Universidade do Minho
Escola de Ciências

Nelson Filipe Sá Costa

Análise multivariada de dados sobre tipologia de produtos numa empresa

Outubro de 2022

Análise multivariada de dados sobre tipologia de produtos numa empresa
Nelson Filipe Sá Costa
UMinho | 2022

Universidade do Minho
Escola de Ciências

Nelson Filipe Sá Costa

Análise multivariada de dados sobre tipologia de produtos numa empresa

Dissertação de Mestrado
em Matemática e Computação

Trabalho efetuado sob a orientação do(a)
Professora Doutora Inês Sousa
Doutor Luís Rodrigues

Outubro de 2022

Agradecimentos

É necessário aqui deixar o meu profundo reconhecimento a todos aqueles que de algum modo tornaram exequível este trabalho de dissertação de mestrado apresentado à Universidade do Minho, e sem os quais nunca teria sido possível ser levado a cabo.

Agradeço à Professora Doutora Inês Sousa, por ter aceite o desafio de ser minha orientadora. Acima de tudo agradeço-lhe por ter sido uma das pessoas que acrescentou significado na minha vida nesta fase académica, nomeadamente o gosto pela estatística e análise de dados. As suas competências científicas inspiraram-me a desenvolver este trabalho com um certo cuidado e rigor científico.

Quero agradecer a toda a equipa da empresa Litel, Lda, por me ter recebido tão bem na empresa, e ter-me lançado este desafio. Destaco, a amabilidade do Doutor Ricardo Carneiro, e da Engenheira Inês Almeida, em disponibilizarem o espaço, e se mostrarem disponíveis em qualquer situação. Mais agradeço ao Doutor Luís Rodrigues, por guiar o meu trabalho, e mostrar-se sempre disponível em discutir ideias, de modo a aprimorar a qualidade da dissertação.

Aos meus pais, por me permitirem realizar o mestrado, é lhes devida uma palavra de agradecimento, especialmente à minha mãe que sempre demonstrou afeto, cuidado, e me incentivou a encarar os desafios da vida, como oportunidades para alcançar o sucesso.

Agradeço a todos os meus amigos, por me restituírem a força e ânimo nos momentos menos bons, em especial ao Hélder Vieira, por sempre se ter mostrado presente na minha vida, à professora Alzira Mota, por me ter impulsionado o desejo de realizar o mestrado, e à Ana Maria Dias pela sua generosidade.

Agradeço no geral a todas as pessoas que contribuíram para que a minha jornada académica se tornasse inesquecível.

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

Análise multivariada de dados sobre tipologia de produtos numa empresa

As pequenas, médias e grandes empresas recolhem os seus dados esperando que estes de alguma forma gerem valor comercial. O *machine learning* tem a capacidade de fornecer informação valiosa a partir dos dados, servindo de principal estratégia de vendas para que a empresa consiga alcançar maior lucro.

O objetivo desta dissertação é analisar dados de vendas da empresa Litel, Lda, procurando extrair padrões de vendas relativos a produtos (caixas e sacos) que compartilham similaridades entre si. Para alcançar este objetivo são apresentadas metodologias de aprendizagem não supervisionada que permitem traçar uma segmentação de grupos, nomeadamente Análise de Componentes Principais (PCA), algoritmos de *k-Means* e *clustering* hierárquico.

O PCA foi usado para identificar variáveis correlacionadas e identificar padrões ocultos nos dados, nomeadamente famílias de produtos com similaridades nas vendas. Foram aplicados os métodos do cotovelo, da silhueta e dos 30 índices para escolha do melhor número de *clusters*, de modo a identificar o número ótimo de *clusters*. Foram ainda aplicados métodos de validação, de modo a identificar o algoritmo de *clustering* com melhor desempenho. Através das medidas de estabilidade foi avaliada a consistência dos *clusters*, e através do coeficiente cofenético foi determinada a combinação de métodos que refletiu num melhor agrupamento de dados. Após realizar as validações anteriores foram implementados os respetivos algoritmos de *clustering*. De um modo geral, todos os algoritmos de *clustering* segmentaram os dados de uma forma bastante satisfatória podendo destacar a ótima performance do algoritmo *clustering* hierárquico método *"average"*.

Palavras-Chave: *Machine Learning*, PCA, *k-Means*, *Clustering* Hierárquico, Vendas.

Multivariate analysis of data on typology of products in a company

Small, medium and large companies collect their data hoping that it will somehow generate commercial value. Machine learning has the ability to provide valuable information from data, serving as the main sales strategy for the company to achieve greater profit.

The objective of this dissertation is to analyze sales data from the company Litel, Lda, seeking to extract sales patterns related to products (boxes and bags) that share similarities with each other. In order to achieve this objective, unsupervised learning methodologies are presented, which allow for a segmentation of groups, namely Principal Component Analysis (PCA), k-Means algorithms and hierarchical clustering.

The PCA was used to identify correlated variables and identify hidden patterns in the data, namely product families with similar sales. The elbow, silhouette and 30 indices methods were applied to choose the best number of clusters, in order to identify the optimal number of clusters. Validation methods were also applied in order to identify the clustering algorithm with the best performance. Through the stability measures, the consistency of the clusters was evaluated, and through the cophenetic coefficient, the combination of methods that reflected a better data grouping was determined. After carrying out the previous validations, the respective clustering algorithms were implemented. In general, all clustering algorithms segmented the data in a very satisfactory way, highlighting the excellent performance of the hierarchical clustering algorithm ”average”method.

Keywords: Machine Learning, PCA, k-Means, Hierarchical Clustering, Sales

Índice

Lista de Figuras

Lista de Tabelas

Lista de Acrónimos

IA Inteligência Artificial

PCA Principal Component Analysis

APN Average proportion of non-overlap

AD Average distance

ADM Average distance between means

FOM Figure of merit

Capítulo 1

Introdução

No âmbito da unidade curricular de dissertação do mestrado em Matemática e Computação, foi proposta a realização de um projeto académico. Este projeto tem o seu foco aplicado a *machine learning*, e surgiu de uma parceria entre a Universidade do Minho, e a empresa Litel, Lda. A empresa Litel, Lda, dispõe de uma base de dados referente a vendas de produtos, entre 2015 e 2021. O objetivo passa pela realização de uma análise multivariada de dados sobre tipologia de produtos da empresa.

Os dados têm ganhado importância na sociedade ao longo dos anos. A sua aquisição e estudo, proporcionam às empresas boas tomadas de decisão, desde áreas como recursos humanos, que permite a seleção e recrutamento dos melhores trabalhadores, e de áreas como marketing que permitem realizar a segmentação de mercado, de forma a encontrar consumidores que estão prontos para comprar (acelerando o processo de vendas). O uso de forma eficiente de dados, agiliza o processo de venda, economizando custos. Desta forma, analisar os dados e incorporá-los na estratégia de vendas, é o papel do gestor (Leonard (2022)).

O *machine learning* é um subcampo da inteligência artificial (IA) que envolve a análise de dados que automatiza a construção de modelos analíticos. Faz uso de algoritmos que aprendem iterativamente a reconhecer padrões complexos e tomar decisões informadas em base em dados, sem ser programado onde pesquisar (Han et al. (2011)).

O *data mining* é o processo de extração e descoberta de padrões em grandes conjuntos de dados. A sua base compreende três disciplinas científicas: a estatística (o estudo numérico das relações entre os dados), a IA (a inteligência exibida por *softwares* e/ou máquinas) e o *machine learning*. O *data*

mining é projetado para extraír regras de grandes conjuntos de dados, enquanto o *machine learning* ensina o computador a aprender e compreender os parâmetros fornecidos (SAS (2022)).

O *machine learning* é a ferramenta ideal para fornecer *insights* precisos sobre os negócios, e reconhecer padrões em grandes quantidades de dados. Isso permite que as equipas comerciais identifiquem as melhores oportunidades de vendas o mais rápido possível.

Através de algoritmos, a IA acompanha equipas de marketing e vendas para proceder à deteção de potenciais clientes valiosos para a empresa.

Uma das vantagens com recurso a esta ferramenta, é que podem ser recolhidas informações de vendas para análise preditiva de clientes, melhorando o desempenho da equipa de vendas.

Outro aspeto vantajoso, é a deteção de ações que têm maior probabilidade de fechar uma venda, ajudando gestores e vendedores a realizar vendas rentáveis sem desperdiçar tempo em ações com menos probabilidade de sucesso.

Além disso, o *machine learning* permite que os vendedores economizem tempo em tarefas manuais, de previsão e de emissão de relatórios. Assim, as equipas de vendas podem garantir um bom serviço.

O *machine learning* também pode reduzir significativamente o processo de *onboarding* para novos vendedores, ou seja, o tempo e os recursos necessários para treino e orientação que permitem a incorporação ideal na empresa.

Promove maior rapidez na aprendizagem e compreensão do funcionamento da empresa e seus objetivos. Enquanto um novo vendedor pode precisar de meses para entender e vender corretamente um produto, um sistema de *machine learning* pode orientar o vendedor, e facilitar o desempenho das suas funções de forma eficaz (Catalog Player Smart Content (2022)).

1.1 Objetivos e Motivação

A Litel, Lda, tem como necessidade traçar uma estratégia de vendas para obter maior lucro na comercialização dos seus produtos. Desta forma, pretende-se saber quais são os produtos com mais e menos vendas, de forma a ser possível uma melhor tomada de decisão por parte da direção comercial da empresa Litel, Lda. O presente trabalho de dissertação tem como principal objetivo a análise de uma base de dados relativa a vendas de produtos da empresa Litel, Lda, nomeadamente sacos e caixas. Este objetivo geral tem como objetivos específicos:

- Realizar uma análise exploratória dos dados;
- Gerar gráficos que permitam aferir o comportamento das vendas de produtos ao longo do tempo, evidenciando quais são os produtos com mais e menos vendas;
- Aplicar uma Análise de Componentes Principais sobre as variáveis dos dados;
- Aplicar algoritmos de *clustering* de forma a agrupar os dados de vendas;
- Avaliar e discutir quais os melhores métodos aplicados;
- Desenvolver habilidades com o uso do *software* R na implementação de código de forma a cumprir os objetivos anteriores.

1.2 Apresentação da empresa

A Litel, Lda (Figura 1), foi fundada em 1974 na Trofa, e apresenta-se como uma empresa familiar, com 80 colaboradores, que desenvolve e fabrica soluções para embalagem. Desde a sua fundação que a Litel aposta na qualidade e inovação, apresentando novos produtos e soluções, motivo pela qual se tem diferenciado da concorrência. Com mais de 40 anos de experiência na área de embalagem, dispõe de um leque variado de produtos em papel, com elevada qualidade. Utiliza para o efeito a mais recente tecnologia e recursos humanos qualificados, para assegurar qualidade, inovação e preços competitivos (Litel (2022b)). Na Tabela 1 é feita uma descrição cronológica da evolução da empresa (Litel (2022a)).

Figura 1: Exterior das instalações da empresa, Litel, Lda (fonte: Litel (2022b))

Tabela 1: Descrição cronológica da evolução da empresa Litel, Lda (fonte: Litel (2022a))

Ano	**Descrição**
1974	A Litel foi fundada;
1995	Houve mudança de instalações para outras de maior dimensão $3500m^2$;
1996	Implementa-se o Sistema de Gestão da Qualidade (NP EN ISSO 90001);
1998	Aquisição de 95 % capital de uma unidade de produção de caixas de cartão canelado;
1999	Implementação do sistema de Gestão Ambiental (NP EN ISSO 14001);
2005	Aquisição da 1ª linha automática de produção de sacos de papel e início do processo de internacionalização;
2011	Certificação da Cadeia de Responsabilidade $PEFC^{TM}$ e aquisição de uma nova unidade industrial com 5000 m^2 ;
2015	Nova imagem corporativa;
2017	Aquisição de tecnologia de aspiração industrial (diminuição do impacte ambiental) ;
2018	Alargamento da área de negócio com aposta no mercado de sacos de papel;
2019	Certificação da Cadeia de Custódia FSC® e participação na campanha "Recicle Sempre"(SPV);
2020/2021	Aquisição de novas máquinas (Remodelação de setor de caixas e sacos).

A Litel é uma empresa direcionada para a produção de sacos de papel com asa em cordão rígido/algodão, sacos de gama luxo, caixas e envelopes para oferta, caixas de transporte, entre outros artigos (Figura 2 e 3).

Figura 2: Sacos comercializados na empresa, Litel, Lda (fonte: Litel (2022b))

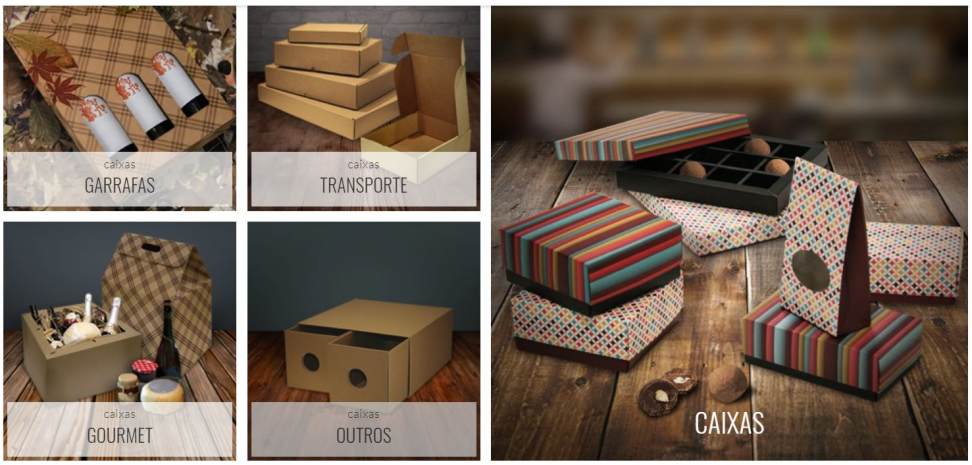

Figura 3: Caixas comercializadas na empresa, Litel, Lda (fonte: Litel (2022b))

Estes materiais podem ter variadas medidas, quantidades e tipos de qualidades (papel Kraft, Kraft Branco, papel Couché, papel Verjurado, cartolina e microcanelado) sob impressão *Offset* ou Flexográfica e acabamentos diversos (relevo, plasticização brilho/mate, verniz, termo estampagem). O comércio de sacos/caixas está direcionado, efetivamente para o setor vitivinícola, *e-commerce*, setor da moda e outros (Litel (2022b)).

A Litel apresenta-se com uma área geográfica nacional e internacional, nomeadamente em: Portugal, Espanha, França, Inglaterra, Alemanha, e outros países da U.E. (Litel (2022c)).

A empresa tem como clientes determinados setores, tais como: bebidas, alimentação, eletrónica, comércio e distribuidoras, entre outros.

O objetivo da Litel, Lda. é de ser reconhecida como empresa de referência na área de embalagens/-sacos e criar valor para o acionista.

Quanto à estratégia e planeamento organizacional, passa pelos seguintes objetivos (Litel (2022d)):

- Ser reconhecida como empresa de referência na área de embalagens/sacos;
- Apostar no desenvolvimento sustentável ao nível económico, social e ambiental;
- Estabelecer estratégias de prevenção contínua de riscos para a segurança, saúde e impactes para o ambiente;
- Otimizar os recursos;
- Prevenção da Poluição;
- Valorizar e Motivar os Recursos Humanos;
- Melhoria Contínua.

1.3 Estrutura

Este trabalho de dissertação está organizado em 7 capítulos como se descreve a seguir.

Neste Capítulo 1, está patente a introdução, na qual é feita uma breve apresentação do tema a ser tratado, os objetivos que se pretendem cumprir, e a apresentação da empresa.

No Capítulo 2 é realizada uma revisão de literatura na área de vendas, onde são apresentados estudos realizados no âmbito da área de vendas, e também é feita uma revisão de literatura na área do *machine learning*, principalmente focada na vertente de aprendizagem não supervisionada.

No Capítulo 3, são apresentados os fundamentos estatísticos, em que se enunciam algoritmos utilizados na vertente prática da dissertação.

No Capítulo 4 são descritos os dados fornecidos pela empresa, e as várias etapas do pré processamento.

No Capítulo 5 é realizada uma análise exploratória dos dados de venda, em que se apresentam estatísticas descritivas das variáveis da base de dados, histogramas, e gráficos que traçam o comportamento de produtos ao longo do tempo.

No Capítulo 6 são apresentados e discutidos os resultados da aplicação de algoritmos, nomeadamente análise de componentes principais, *k-Means* e *clustering* hierárquico.

No Capítulo 7 são aferidas as principais conclusões do trabalho de dissertação, e são sugeridas ideias a considerar para futuros trabalhos.

Capítulo 2

Revisão de literatura

2.1 Área das Vendas

A estratégia de vendas correta é crucial nos negócios para aumentar o valor das vendas. A venda de um produto pode ser afetada por diversos fatores. Taufiq Luthfi (2009) utiliza *data mining* com algoritmo de associação para criar um sistema que determina o padrão de vendas de produtos, e pode ser utilizado para desenvolver novas estratégias de vendas. A partir deste estudo, concluiu-se que o empreendedor pode criar uma estratégia de vendas tendo em conta a relação entre os produtos da empresa (Taufiq Luthfi (2009)).

A implementação de *data mining* nas vendas de produtos de bebidas na Pepsi Cola Indobeverages PT foi realizada por Irdiansyah (2010). Irdiansyah (2010) usa técnicas de agrupamento para identificar objetos que compartilham certas características e usa essas características como 'vetores de características' ou 'centróides'. Com base na pesquisa, verificou-se que o método de agrupamento ajuda a Pepsi Cola Indobeverages PT a obter uma visão geral das decisões de negócio para obter os padrões de vendas dos produtos fabricados (Irdiansyah (2010)).

Arga Felani (2015) determinou a estratégia de venda de alimentos e bebidas na Toserba Lestari Baru Gemolong. Ela comparou três métodos de *data mining* para determinar estratégias de vendas: Árvore de Decisão, *K-Means Clustering* e Regressão Linear. O estudo foi realizado com um grupo

de dados para determinar a percentagem de valor de precisão (métrica utilizada para indicar a relação entre as previsões positivas realizadas corretamente e todas as previsões positivas (incluindo as falsas)), *recall* (métrica utilizada para indicar a relação entre as previsões positivas realizadas corretamente e todas as previsões que realmente são positivas) e *accuracy* (métrica usada para indicar a relação entre as previsões realizadas corretamente e a totalidade de previsões). O estudo concluiu que o método *K-Means Clustering* teve melhor valor do que outros métodos em precisão e *accuracy*, enquanto o método de Regressão Linear teve melhor valor de *recall* (Arga Felani (2015)).

A análise de *clusters* para segmentação de mercado tem sido estudada e amplamente utilizada em inúmeras aplicações nas mais diversas áreas de negócios usando diversos métodos de *clustering*. Punj and Stewart (1983) compilaram uma lista de 20 exemplos onde a análise de *clusters* foi usada para segmentação de mercado. Dolnicar (2003) examinou como a segmentação de mercado por análise de *clusters* é geralmente realizada e quais os pontos menos positivos de tentativas anteriormente realizadas. Dos 243 artigos incluídos nesse estudo, o tamanho médio da amostra é de 698 com um máximo de 20000. O número médio de variáveis é de 11,5 com um máximo de 66. Um conjunto de dados com tamanho de amostra superior a 38000 com mais de 1700 variáveis foram usadas neste estudo. Isto é significativamente maior tanto no tamanho da amostra, quanto no número de variáveis em comparação com tentativas anteriores. Os algoritmos de agrupamento retornam sempre resultados independentemente do tamanho da amostra, mas quanto mais variáveis forem usadas, maior será o tamanho da amostra necessário para encontrar agrupamentos naturais.

Além disso, Dolnicar (2003) descobriu que o método de variância mínima de *Ward* e o *K-means* são os métodos de agrupamento mais usados na literatura. Nenhum algoritmo único é adequado para todos os problemas, portanto, o pesquisador precisa garantir que o método escolhido seja adequado para a tarefa de segmentação. Como medida de similaridade, a distância euclidiana foi utilizada em 96% dos casos. A distância euclidiana geralmente é adequada, mas possui pontos menos positivos quando se tratam de dados ordinais. Se a distância euclidiana for usada com dados categóricos, assume-se uma distância igual para os intervalos entre as categorias, o que é uma suposição duvidosa (Dolnicar (2003)). Strehl et al. (2000) apontaram que a distância cosseno é mais apropriada que a distância euclidiana quando os dados são esparsos.

2.2 *Machine Learning*

Segundo Domingos (2015), o conhecimento provinha apenas de três fontes: evolução (que é o conhecimento codificado no nosso DNA), experiência (o conhecimento codificado nos nossos neurónios) e cultura (o conhecimento que se adquire comunicando com outros pessoas, lendo livros, artigos, etc). Atualmente existe uma quarta fonte de conhecimento no planeta: os computadores. De acordo com Yann et al. (1998), "a maior parte do conhecimento no mundo no futuro será extraído e armazenado em máquinas".

Os computadores e as máquinas geralmente têm cinco formas de descobrir novo conhecimento, que é a premissa para definir o que é *machine learning*:

- Preencher lacunas no conhecimento existente;
- Imitar o cérebro;
- Simular a evolução;
- Reduzir sistematicamente a incerteza;
- Notar semelhanças entre o antigo e o novo

Os modelos de *machine learning* recebem um conjunto de dados e inferem informações sobre as propriedades dos dados – e essas informações permitem que eles façam previsões sobre outros dados que possam surgir no futuro. Isto é possível porque quase todos os dados não aleatórios contêm padrões, e esses padrões permitem que a máquina generalize (Segaran (2007)).

De acordo com Li (2018), existem distintos problemas de *machine learning*. Esses problemas estão agrupados em três categorias e são apresentadas na Figura 4. São conhecidas como Aprendizagem Supervisionada, Aprendizagem Não Supervisionada e Aprendizagem por Reforço.

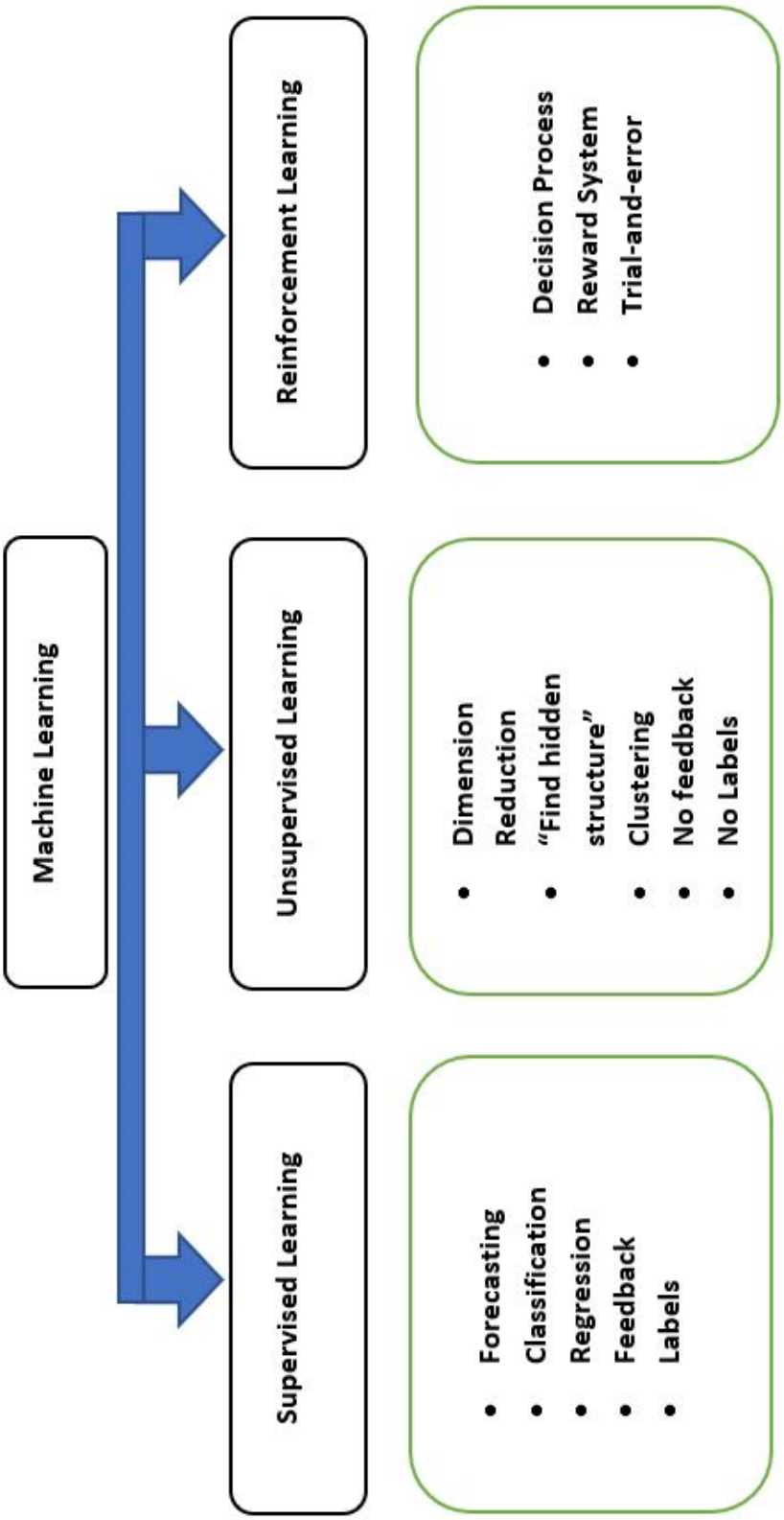

Figura 4: Diferentes tipos de *machine learning* (Adaptado de: NomidI (2022))

2.2.1 Aprendizagem não supervisionada

A aprendizagem não supervisionada (Figura 5) ainda carece de implementação em uma escala mais ampla, mas, de acordo com Loon (2022), ela denotará o verdadeiro potencial do *machine learning* no futuro. Esta abordagem é utilizada quando não existe categorização dos dados. Os humanos não desempenham nenhum papel no ensino do algoritmo, portanto, a máquina deve encontrar os padrões intrínsecos subjacentes aos dados por si só, sem conjuntos de dados de treino ou saídas conhecidas. O algoritmo aprende apenas com as informações que tem em mãos, tentando reconhecer todos os objetos semelhantes e agrupa-os. Todas as *labels* que os objetos recebem são criados pela própria máquina com suas principais subcategorias (Loon (2022)). De seguida apresentam-se dois métodos deste tipo de aprendizagem não supervisionada:

- *Principal Component Analysis*: alguns conjuntos de dados brutos têm recursos redundantes ou irrelevantes, o que às vezes tende a aumentar muito a dimensão dos dados, omitindo o seu relacionamento verdadeiro e latente. Reduzir a dimensionalidade ajuda a evitar isso.
- *Clustering*: Agrupa os dados em *clusters* (ou grupos) de acordo com as semelhanças identificadas pelo algoritmo. O conjunto de dados é segmentado em diferentes grupos abrindo caminho para uma análise mais apurada visando descobrir os seus padrões intrínsecos.

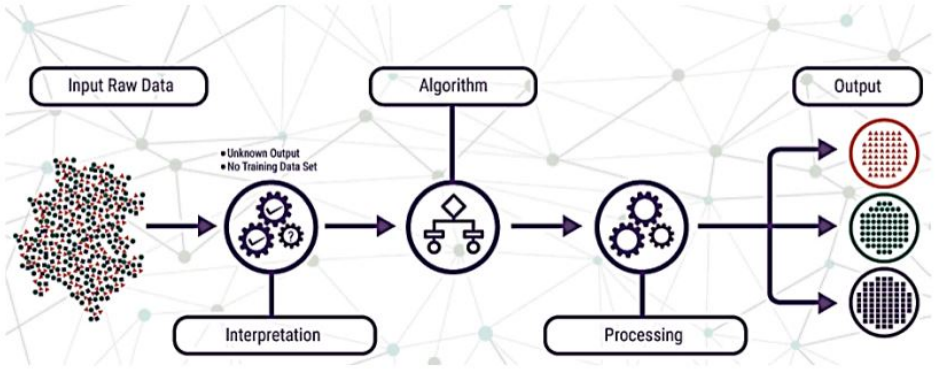

Figura 5: Fluxograma de Aprendizagem Não Supervisionada (fonte: Loon (2022))

Clustering

Clustering é a tarefa de agrupar os dados de tal forma que os objetos num *cluster* sejam mais semelhantes entre si do que os de outros *clusters*. A semelhança entre os pontos de dados pode ser medida usando um ou mais parâmetros, como distâncias entre pontos de dados, densidade de pontos de dados ou com base em outras distribuições estatísticas. O *clustering* pode ser definido então como um problema de otimização multi-objetivo. Existem diferentes tipos de algoritmos de *clustering*, como o *clustering* hierárquico, o baseado em centróides, como o algoritmo *k-Means*, o baseado em densidade como o DBSCAN, e outros. Um típico processo de análise de *clusters* pode ser descrito através das etapas que estão presentes na Figura 6:

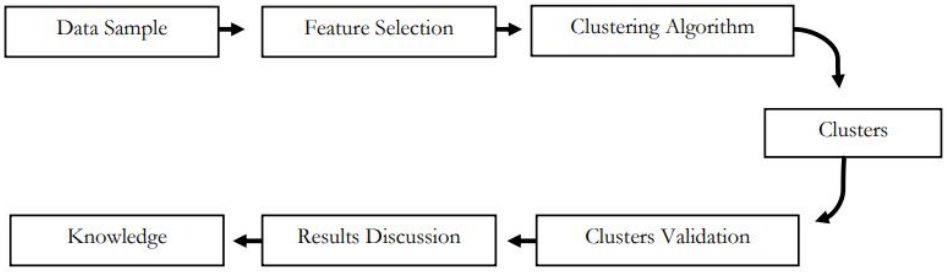

Figura 6: Processo de análise de *clusters* (fonte: Xu and Wunsch (2005))

Principal Component Analysis

Segundo Johnson (1992), o *Principal Component Analysis (PCA)* é mais um meio para se atingir um resultado final do que um resultado por si só. Esta técnica é normalmente utilizada para alcançar alguns objetivos como a redução de dimensionalidade de uma base de dados com muitas variáveis aplicando uma tranformação para a base formada pelas componentes principais, mantendo a maior parte da variância original. Outra utilidade da PCA é explicitar correlações ocultas entre as variáveis, por meio de visualizações do resultado obtido pela aplicação da técnica. Esta nova apresentação dos dados pode auxiliar na interpretação dos mesmos, ao evidenciar tendências, tornar clara a relevância de variáveis para a variância original dos dados e mostrar redundâncias no conjunto original de variáveis.

Capítulo 3

Fundamentos Estatísticos

3.1 *Principal Component Analysis*

A técnica de análise de componentes principais é uma ferramenta estatística que tem como objetivo descrever a estrutura de variância e covariância de um conjunto de variáveis ou dimensões, por meio de combinações lineares dos membros desse conjunto. A PCA aplicada às amostras de medições de um dado sistema, mostra nos como e com qual importância essas dimensões contribuem para a variabilidade dos valores de medição, muitas vezes revelando relações ocultas entre elas. Além disso, o PCA é uma ferramenta utilizada para reduzir redundâncias e reduzir a dimensionalidade do conjunto de variáveis usadas para observar um sistema por meio da criação de uma nova base, cujos componentes são linearmente independentes e em menor número, a partir das principais componentes apontadas pela PCA entre o conjunto inicial de dimensões. Esses novos componentes são ordenados de modo a reter a maior parte da variabilidade original dos primeiros componentes. Assim, a principal componente (PC) resultante da aplicação da PCA representa o eixo da nova base com a maior dispersão dos dados originais.

3.1.1 Algoritmo

De seguida apresentam-se os passos necessários para aplicar o algoritmo de PCA (Andrecut (2009)):

1. Organizar os dados das medições numa matriz n x m, onde m é o número de variáveis medidas, ou dimensões, e n é o número de amostras.

2. Caso seja necessário, dividir as medições de cada dimensão pelo seu desvio padrão para normalizá-las e evitar a sensibilidade da PCA à diferença de escala entre dimensões.

3. Calcular a matriz de covariância da matriz resultante dos passos anteriores (caso o passo 2 tenha sido efetuado, essa matriz será a de correlações).

4. Calcular os autovetores e autovalores associados à matriz de covariância.

5. Ordenar os autovetores de acordo com os autovalores associados. Deste modo, o primeiro autovetor é o componente principal, o segundo é o segundo componente principal, e assim por diante.

6. Descartar as componentes de menor relevância. Para tal, definir o percentual da variância original que deve ser mantido e escolher as componentes principais de modo que a soma dos autovalores associados seja maior ou igual a esse percentual.

3.1.2 Descrição Matemática

Seja o vetor $X' = \begin{bmatrix} X_1 & X_2 & \cdots & X_p \end{bmatrix}$ com matriz de covariância Σ , cujos autovalores são $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ e os autovetores associados $\epsilon_1, \epsilon_2, \cdots, \epsilon_p$.

Seja $a = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix}$, então $Y = a^t X = a_1 X_1 + a_2 X_2 + \cdots + a_p X_p$ é uma combinação linear dos elementos do vetor X.

Agora considerem-se as combinações lineares:

$$Y_1 = a_1^t X = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p$$

$$Y_2 = a_2^t X = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p$$

$$\vdots$$

$$Y_p = a_p^t X = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p$$

Temos que:

$Var(Y_i)$ = $a_i^t\ \Sigma\ a_i$, i= 1, 2, $\cdots$, p

$Covar(Y_i, Y_k)$ = $a_i^t\ \Sigma\ a_k$, i, k= 1, 2, $\cdots$, p

As componentes principais são as combinações lineares não correlacionadas $Y_1, Y_2, \cdots, Y_p$ para as quais a variância é a maior possível, decrescendo de Y_1 a Y_p. Escolhendo os valores de a_i como sendo os autovetores ϵ_i da matriz de covariância Σ do vetor X, temos que a i-ésima componente principal , Y_i, é:

$$Y_i = \epsilon_1^t X = \epsilon_{i1}X_1 + \epsilon_{i2}X_1 + \cdots + \epsilon_{ip}X_p, \qquad i = 1, 2, \cdots, p$$

E então:

$Var(Y_i)$ = $\epsilon_i^t\ \Sigma\ \epsilon_i$ = λ_i , i= 1, 2, $\cdots$, p

$Covar(Y_i, Y_k)$ = $\epsilon_i^t\ \Sigma\ \epsilon_k$ = 0 , i, k= 1, 2, $\cdots$, p

Para escolher as componentes principais de modo a manter a proporção P_v da variância original dos dados, deve-se escolher as componentes $Y_1, \cdots, Y_k$ tais que:

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \geq P_v$$

Ou seja cada componente principal, Y_k representa uma proporção da variância de:

$$\frac{\lambda_k}{\sum_{i=1}^{p} \lambda_i}$$

Mais detalhes sobre fundamentos matemáticos do PCA podem ser consultados na referência Johnson (1992).

3.2 *Clustering*

3.2.1 *k-Means*

O *k-Means* é um dos métodos de agrupamento, e é utilizado quando existem dados não rotulados. As técnicas de agrupamento consistem em iterar um conjunto de dados automaticamente pelo seu grau de similaridade (Steinbach et al. (2000)). O grau de similaridade depende da definição do problema e do algoritmo utilizado. Os algoritmos de agrupamento mais comuns são os particionais e os hierárquicos. A forma mais simples de *clustering* é o *clustering* particional, que visa particionar um determinado conjunto de dados em subconjuntos disjuntos (*clusters*) para que critérios específicos de *clustering* sejam otimizados. O *k-Means* é um algoritmo particional e minimiza o erro de agrupamento.

No *k-Means* subdividem-se os pontos de dados de um determinado conjunto em *clusters* com base nos valores médios mais próximos. De modo a determinar a divisão ótima desses pontos de dados em *clusters*, a distância entre os pontos deve ser minimizada. O objetivo deste algoritmo é minimizar uma função objetivo, neste caso uma função de erro quadrado. A função objetivo é definida como:

$$M(P, C) = \sum_{k=1}^{K} \sum_{i \in P_k} \left\| x_i - c_k \right\|^2$$

Onde P é uma partição *k-cluster* do conjunto de objetos representado por vetores x_i ($i \in I$) no espaço de variáveis N-dimensional, consistindo em *clusters* não vazios e não sobrepostos M_k, cada um com um centróide c_k (k=1,2,...K) (Kodinariya and Makwana (2013)). Para atribuir cada ponto de dados a um dos k *clusters* com base na similaridade de recursos do conjunto de dados, o algoritmo *k-Means* funciona iterativamente e os resultados finais são os centróides do k *clusters* (Figura 7) (que podem ser usados para rotular novos dados). Investigar os pesos dos recursos do centróide pode ser usado para entender qualitativamente o tipo de grupo que cada *cluster* representa (Hamerly and Elkan (2004)).

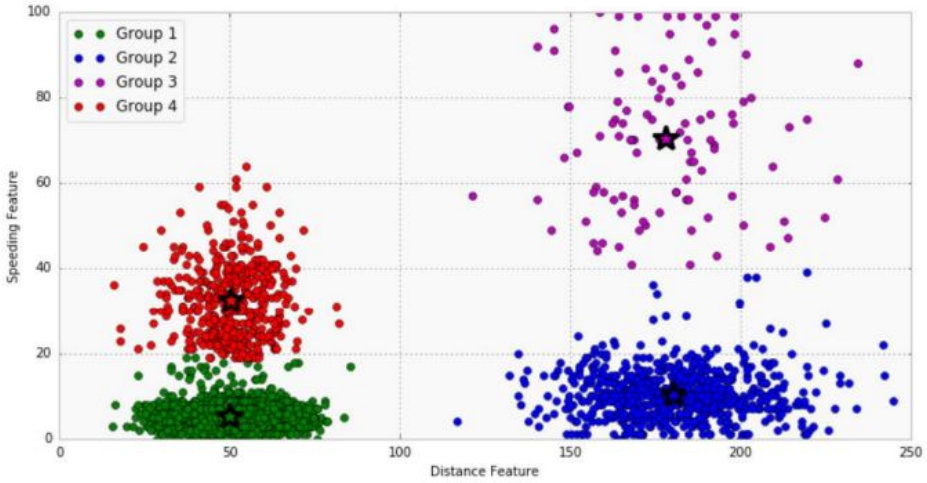

Figura 7: Exemplos de centróides formados a partir de iteração do *k-Means* (resultados finais) (fonte: Hamerly and Elkan (2004))

O algoritmo começa com uma estimativa para k centróides, que podem ser gerados aleatoriamente e, em seguida, iterados em duas etapas:

1. Atribuição de dados: Com base no quadrado da distância euclidiana – "a distância é calculada encontrando o quadrado da distância entre cada *score*, somando os quadrados e encontrando a raiz quadrada da soma" (Oyelade et al. (2010)) – cada ponto de dados é atribuído ao seu centróide mais próximo, e cada centróide descreve um dos *clusters*. Formalmente, se c_i é o grupo de centróides no conjunto C, cada ponto de dados x é atribuído a um *cluster* com base em:

$$argmin_{c_i \in C} dist(c_i - x)^2$$

Onde, $dist\ (c_i - x)^2$ representa a distância euclideana.

2. Atualização de centróides: Reatribuição dos centróides formados na etapa anterior. Essa reatribuição é realizada recalculando a média de todos os pontos de dados atribuídos ao *cluster* desse centróide, com base em:

$$c_i = \frac{1}{|D_i|}\Sigma_{x_i \in D_i} x_i$$

Existem mais medidas de distância que podem ser utilizadas em algoritmos de agrupamento p. ex.: *minkowski*, *manhattan* e *mahalanobis*. A distância é, frequentemente, escolhida de acordo com o tipo de dados. O *k-Means* funciona atribuindo pontos de dados ao centróide mais próximo usando a distância euclidiana dos pontos de dados ao centróide. Este algoritmo baseia-se implicitamente na distância euclidiana, uma vez que a soma dos desvios ao quadrado do centroide é igual à soma das distâncias quadráticas euclidianas dividida pelo número de pontos de dados. O próprio termo "centroide" vem da geometria euclidiana (Oyelade et al. (2010)).

As duas etapas funcionam iterativamente até que um critério seja satisfeito, o que significa que a soma da distância é minimizada e a função objetivo de *k-Means* é alcançada (Selim and Ismail (1984)). O algoritmo *k-Means* descrito determina os *clusters* para os k escolhidos à priori. Não existe um método específico para determinar o valor de k, mas existem diferentes técnicas que podem ser utilizadas.

Número ótimo de *clusters*

Uma observação importante no início do processo *k-Means* é a necessidade de dar a entrada da quantidade de k, que representa o número de *clusters* nos dados. Essa entrada será imperativa na qualidade dos *clusters*, principalmente quando a base de dados tiver mais de três variáveis. Existem alguns métodos que validam os números de *clusters*. Um desses exemplos é o método de Elbow, que é um método visual (Kodinariya and Makwana (2013)). O método de Elbow existe com base na ideia de que se deve escolher um número de *clusters* de modo, a que a adição de outro *cluster* não forneça uma melhor modelação dos dados (Bholowalia and Kumar (2014)).

Critério de convergência do método de Elbow

A soma dos quadrados dentro dos *clusters* é representada graficamente em relação ao número de *clusters*. Os primeiros *clusters* adicionam muita informação, mas em algum momento, o ganho marginal decai drasticamente e dará um ângulo no gráfico. O "k" correto, ou seja, o número de *clusters* é escolhido neste ponto, daí o "critério de *Elbow*" (Bholowalia and Kumar (2014)). Inicialmente começa com 2 *clusters* (k=2) e continua aumentando em cada iteração em 1, calculando os *clusters* e o custo de treino (Kodinariya and Makwana (2013)). Em algum momento, o custo de encontrar o número de *clusters* (k) cairá drasticamente e, neste ponto, atingirá o valor desejado de k. Isso significa que, a partir deste ponto, o aumento do número de *clusters* leva o novo *cluster* a um ponto muito próximo de um já existente. Então, este ponto é chamado de ponto de estabilização porque é o ponto onde o critério de convergência é alcançado. A Figura 8 contém uma representação gráfica do método. Nesta representação gráfica observa-se que o ponto de estabilização é obtido em k=3.

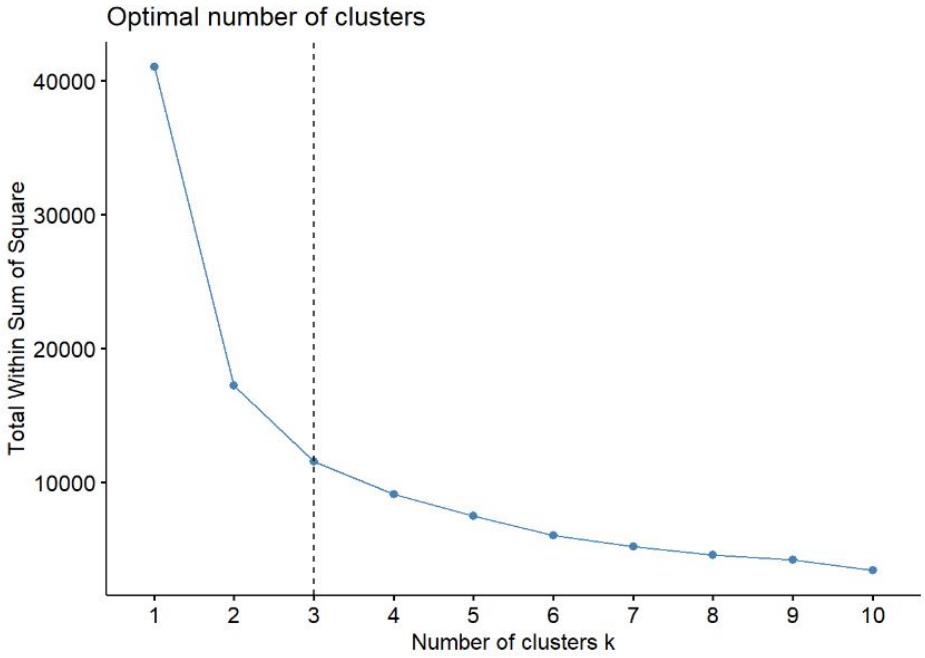

Figura 8: Representação gráfica do método de Elbow (fonte: Dias (2022))

O aumento no número de *clusters* reduz a distância aos pontos de dados, então o aumento de k diminui a métrica de *Elbow* até ao extremo de chegar a zero quando o valor de k for igual ao número de

pontos dos dados. Por esta razão, a função de k é representada graficamente como a distância média ao centróide, e o ponto do cotovelo (ponto de estabilização) é usado para determinar o valor de k.

Método da Silhueta

Outro método gráfico que permite a determinação do número ideal de *clusters* é o chamado método da silhueta. O método da silhueta média calcula a silhueta média das observações para diferentes valores de k. O número ideal de *clusters* k é aquele que maximiza a silhueta média numa dada quantidade de *clusters* k (Matt (2022)). Na Figura 9 observa-se uma representação gráfica do método. O ponto em que existe a maximização da silhueta média para uma dada quantidade de *clusters* é em k=3.

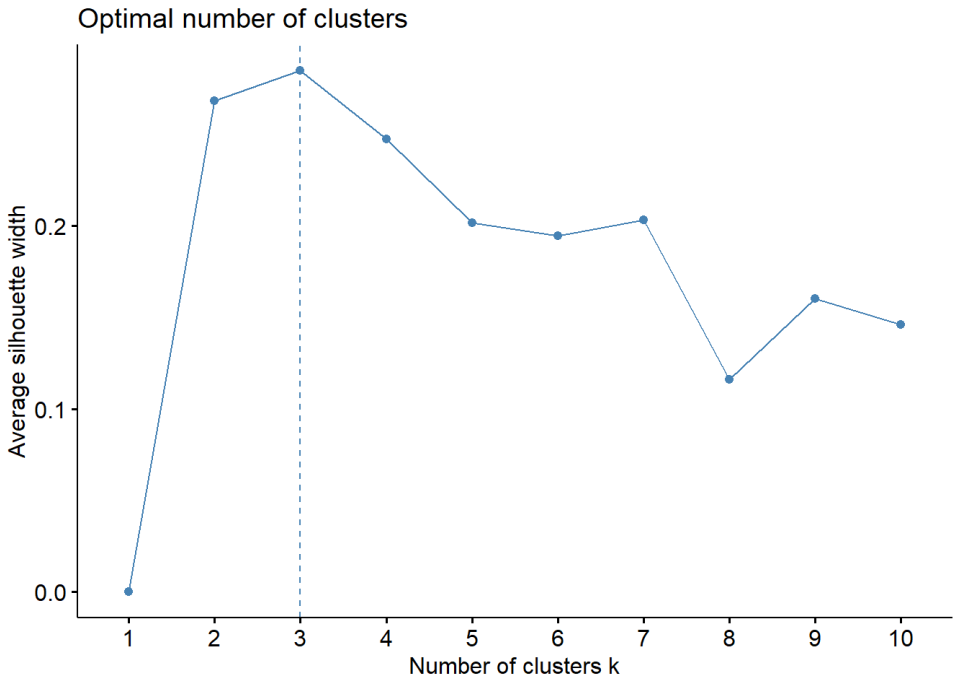

Figura 9: Método da silhueta (fonte: Ross (2022))

Vantagens do algoritmo *k-Means*

- O *k-Means* é um algoritmo de *clustering* simples, e pode ser implementado facilmente.
- O *k-Means* apenas computa e compara distância entre os pontos de dados e os *clusters* de agrupamento. Assim, pode ser computacionalmente mais rápido do que o agrupamento hierárquico, e tem tempo de complexidade $O(n)$, onde n é o número de amostras de dados.
- Pode ser aplicado a grandes conjuntos de dados.
- Pode adaptar-se facilmente a novas amostras de dados (Yse (2022)).

Desvantagens do algoritmo *k-Means*

- O número de *clusters*, k, deve ser especificado manualmente.
- Os resultados do *clustering* podem variar dependendo dos valores iniciais. O *k-Means* também seleciona aleatoriamente os centróides iniciais para os k *clusters*. Portanto, os resultados podem ser diferentes de uma execução para outra, não havendo consistência.
- O *k-Means* tem dificuldade em agrupar conjuntos de dados de tamanhos e densidades variados.
- O *k-Means* não pode identificar *outliers*. Os *outliers* ou ruídos do conjunto de dados podem afetar o processo de agrupamento, pois o *cluster* pode arrastar os *outliers* ou os *outliers* podem se tornar um *cluster* (Yse (2022)).

3.2.2 Hierárquico

Os algoritmos de agrupamento hierárquico têm como objetivo construir uma hierarquia de agrupamento. Normalmente, funcionam bem para um conjunto de dados com *nested clusters*, por exemplo dados geométricos. Começa com alguns *clusters* iniciais e gradualmente passa a convergir para a solução. O agrupamento hierárquico tem duas categorias: aglomerativa e divisiva. A abordagem aglomerativa inicialmente toma cada ponto de dados como um *cluster* individual e combina iterativamente os *clusters* até que o *cluster* final contenha todos os pontos de dados dentro dele.

A abordagem que combina os *clusters* desta forma, também é chamada de abordagem *bottom-up approach*. No agrupamento aglomerativo, as técnicas são de agrupamento divisivo e seguem o fluxo de cima para baixo que começa a partir de um único *cluster* contendo todos os pontos de dados e divide iterativamente o *cluster* em grupos menores até que cada *cluster* contenha um ponto de dados. O algoritmo de agrupamento hierárquico aglomerativo inclui as seguintes etapas.

1. Como passo inicial, o algoritmo toma cada ponto de dados como um único *cluster* e é decidida uma matriz de proximidade específica para determinar a distância entre os *clusters*. Existem quatro funções de distância disponíveis para a matriz de proximidade: *single linkage* (min), *average linkage*, *complete linkage* e *ward* (max) (Reddy (2022)). *Single linkage* significa que a distância entre dois *clusters* é definida como a distância mínima entre um ponto do primeiro *cluster* e outro ponto do segundo *cluster*. *Complete linkage* toma uma distância máxima de dois pontos de dados como a distância entre dois *clusters*. *Average linkage* calcula a distância de todos os pontos de dados do primeiro *cluster* com todos os outros do segundo *cluster*, e considera a distância média como a distância entre os *clusters*. *Ward* é semelhante à *average linkage*, exceto que usa a soma dos quadrados para calcular a distância entre os pontos.

2. Para encontrar o par de *clusters* mais próximo, é calculada a similaridade (distância) entre cada um dos *clusters*.

3. Em seguida, os *clusters* semelhantes são combinados para formar um *cluster* de acordo com a função de distância.

4. A iteração nas etapas 2 e 3 continua até que todos os pontos de dados sejam combinados num último *cluster*.

Em geral, o agrupamento hierárquico forma uma única árvore de agrupamentos onde cada nó representa os agrupamentos e cada ponto de dados começa como uma folha da árvore. A raiz da árvore é o *cluster* final contendo todos os pontos de dados. A Figura 10 mostra como o agrupamento hierárquico corta os k *clusters* do *cluster* final (árvore completa). Na figura, o algoritmo forma sucessivamente uma única árvore de *clusters* e depois corta em um determinado nível k, resultando em 4 *clusters*.

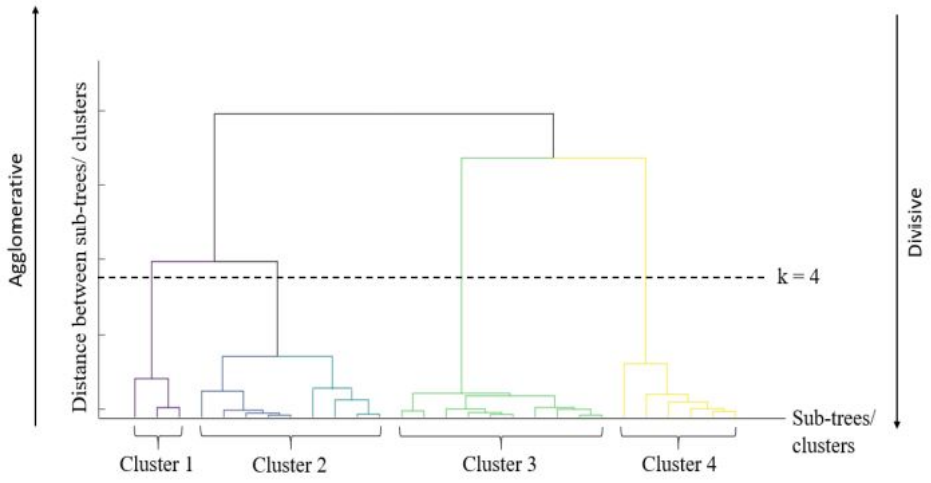

Figura 10: Dendrograma - Técnica de agrupamento do algoritmo de *clustering* hierárquico (fonte: Reddy (2022))

Vantagens do *clustering* hierárquico

- O número de *clusters* não precisa necessariamente ser especificado.
- Assim como o *k-Means*, os algoritmos de agrupamento hierárquico são fáceis de implementar.
- A estrutura hierárquica de uma árvore de *clusters* (dendrograma), pode ajudar a decidir o número de *clusters* (Naik (2022)).

Desvantagens do *clustering* hierárquico

- A principal desvantagem do agrupamento hierárquico é a sua complexidade de tempo. O agrupamento hierárquico tem uma complexidade de tempo de $O(n^2 log(n))$ (em que n é o número de pontos de dados), que é relativamente maior em comparação com outros algoritmos.
- Não há retrocessos, o que significa que uma vez que um *cluster* é criado, os pontos de dados de associação não podem ser movidos.
- Dependendo da escolha da matriz de distância, esta pode ser sensível a ruídos e *outliers*. Além disso, pode enfrentar dificuldades em lidar com *clusters* de diferentes tamanhos e formatos convexos (Naik (2022)).

Capítulo 4

Descrição dos dados e pré-processamento

Neste capítulo são apresentados os dados, e são descritas as etapas do pré-processamento de dados. Os dados que foram disponibilizados pela empresa encontram-se dispostos no formato .xls do Excel, com células agregadas na tipologia de produtos e países, necessitando de um posterior tratamento de dados. Para além do uso da ferramenta Excel, será utilizado o *software* R para análise exploratória de dados e implementação de algoritmos de *machine learning*.

A base de dados, que se pretende analisar é composta no total por 17 variáveis e 3465 observações registadas ao longo de 7 anos. A base de dados possui na sua composição 5 variáveis categóricas, entre as quais, $Referencia$, $Tipo$, $Codigo$, $Medida$ e Ano e ainda 12 variáveis quantitativas, Qt_Pt, Qt_Sp, Qt_Fr, Qt_Eng, Qt_Ger, Qt_Tot, Vl_Pt, Vl_Sp, Vl_Fr, Vl_Eng, Vl_Ger e Vl_Tot, em que Pt= Portugal, Sp= Espanha, Fr= França, Eng= Inglaterra, Ger= Alemanha e Tot= Total. Na Tabela 2 é apresentada informação sobre cada uma das variáveis.

Tabela 2: Descrição das variáveis do *dataset*

Variáveis	**Descrição**
$Referencia$	Refere-se à família de produtos, i.e. um determinado produto é caraterizado por um grupo específico
$Tipo$	Refere-se à tipologia do produto, isto é, carateriza o produto como sendo um saco ou uma caixa
$Codigo$	Permite identificar o produto na sua essência, ou seja, é atribuído um identificador único
$Medida$	Dispõe informação sobre as dimensões do produto (mm)
Ano	Identifica o ano em que os dados foram recolhidos
Qt_Pais[1]	Refere-se à quantidade de vendas realizada pela empresa de um determinado produto num dado país (unidades)
Vl_Pais[2]	Refere-se ao valor das vendas obtido pela empresa de um determinado produto num dado país (euros)

[1] $Qt\_Pais \in \{Qt\_Pt, Qt\_Sp, Qt\_Fr, Qt\_Eng, Qt\_Ger, Qt\_Tot\}$
[2] $Vl\_Pais \in \{Vl\_Pt, Vl\_Sp, Vl\_Fr, Vl\_Eng, Vl\_Ger, Vl\_Tot\}$

As variáveis quantitativas são relativas a um determinado país (Portugal, Espanha, França, Alemanha, Inglaterra), e à totalidade de vendas para todos os países. Assim para cada caso, encontramos 2 variáveis associadas (quantidade e valor de vendas).

Inicialmente, foi realizado o tratamento dos dados fornecidos no formato .xls do Excel. Realizou-se a desagregação de colunas relativas à família de produtos, desagregação de colunas relativas às quantidades e valores de vendas em cada país e remoção das colunas que não foram relevantes para análise. O tratamento de dados inicial é crucial para o carregamento de dados no *software* R, e também para que seja possível extrair informação relevante sobre um determinado produto (isolando as entradas). A empresa forneceu 7 bases de dados, em que cada base de dados é referente a um ano entre 2015 e 2021. Na Figura 11 é possível visualizar uma partição do dataset, em que se observam várias colunas agregadas.

Ref	Cor	cod.	Medida	PORTUGAL		ESPANHA		FRANÇA		INGLATERRA		ALEMANHA		TOTAL		comparação
				Qt.	valor	Qt.	valor	Qt.	valor	Qt.	valor	Qt.	valor	Qt.	valor	2013
COLORIS	Vermelho	061108	180x80x240	19 600	1 803,06 €	12 300	1 101,55 €	35 400	3 124,18 €	0	0,00 €	800	77,97 €	68 100	6 106,76 €	13%
		061008	250x100x320	15 500	2 019,77 €	11 950	1 356,04 €	26 150	2 956,13 €	0	0,00 €	500	61,84 €	54 100	6 393,78 €	-19%
		061040	320x120x420	12 150	2 142,38 €	8 300	1 265,16 €	0	0,00 €	0	0,00 €	500	83,70 €	20 950	3 491,24 €	-33%
		061224	350x140x440	500	102,34 €	3 000	531,00 €	8 250	1 394,25 €	0	0,00 €	0	0,00 €	11 750	2 027,59 €	36%
		061203	440x150x500	4 200	1 092,67 €	3 200	702,86 €	2 600	573,74 €	0	0,00 €	0	0,00 €	10 000	2 369,27 €	36%
COLORIS	Verde Claro	061101	180x80x240	12 350	1 113,64 €	12 400	1 124,76 €	60 250	5 365,84 €	0	0,00 €	400	37,57 €	85 400	7 641,81 €	-6%
		061007	250x100x320	14 300	1 814,89 €	13 350	1 539,32 €	36 250	4 206,58 €	0	0,00 €	250	30,92 €	64 150	7 591,71 €	-7%
		061039	320x120x420	2 820	501,96 €	9 550	1 471,86 €	6 500	996,75 €	0	0,00 €	250	41,85 €	19 120	3 012,42 €	-35%
		061223	350x140x440	1 050	218,39 €	2 500	440,10 €	12 750	2 184,36 €	0	0,00 €	250	50,00 €	16 550	2 892,85 €	9%
		061202	440x150x500	3 070	796,13 €	5 100	1 145,82 €	750	172,12 €	0	0,00 €	0	0,00 €	8 920	2 114,07 €	38%
COLORIS	Violeta	061215	180x80x240	4 000	385,58 €	200	17,29 €	22 250	1 931,49 €	0	0,00 €	0	0,00 €	26 450	2 334,36 €	-8%
		061192	250x100x320	2 250	285,28 €	3 200	370,04 €	16 750	1 929,31 €	0	0,00 €	0	0,00 €	22 200	2 584,63 €	-15%
		061196	320x120x420	250	44,38 €	0	0,00 €	250	45,00 €	0	0,00 €	0	0,00 €	500	89,38 €	-95%
		061230	350x140x440	250	51,17 €	750	135,00 €	7 750	1 293,24 €	0	0,00 €	0	0,00 €	8 750	1 479,41 €	8%
		061208	440x150x500	500	128,25 €	1 750	395,50 €	2 100	449,76 €	0	0,00 €	0	0,00 €	4 350	973,51 €	119%

Figura 11: Excerto da base de dados no formato .xls do Excel (Ano: 2015)

Sobre o $dataset$ original, para além da desagregação de colunas e entradas, foi removida a coluna da cor, e a coluna da comparação com os valores de vendas de anos anteriores por decisão da empresa. Para além disso, foram geradas duas novas colunas que permitem identificar as entradas da base de dados como produtos do tipo caixa, e produtos do tipo saco, e identificar o ano de registo dos valores de venda do produto. Após realizar o tratamento de dados sobre o dataset original foram geradas e guardadas novas bases de dados com o formato .csv (Figura 12).

Referencia	Tipo	Codigo	Medida	Qt_Pt	Vl_Pt	Qt_Sp	Vl_Sp	Qt_FR	Vl_Fr	Qt_Eng	Vl_Eng	Qt_Ger	Vl_Ger	Qt_Tot	Vl_Tot	Ano
ANTIQUA	CAIXA	14205	340x290x5	1315	1900.28	165	201.3	90	112.5	0	0	0	0	1570	2214.08	2015
ELISÉE CL/	CAIXA	14586	310x40x16	450	108.13	0	0	100	21.6	0	0	0	0	550	129.73	2015
PORTET CI	CAIXA	14589	295x95x3(	127	71.91	475	213.79	50	23.5	0	0	0	0	652	309.2	2015
PORTET CI	CAIXA	14592	400x80x3(	0	0	0	0	0	0	0	0	0	0	0	0	2015
ELEGANCE	CAIXA	14599	380X310X(	0	0	0	0	0	0	0	0	0	0	0	0	2015
ELEGANCE	CAIXA	14600	380X310X(	0	0	0	0	0	0	0	0	0	0	0	0	2015
ELEGANCE	CAIXA	14601	380X310X(	0	0	0	0	0	0	0	0	0	0	0	0	2015
ELEGANCE	CAIXA	14602	460X360X8	0	0	0	0	0	0	0	0	0	0	0	0	2015
ELEGANCE	CAIXA	14603	460X360X8	0	0	0	0	0	0	0	0	0	0	0	0	2015
ELEGANCE	CAIXA	14604	460X360X8	0	0	0	0	0	0	0	0	0	0	0	0	2015
VINTAGE	CAIXA	14616	350x290x1	546	1549.99	15	38.7	0	0	0	0	0	0	561	1588.69	2015
GOURMET	CAIXA	14618	180x95x1(	438	480.24	0	0	0	0	0	0	0	0	438	480.24	2015
GOURMET	CAIXA	14619	180x95x1(	333	358.48	0	0	0	0	0	0	0	0	333	358.48	2015
GOURMET	CAIXA	14621	205x205x1	499	651.48	15	17.1	0	0	0	0	0	0	514	668.58	2015

Figura 12: Excerto da base de dados no formato .csv do Excel (Ano: 2015)

As variáveis explicativas deste estudo são $Referencia$, $Tipo$, $Medida$, $Codigo$ e Ano, e as variáveis de interesse são Qt_Pais e Vl_Pais. As bases de dados de todos os anos foram carregadas no ambiente $\mathbb{R}$, e foram unidas numa única base de dados com recurso ao comando $rbind()$. A base de dados original foi dividida em duas novas bases de dados (uma base de dados para as caixas, e outra base de dados para sacos) através da variável $Tipo$. Na Figura 13 está presente um gráfico de barras que apresenta o número de observações de caixas e de sacos na base de dados original. O *dataset* é composto por 1294 observações referentes a caixas, e 2171 observações referentes a sacos.

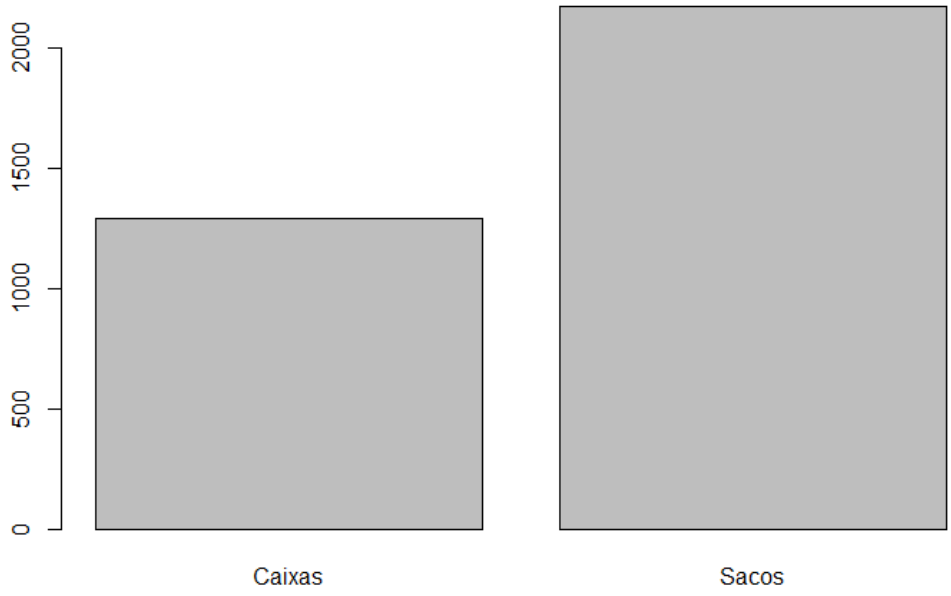

Figura 13: *Barplot* - Nº de observações de caixas e sacos na base de dados

De modo geral, o comportamento de vendas na empresa Litel, Lda, tem sido distinto ao longo do tempo. De 2015 a 2017 existiu um decaimento na quantidade de vendas. De 2017 a 2019 é possível observar que houve um aumento na quantidade de vendas dos produtos. De 2019 a 2020 houve novamente um decréscimo na quantidade de vendas, e em 2021 houve um pico máximo na quantidade de vendas. Na Figura 14, está presente o comportamento de vendas totais na empresa Litel, Lda, de todos os produtos (caixas e sacos) ao longo do tempo.

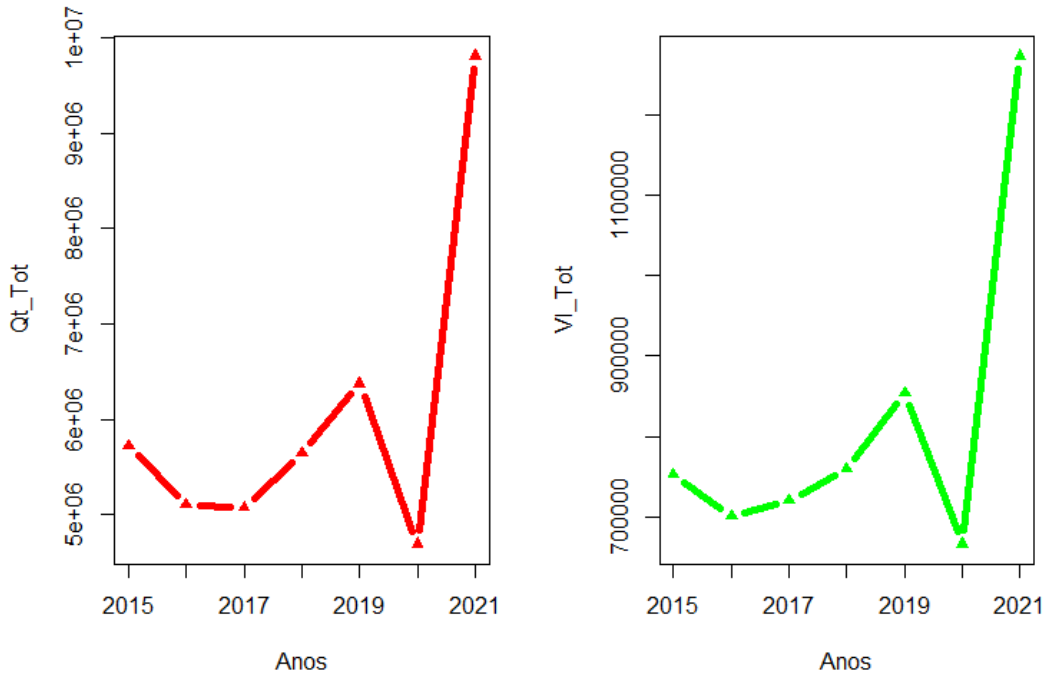

Figura 14: *Plot* de quantidades totais e valores totais de vendas (resp.gráfico da esquerda e gráfico da direita) de todos os produtos ao longo do tempo

O $spaghetti\ plot$ da Figura 15 mostra que existem 6 produtos na empresa que ao longo do tempo tiveram quantidades de vendas acima das 200000 unidades. Os restantes produtos foram vendidos sensivelmente a baixo das 200000 unidades por ano.

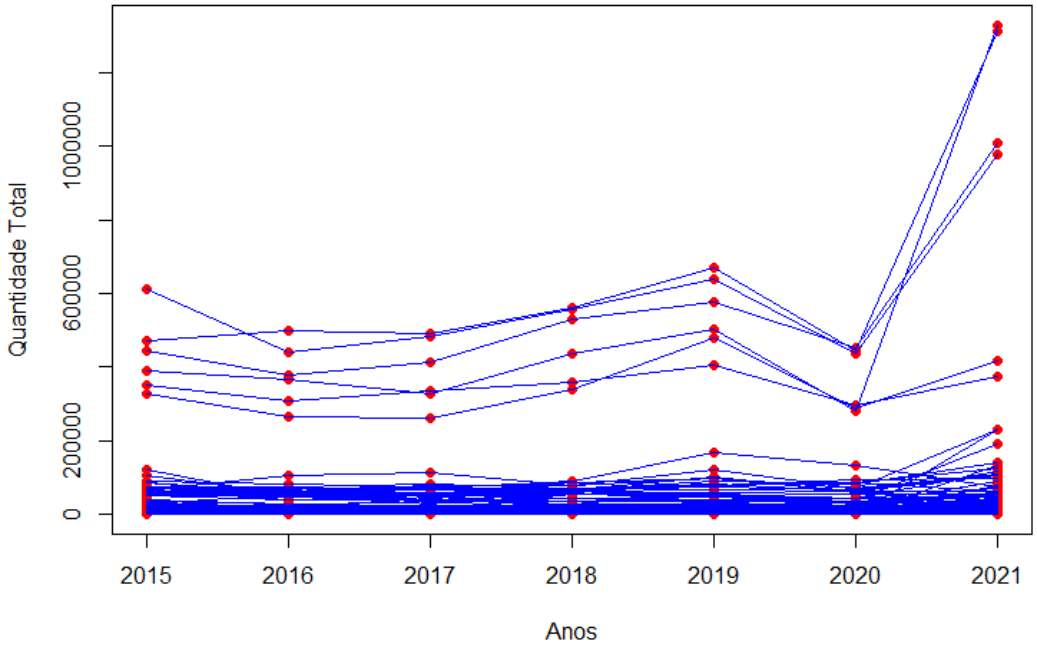

Figura 15: $Spaghetti\ Plot$ do comportamento dos produtos ao longo do tempo

Capítulo 5

Análise exploratória dos dados

5.1 Base de dados das caixas

A Tabela 3, inclui as diversas categorias da variável $Referencia$, e a quantidade de observações para cada categoria na base de dados das caixas.

Tabela 3: Número de observações das categorias da variável $Referencia$ (base de dados das caixas)

Categoria	**Nº obs**	**Categoria**	**Nº obs**	**Categoria**	**Nº obs**	**Categoria**	**Nº obs**
AFFERO	42	CUBE	34	GOURMET HAUTE M	42	GOURMET TARTAN S	2
AFFERO TARTAN	6	CUBE XMAS	4	GOURMET HAUTE S	42	GOURMET TARTAN XL	2
AFFERO XMAS	6	DELICIOUS	20	GOURMET L	28	GOURMET TRIANGULUM	10
ALBUS	35	DOMINUS com janela	10	GOURMET LINGOT	10	GOURMET XL	28
ANTIQUA	7	DOMINUS sem janela	15	GOURMET M	28	GRAN GOURMET	8
ARX	7	DURIUS	10	GOURMET NATURAE L	2	KRAFT	35
AVEL	42	ELEGANCE	22	GOURMET NATURAE M	2	LINEAS AMARELO	35
B2C PACKAGING	70	ELISEE CLASSIC	14	GOURMET NATURAE S	2	LINEAS OURO	25
BOTTLE B2C	21	ELISEE HYPE	112	GOURMET NATURAE XL	2	LINEAS PRETO	35
CALIX com janela	10	FILUM	20	GOURMET S	28	LINEAS VERDE	35
CALIX sem janela	15	GOURMET DOMINUS	16	GOURMET TARTAN L	2	LINEAS VERMELHO	35
CLASSIC	35	GOURMET HAUTE L	42	GOURMET TARTAN M	2	MOORISH	15
NODUS com janela	10	NODUS sem janela	15	PORTET CLASSIC	28	SCOTUS GIFT	15
SMART BOX	30	SPARKLING	35	STANDARD	49	SUPPORTO	21
VINTAGE	21						

De modo, a entender melhor os dados, torna-se fundamental conhecer algumas estatísticas descritivas. Na Tabela 4, são apresentadas estatísticas descritivas das variáveis qualitativas do *dataset* de caixas.

Tabela 4: Estatísticas descritivas das variáveis qualitativas (base de dados das caixas)

	Referencia	**Medida**	**Codigo**	**Ano**
Moda	"ELISEE HYPE"	"180x90x400", "270x90x400"	"15470"	"2020", "2021"
Nº de categorias	57	92	213	7

De acordo com a Tabela 4, verifica-se que a variável $Referencia$ contém cerca de 57 categorias distintas de caixas, e a referência que apresenta maior número de observações é a "*ELISEE HYPE*". Existem 92 medidas distintas de caixas, e as dimensões mais frequentes são "180x90x400", e "270x90x400". No que respeita, à variável $Codigo$, verifica-se que existem na base de dados 213 tipos de caixas diferentes, sendo que a caixa com o código "15470", é a que apresenta um maior número de observações. A base de dados remete a dados recolhidos durante um período de 7 anos, sendo que 2020 e 2021 foram os anos em que se obteve um maior número de observações. A Tabela 5 indica o número de observações da base de dados das caixas em cada ano.

Tabela 5: Número de observações em cada ano (base de dados das caixas)

	2015	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**
Nº de observações	160	160	198	183	183	205	205

Na Tabela 6 e Tabela 7 são apresentadas as principais estatísticas descritivas das variáveis quantitativas.

Tabela 6: Descrição estatística das variáveis Qt_Pais (base de dados das caixas)

	Qt_Pt	Qt_Sp	Qt_Fr	Qt_Eng	Qt_Ger	Qt_Tot
Mínimo	0,00	-2200,00	-210,00	0,00	0,00	-25,00
1º Quartil	10,00	0,00	0,00	0,00	0,00	75,00
Mediana	200,00	10,00	0,00	0,00	0,00	441,50
Média	589,50	332,10	425,00	0,00	2,06	1528,90
3º Quartil	750,00	200,00	2,00	0,00	0,00	1425,00
Máximo	8101,00	27350,00	86715,00	0,00	800,00	89690,00
Desvio Padrão	987,75	1261,04	3285,54	0,00	26,46	4116,87
Variância	975641,20	1590218,00	10794751,00	0,00	700,18	16948616,00

Tabela 7: Descrição estatística das variáveis Vl_Pais (base de dados das caixas)

	Vl_Pt	Vl_Sp	Vl_Fr	Vl_Eng	Vl_Ger	Vl_Tot
Mínimo	-0,69	-1016,40	-134,41	0,00	0,00	-64,25
1º Quartil	6,22	0,00	0,00	0,00	0,00	47,56
Mediana	124,18	6,97	0,00	0,00	0,00	277,88
Média	368,01	178,69	253,52	0,00	0,87	960,67
3º Quartil	459,59	127,62	0,79	0,00	0,00	986,99
Máximo	8977,03	8162,50	64630,23	0,00	330,40	67113,54
Desvio Padrão	637,43	497,62	2160,49	0,00	10,71	2553,56
Variância	406314,40	247628,80	4667731,00	0,00	114,78	6520644,00

Através da análise da Tabela 6 e Tabela 7, observa-se que ao longo dos setes anos em que foram recolhidos os dados, não se realizaram vendas de caixas para a Inglaterra. Verifica-se ainda a existência de notas de crédito na base de dados (correspondente aos valores mínimos negativos). A base de dados não possui *missing values*, e encontra-se devidamente preenchida em todos os campos.

Na Figura 16 e Figura 17, estão presentes histogramas e *boxplots* para as variáveis Qt_Pt e Qt_Tot . Pela análise da variável Qt_Pt, observa-se que existe um grande número de observações com quantidade de vendas baixas, e um reduzido número de observações quando a quantidade de vendas é alta. O raciocínio é análogo para a variável Qt_Tot . As variáveis Vl_Pt e Vl_Tot mostram que existe uma forte relação com Qt_Pt e Qt_Tot (respetivamente), isto é, existem poucas observações com elevado valor de vendas, e muitas observações com um baixo valor de vendas (Figura 18 e Figura 19).

Os histogramas e *boxplots* das restantes variáveis quantitativas, encontram-se na secção A dos anexos desta dissertação.

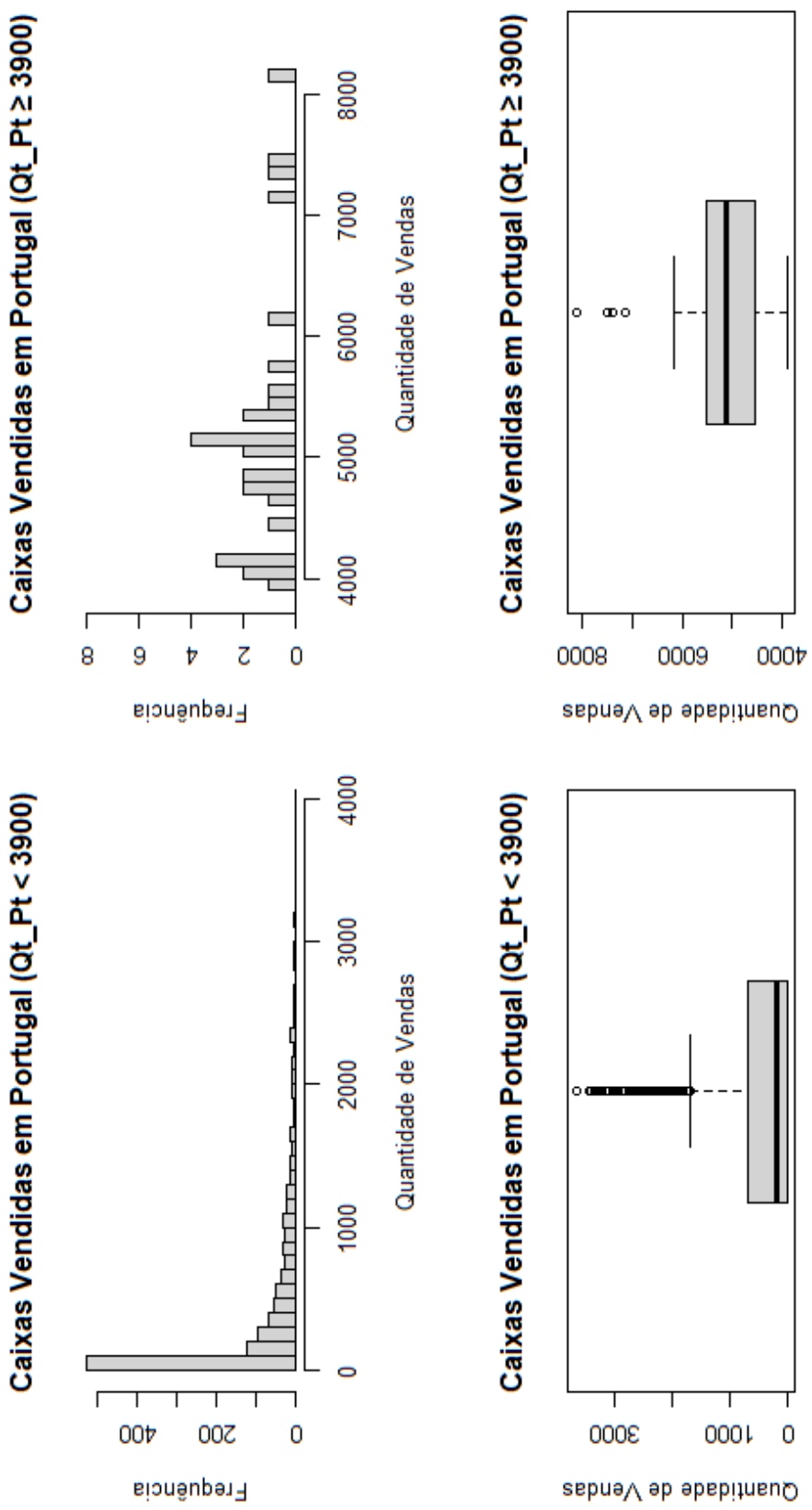

Figura 16: Histograma e *boxplot* para a variável Qt_Pt

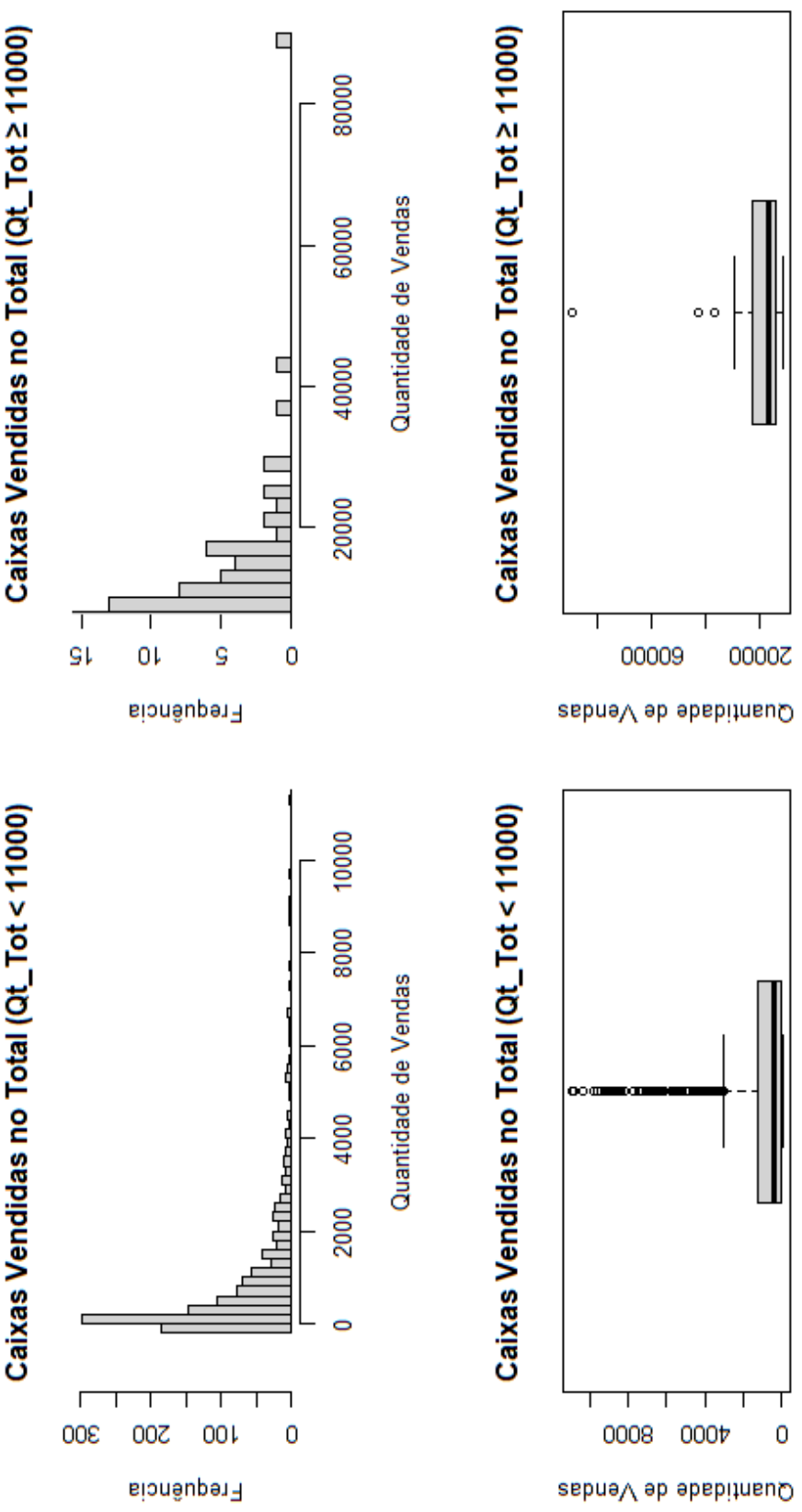

Figura 17: Histograma e *boxplot* para a variável Qt_Tot

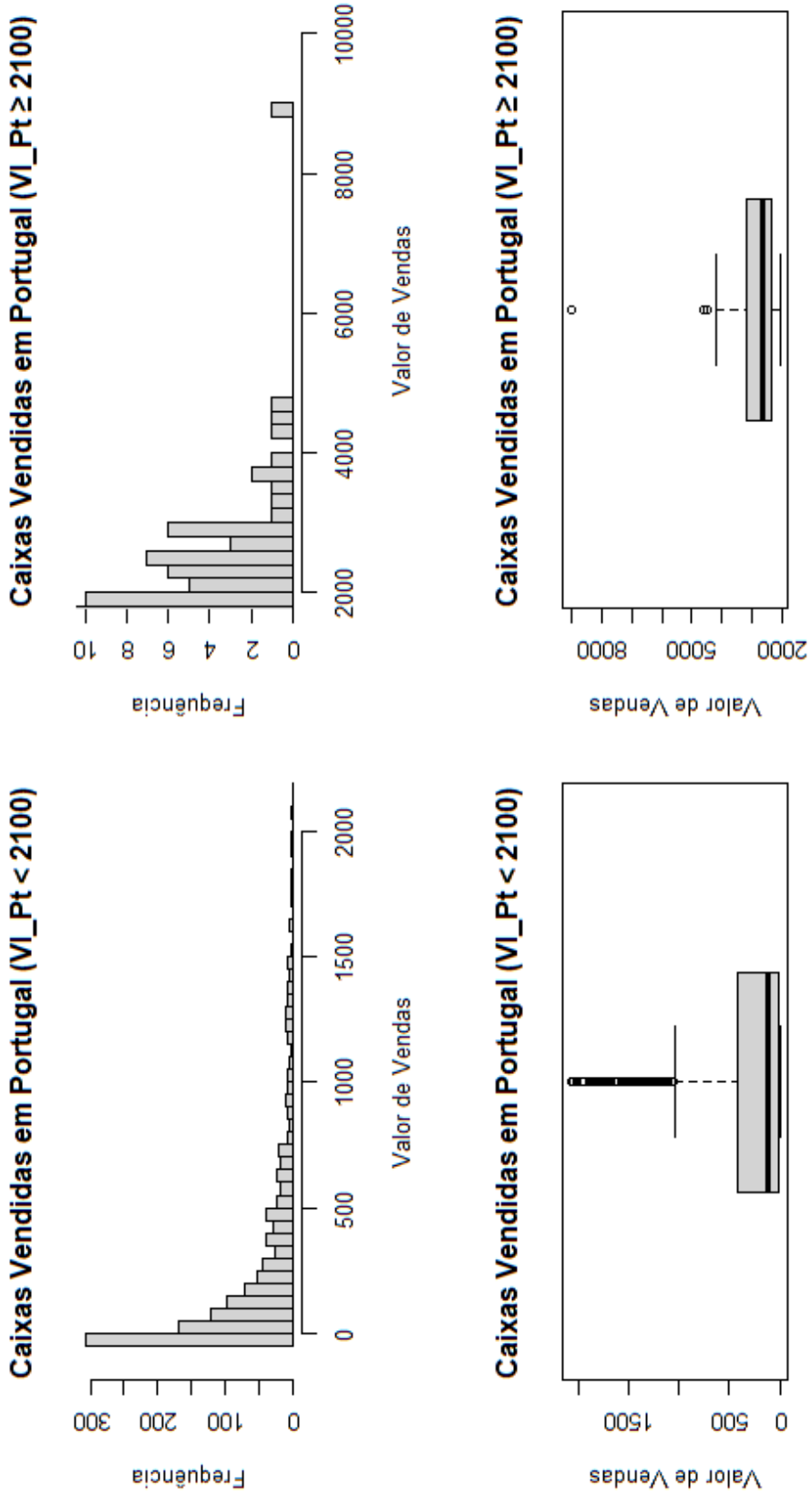

Figura 18: Histograma e *boxplot* para a variável Vl_Pt

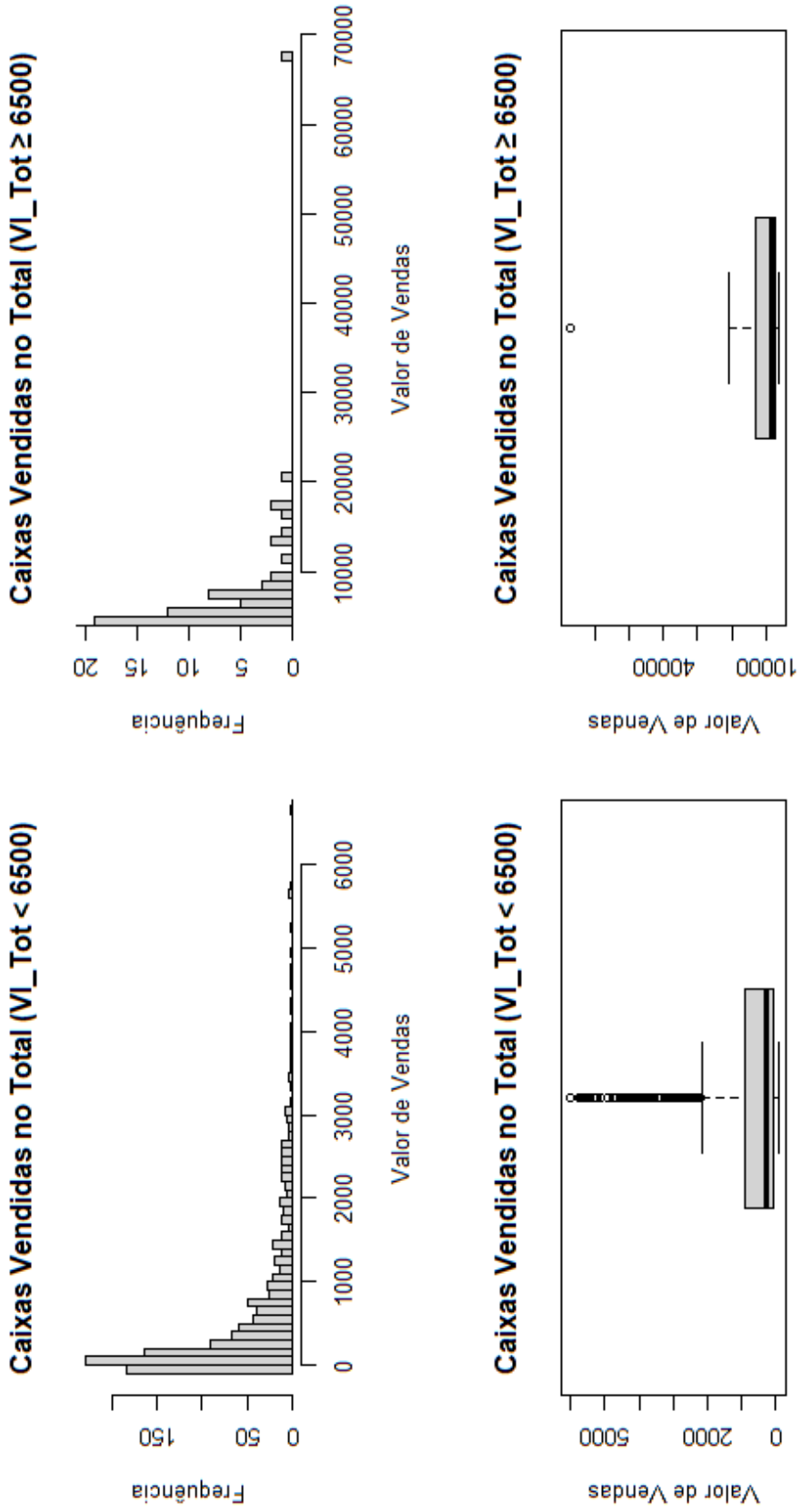

Figura 19: Histograma e *boxplot* para a variável Vl_Tot

5.1.1 Top de vendas

O $spaghetti\ plot$ da Figura 20 apresenta o comportamento das vendas de caixas ao longo do tempo. Visualmente, consegue-se observar que existem caixas que são mais vendidas que as restantes. E existe uma determinada caixa que obteve um pico máximo de vendas no ano 2021, correspondente a mais de 80000 unidades vendidas. De modo, a conseguir identificar o comportamento de vendas de uma determinada caixa, é apresentado de seguida um top de vendas, em que consta as caixas mais vendidas e menos vendidas, enquadrado num certo nível de vendas. Um nível de vendas remete a um intervalo de quantidade de vendas, no caso, representa uma porção de vendas que é analisada, extraindo a informação das caixas que se localizam nos valores mínimos e máximos desses intervalos. A base de dados das caixas foi repartida, em três níveis de vendas:

- Nível 1 - caixas vendidas em menor quantidade (Qt_Tot < 1000)
- Nível 2 - caixas vendidas em quantidade intermédia (1000 $\leq Qt\_Tot$ < 10000)
- Nível 3 - caixas vendidas em maior quantidade (Qt_Tot > 10000)

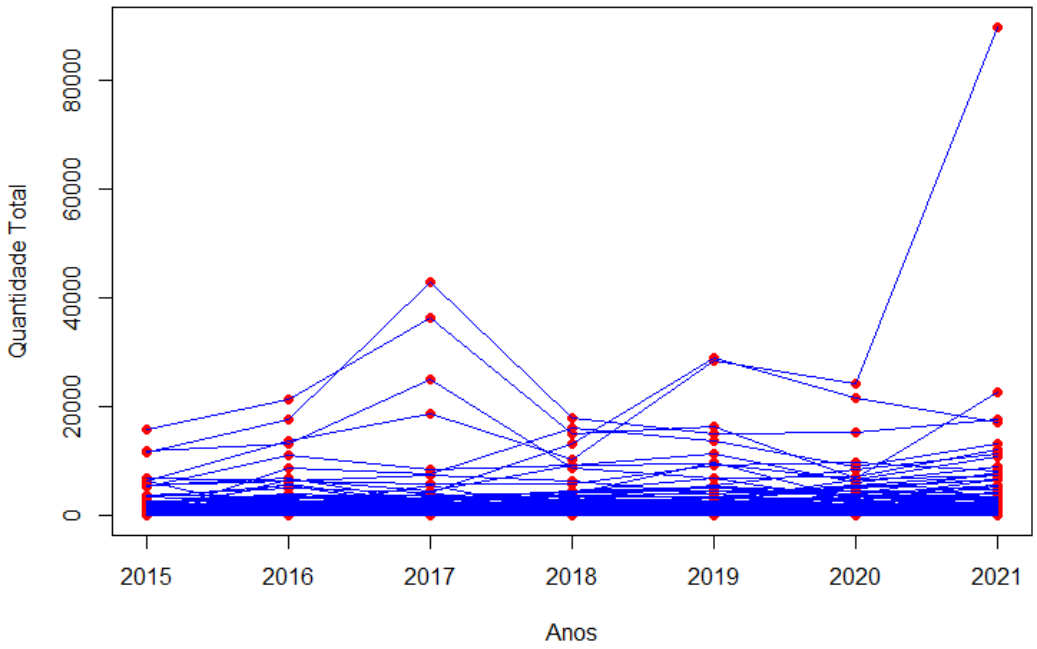

Figura 20: $Spaghetti\ Plot$ do comportamento das caixas ao longo do tempo

Na Figura 21 observa-se que no nível 1 de vendas a caixa *B2C PACKAGING (15798)* é a que mais se destaca com uma ascendente evolução nas vendas ao longo do tempo. Em 2015 foram vendidas menos de 2000 unidades, e desde aí tem tido uma tendência crescente de vendas ao longo dos anos, alcançando um pico de vendas com mais de 6000 unidades vendidas em 2021. Neste TOP5 pode-se referir que as caixas *SPARLING(15419)*, *ELISÉE HYPE (15075) e VINTAGE (14616)*, têm tido vendas mais estáveis ao longo do tempo.

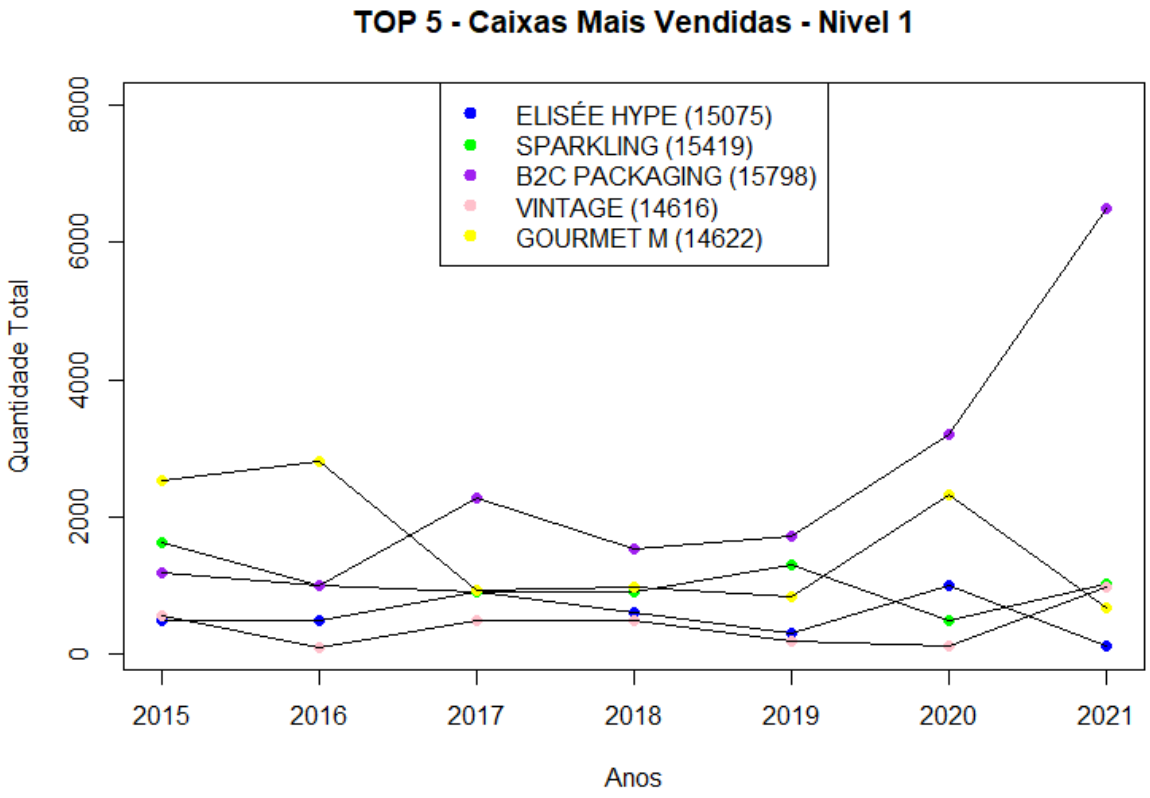

Figura 21: Caixas mais vendidas (Nível 1 de Vendas)

No nível 2 de vendas, é notório que as caixas que fazem parte do TOP5, têm comportamentos distintos. Por exemplo, a caixa *KRAFT (15834)* e a *KRAFT (15836)* possuíam em 2015 quantidades nulas de vendas, e aumentaram bastante as suas vendas ao longo do tempo. 2017 e 2020, destacam-se como anos em que houve um ligeiro decréscimo nas quantidades de vendas destes dois produtos. A caixa *LINEAS PRETO (15467)*, apresentou um pico mínimo de vendas em 2016, e até 2019 as suas vendas aumentaram, apresentando de seguida um decréscimo em 2020 (Figura 22).

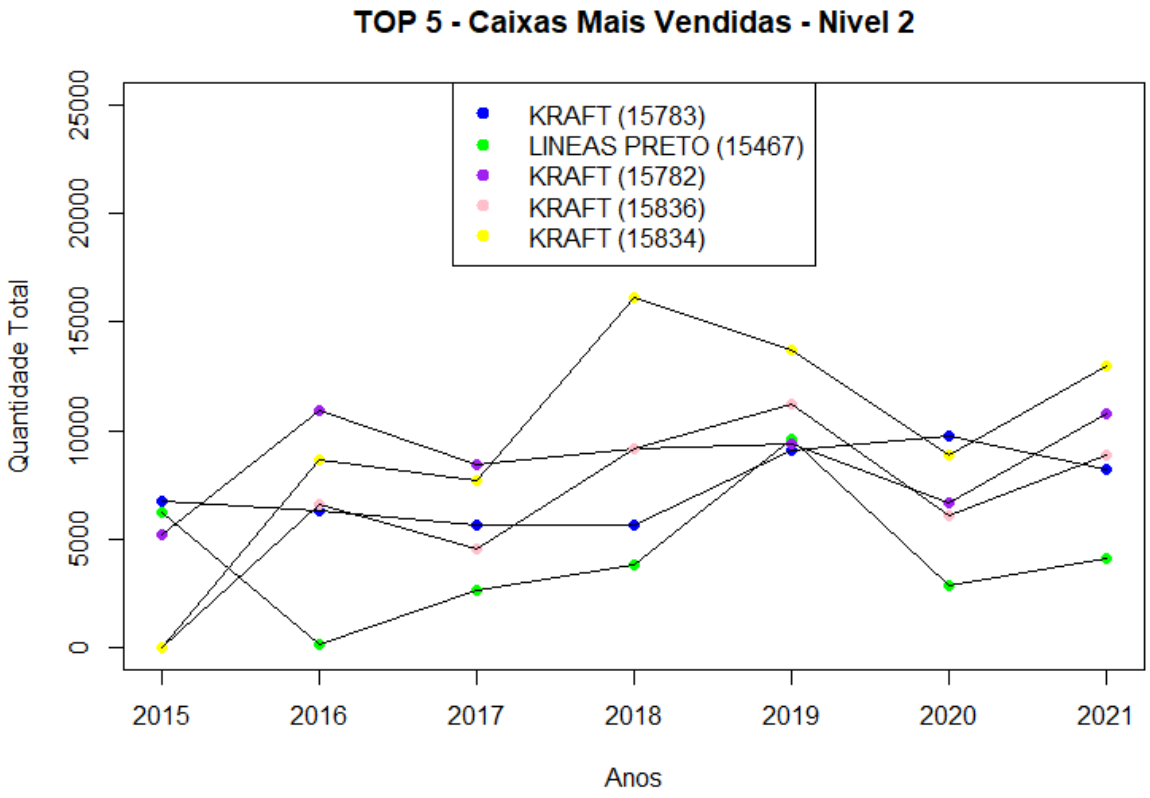

Figura 22: Caixas mais vendidas (Nível 2 de Vendas)

No nível 3 de vendas pode ser destacado o aumento expressivo das vendas da caixa *GOURMET M (14766)* de 2015 para 2021. Este produto passou de menos de 20000 unidades vendidas por ano para mais de 80000 unidades por ano. Os demais produtos não detém variações bruscas ao longo do tempo. A caixa *CLASSIC (15434)*, em 2015 era o segundo produto mais vendido deste TOP5, e passou em 2021 a ser dos produtos menos vendidos, mantendo um comportamento relativamente estável ao longo do tempo (Figura 23).

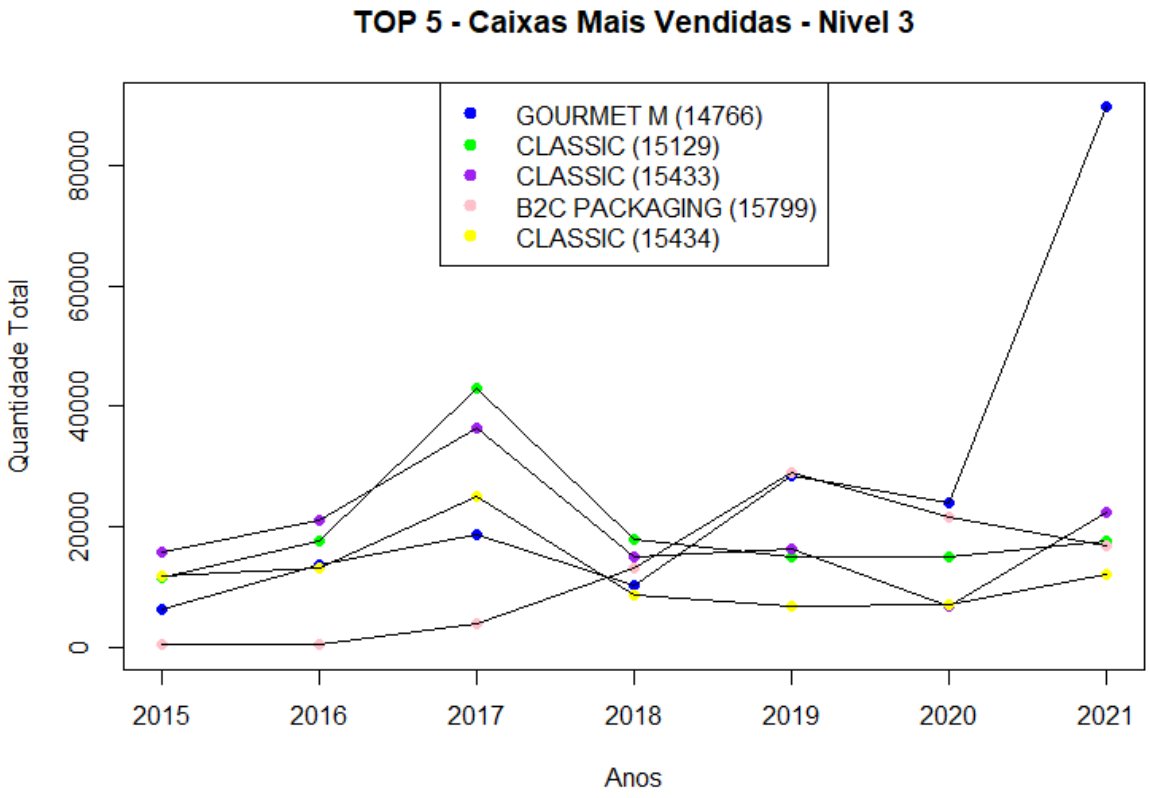

Figura 23: Caixas mais vendidas (Nível 3 de Vendas)

Quanto ao TOP5 das caixas menos vendidas no nível 1, pode ser destacado que a caixa *LINEAS OURO (15468)* foi dos produtos mais vendidos neste TOP5, aumentou as suas vendas no ano 2016, e sofreu um descréscimo acentuado em 2017 para valores praticamente nulos. Esses valores mantiveram-se durante os anos que se seguiram. Em 2018 destaca-se uma tendência nula de vendas para todos os produtos deste TOP5. Com a exceção da caixa *LINEAS OURO (15468)*, os restantes produtos mantiveram as suas vendas a baixo das 750 unidades ao longo do tempo (Figura 24).

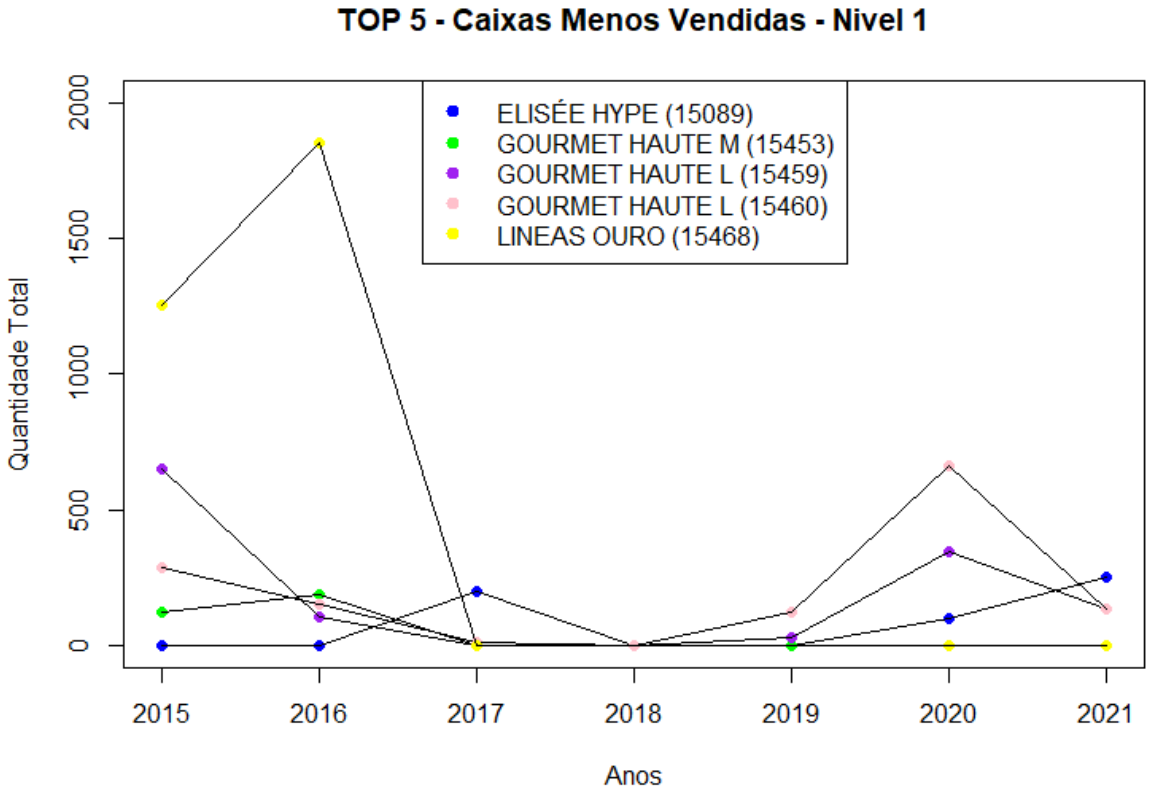

Figura 24: Caixas menos vendidas (Nível 1 de Vendas)

No TOP5 de caixas menos vendidas no nível 2, observa-se que apenas as caixas *SPARKLING (15419)* e *PORTET CLASSIC (14746)* têm observações em todos os anos. Neste TOP5 destacam-se quantidades mínimas de vendas para as caixas *CALIX sem janela (16389)* e *NODUS sem janela (16386)*, em 2018 e 2021 respetivamente. As únicas duas observações presentes da caixa *AFFERO XMAS (16981)* representam alguma estabilidade nas vendas deste produto (Figura 25).

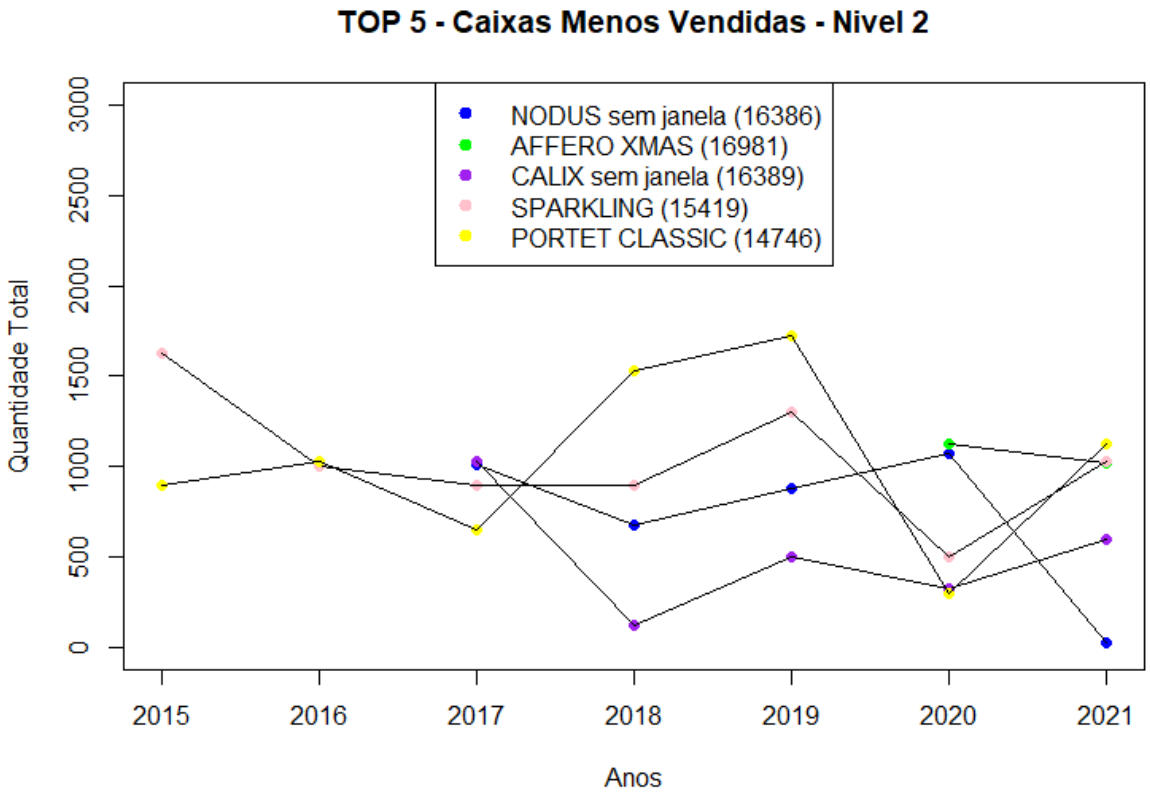

Figura 25: Caixas menos vendidas (Nível 2 de Vendas)

Através da Figura 26 é possível observar que no nível 3 de vendas as caixas menos vendidas são as caixas *KRAFT (15836)*, *B2C PACKAGING (15804)* e *KRAFT (15782)*. Neste TOP5, verifica-se que as caixas que têm tido uma maior heterogeneidade de comportamentos ao longo do tempo são as caixas *GOURMET M (14766)* e *CLASSIC(15129)*. No caso da primeira verifica-se que os principais picos de vendas ocorreram em 2019 e 2021, sendo que em 2021 ocorreu um pico máximo de vendas. No caso da segunda, o principal pico de vendas ocorreu em 2017.

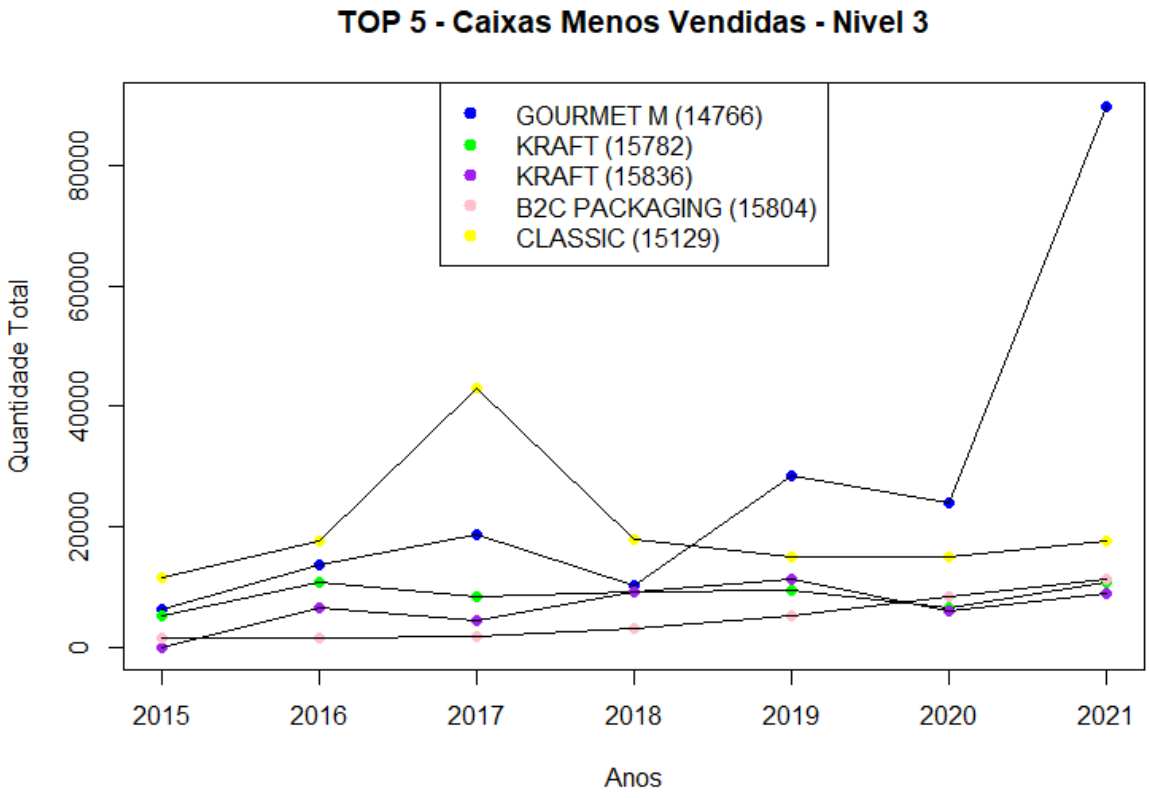

Figura 26: Caixas menos vendidas (Nível 3 de Vendas)

5.2 Base de dados dos sacos

A Tabela 8, inclui as diversas categorias da variável $Referencia$, e a quantidade de observações para cada categoria na base de dados dos sacos.

Tabela 8: Número de observações das categorias da variável $Referencia$ (base de dados dos sacos)

Categoria	**Nº obs**	**Categoria**	**Nº obs**	**Categoria**	**Nº obs**	**Categoria**	**Nº obs**
BACUS	42	CHRISTMAS NATURA	49	GEMINUS 1	25	NATURA	36
BLOOM	15	COLORIS	487	KRAFT GIFT	84	NATURA VIDE	14
BOTTLE B2C	7	COLORIS GIFT	140	KRAFT LISO	302	PRISMA	14
BOUTIQUE	10	COLORIS GLOSS	22	KRAFT PURUS	21	SACOS TAKE AWAY	12
BOW	24	COLORIS WIDE	98	KRAFT VERJURADO	192	SALDOS	14
CHIC	10	CONCEPTA	60	LEATHER	15	SCOTUS	15
CHRISTMAS CLASSIC	49	DIVISORIA	7	LUXURY	60	TEMPORA ANNI	30
CHRISTMAS FUN	49	ECO	42	LUXURY BASIC	20	VALENTIN	14
CHRISTMAS SHINY	28	FASHION	10	LUXURY KRAFT	20	VINIS	14
CHRISTMAS SNOWMAN	49	GEMINUS	21	LUXUS	36	VINIS CLASSIC	7
WINE	7						

De seguida, são apresentadas estatísticas descritivas das variáveis qualitativas do *dataset* de sacos.

Tabela 9: Estatísticas descritivas das variáveis qualitativas (base de dados dos sacos)

	Referencia	**Medida**	**Codigo**	**Ano**
Moda	"COLORIS"	"250x100x320"	"61472", "61473", "61474"	"2017"
Nº de categorias	41	50	347	7

De acordo com a Tabela 9, verifica-se que a variável $Referencia$ contém cerca de 41 categorias distintas de sacos, e a referência que apresenta maior número de observações é a *"COLORIS"*. Existem 50 medidas distintas de sacos, e as dimensões mais frequentes são "250X100X320". No que respeita, à variável $Codigo$, verifica-se que existem na base de dados 347 tipos de sacos diferentes, sendo que os sacos com os códigos "61472", "61473", "61474"apresentam um número de observações superior aos demais. A base de dados remete a dados recolhidos durante um período de 7 anos, sendo que 2017 foi o ano em que se obteve um maior número de observações. A Tabela 10 indica o número de observações da base de dados dos sacos em cada ano.

Tabela 10: Número de observações em cada ano (base de dados dos sacos)

	2015	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**
Nº de observações	296	296	343	289	289	329	329

De modo a entender o comportamento das variáveis de interesse do *dataset* de sacos, são apresentadas de seguida tabelas com as principais estatísticas descritivas das variáveis quantitativas:

Tabela 11: Descrição estatística das variáveis Qt_Pais (base de dados dos sacos)

	Qt_Pt	Qt_Sp	Qt_Fr	Qt_Eng	Qt_Ger	Qt_Tot
Mínimo	0,00	0,00	-250,00	0,00	0,00	0,00
1º Quartil	0,00	0,00	0,00	0,00	0,00	211,00
Mediana	800,00	75,00	0,00	0,00	0,00	2000,00
Média	6467,00	4191,00	6878,00	752,10	33,77	18626,00
3º Quartil	3250,00	1500,00	1500,00	0,00	0,00	7802,00
Máximo	271036,00	254133,00	1054450,00	120000,00	20000,00	1329250,00
Desvio Padrão	25373,83	20067,44	39132,55	5864,04	456,49	77653,31
Variância	643831364,00	402702222,00	1531356607,00	34386973,00	208382,90	6030036857,00

Tabela 12: Descrição estatística das variáveis Vl_Pais (base de dados dos sacos)

	Vl_Pt	Vl_Sp	Vl_Fr	Vl_Eng	Vl_Ger	Vl_Tot
Mínimo	0,00	0,00	-61,50	0,00	0,00	0,00
1º Quartil	0,00	0,00	0,00	0,00	0,00	53,90
Mediana	165,00	19,20	0,00	0,00	0,00	433,00
Média	757,30	472,30	685,10	89,78	4,01	2064,30
3º Quartil	577,50	253,20	258,20	0,00	0,00	1291,40
Máximo	25498,10	23807,10	67369,80	16416,00	1626,80	108995,20
Desvio Padrão	2401,52	1886,09	2951,92	749,38	39,96	6775,49
Variância	5767305,00	3557326,00	8713842,00	561574,90	1596,66	45907319,00

Através da análise da Tabela 11 e 12, verifica-se a existência de notas de crédito na base de dados (correspondente aos valores mínimos negativos). A base de dados não possui *missing values*, e encontra-se devidamente preenchida em todos os campos.

Nas Figuras 27 e 28, estão presentes histogramas e *boxplots* para as variáveis Qt_Pt e Qt_Tot . Pela análise da variável Qt_Pt, observa-se que existe um grande número de observações com quantidade de vendas baixas, e um reduzido número de observações quando a quantidade de vendas é alta. O raciocínio é análogo para a variável Qt_Tot . As variáveis Vl_Pt e Vl_Tot mostram que existe uma forte relação com Qt_Pt e Qt_Tot (respetivamente), isto é, existem poucas observações com elevado valor de vendas, e muitas observações com um baixo valor de vendas (Figuras 29 e 30).

Os histogramas e *boxplots* das restantes variáveis quantitativas, encontram-se na secção A dos anexos desta dissertação.

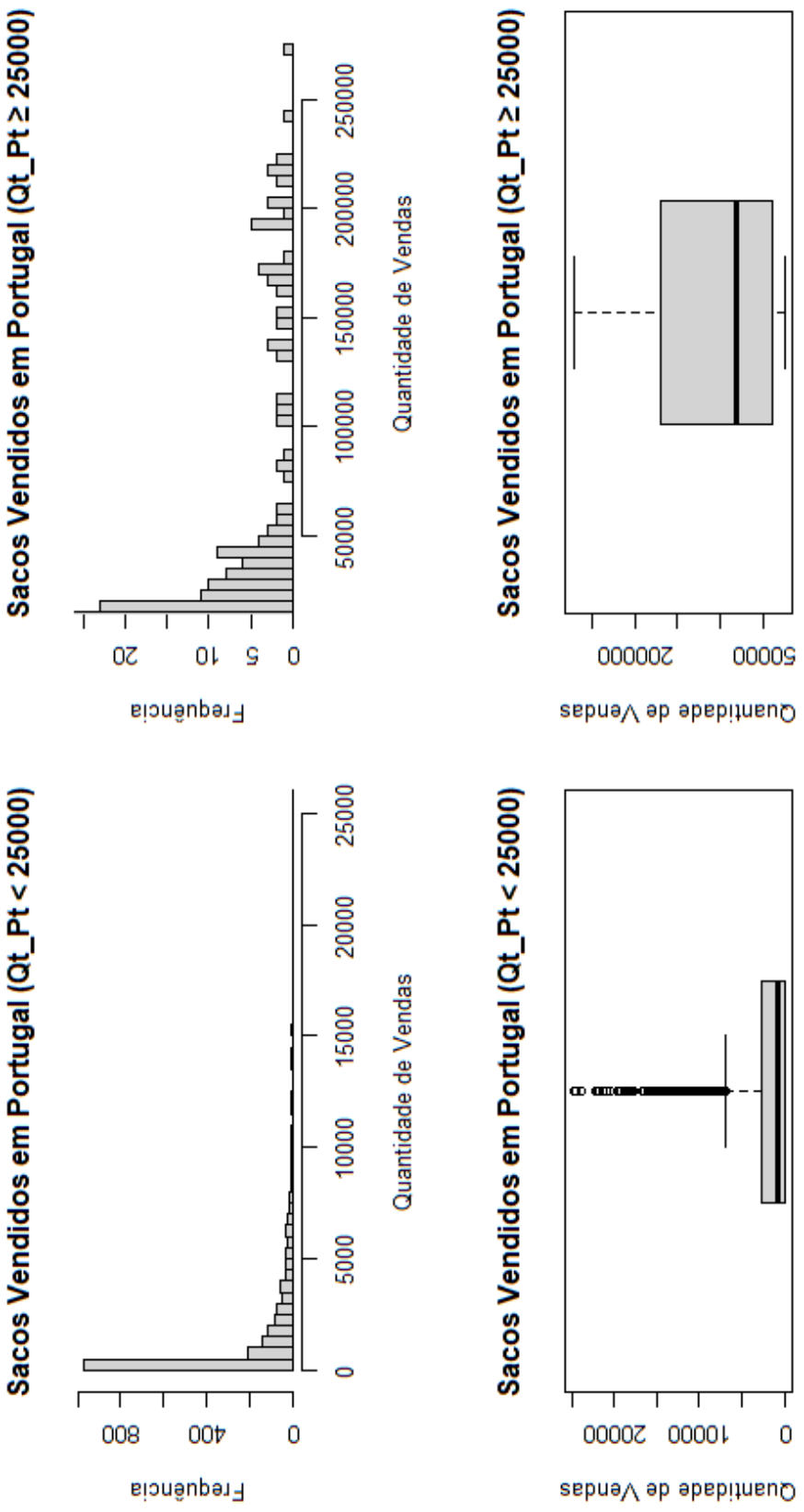

Figura 27: Histograma e *boxplot* para a variável Qt_Pt

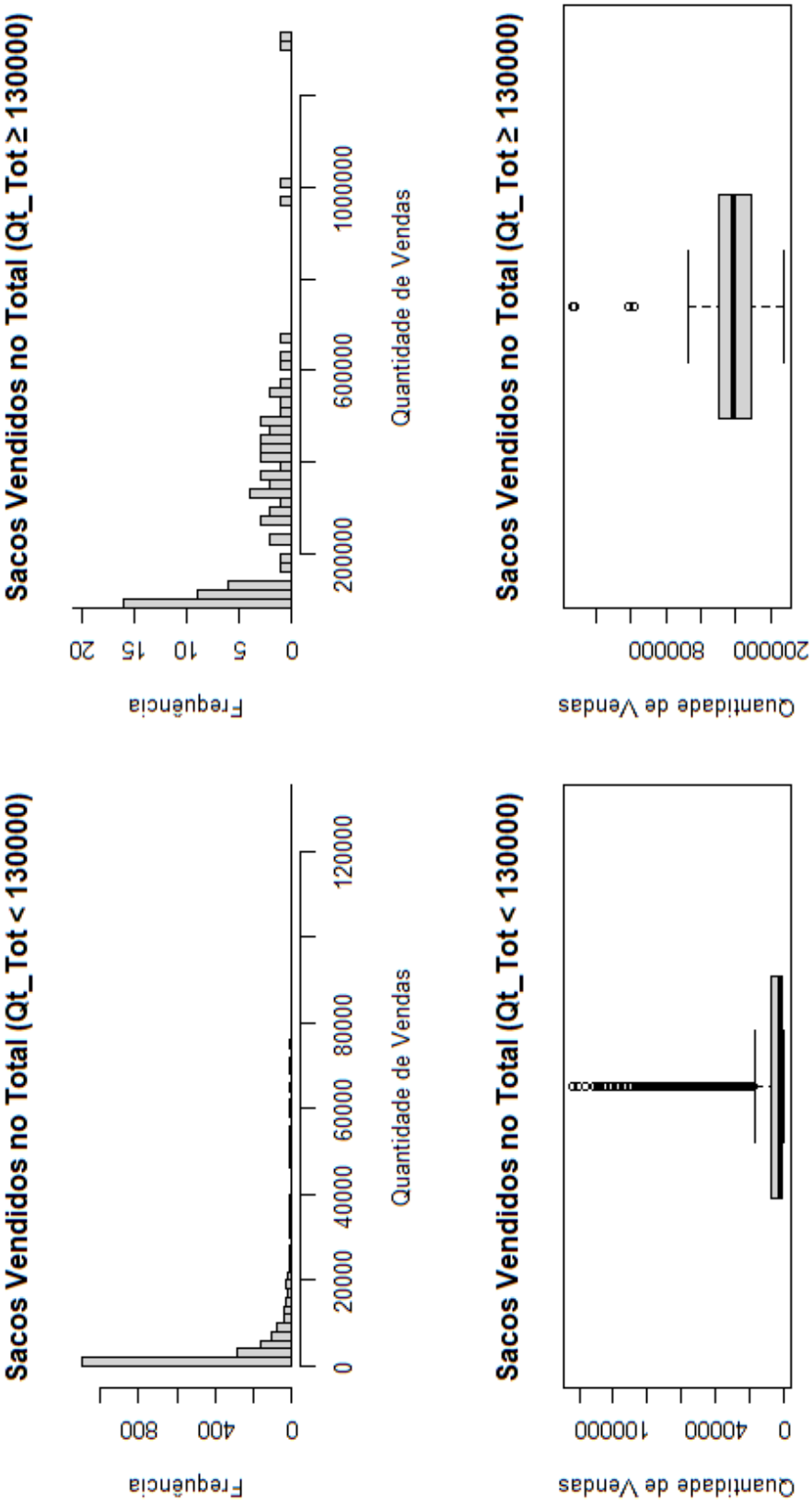

Figura 28: Histograma e *boxplot* para a variável Qt_Tot

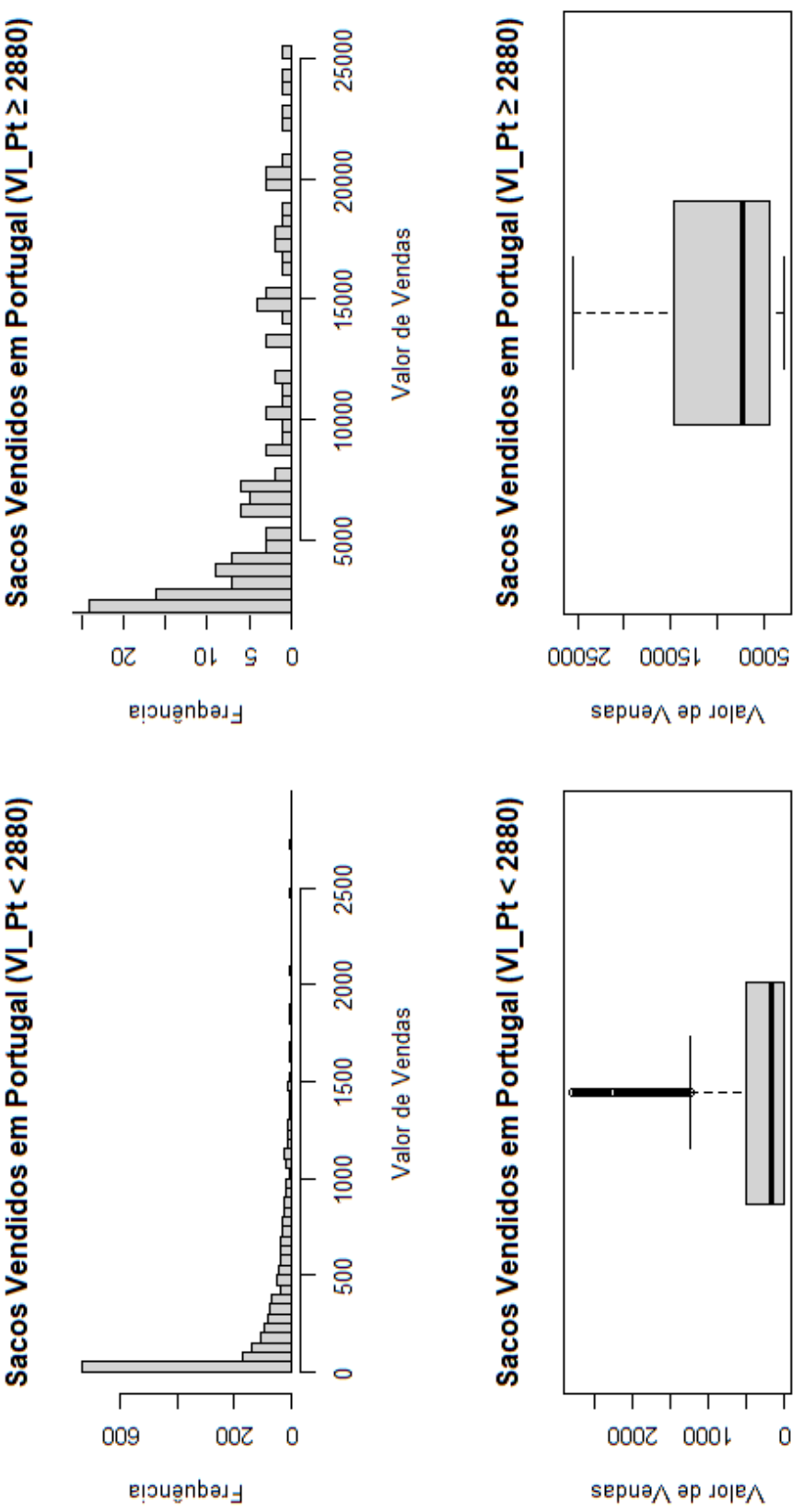

Figura 29: Histograma e *boxplot* para a variável Vl_Pt

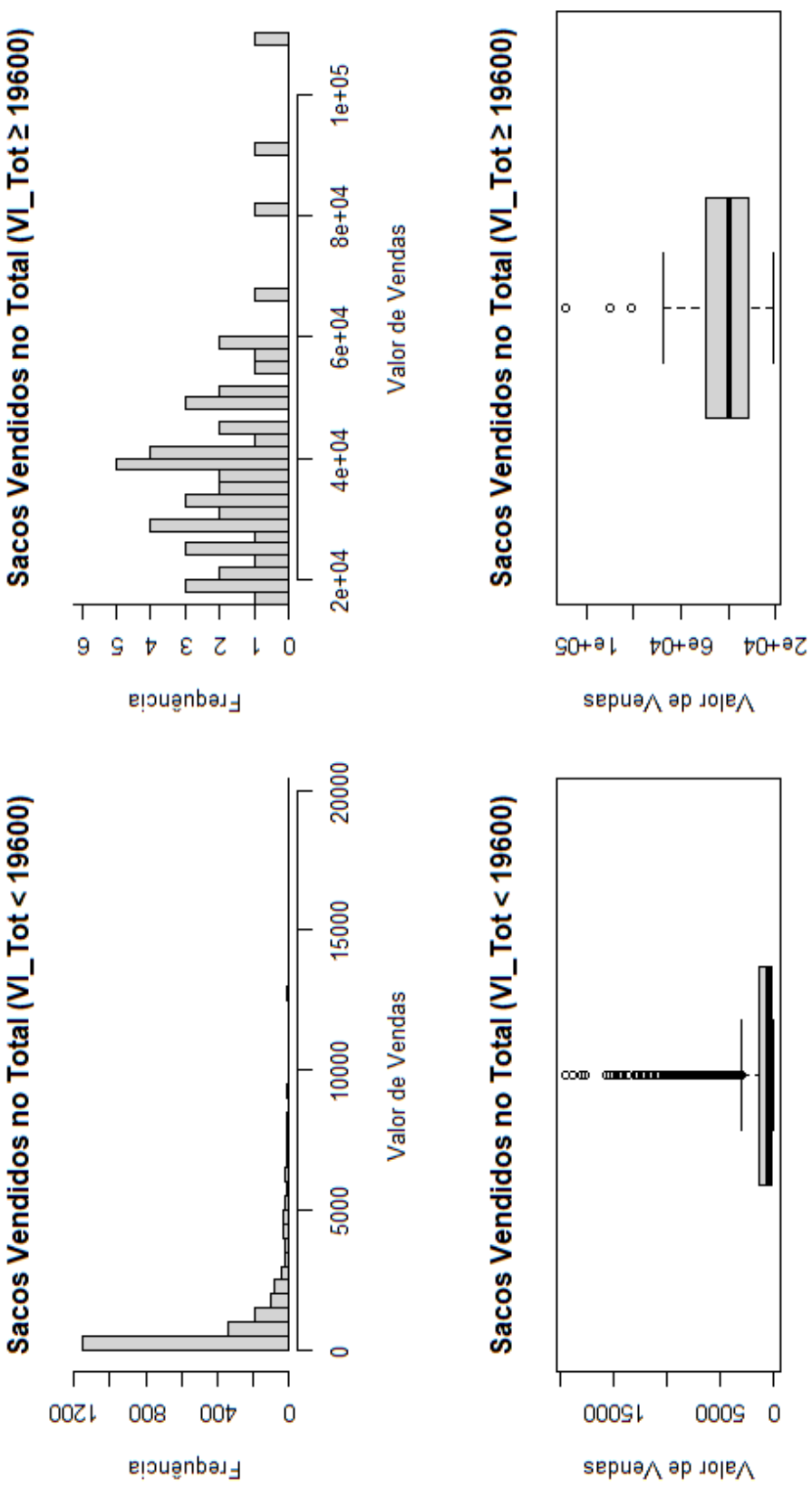

Figura 30: Histograma e *boxplot* para a variável Vl_Tot

5.2.1 Top de vendas

O $spaghetti\ plot$ da Figura 31 apresenta o comportamento das vendas dos sacos ao longo do tempo. Visualmente, consegue-se observar que existem sacos que são mais vendidos que os restantes. E existem dois sacos que obtiveram um pico máximo de vendas no ano 2021, correspondente a mais de 1000000 unidades vendidas. De modo, a conseguir identificar o comportamento de vendas de um determinado saco, é apresentado de seguida um top de vendas, em que consta os sacos mais vendidos e menos vendidos, enquadrado num certo nível de vendas. A base de dados dos sacos foi repartida, em três níveis de vendas:

- Nível 1 - sacos vendidos em menor quantidade (Qt_Tot < 5000)
- Nível 2 - sacos vendidos em quantidade intermédia (5000 $\leq Qt\_Tot$ < 40000)
- Nível 3 - sacos vendidos em maior quantidade (40000 $\leq Qt\_Tot$ < 250000)

Para além destes níveis de vendas, são identificados os seis sacos com mais lucro para a empresa durante os últimos sete anos (Qt_Tot > 250000).

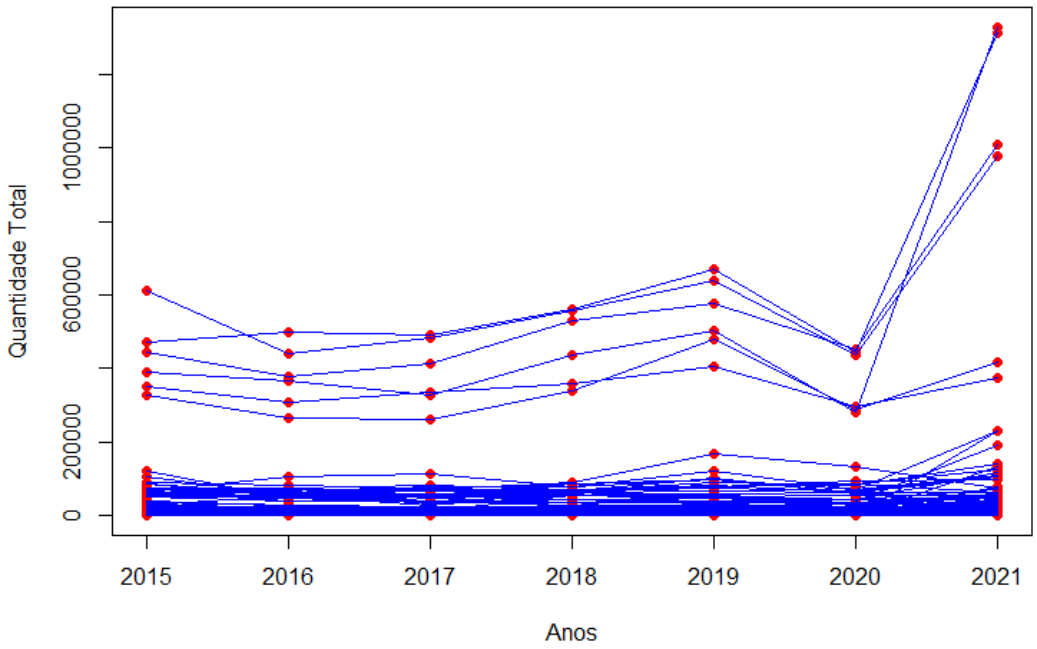

Figura 31: Spaghetti Plot do comportamento dos sacos ao longo do tempo

Na Figura 32 são apresentados os seis sacos com mais lucro para empresa durante os últimos sete anos. Os sacos *COLORIS (61000)* E *KRAFT LISO (61068)* são os dois sacos que mais se venderam na Litel, Lda, nos últimos anos, sendo que *KRAFT LISO (61068)* e *KRAFT LISO (61120)* obtiveram picos máximos de vendas em 2021. Estes picos em 2021 revelam irregularidades nas vendas, uma vez que devido à situação pandémica do Covid-19 em 2020, muitas das encomendas deste ano transitaram para o ano 2021, compensando a quebra de vendas do ano anterior.

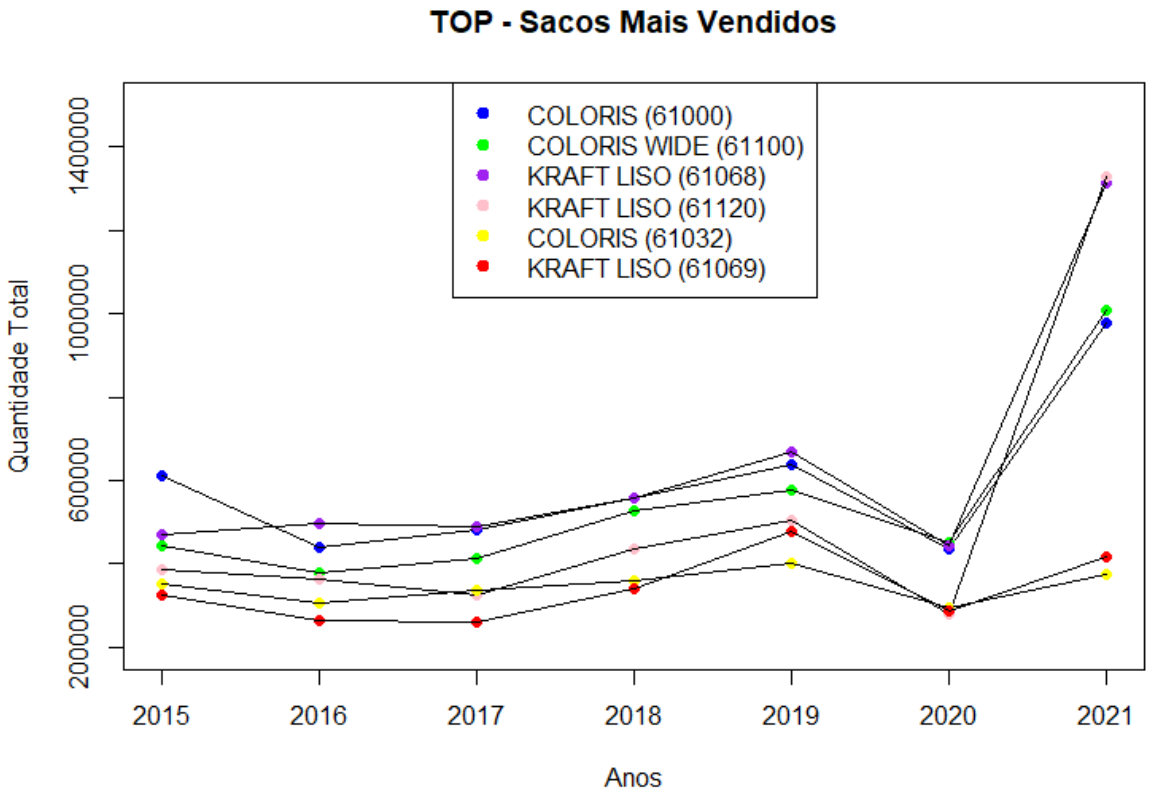

Figura 32: Sacos com vendas mais significativas ao longo do tempo

Na Figura 33 observa-se que os dois sacos mais vendidos do nível 1 em 2015 são *KRAFT GIFT (30666)* e *COLORIS (61204)*. Estes sacos têm um comportamento idêntico ao longo do tempo, e em 2021 apresentam um pico mínimo de vendas. O saco *GEMINUS 1 (62073)* possui observações desde o ano 2017 a 2021, e a sua tendência de vendas tem sido positiva ao longo dos anos.

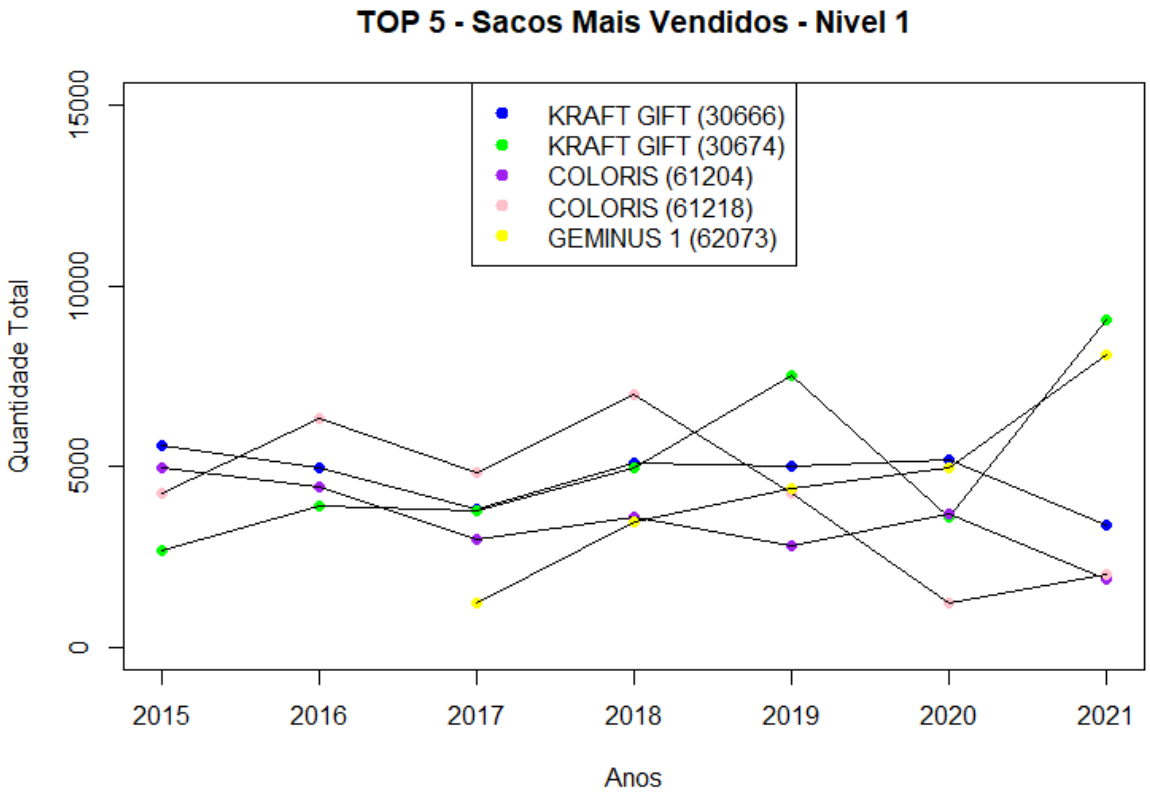

Figura 33: Sacos mais vendidos (Nível 1 de Vendas)

Relativamente ao nível 2 de vendas de sacos mais vendidos (Figura 34), verifica-se que todos os sacos são da família de produtos *COLORIS*. O saco *COLORIS (61007)* possui dois picos de vendas no ano 2015 e 2021. O saco *COLORIS (61458)* destaca-se neste TOP5 como sendo o saco com quantidades vendidas mais baixas relativamente aos restantes sacos deste TOP de vendas.

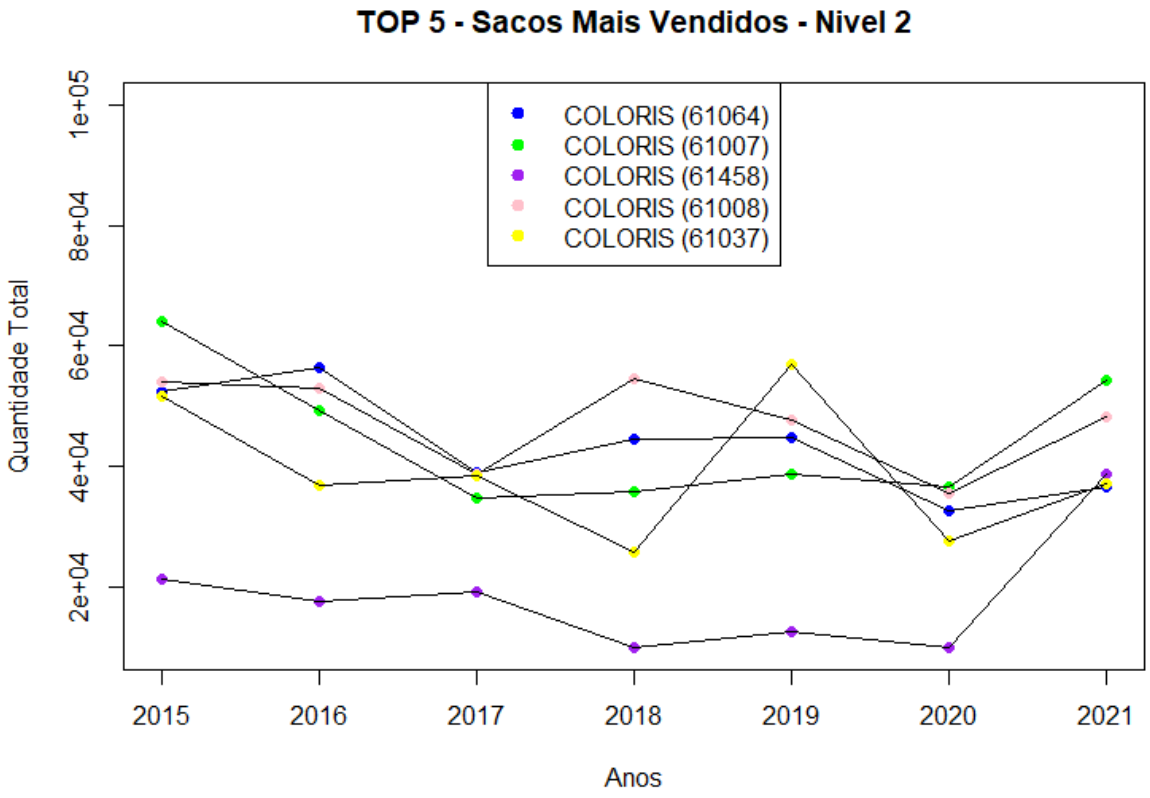

Figura 34: Sacos mais vendidos (Nível 2 de Vendas)

Na Figura 35 observa-se que com a exceção dos *SACOS TAKE AWAY*, todos os restantes sacos do TOP5 possuem quantidade de vendas acima de 50000 unidades ao longo dos anos. Os *SACOS TAKE AWAY* possuem apenas duas observações referentes a 2020 e 2021, que revelam uma tendência crescente de vendas. O saco *KRAFT VERJURADO (61042)* apresentou um pico de vendas em 2019 de cerca de mais de 150000 unidades vendidas, e desde aí apresentou uma tendência decrescente nas suas vendas (quantidade de vendas a baixo das 100000 unidades em 2021).

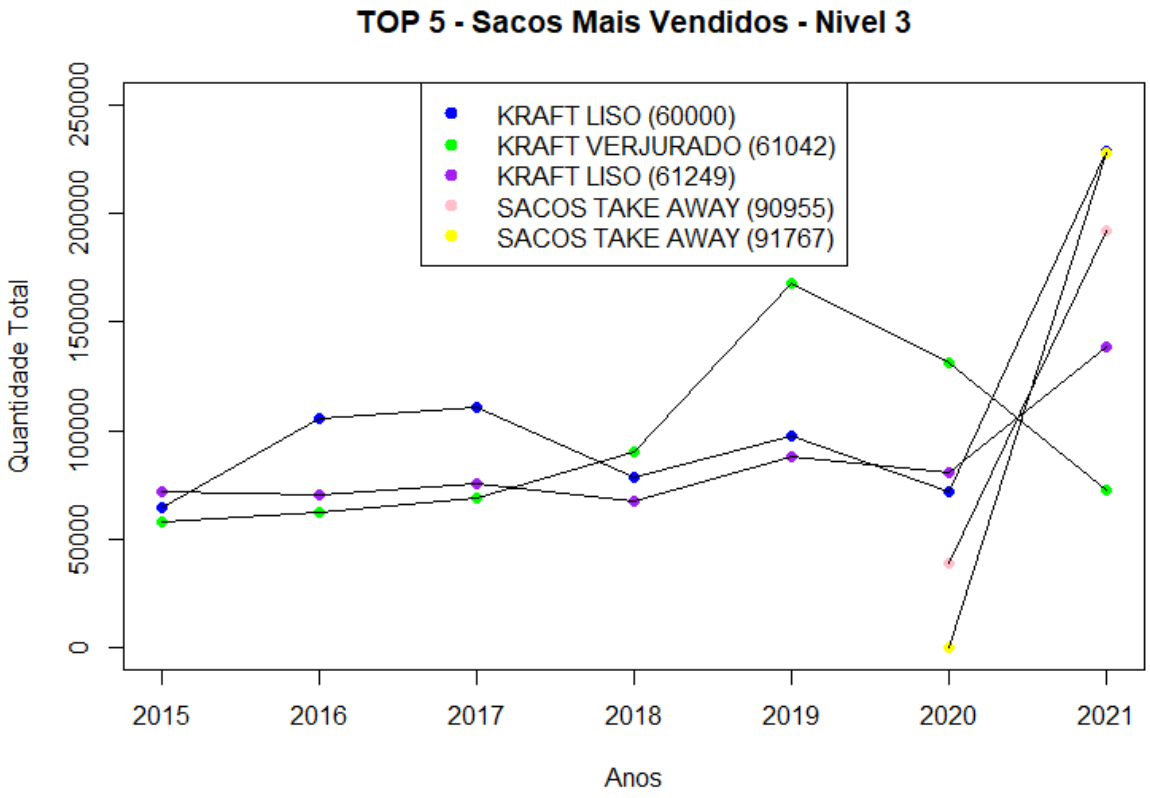

Figura 35: Sacos mais vendidos (Nível 3 de Vendas)

Os sacos menos vendidos do nível 1 (Figura 36), são aqueles que praticamente não trazem lucro para a empresa (vendas a baixo das 2000 unidades por ano). Neste TOP5 verificam-se picos mínimos de vendas em 2016 dos sacos *KRAFT GIFT (30691)*, *COLORIS GIFT (30681)* e *COLORIS GIFT (30689)*. De 2017 até ao ano de 2021 todos os produtos obtiveram vendas a baixo das 500 unidades, com a exceção de *COLORIS GIFT (30681)* no ano de 2019, que obteve quantidade de vendas ligeiramente a cima das 500 unidades.

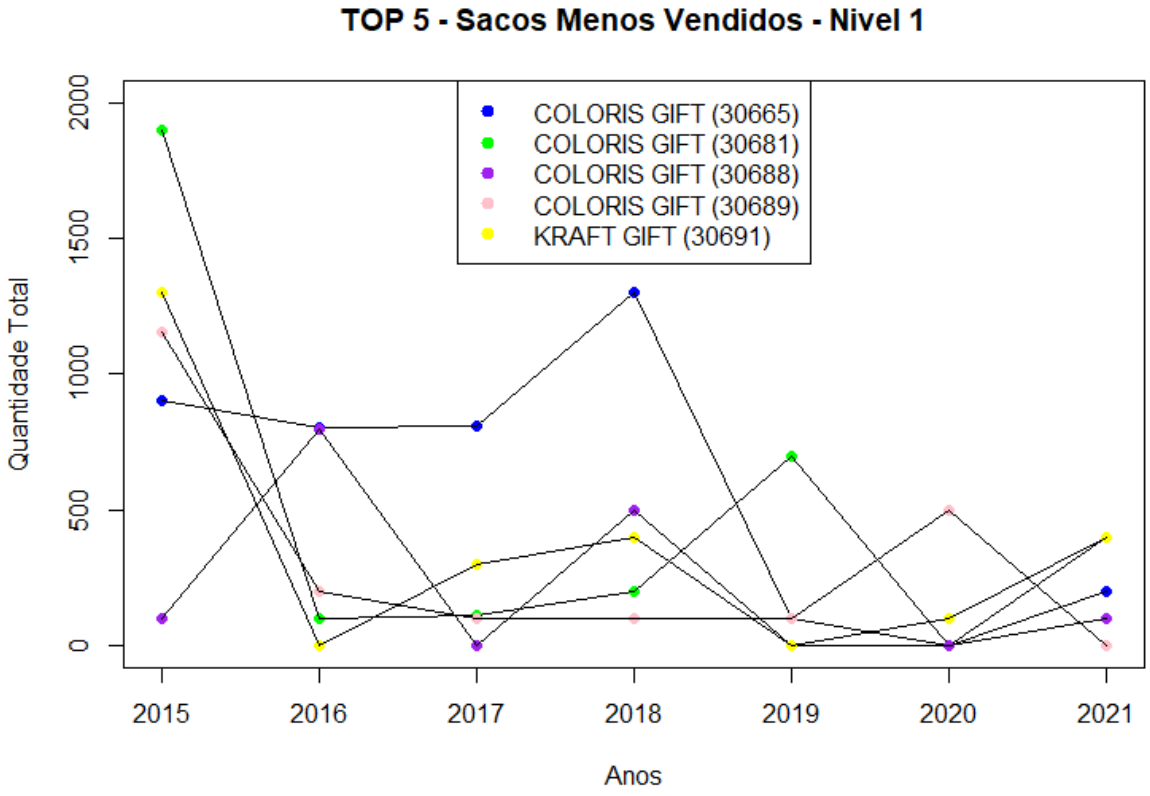

Figura 36: Sacos menos vendidos (Nível 1 de Vendas)

Na Figura 37 estão presentes os sacos menos vendidos do nível 2. De 2015 a 2020 os três sacos que mais se destacaram neste nível de vendas é *BACUS (61299)*, *KRAFT GIFT (30666)* e *CHRISTMAS SNOWMAN (61552)*, sendo que em 2018 o saco *BACUS (61299)* apresentou um pico máximo de vendas. No ano 2021, dentro deste nível de vendas aquele que mais se destacou na quantidade de vendas foi o saco *LUXURY BASIC (68705)*.

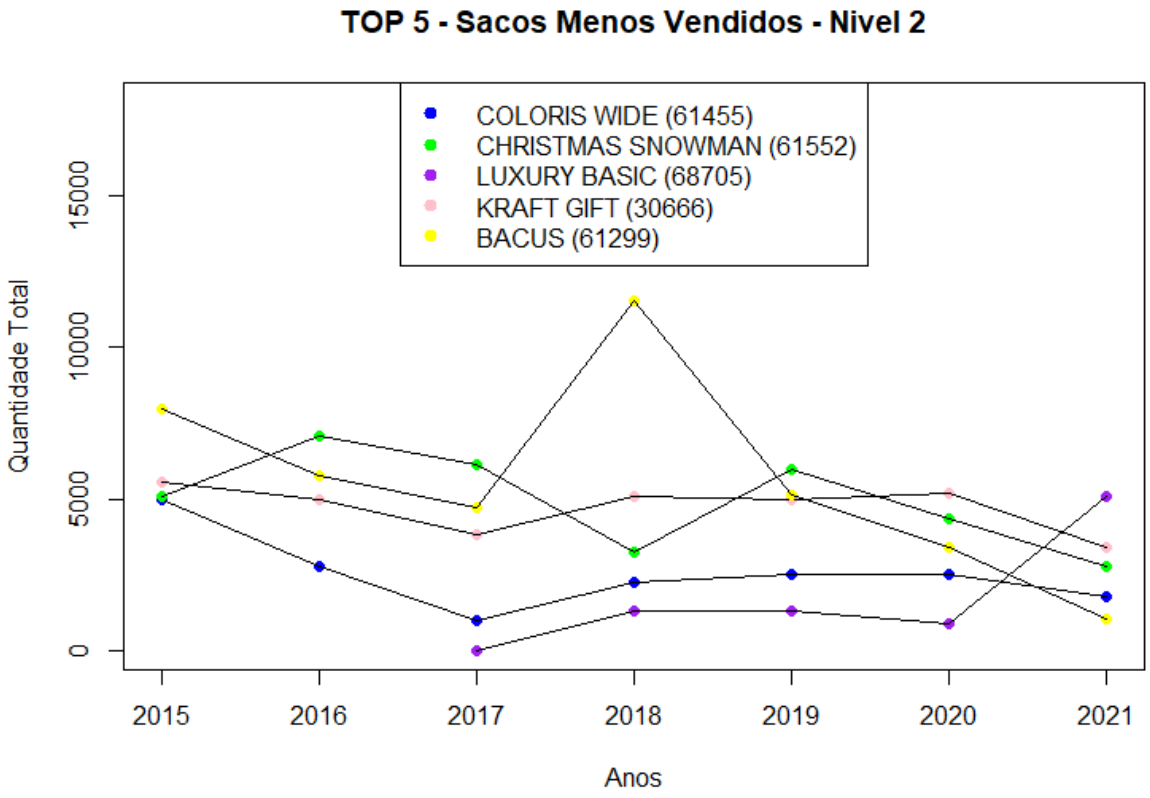

Figura 37: Sacos menos vendidos (Nível 2 de Vendas)

No nível 3 de vendas os 5 sacos menos vendidos foram 4 sacos da família de produtos *COLORIS*, e o saco *KRAFT LISO (61433)*, sendo que este último apresentou ao longo do tempo os valores mais baixos de quantidades de vendas. Os picos máximos mais expressivos são dos sacos *COLORIS (61216)* e *KRAFT LISO (61433)* no ano 2021 (Figura 38).

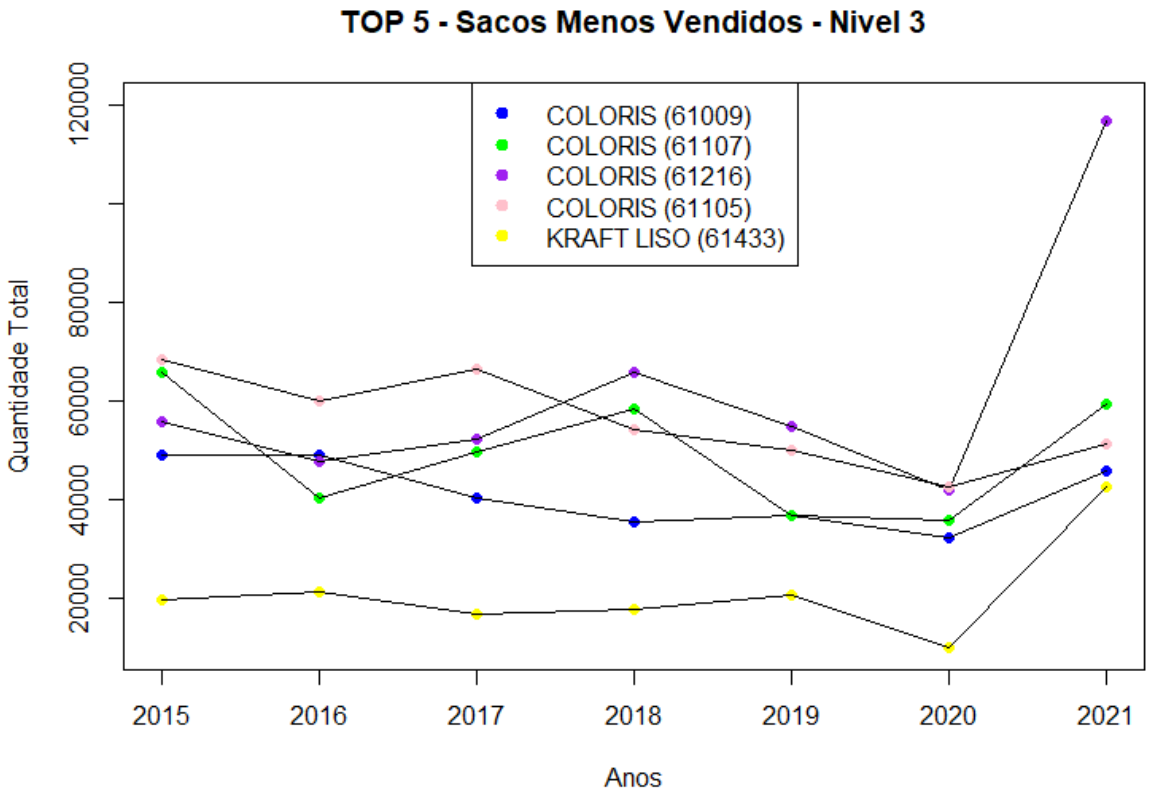

Figura 38: Sacos menos vendidos (Nível 3 de Vendas)

Capítulo 6

PCA e Análise de Clusters

6.1 Base de dados das caixas

6.1.1 PCA

O PCA é uma técnica multivariada que transforma dados em variáveis de número igual ou inferior à amostra inicial denominados "componentes principais". Os componentes principais, correspondem à combinação dos dados originais. Normalmente ocorre uma redução de dimensionalidade dos dados originais em dois ou três componentes, e são identificadas as direções pelas quais a variabilidade dos dados é máxima. O PCA ajuda a identificar padrões ocultos dos dados, reduzir a dimensionalidade, pela diminuição de redundância nos dados, e identifica variáveis correlacionadas.

Inicialmente foram selecionadas apenas as variáveis quantitativas do *dataset* de caixas (excluíram-se os dados da Inglaterra devido à ausência de vendas de caixas para este país), e posteriormente procedeu-se à normalização desses dados.

A variabilidade explicada em cada um dos componentes é medida pelos "autovalores" (*eigenvalues*), que podem ser extraídos utilizando a função $get\_eigenvalue()$ do $\mathbb{R}$. A Tabela 13 contém a informação retida através da aplicação desta função.

Tabela 13: CP da ACP normalizada

Componente	**Valor Próprio**	**Variância (%)**	**Acumulada (%)**
1	2.67169520	33.3961900	33.3961900
2	1.97824564	24.7280705	58.12426
3	1.83066360	22.8832950	81.00756
4	1.20867820	15.1084774	96.11603
5	0.19202322	2.4002902	98.51632
6	0.07233753	0.9042191	99.42054
7	0.02376034	0.2970043	99.71755
8	0.02259627	0.2824534	100.00000

Os autovalores superiores a 1 indicam que a variância do componente é superior ao que representaria a variância dos dados originais, sendo possível utilizar inclusive como ponto de corte para decidir quantos componentes utilizar. Observa-se que no exemplo acima, foram criados 8 componentes principais, dos quais os três primeiros componentes explicam 81,00756% da variabilidade total dos dados. O $scree\ plot$ da Figura 39 faz uma representação gráfica das componentes principais em relação à variabilidade que cada uma delas é capaz de explicar.

Nas Figuras 40, 41 e 42 é possível observar a contribuição de cada uma das variáveis em cada uma das componentes. A primeira componente principal recebe maior contribuição das variáveis Vl_Sp , Qt_Pt, Qt_Sp e Vl_Pt. Em relação à segunda componente, Qt_Ger e Vl_Ger são as que mais contribuem. A terceira componente tem uma maior contribuição da variável Vl_Fr e Qt_Fr.

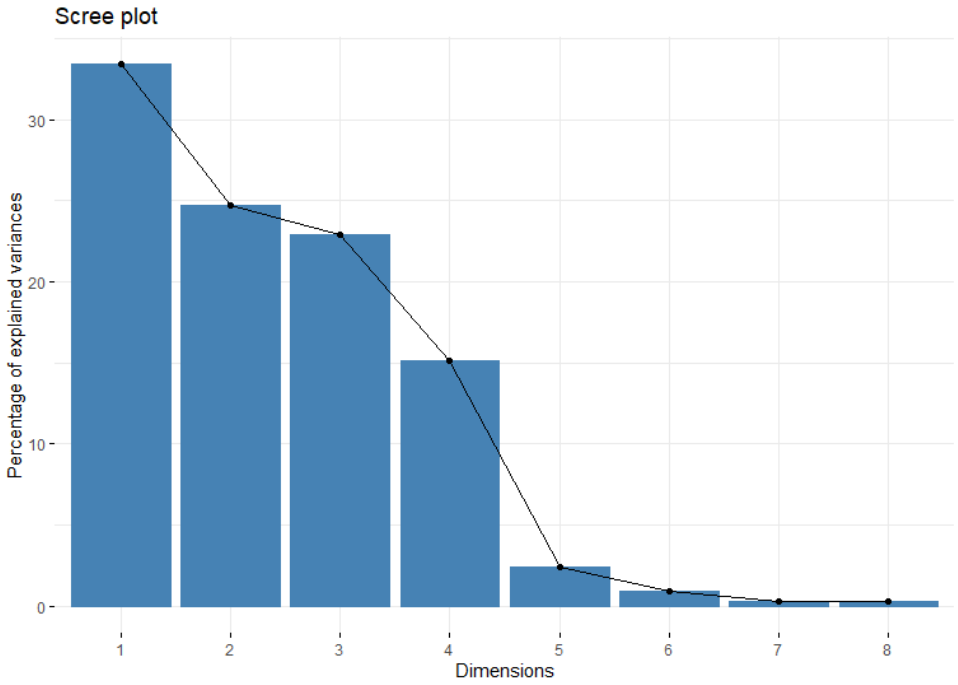

Figura 39: $Scree\ plot$ das CP

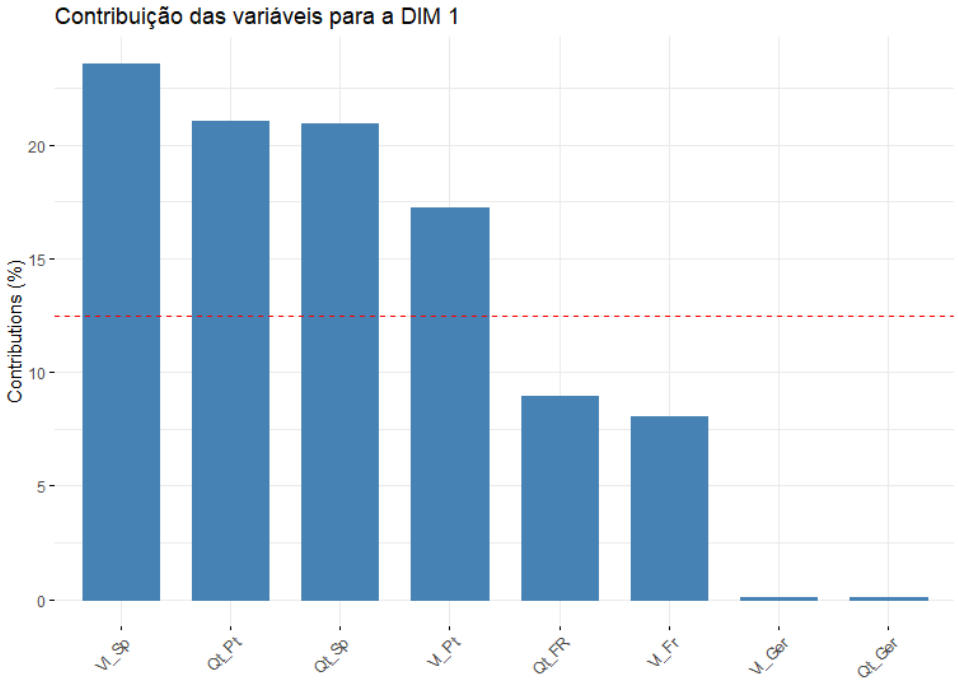

Figura 40: Importância das variáveis para a componente 1

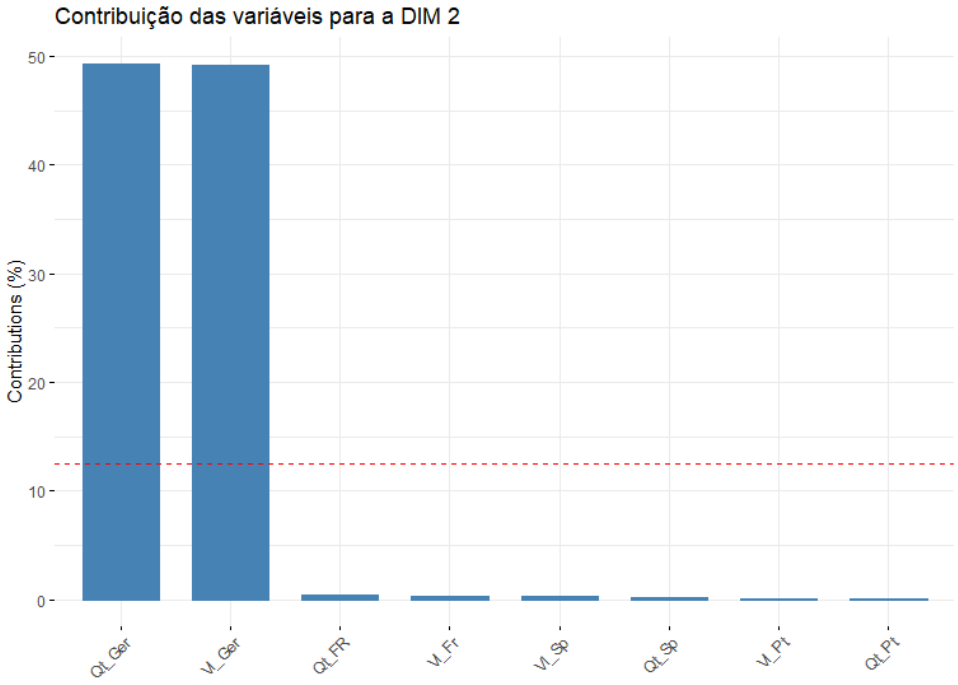

Figura 41: Importância das variáveis para a componente 2

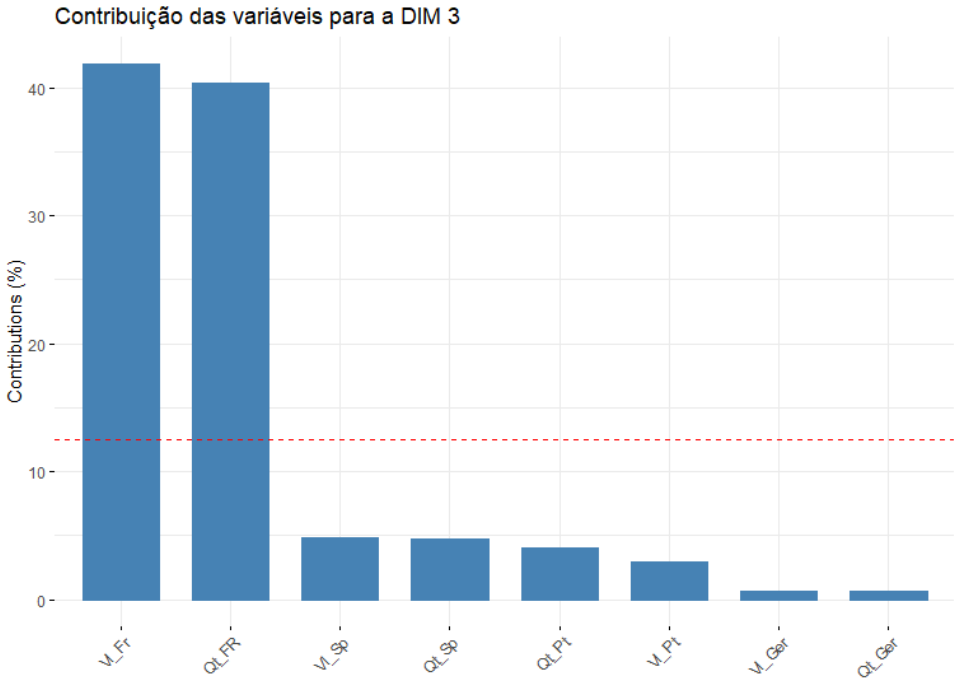

Figura 42: Importância das variáveis para a componente 3

Usando as variáveis das CP 1 e 2 consegue-se perceber alguma sobreposição nos dados, embora existam determinadas observações que parecem ter um padrão completamente distinto das restantes (Figura 43).

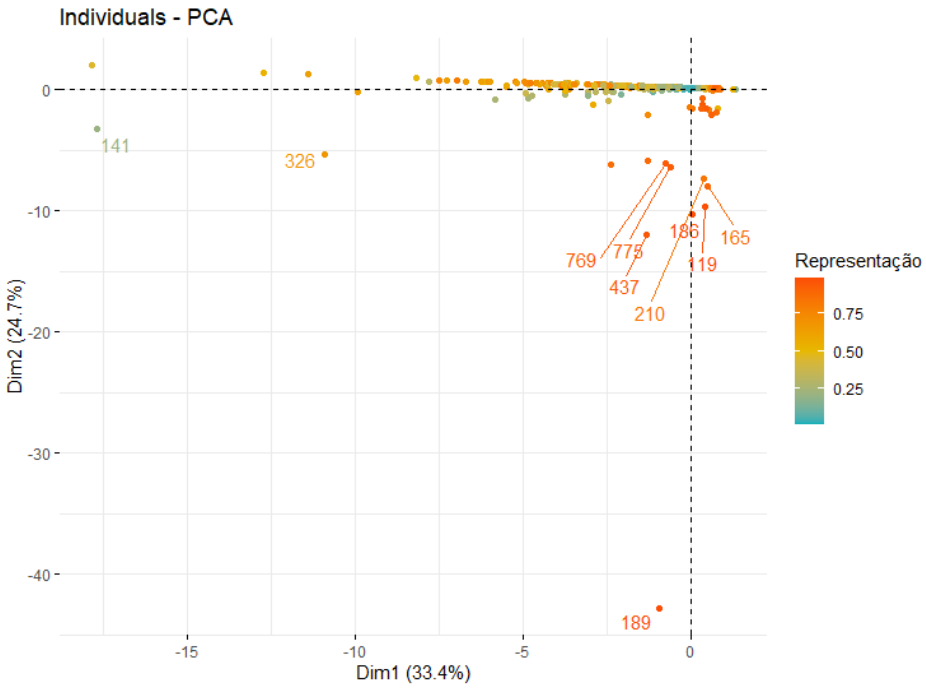

Figura 43: Gráfico de indivíduos

Uma outra forma de representar os dados no PCA é através de *biplots*. *Biplots* são gráficos bidimensionais que mostram as observações e as variáveis de um conjunto de dados, onde é possível ver a contribuição de cada variável para o espaço nas componentes principais. Quando o ângulo é maior, o cosseno é menor. Se o ângulo for maior que 90°, o cosseno é negativo, logo a correlação é negativa. As variáveis iniciais estão representadas no espaço das duas primeiras componentes principais, por autovetores (setas), e a correlação (cor) indica a contribuição (Figura 44). As duas componentes explicam 58.12426% da variância total.

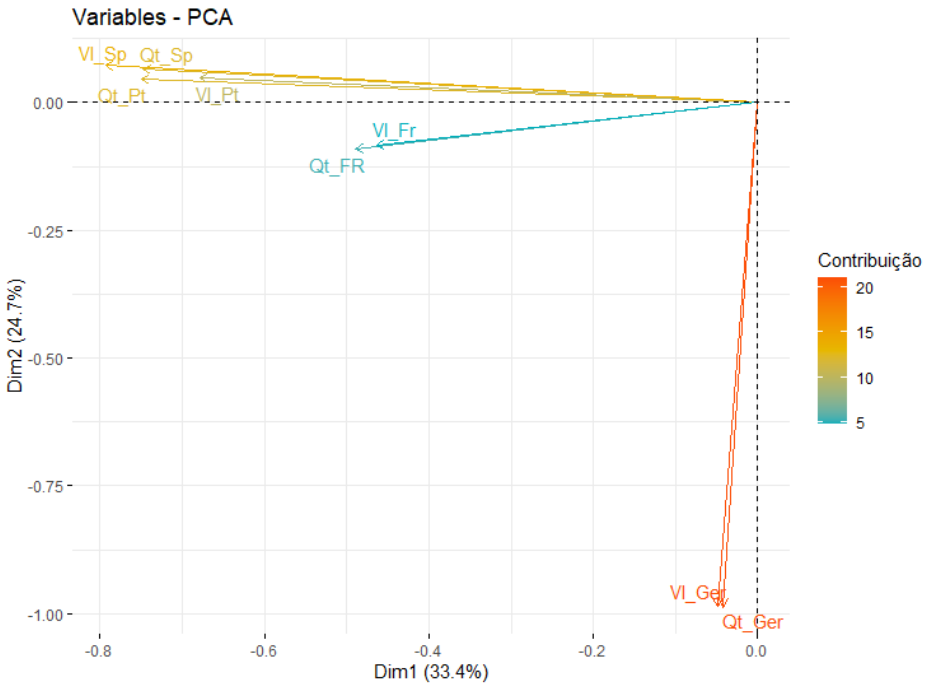

Figura 44: Gráfico de variáveis

Através da figura verifica-se que a correlação entre algumas das variáveis é muito elevada, por exemplo Vl_Sp , Qt_Sp, Qt_Pt e Vl_Pt. Estas variáveis apontam na mesma direção que o componente 1, reforçando a sua importância para este componente. Por outro lado, verifica-se que as variáveis Qt_Ger e Vl_Ger apontam na direção do componente 2, que é a que mais contribui. Na Figura 45, está presente um biplot com a representação das observações e variáveis do conjunto de dados.

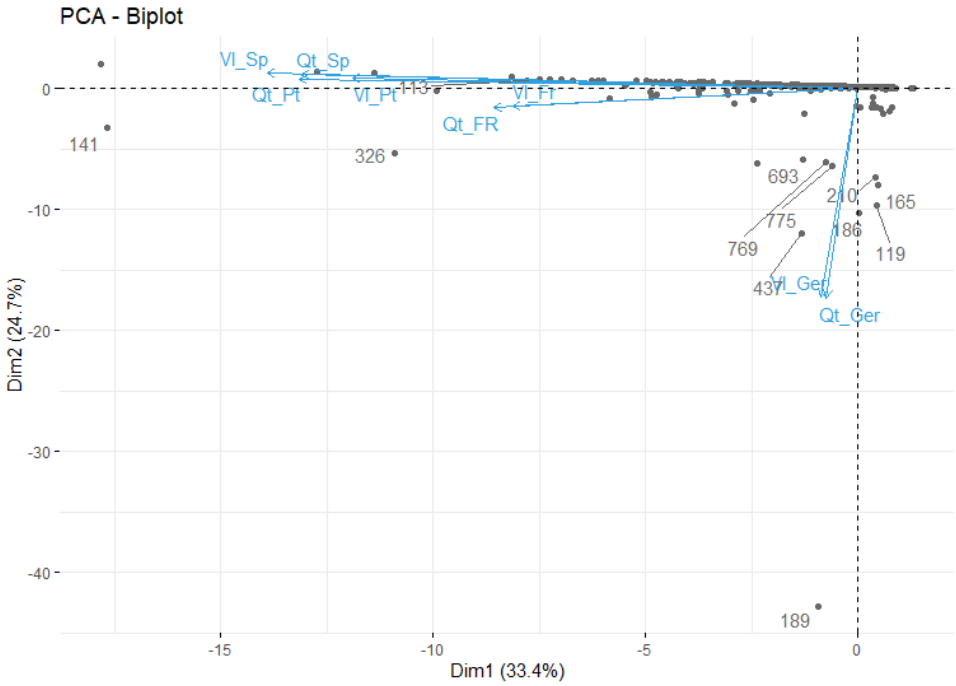

Figura 45: *Biplot* - Gráfico de variáveis e de indivíduos (amostras)

A Figura 46 contém uma representação gráfica com as *labels* de família de produtos. De um modo geral, consegue-se visualizar que existe alguma sobreposição na representação dos dados. No entanto, facilmente se conseguem discriminar alguns grupos de caixas com similaridades nas vendas, nomeadamente os produtos que se encontram a cor de rosa, verde e laranja.

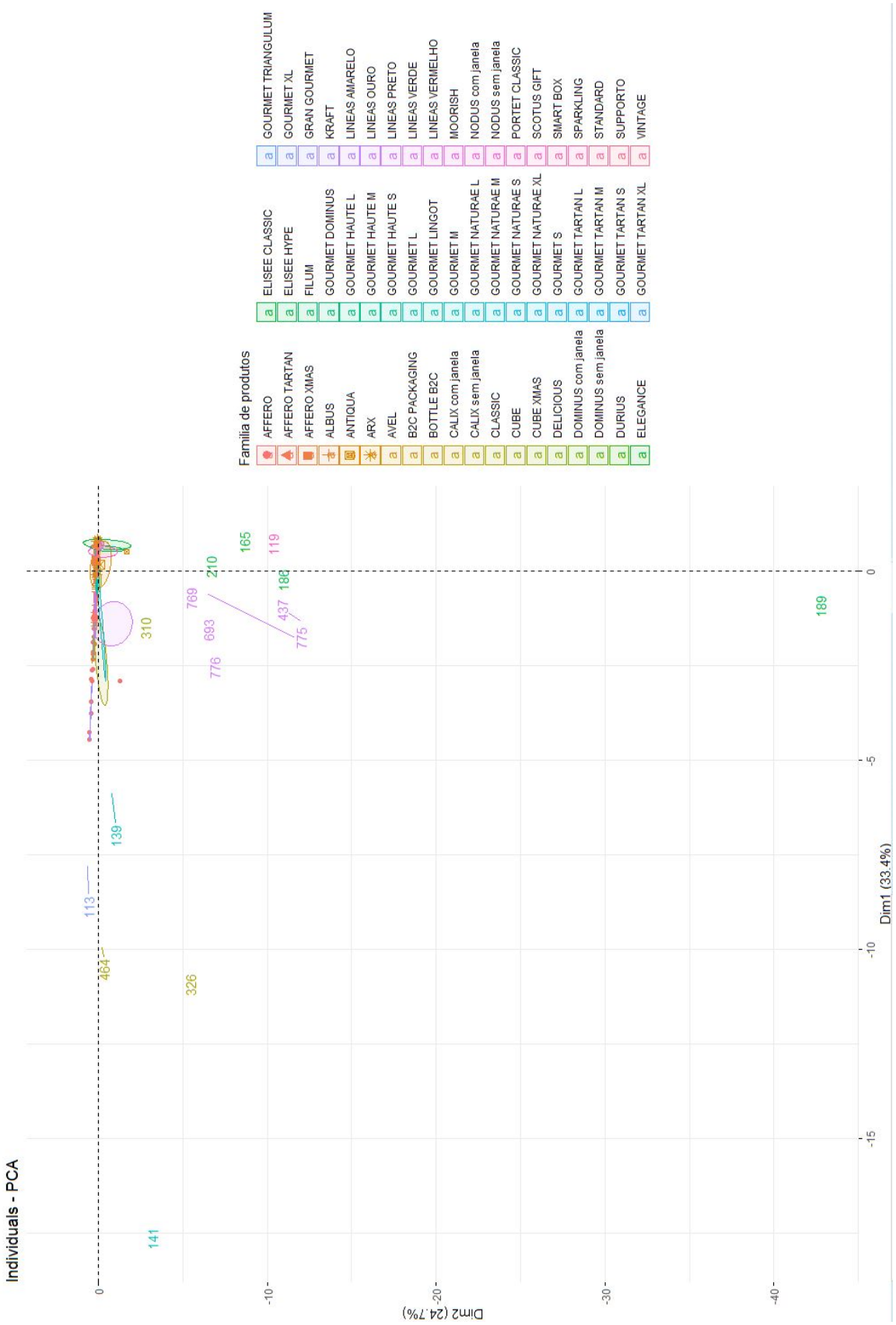

Figura 46: Representação gráfica com as *labels* de famílias de produtos

6.1.2 *k-Means*

Para aplicar o *k-Means* torna-se necessário determinar o número ótimo de *clusters*. O número ótimo de *clusters* é subjetivo, e depende do método usado para medir similaridades e dos parâmetros usados para partição. Para realizar esta validação foram usados métodos diretos que consistem num critério de otimização tais como soma dos quadrados dentro do *cluster* e silhueta média. Os métodos correspondentes são o método do cotovelo e silhueta respetivamente. Outro método que foi usado para determinar o número ótimo de *clusters* foi os 30 índices para escolha do melhor número de *clusters*, com recurso à função $NbClust()$ do $\mathbb{R}$. Na Figura 47 está presente a representação gráfica do método do cotovelo. Neste método torna-se fundamental localizar o ponto em que a soma dos quadrados dentro do *cluster* é minimizada. Para este método foi identificado um número ótimo de 5 *clusters* (k=5), embora a representação do cotovelo para os 10 *clusters* não seja muito precisa.

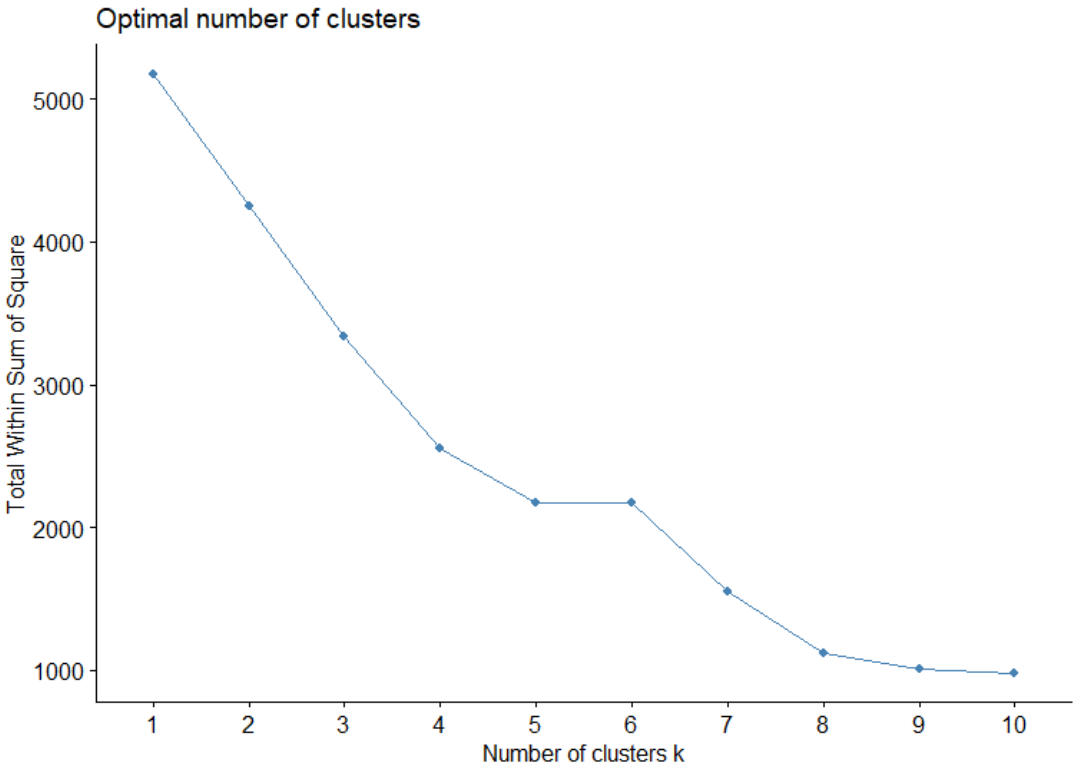

Figura 47: Método do cotovelo

O método da silhueta calcula a silhueta média de observações para diferentes valores de k. O número ótimo de *clusters* k é aquele que maximiza a silhueta média numa dada quantidade de *clusters*. Na Figura 48 está presente a representação gráfica do método da silhueta. Neste método o ponto em que existe uma maximização da silhueta média dentro do *clusters* é um valor de k=2.

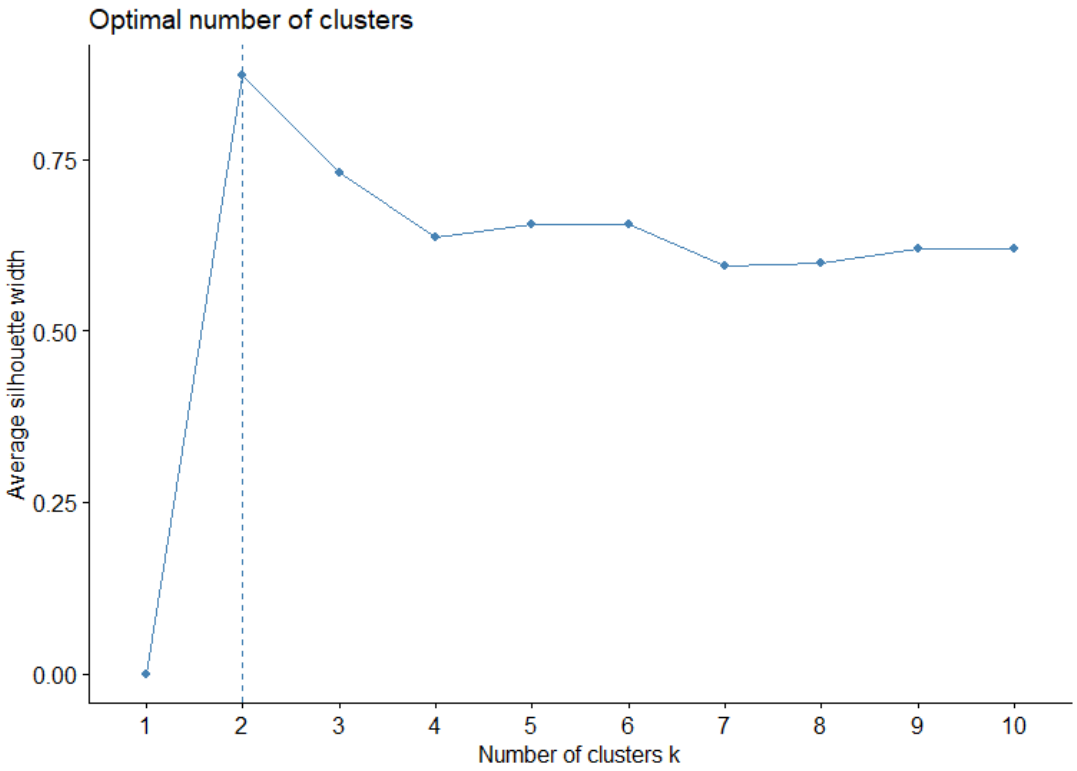

Figura 48: Método da silhueta

Outro método que foi usado para investigar o número ótimo de *clusters* foi os 30 índices para escolha do melhor número de *clusters*, com recurso à função $NbClust()$.

Este método fornece 30 índices para determinar o número ótimo de *clusters* e propõe o melhor esquema de *clustering* a partir dos diferentes resultados obtidos com a variação de todas as combinações de número de *clusters*, medidas de distância e métodos de *clustering*. O método consegue calcular simultaneamente todos os índices e determinar o número de *clusters*. Na Figura 49, está patente o método, e verifica-se que o número ótimo de *clusters* é obtido para k=2.

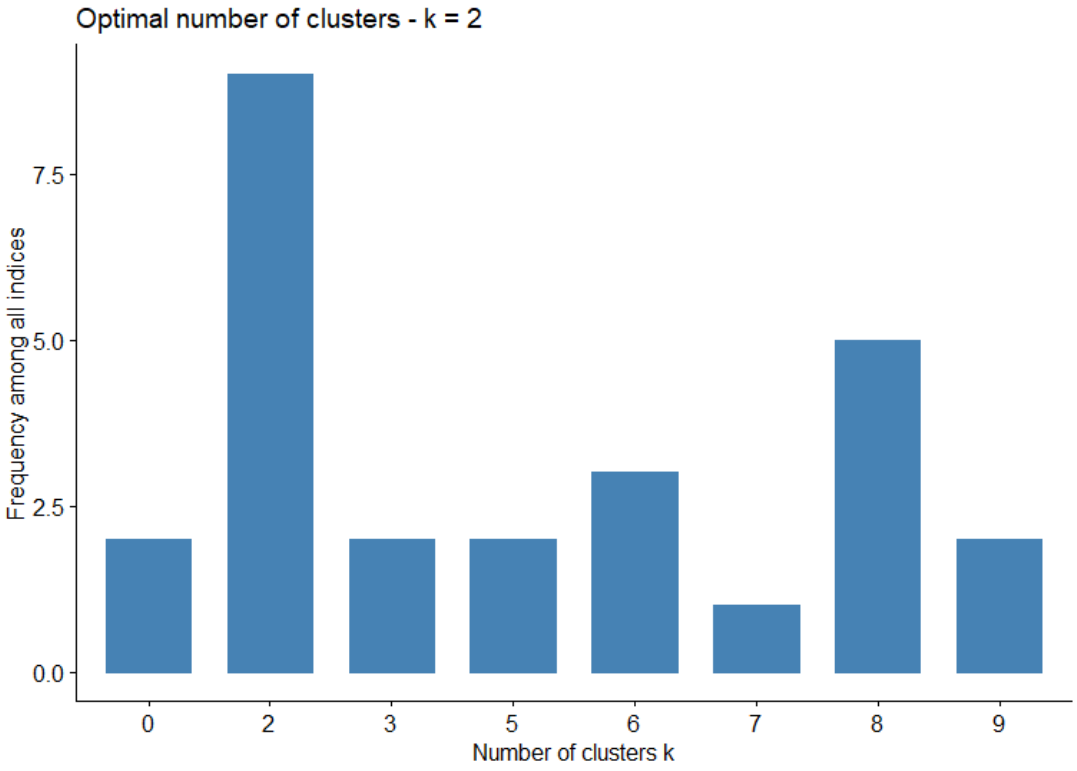

Figura 49: Recurso à função $NbClust()$ para determinar o melhor número de *clusters*

Após a determinação do número ótimo de *clusters* pelos métodos anteriormente mencionados, podemos proceder à implementação do algoritmo *k-Means* com os números de *clusters* recomendados. A Figura 50 apresenta o resultado do algoritmo *k-Means* para k= 2. Com esta representação gráfica, é possível denotar que existem dois grupos bem segmentados. O algoritmo foi capaz de agrupar caraterísticas semelhantes dentro de um mesmo *cluster*.

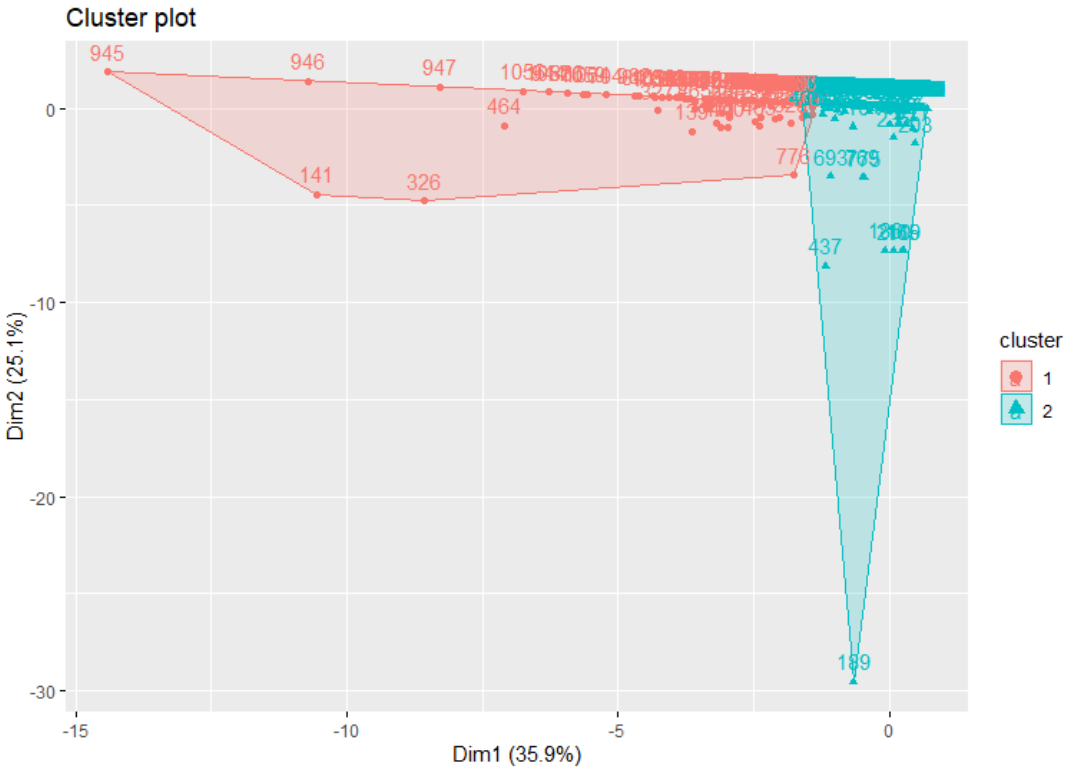

Figura 50: Resultado do *k-Means* com 2 *clusters*

Para o caso de k=5, são identificados alguns grupos, embora estes apresentem mais sobreposição entre eles (Figura 51).

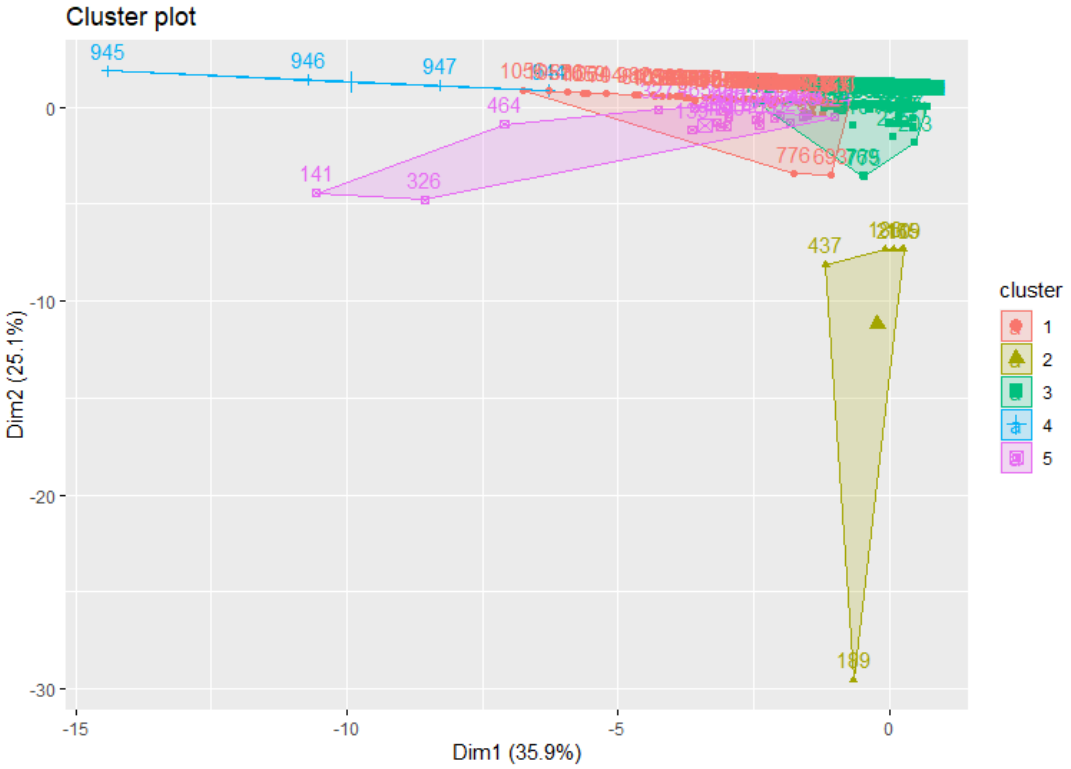

Figura 51: Resultado do *k-Means* com 5 *clusters*

6.1.3 *Clustering* hierárquico

Escolher o melhor método de *clustering* para um conjunto de dados pode ser uma tarefa difícil. No entanto, existem técnicas que permitem determinar o melhor método a ser utilizado. Através do pacote cl_Valid do $\mathbb{R}$, pode-se comparar simultaneamente múltiplos algoritmos de *clustering* através de uma simples função identificando a melhor abordagem de *clustering* e o número ótimo de *clusters*. A Figura 52, refere que o método de *clustering* que obtém um melhor desempenho é o *clustering* hierárquico com 3 *clusters* em cada caso (*Connectivity, Dunn e Silhouette*). Independentemente do algoritmo de agrupamento, o número ótimo de *clusters* é 3 usando as três medidas.

```
Clustering Methods:
 hierarchical kmeans

Cluster sizes:
 3 4 5 6

Validation Measures:
                                 3       4       5       6

hierarchical Connectivity   5.8579 11.2821 15.8067 17.9734
             Dunn           0.7596  0.3523  0.3271  0.3569
             Silhouette     0.9488  0.9079  0.8846  0.8826
kmeans       Connectivity   5.8579 15.8067 16.3492 18.5159
             Dunn           0.7596  0.2160  0.1047  0.1143
             Silhouette     0.9488  0.9014  0.8024  0.8023

Optimal Scores:

             Score  Method       Clusters
Connectivity 5.8579 hierarchical 3
Dunn         0.7596 hierarchical 3
Silhouette   0.9488 hierarchical 3
```

Figura 52: Validação do melhor algoritmo de *clustering* (consola do $\mathbb{R}$)

As medidas de estabilidade são uma versão especial de medidas internas, que avalia a consistência de um resultado de agrupamento comparando-o com os agrupamentos obtidos após a remoção de cada coluna, uma de cada vez. As medidas de estabilidade de *clusters* incluem:

- *Average proportion of non-overlap (APN)* mede a proporção média de observações não colocadas no mesmo *cluster* agrupadas com base nos dados completos e agrupadas com base nos dados com uma única coluna removida.
- *Average distance (AD)* mede a distância média entre observações colocadas no mesmo *cluster* em ambos os casos (conjunto de dados completos e removendo uma coluna).
- *Average distance between means (ADM)* mede a distância média entre centros de *clusters* para observações colocadas no mesmo *cluster* em ambos os casos.
- *Figure of merit (FOM)* mede a variância média intra-*cluster* da coluna removida, onde o agrupamento é baseado nas restantes colunas.

Os valores baixos de APN, ADM e FOM, correspondem a *clusters* altamente consistentes (estas medidas variam entre 0 e 1). AD varia entre 0 e infinito, e valores baixos são também preferidos. De acordo, com a Figura 53 para a medida de APN, o *clustering* hierárquico com 4 *clusters* é o que obtém um melhor *score*.

	Score	Method	Clusters
APN	0.001833594	hierarchical	4
AD	1.170807285	kmeans	6
ADM	0.044359211	hierarchical	3
FOM	0.980297242	kmeans	6

Figura 53: Medidas de estabilidade de *clusters* (consola do R)

A Figura 54, revela qual combinação representa um melhor agrupamento de dados. Nesta análise, é tida em conta o valor do coeficiente cofenético que reflete na qualidade do agrupamento. Com valores mais próximos de 1, os *clusters* refletem de forma mais precisa os dados (bom agrupamento). De um modo geral, valores acima de 0.75 são considerados bons. A combinação que representa o melhor agrupamento dos dados de acordo com o resultado da consola do R, é a junção da distância euclideana com o método *average*, que possui um coeficiente cofenético de cerca de 0.966.

```
> correlacao <- function(x,y){
+   res.dist <- dist(dfcaixas, method = methodsDist[x])
+   res.hc <- hclust(res.dist, method = methodsHc[y])
+   res.coph <- cophenetic(res.hc)
+   cor(res.dist, res.coph)
+ }
> for(i in 1:3){
+   for (j in 1:3){
+     a <- paste("D:", methodsDist[i], "Hc:", methodsHc[j], "Cor:", correlacao(i, j))
+     print(a)
+   }
+ }
[1] "D: euclidean Hc: single Cor: 0.883753390738331"
[1] "D: euclidean Hc: complete Cor: 0.923648801032995"
[1] "D: euclidean Hc: average Cor: 0.966331597947432"
[1] "D: maximum Hc: single Cor: 0.885735013457162"
[1] "D: maximum Hc: complete Cor: 0.922997016212569"
[1] "D: maximum Hc: average Cor: 0.964670970962153"
[1] "D: manhattan Hc: single Cor: 0.87086907176656"
[1] "D: manhattan Hc: complete Cor: 0.936346786541579"
[1] "D: manhattan Hc: average Cor: 0.961917710569196"
```

Figura 54: Combinações que revelam um melhor agrupamento dos dados (consola do R)

É possível validar o agrupamento realizado de forma interna. Observa-se na Figura 55 que a divisão em 3 *clusters* foi a melhor escolha.

```
$All.index
     3      4      5      6
0.9503 0.9094 0.8862 0.8849

$Best.nc
Number_clusters     Value_Index
         3.0000          0.9503
```

Figura 55: Validação interna do agrupamento ($index$: $silhouette$)

Após realizar as validações necessárias, é possível implementar o algoritmo de *clustering* hierárquico correspondente. A Figura 56 mostra o resultado do dendrograma obtido com as combinações de *clustering* hierárquico (distância euclideana e método *average*). Em 3 grupos é possível denotar que dois deles são formados por um único elemento.

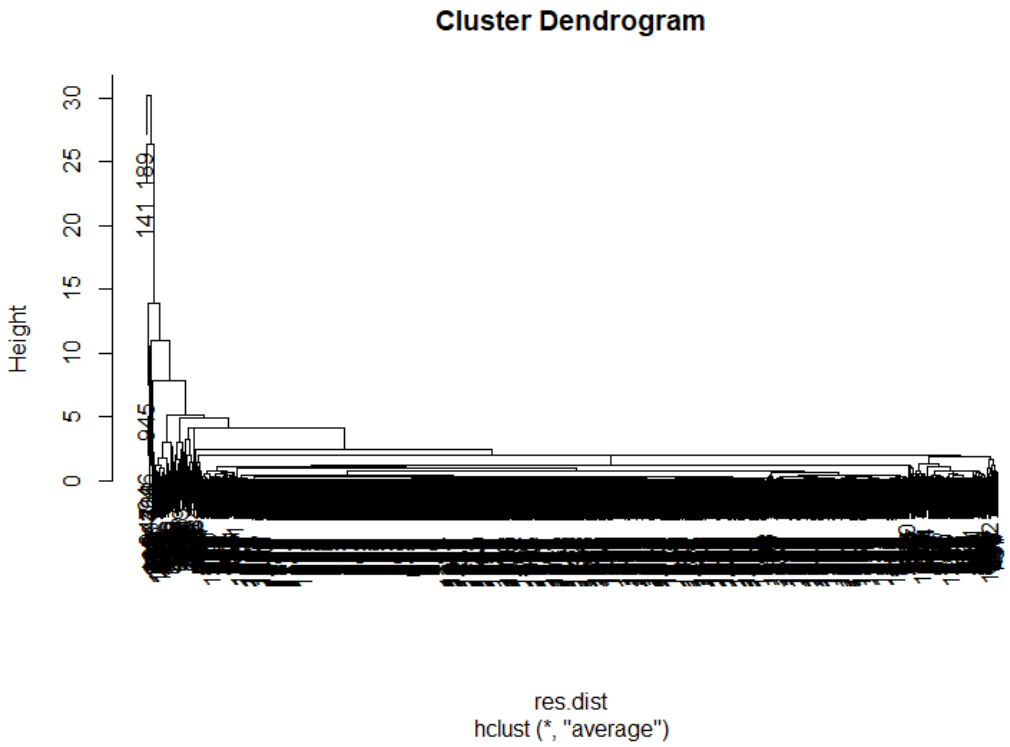

Figura 56: Dendrograma obtido através de *clustering* hierárquico com método *"average"*

A Figura 57 apresenta os três *clusters* obtidos com a aplicação do *clustering* hierárquico. Visualmente verifica-se que a maioria das observações ficaram agrupadas num único *cluster*, devido à proximidade que mantém entre si.

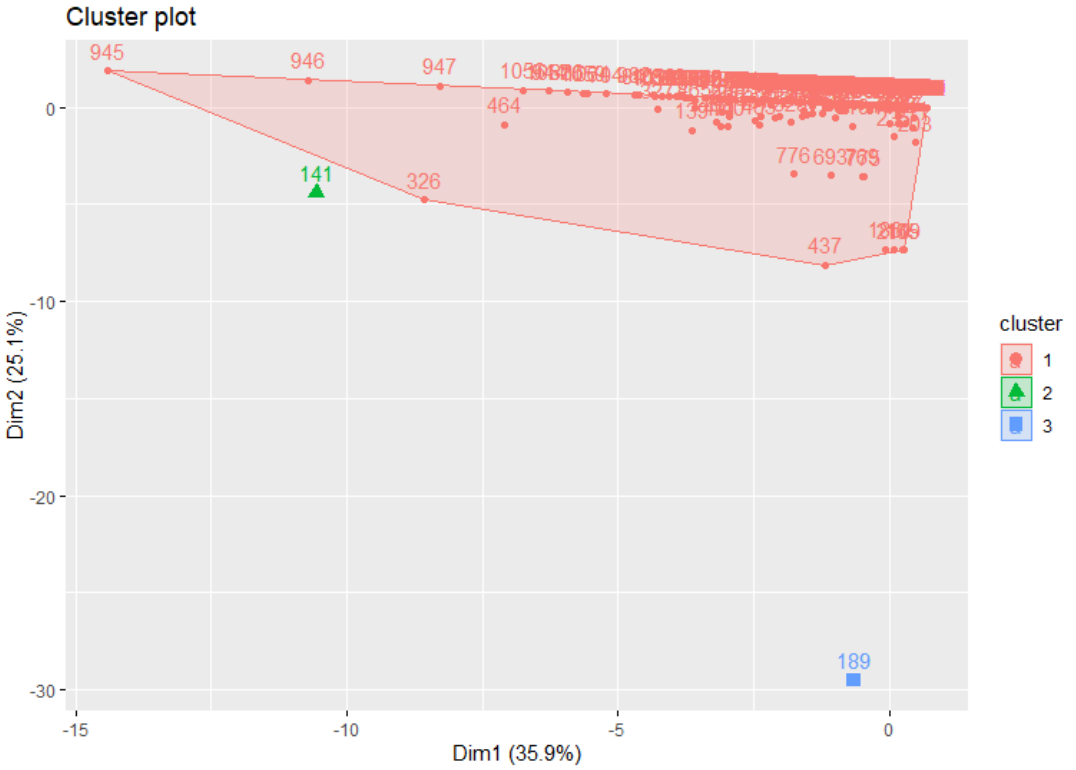

Figura 57: Representação dos *clusters* obtidos com *clustering* hierárquico (método *"average"*)

Outra tentativa efetuada, para o algoritmo de *clustering* hierárquico foi mudar o método para *ward*, e detetar visualmente através do dendrograma o número de *clusters* "ideal". A Figura 58 apresenta o resultado do dendrograma obtido. O dendrograma foi cortado em 3 *clusters*, e os mesmos podem ser visualizados na Figura 59.

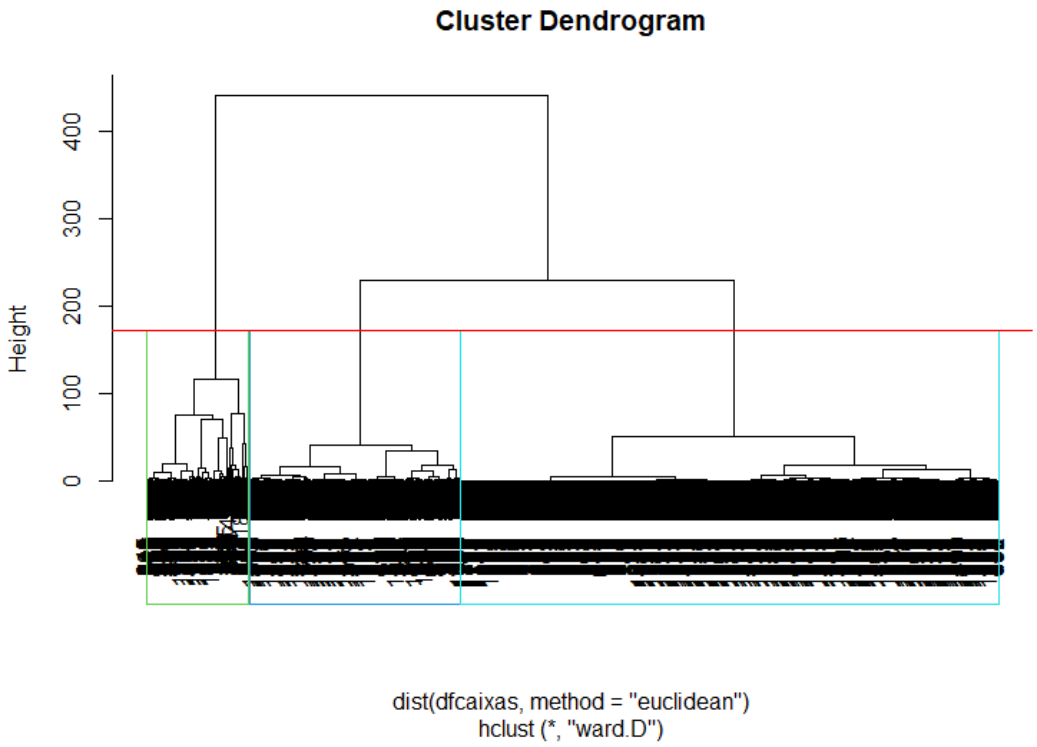

Figura 58: Dendrograma com o método *"ward"*

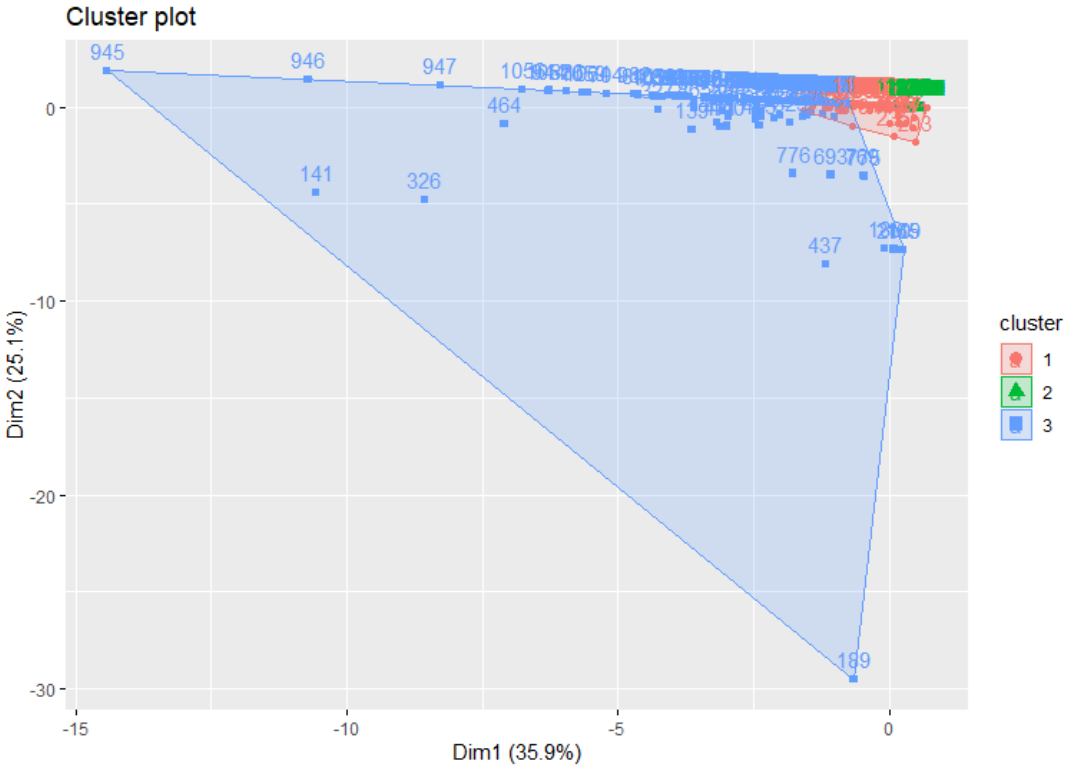

Figura 59: Representação dos *clusters* obtidos com o método de *clustering* hierárquico *"ward"*

A Tabela 14 contém a quantidade de observações que está presente em cada *cluster*. Verifica-se que o *cluster* com maior quantidade de observações é o *cluster* 2, que apresenta grandes quantidades de vendas similares entre si. O *cluster* 3 é aquele que apresenta menor número de observações.

Tabela 14: Quantidade de observações em cada *cluster*

Cluster	**n**
1	321
2	817
3	156

6.1.4 Avaliação de agrupamentos

Analisando os *plots* de silhueta dos algoritmos de *k-Means* (2 e 5 *clusters*), e *clustering* hierárquico (método *average e ward*), consegue-se concluír que o *clustering* hierárquico método *average* produziu melhores agrupamentos (coeficiente de silhueta de cerca de 0.95). No entanto este algoritmo agrupou praticamente todos os dados num único *cluster*, o que não produziu de certa forma os melhores resultados. O algoritmo de *k-Means* para k=2 obteve um bom coeficiente de silhueta, e permitiu definir bem os agrupamentos. Apesar do algoritmo de *clustering* hierárquico método *ward*, obter o coeficiente de silhueta mais baixo, permitiu segmentar de forma satisfatória os dados em três *clusters* (Figuras 60, 61, 62 e 63).

Figura 60: Validação *silhouette* para k=2 (*k-Means*)

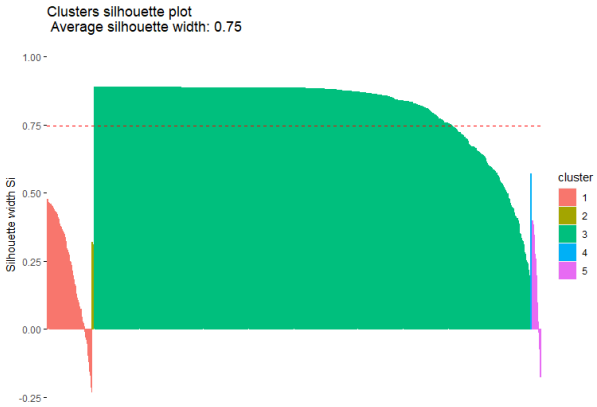

Figura 61: Validação *silhouette* para k=5 (*k-Means*)

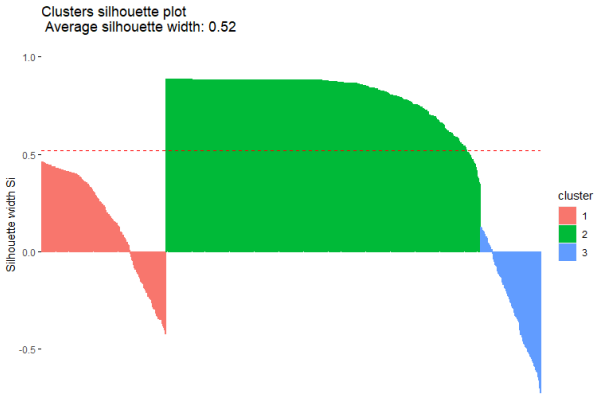

Figura 62: Validação *silhouette* para k=3 (*Clustering* hierárquico - Método *"Ward"*)

Figura 63: Validação *silhouette* para k=3 (*Clustering* hierárquico - Método *"Average"*)

6.2 Base de dados dos sacos

6.2.1 PCA

Inicialmente foram selecionadas apenas as variáveis quantitativas do *dataset* de sacos, e posteriormente procedeu-se à normalização desses dados.

A variabilidade explicada em cada um dos componentes é medida pelos "autovalores" (*eigenvalues*), que podem ser extraídos utilizando a função $get\_eigenvalue()$ do $\mathbb{R}$. A Tabela 15 contém a informação retida através da aplicação desta função.

Tabela 15: CP da ACP normalizada

Componente	**Valor Próprio**	**Variância (%)**	**Acumulada (%)**
1	5.020530291	50.20530291	50.20530
2	1.929853352	19.29853352	69.50384
3	1.884670872	18.84670872	88.35055
4	0.844424758	8.44424758	96.79479
5	0.185540243	1.85540243	98.65020
6	0.057428770	0.57428770	99.22448
7	0.035678172	0.35678172	99.58126
8	0.020177187	0.20177187	99.78304
9	0.017014063	0.17014063	99.95318
10	0.004682293	0.04682293	100.00000

Os autovalores superiores a 1 indicam que a variância do componente é superior ao que representaria a variância dos dados originais, sendo possível utilizar inclusive como ponto de corte para decidir quantos componentes utilizar. Observa-se que foram criados 10 componentes principais, dos quais os três primeiros componentes explicam 88,35055% da variabilidade total dos dados. O $scree\ plot$ da Figura 64 faz uma representação gráfica das componentes principais em relação à variabilidade que cada um deles é capaz de explicar. Nas Figuras 65, 66 e 67 é possível observar a contribuição de cada uma das variáveis em cada uma das componentes. A primeira componente principal recebe maior contribuição das variáveis, Qt_Sp, Vl_Sp, Qt_Pt, Vl_Pt, Vl_Fr e Qt_FR. Em relação à segunda componente, Vl_Ger e Qt_Ger são as que mais contribuem. A terceira componente tem uma maior contribuição das variáveis Vl_Eng e Qt_Eng.

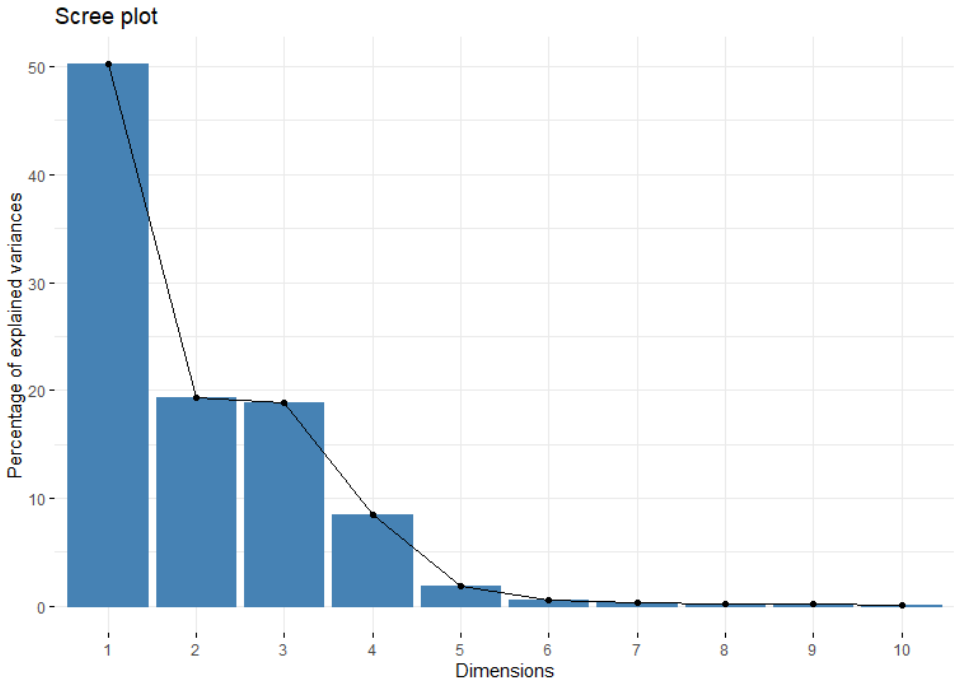

Figura 64: $Scree\ plot$ das CP

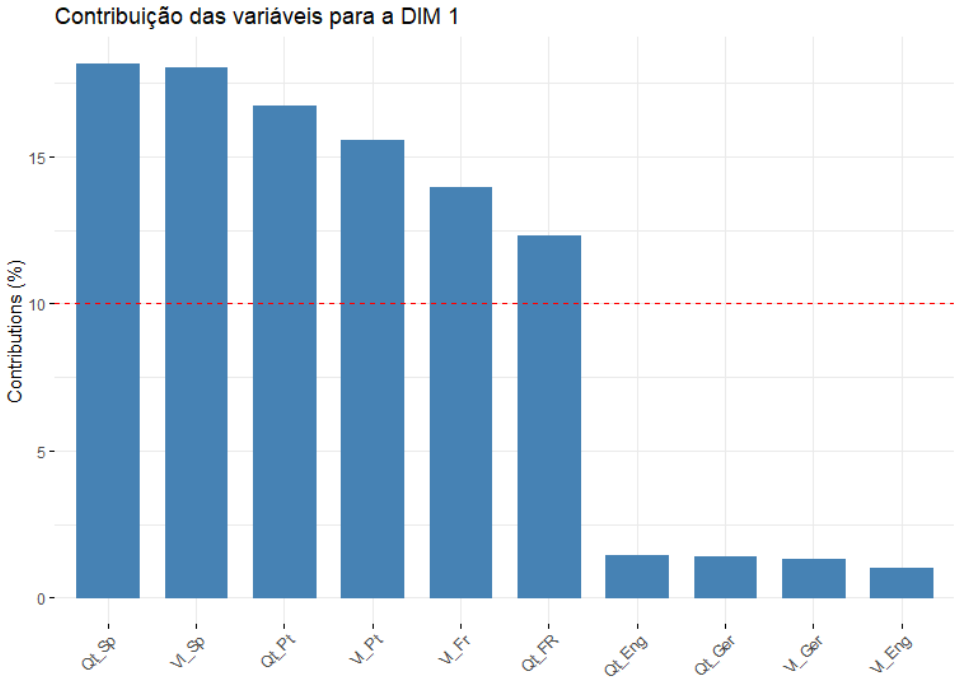

Figura 65: Importância das variáveis para a componente 1

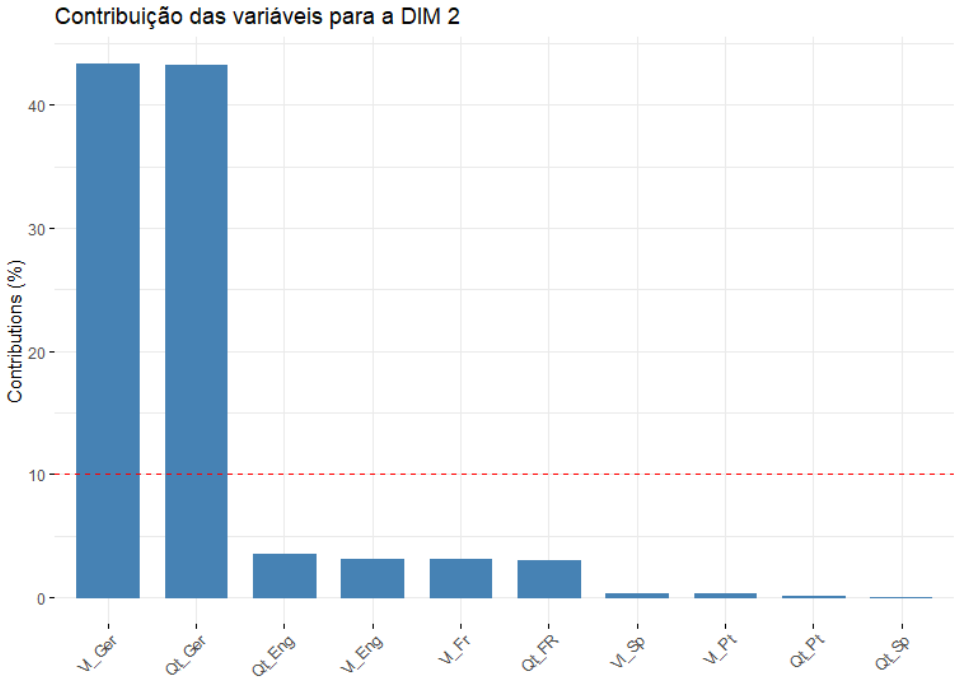

Figura 66: Importância das variáveis para a componente 2

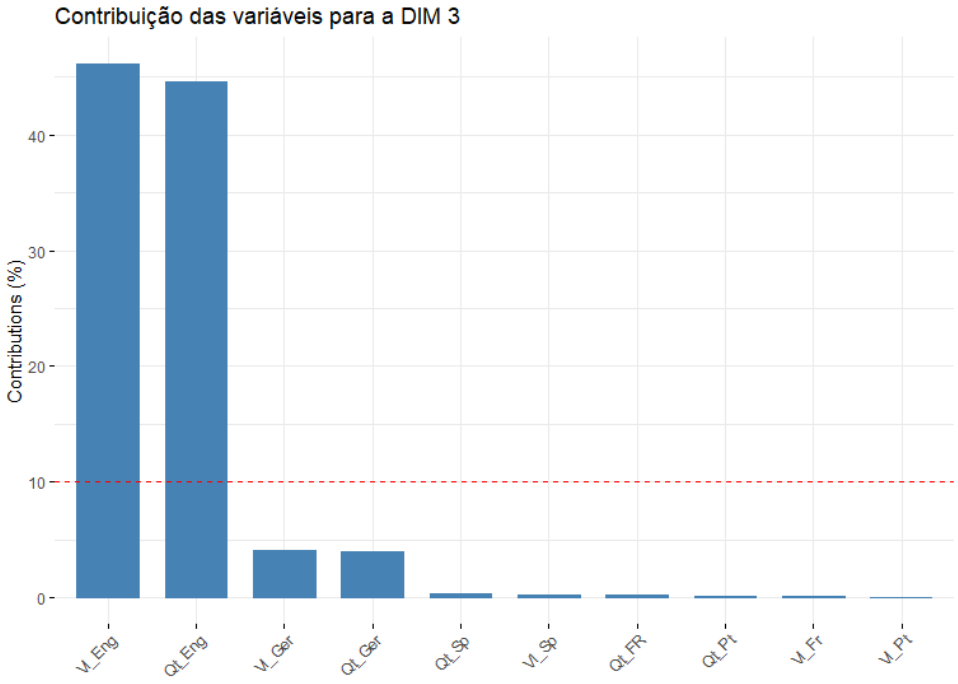

Figura 67: Importância das variáveis para a componente 3

Usando as variáveis das componentes principais 1 e 2 consegue-se perceber que existem determinadas observações que possuem um padrão completamente distinto das restantes (Figura 68).

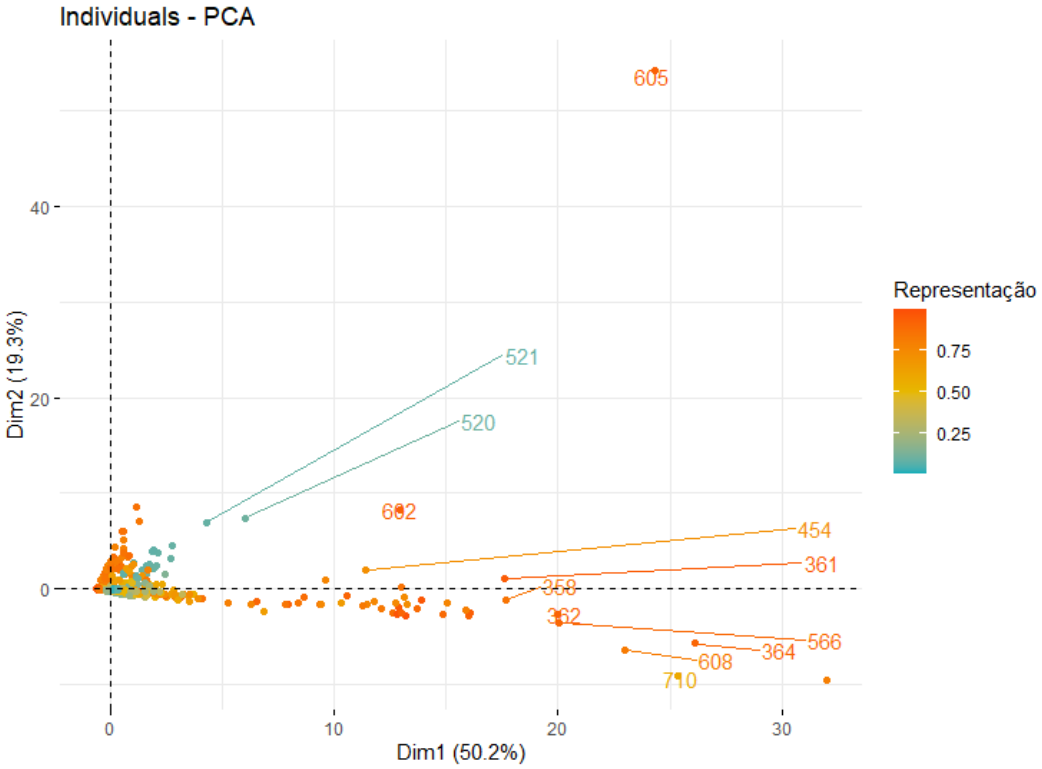

Figura 68: Gráfico de indivíduos

As variáveis iniciais estão representadas no espaço das duas primeiras componentes principais, por autovetores (setas), e a correlação (cor) indica a contribuição (Figura 69). As duas componentes explicam 69.50384% da variância total.

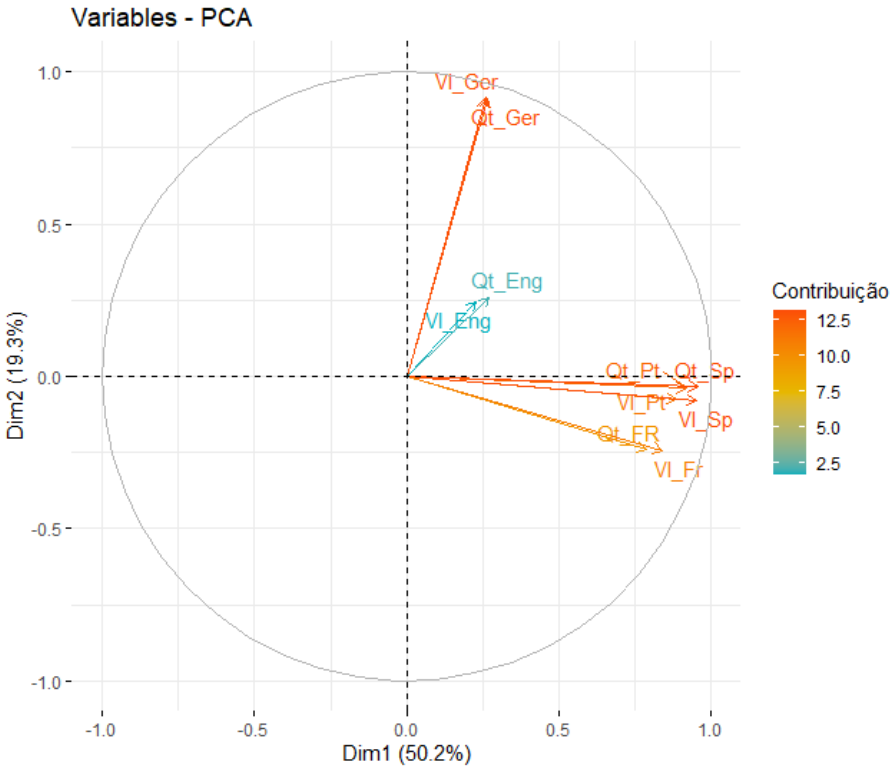

Figura 69: Gráfico de variáveis

Através da figura verifica-se que a correlação entre algumas das variáveis é muito elevada, por exemplo Vl_Sp , Qt_Sp, Qt_Pt e Vl_Pt. Estas variáveis juntamente com Qt_Fr e Vl_Fr apontam na mesma direção que o componente 1, reforçando a sua importância para este componente. Por outro lado, verifica-se que as variáveis Qt_Ger e Vl_Ger apontam na direção do componente 2, que é a que mais contribui. Na Figura 70, está presente um *biplot* com a representação das observações e variáveis do conjunto de dados.

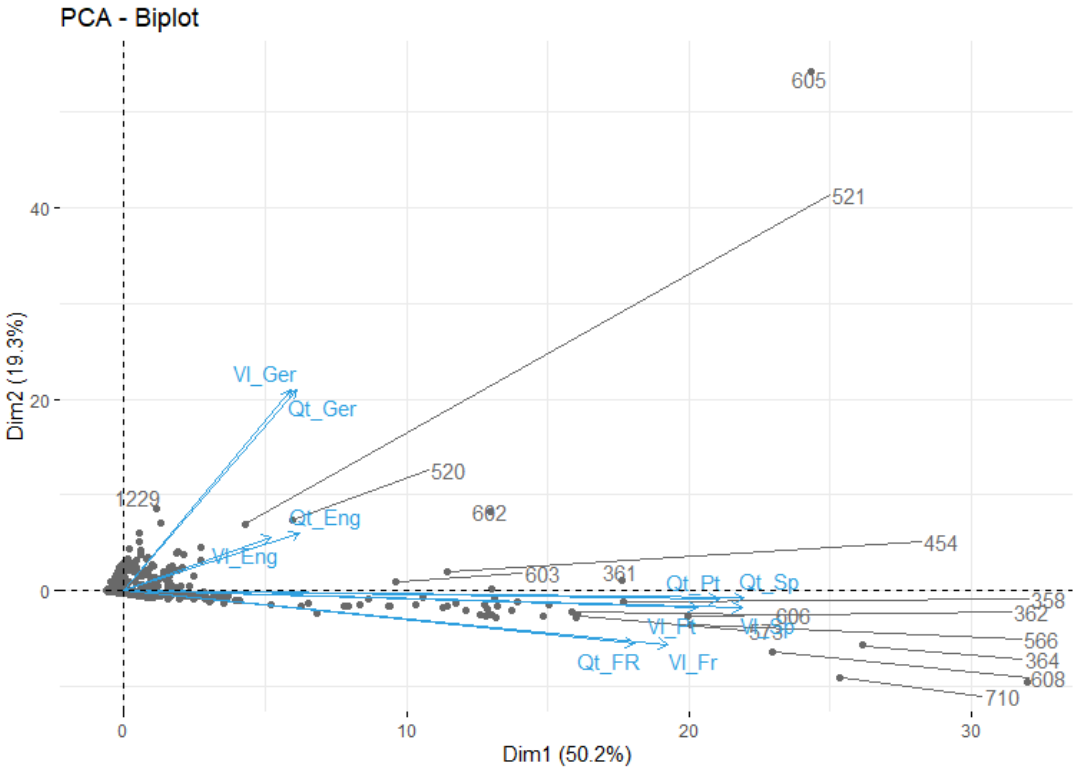

Figura 70: *Biplot* - Gráfico de variáveis e de indivíduos (amostras)

A Figura 71 contém uma representação gráfica com as *labels* de família de produtos. De um modo geral, consegue-se visualizar que existe sobreposição na representação dos dados. No entanto, é possível discriminar alguns grupos de sacos com similaridades nas vendas, nomeadamente os produtos que se encontram a cor de rosa, e laranja.

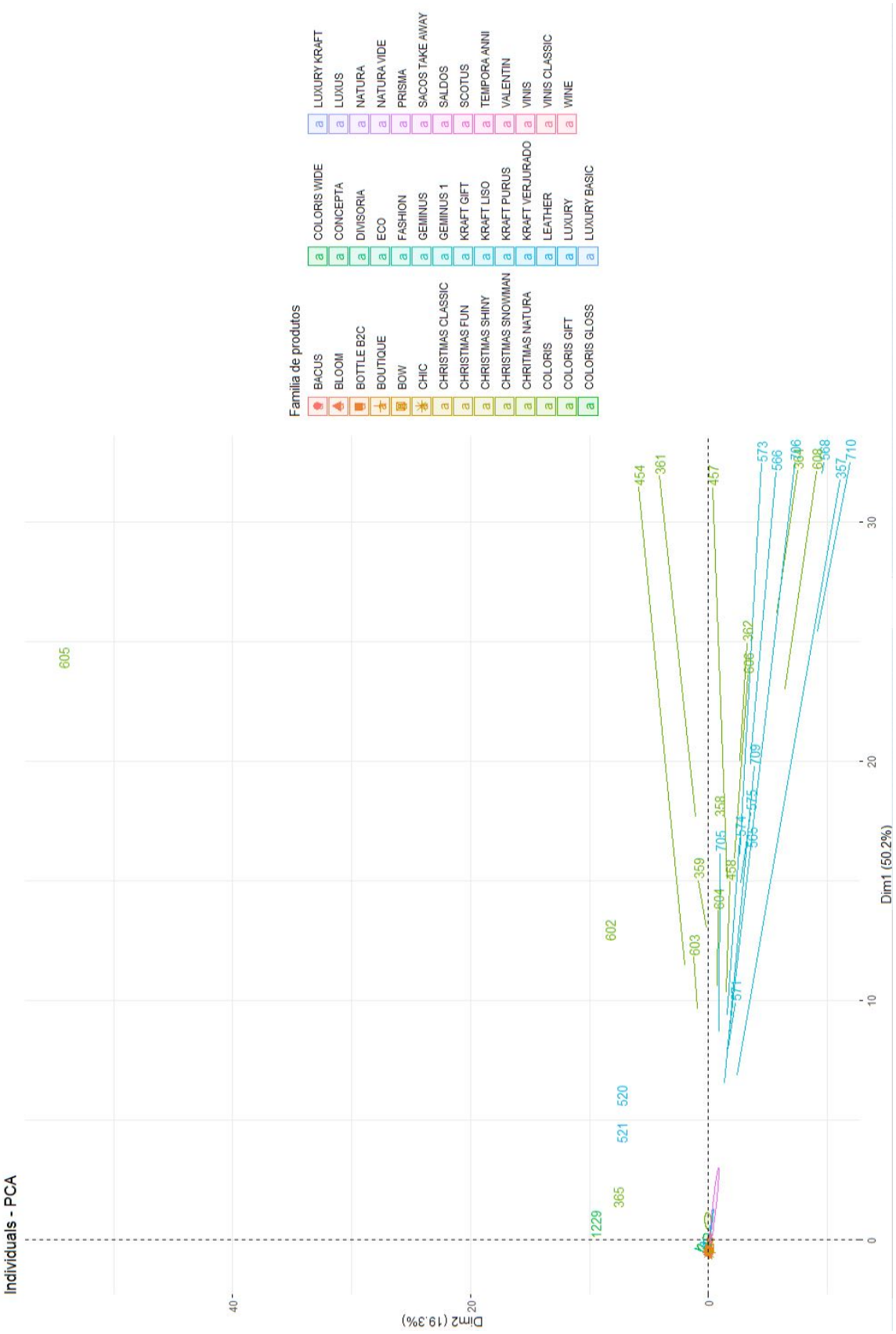

Figura 71: Representação gráfica com as *labels* de famílias de produtos

6.2.2 *k-Means*

Para aplicar o *k-Means* torna-se necessário determinar o número ótimo de *clusters*. Para realizar esta validação foram usados métodos diretos que consistem num critério de otimização tais como soma dos quadrados dentro do *cluster* e silhueta média. Os métodos correspondentes são o método do cotovelo e silhueta respetivamente. Outro método que foi usado para determinar o número ótimo de *clusters* foi os 30 índices para escolha do melhor número de *clusters*, com recurso à função $NbClust()$ do R. Na Figura 72 está presente a representação gráfica do método do cotovelo. Neste método torna-se fundamental localizar o ponto em que a soma dos quadrados dentro do *cluster* é minimizada. Para este método foi identificado um número ótimo de 3 *clusters* (k=3).

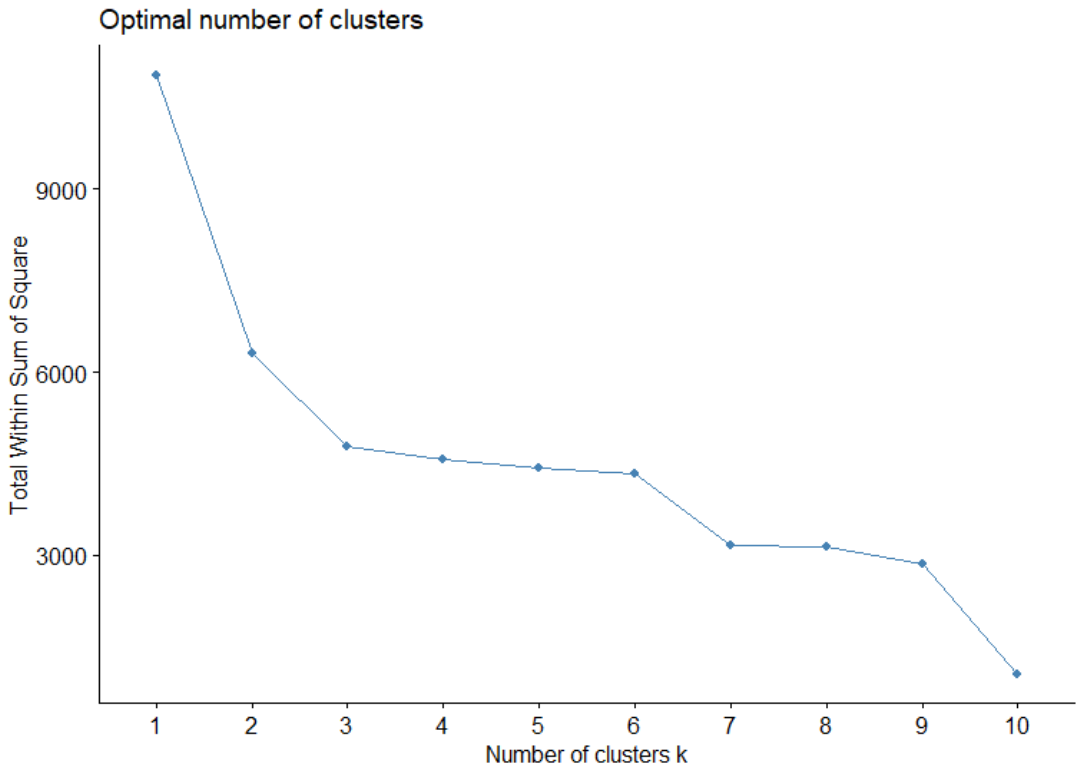

Figura 72: Método do cotovelo

O método da silhueta calcula a silhueta média de observações para diferentes valores de k. O número ótimo de *clusters* k é aquele que maximiza a silhueta média numa dada quantidade de *clusters*. Na Figura 73 está presente a representação gráfica do método da silhueta. Neste método o ponto em que existe uma maximização da silhueta média dentro do *cluster* é um valor de k=2.

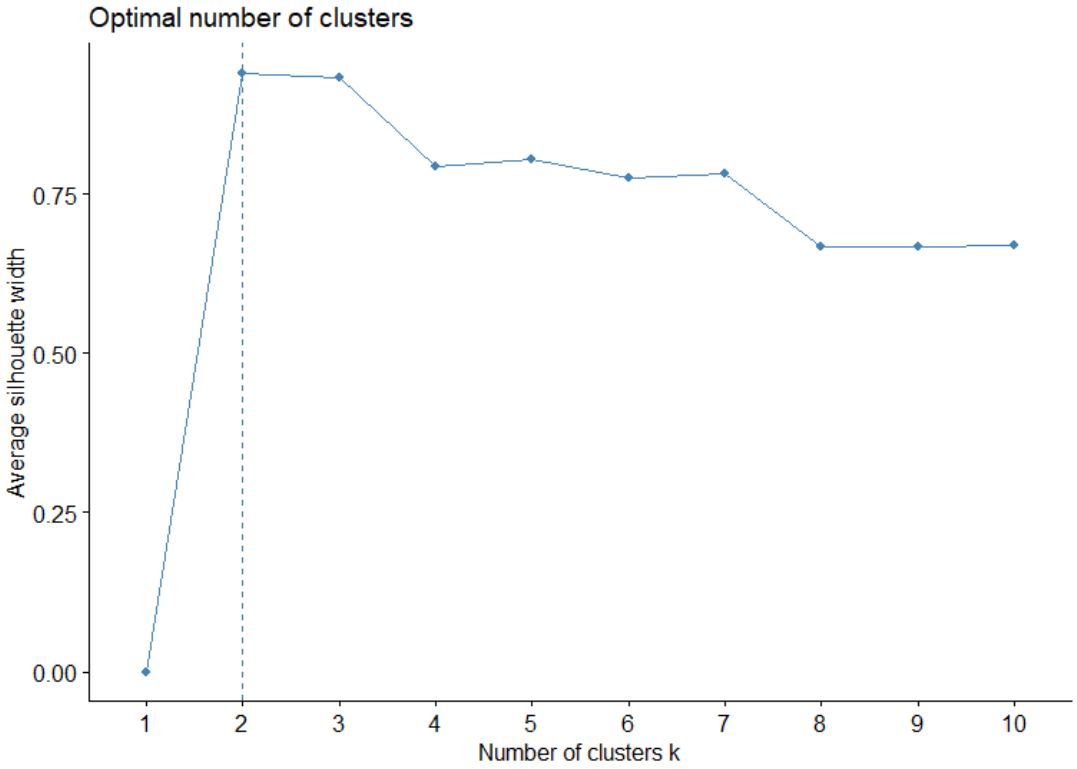

Figura 73: Método da silhueta

Outro método que foi usado para investigar o número ótimo de *clusters* foi os 30 índices para escolha do melhor número de *clusters*, com recurso à função $NbClust()$.

Este método fornece 30 índices para determinar o número ótimo de *clusters* e propõe o melhor esquema de *clustering* a partir dos diferentes resultados obtidos com a variação de todas as combinações de número de *clusters*, medidas de distância e métodos de *clustering*. O método consegue calcular simultaneamente todos os índices e determinar o número de *clusters*. Na Figura 74, está patente o método, e verifica-se que o número ótimo de *clusters* é obtido para k=3.

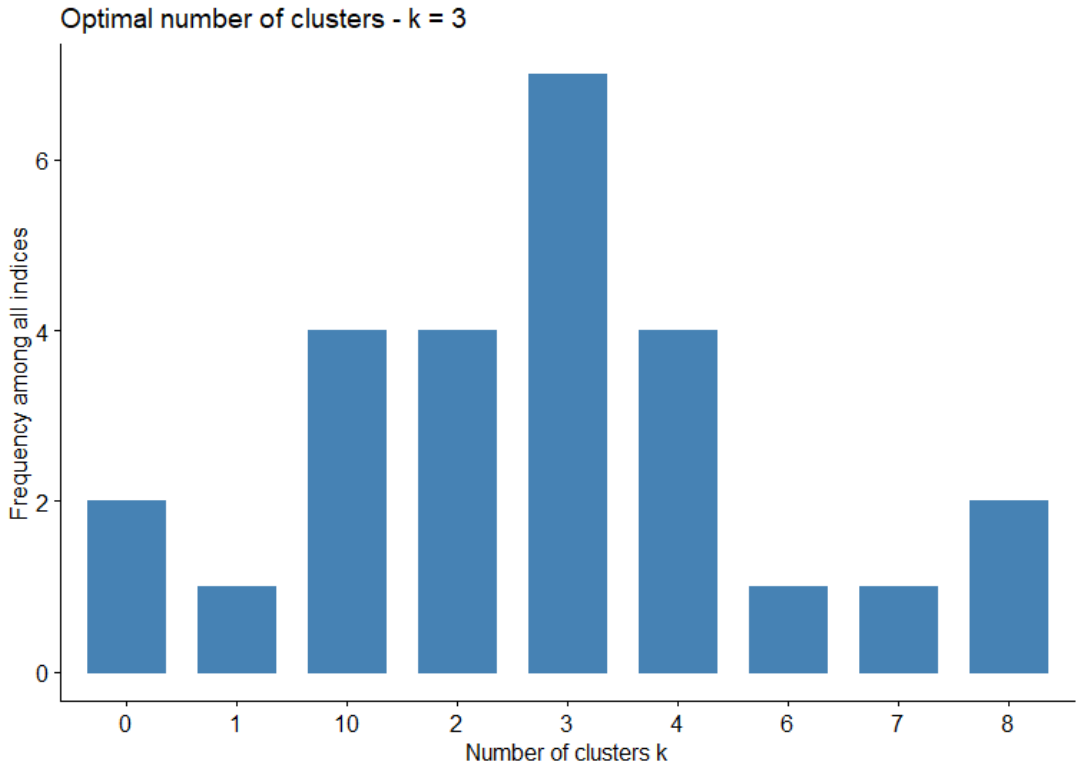

Figura 74: Recurso à função $NbClust()$ para determinar o melhor número de *clusters*

Após a determinação do número ótimo de *clusters* pelos métodos anteriormente mencionados, podemos proceder à implementação do algoritmo *k-Means* com os números de *clusters* recomendados. A Figura 75 apresenta o resultado do algoritmo *k-Means* para k= 2. Com esta representação gráfica, é possível denotar que existem dois grupos bem segmentados. O algoritmo foi capaz de agrupar caraterísticas semelhantes dentro de um mesmo *cluster*.

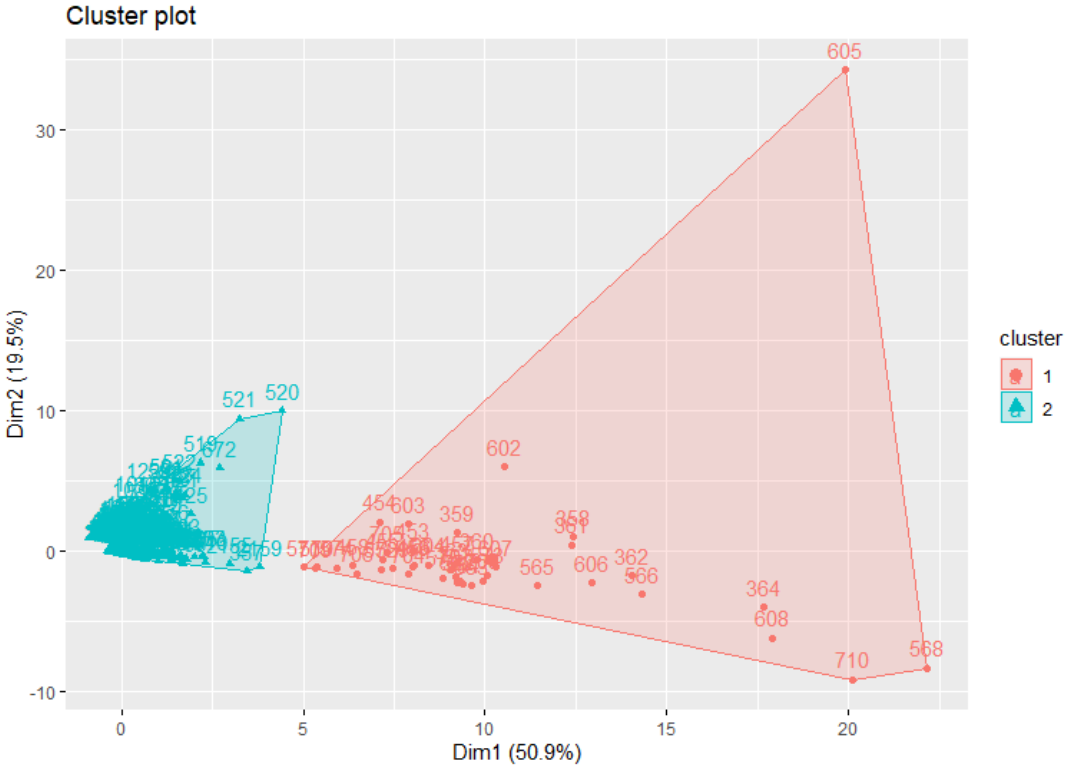

Figura 75: Resultado do *k-Means* com 2 *clusters*

No caso da aplicação do algoritmo *k-Means* (k=3), o *cluster* 2 é formado apenas por um elemento. A dissimilaridade entre este dado e os restantes é elevada, e por isso não foi agregado em nenhum dos *clusters* 1 e 3 (Figura 76).

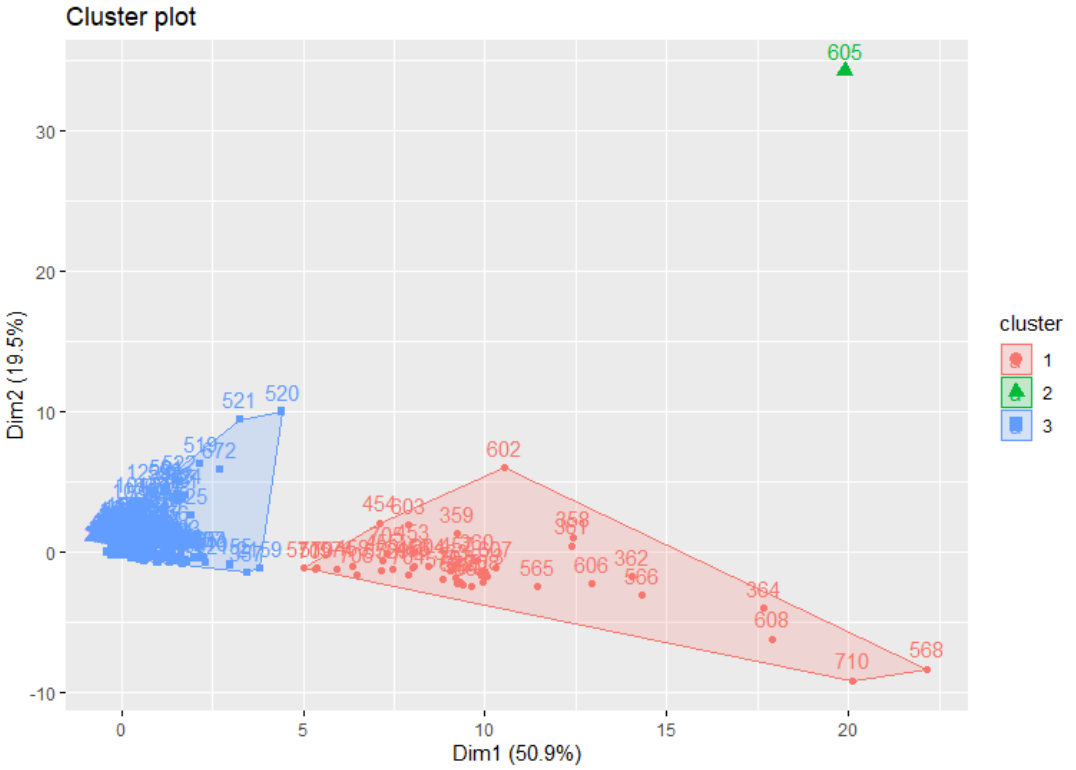

Figura 76: Resultado do *k-Means* com 3 *clusters*

6.2.3 *Clustering* hierárquico

Através do pacote cl_Valid do $\mathbb{R}$, pode-se comparar simultaneamente múltiplos algoritmos de *clustering* através de uma simples função identificando a melhor abordagem de *clustering* e o número ótimo de *clusters*. A Figura 77, refere que o método de *clustering* que obtém um melhor desempenho é o *clustering* hierárquico com 3 *clusters* para as medidas de validação *Dunn e Silhouette*. No caso da medida de validação *Connectivity*, o método com melhor desempenho é o *k-Means* com 3 *clusters*. Independentemente do tipo de algoritmo de agrupamento, o número ótimo de *clusters* é 3 usando as três medidas.

```
Clustering Methods:
 hierarchical kmeans

Cluster sizes:
 3 4 5 6

Validation Measures:
                                  3       4       5       6

hierarchical Connectivity   7.7782 11.7361 15.1028 17.0829
             Dunn           0.4029  0.3196  0.3196  0.1723
             Silhouette     0.9581  0.9513  0.9476  0.9343
kmeans       Connectivity   4.9091 10.6909 15.5401 18.9067
             Dunn           0.0894  0.0383  0.0645  0.0645
             Silhouette     0.9367  0.9339  0.9363  0.9359

Optimal Scores:

             Score  Method       Clusters
Connectivity 4.9091 kmeans       3
Dunn         0.4029 hierarchical 3
Silhouette   0.9581 hierarchical 3
```

Figura 77: Validação do melhor algoritmo de *clustering* (consola do $\mathbb{R}$)

Os valores baixos de *APN, ADM, FOM e AD*, correspondem a *clusters* altamente consistentes. De acordo, com a Figura 78 para a medida de APN, o *clustering* hierárquico com 3 *clusters* é o que obtém um melhor *score*.

```
             Score       Method Clusters
APN 0.0007363071 hierarchical        3
AD  0.6019387477       kmeans        6
ADM 0.0344071277 hierarchical        3
FOM 0.7243011188       kmeans        6
```

Figura 78: Medidas de estabilidade de *clusters* (consola do $\mathbb{R}$)

A Figura 79, revela qual a combinação que representa um melhor agrupamento de dados. Nesta análise, é tida em conta o valor do coeficiente cofenético que reflete na qualidade do agrupamento. Com valores mais próximos de 1, os *clusters* refletem de forma mais precisa os dados (bom agrupamento). De um modo geral, valores acima de 0.75 são considerados bons. A combinação que representa o melhor agrupamento dos dados de acordo com o resultado da consola do $\mathbb{R}$, é a junção da distância euclideana com o método *average*, que possui um coeficiente cofenético de cerca de 0.9773.

```
> correlacao <- function(x,y){
+   res.dist <- dist(dfsacos, method = methodsDist[x])
+   res.hc <- hclust(res.dist, method = methodsHc[y])
+   res.coph <- cophenetic(res.hc)
+   cor(res.dist, res.coph)
+ }
> for(i in 1:3){
+   for (j in 1:3){
+     a <- paste("D:", methodsDist[i], "Hc:", methodsHc[j], "Cor:", correlacao(i, j))
+     print(a)
+   }
+ }
[1] "D: euclidean Hc: single Cor: 0.869130570228446"
[1] "D: euclidean Hc: complete Cor: 0.947584430746842"
[1] "D: euclidean Hc: average Cor: 0.977336592608021"
[1] "D: maximum Hc: single Cor: 0.855169615500611"
[1] "D: maximum Hc: complete Cor: 0.949124066423108"
[1] "D: maximum Hc: average Cor: 0.973101580753166"
[1] "D: manhattan Hc: single Cor: 0.894599585905326"
[1] "D: manhattan Hc: complete Cor: 0.876227382442906"
[1] "D: manhattan Hc: average Cor: 0.975023641361753"
```

Figura 79: Combinações que revelam um melhor agrupamento dos dados (consola do $\mathbb{R}$)

É possível validar o agrupamento realizado de forma interna. Observa-se na Figura 80 que a divisão em 3 *clusters* foi a melhor escolha.

```
$All.index
     3      4      5      6
0.9586 0.9518 0.9481 0.9347

$Best.nc
Number_clusters     Value_Index
         3.0000          0.9586
```

Figura 80: Validação interna do agrupamento ($index$: $silhouette$)

Após realizar as validações necessárias, é possível implementar o algoritmo de *clustering* hierárquico correspondente. A Figura 81 mostra o resultado do dendrograma obtido com as combinações de *clustering* hierárquico (distância euclideana e método *average*). Em 3 grupos é possível denotar que um deles é formado apenas por um único elemento.

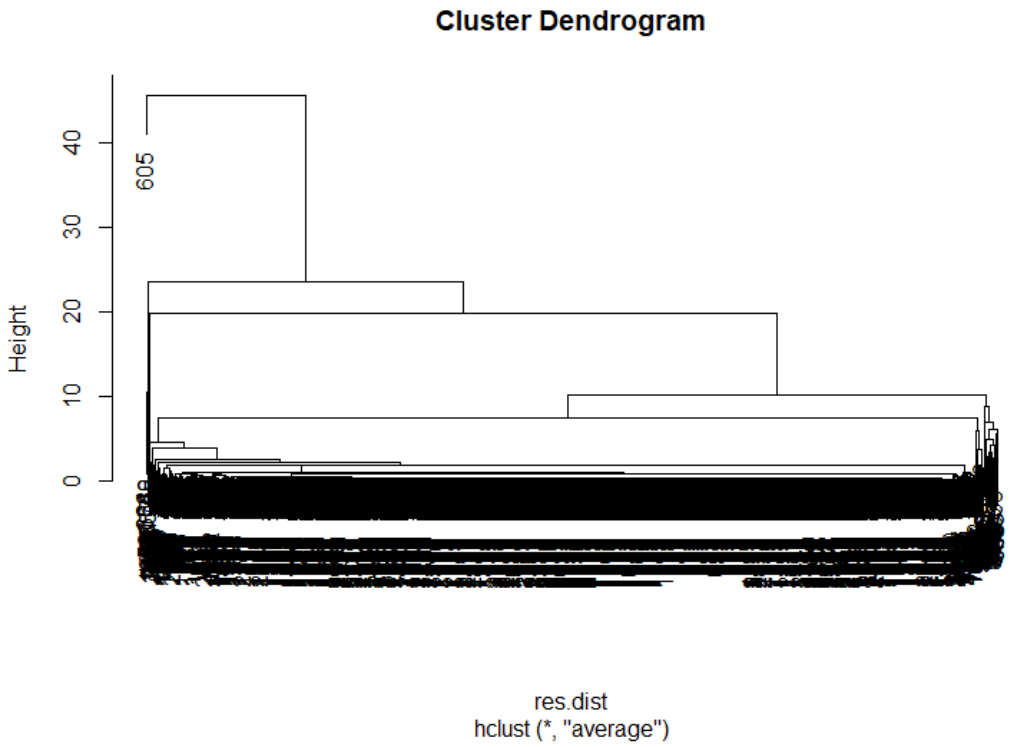

Figura 81: Dendrograma obtido através de *clustering* hierárquico com método *"average"*

A Figura 82 apresenta os três *clusters* obtidos com a aplicação do *clustering* hierárquico. Visualmente verifica-se que a maioria das observações ficaram agrupadas no *cluster* 1, devido à proximidade que mantém entre si. O *cluster* 2 é formado por apenas quatro elementos e o *cluster* 3 com apenas um elemento. Estes *clusters* não apresentam sobreposição entre si, o que permite discriminar bem os grupos.

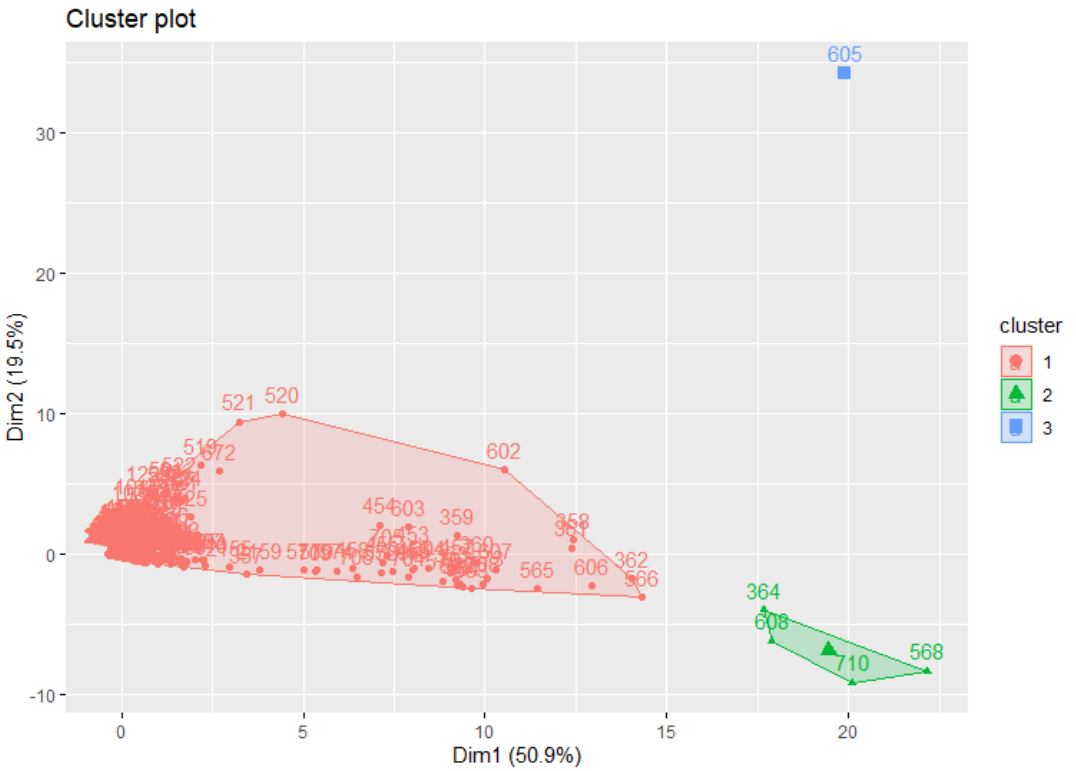

Figura 82: Representação dos *clusters* obtidos com *clustering* hierárquico (método *"average"*)

Outra tentativa efetuada, para o algoritmo de *clustering* hierárquico foi mudar o método para *ward*, e detetar visualmente através do dendrograma o número de *clusters* "ideal". A Figura 83 apresenta o resultado do dendrograma obtido. O dendrograma foi cortado em 3 *clusters*, e os mesmos podem ser visualizados na Figura 84.

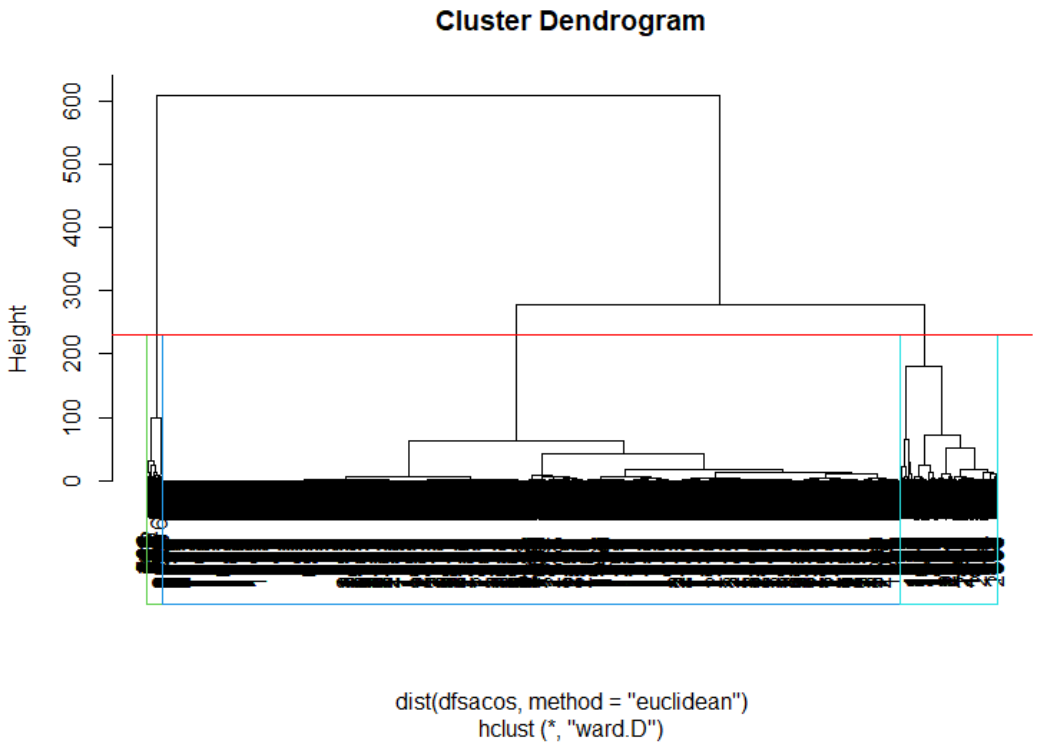

Figura 83: Dendrograma com o método *"ward"*

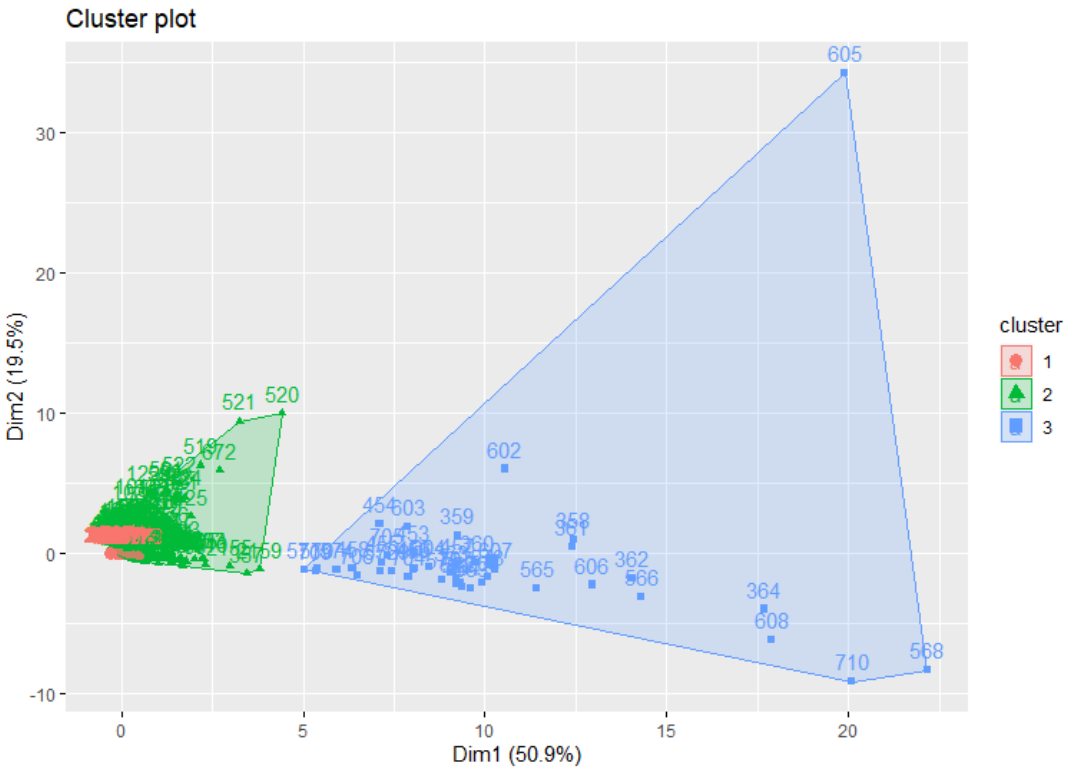

Figura 84: Representação dos *clusters* obtidos com o método de *clustering* hierárquico *"ward"*

A Tabela 16 contém a quantidade de observações que está presente em cada *cluster*. Verifica-se que o *cluster* com maior quantidade de observações é o *cluster* 1, que apresenta grandes quantidades de vendas similares entre si. O *cluster* 3 é aquele que apresenta menor número de observações.

Tabela 16: Quantidade de observações em cada *cluster*

Cluster	**n**
1	1879
2	250
3	42

6.2.4 Avaliação de agrupamentos

Analisando os *plots* de silhueta dos algoritmos de *k-Means* (2 e 3 *clusters*), e *clustering* hierárquico (método *average e ward*), consegue-se concluír que o *clustering* hierárquico método *average* produziu melhores agrupamentos (coeficiente de silhueta de cerca de 0.96). O algoritmo de *k-Means* para k=2 obteve um coeficiente de silhueta (0.94), e permitiu segmentar bem os grupos. Apesar do algoritmo de *clustering* hierárquico método *ward*, obter o coeficiente de silhueta mais baixo, e apresentar alguma sobreposição entre grupos, permitiu segmentar de forma satisfatória um dos grupos (Figuras 85, 86, 87 e 88).

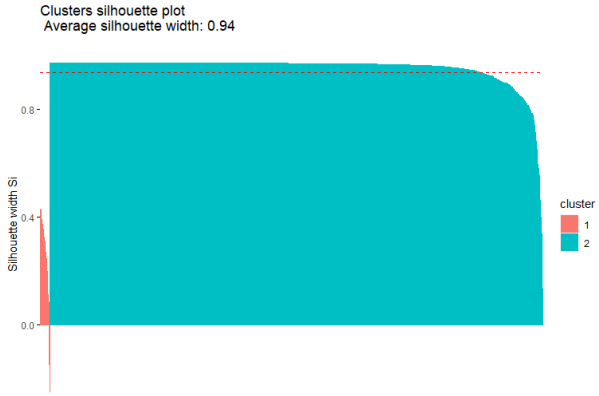

Figura 85: Validação *silhouette* para k=2 (*k-Means*)

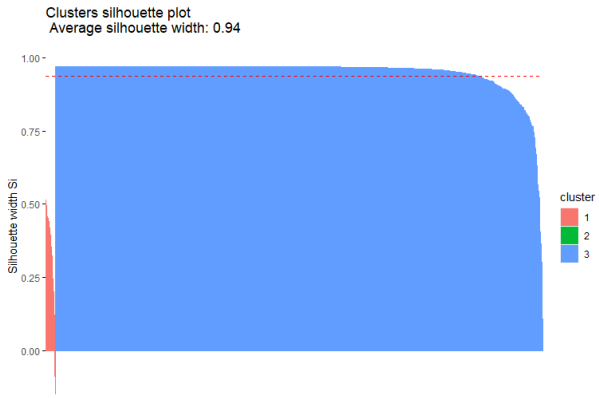

Figura 86: Validação *silhouette* para k=3 (*k-Means*)

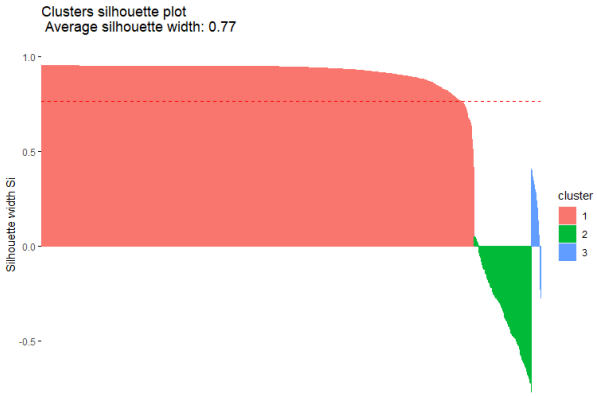

Figura 87: Validação *silhouette* para k=3 (*Clustering* hierárquico - Método *"ward"*)

Figura 88: Validação *silhouette* para k=3 (*Clustering* hierárquico - Método *"average"*)

Capítulo 7

Conclusão e trabalho futuro

Com a realização desta dissertação em ambiente empresarial foi possível colocar em prática conhecimentos de *machine learning*, mais precisamente na vertente de aprendizagem não supervisionada.

Com a análise exploratória dos dados verificou-se que dependendo do país as quantidades de vendas dos produtos diferiam, pelo que Portugal e Espanha apresentavam quantidades de vendas maiores que os restantes países. As vendas de caixas para Inglaterra não foram evidenciadas em nenhuma das observações da base de dados, o que denota que o mercado Inglês não apostou neste tipo de produtos ao longo dos anos. Relativamente aos sacos, este é um dos produtos que gera mais lucro para a Litel, Lda, apresentando máximos de vendas na ordem de 10^6 unidades vendidas por ano. Ao longo dos anos, verifica-se que a empresa Litel, Lda, apresentou um decréscimo nas vendas de 2015 a 2016, e um aumento nas vendas de 2016 até 2019. Em 2020, verificou-se uma quebra nas vendas bastante acentuada devido à pandemia do Covid-19. Posteriormente, em 2021 as vendas aumentaram substancialmente atingindo um pico máximo de vendas.

Numa análise geral, e considerando toda a base de dados das caixas conclui-se que as caixas mais vendidas são *GOURMET M (14766)*, *CLASSIC (15129, 15433, 15434)* e *B2C PACKAGING (15799)*. Relativamente às caixas menos vendidas destacam-se *ELISÉE HYPE (15089)*, *GOURMET HAUTE M (15453)*, e *GOURMET HAUTE L (15459, 15460)*. Quanto à base de dados dos sacos, os que apresentam maiores quantidade de vendas são *COLORIS (61000, 61032)*, *COLORIS WIDE (61100)* e *KRAFT LISO (61068, 61120)*. Por outro lado, os sacos com menos vendas são *COLORIS GIFT (30665, 30681, 30688, 30689)* e *KRAFT GIFT (30691)*.

A aplicação do PCA sobre as variáveis das bases de dados de caixas e sacos permitiu reduzir redundâncias nos dados, identificar variáveis correlacionadas, reduzir a dimensionalidade do *dataset*, e identificar padrões de venda ocultos nos dados. As componentes principais permitiram expor os dados de forma a evidenciar famílias de produtos com similaridade de vendas. No caso do *dataset* das caixas foi possível identificar vários padrões de venda similares entre famílias de produtos. Como exemplo, pode ser destacado o padrão de vendas formado pelas famílias *VINTAGE*, *SUPPORTO*, *SMART BOX* e *SPARKLING*. Um outro padrão de vendas surge entre as famílias de produtos *ELISEE CLASSIC*, *ELISEE HYPE*, *FILUM* e *GOURMET DOMINUS*. No caso do *dataset* dos sacos, pode ser mencionado o padrão de vendas formado pelas famílias *BACUS*, *BLOOM* e *BOTTLE B2C*. Outro exemplo de padrão de vendas similares é entre as famílias *WINE*, *VINIS CLASSIC*, *VINIS* e *VALENTIN*.

Foram aplicados ainda algoritmos de *clustering*, para agrupar os dados de vendas. Uma das tarefas primordiais, foi a determinação do número ótimo de *clusters*, através de vários métodos (cotovelo, silhueta, 30 índices para escolher o melhor número de *clusters*). Foram validados todos os métodos, e procedeu-se à identificação do número ótimo de *clusters*, de forma a implementar os respetivos algoritmos. Os resultados obtidos com *clustering* foram no geral bastante satisfatórios, pelo que foi possível identificar grupos, e testar algoritmos de *k-Means* e *clustering* hierárquico variando diferentes parâmetros. Na base de dados das caixas e sacos, destacam-se os algoritmos *k-Means* (k=2), e *clustering* hierárquico - método "*average*"com os melhores resultados na tarefa de segmentação de *clusters*.

Como trabalhos futuros sugere-se: a interpretação dos *clusters* obtidos no contexto do problema, a geração de novas variáveis explicativas que permitam realizar uma análise da segmentação de grupos de forma mais eficaz, a recolha de dados por parte da empresa de forma mais consistente ao longo do tempo, para que seja possível fazer uma análise de *forecasting* por séries temporais, e a criação de uma interface gráfica que permita ao utilizador carregar o *dataset*, de forma a serem disponibilizados histogramas e outros recursos estatísticos exploratórios para realizar uma análise automática. Em suma, todos os objetivos deste trabalho foram cumpridos.

Referências

Andrecut, M. (2009). Parallel gpu implementation of iterative pca algorithms. *Journal of Computational Biology*, 16(11):1593–1599.

Arga Felani, D. (2015). Perbandingan 3 metode dalam data mining untuk menentukan strategi penjualan produk makanan dan minuman pada toserba lestari baru gemolong.

Bholowalia, P. and Kumar, A. (2014). Ebk-means: A clustering technique based on elbow method and k-means in wsn. *International Journal of Computer Applications*, 105(9).

Catalog Player Smart Content, B. S. (cited March 2022). Five ways machine learning can improve your sales. `https://catalogplayer.com/en/uncategorized/five-ways-machine-learning-can-improve-your-sales/9872/`.

Dias, G. (cited May 2022). Análise de cluster - método do cotovelo. `https://rpubs.com/diascodes/770518`.

Dolnicar, S. (2003). Using cluster analysis for market segmentation-typical misconceptions, established methodological weaknesses and some recommendations for improvement. 2003.

Domingos, P. (2015). The master algorithm: How the quest for the ultimate learning machine will remake our world.

Hamerly, G. and Elkan, C. (2004). Learning the k in k-means. *Advances in neural information processing systems*, pages 281–288.

Han, J., Kamber, M., and Pei, J. (2011). Data mining: Concepts and techniques.

Irdiansyah, E. (2010). Penerapan data mining pada penjualan produk minuman di pt. pepsi cola indobeverages menggunakan metode clustering.

Johnson, A. (1992). Applied multivariate statistical analysis. *Prentice hall Englewood Cliffs*, 4(11).

Kodinariya, T. M. and Makwana, P. R. (2013). Review on determining number of cluster in k-means clustering. *International Journal of Advance Research in Computer Science and Management Studies*, 1(6):90–95.

Leonard, K. (cited May 2022). The role of data in business. https://smallbusiness.chron.com/role-data-business-20405.html.

Li, H. (2018). Which machine learning algorithm should i use?".

Litel, L. (cited September 2022a). História da litel. https://www.litelonline.com/pt/empresa/sobre-a-litel/historia_341.html.

Litel, L. (cited September 2022b). litel - soluçoes para embalagem. https://www.litel.pt/.

Litel, L. (cited September 2022c). Mercados da litel. https://www.litelonline.com/pt/empresa/sobre-a-litel/mercados_340.html.

Litel, L. (cited September 2022d). Politica e sustentabilidade da litel. https://www.litelonline.com/pt/empresa/sobre-a-litel/sustentabilidade_339.html.

Loon, R. V. (2022). Machine learning explained: Understanding supervised, unsupervised reinforcement learning. http://www.ronaldvanloon.com/machine-learning-explained-understanding-supervised-unsupervised-learning/.

Matt, O. (cited May 2022). 10 tips for choosing the optimal number of clusters. https://towardsdatascience.com/10-tips-for-choosing-the-optimal-number-of-clusters-277e93d72d92.

Naik, A. (2022). Hierarchical clustering algorithm. https://sites.google.com/site/dataclusteringalgorithms/hierarchical-clustering-algorithm.

Nomidl (cited May 2022). What are different types of machine learning algorithms? https://www.nomidl.com/machine-learning/machine-learning-interview-questions-part-1/.

Oyelade, O. J., Oladipupo, O. O., and Obagbuwa, I. C. (2010). Application of k-means clustering algorithm for prediction of students academic performance. *International Journal of Computer Science and Information Security*, 7(1):292–295.

Punj, G. and Stewart, D. W. (1983). Cluster analysis in marketing research: Review and suggestions for application. *Journal of marketing research*, pages 134–148.

Reddy, C. (2022). Understanding the concept of hierarchical clustering technique. https://towardsdatascience.com/understanding-the-concept-of-hierarchical-clustering-technique-c6e8243758e.

Ross (cited May 2022). Análise de cluster: diana, daisy e pam. https://blog.metodosquantitativos.com/cluster/.

SAS (cited May 2022). O que é a mineração de dados? https://www.sas.com/pt_br/insights/analytics/mineracao-de-dados.html.

Segaran, T. (2007). Programming collective intelligence.

Selim, S. Z. and Ismail, M. A. (1984). K-means-type algorithms: A generalized convergence theorem and characterization of local optimality. *IEEE Transactions on pattern analysis and machine intelligence*, 1:81–87.

Steinbach, M., Karypis, G., and Kumar, V. (2000). A comparison of document clustering techniques. *KDD workshop on text mining*, 400(1):525–526.

Strehl, A., Ghosh, J., and Mooney, R. (2000). Impact of similarity measures on web-page clustering. *In Workshop on Artificial Intelligence for Web Search (AAAI 2000)*, pages 58–64.

Taufiq Luthfi, E. (2009). L-moments: Penerapan data mining algoritma asosiasi. *JURNAL DASI*, (2).

Xu, R. and Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3):645–678.

Yann, L., Bottou, L., Orr, G. B., and Muller, K.-R. (1998). Efficient backprop. *Neural Networks: Tricks of the Trade*, pages 9–48.

Yse, D. L. (2022). A complete guide to k-means clustering algorithm. `https://www.kdnuggets.com/2019/05/guide-k-means-clustering-algorithm.html`.

Anexo A

Gráficos complementares

A.1 Quantidade de Vendas (Qt_Pais)

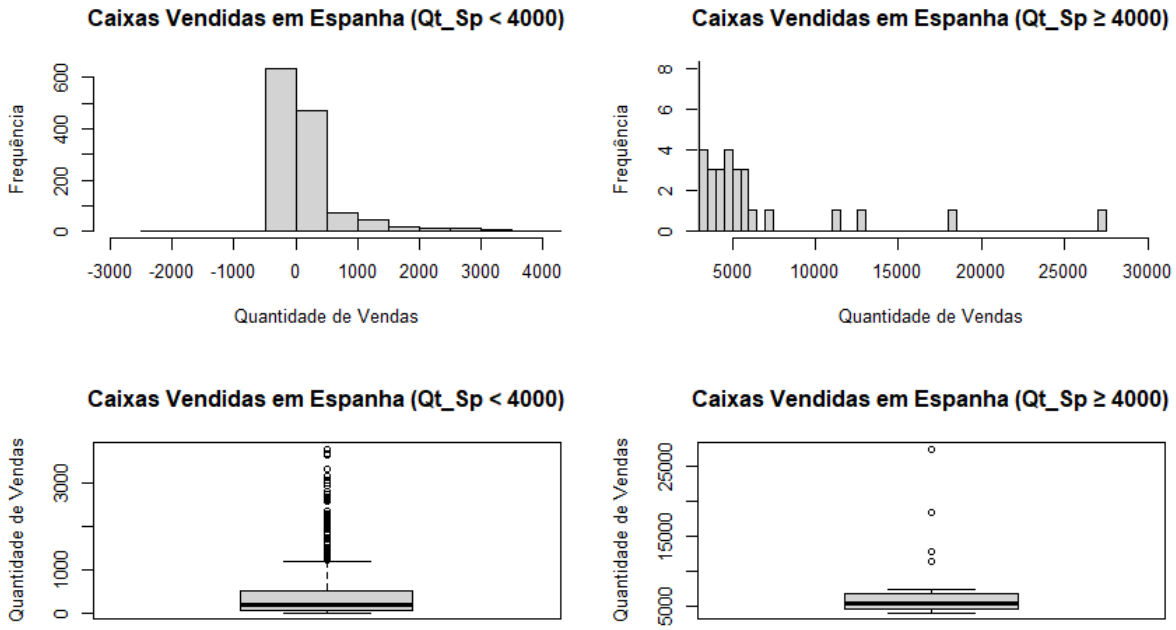

Figura A.1.1: Histograma e *boxplot* para a variável Qt_Sp

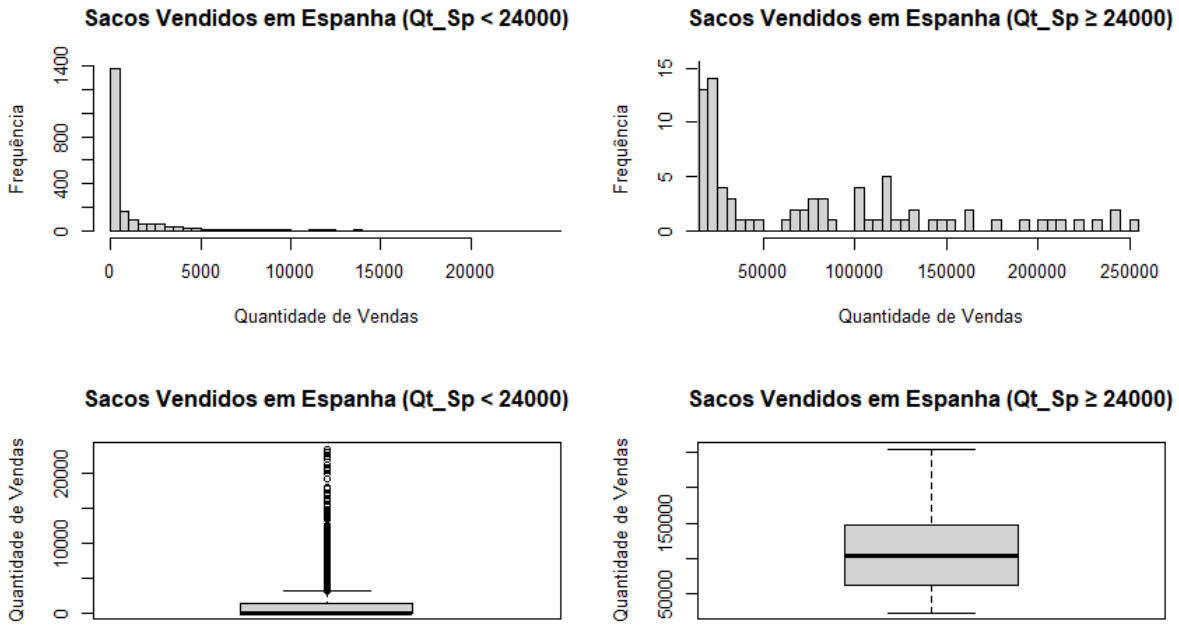

Figura A.1.2: Histograma e *boxplot* para a variável Qt_Sp

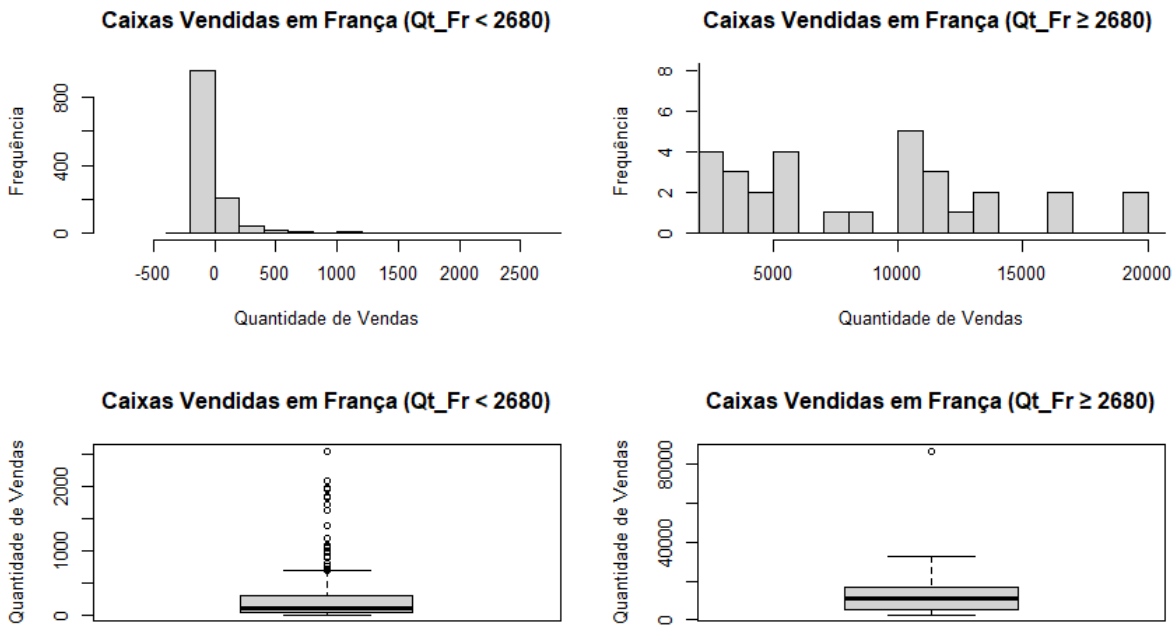

Figura A.1.3: Histograma e *boxplot* para a variável Qt_Fr

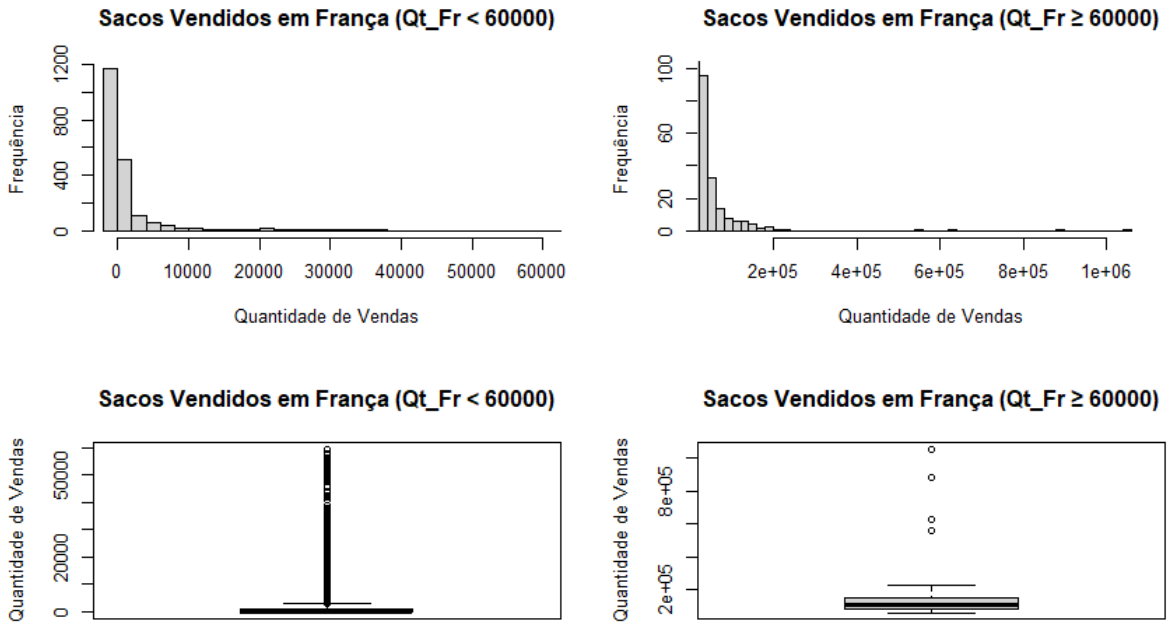

Figura A.1.4: Histograma e *boxplot* para a variável Qt_Fr

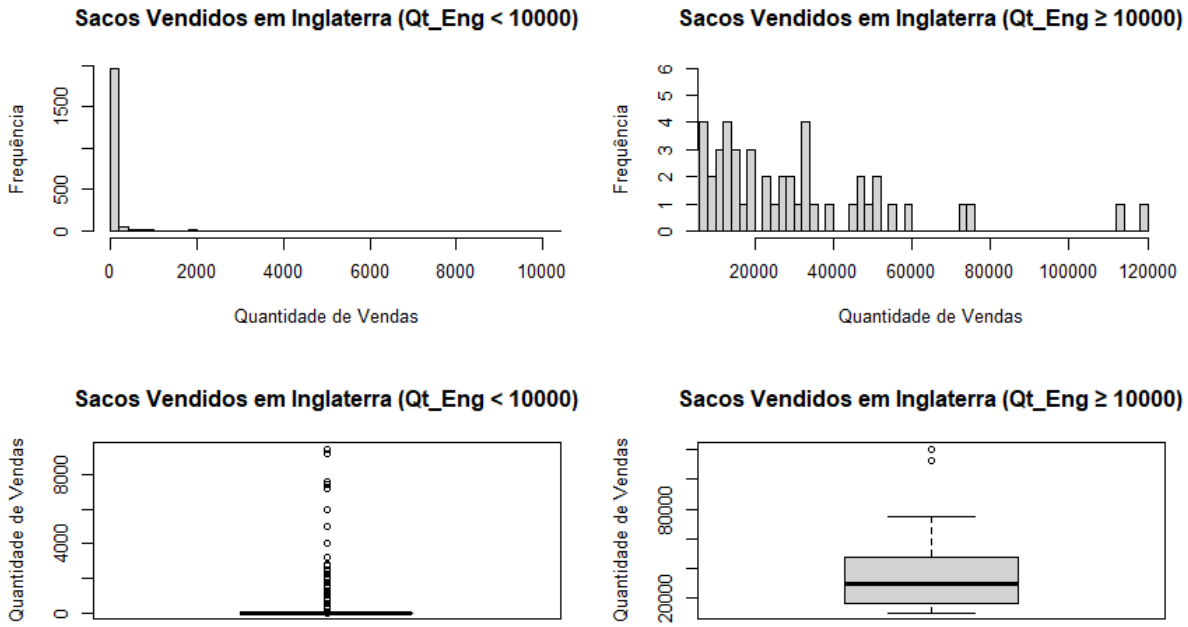

Figura A.1.5: Histograma e *boxplot* para a variável Qt_Eng

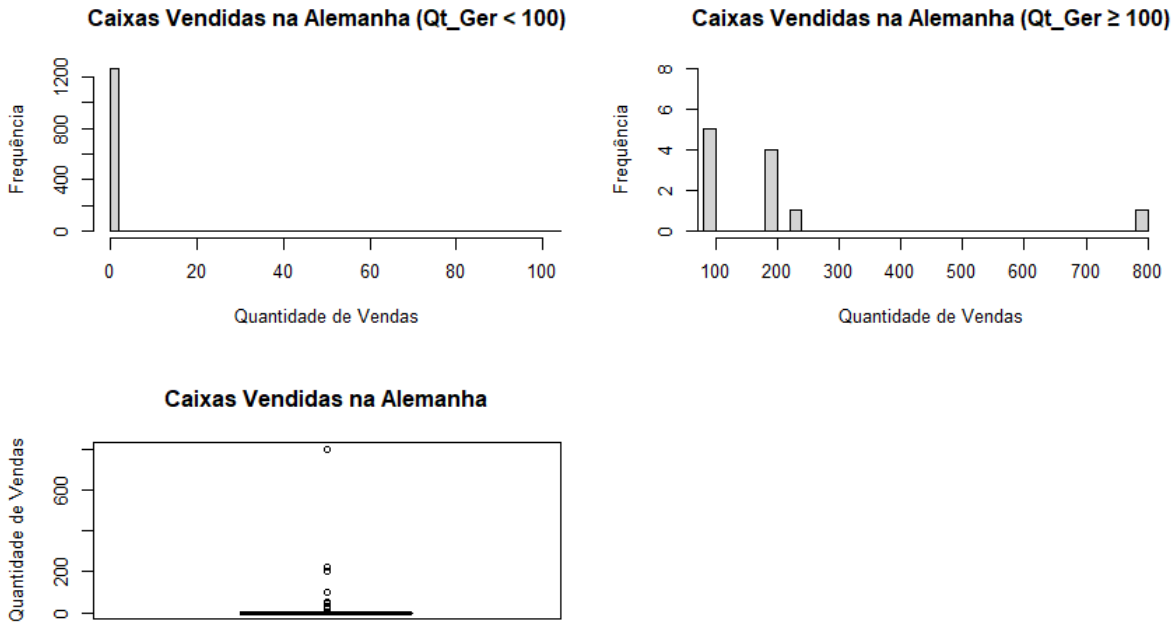

Figura A.1.6: Histograma e *boxplot* para a variável Qt_Ger

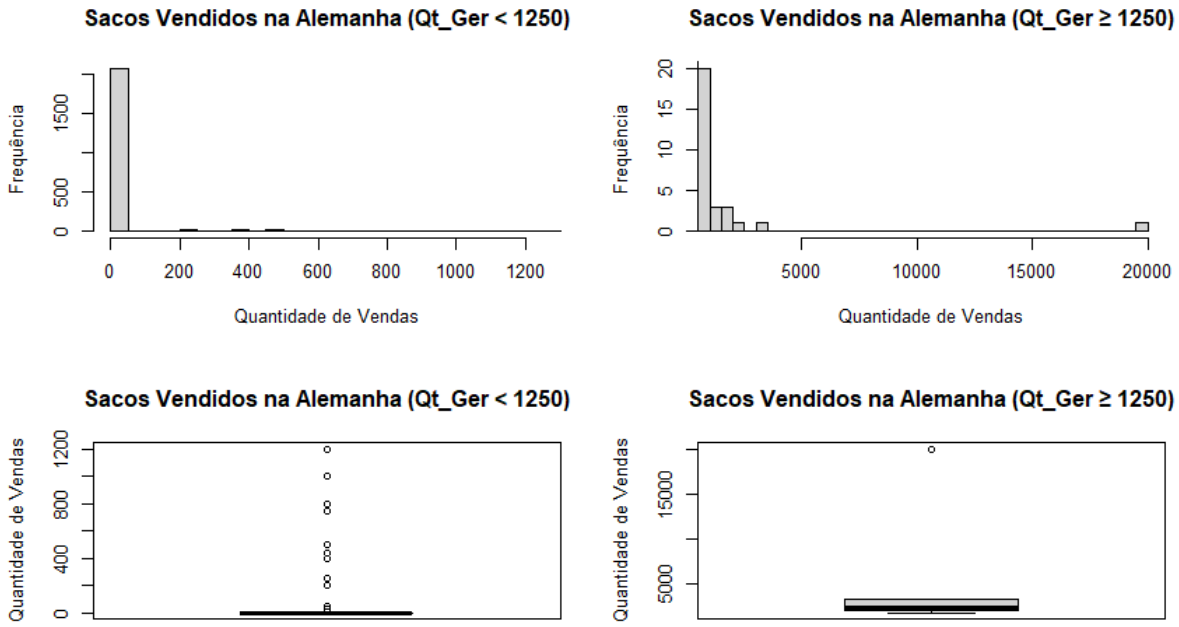

Figura A.1.7: Histograma e *boxplot* para a variável Qt_Ger

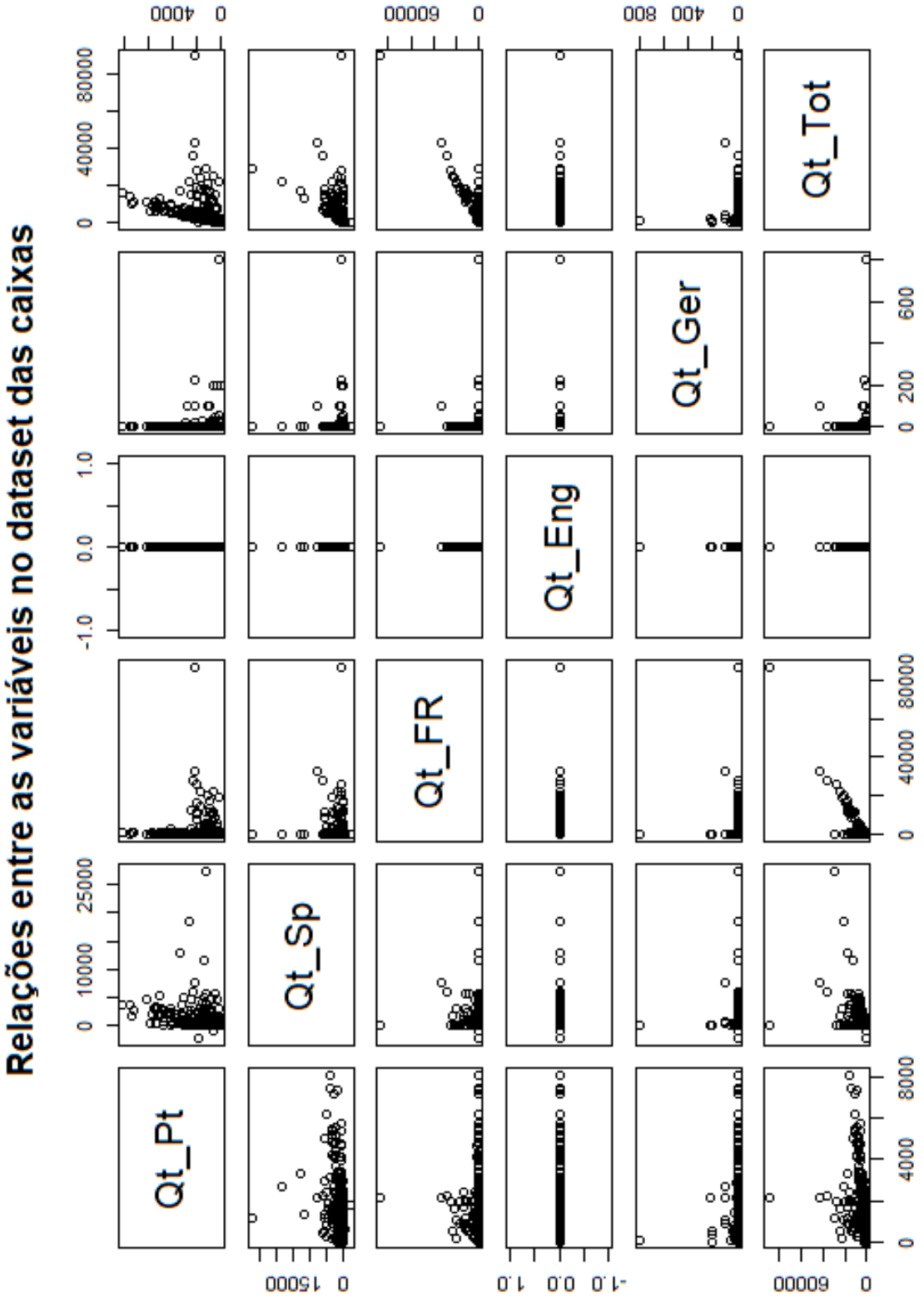

Figura A.1.8: *Scatterplot* do *dataset* de caixas (variáveis da quantidade de vendas)

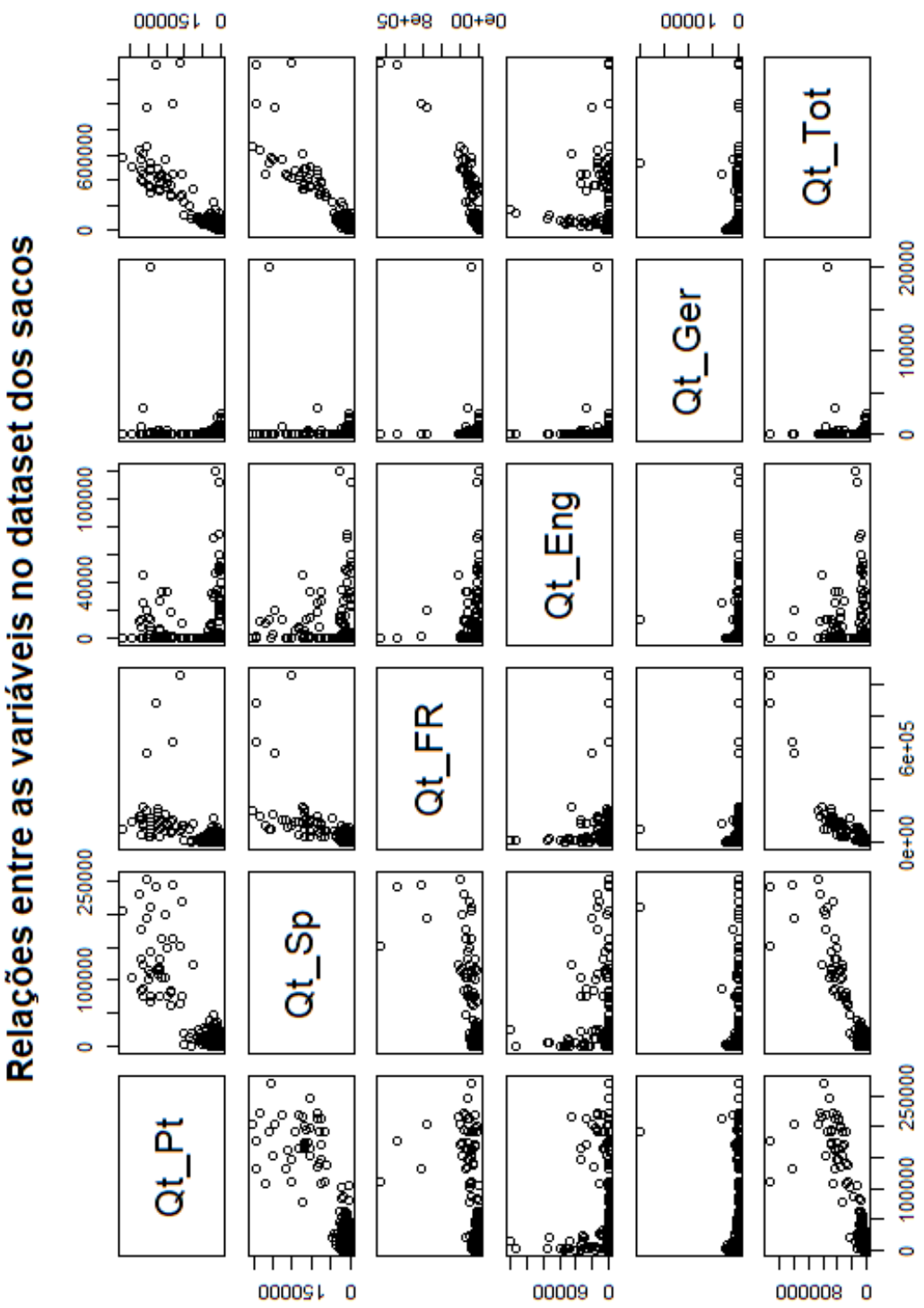

Figura A.1.9: $Scatterplot$ do *dataset* de sacos (variáveis da quantidade de vendas)

A.2 Valor de Vendas (Vl_Pais)

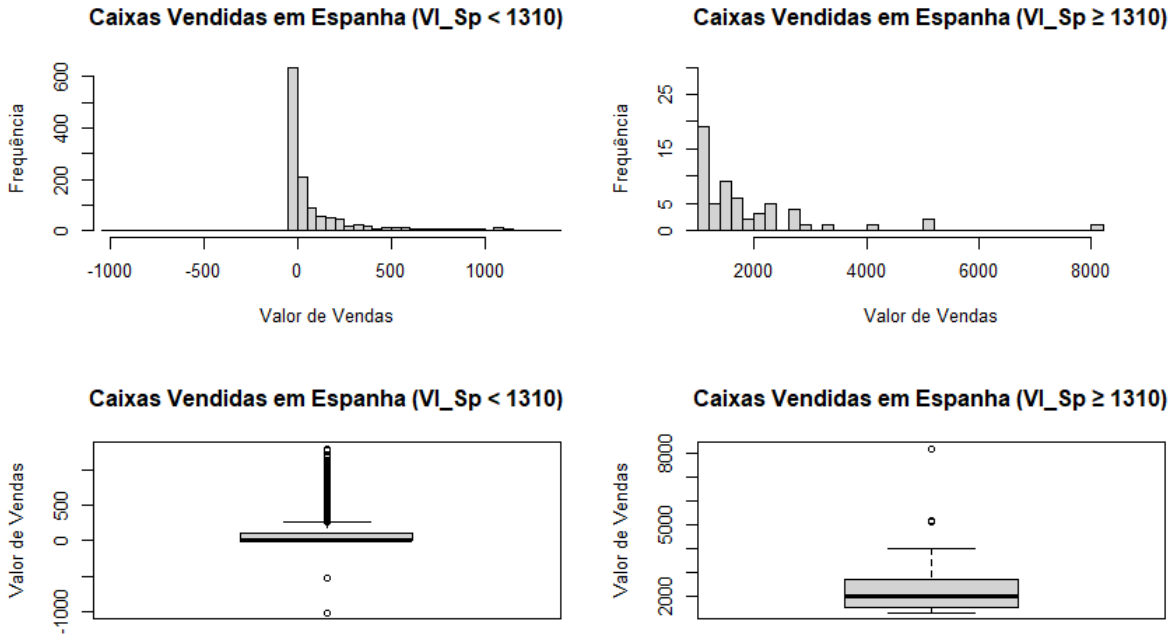

Figura A.2.1: Histograma e *boxplot* para a variável Vl_Sp

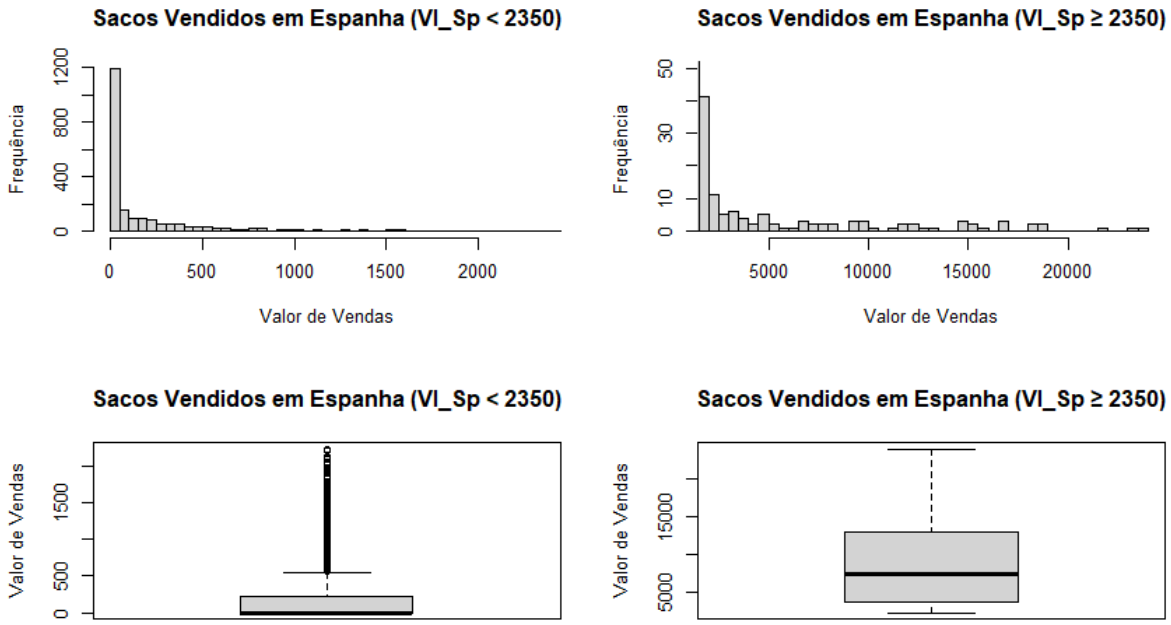

Figura A.2.2: Histograma e *boxplot* para a variável Vl_Sp

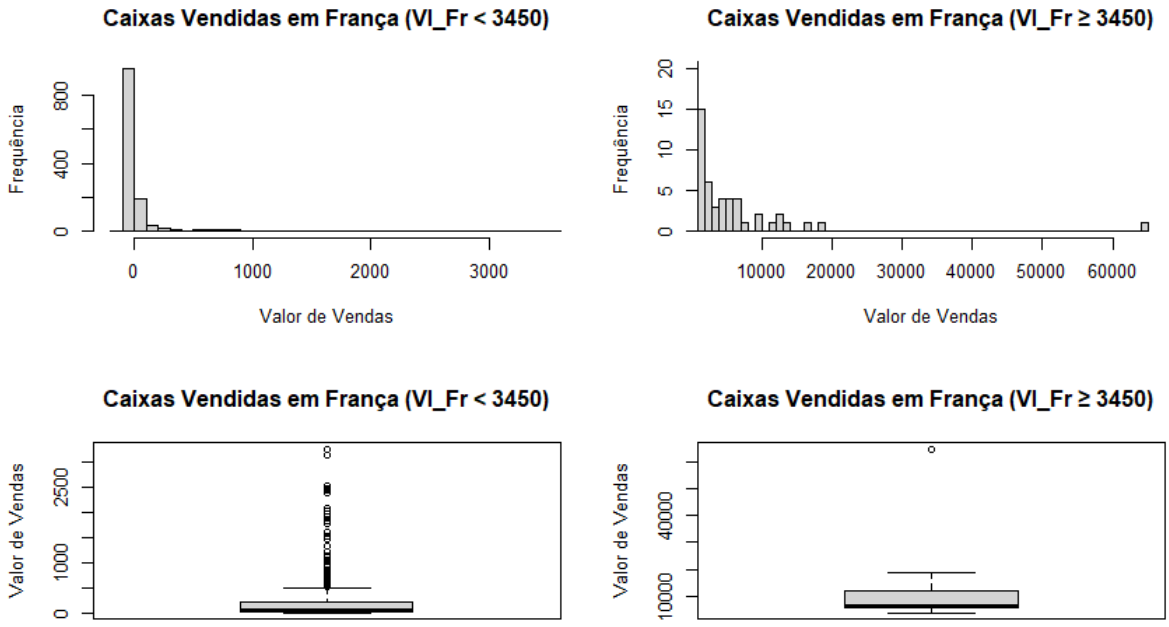

Figura A.2.3: Histograma e *boxplot* para a variável Vl_Fr

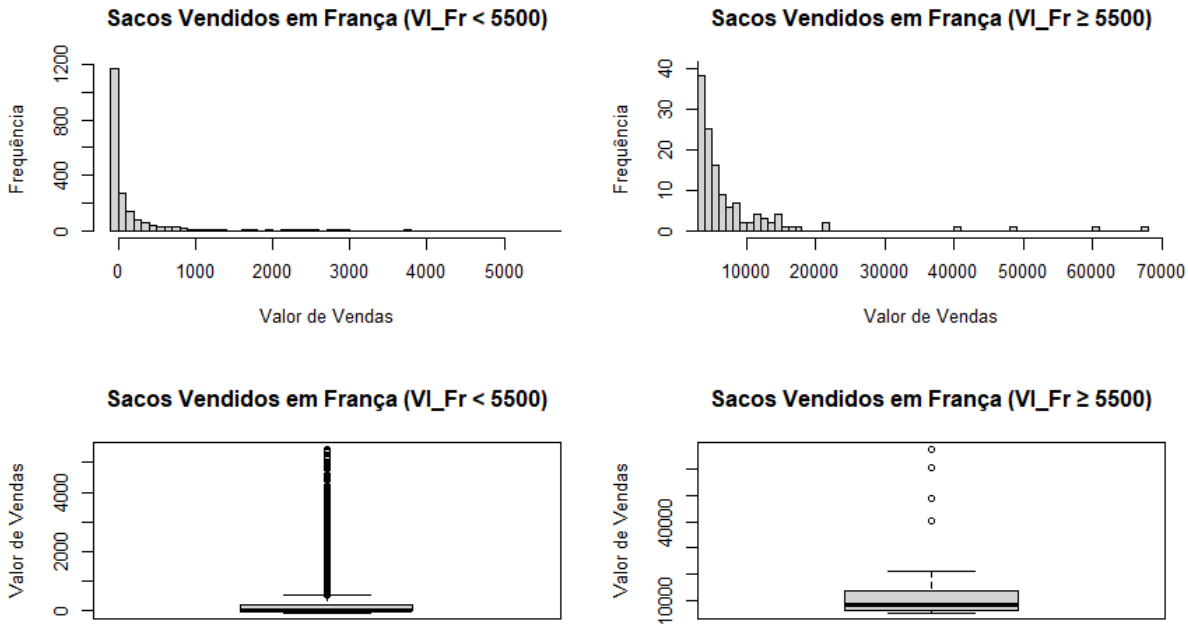

Figura A.2.4: Histograma e *boxplot* para a variável Vl_Fr

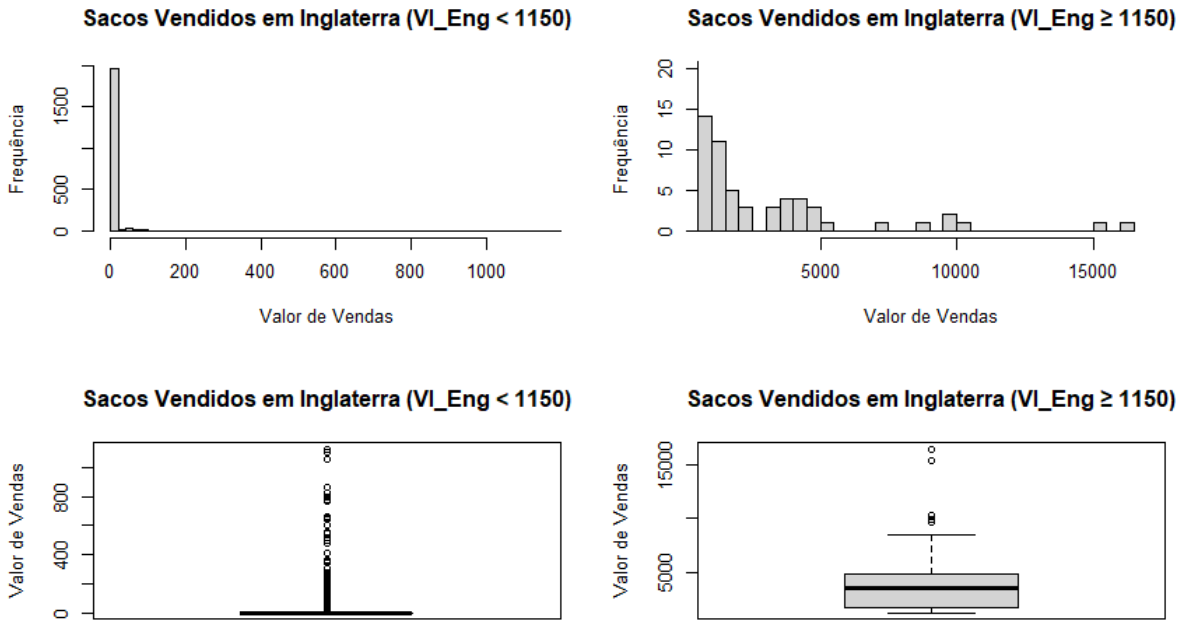

Figura A.2.5: Histograma e *boxplot* para a variável Vl_Eng

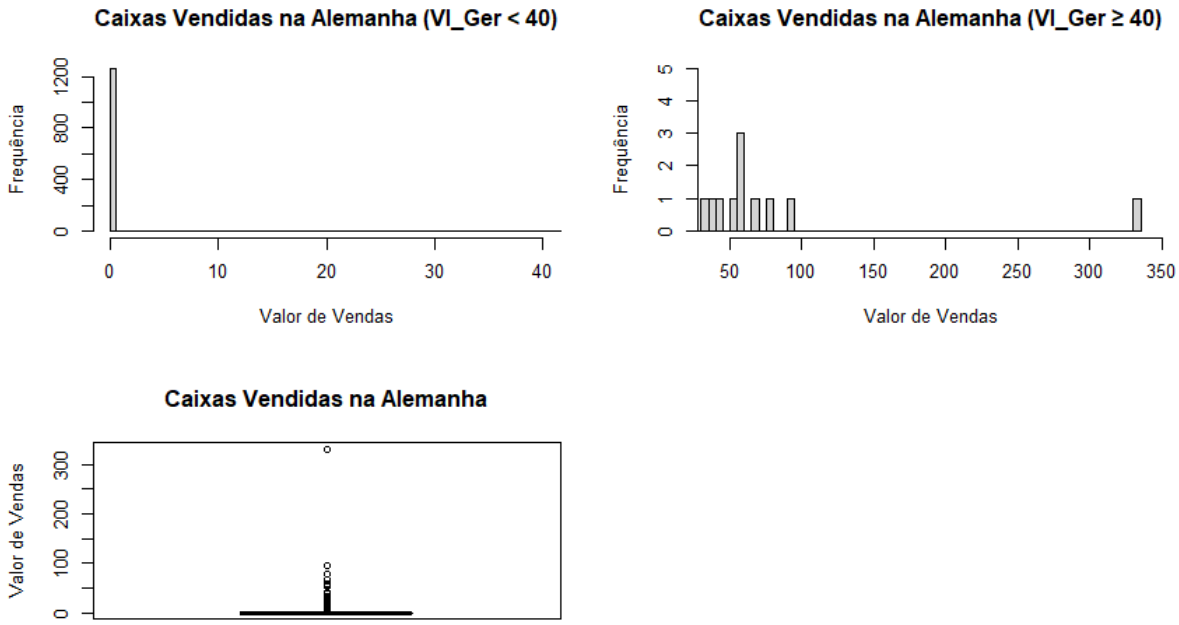

Figura A.2.6: Histograma e *boxplot* para a variável Vl_Ger

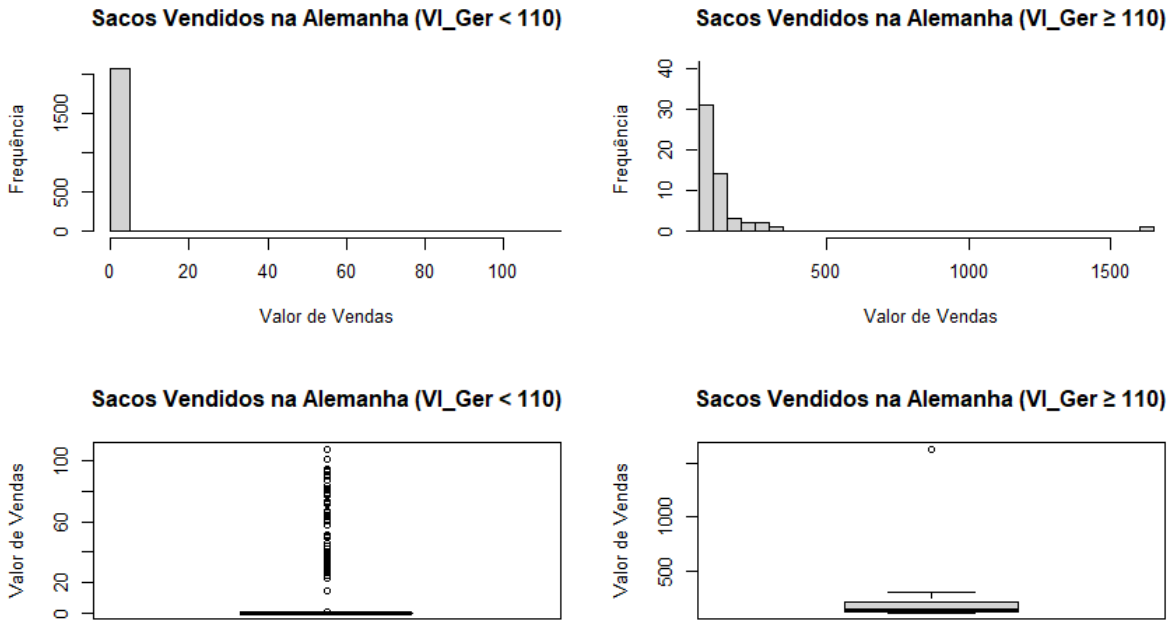

Figura A.2.7: Histograma e *boxplot* para a variável Vl_Ger

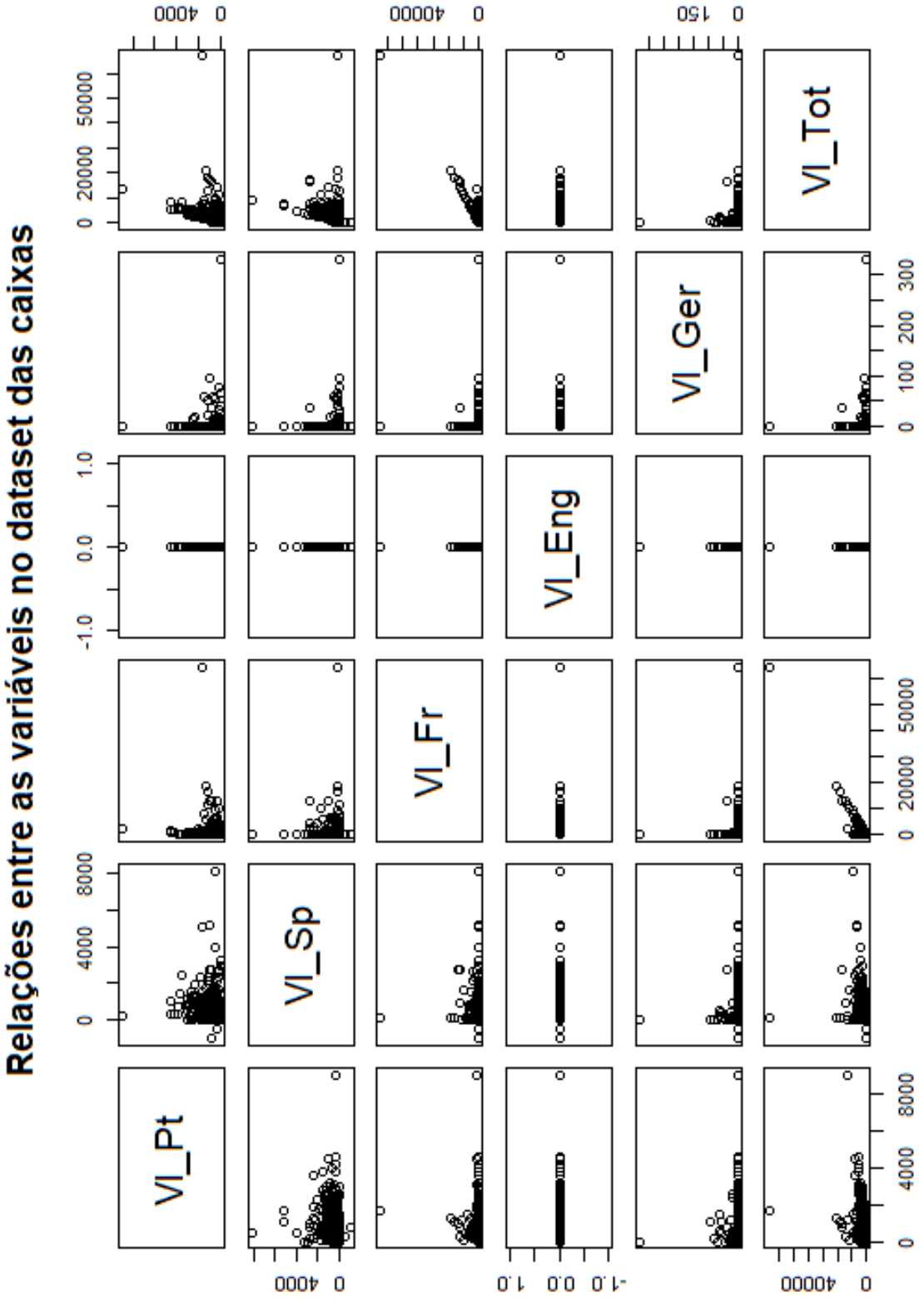

Figura A.2.8: *Scatterplot* do *dataset* de caixas (variáveis do valor de vendas)

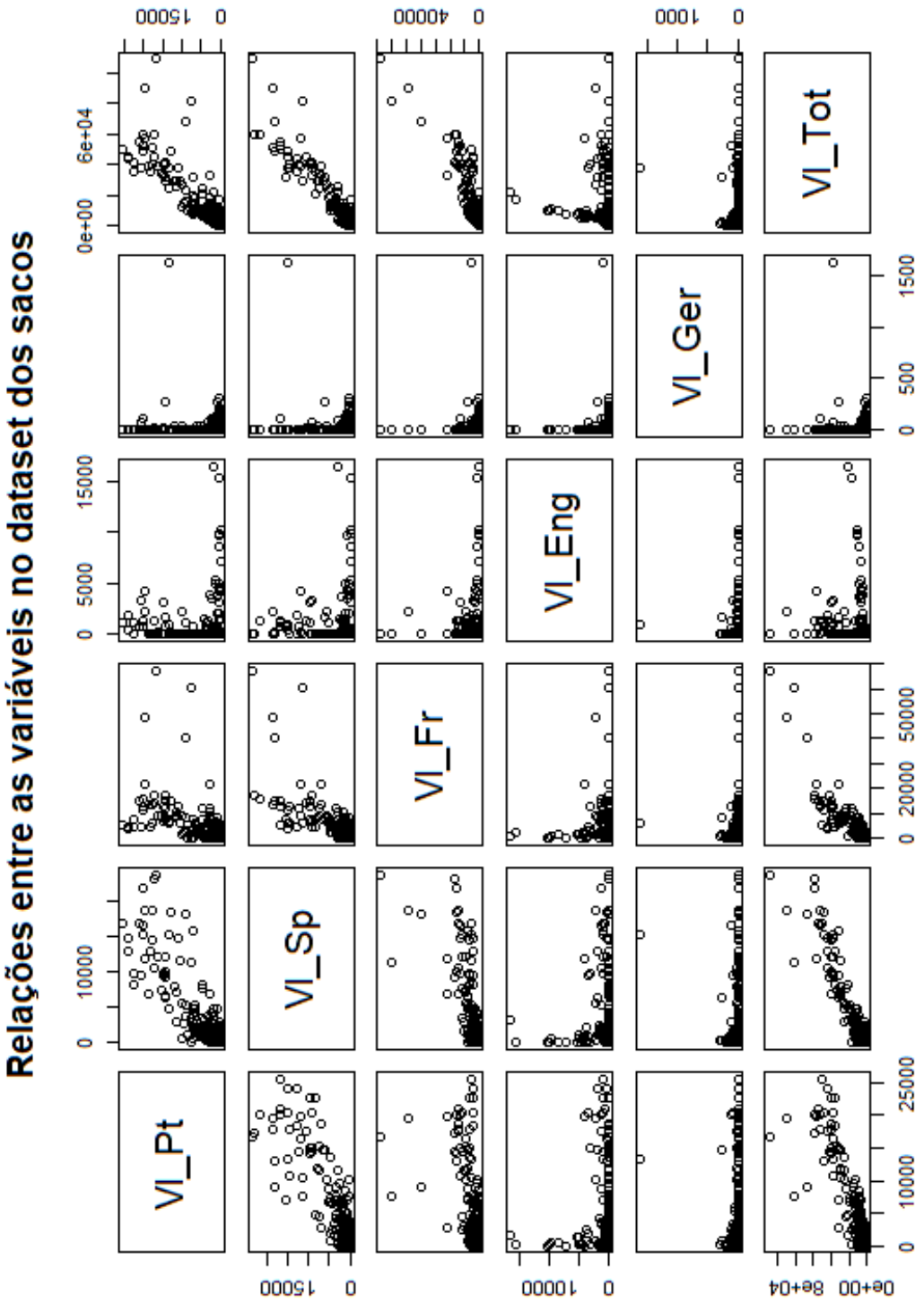

Figura A.2.9: $Scatterplot$ do *dataset* de sacos (variáveis do valor de vendas)

Anexo B

Código desenvolvido no R

O código desenvolvido para este trabalho encontra-se disponível e pode ser consultado no link:
`https://github.com/nelsonkosta/Projeto`