DETEÇÃO DE INDÍCIOS DE FRAUDE

NA INDÚSTRIA DO RETALHO

por

Ricardo Jorge Alves de Oliveira

Tese de Mestrado em Modelação, Análise de Dados e Sistemas de Apoio à Decisão pela Faculdade de Economia do Porto

Orientada por: Professor Dr. João Manuel Portela da Gama

Co-orientada por: Professora Dra. Rita Paula Almeida Ribeiro

2015

Nota Biográfica

Ricardo Oliveira nasceu na cidade do Porto em 1978. Concluiu a licenciatura em Informática de Gestão em 2003 na Universidade do Minho, pólo de Guimarães.

Após dois breves trabalhos como *webdesigner*, em 2004 surge a 1ª experiência profissional na Accenture (Lisboa) durante um ano, nas funções de analista / programador.
Participa em projetos de desenvolvimento/manutenção de aplicações informáticas para a Direção Geral de Contribuições e Impostos: SEF (Sistema de Execuções Fiscais), SLC (Sistema Local de Cobranças) e Reclamações Graciosas.

Em 2005 inicia funções numa empresa de retalho, como analista de dados. Desde então as suas atividades principais têm sido: análise de informação de negócio, análise da eficiência e eficácia dos processos, deteção de casos anómalos ou estranhos, possíveis fraudes ou incumprimento de procedimentos.

Em 2009 obteve a certificação de "Certified Information Systems Auditor" pelo instituto internacional ISACA.

Agradecimentos

Tendo chegado ao final deste trabalho, cabe-me agradecer a todas as pessoas que foram importantes ao longo deste caminho:

À minha esposa Célia, pela paciência e compreensão ao longo destes dois anos do curso.

Aos meus familiares, amigos e colegas de curso, pelo apoio e conselhos que me foram transmitindo ao longo do tempo, levando-me a tomar as melhores decisões possíveis.

Às chefias do meu departamento, pelo apoio e incentivo à inscrição neste mestrado, e também por terem tornado esta dissertação possível.

Por fim e não menos importante, o agradecimento aos professores João Gama e Rita Ribeiro, pelo interesse demonstrado no tema escolhido, pela experiência, sugestões, visão e melhorias ao longo do trabalho, tornando possível o seu resultado final.

Este trabalho foi apoiado pela Comissão Europeia, no âmbito do projeto MAESTRA (Grant number ICT-2013-612944) a quem também envio o meu agradecimento.
Thanks to the support of project MAESTRA (Grant number ICT-2013-612944) funded by European Commission.

Resumo

A fraude é omnipresente a todas as organizações, sendo que nenhuma está imune à sua ameaça. O seu custo é equivalente a um *iceberg* financeiro, onde as perdas financeiras são visíveis, mas existem outros custos indiretos, como perdas de produtividade, danos de imagem ou reputacionais.

Estima-se que todas as organizações possam perder 5% das suas receitas em cada ano, devido à fraude. A taxa de crime económico reportada mundialmente tem vindo a aumentar desde 2009, sendo que um terço das organizações reportou ter sido vítima de crime económico. Recentemente, metade das organizações da indústria do Retalho reportou ter sido vítima de crime económico.

A área de "*Data Analytics*" onde se inclui o *Data Mining*, foi considerada como o método mais eficaz para a deteção da fraude, tendo sido destacada por 25% das organizações que já sofreram crimes económicos.

Esta dissertação propôs-se desenvolver um modelo de aprendizagem a partir de dados de devoluções de artigos, num universo de 40 lojas de um retalhista, na procura de deteção de casos anómalos ou estranhos (designados por *outliers*), que possam ser fraude ou não.

Esse modelo é criado através da aplicação de técnicas de *Data Mining*, sendo que a deteção dos *outliers* é obtida através de representação gráfica (*box-plot*), através do estudo das variáveis do conjunto de dados.

A análise das variáveis é realizada de forma individual (univariada) e em combinação (bivariada e multivariada) entre as mesmas, produzindo resultados relevantes, mas distintos entre si.

É destacada a importância das análises multivariadas, através do desenvolvimento de árvores de regressão, pois permitem identificar *outliers* de contexto. Estes casos são considerados de difícil deteção, sendo apenas observados num determinado contexto, definido por um conjunto de regras associadas às variáveis.

Palavras-chave: Fraude; *Data Mining*; Análises Univariada/Bivariada/Multivariada; Árvores de Regressão; *Outlier*s; Gráficos *Box-Plot*; Devoluções de Artigos.

Abstract

Fraud is omnipresent to all organizations, and no one is immune to its threat. The cost of fraud is equivalent to a financial iceberg, where the financial losses are visible, but there are other indirect costs such as lost productivity, image or reputational damage.

It is estimated that every organization can lose 5% of their revenue each year due to fraud. The economic crime rate reported worldwide has increased since 2009, with one third of organizations reported have been victims of economic crime. Recently, half of the retail industry organizations reported having been victims of economic crime.

The area of "Data Analytics" which includes Data Mining was considered the most effective method for fraud detection, and was highlighted by 25% of organizations that have suffered economic crimes.

This dissertation proposed to develop a learning model from returned items (to stores), in a total of 40 stores from a retailer, trying to find and detecting abnormal or unusual cases (called outliers), which can be fraud or not.

This model is created through the application of data mining techniques, and the detection of outliers is obtained through graphical representation (box-plot), by studying the data set variables.

This analysis of the variables is carried out individually (univariate) and in combination (bivariate and multivariate) between them, to produce relevant results, but distinct from each other.

It is highlighted the importance of multivariate analysis, through the development of regression trees, that can lead to identifying context outliers. These cases are considered of difficult detection, being observed only in a given context, defined by a set of rules associated with the variables.

Keywords: Fraud; Data Mining; Univariate/Bivariate/Multivariate Analysis; Regression Trees; Outliers; Box-Plot graph; Item returns.

Índice Geral

Índice de Figuras

Índice de Tabelas

Índice de Abreviaturas

ACFE – Association of Certified Fraud Examiners

AIQ - Amplitude Interquartílica

CFE – Certified Fraud Examiners

IIA – The Institute of Internal Auditors

IPAI - Instituto Português de Auditoria Interna

POS – Point of Sales

PWC - PricewaterhouseCoopers

Capítulo 1: Introdução

Neste capítulo é realizado um enquadramento ao tema da deteção de indícios de fraude, a motivação, o interesse pessoal e profissional pela sua escolha. É feita uma descrição dos objetivos da dissertação em termos do que se pretende realizar e das metas a alcançar, descrevendo as técnicas aplicadas, quer por referência e sugestão de autores, quer também por aplicar outros métodos e análises diferentes e de certa forma inovadoras. Por fim é referido como o trabalho pretende ser útil e contribuir para a deteção de anomalias ou casos estranhos., podendo ser fraude ou não.

1.1 Motivação

Este trabalho de dissertação enquadra-se no âmbito do mestrado de Modelação, Análise de Dados e Sistemas de Apoio à Decisão onde, durante o 1º ano letivo, se aprofundou e desenvolveu o estudo de práticas de *Data Mining* para extração de conhecimento a partir de dados. Desse modo, a escolha do tema para dissertação visou a colocação em prática dos conhecimentos adquiridos durante o curso de mestrado, tendo essa escolha recaído sobre como o *Data Mining* pode ser importante na deteção de anomalias ou casos estranhos, constituindo-se como indícios de possível fraude nas organizações.

O meu interesse pessoal e profissional pela deteção de indícios de fraude, advém da minha experiência e interesse profissional. Desde 2005 tenho desempenhado funções de analista de dados numa empresa de retalho, com o foco na identificação de padrões e criação de alarmísticas em cenários dinâmicos e que apresentam riscos elevados de falhas de controlo interno, incumprimento de procedimentos ou deteção de situações que possam indiciar fraude.

O tema da fraude sempre me despertou interesse e curiosidade, pelas suas diferentes perspetivas de análise, quer pelo perfil e motivação do autor da fraude, bem como pelo "modus operandi" levado a cabo, não tanto pelo interesse sociológico, mas mais pela utilidade dessa informação para a prevenção, deteção e investigação da fraude.

1.2 Objetivos

De forma tradicional, a deteção de indícios de fraude no retalho tem sido efetuada (na componente informática) através de pesquisas ad-hoc em base de dados ou através da construção de relatórios que despoletem alarmísticas de situações "estranhas". Trata-se de técnicas que incidem muito sobre aspetos quantitativos e estatísticos dos dados, mas que carecem de uma aprendizagem mais eficiente e eficaz sobre a extração de conhecimento a partir dos dados, para além do seu processo ser de criação manual.

O objetivo desta dissertação passa por aplicar outro tipo de técnicas para obtenção de padrões de fraude, mais concretamente técnicas de *Data Mining* e *Machine Learning* que explorem e "aprendam" sobre os dados. Estas técnicas permitem a identificação de diferentes tipos de *outliers*, designados casos raros ou estranhos.

Através de análise estatística e gráfica nas aplicações SPSS e R, pode-se obter uma representação dos casos com uma maior probabilidade de serem anómalos ou estranhos, que podem ser fraude ou não.

O conjunto de dados alvo de estudo nesta dissertação foi obtido a partir de transações de devoluções de artigos, num universo de 40 lojas de um retalhista. Os dados correspondem a três meses de transações, que ocorreram entre os meses de Dezembro-2014 e Fevereiro-2015, com exceção dos dias 24-Dezembro e 01-Janeiro, em que as lojas não estiveram abertas ao público.

Por se tratar de dados reais, informação como o número da loja e do supervisor responsável pelo registo da devolução, foram mascarados como forma de garantir a confidencialidade dos mesmos, não sendo neste trabalho referidos nomes de entidades ou pessoas.

Os resultados que serão obtidos através da implementação de métodos de deteção de *outliers* são considerados apenas indícios de possível fraude, carecendo de uma investigação específica posterior para confirmação ou não dessa ocorrência. Nestes casos deve sempre imperar o bom senso e uma restrição ao analista e investigador de apenas resumir-se aos factos e não a julgamentos de caráter subjetivo.

1.3 Contribuições

Este trabalho visa desenvolver um modelo de aprendizagem semi-automático (não supervisionado) sobre os dados em estudo e que permita a um analista identificar casos anómalos, através da procura de *outliers* que possam revelar indícios de possível fraude.
Para além da realização de análises univariada e bivariada, o trabalho pretende demonstrar como as análises multivariadas são muito importantes na obtenção de um conhecimento mais profundo das variáveis, através da identificação de *outliers* de contexto, que de outra forma não seriam detetáveis.
De uma forma global, pode-se dizer que o contributo principal passa pela demonstração a auditores, investigadores ou analistas de fraude, sobre a importância do uso das ferramentas e tecnologias de *Data Mining* no seu trabalho de investigação.

1.4 Organização

No capítulo 2 é desenvolvido o tema proposto, começando por uma descrição do conceito de fraude, a estrutura da fraude em termos de categorias e subcategorias, a explicação da teoria do triângulo da fraude e a apresentação do perfil mais comum dos infratores que cometem fraude, através dos casos conhecidos e investigados por especialistas, com a ilustração de estatísticas sobre esse mesmo perfil. Ainda neste capítulo é apresentado como diversos autores consideram as técnicas e metodologias de *Data Mining* como importantes na deteção de anomalias ou casos estranhos. No capítulo 3 são desenvolvidas análises sobre um conjunto de dados de devoluções de artigos, com análises estatísticas sobre as variáveis qualitativas e quantitativas. É efetuada uma análise univariada para deteção de *outliers* do tipo global através de gráficos "*box-plot*". De forma complementar são realizadas análises bivariadas (duas variáveis) pela combinação das várias variáveis entre si. Pretende-se aqui identificar novos *outliers*, que não tivessem sido observados ou destacados na análise univariada. Por fim e de forma mais aprofundada são realizadas análises multivariadas, onde são construídas árvores de regressão como forma de estudar as duas variáveis quantitativas, através da combinação das restantes variáveis qualitativas em simultâneo. Em cada nó/folha da árvore são criados gráficos "*box-plot*" para observar a presença de novos *outliers* (de contexto), ainda não identificados noutras análises.

Capítulo 2: A Fraude e o uso do *Data Mining*

2.1 A Fraude - introdução

Nas normas internacionais para a prática profissional de Auditoria Interna do *The Institute of Internal Auditors* (IIA) encontra-se uma definição do conceito de fraude (2012:19):

"Quaisquer atos ilegais caracterizados por desonestidade, dissimulação ou quebra de confiança. Estes atos não implicam o uso de ameaça de violência ou de força física. As fraudes são praticadas por partes e organizações a fim de se obter dinheiro, propriedade ou serviços; para evitar pagamento ou perda de serviços; ou para garantir vantagem pessoal ou em negócios."

No relatório publicado pela *Association of Certified Examiners* (ACFE) em 2014 "*Report to the Nations on Occupational Fraud and Abuse*" a fraude é descrita como algo omnipresente sendo que nenhuma entidade está imune à sua ameaça, embora ainda haja muitas organizações com a convicção de pensamento de que a fraude só acontece aos outros. Por isso e nesse sentido, a ACFE tem procurado combater este paradigma, considerando relevante tornar público o custo da fraude, a sua natureza universal, bem como as suas tendências (ACFE, 2014, p. 6).

Esta associação foi fundada em 1988 no Texas - Estados Unidos da América, e é a maior organização mundial antifraude e a principal fornecedora de formação e educação antifraude, estando a reduzir a mesma nas empresas a nível mundial e a ganhar a confiança das pessoas na integridade e objetividade da profissão (Wells, J. T., 2009, p. 15). Segundo informação do seu sítio institucional (ACFE, 2014), atualmente a associação beneficia da colaboração de cerca de 75 mil membros. O seu presidente honorário e fundador - Joseph T. Wells, é criminologista e antigo agente do FBI, investiga e dá conferências a grupos empresariais e profissionais acerca de temas relacionados com fraudes, já obteve prémios máximos de literatura, e durante a última década, já foi nomeado para a lista das 100 pessoas mais influentes na área da contabilidade da *Accounting Today* (Wells, J. T., 2009, p. 15).

2.2 A Fraude Ocupacional

A fraude pode ser classificada como organizacional (funcionários de uma organização que cometem fraude no interesse da organização) ou ocupacional (funcionários de uma organização que cometem fraude contra ela própria).

Desde o ano de 1996 que a ACFE publica, de dois em dois anos, um estudo sobre a compreensão e conhecimento da ocorrência da fraude ocupacional e os seus impactos financeiros nas organizações por todo o mundo. A edição de 2014 é baseada em 1483 casos de fraudes ocupacionais, reportadas à ACFE por examinadores de fraude certificados (membros da própria associação) que as investigaram (ACFE, 2014, p. 2) em mais de 100 países. A análise destes casos fornece uma visão fundamental sobre como a fraude é cometida, como pode ser detetada e como as organizações podem reduzir as suas vulnerabilidades ao risco da sua ocorrência. Segundo este relatório, a fraude ocupacional é descrita como o abuso ou má aplicação dos recursos ou ativos de uma organização por parte de um indivíduo para proveito próprio (ACFE, 2014, p. 6).

Infelizmente, a própria natureza das fraudes leva a que os seus custos e a sua existência não sejam revelados, mesmo que nalguns casos até sejam detetadas e investigadas. O custo da fraude é equivalente a um *iceberg* financeiro onde as perdas financeiras são claramente visíveis, mas existem outros custos indiretos, como perdas de produtividade, danos de imagem ou reputacionais e custos de investigação (ACFE, 2014, p. 8).

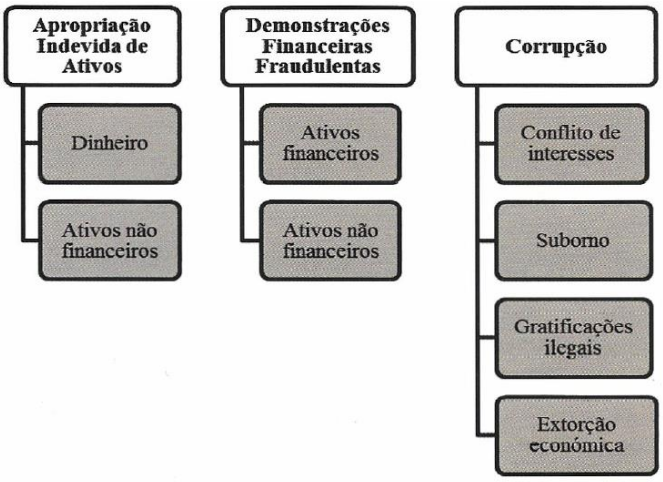
Este relatório refere que se estima que todas as organizações possam perder 5% das suas receitas em cada ano, devido à fraude (ACFE, 2014, p. 8). Esta percentagem aplicada ao produto bruto mundial de 2013 pode corresponder a perdas projetadas de 3,7 triliões de dólares (aproximadamente 3 triliões de euros). Este valor deve ser usado como *benchmark*, mas não deve ser interpretado como o custo real da fraude.

Segundo o mesmo relatório (ACFE, 2014, p. 11), a fraude ocupacional pode ser estruturada em 3 categorias principais: Apropriação Indevida de Ativos, Demonstrações Financeiras Fraudulentas e Corrupção, em que cada uma pode ainda conter várias subcategorias, conforme a figura 2.1 e 2.2.

O âmbito desta dissertação irá incidir o seu foco na deteção de indícios de fraude sobre as subcategorias “Desembolsos fraudulentos” e “Furto”, conforme ilustrado na figura 2.2, dado estarem diretamente relacionadas com o processo de Devolução de artigos nas

lojas. As estatísticas apresentadas mais á frente serão mais focadas quer neste tipo da fraude, quer em termos do tipo de indústria (neste caso, retalho) e uma visão sobre os dados conhecidos nos países europeus (mais próximos da realidade portuguesa).

Figura 2.1 - Árvore da Fraude (versão simplificada)



Fonte: IPAI, 2014, p. 15

Figura 2.2 – As 5 sub-categorias de Fraude por "Apropriação Indevida de Ativos"

Nota: no Anexo I, e em particular na figura A1.1, pode ser consultada a árvore da fraude, na versão original e detalhada, publicada pela ACFE em 2014.

Numa análise à frequência dos esquemas mais comuns de fraude sobre a indústria do retalho ("Retail"), verifica-se que são os Ativos não financeiros ("Non-Cash") os mais frequentes com 33,8% e o Furto de dinheiro ("Cash on Hand") com 22,1%:

Figura 2.3 – Esquemas mais frequentes de Fraude, na indústria do retalho

Industry/ Scheme	Banking and Financial Services	Government and Public Administration	Manufacturing	Health Care	Education	Retail	Insurance	Oil and Gas	Transportation and Warehousing	Services (Other)	Construction	Religious, Charitable or Social Services
Cases	244	141	116	100	80	77	62	49	48	45	43	40
Billing	5.7%	19.1%	22.4%	29.0%	33.8%	10.4%	17.7%	24.5%	33.3%	28.9%	34.9%	32.5%
Cash Larceny	13.1%	10.6%	6.0%	12.0%	6.3%	15.6%	6.5%	2.0%	2.1%	11.1%	14.0%	7.5%
Cash on Hand	18.9%	12.1%	7.8%	16.0%	16.3%	22.1%	1.6%	2.0%	10.4%	11.1%	7.0%	12.5%
Check Tampering	5.7%	5.7%	7.8%	21.0%	10.0%	7.8%	4.8%	4.1%	20.8%	17.8%	27.9%	35.0%
Corruption	37.3%	36.2%	54.3%	37.0%	36.3%	22.1%	33.9%	57.1%	29.2%	35.6%	46.5%	30.0%
Expense Reimbursements	4.1%	12.8%	7.8%	23.0%	31.3%	3.9%	4.8%	14.3%	14.6%	17.8%	27.9%	32.5%
Financial Statement Fraud	10.2%	5.0%	13.8%	8.0%	10.0%	6.5%	3.2%	12.2%	10.4%	6.7%	11.6%	7.5%
Non-Cash	13.1%	17.7%	34.5%	12.0%	12.5%	33.8%	12.9%	16.3%	33.3%	17.8%	20.9%	15.0%
Payroll	5.3%	15.6%	8.6%	15.0%	16.3%	5.2%	8.1%	6.1%	16.7%	6.7%	18.6%	20.0%
Register Disbursements	2.5%	0.7%	2.6%	3.0%	5.0%	13.0%	0.0%	0.0%	4.2%	6.7%	2.3%	2.5%
Skimming	5.7%	11.3%	4.3%	18.0%	20.0%	18.2%	22.6%	2.0%	6.3%	33.3%	7.0%	12.5%

Fonte: ACFE, 2014, p. 29

2.3 Perfil do(s) autor(es) da Fraude

Para melhor se compreender o perfil do autor de uma fraude, é importante considerar os três *drivers* da fraude, conforme é descrito por um estudo realizado pela empresa KPMG (KPMG, 2013, p. 5), sendo eles: oportunidade, pressão (ou motivação) e justificação (ou racionalização).

Figura 2.4 – O triângulo da Fraude

Fonte: Wells, J. T., 2009, p. 24

Este estudo refere que "as pessoas cometem fraude quando os três elementos acontecem em simultâneo". Estes três *drivers* são parte de uma metodologia que foi desenvolvida por investigadores de fraude na década de 50.

A KPMG descreve o seguinte cenário: "O potencial infrator vê uma porta aberta como uma oportunidade. O motivo e o racional fazem-no mover-se em direção à porta e a capacidade fá-lo atravessá-la" (KPMG, 2013, p. 5).

De acordo com a teoria do triângulo da fraude, aqueles que cometem fraude ocupacional tendem a sentir necessidades financeiras, oportunidade e racionalização, sendo por esse motivo que a deteção é um dos principais fatores de sucesso na prevenção da fraude por eliminar a perceção de oportunidade por parte do fraudador (ACFE, 2014, p. 18).

Oportunidade

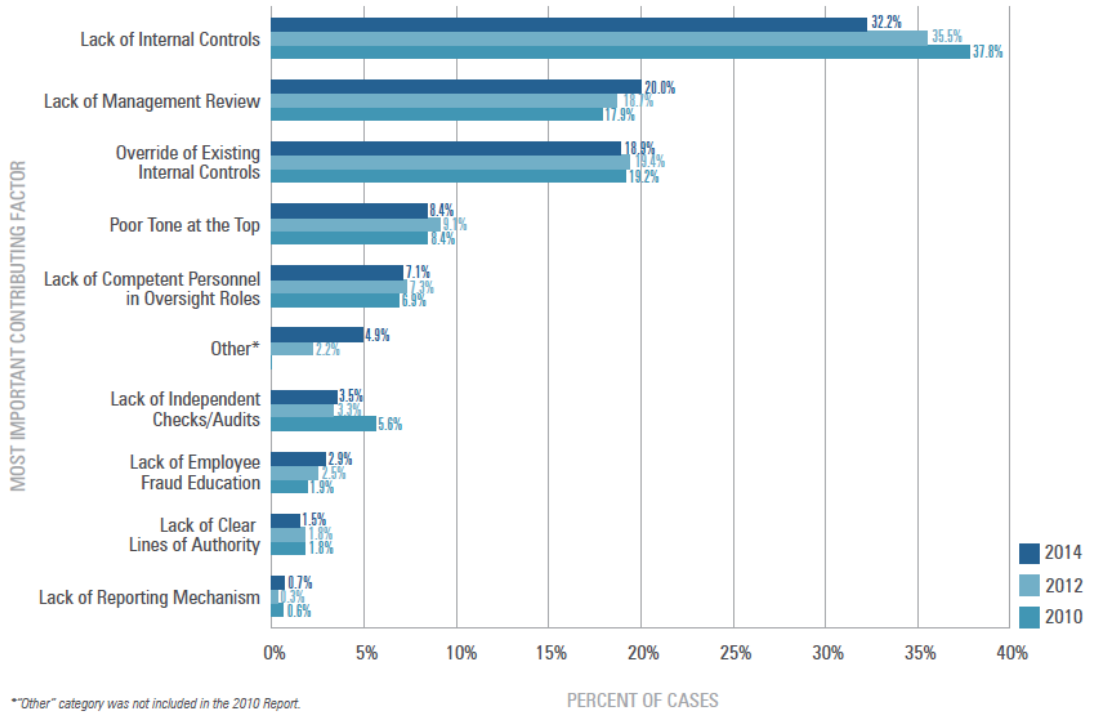
A KPMG refere que "As pessoas não cometem fraude sem surgir a oportunidade para a cometer" (KPMG, 2013, p. 6). No seu estudo de 2013 verificou-se que na maioria dos casos de fraude observados, o infrator já trabalhava na organização há mais de 6 anos, o que significa que um infrator não entra para uma organização com o objetivo *a priori* de

cometer fraude. Porém mudanças em circunstâncias pessoais, como a pressão para atingir objetivos considerados "irrealistas", podem criar condições para a ocorrência da fraude. A fraude pode ser cometida por pessoas que já se sentem confortáveis nas suas funções e também em relação à própria organização (KPMG, 2013, p. 6)

Para a KPMG "não existe um perfil de personalidade que cometa fraude, mas todos os tipos de pessoas podem cometer fraude se a oportunidade surgir" (KPMG, 2013, p. 14)

A figura 2.5 mostra como essas oportunidades podem ser aproveitadas pelas pessoas, sendo que no estudo realizado pela ACFE, verifica-se que em 1/3 dos casos observados pelos CFEs foram observadas falhas ou fraquezas ao nível dos controlos internos para prevenir a ocorrência da fraude. Este facto reforça a importância da existência de controlos anti-fraude, sendo que, no caso particular de empresas de pequenas dimensões ainda assume uma maior importância (ACFE, 2014, p. 39):

Figura 2.5 – Principal falha de controlo interno observada nas organizações



Fonte: ACFE, 2014, p. 39

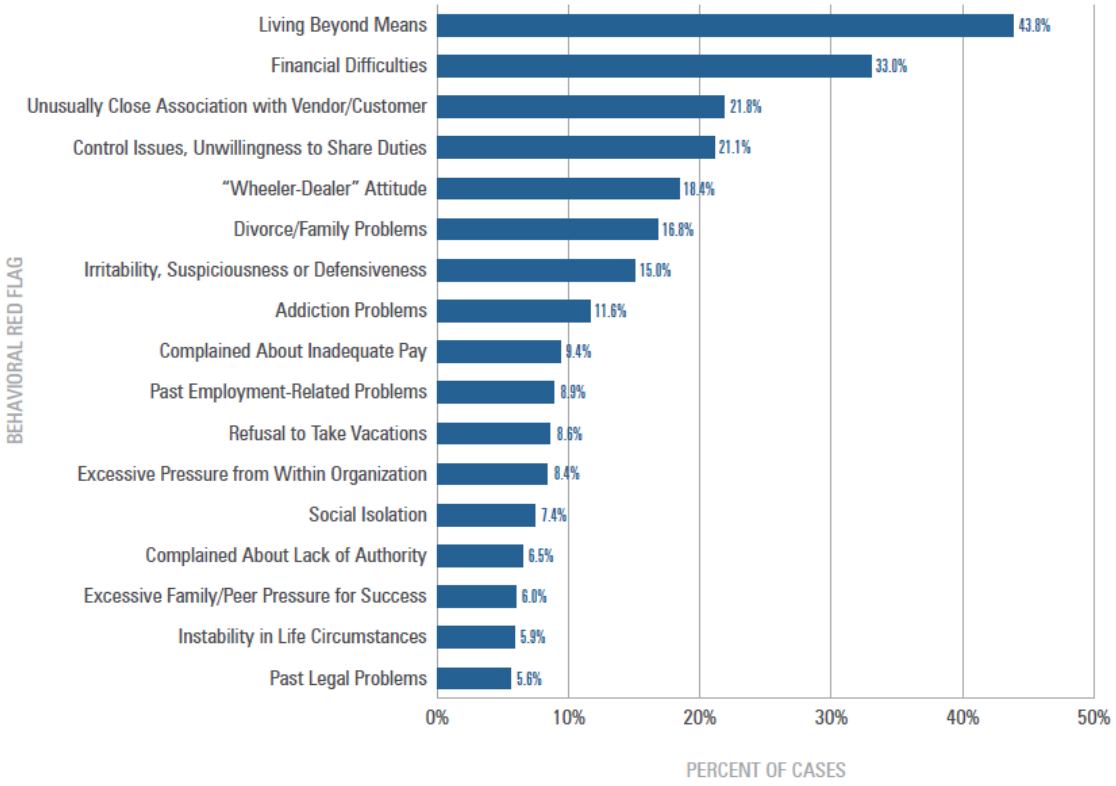
Porém, mesmo existindo um forte sistema de controlos internos nas organizações, isso não previne a ocorrência de todas as fraudes (KPMG, 2013, p. 7).
Nalguns casos, o infrator pode ser alguém dentro da organização que perceba como estão implementados os controlos internos e saber como os "ultrapassar" ou, até ser conhecedor de uma falha de controlo, e explorar essa falha.

A KPMG considera a "capacidade" como uma componente do *driver* da oportunidade, sendo que consiste nos atributos (personalidade e competência) do infrator que lhe permitem explorar a oportunidade, quando esta surge (KPMG, 2013, p. 7).
Por exemplo, quanto mais alta a posição hierárquica do infrator na organização, maior a sua possibilidade de "ultrapassar" certos controlos implementados (KPMG, 2013, p. 7).
Os críticos do triângulo da fraude argumentam que este, por si só, não pode explicar a fraude porque duas das suas características como a Motivação e a Justificação, não são observáveis. Por isso em 2004 foi apresentado um modelo chamado "Diamante da Fraude", que segundo os seus autores, casos conhecidos de fraudes multimilionárias em demonstrações financeiras, não teriam ocorrido sem as pessoas terem a capacidade adequada para executar os detalhes da fraude (IPAI, 2014, p. 17).

Pressão/Motivação

"A fraude como qualquer outro crime, requere um motivo, sendo a razão mais frequente o financeiro" (KPMG, 2013, p. 8).
No inquérito realizado pela ACFE, foram analisados indicadores sobre o comportamento exibido pelos infratores antes das suas fraudes serem detetadas. Pelo menos um alerta vermelho foi identificado em 92% dos casos, sendo que em 64% dos casos o infrator exibiu sinais estranhos em dois ou mais alertas vermelhos (ACFE, 2014, p. 59).
A figura 2.6 mostra como sinais de uma vida para além das posses, ou dificuldades financeiras, proximidade invulgar com fornecedores ou clientes, ou pessoas pouco dadas a delegar tarefas, se tornam alertas importantes.

Figura 2.6 – Comportamentos "estranhos" demonstrados pelos infratores



Fonte: ACFE, 2014, p. 59

Justificação/Racional

Os infratores encontram muitas vezes motivos racionais para as suas práticas, podendo estas estar relacionadas com fúria, medo, sentimentos de injustiça (ex: salário abaixo do expectável) (KPMG, 2013, p. 8).

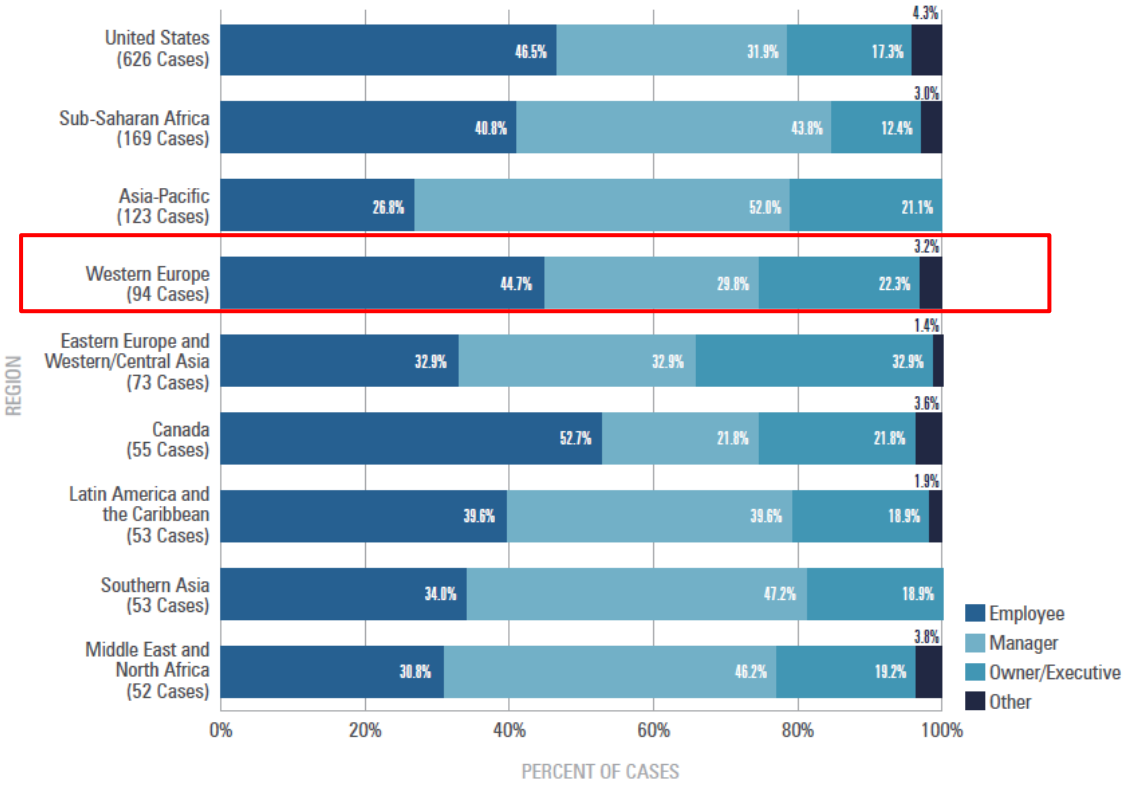
Porém, tem sido observado por investigadores da KPMG que o motivo para a fraude é muito determinado pelo contexto ético e cultural, variando bastante de país para país. (KPMG, 2013, p. 9). As regulamentações governamentais, bem como políticas para reforçar os padrões éticos são por isso importantes.

Verifica-se então que o valor da informação sobre o perfil do autor de uma fraude, permite analisar, identificar e quantificar onde reside o maior risco de fraude dentro de uma organização: qual a sua função/departamento, a sua idade, antiguidade na organização, o seu passado criminal e comportamentos "estranhos" detetados (ACFE, 2014, p. 40).

Estes dados têm sido recolhidos pela ACFE ao longo dos anos, e permitem analisar de uma forma consistente, a tendência e as evoluções dos padrões comportamentais da pessoa que comete a fraude.

Começando pela análise ao cargo que o indivíduo fraudulento ocupa na organização, verifica-se pela figura 2.7, que na Europa Ocidental, cerca de 45% dos casos observados de fraude foram realizados por colaboradores da organização num nível mais baixo em termos hierárquicos, de 30% por parte de chefias ou gestores, e de 22% pelos próprios donos ou administradores:

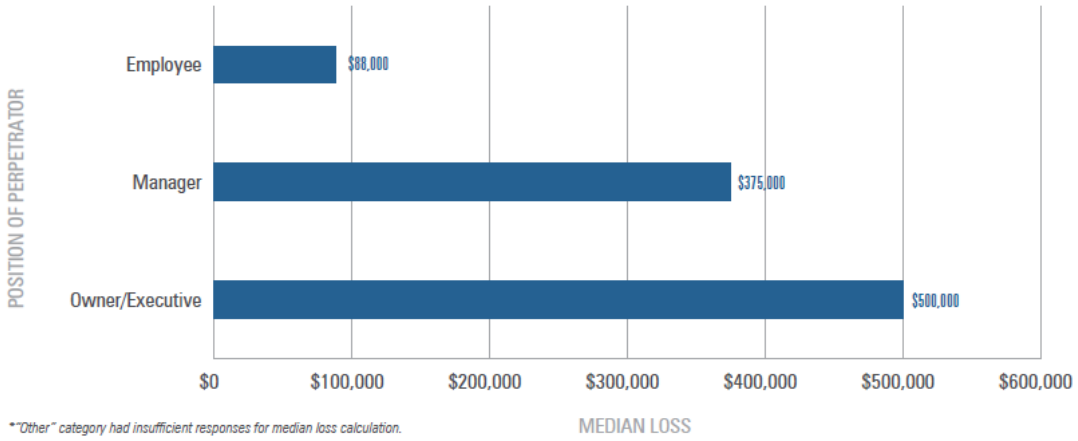
Figura 2.7 – Posição hierárquica do indivíduo que comete a fraude



Fonte: ACFE, 2014, p. 42

Se a probabilidade de ocorrência de fraude parece estar mais relacionada com os níveis hierárquicos mais inferiores de uma organização, porém, o seu impacto em termos de custo de perda é inversamente proporcional, dado que uma fraude cometida por alguém do topo de uma organização, pode implicar perdas muito mais significativas, conforme se pode comprovar pela figura 2.8:

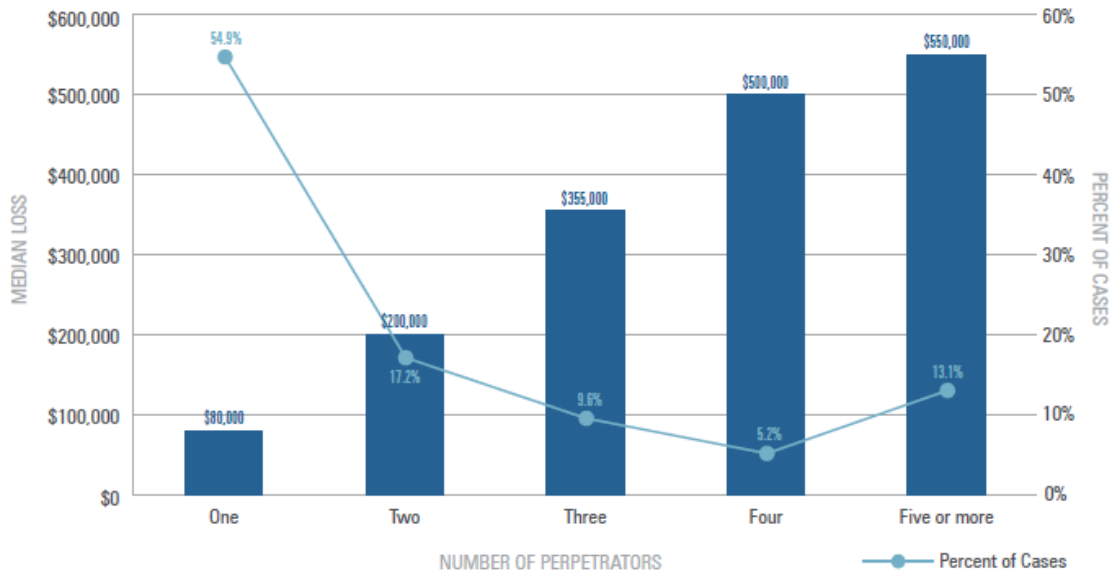
Figura 2.8 – Perda média estimada baseada na posição hierárquica do indivíduo



Fonte: ACFE, 2014, p. 44

Numa outra análise, verifica-se que mais de metade das fraudes estudadas pela ACFE e publicadas no seu último relatório são cometidas por apenas um indivíduo. No entanto, quando as fraudes são cometidas por dois ou mais infratores, as perdas estimadas em termos médios sobem consideravelmente, conforme se pode observar na figura 2.9:

Figura 2.9 – Número de infratores – frequência e perda média estimada



Fonte: ACFE, 2014, p. 46

Este facto está relacionado com a circunstância que advém da possibilidade de quando pelo menos duas pessoas estão envolvidas num conluio entre si que envolve um esquema de fraude, possuírem uma vantagem sobre "contornar" controlos internos e verificações independentes, o que permite a realização de transações que envolvem quantias de dinheiro mais elevadas (ACFE, 2014, p. 46).

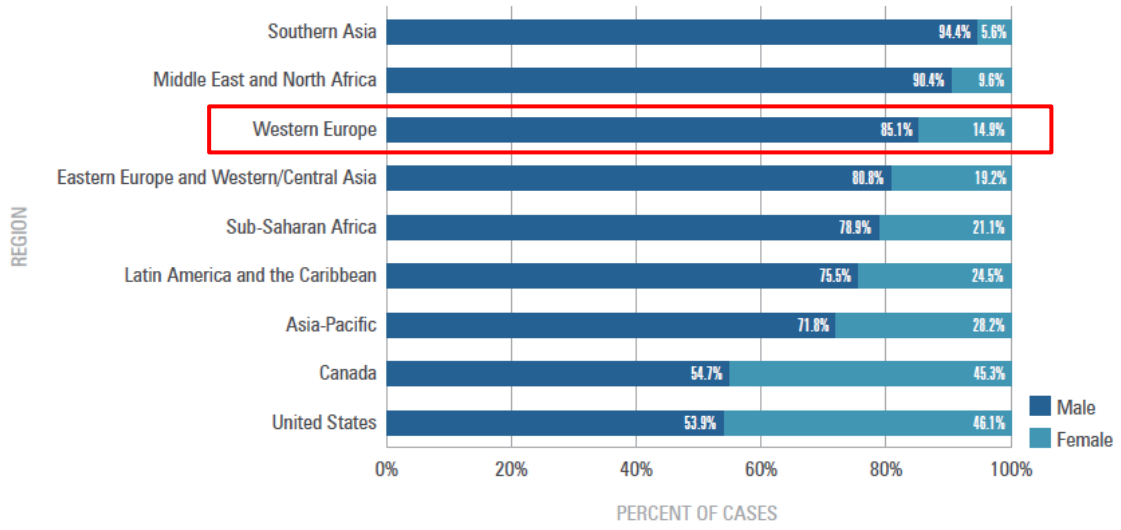
A KPMG refere no seu estudo que muitos infratores preferem "atuar" sozinhos, porque dessa forma, não têm de confiar noutras pessoas para manter silêncio, nem dividir "lucros", apesar de muitas fraudes requererem colaboração (KPMG, 2013, p. 14).

A fraude é muitas vezes complexa para poder ser executada por apenas uma pessoa. É necessário muitas vezes que mais alguém possa "fechar os olhos", ou fornecer uma *password*, ou até falsificar documentos (KPMG, 2013, p. 14).

O conluio pode ocorrer dentro e fora das organizações, sendo que a fraude envolvendo terceiras entidades são difíceis de detetar. Uma forma comum de conluio é ao nível de compras de bens, como por exemplo, faturas "inflacionadas" (KPMG, 2013, p. 15).

A percentagem de infratores masculinos e femininos varia substancialmente conforme a região/continente onde ocorre a fraude. Na Europa Ocidental, 85% das fraudes observadas pela ACFE foram cometidas por indivíduos do sexo masculino:

Figura 2.10 – Género do infrator baseado na região



Fonte: ACFE, 2014, p. 49

A consultora KPMG publicou em 2013 um relatório relativo a uma análise de 596 infratores de empresas investigadas entre 2011 e 2013, em que procura responder à pergunta sobre "Quem é o infrator típico nas organizações?" (KPMG, 2013, p.2).
Com base no perfil dos 596 casos estudados naqueles anos, as conclusões foram as de que o infrator típico tem entre 36 e 45 anos de idade, atua contra a sua própria organização e exerce funções executivas, financeiras, operações, *marketing* ou vendas, e encontra-se em funções na organização há mais de 6 anos (KPMG, 2013, p. 2).

É importante referir que grandes fraudes já levaram à queda de organizações inteiras, perdas significativas de investimentos e perda de confiança nos mercados de capitais. São exemplos os escândalos financeiros ocorridos nos EUA, como os casos da Enron, da WorldCom, Lehman Brothers, e em Portugal, dos casos do BPN e do BPP (IPAI, 2014, p. 14).
Mesmo que as fraudes não levem à queda das organizações, causam sempre impactos negativos, principalmente ao nível da reputação, imagem e perda de confiança por partes das pessoas nas respetivas organizações (IPAI, 2014, p. 14).

Um outro estudo sobre a fraude - "Global Economic Crime Survey 2014", publicado em 2014 pela PricewaterhouseCoopers foi realizado a mais de 5 mil representantes de organizações, de mais de 95 países de todo o mundo. Uma das suas conclusões é de que a **taxa de crime económico reportada tem vindo a aumentar**, de 30% em 2009 para **37% em 2014**, e que uma em cada três organizações reportou ter sido vítima de crime económico.
Em 2014, praticamente **metade** (49%) **das organizações da indústria do Retalho reportou ter sido vítima de crime económico**, sendo um dos tipos de indústria com maior taxa. Em dois terços destes casos, o autor da fraude foi interno à organização.

Portugal surge como um dos países do mundo com menos casos de crimes económicos reportados, com apenas 12% e ao nível de países como a Arábia Saudita e a Dinamarca. Segunda a PWC isto pode significar falta de controlos ou meios que ajudem na deteção da fraude.

Carlos Pimenta refere que "em Portugal existe um conjunto de fatores (institucionais, culturais, cognitivos e outros) que impedem uma quantificação rigorosa da fraude" (Pimenta C., 2009, p. 4).

É por isso importante serem investidas e desenvolvidas abordagens de deteção de fraude, tais como o investimento e aposta em analistas de dados e *Data Mining* que forneçam às organizações meios de identificar casos fraudulentos (KPMG, 2013, p. 3).
Os benefícios dessas análises de dados podem incluir uma identificação de conluios entre pessoas ou organizações, bem como de análise de transações suspeitas em menos tempo e de forma mais eficiente e com menores custos do que usando as técnicas de amostragem forenses mais tradicionais (KPMG, 2006, p. 16).

Segundo a Deloitte, os retalhistas irão provavelmente ter que intensificar o ritmo da inovação em termos da prevenção da fraude e atividades de deteção, no caso de estarem dispostos a recuperar a margem que está a ser perdida para a fraude. É tempo da indústria do retalho considerar como as novas tecnologias como a análise de dados ("*data analytics*") pode ajudar a detetar mais fraude, num ambiente económico cada vez mais competitivo (Deloitte, 2013, p. 1).
A **área de "D*ata Analytics*"** onde se inclui o *Data Mining*, **foi considerada como o método mais eficaz para a deteção de fraude**, tem sido destacada por 25% das organizações que já sofreram crimes económicos (PWC, 2014, p. 45).

2.4 Subcategorias de fraude na categoria "Apropriação Indevida de Ativos"

Dentro da categoria "Apropriação Indevida de Ativos" existem várias sub-categorias (esquemas de fraude), sendo apresentadas de seguida em termos da sua descrição e riscos associados:

Dinheiro

Furto (de Dinheiro)

Wells indica que a caixa registadora é geralmente o ponto mais comum de acesso ao dinheiro para os empregados de uma loja, tornando-se assim um foco onde esquemas de fraude podem ocorrer. De facto, imensas transações envolvem o manuseamento do dinheiro de caixa, por isso, o fraudador pode conseguir retirar dinheiro da caixa sem ser apanhado (Wells, J. T., 2009, p. 138). Um modo simples de um empregado evitar ser detetado é roubar pouco dinheiro de cada vez, uma vez que os montantes em falta são pequenos, logo as diferenças de caixa podem ser atribuídas a erros e não a furto (Wells, J. T., 2009, p. 140).

Desembolsos fraudulentos (de caixa)

Falsas devoluções é um dos esquemas simples relacionados com os desembolsos fraudulentos relacionados com uma caixa registadora (Wells, J. T., 2009, p. 205).
Um reembolso (devolução) é efetuado e registado numa caixa quando um cliente pretende devolver um artigo que comprou na loja, em que o cliente será reembolsado pelo preço de compra do respetivo artigo, e este será reposto na loja ou no armazém da loja, com o incremento de uma unidade no seu *stock*. No entanto, se a transação de devolução for fictícia, nenhuma mercadoria é de facto devolvida, ficando o *stock* da empresa e da loja sobreavaliado.

Sonegação

Dentro da subcategoria Dinheiro, encontramos um tipo de fraude descrito como "Sonegação", que Wells no livro do Manual da Fraude descreve como "a remoção de dinheiro antes do seu lançamento num sistema de contabilidade" (Wells, J. T., 2009, p. 94), sendo conhecidas por fraudes *off-book*, por não deixarem rasto, dado os fundos serem recebidos mas a empresa lesada não ter consciência de que o dinheiro foi recebido.

Wells refere ainda que as pessoas que lidam diretamente com os clientes, são as candidatas mais prováveis a este tipo de fraude (Wells, J. T., 2009, p. 94), sendo que o

esquema mais básico de sonegação acontece quando um operador de caixa atende um cliente, recebe o pagamento, mas não regista a venda.

Outro tipo de sonegação pode ocorrer quando o próprio operador de caixa sonega o valor do desconto a que o cliente teria direito para proveito próprio (Wells, J. T., 2009, p. 107). Porém, Wells refere que "os esquemas de sonegação são, por norma, mais fáceis de ocultar do que a maioria de outros tipos de fraude ocupacional" (Wells, J. T., 2009, p. 118). Wells conclui que os procedimentos de controlo interno são a chave para evitar a ocorrência destes esquemas de fraude (Wells, J. T., 2009, p. 125).

Ativos não financeiros

O furto é o tipo mais básico de roubo de inventário, sendo que na maioria dos casos são cometidos por empregados com acesso ao armazém (Wells, J. T., 2009, p. 314).
A maioria dos métodos de ocultação referem-se a ajustes ou alterações nos registos de *stock*, alterando o respetivo *stock* teórico ou procedendo a uma contagem errada do *stock* real, sendo que alguns fraudadores tentam dar a impressão de que existe mais *stock* em armazém do que na realidade existe (Wells, J. T., 2009, p. 330). Por exemplo, caixas vazias podem ser empilhadas nas prateleiras do armazém, para criar a ilusão de *stock* suplementar.

A nível mundial, existe um estudo sobre o efeito económico do furto e a disponibilidade de mercadoria, designado por "Barómetro Mundial do Furto no Retalho" sobre dados de mais de duzentos retalhistas mundiais, publicado pela Checkpoint Systems, Inc. em 2014. O estudo completo pode ser solicitado através do seguinte endereço:
http://www.globalretailtheftbarometer.com/index.html

De forma resumida pode-se referir que em 2013 a taxa de perda desconhecida (em % das vendas) em Portugal foi de 1,18% situando-se abaixo da taxa global que é de 1,29%. A taxa de Portugal corresponde a 614 milhões de dólares de perda desconhecida (Checkpoint Systems, Inc., 2014, p. 17).
Em Portugal a taxa de furto na loja é de 50%, sendo superior à média da Europa que é de 39% (Checkpoint Systems, Inc., 2014, p. 27).

2.5 O uso do *Data Mining* na deteção de indícios de fraude

2.5.1 Definição de *Data Mining* e suas vantagens

"Fazer mais e melhor com menos recursos" tem sido uma das frases mais ouvidas nos últimos anos, praticamente em todos os departamentos nas organizações, e também entre os investigadores da fraude. O atual ambiente de negócio nas organizações exige aumentos de produtividade em todas as áreas da organização (Codere, D., 2009, p. 41).

Vivemos atualmente na era da informação, onde as organizações estão sobrecarregadas de dados. Cada vez mais informação é armazenada nas bases de dados, em que transformar esses dados em conhecimento cria uma necessidade de procurar poderosas ferramentas de análise (Jans, M. *et al.*, 2007, p. 5).

À medida que o uso das tecnologias de informação é cada vez maior, também maior deverá ser o uso destas tecnologias na deteção da fraude, permitindo aos investigadores um foco maior nas áreas de maior risco (Codere, D., 2009, p. 41).

As técnicas de análise de dados usadas no passado incidiam muito sobre aspetos quantitativos e estatísticos em relação às características dos dados. Estas técnicas são úteis na interpretação dos dados e podem ajudar a obter uma visão mais aprofundada, e criar conhecimento a partir daí. Mas embora estas técnicas de análise de dados tradicionais consigam extrair conhecimento a partir dos dados, o seu processo é criado por analistas, sendo seu processo de criação manual (Jans, M. *et al.*, 2007, p. 5).

Para ultrapassar estas limitações, deve ser desenvolvido um sistema de análise de dados com recurso a conhecimentos de fundo, e que envolvam tarefas de raciocínio a partir dos dados fornecidos. Para atingir este objetivo, pesquisadores têm-se focado na área de *Machine Learning*, que é uma disciplina científica que explora a construção e o estudo de algoritmos que "aprendem" a partir de dados. Esses algoritmos operam a partir de modelos baseados nos *inputs* e usam-nos para criar previsões ou decisões.

Assim emergiu uma nova área frequentemente designada por *Data Mining* e descoberta de conhecimento. *Data Mining* pode ser definido como o processo de descoberta de padrões nos dados, podendo o processo ser automático ou semi-automático (mais comum). Os padrões identificados devem ter significado por forma a trazerem vantagens, principalmente económicas (Jans, M. *et al.*, 2007, p. 6).

Data Mining é assim uma forma de descobrir conhecimento nas vastas bases de dados que as organizações dispõem (Jans, M. *et al.*, 2007, p. 6).

A deteção de fraude é uma área importante para a aplicação prática das técnicas de *Data Mining*, tendo em conta as consequências económicas e sociais que estão normalmente associadas a essas atividades ilegais (Torgo, L., 2010, p. 165).

A empresa SAS num artigo publicado sobre o uso de técnicas de *Data Mining* para deteção de fraude, define alguns passos importantes neste processo (Kristin, R. N. *et al.*,, 1999, p. 1):

1. Definição do problema
2. Extração de dados
3. Estratégia de modelação
4. Análise de resultados

A definição do processo de negócio a analisar é o primeiro e o passo mais importante na utilização do D*ata Mining*, devendo-se definir aqui quais os objetivos da análise de forma clara e como os resultados vão ser medidos (Kristin, R. N. *et al.*,, 1999, p. 1).

Estando assim definido qual o processo de negócio a analisar para o plano de deteção de fraude, deve-se então proceder à identificação dos dados necessários a essa análise, isto é, quais as respetivas base de dados, campos pretendidos, devendo esta tarefa ser executada com prudência, pois as ferramentas de *Data Mining* só são úteis caso os dados estejam completos e não enviesados ou distorcidos (Kristin, R. N. *et al.*,, 1999, p. 2).

O *Data Mining* traz a capacidade de considerar e analisar milhares de transações em menos tempo, com mais eficiência e relação custo-eficácia do que usando técnicas mais tradicionais de amostragem forenses (KPMG, 2006, p. 16).

O fenómeno do *Big Data* trouxe um aumento vertiginoso do volume, variedade e velocidade (3Vs) de informações de negócio, que mudou a forma como as principais empresas gerem os seus desafios de conformidade legal e investigam comportamento aberracionais ou estranhos (EY, 2014, p. 1).

2.5.2 Aprendizagem Supervisionada vs Não Supervisonada

Em termos de estratégias de *Data Mining*, existem duas amplamente conhecidas: aprendizagem supervisionada e não supervisionada.

Os métodos supervisionados são implementados quando existe uma variável "alvo" através da qual se pretende efetuar previsões, a partir de outras variáveis de entrada. Numa investigação de fraude, exemplos de registos com casos de fraude e de não-fraude são necessários, ou seja, todos os registos disponíveis têm de ser classificados como "fraudulento" ou "não-fraudulento" (Jans, M. *et al.*, 2007, p. 7).

Depois de construído o modelo usando esses dados como o conjunto de treino, os novos casos analisados devem ser classificados conforme os rótulos anteriores. Claro que isto implica que haja uma confiança na correta classificação dos dados do conjunto de treino, sendo esta confiança a sustentabilidade do próprio modelo.

Este método é apenas aplicável na deteção de fraude de um determinado tipo, se já ocorreu e existem dados sobre a mesma ocorrência. Em contraste, os métodos não supervisionados não usam registos já classificados (Singh, K. *et al.*, 2012, p. 6).

Os métodos não supervisionados são aplicados nos casos em que não temos uma variável "alvo", apesar de existirem as variáveis de entrada (Kristin, R. N. *et al.*, 1999, p. 3).

Estes métodos buscam análises sobre contas, clientes, fornecedores, etc, que tenham comportamentos considerados "estranhos" pela obtenção de *outputs* suspeitos ou de anomalias gráficas.

Por fim, e para ambos os métodos, deve-se avaliar os resultados validando se os objetivos definidos foram alcançados, sendo que no caso de não estarem a ser alcançados, deve-se então efetuar uma reformulação ou ajustamento ao modelo de forma a garantir resultados mais efetivos, como por exemplo, reduzir ou eliminar os falsos positivos.

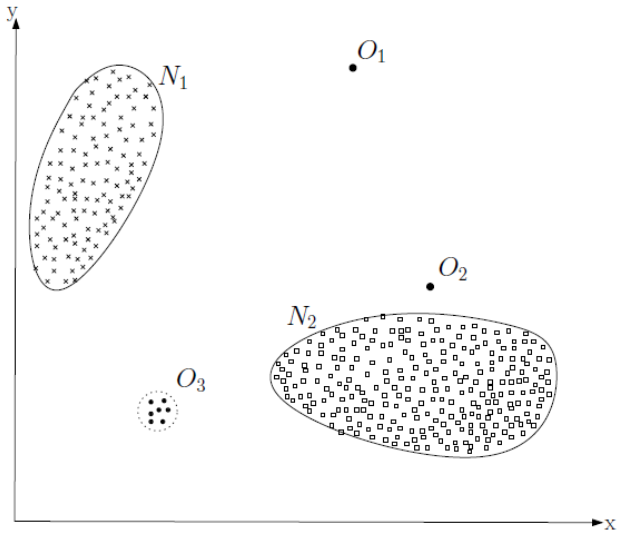
A utilização de qualquer um dos dois métodos, apenas dá uma indicação da probabilidade de ocorrência de fraude, isto porque nenhuma análise estatística por si só, pode assegurar que um objeto particular em análise é fraudulento. Apenas indica que o objeto em estudo é mais provável de ser fraudulento do que outros objetos (Jans, M. *et al.*, 2007, p. 7). É impossível ter a certeza absoluta sobre a legitimidade da intenção de quem efetua uma transação (Phua, C. *et al.*, 2010, p. 1).

2.5.3 Deteção de *Outliers*

O que é um *Outlier*? Segundo Hawkins (1980), é uma observação num conjunto de dados, que é suficientemente diferente ou aberrante dos restantes elementos e que levanta suspeita de ser causado por um mecanismo diferenciado.

A deteção de *outliers* refere-se ao problema de identificar padrões nos dados alvo de análise e que não estão em conformidade com um comportamento esperado como normal. Estes padrões anómalos são muitas vezes referidos como anomalias, observações discordantes, exceções, falhas, defeitos, aberrações, ruído, erros e peculiaridades nos diferentes domínios a que se aplicam (Chandola, V. *et al.*, 2009, p. 1).

Figura 2.11 – Exemplo de *outliers* num conjunto de dados de 2 dimensões



Fonte: Chandola et al, 2009, p. 2

Na figura 2.11 é possível observar-se duas zonas consideradas normais (N1 e N2), sendo que O1 e O2 representam duas observações remotas e O3 significa uma região distante.

Os *outliers* estão fora das zonas consideradas normais. Para além disso, os *outliers* existem em praticamente todos os conjuntos reais de dados (Chandola, V. *et al.*, 2009, p. 2).

Os *outliers* podem ser classificados em três categorias (Chandola, V. *et al.*, 2009, p. 8):

***Outliers* - Tipo 1 (Global)**

São o tipo mais simples de *outliers* e o foco da maioria dos esquemas de deteção de *outliers* existentes. As técnicas que detetam este tipo de *outlier* analisam a relação de uma observação individual em relação às restantes observações. Podem ser detetados sobre qualquer tipo de dados.

***Outliers* - Tipo 2 (Contexto)**

Este tipo de *outlier* também se trata de uma instância individual, mas a diferença para o tipo 1, é que o tipo 2 pode não ser considerado um *outlier* se estiver num contexto diferente. Ou seja, os *outliers* deste tipo são definidos em função do contexto.
A região de uma loja é um exemplo de um atributo de contexto.

***Outliers* - Tipo 3 (Espacial ou Sequencial)**

Este tipo de *outliers* ocorre quando um subconjunto das instâncias dos dados observados é periférico ou distante, em relação ao conjunto global dos dados. Este tipo de *outlier* apenas tem significado quando os dados têm uma natureza espacial ou sequencial. Uma característica deste tipo de *outlier* é a sua continuidade temporal, sendo um elemento chave (Manish, G. *et al.*, 2014, p. 1).

A análise de *outliers* temporais investiga anomalias no comportamento dos dados, ao longo do tempo (Manish, G. *et al.*, 2014, p. 1). Alguns exemplos dessa utilização: mercados financeiros, diagnósticos de sistemas (mecânicos, aeronaves), dados biológicos, ou transações de clientes.
Outras aplicações possíveis são descritas por Zhang, Y. *et al.* (2007) e por Hodge, V. *et al.* (2004), como aplicáveis nas seguintes áreas: deteção de intrusão em sistemas informáticos, monitorização ambiental, cuidados de saúde e controlo de logística e transportes.

2.5.4 Tipos de variáveis: qualitativas vs quantitativas

De seguida, e de forma mais pormenorizada, são descritas como podem ser analisadas as variáveis, consoante sejam qualitativas (texto) ou quantitativas (numéricas):

Variáveis qualitativas são aquelas cujos resultados são possíveis de ser observados na forma de categoria, qualidade ou característica. Estas variáveis podem ser analisadas segundo a distribuição de frequências que consiste na organização dos dados de acordo com as ocorrências dos diferentes resultados observados. A contagem de quantos elementos existem em cada categoria forma uma distribuição de frequência dos dados dessa variável, que pode ser apresentada em tabela ou gráfico. As frequências podem ainda ser apresentadas em forma absoluta ou forma relativa.

Como forma de conhecermos melhor as **Variáveis quantitativas** do conjunto de dados, várias características devem ser observadas, podendo ser agrupadas da seguinte forma:

1. Medidas de localização – estatísticas que resumem a informação da variável, dando informação do centro da distribuição dos dados:
    a. Média aritmética: indica o centro de um conjunto de valores, porém esta métrica perde a sua utilidade quando existem valores discrepantes (*outliers*).
    b. Mediana: centro posicional da distribuição, definida a partir da sucessão ordenada de observações, representando assim o valor central da variável.
    c. Moda: define o valor mais frequente da variável observada.
    d. Quartis (ou Percentis): correspondem a medidas de localização de tendência não central. Os quartis dividem a variável em 4 secções com igual frequência, correspondendo o Quartil 2 à mediana.

2. Medidas de dispersão – determinação da variabilidade ou dispersão da variável, relativamente à medida de localização do centro dos dados. Quanto mais heterogénea é a informação, maiores são os desvios.
    a. Amplitude: diferença entre o valor máximo e mínimo.
    b. Variância: estabelece os desvios em relação à média aritmética;
    c. Desvio-padrão: identifica a regularidade dos dados, correspondendo à raiz quadrada da variância. Quanto menor o seu valor, mais regular são os dados.

d. Amplitude Interquartílica (AIQ): diferença entre o 3º quartil e o 1º quartil. A AIQ engloba 50% das observações mais centrais.

3. Medidas de forma – permitem comprovar se uma distribuição de frequência tem características especiais como simetria, assimetria e nível de concentração:
   a. Assimetria/*Skewness*: indica o grau de distorção da distribuição em relação a uma distribuição simétrica. Revela o enviesamento que a distribuição apresenta relativamente à média. Se a distribuição se concentrar no lado esquerdo com uma longa cauda para a direita, o indicador toma valores maiores do que zero e a distribuição diz-se com enviesamento positivo.
   b. Achatamento/*Kurtose*: indica a forma da curva da distribuição em relação ao seu achatamento. A forma da curva de distribuição em relação à *kurtose* pode ser leptocúrtica (valores maiores do que 0), mesocúrtica (valores próximos de 0) ou platicúrtica (valores menos do que 0).

Com base nos resultados acima, é possível analisar as respetivas distribuições de frequências. Esta distribuição pode ser representada graficamente por um par de eixos cartesianos, sendo que o eixo horizontal representa a variável e o eixo vertical, as frequências. A distribuição de frequências de variáveis contínuas é feita dividindo a amplitude total dos dados (diferença entre o maior e o menor valor) em vários intervalos, denominados classes. Esses intervalos devem ser mutuamente exclusivos, exaustivos e de preferência ter o mesmo tamanho.

O histograma é a forma mais usual de apresentação de distribuições de frequências de uma variável contínua. São retângulos justapostos, feitos sobre as classes da variável em estudo (Abbott, D., 2014, p. 66). A área dos retângulos é igual ou proporcional à frequência observada da correspondente classe.

Podem ser efetuados testes estatísticos para aferir a normalidade dos dados, sendo que um deles é o teste não-paramétrico *Kolmogorov-Smirnov*. Este teste destina-se a averiguar se as variáveis seguem uma distribuição normal, através da hipótese nula das variáveis seguirem uma distribuição normal.

Através das estatísticas é possível analisar as diferenças inter-quartis e a presença de *outliers*, através do gráfico "Diagrama de extremos e quartis" (*box-plot*), que se representa por um retângulo e que evidencia o desvio inter-quartílico.

Este retângulo representa a faixa dos 50% dos valores mais típicos da distribuição. O retângulo é dividido no valor correspondente à mediana, assinalando o quartil inferior (Q1), a mediana (Q2) e o quartil superior (Q3). O valor de AIQ corresponde à diferença entre o 3º quartil (Q3) e o 1º quartil (Q1).

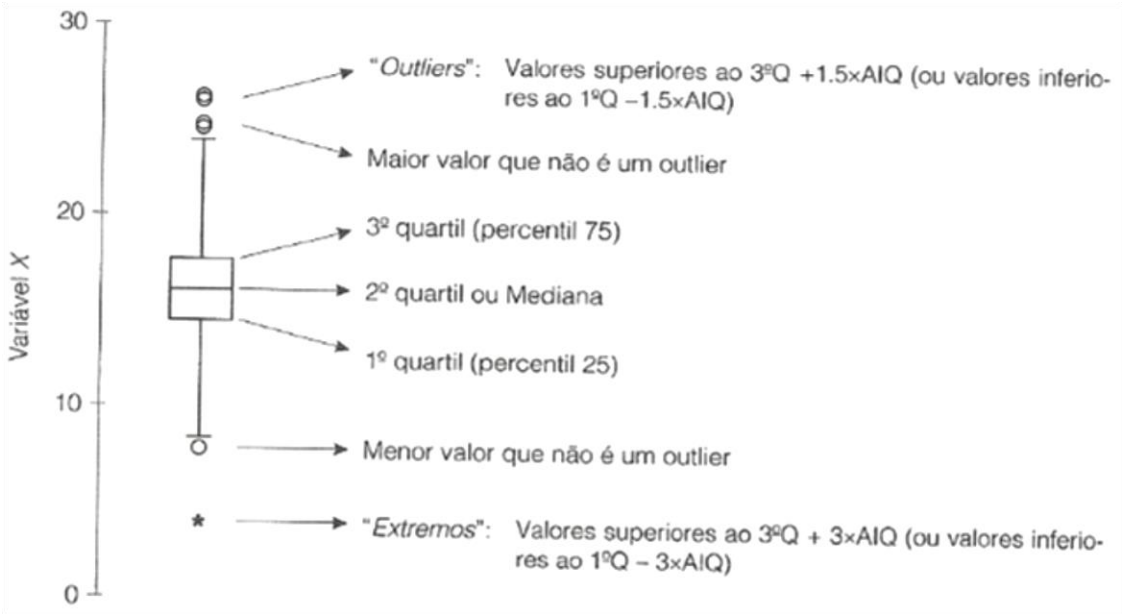
Os *outliers* são aqueles cujo valor é superior ao valor da expressão (Q3 + 1,5*AIQ) ou inferior ao valor da expressão (Q1 - 1,5*AIQ)

O *outlier* pode ser considerado de "moderado" se o seu valor estiver entre:

(Q3 + 1,5*AIQ) e (Q3 + 3*AIQ) ou (Q1 – 1,5*AIQ) e (Q1 – 3*AIQ)

O *outlier* pode ser considerado de "extremo" se o seu valor superior a: (Q3 + 3*AIQ) ou (Q1 – 3*AIQ).

Figura 2.12 – Diagrama de extremos e quartis (*box-plot*)



Fonte: Marôco, J., 2011, p. 28

A figura 2.12 exemplifica de forma gráfica o referido anteriormente, sendo um dos gráficos mais utilizados para descrever uma variável em estudo. Este gráfico pode ser gerado nas aplicações IBM SPSS, R ou Knime.

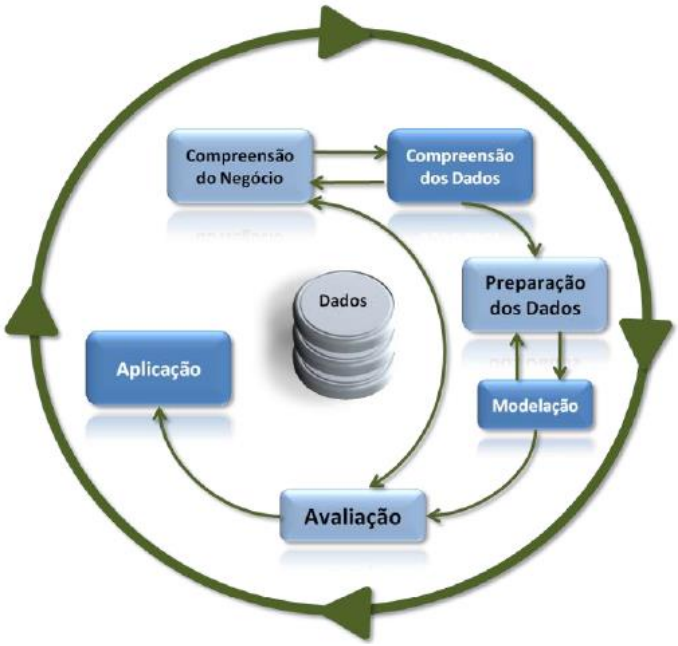
2.5.5 Metodologia usada em *Data Mining*

Nos últimos anos tem-se verificado um crescimento rápido do volume e da dimensão das bases de dados, causado principalmente pela queda no custo de armazenamento dos dados. A análise desses volumes de dados começou a ficar humanamente impossível utilizando os sistemas existentes.
Neste cenário surge, no final da década de 1980, um novo ramo da computação, a descoberta de conhecimento em bases de dados.

O processo de extração de conhecimento em bases de dados (*KDD - Knowledge Discovery in Databases*) é definido como sendo um processo iterativo, não trivial, de identificação de padrões válidos, novos, potencialmente úteis, compreensíveis e embebidos nos dados, envolvendo vários passos, e com decisões feitas pelo analista.
Na tentativa de tornar a extração de conhecimento das bases de dados uma realidade, e à medida que o mercado mostrou interesse pela área do *Data Mining*, tornou-se necessária a criação de uma nova abordagem, para demonstrar o quanto esta área era importante para ser parte do processo de negócio.

Em meados da década de 90, as empresas reuniram esforços para a criação de um processo padrão de *Data Mining*, denominado CRISP-DM (*Cross-Industry Standard Process for Data Mining*). Cada empresa contribuiu com suas experiências para o desenvolvimento de um padrão que pudesse ser aplicado a qualquer tipo de problema.
A metodologia CRISP-DM é descrita como um processo hierárquico, que contém seis fases, conforme se verifica na figura 2.13: compreensão do negócio, compreensão dos dados, preparação dos dados, modelação, avaliação e implementação.

Figura 2.13 – Ciclo de vida da metodologia CRISP-DM



Compreensão do Negócio: Na fase inicial procura-se identificar quais os objetivos e necessidades em termos de perspetivas do negócio, e converter as mesmas num problema de *Data Mining* pela definição de plano inicial de resolução do problema.

Compreensão dos Dados: Nesta fase seleciona-se os dados relevantes para o problema e identifica-se possíveis problemas na qualidade dos dados.

Preparação dos Dados: Nesta fase desenrola-se o processo de construção do conjunto de dados alvo de análise. São desenvolvidas tarefas como: seleção de tabelas, registos e atributos, transformação, limpeza dos dados, para a posterior utilização dos dados na aplicação de algoritmos de *Data Mining*.

Modelação: Aqui pretende-se selecionar e aplicar técnicas de modelação. Existem várias técnicas que podem ser utilizadas sobre o mesmo problema. Algumas dessas técnicas possuem requisitos específicos na formatação dos dados, sendo por isso muitas vezes necessário regressar à fase anterior, de preparação de dados.

Avaliação: Após a modelação, o modelo construído é avaliado, tendo em conta os requisitos do negócio, e é realizada a escolha do melhor modelo.

Implementação: Depois de construído e avaliado, o modelo pode ser implementado. Esta fase pode colidir com uma simples elaboração de relatórios para o apoio à tomada de decisão.

Capítulo 3: Deteção de indícios de fraude sobre devoluções de artigos

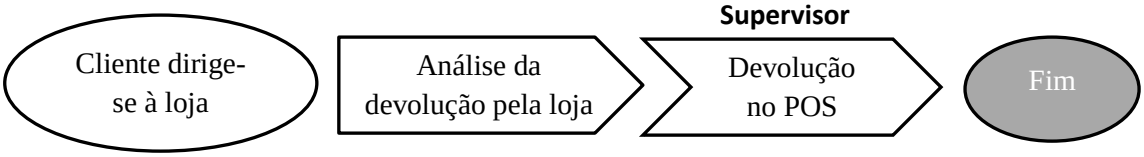
3.1 Objetivo e Âmbito

O objetivo deste trabalho de dissertação visa implementar técnicas de *Data Mining*, para detetar casos anómalos ou estranhos através da identificação de *outliers,* que possam revelar indícios de possível fraude, sobre um conjunto de dados referente ao processo de devoluções de artigos nas lojas.

Pelo recurso à metodologia CRISP-DM, a primeira fase do projeto deve ser o conhecimento do negócio.

Dentro do prazo de satisfação do cliente após a compra de um artigo, é possível ao cliente devolver o artigo, mediante apresentação do talão de compra desde o que artigo se encontre nas devidas condições. A devolução é efetuada por um supervisor da loja (chefia intermédia), que efetua o registo num POS (*Point of Sales*), reembolsando o cliente, conforme fluxo do processo:

Figura 3.1 – Processo das devoluções de artigos



Existem aqui dois riscos associados, sendo que um deles é o da devolução poder ser fictícia e do supervisor realizar a devolução para efetuar um reembolso fraudulento em proveito próprio, não havendo lugar a entrada de artigo em *stock* físico (havendo só em *stock* teórico). Outro risco é o do furto do artigo devolvido, sendo que neste caso a devolução foi efetuada por vontade do cliente, mas após o reembolso ao cliente, o artigo devolvido pelo cliente embora dê entrada no *stock* teórico da loja, fisicamente "desaparece".

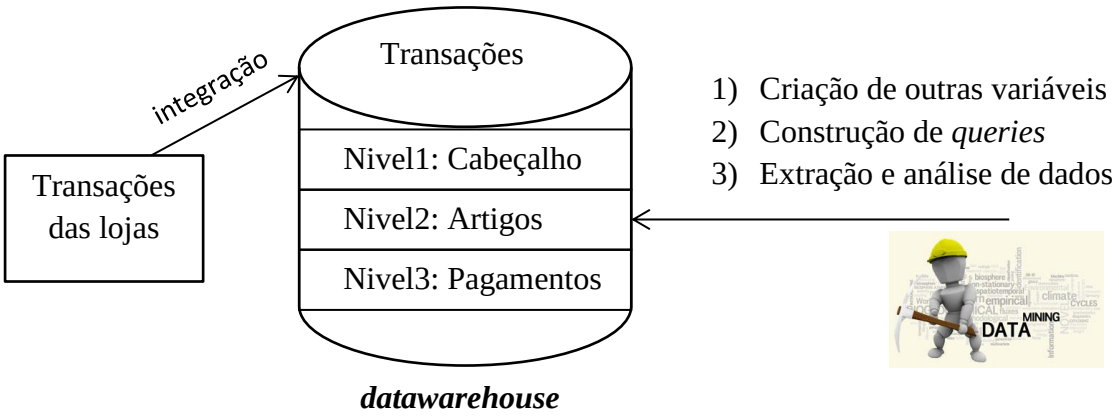
Embora existam outros esquemas de fraude possíveis de serem detetados, o âmbito deste trabalho foi definido por uma proposta direcionada para a utilização das técnicas de *Data Mining* na deteção de *outliers*, com recurso a um modelo de aprendizagem.

Será usada a técnica de aprendizagem do tipo não supervisionada, pelo facto de não termos uma variável "alvo" como classificador de cada caso, e apenas existirem as variáveis de entrada.

3.2 Caracterização do conjunto de dados

O conjunto de dados alvo aqui de análise, foi obtido através do desenvolvimento de uma *query* no sistema de informação (*datawarehouse*) do retalhista. Este sistema recebe diariamente as transações que ocorrem nos POS das lojas.
Em termos de arquitetura, estas transações são compostas por 3 níveis, sendo que o 1º designa-se por "cabeçalho", contendo informação como a data, hora, loja, pos, nº talão e operador/supervisor. No 2º nível é referente aos dados das linhas dos artigos que compõem a transação, e no 3º nível os meios de pagamento associados.

Figura 3.2 – Arquitetura das transações no sistema de informação



datawarehouse

Antes da extração dos dados, o 1º passo consistiu na proteção de duas variáveis a serem analisadas, com a criação de uma "mascara" sobre os campos "Loja" e

"Nr_Supervisor". Desta forma foi possível garantir a confidencialidade dos mesmos e a não adulteração dos dados reais.

No 2º passo, com base na informação do cabeçalho da transação, foram criadas três variáveis qualitativas:

Tabela 3.1 – Criação de 3 variáveis qualitativas

Variável	Descrição da Variável
Zona País	Para cada loja é identificada a região pertencente: - Norte (Norte Portugal atè Gaia) - Centro (Gaia atè Lisboa) - Sul (Lisboa até Algarve)
Dia semana	Classificação consoante a data do registo da devolução: - Dias úteis - Fds ou feriado (fim de semana ou feriado)
Horario registo	Classificação consoante a hora do registo da devolução: - Manhã (Da abertura da loja até às 13h) - Tarde (Das 13h até às 19h) - Noite (Das 19h até ao fecho da loja)

Por fim, foi desenvolvida uma *query* onde foram selecionadas as 5 variáveis referidas acima, e adicionadas outras já existentes no sistema corporativo como a "Unidade Negocio", "Total artigos devolvidos" e "Valor médio artigos devolvidos".

Como filtros aplicados à *query* definiu-se que o âmbito seriam as transações de devoluções de artigos, num universo de 40 lojas de um retalhista. O período de dados extraído correspondeu a três meses de transações, que ocorreram entre os meses de Dezembro-2014 e Fevereiro-2015, com exceção dos dias 24-Dezembro e 01-Janeiro, em que as respetivas lojas não estiveram abertas ao público.

No quadro seguinte são apresentadas as 8 variáveis que compõem o conjunto de dados que será a seguir alvo de análise:

Tabela 3.2 – As 8 variáveis do conjunto de dados

			SPSS	
Variável	**Descrição da Variável**	**Tipo de Variável**	**Tipo**	**Medida**
Loja	Código associado a cada loja (irrepetível / diferenciador)	Qualitativa	String	Nominal
Nr Supervisor	Código associado a cada supervisor (irrepetível / diferenciador)	Qualitativa	String	Nominal
Zona País	"Norte", "Centro" ou "Sul"	Qualitativa	String	Nominal
Dia semana	"Dias úteis" ou "Fds ou feriado"	Qualitativa	String	Nominal
Horario registo	"Manhã", "Tarde" ou "Noite"	Qualitativa	String	Nominal
Unidade Negocio	Descrição do tipo de artigo. Existem 18 unidades de negócio como: - Peixaria, Talho, Bebidas, Congelados, Padaria, Cultura, Lazer, etc	Qualitativa	String	Nominal
Total artigos devolvidos	Total de artigos devolvidos pelo supervisor no periodo analisado	Quantitativa	Numeric	Scale
Valor médio artigos devolvidos	Valor médio dos artigos devolvidos pelo supervisor no periodo analisado	Quantitativa	Custom Currency	Scale

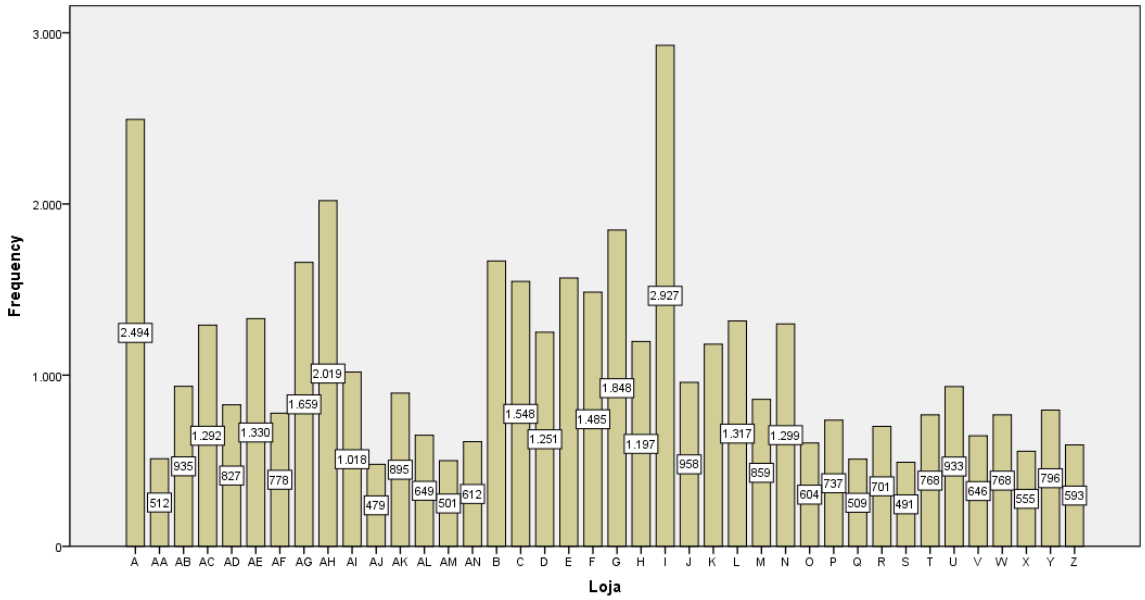
3.3 Análise estatística das variáveis

Depois de referido como foram obtidos os dados, a fase seguinte passa pela realização das análises estatísticas sobre os dados. Os dados são observações de variáveis qualitativas e quantitativas, porém as técnicas de análise são diferentes para cada tipo de variável. Este conjunto de dados é composto por 43.206 observações e 8 variáveis, sendo 6 qualitativas e 2 quantitativas.

3.3.1 Variáveis qualitativas

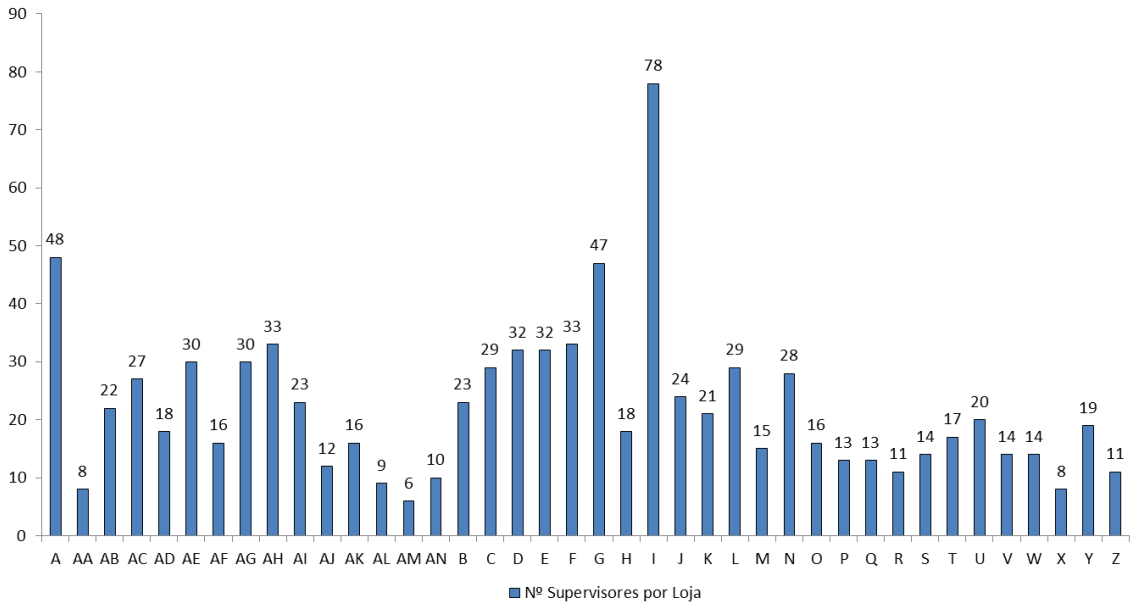
Conforme referido anteriormente, o conjunto de dados é referente a um universo de 40 lojas, mas conforme se verifica na figura 3.3, o nº de observações da variável "Loja" é bastante distinto, onde se destaca a loja "I" com o maior número (2927 registos) e a loja "AJ" com o menor número (479 registos), sendo a média de 1080 registos por loja.
Apesar de todas as lojas pertenceram à mesma cadeia e terem a mesma tipologia, o nº de clientes e o volume de vendas é diferente de loja para loja, o que se traduz num maior ou menor nº de devoluções registadas.

Figura 3.3 – Gráfico de frequências da variável "Loja"



Efetuando uma análise pela variável "Nr Supervisor" verifica-se que o conjunto de dados é constituído por devoluções efetuadas por 887 supervisores distintos, das 40 lojas. Sendo impossível de representar graficamente a distribuição de frequências por supervisor, optou-se por efetuar uma análise do nº de supervisores por loja:

Figura 3.4 – Gráfico do nº de Supervisores por Loja



Sendo a média de 22 supervisores por loja, observa-se no entanto que existe uma loja "I" com 78 supervisores com devoluções efetuadas no período observado, e uma loja com apenas 6 supervisores nessas condições. De referir que não existe uma política nas lojas que estabeleça o nº mínimo ou máximo de devoluções, mas é recomendado que esteja atribuído a colaboradores com a função de supervisor, e que os cartões não sejam partilhados.

Efetuando agora uma análise à variável "Zona País" observa-se que as 40 lojas estão distribuídas em 3 zonas geográficas, sendo que a zona Sul representa cerca de 45% do total das lojas em estudo:

Tabela 3.3 – Nº de lojas por "Zona País"

Zona do País	freq. abs.	freq. rel. (%)
Norte	11	28%
Centro	11	28%
Sul	18	45%
Total	**40**	**100%**

Analisando agora as frequências da variável "Zona País", verifica-se que 50% das observações são referentes às lojas da zona Sul:

Tabela 3.4 – Frequência da variável "Zona País"

Zona País	freq. abs.	freq. rel. (%)
Norte	12232	28%
Centro	9538	22%
Sul	21436	50%
Total	**43206**	**100%**

Na tabela 3.5 pretende-se analisar se no conjunto de dados existem mais observações em "Dias úteis" do que em comparação com "Fins de semana ou feriados". Apesar de na realidade os fins de semana e feriados corresponderem a cerca de 30% de uma semana, a sua frequência relativa é de 44%, o que permite concluir que existe uma maior tendência a se realizarem mais devoluções ao fim de semana do que comparativamente aos dias úteis de trabalho:

Tabela 3.5 – Frequência da variável "Dia semana"

Dia semana	freq. abs.	freq. rel. (%)
Dias úteis	24006	56%
Fds ou feriado	19200	44%
Total	**43206**	**100%**

Na tabela 3.6 é feita uma análise da distribuição de frequências sobre o horário onde é feito o registo da devolução. Verifica-se que 42% das observações são referentes ao período da tarde, podendo estar relacionado com o facto de ser o período com maior nº de horas associado (das 13h às 19h):

Tabela 3.6 – Frequência da variável "Horario registo"

Horario registo	freq. abs.	freq. rel. (%)
Manhã	12727	29%
Tarde	18060	42%
Noite	12419	29%
Total	**43206**	**100%**

Na tabela 3.7 é efetuada uma análise de frequências à variável "Unidade de Negócio", que consiste na descrição do tipo de produto. Verifica-se que as unidades "Congelados" e "TakeAway" representam menos de 4% de frequência relativa, o que indica que são tipos de artigos menos suscetíveis de serem devolvidos.

Por outro lado, produtos do tipo "Higiene e Beleza" e "Mercearia Doce" são os que apresentam frequências relativas mais altas, com 6,9%:

Tabela 3.7 – Frequência da variável "Unidade de Negócio"

Unidade Negocio	freq. abs.	freq. rel. (%)
Bebidas	2714	6,3%
Bricolage e Auto	2326	5,4%
Casa	2655	6,1%
Charcutaria&Queijos	2532	5,9%
Congelados	1499	3,5%
Cultura	2642	6,1%
Frutas e Legumes	2571	6,0%
Higiene e Beleza	2966	6,9%
Lacticínios	2812	6,5%
Lazer	2344	5,4%
Limpeza do Lar	2612	6,0%
Mercearia Doce	2964	6,9%
Mercearia Salgada	2732	6,3%
Padaria	2366	5,5%
Peixaria	2119	4,9%
Pets&Plants	1800	4,2%
TakeAway	1462	3,4%
Talho	2090	4,8%
Total	**43206**	**100%**

3.3.2 Variáveis quantitativas

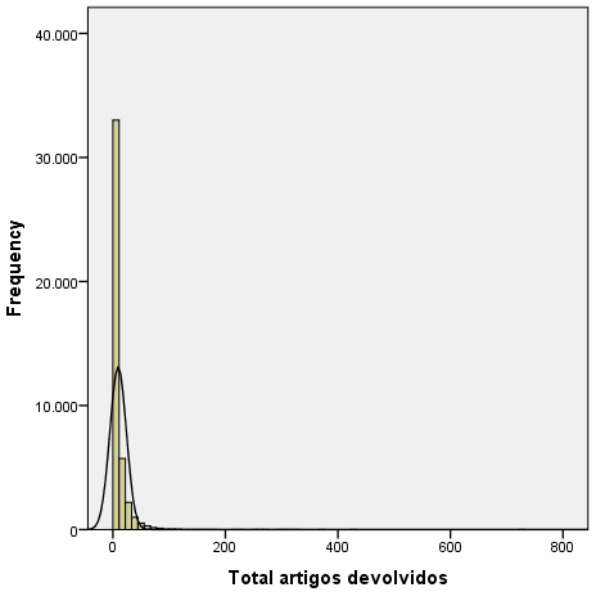
Na tabela 3.8 apresenta-se o resultado das medidas de localização, dispersão e de forma, aplicadas às variáveis quantitativas presentes no conjunto de dados: "Total artigos devolvidos" e "Valor médio artigos devolvidos".

Tabela 3.8 – Medidas estatísticas das variáveis quantitativas

Medidas			Variáveis quantitativas em análise	
			Total artigos devolvidos	Valor médio artigos devolvidos
Nº observações			43.206	
Localização	Média		9,3	8,26 €
	Mediana		4	4,64 €
	Moda		1	1,99 €
	Quartis	25	2	2,48 €
		50	4	4,64 €
		75	11	9,47 €
Dispersão	Minimo		1	0,02 €
	Máximo		725	1.167,12 €
	Amplitude		724	1.167,10 €
	Desvio Padrão		14,6	14,07 €
	Variância		213	197,85 €
Forma	Skewness		7,9	21,87 €
	Desvio padrão do Skewness		0,01	0,01
	Kurtosis		199,1	1.269,80 €
	Desvio padrão do Kurtosis		0,02	0,02

De forma complementar, é possível analisar os seguintes histogramas para as variáveis quantitativas. Nestes histogramas é desenhada a linha de distribuição normal o que permite evidenciar as medidas de assimetria e achatamento:

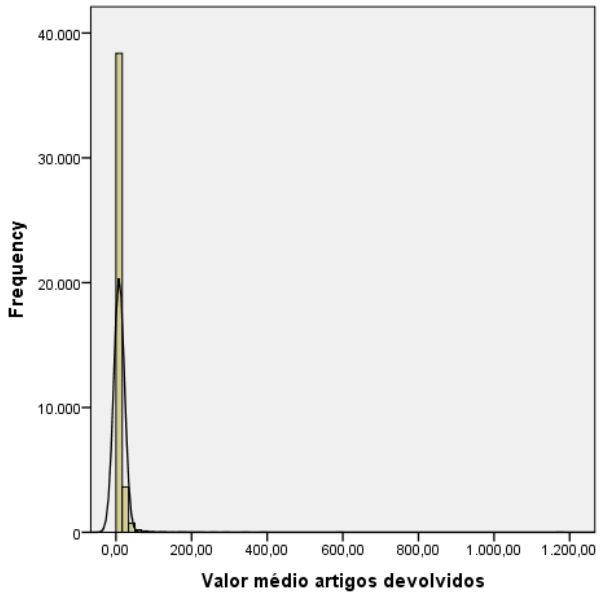
Figura 3.5 – Histograma da variável "Total artigos devolvidos"



Na figura 3.5, a variável "Total artigos devolvidos" mostra uma amplitude considerável entre os seus valores extremos, bem como uma variância elevada de 213.

A distribuição tem um enviesamento positivo, o que indica que apesar da grande amplitude, o nº de artigos devolvidos localiza-se principalmente no 1º intervalo (entre 1 e 11).

Figura 3.6 – Histograma da variável "Valor médio artigos devolvidos"



Na figura 3.6, a variável "Valor médio artigos devolvidos" mostra uma amplitude considerável entre o seu valor mínimo e máximo, bem como uma variância elevada de 197,85€.

A distribuição tem um enviesamento positivo, o que indica que apesar da grande amplitude, o valor médio dos artigos devolvidos localiza-se logo no 1º intervalo (entre 0,02€ e 16,67€).

De seguida é realizado um teste às variáveis quantitativas, a fim de se verificar se as mesmas seguem uma distribuição normal. Pelos valores obtidos na figura 3.7, no valor do *p-value* ("Asymp. Sig."), sendo os ambos inferiores ao nível de significância de 0.05 , a hipótese nula é rejeitada nos dois casos, logo pode-se concluir que as variáveis não seguem uma distribuição normal.

Figura 3.7 – Teste Kolmogorov-Smirnov

One-Sample Kolmogorov-Smirnov Test

		Total artigos devolvidos	Valor médio artigos devolvidos
N		43206	43206
Normal Parameters[a,b]	Mean	9,34	8.2616
	Std. Deviation	14,593	14.06605
Most Extreme Differences	Absolute	,284	,286
	Positive	,225	,224
	Negative	-,284	-,286
Kolmogorov-Smirnov Z		59,009	59,370
Asymp. Sig. (2-tailed)		,000	,000

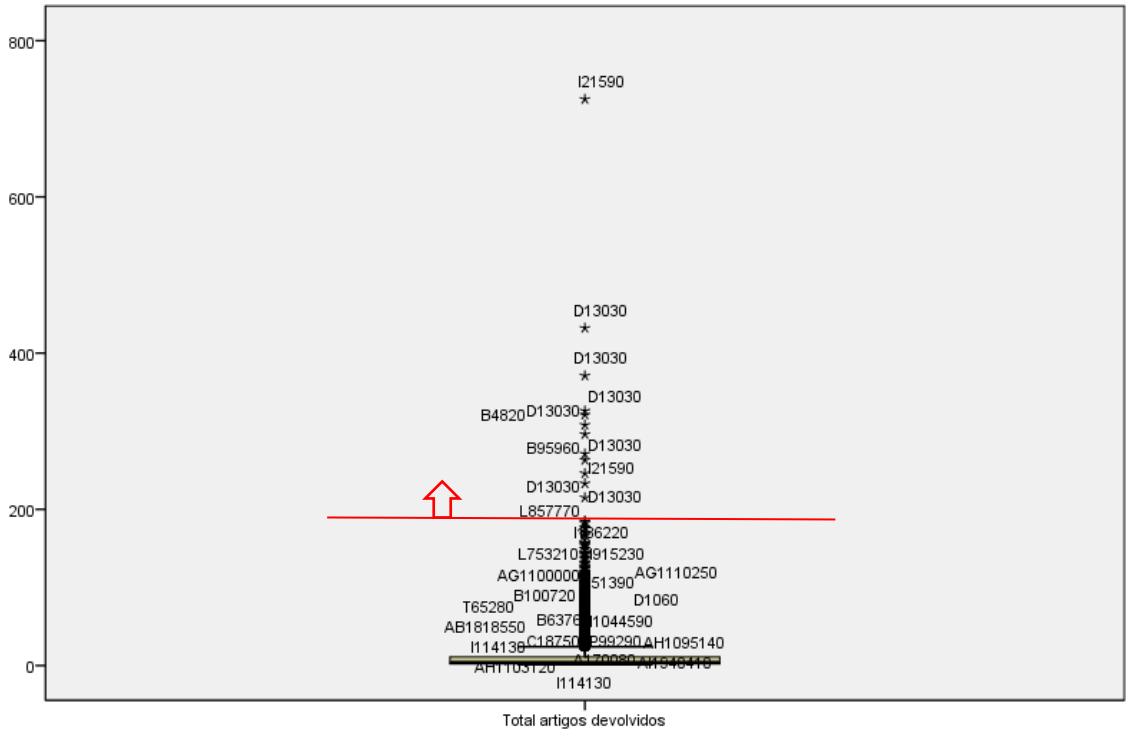
a. Test distribution is Normal.

b. Calculated from data.

3.4 Deteção de *outliers* por análises univariadas

Neste ponto dá-se início à identificação de *outliers* através da representação gráfica *box-plot* na aplicação SPSS (também possível no R ou Knime). Nesta fase são realizadas análises apenas sobre cada uma das variáveis quantitativas, e destacados apenas os *outliers* extremos:

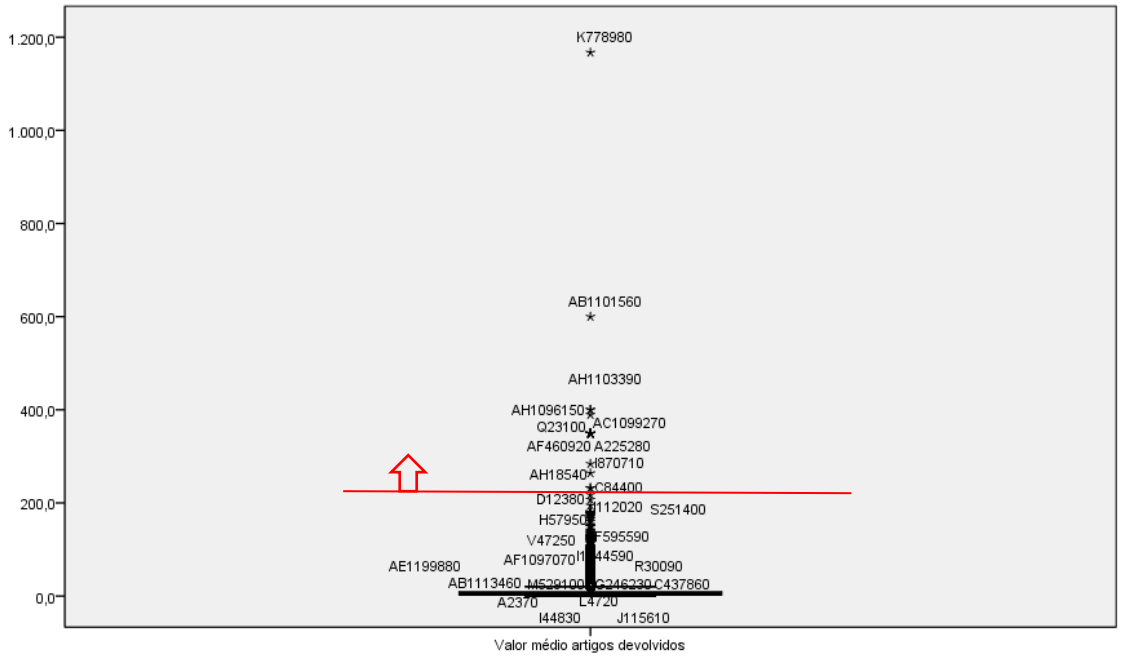
Figura 3.8 – *Box-Plot*: "Total artigos devolvidos"



Observa-se que o grupo é pouco homogéneo, pelo que o número de *outliers* globais extremos é muito elevado, dado existirem 1695 *outliers* de valores acima de 38, obtido através da expressão (11 + 3 * (11-2)).

Desse modo, optou-se por destacar apenas **12 *outliers* globais extremos**, visualmente mais significativos, com total de artigos devolvidos superior a 200 unidades: dois do supervisor "I21590", sete do supervisor "D13030" e um de cada um dos supervisores "B4820", "B95960" e "L857770".

Figura 3.9 – *Box-Plot*: "Valor médio artigos devolvidos"



Observa-se que o grupo é pouco homogéneo, pelo que o número de *outliers* globais extremos é muito elevado, dado existirem 1467 *outliers* de valores acima de 30.5, obtido através da expressão (9.5 + 3 * (9.5-2.5)). Desse modo, optou-se por destacar apenas **11 *outliers* globais extremos**, visualmente mais significativos, com valores médios de artigos devolvidos superiores a 225€: supervisor "K778980", "AB1101560", "AH1103390", "AH1096150", "AC1099270", "Q23100", "AF460920", "A225280", "I870710", "AH18540" e "C84400".

Conclusão: Através das análises univariadas pelos gráficos *box-plot* foi possível identificar *outliers* globais, sendo que cada uma das variáveis quantitativas revelou *outliers* distintos. Contudo, estas análises podem ser consideradas muito "macro" por darem apenas uma visão geral das variáveis, não revelando *outliers* de contexto. No quadro seguinte é apresentado um resumo dos 23 *outliers* identificados até aqui:

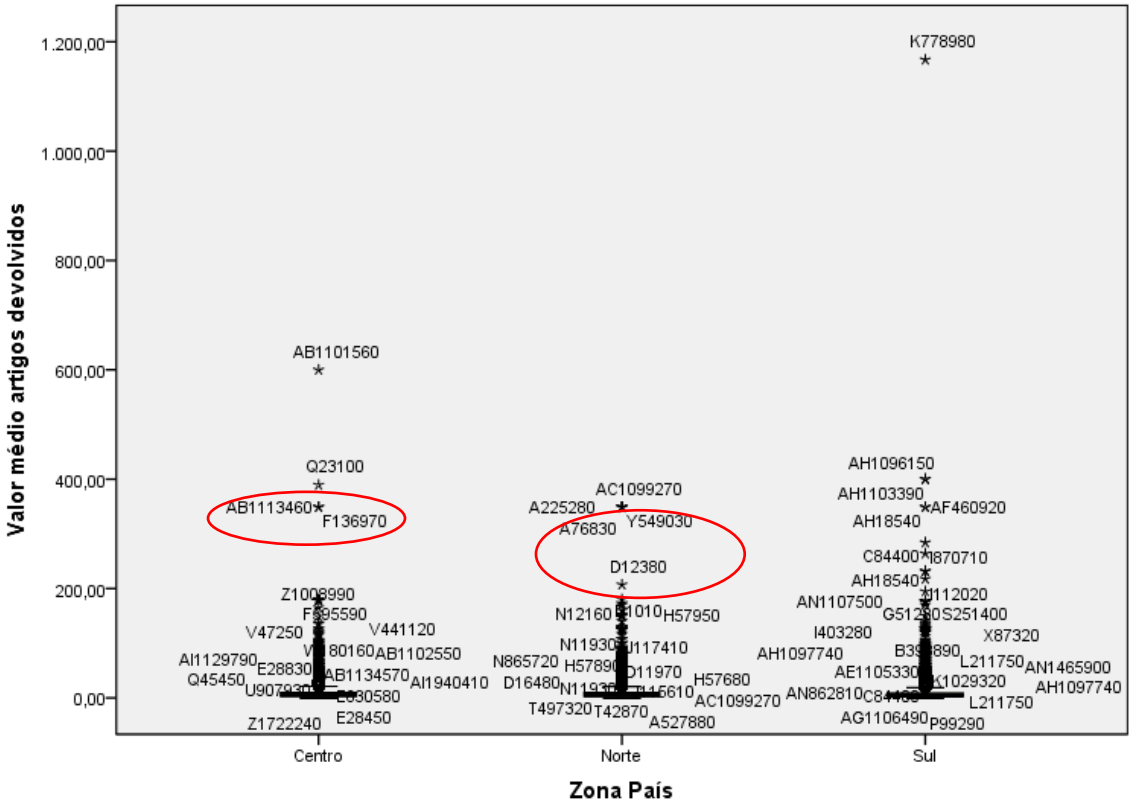
Tabela 3.9 – *Outliers* extremos globais identificados nas análises univariadas

Análises univariadas	***Outliers* globais extremos identificados**	
	Descrição	**Total**
Total artigos devolvidos	I21590 {2} | D13030 {7} | B4820 | B95960 | L857770	**12**
Valor médio artigos devolvidos	K778980 | AB1101560 | AH1103390 | AH1096150 | AC1099270 | Q23100 AF460920 | A225280 | I870710 | AH18540 | C84400	**11**

3.5 Deteção de *outliers* por análises bivariadas

O objetivo neste ponto passa pela criação de um modelo iterativo, onde se vai realizando análises bivariadas (duas variáveis) pela combinação das várias variáveis entre si. Essa iteração consiste na avaliação de cada gráfico, na perspetiva de qual o valor acrescentado que o mesmo traz em termos de relevância de *outliers* novos, aos já conhecidos pelas análises univariadas. Neste capítulo apenas serão ilustrados os gráficos onde são observados novos *outliers*, sendo que os restantes gráficos podem ser consultados no Anexo II.

Figura 3.10 – *Box-Plot*: "Zona País" vs "Valor médio artigos devolvidos"

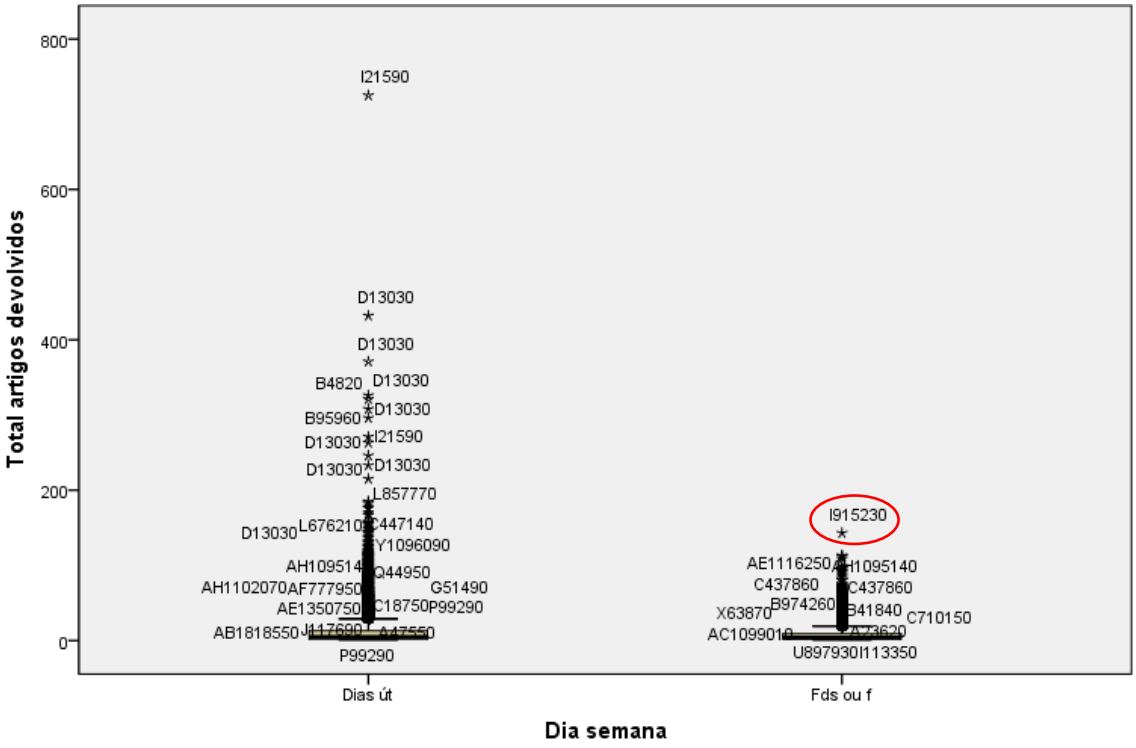


Nesta análise pretende-se verificar e detetar a presença de *outliers* da variável "Valor médio artigos devolvidos" em função da área geográfica onde a loja se situa.

Conclusão: Observa-se que o número de *outliers* é muito elevado, pelo que de forma mais nítida, é possível identificar **5 novos *outliers* extremos** que não foram identificados na análise univariada, para além de se observar os *outliers* globais já conhecidos:

- zona "Centro" - Supervisores "AB1113460" e "F136970"
- zona "Norte" – Supervisores "Y549030", "A76830" e "D12380"

Figura 3.11 – *Box-Plot*: "Dia semana" vs "Total artigos devolvidos"



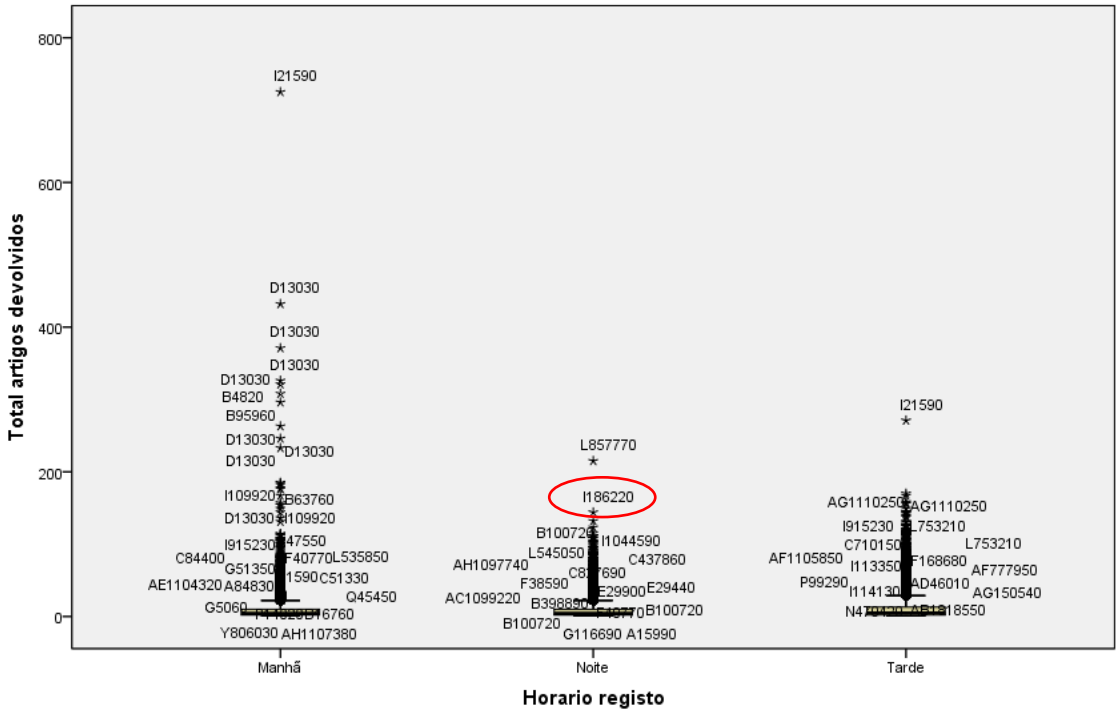
Analisando agora as variáveis "Dia semana" e "Total artigos devolvidos", pretende-se detetar a presença de *outliers*, em função do dia da semana, ou seja, dias úteis ou fins de semana/feriados.

Observa-se que o número de *outliers* é muito elevado, pelo que de forma mais nítida, é possível identificar um novo *outlier* extremo, quando a variável "Dia semana" é "Fins de Semana ou Feriados", no supervisor "I915230".

Nos "Dias úteis" não se identificam *outliers* novos, apenas os casos já conhecidos.

Conclusão: esta análise revelou-se útil pela deteção de **um novo *outlier* extremo**, que não foi identificado na análise univariada, para além de se observar os *outliers* globais já conhecidos.

Figura 3.12 – *Box-Plot*: "Horario registo" vs "Total artigos devolvidos"



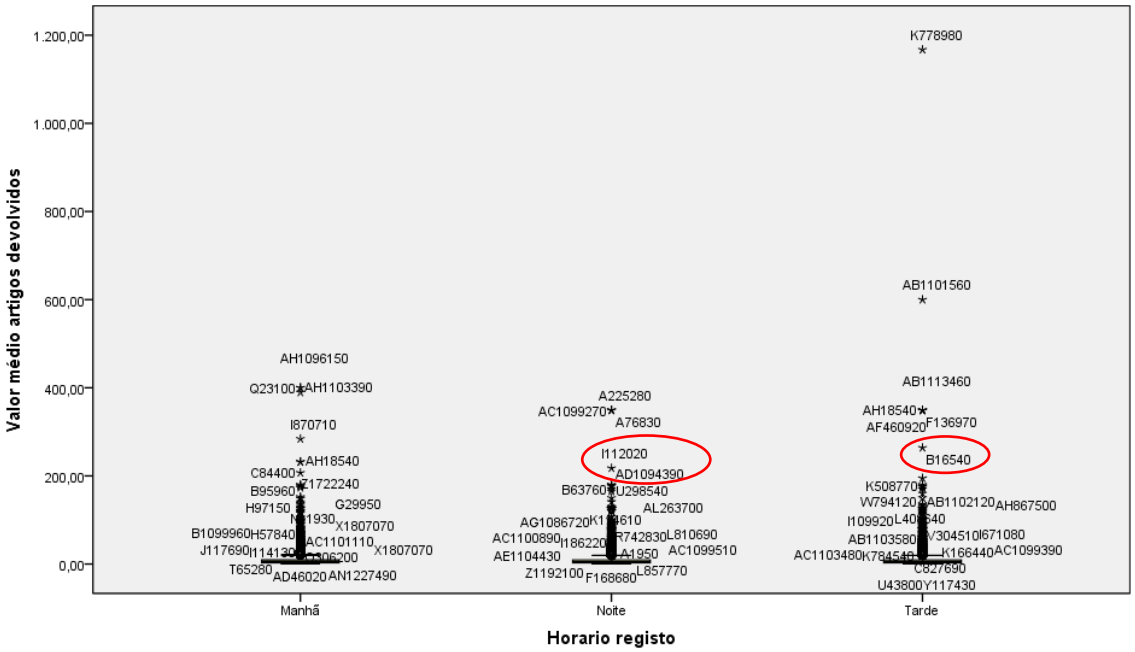
Nesta análise pretende-se verificar e detetar a presença de *outliers* da variável "Total artigos devolvidos" em função do momento temporal em que se verificou o registo da devolução.

Observa-se que o número de *outliers* é muito elevado, pelo que de forma mais nítida, é possível identificar um novo *outlier* extremo, quando a variável "Horario registo" é "Noite", no supervisor "I186220".

Na "Manhã" e "Tarde" não se identificam *outliers* novos, apenas os casos já conhecidos.

Conclusão: esta análise revelou-se útil pela deteção de **um novo *outlier* extremo**, que não foi identificado na análise univariada, para além de se observar os *outliers* globais já conhecidos.

Figura 3.13 – *Box-Plot*: "Horario registo" vs "Valor médio artigos devolvidos"

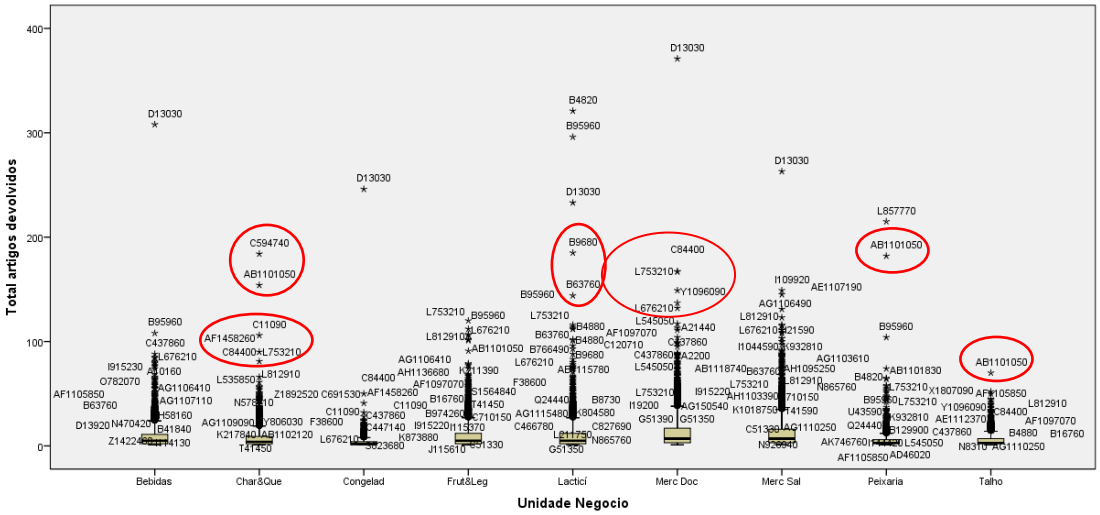


Conclusão: Nesta análise bivariada entre as variáveis "Horario registo" e "Valor médio artigos devolvidos" observa-se que o número de *outliers* é muito elevado, pelo que de forma mais nítida, é possível identificar **três novos *outliers* extremos**: na "Noite" destacam-se dois novos casos no supervisor "I112020" e "AD1094390" e na "Tarde" observa-se um *outlier* associado ao supervisor "B16540".

Na análise bivariada seguinte, pelo facto de existirem 18 unidades de negócio diferentes no conjunto de dados, optou-se por efetuar uma análise gráfica *box-plot* em duas partes (9 unidades de negócio cada), por forma a permitir uma leitura mais percetível.

Na 1ª parte da análise pretende-se verificar e detetar a presença de *outliers* extremos das variáveis "Total artigos devolvidos" e "Valor artigos devolvidos" nas Unidades de Negócio como Bebidas, Charcutaria & Queijos, Congelados, Frutas & Legumes, Lacticínios, Mercearia Doce, Mercearia Salgada, Peixaria e Talho.

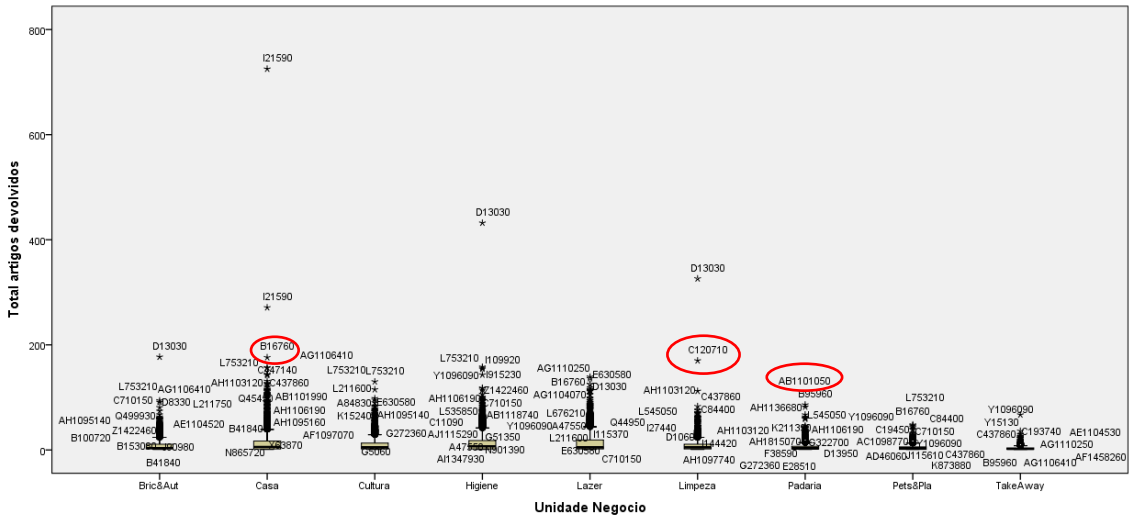
Figura 3.14 – *Box-Plot*: "Unidade Negócio" (parte 1) **vs "Total artigos devolvidos"**



Conclusão: No gráfico da 1ª parte desta análise bivariada observam-se **14 novos *outliers* extremos** nas unidades de negócio "Charcutaria & Queijos", "Lacticinios", "Mercearia Doce", "Peixaria" e "Talho".

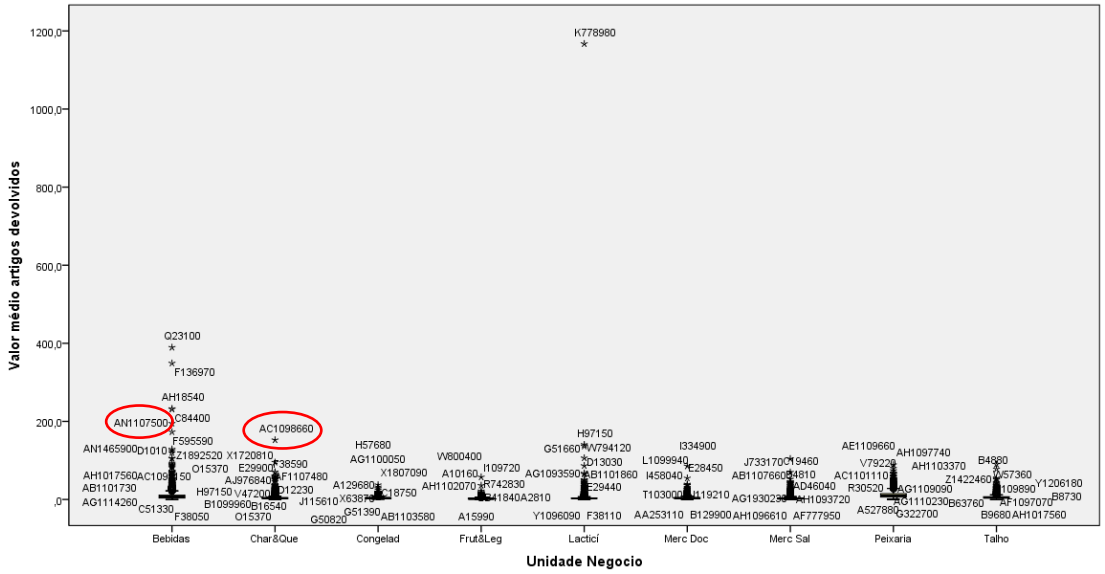
Pretende-se agora analisar as mesmas variáveis nas Unidades de Negócio como Bricolage, Casa, Cultura, Higiene, Lazer, Limpeza, Padaria, Pets&Plants e Take Away.

Figura 3.15 – *Box-Plot*: "Unidade Negócio" (parte 2) **vs "Total artigos devolvidos"**



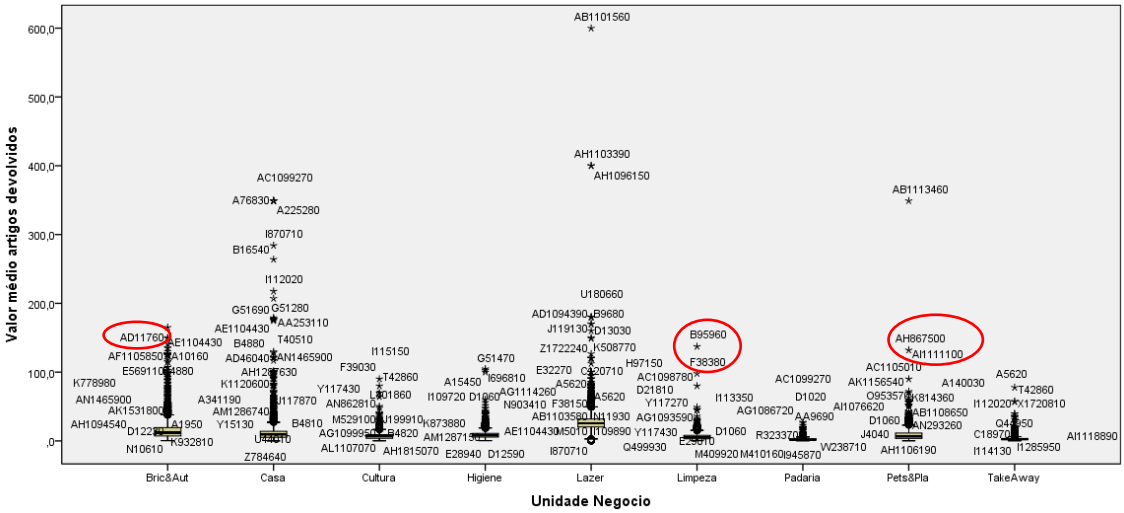
Conclusão: No gráfico da 2ª parte desta análise bivariada observa-se **três novos *outliers* extremos** (supervisores "B16760", "C120710" e "AB1101050").

Figura 3.16 – *Box-Plot*: "Unidade Negócio" (parte 1) **vs "Valor médio artigos devolvidos"**



Conclusão: No gráfico da 1ª parte desta análise bivariada observam-se **2 novos *outliers* extremos** nas unidades de negócio "Bebidas" e "Charcutaria & Queijos".

Figura 3.17 – *Box-Plot*: "Unidade Negócio" (parte 2) **vs "Valor médio artigos devolvidos"**



Conclusão: No gráfico da 2ª parte desta análise bivariada observam-se **cinco novos *outliers* extremos** (supervisores "AD11760", "B95960", "F38380", "AH867500" e "AI1111100").

Após a realização das análises bivariadas entre todas as variáveis, é importante resumir os novos *outliers* extremos identificados, dado que não tinham sido observados nas análises univariadas:

Tabela 3.10 – Novos *Outliers* extremos identificados nas análises bivariadas

Análises bivariadas		Novos *Outliers* extremos identificados	
		Total artigos devolvidos	Valor médio artigos devolvidos
Zona País	Centro	-	AB1113460 | F136970
	Norte		A76830 | Y549030 | D12380
Dia semana	Fins de semana ou feriados	I915230	-
Horario registo	Noite	I186220	I112020 | AD1094390
	Tarde		B16540
Unidade Negocio	Bebidas		AN1107500
	Charcutaria & Queijos	C594740 | AB1101050 | C11090 | AF1458260 | C84440 | L753210	AC1098660
	Lacticinios	B9680 | B63760	
	Mercearia Doce	C84400 | L753210 | Y1096090 | L676210	
	Peixaria	AB1101050	
	Talho	AB1101050	
	Bricolage & Auto		AD11760
	Casa	B16760	
	Limpeza	C120710	B95960 | F38380
	Padaria	AB1101050	
	Pets&Plants		AH867500 | AI1111100
Total		**19**	**15**

Conforme se verifica pela tabela 3.10, através das análises bivariadas, foram identificados **34 novos *outliers* extremos**.

De seguida são apresentadas as vantagens e desvantagens das análises bivariadas em relação às univariadas:

Vantagem: as análises bivariadas permitiram detetar vários *outliers* diferentes dos observadores nas análises univariadas, pelo que enriquecem o trabalho realizado na concretização dos objetivos definidos. Estes novos *outliers* não tinham sido identificados nas análises univariadas.
Cruzar as variáveis, ainda que aos pares, é sempre melhor do que analisar isoladamente cada variável.

Desvantagem: a realização das análises bivariadas requer algum esforço extra para gerar graficamente todas as combinações entre as variáveis disponíveis. Em conjuntos de dados com muitas variáveis pode ser bastante trabalhoso, e não trazer o retorno esperado. Nesse caso a escolha das variáveis mais importantes torna-se um fator crítico de sucesso.

3.6 Deteção de *outliers* por análises multivariadas (árvores de regressão)

Neste ponto pretende-se estudar as duas variáveis quantitativas, através da combinação das variáveis qualitativas em simultâneo (com exceção do "Nr Supervisor" por não se tratar de uma variável descritiva). Dado o nº de variáveis e o nº de instâncias possíveis de cada uma, o nº de combinações torna-se exponencial, o que seria humanamente quase impossível efetuar a análise e teste a todas as combinações possíveis. Desse modo, o recurso a um modelo de aprendizagem, torna-se fundamental para efetuar análise a todas as possíveis combinações, e na seleção das mais relevantes.

A opção recaiu pela utilização de **árvores de regressão**, pela sua capacidade de procurar possíveis soluções em problemas complexos, onde a estratégia passa por ir dividindo em problemas mais simples (Gama, J. *et al.*, 2012, p. 101).

As árvores de regressão são em tudo idênticas às árvores de classificação, sendo que a grande diferença consiste no facto de as folhas da árvore de regressão conterem previsões numéricas e não decisões.

Um dos algoritmos mais conhecidos para a indução de árvores de árvores de classificação e regressão é o algoritmo CART (Breiman, L. *et al.*, 1984). Este algoritmo inicia a sua execução na raiz da árvore de regressão, começando por analisar o conjunto completo dos dados, e calculada a variância da variável objetivo.

Para cada variável, e para cada possível teste no valor da variável, é calculada a redução da variância associada a esse teste. O teste que provoca uma maior redução na variância é escolhido como teste para o nó (Gama, J. *et al.*, 2012, p. 110).

Em cada divisão da árvore, é efetuado esse teste, com a escolha da variável e respetivos valores que melhor separam os dados em dois grupos, e que minimize a variância desse grupo na procura de criar grupos mais homogéneos.

O objetivo passa por minimizar a variância da variável objetivo, à medida que se expande a árvore. No início (na raiz da árvore), o grupo é heterogéneo e tem uma variância elevada, mas á medida que se vai descendo de nível, os grupos vão ficando mais homogéneos e a variância vai diminuindo.

O **desvio** ("deviance") de um nó T é calculado através do total da soma dos quadrados (SS_T):

$$SS_T = \sum (y_i - \bar{y})^2$$

Cada divisão é feita através da maximização entre: a impureza do nó atual T, menos a soma da impureza dos dois nós filho (L e R): $SS_T - (SS_L + SS_R)$.

O processo termina quando já não é possível dividir mais os nós da árvore, isto é, não sendo possível diminuir mais a variância do respetivo grupo.

Se dividirmos este valor pelo tamanho de observações do nó (n), obtemos o valor da variância.

A obtenção das árvores de regressão pelo algoritmo CART está disponível na aplicação R (R Core Team, 2014), através do pacote *rpart* (Therneau, T. *et al.*, 2015) e importando os dados a serem analisados.

```
library(rpart)
dados_dev = read.table("Dados_Devolucoes.csv", sep=";", header=T)
```

Na **função *rpart*** devem ser definidas as variáveis qualitativas e a variável quantitativa objetivo, sendo que neste caso em estudo, pretende-se construir duas árvores de regressão para as variáveis "Total artigos devolvidos" e "Valor médio artigos devolvidos". Na função *rpart* deve ser definido o parâmetro do método como "Anova", por se tratar de árvores de regressão. Outro **parâmetro** fundamental é o "**cp**" que controla a complexidade da árvore. O seu valor define que qualquer divisão na árvore deve diminuir a total falta de ajuste, antes da árvore ser expandida. Este parâmetro permite economizar tempo de processamento ao eliminar divisões que são desnecessárias.

```
fit <- rpart(Total_artigos_devolvidos ~ Loja + Zona + Dia + Horario + UN, method="anova", data=dados_dev, cp=0.0019)
plot(fit, uniform=TRUE, main="Árvore de Regressão sobre Total Artigos Devolvidos", cex=0.8)
text(fit, use.n=TRUE, all=TRUE, cex=.75)
fit
fit$frame
```

```
fit <- rpart(Valor_medio_artigos_devolvidos ~ Loja + Zona + Dia + Horario + UN, method="anova", data=dados_dev, cp=0.0013)
plot(fit, uniform=TRUE, main="Árvore de Regressão sobre Valor médio artigos devolvidos", cex=0.8)
text(fit, use.n=TRUE, all=TRUE, cex=.75)
fit
fit$frame
```

Após a obtenção da cada uma das árvores, em cada nó/folha da árvore serão realizadas análises gráficas *box-plot*, observando se o gráfico revela algum *outlier* até aqui desconhecido, ou apenas os já conhecidos.

3.6.1 Árvore de regressão: variável objetivo "Total artigos devolvidos"

Através da função *rpart* é obtida a árvore de regressão para a variável objetivo "Total artigos devolvidos". A árvore da figura 3.18 é o resultado final de testes realizados na função *rpart*, alterando o parâmetro "cp", de forma a obter folhas da árvore com valores para a variável objetivo o mais elevado possível. Após várias tentativas, o valor "otimizado" do "cp" foi de 0.0019, obtendo-se uma folha da árvore com 16 observações e o valor médio da variável objetivo de 88,81.

Figura 3.18 – Árvore regressão variável objetivo "Total artigos devolvidos"

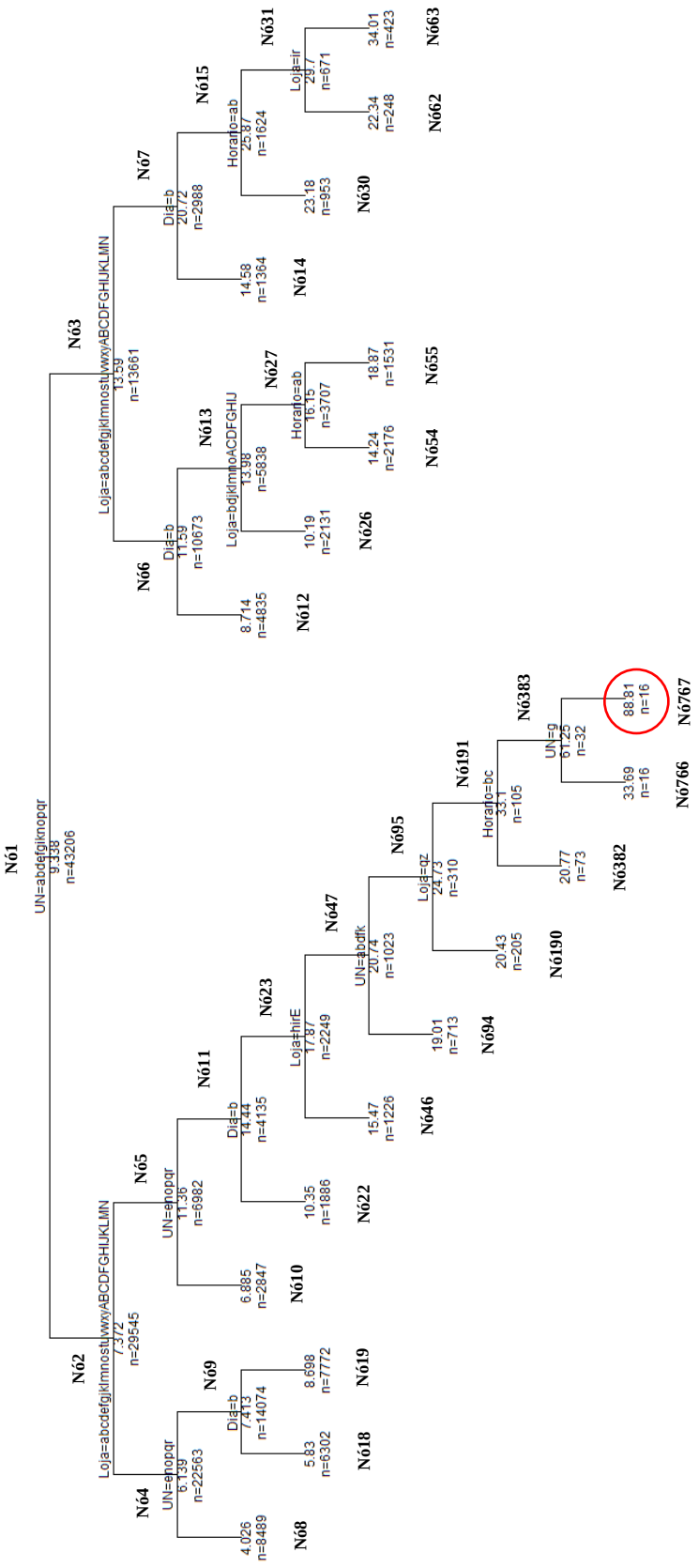


Figura 3.19 – Regras decisão da árvore variável "Total artigos devolvidos"

```
> fit
n= 43206

node), split, n, deviance, yval
      * denotes terminal node

  1) root 43206 9201010.00  9.337569
    2) UN=Bebidas,Bric&Auto,Char&Que,Congelados,Cultura,Frut&Leg,Lacticínios,Limpeza,Padaria,Peixaria,Pets&Plants,TakeAway,Talho 29545 3525370.00  7.37
      4) Loja=A,AA,AB,AC,AD,AE,AF,AI,AJ,AK,AL,AM,AN,E,F,G,H,I,J,K,M,N,O,P,R,S,T,U,V,W,X,Y,Z 22563 1400686.00  6.138856
        8) UN=Congelados,Padaria,Peixaria,Pets&Plants,TakeAway,Talho 8489  222951.40  4.025798 *
        9) UN=Bebidas,Bric&Auto,Char&Que,Cultura,Frut&Leg,Lacticínios,Limpeza 14074 1116969.00  7.413386
         18) Dia=Fds ou feriado 6302  265074.30  5.829737 *
         19) Dia=Dias úteis 7772  823273.80  8.697504 *
      5) Loja=AG,AH,B,C,D,L,Q 6982 1979454.00 11.357780
       10) UN=Congelados,Padaria,Peixaria,Pets&Plants,TakeAway,Talho 2847  281338.70  6.885494 *
       11) UN=Bebidas,Bric&Auto,Char&Que,Cultura,Frut&Leg,Lacticínios,Limpeza 4135 1601965.00 14.437000
         22) Dia=Fds ou feriado 1886  261460.40 10.348890 *
         23) Dia=Dias úteis 2249 1282552.00 17.865270
           46) Loja=AG,AH,D,Q 1226  570995.20 15.467370 *
           47) Loja=B,C,L 1023  696059.30 20.739000
             94) UN=Bebidas,Bric&Auto,Char&Que,Cultura,Limpeza 713  323440.00 19.005610 *
             95) UN=Frut&Leg,Lacticínios 310  365549.70 24.725810
               190) Loja=C,L 205  100228.40 20.434150 *
               191) Loja=B 105  254173.80 33.104760
                 382) Horario=Noite,Tarde 73  27327.04 20.767120 *
                 383) Horario=Manhã 32  190386.00 61.250000
                   766) UN=Frut&Leg 16  12311.44 33.687500 *
                   767) UN=Lacticínios 16  153764.40 88.812500 *
    3) UN=Casa,Higiene,Lazer,Merc Doc,Merc Sal 13661 5314693.00 13.588170
      6) Loja=A,AA,AB,AC,AD,AE,AF,AI,AJ,AK,AL,AM,AN,E,F,G,H,I,J,K,M,N,O,P,R,S,T,U,V,W,X,Y,Z 10673 2977266.00 11.591870
       12) Dia=Fds ou feriado 4835  515053.80  8.713754 *
       13) Dia=Dias úteis 5838 2388991.00 13.975510
         26) Loja=AA,AC,AI,AJ,AK,AL,AM,AN,M,O,P,R,S,T,U,V 2131  272956.40 10.185830 *
         27) Loja=A,AB,AD,AE,AF,E,F,G,H,I,J,K,N,W,X,Y,Z 3707 2067837.00 16.154030
           54) Horario=Manhã,Noite 2176 1158722.00 14.241270 *
           55) Horario=Tarde 1531  889838.20 18.872630 *
      7) Loja=AG,AH,B,C,D,L,Q 2988 2142962.00 20.718880
       14) Dia=Fds ou feriado 1364  351094.10 14.580650 *
       15) Dia=Dias úteis 1624 1697310.00 25.874380
         30) Horario=Manhã,Noite 953  993667.60 23.181530 *
         31) Horario=Tarde 671  686917.20 29.698960
           62) Loja=AH,D 248  160189.90 22.342740 *
           63) Loja=AG,B,C,L,Q 423  505438.90 34.011820 *
```

Em termos de texto, a árvore representa-se conforme a figura 3.19, permitindo uma leitura mais percetível dos nós e das folhas da árvore. Podemos assim transformar toda a árvore em regras de decisão (não é normal denominar estas por regras de regressão).

Na figura 3.19, a 1ª linha indica que o nº de observações usadas para aprendizagem foi de 43206. Abaixo cada linha está identificada com um número, que corresponde a uma identificação do nó. Em cada uma dessas linhas é indicado o nº de casos, o valor do desvio e o valor médio da variável objetivo.

Por exemplo no nó número 767) são classificados 16 casos, que seguem a regra – (Dia="Dias úteis" & Loja="B" & Horario registo="Manhã" & Unidade Negocio="Lacticinios"), prevendo-se um número de artigos devolvidos de 88,8 e com um desvio de 98.

Pretende-se agora estudar a árvore, e em cada nó ir realizando análises do tipo *box-plot* e identificar a presença de novos *outliers* extremos em relação às análises univariada e bivariada. O processo deve ser repetido até concluída a análise a todos os nós e folhas da árvore. O topo da árvore corresponde ao total das observações, pelo que é igual à análise univariada já realizada.

A 1ª divisão da árvore acontece com a variável "Unidade Negócio", separando a árvore em dois nós (2 e 3):

```
2) UN=Bebidas,Bric&Auto,Char&Que,Congelados,Cultura,Frut&Leg,Lacticínios,Limpeza,Padaria,Peixaria,Pets&Plants,TakeAway,Talho
3) UN=Casa,Higiene,Lazer,Merc Doc,Merc Sal
```

De seguida são apresentados os gráficos *box-plot* referentes às 36 regras inerentes a cada nó/folha da árvore. Neste capítulo apenas serão ilustrados os gráficos onde são observados novos *outliers* extremos que não tenham sido identificados nas análises univariadas e bivariadas. Os restantes gráficos podem ser consultados no Anexo III.

Regras de Decisão do Nó 12

SE Unidade Negócio = Casa, Higiene, Lazer, Mercearia Doce, Mercearia Salgada

E Loja = A, AA, AB, AC, AD, AE, AF, AI, AJ, AK, AL, AM, AN, E, F, G, H, I, J, K, M, N, O, P, R, S, T, U, V, W, X, Y, Z

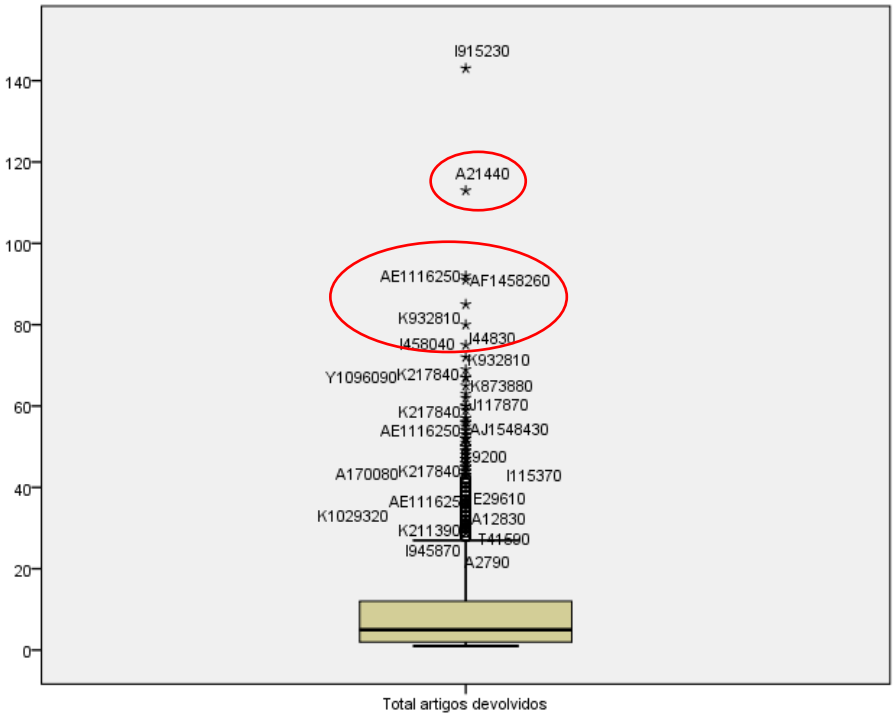
E Dia semana = Fim de semana ou feriado

ENTÃO Total artigos devolvidos **=** 8,71

N = 4835

Desvio = 515053,8

Figura 3.20 – *Box-Plot* do Nó 12, variável "Total artigos devolvidos"



Conclusão: observam-se **seis novos *outliers* extremos** mais significativos, que não foram detetados em análises anteriores. Aplicando a regra descrita acima, o total artigos devolvidos é de cerca de 8.71, no entanto nos casos identificados, as devoluções efetuadas são superiores a 75 unidades. Os valores são superiores em quase 9 vezes o valor da média.

Regras de Decisão do Nó 14

SE Unidade Negócio = Casa, Higiene, Lazer, Mercearia Doce, Mercearia Salgada

E Loja = AG, AH, B, C, D, L, Q

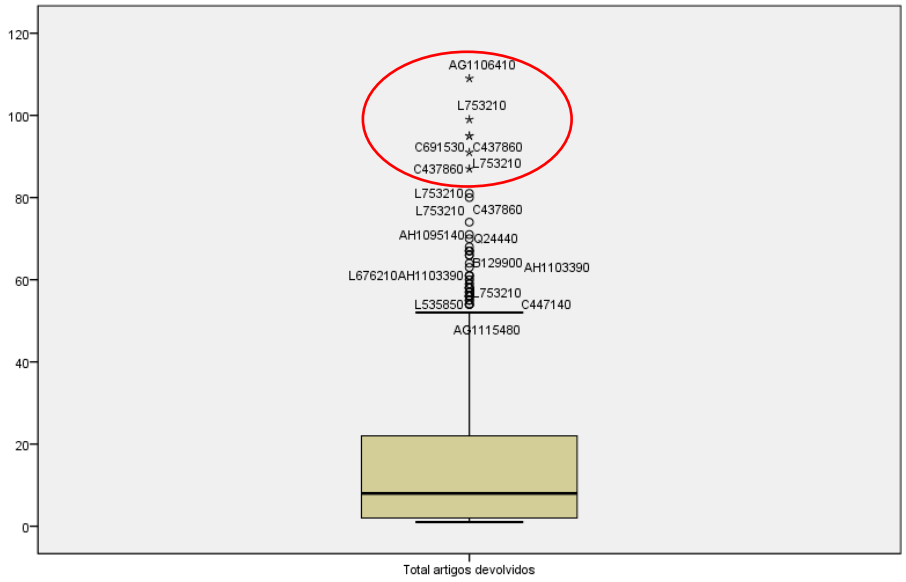
E Dia semana = Fim de semana ou feriado

ENTÃO Total artigos devolvidos **=** 14,58

N = 1364

Desvio = 351094,1

Figura 3.21 – *Box-Plot* do Nó 14, variável "Total artigos devolvidos"



Conclusão: observam-se **seis novos *outliers* extremos** mais significativos, que não foram detetados em análises anteriores. Aplicando a regra descrita acima, o total artigos devolvidos é de cerca de 14.58, no entanto nos casos identificados, as devoluções efetuadas são superiores a 85 unidades. Os valores são superiores em quase 6 vezes o valor da média.

Regras de Decisão do Nó 18

SE Loja = A, AA, AB, AC, AD, AE, AF, AI, AJ, AK, AL, AM, AN, E, F, G, H, I, J, K, M, N, O, P, R, S, T, U, V, W, X, Y, Z

E Unidade Negocio = Bebidas, Bricolage&Auto, Charcutaria&Queijos, Cultura, Frutas&Legumes, Lacticínios, Limpeza

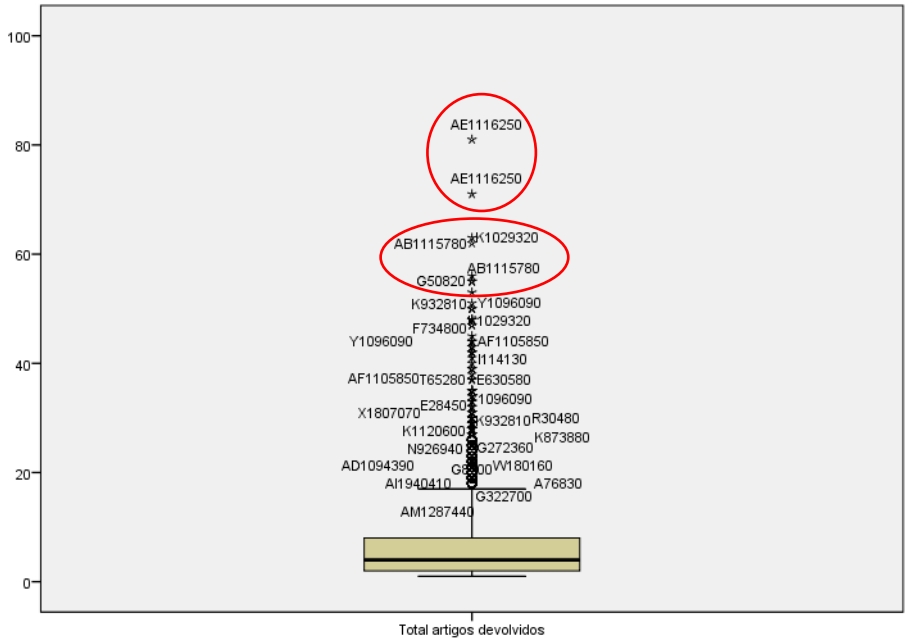
E Dia semana = Fim de semana ou feriado

ENTÃO Total artigos devolvidos **=** 5,83

N = 6302

Desvio = 265074,3

Figura 3.22 – *Box-Plot* do Nó 18, variável "Total artigos devolvidos"



Conclusão: observam-se **seis novos *outliers* extremos** mais significativos, que não foram detetados em análises anteriores. Aplicando a regra descrita acima, o total artigos devolvidos é de cerca de 5.83, no entanto nos casos identificados, as devoluções efetuadas são superiores ou iguais a 55 unidades. Estes valores são superiores em cerca de 9 vezes o valor da média.

Regras de Decisão do Nó 22

SE Loja = AG, AH, B, C, D, L, Q

E Unidade Negocio = Bebidas, Bricolage&Auto, Charcutaria&Queijos, Cultura, Frutas&Legumes, Lacticínios, Limpeza

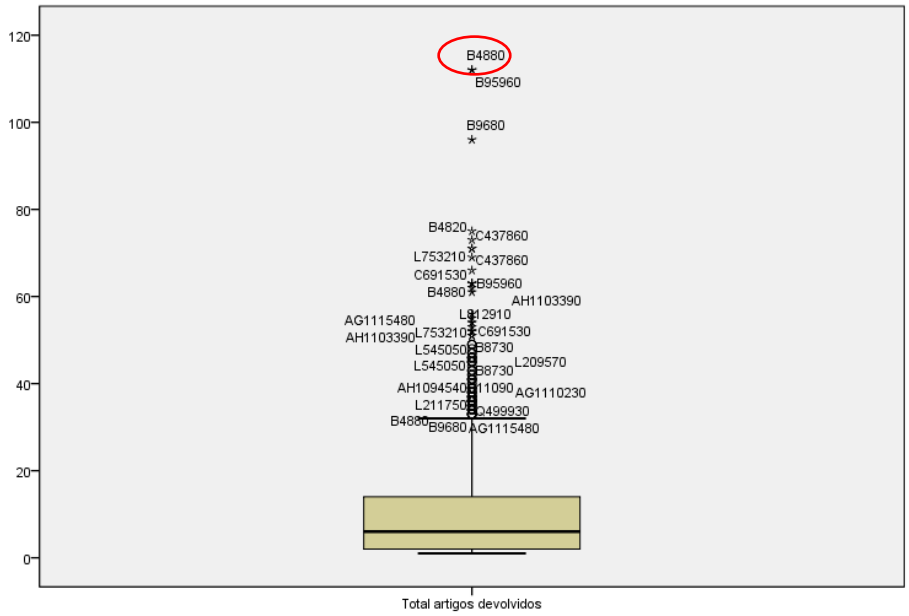
E Dia semana = Fim de semana ou feriado

ENTÃO Total artigos devolvidos **=** 10,35

N = 1886

Desvio = 261460,4

Figura 3.23 – *Box-Plot* do Nó 22, variável "Total artigos devolvidos"



Conclusão: observa-se **um novo *outlier* extremo** mais significativo, que não foi detetado em análises anteriores, no supervisor B4880. Aplicando a regra descrita acima, o total artigos devolvidos é de cerca de 10.35, no entanto no caso identificado, as devoluções efetuadas são de 112 unidades. Este valor é superior em quase de 11 vezes o valor da média.

Regras de Decisão do Nó 26

SE Unidade Negócio = Casa, Higiene, Lazer, Mercearia Doce, Mercearia Salgada

E Dia Semana = Dias úteis

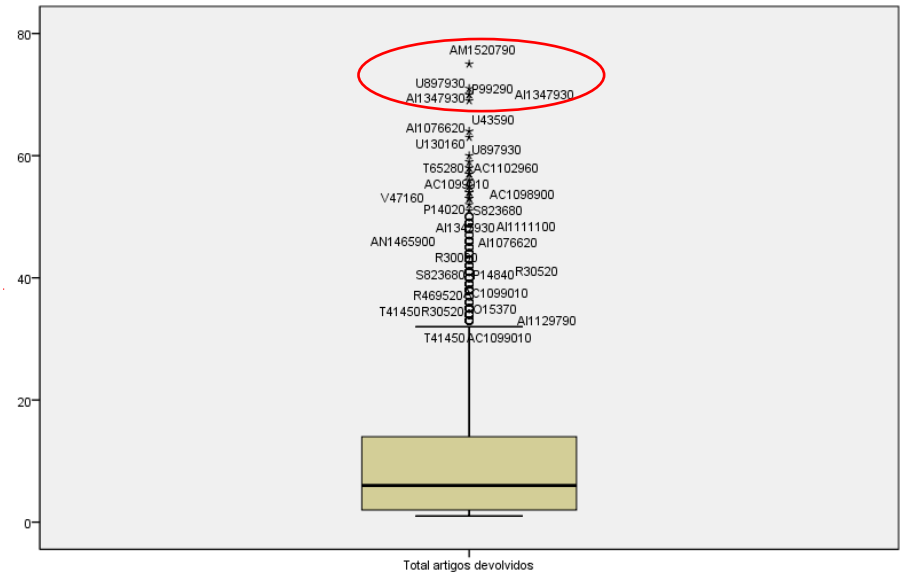
E Loja = AA, AC, AI, AJ, AK, AL, AM, AN, M, O, P, R, S, T, U, V

ENTÃO Total artigos devolvidos **=** 10,19

N = 2131

Desvio = 272956,4

Figura 3.24 – *Box-Plot* do Nó 26, variável "Total artigos devolvidos"



Conclusão: observam-se **cinco novos *outliers* extremos** mais significativos, que não foram detetados em análises anteriores. Aplicando a regra descrita acima, o total artigos devolvidos é de cerca de 10.19, no entanto nos casos identificados, as devoluções efetuadas são superiores ou iguais a 65 unidades. Estes valores são superiores em cerca de 6 vezes o valor da média.

Regras de Decisão do Nó 46

SE Unidade Negócio = Bebidas, Bricolage&Auto, Charcutaria&Queijos, Cultura, Frutas&Legumes, Lacticinios, Limpeza

E Dia Semana = Dias úteis

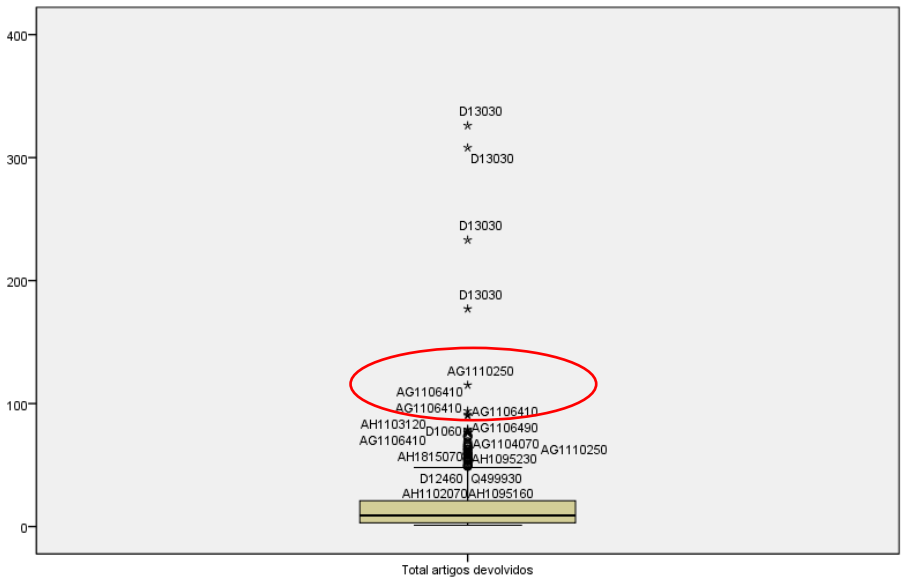
E Loja = AG, AH, D, Q

ENTÃO Total artigos devolvidos **=** 15,47

N = 1226

Desvio = 570995,2

Figura 3.25 – *Box-Plot* do Nó 46, variável "Total artigos devolvidos"



Conclusão: observam-se **quatro novos *outliers* extremos** mais significativos, que não foram detetados em análises anteriores. Aplicando a regra descrita acima, o total artigos devolvidos é de cerca de 15.47, no entanto nos casos identificados, as devoluções efetuadas são superiores ou iguais a 90 unidades. Estes valores são superiores em quase 6 vezes o valor da média.

Regras de Decisão do Nó 55

SE Unidade Negócio = Casa, Higiene, Lazer, Mercearia Doce, Mercearia Salgada

E Dia Semana = Dias úteis

E Loja = A, AB, AD, AE, AF, E, F, G, H, I, J, K, N, W, X, Y, Z

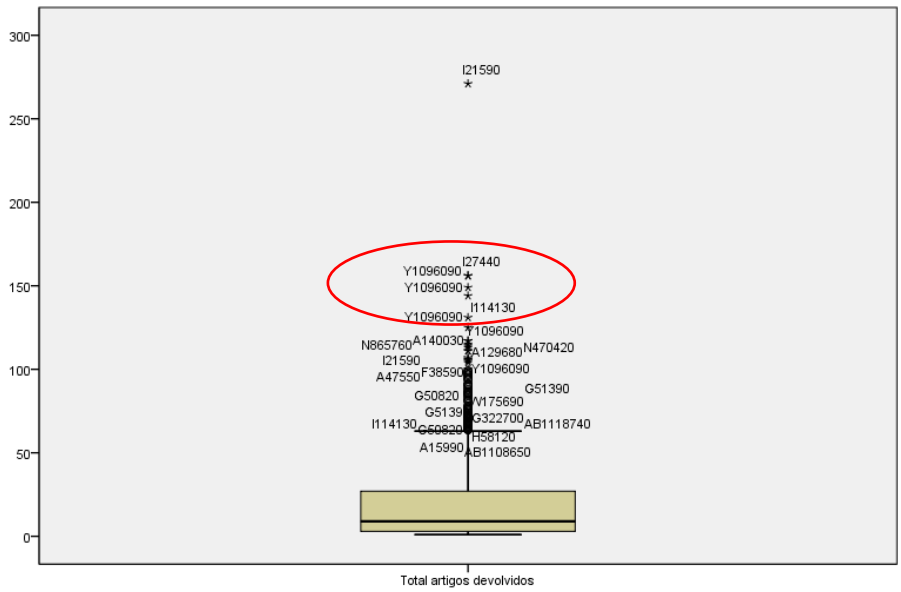
E Horario registo = Tarde

ENTÃO Total artigos devolvidos **=** 18,87

N = 1531

Desvio = 889838,2

Figura 3.26 – *Box-Plot* do Nó 55, variável "Total artigos devolvidos"



Conclusão: observam-se **cinco novos *outliers* extremos** mais significativos, que não foram detetados em análises anteriores. Aplicando a regra descrita acima, o total artigos devolvidos é de cerca de 18.87, no entanto nos casos identificados, as devoluções efetuadas são superiores a 130 unidades. Estes valores são superiores em quase 7 vezes o valor da média.

Regras de Decisão do Nó 62

E Unidade Negocio = Casa, Higiene, Lazer, Mercearia Doce, Mercearia Salgada

E Dia semana = Dias úteis

E Horario registo = Tarde

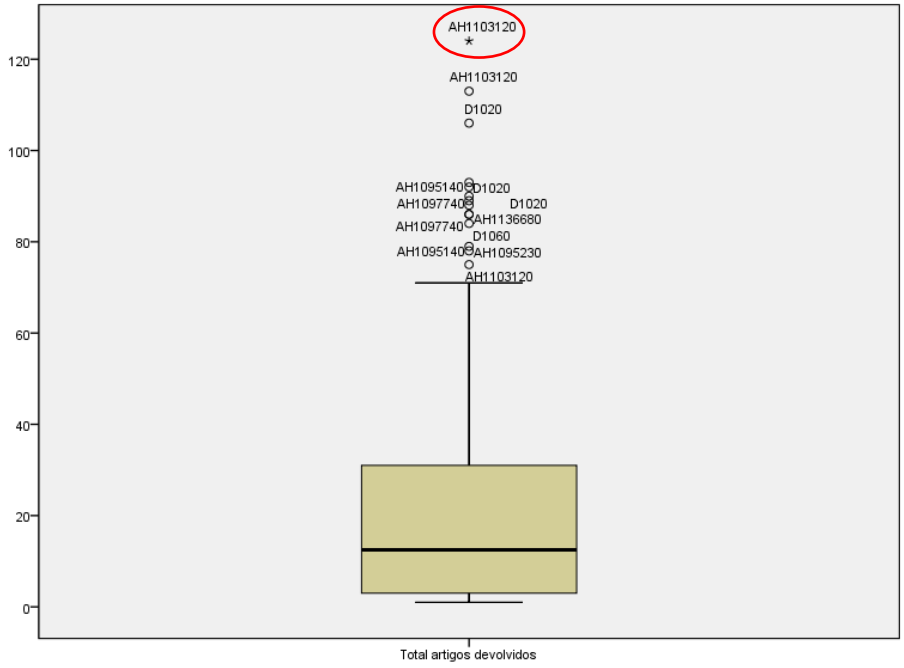
E Loja = AH, D

ENTÃO Total artigos devolvidos **=** 22,34

N = 248

Desvio = 160189,9

Figura 3.27 – *Box-Plot* do Nó 62, variável "Total artigos devolvidos"



Conclusão: observa-se um **novo *outlier* extremo**, que não foi detetado em análises anteriores, no supervisor AH1103120. Aplicando a regra descrita acima, o total artigos devolvidos é de cerca de 22.34, no entanto no caso identificado, as devoluções efetuadas são de 124 unidades. Este valor é superior em cerca de 5,5 vezes o valor da média.

3.6.2 Árvore de regressão: variável objetivo "Valor médio artigos devolvidos"

De forma idêntica à outra variável, a árvore da figura 3.28 é o resultado final de testes realizados na função *rpart*, alterando o parâmetro "cp", de forma a obter folhas da árvore com valores o mais elevado possível. Neste caso o valor "otimizado" do "cp" foi de 0.0013, obtendo-se uma folha da árvore com 10 observações e o valor médio da variável objetivo de 92,02€.

Figura 3.28 – Árvore regressão variável objetivo "Valor médio artigos devolvidos"

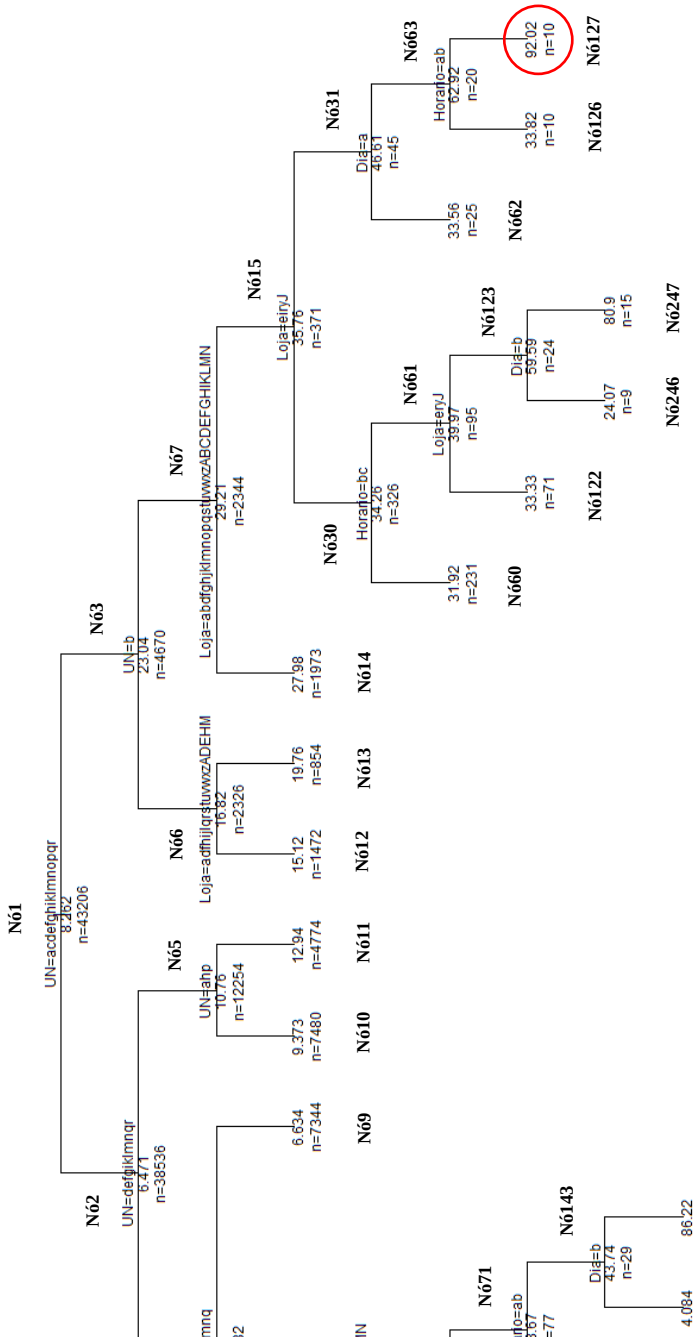


Figura 3.29 – Regras decisão da árvore variável "Valor médio artigos devolvidos"

```
> fit
n= 43206

node), split, n, deviance, yval
      * denotes terminal node

  1) root 43206 8548243.0000  8.262060
    2) UN=Bebidas,Casa,Char&Que,Congelados,Cultura,Frut&Leg,Higiene,Lacticínios,Limpeza,Merc Doc,Merc Sal,Padaria,Peixaria,Pets&Plants,TakeAway,Talho 38536 5179937.0000  6.47
      4) UN=Char&Que,Congelados,Cultura,Frut&Leg,Lacticínios,Limpeza,Merc Doc,Merc Sal,Padaria,TakeAway,Talho 26282 1993308.0000  4.469757
        8) UN=Char&Que,Congelados,Frut&Leg,Lacticínios,Merc Doc,Merc Sal,Padaria,TakeAway 18938 1720307.0000  3.630573
         16) UN=Frut&Leg,Padaria 4937  24465.8400  2.547387 *
         17) UN=Char&Que,Congelados,Lacticínios,Merc Doc,Merc Sal,TakeAway 14001 1688006.0000  4.012523
           34) Loja=A,AA,AB,AC,AD,AE,AF,AG,AH,AI,AJ,AK,AL,AM,AN,B,C,D,E,F,G,H,I,J,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z 13608  332110.1000  3.941199 *
           35) Loja=K 393 1353429.0000  6.482188
             70) UN=Char&Que,Congelados,Merc Doc,Merc Sal,TakeAway 316  2302.8220  3.511709 *
             71) UN=Lacticínios 77 1336895.0000 18.672730
               142) Horario=Manhã,Noite 48  249.3271  3.530208 *
               143) Horario=Tarde 29 1307423.0000 43.736210
                 286) Dia=Fds ou feriado 15  326.1760  4.084000 *
                 287) Dia=Dias úteis 14 1258243.0000 86.220710 *
        9) UN=Cultura,Limpeza,Talho 7344  225273.6000  6.633765 *
      5) UN=Bebidas,Casa,Higiene,Peixaria,Pets&Plants 12254 2855521.0000 10.764100
       10) UN=Bebidas,Higiene,Pets&Plants 7480 1105080.0000  9.372743 *
       11) UN=Casa,Peixaria 4774 1713273.0000 12.944100 *
    3) UN=Bric&Auto,Lazer 4670 2224964.0000 23.039220
      6) UN=Bric&Auto 2326  780794.7000 16.819270
       12) Loja=A,AC,AE,AG,AH,AI,AK,C,D,E,F,G,H,I,J,L,M,P,Q,T,Y 1472  355707.1000 15.115920 *
       13) Loja=AA,AB,AD,AF,AJ,AL,AM,AN,B,K,N,O,R,S,U,V,W,X,Z 854  413455.3000 19.755260 *
      7) UN=Lazer 2344 1264886.0000 29.211400
       14) Loja=A,AA,AC,AE,AF,AG,AI,AJ,AK,AL,AM,AN,B,C,E,F,G,H,I,J,L,M,N,O,P,Q,R,S,T,U,W,X,Y,Z 1973  486453.0000 27.980010 *
       15) Loja=AB,AD,AH,D,K,V 371  759530.8000 35.760000
         30) Loja=AD,AH,D,K,V 326  424464.8000 34.262210
           60) Horario=Noite,Tarde 231  111186.3000 31.916280 *
           61) Horario=Manhã 95  308916.0000 39.966530
           122) Loja=AD,D,K,V 71  38335.1800 33.334230 *
           123) Loja=AH 24  258218.5000 59.587080
             246) Dia=Fds ou feriado 9  658.4663 24.068890 *
             247) Dia=Dias úteis 15  239393.9000 80.898000 *
         31) Loja=AB 45  329036.5000 46.610670
           62) Dia=Dias úteis 25  8734.8460 33.564000 *
           63) Dia=Fds ou feriado 20  310727.0000 62.919000
             126) Horario=Manhã,Noite 10  3445.5150 33.818000 *
             127) Horario=Tarde 10  290344.1000 92.020000 *
```

Na figura 3.29 a árvore está representada em termos de texto, com indicação das regras de decisão e valores dos nós e folhas da árvore.

Tal como na outra variável objetivo, pretende-se estudar a árvore, e em cada nó ir realizando análises do tipo *box-plot* e identificar a presença de novos *outliers* extremos em relação às análises univariada e bivariada. O processo é repetido até concluída a análise a todos os nós e folhas da árvore. O topo da árvore corresponde ao total das observações, pelo que é igual à análise univariada já realizada.

A 1ª divisão da árvore acontece com a variável "Unidade Negócio", separando a árvore em dois nós (2 e 3):

```
2) UN=Bebidas,Casa,Char&Que,Congelados,Cultura,Frut&Leg,Higiene,Lacticínios,Limpeza,Merc Doc,Merc Sal,Padaria,Peixaria,Pets&Plants,TakeAway,Talho
3) UN=Bric&Auto,Lazer
```

De seguida são apresentados os gráficos *box-plot* referentes às 36 regras inerentes a cada nó/folha da árvore. Neste capítulo apenas serão ilustrados os gráficos onde são observados novos *outliers* extremos e que não tenham sido identificados nas análises univariadas e bivariadas. Os restantes gráficos podem ser consultados no Anexo IV.

Regras de Decisão do Nó 5

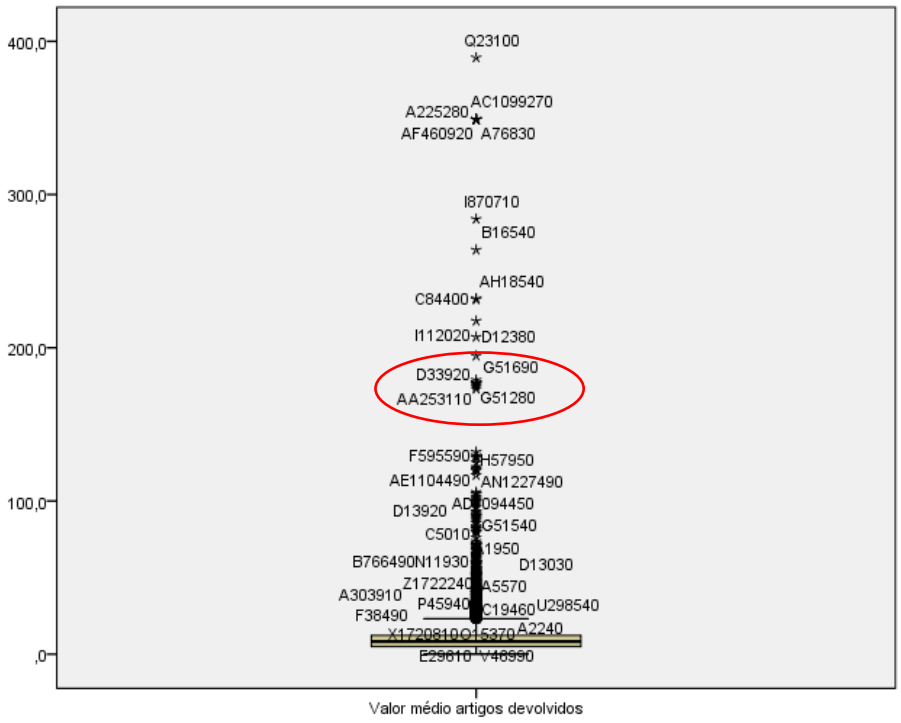
SE Unidade Negócio = Bebidas, Casa, Higiene, Peixaria, Pets&Plants

ENTÃO Valor médio artigos devolvidos = 10,76€

N = 12254

Desvio = 2855521

Figura 3.30 – *Box-Plot* do Nó 5, variável "Valor médio artigos devolvidos"



Conclusão: observam-se **quatro novos *outliers* extremos** mais significativos, que não foram detetados em análises anteriores. Nos produtos de Bebidas, Casa, Higiene, Peixaria ou Pets&Plants, o valor médio dos artigos devolvidos é de cerca de 10,76€, no entanto nos casos identificados, as devoluções efetuadas são superiores a 170€. Estes valores são superiores em cerca de 16 vezes o valor da média.

Regras de Decisão do Nó 6

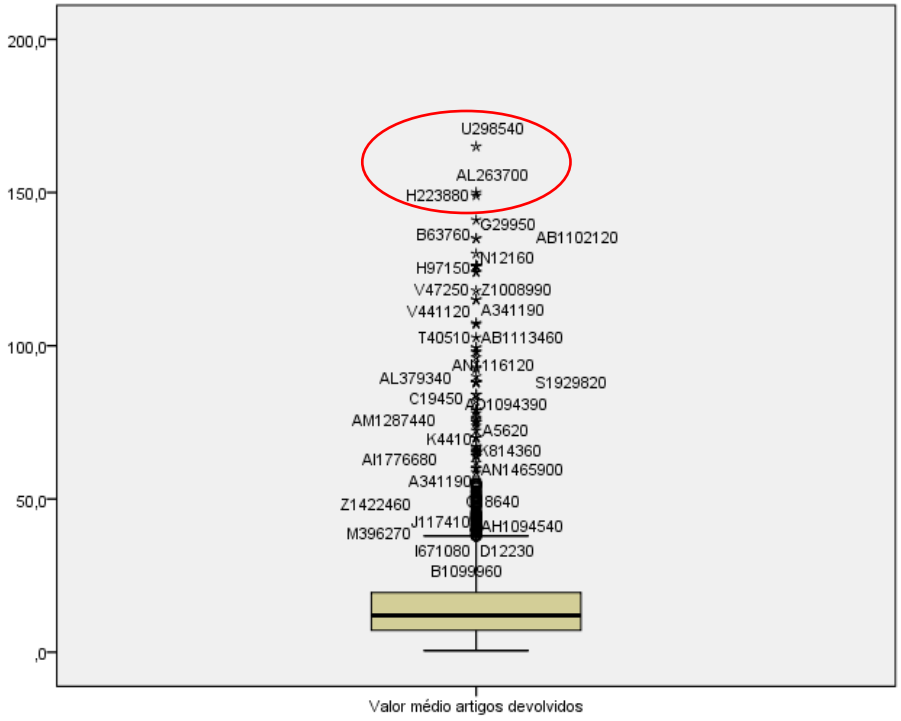
SE Unidade Negócio = Bricolage&Auto

ENTÃO Valor médio artigos devolvidos = 16,82€

N = 2326

Desvio = 780794,7

Figura 3.31 – *Box-Plot* do Nó 6, variável "Valor médio artigos devolvidos"



Conclusão: observam-se **<u>três novos *outliers* extremos</u>** mais significativos, que não foram detetados em análises anteriores. Nos produtos de Bricolage&Auto, o valor médio dos artigos devolvidos é de cerca de 16,82€, no entanto nos casos identificados, as devoluções efetuadas são superiores a 145€. Estes valores são superiores em cerca de 8 vezes o valor da média.

Regras de Decisão do Nó 9

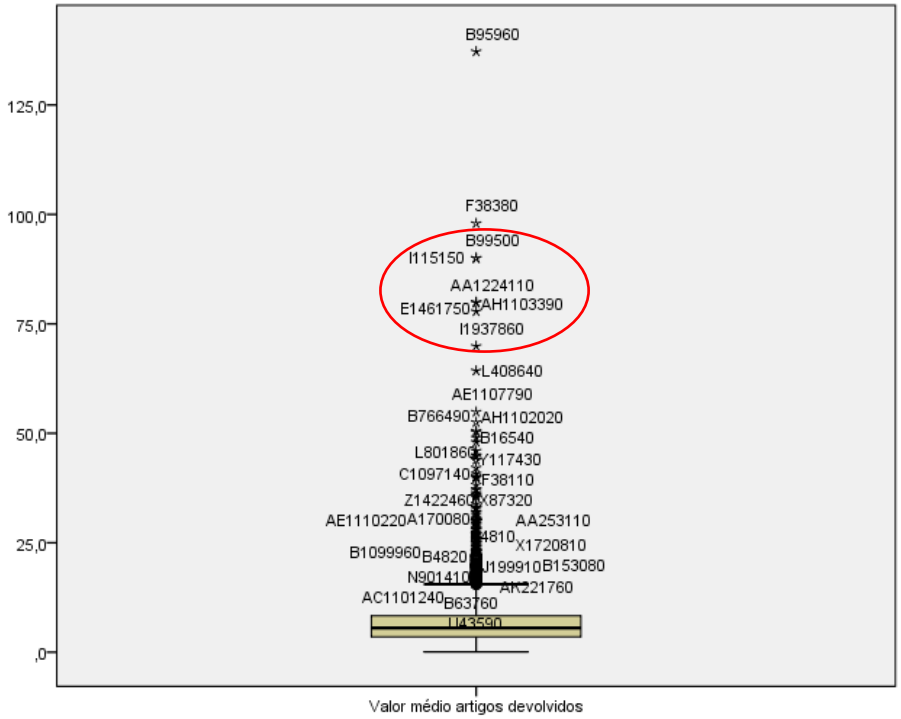
SE Unidade Negócio = Cultura, Limpeza, Talho

ENTÃO Valor médio artigos devolvidos = 6,63€

N = 7344

Desvio = 225273,6

Figura 3.32 – *Box-Plot* do Nó 9, variável "Valor médio artigos devolvidos"



Conclusão: observam-se **seis novos *outliers* extremos** mais significativos, que não foram detetados em análises anteriores. Nos produtos de Cultura, Limpeza ou Talho, o valor médio dos artigos devolvidos é de cerca de 6,63€, no entanto nos casos identificados, as devoluções efetuadas são superiores a 69€. Este valor é superior em cerca de 10 vezes o valor da média.

Regras de Decisão do Nó 10

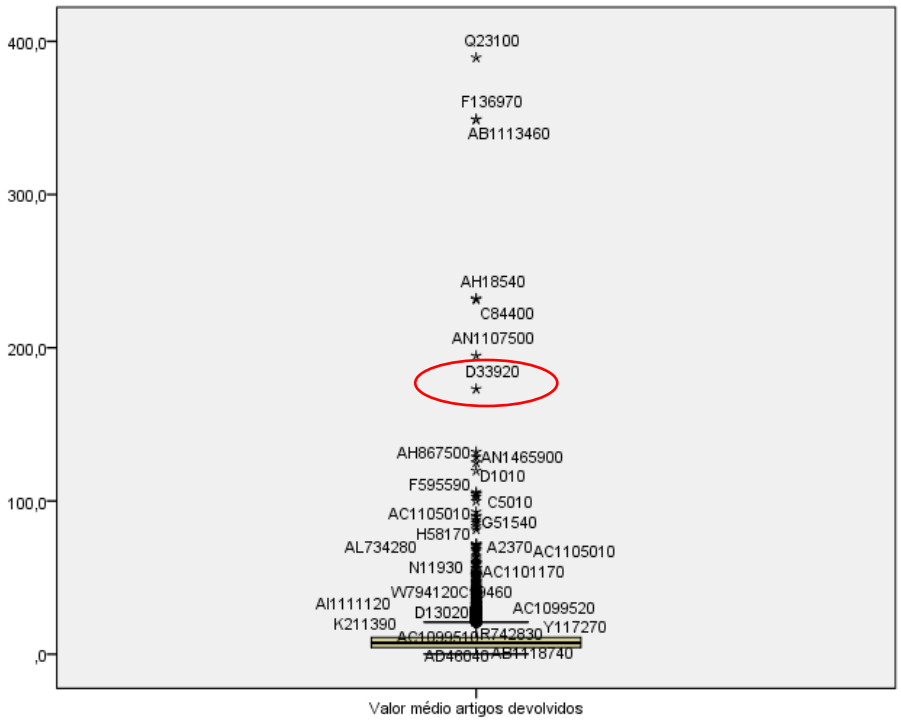
SE Unidade Negócio = Bebidas, Higiene, Pets&Plants

ENTÃO Valor médio artigos devolvidos = 9,37€

N = 7480

Desvio = 1105080

Figura 3.33 – *Box-Plot* do Nó 10, variável "Valor médio artigos devolvidos"



Conclusão: observa-se **um novo *outlier* extremo** mais significativo, que não foi detetado em análises anteriores. Nos produtos de Bebidas, Higiene ou Pets&Plants, o valor médio dos artigos devolvidos é de cerca de 9,37€, no entanto no caso identificado, as devoluções efetuadas é igual a 173€. Este valor é superior em cerca de 18 vezes o valor da média.

Regras de Decisão do Nó 11

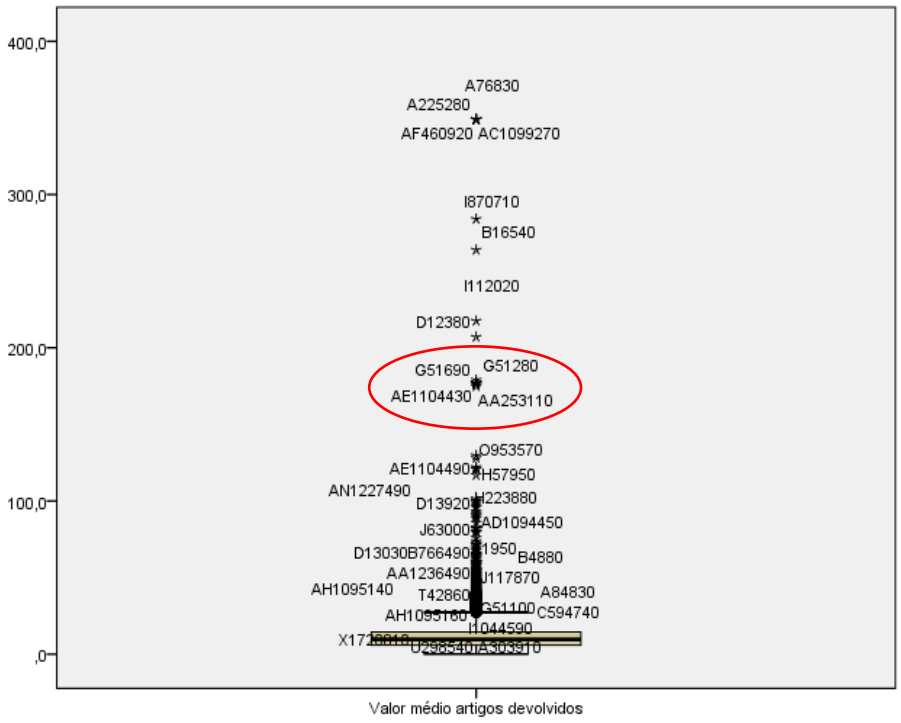
SE Unidade Negócio = Casa, Peixaria

ENTÃO Valor médio artigos devolvidos = 12,94€

N = 4774

Desvio = 1713273

Figura 3.34 – *Box-Plot* do Nó 11, variável "Valor médio artigos devolvidos"



Conclusão: observam-se **quatro novos *outliers* extremos** mais significativos, que não foram detetados em análises anteriores. Nos produtos de Casa ou Peixaria, o valor médio dos artigos devolvidos é de cerca de 12,94€, no entanto nos casos identificados, as devoluções efetuadas são superiores a 170€. Este valor é superior em cerca de 13 vezes o valor da média.

Regras de Decisão do Nó 12

SE Unidade Negócio = Bricolage&Auto

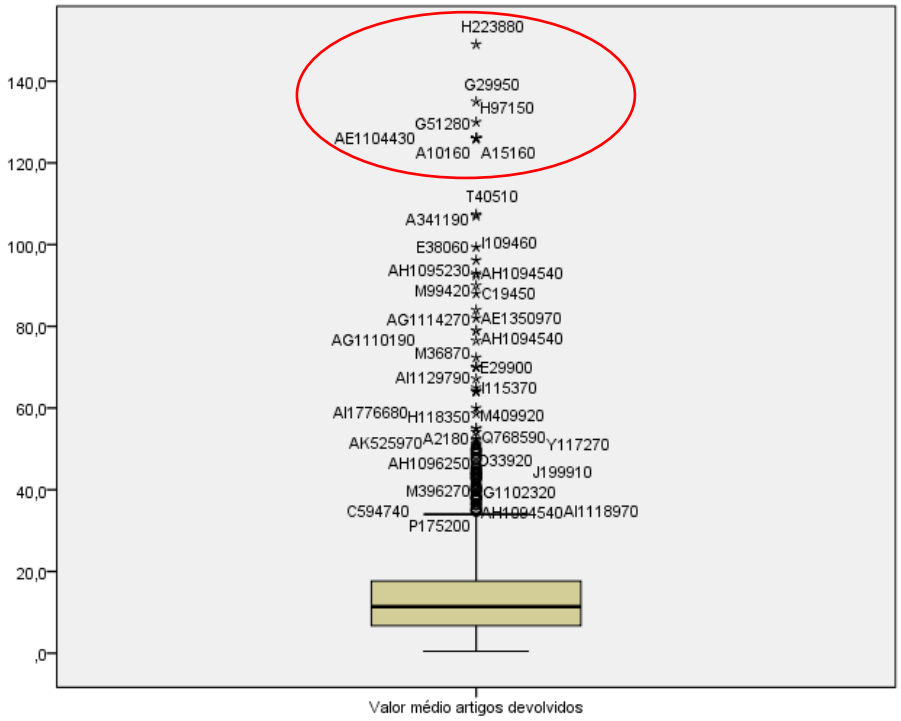
E Loja = A, AC, AE, AG, AH, AI, AK, C, D, E, F, G, H, I, J, L, M, P, Q, T, Y

ENTÃO Valor médio artigos devolvidos = 15,11€

N = 1472

Desvio = 355707,1

Figura 3.35 – *Box-Plot* do Nó 12, variável "Valor médio artigos devolvidos"



Conclusão: observam-se **sete novos *outliers* extremos** mais significativos, que não foram detetados em análises anteriores. Aplicando a regra acima descrita, o valor médio dos artigos devolvidos é de cerca de 15,11€, no entanto nos casos identificados, as devoluções efetuadas são superiores a 125€. Este valor é superior em cerca de 8 vezes o valor da média.

Regras de Decisão do Nó 13

SE Unidade Negócio = Bricolage&Auto

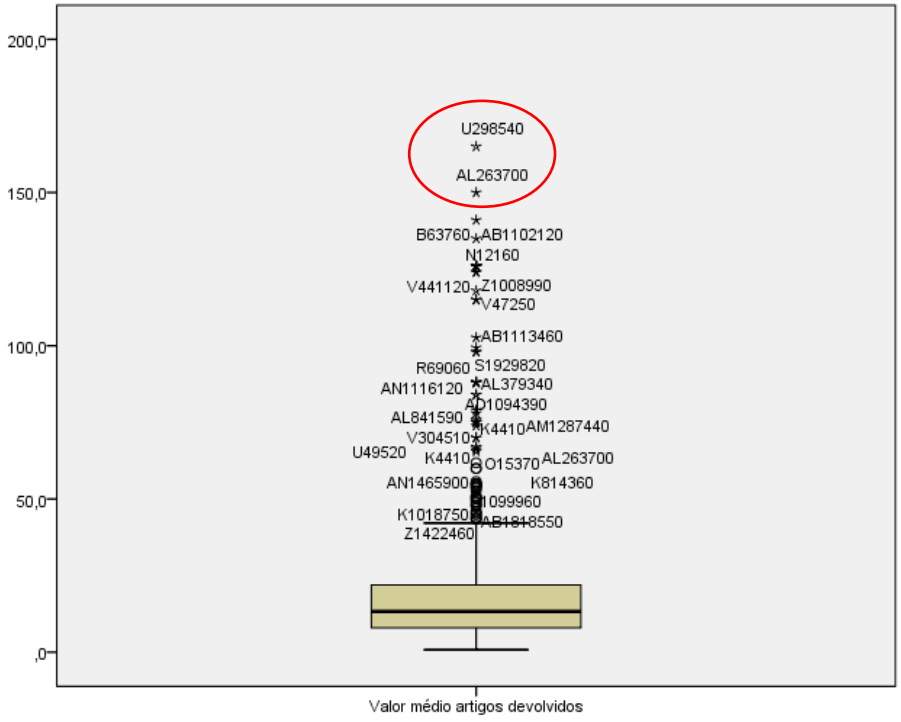
E Loja = AA, AB, AD, AF, AJ, AL, AM, AN, B, K, N, O, R, S, U, V, W, X, Z

ENTÃO Valor médio artigos devolvidos = 19,76€

N = 854

Desvio = 413455,3

Figura 3.36 – *Box-Plot* do Nó 13, variável "Valor médio artigos devolvidos"



Conclusão: observam-se **dois novos *outliers* extremos** mais significativos, que não foram detetados em análises anteriores. Aplicando a regra acima descrita, o valor médio dos artigos devolvidos é de cerca de 19,76€, no entanto nos casos identificados, as devoluções efetuadas são superiores a 145€. Este valor é superior em cerca de 7 vezes o valor da média.

Regras de Decisão do Nó 14

SE Unidade Negócio = Lazer

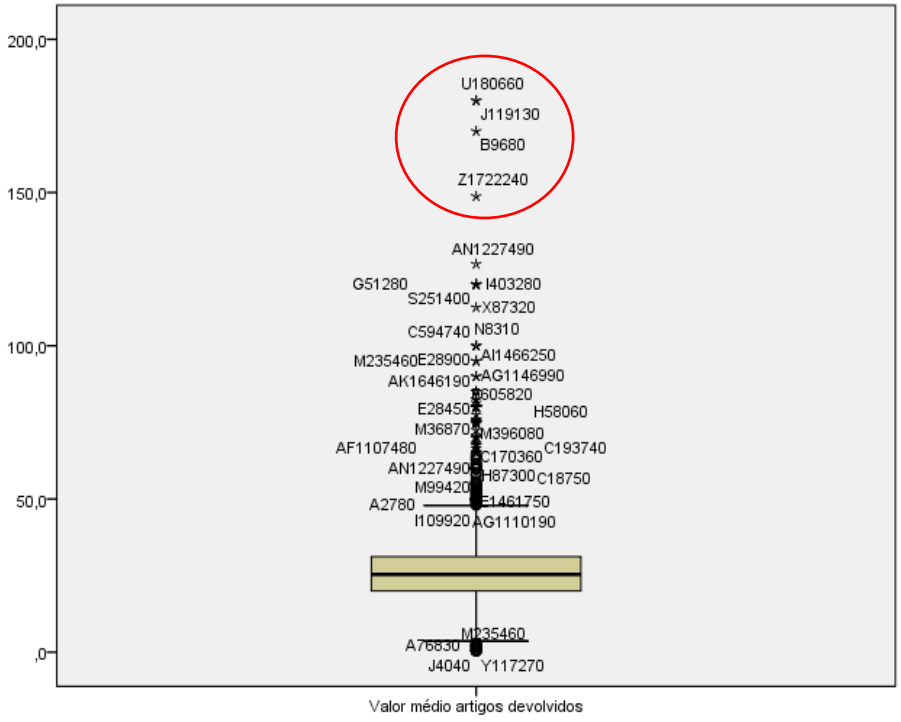
E Loja = A, AA, AC, AE, AF, AG, AI, AJ, AK, AL, AM, AN, B, C, E, F, G, H, I,J, L, M, N, O, P, Q, R, S, T, U, W, X, Y, Z

ENTÃO Valor médio artigos devolvidos **=** 27,98€

N = 1973

Desvio = 486453

Figura 3.37 – *Box-Plot* do Nó 14, variável "Valor médio artigos devolvidos"



Conclusão: observam-se **quatro novos *outliers* extremos** mais significativos, que não foram detetados em análises anteriores. Aplicando a regra acima descrita, o valor médio dos artigos devolvidos é de cerca de 27,98€, no entanto nos casos identificados, as devoluções efetuadas são superiores a 145€. Este valor é superior em cerca de 5 vezes o valor da média.

Regras de Decisão do Nó 15

SE Unidade Negócio = Lazer

E Loja = AB, AD, AH, D, K, V

ENTÃO Valor médio artigos devolvidos = 35,76€

N = 371

Desvio = 759530,8

Figura 3.38 – *Box-Plot* do Nó 15, variável "Valor médio artigos devolvidos"

Conclusão: observam-se **<u>três novos *outliers* extremos</u>** mais significativos, que não foram detetados em análises anteriores. Aplicando a regra acima descrita, o valor médio dos artigos devolvidos é de cerca de 35,76€, no entanto nos casos identificados, as devoluções efetuadas são superiores a 145€. Este valor é superior em cerca de 4 vezes o valor da média.

Regras de Decisão do Nó 16

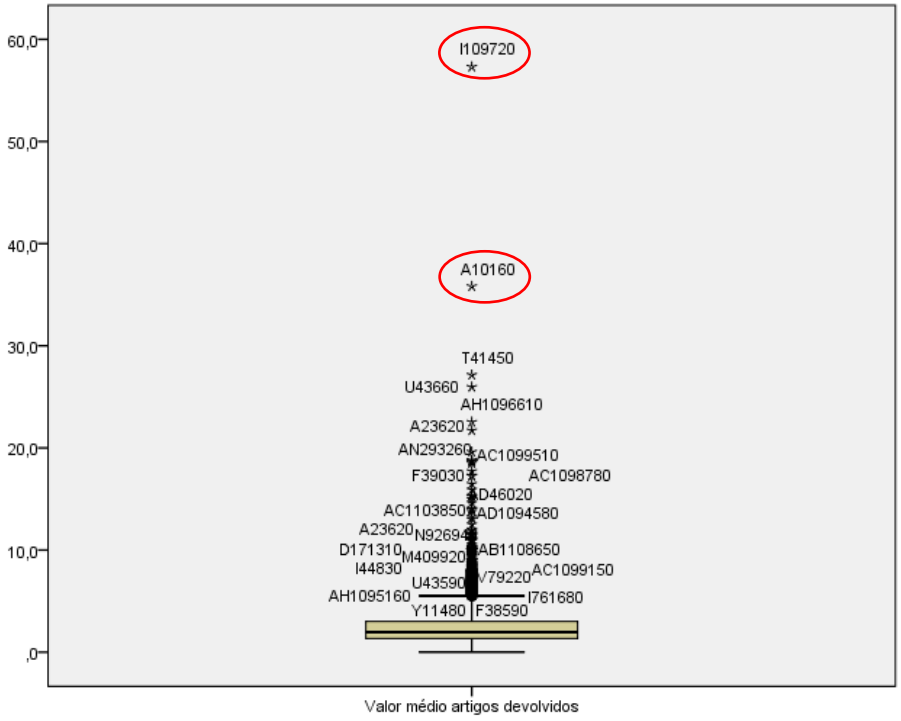
SE Unidade Negócio = Frutas&Legumes, Padaria

ENTÃO Valor médio artigos devolvidos = 2,55€

N = 4937

Desvio = 24465,84

Figura 3.39 – *Box-Plot* do Nó 16, variável "Valor médio artigos devolvidos"



Conclusão: observam-se **dois novos *outliers* extremos** mais significativos, que não foram detetados em análises anteriores. Nos produtos de "Frutas&Legumes" ou de "Padaria", o valor médio dos artigos devolvidos é de cerca de 2,55€, no entanto nos dois casos identificados, as devoluções efetuadas são superiores a 35€. Estes valores são superiores em cerca de 13,7 vezes o valor da média.

Regras de Decisão do Nó 30

SE Unidade Negócio = Lazer

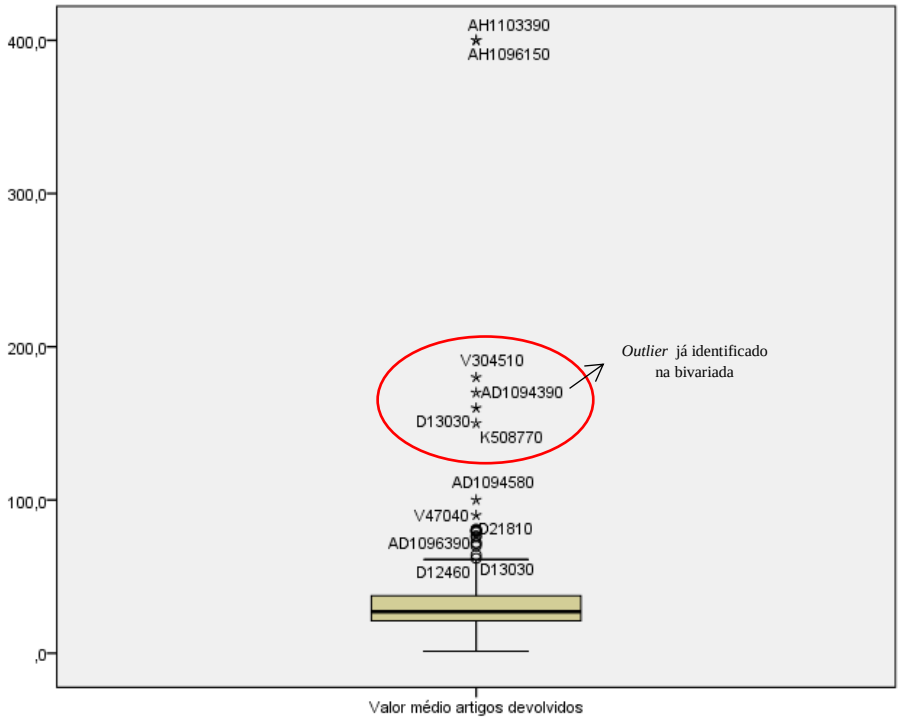
E Loja = AD, AH, D, K, V

ENTÃO Valor médio artigos devolvidos = 34,26€

N = 326

Desvio = 424464,8

Figura 3.40 – *Box-Plot* do Nó 30, variável "Valor médio artigos devolvidos"



Conclusão: observam-se **<u>três novos *outliers* extremos</u>** mais significativos, que não foram detetados em análises anteriores. Aplicando a regra acima descrita, o valor médio dos artigos devolvidos é de cerca de 34,26€, no entanto nos casos identificados, as devoluções efetuadas são superiores a 145€. Este valor é superior em cerca de 4 vezes o valor da média.

Regras de Decisão do Nó 34

SE Unidade Negócio = Charcutaria&Queijos, Congelados, Lacticínios, Mercearia Doce, Mercearia Salgada, TakeAway

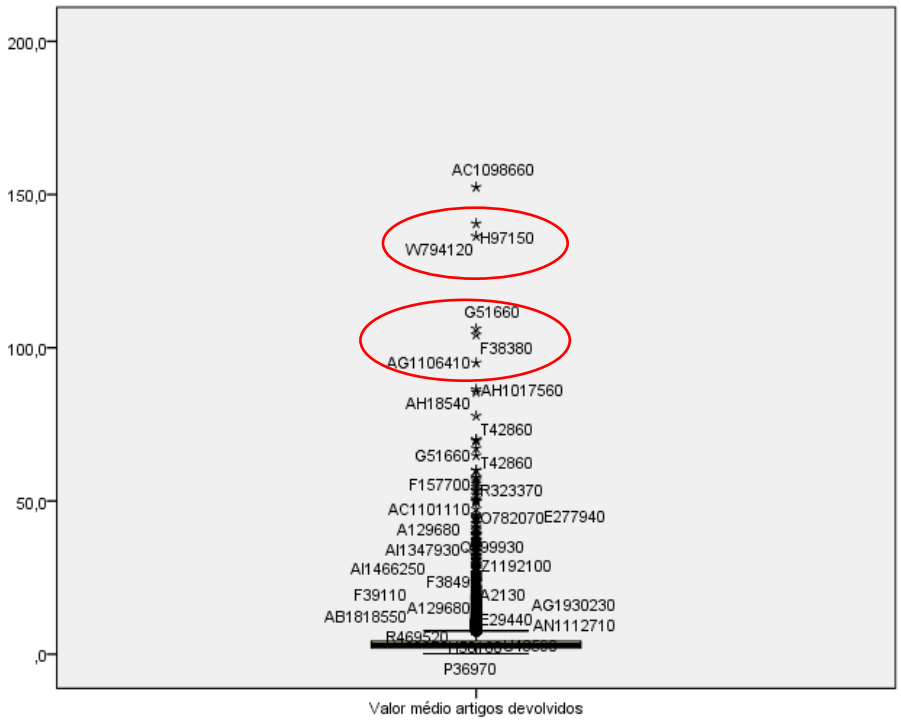
E Loja = A, AA, AB, AC, AD, AE, AF, AG, AH, AI, AJ, AK, AL, AM, AN, B, C, D, E, F, G, H, I, J, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y, Z

ENTÃO Valor médio artigos devolvidos = 3,94€

N = 13608

Desvio = 332110,1

Figura 3.41 – *Box-Plot* do Nó 34, variável "Valor médio artigos devolvidos"



Conclusão: observam-se **cinco novos *outliers* extremos** mais significativos, que não foram detetados em análises anteriores. Aplicando a regra descrita acima, o valor médio dos artigos devolvidos é de cerca de 3,94€, no entanto nos casos identificados, as devoluções efetuadas são superiores a 95€. Estes valores são superiores em cerca de 24 vezes o valor da média.

Regras de Decisão do Nó 60

SE Unidade Negócio = Lazer

E Loja = AD, AH, D, K, V

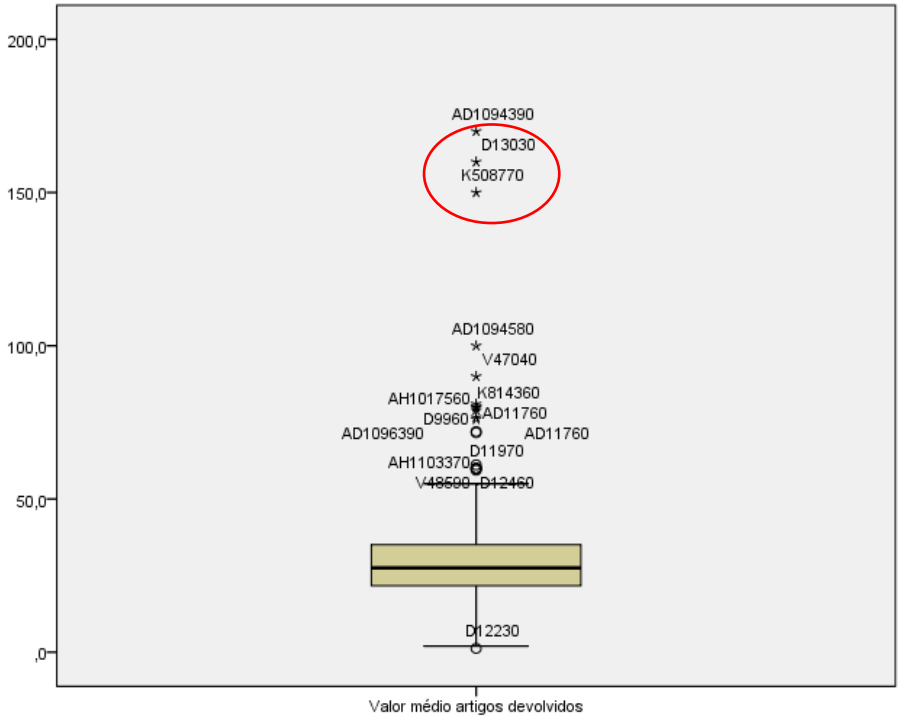
E Horario registo = Tarde, Noite

ENTÃO Valor médio artigos devolvidos = 31,92€

N = 231

Desvio = 111186,3

Figura 3.42 – *Box-Plot* do Nó 60, variável "Valor médio artigos devolvidos"



Conclusão: observam-se **dois novos *outliers* extremos** mais significativos, que não foram detetados em análises anteriores. Nos produtos de "Lazer", entre o horário da tarde e noite, e nas lojas AD, AH, D, K, V, o valor médio dos artigos devolvidos é de cerca de 31,92€, no entanto nos casos identificados, as devoluções efetuadas são superiores ou iguais a 150€. Estes valores são superiores em cerca de 4,7 vezes o valor da média.

Regras de Decisão do Nó 61

SE Unidade Negócio = Lazer

E Loja = AD, AH, D, K, V

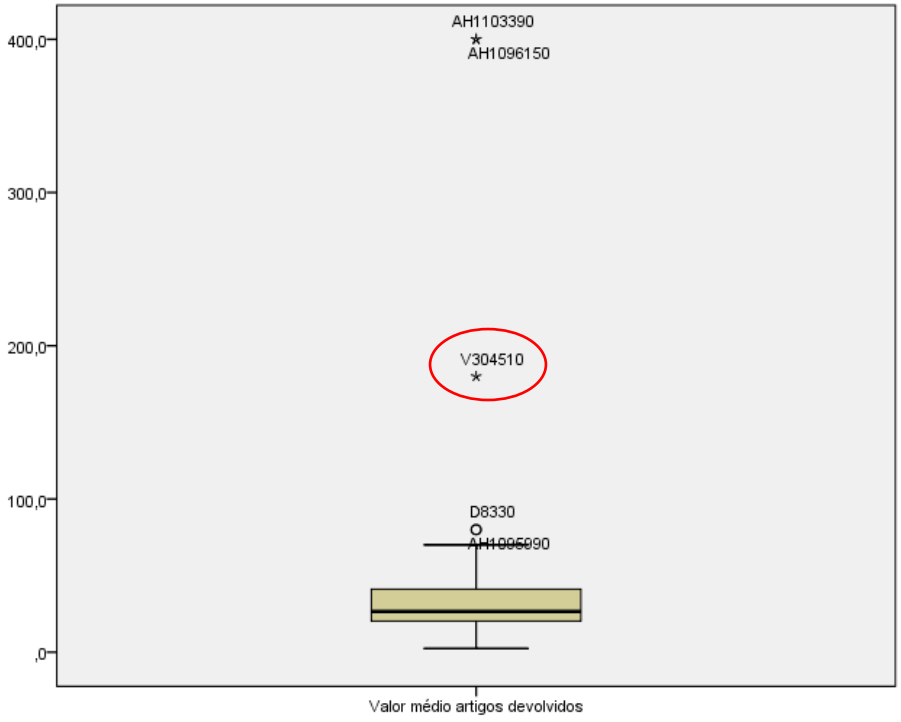
E Horario registo = Manhã

ENTÃO Valor médio artigos devolvidos **=** 39,97€

N = 95

Desvio = 308916

Figura 3.43 – *Box-Plot* do Nó 61, variável "Valor médio artigos devolvidos"



Conclusão: observa-se **um novo *outlier* extremo**, que não foi detetado em análises anteriores. Aplicando a regra acima descrita, o valor médio dos artigos devolvidos é de cerca de 39,97€, no entanto no caso identificado, as devoluções efetuadas são iguais a 180€. Estes valores são superiores em cerca de 4,5 vezes o valor da média.

Regras de Decisão do Nó 62

SE Unidade Negócio = Lazer

E Loja = AB

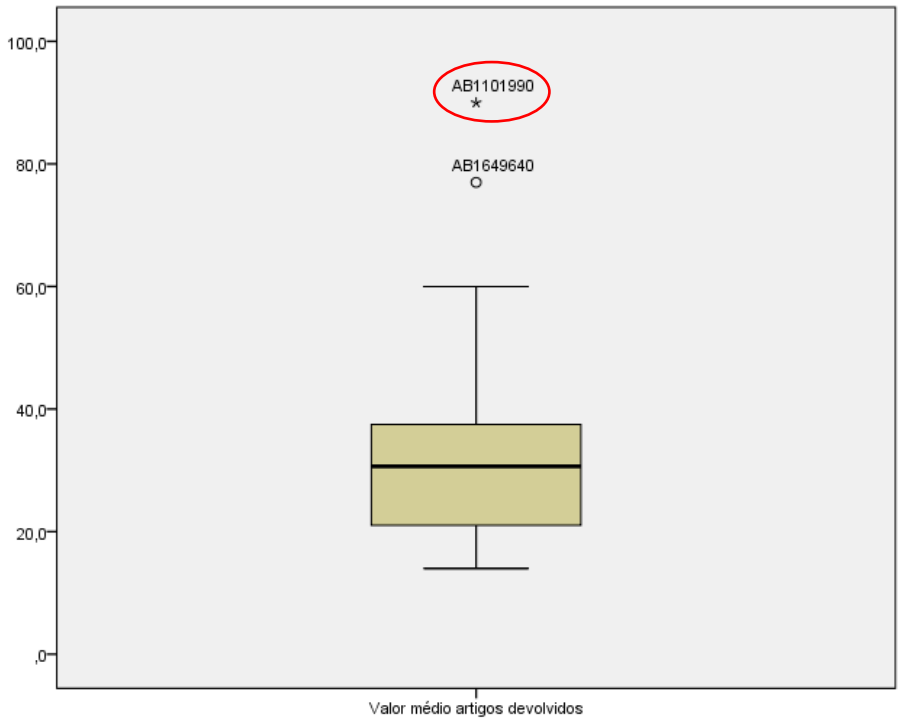
E Dia semana = Dias úteis

ENTÃO Valor médio artigos devolvidos = 33,56€

N = 25

Desvio = 8734,846

Figura 3.44 – *Box-Plot* do Nó 62, variável "Valor médio artigos devolvidos"



Conclusão: observa-se **<u>um novo *outlier* extremo</u>**, que não foi detetado em análises anteriores, no supervisor AB1101990. Nos produtos de "Lazer", nos dias úteis e na loja AB, o valor médio dos artigos devolvidos é de cerca de 33,56€, no entanto no caso identificado, as devoluções efetuadas são cerca de 90€. Este valor é superior em cerca de 2,7 vezes o valor da média.

Regras de Decisão do Nó 70

SE Loja = K

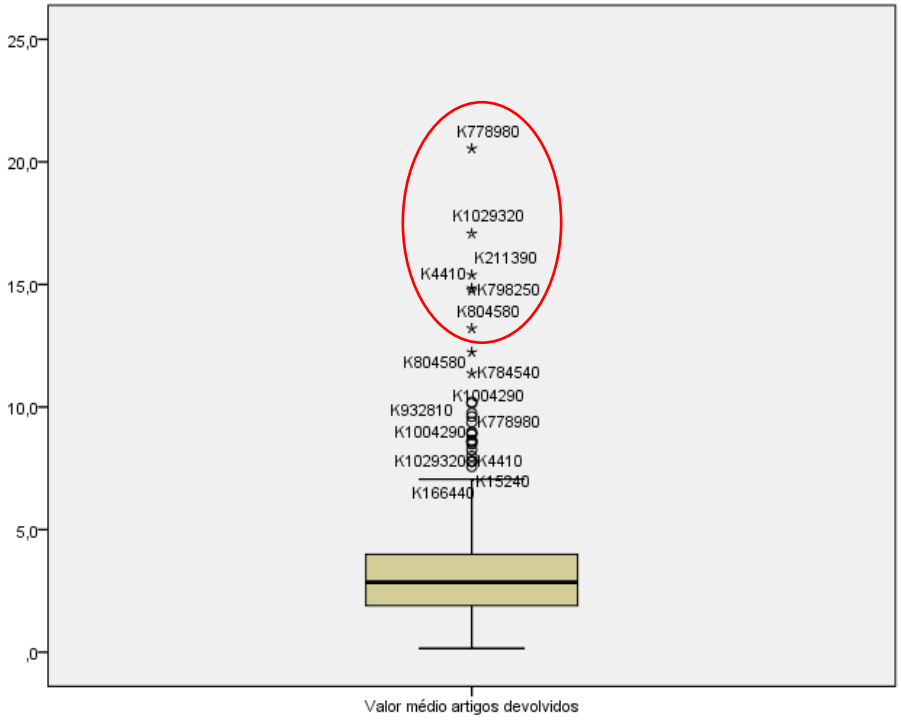
E Unidade Negócio = Charcutaria&Queijos, Congelados, Mercearia Doce, Mercearia Salgada, TakeAway

ENTÃO Valor médio artigos devolvidos = 3,51€

N = 316

Desvio = 2302,822

Figura 3.45 – *Box-Plot* do Nó 70, variável "Valor médio artigos devolvidos"



Conclusão: observam-se **seis novos *outliers* extremos** mais significativos, que não foram detetados em análises anteriores. Aplicando a regra acima descrita, o valor médio dos artigos devolvidos é de cerca de 3,51€, no entanto nos casos identificados, as devoluções efetuadas são superiores ou iguais a 13€. Estes valores são superiores em cerca de 3,7 vezes o valor da média.

Regras de Decisão do Nó 122

SE Unidade Negócio = Lazer

E Horario registo = Manhã

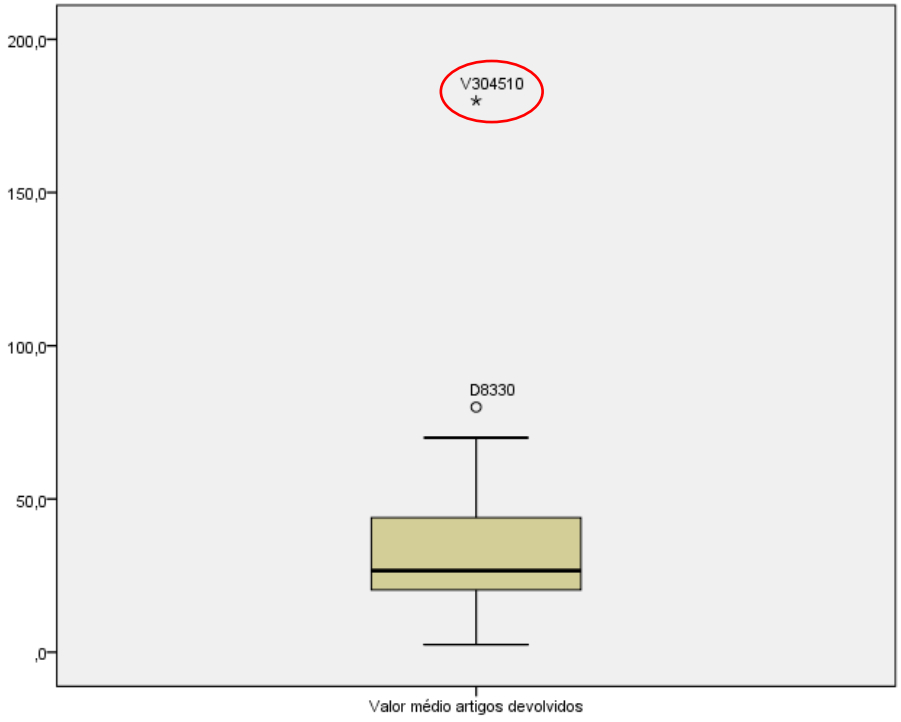
E Loja = AD, D, K, V

ENTÃO Valor médio artigos devolvidos = 33,33€

N = 71

Desvio = 38335,18

Figura 3.46 – *Box-Plot* do Nó 122, variável "Valor médio artigos devolvidos"



Conclusão: observa-se **um novo *outlier* extremo**, que não foi detetado em análises anteriores, no supervisor V304510. Nos produtos de Lazer, durante o horário da manhã e nas lojas AD, D, K e V, o valor médio dos artigos devolvidos é de cerca de 33,33€, no entanto no caso identificado, as devoluções efetuadas são cerca de 178€. Este valor é superior em cerca de 5,4 vezes o valor da média.

Regras de Decisão do Nó 286

SE Loja = K

E Unidade Negócio = Lacticinios

E Horario registo = Tarde

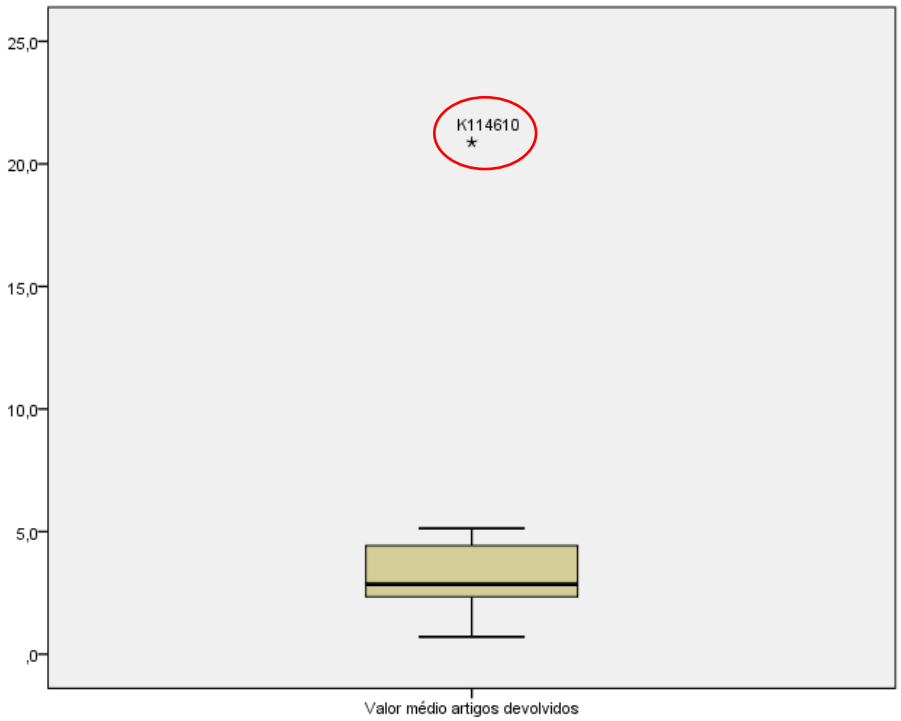
E Dia semana = Fins de semana ou feriado

ENTÃO Valor médio artigos devolvidos = 4,08€

N = 15

Desvio = 326,176

Figura 3.47 – *Box-Plot* do Nó 286, variável "Valor médio artigos devolvidos"



Conclusão: observa-se **<u>um novo *outlier* extremo</u>**, que não foi detetado em análises anteriores, no supervisor K114610. Na loja K, nos produtos de "Lacticinios", nos fins de semana ou feriados e durante o horário da tarde, o valor médio dos artigos devolvidos é de cerca de 4,08€, no entanto no caso identificado, as devoluções efetuadas é cerca de 21€. Este valor é superior em cerca de 5 vezes o valor da média.

Após a realização das análises multivariadas para as duas variáveis objetivo, é importante resumir os novos *outliers* extremos identificados, dado que não tinham sido observados nas análises univariadas e bivariadas:

Tabela 3.11 – Novos *Outliers* extremos identificados nas análises multivariada, variável objetivo: Total artigos devolvidos

Análises multivariadas		Novos *Outliers* extremos identificados (*outliers* de contexto)	
		Variável objectivo: Total artigos devolvidos	Total
Nº do Nó / Folha	12	A21440 | AE1116250 | AF1458260 | K932810 | I44830 | I458040	6
	14	AG1106410 | L753210 {2} | C691530 | C437860 {2}	6
	18	AE1116250 {2} | K1029320 | AB1115780 {2} | G50820	6
	22	B4880	1
	26	AM1520790 | U897930 | P99290 | AI1347930 {2}	5
	46	AG1110250 | AG1106410 {3}	4
	55	I27440 | Y1096090 {3} | I114130	5
	62	AH1103120	1
Total			34

Tabela 3.12 – Novos *Outliers* extremos identificados nas análises multivariada, variável objetivo: Valor médio artigos devolvidos

Análises multivariadas		Novos *Outliers* extremos identificados (*outliers* de contexto)	
		Variável objectivo: Valor médio artigos devolvidos	Total
Nº do Nó / Folha	5	G51690 | D33920 | AA253110 | G51280	4
	6	U298540 | AL263700 | H223880	3
	9	B99500 | I115150 | AA1224110 | AH1103390 | E1461750 | I1937860	6
	10	D33920	1
	11	G51280 | G51690 | AE1104430 | AA253110	4
	12	H223880 | G29950 | H97150 | G51280 | AE1104430 | A15160 | A10160	7
	13	U298540 | AL263700	2
	14	U180660 | J119130 | B9680 | Z1722240	4
	15	V304510 | D13030 | K508770	3
	16	I109720 | A10160	2
	30	V304510 | D13030 | K508770	3
	34	H97150 | W794120 | G51660 | F38380 | AG1106410	5
	60	D13030 | K508770	2
	61	V304510	1
	62	AB1101990	1
	70	K778980 | K1029320 | K211390 | K4410 | K798250 | K804580	6
	122	V304510	1
	286	K114610	1
Total			56

Conforme se verifica pelas tabelas 3.11 e 3.12, através das análises multivariadas, foram identificados **90 novos *outliers* extremos**, que não tinham sido observados nas análises univariada e bivariada.

Capítulo 4: Conclusões e trabalho futuro

Os objetivos deste trabalho de dissertação foram alcançados, pela criação de um modelo de aprendizagem não supervisionada, que permitisse a deteção de casos anómalos ou estranhos na área de devoluções de artigos em loja, que possam ser fraude ou não.
O conjunto de dados analisado foi obtido através de transações reais, num grupo de 40 lojas de um retalhista. O período observado é referente a Dezembro-2014 a Fevereiro-2015.
As devoluções de artigos são um dos esquemas mais comuns de fraude na indústria do retalho, fazendo parte da sub-categoria de "Desembolsos fraudulentos" da árvore da fraude publicada pela maior organização mundial antifraude que é a ACFE.

O plano de trabalho consistiu na deteção de casos anómalos ou estranhos pela identificação de *outliers*, observados através dos gráficos *box-plot*, e pelas análises univariadas, bivariadas e multivariadas, aplicadas sobre as variáveis do conjunto de dados.
As análises univariadas permitiram de uma forma rápida identificar vários *outliers* extremos globais, num grupo de características pouco homogéneas. Embora sejam análises úteis, carecem de maior profundidade na extração de conhecimento, dado que apenas olhamos para uma variável de cada vez.
Desse modo, o passo seguinte foi efetuar análises bivariadas, pela combinação das variáveis, duas a duas. Aqui foi possível explorar e identificar novos casos (*outliers*) não observados anteriormente nas análises univariadas.
Por fim, efetuou-se análises multivariadas como forma de estudar as duas variáveis quantitativas, através da combinação das variáveis qualitativas em simultâneo. Essa análise foi efetuada através do desenvolvimento de árvores de regressão, tendo revelado novos casos desconhecidos até então.

Destes três tipos de análise, a multivariada é sem dúvida a mais rica em termos de extração de conhecimento, dado que para além de identificar *outliers* globais, consegue descer a um nível de detalhe muito interessante, e revelar casos bem "escondidos", considerados como *outliers* de contexto.

Por exemplo, só assim foi possível identificar dois *outliers* extremos (ver figura 3.39), num contexto de devoluções de produtos de "Frutas&Legumes" ou "Padaria", cujo valor médio ronda os 2,55€, mas no entanto dois supervisores realizaram devoluções a um valor médio superior a 35€. Este valor é superior em cerca de 13,7 vezes o valor da média. Este é um tipo de *outlier* de contexto, que dificilmente seria observado nas análises univariada e bivariada, dai a importância das análises multivariadas.

Em termos de resumo final dos resultados alcançados, deve-se referir que foram identificados cerca de 147 *outliers* extremos (total entre globais e de contexto, pelos três tipos de análises), o que confirma o objetivo do trabalho de detetar casos anómalos ou estranhos. Porém a confirmação de cada caso sobre a conclusão se é fraude ou não, só é possível através de uma averiguação e investigação nas respetivas lojas, sendo que esse processo pode envolver técnicas de investigação forense.
Desse modo o âmbito desta dissertação termina nesta fase de deteção de casos anómalos ou estranhos.

Proposta para trabalho futuro:

Perante a confirmação e satisfação com os resultados obtidos, o trabalho futuro mais imediato passa pela sistematização do modelo apresentado, para permitir a identificação mais periódica de situações estranhas. Os gráficos de deteção de *outliers* são bastante intuitivos e podem ser desenvolvidos no SPSS ou no R.
Um outro desafio passa também pela obtenção de novas variáveis associadas ao processo das devoluções de artigos, que possam enriquecer e complementar a análise.
A inclusão de variáveis sobre o *stock* dos artigos devolvidos pode ser interessante, dado que um dos riscos das devoluções é o furto do artigo devolvido, após o registo da devolução e reembolso ao cliente.
Seria também interessante analisar outros conjuntos de dados, com dimensão temporal, explorar e detetar casos anómalos ou estranhos, tal como efetuado neste trabalho.
O modelo desenvolvido neste trabalho é extensível e flexível para outros esquemas de possível fraude, como anulação de artigos durante a venda ou anulação de transações durante a venda, etc.

Bibliografia

Abbott, D. (2014), "Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst", Wiley, 1ª Edição

ACFE (2014), "Report to the Nations on Occupational Fraud and Abuse", http://www.acfe.com/rttn/docs/2014-report-to-nations.pdf, acedido em 25 de Outubro 2014.

ACFE (2014), "About the ACFE", http://www.acfe.com/about-the-acfe.aspx, acedido em 15 Novembro 2014.

ACFE (2014), "ACFE Leadership", http://www.acfe.com/bio-jwells.aspx, acedido em 15 Novembro 2014.

Breiman, L., Friedman, J. H., Olshen, R. A. e C. J. Stone (1984), "Classification and Regression Tress", Wadsworth International Group, Belmont, CA

Chandola, V., Kumar, V. e A. Banerjee (2009), "Outlier Detection: A Survey", ACM Computing Surveys (CSUR), Volume 41 Issue 3, Artigo Nº 15

Checkpoint Systems, Inc. (2014), "Barómetro Mundial do Furto no Retalho 2013-2014", Primeira Edição: Outubro de 2014, USA

Coderre, D. (2009), "Computer-Aided Fraud Prevention & Detection", John Wiley & Sons

Deloitte (2013), "Shrinking retail shrink: Using analytics to help detect fraud and grow margins", publicado no seu sítio institucional em 2013, acedido em 15 Novembro 2014

EY (2014), "Big risks require big data thinking", publicado no seu sítio institucional em 2014, acedido em 11 Dezembro 2014

Gama, J., Carvalho, A., Faceli, K., Lorena, C. e M. Oliveira (2012), "Extração de Conhecimento de Dados: Data Mining", Edições Sílabo, 1ª Edição

Hawkins, D. M. (1980), "Identification of Outliers", Chapman and Hall London

Hodge, V. e J. Austin (2004), "A Survey of Outlier Detection Methodologies", Artificial Intelligence Review, Volume 22

IIA (2012), "Normas Internacionais para a Prática Profissional de Auditoria Interna", Altamonde Springs, USA: The Institute of Internal Auditors.

IPAI (2014), "Auditoria interna: Contributo para a deteção e prevenção de fraude nas organizações", *Auditoria Interna*, Nº 5, pp. 13-24

Jans, M., Lybaert N. e K. Vanhoof (2007), "Data mining for fraud detection: Toward an improvement on internal control systems ?", 30º Annual Congress European Accounting Association (EAA 2007), Lisboa

KPMG (2006), "Fraud Risk Management: Developing a Strategy for Prevention, Detection, and Response", publicado no seu sítio institucional em 10/05/2006, acedido em 15 Novembro 2014

KPMG (2013), "Global profiles of the fraudster", publicado no seu sítio institucional em 11/04/2013, acedido em 15 Novembro 2014

Kristin, R. N. e I. P. Matkovsky (1999), "Using Data Mining Techniques for Fraud Detection", SAS Institute Inc. and Federal Data Corporation

Manish, G., Gao, J., Aggarwal, C. e H. Jiawei (2014), "Outlier Detection for Temporal Data: A Survey", IEEE Transactions on knowledge and data engineering, Vol. 25, Nº 1

Marôco, J. (2011), "Análise Estatística com o SPSS Statistics", Report Number, 5ª Edição

PWC (2014), "Global Economic Crime Survey 2014", publicado no seu sítio institucional em 14/02/2014, acedido em 15 Novembro 2014

Pimenta, C. (2009), "Esboço de Quantificação da Fraude em Portugal", Observatório de Economia e Gestão de Fraude, Working Papers Nº3, Edições Húmus

Phua, C., Lee V., Smith K. e R. Gayler (2010), "A Comprehensive Survey of Data Mining-based Fraud Detection Research", CoRR

R Core Team (2014), "R: A language and environment for statistical computing", R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org/

Singh, K. e S. Upadhyaya (2012), "Outlier Detection: Applications and Techniques", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3

Therneau, T., Atkinson B. e B. Ripley (2015), "rpart: Recursive Partitioning and Regression Trees", R package version 4.1-10, http://CRAN.R-project.org/package=rpart

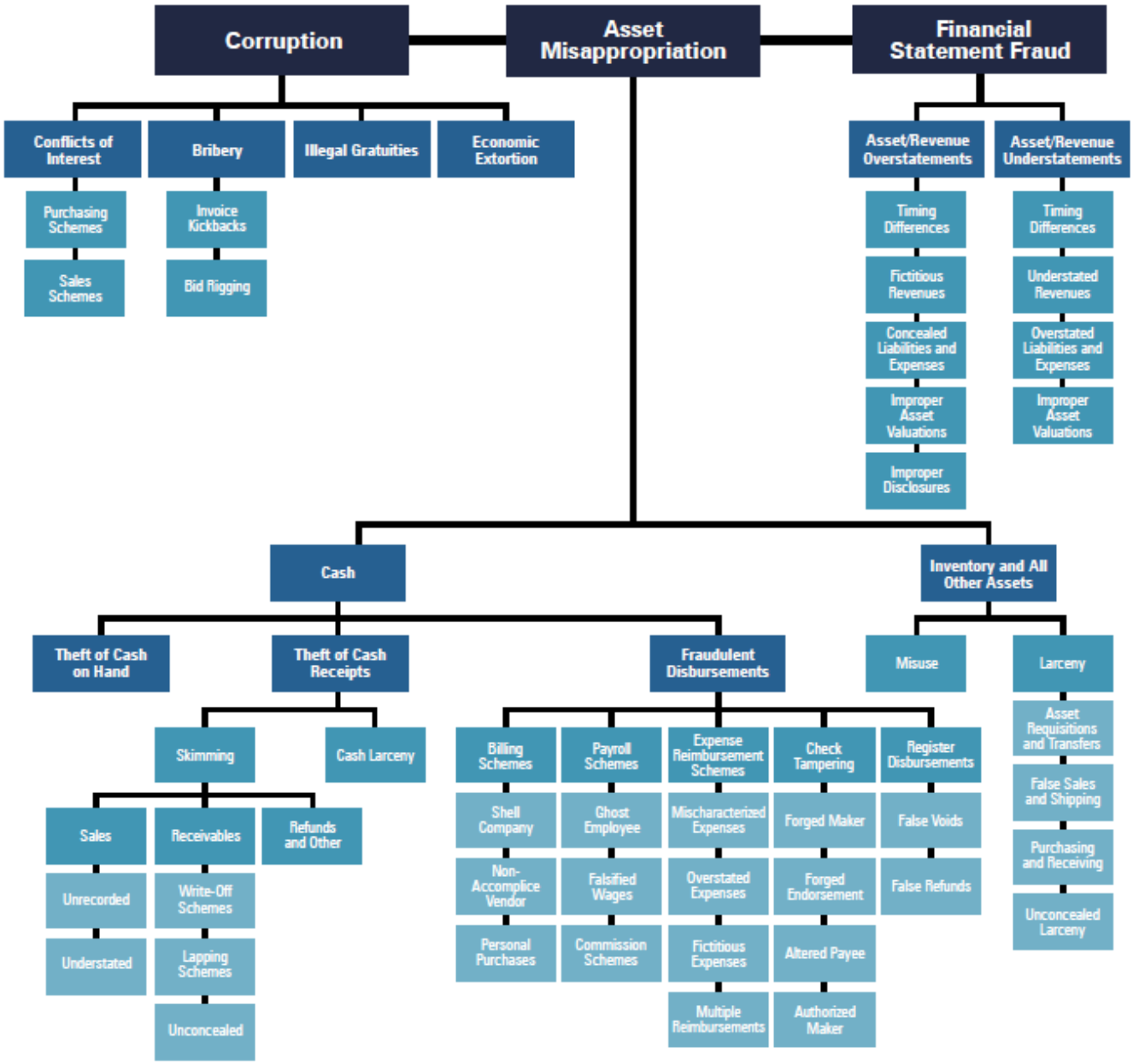
Torgo, L. (2010), "Data Mining with R: Learning with Case Studies", Chapman and Hall/CRC, 1ª Edição

Wells, J. T. (2009), "Manual da Fraude na Empresa", Almedina

Zhang, Y., Meratnia, N. e P. Havinga (2007), "A Taxonomy Framework for Unsupervised Outlier Detection Techniques for Multi-Type Data Sets", Centre for Telematics and Information Technology University of Twente

ANEXO I – Árvore da fraude

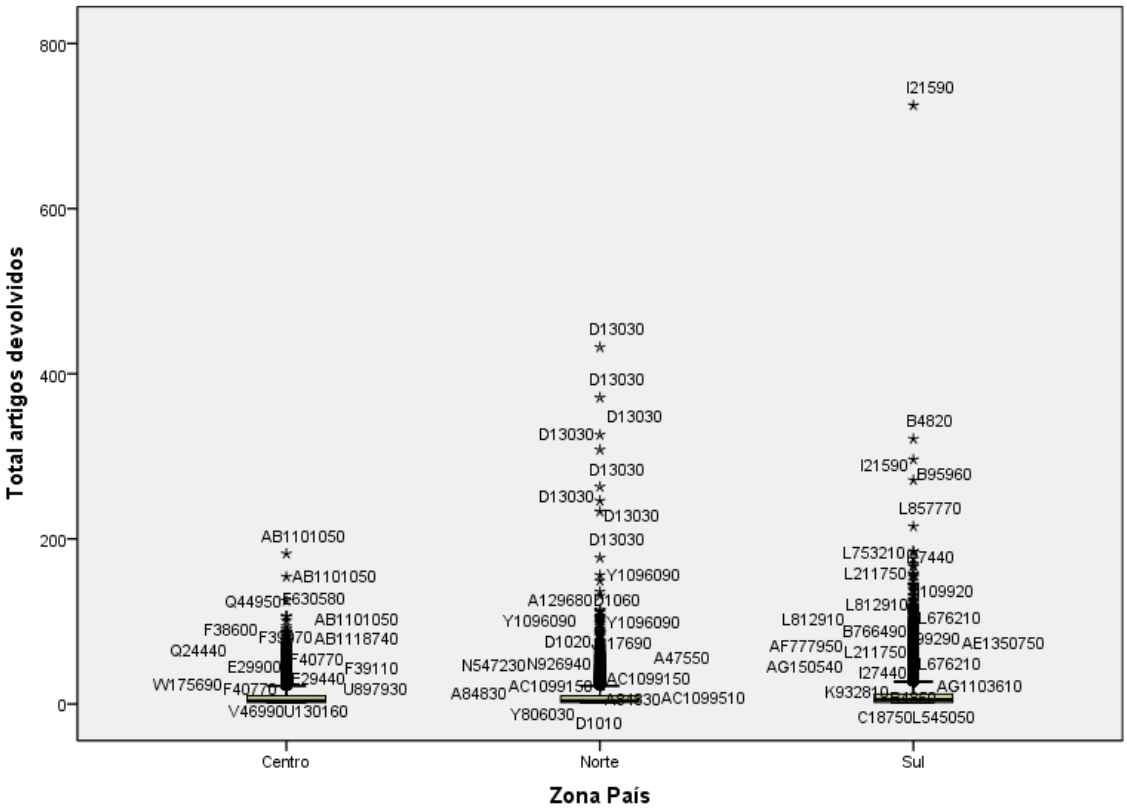
Figura A1.1 - Árvore da Fraude (versão completa)



Fonte: ACFE, 2014, p. 11

ANEXO II - Análises bivariadas

Figura A2.1 – *Box-Plot*: "Zona País" vs "Total artigos devolvidos"

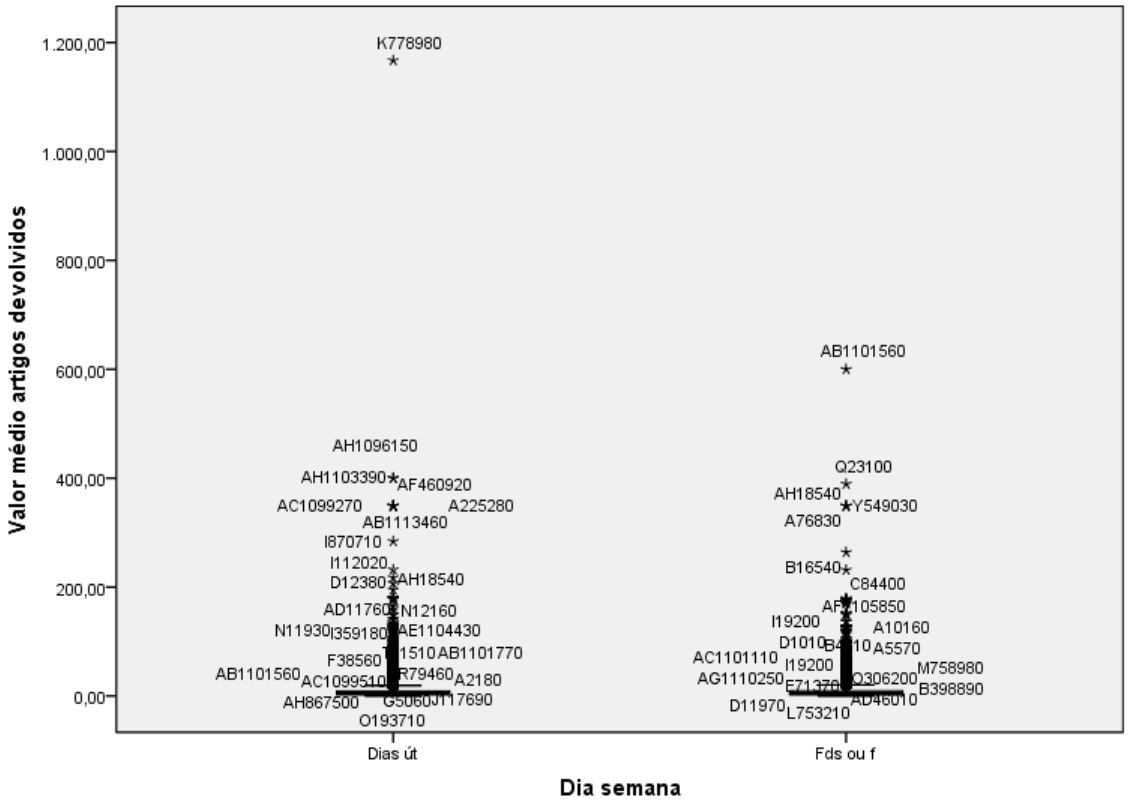


Nesta análise pretende-se verificar e detetar a presença de *outliers* da variável "Total artigos devolvidos" em função da área geográfica onde a loja se situa.

Observa-se que o número de *outliers* é muito elevado, pelo que de forma nítida não é possível identificar novos *outliers*.

Conclusão: esta análise <u>não revelou a existência de novos *outliers*</u>, apenas se observa os *outliers* já conhecidos nas análises univariada e bivariada do capítulo 3.

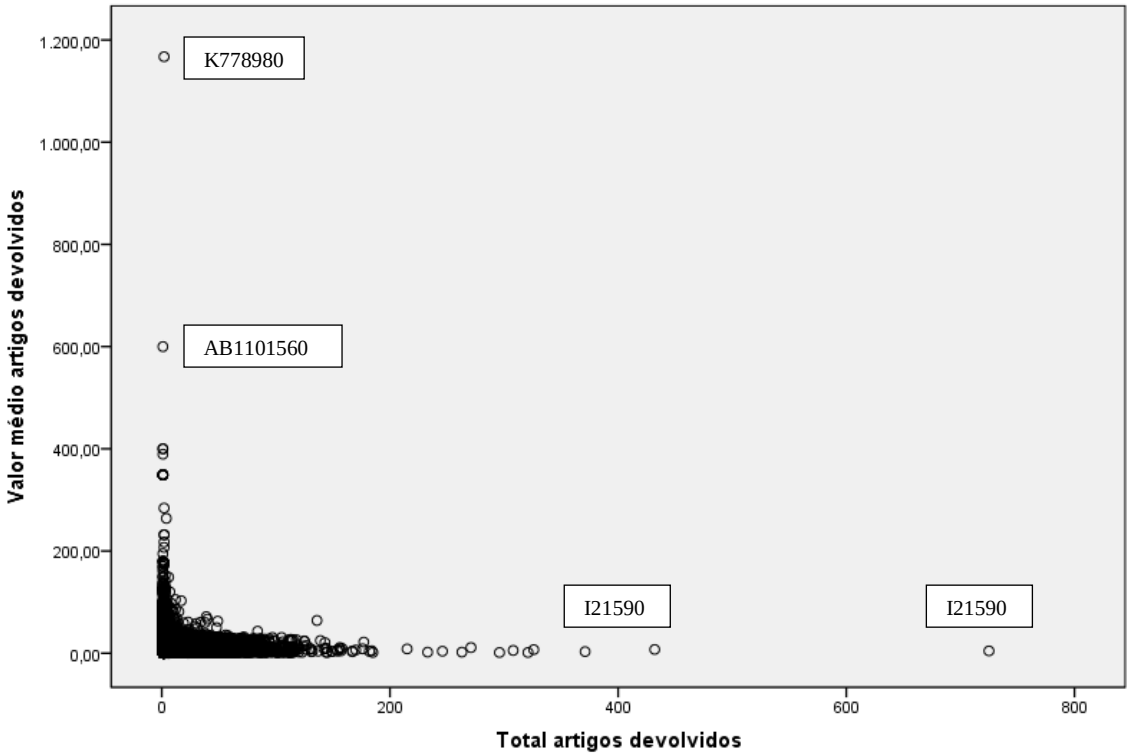
Figura A2.2 – *Box-Plot*: "Dia semana" vs "Valor médio artigos devolvidos"



Nesta análise pretende-se verificar e detetar a presença de *outliers* da variável "Valor médio artigos devolvidos" em função do dia da semana, ou seja, dias úteis ou fins de semana/feriados.

Conclusão: esta análise não revelou a existência de novos *outliers*, apenas se observa os *outliers* já conhecidos nas análises univariada e bivariada do capítulo 3.

Figura A2.3 – *Scatter-Plot*: "Total artigos devolvidos" vs "Valor médio artigos devolvidos"



Nesta análise pretende-se cruzar as duas variáveis quantitativas e verificar se existem novos *outliers* ainda não detetados, através de um gráfico chamado de "scatter-plot".

Conclusão: esta análise não revelou a existência de novos *outliers*, tendo detetado *outliers* importantes, mas já conhecidos das análises univariadas.

ANEXO III - Análises multivariadas: variável objetivo "Total artigos devolvidos"

Regras de Decisão do Nó 2

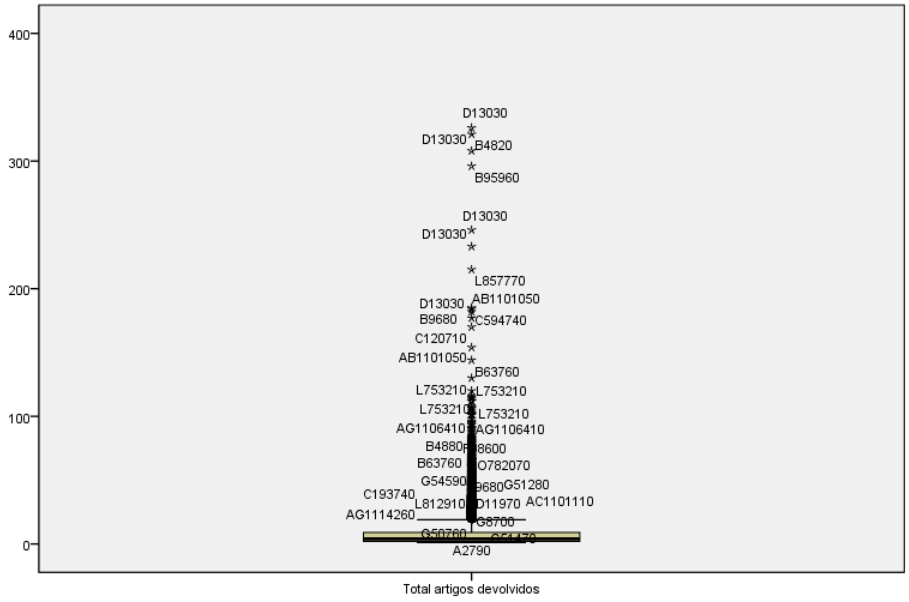
SE Unidade Negócio = Bebidas, Bricolage&Auto, Charcutaria&Queijos, Congelados, Cultura, Frutas&Legumes, Lacticinios, Limpeza, Padaria, Peixaria, Pets&Plants, TakeAway, Talho

ENTÃO Total artigos devolvidos **=** 7,37

N = 29545

Desvio = 3525370

Figura A3.1 – *Box-Plot* do Nó 2, variável "Total artigos devolvidos"



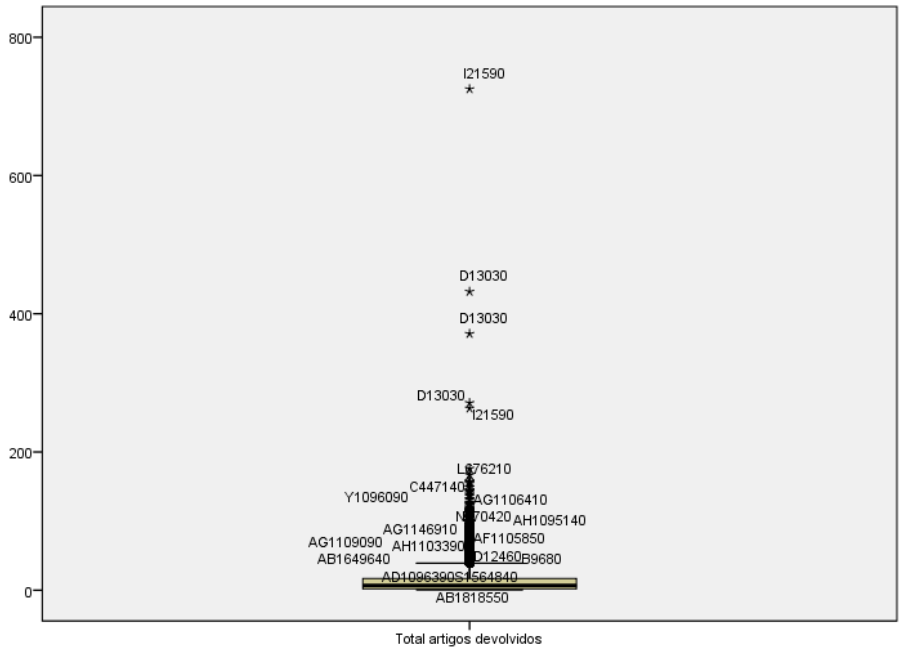
Nota: esta análise não revelou a existência de novos *outliers* extremos significativos, apenas os já conhecidos.

Regras de Decisão do Nó 3

SE Unidade Negócio = Casa, Higiene, Lazer, Mercearia Doce, Mercearia Salgada
ENTÃO Total artigos devolvidos **=** 13,59

N = 13661
Desvio = 5314693

Figura A3.2 – *Box-Plot* do Nó 3, variável "Total artigos devolvidos"



Nota: esta análise não revelou a existência de novos *outliers* extremos significativos, apenas os já conhecidos.

Regras de Decisão do Nó 4

SE Unidade Negócio = Bebidas, Bricolage&Auto, Charcutaria&Queijos, Congelados, Cultura, Frutas&Legumes, Lacticinios, Limpeza, Padaria, Peixaria, Pets&Plants, TakeAway, Talho

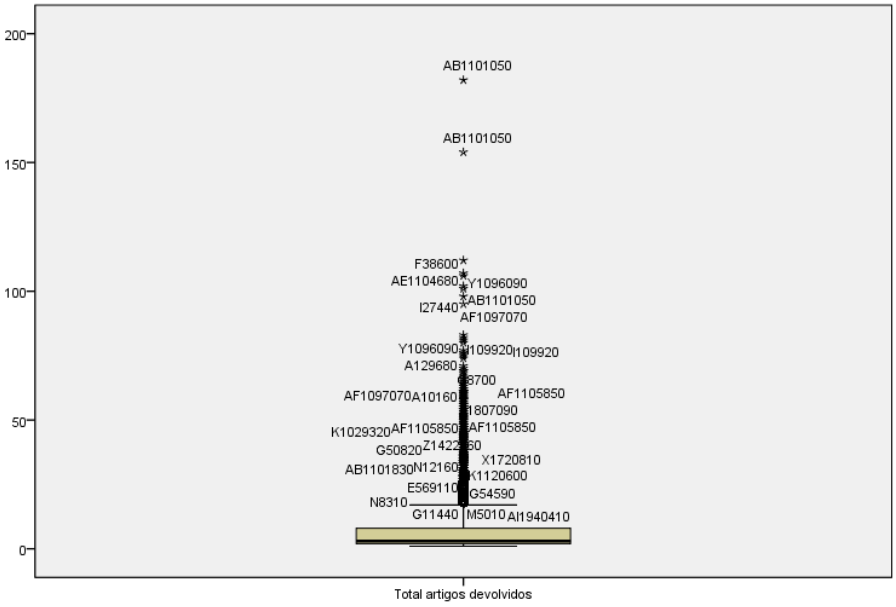
E Loja = A, AA, AB, AC, AD, AE, AF, AI, AJ, AK, AL, AM, AN, E, F, G, H, I, J, K, M, N, O, P, R, S, T, U, V, W, X, Y, Z

ENTÃO Total artigos devolvidos **=** 6,14

N = 22563

Desvio = 1400686

Figura A3.3 – *Box-Plot* do Nó 4, variável "Total artigos devolvidos"



Nota: esta análise não revelou a existência de novos *outliers* extremos significativos, apenas os já conhecidos.

Regras de Decisão do Nó 5

SE Unidade Negócio = Bebidas, Bricolage&Auto, Charcutaria&Queijos, Congelados, Cultura, Frutas&Legumes, Lacticinios, Limpeza, Padaria, Peixaria, Pets&Plants, TakeAway, Talho

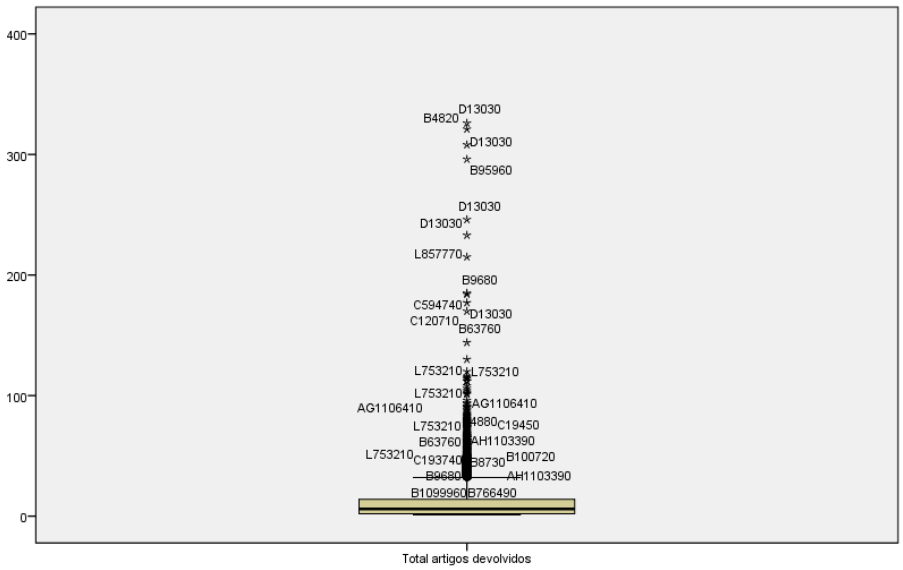
E Loja = AG, AH, B, C, D, L, Q

ENTÃO Total artigos devolvidos **=** 11,36

N = 6982

Desvio = 1979454

Figura A3.4 – *Box-Plot* do Nó 5, variável "Total artigos devolvidos"



Nota: esta análise não revelou a existência de novos *outliers* extremos significativos, apenas os já conhecidos.

Regras de Decisão do Nó 6

SE Unidade Negócio = Casa, Higiene, Lazer, Mercearia Doce, Mercearia Salgada

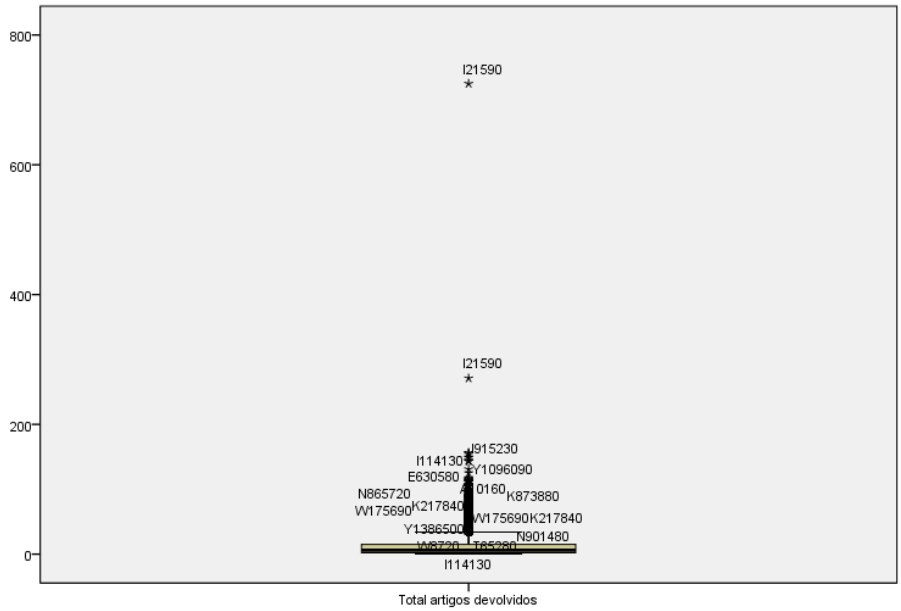
E Loja = A, AA, AB, AC, AD, AE, AF, AI, AJ, AK, AL, AM, AN, E, F, G, H, I, J, K, M, N, O, P, R, S, T, U, V, W, X, Y, Z

ENTÃO Total artigos devolvidos **=** 11,59

N = 10673

Desvio = 2977266

Figura A3.5 – *Box-Plot* do Nó 6, variável "Total artigos devolvidos"



Nota: esta análise não revelou a existência de novos *outliers* extremos significativos, apenas os já conhecidos.

Regras de Decisão do Nó 7

SE Unidade Negócio = Casa, Higiene, Lazer, Mercearia Doce, Mercearia Salgada

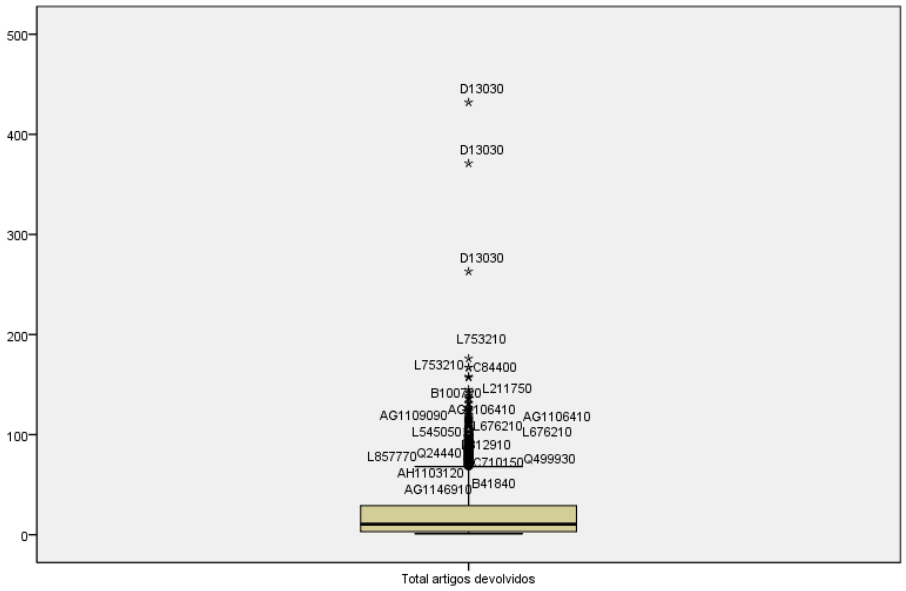
E Loja = AG, AH, B, C, D, L, Q

ENTÃO Total artigos devolvidos **=** 20,72

N = 2988

Desvio = 2142962

Figura A3.6 – *Box-Plot* do Nó 7, variável "Total artigos devolvidos"



Nota: esta análise não revelou a existência de novos *outliers* extremos significativos, apenas os já conhecidos.

Regras de Decisão do Nó 8

SE Loja = A, AA, AB, AC, AD, AE, AF, AI, AJ, AK, AL, AM, AN, E, F, G, H, I, J, K, M, N, O, P, R, S, T, U, V, W, X, Y, Z

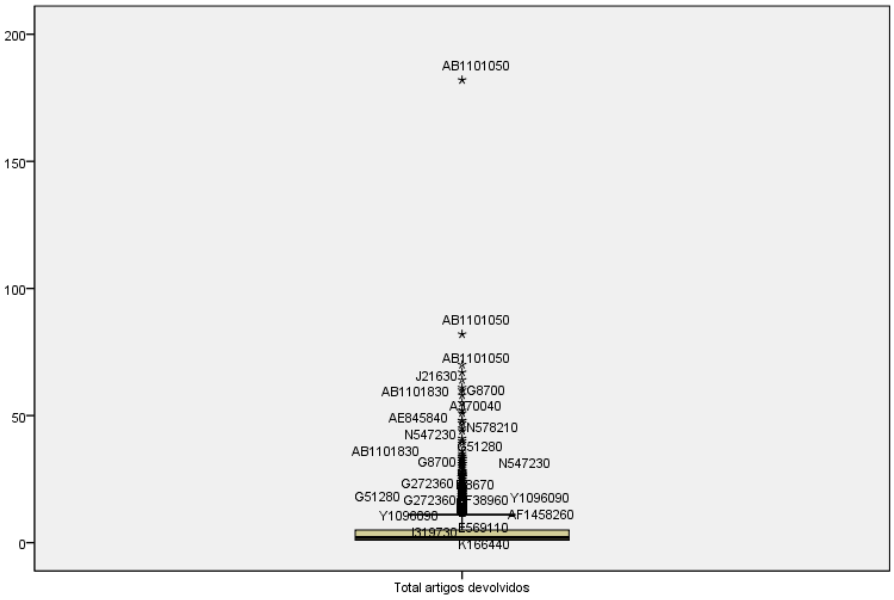
E Unidade Negócio = Congelados, Padaria, Peixaria, Pets&Plants, TakeAway, Talho

ENTÃO Total artigos devolvidos **=** 4,02

N = 8489

Desvio = 222951,4

Figura A3.7 – *Box-Plot* do Nó 8, variável "Total artigos devolvidos"



Nota: esta análise não revelou a existência de novos *outliers* extremos significativos, apenas os já conhecidos.

Regras de Decisão do Nó 9

SE Loja = A, AA, AB, AC, AD, AE, AF, AI, AJ, AK, AL, AM, AN, E, F, G, H, I, J, K, M, N, O, P, R, S, T, U, V, W, X, Y, Z

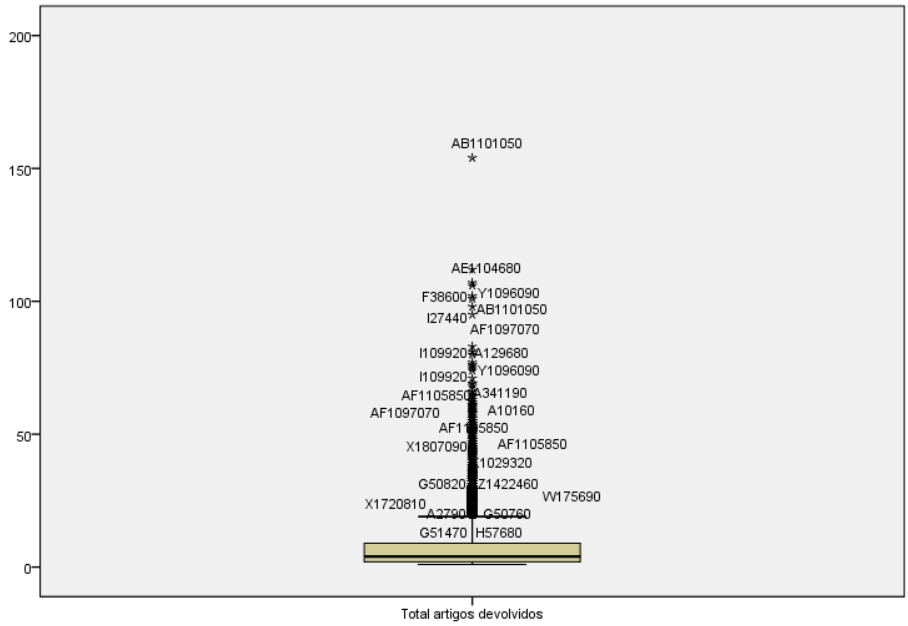
E Unidade Negócio = Bebidas, Bricolage&Auto, Charcutaria&Queijos, Cultura, Frutas&Legumes, Lacticinios, Limpeza

ENTÃO Total artigos devolvidos **=** 7,41

N = 14074

Desvio = 1116969

Figura A3.8 – *Box-Plot* do Nó 9, variável "Total artigos devolvidos"



Nota: esta análise não revelou a existência de novos *outliers* extremos significativos, apenas os já conhecidos.

Regras de Decisão do Nó 10

SE Loja = AG, AH, B, C, D, L, Q

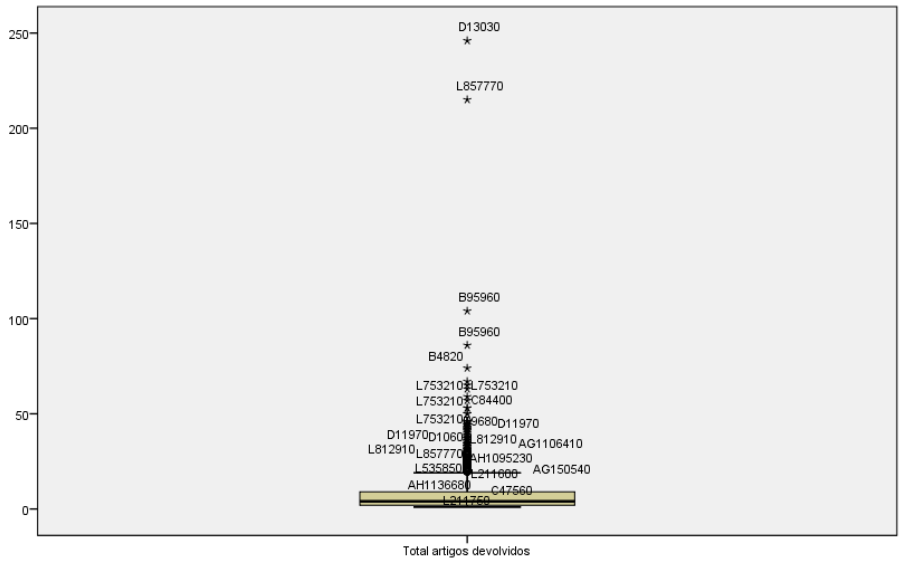
E Unidade Negócio = Congelados, Padaria, Peixaria, Pets&Plants, TakeAway, Talho

ENTÃO Total artigos devolvidos **=** 6,89

N = 2847

Desvio = 281338,7

Figura A3.9 – *Box-Plot* do Nó 10, variável "Total artigos devolvidos"



Nota: esta análise não revelou a existência de novos *outliers* extremos significativos, apenas os já conhecidos.

Regras de Decisão do Nó 11

SE Loja = AG, AH, B, C, D, L, Q

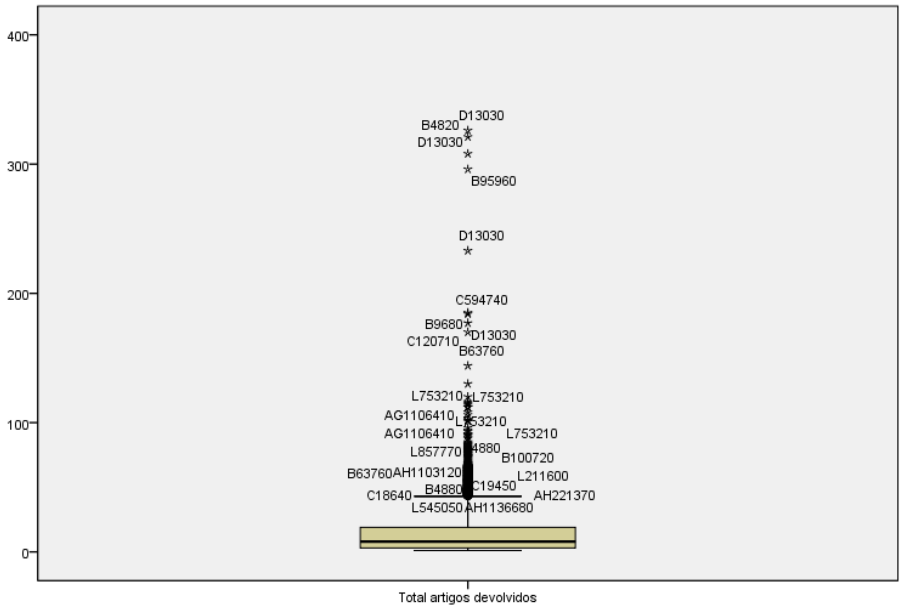
E Unidade Negócio = Bebidas, Bricolage&Auto, Charcutaria&Queijos, Cultura, Frutas&Legumes, Lacticinios, Limpeza

ENTÃO Total artigos devolvidos **=** 14,44

N = 4135

Desvio = 1601965

Figura A3.10 – *Box-Plot* do Nó 11, variável "Total artigos devolvidos"



Nota: esta análise não revelou a existência de novos *outliers* extremos significativos, apenas os já conhecidos.

Regras de Decisão do Nó 13

SE Unidade Negócio = Casa, Higiene, Lazer, Mercearia Doce, Mercearia Salgada

E Loja = A, AA, AB, AC, AD, AE, AF, AI, AJ, AK, AL, AM, AN, E, F, G, H, I, J, K, M, N, O, P, R, S, T, U, V, W, X, Y, Z

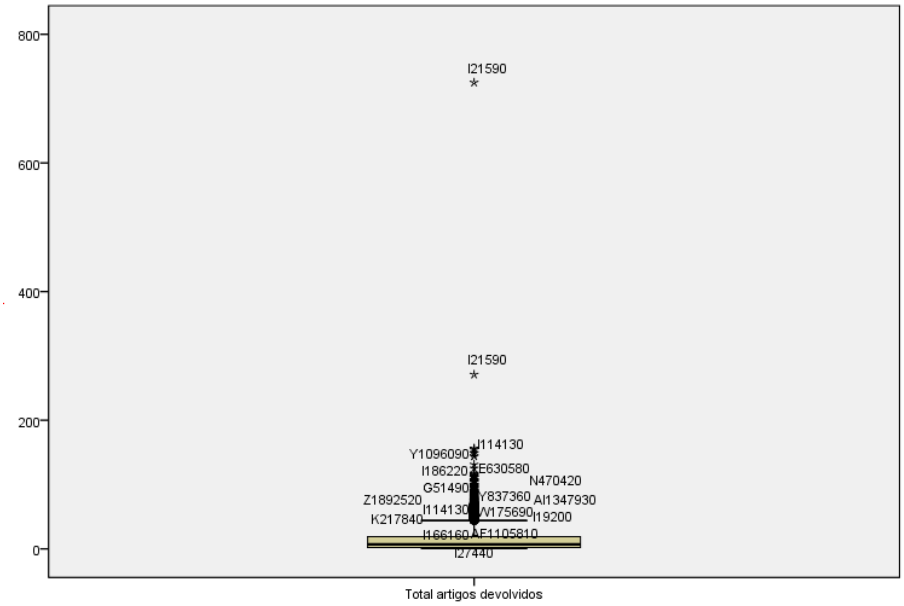
E Dia Semana = Dias úteis

ENTÃO Total artigos devolvidos **=** 13,98

N = 5838

Desvio = 2388991

Figura A3.11 – *Box-Plot* do Nó 13, variável "Total artigos devolvidos"



Nota: esta análise não revelou a existência de novos *outliers* extremos significativos, apenas os já conhecidos.

Regras de Decisão do Nó 15

SE Unidade Negócio = Casa, Higiene, Lazer, Mercearia Doce, Mercearia Salgada

E Loja = AG, AH, B, C, D, L, Q

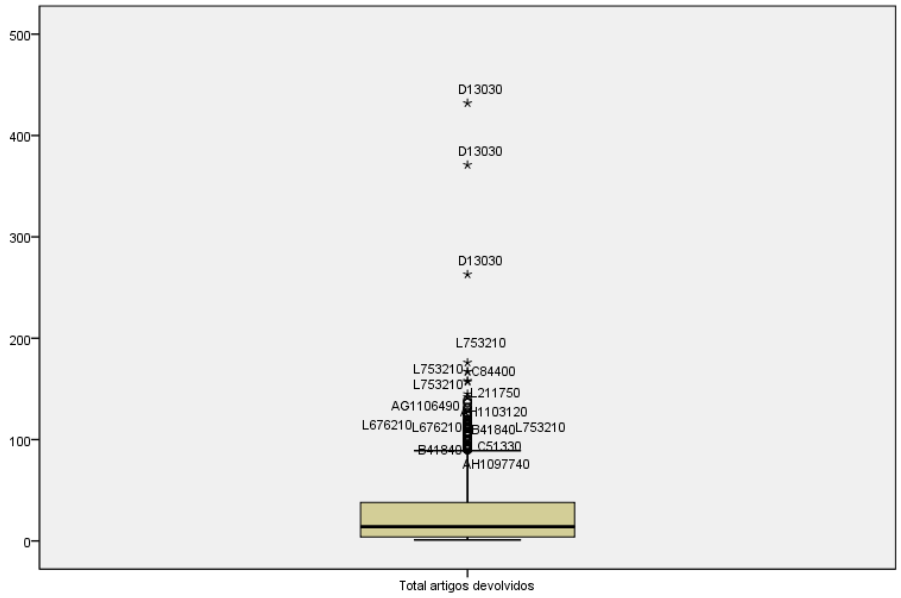
E Dia semana = Dias úteis

ENTÃO Total artigos devolvidos **=** 25,87

N = 1624

Desvio = 1697310

Figura A3.12 – *Box-Plot* do Nó 15, variável "Total artigos devolvidos"



Nota: esta análise não revelou a existência de novos *outliers* extremos significativos, apenas os já conhecidos.

Regras de Decisão do Nó 19

SE Loja = A, AA, AB, AC, AD, AE, AF, AI, AJ, AK, AL, AM, AN, E, F, G, H, I, J, K, M, N, O, P, R, S, T, U, V, W, X, Y, Z

E Unidade Negócio = Bebidas, Bricolage&Auto, Charcutaria&Queijos, Cultura, Frutas&Legumes, Lacticinios, Limpeza

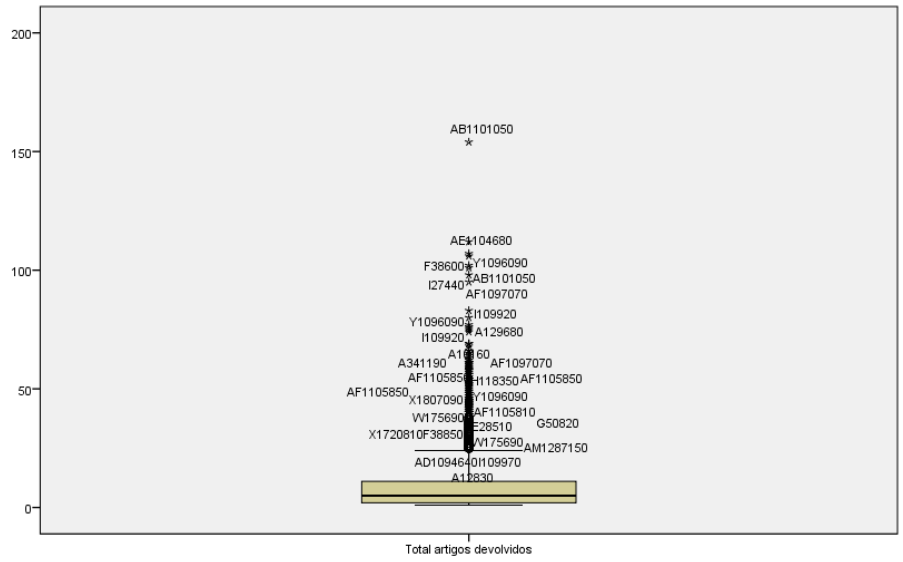
E Dia semana = Dias úteis

ENTÃO Total artigos devolvidos **=** 8,7

N = 7772

Desvio = 823273,8

Figura A3.13 – *Box-Plot* do Nó 19, variável "Total artigos devolvidos"



Nota: esta análise não revelou a existência de novos *outliers* extremos significativos, apenas os já conhecidos.

Regras de Decisão do Nó 23

SE Loja = AG, AH, B, C, D, L, Q

E Unidade Negócio = Bebidas, Bricolage&Auto, Charcutaria&Queijos, Cultura, Frutas&Legumes, Lacticinios, Limpeza

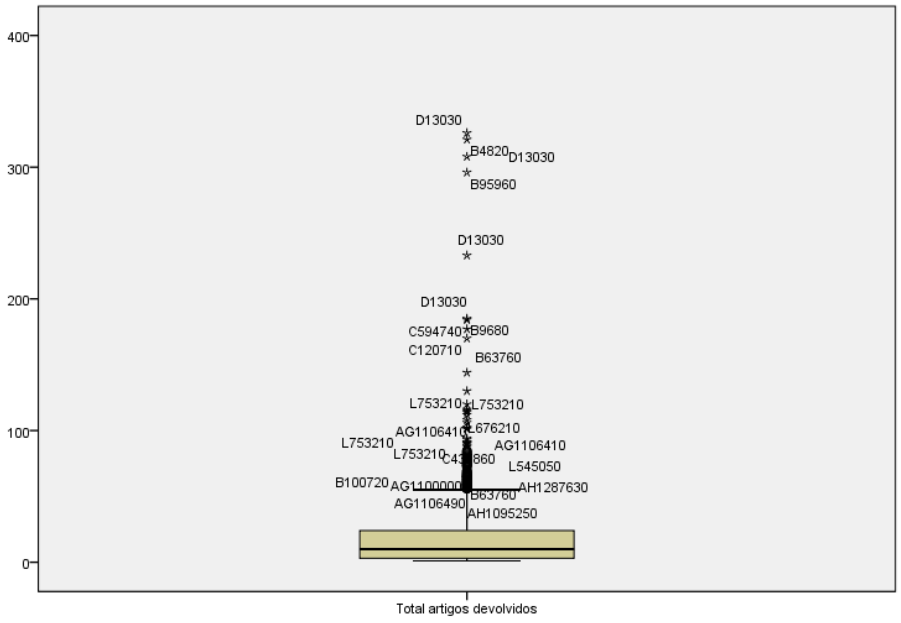
E Dia Semana = Dias úteis

ENTÃO Total artigos devolvidos **=** 17,87

N = 2249

Desvio = 1282552

Figura A3.14 – *Box-Plot* do Nó 23, variável "Total artigos devolvidos"



Nota: esta análise não revelou a existência de novos *outliers* extremos significativos, apenas os já conhecidos.

Regras de Decisão do Nó 27

SE Unidade Negócio = Casa, Higiene, Lazer, Mercearia Doce, Mercearia Salgada

E Dia Semana = Dias úteis

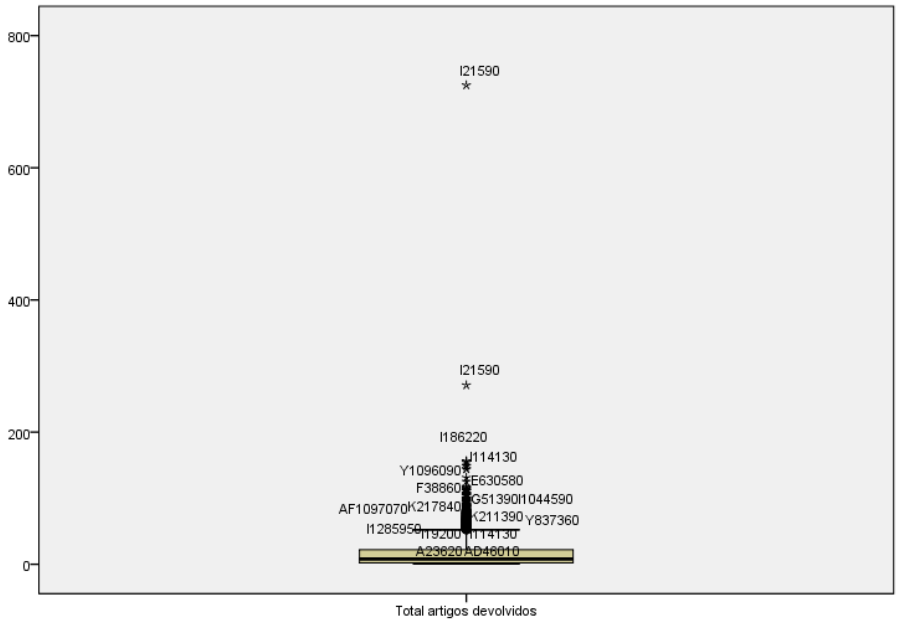
E Loja = A, AB, AD, AE, AF, E, F, G, H, I, J, K, N, W, X, Y, Z

ENTÃO Total artigos devolvidos **=** 16,15

N = 3707

Desvio = 2067837

Figura A3.15 – *Box-Plot* do Nó 27, variável "Total artigos devolvidos"



Nota: esta análise não revelou a existência de novos *outliers* extremos, apenas os já conhecidos.

Regras de Decisão do Nó 30

SE Unidade Negócio = Casa, Higiene, Lazer, Mercearia Doce, Mercearia Salgada

E Loja = AG, AH, B, C, D, L, Q

E Dia semana = Dias úteis

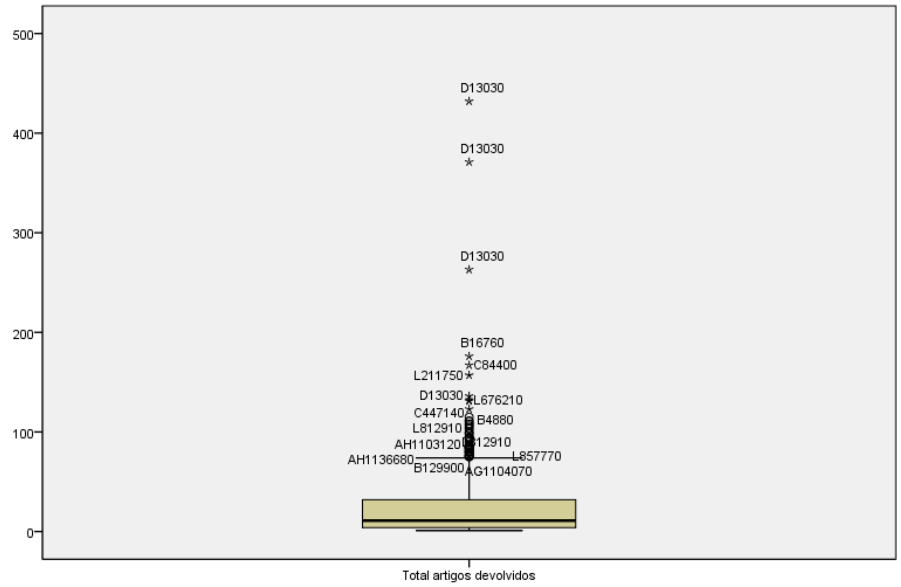
E Horario registo = Manhã, Noite

ENTÃO Total artigos devolvidos **=** 23,18

N = 953

Desvio = 993667,6

Figura A3.16 – *Box-Plot* do Nó 30, variável "Total artigos devolvidos"



Nota: esta análise não revelou a existência de novos *outliers* extremos significativos, apenas os já conhecidos.

Regras de Decisão do Nó 31

SE Unidade Negócio = Casa, Higiene, Lazer, Mercearia Doce, Mercearia Salgada

E Loja = AG, AH, B, C, D, L, Q

E Dia semana = Dias úteis

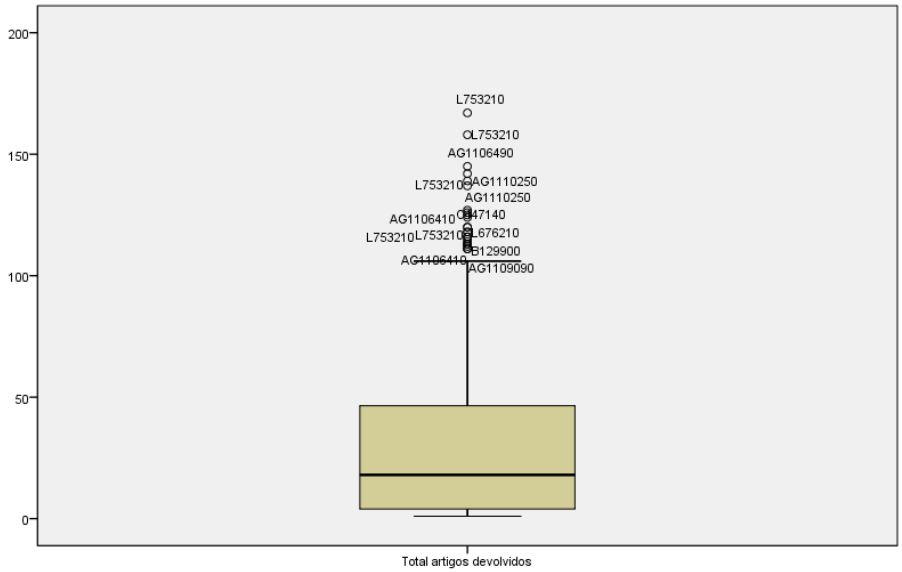
E Horario registo = Tarde

ENTÃO Total artigos devolvidos **=** 29,7

N = 671

Desvio = 686917,2

Figura A3.17 – *Box-Plot* do Nó 31, variável "Total artigos devolvidos"



Nota: esta análise não identificou a presença de *outliers* extremos.

Regras de Decisão do Nó 47

SE Unidade Negócio = Bebidas, Bricolage&Auto, Charcutaria&Queijos, Cultura, Frutas&Legumes, Lacticinios, Limpeza

E Dia Semana = Dias úteis

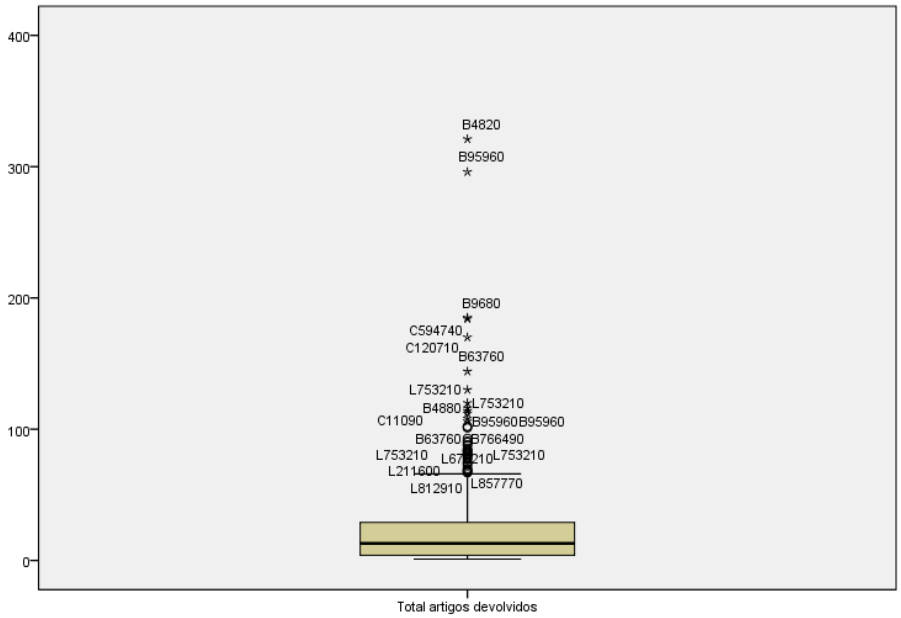
E Loja = B, C, L

ENTÃO Total artigos devolvidos **=** 20,74

N = 1023

Desvio = 696059,3

Figura A3.18 – *Box-Plot* do Nó 47, variável "Total artigos devolvidos"



Nota: esta análise não revelou a existência de novos *outliers* extremos significativos, apenas os já conhecidos.

Regras de Decisão do Nó 54

SE Unidade Negócio = Casa, Higiene, Lazer, Mercearia Doce, Mercearia Salgada

E Dia Semana = Dias úteis

E Loja = A, AB, AD, AE, AF, E, F, G, H, I, J, K, N, W, X, Y, Z

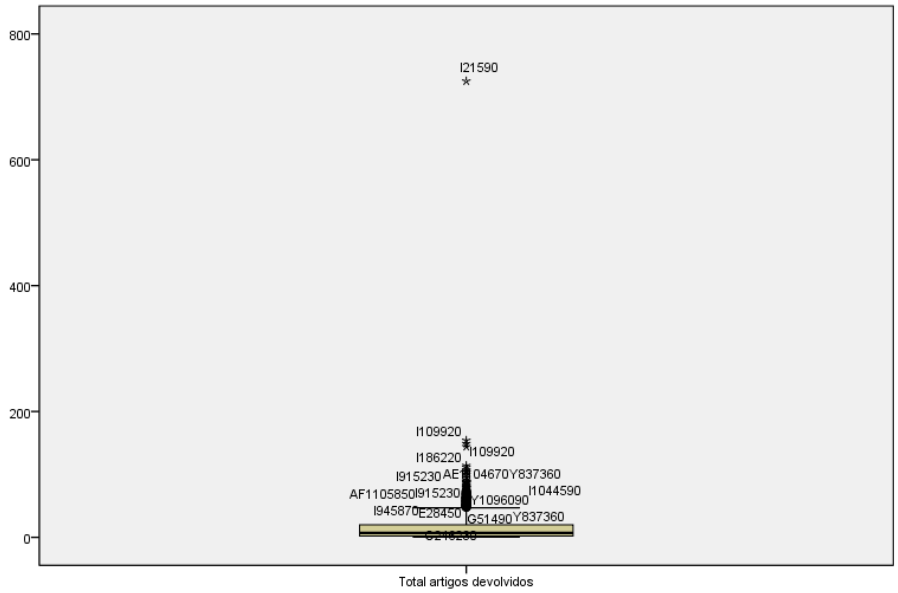
E Horario registo = Manha, Noite

ENTÃO Total artigos devolvidos **=** 14,24

N = 2176

Desvio = 1158722

Figura A3.19 – *Box-Plot* do Nó 54, variável "Total artigos devolvidos"



Nota: esta análise não revelou a existência de novos *outliers* extremos significativos, apenas os já conhecidos.

Regras de Decisão do Nó 63

SE Unidade Negócio = Casa, Higiene, Lazer, Mercearia Doce, Mercearia Salgada

E Dia semana = Dias úteis

E Horario registo = Tarde

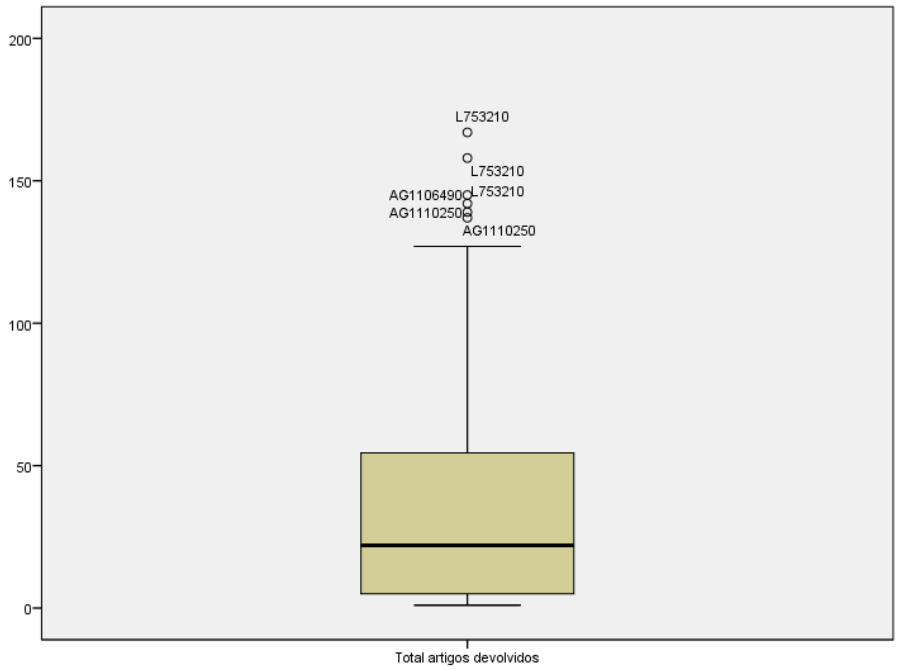
E Loja = AG, B, C, L, Q

ENTÃO Total artigos devolvidos **=** 34,01

N = 423

Desvio = 505438,9

Figura A3.20 – *Box-Plot* do Nó 63, variável "Total artigos devolvidos"



Nota: esta análise não identificou a presença de *outliers* extremos.

Regras de Decisão do Nó 94

SE Dia Semana = Dias úteis

E Loja = B, C, L

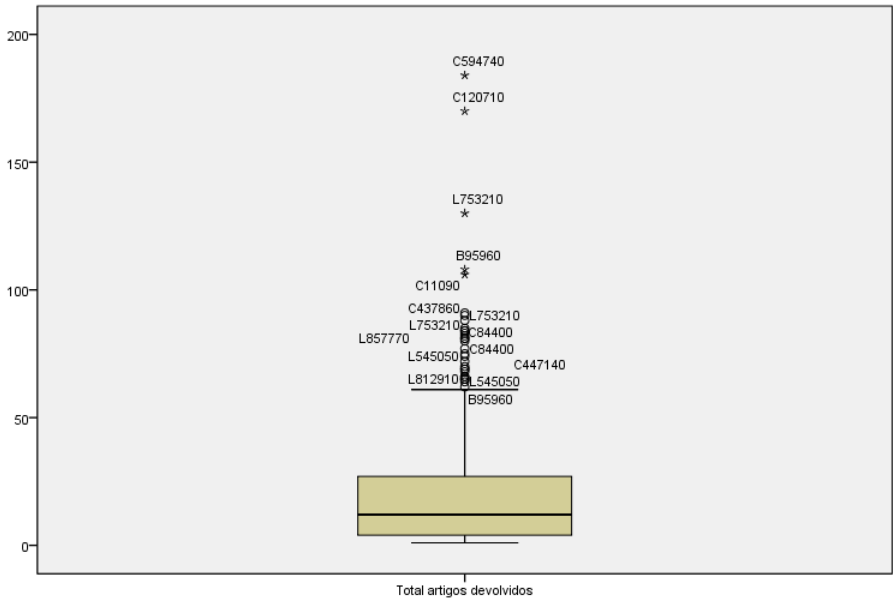
E Unidade Negócio = Bebidas, Bricolage&Auto, Charcutaria&Queijos, Cultura, Limpeza

ENTÃO Total artigos devolvidos **=** 19

N = 713

Desvio = 323440

Figura A3.21 – *Box-Plot* do Nó 94, variável "Total artigos devolvidos"



Nota: esta análise não revelou a existência de novos *outliers* extremos significativos, apenas os já conhecidos.

Regras de Decisão do Nó 95

SE Dia Semana = Dias úteis

E Loja = B, C, L

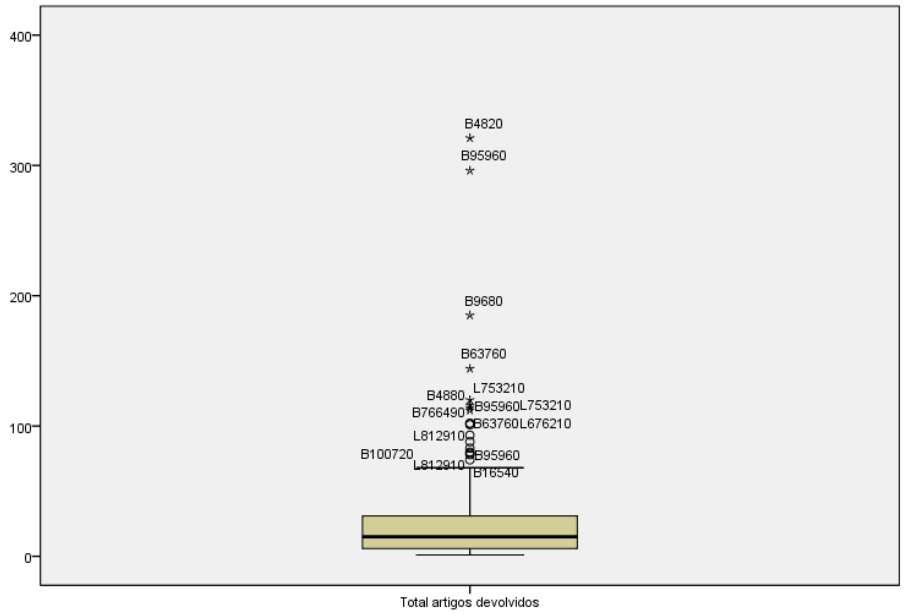
E Unidade Negócio = Frutas&Legumes, Lacticinios

ENTÃO Total artigos devolvidos **=** 24,73

N = 310

Desvio = 365549,7

Figura A3.22 – *Box-Plot* do Nó 95, variável "Total artigos devolvidos"



Nota: esta análise não revelou a existência de novos *outliers* extremos significativos, apenas os já conhecidos.

Regras de Decisão do Nó 190

SE Dia Semana = Dias úteis

E Unidade Negócio = Frutas&Legumes, Lacticinios

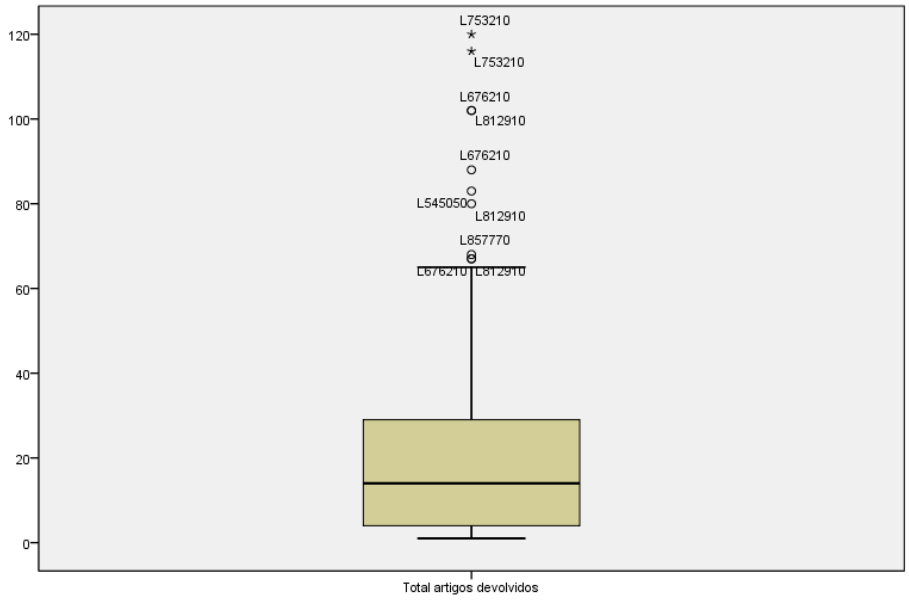
E Loja = C, L

ENTÃO Total artigos devolvidos **=** 20,43

N = 205

Desvio = 100228,4

Figura A3.23 – *Box-Plot* do Nó 190, variável "Total artigos devolvidos"



Nota: esta análise não revelou a existência de novos *outliers* extremos, apenas os já conhecidos.

Regras de Decisão do Nó 191

SE Dia Semana = Dias úteis

E Unidade Negócio = Frutas&Legumes, Lacticinios

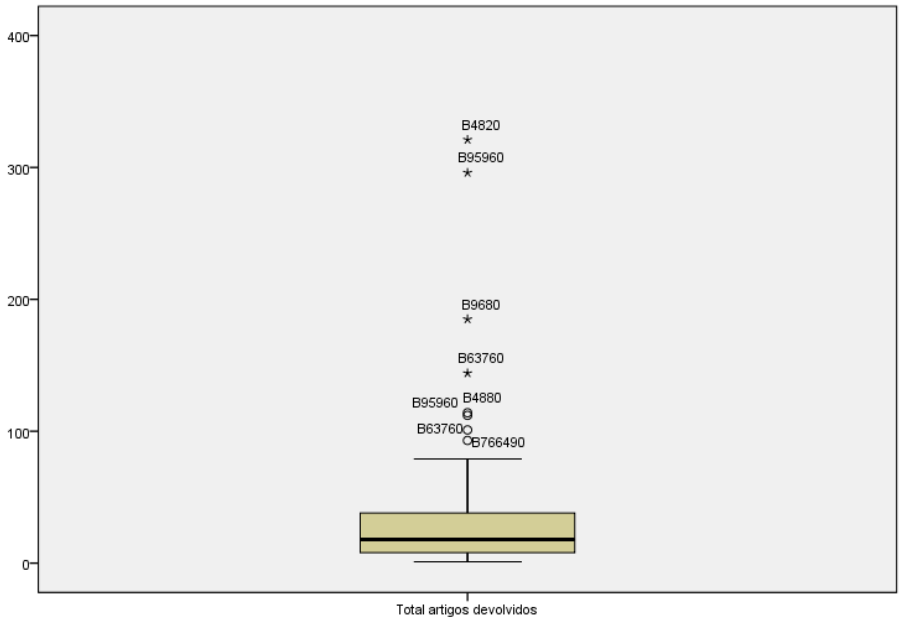
E Loja = B

ENTÃO Total artigos devolvidos **=** 33,1

N = 105

Desvio = 254173,8

Figura A3.24 – *Box-Plot* do Nó 191, variável "Total artigos devolvidos"



Nota: esta análise não revelou a existência de novos *outliers* extremos, apenas os já conhecidos.

Regras de Decisão do Nó 382

SE Dia Semana = Dias úteis

E Unidade Negócio = Frutas&Legumes, Lacticinios

E Loja = B

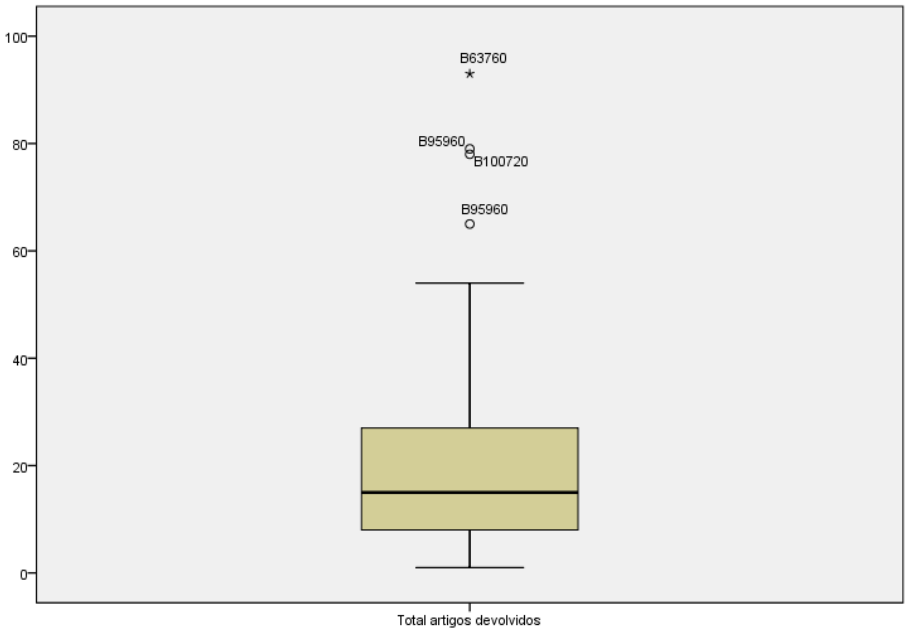
E Horario registo = Tarde, Noite

ENTÃO Total artigos devolvidos **=** 20,77

N = 73

Desvio = 27327,04

Figura A3.25 – *Box-Plot* do Nó 382, variável "Total artigos devolvidos"



Nota: esta análise não revelou a existência de novos *outliers* extremos, apenas os já conhecidos.

Regras de Decisão do Nó 383

SE Dia Semana = Dias úteis

E Unidade Negócio = Frutas&Legumes, Lacticinios

E Loja = B

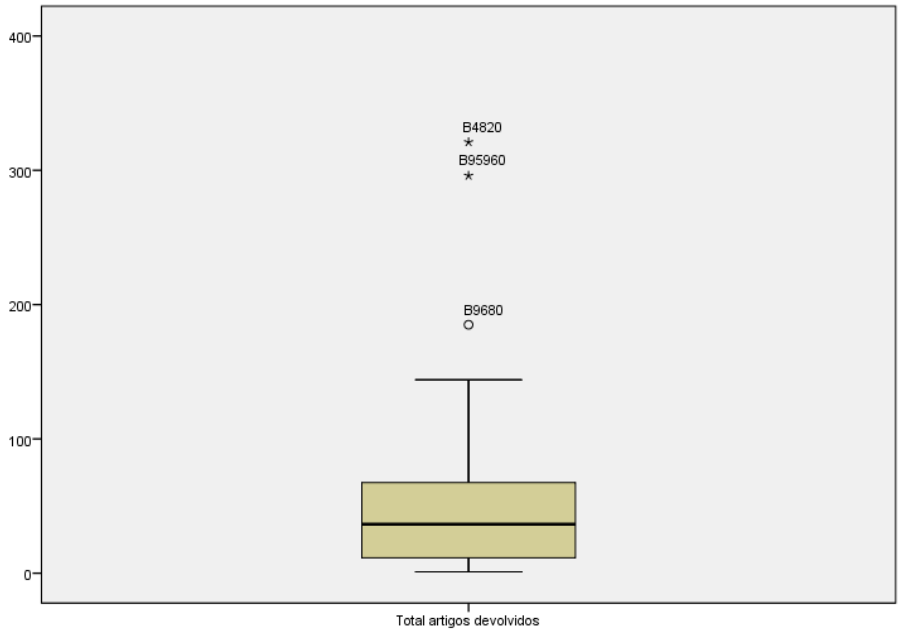
E Horario registo = Manhã

ENTÃO Total artigos devolvidos **=** 61,25

N = 32

Desvio = 190386

Figura A3.26 – *Box-Plot* do Nó 383, variável "Total artigos devolvidos"



Nota: esta análise não revelou a existência de novos *outliers* extremos, apenas os já conhecidos.

Regras de Decisão do Nó 766

SE Dia Semana = Dias úteis

E Loja = B

E Horario registo = Manhã

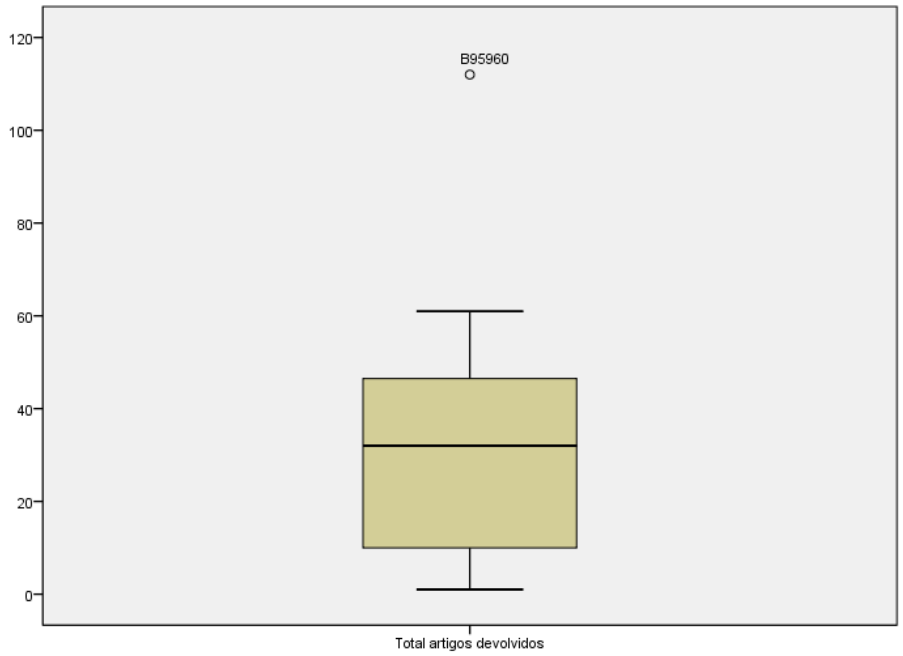
E Unidade Negócio = Frutas&Legumes

ENTÃO Total artigos devolvidos **=** 33,69

N = 16

Desvio = 12311,44

Figura A3.27 – *Box-Plot* do Nó 766, variável "Total artigos devolvidos"



Nota: esta análise não identificou a presença de *outliers* extremos.

Regras de Decisão do Nó 767

SE Dia Semana = Dias úteis

E Loja = B

E Horario registo = Manhã

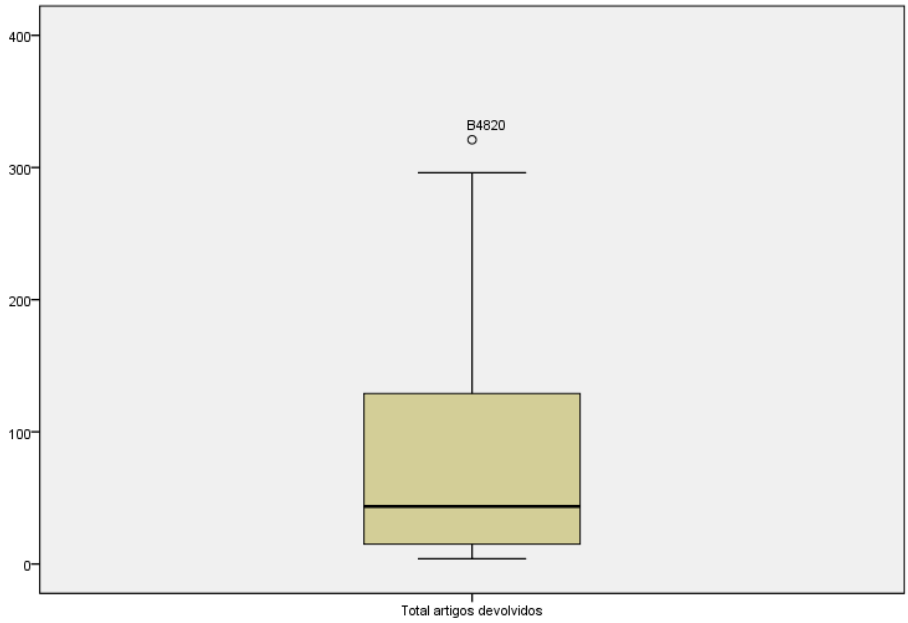
E Unidade Negócio = Lacticinios

ENTÃO Total artigos devolvidos **=** 88,81

N = 16

Desvio = 153764,4

Figura A3.28 – *Box-Plot* do Nó 767, variável "Total artigos devolvidos"



Nota: esta análise não identificou a presença de *outliers* extremos.

ANEXO IV - Análises multivariadas: variável objetivo "Valor médio artigos devolvidos"

Regras de Decisão do Nó 2

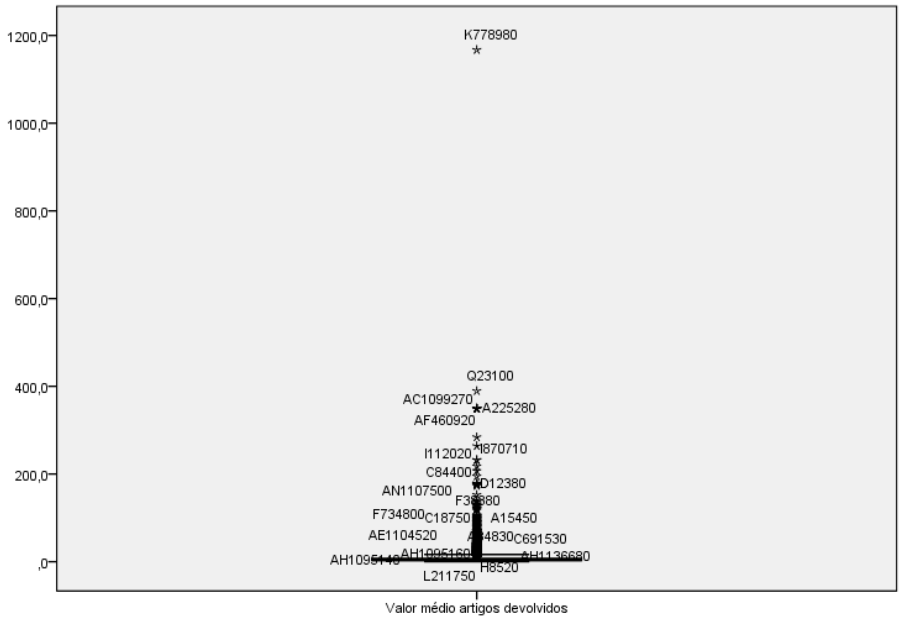
SE Unidade Negócio = Bebidas, Casa, Charcutaria&Queijos, Congelados, Cultura, Frutas&Legumes, Higiene, Lacticinios, Limpeza, Mercearia Doce, Mercearia Salgada, Padaria, Peixaria, Pets&Plants, TakeAway, Talho

ENTÃO Valor médio artigos devolvidos **=** 6,47€

N = 38536

Desvio = 5179937

Figura A4.1 – *Box-Plot* do Nó 2, variável "Valor médio artigos devolvidos"



Nota: esta análise não revelou a existência de novos *outliers* extremos significativos, apenas os já conhecidos.

Regras de Decisão do Nó 3

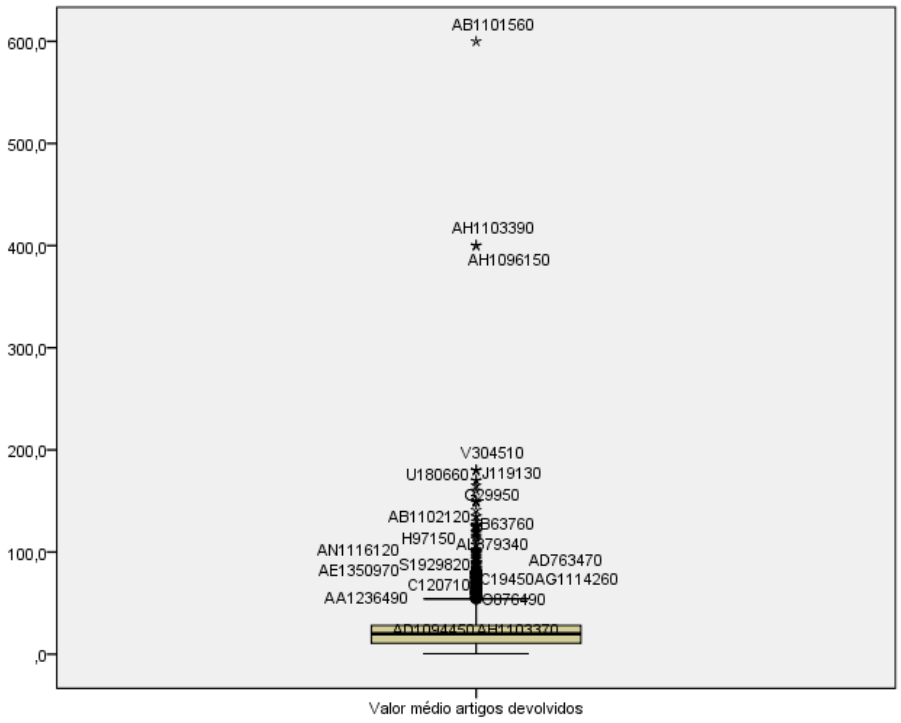
SE Unidade Negócio = Bricolage&Auto, Lazer

ENTÃO Valor médio artigos devolvidos = 23,04€

N = 4670

Desvio = 2224964

Figura A4.2 – *Box-Plot* do Nó 3, variável "Valor médio artigos devolvidos"



Nota: esta análise não revelou a existência de novos *outliers* extremos significativos, apenas os já conhecidos.

Regras de Decisão do Nó 4

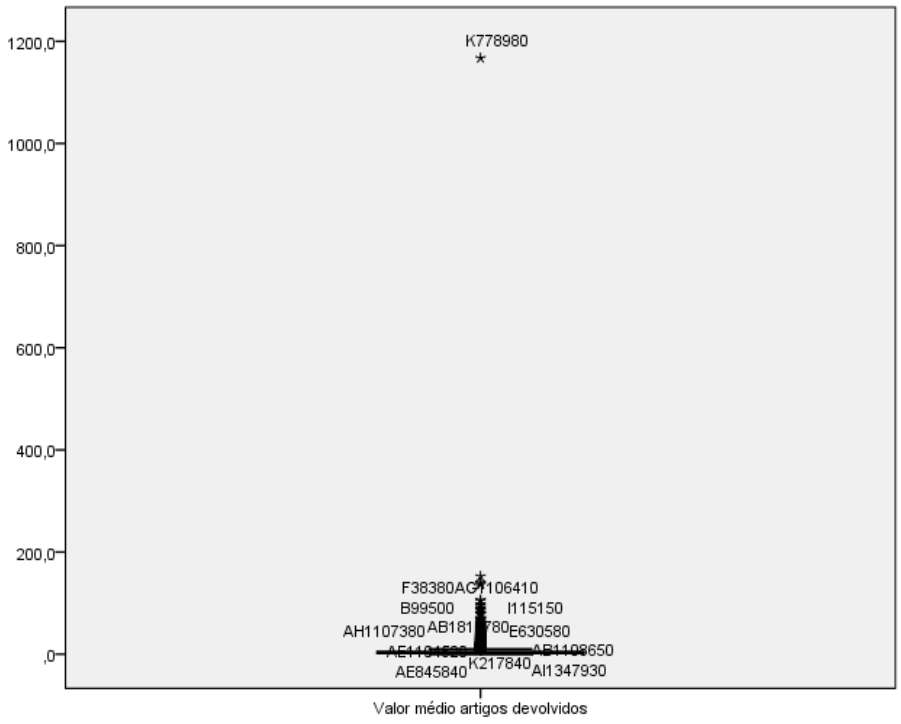
SE Unidade Negócio = Charcutaria&Queijos, Congelados, Cultura, Frutas&Legumes, Lacticinios, Limpeza, Mercearia Doce, Mercearia Salgada, Padaria, TakeAway, Talho

ENTÃO Valor médio artigos devolvidos = 4,47€

N = 26282

Desvio = 1993308

Figura A4.3 – *Box-Plot* do Nó 4, variável "Valor médio artigos devolvidos"



Nota: esta análise não revelou a existência de novos *outliers* extremos significativos, apenas os já conhecidos.

Regras de Decisão do Nó 7

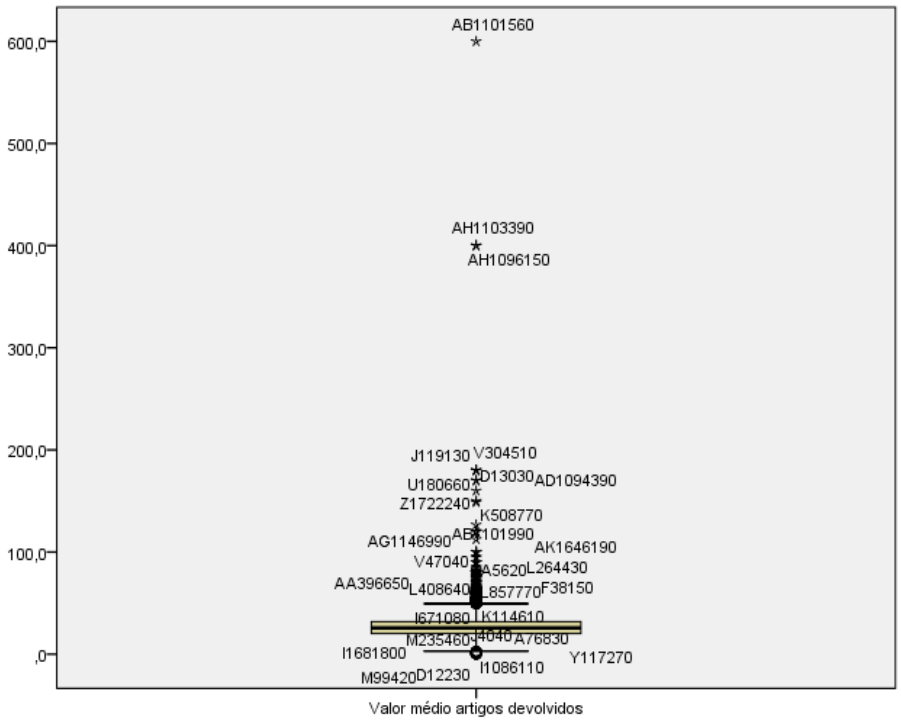
SE Unidade Negócio = Lazer

ENTÃO Valor médio artigos devolvidos = 29,21€

N = 2344

Desvio = 1264886

Figura A4.4 – *Box-Plot* do Nó 7, variável "Valor médio artigos devolvidos"



Nota: esta análise não revelou a existência de novos *outliers* extremos significativos, apenas os já conhecidos.

Regras de Decisão do Nó 8

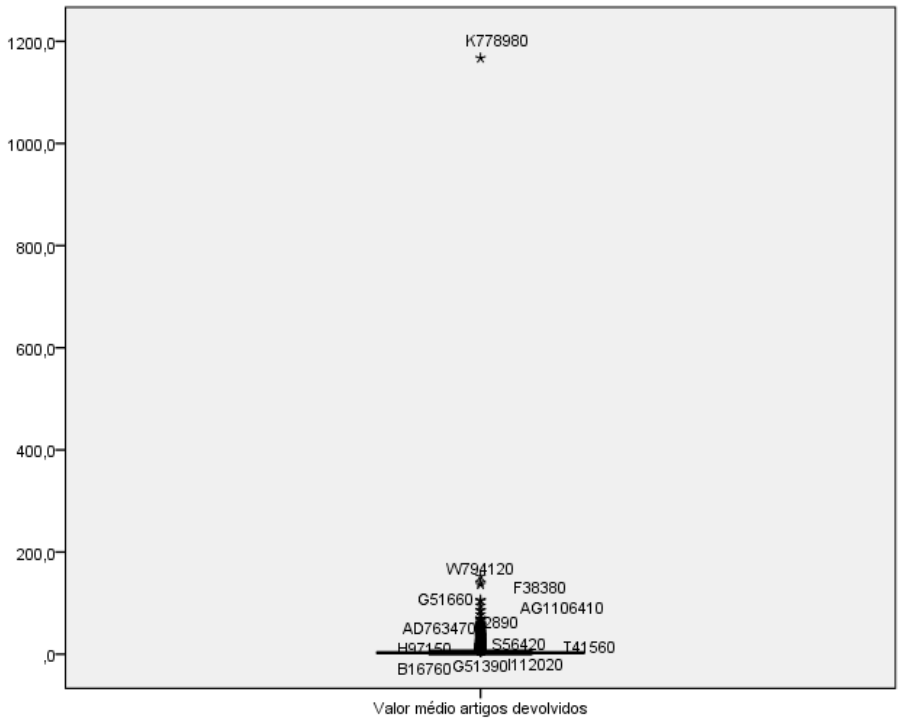
SE Unidade Negócio = Charcutaria&Queijos, Congelados, Frutas&Legumes, Lacticinios, Mercearia Doce, Mercearia Salgada, Padaria, TakeAway

ENTÃO Valor médio artigos devolvidos **=** 3,63€

N = 18938

Desvio = 1720307

Figura A4.5 – *Box-Plot* do Nó 8, variável "Valor médio artigos devolvidos"



Nota: esta análise não revelou a existência de novos *outliers* extremos significativos, apenas os já conhecidos.

Regras de Decisão do Nó 17

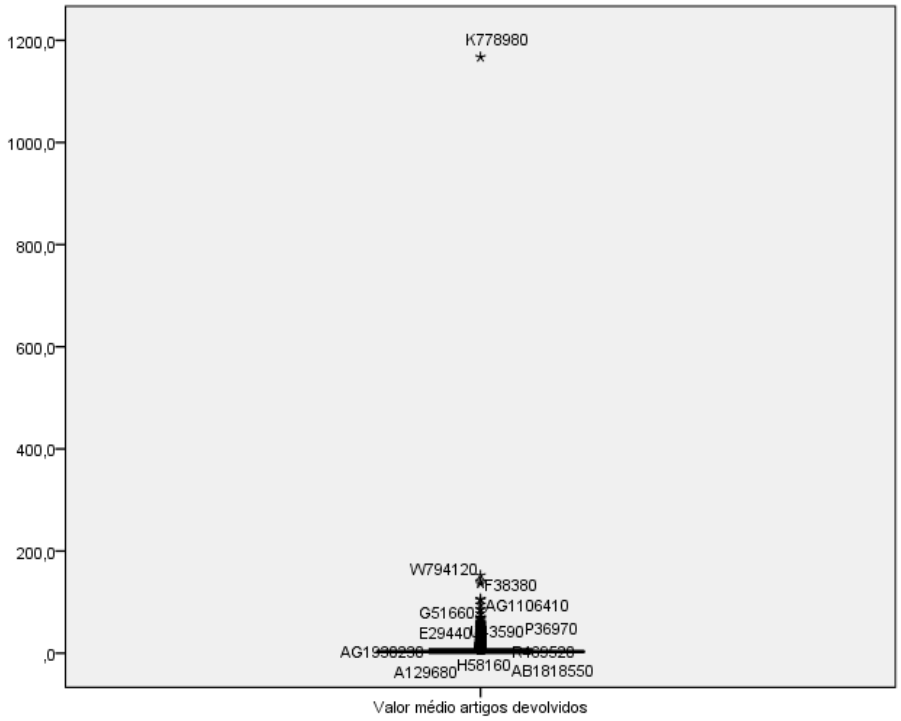
SE Unidade Negócio = Charcutaria&Queijos, Congelados, Lacticinios, Mercearia Doce, Mercearia Salgada, TakeAway

ENTÃO Valor médio artigos devolvidos = 4,01€

N = 14001

Desvio = 1688006

Figura A4.6 – *Box-Plot* do Nó 17, variável "Valor médio artigos devolvidos"



Nota: esta análise não revelou a existência de novos *outliers* extremos significativos, apenas os já conhecidos.

Regras de Decisão do Nó 31

SE Unidade Negócio = Lazer

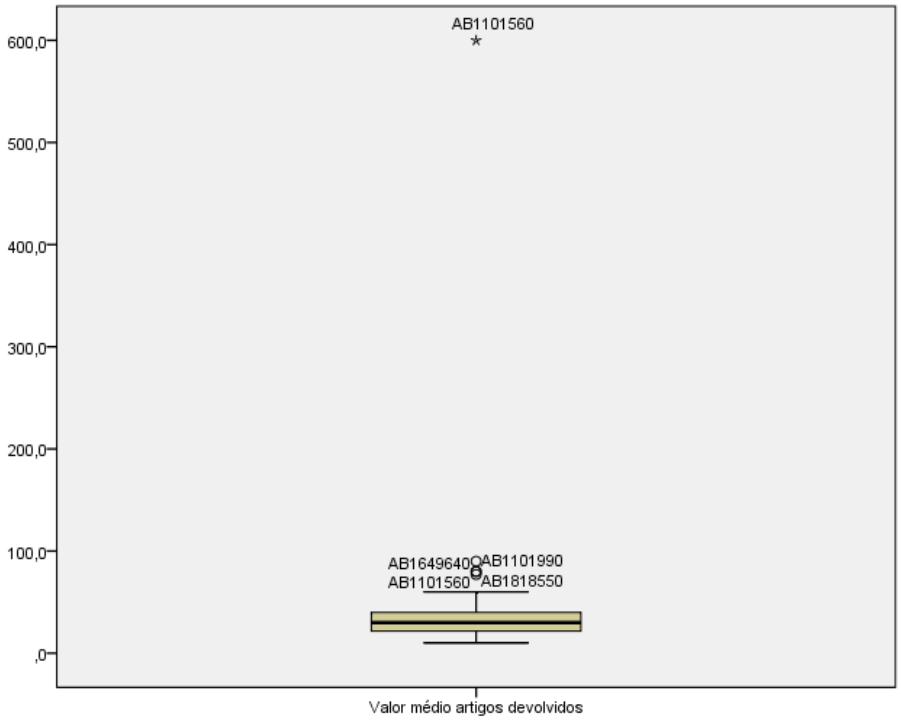
E Loja = AB

ENTÃO Valor médio artigos devolvidos **=** 46,61€

N = 45

Desvio = 329036,5

Figura A4.7 – *Box-Plot* do Nó 31, variável "Valor médio artigos devolvidos"



Nota: esta análise não revelou a existência de novos *outliers* extremos, apenas o já conhecido.

Regras de Decisão do Nó 35

SE Unidade Negócio = Charcutaria&Queijos, Congelados, Lacticinios, Mercearia Doce, Mercearia Salgada, TakeAway

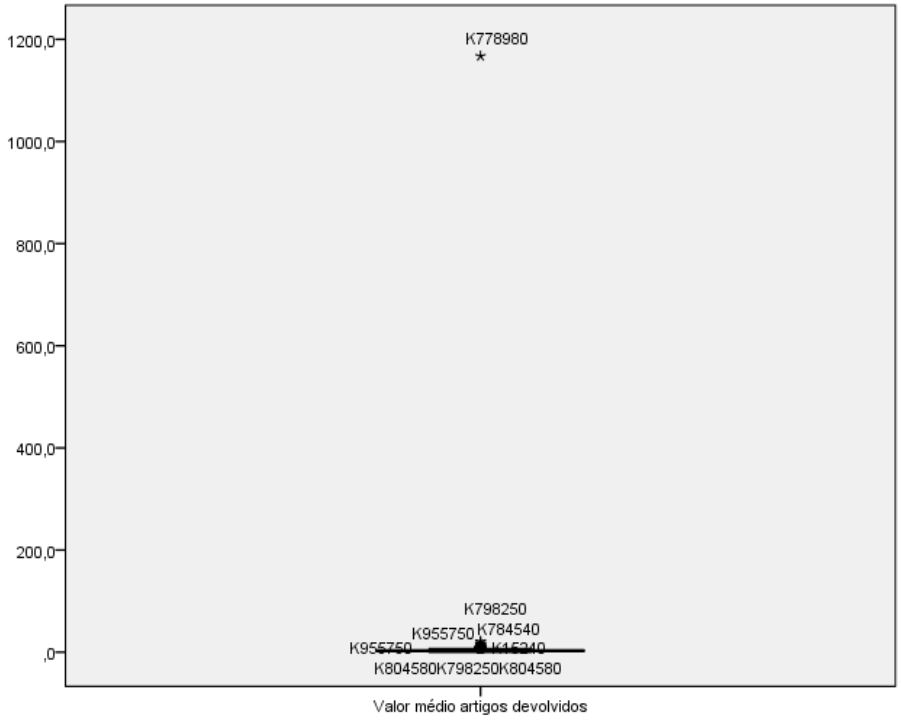
E Loja = K

ENTÃO Valor médio artigos devolvidos = 6,48€

N = 393

Desvio = 1353429

Figura A4.8 – *Box-Plot* do Nó 35, variável "Valor médio artigos devolvidos"



Nota: esta análise não revelou a existência de novos *outliers* extremos significativos, apenas os já conhecidos.

Regras de Decisão do Nó 63

SE Unidade Negócio = Lazer

E Loja = AB

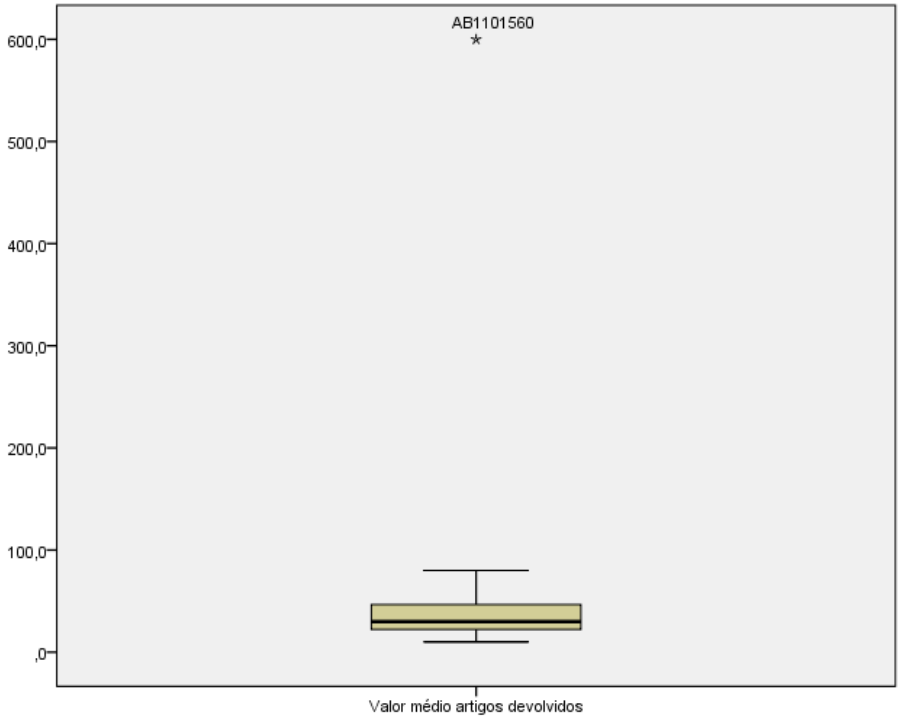
E Dia semana = Fim de semana ou feriado

ENTÃO Valor médio artigos devolvidos **=** 62,92€

N = 20

Desvio = 310727

Figura A4.9 – *Box-Plot* do Nó 63, variável "Valor médio artigos devolvidos"



Nota: esta análise não revelou a existência de novos *outliers* extremos, apenas o já conhecido.

Regras de Decisão do Nó 71

SE Loja = K

E Unidade Negócio = Lacticinios

ENTÃO Valor médio artigos devolvidos = 18,67€

N = 77

Desvio = 1336895

Figura A4.10 – *Box-Plot* do Nó 71, variável "Valor médio artigos devolvidos"

Nota: esta análise não revelou a existência de novos *outliers* extremos significativos, apenas os já conhecidos.

Regras de Decisão do Nó 123

SE Unidade Negócio = Lazer

E Horario registo = Manhã

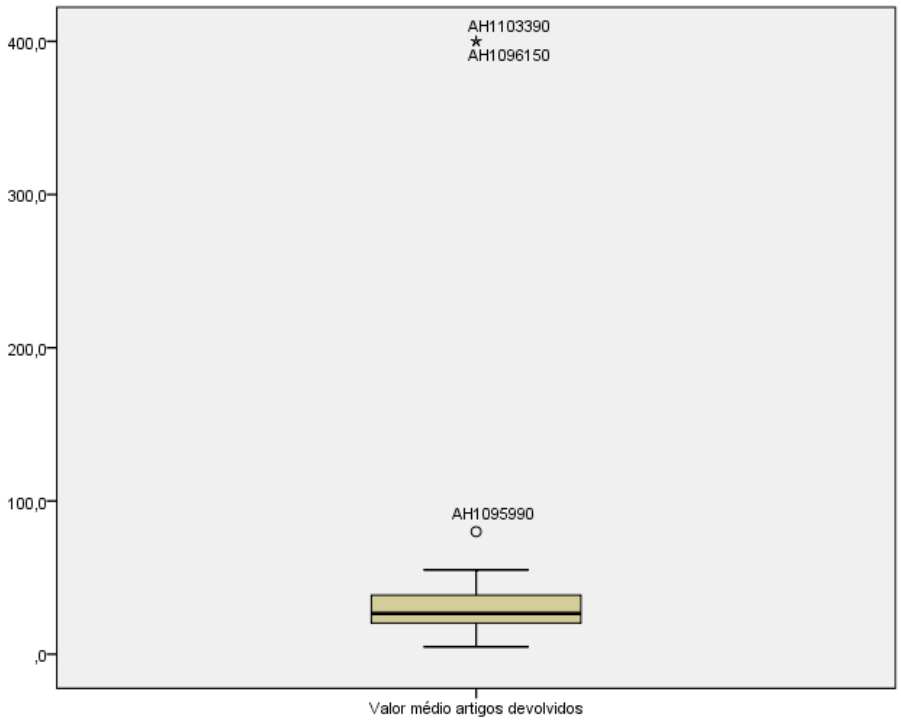
E Loja = AH

ENTÃO Valor médio artigos devolvidos **=** 59,59€

N = 24

Desvio = 258218,5

Figura A4.11 – *Box-Plot* do Nó 123, variável "Valor médio artigos devolvidos"



Nota: esta análise não revelou a existência de novos *outliers* extremos significativos, apenas os já conhecidos.

Regras de Decisão do Nó 126

SE Unidade Negócio = Lazer

E Loja = AB

E Dia semana = Fim de semana ou feriado

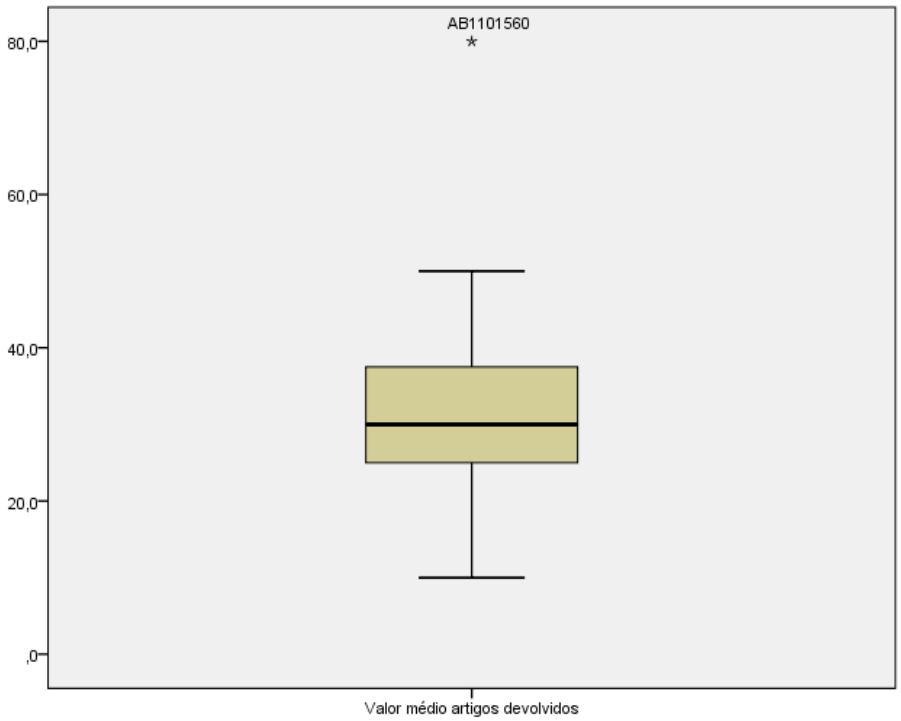
E Horario registo = Manhã, Noite

ENTÃO Valor médio artigos devolvidos **=** 33,82€

N = 10

Desvio = 3445,515

Figura A4.12 – *Box-Plot* do Nó 126, variável "Valor médio artigos devolvidos"



Nota: esta análise não revelou a existência de novos *outliers* extremos significativos, apenas os já conhecidos.

Regras de Decisão do Nó 127

SE Unidade Negócio = Lazer

E Loja = AB

E Dia semana = Fim de semana ou feriado

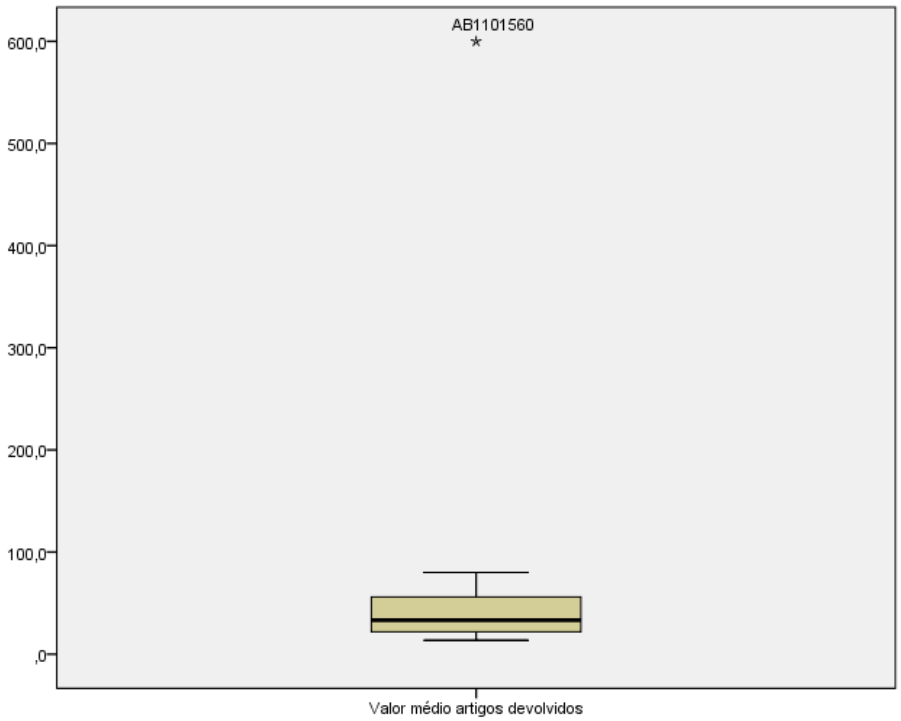
E Horario registo = Tarde

ENTÃO Valor médio artigos devolvidos **=** 92,02€

N = 10

Desvio = 290344,1

Figura A4.13 – *Box-Plot* do Nó 127, variável "Valor médio artigos devolvidos"



Nota: esta análise não revelou a existência de novos *outliers* extremos significativos, apenas os já conhecidos.

Regras de Decisão do Nó 142

SE Loja = K

E Unidade Negócio = Lacticinios

E Horario registo = Manhã, Noite

ENTÃO Valor médio artigos devolvidos = 3,53€

N = 48

Desvio = 249,3271

Figura A4.14 – *Box-Plot* do Nó 142, variável "Valor médio artigos devolvidos"

Nota: esta análise não identificou *outliers* extremos.

Regras de Decisão do Nó 143

SE Loja = K

E Unidade Negócio = Lacticinios

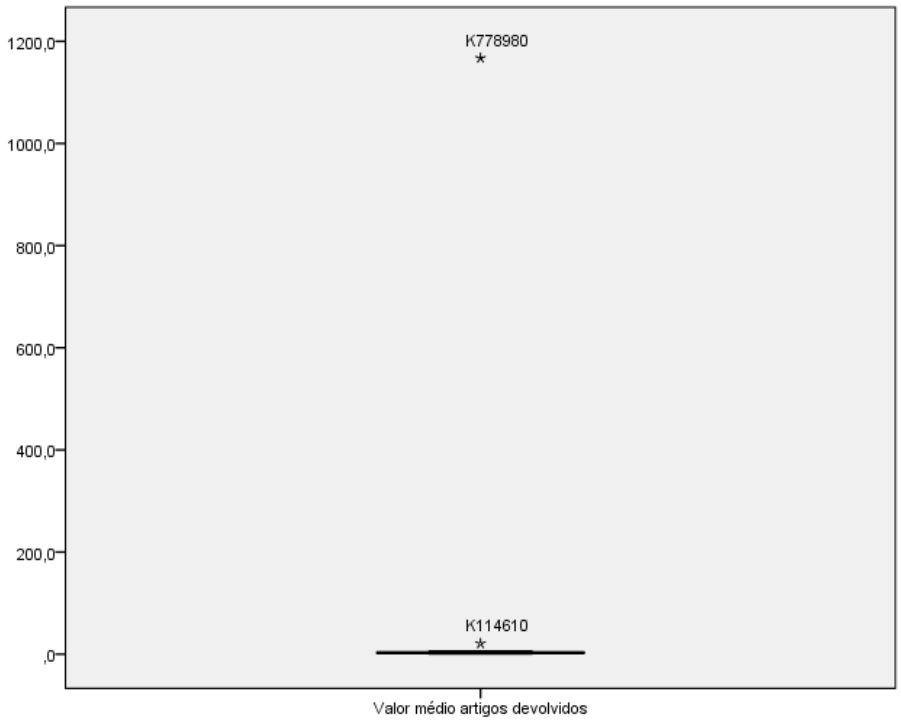
E Horario registo = Tarde

ENTÃO Valor médio artigos devolvidos = 43,74€

N = 29

Desvio = 1307423

Figura A4.15 – *Box-Plot* do Nó 143, variável "Valor médio artigos devolvidos"



Nota: esta análise não revelou a existência de novos *outliers* extremos significativos, apenas os já conhecidos.

Regras de Decisão do Nó 246

SE Unidade Negócio = Lazer

E Horario registo = Manhã

E Loja = AH

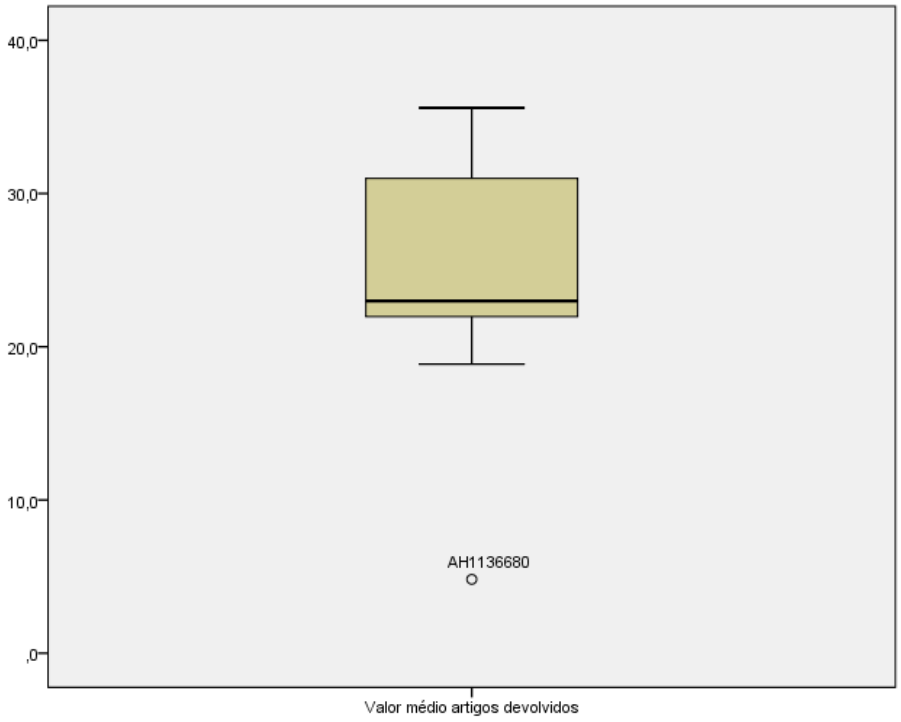
E Dia semana = Fim de semana ou feriado

ENTÃO Valor médio artigos devolvidos **=** 24,07€

N = 9

Desvio = 658, 4663

Figura A4.16 – *Box-Plot* do Nó 246, variável "Valor médio artigos devolvidos"



Nota: esta análise não identificou *outliers* extremos.

Regras de Decisão do Nó 247

SE Unidade Negócio = Lazer

E Horario registo = Manhã

E Loja = AH

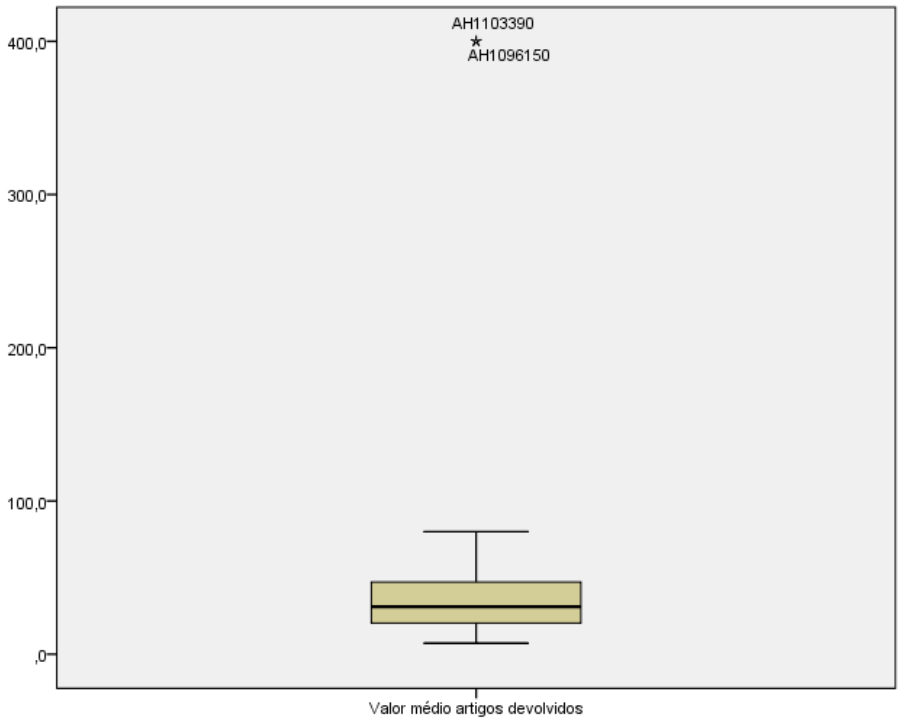
E Dia semana = Dias úteis

ENTÃO Valor médio artigos devolvidos **=** 80,90€

N = 15

Desvio = 239393,9

Figura A4.17 – *Box-Plot* do Nó 247, variável "Valor médio artigos devolvidos"



Nota: esta análise não revelou a existência de novos *outliers* extremos significativos, apenas os já conhecidos.

Regras de Decisão do Nó 287

SE Loja = K

E Unidade Negócio = Lacticinios

E Horario registo = Tarde

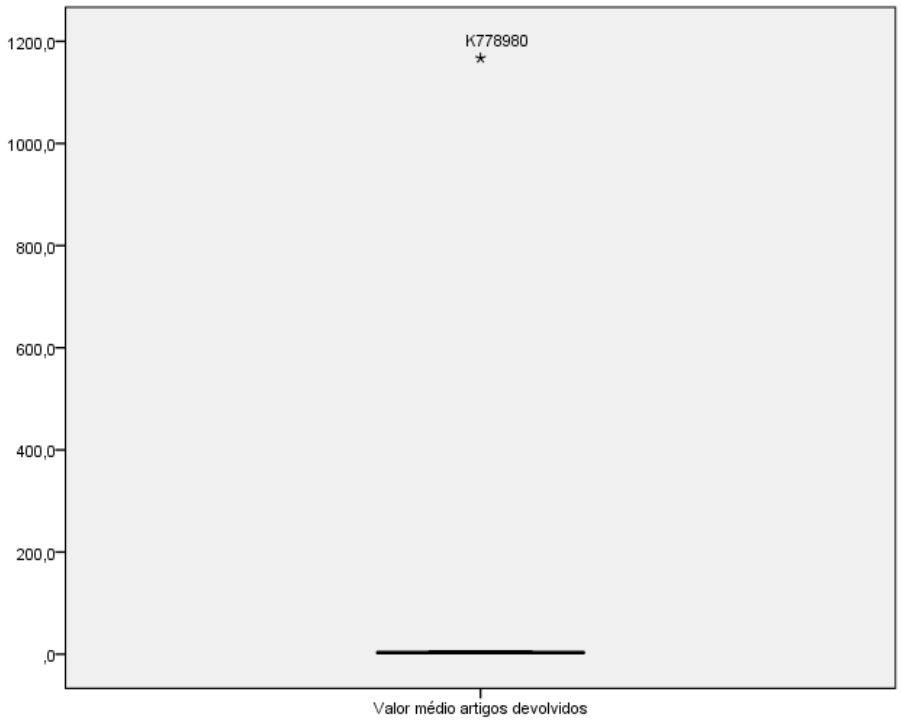
E Dia semana = Dias úteis

ENTÃO Valor médio artigos devolvidos = 86,22€

N = 14

Desvio = 1258243

Figura A4.18 – *Box-Plot* do Nó 287, variável "Valor médio artigos devolvidos"



Nota: esta análise não revelou a existência de novos *outliers* extremos significativos, apenas o já conhecido.