

UNIVERSIDADE FEDERAL DO PARANÁ

LUCAS ALEJANDRO TERRES

GEOGRAFIA DOS PREÇOS EM CURITIBA: UMA ANÁLISE EXPLORATÓRIA DA
DISPERSÃO DE PREÇOS NO VAREJO A PARTIR DE DADOS ABERTOS

CURITIBA
2025

LUCAS ALEJANDRO TERRES

A GEOGRAFIA DOS PREÇOS EM CURITIBA: UMA ANÁLISE EXPLORATÓRIA DA
DISPERSÃO DE CUSTOS NO VAREJO A PARTIR DE DADOS PÚBLICOS

Trabalho de Conclusão de Curso apresentado ao curso de Graduação em Gestão da Informação, Setor de Ciências Sociais Aplicadas, Universidade Federal do Paraná, como requisito parcial à obtenção do título de Bacharel em Gestão da Informação.

Orientador: Prof. Dr. José Simão de Paula Pinto

CURITIBA

2025

AGRADECIMENTOS

A finalização deste trabalho de conclusão de curso representa o fim de um ciclo importante, uma conquista que foi possível graças ao apoio e à contribuição de diversas pessoas a quem sou profundamente grato.

Ao meu orientador, Prof. Dr. José Simão de Paula Pinto, agradeço pela orientação segura e pela disponibilidade fundamental em todas as etapas deste projeto.

Estendo meus agradecimentos ao corpo docente do curso de Gestão da Informação. Os ensinamentos compartilhados em sala de aula construíram a base técnica e analítica que tornou a execução deste trabalho viável e recompensadora.

À minha família e aos meus amigos, meu muito obrigado pelo apoio incondicional, pela paciência nos momentos de ausência e pelo incentivo constante que tornou toda a jornada da graduação mais leve e significativa.

RESUMO

A dinâmica de preços no varejo alimentício impõe desafios significativos aos consumidores, acarretando despesas inerentes à busca por melhores condições de compra que decorrem da alta dispersão de preços. Posto isso, este trabalho tem como objetivo principal realizar uma análise exploratória multidimensional da dispersão de preços em supermercados no município de Curitiba, utilizando dados públicos da plataforma 'Clique Economia' para o ano de 2024. A metodologia utilizada consistiu no desenvolvimento de um processo de Extração, Transformação e Carga (ETL), que incluiu a normalização de endereços e a criação de uma hierarquia de categorização de produtos para mais de 700.000 registros de preços. A análise dos dados foi estruturada em três eixos: geográfico (por bairro), competitivo (por rede de supermercado) e temporal (evolução mensal). Os resultados revelam uma acentuada dispersão de preços entre os bairros da cidade, identificando 'ilhas de custo de vida' mais elevadas e, por outro lado, regiões mais econômicas no que tange às despesas no varejo alimentício. Foi possível, ainda, mapear as distintas estratégias de precificação das redes de varejo, que competem de forma mais agressiva em categorias de produtos específicas. Como síntese, o estudo demonstra o potencial de análise dos dados abertos como ferramenta de empoderamento do consumidor, culminando na apresentação de *insights* práticos para otimizar as decisões de compra.

Palavras-chave: Dispersão de Preços. Análise Exploratória de Dados. Varejo. Geografia dos Preços. Dados Abertos.

ABSTRACT

The complex dynamics of pricing in the food retail sector impose significant challenges on consumers, entailing search costs inherent to the pursuit of better purchasing conditions, which arise from high price dispersion. Therefore, this study's main objective is to perform a multidimensional exploratory analysis of price dispersion in supermarkets within the municipality of Curitiba, utilizing public data from the 'Clique Economia' platform for the year 2024. The methodology employed consisted of developing an Extraction, Transformation, and Load (ETL) process, which included address normalization and the creation of a hierarchical product categorization system for over 700,000 price records. The data analysis was structured along three axes: geographic (by neighborhood), competitive (by supermarket chain), and temporal (monthly evolution). The results reveal a marked price dispersion among the city's neighborhoods, identifying more expensive 'cost-of-living islands' alongside more economical regions for food retail expenses. Furthermore, it was possible to map the distinct pricing strategies of retail chains, which compete more aggressively in specific product categories. In summary, the study demonstrates the potential of open data analysis as a tool for consumer empowerment, culminating in the presentation of practical insights to optimize purchasing decisions.

Keywords: Price Dispersion. Exploratory Data Analysis. Retail. Price Geography. Open Data.

LISTA DE FIGURAS

FIGURA 1 - INTERFACE DE CONSULTA DO SITE CLIQUE ECONOMIA	25
FIGURA 2 - DEMONSTRAÇÃO DO PROCESSO DE ETL.....	28
FIGURA 3 - EXEMPLO DA ESTRUTURA DE ARQUIVOS DIÁRIOS NO REPOSITÓRIO DE DADOS ABERTOS.	29
FIGURA 4 – RELATÓRIO DE VALIDAÇÃO ESTRUTURAL PRELIMINAR DO DATASET.....	35

LISTA DE QUADROS

QUADRO 1 - DICIONÁRIO DE DADOS DE PRODUTOS.....	27
---	----

LISTAS DE TABELAS

TABELA 1 – DISTRIBUIÇÃO DOS PREÇOS REGULARES OBSERVADOS.....	44
TABELA 2 – MELHOR CUSTO DA “CESTA ESSENCIAL” POR BAIRRO E REDE DE SUPERMERCADO.....	65

LISTA DE GRÁFICOS

GRÁFICO 1 – BOXPLOT DOS PREÇOS REGULARES OBSERVADOS.....	45
GRÁFICO 2 – RANKING DAS 20 SUBCATEGORIAS COM MAIOR VOLATILIDADE DE PREÇOS.....	46
GRÁFICO 3 – RANKING DE BAIROS POR CUSTO DA CESTA ESSENCIAL.....	49
GRÁFICO 4 – HEATMAP DE PREÇOS: TOP 15 BAIROS VS. TOP 20 SUBCATEGORIAS (PREÇO MEDIANO).....	52
GRÁFICO 5 – MATRIZ DE COMPETITIVIDADE: PREÇO MEDIANO POR REDE E MACRO CATEGORIA.....	59
GRÁFICO 6 – EVOLUÇÃO DO PREÇO MEDIANO POR CATEGORIA AO LONGO DE 2024.....	62

SUMÁRIO

1 INTRODUÇÃO.....	8
1.1 PROBLEMA.....	9
1.2 OBJETIVOS.....	9
1.2.1 Objetivo Geral.....	9
1.2.2 Objetivos específicos.....	10
1.3 JUSTIFICATIVA.....	10
1.3.1 Justificativa Social e Econômica.....	10
1.3.2 Justificativa Tecnológica.....	11
1.3.3 Justificativa Acadêmica e Científica.....	11
1.3.4 Justificativa Pessoal.....	12
1.4 DELIMITAÇÃO DE PESQUISA.....	12
2 FUNDAMENTAÇÃO TEÓRICA.....	13
2.1 A TEORIA DA DISPERSÃO DE PREÇOS E A ASSIMETRIA DE INFORMAÇÃO..	13
2.2 ANÁLISE ESPACIAL E A DESCRIÇÃO DA GEOGRAFIA DOS PREÇOS.....	15
2.3 ESTRATÉGIAS DE PRECIFICAÇÃO E COMPETIÇÃO NO VAREJO.....	18
2.4 PRINCÍPIOS DA ANÁLISE EXPLORATÓRIA DE DADOS (AED).....	19
2.5 DADOS ABERTOS: FUNDAMENTOS TÉCNICOS E APLICAÇÕES NA GESTÃO PÚBLICA.....	22
3 METODOLOGIA.....	24
3.1 FONTE DE DADOS: O PROGRAMA CLIQUE ECONOMIA.....	24
3.2 PROCESSO ETL (EXTRACT, TRANSFORM, LOAD).....	28
3.2.1 Extração.....	29
3.2.2 Transformação e limpeza dos dados.....	30
3.2.2.1 Enriquecimento das informações.....	32
3.2.2.2 Validação preliminar do dataset.....	34
3.2.3 Carregamento dos dados.....	36
3.3 DEFINIÇÃO DA "CESTA DE CATEGORIAS ESSENCIAIS".....	37
3.4 METODOLOGIA DE ANÁLISE EXPLORATÓRIA E ESCOLHA DAS MÉTRICAS ESTATÍSTICAS.....	38
3.4.1 Ferramenta Computacional Adotada.....	38
3.4.2 Eixos de Análise.....	39
3.5 LIMITAÇÕES METODOLÓGICAS E CONSIDERAÇÕES ÉTICAS.....	41
3.5.1 Limitações Metodológicas.....	41
3.5.2 Considerações Éticas.....	42
4 ANÁLISE E DISCUSSÃO DOS RESULTADOS.....	43
4.1. CARACTERIZAÇÃO DA DISPERSÃO DE PREÇOS EM CURITIBA.....	43
4.1.1 Visão geral dos preços.....	43

4.1.2 As Categorias Mais Voláteis.....	45
4.2. A DIMENSÃO GEOGRÁFICA: ONDE SE COMPRA MAIS CARO E MAIS BARATO?.....	47
4.2.1. Ranking de custo de vida por bairro.....	48
4.2.2. A "Vocação" dos Bairros.....	51
4.3. A DIMENSÃO COMPETITIVA: AS ESTRATÉGIAS DAS REDES DE VAREJO.....	54
4.3.1. Ranking de custo por rede de supermercado.....	55
4.3.2. Matriz de posicionamento estratégico.....	58
4.4. A DIMENSÃO TEMPORAL: A EVOLUÇÃO DOS PREÇOS EM 2024.....	61
4.4.1. A "inflação" por categoria.....	61
4.5. SÍNTESE DOS RESULTADOS: O GUIA DO CONSUMIDOR CURITIBANO.....	64
4.5.1. O melhor custo-benefício por bairro.....	64
5 CONSIDERAÇÕES FINAIS.....	67
5.1 ALCANCE DOS OBJETIVOS.....	68
5.2 TRABALHOS FUTUROS.....	69
REFERÊNCIAS.....	71
APÊNDICE 1 - SCRIPT DE EXTRAÇÃO E CONCATENAÇÃO DAS BASES.....	75
APÊNDICE 2 - TABELA DE ENDEREÇOS.....	77
APÊNDICE 3 - SCRIPT DE PRÉ-PROCESSAMENTO E CLASSIFICAÇÃO DE DADOS.....	78
APÊNDICE 4 - SCRIPT DE VALIDAÇÃO DE QUALIDADE DO DATASET FINAL..	83
APÊNDICE 5 - SCRIPT DE ANÁLISE DESCRITIVA E VISUALIZAÇÃO DE PREÇOS.....	88
APÊNDICE 6 - HEATMAP DE COMPETITIVIDADE DE PREÇO MEDIANO POR..	89
REDE E SUBCATEGORIA.....	89
APÊNDICE 7 - EVOLUÇÃO DO PREÇO MEDIANO DAS PRINCIPAIS SUBCATEGORIAS.....	90

1 INTRODUÇÃO

A dinâmica de preços no varejo alimentício representa um desafio constante para os consumidores, especialmente em um cenário de instabilidade econômica, no qual otimizar o poder de compra torna-se cada vez mais necessário. Em mercados urbanos como Curitiba, a dispersão de preços para produtos idênticos ou similares compromete a eficiência do consumo, exigindo do cidadão tempo e recursos para localizar as melhores ofertas. Segundo a Confederação Nacional de Dirigentes Lojistas (CNDL, 2023), o preço é o principal fator que influencia a decisão de compra do consumidor brasileiro, o que reforça a importância de ferramentas que promovam transparência e reduzam os chamados “custos de busca”.

Neste contexto, os dados abertos da plataforma "Clique Economia", da Prefeitura de Curitiba, são oportunos para investigar a dispersão de preços no varejo da cidade. Assim, este trabalho busca transformar os dados disponibilizados por essa plataforma em inteligência acionável, respondendo à questão central: Como a variação espacial dos preços de produtos de consumo no varejo alimentício de Curitiba pode ser analisada com base em dados públicos, de modo a promover o empoderamento do consumidor?

O objetivo geral é realizar uma análise exploratória multidimensional da dispersão de preços no varejo curitibano em 2024, com foco em: mapear a variação de preços entre bairros (eixo geográfico); comparar o posicionamento das principais redes de supermercados (eixo competitivo); identificar categorias com maior ou menor volatilidade (eixo temporal); e sintetizar os achados em um guia prático.

A metodologia baseou-se em Análise Exploratória de Dados (AED), com um processo robusto de Extração, Transformação e Carga (ETL) sobre mais de 700 mil registros. Incluiu normalização de endereços e criação de uma hierarquia de categorias de produtos para permitir análises agregadas. As análises foram estruturadas nos três eixos — geográfico, competitivo e temporal — utilizando Python para manipulação e Matplotlib/Seaborn para visualizações.

Este trabalho reforça a importância da análise de preços como ferramenta de economia e planejamento, evidenciando o valor social da gestão da informação ao transformar dados públicos em benefício prático para a sociedade.

1.1 PROBLEMA

A significativa variação de preços no varejo alimentício de Curitiba é um fenômeno documentado, com reportagens indicando discrepâncias que podem ultrapassar 200% para itens idênticos entre estabelecimentos (Tribuna PR, 2021). Essa acentuada dispersão de preços impõe um desafio substancial para o consumidor, que, na busca por otimizar seu orçamento, depara-se com uma considerável assimetria de informação. A ausência de uma fonte de dados consolidada, estruturada e de fácil acesso eleva os "custos de busca" – o tempo e os recursos que o consumidor precisa investir para encontrar a melhor oferta.

Em um contexto de variações frequentes nos preços de itens essenciais, essa dificuldade transcende a mera conveniência, impactando diretamente o poder de compra das famílias. As decisões de consumo acabam, muitas vezes, sendo guiadas por percepção, hábito ou pela proximidade geográfica do estabelecimento, em detrimento de uma escolha economicamente mais racional e informada.

Portanto, a falta de uma análise clara sobre os padrões que regem essa variação constitui o problema central desta pesquisa. Diante disso, a questão que norteia este trabalho é: **Como a variação espacial dos preços de produtos de consumo no varejo alimentício de Curitiba pode ser analisada com base em dados públicos, de modo a promover o empoderamento do consumidor?**

1.2 OBJETIVOS

1.2.1 Objetivo Geral

Analisar a dinâmica de preços no varejo alimentício de Curitiba para identificar padrões geográficos, competitivos e temporais, a fim de gerar *insights* que auxiliem a tomada de decisão do consumidor.

1.2.2 Objetivos específicos

- a) Mapear a variação de preços de uma cesta de categorias essenciais de consumo entre os bairros de Curitiba.
- b) Comparar o posicionamento de preço das principais redes de supermercado, identificando suas estratégias por categoria de produto.
- c) Identificar as categorias de produtos com maior e menor estabilidade de preços ao longo do período analisado.
- d) Sintetizar os achados em um guia prático e acionável para o consumidor curitibano.

1.3 JUSTIFICATIVA

O estudo sobre a variação de preços de uma cesta de categorias essenciais de consumo em Curitiba, com base na base de dados pública Clique Economia, justifica-se por múltiplas dimensões: social, econômica, tecnológica e acadêmica.

1.3.1 Justificativa Social e Econômica

Em contextos de possíveis variações regionais no custo dos alimentos, compreender a dinâmica de precificação no varejo pode contribuir para o planejamento financeiro das famílias. Segundo dados do Instituto Brasileiro de Geografia e Estatística (IBGE), a alimentação no domicílio tem registrado altas significativas que pressionam o orçamento doméstico (UOL Economia, 2024). Este estudo visa contribuir com a redução dessa assimetria informacional ao mapear a dispersão de preços no varejo em Curitiba, promovendo maior transparência e transformando dados públicos em subsídios para decisões de consumo mais conscientes e econômicas. Ao decodificar a dinâmica de preços, este trabalho empodera o cidadão e, ao fomentar a transparência, pode estimular a competitividade no setor varejista local.

1.3.2 Justificativa Tecnológica

Este estudo fundamenta-se na aplicação de um *pipeline* de Ciência de Dados para transformar um grande volume de dados brutos e heterogêneos em uma base de conhecimento estruturada e acionável. Utilizando a linguagem Python e seu ecossistema de bibliotecas (como Pandas, Matplotlib e Seaborn), foi desenvolvido um processo de Extração, Transformação e Carregamento (ETL) que abrangeu desde a limpeza e validação de mais de 700.000 registros até complexas etapas de enriquecimento de dados, como a normalização geográfica de endereços e o desenvolvimento iterativo de uma detalhada hierarquia de categorização de produtos. O projeto reforça, portanto, o papel da tecnologia e da Gestão da Informação na produção de conhecimento aplicado a problemas urbanos, alinhando-se ao que Shmueli et al. (2017) destacam sobre como a organização e o tratamento de dados públicos ampliam as possibilidades de geração de valor social a partir da informação.

1.3.3 Justificativa Acadêmica e Científica

Do ponto de vista acadêmico, o trabalho contribui para o campo da Gestão da Informação ao explorar um banco de dados público com granularidade espacial e temporal elevada, cuja utilização ainda é incipiente na literatura. A pesquisa promove a aplicação de fundamentos da análise de dados a uma problemática concreta da economia urbana, reforçando o papel da Gestão da Informação na interface entre dados governamentais e a tomada de decisão cidadã. Ao adotar uma abordagem local e empírica, o estudo evidencia como a análise territorializada dos preços pode revelar assimetrias de mercado e padrões de consumo, contribuindo para o avanço do conhecimento sobre a dinâmica informacional em ambientes urbanos.

1.3.4 Justificativa Pessoal

A escolha deste tema surgiu de uma curiosidade pessoal sobre como os preços variam de um lugar para outro na cidade e de um desejo prático de entender melhor onde e como tomar decisões mais vantajosas no consumo cotidiano. Percebi que, muitas vezes, fazemos escolhas de compra sem saber se existem diferenças significativas de custo entre bairros ou estabelecimentos. Isso despertou meu interesse em explorar, com base em dados públicos, como essas variações se distribuem geograficamente.

Ao longo do curso, desenvolvi afinidade com ferramentas de análise e visualização de dados, e vi nesta pesquisa uma oportunidade de aplicar esses conhecimentos em algo próximo do meu dia a dia, mas com relevância social mais ampla. Encaro este trabalho como uma experiência concreta de aprendizado, que alia o conhecimento técnico à prática informada — tanto para mim quanto, potencialmente, para outros consumidores.

1.4 DELIMITAÇÃO DE PESQUISA

Esta pesquisa se concentra na análise dos dados de preços coletados pela plataforma "Clique Economia" da Prefeitura de Curitiba durante o ano de 2024. O escopo geográfico está delimitado aos estabelecimentos localizados no município de Curitiba presentes na base de dados. A análise de produtos abrange as categorias de consumo mais relevantes, criadas através de um processo iterativo de classificação e agrupamento desenvolvido neste trabalho. A pesquisa possui caráter exploratório-descritivo, não se aprofundando em modelagem preditiva, focando na extração de padrões e *insights* a partir dos dados históricos.

2 FUNDAMENTAÇÃO TEÓRICA

O presente capítulo tem como objetivo estabelecer os fundamentos conceituais que embasam a proposta analítica deste trabalho, conectando o fenômeno empírico da dispersão de preços no varejo às principais teorias da economia da informação, da geografia urbana do consumo, da precificação competitiva e da análise de dados aplicada à gestão pública. A fundamentação teórica não apenas justifica a relevância do objeto de estudo, como também orienta a interpretação dos resultados apresentados nos capítulos analíticos.

A estrutura foi organizada em cinco eixos principais: (1) os fundamentos da dispersão de preços como sintoma de falhas de mercado associadas à assimetria de informação; (2) a desigualdade espacial no acesso ao consumo e os efeitos territoriais na formação de preços; (3) as estratégias de precificação adotadas por diferentes formatos varejistas, como supermercados e atacarejos; (4) o papel da Análise Exploratória de Dados (AED) como abordagem investigativa para extração de padrões em grandes bases de dados; e (5) o potencial dos dados abertos como instrumentos de transparência, controle social e suporte à tomada de decisão. Cada um desses pilares é desenvolvido nos subtópicos a seguir.

2.1 A TEORIA DA DISPERSÃO DE PREÇOS E A ASSIMETRIA DE INFORMAÇÃO

Em mercados perfeitamente competitivos com informação completa, produtos homogêneos tenderiam a ser transacionados todos pelo mesmo preço – a chamada Lei do Preço Único. Na prática, contudo, observa-se ampla dispersão de preços mesmo para bens idênticos, especialmente no varejo. Como apontam Pindyck e Rubinfeld (2006, p. 7), a maioria dos mercados de varejo é caracterizada por um grau considerável de dispersão de preços. Essa discrepância entre teoria e evidência reflete a presença de informação imperfeita no mercado.

Nesse contexto, George Stigler, em seu trabalho pioneiro *The Economics of Information*, lançou as bases para compreender teoricamente a existência de dispersão de preços em mercados concorrenciais com informação assimétrica. Segundo Stigler (1961, p. 214), a dispersão de preços é uma manifestação – e, de

fato, uma medida – da ignorância no mercado. O fator fundamental por trás desse fenômeno é o custo de busca de informações pelo consumidor: para encontrar o menor preço disponível, o comprador precisa pesquisar entre diferentes vendedores, o que envolve tempo, esforço e, eventualmente, custos financeiros. O consumidor racional interrompe a busca quando o custo marginal da investigação excede o benefício esperado de localizar uma oferta melhor. Essa diferença de comportamento entre consumidores ajuda a explicar a coexistência de múltiplos preços para um mesmo bem.

A assimetria de informação também está relacionada ao poder de mercado. Em mercados com informação perfeita, todos os compradores saberiam instantaneamente qual vendedor oferece o menor preço, forçando os demais a igualarem esse preço. Com informação imperfeita, vendedores passam a deter poder de mercado local sobre consumidores mal informados. Isso resulta em preços diferentes para o mesmo produto, conforme o nível de informação do comprador.

Varian (1980) formalizou essa lógica ao propor um modelo em que parte dos consumidores é informada e outra parte não é. As firmas então adotam estratégias mistas de precificação, oscilando entre preços baixos para atrair os bem-informados e preços altos para capturar excedente dos consumidores desinformados. O resultado é uma dispersão de preços persistente mesmo em equilíbrio competitivo. Segundo o autor, a variabilidade de preços é um reflexo racional do comportamento estratégico de empresas em um mercado com diferentes perfis de consumidores.

Por fim, a teoria da informação assimétrica também ajuda a explicar falhas de mercado. Akerlof (1970), no clássico artigo "The Market for Lemons", demonstrou como a falta de informação sobre qualidade leva à seleção adversa. Aplicado ao contexto da dispersão de preços, consumidores com menor acesso à informação pagam mais por bens idênticos, resultando em perda de bem-estar e ineficiência alocativa.

Em suma, a teoria econômica desenvolvida por autores como Stigler (1961), Varian (1980) e Akerlof (1970) oferece um arcabouço robusto para compreender a dispersão de preços como um fenômeno associado à assimetria de informação e aos custos de busca. A redução desses custos, por meio de políticas de transparência, comparadores de preços e acesso a dados abertos, tende a mitigar a dispersão e promover maior eficiência no mercado.

2.2 ANÁLISE ESPACIAL E A DESCRIÇÃO DA GEOGRAFIA DOS PREÇOS

A estruturação espacial do consumo nas cidades revela profundas desigualdades geográficas de preços e acesso a bens essenciais. A geografia do consumo – isto é, a forma como atividades de comércio e hábitos de compra se distribuem no território urbano – está intrinsecamente ligada à segmentação espacial do varejo. Nas últimas décadas, o capital comercial varejista tornou-se um agente central na produção do espaço urbano em múltiplas escalas, revelando desigualdades em termos de poder de uso e apropriação dos territórios, segundo Pereira (2020, p. 14). Em outras palavras, a localização das atividades de varejo não é aleatória: grandes empresas tendem a usar o território de forma seletiva, privilegiando áreas de maior rentabilidade e infraestrutura, enquanto comunidades menos favorecidas ficam à margem dessa rede de ofertas (Pintaudi, 1999, p. 61). Esse uso corporativo seletivo do espaço reforça disparidades preexistentes, produzindo “ilhas” de abundância de consumo em alguns bairros e vazios ou precariedade comercial em outros. Assim, lógicas de mercado e de acumulação capitalista acabam moldando um espaço urbano fragmentado e polarizado, no qual convivem lugares de sofisticação supérflua ao lado de locais marcados pela dificuldade de reprodução da vida cotidiana (SANTOS, 2001, p. 47).

Do ponto de vista teórico, é importante ressaltar a contribuição de Milton Santos sobre a dualidade dos circuitos econômicos urbanos. Em "O Espaço Dividido", Santos (1979, p. 84) descreve a coexistência de um circuito superior – composto pelo grande comércio modernizado, apoiado em tecnologia avançada e redes oligopolistas – e um circuito inferior, formado por atividades de pequena escala voltadas ao consumo popular local. O circuito inferior atende sobretudo à população de menor renda e mobilidade, estando indiretamente subordinado ao circuito moderno, já que muitos produtos vendidos na periferia são fornecidos pelas indústrias e distribuidoras do circuito superior. Essa teoria evidencia como a segmentação espacial do varejo reflete e reforça a segregação socioeconômica: áreas nobres da cidade são atendidas por redes varejistas modernas, enquanto bairros periféricos muitas vezes dependem de comércio de menor porte. Ainda que os dois circuitos interajam – afinal, mesmo os mais pobres consomem bens

modernos em algum grau – há uma diferença qualitativa na forma de abastecimento e no padrão de preços entre essas duas esferas. Essa diferenciação se traduz em desigualdade geográfica de preços: o lugar onde se consome influencia quanto se paga e o que se tem acesso.

Um efeito perverso dessa dinâmica urbana são as chamadas “ilhas de preços”, áreas do espaço urbano onde o custo de vida é significativamente mais alto em comparação com outras regiões. As “ilhas de preços” se manifestam de forma clara no varejo alimentar. Em bairros de alto padrão socioeconômico, por exemplo, é comum que supermercados, mesmo pertencendo à mesma rede, pratiquem preços mais elevados, aproveitando-se do maior poder aquisitivo local e da menor sensibilidade dos consumidores a variações de preço.

Um estudo de grande alcance sobre o tema, conduzido pela FGV IBRE em parceria com a Horus (2023), quantificou essa distorção. Ao analisar os preços de uma cesta básica de consumo em diferentes bairros da Grande São Paulo, a pesquisa revelou uma variação de até 36% no valor dos mesmos produtos dependendo da localidade da loja. O estudo identificou que fatores como o perfil de renda da vizinhança, o nível de concorrência local e os custos operacionais, como o aluguel, são determinantes para essas estratégias de precificação segmentada. Essa prática demonstra como as empresas se adaptam à geografia socioeconômica da cidade, maximizando seus lucros com base na capacidade de pagamento de cada público-alvo.

Paradoxalmente, comunidades de baixa renda também podem enfrentar preços mais altos relativos à sua renda, configurando outro tipo de “ilha de preço”. Em periferias desassistidas por grandes redes varejistas, os moradores dependem de mercearias de bairro, pequenos mercadinhos e comércio informal, os quais operam com menor economia de escala. Nesses contextos, verifica-se frequentemente o fenômeno de que “o pobre paga mais caro” por determinados itens básicos. Seja pela falta de concorrência ou pelo custo logístico de abastecimento em áreas afastadas, muitos produtos acabam tendo preço unitário superior ao encontrado em grandes supermercados de áreas centrais.

Outro conceito fundamental nesse debate é o de “desertos alimentares”. Esses são definidos como regiões urbanas ou rurais onde o acesso a alimentos saudáveis e nutritivos é limitado, especialmente para populações de baixa renda. De

acordo com o Departamento de Agricultura dos Estados Unidos (USDA, 2021), considera-se deserto alimentar uma área de baixa renda em que pelo menos 500 pessoas, ou 33% da população local, vivem a mais de 1,6 km de um supermercado (em áreas rurais, essa distância é de 16 km). Segundo dados divulgados pela *Folha de S. Paulo* (2025), cerca de 25% dos habitantes das capitais brasileiras vivem em desertos alimentares, incluindo aproximadamente 9,7% da população de Curitiba.

As implicações sociais dos desertos alimentares são graves. Vigitel (2023, p. 17) aponta que moradores dessas áreas consomem menos alimentos frescos e mais produtos ultraprocessados, o que contribui para altos índices de doenças crônicas. Além disso, há um custo financeiro e temporal: moradores de desertos alimentares frequentemente precisam deslocar-se grandes distâncias para encontrar alimentos in natura, reforçando sua exclusão econômica e territorial.

Sob uma perspectiva crítica, Harvey (2014, p. 22) argumenta que o capitalismo tende a produzir e explorar desigualdades espaciais conforme busca novos nichos de acumulação, resultando em cidades fragmentadas. Para o autor, a luta pelo “direito à cidade” inclui o acesso equitativo a bens essenciais. Assim, desigualdades de precificação e abastecimento devem ser compreendidas como reflexos diretos de políticas urbanas omissas e interesses econômicos concentrados.

Por fim, estudos de caso em cidades brasileiras (Pereira, 2020, p. 8; Pintaui, 1999, p. 65) demonstram como a expansão seletiva de redes varejistas reorganiza o espaço urbano, muitas vezes reproduzindo as desigualdades existentes. Mesmo quando atingem periferias, as grandes redes nem sempre garantem preços acessíveis ou localizações convenientes para os moradores. Diante desse cenário, políticas públicas que incentivem a instalação de mercados populares ou feiras em regiões vulneráveis, bem como o monitoramento sistemático da variação espacial de preços, tornam-se ferramentas essenciais para promover equidade no acesso ao consumo.

2.3 ESTRATÉGIAS DE PRECIFICAÇÃO E COMPETIÇÃO NO VAREJO

A estratégia de precificação constitui um elemento central na competitividade do varejo, especialmente no setor alimentício, caracterizado por margens reduzidas e alta sensibilidade do consumidor a variações de preço. No contexto brasileiro, destaca-se a coexistência de dois formatos dominantes de varejo: o atacarejo (*cash and carry*) e o supermercado tradicional. Cada um adota posicionamentos de mercado distintos, com diferentes estruturas operacionais e abordagens de precificação.

O atacarejo segue uma lógica de preços baixos e alto volume de vendas, operando com estruturas enxutas, sortimento reduzido e serviços limitados, o que resulta em menores custos operacionais (INFOPRICE, 2025). Esse modelo permite aplicar a estratégia conhecida como *Everyday Low Price* (EDLP), que mantém preços continuamente baixos, sem depender de promoções frequentes. Segundo Levy e Weitz (2001, p. 34), o EDLP proporciona vantagens como previsibilidade de demanda, fidelização do consumidor e maior eficiência logística, embora sua eficácia dependa de elevados volumes de venda para compensar as margens reduzidas.

Por outro lado, supermercados tradicionais adotam com frequência a estratégia *High-Low* (Hi-Lo), baseada em preços regulares mais altos, intercalados com promoções agressivas em produtos selecionados. De acordo com Hoch, Drèze e Purk (1994, p. 83), essa abordagem visa atrair clientes por meio de descontos pontuais, ao mesmo tempo em que maximiza margens em produtos fora de oferta. A dinâmica Hi-Lo exige segmentação de produtos entre itens “chamariz” e itens com maior margem, além de investimentos constantes em publicidade e *marketing* promocional (LEVY, 2001, p. 38).

Comparativamente, o EDLP tende a fidelizar consumidores que valorizam estabilidade de preços e economia no longo prazo, enquanto o Hi-Lo atrai consumidores oportunistas, dispostos a adaptar seus hábitos de compra às promoções. Rajiv, Dutta e Dhar (2002, p. 219) destacam que essas estratégias refletem diferentes comportamentos de consumo e coexistem como modelos dominantes no varejo contemporâneo.

A expansão dos atacarejos, que cresceram de 27% para 46% da fatia de faturamento do setor em seis anos (FURTADO; TAMASO, 2024, p. 4), pressionou supermercados tradicionais a reavaliar suas estratégias, aumentando o número de promoções e otimizando suas operações. Essa concorrência direta resultou em um ambiente marcado por uma verdadeira guerra de preços, intensificando a dispersão de preços no espaço urbano.

Por fim, compreender os mecanismos por trás das estratégias EDLP e Hi-Lo permite analisar como diferentes formatos de loja influenciam a variação de preços entre bairros e grupos sociais. Famílias com acesso a atacarejos ou possibilidade de compras em maior volume conseguem economizar de forma significativa, enquanto consumidores com acesso restrito a mercados menores podem arcar com preços proporcionalmente mais elevados. Essa lógica acentua a desigualdade no acesso ao consumo e reforça a importância de se investigar a precificação como fator de estruturação socioespacial do varejo urbano.

2.4 PRINCÍPIOS DA ANÁLISE EXPLORATÓRIA DE DADOS (AED)

A Análise Exploratória de Dados (AED) caracteriza-se por ser uma etapa inicial e essencialmente orientada à descoberta no processo analítico. Diferentemente das abordagens confirmatórias – em que há hipóteses pré-definidas a serem testadas – a AED parte da premissa de que o pesquisador deve “deixar os dados falarem por si” antes de impor qualquer modelo ou suposição teórica rígida. Conforme Tukey (1977, p. 1), pioneiro do conceito de EDA (Exploratory Data Analysis), a exploração de dados assemelha-se ao trabalho de um detetive em busca de pistas e evidências, ao passo que a análise confirmatória se assemelha ao trabalho judicial, avaliando e julgando evidências já coletadas. Essa metáfora ressalta que, na AED, o investigador navega pelos dados sem preconceitos, examinando distribuições e relações de forma livre, a fim de revelar padrões ou irregularidades que não eram antecipados inicialmente. Essa liberdade investigativa é fundamental especialmente quando não se dispõe de uma hipótese prévia clara – caso comum em estudos sobre fenômenos complexos como dispersão de preços urbanos.

A literatura metodológica sustenta que pesquisas exploratórias têm como objetivo principal proporcionar maior familiaridade com o problema, tornando-o mais explícito e possibilitando, a partir daí, a geração de hipóteses (Gil, 2008, p. 41). Em consonância com esse princípio, a AED atua como uma metodologia de reconhecimento de terreno no âmbito dos dados: trata-se de olhar com atenção para o comportamento das variáveis antes de qualquer inferência estatística formal. Isso envolve técnicas descritivas e visuais (histogramas, boxplots, mapas temáticos, entre outros) que permitem identificar padrões, heterogeneidades, valores atípicos (outliers) e correlações não evidentes à primeira vista. Por meio da inspeção gráfica e estatística inicial, o analista pode perceber, por exemplo, se há agrupamentos de preços por região da cidade, divergências marcantes de preço para determinados produtos ou estabelecimentos, ou ainda tendências sazonais que passem despercebidas em uma mera tabela de números. A identificação desses aspectos é crucial em estudos de dispersão de preços e desigualdades geográficas, pois evidencia fenômenos como variações de preço sistemáticas entre bairros ou redes de supermercado – informações que podem estar relacionadas a diferenças socioeconômicas ou a estratégias de precificação diferenciadas.

Autores clássicos e contemporâneos ressaltam a importância dessa fase exploratória não apenas para descobrir padrões escondidos, mas também para orientar as etapas subsequentes da pesquisa. Behrens (1997, p. 131), por exemplo, aponta que a AED tem entre seus objetivos centrais sugerir hipóteses explicativas para os fenômenos observados e avaliar os pressupostos que fundamentarão a análise estatística confirmatória. Além disso, a AED auxilia na seleção de ferramentas e técnicas mais adequadas para tratar os dados em questão e pode prover subsídios para futuras coletas de dados ou experimentos direcionados. Em outras palavras, ao dissecar os dados em busca de regularidades e exceções, o pesquisador adquire embasamento para decidir quais modelos estatísticos ou econométricos poderão ser empregados adiante e quais variáveis ou relações merecem maior atenção. Essa fase investigativa também serve para verificar suposições – por exemplo, se os dados cumprem requisitos de normalidade, homogeneidade ou independência – antes de avançar para testes de hipótese formais.

No contexto de estudos urbanos, de consumo ou de comportamento de

preços – a exemplo da investigação sobre dispersão de preços no varejo alimentício de Curitiba – a AED desempenha um papel indispensável quando não se parte de uma hipótese claramente formulada. Nesses casos, a análise exploratória funciona como uma abordagem de reconhecimento: o pesquisador explora os dados de preço coletados na cidade para descobrir padrões geográficos e temporais sem ideia pré-concebida do que irá encontrar. Essa abordagem guiada pelos dados pode revelar heterogeneidades espaciais importantes, como bairros onde os preços são sistematicamente mais altos ou mais baixos, bem como identificar anomalias (por exemplo, um supermercado específico praticando valores fora do padrão médio do mercado). Tais descobertas iniciais são valiosas por si só – lançando luz sobre a desigualdade geográfica de preços e possivelmente sobre diferentes estratégias de precificação em operação – e também servem de alicerce para etapas posteriores da pesquisa. Por meio da AED, fenômenos antes ocultos emergem, permitindo ao pesquisador formular perguntas e hipóteses mais refinadas a serem examinadas em análises confirmatórias ou modelagens preditivas subsequentes. De fato, a AED complementa a análise confirmatória ao expandir a compreensão do conjunto de dados e muitas vezes antecede o desenvolvimento de modelos explicativos robustos (Behrens, 1997, p. 131).

Em suma, a Análise Exploratória de Dados consolida-se como uma metodologia inicial de descoberta imprescindível em pesquisas quantitativas complexas. Ao priorizar a investigação livre de pressupostos rígidos, a AED permite que o pesquisador identifique padrões, discrepâncias e relações sutis nos dados de preços que poderiam passar despercebidos de outra forma. Esse conhecimento exploratório inicial enriquece a fundamentação do estudo, oferecendo insumos empíricos para a construção teórica e orientando decisões analíticas nas etapas seguintes. No contexto da dispersão de preços no varejo alimentar – permeado por desigualdades espaciais e diferentes comportamentos de mercado – a AED se integra de modo orgânico ao propósito da pesquisa, fornecendo uma visão panorâmica e esclarecedora do fenômeno antes de se avançar para testes estatísticos e inferências causais mais aprofundadas.

2.5 DADOS ABERTOS: FUNDAMENTOS TÉCNICOS E APLICAÇÕES NA GESTÃO PÚBLICA

A disponibilização de dados abertos — informações públicas em formato digital, acessível e reutilizável — consolidou-se como prática essencial para a promoção da transparência, da inovação e do controle social. Segundo a *Open Knowledge Foundation* (2015, p. 4), dados são considerados abertos quando podem ser livremente acessados, utilizados, modificados e redistribuídos por qualquer pessoa, estando sujeitos apenas à exigência de atribuição e compartilhamento sob a mesma licença. No Brasil, esse princípio foi fortalecido pela Lei de Acesso à Informação (Lei nº 12.527/2011) e, mais especificamente, pelo Decreto nº 8.777/2016, que institui a Política de Dados Abertos na administração pública federal, promovendo diretrizes para que bases de dados estejam disponíveis em formatos não proprietários e legíveis por máquina (BRASIL, 2016, p. 1).

A abertura de dados segue três princípios fundamentais: acessibilidade, reusabilidade e interoperabilidade. A acessibilidade refere-se à possibilidade de o dado ser encontrado e acessado por qualquer interessado; a reusabilidade implica o uso de formatos abertos e licenças permissivas; e a interoperabilidade pressupõe o uso de padrões técnicos que viabilizem a integração entre diferentes sistemas e bancos de dados (BRASIL, 2016, p. 2). Nesse contexto, dados abertos de qualidade devem ser estruturados, atualizados, possuir metadados e estar acompanhados de documentação adequada.

Do ponto de vista técnico e científico, os dados abertos têm ampliado significativamente o campo das análises exploratórias, geoespaciais e econômicas, promovendo ganhos de eficiência e confiabilidade. Conforme assinalam Coutinho e Freitas (2021), o acesso a bases públicas de dados permite que cidadãos, jornalistas, acadêmicos e organizações da sociedade civil formulem diagnósticos independentes, revelem padrões ocultos e proponham soluções fundamentadas. Isso torna o cidadão não apenas receptor de informações, mas também agente ativo no aprimoramento das políticas públicas.

No campo da transparência e da responsabilização pública, os dados abertos funcionam como instrumentos de fiscalização descentralizada. Conforme aponta a

OCDE (2021), a abertura de dados públicos amplia a capacidade de monitoramento por múltiplos atores — como cidadãos, imprensa e organizações da sociedade civil — fortalecendo a accountability vertical e horizontal, estimulando boas práticas administrativas e coibindo desvios. Além disso, iniciativas de governo aberto tendem a fomentar ambientes de maior confiança entre o poder público e a sociedade

Na esfera urbana, o potencial dos dados abertos é ainda mais evidente. Bases de dados com recorte geográfico — como endereços de escolas, postos de saúde, acidentes de trânsito ou pesquisas de preços — permitem identificar desigualdades territoriais e apoiar ações corretivas. Como demonstram Cunha e Baracho (2019) em seu estudo com dados de acidentes em Belo Horizonte, a análise de informações abertas é fundamental para subsidiar políticas de cidades inteligentes, permitindo diagnósticos precisos e o planejamento de intervenções para melhorar a segurança e a qualidade de vida urbana.

O presente trabalho exemplifica esse potencial ao utilizar uma base aberta da Prefeitura de Curitiba — a plataforma Clique Economia — para investigar a dispersão de preços no varejo alimentar da cidade. Trata-se de um exemplo concreto de como dados públicos, quando bem estruturados e acessíveis, possibilitam análises ricas e fundamentadas. A reutilização dessas informações por pesquisadores, gestores e consumidores amplia o impacto da coleta governamental de dados, transformando-os em insumo para o desenvolvimento social e econômico.

Em síntese, os dados abertos representam um pilar técnico e democrático para a construção de cidades mais justas, inteligentes e transparentes. Sua integração com ferramentas analíticas modernas — como a análise exploratória de dados — reforça o papel da informação como bem público e vetor de transformação social.

3 METODOLOGIA

Este capítulo detalha o conjunto de procedimentos técnicos e analíticos empregados para transformar os dados brutos da plataforma "Clique Economia" em uma base de conhecimento estruturada e apta a gerar os *insights* apresentados neste trabalho. A metodologia adotada segue um pipeline de Ciência de Dados, compreendendo as etapas de tratamento e enriquecimento de dados (ETL), a construção de um sistema de classificação e a definição das métricas de análise.

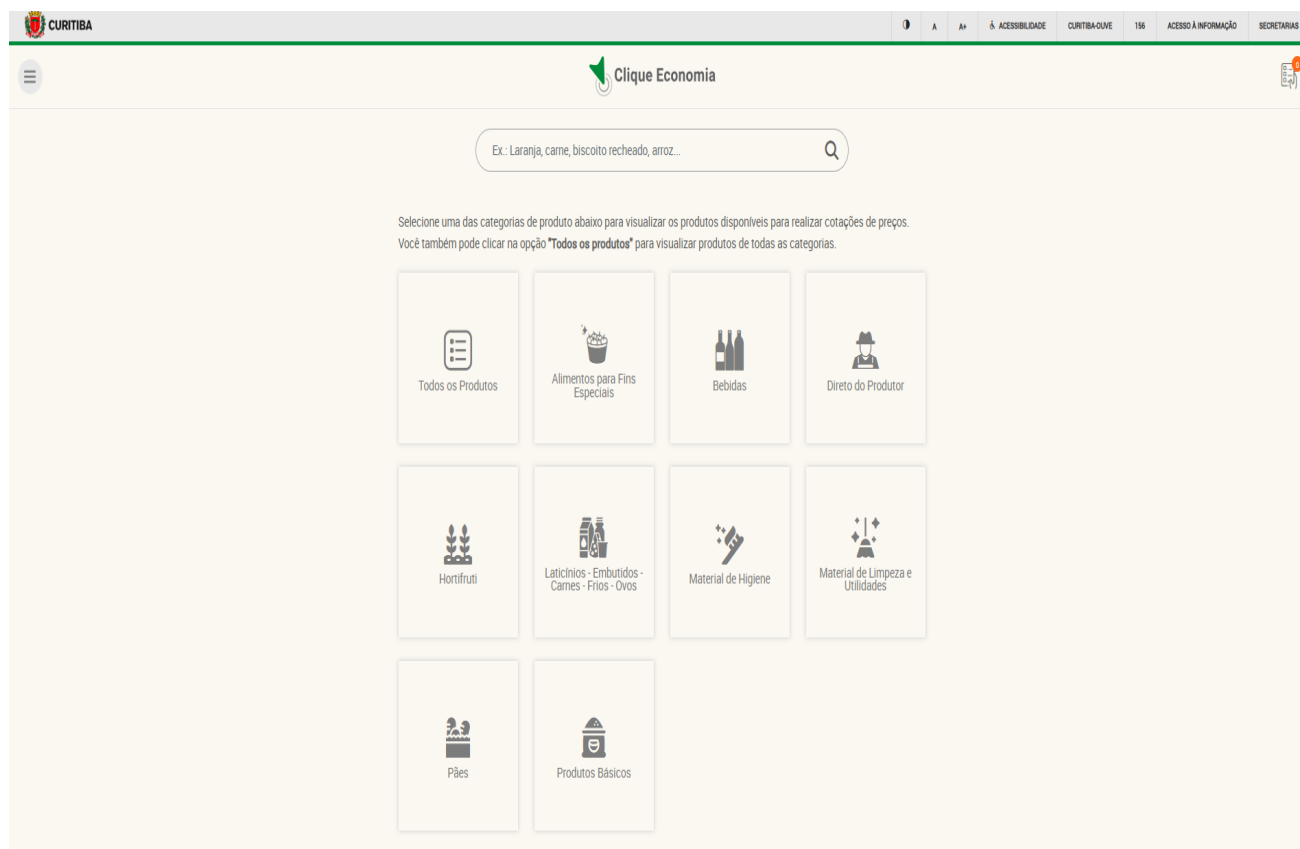
Para atender aos objetivos propostos, a pesquisa foi caracterizada como exploratória, visando aprofundar o conhecimento sobre os dados de preços, e descritiva, buscando detalhar os padrões e características encontradas. A abordagem metodológica é de natureza quantitativa, fundamentada nos dados abertos coletados da plataforma "Clique Economia". O tratamento e a análise dos dados foram realizados em duas etapas principais: primeiramente, um processo de Extração, Transformação e Carga (ETL) para higienização e estruturação, seguido por uma análise exploratória de dados (EDA) para identificar tendências, anomalias e padrões relevantes.

3.1 FONTE DE DADOS: O PROGRAMA CLIQUE ECONOMIA

A matéria-prima desta pesquisa são os dados de cotações de preços de supermercados disponibilizados publicamente pela Prefeitura de Curitiba, por meio da iniciativa Clique Economia. Este programa, coordenado pela Secretaria Municipal de Segurança Alimentar e Nutricional (SMSAN), tem como objetivo central promover a transparência de preços e auxiliar o consumidor curitibano em suas decisões de compra.

O programa realiza uma pesquisa diária de preços em supermercados de médio e grande porte da cidade, cobrindo, em média, 700 itens que englobam produtos da cesta básica, bebidas, laticínios, carnes, hortifrúti, e materiais de higiene e limpeza. As informações são coletadas diretamente nas gôndolas e, para o público geral, são disponibilizadas através de uma interface de consulta no site do programa e no aplicativo Curitiba App, como ilustrado na Figura 1.

FIGURA 1 - INTERFACE DE CONSULTA DO SITE CLIQUE ECONOMIA



FONTE: PREFEITURA DE CURITIBA(2025).

Apesar da existência de uma interface de consulta que permite buscas pontuais de produtos, sua estrutura e funcionalidade apresentam limitações para a realização de análises agregadas, temporais e espaciais. Os dados brutos são disponibilizados em formato de planilhas (.csv), o que possibilita sua extração e processamento por ferramentas analíticas especializadas. É a partir desses dados

primários que esta pesquisa se desenvolve. Para este estudo, foram extraídos, consolidados e tratados os registros diários correspondentes ao ano de 2024, totalizando um volume superior a 700 mil cotações de preços. O Quadro 1, abaixo, detalha as principais variáveis utilizadas neste trabalho, provenientes da base de dados de cotações.

QUADRO 1 - DICIONÁRIO DE DADOS DE PRODUTOS

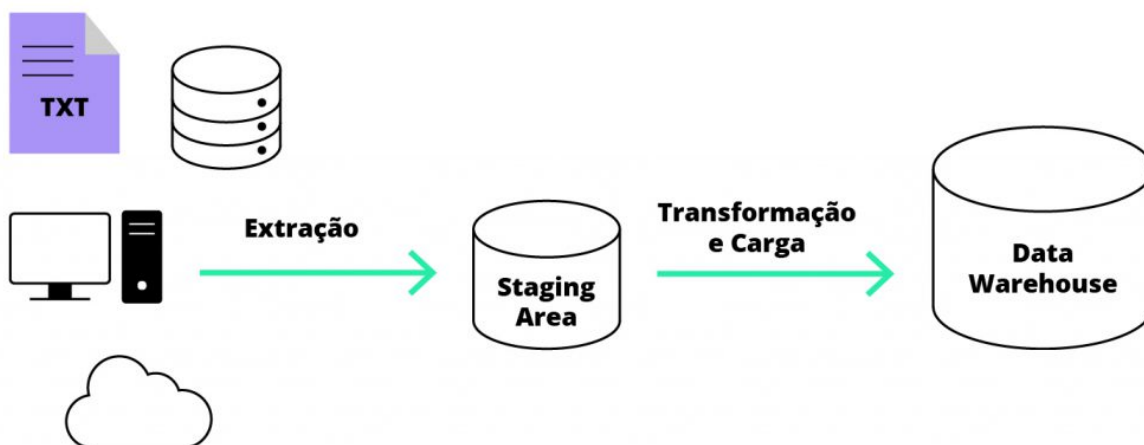
Campo	Descrição	Tipo	Exemplo
data_pesquisa	Data em que a pesquisa foi realizada	Data	"09/12/2024"
id_empresa	Código único da empresa (se disponível)	Texto	"123"
rede	Nome da rede de mercados ou estabelecimentos	Texto	"Condor"
endereco_completo	Endereço completo do estabelecimento	Texto	"Rua Doutor C��zar Pernetta, 483"
id_produto	Identificador ��nico do produto	Texto	"199"
descricao	Descri��o original do produto	Texto	"Caf�� Sol��vel"
preco_regular	Pre��o do produto na g��ndola	Num��rico	"7.49"

FONTE: ELABORADO PELO AUTOR(2025).

3.2 PROCESSO ETL (*EXTRACT, TRANSFORM, LOAD*)

O processo de ETL (Extrair, Transformar, Carregar) é um dos pilares da gestão de dados em projetos analíticos, pois garante que os dados sejam coletados, organizados e preparados para gerar *insights* de forma confiável. Segundo Kimball e Ross (2013), o ETL é a espinha dorsal de qualquer arquitetura de dados, sendo essencial para transformar dados brutos em informações úteis para a tomada de decisão. A Figura 2 ilustra de forma esquemática as etapas do processo, iniciando pela extração de dados de diversas fontes, passando pela transformação, que ocorre em uma área intermediária (*staging area*), onde os dados são limpos, padronizados e enriquecidos, até o carregamento final em um ambiente analítico. No contexto deste trabalho, um *pipeline* ETL foi aplicado para processar a base "Clique Economia", garantindo que os dados estivessem estruturados e prontos para a análise exploratória.

FIGURA 2 - DEMONSTRAÇÃO DO PROCESSO DE ETL

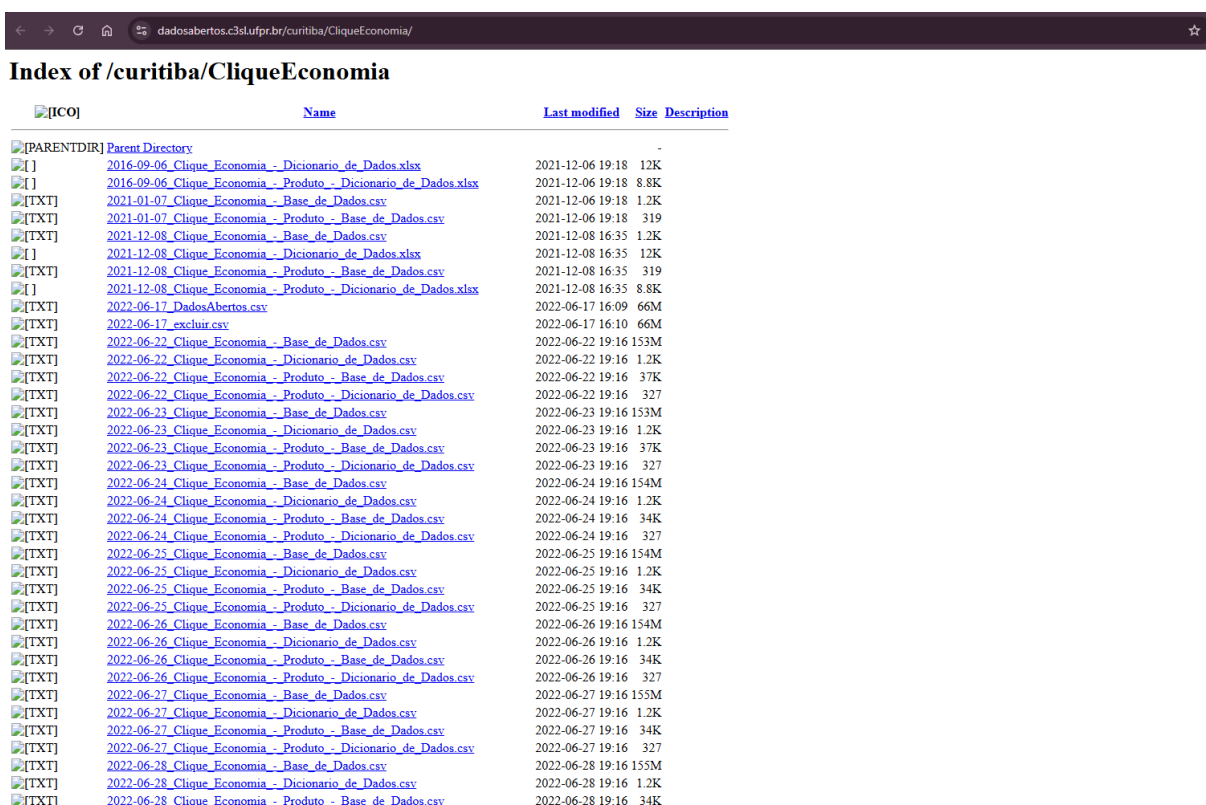


FONTE: ADAPTADO DE MJV INNOVATION(2024).

3.2.1 Extração

A primeira fase do *pipeline* metodológico consistiu na Extração dos dados do repositório público do "Clique Economia", hospedado no portal de dados abertos da C3SL/UFPR. A fonte de dados é caracterizada por uma alta granularidade temporal, com um novo arquivo em formato CSV (.csv) sendo disponibilizado para cada dia de pesquisa, como exemplificado na Figura 3. Para o período definido neste estudo (o ano de 2024), isso representou o desafio inicial de lidar com centenas de arquivos individuais.

FIGURA 3 - EXEMPLO DA ESTRUTURA DE ARQUIVOS DIÁRIOS NO REPOSITÓRIO DE DADOS ABERTOS.



	Name	Last modified	Size	Description
[PARENTDIR]	Parent Directory	-	-	-
[I]	2016-09-06_Clique_Economia_-_Dicionario_de_Dados.xlsx	2021-12-06 19:18	12K	
[I]	2016-09-06_Clique_Economia_-_Produto_-_Dicionario_de_Dados.xlsx	2021-12-06 19:18	8.8K	
[TXT]	2021-01-07_Clique_Economia_-_Base_de_Dados.csv	2021-12-06 19:18	1.2K	
[TXT]	2021-01-07_Clique_Economia_-_Produto_-_Base_de_Dados.csv	2021-12-06 19:18	319	
[TXT]	2021-12-08_Clique_Economia_-_Base_de_Dados.csv	2021-12-08 16:35	1.2K	
[I]	2021-12-08_Clique_Economia_-_Dicionario_de_Dados.xlsx	2021-12-08 16:35	12K	
[TXT]	2021-12-08_Clique_Economia_-_Produto_-_Base_de_Dados.csv	2021-12-08 16:35	319	
[I]	2021-12-08_Clique_Economia_-_Produto_-_Dicionario_de_Dados.xlsx	2021-12-08 16:35	8.8K	
[TXT]	2022-06-17_DadosAbertos.csv	2022-06-17 16:09	66M	
[TXT]	2022-06-17_excluir.csv	2022-06-17 16:10	66M	
[TXT]	2022-06-22_Clique_Economia_-_Base_de_Dados.csv	2022-06-22 19:16	153M	
[TXT]	2022-06-22_Clique_Economia_-_Dicionario_de_Dados.csv	2022-06-22 19:16	1.2K	
[TXT]	2022-06-22_Clique_Economia_-_Produto_-_Base_de_Dados.csv	2022-06-22 19:16	37K	
[TXT]	2022-06-22_Clique_Economia_-_Produto_-_Dicionario_de_Dados.csv	2022-06-22 19:16	327	
[TXT]	2022-06-23_Clique_Economia_-_Base_de_Dados.csv	2022-06-23 19:16	153M	
[TXT]	2022-06-23_Clique_Economia_-_Dicionario_de_Dados.csv	2022-06-23 19:16	1.2K	
[TXT]	2022-06-23_Clique_Economia_-_Produto_-_Base_de_Dados.csv	2022-06-23 19:16	37K	
[TXT]	2022-06-23_Clique_Economia_-_Produto_-_Dicionario_de_Dados.csv	2022-06-23 19:16	327	
[TXT]	2022-06-24_Clique_Economia_-_Base_de_Dados.csv	2022-06-24 19:16	154M	
[TXT]	2022-06-24_Clique_Economia_-_Dicionario_de_Dados.csv	2022-06-24 19:16	1.2K	
[TXT]	2022-06-24_Clique_Economia_-_Produto_-_Base_de_Dados.csv	2022-06-24 19:16	34K	
[TXT]	2022-06-24_Clique_Economia_-_Produto_-_Dicionario_de_Dados.csv	2022-06-24 19:16	327	
[TXT]	2022-06-25_Clique_Economia_-_Base_de_Dados.csv	2022-06-25 19:16	154M	
[TXT]	2022-06-25_Clique_Economia_-_Dicionario_de_Dados.csv	2022-06-25 19:16	1.2K	
[TXT]	2022-06-25_Clique_Economia_-_Produto_-_Base_de_Dados.csv	2022-06-25 19:16	34K	
[TXT]	2022-06-25_Clique_Economia_-_Produto_-_Dicionario_de_Dados.csv	2022-06-25 19:16	327	
[TXT]	2022-06-26_Clique_Economia_-_Base_de_Dados.csv	2022-06-26 19:16	154M	
[TXT]	2022-06-26_Clique_Economia_-_Dicionario_de_Dados.csv	2022-06-26 19:16	1.2K	
[TXT]	2022-06-26_Clique_Economia_-_Produto_-_Base_de_Dados.csv	2022-06-26 19:16	34K	
[TXT]	2022-06-26_Clique_Economia_-_Produto_-_Dicionario_de_Dados.csv	2022-06-26 19:16	327	
[TXT]	2022-06-27_Clique_Economia_-_Base_de_Dados.csv	2022-06-27 19:16	155M	
[TXT]	2022-06-27_Clique_Economia_-_Dicionario_de_Dados.csv	2022-06-27 19:16	1.2K	
[TXT]	2022-06-27_Clique_Economia_-_Produto_-_Base_de_Dados.csv	2022-06-27 19:16	34K	
[TXT]	2022-06-27_Clique_Economia_-_Produto_-_Dicionario_de_Dados.csv	2022-06-27 19:16	327	
[TXT]	2022-06-28_Clique_Economia_-_Base_de_Dados.csv	2022-06-28 19:16	155M	
[TXT]	2022-06-28_Clique_Economia_-_Dicionario_de_Dados.csv	2022-06-28 19:16	1.2K	
[TXT]	2022-06-28_Clique_Economia_-_Produto_-_Base_de_Dados.csv	2022-06-28 19:16	34K	

FONTE: DADOSABERTOS.C3SL.UFPR.BR (2025).

Para executar essa tarefa de forma eficiente, escalável e replicável, foi desenvolvido um script em Python, utilizando as bibliotecas *requests*, *os* e *pandas*, responsáveis, respectivamente, por requisições HTTP, interação com o sistema operacional e manipulação de dados. O *script* foi projetado para, primeiramente,

listar todos os arquivos disponíveis no repositório. Em seguida, aplicou-se um conjunto de filtros para selecionar apenas os arquivos pertinentes ao recorte desta pesquisa. Foram definidas duas regras de seleção:

1. **Filtro Temporal:** Apenas arquivos cujo nome continha o marcador "2024" foram considerados, delimitando o período de interesse.
2. **Filtro de Conteúdo:** Foram selecionados exclusivamente os arquivos que continham o termo "Base_de_Dados" na nomenclatura, excluindo arquivos complementares, como dicionários de dados ou metadados auxiliares.

Após a aplicação dos filtros, o *script* procedeu com o *download* automatizado dos arquivos selecionados, armazenando-os de forma organizada em um diretório local. Durante esse processo, foi gerado um *log* simples contendo a lista dos arquivos baixados, datas de coleta e eventuais falhas de conexão. Mecanismos básicos de reconexão foram implementados para garantir a completude da coleta em caso de erro de rede.

Embora as etapas de unificação e tratamento dos dados sejam detalhadas posteriormente, na fase de Transformação, o processo de extração resultou em um repositório local completo e organizado, reunindo os arquivos brutos com as cotações diárias referentes a todo o ano de 2024, fundamentais para o desenvolvimento das análises subsequentes.

A versão final do script de extração, com detalhamento de parâmetros e configurações, encontra-se documentada no Apêndice 1 deste trabalho.

3.2.2 Transformação e limpeza dos dados

O primeiro passo da fase de Transformação foi a unificação programática de todos os arquivos CSV diários, previamente extraídos, em uma única estrutura de dados. Essa tarefa foi realizada por meio de um script desenvolvido em Python, utilizando principalmente a biblioteca *pandas*, amplamente reconhecida por sua eficiência no processamento de grandes volumes de dados tabulares (McKinney, 2012).

O script foi projetado para iterar automaticamente sobre os arquivos armazenados localmente, realizando a leitura sequencial de cada CSV e, posteriormente, concatenando-os em um único DataFrame. Essa consolidação resultou em um *dataset* inicial contendo mais de 700 mil registros, correspondente ao total de cotações diárias registradas ao longo do ano de 2024. O código-fonte completo desta rotina de unificação encontra-se documentado no Apêndice 1 deste trabalho.

Imediatamente após a unificação, foi conduzida uma pré-validação e limpeza estrutural, com o objetivo de garantir a integridade básica da base de dados antes da aplicação de transformações mais complexas. As principais ações realizadas nesta etapa incluíram:

- **Exclusão de Colunas Irrelevantes:** Remoção de campos não utilizados nas análises, como `preco_atacado`, `preco_promocao`, `preco_fidelidade` e `codigo_categoria`, a fim de otimizar a estrutura e reduzir o volume de memória processado nas etapas seguintes.
- **Tratamento de Valores Nulos:** Eliminação de registros com ausência de informações essenciais em colunas críticas, como `preco_regular` ou `descricao`, cuja falta inviabilizaria análises relacionadas a preços e categorização de produtos.
- **Filtragem de Dados Inválidos:** Exclusão de registros contendo preços zerados ou negativos, interpretados como erros de entrada ou falhas durante o processo de coleta original. A manutenção desses valores traria distorções estatísticas significativas.
- **Remoção de Duplicatas:** Realização de um processo de verificação para identificação e exclusão de registros duplicados, evitando a duplicação artificial de contagens e análises estatísticas enviesadas.
- **Correção de Tipos de Dados:** Cada coluna foi rigorosamente convertida para o tipo de dado apropriado. Destacam-se: a conversão da variável `data_pesquisa` para o formato `datetime`; a padronização de identificadores como `id_produto` e `id_empresa` para o tipo `string`, a fim de evitar problemas com notação científica; e a transformação da coluna `preco_regular` para o tipo

float, após a substituição dos separadores decimais (de vírgula para ponto).

Finalizada essa etapa, obteve-se uma base de dados consolidada e estruturada, com qualidade adequada para as próximas fases de processamento, que envolveram a normalização e o enriquecimento das informações.

3.2.2.1 Enriquecimento das informações

Concluída a limpeza estrutural, a etapa seguinte da transformação foi o enriquecimento dos dados, processo que teve como principal objetivo agregar novas variáveis analíticas ao *dataset*. Esta fase foi fundamental para criar os atributos que permitiram as análises geográficas, comparativas e categóricas que sustentam os objetivos deste trabalho. Seguindo as melhores práticas de engenharia de dados descritas por Kimball e Ross (2013), as atividades de enriquecimento foram organizadas nas seguintes subetapas.

a) Normalização Geográfica dos Endereços: Para possibilitar as análises espaciais, tornou-se imperativo estruturar a informação de localização de cada estabelecimento de forma padronizada. A variável original `endereco_completo`, um campo de texto livre e não padronizado, impedia qualquer agrupamento direto por região. Para resolver este desafio, foi realizado um processo de normalização manual para os 104 endereços únicos identificados na base de dados. Em uma planilha auxiliar, cada endereço textual foi cuidadosamente decomposto em seus campos específicos: `logradouro_tipo`, `logradouro_nome`, `numero`, `bairro`, `cidade` e `estado`. Durante esta etapa, todos os bairros foram conferidos e ajustados para garantir conformidade com a divisão administrativa oficial de Curitiba. Este processo manual resultou na criação de um "gabarito de endereços" de alta precisão. Por fim, este gabarito foi integrado programaticamente à base principal de cotações por meio da variável `id_empresa`, garantindo a vinculação correta entre cada estabelecimento e sua localização estruturada. O arquivo resultante deste processo (`endereco_normalizado_COM_ID.csv`) encontra-se documentado no Apêndice 2.

b) **Padronização das Redes de Varejo:** Uma análise preliminar revelou que uma mesma rede de supermercados era registrada com múltiplas nomenclaturas na base de dados (ex: “Condor”, “Hipermercado Condor”, “Condor João Bettega”). Para garantir a consistência e a validade das análises comparativas, foi desenvolvido um dicionário de padronização que aplicou uma nomenclatura única e representativa para cada grupo varejista. Essa rotina de mapeamento foi implementada diretamente no *script* de enriquecimento, assegurando a consistência em todas as análises subsequentes.

c) **Desenvolvimento do Sistema de Categorização Hierárquica de Produtos:** A categorização dos produtos foi, sem dúvida, a etapa mais desafiadora e de maior contribuição metodológica dentro da fase de enriquecimento. A variável descrição, por se tratar de um campo de texto livre, apresentava elevada diversidade terminológica, com variações, abreviações e descrições ambíguas, o que inviabilizava qualquer tentativa de análise agregada por tipo de produto. Para superar esse obstáculo, foi desenvolvido um sistema híbrido de classificação, implementado em Python, que combinou regras manuais e lógicas automáticas. Inicialmente, identificaram-se casos mais complexos e sujeitos a erro de contexto, como por exemplo o produto “leite de coco”, que poderia ser erroneamente classificado como “leite” caso fosse utilizada apenas uma busca simples por palavras-chave. Para esses casos específicos, foi criado um gabarito de exceções, conhecido como “Gabarito de Ouro”, no qual cada produto ambíguo foi manualmente revisado e classificado com prioridade máxima.

Após a aplicação das exceções, os registros restantes foram submetidos a um dicionário de regras, construído com base em palavras-chave representativas de cada categoria. A estruturação desse dicionário seguiu o princípio da hierarquia de especificidade, ou seja, termos compostos como “creme de leite” foram processados antes de termos mais genéricos como “leite”, reduzindo assim a ocorrência de erros de classificação. Essa lógica foi operacionalizada por meio de expressões regulares (*regex*), permitindo que apenas ocorrências exatas de palavras inteiras fossem consideradas, evitando falsos positivos na categorização.

Ao final do processo, o *dataset* passou a contar com duas novas variáveis categóricas: *categoria_n1*, representando a macrocategoria dos produtos, e *categoria_n2*, detalhando as subcategorias específicas. O código-fonte utilizado para implementar este sistema de categorização encontra-se integralmente documentado no Apêndice 3, garantindo a transparência e a reprodutibilidade metodológica desta etapa.

3.2.2.2 Validação preliminar do dataset

Concluída a etapa de unificação e enriquecimento da base de dados, foi conduzida uma validação estrutural preliminar, com o objetivo de assegurar a consistência, integridade e qualidade do dataset antes da execução das análises exploratórias e estatísticas finais. Para essa finalidade, foi desenvolvido um script de diagnóstico em Python, cujo relatório detalhado encontra-se documentado no Apêndice 4 deste trabalho. O processo de validação abrangeu cinco dimensões principais:

1. **Volume de Dados:** Ao final da extração e unificação, o dataset consolidado apresentou um total de 740.670 registros, representando cotações diárias realizadas ao longo de todo o ano de 2024.
2. **Verificação de Valores Nulos:** Foi confirmada a ausência de valores nulos em todas as colunas essenciais para a análise, incluindo *preco_regular*, *descricao*, *data_pesquisa* e *id_produto*. Esse resultado atesta a completude estrutural mínima necessária para análises estatísticas e geográficas.
3. **Deteção de Duplicatas:** A checagem por duplicatas revelou a existência de 288 registros completamente duplicados (linhas idênticas). Embora o volume seja pequeno em relação ao total da base ($< 0,05\%$), foi registrada a recomendação de exclusão desses registros para evitar distorções em análises agregadas.
4. **Análise de Anomalias de Preço (*Outliers*):** Foram calculadas estatísticas descritivas para a variável *preco_regular*, com destaque para os seguintes valores:

Média = R\$ 10,72 | Mediana = R\$ 6,99 | Desvio padrão = R\$ 10,50 | Mínimo = R\$ 0,08 | Máximo = R\$ 139,00.

Os dez preços mais elevados foram inspecionados individualmente. A análise constatou que os valores máximos referem-se a produtos de alto valor agregado, como fórmulas infantis e cortes de carne nobre, descartando a hipótese de erro de digitação.

5. **Consistência da Categorização e Geolocalização:** A avaliação da distribuição nas colunas categoria_n1 e categoria_n2 indicou uma alocação satisfatória, com ausência de registros classificados como "OUTROS". Além disso, todos os registros apresentaram endereço completo, com bairro devidamente identificado, o que garante a viabilidade para as análises espaciais propostas.

A Figura 4, apresentada a seguir, exibe a saída integral do relatório de validação de qualidade gerado pelo script:

FIGURA 4 – RELATÓRIO DE VALIDAÇÃO ESTRUTURAL PRELIMINAR DO DATASET

```

Iniciando Script Final de Validação de Qualidade...
Dataset com 740670 linhas carregado com sucesso.

=====
--- 1. VERIFICAÇÃO DE VALORES NULOS ---
=====
✅ STATUS: Ótimo! Nenhuma coluna possui valores nulos.

=====
--- 2. VERIFICAÇÃO DE DUPLICATAS ---
=====
Número de linhas completamente duplicadas encontradas: 288
AVISO: Recomenda-se investigar se essas duplicatas são esperadas ou se devem ser removidas.

=====
--- 3. VERIFICAÇÃO DE ANOMALIAS DE PREÇO ---
=====

[3.1] Estatísticas Descritivas da coluna 'preco_regular':
count      740670.00
mean        10.72
std         10.50
min          0.08
25%          4.39
50%          6.99
75%         12.99
max         139.00
Name: preco_regular, dtype: float64

[3.2] Top 10 preços mais altos no dataset para verificação manual:
   descricao      rede      bairro      preco_regular
0  FORMULA INFANTIL -APTAMIL 2 -800 GR  MICHEL      cic      139.00
1  FORMULA INFANTIL -APTAMIL PROFUTURA 2 -800 GR  MICHEL      cic      139.00
2  FORMULA INFANTIL -APTAMIL 2 -800 GR  MICHEL      cic      139.00
3  FORMULA INFANTIL -APTAMIL PROFUTURA 2 -800 GR  MICHEL      cic      139.00
4  CARNE BOVINA FILE MIGNON S/ OSSO -( + ) BARATO... FESTIVAL  xaxim      129.98
5  FORMULA INFANTIL -APTANUTRI PROFUTURA 3 -800 GR  ANGELONI  bigorriho  128.98
6  CARNE BOVINA FILE MIGNON S/ OSSO -( + ) BARATO... ANGELONI  bigorriho  121.98
7  FORMULA INFANTIL -APTANUTRI PROFUTURA 3 -800 GR  ANGELONI  bigorriho  120.98
8  FORMULA INFANTIL -APTANUTRI PROFUTURA 3 -800 GR  ANGELONI  agua verde  120.98
9  FORMULA INFANTIL -APTANUTRI PROFUTURA 3 -800 GR  ANGELONI  agua verde  120.98

=====
--- 4. VERIFICAÇÃO DA CONSISTÊNCIA DAS CATEGORIAS ---
=====

[4.1] Contagem de itens por Macro Categoria (Nível 1):
categoria_n1
MERCERIA      299138
HORTIFRUTI    136536
LATICINIOS_E_FRIOS  100217
BEBIDAS       54621
LIMPEZA_E_UTILIDADES  54135
CARNES        50257
HIGIENE_PESSOAL  24284
PADARIA       14847
INFANTIL      6635
Name: count, dtype: int64

[4.2] Total de produtos que restaram na subcategoria 'OUTROS': 0
✅ STATUS: Perfeito! Nenhum produto foi classificado como 'OUTROS'.

=====
--- 5. VERIFICAÇÃO DA INTEGRIDADE DOS ENDEREÇOS ---
=====

Total de registros sem informação de bairro: 0
✅ STATUS: Perfeito! Todos os registros possuem endereços completos.

=====
--- VALIDAÇÃO DE INCONSISTÊNCIAS CONCLUÍDA ---
=====

```

FONTE: ELABORADO PELO AUTOR (2025).

O conjunto dessas verificações preliminares oferece evidências suficientes de que o dataset consolidado encontra-se estruturalmente íntegro, consistente e apto a sustentar com segurança as análises descritivas, espaciais e comparativas apresentadas nos capítulos seguintes.

3.2.3 Carregamento dos dados

A etapa final do pipeline ETL consistiu no Carregamento (*Load*), cujo objetivo foi persistir o *DataFrame* final, já tratado e enriquecido, em um formato de arquivo estável, portátil e otimizado para as ferramentas de análise e visualização utilizadas neste trabalho.

Antes da exportação, foram realizados procedimentos finais de governança dos dados para garantir a qualidade do artefato final. Variáveis transitórias criadas durante a etapa de transformação, como “descricao_norm”, foram removidas para evitar redundância. As colunas do dataset foram então logicamente reordenadas para facilitar a interpretação e a manipulação nas fases seguintes.

Optou-se por salvar o dataset final no formato CSV (Comma-Separated Values), utilizando o ponto e vírgula (;) como delimitador, devido à sua ampla compatibilidade. A codificação UTF-8-BOM foi especificamente escolhida para garantir a correta interpretação de caracteres especiais e acentuação pela ferramenta de análise escolhida (python).

Como última verificação de integridade, aplicou-se uma rotina para eliminar quaisquer registros completamente duplicados que pudessem ter surgido durante o processo, assegurando a unicidade de cada linha no dataset final.

O processo foi concluído com a exportação do arquivo definitivo, nomeado *CliqueEconomia_TCC_DATASET_PRONTO.csv*, que se tornou a fonte única e confiável (*single source of truth*) para todo o Capítulo 4: Análise e Discussão dos Resultados. O script Python contendo a rotina exata de carregamento, com todos os seus parâmetros, encontra-se documentado no Apêndice 4.

3.3 DEFINIÇÃO DA "CESTA DE CATEGORIAS ESSENCIAIS"

Com o intuito de estabelecer um indicador de custo que fosse comparável entre as diferentes regiões e redes de varejo, este trabalho adotou uma abordagem orientada por dados (*data-driven*), em detrimento de utilizar um modelo teórico externo ou uma cesta básica oficial previamente estabelecida. A escolha metodológica teve como premissa construir uma métrica que refletisse, de maneira fiel, o sortimento efetivamente disponível no dataset e o comportamento de uma compra doméstica típica.

Assim, foi concebido o conceito de “Cesta de Categorias Essenciais”, um construto analítico elaborado para representar o custo médio de uma aquisição que envolvesse os principais grupos de consumo alimentar e de necessidades básicas. Essa cesta foi composta pelas cinco macrocategorias (*categoria_n1*) previamente definidas durante o processo de enriquecimento dos dados:

- **MERCEARIA:** Inclui alimentos não perecíveis e de longa duração, como arroz, feijão, açúcar, café, óleo, massas, enlatados, molhos, farinhas, biscoitos, condimentos e conservas.
- **HORTIFRUTI:** Abrange frutas, legumes, verduras e ovos, refletindo a parte mais perecível e fresca da alimentação diária.
- **CARNES:** Contempla cortes de carne bovina, suína, de frango e outros tipos de proteína animal vendidos nas gôndolas refrigeradas dos supermercados.
- **LATICINIOS_E_FRIOS:** Engloba leite, queijos, iogurtes, manteigas, margarinas e embutidos, como presunto e mortadela.
- **LIMPEZA_E_UTILIDADES:** Inclui itens de limpeza doméstica e utilidades básicas, como detergentes, sabão, desinfetantes, papel higiênico e utensílios simples de uso cotidiano.

A seleção dessas categorias de produtos foi um passo metodológico para delimitar um escopo de análise abrangente, simulando uma cesta de compras que engloba itens de diferentes setores de um supermercado (como mercearia, hortifrúti, higiene e açougue). O cálculo do custo da cesta, utilizado como principal indicador

comparativo entre bairros e redes de varejo, foi estruturado como um índice composto, definido pela soma dos preços medianos de cada uma das cinco macrocategorias. A escolha pela mediana como medida de tendência central foi intencional e metodologicamente fundamentada, visando reduzir o efeito de valores extremos (outliers), comuns em dados de preços no varejo (FÁVERO; BELFIORE, 2017).

Essa metodologia resultou em um *benchmark* internamente consistente, capaz de minimizar o viés decorrente da assimetria de sortimento entre os diferentes estabelecimentos. Desse modo, a “Cesta de Categorias Essenciais” tornou-se uma variável-chave para as análises comparativas de dispersão de preços e competitividade que compõem os capítulos analíticos deste trabalho.

3.4 METODOLOGIA DE ANÁLISE EXPLORATÓRIA E ESCOLHA DAS MÉTRICAS ESTATÍSTICAS

Com o dataset consolidado, limpo e enriquecido, a etapa seguinte deste trabalho concentrou-se na aplicação de técnicas de Análise Exploratória de Dados (EDA – *Exploratory Data Analysis*), com o objetivo de identificar padrões, variações e comportamentos relevantes na estrutura de preços do varejo alimentício curitibano.

A escolha pela EDA como abordagem central foi fundamentada na necessidade de compreender a estrutura interna dos dados antes de qualquer modelagem mais avançada, conforme recomendado por Tukey (1977) e reforçado por Shmueli et al. (2017), que destacam a importância da EDA em projetos com grandes volumes de dados e com múltiplas variáveis categóricas e numéricas.

3.4.1 Ferramenta Computacional Adotada

Todas as análises foram realizadas utilizando a linguagem Python, executada localmente no ambiente de desenvolvimento Visual Studio Code (VS Code). As rotinas analíticas foram desenvolvidas integralmente em formato de scripts, sem o uso de notebooks interativos, o que permitiu maior controle sobre o fluxo de execução e o versionamento das etapas de processamento. As principais bibliotecas

utilizadas foram:

- **Pandas:** Para manipulação e análise de estruturas tabulares.
- **Numpy:** Para cálculos estatísticos e operações numéricas.
- **Matplotlib e seaborn:** Para criação de gráficos exploratórios.
- **Re (expressões regulares):** Durante a preparação e categorização de dados.

Essa escolha se deu pela capacidade dessas ferramentas em lidar com grandes volumes de dados de forma eficiente, além de oferecerem recursos adequados para a execução de análises exploratórias e a geração de visualizações estatísticas.

3.4.2 Eixos de Análise

A estrutura analítica do trabalho foi organizada a partir de três principais dimensões de análise, alinhadas aos objetivos específicos previamente definidos:

- **Análise Geográfica:** Focada na comparação do custo da cesta entre os diferentes bairros de Curitiba, utilizando a variável “bairro” como chave de agrupamento.
- **Análise Competitiva:** Direcionada à comparação entre as principais redes de supermercado (rede), com o objetivo de avaliar diferenças nas políticas de precificação.
- **Análise Temporal:** Explorando a variação dos preços ao longo dos meses de 2024, a partir da variável “data_pesquisa”, para identificar sazonalidades e tendências.

3.4.3 Métricas Estatísticas Seleccionadas

Para cada eixo de análise, foram escolhidas as seguintes métricas estatísticas, com base em sua capacidade de representar centralidade, dispersão e variações relativas:

- **Mediana:** Seleccionada como medida de tendência central, a mediana foi aplicada para reduzir o impacto de valores extremos (outliers), comuns em dados de preços no varejo (FÁVERO; BELFIORE, 2017).
- **Desvio-Padrão e Coeficiente de Variação (CV):** Essas medidas de dispersão foram utilizadas para avaliar a volatilidade dos preços dentro de cada bairro, rede ou categoria de produto. O CV, por ser uma métrica relativa, permitiu comparar a variação de preços mesmo entre categorias com níveis absolutos de preço muito diferentes.
- **Rankings:** Foram elaborados rankings para identificar os bairros mais baratos, as redes com menor custo médio e as categorias de produtos com maior ou menor estabilidade de preços.

3.4.4 Visualização dos Resultados

Para facilitar a interpretação dos achados, os resultados da análise exploratória foram apresentados por meio de diferentes formas de visualização:

- **Gráficos de Linha:** Para representar a evolução temporal dos preços.
- **Boxplots:** Para demonstrar a dispersão e presença de outliers por bairro e por rede.
- **Tabelas de Ranking:** Para comparação direta entre bairros e redes.

- **Heatmaps:** Aplicados especialmente nas análises cruzadas entre redes e categorias, evidenciando a matriz de competitividade.

Todos os gráficos e tabelas gerados encontram-se apresentados e discutidos no Capítulo 4 – Análise e Discussão dos Resultados.

3.5 LIMITAÇÕES METODOLÓGICAS E CONSIDERAÇÕES ÉTICAS

3.5.1 Limitações Metodológicas

Ao longo deste trabalho, algumas limitações metodológicas foram identificadas e são reconhecidas a fim de assegurar a transparência científica.

A primeira delas refere-se à própria estratégia de estruturação e classificação dos dados. Embora o processo de enriquecimento e categorização tenha sido conduzido de forma criteriosa, a ausência de um critério previamente validado ou de uma taxonomia oficial para a classificação dos produtos demandou o uso de soluções pragmáticas e iterativas. A elaboração do sistema de classificação baseou-se predominantemente em regras heurísticas e em revisões manuais sucessivas, o que, apesar de garantir resultados satisfatórios, pode ter introduzido vieses ou subjetividades em algumas decisões de categorização. Caso um método mais rigoroso e previamente estruturado tivesse sido adotado desde o início, a etapa de enriquecimento teria apresentado maior eficiência e menor margem para interpretações individuais.

Outro aspecto relevante foi o recorte temporal da análise, limitado exclusivamente ao ano de 2024. Essa definição, embora tenha atendido aos objetivos do estudo e às condições de viabilidade operacional, restringe os resultados ao comportamento dos preços em um único período anual, inviabilizando comparações com séries históricas ou com ciclos econômicos mais amplos.

Além disso, a opção por uma abordagem estritamente exploratória e descritiva, sem a aplicação de técnicas inferenciais ou preditivas, representa uma limitação metodológica. Tal escolha refletiu o escopo definido para esta pesquisa, mas restringiu as possibilidades de generalização ou previsão de comportamento.

futuro dos preços.

3.5.2 Considerações Éticas

Do ponto de vista ético, todas as etapas deste estudo foram conduzidas em conformidade com os princípios de uso responsável de dados públicos e com as boas práticas em Ciência de Dados. As informações utilizadas são provenientes de bases abertas, públicas e anonimizadas, disponibilizadas oficialmente pelo portal de Dados Abertos da Prefeitura de Curitiba e pelo Centro de Computação Científica e Software Livre da UFPR (C3SL).

Em nenhuma fase da pesquisa foram manipulados dados de caráter pessoal, sensível ou que pudessem identificar consumidores ou representantes comerciais de forma individualizada. Todas as análises tiveram caráter exclusivamente acadêmico e voltado ao interesse público, respeitando os princípios de transparência, integridade e responsabilidade social na utilização de informações governamentais.

Por fim, todas as etapas do processo analítico, incluindo os scripts, rotinas de tratamento e regras de categorização, foram documentadas nos apêndices deste trabalho, garantindo a reprodutibilidade e a auditabilidade das análises, em consonância com os preceitos da ciência aberta e da ética em pesquisa com dados secundários.

4 ANÁLISE E DISCUSSÃO DOS RESULTADOS

Este capítulo apresenta a análise e discussão dos resultados obtidos a partir do *dataset* tratado e enriquecido conforme a metodologia descrita no capítulo 3. Com o objetivo de responder à questão de pesquisa, a exposição dos achados foi estruturada para, primeiramente, caracterizar a dispersão e a volatilidade dos preços entre as categorias de produtos. Em seguida, a análise aprofunda-se nas dimensões geográfica e competitiva, mapeando os custos por bairro e as estratégias das redes de varejo. Por fim, examina-se a evolução temporal dos preços ao longo de 2024, culminando na síntese dos resultados em um guia prático para o consumidor.

4.1. CARACTERIZAÇÃO DA DISPERSÃO DE PREÇOS EM CURITIBA

A primeira etapa da análise consiste em apresentar um panorama geral dos preços observados no dataset e, em seguida, aprofundar na identificação de padrões de instabilidade que afetam a previsibilidade dos custos no varejo curitibano. Essa abordagem permite compreender, inicialmente, a escala e a distribuição dos preços, para então avançar na análise das subcategorias mais voláteis. Como discutido na fundamentação teórica (STIGLER, 1961), a dispersão de preços é uma característica típica de mercados com assimetria de informação — e seu mapeamento constitui o ponto de partida para os achados deste capítulo.

4.1.1 Visão geral dos preços

Além da identificação das categorias mais voláteis, é importante estabelecer um panorama geral dos preços praticados no varejo curitibano ao longo de 2024. Para isso, foram calculadas estatísticas descritivas clássicas da variável `preco_regular`, conforme apresentadas a seguir:

TABELA 1 – DISTRIBUIÇÃO DOS PREÇOS REGULARES OBSERVADOS

Estatística	Valor
Contagem (n)	740.670
Média	R\$ 10,72
Mediana (50%)	R\$ 6,99
Desvio padrão	R\$ 10,50
Mínimo	R\$ 0,08
1º quartil (25%)	R\$ 4,39
3º quartil (75%)	R\$ 12,99
Máximo	R\$ 139,00

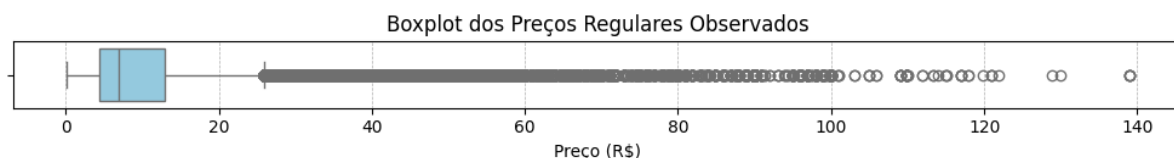
FONTE: ELABORADO PELO AUTOR (2025).

Esses dados revelam uma clara assimetria positiva na distribuição dos preços: enquanto a mediana está fixada em R\$ 6,99, indicando que metade dos produtos custa abaixo desse valor, o preço máximo registrado atinge R\$ 139,00 — quase vinte vezes maior. Esse desnível reflete a amplitude extrema da cesta analisada, que contempla desde produtos de baixíssimo custo unitário, como hortaliças vendidas por unidade, até itens com valor agregado elevado, como fórmulas infantis especiais, carnes *premium* e produtos importados.

Além disso, a grande dispersão nos preços é reforçada pelo desvio padrão de R\$ 10,50, valor quase equivalente à média e fortemente influenciado por *outliers*. Essa alta variabilidade confirma a inadequação da média aritmética para esta análise, validando a escolha da mediana como métrica central de comparação.

Para visualizar essa distribuição de forma mais intuitiva, o gráfico 1 apresenta o boxplot dos preços regulares observados:

GRÁFICO 1 – BOXPLOT DOS PREÇOS REGULARES OBSERVADOS



FONTE: ELABORADO PELO AUTOR (2025).

A visualização mostra que a maior parte dos preços está concentrada abaixo de R\$ 20,00, com o corpo da caixa (entre o 1º e o 3º quartil) cobrindo a faixa entre R\$ 4,39 e R\$ 12,99. A presença de uma cauda longa à direita, acompanhada por uma grande quantidade de outliers, evidencia a existência de valores atípicos que distorcem a distribuição — muitos deles referentes a nichos específicos do varejo alimentar curitibano.

Essa análise reforça a importância de se utilizar medidas robustas e apropriadas para representar o custo de vida do consumidor, e serve como base empírica para as decisões metodológicas adotadas neste estudo.

As estatísticas descritivas e a visualização gráfica foram geradas por meio de rotinas desenvolvidas em Python, cujo código encontra-se documentado no Apêndice 5 deste trabalho.

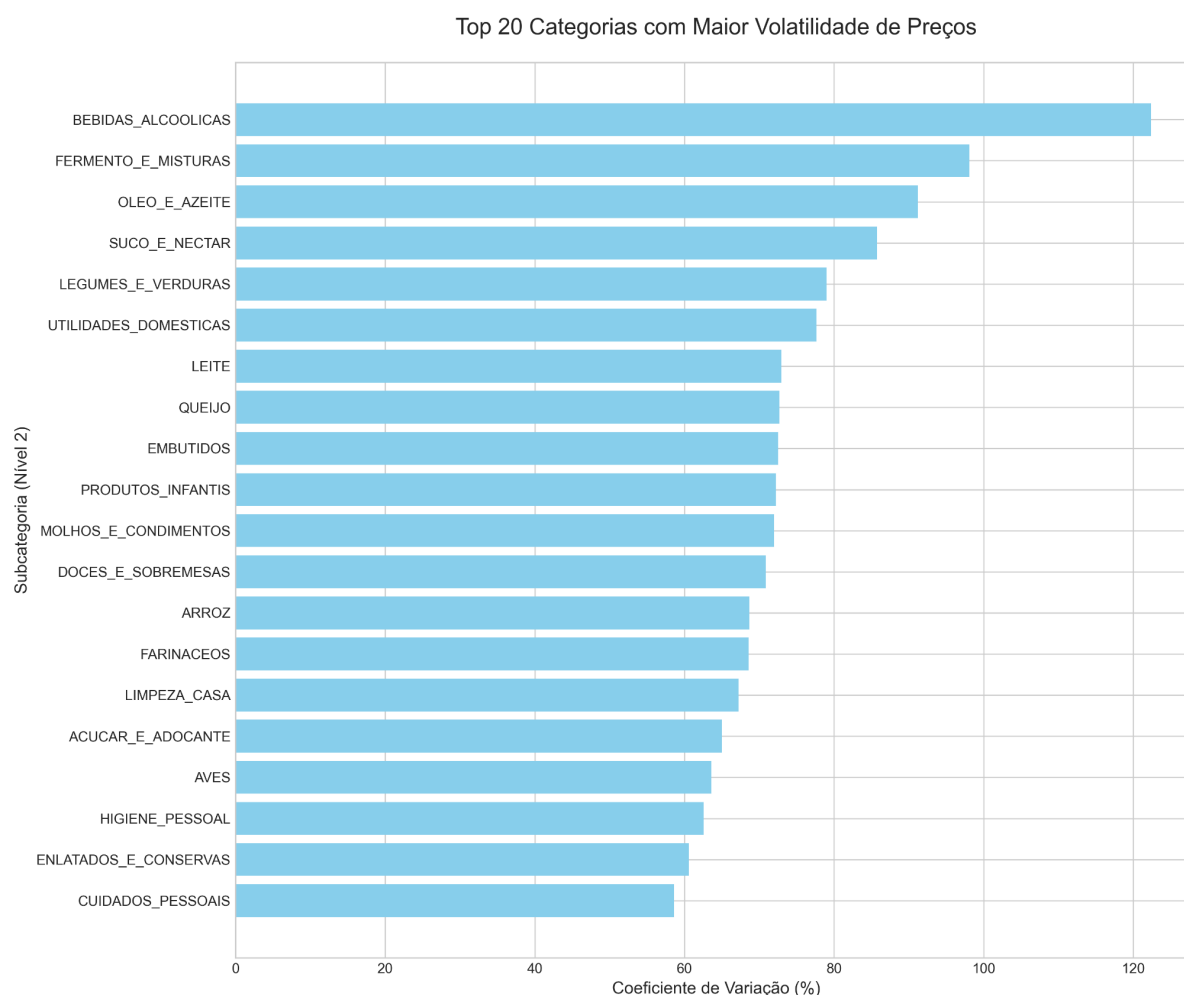
4.1.2 As Categorias Mais Voláteis

Para mensurar a instabilidade dos preços de forma comparável entre produtos com diferentes faixas de valor, utilizou-se o Coeficiente de Variação (CV). Essa métrica relativa — calculada pela razão entre o desvio padrão e a média — permite identificar quais subcategorias apresentam maior dispersão em torno de seu valor médio. Quanto maior o CV, maior a variabilidade dos preços praticados para aquela subcategoria, indicando que a pesquisa de preços pelo consumidor pode gerar maior economia.

A análise do CV para todas as subcategorias de produto (variável

categoria_n2) demonstrou que os produtos se agrupam em distintas faixas de variação de preço, conforme ilustrado no Gráfico 2.

GRÁFICO 2 – RANKING DAS 20 SUBCATEGORIAS COM MAIOR VOLATILIDADE DE PREÇOS



FONTE: ELABORADO PELO AUTOR (2025).

Observa-se que o topo do ranking é dominado por categorias como BEBIDAS_ALCOOLICAS, FERMENTO_E_MISTURAS e OLEO_E_AZEITE, todas com coeficientes superiores a 80%, evidenciando forte oscilação de preços entre estabelecimentos e ao longo do tempo. A presença de LEGUMES_E_VERDURAS e SUCO_E_NECTAR entre as posições mais elevadas corrobora a expectativa de que itens perecíveis, sazonais e sensíveis à oferta apresentam maior volatilidade.

Surpreende, contudo, a inclusão de categorias não perecíveis como

UTILIDADES_DOMESTICAS e PRODUTOS_INFANTIS, sugerindo que fatores como políticas promocionais, segmentação de público e variação de marcas podem também exercer forte influência sobre os preços.

A maioria das subcategorias do ranking apresenta CVs superiores a 60%, o que demonstra que a volatilidade é um fenômeno disseminado no varejo curitibano — não se restringindo a um único tipo de produto. Este resultado reforça a relevância da presente pesquisa, ao evidenciar os pontos de maior assimetria de informação e, portanto, maior potencial de economia para o consumidor bem informado.

A rotina de cálculo dos coeficientes de variação foi desenvolvida em Python, conforme detalhado no Apêndice 5 deste trabalho.

4.2. A DIMENSÃO GEOGRÁFICA: ONDE SE COMPRA MAIS CARO E MAIS BARATO?

A dimensão territorial exerce influência direta sobre os preços praticados no varejo, refletindo não apenas diferenças estruturais entre bairros, mas também a segmentação estratégica do mercado e o poder aquisitivo da população local. No contexto da cidade de Curitiba, marcada por uma urbanização heterogênea e pela coexistência de diferentes perfis socioeconômicos, torna-se essencial analisar a variação espacial dos preços como forma de revelar disparidades no acesso a bens essenciais.

Nesta seção, busca-se responder à pergunta: onde se paga mais caro — ou mais barato — pela mesma cesta de produtos? Para isso, foram utilizadas as cotações diárias da base tratada para calcular o custo da "Cesta de Categorias Essenciais" em cada bairro com representatividade estatística suficiente. A comparação entre bairros foi estruturada em dois níveis: inicialmente, com o ranking geral de custo da cesta por bairro (Seção 4.2.1), seguido de uma análise mais granular que identifica as vocações de preço por categoria (Seção 4.2.2), revelando especializações ou vantagens relativas em determinados grupos de consumo.

Essas análises permitem não apenas evidenciar desigualdades locais, mas também oferecer informações úteis para o planejamento de compras e políticas

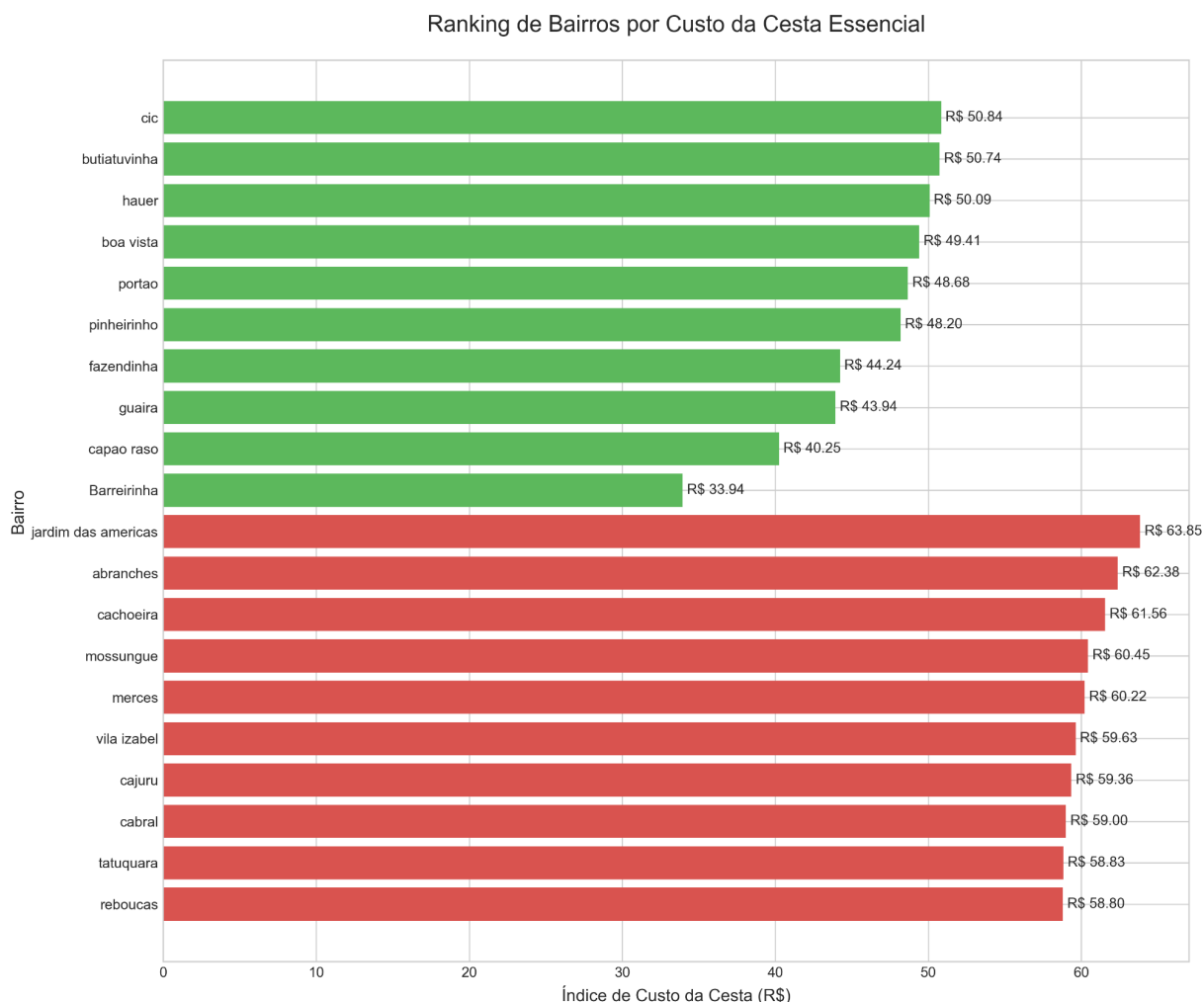
públicas voltadas à segurança alimentar.

4.2.1. Ranking de custo de vida por bairro

A comparação do custo da “Cesta de Categorias Essenciais” entre os bairros de Curitiba permite identificar diferenças expressivas na experiência de compra do consumidor, determinadas não apenas pela renda ou perfil socioeconômico, mas também pela disponibilidade, variedade e política de preços locais. Essa análise é particularmente relevante para entender a desigualdade espacial do acesso a bens essenciais, que impacta diretamente a qualidade de vida e o poder de compra dos habitantes da cidade.

O gráfico 3 apresenta o ranking dos 20 bairros com maior representatividade no dataset, com base na mediana dos preços registrados ao longo de 2024. Para fins de clareza visual, os bairros foram divididos entre os 10 com menor custo (em verde) e os 10 com maior custo (em vermelho), com base no índice composto da Cesta de Categorias Essenciais, conforme definido na Metodologia (seção 3.3.1).

GRÁFICO 3 – RANKING DE BAIRROS POR CUSTO DA CESTA ESSENCIAL



FONTE: ELABORADO PELO AUTOR (2025).

No grupo de menor custo, destacam-se bairros como Barreirinha (R\$ 33,94), Capão Raso (R\$ 40,25) e Guaíra (R\$ 43,94). Esses valores apontam para maior acessibilidade econômica da cesta de produtos analisada, o que pode representar uma vantagem real para consumidores locais que dependem do varejo físico. No entanto, é fundamental interpretar esses resultados com cautela metodológica. Em bairros como a Barreirinha, por exemplo, o menor índice pode estar parcialmente associado a uma oferta menos diversificada de produtos — seja pela menor presença de redes *premium*, seja pela escassez de itens de alto valor agregado — o que impacta diretamente a composição da mediana calculada. Em outras palavras, o custo mais baixo pode refletir tanto uma realidade de preços mais acessíveis quanto

uma restrição no sortimento disponível.

Por outro lado, entre os bairros com maior custo, figuram Jardim das Américas (R\$ 63,85), Abranches (R\$ 62,38) e Cachoeira (R\$ 61,56). A diferença entre o bairro mais barato e o mais caro ultrapassa R\$ 29,00 — uma variação de 85% — evidenciando a intensidade da disparidade espacial no custo de vida alimentar em Curitiba. Essa diferença é significativa o suficiente para indicar que, para um mesmo perfil de consumo, o local de residência pode representar um impacto financeiro relevante ao longo do mês.

Essa variação não se explica apenas por fatores de renda ou localização geográfica, mas também por elementos estruturais, como:

- **O perfil dos estabelecimentos predominantes:** bairros com concentração de hipermercados e supermercados de redes *premium* tendem a apresentar preços mais elevados em produtos não promocionais;
- **A densidade comercial e a concorrência local:** regiões com menor presença de redes concorrentes frequentemente têm menor pressão competitiva sobre os preços;
- **O mix de produtos disponíveis:** bairros que oferecem mais produtos de alto valor (como carnes nobres, fórmulas infantis, produtos importados) podem elevar a mediana da cesta, ainda que haja itens baratos também disponíveis.

Dessa forma, o ranking da cesta não deve ser interpretado de forma isolada como indicador de "eficiência de mercado", mas sim como um retrato composto da realidade econômica, logística e comercial de cada bairro.

Do ponto de vista prático, esses achados reforçam a importância de incorporar dimensões territoriais às políticas públicas de segurança alimentar e abastecimento urbano. Mapear os bairros mais vulneráveis — não apenas por renda, mas por custo efetivo de consumo — pode embasar iniciativas como feiras itinerantes, subsídios cruzados ou incentivos fiscais à instalação de novas unidades varejistas em regiões com preços elevados ou baixa oferta.

Por fim, o presente trabalho aprofunda essa leitura na seção seguinte (4.2.2), com uma análise detalhada das subcategorias de produtos por bairro, buscando

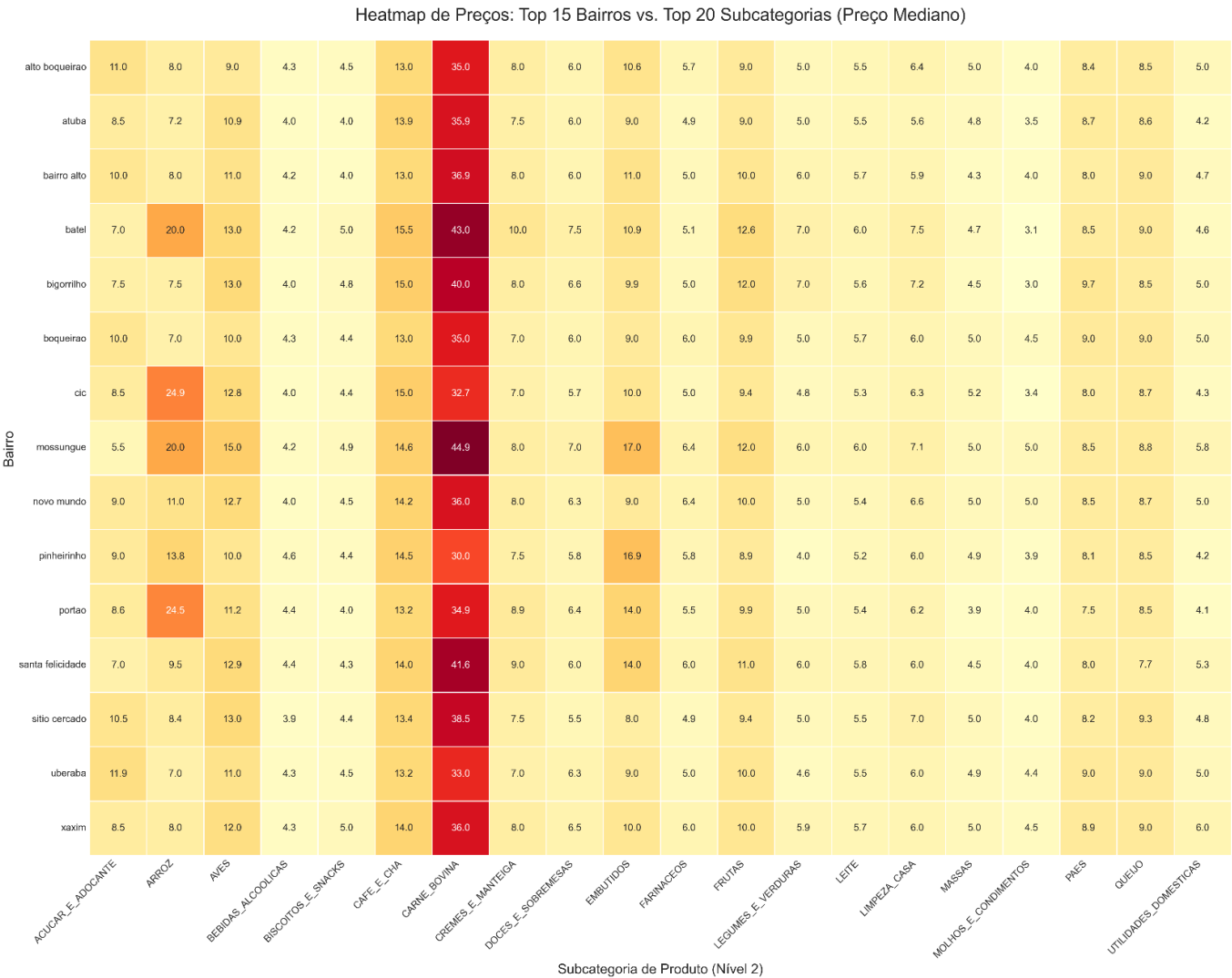
revelar não apenas onde se compra mais barato ou caro em termos agregados, mas quais bairros são mais vantajosos para cada tipo específico de item.

4.2.2. A "Vocação" dos Bairros

Para além do custo total da cesta essencial, uma análise mais detalhada revela padrões específicos de precificação por tipo de produto em cada bairro, configurando o que pode ser interpretado como 'vocação de consumo'. Essa abordagem permite identificar se há bairros que se destacam por oferecer preços sistematicamente mais baixos em determinadas categorias — como carnes, hortifrúti ou laticínios — ainda que o custo geral da cesta não seja o mais vantajoso.

O gráfico 4 apresenta um heatmap de preços medianos por subcategoria (nível 2) nos 15 bairros com maior volume de dados, representando a diversidade geográfica da cidade. As cores indicam os valores medianos observados em cada cruzamento entre bairro e subcategoria, permitindo comparar o comportamento de preços de forma visual e imediata.

GRÁFICO 4 – HEATMAP DE PREÇOS: TOP 15 BAIRROS VS. TOP 20 SUBCATEGORIAS (PREÇO MEDIANO)



FONTE: ELABORADO PELO AUTOR (2025).

A análise do heatmap revela evidências de especializações locais. Por exemplo:

- O bairro CIC, já identificado no ranking anterior como uma das regiões com menor custo da cesta essencial, apresenta valores particularmente baixos em categorias como CARNE_BOVINA (R\$32,70), MASSAS (R\$5,20) e

MOLHOS_E_CONDIMENTOS (R\$3,40), o que contribui diretamente para sua atratividade enquanto polo de compras econômicas em produtos básicos e proteínas animais.

- Em contraste, o bairro Mossunguê — um dos mais caros — exhibe preços elevados em subcategorias como FRUTAS (R\$12,00), CREMES_E_MANTEIGA (R\$8,00) e DOCES_E_SOBREMESAS (R\$7,00), o que sugere um perfil de sortimento mais sofisticado e um mercado voltado a consumidores com maior poder aquisitivo.
- Bairros como o Boqueirão e o Santa Felicidade demonstram certa estabilidade de preços, sem grandes picos ou vales nas subcategorias, o que pode indicar maior diversidade de estabelecimentos e concorrência equilibrada dentro do próprio bairro.

Essa segmentação revela que nem sempre o bairro com o menor custo total será a melhor escolha para todos os perfis de consumo. Por exemplo, um consumidor cuja principal preocupação seja a aquisição de laticínios pode encontrar melhores oportunidades no Pinheirinho (LEITE = R\$5,20; QUEIJO = R\$8,50) do que em bairros mais baratos no *ranking* geral. Isso reforça a ideia de que a vocação dos bairros deve ser considerada na decisão de compra, especialmente quando o deslocamento entre regiões for viável.

Além disso, há indícios de que os bairros com maior renda per capita tendem a apresentar preços mais elevados em categorias não essenciais ou mais sensíveis à diferenciação de qualidade, como PRODUTOS_INFANTIS, UTILIDADES_DOMESTICAS e CUIDADOS_PESSOAIS. Isso pode refletir uma menor elasticidade de preço entre os consumidores dessas regiões, bem como uma oferta mais diversificada e *premium*, elevando a mediana de preços mesmo entre produtos similares.

Por outro lado, bairros como Barreirinha, que ocupam o topo do ranking de menor custo, devem ser interpretados com cautela. Apesar do resultado aparentemente vantajoso, há a possibilidade de que esse valor seja influenciado por uma menor diversidade de sortimento, o que reduz a amostra de preços considerados para certas categorias. Isso não significa, necessariamente, que o

bairro seja sempre o mais barato, mas sim que os produtos disponíveis tendem a apresentar valores inferiores aos praticados em outras regiões — o que pode ser tanto um reflexo de eficiência de preço quanto de oferta limitada ou ausência de categorias mais caras no *mix* local.

Do ponto de vista metodológico, a análise foi realizada com base nas medianas por subcategoria e bairro, o que reduz o impacto de valores extremos e torna os comparativos mais robustos. Ainda assim, recomenda-se cautela na interpretação dos dados, principalmente nos casos em que a quantidade de registros por subcategoria foi reduzida — uma situação controlada pela filtragem aplicada durante a fase de enriquecimento e validação do dataset (Capítulo 3).

Finalmente, os achados deste tópico não apenas corroboram a existência de desigualdade espacial nos preços ao consumidor, como também apontam caminhos para políticas públicas e decisões de planejamento urbano. Iniciativas como feiras móveis, subsídios logísticos ou incentivos fiscais à abertura de mercados em regiões carentes podem ser mais bem desenhadas a partir de evidências como as aqui apresentadas, com foco em ampliar o acesso a produtos básicos a preços justos.

4.3. A DIMENSÃO COMPETITIVA: AS ESTRATÉGIAS DAS REDES DE VAREJO

Além das diferenças geográficas já discutidas, a compreensão das dinâmicas de preços no varejo alimentar requer uma análise cuidadosa da atuação e das estratégias das redes varejistas. Redes diferentes adotam estratégias distintas de precificação, *mix* de produtos e segmentação de mercado, influenciando diretamente os preços finais observados nas gôndolas. Dessa forma, torna-se essencial entender como cada rede posiciona-se em relação às concorrentes, buscando captar diferentes perfis de consumidores e diferentes níveis de poder aquisitivo.

Este tópico investiga a dimensão competitiva do mercado varejista de Curitiba a partir de duas perspectivas principais. A primeira é a comparação direta entre as redes, com base no custo geral da "Cesta de Categorias Essenciais", para determinar quais empresas oferecem sistematicamente melhores condições para compras completas e regulares. A segunda perspectiva aprofunda essa análise, identificando especializações estratégicas em categorias específicas de consumo,

revelando o que se denomina "matriz de posicionamento estratégico".

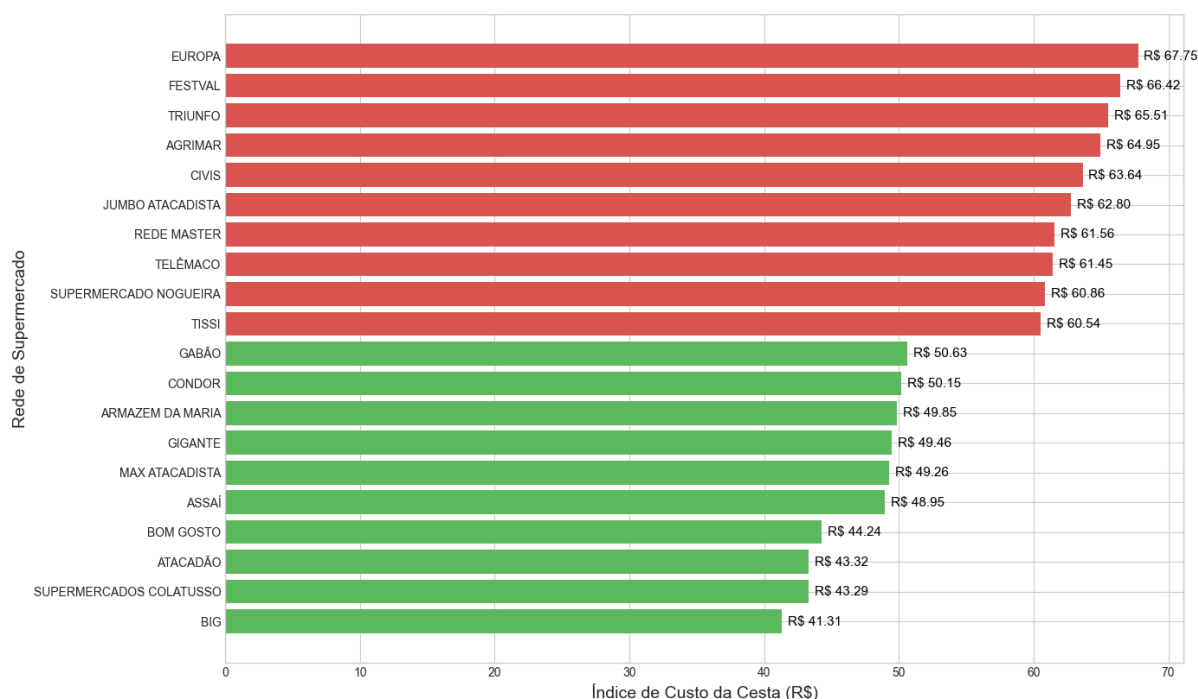
Essas informações permitem entender quais redes se destacam por preços competitivos em segmentos específicos, e quais utilizam estratégias de preços *premium*, com sortimento ampliado e produtos de maior valor agregado. A análise detalhada auxilia não apenas o consumidor em suas decisões cotidianas, mas também fornece subsídios estratégicos para gestores e formuladores de políticas públicas, especialmente no que diz respeito à promoção da concorrência e do acesso a produtos essenciais.

A seguir, são apresentados o *ranking* comparativo das redes de supermercado e a matriz estratégica de preços por categoria.

4.3.1. *Ranking* de custo por rede de supermercado

A dimensão competitiva do varejo alimentar é fortemente influenciada pela estratégia adotada pelas redes de supermercado, impactando diretamente o custo final para o consumidor. Com o objetivo de investigar tais estratégias, construiu-se o *ranking* duplo apresentado no gráfico 5, identificando as 10 redes com maior e menor custo para aquisição da "Cesta de Categorias Essenciais" definida por este estudo, representada pela soma das medianas de preços das categorias analisadas.

GRÁFICO 5 – RANKING DUPLO DAS REDES DE SUPERMERCADO POR CUSTO DA CESTA ESSENCIAL (TOP 10 MAIS BARATOS E MAIS CAROS)



FONTE: ELABORADO PELO AUTOR (2025).

A análise dos preços mais elevados destaca um grupo de estabelecimentos liderado pela rede Europa (R\$ 67,75), seguida por Festval (R\$ 66,42) e Triunfo (R\$ 65,51). Neste grupo, é notório que redes como Festval e Triunfo possuem um claro posicionamento de mercado voltado a segmentos de maior poder aquisitivo, caracterizando-se pela oferta de produtos diferenciados, serviços adicionais e uma experiência de compra *premium*. Esse modelo de negócio, que prioriza o valor agregado, reflete-se em uma precificação mais elevada da cesta e em custos superiores para o consumidor.

Por outro lado, as redes com o menor custo apresentam um perfil operacional distinto, marcado principalmente por estratégias de alta eficiência e volume elevado, frequentemente associados ao modelo de atacarejo ou hipermercados. Nesse contexto, destaca-se a rede BIG (R\$41,31), que figura como a opção mais econômica do ranking. Também merecem menção as redes Supermercados

Colatusso (R\$43,29) e Atacadão (R\$43,32), que se destacam por oferecer preços notadamente mais baixos, baseados na venda por volume e na redução de custos operacionais, tais como infraestrutura simplificada, sortimento reduzido e logística otimizada.

A disparidade de preços entre as redes mais caras e mais baratas ultrapassa os R\$ 26,00, uma variação considerável que ilustra claramente a importância da pesquisa de preços pelos consumidores para uma melhor gestão financeira doméstica. Adicionalmente, o consumidor pode extrair desses resultados uma compreensão mais detalhada sobre quais redes oferecem melhores vantagens econômicas dependendo de suas preferências de compra e da composição específica da cesta de consumo.

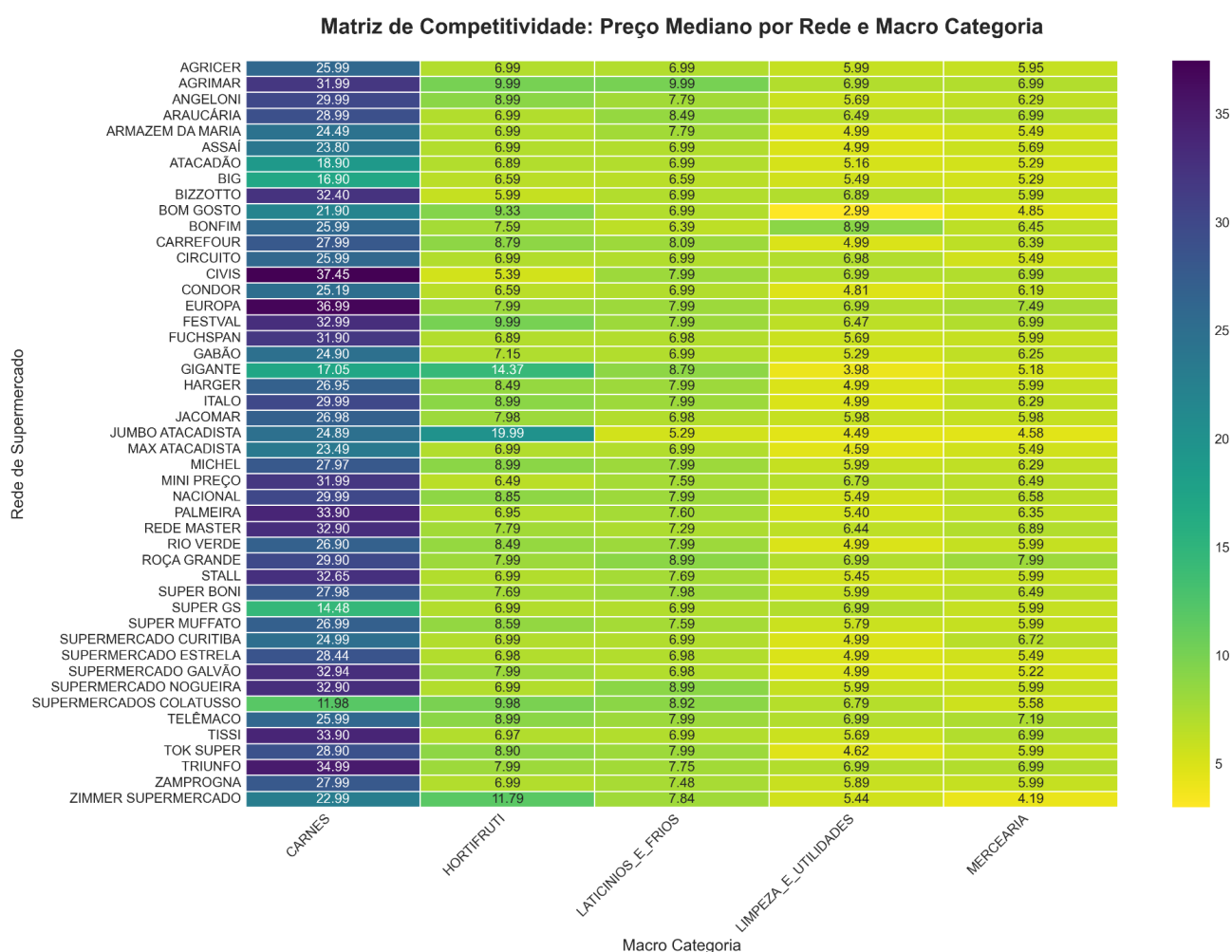
É importante salientar que esse ranking não é um indicativo isolado de desempenho ou eficiência absoluta de uma rede sobre outra. Redes com maior custo podem ser preferidas por consumidores específicos que valorizam a experiência de compra, exclusividade ou conveniência acima da simples economia financeira. Contudo, para a parcela da população cuja sensibilidade ao preço é um fator decisivo, a identificação das redes com menor custo oferece uma orientação clara e prática, contribuindo diretamente para decisões mais conscientes e otimizadas.

Assim, os resultados desta seção trazem não apenas uma compreensão mais ampla sobre o cenário competitivo do varejo alimentar curitibano, mas também indicam quais são os principais *players* de cada extremo do espectro de preços. Essa análise estratégica fornece fundamentos importantes para as seções seguintes, especialmente no detalhamento da matriz de posicionamento estratégico das redes, que revela quais categorias de produtos são efetivamente responsáveis pelos diferentes posicionamentos de preço das redes avaliadas.

4.3.2. Matriz de posicionamento estratégico

Para aprofundar a compreensão das estratégias competitivas adotadas pelas redes de supermercados em Curitiba, foi elaborada uma matriz de competitividade, apresentada no gráfico 5. Esse gráfico, no formato de heatmap, detalha visualmente o preço mediano praticado por cada rede nas cinco principais macrocategorias analisadas: Carnes, Hortifruti, Laticínios e Frios, Limpeza e Utilidades e Mercadoria. Essa representação permite identificar padrões estratégicos distintos, revelando com clareza quais redes se posicionam de maneira competitiva em categorias específicas e quais adotam uma estratégia geral com preços mais elevados.

GRÁFICO 5 – MATRIZ DE COMPETITIVIDADE: PREÇO MEDIANO POR REDE E MACRO CATEGORIA



FONTE: ELABORADO PELO AUTOR (2025).

Uma primeira leitura da matriz de competitividade revela uma segmentação clara entre as estratégias adotadas pelas diferentes redes de varejo. Enquanto algumas redes adotam políticas de precificação mais agressivas em múltiplas categorias, outras concentram seus esforços em nichos específicos, e há ainda aquelas que mantêm preços sistematicamente mais elevados, sinalizando uma possível estratégia de diferenciação com foco em qualidade percebida ou experiência de compra superior.

Entre os grupos com posicionamento mais competitivo em termos de preço, destaca-se a rede BIG, que apresenta o menor preço mediano na categoria

Mercearia (R\$ 5,29), além de valores bastante atrativos em Limpeza e Utilidades e Carnes. Esse padrão de precificação sugere a adoção de uma política de *Everyday Low Price* (EDLP), uma estratégia recorrente entre grandes redes que buscam fidelizar consumidores sensíveis a preço, oferecendo descontos consistentes e previsíveis (HOCH et al., 1994).

Outro destaque importante são os chamados atacarejos, como Atacadão, Assaí e Max Atacadista, que mantêm preços medianos visivelmente mais baixos nas categorias de maior peso no consumo doméstico, especialmente em Carnes e Mercearia. A competitividade dessas redes confirma sua vocação estrutural para operar com margens enxutas, alto giro de estoque e forte apelo ao consumo em volume (LAL; RAO, 1997). Sua relevância no cenário curitibano reforça o papel estratégico desses formatos, sobretudo em períodos de inflação persistente e perda de poder de compra da população.

No outro extremo, redes como Europa, Festival e Triunfo aparecem de forma recorrente entre as mais onerosas em praticamente todas as macro categorias analisadas. Por exemplo, a rede Europa apresenta preços expressivamente acima da média em Carnes (R\$36,99), Hortifruti (R\$7,99) e Mercearia (R\$7,49). Tais valores indicam um posicionamento voltado à diferenciação por qualidade, variedade de sortimento e experiência de compra superior, voltado a públicos com maior poder aquisitivo.

Entre os modelos intermediários, destacam-se redes como Condor e Angeloni, que adotam uma estratégia híbrida: combinam preços competitivos em categorias estratégicas com níveis medianos ou moderadamente elevados em outras. O Condor, por exemplo, mantém preços bastante agressivos em Limpeza e Utilidades (R\$4,81) e Carnes (R\$25,19), ao passo que apresenta comportamento mais neutro em categorias como Hortifruti ou Laticínios. Esse modelo permite atrair diferentes perfis de consumidores ao mesmo tempo em que protege margens operacionais por meio de equilíbrio no mix de preços.

Vale ainda mencionar casos como o da rede Gigante, que se destaca pela precificação extremamente competitiva na categoria Mercearia (R\$3,98), o que pode refletir uma estratégia de atração para itens de alto giro, funcionando como porta de entrada para outras compras e reforçando o apelo popular do estabelecimento. Ainda assim, é necessário cautela na leitura de redes de menor escala ou com

menor presença amostral na base de dados, pois resultados atípicos podem decorrer de limitações no volume ou diversidade dos registros capturados.

De modo geral, o heatmap de competitividade permite compreender com maior clareza como as redes disputam o mercado curitibano não apenas por preço global, mas por diferenciação estratégica em categorias específicas. Essa análise oferece subsídios relevantes tanto para consumidores — ao otimizarem suas decisões de compra — quanto para gestores e formuladores de política pública interessados em ampliar a transparência e a eficiência do mercado varejista alimentar.

Como extensão da matriz de competitividade por macrocategorias, foi desenvolvida uma análise complementar detalhada, cruzando as 15 redes com maior presença no conjunto de dados com as 20 subcategorias mais representativas do consumo cotidiano. Essa matriz ampliada — apresentada no Apêndice 6 — permite identificar com maior precisão os focos estratégicos de cada rede, revelando, por exemplo, quais varejistas praticam os menores preços em subcategorias como óleo e azeite, arroz, frutas ou produtos de limpeza. Essa visualização mais granular contribui para um entendimento aprofundado da diferenciação entre as redes e complementa as evidências discutidas neste tópico.

Na próxima seção, a análise volta-se à dimensão temporal da precificação, abordando a evolução dos preços ao longo de 2024 e destacando as tendências e oscilações sazonais que impactam o consumidor.

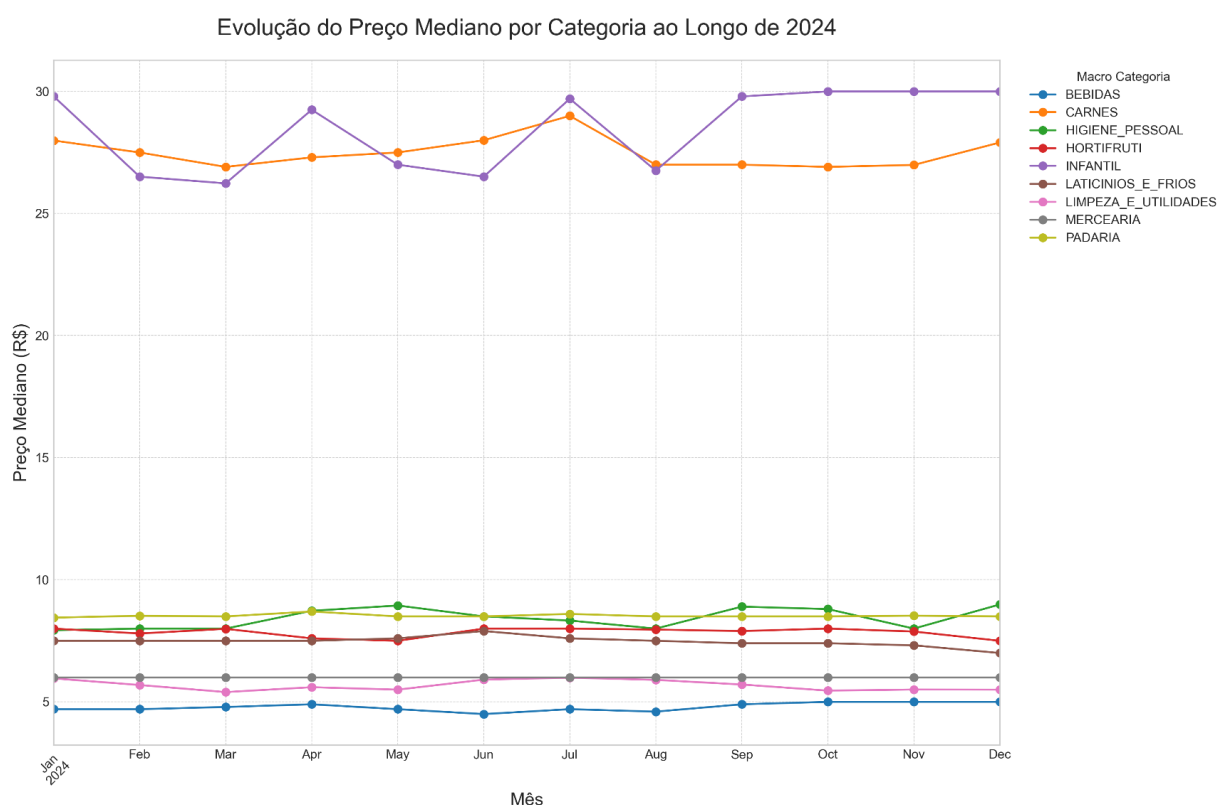
4.4. A DIMENSÃO TEMPORAL: A EVOLUÇÃO DOS PREÇOS EM 2024

4.4.1. A "inflação" por categoria

Embora o termo “inflação” costume ser associado a índices oficiais de variação de preços, como o IPCA (Índice Nacional de Preços ao Consumidor Amplo), esta seção adota uma perspectiva empírica baseada nos dados de mercado coletados ao longo de 2024. O objetivo é observar a evolução do preço mediano das macro categorias ao longo dos meses, buscando identificar tendências, picos sazonais e estabilidade relativa nos diferentes grupos de consumo.

O gráfico 6 apresenta a variação mensal dos preços medianos para cada macro categoria do conjunto de dados, permitindo visualizar de forma direta o comportamento agregado das principais famílias de produtos alimentícios e de uso doméstico.

GRÁFICO 6 – EVOLUÇÃO DO PREÇO MEDIANO POR CATEGORIA AO LONGO DE 2024



FONTE: ELABORADO PELO AUTOR (2025).

A leitura do gráfico revela um padrão de estabilidade acentuada na maioria das categorias, com pequenas oscilações em torno de uma média estável ao longo dos 12 meses. Categorias como Limpeza e Utilidades, Mercearia, Bebidas, Higiene Pessoal e Padaria apresentaram variações inferiores a R\$ 0,50 no preço mediano durante todo o período analisado, sugerindo um controle relativamente rígido de preços por parte do setor varejista, ou baixa sensibilidade a fatores sazonais de oferta e demanda.

Em contraste, duas macro categorias se destacam com variações mais pronunciadas: Carnes e Infantil. No caso das carnes, observa-se um aumento mais expressivo entre os meses de junho e julho, possivelmente relacionado a fatores sazonais, como demanda aumentada durante o inverno, feriados regionais ou instabilidades na cadeia de fornecimento. Já a categoria Infantil demonstra um comportamento de alta progressiva no segundo semestre, alcançando em dezembro o valor mais elevado do ano, o que pode estar associado tanto à menor elasticidade da demanda quanto ao lançamento de novos produtos ou ajustes estratégicos de preço no fim de ano.

Essas oscilações, embora modestas em termos absolutos, refletem movimentos relevantes dentro do contexto da precificação estratégica no varejo. Como apontam Levy et al. (1997), muitos varejistas evitam repassar pequenas variações de custo ao consumidor com frequência, a fim de manter a percepção de estabilidade e previsibilidade. Isso ajuda a explicar a suavidade geral observada na série de preços medianos, mesmo diante de eventuais pressões inflacionárias nos bastidores da cadeia de suprimentos.

O uso da mediana como métrica central para esse tipo de análise mostra-se, mais uma vez, apropriado. Ela atenua o impacto de valores extremos e possibilita uma leitura mais robusta da tendência central, sobretudo em categorias com forte dispersão de preços.

Para um exame mais detalhado, foi produzido um painel com a evolução mensal do preço mediano das principais subcategorias — como leite, arroz, carnes, embutidos, frutas e hortaliças. Esse painel completo, disponível no Apêndice 7, aprofunda a leitura das dinâmicas sazonais internas de cada grupo, evidenciando comportamentos específicos muitas vezes ocultos na análise agregada por macro categorias.

Em resumo, a análise temporal dos preços indica que, ao longo de 2024, o comportamento geral dos preços foi marcado por estabilidade relativa, com variações concentradas em poucas categorias específicas. Embora o termo “inflação” tenha sido utilizado de forma interpretativa, os dados sugerem que, no nível varejista, o consumidor curitibano não enfrentou aumentos abruptos de preços na maioria dos segmentos acompanhados. A próxima seção sintetiza os principais achados do trabalho e propõe uma aplicação prática dos resultados para orientar

escolhas mais eficientes de consumo na cidade.

4.5. SÍNTESE DOS RESULTADOS: O GUIA DO CONSUMIDOR CURITIBANO

Depois de mapeadas as variações espaciais, setoriais e temporais nos preços praticados pelo varejo curitibano, esta seção final propõe uma aplicação prática dos achados em benefício direto do consumidor. Com base na combinação dos dados por bairro e por rede de supermercado, foi possível identificar, para cada região analisada, qual rede oferece o menor custo da “Cesta de Categorias Essenciais”, compondo assim um guia de referência local que sintetiza as evidências apresentadas ao longo do estudo.

Essa informação, sistematizada e agregada, visa oferecer ao consumidor curitibano subsídios objetivos para otimizar suas decisões de compra, considerando o fator local e a escolha da rede de supermercado mais econômica em sua vizinhança. A consolidação foi realizada com base na mediana dos preços de cada macro categoria nos cruzamentos entre bairros e redes, garantindo robustez estatística mesmo em casos com leve assimetria ou dispersão.

4.5.1. O melhor custo-benefício por bairro

A consolidação das análises possibilitou identificar, para cada bairro com representatividade no conjunto de dados, qual rede de supermercado ofereceu o menor custo da “Cesta de Categorias Essenciais” ao longo de 2024. Esse cruzamento entre bairros, redes de varejo e respectivos valores medianos é apresentado na tabela 2.

TABELA 2 – MELHOR CUSTO DA “CESTA ESSENCIAL” POR BAIRRO E REDE DE SUPERMERCADO

Bairro	Rede mais barata	Custo cesta em R\$
abranches	PALMEIRA	60.47
agua verde	ANGELONI	57.55
alto boqueirao	ATACADÃO	40.2
atuba	ASSAÍ	51.45
bacacheri	NACIONAL	58.1
bairro alto	ATACADÃO	52.06
Barreirinha	TRIUNFO	33.94
batel	NACIONAL	57.85
bigorrilho	SUPER MUFFATO	56.85
boa vista	CONDOR	49.41
boqueirao	AGRICER	50.64
butiatuvinha	CONDOR	50.74
cabral	CARREFOUR	59
cachoeira	REDE MASTER	61.56
cajuru	MINI PREÇO	59.36
campo comprido	CONDOR	50.73
campo de santana	JACOMAR	52.9
capao da imbuia	TOK SUPER	57.37
capao raso	ATACADÃO	37.25
centro	NACIONAL	57.53
cic	ASSAÍ	41.98
fazendinha	BOM GOSTO	44.24
guaira	ATACADÃO	43.94
hauer	MAX ATACADISTA	47.29
jardim das americas	NACIONAL	63.85
merces	FESTVAL	60.22
mossungue	CARREFOUR	52.43
novo mundo	SUPER GS	38.94
parolin	ASSAÍ	51.25
pilarzinho	CONDOR	51.75
pinheirinho	ASSAÍ	41.26
portao	ATACADÃO	42.55
reboucas	CARREFOUR	58.8
santa felicidade	ATACADÃO	44.06
sitio cercado	SUPERMERCADO GALVÃO	23.16
taruma	HARGER	56.36
tatuquara	TISSI	55.42
uberaba	CONDOR	49.25
vila izabel	FESTVAL	59.63
xaxim	BIG	41.31

FONTE: ELABORADO PELO AUTOR (2025).

Os resultados revelam um cenário descentralizado, com diferentes redes ocupando a posição de menor preço em distintas regiões da cidade. Redes de atacarejo, como Atacadão e Assaí, destacam-se por apresentarem os menores custos medianos em bairros como Capão Raso, Portão, CIC e Pinheirinho, o que reflete seu posicionamento estratégico focado em volume e preços competitivos.

Por outro lado, redes regionais também demonstram protagonismo em determinadas localidades. É o caso do Supermercado Palmeira no bairro Abranches, do Jacomar em Campo de Santana e do Super GS em Novo Mundo, que lideram localmente em termos de menor custo da cesta, evidenciando que a competitividade de preços não se restringe exclusivamente às grandes redes nacionais.

O menor valor observado em toda a amostra foi registrado no Sítio Cercado, onde a rede Supermercado Galvão ofertou a cesta por R\$23,16. Embora expressivo, esse resultado deve ser interpretado com cautela, uma vez que fatores como menor sortimento, mix de produtos limitado ou promoções pontuais podem influenciar diretamente na composição final da cesta naquele local.

Na extremidade superior do ranking, bairros como Jardim das Américas, Cachoeira e Mercês apresentaram custos medianos superiores a R\$60,00, sinalizando uma variação expressiva no custo de vida básico entre as diferentes regiões da cidade. Esses dados reforçam a relevância de análises geográficas aplicadas ao consumo e sua utilidade na formulação de políticas públicas de segurança alimentar e combate à desigualdade.

Assim, os achados desta seção oferecem ao consumidor uma referência prática, baseada em dados reais, para otimizar decisões de compra com base no local de residência e no perfil de redes disponíveis em sua vizinhança.

5 CONSIDERAÇÕES FINAIS

O presente trabalho teve como propósito analisar, com base empírica e metodológica consistente, a estrutura de preços dos produtos essenciais ofertados em supermercados da cidade de Curitiba, a partir do aproveitamento integral da base de dados do programa municipal “Clique Economia”. Valendo-se de ferramentas de ciência de dados e de uma análise exploratória e descritiva, a pesquisa avançou na compreensão dos padrões de precificação no varejo alimentar urbano. Os resultados confirmaram a acentuada dispersão de preços, revelando expressivas desigualdades geográficas e competitivas que impactam diretamente o cotidiano do consumidor curitibano.

Ao longo da investigação, foi possível estruturar um *pipeline* de tratamento e enriquecimento de dados que, partindo de registros brutos, permitiu a geração de variáveis analíticas com alto grau de relevância prática. O esforço metodológico na normalização de endereços, padronização de redes de varejo e categorização hierárquica de produtos constituiu uma das principais contribuições técnicas da pesquisa, criando uma base confiável e analiticamente robusta para os estudos subsequentes.

A análise dos resultados revelou um cenário de dispersão de preços generalizada, com destaque para subcategorias como óleo e azeite, fermento e misturas, legumes e verduras, e produtos infantis, que apresentaram altos coeficientes de variação. Tais achados reforçam a hipótese teórica de que mercados com assimetria de informação e baixa transparência de preços produzem oportunidades de arbitragem relevantes para o consumidor informado (STIGLER, 1961).

No plano espacial, identificou-se uma notável discrepância no custo da cesta de categorias essenciais entre os bairros de Curitiba, com variações que superam R\$30, mesmo dentro de uma mesma cidade. Essa diferença evidencia desigualdades no acesso a produtos básicos, e aponta para a necessidade de políticas públicas mais direcionadas à garantia da segurança alimentar em regiões com maior vulnerabilidade econômica.

A dimensão competitiva, por sua vez, evidenciou o comportamento

estratégico das redes de supermercado. Algumas adotam posturas agressivas com preços baixos consistentes, como o Atacadão, Assaí e BIG; outras operam em nichos *premium*, como Festival e Triunfo; e algumas redes de médio porte demonstraram estratégias seletivas que refletem tanto contexto local quanto posicionamento de mercado. A análise cruzada entre categorias e empresas, realizada por meio da matriz de posicionamento, proporcionou uma leitura refinada dessas estratégias, com implicações diretas para o entendimento das dinâmicas de precificação na capital paranaense.

No recorte temporal, os preços revelaram estabilidade relativa ao longo de 2024, com exceção de variações sazonais pontuais em categorias como Carnes e Infantil. Essa estabilidade, embora aparente, reflete estratégias deliberadas do setor varejista para manter a previsibilidade e evitar percepções de inflação por parte do consumidor — uma conduta bem documentada na literatura de precificação (LEVY et al., 1997).

Por fim, a construção da "Cesta de Categorias Essenciais" — uma métrica composta e orientada por dados — e sua aplicação em comparações entre bairros e redes, permitiu chegar a um diagnóstico prático do custo de vida alimentar em Curitiba. A consolidação desses dados em um guia de custo-benefício por bairro representa uma contribuição de utilidade direta para o cidadão, auxiliando-o em decisões de consumo mais informadas e racionais.

5.1 ALCANCE DOS OBJETIVOS

Os objetivos delineados no início desta pesquisa foram plenamente atingidos, conforme evidenciado pelos resultados obtidos ao longo da investigação. O objetivo geral — analisar exploratoriamente a dinâmica de preços no varejo alimentício de Curitiba — foi alcançado com êxito por meio de uma abordagem metodológica estruturada e de análises multifacetadas, que permitiram revelar padrões relevantes sob as dimensões geográfica, competitiva e temporal.

O primeiro objetivo específico, de mapear a variação de custos de uma cesta de consumo representativa entre os diferentes bairros da cidade, foi cumprido com a construção de rankings e visualizações que evidenciaram desigualdades territoriais

no acesso a preços mais baixos. A análise revelou diferenças expressivas no custo da cesta entre regiões, apontando para possíveis assimetrias estruturais no varejo urbano curitibano.

O segundo objetivo específico, que previa a comparação entre as principais redes de supermercado em termos de posicionamento de preços, foi atendido com o desenvolvimento de matrizes de posicionamento estratégico e rankings por categoria. A investigação permitiu distinguir estratégias comerciais distintas entre atacarejos e supermercados tradicionais, e identificou padrões de precificação por rede e tipo de produto, embasando-se em conceitos clássicos como EDLP e Hi-Lo.

O terceiro objetivo específico — identificar as categorias de produtos com maior e menor volatilidade ao longo do período — foi atingido com a aplicação do coeficiente de variação às subcategorias do dataset. O resultado forneceu uma hierarquia clara da instabilidade de preços, sinalizando quais produtos exigem maior atenção por parte dos consumidores em função da oscilação nos valores registrados.

Por fim, o quarto objetivo específico, voltado à elaboração de uma síntese aplicável dos achados em formato de guia prático para o consumidor, materializou-se na construção de uma tabela cruzada que aponta a rede de supermercado com menor custo da cesta essencial em cada bairro. Essa entrega consolida o conhecimento gerado ao longo do estudo em um recurso acessível e funcional, contribuindo com a autonomia informacional do consumidor curitibano.

Dessa forma, os resultados da pesquisa não apenas respondem às questões que nortearam este estudo, como também ampliam o entendimento sobre os determinantes da dispersão de preços e sua relação com o território urbano, a estrutura competitiva do varejo e o comportamento do consumidor.

5.2 TRABALHOS FUTUROS

A pesquisa abre diversas possibilidades de aprofundamento e continuidade. Em primeiro lugar, recomenda-se a expansão da base de dados para múltiplos anos, o que permitiria uma análise longitudinal mais completa dos efeitos inflacionários e sazonais no varejo local. Outra possibilidade promissora é a incorporação de

variáveis socioeconômicas — como renda média por bairro, densidade populacional ou presença de programas sociais — de modo a avaliar o impacto da estrutura de preços sobre diferentes camadas da população.

Do ponto de vista técnico, há potencial para a adoção de métodos preditivos, como séries temporais ou redes neurais, para estimar variações futuras de preços e antecipar tendências de mercado. Além disso, recomenda-se a criação de um painel interativo de consulta pública — em formato de dashboard — que possa ser integrado ao site da Prefeitura ou ao aplicativo *Curitiba App*, promovendo a transparência e o empoderamento do consumidor.

Por fim, a metodologia aqui aplicada pode ser adaptada a outros municípios, sobretudo aqueles que já possuem iniciativas de divulgação de preços. Com isso, seria possível construir um sistema comparativo nacional de monitoramento de preços no varejo, promovendo maior equidade e eficiência na política pública de segurança alimentar.

REFERÊNCIAS

AKERLOF, George A. The Market for "Lemons": Quality Uncertainty and the Market Mechanism. **Quarterly Journal of Economics**, v. 84, n. 3, p. 488-500, ago. 1970.

BEHRENS, John T. Principles and Procedures of Exploratory Data Analysis. **Psychological Methods**, Washington, DC, v. 2, n. 2, p. 131-160, 1997.

BRASIL. Decreto n. 8.777, de 11 de maio de 2016. Institui a Política de Dados Abertos no âmbito do Poder Executivo federal. **Diário Oficial da União**, Brasília, DF, 11 maio 2016. Disponível em:

https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/decreto/d8777.htm.

Acesso em: 2 dez. 2024.

BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde e Ambiente. **Vigitel Brasil 2023**: vigilância de fatores de risco e proteção para doenças crônicas por inquérito telefônico. Brasília, DF: Ministério da Saúde, 2023. Disponível em: https://bvsms.saude.gov.br/bvs/publicacoes/vigitel_brasil_2023.pdf. Acesso em: 12 jul. 2025.

CONFEDERAÇÃO NACIONAL DE DIRIGENTES LOJISTAS (CNDL). Preço é o fator que mais influencia a decisão de compra do consumidor brasileiro, aponta pesquisa exclusiva. **CNDL**, 2023. Disponível em: <https://cndl.org.br/varejosa/preco-e-o-fator-que-mais-influencia-a-decisao-de-compra-do-consumidor-brasileiro-aponta-pesquisa-exclusiva/>. Acesso em: 2 dez. 2024.

COUTINHO, Eduardo D.; FREITAS, Angilberto S. Valor público via tecnologias desenvolvidas com dados governamentais abertos: o caso Operação Serenata de Amor. **Revista de Administração Mackenzie**, São Paulo, v. 22, n. 6, p. 1-27, 2021. Edição especial. Disponível em: <https://doi.org/10.1590/1678-6971/eRAMD210079>. Acesso em: 10 jan. 2025.

CUNHA, I. B. A.; BARACHO, R. M. A. Dados abertos e suas aplicações em cidades inteligentes. **Múltiplos Olhares em Ciência da Informação**, Belo Horizonte, v. 9, n. 2, p. 1-18, 2019. DOI: 10.35699/2237-9777.2019.11776. Disponível em:

<https://periodicos.ufmg.br/index.php/moci/article/view/19143>. Acesso em: 12 mai. 2025.

FÁVERO, L. P.; BELFIORE, P. **Manual de análise de dados: estatística e modelagem multivariada com Excel, SPSS e Stata**. Rio de Janeiro: Elsevier, 2017.

FOLHA DE S. PAULO. Maioria das capitais tem ao menos 1 em 4 moradores vivendo em desertos alimentares. **Folha de S. Paulo**, 2025. Disponível em: <https://www1.folha.uol.com.br/folha-social-mais/2025/05/maioria-das-capitais-tem-ao-menos-1-em-4-moradores-vivendo-em-desertos-alimentares.shtml>. Acesso em: 2 jun. 2025.

FURTADO, B.; TAMASO, R. Atacarejo, redes regionais e conveniência: as tendências para o varejo de alimentos no Brasil. **Brazil Journal**, 2024. Disponível em: <https://braziljournal.com/on-business/atacarejo-redes-regionais-e-conveniencia-as-tendencias-para-o-varejo-de-alimentos-no-brasil/>. Acesso em: 3 fev. 2025.

GIL, Antônio Carlos. **Métodos e técnicas de pesquisa social**. 6. ed. São Paulo: Atlas, 2008.

HARVEY, David. **Cidades Rebeldes: do direito à cidade à revolução urbana**. Trad. Jeferson Camargo. São Paulo: Martins Fontes, 2014.

HOCH, S. J.; DRÈZE, X.; PURK, M. E. EDLP, Hi-Lo, and margin arithmetic. **Journal of Marketing**, v. 58, n. 4, p. 16-27, 1994.

INFOPRICE. O crescimento do atacarejo e o que aprender com o modelo. **InfoPrice**, 16 fev. 2024. Disponível em: <https://infoprice.co/blog/o-crescimento-do-atacarejo-e-o-que-aprender-com-o-modelo/>. Acesso em: 20 mar. 2025.

INSTITUTO BRASILEIRO DE ECONOMIA; HORUS INTELIGÊNCIA DE MERCADO. **Cesta de Consumo: Análise de preços e variação por zonas de São Paulo**. Rio de Janeiro: FGV IBRE, 2023. Disponível em: <https://portal.fgv.br/noticias/cesta-consumo-fgv-ibre-horus-registra-variacao-36-entre->

[zonas-sao-paulo](#). Acesso em: 11 fev. 2025.

KIMBALL, R.; ROSS, M. **The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling**. 3. ed. New York: Wiley, 2013.

LAL, R.; RAO, R. S. Supermarket competition: the case of every day low pricing. **Marketing Science**, v. 16, n. 1, p. 60-80, 1997.

LEVY, D.; BERGEN, M.; DUTTA, S.; VENABLE, R. The magnitude of menu costs: direct evidence from large U.S. supermarket chains. **Quarterly Journal of Economics**, v. 112, n. 3, p. 791-825, 1997.

LEVY, M.; WEITZ, B. A. **Retailing Management**. 4. ed. New York: McGraw-Hill, 2001.

MCKINNEY, W. **Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython**. Sebastopol, CA: O'Reilly Media, 2012.

MJV INNOVATION. O que é ETL e como funciona?. **MJV Innovation**, 2024. Disponível em: <https://www.mjvinnovation.com/pt-br/blog/o-que-e-etl-como-funciona/>. Acesso em: 2 dez. 2024.

OPEN KNOWLEDGE FOUNDATION. **Open Data Handbook**. 2015. Disponível em: <https://opendatahandbook.org>. Acesso em: 18 dez. 2024.

ORGANIZAÇÃO PARA COOPERAÇÃO E DESENVOLVIMENTO ECONÔMICO (OCDE). **The economic and social impact of open government: enhancing access and inclusion through public data**. Paris: OECD Publishing, 2021. Disponível em: <https://www.oecd.org/gov/open-government/impact-of-open-government.pdf>. Acesso em: 12 mai. 2025.

PEREIRA, C. S. S. Produção do espaço, transformações do capitalismo e estratégias imobiliárias das empresas varejistas. **Mercator**, Fortaleza, v. 19, e19004, 2020.

PINDYCK, R. S.; RUBINFELD, D. L. **Microeconomia**. 6. ed. São Paulo: Pearson Prentice Hall, 2006.

PINTAUDI, S. M. A cidade e as formas do comércio. *In*: CARLOS, A. F. A. (Org.). **Novos caminhos da geografia**. São Paulo: Contexto, 1999. p. 59-70.

PREFEITURA MUNICIPAL DE CURITIBA. **Clique Economia**: base de dados de preços do varejo alimentício de Curitiba. Curitiba, 2025. Disponível em: <https://dadosabertos.c3sl.ufpr.br/curitiba/CliqueEconomia/>. Acesso em: 3 jan. 2025.

RAJIV, Surendra; DUTTA, Shantanu; DHAR, Sanjay K. Asymmetric Store Positioning and Promotional Advertising Strategies: Theory and Evidence. **Marketing Science**, Hanover, v. 21, n. 1, p. 74-96, 2002.

SANTOS, M. **O Espaço Dividido**: os dois circuitos da economia urbana. Rio de Janeiro: Francisco Alves, 1979.

SANTOS, M.; SILVEIRA, M. L. **O Brasil**: território e sociedade no início do século XXI. Rio de Janeiro: Record, 2001.

SHMUELI, G. *et al.* **Data Mining for Business Analytics**: Concepts, Techniques, and Applications in R. New York: Wiley, 2017.

STIGLER, George J. The Economics of Information. **Journal of Political Economy**, v. 69, n. 3, p. 213-225, jun. 1961.

TRIBUNA PR. Variação de preços dos alimentos chega a 202% em Curitiba. **Tribuna PR**, 2021. Disponível em: <https://www.tribunapr.com.br/noticias/curitiba-regiao/variacao-de-precos-dos-alimentos-chega-a-202-em-curitiba-veja-tabela-completa/>. Acesso em: 1 mar. 2024.

TUKEY, J. W. **Exploratory Data Analysis**. Reading: Addison-Wesley, 1977.

UNITED STATES DEPARTMENT OF AGRICULTURE (USDA). **Food Access Research Atlas**. Washington, D.C., 2021. Disponível em: <https://www.ers.usda.gov/data-products/food-access-research-atlas/>. Acesso em: 11 fev. 2025.

UOL ECONOMIA. Alimentação no domicílio teve maior alta no IPCA desde janeiro, diz IBGE. **UOL Economia**, 2024. Disponível em:

<https://economia.uol.com.br/noticias/estadao-conteudo/2024/12/10/alimentacao-no-domicilio-teve-maior-alta-no-ipca-desde-janeiro-diz-ibge.htm>. Acesso em: 12 dez. 2024.

VARIAN, Hal R. A model of sales. **American Economic Review**, v. 70, n. 4, p. 651-659, set. 1980.

APÊNDICE 1 - SCRIPT DE EXTRAÇÃO E CONCATENAÇÃO DAS BASES

```
import os
import requests
from bs4 import BeautifulSoup
import pandas as pd
import sys

# --- 1. CONFIGURAÇÕES ---
# URL do repositório de dados abertos
URL_DADOS = 'https://dadosabertos.c3sl.ufpr.br/curitiba/CliqueEconomia/'

# Pasta local para salvar os arquivos baixados e o resultado final
PASTA_SAIDA = 'dados_clique_economia'

# Nome do arquivo CSV consolidado que será gerado
ARQUIVO_FINAL = 'CliqueEconomia_Consolidado_2024.csv'

# --- 2. FUNÇÃO PARA LISTAR ARQUIVOS NO REPOSITÓRIO ---
def listar_arquivos_csv(url):
    """
    Acessa a URL do repositório e extrai todos os links de arquivos CSV.
    """
    print(f"Acessando o repositório em: {url}")
    try:
        response = requests.get(url)
        response.raise_for_status() # Lança um erro para status ruins (ex: 404)
        soup = BeautifulSoup(response.text, 'html.parser')
        # Encontra todas as tags 'a' (links) cujo href termina com '.csv'
        links_csv = [a['href'] for a in soup.find_all('a') if a['href'].endswith('.csv')]
        print(f"Encontrados {len(links_csv)} arquivos CSV no total.")
        return links_csv
    except requests.exceptions.RequestException as e:
        print(f"ERRO: Falha ao acessar a URL. Verifique sua conexão. Detalhe: {e}")
        sys.exit()

# --- 3. EXECUÇÃO DO PROCESSO DE EXTRAÇÃO ---
if __name__ == "__main__":
    # Garante que a pasta de saída exista
    if not os.path.exists(PASTA_SAIDA):
        os.makedirs(PASTA_SAIDA)
        print(f"Pasta '{PASTA_SAIDA}' criada.")

    # Etapa de Extração
    todos_os_arquivos = listar_arquivos_csv(URL_DADOS)

    # Filtra apenas os arquivos relevantes (ano 2024 e de 'Base_de_Dados')
    print("\nFiltrando arquivos para o ano de 2024...")
    arquivos_para_baixar = [
        f for f in todos_os_arquivos
        if "2024" in f and "Base_de_Dados" in f
    ]
    print(f"Selecionando {len(arquivos_para_baixar)} arquivos para download.")

    # Loop para baixar cada arquivo CSV filtrado
    print("\nIniciando download dos arquivos...")
    for nome_arquivo in arquivos_para_baixar:
        caminho_local = os.path.join(PASTA_SAIDA, nome_arquivo)
        # Verifica se o arquivo já existe para não baixar novamente
        if not os.path.exists(caminho_local):
            url_arquivo = URL_DADOS + nome_arquivo
            print(f"Baixando '{nome_arquivo}'...")
            try:
                r = requests.get(url_arquivo, stream=True)
                r.raise_for_status()
                with open(caminho_local, 'wb') as f:
                    for chunk in r.iter_content(chunk_size=8192):
                        f.write(chunk)
            except requests.exceptions.RequestException as e:
```

```

        print(f"    -> Falha ao baixar {nome_arquivo}. Erro: {e}")
    else:
        print(f"    -> Arquivo '{nome_arquivo}' já existe. Pulando.")
    print("Download concluído.")

# Etapa de Consolidação
print("\nIniciando consolidação dos arquivos em um único CSV...")
lista_dfs = []
arquivos_baixados = [f for f in os.listdir(PASTA_SAIDA) if f.endswith('.csv') and "2024" in f and
"Base_de_Dados" in f]

for nome_arquivo in arquivos_baixados:
    caminho_completo = os.path.join(PASTA_SAIDA, nome_arquivo)
    try:
        # Lê cada CSV. O encoding 'latin-1' é mais robusto para dados governamentais brasileiros.
        df_temp = pd.read_csv(caminho_completo, delimiter=';', encoding='latin-1', low_memory=False)
        lista_dfs.append(df_temp)
    except Exception as e:
        print(f"Erro ao ler o arquivo {nome_arquivo}: {e}")

if lista_dfs:
    # Concatena todos os DataFrames da lista em um único DataFrame
    df_consolidado = pd.concat(lista_dfs, ignore_index=True)
    print(f"Total de {len(df_consolidado)} registros consolidados.")

    # Salva o resultado final
    caminho_final = os.path.join(PASTA_SAIDA, ARQUIVO_FINAL)
    df_consolidado.to_csv(caminho_final, index=False, sep=';', encoding='utf-8-sig')
    print(f"\n✅ Processo concluído! Arquivo consolidado salvo em: '{caminho_final}'")
else:
    print("\nNenhum dado para consolidar. Processo encerrado.")

```

APÊNDICE 2 - TABELA DE ENDEREÇOS

id_empre	endereço_completo	logradouro	logradouro_nome	numero	bairro	cidade	estado
113	alameda doutor carlos de carvalho 2050 - bigorinho - curitiba/pr	al	Doutor Carlos de Carvalho	2050	bigorinho	curitiba	pr
49640849	av juscelino kubitschek de oliveira, 2297 - cidade industrial - curitiba/pr	av	Juscelino Kubitschek de Oliveira	2297	cic	curitiba	pr
41762770	av. juscelino kubitschek de oliveira, 2511 - cidade industrial de curitiba - curitiba/pr	av	Juscelino Kubitschek de Oliveira	2511	cic	curitiba	pr
160	av. presidente wenceslau braz, n.º 1046 - curitiba/pr	av	presidente wenceslau braz	1046	guaira	curitiba	pr
103	av. rep. argentina, 900 - curitiba/pr	av	Rep. Argentina	900	agua verde	curitiba	pr
26614363	avenida alcir martins bastos 483 - fazendinha - curitiba/pr	av	Alcir Martins Bastos	483	fazendinha	curitiba	pr
4555459	avenida anita garibaldi, 4654 - curitiba/pr	av	Anita Garibaldi	4654	Barreirinha	curitiba	pr
22	avenida Brasília 6900 - novo mundo - curitiba/pr	av	Brasília	6900	novo mundo	curitiba	pr
40523255	avenida caviúna - bairro alto - curitiba/pr	av	Caviúna	472	bairro alto	curitiba	pr
76	avenida comendador franco 6005 - uberaba - curitiba/pr	av	Comendador Franco	6005	uberaba	curitiba	pr
96	avenida francisco krüger 6646 - cachoeira - curitiba/pr	av	Francisco Krüger	6646	cachoeira	curitiba	pr
45	avenida Manoel Ribas 8243 - butiatuvinha - curitiba/pr	av	Manoel Ribas	8243	butiatuvinha	curitiba	pr
157	avenida Marechal Floriano Peixoto 10520 - alto boqueirão - curitiba/pr	av	Marechal Floriano Peixoto	10520	alto boqueirão	curitiba	pr
48	avenida Marechal Floriano Peixoto 3031 - rebouças - curitiba/pr	av	Marechal Floriano Peixoto	3031	rebouças	curitiba	pr
66898846	avenida Marechal Floriano Peixoto, 5350 - hauer - curitiba/pr	av	Marechal Floriano Peixoto	5350	hauer	curitiba	pr
74244342	avenida Marechal Floriano Peixoto, 6400 - hauer - curitiba/pr	av	Marechal Floriano Peixoto	6400	hauer	curitiba	pr
85	avenida Marechal Mascarenhas de Moraes 1463 - atuba - curitiba/pr	av	Marechal Mascarenhas de Moraes	1463	atuba	curitiba	pr
86	avenida Marechal Mascarenhas de Moraes 312 - atuba - curitiba/pr	av	Marechal Mascarenhas de Moraes	312	atuba	curitiba	pr
55341648	avenida Paraná 1250 - cabral - curitiba/pr	av	Paraná	1250	cabral	curitiba	pr
175	avenida prefeito Erasto Gaertner 2194 - Bacacheri - curitiba/pr	av	Prefeito Erasto Gaertner	2194	Bacacheri	curitiba	pr
107	avenida presidente Arthur da Silva Bernardes 2200 - portão - curitiba/pr	av	presidente Arthur da Silva Bernardes	2200	portão	curitiba	pr
100	avenida presidente Kennedy 1000 - parolin - curitiba/pr	av	presidente Kennedy	1000	parolin	curitiba	pr
75	avenida senador Salgado Filho 4487 - uberaba - curitiba/pr	av	senador Salgado Filho	4487	uberaba	curitiba	pr
51	avenida sete de setembro 4568 - batel - curitiba/pr	av	sete de setembro	4568	batel	curitiba	pr
42	avenida Ver Toaldo Túlio, 2780 - curitiba/pr	av	Vereador Toaldo Túlio	2780	santa felicidade	curitiba	pr
43	avenida Ver Toaldo Túlio, nº2184 - novo mundo - curitiba/pr	av	Vereador Toaldo Túlio	2184	novo mundo	curitiba	pr
133	avenida Vicente Machado 18 - centro - curitiba/pr	av	avenida Vicente Machado	18	centro	curitiba	pr
73	avenida Victor Ferreira do Amaral, 1088 - curitiba/pr	av	avenida Victor Ferreira do Amaral	1088	Taruma	curitiba	pr
72	avenida Victor Ferreira do Amaral, 816 - Tarumã - curitiba/pr	av	avenida Victor Ferreira do Amaral	816	Taruma	curitiba	pr
381180	avenida Winston Churchill, 1734 - Capão Raso - curitiba/pr	av	Winston Churchill	1734	Capão Raso	curitiba	pr
76104300	avenida Winston Churchill, 2170 - Capão Raso - curitiba/pr	av	Winston Churchill	2170	Capão Raso	curitiba	pr
163	br 4761, 1801 - atuba - curitiba/pr	rod	br 4761	1801	atuba	curitiba	pr
162	br-116, 3041 - bairro alto - curitiba/pr	rod	br 116	3041	bairro alto	curitiba	pr
35	governador Ney Braga, 13807 - Xaxim - curitiba/pr	av	Governador Ney Braga	13807	Xaxim	curitiba	pr
43940473	r. eduardo pinto da rocha - 81750-450 - curitiba/pr	rua	eduardo pinto da rocha - 81750-450 - curitiba/pr	2629	sítio cercado	curitiba	pr
80258131	r. emanoel ernesto bertoldi, 1025 - campo de santana - curitiba/pr	rua	emanoel ernesto bertoldi	1025	campo de santana	curitiba	pr
28	r. francisco derosso, 3300 - curitiba/pr	rua	francisco derosso	3300	Xaxim	curitiba	pr
78	r. José de Oliveira Franco, 2623 - curitiba/pr	rua	José de Oliveira Franco	2623	bairro alto	curitiba	pr
33223745	r. João Bettega, 1930 - novo mundo - curitiba/pr	rua	João Bettega	1930	novo mundo	curitiba	pr
158	r. João Bettega, 2576 - portão - curitiba/pr	rua	João Bettega	2576	portão	curitiba	pr
52959500	r. Ourizona, 641 - sítio cercado - curitiba/pr	rua	Ourizona	641	sítio cercado	curitiba	pr
78743218	r. Ourizona, 641 - sítio cercado - curitiba/pr	rua	Ourizona	641	sítio cercado	curitiba	pr
159	rodovia br-116 - Tarumã - curitiba/pr	rod	br 116	102	hauer	curitiba	pr
83033673	rodovia br-116, 18.580 - pinheirinho - curitiba/pr	rod	br-116	18580	pinheirinho	curitiba	pr
92033480	rua agostinho merlin 586 - portão - curitiba/pr	rua	agostinho merlin	586	portão	curitiba	pr
1587517	rua Ailton Ribeiro Marques - Tatuquara - curitiba/pr	rua	Olívio José Bocchi	400	Tatuquara	curitiba	pr
80	rua Alberico Flores Bueno 886 - bairro alto - curitiba/pr	rua	Alberico Flores Bueno	886	bairro alto	curitiba	pr
38	rua Alfredo Pujol, 38 - cidade industrial - curitiba/pr	rua	Alfredo Pujol	38	cic	curitiba	pr
94	rua Amaury Lange Silvério 1130 - Pilarzinho - curitiba/pr	rua	Amaury Lange Silvério	1130	Pilarzinho	curitiba	pr
67	rua Ana Berta Roskamp, 10 - curitiba/pr	rua	Ana Berta Roskamp	10	Jardim das Americas	curitiba	pr
3472158	rua Antônio Zak, 363 - alto boqueirão - curitiba/pr	rua	Antônio Zak	363	alto boqueirão	curitiba	pr
56	rua Brigadeiro Franco 641 - mercês - curitiba/pr	rua	Brigadeiro Franco	641	Mercês	curitiba	pr
88	rua Carmelina Cavassin 201 - Abranches - curitiba/pr	rua	Carmelina Cavassin	201	Abranches	curitiba	pr
89	rua Carmelina Cavassin 850 - Abranches - curitiba/pr	rua	Carmelina Cavassin	850	Abranches	curitiba	pr
80695901	rua Celeste Tortato Gabardo, 1820 - sítio cercado - curitiba/pr	rua	Celeste Tortato Gabardo	1820	sítio cercado	curitiba	pr
52	rua Coronel Dulcídio - Batel - curitiba/pr	rua	Coronel Dulcídio	880	Batel	curitiba	pr
30	rua Coronel Luiz José dos Santos, 3241 - Xaxim - curitiba/pr	rua	Coronel Luiz José dos Santos	3241	Xaxim	curitiba	pr
70	rua Delegado Leopoldo Belczak 1473 - Capão da Imbuia - curitiba/pr	rua	Delegado Leopoldo Belczak	1473	Capão da Imbuia	curitiba	pr
46	rua Deputado Heitor Alencar Furtado 1210 - Mossunguê - curitiba/pr	rua	Deputado Heitor Alencar Furtado	1210	Mossunguê	curitiba	pr
26	rua Desembargador Antônio de Paula - 2765 - Boqueirão - curitiba/pr	rua	Desembargador Antônio de Paula	2765	Boqueirão	curitiba	pr
2711048	rua dos Bem-te-vis - novo mundo - curitiba/pr	rua	dos Bem-te-vis	295	novo mundo	curitiba	pr
66522563	rua Doutor César Permetta, 295 - Tatuquara - curitiba/pr	rua	Doutor César Permetta	483	Tatuquara	curitiba	pr
29319726	rua Doutor César Permetta, 483 - Tatuquara - curitiba/pr	rua	Doutor César Permetta	3605	Tatuquara	curitiba	pr
106	rua Eduardo Carlos Pereira 3605 - novo mundo - curitiba/pr	rua	Eduardo Carlos Pereira	2836	novo mundo	curitiba	pr
2	rua Eduardo Pinto da Rocha 2836 - sítio cercado - curitiba/pr	rua	Eduardo Pinto da Rocha	2836	sítio cercado	curitiba	pr
124	rua Eduardo Pinto da Rocha, 469 - alto boqueirão - curitiba/pr	rua	Eduardo Pinto da Rocha	469	alto boqueirão	curitiba	pr
130	rua Eduardo Sprada, 5105 - curitiba/pr	rua	Eduardo Sprada	5105	Tatuquara	curitiba	pr
26094664	rua Ernesto Germano Francisco Hannemann, 174 - Tatuquara - curitiba/pr	rua	Ernesto Germano Francisco Hannemann	174	Tatuquara	curitiba	pr
66	rua Eunice Bettini Bartoszek, 1930 - Uberaba - curitiba/pr	rua	Eunice Bettini Bartoszek	1930	Uberaba	curitiba	pr
90676964	rua Francisco Claudino Ferreira - sítio cercado - curitiba/pr	rua	Francisco Claudino Ferreira	1244	sítio cercado	curitiba	pr
64735871	rua Francisco Derosso - Xaxim - curitiba/pr	rua	Francisco Derosso	2652	Xaxim	curitiba	pr
29	rua Francisco Derosso, 2973 - Xaxim - curitiba/pr	rua	Francisco Derosso	2973	Xaxim	curitiba	pr
27	rua Francisco Derosso, 4693 - alto boqueirão - curitiba/pr	rua	Francisco Derosso	4693	alto boqueirão	curitiba	pr
111	rua General Potiguara 2895 - novo mundo - curitiba/pr	rua	General Potiguara	2895	novo mundo	curitiba	pr
17	rua Izaac Ferreira da Cruz, 1440 - pinheirinho - curitiba/pr	rua	Izaac Ferreira da Cruz	1440	pinheirinho	curitiba	pr
5	rua Izaac Ferreira da Cruz, 2941 - sítio cercado - curitiba/pr	rua	Izaac Ferreira da Cruz	2941	sítio cercado	curitiba	pr
6	rua Izaac Ferreira da Cruz, 3868 - curitiba/pr	rua	Izaac Ferreira da Cruz	3868	sítio cercado	curitiba	pr
110	rua José Benedito Cottolengo 1191 - campo Comprido - curitiba/pr	rua	José Benedito Cottolengo	1191	campo Comprido	curitiba	pr
109	rua João Alencar Guimarães, 2126 - curitiba/pr	rua	João Alencar Guimarães	2126	campo Comprido	curitiba	pr
156	rua João Bettega, 4280 - novo mundo - curitiba/pr	rua	João Bettega	4280	novo mundo	curitiba	pr
40	rua João Dembinski 1410 - cidade industrial - curitiba/pr	rua	João Dembinski	1410	cic	curitiba	pr
84	rua Lodovico Geronazzo 805 - Boa Vista - curitiba/pr	rua	Lodovico Geronazzo	805	Boa Vista	curitiba	pr
52519003	rua Lothario Boutin, 554 - pinheirinho - curitiba/pr	rua	Lothario Boutin	554	pinheirinho	curitiba	pr
25	rua Maestro Carlos Frank 2812 - boqueirão - curitiba/pr	rua	Maestro Carlos Frank	2812	boqueirão	curitiba	pr
128	rua Manoel Máximo dos Santos, 141 - Cajuru - curitiba/pr	rua	Manoel Máximo dos Santos	141	Cajuru	curitiba	pr
136	rua Marcos Antônio Malucelli, 947 - curitiba/pr	rua	Marcos Antônio Malucelli	947	cic	curitiba	pr
93424208	rua Marcos Bertoldi, 700 - campo de santana - curitiba/pr	rua	Marcos Bertoldi	700	campo de santana	curitiba	pr
115	rua Martin Afonso 2162 - bigorinho - curitiba/pr	rua	Martin Afonso	2162	bigorinho	curitiba	pr
90	rua Mateus Leme, 4556 - novo mundo - curitiba/pr	rua	Mateus Leme	4556	novo mundo	curitiba	pr
71533540	rua Nicola Pellanda 1346 - pinheirinho - curitiba/pr	rua	Nicola Pellanda	1346	pinheirinho	curitiba	pr
12455124	rua Nova Aurora, 2509 - novo mundo - curitiba/pr	rua	Nova Aurora	2509	novo mundo	curitiba	pr
56602246	rua Ourizona, 3159 - sítio cercado - curitiba/pr	rua	Ourizona	3159	sítio cercado	curitiba	pr
41	rua Padre José Lopacinski 740 - cidade industrial - curitiba/pr	rua	Padre José Lopacinski	740	cic	curitiba	pr
82681899	rua Professor Alcacyr Munhoz Mader, 4538 - cidade industrial de curitiba - curitiba/pr	rua	Professor Alcacyr Munhoz Mader	4538	cic	curitiba	pr
44	rua Professor Francisco Zardo 58 - santa felicidade - curitiba/pr	rua	Professor Francisco Zardo	58	santa felicidade	curitiba	pr
117	rua Professor Pedro Viriato Parigot de Souza 600 - Mossunguê - curitiba/pr	rua	Professor Pedro Viriato Parigot de Souza	600	Mossunguê	curitiba	pr
104	rua Professor Álvaro Jorge 75 - Vila Izabel - curitiba/pr	rua	Professor Álvaro Jorge	75	Vila Izabel	curitiba	pr
32	rua Raposo Tavares 1236 - Pilarzinho - curitiba/pr	rua	Raposo Tavares	1236	Pilarzinho	curitiba	pr
93	rua Raposo Tavares 250 - Pilarzinho - curitiba/pr	rua	Raposo Tavares	250	Pilarzinho	curitiba	pr
4120472	rua Raul Pompéia - cidade industrial - curitiba/pr	rua	Raul Pompéia	330	cic	curitiba	pr
84346287	rua Tijucas do Sul, 1491 - sítio cercado - curitiba/pr	rua	Tijucas do Sul	1491	sítio cercado	curitiba	pr
15039153	rua Uriel Nogueira dos Santos - campo de santana - curitiba/pr	rua	Uriel Nogueira dos Santos	632	campo de santana	curitiba	pr
64677704	rua Visconde do Serro Frio - novo mundo - curitiba/pr	rua	Visconde do Serro Frio	164	novo mundo	curitiba	pr
0	rua William Booth 2623 - boqueirão - curitiba/pr	rua	William Booth	2623	Boqueirão	curitiba	pr

APÊNDICE 3 - SCRIPT DE PRÉ-PROCESSAMENTO E CLASSIFICAÇÃO DE DADOS

```
import pandas as pd
import re
import sys

# =====
# --- 1. CONFIGURAÇÃO E DICIONÁRIOS DE CLASSIFICAÇÃO ---
# =====
print("Iniciando a operação final de preparação de dados...")

# --- Definição dos Caminhos de Arquivo ---
# Ajuste esses caminhos conforme a localização dos seus dados.
caminho_principal = r'S:\CliqueEconomia_dataset_variaveisnormalizadas_V2.csv'
caminho_enderecos_com_id = r'S:\endereço_normalizado_COM_ID.csv'
caminho_gabarito_ouro = 'Classificacao_Refinada_de_Produtos.csv'
caminho_saida_definitivo = r'S:\CliqueEconomia_DATASET_PRONTO_PARA_TCC.csv'

# --- Dicionários de Mapeamento para Classificação de Produtos ---

# CAMADA 1: GABARITO DE OURO
# Este dicionário prioriza correções manuais e exceções específicas.
gabarito_ouro_map = {
    'leite condensado': 'DOCES_E_SOBREMESAS',
    'doce de leite': 'DOCES_E_SOBREMESAS',
    'creme de leite': 'CREME_E_MANTEIGA',
    'leite de coco': 'DOCES_E_SOBREMESAS',
    'composto lacteo': 'PRODUTOS_INFANTIS',
    'leite fermentado': 'IOGURTE',
    'bebida lactea': 'IOGURTE',
    'batata doce': 'LEGUMES_E_VERDURAS',
    'batata palha': 'BISCOITOS_E_SNACKS',
    'batata frita': 'BISCOITOS_E_SNACKS',
    'couve manteiga': 'LEGUMES_E_VERDURAS',
    'cebola para conserva': 'LEGUMES_E_VERDURAS',
    'milho verde': 'LEGUMES_E_VERDURAS',
    'farinha de arroz': 'FARINACEOS',
    'pao fatiado': 'PAES',
    'pao bisnaguinha': 'PAES',
    'pao frances': 'PAES',
    'caldo de galinha': 'MOLHOS_E_CONDIMENTOS',
    'atum ralado': 'ENLATADOS_E_CONSERVAS',
    'sardinha molho': 'ENLATADOS_E_CONSERVAS',
    'sabao em po': 'LIMPEZA_ROUPA',
    'sabao em barras': 'LIMPEZA_ROUPA',
    'suco integral': 'SUCO_E_NECTAR',
    'po p/ refresco': 'SUCO_E_NECTAR',
    'po p/ bebida': 'SUCO_E_NECTAR',
    'filtro de papel': 'UTILIDADES_DOMESTICAS',
    'filtro p/cafe': 'UTILIDADES_DOMESTICAS',
    'limpa vidros': 'LIMPEZA_CASA',
    'geleia de morango': 'DOCES_E_SOBREMESAS',
    'geleia de frutas': 'DOCES_E_SOBREMESAS',
    'margarina cremosa': 'CREME_E_MANTEIGA',
    'manteiga primeira qualidade': 'CREME_E_MANTEIGA',
    'formula infantil': 'PRODUTOS_INFANTIS',
    'nestogeno': 'PRODUTOS_INFANTIS',
    'cafe vacuo': 'CAFE_E_CHA',
    'salsichao': 'EMBUTIDOS',
    'charque': 'CARNES',
    'farofa': 'FARINACEOS',
    'bolinho recheado': 'BISCOITOS_E_SNACKS',
    'tempero completo': 'MOLHOS_E_CONDIMENTOS',
    'tempero alho': 'MOLHOS_E_CONDIMENTOS',
    'nata': 'CREME_E_MANTEIGA',
    'cereal alimentacao infantil': 'PRODUTOS_INFANTIS',
    'creme de hidratacao': 'CUIDADOS_PESSOAIS',
    'cachaca': 'BEBIDAS_ALCOOLICAS',
    'salsa': 'LEGUMES_E_VERDURAS',
    'cheiro verde': 'LEGUMES_E_VERDURAS',
}
```

```

    'hortalica': 'LEGUMES_E_VERDURAS',
    'saponaceo cremoso': 'LIMPEZA_CASA',
    'pipoca para microondas': 'BISCOITOS_E_SNACKS',
    'aveia em flocos': 'CEREAIS_E_GRANOLAS',
    'barra de nuts': 'CEREAIS_E_GRANOLAS',
    'barra de cereais': 'CEREAIS_E_GRANOLAS',
    'champignon': 'ENLATADOS_E_CONSERVAS',
    'grao de bico': 'FEIJAO',
    'maria mole': 'DOCES_E_SOBREMESAS'
}

# CAMADA 2: DICIONÁRIO PRINCIPAL
# Contém as regras de classificação gerais, aplicadas após o gabarito.
keyword_map_principal = {
    'QUEIJO': ['queijo', 'mussarela', 'prato', 'ricota', 'requeijao', 'parmesao', 'minas', 'ralado', 'cream cheese'],
    'IOGURTE': ['iogurte', 'yakult', 'petit suisse'],
    'EMBUTIDOS': ['presunto', 'mortadela', 'salame', 'linguica', 'calabresa', 'bacon', 'salsicha', 'fatiado', 'embutido',
'morcele'],
    'CONGELADOS': ['hamburguer', 'congelada', 'congelado'],
    'AVES': ['frango', 'coxa', 'sobrecosta', 'peito', 'asa', 'ave', 'galinha', 'sambiquira'],
    'CARNE_BOVINA': ['bovina', 'alcatra', 'patinho', 'coxao', 'picanha', 'molda', 'costela', 'osso', 'vacuo', 'pedaco', 'file'],
    'CARNE_SUINA': ['suina', 'pernil', 'lombo', 'bisteca', 'panceta', 'banha'],
    'ARROZ': ['arroz', 'polido', 'parboilizado'],
    'FEIJAO': ['feijao', 'preto', 'carioca', 'grao'],
    'ACUCAR_E_ADOCANTE': ['acucar', 'adocante', 'doce'],
    'CAFE_E_CHA': ['cafe', 'cha', 'cappuccino', 'erva', 'mate'],
    'OLEO_E_AZEITE': ['oleo', 'soja', 'azeite'],
    'FARINACEOS': ['farinha', 'fuba', 'trigo', 'polenta', 'polvilho', 'tapioca'],
    'MASSAS': ['macarrao', 'nhoque', 'lasanha', 'massa', 'pastel', 'sopao'],
    'MOLHOS_E_CONDIMENTOS': ['molho', 'extrato', 'maionese', 'ketchup', 'mostarda', 'sal', 'especiarias', 'vinagre', 'catchup',
'caldo', 'tempero', 'vinagrete'],
    'BISCOITOS_E_SNACKS': ['biscoito', 'bolacha', 'torrada', 'salgadinho', 'waffer', 'cookie', 'cookies', 'pipoca'],
    'ENLATADOS_E_CONSERVAS': ['milho', 'ervilha', 'seleta', 'atum', 'sardinha', 'palmito', 'azeitona', 'conserva', 'pessegos'],
    'DOCES_E_SOBREMESAS': ['chocolate', 'achocolatado', 'bombom', 'bala', 'goiabada', 'gelatina', 'pudim', 'mel', 'barra'],
    'FERMENTO_E_MISTURAS': ['fermento', 'biologico', 'mistura', 'bolo', 'po'],
    'CEREAIS_E_GRANOLAS': ['granola', 'castanha', 'nozes', 'amendoim', 'lentilha', 'sagu', 'quixerinha', 'aveia', 'nuts'],
    'PAES': ['pao', 'bisnaga'],
    'FRUTAS': ['banana', 'maca', 'laranja', 'mamao', 'uva', 'limao', 'frutas', 'manga', 'morango', 'maracuja', 'melao', 'abacaxi',
'goiaba', 'abacate', 'pera', 'pessego', 'kiwi', 'ponkan', 'ameixa', 'caqui', 'melancia', 'mimosa', 'jabuticaba', 'figo'],
    'LEGUMES_E_VERDURAS': ['batata', 'cebola', 'alho', 'tomate', 'cenoura', 'alface', 'couve', 'brocolis', 'abobora', 'pepino',
'pimentao', 'repolho', 'chuchu', 'beterraba', 'berinjela', 'abobrinha', 'kabotia', 'moranga', 'agriao', 'rucula', 'acelga', 'aipim',
'cebolinha', 'gengibre', 'inhame', 'rabanete', 'quiabo', 'vagem', 'pinhao'],
    'OVOS': ['ovos', 'ovo'],
    'BEBIDAS_ALCOOLICAS': ['cerveja', 'vinho', 'conhaque', 'energetico', 'vodka', 'whisky'],
    'SUCO_E_NECTAR': ['suco', 'nectar', 'del valle', 'tang'],
    'REFRIGERANTE': ['refrigerante', 'coca-cola', 'guarana', 'pepsi', 'soda', 'pet'],
    'AGUA': ['agua mineral'],
    'LIMPEZA_ROUPA': ['sabao', 'amaciante', 'lava roupas', 'tira manchas'],
    'LIMPEZA_CASA': ['agua sanitaria', 'desinfetante', 'multiuso', 'limpador', 'detergente', 'cera', 'alcool', 'saponaceo'],
    'UTILIDADES_DOMESTICAS': ['filtro', 'esponja', 'aco', 'saco', 'lixo', 'plastico', 'fosforo', 'papel toalha', 'guardanapo',
'rolo', 'folha', 'vela', 'vassoura', 'toalha', 'rodo', 'lampada'],
    'HIGIENE_BUCAL': ['creme dental', 'pasta de dente', 'escova dental', 'fio dental', 'enxaguante'],
    'CUIDADOS_PESSOAIS': ['sabonete', 'shampoo', 'condicionador', 'absorvente', 'desodorante', 'locacao', 'gel'],
    'PAPEL_HIGIENICO': ['papel higienico'],
    'PRODUTOS_INFANTIS': ['fralda', 'lenco umedecido'],
    'LEITE': ['leite'],
    'CREME_E_MANTEIGA': ['manteiga', 'margarina'],
    'COCO': ['coco'],
}

# MAPA DE MACROCATEGORIAS
# Mapeia as categorias de nível 2 para categorias de nível 1.
macro_map = {
    'LEITE': 'LATICINIOS_E_FRIOS', 'QUEIJO': 'LATICINIOS_E_FRIOS', 'IOGURTE': 'LATICINIOS_E_FRIOS', 'CREME_E_MANTEIGA':
'LATICINIOS_E_FRIOS', 'EMBUTIDOS': 'LATICINIOS_E_FRIOS',
    'CONGELADOS': 'CONGELADOS_E_FRIOS', 'CARNE_BOVINA': 'CARNES', 'AVES': 'CARNES', 'CARNE_SUINA': 'CARNES', 'CARNES': 'CARNES',
    'ARROZ': 'MERCEARIA', 'FEIJAO': 'MERCEARIA', 'ACUCAR_E_ADOCANTE': 'MERCEARIA', 'CAFE_E_CHA': 'MERCEARIA', 'OLEO_E_AZEITE':
'MERCEARIA', 'FARINACEOS': 'MERCEARIA', 'MASSAS': 'MERCEARIA', 'MOLHOS_E_CONDIMENTOS': 'MERCEARIA', 'BISCOITOS_E_SNACKS':
'MERCEARIA', 'ENLATADOS_E_CONSERVAS': 'MERCEARIA', 'DOCES_E_SOBREMESAS': 'MERCEARIA', 'FERMENTO_E_MISTURAS': 'MERCEARIA',
'CEREAIS_E_GRANOLAS': 'MERCEARIA',
    'PAES': 'PADARIA',
}

```

```

'LEGUMES_E_VERDURAS': 'HORTIFRUTI', 'FRUTAS': 'HORTIFRUTI', 'OVOS': 'HORTIFRUTI', 'COCO': 'HORTIFRUTI',
'REFRIGERANTE': 'BEBIDAS', 'SUCO_E_NECTAR': 'BEBIDAS', 'AGUA': 'BEBIDAS', 'BEBIDAS_ALCOOLICAS': 'BEBIDAS',
'LIMPEZA_ROUPA': 'LIMPEZA_E_UTILIDADES', 'LIMPEZA_CASA': 'LIMPEZA_E_UTILIDADES', 'UTILIDADES_DOMESTICAS':
'LIMPEZA_E_UTILIDADES',
'HIGIENE_BUCAL': 'HIGIENE_PESSOAL', 'CUIDADOS_PESSOAIS': 'HIGIENE_PESSOAL', 'PAPEL_HIGIENICO': 'HIGIENE_PESSOAL',
'PRODUTOS_INFANTIS': 'INFANTIL', 'OUTROS': 'OUTROS'
}

# =====
# --- FUNÇÕES AUXILIARES ---
# =====
def normalizar_texto(texto):
    """
    Converte o texto para minúsculas e remove caracteres acentuados.
    Essencial para padronizar descrições antes da classificação.
    """
    texto = str(texto).lower()
    mapa_acentos = {'á': 'a', 'à': 'a', 'â': 'a', 'ã': 'a', 'é': 'e', 'ê': 'e', 'í': 'i', 'î': 'i', 'ó': 'o', 'ô': 'o', 'õ': 'o', 'ç': 'c', 'ú': 'u', 'û': 'u'}
    for acento, sem_acento in mapa_acentos.items():
        texto = texto.replace(acento, sem_acento)
    return texto

def categorizar_produto_final(descricao_norm):
    """
    Classifica a descrição normalizada de um produto usando a hierarquia de duas camadas:
    1. Gabarito de Ouro para exceções.
    2. Dicionário Principal para regras gerais.
    Retorna 'OUTROS' se nenhuma categoria for encontrada.
    """
    # Aplica a primeira camada: Gabarito de Ouro
    for keyword, categoria in gabarito_ouro_map.items():
        if re.search(r'\b' + re.escape(keyword) + r'\b', descricao_norm):
            return categoria

    # Se não encontrou no gabarito, aplica a segunda camada: Dicionário Principal
    for categoria, keywords in keyword_map_principal.items():
        for keyword in keywords:
            if re.search(r'\b' + re.escape(keyword) + r'\b', descricao_norm):
                return categoria
    return 'OUTROS' # Caso o produto não se encaixe em nenhuma categoria definida.

def padronizar_rede(nome_rede):
    """
    Padroniza nomes de redes de supermercados para consistência.
    A correspondência é feita por substrings para flexibilidade.
    """
    nome_rede_norm = str(nome_rede).upper()
    if 'CONDOR' in nome_rede_norm: return 'CONDOR'
    if 'MUFFATO' in nome_rede_norm: return 'SUPER MUFFATO'
    if 'NOGUEIRA' in nome_rede_norm: return 'SUPERMERCADO NOGUEIRA'
    if 'MASTER' in nome_rede_norm: return 'REDE MASTER'
    if 'CURITIBA' in nome_rede_norm: return 'SUPERMERCADO CURITIBA'
    return nome_rede_norm

# =====
# --- FLUXO PRINCIPAL DO SCRIPT ---
# =====
try:
    # --- ETAPA 1: CARREGAMENTO DE DADOS ---
    print("Carregando arquivos de entrada...")
    # Carrega o dataset principal de produtos.
    df = pd.read_csv(caminho_principal, delimiter=';', encoding='utf-8-sig')
    # Carrega o dataset com informações de endereço das empresas.
    df_enderecos = pd.read_csv(caminho_enderecos_com_id, delimiter=';', encoding='utf-8-sig')

    # Tenta carregar o gabarito manual de produtos.
    # Se o arquivo não for encontrado, o processo continua sem ele, usando apenas os dicionários internos.
    try:
        df_gabarito_ouro_file = pd.read_csv(caminho_gabarito_ouro, delimiter=';', encoding='latin-1')
        print("Arquivo de gabarito manual carregado com sucesso.")
        # Prepara o mapa do gabarito, normalizando descrições e categorias.

```

```

df_gabarito_ouro_file.columns = df_gabarito_ouro_file.columns.str.strip()
df_gabarito_ouro_file.dropna(subset=['descricao', 'categoria_correta'], inplace=True)
df_gabarito_ouro_file['descricao_norm'] = df_gabarito_ouro_file['descricao'].apply(normalizar_texto)
df_gabarito_ouro_file['categoria_correta_norm'] = df_gabarito_ouro_file['categoria_correta'].str.upper().str.replace(' ',
'_')

# Atualiza o dicionário `gabarito_ouro_map` com as entradas do arquivo, garantindo prioridade.
gabarito_ouro_map.update(df_gabarito_ouro_file.set_index('descricao_norm')['categoria_correta_norm'].to_dict())
except FileNotFoundError:
    print(f"AVISO: Arquivo '{caminho_gabarito_ouro}' não encontrado. A classificação manual não será aplicada.")

# --- ETAPA 2: PRÉ-PROCESSAMENTO E LIMPEZA ---
print("Realizando limpeza e padronização dos dados...")
# Remove linhas com valores ausentes críticos para análise.
df.dropna(subset=['preco_regular', 'descricao'], inplace=True)
# Converte a coluna 'preco_regular' para numérica, tratando vírgulas como decimais.
df['preco_regular'] = pd.to_numeric(df['preco_regular'].astype(str).str.replace(',', '.'), errors='coerce')
# Filtra produtos com preço regular maior que zero.
df = df[df['preco_regular'] > 0]
# Garante que 'id_empresa' seja tratado como string para junção.
df['id_empresa'] = df['id_empresa'].astype(str)
df_enderecos['id_empresa'] = df_enderecos['id_empresa'].astype(str)
# Aplica a padronização nos nomes das redes de supermercado.
df['rede'] = df['rede'].apply(padronizar_rede)

# --- ETAPA 3: INTEGRAÇÃO DE DADOS DE ENDEREÇO ---
print("Juntando informações de endereço...")
# Combina o DataFrame principal com o de endereços usando 'id_empresa'.
df_final = pd.merge(df, df_enderecos, on='id_empresa', how='left')

# --- ETAPA 4: CLASSIFICAÇÃO DE PRODUTOS ---
print("Classificando todos os produtos...")
# Normaliza as descrições dos produtos para facilitar a classificação.
df_final['descricao_norm'] = df_final['descricao'].apply(normalizar_texto)
# Aplica a função de categorização para gerar a categoria de nível 2.
df_final['categoria_n2'] = df_final['descricao_norm'].apply(categorizar_produto_final)
# Mapeia as categorias de nível 2 para as macrocategorias de nível 1.
df_final['categoria_n1'] = df_final['categoria_n2'].map(macro_map)
# Garante que qualquer categoria de nível 1 não mapeada seja 'OUTROS'.
df_final['categoria_n1'].fillna('OUTROS', inplace=True)

# --- ETAPA 5: AJUSTES FINAIS E EXPORTAÇÃO ---
print("Realizando limpeza final (duplicatas e organização)...")
# Remove quaisquer linhas duplicadas que possam ter surgido durante o processo.
df_final.drop_duplicates(inplace=True)

# Define a ordem desejada das colunas no dataset final.
ordem_colunas = [
    'data_pesquisa', 'id_empresa', 'rede', 'bairro', 'id_produto', 'descricao',
    'categoria_n1', 'categoria_n2', 'preco_regular', 'logradouro_tipo',
    'logradouro_nome', 'numero', 'cidade', 'estado'
]

# Remove colunas intermediárias que não são necessárias no resultado final.
df_final.drop(columns=['descricao_norm', 'endereco_completo'], errors='ignore', inplace=True)
# Reorganiza as colunas do DataFrame.
df_final = df_final[[col for col in ordem_colunas if col in df_final.columns]]

# Salva o DataFrame processado em um novo arquivo CSV.
df_final.to_csv(caminho_saida_definitivo, index=False, sep=';', encoding='utf-8-sig')

# --- RESULTADOS FINAIS ---
# Verifica a quantidade de itens que não foram categorizados.
outros_final = df_final[df_final['categoria_n2'] == 'OUTROS']

print("\n-----")
print("🏆 PROCESSO CONCLUÍDO COM SUCESSO! 🏆")
print(f"Dataset definitivo salvo em: {caminho_saida_definitivo}")
print(f"Total de linhas no arquivo final: {len(df_final)}")
print(f"Total de itens não categorizados ('OUTROS'): {len(outros_final)}")
if not outros_final.empty:
    print("\nTop 10 descrições de produtos classificados como 'OUTROS':")
    print(outros_final['descricao'].value_counts().head(10).to_string())

```

```
print("-----")  
  
except Exception as e:  
    print(f"\nOcorreu um erro inesperado durante a execução do script: {e}")  
    sys.exit(1) # Sai com código de erro para indicar falha.
```

APÊNDICE 4 - SCRIPT DE VALIDAÇÃO DE QUALIDADE DO DATASET FINAL

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import os
import sys

#=====
# --- 1. CONFIGURAÇÃO E CARREGAMENTO ---
#=====

print("Iniciando Script de Análise Final e Detalhada para o TCC...")

# --- Caminho do Dataset Final ---
caminho_dataset_pronto = r'D:\CliqueEconomia_TCC_DATASET_PRONTO.csv'

# --- Define o caminho absoluto para salvar os resultados ---
pasta_resultados = r'D:\resultados_tcc'
if not os.path.exists(pasta_resultados):
    os.makedirs(pasta_resultados)
print(f"Todos os resultados serão salvos na pasta: '{pasta_resultados}'")

# --- Definições ---
CESTA_ESSENCIAL = [
    'MERCERIA', 'HORTIFRUTI', 'CARNES',
    'LATICINIOS_E_FRIOS', 'LIMPEZA_E_UTILIDADES'
]

# Lista de subcategorias chave para a nova análise detalhada
SUBCATEGORIAS_CHAVE = [
    'LEITE', 'ARROZ', 'FEIJAO', 'CARNE_BOVINA', 'AVES', 'PAES', 'OVOS', 'FRUTAS', 'LEGUMES_E_VERDURAS'
]

# --- Carregamento dos Dados ---
try:
    df = pd.read_csv(caminho_dataset_pronto, delimiter=';', encoding='utf-8-sig')
    df['data_pesquisa'] = pd.to_datetime(df['data_pesquisa'], errors='coerce')
    df.dropna(subset=['data_pesquisa'], inplace=True)
    print(f"Dataset com {len(df)} linhas carregado e datas validadas com sucesso.")
except FileNotFoundError:
    print(f"ERRO CRÍTICO: Arquivo '{caminho_dataset_pronto}' não encontrado.")
    sys.exit()

df_cesta = df[df['categoria_n1'].isin(CESTA_ESSENCIAL)].copy()

#=====
# --- TEMAS 1, 2, 3 e 4: ANÁLISES MACRO (GERA ARQUIVOS 1-6) ---
#=====

print("\nGerando Insights Macro (Temas 1 a 4)...")

# TEMA 1: Análise Geográfica por Bairro
print(" -> Gerando Análise Geográfica por Bairro...")
preco_mediano_bairro = df_cesta.groupby(['bairro', 'categoria_n1'])['preco_regular'].median().unstack()
custo_cesta_bairro = preco_mediano_bairro.sum(axis=1).sort_values(ascending=False)
tabela_ranking_bairros = custo_cesta_bairro.round(2).reset_index(name='indice_custo_cesta_R$')
tabela_ranking_bairros.to_csv(os.path.join(pasta_resultados, '1_tabela_ranking_bairros.csv'), index=False, sep=';', encoding='utf-8-sig')
top_10_caros = tabela_ranking_bairros.head(10).sort_values(by='indice_custo_cesta_R$')
top_10_baratos = tabela_ranking_bairros.tail(10).sort_values(by='indice_custo_cesta_R$')
bairros_plot = pd.concat([top_10_caros, top_10_baratos])
```

```

plt.style.use('seaborn-v0_8-whitegrid')
fig, ax = plt.subplots(figsize=(12, 10))
colors = ['#d9534f']*10 + ['#5cb85c']*10
bars = ax.barh(bairros_plot['bairro'], bairros_plot['indice_custo_cesta_R$'], color=colors)
ax.set_title('Ranking de Bairros por Custo da Cesta Essencial', fontsize=16, pad=20)
ax.set_xlabel('Índice de Custo da Cesta (R$)', fontsize=12)
ax.set_ylabel('Bairro', fontsize=12)
ax.bar_label(bars, fmt='R$ %.2f', padding=3, fontsize=10)
plt.tight_layout()
plt.savefig(os.path.join(pasta_resultados, '1_grafico_ranking_bairros.png'), dpi=300)
plt.close()

# TEMA 2: Análise Competitiva entre Redes
print(" -> Gerando Análise Competitiva entre Redes...")
preco_mediano_rede = df_cesta.groupby(['rede', 'categoria_n1'])['preco_regular'].median().unstack()
custo_cesta_rede = preco_mediano_rede.sum(axis=1).sort_values(ascending=False)
tabela_ranking_redes = custo_cesta_rede.round(2).reset_index(name='indice_custo_cesta_R$')
tabela_ranking_redes.to_csv(os.path.join(pasta_resultados, '2_tabela_ranking_redes.csv'), index=False, sep=';',
encoding='utf-8-sig')
matriz_competitividade = preco_mediano_rede.fillna(0).round(2)
matriz_competitividade.to_csv(os.path.join(pasta_resultados, '3_tabela_matriz_competitividade.csv'), sep=';',
encoding='utf-8-sig')
fig, ax = plt.subplots(figsize=(14, 12))
sns.heatmap(matriz_competitividade, annot=True, fmt=".2f", cmap="viridis_r", linewidths=.5, ax=ax)
ax.set_title('Matriz de Competitividade: Preço Mediano por Rede e Categoria', fontsize=16, pad=20)
ax.set_xlabel('Macro Categoria', fontsize=12)
ax.set_ylabel('Rede de Supermercado', fontsize=12)
plt.xticks(rotation=45, ha='right')
plt.yticks(rotation=0)
plt.tight_layout()
plt.savefig(os.path.join(pasta_resultados, '3_grafico_matriz_competitividade.png'), dpi=300)
plt.close()

# TEMA 3: Análise Temporal e de Volatilidade
print(" -> Gerando Análise Temporal e de Volatilidade...")
stats_preco_categoria = df.groupby('categoria_n2')['preco_regular'].agg(['mean', 'std']).dropna()
stats_preco_categoria['coef_variacao_%'] = (stats_preco_categoria['std'] / stats_preco_categoria['mean']) * 100
tabela_volatilidade = stats_preco_categoria.sort_values(by='coef_variacao_%', ascending=False)
tabela_volatilidade.round(2).to_csv(os.path.join(pasta_resultados, '4_tabela_ranking_volatilidade.csv'), sep=';',
encoding='utf-8-sig')
top_20_volateis = tabela_volatilidade.head(20).sort_values(by='coef_variacao_%')
fig, ax = plt.subplots(figsize=(12, 10))
ax.barh(top_20_volateis.index, top_20_volateis['coef_variacao_%'], color='skyblue')
ax.set_title('Top 20 Categorias com Maior Volatilidade de Preços', fontsize=16, pad=20)
ax.set_xlabel('Coeficiente de Variação (%)', fontsize=12)
ax.set_ylabel('Subcategoria (Nível 2)', fontsize=12)
plt.tight_layout()
plt.savefig(os.path.join(pasta_resultados, '4_grafico_volatilidade.png'), dpi=300)
plt.close()

# --- ALTERAÇÃO AQUI: Lógica para plotar a evolução de cada categoria ---
df['mes'] = df['data_pesquisa'].dt.to_period('M')
# Calcula o preço mediano de cada categoria para cada mês
preco_evolucao_categoria = df.groupby(['mes', 'categoria_n1'])['preco_regular'].median().unstack()

# Salva a tabela detalhada
tabela_evolucao_detalhada = preco_evolucao_categoria.round(2)
tabela_evolucao_detalhada.index = tabela_evolucao_detalhada.index.to_timestamp() # Converte o índice para um formato
mais legível
tabela_evolucao_detalhada.to_csv(os.path.join(pasta_resultados, '5_tabela_evolucao_por_categoria.csv'), sep=';',
encoding='utf-8-sig')

```

```

print(" -> Tabela '5_tabela_evolucao_por_categoria.csv' salva.")

# Gera o novo gráfico de linhas com uma linha para cada categoria
fig, ax = plt.subplots(figsize=(16, 9))
preco_evolucao_categoria.plot(ax=ax, marker='o', linestyle='--')

ax.set_title('Evolução do Preço Mediano por Categoria ao Longo de 2024', fontsize=18, pad=20)
ax.set_xlabel('Mês', fontsize=14)
ax.set_ylabel('Preço Mediano (R$)', fontsize=14)
ax.grid(True, which='both', linestyle='--', linewidth=0.5)
ax.legend(title='Macro Categoria', bbox_to_anchor=(1.02, 1), loc='upper left')
plt.xticks(rotation=45)
plt.tight_layout(rect=[0, 0, 0.85, 1])
plt.savefig(os.path.join(pasta_resultados, '5_grafico_evolucao_categorias.png'), dpi=300)
print(" -> Gráfico '5_grafico_evolucao_categorias.png' (detalhado por categoria) salvo.")
plt.close()
# --- FIM DA ALTERAÇÃO ---

# TEMA 4: Síntese para o Consumidor
print(" -> Gerando Guia do Consumidor Inteligente...")
# A lógica para a tabela 6 permanece a mesma, pois ela se baseia na cesta essencial geral.
preco_mediano_bairro_rede_cat = df_cesta.groupby(['bairro', 'rede', 'categoria_n1'])['preco_regular'].median()
custo_por_cat = preco_mediano_bairro_rede_cat.unstack()
custo_cesta_bairro_rede = custo_por_cat.sum(axis=1, min_count=len(CESTA_ESSENCIAL) - 1)
custo_df = custo_cesta_bairro_rede.reset_index(name='custo_cesta')
idx_menor_custo = custo_df.loc[custo_df.groupby('bairro')['custo_cesta'].idxmin()]
tabela_guiia_consumidor = idx_menor_custo.rename(columns={'rede': 'rede_mais_barata', 'custo_cesta':
'custo_minimo_cesta_R$'}).round(2)
tabela_guiia_consumidor.to_csv(os.path.join(pasta_resultados, '6_tabela_guiia_consumidor.csv'), index=False, sep=';',
encoding='utf-8-sig')

print(" -> Insights Macro (1 a 6) gerados com sucesso.")

#=====
# --- TEMA 5: ANÁLISE DETALHADA POR SUBCATEGORIA (HEATMAPS) ---
#=====
print("\nGerando Insight 5: Heatmaps Detalhados por Subcategoria...")

# Identificar as redes, bairros e subcategorias mais relevantes para focar a análise
top_15_redes = df['rede'].value_counts().nlargest(15).index
top_15_bairros = df['bairro'].value_counts().nlargest(15).index
top_20_subcategorias = df['categoria_n2'].value_counts().nlargest(20).index

# Filtrar o DataFrame para conter apenas esses dados mais relevantes
df_filtrado = df[
    df['rede'].isin(top_15_redes) &
    df['bairro'].isin(top_15_bairros) &
    df['categoria_n2'].isin(top_20_subcategorias)
]

print(f" -> Foco da análise: {len(top_15_redes)} principais redes, {len(top_15_bairros)} principais bairros e
{len(top_20_subcategorias)} principais subcategorias.")

# --- Heatmap: Rede vs. Subcategoria ---
matriz_rede_subcat = df_filtrado.pivot_table(index='rede', columns='categoria_n2', values='preco_regular',
aggfunc='median')
matriz_rede_subcat.round(2).to_csv(os.path.join(pasta_resultados, '11_tabela_heatmap_rede_vs_subcategoria.csv'),
sep=';', encoding='utf-8-sig')
print(" -> Tabela '11_tabela_heatmap_rede_vs_subcategoria.csv' salva.")

fig, ax = plt.subplots(figsize=(22, 14))

```

```

sns.heatmap(matriz_rede_subcat, annot=True, fmt=".2f", cmap="YlGnBu", linewidths=.5, ax=ax)
ax.set_title('Heatmap de Preços: Top 15 Redes vs. Top 20 Subcategorias (Preço Mediano)', fontsize=18, pad=20)
ax.set_xlabel('Subcategoria de Produto (Nível 2)', fontsize=14)
ax.set_ylabel('Rede de Supermercado', fontsize=14)
plt.xticks(rotation=45, ha='right')
plt.yticks(rotation=0)
plt.tight_layout()
plt.savefig(os.path.join(pasta_resultados, '11_grafico_heatmap_rede_vs_subcategoria.png'), dpi=300)
print(" -> Gráfico '11_grafico_heatmap_rede_vs_subcategoria.png' salvo.")
plt.close()

# --- Heatmap: Bairro vs. Subcategoria ---
matriz_bairro_subcat = df_filtrado.pivot_table(index='bairro', columns='categoria_n2', values='preco_regular',
aggfunc='median')
matriz_bairro_subcat.round(2).to_csv(os.path.join(pasta_resultados, '12_tabela_heatmap_bairro_vs_subcategoria.csv'),
sep=';', encoding='utf-8-sig')
print(" -> Tabela '12_tabela_heatmap_bairro_vs_subcategoria.csv' salva.")

fig, ax = plt.subplots(figsize=(22, 14))
sns.heatmap(matriz_bairro_subcat, annot=True, fmt=".1f", cmap="YlOrRd", linewidths=.5, ax=ax)
ax.set_title('Heatmap de Preços: Top 15 Bairros vs. Top 20 Subcategorias (Preço Mediano)', fontsize=18, pad=20)
ax.set_xlabel('Subcategoria de Produto (Nível 2)', fontsize=14)
ax.set_ylabel('Bairro', fontsize=14)
plt.xticks(rotation=45, ha='right')
plt.yticks(rotation=0)
plt.tight_layout()
plt.savefig(os.path.join(pasta_resultados, '12_grafico_heatmap_bairro_vs_subcategoria.png'), dpi=300)
print(" -> Gráfico '12_grafico_heatmap_bairro_vs_subcategoria.png' salvo.")
plt.close()

#=====
# --- TEMA 6: RANKINGS DETALHADOS POR SUBCATEGORIA (TEXTO) ---
#=====
print("\nGerando Insight 6: Rankings Detalhados por Subcategoria...")

# Prepara as tabelas de preço mediano por subcategoria (usando o df original completo)
preco_mediano_bairro_subcat_full = df.groupby(['bairro', 'categoria_n2'])['preco_regular'].median().unstack()
preco_mediano_rede_subcat_full = df.groupby(['rede', 'categoria_n2'])['preco_regular'].median().unstack()

# Salva as tabelas completas para consulta
preco_mediano_bairro_subcat_full.round(2).to_csv(os.path.join(pasta_resultados,
'7_tabela_detalhada_bairros_por_subcategoria.csv'), sep=';', encoding='utf-8-sig')
print(" -> Tabela '7_tabela_detalhada_bairros_por_subcategoria.csv' salva.")
preco_mediano_rede_subcat_full.round(2).to_csv(os.path.join(pasta_resultados,
'8_tabela_detalhada_redes_por_subcategoria.csv'), sep=';', encoding='utf-8-sig')
print(" -> Tabela '8_tabela_detalhada_redes_por_subcategoria.csv' salva.")

# Gera os rankings para as subcategorias chave e imprime no terminal
for subcategoria in SUBCATEGORIAS_CHAVE:
    print("\n" + "-"*70)
    print(f"--- RANKINGS DETALHADOS PARA A SUBCATEGORIA: {subcategoria} ---")
    print("-"*70)

    # Verifica se a subcategoria existe nos dados antes de tentar acessá-la
    if subcategoria in preco_mediano_bairro_subcat_full.columns and subcategoria in
preco_mediano_rede_subcat_full.columns:
        # Ranking de Bairros
        ranking_bairro = preco_mediano_bairro_subcat_full[subcategoria].dropna().sort_values(ascending=False)
        print(f"\n[ Ranking de Bairros para '{subcategoria}' ]")

```

```

print("\nTop 5 Bairros MAIS CAROS:")
print(ranking_bairro.head(5).round(2).reset_index(name='preco_mediano_R$'))
print("\nTop 5 Bairros MAIS BARATOS:")
print(ranking_bairro.tail(5).sort_values().round(2).reset_index(name='preco_mediano_R$'))

# Ranking de Redes
ranking_rede = preco_mediano_rede_subcat_full[subcategoria].dropna().sort_values(ascending=False)
print(f"\n[ Ranking de Redes para '{subcategoria}' ]")
print("\nTop 5 Redes MAIS CARAS:")
print(ranking_rede.head(5).round(2).reset_index(name='preco_mediano_R$'))
print("\nTop 5 Redes MAIS BARATAS:")
print(ranking_rede.tail(5).sort_values().round(2).reset_index(name='preco_mediano_R$'))
else:
    print(f"AVISO: Não há dados suficientes para gerar rankings para a subcategoria '{subcategoria}'.")

print("\n-----")
print("✅ ANÁLISE COMPLETA E DETALHADA CONCLUÍDA! ✅")
print(f"Todos os arquivos foram salvos na pasta '{pasta_resultados}'.")
print("-----")

```

APÊNDICE 5 - SCRIPT DE ANÁLISE DESCRITIVA E VISUALIZAÇÃO DE PREÇOS

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# =====
# 1. CARREGAMENTO DO DATASET
# =====
caminho_arquivo = r'D:\resultados_tcc\3_tabela_matriz_competitividade.csv'

# Carrega o arquivo
df = pd.read_csv(caminho_arquivo, sep=';', encoding='utf-8-sig')

# =====
# 2. ESTATÍSTICAS DESCRITIVAS
# =====
print("=== ESTATÍSTICAS DESCRITIVAS DA VARIÁVEL 'preco_regular' ===")
estatisticas = df['preco_regular'].describe()
print(estatisticas)

# =====
# 3. PLOTAGEM DO BOXPLOT
# =====
plt.figure(figsize=(10, 1.5))
sns.boxplot(x=df['preco_regular'], color='skyblue')
plt.title('Boxplot dos Preços Regulares Observados')
plt.xlabel('Preço (R$)')
plt.grid(True, axis='x', linestyle='--', linewidth=0.5)
plt.tight_layout()
plt.savefig('boxplot_preco_regular.png', dpi=300)
plt.show()
```

APÊNDICE 6 - HEATMAP DE COMPETITIVIDADE DE PREÇO MEDIANO POR REDE E SUBCATEGORIA



APÊNDICE 7 - EVOLUÇÃO DO PREÇO MEDIANO DAS PRINCIPAIS SUBCATEGORIAS

Evolução do Preço Mediano das Principais Subcategorias (Curitiba - 2024)

