KEVIN TAKANORI SIBATA

SEGMENTAÇÃO DE CLIENTES DE UMA EMPRESA DE MODELO DE ASSINATURA ATRAVÉS DA ANÁLISE DE CLUSTERS

São Paulo

2017

KEVIN TAKANORI SIBATA

SEGMENTAÇÃO DE CLIENTES DE UMA EMPRESA DE MODELO DE ASSINATURA ATRAVÉS DA ANÁLISE DE CLUSTERS

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

São Paulo

2017

KEVIN TAKANORI SIBATA

SEGMENTAÇÃO DE CLIENTES DE UMA EMPRESA DE MODELO DE ASSINATURA ATRAVÉS DA ANÁLISE DE CLUSTERS

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

Orientador: Prof. Dr. Davi Noboru Nakano

São Paulo

2017

FICHA CATALOGRÁFICA

Sibata, Kevin Takanori

Segmentação de clientes de uma empresa de modelo de assinatura através da análise de clusters / K. T. Sibata -- São Paulo, 2017.

127 p.

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

1.Segmentação de mercado 2.Análise de conglomerados I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Produção II.t.

Dedico este trabalho à minha família, por todo apoio ao longo da minha trajetória.

AGRADECIMENTOS

Aos meus familiares, por sempre me apoiarem nos momentos mais difíceis e dedicarem esforços para me ajudar a realizar minhas conquistas.

Aos meus colegas da USP, POLI e Engenharia de Produção, pelos momentos de alegria e de superação. São muitos os amigos que participaram da minha jornada acadêmica, mas gostaria de agradecer principalmente para Cesar, Diogo, Erik, Sandro, Pedro, Gabriel e Bianca, melhores amigos da faculdade que levo para vida toda.

Aos meus colegas de trabalho da Best Berry, a empresa que acreditou no meu potencial profissional e me estimula a sempre dar o meu melhor. Em especial, Roberto, Alberto e Na, meus grandes mentores.

Por fim, mas não menos importante, ao Prof Davi Nakano, pela paciência e orientação, essenciais para que este trabalho fosse desenvolvido.

“Imagination is more important than knowledge. For knowledge is limited, whereas imagination embraces the entire world, stimulating progress, giving birth to evolution.”

(Albert Einstein)

RESUMO

O acesso rápido e prático a uma ampla gama de informações permitiu que os clientes consigam avaliar melhor suas escolhas e obter seus produtos de forma simples e a um preço menor. Consequentemente, as empresas precisaram também se adaptar a este novo contexto, principalmente na entrega de valor para seus clientes e na construção de relacionamento. Kotler & Armstrong (2015) destacam que a era digital propiciou que as empresas consigam aprender mais sobre os clientes e rastreá-los, possibilitando a criação de produtos e serviços com maior grau de customização. Nesse sentido, a segmentação do mercado é uma etapa essencial para a definição dos grupos a serem desenvolvidos e na geração de informações para realizar esta diferenciação. Os potenciais ganhos com tal alternativa são a redução em custo, maior eficiência no uso de recursos e a maior rapidez de ação e resposta.

O presente trabalho realizou o estudo dos segmentos de clientes de uma empresa de modelo de assinatura. Seu objetivo era a identificação dos segmentos mais rentáveis e na elaboração das estratégias de Marketing para melhorar a eficiência no uso dos recursos e aumentar o faturamento. Para tanto, foi empregada a ferramenta de análise de clusters com as informações dos clientes contidas no banco de dados da empresa. Após a obtenção, caracterização e validação dos 6 segmentos e suas iniciativas, o projeto foi avaliado pela empresa. A recepção foi positiva, resultando na discussão da implementação das iniciativas e também no estudo de outras ações baseadas nesse trabalho.

Palavras-chave: segmentação de clientes; estratégia de Marketing; análise de clusters; modelo de assinatura.

ABSTRACT

The fast and practice access to a wide range of information has allowed customers to evaluate better their choices and to obtain their products simply and at a lower price. Consequently, companies also needed to adapt to this new context, mainly in delivering value to their customers and building relationships. Kotler & Armstrong (2015) point out that the digital era has allowed companies to learn more about their customers and to track them, making possible the creation of products and services with a higher degree of customization. In this sense, market segmentation is an essential step for defining the groups to be developed and generating information to accomplish this differentiation. The potential gains from this alternative are the reduction in cost, greater efficiency in the use of resources and the greater speed of action and response.

The present work carried out the study of the customer segments of a subscription model company. Its objective was to identify the most profitable segments and to elaborate the Marketing strategies to improve the efficiency in the use of the resources and to increase the revenue. To do so, the cluster analysis tool was used with the information of the clients contained in the company's database. After obtaining, characterizing and validating the 6 segments and their initiatives, the project was evaluated by the company. The reception was positive, resulting in the discussion of the implementation of the initiatives and also in the study of other actions based on this work.

Keywords: customer segmentation; Marketing strategy; cluster analysis; subscription model.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

SUMÁRIO

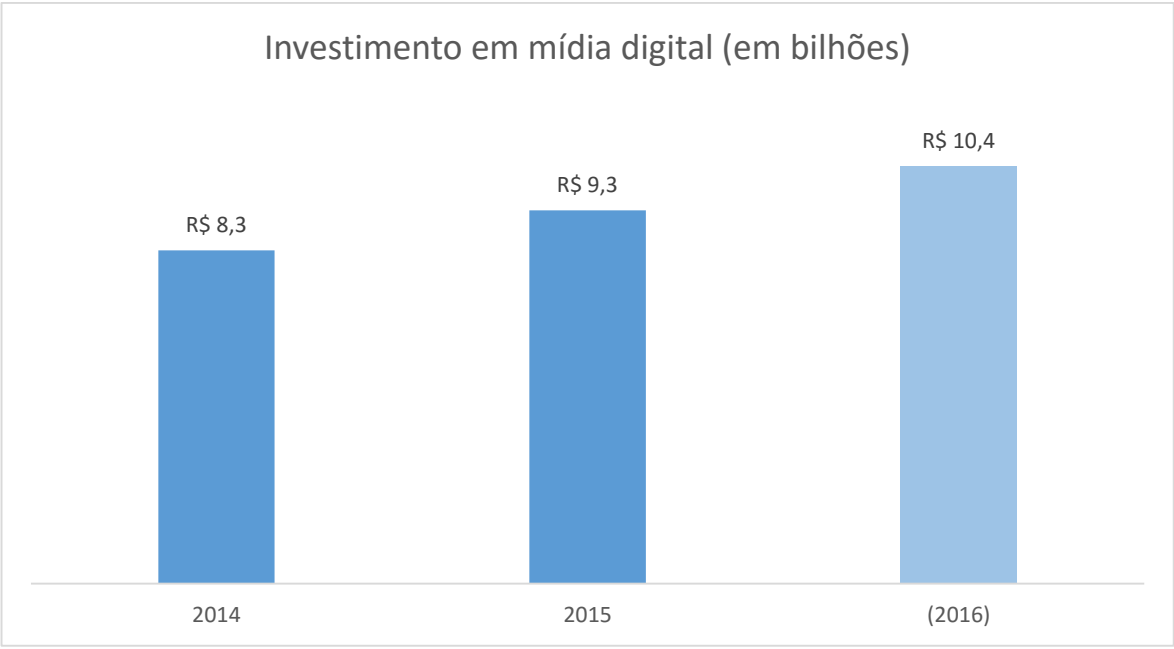
1. INTRODUÇÃO

O avanço acelerado da tecnologia promoveu grandes mudanças na maneira como vivemos. O acesso rápido e prático a uma ampla gama de informações permitiu que os clientes consigam avaliar melhor suas escolhas e obter seus produtos de forma simples e a um preço menor. Consequentemente, as empresas precisaram também se adaptar a este novo contexto, principalmente na entrega de valor para seus clientes e na construção de relacionamento.

Kotler & Armstrong (2015) citam a geração de novas ferramentas de comunicação, propaganda e construção de relacionamento, tais como a propaganda online, as redes sociais e os aplicativos para *smartphones*. Além disso, os autores destacam que esta era digital propiciou que as empresas consigam aprender mais sobre os clientes e rastreá-los, possibilitando a criação de produtos e serviços com maior grau de customização.

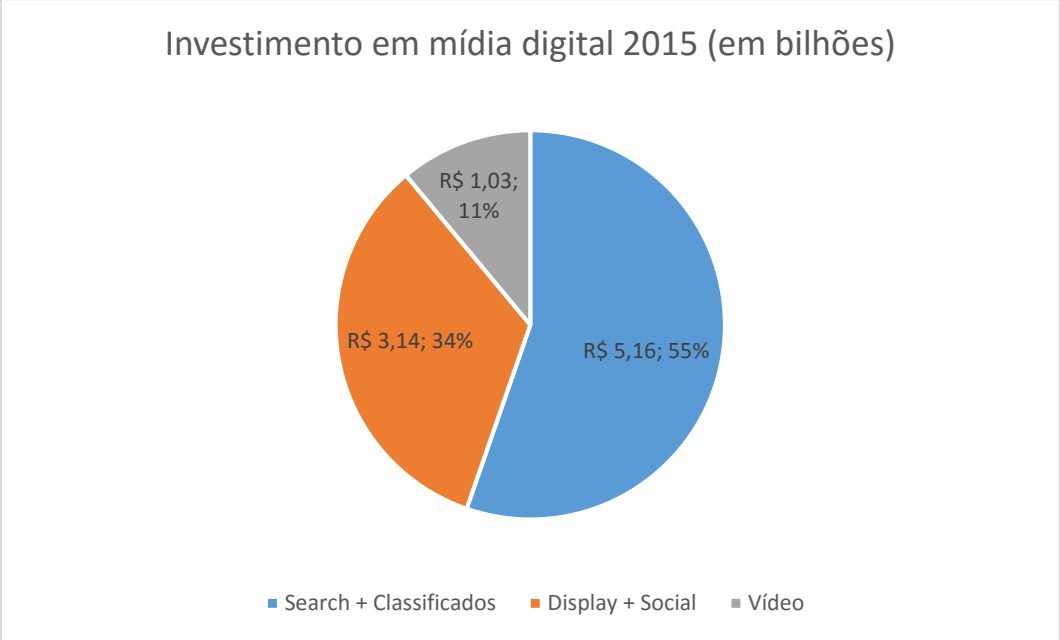
De acordo com o IAB Brasil (*Interactive Advertising Bureau*), o investimento em mídia digital foi de R$9,3 bilhões em 2015, um crescimento de 14% em relação ao ano anterior. A projeção para 2016 é de R$10,4 bilhões, ressaltando assim a importância do meio digital para o Marketing das empresas. A Figura 1 ilustra este crescimento enquanto que Figura 2 é a distribuição do investimento de verbas publicitárias em 2015.

Figura 1: Investimento em mídia digital.



Fonte: Adaptado de IAB Brasil.

Figura 2: Distribuição do investimento em mídia digital em 2015.



Fonte: Adaptado de IAB Brasil.

Esta tendência, chamada de Marketing Direto, consiste em se conectar, de maneira direta, a consumidores cuidadosamente definidos como alvo, muitas vezes em uma base individual e interativa (Kotler & Armstrong, 2015). Os principais ganhos que tornam esta alternativa interessante são a redução em custo, maior eficiência no uso de recursos e a maior rapidez de ação e resposta. Aliado ao contexto online atual, o Marketing Direto Online é amplamente utilizado por empresas como a Amazon.com, o eBay, a Priceline, a Netflix e a GEICO.

Para tanto, um bom banco de dados de clientes é essencial para que as empresas consigam as informações necessárias, tanto de clientes individuais, existentes ou ainda potenciais. Diversos tipos de dados podem ser arquivados, tais como dados geográficos (endereço, região), demográficos (idade, renda, membros da família, datas de aniversário), psicográficos (atividades, interesses e opiniões) e de comportamento de compra (preferências e análises de periodicidade, frequência e valor monetário das compras passadas) (Kotler & Armstrong, 2015).

Kotler & Armstrong (2015) defendem que existem diversas aplicações do banco de dados. Além do uso das informações dos clientes para o ajuste nas ofertas e comunicações ao mercado de acordo com as características dos segmentos ou indivíduos, há a possibilidade da localização de bons clientes potenciais e a geração de *leads* de vendas.

Baseado em todos estes potenciais ganhos que podem ser obtidos através do Marketing Direto Online e o uso das informações armazenadas no banco de dados de clientes, o presente trabalho se propõe a explorar esta alternativa em uma empresa. Para tanto, será realizada a análise estatística dos principais perfis de clientes da empresa (segmentos de clientes) com o auxílio da ferramenta de análise de clusters. A partir deste aprendizado, espera-se a elaboração de abordagens mais eficazes e eficientes para a aquisição de novos clientes com características semelhantes aos segmentos.

2. DESCRIÇÃO DA EMPRESA

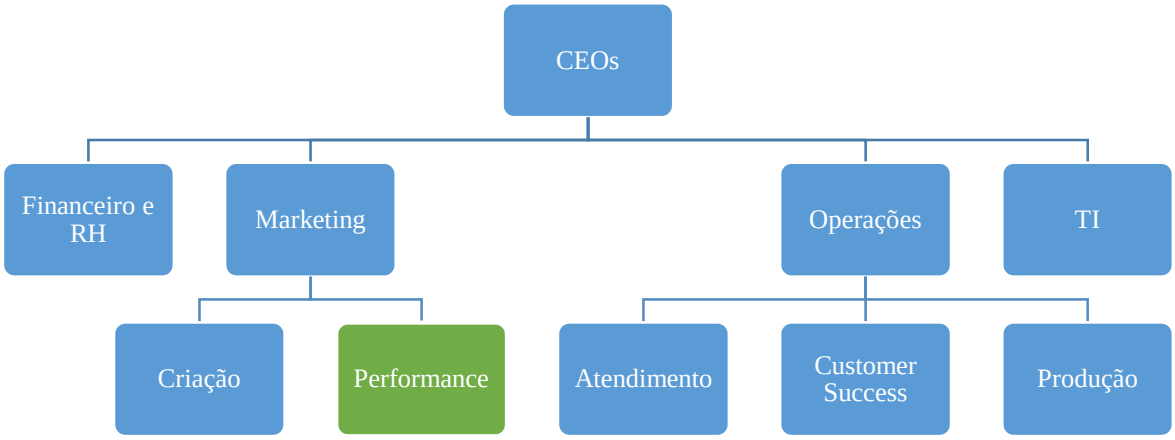
A empresa que será desenvolvido o trabalho é a Best Berry, uma empresa que oferece um serviço de assinatura de *snacks* saudáveis. Os assinantes pagam uma taxa de mensal conforme o plano escolhido (10 *snacks* com 5 variações de sabores por caixa, ou 18 *snacks* e 6 variações de sabores por caixa) e recebem todos os meses uma caixa com os produtos. A Figura 3 apresenta o produto da empresa.

Figura 3: Produto da Best Berry.

Fonte: Site da Best Berry.

A empresa é considerada uma *startup*, que foi criada no final do ano de 2014. Atualmente, conta com aproximadamente 30 funcionários, os quais estão divididos nas seguintes áreas (Figura 4):

Figura 4: Organograma da empresa.



- CEOs: sócios-fundadores da empresa.
- Financeiro e RH: Cuidam das operações financeiras e dos processos de recursos humanos.
- Marketing: Área responsável pela aquisição de novos clientes.
    - Criação: Time de planejamento e execução de campanhas de Marketing, ações de promoção da marca, relacionamento de parcerias e geração de peças de design.
    - Performance: Atuam no investimento em mídias online pagas para aquisição de clientes e melhorias de usabilidade do site. É a área em que o trabalho foi desenvolvido.
- Operações: Área responsável pelo atendimento ao cliente, entrega do produto e retenção de assinantes.
    - Atendimento: Equipe de SAC, responsável por tirar dúvidas e encontrar soluções de clientes ou potenciais clientes.
    - *Customer Success*: Time de planejamento e execução de ações para satisfação e retenção dos clientes ativos.
    - Produção: Realizam o planejamento de compra de insumos, e a produção de *snacks* e caixas dos clientes.
- TI: Equipe de desenvolvimento do site e de ferramentas online internas da empresa.

Em relação ao Marketing de Performance, área em que foi desenvolvida o trabalho, a empresa atua em vários canais de aquisição de novos assinantes, sendo que os principais são:

- Facebook e Instagram: Divulgação de anúncios nas redes sociais do grupo Facebook (Figura 5).
- Google Adwords: Divulgação do link do site em certas palavras chave utilizadas na pesquisa do Google (Figura 6).
- Google Display: Divulgação de banners na rede de display do Google, tais como matérias em blogs, sites de notícias, entre outros (Figura 7).
- Email Marketing: Emails que são enviados para os potenciais clientes que não finalizaram a sua assinatura (Figura 8), e para os leads que se cadastraram através dos *pop-ups* (Figura 9).
- Afiliados: Rede de anunciantes que recebem comissão por assinatura trazida. Os canais de mídia empregados variam conforme o afiliado.

- Orgânico: Inclui os canais de mídia com links não pagos, como por exemplo acesso por meio de portais de notícia, blogs, pesquisa não paga no Google, acesso direto via URL do site.

Figura 5: Exemplo de anúncio no Facebook.

Figura 6: Exemplo de anúncio no Google Adwords.

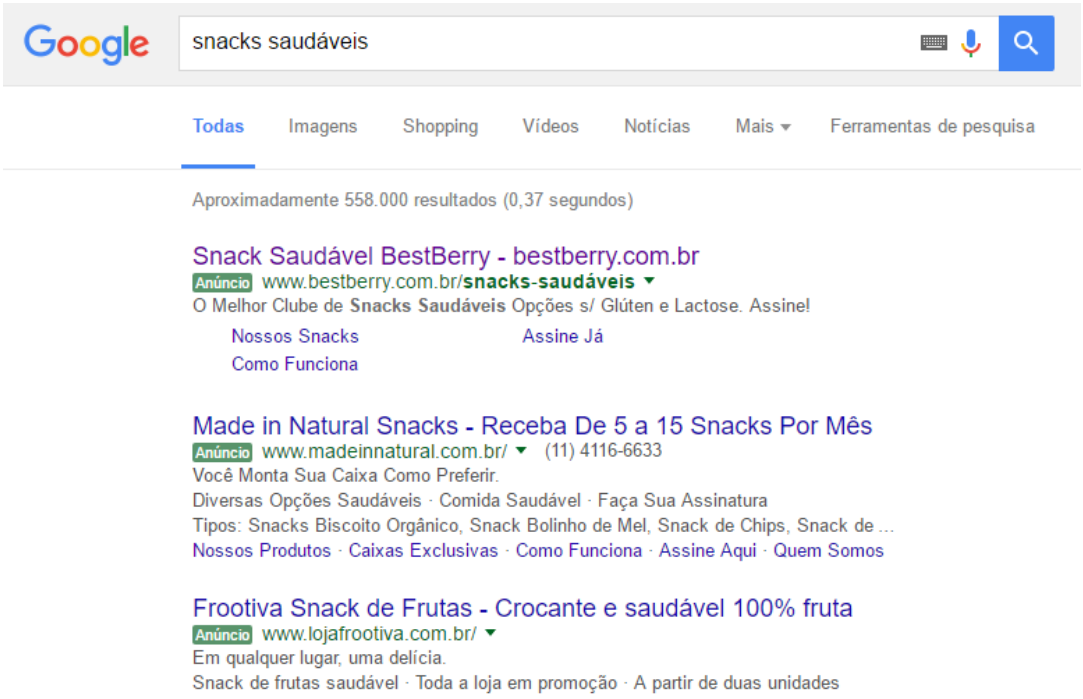


Figura 7: Exemplo de banner de Google Display.

Figura 8: Exemplo de email marketing.

Figura 9: Exemplo de *pop-up* para captação de leads.

Fonte: Site da Best Berry.

Através de análises prévias sobre o perfil dos assinantes, a Best Berry definiu seu público alvo com as seguintes características: gênero feminino, faixa etária de 22 a 55 anos, habitam na região Sudeste do país, classe social B ou maior, pessoas que procuram um estilo de vida mais saudável e bem estar. Todas as campanhas de Marketing para a aquisição de novos assinantes utilizam estas informações para a elaboração dos anúncios.

Com o auxílio das ferramentas de rastreamento, são trabalhados também os visitantes recentes no site, segmentados conforme a etapa do funil de compra em que o usuário finalizou sua interação, como por exemplo a página de escolha do plano, cadastro, preenchimento dos dados de pagamento.

2.1.Motivação e definição do problema

Como qualquer outra empresa, a Best Berry deseja crescer sua carteira de assinantes, aumentando assim seu faturamento. Tratando-se de um modelo de negócios de assinatura, o crescimento da carteira pode ser dado como:

$$crescimento = novos\ clientes - clientes\ que\ cancelaram$$

Uma das vertentes considerada pela empresa para atuar no crescimento é o estudo detalhado sobre a sua base de clientes, a qual contém diversas informações pouco exploradas contidas nos bancos de dados.

Em termos de aquisição de novos clientes, apesar da empresa apresentar uma estratégia de público alvo bem direcionada, abordando características demográficas e psicográficas, nota-se que a segmentação explora pouco as informações contidas sobre o cliente. Dentre as três dimensões de dados encontrados nos bancos de dados, a referente ao comportamento de compra foi pouco utilizada. Segundo Brito et al (2015), quando dados de comportamento estão disponibilizados, então é possível implementar uma segmentação mais refinada.

Além disso, a análise dos clientes permite verificar se o público alvo da empresa é de fato o mais atraente, ou se existe algum outro segmento que poderia ser atendido. A segmentação de clientes permite um estudo preliminar para elaboração de ações de retenção mais direcionadas.

Diante de todas estas oportunidades ilustradas, permitiu-se então o desenvolvimento do trabalho em conjunto com a empresa.

3. REVISÃO BIBLIOGRÁFICA

3.1.Conceitos de Marketing

Segundo Kotler & Armstrong (2015), o Marketing é o processo pelo qual as empresas criam valor para os clientes e constroem fortes relacionamentos com eles para capturar valor deles em troca. Os autores definem que o Marketing apresenta dois principais objetivos: i) atrair novos clientes, prometendo valor superior; ii) manter e cultivar os clientes atuais, entregando satisfação.

Para os próximos tópicos, será detalhada a visão de Marketing como um processo e o papel da segmentação de mercado, a qual apresenta uma lógica semelhante à segmentação de clientes.

3.1.1. O Processo de Marketing

Como explicado anteriormente, Kotler & Armstrong (2015) entendem que o Marketing constitui um processo, cujo modelo é formado por cinco etapas (Figura 10).

As quatro primeiras etapas do processo de marketing se concentram em criar valor para os clientes.

Primeiramente, procura-se entender o mercado através da pesquisa das necessidades dos clientes e da administração das informações de Marketing.

A segunda etapa é a elaboração da Estratégia de Marketing orientada para o cliente. Para tanto, são respondidas duas questões. A primeira pergunta é: “A quais clientes atenderemos?”. Dado que é impossível agradar todos os clientes de forma igualitária, as empresas devem concentrar os recursos nos clientes que são mais lucrativos e que ela consegue atender melhor. Para tanto, são realizadas a segmentação de mercado e a seleção de mercado-alvo. A segunda é: “Como podemos atender melhor aos clientes-alvo?”. Define-se uma proposição de valor que representas os valores a serem entregues para conquistar os clientes-alvo (diferenciação e posicionamento).

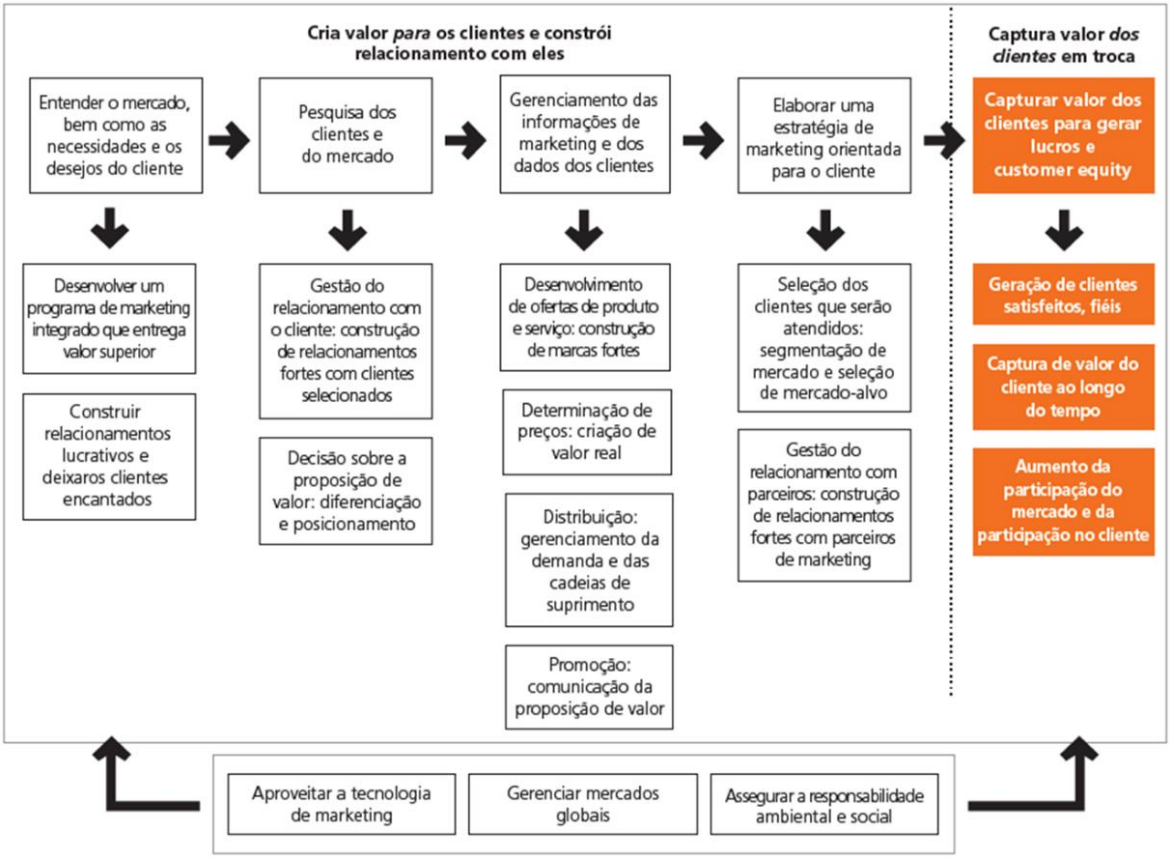
Finalizada a Estratégia de Marketing, constrói-se um programa de marketing integrado, formado pelos quatro elementos do mix de marketing (produto, preço, promoção e praça), o qual materializa a estratégia em valor real para os clientes.

A quarta etapa é a construção de relacionamentos lucrativos e com valor para os clientes-alvo. Para tanto, a empresa utiliza o conhecimento obtido através da gestão do relacionamento com cliente e também do estabelecimento de vínculos com os parceiros de marketing.

Na etapa final, a empresa colhe as recompensas de seu forte relacionamento com os clientes ao capturar valor deles. Este estado é decorrente da repetição de compra dos clientes altamente satisfeitos.

Os autores incluem mais três aspectos decorrentes do contexto contemporâneo de constantes mudanças. Ao construir relacionamento com os clientes e os parceiros, elas devem aproveitar as tecnologias de marketing, explorar as oportunidades globais e certificar-se de que estão agindo de maneira ética e socialmente responsável.

Figura 10: Modelo do Processo de Marketing.



Fonte: Kotler & Armstrong (2015).

3.1.2. Segmentação de mercado

Os consumidores apresentam caraterísticas diferentes entre si e dada a limitação de recursos e competências das empresas, é impossível que ela consiga atender a todos os perfis de clientes de forma igualitária. Neste capítulo, serão apresentados os principais conceitos relacionados à segmentação de mercado.

3.1.2.1.Perspectiva histórica

Conforme estudado por Smith (1956), um dos pioneiros na literatura em segmentação de mercado, a teoria de competição perfeita e puro monopólio não se adequam ao cenário de negócios atual, sendo o mais comum a presença de mercados imperfeitos. A competição perfeita assume homogeneidade dos componentes de mercado, tanto da demanda quanta a oferta. Segundo o autor, em termos de oferta, a presença de diversidade em mercados é decorrente dos seguintes pontos:

- Variações nos equipamentos de produção e métodos ou processos usados por diferentes manufaturas de produtos projetados para o mesmo ou similar uso;
- Recursos especializados ou superiores utilizados com maior preferência por manufatureiros bem situados;
- Progresso desigual entre competidores em design, desenvolvimento, e melhoria de produtos;
- Inabilidade de manufatureiros de algumas indústrias em eliminar variações de produto apesar da aplicação de técnicas de controle de qualidade;
- Variações nas estimativas dos produtores da natureza da demanda de mercado em relação a sensibilidade de preço, cor, material, ou tamanho do pacote.

Com relação a demanda, a estratégia de Marketing apresentava uma abordagem convergente, em que as demandas individuais pela variedade de produtos eram atendidas por uma única ou limitada oferta ao mercado, a qual era atingida pela diferenciação do produto através de publicidade e promoção. Entretanto, em alguns casos, era necessário aceitar a divergência da demanda em termos de característica de mercado, e então ajustar as linhas de produtos e estratégias de Marketing de acordo com essa. Esta falta de homogeneidade da

demanda, segundo Smith (1956), pode ser baseada em diferentes costumes, desejo por variedade, ou desejo por exclusividade ou pode surgir de diferenças básicas das necessidades de usuário.

Diante deste cenário heterogêneo, Smith (1956) destaca duas estratégias para explorar tais oportunidades: diferenciação de produto e segmentação de mercado. O resultado delas pode ser parecido, tais como diferenças nos produtos, imagem, distribuição e/ou promoção, no entanto, a diferenciação de produto parte da mudança na oferta, o que implica na adaptação da demanda conforme a variedade da oferta. A segmentação de mercado inicia com estudo do mercado e suas necessidades para que então seja elaborada a oferta para cada segmento. De forma resumida, a diferenciação de produto é uma abordagem "de dentro para fora" e a segmentação de mercado uma "de fora para dentro" (Evans, 2004).

Vale ressaltar que, para a segmentação, nem sempre é necessário desenvolver um produto diferente para cada segmento. Por exemplo, é possível estabelecer uma política de preços para um mesmo produto (gasolina, energia elétrica, passagem de trem) ou para um segmento baseado em níveis de compra repetida ou fidelidade (Evans, 2004). Outro ponto importante é o excesso da segmentação, chamado por Evans (2004) de fragmentação do mercado; ou seja, a idealização de segmentos muito pequenos, que não rentáveis e tornam-se ineficientes.

Entretanto, o avanço tecnológico permite uma análise mais detalhada do comportamento do cliente e em nível individual, possibilitando a construção do relacionamento a longo prazo com os clientes que mais contribuem para a posição financeira da empresa (Evans, 2004).

3.1.2.2.Definição e objetivo

De acordo com Blythe (2005), o princípio básico da segmentação de mercado é que os mercados não são homogêneos e que faz sentido, em termos comerciais, diferenciar as ofertas para diferentes grupos de clientes.

O objetivo da segmentação é identificar um grupo de pessoas que possui uma(s) necessidade(s) que pode ser satisfeita por um único produto, para então concentrar os esforços de marketing da empresa da melhor forma efetiva e econômica (Evans, 2004).

Dentre as vantagens da segmentação, Blythe (2005) enuncia as seguintes (Quadro 1):

Quadro 1: Vantagens da segmentação.

Vantagem	Explicação
Análise do cliente	Através da segmentação, a empresa consegue entender melhor seus melhores clientes.
Análise do concorrente	É mais fácil reconhecer e enfrentar a concorrência concentrando-se em uma pequena parte do mercado.
Alocação efetiva de recurso	Os recursos escassos das empresas podem ser concentrados com maior efetividade em poucos clientes, ao invés de difundi-los ao longo das massas.
Planejamento estratégico de Marketing	Planejar se torna mais fácil quando a empresa tem uma clara imagem de seus melhores clientes.
Expansão do mercado	Uma boa segmentação pode aumentar o tamanho do mercado trazendo novos clientes, os quais se enquadram no perfil típico de cliente, mas não reconheciam o produto.

Fonte: Adaptado de Blythe (2005).

3.1.2.3.Bases de segmentação

Blythe (2005) categoriza as variáveis de segmentação em 4 grupos. Kotler & Armstrong (2015) propõe de forma semelhante, acrescentando mais um grupo, resultando em:

- Segmentação geográfica: Divisão de um mercado em diferentes unidades geográficas, como países, regiões, estados, cidades ou até mesmo bairros.
- Segmentação demográfica: Divisão de um mercado em segmentos com base em variáveis como idade, estágio no ciclo de vida, sexo, renda, ocupação, grau de instrução, religião, etnia e geração.
- Segmentação psicográfica: Divisão de um mercado em diferentes grupos com base na classe social, no estilo de vida ou em traços da personalidade.
- Segmentação comportamental: Divisão de um mercado em segmentos com base no conhecimento que os consumidores possuem sobre um produto, nas atitudes que têm direcionadas a ele, no uso que fazem desse produto e em suas reações a ele.
- Segmentação por benefício: Divisão de um mercado em segmentos de acordo com os diferentes benefícios que os consumidores procuram em um produto.

O Quadro 2 apresenta os grupos e as variáveis de forma resumida.

Quadro 2: Principais variáveis de segmentação para mercados consumidores.

Variável de segmentação	Exemplos
Geográfica	Países, regiões, estados, cidades, bairros, densidade populacional (urbana, suburbana, rural), clima
Demográfica	Idade, estágio no ciclo de vida, sexo, renda, ocupação, grau de instrução, religião, etnia, geração
Psicográfica	Classe social, estilo de vida, personalidade
Comportamental	Ocasiões
Benefícios	Status do usuário

Fonte: Kotler & Armstrong (2015).

Evans (2004) defende que as abordagens de segmentação podem ser classificadas em objetivas ou subjetivas. Uma base objetiva pode ser mensurada sem ambiguidade ou obtida por registros de transações. A subjetiva precisa ser mensurada com os próprios respondentes e são geralmente "construídas mentalmente", como as atitudes e intenções.

As bases de segmentação podem também apresentar níveis (Quadro 3). No nível geral, a segmentação é baseada nas características permanentes ou de longo termo do clientes, as quais são iguais para diferentes produtos, serviços ou situações de uso. Para a segmentação de domínio específico, existem diferentes classes de produtos e domínios de consumo. Finalmente, no caso do nível específico, os clientes são segmentados, como por exemplo, em usuários experientes ou inexperientes de marcas específicas.

Quadro 3: Classificação das variáveis de segmentação.

	Objetiva	Subjetiva
Nível geral (consumo)	Idade, Nível educacional, área geográfica	Estilo de vida, valores gerais, personalidade
Nível domínio específico (classe de produto)	Frequência de uso, substituição, complementariedade	Percepção, atitude, preferência, interesses, opiniões, valores de domínio específico
Nível específico (marca)	Lealdade a marca (comportamental), frequência de uso	Lealdade a marca (atitude), preferência de marca, intenção de compra

Fonte: Adaptado de Evans (2004).

3.1.2.4. Validação dos segmentos

Como existem diferentes formas de segmentar um mercado, já que cada caso apresenta seu conjunto único de variáveis e há muitas maneiras de se realizar o agrupamento, é importante avaliar a eficácia de uma determinada segmentação. Segundo Blythe (2005) e Kotler & Armstrong (2015), os segmentos obtidos são válidos se os mesmos apresentarem os seguintes requisitos:

- Mensuráveis: o tamanho, o poder de compra e o perfil dos segmentos podem ser mensurados.
- Acessíveis: os segmentos de mercado podem ser alcançados e atendidos de maneira eficiente.
- Substanciais: os segmentos de mercado são grandes e lucrativos o suficiente para serem atendidos. Um segmento deve ser o maior grupo homogêneo possível, que compense o desenvolvimento de um programa de marketing sob medida para ele.
- Diferenciáveis: os segmentos são conceitualmente distintos e respondem de maneira diferente a programas e elementos do mix de marketing diversos.
- Acionáveis: podem ser desenvolvidos programas eficientes para atrair os segmentos e atender a eles.

Evans (2004) apresenta uma validação semelhante à proposta anterior, sendo a diferença decorrente da inclusão da questão comportamental dos segmentos. Segundo o autor, os segmentos devem apresentar os seguintes critérios:

- Tipificando os segmentos
    - Identificação: Diferenciação do segmento dos demais segmentos.
    - Mensurabilidade: Identificação dos segmentos em termos de diferenças em características individuais e familiares ou outras características “mensuráveis” devem ser possíveis.
- Homogeneidade
    - Variação: Heterogeneidade entre os segmentos em termos de resposta comportamental.
    - Estabilidade: Segmentos devem ser relativamente estáveis ao longo do tempo e a mudança de clientes de um segmento para outro não deve ser frequente.

    - Congruência: Homogeneidade dentro dos segmentos em termos de respostas comportamentais.
- Utilidade
    - Acessibilidade: Segmentos devem ser acessíveis em termos de comunicação de mídia e canais de distribuição. Ou seja, é possível alcançar o segmento.
    - Substancialidade: Segmentos devem ter tamanho suficiente para permitir ações de marketing específicas. Isso não significa que os segmentos devem ser especificamente grandes, mas rentáveis o suficiente.
- Critérios estratégicos:
    - Potencial: Os segmentos devem ter potencial suficiente para os objetivos de marketing (por exemplo, rentabilidade).
    - Atratividade: Segmentos devem ser atraentes estruturalmente para o produtor.

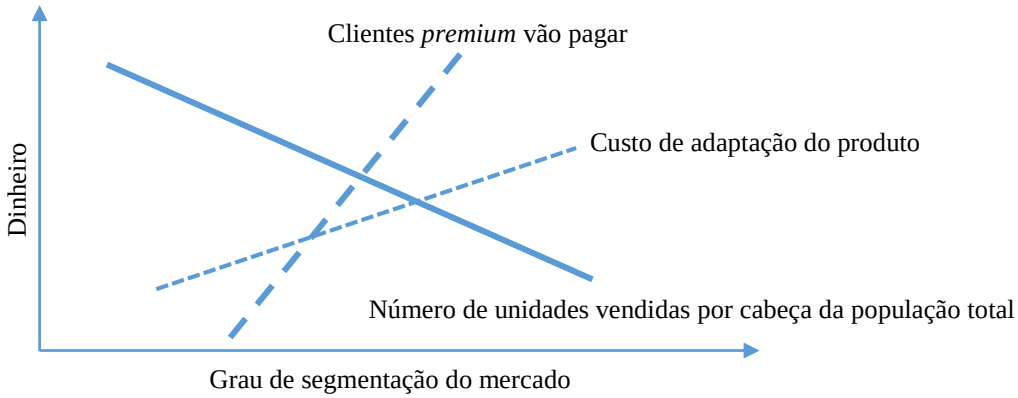
Como mencionado anteriormente, a segmentação realizada de forma correta aumenta o retorno da empresa. Blythe (2005) analisa os principais *trade-offs* a serem considerados para verificar a efetividade da segmentação (Figura 11).

Através da segmentação, a empresa conseguirá atender melhor os clientes *premium* com os produtos mais adequados e com uma comunicação mais apropriada. Sendo assim, os clientes estarão mais dispostos a pagar mais por este valor extra.

O segmento será rentável desde que o preço pago pelos clientes *premium* for maior do que o custo de manufatura necessário para efetuar a adaptação. Outro *trade-off* importante a ser lembrado é que, apesar do maior preço pago, quanto maior a segmentação, menor é o mercado em relação ao total e, consequentemente, o número de unidades vendidas por pessoa decai.

Resumidamente, quando o custo de adaptação for maior que o preço pago pelos clientes *premium*, a adaptação não deve ser feita; e quando o preço pago é maior do que o custo de adaptação, pode ser vantajoso realizar a mudança, porém, a empresa deve considerar a redução do volume de vendas.

Figura 11: Análise de trade-off da segmentação.



Fonte: Adaptado de Blythe (2005).

3.1.3. Resumo do capítulo

O objetivo deste capítulo foi a introdução dos conceitos de Marketing, os quais estão relacionados ao desenvolvimento do trabalho. Iniciou-se a discussão com a definição do Marketing e seus dois objetivos, sendo que o primeiro trata da aquisição de clientes, que é o tema central deste trabalho.

Segundo Kotler & Armstrong (2015), o Marketing pode ser visto como um processo dividido em cinco etapas, as quais foram detalhadas para o entendimento geral e o contexto pelo qual a segmentação dos clientes está inserida.

Por fim, foi discutido a segmentação de clientes, que, caso seja realizado de forma adequada e considerado os *trade-offs*, permite o uso mais eficiente e eficaz dos recursos e ações de Marketing de modo a maximizar o retorno da empresa. Os principais pontos para a realização desta tarefa são a determinação das variáveis para o entendimento dos clientes, o método de segmentação (o qual varia conforme o contexto da empresa) e a validação dos segmentos.

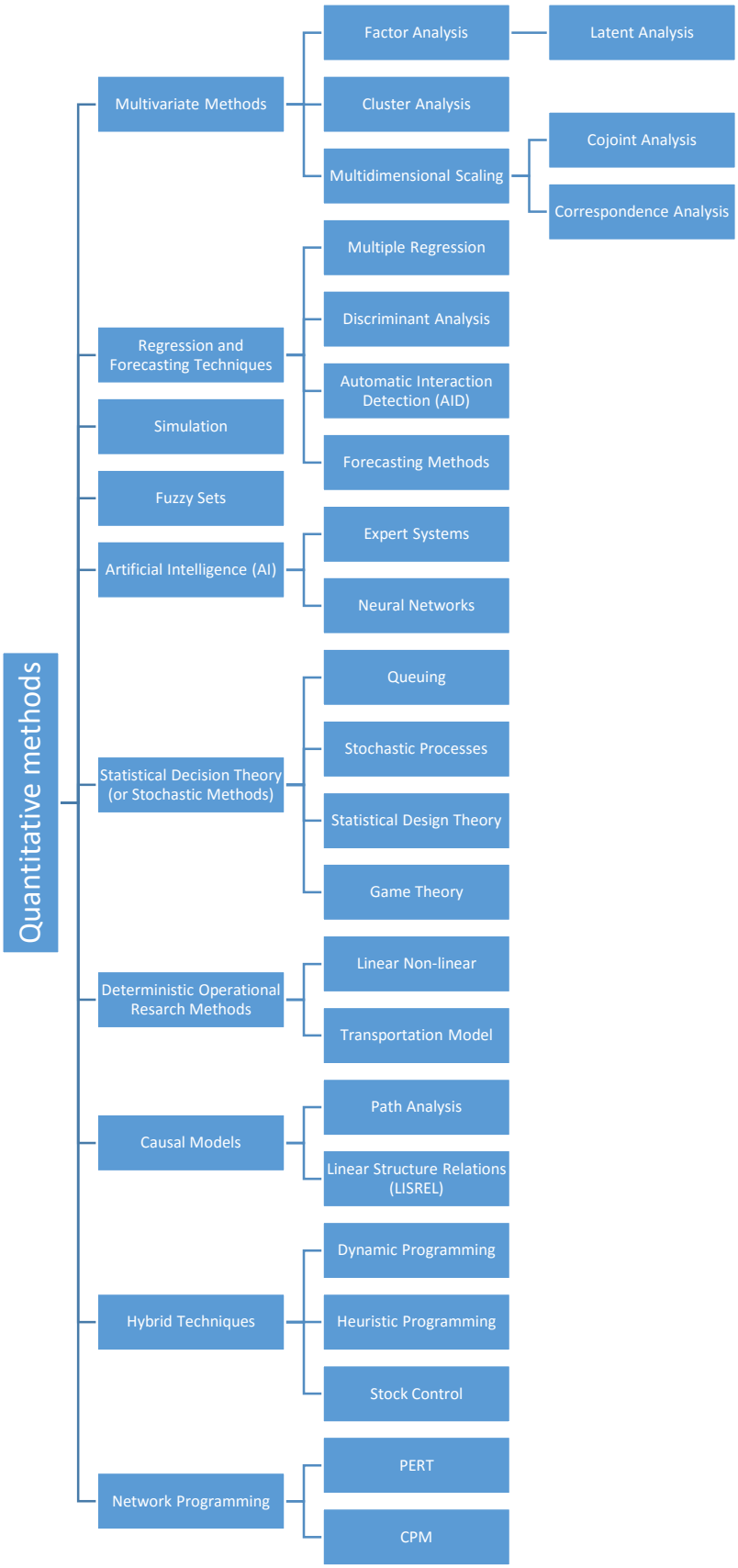
3.2.Métodos quantitativos em Marketing

Segundo Moutinho & Meidan (2004), o uso tardio de ferramentas e métodos quantitativos em Marketing se deve a quatro principais fatores:

- Complexidade dos fenômenos de Marketing: Quando o ambiente é estimulado, as repostas tendem a serem não lineares, a exibir efeito limiar (um nível mínimo de estímulo precisa ser aplicado antes que a resposta ocorra), a ter efeito de repercussão (por exemplo, a resposta do anúncio deste período ocorrerá no futuro) e a decair com o tempo pela ausência de estímulos seguintes.
- Efeitos de interação das variáveis de Marketing: Significa que o impacto de uma única e controlada variável de Marketing é difícil de ser determinado por conta das interações da variável com o ambiente e com outras varáveis. De fato, a maioria das variáveis no Marketing são interdependentes e inter-relacionadas.
- Problemas de mensuração no Marketing: É comum a dificuldade para mensurar diretamente a resposta dos consumidores a certos estímulos e, por conseguinte, técnicas indiretas são empregadas.
- Instabilidade das relações de Marketing: O relacionamento entre respostas de Marketing e as variáveis de decisão do Marketing tende a ser instável devido a mudanças no gosto, atitude, expectativa e outros. Estes fatores tornam contínuas as mensurações de mercado e a revisão das decisões cruciais.

Existem diversas ferramentas que podem ser aplicadas no Marketing. Moutinho & Meidan (2004) apresentam uma taxonomia, a qual contém as principais delas (Figura 12). Cada método possui a sua própria literatura e, para não tornar o assunto muito extenso, serão apresentados breves resumos sobre elas, conforme a descrição destes autores. O Capítulo 3.3 entra em maiores detalhes sobre a ferramenta empregada neste trabalho, cuja justificativa se encontra no Capítulo 3.2.1.

Figura 12: Principais métodos quantitativos de Marketing.



Fonte: Adaptado de Moutinho & Meidan (2004).

Conforme Moutinho & Meidan (2004), os métodos multivariados (*Multivariate Methods*) são os mais empregados na área de Marketing. O objetivo deles é tentar investigar a relação e os padrões das decisões de Marketing que emergem como resultado da interação e interdependência entre as variáveis ao mesmo tempo. Os principais métodos relacionados são a análise de fatores, a análise latente, a análise de cluster, o escalonamento multidimensional, a análise conjunta e a análise de correspondência.

A análise de fatores (*factor analysis*) tem como objetivo a identificação de relacionamentos entre variáveis de modo a estabelecer a influência delas. Dentre as aplicações em Marketing, pode-se citar a determinação de imagens de Marketing corporativo, estudo do comportamento do consumidor e atitudes. Em relação as vantagens deste método, destaca-se a redução de dados e a identificação dos fatores que subscrevem as características dos dados. A principal limitação é que o método exige o uso de dados contínuos.

A análise latente (*latent analysis*) é um método empregado para investigação de fatores manifestantes e latentes através da estimativa dos parâmetros latentes. Pesquisa de segmentação e análise de estrutura de mercado são possíveis aplicações na área de Marketing. Além disso, seu ponto forte é na investigação de sistema causais envolvendo variáveis latentes. A limitação do método é na estimativa destas variáveis latentes.

Análise de cluster (*cluster analysis*) trata do desenvolvimento de medidas de similaridade ou dissimilaridade (coeficientes), ou medidas de distância, para estabelecer associação de clusters. Primariamente, é empregado em Marketing para estudos de segmentação e estratégia. A principal vantagem do método é a classificação, como por exemplo de marcas, produtos, clientes, distribuidores, etc. Dentre as limitações, tem-se que diferentes métodos de clusterização podem gerar diferentes clusters.

Escalonamento multidimensional (*multidimensional scaling*) é baseado no cálculo de proximidade (ou, alternativamente, de dominância) entre atributos/variáveis e respondentes. Pesquisa de mercado, análise de *market share*, segmentação de mercado, posicionamento de marca são as principais aplicações na área de Marketing. Sua vantagem está no fato de apresentar toda a estrutura de variáveis, facilitando a visualização e interpretação dos relacionamentos/similaridades entre os dados. Sua barreira consiste em que diferentes pacotes de softwares são necessários para diferentes tipos de dados de entrada.

Análise conjunta (*cojoint analysis*) realiza a mensuração dos julgamentos psicológicos pela mensuração do efeito de junção de duas ou mais variáveis independentes sob uma variável

dependente. É empregado em pesquisa de consumidor e avaliação de anúncios no Marketing. O ponto forte está no cálculo de preferências, além de ser adequado para design de produto e mensuração de atitude. Sua limitação está na premissa de que mensura primeiro a utilidade ao invés do comportamento.

Análise de correspondência (*correspondence analysis*) é uma técnica gráfica para representar tabelas multidimensionais. Foi empregada em estudos de funções de vendas em agências de bancos, segmentos de mercado, rastrear imagens de marca. É destacada por poder ser rápida e fácil para interpretar, usada para análise de dados binários, discretos e/ou contínuos, além de facilitar tanto a comparação da distância quadrada dentre e entre os conjuntos. Suas aplicações são limitadas em Marketing por causa da falta de software adequado.

A análise de regressão (*regression analysis*) desenvolve uma função expressando a associação (ou relacionamento) entre variáveis dependentes e independentes. Aplicado em estudo de Marketing para segmentação, análise do comportamento do consumidor, previsão de vendas. As vantagens do método são: i) Permite previsões sobre uma variável dependente; ii) Fornece medidas de associação entre variáveis independentes e algumas importantes variáveis dependentes de Marketing. Para as limitações, tem-se o requerimento do ajuste da linha de regressão e determinação dos parâmetros. Isto pode ser complexo e gerar alguns erros.

Detecção de interação automática (*automatic interaction detection*) é uma rotina sequencial baseada em computador que tenta classificar objetos em grupos. Empregado em Marketing para análise de segmentos de mercado, avaliação dos efeitos de anúncio nas vendas do varejo, previsão da lealdade a marca, previsão de vendas, etc. É um método adequado para identificação de diferentes variáveis afetando os segmentos de mercado; determinação da importância de cada variável independente e a forma em que afeta a variável dependente. É menos poderoso do que regressão. O tamanho mínimo do grupo não pode ser menor do que 30, e o tamanho da amostra original deve ser grande.

A análise discriminante (*discriminant analysis*) realiza a maximização da relação da variância entre médias de grupo, variância não dentre grupo. Os estudos de Marketing associados são em previsão da lealdade a marca, clientes inovadores, aprovação/desaprovação de um serviço (ou produto), etc. Destaca-se por permitir previsões de variáveis dependentes. As limitações estão na identificação da significância estatística da função discriminante e que a análise múltiplos discriminantes requer programa de computador.

Simulação (*Simulation*) é a condução de experimentos usando um modelo para simular condições de trabalho de sistemas reais. Foi empregado no Marketing para: (a) Planejamento de Marketing; (b) Monitoramento e controle, operações de Marketing; (c) Distribuição, comportamento do consumidor, varejo, recrutamento, anúncio. Dentre os pontos fortes, tem-se: (a) Método muito flexível e simples entendido facilmente por gestores; (b) Economiza tempo e recursos; (c) Simulação possui diversas aplicações no campo de Marketing. Em termos de limitações, pode-se citar: (a) Cálculo aritmético tedioso; (b) Custo de tempo computacional relevante.

Conjuntos difusos (*fuzzy sets*) consiste essencialmente em um processo de modelagem factual que tenta o ajuste fino a expressão de conhecimento. É feito utilizando uma escala linguística que descreve as características sob cada uma das principais dimensões do modelo para formar conjuntos *fuzzy*; uma agregação hierárquica baseada em operadores agregadores *fuzzy*; e um hipercubo conceitual para determinar o *rank* e tamanho do *rank* dos resultados. Inclui o conceito de função de adesão (entre 0 e 1). Foi empregado em Marketing para modelagem do comportamento do consumidor, planejamento de Marketing, teste de novo produto, teste de preço percebido, pesquisa de efeitos de comunicação de Marketing. Seu ponto forte está na flexibilidade a qual acomoda um grau de incerteza ou *fuzziness*, no diagnóstico. Esta *fuzziness* é de fato enaltecida como realista em expressão dos julgamentos humanos. Apresenta dificuldade na escala de mensuração e estimativa dos descritivos bipolares, na escala linguística para características descritivas, e na descrição de valores para parâmetros do modelo.

Inteligência Artificial (*Artificial Intelligence*) é um programa de computador que expressa o processo de raciocínio através da modelagem de relacionamentos entre variáveis. Suas aplicações estão em pesquisa de Marketing, teste de Marketing, precificação, seleção de site, Marketing de turismo e Marketing internacional. É um método flexível, capaz de explicar raciocínio das interações. Tem dificuldades na construção do modelo de sistema especialista.

Redes neurais (*neural networks*) é um método de uso de dados estruturados de entrada e saída para desenvolver padrões que replicam a tomada de decisão humana. Emprega um procedimento estatístico de ajustes iterativos de pesos. Existem aplicações em Marketing para comportamento do consumidor, modelagem de preço, planejamento de mídia e segmentação de mercado. As vantagens estão na capacidade de reaprendizado, além de poder trazer junto análises psicométricas e econométricas. É um método de baixa exatidão e é mais difícil de interpretar do que os sistemas especialistas acima.

Teoria das filas (*queuing*) é a análise de distribuição de probabilidade de dados (empiricamente coletados em como os principais fatores/variáveis afetarão a situação problema em análise). É uma análise de sistemas de fila para determinar o nível/performance de serviço. Suas aplicações em Marketing são: (a) Otimização: equipe de vendas, número de *checkouts*, número de atendentes, etc; (b) Minimizar os custos de estoque; adequado e usado amplamente por cadeias de lojas, supermercados, lojas de departamento, estações de petróleo, escritório de passagens aéreas, portos, aeroportos, etc. Dentre suas vantagens, pode-se citar: (a) Prevê como diferentes sistemas de Marketing operam; (b) Permite a expressão explícita relacionada ao design do sistema para o tamanho e frequência das filas, tempo de espera, etc. Para as limitação, tem-se: (a) Deve ser operada por um período de tempo suficiente para atingir a solução de estado; (b) Relutância do gestor para confiar no método.

Processo estocástico (*stochastic process*) consiste em experimento aleatório em que ocorre ao longo do tempo e cujo resultado é determinado por chance. É uma análise de sistemas com variáveis/componentes incertos. Aplicado em Marketing em: (a) Construção de escolha, modelos para verificar lealdade do cliente; (b) Prevê decisões de compra e probabilidade de compras futuras. Seus pontos fortes são na capacidade de prever o fluxo de clientes e a probabilidade de compra futura. É um método adequado somente para previsões de curto período.

A teoria dos jogos (*game theory*) solução de jogo de soma constante, uso de um critério máximo para determinar, por exemplo, alocação de verba/recursos. É uma análise teórica de competição/coalisão entre empresas. No Marketing, é visto em tomada de decisão para empresas de varejo, principalmente em: precificação, determinação de estoque de produto e anúncio, alocação de verba, também para decisão melhor em processos de negociação. É um método que se destaca por: (a) Auxiliar gestão de tomada de decisão; (b) Sugerir uma útil abordagem analítica para problemas de competição, como: precificação, anúncio, despesas e decisões de produto. Sua limitação está no fato de não ter muito poder de previsão comparado a outras técnicas quantitativas.

Programação linear (*linnear programming*) é um método baseado em objetivo e funções de restrições lineares. Suas aplicações em Marketing estão em: (a) Anúncios, espaço, alocação de mix de mídia otimizado; (b) Problemas de distribuição, localização de sítio; (c) Alocação de verba, decisão de novos produtos; (d) Combinação de mix de produtos; (e) Decisões de mix de Marketing. É um método vantajoso para: (a) Maximizar rentabilidade de alocações, sujeito a restrições; (b) Minimizar custos; (c) Auxilia gestão de tomada de decisão. Dentre as

dificuldades, pode-se citar: (a) Dificuldade em obter e formular as várias funções; (b) Restrições devem ser alteradas o mais rápido possível para mudanças de fatores externos e/ou internos.

Modelo de transporte (*transportation model*) é baseado em uma matriz de transporte/alocação visando o mínimo custo, rota, quantidade fornecida, etc. É utilizada no Marketing para alocação de recursos, fornecimento, através da redução dos custos de transporte. Adequado particularmente para lojas de departamento, empresas de empréstimo de caminhões, companhias de transporte. É bastante adequado para tomada de decisão gerencial; contudo, é um método pouco preciso no longo prazo como um resultado da mudança nos custos.

Programação não linear (*non-linear programming*) se baseia em funções objetivo não lineares e relações de restrição não lineares. As aplicações em Marketing para este métodos são encontrar o máximo retorno na pesquisa de um novo produto, sujeito a restrição de verba. As vantagens são: (a) Quando as relações são não lineares; (b) Quando a função objetivo é não linear enquanto as restrições são não lineares. A principal dificuldade está em estabelecer relações não lineares.

Os modelos causais (*causal models*) são relativamente novos em Marketing, e apresenta dois métodos principais: LISREL e análise de caminho.

LISREL (*linear structural relations*) é uma modelagem de equações estrutural, que permite a decomposição das relações entre variáveis e testa modelas causais que envolvem tanto variáveis observáveis quanto inobserváveis. Utiliza-se em estudos de comportamento do consumidor, venda pessoal, estratégia de Marketing, Marketing internacional. Fornece uma abordagem integral para análise de dados e construção de teoria. O método facilmente lida com erros na medição. Habilidade em juntar análises psicométricas e econométricas. Contudo, requer uma teoria prévia para análise estrutural.

Análise de caminho (*path analysis*) fornece meios para estudar os efeitos diretos e indiretos das variáveis, através da informação quantitativa baseada nos dados qualitativos de relações causais. Suas aplicações estão na área de Marketing de turismo. A principal vantagem é o resultado gráfico do padrão de relações causais. A desvantagem do método é que ele assume relações entre variáveis como lineares.

Programação dinâmica (*dynamic programming*) é um procedimento de otimização recursiva; trata-se de uma otimização passo-a-passo. É empregado em solução de problemas de seleção de mídia; distribuição (minimização dos custos de transporte; distribuição do time de

vendas para várias territórios de vendas). Suas vantagens são: (a) Maximizar objetivo do período planejado; (b) Introduz novos fatores, por exemplo, “tempo de esquecimento”, “acúmulo ou intersecção”; (c) Amplo potencial de aplicação na indústria. Entretanto, o procedimento de programação é relativamente complexo; dificuldades computacionais.

Programação heurística (*heuristic programming*) é um procedimento guiado de pesquisa ordenada através do uso de regra geral. Baseado em “abordagem marginal” ou tentativa e erro. Suas aplicações na área de Marketing são: seleção de mídia e agendamento; localização de armazém; alocação de time de vendas; decisão do número de itens da linha de produtos; adequado para fazer decisões de promoção de produtos. O método se destaca em: (a) Método bom, flexível, simples e barato; (b) Combina a análise com o estilo de tomada de decisão e o raciocínio usado por gestores. Contudo, a principal falha do método é que ele não garante a solução ótima.

PERT e CPM apresentam uma ampla gama de atividades críticas que dever ser seguidas e coordenadas. PERT reconhece incertezas no tempo necessário para completar atividades enquanto que CPM lida apenas com o fator tempo. CPM lida também com *trade-offs* de tempo-custo. São métodos empregados para planejamento, agendamento e controle de projetos complexos de Marketing, por exemplo, construção de novas lojas, desenvolvimento de novos produtos, comercialização de produtos, relacionamentos de anúncio-vendas, planejamento de distribuição. Suas vantagens são: (a) Sequências e tempo de atividades são consideradas, responsabilidades alocadas e coordenação de projetos grandes/complexos de Marketing; (b) Tempo de projeto pode ser previsto e tempo de finalização pode ser encurtado. As dificuldades são: (a) Dificuldade na estimativa de custos e tempo de forma precisa, particularmente para novos projetos; (b) Válido apenas quando funções e atividades podem de fato serem separadas.

3.2.1. Escolha do melhor método para segmentação

No Capítulo 3.2, foram discutidos diversos métodos e técnicas que podem ser empregados para análises de Marketing. No caso da segmentação, conforme a taxonomia de Moutinho & Meidan (2004), podem ser destacadas as seguintes opções: (i) *Latent analysis*; (ii) *Cluster analysis*; (iii) *Multidimensional scaling*; (iv) *Correspondence analysis*; (vi) *Regression analysis*; (vii) *Automatic interaction detection*; (viii) *Neural networks*.

Conforme apresentados por Moutinho & Meidan (2004), todos os métodos podem ser aplicados para a segmentação de mercado. Comparada aos demais métodos, em que a vantagem é dada pela análise do impacto das variáveis, a análise de clusters se destaca pelo caráter de agrupamento, permitindo assim a classificação dos dados. A análise de fatores, cujo benefício está na redução de dados, poderia ser uma alternativa a ser aplicada; no entanto, a sua limitação em lidar apenas com dados contínuos enfraquece a análise desejada.

Tendo em vista a sinergia entre a classificação e o objetivo do trabalho, e a capacidade de tratamento com variáveis não só contínuas, optou-se pela adoção da análise de clusters.

3.3.Análise de clusters

Segundo Tan et al (2005), a análise de clusters consiste no agrupamento de objetos de dados baseado nas suas informações e suas relações. O objetivo é que os objetos de um grupo sejam similares (ou relacionados) entre si e diferentes (ou não relacionados) dos outros grupos. Quanto maior a similaridade (ou homogeneidade) no grupo e maior a diferença entre os grupos, melhor ou mais distinta é a clusterização. Kaufman & Rousseeuw (1990) possuem um definição mais simples: a análise de clusters é a arte de encontrar grupos em dados.

Dentre as áreas de aplicação da Análise de Clusters, podem ser citadas (Tan et al, 2005): psicologia e outras ciências sociais, biologia, estatística, reconhecimento de padrões, recuperação de informações, *machine learning* e mineração de dados. Além disso, os autores citam que na área de negócios: dado que as empresas atualmente coletam uma grande quantidade de informações de seus clientes (tanto atuais quanto potenciais), a análise de clusters pode ser utilizada para segmentar clientes em pequenos grupos, facilitando análises adicionais e as atividades de Marketing.

Em relação ao Marketing, Punj & Stewart (1983) destacam quatro aplicações para a análise de clusters: i) segmentação de mercado; ii) entendimento do comportamento de compra através da identificação de grupos homogêneos de compradores; iii) desenvolvimento oportunidades de potenciais novos produtos; iv) seleção do mercado para teste; v) redução de dados por meio de agrupamentos, com o objetivo de facilitar a gestão.

3.3.1. Algoritmos de clusterização

Punj & Stewart (1983) e Donilcar (2003) mostram que existem diversos algoritmos para a clusterização. Em conjunto com Kaufman & Rousseeuw (1990), os algoritmos podem ser classificados em duas grandes categorias: métodos de particionamento iterativo (métodos não hierárquicos) e métodos hierárquicos. A primeira apresenta diferentes vertentes, sendo que o algoritmo mais utilizado é o *K-means*, o qual será melhor detalhado para efeito de ilustração. A segunda categoria é um conjunto de técnicas com uma lógica bem semelhante, a qual é dividida em aglomerativa e divisiva. Para simplificar, será exposto o método hierárquico aglomerativo e algumas de suas variações.

Em termos de uso para a segmentação de mercado, o levantamento de Donilcar indica que ambos os métodos são empregados com frequência semelhante, com um pequeno favorecimento ao método de particionamento em relação ao hierárquico (46% e 44%, respectivamente).

3.3.1.1. Algoritmo *K-means*

O algoritmo *K-means* é um exemplo de método não hierárquico de clusterização. Conforme o estudo de Donilcar (2003), é o mais empregado nas pesquisas de análises de clusters quando se trata de segmentação de mercado (Tabela 1).

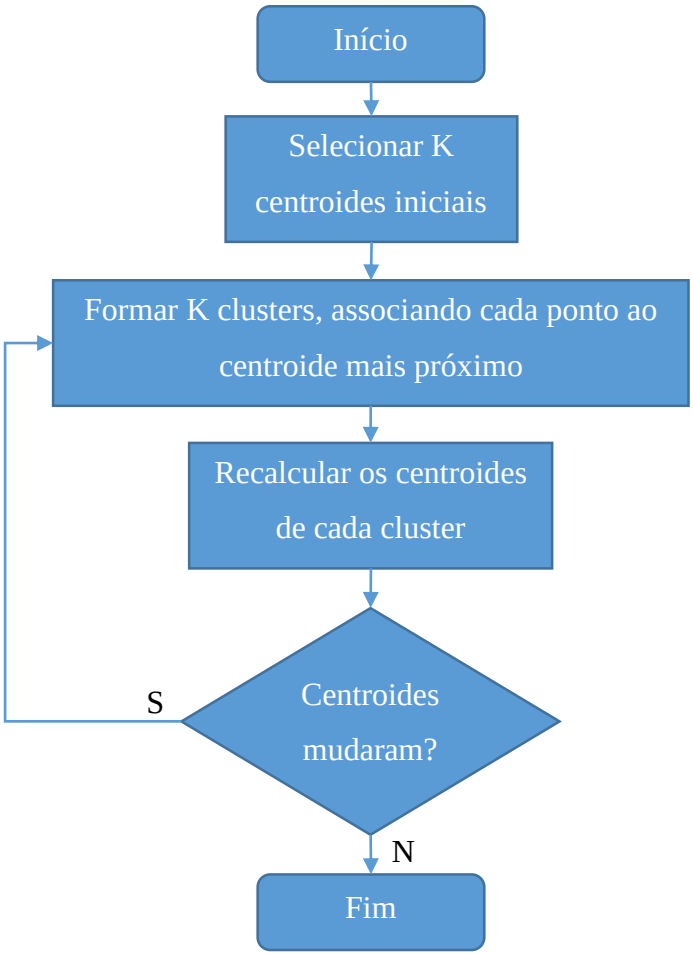
Tabela 1: Frequência dos métodos de particionamento.

Método	**Frequência**	**Percentual**
K-means	68	76
Não declarado	17	19
RELOC	1	1
Cooper-Lewis	1	1
Redes neurais	3	3

Fonte: Adaptado de Donilcar (2003).

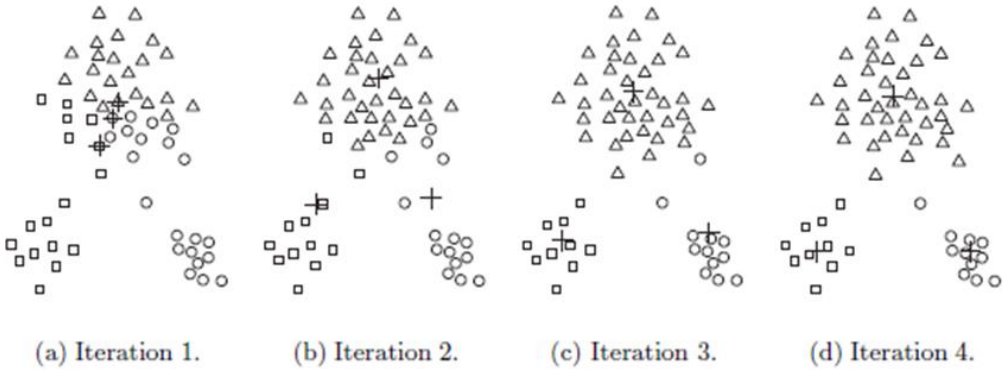
Dado um parâmetro K, o algoritmo procura agrupar os pontos de modo a serem obtidos K clusters, que serão representados pelos centroides. A Figura 13 apresenta esquematicamente o algoritmo e a Figura 14 exemplifica graficamente as iterações até a obtenção dos clusters finais.

Figura 13: Algoritmo *K-means*.



Fonte: Adaptado de Tan et al (2005).

Figura 14: Exemplo de iterações do algoritmo *K-means*.



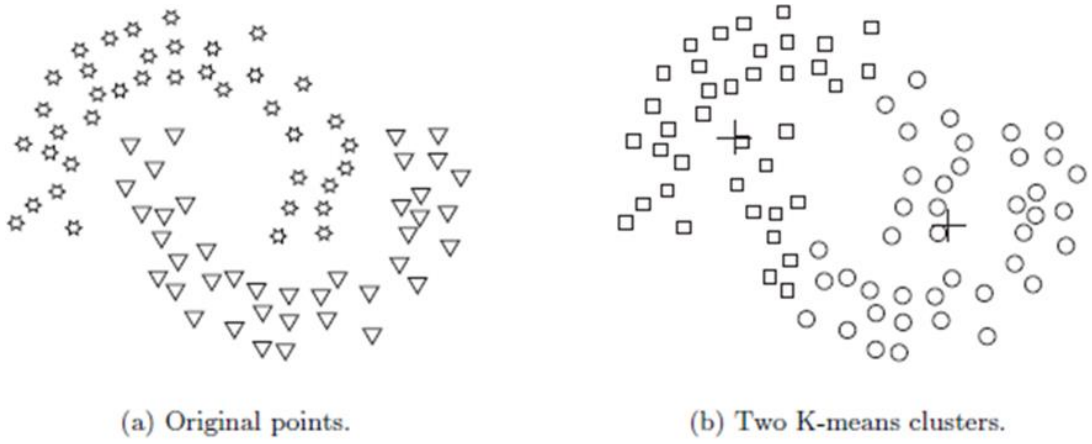
Fonte: Tan et al (2005).

Primeiramente, são escolhidos os *K* centroides iniciais, lembrando que *K* é um parâmetro dado pelo usuário, chamado de número de clusters desejado. Cada ponto é então associado ao centroide mais próximo (o conjunto de pontos associados ao centroide forma um cluster). O centroide de cada cluster é então atualizado baseado nos pontos associados. Repete-se os passos anteriores até a convergência dos centroides.

Conforme destaca Tan et al (2005) e Maimon & Rokach (2005), o algoritmo *K-means* é simples e pode ser aplicado para uma ampla variedade de tipos de dados. Apesar do caráter iterativo, é um método bastante eficiente em termos computacionais.

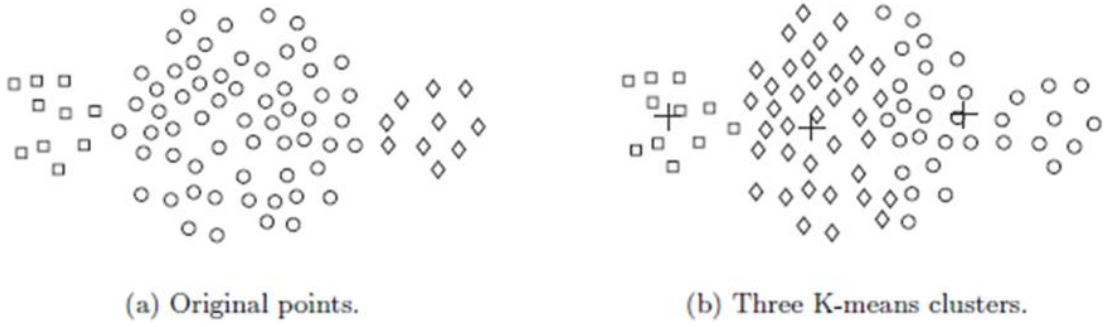
Dentre as fraquezas do método, os autores citam a restrição de aplicação para dados que tenham a noção de centroide, a sensibilidade a *outliers*, o risco de obtenção de clusters vazios (decorrente da má escolha dos centroides iniciais), e a dificuldade na formação de clusters naturalmente não globulares (Figura 15) ou de diferentes tamanhos (Figura 16) ou densidades (Figura 17).

Figura 15: *K-means* com clusters não globulares.



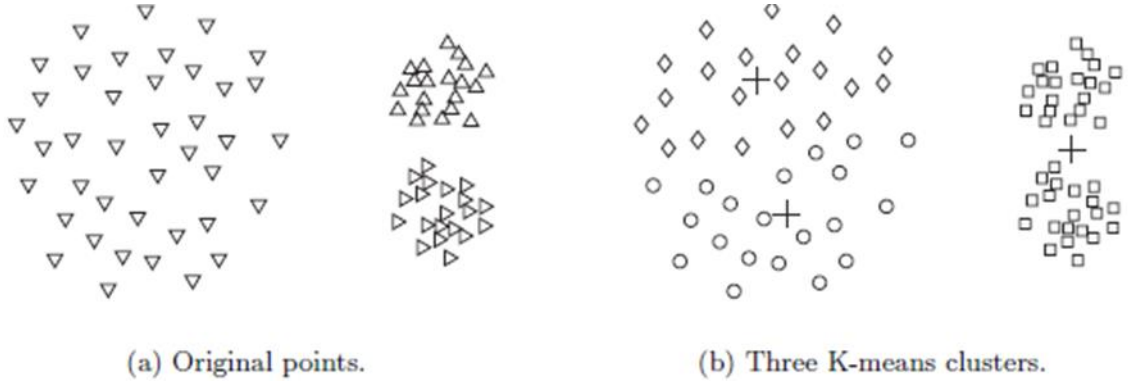
Fonte: Tan et al (2005).

Figura 16: *K-means* com clusters de tamanhos diferentes.



Fonte: Tan et al (2005).

Figura 17: *K-means* com clusters de densidades diferentes.



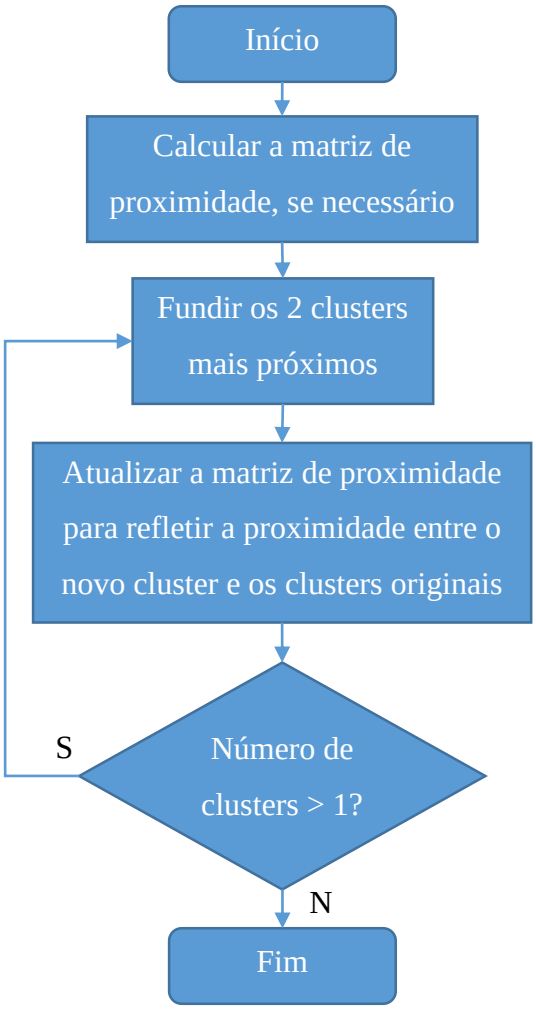
Fonte: Tan et al (2005).

3.3.1.2.Métodos hierárquicos aglomerativos

Os métodos hierárquicos podem ser classificados em aglomerativos e divisivos. Os primeiros partem da situação em que todos os pontos são clusters individuais e, após cada iteração, os clusters mais próximos são fundidos (no limite, tem-se apenas um único cluster que contém todos os pontos). Além da definição da avaliação de semelhança, o método hierárquico aglomerativo necessita também de um critério para comparar 2 clusters entre si. Os métodos hierárquicos divisivos iniciam na situação contrária: um único cluster, que contém todos os pontos, é dividido em cada etapa (o último passo resulta em clusters individuais). Para este caso, deve-se estabelecer o critério de escolha do cluster a ser dividido e como será realizada a divisão.

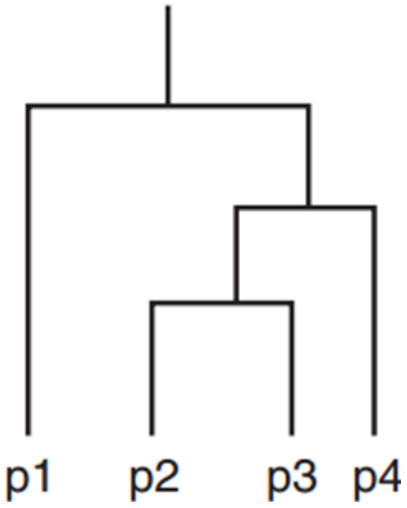
Neste capítulo, serão abordados os métodos hierárquicos aglomerativos, cujo algoritmo está representado na Figura 18. Graficamente, a ferramenta mais empregada para representar o resultado obtido é o dendograma (Figura 19).

Figura 18: Algoritmo do método hierárquico aglomerativo.



Fonte: Adaptado de Tan et al (2005).

Figura 19: Exemplo de dendograma.

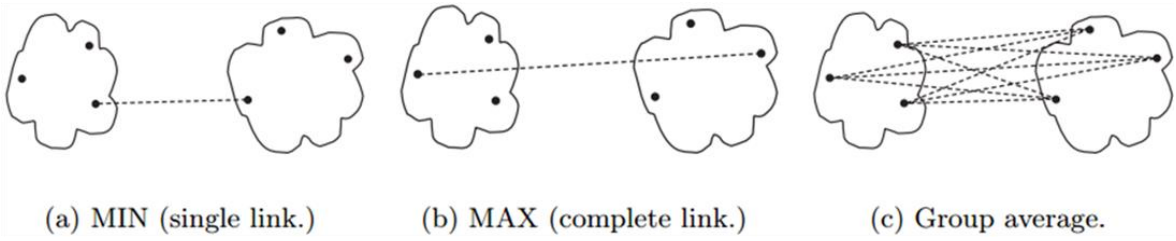


Fonte: Tan et al (2005).

Para a avaliação entre dois clusters existem cinco abordagens principais (sendo que as três primeiras estão ilustradas graficamente na Figura 20):

- *Single linkage*: a proximidade entre dois clusters é dada pela menor distância entre dois pontos de clusters diferentes.
- *Complete linkage*: a proximidade entre dois clusters é dada pela maior distância entre dois pontos de clusters diferentes.
- *Group average*: a proximidade entre dois clusters é dada pela média das distâncias dois a dois de todos os pontos de clusters diferentes.
- *Centroid*: a proximidade entre dois cluster é dada pela distância entre os centroides dos clusters.
- Método de Ward: a proximidade entre dois cluster é dada em termos da soma dos erros quadrados (SSE). Neste caso, prefere-se a fusão com menor SSE.

Figura 20: Definições de proximidade entre os clusters.



Fonte: Tan et al (2005).

A Tabela 2 contém o levantamento de Donilcar (2003) sobre a utilização das abordagens citadas no estudo de segmentação de mercado.

Tabela 2: Utilização dos métodos para clusterização hierárquica aglometativa.

Método	**Frequência**	**Percentual**
Single linkage	5	6
Complete linkage	8	10
Average linkage	6	7
Nearest centroid sorting	5	6
Ward	47	57
Não declarado	8	10
Múltiplos	4	5

Fonte: Adaptado de Donilcar (2003).

Segundo Tan et al (2005), o ponto positivo deste método é justamente a criação de uma hierarquia. Alguns pontos chaves sobre o método, os quais são também reforçados por Maimon & Rokach (2005), são o alto custo computacional (tanto em termos de processamento quanto de armazenamento); o tratamento com clusters de tamanhos diferentes; o processo de fusão é irreversível; a ausência do problema de escolha dos pontos iniciais.

3.3.2. Medidas de distância e de semelhança

Para efetuar a associação dos pontos e a obtenção dos clusters, é essencial a definição de uma medida de distância ou de semelhança. No entanto, a escolha varia conforme a natureza dos dados, sendo que cada caso apresenta uma medida mais apropriada. A seguir, serão citadas as principais medidas empregadas, conforme Tan et al (2005) e Maimon & Rokach (2005).

A medida de distância mais empregada é a Distância Euclidiana (L_2). Sejam os pontos $P = (p_1, p_2, \dots, p_n)$ e $Q = (q_1, q_2, \dots, q_n)$, a distância é dada por:

$$d(P,Q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

No caso de variáveis binárias, pode-se utilizar o coeficiente de correspondência simples, dado por:

$$d(x_i, x_j) = \frac{r + s}{q + r + s + t}$$

Onde:

- r é o número de atributos com valor 1 para x_i e x_j;
- t é o número de atributos com o valor 0 para x_i e x_j;
- r e s são o número de atributos que não são iguais para x_i e x_j.

Para variáveis nominais, é possível transformar cada estado da variável em uma variável binária (e utilizar a mesma métrica do caso anterior), ou então realizar a correspondência simples:

$$d(x_i, x_j) = \frac{p - m}{p}$$

Onde:

- p é o número de atributos;
- m é o número de correspondências.

Tratando-se de similaridade, pode-se citar a Medida de Cosseno, dada por:

$$s(x_i, x_j) = \frac{x_i^T \cdot x_j}{\|x_i\| \cdot \|x_j\|}$$

Outra medida de similaridade que pode ser utilizada é a Medida de Jaccard:

$$s(x_i, x_j) = \frac{x_i^T \cdot x_j}{\|x_i\|^2 + \|x_j\|^2 - x_i^T \cdot x_j}$$

Entretanto, é comum uma base de dados apresentar diversos tipos de variáveis. Segundo Kaufman & Rousseeuw (1990), a medida de Gowers trata de forma adequada a existência de variáveis do tipo contínua, nominal e binária. Supondo que o conjunto de dados apresenta p variáveis de tipos variados, então a dissimilaridade $d(i,j)$ entre os objetos i e j é definida como:

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

Onde:

- $\delta_{ij}^{(f)}$ é igual 1 se ambos x_{if} e x_{jf} para a f-pesima variável existem, e é igual a 0 caso contrário;
- $\delta_{ij}^{(f)}$ também é 0 quando a variável f é um atributo binário assimétrico e objetos i e j constituem uma correspondência 0-0;
- $d_{ij}^{(f)}$ é a contribuição da f-ésima variável para a dissimilaridade entre i e j.

Caso a variável f seja binária ou nominal, então $d_{ij}^{(f)}$ é definido como:

$$\begin{cases} d_{ij}^{(f)} = 1 & se\ x_{if} \neq x_{jf} \\ d_{ij}^{(f)} = 0 & se\ x_{if} = x_{jf} \end{cases}$$

Se a variável f é contínua, então $d_{ij}^{(f)}$ é dado por:

$$d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{R_f}$$

Onde R_f é o intervalo da variável f, definido como:

$$R_f = \max_h x_{hf} - \min_h x_{hf}$$

Onde h percorre por todos objetos existentes para a variável f.

Em relação ao cálculo da proximidade dos clusters no método hierárquico aglomerativo, a fórmula de Lance-Williams engloba os casos mencionados (Tabela 3). Ela é dada por:

$$p(R, Q) = \alpha_A p(A, Q) + \alpha_B p(B, Q) + \beta p(A, B) + \gamma |p(A, Q) - p(B, Q)|$$

Onde:

- $p(R, Q)$ é a função de proximidade entre os clusters R e Q;
- R é o cluster resultante da fusão entre os clusters A e B;
- α_A, α_B, β, γ são os coeficientes da fórmula (vide Tabela 3);
- m_A, m_B, m_Q são o número de pontos nos clusters A, B e Q, respectivamente.

Tabela 3: Coeficientes de Lance-Williams.

Método de clusterização	$\boldsymbol{\alpha_A}$	$\boldsymbol{\alpha_B}$	$\boldsymbol{\beta}$	$\boldsymbol{\gamma}$
Single Link	½	½	0	$-½$
Complete Link	½	½	0	½
Group Average	$\frac{m_A}{m_A+m_B}$	$\frac{m_B}{m_A+m_B}$	0	0
Centroid	$\frac{m_A}{m_A+m_B}$	$\frac{m_B}{m_A+m_B}$	$\frac{-m_A m_B}{(m_A+m_B)^2}$	0
Ward	$\frac{m_A+m_Q}{m_A+m_B+m_Q}$	$\frac{m_B+m_Q}{m_A+m_B+m_Q}$	$\frac{-m_Q}{m_A+m_B+m_Q}$	0

Fonte: Tan et al (2005).

3.3.3. Validação da clusterização

Tão importante quanto a seleção do método de clusterização, a validação dos resultados obtidos define se o modelo escolhido representa bem os dados, além de permitir a comparação entre os modelos.

As medidas euclidianas, como dito anteriormente, são as mais comuns nos estudos de análise de cluster e o principal indicador associada a elas é a soma dos quadrados dos erros (ESS), que é dada por:

$$ESS = \sum_{\nu=1}^{k} ESS(C_\nu) = \sum_{i \in \nu} \sum_{f=1}^{p} \left(x_{if} - \bar{x}_f(\nu) \right)^2$$

Onde:

- k se refere ao número de clusters
- C_ν é o cluster ν
- p é número de variáveis do modelo
- x é o objeto a ser clusterizado
- $\bar{x}(\nu)$ é o centroide do cluster ν

Kaufmann & Rousseeuw (1990) propõe o conceito de silhuetas (do original *silhouettes*), que identificam a adesão dos objetos ao cluster.

Silhuetas são construídas da seguinte forma: para cada objeto i o valor $s(i)$ é definido e então estes números são combinadas em um gráfico. Para definir $s(i)$, tem-se A o cluster em que o objeto i foi associado e então calcula-se:

$$a(i) = dissimilaridade\ média\ de\ i\ para\ todos\ os\ demais\ objetos\ de\ A$$

Isto só pode ser feito quando A contém outros objetos além de i, logo assume-se que A não é um cluster unitário.

Considere qualquer cluster C diferente de A e define-se:

$$d(i, C) = dissimilaridade\ média\ de\ i\ para\ todos\ os\ objetos\ de\ C$$

Após o cálculo de $d(i, C)$ para todos os clusters $C \neq A$, seleciona-se o menor deles:

$$b(i) = \min_{C \neq A} d(i, C)$$

O cluster B que contém o valor mínimo associado (ou seja, $d(i, B) = b(i)$), é chamado de vizinho do objeto (i). É a segunda melhor escolha para o objeto i: se o cluster A for descartado, o cluster B seria o mais próximo de i. Note que a construção de $b(i)$ depende da disponibilidade dos clusters diferentes de A, o que explica a não definição das silhuetas para $k = 1$.

O número $s(i)$ é obtido pela combinação de $a(i)$ e $b(i)$:

$$s(i) = \begin{cases} 1 - \dfrac{a(i)}{b(i)} & se\ a(i) < b(i) \\ 0 & se\ a(i) = b(i) \\ \dfrac{b(i)}{a(i)} - 1 & se\ a(i) > b(i) \end{cases}$$

É possível escrevê-la em uma fórmula:

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

Quando o cluster A conter apenas um único objeto, não é claro como $a(i)$ deve ser definido, então apenas assume-se $s(i) = 0$. Esta escolha é arbitrária, mas o valor zero parece ser o mais neutro. Através das definições anteriores, pode-se verificar que para cada objeto i:

$$-1 \leq s(i) \leq 1$$

Observando a definição e o intervalo de valores, pode-se que:

- Quando $s(i)$ é próximo de 1, significa que i está bem classificado, ou seja, a segundo melhor cluster B não tão próximo quanto a escolha atual A.
- Se $s(i)$ é próximo de zero, então $a(i)$ e $b(i)$ são aproximadamente iguais; logo, não é claro se i deveria ser associado a A ou B.
- O pior caso ocorre quando $s(i)$ é próximo de -1. $a(i)$ é muito maior do que $b(i)$, então na média, i é mais próximo de B do que de A. Dessa maneira, seria mais natural associar o objeto i ao cluster B.

A Tabela 4 contém a interpretação subjetiva sobre a clusterização obtida em função do coeficiente de silhueta (SC) da análise (Kaufmann & Rousseeuw, 1990). Este coeficiente se refere a média da largura da silhueta para todo o conjunto de dados; ou seja, a média de $s(i)$ para $i = 1, 2, \ldots, n$.

Tabela 4: Interpretação subjetiva do SC.

SC	**Interpretação proposta**
0,70 ⊢ 1,00	Uma estrutura robusta foi encontrada
0,50 ⊢ 0,70	Uma estrutura razoável foi encontrada
0,25 ⊢ 0,50	A estrutura é fraca e pode ser artificial; tentar outros métodos neste conjunto de dados
≤ 0,25	Nenhuma estrutura substancial foi encontrada

Fonte: Adaptado de Kaufmann & Rousseeuw (1990).

3.3.4. Aplicações de análise de clusters para segmentação

Para ilustrar o potencial da análise de cluster para a segmentação, foram estudados 3 artigos. Abaixo, tem-se uma breve descrição sobre cada um deles.

3.3.4.1.Estudo da Bivolino

O primeiro deles foi um estudo realizado por Brito et al (2015) com o auxílio da empresa Bivolino, uma fabricante de camisas personalizadas que comercializa através de seu site. Foram dois os objetivos estabelecidos pelos autores.

O primeiro deles, de caráter metodológico, foi a aplicação de duas técnicas de data mining para resolver o problema de segmentação e o teste da extensão da complementariedade entre os eles para explicar diferentes aspectos do mercado. A primeira etapa consistiu na obtenção dos segmentos de mercado e, para tanto, utilizou-se a técnica *K-medoids*. A segunda etapa, para realizar a caracterização dos subgrupos de observações com distribuição raras, utilizou o método de descoberta de subgrupo CN2-SD, sendo que, de acordo com os autores, nunca foi empregado com tal propósito.

O segundo objetivo deste estudo foram essencialmente gerenciais, sobre os benefícios da segmentação: i) externamente, auxiliar a empresa a redefinir sua estratégia de comunicação, particularmente em relação às promoções de venda; ii) internamente, através da correspondência entre os produtos e as preferências dos clientes, ajudar a redefinir o design do produto, ajustando o processo de manufatura e acelerando a entrega. Para atingir tais benefícios, é crucial obter a caracterização dos segmentos de mercado baseados nos atributos dos produtos preferidos com o perfil do cliente.

Os dados utilizados no estudo de Brito et al (2015) foi a base de dados da empresa Bivolino, a qual contava com total de 10775 pedidos de clientes. As variáveis disponíveis estavam agrupadas em 5 grupos: i) características do produto (tipo de tecido, cor do tecido, tipo de colarinho, estrutura do tecido); ii) demográfico e biométrico (gênero, faixa etária, tamanho do colarinho, índice de massa corporal); iii) geográfico (país/nacionalidade); iv) psicográfico (estilo de vida, propósito); v) comportamental (sensibilidade de preço).

O *K-medoids*, empregado na primeira etapa, foi escolhido como alternativa ao *K-means* pois o segundo só pode ser aplicado para dados numéricos. O *K-medoids* é baseado em dissimilaridades entre os pares de objetos, permitindo a aplicação para dados mistos, como é caso deste estudo. Além disso, o *K-medoids* usa objetos representativos como pontos de referência, enquanto que os obtidos pelo *K-means* podem não ser observáveis. A implementação se deu pelo uso do software Rapid Miner.

Esta primeira etapa foi realizada em dois passos, sendo o primeiro apenas com as características do produto e o segundo aplicando para todas as variáveis, porém, a exclusão de alguns dados, considerados como *outliers*. Foram obtidos 6 clusters, através de experimentos preliminares, representados por seus *medoids*. Não foi relatado com clareza a validação da clusterização e a determinação do número de clusters.

A técnica de descoberta de subgrupo tem como objetivo a investigação de subgrupos da população que são estatiscamente mais interessantes e incomuns, ou seja, com distribuição estatística que mostra uma característica única em relação a distribuição global da propriedade sob investigação. A ferramenta empregada foi também o software Rapid Miner. Como este método está fora do escopo deste trabalho, optou-se pela omissão de seu detalhamento. Contudo, é importante ressaltar que a justificativa para o uso deste método não foi esclarecida neste estudo.

Para a segunda etapa, foram considerados 7066 pedidos, pois foram excluídos aqueles relacionados ao gênero feminino e também de outros países com parcela minoritária. Como resultado, foram obtidos 10 subgrupos a serem trabalhados, sendo 4 de interesse para o Marketing e 6 para o Design.

3.3.4.2.Estudo da biblioteca da faculdade privada de Taiwan

Hsu et al (2012) propuseram uma metodologia de segmentação para identificar similaridades entre clientes, baseado no conceito de hierarquia de itens. Foram analisados dados transacionais da biblioteca de uma faculdade privada de Taiwan. O período de análise foi de 3 meses entre janeiro e março de 2009, e as variáveis foram a sequência das transações por cliente, e os itens da mesma sequência.

Primeiro, o estudo relatou a medida de dissimilaridade entre dois dados transacionais, a qual pode ser verificada com mais detalhes no artigo.

Em seguida, foi detalhado o conceito de hierarquia e a clusterização hierárquica. A escolha deste método da análise de cluster foi baseada em três argumentos. O primeiro deles é a robustez do método, dada pela não necessidade de determinação de um valor inicial para o algoritmo, como ocorre nos métodos de particionamento iterativo (*K-means* e *K-medoids*). O segundo argumento é que a natureza dos dados é atendida pelo método, sendo duas características que se destacam: o comprimento dos pontos de referência não são iguais; os dados empregados na análise não são contínuos.

O terceiro ponto destacado pelo artigo é a validação da clusterização, determinada pela avaliação do número ideal de clusters. Foram três métricas empregadas para essa fase: SVM (*Silhouette Validation Method*), *C index* e *isolation index*. Além disso, os autores criaram mais uma métrica pra validação da clusterização, resultado da ponderação dos 3 indicadores (rotulado de *average index*). Os detalhes de cada métrica foram omitidos para não alongar este trabalho.

Após a implementação do modelo, foi determinado que a clusterização com 8 clusters seria a mais apropriada.

3.3.4.3.Estudo do Carrefour de Taiwan

O terceiro artigo estudado foi do trabalho de Liao et al (2011), que foi realizado em parceria com a empresa Carrefour em Taiwan.

O trabalho combinou as informações de clientes que compram online e recebem os produtos em casa. Os clientes foram divididos em clusters pela análise de clusters, e o catálogo de produtos foi especificado de acordo com as preferências de consumo do cluster. Dessa forma, desejava-se aumentar a atratividade do catálogo de produtos para os consumidores.

Para coleta de dados, foi aplicado um questionário de 9 seções com consumidores que compram produtos frescos e produtos não perecíveis. A seção 1 tratava de comportamento e motivação de compra dos consumidores. As seções 2-5 perguntava sobre os alimentos frescos, de forma a determinar suas preferências sobre esta categoria. A seção 6 questionava a compra de produtos não frescos, de forma análoga ao caso anterior. A seção 7 explorava a questão da

entrega em domicílio, com o objetivo de descobrir os tipos de produtos seriam os mais apropriados para este serviço, conforme os consumidores. A seção 8 investigava o comportamento de compra online, verificando se os consumidores tinham alguma experiência com esta plataforma. Finalmente, a seção 9 tratava de informações básicas sobre os consumidores. A pesquisa foi realizada entre julho e outubro de 2008, resultando em uma amostra válida de 352 respostas.

O método de análise de clusters empregado foi uma variação do *K-means*. Contudo, não foram detalhados os motivos para a escolha deste método. O desenvolvimento dos clusters se deu em 2 etapas, sendo a primeira utilizando as variáveis relacionadas às "informações básicas dos consumidores" e a segunda com o conjunto de "comportamento e motivação dos consumidores".

Foram 3 clusters obtidos neste estudo, sendo a frequência de consumo a principal característica que os diferencia. Além disso, foi possível levantar as categorias de produtos mais relevantes para cada cluster. A validação dos resultados não foi abordada no artigo.

3.3.5. Resumo do capítulo

Este capítulo apresentou a análise de clusters e seus principais conceitos. Conforme exemplificado por Tan et al (2005), é uma das ferramentas que pode ser utilizada para a realização da segmentação dos clientes.

As principais questões a serem desenvolvidas são a escolha do algoritmo mais apropriado e a definição da métrica de distância ou similaridade, as quais devem ser respondidas após a análise das informações disponíveis. Outra etapa igualmente importante é a validação da clusterização obtida.

Três estudos foram analisados de modo a verificar o potencial da ferramenta para a segmentação e também servir como base para o estabelecimento da metodologia do trabalho.

4. METODOLOGIA

4.1.Modelo de análise

4.1.1. Variáveis do modelo e coleta de dados

Diversos estudos demonstraram que a base de dados das empresas podem gerar bons resultados para análise de clusters.

Liao et al (2011) realizaram um trabalho com a base de dados da empresa Carrefour de Taiwan. As informações estavam organizadas em diversas tabelas de dados. Contudo, 3 clusters obtidos foram resumidos a basicamente 6 variáveis

- Gênero;
- Idade;
- Nível de escolaridade;
- Área de atuação do trabalho;
- Média mensal da renda familiar;
- Frequência de consumo.

Hsu et al (2012) focaram em dados transacionais de uma biblioteca em Taiwan, utilizando: número de sequência do registro; identificação do leitor; identificação do livro; nome do livro; número da categoria do livro; data do registro. Contudo, os autores optaram pelo pré processamento dos dados, resultando na aplicação apenas para 2 variáveis:

- Sequência da transação do cliente;
- Sequência dos itens da transação.

Brito et al (2015), com auxílio da empresa Bivolino, que atua na fabricação de camisas, valeram-se de 10775 dados de pedidos de clientes, que foram resumidos em 10 variáveis, agrupadas em 5 tipos:

- Caraterísticas de produto
    - Tipo de tecido
    - Cor do tecido
    - Tipo de colarinho

    - Estrutura do tecido
- Demográfico e biométrico (quem são eles)
    - Gênero
    - Faixa etária
    - BMI (índice de massa corporal)
- Geográfico (onde eles moram)
    - País/Nacionalidade
- Psicográfico (como eles se comportam)
    - Estilo de vida
- Comportamental (por que eles compram)
    - Sensibilidade a preço

Um importante quesito nesta etapa é a determinação do número de variáveis a serem inclusas no modelo, dado a limitação do tamanho da amostra de dados. Segundo Donilcar (2003), não foi ainda estabelecido uma metodologia apropriada para a relação entre número de variáveis e tamanho da amostra necessário. Em seu estudo, a autora destaca que Anton Formann (1982) é um dos poucos autores a propor tal relação, onde o tamanho mínimo da amostra deve ser aproximadamente 2^k, sendo k o número de variáveis na base de segmentação. Contudo, o autor sugere que o tamanho ideal para uma análise deveria respeitar a relação $5.(2^k)$; entretanto, esta regra desqualifica a maioria dos estudos publicados na área (Donilcar, 2003).

Para a escolha das variáveis, optou-se também pela utilização da base de dados da empresa. Neste, podem ser obtidas informações de cadastro para entrega e dados referentes ao produto enviado mensalmente. O Quadro 4 abaixo apresenta as variáveis disponíveis.

Dado que o banco de dados sofreu diversas adaptações desde a sua primeira versão com o lançamento da empresa, visando manter a homogeneidade das informações dos assinantes, foram coletados apenas os dados referentes ao ano de 2016. O tamanho da amostra coletada foi de 9940 resultados.

Valendo-se da relação de Formann, o número recomendado de variáveis do modelo deveria ser 13 no cenário real:

$$2^k \leq 9940 \therefore k \leq 13$$

Já no cenário ideal, 10 variáveis seriam permitidas:

$$5.(2^k) \leq 9940 \therefore k \leq 10$$

Quadro 4: Variáveis do modelo.

Variável	Tipo	Descrição	Exemplo de dado
created	Contínuo	Data de criação da assinatura	03/01/2016
age	Inteiro	Idade do assinante	31
gender	Nominal	Gênero do assinante	Feminino
active	Nominal	Status da assinatura	Ativo
plan	Nominal	Plano do assinante	18 snacks
box	Inteiro	Número de caixas recebidas	4
snacks	Inteiro	Número de snacks diferentes recebidos	15
region	Nominal	Estado do endereço do assinante	São Paulo
coupon	Nominal	Indica se o assinante utilizou cupom de desconto	Sim
channel	Nominal	Canal de mídia de aquisição do assinante	Facebook Anúncios

Fonte: Best Berry.

4.1.2. Definição da métrica de clusterização

Os dados obtidos apresentam variáveis nominais e contínuas. Sendo assim, conforme indicado por Kaufman & Rousseeuw (1990), a medida para geração de clusters escolhida foi a medida de Gowers, cujo detalhamento se encontra no Capítulo 3.3.2.

4.1.3. Definição do algoritmo de clusterização

Conforme explicado no Capítulo 3.3.1, ambos os métodos de particionamento e hierárquicos possuem frequência de uso semelhante para a segmentação de mercado. Contudo, nos casos estudados em Liao et al (2011), Hsu et al (2012) e Brito et al (2015), todos empregaram o algoritmo de particionamento *K-means*, sendo que Brito et al (2015) empregou uma das variações chamada *K-medoids*.

Um outro motivo para se optar pelo uso dos algoritmos de particionamento é o tamanho da base de dados. Dado que foram coletados cerca de 10000 resultados, o custo computacional necessário para executar os métodos hierárquicos seria muito elevado (Tan et al, 2005; Maimon & Rokach, 2005). Como os dados contém variáveis nominais, é difícil utilizar a noção de média, o que descarta o uso do *K-means*, assim como fora apontado na pesquisa de Brito et al (2015).

Além disso, o método *K-medoids* é mais robusto do que os métodos que utilizam a soma de quadrados, como é o caso do *K-means*. Apesar da simplicidade computacional do último, sua sensibilidade a *outliers* enfraquece o método (Kaufmann & Rousseeuw, 1990).

Finalmente, isso conclui que o melhor algoritmo a ser empregado é o *K-medoids*. No pacote "cluster" do software R, a função PAM é a responsável por realizar este algoritmo. Os detalhes do *K-medoids* e do PAM estão no Anexo A.

4.2.Validação do modelo

Conforme apontado no Capítulo 3.1.2.4, a segmentação, resultada da análise de clusters, precisa ser validada para justificar seu investimento. Para tanto, o resultado será avaliado de duas formas.

A primeira é a validação matemática da clusterização obtida. O coeficiente SC e as faixas propostas por Kaufmann & Rousseeuw são apropriadas para o caso em estudo.

A segunda é de caráter qualitativo sobre os segmentos obtidos, que devem seguir os critérios Blythe (2005) e Kotler & Armstrong (2015): (i) mensuráveis; (ii) acessíveis; (iii) substanciais; (iv) diferenciáveis; (v) acionáveis.

4.3.Elaboração das estratégias dos segmentos

Por fim, mediante o resultado dos clusters e dos segmentos obtidos, é possível realizar a recomendação de possíveis estratégias de Marketing direcionadas para cada grupo, de modo a aumentar a rentabilidade da empresa, conforme especificado no objetivo do trabalho.

5. RESULTADOS

5.1.Matriz de dissimilaridade

O software utilizado para o tratamento dos dados foi o R, amplamente utilizado para computação estatística e gráficos. Além disso, o pacote "cluster" foi empregado, o qual contém as ferramentas apresentas por Kaufman & Rousseeuw (1990).

A primeira etapa do modelo consistiu no cálculo das dissimilaridades, segundo o modelo de Gowers. Para tanto, o pacote "cluster" contém a função DAISY, a qual realiza o cálculo da matriz de dissimilaridade. Ao dados foram importados através de um arquivo CSV. Todos os comandos empregados no software R necessários para realizar a análise se encontra disponível no Anexo B.

Por conta do tamanho da matriz de ordem 9940, não foi possível encontrar alguma representação deste resultado de forma sucinta.

5.2.Algoritmo PAM

Para os métodos não hierárquicos como o *K-medoids*, é necessário fornecer além das dissimilaridades, o parâmetro k número de clusters, sendo que este número pode variar de 2 (mínimo de clusters) até $n - 1$, onde n é o número de registros.

Contudo, foi imposto que o valor máximo de k deveria ser máximo até 20. Dentre os principais motivos, estão o esforço computacional para gerar todas possibilidades e o *trade-off* da segmentação muito específica em termos de recursos da empresa.

A Tabela 5 contém o resultado das iterações do algoritmo PAM em função do número de clusters fornecido. A métrica utilizada para representar a qualidade da iteração e a sua validação é o SC, o qual fora detalhado no Capítulo 3.3.3. Para quebrar o desenvolvimento do estudo, optou-se pelo aprofundamento dos resultados apenas para a melhor clusterização, a qual será apresentada a seguir. Para maiores informações sobre as demais iterações, basta observar o Anexo C.

Tabela 5: SC em função do parâmetro k.

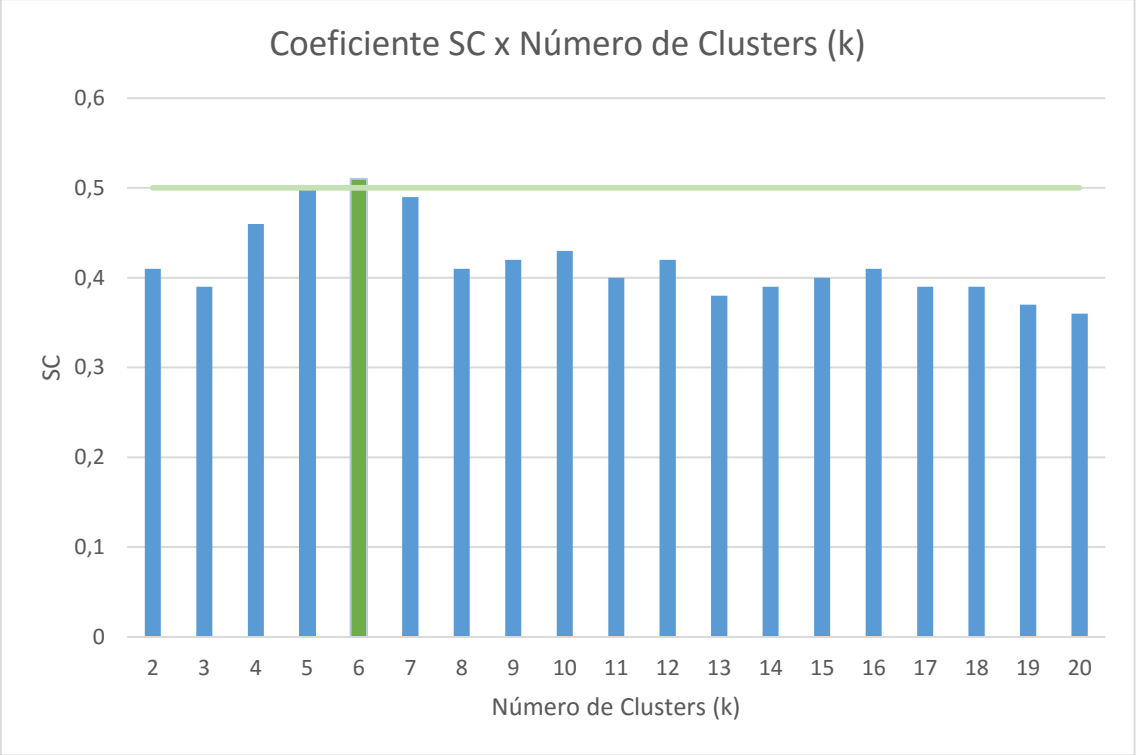
Número de Clusters	Coeficiente SC
2	0,41
3	0,39
4	0,46
5	0,50
6	0,51
7	0,49
8	0,41
9	0,42
10	0,43
11	0,40
12	0,42
13	0,38
14	0,39
15	0,40
16	0,41
17	0,39
18	0,39
19	0,37
20	0,36

5.3.Escolha da melhor clusterização

De acordo com Kaufman & Rousseeuw (1990), o modelo com SC contido no intervalo]0,50; 0,70] indica uma boa clusterização (Tabela 4). Verificando a Tabela 5, nota-se que o a clusterização aprovada mediante tal critério é a de 6 clusters.

Pode-se dizer então que a melhor segmentação obtida é para $k = 6$ e que a estrutura obtida na clusterização representa bem os dados. A Figura 21 ilustra os clusters obtidos em relação ao valor SC aceitável.

Figura 21: Relação entre coeficiente SC dos clusters e o valor de validação.



5.4.Detalhamento dos clusters

Definida a melhor clusterização, é possível verificar com mais detalhes o resultado obtido. A Tabela 6 contém a silhueta média dos clusters e mostra que todos os clusters obtidos atenderiam o critério de Kaufmann & Rousseeuw (1990), com exceção claro do cluster 2, cuja silhueta média está abaixo de 0,50. Dessa forma, é possível afirmar que a qualidade do cluster 2 é a menor.

Tabela 6: Silhueta média dos clusters.

Cluster	**Silhueta média**
1	0,52
2	0,49
3	0,52
4	0,51
5	0,53
6	0,53
GERAL	**0,51**

A Figura 22 apresenta a distribuição das silhuetas e a principal questão apontada por este gráfico é a ausência de componentes com a silhueta negativa. Isso significa que os objetos pertencentes a seu respectivo *medoid* é de fato a melhor solução; ou seja, nenhum objeto foi categorizado a um cluster errado.

Observando as dissimilaridades (Tabela 7), o primeiro ponto que chama atenção é a separação dos clusters baixa. Isso indica que existe uma certa proximidade entre os clusters e que a clusterização obtida não está bem isolada.

Outro ponto a ser notado é que a dissimilaridade média em relação aos *medoids* é baixa, justificando assim o agrupamento dos clusters. No entanto, o diâmetro deles (maior dissimilaridade dentro do cluster) e a dissimilaridade máxima em relação ao *medoid* é relativamente grande. Sendo assim, pode-se concluir que a maioria dos componentes dentro dos clusters estão próximos de seus respectivos *medoids*, contudo, existem uma pequena parcela que está distante do *medoid*. Finalmente, isso justifica a baixa separação entre os clusters.

O Quadro 5 contém os *medoids* dos 6 clusters obtidos, os quais representam melhor seus integrantes. A seguir serão detalhados cada um dos clusters com base no *medoid* e da distribuição dos dados conforme as variáveis do modelo.

Figura 22: Gráfico da distribuição da silheuta para 6 clusters.

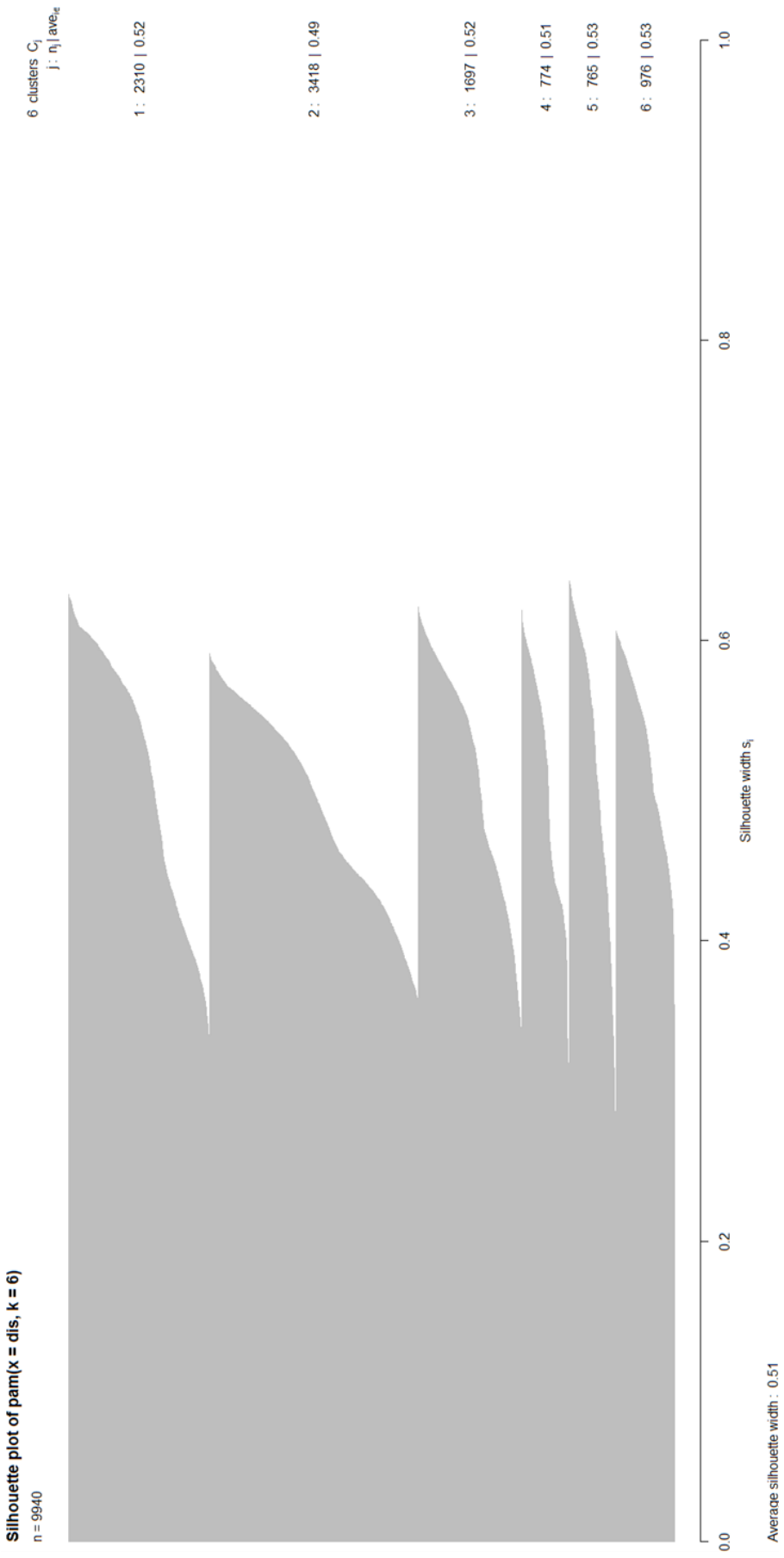


Tabela 7: Caracterização das dissimilaridades dos clusters.

cluster	size	max_diss	av_diss	diameter	separation
1	2310	0,19	0,07	0,31	0,10
2	3418	0,19	0,08	0,30	0,10
3	1697	0,35	0,09	0,46	0,10
4	774	0,33	0,11	0,54	0,10
5	765	0,27	0,07	0,36	0,10
6	976	0,22	0,07	0,35	0,10

Quadro 5: *Medoids* da clusterização.

	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5	cluster 6
age	32	36	38	26	33	24
gender	female	female	female	male	female	female
box	2	3	8	2	1	2
created	12/10/2016	17/05/2016	26/06/2016	14/06/2016	23/05/2016	23/08/2016
plan	18 snacks	18 snacks	18 snacks	10 snacks	10 snacks	18 snacks
status	Cancelado	Cancelado	Ativo	Cancelado	Cancelado	Cancelado
channel	Facebook Anúncios	Facebook Anúncios	Facebook Anúncios	Orgânico	Facebook Anúncios	Google Anúncios
snacks	8	9	24	8	5	8
coupon	yes	no	no	yes	yes	yes
region	Sudeste	Sudeste	Sudeste	Sudeste	Sudeste	Sudeste

5.4.1. Cluster 1: Experimentadoras

O cluster 1 é formado apenas por ex-assinantes do gênero feminino da Best Berry. Observando seu representante do Quadro 13 e a Figura 23, verifica-se que é uma base de clientes em que predomina a faixa etária jovem. Apesar do *medoid* ser de 32 anos, nota-se a maior concentração na faixa de 25 a 29, que consiste em 30% do grupo. A idade média do grupo é ligeiramente superior ao *medoid*: 35 anos.

Através da Figura 24, nota-se que o número de caixas recebidas do grupo é baixa, com média de 2,3 caixas, sendo o *medoid* de 2 caixas. Outra curiosidade percebida foi que a distribuição de caixas recebidas por este cluster pode ser modelado por uma função logarítmica, com coeficiente de determinação $R^2 = 0{,}98$, próximo de 1, indicando uma boa aproximação. Tal comportamento é decorrente do modelo de assinatura adotado pela empresa, sendo que esta informação é uma boa estimativa para a taxa de cancelamento deste perfil.

A pouca variedade de *snacks* (Figura 25) é decorrente do número baixo de caixas recebidas, com uma grande concentração de 6 *snacks* (43% dos componentes do cluster 1). Além disso, foi identificado um caso com 5 variações *snacks*, que pode ser considerado um *outlier*, dado que um assinante com o plano de 18 *snacks* deveria ter recebido uma variação de 6 tipos.

Analisando o comportamento de compra, todos os componentes utilizaram algum cupom de desconto na sua assinatura. Além disso, nota-se pela Figura 26 que o período de maior aquisição deste perfil foi no mês de outubro e novembro (juntos, representam 43% das aquisições do ano). Coincidentemente, é o período das ações de *Black Friday*, em que a Best Berry e muitas empresas de varejo trabalham com descontos agressivos.

Outra característica deste cluster é que o canal de aquisição destes clientes foi através dos anúncios do Facebook, compreendendo 100% do cluster.

A Figura 27 revela a maior concentração desses clientes na região Sudeste (80%), conforme percebido pelo *medoid* do cluster, com uma pequena parcela pertencente à região Sul (14%). As demais regiões não são o principal foco da empresa, justificando assim a participação de apenas 6% delas.

Por meio destas informações, pode-se deduzir que o perfil do cluster 1 é formado pelas **Experimentadoras**: mulheres jovens que optaram pelo plano com maior variedade de *snacks* e tiveram curiosidade pelo produto. A compra da primeira caixa se deve apenas pela disponibilidade de um cupom de desconto. Após experimentarem o produto, concluíram que a manutenção da assinatura não era vantajoso.

Figura 23: Distribuição da faixa etária do cluster 1.

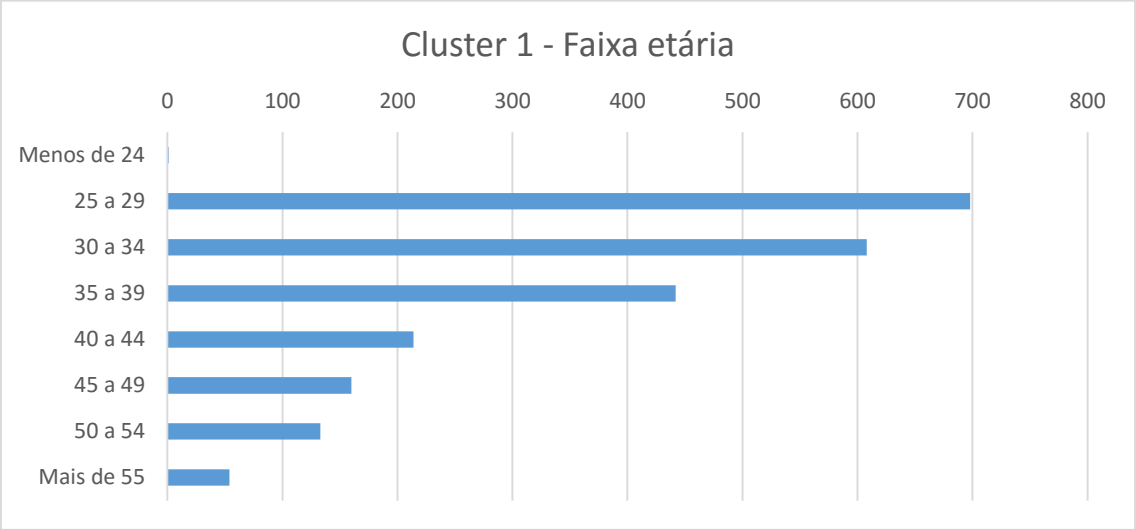


Figura 24: Distribuição de caixas recebidas do cluster 1.

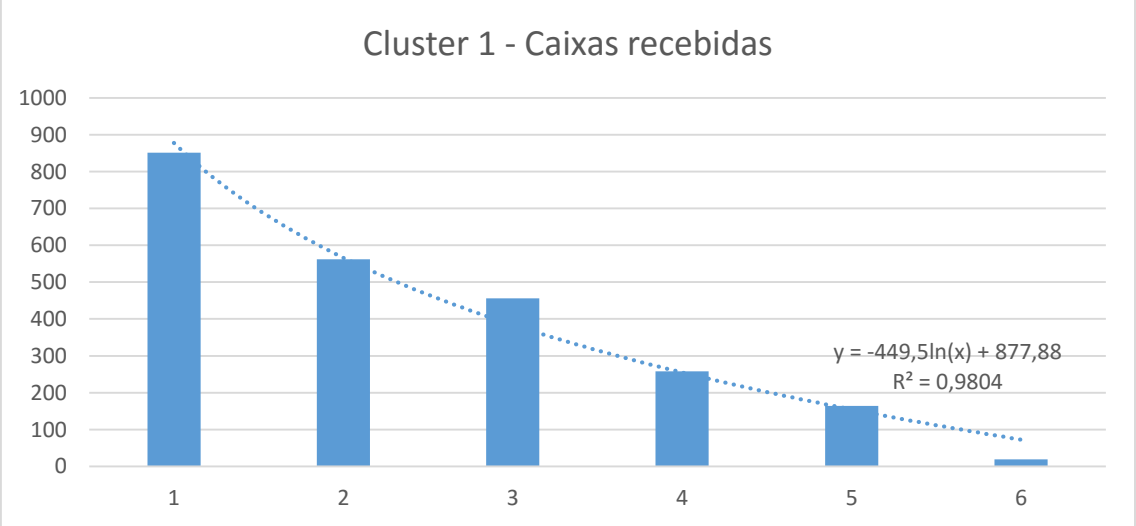


Figura 25: Distribuição de snacks recebidos do cluster 1.

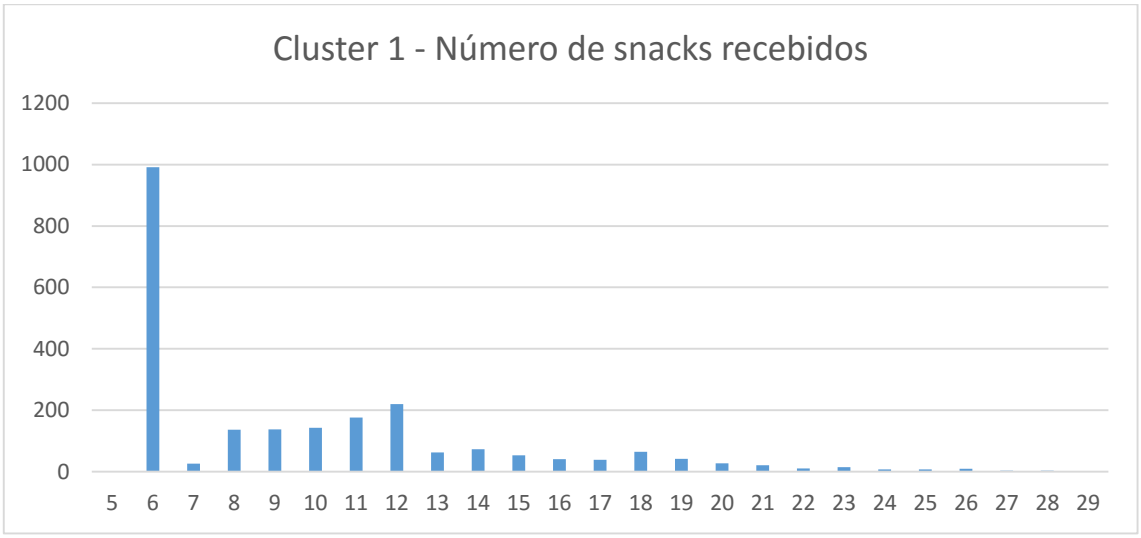


Figura 26: Distribuição da data de criação da assinatura do cluster 1.

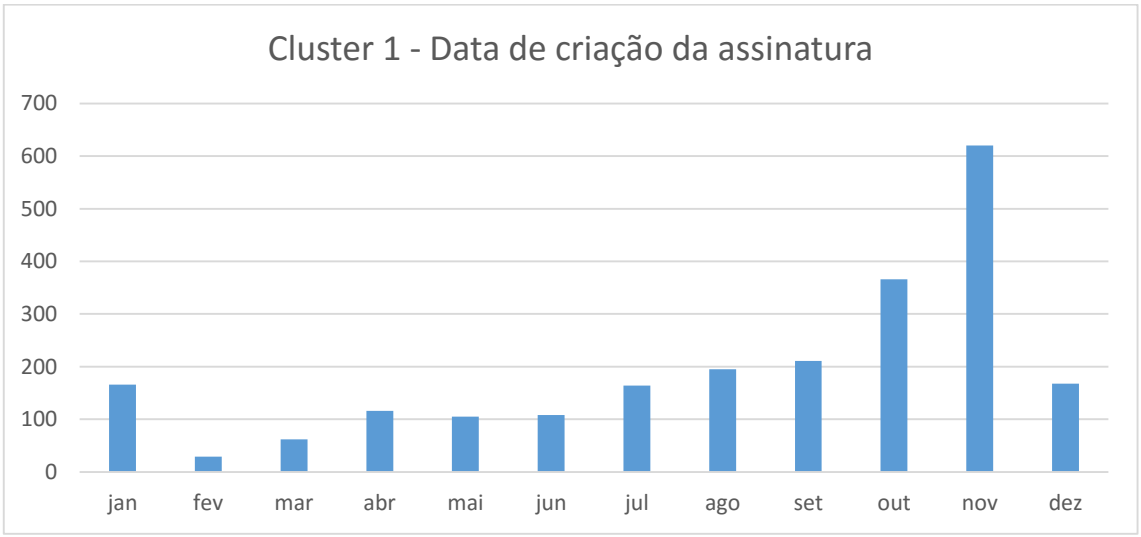
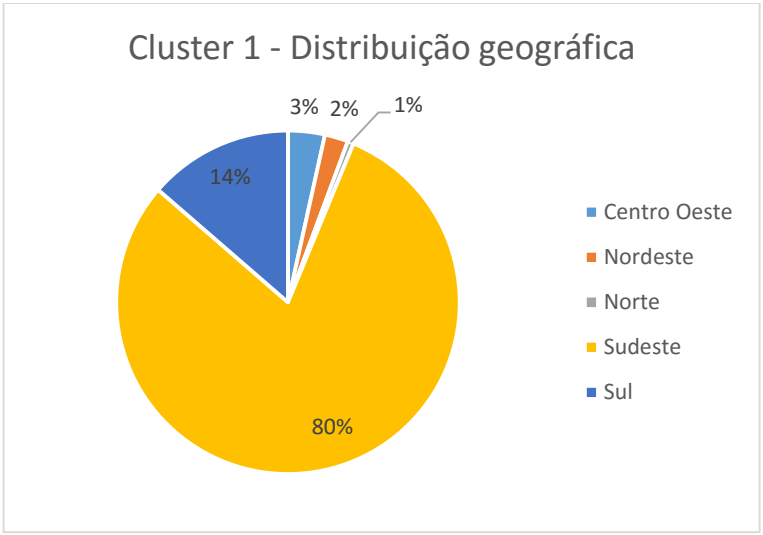


Figura 27: Distribuição geográfica do cluster 1.



5.4.2. Cluster 2: Quase assinantes

O cluster 2 se assemelha com o cluster 1 pelo fato de ambos apresentarem assinantes do gênero feminino, status da assinatura cancelado, plano de 18 *snacks* e o canal de aquisição são anúncios do Facebook.

Contudo, analisando a Figura 28, verifica-se que a distribuição de idade é diferente comparada ao cluster 1: a maior concentração está na faixa de 30 a 39 anos (25% do cluster). A média da idade do cluster 2 é de 37 anos, próxima à idade de seu *medoid* (36 anos).

O *medoid* do cluster 2 apresenta um número de caixas ligeiramente superior em relação ao cluster 1, e o mesmo é verificado comparando a média (2,7 caixas). A principal diferença notada é na distribuição de caixas recebidas, em que o cluster 2 contém uma queda menos acentuada, conforme a Figura 29. Foi possível detectar uma tendência linear entre o número de assinantes e o mês, com $R^2 = 0{,}97$. Por fim, o último ponto a ser ressaltado é a quebra entre os meses 3 e 4 (36%), e entre os meses 5 e 6 (53%). A taxa de cancelamento para estes períodos é bem alta para os padrões da empresa, e poderia ser feita uma análise mais detalhada sobre os motivos.

Em termos de variação de *snacks* (Figura 30), a cluster 2 apresenta uma distribuição mais uniforme na faixa de 8 a 12 *snacks*, porém, sem muita diferença quando comparado ao cluster 1.

Outro ponto que diferencia este cluster em relação ao cluster 1 é a ausência de cupom de cupom de desconto na assinatura.

Analisando a Figura 31, a maior concentração de vendas ocorre nos meses de agosto e setembro (15% e 13%, respectivamente). Nos meses entre janeiro e maio, as aquisições representam entre 10 e 12% do ano. Outro fato curioso é a ausência de aquisições durante os meses de novembro e dezembro.

Em relação a distribuição geográfica (Figura 32), ainda predomina a região Sudeste (73%) e uma parcela menor do Sul (14%, a mesma proporção do cluster 1). Contudo, nota-se uma maior participação das regiões Centro Oeste, Nordeste (7% e 5%, respectivamente).

O cluster 3 será detalhado a seguir, porém é possível inferir que este possui características bem semelhantes ao cluster 2, com a grande diferença do status da assinatura. Dessa forma, pode-se categorizar o cluster 2 como **Quase assinantes**. É formado por moças mais maduras, com uma situação financeira mais estável a qual permitiu criar a assinatura sem o uso de desconto. Apesar do produto ter uma sinergia com suas necessidades, a experiência com o produto não atendeu suas expectativas, provocando o cancelamento da assinatura.

Figura 28: Distribuição da faixa etária do cluster 2.

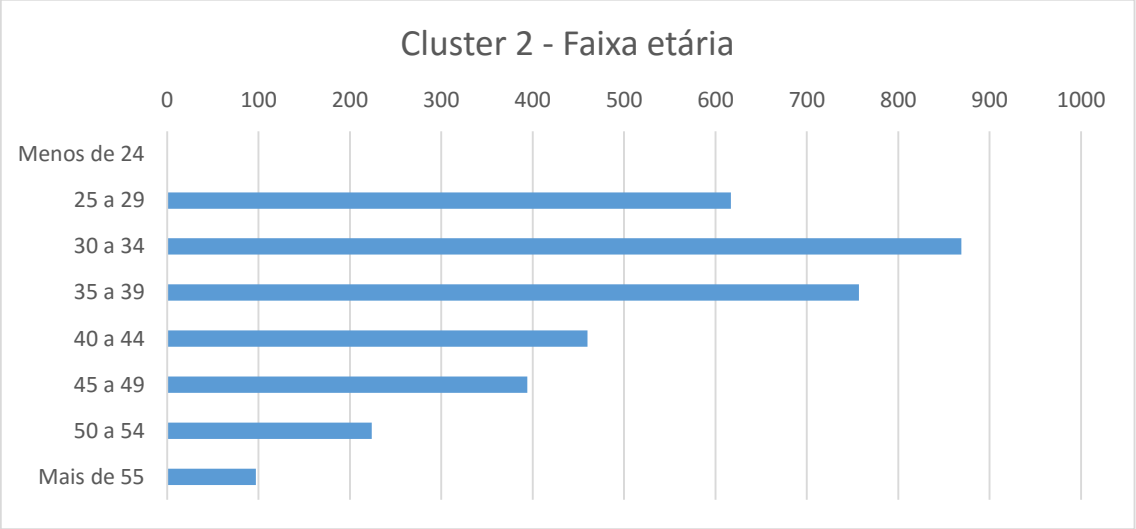


Figura 29: Distribuição de caixas recebidas do cluster 2.

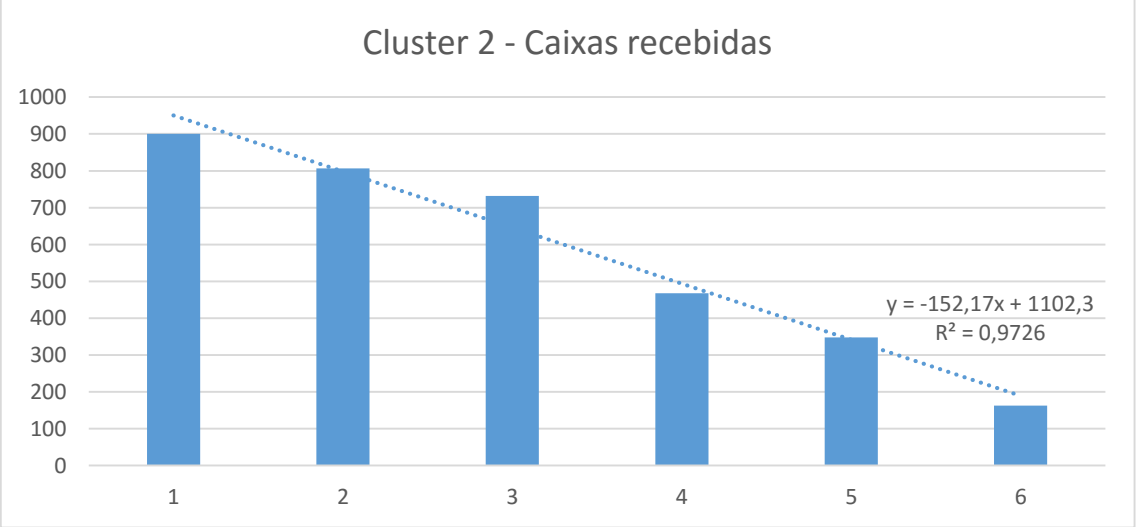


Figura 30: Distribuição de snacks recebidos do cluster 2.

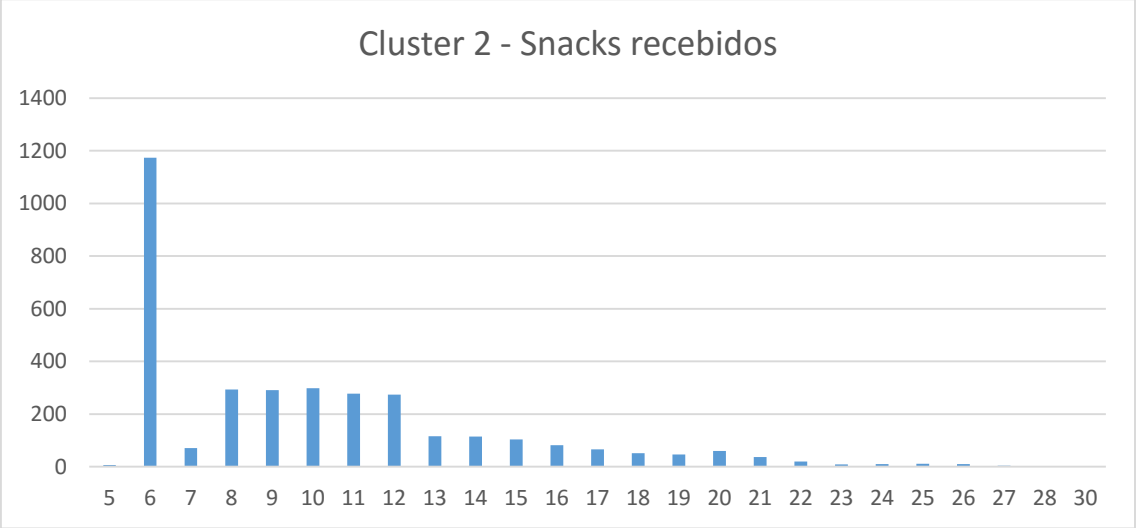


Figura 31: Distribuição da data de criação da assinatura do cluster 2.

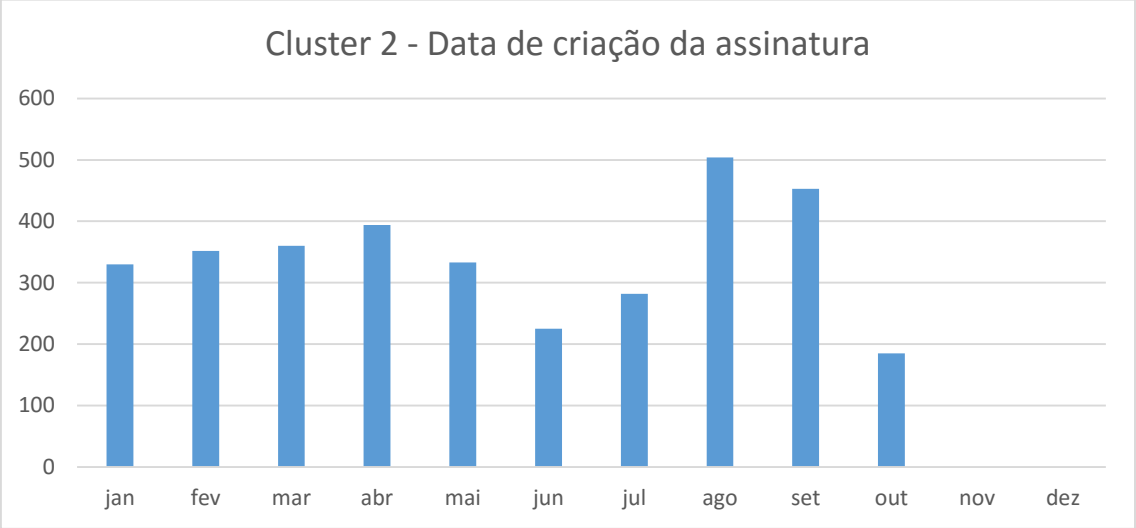
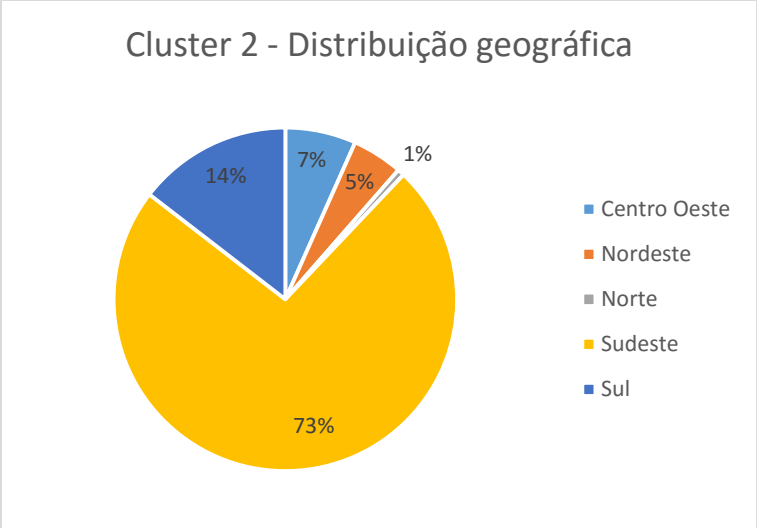


Figura 32: Distribuição geográfica do cluster 2.



5.4.3. Cluster 3: Mina de Ouro

É o cluster mais atrativo, pois contém todas as assinaturas ativas de clientes do gênero feminino. Felizmente, segundo o posicionamento da empresa, é o grupo que mais se assemelha ao público alvo.

Analisando a distribuição da faixa etária (Figura 33), é possível notar que trata-se de um público mais maduro, com maior representatividade entre 30 e 44 anos (43% dos componentes do cluster). Seu representante (*medoid*) é de 38 anos e a média do cluster é bem próxima, de 39 anos.

Não foi preciso o uso de cupom para estimular a assinatura e tais assinantes aderiram ao plano de 18 *snacks*.

Por ser uma base ativa, o número de caixas recebidas é elevado (Figura 34). O *medoid* acusa 8 caixas recebidas e a média do cluster é de 9, sendo novamente bem coerentes entre si. A faixa entre 7 e 9 caixas representa 51% dos componentes do cluster.

Como resultado do perfil de caixas recebidas, nota-se uma ampla variedade de *snacks* experimentados (Figura 35). A média é de 25 tipos de *snacks*, compatível com os 24 do *medoid*. A maior concentração ocorre na faixa de 20 a 25 *snacks*, em que se percebe a frequência maior do que 100. Isso resulta em uma parcela de 48% do cluster.

É interessante notar uma possível sazonalidade na aquisição deste perfil de cliente: os maiores volumes de vendas ocorrem nos inícios de semestre (janeiro e agosto), sendo que ambos ultrapassam a marca de 200 aquisições, e ocorre uma queda entre eles, com menor volume em junho (5% do volume do ano) e dezembro (apenas 2% do total). Tal distribuição pode ser observada na Figura 36.

Trata-se de um cluster com mais variação em termos de canais de aquisição, apesar de 95% delas serem via anúncios do Facebook. A parcela menor se divide em anúncios do Google e em mídia orgânica não paga (Figura 37).

A distribuição geográfica segue um padrão semelhante aos clusters anteriores, com 78% na região Sudeste, 13% na região Sul e 9% nas demais regiões (Figura 38).

As informações mostram que este é o principal cluster da empresa e que entende a proposta de valor do produto, sendo que tal inferência é percebida pelo plano mais completo de 18 *snacks* e a ausência do uso de cupom de desconto. O público tem uma idade mais avançada em relação ao cluster 2, reafirmando a importância de uma maior estabilidade pessoal para a manutenção da assinatura. Dessa forma, pode-se dizer que este segmento é a verdadeira **Mina de Ouro**.

Figura 33: Ditribuição da faixa etária do cluster 3.

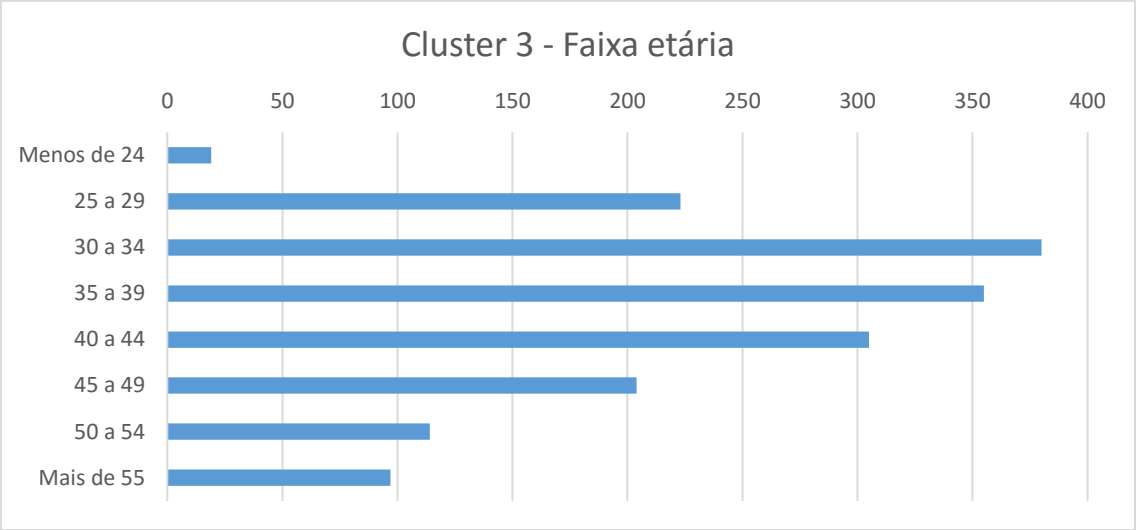


Figura 34: Distribuição de caixas recebidas do cluster 3.

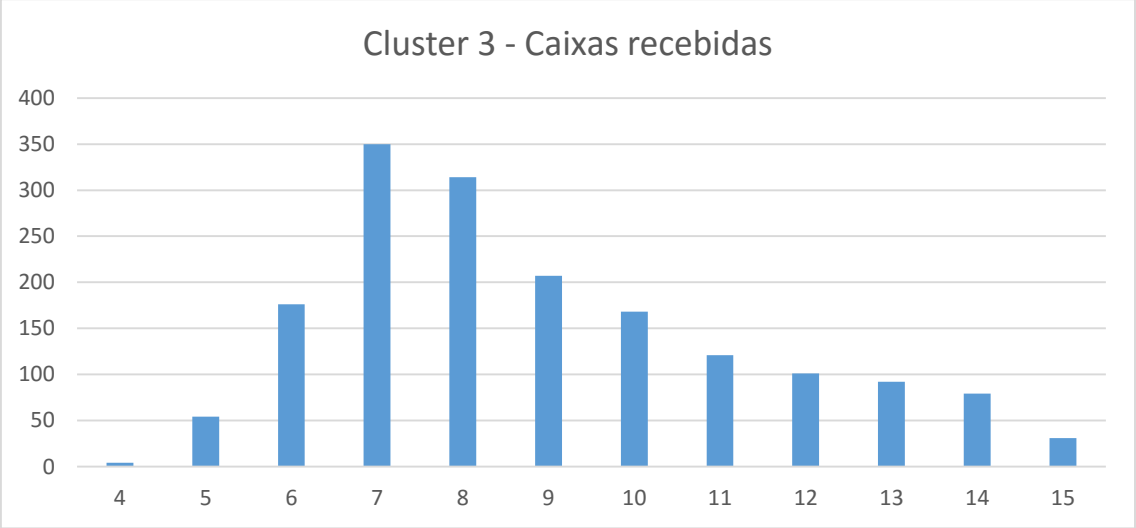


Figura 35: Distribuição de snacks recebidos do cluster 3.

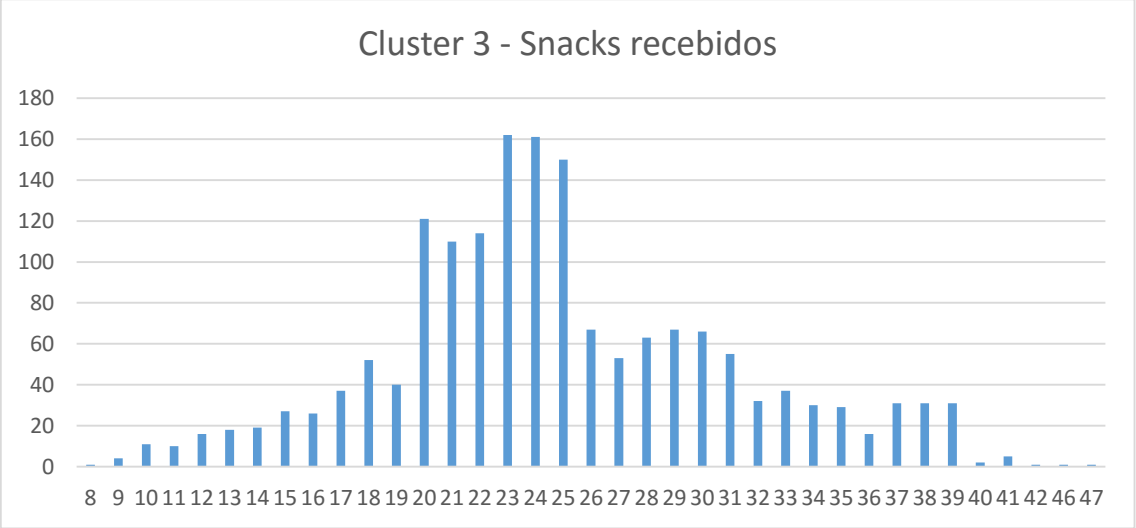


Figura 36: Distribuição da data de criação da assinatura do cluster 3.

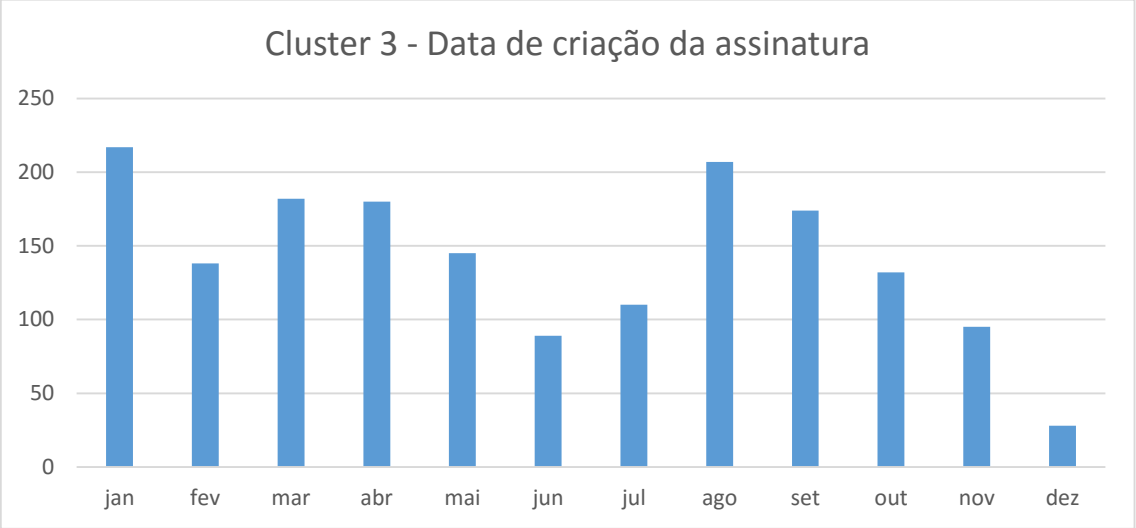


Figura 37: Distribuição dos canais de aquisição do cluster 3.

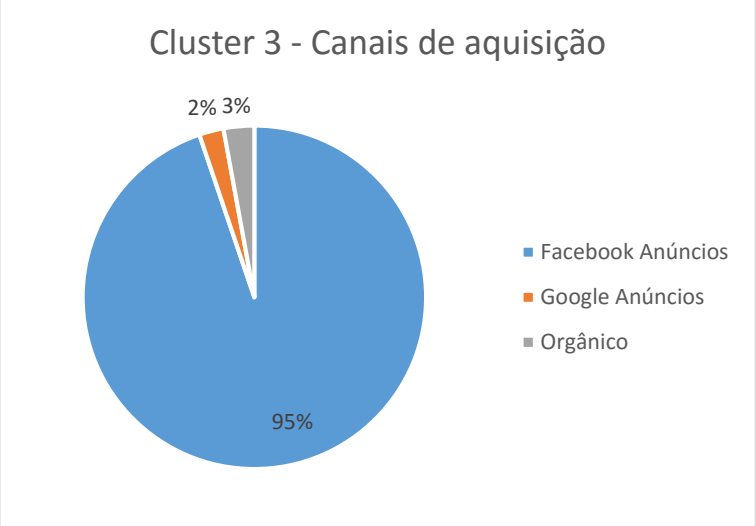
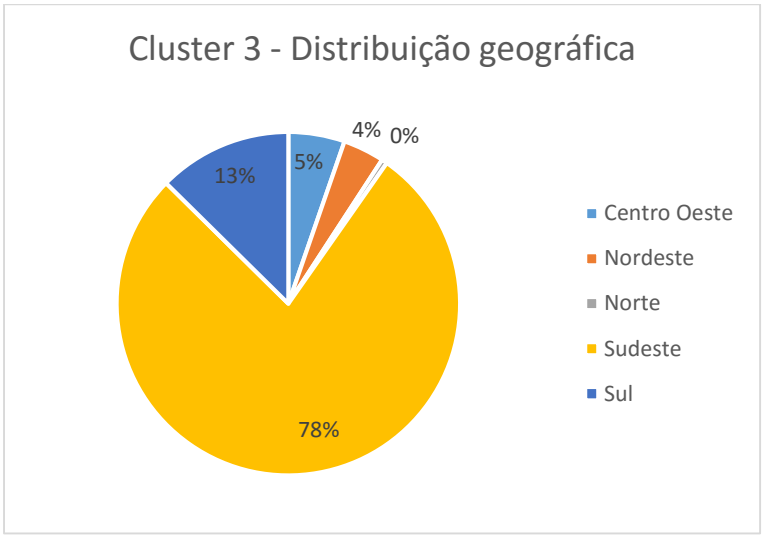


Figura 38: Distribuição geográfica do cluster 3.



5.4.4. Cluster 4: Vaidosos

O quarto cluster representa a base de assinantes do gênero masculino. Todos estes assinantes são do plano de 10 *snacks* e, em sua grande maioria, utilizaram algum cupom de desconto.

Outra característica importante deste cluster é a faixa etária relativamente baixa, com o principal intervalo de 20 a 34 anos, representando 94% dos componentes do cluster (Figura 39).

Em relação ao status da assinatura, prevalece o cancelamento com 97% dos componentes do cluster (Figura 40).

O número de caixas recebidas deste cluster é baixo (Figura 41). A média do cluster é de 3 caixas recebidas, sendo que o *medoid* é de 2 caixas. A maior concentração é de apenas 1 caixa (32% do cluster). Assim como o cluster 1 e 2, é possível traçar uma curva de tendência que relaciona o número de caixas ao longo do tempo. Neste caso, a melhor aproximação foi a exponencial, com $R^2 = 0{,}96$, indicando um bom modelo.

A pouca variação de *snacks* é, novamente, decorrente do perfil de caixas recebidas (Figura 42). Neste caso, o plano de 10 *snacks* provoca uma maior concentração em 5 tipos de *snacks* recebidos, o que equivale a 33% do cluster. Dado que neste cluster tem-se a presença de usuários ativos, cuja tendência é de apresentar uma variedade de *snacks* maior, a média é maior de 11 *snacks*. Comparado com o *medoid* de 8 *snacks*, percebe-se a importância de verificar a distribuição da Figura 42.

Este cluster apresenta maior variedade em termos de canais de aquisição (Figura 43). Ele contém Orgânico (72%), Email e Afiliados (28%), sendo que o último canal está compreendido apenas neste cluster.

As principais datas de criação da assinatura deste cluster (Figura 44) é nos meses de outubro e novembro (26% somando os dois meses), semelhante ao cluster 1. Nos demais meses, o volume de aquisições é aproximadamente constante.

Na Figura 45, a região Sudeste é possui uma maior distribuição em relação aos demais clusters, com uma participação de 85%. A região Sul é consideravelmente menor no cluster 4, com apenas 6%, sendo este outro diferencial deste grupo. A região Centro Oeste é maior do que a Sul, com 7% e as regiões Nordeste e Norte são também a minoria, como nos demais clusters, com apenas 2%.

Conforme visto por seu *medoid*, as características mais marcantes são o gênero masculino e o perfil jovem. Dado que o produto da Best Berry é relacionado à saúde e bem estar, faz sentido a rotulação de **Vaidosos** para o cluster 4. Justamente pelo fato do produto não ser projetado para o público masculino e jovem, é natural que este perfil se assemelhe aos demais clusters de cancelados. O grupo apresenta uma minoria de 23 clientes ativos (3% do grupo), e, por conta deste tamanho pequeno, não se justifica tratar este subgrupo de forma diferenciada.

Figura 39: Distribuição da faixa etária do cluster 4.

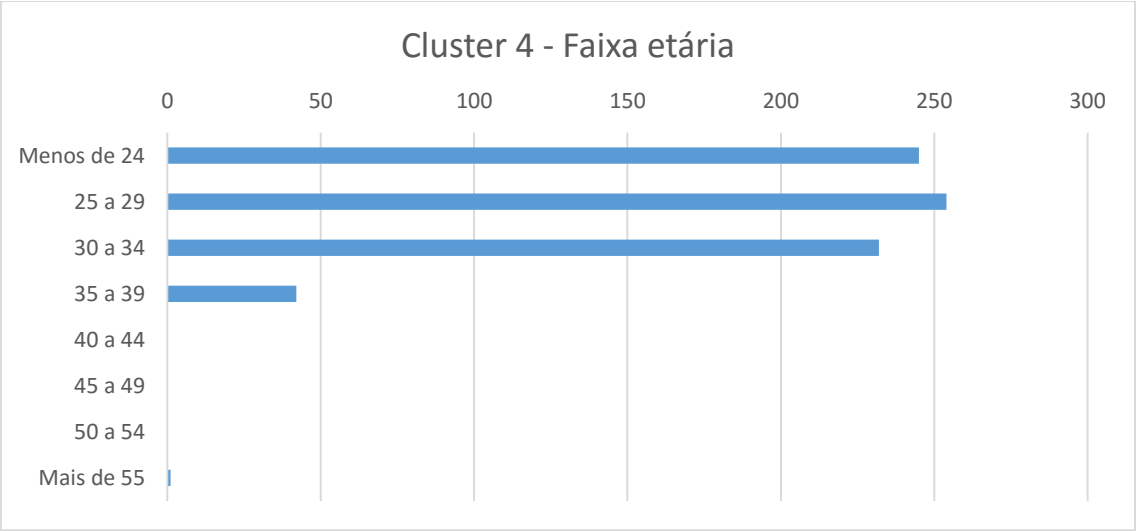


Figura 40: Distribuição do status da assinatura do cluster 4.

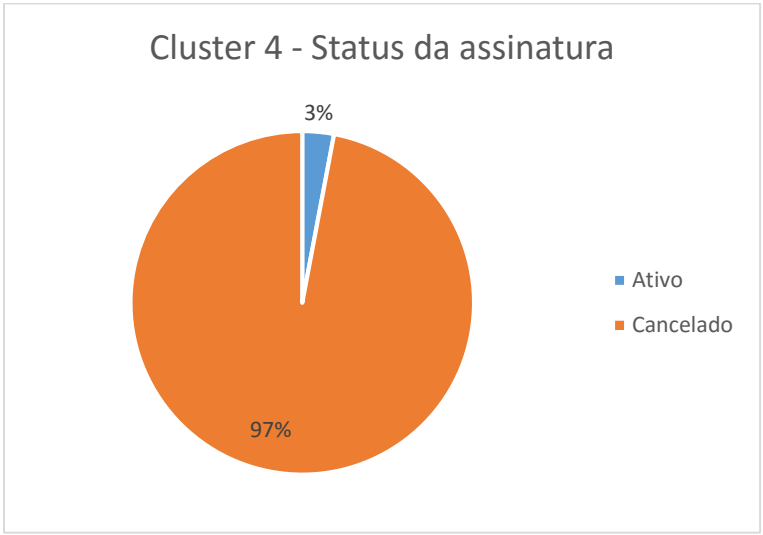


Figura 41: Distribuição de caixas recebidas do cluster 4.

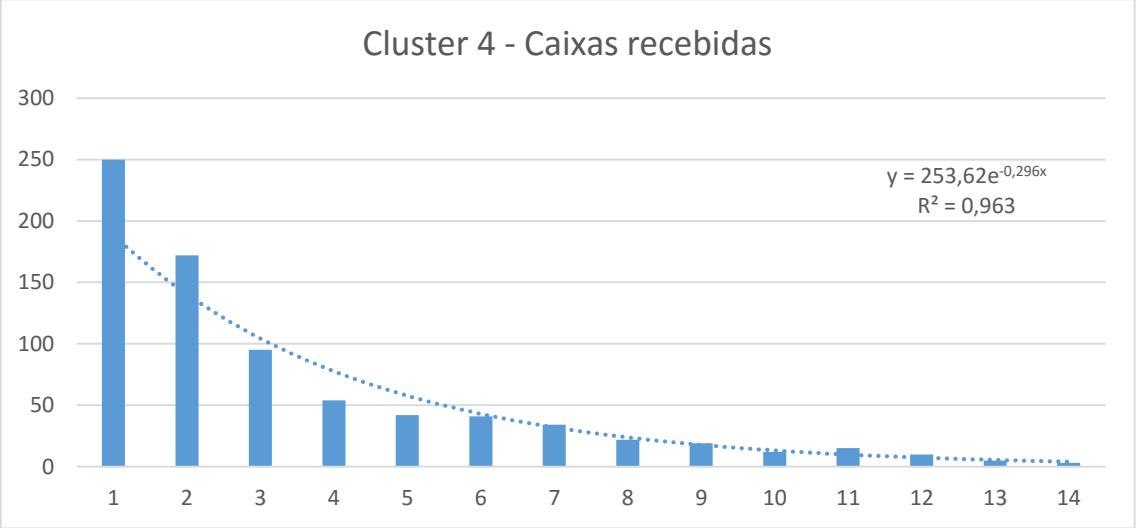


Figura 42: Distribuição de snacks recebidos do cluster 4.

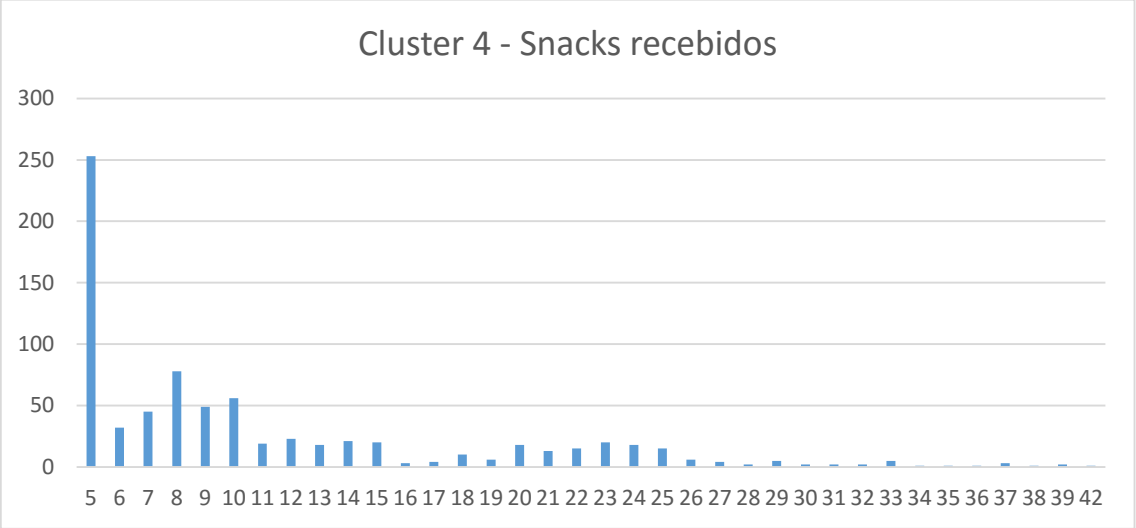


Figura 43: Distribuição dos canais de aquisição do cluster 4.

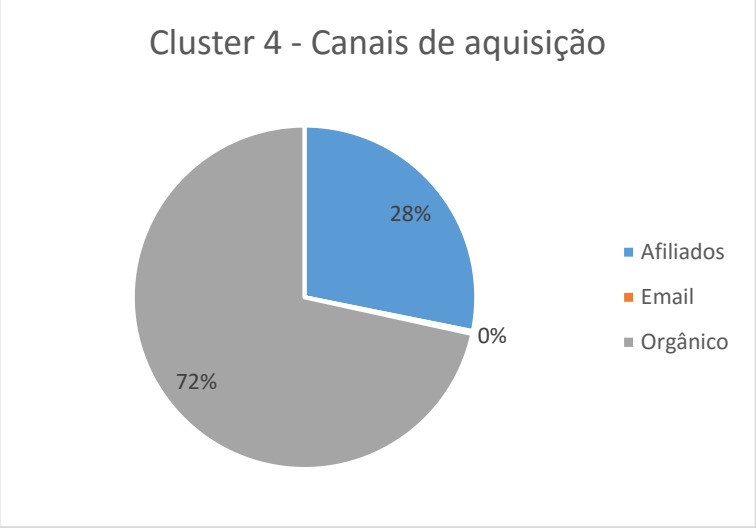


Figura 44: Distribuição da data de criação da assinatura do cluster 4.

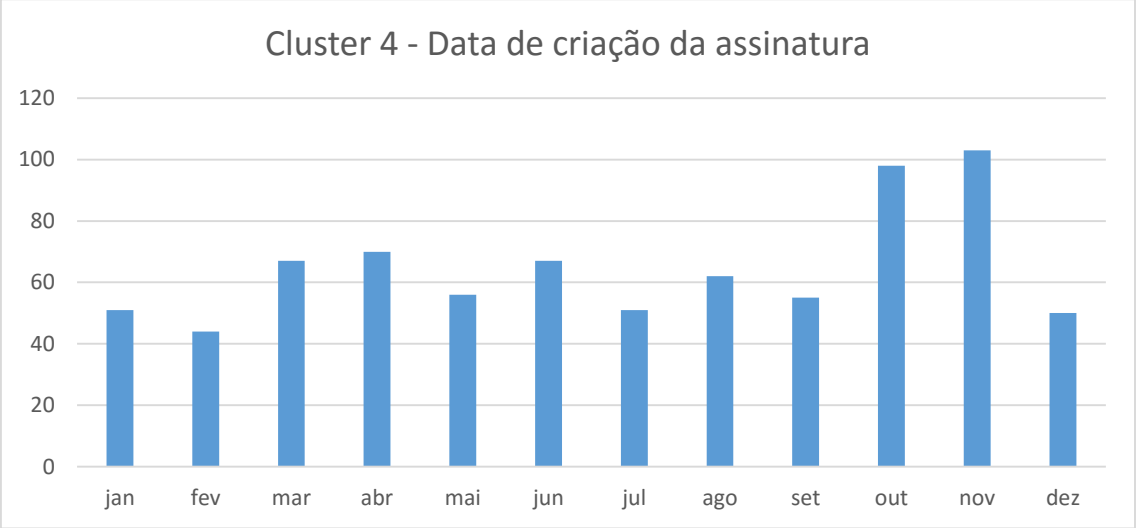
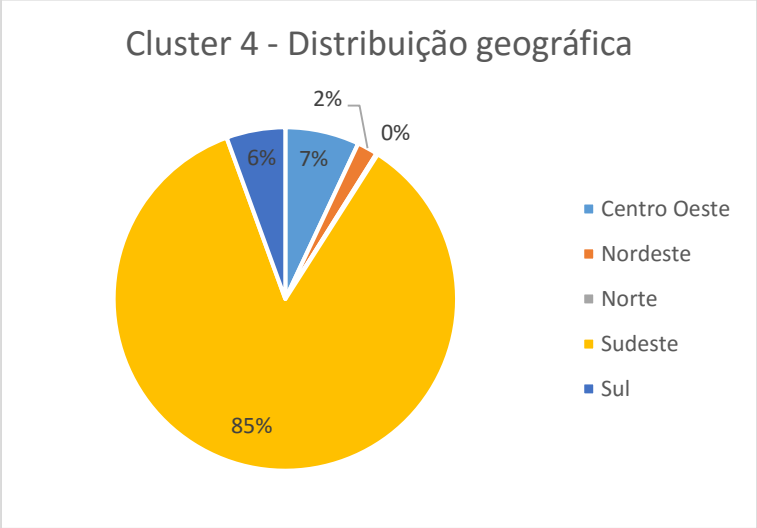


Figura 45: Distribuição geográfica do cluster 4.



5.4.5. Cluster 5: Caçadoras de Descontos

O cluster 5 é outro tipo de perfil feminino de assinantes da empresa. Trata-se de mais um clusters com assinantes já cancelados, que aderiram ao plano de 10 *snacks* através do uso de cupom de desconto.

É um público jovem (Figura 46), semelhante ao cluster 1. O *medoid* é de 33 anos, próxima à média do cluster de 35 anos. A faixa de maior concentração é de 25 a 39 anos, o que representa um percentual de 78%.

Verifica-se que o número de caixas recebidas é bem baixo através do *medoid* com valor de 1 caixa, da média de 1,6 caixa e da distribuição presente na Figura 47. É o cluster menor amplitude de caixas recebidas, com o valor máximo de apenas 4 caixas. Novamente, é possível verificar uma tendência exponencial do número de caixas com o tempo. O coeficiente de determinação é de $R^2 = 0{,}97$.

A variação predominante é de 5 *snacks* (56% do cluster), que é justamente referente à compra de 1 caixa (Figura 48). O *medoid* também é de 5 *snacks* e a média do grupo é de 7 *snacks*.

O perfil da data de assinatura (Figura 49) difere dos demais clusters por conta do maior volume do primeiro semestre do ano (59% das assinaturas no ano), com exceção do mês de novembro, que representa 14% do ano e 33% do segundo semestre.

Em termos de canais de aquisição, além dos anúncios no Facebook (96% das aquisições), existe uma pequena parcela de Email (4%) (Figura 50).

A distribuição geográfica segue o mesmo padrão dos clusters anteriores, com a maior presença da região Sudeste (80%) e Sul (13%) (Figura 51).

Por conta do comportamento de compra do plano mais simples (10 *snacks*) e uso do cupom, acredita-se que o fator principal para este cluster é o preço do produto. Dado que a assinatura efetuada é a com menor preço possível da Best Berry, o nome **Caçadoras de Descontos** reflete bem esta característica. Por consequência, é um cluster que dificilmente mantém a assinatura, tornando-se pouco atraentes em termos de negócio.

Figura 46: Distribuição da faixa etária do cluster 5.

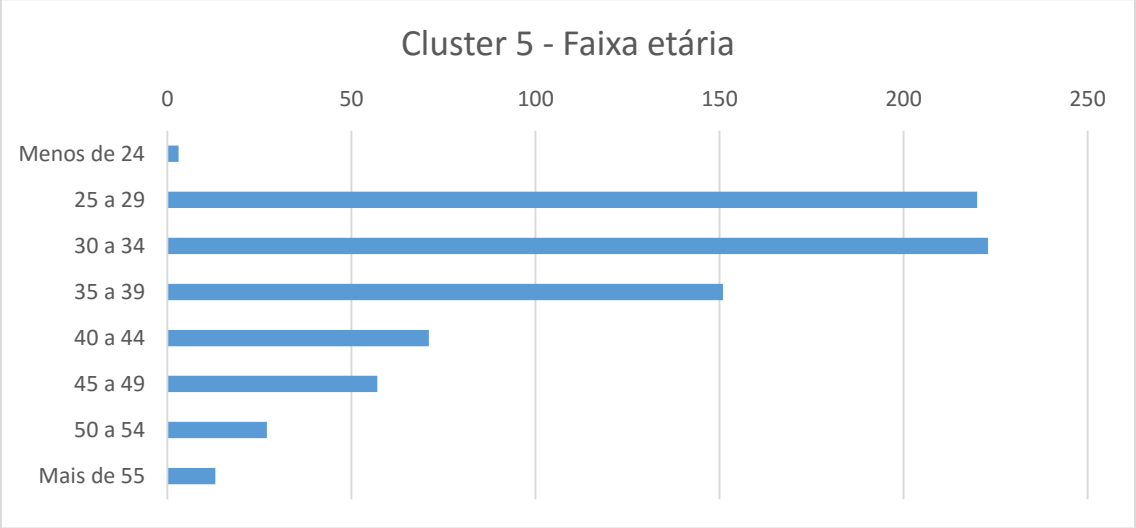


Figura 47: Distribuição de caixas recebidas do cluster 5.

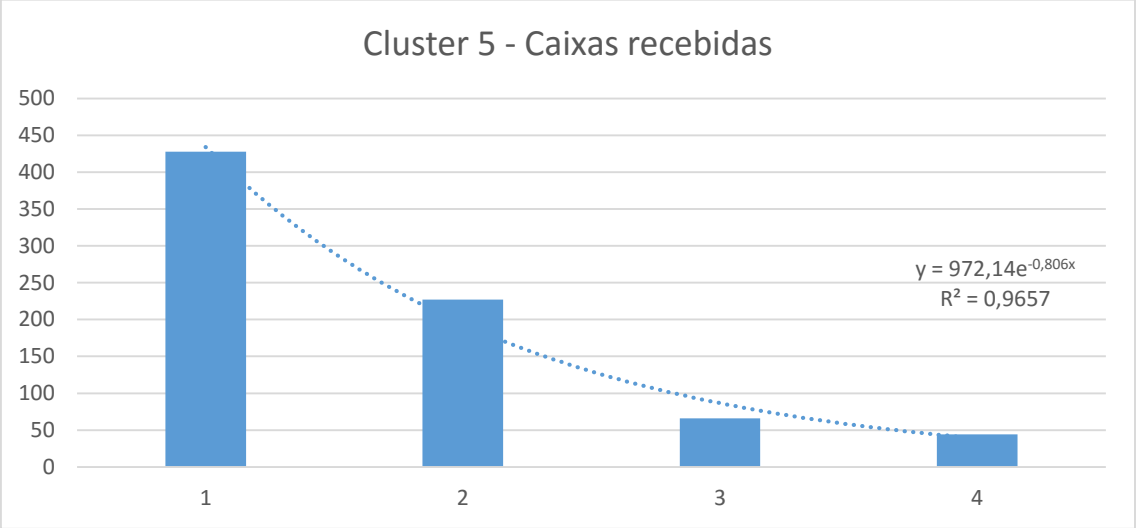


Figura 48: Distribuição de snacks recebidos do cluster 5.

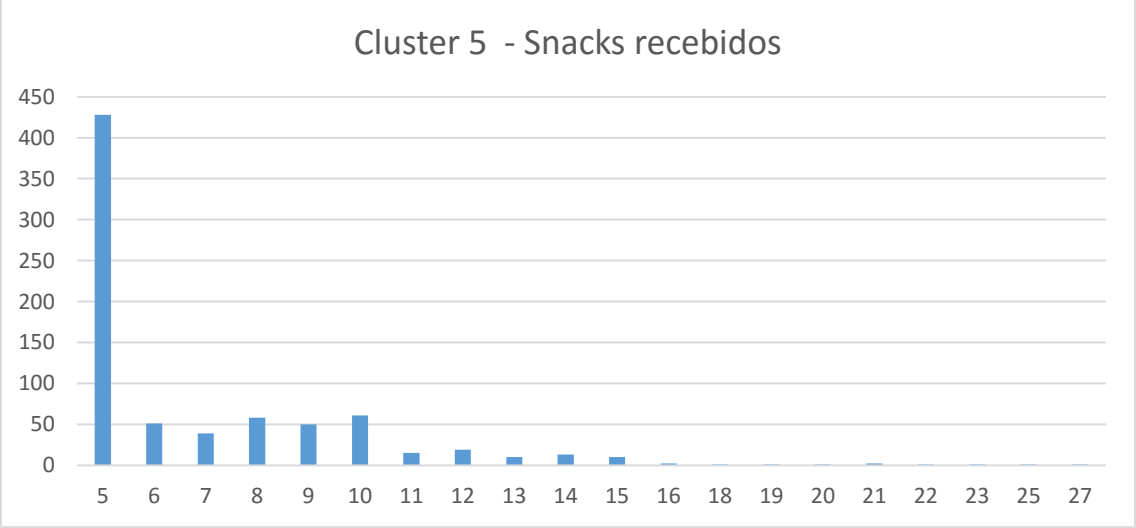


Figura 49: Distribuição da data de criação da assinatura do cluster 5.

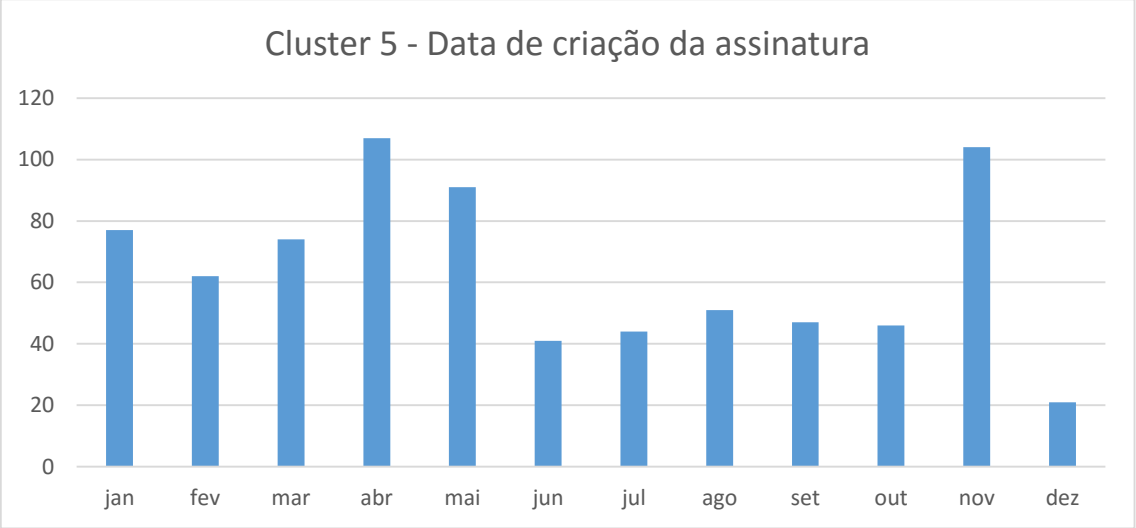


Figura 50: Canais de aquisição do cluster 5.

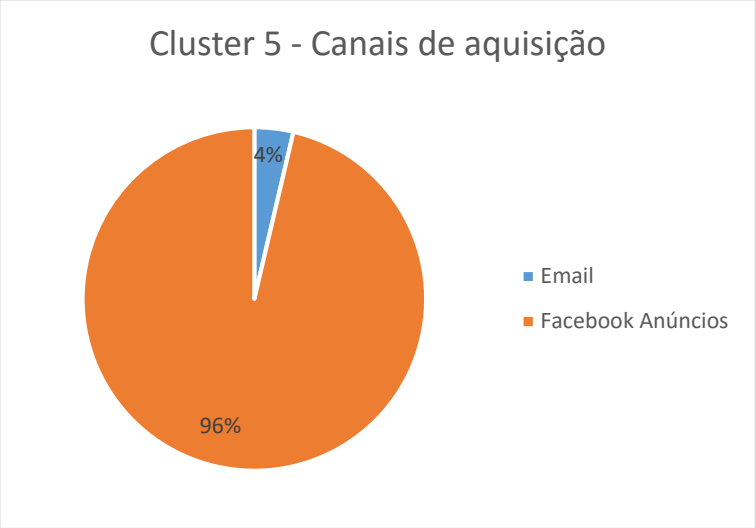
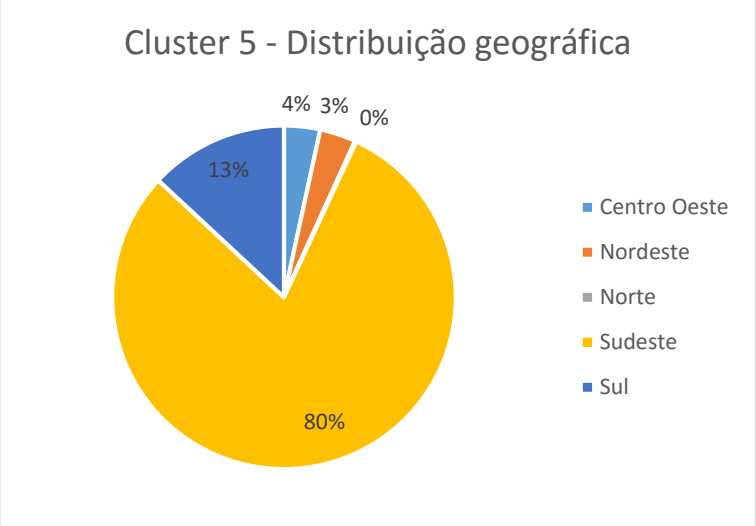


Figura 51: Distribuição geográfica do cluster 5.



5.4.6. Cluster 6: #BestBerry

O cluster 6 é caracterizado por apresentar apenas assinantes do gênero feminino, com o plano de 18 *snacks* e o uso de cupom de desconto. Além disso, todas as integrantes deste cluster foram obtidas através do uso dos anúncios do Google.

Um outro fator determinante para a caracterização deste cluster é a faixa etária jovem destas assinantes (Figura 52). É o cluster com menor amplitude de valores para esta variável, variando entre 20 e 25 anos. Seu *medoid* é de 24 anos e a média do cluster é de 23 anos.

Em termos de comportamento de compra, são poucas as assinantes ainda ativas, apenas 2% do grupo (Figura 53).

Assim como os demais clusters com assinantes cancelados, o número de caixas recebidas é baixo, sendo que o *medoid* e a média são de 2 caixas. Foi possível também traçar a linha de tendência exponencial para o número de caixas em função dos meses, com o parâmetro $R^2 = 0{,}94$. A Figura 54 apresenta a distribuição desta variável para o cluster 6.

Em relação ao número de *snacks* experimentados, a distribuição segue semelhante ao cluster 1 (Figura 55). O *medoid* é de 8 tipos de *snacks*, próximo à média do cluster de 9 *snacks*. Predomina o número de 6 *snacks*, referente à quantidade do plano de 18. Isso representa uma parcela de 48% dos componentes do grupo.

Outro fenômeno interessante é o volume de assinaturas maior no segundo semestre do ano (Figura 56). São 34% das aquisições para o primeiro semestre e 66% para o segundo semestre.

Novamente, a distribuição geográfica indica a maior presença de assinantes na região Sudeste do país (Figura 57). Contudo, é uma distribuição mais diversificada, com 70% região Sudeste, 18% região Sul e 12% nas demais regiões.

#BestBerry foi o nome escolhido para representar este cluster. As principais motivações para esta rotulação foram a faixa etária muito jovem e o canal de aquisição de anúncios do Google. Juntos, eles indicam que se tratam de um público mais adepto à tecnologia e que realiza uma pesquisa mais intensa para tomar suas decisões de compra. Contudo, como ainda não estão em um período estável para manter a assinatura, experimentam alguns *snacks* e optam por finalizar a assinatura.

Figura 52: Distribuição da faixa etária do cluster 6.

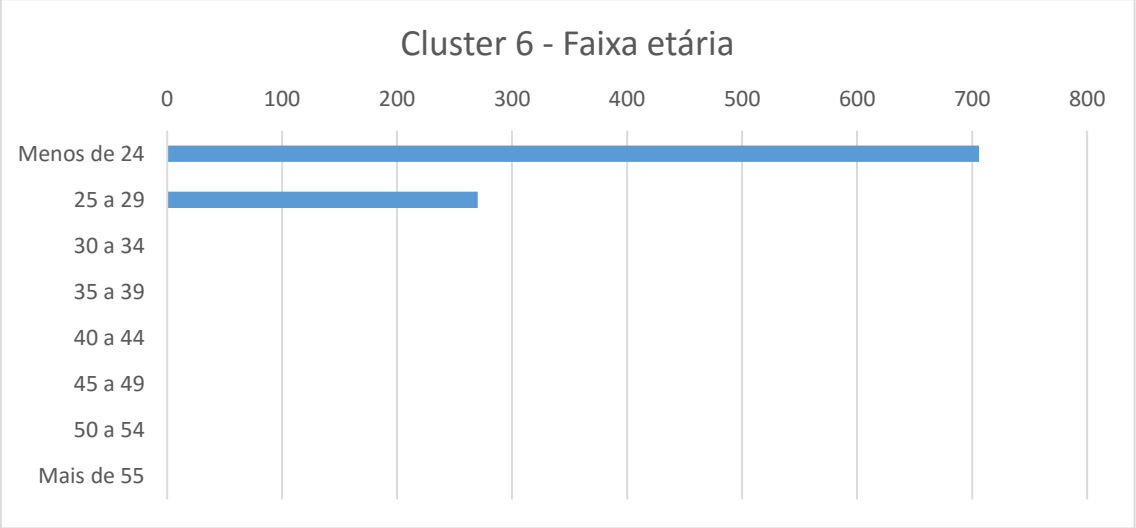


Figura 53: Distribuição do status da assinatura do cluster 6.

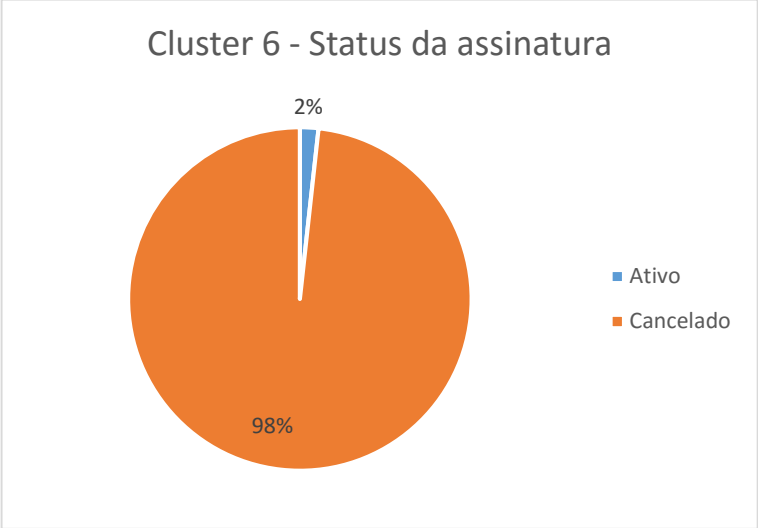


Figura 54: Distribuição de caixas recebidas do cluster 6.

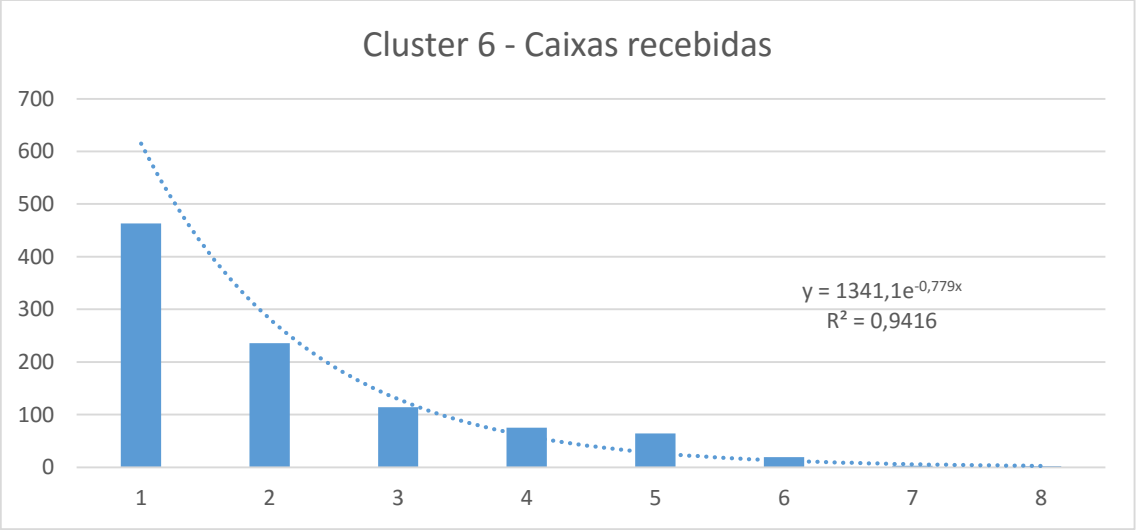


Figura 55: Distribuição de snacks recebidos do cluster 6.

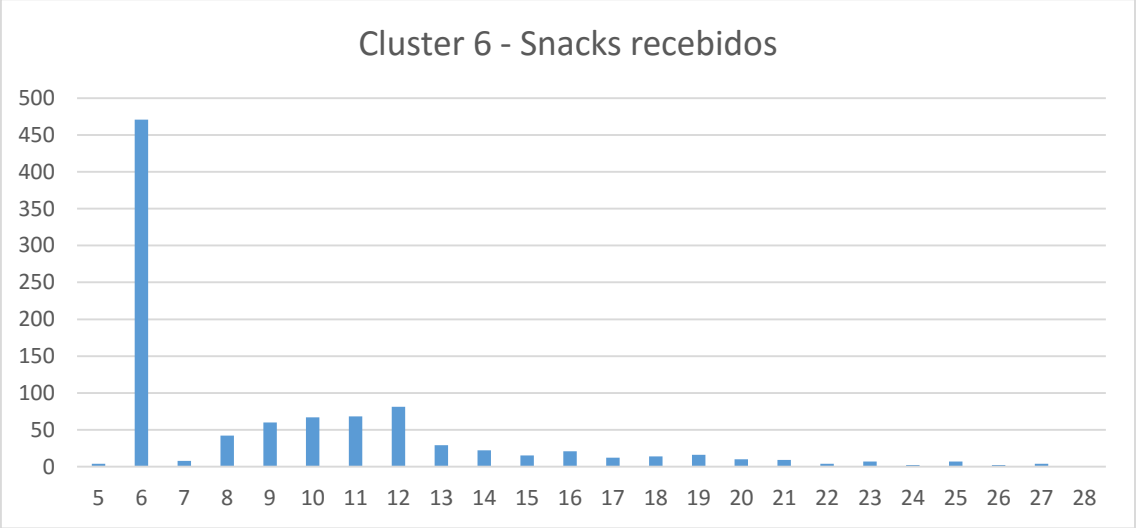


Figura 56: Distribuição da data de criação da assinatura do cluster 6.

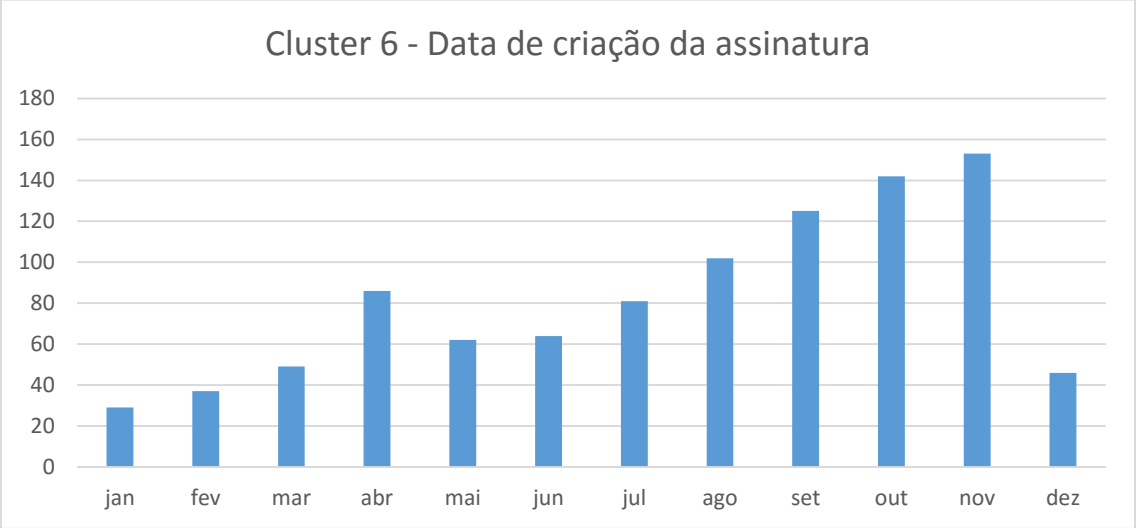
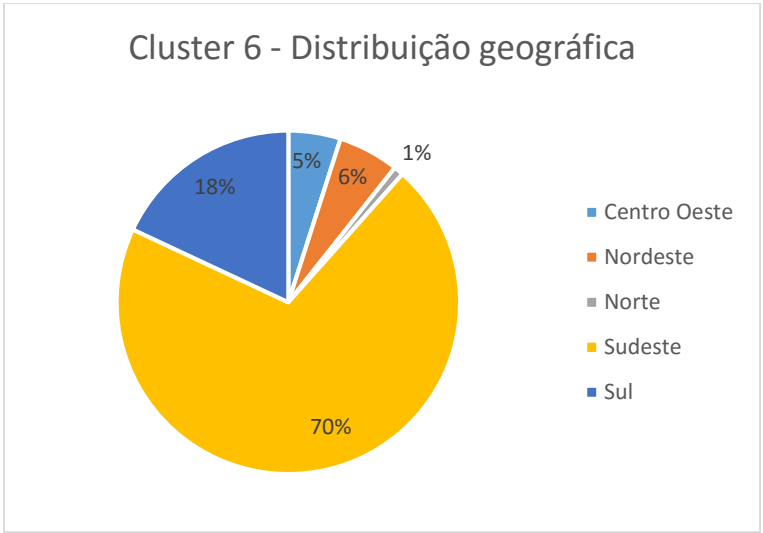


Figura 57: Distribuição geográfica do cluster 6.



5.5.Validação qualitativa dos segmentos

Após a confirmação quantitativa e matemática dos clusters obtidos, é necessário realizar a transição destes elementos para o conceito de segmento de mercado. Parte desta etapa foi realizada através do detalhamento e interpretação dos resultados no capítulo anterior, porém, utilizando os critérios de Blythe (2005) e Kotler & Armstrong (2015), é possível confirmar de fato se a clusterização resultou em uma segmentação válida. Abaixo segue a avaliação:

- Mensuráveis: Sim, através da associação dos clientes aos seus respectivos clusters, é possível mensurar suas propriedades.
- Acessíveis: Sim, por meio das ferramentas de segmentação das plataformas digitais, é possível fazer a definição dos segmentos de uma forma precisa. Além disso, as opções de rastreio auxiliam na validação.
- Substanciais: Sim, os segmentos de clientes clusterizados foram relativamente grandes para investimento de ações, sugerindo que as campanhas do mercado sigam a mesma proporção em escala maior. Nos casos em que o segmento não é atraente, é possível realizar o pensamento oposto: a exclusão e o não investimento.
- Diferenciáveis: Sim, a análise quantitativa revelou que cada cluster apresenta características distintas que justificam seu agrupamento (apesar da análise de dissimilaridade revelar que esta diferença não é tão robusta).
- Acionáveis: Sim, conforme será detalhado no posicionamento estratégico, as novas ações apresentam uma complexidade compatível com o já praticado atualmente pela empresa.

Por fim, pode-se dizer que o resultado obtido é de alta qualidade, por conta de sua afirmação tanto em termos matemáticos como cluster, quanto como segmentos para análise qualitativa.

Sendo assim, é possível então trabalhar em cima da caracterização dos clusters/segmentos para verificar quais as melhores ações a serem tomadas em cada um deles.

5.6.Posicionamento estratégico de Marketing dos segmentos

Mediante as informações dos segmentos e sua validação, é possível então realizar o posicionamento estratégico de Marketing em relação a eles; ou seja, definir quais seriam as ações para cada segmento.

Antes disso, utilizando as funções de caixas recebidas em função do mês e as informações sobre os planos, é possível realizar uma simulação para comparar os segmentos cancelados em termos de geração de receita. Neste caso, foi considerada uma base inicial de 1000 assinantes para cada segmento e um intervalo de tempo de 4 meses, por conta da limitação de escopo do segmento 5. A Tabela 8 contém o resultado desta análise e, logo abaixo, seguem as principais ações que poderiam ser trabalhadas em cada caso.

Tabela 8: Simulação de receita para os segmentos de assinantes cancelados.

Base de clientes por mês	**Segmento 1**	**Segmento 2**	**Segmento 4**	**Segmento 5**	**Segmento 6**
1	1000	1000	1000	1000	1000
2	688	848	744	447	459
3	506	696	553	199	211
4	377	543	411	89	97
Assinaturas cobradas total	2571	3087	2708	1735	1766
Receita por assinatura cobrada	R$ 99,90	R$ 99,90	R$ 79,50	R$ 79,50	R$ 99,90
Receita total	R$ 256.889	R$ 308.389	R$ 215.325	R$ 137.951	R$ 176.427

O segmento 1 trata das ex-assinantes que provaram o produto após um estímulo com o cupom de desconto, mas que não obteve satisfação suficiente para manter a assinatura. Dado que são clientes que já experimentaram o produto e tiveram uma aceitação parcial dele, é possível planejar ações de reativação. Em termos de campanhas de aquisição, como a empresa já possui as informações destes clientes, faz sentido investir em mídias mais baratas, como por exemplo o Email. O conceito das campanhas deve ser diferenciado, o qual poderia atacar o motivo do cancelamento. Poderia também ser oferecido um plano com preço menor, estimulando a facilidade em manter a assinatura.

O segmento 2 apresenta uma boa semelhança com o cluster 3, com a exceção de que os seus assinantes já efetuaram o cancelamento, e com o cluster 1, em termos experiência com o produto. Dentre os segmentos de cancelados, é aquele com maior potencial de geração de receita (Tabela 8). Portanto, as ações de reativação, uso de mídias mais baratas e oferecimento de produtos diferenciado fazem sentido neste caso também. Contudo, como a aceitação deste segmento é maior, a prioridade de tais campanhas deveria de maior neste grupo do que do segmento 1.

O segmento 3 são os clientes ativos da base e, portanto, são aqueles que enxergam valor no produto e trazem maior retorno para a empresa. Três frentes podem partir da análise deste cluster. A primeira é a prospecção de mais clientes semelhantes ao perfil deles. Para tanto, além das informações obtidas pelos clusters, pode-se investir em uma análise mais profunda sobre os perfis destes assinantes, de modo a melhorar ainda mais a eficiência das campanhas. A segunda opção é o desenvolvimento dos demais canais. Como visto na Figura 37, a maior parte das aquisições são através dos anúncios no Facebook e uma pequena parcela é atrelada a anúncios do Google e Orgânico. Dado que Orgânico não é uma mídia paga, fica então a possibilidade de melhorar as campanhas através do canal do Google. A terceira frente é na retenção da carteira. Como explicado anteriormente no Capítulo 2.1, o crescimento da empresa depende também da diminuição dos cancelados. Dado que este segmento é o mais rentável, é esperado que a prioridade de tais ações seja alta para ele.

O segmento 4, cuja principal característica é a presença de assinantes do gênero masculino, não é abordado no público alvo da empresa e a análise de cluster confirma a motivação para esta estratégia. Apresentam uma média de caixas recebidas baixa, um ticket médio baixo (plano com menos *snacks*) e representam apenas 7,8% da base. Portanto, a estratégia recomendada para este cluster está na exclusão deste segmento nas campanhas de aquisição. Além disso, é importante notar que todas aquisições pelos canais de Afiliados estão presentes neste segmento. Sendo assim, a outra ação válida é reavaliar este canal e definir qual seria o orçamento ideal ou se a empresa deve continuar com o mesmo. Uma terceira frente possível seria a estratégia de *upsell*, que consiste em estimular este segmento a trocar o plano de 10 *snacks*, menor receita, para o de 18 *snacks*, maior receita. Isso se justifica pela simulação realizada na Tabela 8, em que a taxa de cancelamento do segmento 4 é a segunda mais baixa.

O segmento 5 segue a mesma lógica do segmento 4, só que o público é feminino e é adquirido principalmente pelas campanhas no Facebook (98% do cluster). Por ser o segmento com menor potencial de receita (Tabela 8), a melhor estratégia é justamente a exclusão do segmento nas campanhas.

O segmento 6 é caracterizado pelas jovens assinantes que se mostraram interessadas pelo produto principalmente pelas ações da marca e recomendação de terceiros. Apesar de não ser um segmento atraente em termos de rentabilidade, é bastante engajado na procura de informações, característica percebida pelo canal de aquisição de anúncios do Google. Aliada à sua facilidade em uso da tecnologia, as campanhas de promoção da marca (*branding*) tornam-se bastante atraentes para este público. A principal função deste segmento seria na amplificação das campanhas de marca, de forma a gerar o interesse no produto para os demais clusters. Aliada a esta estratégia, é preciso realizar também a exclusão deste segmento nas campanhas de venda direta, pois é um segmento com baixo potencial de geração de receita (Tabela 8).

O Quadro 6 sintetiza a análise dos segmentos obtidos e o seu respectivo posicionamento sugerido pelo autor.

Quadro 6: Segmentos e sugestão de posicionamento estratégico de Marketing.

Segmento/Cluster	Principais características	Estratégia de Marketing
1 – Experimentadoras	Mulheres que provaram o produto após um estímulo com o cupom de desconto, mas que não obteve satisfação suficiente para manter a assinatura.	• Campanhas de reativação da base de clientes. • Impactar público através de mídias mais baratas (Email). • Desenvolver produto diferenciado.
2 – Quase assinantes	Perfil semelhante ao segmento 3, porém, a empresa falhou em conquistar tais clientes. Experiência com o produto parecida com o segmento 1, porém, a aceitação é maior.	• Campanhas de reativação da base de clientes. • Impactar público através de mídias mais baratas (Email). • Desenvolver produto diferenciado.
3 – Mina de Ouro	Clientes mais rentáveis e fiéis. Público mais maduro, estabilidade financeira suficiente para manter a assinatura sem o uso de cupom de desconto.	• Aquisição de novos clientes com mesmo perfil. • Desenvolver outros canais além do Facebook. • Campanhas de retenção de clientes.
4 – Vaidosos	Público masculino e jovem, pouco considerado na elaboração do produto e das campanhas em geral.	• Exclusão deste segmento nas campanhas. • Reavaliação do canal de Afiliados. • Estimular o *upsell*.
5 – Caçadoras de Descontos	Público feminino com principal interesse em experimentar o produto gastando o mínimo possível.	• Exclusão deste segmento nas campanhas.
6 – #BestBerry	Moças muito jovens interessadas no produto. Influenciadas pela marca e avaliação de terceiros.	• Campanhas de *branding* de modo a transmitir o interesse para os demais públicos. • Exclusão deste segmento nas campanhas de venda direta.

5.7.Avaliação dos gestores

Após a elaboração dos segmentos e das estratégias de Marketing, o resultado do trabalho foi apresentado para os gestores da empresa. Participaram desta conversa os 2 sócios fundadores da empresa e uma funcionária da área de Performance.

A clusterização obtida foi bastante elogiada, por conta do trabalho de categorização e também da validação dos segmentos, permitindo assim que a empresa consiga basear suas ações neste modelo.

Um ponto questionado pelos avaliadores foi a fraca separação entre os segmentos, que pode ser notado pelos *medoids*. Sugeriu-se que um estudo com outras variáveis de natureza comportamental fosse abordado, justamente para tentar refinar mais a distinção entre os clusters e auxiliar na definição dos conceitos a serem trabalhados nas campanhas.

Sobre as iniciativas propostas, as estratégias foram bem aceitas e se mostraram apropriadas para cada segmento. Duas iniciativas apresentaram pouca adesão da equipe avaliadora. A primeira delas foi a referente ao segmento 6 sobre as campanhas com foco em branding. O problema relatado foi na avaliação do impacto desta iniciativa, cujas métricas e benefícios não são muito claros. A segunda iniciativa foi a da estratégia de *upsell*, por conta da já existente dificuldade em atender o público masculino do cluster 4.

Finalmente, a simulação de receita por segmento, apesar de não ter sido o foco deste trabalho, recebeu uma avaliação positiva pelo fato de ser um resultado a ser discutido para elaboração de frentes para outras áreas da empresa.

6. CONCLUSÃO

O presente trabalho tinha como objetivo a segmentação da base de clientes de modo a identificar melhorias na otimização dos recursos de Marketing, aumentando assim a rentabilidade da empresa. Através do modelo elaborado, pode-se dizer que tal objetivo foi cumprido: foram obtidos os segmentos de mercado bem definidos e a proposição de estratégias de Marketing para cada um deles.

Os 6 segmentos resultantes da análise de clusters mostraram-se bastante representativos, através das validações quantitativa (clusterização) e qualitativa (segmentação). Pode-se notar que a empresa trabalha com diversos perfis de clientes e que uma abordagem mais segmentada faz total sentido em ser empregada. Dentre as ações propostas, são várias as alternativas consideradas: desde a simples diferenciação da oferta até o não investimento no segmento.

Sobre o tema de segmentação de mercado, este trabalho destaca a importância do uso das ferramentas quantitativas para a realização de análises. Conforme citado por Moutinho & Meidan (2004), o uso destes métodos em Marketing é recente, por conta da complexidade dos fenômenos e de sua mensuração. Contudo o avanço tecnológico do rastreamento para obtenção de dados e dos métodos computacionais torna mais acessível o emprego destas ferramentas, e a tendência é que seja cada vez mais incorporado na rotina dos profissionais da área.

Sob a perspectiva da análise de clusters, este estudo pode servir como uma aplicação dos métodos. Através da revisão bibliográfica, percebe-se que esta área de estudo também precisa ser avançada, principalmente na questão de consolidação da literatura. Assim como visto por Donilcar (2003), isso afeta os estudos práticos da análise de clusters para a segmentação de mercado, sendo que em muitos casos eles carecem de maior transparência na seleção do algoritmo de clusterização e da validação dos resultados, dois pontos atacados por este presente estudo.

Para a empresa, o estudo atinge o proposto em identificação dos segmentos e proposta de melhorias. Conforme visto nos resultados, a análise prévia dos gestores valida o estudo, viabilizando assim muito das frentes sugeridas. Os próximos passos seriam o detalhamento de cada frente validada, a definição dos planos de ação e a implementação propriamente dita. Além disso, o estudo em si levanta novas questões a serem avaliadas, como por exemplo um maior refinamento dos segmentos obtidos; o que ressalta ainda mais a importância deste trabalho para a empresa.

Por fim, o trabalho em si fornece base para outros estudos. Apesar do escopo estar limitado ao Marketing, os resultados obtidos podem ser empregados em áreas como CRM, retenção de clientes, análise do comportamento do consumidor, desenvolvimento de produto, entre outros. Na própria área de clusterização, o modelo pode ser evoluído, melhorando ainda mais a robustez e fornecendo mais análises. Algoritmos mais complexos e outras formas de dissimilaridade poderiam sem empregadas futuramente. O estudo interno dos clusters, valendo-se de variáveis mais específicas é também um alternativa atraente.

Um ponto essencial para o desenvolvimento do trabalho é o apoio da gestão, que identificou junto com o autor e o professor orientador, a oportunidade de explorar o estudo, e também no total suporte de fornecimento de dados. Através desta colaboração, foi possível desenvolver um modelo robusto e uma análise enriquecedora, com alto potencial de implementação.

REFERÊNCIAS BIBLIOGRÁFICAS

BEST BERRY. Site da empresa. Disponível em: <http://bestberry.com.br/>. Último acesso em 28/05/2017 às 19:45.

BIVOLINO. Site da empresa. Disponível em: <http://www.bivolino.com/en/default.html>. Último acesso em 11/06/2017 às 18:24.

BLYTHE, J. Essential of marketing. 3ª ed. Pearson Prentice Hall, 2005.

BRITO, P.; SOARES, C.; ALMEIDA, S.; MONTE, A.; BYVOET, M. Customer segmentation in a large database of an online customized fashion business. Robotics and Computer-Integrated Manufacturing 36 (2015) 93-100.

CRAN. Package 'cluster'. Version 2.0.6. March 16, 2017. Disponível em: <https://cran.r-project.org/web/packages/cluster/cluster.pdf >. Último acesso em 28/05/2017 às 16:40.

DONILCAR, S. Using cluster analysis for market segmentation - typical misconceptions, established methodological weaknesses and some recommendations for improvement. Australasian Journal of Market Research, 2003, 11(2), 5-12.

EVANS, M. Marketing segmentation. In: BARKER, M. The Marketing Book. Butterworth-Heinemann, 2003. Cap. 10, p. 246-284.

HSU, F.; LU, L.; LIN, C. Segmenting customers by transaction data with concept hierarchy, Expert Systems with Applications 39 (6) (2012) 6221–6228.

IAB BRASIL. Número de Investimento 2016. Disponível em: <http://iabbrasil.net/assets/upload/boas_praticas/1457447232.pdf>. Último acesso em 29/11/2016 às 11:35.

KAUFMAN, L.; ROUSSEEUW, P. Finding groups in data – An introduction to cluster analysis. New York. John Wiley & Sons, 1990.

KOTLER, P.; ARMSTRONG, G. Princípios de Marketing. 15ª ed. São Paulo: Pearson Prentice Hall, 2015.

LIAO, S.; CHEN, Y.; LIN, Y. Mining customer knowledge to implement online shopping and home delivery for hypermarkets. Expert Systems with Applications 38 (2011) 3982-3991.

MAIMON, O; ROKACH, L. Data Mining and Knowledge Discovery Handbook. In:_____. Clustering Methods. Springer US, 2005. Cap. 15, p. 197-245.

MOUTINHO, L; MEIDAN, A. Quantitative methods in marketing. In: BARKER, M. The Marketing Book. Butterworth-Heinemann, 2003. Cap. 9, p. 246-284.

PUNJ, G.; STEWART, D. Cluster Analysis in Marketing Research: Review and Suggestions for Application. Journal of Marketing Research, Vol. 20, No. 2 (May, 1983), pp. 134-148.

TAN, P.; STEINBACH, M.; KUMAR, V. Introduction to Data Mining. 1ª ed. Pearson Prentice Hall, 2005.

SMITH, W. Product Differentiation and Market Segmentation as Alternative Marketing Strategies. Journal of Marketing, Vol. 21, No. 1 (Jul, 1956), pp. 3-8.

ANEXO A: MÉTODO K-MEDOIDS E O ALGORITMO PAM

Conforme indicado por Kaufmann & Rousseeuw (1990), o algoritmo PAM (*Partitioning Around Medoids*) é baseado na procura dos k objetos representativos entre os objetos de um conjunto de dados. Tais representantes são chamados de *medoids* pelos autores, mas podem também ser encontrados como *centrotypes* na literatura. Após encontrar os *medoids*, os k clusters são construídos pela associação de cada objeto do conjunto de dados ao objeto representativo mais próximo. Em termos matemáticos, pode-se dizer que o PAM tem como objetivo minimizar a soma das dissimilaridades.

Por conta desta definição, pode-se considerar que o PAM é um tipo de método *K-medoids* (que é também chamado de *K-median*).

O algoritmo PAM é dividido em 2 fases (Kaufmann & Rousseeuw, 1990). A primeira delas é a BUILD, que procura obter uma clusterização inicial através da seleção sucessiva dos medoids até que k objetos sejam encontrados. O primeiro objeto é aquele cuja soma de dissimilaridade entre todos os demais objetos seja o menor possível. O próximo objeto é aquele que diminui a função objeto o máximo possível, o qual é encontrado da seguinte maneira:

- Considere um objeto i que ainda não foi selecionado.
- Considere um objeto não selecionado j e calcule a diferença entre sua dissimilaridade D_j com o objeto previamente selecionado mais semelhante e sua dissimilaridade $d(i,j)$ com o objeto i.
- Se esta diferença for positiva, objeto j vai contribuir com a decisão de selecionar o objeto i. Sendo assim, calcula-se

$$C_{ji} = \max\left(D_j - d(i,j), 0\right)$$

- Calcula-se o ganho total obtido selecionando-se o objeto i:

$$\sum_j C_{ji}$$

- Escolhe-se o objeto i ainda não selecionado que:

$$\underset{i}{\text{maximizes}} \sum_j C_{ji}$$

- Continuar processo até que k objetos sejam encontrados.

A segunda etapa é chamada de SWAP, cujo objetivo é melhorar os *medoids* e, consequentemente, a clusterização. Isto é realizado considerando todos os pares de objetos (i, h) em que o objeto i foi selecionado e o objeto h ainda não foi. É determinado qual o efeito obtido na clusterização quando a troca é realizada, ou seja, quando o objeto i não é mais um objeto representativo, mas o objeto h sim. Tal efeito é dado pela soma de dissimilaridades entre cada objeto e o objeto representativo mais próximo.

Kaufmann & Rousseeuw (1990) enuncia os seguintes passos para calcular o efeito de troca entre i e h:

1) Considere um objeto não selecionado j e calcule sua contribuição C_{jih} para a troca:
   a) Se j é mais distante de i e h do que de algum outro objeto representativo, C_{jih} é zero.
   b) Se j não é mais distante de i do que qualquer outro objeto representativo selecionado ($d(i, j) = D_j$), duas situações devem ser consideradas:

   b1) j é mais próximo de h do que o segundo objeto representativo mais próximo

   $$d(j, h) < E_j$$

   onde E_j é a dissimilaridade entre j e o segundo objeto representativo mais próximo. Neste caso, a contribuição do objeto j para a troca entre os objetos i e h é

   $$C_{jih} = d(j, h) - d(j, i)$$

   b2) j é tão distante de h quanto é do segundo objeto representativo mais próximo

   $$d(j, h) \geq E_j$$

   Neste caso, a contribuição do objeto j para a troca é

   $$C_{jih} = E_j - D_j$$

   Deve ser observado que na situação b1 a contribuição C_{jih} pode ser tanto positiva quanto negativa, dependendo da posição relativa dos objetos j, h e i. Somente se o objeto j for mais próximo do objeto i do que h é que a contribuição é positiva, o que indica que a troca não é favorável do ponto de vista do objeto j. Por outro lado, na situação b2 a contribuição será sempre positiva porque não

pode ser vantajoso substituir i por um objeto h mais distante de j do que o segundo objeto representativo mais próximo.

c) j é mais distante do objeto i do que de pelo menos um dos outros objetos representativos mas mais próximo de h do que qualquer objeto representativo. Neste caso, a contribuição de j para a troca é

$$C_{jih} = d(j,h) - D_j$$

2) Calcular o resultado total da troca adicionando as contribuições C_{jih}

$$T_{ih} = \sum_j C_{jih}$$

Os próximos passos definem se é necessário continuar a troca

3) Selecione o par (i,h) que

$$\underset{i,h}{\text{minimizes}}\, T_{ih}$$

4) Se o mínimo T_{ih} é negativo, a troca é feita e o algoritmo retorna ao passo 1. Se o mínimo T_{ih} é positivo ou zero, o valor do objetivo não pode diminuir através da troca e o algoritmo para.

Note que todas as potenciais trocas são consideradas e que os resultados do algoritmo não dependem da ordem dos objetos no arquivo de entrada (exceto no caso em que algumas distâncias entre objetos estão ligadas).

ANEXO B: COMANDOS NO SOFTWARE R

Este anexo contém os comandos empregados para a realização da análise de clusters no software estatístico R. Vale ressaltar que é necessário instalar o pacote “cluster” para conseguir rodar os dados.

Conforme mostrado no Anexo C, os comandos foram empregados para obter as variações de cluster para k de 2 a 20. Para tornar este documento mais simples, será mostrado apenas a lista de comandos efetuada para obter os resultados de $k = 6$.

```
dados = read.csv("D:/Shibata/Documents/~KEVIN/~POLI/dados.csv", TRUE, ";");
dis <- daisy(dados);
pam6 <- pam(dis, 6);
pam6 $ silinfo $ avg.width;
pam6 $ silinfo $ clus.avg.widths;
pam6 $ clusinfo;
pam6 $ objective;
```

ANEXO C: RESULTADO DAS CLUSTERIZAÇÕES

Este anexo apresenta os resultados das demais clusterizações aplicada em função do parâmetro k

Resultados $k = 2$

```
> pam2 $ silinfo $ clus.avg.widths; pam2 $ silinfo $ avg.width
[1] 0.3328268 0.4772428
[1] 0.4071706
```

```
> pam2 $ objective
    build      swap
0.1470192 0.1447566
```

```
> pam2 $ clusinfo
     size  max_diss   av_diss  diameter separation
[1,] 4823 0.6608062 0.1594223 0.7780061        0.1
[2,] 5117 0.5002463 0.1309335 0.7543183        0.1
```

Resultados $k = 3$

```
> pam3 $ silinfo $ clus.avg.widths; pam3 $ silinfo $ avg.width
[1] 0.2013683 0.6036904 0.4930274
[1] 0.3902025
```

```
> pam3 $ objective
    build      swap
0.1199681 0.1189170
```

```
> pam3 $ clusinfo
     size  max_diss    av_diss  diameter separation
[1,] 4802 0.5605534 0.15770726 0.6884750 0.02604803
[2,] 3420 0.3596521 0.07646172 0.5431869 0.10000000
[3,] 1718 0.5689117 0.09500906 0.6974369 0.02604803
```

Resultados $k = 4$

```
> pam4 $ silinfo $ clus.avg.widths; pam4 $ silinfo $ avg.width
[1] 0.3250142 0.5703156 0.5138274 0.5817954
[1] 0.462123
```

```
> pam4 $ objective
     build       swap
0.09878848 0.09773241
```

```
> pam4 $ clusinfo
     size  max_diss    av_diss  diameter separation
[1,] 4030 0.3528583 0.11606001 0.5895822 0.02604803
[2,] 3418 0.1886830 0.07629946 0.2973645 0.10000000
[3,] 1699 0.3509120 0.09062085 0.5082445 0.02604803
[4,]  793 0.3740047 0.11220919 0.6153798 0.03364755
```

Resultados $k = 5$

```
> pam5 $ silinfo $ clus.avg.widths; pam5 $ silinfo $ avg.width
[1] 0.3999069 0.5387234 0.5150502 0.5750624 0.5747974
[1] 0.4984446
```

```
> pam5 $ objective
     build       swap
0.08799121 0.08668245
```

```
> pam5 $ clusinfo
     size  max_diss    av_diss  diameter separation
[1,] 3056 0.3537565 0.09459399 0.4714524 0.03364755
[2,] 3418 0.1886830 0.07629946 0.2973645 0.10000000
[3,] 1697 0.3509120 0.09040207 0.4555976 0.10000000
[4,]  793 0.3740047 0.11220919 0.6153798 0.03364755
[5,]  976 0.2175186 0.07106406 0.3502125 0.10169492
```

Resultados $k = 6$

```
> pam6 $ silinfo $ clus.avg.widths; pam6 $ silinfo $ avg.width
[1] 0.5152433 0.4919348 0.5150502 0.5079214 0.5251320 0.5264851
[1] 0.5084901
```

```
> pam6 $ objective
     build       swap
0.07936462 0.07815097
```

```
> pam6 $ clusinfo
     size  max_diss    av_diss  diameter separation
[1,] 2310 0.1932855 0.06847157 0.3083214  0.1000000
[2,] 3418 0.1886830 0.07629946 0.2973645  0.1000000
[3,] 1697 0.3509120 0.09040207 0.4555976  0.1000000
[4,]  774 0.3261063 0.10729943 0.5394966  0.1048433
[5,]  765 0.2701582 0.06802507 0.3586846  0.1002849
[6,]  976 0.2175186 0.07106406 0.3502125  0.1016949
```

Resultados $k = 7$

```
> pam7 $ silinfo $ clus.avg.widths; pam7 $ silinfo $ avg.width
[1] 0.4871091 0.5190634 0.4721693 0.5079010 0.5225098 0.5242986 0.2800945
[1] 0.4852916
```

```
> pam7 $ objective
     build       swap
0.07361220 0.07287291
```

```
> pam7 $ clusinfo
     size  max_diss    av_diss  diameter  separation
[1,] 2251 0.1932855 0.06611289 0.3083214 0.003959631
[2,] 2710 0.1847937 0.05807151 0.2705780 0.001979816
[3,] 1697 0.3509120 0.09040207 0.4555976 0.100000000
[4,]  774 0.3261063 0.10729943 0.5394966 0.104843305
[5,]  765 0.2701582 0.06802507 0.3586846 0.100284900
[6,]  976 0.2175186 0.07106406 0.3502125 0.101694915
[7,]  767 0.1806102 0.07862200 0.3437043 0.001979816
```

Resultados $k = 8$

```
> pam8 $ silinfo $ clus.avg.widths; pam8 $ silinfo $ avg.width
[1] 0.4508945 0.2967560 0.4423470 0.5078340 0.5196013 0.5226618 0.2934700 0.3174624
[1] 0.4144112
```

```
> pam8 $ objective
     build       swap
0.06988678 0.06921247
```

```
> pam8 $ clusinfo
     size  max_diss    av_diss  diameter  separation
[1,] 2226 0.1932855 0.06516779 0.3083214 0.003689217
[2,] 1509 0.1742203 0.04960529 0.2517370 0.001424501
[3,] 1697 0.3509120 0.09040207 0.4555976 0.100000000
[4,]  774 0.3261063 0.10729943 0.5394966 0.104843305
[5,]  765 0.2701582 0.06802507 0.3586846 0.100284900
[6,]  976 0.2175186 0.07106406 0.3502125 0.101694915
[7,]  679 0.1753337 0.07029696 0.3195836 0.003689217
[8,] 1314 0.1645510 0.04753626 0.2414223 0.001424501
```

Resultados $k = 9$

```
> pam9 $ silinfo $ clus.avg.widths; pam9 $ silinfo $ avg.width
[1] 0.5492107 0.2875394 0.4386340 0.5047679 0.4895775 0.4743262 0.4039333 0.2992916 0.3105848
[1] 0.4235823
```

```
> pam9 $ objective
     build       swap
0.06661427 0.06604239
```

```
> pam9 $ clusinfo
      size  max_diss    av_diss  diameter   separation
 [1,] 1905 0.1830581 0.05296141 0.2783372 0.0048433048
 [2,] 1509 0.1742203 0.04960529 0.2517370 0.0014245014
 [3,] 1697 0.3509120 0.09040207 0.4555976 0.1000000000
 [4,]  774 0.3261063 0.10729943 0.5394966 0.1048433048
 [5,]  741 0.2710947 0.06482716 0.3586846 0.0037037037
 [6,]  934 0.2175186 0.06793324 0.3502125 0.0005698006
 [7,]  595 0.1653556 0.06120099 0.2537544 0.0051137187
 [8,]  471 0.1704180 0.07195251 0.3211016 0.0005698006
 [9,] 1314 0.1645510 0.04753626 0.2414223 0.0014245014
```

Resultados $k = 10$

```
> pam10 $ silinfo $ clus.avg.widths; pam10 $ silinfo $ avg.width
[1] 0.5492107 0.2875394 0.4386340 0.4561580 0.4893301 0.4743262 0.4039333 0.7426076 0.2992916
0.3105848
[1] 0.4261187
```

```
> pam10 $ objective
     build       swap
0.06376701 0.06302939
```

```
> pam10 $ clusinfo
      size  max_diss    av_diss  diameter   separation
 [1,] 1905 0.1830581 0.05296141 0.2783372 0.0048433048
 [2,] 1509 0.1742203 0.04960529 0.2517370 0.0014245014
 [3,] 1697 0.3509120 0.09040207 0.4555976 0.1000000000
 [4,]  554 0.3062243 0.08090812 0.4580879 0.1000000000
 [5,]  741 0.2710947 0.06482716 0.3586846 0.0037037037
 [6,]  934 0.2175186 0.06793324 0.3502125 0.0005698006
 [7,]  595 0.1653556 0.06120099 0.2537544 0.0051137187
 [8,]  220 0.1569194 0.03762462 0.2719901 0.1000000000
 [9,]  471 0.1704180 0.07195251 0.3211016 0.0005698006
[10,] 1314 0.1645510 0.04753626 0.2414223 0.0014245014
```

Resultados $k = 11$

```
> pam11 $ silinfo $ clus.avg.widths; pam11 $ silinfo $ avg.width
[1] 0.5492107 0.2875394 0.2312976 0.4561580 0.4893301 0.4742615 0.2975983 0.4039333 0.7426076
0.2992916 0.3105848
[1] 0.3967116
```

```
> pam11 $ objective
     build       swap
0.06148342 0.06019059
```

```
> pam11 $ clusinfo
      size  max_diss    av_diss  diameter   separation
 [1,] 1905 0.1830581 0.05296141 0.2783372 0.0048433048
 [2,] 1509 0.1742203 0.04960529 0.2517370 0.0014245014
 [3,]  798 0.2979788 0.07808714 0.3932635 0.0034188034
 [4,]  554 0.3062243 0.08090812 0.4580879 0.1000000000
 [5,]  741 0.2710947 0.06482716 0.3586846 0.0037037037
 [6,]  934 0.2175186 0.06793324 0.3502125 0.0005698006
 [7,]  899 0.2288815 0.06994556 0.3970979 0.0034188034
 [8,]  595 0.1653556 0.06120099 0.2537544 0.0051137187
 [9,]  220 0.1569194 0.03762462 0.2719901 0.1000000000
[10,]  471 0.1704180 0.07195251 0.3211016 0.0005698006
[11,] 1314 0.1645510 0.04753626 0.2414223 0.0014245014
```

Resultados $k = 12$

```
> pam12 $ silinfo $ clus.avg.widths; pam12 $ silinfo $ avg.width
[1] 0.5444425 0.3586839 0.2312976 0.4561580 0.4877913 0.4731907 0.2969662 0.5229513 0.4269039
0.7426076 0.2968863 0.3284405
[1] 0.4164326
```

```
> pam12 $ objective
     build       swap
0.05941584 0.05809715
```

```
> pam12 $ clusinfo
      size  max_diss    av_diss  diameter   separation
 [1,] 1894 0.1830581 0.05239593 0.2783372 0.0048433048
 [2,] 1407 0.1715807 0.04217479 0.2544842 0.0008547009
 [3,]  798 0.2979788 0.07808714 0.3932635 0.0034188034
 [4,]  554 0.3062243 0.08090812 0.4580879 0.1000000000
 [5,]  741 0.2710947 0.06482716 0.3586846 0.0037037037
 [6,]  934 0.2175186 0.06793324 0.3502125 0.0005698006
 [7,]  899 0.2288815 0.06994556 0.3970979 0.0034188034
 [8,]  523 0.1599143 0.05151363 0.2537544 0.0045294316
 [9,]  272 0.1648269 0.05686819 0.3154588 0.0045294316
[10,]  220 0.1569194 0.03762462 0.2719901 0.1000000000
[11,]  471 0.1704180 0.07195251 0.3211016 0.0005698006
[12,] 1227 0.1499596 0.04305361 0.2324741 0.0008547009
```

Resultados $k = 13$

```
> pam13 $ silinfo $ clus.avg.widths; pam13 $ silinfo $ avg.width
[1] 0.3947059 0.3590647 0.2312976 0.4561580 0.4400695 0.4314080 0.2969675 0.5232217 0.4124486
0.7426076 0.2958592 0.3286021 0.3850650
[1] 0.3790343
```

```
> pam13 $ objective
     build       swap
0.05737560 0.05566833
```

```
> pam13 $ clusinfo
      size  max_diss    av_diss  diameter   separation
 [1,]  619 0.1757435 0.04231951 0.2298835 0.0005698006
 [2,] 1407 0.1715807 0.04217479 0.2544842 0.0008547009
 [3,]  798 0.2979788 0.07808714 0.3932635 0.0034188034
 [4,]  554 0.3062243 0.08090812 0.4580879 0.1000000000
 [5,]  731 0.2710947 0.06379763 0.3586846 0.0076778212
 [6,]  937 0.2175186 0.06856001 0.3502125 0.0014245014
 [7,]  899 0.2288815 0.06994556 0.3970979 0.0034188034
 [8,]  523 0.1599143 0.05151363 0.2537544 0.0045294316
 [9,]  275 0.1711820 0.05809953 0.3218138 0.0045294316
[10,]  220 0.1569194 0.03762462 0.2719901 0.1000000000
[11,]  444 0.1677145 0.06616835 0.3253082 0.0014245014
[12,] 1227 0.1499596 0.04305361 0.2324741 0.0008547009
[13,] 1306 0.1933814 0.04097321 0.2489984 0.0005698006
```

Resultados $k = 14$

```
> pam14 $ silinfo $ clus.avg.widths; pam14 $ silinfo $ avg.width
[1] 0.3947059 0.3590647 0.2688858 0.4561580 0.4400695 0.3545006 0.4312981 0.5232217 0.3228130
0.4124486 0.7426076 0.2958592 0.3286021 0.3850650
[1] 0.3858581
```

```
> pam14 $ objective
     build       swap
0.05546795 0.05382891
```

```
> pam14 $ clusinfo
      size  max_diss    av_diss  diameter   separation
 [1,]  619 0.1757435 0.04231951 0.2298835 0.0005698006
 [2,] 1407 0.1715807 0.04217479 0.2544842 0.0008547009
 [3,]  677 0.2812814 0.06455366 0.3803122 0.0034188034
 [4,]  554 0.3062243 0.08090812 0.4580879 0.1000000000
 [5,]  731 0.2710947 0.06379763 0.3586846 0.0076778212
 [6,]  258 0.2346871 0.07443551 0.3798590 0.0170650442
 [7,]  937 0.2175186 0.06856001 0.3502125 0.0014245014
 [8,]  523 0.1599143 0.05151363 0.2537544 0.0045294316
 [9,]  762 0.2288815 0.05774749 0.3970979 0.0034188034
[10,]  275 0.1711820 0.05809953 0.3218138 0.0045294316
[11,]  220 0.1569194 0.03762462 0.2719901 0.1000000000
[12,]  444 0.1677145 0.06616835 0.3253082 0.0014245014
[13,] 1227 0.1499596 0.04305361 0.2324741 0.0008547009
[14,] 1306 0.1933814 0.04097321 0.2489984 0.0005698006
```

Resultados $k = 15$

```
> pam15 $ silinfo $ clus.avg.widths; pam15 $ silinfo $ avg.width
[1] 0.4092910 0.3588281 0.2688858 0.4561580 0.4317401 0.3545006 0.5183764 0.5230222 0.4294478
0.3228130 0.4204903 0.7426076 0.3999233 0.3284686 0.3710361
[1] 0.3956835
```

```
> pam15 $ objective
     build       swap
0.05365580 0.05202006
```

```
> pam15 $ clusinfo
      size  max_diss    av_diss  diameter   separation
 [1,]  616 0.1393224 0.04174391 0.2133017 0.0005698006
 [2,] 1407 0.1715807 0.04217479 0.2544842 0.0008547009
 [3,]  677 0.2812814 0.06455366 0.3803122 0.0034188034
 [4,]  554 0.3062243 0.08090812 0.4580879 0.1000000000
 [5,]  735 0.2710947 0.06417914 0.3586846 0.0025641026
 [6,]  258 0.2346871 0.07443551 0.3798590 0.0170650442
 [7,]  743 0.1924415 0.05060856 0.3318771 0.0136141636
 [8,]  523 0.1599143 0.05151363 0.2537544 0.0045294316
 [9,]  233 0.1709112 0.05765964 0.2599367 0.0136141636
[10,]  762 0.2288815 0.05774749 0.3970979 0.0034188034
[11,]  273 0.1530939 0.05725071 0.2904776 0.0045294316
[12,]  220 0.1569194 0.03762462 0.2719901 0.1000000000
[13,]  395 0.1911750 0.05957048 0.3317298 0.0025641026
[14,] 1227 0.1499596 0.04305361 0.2324741 0.0008547009
[15,] 1317 0.1671968 0.04166835 0.2634527 0.0005698006
```

Resultados $k = 16$

```
> pam16 $ silinfo $ clus.avg.widths; pam16 $ silinfo $ avg.width
[1] 0.4154641 0.3766681 0.2688858 0.4561580 0.4306456 0.3545006 0.5181528 0.5778726 0.4294478
0.3226886 0.5674857 0.5360287 0.7426076 0.4061689 0.3661335 0.3684559
[1] 0.4114227
```

```
> pam16 $ objective
     build       swap
0.05220435 0.05060473
```

```
> pam16 $ clusinfo
      size  max_diss    av_diss  diameter   separation
 [1,]  613 0.1127481 0.04128549 0.2040648 0.0005698006
 [2,] 1357 0.1495209 0.03883387 0.2327239 0.0008547009
 [3,]  677 0.2812814 0.06455366 0.3803122 0.0034188034
 [4,]  554 0.3062243 0.08090812 0.4580879 0.1000000000
 [5,]  735 0.2710947 0.06417914 0.3586846 0.0025641026
 [6,]  258 0.2346871 0.07443551 0.3798590 0.0170650442
 [7,]  743 0.1924415 0.05060856 0.3318771 0.0136141636
 [8,]  500 0.1293768 0.04771908 0.2248774 0.0106430606
 [9,]  233 0.1709112 0.05765964 0.2599367 0.0136141636
[10,]  762 0.2288815 0.05774749 0.3970979 0.0034188034
[11,]  169 0.1782731 0.04752770 0.2308375 0.0188237688
[12,]  242 0.1530939 0.04860288 0.2660689 0.0053841325
[13,]  220 0.1569194 0.03762462 0.2719901 0.1000000000
[14,]  392 0.1911750 0.05872164 0.3317298 0.0025641026
[15,] 1168 0.1440064 0.03872937 0.2217963 0.0008547009
[16,] 1317 0.1671968 0.04166835 0.2634527 0.0005698006
```

Resultados $k = 17$

```
> pam17 $ silinfo $ clus.avg.widths; pam17 $ silinfo $ avg.width
[1] 0.3918142 0.3766564 0.2688858 0.4561580 0.4153875 0.3545006 0.5153051 0.5777943 0.4294478
0.3226886 0.5746021 0.5401895 0.7426076 0.4261075 0.3661145 0.3373024 0.1782657
[1] 0.3940078
```

```
> pam17 $ objective
     build       swap
0.05107936 0.04943419
```

```
> pam17 $ clusinfo
      size  max_diss    av_diss  diameter   separation
 [1,]  554 0.1127481 0.03993177 0.1988645 0.0011396011
 [2,] 1357 0.1495209 0.03883387 0.2327239 0.0008547009
 [3,]  677 0.2812814 0.06455366 0.3803122 0.0034188034
 [4,]  554 0.3062243 0.08090812 0.4580879 0.1000000000
 [5,]  735 0.2710947 0.06417914 0.3586846 0.0025641026
 [6,]  258 0.2346871 0.07443551 0.3798590 0.0170650442
 [7,]  743 0.1924415 0.05060856 0.3318771 0.0136141636
 [8,]  500 0.1293768 0.04771908 0.2248774 0.0106430606
 [9,]  233 0.1709112 0.05765964 0.2599367 0.0136141636
[10,]  762 0.2288815 0.05774749 0.3970979 0.0034188034
[11,]  168 0.1566493 0.04674945 0.2174614 0.0188237688
[12,]  241 0.1530939 0.04820014 0.2660689 0.0079627215
[13,]  220 0.1569194 0.03762462 0.2719901 0.1000000000
[14,]  374 0.1645106 0.05418044 0.2972800 0.0025641026
[15,] 1168 0.1440064 0.03872937 0.2217963 0.0008547009
[16,]  711 0.1685692 0.03095072 0.2120282 0.0011396011
[17,]  685 0.1737261 0.04014780 0.2400299 0.0019943020
```

Resultados $k = 18$

```
> pam18 $ silinfo $ clus.avg.widths; pam18 $ silinfo $ avg.width
[1] 0.3918142 0.3766564 0.2077717 0.4561580 0.4153875 0.3716100 0.5152374 0.2387775 0.5777943
0.4294478 0.5746021 0.5401895 0.7426076 0.4261075 0.3661145 0.3369000 0.3373024 0.1782657
[1] 0.3902539
```

```
> pam18 $ objective
     build       swap
0.04999662 0.04848156
```

```
> pam18 $ clusinfo
      size  max_diss    av_diss  diameter   separation
 [1,]  554 0.1127481 0.03993177 0.1988645 0.0011396011
 [2,] 1357 0.1495209 0.03883387 0.2327239 0.0008547009
 [3,]  420 0.1999148 0.05674907 0.3549620 0.0059684195
 [4,]  554 0.3062243 0.08090812 0.4580879 0.1000000000
 [5,]  735 0.2710947 0.06417914 0.3586846 0.0025641026
 [6,]  242 0.2542959 0.06946987 0.3796493 0.0105471741
 [7,]  743 0.1924415 0.05060856 0.3318771 0.0136141636
 [8,]  446 0.2683768 0.05980463 0.3803122 0.0052299552
 [9,]  500 0.1293768 0.04771908 0.2248774 0.0106430606
[10,]  233 0.1709112 0.05765964 0.2599367 0.0136141636
[11,]  168 0.1566493 0.04674945 0.2174614 0.0188237688
[12,]  241 0.1530939 0.04820014 0.2660689 0.0079627215
[13,]  220 0.1569194 0.03762462 0.2719901 0.1000000000
[14,]  374 0.1645106 0.05418044 0.2972800 0.0025641026
[15,] 1168 0.1440064 0.03872937 0.2217963 0.0008547009
[16,]  589 0.1911612 0.05114160 0.3370622 0.0052299552
[17,]  711 0.1685692 0.03095072 0.2120282 0.0011396011
[18,]  685 0.1737261 0.04014780 0.2400299 0.0019943020
```

Resultados $k = 19$

```
> pam19 $ silinfo $ clus.avg.widths; pam19 $ silinfo $ avg.width
[1] 0.3918142 0.3132024 0.2077717 0.4561580 0.4153875 0.3716100 0.5152374 0.2387775 0.5644499
0.4294478  0.5622228  0.5285258  0.7426076  0.4261066  0.3238739  0.2292433  0.3369000  0.3373024
0.1782657
[1] 0.3703686
```

```
> pam19 $ objective
     build       swap
0.04898564 0.04753788
```

```
> pam19 $ clusinfo
      size  max_diss    av_diss  diameter  separation
 [1,]  554 0.1127481 0.03993177 0.1988645 0.001139601
 [2,] 1253 0.1476542 0.03734028 0.2272587 0.001139601
 [3,]  420 0.1999148 0.05674907 0.3549620 0.005968420
 [4,]  554 0.3062243 0.08090812 0.4580879 0.100000000
 [5,]  735 0.2710947 0.06417914 0.3586846 0.002564103
 [6,]  242 0.2542959 0.06946987 0.3796493 0.010547174
 [7,]  743 0.1924415 0.05060856 0.3318771 0.013614164
 [8,]  446 0.2683768 0.05980463 0.3803122 0.005229955
 [9,]  500 0.1252982 0.04771092 0.2250985 0.010643061
[10,]  233 0.1709112 0.05765964 0.2599367 0.013614164
[11,]  168 0.1566493 0.04674945 0.2174614 0.018823769
[12,]  241 0.1530939 0.04820014 0.2660689 0.007962722
[13,]  220 0.1569194 0.03762462 0.2719901 0.100000000
[14,]  374 0.1645106 0.05418044 0.2972800 0.002564103
[15,]  725 0.1504632 0.03066665 0.1939371 0.001139601
[16,]  547 0.1334616 0.03571621 0.2091817 0.002549616
[17,]  589 0.1911612 0.05114160 0.3370622 0.005229955
[18,]  711 0.1685692 0.03095072 0.2120282 0.001139601
[19,]  685 0.1737261 0.04014780 0.2400299 0.001994302
```

Resultados $k = 20$

```
> pam20 $ silinfo $ clus.avg.widths; pam20 $ silinfo $ avg.width
[1] 0.3918142 0.3132024 0.2077717 0.4321795 0.4152929 0.3716100 0.5152374 0.2387775 0.5644499
0.4294478 0.5622228 0.1793920 0.5285258 0.7220832
[15] 0.4261066 0.3238739 0.2292433 0.3369000 0.3373024 0.1782657
[1] 0.36262
```

```
> pam20 $ objective
     build       swap
0.04803732 0.04663282
```

```
> pam20 $ clusinfo
      size  max_diss    av_diss  diameter  separation
 [1,]  554 0.1127481 0.03993177 0.1988645 0.001139601
 [2,] 1253 0.1476542 0.03734028 0.2272587 0.001139601
 [3,]  420 0.1999148 0.05674907 0.3549620 0.005968420
 [4,]  320 0.1823189 0.05451406 0.3430989 0.006741720
 [5,]  735 0.2710947 0.06417914 0.3586846 0.002564103
 [6,]  242 0.2542959 0.06946987 0.3796493 0.010547174
 [7,]  743 0.1924415 0.05060856 0.3318771 0.013614164
 [8,]  446 0.2683768 0.05980463 0.3803122 0.005229955
 [9,]  500 0.1252982 0.04771092 0.2250985 0.010643061
[10,]  233 0.1709112 0.05765964 0.2599367 0.013614164
[11,]  168 0.1566493 0.04674945 0.2174614 0.018823769
[12,]  234 0.2877320 0.07855650 0.4086550 0.006741720
[13,]  241 0.1530939 0.04820014 0.2660689 0.007962722
[14,]  220 0.1569194 0.03762462 0.2719901 0.100000000
[15,]  374 0.1645106 0.05418044 0.2972800 0.002564103
[16,]  725 0.1504632 0.03066665 0.1939371 0.001139601
[17,]  547 0.1334616 0.03571621 0.2091817 0.002549616
[18,]  589 0.1911612 0.05114160 0.3370622 0.005229955
[19,]  711 0.1685692 0.03095072 0.2120282 0.001139601
[20,]  685 0.1737261 0.04014780 0.2400299 0.001994302
```