Processo Seletivo Estágio em Dados

Desafio Dados Itaú Unibanco: Análise de dados de *E-commerce*

Vinicius Tognetti

Junho, 2024

Sumário

1 Sobre os dados

Os dados são referentes a vendas realizados por um *e-commerce*, são 25000 observações de 14 variáveis referentes a cada compra realiza por cada cliente - todas com 100% de preenchimento. As variáveis são as seguintes:

- **Costumer ID** *(Variável Qualitativa Nominal)*: é a identificação única de cada cliente do *e-commerce* que realizou uma compra.
- **Purchase Date** *(Variável Temporal)*: data em que o cliente realizou a compra.
- **Product Category** *(Variável Qualitativa Nominal)*: categoria do produto comprado. Assume os valores **Books** (livros), **Clothing** (roupas), **Eletronics** (eletrônicos) ou **Home** (doméstico).
- **Product Price** *(Variável Quantitativa Contínua)*: é o preço unitário do produto comprado pelo cliente.
- **Quantity** *(Variável Quantitativa Discreta)*: é a quantidade do produto em questão que o cliente comprou.
- **Total Purchase Amount** *(Variável Quantitativa Contínua)*: é o valor da compra desde a compra atual.
- **NPS** *(Variável Quantitativa Discreta)*: é a nota dada pelo clinete sobre a compra. Assume valores de 0 (pior) a 10 (melhor).
- **Costumer Age** *(Variável Quantitativa Discreta)*: é a idade do consumidor. Assume valores entre 18 e 70 (inclusive).
- **Gender** *(Variável Qualitativa Nominal)*: é o gênero do cliente. Assume valor **Female** (mulher) ou **Male** (homem).
- **Source** *(Variável Qualitativa Nominal)*: é o canal pelo qual o cliente foi redirecionado ao site. Assume valores **FaceBook campaign** (campanha de marketing no FaceBook), **Instagram Campign** (campanha de marketing no Instagram), **Organic Search** (busca orgânica, realizada pelo cliente) ou **SEM** (marketing de mecanismo de busca).
- **Country** *(Variável Qualitativa Nominal)*: é o país do de residência do cliente que realizou a compra. Pode ser EUA ou Canadá.
- **State** *(Variável Qualitativa Nominal)*: é o estado de residência do cliente que realizou a compra. Assume como valores os estados de EUA e de Canadá.
- **Latitude** *(Variável Geográfica)*: é a latitude do estado do cliente.
- **Longituite** *(Variável Geográfica)*: é a longitude do estado do cliente.

Antes de iniciar a análise dos dados, já podemos ter algumas ideias de como conduzí-la para responder às perguntas propostas.

Primeiro, para identificar os produtos mais vendidos nos últimos 3 anos, podemos filtrar a base por data de compra, agrupar por produto e somar as quantidade vendidas de cada um deles.

Segundo, para encontrar o(s) produtos mais caro(s) e o(s) mais barato(s), podemos encontrar aquele(s) em que o valor unitário do produto é o máximo presente no banco de dados - análogo para o mais barato.

Terceiro, para encontrar as categorias mais e menos vendidas e as mais e menos caras, podemos agrupar por categorias realizando (1) a soma das quantidades e (2) a média dos valores unitários dos produtos e encontrando os valores máximos e mínimos.

Quarto, para descobrir o(s) produto(s) com o melhor e o pior NPS, podemos prosseguir da mesma maneira que no problema de encontrar os produtos mais caros, só que usando como critério o NPS.

Quinto - por último - para descobrir o tipo de público (em gênero e idade) e o canal ideal para vender determinado produto, podemos, primeiro, criar uma variável do total da compra por cliente - fazemos isso multipicando o valor unitário do produto pela quantidade comprada - chamaremos essa variável de **ticket**, depois agruparemos por ID de cliente usando a média (calculando o ticket médio de cada cliente). Em seguida, podemos visualizar graficamente o impacto das interações entre gênero, idade e tipo de produto no ticket médio para identificar se seria mais apropriado clusterizar os clientes de acordo com as categorias de produtos ou analisar separadamente cada uma das categorias usando regressão múltipla..

Essa é a linha de pensamento que pretendo seguir aqui, seus desdobramentos (caso tenham) seram descritos dentro da seção referente a cada resposta.

Vamos lá.

2 Respostas às perguntas

2.1 Quais os produtos mais vendidos considerando os últimos 3 anos?

Realizando os filtros necessários no banco de dados, encontram-se 35091 vendas que tiveram a quantidade máxima (5). Se os agrupar por categoria temos a seguinte tabela e

gráfico de barras.

Categoria do Produto	**Quantidade Vendida nos Últimos 3 anos**
Clothing (Vestuário)	52865
Books (Livros)	52275
Eletronics (Eletrônicos)	35230
Home (Domésticos)	35085

Tabela 1: Tabela com a quantidade de produtos vendidos nos últimos 3 anos por categoria.

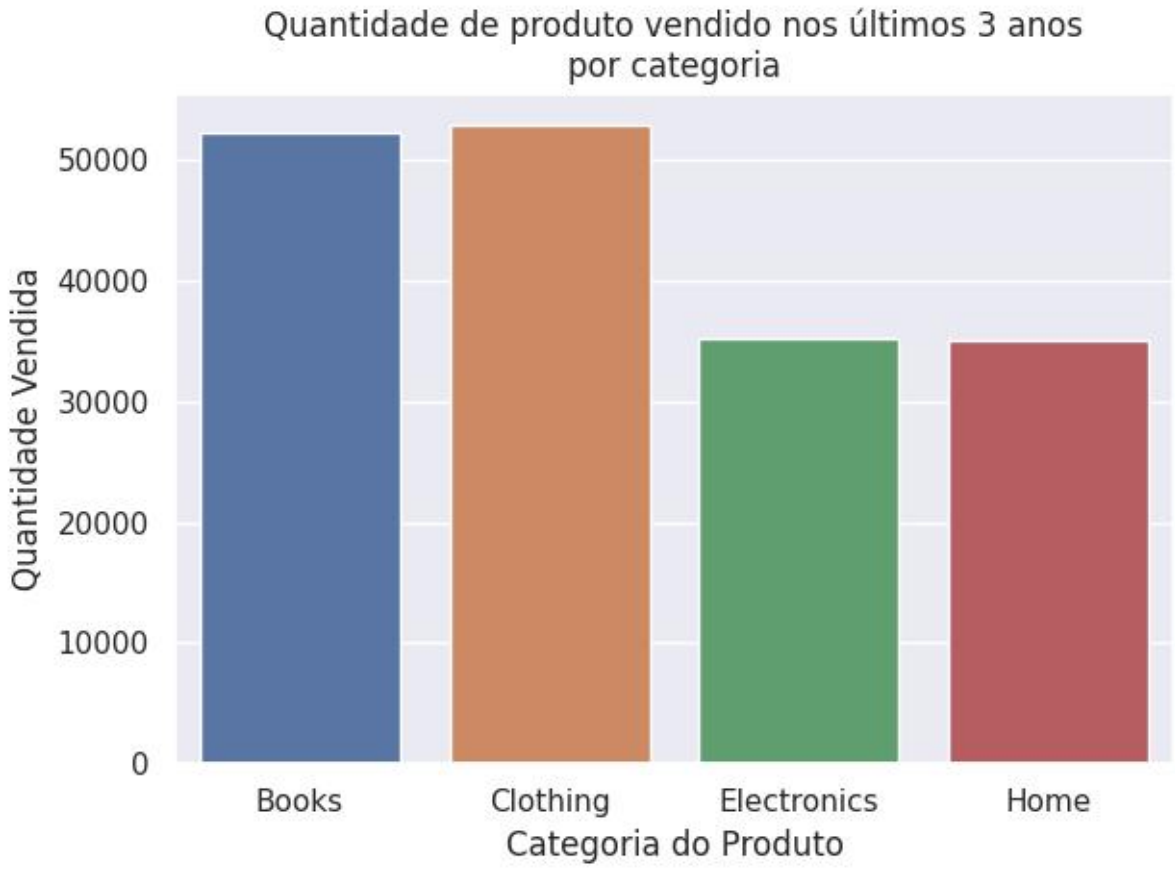

Figura 1: Gráfico de barras com as quantidades de produtos vendidos por categoria nos últimos 3 anos.

E podemos ver que os produtos mais vendidos nos últimos 3 anos foram artigos de

vestuário.

2.2 Qual o produto mais caro e o mais barato?

Realizando a contagem de quantos vendas foram concluídas pelos valores máximo e mínimo obtemos 474 e 493, respectivamente. Vamos, então, agrupá-las por categoria. Fazendo isso, obtemos a seguinte tabela e gráfico.

Categoria do Produto	**Mínimo**	**Máximo**
Clothing (Vestuário)	10	500
Books (Livros)	10	500
Eletronics (Eletrônicos)	10	500
Home (Domésticos)	10	500

Tabela 2: Tabela com os valores máximos e mínimos em cada categoria de produto.

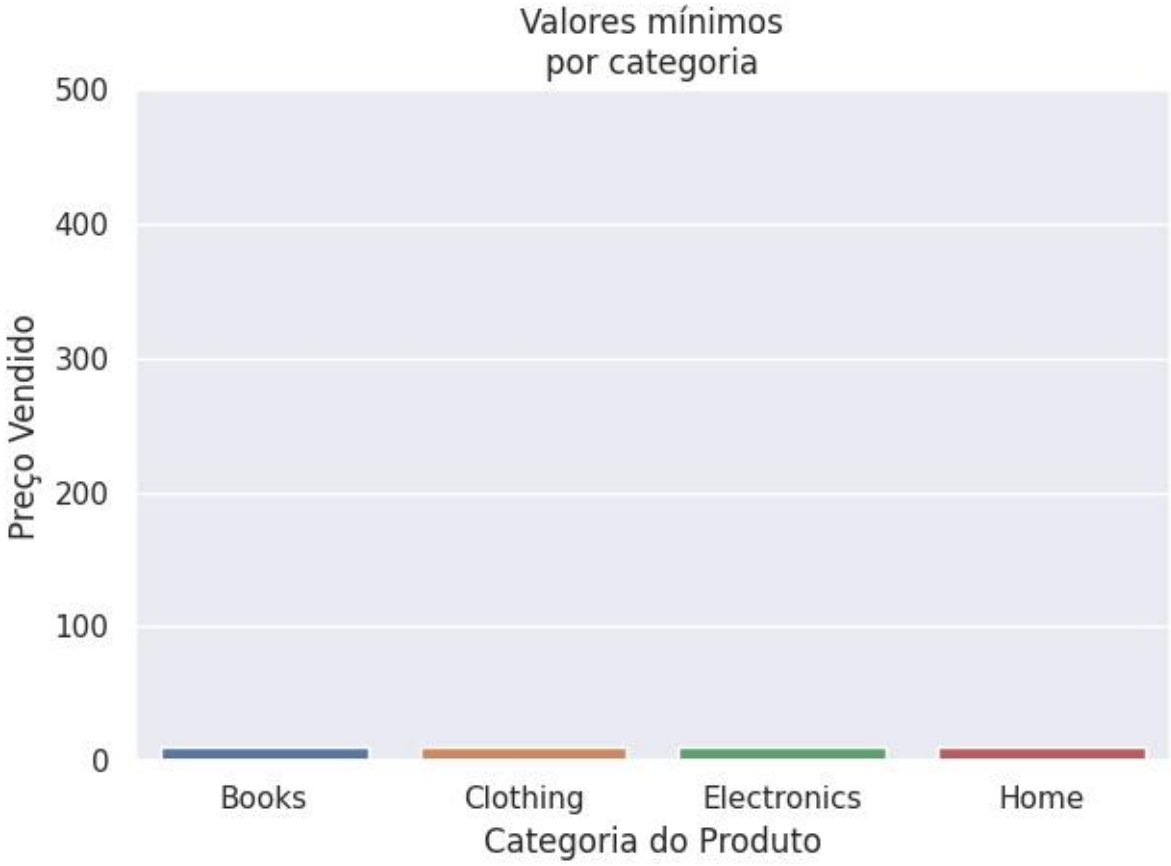

Figura 2: Gráfico de barras com os valores mínimos por categoria.

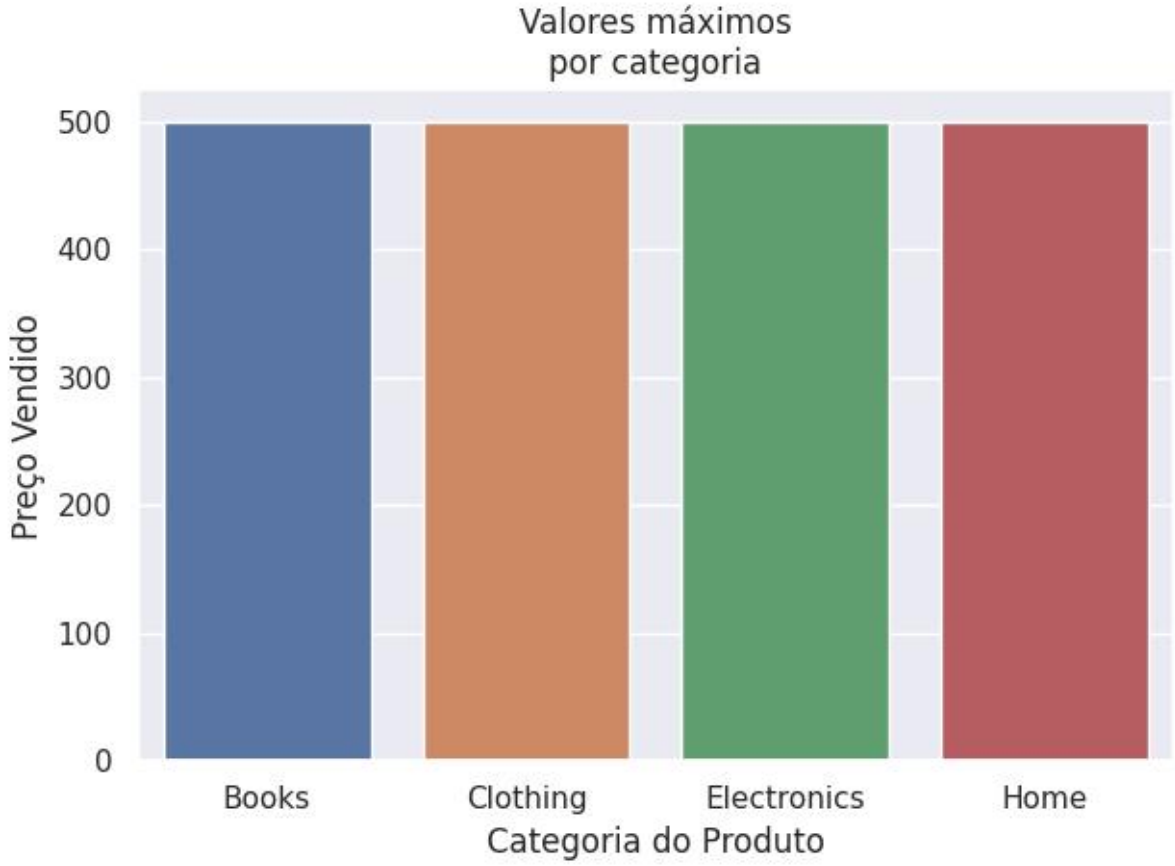

Figura 3: Gráfico de barras com os valores máximos por categoria.

Como podemos ver pela tabela e pelos gráficos, os produtos mais caros e mais baratos estão presentes em todas as categorias (aqui podemos perceber, também, que provavelmente os dados foram gerados).

2.3 Qual a categoria de produto mais vendida e menos vendida? Quais as categorias mais e menos caras?

Agrupando por categorias e calculando (1) a média dos preços e (2) a quantiade total por categoria, obtemos a seguinte tabela e gráficos.

Categoria do Produto	Preço Médio	Quantidade Total
Clothing (Vestuário)	254	17357
Books (Livros)	255	17117
Eletronics (Eletrônicos)	255	11268
Home (Domésticos)	255	11351

Tabela 3: Tabela com os preços e quantidades médias por categoria de produto.

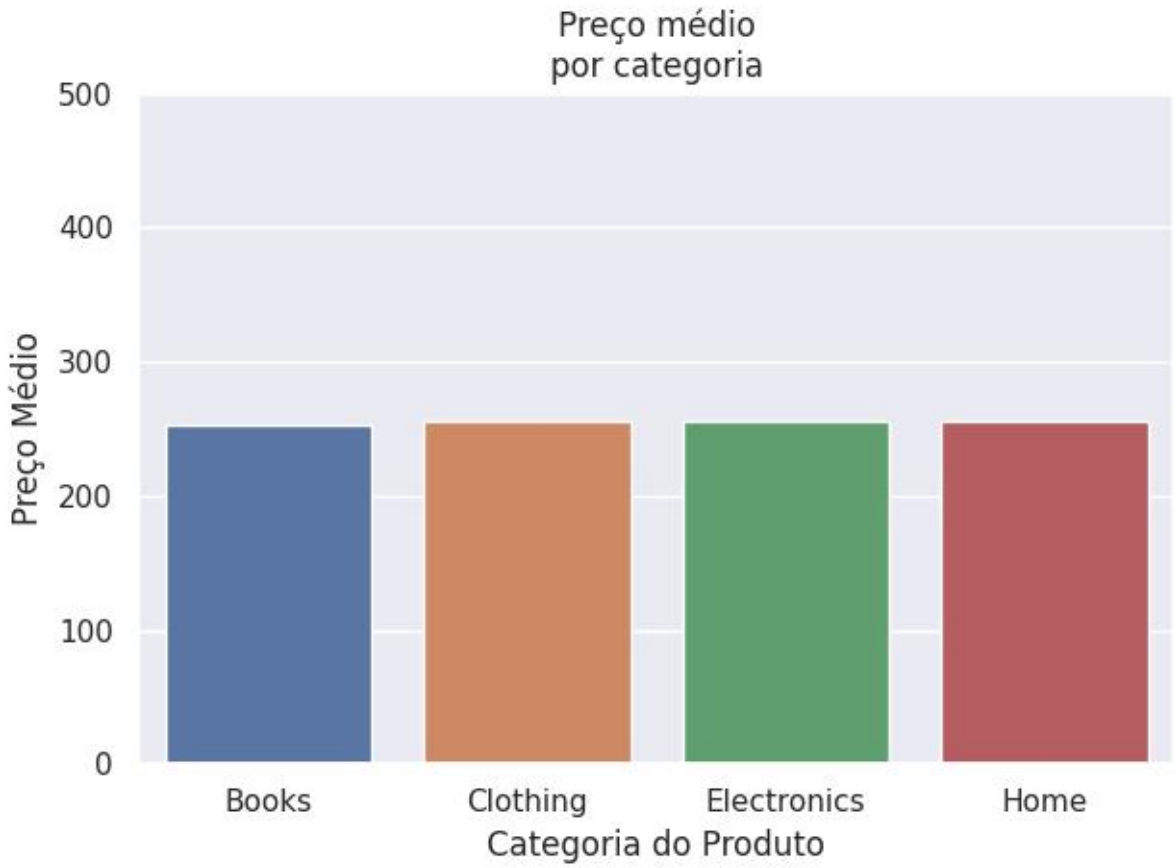

Figura 4: Gráfico de barras dos preços médios por categoria.

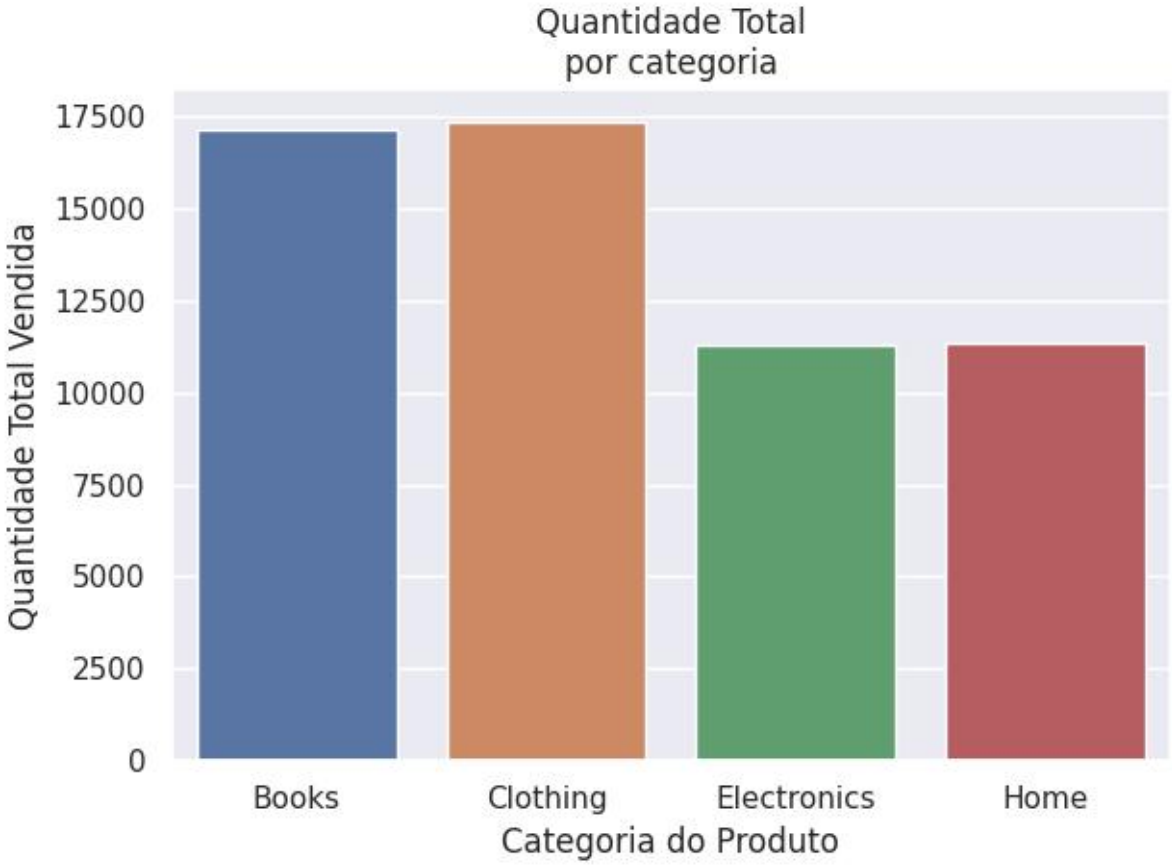

Figura 5: Gráfico de barras das quantidades totais por categoria.

Como podemos observa na tabela e nos gráficos, a categoria mais vendida é a de vestuário (Clothing), e, como os preços médios por categoria são iguais, não há uma categoria mais cara ou mais barata.

2.4 Qual o produto com o melhor e o pior NPS?

Agrupando os NPS médios por categoria temos a seguinte tabela e gráfico.

Categoria do Produto	NPS
Clothing (Vestuário)	5
Books (Livros)	5
Eletronics (Eletrônicos)	5
Home (Domésticos)	5

Tabela 4: Tabela com os NPS médios por categoria do produto.

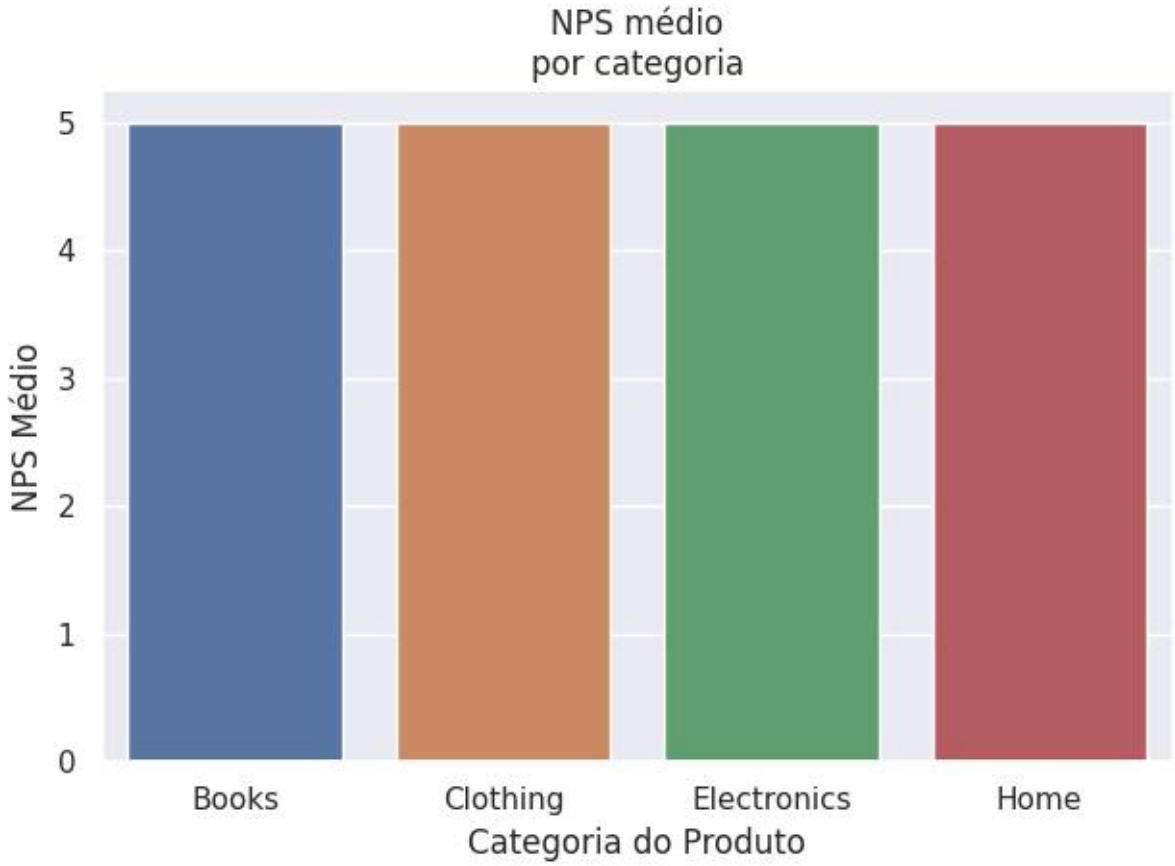

Figura 6: Gráfico de barras do NPS médio por categoria de produto.

Como podemos ver, os NPS médios são iguais, então não á um porduto com o melhor ou pior NPS.

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

Aqui a ideia é que temos várias compras para cada cliente, então vamos calcular o valor total de cada compra (ticket) e calcular o valor média de compra pra cada cliente (ticket médio). Com base nas informações sobre cada cliente (gênero, idade e canal) vamos procurar entender qual delas têm mais influência no ticket médio - bem como qual o valor ótimo simultâneo delas - a fim de determinar a configuração que máximiza o ticket médio para cada categoria de produto.

Vamos começar procurando indícios de que há ou não efeito de cada uma das informações sobre o cliente individualmente no ticket médio.

Vejamos um gráfico do ticket médio por idade para cada uma das categorias de produto.

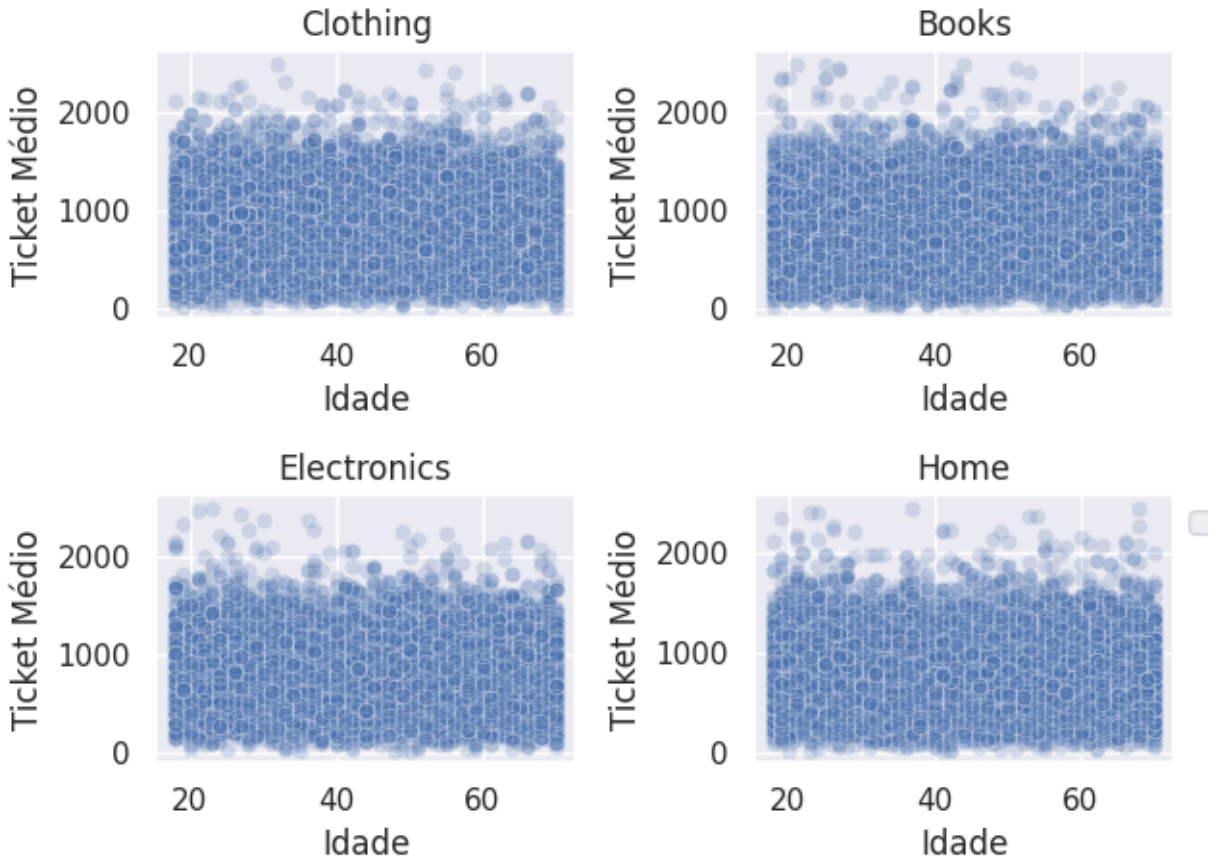

Figura 7: Gráficos de dispersão entre a idade e o ticket médio por categoria de produto.

Note que os valores do ticket médio não alteram conforme a idade muda - bem como

não exibem nenhum padrão - o que nos indica que não existe uma relação entre o ticket médio e a idade para nenhuma das categorias de produtos.

Agora, investiguemos a relação entre ticket médio e gênero.

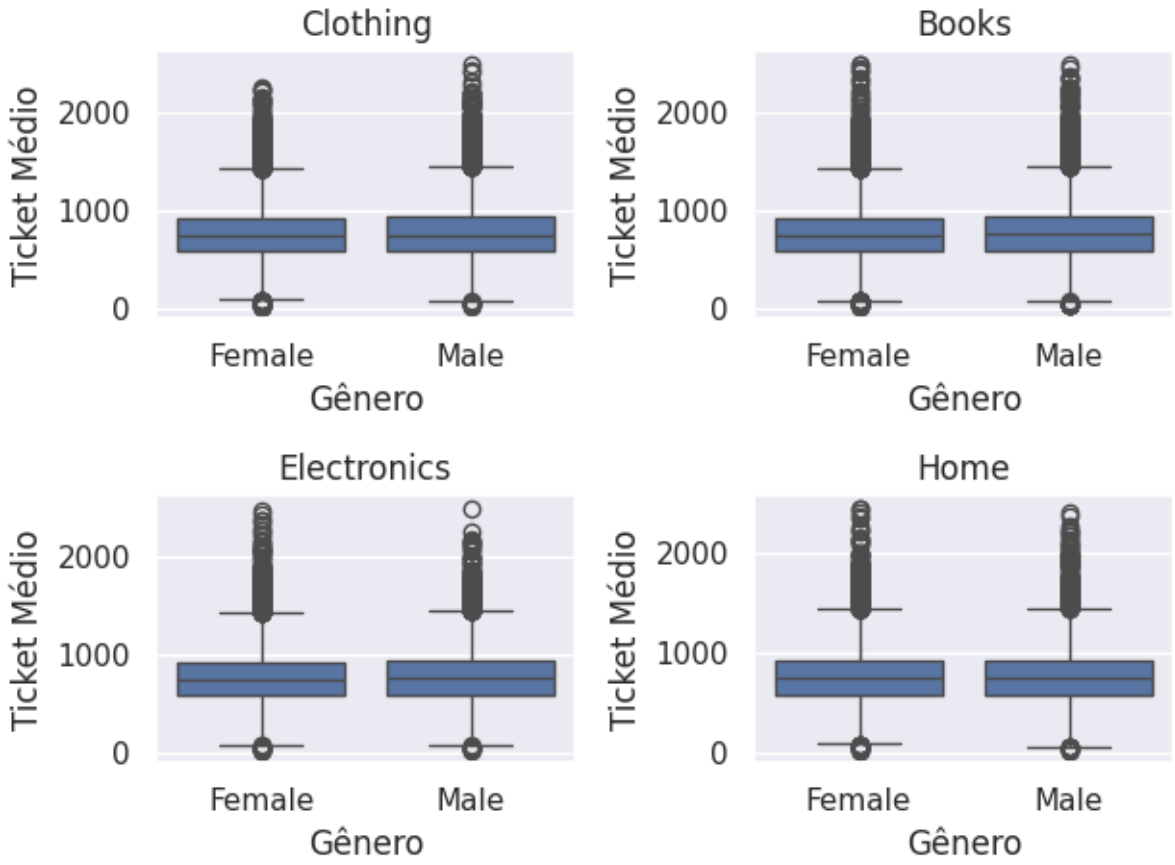

Figura 8: Boxplots do ticket médio separado por gênero para cada categoria de produto.

Podemos ver aqui que, para todas as categorias de produtos, os boxplots para os dois gêneros têm comportamentos praticamente iguais, o que indica que também não há relação entre o gênero e o ticket médio.

Por fim, vejamos ticket médio e canal.

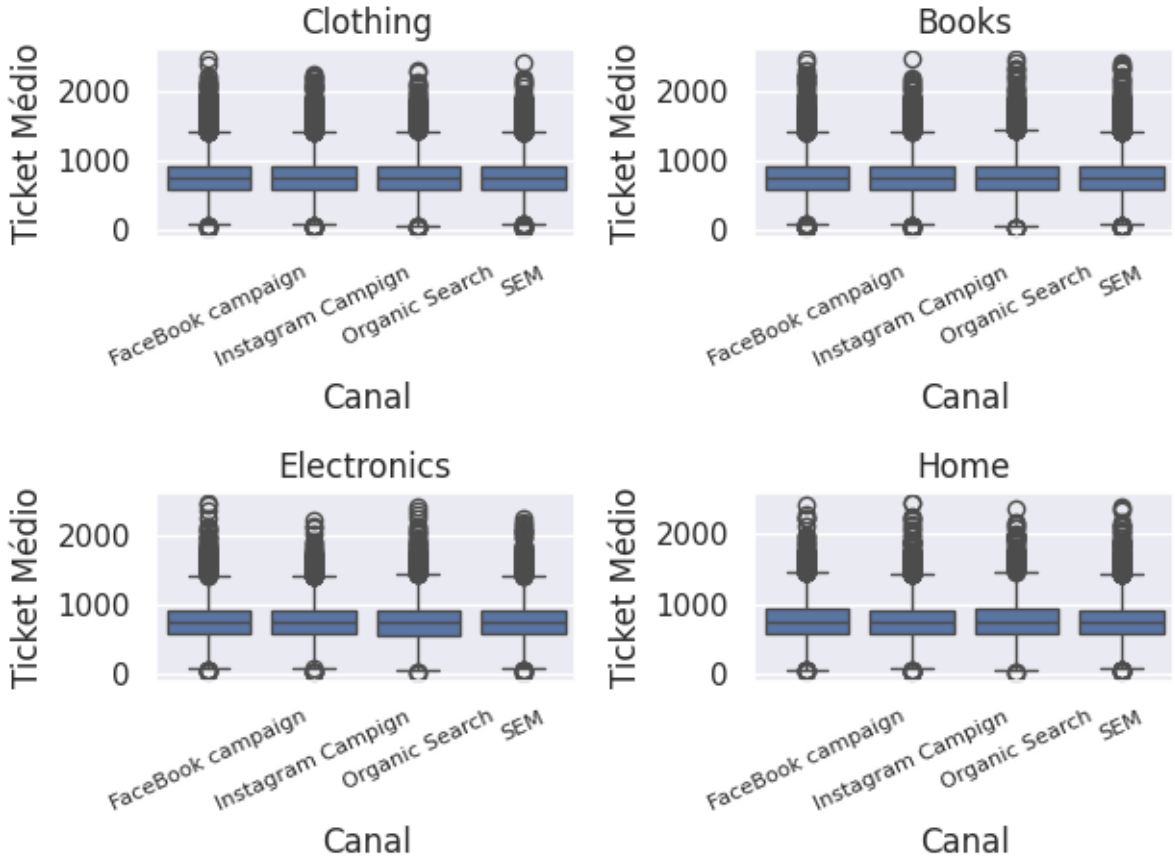

Figura 9: Boxplots do ticket médio separado por canal para cada categoria de produto.

Aqui, também, os boxplots tem comportamentos bem parecidos para cada canal dentro de cada categoria de produto, o que nos indica que não há uma relação entre o canal e o ticket médio.

Assim, para as variáveis observados individualmente, não há indícios de que haja qualquer relação entre elas e o ticket médio.

Seguiremos para investigar se as interações gênero-idade, gênero-canal e canal-idade têm alguma relação com o ticket médio (Obs: é possível que a interação entre covariáveis tenha efeito na variável resposta mesmo que individualmente não exista efeito. o mesmo vale para a interação entre as três, que investigaremos mais a frente também).

Começando pela interação gênero-idade.

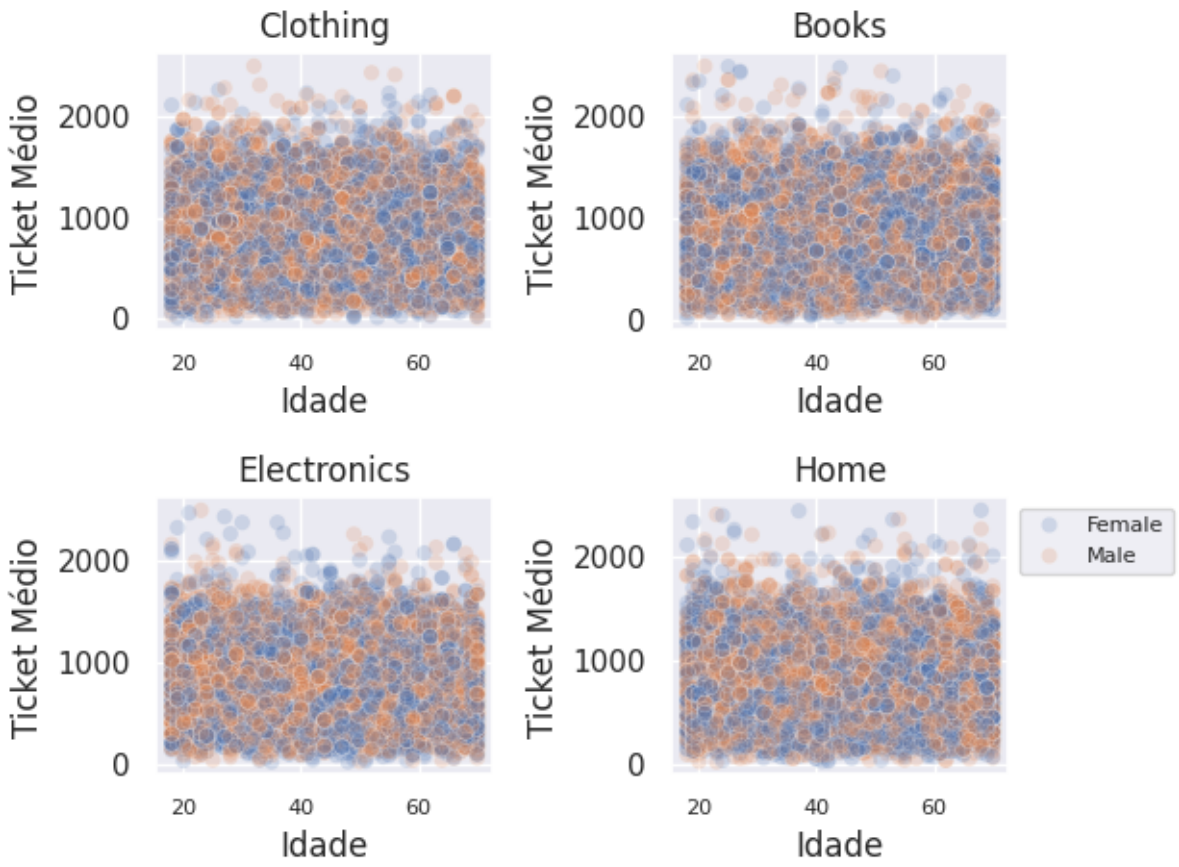

Figura 10: Gráficos de dipersão visualizando a relação entre a interação gênero-idade e o ticket médio.

Aqui temos os pontos do gráfico de dispersão coloridos de acordo com o gênero do cliente. Note que não se apresenta nenhum padrão no gráfico, o que indica que não há relação entre a interação gênero-idade com o ticket médio.

Caso houvesse algum tipo de relação observaríamos algo como os pontos em azul em torno de um valor de ticket médio diferente dos em laranja ou os pontos em azul formando uma relação linear positiva com o ticket médio enquanto os em laranja formariam uma relação linear negativa, por exemplo. Nada desse tipo ocorre nesses gráficos.

Agora, vejamos a interação gênero-canal.

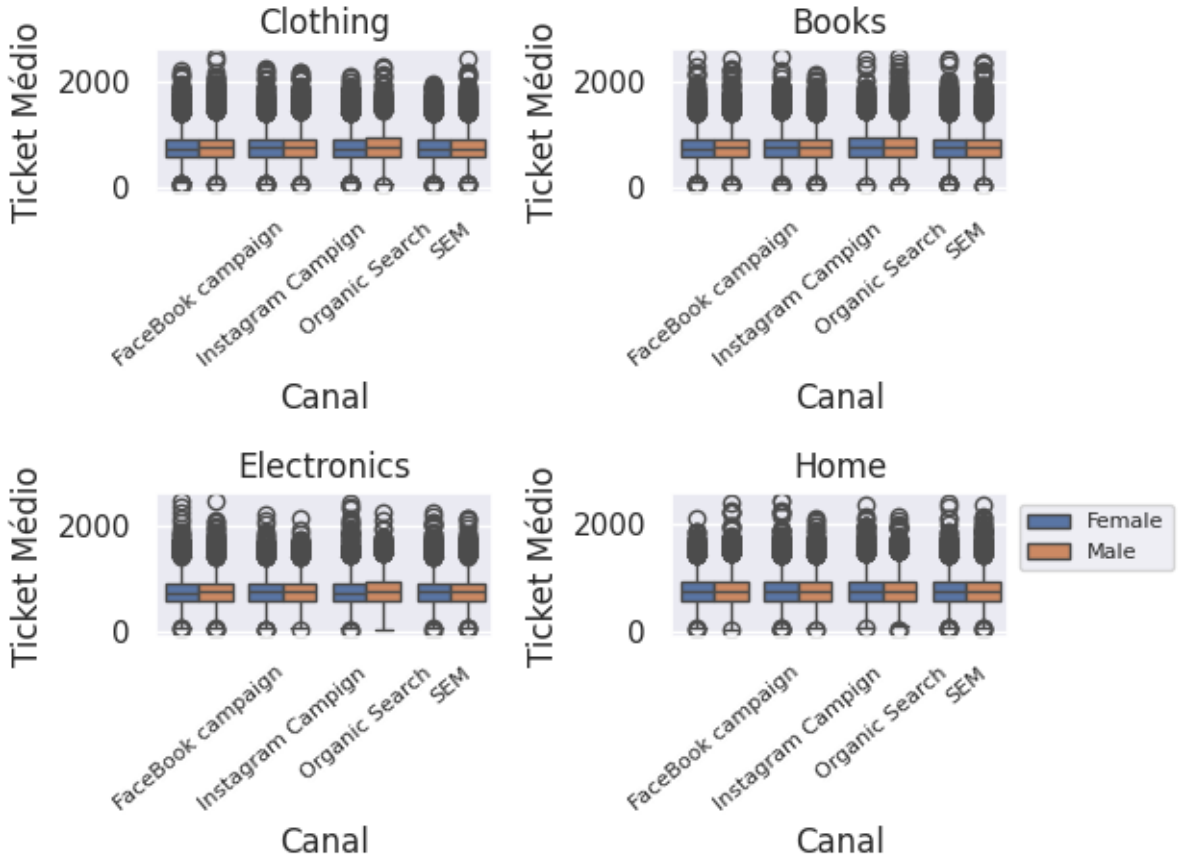

Figura 11: Boxplots dos tickets médios por categoria de produto, agredados por canal e gênero.

Interpretando o gráficos, podemos ver que para todas as categorias os boxplots dos dois gêneros dentro de cada um dos canais têm o mesmo comportamento, o que implica que não há indícios de que haja qualquer relação entre a interação canal-gênero e o ticket médio.

Por último, vejamos a interação canal-idade.

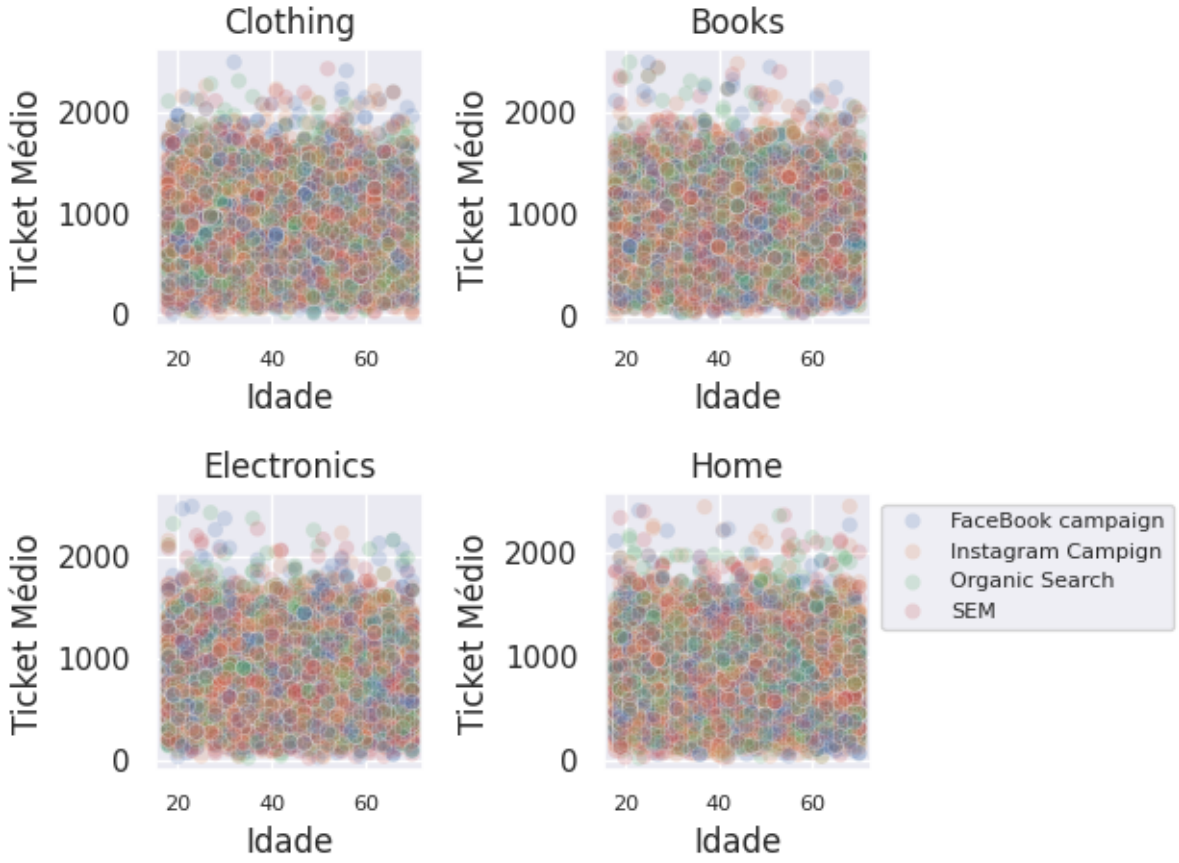

Figura 12: Gráficos de dipersão visualizando a relação entre a interação canal-idade e o ticket médio.

Aqui temos os pontos coloridos de acordo com o canal. Como ocorreu na interação gênero-idade, aqui também os gráficos não exibem qualquer padrão visual, sugerindo que não há relação entre o ticket médio e a interação canal-idade.

Finalizando a análise gráfica, veremos a interação gênero-canal-idade.

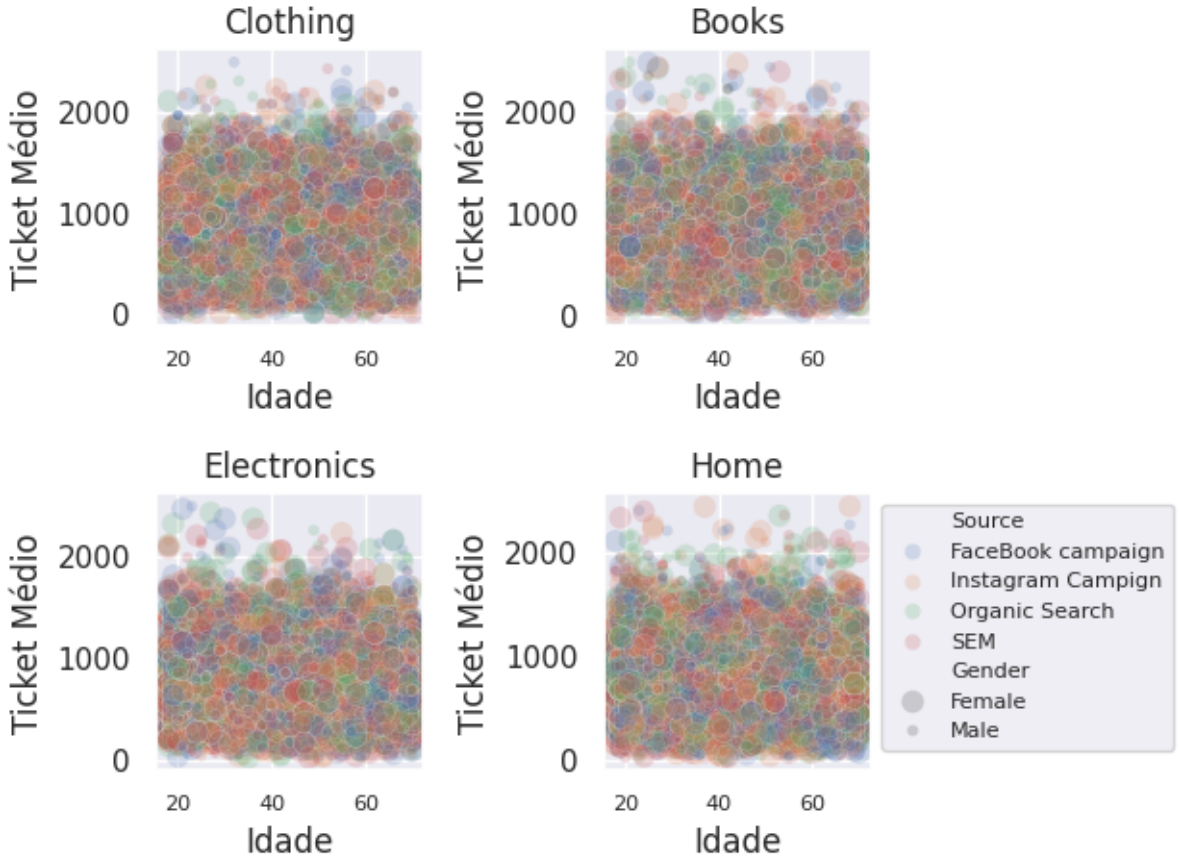

Figura 13: Gráficos de dipersão visualizando a relação entre a interação gênero-canal-idade e o ticket médio

Aqui temos os pontos coloridos por canal e seu tamanho codificado de acordo com o gênero do cliente. Veja que não há padrões entre os pontos, as cores e os tamanhos, o que nos indica que a interação gênero-canal-idade não tem relação com o ticket médio.

De acordo com isso tudo que foi observados nos gráficos, não há indícios de que exista alguma diferença no ticket médio que dependa das informações sobre o cliente, é dizer, tanto faz qual a combinação entre gênero, idade e canal para cada categoria de produto pois parece que o ticket médio esperado será o mesmo para qualquer situação possível.

Realizando um regressão linear múltipla tomando como resposta o tiket médio e como covariáveis a idade, o gênero e o canal, obtemos a seguinte tabela de p-valores dos coeficientes. A referência é gênero feminino e canal FaceBook.

	Coeficiente	**Clothing**	**Books**	**Home**	**Eletronics**
	Constante	0	0	0	0

Tabela 5 continuada da página anterior

	Coeficiente	Books	Clothing	Eletronics	Home
Idade	Idade	0.260	0.973	0.641	0.477
Gênero	Masculino	**0.031**	0.460	0.241	0.129
Canal	Instagram	0.580	0.757	0.746	0.875
	Orgânico	0.638	0.363	0.475	0.722
	SEM	0.112	0.535	0.685	0.284
Gênero-Idade	Masculino-Idade	0.057	0.392	0.334	0.134
Canal-Idade	Insta-Idade	0.077	0.579	0.964	0.934
	Org-Idade	0.961	0.548	0.819	0.523
	SEM-Idade	0.107	0.647	0.901	0.262
Canal-Gênero	Insta-M	0.125	0.482	**0.054**	0.734
	Org-M	0.820	0.437	0.821	0.490
	SEM-M	0.661	0.829	0.874	0.373
Canal-Gênero-Idade	Insta-M-Idade	0.167	0.322	0.088	0.552
	Org-M-Idade	0.835	0.643	0.931	0.259
	SEM-M-Idade	0.750	0.767	0.609	0.303

Tabela 5: Tabela com os p-valores dos coeficientes da regressão linear.

Essa tabela representa os p-valores das hipóteses

$$H_0 : \text{não há efeito da covariável no ticket médio}$$
$$H_1 : \text{há efeito da covariável no ticket médio}$$

Com os p-valores mais significativos destacados.

Daí, auxiliado dos resultados obtidos em cada regressão, podemos ver que

- Para os livros, é melhor focar no público masculino (independente de idade e canal). Apresentam um ticket médio R$ 28 maiores que o público feminino, em média.
- Para os domésticos, é mais interessante **evitar** o público masculino obtido via Instagram (independente da idade). Apresentam um ticket médio R$ 40 **menor** que o público masculino obtido por outros canais e o público feminino obtido via quaisquer canais (independente da idade), em média.
- É interessante ressaltar que o ticket médio em geral (com exceção dos dois casos citados aqui) é entre R$ 741 e R$ 766.

Para as outras categorias, não existe - baseado na amostra coletada - uma combinação ótima de características de cliente para melhor/piorar o ticket médio.

Seguem as regressões para cada categoria. ($Y_{\text{Categoria}}$ indica o valor esperado do ticket médio da categoria).

(Livros) $Y_{\text{Livros}} = 741.51 + 27.61\mathbb{I}(\text{Gênero} = \text{Masculino})$

(Vestuário) $Y_{\text{Vestuário}} = 765.25$

(Domésticos) $Y_{\text{Domésticos}} = 766.02 - 40.54\mathbb{I}(\text{Gênero} = \text{Masculino} \ \& \ \text{Canal} = \text{Instagram})$

(Eletrônicos) $Y_{\text{Eletrônicos}} = 755.78$