



**PUC Minas**

**PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS**

**NÚCLEO DE EDUCAÇÃO A DISTÂNCIA**

**Pós-graduação *Lato Sensu* em Analytics e Business Intelligence**

**SUPER LOJA**

Marília Figueiredo Santos

Belo Horizonte

2023

# SUMÁRIO

<b>1. Introdução .....</b>	<b>3</b>
1.1. O problema proposto .....	4
<b>2. Coleta de Dados .....</b>	<b>5</b>
<b>3. Processamento/Tratamento de Dados.....</b>	<b>7</b>
3.1. Ferramentas utilizadas.....	7
3.1.2 Bibliotecas utilizadas.....	8
<b>3. Integração, Tratamento e Carga de Dados .....</b>	<b>8</b>
3.1. Fontes de Dados.....	8
3.2. Obtendo dados .....	9
3.2.1 Dataset: Superloja 2023 PBI.csv e Orçamento de Vendas.....	9
<b>4. Análise e Exploração dos Dados.....</b>	<b>17</b>
4.1 Exploração dos dados .....	18
4.2 Análises com novas colunas.....	19
<b>5. Modelo Canvas.....</b>	<b>27</b>
<b>6. Links .....</b>	<b>27</b>
<b>REFERÊNCIAS .....</b>	<b>28</b>

## 1. Introdução

A presente pesquisa aborda o tema de uma Super Loja do seguimento mobiliário corporativo e consumidor que atua em todo EUA nas regiões Sudeste, Leste, Oeste e Central, consiste nas vendas de produtos de escritório, moveis e tecnologia nas quais os atores se relacionam por um meio eletrônico. As mudanças representadas pelas transações eletrônicas impactaram em crescimento econômico, na geração de empregos e criaram relações singulares entre o setor produtivo e seus consumidores.

O comércio por intermédio eletrônico advém da evolução das tecnologias de comunicação e informação e está acessível a praticamente todo tipo de empresa, dos mais variados tamanhos e setores, em qualquer parte do mundo. A integração proposta pela globalização e pelas novas tecnologias da informação causou a tendência de aumento desse tipo de comércio no ambiente interno e externo.

A internet tem se tornado parte da vida e do cotidiano de grande parte dos cidadãos. Há um entendimento global que ela deveria ser alçada a um direito humano fundamental e universal. As organizações, por sua vez, utilizam cada vez mais a tecnologia de informação tanto no ambiente organizacional, quanto na relação com seus clientes. As mudanças de paradigmas causadas pela globalização e pela quebra de barreiras físicas no mundo eletrônico justificam o estudo e a aplicação do comércio eletrônico como ferramenta fundamental de aquisição e sustentação de novos mercados pelas empresas. A mineração de dados é uma das principais ferramentas para obter dados relevantes e realizar um mapeamento mais aguçado para ajudar na tomada de decisões.

Esse trabalho objetiva aplicar técnicas de Mineração de Dados e Modelos Preditivos em vendas a fim de classificar e encontrar padrões extrair algumas informações e gerar insights que podem influenciar em estratégias de vendas e vendas.

A estratégia metodológica utilizada no presente trabalho foi um estudo descritivo e exploratório, com análise dos dados através de uma abordagem quantitativa. Primeiramente buscou-se analisar as informações das vendas da

Super Loja. Essas informações foram obtidas no site data.world. As palavras-chave utilizadas nas pesquisas foram: Super Lojas, comércio eletrônico, lojas virtuais.

Posteriormente à coleta das informações, foi realizada a leitura e seleção do material e em seguida foi efetuada uma análise com o objetivo de compreender e estender o conhecimento sobre as vendas da Super Loja e, assim, elaborar o referencial teórico da investigação. O Trabalho foi feito utilizando o Google Colab.

### **1.1. O problema proposto**

Neste trabalho será utilizado Análise Exploratória e Modelagem Preditiva para extração de informações importantes do conjunto de dados da Super Loja com o objetivo de realizar uma análise visando o aumento da receita e crescimento do negócio. Para isso todos os atributos serão classificados com um grau de importância, desta maneira conseguimos analisar os resultados e utilizá-los em previsões futuras.

Para isso, serão analisados o conjunto de dados da Super Loja, disponibilizados no site da data.world. Os principais objetivos dessa análise são:

- Realizar uma análise nos dados das vendas da Super Loja para tomar decisões estratégicas na venda desses produtos.
- Os dados que serão analisados, foram coletados do site da data.world. Foi necessário coletar algumas informações separadamente, são elas:
  1. Dataset da Super Loja: neste dataset é apresentado informações sobre as vendas dos produtos por categorias, contendo informações como: seguimento, nome do cliente, cidade, região, quantidade.
  2. Dataset do Orçamento da Super Loja: neste dataset é apresentado o valor do orçamento das vendas.
- As análises realizadas, têm como objetivo encontrar padrões, métricas e tendências que auxiliarão no entendimento das bases trabalhadas. E assim

poderemos indicar quais características da Super Loja que a tornam uma empresa com crescente receita e um negócio durável.

- Em relação a aspectos geográficos, os resultados obtidos destinam-se exclusivamente a vendas online ocorridas no site da Super Loja.
- Os dados coletados são de 2020 á 2023, ordenados pela data de publicação crescente.

<sup>1</sup> <https://data.world/ehughes/superstore-sales-2023/wortspace/file?filename=Supertore+2023.csv>

<sup>2</sup><https://data.world/mlongoria/superstore-sales-budget-2023/workspace/file?filename=Superstore+2023+PBI.csv>

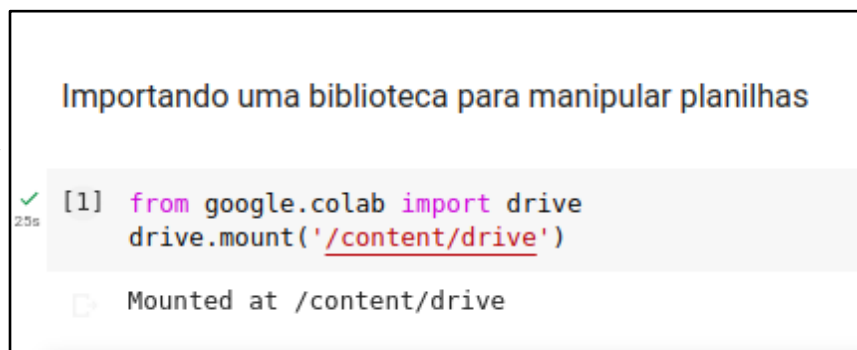
## 2. Coleta de Dados

O Trabalho foi desenvolvido no Google Colab, onde a coleta de dados foi separada em duas partes: a primeira e a segunda com dados do data.world com dados da Super Loja e Orçamento de vendas. Como observação, os dados já vieram formatados em csv, por este motivo não foi necessário nenhuma conversão de formato.

No Google Colab, onde o projeto foi executado, a importação foi feita conforme abaixo:

Importar o dataset do data.world, o qual foi feito download, para isso foi necessário utilizar uma importação do google, selecionar o arquivo (fazendo upload para o Google Drive) e preparando ele para o Pandas fazer a leitura.

**Figura 1:**  
Importando  
arquivo csv



```

Importando uma biblioteca para manipular planilhas

[1] from google.colab import drive
     drive.mount('/content/drive')

Mounted at /content/drive
  
```

Fonte: Autor

## 2.1 Listagem descritiva das colunas:

A tabela abaixo mostra uma descrição da estrutura encontrada no Dataset superLoja.csv, exibindo o nome do atributo, tipo e sua descrição.

**Tabela 1:** Estrutura Dataset superLoja.csv

Nome da coluna/campo	Descrição	Tipo
index	Índice	float
Row ID	ID Pedido	float
Order ID	Tipo de Cliente	object
Order Date	Data do pedido	object
Ship Date	Data do Envio	object
Ship Mode	Modo de Envio	object
Customer ID	ID CLiente	object
Customer Name	Nome do CLiente	object
Segment	Seguimento	object
Country/Region	País	object
City	Cidade	object
State	Estado	object
Postal Code	Código Postal	object
Region	Região	object
Product ID	ID do Produto	object
Category	Categoria	object
Sub-Category	Sub Categoria	object
Product Name	Nome do Produto	object
Sales	Vendas	float
Quantity	Quantidade	float
Discount	Desconto	float

Fonte: Autor

O Dataset OrçamentoVendas.csv contém informações referentes ao Orçamento de vendas em python. Ele contém a seguinte estrutura:

**Tabela 2:**  
Orçamento  
Vendas.csv

Nome da coluna/campo	Descrição	Tipo
row_id	ID Pedido	integer
order_date	Tipo de Cliente	object
region	Região	object
product_name	Nome do Produto	object
budget	Orçamento	float

Fonte: Autor

Dessa forma os dados estão preparados para serem trabalhados e posteriormente analisados.

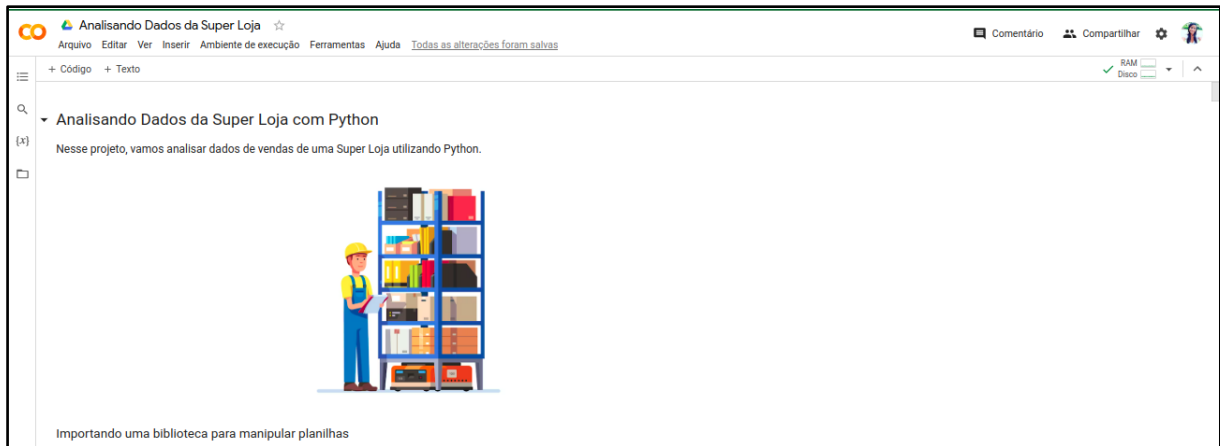
### 3. Processamento/Tratamento de Dados

Nessa seção será apresentado todas as ferramentas e bibliotecas utilizadas para o processamento e o tratamento dos dados.

#### 3.1. Ferramentas utilizadas

Como ferramenta para desenvolvimento dos scripts em python, foi escolhido a ferramenta em nuvem do Google Colaboratory, (figura 4), disponível em <https://colab.research.google.com/>

**Figura 2:** Captura da tela do Google Colab



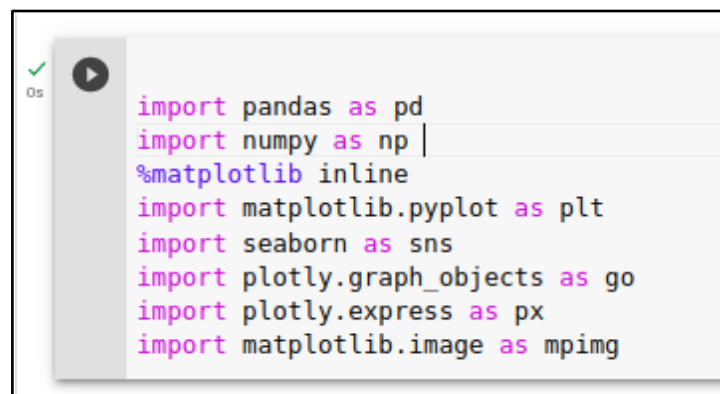
Fonte: Autor

Foi escolhida a ferramenta Google Colab, pois possui as principais ferramentas e bibliotecas para realizarmos toda a codificação necessária para análise e tratamento dos dados. Esta ferramenta nos permite unir código e texto, facilitando nossa organização no projeto.

### 3.1.2 Bibliotecas utilizadas

Para realizar o processamento e o tratamento dos dados, foi necessário importar algumas bibliotecas conforme a Figura 5 abaixo.

**Figura 3:** Captura da tela da importação das bibliotecas



Fonte: Autor

## 3. Integração, Tratamento e Carga de Dados



### 3.1. Fontes de Dados

Descrição das bases de dados ou arquivos utilizados pelo projeto como fonte, apresentando possíveis diagramas dos bancos de dados relacionais.

A tabela 2 abaixo mostra de forma detalhada as bibliotecas importadas.

**Tabela 2:** Bibliotecas utilizadas

Bibliotecas	Descrição	Comando(s) utilizados
Pandas	Pacote de ferramentas para análise de dados e manipulação, construída sobre a base da linguagem de programação python. Informações: <a href="https://pandas.pydata.org/">https://pandas.pydata.org/</a>	<code>import pandas as pd</code>
Numpy	Pacote de ferramentas utilizada para realizar cálculos em Arrays Multidimensionais, construída sobre a base da linguagem de programação python. Informações: <a href="https://numpy.org/">https://numpy.org/</a>	<code>import numpy as np</code>
Matplotlib	Pacote de ferramentas utilizada para criação de gráficos e visualização de dados, construída sobre a base da linguagem de programação python. Informações: <a href="https://matplotlib.org/">https://matplotlib.org/</a>	<code>import matplotlib.pyplot as plt</code> <code>%matplotlib inline</code> <code>import matplotlib.image as mpimg</code>
Seaborn	Pacote de ferramentas utilizadas para criação de gráficos e visualização de dados de alto nível baseada na lib Matplotlib. Informações: <a href="https://docs.python.org/pt-br/3/library/datetime.html">https://docs.python.org/pt-br/3/library/datetime.html</a>	<code>import seaborn as sns</code>
Plotly graph	Pacote de ferramentas utilizadas para criação de gráficos. <a href="https://plotly.com/python/graph-objects/">https://plotly.com/python/graph-objects/</a>	<code>import plotly.graph_objects as go</code>
plotly.express	Pacote de ferramentas utilizadas para criação de figuras. <a href="https://plotly.com/python/plotly-express/">https://plotly.com/python/plotly-express/</a>	<code>import plotly.express as px</code>

Fonte: Autor

### 3.2. Obtendo dados

Nessa seção será apresentado como foi coletado os dados após a execução do notebook Analisando Dados da Super Loja, salvo no google drive arquivo csv.

#### 3.2.1 Dataset: Superloja 2023 PBI.csv e Orçamento de Vendas

Na Figura abaixo é apresentado a parte do código utilizada para obter os dados salvos em formato csv e transformá-los em um formato Data Frame para iniciarmos a nossa análise.

Esses dois datasets foram unificados, onde foram realizadas as análises diante do cenário da Super Loja, ou seja, verificar a correlação com as vendas e o orçamento de vendas.



mazenando dados df1

Fonte: Autor

Para esse caso, foi criado a variável df1 para trazer os dados da super loja e armazenar no dataframe.

O resultado em df1 dos dados coletados,selecionando apenas os 5 primeiros registros encontrados através do comando df1.head().

**Figura 5:** Exbindo dados dataset do df1

Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	Country/Region	City	Postal Code	Region	Product ID	Category	Sub-Category	Product Name	Sales	Quantity	Discount	Profit
1	CA-2020-152156	08/11/2022 00:00:00	11/11/2020	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	42420.0	South	FUR-BO-10001798	Furniture	Bookcases	Bush Somerset Collection Bookcase	261.9600	2	0.00	41.9136
2	CA-2020-152156	08/11/2022 00:00:00	11/11/2020	Second Class	CG-12520	Claire Gute	Consumer	United States	Henderson	42420.0	South	FUR-CH-10000454	Furniture	Chairs	Hon Deluxe Fabric Upholstered Stacking Chairs,...	731.9400	3	0.00	219.5820
3	CA-2020-138688	12/06/2022 00:00:00	16/06/2020	Second Class	DV-13045	Darrin Van Huff	Corporate	United States	Los Angeles	90036.0	West	OFF-LA-10000240	Office Supplies	Labels	Self-Adhesive Address Labels for Typewriters b...	14.6200	2	0.00	6.8714
4	US-2019-108966	11/10/2021 00:00:00	18/10/2019	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	33311.0	South	FUR-TA-10000577	Furniture	Tables	Bretford CR4500 Series Slim Rectangular Table	957.5775	5	0.45	-383.0310
5	US-2019-108966	11/10/2021 00:00:00	18/10/2019	Standard Class	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	33311.0	South	OFF-ST-10000760	Office Supplies	Storage	Eldon Fold 'N Roll Cart System	22.3680	2	0.20	2.5164

aws x 21 columns

Fonte: Autor

A extração apenas do Orçamento de vendas é executada conforme o script abaixo.

**Figura 6:** Armazenando dados df2

```
df2 = pd.read_csv('https://query.data.world/s/tsf2jsw7te3hb7nne6ukywo1fzfthj?dws=00000')
df2.head()
```

Fonte: Autor

O resultado em df2 dos dados coletados,selecionando apenas os 5 primeiros registros encontrados através do comando df2.head().

**Figura 7:** Exibindo dados dataset do df2

	row_id	order_date	region	product_name	budget
0	1	2022-11-08 00:00:00	South	Bush Somerset Collection Bookcase	261.9600
1	2	2022-11-08 00:00:00	South	Hon Deluxe Fabric Upholstered Stacking Chairs,...	731.9400
2	3	2022-06-12 00:00:00	West	Self-Adhesive Address Labels for Typewriters b...	14.6200
3	4	2021-10-11 00:00:00	South	Bretford CR4500 Series Slim Rectangular Table	957.5775
4	5	2021-10-11 00:00:00	South	Eldon Fold 'N Roll Cart System	22.3680

Fonte: Autor

### 3.3 Informações dos DataFrames

Para cada objeto DataFrame foi utilizado a função info() para visualizarmos algumas informações como a quantidade de registros, quantidade de colunas, informações de cada coluna e o tipo dela. Podemos observar na Figura abaixo, que no DataFrame df1 foram encontrados 9994 registros com um total de 21 colunas.

**Figura 8:**  
Exibindo informações  
DataFrame df1

```
[18] #Vesualizando informações sobre os dados antes da formatação
df1.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9994 entries, 0 to 9993
Data columns (total 21 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Row ID                 9994 non-null  int64
1   Order ID               9994 non-null  object
2   Order Date             9994 non-null  object
3   Ship Date              9994 non-null  object
4   Ship Mode              9994 non-null  object
5   Customer ID            9994 non-null  object
6   Customer Name          9994 non-null  object
7   Segment                9994 non-null  object
8   Country/Region         9994 non-null  object
9   City                   9994 non-null  object
10  State                  9994 non-null  object
11  Postal Code            9983 non-null  float64
12  Region                 9994 non-null  object
13  Product ID             9994 non-null  object
14  Category               9994 non-null  object
15  Sub-Category           9994 non-null  object
16  Product Name           9994 non-null  object
17  Sales                  9994 non-null  float64
18  Quantity               9994 non-null  int64
19  Discount               9994 non-null  float64
20  Profit                 9994 non-null  float64
dtypes: float64(4), int64(2), object(15)
memory usage: 1.6+ MB
```

Fonte: Autor

Para o DataFrame com os valores da tabela orçamento de vendas, temos o seguinte resultado:

**Figura 9:**  
Exibindo  
informações  
DataFrame df2

```
[8] #Vesualizando informações sobre os dados antes da formatação
df2.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9977 entries, 0 to 9976
Data columns (total 5 columns):
#   Column                Non-Null Count  Dtype
---  -
0   row_id                 9977 non-null  int64
1   order_date             9977 non-null  object
2   region                 9977 non-null  object
3   product_name           9977 non-null  object
4   budget                 9977 non-null  float64
dtypes: float64(1), int64(1), object(3)
memory usage: 389.9+ KB
```

Fonte: Autor

Logo podemos ver que foram encontrados 9977 registros a quantidade do DataFrame da super loja, porém somente com cinco colunas.

### 3.4 Tratamento dos Dados

Nessa seção será detalhado cada passo realizado no tratamento dos dados antes de começarmos as análises.

#### 3.4.1 Removendo as colunas

Após realizado análise em cima dos DataFrames df1 e df2, verificou-se que existem colunas que não serão necessárias e não terão impacto em análises futuras. Desta forma essas colunas foram removidas conforme a Figura 10 e 11.

**Figura 10:** Remoção colunas DataFrame df1

```
#Excluindo colunas
del df1['index']
del df1['Order ID']
del df1['Ship Mode']
del df1['Country/Region']
del df1['Postal Code']
del df1['Product ID']
del df1['Product Name']
del df1['Customer ID']
```

Fonte: Autor

**Figura 11:** Remoção colunas DataFrame df2

```
#Excluindo colunas
del df2['row_id']
del df2['order_date']
del df2['region']
del df2['product_name']
```

Fonte: Autor

#### 3.4.2 Unindo os DataFrames

Antes de realizarmos o tratamento em cada Dataset, para facilitar foi feita a união dos DataFrames. Representado na figura 12 abaixo:

**Figura 12:** Unindo os DataFrame: superLoja

```
# Unindo os datasets
superLoja = pd.concat([df1, df2])
superLoja.head()
```

Fonte: Autor

### 3.4.3 Renomeando as colunas

Para deixar nossos dados formatados e padronizados, algumas colunas foram renomeadas utilizando a função `rename()`, conforme a Figura 13 abaixo.

```
#Traduzindo as colunas do Dataset

superLoja.rename(columns={
    'Row ID': 'ID_Pedido',
    'Order Date': 'Data_pedido',
    'Ship Date': 'Data_envio',
    'Customer Name': 'Nome_cliente',
    'Segment': 'Seguimento',
    'City': 'Cidade',
    'State': 'Estado',
    'Region': 'Regiao',
    'Category': 'Categoria',
    'Sub-Category': 'SubCategoria',
    'Sales': 'Vendas',
    'Quantity': 'Quantidade',
    'Discount': 'Desconto',
    'Profit': 'Lucro',
    'budget': 'Orçamento', }, inplace = True)
```

**Figura 13:**

Renomeando as colunas DataFrame: superLoja

Fonte: Autor

### 3.4.4 Traduzindo as linhas do dataset

No entanto, também para deixar os dados formatados e padronizados, algumas linhas foram renomeadas, conforme a Figura 14 abaixo.

**Figura 14:** Renomeando as colunas DataFrame: superLoja

```
[15] #Traduzindo as linhas do dataset
superLoja['Categoria'].replace({'Technology':'Tecnologia', 'Office Supplies':'Material de Escritório', 'Furniture':'Móveis'}, inplace = True)
superLoja['Seguimento'].replace({'Consumer':'Consumidor', 'Corporate':'Corporativo',
                                'Home Office':'Escritório em casa'}, inplace = True)
superLoja['SubCategoria'].replace({'Accessories':'Acessórios', 'Appliances':'Eletrodomésticos', 'Art':'Art', 'Binders':'Fichários',
                                   'Bookcases':'Estantes', 'Chairs':'Cadeiras', 'Copiers':'Copiadoras', 'Envelopes':'Envelopes',
                                   'Fasteners':'Fechos', 'Furnishings':'Mobiliária', 'Labels':'Etiquetas', 'Machines':'Máquinas',
                                   'Paper':'Papel', 'Phones':'Telefones', 'Storage':'Armazenar', 'Supplies':'Suprimentos',
                                   'Tables':'Tabelas' }, inplace = True)
superLoja['Regiao'].replace({'Central': 'Central', 'East':'Leste', 'West':'Oeste', 'South':'Sul'}, inplace = True)
```

Fonte: Autor

### 3.4.4 Padronizando os tipos das colunas

Para deixar os tipos de cada coluna padronizados, foi realizada a conversão dos tipos em algumas colunas utilizando a função `astype()` e outros em `to_datetime`, conforme Figura 15.

**Figura 15:** Padronizando colunas DataFrame: superLoja

```
[59] # Convertendo dados que estão como object para string
superLoja['Nome_cliente'] = superLoja['Nome_cliente'].astype('string')
superLoja['Seguimento'] = superLoja['Seguimento'].astype('string')
superLoja['Cidade'] = superLoja['Cidade'].astype('string')
superLoja['Estado'] = superLoja['Estado'].astype('string')
superLoja['Regiao'] = superLoja['Regiao'].astype('string')
superLoja['Categoria'] = superLoja['Categoria'].astype('string')
superLoja['SubCategoria'] = superLoja['SubCategoria'].astype('string')
```

Fonte: Autor

**Figura 16:** Padronizando colunas DataFrame: superLoja

```
[66] # Convertendo dados que estão como object para Datetime
superLoja["Data_pedido"] = pd.to_datetime(superLoja["Data_pedido"])
superLoja["Data_envio"] = pd.to_datetime(superLoja["Data_envio"])
superLoja.dtypes
```

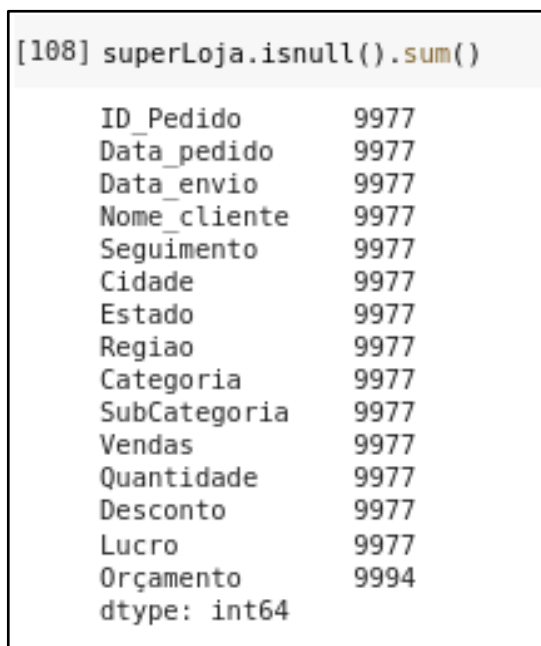
Fonte: Autor

As colunas `Nome_cliente`, `Seguimento`, `Cidade`, `Estado`, `Região`, `Categoria` e `Subcategoria` foram convertidas para `string`. O valor da coluna `Data_pedido` e `Data_envio` estava em `object`, foi convertido para um formato `Datetime`, correspondendo as datas das vendas.

### 3.4.5 Informações dos DataFrames

Para certificar que os dados estão padronizados e que não existem valores nulos no DataFrame, foi executado a função `isnull()` onde recupera todos os registros que estão com valores nulo e logo em seguida aplicado a função `sum()`, pois somamos a quantidade de registros nulos da coluna correspondente, conforme a Figura 18.

**Figura 18:** Verificando valores nulos



```
[108] superLoja.isnull().sum()
```

ID_Pedido	9977
Data_pedido	9977
Data_envio	9977
Nome_cliente	9977
Seguimento	9977
Cidade	9977
Estado	9977
Regiao	9977
Categoria	9977
SubCategoria	9977
Vendas	9977
Quantidade	9977
Desconto	9977
Lucro	9977
Orçamento	9994
dtype:	int64

Fonte: Autor

O resultado obtido apresenta a listagem de todas as colunas e o valor da soma de valores nulos obtidos. Neste caso todas as colunas possuem registro com valores nulos, por isso os valores aparecem com o total de registros.

Por conta desse concat, os dados que não têm dados correspondentes acabaram por ficarem como NaN em todos os campos. Para isso, foi utilizado a função `fillna()` para preencher com zeros esses casos.

**Figura 18:** Substituindo valores NaN por zero



```
# Obtendo uma amostra do conjunto de dados
superLoja.fillna(0, inplace = True)
superLoja.sample(5)
```

	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	Country/Region	City	...	Region	Product ID	Category	Sub-Category	Product Name	Sales	Quantity	Discount	Profit
7481	7482.0	CA-2020-124583	01/09/2022 00:00:00	03/09/2020	Second Class	LB-16795	Laurel Beltran	Home Office	United States	Huntington Beach	...	West	OFF-EN-10002500	Office Supplies	Envelopes	Globe Weis Peel & Seal First Class Envelopes	12.780	1.0	0.0	5.7510
9907	0.0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0.000	0.0	0.0	0.0000
7942	7943.0	CA-2021-134194	25/12/2023 00:00:00	01/01/2022	Standard Class	GA-14725	Guy Armstrong	Consumer	United States	Dallas	...	Central	OFF-BI-10003684	Office Supplies	Binders	Wilson Jones Legal Size Ring Binders	39.582	9.0	0.8	-59.3730
3182	3183.0	CA-2021-152912	09/11/2023 00:00:00	12/11/2021	Second Class	BM-11650	Brian Moss	Corporate	United States	Columbia	...	East	TEC-AC-10004666	Technology	Accessories	Maxell iVDR EX 500GB Cartridge	826.620	3.0	0.0	355.4466
3227	3228.0	CA-2018-108189	02/10/2020 00:00:00	05/10/2018	First Class	ES-14080	Erin Smith	Corporate	United States	Tempe	...	West	TEC-PH-10001557	Technology	Phones	Pyle PMP37LED	230.376	3.0	0.2	20.1579

5 rows x 22 columns

Fonte: Autor

Dessa forma, agora os dados estão preparados para uma melhor exploração, e nesse caso, verificar a correlação entre as vendas da super Loja e orçamento o de vendas.

**Figura 19:** Verificando valores nulos após tratamento

```
[115] #Verificando se existe algum valor nulo após o tratamento dos dados
superLoja.isnull().sum()
```

ID_pedido	0
Data_pedido	0
Data_envio	0
Nome_cliente	0
Seguimento	0
Cidade	0
Estado	0
Regiao	0
Categoria	0
SubCategoria	0
Vendas	0
Quantidade	0
Desconto	0
Lucro	0
Orçamento	0
dtype: int64	

Fonte: Autor

Para exibir as informações do DataFrame de forma detalhada, utilizamos a função `info()`, ela retorna informações importantes de cada coluna, tais como: nome da coluna, tipo da coluna e se a coluna aceita valores nulos, conforme a Figura 20 abaixo.

**Figura 20:** Verificando valores nulos após tratamento

```
[129] #informações sobre os dado já formatados
superLoja.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 19971 entries, 0 to 9976
Data columns (total 15 columns):
#   Column                Non-Null Count  Dtype  
---  -
0   ID_Pedido              19971 non-null  float64
1   Data_pedido            19971 non-null  datetime64[ns]
2   Data_envio             19971 non-null  datetime64[ns]
3   Nome_cliente           19971 non-null  string  
4   Seguimento            19971 non-null  string  
5   Cidade                 19971 non-null  string  
6   Estado                 19971 non-null  string  
7   Regiao                 19971 non-null  string  
8   Categoria              19971 non-null  string  
9   SubCategoria           19971 non-null  string  
10  Vendas                  19971 non-null  float64
11  Quantidade              19971 non-null  float64
12  Desconto                19971 non-null  float64
13  Lucro                   19971 non-null  float64
14  Orçamento               19971 non-null  float64
dtypes: datetime64[ns](2), float64(6), string(7)
memory usage: 2.4 MB
```

Fonte: Autor

Foi utilizado a função `shape()`, onde podemos visualizar as dimensões do DataFrame.

**Figura 21:** Informações das dimensões do DataFrame

```
[130] #Verificar o número de linhas e colunas
superLoja.shape

(19971, 15)
```

Fonte: Autor

Com isso obtemos como resultado 19.971 linhas e 15 colunas.

## 4. Análise e Exploração dos Dados

Nessa seção será mostrado todas as análises e exploração dos dados tratados anteriormente. Analisaremos as ocorrências, padrões e informações importantes que levantamos do dataset.

### 4.1 Exploração dos dados

Para iniciar a exploração dos dados foi feita, inicialmente a descrição estatística deles:

**Figura 21:** Descrição estatística do DataFrame

```
[189] #Descrição estatística da Super Loja
superLoja.describe()
```

	ID_Pedido	Vendas	Quantidade	Desconto	Lucro	Orçamento	Receita	Receita/Vendas
count	9994.000000	9994.000000	9994.000000	9994.000000	9994.000000	9977.000000	9994.000000	9994.000000
mean	4997.500000	229.858001	3.789574	0.156203	28.656896	230.249660	1149.495905	3.789574
std	2885.163629	623.245101	2.225110	0.206452	234.260108	626.467055	3898.666090	2.225110
min	1.000000	0.444000	1.000000	0.000000	-6599.978000	0.444000	0.444000	1.000000
25%	2499.250000	17.280000	2.000000	0.000000	1.728750	17.280000	48.693500	2.000000
50%	4997.500000	54.490000	3.000000	0.200000	8.666500	54.480000	183.680000	3.000000
75%	7495.750000	209.940000	5.000000	0.200000	29.364000	209.880000	763.201500	5.000000
max	9994.000000	22638.480000	14.000000	0.800000	8399.976000	22638.480000	135830.880000	14.000000

Fonte: Autor

Nessa descrição podemos ver que a mínima e máxima modificaram bastante, o que já indica uma grande movimentação e crescimento da receita.

Por conseguinte, foi analisada a somatória das vendas por região.

**Figura 22:** Vendas por região

```
[173] superLoja.groupby("Regiao").sum()
```

<ipython-input-173-0914b0e41b32>:1: FutureWarning: The default value of nu  
superLoja.groupby("Regiao").sum()

	ID_Pedido	Vendas	Quantidade	Desconto	Lucro	Orçamento
Regiao						
Central	11685963.0	501239.8908	8780.0	558.34	39706.3625	0.0
Leste	14073919.0	678781.2400	10618.0	414.00	91522.7800	0.0
Oeste	15971838.0	725457.8245	12266.0	350.20	108418.4489	0.0
Sul	8213295.0	391721.9050	6209.0	238.55	46749.4303	0.0

Fonte: Autor

Nesse caso, é possível visualizar que a região Oeste foi a que obteve a maior quantidade de vendas e consequentemente o maior percentual de lucro.

Além disso, foi verificado a média de vendas, segue figura abaixo:

**Figura 23: Média vendas**

```
[182] #Média vendas
superLoja["Vendas"].mean()

229.85800083049827
```

Fonte: Autor

Podemos verificar que a média de vendas foi boa, comparado com cada região.

## 4.2 Análises com novas colunas

Após as análises realizadas, concluímos que a receita também pode ser considerada um atributo relevante para tomar decisões estratégicas na venda desses produtos. A receita está diretamente ligada ao dinheiro arrecadado pelas vendas dos determinados produtos da super loja e isso acaba tendo impacto no seu lucro líquido, figura 24. Neste caso, foi criada também a receita de vendas figura 25.

**Figura 24: Coluna de receitas**

```
[236] # Criando a coluna de receita
superLoja["Receita"] = superLoja["Vendas"].mul(superLoja["Quantidade"])
```

```
[237] superLoja.head()
```

	ID_Pedido	Data_pedido	Data_envio	Nome_cliente	Seguimento	Cidade	Estado	Regiao	Categoria	SubCategoria	Vendas	Quantidade	Desconto	Lucro	Orçamento	Receita
0	1.0	2022-08-11	2020-11-11	Claire Gute	Consumidor	Henderson	Kentucky	Sul	Móveis	Estantes	261.9600	2.0	0.00	41.9136	0.0	523.9200
1	2.0	2022-08-11	2020-11-11	Claire Gute	Consumidor	Henderson	Kentucky	Sul	Móveis	Cadeiras	731.9400	3.0	0.00	219.5820	0.0	2195.8200
2	3.0	2022-12-06	2020-06-16	Darrin Van Huff	Corporativo	Los Angeles	California	Oeste	Material de Escritório	Etiquetas	14.6200	2.0	0.00	6.8714	0.0	29.2400
3	4.0	2021-11-10	2019-10-18	Sean O'Donnell	Consumidor	Fort Lauderdale	Florida	Sul	Móveis	Tabelas	957.5775	5.0	0.45	-383.0310	0.0	4787.8875
4	5.0	2021-11-10	2019-10-18	Sean O'Donnell	Consumidor	Fort Lauderdale	Florida	Sul	Material de Escritório	Armazenar	22.3680	2.0	0.20	2.5164	0.0	44.7360

Fonte: Autor

**Figura 25: Coluna de receitas/vendas**

```
[238] # Criando a coluna de receita de vendas
superLoja["Receita/Vendas"] = superLoja["Receita"] / superLoja["Vendas"]
superLoja.head()
```

	ID_Pedido	Data_pedido	Data_envio	Nome_cliente	Seguimento	Cidade	Estado	Regiao	Categoria	SubCategoria	Vendas	Quantidade	Desconto	Lucro	Orçamento	Receita	Receita/Vendas
0	1.0	2022-08-11	2020-11-11	Claire Gute	Consumidor	Henderson	Kentucky	Sul	Móveis	Estantes	261.9600	2.0	0.00	41.9136	0.0	523.9200	2.0
1	2.0	2022-08-11	2020-11-11	Claire Gute	Consumidor	Henderson	Kentucky	Sul	Móveis	Cadeiras	731.9400	3.0	0.00	219.5820	0.0	2195.8200	3.0
2	3.0	2022-12-06	2020-06-16	Darrin Van Huff	Corporativo	Los Angeles	California	Oeste	Material de Escritório	Etiquetas	14.6200	2.0	0.00	6.8714	0.0	29.2400	2.0
3	4.0	2021-11-10	2019-10-18	Sean O'Donnell	Consumidor	Fort Lauderdale	Florida	Sul	Móveis	Tabelas	957.5775	5.0	0.45	-383.0310	0.0	4787.8875	5.0
4	5.0	2021-11-10	2019-10-18	Sean O'Donnell	Consumidor	Fort Lauderdale	Florida	Sul	Material de Escritório	Armazenar	22.3680	2.0	0.20	2.5164	0.0	44.7360	2.0

Fonte: Autor

Para realizar o cálculo da receita de vendas, pegamos a receita e dividimos pelas vendas e desta forma conseguimos comparar a quantidade de vendas por ano.

No entanto, na descrição abaixo podemos verificar a mínima e máxima das receitas, valor bem considerável, se comparada a média de vendas.

**Figura 27:** Coluna de receita min e max

```
[281] # Retornando a maior receita
superLoja["Receita"].max()

135830.88

[282] # Retornando a menor receita
superLoja["Receita"].min()

0.444
```


Fonte: Autor

Já, nas figuras abaixo é possível verificar as 3 maiores receitas e as 3 menores receitas, com isso percebe-se que o resultado da operação do negócio é bem lucrativo.

**Figura 28:** Três maiores e Três menores receitas

[283] # Descobrindo as 3 maiores receitas  
superLoja.nlargest(3, "Receita")

	ID_Pedido	Data_pedido	Data_envio	Nome_cliente	Seguimento	Cidade	Estado	Regiao	Categoria	SubCategoria	Vendas	Quantidade	Desconto	Lucro	Orçamento	Receita	Receita/Venda	
	2697	2698.0	2020-03-18	2018-03-23	Sean Miller	Escritório em casa	Jacksonville	Florida	Sul	Tecnologia	Máquinas	22638.48	6.0	0.5	-1811.0784	NaN	135830.88	6
	9039	9040.0	2022-12-17	2020-12-21	Adrian Barton	Consumidor	Detroit	Michigan	Central	Material de Escritório	Fichários	9892.74	13.0	0.0	4946.3700	NaN	128605.62	13
	6826	6827.0	2022-02-10	2020-09-10	Tamara Chand	Corporativo	Lafayette	Indiana	Central	Tecnologia	Copiadoras	17499.95	5.0	0.0	8399.9760	NaN	87499.75	5



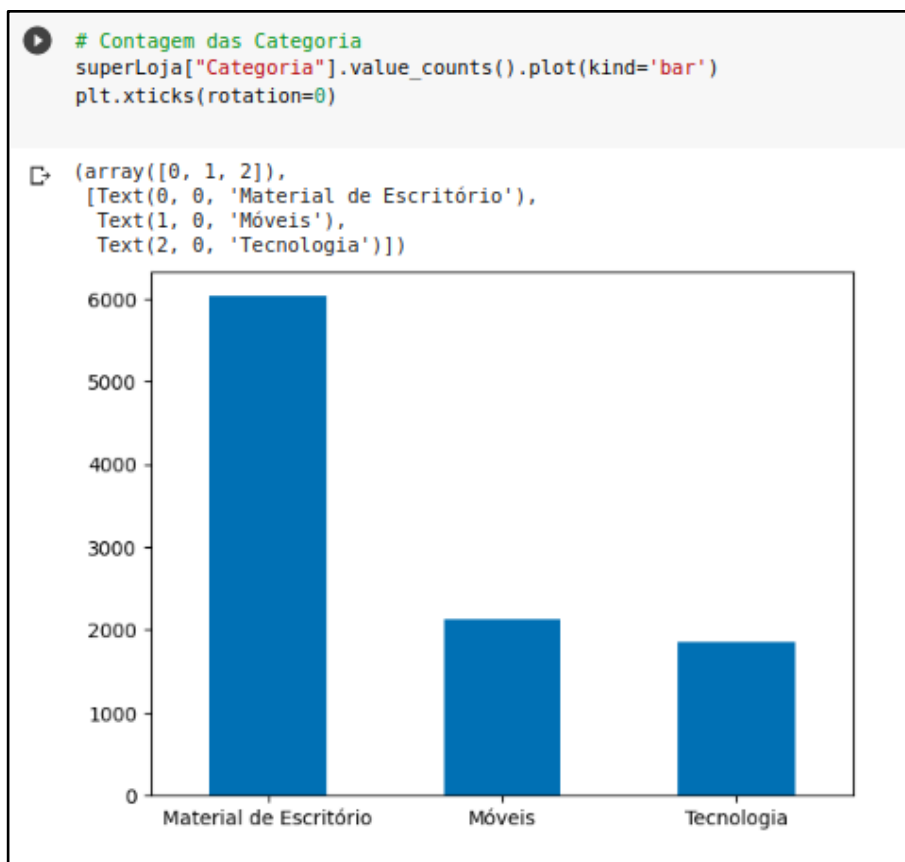
< | >

[284] # Descobrindo a 3 menores receitas  
superLoja.nsmallest(3, "Receita")

	ID_Pedido	Data_pedido	Data_envio	Nome_cliente	Seguimento	Cidade	Estado	Regiao	Categoria	SubCategoria	Vendas	Quantidade	Desconto	Lucro	Orçamento	Receita	Receita/Vendas	
	4101	4102.0	2023-06-19	2021-06-23	Zuschuss Carroll	Consumidor	Houston	Texas	Central	Material de Escritório	Eletrodomésticos	0.444	1.0	0.8	-1.1100	NaN	0.444	1.0
	9292	9293.0	2023-02-03	2021-02-03	Roland Schwarz	Corporativo	Waco	Texas	Central	Material de Escritório	Fichários	0.556	1.0	0.8	-0.9452	NaN	0.556	1.0
	8658	8659.0	2022-06-21	2020-06-25	Ken Brennan	Corporativo	Chicago	Illinois	Central	Material de Escritório	Fichários	0.836	1.0	0.8	-1.3376	NaN	0.836	1.0

Fonte: Autor

De acordo com o gráfico abaixo é possível verificar a contagem dos produtos vendidos na Super Loja.

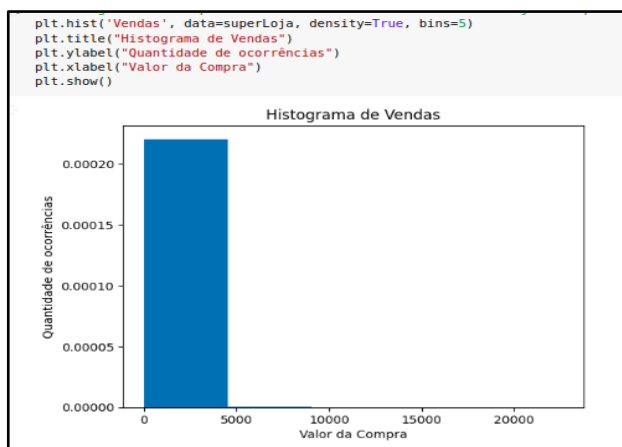


**Figura 28:**  
Contagem das categorias

Fonte: Autor

Nesse sentido, o grupo de material de escritório vem em primeiro lugar com o maior número de produtos, em seguida vem os móveis e por último os produtos de tecnologia.

**Figura 29: Histograma**

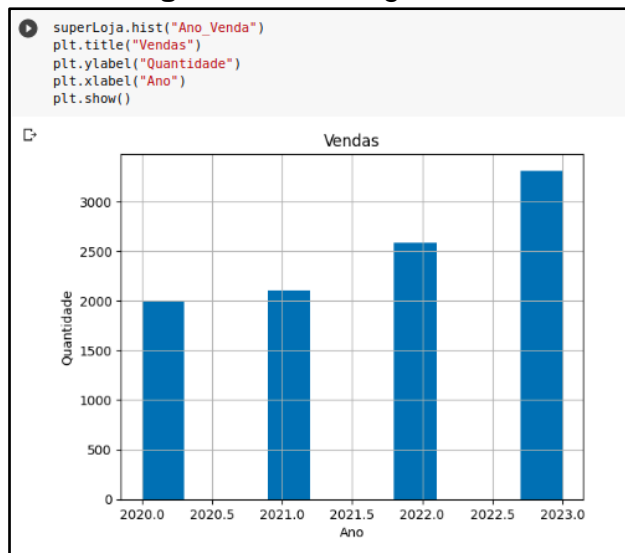


Fonte: Autor

O histograma está representando os valores absolutos no eixo y. Para plotar em termos de frequência relativa, acrescenta-se `density=True` dentro da função, de acordo com a figura 29.

De acordo com o histograma de vendas é possível verificar que as vendas foram crescentes de 2020 á 2023.

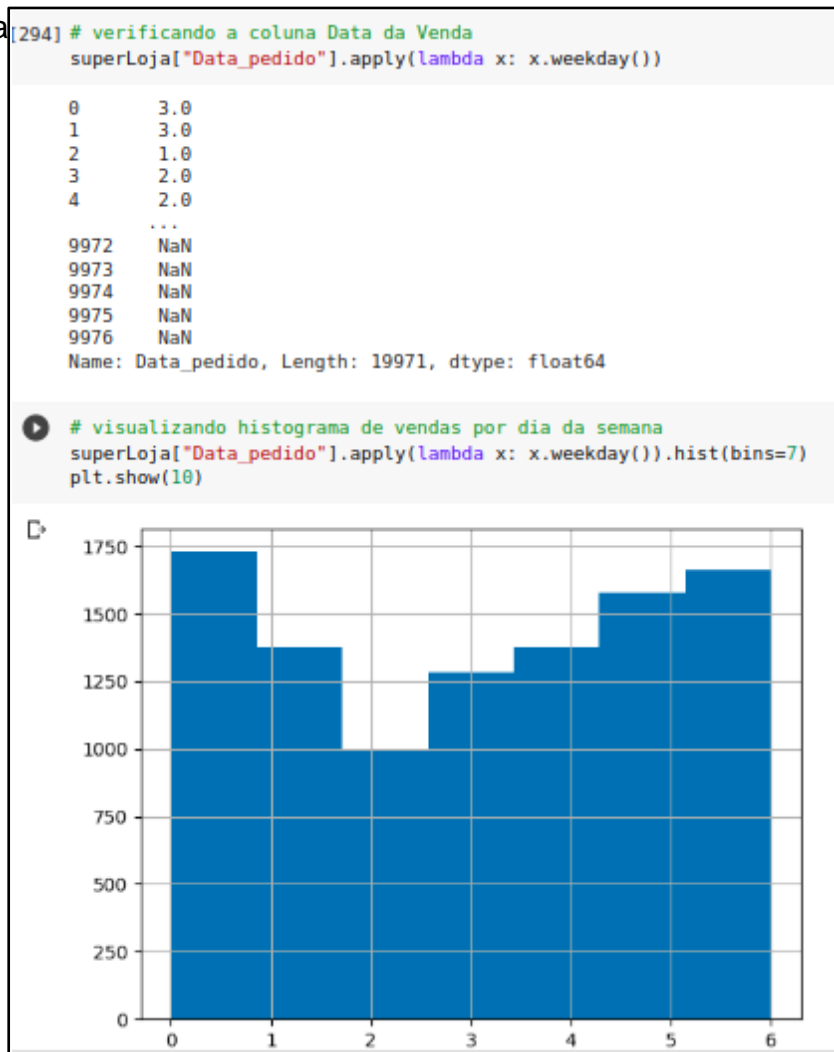
**Figura 30: Histograma de vendas**



Fonte: Autor

Observando o gráfico podemos perceber uma crescente movimentação do índice de vendas ao longo do tempo. Mas olhando para 2023 percebemos esse grande aumento, que está ligada ao período.

**Figura 31:** Histograma de vendas por dia da semana

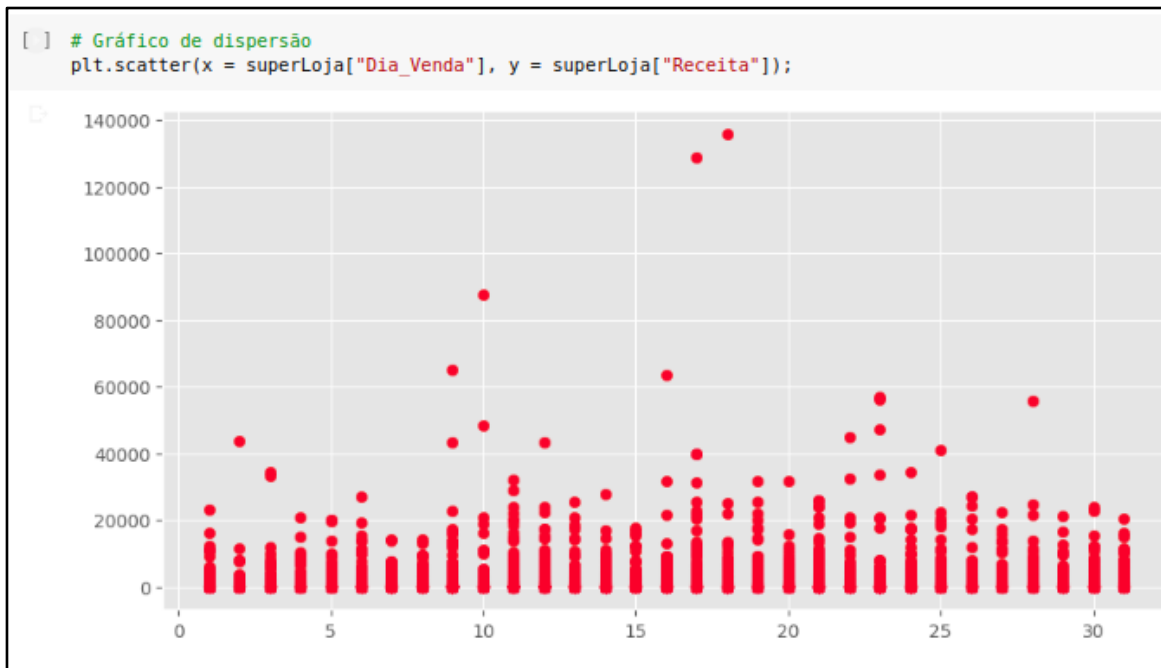


Fonte: Autor

Quando olhamos apenas os 7 dias da semana podemos ver que segue bastante a tendência apresentada, oscilando pouco.



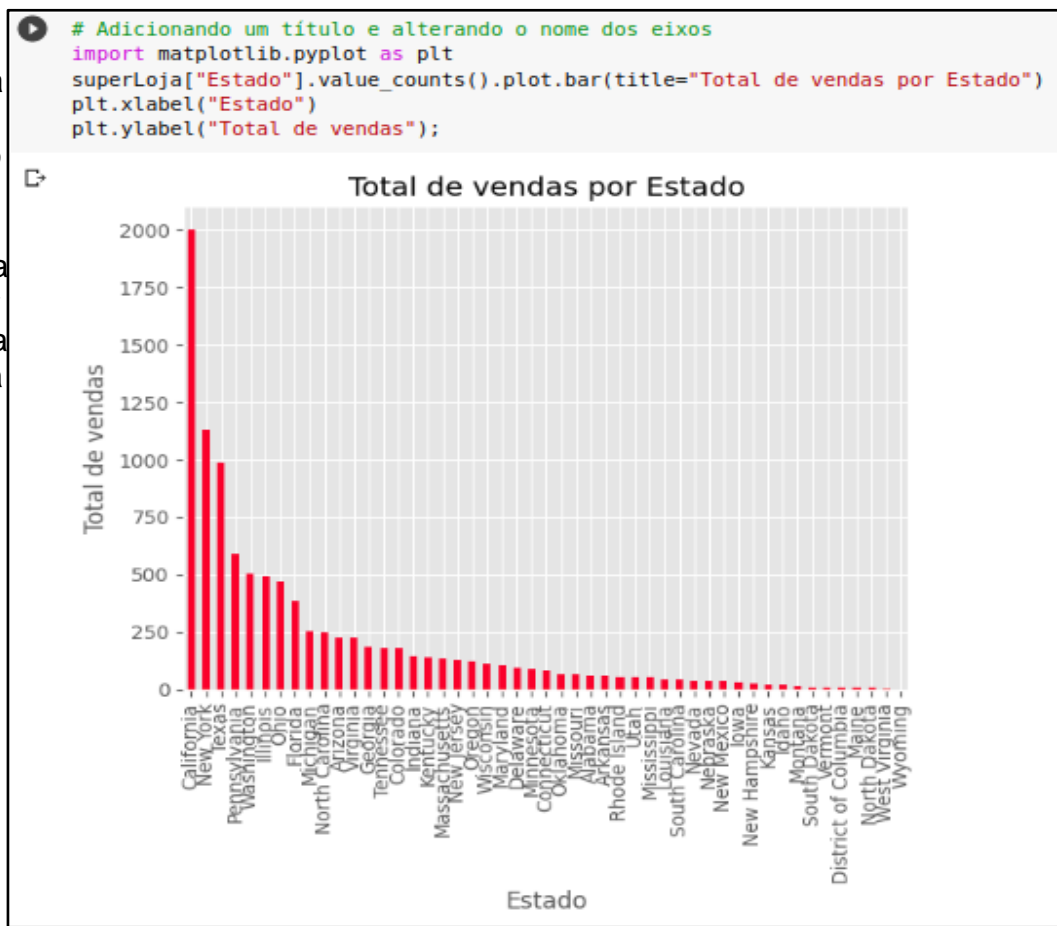
Já quando olhamos 30 dias, o equivalente ao mês completo, respectivamente, que é justamente o período em que as empresas fecham os resultados, podemos ver uma oscilação um pouco maior, com grande desempenho nas vendas. Ainda observando a janela de 30 dias, podemos ver que na quinzena do mês houve uma alta das vendas.



**Figura 32:** Histograma de vendas por dia da semana

Fonte: Autor

**Figura 33:** Histograma de vendas por dia da semana



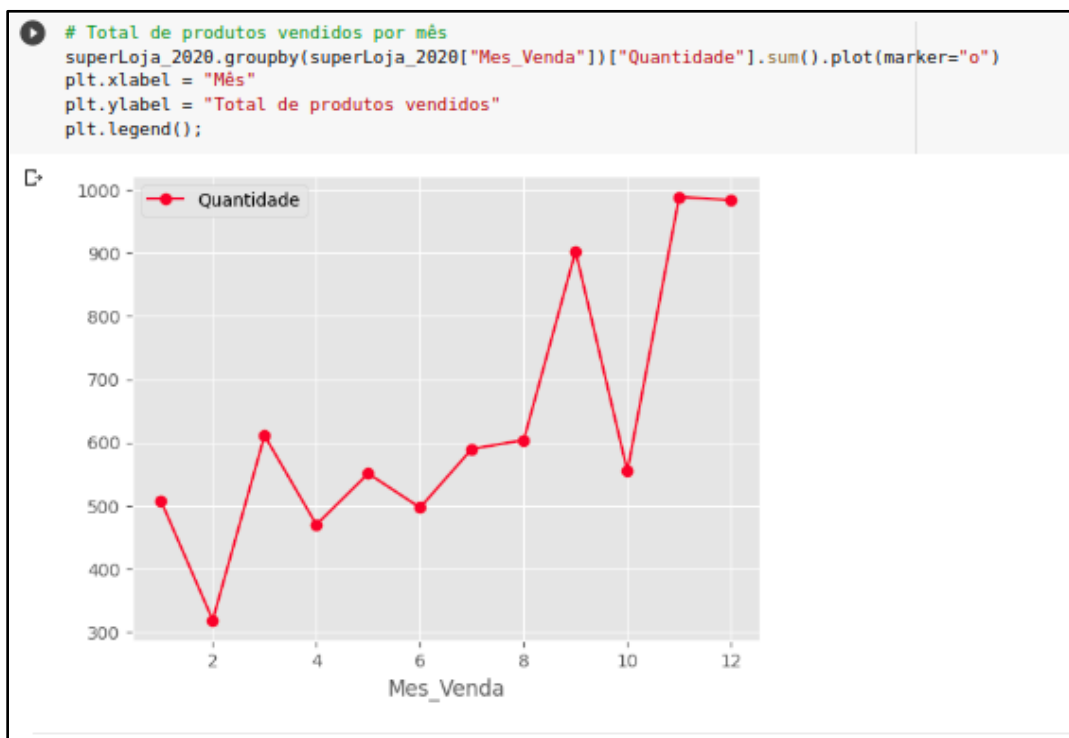
Fonte: Autor

Conforme o gráfico acima, os Estados com maior índice de vendas foram em disparada a Califórnia região Oeste, em seguida New York e Pensilvânia ambas da região Leste.

Segundo Silva (2009, p. 26) o crescimento acelerado do comércio eletrônico faz a internet alcançar diariamente praticamente todos os pontos do território nacional. Conforme ela se torna disponível nas cidades do interior, aqueles usuários (de locais remotos), começam a ter a possibilidade de comprar produtos que não estão disponíveis no local onde vivem.

A internet se converteu em parte da vida dos cidadãos modernos e, conforme o custo de acesso diminuiu, a quantidade de usuários aumentou. Desse modo as empresas possuem um mercado disponível de diferentes oportunidades de crescimento e de melhoria dos seus negócios. Praticamente não há barreiras para o alcance da internet e, com o aparecimento das redes sociais, os hábitos de consumo se transformaram rapidamente. Esse fato também gera uma consequência colateral e um risco que as empresas devem ter em mente: a internet tornou os consumidores volúveis em razão da quantidade de informação que recebe (BARRIENTOS, 2015, p. 19).

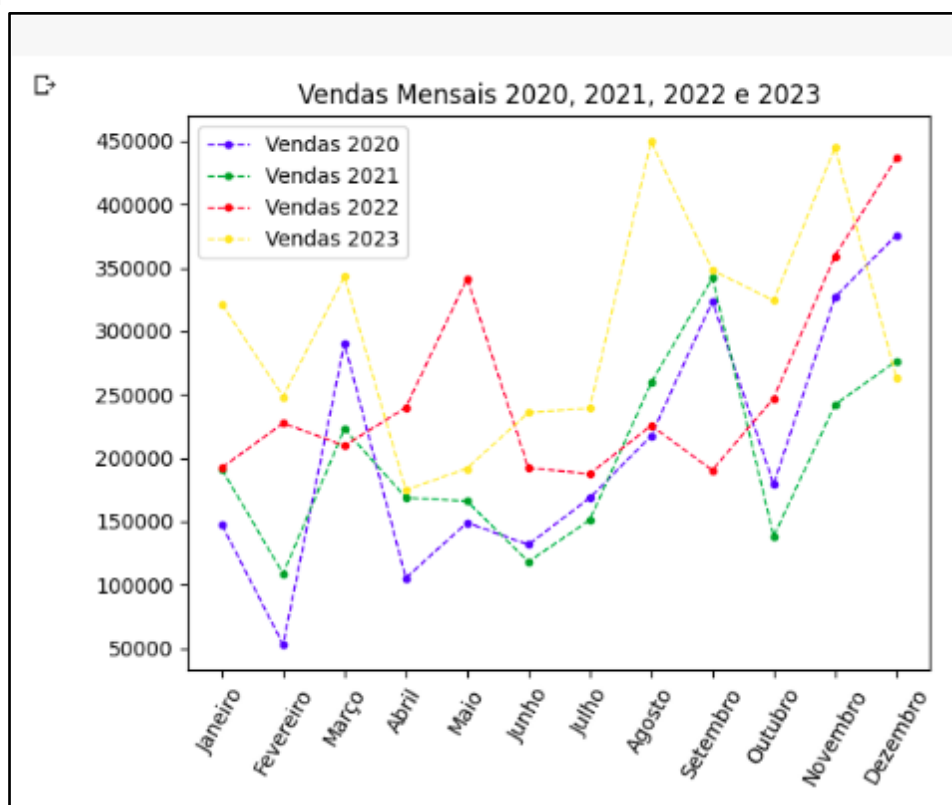
**Figura 34:** Histograma de vendas por dia da semana



Fonte: Autor

Considerando que o conjunto de dados possui registros de Janeiro a Dezembro de 2020, os três melhores meses de venda foram Setembro, Novembro e Dezembro.

**Figura 35:** Histograma de vendas por dia da semana



Fonte: Autor

As vendas do ano de 2020 apresentaram crescimento em março, agosto e um crescimento mais significativo em dezembro, 2021 começou o ano melhor que 2020, mantendo um aumento constante em cada mês, teve uma leve queda em junho, mas se recuperou no mês seguinte. Já no ano de 2022 apresentaram um crescimento significativo desde fevereiro; a força de vendas aumenta em maio e se mantém na média em constante crescimento até o final de dezembro. Para o mês de janeiro de 2023 atingindo 321 milhões, devido às condições normais de mercado, uma queda é evidente em fevereiro e se estabiliza no final do primeiro semestre do ano, de agosto a outubro a queda nas vendas retorna próxima às vendas do 2021, mas aumenta fortemente em novembro, com vendas recordes.

## 5. Modelo Canvas

Título: Super Loja		
<b>Problema</b> Analisar o dataset de vendas e investigar atributos da Super Loja que a tornam uma empresa com crescente receita.	<b>Resultados e Previsões</b> Avaliar os atributos da Super Loja, extrair informações e gerar insights que podem influenciar em estratégias	<b>Aquisição de Dados</b> Os dados de ambos os datasets superstore-sales-2023 e superstore-sales-budget-2023 foram coletados do site da data.world.
<b>Modelagem</b> Realizado análises no dataset coletado, tanto de forma gráfica quanto análise descritiva dos dados utilizando a biblioteca Pandas em Python.	<b>Avaliação do Modelo</b> Para avaliação dos resultados obtidos no modelo de vendas, foram avaliados o gráfico de vendas mensais, a tabela de vendas anuais e o histograma de vendas semais e diário, conforme o notebook em Python no diretório deste projeto.	<b>Preparação dos Dados</b> Após a união dos datasets, os dados foram tratados, as colunas foram renomeadas, os dados duplicados foram removidos e dados desnecessários para a análise também foram removidos.

6.  
Lin  
ks

T  
odos  
os  
códig  
os  
desen  
volvid  
os e a

documentação utilizada são disponibilizados no repositório do Github.

<https://github.com/mariliafigueiredo/Trabalho-Puc---BI>

## REFERÊNCIAS

BARRIENTOS, Pedro. *Marketing + internet = e-commerce: oportunidades desafios*. Artículo de investigación. **Finanz. polit. econ.**, ISSN: 2248-6046, Lima, Perú Vol. 9, No. 1, pp. 41-56, 2017. (acessado em 10 de abril de 2023).

SILVA, Leandro. **Aumente suas Vendas com E-commerce**. Rio de Janeiro: Editora Ciência Moderna Ltda, 2009.

TRANSFORMAÇÃO DIGITAL. **Quais as vantagens do big data em vendas?** Disponível em <http://www.mma.gov.br/sitio/index.php?ido=conteudo.monta&idEstrutura=18>. Acesso em: 21/0/2023.