BlueEdTech
Bootcamp de Dados

Empresa Parceira: HVAR

Relatório da Sprint 1: Análise exploratória e levantamento de hipóteses

Equipe: Os Outliers

Integrantes:
Antonio Duarte Marcos Junior
Lucilio Fontes Moura
Gustavo Ramos Abreu

Setembro de 2022

SUMÁRIO

1 Introdução

A precificação dinâmica ou precificação diferencial é uma estratégia por parte das empresas para precificação de um produto para clientes distintos a fim de maximizar os lucros. Há décadas é utilizada por companhias aéreas e hotéis, entretanto, esses preços eram estabelecidos durante uma temporada e durante um longo tempo [1], mas recentemente, com a inteligência artificial, big data os preços estão sendo definidos de forma automática, com base no banco de dados da empresa e no perfil do usuário, com a crescente acesso ao uso de computadores e celulares, isso se torna cada vez mais necessário, como mostra o gráfico abaixo acerca do uso de celulares [2]:

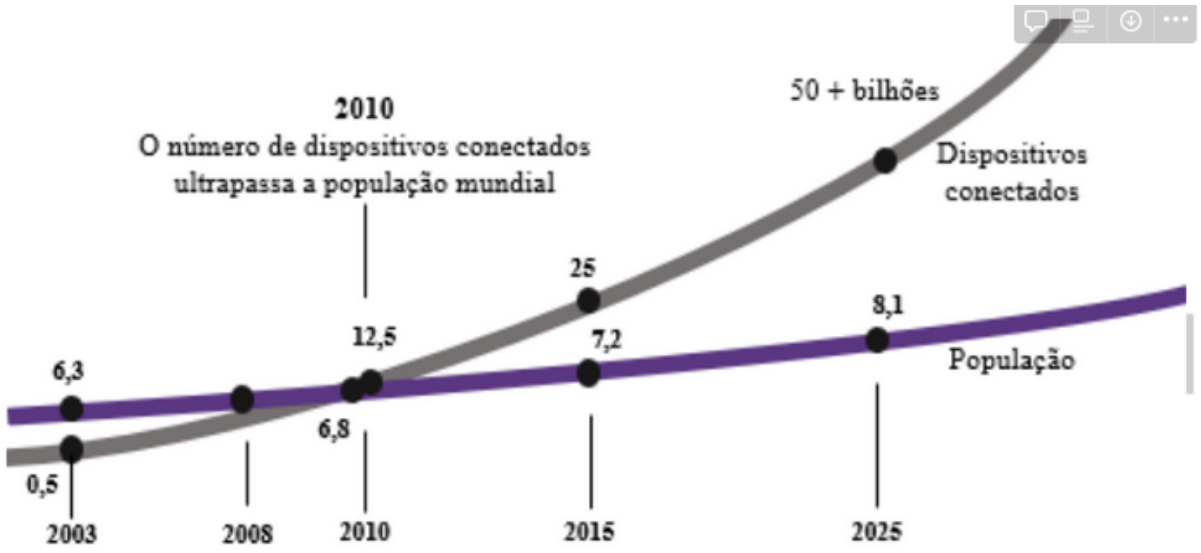


Figura 3 – Crescimento exponencial dos dispositivos conectados
Fonte: Statistitches Bundesamt; Deutsche Bundesbank; Prognos; Thomas Nippperde; McKinsey. In: Seminário Growth for Strategies, 2017, p. 5.

Entre as principais vantagens, temos:

- Busca por melhor rentabilidade;
- Previsão de cenários de venda, rentabilidade e giro de estoque, tendo assim uma maior rapidez;
- Mapeamento do preço da concorrência;
- Gerar dados aos gestores, foco nos itens mais vendidos.

Ou seja, ela é adequada à natureza temporal e se baseia no princípio econômico de ajuste entre oferta e demanda, possibilitando a maximização dos lucros com a obtenção de melhores margens em períodos de alta demanda, e melhor aproveitamento da capacidade instalada em períodos de baixa demanda [1].

2 Levantamento de hipóteses

2.1 Hipótese 1: Os preços dos produtos anunciados no site têm relação com a data do anúncio?

Existem alguns produtos com forte sazonalidade, como roupas de estação, e tal comportamento pode ter efeito direto no preço dos produtos. É esperado que os produtos da estação tenham maior procura e consequentemente preços maiores do que quando estão fora da estação com menor procura. Com base nestes conceitos será investigado se há produtos no dataset tem sazonalidade marcante nos preços.

2.2 Hipótese 2: Qual o conjunto de atributos mais relevantes para a precificação de um produto?

Dentre os atributos disponíveis no dataset deve existir uma melhor combinação de parâmetros com maior influência na precificação dos produtos. Será investigado então quais são estes atributos. A investigação irá identificar um conjunto de atributos ótimo, reduzindo a quantidade de variáveis para se trabalhar e diminuição da complexidade dos modelos.

2.3 Hipótese 3: Os preços são influenciados tanto pelo tempo quanto pelos atributos do produto?

A hipótese 1 visa avaliar a variação do preço de um produto apenas em função do tempo e a hipótese 2 apenas em função das características do produto. Porém, é possível que a combinação das duas metodologias possa produzir melhores resultados que o uso de cada uma individualmente.

2.4 Hipótese 4: A marca é um atributo importante para a definição do preço de um produto?

Produtos com características semelhantes podem ter faixas de preços bem distintas devido ao atributo marca. Separar e agrupar os produtos por marcas, em especial aquelas que agregam mais valor aos produtos, pode ser uma boa prática para precificar os produtos.

2.5 Hipótese 5: As condições de um produto são relevantes para precificar um produto?

É esperado que produtos em melhor estado de conservação tenham preços maiores do que os produtos mal conservados. Este produto possui produtos novos e usados, então é importante poder diferenciá-los. Produtos de coleção e vintage em bom estado de conservação podem até mesmo ter preços maiores do que os de mesmas categorias classificados como novos.

3 Base de dados

A base de dados utilizada é formada por anúncios de produtos do Mercari Price Suggestion Challenge do Kaggle. O dataset contém 1.481.661 anúncios de 1.081.584 de produtos. O Mercari, é um dos maiorres aplicativos de compras do Japão. Com este desafio eles gostariam de oferecer sugestões de preços aos vendedores, mas isso é difícil porque seus vendedores podem colocar praticamente qualquer coisa, ou qualquer pacote de coisas, no mercado da Mercari.

Dicionário de dados

Variável	**Descrição**
train_id	Identificador único de cada anúncio.
name	Título do anúncio. Foram removidas as informações de preços dos anúncios para evitar vazamento de informação. Os preços removidos foram preenchidos com [rm].
item_condition_id	Condição do item fornecido pelo vendedor. Têm valores variados entre 1 e 5 com a seguinte descrição: 1 - ótimo; 2 - bom; 3 - regular; 4 - ruim; 5 - Péssimo.
brand_name	Marca do produto.
category	Lista de categorias às quais o produto pertence.
price	Preço de venda do produto em dólares.
shipping	1 - caso a taxa de envio seja paga pelo vendedor. 0 - caso a taxa de envio seja paga pelo consumidor.
item_description	Deve conter descrição detalhada do produto. Foram removidas as informações de preços das descrições para evitar vazamento de informação. Os preços removidos foram preenchidos com [rm].

4 Análise exploratória

4.1 Inconsistência dos dados

Variável “name”

A variável “name” deveria conter apenas os nomes dos produtos. Mas os usuários a utilizam como título do anúncio e assim colocam nomes pouco descritivos dos produtos como “bundle” (pacote), “brand new” (marca nova). Alguns dos valores contém emoticons. Foi verificado que há valores repetidos, porém com grafias diferentes, por exemplo: bundle, Bundle, BUNDLE.

Para tratar estas inconsistências foram aplicados os seguintes passos:

1º Converter todos os “names” para apenas minúsculas.

2º Foram removidos todos os caracteres especiais, e acentos.

3º Alguns usuários colocaram o preço no produto já no anúncio do produto. Quanto isto ocorreu o fornecedor do dataset substituiu o preço do produto por [rm]. Considerou-se que esta informação não é relevante para a proposta do projeto e as ocorrências de [rm] foram removidas dos nomes dos anúncios.

Variável “category”

Os anúncios do dataset podem ter até 5 categorias diferentes. Como 95% dos itens possuem apenas 3 categorias, para facilitar as análises futuras serão consideradas as três primeiras categorias de cada anúncio. A coluna “category” foi então dividida em três outras colunas: category_1, category_2 e category_3.

Foi observado que 42% dos anúncios não possuem informação de categoria. Como esta é uma parcela com tamanho significativo e não temos controle sobre como o usuário fornece esta informação, optou-se por manter-se estes registros sem categoria. Porém os valores nulos foram preenchidos com o termo “No Category”.

Coluna “brand_name”

Na coluna “brand_name” também há um quantitativo relevante de valores nulos. Optou-se por manter-se os dados, pelos mesmos motivos da coluna “category”, e os valores nulos foram preenchidos com “No Brand”.

Coluna “item_description”

A coluna “item_description” deveria conter a descrição técnica do produto, mas, novamente, como não há controle sobre como o usuário preenche esta informação, há valores pouco relevantes. Algumas descrições são preenchidas apenas com emoticons. O tratamento para estas colunas seguiu metodologia semelhante ao da coluna name:

1º Colocar todos os nomes em apenas minúsculas.

2º Remover caracteres especiais e pontuação.

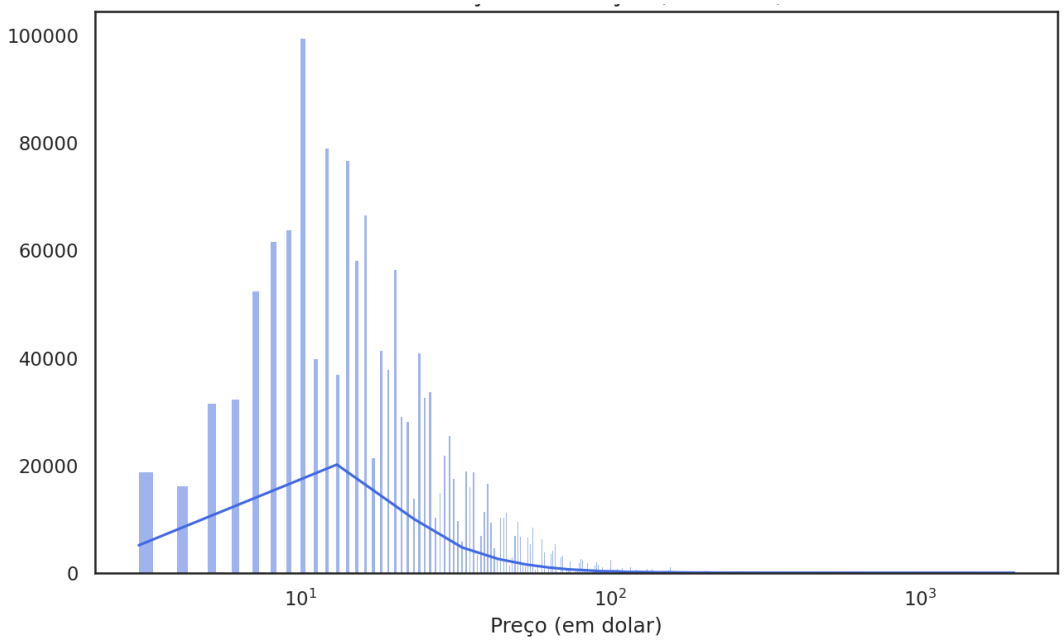
3º Comentários vazios foram preenchidos com “No comment yet”.

4.2 Análise dos dados

Quanto ao título dos anúncios, idealmente os anunciantes deveriam colocar um texto descritivo do produto. Porém, como os anunciantes são livres para preencherem o que quiserem, muitos dos títulos têm textos genéricos. Um exemplo são os anúncios com o termo “bundle” (pacote) sendo o com maior número de registros com 3.455 anúncios.

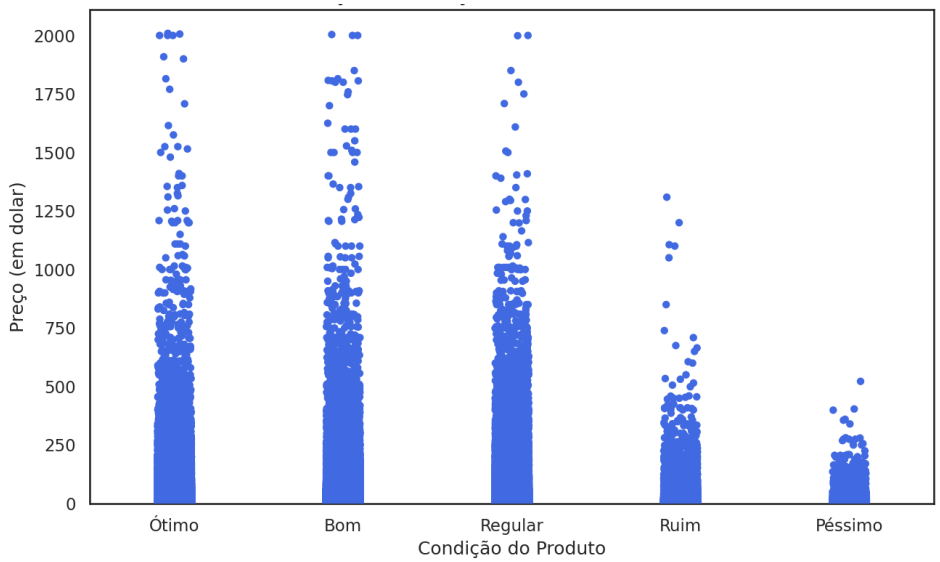
A Figura 1 mostra o histograma de preços dos produtos, devido à grande dispersão dos dados for necessário utilizar a escala logar timica no eixo x para melhor visualização dos dados. Os preços dos anúncios estão concentrados da faixa de valores até 100 dólares, como pode ser na Figura. Contudo, existem produtos que chegam a mais de 2000 dólares. Este resultado sugere que o forco de mercado da empresa são os produtos de populares de baixo preço.

Figura 1 - Distribuição dos Preços dos Produtos



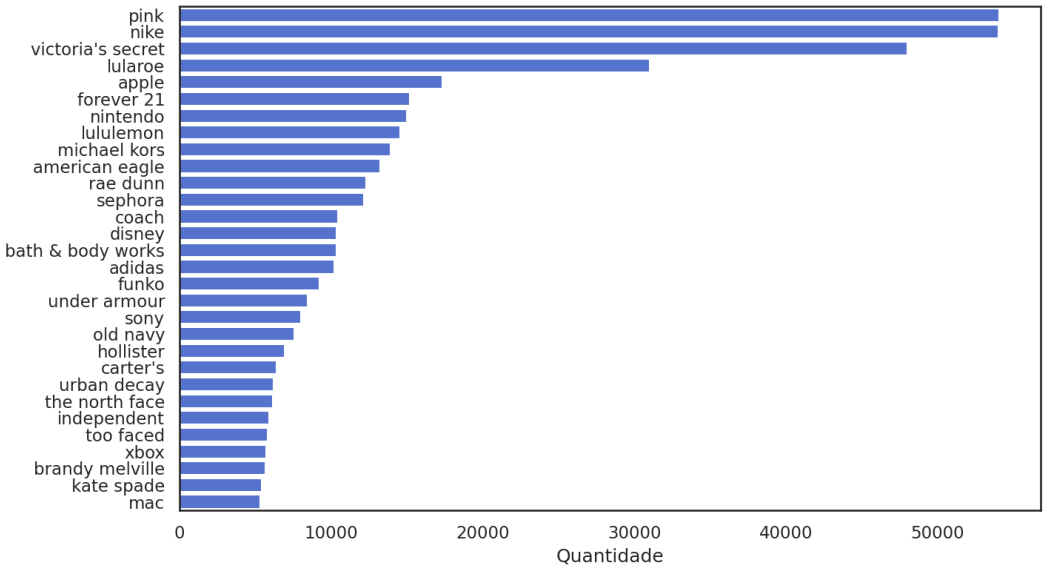
A Figura 2 mostra a relação existente entre o preço do produto e sua qualidade. Como é de se esperar, os produtos de maior qualidade tendem a ter preços maiores que superiores. Os itens anunciados, com condição entre ótimo e regular, concentram seus preços na faixa de até 150 dólares, com alguns itens chegando a casa dos 2.000 dólares. Já os itens com pior qualidade (ruim, péssimo) concentram-se na faixa de até 250 dólares. O gráfico da Figura 1 também indica que a maioria dos anúncios são de produtos com qualidade entre Ótimo e Regular.

Figura 2 - Distribuição dos preços pela qualidade do produto.



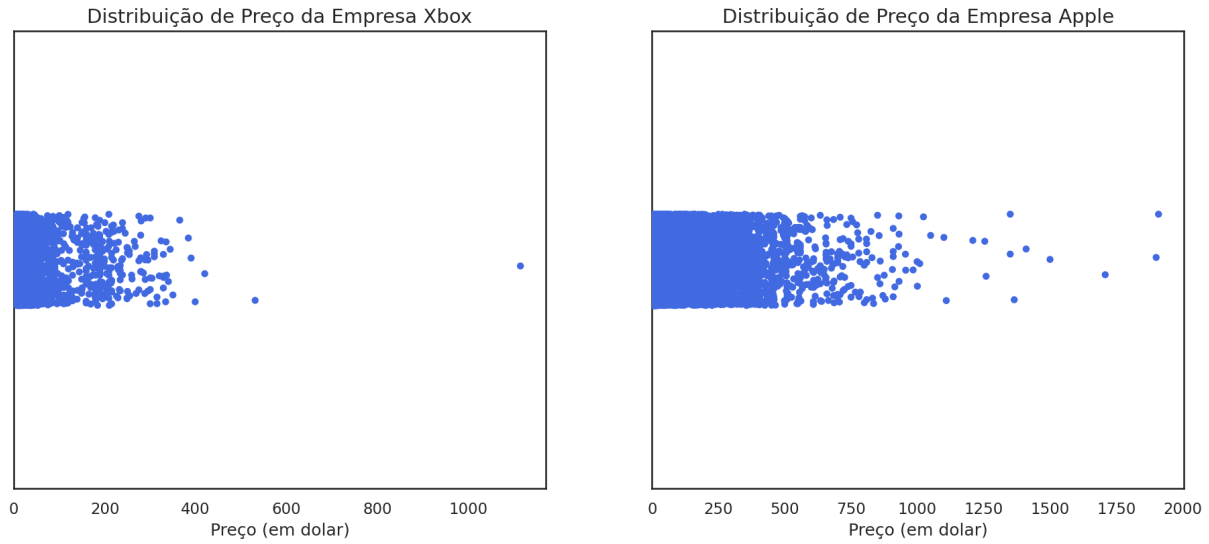
No gráfico da Figura 3 são exibidas as marcas mais comuns do dataset. Às três principais marcas, considerando-se a quantidade de anúncios, são Pink, Nike e Victoria's Secret. No geral, as principais marcas no dataset são relacionadas a itens de moda e produtos do lar. Estes dados sugerem então que o principal mercado de atuação da empresa é o de moda.

Figura 3 - As 30 marcas com maior quantidade de anúncios



das 30 empresas mais populares do dataset form selecionadas algumas para verificar a distribuição de seus preços, onde no exemplo da Figura 4 estão as empresas Xbox e Apple.

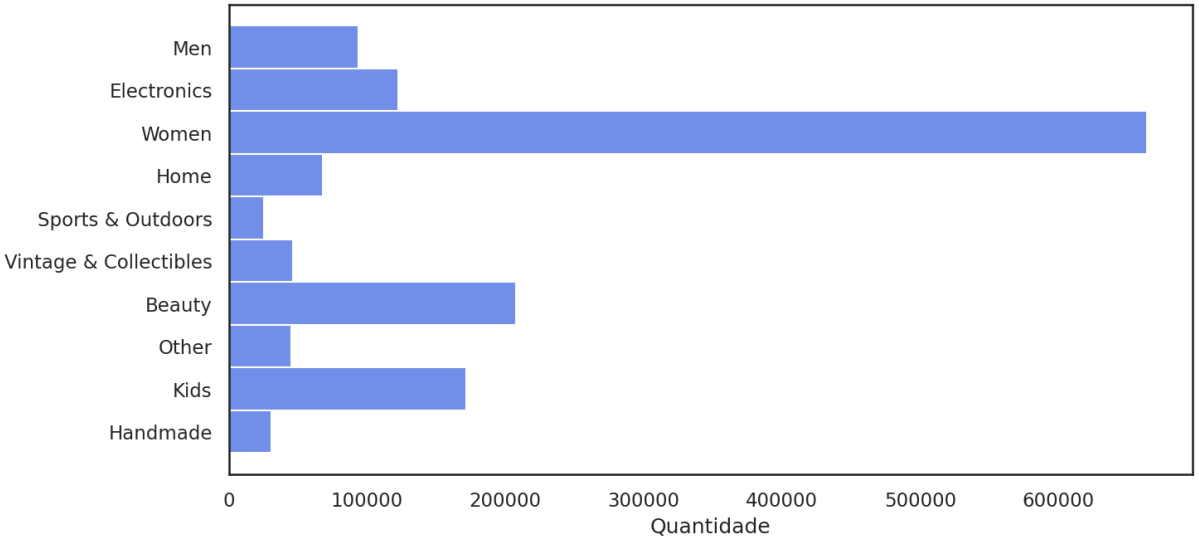
Figura 4 - Distribuição dos preços das empresas Xbox e Apple.



Percebe-se que a maioria dos preços dos produtos de Xbox estão até 400 dólares, enquanto a Apple a maioria dos preços dos produtos estão até mil dólares, mostrando que a marca tem uma influência no valor do produto, o intuito é verificar uma análise mais aprofundada nas próximas sprints acerca disso.

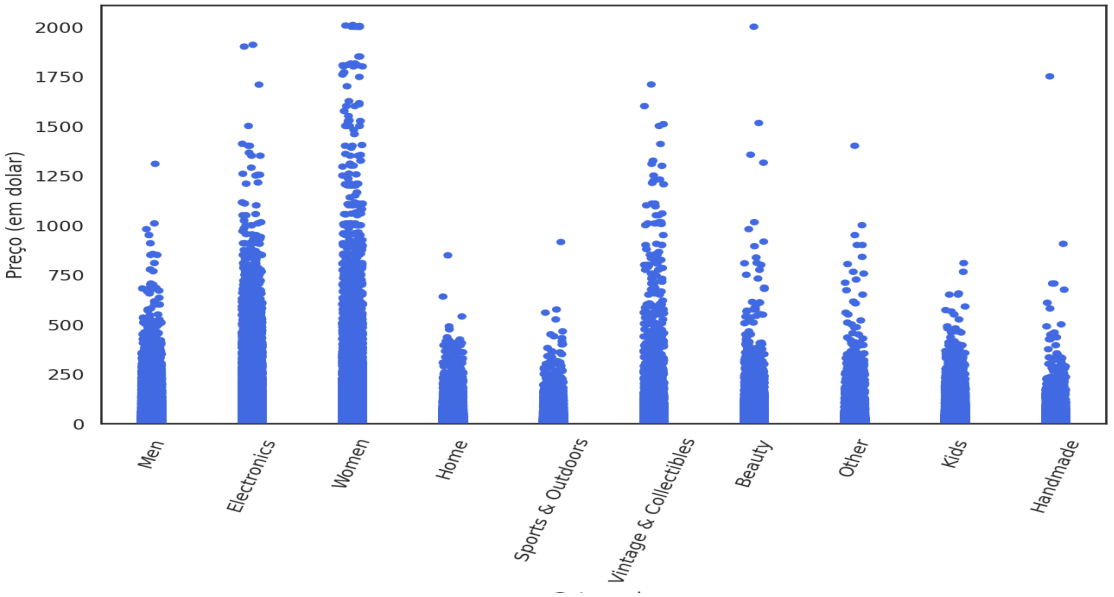
No dataset, para cada anúncio, são atribuídas 3 categorias. A primeira categoria (categoria principal) é a que define o segmento mais geral do produto (contém 11 classificações) e duas subcategorias que referem-se às características mais específicas. Na figura 3 é mostrada a distribuição de anúncios para a categoria principal dos anúncios. Devido a grande quantidade de classes nas subcategorias (123 classes para a primeira subcategoria e 803 classes para a segunda subcategoria) elas não foram avaliadas neste relatório. O segmento com maior quantidade de anúncios é o Feminino (Women) com mais de 600 mil itens anunciados. Em seguida está o segmento de produtos de beleza (Beauty) com 200 mil anúncios. Estas informações em conjunto com a distribuição das marcas indicam que o principal cliente da empresa são mulheres.

Figura 5 - Distribuição dos anúncios segundo a Categoria Principal



Na Figura 6 é exibida a distribuição de preços segundo a Categoria Principal dos anúncios. Os produtos destinados ao público Feminino (Women) têm os maiores valores, com máximos acima dos 2.000 dólares e com concentração na faixa de até 1.000 dólares. Outras duas classes se destacam pela maior variação de preços entre os anúncios: Eletrônicos (Electronics) e Vintage e Colecionáveis (Vintage e Collectibles). As classes com menor dispersão são a de produtos para o lar (Home) e esportivos (Sport & Outdoors) às que concentram os preços na faixa de até 500 dólares.

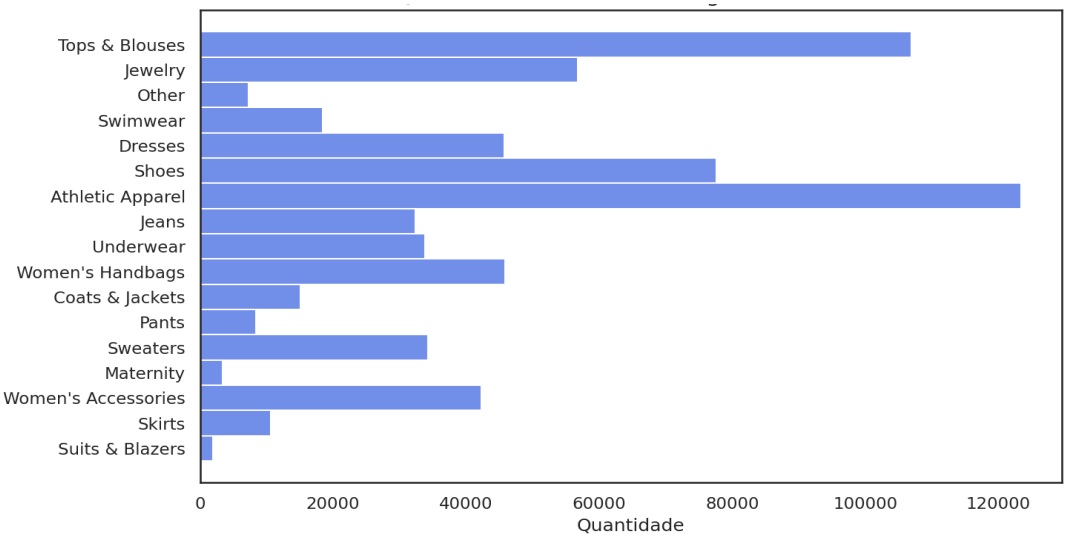
Figura 6 - Distribuição dos preços segundo a Categoria Principal



Na Figura 7 é feita uma análise mais detalhada dos produtos da categoria Women. A maioria das sub-categorias presentes nesta classe são de peças de

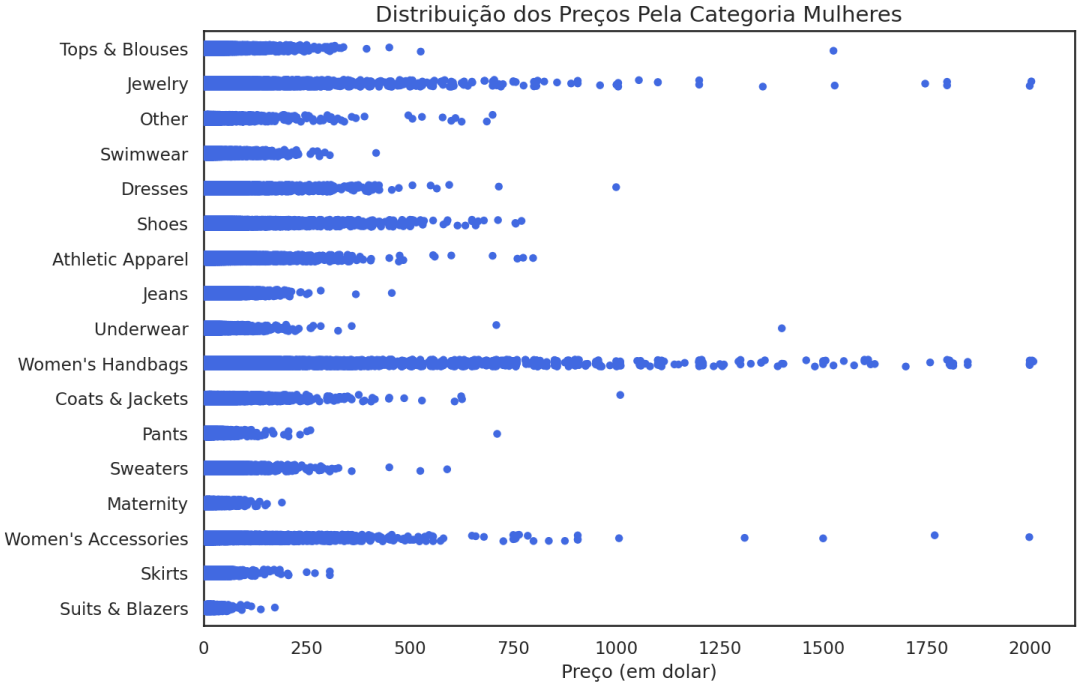
vestuário. Individualmente as subcategorias com maior quantidade de anúncios são: Athletic Apparel e Tops & Blouses.

Figura 7 - Distribuição das subcategorias dos itens da categoria Women



Quanto a distribuição dos preços (Figura 7) dos anúncios da Categoria Women os acessórios, como bolsas(Handbags), jóias(Jewelry) e outros acessórios (Accessories), tendem ter preços maiores que as peças de vestuário. Os menores preços médios são observados para produtos das sub-categorias: Pants, Maternity, Skirts e Suits & Blazers.

Figura 8 - Distribuição de preços das subcategorias dos itens da categoria Women.



No Anexo são exibidos os mesmos gráficos das Figuras 7 e 8, porém para as demais categorias principais.

5. Modelagem

O modelo a ser desenvolvido neste projeto precisa atender ao objetivo principal de precificar novos produtos cadastrados em um site de e-commerce. Com base nas hipóteses levantadas e em alguns estudos realizados sobre precificação dinâmica, algumas ideias de técnicas para modelagem foram identificadas para resolver problemas específicos. Caso confirmadas as hipóteses, o modelo precisa poder agrupar produtos por similaridade, considerar a data de anúncio, identificar a marca e as condições dos produtos cadastrados.

Um modelo simplificado foi desenvolvido inicialmente com intuito de servir de referencial para avaliação do desempenho de modelos mais avançados posteriormente. Ele basicamente identifica produtos que possuem as mesmas categorias do produto e realiza uma média dos preços de todos eles para encontrar um preço para o novo produto cadastrado. Este algoritmo inicial é baseado na hipótese de que produtos similares devem possuir preços similares.

Um segundo modelo também foi conceituado, para este segundo modelo planeja-se utilizar técnicas de NLP. Baseando-se na hipótese de que o atributo nome e a descrição dos anúncios carregam informação importante para definir o preço de um produto, o modelo buscará precificar os produtos utilizando tais informações. Este modelo também deverá considerar as informações de categorias, marcas, descrição e frete para a atribuição de preços. Neste modelo serão utilizadas a técnica de TfidfVectorizer e a distância cosseno para cálculo de similaridade. Planeja-se ainda, com este modelo, sugerir uma faixa de preços em que o produto possa estar.

A similaridade de cosseno é um método que utiliza o ângulo de cosseno formado entre dois vetores para identificar a semelhança entre eles. Ao considerar a função cosseno, conforme o ângulo se aproxima de 0, seu valor é 1, e, portanto, isso significa que os dois vetores estão sobrepostos e são o máximo similares, já quando o ângulo se aproxima de 180, seu valor é -1, e, portanto, os vetores estão precisamente opostos, representando o máximo de diferença entre os vetores [3].

Foi possível utilizar a ferramenta de similaridade de cosseno graças ao uso da função TfidfVectorizer da biblioteca Sklearn para transformar todo o conteúdo textual em vetores. O TfidVectorizer se baseia no cálculo dos valores TF-IDF, que basicamente é o produto da frequência de um termo em um determinado documento multiplicado pela frequência inversa do mesmo termo em todos os documentos. [4].

6. Próximos passos

Para a próxima sprint planeja-se aprofundar o entendimento do negócio, através do diálogo com membros da empresa solicitante e feedbacks dos professores. Pois, quanto melhor se entender quais as “dores” do negócio e demandas do solicitante, mais adequadas à realidade da empresa serão as soluções geradas.

Serão exploradas as hipóteses de dependências, tendências dos preços em função da série temporal e da quantidade de itens no estoque.

Serão testados modelos de agrupamento dos itens como: K-means, DBSCAN e Meanshift.

Será feita a avaliação de desempenho dos primeiros modelos propostos.

Referências

[1] SANTOS, F. A. do N.; MAYER, V. F.; MARQUES, O. R. B. Precificação dinâmica e percepção de justiça em preços: um estudo sobre o uso do aplicativo Uber em viagens. **Turismo: Visão e Ação**, v. 21, p. 239–264, 8 maio 2020.

[2] VAN DER POEL, N. **Precificação dinâmica como uma ferramenta para administrar preços e vendas no varejo on-line : um estudo de caso na Netshoes**. Dissertação de Mestrado—São Paulo: Fundação Getúlio Vargas, 2020.

[3] ARORA, S. **Semelhança de cosseno em Python**. Disponível em: <https://www.delftstack.com/pt/howto/python/cosine-similarity-between-lists-python/>. Acesso em: 16 set. 2022.

[4] VAZ, A. L. **TF-IDF — algoritmo de recomendação**. **Data Hackers**, 9 jul. 2022. Disponível em: <https://medium.com/data-hackers/tf-idf-algoritmo-de-recomenda%C3%A7%C3%A3o-6c3cbd55e439>. Acesso em: 16 set. 2022

ANEXO

Figura A1 - Distribuição das sub-categorias dos itens da categoria Beauty

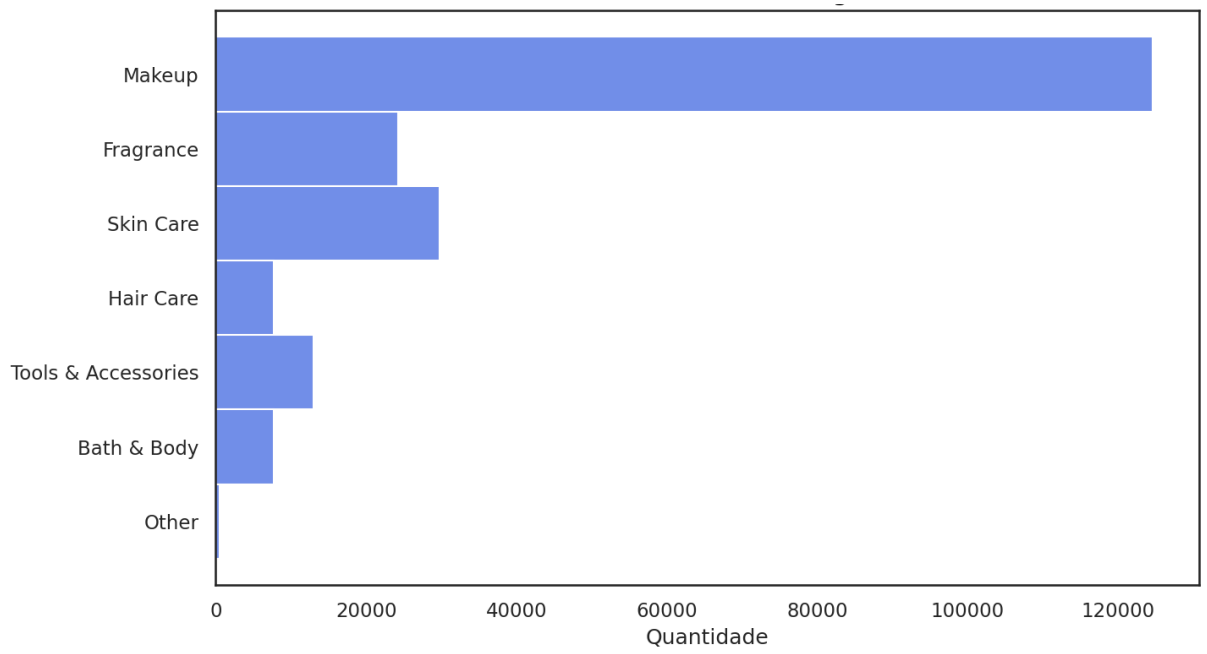


Figura A2 - Distribuição de preços das sub-categorias dos itens da categoria Beauty.

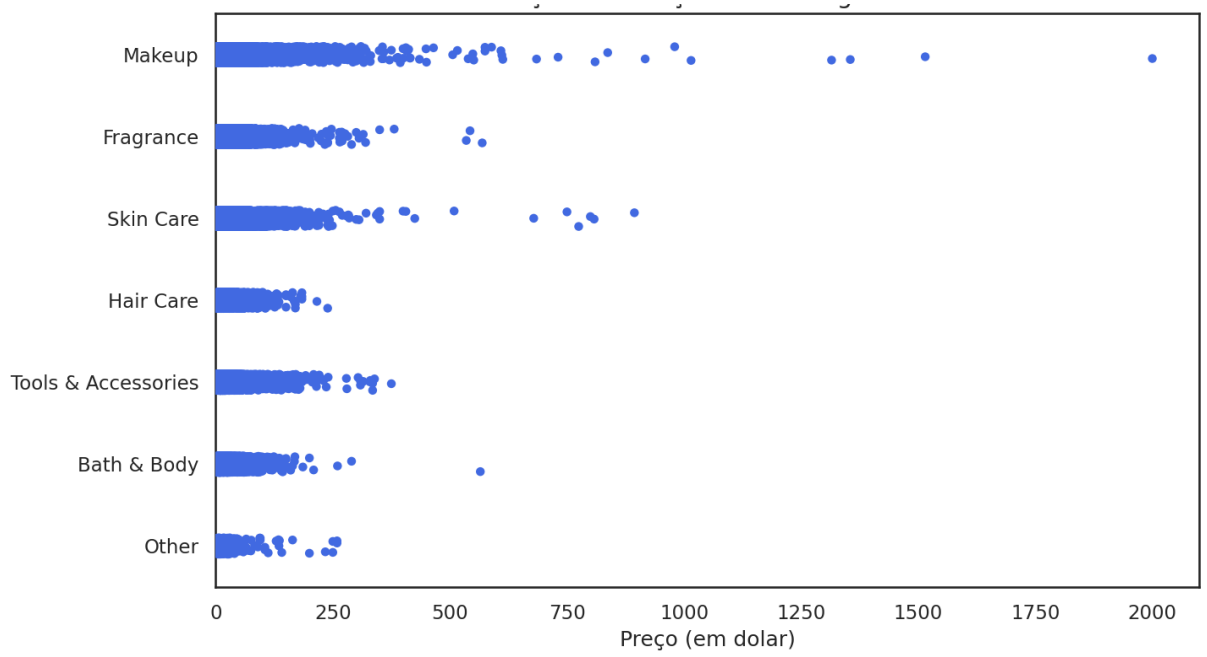


Figura A3 - Distribuição das sub-categorias dos itens da categoria Men

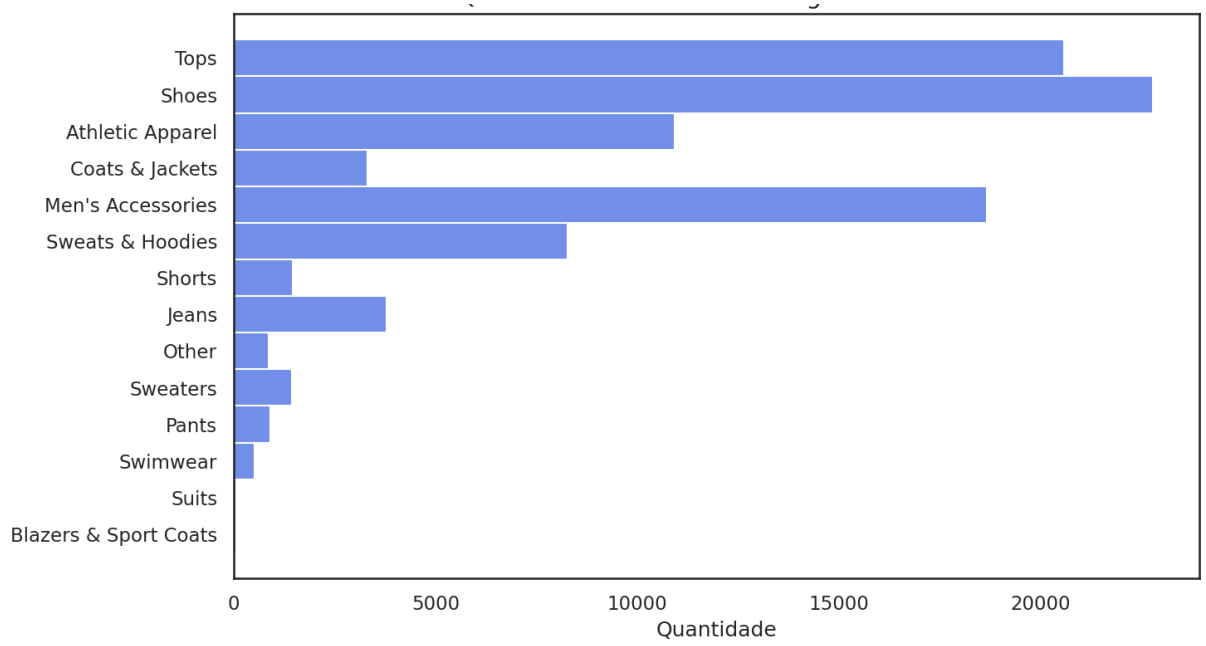


Figura A4 - Distribuição de preços das sub-categorias dos itens da categoria Men.

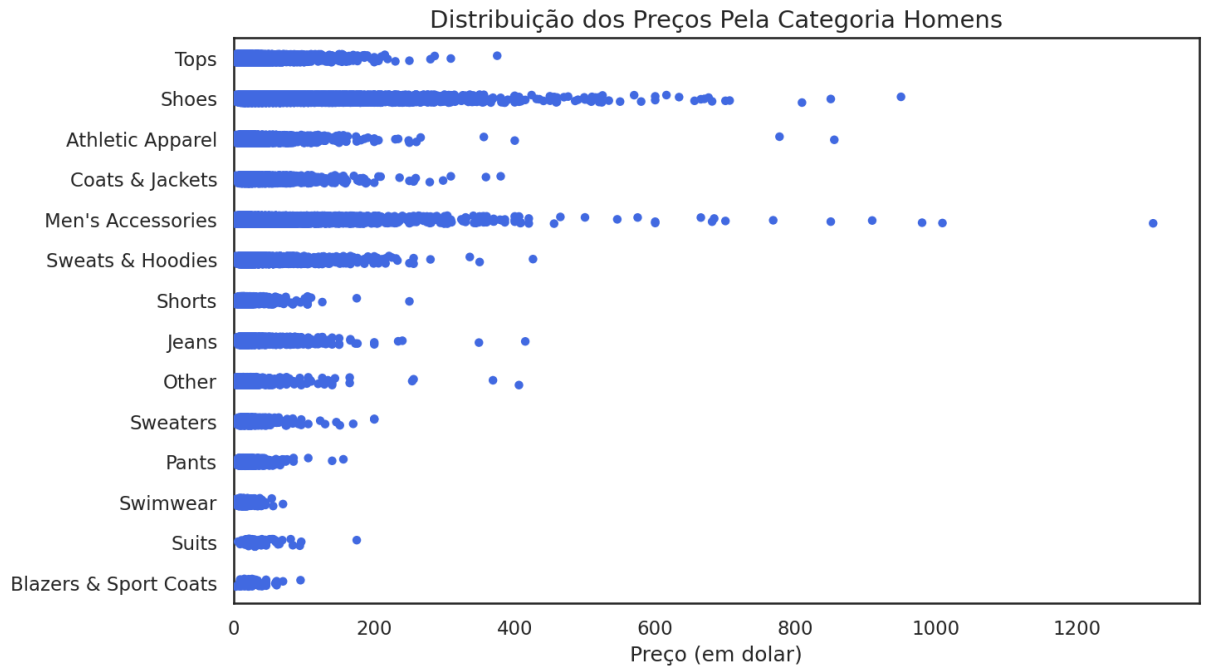


Figura A5 - Distribuição das sub-categorias dos itens da categoria Electronics.

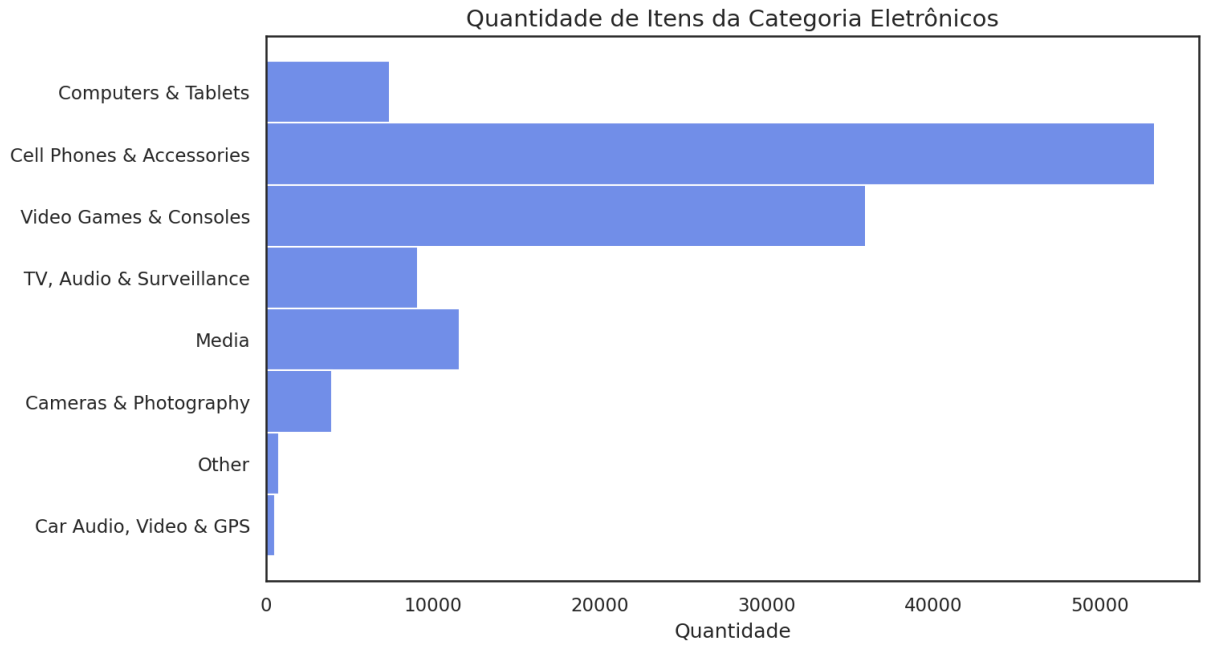


Figura A6 - Distribuição de preços das sub-categorias dos itens da categoria Electronics.

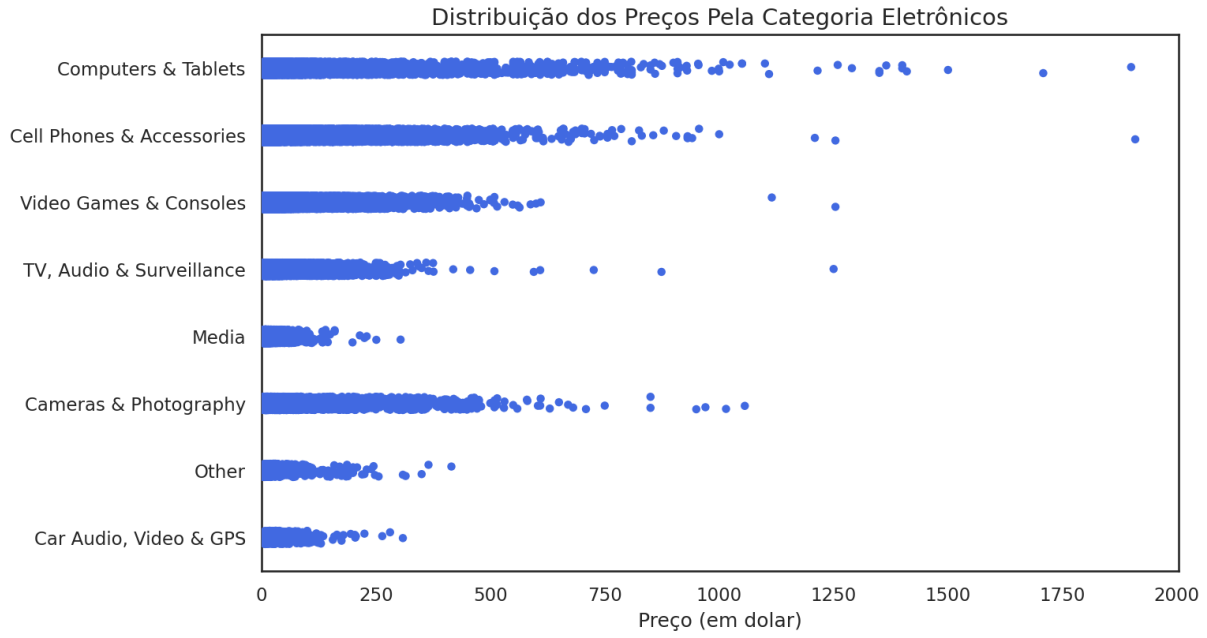


Figura A7 - Distribuição das sub-categorias dos itens da categoria Home.

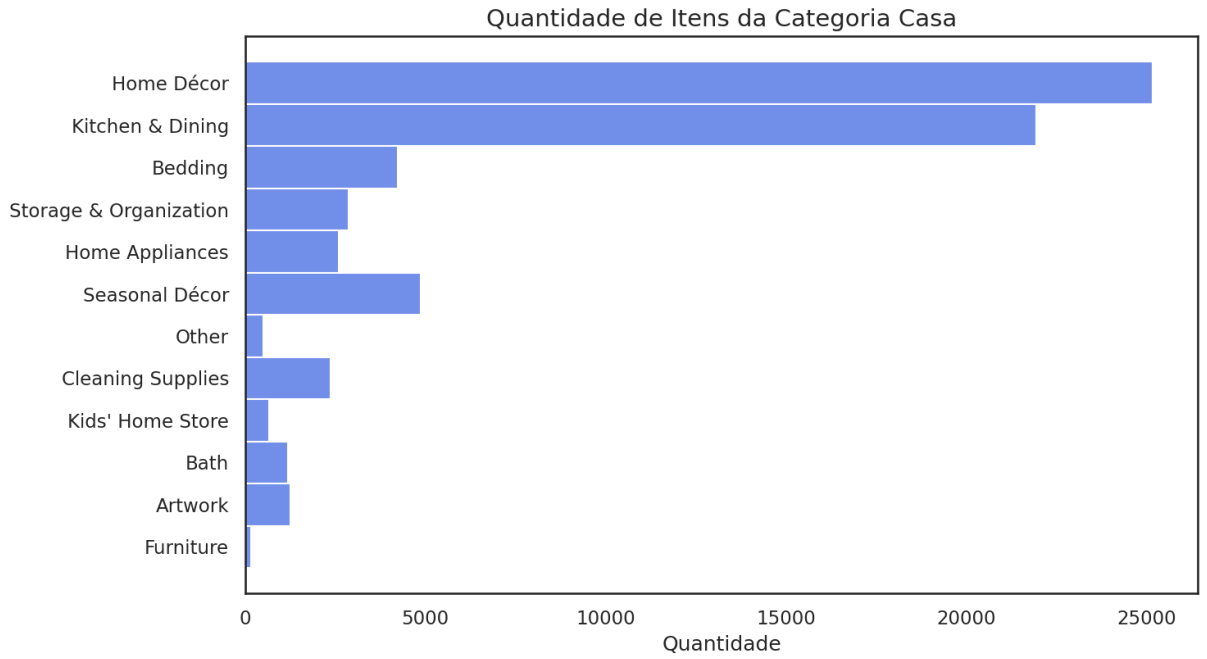


Figura A8 - Distribuição de preços das sub-categorias dos itens da categoria Home

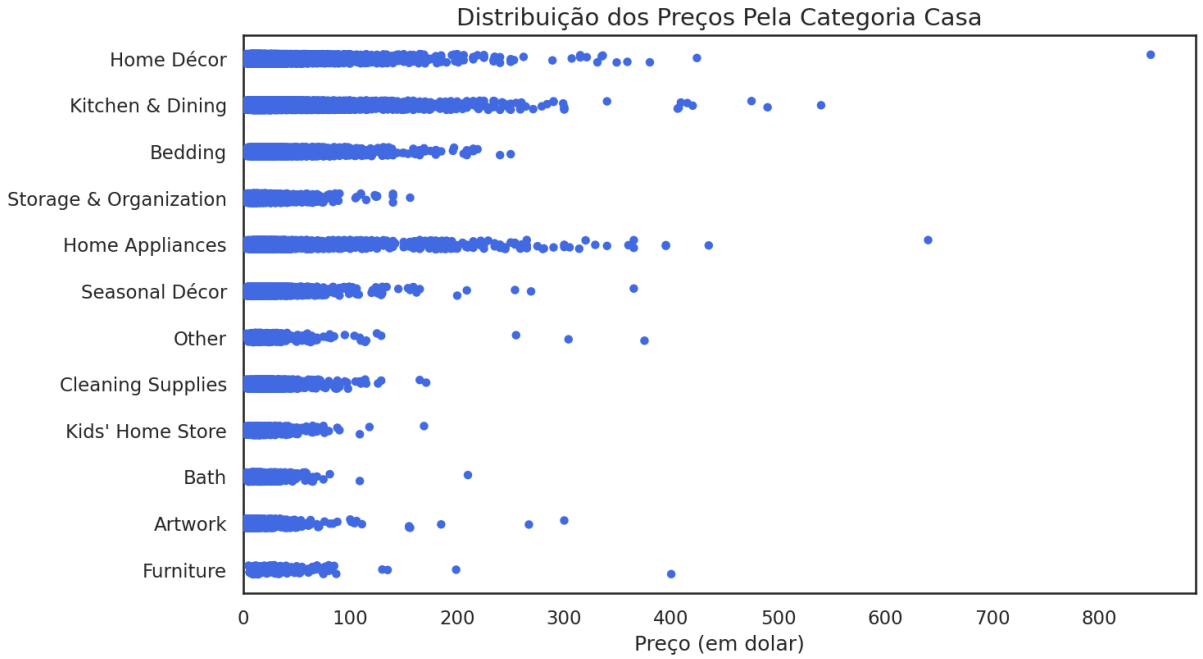


Figura A9 - Distribuição das sub-categorias dos itens da categoria Sports & Outdoor.

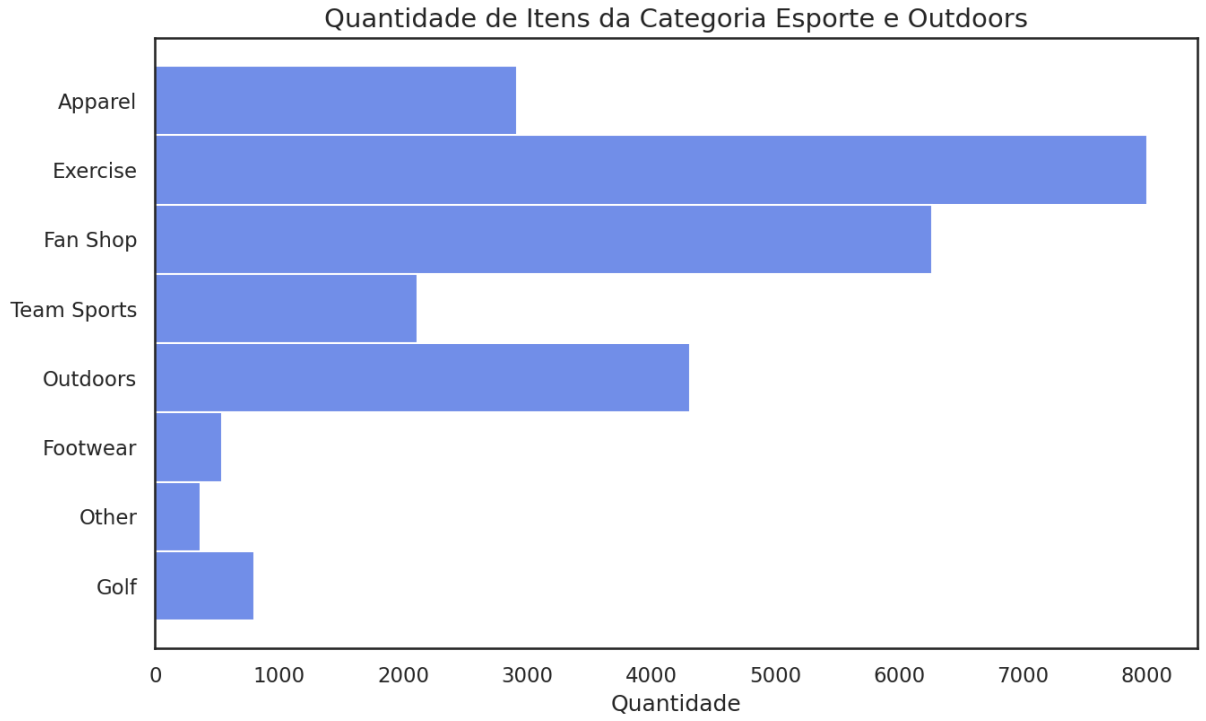


Figura A10 - Distribuição de preços das sub-categorias dos itens da categoria sports & Outdoor.

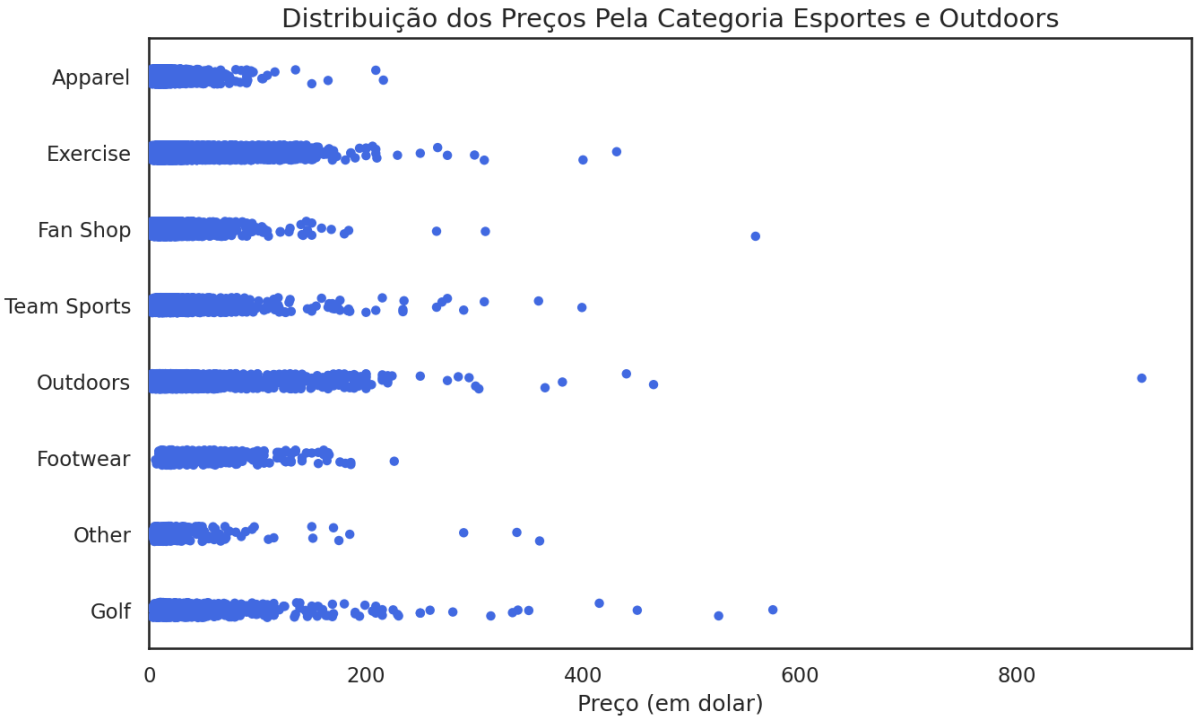


Figura A11 - Distribuição das sub-categorias dos itens da categoria Vintage & Collectionable.

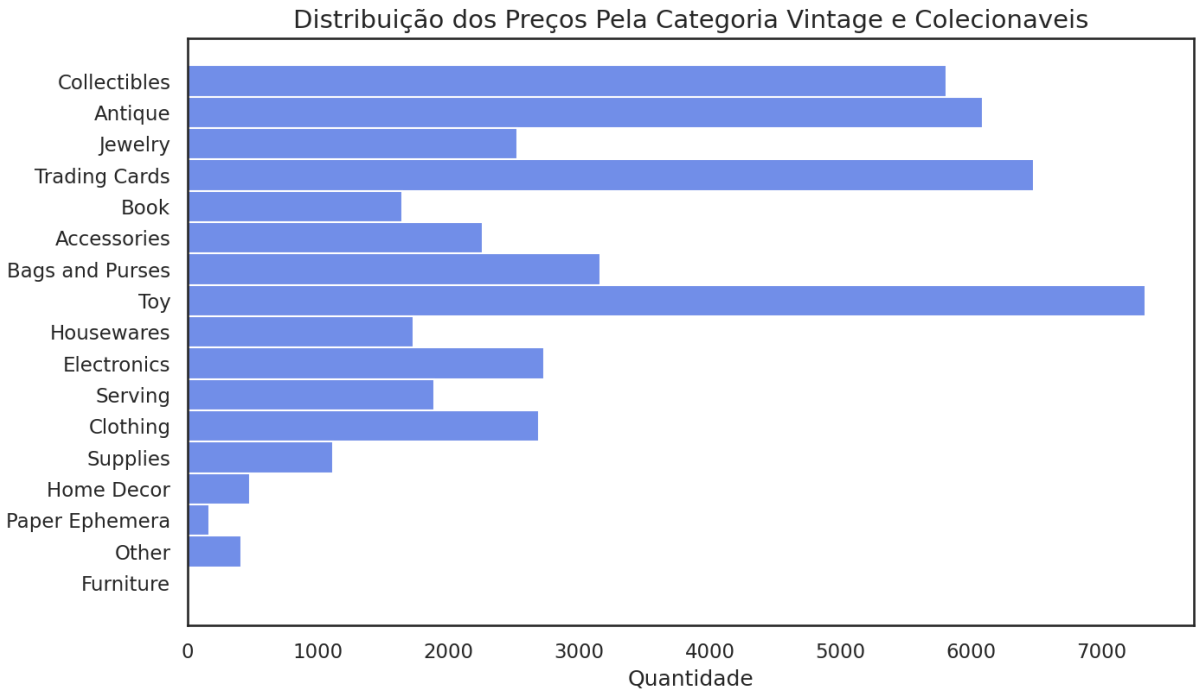


Figura A12 - Distribuição de preços das sub-categorias dos itens da categoria Vintage & Collectionable.

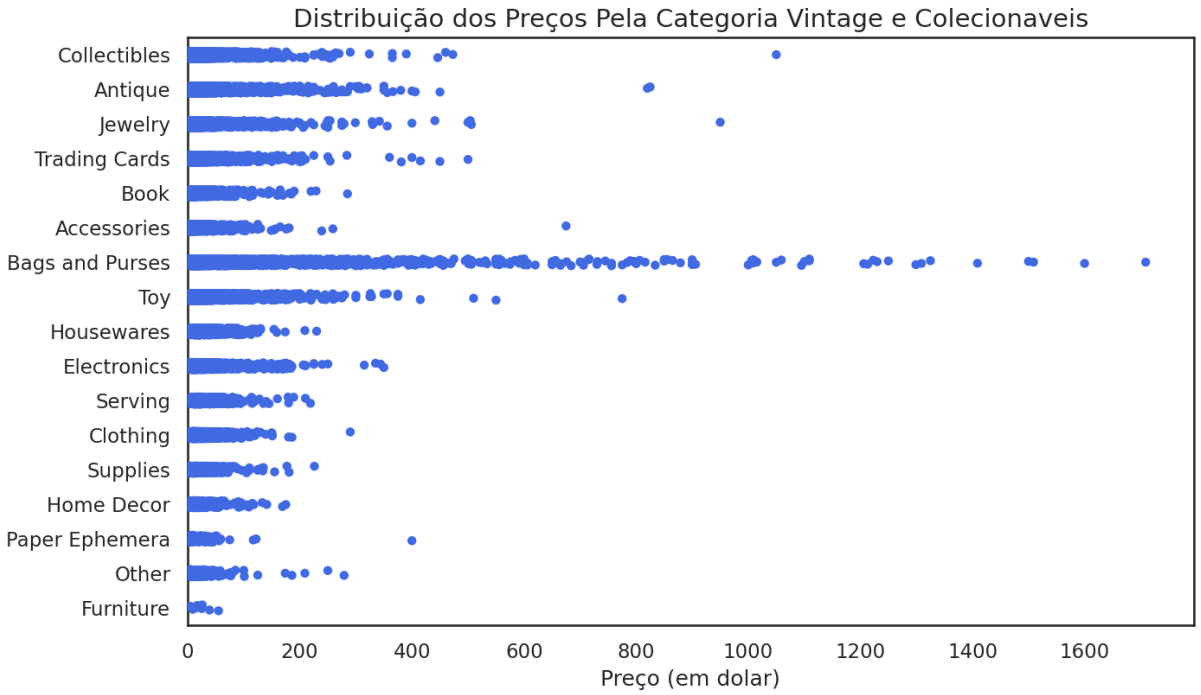


Figura A13 - Distribuição das sub-categorias dos itens da categoria Kids.

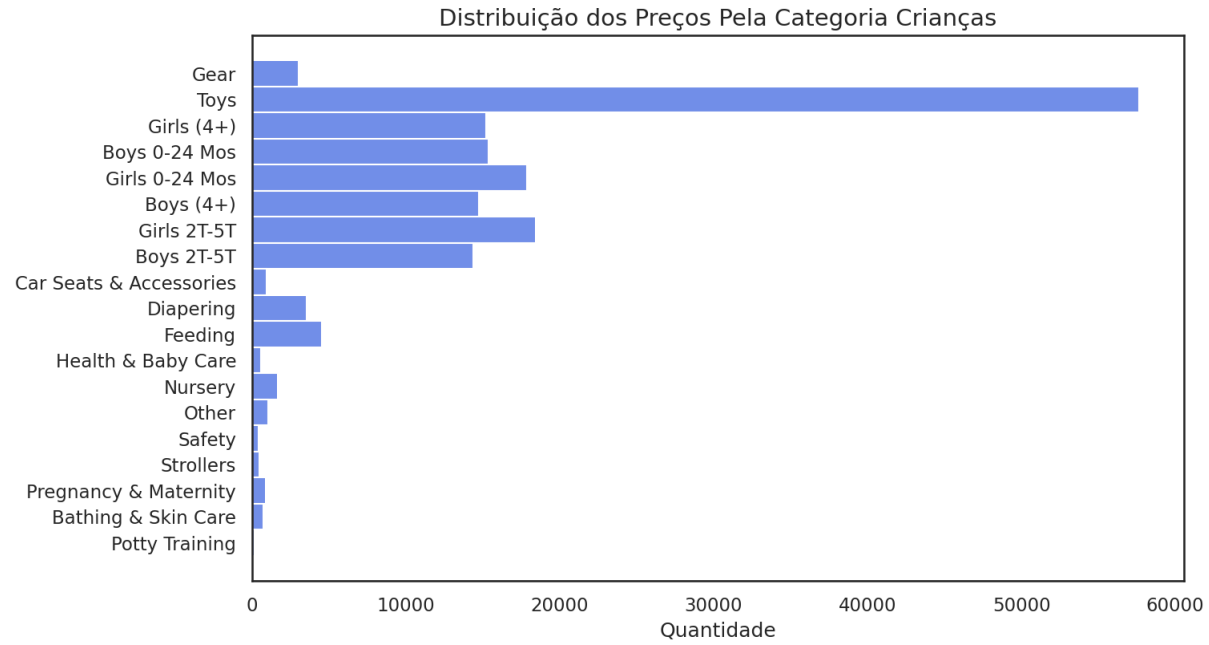


Figura A14 - Distribuição de preços das sub-categorias dos itens da categoria Kids.

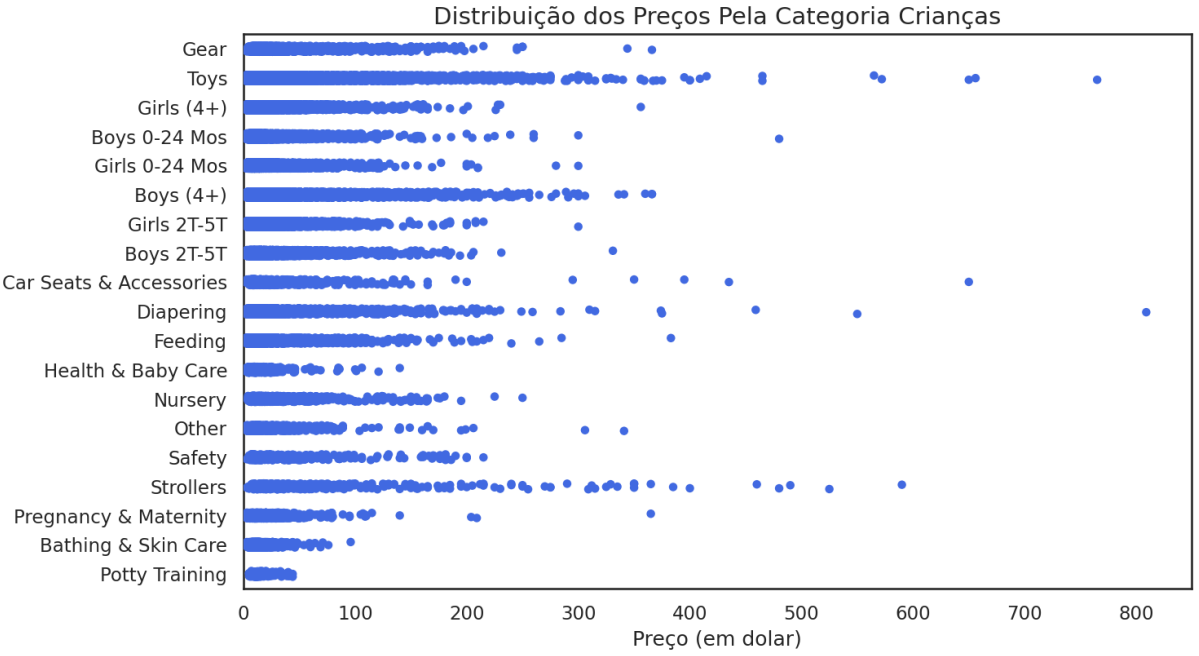


Figura A15 - Distribuição das sub-categorias dos itens da categoria Handmade

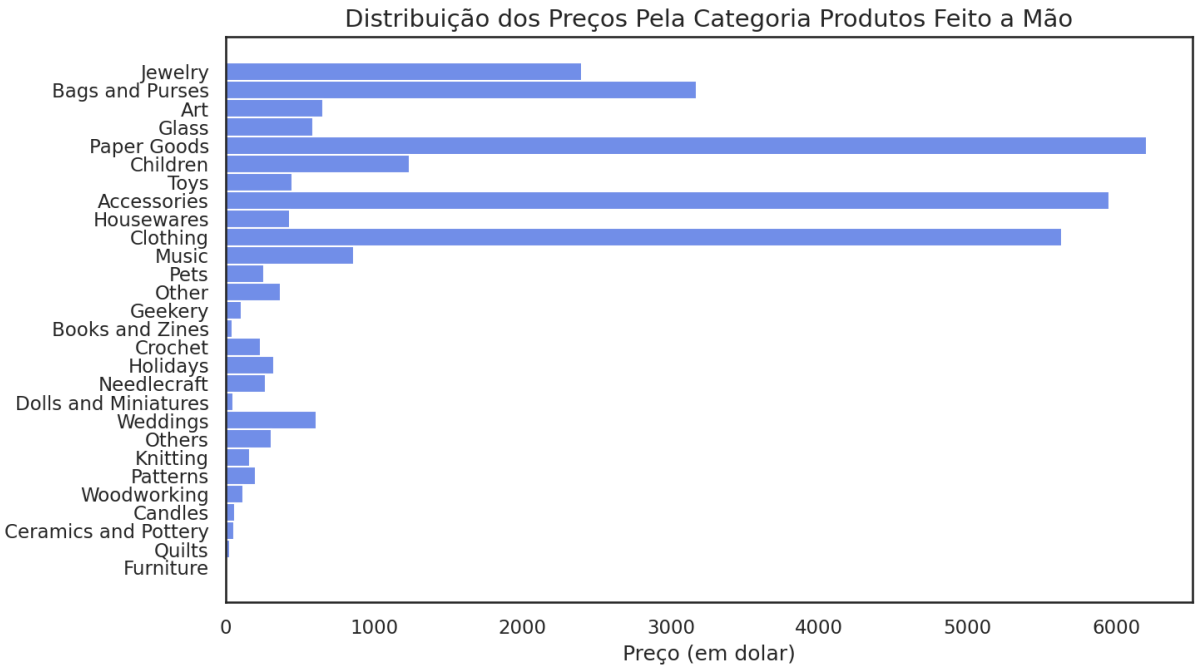


Figura A16 - Distribuição de preços das sub-categorias dos itens da categoria Handmade.

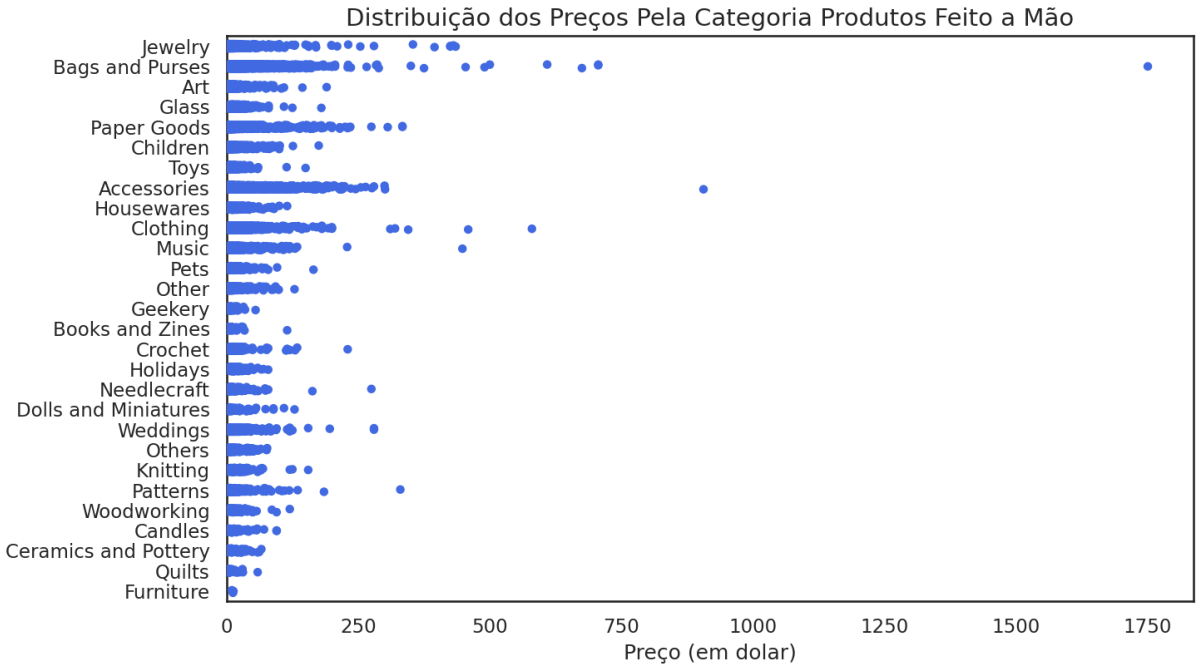


Figura A17 - Distribuição das sub-categorias dos itens da categoria Other

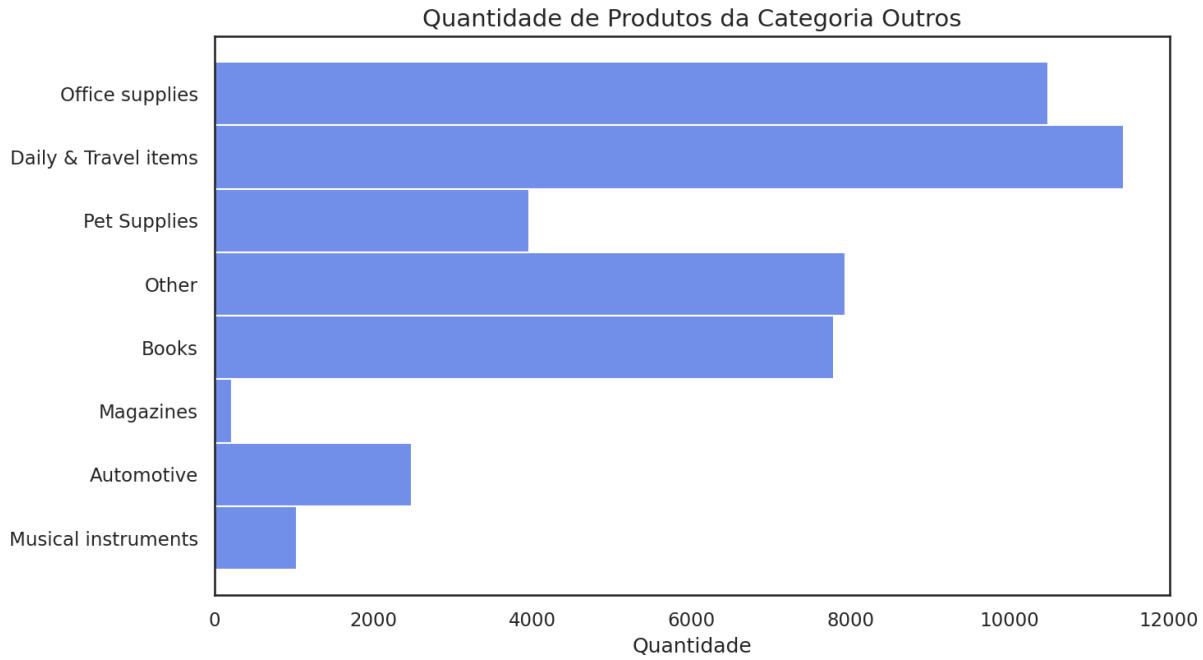


Figura A18 - Distribuição de preços das sub-categorias dos itens da categoria Other.