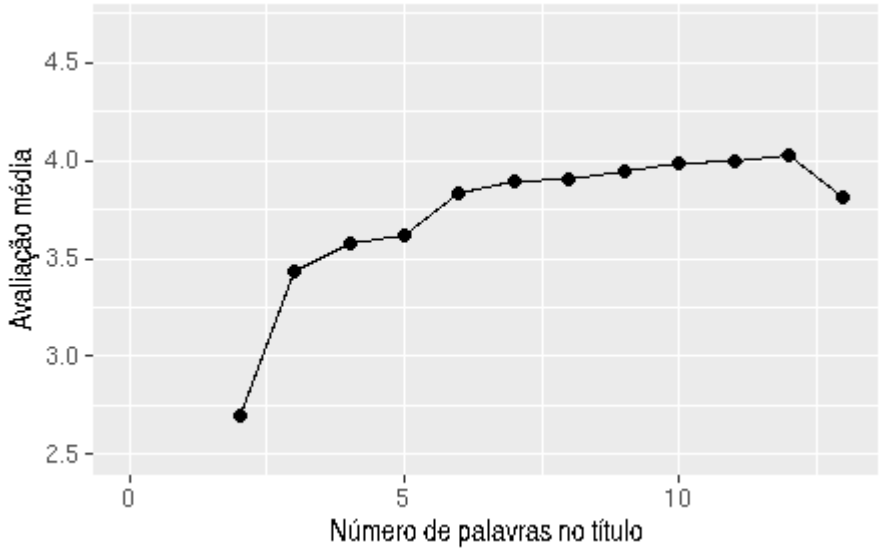
Omitimos as receitas com mais de 30 ingredientes porque o número de observações era muito baixo, o que torna a variância da média muito alta.



Observamos que receitas com mais ingredientes costumam ser mais bem avaliadas.

Tamanho do título

Analisamos também o efeito do tamanho do título das receitas. No gráfico abaixo, plotamos a média da avaliação das receitas de acordo com o número de palavras no título. Omitimos as observações com título maior que 14 palavras porque o número de observações era muito baixo, o que torna a variância da média muito alta.



Observamos que receitas com título mais longo tendem a ser mais bem avaliadas.

Comentários

A maior dificuldade desse desafio foi não haver um propósito claro da análise dos dados.

Além disso, não haviam dados da quantidade de porções para cada receita e nem das unidades de medida das variáveis numéricas, dificultando uma análise do valor nutricional, pois não há como calcular o valor de cada nutriente por porção.

Teste técnico de modelagem

Classificação

Descrição do desafio

Nesse desafio, recebemos um conjunto de dados com três variáveis:

- **target**: uma variável binária;
- **x1**: uma variável contínua e
- **x2**: uma variável contínua.

Não sabemos se essas variáveis têm interpretação real ou se seus valores foram simulados.

Nosso objetivo é comparar a performance de dois modelos de previsão da variável **target**:

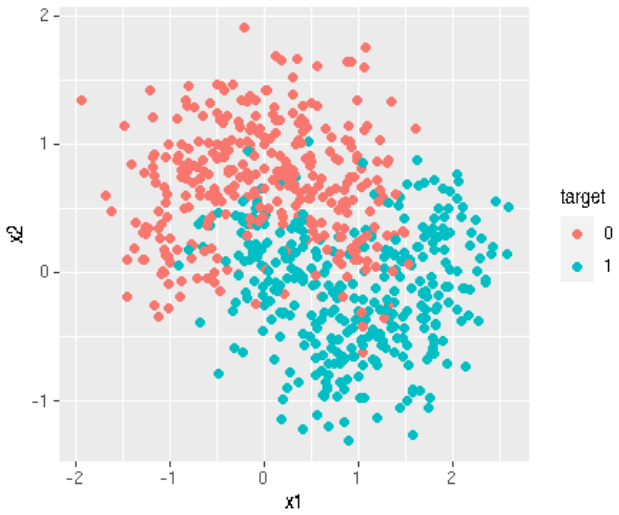
- um modelo que prevê a variável **target** apenas por meio da variável **x1**, que chamaremos de “modelo x1” e
- um modelo que prevê a variável **target** apenas por meio da variável **x2**, que chamaremos de “modelo x2”.

Análise exploratória dos dados

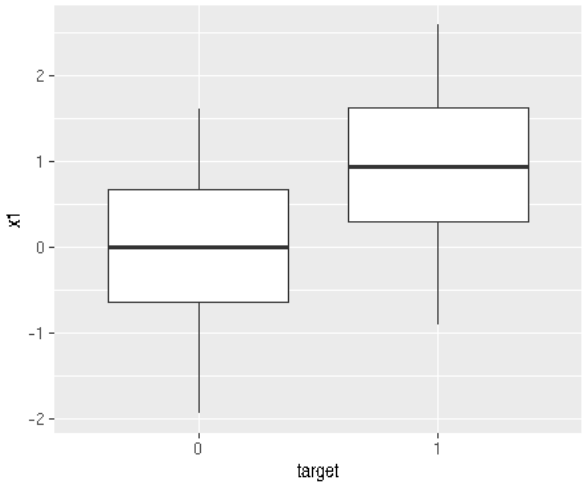
Plotamos um gráfico de dispersão das variáveis **x1** e **x2**. Cada ponto foi colorido de acordo com o valor assumido pela variável **target**.

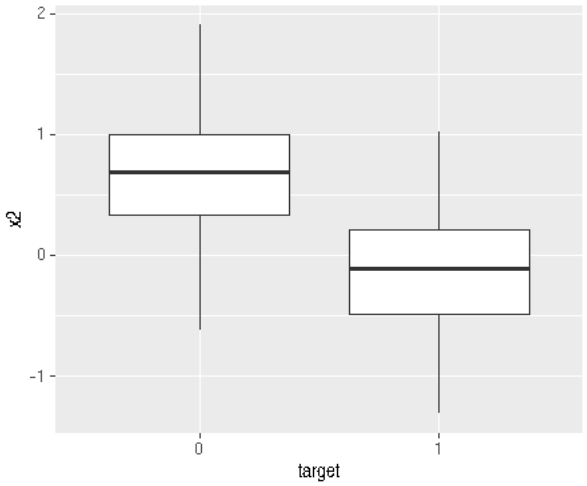
Alguns *insights* que temos desse gráfico são:

1. Não parece haver correlação significativa entre **x1** e **x2**.
2. Valores de **target** iguais a 0 são associados a valores menores de **x1** e maiores de **x2**.
3. Valores de **target** iguais a 1 são associados a valores maiores de **x1** e menores de **x2**.
4. Tanto **x1** como **x2** parecem ser significativos para prever o valor de **target**.



Podemos visualizar os *insights* 2 e 3 de maneira mais clara nos *boxplots* abaixo:

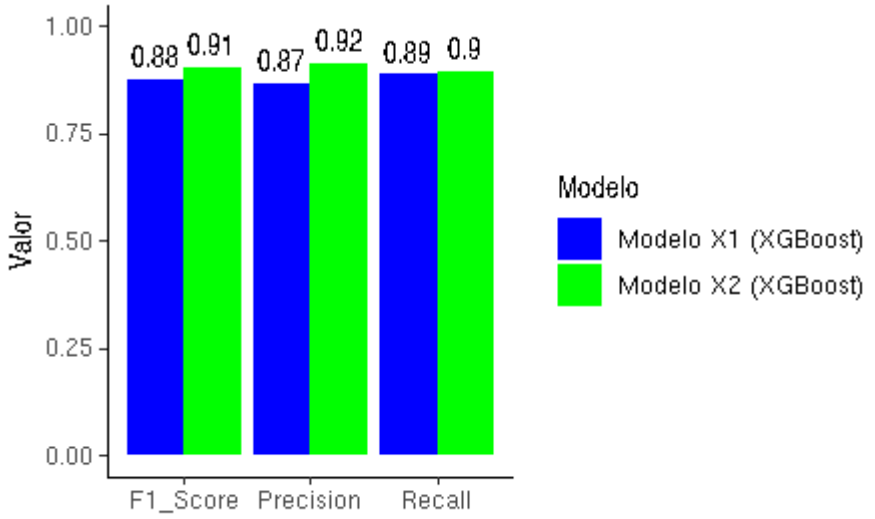




Modelagem e comparação de modelos

Modelamos esses dados usando XGBoost. Criamos um modelo que usa apenas a variável **X1** e outro que usa apenas a variável **X2**.

Comparamos os valores de *precision*, *recall* e F1-score de ambos os modelos. Os resultados estão exibidos no gráfico abaixo:



Concluímos que o modelo que usa apenas a variável **X2** é mais eficiente que o modelo que usa apenas **X1** porque:

1. O modelo x2 possui um valor maior de *precision*. Isso significa que, usando esse modelo, a proporção de observações classificadas como positivas que de fato são positivas é maior que a do modelo x1.
2. O modelo x2 possui um valor maior de *recall*. Isso significa que, usando esse modelo, a proporção de observações positivas corretamente identificadas como positivas é maior que a do modelo x1.
3. O modelo x2 possui um valor maior de *F1-score*, que é uma média harmônica entre o *precision* e o *recall*.

Comentários

Esse desafio teve objetivos claros; os dados estavam completos e o modelo usado obteve um resultado bastante satisfatório.

Outros modelos que poderiam ser utilizados seriam:

- Regressão logística
- Árvore de decisão
- *Random Forest*
- *Support Vector Machines*

Testamos todos esses modelos, mas seus resultados foram inferiores aos obtidos com o XGBoost. Os scripts estão disponíveis no repositório do GitHub.

Previsão

Descrição do desafio

Neste desafio, recebemos um conjunto de dados com observações de quantidades de diversos produtos durante um período de tempo de 364 dias. O número de observações de cada produto é diferente. Não sabemos o que são esses produtos ou se a “quantidade” é uma quantidade em estoque ou de vendas.

Nosso objetivo é:

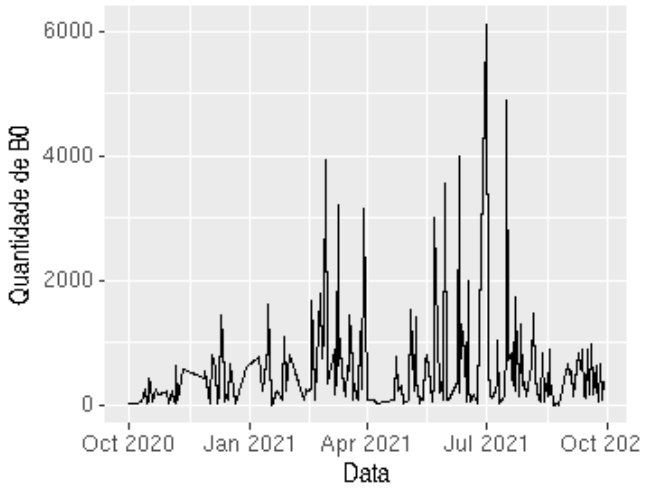
- escolher algum dos produtos da lista e
- propor um modelo de séries temporais para prever a quantidade do produto.

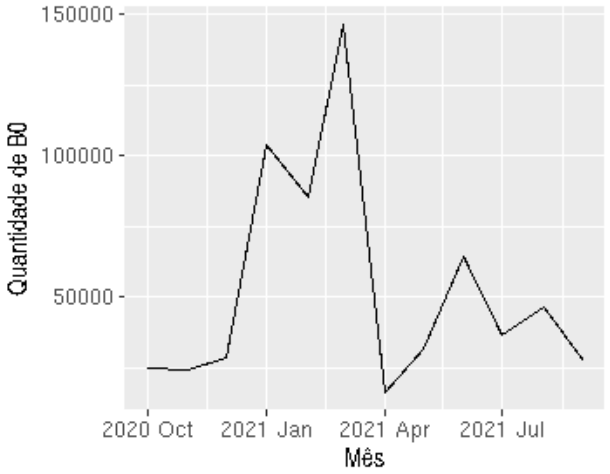
Decidimos por:

- escolher o produto B0 e
- prever a quantidade de B0 para os últimos quatro meses.

Análise exploratória dos dados

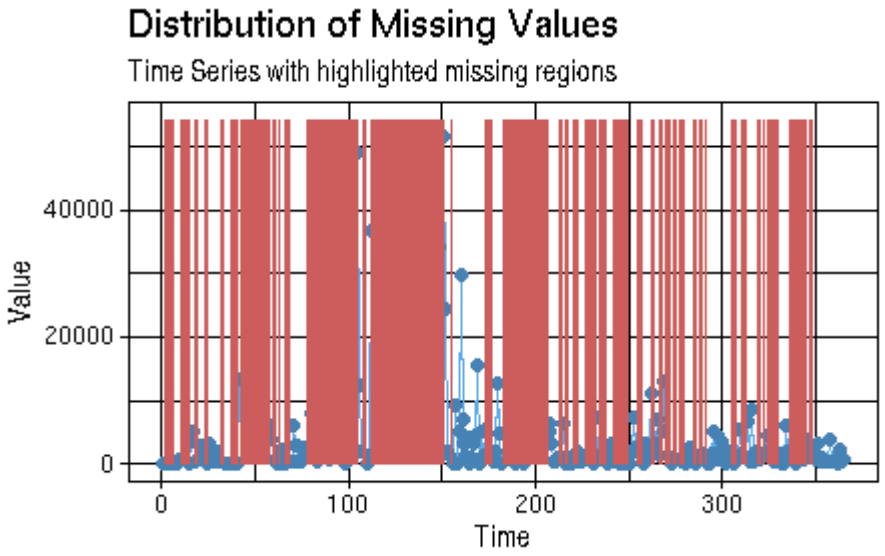
Abaixo, plotamos a quantidade de B0 a nível diário e a nível mensal.





Há vários valores faltantes implícitos nesse conjunto de dados. Isto é, não há observações para várias datas no espaço de tempo decorrido do início ao fim das observações.

Tornamos esses valores ausentes explícitos. Ou seja, adicionamos as datas ausentes no conjunto e preenchemos o valor da variável de quantidade do produto como **NA**. Abaixo, plotamos os dados ausentes em vermelho.

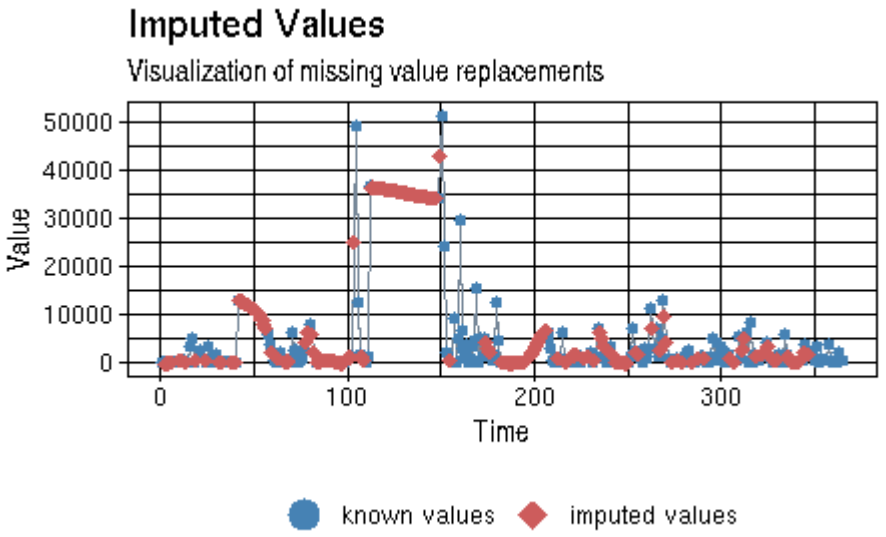


Alguns *insights* obtidos nestes gráficos são:

1. Há uma alta variabilidade a nível diário da quantidade de B0.
2. Houve uma demanda maior por B0 nos primeiros meses de 2021, seguido de uma queda em abril/2021 e nos meses seguintes.
3. Há muitos dados ausentes. Esses dados estão bem espalhados pelo espaço de tempo observado, mas há alguns períodos de tempo com vários valores ausentes consecutivos.

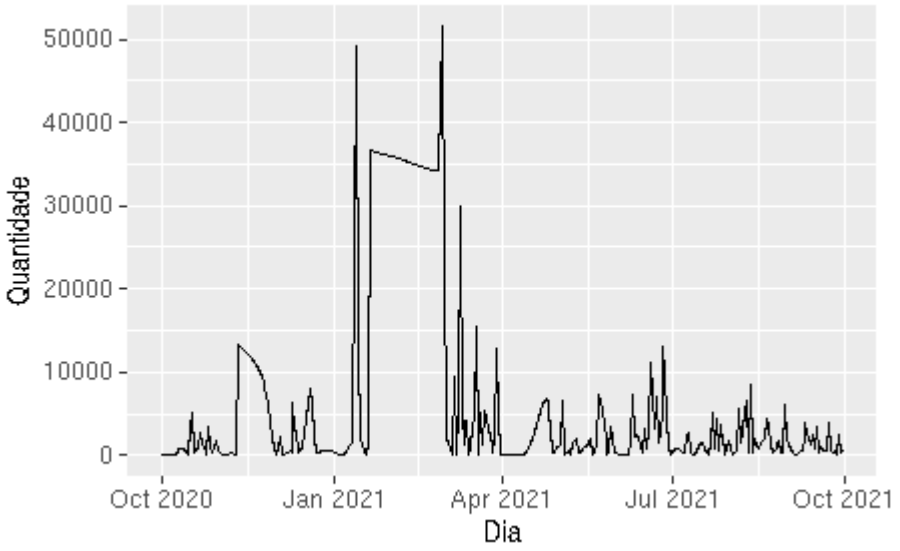
Optamos por interpolar os dados usando a interpolação de Stineman. Optamos por esse método porque ele é mais adequado para dados sem tendência e com picos, o que é o caso desses dados.

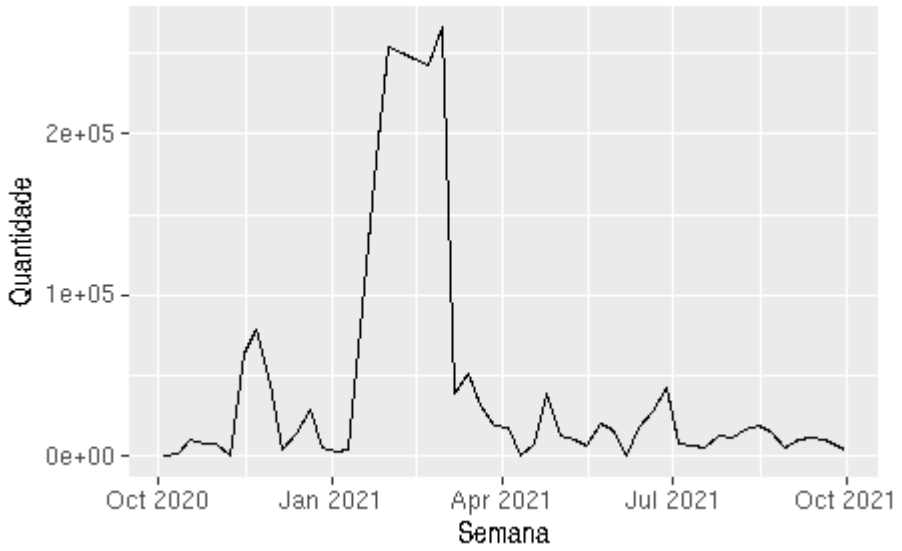
Abaixo, plotamos um gráfico dos dados reais e interpolados. Dados reais são representados por pontos azuis e interpolados, por losangos vermelhos

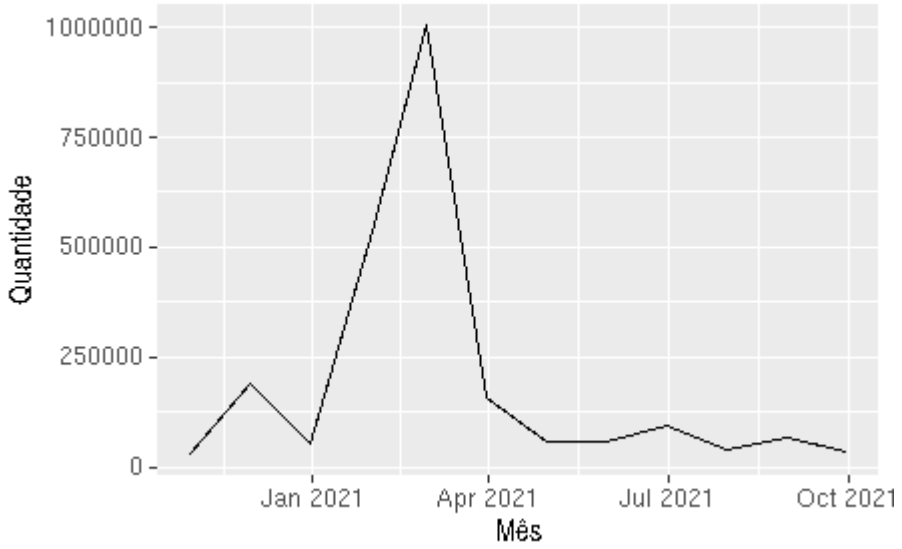


Em cinco observações, os valores interpolados foram negativos. Como isso não é possível para uma variável de quantidade de produto, substituímos os valores negativos por zero.

Abaixo, plotamos o gráfico de série temporal por dia, semana e mês com os dados interpolados.







Modelagem

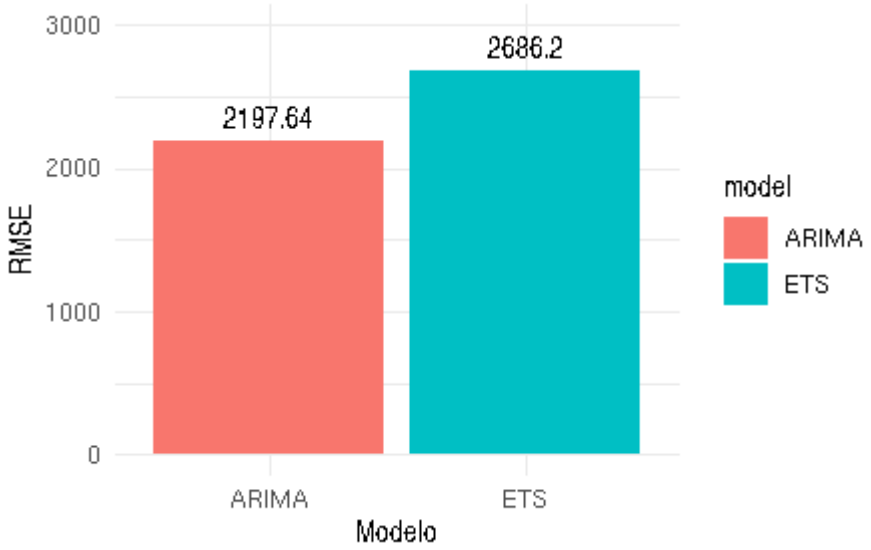
Optamos por modelar os dados semanais porque não têm uma variância tão grande quanto os diaŕios e não têm uma quantidade de dados tão baixa quanto os mensais.

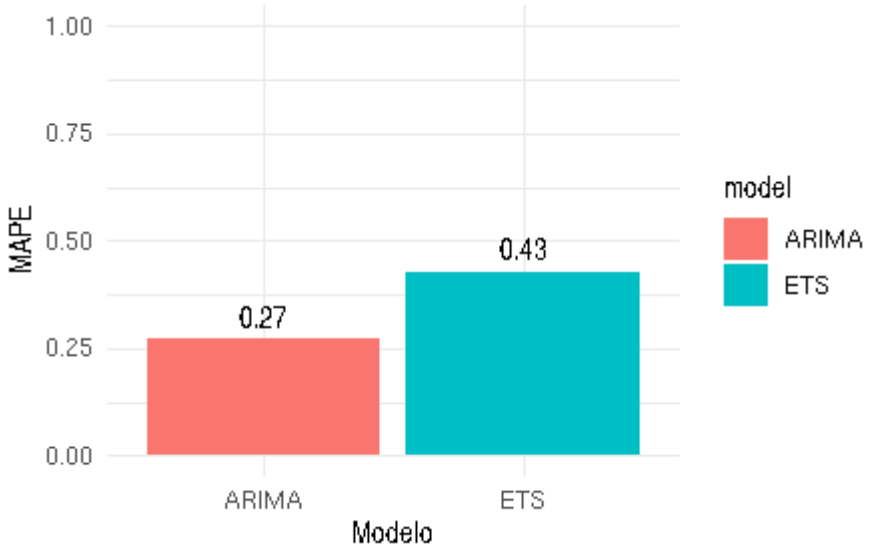
Removemos dados relativos às quatro semanas e modelamos os dados usando ETS e ARIMA.

O melhor modelo ETS, usando a maximização de verossimilhança, é o ETS(M, Ad, N). Isto é, com componente sazonal multiplicativo, componente de tendência aditivo e sem componente de nível. Também testamos o melhor modelo usando a minimização dos erros quadráticos, mas os resultados foram inferiores.

O melhor modelo ARIMA, de acordo com a minimização do AIC, é o modelo ARIMA(1,0,0). Isto é, um modelo com componente autorregressivo de ordem 1, sem componentes de média móvel e sem diferenciar a série temporal.

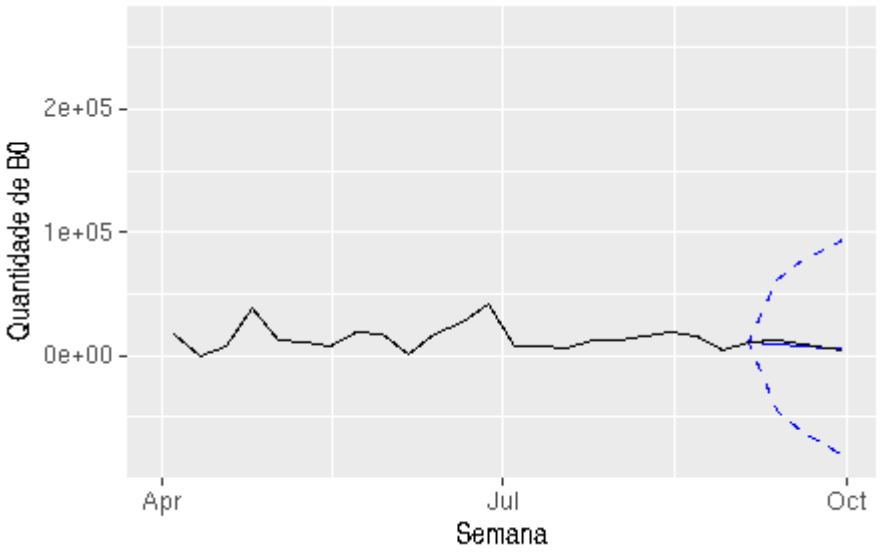
Comparamos os modelos de acordo com as métricas propostas.





Observamos que o modelo ARIMA obteve um desempenho melhor.

Plotamos um gráfico com os valores preditos e reais da série temporal. O valor predito está representado pela linha azul contínua e os intervalos de confiança de 80%, pelas linhas azuis tracejadas. Para facilitar a visualização, exibimos apenas os dados a partir de abril de 2021.



Observamos que os valores preditos foram muito próximos dos previstos e sempre dentro do intervalo de confiança.

Os valores preditos, atuais e intervalos de confiança estão disponíveis na tabela abaixo:

Valor predito	Lim. inf. 80	Lim. sup. 80	Valor real
9025.331	-43887.50	61938.16	12188
7870.348	-62335.11	78075.80	9915
6863.170	-74055.31	87781.64	6789
5984.881	-82213.40	94183.17	3720

Comentários

Outras técnicas que poderiam ser aplicadas seriam:

- Redes neurais. Não usamos esse modelo porque é demasiadamente complexo para um problema relativamente simples.

Regressão

Descrição do desafio

Nesse desafio, recebemos um conjunto de dados com oito variáveis:

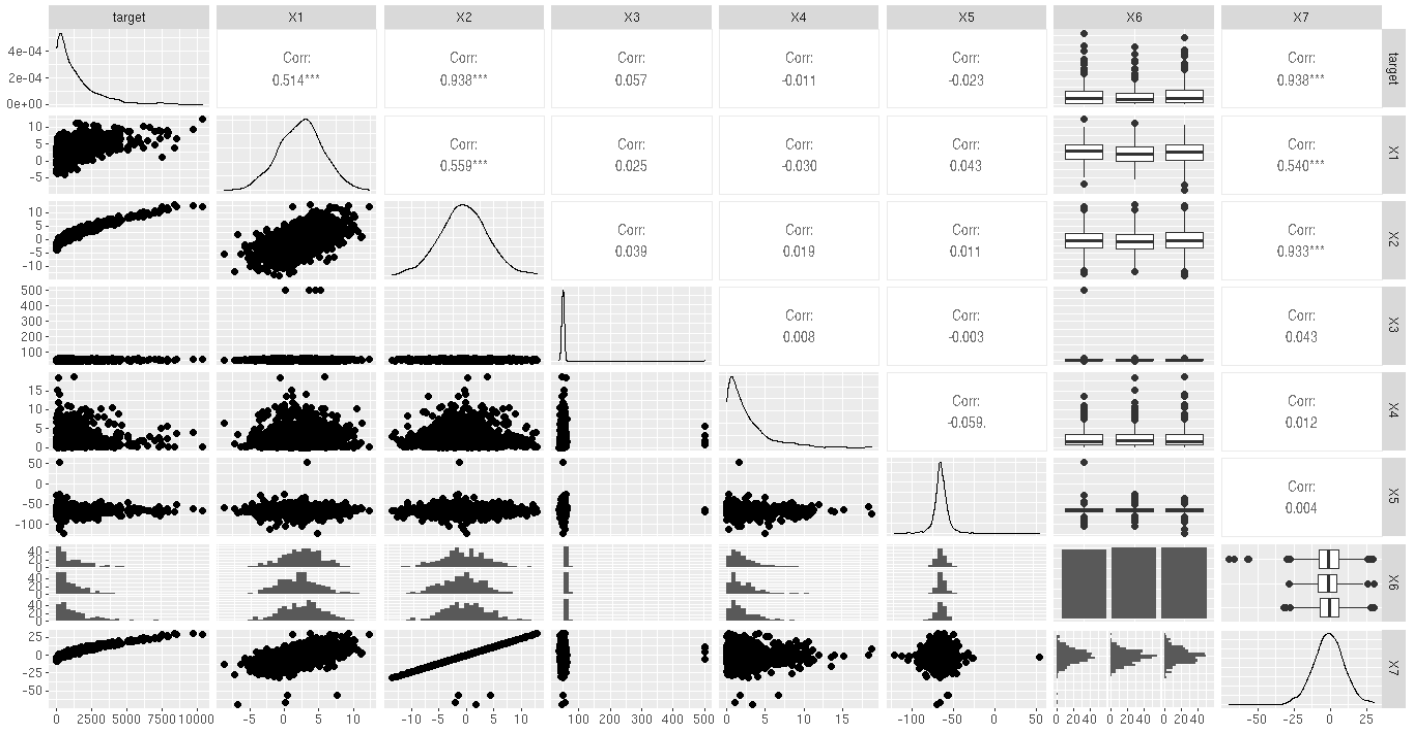
- **target**: uma variável contínua que queremos prever;
- **X1**, **X2**, **X3**, **X4**, **X5** e **X7**: variáveis contínuas;
- **X6**: uma variável que assume apenas 3 valores e por isso, foi considerada categórica.

Não sabemos se essas variáveis têm interpretação real ou se seus valores foram simulados.

Nosso objetivo é criar um modelo de regressão para modelar a variável **target** usando as outras variáveis ou um subconjunto delas.

Análise exploratória de dados

Plotamos um conjunto de gráficos para analisar a relação da variável **target** com as variáveis independentes e das variáveis independentes entre si.



Alguns *insights* extraídos desses gráficos são:

1. Há uma correlação muito alta entre as variáveis **X2** e **target**, assim como **X7** e **target**, como evidenciado pelos gráficos de dispersão e pelos valores das correlações.
    a. Essa correlação é aproximadamente linear, o que indica que é apropriado modelarmos esses dados com uma regressão linear.

2. **X1** tem correlação considerável com **target**, **X2** e **X7**, como evidenciado pelo gráfico de dispersão e pelo valor da correlação**.**
3. As outras variáveis possuem uma correlação baixa entre si mesmas e também com a variável **target**, como evidenciado pelos gráficos de dispersão e pelos valores das correlações.

Após uma análise mais profunda, vimos que, na exceção de uma observação, as variáveis **X2** e **X7** são linearmente dependentes. Essas variáveis seguem a relação **X7** = **X2** * **2,35.**

Modelagem e comparação de modelos

A dependência linear entre as variáveis **X7** e **X2** implica que:

1. As duas variáveis não podem estar juntas num mesmo modelo de regressão linear.
2. Um modelo de regressão que usa apenas uma das duas variáveis fará previsões idênticas.

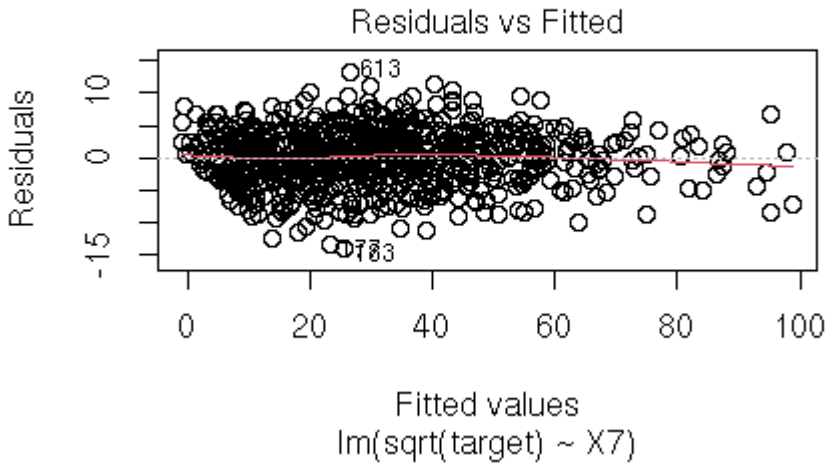
Após a análise de vários modelos, concluímos que um dos modelos mais adequados para esses dados é dado pela fórmula $\sqrt{target}$ = **X7** + ε, em que ε é um erro com distribuição normal. Isso porque ele se adequa às suposições do modelo linear e possui R^2 alto, aproximadamente 0,95.

Para criar esse modelo, supomos que a variável **target** só pode assumir valores não-negativos. Essa suposição é razoável porque tanto no conjunto de dados de teste como no de treino, **target** não assume nenhum valor negativo.

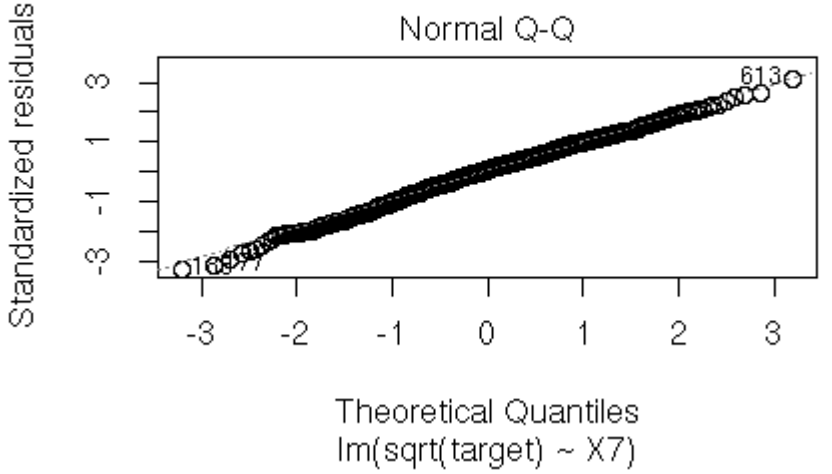
Decidimos usar o modelo com apenas uma variável porque:

- A variável **X1**, que era a única variável dependente com correlação significativa com **target**, possui correlação significativa com **X7**. Adicioná-la poderia trazer multicolinearidade ao modelo.
- Fora **X1**, **X2** e **X7**, as outras variáveis não apresentam impacto significativo em **target**.
- Como o valor de R^2 do modelo com apenas uma variável é alto, há um risco de *overfitting* se acrescentarmos mais variáveis ao modelo.

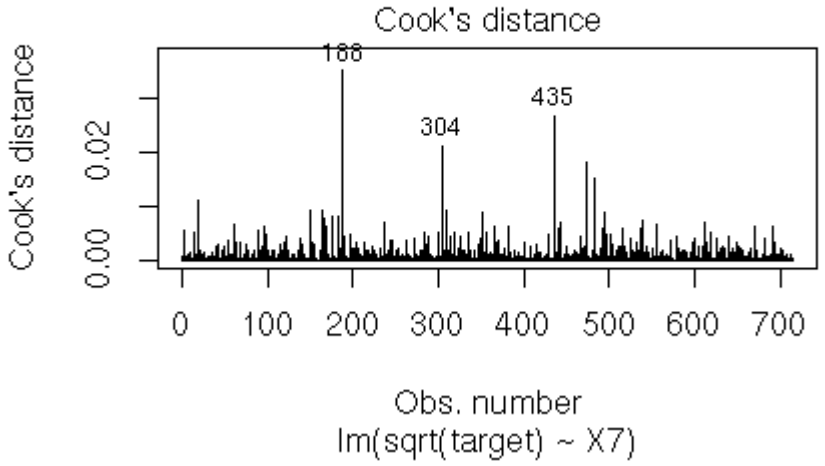
Abaixo, plotamos os gráficos de diagnóstico dessa regressão linear.



No gráfico de resíduos por valor predito, vemos que não há evidência de heterocedasticidade ou de não-linearidade.



No gráfico QQ dos resíduos, vemos que os resíduos seguem distribuição normal.



Finalmente, no gráfico da distância de Cook, vamos que há apenas três observações que são consideradas influentes.

Uma tabela com os valores de RMSE, R^2 e correlação entre valores preditos e reais com o modelo sugerido está abaixo:

RMSE	**R^2**	**Correlação**
251.20	0.95	0.98

Comentários

Esse desafio teve objetivos claros. Haviam valores ausentes no conjunto de dados, porém isso não impossibilitou a criação de um modelo eficiente.

Não sabemos a interpretação da variável **target**. Por isso, tivemos que supor que ela não pode assumir valores negativos. Se soubéssemos sua interpretação, essa suposição poderia ser descartada.

Poderíamos ter usado modelos baseados em árvores, como árvores de regressão, árvores de decisão, *Random Forest*, XGBoost, entre outros. No entanto:

- Esses modelos são mais complexos que a regressão linear.
- Esses modelos não são treinados para otimizar o R^2, que era uma das métricas solicitadas pelo desafio.
- Esses modelos não supõem uma relação linear entre as variáveis, mas nesse caso, essa suposição é apropriada.