UNIVERSIDADE PRESBITERIANA MACKENZIE
RAYAN CRHISTOFER GOMES DA SILVA (10408039)

ANÁLISE EXPLORATÓRIA DE DADOS DE UM SUPERMERCADO

SÃO PAULO
2024

SUMÁRIO

1. INTRODUÇÃO

A análise de dados é uma ferramenta essencial para a compreensão de padrões de consumo e para a formulação de estratégias empresariais mais eficazes. No contexto de supermercados, a capacidade de extrair insights a partir das vendas realizadas pode representar uma vantagem competitiva significativa, permitindo a identificação de tendências, otimização de recursos e

melhoria da experiência do cliente. Este projeto tem como objetivo realizar uma análise estatística e visual detalhada das transações realizadas em três filiais de um supermercado, durante os meses de janeiro, fevereiro e março de 2019.

Utilizando um dataset real e rico em informações, esta análise busca compreender o comportamento dos consumidores em termos de volume de vendas, métodos de pagamento, categorias de produtos e ticket médio por tipo de cliente. Além disso, pretende-se explorar a distribuição temporal das vendas, verificando os períodos de maior movimentação e os produtos mais rentáveis, com foco na identificação de padrões que possam subsidiar decisões estratégicas no ambiente empresarial.

A metodologia utilizada combina técnicas de limpeza e preparação de dados, modelagem estatística e visualizações gráficas, implementadas por meio de bibliotecas amplamente reconhecidas, como **Pandas**, **Matplotlib** e **Seaborn**. O estudo também se concentra na avaliação de métricas financeiras, como o cálculo da margem de lucro por produto e a análise do comportamento de compra de clientes fidelizados versus clientes normais.

Os resultados esperados incluem uma visão consolidada do desempenho de vendas por filial, uma análise das preferências dos consumidores e insights sobre a eficiência das estratégias de pagamento e fidelização. Este trabalho, ao integrar análise descritiva e inferencial, pretende servir como base para reflexões e tomadas de decisão em um setor altamente competitivo e dinâmico, como o varejo de supermercados.

2. PREMISSAS DO PROJETO

2.1 Apresentação dos dados

O conjunto de dados utilizado neste projeto contém informações detalhadas sobre transações comerciais realizadas em uma rede de filiais de uma empresa de varejo. As variáveis

presentes no dataset foram selecionadas para fornecer uma visão abrangente das compras realizadas, permitindo uma análise profunda do comportamento dos consumidores e do desempenho operacional das filiais. A seguir, são apresentadas as principais colunas do dataset, organizadas de acordo com suas características e relevância analítica:

- **Informações Temporais:**

- **data_compra**: Representa a data em que a transação foi realizada. Esta variável é essencial para a identificação de tendências temporais, como sazonalidade ou picos de vendas em determinados períodos.
- **hora_compra**: Indica o horário da transação, permitindo a análise de horários de maior movimento e identificação de padrões de compra ao longo do dia.

- **Dados sobre as Filiais:**

- **filial**: Identifica a unidade responsável pela transação, podendo ser uma das três filiais da empresa (A, B ou C). Essa variável é fundamental para analisar diferenças regionais ou operacionais entre as filiais, fornecendo insights sobre o desempenho comparativo das unidades.

- **Informações sobre os Produtos:**

- **categoria_produto**: Indica a categoria do produto adquirido, como alimentos, bebidas ou produtos de limpeza. Essa variável é crucial para compreender as preferências dos clientes por diferentes segmentos de produtos, o que pode impactar decisões estratégicas de marketing e estoque.

- **quantidade**: Representa o número de unidades adquiridas de um produto em uma transação. Essa informação é importante para avaliar volumes de vendas e entender a demanda por produtos específicos.

- **valor_total**: Refere-se ao valor total da transação em moeda local. É uma das variáveis mais relevantes, pois permite a análise do faturamento e do desempenho financeiro da empresa, além de possibilitar a avaliação do ticket médio de compra.

- **Dados do Cliente:**

- **faixa_etaria**: Classifica os clientes por intervalos de idade (ex.: 18-25, 26-35). Essa informação é útil para a análise de consumo e para a segmentação do mercado com base nas preferências de diferentes faixas etárias.
- **gênero**: Indica o gênero do cliente, categorizado como masculino ou feminino. Esta variável permite analisar padrões de consumo relacionados ao gênero, o que pode auxiliar em estratégias de marketing mais direcionadas.
- **frequencia_compras**: Mede a assiduidade do cliente em realizar compras durante o período analisado. Essa variável é essencial para identificar clientes regulares e ocasionais, permitindo segmentar os consumidores de acordo com seu nível de fidelidade e envolvimento com a marca.

- **Dados de Pagamento:**

- **forma_pagamento**: Especifica o método de pagamento utilizado em cada transação, como cartão de crédito, débito ou dinheiro. Essa variável é importante para entender as preferências dos consumidores em relação aos métodos de pagamento, além de fornecer insights sobre as práticas financeiras da empresa.

- **Informações de Fidelidade:**

- **participacao_fidelidade**: Indica se o cliente é participante de um programa de fidelidade da empresa (sim ou não). Esta variável é relevante para analisar a eficácia dos programas de fidelização no aumento de vendas e no fortalecimento do relacionamento com o cliente.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Invoice ID,Branch,City,Customer type,Gender,Product line,Unit price,Quantity,Tax 5%,Total,Date,Time,Payment,cogs,gross margin percentage,gross income,Rating														
2	750-67-8428,A,Yangon,Member,Female,Health and beauty,74.69,7,26.1415,548.9715,1/5/2019,13:08,Ewallet,522.83,4.761904762,26.1415,9.1														
3	226-31-3081,C,Naypyitaw,Normal,Female,Electronic accessories,15.28,5,3.82,80.22,3/8/2019,10:29,Cash,76.4,4.761904762,3.82,9.6														
4	631-41-3108,A,Yangon,Normal,Male,Home and lifestyle,46.33,7,16.2155,340.5255,3/3/2019,13:23,Credit card,324.31,4.761904762,16.2155,7.4														
5	123-19-1176,A,Yangon,Member,Male,Health and beauty,58.22,8,23.288,489.048,1/27/2019,20:33,Ewallet,465.76,4.761904762,23.288,8.4														
6	373-73-7910,A,Yangon,Normal,Male,Sports and travel,86.31,7,30.2085,634.3785,2/8/2019,10:37,Ewallet,604.17,4.761904762,30.2085,5.3														
7	699-14-3026,C,Naypyitaw,Normal,Male,Electronic accessories,85.39,7,29.8865,627.6165,3/25/2019,18:30,Ewallet,597.73,4.761904762,29.8865,4.1														
8	355-53-5943,A,Yangon,Member,Female,Electronic accessories,68.84,6,20.652,433.692,2/25/2019,14:36,Ewallet,413.04,4.761904762,20.652,5.8														
9	315-22-5665,C,Naypyitaw,Normal,Female,Home and lifestyle,73.56,10,36.78,772.38,2/24/2019,11:38,Ewallet,735.6,4.761904762,36.78,8														
10	665-32-9167,A,Yangon,Member,Female,Health and beauty,36.26,2,3.626,76.146,1/10/2019,17:15,Credit card,72.52,4.761904762,3.626,7.2														
11	692-92-5582,B,Mandalay,Member,Female,Food and beverages,54.84,3,8.226,172.746,2/20/2019,13:27,Credit card,164.52,4.761904762,8.226,5.9														
12	351-62-0822,B,Mandalay,Member,Female,Fashion accessories,14.48,4,2.896,60.816,2/6/2019,18:07,Ewallet,57.92,4.761904762,2.896,4.5														
13	529-56-3974,B,Mandalay,Member,Male,Electronic accessories,25.51,4,5.102,107.142,3/9/2019,17:03,Cash,102.04,4.761904762,5.102,6.8														
14	365-64-0515,A,Yangon,Normal,Female,Electronic accessories,46.95,5,11.7375,246.4875,2/12/2019,10:25,Ewallet,234.75,4.761904762,11.7375,7.1														
15	252-56-2699,A,Yangon,Normal,Male,Food and beverages,43.19,10,21.595,453.495,2/7/2019,16:48,Ewallet,431.9,4.761904762,21.595,8.2														
16	829-34-3910,A,Yangon,Normal,Female,Health and beauty,71.38,10,35.69,749.49,3/29/2019,19:21,Cash,713.8,4.761904762,35.69,5.7														
17	299-46-1805,B,Mandalay,Member,Female,Sports and travel,93.72,6,28.116,590.436,1/15/2019,16:19,Cash,562.32,4.761904762,28.116,4.5														
18	656-95-9349,A,Yangon,Member,Female,Health and beauty,68.93,7,24.1255,506.6355,3/11/2019,11:03,Credit card,482.51,4.761904762,24.1255,4.6														
19	765-26-6951,A,Yangon,Normal,Male,Sports and travel,72.61,6,21.783,457.443,1/1/2019,10:39,Credit card,435.66,4.761904762,21.783,6.9														
20	329-62-1586,A,Yangon,Normal,Male,Food and beverages,54.67,3,8.2005,172.2105,1/21/2019,18:00,Credit card,164.01,4.761904762,8.2005,8.6														
21	319-50-3348,B,Mandalay,Normal,Female,Home and lifestyle,40.3,2,4.03,84.63,3/11/2019,15:30,Ewallet,80.6,4.761904762,4.03,4.4														
22	300-71-4605,C,Naypyitaw,Member,Male,Electronic accessories,86.04,5,21.51,451.71,2/25/2019,11:24,Ewallet,430.2,4.761904762,21.51,4.8														
23	371-85-5789,B,Mandalay,Normal,Male,Health and beauty,87.98,3,13.197,277.137,3/5/2019,10:40,Ewallet,263.94,4.761904762,13.197,5.1														
24	273-16-6619,B,Mandalay,Normal,Male,Home and lifestyle,33.2,2,3.32,69.72,3/15/2019,12:20,Credit card,66.4,4.761904762,3.32,4.4														
25	636-48-8204,A,Yangon,Normal,Male,Electronic accessories,34.56,5,8.64,181.44,2/17/2019,11:15,Ewallet,172.8,4.761904762,8.64,9.9														
26	549-59-1358,A,Yangon,Member,Male,Sports and travel,88.63,3,13.2945,279.1845,3/2/2019,17:36,Ewallet,265.89,4.761904762,13.2945,6														
27	227-03-5010,A,Yangon,Member,Female,Home and lifestyle,52.59,8,21.036,441.756,3/22/2019,19:20,Credit card,420.72,4.761904762,21.036,8.5														
28	649-29-6775,B,Mandalay,Normal,Male,Fashion accessories,33.52,1,1.676,35.196,2/8/2019,15:31,Cash,33.52,4.761904762,1.676,6.7														
29	189-17-4241,A,Yangon,Normal,Female,Fashion accessories,87.67,2,8.767,184.107,3/10/2019,12:17,Credit card,175.34,4.761904762,8.767,7.7														
30	145-94-9061,B,Mandalay,Normal,Female,Food and beverages,88.36,5,22.09,463.89,1/25/2019,19:48,Cash,441.8,4.761904762,22.09,9.6														
31	848-62-7243,A,Yangon,Normal,Male,Health and beauty,24.89,9,11.2005,235.2105,3/15/2019,15:36,Cash,224.01,4.761904762,11.2005,7.4														
32	871-79-8483,B,Mandalay,Normal,Male,Fashion accessories,94.13,5,23.5325,494.1825,2/25/2019,19:39,Credit card,470.65,4.761904762,23.5325,4.8														
33	149-71-6266,B,Mandalay,Member,Male,Sports and travel,78.07,9,35.1315,737.7615,1/28/2019,12:43,Cash,702.63,4.761904762,35.1315,4.5														
34	640-49-2076,B,Mandalay,Normal,Male,Sports and travel,83.78,8,33.512,703.752,1/10/2019,14:49,Cash,670.24,4.761904762,33.512,5.1														
35	595-11-5460,A,Yangon,Normal,Male,Health and beauty,96.58,2,9.658,202.818,3/15/2019,10:12,Credit card,193.16,4.761904762,9.658,5.1														

Figura 1: Visualização do dataset ”supermarket_sales.csv”.

Essas colunas fornecem uma base robusta de informações que possibilitam a realização de análises detalhadas sobre o comportamento do consumidor, as operações das filiais, e o impacto de fatores como horário de compra, características demográficas e formas de pagamento nas vendas. A riqueza e diversidade dos dados tornam este conjunto adequado para a aplicação de técnicas analíticas avançadas, como modelagem preditiva e segmentação de mercado.

2.2 Objetivos e metas

O objetivo principal deste projeto é explorar os dados de vendas da empresa para obter uma compreensão mais profunda dos padrões de consumo, a fim de identificar insights que possam otimizar as operações comerciais e melhorar as estratégias de vendas. Em primeiro lugar, será realizada uma análise detalhada das vendas por filial, categoria de produto e método de

pagamento, visando entender como esses fatores influenciam o comportamento de compra dos consumidores. A partir dessa exploração, o projeto buscará identificar o ticket médio por tipo de cliente, permitindo uma melhor segmentação e personalização das ofertas.

Além disso, será realizada uma análise das margens de lucro de cada linha de produto, de modo a avaliar quais itens apresentam maior rentabilidade e quais podem necessitar de ajustes em sua estratégia de precificação ou promoção. Outro ponto importante será a investigação das tendências temporais, com foco nos dias e horários de maior movimento nas vendas, o que permitirá à empresa ajustar suas estratégias de marketing e operações de estoque de maneira mais eficaz. Por fim, todos os resultados serão visualizados de forma clara e intuitiva, utilizando gráficos e dashboards que facilitem a compreensão das principais descobertas, tornando as informações acessíveis para decisões estratégicas rápidas e informadas.

3. DEFINIÇÃO DO PRODUTO ANALÍTICO

O produto analítico desenvolvido neste projeto visa fornecer insights valiosos sobre os padrões de consumo e desempenho das vendas de uma empresa. Esse produto será constituído por uma série de análises que buscam entender o comportamento de compra dos consumidores, a rentabilidade de cada linha de produto, as variações nas vendas em função do tempo e as preferências de pagamento.

3.1 Bibliotecas e repositórios

Para a execução deste projeto, foram selecionadas diversas bibliotecas do ecossistema Python, visando uma análise eficiente e detalhada dos dados:

- **Pandas**: Biblioteca essencial para manipulação e análise de dados, utilizada para carregamento, limpeza e preparação dos dados.
- **NumPy**: Utilizada para operações matemáticas e de cálculo em larga escala, como agregações e cálculos de métricas como ticket médio e margem de lucro.
- **Matplotlib e Seaborn**: Bibliotecas de visualização gráfica, utilizadas para criar gráficos informativos e intuitivos, como gráficos de barras, linhas e pizza.

- **Scikit-learn**: Embora este projeto não envolva modelos preditivos, essa biblioteca pode ser útil para análises futuras, como segmentação de clientes com base em clustering ou construção de modelos de previsão.
- **Plotly**: Usada para a criação de visualizações interativas, permitindo ao usuário explorar os resultados de forma dinâmica.

3.2 Definição da base de dados e análise exploratória de dados

A base de dados utilizada neste projeto consiste em um conjunto de registros de vendas da empresa, que inclui informações detalhadas sobre cada transação realizada. Essas informações abrangem dados de clientes, produtos vendidos, valores de venda, métodos de pagamento, além de informações temporais relacionadas à data e hora das transações. A base de dados permite realizar análises multifacetadas, desde a segmentação dos consumidores até a análise de rentabilidade de produtos e tendências de consumo.

A análise exploratória dos dados (AED) tem como objetivo a compreensão inicial da base de dados, identificando padrões, outliers e relacionamentos importantes entre as variáveis. O processo de AED incluiu as seguintes etapas:

1. **Análise Descritiva**: Cálculo de métricas como média, mediana, desvio padrão e quantis, para entender a distribuição das variáveis.
2. **Visualização de Dados**: Gráficos e tabelas foram criados para observar a distribuição das vendas por filial, categoria de produto e método de pagamento, entre outros.
3. **Identificação de Dados Faltantes e Outliers**: Foi realizada uma verificação detalhada de dados ausentes, e as estratégias de imputação foram aplicadas, quando necessário. Outliers também foram identificados e analisados para garantir que não interferissem nas análises.
4. **Tratamento de Dados**: Além da imputação de valores ausentes, foi realizado um processo de normalização e transformação de variáveis, como a conversão de datas para um formato mais conveniente para análise temporal.
5. **Criação de Novas Variáveis**: A partir da análise exploratória, novas variáveis foram criadas, como a identificação do ticket médio por cliente e a margem de lucro por produto, utilizando cálculos simples e funções agregadas.

3.3 Tratamento da base de dados (Preparação e treinamento)

O tratamento da base de dados foi uma etapa essencial para garantir a qualidade e a consistência dos dados, visando otimizar a análise e a obtenção de insights relevantes. Inicialmente, foram identificados e tratados os dados faltantes por meio da imputação, utilizando a média ou mediana para variáveis numéricas e a moda ou categorias específicas para variáveis categóricas. A remoção de duplicatas também foi realizada para eliminar registros redundantes que poderiam distorcer os resultados.

As variáveis temporais, como data e hora das transações, foram transformadas para o formato datetime, permitindo a extração de atributos adicionais, como dia da semana, mês, ano e hora do dia, facilitando a análise de padrões sazonais e de pico de vendas. Além disso, novas variáveis derivadas, como o ticket médio por cliente e a margem de lucro por produto, foram criadas para agregar valor à análise de rentabilidade e comportamento de compra.

O tratamento de variáveis categóricas foi feito por meio de **One-Hot Encoding** e **Label Encoding**, conforme necessário, para garantir a compatibilidade dos dados com os modelos analíticos. A detecção de outliers foi realizada para garantir que valores extremos não interferissem nas conclusões, e a consistência dos dados foi verificada para garantir que as variáveis estivessem dentro dos limites esperados. Essas ações de limpeza e transformação permitiram a preparação adequada da base de dados para análises exploratórias e de tendências, fundamentando o processo decisional da empresa.

4. ANÁLISE EXPLORATÓRIA DE DADOS COM PYTHON

4.1 Exploração e tratamento de dados

Nesta seção, aplicamos métodos estatísticos e de visualização para investigar as principais questões levantadas na análise dos dados de vendas. A modelagem estatística e a análise exploratória tiveram como objetivo identificar padrões de comportamento e verificar hipóteses com base em aspectos específicos das colunas do dataset, como filial, cidade, tipo de cliente, linha de produto, método de pagamento, análise temporal e lucro bruto.

A seguir, detalhamos a manipulação e visualização dos dados, além de apresentarmos as principais análises realizadas. Neste trecho de código, realizamos a exploração inicial do dataset com o objetivo de obter uma visão geral de sua estrutura e das características principais dos dados:

Estrutura do DataFrame (df.info()): O comando df.info() fornece informações essenciais sobre o DataFrame, incluindo o número de registros, os nomes e tipos de cada coluna, além de indicar a presença de valores ausentes. Essa etapa é crucial para identificar potenciais problemas, como colunas com dados faltantes ou tipos de dados inadequados para análises posteriores.

Estatísticas Descritivas (df.describe()): O comando df.describe() gera estatísticas descritivas das variáveis numéricas, apresentando informações como média, desvio padrão, valores mínimo e máximo, além dos percentis. Esta análise é útil para compreender a distribuição dos dados, identificar padrões e detectar possíveis outliers ou valores atípicos.

Ambos os comandos são fundamentais para a fase de exploração inicial, pois proporcionam uma visão geral do dataset, permitindo a identificação de possíveis inconsistências e ajudando a guiar as etapas seguintes de análise e modelagem.

```
[11] # Carregar os dados do arquivo CSV no DataFrame
     df = pd.read_csv('supermarket_sales.csv')

     # Exibir as primeiras linhas para confirmar que o DataFrame foi carregado corretamente
     print(df.head())

     # Informações gerais sobre o DataFrame
     print(df.info())

     # Estatísticas descritivas
     print(df.describe())
```

Figura 2: carregando dados do Dataframe.

```
    Invoice ID Branch       City Customer type  Gender  \
0  750-67-8428      A     Yangon        Member  Female
1  226-31-3081      C  Naypyitaw        Normal  Female
2  631-41-3108      A     Yangon        Normal    Male
3  123-19-1176      A     Yangon        Member    Male
4  373-73-7910      A     Yangon        Normal    Male

             Product line  Unit price  Quantity   Tax 5%     Total       Date  \
0       Health and beauty       74.69         7  26.1415  548.9715   1/5/2019
1  Electronic accessories       15.28         5   3.8200   80.2200   3/8/2019
2      Home and lifestyle       46.33         7  16.2155  340.5255   3/3/2019
3       Health and beauty       58.22         8  23.2880  489.0480  1/27/2019
4       Sports and travel       86.31         7  30.2085  634.3785   2/8/2019

    Time      Payment    cogs  gross margin percentage  gross income  Rating
0  13:08      Ewallet  522.83                 4.761905       26.1415     9.1
1  10:29         Cash   76.40                 4.761905        3.8200     9.6
2  13:23  Credit card  324.31                 4.761905       16.2155     7.4
3  20:33      Ewallet  465.76                 4.761905       23.2880     8.4
4  10:37      Ewallet  604.17                 4.761905       30.2085     5.3
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 17 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Invoice ID               1000 non-null   object
 1   Branch                   1000 non-null   object
 2   City                     1000 non-null   object
 3   Customer type            1000 non-null   object
 4   Gender                   1000 non-null   object
 5   Product line             1000 non-null   object
 6   Unit price               1000 non-null   float64
 7   Quantity                 1000 non-null   int64
 8   Tax 5%                   1000 non-null   float64
 9   Total                    1000 non-null   float64
 10  Date                     1000 non-null   object
 11  Time                     1000 non-null   object
 12  Payment                  1000 non-null   object
 13  cogs                     1000 non-null   float64
 14  gross margin percentage  1000 non-null   float64
 15  gross income             1000 non-null   float64
 16  Rating                   1000 non-null   float64
```

Figura 3: Informações gerais sobre o DataFrame.

Para assegurar a qualidade e integridade dos dados, foram realizadas algumas etapas essenciais de limpeza:

Remoção de Dados Faltantes: Utilizamos o comando dropna() para excluir qualquer linha que contivesse valores ausentes no DataFrame. Esse procedimento garante que as análises subsequentes sejam feitas com dados completos e consistentes.

Remoção de Duplicatas:

Inicialmente, removemos duplicatas com base na coluna **Invoice ID**, pois cada transação deve ser única, evitando que o mesmo registro de venda seja contado mais de uma vez. Em seguida, foi realizada uma verificação adicional para assegurar que não restassem duplicatas completas no DataFrame.

Essas etapas são cruciais para prevenir distorções nos resultados da análise e assegurar que o dataset reflita com precisão as transações reais de vendas.

```
[12] # Remover dados faltantes
     df.dropna(inplace=True)

     # Remover duplicatas com base no Invoice ID
     df.drop_duplicates(subset='Invoice ID', inplace=True)

     # Remover duplicatas
     df.drop_duplicates(inplace=True)
```

Figura 4: Tratamento dos dados.

Agora, realizaremos um ajuste no formato de data, convertendo a coluna original para o tipo **datetime**, a fim de padronizar e facilitar a manipulação temporal dos dados. Além disso, criaremos uma nova coluna com o formato de data no padrão **pt-br** (dia/mês/ano), visando tornar as informações mais acessíveis e compreensíveis para usuários que preferem esse formato. Este ajuste é fundamental para garantir que a análise temporal dos dados seja realizada de maneira consistente e que as datas sejam apresentadas de forma intuitiva, especialmente em relatórios e visualizações.

```
[13] # Converter a coluna 'Date' para o tipo datetime
     df['Date'] = pd.to_datetime(df['Date'], format='%m/%d/%Y')

     # Criar a coluna com o formato '05/01/2019' (pt-BR)
     df['Date_pt_br'] = df['Date'].dt.strftime('%d/%m/%Y')

     # Criar a coluna com o formato '01-05-2019'
     df['Date'] = df['Date'].dt.strftime('%m-%d-%Y')

     # Exibir o DataFrame resultante
     df.head()
```

Figura 5: Conversão da coluna ”Date” para o tipo datatime.

4.2 Modelagem dos dados

Nesta seção, utilizamos métodos estatísticos para investigar as questões-chave levantadas durante a análise dos dados de vendas. A modelagem estatística e a análise exploratória têm como objetivo identificar padrões de comportamento e validar hipóteses com base em variáveis específicas do dataset, como:

- **Filial e Cidade**: Avaliamos o volume de vendas por filial e cidade para identificar as regiões com melhor desempenho.
- **Tipo de Cliente**: Comparamos os clientes "Member" e "Normal" para entender qual grupo apresenta maior ticket médio e frequência de compras.
- **Linha de Produto**: Investigamos as categorias de produtos mais populares e realizamos uma comparação do ticket médio entre as diferentes linhas de produtos.
- **Método de Pagamento**: Analisamos os métodos de pagamento mais utilizados e verificamos sua relação com o valor total das compras.
- **Análise Temporal**: Identificamos padrões de vendas em determinados dias e horários de pico, buscando entender comportamentos sazonais e tendências temporais.
- **Lucro Bruto**: Calculamos a lucratividade média por produto e categoria para avaliar o desempenho financeiro das transações.

Na **Análise de Vendas por Filial**, utilizamos a função groupby() para agrupar os dados pela coluna "Branch", que representa as filiais A, B e C. Em seguida, aplicamos a função sum() sobre a coluna "Total" para calcular o valor total das vendas em cada filial. Essa abordagem permite a comparação do desempenho entre as filiais, destacando qual delas obteve o maior volume de vendas no período analisado.

Esta análise é crucial para identificar pontos fortes e fracos entre as unidades, auxiliando na formulação de estratégias de negócios.

```
[14] valor_total_vendas_filial = df.groupby('Branch')['Total'].sum()
     print(valor_total_vendas_filial)

     Branch
     A    106200.3705
     B    106197.6720
     C    110568.7065
     Name: Total, dtype: float64
```

Figura 6: Análise de vendas por filial.

Observa-se um equilíbrio no valor das vendas entre as três filiais, com exceção da filial C, que apresenta uma diferença de aproximadamente 4 mil unidades monetárias a mais em vendas quando comparada às filiais A e B. Essa leve discrepância pode ser indicativa de fatores específicos, como estratégias de marketing, localização geográfica ou tipos de produtos vendidos, que influenciam diretamente o volume de vendas em cada filial. A análise detalhada dessa diferença pode fornecer insights valiosos sobre o desempenho regional e ajudar a direcionar estratégias de vendas mais eficazes para otimizar os resultados em todas as filiais.

```
[15] media_vendas_cliente = df.groupby('Customer type')['Total'].mean()
     print(media_vendas_cliente)

     Customer type
     Member    327.791305
     Normal    318.122856
     Name: Total, dtype: float64
```

Figura 7: Média de vendas.

Ao analisar a média de vendas por tipo de cliente, observa-se que os clientes classificados como "Membros" apresentam um valor médio de compras superior em relação aos clientes "Normais". Esse comportamento sugere que o programa de membros, provavelmente relacionado a um sistema de fidelidade ou benefícios exclusivos como descontos, pode estar incentivando os clientes a realizarem compras de maior valor.

A diferença no ticket médio entre os dois grupos pode refletir a eficácia do programa de fidelidade em aumentar o volume de compras, fidelizar clientes e promover um relacionamento mais duradouro com a marca.

```
[18] # Importar a biblioteca seaborn e matplotlib
     import seaborn as sns
     import matplotlib.pyplot as plt

     # Certifique-se de que o DataFrame 'df' foi carregado corretamente e contém a coluna 'Payment'

     # Criar o gráfico de barras para a distribuição dos métodos de pagamento
     sns.countplot(x='Payment', data=df)

     # Adicionar título ao gráfico
     plt.title('Distribuição de Métodos de Pagamento')

     # Exibir o gráfico
     plt.show()
```

Figura 8: Gerando um gráfico da distruibuição de métodos de pagamento.

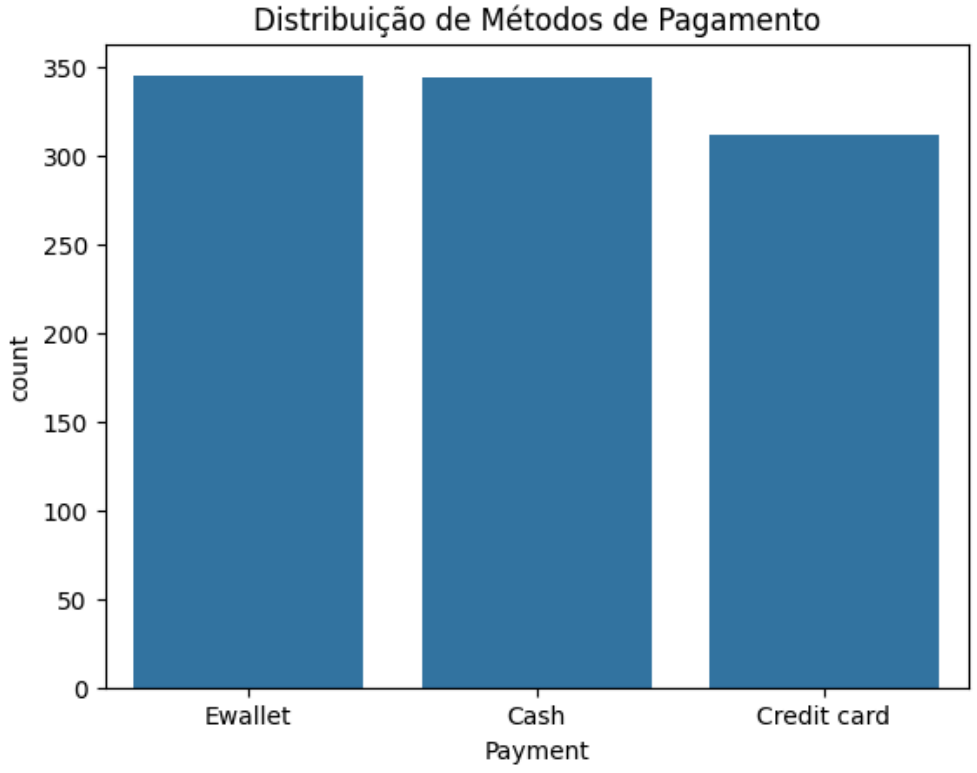

Figura 9: Gráfico “Distribuição de Métodos de Pagamento”

Ao analisar a distribuição dos métodos de pagamento, é possível observar um equilíbrio notável entre as opções disponíveis. Embora o cartão de crédito tenha registrado o menor número de pedidos, a diferença em relação aos outros métodos de pagamento é mínima. Isso sugere que, embora o cartão de crédito seja menos utilizado em termos de volume de transações, ele ainda representa uma parte significativa do total de vendas. Essa leve disparidade pode ser atribuída a uma série de fatores, como preferências pessoais dos clientes, facilidades oferecidas por outros métodos de pagamento (como boletos ou carteiras digitais), ou até mesmo políticas de incentivos, como descontos em pagamentos à vista. Para uma análise mais detalhada, seria interessante verificar a correlação entre o método de pagamento e o valor total das compras, já que métodos como cartão de crédito podem ter uma associação mais forte com compras de maior valor devido à possibilidade de parcelamento.

```
[19] valor_venda_media_produto = df.groupby('Product line')['Total'].mean().sort_values(ascending=False)
     print(valor_venda_media_produto)

Product line
Home and lifestyle        336.636956
Sports and travel         332.065220
Health and beauty         323.643020
Food and beverages        322.671517
Electronic accessories    319.632538
Fashion accessories       305.089298
Name: Total, dtype: float64
```

Figura 10: Valor de venda média por produto.

A análise da média de vendas por categoria de produtos revela um equilíbrio notável entre as diferentes categorias. No entanto, os produtos da categoria "Casa e Bem-estar" se destacam com a maior média de vendas, indicando uma tendência favorável de consumo dentro desse segmento. Essa categoria pode estar associada a itens de maior valor agregado ou que atendem a necessidades contínuas dos consumidores, como móveis, decoração, utensílios domésticos e produtos de cuidado pessoal.

Esse resultado sugere que, embora haja um equilíbrio geral nas vendas das demais categorias, a de "Casa e Bem-estar" demonstra um potencial de consumo superior, possivelmente refletindo comportamentos de compra sazonais ou uma forte demanda por produtos que melhoram a qualidade de vida no lar. Para uma compreensão mais profunda, seria interessante investigar a correlação entre o ticket médio dessa categoria e as características demográficas dos consumidores, como faixa etária ou localização geográfica, além de observar eventuais promoções ou tendências de mercado que possam ter influenciado esse desempenho.

```
[20] lucro_medio_produto = df.groupby('Product line')['gross income'].mean().sort_values(ascending=False)
     print(lucro_medio_produto)

     Product line
     Home and lifestyle        16.030331
     Sports and travel         15.812630
     Health and beauty         15.411572
     Food and beverages        15.365310
     Electronic accessories    15.220597
     Fashion accessories       14.528062
     Name: gross income, dtype: float64
```

Figura 11: Lucro médio por produto.

Na etapa anterior, definimos a coluna **Date** como o índice do DataFrame, facilitando a análise temporal dos dados. Esse ajuste permite que a data seja utilizada de forma mais eficaz, especialmente em gráficos que exigem uma linha do tempo clara, como aqueles que visualizam tendências ao longo de períodos específicos.

Agora, iremos realizar uma análise do comportamento das vendas ao longo do tempo, considerando uma distribuição semanal. Essa abordagem nos permitirá observar padrões sazonais, identificar picos de vendas em determinadas semanas e comparar o desempenho em diferentes períodos. A segmentação por semana também pode revelar influências externas, como promoções semanais, feriados ou campanhas de marketing, que podem impactar o volume de vendas de maneira significativa.

Ao agrupar as vendas por semana, será possível obter uma visão mais granular das flutuações no comportamento de compra dos consumidores, facilitando a identificação de tendências de longo prazo ou anomalias pontuais. Essa análise será crucial para o planejamento estratégico de vendas e para a otimização de campanhas futuras.

```
[21] # Converter a coluna 'Date' para datetime
     df['Date'] = pd.to_datetime(df['Date'], format='%m-%d-%Y')

     # Definir a coluna 'Date' como índice para facilitar o agrupamento semanal
     df.set_index('Date', inplace=True)
```

Figura 12: análise do comportamento das vendas.

```
[22] # Agrupar a quantidade de vendas por semana
     qtd_vendas_por_semana = df['Total'].resample('W').count()

     # Plotar o total de vendas semanais
     qtd_vendas_por_semana.plot(figsize=(10, 5), title="Total de Vendas Semanais")
     plt.xlabel("Data")
     plt.ylabel("Vendas Totais")
     plt.show()
```

Figura 13: Agrupamento da quantidade de vendas por semana.

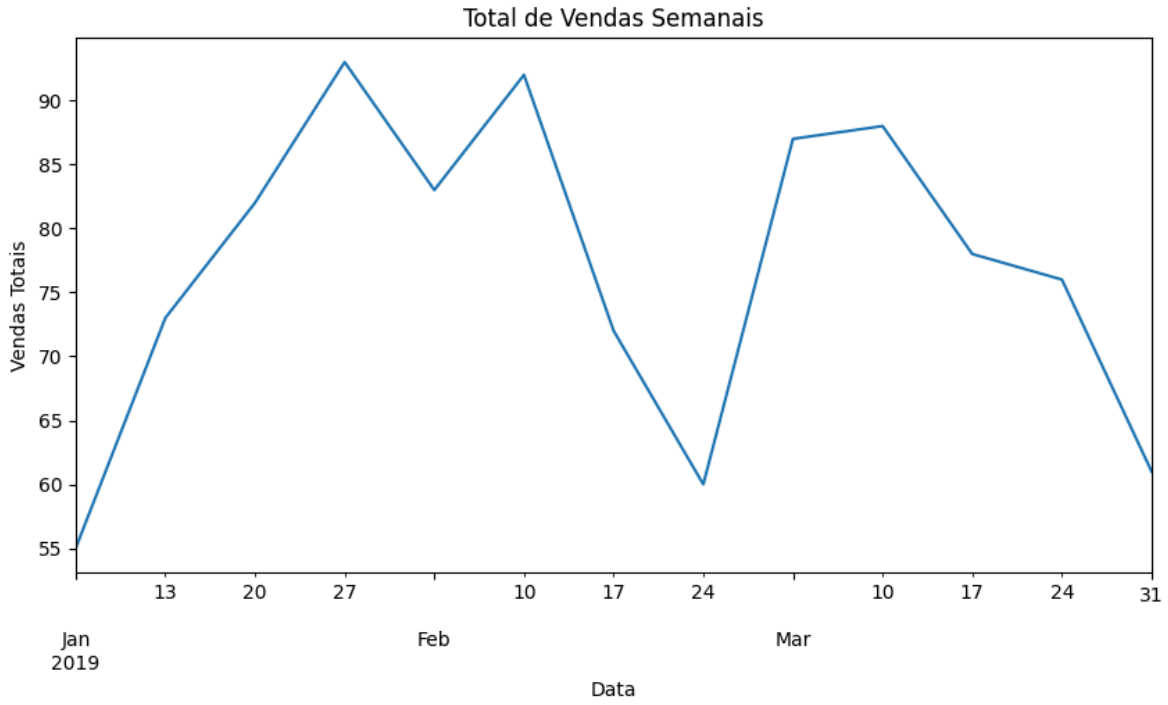

Figura 14: Gráfico “Total de Vendas Semanais”.

Ao analisar o primeiro gráfico, podemos observar um claro pico de vendas entre os dias 20 e 27 de janeiro, seguido por uma leve diminuição no início de fevereiro. Esse pico pode ser atribuído a fatores como o retorno das compras após o período de festas de fim de ano e

promoções de início de ano, o que leva a um aumento significativo na demanda. Além disso, há uma queda nas vendas nas últimas semanas de fevereiro e de março, o que sugere que os consumidores tendem a concentrar suas compras nas primeiras semanas de cada mês. Essa diminuição nas vendas nas últimas semanas pode estar relacionada a uma diminuição no poder de compra após o pagamento de contas e a estabilidade das finanças mensais.

O comportamento atípico observado em janeiro, com um volume de vendas elevado ao longo de todo o mês, é um reflexo de fatores sazonais, como o período de férias e o início do ano. Durante esse período, muitos consumidores têm mais tempo disponível para realizar compras, seja por lazer ou pela necessidade de adquirir produtos após o término de festividades. Além disso, o início do ano é um momento propício para consumidores que buscam novas ofertas e descontos pós-festas, o que impulsiona o volume de vendas. Outro fator que pode influenciar esse aumento é o planejamento de consumo para o ano, quando as pessoas aproveitam as promoções e fazem compras para o início do ciclo.

Esse tipo de análise temporal ajuda a identificar tendências sazonais e pode ser utilizado para planejar estratégias de marketing, campanhas promocionais e gerenciamento de estoque, garantindo que a empresa esteja preparada para picos de demanda e possa ajustar suas operações de forma eficiente.

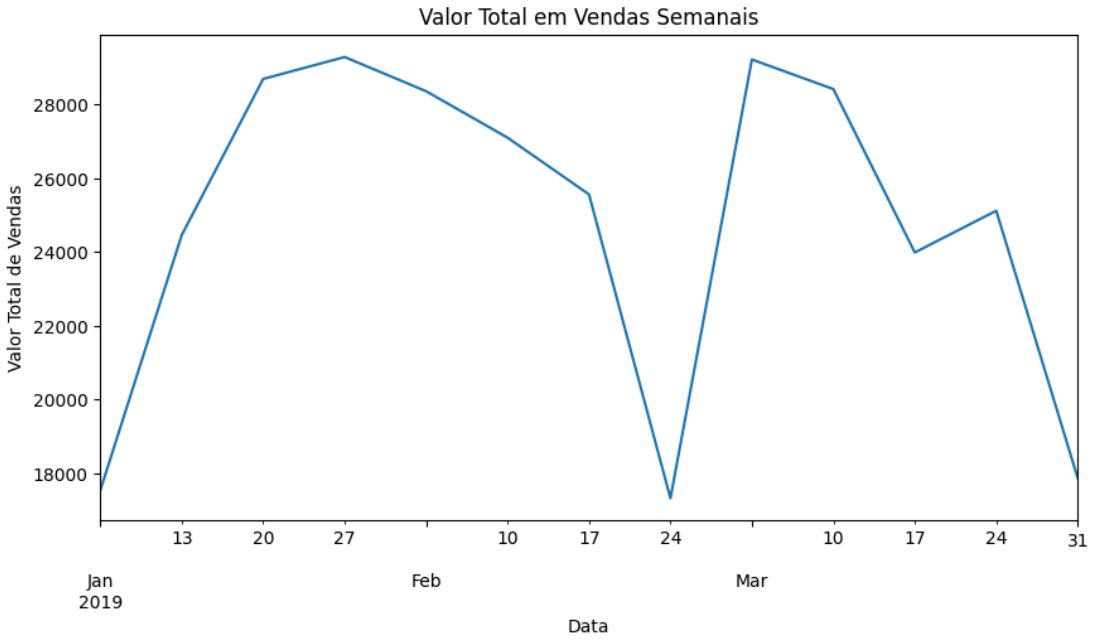

Figura 15: Gráfico “Valor Total em Vendas Semanais”.

Ao analisar o segundo gráfico, que apresenta o valor das vendas em períodos regulares, é possível identificar um padrão bastante semelhante ao observado no gráfico de quantidade de vendas. Esse comportamento reforça a ideia de que os consumidores tendem a concentrar suas compras nas primeiras semanas de cada mês, o que resulta em um aumento significativo no volume de vendas e, consequentemente, no valor total das transações durante esses períodos.

Esse padrão de concentração de compras nas primeiras semanas pode ser explicado por diversos fatores, como o recebimento de salários no início do mês, o que proporciona aos consumidores maior poder de compra. Além disso, a necessidade de abastecer a despensa para o mês e aproveitar ofertas promocionais no começo do ciclo mensal são fatores que contribuem para esse comportamento. Esse aumento nas vendas nas primeiras semanas é refletido também no valor total, indicando que o pico de compras não se limita apenas à quantidade de itens adquiridos, mas também está associado a transações de maior valor.

Para os supermercados e estabelecimentos de varejo, esse padrão de compras oferece insights valiosos para o planejamento de estoque, estratégias de precificação e campanhas promocionais. Por exemplo, as lojas podem focar suas campanhas e promoções nas primeiras semanas de cada mês, aproveitando o maior volume de compras e aumentando a margem de lucro durante esses períodos. Além disso, o acompanhamento dessas flutuações pode ajudar na gestão de estoques, garantindo que os produtos mais procurados estejam disponíveis quando a demanda é mais alta.

5. CONCLUSÃO

A análise dos dados revelou padrões consistentes no comportamento de compra dos clientes e no desempenho das filiais, proporcionando uma base sólida para decisões estratégicas no contexto do varejo. A seguir, apresentam-se os principais insights derivados da análise, abordando aspectos como a distribuição de vendas entre filiais, as preferências de pagamento dos clientes, o comportamento de compra em relação aos dias e períodos do mês, e o desempenho das categorias de produto. Esses resultados fornecem uma compreensão mais aprofundada do comportamento do consumidor, com implicações diretas para a formulação de estratégias de negócio.

Equilíbrio de Vendas entre Filiais: A análise do desempenho das filiais indicou um equilíbrio nas vendas totais entre as três unidades, com uma pequena variação observada na filial C, que superou as demais filiais em aproximadamente 4 mil unidades monetárias. Esse equilíbrio sugere uma distribuição uniforme de clientes e demanda entre as filiais, permitindo que a organização aloque seus recursos de maneira equitativa entre as unidades, sem grandes discrepâncias no volume de vendas.

Tipo de Cliente e Programa de Fidelidade: A análise da média de vendas por tipo de cliente revelou que os membros de programas de fidelidade, ou aqueles que têm acesso a descontos especiais, apresentam um ticket médio superior em comparação aos clientes não fidelizados. Esse padrão sugere que os clientes fidelizados tendem a gastar mais em cada transação, possivelmente devido aos benefícios associados à adesão ao programa, como descontos exclusivos e ofertas personalizadas. Esse comportamento pode ser utilizado como base para reforçar estratégias de fidelização, buscando aumentar o volume de compras desses clientes ao longo do tempo.

Métodos de Pagamento: A distribuição dos métodos de pagamento foi amplamente equilibrada, com uma leve predominância de algumas opções em relação a outras. O cartão de crédito foi identificado como o método menos utilizado, embora a diferença entre os métodos de pagamento fosse pequena. Esse dado reflete uma preferência diversificada por parte dos consumidores, que têm a flexibilidade de escolher o método que melhor se adequa às suas necessidades e conveniência. Para os gestores de varejo, esse equilíbrio sugere que a oferta de diversas opções de pagamento é uma estratégia eficaz para atender às expectativas de diferentes perfis de clientes.

Categorias de Produto: A análise das médias de vendas por categoria de produto revelou um padrão equilibrado, com a categoria "Casa e Bem-Estar" destacando-se levemente como a de maior ticket médio. Esse dado sugere que há uma demanda consistente por diversos tipos de produtos, com uma leve preferência por itens voltados ao uso doméstico e ao bem-estar. Essa informação é relevante para os gestores de estoque e marketing, pois indica quais categorias têm um desempenho superior em termos de valor médio por transação, podendo direcionar campanhas e promoções para essas categorias de forma mais eficaz.

Médias de Lucro: As médias de lucro por categoria seguiram um padrão similar às médias de vendas, evidenciando um equilíbrio nas margens de lucro entre as diferentes linhas de produtos. Essa constatação é importante, pois demonstra uma estabilidade na lucratividade do supermercado, mesmo em um cenário de variação de vendas. Manter essa estabilidade nas margens de lucro é essencial para o planejamento financeiro e estratégico da empresa, garantindo que os lucros não sejam excessivamente impactados por flutuações de vendas de produtos com margens mais baixas.

Padrões de Compra ao Longo do Tempo: O comportamento temporal das vendas revelou um pico significativo entre os dias 20 e 27 de janeiro, seguido de uma leve queda nas primeiras semanas de fevereiro. Esse padrão de concentração de compras nas primeiras semanas de cada mês foi observado ao longo de vários períodos analisados. Além disso, uma queda acentuada nas últimas semanas de fevereiro e março sugere que os clientes tendem a antecipar suas compras nos primeiros dias do mês, o que pode ser influenciado por fatores como o recebimento de salários e a necessidade de abastecimento inicial de produtos. O comportamento atípico observado em janeiro, com um volume de vendas significativamente maior, pode ser atribuído ao período de férias e ao início do ano, quando os consumidores, em geral, dispõem de mais tempo e necessidade de realizar compras. Esse pico pode refletir também a intenção dos consumidores de iniciar o ano com a aquisição de novos produtos ou aproveitar ofertas de início de ano.

Esses insights oferecem uma base estratégica para otimizar a gestão de estoque, aprimorar o atendimento ao cliente e desenvolver campanhas promocionais mais eficazes. Ao identificar os períodos de maior demanda e os comportamentos de compra específicos, a empresa pode implementar estratégias direcionadas, como o aumento de ofertas e descontos nas primeiras semanas de cada mês, bem como fortalecer seus programas de fidelização para maximizar a receita e a lucratividade em momentos chave. Além disso, a análise pode servir como alicerce para um planejamento de marketing mais assertivo e para a definição de ações específicas que atendam às preferências e necessidades dos diferentes perfis de consumidores.

6. REPOSITÓRIO GITHUB

Todos os arquivos e dados, como o dataset e a análise de dados exploratória com Python (Google Colab) utilizados neste trabalho serão armazenados no GitHub [2].

7. REFERÊNCIAS BIBLIOGRÁFICAS

[1] AUNG PYA AEP. **Supermarket sales**. Kaggle, 2020. Disponível em: https://www.kaggle.com/datasets/aungpyaeap/supermarket-sales?resource=download. Acesso em: 20 nov. 2024.

[2]GITHUB. Disponível em: https://github.com/RayanCrhistofer/Analise-Explorat-ria-de-Dados-de-um-Supermercado.git

[3] Análise exploratória de dados com Python. Google Colab. Disponível em: https://colab.research.google.com/drive/1dfSOIpUXQDZiVdZz4jGLfmg6Pd2nNytZ?usp=sharing