Aplicação de técnicas multivariadas na gestão do volume de cerveja vendido

Luigi das Chagas Dondoni – dondoni.luigi@gmail.com

Prof. Michel José Anzanello, Ph.D – anzanello@producao.ufrgs.br

Resumo

A principal métrica de performance do mercado de cerveja é o volume vendido da bebida. Todavia, em virtude da maginitude desse segmento no Braisl, a compreensão desse indicador se demonstra um desafio, tanto pelo montante de dados disponíveis para análise, quanto pelos fatores externos associados à venda de cerveja. Este artigo propõe a utilização de duas técnicas de análise multivariada (TAM) com vistas a interpretar o impacto de um conjunto de variáveis independentes sobre a venda de cerveja no mercado de Porto Alegre. A base de pontos de venda (PDVs), tipicamente formada por elevado número de observações, foi agrupada através de técnicas de clusterização com base nos perfis dos PDVs. Em seguida, modelos de regressão linear múltipla foram gerados para cada um dos agrupamentos. Os modelos apresentaram boa capacidade preditiva uma vez que métricas como mean square error ficaram entre 0,004 e 0,016 e o R^2 ajustado entre 0,86 e 0,97. Percebeu-se que, para determinado grupo de clientes, é mais vantajoso conceder maior prazo de pagamento do que promocionar o preço dos produtos, enquanto que, para outro cluster, o verão tem impacto negativo nas vendas devido à movimentação da população local para áreas litorâneas.

Palavras-chave: *venda de cerveja*; *clusterização*; *regressão linear múltipla*; *k-means*; *interpreção de coeficientes*.

1. Introdução

A crescente concorrência no mercado nacional demanda das empresas a estruturação de um sistema de gestão que permeie e alinhe todos os níveis da empresa (estratégico, tático e operacional) (Muller,2003). No mercado de cervejas brasileiro, responsável pela terceira maior produção mundial (cerca 14,1 bilhões de litros anuais), um dos principais indicadores de desempenho é o volume de cerveja vendido; todavia, por ser tratar de um mercado de larga escala com cerca de 1,2 milhões de pontos de venda, a compreensão das variáveis que impactam o volume de cerveja vendido (por exemplo, preço do produto, temperatura e índice pluviométrico diário, bem como elasticidade de cada parâmetro) torna-se uma tarefa complexa em virtude do volume de dados disponíveis.

A não compreensão dos parâmetros e suas correlações com a variável de resposta (volume de cerveja vendido) é um problema estratégico que dificulta o planejamento da empresa para atingir seus objetivos (Falconi, 2014). No contexto da cerveja, verificam-se dois grandes problemas. O primeiro tem cunho logístico e consiste na ruptura de estoque devido às oscilações na demanda que não foram previstas, resultando na indisponibilidade do produto ao cliente e, consequentemente, no decréscimo dos indicadores de receita, volume de vendas e *market share*. O segundo problema decorre de descontos ineficientes que objetivam incrementar o volume de vendas, o qual pode não ser atingido em virtude de se considerar apenas um único fator associado à variável de resposta (ocasionando, assim, perda de faturamento). É nesse cenário que as técnicas de análise multivariadas (TAM) se tornam úteis para melhor compreender as correlações entre as variáveis e seus impactos sobre a variável de resposta (volume vendido), uma vez que viabilizam a obtenção de conhecimento gerencial a partir de base de dados extensas (Hair Jr et al., 2010), sintetizando problemas complexos (Rencher, 2002)

Este artigo propõe a aplicação de duas técnicas de TAM, clusterização e regressão linear múltipla, com vistas a melhor gerir o volume de cerveja vendido pelo Centro de Distribuição situado em Porto Alegre visando ao incremento de vendas. Para tanto, os clientes foram inicialmente agrupados com base em variáveis que incluem número de cervejas vendidas no estabelecimento, prazo de compra e volume de compra. A ideia foi gerar agrupamentos de clientes com perfis similares e potencial de reação equivalente em termos de vendas quando incentivados por determinada estratégia. As técnicas de clusterização aqui utilizadas reduzem a complexidade do problema a ser analisado (Taboada et al., 2007), pois agrupam amostras que apresentem características semelhantes em *clusters*; assim, cada *cluster* representa um grupo de clientes que apresentam características comuns entre si (Anzanello et al., 2014). Na sequência, gerou-se, para cada *cluster* de clientes, um modelo de regressão linear múltipla para compreender como cada variável independente impacta no volume de cerveja vendido em cada segmento de cliente. Ademais, os modelos preditivos também utilizaram variáveis como, volume de cerveja mensal médio dos últimos 3 meses, precipitação diária, preço por litro de cerveja, prazo de pagamento além de variáveis categóricas para indicar a estação do ano (verão, inverno e outono/primavera).

Objetiva-se, com as análises propostas, compreender as correlações existentes entre a variável dependente (volume de cerveja) e as variáveis independentes para cada *cluster*

de clientes através da comparação dos resultados obtidos entre os grupos. A partir disso, a empresa poderá planejar de forma mais eficiente os recursos associados à venda de cerveja com base no comportamento de cada segmento de cliente, tanto no âmbito comercial quanto no âmbito da cadeia produtiva.

Este artigo, além da introdução, estrutura-se da seguinte forma. A segunda seção corresponde a uma revisão teórica acerca dos tópicos abordados. A terceira seção apresentará a metodologia utilizada na análise dos dados e, por fim, a quarta seção discutirá os resultados e as conclusões obtidas com o presente trabalho.

2. Referencial Teórico

Esta seção objetiva apresentar e contextualizar os assuntos relativos ao escopo deste artigo. Primeiramente, apresenta-se como ferramentas de Análise Multivariada estão sendo utilizadas no segmento de alimentos. Em seguida serão detalhadas as duas ferramentas utilizadas neste projeto: a Clusterização e a Regressão Linear Múltipla.

2.1 Ferramentas Multivariadas no segmento de alimentos:

As ferramentas de Análise Multivariada estão presentes nos mais diversos contextos de pesquisas. Rencher (2002) afirma que o caráter exploratório da TAM, isto é, auxiliar através de modelos matemáticos a compreensão da relação causa-efeito dos fenômenos pesquisados, potencializa a sua presença em diferentes campos de pesquisas tais como educação, biologia e negócios. No que se refere ao segmento de alimentos, percebe-se um aumento na publicação de artigos que envolvam ferramentas de análise multivariada, os quais apresentam dois grandes focos: Análises de qualidade e Análise de segurança.

No campo da análise de segurança tanto técnicas qualitativas quanto técnicas quantitativas estão sendo pesquisadas. Segundo Callao e Ruisánchez (2018), as técnicas qualitativas configuram-se como boa opção para problemas de fraude alimentar que não possam ser solucionados com apenas uma variável. Todavia, Callao e Ruisánchez (2018) afirmam que as técnicas de cunho qualitativo ainda não possuem consenso claro no que diz respeito às etapas de validação, sendo este ainda dominado pelas técnicas quantitativas. As técnicas quantitativas para análise de segurança são úteis para detectar casos nos quais as empresas tentam baratear seus produtos se utilizando de insumos mais baratos que, no paladar do cliente, não apresentam diferenças significativas no sabor do produto; contudo, modificando suas características nutricionais. Por exemplo, Sikorska et al., (2018) utilizou análises de componentes principais (PCA) para então aplicar

técnicas de classificação e regressão em conjunto com análises espectrais de sucos de maçã para determinar a quais categorias estes deveriam pertencer.

No que diz respeito aos estudos de análise de qualidade no segmento de alimentos, as técnicas multivariadas quantitativas são as mais utilizadas. A metodologia destes estudos é semelhante à aplicada nas análises de segurança, porém o objetivo é predizer se a qualidade do produto final será atendida conforme os insumos utilizados na produção. Para tanto, técnicas como classificação de produtos e regressão também são utilizadas nesse campo de pesquisa. Zervos e Albert (1992) classificaram amostras de produtos através do método *nearest neighbor* utilizando como parâmetros iniciais *outputs* da análise PCA, em seguida através de regressões lineares múltiplas determinaram o impacto de cada uma das variáveis independentes adotadas na variável de resposta (*off-flavors*).

Por outro lado, algumas pesquisas estão mais voltadas para a compreensão do comportamento dos clientes e suas preferências de consumo como mostra o estudo realizado por Wang et al. (2013) o qual utiliza regressões logísticas para entender preferência de compras dos clientes no mercado de alimentos. Além disso, alguns poucos estudos abordam temas mais gerenciais dentro do segmento de alimentos, utilizando técnicas multivariadas para obter informação gerencial no segmento de alimentos e assim melhor estruturar as etapas da cadeia estudada. Caniato et al. (2005) utilizou modelos de clusterização para segmentar clientes e assim desenvolver modelos de previsão de demanda com objetivo de adequar os modelos preditivos à variabilidade existente no processo de vendas. O estudo concluiu que o uso de métodos de clusterização, como o *k-means*, associados a modelos de regressão linear múltipla apresentam vantagens para previsão da demanda em cenários nos quais há fatores de variabilidade tais como a sazonalidade na demanda e a oferta de descontos no preço do produto estudado, auxiliando a programação do setor logístico do caso estudado. Todavia, não foram encontrados na literatura estudos que utilizassem técnicas multivariadas no contexto de inteligência comercial.

2.2 Clusterização

Os métodos de clusterização tem como principal objetivo segmentar uma base de dados em conjuntos denominados *clusters*, de forma a maximizar tanto a homogeneidade das observações dentro de um mesmo conjunto quanto a heterogeneidade de *clusters* distintos (James et al., 2017). Usualmente, a Distância Euclidiana é utilizada como métrica de semelhança ou discrepância entre as observações da base de dados em um espaço com *N*

dimensões, sendo *N* o número de variáveis utilizadas para caracterizar as observações da base de dados (Hair Jr et al., 2010). O grande benefício da aplicação de técnicas de clusterização é o auxílio na tomada de decisões, uma vez que decompõem uma grande base de dados em *clusters*, fragmentando o problema sem perda significativa de informação (Taboada et al., 2007).

Dentre os algoritmos de clusterização se destacam dois grandes grupos: os hierárquicos e os não hierárquicos (Anzanello et al., 2014). Algoritmos hierárquicos resultam em uma representação gráfica dos *clusters*-, denominada de dendrograma a partir da qual é possível determinar o melhor número de agrupamentos para o problema que está sendo analisado (Rencher, 2002). A obtenção do dendrograma decorre do agrupamento das observações mais semelhantes sendo normalmente utilizada como métrica de semelhança a Distância Euclidiana, assim, inicia-se com cada observação (n) representando um *cluster*, a partir de então as observações mais semelhantes são agrupadas gerando n-1 clusters. O algoritmo continua agrupando os clusters mais similares até que se obtenha apenas um único *cluster* (James et al., 2017). De tal forma, os *clusters* são obtidos através da combinação de *clusters* imediatamente abaixo, de maneira que, cada *cluster* é uma observação e o *cluster* final resulta do agrupamento de todos os demais níveis (Hastie et al., 2008).

No que diz respeito aos métodos não hierárquicos, destaca-se o *k-means*, reconhecido como o método mais popular entre os não hierárquicos (Anzanello et al., 2011). O primeiro passo desse algoritmo é determinar o número *k* de grupos antes de qualquer outra etapa (Hastie et al.,2008). Essa decisão pode ser embasada em uma prévia classificação hierárquica pela interpretação do dendrograma resultante (Hair Jr et al., 2010), ou com base na opinião qualitativa de especialistas valendo-se de características dos dados que auxiliem a predefinição do número de *clusters* (Moraes, 2017). Por fim, ressalta-se que a escolha do número de *clusters* e, por conseguinte, dos centroides, dado que cada *cluster* deve possuir um único centroide, é um fator determinante para o resultado dos agrupamentos e deve ser revisitado pelo pesquisador (Rencher, 2002; Hastie et al., 2008; James et al., 2017).

Após a definição do número de *clusters*, a operacionalização do algoritmo *k-means* pode ser resumido em três macro etapas (Zhao et al., 2018). Primeiramente é preciso determinar os centroides iniciais de cada um dos *clusters*; recomenda-se que os centroides apresentem a maior distância possível entre si, dado que o posicionamento dos centroides

influencia o resultado final da clusterização (Rencher, 2002). Em seguida, cada observação é associada ao centroide mais próximo, conforme o racional de similaridade escolhido pelo pesquisador. Por fim, recalcula-se o centroide de cada *cluster*. As macro etapas devem ser revisadas até que a segunda e a terceira etapa convirjam (Zhao et al., 2018). Ademais, conforme (Rencher, 2002; Hastie et al., 2008; James et al., 2017; Zhao et al, 2018) duas propriedades devem ser satisfeitas terminada a aplicação do algoritmo:

i. Todas observações devem pertencer a algum *cluster*.
ii. Cada observação deve pertencer somente a um único *cluster*.

Por fim, é necessário aferir a qualidade da clusterização obtida, uma vez que condições iniciais podem ser definidas de maneira intuitiva. Deste modo, a qualidade da clusterização pode ser mensurada por meio do *Silhouette Index* (*SI*) métrica que apresenta valores entre -1 e +1, sendo -1 uma segmentação ineficiente dos dados e +1 uma segmentação ideal. Essa métrica apresenta o nível de semelhança de cada observação frente ao *cluster* ao qual ela foi atribuída, sendo o *Silhouette Index* estimado pela equação (1):

$$SI_j = \frac{b_j - a_j}{máx\{b_j, a_j\}} \quad (1)$$

Onde:

$a_j =$ Distância média entre a observação j e as demais observações do *cluster* e

$b_j =$ Distância média entre a observação j e as demais observações do *cluster* mais próximo.

Zotteri (2005) ressalta a importância do nível de agregação das informações para a *performance* de modelos preditivos de demanda, uma vez que seu estudo apresentou melhora nos resultados após clusterizar os dados. Neto (2013) exemplificou isto ao obter incrementos na qualidade de predição de faturamento de lojas no varejo quando comparado à dados não clusterizados.

2.3 Regressão Linear Múltipla

Após segmentar a base de clientes em *clusters* é necessário compreender como cada variável preditora (entendida como variável independente ou *x*) influencia na variável de resposta (variável dependente ou *y*) para cada um dos segmentos de clientes. Quando o

problema analisado envolve apenas uma variável independente para predição de uma variável de resposta, utiliza-se a técnica estatística de regressão linear simples, já para problemas que envolvam duas ou mais variáveis independentes é utilizada a regressão linear múltipla (Hair Jr et al., 2010). Assim, a regressão linear múltipla equaciona a relação entre todas as variáveis independentes e a variável dependente a partir de uma base de dados relativos ao fenômeno estudado (Shu et al., 2011). A partir da combinação linear das variáveis independentes e seus respectivos pesos é possível representar a variável dependente e estimar seu valor (Rencher, 2002).

Para construir um modelo de regressão linear múltipla atribui-se a cada variável independe um coeficiente. Esses coeficientes, comumente chamados de pesos, denotam a contribuição relativa da variável independente na predição da variável de resposta (Rencher, 2002). O objetivo do modelo de regressão é aproximar as distâncias entre as observações do fenômeno estudado com o plano ou reta, dependendo do número de variáveis adotadas para o problema, oriundo da regressão. Ademais, o modelo genérico de uma regressão linear múltipla é apresentado através da equação (2):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \cdots + \beta_n x_n + \varepsilon \quad (2)$$

Onde:

x_n representa a n-ésima variável independente selecionada para o modelo;

β_n representa o coeficiente referente à n-ésima variável independente;

ε representa o erro associado ao modelo regressivo.

Após a construção do modelo de regressão é preciso verificar a qualidade das predições retornadas pelo modelo, ou seja, o quão próximas estão as predições dos dados observados. Dado que $\boldsymbol{y}$ e $\boldsymbol{x}$ são conhecidos, (Rencher, 2002) recomenda o uso do método dos mínimos quadrados, o qual tem por objetivo minimizar a soma dos quadrados dos erros (SEE) de todo o conjunto de observações comparados aos valores retornados pelo modelo de regressão; ver equação (3) (Hair Jr et al.,2010):

$$SSE = \sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0 + \sum_{j=1}^{q} \hat{\beta}_j x_{ij})\right)^2 \quad (3)$$

Na qual, $\widehat{\boldsymbol{\beta}}$ é dada por (4):

$$\widehat{\boldsymbol{\beta}}' = (X'X)^{-1}X'y \quad (4)$$

É necessário mensurar a qualidade da predição resultante do modelo regressivo, ou seja, o quão aderente são os valores preditos quando comparados com as observações. Para tal, utiliza-se o coeficiente de determinação denominado de R^2, obtido através do quociente da soma dos resíduos da diferença (SQR) pela soma dos quadrados totais (SQT), como representado em (5), onde SQE representa a soma dos quadrados dos resíduos do conjunto de variações (Hair Jr et al.,2010):

$$R^2 = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT} \quad (5)$$

A métrica acima pode apresentar valores entre zero e um, sendo zero um indicador de baixa aderência do modelo aos dados observados e $R^2 = 1$ representa a total aderência do modelo aos dados; todavia, ressalta-se que resultados muito próximos ou iguais a um caracterizam o chamado *overfitting*, ou seja, quando o modelo se adequada perfeitamente aos dados observados, porém não apresenta boa capacidade preditiva para dados futuros Hair Jr et al., 2010).

Estudos ressaltam à importância de técnicas de regressão como modelos de demanda no gerenciamento de preços, para Kunz (2014) o modelo regressivo é a peça fundamental para um bom gerenciamento preços. Enquanto que Sassi (2011) e Rencher (2002) destacam que o modelo de regressão apresenta o efeito avulso de cada uma das variáveis independentes no resultado final da predição, sendo esta uma técnica eficiente para analises de elasticidades dos parâmetros em cada *cluster* de cliente proporcionando maior embasamento para tomada de decisões no que diz respeito ao atingimento das metas de volume de cerveja vendido.

3. Método

O presente artigo tem como objetivo gerenciar o volume de cerveja vendido a partir da compreensão do impacto de um conjunto de variáveis independentes (sendo elas: preço por litro, precipitação, prazo de pagament e uma variável categórica indicando a estação do ano) no volume de cerveja vendido (variável dependente) em cada um dos segmentos de clientes. De posse do entendimento de tais relações, pretende-se conceber diferentes estratégias para incrementar os volumes vendidos em cada segmento. Para tanto foram realizadas quatro etapas: (i) coleta e tratamento dos dados para clusterização da base de clientes em segmentos distintos, (ii) clusterização da base de clientes, (iii) geração de modelos regressivos para cada cluster, e (iv) concepção de estratégias alternativas para aumentar o volume de cerveja vendido. Tais etapas são agora detalhadas.

Etapa 1 - Coleta e tratamento dos dados para clusterização de clientes

Primeiramente serão extraídos do banco de dados da empresa em análise dados referentes à base de clientes da cidade de Porto Alegre, Rio Grande do Sul, uma vez que é a região com maior representatividade no faturamento da empresa no estado. Estes dados possuem caráter exclusivamente quantitativo e serão utilizados para agrupar a base de clientes em diferentes clusters com base em variáveis que incluem volume de cerveja vendido, prazo de pagamento e número médio de cervejas da empresa que o cliente trabalha (arranjo dos dados ilustrados na Tabela 1):

Tabela 1 – Arranjo dos dados para clusterização dos clientes

Cliente	X1 (Volume)	X2 (Prazo)	X3 (SKUs)
Cliente 1	Valor numérico *a*	Valor numérico *b*	Valor numérico *c*
Cliente 2	Valor numérico *d*	Valor numérico *e*	Valor numérico *f*
⋮	⋮	⋮	⋮
Cliente *n*	Valor numérico *g*	Valor numérico *h*	Valor numérico *i*

Na sequência, os dados serão normalizados para evitar que diferentes magnitudes das variáveis interfiram na clusterização dos clientes, dado que os métodos de agrupamento utilizam métricas de distância para segmentar os dados (por exemplo, o prazo, que é tipicamente dado em dias, não seja dominado pela variável volume, dada em hectolitros). Por fim, será analisada a presença de dados espúrios, ou seja, possíveis *outliers* que possam prejudicar a consistência do agrupamento.

Etapa 2 - Clusterização da base de clientes

Em seguida, são aplicados dois métodos de clusterização em vistas a agrupar os clientes, ambos utilizando as variáveis detalhadas na Tabela 1. Inicialmente, os clientes serão agrupados hierarquicamente gerando um dendrograma, o qual permite definir o número de *clusters* a ser gerado. Na sequência, através do método *k-means*, realiza-se o agrupamento utilizando o número de *clusters* sugerido pelo dendrograma. Ressalta-se que especialistas de mercado da empresa estarão envolvidos em ambas as etapas de clusterização a fim de validar qualitativamente os agrupamentos gerados. Adicionalmente, a qualidade dos agrupamentos gerados será avaliada através da métrica *Silhouette Index*.

Etapa 3 – Geração de modelos regressivos para cada cluster

Nesta etapa é feita uma nova coleta de dados do sistema da empresa, desta vez para gerar modelos regressivos com vistas à predição do volume de cerveja a ser comercializado

para cada um dos clusters gerados na etapa 2. Foram utilizados dados de caráter quantitativo, tais como volume mensal médio dos últimos 3 meses, preço por litro de cerveja, prazo de pagamento e precipitação do dia e uma variável categórica indicando a estação do ano (verão,inverno e primavera/outono), os dados quantitativos foram tratados (normalizados) para que o modelo fosse contruído. O objetivo da regressão linear múltipla será identificar o impacto de cada uma das variáveis independentes citadas na variável dependente (volume de cerveja vendido). Para verificar a capacidade preditiva dos modelos regressivos gerados os dados são segmentados em dois grupos: grupo de treino, no qual o modelo é construído, e grupo de teste, no qual é avaliada a precisão das predições aos valores observados através de métricas como o R^2 na base de treino e do *mean square error* (*MSE)* na base de teste.

Etapa 4 - Estratégias para aumentar o volume de cerveja vendido

Finalizada a etapa de construção dos modelos regressivos, será feita uma análise das variáveis independentes que possuem maior impacto na variável de resposta em cada um dos clusters. Assim, será possível avaliar se determinada variável tem impacto semelhante em todos os clusters ou se há variáveis que impactam de forma heterogênea o volume de cerveja vendida em cada cluster. A partir da Tabela 2 (que compara os coeficientes entre os clusters), espera-se formular estratégias gerenciais para cada cluster utilizando as variáveis com potencial maior impacto sobre o volume de cerveja vendido no cluster.

Tabela 2 – Compilação dos coeficientes das regressões para fins de identificação das variáveis mais impactantes sobre o volume de cerveja comercializado.

Cluster	X1 (Volume)	X2 (Prazo)	X3 (Precipitação)	X4 (Verão)	X5 (Inverno)	X6 (Vol. Médio U3M)
Cluster 1	Coeficiente *C11*	Coeficiente *C12*	Coeficiente *C13*	Coeficiente *C14*	Coeficiente *C15*	Coeficiente *C16*
Cluster 2	Coeficiente *C21*	Coeficiente *C22*	Coeficiente *C23*	Coeficiente *C24*	Coeficiente *C25*	Coeficiente *C26*
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Cluster *n*	Coeficiente *Cn1*	Coeficiente *Cn2*	Coeficiente *Cn3*	Coeficiente *Cn4*	Coeficiente *Cn5*	Coeficiente *Cn6*

4 Resultados

A empresa estudada nesse artigo está presente em todo território nacional, fabricando e comercializando duas grandes famílias de produtos: cervejas e refrigerantes. No que diz respeito ao mercado de cervejas, foco deste estudo, apresenta uma grande variedade de marcas e embalagens, representando uma das maiores fatias do mercado nacional e um dos maiores volumes comercializados, em Porto Alegre, área de estudo deste artigo, a empresa atende cerca de 7000 clientes, faturando cerca de 1 milhão de reais por dia.

Primeiramente, será apresentada a *clusterização* dos clientes em suas duas etapas, hierárquica e não-hierárquica, posteriormente as métricas de desempenho dos modelos de regressão aplicadas a cada *cluster* de clientes e, por fim, os coeficientes serão analisados com o intuito de discutir estratégias alternativas para aumentar o volume de cerveja vendido.

4.1 *Clusterização*

Com o auxílio de especialistas da empresa, foram selecionados 3 parâmetros para segmentar os 6.626 pontos de venda (PDVs). Os parâmetros de classificação, bem como suas características são apresentados na Tabela 3.

Tabela 3 – Descrição dos parâmetros utilizados na *clusterização* da base de clientes

Parâmetro	Unidade	Descrição
Volume de cerveja vendido	Hectolitros (hl)	Volume médio dos últimos 3 meses do período estudados
Prazo de pagamento	Dias	Prazo médio de pagamento dos últimos 3 meses estudados
Número médio de cervejas	SKUs	# médio de SKUs distintos que o PDV comprou nos últimos 3 meses estudados

Os parâmetros foram normalizados para, então, avaliar o número adequado de *clusters* a ser utilizado no método *k-means* através de um dendrograma. Este foi construído através de técnicas de *clusterização* hierárquicas utilizando como racional de segmentação a distância euclidiana entre os PDVs. Através da análise do dendrograma (Figura 1), percebe-se que 3, 4 e 6 *clusters* aparecem como sugestões de número de clusters a serem gerados via algoritmo *k-means*, visto que apresentam o maior grau de separação visual entre as observações. Tais alternativas de número de clusters (*k*) foram testados via k-means; os valores de Silhouette Index são apresentados na Tabela 4 em ordem decrescente.

Figura 1 – Dendrograma da base de clientes

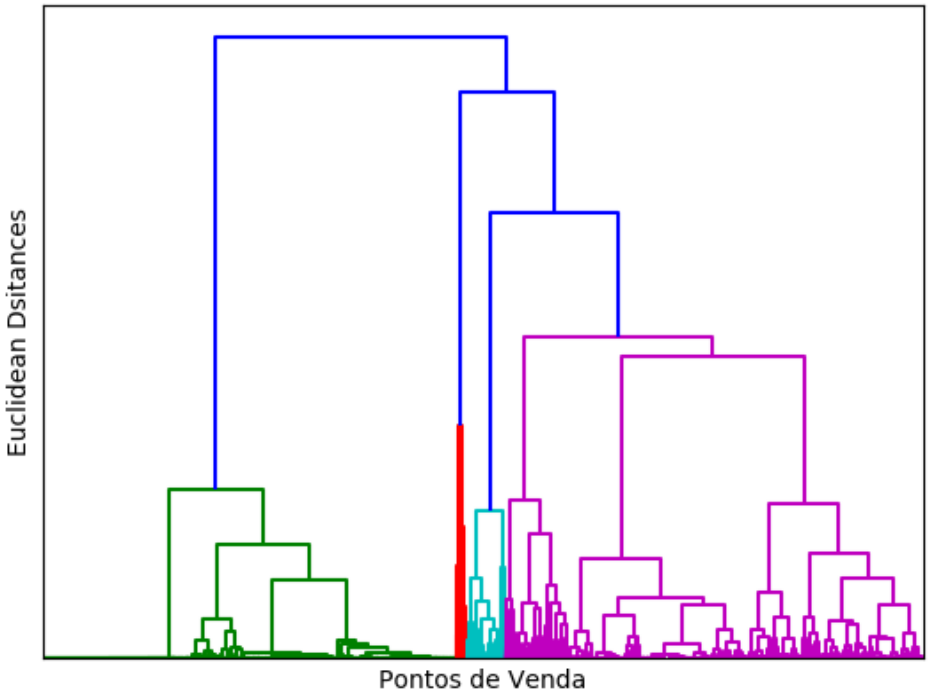

Tabela 4 – Cenários de *clusterização*

Cenário	Silhoutte Index (SI)
k-means com 4 Clusters	**0,4902**
k-means com 3 Clusters	**0,4605**
k-means com 6 Clusters	**0,4597**

Com base na Tabela 4, optou-se pelo cenário com 4 *clusters* não só pelo fato deste apresentar o maior valor de *SI*, mas também pela análise qualitativa realizada em conjunto com especialistas da empresa apontar tal número como o mais favorável na identificação de aspectos relevantes para segmentação dos clientes. Com base em uma análise qualitativa dos agrupamento formados, percebeu-se que o cenário com 4 agrupamentos dividiu a base de clientes em grandes redes, pequenas redes, clientes especiais (VIPs) e mercearias (usualmente chamadas de tradicionais), diferentemente do cenário com 3 *clusters* no qual clientes VIPs e pequenas redes configuram um único *cluster*. Isso se deve ao fato de clientes especiais, assim como pequenas redes, apresentarem um volume de cerveja vendido representativo bem como maiores prazos de pagamento; porém, tais clientes trabalham com uma variedade menor de *SKU*s quando comparado às pequenas

redes. Contudo, por mais que a diferença entre os *clusters* seja pequena, tanto o processo de venda quanto a experiência do consumidor final nos clientes em questão apresenta grande diferença o que, empiricamente, invalida a hipótese de agrupar clientes especiais e pequenas redes como um único *cluster*.

O cenário com 6 agrupamentos de clientes apresentou a mesma lógica que o cenário ótimo, mas requer a segmentação do *cluster* das mercearias e dos clientes especiais em PDVs grandes e pequenos no que diz respeito ao parâmetro volume de cerveja vendido. Esta divisão de clientes especiais e mercearias não apresenta relevância prática, segundo os especialistas, uma vez que tanto o processo de vendas quanto a experiência do consumidor final nesses tipos de PDVs são equivalentes, não havendo diferença entre um ponto de venda pequeno e um grande sob a perspectiva do consumidor. Ademais, foi destacado pelos especialistas da empresa o fato de esta segmentação não contemplar variáveis geográficas do PDV, pois por se tratar do mesmo tipo de comércio é necessário que se leve em consideração fatores externos na classificação dos clientes e não apenas dados internos como é o caso da presente análise. Por fim, a Tabela 5 apresenta, de modo resumido, a configuração final da *clusterização* da base de clientes de Porto Alegre.

Tabela 5 – Configuração final da clusterização da base de clientes de Porto Alegre

				Centróides		
SI	Cluster	Total PDVs	Tipo de PDVs	Volume de cerveja vendido	Prazo de pagamento	Número médio de cervejas
0,4902	1	1.832	Pequenas redes	0,14	0,13	1,02
	2	4.023	Tradicionais	-0,22	-0,46	-0,55
	3	81	Grandes Redes	7,29	2,25	2,38
	4	690	VIPs	0,04	0,69	0,26

Analisando os resultados em conjunto com os especialistas da empresa foi observado que, além do resultado satisfatório na métrica de agrupamento adotada (*SI*), os *clusters* originados retrataram com alto grau de relevância as peculiaridades do mercado de cervejas de Porto Alegre. Tal mercado apresenta grande número de mercearias de bairro com pouca representatividade em relação ao volume de cerveja vendido e ao número médio de cervejas trabalhadas, fato observado pelo posicionamento dos centróides do *cluster* em questão. Além disso, é possível observar a representativadidade do parâmetro prazo de pagamento apresentada tanto pelas grandes redes quanto para os clientes VIPs,

característica que reflete o comportamento das lojas desses *clusters* de buscarem trabalhar com ciclo financeiro pequeno e positivo.

4.2 Geração de modelos de regressão linear múltipla

Após a *clusterização* clientes em 4 segmentos, foram coletados dados para gerar regressões lineares múltiplas para cada um dos *clusters*. Para tanto, foi utilizado como intervalo de estudo o ano de 2018 com os dados agregados por dia. Após analisar o Gráfico 1, a qual apresenta o volume de cerveja vendido por mês, optou-se pela inserção de duas variáveis do tipo *dummy* para representar a estação do ano, assim sendo uma variável representa a estação verão e a outra a estação inverno (quando a estação da observação é outuno ou primavera as duas variáveis recebem valor nulo), uma vez que é possível inferir que se trata de um fenômeno sazonal.

Gráfico 1 – Volume de cerveja vendido por mês

Na sequência, as demais variáveis foram normalizadas para então gerar um modelo de regressão para cada *cluster*. A Tabela 6 apresenta estatísticas resultantes dos modelos para cada um dos *clusters*, na porção de treino – 75% das observações – foi mensurado o R^2 e o R^2 Ajustado e para a porção de teste – 25% das observações – foi utilizado o MSE (*mean square error*). Para tanto, utilizou-se uma base de dados com 200000 amostras acerca do fenômeno estudado.

Tabela 6 – Estatísticas da Regressão

Cluster	Tipo de PDVs	MSE	R^2	R^2 ADJ
1	Pequenas redes	0,007	0,973	0,971
2	Tradicionais	0,006	0,867	0,866
3	Grandes Redes	0,004	0,972	0,972
4	VIPs	0,016	0,939	0,935

Observando o *mean square error* resultante do ajuste da equação às observações em cada *cluster*, é possível inferir que todos os modelos apresentaram boa capacidade preditiva. O agrupamento das grandes redes apresentou o menor *MSE*; isso se deve ao fato dos PDVs dentro desse grupo serem mais semelhantes entre si no que diz respeito à experiência oferecida ao consumidor final, tornando as observações mais homogêneas entre os diferentes PDVs. Já o agrupamento dos clientes especiais apresentou o maior *MSE*, fato que também pode ser explicado pela experiência oferecida ao consumidor final, uma vez que nesse *cluster* os PDVs apresentam maior heterogeneidade no serviço ofertado. No que diz respeito ao R^2 e R^2 ajustado, ambos aplicados na porção de treino, os valores alcançados em todos os agrupamentos indicam boa adequação aos dados por parte do modelo, dado que ambas as métricas apresentaram valores entre 0,86 e 0,97, corroborando ao fato de que as variáveis adotadas explicam de modo satisfatório a venda de cerveja no mercado de Porto Alegre. Ademais, todas as variáveis, com exceção das variáveis categóricas no agrupamento das grandes redes, apresentaram *p-value* adequado, ou seja $p < 0,05$, como pode ser observado na Tabela 7.

Tabela 7 – *p-value* das variáveis

Cluster	Tipo de PDVs	Vol. Médio U3M	R$/l	Prazo	Chuva	Verão	Inverno
1	Pequenas redes	0	0	7,70E-23	2,60E-12	1,15E-04	3,45E-02
2	Tradicionais	0	0	2,90E-35	2,92E-03	7,70E-10	8,56E-03
3	Grandes Redes	0	1,00E-257	1,30E-28	9,07E-03	3,04E-01	1,83E-01
4	VIPs	0	2,10E-02	3,56E-04	3,99E-02	2,94E-03	6,59E-03

4.3 Análise dos coeficientes do modelo de regressão

A Tabela 8 apresenta os coeficientes associados a cada uma das variáveis para cada um dos *clusters*. A partir da interpretação dos coeficientes é possível obter conclusões acerca do impacto de cada variável no volume de cerveja vendido como, por exemplo, a predominância da variável volume mensal médio dos últimos 3 meses em todos os agrupamentos. Essa variável caracteriza o padrão de consumo do PDV incorporando

variáveis mais complexas de obtenção, tais como competitividade da área geográfica do cliente, bem como a margem com a qual ele trabalha nas vendas de cerveja (sendo coerente a sua predominância).

Tabela 8 – Coeficientes da Regresão

Cluster	Tipo de PDVs	Vol. Médio U3M	R$/l	Prazo	Chuva	Verão	Inverno
1	Pequenas redes	0,908	-0,14	0,003	0,004	0,002	0,001
2	Tradicionais	0,847	-0,13	0,005	-0,001	0,004	-0,001
3	Grandes Redes	0,873	-0,02	0,06	0,004	0,001	-0,002
4	VIPs	0,981	-0,2	0,02	-0,03	-0,06	0,05

Além disso, todos os *clusters* apresentam correlação positiva entre a variável prazo e a variável dependente, fato que atualmente é subaproveitado na empresa estudada. Analisando especifacamente as grandes redes, se observa que a variável prazo apresenta segundo maior valor absoluto, evidenciando que é mais efetivo para esse agrupamento conceder maior prazo de pagamentos do que promocionar o preço dos produtos. Para os clientes VIPs, se destaca o efeito negativo associado à variável verão, explicada pelo fato da população da cidade estudada viajar para a regiões litorâneas.

Outros dois pontos importantes são a sensibilidade do fator preço para esse tipo de cliente, sendo o *cluster* mais sensível à essa variável e o efeito positivo associado ao inverno o que pode ser explicado pelo tipo de cerveja vendido – líquidos com maior teor alcólico – em estabelecimentos desse perfil. Tanto os tradicionais quanto as pequenas redes apresentaram coeficientes semelhantes, sendo a única diferença significativa a sensibilidade dos tradicionais a variáveis como chuva e inverno. Nesse cenário, além de revisões na política de preço para esses clientes, uma alternativa seria proporcionar diferentes margens para os pontos de venda no intuito de que os produtos fiquem mais atrativos para os consumidores finais, incrementando a demanda dos PDVs.

5 Conclusão

Este artigo teve o objetivo de compreender, através de ferramentas de análise multivariada, o impacto de variáveis independentes (preço por litro de cerveja, prazo de pagamento, precipitação, volume mensal médio dos últimos 3 meses e estação do ano) sobre o volume de venda de cerveja no mercado de Porto Alegre. A partir dessas análises, objetivou-se então identificar possíveis estratégias para aumentar o volume de cerveja vendida, melhorando a principal métrica das empresas desse segmento. Para tanto, cerca

de 7 mil clientes foram segmentados em quatro grupos através do método *k-means*, aos quais foram aplicados modelos de regressão linear múltipla.

A *clusterização* dos clientes resultou em quatro perfis distintos de clientes (grandes redes, pequenas redes, tradicionais (mercearias) e clientes especiais (VIPs)). As variáveis utilizadas no agrupamento foram volume mensal médio dos últimos 3 meses do ano de 2018, número médio de produtos distintos comprados nos últimos 3 meses de 2018 e prazo de pagamento médio nos últimos 3 meses de 2018. Tal segmentação apresentou um *Silhouette Index* de 0,4902, indicando uma boa *clusterização* da base a partir das variáveis adotadas.

Os modelos preditivos foram gerados a partir das variáveis volume mensal médio dos últimos 3 meses, preço por litro, precipitação, prazo de pagamento e de duas variáveis categoricas uma para a estação verão e outra pra estação inverno; quando a observação pertencia à primavera ou ao outono ambas variáveis categóricas assumiam valor nulo. Foram utilizadas cerca de 200.000 observações, sendo 75% delas voltadas para base de treino e 25% para a base de teste. Na base de teste adotou-se o *mean square error* para aferir a capacidade preditiva do modelo, os valores variaram entre 0,004 e 0,016, para a base de teste foram comparados R^2 e R^2 ajustado, os quais variaram, respectivamente, entre 0,867 e 0,973 e 0,866 e 0,972 atestando a qualidade dos modelos, bem como da *clusterização*. Todas as variáveis apresentaram *p-value* que às caracterizaram como relevantes para o modelo, com exceção das variáveis categóricas para o *cluster* das grandes redes.

A partir desses resultados foi possível identificar que, para as grandes redes, é mais vantojoso conceder maior prazo de pagamento do que promocionar o preço da cerveja. Tal ação aumenta não só o volume de cerveja vendido, mas também o faturamento da empresa. Também ficou evidente a sensibilidade dos clientes VIPs ao preço, sendo o *cluster* das grandes redes aquele em que esse coeficiente obteve maior valor absoluto. Outro ponto importante é o efeito negativo que o verão apresenta em virtude da população local viajar para áreas litorâneas, reduzindo a venda de cerveja para o grupo. Tanto as pequenas redes quanto os tradicionais apresentaram como variável mais relevante o preço e, por serem um mercado mais pulverizado, recomenda-se que atributos como a margem com a qual os pontos de venda trabalham seja sugerida pela empresa, objetivando a redução do preço final ao consumidor e elevando o volume de cerveja vendido sem alterar o preço para os PDVs.

A metodologia aplicada nesse estudo de caso se mostrou adequada para os fins propostos, tanto no que diz respeito à métricas de desempenho dos modelos quanto as inferências obtidas através dos resultados. Para trabalhos futuros, recomenda-se que dados como número de PDVs concorrentes na região e a margem com a qual o PDV trabalha sejam utilizados na tentativa compreender com maior profundidade os fatores externos que influenciam na venda de cerveja.

Referências

ANZANELLO, Michel J.; FOGLIATTO, Flavio S. **Selecting the best clustering variables for grouping mass-customized products involving workers' learning.** Int. J. Production Economics 130, pp. 268-276, 2011.

ANZANELLO, Michel J.; ORTIZ, Rafael S.; LIMBERGER, Renata; MARIOTTI, Kristiane. **Performance of some supervised and unsupervised multivariate techniques for grouping authentic and unauthentic Viagra and Cialis**. Egyptian Journal of Forensic Sciences 4, pp. 83-89, 2014.

CALLAO, Maria P.; RUISÁNCHEZ, Itziar. **An overview of multivariate qualitative methods for food fraud detection.** Food Control 86, pp. 283-293, 2018.

CANIATO, Federico; KALCHSCHMIDT, Matteo; RONCHI, Stefanno; VERGANTI, Roberto; ZOTTERI, Giulio. **Clustering customers to forecast demand.** Production Planning & Control: The Management of Operations, 16:1, pp. 32-43, 2005.

FALCONI, Vicente. **O Verdadeiro Poder,** 2014.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. **The elements of statistical learning: Data Mining, Inference, and Prediction.** 2. ed. New York: Springer series in statistics, 2008.

HAIR JR, Joseph F.; BLACK, William C.; BABIN, Barry J. **Multivariate Data Analysis: A Global Perspective**, Pearson Education 2010.

JAMES, Gareth; WITTE, Daniela; HASTIE, Trevor; TIBSHIRANI, Robert. **An Introduction to Statistical Learning.** Nova Iorque: Springer 2013 (8a edição 2017).

KUNZ, Timo P.; CRONE, Sven F. **Demand Models for The Static Retail Price Optimization Problem – A Revenue Management Perspective**. Lancaster University SCOR'14, pp. 101-125, 2014.

MORAES, Renan M. **Agrupamento de Clientes e Composição das Equipes de Venda do Grupo RBS Através de Técnicas Multivariadas.** 2017. Dissertação para título de Mestre (Mestre em Engenharia de Produção) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2017. Orientador: Prof. Michel José Anzanello, Ph.D].

MÜLLER, Cláudio J. **Modelo de Gestão Integrando Planejamento Estratégico, Sistemas de Avaliação de Desempenho e Gerenciamento de Processos (MEIO – Modelo de Estratégia, Indicadores e Operações)**, 2003

NETO, Luis B.; MELLO, João Carlos.; LEITE, Vivian P.; CHIERALLA, Rachel C.; ALVES, Laura A. **Previsão de Faturamento para Lojas do Setor de Varejo com Redes Neurais.** Revista Eletrônica Pesquisa Operacional para o Desenvolvimento, p.1-13, 2013.

RENCHER Alvin C. **Methods of Multivariate Analysis**. Nova Iorque, Estados Unidos. Editora: Wiley- Interscience 2002 (2ª edição).

SASSI, Cecília P.; PEREZ, Felipe G.; MYAZATO, Xiao Ye; FERREIRA-SILVA, Paulo H.; LOUZADA, Francisco. **Modelos de Regressão Linear Múltipla Utilizando os Softwares R e Statistica: Uma Aplicação a Dados de Conservação de Frutas.** São Paulo, 2011.

SHU, Yuqin; LAM, Nina S. N. **Spatial Disaggregation of Carbon Dioxide Emissions from Road Traffic Based on Multiple Linear Regression Model.** Atmospheric Environment 45, pp. 634-640, 2011.

SIKORSKA, Ewa; KHMELINSKII, Igor; WLODARSKA, Katarzyna. **Authentication of apple juice categories based on multivariate analysis of the synchronous fluorescence spectra.** Food Control 86, pp. 42-49, 2018.

TABOADA, Heide A.; COIT, David W. **Data Clustering of Solutions for Multiple Objective System Reliability Optimization Problems**, 2007.

WANG, Holly H.; ZHANG, Xu; ORTEGA, David L.; WIDMAR, Nicole J. O. **Information on food safety, consumer preference and behavior: The case study os seafood in the US.** Food Control 33, pp. 293-300, 2013.

ZERVOS, Constantine; ALBERT, Richard H. **Chemometrics: The Use of Multivariate Methods for the Determination and Characterization of Off-Flavors.** Developments in Food Science 28, pp. 669-742, 1992.

ZHAO, Wan-Lei; DENG, Cheng-Hao; NGO, Chong-Wah. **K-means: A revisit.** Neurocomputing 291, pp. 195-206, 2018.

ZOTTERI, Giulio; KALCHSCHMIDT, Matteo; CANIATO, Federico. **The impact of aggregation level on forecasting performance.** International Journal of Production Economics, 93-94:479–491, January 2005.