

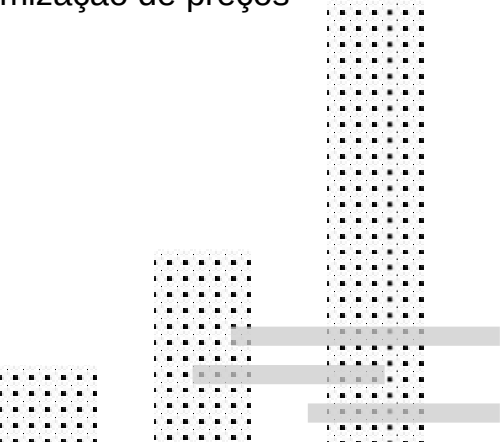
Precificação em varejo

Sérgio Anéfalos Pereira

Caso de estudo – Grupo SBF
Maio 2024

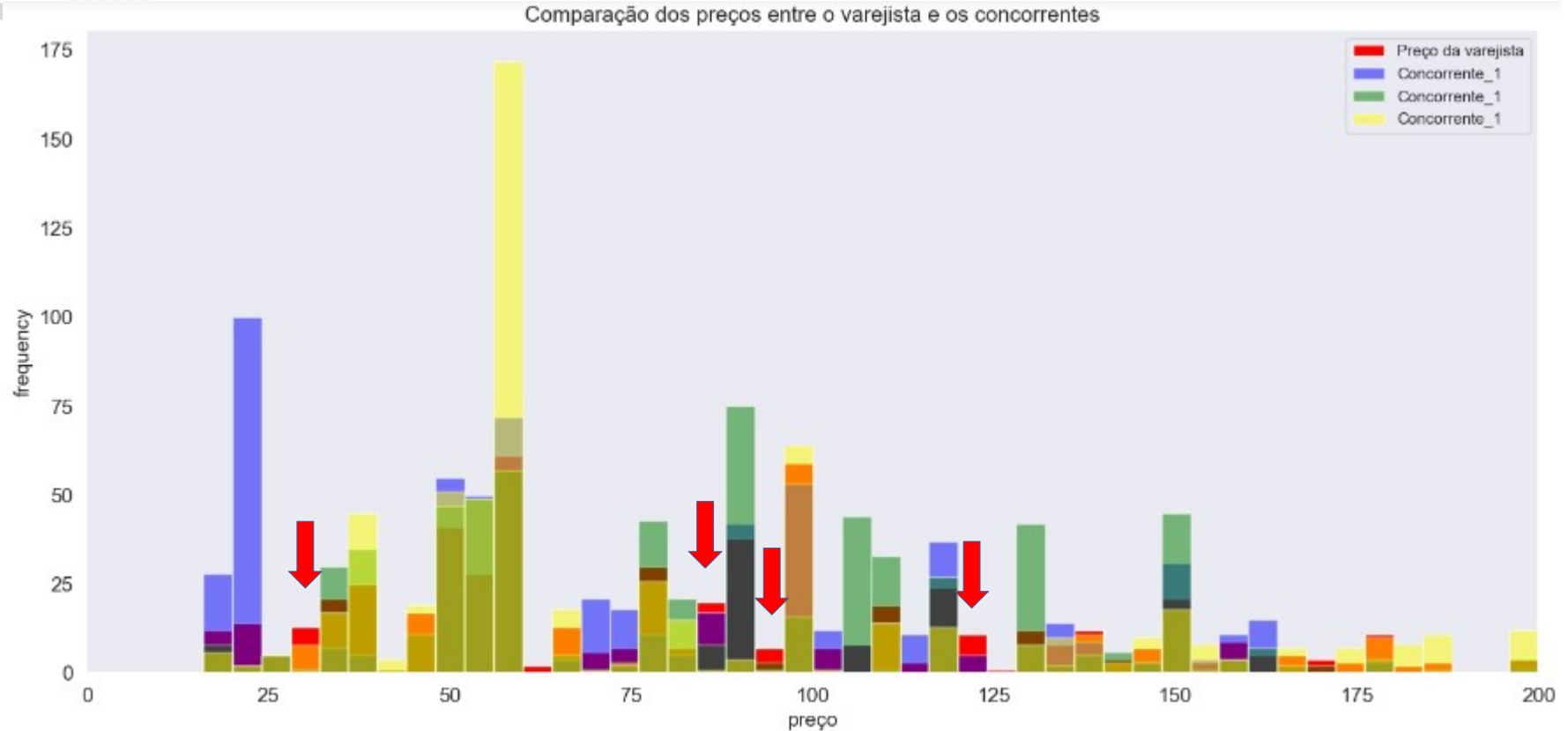


Table of contents

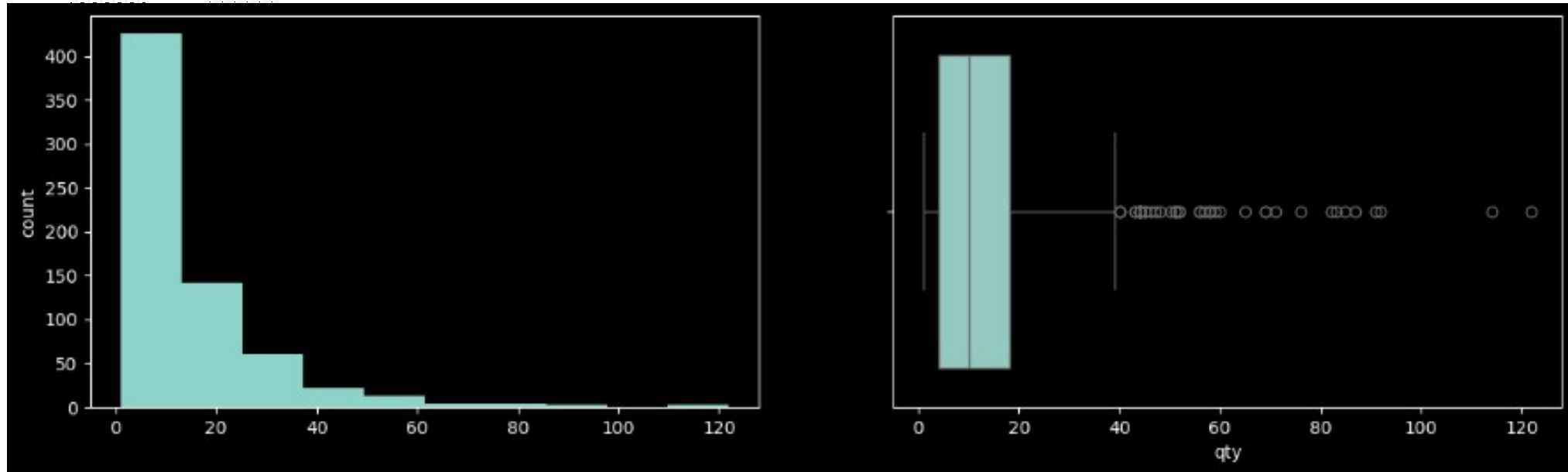
- análise exploratória dos dados
 - estudo da competitividade da varejista nas diversas categorias de produtos
 - análise do impacto das variáveis no preço de venda
 - análise de demanda de produtos
 - criação de um modelo preditivo para os preços dos produtos
 - modelo de otimização da política de preços praticada
 - sugestão de novas variáveis melhorar as previsões de preço e demanda e otimização de preços
 - pipeline de dados para os modelos criados
 - pontos críticos de monitoramento da pipeline
 - Conclusão e impressões gerais
- 

O objetivo é realizar a análise exploratória dos dados para se entender melhor, entre outras coisas:

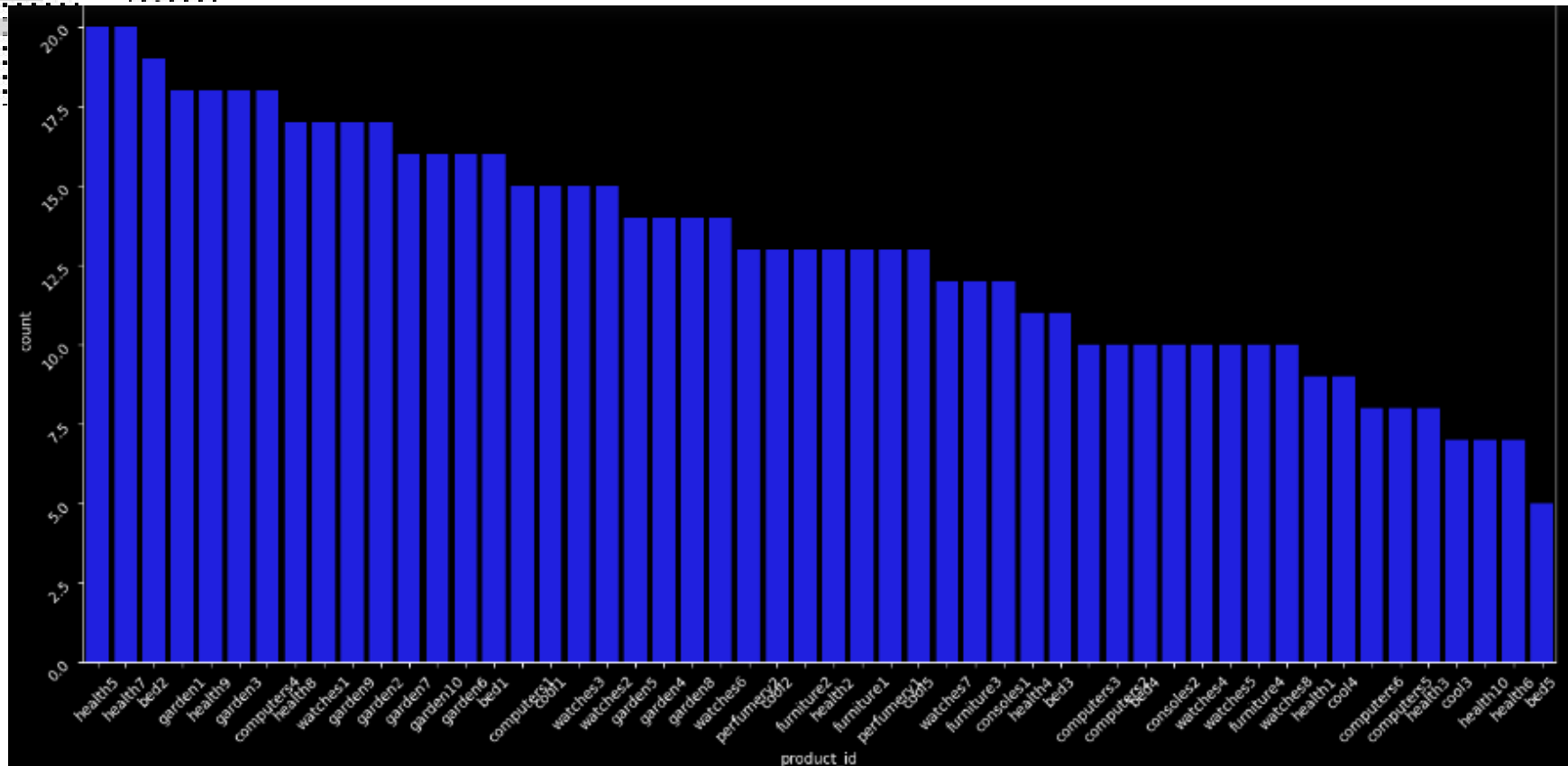
- Qualidade dos dados
- Existência de valores nulos (NULL)
- Valores faltantes (missing values)
- Range dos dados
- Identificação dos tipos de variáveis (dtype=object, int64, float64, etc)
- Dados duplicados
- Possíveis correlações
- Entre outras ...



- Comparação entre os preços da varejista com seus concorrentes
- Nota-se que no geral, os preços praticados pela varejista é inferior aos dos concorrentes. Apenas em alguns casos, a cor vermelha se destaca (indicados pelas setas vermelhas no gráfico)

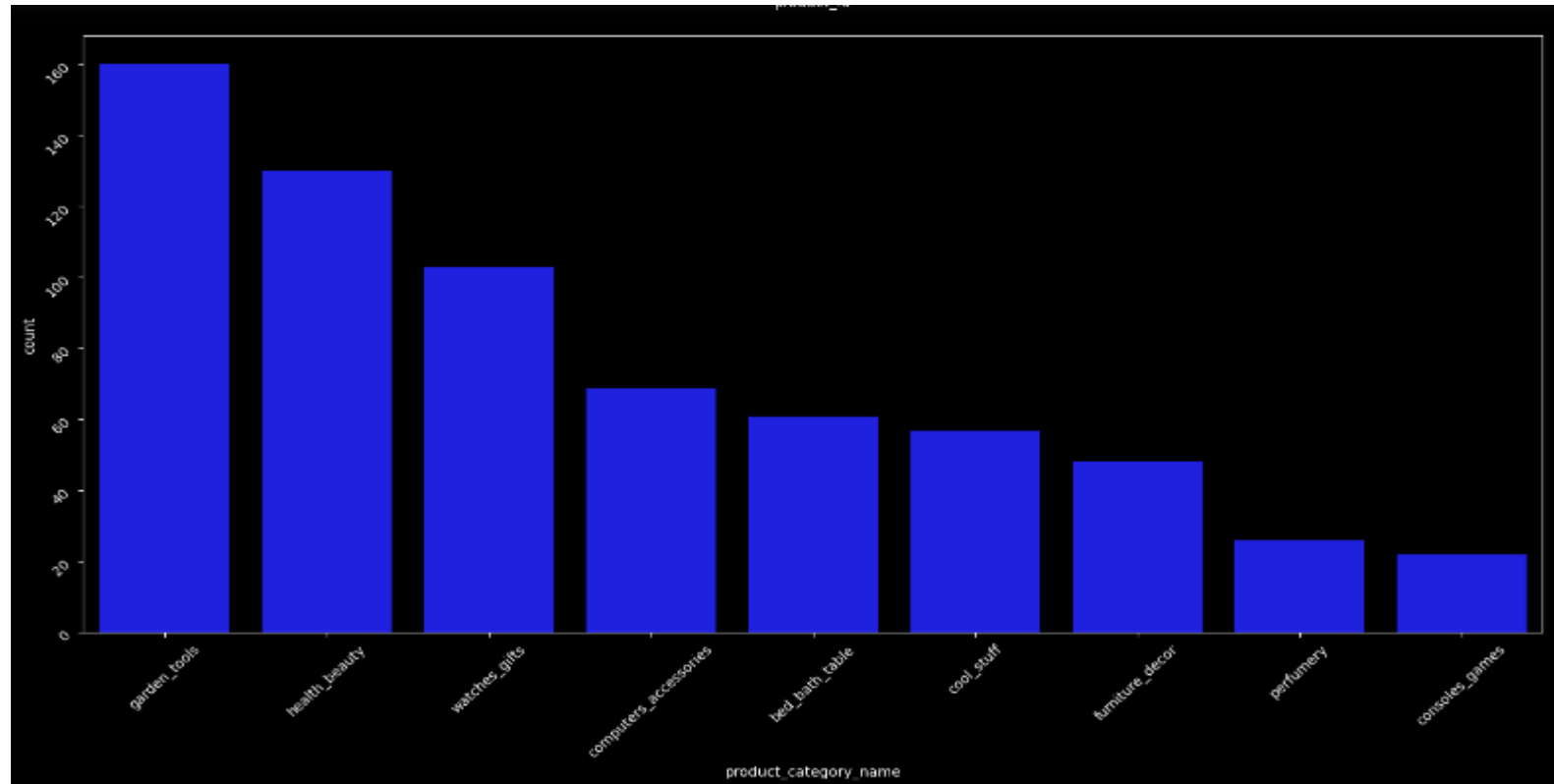


- Distribuição da variável “qty” (qty – venda mensal por produto) (plot à esquerda)
- Gráfico tipo Box (box plot) (à direita) – uma das informações que se pode extrair dele é a presença de outliers (círculos representados à direita – valores maiores que $qty=40$). Mas não neste caso.
- Neste dataset, especificamente, todos os produtos estão presentes nesta variável. Para se chegar a qualquer conclusão sobre os outliers, o dataset deve ser trabalhado de modo a ser analisado produto a produto. O mesmo vale para os preços, por exemplo



Comparação da quantidade de venda de todos os produtos

- Os 3 produtos mais vendidos são - health5, health7 e bed2
- Os 3 produtos menos vendidos são - health10, health6 e bed5



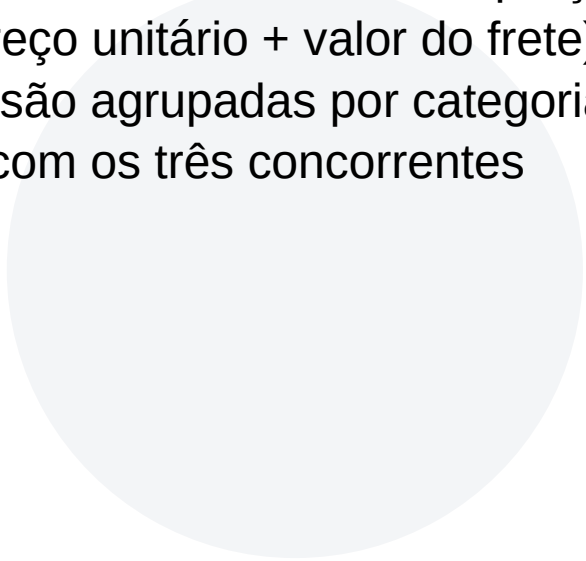
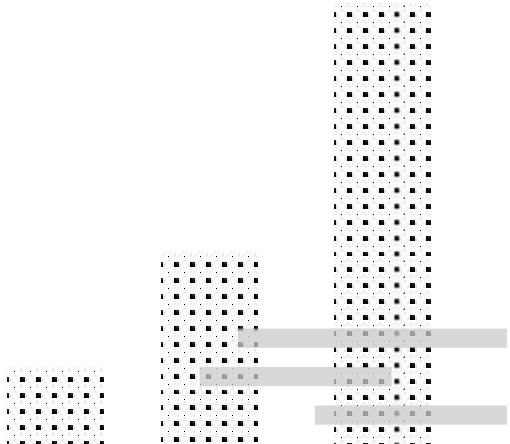
Porém, quando comparamos da quantidade de venda por categoria:

- "garden_tools" vem em primeiro lugar, seguido por "health_beauty"

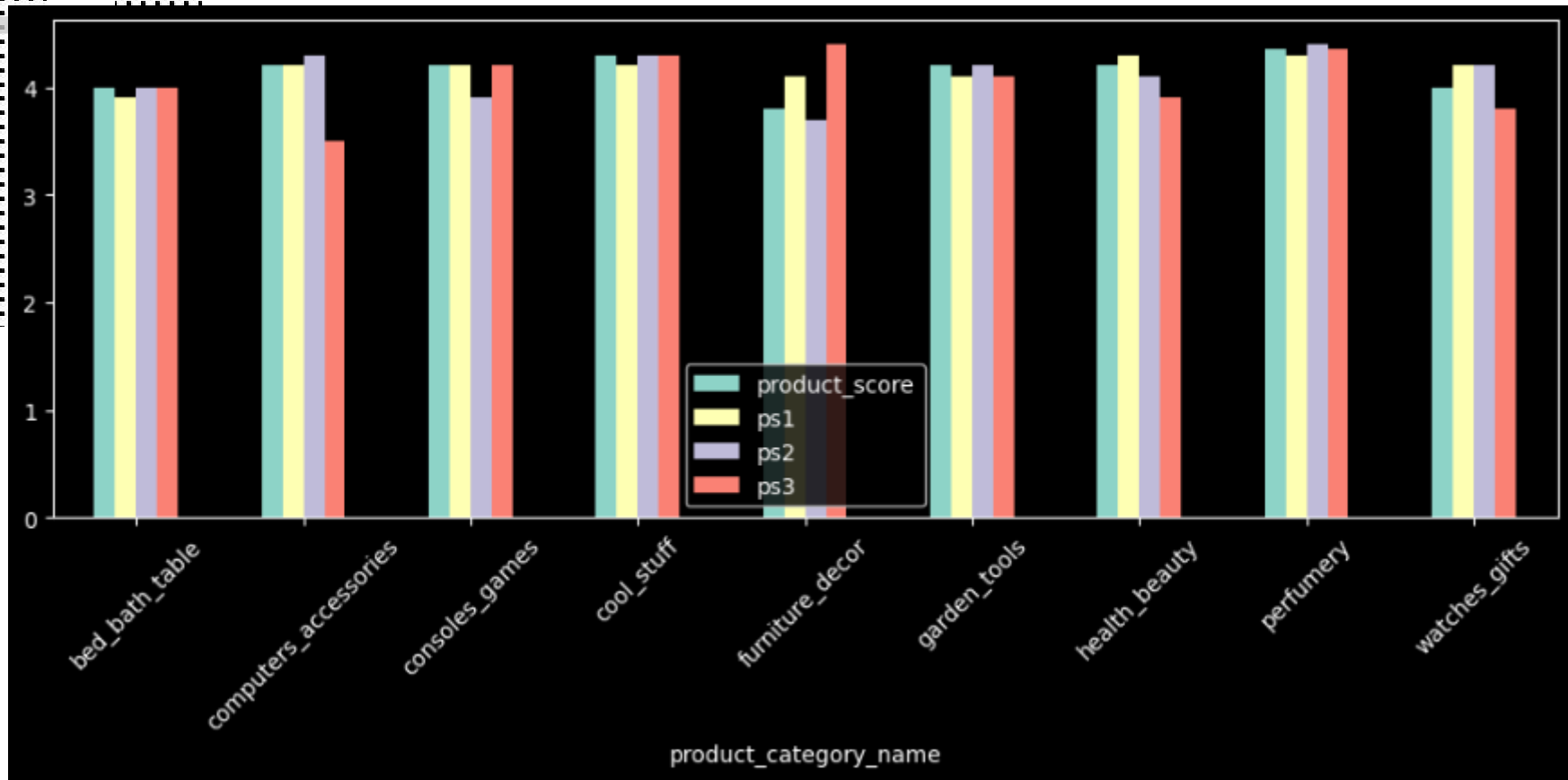
Considerações

- O dataset tem 676 linhas e 28 colunas
- Não tem nenhum valor nulo ou missing
- Os dados cobrem um período de 2 anos (2017-2018), incompletos
- Quando comparados os valores de “unit_price” do varejista com seus 3 concorrentes, nota-se que o valor máximo do varejista é mais elevado do que seus concorrentes em alguns casos
- Dos histogramas e box plots, nota-se uma grande assimetria nos dados e presença de outliers. Como o data set contém todas as categorias e subcategorias de produtos (product_id e product_category_name) não se pode tirar conclusões sobre outliers neste formato. Quando agruparmos os dados por "id" e "categoria" uma análise mais conclusiva pode ser feita
- Os 3 produtos mais vendidos são health5, health7 e bed2
- Entretanto, quando olhamos os produtos por categoria, "garden_tools" vem em primeiro lugar, seguido por "health_beauty"
- unit_price e lag_price são altamente correlacionados positivamente

O objetivo nesta etapa é determinar em quais categorias de produtos a varejista têm uma posição mais competitiva e em quais a posição é menos competitiva:

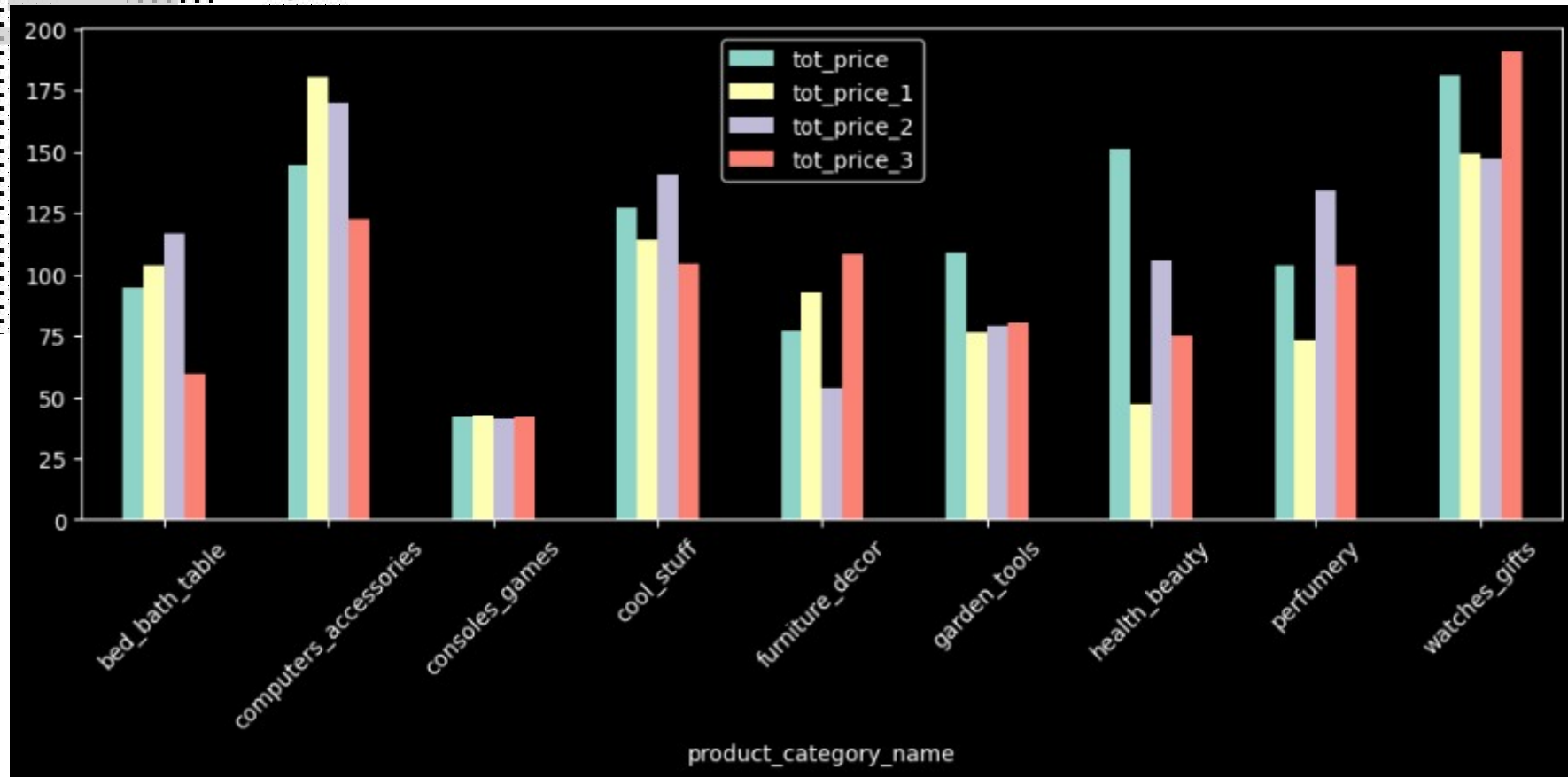
- Nesta análise, as variáveis estudadas são a avaliação do produto (product_score) e uma nova variável – preço total (tot_price) (preço total = preço unitário + valor do frete)
 - Estas variáveis são agrupadas por categoria
 - E comparadas com os três concorrentes
- 
- 

Estudo da competitividade da varejista nas diversas categorias de produtos



Comparando-se as avaliações dos produtos, nenhuma diferença significativa é observada, não sendo possível nenhuma conclusão sobre uma possível vantagem ou desvantagem relacionada às avaliações

Estudo da competitividade da varejista nas diversas categorias de produtos



Neste caso, as diferenças são significativas. Estas diferenças indicam um preço mais (ou menos) competitivo entre os varejistas. Por exemplo, nas categorias bed_bath_table, computers_accessories, furniture_decor, a varejista têm preços mais competitivos

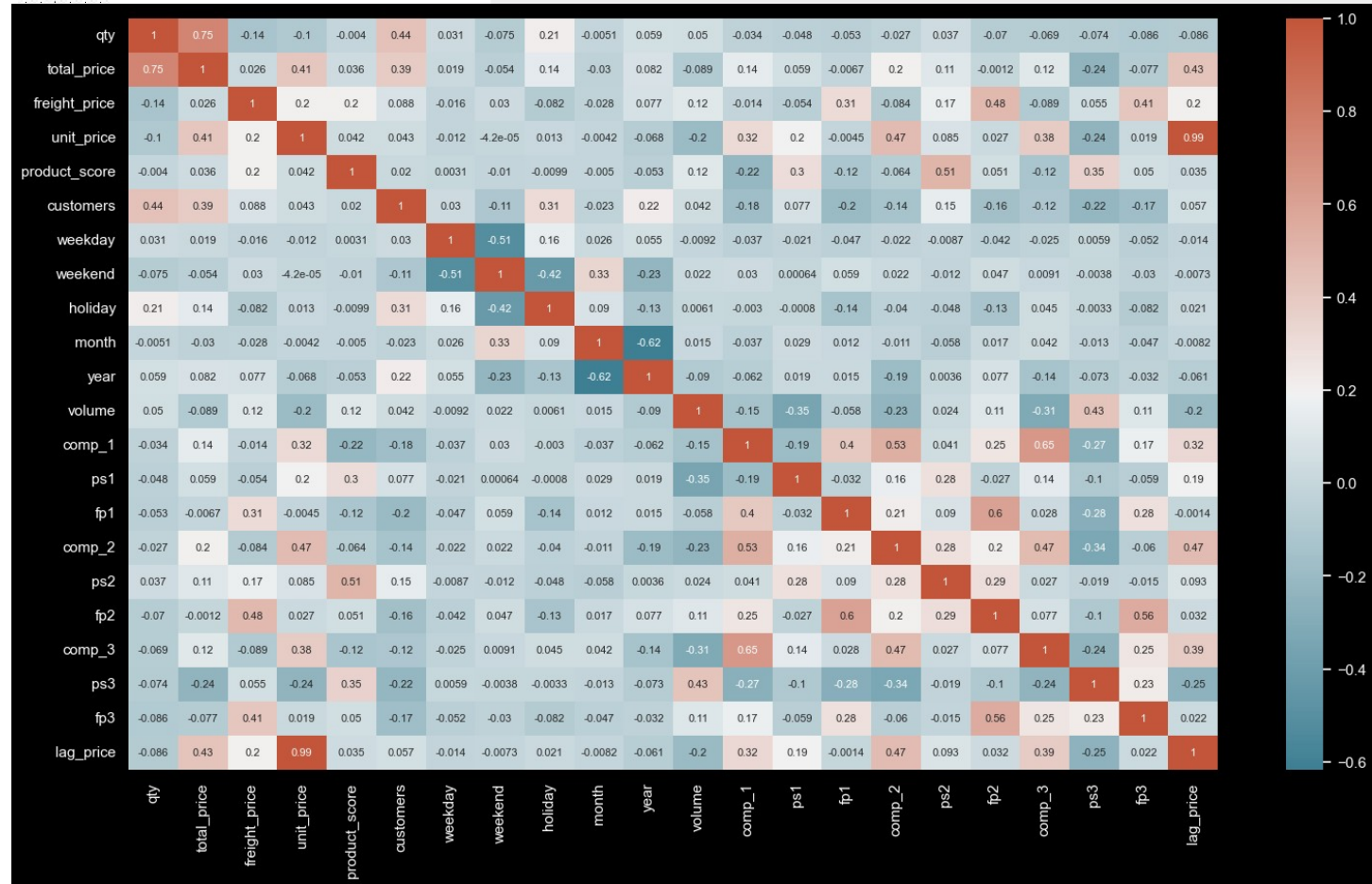
Considerações

- A primeira ação foi agregar os dados em produtos e categorias. Como devemos avaliar a competitividade por categoria entre o varejista e seus comcorrentes, foram selecionadas as variáveis mais relevantes (preço unitário, preço do frete e avaliação dos produtos). Em seguida, os dados foram agrupados por categoria e criada uma nova variável - preço total (preço unitário + preço do frete).
 - Comparando as avaliações por categoria entre o varejista e outras 3 lojas, nenhuma diferença significativa foi observada.
 - Quando se compara os preços totais, nota-se diferenças significativas. Estas diferenças poderiam afetar a competitividade do varejista, pois em muitos casos, os preços de seus produtos são mais elevados do que os dos seus concorrentes.
- 1) Em quais categorias de produtos a varejista tem uma posição mais competitiva?
 - bed_bath_table, computers_accessories, furniture_decor - seus preços são mais vantajosos em relação a maioria dos concorrentes
 - 2) Em quais categorias de produtos a varejista tem uma posição menos competitiva?
 - health_beauty, garden_tools - nestas duas categorias, o varejista tem os preços mais altos do que todos seus concorrentes

Nesta etapa, o objetivo é estudar o impacto das outras variáveis no preço de venda do produto:

- Foram analisadas as correlações entre as variáveis para se verificar como elas se relacionam
- A análise de correlação é uma forma descritiva que mede se há e qual o grau de dependência entre variáveis, ou seja, o quanto uma variável interfere em outra. Essa medida de grau de relação é medida através de coeficientes.
- Estes coeficientes podem assumir valores entre -1 e 1. Quanto mais próximos dos extremos, maior é a correlação entre as variáveis
- Existem três possíveis cenários:
 - 1) **Correlação positiva:** quando duas variáveis que possuem correlação crescem ou decrescem juntas, ou seja, que possuem uma relação direta;
 - 2) **Correlação negativa:** quando duas variáveis que possuem correlação mas quando uma variável cresce a outra decresce, ou vice-versa;
 - 3) **Não ter correlação:** quando o crescimento ou decrescimento de uma variável não tem efeito sobre outra variável.

Análise do impacto das variáveis no preço de venda



- Esta figura representa um mapa de calor (heat map). É uma ferramenta muito útil que nos permite ter acesso muito rápido às correlações de modo visual.
- Como o texto é muito pequeno, uma tabela será apresentada e discutida a seguir

Análise do impacto das variáveis no preço de venda																							
	qty	total_price	freight_price	unit_price	product_score	customers	weekday	weekend	holiday	month	year	volume	comp_1	ps1	fp1	comp_2	ps2	fp2	comp_3	ps3	fp3	lag_price	
qty	1.00	0.75	-0.14	-0.10	-0.00	0.44	0.03	-0.08	0.21	-0.01	0.06	0.05	-0.03	-0.05	-0.05	-0.03	0.04	-0.07	-0.07	-0.07	-0.09	-0.09	
total_price	0.75	1.00	0.03	0.41	0.04	0.39	0.02	-0.05	0.14	-0.03	0.08	-0.09	0.14	0.06	-0.01	0.20	0.11	-0.00	0.12	-0.24	-0.08	0.43	
freight_price	-0.14	0.03	1.00	0.20	0.20	0.09	-0.02	0.03	-0.08	-0.03	0.08	0.12	-0.01	-0.05	0.31	-0.08	0.17	0.48	-0.09	0.05	0.41	0.20	
unit_price	-0.10	0.41	0.20	1.00	0.04	0.04	-0.01	-0.00	0.01	-0.00	-0.07	-0.20	0.32	0.20	-0.00	0.47	0.09	0.03	0.38	0.24	0.02	0.99	
product_score	-0.00	0.04	0.20	0.04	1.00	0.02	0.00	-0.01	-0.01	-0.00	-0.05	0.12	-0.22	0.30	-0.12	-0.06	0.51	0.05	-0.12	0.35	0.05	0.04	
customers	0.44	0.39	0.09	0.04	0.02	1.00	0.03	-0.11	0.31	-0.02	0.22	0.04	-0.18	0.08	-0.20	-0.14	0.15	-0.16	-0.12	-0.22	-0.17	0.06	
weekday	0.03	0.02	-0.02	-0.01	0.00	0.03	1.00	-0.51	0.16	0.03	0.06	-0.01	-0.04	-0.02	-0.05	-0.02	-0.01	-0.04	-0.03	0.01	-0.05	-0.01	
weekend	-0.08	-0.05	0.03	-0.00	-0.01	-0.11	-0.51	1.00	-0.42	0.33	-0.23	0.02	0.03	0.00	0.06	0.02	-0.01	0.05	0.01	-0.00	-0.03	-0.01	
holiday	0.21	0.14	-0.08	0.01	-0.01	0.31	0.16	-0.42	1.00	0.09	-0.13	0.01	-0.00	-0.00	-0.14	-0.04	-0.05	-0.13	0.04	-0.00	-0.08	0.02	
month	-0.01	-0.03	-0.03	-0.00	-0.00	-0.02	0.03	0.33	0.09	1.00	-0.62	0.01	-0.04	0.03	0.01	-0.01	-0.06	0.02	0.04	-0.01	-0.05	-0.01	
year	0.06	0.08	0.08	-0.07	-0.05	0.22	0.06	-0.23	-0.13	-0.62	1.00	-0.09	-0.06	0.02	0.02	-0.19	0.00	0.08	-0.14	-0.07	-0.03	-0.06	
volume	0.05	-0.09	0.12	-0.20	0.12	0.04	-0.01	0.02	0.01	0.01	-0.09	1.00	-0.15	-0.35	-0.06	-0.23	0.02	0.11	-0.31	0.43	0.11	-0.20	
comp_1	-0.03	0.14	-0.01	0.32	-0.22	-0.18	-0.04	0.03	-0.00	-0.04	-0.06	-0.15	1.00	-0.19	0.40	0.53	0.04	0.25	0.65	-0.27	0.17	0.32	
ps1	-0.05	0.06	-0.05	0.20	0.30	0.08	-0.02	0.00	-0.00	0.03	0.02	-0.35	-0.19	1.00	-0.03	0.16	0.28	-0.03	0.14	-0.10	-0.06	0.19	
fp1	-0.05	-0.01	0.31	-0.00	-0.12	-0.20	-0.05	0.06	-0.14	0.01	0.02	-0.06	0.40	-0.03	1.00	0.21	0.09	0.60	0.03	-0.28	0.28	-0.00	
comp_2	-0.03	0.20	-0.08	0.47	-0.06	-0.14	-0.02	0.02	-0.04	-0.01	-0.19	-0.23	0.53	0.16	0.21	1.00	0.28	0.20	0.47	-0.34	-0.06	0.47	
ps2	0.04	0.11	0.17	0.09	0.51	0.15	-0.01	-0.01	-0.05	-0.06	0.00	0.02	0.04	0.28	0.09	0.28	1.00	0.29	0.03	-0.02	-0.01	0.09	
fp2	-0.07	-0.00	0.48	0.03	0.05	-0.16	-0.04	0.05	-0.13	0.02	0.08	0.11	0.25	-0.03	0.60	0.20	0.29	1.00	0.08	-0.10	0.56	0.03	
comp_3	-0.07	0.12	-0.09	0.38	-0.12	-0.12	-0.03	0.01	0.04	0.04	-0.14	-0.31	0.65	0.14	0.03	0.47	0.03	0.08	1.00	-0.24	0.25	0.39	
ps3	-0.07	-0.24	0.05	-0.24	0.35	-0.22	0.01	-0.00	-0.00	-0.01	-0.07	0.43	-0.27	-0.10	-0.28	-0.34	-0.02	-0.10	-0.24	1.00	0.23	-0.25	
fp3	-0.09	-0.08	0.41	0.02	0.05	-0.17	-0.05	-0.03	-0.08	-0.05	-0.03	0.11	0.17	-0.06	0.28	-0.06	-0.01	0.56	0.25	0.23	1.00	0.02	
lag_price	-0.09	0.43	0.20	0.99	0.04	0.06	-0.01	-0.01	0.02	-0.01	-0.06	-0.20	0.32	0.19	-0.00	0.47	0.09	0.03	0.39	-0.25	0.02	1.00	

Análise do impacto das variáveis no preço de venda

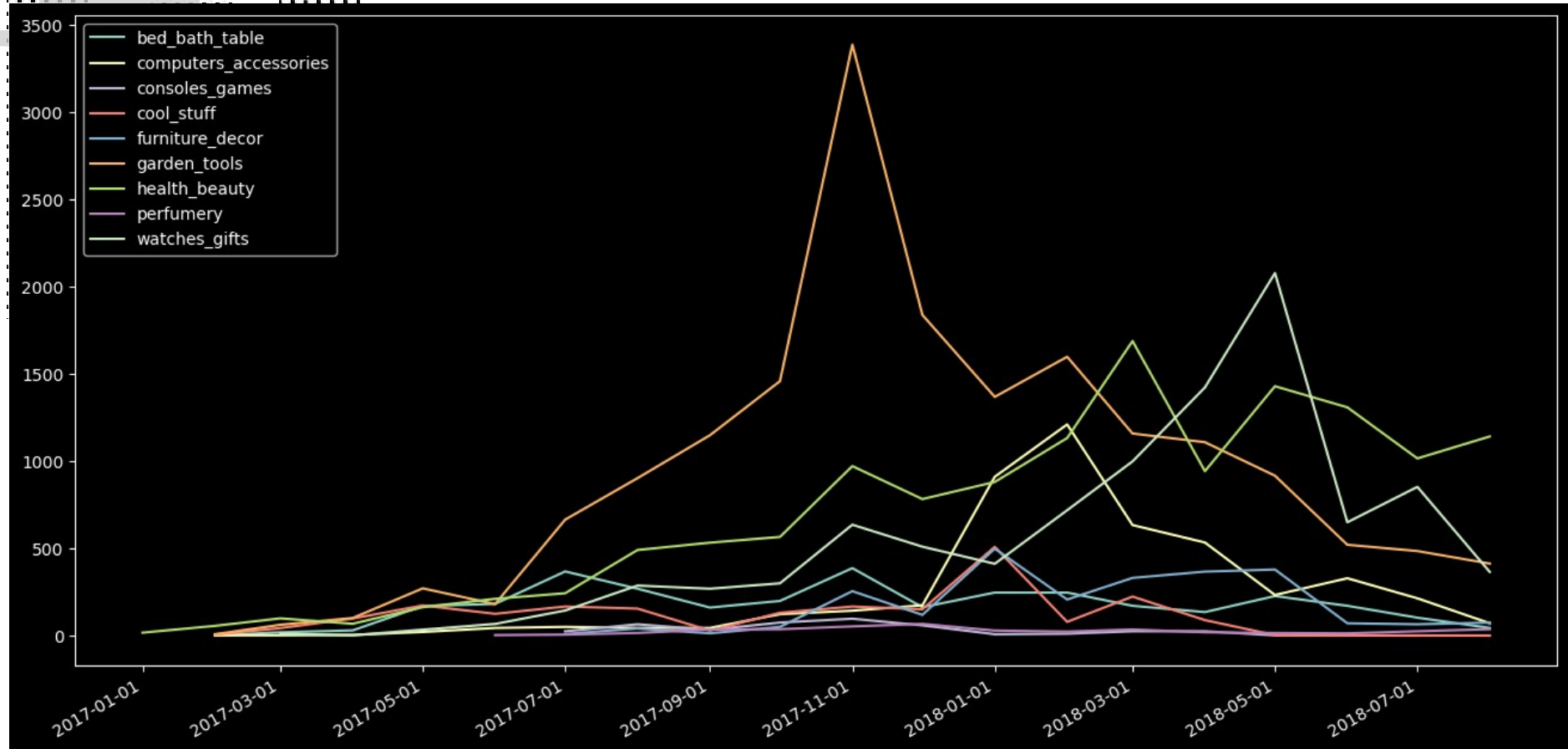
	qty	total_price	freight_price	unit_price	product_score	customers	weekday	weekend	holiday	month	year	volume	comp_1	ps1	fp1	comp_2	ps2	fp2	comp_3	ps3	fp3	lag_price
qty	1.00	0.75	-0.14	-0.10	-0.00	0.44	0.03	-0.08	0.21	-0.01	0.06	0.05	-0.03	-0.05	-0.05	-0.03	0.04	-0.07	-0.07	-0.07	-0.09	-0.09
total_price	0.75	1.00	0.03	0.41	0.04	0.39	0.02	-0.05	0.14	-0.03	0.08	-0.09	0.14	0.06	-0.01	0.20	0.11	-0.00	0.12	-0.24	-0.08	0.43
freight_price	-0.14	0.03	1.00	0.20	0.20	0.09	-0.02	0.03	-0.08	-0.03	0.08	0.12	-0.01	-0.05	0.31	-0.08	0.17	0.48	-0.09	0.05	0.41	0.20
unit_price	-0.10	0.41	0.20	1.00	0.04	0.04	-0.01	-0.00	0.01	-0.00	-0.07	-0.20	0.32	0.20	-0.00	0.47	0.09	0.03	0.38	0.24	0.02	0.99
product_score	-0.00	0.04	0.20	0.04	1.00	0.02	0.00	-0.01	-0.01	-0.00	-0.05	0.12	-0.22	0.30	-0.12	-0.06	0.51	0.05	-0.12	0.35	0.05	0.04
customers	0.44	0.39																	-0.16	-0.12	-0.22	-0.17
weekday	0.03	0.02																	-0.04	-0.03	0.01	-0.05
weekend	-0.08	-0.05																	0.05	0.01	-0.00	-0.03
holiday	0.21	0.14																	-0.13	0.04	-0.00	-0.08
month	-0.01	-0.03																	0.02	0.04	-0.01	-0.05
year	0.06	0.08																	0.08	-0.14	-0.07	-0.03
volume	0.05	-0.09																	0.11	-0.31	0.43	0.11
comp_1	-0.03	0.14																	0.25	0.65	-0.27	0.17
ps1	-0.05	0.06																	-0.03	0.14	-0.10	-0.06
fp1	-0.05	-0.01																	0.60	0.03	-0.28	0.28
comp_2	-0.03	0.20																	0.20	0.47	-0.34	-0.06
ps2	0.04	0.11																	0.29	0.03	-0.02	-0.01
fp2	-0.07	-0.00																	1.00	0.08	-0.10	0.56
comp_3	-0.07	0.12																	0.08	1.00	-0.24	0.25
ps3	-0.07	-0.24	0.05	-0.24	0.35	-0.22	0.01	-0.00	-0.00	-0.01	-0.07	0.43	-0.27	-0.10	-0.28	-0.34	-0.02	-0.10	-0.24	1.00	0.23	-0.25
fp3	-0.09	-0.08	0.41	0.02	0.05	-0.17	-0.05	-0.03	-0.08	-0.05	-0.03	0.11	0.17	-0.06	0.28	-0.06	-0.01	0.56	0.25	0.23	1.00	0.02
lag_price	-0.09	0.43	0.20	0.99	0.04	0.06	-0.01	-0.01	0.02	-0.01	-0.06	-0.20	0.32	0.19	-0.00	0.47	0.09	0.03	0.39	-0.25	0.02	1.00

Desta análise, podemos concluir que quatro variáveis tem algum impacto no preço de venda:

- Os preços dos concorrentes:
 - ✓ comp_1 (coef = 0.32)
 - ✓ comp_2 (coef = 0.47)
 - ✓ comp_3 (coef = 0.38)
 - ✓ Existe uma correlação positiva, embora seja fraca
- E o lag_price (preço de venda referente ao mês anterior)
 - ✓ lag_price (coef = 0.99)
 - ✓ Existe uma correlação positiva e é extremamente forte

O objetivo é fazer uma análise de demanda de produtos, respondendo as seguintes questões:

- É possível prever qual será a demanda por exemplo, por categoria de produto?
- E por produto individual?
- Para tal estudo, iremos estudar a variação da demanda ao longo do tempo (série histórica) e utilizaremos alguns modelos de séries temporais para modularmos esta variação
- Primeiro, os dados serão agrupados por categoria, e depois por produto



- Distribuição da demanda ao longo do tempo para as diversas categorias
- Para a análise, a categoria “health_beauty” foi utilizada

Metodologia de média móvel (Moving Average Methodology)

O método da média móvel é muito utilizado em séries temporais. A média móvel (MA) (ou) média móvel é calculada utilizando valores médios da série temporal em k períodos.

- Média Móvel Simples (SMA),
- Média Móvel Cumulativa (CMA)
- Média Móvel Exponencial (EMA)

Média Móvel Simples (SMA)

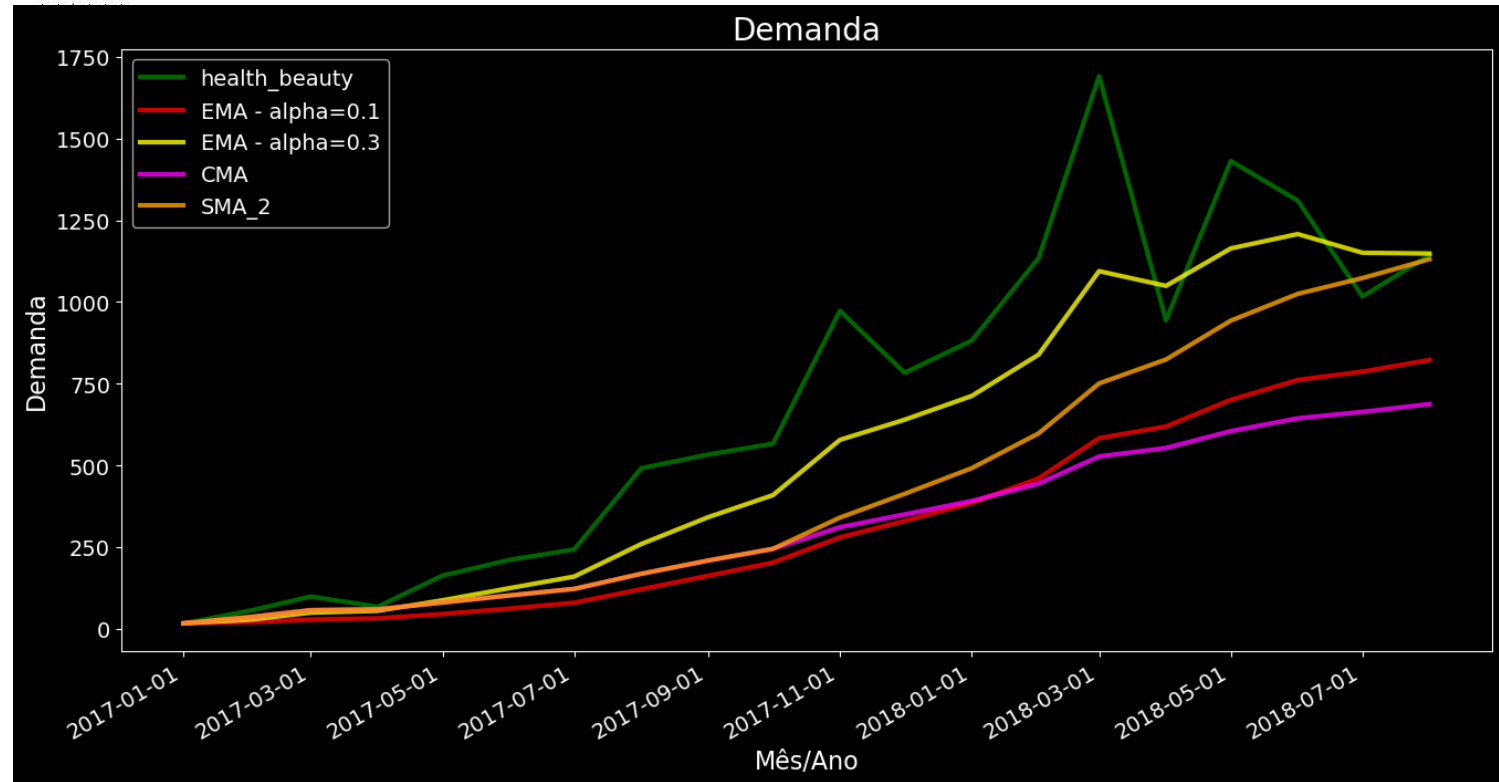
A Média Móvel Simples (SMA) calcula a média não ponderada dos pontos anteriores.

Média Móvel Cumulativa (CMA)

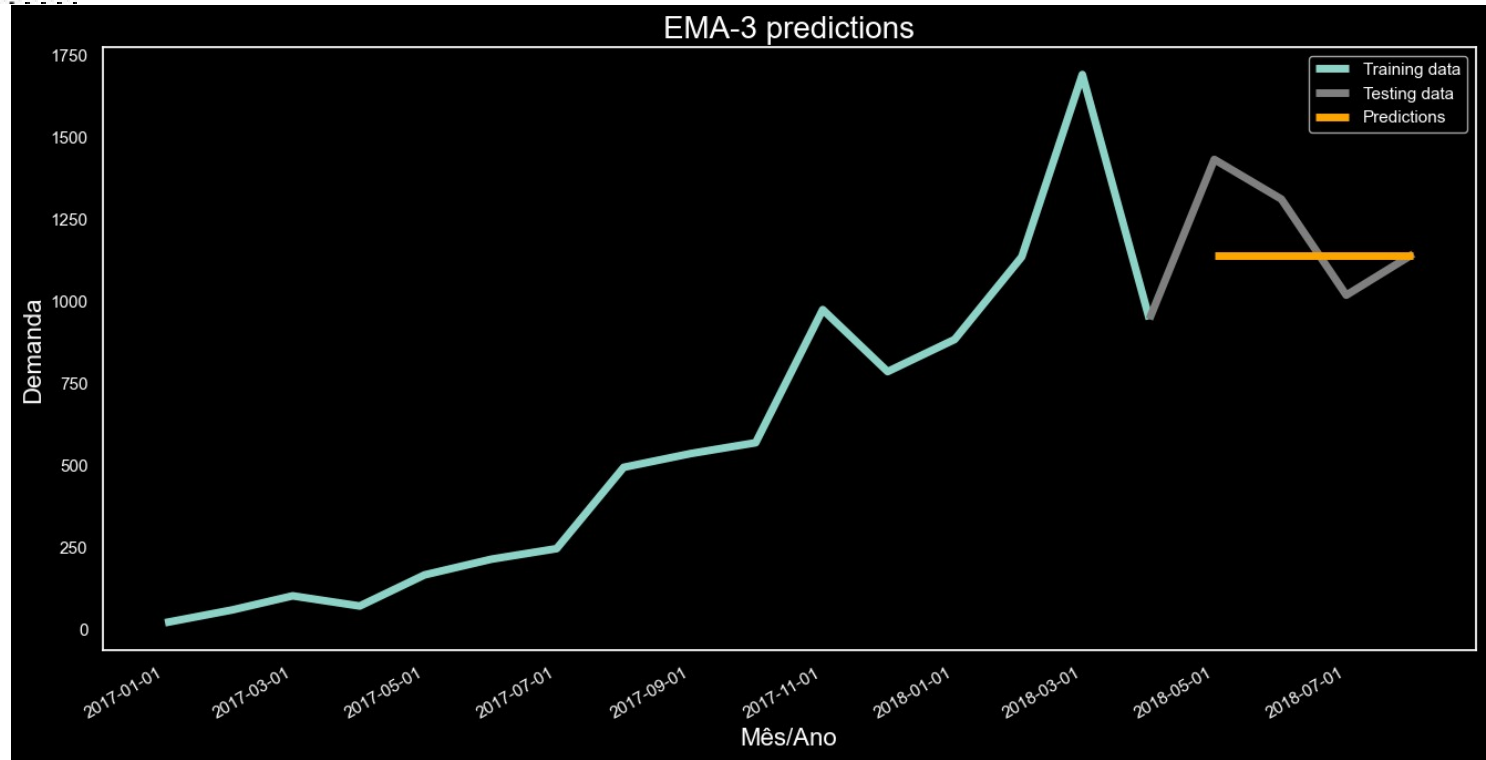
O CMA é a média não ponderada dos valores anteriores até o momento atual.

Média Móvel Exponencial (EMA)

A EMA é usada principalmente para identificar tendências e filtrar ruídos. O peso dos elementos diminui gradualmente ao longo do tempo. Isso significa que dá peso aos pontos de dados recentes, não aos históricos.



Os três métodos citados anteriormente foram aplicados para reproduzir os dados. Embora eles razoavelmente reproduzem a tendência, eles são incapazes de reproduzir as modulações em sua amplitude



Foi feita uma tentativa de utilizar o EMA para fazer as previsões. Os dados foram divididos em dois conjuntos – treino (azul) e teste (cinza). A linha laranja representa o modelo treinado sendo aplicado no conjunto de teste. Claramente, o modelo (laranja) é incapaz de reproduzir o conjunto de teste (cinza). Outro tipo de modelagem se faz necessária.

Também foram testados outros modelos mais robustos para a análise de séries temporais.

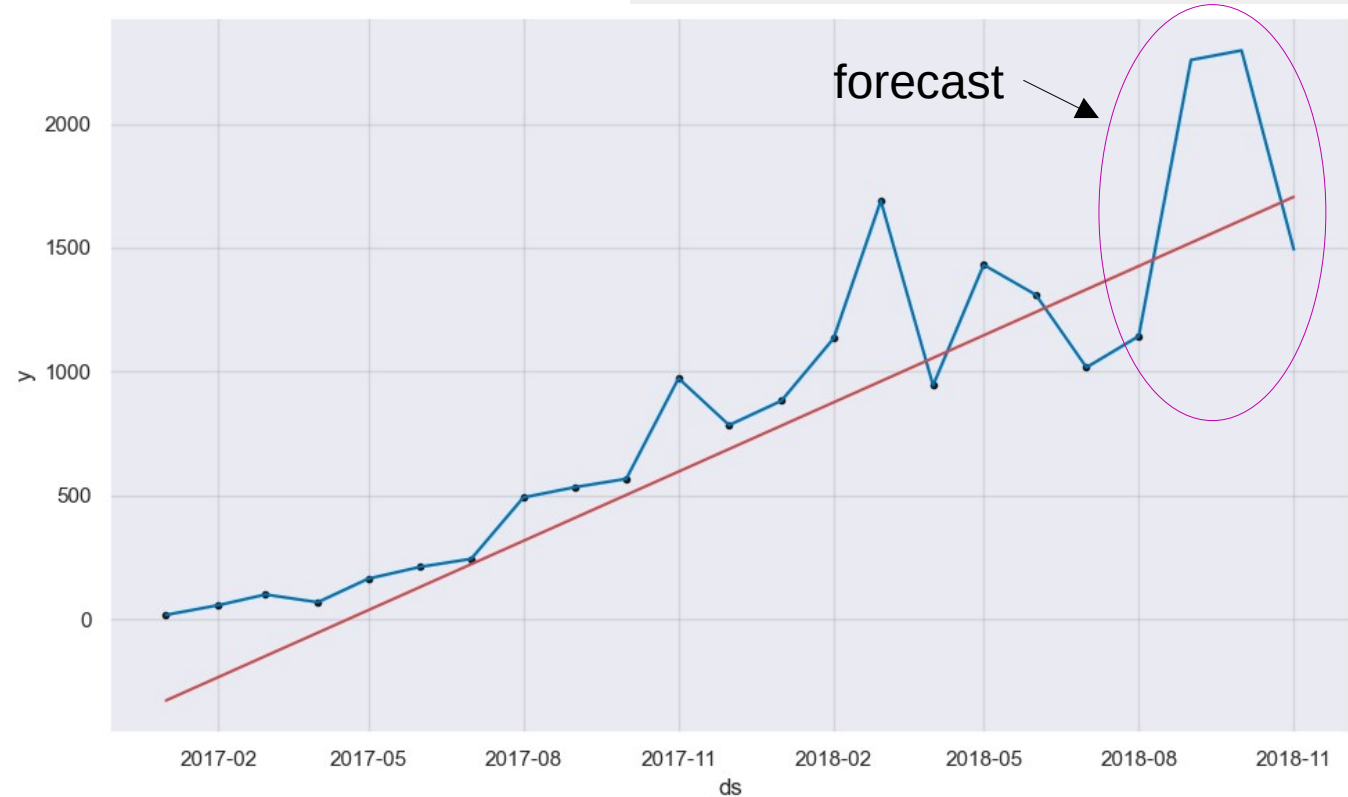
- Autoregressive Moving Average (ARMA)
- Autoregressive Integrated Moving Average (ARIMA)
- Seasonal Autoregressive Integrated Moving Average (SARIMA)

Infelizmente, o dataset disponível é referente a um intervalo de tempo muito limitado (~20 meses) o que inviabiliza a identificação de qualquer modulação ou efeitos sazonais.

Prophet Model

- Prophet, ou “Facebook Prophet”, é uma biblioteca de código aberto para previsão de séries temporais univariadas (uma variável) desenvolvida pelo Facebook.
- Ele foi projetado para ser fácil e totalmente automático. Dada uma série temporal o modelo retorna uma previsão (forecast).

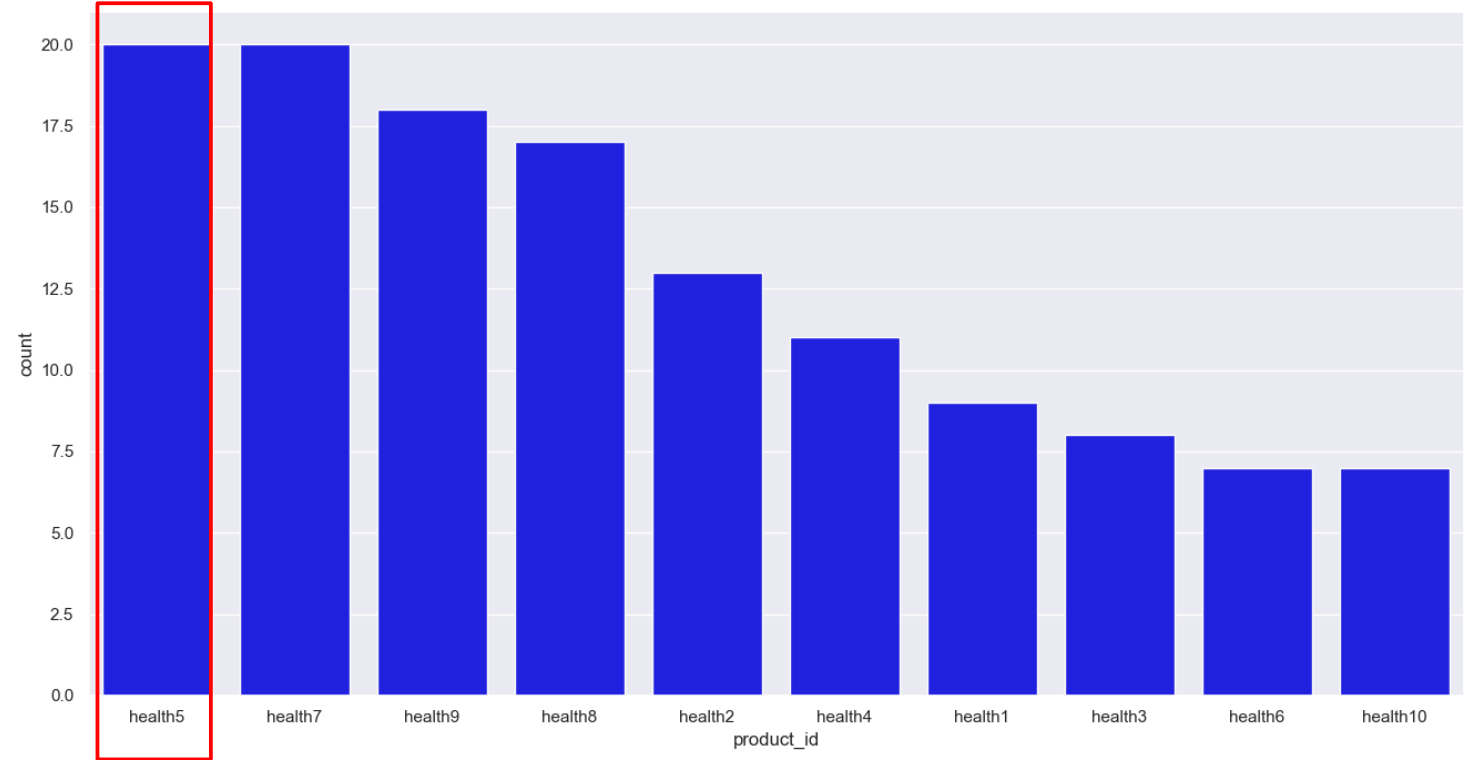
Análise de demanda por categoria



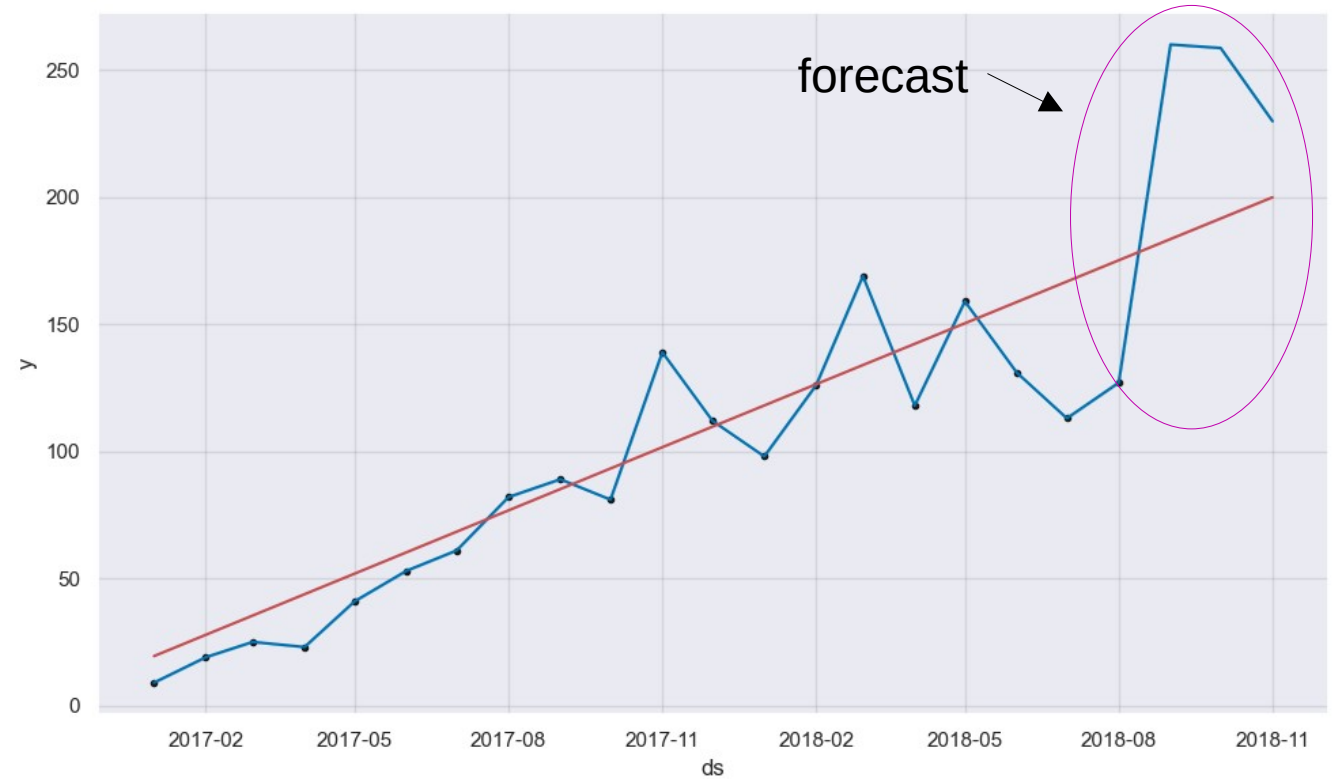
A performance do Prophet se mostrou muito superior aos outros modelos. Uma das razões é o fato de você não ter que dividir o dataset em treino e teste. Com estes dados (20 meses) você não consegue treinar um bom modelo, pois o dataset de treino é muito diferente do dataset de teste. A modulação não se repete. Veja por exemplo, o período entre 02/2017-05/2017 comparado com 02/2018-05/2018.

Análise de demanda por produto

Bar plot for product_id



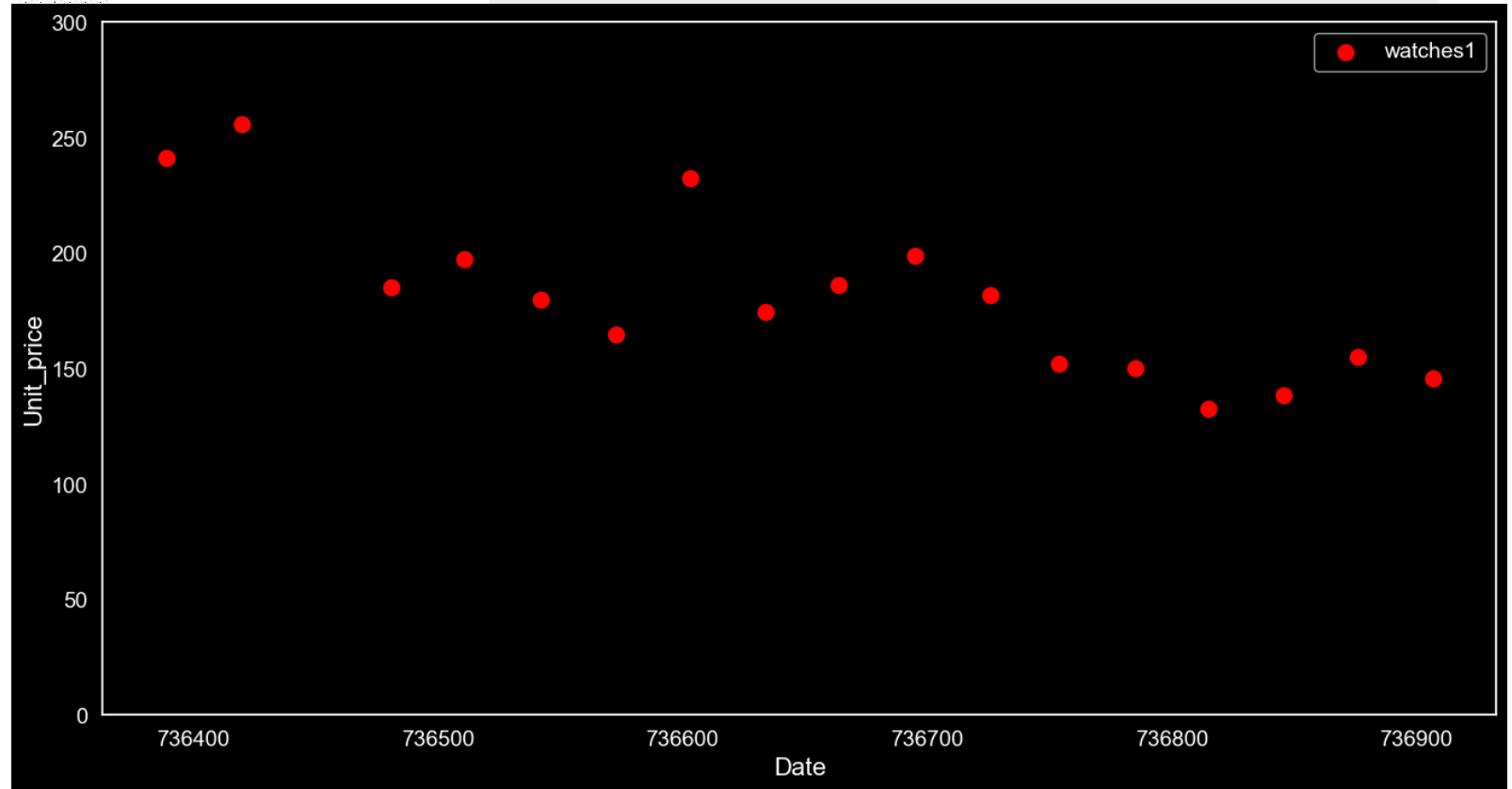
Na análise da demanda por produto, foi selecionado o `product_id = health5`, dentro da mesma categoria, por ele ter um maior número de eventos (como visto no gráfico acima).



Na análise de demanda por produto, o mesmo procedimento foi adotado utilizando o modelo Prophet

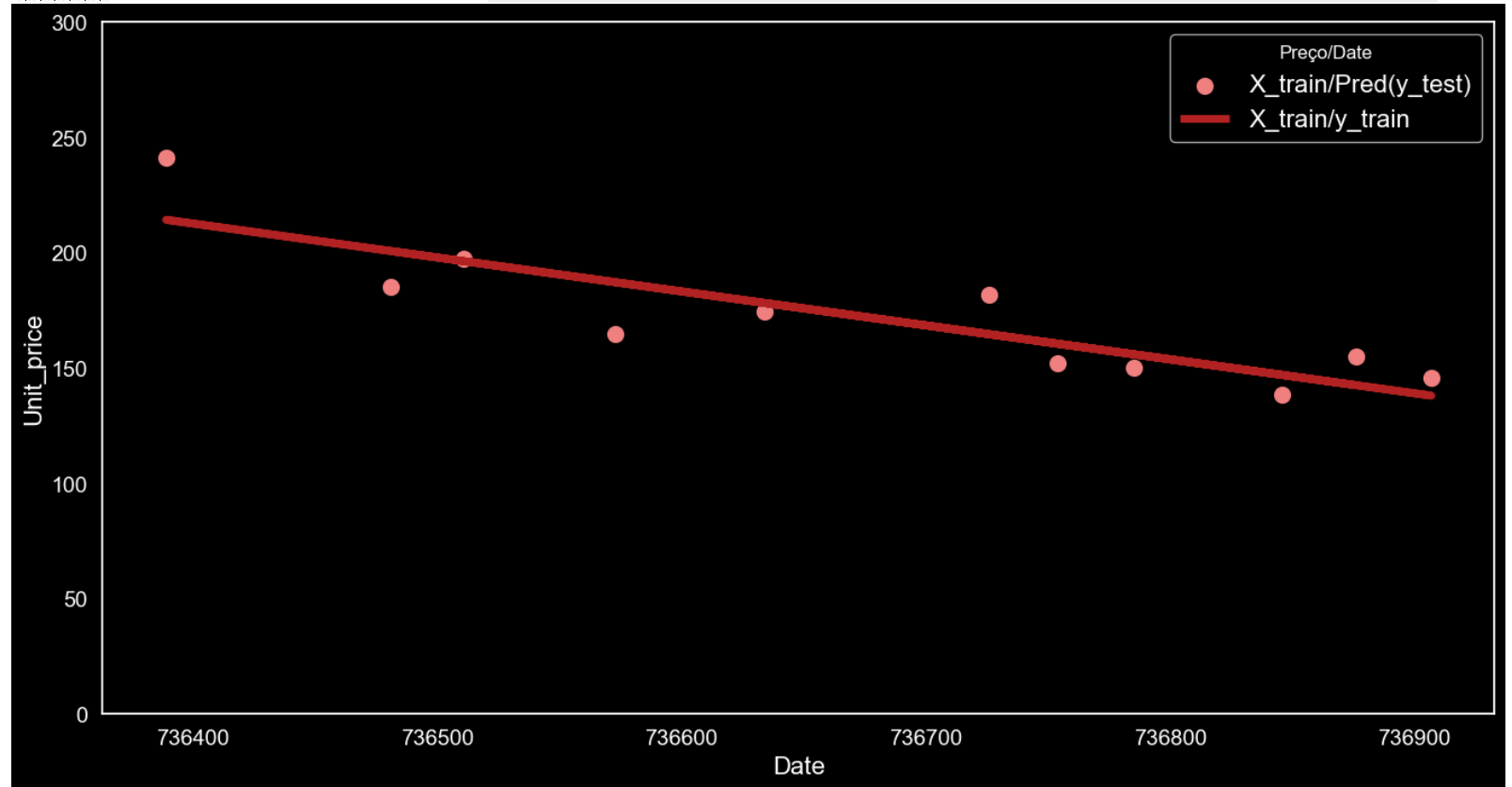
- O objetivo é o desenvolvimento de um modelo preditivo para os preços dos produtos
- Durante a análise exploratória, notou-se uma relação aproximadamente linear, decrescente, entre os preços dos produtos ao longo do tempo. Como primeira solução, vamos utilizar este comportamento para prever os preços dos produtos
- O modelo de regressão linear simples foi escolhido. A regressão linear simples é um modelo de regressão linear com uma única variável explicativa.
- Na regressão linear simples, prevemos valores de uma variável com base nos valores de outra. A variável Y é a variável que estamos prevendo (preço). A variável preditora X é a variável com a qual estamos fazendo nossas previsões (data).
- O produto escolhido para a criação do modelo foi “watches1”, dentro da categoria “watches_gifts”

Modelo preditivo para os preços dos produtos



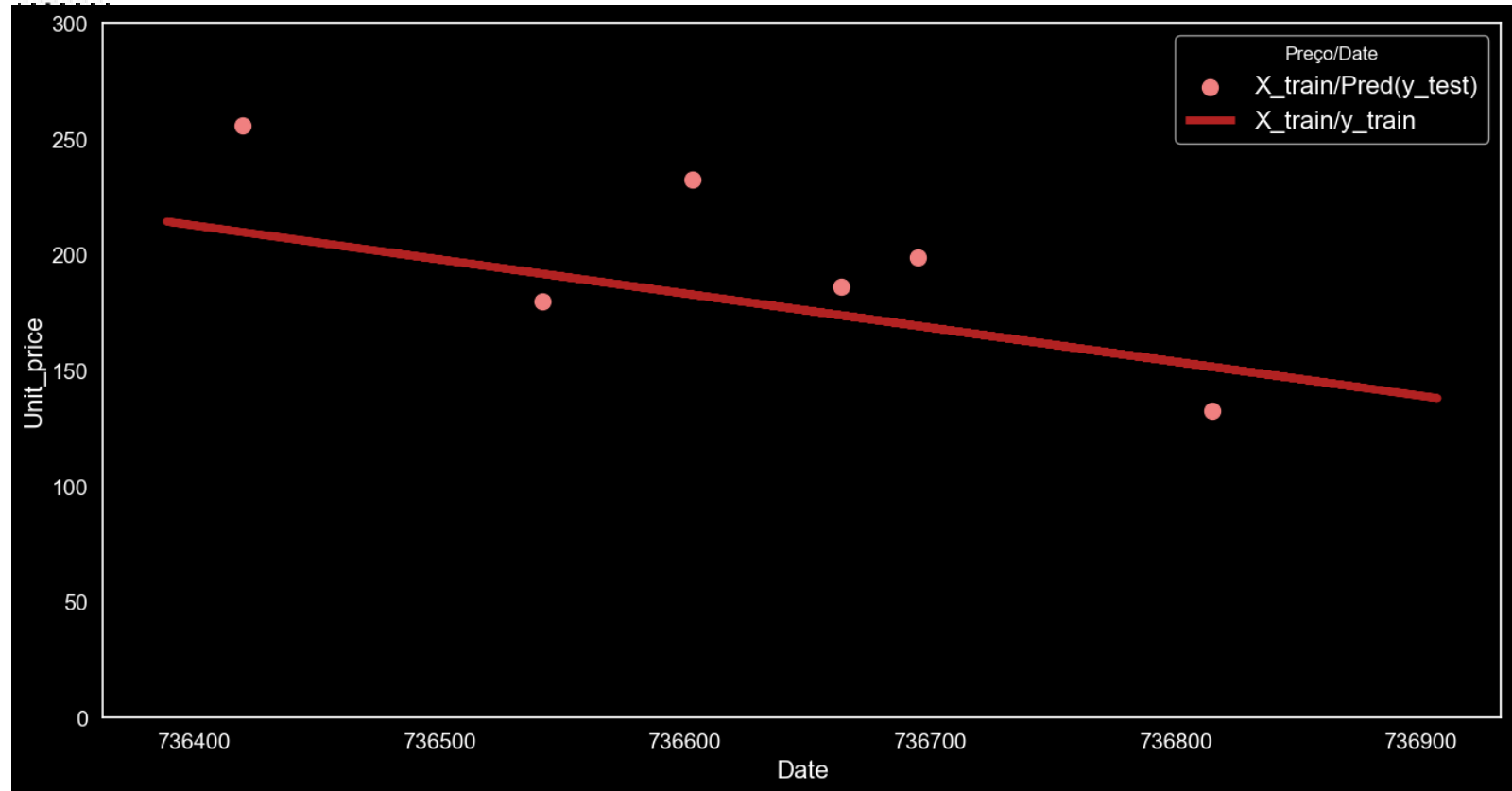
O gráfico acima, demonstra a relação entre os preços e o período (Mês/Ano) para o produto “watches1”

Modelo preditivo para os preços dos produtos



- O dataset foi dividido em treino (80%) e teste (20%)
- O gráfico acima mostra os dados do treino (círculos) e os valores preditos (linha)

Modelo preditivo para os preços dos produtos

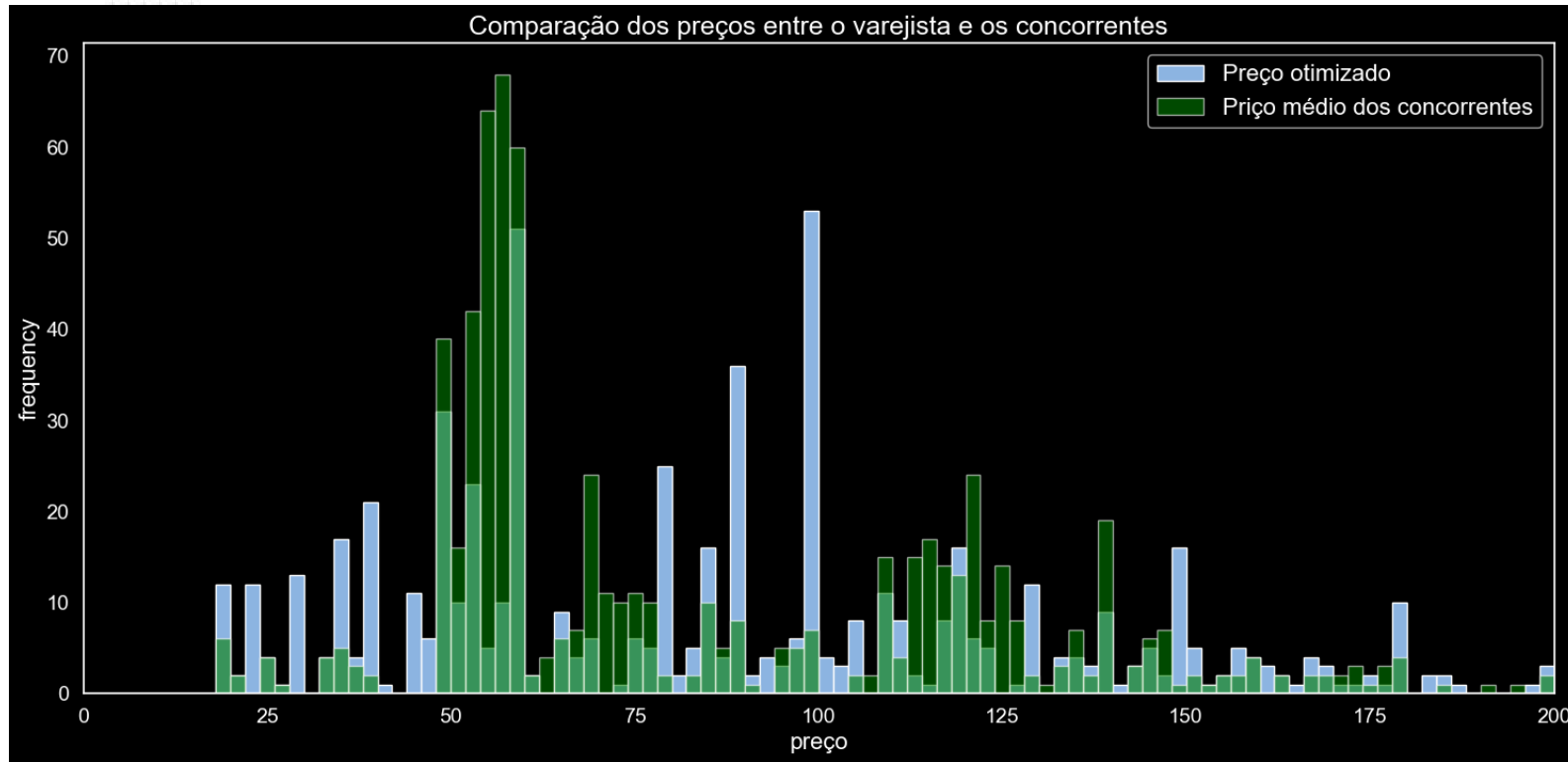


- O gráfico acima mostra o dataset de teste (círculos) e os valores preditos (linha)
- Usando a equação da reta $y = mx + c$, e os parâmetros c (y-intercept) e m (slope/coefficient) diretamente do modelo de regressão, podemos prever os preços futuros

- Nesta etapa, o objetivo é criar um modelo de otimização para a política de preços praticada pela varejista e identificar as vantagens da política sugerida
- Queremos otimizar os preços dos produtos baseado nos dados presentes no dataset disponível. Não temos informações sobre os consumidores, mas temos dados históricos de venda
- As variáveis selecionadas para desenvolver o modelo são:
 - preço do produto, preço do frete, quantidade de fotos do produto, avaliação do produto, preço do concorrente 1, preço do concorrente 2 e preço do concorrente 3
- E como queremos otimizar o preço, a variável alvo é:
 - Quantidade de vendas por produto
- Nesta solução, um modelo de regressão linear foi utilizado

Modelo de otimização da política de preços praticada

Um ajuste personalizado também foi aplicado com base na margem desejada e nas restrições de preço mínimo. Os valores aqui utilizados foram 10% do preço como margem e como preço mínimo, 90% do preço. O preço final ajustado é determinado tomando-se o valor máximo entre o preço ajustado do produto, o preço mínimo e um preço baseado na margem desejada.



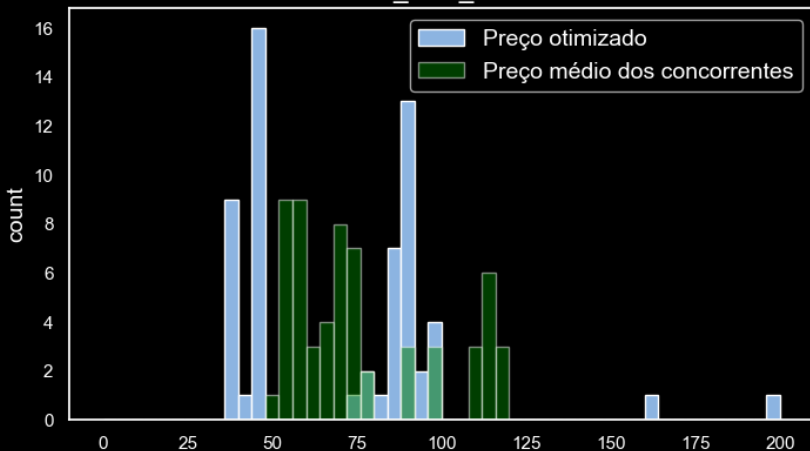
Neste gráfico, temos a comparação entre os preços otimizados e a média dos preços para o mesmo produtos dos concorrentes. Nota-se que os preços otimizados da varejista são muito competitivos, mesmo sendo mais elevados do que os preços antes da otimização, maximizando a venda.

Modelo de otimização da política de preços praticada

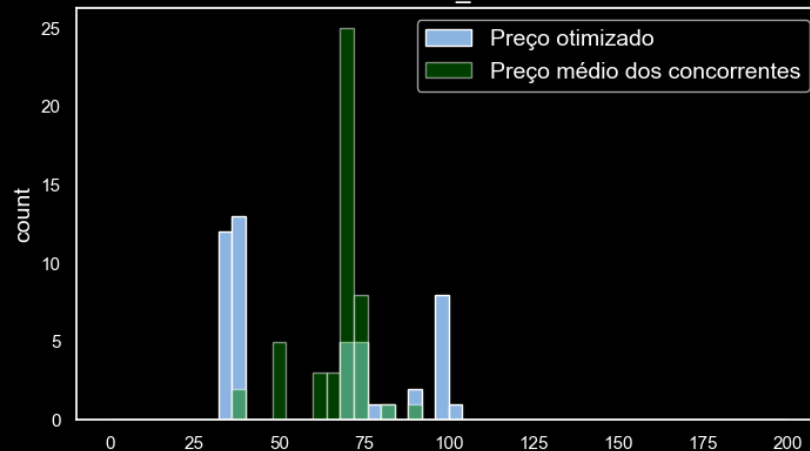
Mesma comparação do gráfico anterior, mas separados por categoria:

- bed_bath_table
- furniture_decor
- health_beauty
- watches_gifts

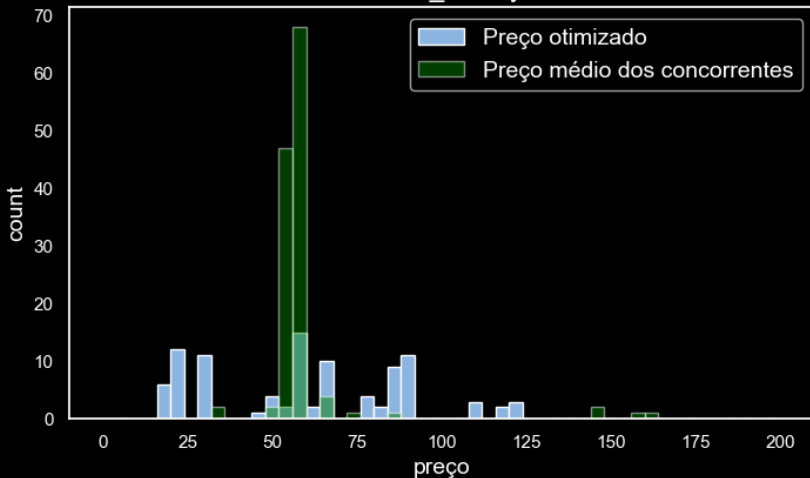
bed_bath_table



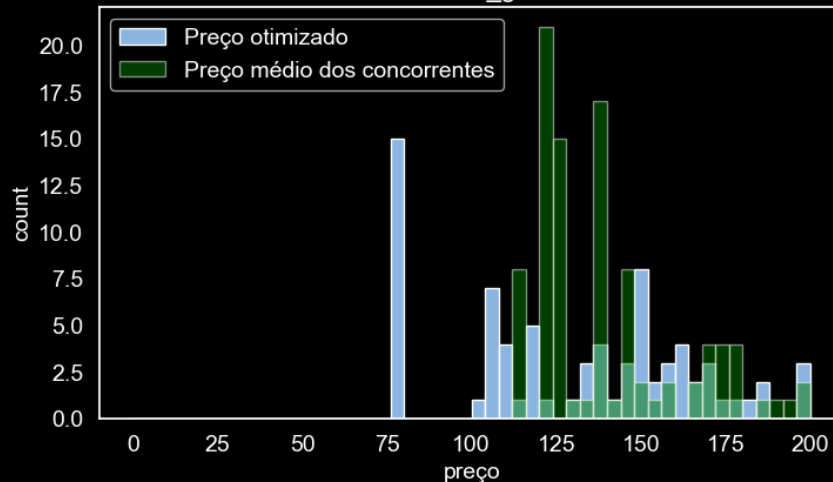
furniture_decor

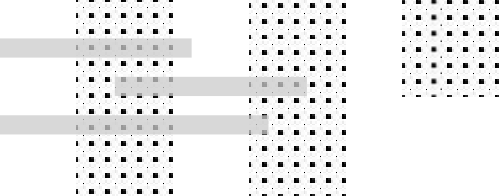


health_beauty



watches_gifts





Sugestão de novas variáveis melhorar as previsões de preço e demanda e otimização de preços

- Demanda histórica: O histórico de vendas é muito limitado. É necessário um período mais longo e mais detalhado – diário ou pelo menos, semanal
- Custo de produção/aquisição: Conhecer os custos associados à produção ou aquisição dos produtos é importante para garantir que os preços estabelecidos sejam lucrativos.
- Elasticidade de preço: A elasticidade de preço da demanda indica como a quantidade demandada de um produto muda em resposta a mudanças no preço. Essa informação é muito útil para ajustar os preços de forma ótima.
- Segmentação de clientes: Diferentes segmentos de clientes podem ter sensibilidades de preço distintas. A personalização dos preços com base no comportamento do cliente pode ser vantajoso.
- Feedback dos clientes: Além da nota de avaliação, comentários dos clientes, pesquisas de satisfação e outras formas de feedback podem ser úteis
- Estratégias de desconto e promoção: Estudar o impacto de descontos e promoções na demanda e no lucro pode fornecer informações a determinar a eficácia dessas estratégias e otimizar sua utilização.



Independente do ambiente utilizado (GCP, AWS, AZURE, etc) o pipeline de dados deve ter a seguinte estrutura:

- **Aquisição de dados**

- Identificar as fontes de dados relevantes, como registros de vendas, dados de inventário, dados de preços, dados meteorológicos, dados econômicos, entre outros.

- **Limpeza e pré-processamento de dados**

- Remover dados duplicados, ausentes ou inconsistentes.
- Transformar os dados em um formato adequado para análise, como séries temporais para dados de vendas e preços.

- **Engenharia de variáveis (feature engineering)**

- Criar (se necessário) novas variáveis a partir dos dados existentes que possam ser relevantes para os modelos, como médias móveis, sazonalidade, indicadores de eventos especiais, entre outros.
- Criar variáveis de lag (atraso), que representam valores anteriores de vendas e preços, pois podem ser úteis em modelos de previsão temporal.

- **Divisão de dados**

- Separar os dados em conjuntos de treinamento, validação e teste.

Continua no próximo slide ...



- **Modelagem**

- Escolher os algoritmos de Machine Learning mais adequados para o problema, como modelos de séries temporais, modelos lineares, árvores de decisão, redes neurais, entre outros.
- Treinar vários modelos com diferentes hiperparâmetros e técnicas de regularização para encontrar o melhor desempenho

- **Avaliação do modelo**

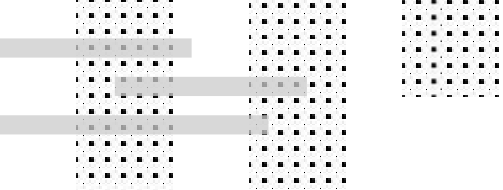
- Avaliar o desempenho dos modelos usando métricas adequadas escolhidas em conjunto com o time de negócio
- Comparar o desempenho dos diferentes modelos e selecionar o melhor modelo com base nas métricas de avaliação

- **Deploy e monitoramento**


- Fazer o deploy do modelo em um ambiente de produção, onde possa ser usado para fazer previsões em tempo real
- Monitorar regularmente o desempenho do modelo (retreinar se necessário)

- **Iteração e melhoria contínua**

- Analisar constantemente os resultados e sempre buscar maneiras de melhorar o pipeline de dados, os modelos e as estratégias de previsão com base no feedback do time de negócios.



Pontos críticos de monitoramento da pipeline

- Monitorar a qualidade dos dados de entrada, como a presença de valores ausentes, duplicados ou inconsistentes.
 - Verificar se as fontes de dados estão enviando dados conforme o esperado e se não houve interrupções na coleta de dados
 - Monitorar se não houve mudanças nas características dos dados que possam afetar a eficácia dos modelos
 - Monitorar o processo de treinamento do modelo para detectar possíveis problemas, como instabilidades durante o treinamento, tempos de treinamento anormalmente longos ou falhas no processo de treinamento
 - Verificar se os hiperparâmetros estão sendo ajustados corretamente e se não houve alterações significativas nos resultados de treinamento ao longo do tempo
 - Acompanhar regularmente as métricas de desempenho do modelo nos conjuntos de validação e teste
 - Identificar variações significativas nas métricas de avaliação que possam indicar deterioração do desempenho do modelo
 - Detectar qualquer degradação do desempenho do modelo em tempo real e implementar alertas para notificar sobre problemas.
 - Coletar feedback dos usuários sobre as previsões do modelo e monitorar a satisfação do usuário.
 - Realizar validação contínua do modelo com novos dados para garantir que continue a produzir previsões precisas ao longo do tempo.
- 

Conclusão e impressões gerais

- Realmente acho que é um case bem completo, capaz de avaliar várias características do candidato e representa um problema do mundo real
- O tempo estabelecido para a entrega do case é realista se você puder se dedicar apenas para ele. Caso você tenha outras atividades e responsabilidades, você tem que compensar em outros horários
- A qualidade dos dados é boa, embora limitados. O intervalo de tempo para análise de séries temporais é muito reduzido, o que afeta a qualidade da modelagem e os erros associados à previsão (forecast)
- A única variável que não encontrei na descrição das features foi “volume”
- Um conhecimento prévio, ou experiência prévia, nesta indústria específica, facilita a análise dos dados. Tanto na metodologia empregada como também na identificação de novas variáveis que melhorem o desempenho dos modelos criados
- No geral, foi um exercício bem gratificante. Aprendi bastante.