Universidade Federal de Uberlândia
Faculdade de Matemática

Bacharelado em Estatística

APLICAÇÃO DE ANÁLISE DE CONGLOMERADOS PARA A SEGMENTAÇÃO DE CLIENTES

Gabriela Gonçalves Peres

Uberlândia-MG
2017

Gabriela Gonçalves Peres

APLICAÇÃO DE ANÁLISE DE CONGLOMERADOS PARA A SEGMENTAÇÃO DE CLIENTES

Trabalho de conclusão de curso apresentado à Coordenação do Curso de Bacharelado em Estatística como requisito parcial para obtenção do grau de Bacharel em Estatística.

Orientador: Profº. Dr. Lúcio Borges de Araújo

Uberlândia-MG

2017

Universidade Federal de Uberlândia
Faculdade de Matemática

Coordenação do Curso de Bacharelado em Estatística

A banca examinadora, conforme abaixo assinado, certifica a adequação deste trabalho de conclusão de curso para obtenção do grau de Bacharel em Estatística.

Uberlândia, ____ de _____ de 20____

BANCA EXAMINADORA

_____

Prof^{o}. Dr. Lúcio Borges de Araújo

_____

Prof^{a}. Dra. Aurélia Aparecida de Araújo Rodrigues

_____

Prof^{a}. Ma. Patrícia Viana da Silva

Uberlândia-MG

2017

Agradecimentos

Agradeço à Deus por ter me abençoado com saúde e ter me dado condições de realizar cada etapa da graduação. Foram anos de aprendizado, amadurecimento e percepções. Agradeço também, minha família pela aceitação da minha escolha e por todo apoio nessa trajetória.

Um agradecimento com carinho à todos os amigos que fiz nesse curso. Tantas pessoas incríveis que conheci. Algumas eu pude me aproximar mais, outras conheci de forma mais superficial, sou grata a cada momento. Um agradecimento especial ao João Flávio que esteve comigo como grande amigo e parceiro desde o começo, tenho completa admiração pela pessoa, pelo profissional, pelo amigo e pelo companheiro que você é. Michelle Silva, grande amiga que terei para a vida, obrigada por dividir tantos momentos comigo, torço com muito amor pelo seu sucesso e felicidade. Também agradeço pela amizade e parceria do Mateus Aguiar, do Victor Moreira e da Gabriela Bolaina, que além de nossos momentos na vida acadêmica, juntos, demos o primeiro passo na vida profissional, o que me acrescentou tanto conhecimento quanto admiração por vocês.

Agradeço à todos os professores por compartilharem suas experiências e conhecimentos, pela paciência e dedicação. Em especial, agradeço à Patrícia Viana que trouxe melhorias para o curso, que me mostrou o quanto eu precisava estar mais atenta, agradeço por usar o seu espírito maternal para nos alertar e nos trazer para a realidade, buscando que sejamos pessoas melhores e menos acomodadas. Um agradecimento especial à Aurélia Araújo, que foi minha orientadora de Iniciação Científica, por tanta paciência e compreensão nas nossas pesquisas e nas etapas que passamos juntas. Tenho também um forte sentimento de gratidão e admiração à professora Priscila Neves, por ser tão apaixonada por sua profissão e pela forma atenciosa que conduz suas aulas. E claro, um forte agradecimento ao Lúcio Borges, por ser tão solícito, sereno e compreensivo como orientador e como professor. Obrigada por conseguir passar seu conhecimento de forma clara e palpável, e por fazer parte do fim desse ciclo.

Que Deus abençoe todos nós e que sempre haja motivação para seguirmos em frente com nossos objetivos e sonhos.

Resumo

Segmentação de mercado é dividir um mercado em grupos de compradores potenciais que tenham semelhantes necessidades e desejos, percepções de valores ou comportamentos de compra. A segmentação de mercado representa um esforço para o aumento da precisão de alvo de uma empresa, fazendo sentido ao ajudar a corporação a penetrar mais a fundo nos mercados que escolheu como prioritários e ao contribuir para aumentar a diferenciação entre produtos e serviços. O presente trabalho teve como objetivo reestruturar a classificação dos clientes de uma empresa atacadista situada em Uberlândia, Minas Gerais, por meio da análise de conglomerados, que tem como finalidade realizar uma classificação de acordo com as relações naturais contidas na amostra, formando grupos de objetos por similaridade. Os objetos são agrupados de acordo com a semelhança em relação a algum critério pré-determinado. Foi de interesse da empresa que os clientes fossem agrupados de acordo com suas respectivas necessidades e reciprocidade com a corporação, com valor pré-determinado de três grupos de clientes: simples, intermediário e de valor agregado. O que possibilita especificar e direcionar a forma de atendimento à cada tipo de comprador, com mais precisão nos produtos e serviços que serão oferecidos, garantindo assim uma maior satisfação do cliente, e ainda, possibilitando a otimização de gastos no atendimento e uma venda mais assertiva.

Palavras-chave: *Cluster*, agrupamento, atacadista.

Abstract

Market segmentation is to divide a market into groups of potential buyers who have similar needs and desires, perceptions of values or buying behaviors. Market segmentation represents an effort to increase a company's target accuracy, making sense by helping the corporation penetrate deeper into the markets it has chosen as priorities and by helping the increment of the differentiate between products and services. The objective of this work was to restructure the classification of the customers of a wholesale company located in Uberlândia, Minas Gerais, through Cluster analysis, whose purpose is to perform a classification according to the natural relations contained in the sample, forming groups of objects by similarity. The objects are grouped according to the similarity to some predetermined criterion. It was in the companys interest that customers were grouped according to their respective needs and reciprocity with the corporation, with pre-determined value of three groups of customers: simple, intermediate and value-added. This makes it possible to specify and direct the manner of service to each type of buyer, with more precision in the products and services that will be offered, thus guaranteeing a greater customer satisfaction, and also, allowing the optimization of expenses in the service and a more assertive sale.

Keywords: Clustering, grouping, wholesaler.

Sumário

Lista de Tabelas

1. Introdução

As mudanças no cenário econômico mundial em direção à globalização e à competitividade, em sua grande maioria, exigem maior agilidade das empresa [13]. A sobrevivência das organizações requer adaptação, mesmo que, geralmente, as estruturas organizacionais e práticas gerenciais não sejam criadas no ritmo de mudanças e, sim, para funcionarem com estabilidade [10]. A disputa de mercado potencializa cada vez mais a conquista de novos clientes e a fidelização dos vigentes. É preciso conhecer o cliente, saber suas preferências, compreender suas necessidades e atendê-los com astúcia, atributos que são consolidados pelo relacionamento constante nas pré-vendas, vendas e pós-vendas.

Segmentação de mercado é o processo de reconhecer os diversos grupos de clientes, com características relativamente homogêneas, definindo assim a melhor forma como cada um será abordado. Tratando-se das vendas por atacado, segundo Alves [2], cada cliente é um mundo à parte, que merece e exige atenção e dedicação exclusivas. No entanto, é preciso estar alerta, dado que nem todos os segmentos são rentáveis o suficiente ou mesmo desejáveis, devido aos riscos, por exemplo. O que requer a necessidade de escolher os mercados-alvo após a segmentação.

Para que a segmentação aconteça é necessário identificar os fatores que afetam as decisões de compras dos consumidores, atendendo cinco requisitos: o segmento deve ser identificável, mensurável, acessível, rentável e estável. Além disso, as empresas podem adotar cinco níveis de segmentação: *marketing* de massa, *marketing* de segmento, *marketing* de nicho, *marketing* local e *marketing* individual, os quais serão descritos a seguir:

O *marketing* de massa é quando ocorre uma oferta de produtos com padrão satisfatório para a maioria dos consumidores, utilizando-se do apoio de revendedores para alcançar o mercado juntamente com grandes campanhas de publicidade e de promoção para liderar e guiar o mercado maciço, tendo, ainda, os preços fixados em um nível acessível [9].

No caso do *marketing* de segmento, a empresa reconhece que os compradores diferem em seus desejos, poder de compra, localizações geográficas, atitudes e hábitos de compra. A segmentação é um ponto intermediário entre *marketing* de massa e *marketing* individual. Presume-se que os consumidores pertencentes a um segmento sejam bastante similares em desejos e necessidades [9]. O qual é adotado pela empresa do presente estudo.

No *marketing* de nicho há um grupo mais restrito de compradores, tipicamente um pequeno mercado. Comumente as empresas identificam nichos dividindo segmentos em subsegmentos ou definindo um grupo formado por um conjunto de traços que podem buscar uma combinação especial de benefícios [9].

O *marketing* local é característico com programas preparados sob medida conforme as necessidades e desejos de grupos de consumidores locais (áreas comerciais, vizinhanças, até lojas individuais) [9].

O último nível de segmentação, o *marketing* individual, o fabricante prepara a oferta, a logística e as condições financeiras sob medida a cada cliente [9].

Frequentemente muitas empresas entram no mercado, novos produtos e diversos serviços são lançados. A cada dia a concorrência se fortifica uma vez que o atacado a cada ano vem aumentando seu faturamento e atraindo cada vez mais investimento [15]. Conhecer o público alvo é a melhor maneira de conduzir as ações de forma adequada e, com isso, obter melhores resultados para o negócio. É fundamental saber como atrair o cliente com estratégias de *marketing* bem direcionadas e com vendedores preparados para uma abordagem efetiva.

A empresa fornecedora de dados (EFD) desse estudo é um comércio atcadista. Visando oferecer um atendimento diferenciado à sua carteira de clientes, a EFD especializa sua equipe de representantes comerciais para trabalharem de forma direcionada de acordo com os segmentos de clientes, segundo informações de um gestor interno da mesma, o qual preferiu não ser identificado.

1.1 Objetivos Gerais e Específico

A classificação dos clientes da EFD está inclusa nos fatores para serem melhorados em sua gestão. A carteira de clientes da EFD além de numericamente vasta, é uma carteira de clientes heterogêneos. É necessário ter funcionários com qualificação para atender toda a demanda da melhor forma de se relacionar e gerando menos custos o possível. Como missão, a EFD visa suprir a carência do mercado de vender por unidade para pequenos e médios varejistas que não têm necessidade de fazer compras volumosas para suas lojas. Isso implica que o público alvo da EFD varia do pequeno ao grande varejista, de qualquer região do Brasil. Assim sendo, foi constatada a necessidade de atualizar a classificação de clientes já existente, para fortificar a importância do atendimento personalizado, que só é possível quando se conhece as classes atendidas. A EFD tem o objetivo de reduzir de seis para três agrupamentos de clientes.

Com a nova classificação de clientes, a EFD almeja poder oferecer um profissional para atender o cliente com o perfil que ele necessita. Sendo possível passar instruções e orientações sólidas e alcançáveis para seus vendedores. Gerando uma satisfação melhor tanto de seus clientes quanto de seus funcionários.

Esse estudo tem o objetivo geral de atender as expectivas da EFD, fazendo a classificação dos clientes em três grupos. Especificamente, esse estudo tem o objetivo de realizar a classificação dos clientes através da análise de conglomerados, por meio da técnica não-hierárquica de K-médias.

2. Materiais e Métodos

Foi solicitado pela EFD, antes de iniciar a análise para refazer a classificação dos clientes, que fossem determinados três grupos, com a intenção de que cada um tenha grau diferente de necessidade de atendimento e de produtos. Uma vez que foi estipulado o número de *clusters*, esse estudo foi então, embasado em uma análise de conglomerados de método não-hierárquico.

2.1 Apresentação do Problema

Um cliente é a pessoa mais importante do mundo, quer ele se comunique pessoalmente, quer ele se comunique por canais alternativos. O cliente é quem leva seus desejos à empresa, a qual deve, lidar com ele de maneira lucrativa para ambas as partes. Conhecendo o cliente é a melhor forma de satisfazê-lo e, até mesmo, antecipar-lhe em suas necessidades [9].

Na busca de uma melhor posição competitiva no mercado, as empresas consideradas visionárias não utilizam mais, com alta prioridade, seu tempo ocupando-se apenas em conquistar eficiência operacional. Dentre todas as estratégias, arregimenta-se para tornar-se mais competitivo no mercado, a operação de atendimento a clientes [4].

Além da necessidade de estar em constante evolução e em processos de melhorias, a EFD se encontrou numa situação problemática na antiga segmentação de seus clientes: havia seis grupos distintos de clientes e com essa quantidade, o atendimento de cada classe não estava sendo diferenciado. Em outras palavras, esse agrupamento não estava gerando benefícios nem para a empresa, nem para o cliente. Então, para adequedar promoções, ofertas e estratégias de vendas, surgiu o anseio de reestruturar os grupos de clientes.

2.2 Relacionamento Entre Cliente e Empresa

Os antigos grupos de clientes foram segmentados de acordo com seu relacionamento com a empresa. Se eles tinham uma necessidade maior de volume de produtos e serviços, então, eles eram denominados mais "recíprocos" , quanto mais variedade de marcas e itens diferentes fossem comprados, maior seria o grau de "necessidades" . A figura 2.1 mostra quais as classificações existiam e, também, a relação de reciprocidade e necessidade, que era utilizada para o *clustering*.

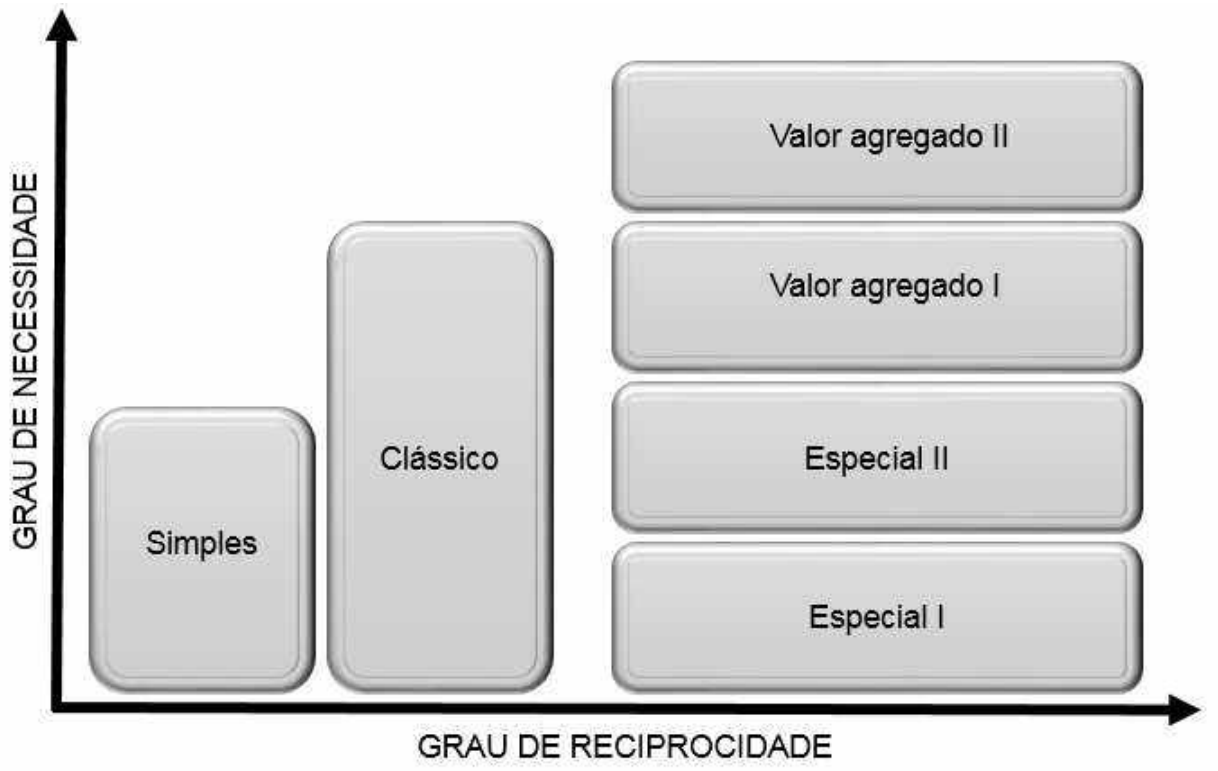

Figura 2.1: Agrupamento de Clientes de acordo com os graus de Reciprocidade e de Necessidade

Os próximos tópicos descrevem os *Clusters* antigos de clientes (CAC) para melhor entender a representação gráfica vista no tópico anterior:

1. CAC Simples: Grupo de clientes que compravam um baixo volume de mercadorias, não tinham necessidade de comprar serviços. Esse grupo também não comprava uma grande variedade de marcas e nem uma grande diversidade de itens diferentes. Em outras palavras, apresentavam baixo grau de reciprocidade e de necessidades.

2. CAC Clássico: Grupo de clientes que compram um volume maior de mercadorias, em comparação ao CAC Simples e que já compram algum tipo de serviço oferecido pela EFD. Esse grupo já compra uma variedade moderada de marcas e de itens diversos. Sendo assim, esses clientes apresentavam o grau intermediário de reciprocidade e de necessidades.

3. CAC Especial I: Grupo de clientes que compravam um alto volume de mercadorias e de serviços, no entanto, sem muita variedade de marcas ou de itens diversos. Ou seja, alto grau de reciprocidade e baixo grau de necessidades.

4. CAC Especial II: Grupo de clientes que compravam um alto volume de mercadorias e de serviços, com uma variedade moderada de marcas e de itens diversos. Ou seja, alto grau de reciprocidade e grau intermediário-baixo de necessidades.

5. CAC Valor Agregado I: Grupo de clientes que compravam um alto volume de mercadorias e de serviços, com uma variedade moderada, um pouco mais elevada do que o CAC

Especial II, de marcas e de itens diversos. Ou seja, alto grau de reciprocidade e grau intermediário-alto de necessidades.

6. CAC Valor Agregado II: Grupo de clientes que compravam um alto volume de mercadorias e de serviços, com uma variedade alta de marcas e de itens diversos. Ou seja, alto grau de reciprocidade e de necessidades.

Essa quantia de grupos, na visão da EFD, estava deixando a desejar a atenção direcionada ao cliente. Seria melhor e mais viável reduzir o número de *clusters* para poder aprimorar as abordagens com os clientes e, também, para adequar os profissionais que irão atendê-los. A nova classificação dos grupos foi realizada por meio da análise de conglomerados.

2.3 Base de Dados

Foi utilizado um banco de dados de 10.374 clientes varejistas de Minas Gerais, que representam 20% do total de clientes atendidos nesse ramo de negócio da empresa atacadista em estudo, entre julho 2016 e agosto de 2017. Dez variáveis foram utilizadas para a segmentação dos clientes, sendo elas:

Tempo de casa: mensuração de quantos meses o cliente compra da empresa em questão;

Checkouts: quantidade de caixas de pagamento na loja do cliente;

Limite de crédito: qual o valor de crédito o cliente tem para realizar compras na empresa;

Média de compra: média mensal de compra do cliente no período determinado;

Frequência de compras: tempo médio em dias para o cliente realizar uma compra;

Média de itens: média mensal da quantidade de itens diversos;

Quantidade de pedidos: quantidade total de pedidos – por cliente – realizados em todo o período do estudo;

Valor da venda: quantidade total da venda – por cliente – realizada durante o ano fechado em análise;

Valor primeira compra: qual foi o valor da primeira compra do cliente;

Valor última compra: qual foi o valor da última compra do cliente.

Em oportunidades futuras, a EFD irá ampliar essa reestruturação para todos os segmentos e todos os estados.

2.4 FUNDAMENTAÇÃO TEÓRICA

Inicialmente, foi realizada uma análise descritiva básica composta por média, mediana, desvio padrão, mínimo e máximo de cada variável. Em seguida, a análise de agrupamento pelo método não hierárquico de k-médias foi feita com o número de três grupos pré-estabelecido pela empresa. As etapas seguidas para a análise de conglomerados foram:

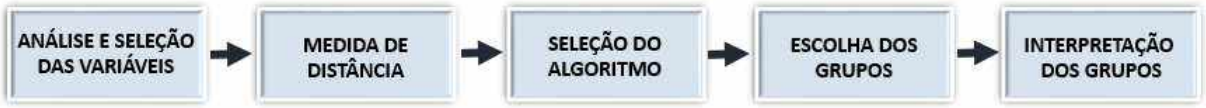

Figura 2.2: Etapas da análise de conglomerados

2.4.1 ANÁLISE DE CONGLOMERADOS

A análise de conglomerados ou de *cluster*, busca solucionar o seguinte problema: dada uma amostra de n indivíduos, cada um dos quais caracterizados por p variáveis; um critério deve ser criado para agrupar os indivíduos em classes ou *clusters*, de forma que os indivíduos que possuam características semelhantes estejam na mesma classe. Sendo assim, a análise de conglomerados objetiva alocar indivíduos em grupos mutuamente exclusivos, isto é, um indivíduo pertence apenas à um grupo, tal que os elementos do mesmo grupo são mais parecidos quanto possível uns com outros, enquanto indivíduos em grupos diferentes são heterogêneos [20].

A análise de conglomerados, é uma técnica de avaliação ou análise de interdependência que busca agrupar os elementos conforme sua estrutura natural [6]. Em termos gerais, pode-se dizer que uma análise de conglomerados terá êxito se trouxer informações para aprimorar agrupamentos previamente determinados em um conjunto de dados, ou ajudar a formalizar sua estrutura hierárquica [1]. Esta é uma importante técnica exploratória, uma vez que, ao estudar a estrutura natural de grupos, permite reduzir a dimensionalidade dos dados, identificar valores atípicos (*outliers*) e levantar hipóteses relacionadas às associações dos indivíduos [8].

Essa técnica visa segregar indivíduos (ou variáveis) em grupos homogêneos internamente, heterogêneos entre si e mutuamente exclusivos (não existe mais de um grupo com as mesmas particularidades), a partir de determinadas características conforme uma medida de similaridade, que mede quão semelhantes são dois indivíduos, ou de dissimilaridade, que mede quão diferentes são dois indivíduos. [6].

Cabe destacar que ao realizar uma análise de conglomerados, não há distinção se as variáveis são ou não relevantes para o estudo, ficando à serviço do pesquisador. Nesse sentido, a inclusão de variáveis não representativas ou a presença de multicolinearidade podem distorcer o resultado da pesquisa. A multicolinearidade interfere na ponderação das medidas de similaridade [6]. Uma maneira de mitigar seus efeitos, é aderir a medida de distância de Mahalanobis (D^2), a qual além de padronizar os dados estabelecendo uma escala em termos de desvio padrão, soma a covariância acumulada dentro dos grupos, ajustando as intercorrelações entre as variáveis,

sendo, então, uma medida de distância comparável ao (R^2) da regressão.

Outro ponto importante a ser ressaltado, é que essa técnica é sensível à inclusão de variáveis com comportamentos atípicos, ou seja, com a presença de *outliers*. Os *outliers* podem ser definidos como observações que saem do padrão esperado em cada variável, isto é, referem-se a observações com características muito destoantes dos demais membros da população, podendo ser prejudicial a qualidade dos resultados. Sendo assim, recomenda-se a verificação da existência de *outliers*, e a decisão do pesquisador se os mesmos devem ou não permanecer na base de dados [5].

Após a seleção e tratamento do banco de dados, a próxima etapa que o pesquisador irá se deparar, relaciona-se à escolha da medida de similaridade a ser utilizada no estudo. O conceito de similaridade na análise de conglomerados, é de vital importância, uma vez que a identificação de agrupamentos de variáveis só é possível ao adotar alguma medida de semelhança que permita a comparação entre as variáveis. Nessa técnica, as observações são agrupadas de acordo com algum tipo de métrica de distância (e as variáveis conforme medidas de correlação ou associação). A análise teórica das relações de semelhança tem sido dominada pelos modelos geométricos [14]. Eles representam os objetos como pontos em um espaço de coordenadas, de forma que as dissimilaridades observadas entre objetos correspondam às distâncias métricas entre os respectivos pontos. Para separar os grupos similares por uma determinada característica, com base em duas variáveis, pode-se plotar um gráfico de dispersão e verificar quais observações apresentam comportamentos semelhantes. Isto é, alocar as observações em grupos conforme o grau de similaridade, geometricamente falando. Para uma grande quantidade de variáveis, não é possível identificar visualmente os grupos, devido à limitação gráfica de espaço tridimensional, no entanto, outros critérios de aglomeração, tais como medidas de distância e outras medidas de similaridade, podem ser utilizados. De maneira ampla, as medidas de similaridade ou distância podem ser classificadas em três tipos:

1. Medidas de distância: medida de separação entre dois pontos, utilizada para variáveis métricas.

2. Medidas correlacionais: permitem trabalhar com variáveis categóricas.

3. Medidas de associação:quando os indivíduos são agrupados com base nos coeficientes de correlação ou de outras medidas de associação.

2.4.2 Função de Distância

A análise de conglomerados baseia-se normalmente em uma função de similaridade, função esta que recebe dois objetos e retorna a distância entre eles [11]. Para serem determinados por uma métrica de qualidade, os grupos devem apresentar alta homogeneidade interna e alta heterogeneidade externa. Os indivíduos podem também ser chamados de objetos, exemplos, t-uplas ou registros. Cada indivíduo representa uma entrada de dados que pode ser constituída por um vetor de atributos que são campos numéricos ou categóricos.

Os métodos de agrupamento assumem que todos os relacionamentos relevantes entre os objetos podem ser descritos por uma matriz contendo uma medida de dissimilaridade ou de proximidade entre cada par de objetos.

Cada entrada d_{ij} na matriz 2.2 incide em um valor numérico que evidencia quão próximos os objetos i e j são. Algumas métricas calculam a similaridade, outras calculam a dissimilaridade. Os coeficientes de distância são funções $d : \Gamma \times \Gamma \longrightarrow \mathbb{R}$, em que Γ representa o conjunto de objetos sendo trabalhado pelo pesquisador. Seja n o número de indivíduos e p o número de variáveis, é possível realizar a transformação da matriz de dados, dada por:

$$\Gamma = \begin{pmatrix} x_{1,1} & \cdots & x_{1,f} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i,1} & \cdots & x_{i,f} & \cdots & x_{i,p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,f} & \cdots & x_{n,p} \end{pmatrix} \tag{2.1}$$

em que uma matriz de distâncias, dada por:

$$d = \begin{pmatrix} 0 & & & & \\ d_{2,1} & 0 & & & \\ d_{3,1} & d_{3,2} & 0 & & \\ \vdots & \vdots & \vdots & \ddots & \\ d_{n,1} & d_{n,2} & \cdots & \cdots & 0 \end{pmatrix} \tag{2.2}$$

onde a entrada $d_{i,j}$ representa exatamente a distância entre os elementos i e j.

As funções de (dis)similaridade obedecem as seguintes propriedades:

i. $d_{ii} = 0$. A diagonal principal é nula, uma vez que se trata da distância entre o mesmo indivíduo.

ii. $d_{ij} > 0, \ \forall \ x_i, \ x_j \in \Gamma$.

iii. $d_{ij} = d_{ji}, \ \forall \in \Gamma$. Esta regra afirma que a distância entre dois elementos não varia, não importando o ponto a partir do qual ela é medida. Por isto, a matriz 2.2 é mostrada sendo triangular inferior. Uma vez que ela é simétrica, os valores acima da diagonal principal estão implicitamente definidos.

iv. $d_{ij} + d_{jk} \geq d_{ik}, \ \forall \in \Gamma$. Este critério é conhecido como a desigualdade triangular e refere-se ao teorema que afirma que, num triângulo, o comprimento de um dos lados é sempre inferior à soma dos comprimentos dos outros dois lados.

Uma das principais medidas de dissimilaridade usadas nos algoritmos de agrupamento é a distância Euclidiana entre dois pontos, dada por:

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2} \tag{2.3}$$

2.4.3 Distância de Mahalanobis

A distância de Mahalanobis é uma métrica que se difere da distância Euclidiana por levar em consideração a correlação entre os conjuntos de dados, além disso, não é necessário que as variáveis tenham a mesma unidade de medida. Sua fórmula para a distância entre dois vetores $(\boldsymbol{x_i}, \boldsymbol{x_j})$ que possuem uma matriz de covariância $\boldsymbol{\Sigma}$ — matriz simétrica que sumariza a covariância entre $\boldsymbol{p}$ variáveis — é dada por:

$$\boldsymbol{d(x_i, x_j) = \sqrt{(x_i - x_j)^t \Sigma^{-1} (x_i - x_j)}} \tag{2.4}$$

O uso da distância de Mahalanobis corrige algumas das limitações da distância Euclidiana, pois leva em consideração automaticamente a escala dos eixos coordenados. Como ponto negativo, as matrizes de covariância podem ser difíceis de determinar e a memória e o tempo de computação crescem de forma quadrática com o número de características.

A métrica de Mahalanobis também é usada para medir a distância entre um elemento $\boldsymbol{x}$ e um grupo de elementos cuja média seja dada por $\boldsymbol{\mu = (\mu_1, \mu_2, ..., \mu_n)}$ e que possua uma matriz de covariância dada por $\boldsymbol{\Sigma}$ (matriz quadrada que contém as variâncias e covariâncias associadas às variáveis em estudo). Neste caso, a distância é dada pela fórmula:

$$\boldsymbol{d = \sqrt{(x_i - \mu)^t \Sigma^{-1} (x_i - \mu)}} \tag{2.5}$$

Conceitualmente, é como se estivesse sendo avaliada a pertinência de um elemento não só por sua distância ao centro (média) do grupo, mas também, pela variabilidade do mesmo, determinando assim a distância do indivíduo $\boldsymbol{x_i}$ em termos de uma comparação com o desvio padrão do agrupamento. Quanto maior for o valor desta métrica, maior o número de desvios padrões que um indivíduo está distante do centro do agrupamento, e menor sua chance pertencer no mesmo.

2.4.4 Métodos Hierárquicos x Não-Hierárquicos

Os métodos hierárquicos e não-hierárquicos se referem principalmente a maneira como os dados são divididos ou organizados. Os métodos hierárquicos constroem uma hierarquia de partições, enquanto os não-hierárquicos constroem uma partição dos dados [3].

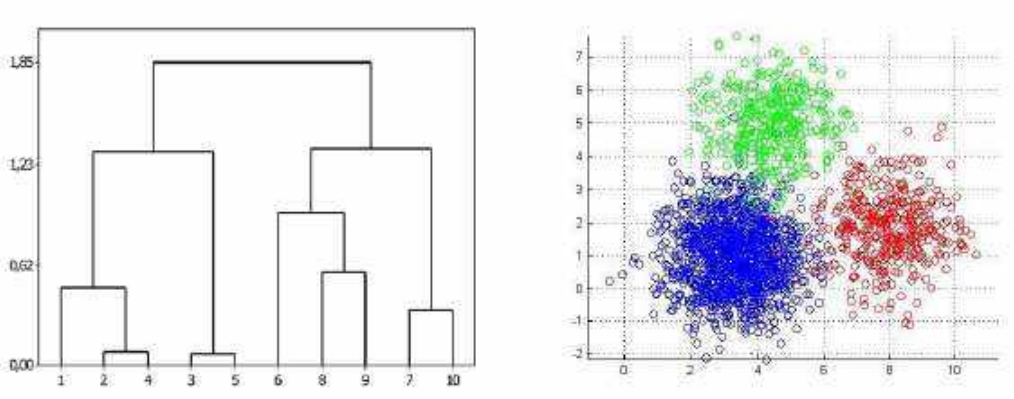

Figura 2.3: Método Hierárquico x Método Não-Hierárquico [3]

Os algoritmos hierárquicos criam uma hierarquia de relacionamentos entre os indivíduos usando uma medida de distância. Os algoritmos não-hierárquicos separam os indivíduos em grupos baseando-se nas características que eles possuem [3]. Dado que a EFD solicitou três grupos de clientes, será utilizado um método não-hierárquico.

2.4.5 MÉTODOS NÃO-HIERÁRQUICOS

Os métodos não hierárquicos têm como objetivo encontrar diretamente uma partição de $\boldsymbol{n}$ indivíduos em $\boldsymbol{k}$ grupos, de modo que a partição atenda dois requisitos básicos: semelhança interna, também chamada de "coesão", e isolamento dos clusters formados, também denominado como "separação" dos grupos [12]. Para encontrar a melhor partição de ordem $\boldsymbol{k}$, deve ser empregado algum critério de qualidade de partição. Computacionalmente, é impossível criar todas as partições possíveis de ordem $\boldsymbol{k}$ e, após conhecer essas partições, eleger a mais adequada. Sendo assim, são necessários processos que investiguem uma parte das partições possíveis com o objetivo de achar a partição quase ótima.

Há vários aspectos em que os métodos não hierárquicos se diferenciam dos hierárquicos. Em primeiro lugar, os métodos não-hierárquicos requerem que o pesquisador especifique previamente o número $\boldsymbol{k}$ de grupos desejado, o oposto das técnicas hierárquicas. A cada estágio do agrupamento, os novos grupos podem ser formados pela divisão ou junção de grupos já combinados em etapas anteriores, quando é utilizado um método não-hierárquico. É comum que os algoritmos computacionais utilizados nos métodos não-hierárquicos sejam do tipo iterativo e, comparando com os métodos hierárquicos, existe uma maior capacidade de análise de conjunto de dados com um grande número de observações.

Os métodos das K-Médias utilizado para partições sem sobreposição dos dados, o Fuzzy c-Médias que é um método de agrupamento com sobreposição (nesse caso os indivíduos pertencem a todos os grupos com diferentes graus de pertinência) e as redes neurais artificiais aplicadas à análise de conglomerados, que é um método que tem a capacidade de lidar com dados imprecisos, incompletos ou totalmente novos, são exemplos dos métodos não-hierárquicos [17]. Nesse estudo, os indivíduos devem pertencer à apenas um grupo e os dados são precisos,

então será utilizado o método das K-Médias.

2.4.6 K-Médias

O método das K-Médias é, de acordo com Hartigan e Wong [7], um dos mais conhecidos e utilizados para problemas práticos. E pode ser resumido em quatro etapas conforme é apresentado na figura 2.4:

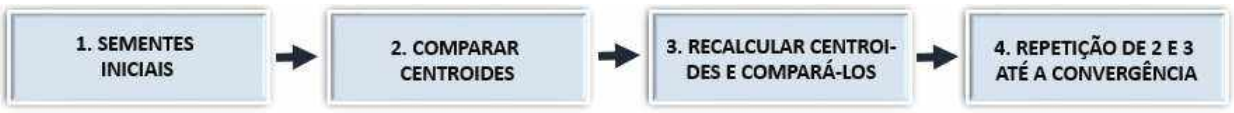

Figura 2.4: Algoritmo resumido do método das K-Médias

De forma simples, cada elemento amostral é alocado àquele cluster cujo centroide (vetor de médias amostral) é o mais próximo do vetor de valores observados para o respectivo elemento. O processo é composto de quatro passos:

i. O pesquisador escolher $\boldsymbol{k}$ centroides, também chamados de "sementes" ou "protótipos", para ser iniciado o processo de partição.

ii. Então, cada elemento do conjunto de dados será comparado com cada centroide, por meio de uma medida de distância. O elemento é alocado ao grupo cuja distância é a menor.

iii. Após o passo (ii) ter sido aplicado para todos os $\boldsymbol{n}$ elementos amostrais, recalcula-se os valores dos centroides para cada grupo formado. Em seguida, repete-se o passo (ii), levando em consideração os centroides novos dos grupos.

iv. Deve-se repetir os passos (ii) e (iii) até que todos os indivíduos amostrais estejam bem alocados em seus respectivos grupos. Em outras palavras, até que não seja necessária nenhuma outra realocação de indivíduos, dado que cada indivíduo estará no grupo em que a distância dele para o centroide do grupo, será a menor.

A escolha das sementes iniciais de agrupamento, influencia no agrupamento final, então, o pesquisador precisa ficar atento na escolha das sementes. Grande parte dos *softwares* estatísticos tem como padrão escolher as sementes iniciais pelo uso das $\boldsymbol{k}$ primeiras observações do banco de dados. No entanto, o pesquisador pode deixar especificado quais sementes deseja utilizar para iniciar o algoritmo. Tal procedimento pode trazer bons resultados quando os $\boldsymbol{k}$ primeiros elementos amostrais são discrepantes entre si e não é recomendável quando são semelhantes entre si.

2.4.7 Critério de convergência

O valor padrão do critério de convergência utilizado para saber quantas iterações, quantas vezes os centroides serão recalculados e os indivíduos realocados, é 0,0001. As iterações param

quando a variação relativa máxima nas sementes do grupo é menor ou igual ao critério de convergência. A mudança relativa em uma semente do grupo é a distância entre a semente velha e a nova semente dividida por um fator de escala, que nesse caso é a distância mínima entre as sementes iniciais [16].

2.4.8 TESTE DO PSEUDO-F

O pseudo-F está baseado na variância entre os grupos, a cada nível de agregação. Assim, um valor alto do teste é desejável e implicaria na rejeição à hipótese de homogeneidade entre os grupos criados [19]. A estatística do pseudo-F é dada por:

$$F = \frac{\frac{R^2}{c-1}}{\frac{1-R^2}{n-c}}$$

onde R^2 é o coeficiente de determinação global observado, c é o número de grupos e n é o número de observações.

2.5 SOFTWARE PARA ANÁLISE

Todo o processo de manipulação dos dados e todas as análises, foram realizadas no *software SAS Enterprise Guide*. O código da análise de conglomerdos está anexado no apêndice 1.

3. RESULTADOS

Na tabela 3.1 são apresentadas as estatísticas descritivas das variáveis. De acordo com os dados, os clientes possuem em média 14 anos de relacionamento com a EFD, ao menos 50% desses, tem apenas um *checkout*. Os clientes podem ter ou não ter limite de crédito, sendo que o máximo foi de R$1.000.000,00. No período pesquisado de um ano, os clientes compraram em média 38 vezes, sendo R$1.448,63 o valor médio de compra. Observa-se que a média do valor da última compra realizada, está maior do que o da primeira.

Tabela 3.1: Análise descritiva das variáveis para a nova reestruturação de clientes

Variável	$\bar{x}$	**Mediana**	s	**Mínimo**	**Máximo**
Tempo de casa	169,35	146	113,71	0	531
Checkouts	2	1	3,31	1	90
Limite de crédito	23.415,55	10.920	32.109	0	1.000.000
Média de compra	1.448,63	919,7	2372	32,11	71.200,99
Frequência de compras	38,31	29	40,38	1	407
Média de itens	14	12	11,4	1	170
Quantidade de pedidos	15	10	18,3	1	213
Valor da venda	34.525,84	8.844,27	132.508	32,11	5.642.711,42
Valor da primeira compra	1.324,38	562,67	5.421	47,23	259.898,66
Valor da última compra	1.420,04	771,22	2.802	126,30	9.6013,20

Na tabela 3.2 apresentam-se as estimativas iniciais de cada variável em cada grupo, gerados aleatoriamente pelo *SAS*. Na tabela 3.3 observa-se as iterações necessárias para satisfazer o critério de convergência.

Tabela 3.2: Estimativas iniciais da análise de conglomerados por K-Médias

Variáveis	Grupos		
	1	**2**	**3**
Tempo de casa	171	26	111
Checkout	5	20	4
Limite de crédito	1.000.000	500.000	0
Media de compras	26.870	20.810	32
Frequencia de compras	2	3	0
Média de itens	40	29	1
Quantidade de itens	210	128	1
Valor da venda	5.642.711	2.663.768	32
Valor primeira compra	467	259.898	984
Valor última compra	1.064	17.508	493

Tabela 3.3: Histórico de iterações

Iteração	Critério	Mudança Relativa nas Estimativas		
		1	**2**	**3**
1	32686	0	0,1879	0,0141
2	29165,4	0,3114	0,1622	0,000303
3	27525,8	0	0,1218	0,000486
4	26326,3	0	0,0522	0,000295
5	25981,2	0	0,0457	0,00027
6	25658,1	0,2224	0,0456	0,000244
7	24659,9	0,2123	0,0618	0,000418
8	23485,8	0	0,0332	0,000421
9	23152,8	0,0876	0,0192	0,000203
10	22764	0,1169	0,0221	0,000197
11	22220,1	0,0819	0,0186	0,000182
12	21953,4	0	0,0139	0,000243

O resumo dos grupos obtidos é apresentado na tabela 3.4. Observa-se que o primeiro grupo é o de menor frequência (10 elementos) e o que apresenta maior distância entre os outros centroides. O segundo grupo tem uma frequência um pouco maior em comparação ao primeiro, esse grupo é o mais próximo do terceiro e o terceiro grupo reciprocamente é o mais próximo do segundo. Em analogia à EFD, espera-se que o primeiro grupo seja de clientes maiores, dado que eles são minoria na empresa. Os clientes de médio e pequeno porte são os mais recorrentes, o que é esperado resultar nos grupos dois e três respectivamente.

Tabela 3.4: Resumo dos grupos

Cluster	n	Grupo mais próximo	Distância entre os centroides
1	10	2	2318483
2	179	3	584811
3	10185	2	584811

Observa-se a abaixo, os coefientes de determinação $\boldsymbol{R^2}$ na tabela 3.5. O $\boldsymbol{R^2}$ é uma medida descritiva da qualidade do ajuste obtido, basicamente, ele indica quanto o modelo foi capaz de explicar os dados coletados, variando entre 0 e 1, indicando, em porcentagem, o quanto o modelo consegue explicar os valores observados [18]. Algumas das variáveis são melhores diferenciadas pelos grupos, como o valor de venda, o limite de crédito e a quantidade de pedidos. De modo geral, 74% da variação dos dados, é explicada pelos *clusters*.

Tabela 3.5: Coeficiente de determinação por variável

Variável	$\boldsymbol{R^2}$
Tempo de casa	0,000302
Checkouts	0,095692
Limite de crédito	0,458284
Média de compra	0,247997
Frequência de compras	0,011843
Média de itens	0,076709
Quantidade de pedidos	0,302200
Valor da venda	0,762293
Valor primeira compra	0,063155
Valor última compra	0,131737
NO GERAL	0,743947

As médias das variáveis em seus respectivos grupos podem ser observadas na tabela 3.6. Pode-se observar que o grupo 1 apresenta, em média, os maiores valores de limite de crédito, média, frequência, venda total, primeira e última compra, e ainda, a maior média de quantidade de *checkouts*. Já o grupo 2 apresenta valores intermediários e uma média inferior de quantidade de *checkouts*. O grupo 3, apresenta maior frequência de compra, no entanto, os demais valores são os mais baixos em comparação aos grupos anteriores.

Tabela 3.6: Médias dos grupos

Variáveis	Grupos		
	1	**2**	**3**
Tempo de casa	206,60	181,38	169,10
Checkout	13	9	2
Limite de crédito	561.500,00	127.762,87	21.053,34
Média de compra	25.480,15	8.331,85	1.304,06
Frequência de compras	3	6	38
Média de itens	23	28	14
Quantidade de pedidos	135	85	13
Valor da venda	2.874.092	596.709	21.857
Valor primeira compra	28.457	9.375	1.156
Valor última compra	18.221	7.991	1.288

Nesse caso, tem-se que a estatística $\boldsymbol{pseudo-F = 15066,15}$. Conclui-se que há diferença significativa entre os grupos, logo, um agrupamento não é igual ao outro. A análise de conglomerados para a reestruturação dos clientes da EFD mostrou-se eficaz atendendo o objetivo da mesma: uma divisão de três grupos com características heterogêneas:

i **Grupo I**: Clientes de valor agregado - clientes que estão há um bom tempo comprando da EFD, que apresentam grande porte, frequentemente compram pouco, mas, suas compras são de grande faturamento e sua necessidade de itens diversos é maior. Esses são a minoria de clientes em termo de volume, no entanto, são os que mais dão lucro à EFD. Por terem uma relação sólida com a EFD e por apresentarem um perfil lucrativo, o limite de crédito desses clientes é mais alto dos o dos demais. Esses clientes não têm a necessidade de estarem sendo atendidos com muita frequência, mas, eles têm a necessidade de um atendimento personalizado e exclusivo.

ii **Grupo II**: Clientes intermediários - clientes que tem um tempo intermediário com a EFD, de médio porte, com frequencia de compras também intermediária. Esses clientes compram itens diversos para garantir aos seus clientes as novidades e variabilidade do mercado. São clientes que já foram pequenos e estão em desenvolvimento. Eles possuem um limite de crédito intermediário. Esses clientes são atendidos com mais frequência do que os de valor agregado, não em busca apenas de produto, mas também, de consultorias de melhorias e para seu desenvolvimento.

iii **Grupo III**: Clientes simples - clientes que estão ou não há muito tempo comprando da EFD. Esses clientes tem uma maior frequência de compra do que os demais, haja vista que eles compram o necessário, sem muita necessidade diversos itens. São clientes pequenos, que compram pouco volume e pouca variedade. Esses clientes não necessitam de muita formalidade em seu atendimento, eles necessitam do básico à sua disposição. São a grande maioria, em volume, da clientela da EFD.

Tabela 3.7: Análise descritiva dos grupos após o agrupamento por K-Médias

Grupo	n	**Variáveis**	$\bar{x}$	s	**Mínimo**	**Máximo**
1	10	Tempo de casa	207	131	26	393
		Checkouts	14	9	1	25
		Limite de crédito	561.500	194.708	290.000	1.000.000
		Média de compra	25.480	14.403	9.732	53.420
		Frequência de compras	4	2	2	10
		Média de itens	23	13	7	40
		Quantidade de pedidos	136	57	41	210
		Valor da venda	2.874.092	1.152.686	1.983.609	5.642.711
		Valor primeira compra	28.457	81.352	468	259.899
		Última_compra	18.221	24.209	308	79.658
2	179	Tempo de casa	181	115	29	496
		Checkouts	10	7	1	60
		Limite de crédito	127.763	74.780	0	455.000
		Média de compra	8.332	6.898	2.519	71.201
		Frequência de compras	6	4	2	33
		Média de itens	38	23	3	117
		Quantidade de pedidos	85	36	12	213
		Valor da venda	596.710	288.520	270.791	1.670.316
		Valor primeira compra	9.375	22.881	103	227.713
		Valor última compra	7.991	9.018	294	91.105
3	10185	Tempo de casa	169	114	0	531
		Checkouts	2	3	1	90
		Limite de crédito	21.053	20.920	0	200.000
		Média de compra	1.304	1.815	32	70.712
		Frequência de compras	39	41	0	407
		Média de itens	14	11	1	170
		Quantidade de pedidos	14	15	1	116
		Valor da venda	21.858	40.274	32	323.453
		Valor primeira compra	1.156	3.612	47	149.514
		Valor última compra	1.288	2.237	126	96.013

4. Conclusões

A EFD adota o *marketing* de segmento, reconhecendo que seus clientes se diferem em seus desejos, poder de compra, localização, atitudes e hábitos de compra. Ela se propõe geograficamente à atender em todo o território nacional, todos os tipos e tamanhos de segmentos comerciais, independente do estilo de vida de cada um. Na EFD a seleção do mercado alvo é feita por cobertura ampla de mercado, ela visa atender a todos os grupos de consumidores com todos os produtos que possam necessitar. A análise de aglomerados para a reestruturação dos clientes da EFD mostrou-se eficaz atendendo o objetivo da mesma de dividí-los em três grupos com características heterogêneas: clientes simples, clientes intermediários e cliente de valor agregado.

Assim sendo, a EFD considerou que a análise de conglomerados por K-Médias foi eficaz para seu problema, nesse estudo feito para o varejo alimentar em Minas Gerais. Como planos para o futuro, a EFD visa manter o padrão de três grupos por segmento (varejo alimentar, farmácia, agro negócios, materiais de construção, eltro eletrônicos e telefonia), mantendo também, a segmentação por unidade federativa. E assim, poder melhorar a forma de atendimento ao cliente e se sustentar no mercado.

Referências Bibliográficas

[1] ABRAM, G. e TREINISH, L.: *An Extended Data-Flow Architecture for Data Analysis and Vizualization*. New York: Computer Graphics, 1ª ed., 1999.

[2] ALVES, M. E. e CAMAROTTO, M. R.: *Comunicação Integrada de Marketing*. IESDE Brasil S.A., 1ª ed., 2012.

[3] CAMPELLO, R.: *Análise de agrupamento*. ICMC/USP, 9(1):2–45, 2015. www.facom.ufu.br/~backes/pgc204/Aula09-Agrupamentos.pdf.

[4] COSTA, R. M.: *Estratégias competitivas e desempenho econômico: o caso da indústria automobilística brasileira de 1986 a 2007*. Dissertação de Mestrado, 2008.

[5] FÁVERO, L. P. ; BELFIORE, P. S. F. e CHAN, B. L.: *Análise de dados: modelagem multivariada para tomada de decisões*. Rio de Janeiro: Elsevier, 1ª ed., 2009.

[6] HAIR, J.F.; ANDERSON, R. E. T. R. e BLACK, W.: *Análise multivariada de dados*. Porto Alegre: Bookman, 5ª ed., 2005.

[7] HARTIGAN, J. A. e WONG, M. A.: *Algorithm AS 136: A k-means clustering algorithm*. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1):100–108, 1979.

[8] JOHNSON, R. e WICHERN, D.: *Applied multivariate statistical analysis*. Upper Saddle River: Pearson education, 6ª ed., 2007.

[9] KOTLER, P.: *Marketing para o Século XXI*. Ediouro, 1ª ed., 2009.

[10] LIMA, C. A. G. de: *Segmentação de Mercado de Clientes Pessoas Jurídicas do Banco do Brasil no Pilar Atacado*. Monografia de MBA, 2005.

[11] LINDEN, R.: *Técnicas de Agrupamento*. Revista de Sistemas de Informação da FSMA, 4(4):18–36, 2009. http://www.fsma.edu.br/si/edicao4/FSMA_SI_2009_2_Tutorial.pdf.

[12] MINGOTI, S.: *Análise de dados através de métodos de estatística multivariada*. UFMG, 1ª ed., 2007.

[13] PEREIRA, H. J.: *Os novos modelos de gestão: análise e algumas práticas em empresas brasileiras*. Tese de Doutorado, 1995.

[14] REIS, E.: *Estatística multivariada*. Lisboa: Sílabo, 2ª ed., 2001.

[15] ROCCA, G. L.: *A Importância da Aplicação do Marketing nos Segmentos de Atacado e Varejo*, 2008. https://www.webartigos.com/artigos/a-importancia-da-aplicacao-do-marketing-nos-segmentos-de-atacado-e-varejo/11021, acessado em 10/08/2017.

[16] SAS: *PROC FASTCLUS: PROC FASTCLUS Statement :: SAS/STAT(R)*. https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_fastclus_sect005.htm, acessado em 29/11/2017.

[17] SCHEER, J. F.; HINES, R. J. O. e M., K. K.: *Classification of Dive Profiles: A Comparison of Statistical Clustering Techniques and Unsupervised Artifical Neural Networks*. Journal of Agricultural, Biological, and Environmental Statistics, 3(4):383–404, 1998. http://www.jstor.org/stable/1400572.

[18] SHIKAMURA, S.: *Coeficiente de determinação*, 2005. http://leg.ufpr.br/~silvia/CE701/node83.html, acessado em 15/11/2017.

[19] T. CALIŃSKI, J. H.: *A dendrite method for cluster analysis. Communications in Statistics-theory and Methods*. 3ª ed., 1974.

[20] VALLI, M.: *Análise de Cluster*. Augusto Guzzo Revista Acadêmica, 5(4):77–87, 2012. http://www.fics.edu.br/index.php/augusto_guzzo/article/view/107.

5. APÊNDICE

```
DATA _NULL_;
     dsid = OPEN("WORK.BASE_DE_DADOS_TCC", "I");
     dstype = ATTRC(DSID, "TYPE");
     IF TRIM(dstype) = " " THEN
          DO;
          CALL SYMPUT("_EG_DSTYPE_", "");
          CALL SYMPUT("_DSTYPE_VARS_", "");
          END;
     ELSE
          DO;
          CALL SYMPUT("_EG_DSTYPE_", "(TYPE=""" || TRIM(dstype) || """)");
          IF VARNUM(dsid, "_NAME_") NE 0 AND VARNUM(dsid, "_TYPE_") NE 0 THEN
               CALL SYMPUT("_DSTYPE_VARS_", "_TYPE_ _NAME_");
          ELSE IF VARNUM(dsid, "_TYPE_") NE 0 THEN
               CALL SYMPUT("_DSTYPE_VARS_", "_TYPE_");
          ELSE IF VARNUM(dsid, "_NAME_") NE 0 THEN
               CALL SYMPUT("_DSTYPE_VARS_", "_NAME_");
          ELSE
               CALL SYMPUT("_DSTYPE_VARS_", "");
          END;
     rc = CLOSE(dsid);
     STOP;
RUN;
```

```
DATA WORK.SORTTempTableSorted &_EG_DSTYPE_ / VIEW=WORK.SORTTempTableSorted;
     SET WORK.BASE_DE_DADOS_TCC;
RUN;
TITLE;
TITLE1 "Cluster Analysis Results";
FOOTNOTE;
FOOTNOTE1 "Generated by the SAS System (& SASSERVERNAME, &SYSSCPL) on %TRIM(%QSYSFUNC(DATE(),
          NLDATE20.)) at %TRIM(%SYSFUNC(TIME(), TIMEAMPM12.))";
```

```
PROC FASTCLUS DATA=WORK.SORTTempTableSorted
     MAXC=3
     MAXITER=50
     REPLACE=FULL
     OUT=WORK.CLKMKMeansDataBASE_DE_DADOS_TCC(LABEL="K-means cluster data for WORK.BASE_DE_DADOS_TCC")
     ;
     VAR Tempo_de_casa QtdCheckOut "Limite_crédito"n MediaCompras FrequenciaCompra_DIAS MediaItensPedido
          QtdPedidos Valor_venda Primeira_compra "Última_compra"n;
RUN;
```