PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS
ESCOLA POLITÉCNICA
GRADUAÇÃO EM ENGENHARIA DE COMPUTAÇÃO

CIÊNCIA DE DADOS APLICADA AO SETOR VAREJISTA

PEDRO DOUGLAS OLIVEIRA BEZERRA

GOIÂNIA
2021

PEDRO DOUGLAS OLIVEIRA BEZERRA

CIÊNCIA DE DADOS APLICADA AO SETOR VAREJISTA

Trabalho de Conclusão de Curso apresentado à Escola Politécnica, da Pontifícia Universidade Católica de Goiás, como parte dos requisitos para a obtenção do título de Bacharel em Engenharia de Computação.

Orientador: Prof. Me. André Luiz Alves

GOIÂNIA

2021

RESUMO

Por conta do crescimento assíduo dos dados produzidos diariamente, é evidente o crescimento de áreas de estudos relacionados a este fenômeno, a ciência de dados por sua vez é uma destas áreas utilizada para soluções de perguntas que ainda não foram possíveis obter resposta de forma tão clara, por meios analíticos tradicionais e isso é utilizado para todos os segmentos do mercado, porém tem uma adesão maior no campo do varejo e mercado de consumo. Contudo o varejo por sua vez representa toda e qualquer venda de consumo de bens ou serviços direcionadas aos consumidores finais para o consumo pessoal. De modo que a ciência de dados tem como seu objetivo o auxílio na tomada de decisão baseada em dados, onde se procura por habilidades de modelagem de dados, para conversão em informações relevantes, auxiliando assim principalmente no processo de tomada de decisão de empresas do setor varejista, trazendo possíveis contribuições para o desenvolvimento do varejo.

Palavras chaves: *Ciência de Dados, Varejo, Análise de Dados*.

ABSTRACT

Due to the steady growth of data obtained daily, the growth of study areas related to this phenomenon is evident, data science, in turn, is one of those areas used to solve questions that have not yet been answered so clearly, by traditional analytical means and this is used for all market segments, but it has a greater adherence in the field of retail and consumer market. However, retail in turn represents all consumer sales of goods or services directed to final consumers for personal consumption. So, data science has as its objective to aid in taking a database, where data modeling skills are searched for, for conversion into relevant information, thus helping mainly in the decision-making process of companies in the up-to-date sector, bringing possible contributions to the development of retail.

Keywords: *Data Science, Retail, Data Analysis.*

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS

ABCOMM	Associação Brasileira de Comércio Eletrônica
SBVC	Sociedade Brasileira de Varejo e Consumo

SUMÁRIO

1.INTRODUÇÃO

Cotidianamente se produz enormes quantidades de dados, devido ao aumento constante das tecnologias de informação e comunicação. Por conta desse crescimento notório, existe um debate constante no campo da ciência da informação, principalmente, em como a utilização dos dados por meio de rastros digitais que são produzidos por artefatos computacionais (celulares, relógios inteligentes, cartões de créditos etc.). Por este motivo convém a necessidade de um espaço interdisciplinar para o estudo destas questões polêmicas sobre dados, informações e conhecimento (RAUTENBERG E CARMO, 2019).

Apoiando nesta ideia de mundo, tendo em vista o avanço da Internet, a população em geral vem produzindo cada vez mais dados nas mais diversas plataformas digitais (Twitter, Facebook, Instagram etc.). Por conta disto são coletadas e armazenadas imensas quantidades de dados, como sinais de imagens, registros, vídeos e post (Rautenberg e Carmo, 2019). Devido a esse alto volume de dados que são gerados hoje em dia é possível criar modelos para prever diversas situações e a ciência de dados é um meio técnico para análise de dados, e facilita tomada de decisões fundamentadas, decidir ações de mercado, e também se preciso, construir produtos e serviços para seus clientes.

A ciência de dados é utilizada para soluções de perguntas que não foram possíveis obter as repostas de forma tão clara por meios analíticos tradicionais e isso é utilizado para todos os segmentos do mercado, porém tem uma adesão maior no campo do varejo e mercado de consumo (LUMINATTI, 2018).

O varejo simboliza uma atividade comercial encarregada de prover mercadorias e serviços pretendidos pelos consumidores ou também pode ser determinado como um negócio que compra produtos de distribuidores, atacadistas e entre outros produtores e comercializa diretamente com os consumidores finais. (Gouveia, et al, 2011). Em vista disso o setor varejista é um dos principais setores da economia mundial, que está movendo dezenas de trilhões de dólares anualmente.

A análise de dados no varejo consegue orientar as empresas a monitorizar tendências de seus consumidores, utilizando análises em perfis de clientes no varejo para encontrar, compreender e atuar com informações significativas, dentre esses os padrões de clientes online e físicos (JUNQUEIRA, 2020).

A função do cientista de dados é principalmente produzir estudos amplos e ponderados, utilizando técnicas de análises estatísticas e métodos computacionais avançados, dessa forma auxiliando a criação de *insights* para diversas áreas da empresa, ou seja, é possível responder

diversos questionamentos utilizando as técnicas adequadas.

É relevante estudar esse tema pois a ciência de dados proporciona especificar dados decisivos às empresas. Visto que permite compreender altamente a dinâmica de mercado atual a partir dos dados, apoiando-se nos inúmeros dados que são coletados e analisados. Tendo em vista isto, os modelos analíticos apresentam mais clareza nas tomadas de decisões, também apresentam as melhores escolhas a serem tomadas, utilizando soluções prescritivas ou quais melhores previsões utilizando soluções preditivas e identificando recomendações da Ciência de dados que podem melhorar os resultados de empresas do setor varejista.

O trabalho foi estruturado seguindo a metodologia do Raunteberg e do Carmo (2019), ou seja, a partir do ciclo de dados apresentado por eles em sua pesquisa, foi feito a aplicação das etapas, para que assim seja possível extrair informações importantes dos dados da empresa.

2. REFERENCIAL TEÓRICO

2.1 Ciência de dados

A ciência de dados é um campo de estudo que trata obter conhecimento e informação, a partir de dados estruturados ou não, utilizando métodos e modelos computacionais e estáticos visando auxiliar na tomada de decisões de instituições (Fontes, 2020). Ainda que o nome ciência de dados vincula intensamente a ideia do estudo de banco de dados, ciência da computação, na área deste campo de estudo também é necessário habilidades não matemáticas, as fundamentais habilidades destacadas pelo breve estudo de caso incluem habilidades de comunicação, habilidades de análise de dados e habilidades de raciocínio (STANTON, 2013).

Dentre os objetivos desta área pode se destacar o auxílio na tomada de decisão baseada em dados, para isto são necessárias diversas competências. Uma analogia para melhor entendimento do que é ciência de dados seria visualizar ela como uma "caixa de ferramentas" onde se procura por habilidades de modelagem de dados, para manipular os dados não estruturados e em larga quantidade; o aprendizado de máquina para utilizar a verificação de padrões e realizar abstrações profundas; a matemática e estatística para desenvolver modelos e distribuições sólidas; e a análise de dados para conseguir identificar tendências, tanto de mercado como de comportamento para assim auxiliando na tomada de decisão.

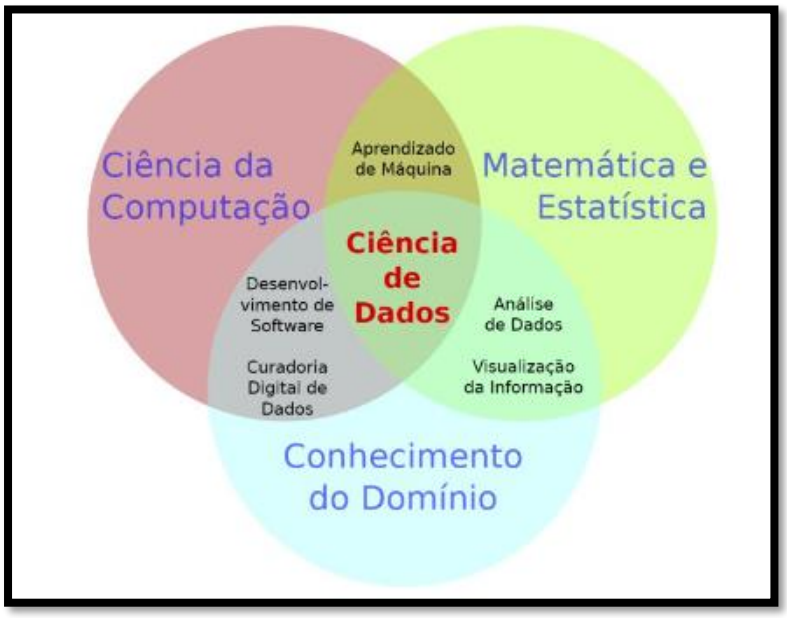
Os domínios do conhecimento da ciência de dados conforme abordado por Rautenberg e Carmo (2019), são ressaltadas habilidades na área da ciência da computação, tendo que principalmente os dados são armazenados e processados por computadores. Nesta situação os ambientes computacionais são ferramentas essenciais para a implementação do aprendizado de máquina e para acarretar a curadoria digital e para interfaces de visualização da informação. É importante saber como utilizar tecnologias para a acessar e converter os dados para ser possível abstrair em uma informação útil.

Conhecimento no que se refere a Matemática e Estatística e necessário para desenvolver atividades de Análise de Dados, ou seja os profissionais que atuam na ciência de dados devem conhecer bem a forma com que os Algoritmos de Aprendizado de máquina funcionam, também como interpretar os resultados destes algoritmos estatisticamente. De forma interdisciplinar pode ser considerado como a interpretação é facilitada pela Visualização da Informação.

Para conseguir obter o devido sucesso das possíveis soluções de Ciência de Dados, o conhecimento no que se refere ao Domínio do problema, deve ser amplamente utilizado na abordagem do processo de Tomada de Decisão. Dessa forma as soluções de Ciência de Dados são designadas para que seja formulada hipóteses e informações que se adere ao processo de

decisão. Abaixo podemos visualizar na figura 1 a interdiscipliaridade da Ciência de dados.

Figura 1: Interdisciplinaridade da Ciência de Dados



Fonte: Rautenberg e Carmo (2019).

Para que seja possível aplicar a ciência de dados é necessário primeiramente entender como este ciclo de vida dos dados ocorre nas empresas. Conforme abordado por Rautenberg e Carmo (2019). Abaixo é descrito o ciclo de vida da ciência de dados:

Obtenção de Dados: Inicia a realização de operações para avaliação e seleção de dados primários e seus metadados a partir de processamentos de arquivos tantos de texto ou de monitoramentos de base de dados, e de dados proveniente da web, dentre outros.

Ingestão de Dados: Se trata de uma conversão e carga dos dados de origem, obtido a partir de fontes diversas e formatos diversificados, em uma base de dados centralizada. Esta etapa tem como ventura organizar e representar os dados, a tal que é utilizado os recursos de pré-processamento em um único repositório de dados principal.

Exploração de Dados: Favorece a aplicação de estudos iniciais, para assim estabelecer deduções iniciais dos dados que foram disponibilizados, tendo em vista a informação requisitada. Por tanto está etapa é de extra importância para que seja possível definir o fluxo do trabalho, estabelecendo assim o roteiro de trabalho e como relacionar os dados primários à

informação relevante.

Definição dos Parâmetros: Está etapa está profundamente ligada as preferências necessárias para ocupação do(s) algoritmo(s) de Aprendizado de Máquina. Podemos definir está etapa em alguns processos, iniciando o primeiro em converter os dados de entrada de acordo com requisitos do algoritmo de aprendizado de máquina. O segundo em converter os dados de saída, para que assim seja possível ser legível ao ser humano. O terceiro em definir os intervalos dos parâmetros de entradas a serem relevantes. O quarto em definir os parâmetros de parada do algoritmo de aprendizado de máquina e o último em definir o nível de confiabilidade exigido da resposta que foi gerada, dentre outros processos podem ser tomados.

Implementação do Modelo: Tende-se a se utilizar algoritmos de Aprendizado de Máquina para definir modelos a partir dos dados iniciais de entrada e os dados de saída. Isto está envolvido nas definições das estratégias de treinamento e teste definidas para o algoritmo de Aprendizado de Máquina, para assim definir as estratégias mais adequadas dentre aqueles requisitos que foram avaliados. Após se obter o resultado, deve-se então se abstrair em um modelo represente melhor as características dos dados utilizados.

Utilização do Modelo: Após já ter definido o modelo, então irá poder utilizá-lo para concluir informações sobre os dados utilizados no ambiente de produção. Por tanto irá confirmar a habilidade de generalização do modelo, desta forma o modelo poderá ser empregado em Tarefas Intensivas em Conhecimento.

Tomada de Decisão: A partir dos resultados gerados pelo a utilização do modelo, com o seu conhecimento especializado em análise de dados, o gestor tem fundamentação para suas decisões tomadas. Uma parte importantíssima nesta etapa envolve uma elaboração da apresentação dos dados então obtidos, e da visualização das informações a partir de gráficos e relatórios. Assim auxiliando na visualização de *insights* mais claros, nas atividades dos tomadores de decisão.

Essa jornada do ciclo de vida da ciência de dados pode ser visualizada conforme abaixo na figura 2.

Figura 2: Ciclo de vida da Ciência de Dados

Fonte: Rautenberg e Carmo (2019).

Tendo em conta as etapas relatadas, é possível extrair informação útil mediante a dados brutos, ou seja, pode ser definido como um processo interativo. Partindo deste ponto, as pessoas envolvidas podem auxiliar a formular premissas iniciais a respeito do problema, e então adicionar novos dados assim refinando as ideias e soluções. Em outras palavras, a partir de um conjunto bruto e volumoso de dados, e assim sendo possível encontrar outra forma para sua apresentação, mas abstrata e útil assim auxiliando ainda mais no processo de Tomada de Decisão.

2.2 Varejo

O varejo tem seu início nos Estados Unidos, por volta do século XIX, quando ocorreu o surgimento dos chamados *general stores,* ou também chamadas de lojas de mercadorias gerais, onde era comercializado mercadorias como: tecidos, armas e munições, alimentos entre outras. No ano de 1886 começou a Sears, um varejo que se vendia mediante catálogos, que futuramente se transformou em uma loja de departamentos (GOUVEIA, ET AL, 2011).
Pode-se definir o varejo por um processo de compra de mercadorias em grandes quantidades de distribuidores atacadistas dentre outros fornecedores e após isso a venda em quantidades menores para o consumidor final.

A definição de varejo segundo Parente (2000), corresponde nas atividades que compõe

o procedimento de venda de serviços e produtos, para que assim atendam às necessidades do consumidor final. Também é definido por Giuliani (2003) que o varejo é um comerciante que comercializa serviços e produtos, tanto para o uso pessoal ou familiar, aos seus consumidores, definindo assim como o último comerciante de um canal de consumidores.

Conforme abordado por Kotler (1998) o varejo inclui todas as atividades de vendas de serviços e produtos, que são direcionadas ao consumidor final.

Portanto o varejo compõe todas as atividades que se referem à venda de serviços e produtos para consumidores finais, para o uso pessoal, não relacionando a negócios, ou seja, toda e qualquer empresa que está vendendo produtos e serviços está praticando o varejo. Por isso existe inúmeros meios para praticar a venda do varejo, como por exemplo por telefone, por máquinas de vendas, de formas pessoais, nas ruas ou nas casas de seus consumidores, e por meio da Internet.

Ainda sobre os tipos de meios que se caracteriza varejo, as vendas na Internet que são normalmente denominadas pelo termo varejo online, de acordo com as informações da Associação Brasileira de Comércio Eletrônica (ABCOMM), apenas esse segmento do varejo apresentou uma receita de R$ 35 bilhões, isto apenas no primeiro semestre de 2019. Uma melhor definição para o varejo online seria a comercialização de serviços ou produtos que acontece na Internet. Portanto como sempre seguindo a definição de varejo sem intermediários e com o foco no consumidor final.

As compras e vendas pela Internet é um fenômeno corriqueiramente mais comum nos últimos anos. Quem poderia imaginar que tal fenômeno poderia permitir as pessoas fazer compras no supermercado entre outros estabelecimentos, sem que seja necessário sair de casa, (Santos, 2020). Segundo um levantamento que foi feito pela empresa BigdataCorp em conjunto com a PayPal Brasil, o setor do comércio eletrônico também definido por *e-commerce* teve um crescimento de 37,5% dos anos de 2018 e 2019, conforme abordado pelo estudo este crescimento foi o maior desde 2014.

O comercio eletrônico, ou *e-commerce*, que também engloba o varejo online, segundo o estudo feito pela Sociedade Brasileira de Varejo e Consumo (SBVC), a tendência para o crescimento desta área seria 43% para o ano de 2020. Porém conforme abordado por Santos (2020) por conta da pandemia do novo coronavírus, não se sabe afirmar com dados concretos e reais, qual será o impacto dessa crise no comercio eletrônico, entretanto, este é um cenário de perspectivas mistas. Isto está acontecendo porque, da mesma forma que economia está em recessão e os consumidores estão tendo que ficar mais em casa, se focando apenas na alimentação e saúde, foi decretado em todo o mundo medidas de distanciamento social. Com

estas medidas de isolamentos algumas lojas físicas tiveram que ser fechadas, com isso os estabelecimentos tiveram prejuízos. Contudo teve um aumento bastante significativo nas vendas online devido várias restrições por conta das medidas de isolamento e contenção do vírus.

3. PROCEDIMENTOS METODOLÓGICOS

Quanto à natureza dessa pesquisa ela se dá como trabalho original:

> "Busca apresentar conhecimento novo a partir de observações e teorias construídas para explicá-las. Assume-se a nova informação como relevante quando ela tem implicação na forma como se entendem os processos e sistemas ou quando tem implicação prática na sua realização." (WAZLAWICK, 2014, p.22).

Referente aos objetivos, esta pesquisa trata-se de uma pesquisa exploratória e descritiva, conforme descrito por Wazlawick (2014), tem o objetivo de estudar os dados mais consistentes sobre uma determinada realidade e compreender anomalias que não sejam conhecidas, e que podem então ser a base para uma pesquisa mais elaborada.

No que se refere aos procedimentos técnicos é uma pesquisa documental pois visa a análise de dados e documentos que ainda não foram sistematizados e publicados (Wazlawick, 2014). Justifica-se a escolha do campo de ciência de dados pois é uma área com enorme potencial e ainda não foi amplamente explorada.

No decorrer deste trabalho foram realizadas pesquisas exploratórias que serviram para definir a metodologia utilizada neste trabalho. O capítulo seguinte será dividido conforme as etapas da metodologia abordado por Rautenberg e Carmo (2019), que foi referenciada no tópico de ciência de dados.

4. APLICAÇÃO DA CIÊNCIA DE DADOS AO SETOR VAREJISTA

Neste capítulo é abordada a aplicação da ciência de dados no setor varejista, contudo é utilizado a mesma metodologia abordada por Rautenberg e Carmo (2019), será demonstrado o mesmo ciclo de dados conforme visto na figura 2.

4.1 Obtenção dos Dados.

Nesta etapa da obtenção dos dados inclui a coleta, a exploração e a verificação da qualidade dos dados obtidos. As ferramentas empregadas nesta etapa foram utilizadas em um sistema Windows 7 de 64 Bits. As ferramentas empregadas foram:

Google Chrome: Navegador de Internet utilizado para acessar a ferramentas de modelagem dos dados:

Google Colab: Ferramenta de criação de códigos em *Python* no navegador;

Pandas: Biblioteca do *Python* usada para manipulação dos dados;

Python: Linguagem de programação;

Matplotlib: Biblioteca do Python utilizada para plotar gráficos;

Kaggle: Site que disponibiliza o conjunto de dados da empresa Rossmann;

Facebook *Prophet*: permite a previsão de séries temporais baseado em regressão aditiva;

4.2 Coleta dos dados

Na etapa de coleta de dados foram utilizados dados históricos de vendas referente a empresa Rossmann. Os dados foram obtidos e disponibilizados a partir da plataforma Kaggle, existem informações de vendas entre os anos de 2013 a 2015, tanto informações de vendas e histórico de clientes nas lojas nesse período, quanto informações das lojas como por exemplo a distância do concorrente mais próximo.

O estudo de caso contém dois conjuntos de dados em arquivos no formato csv:

train.csv: este arquivo possui 1017209 linhas e 9 colunas. Este conjunto de dados tem informações relacionadas as vendas, essas informações serão detalhadas posteriormente.

store.csv: este arquivo possui 1115 linhas e 10 colunas. As linhas correspondem às lojas, ou seja, existem um total de 1115 lojas nesta base de dados, o restante das colunas com informações relacionadas a loja, um desses exemplos o tamanho e tipo da loja, entre outras

informações que será abordada posteriormente no tópico de descrição de dados.

Nesta pesquisa serão utilizados ambos conjuntos de dados. A partir destes é possível identificar insights que possam auxiliar na tomada de decisão.

4.2.1 Descrição dos dados

Nesta seção serão apresentados os atributos obtidos a partir das bases de dados mencionadas anteriormente. Será feito um estudo estatístico sobre alguns desses registros. Estas são as descrições de cada atributo relacionado a tabela 1.

Tabela 1 Relação de atributos retirados da base de dados *store*.

Atributos	**Descrição**
Store	identificador único da loja
*Store*Type	tipo da loja (a,b,c,d)
Assortment	Determina o tamanho da loja a=basic, b=extra, c=extended
CompetitionDistance	distância para a loja concorrente mais perto em metros
CompetitionOpenSinceMonth., *CompetitionOpenSinceYear*	Mês e Ano que foi feito a abertura do concorrente mais próximo.
Promo2	Promoção Adicional da loja.
Promo2SinceWeek, Promo2SinceYear	Semana e Ano que ocorreu a promoção adicional.

Fonte: Autoria própria (2021).

É de extrema importância ter o entendimento de cada a atributo para que assim possa ser feita a análise de forma concisa, sem ignorar cada informação que aquele atributo possa representar. É necessária essa descrição para que assim seja possível fazer as análises futuras de forma detalhada, ou seja um bom entendimento da descrição dos dados iniciais irá influenciar diretamente nos procedimentos futuros.

Abaixo será feito a descrição de dados referente ao conjunto de dados *train*, a partir dessa descrição que é possível compreender o que cada coluna representa, para que assim seja possível prosseguir com as análises.

Tabela 2 Relação de atributos retirados da base de dados *train*.

Atributos	**Descrição**
Id	Identificador da transação
Store	identificador único da loja.
DayOfWeek	Identificador referente ao dia da semana
Sales	vendas/dia (objetivo).
Customers	número de clientes no dia.
Open	Operador booleano que indica se a loja estava aberta ou fechada.
Promo	se existe uma promoção no dia
StateHoliday	feriado
SchoolHoliday	feriado escolar

Fonte: Autoria própria (2021).

A figura 3 representa os 5 primeiros registros de cada conjunto de dados para melhor visualização dos dados e o que contém cada atributo, e como possivelmente estão distribuídos.

Figura 3: Os 5 primeiros registros da base de dados *train*.

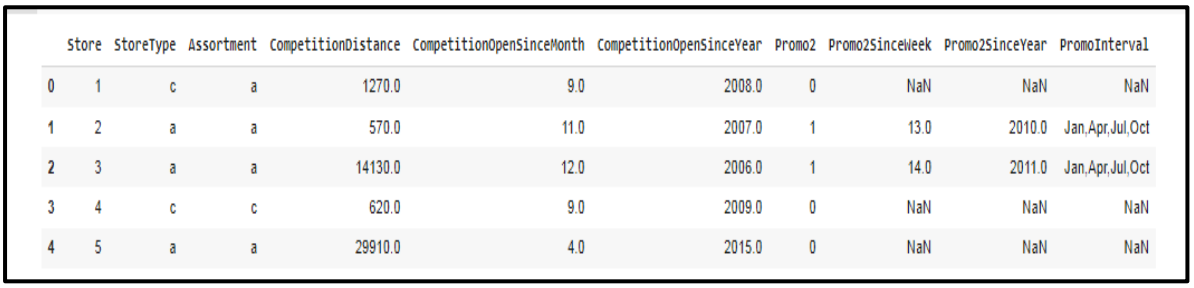
	Store	DayOfWeek	Date	Sales	Customers	Promo	StateHoliday	SchoolHoliday
0	1	5	2015-07-31	5263	555	1	0	1
1	2	5	2015-07-31	6064	625	1	0	1
2	3	5	2015-07-31	8314	821	1	0	1
3	4	5	2015-07-31	13995	1498	1	0	1
4	5	5	2015-07-31	4822	559	1	0	1

Fonte: Autoria própria (2021).

Conforme a figura 3, pode ser visualizado inicialmente a distribuição desses atributos de uma forma mais objetiva.

Posteriormente será também abordado algumas análises estatísticas referente ao conjunto de dados *store*, para que possa ser identificado se existe alguns dados fora do padrão.

Figura 4: Os 5 primeiros registros da base de dados *store*.



	Store	StoreType	Assortment	CompetitionDistance	CompetitionOpenSinceMonth	CompetitionOpenSinceYear	Promo2	Promo2SinceWeek	Promo2SinceYear	PromoInterval
0	1	c	a	1270.0	9.0	2008.0	0	NaN	NaN	NaN
1	2	a	a	570.0	11.0	2007.0	1	13.0	2010.0	Jan,Apr,Jul,Oct
2	3	a	a	14130.0	12.0	2006.0	1	14.0	2011.0	Jan,Apr,Jul,Oct
3	4	c	c	620.0	9.0	2009.0	0	NaN	NaN	NaN
4	5	a	a	29910.0	4.0	2015.0	0	NaN	NaN	NaN

Fonte: Autoria própria (2021).

Após a visualização inicial dos dados será feito as primeiras análises, como o número de valores disponíveis nos atributos, após isso será feito uma análise estatística para descobrimos a distribuição desses valores.

A análise será iniciada pela base de dados *train*, essa base de dados possui as informações relacionados as vendas feitas pelas lojas, a seguir seus atributos e os valores que compõe a mesma:

DayOfWeek: está coluna conforme descrito na tabela 2 descreve o dia da semana que ocorreu a venda, este atributo está na base de dados com um intervalo que vai de 1 a 7, com o número 1 representando o domingo terminando no número 7 que significa o sábado e assim respectivamente;

Open: responsável por representar se a loja está em funcionamento ou inativa, os valores disponíveis na base para este atributo são apenas 0 e 1, 0 para quando a loja estiver fechada é 1 para quando a loja estiver aberta;

Promo: informa se existe promoção na loja no dia, esse atributo é representado somente por 0 ou 1, quando é 0 informa que não houve promoção e 1 informa que houve promoção no dia.

StateHoliday: representa se existe ou não um feriado, ele está distribuído em variáveis “a” que se refere ao feriado público, a variável “b” que se refere a Páscoa, “c” referente ao Natal e “0” quando não houver nenhum feriado;

SchoolHoliday: atributo que informa se existe ou não um feriado escolar, ele está distribuído nos valores “0” quando não houver feriado escolar e “1” quando houver feriado escolar.

Sales: atributo que representa as vendas no dia, o importante a ser informado neste atributo que esse valor está descrito em Euros;

Customers: retrata o número de clientes que visitou a loja naquele dia;

Em seguida na figura 5 será apresentado algumas análises estáticas sobre cada coluna

da base de dados *train*, para termos uma breve noção de cada atributo, como por exemplo os valores máximos e mínimos de cada atributos e outra informação importante como a média.

Figura 5: Análise estatísticas das colunas referente a base de dados *train*.

	Store	DayOfWeek	Sales	Customers	Open	Promo	SchoolHoliday
count	1.017209e+06	1.017209e+06	1.017209e+06	1.017209e+06	1.017209e+06	1.017209e+06	1.017209e+06
mean	5.584297e+02	3.998341e+00	5.773819e+03	6.331459e+02	8.301067e-01	3.815145e-01	1.786467e-01
std	3.219087e+02	1.997391e+00	3.849926e+03	4.644117e+02	3.755392e-01	4.857586e-01	3.830564e-01
min	1.000000e+00	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
25%	2.800000e+02	2.000000e+00	3.727000e+03	4.050000e+02	1.000000e+00	0.000000e+00	0.000000e+00
50%	5.580000e+02	4.000000e+00	5.744000e+03	6.090000e+02	1.000000e+00	0.000000e+00	0.000000e+00
75%	8.380000e+02	6.000000e+00	7.856000e+03	8.370000e+02	1.000000e+00	1.000000e+00	0.000000e+00
max	1.115000e+03	7.000000e+00	4.155100e+04	7.388000e+03	1.000000e+00	1.000000e+00	1.000000e+00

Fonte: Autoria própria (2021).

É importante fazer essa breve análise para identificar se existe algum valor fora do padrão. Por isso verificar os valores máximos e os mínimos a fim de identificar se os valores estão coesos. Pode-se então retirar algumas informações importante sobre essas estáticas como por exemplo, média da variável *Sales* que representa as vendas no dia, ou seja podemos concluir que a média de atributo representa € 5.773,00 por dia, outra informação importante é informar que a maior venda feita em um dia foi de € 41.551,00.

Outra informação que é possível retirar é que o valor mínimo de vendas é € 0,00 isso condiz com os dados pois as lojas podem estar fechadas neste dia, ou seja não houve nenhuma venda. Outra variável importante a ser verificada é a Customers, essa variável representa os números de clientes que visitou a loja no dia, ou seja, como pode-se ver essa variável está fortemente relacionada com a variável *Sales*. Quando não houve nenhuma venda também não houve nenhum cliente em loja e porventura a loja estava fechada.

Continuamente pode-se abordar a base de dados *store*, que possue os registros e informações referente às lojas, a seguir seus atributos e os valores que compõe a mesma:

***Store*:** este atributo representa as 1115 lojas, ou seja, é o identificador das lojas. Possue valores que se iniciam em 1 com o valor máximo de até 1115.

CompetitionDistance: representa a distância em metros dos correntes diretos mais próximas a loja.

CompetitionOpenSinceYear/ CompetitionOpenSinceMonth: está variável informar a data que a loja concorrente foi aberta, é representada pelo o ano é pelo mês.

Promo2: informa se existe promoção adicional em todas as lojas, possui apenas os valores 0 não está participando, e o valor 1 que significa que está participando.

Promo2Since: está coluna está relacionada a data quando a loja começou a participar da Promo2.

PromoInterval: este atributo tem o intervalo consecutivos que a Promo2 é iniciada(meses). Exemplo: "*Feb, May, Aug, Nov*" indica que cada "*round*" da promoção começa em fevereiro, maio, agosto, novembro.

A figura 6 representa uma breve análise sobre o conjunto de dados *store* e seus atributos:

Figura 6: Análise estatísticos das colunas referente a base de dados *store*.

	Store	CompetitionDistance	CompetitionOpenSinceMonth	CompetitionOpenSinceYear	Promo2	Promo2SinceWeek	Promo2SinceYear
count	1115.00000	1112.000000	761.000000	761.000000	1115.000000	571.000000	571.000000
mean	558.00000	5404.901079	7.224704	2008.668857	0.512108	23.595447	2011.763573
std	322.01708	7663.174720	3.212348	6.195983	0.500078	14.141984	1.674935
min	1.00000	20.000000	1.000000	1900.000000	0.000000	1.000000	2009.000000
25%	279.50000	717.500000	4.000000	2006.000000	0.000000	13.000000	2011.000000
50%	558.00000	2325.000000	8.000000	2010.000000	1.000000	22.000000	2012.000000
75%	836.50000	6882.500000	10.000000	2013.000000	1.000000	37.000000	2013.000000
max	1115.00000	75860.000000	12.000000	2015.000000	1.000000	50.000000	2015.000000

Fonte: Autoria própria (2021).

De acordo com a figura 6 essas estáticas sobre conjunto de dados *store*, temos alguns *insights* igualmente importantes que podemos adotar, como por exemplo a distância de uma loja para o concorrente mais próximo podemos ver na variável CompetitionDistance, que a média de distância de um concorrente mais próximo é de 5.4 km, e a distância mínima é de apenas 20 metros. Por fim terminamos as análises iniciais dos valores desses atributos.

4.3 Ingestão dos dados

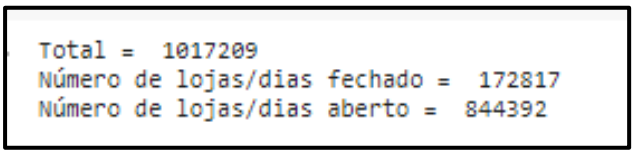
Como esta etapa está relacionado com a conversão e carga dos dados de origem, obtido a partir de fontes diversas e formatos diversificados, como o estudo feito anteriormente dos atributos encontrados na base, todas as informações são da mesma fonte e do mesmo tipo de arquivo, não terá necessidade da conversão de dados. O que pode ser feito posteriormente será a junção de ambas as tabelas para verificar a correlação desses atributos, pois assim é possível identificar quais serão utilizados como parâmetros para o modelo.

4.4 Exploração dos dados

Nesta seção será abordado a análise exploratória a fim de encontrar correlações entre atributos e obter um entendimento mais claro sobre cada papel que cada dado pode ser atribuído.

Uma consideração importante que deve ser tomada, seria a verificação da coluna *Open* na base de dados *train*, conforme a figura 7, é possível verificar a quantidade de registro referente a loja aberta e fechada. Pois como nosso objetivo é verificar os fatores que influenciam as vendas, não faz sentido usar dados quando a loja não estiver aberta, pois logo não houve vendas nesses dados.

Figura 7: Quantidade de registros na coluna *Open* referente a base de dados *train.*



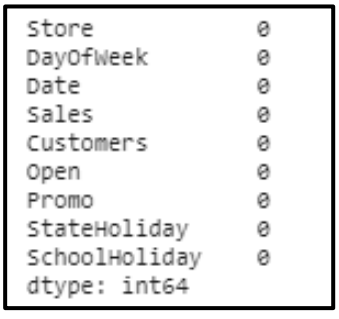
Fonte: Autoria própria (2021).

Outra informação importante que exige cuidado na análise das bases dados são os valores nulos, ou seja, não existe nenhuma informação na linha quando ela possue valor nulo.

Por tanto umas das formas de preencher os valores nulos, o mais comum a ser feito é substituir esses valores pelas médias encontradas, porém isso não se aplica a todas as variáveis como por exemplo as variáveis do tipo data, não faz sentido utilizar esse conceito por tanto caso isso ocorra, deve ser substituído esses valores por zeros.

Conforme descrito na figura 8, o conjunto de dados *train* não possui atributos nulos.

Figura 8: Relação de atributos nulos para cada coluna da base *train*.



Fonte: Autoria própria (2021).

Em seguida ser feito a mesma verificação para a base de dados *store*, para que seja possível prosseguir com as análises, pois caso exista valores nulos, irá afetar possíveis resultados obtidos, por isso é de extrema importância verificar essa informação, conforme a

figura 9 é possível validar essa informação.

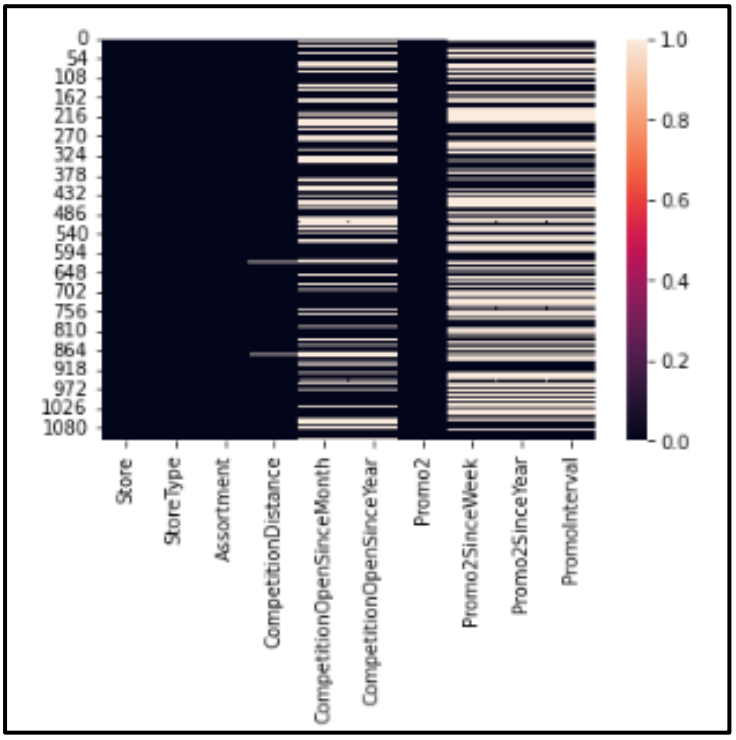
Figura 9: Relação de atributos nulos para cada coluna da base *store.*

```
Store                        0
StoreType                    0
Assortment                   0
CompetitionDistance          3
CompetitionOpenSinceMonth  354
CompetitionOpenSinceYear   354
Promo2                       0
Promo2SinceWeek            544
Promo2SinceYear            544
PromoInterval              544
dtype: int64
```

Fonte: Autoria própria (2021).

Como essa base de dados possui elementos nulos, será realizada uma verificação mais detalhada, em seguida na figura 10, será gerado um gráfico de calor, também conhecido por *heatmap* para visualização desses dados nulos e a sua distribuição.

Figura 10: Gráfico de calor da distribuição dos valores nulos no conjunto *store.*

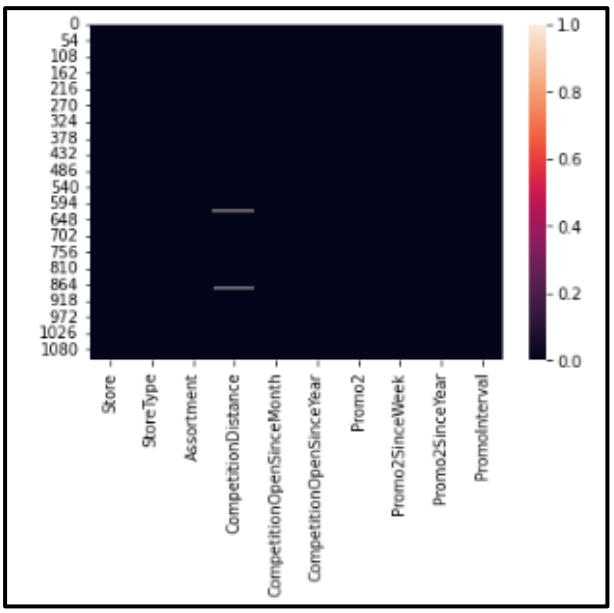


Fonte: Autoria própria (2021).

De acordo com a figura 10 alguns dos atributos há muitos valores nulos, a distribuição desses valores está relacionada principalmente a variável Promo2, pois provavelmente quando está variável possuía o valor 0, as colunas que são relacionadas com ela não foram preenchidas.

As colunas referentes a datas : *CompetitionOpenSinceYear, CompetitionOpenSinceMonth, Promo2SinceWeek, Promo2SinceYear, PromoInterval*; como descrito na seção anterior normalmente quando os valores são nulos as variável são preenchidas com a média da distribuição, porém como essas variáveis são datas, não faz sentido fazer essa distribuição, portanto iremos preencher essas colunas onde o valor for nulo para o número 0. Após o preenchimento é possível verificar na figura 11 o mapa de calor referente as colunas com valores nulos da base de dados *store*.

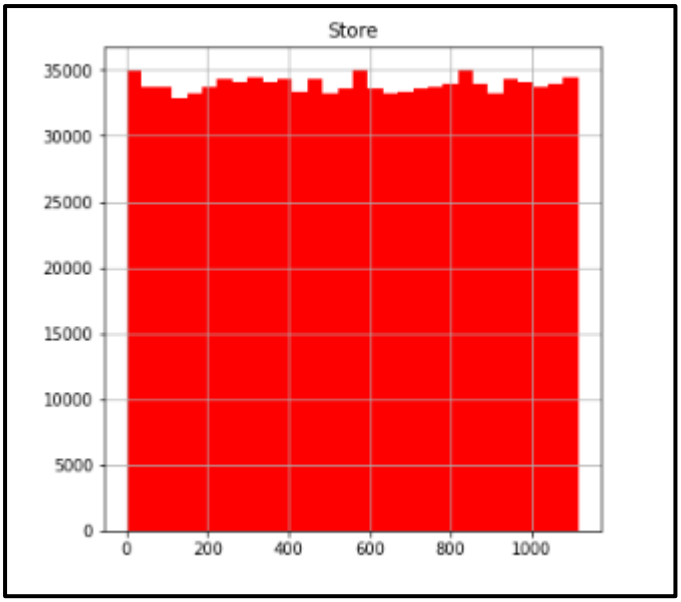
Figura 11: Gráfico de calor da distribuição dos valores nulos no conjunto *store* após o preenchimento das colunas referente a datas.



Fonte: Autoria própria (2021).

Em seguida, será criado um histograma na figura 12, o que é uma espécie de gráfico de barras que demonstra uma distribuição de frequências. Será feito para cada coluna começando pela base de dados *train*. Essa visualização auxilia com a representação gráfica dos conjuntos de dados de forma mais amigável, tornando mais fácil a visualização de onde a maioria dos valores se concentra.

Figura 12: Histograma da coluna *Store*, referente a base de dados *train.*

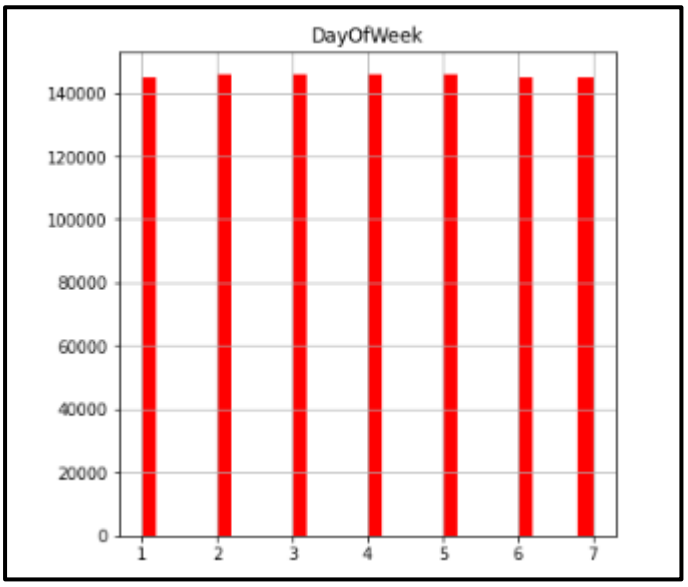


Fonte: Autoria própria (2021).

Como esta coluna é referente a cada identificador único das lojas, temos essa distribuição dos valores de forma uniforme representando cada loja. Não é um atributo levado em consideração para essa análise.

Em seguida será também criado na figura 13, o histograma referente a coluna *DayOfWeek*.

Figura 13: Histograma da coluna *DayOfWeek*, referente a base de dados *train*.



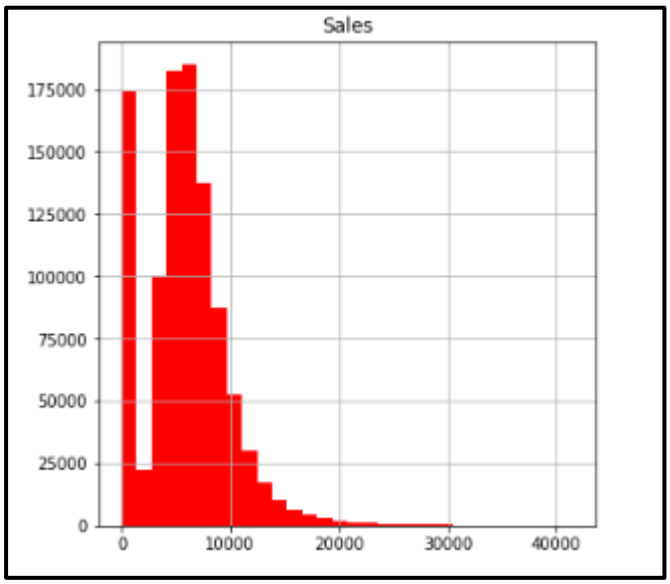
Fonte: Autoria própria (2021)

Os valores referentes a está coluna *DayOfWeek*, está totalmente coesa pois está

distribuída igualmente nos dias da semana que ela representa.

Abaixo será apresentado na figura 14 o histograma da coluna *Sales*.

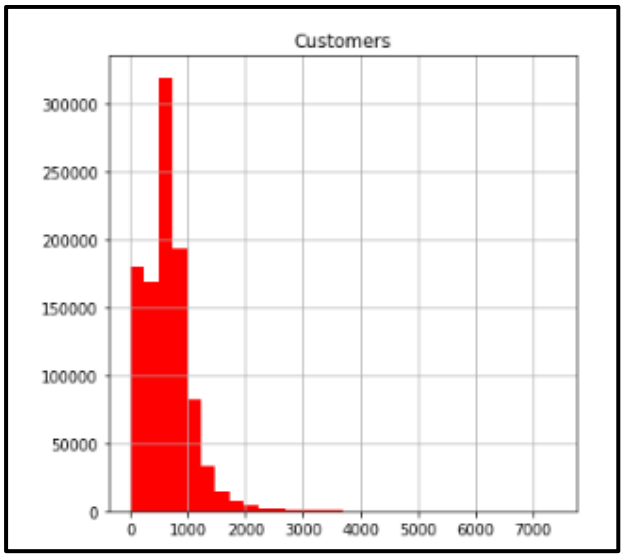
Figura 14: Histograma da coluna Sales, referente a base de dados *train*.



Fonte: Autoria própria (2021).

A distribuição dessa coluna, que representa as vendas está com picos entre o valor 0 e entre 5 a 6 mil euros, ou seja, a distribuição faz sentido com a média que calculamos anteriormente. Na figura 15 também possível identificar a distribuição dos registros na coluna *Customers*.

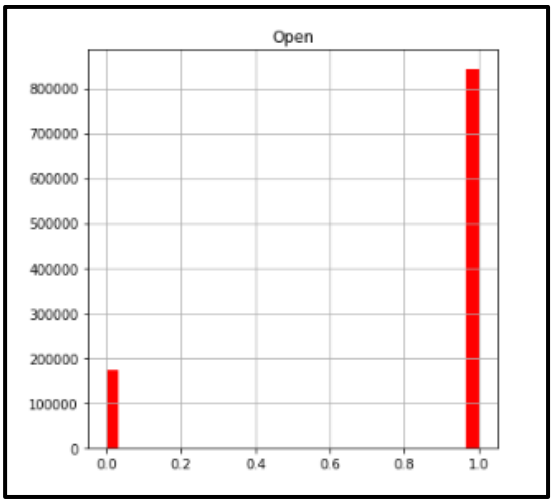
Figura 15: Histograma da coluna *Customers*, referente a base de dados *train*.



Fonte: Autoria própria (2021).

Pode ser verificado também que o a distribuição está condizente com a média calculada anteriormente, há uma média de clientes por dia entorno de 600. Em seguida pode-se visualizar na figura 16 a distribuição da coluna *Open*.

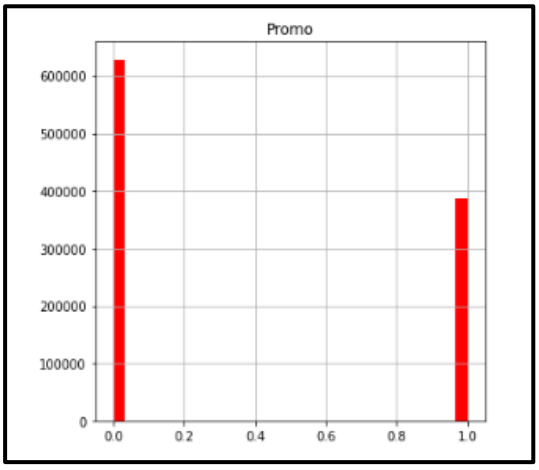
Figura 16: Histograma da coluna *Open*, referente a base de dados *train.*



Fonte: Autoria própria (2021).

A partir desse histograma é possível identificar que essa coluna está distribuída nos valores 0 e 1, ou seja, a loja possui mais registros com ela aberta do que quando está inativa. Essa informação é interessante pois é possível utilizar somente os dados com a loja aberta já que assim podemos diminuir o tamanho dos registros, que auxilia no custo computacional do modelo. E na figura 17 pode ser visto a distribuição da coluna *Promo.*

Figura 17: Histograma da coluna Promo, referente a base de dados *train*

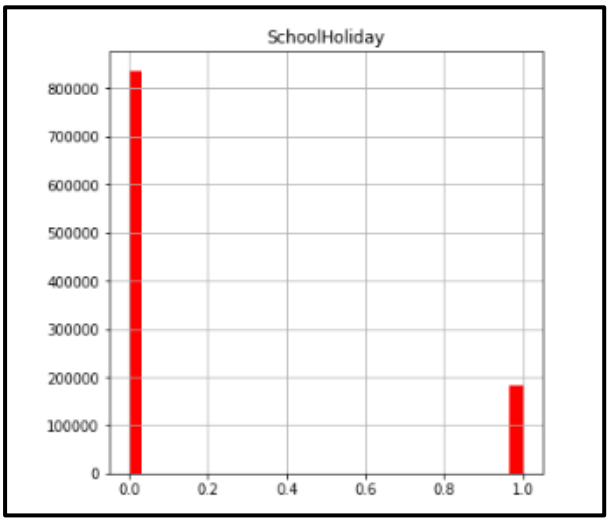


Fonte: Autoria própria (2021).

Diferentemente da coluna Open, a coluna Promo possui mais registros quando não houve promoção do que o contrário, isso é importante devido ao próximo passo que é verificar se a coluna tem uma correlação forte relacionada a colunas *Sales*, que indica as vendas.

Na figura 18 é criado o histograma da coluna *SchoolHoliday*, para que assim seja possível visualizar sua distribuição.

Figura 18: Histograma da coluna SchoolHoliday, referente a base de dados *train.*

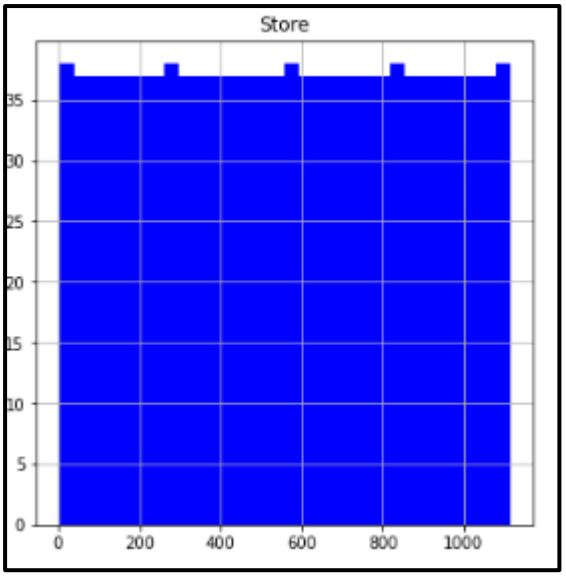


Fonte: Autoria própria (2021).

Conforme descrito na figura 18, está sendo verificado que a variável *SchoolHoliday* que também informar se houve ou não feriado escolar, está mais distribuída para o valor 0, também está correto pois existe menos feriados do que dias letivos.

Em seguida a partir da figura 19, será realizada a mesma análise nas colunas referente ao conjunto de dado *store*, que possui informação referente a loja.

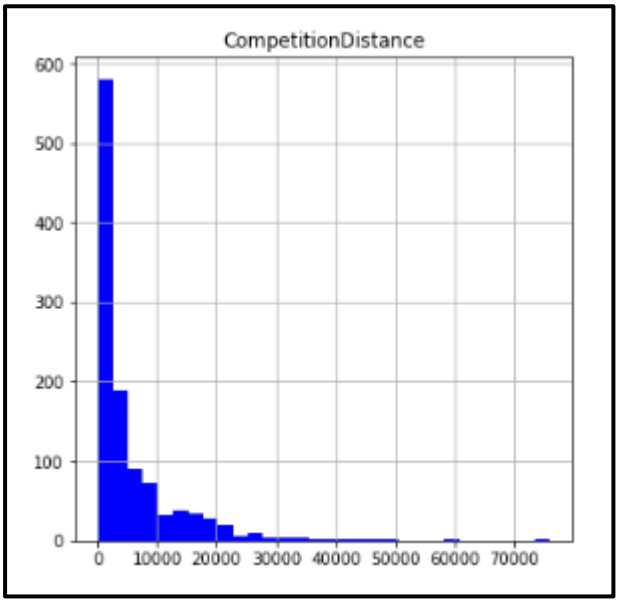
Figura 19: Histograma da coluna *Store*, referente a base de dados *store*.



Fonte: Autoria própria (2021).

Diferente do histograma referente ao conjunto *train*, por mais que a coluna *Store* seja a mesma, a distribuição é levemente diferente pois existe quantidades variadas de registros, de certa forma a distribuição continua uniforme, pois representa o identificador único de cada loja. Na figura 20 criado o histograma referente a coluna *CompetitionDistance*.

Figura 20: Histograma da coluna *CompetitionDistance*, referente a base de dados *store*.

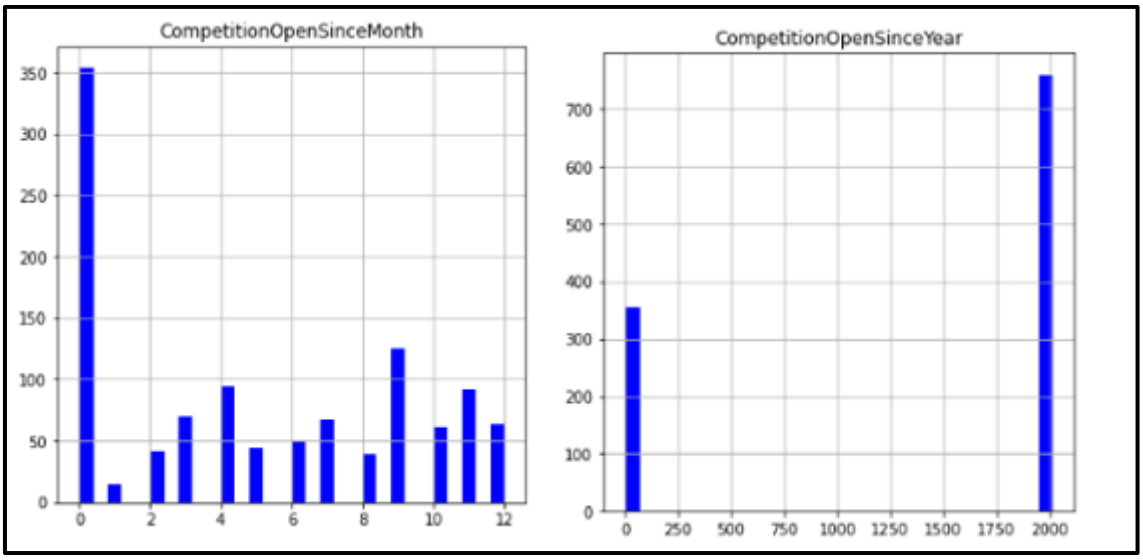


Fonte: Autoria própria (2021).

Referente a coluna CompetitionDistance está relacionada a distância do concorrente

mais próximo, ela segue o padrão conforme análise estática abordada anteriormente e tem uma distribuição maior nos períodos entre 0 e 5 km. Na figura 21 é possível visualizar a distribuição do atributo *CompetitionOpen.*

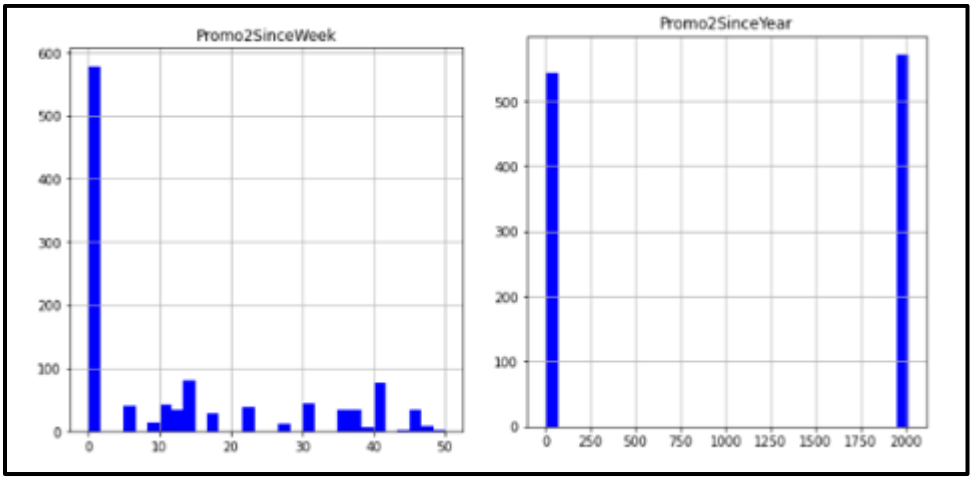
Figura 21: Histograma da coluna *CompetitionOpen*, referente a base de dados *store*.



Fonte: Autoria própria (2021).

Estes gráficos se referem ao mês e o ano onde as lojas concorrente foram abertas, um ponto que deve ser destacado é a quantidade alta dos registros com valores 0, isso ocorreu devido ao tratamento de dados que foi feto anteriormente alterando os valores de dados nulos para o valor 0. Em seguida na figura 22 será possível visualizar a distribuição da coluna *Promo2Since.*

Figura 22: Histograma da coluna *Promo2Since*, referente a base de dados *store*.

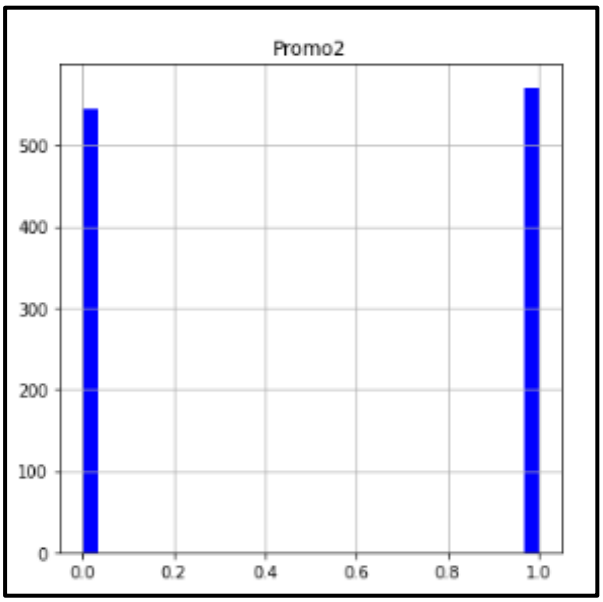


Fonte: Autoria própria (2021).

Estes histogramas são referentes a semana e o ano que ocorreu a promoção adicional,

pois está relacionada com a coluna *Promo2* que representa uma promoção adicional para todas as lojas. Na figura 23 é possível visualizar a distribuição da coluna *Promo2.*

Figura 23: Histograma da coluna Promo2, referente a base de dados *store*.



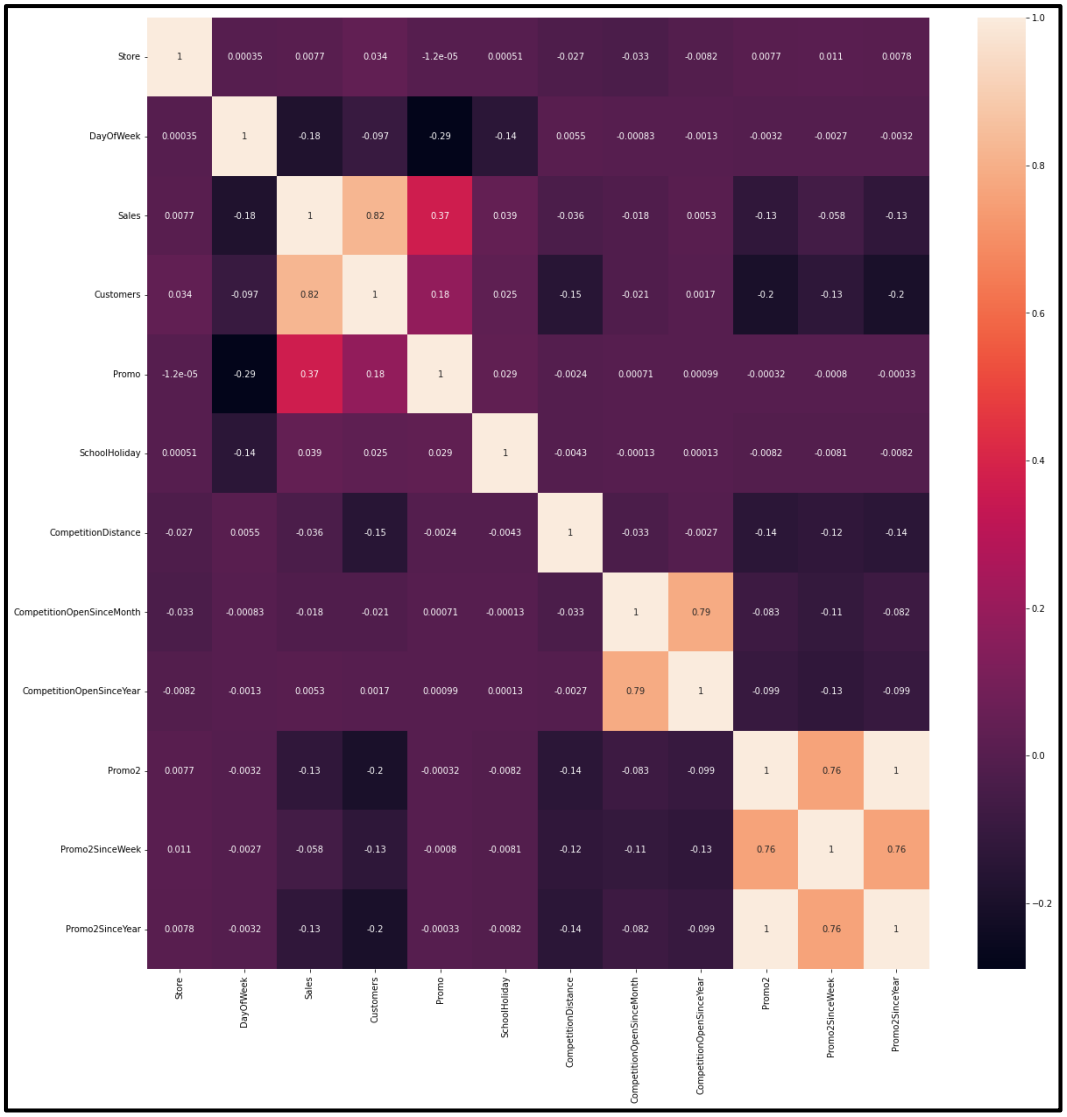
Fonte: Autoria própria (2021).

O histograma referente a figura 23 representa a coluna *Promo2* que demonstra se ocorreu ou não a promoção adicional na loja, é visível que a distribuição está bem balanceada. Esta verificação dos histogramas de quase todas as colunas é importante para o entendimento da distribuição desses valores em ambos conjuntos para assim auxiliar no entendimento completo da base, e que informações é possível retirar desses dados.

Após essa análise será realizado a junção dos conjuntos de dados *store* e *train* em um conjunto de dados único, este novo conjunto irá ser renomeado de *store_train_all* para assim termos todas as informações em um único conjunto de dados, dessa forma pode ser compreendido quais são as variáveis que tem valores mais próximo a 1, referente as suas correlações com a coluna *Sales*, e consequentemente com os clientes que visitam as lojas. Uma matriz de correlação é simplesmente uma tabela que exibe os coeficientes de correlação para diferentes variáveis. A matriz mostra entre todos os pares de valores possíveis em uma tabela. É uma ferramenta poderosa para resumir um grande conjunto de dados para identificar e visualizar padrões nos dados fornecidos.

A seguir será apresentado na figura 24, a matriz de correlação entre as colunas que por sua vez é referente ao conjunto de dados *store_train_all*, este conjunto conforme descrito anteriormente se refere aos conjuntos de dados *train* e *store* mesclado.

Figura 24: Matriz de correlação entre as colunas do conjunto de dados *store_train_all*.



Fonte: Autoria própria (2021).

Na figura 24 foi utilizado o mapa de calor, pois é uma forma mais rápida de visualizar as correlações, quanto mais claro ou mais escuro, ou seja, os dois extremos estão relacionados ao nível mais alto de correlação.

Ressalta-se que está matriz possui valores vão de -1 até 1, e quanto mais próximo de 1 e -1, mais forte é essa correlação. Neste momento iremos focar na variável *Sales*, pois ela representa as vendas e esse será nosso objetivo.

Na figura 25 será possível visualizar essa informação pois é observar apenas as correlações referente a coluna *Sales*. Ressaltando que a coluna Sales é nosso objetivo, pois ela representa o número de vendas.

Figura 25: Relação de correlação da variável Sales para as demais variáveis.

```
DayOfWeek                   -0.178736
Promo2SinceYear             -0.127621
Promo2                      -0.127596
Promo2SinceWeek             -0.058476
CompetitionDistance         -0.036343
CompetitionOpenSinceMonth   -0.018370
CompetitionOpenSinceYear     0.005266
Store                        0.007710
SchoolHoliday                0.038617
Promo                        0.368145
Customers                    0.823597
Sales                        1.000000
Name: Sales, dtype: float64
```

Fonte: Autoria própria (2021).

Os valores que mais chamam a atenção nesta figura são os da correlação entre as Colunas *Sales* e *Customers*, está sendo de 0.82. Ou seja, é uma correlação forte, quanto mais clientes na loja mais vendas existem.

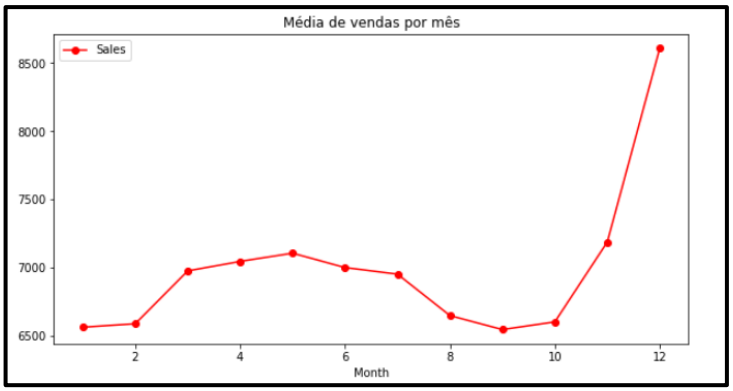
Outro ponto também importante que deve ser levado em consideração é a correlação entre a variável Sales – *Promo*. Pois as variáveis têm 0.37. Portanto é uma correlação moderada, o que implica que a promoção tem sim um impacto direto nas vendas.

O próximo ponto que deve ser notado é a correlação entre a variável Sales – *Promo2*, entre essas duas variáveis é de apenas -0.13. Ou seja, é uma correlação baixa, então podemos concluir que essa promoção adicional não tem um impacto direto nas vendas.

Seguindo com a análise exploratória dos dados, devemos entender claramente cada dado do que se trata e que informações podemos retirar deles.

É interessante fazer a fragmentação da coluna *date* em novos campos como o dia, mês e ano. Essa separação é importante para permitir fazer uma análise das vendas, a partir da figura 26.

Figura 26: Médias de vendas por mês.

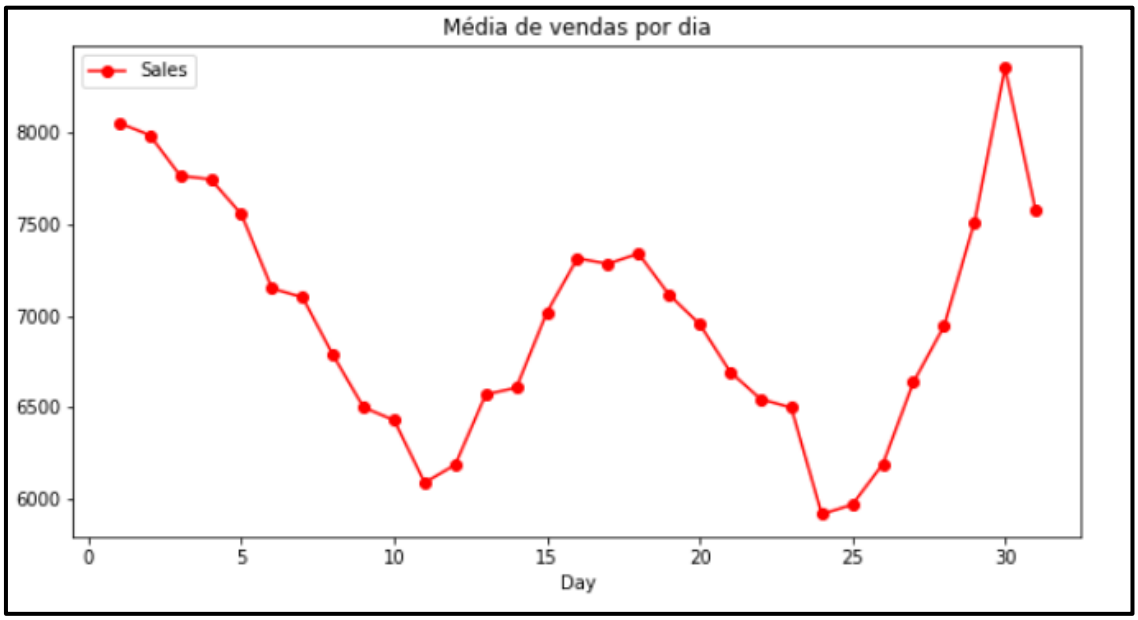


Fonte: Autoria própria (2021).

Em alusão na figura 26, a distribuição de venda tem um pico sempre nos últimos meses

do ano, provavelmente atrelado aos feriados de fim de ano. Porém isso será discutido posteriormente. Na figura 27 é possível visualizar as vendas por dia em relação ao mês.

Figura 27: Médias de vendas por dia.



Fonte: Autoria própria (2021).

A Figura 27 representa as vendas por dia, existe dois picos centrais o início do mês e logo ao final, pode ser visto um terceiro pico no meio do mês, e algumas baixas nas vendas entre o dia 10 a 13 e entre o dia 20 a 25, ou seja esse gráfico pode ser de grande ajuda para os tomadores de decisões para entender quais dias no mês necessitam de mais produtos nas lojas.

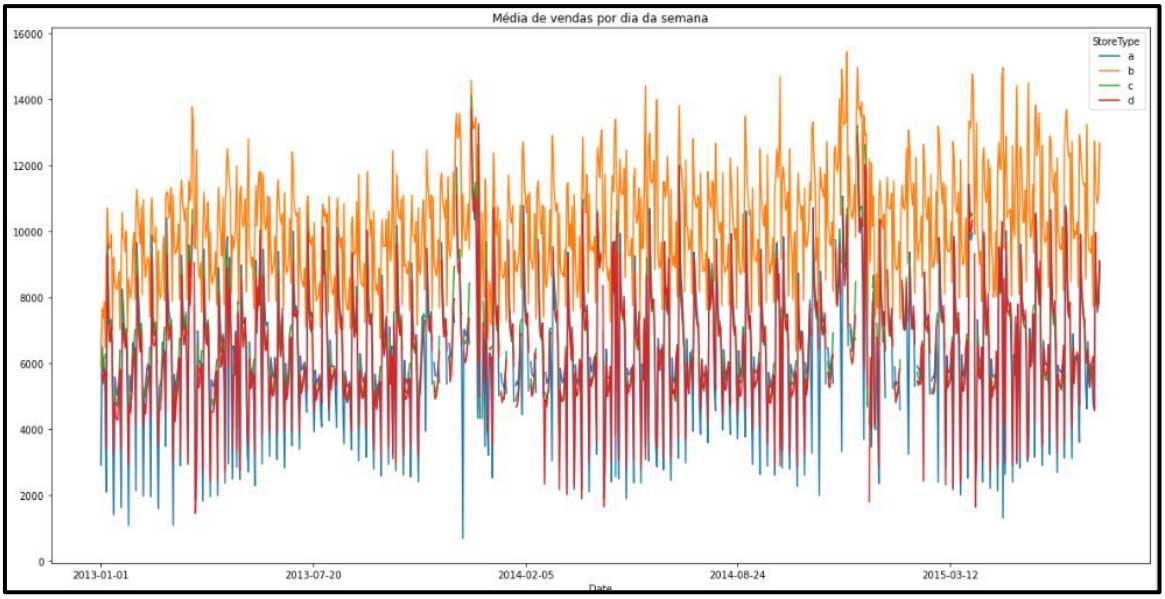
A partir da 28 é possível visualizar as vendas por dia da semana.

Figura 28: Médias de vendas por dia da semana.

Fonte: Autoria própria (2021).

É interessante observar o gráfico de média de vendas por dias a partir da figura 28, pois podemos ver quais dias tem picos ou declínios nas vendas, os dias que se destacam mais são domingo e sábado, pois são referentes aos fins de semana. Na figura 29 é será criado o gráfico de vendas, diferenciando os tipos de lojas.

Figura 29: Médias de cliente em toda base de dados.



Fonte: Autoria própria (2021).

Neste gráfico é possível ver a distribuição de vendas referente aos tipos de loja a, b, c e d. Pode se notar que as do tipo b se sobressaem sobre as demais, ou seja as loja do tipo b tem mais vendas que o restante, e o tipo que tem menos vendas são do tipo a.

Apenas com esta investigação encontramos pontos bastante interessante que ajudariam os tomadores de decisões, para desenvolver *insights* que auxiliariam no desenvolvimento da empresa como um todo.

4.5 Definição de parâmetros

Nesta seção será abordada os parâmetros necessários para que possa utilizada o algoritmo de aprendizado de máquina. Nesta pesquisa o objetivo principal é que seja possível fazer a previsão de vendas futuras a partir dos dados antigos disponibilizados na base de dados, será feita a partir de cada loja, será utilizado a loja 10 como exemplo, porém o modelo poderia ser atribuído para todas as 1115 lojas.

Foi adota a ferramenta *Facebook Prophet*, é uma ferramenta que auxilia a utilização

desses algoritmos, utilizando series temporárias para desenvolver essas previsões. Esta ferramenta permite utilizar tanto a função de regressão quanto a função previsão, será utilizado então a função de previsão. Normalmente, a previsão utiliza o valor das vendas anteriores para que assim seja possível prever os valores de vendas futuras, enquanto a regressão usa os atributos que compõem uma transação de venda para assim prever os novos valores das vendas, a partir da análise exploratórios os atributos a serem utilizados serão os dados temporárias, as datas das vendas, e o *store*, que se refere ao identificador de cada loja, e também um atributo chave será os feriados.

Diferente de outros algoritmos convencionais a ferramenta *Facebook Prophet*, consegue trabalhar muito bem com os feriados pois esta ferramenta não define os feriados como pontos fora da curva, normalmente chamados em ciência de dados de *outlier*, observações que se diferem muito das demais observações, normalmente esses *outlier* é descartado das previsões, por ser valores fora da realidade. Porém neste caso os feriados não são pontos fora da curva e eles devem ser mantidos para previsão.

4.6 Implementação do modelo

Nesta seção será demostrado como foi feito a implementação do modelo.

Umas das vantagens de utilizar a ferramenta *Facebook Prophet*, é de não ser necessário criar uma base de dados complexa para fazer a criação do modelo. Conforme descrito na seção anterior será construído o modelo para criar previsões de vendas de lojas especificas, ou seja, será feito a previsão de uma determinada loja em vez de realizar a previsões de vendas de todas as lojas , que seria necessário muito custo computacional, e também isso nos ajuda a focar em uma determinada loja.

O primeiro passo conforme descrito anteriormente é definir o identificador que neste caso vai ser a coluna *Store*, que irá representar a loja que deve ser feito a previsão das vendas, outro parâmetro será o período ou seja, quantos dias irá ter a previsão, outro parâmetro que deve ser utilizado neste caso são os feriados ou em inglês *holidays*, para utilizar esse parâmetro em especifico, foi necessário retirar as colunas referentes ao feriado e criar um novo conjunto de dados, apenas para os feriados, para ser utilizados. Após ser definido o parâmetro a ser utilizado da base de dados, estes devem ser também ajustados para a ferramenta o *Facebook Prophet*, por isso deve ser renomeado as colunas *Date* e *Sales*, a *Date* que se refere a data para ds, e o objetivo que é as vendas a coluna *Sales*, renomear para y. Outro passo será ordenar as vendas pela data.

Aplicando o método *make_future_dataframe* do *Prophet*, na figura 30 foi criado um novo conjunto de dados com as datas futuras., após isso foi feito a plotagem dessas vendas, e a plotagem dos recursos desse novo conjunto de dados o *forecast*, que se refere as previsões.

Figura 30: Modelo de previsões criado em *python.*

```
def sales_prediction(store_id, sales_df, holidays, periods):
  sales_df = sales_df[sales_df['Store'] == store_id]
  sales_df = sales_df[['Date', 'Sales']].rename(columns = {'Date': 'ds', 'Sales': 'y'})
  sales_df = sales_df.sort_values(by = 'ds')

  model = Prophet(holidays=holidays)
  model.fit(sales_df)
  future = model.make_future_dataframe(periods = periods)
  forecast = model.predict(future)
  figure1 = model.plot(forecast, xlabel = 'Data', ylabel = 'Vendas')

  return sales_df, forecast,model
```

Fonte: Autoria própria (2021).

Conforme a Figura 30, o modelo retornar o conjunto de dados das vendas com os nomes das colunas que serão utilizadas como parâmetro, e o conjunto de dados das previsões, variável *model* referente ao modelo do *Prophet*, será importante retornar, pois ela será utilizada como parâmetro para validação cruzada, o método utilizado para identificar a precisão do modelo.

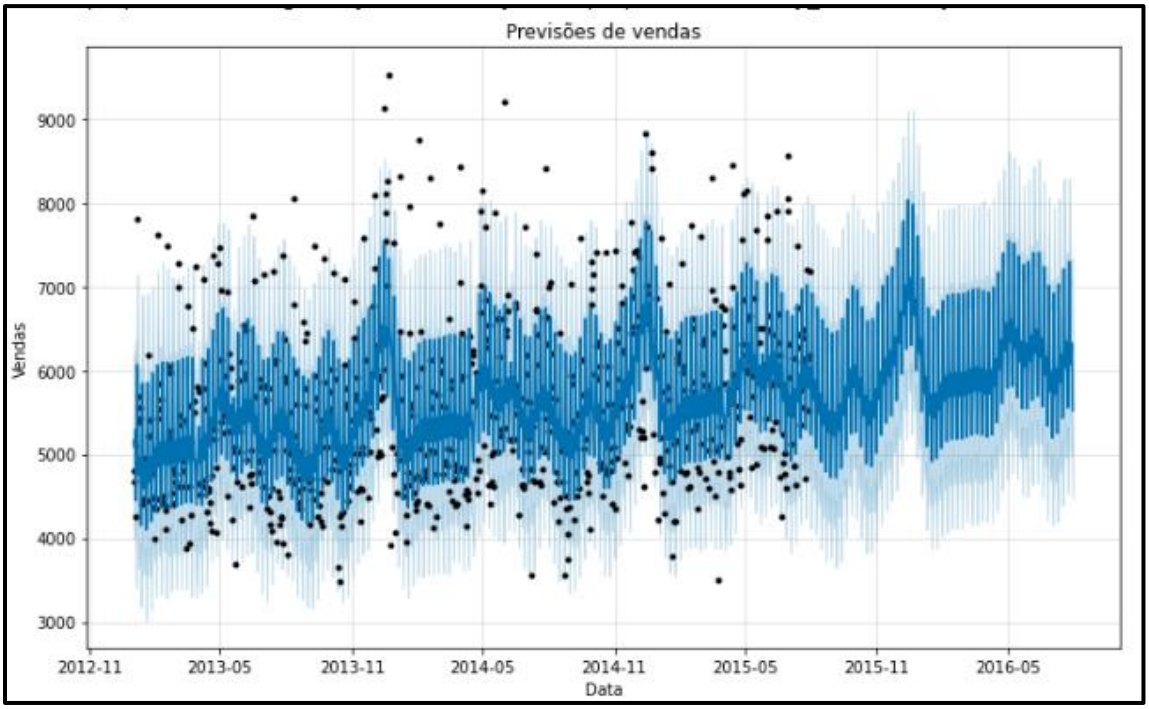
4.7 Utilização do modelo

Após já ter definido o modelo, é possível utilizá-lo para concluir informações sobre os dados aplicado no ambiente de produção.

As previsões irão ser feitas sobre a coluna ds que se refere as datas. Conforme já abordado na seção anterior precisasse definir o período e a loja que tem como objetivo a previsão, conforme demostrado abaixo foi feito a escolha da loja 10, e a previsão de um período de 365 dias.

Logo então irá ser plotado na figura 31 a previsão chamando o método *Prophet*.plot() e passando o conjunto de dados de previsões como argumento. O algoritmo consegue predizer bem a tendência, apesar dos pontos fora da curva atrapalharem a estimativa de incerteza, mas não tem tanto impacto assim na média previsão principal.

Figura 31: Gráfico de previsões das vendas



Fonte: Autoria própria (2021).

Este gráfico contém todas as vendas que foram realizadas e as previsões. Desta forma é fácil visualizar as previsões das vendas que realmente aconteceram, prestando atenção nos pontos pretos, onde não existe mais os pontos são onde as vendas antigas terminam e as previsões começam, exatamente no dia 01-08-2015 começam as previsões, e terminam no dia 30-07-2016 um ano depois. Com isso o modelo conseguiu fazer a previsão de vendas com os dados dos conjuntos obtidos a partir da plataforma Kaggle.

Porém para o entendimento de quão preciso são essas previsões precisamos utilizar alguma forma de validação, a forma de validação utilizada para este modelo, é a **validação cruzada** que é disponibilizada também no Facebook *Prophet* será feito a abordagem desse método no próximo tópico.

4.7.1 Validação cruzada

Esse método do ***Prophet*** de validação cruzada pode ser feito automaticamente para uma

faixa de cortes históricos utilizando a função *cross_validation*. E então é feito a especificação do horizonte de previsão ***horizon*** e logo em seguida, de forma opcional, o tamanho do período de treinamento inicial ***initial*** e o espaçamento entre as datas de corte que é o ***period,*** é possível observar a saída da função a partir da figura 32. Porém por ser opcional, ou seja, por padrão, o período de treinamento inicial é definido como três vezes o horizonte e os cortes são feitos a cada meio horizonte (Sean J. Taylor e Letham, 2017). Nesta pesquisa foi utilizado o horizonte de previsões (*horizon*) de 180 dias, e os outros parâmetros foram utilizados o padrão do método.

Figura 32: Saída da função *cross_validation* do modelo utilizado.

	ds	yhat	yhat_lower	yhat_upper	y	cutoff
0	2014-08-06	5576.248409	4595.670547	6541.841238	5811	2014-08-05
1	2014-08-07	5592.267071	4524.504333	6639.660268	5452	2014-08-05
2	2014-08-08	5833.951112	4797.855078	6920.955072	5683	2014-08-05
3	2014-08-09	5228.776189	4264.346051	6210.812824	4420	2014-08-05
4	2014-08-11	6945.275650	5886.188386	7970.690573	5661	2014-08-05
...	...	...	...	...	...	...
444	2015-07-27	7014.613080	6002.406928	8112.570212	7212	2015-02-01
445	2015-07-28	6168.691560	5104.808418	7297.643917	6140	2015-02-01
446	2015-07-29	5719.112270	4779.313174	6832.878849	5524	2015-02-01
447	2015-07-30	5754.263131	4642.085237	6822.501464	6186	2015-02-01
448	2015-07-31	5994.039725	4923.614660	7075.203919	7185	2015-02-01

449 rows × 6 columns

Fonte: Autoria própria (2021).

A saída da função *cross_validation* é um quadro de dados com os valores reais a coluna y é os valores de previsão fora da amostra a coluna yhat, também tem as colunas yhat_lower e yhat_upper o intervalo de incertezas, e em particular, uma previsão é feita para cada ponto observado pela a coluna cutoff, que seria os cortes feitos pela função.

Logo após ser feita a validação cruzada, também é oferecido pelo *Prophet* uma função para avaliar a performance da validação, onde é passado o conjunto de dados que recebeu a validação cruzada e então será retornado o conjunto de dados com as principais métricas de desempenho. É possível observar essas métricas de desempenho pela figura 33.

Figura 33: Métricas de desempenho da avaliação.

	horizon	mse	rmse	mae	mape	mdape	coverage
0	17 days	611836.493188	782.199778	646.107841	0.121256	0.107092	0.818182
1	18 days	662445.369648	813.907470	668.756687	0.129428	0.111383	0.795455
2	19 days	659342.505578	811.999080	662.809752	0.128248	0.111383	0.795455
3	20 days	794247.389387	891.205582	711.708189	0.140162	0.113183	0.772727
4	21 days	842623.663823	917.945349	742.747917	0.145968	0.116369	0.750000

Fonte: Autoria própria (2021).

Neste estudo irá ser focado apenas o MAPE, ou erro percentual absoluto médio, pois é medido o erro em porcentagem também é possível observar a fórmula pela figura 34.

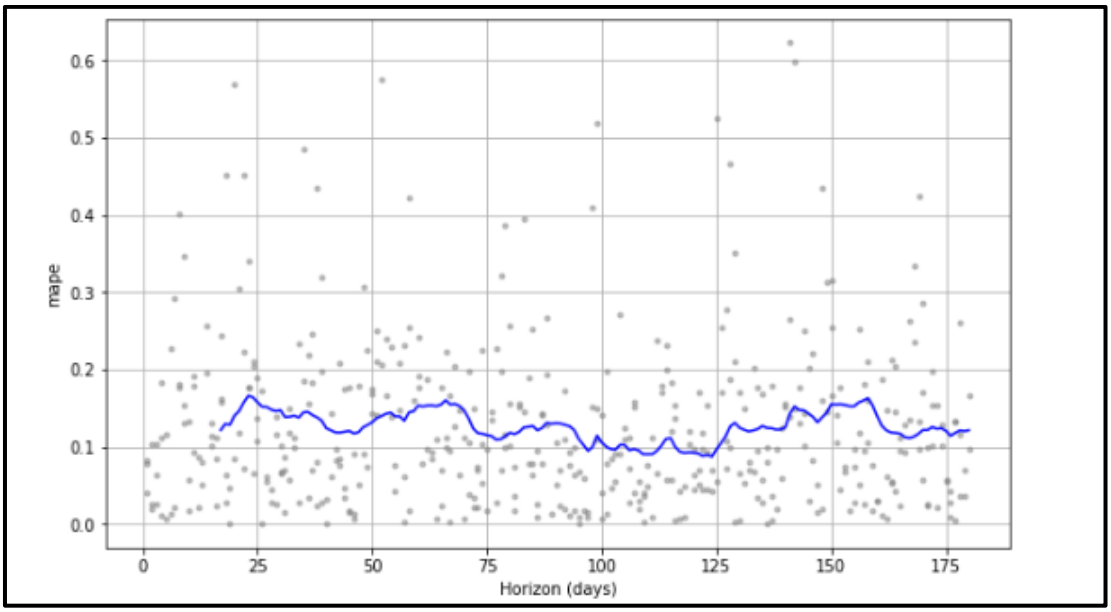
Figura 34: Fórmula do erro percentual absoluto médio

$$\left(\frac{1}{n}\sum\frac{|Real - Previsão|}{|Real|}\right) * 100$$

Fonte: Belge 2012

Na figura 35, será utilizado os valores de mape em um gráfico para visualizar a curva de erro durante o horizonte de tempo que projetamos.

Figura 35: Gráfico do mape em relação ao horizonte de tempo.



Fonte: Autoria própria (2021).

Este erro do modelo poderia provavelmente ser melhorada ajustando parâmetros, incluindo mudanças nos padrões de tendências, sazonalidades e intervalos de incerteza, mas para a pesquisa não ficar tão extensa e complexa, estes pontos ficarão para trabalhos futuros.

5. RESULTADOS OBTIDOS

Nesta seção será exemplificado os possíveis usos da desta pesquisa para auxiliar as empresas, ou seja, como utilizar a pesquisa desenvolvida para a tomada de decisão da empresa, pois é importante a elaboração da apresentação dos dados então obtidos, e da visualização das informações a partir de gráficos e relatórios. Dessa forma auxiliando na visualização de *insights* mais claros, nas atividades dos tomadores de decisão.

Como está pesquisa não visa auxiliar os tomadores de decisões de forma imediata, e foi feito apenas uma exploração dos dados e utilizando os conceitos básicos da ciência de dados para o estudo do tema aplicado ao setor varejista, então pode ser adotado alguns exemplos hipotéticos de auxílios em tomadas de decisões, como os análises feitos na seção de exploração de dados, além do modelo criado. Em vista disso alguns gráficos que foram feitos que auxiliariam em *insights*, como por exemplo na figura 26 que se trata das média por mês, pode ser verificado quais meses existe picos na vendas, assim auxiliando a empresa em quais meses ela necessita de mais estoques de seus produtos, ou quais meses precisam reforçar o número de colaboradores. Outro gráfico que pode auxiliar no insights é o da figura 27, que está se referindo ao número de vendas por dia, pois nele pode-se ter uma noção das vendas em relação ao mês, assim identificando quais dias necessitam de mais anúncios para atrais mais clientes para aumentar os números de vendas, outro exemplo seria o gráfico da figura 28, referente as vendas por dias da semana, com base nisso é possível retirar a informação de quais dias da semana tem maior fluxo de vendas e quais dias precisam ser melhorados. Como esses exemplos citados e outros gráficos feitos na pesquisa é possível auxiliar a empresa em seus eventos diários e em alguns ajustes para possíveis melhorias. Outra grande possível ajuda que deve ser considerado é o modelo desenvolvido pois a partir dele é possível prever vendas de dias futuros auxiliando a empresa a ter essas informações previamente, assim auxiliando com as tomadas de decisões tomadas.

6.CONCLUSÃO

Como explorado no ciclo de vida dos dados, pode-se concluir que os dados se tornaram estratégias importantes para serem utilizadas no processo de tomada de decisão.

Diante disso, a ciência de dados se torna um papel crucial para os negócios e assim também no varejo. Pois é muito mais complexo se retirar informações úteis diante debases de dados brutas, por tanto é necessário utilizar as ferramentas da ciência de dados.

Tendo isso em vista na pesquisa foi desenvolvido a aplicação da ciência de dados no setor varejista utilizando dados da empresa *Rosseman,* a partir desses dados foi possível construir gráficos e relatórios importantes, assim de certa forma auxiliando a empresa no processo de tomada de decisão, para a criação do modelo que previu as vendas a partir de uma loja especifica para prever as vendas durante um período de 365 dias, foi utilizado a ferramenta *Facebook Prophet* que a partir dela é possível fazer a análise de series temporais e de maneira de certa forma até simples de se trabalhar com dados históricos e construir as previsões com base na sua sazonalidades, tendências e feriados ou datas especiais.

Por tanto a partir da pesquisa desenvolvida pode-se concluir que a ciência de dados é uma ferramenta importante que pode auxiliar bastante, a partir de dados já existentes fazer uma análise e identificar pontos que podem ser melhorados nas empresas, ou até mesmo fazer possíveis previsões desses dados, assim auxiliando as empresas a terem fundamentos nos que se basear para tomar decisões importante que podem ajudar no crescimento da mesma.

REFERÊNCIAS

BELGE: Consultoria. In: Avaliando e aprimorando previsões com base nos erros. [S. l.], 1 fev. 2012. Disponível em: http://belge.com.br/blog/2012/02/01/avaliando-e-aprimorando-previsoes-com-base-nos-erros/. Acesso em: 19 nov. 2021.

CAMILA FARANI. Sociedade Brasileira de Varejo e Consumo. E-COMMERCE DEVE CRESCER 43% EM 2020. Disponível em: http://sbvc.com.br/ecommerce-crescer-43-2020/. Acesso em: 16 maio 2021.

FONTES, Thiago. Ciência de dados: o que é, como funciona e qual importância. 2020. Disponível em:https://blog.ccmtecnologia.com.br/post/ciencia-de-dados-o-que-e-como-funciona-qual-importancia. Acesso em: 25 maio 2021.

GIULIANI, A.C. Marketing em um ambiente globalizado. São Paulo: Cobra, 2003, 287p.

GOUVEIA, Fágner Sousa, et al. O MARKETING E SUA IMPORTANCIA PARA O VAREJO. **REVISTA CIENTÍFICA DO ITPAC**, v. 4, p. 12, 01 2011. Disponível em: https://assets.unitpac.com.br/arquivos/revista/41/4. pdf. Acesso em: 16 mai. 2021.

GRUS, Joel. **Data Science do Zero**: Primeiras Regras com o Python. Alta Books, v. 3, f. 168, 2019. 336 p.

JUNQUEIRA, Gabriel. **A era dos Dados e o Varejo**. infovarejo. 2020. Disponível em: https://www.infovarejo.com.br/a-era-dos-dados-e-o-varejo/. Acesso em: 16 mai. 2021.

KAGGLE. Rossmann *Store* Sales: Forecast sales using *store*, promotion, and competitor data. In: Featured Prediction Competition. [S. l.], 7 dez. 2015. Disponível em: https://www.kaggle.com/c/rossmann-*store*-sales/discussion. Acesso em: 19 nov. 2021.

KOTLER, P. Administração de marketing: análise, planejamento, implementação e controle. Tradução de Ailton Bomfim Brandão. 5.ed.São Paulo: Atlas, 1998.725p.

KOTLER, Philip. Administração de Marketing. 10 ed. São Paulo: Pretice Hall, 2000.

LUMINATTI, Carlos Eduardo. **O uso da Ciência de Dados (Data Science) no varejo traz muitos benefícios**. IntelliTi. 2019. Disponível em: https://planejamentodovarejo.com.br/ciencia_de_dados_no_varejo/. Acesso em: 16 mai. 2021.

MAURICIO SALVADOR. Associação Brasileira de Comércio Eletrônico. Faturamento do setor de e-commerce tem alta de 16% no primeiro semestre de 2019. Disponível em: https://abcomm.org/noticias/. Acesso em: 16 maio 2021.

PARENTE, J. Varejo no Brasil: gestão e estratégia. São Paulo: Atlas, 2000.

RAUTENBERG, sandro; CARMO, Paulo Ricardo Viviurka. **BIG DATA E CIENCIA DE DADOS: COMPLEMENTARIEDADE CONCEITUAL NO PROCESSO DE TOMADA DE DECISÃO**. Periódicos ufpb. Paraná, 2019. 12 p. Disponível em: https://periodicos.ufpb.br/ Acesso em: 12 mai. 2021.

SANTOS, Nayara. Varejo online: tudo o que você precisa saber sobre ele! 2020. Disponível

em: https://listenx.com.br/blog/varejo-online/. Acesso em: 16 maio 2021.

STANTON, J. M. Introduction to data science. 2013.

TAYLOR, Sean j.; LETHAM, Benjamin. Forecasting at Scale. ***Prophet***, [*s. l.*], 27 set. 2017. Disponível em: https://peerj.com/preprints/3190v2/. Acesso em: 19 nov. 2021.

THIAGO CHUEIRI. Paypal. Pesquisa "Perfil do E-commerce Brasileiro 2020": ritmo de expansão do total de lojas online no Brasil é superior a 40% ao ano. 2020. Disponível em: https://newsroom.br.paypal-corp.com/. Acesso em: 16 maio 2021.

WAZLAWICK, R. S. Metodologia de pesquisa para ciência da computação. 2. ed. Rio de Janeiro: Elsevier, 2014.