

Walmart



By: Raquel Rocha

QUESTÃO DE NÉGOCIO

O Walmart é uma corporação de varejo que opera como uma rede de supermercados.

Eles possuem os dados de 45 lojas, no período de 2010-02-05 á 2012-11-1. Existem em cada loja 99 departamentos (seção, setor, como exemplo, setor de produtos de limpeza, setor de pesca, setor de banho, etc...).

Também possuem os dados de vendas semanais em cada loja/departamento.

É preciso considerar que o Walmart tem uma política de promoções ao longo do ano, principalmente em datas que precedem feriados importantes (Super Bowl , Labor Day (dia do trabalho), Thanksgiving/Black-Friday e Christmas (natal)).

Com base na ideia de validar as promoções e sua efetividade, o negócio está interessado em saber?

1. Qual é o impacto dos feriados (promoções também) nas vendas das lojas?
2. Existe algum setor que desempenhe melhor?
3. É possível estimar as vendas das lojas por semana em datas futuras de 2012-11-02 a 2013-07-26? Se sim, quais seriam esses valores?

MÉTODO UTILIZADO

O projeto foi desenvolvido através da técnica CRISP-DM

Versão END-TO-END da solução,
Velocidade na entrega de valor,
Mapeamento de todos os possíveis problems.

Passo 01 - Descrição dos dados: Conhecimento dos dados, tipos, métricas estatísticas para identificar outliers, análise das métricas estatísticas e ajustes em features do dataset (preenchimento de NA's).

Passo 02 - Feature Engineering: Desenvolvimento de mapa mental para analisar o fenômeno, as variáveis e os principais aspectos que impactam cada uma delas.

Passo 03 - Filtragem dos dados: Filtragem das linhas e excluir as colunas que não são relevantes para o modelo ou não fazem parte do escopo do negócio.

Passo 04 - Análise Exploratória dos dados: Exploração dos dados para encontrar insights.

Passo 05 - Preparação dos dados: Preparação para as aplicações de modelos de machine learning.

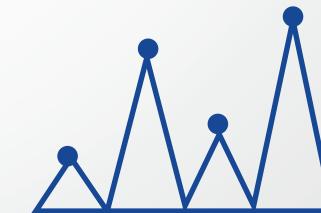
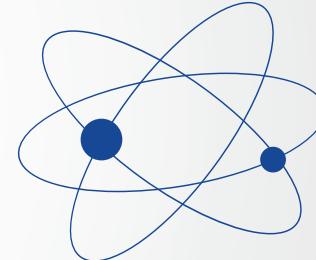
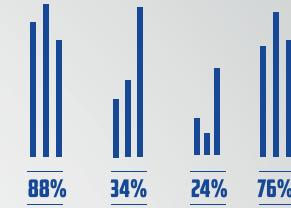
Passo 06 - Modelagem de Machine Learning: Foram realizados testes e treinamentos de alguns modelos de machine learning, para possibilitar a comparação da performance e escolha do modelo ideal para o projeto. Foi utilizada a técnica de Cross Validation para garantir a performance real sobre os dados selecionados.

ETAPAS DO PROJETO

- 01 - Questão de Negócio**
- 02 - Entendimento de Negócio**
- 03 - Coleta de Dados**
- 04 - Limpeza de Dados**
- 05 - Exploração de Dados**
- 06 - Análise de Dados**
- 07 - Apresentação**

Insights

Análise dos Dados



ESTATISTICA DESCRIPTIVA

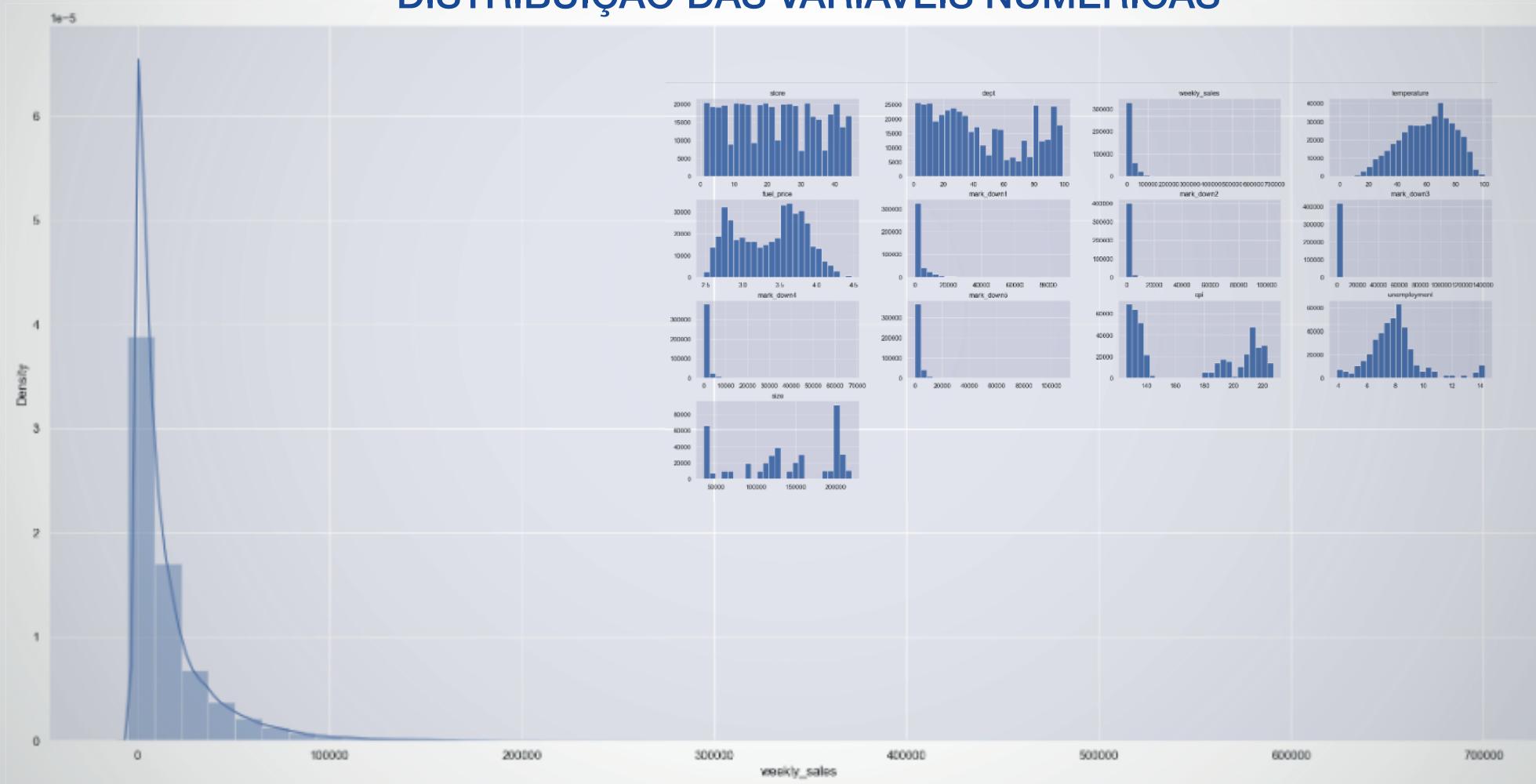
Observações relevantes:

- Valores mínimos negativos nas variáveis weekly_sales e markdown, que devem ser analisados e tratados a posteriori,
- Nenhuma das variáveis numéricas possuem distribuição normal,
- As variáveis: "cpi", "fuel_price" e "size" demonstram uma distribuição bimodal,
- A variável target possui valores negativos, os quais serão tratados em passo posterior.

	min	max	range	mean	median	std	skew	kurtosis	shapiro
store	1.0	45.0	44.0	22.200546	22.0	12.785282	0.077763	-1.146503	not normal
dept	1.0	99.0	98.0	44.260317	37.0	30.492018	0.358223	-1.215571	not normal
weekly_sales	-4988.94	693099.36	698088.3	15981.258123	7612.03	22711.156583	3.262008	21.49129	not normal
temperature	-2.06	100.14	102.2	60.090059	62.09	18.447909	-0.321404	-0.635922	not normal
fuel_price	2.472	4.468	1.996	3.361027	3.452	0.458514	-0.104901	-1.185405	not normal
mark_down1	0.0	88646.76	88646.76	2590.074819	0.0	6052.378756	4.731304	34.917236	not normal
mark_down2	-265.76	104519.54	104785.3	879.974298	0.0	5084.53277	10.645956	145.421293	not normal
mark_down3	-29.1	141630.61	141659.71	468.087665	0.0	5528.866895	14.922341	248.095371	not normal
mark_down4	0.0	67474.85	67474.85	1083.132268	0.0	3894.525326	8.077666	86.242339	not normal
mark_down5	0.0	108519.28	108519.28	1662.772385	0.0	4207.62433	9.964519	183.408065	not normal
cpi	126.064	227.232807	101.168807	171.201947	182.31878	39.159229	0.085219	-1.829714	not normal
unemployment	3.879	14.313	10.434	7.960289	7.866	1.863294	1.183743	2.731217	not normal
size	34875.0	219622.0	184747.0	136727.915739	140167.0	60980.511002	-0.32585	-1.206346	not normal

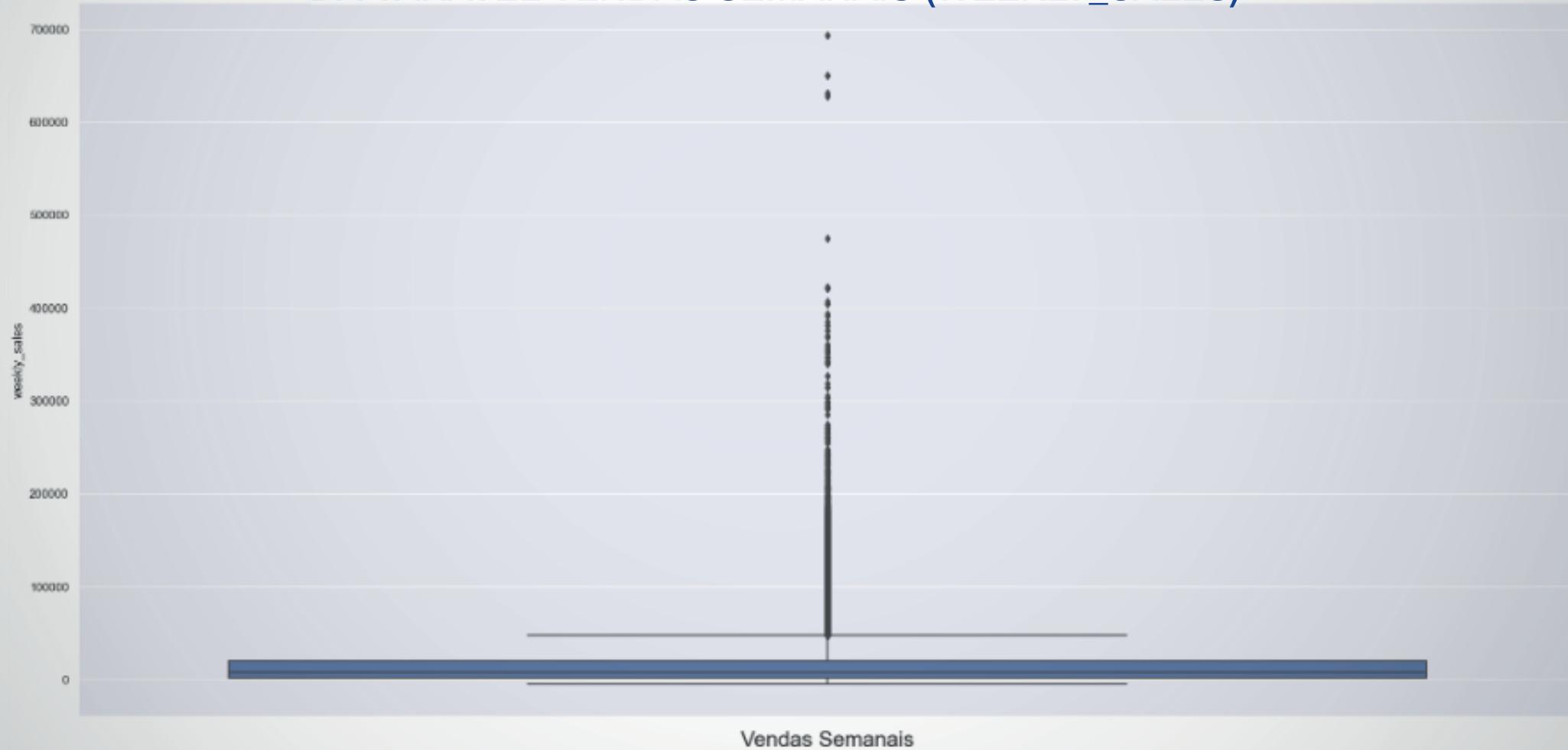
ESTATISTICA DESCRIPTIVA

DISTRIBUIÇÃO DAS VARIÁVEIS NUMÉRICAS



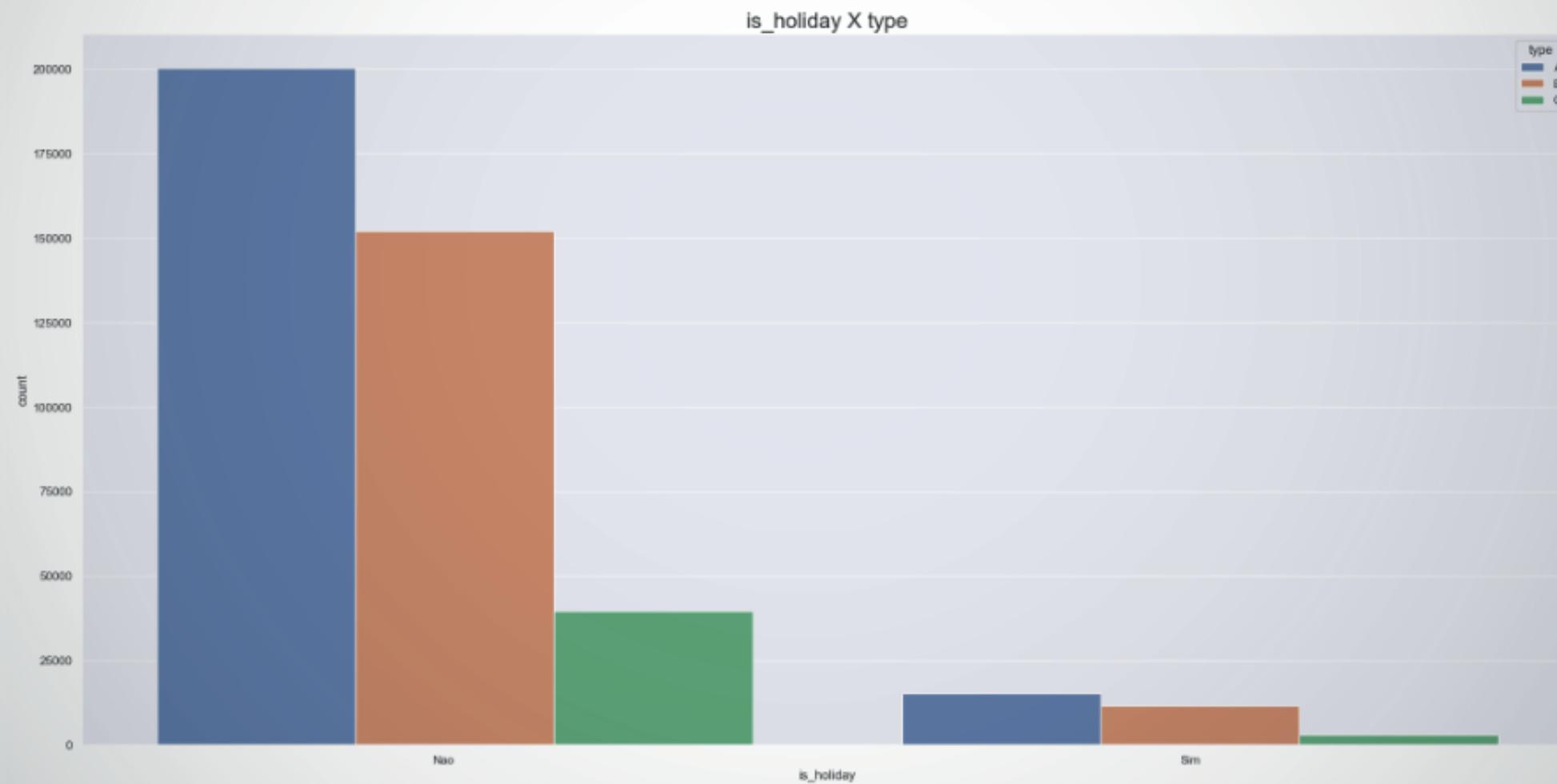
ESTATISTICA DESCRIPTIVA

BOXPLOT DEMONSTRANDO OS OUTLIERS DA VARÁVEL VENDAS SEMANAIS (WEEKLY_SALES)



ESTATISTICA DESCRIPTIVA

DISTRIBUIÇÃO DAS VARIÁVEIS CATEGÓRICAS



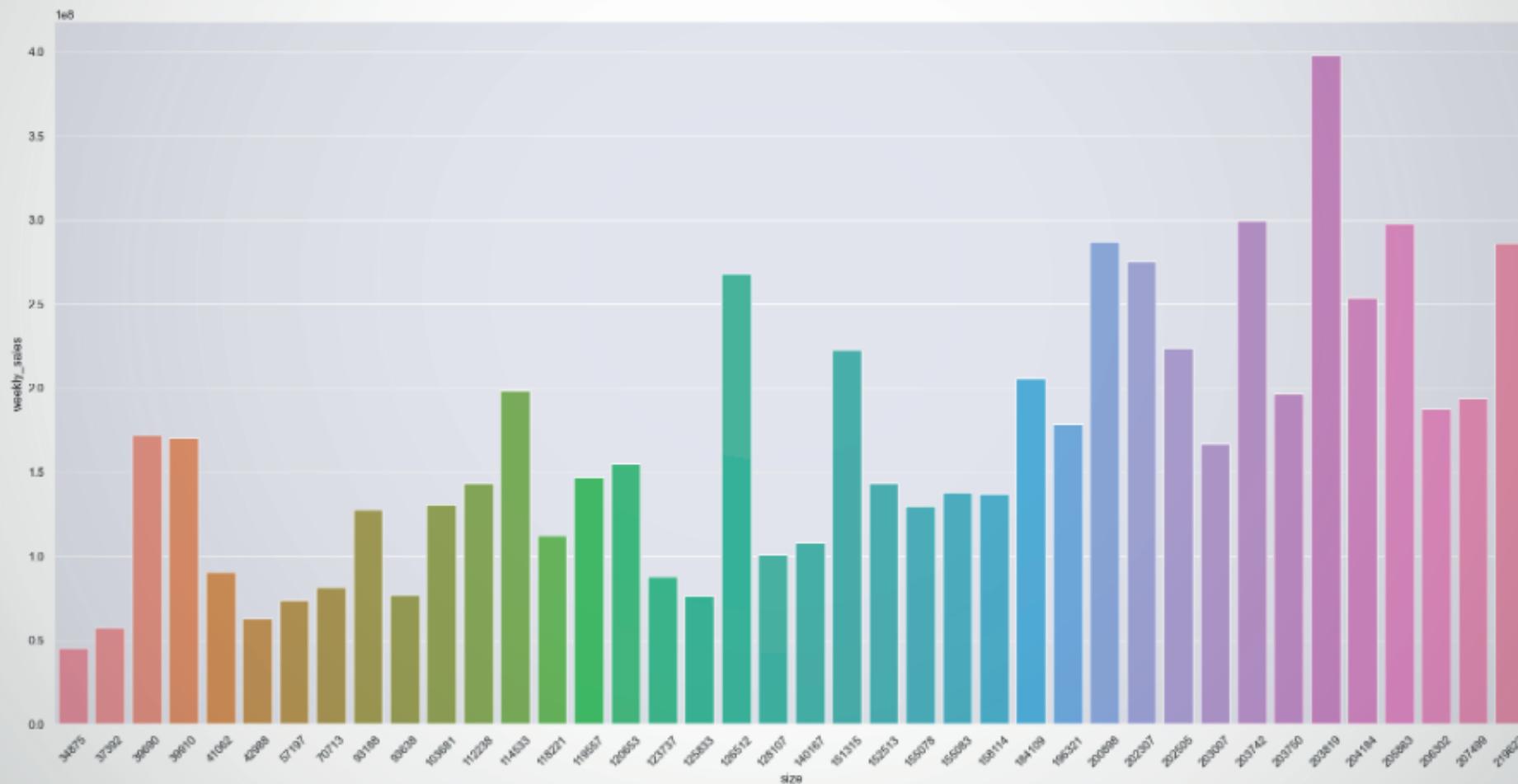
1

Lojas maiores
vendem mais?

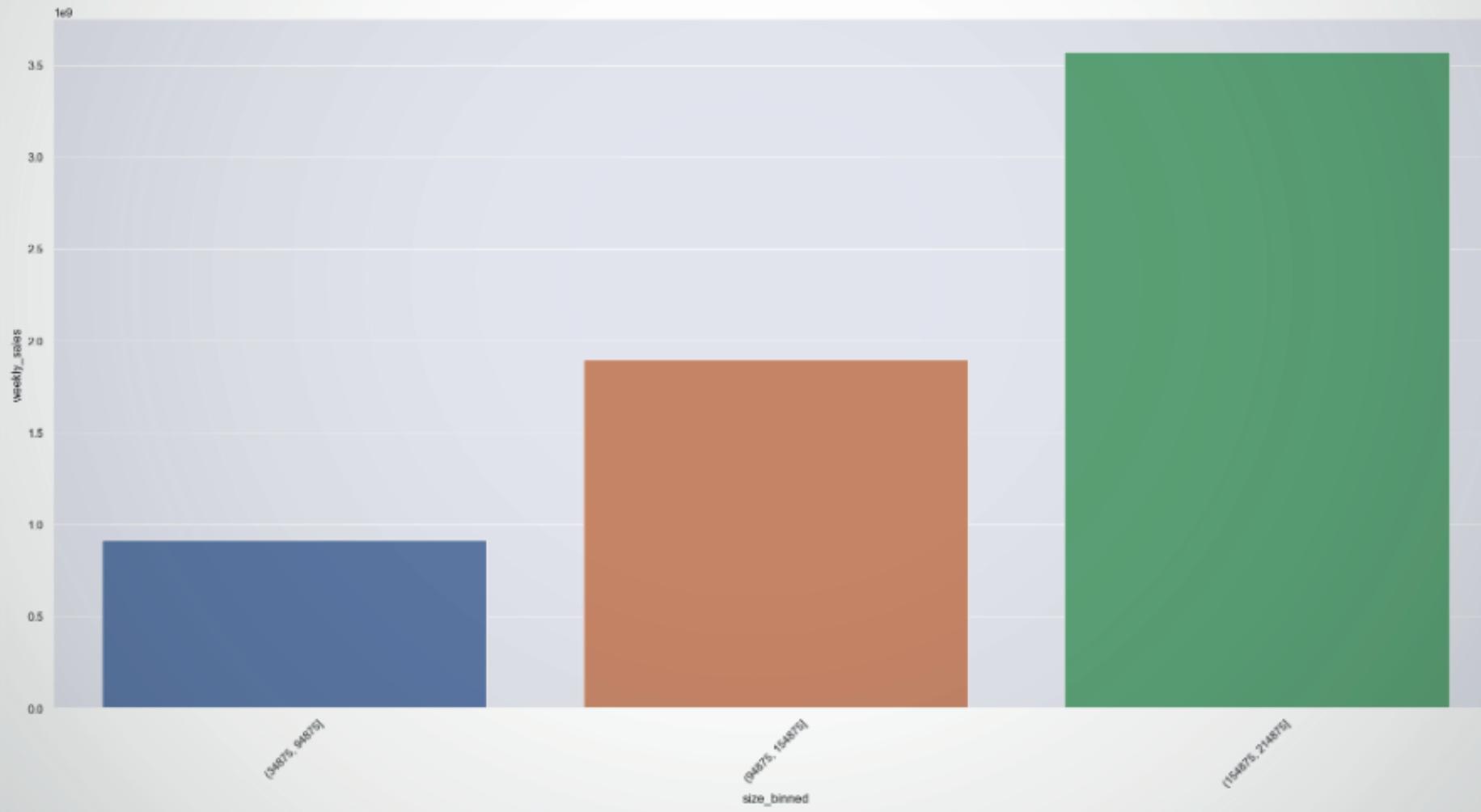
PERGUNTAS DE NEGÓCIO

Resposta:

No gráfico apresentado, a maior distribuição de vendas está nas lojas com tamanho maior, mas não podemos deixar de observar algumas lojas menores com quantidades de vendas bem acima de outras com tamanho próximo.



DISTRIBUIÇÃO DAS LOJAS, AGRUPADAS, EM UM GRÁFICO DE BARRAS



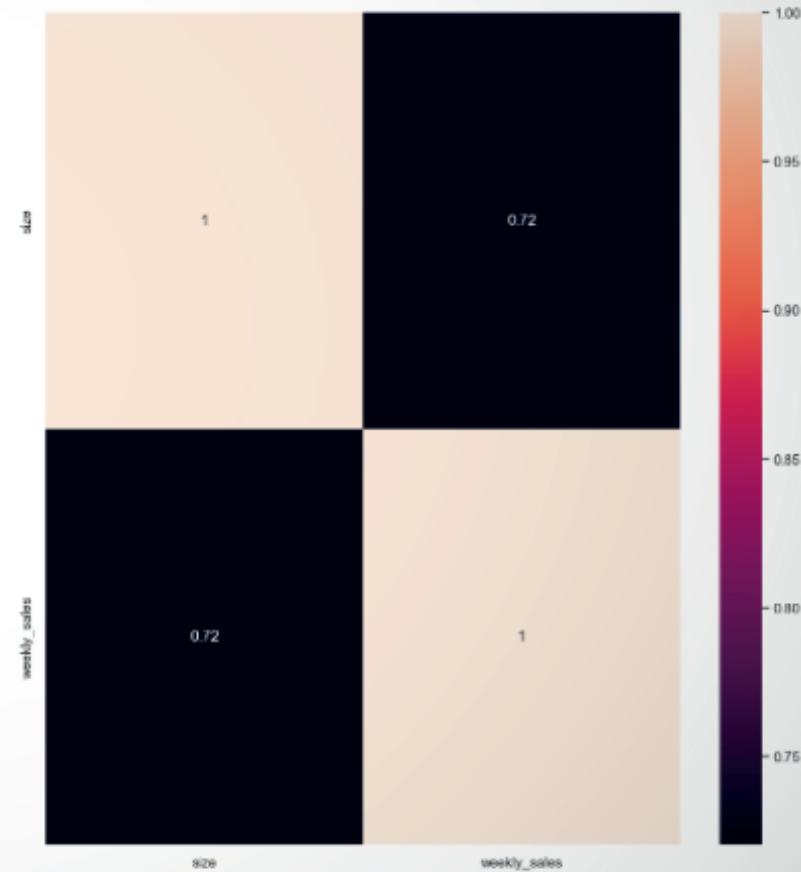
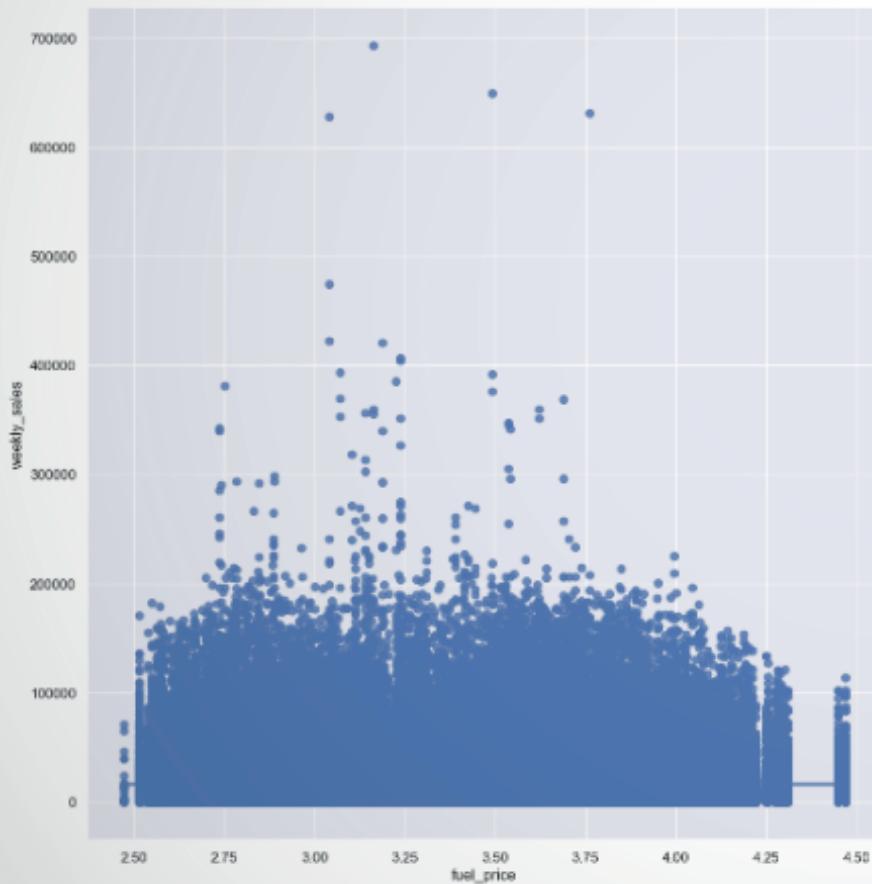
2

Lojas com custo de
combustível mais alto
vendem menos?

PERGUNTAS DE NEGÓCIO

Resposta:

A distribuição não demonstra dispersão forte, quanto as vendas semanais, dessa forma entendo que o custo de combustível da região não é um fator que tenha uma correlação direta com as vendas semanais (demonstrada também no gráfico de correlação de Pearson).



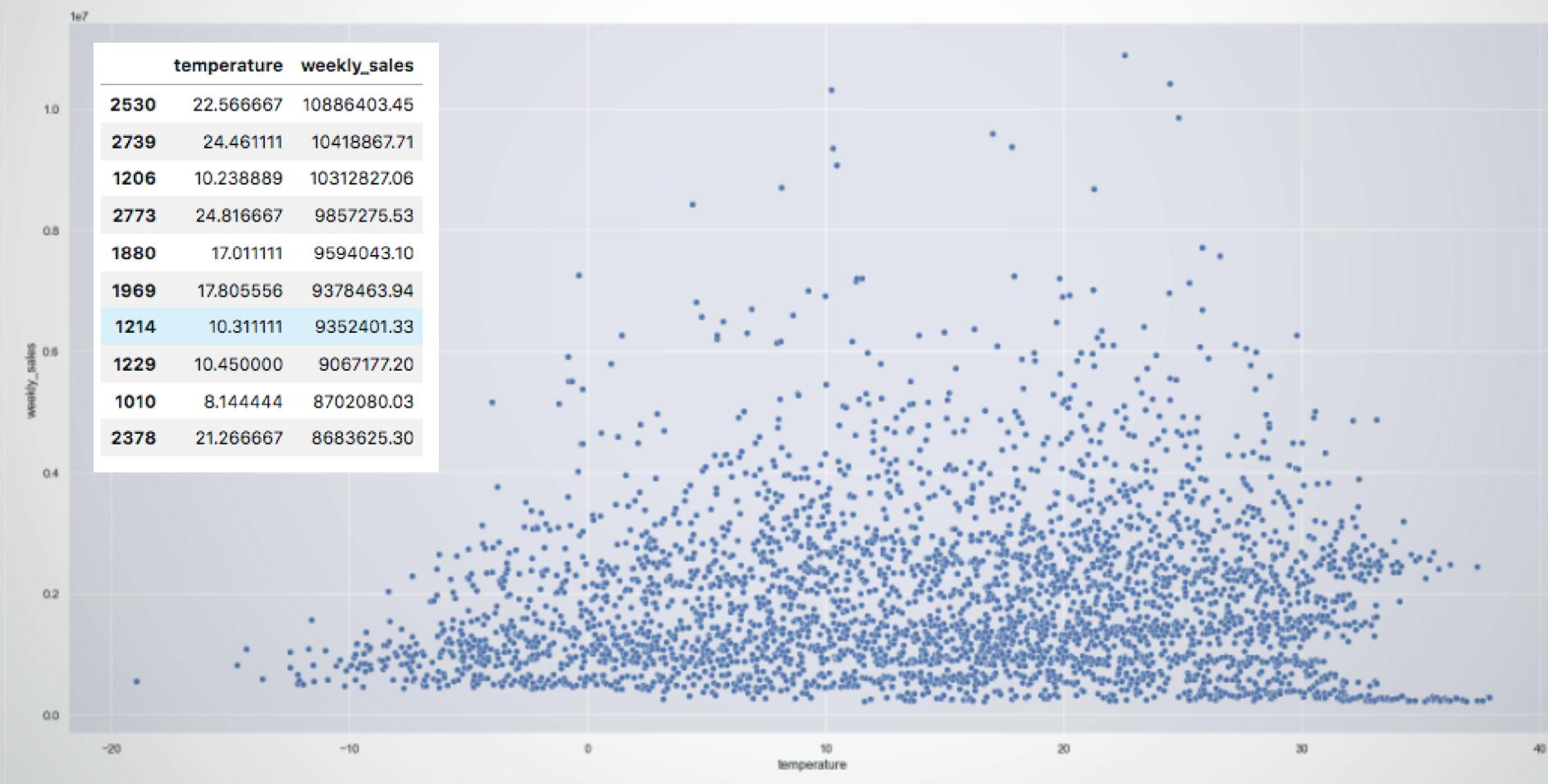
3

Dias mais quentes
tendem a vender menos?

PERGUNTAS DE NEGÓCIO

Resposta:

O comportamento de vendas é impactado com temperaturas abaixo de 0. Apesar do gráfico mostrar distribuições de vendas maior com temperaturas mais altas, não é observado grandes alterações no comportamento das vendas entre temperaturas de 12 a 35 graus celsius.



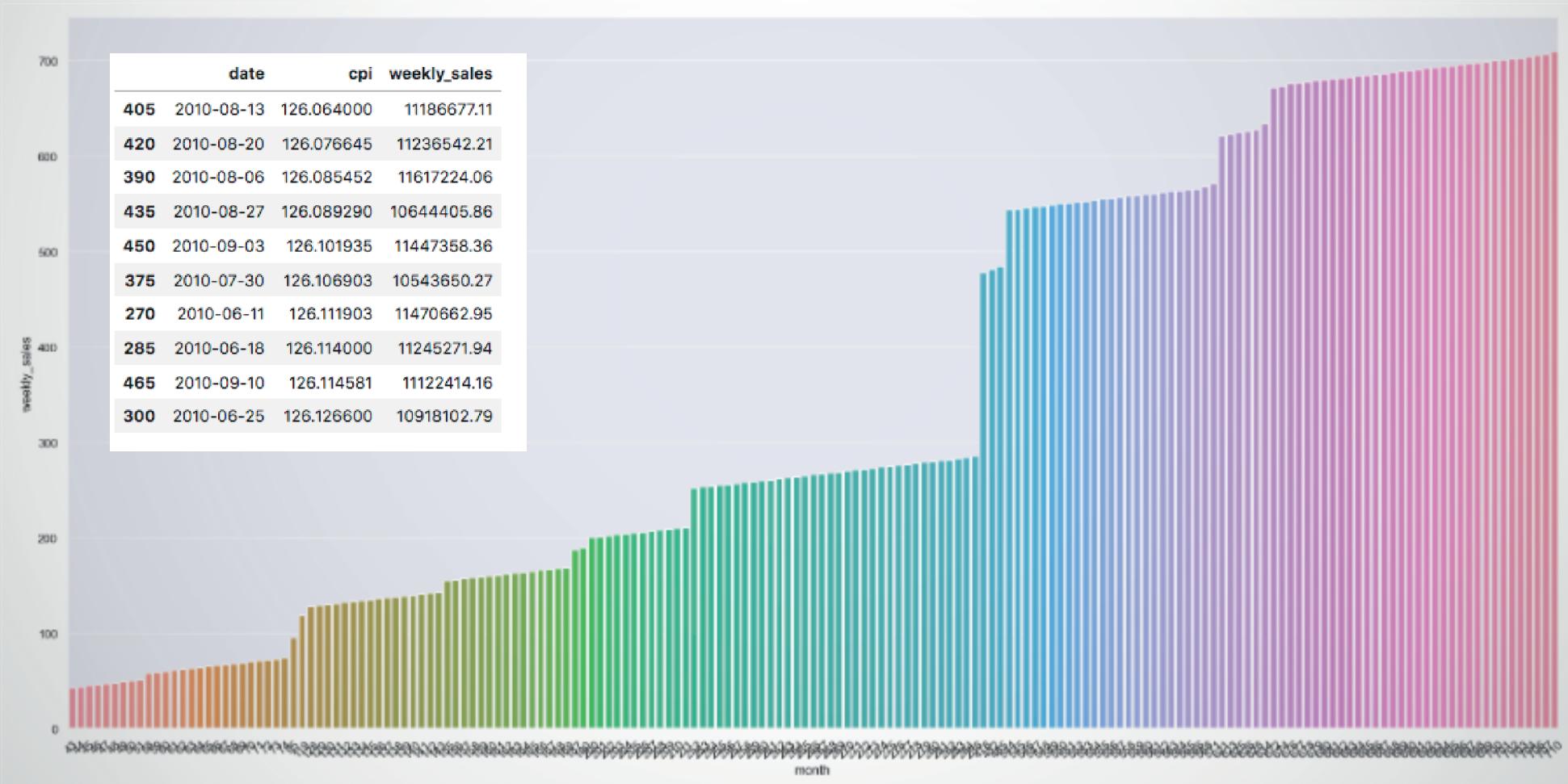


Meses com variação
do índice de preços ao
consumidor maiores
deveriam vender menos?

PERGUNTAS DE NEGÓCIO

Resposta:

O aumento da variação do índice de preços não afetou nas vendas, o gráfico demonstra que ao passar do tempo o comportamento de vendas foi exponencial.



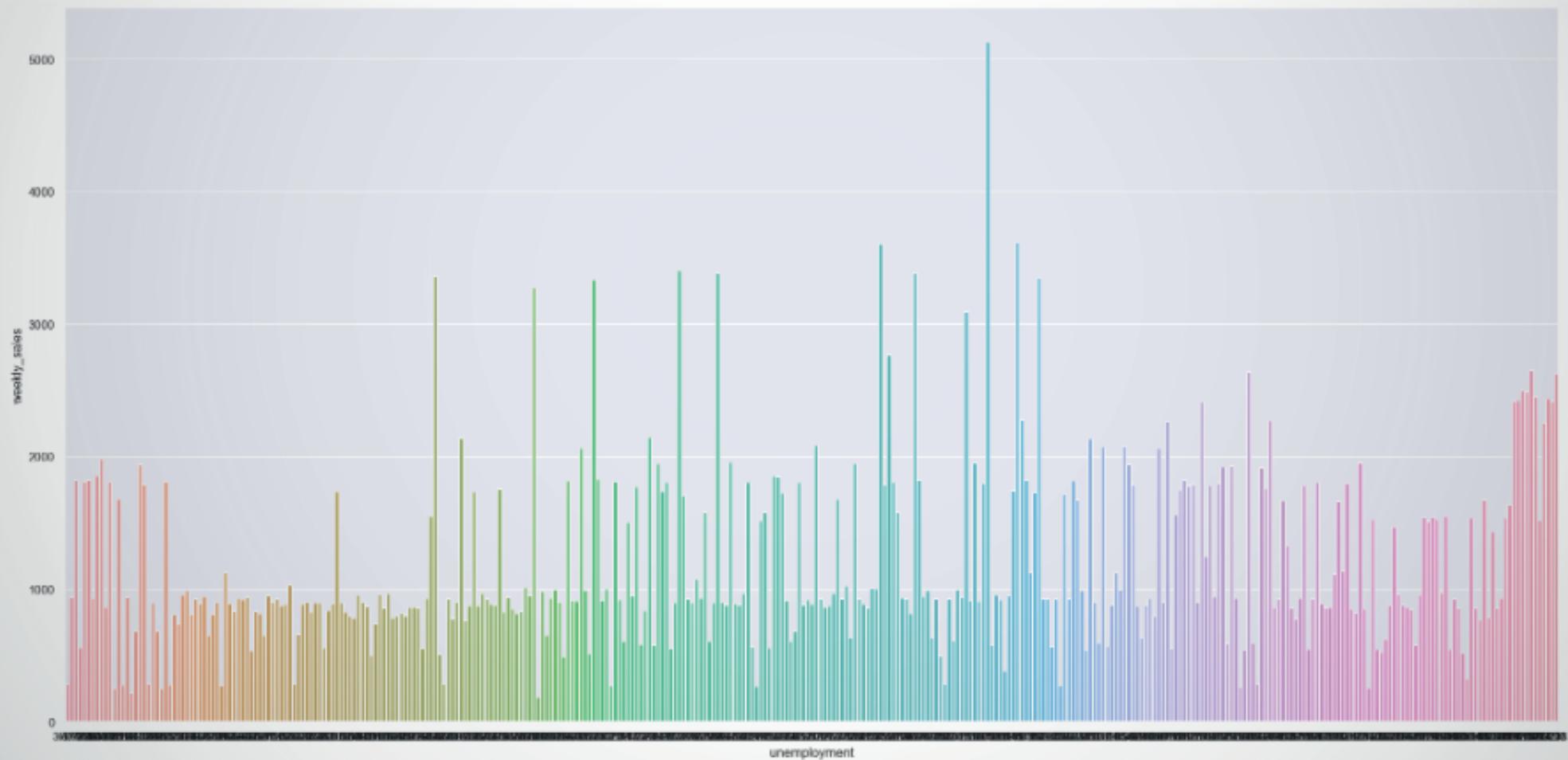
5

Ano em que a taxa de
desemprego foi maior
deveria vender menos?

PERGUNTAS DE NEGÓCIO

Resposta:

Ao longo dos anos o gráfico demonstra picos de vendas esparssas e também que ocorreram pequenos aumentos de vendas mesmo com a taxa de desemprego aumentando, mas no gráfico geral demonstra um comportamento que se manteve uniforme.





6

Lojas em dias de
promoção vendem mais?

PERGUNTAS DE NEGÓCIO

Resposta:

Apesar de vermos picos de vendas, considerando o período com marcação de promoção, o comportamento de vendas não mantêm níveis altos, pois observamos muitas quedas após esses picos de vendas.

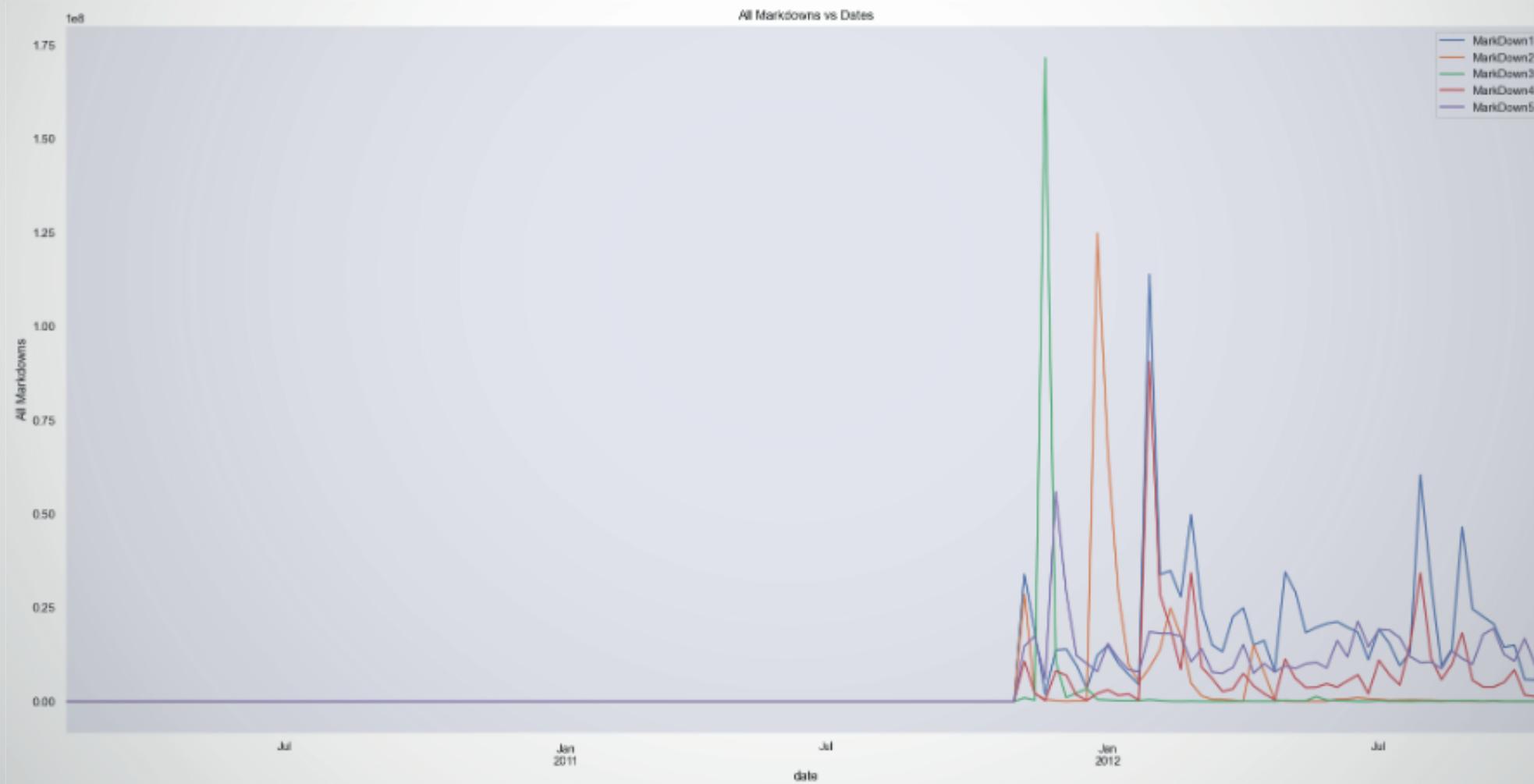
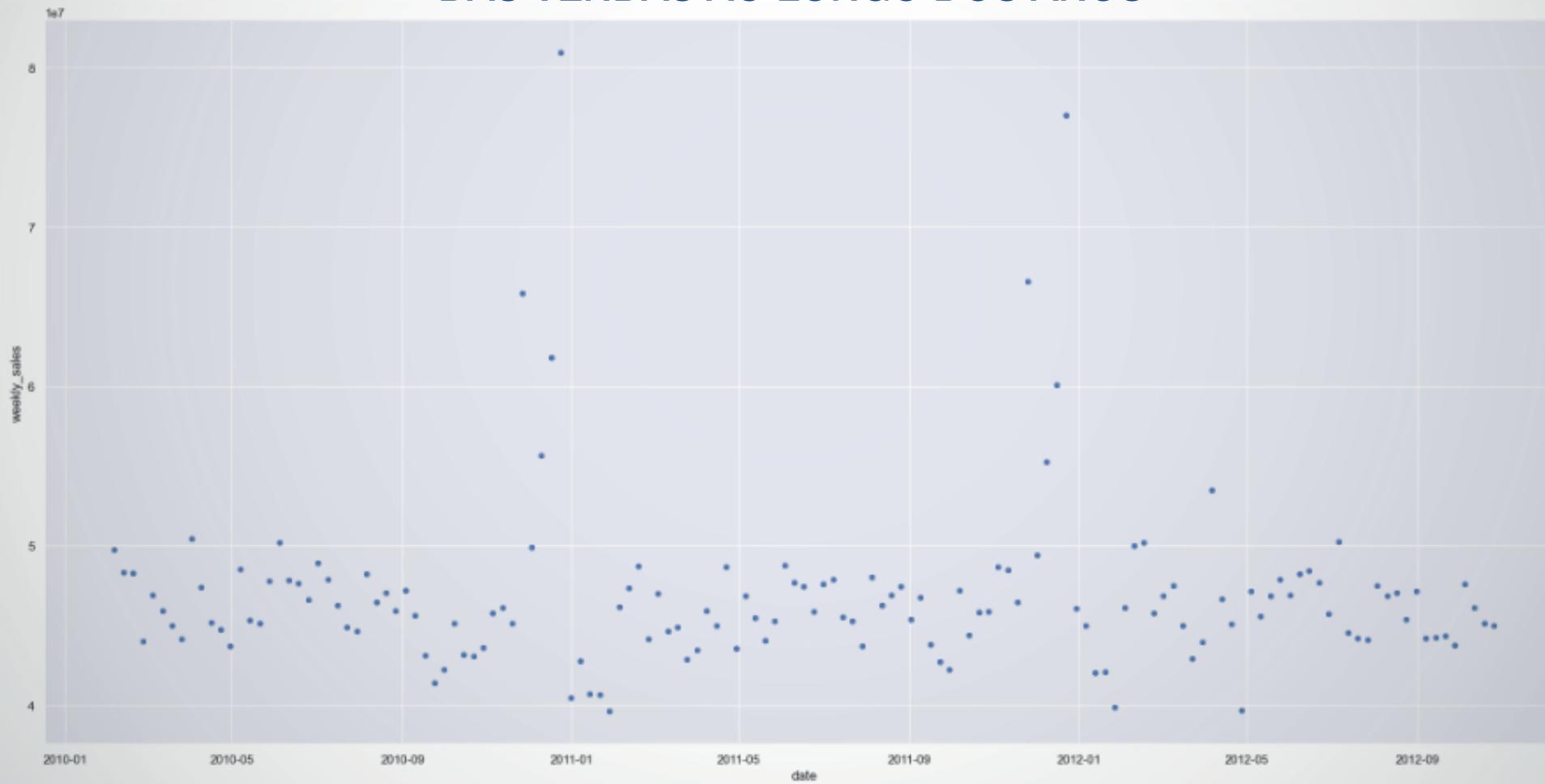


GRÁFICO DE DISPERSÃO DO COMPORTAMENTO DAS VENDAS AO LONGO DOS ANOS



7

Lojas abertas durante
período de feriados
importantes deveriam
vender mais?

PERGUNTAS DE NEGÓCIO

Resposta:

Os dados demonstram que a maior quantidade de vendas ocorrem em semanas sem feriado, apesar disso percebemos que a média de vendas nas semanas com feriados é de 1163.14 maior que a média de vendas nas semanas sem feriado. Essa visão agrupada, esconde um comportamento interessante, nos feriados de fevereiro e setembro ocorre um pequeno aumento nas vendas. Já em novembro, o comportamento das vendas tem um aumento considerável, enquanto em dezembro mostra um comportamento oposto, com baixas expressivas nas vendas. Embora não haja dados suficientes, podemos supor que as vendas para os feriados de dezembro são antecipadas em novembro, no feriado de Thanksgiving que antecede a Black Friday.

Não é feriado durante a semana

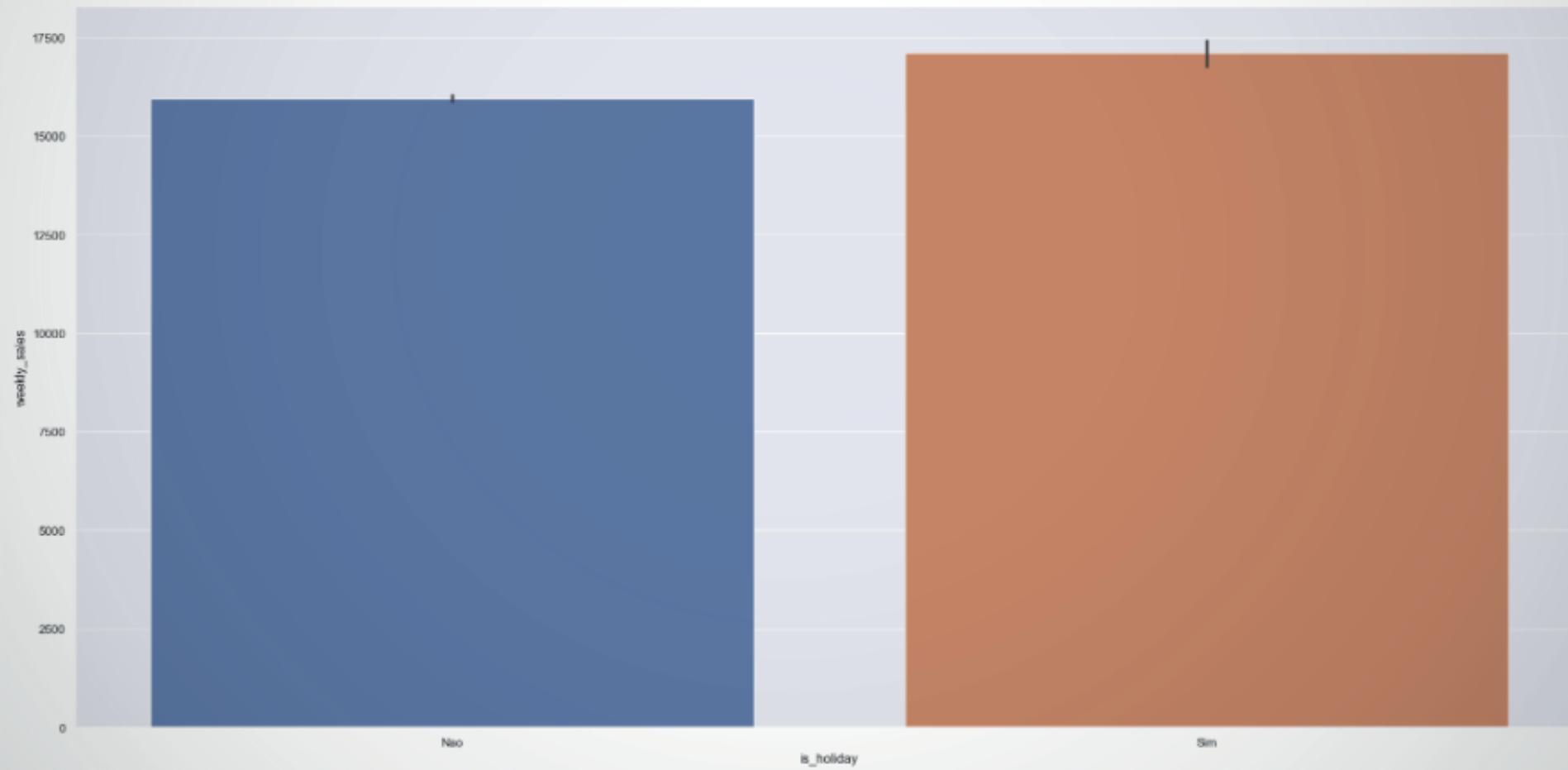
```
count    389434.000000
mean     15944.465963
std      22343.199043
min      0.000000
25%     2115.510000
50%     7632.320000
75%     20183.130000
max     406988.630000
Name: weekly_sales, dtype: float64
```

É feriado durante a semana

```
count    29293.000000
mean     17108.099010
std      27279.158431
min      0.000000
25%     2124.720000
50%     8004.350000
75%     21278.830000
max     693099.360000
Name: weekly_sales, dtype: float64
```

PERGUNTAS DE NEGÓCIO

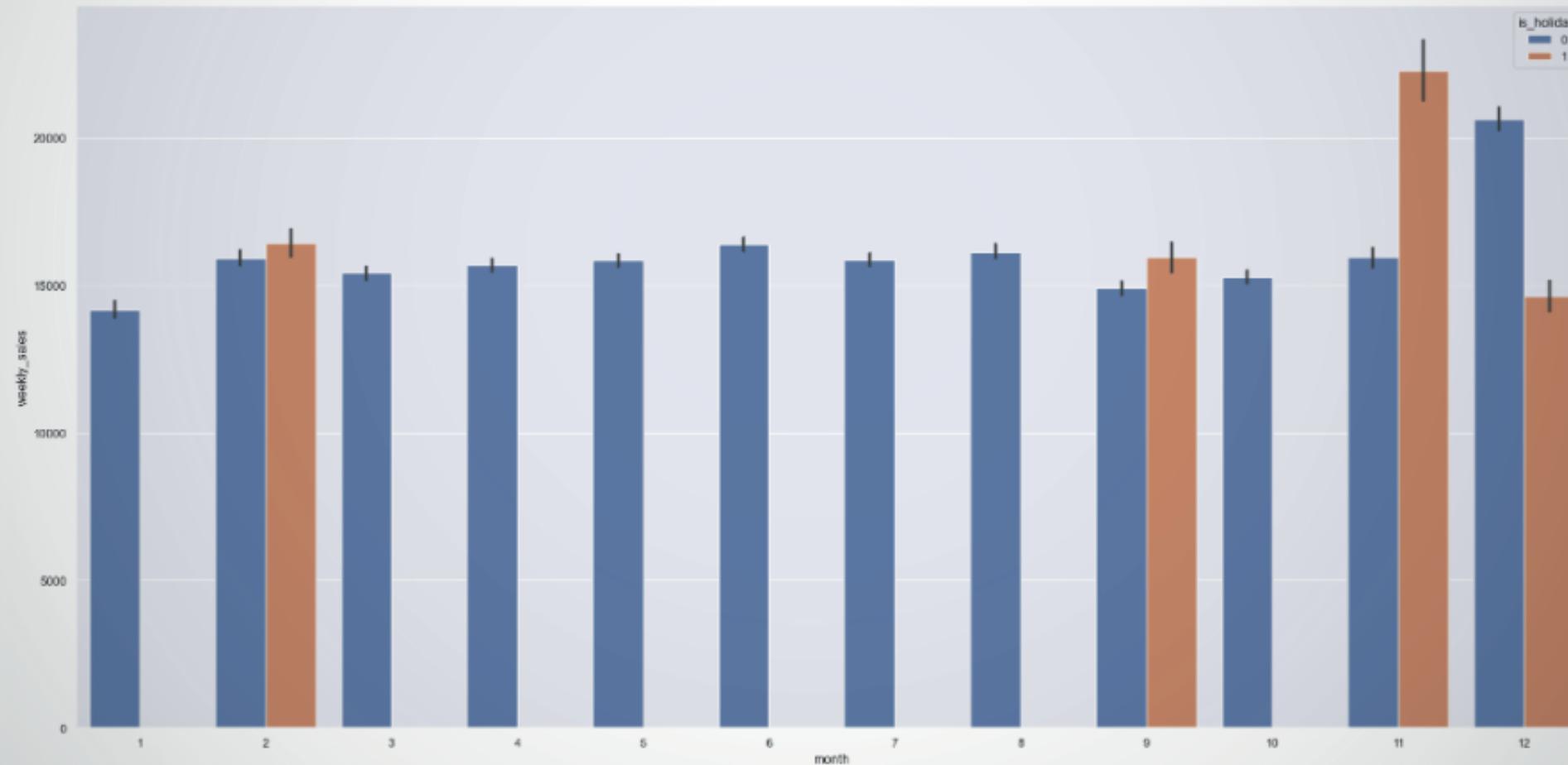
GRÁFICO DE BARRAS VENDAS SEMANAIS X FÉRIADOS*



*SE O FÉRIADO CAIU EM DIA DE SEMANA OU NÃO

PERGUNTAS DE NEGÓCIO

GRÁFICO DE BARRAS VENDAS SEMANAIS X MÊS*



*SE O FÉRIADO CAIU EM DIA DE SEMANA OU NÃO:
0 = NÃO E 1 = SIM

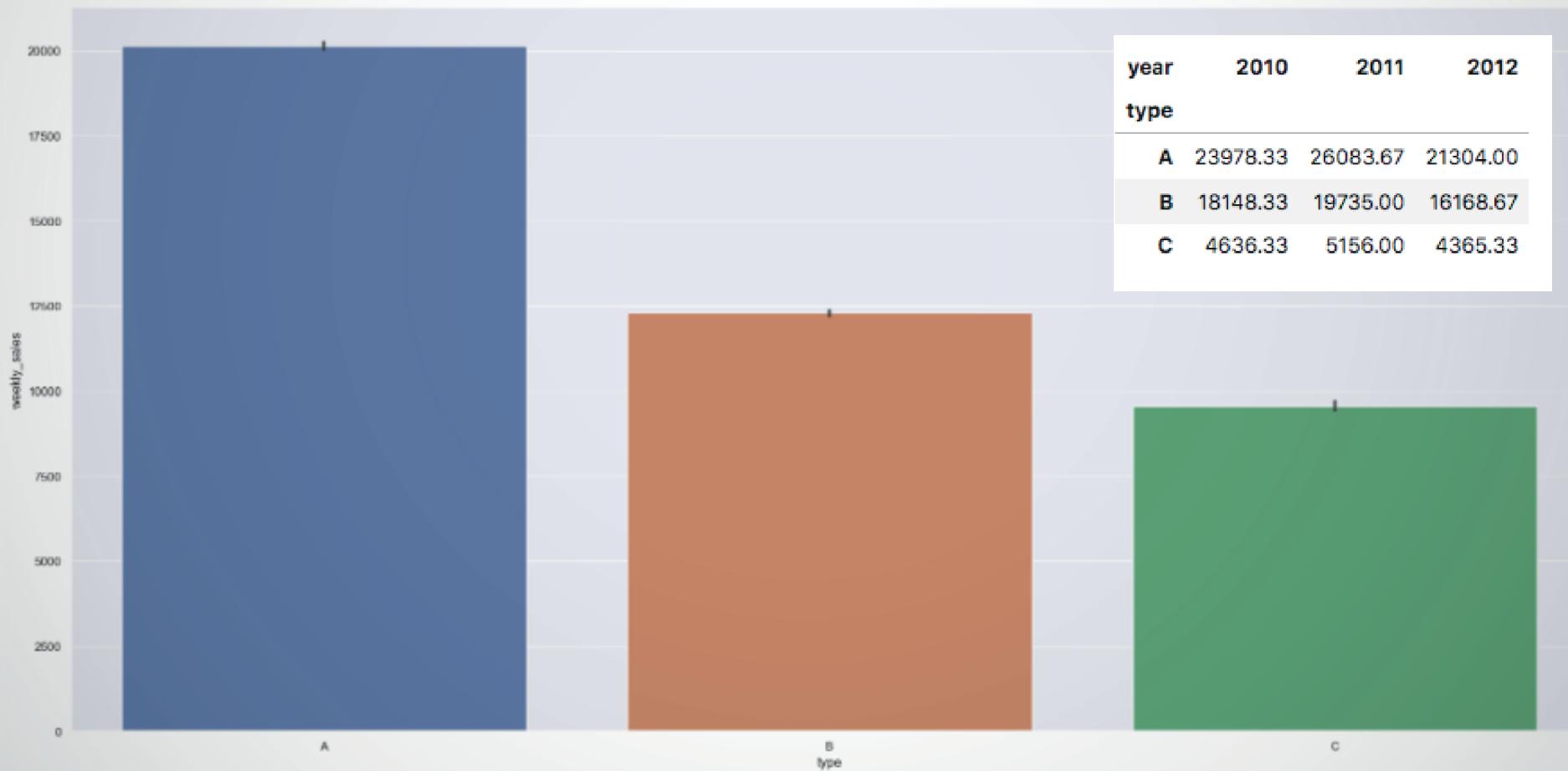
8

Setor de lojas tipo A
tem melhores resultados
com vendas?

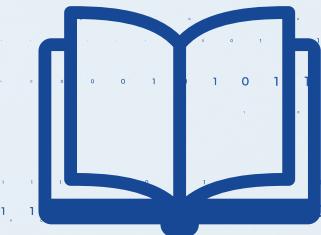
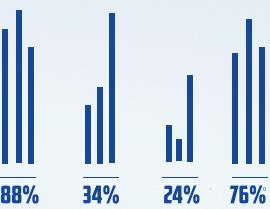
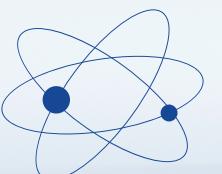
PERGUNTAS DE NEGÓCIO

Resposta:

Pela ordem demonstrada no gráfico, os tipos de lojas com melhores resultados são: A, B, C.
Inclusive ao longo dos anos entre 2010 a 2012, a loja de tipo "A" manteve vendas maiores.



Resultado



PERFORMANCE DO MODELO

A estratégia foi usar modelos lineares com o propósito de relacionar duas variáveis (resposta e explicativa) como também modelos não lineares que permitem ajustes de relações mais complexas.

Desta forma, respectivamente, podemos determinar se as médias dos grupos são diferentes e as causas de variação para entender o comportamento dessas variáveis, como também, fazer previsões fora do domínio observado de uma variável(x).

Para a realização desta etapa do projeto, foram aplicados os seguintes modelos:

Modelos Lineares

Média,
Linear Regression Regularized.

Modelos Não Lineares

Random Forest Regressor,
XGBoost Regressor.

	Model Name	MAE	MAPE	RMSE
0	Linear Regression - Lasso	0.000561	inf	0.000561
0	Random Forest Regressor	0.356232	0.000033	4.039910
0	KNN Regression	345.244671	inf	1505.370371
0	XGBoost Regressor	6385.084315	inf	11540.477985
0	Average Model	13706.385637	440.156251	20559.686148

RESULTADO FINAL

As vendas das semanas futuras estão sendo projetadas com sucesso. A equipe de dirigentes da Wallmar poderá consultar as previsões das vendas, podendo extrair informações do rendimento das lojas para melhores tomadas de decisão sobre suas filiais.

PREDIÇÕES DOS MODELOS

LINEAR REGRESSION - LASSO

	Actual	Predicted
9740	19616.22	19616.2500
9741	44493.61	44494.1201
9742	14288.22	14288.1921
9743	35044.06	35044.4446
9744	19369.52	19369.3878

RANDOM FOREST REGRESSOR

	Actual	Predicted
9740	19616.22	19616.219444
9741	44493.61	44494.1201
9742	14288.22	14288.219443
9743	35044.06	35044.059447
9744	19369.52	19369.519444

K-NEAREST NEIGHBORS

	Actual	Predicted
9740	19616.22	19614.570000
9741	44493.61	44419.368571
9742	14288.22	14305.062857
9743	35044.06	37117.942857
9744	19369.52	19366.947143

XGBOOST REGRESSOR

	Actual	Predicted
9740	19616.22	11362.215820
9741	44493.61	25069.363281
9742	14288.22	8464.820312
9743	35044.06	19870.058594
9744	19369.52	11247.692383

AVERAGE MODEL

is_holiday	weekly_sales	predictions
0	21379.88	12701.951652
0	21961.92	17104.821951
0	59483.14	13412.127627
0	389.60	23822.771574

PRÓXIMOS PASSOS - 2 CICLO

- Iniciar um segundo ciclo para analisar o problema, buscando diferentes abordagens,
- Obter mais dados,
- Trabalhar em combinar diferentes variáveis,

- Utilizar outras técnicas de tratamento nos dados NA,
- Utilizar técnicas de seleção de variáveis por meio do algoritmo Boruta ou “Feature Importance”,

- Fazer um ciclo com Hiperparâmetro para controlar o processo de treinamento do modelo e , consequentemente, conseguindo melhores resultados.

