UNIVERSIDADE DE SÃO PAULO
ESCOLA POLITÉCNICA
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO

MARCELA VON BORSTEL OKUYAMA

APLICAÇÃO DE ANÁLISE DE AGRUPAMENTOS EM ESTABELECIMENTOS COMERCIAIS PARA UMA EMPRESA DE VAREJO

São Paulo
2024

Marcela von Borstel Okuyama

APLICAÇÃO DE ANÁLISE DE AGRUPAMENTOS EM ESTABELECIMENTOS COMERCIAIS PARA UMA EMPRESA DE VAREJO

Trabalho de Formatura apresentado à Escola Escola Politécnica da Universidade de São Paulo para obtenção do Diploma de Engenheira de Produção

Orientadora: Profa. Dra. Linda Lee Ho

SÃO PAULO
2024

Catalogação-na-publicação

Okuyama, Marcela von Borstel
Aplicação de Análise de Agrupamentos em Estabelecimentos Comerciais para uma Empresa de Varejo / M. B. Okuyama – São Paulo, 2024.
95p.

Trabalho de Formatura - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Produção.

1. Análise de agrupamentos 2. Segmentação de estabelecimentos comerciais 3.K-médias 4. Aprendizado não supervisionado I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Produção II.t.

AGRADECIMENTOS

Aos cidadãos de São Paulo, por contribuir para a manutenção da Universidade de São Paulo e, assim, permitir meu acesso ao ensino superior público, gratuito e de qualidade. Espero que eu possa retribuir o investimento de vocês.

À minha família, pelo suporte e incentivo durante meu desenvolvimento.

Aos meu amigos da Poli, por todas as experiências, companherismo e aprendizados durante esses anos. Carrego um pouco de cada um comigo.

À Poli Social, pelo profundo impacto na minha formação como pessoa. Espero continuar contribuindo para a mudança que quero ver no mundo.

À Cris e ao Osni, por serem meu porto seguro nessa jornada. Obrigada por me acolherem.

Ao André Carneiro, por compartilhar a vida comigo. Sem você, nada disso seria possível.

À Profa. Dra. Linda Lee Ho, pela orientação e apoio na execução deste trabalho. Obrigada por todo aprendizado.

Ao meu pai, por ser meu maior exemplo. Gostaria que você estivesse aqui.

RESUMO

OKUYAMA, Marcela von Borstel. Aplicação de Análise de Agrupamentos em Estabelecimentos Comerciais para uma Empresa de Varejo. 2024. Trabalho de formatura (Engenharia de Produção) - Escola Politécnica, Universidade de São Paulo, São Paulo, 2024.

O varejo é um setor altamente competitivo e dinâmico, em que decisões baseadas em dados podem representar vantagens estratégicas. Este trabalho investiga a aplicação de análise de agrupamentos para segmentar estabelecimentos comerciais que utilizam as plataformas digitais de um grande fabricante de bebidas. A segmentação tem como objetivo a caracterização de perfis de estabelecimentos para aprimorar as recomendações de produtos, ações de marketing, fidelização e atendimento. A metodologia CRISP-DM orientou o processo de mineração de dados que utilizou metodologias de aprendizado não supervisionado para definir e avaliar os grupos formados. Os resultados revelaram quatro perfis distintos de comportamento de compra e venda de bebidas alcoólicas entre os estabelecimentos, permitindo identificar oportunidades para intervenções específicas. Conclui-se que a segmentação oferece *insights* estratégicos para a personalização de ações, contribuindo para o fortalecimento da competitividade da empresa.

Palavras-chave: análise de agrupamentos; segmentação de estabelecimentos comerciais; k-médias; aprendizado não supervisionado.

ABSTRACT

OKUYAMA, Marcela von Borstel. Application of Cluster Analysis in Commercial Establishments for a Retail Company. 2024. Graduation project (Industrial Engineering) - Polytechnic School, University of Sao Paulo, Sao Paulo, 2024.

The retail sector is highly competitive and dynamic, where data-driven decisions can confer strategic advantages. This study investigates the application of cluster analysis to segment commercial establishments that utilize the digital platforms of a major beverage manufacturer. The segmentation aims to characterize establishment profiles to enhance product recommendations, marketing initiatives, customer loyalty, and service delivery. The CRISP-DM methodology guided the data mining process, employing unsupervised learning techniques to define and evaluate the resulting groups. The findings revealed four distinct profiles of alcoholic beverage purchase and sales behavior among establishments, enabling the identification of opportunities for targeted interventions. The study concludes that segmentation provides strategic insights for the personalization of actions, thereby contributing to the company's competitive strength.

Keywords: cluster analysis; commercial establishment segmentation; k-means; unsupervised learning.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

Sumário

Capítulo 1

INTRODUÇÃO

O Capítulo 1 deste trabalho oferece uma visão geral da temática e dos objetivos centrais abordados ao longo do trabalho. Primeiramente, apresenta-se a Corporação S.A., a empresa focal desta análise, destacando suas operações no mercado e a relevância das plataformas Corp Delivery e Corp App para a estratégia de negócios (Seção 1.1). Em seguida, são detalhados os motivos e objetivos que sustentam a importância da análise de agrupamentos aplicada aos estabelecimentos comerciais que utilizam essas plataformas (Seção 1.2). Este capítulo culmina com a estrutura do trabalho (Seção 1.3), que orienta o leitor sobre a organização e os temas específicos abordados nas demais seções.

1.1 A Empresa

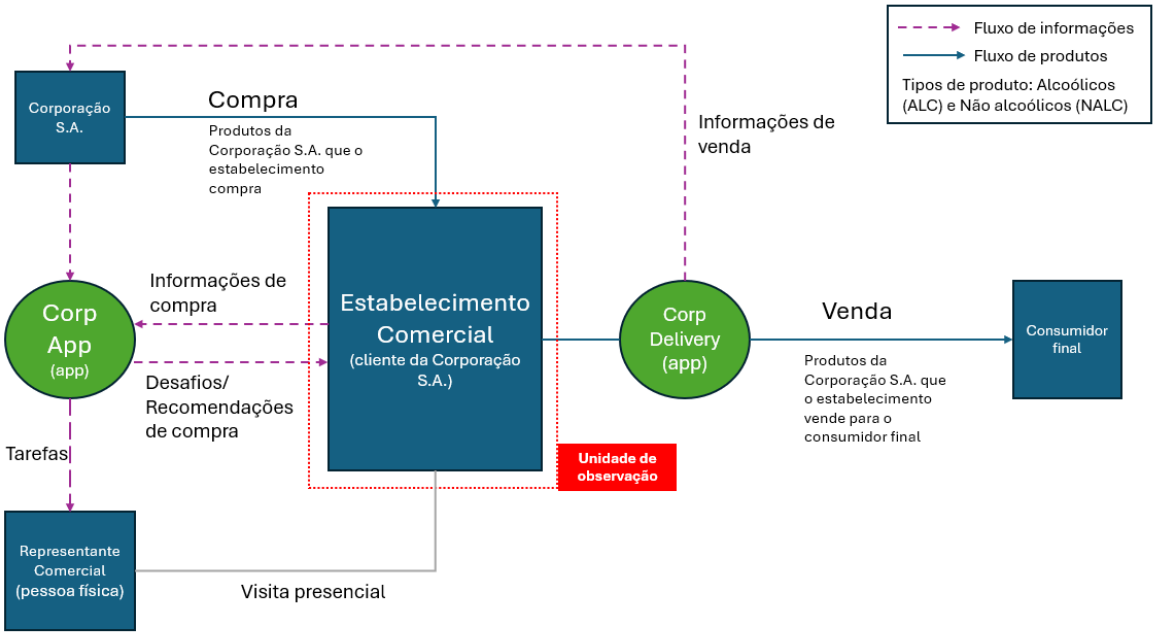
A empresa tratada neste trabalho será denominada como "Corporação S.A." por motivos de confidencialidade. Ela é a maior cervejaria da América Latina em termos de volume de vendas e uma das maiores fabricantes de cerveja do mundo, responsável por fabricar, distribuir e comercializar cerveja, refrigerantes e outras bebidas em 18 países. Empresa de capital aberto na Bolsa de Valores, apresentou um lucro líquido de quase 15 bilhões de reais em 2023. Com aproximadamente um milhão de pontos de venda no Brasil, além do portfólio de produtos (como cerveja, chopes, bebidas mistas, refrigerantes, sucos, isotônicos, energéticos, águas e chás), a Corporação S.A. possui plataformas digitais que conectam negócios e consumidores.

A plataforma de entrega direta ao consumidor, Corp Delivery (nome fictício), conecta estabelecimentos comerciais cadastrados na plataforma aos consumidores para entrega de bebidas geladas e outros produtos, tanto da Corporação S.A. quanto de outras

empresas concorrentes. O Corp Delivery está presente em mais de 700 cidades em todos os 26 estados brasileiros e Distrito Federal e, em 2023, entregou mais de 60 milhões de pedidos, atingindo aproximadamente 6 milhões de usuários ativos mensais.

Sua outra plataforma digital, o Corp App (nome fictício), funciona como um *marketplace* e B2B (*Business-to-Business*) no qual os estabelecimentos comerciais varejistas (clientes da Corporação S.A.) têm acesso a mais de 650 Unidades de Manutenção de Estoque ("SKUs") da empresa e de outras marcas. Ainda, é por meio do Corp App que a organização oferece sugestões de produtos com base no perfil dos clientes e relevância do produto, fornece melhorias no rastreamento de pedidos e de suporte em tempo real. Os dados coletados durante a interação do cliente com a plataforma colabora diretamente com a estratégia de inovação e maior complexidade de portfólio. Além disso, é também por meio dela que acontece a interação das equipes da Corporação S.A. com os representantes de negócio, funcionários que se concentram em ajudar os clientes a melhorar seu desempenho de vendas. Figura 1.1 ilustra o fluxo de produtos e informações na integração entre os agentes e plataformas da Corporação S.A..

Figura 1.1: **Agentes e integração entre plataformas digitais da Corporação S.A.**



Fonte: Elaboração própria

1.2 Motivação e Objetivos

O varejo é um setor altamente competitivo e dinâmico, em que a eficiência na gestão de estabelecimentos comerciais pode representar uma vantagem estratégica. Na Corporação S.A., maior cervejaria da América Latina e uma das líderes globais no mercado de bebidas, a complexidade operacional é ampliada pelo vasto portfólio de produtos e pela diversidade de canais de venda, que incluem plataformas digitais como o Corp Delivery e o Corp App.

Este trabalho se propõe a segmentar estabelecimentos comerciais (cerca de três mil) que utilizam simultaneamente as plataformas Corp App e Corp Delivery da Corporação S.A.. A relevância desta segmentação se dá pelo fato de que tais estabelecimentos fornecem uma visão integrada das dinâmicas de compra e venda de produtos, o que permite à empresa um entendimento mais apurado de seus clientes e a adoção de estratégias mais eficazes. Os resultados deste trabalho serão utilizados pela equipe encarregada dos algoritmos do Corp App, responsáveis pelas tarefas enviadas aos representantes de vendas e pelos desafios, promoções e recomendações de compra para os estabelecimentos, conforme ilustrado na Figura 1.1.

Dentro deste contexto, a análise dos dados provenientes das duas plataformas permitirá o alcance de quatro objetivos específicos, cada um com uma importância estratégica distinta para a equipe da Corporação S.A.:

1. **Caracterização de Perfis**: A identificação e caracterização de perfis distintos de estabelecimentos comerciais é fundamental para compreender o comportamento desses pontos de venda. Estabelecimentos que apresentam volumes elevados de vendas, mas que realizam poucas compras via Corp App, por exemplo, indicam a necessidade de intervenções estratégicas. Nessas situações, a Corporação S.A. pode adotar medidas como a oferta de preços mais competitivos ou o reforço da proposta de valor do Corp App, além de mobilizar seus representantes de vendas para engajar diretamente esses clientes, visando aumentar suas compras e, consequentemente, sua participação no portfólio da empresa.

2. **Identificação de Anomalias**: A identificação de anomalias nos padrões de compra e venda é outro aspecto abordado neste trabalho. Estabelecimentos que, por exemplo, apresentam um volume elevado de vendas no Corp Delivery, mas que não

realizam compras pelo Corp App, podem estar praticando ações comerciais atípicas, como o uso de múltiplos CNPJs. Detectar tais anomalias é uma forma de garantir a conformidade operacional e a assertividade dos dados.

3. **Personalização de Algoritmos para Ações Mais Assertivas**: A compreensão dos diferentes perfis de estabelecimentos comerciais possibilita a personalização de algoritmos que orientem ações de marketing e vendas mais assertivas. Para grandes estabelecimentos com portfólios diversificados, é possível incentivar a compra de produtos em maiores volumes, otimizar estoques e maximizar a eficiência operacional. Em contrapartida, estabelecimentos menores podem ser encorajados a diversificar seus portfólios com produtos incrementais, aumentar sua atratividade e competitividade no mercado, e melhorar sua capacidade de atender à demanda gerada pelo Corp Delivery.

4. **Estratificação de Amostras para Experimentos Aleatorizados**: A segmentação dos estabelecimentos permite também a realização de experimentos aleatorizados, que são essenciais para a inovação e o aprimoramento das estratégias comerciais. Através de modelos de agrupamento, cada segmento pode ser subdividido para receber diferentes tipos de intervenções,o que possibilita à Corporação S.A. identificar quais estratégias geram maior adesão e impacto em diferentes grupos de clientes. Essa abordagem experimental é amplamente utilizada na área para validar hipóteses e direcionar decisões baseadas em dados.

Em suma, o desenvolvimento de um modelo de agrupamento de estabelecimentos comerciais que utiliza as plataformas digitais da Corporação S.A. representa uma ferramenta estratégica para a empresa. Não apenas permite a otimização de operações e o aperfeiçoamento do relacionamento com os clientes, mas também oferece uma base para a realização de experimentos e intervenções baseados em evidências, o que promove a inovação e a competitividade da empresa no mercado de bebidas.

1.3 Estrutura do Trabalho

A estrutura deste trabalho é organizada em cinco capítulos, cada um destinado a explorar aspectos específicos da análise de agrupamentos aplicada a estabelecimentos comerciais. No Capítulo 1, é realizada uma introdução ao tema, contextualizando a

empresa, os objetivos e a justificativa do estudo. O Capítulo 2, Revisão Bibliográfica, discute as bases teóricas e metodológicas que sustentam a análise de agrupamentos e o processo CRISP-DM. O Capítulo 3, Metodologia, descreve os dados, o pré-processamento e as técnicas de análise adotadas para segmentar os pontos de venda. Em Resultados da Análise de Agrupamentos (Capítulo 4), são apresentados os grupos identificados, seguidos de uma interpretação dos perfis e recomendações estratégicas. Finalmente, o Capítulo 5 traz a Conclusão e Próximos Passos, com as considerações finais e sugestões para futuras investigações.

Capítulo 2

REVISÃO BIBLIOGRÁFICA

Neste capítulo, serão explorados os conceitos e metodologias que sustentam a análise de dados e o aprendizado de máquina, conforme delineado nas seções subsequentes. Um modelo denominado *Cross-Industry Standard Process for Data Mining* (CRISP-DM) é o assunto da Seção 2.1, que orienta o processo de mineração de dados através de suas fases estruturadas, desde a compreensão do problema até a implementação dos resultados. Em seguida, será abordada a Análise Exploratória de Dados (AED), destacando sua importância na preparação e análise de dados. A Seção 2.2, dedicada à análise dos dados, aborda técnicas de limpeza e transformação, além da análise exploratória de dados que busca identificar padrões e tendências relevantes.

Após estabelecer as bases da AED, o capítulo discutirá o Aprendizado de Máquina (Seção 2.3), que se tornou uma ferramenta relevante para a análise de dados complexos. Dentro dessa Seção, serão explorados os Modelos de Aprendizado Não Supervisionado, que permitem a identificação de estruturas subjacentes nos dados sem a necessidade de rótulos predefinidos, caso que se aplica ao problema tratado neste trabalho. Para esse caso, será abordada a Análise de Agrupamentos (Seção 2.3.1), uma técnica para segmentar dados em grupos significativos, e, por fim, os índices de valiação de agrupamentos para avaliar a qualidade dos modelos de agrupamentos utilizados. Este capítulo proporcionará uma base teórica que guiará a compreensão e aplicação das técnicas discutidas nas fases posteriores deste trabalho.

2.1 CRISP-DM (*Cross-Industry Standard Process for Data Mining*)

O CRISP-DM (*Cross-Industry Standard Process for Data Mining*) é uma metodologia amplamente adotada no campo da mineração de dados, projetada para fornecer um processo estruturado para planejar, executar e gerenciar projetos de mineração de dados, conforme disposto por Chapman (2000) [1]. Desenvolvido na década de 1990 por um consórcio empresarial e o Instituto de Pesquisa Fraunhofer, o CRISP-DM é constituído por seis fases inter-relacionadas, que são iterativas e podem ser revisadas conforme necessário ao longo do projeto: Compreensão do Negócio (*Business Understanding*), Compreensão dos Dados (*Data Understanding*), Preparação dos Dados (*Data Preparation*), Modelagem (*Modeling*), Avaliação (*Evaluation*) e Implantação (*Deployment*) (CHAPMAN, 2000 [1]).

Na etapa de **Compreensão do Negócio**, o objetivo é compreender o contexto empresarial e identificar problemas ou oportunidades que podem ser abordados com a mineração de dados, com foco em traduzir objetivos e restrições do negócio em problemas de mineração de dados. As atividades desta fase incluem a definição dos objetivos do negócio, de forma a identificar as metas e objetivos gerais do projeto, bem como os resultados esperados. Realiza-se uma avaliação da situação, que pode envolver uma análise SWOT (forças, fraquezas, oportunidades e ameaças), para entender o contexto atual. Em seguida, determinam-se os objetivos de mineração de dados, para isso traduz-se os objetivos do negócio em questões específicas, e elabora-se um plano de projeto detalhado que descreva as fases do projeto, cronogramas, recursos necessários e responsabilidades (CHAPMAN, 2000 [1]).

A fase de **Compreensão dos Dados** visa familiarizar-se com os dados, identificar problemas de qualidade e descobrir percepções iniciais. As atividades incluem a coleta de dados iniciais, para obter os dados necessários para a análise, provenientes de diversas fontes. Em seguida, descrevem-se os dados, documentam-se suas características básicas, como número de registros e tipos de dados. Realiza-se uma exploração dos dados para identificar padrões, anomalias e hipóteses iniciais. Por fim, verifica-se a qualidade dos dados, em termos de completude, precisão e ausência de valores duplicados ou anômalos (CHAPMAN, 2000 [1]).

A **Preparação dos Dados** é uma das etapas mais trabalhosas e cruciais do CRISP-DM, que envolve a transformação dos dados brutos em um formato adequado

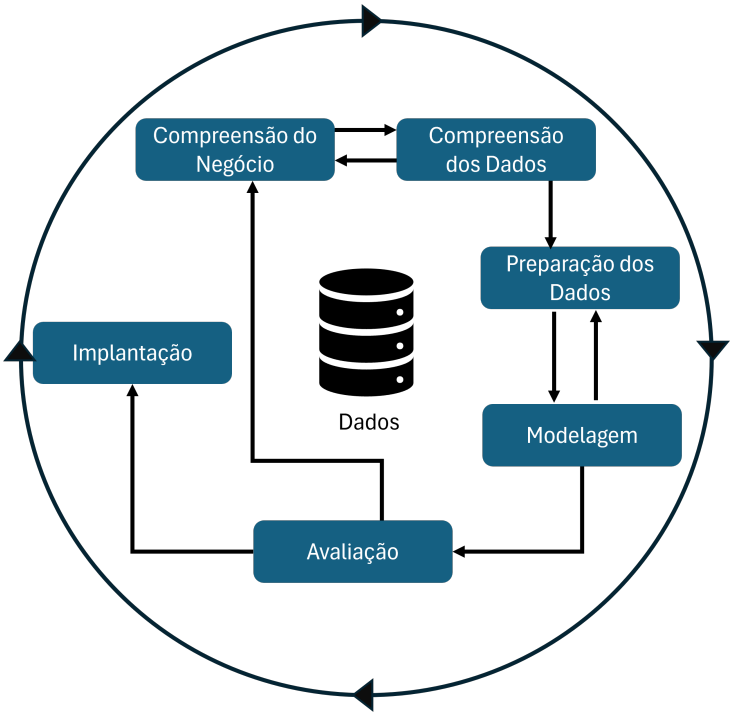
para análise, que melhore sua qualidade e integridade. As atividades principais são a seleção dos dados relevantes para a análise, com base em critérios específicos, e a limpeza dos dados, com a correção ou remoção de dados incompletos, duplicados ou inconsistentes. Também se constroem novos dados, a partir da criação de novas variáveis ou atributos úteis na modelagem. Além disso, integram-se dados provenientes de diferentes fontes para criar um conjunto coeso e formatam-se os dados para atender aos requisitos dos algoritmos de modelagem (CHAPMAN, 2000 [1]).

Na etapa de **Modelagem**, o objetivo é aplicar técnicas de mineração de dados para construir modelos que atendam aos objetivos do projeto. As atividades incluem a seleção das técnicas de modelagem apropriadas, como regressão, árvores de decisão e redes neurais, e a definição de um plano de teste para avaliar a qualidade e robustez dos modelos. Em seguida, constroem-se os modelos aplicando os algoritmos aos dados preparados e avaliam-se os modelos com base em métricas de desempenho, ajustam-se os parâmetros conforme necessário (CHAPMAN, 2000 [1]).

A fase de **Avaliação** visa avaliar a eficácia dos modelos em relação aos objetivos de negócio e verificar se todos os requisitos foram atendidos. As atividades incluem a avaliação dos resultados dos modelos, comparando-os com os objetivos de negócio para verificar se são úteis e aplicáveis. Em seguida, revisa-se o processo de modelagem para identificar possíveis melhorias ou ajustes e determina-se a prontidão dos modelos para implantação no ambiente de produção, ao consider todos os aspectos do negócio (CHAPMAN, 2000 [1]).

Na **Implantação**, o objetivo é implementar os modelos no ambiente de produção e assegurar que eles gerem valor contínuo para o negócio. As atividades principais incluem o planejamento da implantação, a definição dos passos necessários, cronogramas e recursos, e a implementação dos modelos, com a integração aos sistemas de produção e treinamento dos usuários finais. Além disso, estabelece-se um sistema para monitorar o desempenho dos modelos e realizar manutenções periódicas. Os resultados e percepções obtidos são documentados e comunicados às partes interessadas, e realiza-se uma revisão e documentação do projeto, com o registro das lições aprendidas e recomendações para futuros projetos (CHAPMAN, 2000 [1]).

Figura 2.1: **O processo de mineração de dados CRISP-DM.**



Fonte: adaptado de Chapman (2000) [1]

É importante notar que o processo CRISP-DM não é linear, mas iterativo, como mostra a Figura 2.1. Ao longo de todas as fases, pode ser necessário revisar etapas anteriores à medida que novas informações e percepções são descobertas. Essa abordagem flexível assegura que o projeto permaneça alinhado com os objetivos de negócio e possa adaptar-se a mudanças e novos desafios. A metodologia CRISP-DM é altamente valorizada por sua estrutura clara e prática, permite uma execução eficiente e eficaz de projetos de mineração de dados, desde a concepção inicial até a implementação e manutenção contínua dos modelos desenvolvidos (CHAPMAN, 2000 [1]).

2.2 Preparação de Dados

Conforme Morettin e Singer (2021) [2], dados são coletados com o objetivo de obter informações e, comumente, envolvem valores de várias variáveis obtidos da observação de unidades de investigação (como indivíduos, escolas, cidades etc.) que constituem uma amostra de uma população. Para os autores, a Análise Exploratória de Dados (AED) consiste em um conjunto de técnicas e processos destinados a investigar e resumir as principais características de um conjunto de dados. O objetivo principal da AED é, portanto,

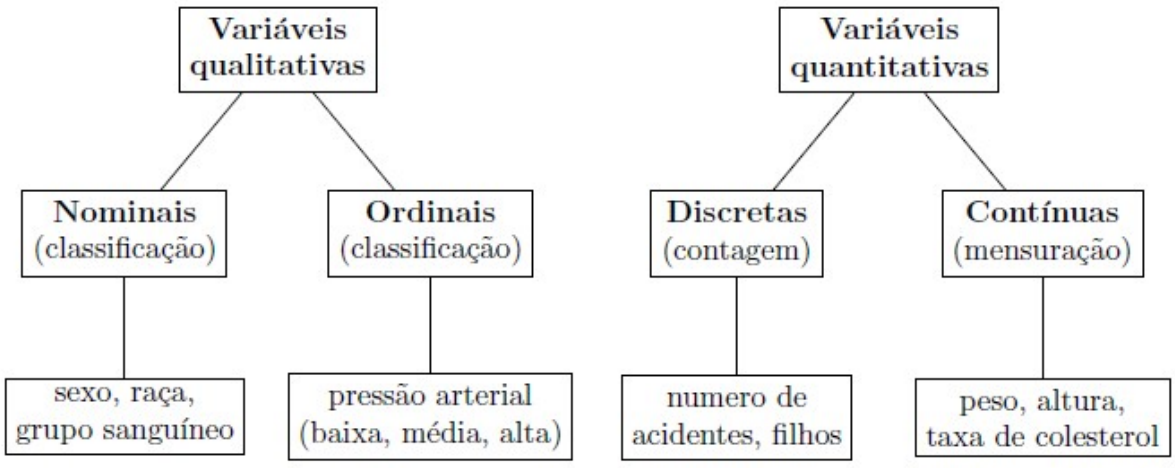
proporcionar uma compreensão inicial dos dados antes da aplicação de métodos analíticos ou de modelagem mais complexos.

Assim, a AED é uma etapa que permite aos analistas identificar padrões, detectar anomalias, testar hipóteses e verificar suposições por meio de estatísticas descritivas e técnicas gráficas. A partir dessa análise, pode-se elaborar uma estratégia apropriada de modelagem estatística ou de aprendizado de máquina, além de detectar *outliers* ou inconsistências nos dados.

2.2.1 Tipo de Dados

Os tipos de variáveis que compõem o conjunto de dados influenciam na escolha das técnicas empregadas na análise descritiva. A classificação empregada por Morettin e Singer (2021) [2] está representada na Figura 2.2.

Figura 2.2: **Classificação de variáveis**



Fonte: Morettin e Singer (2021) [2]

As variáveis qualitativas representam características não numéricas de uma unidade de investigação, como o gênero, por exemplo. Elas podem ser ordinais, quando existe uma hierarquia entre as categorias (como o tamanho de uma escola: pequeno, médio ou grande), ou nominais, quando não há uma ordem estabelecida (como a localização de uma empresa: norte, sul, leste ou oeste).

Já as variáveis quantitativas correspondem a valores numéricos associados à uni-

dade de investigação, como o peso, por exemplo. Essas variáveis podem ser discretas, quando apresentam valores dentro dos números naturais (como o número de produtos em uma loja), ou contínuas, quando podem ter valores dentro dos números reais, como o tempo que um atleta demora para correr 100 metros, por exemplo (MORETTIN E SINGER, 2021 [2]).

2.2.2 Limpeza de Dados

Conjuntos de dados podem apresentar problemas relacionados à qualidade dos dados. Os casos mais comuns desses problemas incluem dados inconsistentes ou incompletos (FACELI et al., 2011 [3]). A presença dessas deficiências em um conjunto de dados pode resultar em análises incorretas.

No caso de **dados incompletos**, que se referem à ausência de valores em algumas variávis de certos objetos, Faceli et al. (2011) [3] apontam que diversas estratégias foram propostas para lidar com essas variáveis, dentre as quais se destacam:

- Remover os objetos que contêm valores ausentes. Essa estratégia é geralmente aplicada quando uma das variáveis ausentes em um objeto é aquele que define sua classe. No entanto, essa abordagem não é recomendada quando poucas variáveis estão faltando, quando a quantidade de variáveis ausentes varia significativamente entre os objetos afetados ou quando a quantidade de objetos restantes for pequena.

- Definir e preencher manualmente os valores para as variáveis ausentes. Essa opção não é prática quando há um número elevado de objetos ou variáveis com dados faltantes.

- Utilizar um método ou heurística para determinar automaticamente os valores para as variáveis com dados ausentes, sendo essa a opção mais utilizada. Diferentes abordagens podem ser aplicadas, como: atribuir um novo valor que indique que a variável tem um valor desconhecido; utilizar a média, moda ou mediana dos valores conhecidos dessa variável; ou empregar um algoritmo indutor para estimar o valor da variável, como, por exemplo, o *Multivariate Imputation by Chained Equations* (MICE) [4].

Por outro lado, os **dados inconsistentes**, que não se alinham ou contradizem os valores de outras variáveis do mesmo objeto, podem ser detectados quando as relações

conhecidas entre as variáveis são violadas ou por meio de técnicas de identificação de ruído. Segundo os autores, algoritmos simples podem verificar automaticamente se as relações existentes entre as variáveis foram comprometidas; ou, quando o conjunto de dados é pequeno, dados inconsistentes podem ser eliminados manualmente.

2.2.3 Exploração dos Dados

Uma vasta quantidade de informações relevantes pode ser obtida de um conjunto de dados por meio de sua análise ou exploração, o que contribui, por exemplo, para a escolha da técnica mais adequada de pré-processamento e aprendizado. Um dos métodos de explorar um conjunto de dados é através da extração de métricas pertencentes à área da estatística descritiva, que sintetiza quantitativamente as principais características do conjunto de dados (FACELLI et al.,2011 [3]).

As técnicas aplicáveis dependem do tipo de variáveis envolvidas (MORETTIN E SINGER, 2021 [2]). Para variáveis qualitativas, é possível utilizar as frequências (absolutas e/ou relativas) das unidades analisadas em cada categoria. A frequência absoluta indica o número de unidades em cada categoria, enquanto a frequência relativa expressa a porcentagem correspondente. Essas distribuições de frequência podem ser visualizadas por meio de gráficos de barras ou diagramas circular.

Já para variáveis quantitativas, em muitas situações deseja-se fazer um resumo dos resultados através de poucos números mais detalhado do que apenas a distribuição de frequências em determinadas classes e, nesses casos, podem-se considerar as medidas de posição (localização ou de tendência cetral), as medidas de dispersão e medidas de forma, entre outras (MORETTIN E SINGER, 2021 [2]).

A **média aritmética** é uma medida de tendência central que representa o valor médio de um conjunto de dados. Outra medida de posição, a **mediana** é o valor que divide o conjunto de dados ordenados em duas partes iguais. Ela não é afetada por valores extremos e é calculada de forma diferente dependendo do número de observações n. Se n é ímpar, a mediana é o valor central, e se n é par, a mediana é a média dos dois valores centrais.

Também muito utilizados como medidas de posição, os **quantis** são valores que dividem um conjunto de dados ordenados em partes iguais. O quantil q_p de ordem p, onde $0 < p < 1$, é o valor tal que $p \times 100\%$ dos dados são menores ou iguais a ele. Quantis

comumente usados são:

- **Quartis**: Dividem os dados em 4 partes iguais. O primeiro quartil (Q_1) é o quantil de 25% ($q_{0.25}$), o segundo quartil (Q_2) é a mediana ($q_{0.50}$) e o terceiro quartil (Q_3) é o quantil de 75% ($q_{0.75}$).
- **Decis**: Dividem os dados em 10 partes iguais.
- **Percentis**: Dividem os dados em 100 partes iguais.

A variância e o desvio padrão são duas medidas de dispersão bastante usadas e referem-se a média de alguma função positiva dos desvios das observações em relação à sua média (MORETTIN e SINGER, 2021 [2]). A **variância** é calculada como a média dos quadrados das diferenças entre cada valor e a média. Entretanto, a unidade de medida da variância é o quadrado da unidade de medida da observação correspondente e, portanto, convém definir outra medida de dispersão que preserve a unidade original. Assim, para manter essa propriedade, utiliza-se a raiz quadrada positiva da variância, conhecida como **desvio padrão**.

Outra medida de dispersão utilizada é a **distância interquartis** ou amplitude interquartis, que é calculada pela diferença entre o quartil 3 e o quartil 1. Ela pode ser utilizada para estimar o desvio padrão de uma distribuição. Além disso, ela pode ser empregada para identificar *outliers*, ou seja, valores que ultrapassem os limites superior e inferior a $Q_3 + 1,5 \times (Q_3 - Q_1)$ e $Q_1 - 1,5 \times (Q_3 - Q_1)$, respectivamente.

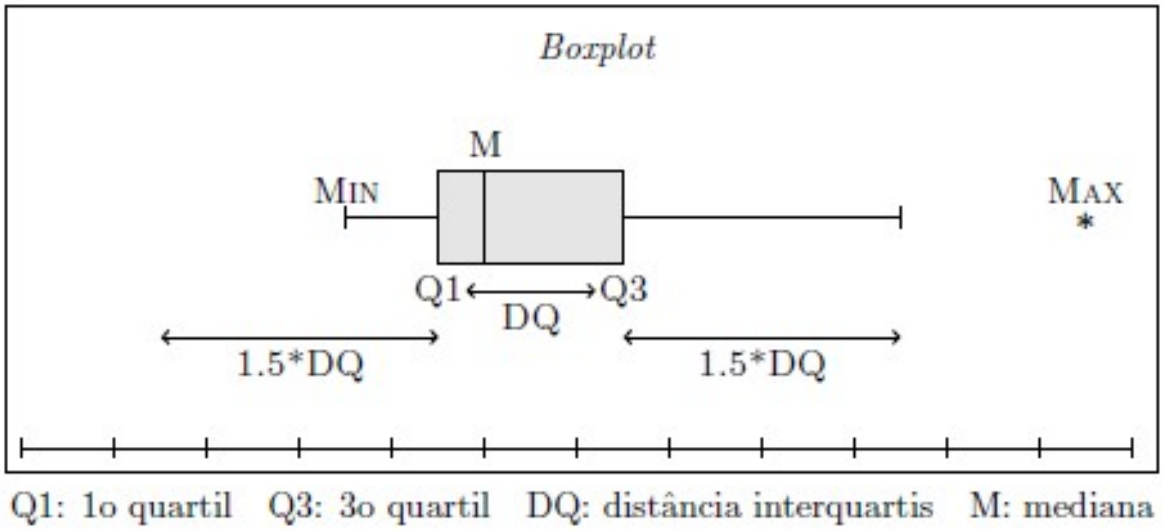
Além das estatísticas descritivas, as representações gráficas complementam a compreensão da distribuição dos dados. Dois gráficos amplamente utilizados para esse propósito são o histograma e o *boxplot*.

O **histograma** é um gráfico construído a partir da distribuição de frequências e é composto de retângulos contíguos que demonstram a distribuição dos dados em intervalos associados a sua base. Essa configuração possibilita a visualização de tendências como a forma da distribuição (simétrica, assimétrica, unimodal, multimodal) e a presença de caudas.

Já o ***boxplot*** é um gráfico que utiliza os quantis como base e serve como uma alternativa ao histograma para resumir a distribuição dos dados. Ele é composto por um retângulo de base definida por Q_1 e Q_3, e por um segmento mais longo entre estes dois números que representa a mediana M. Além disso, dois segmentos traçados a partir de Q_1

e Q_3 dados, com limites determinados por $\min[x_z, Q_3 + 1,5 \times dQ]$ e $\max[x_1, Q_1 - 1,5 \times dQ]$, sendo que $dQ = Q_3 - Q_1$ indica a distância interquartil, x_1 e x_z o menor e o maior valor de x respectivamente. Os valores maiores que o limite superior ou menores que o limite inferior são classificados como valores atípicos ou discrepantes (*outliers*) e são representados por um símbolo, como um asterisco ($*$), por exemplo. Esses elementos estão representados na Figura 2.3.

Figura 2.3: **Detalhes para a construção de *boxplot***



Fonte: Morettin e Singer (2021) [2]

Dados multivariados permitem ainda análises da relação entre dois ou mais atributos ou variáveis. A **correlação** é uma medida estatística que quantifica a força e a direção da relação linear entre duas variáveis, expressa pelo coeficiente de correlação, que varia de -1 a 1. Um coeficiente de 1 indica uma correlação positiva perfeita entre duas variáveis y e x expressa por $y = b \times x$, -1 indica uma correlação negativa perfeita entre duas variáveis dada por $y = -b \times x$, e 0 sugere ausência de relação linear. A correlação pode ser calculada utilizando diversas métricas, como o coeficiente de Pearson, o coeficiente de Spearman e o coeficiente de Kendall, dependendo das características dos dados e do tipo de relação analisada. Uma forma de apresentar a correlação entre cada possível par de variáveis de um conjunto de dados é a **matriz de correlação**. Nesses casos, a matriz terá o valor zero nas posições da diagonal e as correlações nas demais posições (ARTES e BARROSO, 2023 [5]).

2.2.4 Transformação de Variáveis

Diversas metodologias empregadas na inferência estatística são fundamentadas na premissa de que os valores de uma ou mais variáveis de interesse derivam de uma distribuição normal. Entretanto, em várias situações práticas, a distribuição dos dados amostrais pode apresentar assimetria e incluir valores atípicos. Nesses contextos, pode-se considerar uma transformação das observações com o objetivo de obter uma distribuição que seja mais simétrica, aproximando-se, assim, da distribuição normal, conforme indicado por Morettin e Singer (2021) [2]. Segundo os autores, uma tranformação comumente utilizada é:

$$x^{(p)} = \begin{cases} x^p, & \text{se } p > 0 \\ \log(x), & \text{se } p = 0 \\ -x^p, & \text{se } p < 0 \end{cases} \tag{3.21}$$

Essa transformação é adequada para distribuições assimétricas à direita com $0 < p < 1$, uma vez que grandes valores de x decrescem mais em relação à valores pequenos, enquanto valores de $p > 1$ são usados para distribuições assimétricas à esquerda. Normalmente, considera-se uma sequência de valores de p para análise (como a sequência $-3, -2, -1, -\frac{1}{2}, -\frac{1}{3}, -\frac{1}{4}, 0, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, 1, 2, 3$ sugerida por Hinkley (1997) *apud* Morretin e Singer (2021) [2]) e visualização por meio de gráficos, a fim de identificar o valor mais apropriado.

Ainda, conforme indicado por Faceli et al. (2011) [3], a transformação dos valores numéricos de variáveis em outro valor numérico pode ser necessária, especialmente quando há discrepâncias significativas entre os limites inferior e superior das variáveis ou quando múltiplas variáveis estão em escalas distintas. Essas transformações visam evitar que uma variável predomine sobre as demais. A **normalização dos dados** é uma técnica frequentemente aplicada, sendo recomendada quando os limites de valores de diferentes variáveis variam consideravelmente, como em casos em que as unidades estão em diferentes escalas.

Segundo os autores, existem duas abordagens principais para a normalização: por amplitude e por distribuição. A normalização por amplitude se divide em reescala e padronização. Na **normalização por reescala**, ou normalização min-max, os dados

são transformados para que todos os valores fiquem dentro de um intervalo específico, geralmente entre 0 e 1. Essa normalização é útil quando é necessário que os dados estejam em uma escala comum. A fórmula para a normalização por reescala é dada por:

$$\mathbf{x}'_i = New_{min} + \frac{\mathbf{x}_i - \min(\mathbf{X})}{\max(\mathbf{X}) - \min(\mathbf{X})} \times (New_{max} - New_{min}) \tag{2.1}$$

onde:

- $\mathbf{x}'_i$ é a observação i normalizada.
- $\mathbf{x}_i$ é a observação i original.
- $\min(\mathbf{X})$ é o valor mínimo do conjunto de dados $\mathbf{X}$.
- $\max(\mathbf{X})$ é o valor máximo do conjunto de dados $\mathbf{X}$.
- New_{min} é o valor mínimo da nova escala.
- New_{max} é o valor máximo da nova escala.

Para a **normalização por padronização**, é subtraída uma medida de posição a cada valor da variável a ser normalizada e o resultado é, posteriormente, dividido medida de dispersão. Com isso, diferentes variáveis podem possuir limites inferiores e superiores distintos, mas terão valores equivalentes para as medidas de posição e dispersão. Se as medidas de posição e de dispersão forem a média amostral ($\overline{x}$) e o desvio padrão amostral (s), respectivamente, os valores de uma variável são transformados em um novo conjunto com média igual a 0 e variância igual a 1, pela fórmula:

$$\mathbf{x}'_i = \frac{\mathbf{x}_i - \overline{\mathbf{x}}}{s} \tag{2.2}$$

Geralmente, a padronização é preferida à reescala, uma vez que lida melhor com outliers (FACELI et al., 2011 [3]).

Já a **normalização por distribuição** altera a escala dos valores de uma variável, conforme indicado pelos autores. Um exemplo é substituição de cada valor pela sua posição relativa no ranking (após a ordenação dos valores), o que resulta em uma distribuição uniforme quando todos os valores são distintos.

2.3 Aprendizado de Máquina

No campo do aprendizado estatístico, segundo James et al. (2013) [6], os problemas podem ser classificados principalmente em duas categorias: aprendizado supervisionado e não supervisionado. O **aprendizado supervisionado** envolve a existência de um conjunto de dados onde, para cada observação, há uma correspondência clara entre os preditores e uma variável de resposta . O objetivo principal nesse contexto é construir um modelo que seja capaz de prever com precisão a resposta para novas observações, utilizando métodos como regressão linear, regressão logística e técnicas mais avançadas como máquinas de vetores de suporte e *boosting*. Esses métodos são amplamente utilizados devido à sua capacidade de não apenas prever, mas também inferir relações entre variáveis.

Em contrapartida, o **aprendizado não supervisionado** lida com situações em que não há uma variável de resposta associada aos dados. Nesse cenário, o objetivo é explorar as estruturas subjacentes dos dados para descobrir padrões ou agrupamentos ocultos. A Análise de Agrupamentos (*Cluster Analysis*) é um exemplo típico de uma abordagem de aprendizado não supervisionado, em que o foco está em agrupar as observações em grupos relativamente distintos com base em características compartilhadas (JAMES et al., 2013) [6].

A escolha entre aprendizado supervisionado ou não supervisionado depende da natureza dos dados disponíveis e dos objetivos específicos da análise. Enquanto o aprendizado supervisionado é utilizado devido à sua aplicabilidade direta em previsão e inferência, o aprendizado não supervisionado oferece ferramentas para a exploração de dados quando a estrutura subjacente não é claramente definida. Devido à natureza do problema tratado neste trabalho, em que o objetivo é identificar padrões de comportamento de venda e compra entre os estabelecimentos para diferenciar possíveis grupos, a abordagem de Análise de Agrupamentos será explorada.

2.3.1 Análise de Agrupamentos

Conforme indicado por Morettin e Singer (2021) [2] em seu livro, a Análise de Agrupamentos (*Cluster Analysis*) visa agrupar dados em grupos (*clusters*) com base em uma medida de distância, de modo que as distâncias entre os elementos de um mesmo

grupo sejam minimizadas e distância entre grupos maximizada. Essa técnica é frequentemente referida como segmentação de dados e é utilizada para identificar padrões e estruturas nos dados ao reunir informações semelhantes em um espaço específico. A aplicação de uma análise de agrupamentos segue algumas etapas, segundo Artes e Barroso (2023) [5]:

- Definição do critério de parecença: nessa fase são definidas as variáveis que precisam ou não serem transformadas (conforme abordado na Seção 2.2.4), além do critério que será utilizado na definição dos grupos (por exemplo, distância euclidiana entre as observações);
- Formação dos grupos: nesse momento é escolhido o algoritmo que será empregado na definição dos grupos;
- Escolha do número de grupos: o número de grupos que será utilizado na análise pode ser definido a partir de alguns critérios como conhecimento prévio sobre os dados, conveniência de análise ou ainda pode ser determinado posteriormente baseado nos resultados da análise;
- Interpretação dos grupos: ao final, caracteriza-se os grupos formados a partir de medidas de estatísticas descritivas, por exemplo.

Medidas de Parecença

As medidas de parecença são utilizadas para determinar os critérios que definem se dois pontos estão próximos e, assim, podem fazer parte do mesmo grupo. Dessa forma, elas são fundamentais nos algoritmos de análise de agrupamentos. Essas medidas são divididas em dois grupos: medidas de similaridade, em que quanto maior o valor, maior a semelhança entre os objetos; e medidas de dissimilaridade, em que quanto maior o valor, maior é a diferença entre os objetos (ARTES e BARROSo, 2023 [5]).

Para dados numéricos, as distâncias são as medidas de dissimilaridade mais utilizadas, segundo Artes e Barroso (2023) [5]. Dois métodos de cálculo de distâncias utilizados pelos autores (considera que cada observação é representada por um ponto em um espaço euclidiano) são a distância Euclidiana e a distância Manhattan ou quarteirão (*city block*).

Conforme descrito por Artes e Barroso (2023) [5], seja $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ o vetor de observações do indivíduo $i, i = 1, \ldots, n$, no qual x_{ij} representa o valor assumido pela variável x_j para o indivíduo i, as distâncias entre os indivíduos i e k são dadas por:

- **Distância euclidiana**:

$$d_{ik} = \sqrt{(\mathbf{x}_i - \mathbf{x}_k)^T(\mathbf{x}_i - \mathbf{x}_k)} = \sqrt{\sum_{j=1}^{p}(x_{ij} - x_{kj})^2} \tag{2.3}$$

- **Distância Manhattan** ou quarteirão (*city block*):

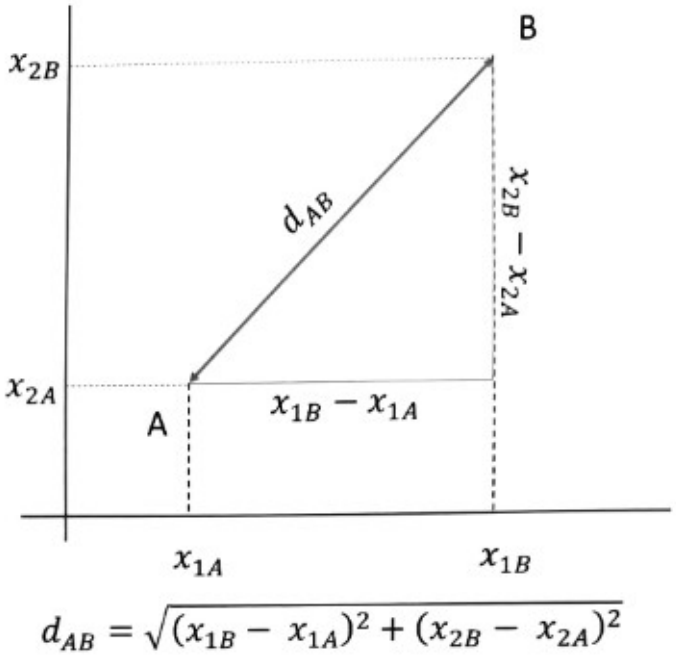
$$d_{ik} = \sum_{j=1}^{p}|x_{ij} - x_{kj}| \tag{2.4}$$

Conforme destacado pelos autores, tanto a distância Euclidiana como a Manhattan são casos particulares da distância de Minkowsky (para m=2 e m=1, respectivamente), definida por:

$$d_{ik} = \left(\sum_{j=1}^{p}|x_{ij} - x_{kj}|^m\right)^{\frac{1}{m}}, m \geq 1 \tag{2.5}$$

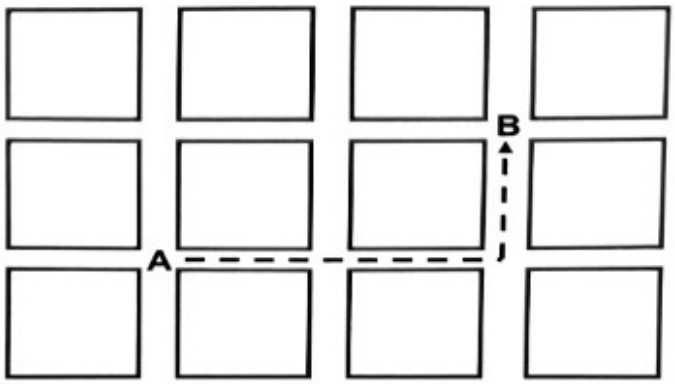
Duas ilustrações do cálculo das distâncias entre dois pontos (A e B) que utiliza a distância euclidiana (Figura 2.4) e a distância Manhattan (Figura 2.5) são apresentadas a seguir.

Figura 2.4: **Distância euclidiana entre os pontos A e B**



Fonte: ARTES e BARROSO (2023) [5]

Figura 2.5: **Distância Manhattan entre os pontos A e B**



Fonte: ARTES e BARROSO (2023) [5]

Quando se trabalha com distâncias em análises de dados, a utilização de variáveis em diferentes escalas pode impactar significativamente os resultados, uma vez que variáveis com escalas maiores podem dominar o cálculo das distâncias, ofuscando a contribuição de variáveis com escalas menores. Isso pode levar a interpretações distorcidas e a agrupamentos inadequados nos dados. Para mitigar esse efeito, a padronização das variáveis é uma prática recomendada, conforme mencionado anteriormente na Seção 2.2.4. Esse

procedimento assegura que cada variável contribua de forma equitativa para a análise, permitindo uma comparação mais equilibrada entre as variáveis.

Estratégias de Agrupamentos

Em seu livro, Artes e Barroso (2023) [5] abordam duas famílias de algoritmos utilizados na formação de agrupamentos: métodos hierárquicos aglomerativos e métodos de partição.

Os **métodos hierárquicos aglomerativos** formam agrupamentos a partir de uma matriz de parecença. O processo começa com a identificação do par de objetos mais semelhantes, que é agrupado e, a partir disso, considerado um único objeto. Em seguida, é criada uma nova matriz de parecença para identificar o par mais semelhante, que formará um novo grupo. Esse processo se repete até que todos os objetos estejam agrupados. A análise do histórico do agrupamento permite inferir o número de grupos existentes nos dados.

Há diversos métodos que seguem esse processo e o que os distingue é a regra para a atualização da matriz de parecença a cada nova união de pares de objetos (ARTES e BARROSO, 2023 [5]). Para ilustrar, com base no exemplo dos autores, considere cinco pontos de dados em um espaço bidimensional: A(1; 2), B(2; 2,5), C(5; 5), D(10; 10) e E(10; 11). A matriz das distâncias euclidianas calculadas entre os pares é dada na Tabela 2.1.

Tabela 2.1: **Matriz de distâncias**

	A	B	C	D	E
A	0				
B	0,50	0			
C	5,00	4,72	0		
D	12,04	11,72	7,07	0	
E	12,73	12,38	7,81	1,00	0

Fonte: Elaboração própria

Como o primeiro passo é identificar os pares mais semelhantes, na tabela observa-se que o par A e B apresenta a menor distância (0,5). Assim, esses dois pontos são

agrupados e esse grupo passa a ser considerado como um novo objeto. A partir dessa nova configuração, deve-se calcular uma nova matriz de distâncias. A forma em que o grupo {A, B} será tratado nos cálculos é o que diferencia os metódos hierárquicos utilizados (ARTES e BARROSO, 2023 [5]).

Entre as alternativas, o **método do vizinho mais próximo** (MVP), ou método de ligação simples, considera a menor distância observada entre um elemento de um grupo G_1 e um elemento de outro grupo G_2. Essa distância é definida como:

$$d[G_1, G_2] = \min_{i \in G_1, k \in G_2} d_{ik}, \tag{2.6}$$

Por outro lado, o **método do vizinho mais distante** (MVD), ou método da ligação completa, utiliza a maior distância observada entre um elemento de G_1 e um de G_2, sendo caracterizado por:

$$d[G_1, G_2] = \max_{i \in G_1, k \in G_2} d_{ik}, \tag{2.7}$$

O **método das médias das distâncias** (MMD) calcula a média das distâncias entre os elementos de G_1 e os de G_2, conforme a seguinte fórmula, em que g_j é a quantidade de objetos no grupo j:

$$d[G_1, G_2] = \sum_{i \in G_1} \sum_{k \in G_2} \frac{d_{ik}}{g_1 g_2} \tag{2.8}$$

Já o **método do centroide** (MC) estabelece que a coordenada de cada grupo é a média das coordenadas de seus objetos. A coordenada obtida é denominada centróide. A distância entre os grupos é calculada utilizando a média das coordenadas, definida como:

$$d[G_1, G_2] = d[\mathbf{c}_1, \mathbf{c}_2]; \text{ com } \mathbf{c}_j = \sum_{i \in G_j} \frac{\mathbf{x}_i}{g_j}, \quad j = 1, 2. \tag{2.9}$$

Para o exemplo proposto, considerando o método do vizinho mais próximo, calcula-se a nova matriz de distâncias, apresentada na Tabela 2.2.

Tabela 2.2: **Matriz de distâncias após o primeiro agrupamento**

	AB	C	D	E
AB	0			
C	4,72	0		
D	12,04	7,07	0	
E	12,73	7,81	1,00	0

Fonte: Elaboração própria

Em que a distância entre A, B e C foi definida como:

1. Distância entre A e C:

$$d(A,C) = \sqrt{(5-1)^2 + (5-2)^2} = \sqrt{(4)^2 + (3)^2} = \sqrt{16+9} = \sqrt{25} = 5$$

2. Distância entre B e C:

$$d(B,C) = \sqrt{(5-1)^2 + (5-2.5)^2} = \sqrt{(4)^2 + (2,5)^2} = \sqrt{16+6,25} = \sqrt{22,25} \approx 4,72$$

3. A distância entre o grupo {A, B} e C é dada pela menor distância calculada:

$$d(\{A,B\},C) = \min(d(A,C), d(B,C)) = \min(5; 4,72) = 4,72$$

Da nova matriz, observa-se que o par de observações mais próximo é entre D e E (1). Assim, eles são agrupados formando o grupo {D, E}. Atualiza-se novamente a matriz de distâncias considerando os grupos {A, B} e {D, E}.

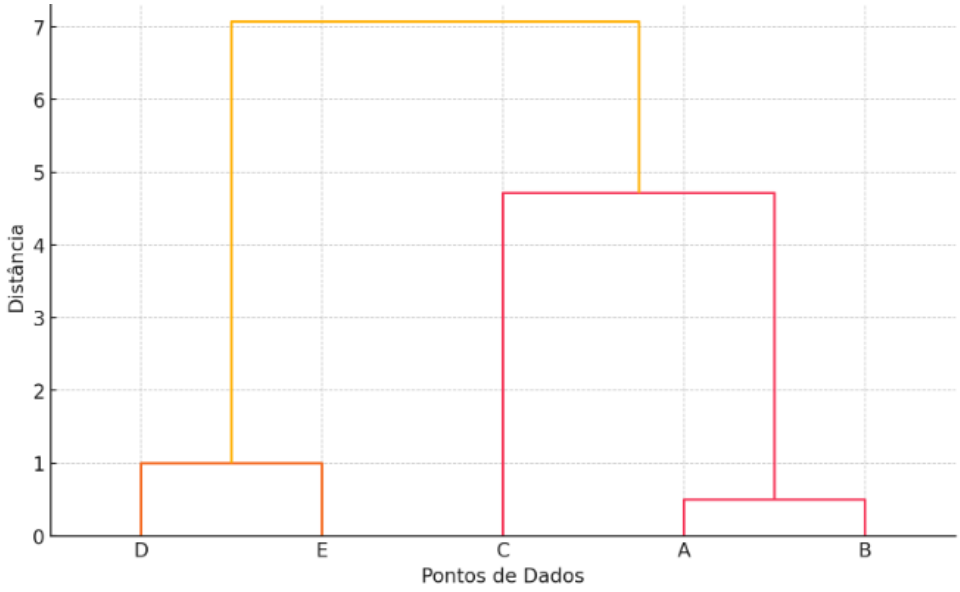
Tabela 2.3: **Matriz de distâncias após o segundo agrupamento**

	AB	C	DE
AB	0		
C	4,72	0	
DE	12,04	7,81	0

Fonte: Elaboração própria

Por fim, da Tabela 2.3, identifica-se que o grupo {A, B} está a 4,72 de distância de C, então a última união será entre esses pares. Para facilitar a identificação do número de grupos, é comumente utilizado o gráfico denominado **dendograma** (Artes e Barroso, 2023 [5]). Neste gráfico, os objetos são dispostos no eixo das abscissas, ao passo que o eixo das ordenadas representa as distâncias em que as uniões foram realizadas. Cada linha horizontal que conecta dois ou mais grupos representa a fusão desses grupos e a altura em que os grupos se unem fornece a distância que levou ao agrupamento. Grupos que se unem em alturas menores são mais semelhantes do que aqueles que se unem em alturas maiores. Assim, ao se analisar o gráfico, procura-se observar grandes saltos, o que indica a união de grupos heterogêneos.

Figura 2.6: **Dendograma resultante**



Fonte: Elaboração própria

O dendrograma resultante (Figura 2.6) mostra as ligações entre os pontos e a hierarquia de agrupamento, que agora inclui todos os pontos A, B, C, D e E, formando um único grupo no final. A partir do gráfico, nota-se um grande aumento na distância entre C e {A,B} e entre {D, E} e {A, B, C}, o que sugere que são agrupamentos muito diferentes. Nesse caso, uma boa solução seria {D, E}, {A, B} e {C}.

Outra família de algoritmos de agrupamentos faz referência aos **métodos de partição**. Nos algoritmos de partição, os grupos finais formam uma partição do conjunto de pontos inicial. Nesses casos, cada ponto é associado ao grupo ao qual melhor se ajusta

seguindo alguma lógica de alocação (MORETTIN e SINGER, 2021 [2]). Conforme indicado por Artes e Barroso (2023) [5], os métodos de partição buscam encontrar a partição em que os grupos apresentem elevada homogeneidade interna e que sejam diferentes entre si. A seguir, são descritos os dois métodos mais utilizados na prática.

Método das k-médias

O critério empregado pelo método das k-médias baseia-se na partição da soma de quadrados total de uma análise de variância e é o algoritmo mais comum nessa classe, proposto por Hartigan e Wong em 1979 (MORETTIN e SINGER, 2021 [2]). O objetivo do k-médias é associar os pontos a k grupos, de forma a minimizar a soma dos quadrados das distâncias dos pontos aos centros dos agrupamentos (centróides). Como forma de diminuir a gama de possíveis partições, esse método demanda que se defina o número de grupos k que devem ser gerados a priori (ARTES e BARROSO, 2023 [5]).

O algoritmo consiste nos seguintes passos:

1. **Inicialização**: o algoritmo começa selecionando aleatoriamente k centróides iniciais (centros dos grupos) a partir do conjunto de dados. Seja $\mathbf{C} = \{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_k\}$ o conjunto de centróides, onde cada $\mathbf{c}_i$ representa o centróide do grupo i.

2. **Atribuição**: nesta etapa, cada ponto de dado $\mathbf{x}_j$ no conjunto de dados é atribuído ao centróide mais próximo. A atribuição baseada na distância euclidiana é definida como:

$$d(\mathbf{x}_j, \mathbf{c}_i) = \|\mathbf{x}_j - \mathbf{c}_i\| = \sqrt{\sum_{m=1}^{p} (x_{jm} - c_{im})^2}$$

   onde $\mathbf{x}_j$ é o j-ésimo ponto dos dados, $\mathbf{c}_i$ é o i-ésimo centróide, e p é o número de dimensões nos dados.

3. **Atualização**: após a atribuição, o algoritmo recalcula os centróides dos grupos com base nas atribuições atuais (média das coordenadas dos pontos). O novo centróide $\mathbf{c}_i$ para o grupo i é computado como a média de todos os pontos de dados atribuídos a esse grupo:

$$\mathbf{c}_i = \frac{1}{g_i} \sum_{x_j \in G_i} \mathbf{x}_j$$

onde $\mathbf{G}_i$ é o conjunto de pontos atribuídos ao grupo i e g_i é o número de pontos no grupo i.

4. **Convergência**: o algoritmo realiza iterativamente as etapas de atribuição e atualização até que a convergência seja alcançada. A convergência pode ser definida como o ponto em que os centróides não mudam significativamente, ou as atribuições dos pontos de dados aos grupos permanecem constantes.

O objetivo geral do algoritmo k-médias é minimizar a soma dos quadrados dentro do grupo, definida como:

$$W(C) = \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in \mathbf{G}_i} \|\mathbf{x}_j - \mathbf{c}_i\|^2$$

Esta função quantifica a compacidade dos grupos. O algoritmo busca minimizar $W(C)$ ao longo do curso das iterações.

Método das k-medoides

O método k-medoides é uma técnica de agrupamento que é semelhante ao algoritmo k-médias, mas com algumas diferenças na forma como lida com os pontos de dados e os centros dos grupos. Nesse método, os centro dos grupos são pontos de dados reais do conjunto de dados, chamados de **medoides**. A medoide de um grupo é determinada pelo objeto que apresenta a menor distância média em relação às demais observaçõs do grupo. Isso contrasta com o k-médias, em que os centros do grupo (calculados pela média das coordenadas dos pontos do grupo) podem ser qualquer ponto no espaço (ARTES e BARROSO, 2023 [5]).

O objetivo principal é minimizar a diferença total (distância) entre as observações e seus respectivos medoides. Assim, seja k o número de grupos a serem formados, o algoritmo busca identificar a partição e respectivas medoides que minimizem a seguinte função, conforme descrito por Artes e Barroso (2023) [5]:

$$C = \sum_{i=1}^{k} \sum_{z \in \mathbf{G}_i} d[\mathbf{m}_i, \mathbf{x}_z] \tag{2.10}$$

em que $\mathbf{G}_i$ representa o conjunto de objetos no grupo i, $d[\mathbf{m}_i, \mathbf{x}_z]$ a distância entre a medoide do grupo i ($\mathbf{m}_i$) e a observação $\mathbf{x}_z$.

O algoritmo inicia com a escolha aleatória de k medoides entre os pontos do

conjunto de dados. Depois, na etapa de **atribuição**, cada ponto de dados é associado ao medoide mais próximo com base na métrica de distância escolhida. Em seguida é feita a etapa de **atualização**, em que, para cada grupo, é calculada a matriz de distância entre todos os pontos e selecionado aquele que minimiza a distância total a todos os outros pontos no grupo para ser o novo medoide. As etapas de atribuição e atualização são repetidas até que os medoides não mudem mais (convergência) ou até que um número pré-definido de iterações seja atingido.

Conforme apresentado por Artes e Barroso (2023) [5], os métodos de k-médias e k-medoides apresentam a flexibilidade de realocar objetos a cada iteração, o que garante uma alocação otimizada. Essa característica os distingue dos métodos hierárquicos, onde uma vez que dois objetos são agrupados, eles permanecem no mesmo grupo durante todo o processo, sem considerar que a introdução de novos elementos aos grupos pode alterar a proximidade entre agrupamentos.

Uma vez que os métodos hierárquicos não requerem conhecimento prévio do número de grupos a serem formados, essa propriedade os torna úteis para a definição inicial de agrupamentos, pode ser utilizados antes da aplicação dos métodos de k-médias ou k-medoides. Entretanto, a aplicação de métodos hierárquicos em grandes conjuntos de dados pode ser computacionalmente custosa e desafiadora na análise dos resultados, além de dificultar a análise por meio de dendogramas. Nessas situações, os métodos de partição, especialmente o k-médias, demonstram-se mais eficazes.

O método das k-médias é mais sensível à presença de *outliers* do que o método de k-medoides. Entretanto, para grandes volumes de dados e valores de k, a complexidade de cada iteração do k-medoides torna-se muito mais custosa do que o k-médias (ARTES e BARROSO, 2023 [5]). Por essas características, considerando que a base de dados tratada neste trabalho possui 3050 observações, a escolha do k-médias se torna mais recomendada.

Definição do Número de Grupos

Segundo Arbelaitz et al.(2013) [7], os índices de validade de grupo (*Cluster Validity Indices*, **CVI**) são métricas que avaliam a qualidade das partições obtidas por algoritmos de agrupamento, com o objetivo de medir o quão bem os grupos representam a estrutura subjacente dos dados. Esses índices têm como objetivo auxiliar na identificação da partição que melhor se ajusta aos dados, ao fornecer uma métrica quantitativa para

comparar diferentes soluções de agrupamento. Entre os CVIs mais utilizados, destacam-se aqueles que medem a coesão intra-grupo (o quão próximos os elementos de um grupo estão entre si) e a separação inter-grupo (o quão distintos são os grupos entre si).

Um dos principais desafios do agrupamento é determinar o número ideal de grupos, frequentemente denotado pela variável k. Em muitos algoritmos, como o k-médias, o valor de k deve ser especificado previamente, o que exige métodos auxiliares para identificar o melhor valor de k. Os CVIs são utilizados nesse processo, pois fornecem critérios para avaliar e comparar diferentes soluções de agrupamento com variados valores de k. O valor que maximiza ou minimiza o índice de validade, dependendo da métrica utilizada, pode ser considerado uma boa opção de partição de grupos.

Em seu trabalho, Arbelaitz et al.(2013) [7] realiza uma extensa análise comparativa de 30 índices de validade de grupos (CVI). Os autores destacam que diferentes algoritmos ou configurações podem gerar diversas partições para o mesmo conjunto de dados, e nenhum algoritmo ou índice de validade é o melhor em todas as situações. A análise abrange tanto conjuntos de dados sintéticos quanto reais, considera diferentes fatores como o número de grupos, dimensionalidade, sobreposição de grupos, densidade e ruído. Entre os principais índices analisados estão ométodo das silhuetas, Davies-Bouldin e Calinski-Harabasz, que apresentam bons resultados em diversos cenários. O método das silhuetas foi o que apresentou o melhor desempenho geral.

Método das Silhuetas

Em seu artigo, Rousseeuw (1987) [8] propõe um método gráfico para interpretar e avalidar análises de grupo. O método das silhuetas é uma técnica que permite avaliar a qualidade dos grupos formados em uma análise de agrupamento, a partir da relação entre a coesão dos grupos (similaridade dentro do grupo) e sua separação (quão distintos os grupos são entre si).

O autor destaca as limitações dos algoritmos comuns de agrupamento, que frequentemente produzem um número fixo de grupos sem fornecer uma visão clara sobre se esses grupos são significativos. Ao contrário dos dendrogramas usados em métodos hierárquicos, os métodos de particionamento, como o k-médias, não oferecem uma representação visual inerente sobre a adequação dos dados aos grupos. O método de silhuetas foi projetado para preencher essa lacuna, oferece uma exibição gráfica que ilustra a qualidade do agrupamento.

O valor do coefciente da silhueta, denotado por $s(i)$, é calculado para cada objeto $\mathbf{x}_i$ no conjunto de dados. Este valor mede quão similar $\mathbf{x}_i$ é aos outros objetos do mesmo grupo em comparação com os objetos no grupo vizinho mais próximo. O valor para um objeto $\mathbf{x}_i i$ é definido da seguinte forma:

- **Coesão dentro do grupo ($a(i)$)**: a distância média entre o objeto i e todos os outros objetos no mesmo grupo $\mathbf{G}_i$. Isso mede quão bem o objeto i está agrupado com os outros membros do seu grupo. Matematicamente, se $\mathbf{G}_i$ for o conjunto de objetos no mesmo grupo de i:

$$a(i) = \frac{1}{g_i - 1} \sum_{j \in G, j \neq i} d(i, j) \tag{2.11}$$

  onde $d(i, j)$ é a dissimilaridade entre os objetos i e j, e g_i é o número de objetos no grupo $\mathbf{G}_i$.

- **Separação do grupo mais próximo ($b(i)$)**: a menor distância média entre o objeto i e todos os objetos em qualquer outro grupo $\mathbf{G}_w$, diferente de $\mathbf{G}_i$. Isso mede quão distante i está do próximo melhor grupo (o grupo vizinho mais próximo):

$$b(i) = \min_{\mathbf{G}_w \neq \mathbf{G}_i} \frac{1}{g_w} \sum_{j \in \mathbf{G}_w} d(i, j) \tag{2.12}$$

- **Valor do coeficiente de silhueta ($s(i)$)**: uma vez que $a(i)$ e $b(i)$ são calculados, o valor do coeficiente de silhueta para o objeto i é dado por:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{2.13}$$

  O valor do coeficiente de silhueta $s(i)$ varia de -1 a 1:

  - $s(i)$ próximo de 1 indica que o objeto está bem agrupado (ou seja, a distância dentro do grupo é muito menor do que sua distância ao grupo vizinho mais próximo).
  - $s(i)$ próximo de 0 significa que o objeto está na fronteira entre dois grupos.
  - $s(i)$ próximo de -1 indica que o objeto provavelmente foi mal classificado, já que está mais próximo de outro grupo do que do seu próprio.

Rousseeuw (1987) [8] propõe o uso do coeficiente de silhueta médio como uma medida geral da qualidade do grupo. O coeficiente de silhueta médio para um grupo $\mathbf{G}_i$ é a média de todos os valores $s(i)$ dos objetos pertencentes ao grupo $\mathbf{G}_i$:

$$S(\mathbf{G}_i) = \frac{1}{g_i} \sum_{i \in \mathbf{G}_i} s(i) \tag{2.14}$$

O coeficiente de silhueta médio para todo o conjunto de dados $S(k)$ é a média dos valores do coeficiente de silhueta em todos os grupos:

$$S(k) = \frac{1}{n} \sum_{i=1}^{n} s(i) \tag{2.15}$$

onde n é o número total de objetos. O valor de $S(k)$ ajuda a determinar o número mais apropriado de grupos. O número ideal de grupos é, normalmente, aquele que maximiza o coeficiente de silhueta médio, pois isso indica que os grupos estão bem separados e são coesos.

Índice Davies-Bouldin

Proposto por Davies e Bouldin (1979) [9], esse índice, denotada como $\bar{R}$, quantifica a similaridade entre os grupos com base na distância dos centróides e nas dispersões dos grupos, fornece um método quantitativo para comparar diferentes soluções de agrupamento.

Para quantificar a dispersão (ou compactação) de um grupo, Davies e Bouldin (1979) [9] definem a medida de dispersão, denotada como D_i, como:

$$S_i = \left(\frac{1}{g_i} \sum_{j=1}^{g_i} |\mathbf{x}_j - \mathbf{c}_i|^q \right)^{\frac{1}{q}} \tag{2.16}$$

onde g_i representa o número de vetores no grupo $\mathbf{G}_i$, $\mathbf{x}_j$ são os pontos de dados no grupo, $\mathbf{c}_i$ é o centróide do grupo, e q é um parâmetro que define a métrica de distância utilizada, sendo $q = 2$ o padrão para a distância euclidiana. Conforme definido pelos autores, D_i é a q-ésima raiz do q-ésimo momento dos pontos no grupo $\mathbf{G}_i$ em relação à sua média. Assim, se $q = 1$, D_i torna-se a distância euclidiana média dos vetores no grupo $\mathbf{G}_i$ ao centróide do grupo i. Por outro lado, para $q = 2$, D_i é o desvio padrão da distância das amostras em um grupo ao seu respectivo centro.

Outro parâmetro utilizado é a medida de distância M_{ij} que calcula a distância

entre os centróides dos grupos $\mathbf{G}_i$ e $\mathbf{G}_j$ e é definida como:

$$M_{ij} = \left(\sum_{z=1}^{p} |\mathbf{c}_{iz} - \mathbf{c}_{jz}|^q \right)^{\frac{1}{q}} \tag{2.17}$$

Deve-se notar que M_{ij} é a métrica Minkowski dos centróides que caracterizam os grupos $\mathbf{G}_i$ e $\mathbf{G}_j$. Quando $q = 1$, M_{ij} reduz-se à distância de Manhattan. Quando $q = 2$, M_{ij} é a distância euclidiana entre centróides.

Usando essas definições, a medida de separação entre grupos R_{ij} pode ser formulada como:

$$R_{ij} = \frac{D_i + D_j}{M_{ij}} \tag{2.18}$$

Essa equação indica a similaridade entre os grupos $\mathbf{G}_i$ e $\mathbf{G}_j$, onde um valor menor de R_{ij} reflete uma melhor separação entre os grupos.

O principal objetivo do Índice Davies-Bouldin consiste em minimizar a similaridade média de cada grupo com seu grupo mais semelhante. Isso é alcançado por meio do cálculo da similaridade máxima para cada grupo:

$$R_i = \max_{j \neq i}(R_{ij}) \tag{2.19}$$

Subsequentemente, o Índice Davies-Bouldin $\bar{R}$ é calculado como:

$$\bar{R} = \frac{1}{k} \sum_{i=1}^{k} R_i \tag{2.20}$$

onde k é o número total de grupos. O conceito de $\bar{R}$ é ser a média sistemática das medidas de similaridade de cada grupo com seu grupo mais semelhante. Assim, um valor mais baixo de $\bar{R}$ indica uma melhor configuração de agrupamento, sugerindo que os grupos estão bem separados entre si e compactos em si.

Indice Calinski-Harabasz

O índice proposto por Calinski e Harabasz (1974) [10], também conhecido como Critério da Razão da Variância, pode ser utilizado para avaliar os modelos de agrupamentos, em que uma pontuação Calinski-Harabasz mais alta está relacionada a um modelo com grupos melhor definidos. O índice é a razão entre a variância entre grupos e a variância dentro dos grupos (em que a variância é definida como a soma dos quadrados das

distâncias). A ideia principal é que um bom agrupamento deve maximizar a variância entre os grupos (ou seja, as diferenças entre os grupos) enquanto minimiza a variância dentro dos grupos (ou seja, as diferenças dentro de cada grupo).

Considere k como o número total de grupos,n como o número total de pontos de dados, g_i como o número de pontos no grupo $\mathbf{G}_i$, $\bar{\mathbf{x}}$ como o centroide global (ou média) de todos os dados, e $\mathbf{c}_i$ como o centroide do grupo $\mathbf{G}_i$. A soma dos quadrados das distâncias entre os grupos (SSB) é calculada pela equação:

$$SSB = \sum_{i=1}^{k} g_i \times ||\mathbf{c}_i - \bar{\mathbf{x}}||^2 \tag{2.21}$$

onde $||\mathbf{c}_i - \bar{\mathbf{x}}||^2$ representa a distância quadrática entre o centroide do grupo $\mathbf{G}_i$ e o centroide global. Em contrapartida, a soma dos quadrados dentro dos grupos (SSW) é dada por:

$$SSW = \sum_{i=1}^{k} \sum_{x_z \in G_i} ||\mathbf{x}_z - \mathbf{c}_i||^2 \tag{2.22}$$

nesta equação, $||\mathbf{x}_z - \mathbf{c}_i||^2$ é a distância quadrática entre um ponto $\mathbf{x}_z$ pertencente ao grupo $\mathbf{G}_i$ e seu centroide.

A partir dessas definições, o índice Calinski-Harabasz é finalmente calculado como:

$$CH = \frac{SSB/(k-1)}{SSW/(n-k)} \tag{2.23}$$

onde $\frac{SSB}{k-1}$ representa a variância média entre os grupos, enquanto $\frac{SSW}{n-k}$ representa a variância média dentro dos grupos. Assim, o índice CH fornece uma medida que pode ser utilizada para avaliar a eficácia do agrupamento.

A separação adequada dos grupos resulta em uma maior dispersão entre eles em relação à dispersão dentro deles, culminando em um índice mais elevado. Assim, ao calcular o índice para diferentes quantidades de grupos, é possível determinar o número que maximiza essa relação, sinalizando um agrupamento mais robusto e distinto, o que facilita a seleção do número de grupos mais apropriado aos padrões dos dados. Por outro lado, valores baixos indicam a possibilidade de sobreposição entre os grupos ou uma inadequação na definição dos grupos, sinalizando que o agrupamento pode não ser

adequado para os dados em questão.

Interpretação dos Grupos

Como colocado por Faceli et al. (2011) [3], a interpretação de grupos envolve a análise de cada grupo em relação aos seus objetos, visando rotulá-los e descrever sua natureza. Este processo vai além da mera descrição, pois permite validar os grupos identificados e explorar avaliações subjetivas com significado prático. Os autores acrescentam que o envolvimento de um especialista é fundamental, pois seu conhecimento sobre os dados facilita a identificação de significados e relações entre os grupos. Além disso, a visualização gráfica dos grupos ajuda a apresentar os resultados de forma clara e intuitiva, como também é mencionado por Artes e Barroso (2023) [5].

Capítulo 3

METODOLOGIA

Neste capítulo é apresentada a estrutura metodológica empregada neste trabalho para a analise de agrupamentos dos estabelecimentos clientes da Corporação S.A.. Essa estrutura foi elaborada para garantir uma abordagem sistemática e rigorosa na coleta e análise de dados.

O capítulo inicia-se com a descrição dos dados, em que são apresentados os conjuntos de informações coletadas, incluem suas características principais, como variáveis, amostras e formatos. Essa etapa é essencial para compreender a natureza dos dados e identificar potenciais desafios que podem surgir durante a análise. Em seguida, é apresentada a análise exploratória, que envolve a aplicação de técnicas estatísticas e visuais para explorar padrões, tendências e anomalias presentes nos dados. Essa análise permite uma visão mais clara sobre as relações entre variáveis e a distribuição dos dados. Por fim, é elaborada a etapa de pré-processamento, que inclui a limpeza dos dados, tratamento de valores ausentes e normalização, preparando-os adequadamente para as fases subsequentes de modelagem e análise.

3.1 Descrição dos Dados Coletados

3.1.1 Unidade de Investigação

A unidade de investigação deste trabalho consiste nos estabelecimentos comerciais que são clientes da Corporação S.A. e possuem cadastro tanto no App Corp, quanto no Corp Delivery. A seleção desses estabelecimentos se justifica pela possibilidade de analisar de forma integrada os dados de compras realizadas no App Corp e as vendas efetuadas

por meio do Corp Delivery dos produtos da organização. Essa abordagem proporciona uma visão do comportamento comercial desses estabelecimentos em relação aos produtos fornecidos pela Corporação S.A., o que permite uma análise das dinâmicas de compra e venda. No total, foram extraídos dados de 3050 estabelecimentos da base que atendem a esse critério.

Para cada estabelecimento incluído na amostra, foram coletados dados de caracterização para contextualizar a análise e permitir uma segmentação detalhada. As variáveis de caracterização incluem o identificador único de cada estabelecimento, o tipo de operação comercial (como bar, restaurante, supermercado, entre outros) e a unidade federativa onde o estabelecimento está localizado. Essas informações são essenciais para compreender as diferenças regionais e operacionais dos estabelecimentos.

3.1.2 Dados de Compra e Venda

A análise dos dados de compras se concentra nas transações realizadas pelos estabelecimentos por meio do App Corp, a plataforma digital de compras da Corporação S.A. As variáveis coletadas incluem o volume total de produtos adquiridos, o número total de compras realizadas, o valor monetário total das compras, o número de SKUs (*Stock Keeping Units*) comprados e o número de SKUs distintos adquiridos. Essas métricas oferecem uma visão detalhada do comportamento de compra dos estabelecimentos, o que permite identificar padrões de aquisição e preferência por determinados produtos.

Além dos dados de compras, também foram coletadas informações sobre as vendas realizadas pelos estabelecimentos por meio do Corp Delivery, a plataforma de entregas da Corporação S.A. As variáveis de interesse incluem o volume total de vendas, o número de transações de venda, o número de SKUs vendidos e o número de SKUs distintos vendidos. Essa análise é essencial para compreender o desempenho dos estabelecimentos na comercialização dos produtos da Corporação S.A. e identificar possíveis correlações entre os produtos comprados e vendidos.

3.1.3 Segmentação dos Dados de Portfólio

Uma das estratégias adotadas neste trabalho foi a segmentação dos dados de compras e vendas em duas categorias: "total" e "portfólio". A Corporação S.A. define um portfólio ideal de produtos que cada estabelecimento deveria manter em estoque, de

acordo com as estratégias comerciais da companhia. Portanto, foi necessário avaliar as variáveis relacionadas aos produtos do portfólio separadamente, o que permite uma análise mais precisa da aderência dos estabelecimentos às recomendações da empresa.

3.1.4 Diferenciação entre Bebidas Alcoólicas e Não Alcoólicas

Outra segmentação relevante aplicada ao conjunto de dados foi a separação entre bebidas alcoólicas e não alcoólicas. A Corporação S.A. adota estratégias distintas para cada uma dessas categorias, refletindo em diferentes dinâmicas de mercado e padrões de consumo. Ao segmentar as variáveis de compra e venda de acordo com o tipo de bebida, é possível identificar comportamentos específicos dos estabelecimentos em relação a cada categoria, o que pode auxiliar no desenvolvimento de estratégias comerciais mais direcionadas.

Após analisar os dados dos estabelecimentos quanto ao tipo de bebidas comercializadas, constatou-se que o volume de compra de bebidas alcoólicas supera 69% em todas as unidades federativas (Tabela 3.1), enquanto o volume de venda dessas bebidas é superior a 86% em cada um deles, o que evidencia a predominância das bebidas alcoólicas nas operações de compra e venda em relação a bebidas não alcoólicas. Considerando as limitações de recursos disponíveis para o estudo, optou-se por restringir o escopo deste trabalho para as bebidas alcoólicas. Essa decisão visa concentrar os esforços analíticos nos produtos que apresentam maior impacto comercial.

Tabela 3.1: **Volume mensal de compra e venda por unidade federativa**

UF	Vol. Venda - Alcoólicos (hl)	Vol. Venda - Total (hl)	Alcoólicos na venda total (%)	Vol. Compra - Alcoolicos (hl)	Vol. Compra - Total (hl)	Alcoólicos na compra total (%)
GO	395	2.906	86	3.444	4.186	82
RS	12.223	26.391	86	15.431	17.879	86
RJ	12.469	89.685	86	57.979	68.202	85
DF	313	2.463	87	2.620	3.080	85
PE	1.130	8.583	87	8.132	10.032	81
SP	9.865	75.132	87	85.570	99.165	86
AL	134	1.132	88	1.370	1.976	69
PB	311	2.587	88	2.282	2.748	83
MG	4.575	41.147	89	39.620	45.514	87
TO	73	733	90	1.429	1.732	83
BA	579	5.532	90	5.929	7.067	84
PR	1.210	12.198	90	13.782	15.491	89
RN	54	584	91	444	512	87
SE	637	1.340	91	1.353	1.598	85
ES	262	2.924	91	2.979	3.348	89
SC	699	7.884	91	7.821	9.078	86
MT	675	7.585	91	12.257	13.481	91
AP	28	348	92	472	520	91
PA	398	5.072	92	5.674	6.493	87
MS	305	3.616	92	5.958	6.645	90
RR	46	692	93	971	1.117	87
AM	296	4.935	94	4.013	5.496	73
CE	358	6.356	94	5.000	5.714	87
AC	2	840	95	1.695	1.827	93
MA	335	6.946	95	4.515	4.797	94
RO	73	1.799	96	2.188	2.317	94
PI	114	2.581	96	3.891	4.089	95

Fonte: Elaboração própria

3.1.5 Dicionário de Variáveis

Para assegurar a clareza e a replicabilidade da análise, foi desenvolvido um dicionário de variáveis conforme sugerido por Morettin (2021)[2]. Este dicionário contém a identificação de cada variável, uma definição detalhada e a unidade. A inclusão do dicionário garante que outras áreas da empresa possam interpretar corretamente os dados e replicar os resultados da análise em estudos futuros. A transparência na definição das

variáveis também fortalece a validade e a confiabilidade dos resultados obtidos.

Para este trabalho foram extraídos os valores de 21 variáveis para caracterizar cada um dos 3050 estabelecimentos da base, conforme Tabela 3.2.

Tabela 3.2: **Dicionário das variáveis extraídas da base de dados**

Grupo	Rótulo	Descrição	Unidade de medida/Categorias
Estabelecimento	id_est	Identificador do estabelecimento comercial	
	tipo_est	Tipo de operação do estabelecimento comercial	ACOUGUE; ATACADO; BAR; HOSPEDAGEM;LANCHONETE; LOJA DE CONVENIENCIA; MERCADO; MERCEARIA; PADARIA; RESTAURANTE; SUBDISTRIBUIDOR; OUTROS;
	uf_est	Unidade Federativa em que o estabelecimento está localizado	AC; AL; AM; AP; BA; CE; DF; ES; GO; MA; MG; MS; MT; PA; PB; PE; PI; PR; RJ; RN; RO; RR; RS; SC; SE; SP; TO
Compra Total	val_c_tot	Valor médio desembolsado pelo estabelecimento para a compra dos produtos alcoólicos da Corporação S.A. por mês	Reais
	n_c_tot	# médio de compras feitas pelo estabelecimento no App Corp por mês que continham pelo menos um produto alcoólico	# inteiro
	vol_c_tot	Volume médio total dos produtos alcoólicos comprados pelo estabelecimento no App Corp por mês	Hectolitros
	sku_c_tot	Qtdd média de SKUs total comprados pelo estabelecimento no App Corp por mês na categoria de produtos alcoólicos	# inteiro
	sku_d_c_tot	Qtdd média de SKUs distintos total comprados pelo estabelecimento no App Corp por mês na categoria de produtos alcoólicos	# inteiro
Compra Portfólio	val_c_port	Valor médio desembolsado pelo estabelecimento para a compra dos produtos alcoólicos pertencentes ao portfólio da Corporação S.A. por mês	Reais
	n_c_port	# médio de compras feitas por mês pelo estabelecimento no App Corp que continham pelo menos um produto alcoólico que pertence ao portfólio ideal	# inteiro
	vol_c_port	Volume médio referente aos produtos que pertencem ao portfólio ideal comprados pelo estabelecimento no App Corp por mês	Hectolitros
	sku_c_port	Qtdd média de SKUs do portfólio comprados pelo estabelecimento no App Corp por mês na categoria de produtos alcoólicos	# inteiro
	sku_d_c_port	Qtdd média de SKUs distintos do portfólio comprados pelo estabelecimento no App Corp por mês na categoria de produtos alcoólicos	# inteiro
Venda Total	n_v_tot	# médio por mês de vendas feitas pelo estabelecimento no Corp Delivery que continham pelo menos um produto alcoólico	# inteiro
	vol_v_tot	Volume médio total de produtos alcoólicos vendidos pelo estabelecimento no Corp Delivery por mês	Hectolitros
	sku_v_tot	Qtdd média de SKUs total vendidos pelo estabelecimento no Corp Delivery por mês na categoria de produtos alcoólicos	# inteiro
	sku_d_v_tot	Qtdd média de SKUs distintos total vendidos pelo estabelecimento no Corp Delivery por mês na categoria de produtos alcoólicos	# inteiro
Venda Portfólio	n_v_port	# médio de vendas feitas por mês pelo estabelecimento no Corp Delivery que continham pelo menos um produto alcoólico que pertence ao portfólio ideal	# inteiro
	vol_v_port	Volume médio referente aos produtos alcoólicos que pertencem ao portfólio ideal vendidos pelo estabelecimento no Corp Delivery por mês	Hectolitros
	sku_v_port	Qtdd média de SKUs do portfólio vendidos pelo estabelecimento no Corp Delivery por mês na categoria de produtos alcoólicos	# inteiro
	sku_d_v_port	Qtdd média de SKUs distintos do portfólio vendidos pelo estabelecimento no Corp Delivery por mês na categoria de produtos alcoólicos	# inteiro

Fonte: Elaboração própria

3.1.6 Período de Coleta de Dados

O período de análise dos dados foi escolhido para evitar distorções causadas pela sazonalidade, um fator comum no setor de varejo. Os dados utilizados foram extraídos de 01 de fevereiro de 2024 a 30 de abril de 2024, considerados pela Corporação S.A. como meses neutros em relação à sazonalidade. Para caracterizar o comportamento mensal de cada estabelecimento, foi utilizada a média desses meses para cada variável. A utilização da média de três meses é uma prática comum na empresa para suavizar variações atípicas e capturar de forma mais fiel o comportamento regular dos estabelecimentos. Essa abordagem permite que os resultados da análise sejam mais representativos e aplicáveis ao longo do tempo. Para exemplificar, a Tabela 3.3 demonstra os dados extraídos de cinco estabelecimentos da base que vai ser utilizada na modelagem.

Tabela 3.3: **Exemplos de estabelecimentos extraídos da base**

Variável	**Exemplo 1**	**Exemplo 2**	**Exemplo 3**	**Exemplo 4**	**Exemplo 5**
id_est	36_429	36_443	36_1355	36_1356	36_1482
tipo_est	Subdistribuidor	Mercado	Outros	Subdistribuidor	Subdistribuidor
uf_est	RJ	RJ	RJ	RJ	RJ
val_c_tot	35	4.978	0	95.805	2.690
n_c_tot	1	3	0	14	2
vol_c_tot	0	5	0	165	2
sku_c_tot	1	19	0	136	9
sku_d_c_tot	1	15	0	41	9
val_c_port	0	3.192	0	54.775	1.079
n_c_port	1	3	0	13	1
vol_c_port	0	4	0	74	1
sku_c_port	0	9	0	54	2
sku_d_c_port	0	7	0	15	2
n_v_tot	416	2.145	128	2.398	1.007
vol_v_tot	24	84	7	148	40
sku_v_tot	439	2.159	130	2.575	1.074
sku_d_v_tot	32	72	17	73	45
n_v_port	307	1.692	102	1.806	735
vol_v_port	17	85	5	91	39
sku_v_port	282	1.504	102	1.701	703
sku_d_v_port	17	36	12	35	22

Fonte: Elaboração própria

Análise de Consistência

Para verificar a completude e a qualidade da base, foram feitos alguns testes. O primeiro foi a verificação dos tipos de dados por variável para garantir o processamento correto. Os resultados estão presentes na Tabela 3.4 e demonstram que todos os dados estão no formato esperado, com 3050 observações (ou linhas) no total.

Conforme descrito na Seção 2.2.2, a segunda verificação foi a presença de valores ausentes por variável (ou coluna). Valores nulos no identificador (id_est), tipo de estabelecimento (tipo_est) ou unidade federativa (uf_est) são considerados inconsistências, uma vez que eles representam as características do estabelecimento. Além disso, para que um estabelecimento faça parte do Corp Delivery, ele precisa estar ativo, ou seja, precisa movimentar as vendas dentro do aplicativo em três meses. Dessa forma, valores nulos nos dados de venda (variáves com o indicador "_v_") também seriam considerados como inconsistentes. Conforme apresentado na Tabela 3.4, não foram identificados valores nulos para nenhum dos dois casos.

Em relação às variáveis de compra (identificadas com "_c_"), valores nulos são esperados devido à natureza do problema. Alguns casos conhecidos pela empresa são em estabelecimentos que atendem ao Corp Delivery e que possuem mais de um CNPJ (que centraliza as compras), e em estabelecimentos que vendem pelo CNPJ da empresa mas compram por meio de um cadastro de pessoa física (CPF). Assim, esses dados não são considerados inconsistentes. Nesse caso, valores nulos são equivalentes a zero, ou seja, pode-se considerar que estabelecimentos que não possuem dados de compra têm um valor de compra igual a zero e, portanto, os valores nulos foram subtituído por zero para permitir a análise estatística e o tratamento desses dados.

Por fim, verificou-se a presença de valores negativos na base, o que representaria uma inconsistência já que os dados se referem a transações de produtos físicos e suas características (volume, valor financeiro, quantidade, etc). Após análise, não foram identificados valores negativos na base.

Tabela 3.4: **Tipo de dados e valores nulos**

Variável	**Tipo de dados na base**	**Quantidade de valores nulos**
id_est	*object*	0
tipo_est	*object*	0
uf_est	*object*	0
val_c_tot	*float*	163
n_c_tot	*float*	156
vol_c_tot	*float*	163
sku_c_tot	*float*	163
sku_d_c_tot	*float*	163
val_c_port	*float*	173
n_c_port	*float*	158
vol_c_port	*float*	173
sku_c_port	*float*	173
sku_d_c_port	*float*	173
n_v_tot	*float*	0
vol_v_tot	*float*	0
sku_v_tot	*float*	0
sku_d_v_tot	*float*	0
n_v_port	*float*	0
vol_v_port	*float*	0
sku_v_port	*float*	0
sku_d_v_port	*float*	0

Fonte: Elaboração própria

3.2 Análise Exploratória dos Dados

3.2.1 Estatística Descritiva

Conforme descrito na Seção 2.2.3, os dados das estatísticas descritivas das variáveis extraídas da base (Tabela 3.2) revelam uma grande heterogeneidade entre os estabelecimentos, tanto em termos de compras quanto de vendas. Enquanto alguns estabelecimentos realizam compras e vendas de grandes volumes, muitos outros operam em uma

escala bem menor. Além disso, o portfólio ideal tem uma participação significativa nas operações dos estabelecimentos, mas a magnitude dessa participação varia amplamente. Isso reforça a necessidade de estratégias personalizadas para atender às necessidades de diferentes perfis de clientes, desde pequenos estabelecimentos até grandes compradores.

Tabela 3.5: **Estatística descritivas das variáveis extraídas da base.**

Variável	Média	Desvio padrão	Mínimo	Quartil 1	Quartil 2	Quartil 3	Máximo
val_c_tot	82.750	128.862	0	21.032	52.917	99.536	2.493.994
n_c_tot	9	5	0	5	8	12	28
vol_c_tot	120	192	0	24	69	147	3.750
sku_c_tot	124	141	0	44	97	162	1.734
sku_d_c_tot	40	22	0	24	41	55	111
val_c_port	45.327	81.087	0	10.058	27.125	51.531	1.596.821
n_c_port	8	5	0	5	7	10	25
vol_c_port	52	95	0	11	30	59	2.203
sku_c_port	58	63	0	22	47	75	822
sku_d_c_port	20	11	0	13	21	28	58
n_v_tot	1.773	1.746	1	753	1.374	2.203	17.698
vol_v_tot	96	122	0	35	66	118	1.796
sku_v_tot	1.948	2.385	1	781	1.447	2.328	35.947
sku_d_v_tot	61	24	1	45	61	78	145
n_v_port	1.334	1.314	1	561	1.030	1.667	13.125
vol_v_port	60	75	0	22	43	75	1.077
sku_v_port	1.303	1.621	1	524	951	1.564	24.392
sku_d_v_port	34	13	1	26	34	43	86

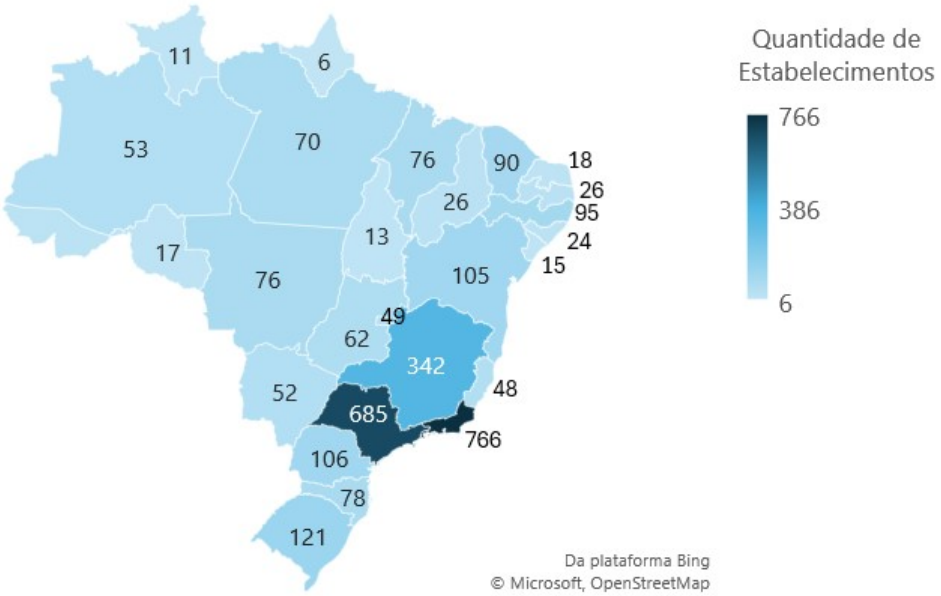
Fonte: Elaboração própria

Como pode-se notar, no geral existem valores muito discrepantes observados pelos máximos. Os *boxplots* na Figura 3.7 demonstram a presença de uma grande quantidade de *outliers* em todas as variáveis. Nesse caso, a mediana oferece uma medida de tendência central mais robusta, uma vez que a média tende a se afastar da maioria dos valores centrais quando há muitos *outliers*, sendo menos representativa. Assim, pelo menos 50% dos estabelecimentos fazem 8 compras por mês, comprando 69hl nesse período, e gastam aproximadamente R$ 53 mil, com 97 SKUs diferentes, sendo 47 do portfólio ideal. Além disso, pelo menos 50% dos estabelecimentos fazem 1374 vendas por mês no app Corp Delivery (mais de 40 vendas por dia), vendendo 66hl no período, com 61 SKUs distintos, sendo 34 do portfólio ideal.

3.2.2 Análise Univariada

A análise da distribuição de estabelecimentos por unidade federativa no Brasil, conforme mostrado na Figura 3.1, revela que a região Sudeste apresenta a maior quantidade de estabelecimentos, totalizando 1.841, com destaque para o Rio de Janeiro, que concentra 766 estabelecimentos, seguido por São Paulo com 685. O Nordeste ocupa a segunda posição com 475 estabelecimentos, sendo Bahia e Maranhão os estados que mais contribuem para esse total, com 105 e 76 respectivamente. O Sul possui 305 estabelecimentos, com Paraná (106) e Rio Grande do Sul (121) se destacando. No Centro-Oeste, a soma é de 239 estabelecimentos, com Mato Grosso (76) e Distrito Federal (49) apresentando os maiores números. A região Norte, com 190 estabelecimentos, apresenta a menor relevância, tendo Amapá (6) e Roraima (11) como os estados com menor representatividade. Essas informações indicam uma concentração significativa de atividades comerciais no Sudeste, refletindo a dinâmica econômica e populacional da região.

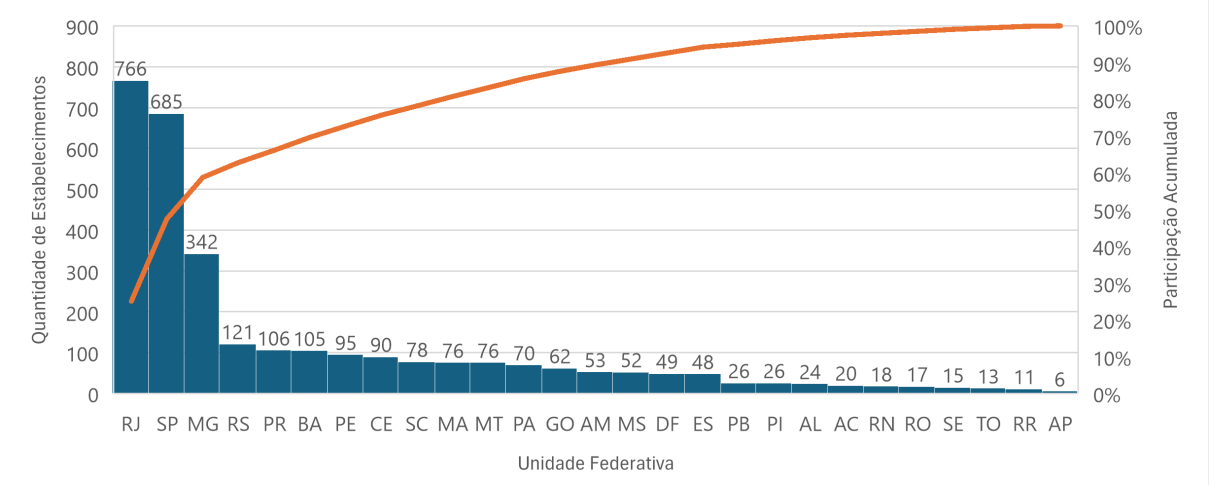
Figura 3.1: **Quantidade de estabelecimentos por unidade federativa**



Fonte: Elaboração própria

Adicionalmente, conforme mostrado na Figura 3.2, os três estados mais relevantes são Rio de Janeiro, São Paulo e Minas Gerais, todos localizados na região Sudeste, concentrando 58,79% dos estabelecimentos do analisados no país.

Figura 3.2: **Diagrama de Pareto da quantidade de estabelecimentos por unidade federativa**

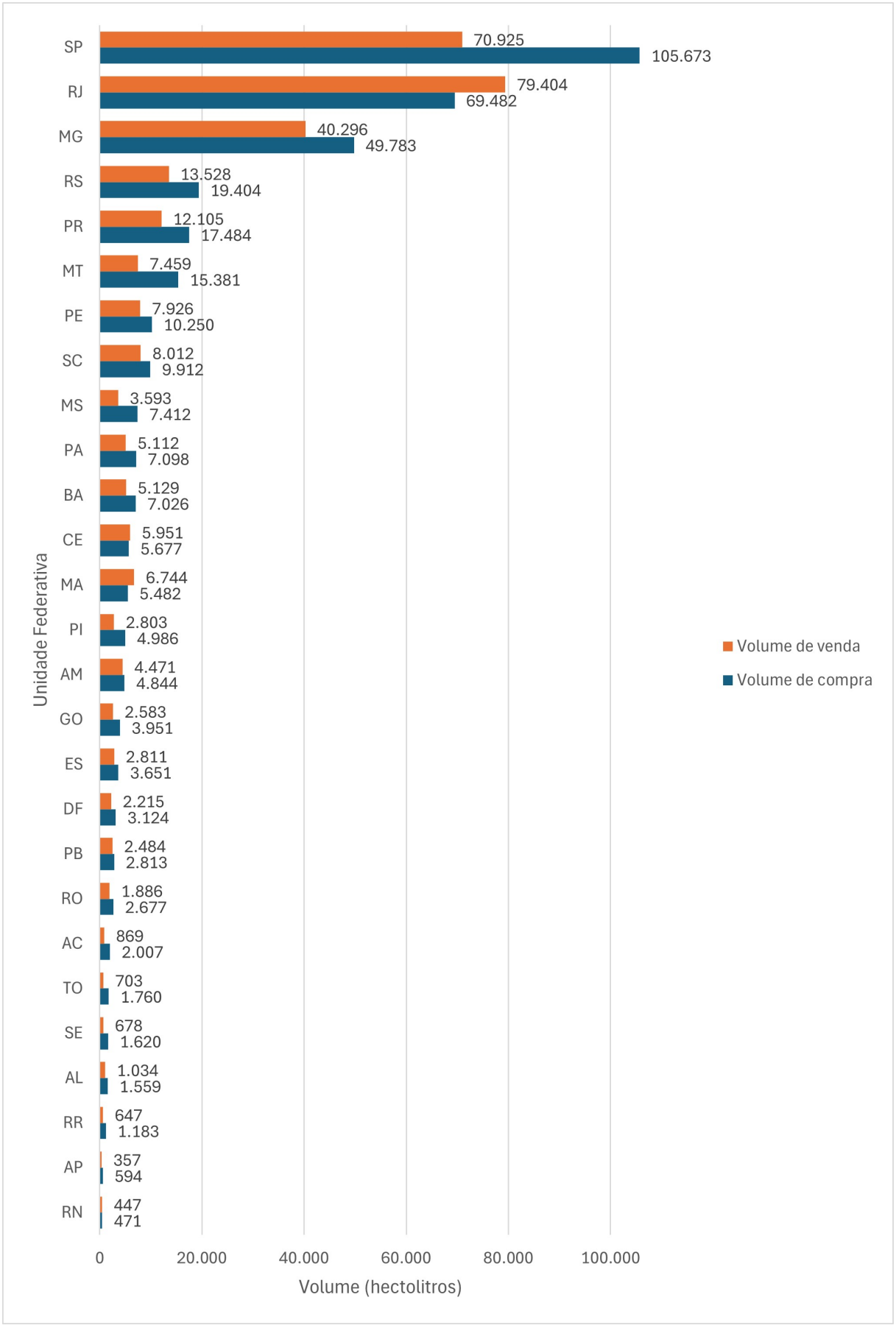


Fonte: Elaboração própria

A análise dos volumes mensais de compra e venda por estabelecimento nas unidades federativas do Brasil (Figura 3.3) demonstra que as unidades federativas que possuem a maior quantidade de estabelecimentos, também apresentam a maiores volumes totais. No entando, o volume médio por estabelecimento (Figura 3.4) mostra uma diversidade significativa nos dados.

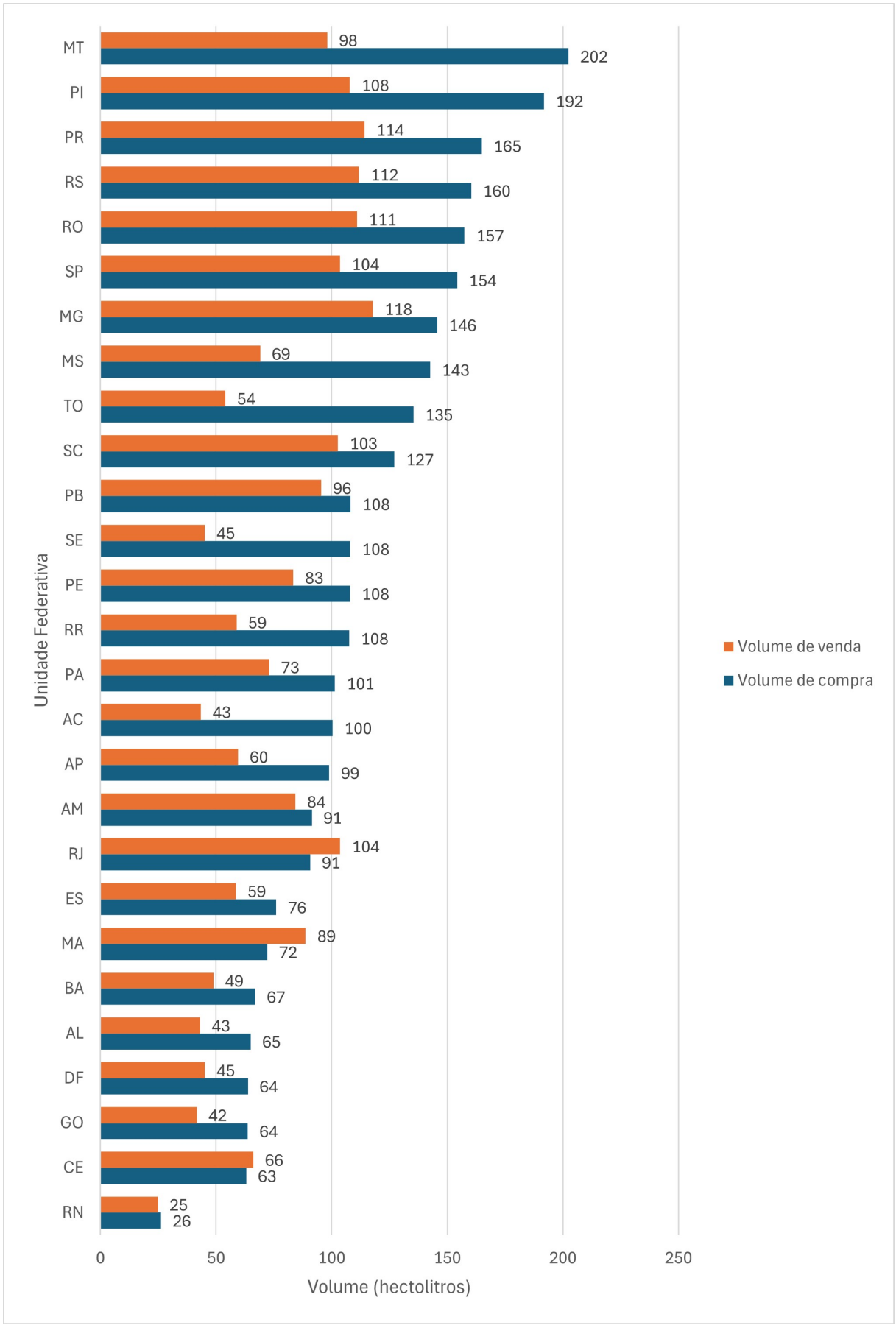
Mato Grosso (MT) lidera com o maior volume de compra médio, com 202 hectolitros por estabelecimento, seguido por Piauí (PI) com 192 hl. Em contraste, o estado do Rio Grande do Norte (RN) apresenta o menor volume médio de compra, com apenas 26 hl. Em relação ao volume de vendas, o estado de Minas Gerais (MG) se destaca, registrando 118 hl por estabelecimento, seguido por Paraná (114 hl), enquanto o Rio Grande do Norte também possui o menor volume de vendas médio, com 25.

Figura 3.3: **Volume mensal total de compra e venda por unidade federativa**



Fonte: Elaboração própria

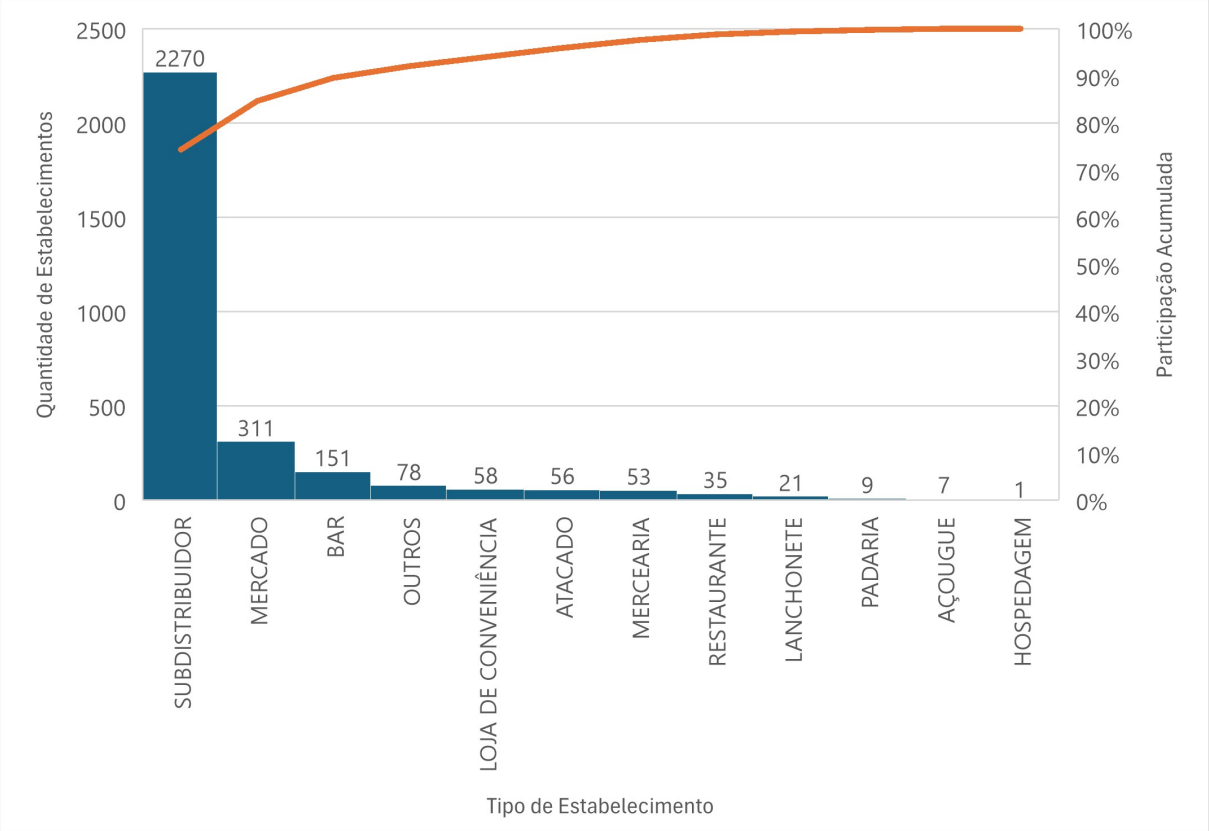
Figura 3.4: **Volume mensal médio de compra e venda por estabelecimento por unidade federativa**



Fonte: Elaboração própria

Em relação ao tipo de estabelecimentos, há a predominância de Subdistribuidores (estabelecimentos que distribuiem para outros estabelecimentos) com 74% da quantidade de estabelecimentos da base analisada, seguidos de Mercado (10,2%) e Bar (5%). Em contrapartida, Padaria (0,3%), Açougue (0,2%) e Hospedagem (0,03%) apresentam as menores quantidades de estabelecimentos, conforme indicado na Figura 3.5.

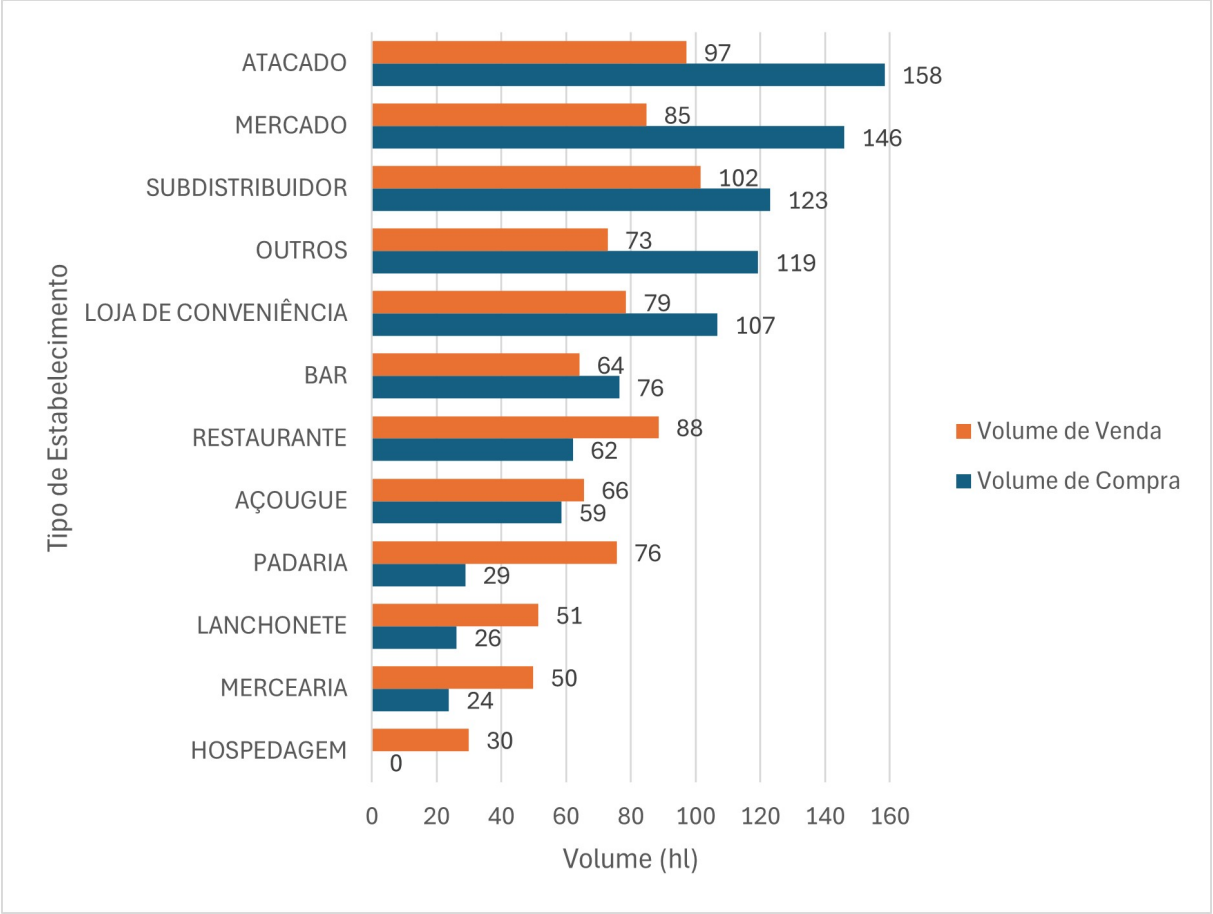
Figura 3.5: **Diagrama de Pareto da quantidade de estabelecimentos por tipo de estabelecimento**



Fonte: Elaboração própria

Considerando o volume mensal de compra e venda, os representantes de Atacado e Mercado possuem os maiores valores médios de compra por estabelecimento, com 158 hl e 146 hl, respectivamente, seguidos de Subdistribuidor (123 hl) e Loja de Conveniência (107 hl). Já em relação ao volume médio mensal de venda por estabelecimento, Subsdistribuidor (102 hl) e Atacado (97 hl) se destacam, como mostrado na Figura 3.6. Apesar da pouca quantidade de estabelecimentos classificados como Padaria, eles apresentam um volume médio de venda considerável (76 hl) comparado às demais categorias.

Figura 3.6: **Volume mensal médio de compra e venda por estabelecimento por tipo de estabelecimento**



Fonte: Elaboração própria

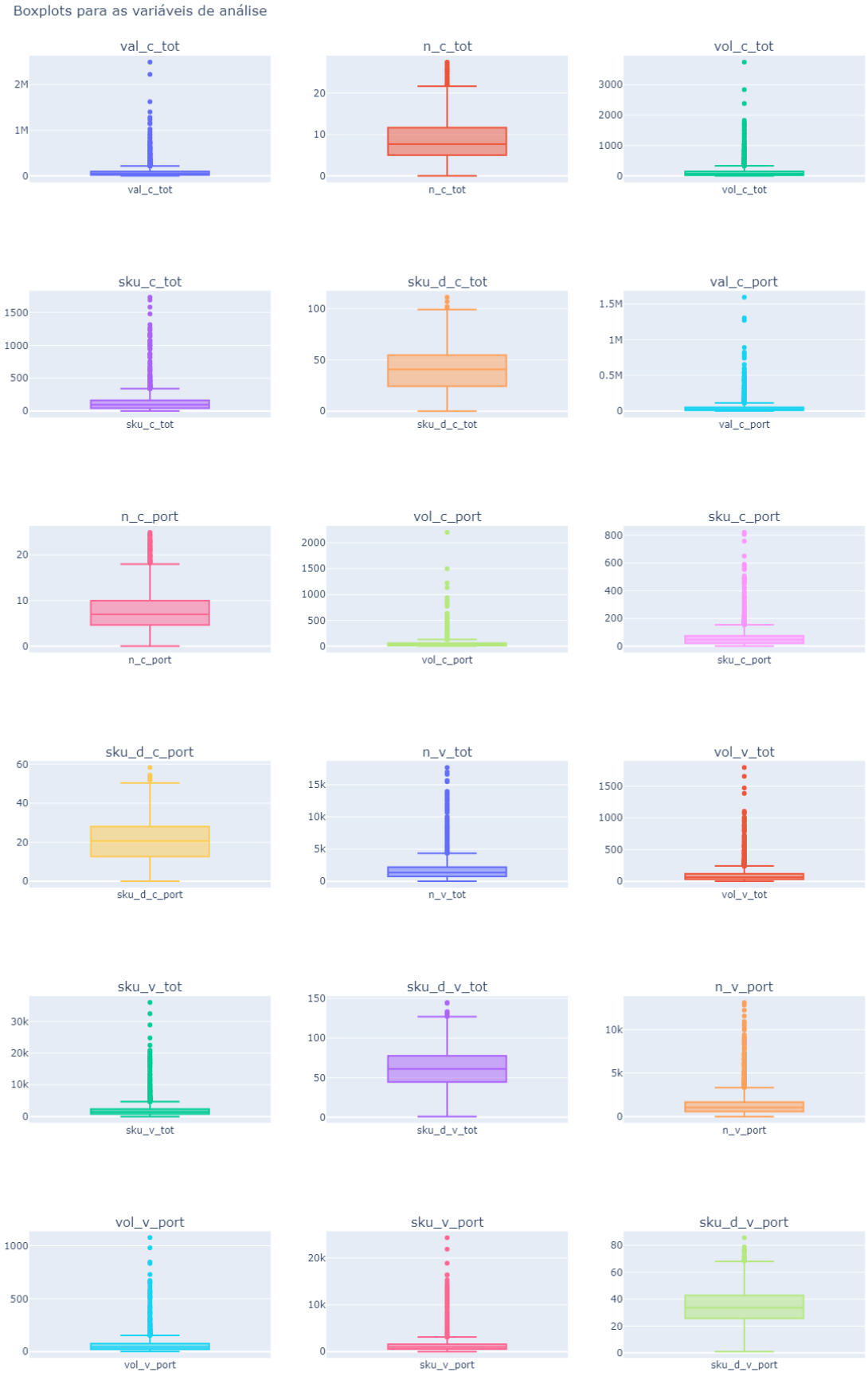
A análise dos *boxplots* das variáveis (Figura 3.7) revela algumas informações importantes sobre as distribuições de compra e venda de produtos nos estabelecimentos. Em várias variáveis, como o valor total de compras (val_c_tot) e volume total de compras (vol_c_tot), pode-se observar a presença de muitos *outliers*, indicando que há uma minoria de estabelecimentos que realiza compras significativamente maiores do que a média, enquanto a maioria dos pontos está concentrada em valores menores.

A maioria das variáveis apresenta distribuições assimétricas, com caudas longas superiores (indicando alguns valores muito altos). Por exemplo, no número total de vendas (n_v_tot), nota-se uma grande concentração de valores em torno da mediana, mas alguns pontos que se distanciam bastante, revelando que alguns poucos estabelecimentos possuem número de vendas muito maiores. A variável relativa a quantidade distinta de

SKUs comprados (sku_d_c_tot) também mostra uma distribuição ampla, sugere que o número de SKUs comprados varia significativamente entre os estabelecimentos.

No geral, esses gráficos mostram uma tendência de variabilidade considerável entre os estabelecimentos, com um pequeno número que realiza operações em volumes ou valores muito maiores que a maioria.

Figura 3.7: ***Boxplot*** **das variáveis extraídas da base**

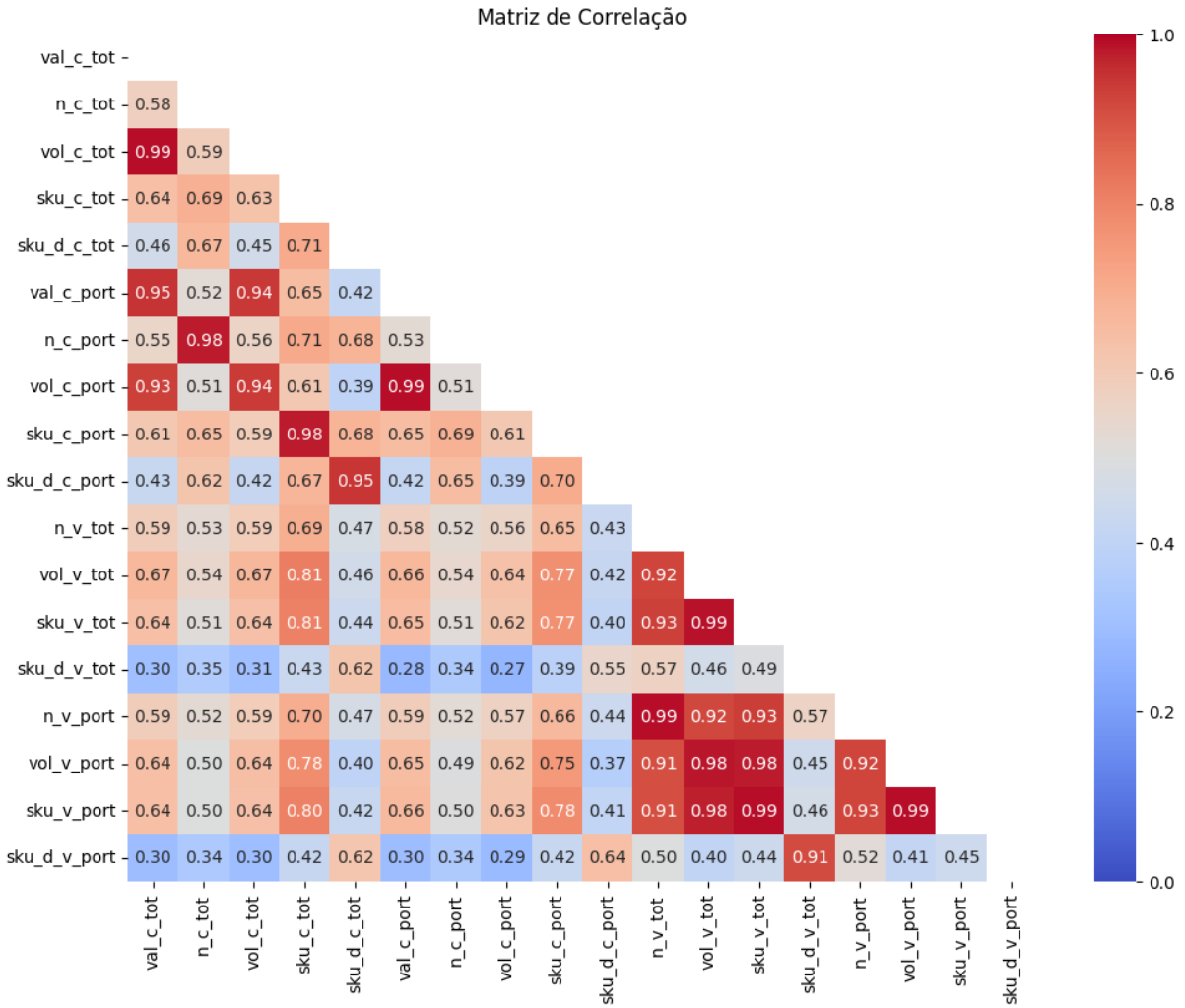


Fonte: Elaboração própria

3.2.3 Correlação

Conforme descrito na Seção 2.2.3, analisar a correlação entre variáveis é importante em processos de mineração de dados, uma vez que uma correlação alta pode indicar redundância de informações, o que pode distorcer o desempenho de algoritmos de agrupamento, que agrupam pontos com base em características comuns. Se variáveis correlacionadas forem incluídas no modelo, elas podem prejudicar a qualidade do agrupamento. Portanto, identificar e tratar essas correlações, eliminando variáveis redundantes ou transformando os dados, melhora a precisão do modelo. Para a base de dados utilizada para este trabalho, a correlação entre as variáveis está representada na Figura 3.8.

Figura 3.8: **Matriz de correlação das variáveis originais**



Fonte: Elaboração própria

Na análise da matriz de correlação, identificou-se que vários pares de variáveis

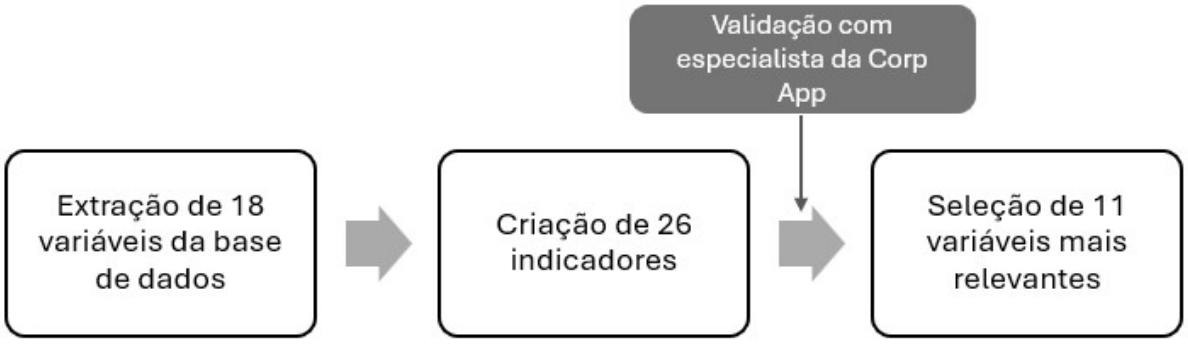
apresentam uma correlação superior a 0,9. Essa alta correlação entre as variáveis que representam a quantidade de produtos do portfólio e a respectiva quantidade total de produtos já era esperada, uma vez que, quanto maior a quantidade total de produtos comprados, mais provável é que haja uma maior presença de produtos do portfólio. Além disso, a relação positiva entre o valor pago nas compras e o volume adquirido também era antecipada, o que reflete o comportamento típico de aumento de custo com o aumento da quantidade comprada.

3.3 Pré-processamento

3.3.1 Seleção das Variáveis

O processo de seleção das variáveis a serem utilizadas no modelo de agrupamento iniciou-se com a extração de 18 variáveis numéricas da base de dados da Corporação S.A., conforme a Tabela 3.2, todas relacionadas aos produtos alcoólicos comercializados pelos estabelecimentos. Considerando a estratégia da empresa, foi realizada uma transformação dessas variáveis, o que resultou na criação de 26 novos indicadores. Dessa forma, o conjunto total passou a contar com 44 variáveis. Essas variáveis foram, então, analisadas por um representante da equipe responsável pelos algoritmos do Corp App, que utilizará os resultados deste trabalho. Após discussões sobre a importância estratégica de cada uma, 11 variáveis foram escolhidas por serem mais relevantes para caracterizar os estabelecimentos com informações importantes para a estratégia da área. Essa seleção visou garantir que o modelo de agrupamento oferecesse *insights* alinhados com os objetivos da empresa.

Figura 3.9: **Processo de seleção das variáveis para o modelo de agrupamento**



Fonte: Elaboração própria

Ao final do processo, ilustrado no fluxagrama
rvao da Figura 3.9, as onze variáveis selecionadas são descritas na Tabela 3.6.

Tabela 3.6: **Variáveis selecionadas para a modelagem**

Rótulo	**Indicador**	**Unidade**	**Descrição**
vol_med_n_c_tot	=vol_c_tot/n_c_tot	hectolitro/compra	Volume médio por compra. Como existe uma equipe da Corporação S.A. dedicada apenas a questões logísticas, essa é uma informação importante.
sku_med_n_c_tot	=sku_c_tot/n_c_tot	hectolitro/compra	Quantidade média de SKUs por compra.
por_vol_port_c_tot	=vol_c_port/vol_c_tot	-	Proporção do volume de SKUs do portfólio em relação ao volume total comprado.
vol_med_n_v_tot	=vol_v_tot/n_v_tot	hectolitro/venda	Volume médio por venda.
por_vol_port_v_tot	=vol_v_port/vol_v_tot	-	Proporção do volume de SKUs do portfólio em relação ao volume total vendido.
vol_c_v_tot	=vol_c_tot/vol_v_tot	-	Volume comprado em relação ao volume vendido (valores abaixo de 1 indicam que há compras de outros fornecedores que não a Corporação S.A.).
sku_d_c_v_port	=sku_d_c_port/sku_d_v_port	-	Quantidade de SKUs distintos do portfólio comprados em relação à quantidade de SKUs distintos do portfólio vendidos (valores abaixo de 1 indicam que há compras de outros fornecedores que não a Corporação S.A.).
n_c_tot	-	compra	Número médio de compras feitas pelo estabelecimento no App Corp por mês que continham pelo menos um produto alcoólico.
vol_c_tot	-	hectolitro	Volume médio total dos produtos alcoólicos comprados pelo estabelecimento no App Corp por mês.
sku_c_tot	-	SKU	Quantidade média de SKUs total comprados pelo estabelecimento no App Corp por mês na categoria de produtos alcoólicos.
vol_v_tot	-	hectolitro	Volume médio total de produtos alcoólicos vendidos pelo estabelecimento no Corp Delivery por mês.

Fonte: Elaboração própria

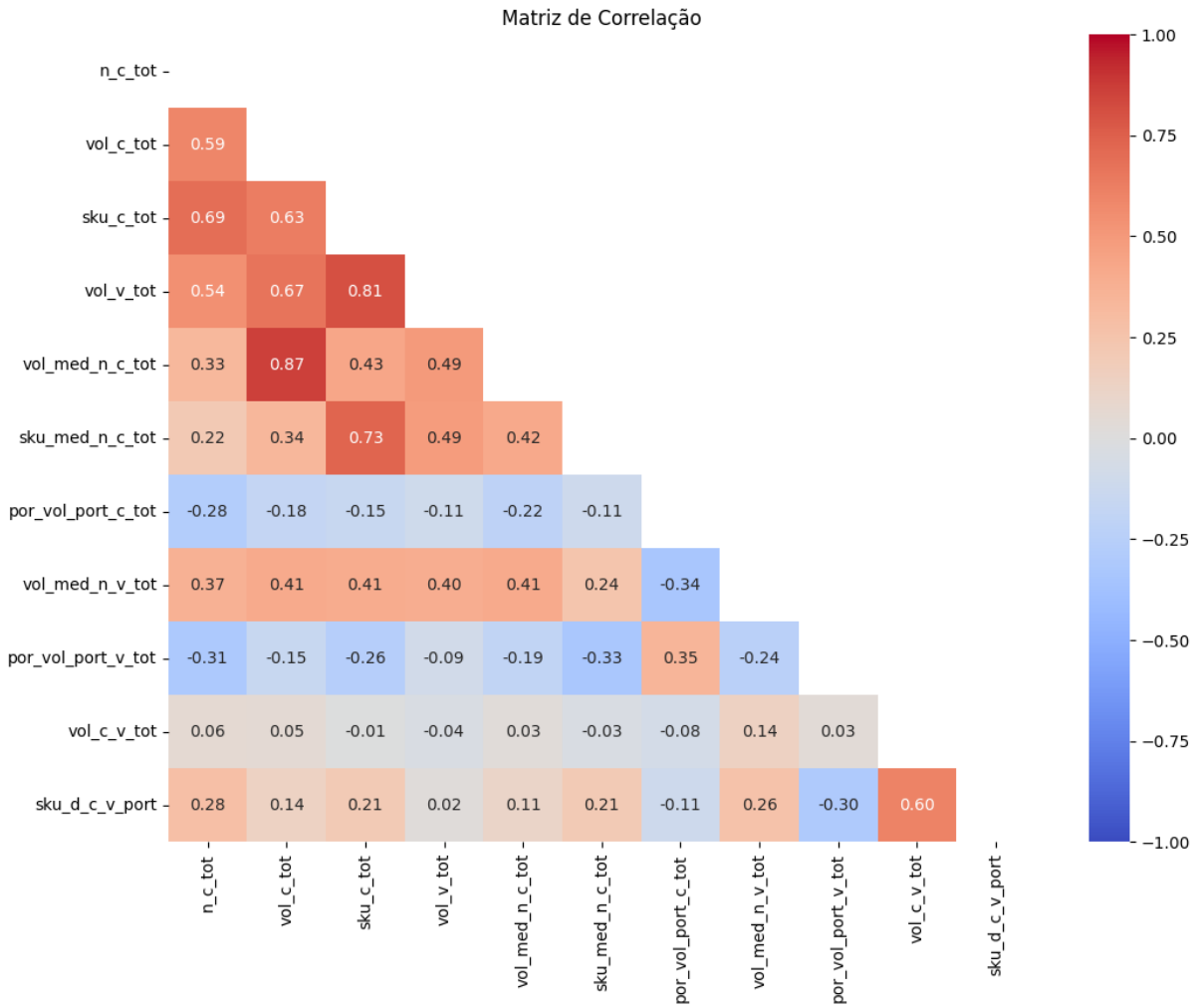
Após a seleção dos indicadores, foi elaborada uma nova matriz de correlação para verificar a existência de correlações altas, que podem indicar redundância de informações. Pode-se notar pela matriz (Figura 3.10) a redução drástica no número de pares com alta correlação em relção às variáveis originalmente coletadas (Figura 3.8) e que nehum par apresenta correlação superior a 0,9.

Apesar da alta correlação entre as variáveis de volume total de compra (vol_c_tot) e de volume médio por compra (vol_med_n_c_tot) e apresentarem alta correlação (0,87), elas contêm informações valiosas que são interpretadas de forma diferente na análise. O volume total de compra indica o tamanho do estabelecimento como cliente da Corporação S.A., enquanto que o volume médio por compra indica a demanda logística pela empresa.

Dessa forma, optou-se por manter as duas variáveis na análise.

Outro caso de alta correlação (0,81) é o das variáveis de volume total de venda (vol_v_tot) e SKUs totais na compra (shu_c_tot). Apesar do valor alto, as informações contidas nessas duas varíaveis são distintas e não apresentam redundância para a análise. Assim, ambas foram mantidas para a aplicação do modelo.

Figura 3.10: **Matriz de correlação para as variáveis selecionadas**



Fonte: Elaboração própria

3.3.2 Análise da Simetria das Distribuições

Ao utilizar algoritmos de agrupamento baseados em medidas de distância, é importante que a distribuição dos dados tenha uma forma mais próxima de uma distribuição normal por diversas razões, conforme descrito na Seção 2.2.4. Primeiramente, muitos desses algoritmos dependem de medidas de distância, como a distância euclidiana, para

agrupar pontos de dados semelhantes. Essas medidas pressupõem que os dados estão distribuídos de maneira relativamente uniforme ao redor dos centros dos grupos. Se os dados forem assimétricos ou contiverem valores extremos, as métricas de distância tornam-se menos confiáveis, e os resultados do agrupamento podem não refletir a verdadeira estrutura dos dados.

Além disso, algoritmos como k-médias assumem que os grupos têm uma forma aproximadamente esférica ou elíptica. Uma distribuição normal, por ser simétrica, ajusta-se bem a essa suposição. Quando os dados são fortemente assimétricos, o algoritmo pode forçar a formação de grupos que não fazem sentido ou pode não captar padrões relevantes. Também é importante considerar o equilíbrio de influência dos dados, uma vez que, em distribuições próximas à normal, os pontos de dados tendem a estar mais próximos da média, com menos valores discrepantes que possam distorcer os resultados. Dados assimétricos ou com alta variabilidade podem levar à formação de grupos desproporcionalmente influenciados por valores extremos, tornando o agrupamento menos preciso para a maioria dos pontos de dados.

Outro aspecto relevante é a interpretabilidade dos resultados. Se os dados seguem uma distribuição normal, é mais fácil interpretar os centros dos grupos e a dispersão de cada um. Em contrapartida, quando os dados são altamente assimétricos ou apresentam caudas longas, os centróides dos grupos podem ser menos significativos e mais difíceis de interpretar. A eficiência dos algoritmos também é beneficiada por uma distribuição mais normalizada, já que muitos algoritmos de agrupamento funcionam de maneira mais eficiente nessas condições. Dados assimétricos ou desbalanceados podem retardar a convergência ou levar a soluções subótimas.

Para este trabalho, foi realizada uma transformação dos dados de cada variável por meio de duas técnicas: a transformação logarítmica de base 10 e a transformação pela raiz cúbica. O objetivo dessas transformações foi reduzir a assimetria e a curtose das distribuições originais, tornando-as mais próximas de uma distribuição normal.

A transformação logarítmica é amplamente utilizada para reduzir a amplitude dos dados, especialmente em casos onde há grandes variações nos valores ou quando a distribuição original apresenta forte assimetria positiva. Para evitar indeterminações no cálculo do logaritmo de zero, os valores iguais a zero nas variáveis foram substituídos por 0,01, uma constante positiva arbitrária, mas suficientemente pequena para não interferir

significativamente na distribuição dos dados. A transformação aplicada foi a seguinte:

$$\mathbf{x}'_i = \log_{10}(\mathbf{x}_i) \tag{3.1}$$

Onde:

- $\mathbf{x}_i$ representa os valores originais da variável,
- $\mathbf{x}'_i$ representa os valores transformados,

A segunda transformação aplicada foi a raiz cúbica dos dados. Esta transformação é útil tanto para dados com assimetria positiva quanto negativa, uma vez que ela não requer que os valores sejam estritamente positivos. A transformação pela raiz cúbica suaviza as variações extremas e pode trazer distribuições altamente assimétricas para um formato mais normalizado. A fórmula utilizada para a transformação foi:

$$\mathbf{x}'_i = \sqrt[3]{\mathbf{x}_i} \tag{3.2}$$

Essa transformação é menos agressiva do que a logarítmica e é mais adequada para variáveis que possuem tanto valores positivos quanto negativos.

Para cada variável, foram gerados três histogramas: um correspondente aos dados originais, outro após a transformação logarítmica e o terceiro após a transformação pela raiz cúbica. Essa abordagem permitiu a comparação visual das distribuições antes e depois das transformações, facilitando a avaliação da normalidade dos dados e auxiliando na escolha da transformação mais apropriada para as análises subsequentes.

3.3.3 Tratamento de *Outliers*

O tratamento de *outliers* é necessário para evitar distorções nos resultados do agrupament, pois podem influenciar negativamente a formação dos grupos, deslocando os centroides e levando a grupos que não representam adequadamente os dados. Ainda, algoritmos de agrupamento, como o k-médias, que utilizam a distância euclidiana como métrica são sensíveis a valores extremos, o que pode aumentar a variância dentro dos grupos e dificultar a interpretação dos grupos.

Neste trabalho, o tratamento de outliers foi realizado por meio do truncamento das distribuições com limites baseados nos percentis de 2,5% e 97,5% da distribuição dos

dados. Essa abordagem visa limitar os valores extremos, que podem distorcer a análise, sem remover observações. O limite inferior foi ajustado para o valor do percentil de 2,5%, enquanto o limite superior foi ajustado para o percentil de 97,5%. Dessa forma, os valores abaixo do limite inferior foram elevados ao valor correspondente a esse percentil, e os valores acima do limite superior foram reduzidos ao valor correspondente ao percentil de 97,5%. Esse método preserva a maior parte dos dados originais, garantindo que os outliers não influenciem negativamente a formação de grupos no modelo k-médias, ao mesmo tempo em que mantém uma maior representatividade da distribuição geral.

3.3.4 Padronização

Antes da aplicação de modelos de agrupamento, é necessário garantir que os dados estejam em uma escala adequada para evitar distorções nos resultados. Nesse contexto, a normalização por padronização é uma técnica utilizada para ajustar as distribuições das variáveis, conforme descrito na Seção 2.2.4.

A padronização transforma os dados para que cada variável tenha uma média igual a zero e um desvio padrão igual a um. Essa técnica é importante quando as variáveis possuem unidades ou escalas diferentes. Em muitos algoritmos de agrupamento, como o k-médias, a distância euclidiana é usada para medir a similaridade entre os pontos de dados. Sem a padronização, variáveis com magnitudes maiores podem influenciar o resultado, fazendo com que variáveis de menor escala tenham pouca relevância no processo de agrupamento. A padronização garante que cada variável tenha o mesmo peso no cálculo das distâncias. Assim, foi aplicada a padronização dos dados para cada variável e tipo de distribuição.

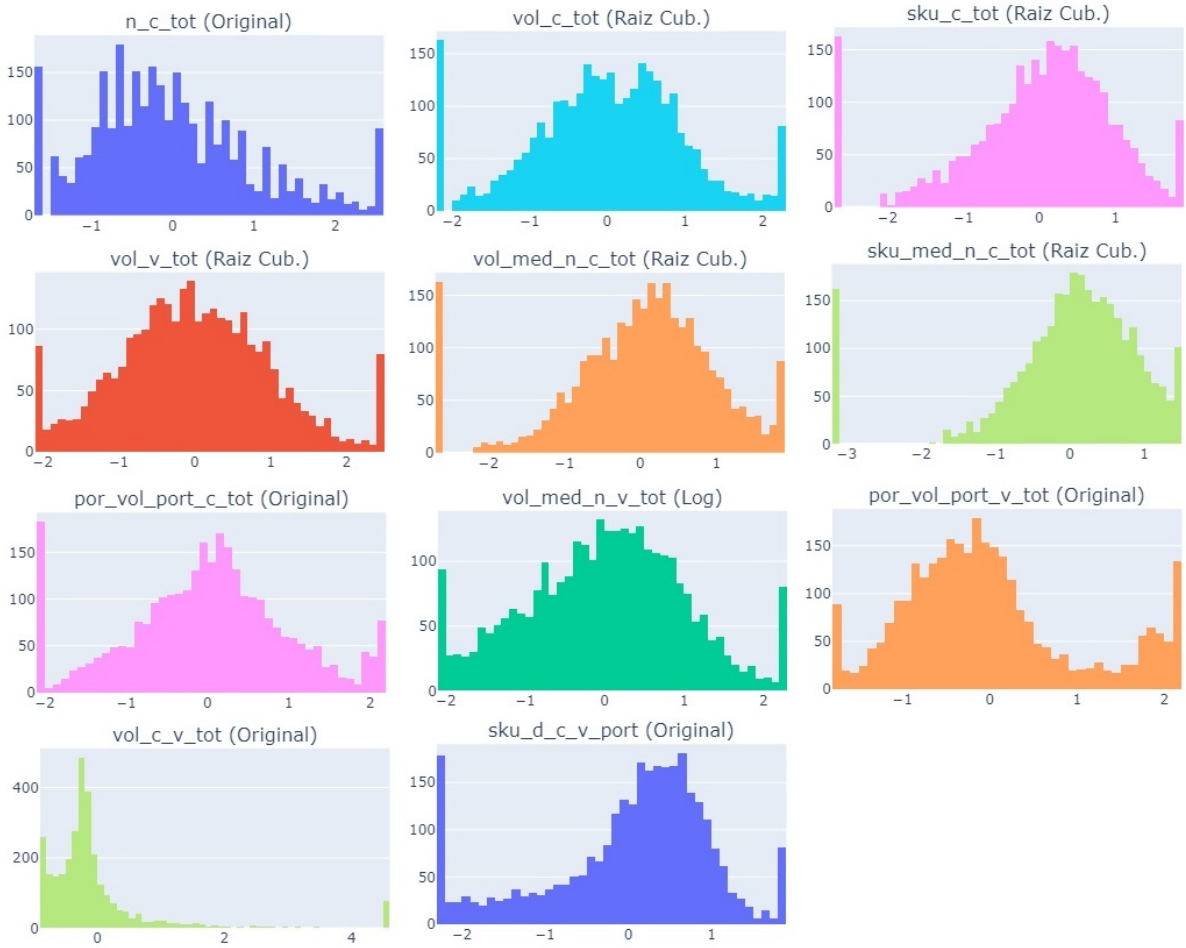
Após a análise visual do formato das distribuições, para cada variável foi selecionada aquela que mais se aproximava de uma distribuição simétrica, conforme indicado na Tabela 3.7. As distribuições escolhidas podem ser visualizadas na Figura 3.11.

Tabela 3.7: **Transformações selecionadas por variável**

Variável	**Transformação Selecionada**
n_c_tot	Sem transformação
vol_c_tot	Transformação pela raiz cúbica
sku_c_tot	Transformação pela raiz cúbica
vol_v_tot	Transformação pela raiz cúbica
vol_med_n_c_tot	Transformação pela raiz cúbica
sku_med_n_c_tot	Transformação pela raiz cúbica
por_vol_port_c_tot	Sem transformação
vol_med_n_v_tot	Transformação logarítmica
por_vol_port_v_tot	Sem transformação
vol_c_v_tot	Transformação pela raiz cúbica
sku_d_c_v_port	Sem transformação

Fonte: Elaboração própria

Figura 3.11: **Distribuições padronizadas escolhidas para cada variável**



Fonte: Elaboração própria

Capítulo 4

RESULTADOS DA ANÁLISE DE AGRUPAMENTOS

No presente capítulo são apresentados os resultados obtidos a partir da análise realizada. Primeiramente, aborda-se a modelagem dos dados (Seção 4.1), com foco na aplicação do algoritmo k-médias para identificação de agrupamentos distintos entre os estabelecimentos. São descritos os gráficos utilizados para avaliação do número ideal de grupos, bem como o processo de seleção do valor de k mais adequado. Na sequência, realiza-se a avaliação dos grupos resultantes (Seção 4.2), com a caracterização detalhada de cada grupo e a sugestão de ações estratégicas específicas para potencializar as práticas comerciais de acordo com o comportamento descrito.

4.1 Formação dos Grupos

Para a obtenção dos resultados da modelagem, aplicou-se o algoritmo k-médias (conforme justificado na Seção 2.3.1), variando o número de grupos k de 2 a 10. A escolha dessa faixa ocorreu devido à necessidade de definir uma quantidade de grupos que permitisse a identificação de perfis distintos de comportamento dos estabelecimentos analisados. Entretanto, um número muito elevado de grupos poderia dificultar a implementação de ações específicas.

Para cada valor de k, foram calculadas três métricas de avaliação de qualidade dos agrupamentos, conforme descrito na Seção 2.3.1:

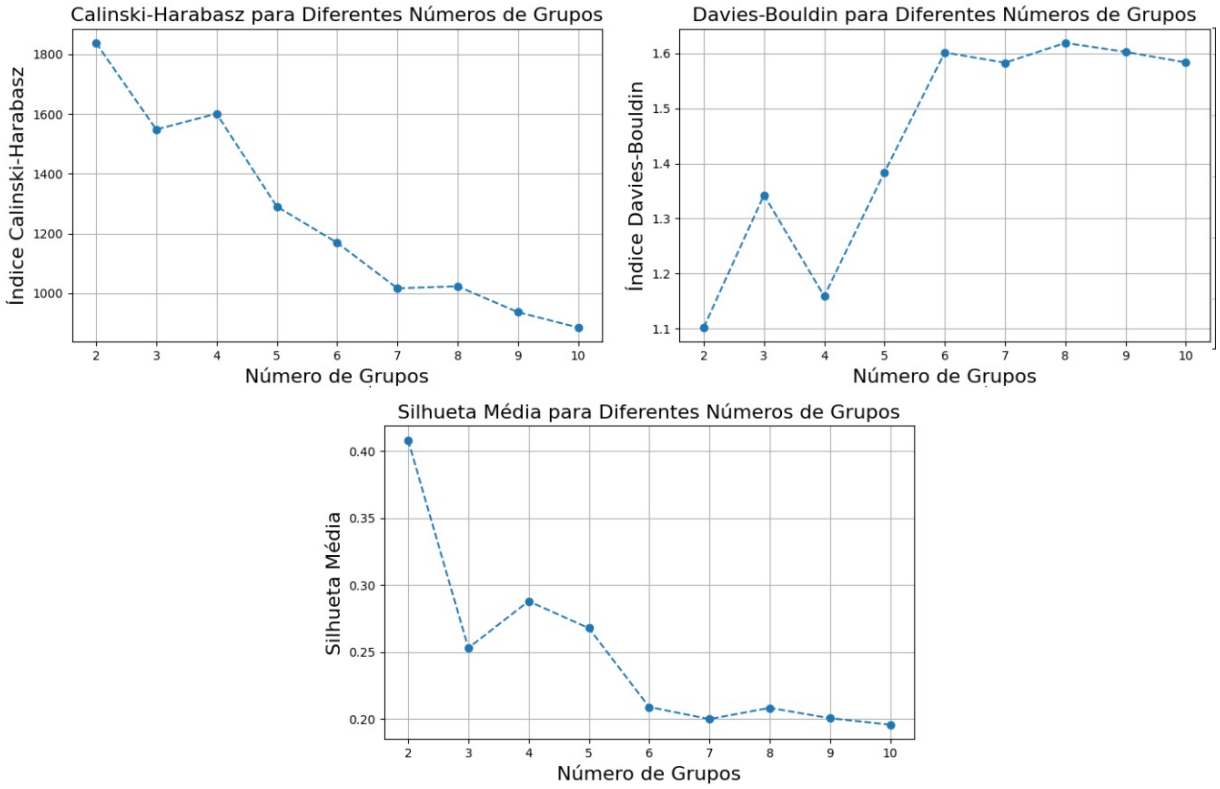
1. **Índice de Calinski-Harabasz**: avalia a relação entre a dispersão dos dados dentro

dos grupos e a distância entre os grupos, com valores mais altos indicando uma melhor estruturação.

2. **Índice de Davies-Bouldin**: mede a similaridade entre os grupos, com valores mais baixos indicando uma maior separação entre os grupos.

3. **Coeficiente de Silhueta**: mede a separação entre os grupos e a coesão dentro de cada grupo. Valores mais altos indicam uma melhor separação entre os grupos.

Figura 4.1: **Índices de validação para agrupamentos utilizando k-médias**



Fonte: Elaboração própria

Tabela 4.1: **Melhores resultados para os índices de validação**

Número de Grupos (k)	**2**	**4**
Índice Calinski-Harabasz	1838,07	1601,12
Índice Davies-Bouldin	1,1	1,16
Silhueta Média	0,41	0,29

Fonte: Elaboração própria

Após a análise da Figura 4.1, verificou-se que $k = 2$ apresentou os melhores resultados nas três métricas, seguido por $k = 4$ (Tabela 4.1). Optou-se por seguir com $k = 4$, pois esse valor oferece uma segmentação mais detalhada e permite uma maior diferenciação entre os grupos, sem gerar uma quantidade de grupos que inviabilize a execução das estratégias comerciais. A quantidade de estabelecimentos selecionados para cada grupo está indicada na Tabela 4.2.

Tabela 4.2: **Quantidade de estabelecimentos por grupo**

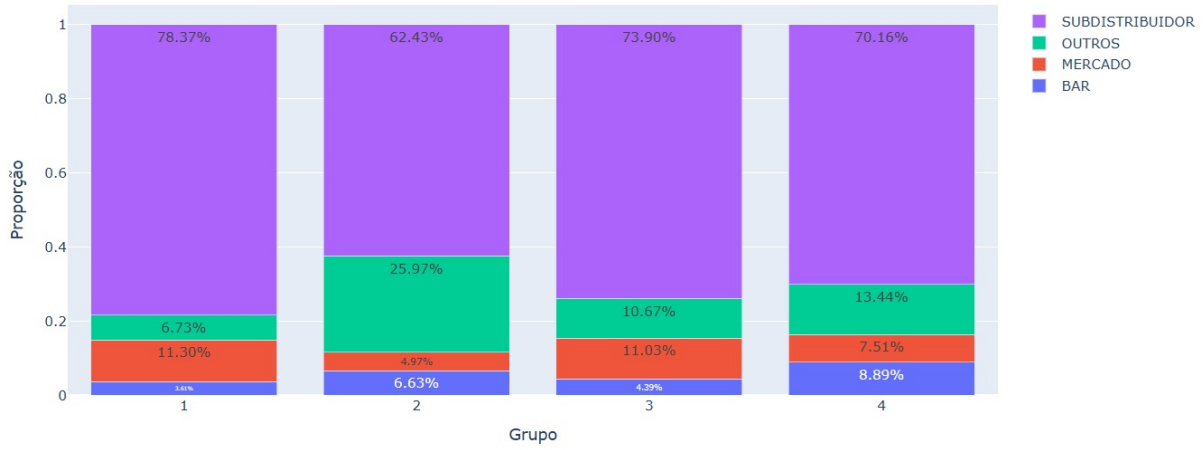
Grupo	Quantidade de estabelecimentos
1	1248
2	181
3	1115
4	506

Fonte: Elaboração própria

Após a definição dos estabelecimentos em cada grupo, foram analisadas suas distribuições em relação ao tipo de estabelecimento e região geográfica. Apesar dessas características não terem sido utilizadas como dados de entrada no algoritmo k-médias, essas informações contribuem para complementam a caracterização dos grupos.

O gráfico que retrata a proporção por **tipo de estabelecimento** (Figura 4.2) revela um predomínio notável dos subdistribuidores, com percentuais acima de 62% entre os grupos. O Grupo 1 conta com a maior representação (78,37%), enquanto o Grupo 4 apresenta 62,43% desse mesmo tipo. A categoria "Outros" (com representantes de todos os nove tipos restantes) se apresenta com uma representatividade modesta, oscilando entre 6,73% e 25,97% nos diferentes grupos, tendo um destaque maior no Grupo 2. Assim, o Grupo 2 apresenta a maior variedade de estabelecimentos proporcionamente em relação aos outros. A análise dos mercados e bares revela uma presença ainda mais limitada, não ultrapassando 12% . O Grupo 1 apresenta a maior proporção de mercados (11,30%), enquanto que o Grupo 4 apresenta a maior proporção de bares (8,89%).

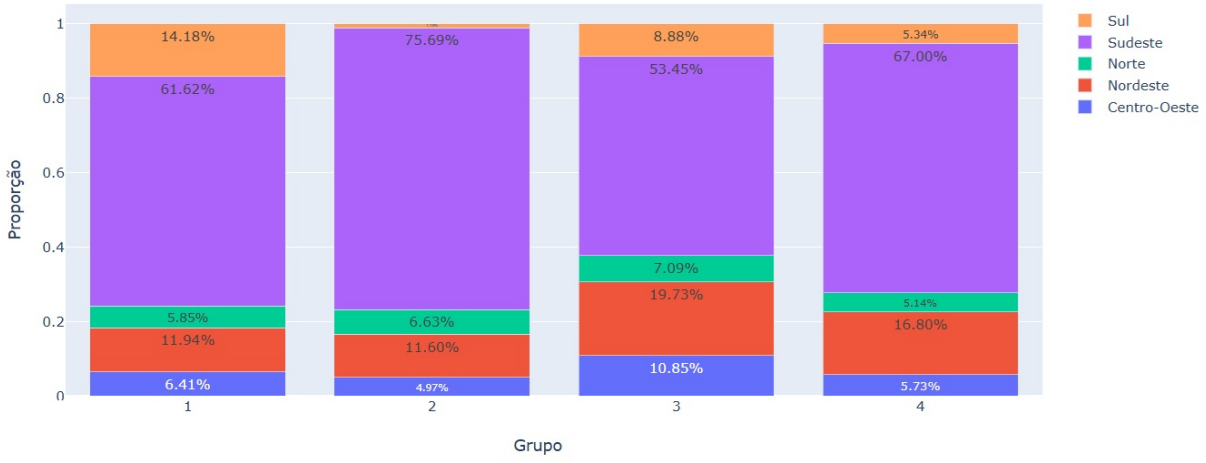
Figura 4.2: **Divisão dos grupos por tipo de estabelecimento**



Fonte: Elaboração própria

Em relação à distribuição por **região geográfica** (Figura 4.3), a região Sudeste mantém uma participação expressiva em todos os grupos, com valores superiores a 53% em todos. Deles, o Grupo 2 apresenta a maior proporção em relação a essa região (75,69%). Os Grupos 3 e 4 apresentam a maior participação da região Nordeste com, respectivamente, 19,73% e 16,80%. Por outro lado, a maior concentração de estabelecimentos localizados na região Sul está no Grupo 1 (14,18%). A participação das regiões Norte e Centro-Oeste apresenta valores modestos, com 7,09% e 10,85%, respectivamente, no Grupo 3 sendo os mais elevados. Assim, o Grupo 3 exibe uma maior diversidade de localidades dos seus estabelecimentos, enquanto o Grupo 1 está mais concentrado no Sudeste.

Figura 4.3: **Divisão dos grupos por região**



Fonte: Elaboração própria

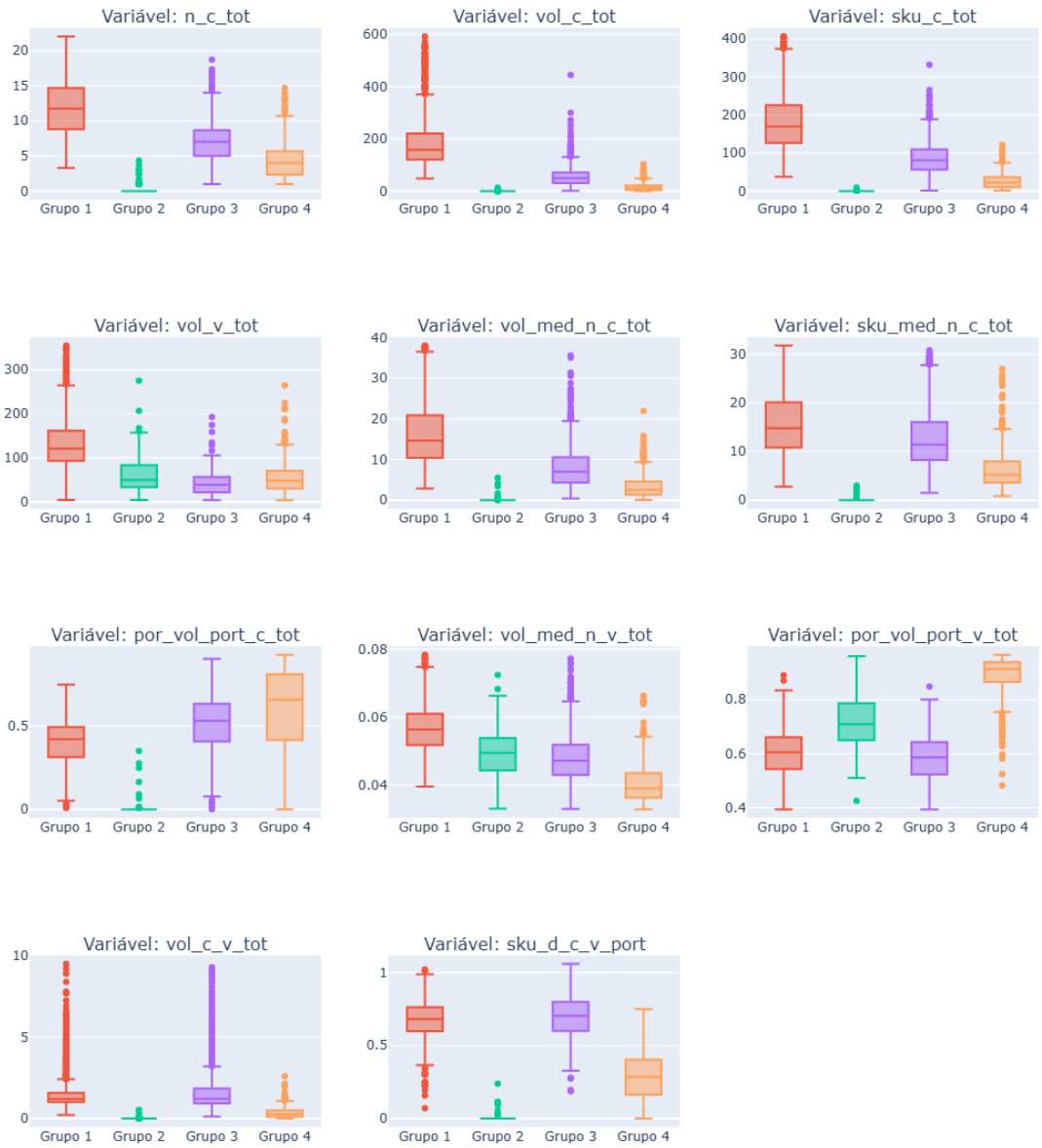
As conclusões extraídas desses gráficos evidenciam um padrão de concentração de estabelecimentos em regiões e tipos específicos. O domínio da região Sudeste e dos subdistribuidores sugere uma centralização das atividades comerciais nessas áreas, o que pode ter implicações para estratégias de marketing e distribuição da Corporação S.A. A escassez de estabelecimentos nas regiões Norte e Centro-Oeste, bem como a baixa presença de mercados e bares, aponta para oportunidades potenciais de expansão e diversificação de atuação nessas áreas. Assim, recomenda-se que futuras pesquisas e ações de marketing considerem esses dados para otimizar a alocação de recursos e o desenvolvimento de estratégias que atendam de forma mais eficaz as particularidades de cada região e tipo de estabelecimento.

4.2 Interpretação dos Perfis dos Grupos e Sugestões Estratégicas

Após a segmentação dos grupos, foi elaborada as estatísticas descritivas para cada grupo para compreender as características centrais de cada um, considerando as variáveis que influenciaram a segmentação, além dos gráficos das distribuições para ilustrar as diferenças de comportamento conforme especificado na Seção 2.3.1. Esse procedimento incluiu o cálculo de medidas como a média, mediana, desvio padrão e distribuição de frequências, como demonstrado na Tabela 4.4. Com base nos resultados obtidos, foi possível criar uma caracterização geral de cada grupo, conforme abordado na Seção 2.3.1, e sugerir ações estratégicas específicas para potencializar os resultados da Corporação

S.A..

Figura 4.4: ***Boxplot* das distribuições das variáveis por grupo**



Fonte: Elaboração própria

GRUPO 1: Comprador Grande com Baixa Aderência ao Portfólio

O grupo 1 representa grandes estabelecimentos que fazem compras frequentes e em grande volume diretamente da Corporação S.A., mas têm uma adesão relativamente baixa ao portfólio ideal da empresa. Nesse grupo também estão associados estabele-

cimentos "gigantes" com valores de compra e venda muito discrepantes da média dos estabelecimentos grandes do grupo (*outliers*), o que impacta na variação interna. Esses estabelecimentos são caracterizados por:

- **Frequência de Compras:** Fazem em média 12,78 compras por mês (variável: n_c_tot), o que demonstra uma alta regularidade nas compras.
- **Volume de Compras:** Compram 237,75 hectolitros por mês (variável: vol_c_tot), o maior volume entre os grupos.
- **Diversificação de SKUs:** Compram uma média de 216 SKUs por mês (variável: sku_c_tot), demonstrando uma grande variedade de produtos comprados.
- **Vendas:** Vendem 168,35 hectolitros por mês (variável: vol_v_tot), um volume alto, refletindo a grande capacidade de vendas.
- **Aderência ao Portfólio Ideal:** A participação do portfólio ideal nas compras é baixa, com 40% (variável: $por\_vol\_port\_c\_tot$).
- **Dependência de Outros Fornecedores:** A relação $vol\_c\_v\_tot$ de 5,93 indica que compram muito mais diretamente da Corporação S.A. do que vendem, com baixa dependência de outros fornecedores.

Considerando as características descritas acima, as ações estratégicas sugeridas para a Corporação S.A. para o grupo 1 são:

- **Promoções Focadas no Portfólio Ideal:** Criar promoções que incentivem a compra de SKUs do portfólio ideal, como descontos especiais para produtos do portfólio ou combos que aumentem a participação do portfólio nas compras.
- **Desafios de Exclusividade:** Lançar desafios que recompensem os estabelecimentos que comprarem 100% pelo Corp App, reduzindo ainda mais a dependência de outros fornecedores.
- **Foco em Aderência ao Portfólio:** Os representantes de vendas podem atuar de forma personalizada, incentivando esses estabelecimentos a diversificar os produtos do portfólio ideal e aumentar sua representatividade nas compras.

GRUPO 2: Varejista Inativo com Baixa Atividade

O grupo 2 representa médios varejistas em volume de venda que fazem pouquíssimas compras diretamente da Corporação S.A. pelo Corp App. Esses estabelecimentos têm uma adesão baixa ao portfólio de compra e mostram-se praticamente inativos na plataforma. Esses estabelecimentos são caracterizados por:

- **Frequência de Compras:** Realizam em média apenas 0,23 compras por mês (variável: n_c_tot), indicando quase nenhuma atividade no Corp App.
- **Volume de Compras:** Compram apenas 0,2 hectolitros por mês (variável: vol_c_tot), o menor volume entre os grupos.
- **Diversificação de SKUs:** Compram uma média de 0,28 SKUs por mês (variável: sku_c_tot), evidenciando baixa diversificação de produtos.
- **Vendas:** Apesar do baixo volume de compras, vendem 60,99 hectolitros por mês (variável: vol_v_tot) no *Corp Delivery*, um volume considerável e próximo do grupo 4, sugerindo que complementam suas vendas com produtos de outros fornecedores.
- **Aderência ao Portfólio Ideal:** A participação de produtos do portfólio ideal nas compras é muito baixa, com 1% (variável: $por\_vol\_port\_c\_tot$), porém a participação de produtos do portfólio nas vendas é alta, com 71% (variável: $por\_vol\_port\_v\_tot$), demonstrando a relevância desses produtos nas vendas e oportunidade de aumentar a participação nas compras.
- **Dependência de Outros Fornecedores:** A relação $vol\_c\_v\_tot$ de 0,01 indica que quase todas as vendas são abastecidas por fornecedores externos à Corporação S.A..

Considerando as características descritas acima, as ações estratégicas sugeridas para a Corporação S.A. para o grupo 2 são:

- **Campanhas de Reativação:** Promoções agressivas podem incentivar esses estabelecimentos a comprar mais diretamente via Corp App, com descontos em primeiras compras e ofertas de reativação.
- **Educação sobre o Portfólio:** Os representantes de vendas podem ser encarregados de realizar visitas focadas em educar os proprietários sobre o valor do portfólio

ideal e como ele pode beneficiar o estabelecimento em termos de margem e diversidade de produtos.

- **Promoções de Frequência:** Incentivar a frequência de compras com desafios mensais que recompensem o aumento da frequência de pedidos, como prêmios ou descontos progressivos.

GRUPO 3: Médio Comprador com Aderência Razoável ao Portfólio

O grupo 3 representa pequenos estabelecimentos que compram diretamente da Corporação S.A., mas têm uma alta adesão ao portfólio ideal, indicando uma maior fidelidade à empresa. Esses estabelecimentos são caracterizados por:

- **Frequência de Compras:** Fazem em média 7,19 compras por mês (variável: n_c_tot), uma frequência alta considerando o volume comprado.
- **Volume de Compras:** Compram 54,81 hectolitros por mês (variável: vol_c_tot), um volume moderado.
- **Diversificação de SKUs:** Compram uma média de 86 SKUs por mês (variável: sku_c_tot), sugerindo uma boa diversidade de produtos comprados.
- **Vendas:** Vendem 38,72 hectolitros por mês (variável: vol_v_tot).
- **Aderência ao Portfólio Ideal:** A participação de produtos do portfólio ideal nas compras é razoável, com 52% (variável: $por\_vol\_port\_c\_tot$), semelhante ao percentual das vendas (57%).
- **Dependência de Outros Fornecedores:** A relação $vol\_c\_v\_tot$ de 7,66 indica que esses estabelecimentos compram muito mais diretamente da Corporação S.A. do que vendem, mostrando baixa dependência de outros fornecedores.

Considerando as características descritas acima, as ações estratégicas sugeridas para a Corporação S.A. para o grupo 3 são:

- **Desafios para Aumentar a Diversificação:** Criar desafios que incentivem esses estabelecimentos a expandirem a diversificação de SKUs comprados e aumentarem o volume de compras.

- **Promoções de Fidelidade:** Como esses estabelecimentos já têm uma alta adesão ao portfólio, a empresa pode oferecer descontos e recompensas em SKUs específicos do portfólio para aumentar ainda mais a participação nas compras.

- **Apoio de Representantes para Expansão:** Os representantes podem auxiliar com estratégias de expansão de volume e portfólio, sugerindo oportunidades de crescimento no número de SKUs adquiridos.

GRUPO 4: Comprador de Pequeno Porte com Adesão Alta ao Portfólio

O grupo 4 apresenta estabelecimentos que fazem compras em quantidades intermediárias, com uma adesão razoável ao portfólio ideal e compras moderadas diretamente da Corporação S.A.. Esses estabelecimentos são caracterizados por:

- **Frequência de Compras:** Realizam em média 4,37 compras por mês (variável: n_c_tot), o que demonstra uma frequência moderada.

- **Volume de Compras:** Compram 15,18 hectolitros por mês (variável: vol_c_tot), posicionando-se como compradores pequenos.

- **Diversificação de SKUs:** Compram uma média de 26 SKUs por mês (variável: sku_c_tot).

- **Vendas:** Vendem 54,21 hectolitros por mês (variável: vol_v_tot), um volume significativo.

- **Aderência ao Portfólio Ideal:** A participação de produtos do portfólio nas compras é de 65% (variável: $por\_vol\_port\_c\_tot$), evidenciando uma adesão alta. Entretanto, a média da porcentagem de adesão dos produtos vendidos ao portfólio é de 90% (por_vol_port_v_tot), indicando que o valor dos produtos comprados pode aumentar.

- **Dependência de Outros Fornecedores:** A relação $vol\_c\_v\_tot$ de 0,35 indica que esses estabelecimentos complementam boa parte de suas vendas com outros fornecedores.

Considerando as características descritas acima, as ações estratégicas sugeridas para a Corporação S.A. para o grupo 4 são:

- **Promoções de Aumento de Volume:** Oferecer descontos progressivos para estimular esses estabelecimentos a aumentarem o volume de compras.
- **Educação e Expansão do Portfólio:** Representantes de vendas podem realizar visitas focadas em educar sobre o portfólio ideal e sugerir SKUs adicionais que podem complementar o mix de produtos, incentivando a diversificação.
- **Incentivos para Aumentar a Frequência:** Criar desafios que recompensem o aumento da frequência de compras no Corp App e premiem quem mantiver uma regularidade alta de pedidos.

Tabela 4.3: **Resumo das características dos grupos e estratégias sugeridas**

Grupo	**Características Principais**	**Estratégias Sugeridas**
Grupo 1: Comprador Grande com Baixa Aderência ao Portfólio	- Alta frequência de compras (12,78 compras/mês) - Maior volume de compras entre os grupos (237,75 hl/mês) - Baixa adesão ao portfólio ideal (40%) - Alta dependência da Corporação (vol_c_v_tot = 5,93)	- Promoções focadas no portfólio ideal para incentivar adesão - Desafios de exclusividade para compras 100% pelo Corp App - Ações de incentivo à diversificação do portfólio por representantes de vendas
Grupo 2: Varejista Inativo com Baixa Atividade	- Frequência de compras muito baixa (0,23 compras/mês) - Volume de compras insignificante (0,2 hl/mês) - Baixa diversificação de SKUs (0,28 SKUs/mês) - Alta dependência de fornecedores externos (vol_c_v_tot = 0,01)	- Campanhas de reativação com promoções e ofertas iniciais - Educação sobre o portfólio ideal realizada por representantes - Promoções de frequência para estimular compras regulares
Grupo 3: Médio Comprador com Aderência Razoável ao Portfólio	- Frequência de compras moderada (7,19 compras/mês) - Volume de compras moderado (54,81 hl/mês) - Aderência ao portfólio ideal razoável (52%) - Boa diversificação de SKUs (86 SKUs/mês)	- Desafios para aumentar diversificação e volume de compras - Promoções de fidelidade para aumentar participação do portfólio - Suporte de representantes para expansão de volume e portfólio
Grupo 4: Comprador de Pequeno Porte com Adesão Alta ao Portfólio	- Frequência moderada de compras (4,37 compras/mês) - Baixo volume de compras (15,18 hl/mês) - Alta adesão ao portfólio ideal nas vendas (90%) - Alta dependência de fornecedores externos para complementação de vendas	- Promoções de aumento de volume com descontos progressivos - Educação sobre portfólio para diversificar mix de produtos - Incentivos para aumentar a frequência de compras

Fonte: Elaboração própria

Tabela 4.4: **Estatística descritiva dos grupos**

Grupo	Variável	Média	Desvio Padrão	Valor mínimo	Quartil 1 (25%)	Quartil 2 (50%)	Quartil 3 (75%)	Valor máximo
1	n_c_tot	12,78	4,94	3,30	9,17	12,10	15,77	27,50
2	n_c_tot	0,23	0,68	0,00	0,00	0,00	0,00	4,33
3	n_c_tot	7,19	2,82	1,00	5,00	7,00	8,67	24,67
4	n_c_tot	4,37	2,62	1,00	2,33	4,00	5,67	14,67
1	vol_c_tot	237,75	253,80	47,93	122,96	163,53	241,01	3.749,53
2	vol_c_tot	0,21	1,15	0,00	0,00	0,00	0,00	12,63
3	vol_c_tot	54,81	36,58	0,53	30,50	49,17	70,88	444,14
4	vol_c_tot	15,18	14,89	0,04	4,01	11,08	21,60	102,46
1	sku_c_tot	215,56	175,53	37,20	129,20	176,40	240,00	1.734,40
2	sku_c_tot	0,28	1,08	0,00	0,00	0,00	0,00	9,00
3	sku_c_tot	86,35	41,92	1,50	56,50	81,33	109,33	331,50
4	sku_c_tot	26,37	19,67	1,00	10,33	22,67	36,58	120,67
1	vol_v_tot	168,35	161,88	0,14	94,47	124,85	176,54	1.795,70
2	vol_v_tot	60,99	43,59	1,26	31,89	49,11	83,03	274,61
3	vol_v_tot	38,72	24,66	0,02	19,22	36,66	55,05	192,84
4	vol_v_tot	54,21	36,57	1,09	29,13	47,23	69,63	264,55
1	vol_med_n_c_tot	18,61	13,91	2,86	10,65	15,31	22,30	182,61
2	vol_med_n_c_tot	0,11	0,60	0,00	0,00	0,00	0,00	5,52
3	vol_med_n_c_tot	8,14	5,63	0,40	4,32	7,05	10,56	88,83
4	vol_med_n_c_tot	3,35	2,94	0,04	1,33	2,56	4,58	21,96
1	sku_med_n_c_tot	17,13	8,73	2,75	11,05	15,36	21,13	68,84
2	sku_med_n_c_tot	0,15	0,49	0,00	0,00	0,00	0,00	3,00
3	sku_med_n_c_tot	12,83	6,28	1,50	8,25	11,43	16,20	43,67
4	sku_med_n_c_tot	6,25	4,13	0,85	3,59	5,24	8,00	27,00
1	por_vol_port_c_tot	0,40	0,13	0,01	0,31	0,42	0,49	0,75
2	por_vol_port_c_tot	0,01	0,04	0,00	0,00	0,00	0,00	0,35
3	por_vol_port_c_tot	0,52	0,17	0,00	0,41	0,53	0,63	0,95
4	por_vol_port_c_tot	0,65	0,27	0,00	0,45	0,71	0,90	1,00
1	vol_med_n_v_tot	0,06	0,01	0,04	0,05	0,06	0,06	0,16
2	vol_med_n_v_tot	0,05	0,01	0,03	0,04	0,05	0,05	0,17
3	vol_med_n_v_tot	0,05	0,01	0,01	0,04	0,05	0,05	0,18
4	vol_med_n_v_tot	0,04	0,01	0,02	0,03	0,04	0,04	0,11
1	por_vol_port_v_tot	0,60	0,09	0,09	0,54	0,60	0,66	0,89
2	por_vol_port_v_tot	0,71	0,12	0,23	0,64	0,71	0,79	1,00
3	por_vol_port_v_tot	0,57	0,09	0,20	0,52	0,58	0,64	0,85
4	por_vol_port_v_tot	0,90	0,08	0,48	0,88	0,92	0,95	1,00
1	vol_c_v_tot	5,93	106,17	0,22	1,02	1,20	1,64	3.205,10
2	vol_c_v_tot	0,01	0,05	0,00	0,00	0,00	0,00	0,54
3	vol_c_v_tot	7,66	93,52	0,11	0,94	1,25	2,04	2.938,99
4	vol_c_v_tot	0,35	0,33	0,00	0,11	0,27	0,50	2,60
1	sku_d_c_v_port	0,69	0,37	0,07	0,60	0,68	0,76	10,00
2	sku_d_c_v_port	0,00	0,02	0,00	0,00	0,00	0,00	0,24
3	sku_d_c_v_port	0,78	0,52	0,19	0,61	0,72	0,82	8,33
4	sku_d_c_v_port	0,29	0,16	0,00	0,16	0,29	0,40	0,75

Fonte: Elaboração própria

Capítulo 5

CONCLUSÃO E PRÓXIMOS PASSOS

Este trabalho abordou a aplicação de análise de agrupamentos para segmentar estabelecimentos comerciais que utilizam as plataformas digitais Corp App e Corp Delivery, da Corporação S.A. O tema central foi a criação de uma segmentação que permita à empresa personalizar o atendimento e incentivar ações estratégicas por meio de intervenções direcionadas. A análise identificou padrões de comportamento nos estabelecimentos, com uma compreensão detalhada das dinâmicas de compra e venda dos pontos de venda.

A relevância do tema está no crescente uso da análise de dados no mercado de varejo, especialmente em empresas com uma ampla base de clientes e complexidade operacional, como a Corporação S.A. A segmentação eficiente dos estabelecimentos é essencial para aumentar a competitividade e a eficiência das ações. Para o pesquisador, este estudo representa um aprofundamento em análise de dados e técnicas de *machine learning* aplicadas a desafios do setor de varejo.

A análise segmentou os estabelecimentos em grupos significativos, fornece informações que possibilitam ações estratégicas direcionadas. Com os perfis identificados, a equipe da Corporação S.A. responsável pelo Corp App pode personalizar os algoritmos para ações mais assertivas para aumentar as compras via Corp App e a fidelidade ao portfólio ideal. Ainda, a segmentação pode ser utilizada na estratificação de amostras em experimentos aleatorizados para identificar quais estratégias geram maior adesão e impacto em diferentes grupos de clientes.

Os resultados indicam que diferentes perfis de estabelecimentos requerem estratégias distintas, permitindo intervenções específicas no Corp App. A segmentação proposta fornece um modelo para potencializar a competitividade da Corporação S.A. no mercado

brasileiro.

Para estudos futuros, sugere-se o uso de outros métodos de agrupamento, como o *k-prototypes*, um algoritmo baseado no k-médias que utiliza tanto dados numéricos como categóricos, para entender o impacto das informações relacionadas a localização (unidade federativa) e ao tipo de estabelecimento na definição dos grupos. Como próximos passos, a empresa deve analisar as ações estratégicas sugeridas para cada grupo e avaliar e prioridade de implementação, considerando o custo, prazo e o retorno de cada ação, como forma de otimizar o uso de recursos e o impacto das intervenções planejadas.

REFERÊNCIAS

1 CHAPMAN, P. Crisp-dm 1.0: Step-by-step data mining guide. In: CRISP-DM 1.0: Step-by-step data mining guide. [s.n.], 2000. Disponível em: <https://api.semanticscholar.org/CorpusID:59777418>.

2 MORETTIN, P. A.; SINGER, J. M. Estatística e Ciência de Dados. [S.l.]: Editora Ltc, 2021. ISBN 9788521638162.

3 FACELI, K. et al. Inteligência artificial: uma abordagem de aprendizado de máquina. [S.l.]: Editora Ltc, 2011. ISBN 9788521618805.

4 BUUREN, S. V.; GROOTHUIS-OUDSHOORN, K. mice: Multivariate imputation by chained equations in r. Journal of statistical software, v. 45, p. 1–67, 2011.

5 ARTES, R.; BARROSO, L. P. Métodos Multivariados de Análise Estatística. [S.l.]: Edgard Blücher Ltda, 2023. ISBN 9786555067026.

6 JAMES, G. et al. An Introduction to Statistical Learning. [S.l.]: Springer New York Heidelberg Dordrecht London, 2013. ISBN 9781461471387.

7 ARBELAITZ, O. et al. An extensive comparative study of cluster validity indices. Pattern Recognition, v. 46, n. 1, p. 243–256, 2013. ISSN 0031-3203. Disponível em: <https://www.sciencedirect.com/science/article/pii/S003132031200338X>.

8 ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, v. 20, p. 53–65, 1987. ISSN 0377-0427. Disponível em: <https://www.sciencedirect.com/science/article/pii/0377042787901257>.

9 DAVIES, D.; BOULDIN, D. A cluster separation measure. Pattern Analysis and Machine Intelligence, IEEE Transactions on, PAMI-1, p. 224 – 227, 05 1979.

10 CALINSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. Communications in Statistics - Theory and Methods, v. 3, p. 1–27, 01 1974.