Desafio - estágio em dados Itaú

Candidato

Nome: Andreas Hukuhara Christe
E-mail: andreash.christe@hotmail.com
CPF: 443.853.418-69

1. Introdução

O projeto consiste em analisar o dataset Ecommerce disponível no kaggle (https://www.kaggle.com/datasets/raziehghahartars/ecommerce?resource=download) e responder o desafio proposto, assim como as 4 perguntas levantadas.

2. Análise Exploratória e higienização da base.

Inicialmente, foi realizado o processo de classificação das variáveis, para melhor entendimento do dos dados disponibilizados.

- Classificação das variáveis:

Variável	Tipo
Customer ID	Quantitativa Discreta
Purchase Date	Qualitativa Nominal
Product Category	Qualitativa Nominal
Product Price	Quantitativa Contínua
Quantity	Quantitativa Discreta
Total Purchase Amount	Quantitativa Contínua
NPS	Quantitativa Discreta
Customer Age	Quantitativa Discreta
Gender	Qualitativa Nominal
Source	Qualitativa Nominal
Country	Qualitativa Nominal
State	Qualitativa Nominal
Latitude	Quantitativa Contínua
Longituide	Quantitativa Contínua

Tabela 01 - Classificação das variáveis

Em seguida, a análise exploratória inicial dos dados nos auxiliou na verificação dos tipos de dados (int64, float64, data, string) para cada uma das colunas, avaliação de valores nulos no dataset e nas estatísticas descritivas (média, quartis, mínimo, máximo e quantidade de observações) de cada variável. Com base nisso, foram feitas as devidas modificações para que estivessem de acordo com a classificação de variáveis proposta. Essa

análise possibilitou uma compreensão da disposição do dataset e a identificação de possíveis outliers.

Após isso, fora realizado o processo de limpeza de dados, caso houvesse linhas com dados nulos ou brancos, e a seleção das colunas relevantes ('Purchase Date','Product Category', 'Product Price', 'Quantity','NPS', 'Customer Age ', 'Gender','Source') para a resposta das questões.

Além disso, foram feitas algumas análises para verificar a distribuição da quantidade de vendas por gênero, como podemos ver na figura a seguir:

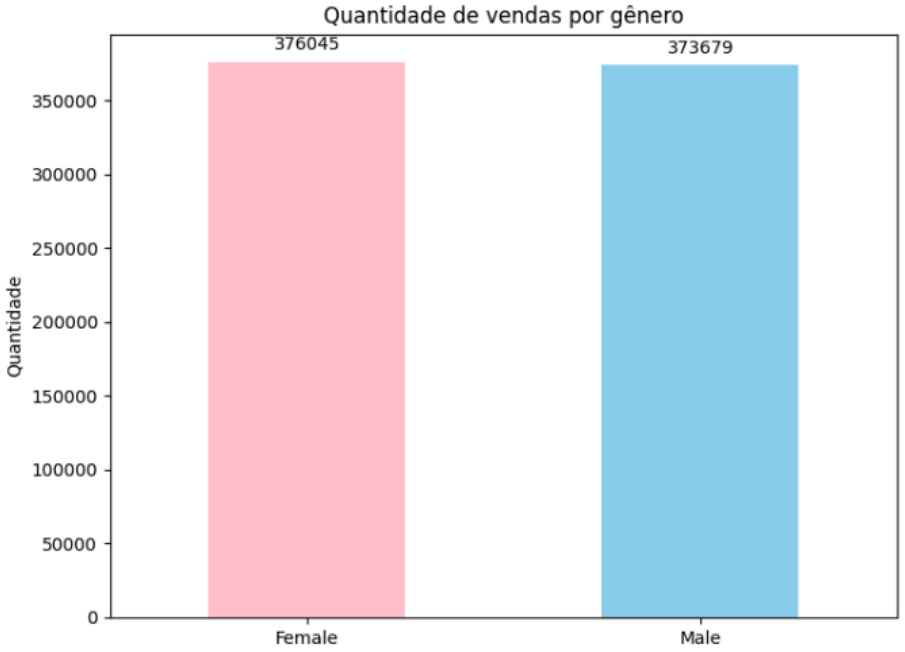

Figura 01 - Quantidade de vendas por gênero

E também de gênero por canal, como podemos ver:

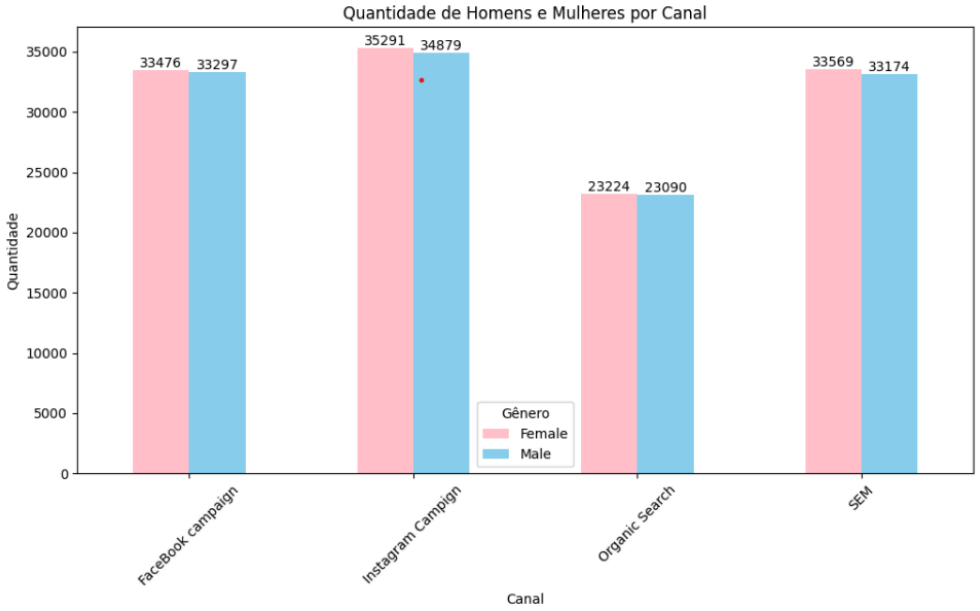

Figura 02 - Quantidade de Homens e Mulheres por Canal

Ademais, para as variáveis numéricas foram analisadas as distribuições dos dados, por meio do histograma, boxplot e o Q-Q plot. Tendo isso em vista, foi possível identificar que não há outliers para as variáveis relevantes no dataset, neste caso, não havendo a necessidade de eliminar dados que possuem um caráter extremo.

Abaixo segue um exemplo dos gráficos citados anteriormente para a variável Customer Age, diante disso, é possível observar que os dados estão distribuídos de maneira similar (homogênea), tendo em vista cada idade presente no dataset e que não há presença de outliers.

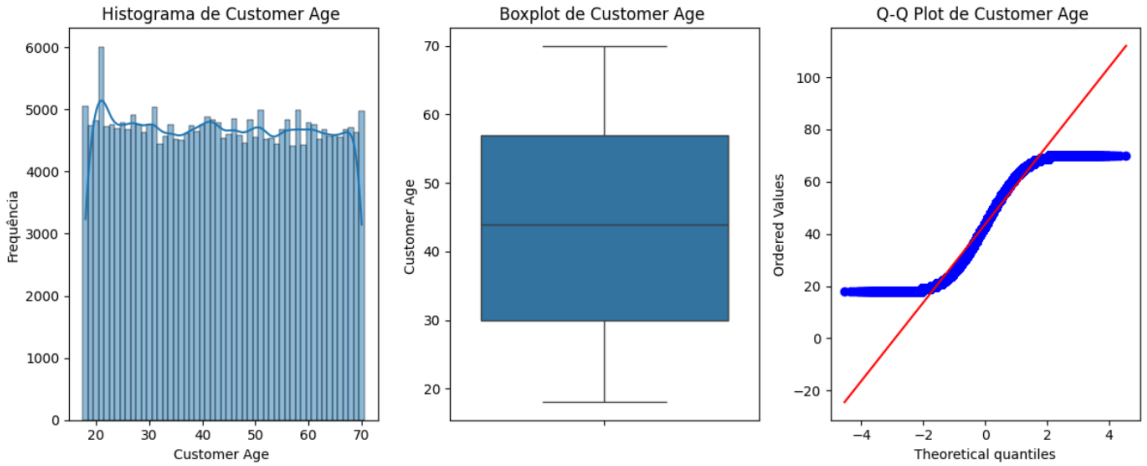

Figura 03 - Histograma, Boxplot, Q-Q plot da variável Customer Age (idade)

Esse mesmo comportamento é observado nas variáveis 'Product Price', 'Quantity', 'NPS' e isso pode ser observado nos gráficos plotados no python jupyter notebook.

3. Respostas

- **Quais os produtos mais vendidos considerando os últimos 3 anos?**

Para essa análise existem diferentes regras de negócios que podem ser interpretadas de diferentes formas, para essa questão em específico, fora considerado a data mais recente de uma compra realizada, ou seja a última compra que foi registrada no dataset, e a partir disso dado um offset de 3 anos para trás para que possa ser respondida essa questão.
Diante disso, obtivemos as seguintes informações dos produtos, nos últimos 3 anos, para cada agrupamento por categoria de produtos:

	Product Category	Quantity
0	Clothing	182696
1	Books	181069
2	Electronics	121867
3	Home	120696

tabela 02 - Quantidade de produtos por categoria para os últimos três anos

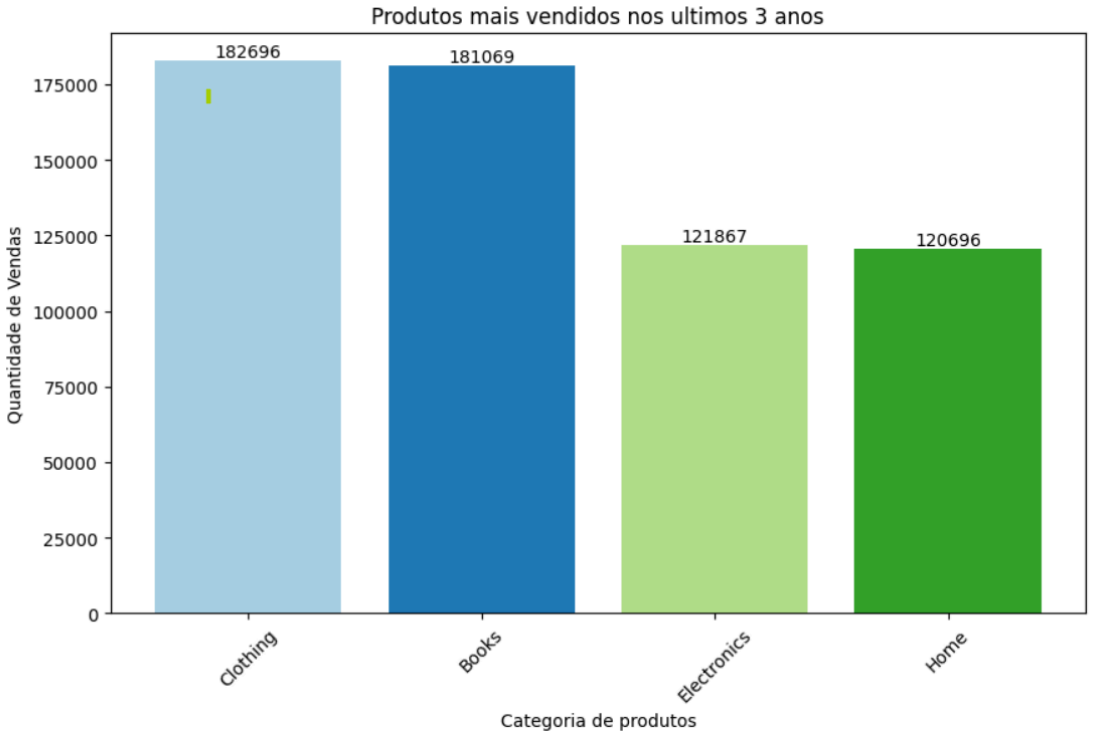

Figura 04 - gráfico de barras dos produtos mais vendidos nos últimos 3 anos

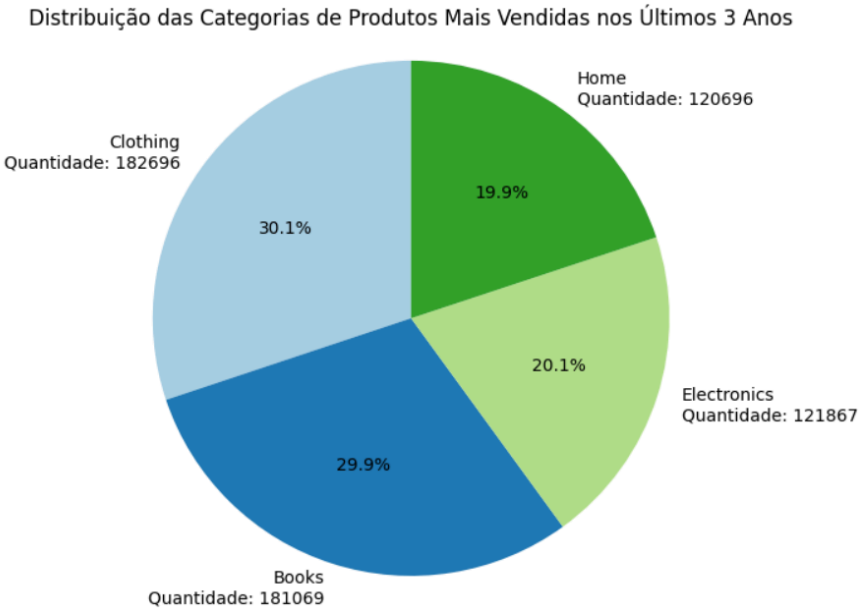

Figura 05 - gráfico de pizza da categoria de produtos mais vendidos nos últimos 3 anos

Resposta: Diante destas informações o produto mais vendido são roupas representando 30,1% das vendas nos últimos 3 anos, e em segundo lugar os livros ,com uma pequena diferença, representando 29,9% das vendas.

Outra maneira de interpretar essa questão seria utilizar a data atual,ou seja, literalmente a data de hoje e realizar o mesmo processo de análise. Contudo, para não ficar repetitivo essa análise não fora contemplada.

Vale ressaltar que o que está sendo analisado é a categoria de produtos, pois o dataset não apresenta a coluna “Produto” e essa observação também é válida para outras análises feitas em que envolvem a coluna Product Category sendo tratada como um produto.

- **Qual o produto mais caro e o mais barato?**

Para analisar qual o produto mais caro e o mais barato a abordagem utilizada foi buscar na coluna Product Price e identificar os valores máximo e mínimo presente nela. Com isso, temos o seguinte resultado, tendo em consideração o produto:

	-	Preço do Produto	Quantidade de Observações	Categorias
0	Maior Valor	500.0	474	Electronics, Home, Books, Clothing
1	Menor Valor	10.0	493	Electronics, Clothing, Books, Home

tabela 03 - Produtos mais caros e mais baratos

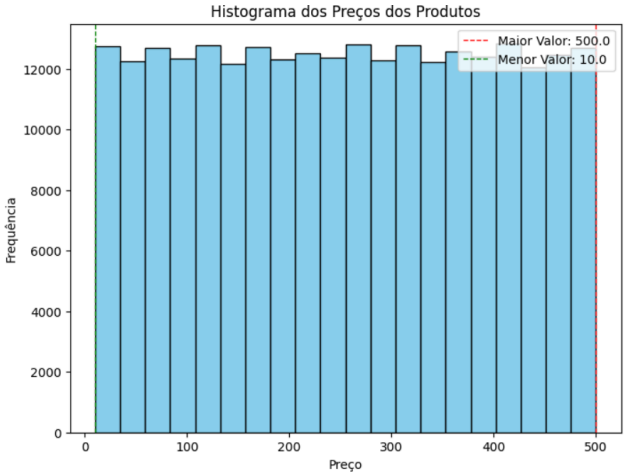

Figura 06 - Histograma dos Preços dos produtos

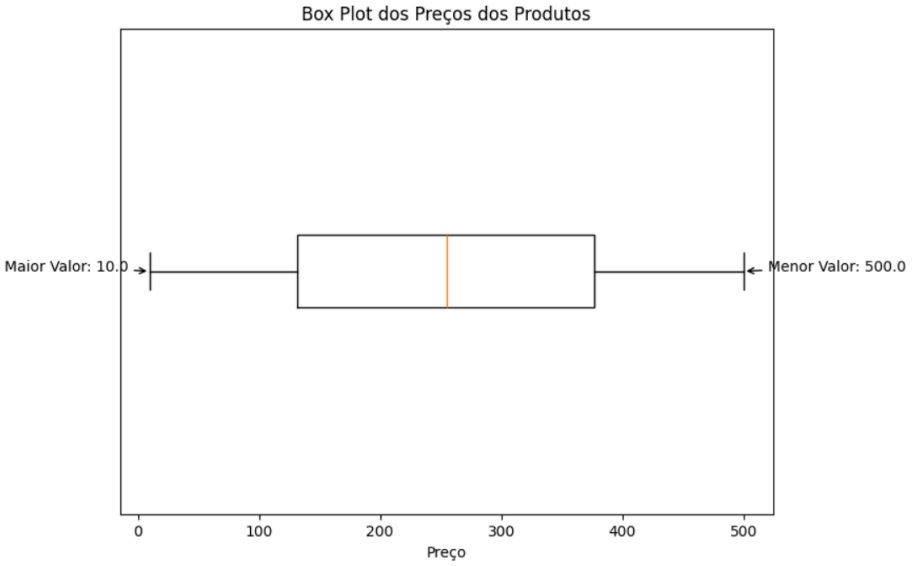

Figura 07 - Boxplot dos preços dos Produtos

Resposta: Tendo a tabela e os gráficos em vista, é possível identificar que há vários representantes para os produtos mais caros, que possuem um valor de 500 (unidade de dinheiro) e estão nas categorias roupas, livros, eletrônicos e casa. Já para os produtos de menores valores temos as mesmas categorias e o valor é de 10 (unidade de dinheiro).

Além disso, é possível verificar que há 474 produtos que possuem o valor de 500 e 493 produtos que possuem o valor de 10. Vale ressaltar que não é possível identificar se dentre esses produtos há a ocorrência de produtos iguais, pois, mais uma vez, o dataset apresenta a categoria dos produtos, não temos informações específicas de cada produto.

- **Qual a categoria de produto mais vendida e menor vendida? Qual a categoria mais e menos cara?**

Para essa análise, agrupamos os produtos por categoria e identificamos a quantidade de produtos vendidas em cada segmento, como podemos nos gráficos e tabelas abaixo:

	Categoria de Produtos	Total Vendido
0	Clothing	225322
1	Books	223876
2	Electronics	150828
3	Home	149698

Tabela 04: tabela de Produtos vendido por categoria.

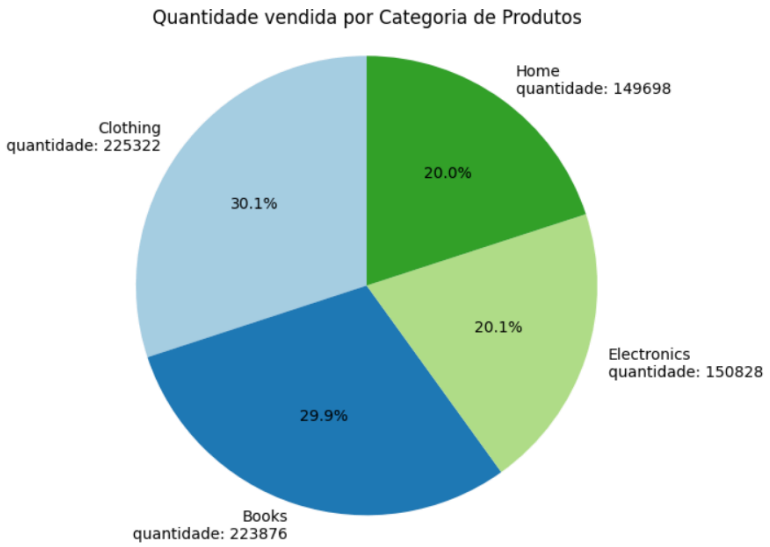

Figura 08 - Gráfico de pizza da quantidade vendida por Categoria de Produtos.

Resposta: A partir da tabela e gráfico acima é notável que a categoria de produto mais vendido são as roupas, com 30,1% das vendas, e o menos vendido são os da categoria casa, com apenas 20% das vendas.

Já para analisar qual a categoria mais cara e mais barata, foi realizada a média aritmética dos preços de cada observação no dataset e com isso obtivemos os seguintes resultados:

	Categoria de Produtos	Media de preco
0	Home	254.84
1	Electronics	254.72
2	Books	254.71
3	Clothing	254.45

Tabela 05: tabela de preço médio por Categoria de produtos.

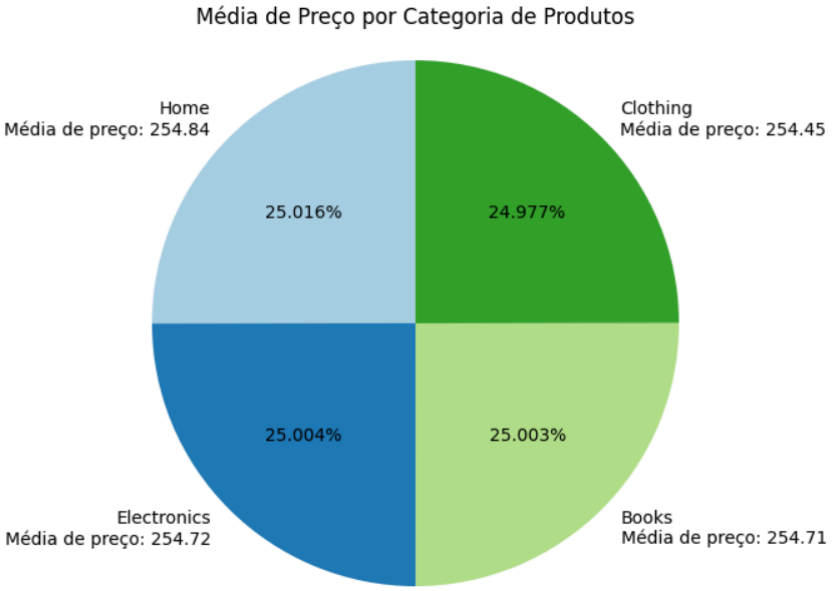

Figura 09 - Gráficos de pizza da média de preço por categoria de produtos.

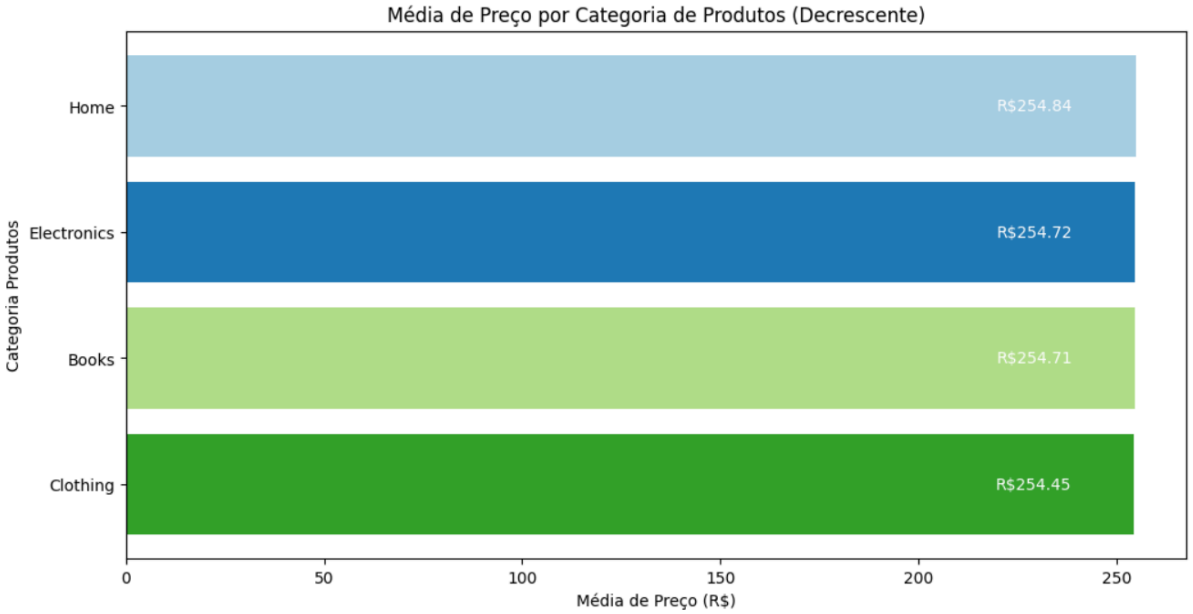

Figura 10 - gráfico de barra de média de preço por Categoria de produtos (Decrescente)

Resposta: Logo, a categoria de produtos que possui uma média maior de preços são os de ‘Casa’ que possuem um valor médio de 254,84 (unidade de dinheiro) e a categoria de produto com menor valor são as ‘roupas’ com valor médio de 254,45 (unidade de dinheiro). Mas vale observar que a diferença entre as quatro categorias são bem pequenas.

- **Qual o produto com melhor e pior NPS?**

Para essa análise foi utilizada o NPS médio para identificar qual o melhor e pior NPS por categoria de produto, segue as tabelas e gráficos analisados:

	Categoria de Produtos	NPS Medio
0	Home	5.011394
1	Books	5.003698
2	Clothing	4.983904
3	Electronics	4.966544

Tabela 06 - tabela de NPS médio ordenado decrescente por categoria de produto.

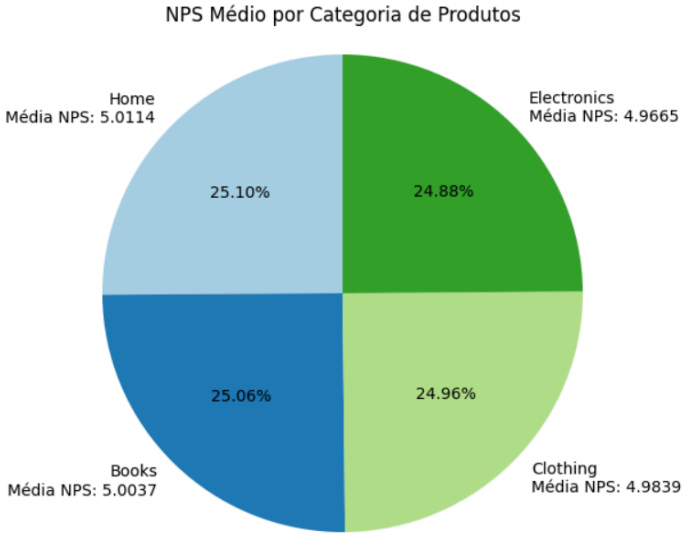

Figura 11 - Gráfico de pizza do NPS médio por categoria de produtos.

Resposta: Diante das informações acima é possível identificar que os produtos da categoria ‘casa’ possuem o melhor NPS médio e os produtos da categoria ‘eletrônicos’ possuem o pior NPS médio. Vale ressaltar que o NPS médio é muito próximo entre as categorias e a quantidade de observações em cada categoria também são semelhantes.

Desafio: Analisando a base de dados, qual o tipo de público(considerando gênero e idade) e o canal ideal para vender determinado tipo de produto?

Para essa análise foi utilizada a análise de cluster no qual consiste em agrupar as tuplas do dataset em grupos que possuem características em comum e para isso foi utilizado o algoritmo de Kmeans que realiza esse processo de agrupamento.

Antes de dar início a análise cluster, vamos fazer algumas análises de correlação entre algumas variáveis quantitativas, para verificar se há correlação.

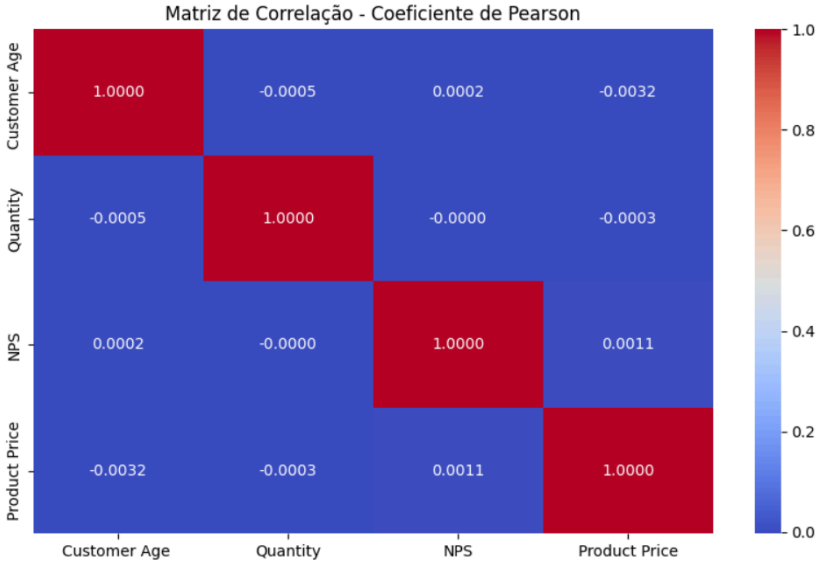

Figura 12 - Matriz de correlação entre variáveis quantitativas

Como podemos observar os valores de correlação são praticamente 0, fator que evidencia a inexistência de correlação entre as variáveis, ou seja, o comportamento de uma variável não influencia em outra variável.

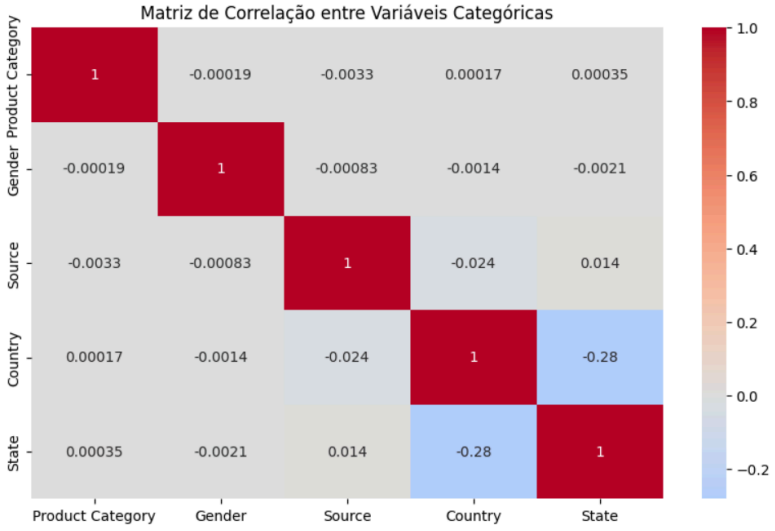

Figura 13 - Matriz de correlação entre variáveis categóricas

Como podemos ver na figura acima, não há correlação entre as variáveis categóricas, assim como ocorre nas variáveis quantitativas, com exceção do país e estado.

Visando identificar diferentes grupos, a análise proposta para esta questão está em realizar agrupamentos considerando grupos e canais, para identificar possíveis semelhanças entre esses atributos, e posteriormente verificar qual produto é vendido dentre esses grupos.

Para definir o número de cluster , foi utilizado o método do Cotovelo, pelo qual nos dá uma estimativa de qual seria o número ideal de cluster, sendo o a tangente do cotovelo o valor ideal, neste caso 5 clusters. Contudo, isso é apenas

uma estimativa, valores como 4 e 6 também podem ser analisados e gerarem informações relevantes.

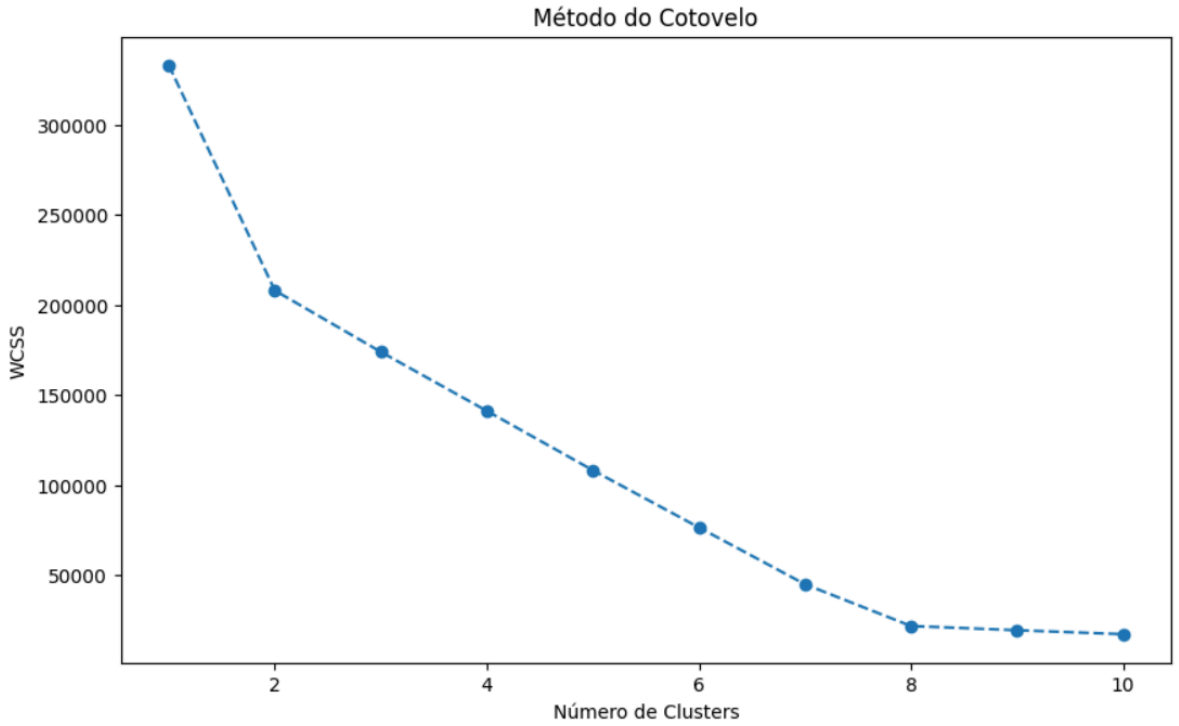

Figura 14 : gráfico método do cotovelo

Nesta apresentação, apresentarei apenas o resultado de 5 cluster, contudo a análise de 4 e 6 clusters também foram feitas, basta apenas mudar o valor na variável indicada abaixo:

```
# Aplicar K-Means com o número ideal de clusters
optimal_clusters = 5  # Escolha o valor ideal baseado no gráfico do cotovelo
kmeans = KMeans(n_clusters=optimal_clusters, init='k-means++', max_iter=300, n_init=10, random_state=0)
data_cleaned_dummy['Cluster'] = kmeans.fit_predict(variables_for_clustering)

# Exibir as primeiras linhas do DataFrame com os clusters
print(data_cleaned_dummy.head())
```

Figura 15 - variável que define o número de cluster.

Com isso, temos os seguintes clusters :

	Cluster	Número de Observações
0	5	89561
1	4	56700
2	3	35291
3	1	34879
4	2	33569
5	Total	250000

Tabela 07 - distribuição dos clusters

Resposta: Com isso, conseguimos tirar algumas informações acerca de cada grupo para identificar, qual público e canal são os ideais para uma determinada venda :

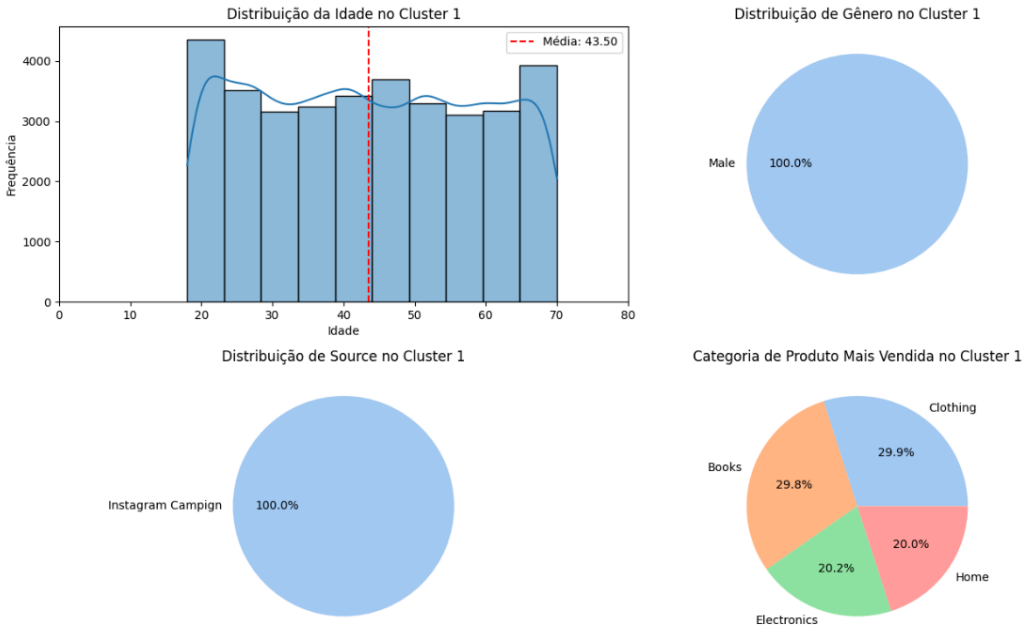

Figura 16 - Informações sobre o cluster 1

Este primeiro Cluster é composto por pessoas com idade média de 43.5 anos, pelo gênero masculino e que utilizam o Instagram Campign como canal de comunicação, e é possível identificar que os produtos mais vendidos nesse público são roupas, seguido por livros e por último produtos de casa.

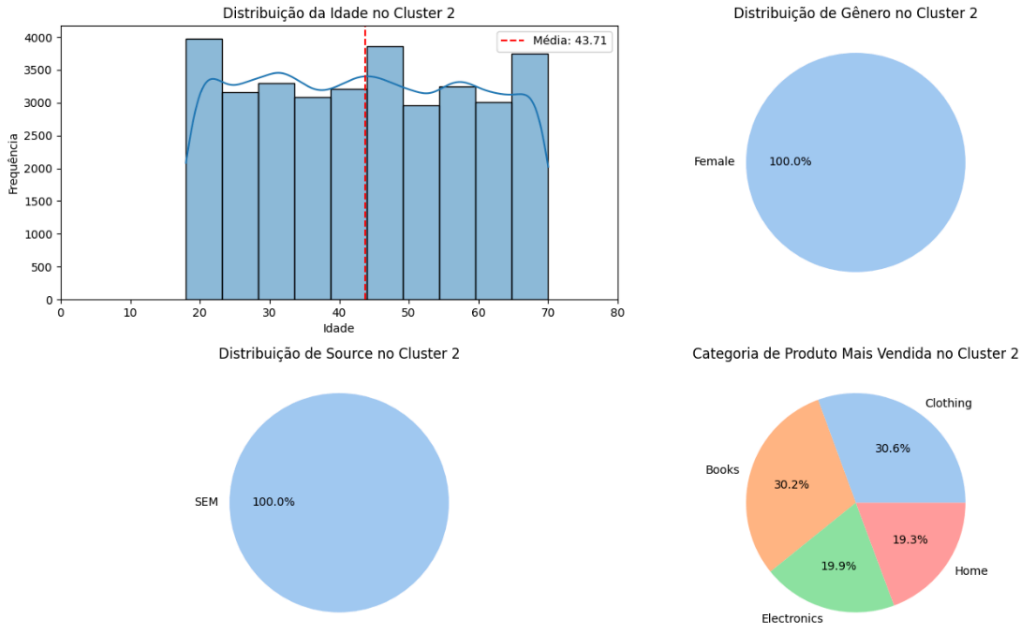

Figura 17 - Informações sobre o cluster 2

O segundo cluster é composto por pessoas com idade média de 43.7 anos, pelo gênero feminino e que utilizam o SEM como canal de comunicação, e é possível identificar que os produtos mais vendidos nesse público são roupas, seguido por livros e por último produtos de casa.

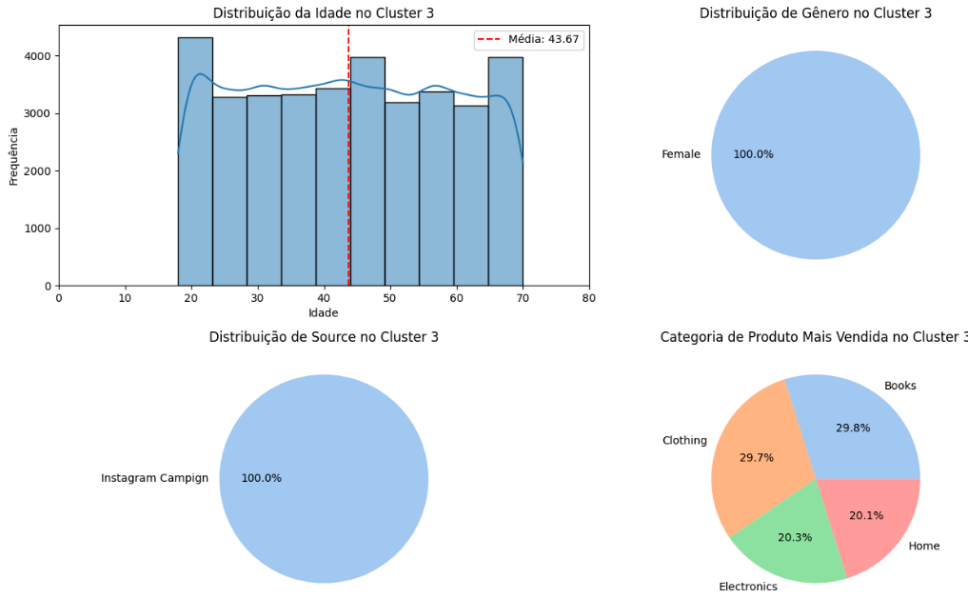

Figura 18 - Informações sobre o cluster 3

O terceiro cluster é composto por pessoas com idade média de 43.7 anos, pelo gênero feminino e que utilizam o Instagram Campign como canal de comunicação, e é possível identificar que os produtos mais vendidos nesse público são livros, seguido por roupas e por último produtos de casa.

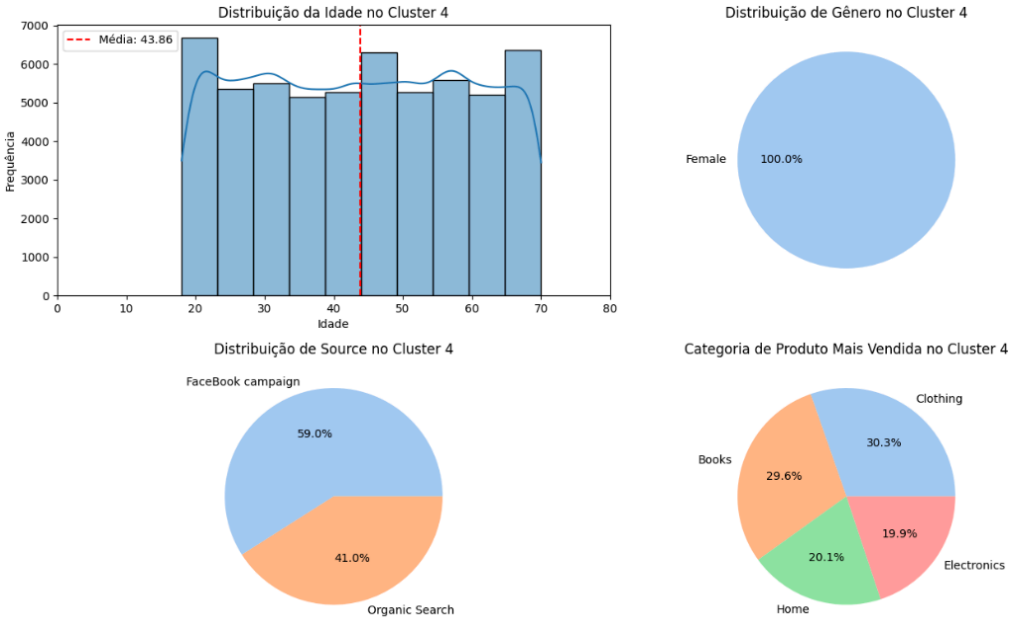

Figura 19 - Informações sobre o cluster 4

O quarto cluster é composto por pessoas com idade média de 43.9 anos, pelo gênero feminino e que utilizam o FacebookCampaign e o Organic Search como canal de comunicação, e é possível identificar que os produtos mais vendidos nesse público são roupas, seguido por livros e por último produtos de eletrônicos.

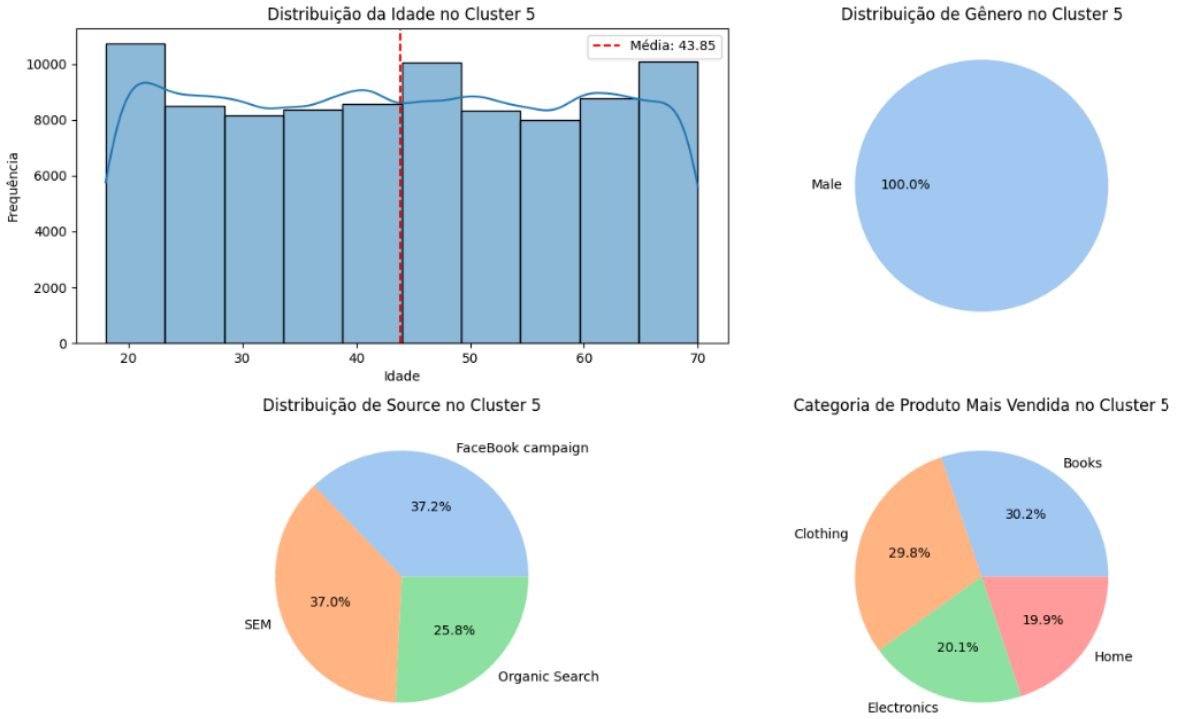

Figura 20 - Informações sobre o cluster 5

O quinto cluster é composto por pessoas com idade média de 43.9 anos, pelo gênero feminino e que utilizam o FacebookCampaign e o Organic Search e SEM como canal de comunicação, e é possível identificar que os produtos mais vendidos nesse público são livros, seguido por roupas e por último produtos de casa.

Conclusão do desafio:

Apesar de identificar grupos e a categoria de produtos mais e menos vendidos em cada um deles, o caráter homogêneo de distribuição dos dados no dataset não nos permite informar, somente com esses dados fornecidos, com maior precisão e maior distinção qual o público alvo e canal para um determinado produto.

Vale ressaltar também que a idade não é um fator determinante no agrupamento dos clusters, haja vista que a idade média dos 5 clusters são muito semelhantes.