UNIVERSIDADE FEDERAL DOS VALES DO JEQUITINHONHA E MUCURI

FACET

SISTEMAS DE INFORMAÇÃO

Inteligência Artificial

Trabalho Prático I

Diamantina

2024

Lucas Aguiar Rodrigues
Gabriel Moreira Siqueira
José Inácio de Moraes Santos
Livia Moreira

Inteligência Artificial
Trabalho Prático I

Trabalho apresentado como atividade avaliativa da disciplina de Inteligência Artificial do Curso de Sistemas de Informação da UFVJM.

Professora: Luciana Pereira de Assis

Diamantina

2024

SUMÁRIO

1 Objetivo

Este trabalho visa aplicar técnicas de inteligência artificial, especificamente algoritmos de clusterização, para analisar transações de vendas online. O foco principal será na comparação e seleção do algoritmo mais adequado entre K-Means, Hierarchical, DBSCAN e Mean Shift. Cada algoritmo será avaliado com base em suas capacidades de identificar padrões de compras e preferências regionais.

É possível, por exemplo, utilizar o algoritmo K-Means para agrupar transações em clusters com base em características como preço, quantidade e categoria do produto, ajudando a revelar padrões de compra, e mesmo tentar validar a qualidade dos clusters formados.

A análise dos clusters permite tentar identificar padrões e tendências, como por exemplo a determinação dos produtos mais populares em diferentes regiões e como essas preferências mudam ao longo do tempo. Além disso, ao agrupar transações por método de pagamento, buscamos entender as preferências regionais em diferentes categorias de produtos. Os comportamentos de compra serão analisados para se tentar identificar padrões. Para representar visualmente essas informações, gráficos serão criados com o uso de bibliotecas como Matplotlib e Seaborn, mostrando a distribuição de vendas por região, métodos de pagamento e categorias de produtos.

Utilizando Inteligência Artificial, mais especificamente tentando agrupar dados sem o uso de rótulos pŕe-estabelecidos por meio do aprendizado de máquina não supervisionado ao analisar transações de vendas online, podemos tentar obter uma maior compreensão das preferências e comportamentos dos clientes. Os possíveis efeitos que ensejam benefícios da clusterização são, por exemplo, permitir otimizar a alocação de estoque, estratégias de marketing e vendas, garantir uma resposta rápida às mudanças do mercado e uma experiência de compra mais alinhada ao perfil dos clientes.

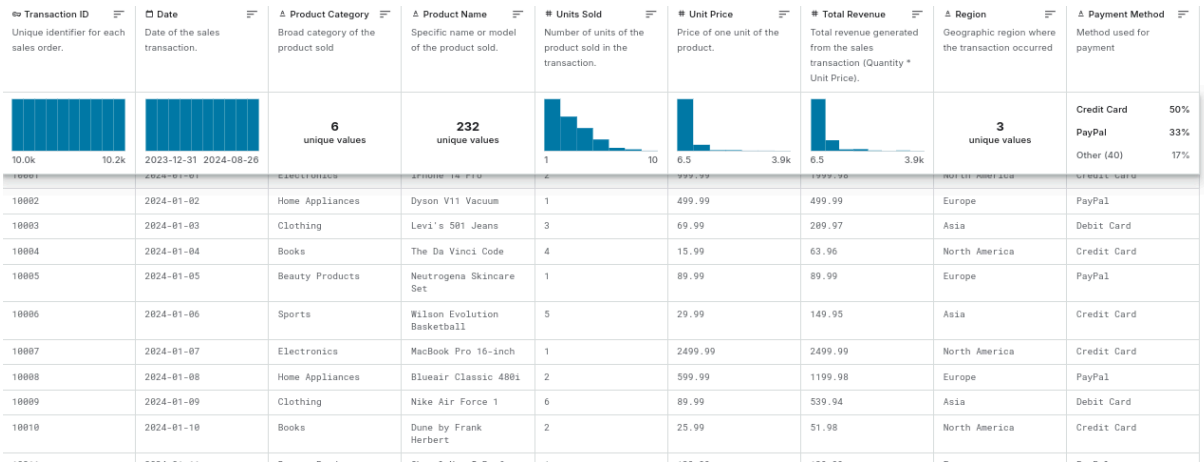
2 Contextualização do Problema

Empresas de vendas online enfrentam o desafio de otimizar suas operações de maneira eficiente e eficaz, dado o vasto e diversificado volume de dados de transações que geram diariamente. Um problema específico que surge é como agrupar diferentes categorias de produtos e padrões de vendas para identificar tendências regionais, preferências de pagamento e comportamentos de compra. A dificuldade reside em analisar e interpretar esses dados complexos e variados, a fim de tomar decisões informadas que melhorem a alocação de estoque, campanhas de marketing e estratégias de vendas.

2.1 Dados do Problema

O dataset selecionado para este trabalho oferece uma visão abrangente sobre transações de vendas online, cobrindo diferentes categorias de produtos. Cada linha representa uma transação individual, incluindo informações detalhadas como ID do pedido, data, categoria, nome do produto, quantidade vendida, preço unitário, preço total, região e método de pagamento. As colunas do dataset são as seguintes:

- **ID do Pedido**: Identificador único de cada transação de venda.
- **Data**: Data em que a transação foi realizada.
- **Categoria**: Categoria geral do produto vendido, como Eletrônicos, Eletrodomésticos, Roupas, Livros, Cosméticos, e Esportes.
- **Nome do Produto**: Nome específico ou modelo do produto vendido.
- **Quantidade**: Número de unidades do produto vendidas na transação.
- **Preço Unitário**: Valor de uma unidade do produto.
- **Preço Total**: Receita gerada pela transação (Quantidade x Preço Unitário).
- **Região** geográfica onde a transação ocorreu (ex.: América do Norte, Europa, Ásia).
- **Método de Pagamento**: Forma de pagamento utilizada (ex.: Cartão de Crédito, PayPal, Cartão de Débito).



Transaction ID Unique identifier for each sales order.	Date Date of the sales transaction.	Product Category Broad category of the product sold	Product Name Specific name or model of the product sold.	Units Sold Number of units of the product sold in the transaction.	Unit Price Price of one unit of the product.	Total Revenue Total revenue generated from the sales transaction (Quantity * Unit Price).	Region Geographic region where the transaction occurred	Payment Method Method used for payment
10.0k 10.2k	2023-12-31 2024-08-26	6 unique values	232 unique values	1 10	6.5 3.9k	6.5 3.9k	3 unique values	Credit Card 50% PayPal 33% Other (40) 17%
10002	2024-01-02	Home Appliances	Dyson V11 Vacuum	1	499.99	499.99	Europe	PayPal
10003	2024-01-03	Clothing	Levi's 501 Jeans	3	69.99	209.97	Asia	Debit Card
10004	2024-01-04	Books	The Da Vinci Code	4	15.99	63.96	North America	Credit Card
10005	2024-01-05	Beauty Products	Neutrogena Skincare Set	1	89.99	89.99	Europe	PayPal
10006	2024-01-06	Sports	Wilson Evolution Basketball	5	29.99	149.95	Asia	Credit Card
10007	2024-01-07	Electronics	MacBook Pro 16-inch	1	2499.99	2499.99	North America	Credit Card
10008	2024-01-08	Home Appliances	Blueair Classic 480i	2	599.99	1199.98	Europe	PayPal
10009	2024-01-09	Clothing	Nike Air Force 1	6	89.99	539.94	Asia	Debit Card
10010	2024-01-10	Books	Dune by Frank Herbert	2	25.99	51.98	North America	Credit Card

Esse conjunto de dados será a base para análises de clusterização, que permitirão identificar padrões nas transações, incluindo tendências de compra por região, preferências de pagamento, e comportamentos de consumo. Essas informações servirão de base para avaliar a aplicação de diferentes algoritmos de clusterização e extrair insights estratégicos para otimização de operações comerciais.

3 Bibliotecas Usadas

1. NumPy (import numpy as np)

- Utilidade: NumPy é fundamental para realizar operações matemáticas e manipulação de arrays de forma eficiente. No contexto do nosso objetivo, NumPy pode ser usado para realizar cálculos necessários na análise de tendências de vendas, preferências de pagamento e comportamento de compra.
- Exemplo de Uso: Cálculo de médias e variâncias de preços de produtos em diferentes regiões para identificar tendências.

2. Pandas (import pandas as pd)

- Utilidade: Pandas é essencial para a manipulação e análise de dados. Ele facilita a leitura de dados de várias fontes, a limpeza, e a análise exploratória. No nosso caso, Pandas ajudará a organizar e manipular o conjunto de dados de transações de vendas online.
- Exemplo de Uso: Importação dos dados de vendas, agrupamento por regiões, categorias de produtos, e métodos de pagamento para análise.

3. Matplotlib (import matplotlib.pyplot as plt)

- Utilidade: Matplotlib é crucial para a visualização de dados. Ele permite criar gráficos que ajudam a entender padrões e tendências nos dados de vendas.
- Exemplo de Uso: Criação de gráficos de dispersão para visualizar a relação entre preços de produtos e a quantidade vendida em diferentes regiões.

4. Seaborn (import seaborn as sns)

- Utilidade: Seaborn, construído sobre o Matplotlib, oferece gráficos estatísticos mais atraentes e fáceis de interpretar. Ele será útil para criar visualizações detalhadas dos dados, facilitando a identificação de padrões complexos.
- Exemplo de Uso: Visualização de mapas de calor para mostrar a correlação entre diferentes categorias de produtos e preferências regionais de pagamento.

5. Locale (import locale)

- Utilidade: A biblioteca locale é usada para configurar e formatar números, datas e moedas conforme as convenções locais. Isso é importante para garantir que os dados sejam apresentados de forma compreensível para os diferentes mercados regionais.
- Exemplo de Uso: Formatação de preços e datas conforme as convenções locais das regiões analisadas.

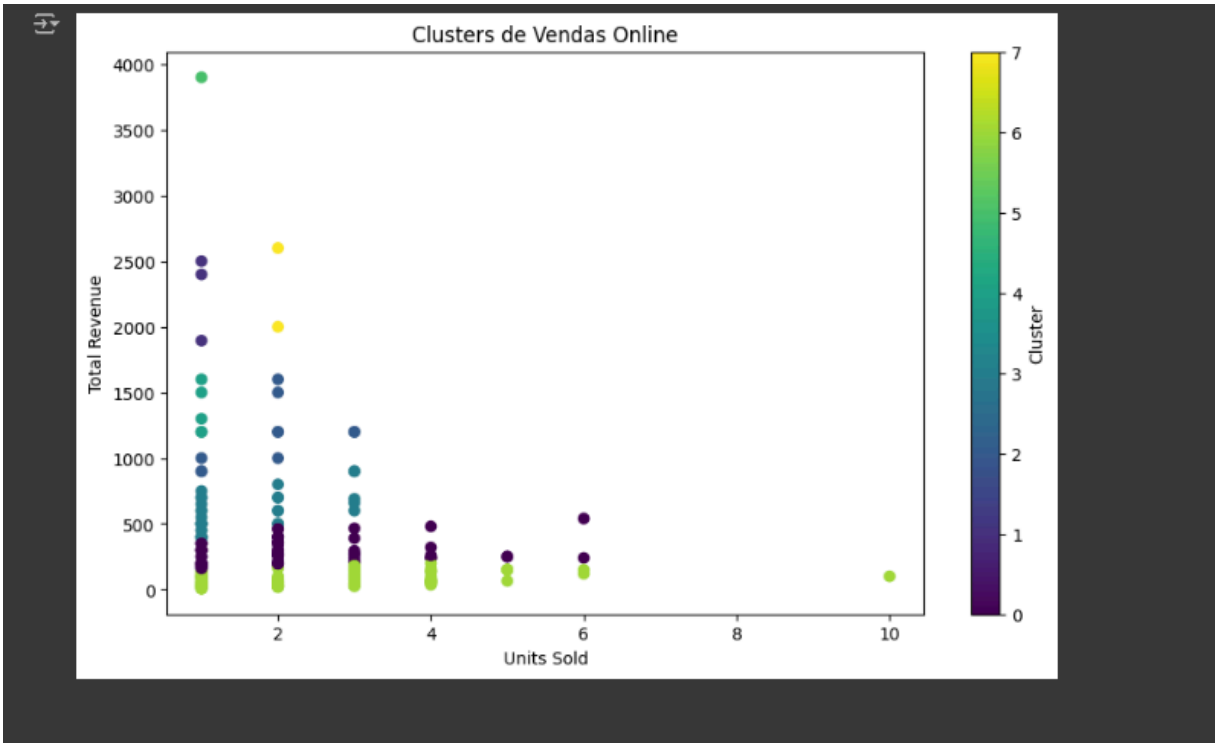
6. KMeans (from sklearn.cluster import KMeans)

- Utilidade: KMeans é um algoritmo de clustering que ajudará a agrupar as transações de vendas em clusters distintos, facilitando a identificação de padrões de compra e tendências regionais.
- Exemplo de Uso: Agrupamento de transações por categorias de produtos para identificar padrões de venda, preferências de pagamento e comportamentos de compra em diferentes regiões.

Aplicação das Bibliotecas

Estas bibliotecas em conjunto permitem a análise detalhada do conjunto de transações de vendas online. Elas ajudam a processar os dados, aplicar técnicas de machine learning para identificar clusters, e visualizar os resultados de forma clara e compreensível. Isso fornece à empresa insights valiosos para otimizar a alocação de estoque, campanhas de marketing e estratégias de vendas, alinhando-se com o objetivo proposto.

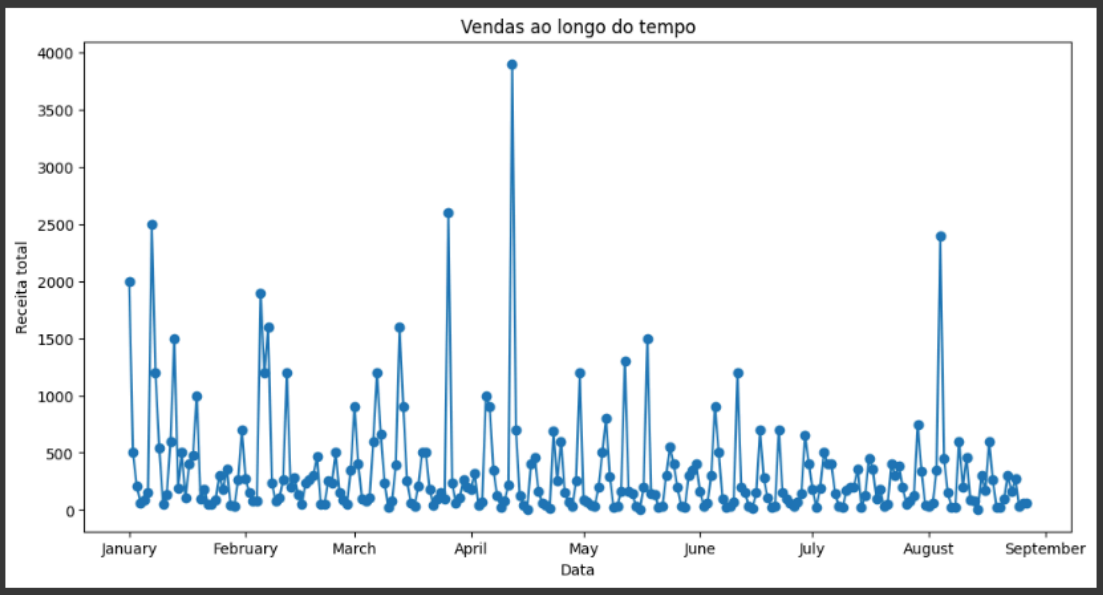
KMeans -


Clusters de Vendas Online
Total Revenue
Units Sold
Cluster
4000
3500
3000
2500
2000
1500
1000
500
0
2
4
6
8
10
7
6
5
4
3
2
1
0

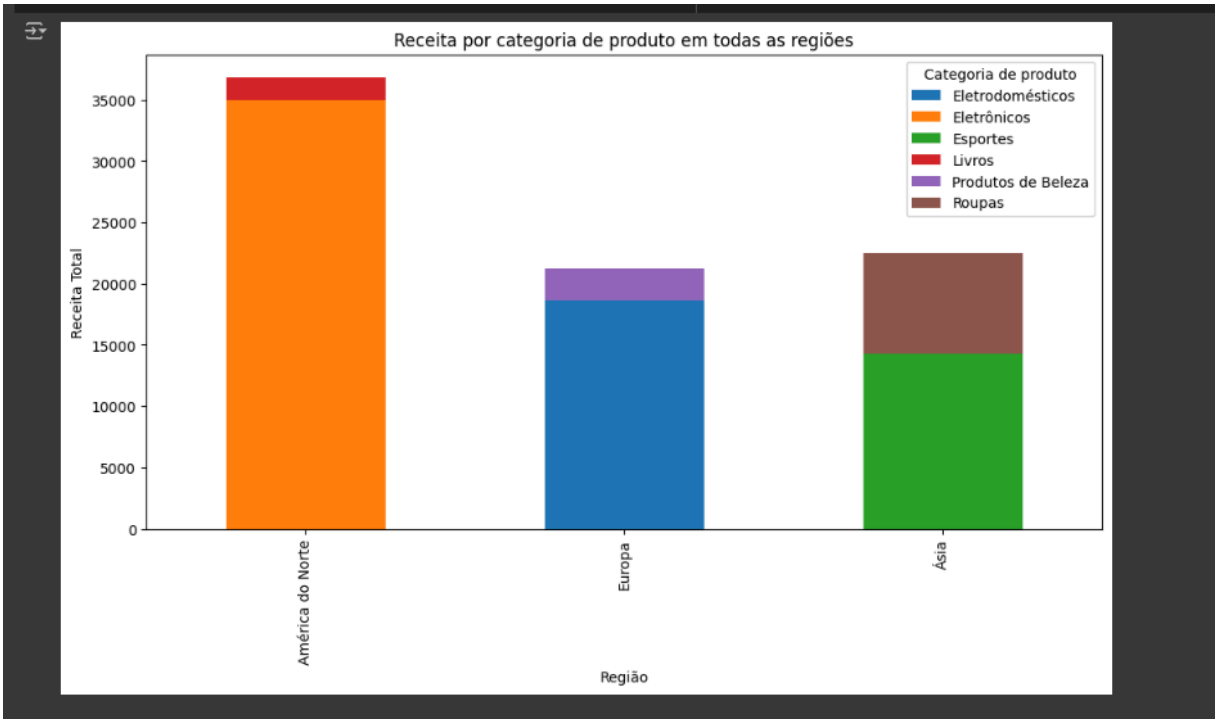
Matplotlib (import matplotlib.pyplot as plt)

Tendências de vendas ao longo do tempo

```
# @title Tendências de vendas ao longo do tempo

# Aggregate sales data by date
sales_trends = df.groupby('Date')['Total Revenue'].sum().reset_index()

# Plot sales trends over time
plt.figure(figsize=(12, 6))
plt.plot(sales_trends['Date'], sales_trends['Total Revenue'], marker='o')

# Formatar o eixo x para mostrar o nome dos meses
plt.gca().xaxis.set_major_formatter(mdates.DateFormatter('%B'))

plt.title('Vendas ao longo do tempo')
plt.xlabel('Data')
plt.ylabel('Receita total')

plt.show()
```

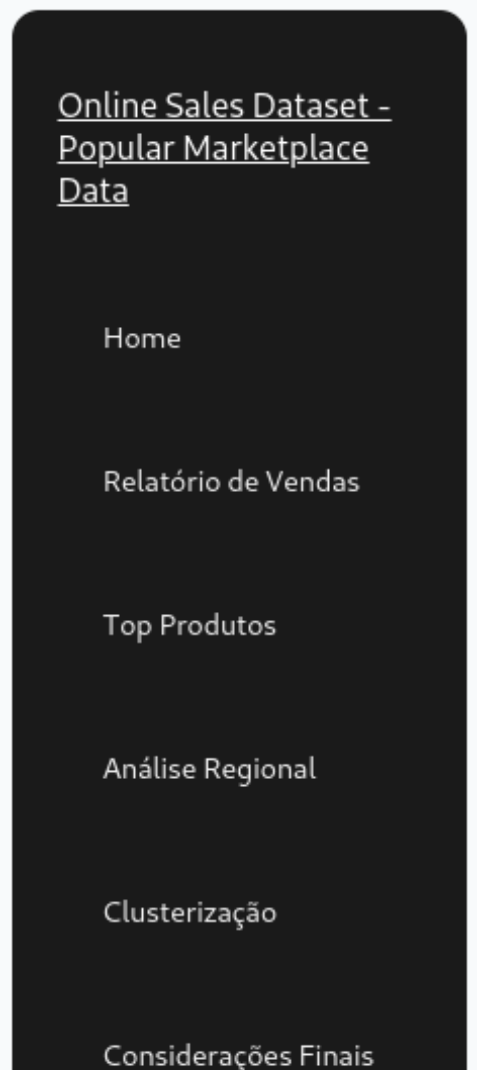


Matplotlib -


Receita por categoria de produto em todas as regiões
Categoria de produto
Eletrodomésticos
Eletrônicos
Esportes
Livros
Produtos de Beleza
Roupas
Receita Total
35000
30000
25000
20000
15000
10000
5000
0
América do Norte
Europa
Ásia
Região

4 Protótipo

O protótipo deu início a uma aplicação web utilizando o framework Django, permitindo a criação de um sistema dinâmico e interativo. Houve a utilização de Django Templates para a renderização das páginas e inclusão de arquivos CSS e JavaScript para estilização e funcionalidades interativas na interface do usuário.
A sidebar do protótipo desempenha um papel crucial na estrutura de navegação e na experiência do usuário, facilitando o acesso às diferentes seções da aplicação.

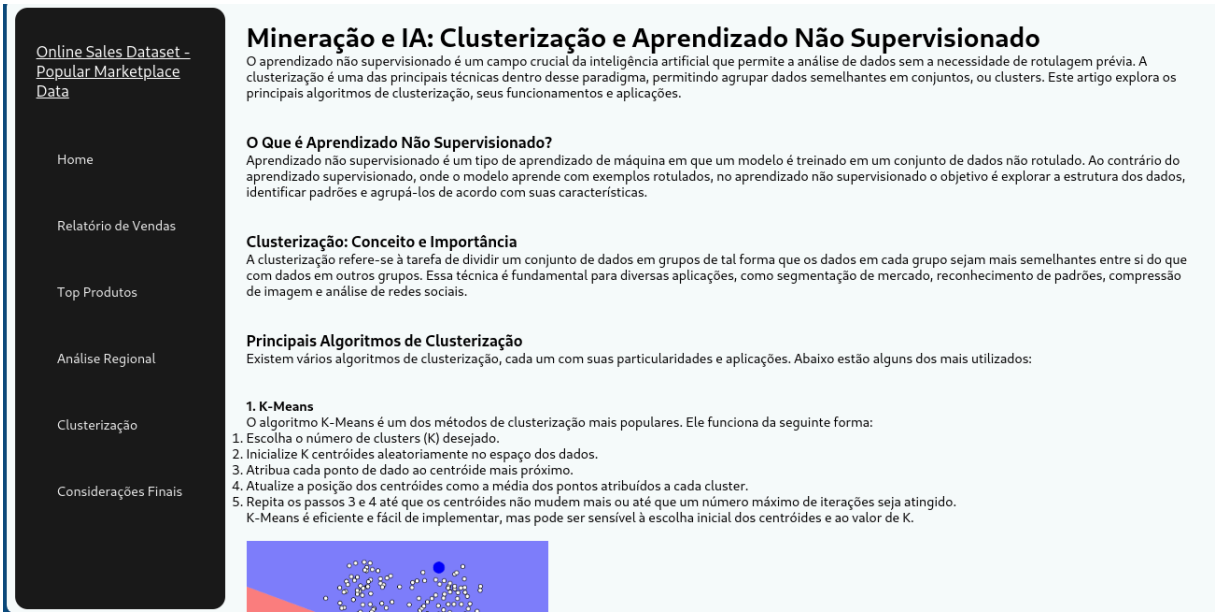


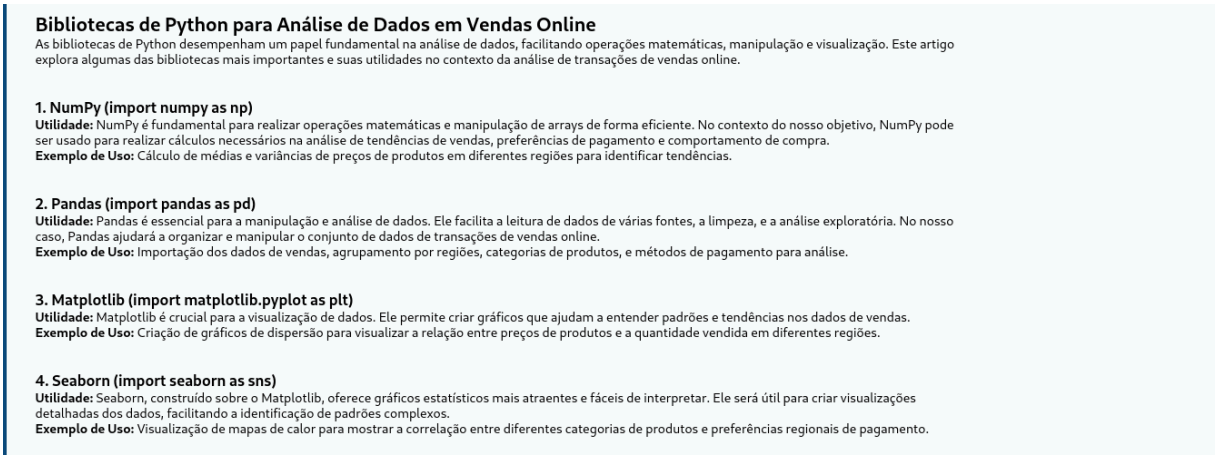
A sidebar é uma barra lateral que fica fixa na interface, proporcionando fácil acesso às diferentes seções do protótipo. Cada item na sidebar está associado a uma rota específica que corresponde a uma seção ou funcionalidade do protótipo. O primeiro item é o hyperlink para o Dataset “Online Sales Dataset - Popular Marketplace Data”.

Até então, as seguintes páginas foram estabelecidas para fazerem parte do sistema final e estão na navbar:

- Home, que leva à visão geral do dashboard com dados gerais do dataset, proporcionando um retorno à página inicial.

- Relatório de Vendas, Análise Regional e Top Produtos em geral serão páginas que irão conter dados frutos da análise de dados e uso dos algoritmos de clusterização no dataset escolhido. Estão em desenvolvimento, conforme o trabalho progride e se busca maiores formas de sinergia para exibir dados advindos da análise via Google Colab.
- Clusterização atualmente é uma seção informativa informando sobre os algoritmos de clusterização e as principais bibliotecas utilizadas até então no desenvolvimento do trabalho. Conforme o trabalho progride, pode ser seção dedicada à análise de clusters.





- Considerações Finais será o painel que irá conter os principais insights e resultados decorrentes da conclusão deste trabalho.

REFERÊNCIAS

VERMA, Shreyansh. Online Sales Dataset - Popular Marketplace Data. Disponível em: https://www.kaggle.com/datasets/shreyanshverma27/online-sales-dataset-popular-marketplace-data. Acesso em: 20 out. 2024.
https://colab.research.google.com/drive/1LwDsQU3UE7R-ymYYdm7sgJQ8wUrvypeF?usp=sharing

https://ealexbarros.medium.com/clusterização-de-produtos-de-um-supermercado-com-python-1d1cb808dcc6