NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTAMENTO DE MATEMÁTICA

MARTA REIS DE CARVALHO

Licenciada em Matemática

ESTUDO DO PERFIL DE CLIENTE DE UMA COMPANHIA DE SEGUROS

ANÁLISE DE CLUSTERS, CÁLCULO DO VALOR DO CLIENTE E PREVISÃO DA PROBABILIDADE DE CHURN PARA O DESENVOLVIMENTO DE CAMPANHAS DIRECIONADAS

MESTRADO EM MATEMÁTICA E APLICAÇÕES

Universidade NOVA de Lisboa
2023, Novembro

NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTAMENTO DE MATEMÁTICA

ESTUDO DO PERFIL DE CLIENTE DE UMA COMPANHIA DE SEGUROS

ANÁLISE DE CLUSTERS, CÁLCULO DO VALOR DO CLIENTE E PREVISÃO DA PROBABILIDADE DE CHURN PARA O DESENVOLVIMENTO DE CAMPANHAS DIRECIONADAS

MARTA REIS DE CARVALHO

Licenciada em Matemática

Orientadora: Ayana Maria Xavier Furtado Mateus
Professora Auxiliar da Faculdade de Ciências e Tecnologia da Universidade NOVA de Lisboa

Coorientador: Frederico Almeida Gião Gonçalves Caeiro
Professor Associado da Faculdade de Ciências e Tecnologia da Universidade NOVA de Lisboa

Júri

Presidente: Doutora Marta Cristina Vieira Faias Mateus
Professora Associada da Faculdade de Ciências e Tecnologia da Universidade NOVA de Lisboa

Arguente: Doutor Miguel dos Santos Fonseca
Professor Auxiliar da Faculdade de Ciências e Tecnologia da Universidade NOVA de Lisboa

Orientadora: Doutora Ayana Maria Xavier Furtado Mateus
Professora Auxiliar da Faculdade de Ciências e Tecnologia da Universidade NOVA de Lisboa

MESTRADO EM MATEMÁTICA E APLICAÇÕES
ESPECIALIDADE EM CIÊNCIA DOS DADOS E DA DECISÃO

Universidade NOVA de Lisboa
2023, Novembro

Estudo do Perfil de Cliente de uma companhia de seguros

Este documento foi gerado com o processador (pdf/Xe/Lua)LaTeX e o modelo NOVAthesis (v6.10.5) Lourenço, 2021.

Agradecimentos

Primeiramente, gostaria de expressar a minha gratidão ao Grupo Future Healhtcare por me ter concebido esta oportunidade. Agradeço, em particular, aos colaboradores envolvidos neste projeto pela ajuda e incentivo na procura de fazer mais e melhor na realização da presente dissertação.

Agradecer igualmente aos meus orientadores, à Professora Ayana Mateus e ao Professor Frederico Caeiro.

Não podia deixar de agradecer a todos os professores, amigos e colegas que me acompanharam ao longo do meu percurso escolar e académico. Agradeço também à FCT NOVA pelos 5 anos do meu percurso académico. Guardarei sempre um carinho por esta instituição, por todos os momentos e aprendizagens que me proporcionou.

Por fim, mas não menos importante, tenho que agradecer à minha família, que tornou tudo isto possível. Em especial, aos meus pais e à minha irmã que sempre me apoiaram e incentivaram não somente durante a fase da dissertação, mas também durante toda a minha formação. Sem vocês, não teria chegado até aqui. Agradeço imensamente pelo acompanhamento diário e incondicional.

Resumo

Ao longo das últimas décadas, verificou-se que o cliente, bem como as suas necessidades e a sua satisfação, são cada vez mais o foco das empresas. Tentar satisfazer e reter os clientes mais lucrativos para as companhias é um tópico com bastante interesse que tem ganho relevância nos últimos tempos.

Desta forma, e tratando-se o Grupo Future Healthcare de uma empresa que comercializa seguros e planos de saúde, é de extrema importância para a companhia estar consciente do valor que cada cliente representa, assim como perceber qual a influência que as suas estratégias de marketing têm na retenção dos seus clientes.

Assim, esta dissertação pretende numa primeira fase identificar os clientes mais lucrativos para a empresa e caracterizá-los, com recurso à análise Recência, Frequência e Valor Monetário (RFM).

Posteriormente, o objetivo passa pela segmentação dos clientes com base no algoritmo agregação por K-Médios (*K-Means*), por forma a obter grupos de clientes com comportamentos semelhantes, e daí concluir sobre o perfil do cliente, e quais as campanhas mais eficazes para cada grupo.

Por fim, recorre-se à regressão logística para averiguar sobre a taxa de retenção, ou seja, pretende-se analisar que grupos são mais suscetíveis de abandonar a empresa, de modo a direcionar as estratégias desenvolvidas e a melhorar a satisfação destes clientes.

Palavras-chave: Análise RFM, Segmentação de Clientes, Análise de Clusters, Taxa de retenção

Abstract

Over the last decades, it has been seen that more and more the customer is the focus, as well as their needs and satisfaction.Trying to satisfy and retain the most profitable customers for companies is a topic with a lot of interest that has been gaining relevance in recent times.

Therefore, as Future Healthcare Group is a company that, among many other things, sells insurance and health plans, it is extremely important for the company to be aware of the value that each client represents, as well as to understand the influence that its marketing strategies have on client retention.

Thus, this dissertation aims, in a first phase, to identify the most profitable customers for the company and characterise them using Recency, Frequency and Monetary Value (RFM) analysis.

Subsequently, the goal is to segment customers based on the *K-Means* algorithm, in order to obtain groups of customers with similar behaviour, and then conclude on the customer profile and which campaigns are most effective for each group.

Finally, logistic regression is used to find out about the retention rate, that is, to discover which groups are more likely to leave the company, so as to direct the strategies developed and improve customer satisfaction.

Keywords: RFM Analysis, Customer Segmentation, Cluster Analysis, Retention Rate

Índice

Índice de Figuras

Índice de Tabelas

Glossário

Churn Rotatividade de clientes ou *Customer Churn* é a percentagem de clientes que deixaram de utilizar um produto ou serviço da empresa *(p. 3)*

Siglas

AHP Analytic Hierarchy Process *(pp. 10, 11, 64)*

CRM Gestão da Relação com o Cliente *(pp. xiii, xv, 5, 7, 12–16)*

DM Data Mining *(pp. xiii, xv, 7, 13–15)*

FH Future Healthcare *(pp. 2, 3, 6, 27)*

RFM Recência, Frequência e Valor Monetário *(pp. vii, xv, 3, 7–11, 16, 17, 21, 22, 61, 63, 64, 67)*

WRFM Análise RFM ponderada *(pp. 10, 11, 22, 64)*

WSS Within Sum of Squares *(pp. 24, 25)*

Capítulo

1 Introdução

A ocorrência de acontecimentos num futuro longínquo ou muito próximo, como daqui a um mês é um perfeito enigma para a Humanidade. A necessidade de proteção e garantia contra tais acontecimentos é uma real preocupação que remota desde a antiguidade, sensivelmente desde há 5000 anos a.C.

Tal como Albert Einstein uma vez disse:

> A Clever Person Solves a Problem.
> A Wise Person Avoids It
>
> *Albert Einstein*

Desta forma, e com o objetivo de tranquilizar as pessoas foi concebida a ideia de seguro que brevemente definindo se trata de um acordo entre a entidade seguradora e o tomador de seguro. Neste e como é definido por «ASF - Apoio ao Consumidor», s.d. a seguradora

> [...] assume a cobertura de determinados riscos comprometendo-se a satisfazer as indemnizações ou a pagar o capital seguro em caso de ocorrência de sinistro, nos termos acordados. Em contrapartida a pessoa ou entidade que celebra o seguro (o tomador do seguro) fica obrigada a pagar ao segurador o prémio correspondente, ou seja, o custo do seguro.

Ao longo dos anos os seguros têm evoluindo e ganho importância sendo que nos dias de hoje o seguro está estruturado de forma a prevenir quase todo o tipo de situações. A aquisição de um seguro precede de uma análise ponderada e por vezes complicada uma vez que existe uma enorme diversidade de seguros que cobrem vários tipos de riscos, assim como as seguradoras que os oferecem (Martinho, 2014).

De toda a oferta existente no mercado segurador, destaca-se o seguro de saúde. É considerado por muitos como um dos mais importantes, visto que segura o nosso maior "bem", a saúde.

Como é muitas vezes dito:

> A saúde não tem preço.

Em conformidade com a CNN Portugal (Guimarães, 2022), de acordo com os dados fornecidos pela Associação Portuguesa de Seguradores, o mercado segurador tem crescido exponencialmente contando com aproximadamente 3 milhões de portugueses cobertos com seguro de saúde nos dias que correm, representando assim cerca de 30% da população portuguesa.

Num ambiente competitivo como é o dos seguros, e tendo em conta o princípio de Pareto isto é, que 80% do lucro provém de apenas 20% dos clientes, constata-se que a construção de uma relação próxima com o cliente e o conhecimento das suas necessidades bem como das suas preferências e comportamentos é uma ferramenta valiosa que permite obter vantagem sobre as restantes empresas.

Tendo em consideração a facilidade de mudança e por oposição a dificuldade de aquisição de novos clientes, é fundamental este conhecimento e a criação destas relações para que seja possível ajustar estratégias de marketing, desenvolver campanhas direcionadas e identificar padrões de consumo.

Michael LeBoeuf, autor empresarial americano e antigo professor na Universidade de Nova Orleães, disse:

> A satisfied customer is the best business strategy of all.
>
> *Michael LeBoeuf*

Posto isto, uma boa gestão da relação com o cliente conduz a uma melhor utilização dos recursos financeiros, que por conseguinte leva a relações duradouras e à sua fidelização.

O Grupo Future Healthcare (FH), é um grupo internacional fundado em 2003 que disponibiliza aos seus clientes um amplo portfólio de serviços no âmbito dos Seguros de Saúde e de Vida, com a finalidade de proporcionar acesso às melhores condições de saúde, vida e bem-estar («Future-Healthcare», s.d.).

Portanto este grupo ao estar ciente das exigências, da competitividade e da evolução do mercado segurador alinhou as suas principais competências que passam pela gestão de risco, tecnologias da informação e distribuição de serviços financeiros com a intenção de inovar e acompanhar o desenvolvimento deste grande mercado.

Dentro de todos os serviços prestados pela Future Healthcare esta dissertação incide sobre os seus planos e seguros de saúde. Atualmente o Grupo FH comercializa vários produtos e tem em vigor várias campanhas comerciais tanto para os seguros como para

os planos de saúde. É de extrema relevância perceber como essas campanhas influenciam a utilização e desistência de alguns dos produtos, tal como entender quais os clientes que maior benefício trazem à empresa, de maneira a conhecer o cliente e assim ponderar sobre as estratégias de marketing.

Para cumprir tais objetivos a empresa dispõe atualmente de análises estatísticas e de um esboço da análise Recência, Frequência e Valor Monetário (RFM). No entanto e apesar dos esforços da FH, os métodos aplicados e a análise presente não são suficientes, já que é praticamente impossível recolher informações detalhadas sobre a taxa de retenção, sobre que clientes são mais suscetíveis de não renovar contrato e sobre quais as campanhas mais eficazes para cada tipo de cliente, ou seja, as informações recolhidas não permitem elaborar estratégias direcionadas e tomar decisões impactantes que conduzam a um conhecimento alargado dos clientes, e com isso a uma vantagem competitiva.

Motivada pela melhoria da relação com o cliente do ponto de vista de o "conhecer" e ter em atenção as suas necessidades, para que se possam desenvolver estratégicas de marketing consistentes, esta dissertação pretende numa primeira fase mais geral identificar os clientes mais lucrativos para a empresa e caracterizá-los. Para esse fim recorrer-se-á à análise RFM e a uma variação da mesma.

Numa fase mais avançada são propostos dois objetivos mais específicos de modo a permitir à empresa complementar o objetivo principal e gerar um conhecimento mais rico e aprofundado sobre os seus clientes.

Assim, numa primeira fase e depois de caracterizados os clientes tenciona-se utilizar métodos de análise de *clusters* para a segmentação dos mesmos, e posteriormente analisar com recurso a técnicas da estatística clássica o efeito das campanhas na utilização\desistência dos produtos. Por último é calculada a probabilidade de Churn tendo como base a regressão logística, de forma a descobrir quais os clientes com maior probabilidade de mudar para a concorrência, e assim prestar-lhes uma maior atenção para evitar esse fim.

Capítulo

2 Revisão da Literatura

Em conformidade com os objetivos propostos, neste capítulo ir-se-á descrever e apresentar os estudos e as respetivas abordagens de diversos autores, no âmbito da Gestão da Relação com o Cliente, em inglês *Customer Relationship Management* (CRM).

2.1 Enquadramento Histórico do conceito CRM

Nos últimos anos a gestão da relação com o cliente (CRM) tem vindo a ganhar importância e interesse junto da comunidade científica (Payne e Frow, 2005). No entanto, e como é apresentado por Ngai, Xiu e Chau, 2009 não existe uma definição precisa aceite universalmente. Apesar disso, pode ser definida como uma disciplina ampla com muitas aplicações que abrange desde o serviço ao cliente até às estratégias de marketing e vendas. Concentra-se na relação cliente-organização, tendo como principal objetivo o cumprimento dos requisitos e a satisfação das necessidades do cliente, a fim de criar uma relação forte, duradoura e, em última análise, vitalícia entre o cliente e a empresa (Cheng e Chen, 2009; Huigevoort, 2015; Ravi, 2007).

Destacam-se a identificação, atração, desenvolvimento e retenção de clientes como sendo os quatro principais elementos da gestão da relação com o cliente.

A identificação dos clientes pode ser vista, como o nome sugere, identificando os clientes mais lucrativos, bem como os potenciais clientes e até mesmo aqueles que se podem perder para a concorrência, por meio da sua análise e segmentação. Por exemplo, a atração de clientes inclui o marketing direto para atrair segmentos-alvo. O desenvolvimento consiste em aumentar a lucratividade do cliente analisando o seu valor vitalício. Por fim a retenção de clientes, considerada a mais importante e a principal preocupação da CRM, inclui a criação de programas de fidelização e marketing individual. Destacam-se as campanhas/atividades direcionadas e a análise de rotatividade entre os programas

de fidelização e a definição dos perfis entre as estratégias de marketing individual (Ngai et al., 2009).

No estudo desenvolvido por Ng e Liu, 2001 estimou-se que custa cinco vezes mais atrair um novo cliente do que manter um cliente existente. Posto isto, e estando as empresas a par da dificuldade e do quão dispendioso é atrair novos clientes consideram importante estabelecer, e desenvolver atividades inovadoras com foco no cliente que ajudem a entender o seu valor, a fim de identificar clientes preciosos e leais (Cheng e Chen, 2009; Chuang e Shen, 2008). Considerando a forte correlação entre lucratividade e retenção referida por Payne, Christopher, Clark e Peck, 1999 e citado por Dursun e Caber, 2016 atender às necessidades dos clientes mencionados acima é vital para a sua retenção, e consequentemente para os resultados de uma empresa.

A gestão do relacionamento com o cliente é considerado pelos autores Blattberg, Kim e Neslin, 2008 como um "primo próximo" do database marketing. Melhor dizendo estes dois conceitos diferem apenas nos tópicos que enfatizam. A gestão da relação com o cliente enfatiza a melhoria das relações com os mesmos, enquanto o database marketing enfatiza o uso das bases de dados. Assim o database marketing pode ser visto como uma possível abordagem usada por grandes empresas para desenvolver relações com os seus clientes. Os seus três principais focos são o uso e a análise dos dados, a produtividade do marketing e a gestão de clientes.

Database Marketing ou Marketing de bases de dados define-se em (Blattberg et al., 2008) como sendo:
"[. . .] the use of customer databases to enhance marketing productivity through more effective acquisition, retention, and development of customers."

Num mundo inovador e tecnológico como é o de hoje, é possível para as empresas armazenar enormes quantidades de dados sobre os seus clientes. Estas bases contém informações demográficas, psicográficas, comportamentais e até informação sobre antigas estratégias de marketing, e possivelmente a reposta dos clientes às mesmas (Blattberg et al., 2008). Desta forma e por si só a informação presente nessas bases parece não ser significativa e suficiente, mas quando analisadas com detalhe e estudadas ao pormenor permitem extrair conhecimento novo e interessante sobre a carteira de clientes.

Posto isto, com fundamento na dimensão da FH, em todos os artigos e conclusões acima apresentadas, em resposta aos objetivos que se pretende cumprir deduz-se que a gestão de excelência das relações com os clientes, sendo este o conceito chave da presente dissertação passa por uma análise pormenorizada das bases de dados.

É claro que a análise das bases de dados é inerente ao seu uso, pelo que irá recorrer-se

à análise RFM e às ferramentas do Data Mining (DM) para realizar essa análise e extrair conhecimento valioso. A escolha destas metodologias é, como se poderá ver mais à frente, a melhor maneira de analisar o tipo de dados presente nesta dissertação.

Com o intuito de perceber a evolução deste grande conceito, consultaram-se três artigos que essencialmente tinham como objetivo identificar e categorizar estudos passados por forma a compreender a tendência de investigação. Destes concluiu-se que o número de artigos publicados tem vindo a aumentar, testemunhando o aumento do interesse da comunidade científica (Meena e Sahu, 2021; Silva de Araújo, Pedron e Picoto, 2018; Soltani e Navimipour, 2016).

Dos três realça-se o artigo escrito por Meena e Sahu, 2021. Neste foram revistos 104 artigos publicados entre janeiro de 2000 e junho de 2020. A maioria dos artigos foi publicado em revistas científicas.

Uma das análises mais interessantes levada a cabo por Meena e Sahu para este estudo foi a classificação dos artigos por indústria. Como se observa pela figura 2.1, a investigação da CRM está generalizada na indústria dos serviços, com destaque para os bancos e as seguradoras.

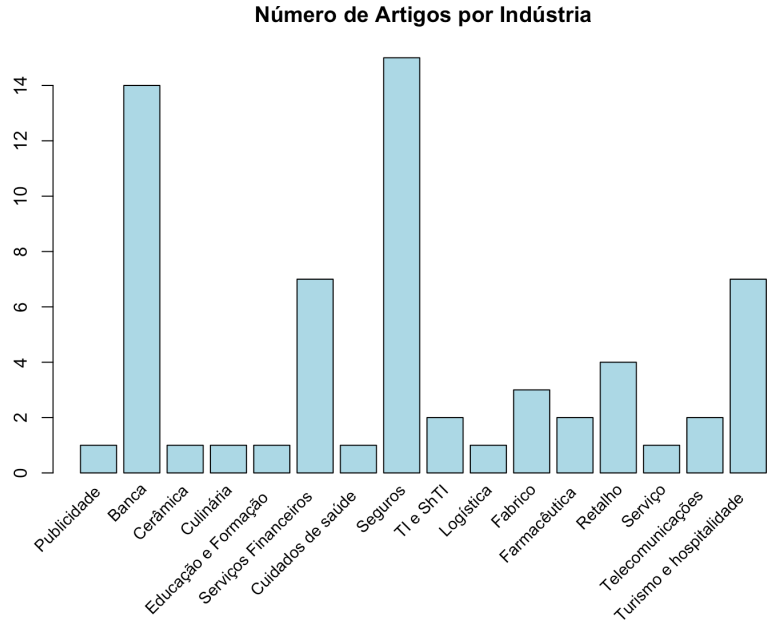


Figura 2.1: Número de artigos por indústria. Retirado de: Meena e Sahu, 2021

2.2 Análise RFM

Agora que se expôs os resultados e a literatura relevante relativos ao conceito chave, CRM, prossegue-se com a literatura pertinente da análise RFM.

Relembra-se que os custos de retenção são muito menores do que os custo de atração, e que nem todos os clientes são igualmente atraentes para a empresa, conforme concluído por Sohrabi e Khanlari, 2007 e citado por Wei, Lin e Wu, 2010. É desta forma aconselhável avaliar primeiro a lucratividade de cada cliente, com a finalidade de identificar os clientes que contribuem significativamente, e ao mesmo tempo identificar os que a empresa não tem intenção de reter, para então alocar os recursos financeiros de forma racional (Wei et al., 2010).

De entre os modelos disponíveis para analisar o valor do cliente, um dos mais conhecidos e amplamente utilizado é o modelo RFM. É um modelo baseado no comportamento, que tem como principal objetivo analisar o comportamento do cliente, para depois tentar prever o seu comportamento futuro. Baseia-se em apenas três fatores R, F e M que dizem respeito a três importantes variáveis relacionadas com a compra, e cujas pontuações atribuídas a esses fatores são "directamente proporcionais à vida e retenção do cliente" (Christy, Umamakeswari, Priyatharsini e Neyaa, 2021).

As iniciais R, F, e M na sua extensão caracterizam a recência, frequência e o valor monetário.

- Recência: quão recentemente um cliente fez uma compra;
- Frequência: com que frequência um cliente faz uma compra;
- Valor monetário: quanto dinheiro um cliente gasta em compras

Uma melhor descrição e explicação do método, assim como os procedimentos à sua aplicação serão discutidos na secção seguinte.

O conceito RFM foi introduzido pela primeira vez no artigo *Optimal selection for direct mail* escrito por Bult e Wansbeek em 1995. Durante o desenvolvimento, os autores conseguiram provar a eficácia da aplicação deste conceito a bases de dados de clientes (Birant, 2011).

Entre a comunidade científica que ao longo dos anos aplicou o modelo RFM, salientam-se os autores Kaymak, Hughes, Hu, Yeh, Cheng e Chen, de entre muitos que consideram este modelo como um dos mais conhecidos, eficazes, simples e poderosos para gerar conhecimento valioso sobre a carteira de clientes (Cheng e Chen, 2009; Dursun e Caber, 2016; Kaymak, 2001). É ainda considerado por Schijns e Schroder como é referido por Cheng e Chen, 2009 "como um modelo há muito familiar para medir a força da relação com o cliente".

A popularidade da análise RFM deve-se a vários fatores, nomeadamente ao facto de ser uma análise simples, intuitiva e que utiliza escalas numéricas e objetivas. Desta forma é possível para os profissionais de marketing retirarem conclusões sem a necessidade de

cientistas de dados. Já entre os investigadores, a sua popularidade deve-se ao facto de utilizar poucos critérios como atributos, o que reduz a complexidade do modelo (Cheng e Chen, 2009).

Devido à constante mudança tanto do comportamento dos clientes como das relações com os próprios, é extremamente importante um modelo que seja capaz de acompanhar essa mudança. Uma vez que o modelo RFM é de fácil aplicação e bastante claro permite que seja fácil a sua atualização, e consequentemente é possível para as empresas estarem sempre a par do valor real e atual do cliente, sendo assim popular também entre o mundo empresarial.

O modelo RFM tem sido utilizado com sucesso pelos investigadores numa grande variedade de áreas como o setor bancário e dos seguros, agências governamentais e indústrias de telecomunicação. Tem ainda uma utilidade muito abrangente, uma vez que permite aos profissionais segmentar os clientes, calcular o seu valor e observar o seu comportamento, como ainda pode ser utilizado com a finalidade de estimar a probabilidade de resposta para cada tipo de oferta (Dursun e Caber, 2016; Wei et al., 2010).

Por exemplo Colombo e Jiang, 1999 utilizaram um modelo RFM estocástico com o objetivo de identificar para cada oferta quais os clientes - alvo. Adotaram uma abordagem em duas fases, onde primeiro eliminaram os clientes que não eram ativos há muito tempo, e posteriormente utilizaram a frequência e o valor monetário para caracterizar e identificar quais os clientes visar.

Já Cheng e Chen, 2009 propuseram um novo procedimento que une os atributos RFM e o algoritmo *K-Means* em teoria de conjuntos aproximados, com a finalidade de segmentar diferentes grupos de clientes.

Ha e Park, 1998 citado por Wei et al., 2010 aplicaram com base em atributos RFM ferramentas do data mining para aumentar a quantidade de vendas de uma loja.

Cho e Moon, 2013 citado por Christy et al., 2021 tinham como principal objetivo a elaboração de um sistema de recomendação personalizado. Com a análise RFM foi possível para os autores extrair o perfil de cliente e assim fornecer uma recomendação mais precisa e fundamentada.

Por último Dursun e Caber, 2016 recorreram à análise RFM de modo a concluir sobre o perfil de clientes rentáveis, no contexto de hotelaria.

Tabela 2.1: Resumo das aplicações da análise RFM.

Autores	Objetivos	Referências
Colombo e Jiang	Identificar os clientes alvo para cada oferta	Colombo e Jiang, 1999
Cheng e Chen	Segmentar diferentes grupos de clientes	Cheng e Chen, 2009
Ha e Park	Aumentar a quantidade de vendas numa loja	Citado por Wei et al., 2010
Cho e Moon	Elaborar um sistema de recomendação personalizado	Citado por Christy et al., 2021
Dursun e Caber	Traçar o perfil de clientes hoteleiros rentáveis	Dursun e Caber, 2016

Apesar de ser vastamente utilizado e popular entre a comunidade científica, como anteriormente se exemplificou e resume na tabela 2.1, certamente há aspetos passíveis de se melhorar e algumas desvantagens que serão detalhadas no seguimento da presente dissertação.

Nesse sentido, os esforços dos investigadores para melhorar o modelo RFM aumentaram significativamente nos últimos anos. Alguns autores propõem aumentar o número de parâmetros, enquanto que outros propõem substituir um dos parâmetros por outro, existindo assim múltiplas versões do modelo (Dursun e Caber, 2016).

Foi sugerido, por exemplo por Cheng e Chen, 2009 a utilização de mais dois parâmetros no modelo original RFM. Propuseram a utilização do modelo RFMTC, onde T e C representam o tempo desde a primeira compra e a probabilidade de *Churn*, respetivamente. Já outra versão foi o modelo TRFM, que foi proposta para lidar com a periodicidade do produto (Birant, 2011).

As versões em que um dos parâmetros foi substituído, foram criadas essencialmente com o propósito de representar melhor os clientes da indústria em estudo. A título de exemplo tem-se o modelo RFD (recência, frequência e duração) e o modelo RFL (recência, frequência e lealdade). O primeiro mostrou-se útil para analisar o comportamento de visitantes de sites online e o segundo é uma adaptação para transações anuais (Birant, 2011).

Uma das versões mais antigas e que difere das restantes é a Análise RFM ponderada (WRFM). Isto porque não se encaixa nem nas versões que utilizam mais parâmetros nem nas que substituem um deles. Contrariamente às restantes, este modelo considera que as medidas R, F e M têm diferentes pesos consoante as características da indústria.

Como Wei et al., 2010 referiram, o modelo RFM ponderado tem em consideração a "importância relativa das variáveis RFM através do algoritmo Analytic Hierarchy Process

(AHP), enquanto que o modelo RFM não ponderado não".

Existem dois artigos com bastante visibilidade neste assunto e que defendem opiniões distintas. Hughes, 1994 considera e defende que cada medida tem o mesmo peso, sendo assim de igual importância. Já Stone, 1995 acredita e justifica que a importância das três variáveis não é igual, tendo por isso diferentes pesos (Cheng e Chen, 2009). Complementando as ideias defendidas por Stone, e como é referido em Wei et al., 2010 Tsai e Chiu, 2004 indicaram ainda que a soma dos pesos deveria ser um.

Por conseguinte, que peso atribuir e qual o parâmetro com maior ponderação foram questões que surgiram. Para dar resposta, Stone, 1995 entre outros, sugerem atribuir maior ponderação à frequência, seguida da recência e com a menor para o valor monetário. Sob a perspectiva de Chuang e Shen, 2008 e citado por Khajvand, Zolfaghar, Ashoori e Alizadeh, 2011 a ponderação mais elevada é atribuída ao valor monetário e a menor à recência.

No entanto e atendendo às particularidades de cada indústria, Liu e Shih, 2005 referido em Wei et al., 2010 em vez de atribuir os pesos arbitrariamente aplicaram o método AHP. Manifestou-se ser bastante positivo na medida em que considera as opiniões, e todas as decisões são tomadas pelo decisor. Como consequência este método é utilizado, como é referido por Khajvand et al., 2011 para "determinar a importância relativa (pesos) das variáveis RFM".

Uma vez que na presente dissertação o objetivo mais geral, passa pela caracterização dos clientes destaca-se o artigo escrito por Hosseini, Maleki e Gholamian, 2010. Com base nos dados de uma empresa fornecedora de equipamento e acessórios para várias fábricas de automóveis, os autores conduziram um estudo sobre a aplicação de técnicas de data mining conjuntamente com a análise RFM com o objetivo de "[. . .] determinar o grau de lealdade do produto para alcançar uma excelente CRM [. . .]". Utilizaram a análise RFM ponderada e não ponderada a fim de comparar a qualidade dos agrupamentos e concluíram uma melhoria na precisão do agrupamento aquando da utilização do modelo WRFM.

Em suma o modelo RFM ponderado, em conjunto com o algoritmo AHP é muitas vezes preferível ao não ponderado por simplesmente ter em atenção as características e a unicidade tanto da indústria como dos objetivos em questão.

Para concluir, do ponto de vista empresarial o benefício da utilização desta análise é significativo uma vez que permite, tal como foi referido anteriormente, calcular a rentabilidade de cada cliente. Em consequência é possível identificar os clientes que criam maior valor, e assim desenvolver estratégias de marketing eficazes e apelativas. Finalmente conseguem atingir taxas de retenção mais elevadas e maior lucro.

2.3 Utilização de técnicas de Data Mining em contexto CRM

Tendo em conta o conceito chave da presente dissertação, de acordo Ngai et al., 2009 a base de construção para qualquer estratégia de CRM eficaz e bem sucedida são as ferramentas de dados e as tecnologias de informação (TI).

Os autores concluíram ainda, com base nos artigos de Berson, Smith e Thearling, 2000, de He, Xu, Huang, e Deng, 2004 e de Teo, Devadoss, e Pan, 2006, que a gestão da relação com o cliente pode ser classificada em operacional e analítica. A operacional diz respeito "à automatização dos processos empresariais", já a analítica "refere-se à análise das características e comportamentos dos clientes de modo a apoiar as estratégias de gestão de clientes da organização" (Ngai et al., 2009).

Como mencionado anteriormente, a capacidade atual de armazenamento de dados é enorme. Muitas organizações recolhem e armazenam informações muito variadas sobre os seus clientes. No entanto, e apesar da extensa informação, é impossível para as organizações converter os dados em conhecimento útil e valioso, uma vez que não dispõem da capacidade para o fazer (Ngai et al., 2009).

Assim sendo é necessário utilizar ferramentas que consigam lidar com dados massivos e deles extrair conhecimento novo e valioso. De acordo com Ngai et al., 2009, uma das formas mais popular de analisar dados de clientes no contexto da CRM são as ferramentas do data mining. Distinguem-se pela positiva das restantes, pois possibilitam o descobrimento de conhecimentos novos e valiosos que estavam ocultos na enorme quantidade de dados.

Apresentado em Ngai et al., 2009, os autores Turban, Aronson, Liang, e Sharda ,2007 definem o data mining como "o processo que utiliza técnicas estatísticas, matemáticas, de inteligência artificial e de aprendizagem por máquinas para extrair e identificar informação útil e subsequentemente adquirir conhecimentos a partir de grandes bases de dados". Poderá então ser visto como o processo de extração e descoberta de padrões em grandes dados.

Como é referido pelos autores Ngai et al., 2009, "a aplicação de ferramentas de data mining em CRM é uma tendência emergente na economia global". A chave para desenvolver uma estratégia de CRM competitiva que proporcione a atração de potenciais clientes e a maximização do valor do cliente, é a análise e compreensão do comportamento assim como das características do cliente.

Como Berson et al., 2000 constatou e Ngai et al., 2009 referiu, as melhores ferramentas de apoio à tomada de diferentes decisões da CRM passam pela escolha apropriada de

entre o leque de ferramentas do data mining , que são excelentes para extrair e identificar informações e conhecimentos úteis de enormes bases de dados de clientes. Como resultado, numa economia centrada no cliente, a utilização de técnicas de extracção de dados em CRM é vantajosa.

Os tipos de modelação de dados que a seguir se apresentam, podem ser realizados por mais do que uma técnica de data mining:

- Associação;
- Classificação;
- Agrupamento;
- Previsão;
- Regressão;
- Descoberta de Sequências;
- Visualização

De seguida é exposto na figura 2.2 um quadro de classificação sobre técnicas de DM em CRM. Para a construção do seguinte quadro os autores basearam-se num artigo de revisão das técnicas de data mining utilizadas na CRM.

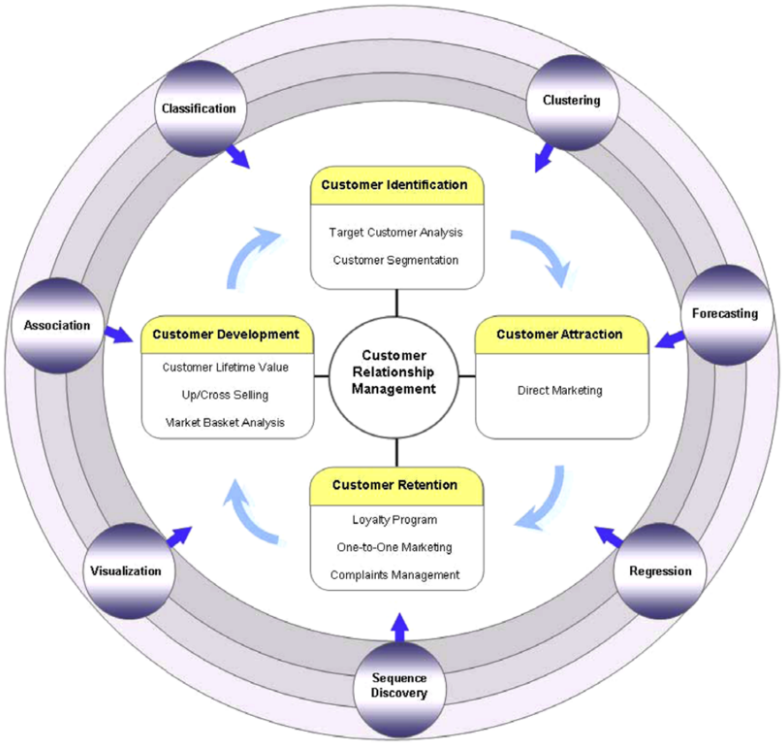


Figura 2.2: Classificação de técnicas de DM em CRM. Retirado de: Ngai, Xiu e Chau, 2009

Como é referido por Ngai et al., 2009, no contexto da CRM "o data mining pode ser utilizado para orientar a tomada de decisões e prever os efeitos das decisões". Por exemplo, para o estudo em questão, dois dos maiores e mais importantes resultados da aplicação do DM, é a possibilidade de calcular a probabilidade de um cliente existente optar por fazer negócio com um concorrente e a de aumentar as taxas de resposta das campanhas de marketing.

A tabela 2.2 apresenta para cada modelo de data mining, uma pequena descrição assim como algumas das técnicas mais utilizadas. Destacam-se a negrito os tipos de modelação que se irão utilizar para dar repostas aos dois objetivos mais específicos.

Tabela 2.2: Tabela de resumo dos tipos de modelação de DM

Tipo de Modelação	**Descrição**	**Técnicas mais utilizadas**
Associação	Encontrar relações entre variáveis de um conjunto de dados	Análises Estatísticas Algoritmos à Priori
Classificação	Atribuir objectos a uma de várias categorias pré-definidas	Redes Neuronais Árvores de Decisão
Agrupamento	Segmentar uma população heterogénea em vários grupos mais homogéneos	Algoritmos de Cluster
Previsão	Estimar o valor futuro com base nos acontecimentos do passado	Redes Neuronais Análise de sobrevivência
Regressão	Estimar as relações entre uma variável dependente e uma ou mais variáveis independentes para depois prever os valores numéricos de um dado conjunto de dados	Regressão Linear Regressão Logística
Descoberta de sequências	Descobrir padrões estatisticamente relevantes entre dados em que os valores são fornecidos numa sequência	Estatística Teoria de Conjuntos
Visualização	Apresentação de dados para que os utilizadores possam visualizar padrões complexos	-

Por forma a ter conhecimento da utilização das ferramentas do data mining na gestão da relação com o cliente, consultou-se o artigo de revisão escrito pelos autores mencionados nesta secção. Neste foram analisados 87 artigos publicados entre 2000 e 2006, e o que se pretendia era dar conhecimento ao leitor "[...] sobre a aplicação de técnicas de mineração de dados no domínio da CRM e as técnicas que são mais frequentemente utilizadas" (Ngai et al., 2009).

Assim infere-se que:

- A retenção de clientes é a dimensão onde mais se utilizam ferramentas do DM, representando 62.1% dos artigos;
- A técnica de DM mais comum são as redes neuronais;
- A revista *Expert Systems with Applications* foi a que mais publicou sobre o assunto, cerca de 73% dos artigos;
- As publicações relacionadas com a aplicação de técnicas de DM em CRM têm aumentado exponencialmente.

Exibe-se, na tabela 2.3 a título de exemplo qual o objetivo e a técnica utilizada, assim como os resultados de três artigos que se inserem na identificação, atração e retenção de clientes, por forma a dar um exemplo prático para três das quatro dimensões do CRM, e também de modo a fundamentar as técnicas escolhidas para abordar os dois objetivos complementares.

Tabela 2.3: Tabela de resumo de três artigos que aplicam técnicas de DM em CRM

Dimensão CRM	**Modelo de DM**	**Técnica escolhida**	**Objetivo**	**Referência**
Identificação	Agrupamento	K-Means	Identificar sub-grupos de clientes que partilhem motivações de compra semelhantes	Dennis, Marsland e Cockett, 2001
Atração	Classificação	Árvores de Decisão	Identificar o conjunto de variáveis relevantes para prever a utilização de cupões	Buckinx, Moons, Van den Poel e Wets, 2004
Retenção	Regressão	Regressão Logística	Prever a deserção de clientes leais	Buckinx e Van den Poel, 2005

Constata-se que as técnicas de DM são amplamente utilizadas em todas as dimensões da CRM, sendo estas uma tendência emergente na indústria. Em particular a análise

de *clusters* e a regressão logística, que são as mais indicadas e as mais utilizadas para a segmentação de clientes e para a análise da probabilidade de rotatividade (probabilidade de *Churn*), respetivamente.

Após a apresentação das diferentes técnicas de data mining e como normalmente são utilizadas no contexto da gestão da relação com o cliente, prosseguir-se-à com a exposição de três artigos bastante relevantes para o seguimento do estudo. Ou seja, serão discutidos dois artigos que dizem respeito à segmentação de clientes utilizando técnicas de agrupamento e outro referente à análise da probabilidade de rotatividade recorrendo à regressão.

2.3.1 Segmentação de Clientes

A segmentação de clientes na CRM é algo com muito valor e bastante feito entre a comunidade científica. São inúmeras as vantagens deste estudo, salientando-se o conhecimento que se obtém sobre os diferentes tipos de clientes, para depois se desenvolverem campanhas de marketing bem sucedidas. Ou seja com a segmentação é possível dividir o grupo de clientes em sub-grupos mais homogéneos, que partilhem entre si características comportamentais e assim obter informação sobre os clientes leais e valiosos, assim como dos clientes perdidos, e também quais as características de um potencial cliente.

Embora se identifiquem diferentes grupos de clientes com a segmentação, é possível explorar ainda mais cada segmento através da realização de perfis de clientes. Cada perfil "descreve os hábitos, necessidades e comportamentos específicos de um grupo de clientes (segmento)" (Walters e Bekker, 2017).

Assim sendo tanto a segmentação como a definição do perfil de cliente são fundamentais para direcionar campanhas de marketing. Tal como os autores Walters e Bekker, 2017 referem "a segmentação é frequentemente utilizada em conjunto com a definição de perfis".

Existe uma abundância de métodos que permitem segmentar um grupo de clientes, desde os tão conhecidos algoritmos de *cluster* até às árvores de classificação e ainda às redes neuronais. Tem-se ainda, no contexto CRM que "um modelo simples e eficaz que pode ser aplicado à segmentação é o modelo RFM" (Gustriansyah, Suhandi e Antony, 2020). Todavia e como se poderá ver na secção 3, a forma como os clientes são divididos não é transversal nem aceite universalmente entre os investigadores.

Contudo e apesar da variedade de modelos, os algoritmos de *cluster* continuam a ser os mais conhecidos e mais utilizados para este fim. Em contexto CRM, são muitas vezes utilizados em conjunto com o modelo RFM, sendo que desta forma é obtido um melhor agrupamento, tendo ainda em atenção as importantes variáveis representativas

do comportamento (R, F e M).

Walters e Bekker, 2017 com o objetivo de identificar estratégias de marketing apropriadas, demonstram em dois problemas, um de exemplo e o outro real, a segmentação e a criação de perfis de clientes.

Para o processo de segmentação começaram por determinar para cada cliente os parâmetros R, F e M, que foram depois codificados através da divisão dos valores de recência, frequência e monetários em 5 categorias. Ao conjunto de dados categóricos aplicaram o algoritmo *K-means*, conseguindo identificar com sucesso 4 e 3 *clusters* para o problema de exemplo e real, respetivamente. Posto isto finalizaram com a extração e criação do perfil de cliente para cada agrupamento encontrado.

Como resultado os autores conseguiram não só associar a segmentação e os perfis de clientes a grandes dados, como ainda melhorar a eficiência das campanhas de marketing.

Já os autores Wu et al., 2020, de forma semelhante aos anteriores aplicaram a um problema real, o modelo RFM conjuntamente com o *K-Means* para realizar a segmentação de clientes. Identificaram 4 *clusters* e para cada um deles caracterizaram os clientes como clientes em risco, de elevado valor, ativos e com potencial. Posteriormente extraíram para cada tipo de cliente as características de compra e desenvolveram estratégias de marketing direcionadas.

Conseguiram comprovar a eficácia do modelo proposto pela melhoria de índices de desempenho da empresa, como por exemplo o número de clientes ativos que aumentou significativamente.

2.3.2 Retenção de Clientes

Tal como já foi referido, a indústria seguradora opera num ambiente bastante competitivo uma vez que o cliente é livre de escolher a sua seguradora, e também muito desafiante dado que as apólices são normalmente renovadas todos os anos. Desta forma os clientes podem facilmente alterar o seu fornecedor de seguros, levando a companhia à perda de um cliente que "pode significar uma perda de receitas durante vários anos" (Spiteri e Azzopardi, 2018).

Juntando às perdas significativas de receitas, o custo elevado de aquisição de novos clientes, quando comparado com o custo de retenção, tem-se que a rotatividade de clientes é um tópico muito importante e de bastante interesse para as empresas, que tem vindo a ganhar popularidade junto da comunidade científica.

A Rotatividade de Clientes ou *Customer Churn*, como é habitualmente designado diz respeito à perda de um cliente existente para a concorrência. Usualmente também pode

ser utilizado para denotar "o movimento dos clientes e determinar aqueles que estão em risco de sair" (Spiteri e Azzopardi, 2018).

Por sua vez, a retenção de clientes tem como principal objetivo "exceder as expectativas dos clientes para que estes se tornem leais à marca", e quando isso não acontece pode ocorrer o efeito oposto, ou seja, a rotatividade de clientes (Huigevoort, 2015).

Para finalizar e apenas a título de curiosidade, tem-se que a taxa de retenção no mercado nacional de seguros tem vindo a diminuir nos últimos anos.

As razões que levam um cliente a cessar contrato são variadíssimas, desde preços mais baratos, melhores qualidades de serviço, preferência pessoal, reclamações que ainda não foram resolvidas, promoções, descontos entre muitas outras. De acordo com Spiteri e Azzopardi, 2018, as variáveis mais significativas para a rotatividade de clientes na indústria seguradora são relacionadas com a apólice e o seu detentor, com reclamações, prémios e ainda com a satisfação do cliente.

Devido à quantidade de clientes a cargo de uma seguradora é bastante complicado para esta estar a par de mudanças no comportamento dos seus clientes, assim como estabelecer quem são os clientes prováveis de abandonar os seus serviços.

Consequentemente, é importante identificar com antecedência os clientes rotativos e compreender as razões que os motivam a sair, para depois elaborar estratégias adequadas que evitem esse fim.

Ao estabelecer antecipadamente os clientes suscetíveis de abandonar a companhia, é criada uma oportunidade para que as empresas se concentrem em encontrar iniciativas, que conduzam à retenção desses clientes, como por exemplo o marketing direto e campanhas direcionadas. É assim possível visar apenas os clientes em questão em vez de tentar chegar a todos eles, e desse modo encorajar a lealdade para com a companhia.

Como exemplo tem-se o artigo escrito por Spiteri e Azzopardi, 2018 cujo principal objetivo é estabelecer quem são os clientes prováveis de sair e quais os fatores associados à rotatividade. Consideram para o seu estudo informações relacionadas com prémios, sinistros, apólices e detentores de apólices. Aos dados aplicaram seis técnicas e a que se mostrou com melhores resultados na previsão da rotatividade foi o algoritmo florestal aleatório, contudo também a regressão logística teve um bom desempenho.

No geral, as técnicas mais utilizadas neste contexto são as árvores de decisão, regressão logística e redes neuronais. Das três as que normalmente apresentam melhores resultados são a regressão logística e as redes neuronais, com destaque para a regressão visto que fornece o conhecimento sobre as variáveis mais importantes para prever a rotatividade dos clientes.

Por fim, com este estudo é possível não só identificar os clientes rotativos mas também desenvolver estratégias mais focalizadas e direcionadas para os mesmos. Como consequência tem-se o aumento da satisfação do cliente, o que poderá levar à sua lealdade.

Capítulo

3 Fundamentação Teórica

No presente capítulo serão descritos com o maior detalhe os conceitos e metodologias que serão utilizados na fase de implementação desta Dissertação.

3.1 Valor do Cliente

Para o cálculo do valor do cliente como já se enunciou, a análise mais conhecida é a análise RFM que será descrita já de seguida.

3.1.1 Análise RFM

A análise RFM pode ser realizada seguindo várias abordagens. Na forma tradicional e usualmente utilizada, os clientes são analisados e posteriormente codificados segundo o "método quintil". Este método divide os clientes em cinco sub-grupos de igual dimensão e atribui a cada um deles um pontuação que varia de um a cinco. As pontuações podem ser assumidas como tendo características únicas, dadas na tabela 3.1.

Tabela 3.1: Descrição das pontuações RFM

Pontuação	Característica
5	Leais
4	Promissores
3	Potenciais
2	Em Risco
1	Perdidos

Na prática funciona da seguinte forma:

1. Os clientes são ordenados por ordem decrescente para a Recência (R), isto é em primeiro lugar têm-se os clientes que compraram mais recentemente e em último os que compraram há mais tempo.

    a) Depois, a lista ordenada de clientes é dividida em quintis iguais. Assim, aos 20% dos clientes que mais recentemente compraram é atribuído o número 5; aos 20% seguintes é atribuído o número 4, e assim por diante.

2. Repete-se o processo mas para a Frequência (F), sendo que neste caso em primeiro estão os que compram mais frequentemente e em último os que menos compras fazem.

3. Por fim ordena-se os clientes tendo em conta o Valor Monetário (M) gasto e repete-se o processo. Assim obtém-se uma lista ordenada, onde os que estão no topo são os que mais gastam em compras e os que estão em últimos são os que menos gastam.

Finalmente, a todos os clientes foi atribuída uma pontuação para as três variáveis (R,F e M) e a sua pontuação pode variar entre (5,5,5), sendo esta a mais elevada, (5,5,4), (5,5,3), ... e (1,1,1) que é a mais baixa. Os clientes com as pontuações mais altas são geralmente os clientes mais rentáveis da empresa.

No modelo RFM do presente estudo e assim como se resume na tabela 3.2, Recência (R) indica o tempo decorrido desde o último sinistro; Frequência (F) mostra o número total de sinistros no período de tempo considerado, e Valor Monetário (M) reflete o custo total com sinistros. Apesar de não se tratarem de compras mas sim de sinistros, o modelo é aplicado seguindo o mesmo molde.

Nota-se, que para a Recência, na lista ordenada em primeiro tem-se os clientes que tiveram sinistros há mais tempo, e assim a pontuação atribuída é mais elevada quanto maior é o valor de R. Ao contrário do que acontece para a Frequência e o Valor Monetário, em que na lista ordenada aparecem em primeiro aqueles que têm menos sinistros e menor valor total gasto, respetivamente. Desta forma, tanto para a variável F e M tem-se que as pontuações a atribuir serão mais elevadados quanto menor for o seu valor.

Tabela 3.2: Descrição das variáveis RFM do presente estudo

Variáveis	Significado	Este estudo define
R	Tempo desde a última compra	Tempo desde o última sinistro
F	Número total de compras	Número total de sinistros
M	Valor total gasto em compras	Valor total gasto em sinistros

3.1.2 Análise WRFM

A análise RFM ponderada (WRFM) segue o procedimento anterior descrito sendo que a única diferença consiste no peso atribuído a cada variável. Isto é, depois de atribuir a

cada cliente uma pontuação de um a cinco para a recência, frequência e valor monetário é calculada uma média ponderada, de acordo com a equação (3.1), obtendo-se o valor que cada cliente representa para a empresa. Quanto mais próximo estiver de um, mais rentável é o cliente para a companhia.

$$VC_j = w_R P_R^j + w_F P_F^j + w_M P_M^j \tag{3.1}$$

onde VC_j representa o valor do cliente j; w_R, w_F e w_M os pesos atribuídos às variáveis R, F e M, respetivamente e P_R^j, P_F^j e P_M^j representam as pontuações atribuídas ao cliente j para as variáveis R, F e M, respetivamente.

Nota-se que a soma dos pesos deverá ser um.

3.2 Análise de *Clusters*

A análise de *Clusters* é umas das técnicas mais conhecidas e utilizadas para a análise de dados. Esta técnica é uma técnica multivariada que pretende agrupar casos em grupos homogéneos, designados por *clusters*, com base nas semelhanças/dissemelhanças existentes entre os dados.

Os algoritmos de *Clustering* estão divididos em dois grandes grupos:

- Hierárquicos: Inicialmente cada caso é considerado como um *cluster* individual. Em cada iteração, os *clusters* semelhantes unem-se com outros *clusters* até que um *cluster* ou k *clusters* sejam formados. Determina-se o número de *clusters* que é possível formar a partir da matriz inicial de dados
- Não hierárquicos: Consiste na divisão dos casos por grupos previamente definidos, ou seja os n casos são divididos pelos k grupos definidos à priori.

As subsecções 3.2.1 e 3.2.2 e 3.2.3 descrevem sumariamente o algoritmo *K-Means*, o método *Silhouette* e o método *Elbow*, estes dois últimos utilizados para escolher o número ótimo de *clusters*.

3.2.1 *K-Means*

O algoritmo *K-Means* insere-se nos algoritmos não hierárquicos, pelo que é um método que divide as n observações em K grupos similares, denominados *clusters*. A divisão é feita atribuindo cada uma das n observações ao *cluster* cujo centróide é o que está mais próximo, sendo este o ponto médio do *cluster*. Este algoritmo pretende agrupar as observações de modo a maximizar a semelhança dentro do mesmo *cluster* e a minimizar a similaridade entre diferentes *clusters*.

A similaridade é calculada utilizando uma função de distância, sendo que no caso do *K-Means* é a distância euclideana. Tem-se ainda que a similaridade entre duas observações

é tanto maior quanto menor for a distância entre eles.

Este método baseia-se num processo iterativo, isto é num processo com duas grandes etapas, a primeira onde os membros do *cluster* são atualizados e a segunda onde são recalculados os centróides.

Seja $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ um conjunto de n observações, onde cada uma das observações é um vetor real de dimensão d, isto é, $\mathbf{x} = (x_1, ..., x_d)'$. Seja também $\mathbf{C} = (C_1, ..., C_k)$ um conjunto de k *clusters* com centróides $\overline{\mathbf{x}} = (\overline{x}_1, ..., \overline{x}_k)$.

Inicialmente o centróide de cada *cluster* é definido de forma aleatória, e de seguida é calculada através da distância euclideana, que se apresenta na equação (3.2), a distância entre cada observação e cada *cluster* de $\mathbf{C}$.

$$d(\mathbf{x}, \overline{\mathbf{x}}) = \sqrt{(\mathbf{x} - \overline{\mathbf{x}})'(\mathbf{x} - \overline{\mathbf{x}})} \tag{3.2}$$

Para cada observação é calculada a distância euclideana entre essa observação e o centróide de cada um dos *clusters*. Posteriormente, o algoritmo agrupa a respetiva observação ao *cluster* cuja distância é a menor, ou seja essa observação será membro do *cluster* cujo centróide está mais próximo de si.

Depois de atribuir todas as observações aos *clusters* existentes são atualizados os centróides. A sua atualização é feita recalculando o ponto médio do *cluster*.

Após atualização é calculada novamente a distância entre cada observação e os novos centróides. Uma vez que os centróides atualizaram é também necessário atualizar os membros de cada *cluster*, assim sendo a distância entre cada observação e os novos centróides é recalculada, e estas são novamente atribuídas ao *cluster* cujo centróide está mais próximo. Este processo repete-se até que não hajam realocações das observações e neste momento está encontrada a menor variabilidade intragrupo denominada por Within Sum of Squares (WSS). A WSS é calculada através do somatório do quadrado das distâncias euclideanas:

$$WSS = \sum_{j=1}^{n} (\mathbf{x}_j - \overline{\mathbf{x}})'(\mathbf{x}_j - \overline{\mathbf{x}}) \tag{3.3}$$

Em suma podemos definir as seguintes etapas:

1. Definir o número de *clusters* k que deve ser formado;
2. Inicializar os centróides de forma aleatória;
3. Calcular a distância entre cada observação e cada centróide;
4. Atribuir cada observação ao *cluster* com o centróide mais próximo;
5. Atualizar os centróides;
6. Recalcular a distância entre cada observação e cada centróide;

7. Atualizar os membros de cada *cluster*;

8. Repetir os passos 6 e 7 até não hajam realocações

3.2.2 *Silhouette*

Uma das formas de definir o número ótimo de *clusters* passa pelo cálculo do índice de *Silhouette*

Assim, para determinar o número ótimo de *clusters* a considerar procede-se da seguinte forma:

1. Para diferentes valores de K, como por exemplo de 1 a 10, realizar a análise de *clusters*;

2. Para cada valor de K calcular

   - a distância média entre cada objeto i e os restantes objetos do mesmo *cluster*, a_i
   - a menor distância média entre cada objeto i e os membros dos *clusters* que não o contêm, b_i
   - o coeficiente Silhouette, dado por

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)}$$

3. Representar graficamente a relação entre k e a média dos coeficientes *Silhouette*

O valor do índice de *Silhouette* varia entre -1 e 1. Se apresentar um valor próximo de 1 então será um bom agrupamento, isto é os objetos estão bem agrupados, contrariamente se for próximo de -1 então as conclusões são as inversas.

Com base na figura elaborada no passo 3, conclui-se que o número ótimo de *clusters* a considerar corresponderá àquele que apresentar a média mais elevada.

3.2.3 *Elbow*

Outra forma de definir o número ótimo de *clusters* passa pela minimização da variabilidade intragrupo (WSS) definida na equação (3.3).

Assim, para determinar o número ótimo de *clusters* a considerar procede-se da seguinte forma:

1. Para diferentes valores de K, como por exemplo de 1 a 10, realizar a análise de *clusters*;

2. Calcular para cada análise a variação total WSS_{total}

$$WSS_{total} = \sum_{j=1}^{k} WSS_j$$

3. Representar graficamente a relação entre k e a variação WSS_{total}

Pela figura elaborada no passo 3, conclui-se que o número ótimo de *clusters* a considerar corresponderá ao ponto onde a linha quebrada apresenta o aspeto de um "cotovelo", a partir do qual o decréscimo da variabilidade intragrupo é negligenciável.

3.3 Regressão Logística

Tal como a regressão linear também a regressão logística tem como objetivo a modelação ou predição de valores de uma variável, usualmente designada variável dependente ou de resposta em função dos valores de outra ou outras variáveis, também conhecidas como variáveis independentes ou regressoras. Essencialmente a regressão logística utiliza uma função logística para modelar uma variável de saída binária.

$$f(x) = \text{Função Logística} = \frac{1}{1 + e^{-x}} \tag{3.4}$$

onde x é a variável independente.

A diferença consiste nos valores que a variável resposta toma, ou seja, uma vez que a regressão logística estima a probabilidade de ocorrência de um evento, como *churn* ou não *churn*, isto é modela uma variável binária, com base num determinado conjunto de dados de variáveis independentes, o resultado da variável resposta é limitado entre 0 e 1.

Outra diferença reside no facto de a regressão logística não requerer uma relação linear entre variáveis de entrada e de saída, isto é resultado de uma transformação conhecida como probabilidade logística, ou o logaritmo natural das probabilidades, que divide a probabilidade de sucesso pela de insucesso. A função logística e a respetiva transformação são dados da seguinte forma:

$$\text{Probabilidades} = \frac{p}{1-p} \rightarrow logit(p) = ln\left(\frac{1}{1-p}\right) \tag{3.5}$$

Seja $logit(p) = mx + b$, então tem-se:

$$logit(p) = mx + b \Rightarrow mx + b = ln\left(\frac{1}{1-p}\right) \tag{3.6}$$

$$\Leftrightarrow e^{mx+b} = \frac{1}{1-p} \Leftrightarrow p = \frac{e^{(mx+b)}}{1 + e^{(mx+b)}} \Rightarrow p = \frac{1}{1 + e^{-(mx+b)}}$$

Os parâmetros são estimados utilizando os estimadores de máxima verosimilhança. Depois de encontrados serão calculadas as probabilidades previstas. Uma vez que a regressão logística modela uma variável binária, temos que “uma probabilidade inferior a 0.5 irá prever 0 enquanto uma probabilidade superior a 0.5 irá prever 1”.

Capítulo

4 Pressupostos e Dados

Neste capítulo ir-se-á apresentar e descrever as Bases de Dados utilizadas. Além disso, irá descrever-se a etapa de Pré-Processamento e todos os passos que conduziram à criação das bases utilizadas no seguimento do estudo. De referir ainda, que se irá apresentar os pressupostos considerados para o efeito.

4.1 Descrição das Bases de Dados

O presente estudo incidiu sobre quatro bases de dados disponibilizadas pela FH. Destas, duas contém informação sobre os clientes dos seus planos e seguros de saúde, e as outras duas informação sobre os atos médicos dos clientes que possuem planos e seguros. De forma resumida tem-se:

- **Base de dados 1 (BD_1)**: Informação sobre os clientes que possuem planos de saúde, constituída por 15 variáveis e 23 576 observações;
- **Base de dados 2 (BD_2)**: Informação sobre os clientes que possuem seguros de saúde, constituída por 18 variáveis e 6 919 observações;
- **Base de dados 3 (BD_3)**: Informação sobre os atos médicos de clientes que possuem planos de saúde, constituída por 16 variáveis e 128 787 observações;
- **Base de dados 4 (BD_4)**: Informação sobre os atos médicos de clientes que possuem seguros de saúde, constituída por 16 variáveis e 35 685 observações;

Nas BD_1 e BD_2 tem-se as seguintes variáveis:

Tabela 4.1: Descrição das variáveis presentes nas bases BD_1 e BD_2

Nome da Variável	Descrição	Tipo
Num_Apolice	Identificador da Apólice	Nominal
Data_Inicio_Ap	Data de início da Apólice	Ordinal
Data_Ultima_Anuidade	Data da última renovação da Apólice	Ordinal
Estado_Apolice	Estado da Apólice	Nominal
Id_Prod	Identificador do Produto	Nominal
Fracionamento	Tipo de pagamento	Nominal
Id_PS	Identificador do Segurado	Nominal
Data_Nascimento	Data de Nascimento	Ordinal
Data_Inicio_PS_Ap	Data de início do segurado na apólice	Ordinal
Data_Fim_PS_Ap	Data de fim do segurado na apólice	Ordinal
P_Comercial_PS	Prémio comercial pago pelo segurado	Contínua
Campanha	Campanha associada à Apólice	Nominal
Protocolo	Descrição da campanha	Nominal
Valor desconto	Valor de desconto oferecido para a Duração da Campanha	Ordinal
Duracao_Campanha	Duração da Campanha	Nominal
Id_Ramo	Identificador do Ramo	Nominal
Nome_Prod	Nome do Produto	Nominal
P_Total_PS	Prémio total pago pelo segurado	Contínua

Nota-se que as três últimas variáveis estão presentes apenas na BD_2.

Nas BD_3 e BD_4 tem-se as variáveis a seguir enumeradas. Nota-se que estas estão presentes em ambas as bases.

Tabela 4.2: Descrição das variáveis presentes nas bases BD_3 e BD_4

Nome da Variável	Descrição	Tipo
Num_Apolice	Identificador da Apólice	Nominal
Id_PS	Identificador do Segurado	Nominal
Identificador_Sin	Identificador do Sinistro	Nominal
AM	Identificador do Ato Médico	Nominal
Descritivo	Descrição do Ato Médico	Nominal
Tipo_Codigo	Tipo de Sinistro	Nominal
Data_Ocorrencia	Data de ocorrência do Sinistro	Ordinal
Grupo_Prestador	Grupo de Prestador	Nominal
Prestador	Hospital/Clínica onde o sinistro ocorreu	Ordinal
Domicilio	Hospital onde o sinistro ocorreu	Ordinal
Servico	Descrição do Sinistro	Nominal
Grupo_Am	Grupo de Ato Médico	Nominal
Especialidade	Especialidade do Sinistro	Nominal
Valor_Total	Valor total do Ato Médico	Contínua
Valor_Seguradora	Valor pago pela Seguradora	Contínua
Mes_Ano_Ocor	Mês e Ano da Data de Ocorrência do sinistro	Ordinal

4.2 Pré-Processamento das Bases de Dados

De forma a preparar as bases para a Análise Exploratória e para a aplicação das metodologias anteriormente referidas, foi necessário realizar um pré-processamento, de forma a corrigir algumas possíveis falhas existentes e também "criar" variáveis necessárias que não estão disponíveis nas bases originais.

Em primeiro lugar, verificou-se a existência de valores em falta, posteriormente a existência de valores que não faziam sentido, ou seja a existência de incoerências e por fim executou-se alguns agrupamentos de forma a reduzir a complexidade das bases de dados. O passo seguinte foi dedicado à criação de variáveis necessárias para o prosseguimento do estudo.

4.2.1 Bases BD_1 e BD_2

Relativamente às bases BD_1 e BD_2, verificou-se incoerências e valores em falta nas seguintes variáveis:

Data de fim da pessoa na apólice → A incoerência deve-se ao facto de existirem pessoas cuja apólice está cancelada e a sua data de fim não estar preenchida. Para estes casos admitiu-se que a data de fim destas pessoas coincide com a data da última anuidade da apólice a que pertence. Para as apólices e segurados que ainda estão ativos, ou seja para aqueles cuja data de fim não está preenchida, foi admitido que estes saíram no dia 30/9/2023 apenas para conseguir retirar algumas das conclusões a seguir apresentadas.

Prémio Comercial → Nesta variável importava perceber quais as pessoas que têm prémio igual a 0, isto porque este cenário só acontece com pessoas com menos de 15 anos. Assim para os prémios iguais a 0, observou-se a idade destas pessoas e chegou-se à conclusão que existiam pessoas com mais de 15 anos com prémio nulo. De forma a não influenciar os resultados decidiu-se eliminar estas pessoas. Em termos percentuais, nos planos estas pessoas representam 2.3% das observações e nos seguros apenas 0.14%.

Para as campanhas em vigor, entendeu-se por bem agrupar campanha semelhantes como por exemplo BlackFriday2020, BlackFriday2021 entre outras por forma a reduzir o número de campanhas existentes e assim simplificar. Consequentemente, as variáveis *Protocolo*, *Valor desconto* e *Duracao_Campanha* sofreram também simplificações. Admitiu-se, entre outros pressupostos que todas as apólices com valor em falta na variável *Campanha* não têm nenhuma campanha associada. Nos planos, após o agrupamento reduziu-se de 84 para 39 campanhas e nos seguros reduziu-se de 18 para 12 campanhas.

Adicionou-se a estas bases as variáveis *Anuidades_Seg_Ativo*, *Anuidades_Ap_Ativa* e *Idade*. As duas primeiras representam, respetivamente, o número de anos que o segurado esteve ativo e o número de anuidades que a apólice esteve ativa. Devido a algumas incoerências, existiam apólices e segurados com valor nulo nestas variáveis, pelo que se assumiu em ambos os casos que tinham estado ativos 1 mês.

4.2.2 Bases BD_3 e BD_4

Nas bases BD_3 e BD_4, verificou-se incoerências e valores em falta nas seguintes variáveis:

Grupo Prestador e Especialidade → Aos sinistros cujo grupo prestador não se encontrava identificado foi atribuído o Grupo "Outro". Da mesma forma aos sinistros cuja especialidade não se encontra especificada foi atribuída a especialidade "Outra".

Valor Total e Valor Seguradora → Verificou-se a existência de valores incoerentes, como valores negativos nestas duas colunas. Identificaram-se 13 e 4 atos médicos com valores negativos na BD_3 e BD_4, respetivamente. A sua existência pode ser justificada por ajustes feitos, no entanto por forma a não influenciar de forma negativa os resultados optou-se por retirar estes atos médicos da base. Em termos percentuais, nos planos estes representam 0.0101% das observações e nos seguros apenas 0.0112%.

Adicionaram-se a estas bases as seguintes variáveis:

- *Num_total_sin* → Número total de sinistros ocorridos de 2018-2021;
- *Num_sin_18* → Número total de sinistros ocorridos em 2018;
- *Num_sin_19* → Número total de sinistros ocorridos em 2019;
- *Num_sin_20* → Número total de sinistros ocorridos em 2020;
- *Num_sin_21* → Número total de sinistros ocorridos em 2021;
- *Media_Sin_ano* → Média de sinistros por ano;
- *Valor_total_gasto* → Valor total gasto pelo cliente desde 2018 a 2021;
- *Valor_total_gasto_ano* → Valor total gasto pelo cliente por ano;
- *Data_Ultima_Ocorrencia* → Data de ocorrência do último sinistro
- *Dias_Decorridos* → Dias decorridos desde a data do último sinistro até à data 31/12/2021

Assumiu-se na variável *Dias_Decorridos*, para as pessoas que não tiveram sinistros o valor de 1460. Representa o número de dias decorridos de 1/1/2018 a 31/12/2021.

4.2.3 Criação das Bases de Dados

De maneira a ir ao encontro do objetivo que se pretende cumprir, escolheu-se fazer duas análises, uma por apólices e outra por clientes.

Uma vez que se tem bases separadas para os clientes que possuem Planos e Seguros, e que estas não têm as mesmas variáveis, decidiu-se separar por Planos e Seguros. Esta decisão permite retirar conclusões separadamente sobre os clientes detentores de planos e seguros, e posteriormente concluir sobre todos os clientes em conjunto.

Como consequência foram criadas quatro bases de dados, a partir das variáveis iniciais e das variáveis adicionais. Estas bases foram criadas com a finalidade de melhor se explorar e aplicar as metodologias mencionadas. Assim, por fim tem-se as seguintes bases:

- **Base de dados representativa das Apólices (BD_AP_PLANOS)**: Informação sobre as apólices de Planos, constituída por 13 348 apólices e 21 variáveis
- **Base de dados representativa das Apólices (BD_AP_SSL)**: Informação sobre as apólices de Seguros, constituída por 3 511 apólices e 24 variáveis
- **Base de dados representativa dos Segurados (BD_SEG_PLANOS)**: Informação sobre os segurados detentores de Planos, constituída por 23 032 segurados e 31 variáveis
- **Base de dados representativa dos Segurados (BD_SEG_SSL)**: Informação sobre os segurados detentores de Seguros, constituída por 6 904 segurados e 32 variáveis

Capítulo

5 Análise Exploratória dos Dados

É de extrema importância, realizar em primeiro lugar uma análise exploratória de dados, uma vez que tem como finalidade a descoberta de padrões, características, tendências, comportamentos atípicos e outliers, permitindo assim entender os dados e as relações existentes entre as diferentes variáveis.

À vista disso, este capítulo segue uma análise aprofundada das apólices de seguro e dos segurados presentes nas bases de trabalho.

5.1 Análise por Apólice

Para se perceber melhor, quais as tendências e comportamentos das apólices, realizou-se uma análise às bases BD_AP_PLANOS e BD_AP_SSL. As conclusões a que se chegou são apresentadas de seguida.

Tal como referido na seção 4, a base dos Planos é constituída por 13348 apólices distintas e a dos Seguros é constituída por 3511 apólices distintas.

5.1.1 Número de Pessoa Seguras

Nos planos, em média cada apólice é constituída por aproximadamente 2 pessoas, no mínimo cada apólice tem associada uma única pessoa, e nesta situação existem 7920 apólices. Por sua vez, 690 pessoas seguras é o máximo que ocorre, sendo que apenas existe uma apólice nesta situação.

Já nos seguros, a média e o mínimo de pessoas seguras por apólice mantém-se. O máximo é de 110 pessoas seguras e existe também uma só apólice nesta situação.

Apresenta-se nas figuras 5.1 e 5.2, a distribuição de pessoas seguras tanto nas apólices de planos como de seguros.

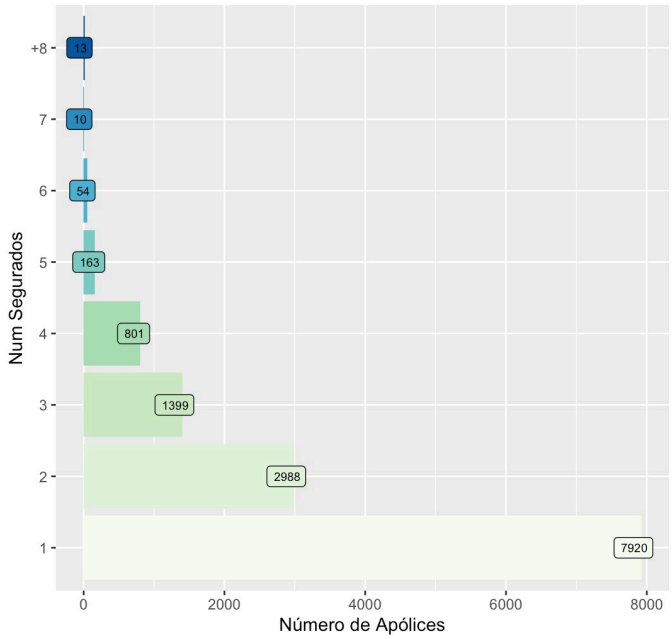


Figura 5.1: Número de apólices (Planos) por número de pessoas seguras

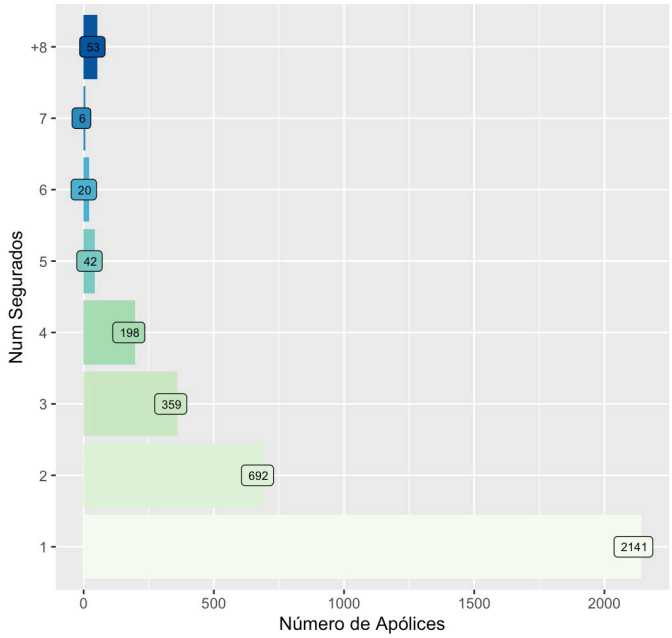
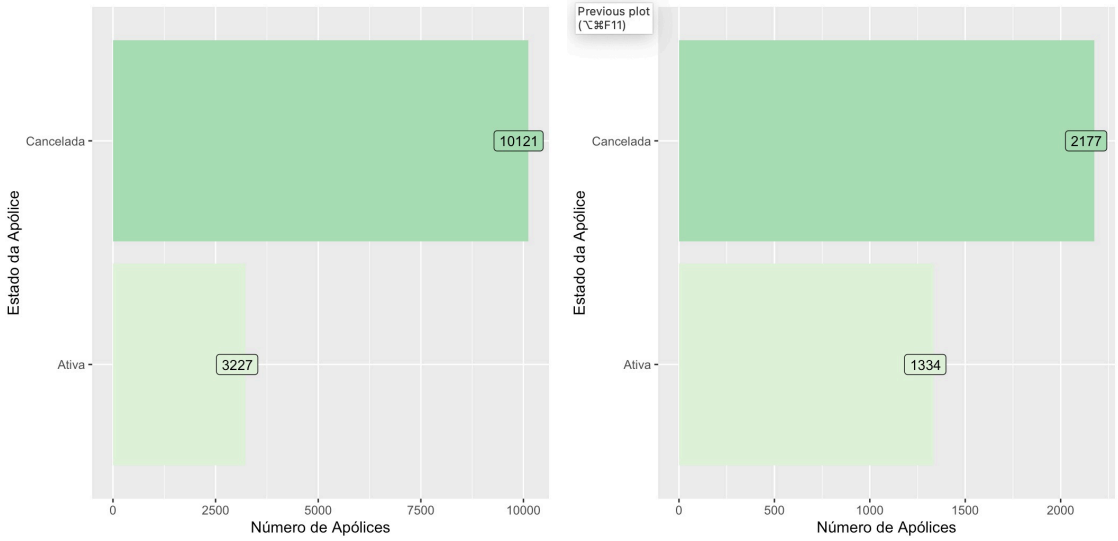


Figura 5.2: Número de apólices (SSL) por número de pessoas seguras

5.1.2 Estado da Apólice

Ao analisar o estado da apólice, presente na figura 5.3, concluiu-se que mais de metade das apólices tanto nos planos como nos seguros se encontram canceladas. Este facto contribui significativamente para a baixa taxa de retenção e pode dever-se a vários fatores,

nomeadamente à falta de campanhas direcionadas o que realça o pouco conhecimento que se detém dos clientes.



(a) Planos (b) SSL

Figura 5.3: Número de Apólices ativas e canceladas

Como se observa na figura 5.3, nos planos as apólices canceladas representam 75.82% e nos seguros 62.01%.

5.1.3 Tipo de produto

A análise desta variável é fundamental para se perceber quais os produtos mais comuns, os menos comuns e posteriormente quais os produtos com melhores resultados, isto é das apólices que estão ativas quais os produtos associados. Permite também aprofundar o conhecimento para na etapa seguinte descrever o perfil do cliente de forma mais consciente.

De seguida, apresenta-se nas figuras 5.4 e 5.5 a distribuição de apólices por produto.

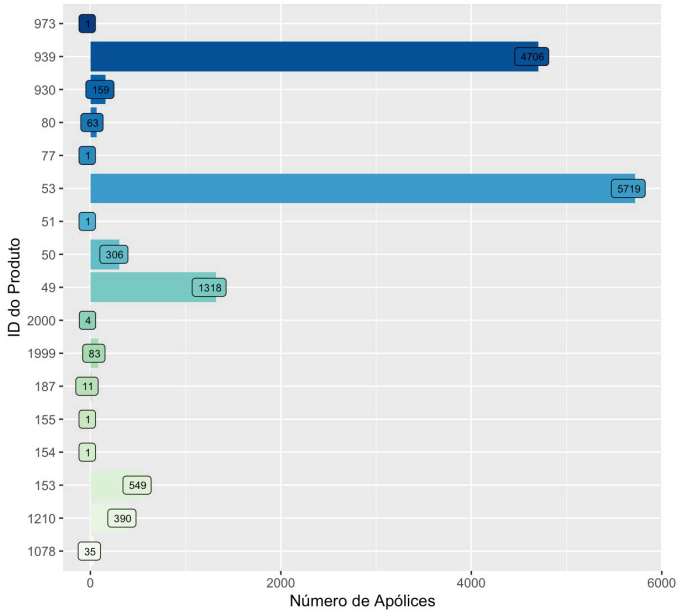


Figura 5.4: Número de apólices (Planos) por produto

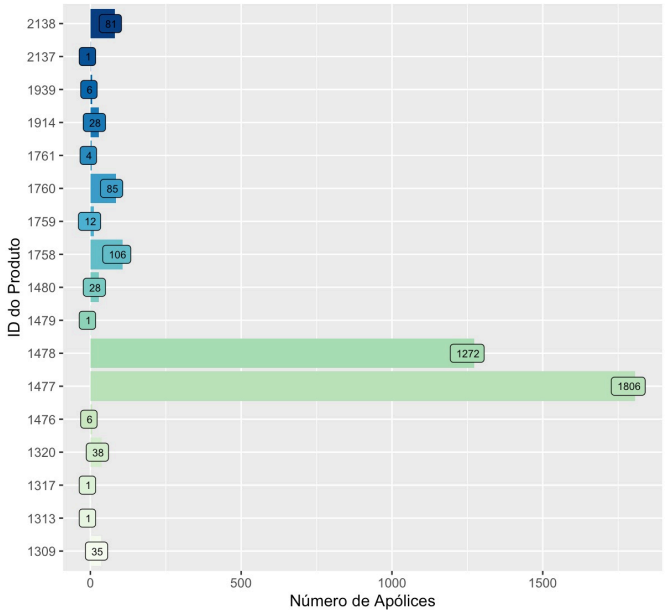


Figura 5.5: Número de apólices (SSL) por produto

Pela análise das figuras 5.4 e 5.5 tem-se relativamente, aos produtos comercializados para planos, é visível que os mais comuns são o 53 e o 939, representando cerca de 42.85% e 35.26% das apólices, respetivamente. Os menos populares são os produtos 77, 973, 51, 155, 154, 2000 e o 187 que representam 0.15% das apólices.

Assim, é possível entender-se que as características dos produtos 53 e 939 são mais apelativas ao grupo de clientes que possui planos e portanto no futuro poder-se-á comercializar produtos que contenham características semelhantes.

De forma análoga, deduz-se que os produtos 77, 51, 155, 2000, 1078 e 1210 são estão associados apenas a apólices canceladas pelo que poderão ser produtos não cativantes e com características de pouco interesse.

Verificou-se ainda que a percentagem de apólices ativas que têm associado os produtos 53 e 939 é 23.68% e 23.01%, respetivamente. De todos os produtos concluiu-se ainda que, o produto com maior percentagem de apólices ativa, cerca de 50% é o 153, e portanto este é também um produto com interesse.

Já nos seguros, os produtos mais comuns são o 1477 e 1478, que representam 51.44% e 36.23%, respetivamente. Contrastando com os referidos, estão os produtos 2137,1479,1317, 1313 e 1761 que se encontram entre os menos vendidos. De igual forma é possível concluir quais as características que mais interesse têm para o grupo de clientes que possui um seguro.

5.1.4 Campanha associada

Para os planos, após o agrupamento ficou-se com 39 campanhas. A distribuição de apólices por campanha é apresentada na figura 5.6. Nota-se, que não se considerou para a elaboração do seguinte gráfico as campanhas que têm menos de 20 apólices associadas.

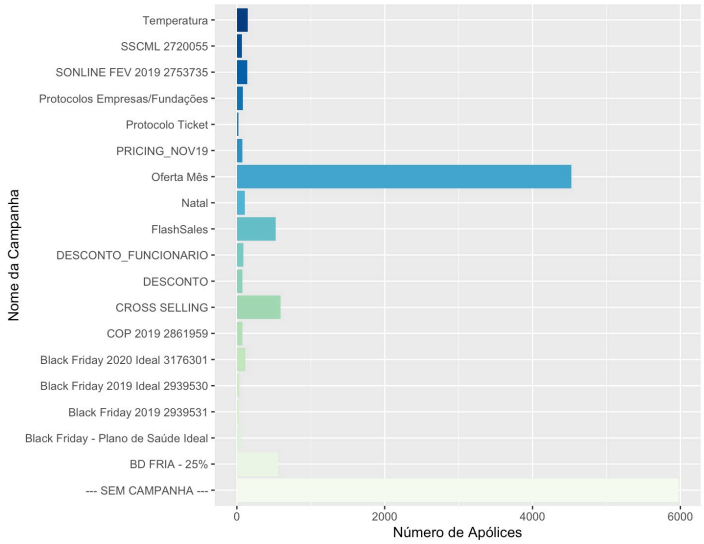


Figura 5.6: Número de apólices (Planos) por campanha

A campanha mais comum é a “oferta mês”, no entanto das que mais interesse têm, dado à sua duração ou valor de desconto, as mais comercializadas são "Cross Selling", "FlashSales"e a "BD Fria-25%". Tal como se esperaria a maioria das apólices não tem nenhuma campanha associada.

No decorrer, realizou-se uma análise mais aprofundada das campanhas com maior interesse, com a finalidade de perceber o efeito que estas têm na durabilidade das apólices

assim como dos segurados.

Relativamente aos seguros, tem-se outro tipo de campanhas, que se apresentam no gráfico 5.7.

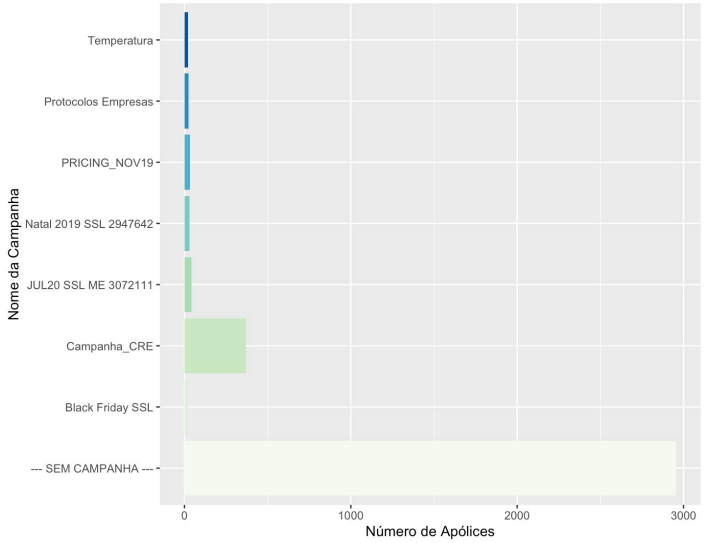


Figura 5.7: Número de apólices (SSL) por campanha

Tal como nos planos, o mais comum é não ter qualquer tipo de campanha associada. Das restantes campanhas, o número de apólices é muito semelhante com exceção da "Campanha_CRE".

5.1.5 Prémio Comercial anual

É importante para o estudo compreender qual o prémio pago quer por pessoa quer por apólice para perceber quais as apólices/pessoas que mais interesse têm para a companhia e assim direcionar estratégias de marketing.

Com esse intuito, para o prémio anual pago por apólice analisou-se as estatísticas básicas, assim como a existência de outliers e também o intervalo onde estão situadas a maior parte das apólices. Apresentam-se estes valores nas tabelas 5.1 e 5.2

Tabela 5.1: Estatísticas básicas do prémio comercial nos Planos

Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
0	84	144	157.10	180	8640

Tabela 5.2: Estatísticas básicas do prémio comercial nos Seguros

Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
0	78.14	100.47	136.60	167.44	4297.37

Nos planos, existem outliers, no entanto a sua presença é justificada pelo facto de esta variável se tratar do prémio comercial por apólice e não por segurado, assim sendo o prémio será tanto mais elevado quanto maior for o número de pessoas por apólice. Como a média de pessoas seguras por apólice é cerca de 2, mas existem apólices com 150 pessoas e até 690, é normal haver uma discrepância significativa. Justifica-se também, de igual forma o valor máximo de 8640€. A existência de outliers, nos seguros justifica-se de forma análoga.

Posteriormente, foram definidas 7 classes e atribui-se uma classe a cada apólice. Abaixo nas figuaras 5.8 e 5.9, estão representadas o número de apólices por classe.

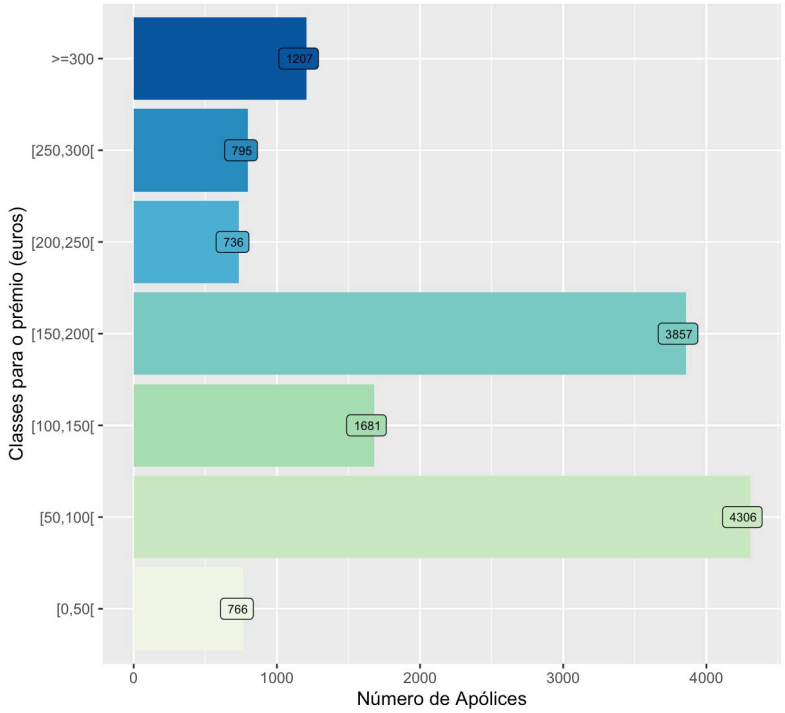


Figura 5.8: Número de apólices (Planos) por classe

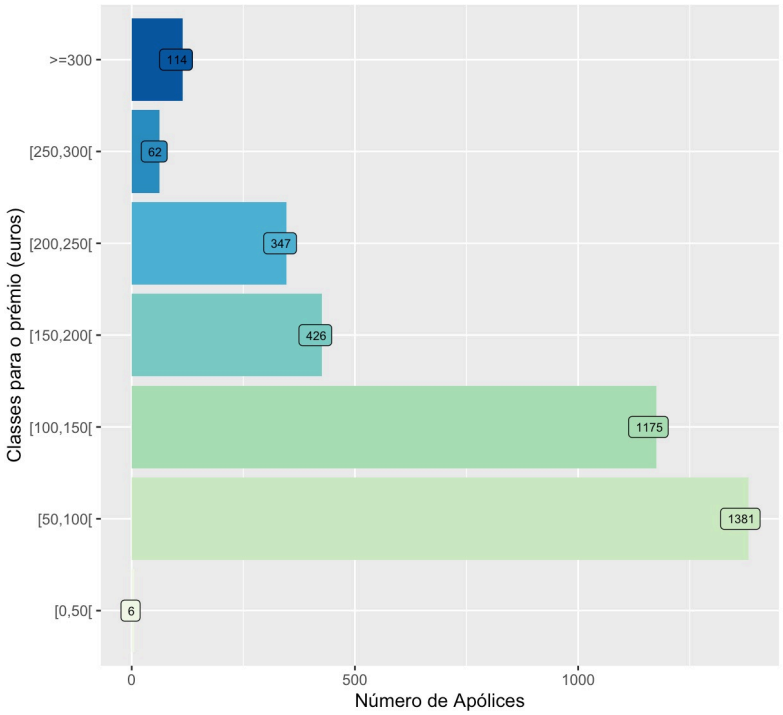


Figura 5.9: Número de apólices (SSL) por classe

Em ambos, tem-se que a classe mais comum é a [50,100[. Destaca-se a baixa concentração de apólices de seguros na classe dos prémios inferiores a 50€, e a concentração algo elevada das apólices de planos na classe dos prémios superiores a 300€. Mais uma vez, não é de suspeitar, dado a variável "Num_Segurados", que tem influência direta nesta variável.

5.1.6 Número de anuidades ativa

A fim de se perceber, qual a duração média de uma apólice, foi calculado para todas as apólices quantos anuidades estiveram ativas. Relembra-se que para as apólices que estão ativas, admitiu-se a data da última anuidade como sendo a 30/9/2023.

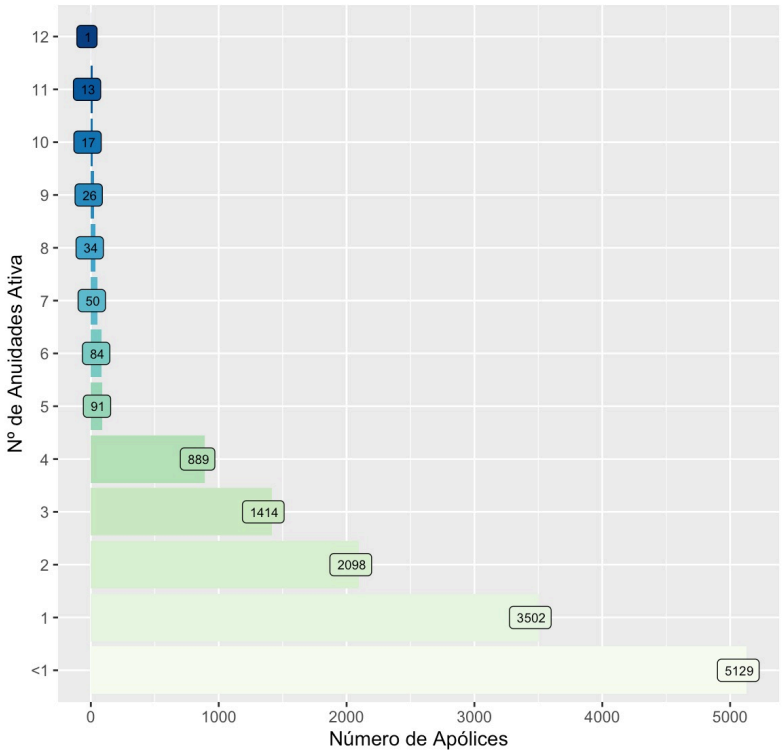


Figura 5.10: Número de apólices (Planos) por número de anuidades ativa

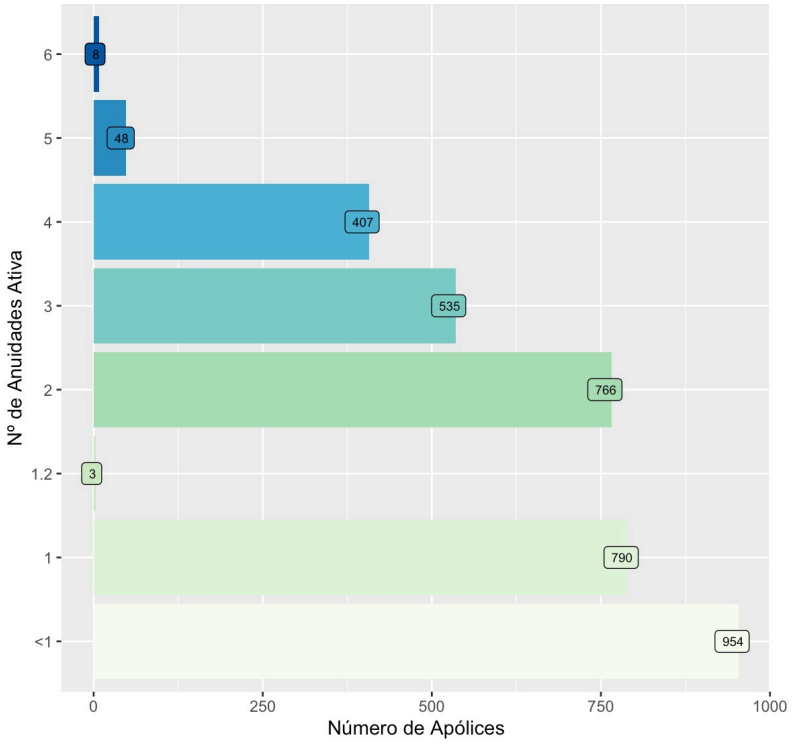


Figura 5.11: Número de apólices (SSL) por número de anuidades ativa

Constata-se pelas figuras 5.10 e 5.11 que uma grande parte das apólices, cerca de 38.43%, esteve ativa menos de 1 ano. Para além disso, mais de metade das apólices permanece ativa

entre uma e quatro anuidades. Nos seguros, por sua vez apenas 27% das apólices esteve ativa menos de um ano, sendo que a grande parte está ativa entre uma e quatro anuidades.

No entanto para uma melhor compreensão do número de anuidades ativa é necessário analisar este indicador para as apólices que já estão anuladas. Assim, apresenta-se em 5.12 e 5.13, a mesma informação apenas para as apólices canceladas.

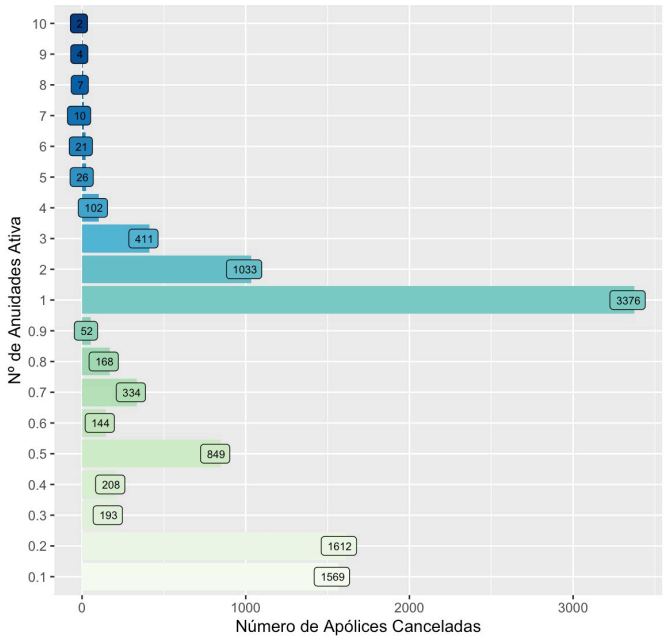


Figura 5.12: Número de apólices canceladas (Planos) por número de anuidades ativa

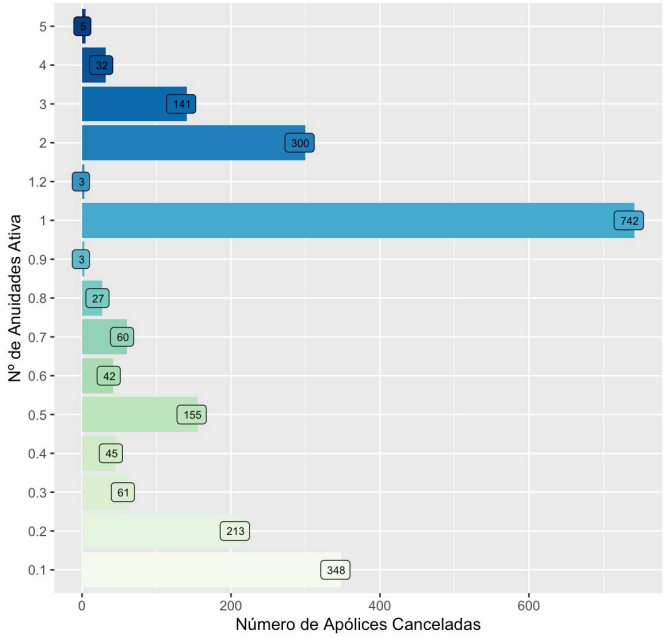


Figura 5.13: Número de apólices canceladas (SSL) por número de anuidades ativa

Com os gráficos anteriores é possível concluir a baixa duração das apólices anuladas, uma vez que as apólices que aparecem nas figuras 5.10 e 5.11 com menor duração, são todas apólices anuladas.

Chama-se também a atenção para dois factos com bastante interesse. Nos planos das 3502 apólices que estiveram ativas 1 anuidade, 3376 são apólices que já se encontram anuladas, já nos seguros das 790 que estiveram ativas 1 ano, 742 são apólices anuladas. De igual forma, nos planos as 5129 apólices com menos de uma anuidade de duração são todas apólices anuladas, o mesmo acontece com as apólices de seguros.

Deste modo, ao perceber os motivos que levaram à anulação, e agindo nesse sentido, é possível aumentar o número de anuidades ativa pelo menos para 2 e assim aumentar as possibilidades de fidelização.

5.1.7 Número de Sinistros total

Pela tabela 5.3 e figura 5.15 tem-se relativamente ao número de sinistros total, no horizonte temporal considerado (período de 2018 a 2021), que a média de sinistros por apólice nos planos é de aproximadamente 5 sinistros por apólice. Ao observar, a figura 5.14, observa-se que o máximo de sinistros ocorridos é claramente um outlier, no entanto como a apólice tem associada 150 pessoas, dá em média 3 sinistros por pessoa, estando assim justificada a sua presença.

Tabela 5.3: Estatísticas básicas do número de sinistros total nos planos

Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
0	0	2	4.62	6	488

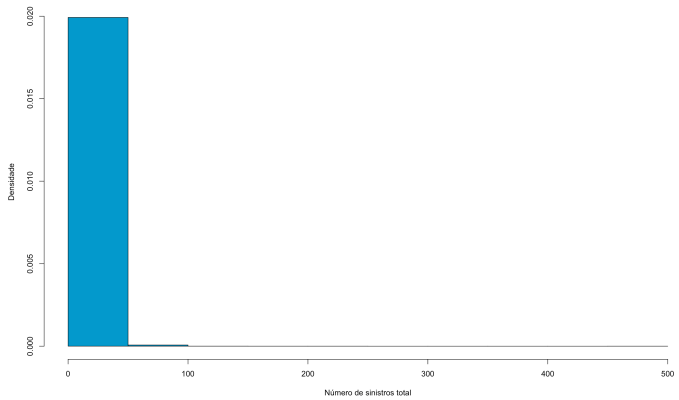


Figura 5.14: Número de sinistros total (Planos)

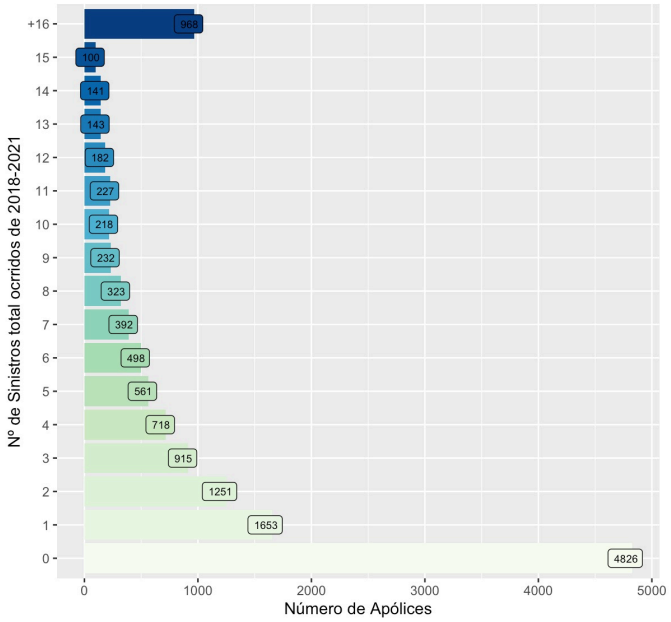


Figura 5.15: Número de apólices (Planos) por número de sinistros

Já nos SSL, pela tabela 5.4 e figuras 5.16 e 5.17, tem-se que o comportamento dos sinistros é bastante semelhante, com a média de 5 sinistros por apólice e a mediana de 2 sinistros. Já o máximo é significativamente mais baixo, dado que o número de pessoas por apólice é também inferior quando comparado com as apólices de planos.

Tabela 5.4: Estatísticas básicas do número de sinistros total nos SSL

Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
0	0	2	4.98	6	88

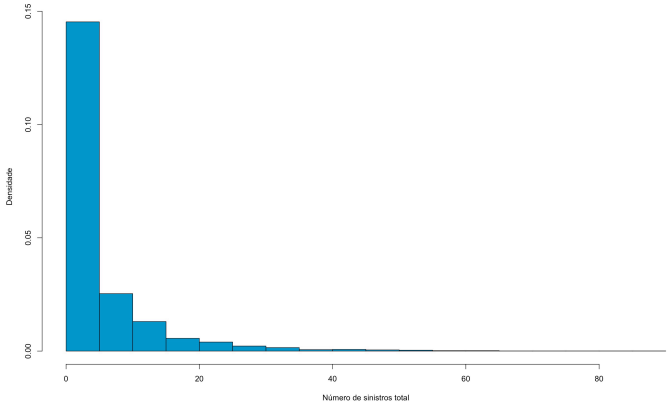


Figura 5.16: Número de sinistros total (SSL)

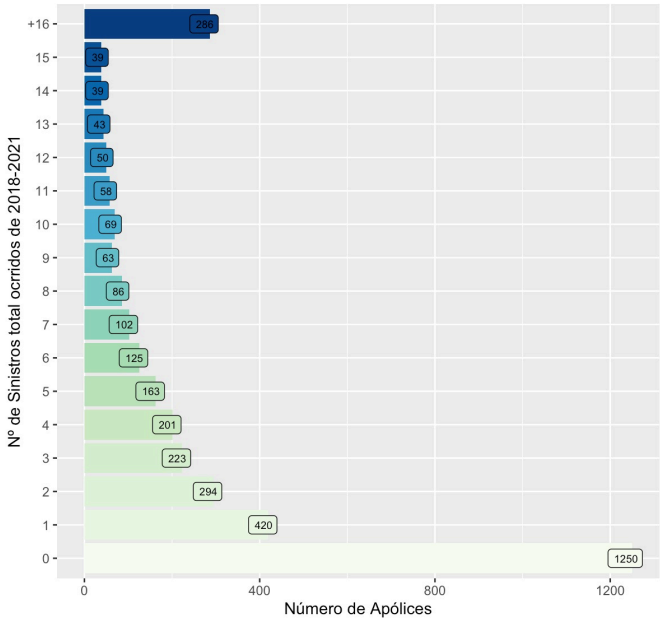


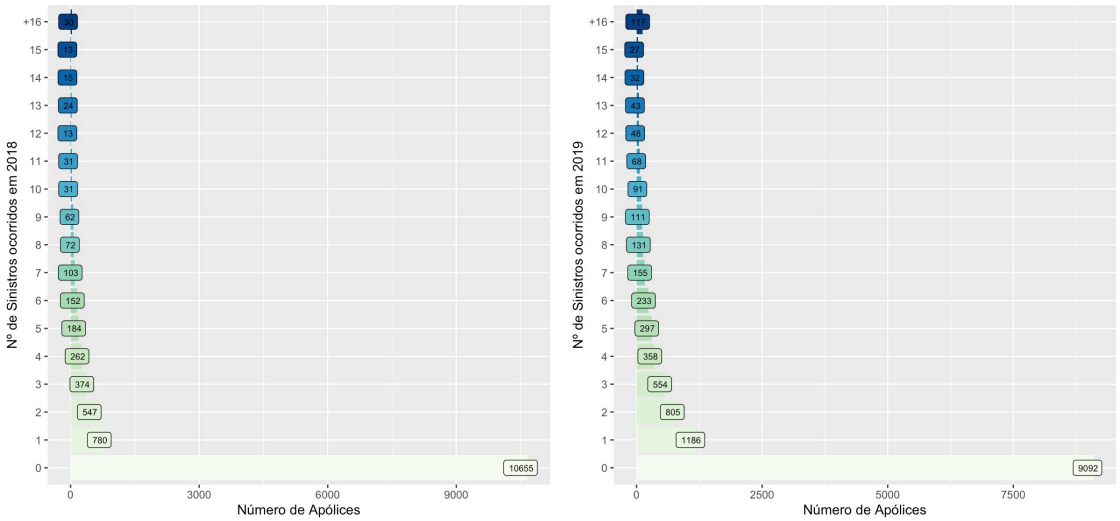
Figura 5.17: Número de apólices (SSL) por número de sinistros

Já de seguida, estuda-se para cada um dos anos considerado o comportamento do número de sinistros para uma melhor compreensão do número de sinistros por ano.

5.1.8 Número de sinistros ocorridos por ano (Planos)

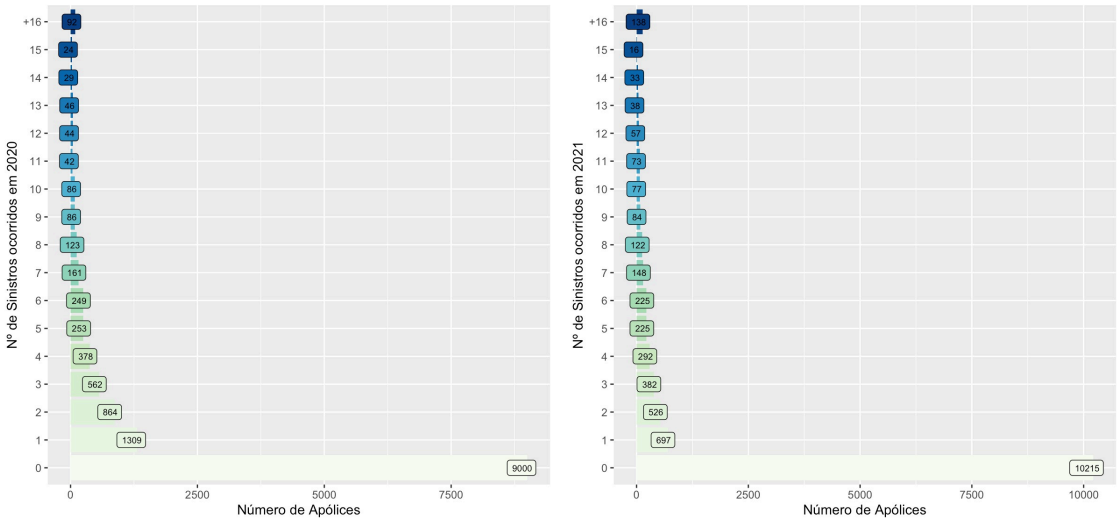
Tabela 5.5: Total e média de sinistros por ano (Planos)

Ano	2018	2019	2020	2021
Média de Sinistros	0.746	1.354	1.296	1.219
Total de Sinistros	9957	18072	17298	16275



(a) Ano 2018 (b) Ano 2019

Figura 5.18: Número de sinistros por apólice



(a) Ano 2020 (b) Ano 2021

Figura 5.19: Número de sinistros por apólice

Como se observa pelas figuras 5.18 e 5.19 e tabela 5.5, pode concluir-se que em 2019 foi o ano em que ocorreram mais sinistros, sendo também em 2019 que a média de sinistros por apólice é maior, tendo-se em média aproximadamente 1 sinistro. Constata-se que o número médio de sinistros por apólice é muito idêntico para os quatro anos estudados pelo que se poderá afirmar que para esta carteira, a média ronda um sinistro por apólice.

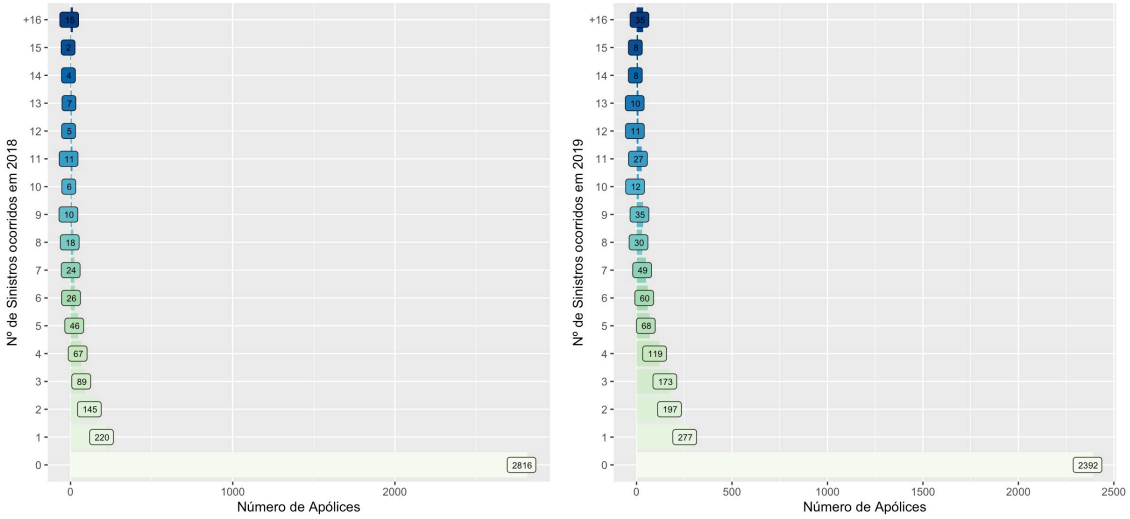
O número total de sinistros ocorridos flutuou, tendo subido em 2019 e descido a partir daí, pelo que será necessário estudar esta variável num horizonte temporal maior para se concluir com melhor precisão qual o intervalo de valores em que se situa. E tal como se esperaria, em todos os anos, o número de apólices sem sinistros é bastante elevado,

representado sempre mais de 67% das apólices.

5.1.9 Número de sinistros ocorridos por ano (SSL)

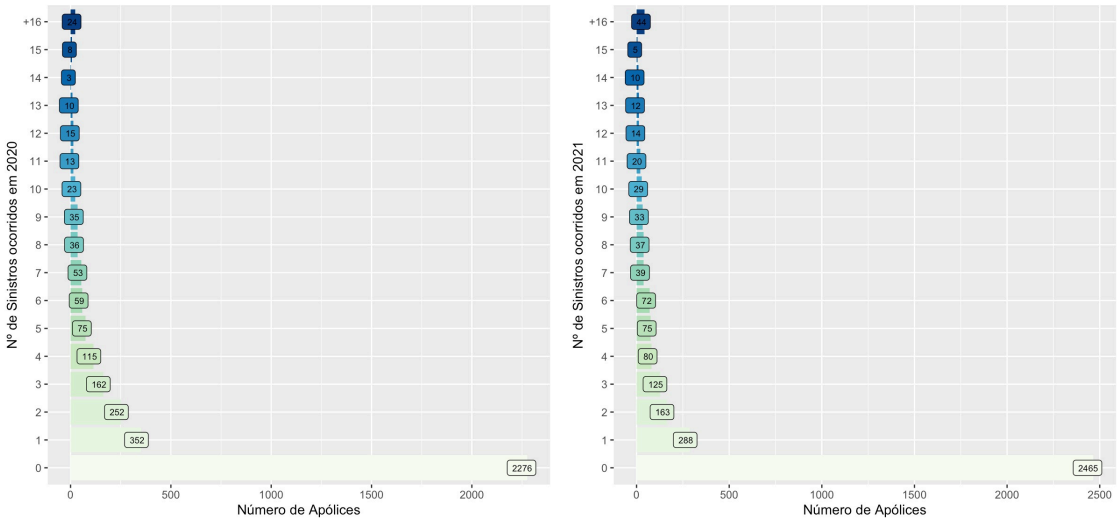
Tabela 5.6: Total e média de sinistros por ano (SSL)

Ano	2018	2019	2020	2021
Média de Sinistros	0.757	1.41	1.395	1.418
Total de Sinistros	2656	4952	4897	4980



(a) Ano 2018 (b) Ano 2019

Figura 5.20: Número de sinistros por apólice



(a) Ano 2020 (b) Ano 2021

Figura 5.21: Número de sinistros por apólice

Já nos seguros com base nas figuras 5.20 e 5.21 e na tabela 5.6, concluiu-se que foi em 2021 que ocorreram mais sinistros, sendo também em 2021 que a média de sinistros por apólice é maior, tendo-se em média aproximadamente 1 sinistro.

Repete-se também as conclusões retiradas relativamente ao número médio de sinistros por apólice, isto é pode afirmar-se que para esta carteira, a média ronda um sinistro por apólice. A diferença evidente reside no número total de sinistros ocorridos, que é bastante inferior ao que acontece nos planos. Este sofre um aumento significativo em 2019, baixa ligeiramente em 2020, e volta a subir em 2021. A partir de 2019, o número de sinistros parece rondar os 4900, no entanto para uma melhor precisão deste valor, é essencial estudar o seu comportamento num horizonte temporal maior. Mais uma vez, tem-se em todos os anos, que o número de apólices sem sinistros é bastante elevado, representado sempre mais de 64% das apólices.

5.2 Análise por Segurado

Nesta secção, de forma semelhante à anterior analisou-se as bases BD_SEG_PLANOS e BD_SEG_SSL para se perceber melhor, quais as tendências e comportamentos das pessoas seguras. As conclusões a que se chegou são apresentadas de seguida.

Tal como referido na seção 4, a base dos Planos é constituída por 23 032 pessoas distintas e a dos Seguros é constituída por 6 904 pessoas distintas.

5.2.1 Campanha associada

A importância de analisar a distribuição de pessoas seguras por campanha deve-se ao facto de um dos principais objetivos passar por traçar o perfil do cliente. Assim ao analisar

o gráfico 5.22, entende-se quais as campanhas mais comuns entre as pessoas. Nota-se, que não se considerou para a elaboração do seguinte gráfico as campanhas que têm menos de 20 segurados associadas.

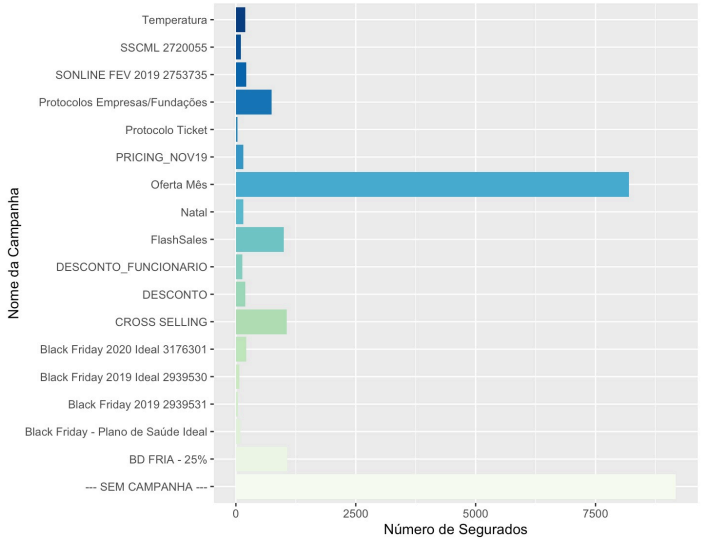


Figura 5.22: Número de pessoas seguras (Planos) por campanha

A campanha mais comum é a "oferta mês", no entanto das que mais interesse têm, dado à sua duração ou valor de desconto, as mais comercializadas são tal como se concluiu na análise por apólice, a campanha de Cross Selling, FlashSales e BD Fria-25%. A maioria das pessoas não tem nenhuma campanha associada.

Relativamente aos seguros a distribuição de pessoas seguras por campanha está presente na figura 5.23.

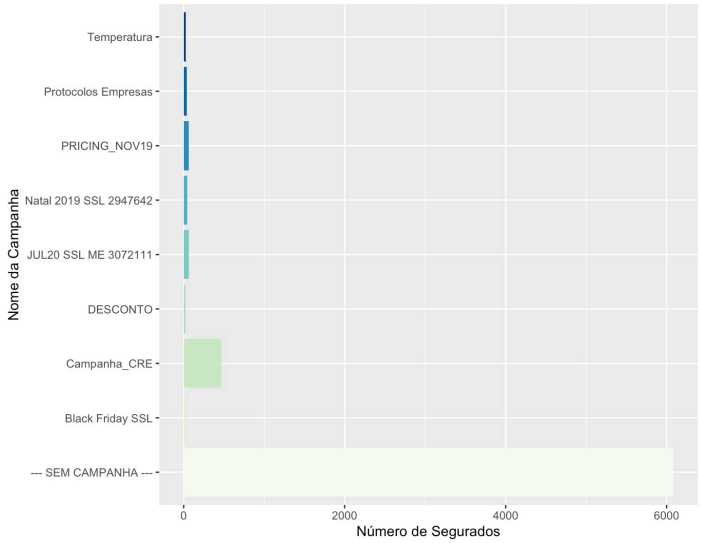


Figura 5.23: Número de pessoas seguras (SSL) por campanha

Assim como na análise por apólices, as conclusões retiradas são exatamente iguais.

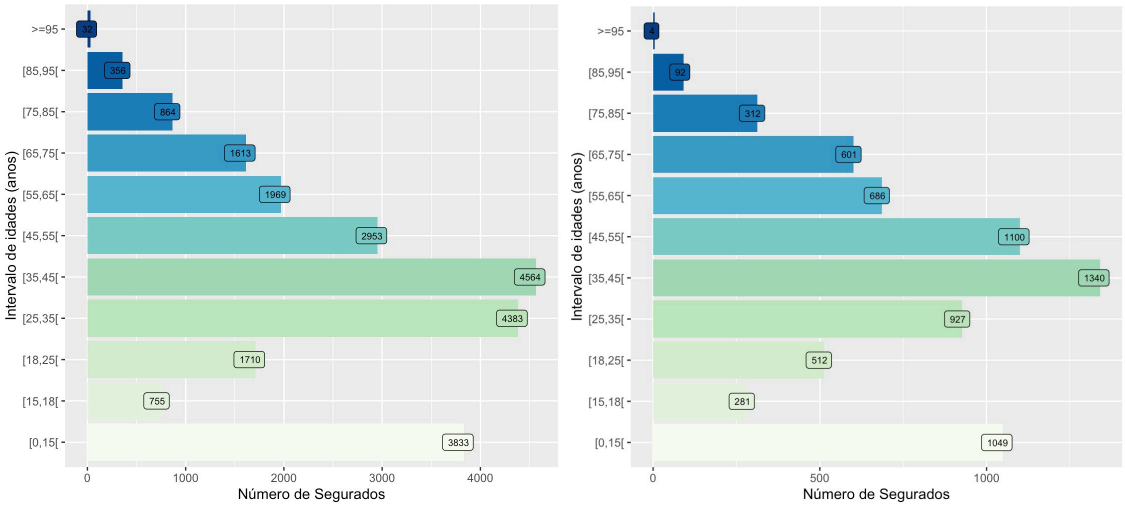
Consta-se assim que, para as campanhas analisar as apólices ou as pessoas seguras é idêntico.

5.2.2 Idade

Para se entender qual o intervalo de idades mais comum na carteira, estudou-se a idade das pessoas seguras. Nos planos, a média de idade é aproximadamente 38 anos, a idade mínima como se esperava é 0 anos e a idade máxima é de 111 anos, existindo apenas uma pessoa nesta situação.

Já nos seguros, a média é aproximadamente 40 anos. O mínimo é de 0 anos e a idade máxima é de 97 anos, situação que ocorre três vezes.

De seguida apresenta-se, na figura 5.24 a distribuição de pessoas seguras por intervalo de idade, tanto nos planos como nos seguros.



(a) Planos (b) SSL

Figura 5.24: Número de pessoas seguras por intervalo de idade

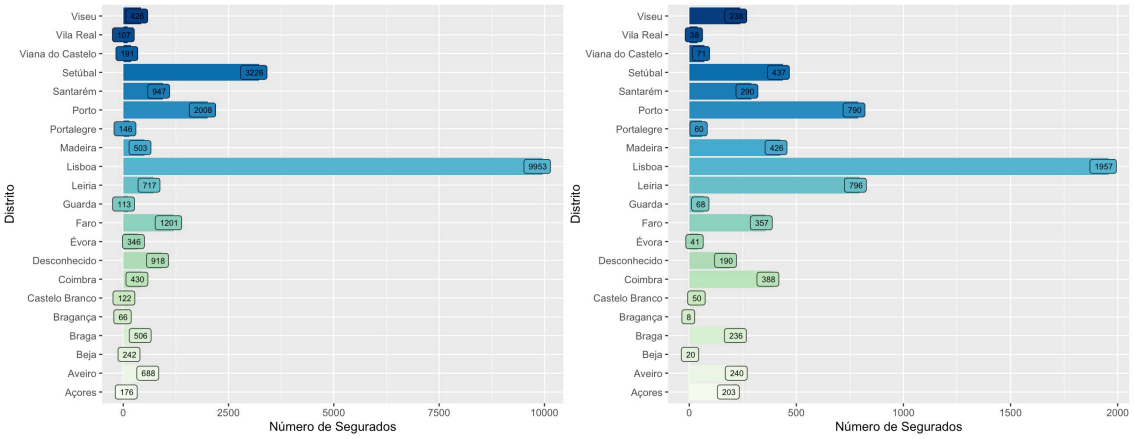
Nos planos verifica-se que a classe de idades mais representativa na carteira em estudo é a classe dos [35,45[, seguida da classe dos [25,35[. Cerca de 39% dos segurados tem entre 25 anos (inclusive) e 45 anos (exclusive). Por sua vez, a classe de idade menos representada na carteira é a classe dos maiores de 95 anos de idade, representam cerca de 0.14% da carteira.

Já nos seguros, a classe mais representativa é também a classe dos [35,45[, seguida neste caso da classe dos [45,55[. Cerca de 35% dos segurados tem entre 35 anos (inclusive) e 55 anos (exclusive). Por sua vez, a classe de idade menos representada na carteira é a classe dos maiores de 95 anos de idade, representam cerca de 0.058% da carteira.

5.2.3 Distrito

Avaliou-se também o distrito das pessoas seguras na carteira em estudo uma vez que esta variável é importante para definir o perfil do cliente e por forma a aprofundar o conhecimento que se detém dos clientes.

De seguida apresenta-se na figura 5.25, a distribuição de pessoas seguras por distrito.



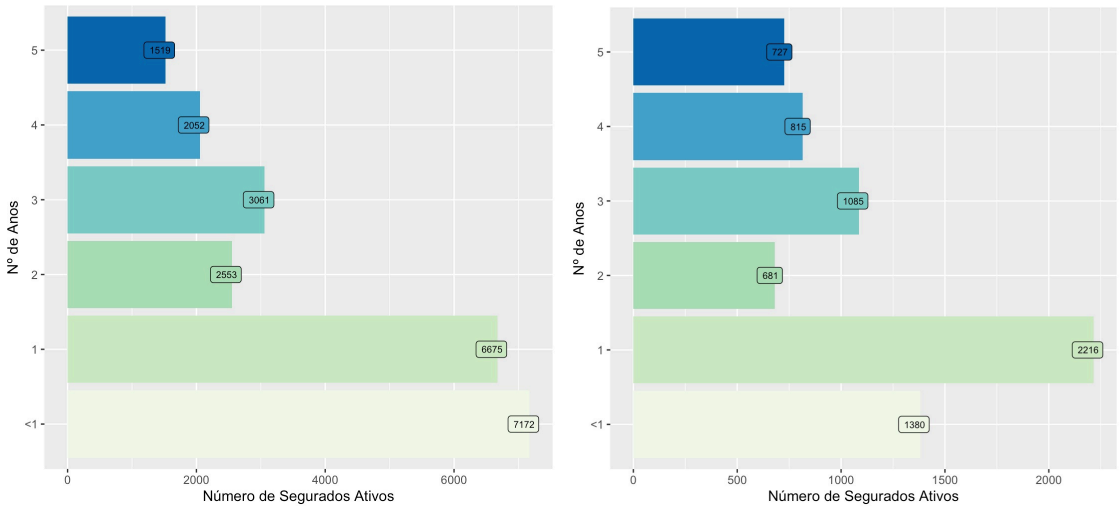
(a) Planos (b) SSL

Figura 5.25: Número de pessoas seguras por distrito

Tanto nos planos como nos seguros, e tal como se previa o distrito de Lisboa é o mais comum. Cerca de 57% e 35% dos clientes são dos distrito de Lisboa e Setúbal. Nos seguros contrariamente ao que acontece nos planos o segundo distrito com mais representatividade é Leiria. O distrito de Bragança é em ambos o distrito menos representado.

5.2.4 Número de anuidades Ativo

A fim de se perceber, qual a duração média de uma pessoa numa apólice, foi calculado para todas as pessoas quantas anuidades estiveram ativas. Relembra-se que para as pessoas que estão ativas, admitiu-se a data de saída da apólice como sendo a 30/9/2023.

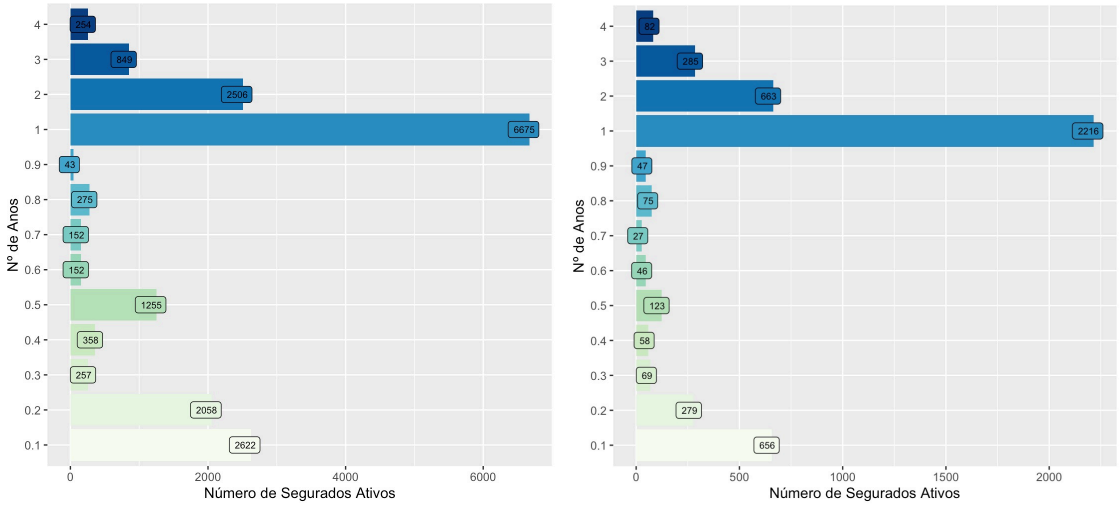


(a) Planos (b) SSL

Figura 5.26: Número de pessoas seguras por número de anuidades ativo

Doa figura 5.26, depreende-se que tanto nos planos como nos seguros, mais de 50% das pessoas seguras permanece ativa no máximo uma anuidade. No entanto, para uma melhor percepção, avalia-se apenas as pessoas que já não se encontram ativas, uma vez que podem existir pessoas ativas a ser consideradas que tenham entrado na anuidade anterior, estando por isso ativas apenas à uma anuidade.

No entanto para uma melhor compreensão do número de anuidades que uma pessoa esteve ativa é necessário analisar este indicador para as pessoas que de facto já anularam a sua subscrição. Assim, apresenta-se na figura 5.27, a mesma informação apenas para as pessoas não ativas.



(a) Planos (b) SSL

Figura 5.27: Número de pessoas seguras não ativas por número de anuidades ativo

Dos gráficos anteriores é possível concluir que das pessoas seguras que permaneceram ativas, uma anuidade ou menos, são todas pessoas que já não se encontram ativas. Desta forma, ao perceber os motivos que levaram à anulação, e agindo nesse sentido, é possível aumentar o número de anuidades ativa e assim aumentar as possibilidades de fidelização.

5.2.5 Prémio Comercial anual

Apresenta-se, de seguida, alguns resultados relacionados com o prémio comercial pago por pessoa. Tal como já foi referido, é importante analisar esta variável por forma a concluir quais os clientes de maior interesse e posteriormente direcionar estratégias de marketing.

Com esse intuito, para o prémio anual pago por pessoa analisou-se as estatísticas básicas, presentes nas tabelas 5.7 e 5.8, assim como a existência de outliers e também o intervalo onde estão situadas a maior parte das pessoas.

Tabela 5.7: Estatísticas básicas do prémio comercial nos Planos

Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
0	71.40	84	91.04	126	384

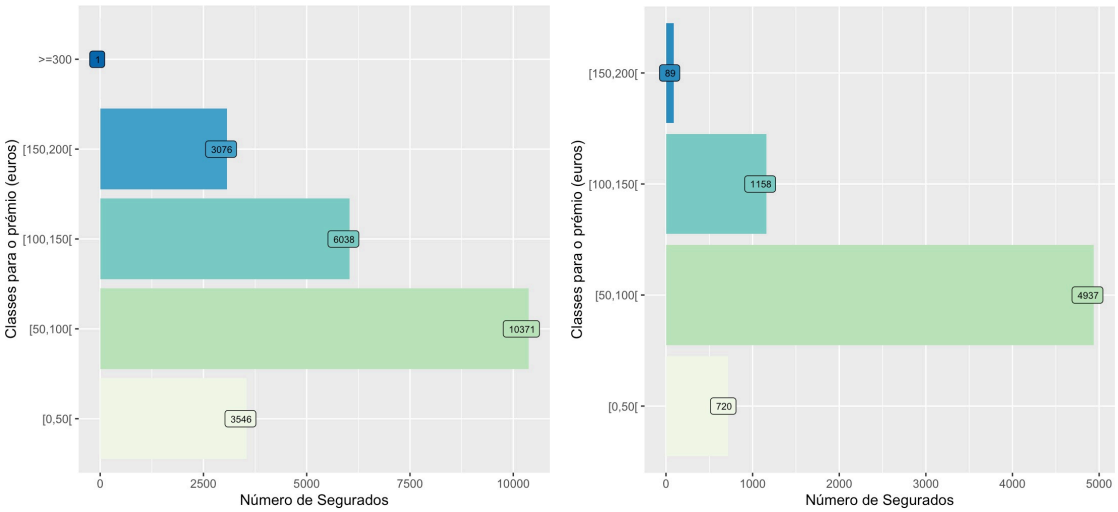
Tabela 5.8: Estatísticas básicas do prémio comercial nos Seguros

Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
0	55.81	70.33	69.47	89.30	180

Nos planos, existe claramente um outlier, no entanto a sua presença, pode dever-se ao facto de se tratar de uma pessoa com uma idade avançada que apenas está ativo há duas anuidades.

Ao analisar, em particular para esta pessoa, constata-se que se trata de um cliente "bom" para a companhia, uma vez que não apresenta sinistros. Assim, é importante identificar este tipo de clientes para depois desenvolver campanhas especiais.

Posteriormente, foram definidas 7 classes e atribui-se uma classe a cada pessoa. Na figura 5.28, estão representadas o número de pessoas por classe.



(a) Planos (b) SSL

Figura 5.28: Número de pessoas seguras por classe

Em ambos, tem-se que a classe mais comum é a [50,100[, seguida da classe dos [100,150[. Cerca de 71.24% e 88.28% dos segurados pagam entre 50 (inclusive) e 150 (exclusive) euros, nos planos e seguros respetivamente.

Destaca-se a baixa concentração de pessoas detentoras de planos na classe dos prémios superiores a 300€ e de pessoas detentoras de seguros na classe dos prémios superiores a 150€.

5.2.6 Número de Sinistros total

Relativamente ao número de sinistros total, no horizonte temporal considerado (período de 2018 a 2021), pela tabela 5.9 retira-se que a média de sinistros por pessoa nos planos é de aproximadamente 3 sinistros por pessoa. Ao observar, a figura 5.29, observa-se que o máximo de sinistros ocorridos é claramente um outlier, no entanto justifica-se a sua presença pelo facto de se tratarem de tratamentos de fisioterapia.

Tabela 5.9: Estatísticas básicas do número de sinistros total nos planos

Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
0	0	1	2.67	3	160

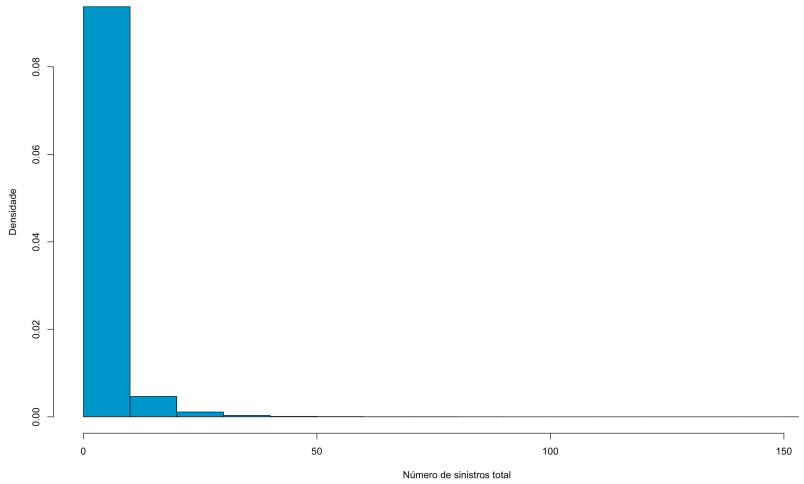


Figura 5.29: Número de sinistros total (Planos)

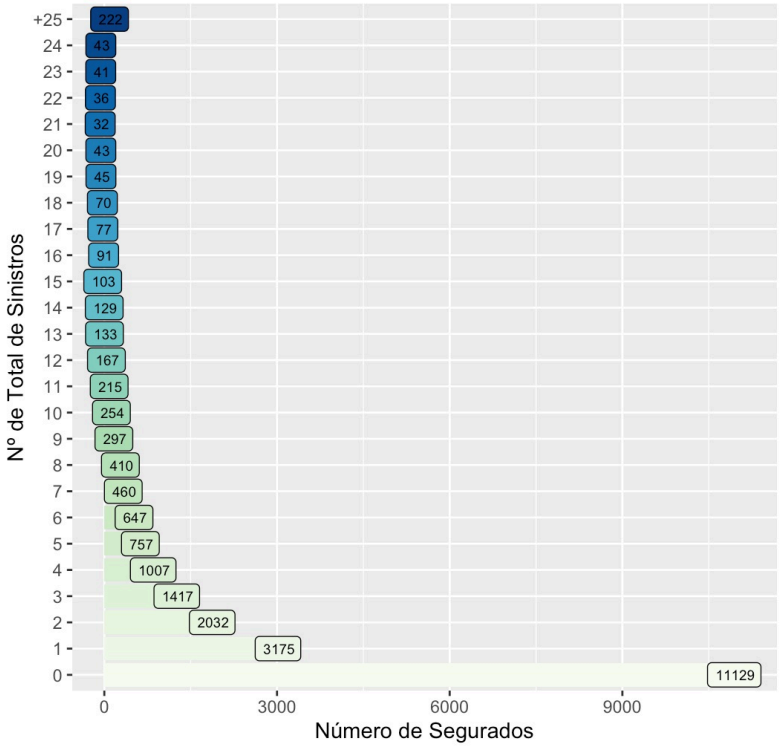


Figura 5.30: Número de pessoas seguras (Planos) por número de sinistros

Pela figura 5.30, observa-se que apesar de existirem 222 pessoas com mais de 25 sinistros, cerca de 77.08% das pessoas tem no máximo 3 sinistros.

Já nos SSL, comprova-se pela tabela 5.10 e figura 5.31 que o comportamento dos sinistros é distinto do que ocorre nos planos. Apesar de a média se manter 3 sinistros por pessoa, tem-se um valor máximo significativamente mais baixo.

Tabela 5.10: Estatísticas básicas do número de sinistros total nos planos

Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
0	0	0	2.53	3	68

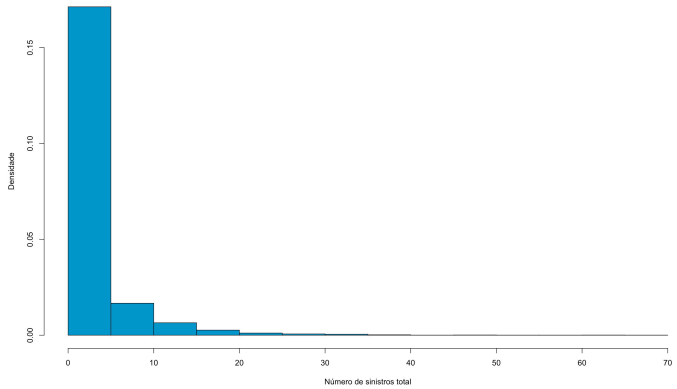


Figura 5.31: Número de sinistros total (SSL)

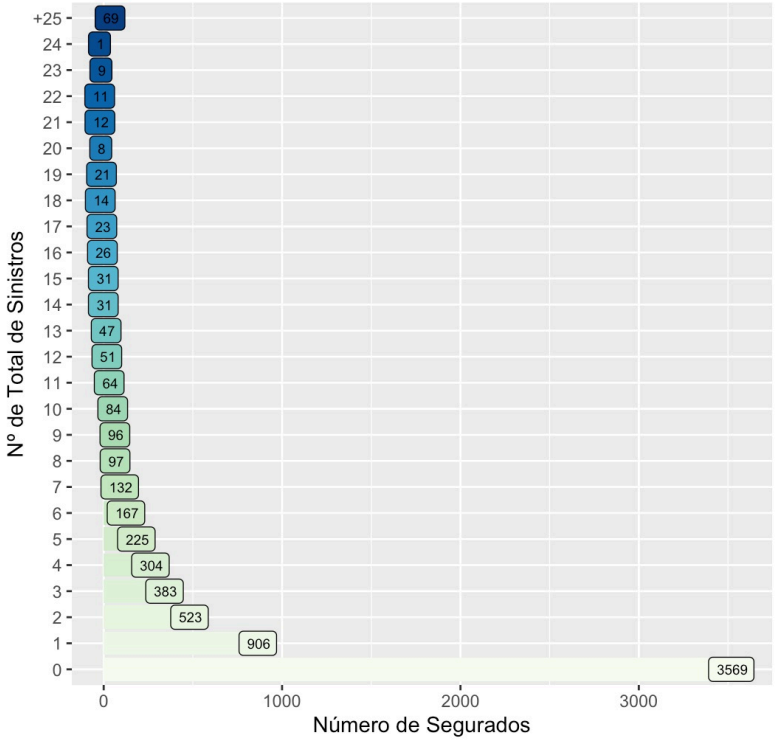


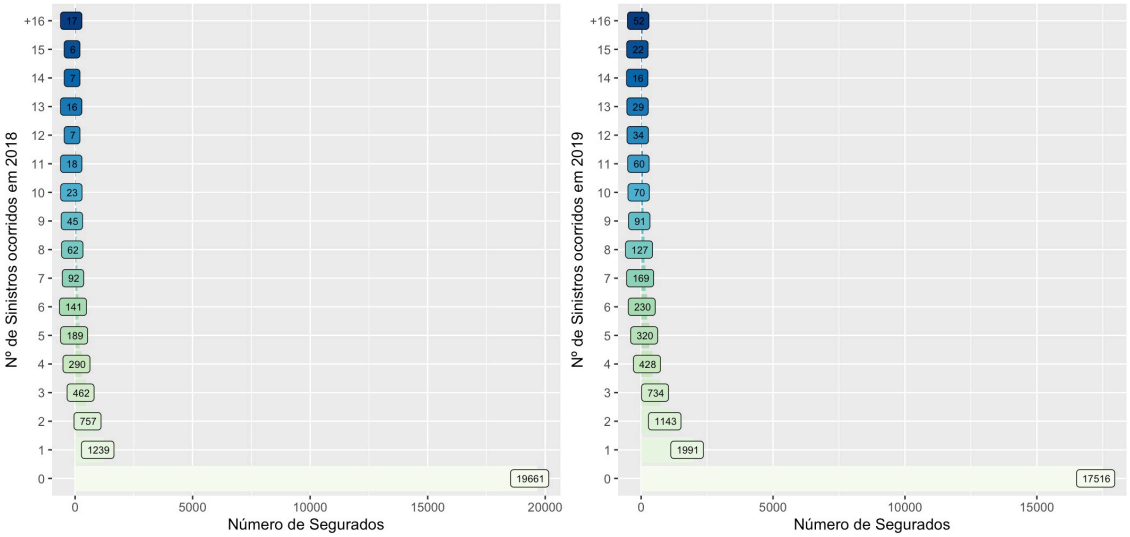
Figura 5.32: Número de pessoas seguras (SSL) por número de sinistros

Pela figura 5.32, observa-se que o número de pessoas com mais de 25 sinistros (69 pessoas) é inferior ao que acontece nos planos. No entanto e tal como como nos planos cerca de 88.27% das pessoas tem no máximo 3 sinistros.

5.2.7 Número de sinistros ocorridos por ano (Planos)

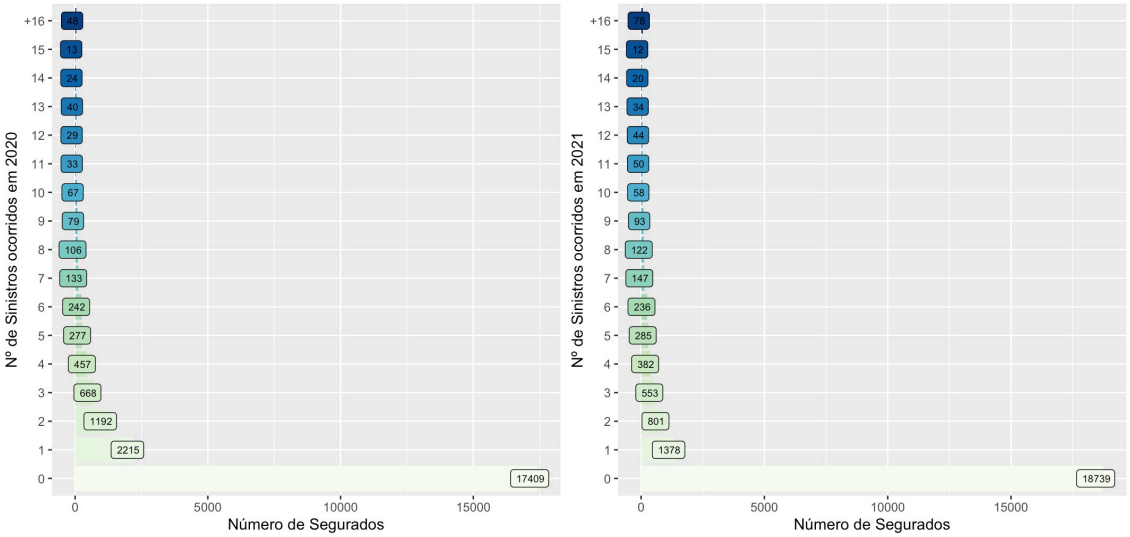
Tabela 5.11: Total e média de sinistros por ano (Planos)

Ano	2018	2019	2020	2021
Média de Sinistros	0.432	0.783	0.749	0.704
Total de Sinistros	9957	18072	17298	16275



(a) Ano 2018 (b) Ano 2019

Figura 5.33: Número de pessoas seguras por número de sinistros



(a) Ano 2020 (b) Ano 2021

Figura 5.34: Número de pessoas seguras por número de sinistros

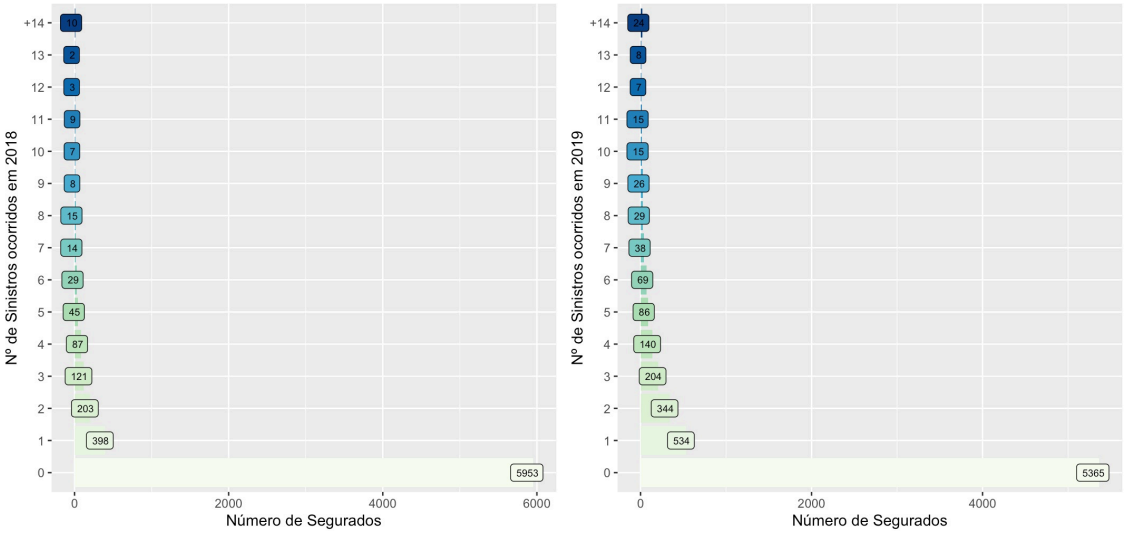
Como se observa pelas figuras 5.33 e 5.34 e tabela 5.11, pode concluir-se que em 2019 foi o ano em que ocorreram mais sinistros, sendo também em 2019 que a média de sinistros por pessoa é maior, tendo-se em média aproximadamente 1 sinistro.

Constata-se que o número médio de sinistros por pessoa é muito idêntico para os três últimos anos estudados pelo que se poderá afirmar que para esta carteira, a média ronda um sinistro por apólice. O número total de sinistros ocorridos flutuou, tendo subido significativamente em 2019 e descido a partir daí, pelo que será necessário estudar esta variável num horizonte temporal maior para se concluir com melhor precisão qual o intervalo de valores em que se situa. E tal como se esperaria e se deseja, em todos os anos, o número de apólices sem sinistros é bastante elevado, representado sempre mais de 75% da carteira.

5.2.8 Número de sinistros ocorridos por ano (SSL)

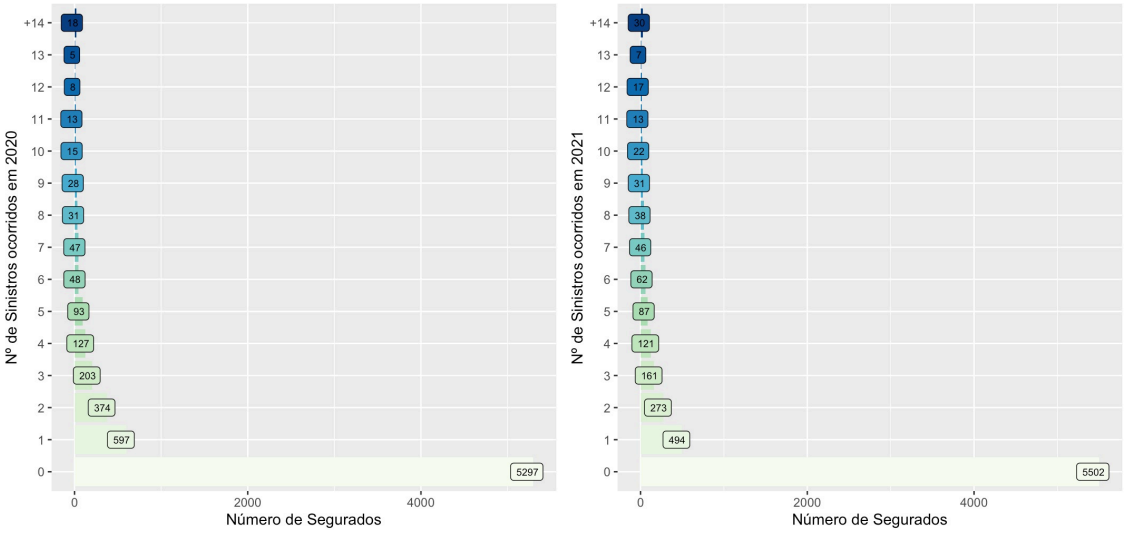
Tabela 5.12: Total e média de sinistros por ano (SSL)

Ano	2018	2019	2020	2021
Média de Sinistros	0.385	0.717	0.708	0.720
Total de Sinistros	2656	4952	4897	4980



(a) Ano 2018 (b) Ano 2019

Figura 5.35: Número de pessoas seguras por número de sinistros



(a) Ano 2020 (b) Ano 2021

Figura 5.36: Número de pessoas seguras por número de sinistros

Já nos seguros, concluiu-se com base nas figuras 5.35 e 5.36 e na tabela 5.12, que foi em 2021 que ocorreram mais sinistros, sendo também em 2021 que a média de sinistros por apólice é maior, tendo-se em média aproximadamente 1 sinistro.

Repete-se as conclusões retiradas relativamente ao número médio de sinistros por pessoa, isto é pode afirmar-se que para esta carteira, a média ronda um sinistro por pessoa. A diferença evidente reside no número total de sinistros ocorridos, que é bastante inferior ao que acontece nos planos. Este sofre um aumento significativo em 2019, baixa ligeiramente em 2020, e volta a subir em 2021. A partir de 2019, o número de sinistros parece rondar os 4900, no entanto para uma melhor precisão deste valor, é essencial estudar

o seu comportamento num horizonte temporal maior. Mais uma vez, tem-se em todos os anos, que o número de apólices sem sinistros é bastante elevado, representado sempre mais de 76% da carteira.

Capítulo

6 Implementação e Resultados

Nesta secção serão expostos os resultados obtidos aquando da implementação das metodologias já descritas. Os mesmos serão também analisados e discutidos com o maior detalhe.

6.1 Análise RFM

Relembra-se, de forma muito sucinta, que a análise RFM tem como objetivo classificar a carteira de clientes, dividindo-os em 5 categorias, com base em três variáveis, R, F e M. Por forma a implementar esta metodologia, construiu-se, com base nas bases de dados originais três variáveis que correspondem às variáveis R, F e M, enunciadas de seguida:

- $R \rightarrow$ Dias_Decorridos
- $F \rightarrow$ Num_total_sin
- $M \rightarrow$ Valor_total_gasto

Uma vez que se tratam de variáveis com escalas e dimensões muitos distintas, foi necessário a normalização dos dados usando a transformação Min-Max

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{6.1}$$

Apresenta-se na tabela 6.1, as cinco primeiras observações presentes na base BD_SEG_PLANOS:

Tabela 6.1: Primeiras cinco observações

ID_PS	...	R	F	M	R_norm	F_norm	M_norm
750208		171	2	0.00	0.1171	0.0125	0.0000
1077501		492	2	65.00	0.3370	0.0125	0.0061
917240		666	1	30.00	0.4562	0.0063	0.0028
769133		1460	0	0.00	1.0000	0.0000	0.0000
1123791		144	6	261.00	0.0986	0.0375	0.0245

Numa primeira e muito breve análise, destaca-se o cliente 769133, dado tratar-se de um cliente sem sinistros nos quatro anos considerados, o que se traduz num valor monetário nulo e no número máximo de dias decorridos desde o último sinistro. Contrastando, tem-se o cliente 1123791, que teve 6 sinistros num total de 261 euros, tendo realizado o último sinistro há apenas 144 dias, sensivelmente há menos de 6 meses.

Posteriormente, foi atribuído a cada cliente uma pontuação de 1 a 5 (variáveis Ponto_R, Pont_F, Pont_M) consoante o valor que apresentam para cada uma das variáveis (R_ norm, F_norm e M_norm).

Assim, e começando pelo número de dias decorridos desde o último sinistros (R), organizou-se por ordem decrescente a carteira de clientes. Tem-se assim, nas primeiras linhas os clientes com o maior número de dias decorridos e nas últimas linhas, os que tiveram sinistros há menos tempo.

É apresentado uma parte, da lista final já ordenada na tabela 6.2.

Tabela 6.2: Parte da lista final já ordenada

ID_PS	R_norm
769133	1
742883	1
[...]	[...]
943574	1
943575	1
[...]	[...]
1206365	0.2781
1175666	0.2774
[...]	[...]
1258741	0
1258749	0

Tal como se esperaria, em primeiro os que apresentam o valor 1 e em último os que apresentam o valor 0 na variável R_norm.

Após ordenar, e de acordo com a maneira usual de aplicar a análise RFM divide-se a lista ordenada em cinco partes iguais. Aos primeiros 20% atribui-se pontuação 5, correspondente a um cliente leal, seguida da pontuação 4 e assim sucessivamente.

Tabela 6.3: Pontuação atribuída à variável R

ID_PS	R_norm	Pont_R
769133	1	5
742883	1	5
[...]	[...]	[...]
943574	1	5
943575	1	4
[...]	[...]	[...]
1206365	0.2781	2
1175666	0.2774	2
[...]	[...]	[...]
1258741	0	1
1258749	0	1

Como se pode ver pela tabela 6.3, existem dois clientes com igual valor na variável R_norm mas com pontuações distintas. Assim, numa versão aprimorada do estudo realizado no presente trabalho, após pontuar a carteira de clientes, é necessário analisar estas discrepâncias, corrigir estes casos e possivelmente utilizar outra forma, que não a do "método quintil " para subdividir a lista ordenada. Uma abordagem aceitável, seria atribuir pontuação 5 a todos os clientes com valor máximo na variável em estudo, pontuação 1 aos clientes com valor mínimo e dividir os restantes em três partes iguais.

Repete-se o processo para as variáveis F_norm e M_norm. A única diferença comparando com a variável R, consiste no facto de para estas duas variáveis se ordenar por ordem crescente, uma vez que se pretende ter clientes com menos sinistros e menor valor gasto nos primeiros lugares. Assim e ilustrando na tabela 6.4, tem-se:

Tabela 6.4: Primeiras cinco observações já com as pontuações RFM

ID_PS	...	R	F	M	R_norm	F_norm	M_norm	Pont_R	Pont_F	Pont_M
750208		171	2	0.00	0.1171	0.0125	0.0000	1	2	5
1077501		492	2	65.00	0.3370	0.0125	0.0061	1	2	2
917240		666	1	30.00	0.4562	0.0063	0.0028	2	3	3
769133		1460	0	0.00	1.0000	0.0000	0.0000	5	5	5
1123791		144	6	261.00	0.0986	0.0375	0.0245	1	1	1

Agora que cada cliente já tem uma pontuação para cada uma das três variáveis, já é possível retirar conclusões sobre o seu comportamento e o seu valor para a companhia, no entanto para facilitar e melhorar a compreensão é calculado um valor único com base nas pontuações atribuídas. Para isso, utilizou-se o método comum, isto é considerar as três variáveis com a mesma importância para a companhia e também a análise WRFM, que considera que as variáveis têm "pesos" distintos consoante a sua importância para a companhia em questão.

Na análise RFM convencional, cada variável contribui de igual forma para o valor do cliente, pelo que este consiste apenas numa média. Obtém-se fazendo:

$$VC_j = \frac{Pont\_R_j + Pont\_F_j + Pont\_M_j}{3} \tag{6.2}$$

Tabela 6.5: Valor do Cliente obtido da análise RFM convencional

ID_PS	...	Pont_R	Pont_F	Pont_M	Média_RFM
750208		1	2	5	2.67
1077501		1	2	2	2
917240		2	3	3	2.67
769133		5	5	5	5
1123791		1	1	1	1

Na análise WRFM, a abordagem escolhida para atribuir os "pesos", foi o método AHP. A escolha deste método, deve-se ao facto de se tratar de um método de fácil compreensão, que incluiu as opiniões do decisor assim como as características da companhia.

Após aplicar o método AHP, cujo processo se encontra no anexo I, inferiu-se que os pesos atribuídos às variáveis R, F e M são 0.083, 0.724 e 0.193, respetivamente. Assim

sendo, o valor do cliente obtém-se fazendo a seguinte média ponderada:

$$VC_j = 0.083 * Pont\_R_j + 0.724 * Pont\_F_j + 0.193 * Pont\_M_j \tag{6.3}$$

Tabela 6.6: Valor do Cliente obtido da análise RFM convencional

ID_PS	...	Pont_R	Pont_F	Pont_M	Média_RFM	Média_WRFM
750208		1	2	5	2.67	2.5
1077501		1	2	2	2	2
917240		2	3	3	2.67	2.92
769133		5	5	5	5	5
1123791		1	1	1	1	1

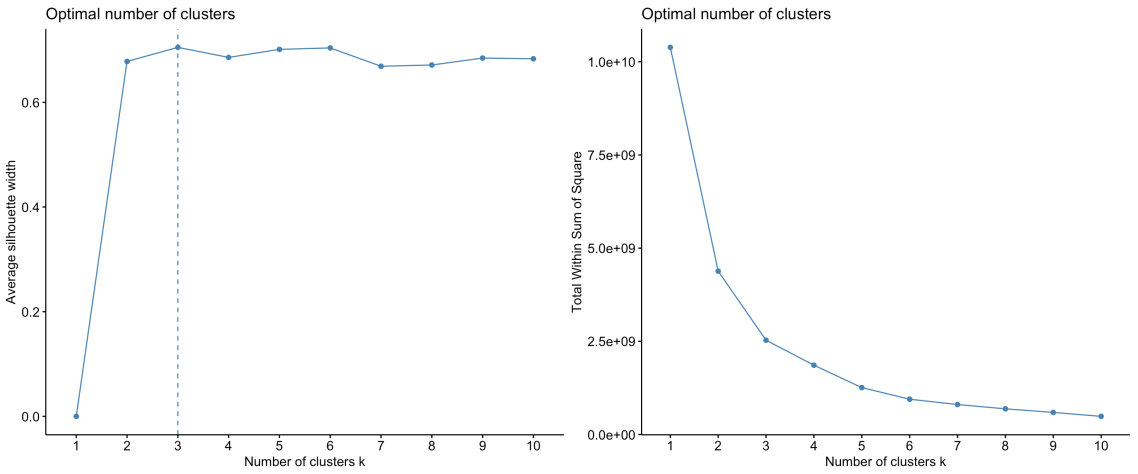
Comparando os valores obtidos e apresentados nas tabelas 6.5 e 6.6, não parece existir grande diferença entre os dois métodos utilizados. O cliente 917240 no primeiro método é classificado com 2.67 e no segundo com 2.92. Apesar de serem valores distintos são ambos próximos de 3, pelo que este cliente será sempre classificado como um potencial cliente.

Finalizando, tem-se para cliente uma pontuação que varia de 1 a 5. É assim possível concluir sobre o potencial de cada cliente e estar sempre a par do valor que este representa para a companhia.

6.2 K-Means

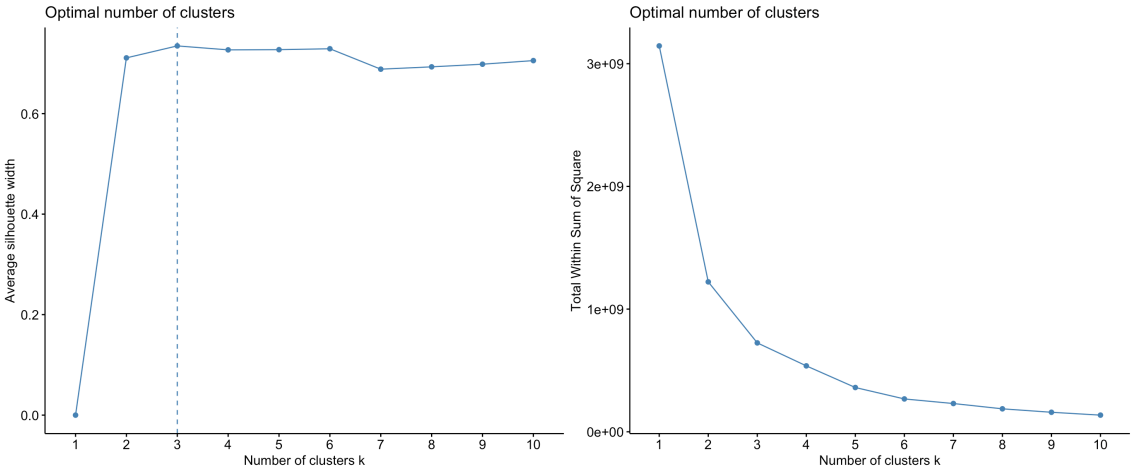
Após se ter classificado todos os clientes, passou-se para a segmentação dos mesmos utilizando o algoritmo *K-Means*.

Começou-se por utilizar dois índices para determinar o número ótimo de agrupamentos (K), como se mostra nas figura 6.1 e 6.2. O número ótimo de agregados será medido utilizando o índice *Silhouette* e o método *Elbow*.



(a) Índice de *Silhouette* (b) Método *Elbow*

Figura 6.1: Número de ótimo de *clusters* a considerar (Planos)



(a) Índice de *Silhouette* (b) Método *Elbow*

Figura 6.2: Número de ótimo de *clusters* a considerar (SSL)

O número de clusters testados começa com $k = 1, 2, 3, ..., 10$ *clusters*. Os resultados da avaliação das figuras anteriores mostram que o número ótimo de *clusters* (k) para o método *k-Means*, nos planos é $k = 3$ e nos SSL é $k = 2$.

Para além dos gráficos 6.1 e 6.2, foi calculado para cada valor de k, o valor médio do índice de *silhouette* que se expõe nas tabela seguintes (6.7 e 6.8).

Tabela 6.7: Valor médio do índice de *silhouette* para diferentes valores de k (Planos)

K	2	3	4	5	6	7	8	9	10
Valor Médio	0.68	0.71	0.69	0.70	0.70	0.67	0.67	0.67	0.68

Tabela 6.8: Valor médio do índice de *silhouette* para diferentes valores de k (SSL)

K	2	3	4	5	6	7	8	9	10
Valor Médio	0.71	0.73	0.73	0.73	0.73	0.69	0.69	0.70	0.71

Observa-se que o valor médio de $k = 3$ e $k = 3, 4, 5, 6$, nos planos e nos SSL, respectivamente é o mais próximo de um, no entanto como a dimensão do *dataset* é elevado decidiu-se considerar 6 *clusters* tanto nos planos como nos SSL. Ao analisar os gráficos em conjunto com os valores médios dos índices de *silhouette* parece razoável assumir $k = 6$.

Depois de definido o número ótimo de *clusters* a considerar, prosseguiu-se com a aplicação do algoritmo *K- Means*. Para o agrupamento foram consideradas as variáveis R, F e M, e posteriormente as variáveis R_norm, F_norm e M_norm, uma vez que se tem variâncias muito distintas devido a diferentes ordens de grandeza.

Os resultados do agrupamento utilizando o método *K-Means* ($k = 6$) na análise RFM podem ser observados nas figuras 6.3 e 6.4.

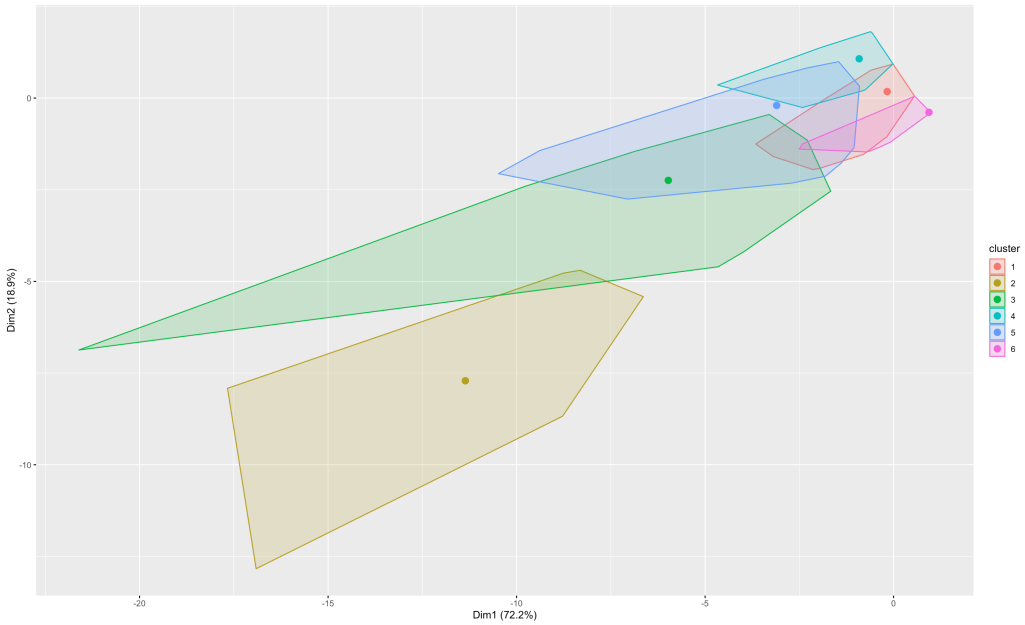


Figura 6.3: Visualização dos resultados de agrupamento utilizando o método *k-Means* ($k = 6$) nos Planos

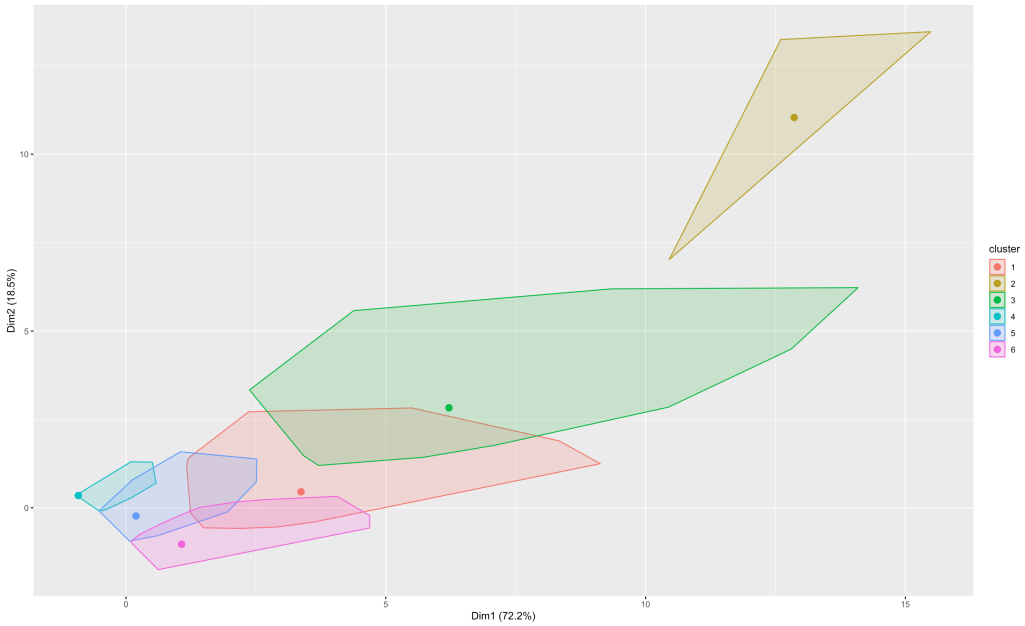


Figura 6.4: Visualização dos resultados de agrupamento utilizando o método *k-Means* ($k = 6$) nos SSL

Antes de analisar com detalhe os agrupamentos anteriormente obtidos, relembra-se a média das variáveis R, F e M nas bases BD_SEG_PLANOS e BD_SEG_SSL, na tabela 6.9, para depois comparar com os valores médios destas mesmas variáveis nos diferentes grupos de clientes encontrados (agrupamentos).

Tabela 6.9: Valor Médio das variáveis R, F e M na base dos Planos (BD_SEG_PLANOS)

	R	F	M
Valor Médio	1009	2.67	160.05

Nos planos, em média cada cliente tem 2.67 sinistros, tendo ocorrido sensivelmente há 1009 dias no valor de 160.05 euros.

Para além dos resultados do *K-Means*, os valores R, F, M foram categorizados em dois grupos com R, F, M ↓ para abaixo da média, e R, F, M ↑ para acima da média. Na fase seguinte, as pontuações RFM foram simbolizadas por ↓, se os valores de R,F e M forem inferiores à média, e por ↑, se os valores de R,F e M forem superiores à média.

Na tabela 6.10 é apresentada não só os valores médios obtidos com o *K-Means* como a dimensão de cada agrupamento e também a categorização anteriormente descrita.

Tabela 6.10: Centóides dos *clusters* obtidos nos Planos e respetiva categorização.

Grupo de Cliente	N	R	F	M	Classificação RFM
1	4231	890.51	3.20	151.93	R↓F↑M↓
2	25	206.44	17.20	6533.46	R↓F↑M↑
3	331	249.21	22.60	2236.21	R↓F↑M↑
4	4886	278.97	4.07	168.91	R↓F↑M↑
5	1309	248.38	13.58	924.57	R↓F↑M↑
6	12250	1444.54	0.19	8.55	RF↓M↓
Total	23032	-	-	-	-

Da tabela 6.10 apresentada é possível concluir que o grupo de clientes 6 é o mais numeroso e o grupo 2 é o mais pequeno. Relativamente ao grupo de clientes de 6 é ainda possível classificá-lo, numa fase inicial como um grupo de potenciais clientes, uma vez que o valor de R é superior à média pelo que os clientes pertencentes a este grupo tiveram sinistros há mais tempo, e os valores de F e M são inferiores à média, que não é algo positivo pois significa que tiveram mais sinistros com valores mais elevados, no entanto os valores não são consideravelmente elevados pelo que com as campanhas adequadas poderá melhorar.

Já o grupo de clientes 3 não é um grupo de muito interesse para a companhia, dado ter valor de R abaixo da média e valores de F e M acima da média.

Já nos SSL, com base na tabela 6.11, em média cada cliente tem 2.53 sinistros, tendo ocorrido sensivelmente há 1019 dias no valor de 152.92 euros.

Tabela 6.11: Valor Médio das variáveis R, F e M na base dos SSL (BD_SEG_SSL)

	R	F	M
Valor Médio	1018.77	2.53	152.92

Tabela 6.12: Centróides dos *clusters* obtidos nos SSL e respetiva categorização.

Grupo de Cliente	N	R	F	M	Classificação RFM
1	392	262.52	14.30	1006.14	R↓F↑M↑
2	4	305.75	9.50	8179.88	R↓F↑M↑
3	68	280.40	20.91	2478.96	R↓F↑M↑
4	3889	1444.61	0.15	6.62	R↑F↓M↓
5	1087	864.72	2.98	141.25	R↓F↑M↓
6	1464	240.66	4.50	191.81	RF↑M↑
Total	6904	-	-	-	-

As conclusões a retirar são semelhantes às apresentadas para os planos. De forma sucinta e com base na tabela 6.12, tem-se que o grupo de clientes 4 e 2 são respetivamente o mais e menos numeroso. À primeira vista o grupo de clientes 4 é o mais interessante do ponto de vista da companhia, podendo ser classificados como clientes leais e promissores. Por outro lado tem-se o terceiro grupo, que apresenta um número médio de sinistros bastante elevando assim como o valor total gasto.

Prosseguindo, segue-se uma análise mais detalhada dos gráficos 6.3 e 6.4.

Para os planos na figura 6.3, o grupo de clientes mais homogéneo é o 6 e o mais heterogéneo é o 4. Existe alguma sobreposição que não é o mais indicado e o que se pretendia, no entanto e apesar disso os grupos encontrados apresentam comportamentos semelhantes como se pode constatar pelas tabelas anteriores e pela análise do perfil de cada grupo de clientes.

Para o primeiro *cluster*, os valores de R ($\overline{R}$ = 890.51) e M ($\overline{M}$ = 151.93) estavam abaixo dos valores médios ao contrário de F ($\overline{F}$ = 3.20) que estava acima. Este cluster, portanto, é simbolizado por R↓F↑M↓.

Os clientes deste cluster pertencem na sua maioria a apólices que se encontram canceladas (84.36%), permanecendo ativos apenas uma anuidade e meia. Assim, este grupo é designado por "Clientes perdidos ou em risco". Os "clientes perdidos ou em risco" gastam menos do que a média, têm sinistros com mais frequência e que ocorreram nos últimos dois anos. Estes clientes são maioritariamente mulheres (41.15%) da zona de Lisboa (47.13%) na faixa etária dos 25-35 anos (22.90%), sendo a média de idades de aproximadamente 38 anos. Em média, o valor médio do cliente, calculado na primeira fase com base na análise RFM situa-se perto dos 2.

O segundo grupo é simbolizado por R↓F↑M↓ e denominado "Clientes de pouco

interesse".

Os clientes deste *cluster* parecem ser do tipo que têm muitos sinistros num curto espaço de tempo, uma vez que o valor R e F são respectivamente inferiores e superiores à média, e de valor bastante elevado. Os "Clientes de pouco interesse"são maioritariamente mulheres (44%) de Lisboa (76%), na faixa etária dos 45-55 anos (44%). Apesar de serem clientes de pouco interesse, com valor de cliente médio perto do 1, são clientes leais à companhia uma vez que 80% das apólices estão ativas e em média cada cliente permanece ativa quase quatro anuidades.

Se a empresa sugerir preços atractivos e competitivos para este grupo, é possível aumentar consideravelmente a permanência e assim suportar os custos associados a este grupo e investir em campanhas direcionadas para outros grupos de maior interesse.

O terceiro grupo é muito idêntico ao anterior, a única diferença significativa reside na faixa etária, sendo que a maioria se situa entre os 25 e os 45 anos. Poderão ser um grupo de maior interesse do que o anterior, uma vez que pertencem a faixas etárias mais jovens, o custo é mais baixo, o prémio comercial anual médio é mais elevado e o número de anuidades ativo é também quase quatro anuidades.

Apesar de os valores de F ($\overline{F}$ = 4.07) e M ($\overline{M}$ = 168.91) do quarto *cluster* estarem acima da média não é significativo pelo que com as campanhas certas poderão vir a tornar-se clientes importantes para a companhia. Este cluster é simbolizado como R↓F↑M↑ e denominado “Potencias Clientes".

Os clientes deste grupo mantém-se ativos durante três anuidades com prémios comerciais elevados e têm uma percentagem de apólices ativas de 52%. São clientes de idades jovens (maioritariamente 25-45 anos) (31.74%) que aproveitaram a campanha “Oferta Mês” do género feminino e da zona da Grande Lisboa.

Assim os “Potenciais clientes "são um segmento importante, cuja tendência de fidelização é elevada, pelo que devem ser sustentadas por algumas ofertas da empresa e campanhas especialmente criadas para eles.

De igual forma também o quinto *cluster*, apresenta valores de F ($\overline{F}$ = 13.58) e M ($\overline{M}$ = 924.27) acima da média, no entanto estes são bastante mais elevados o que se traduz num grupo de clientes de interesse reduzido. É bastante idêntico ao grupo três, nas variáveis R e F mas com custos muito mais reduzidos.

Assim do grupo dois, três este é aquele em que é vantajoso prestar atenção, uma vez que não só é mais numeroso como é melhor nas variáveis RFM.

Por último, os clientes do grupo 6 são os que melhor resultados têm nas variáveis RFM. O número de sinistros assim como o custo associado são bastante baixos pelo que é um de clientes com bastante interesse. Apesar disso apenas se mantém ativos uma anuidade e cerca de 82% das apólices estão canceladas.

É possível reverter esta situação a favor da companhia prestando-lhes atenção redobrada e analisando com regularidade o seu comportamento, por forma a criar campanhas especiais que permitam aumentar o número de anuidades ativos. São clientes na sua maioria jovens da zona de Lisboa e que apresentam um valor do cliente medio perto do 4.

Relativamente aos grupos de clientes encontrados para os SSL, o perfil de clientes define-se de forma semelhante.

6.3 Regressão Logística

Após aplicação da metodologia *K-means*, cujos resultados foram apresentados anteriormente, foi possível identificar 6 grupos distintos de clientes e consequentemente traçar o seu perfil.

Assim, tem-se no final uma classificação RFM, bem como um valor médio de cliente e um perfil para cada grupo. De forma a complementar os resultados obtidos, calcula-se para cada grupo, com base no valor médio de cliente, a probabilidade de churn, isto é de abandonar a companhia.

Para isso utilizou-se, como se refere na secção 3 a regressão logística simples, de notar que a abordagem utilizada utiliza apenas uma variável independente, e no futuro poder-se-ia melhorar ao considerar mais variáveis independentes, ou seja o modelo de regressão logística múltipla.

Adicionou-se uma variável à base nomeada Churn, que é a variável dependente binária a utilizar no modelo de regressão. Esta indica se o cliente abandonou ou não a companhia, isto é terá o valor 0 se o cliente se mantém ativo e o valor 1 caso contrário. A variável independente considerada é a Média_WRFM, estando assim criado um modelo de regressão logística.

Na figura 6.5, apresenta-se a dispersão dos "eventos" e "não-eventos" de Churn em relação à variável Média_WRFM.

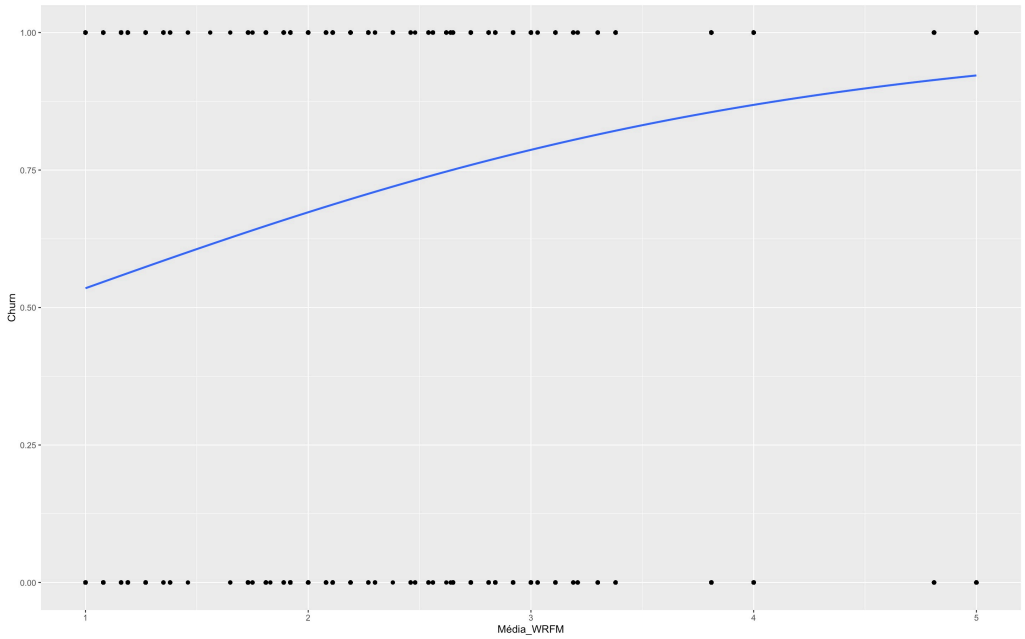


Figura 6.5: Dispersão de eventos e não-eventos

Posteriormente foi aplicado o modelo de regressão logística com a variável dependente Churn e a variável independente Média_WRFM.

Obteve-se a função de ligação estimada pelo modelo presente na equação 6.4.

$$ln\left(\frac{1}{1-p}\right) = -0.4421 + 0.5825 * Media\_WRFM \tag{6.4}$$

Como se observa o coeficiente da variável Média_WRFM é 0.5825. Uma vez que é positivo, conclui-se que quanto maior a média RFM maior será a probabilidade de churn. De igual forma nota-se que esta variável possui importância no modelo de regressão proposto, dado o valor-p ser 0.001.

Por conseguinte, utiliza-se o modelo proposto para prever, com base na média RFM de cada cliente, a probabilidade de Churn. Assim, ilustra-se na figura 6.6 a probabilidade de churn de todos os clientes presente na base de dados:

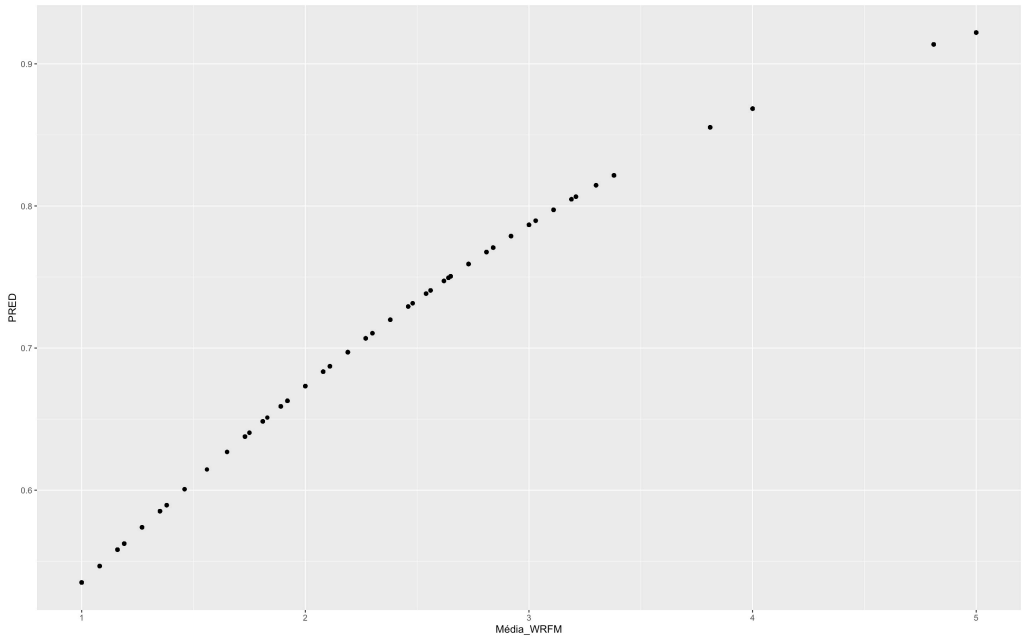


Figura 6.6: Distribuição das probabilidades previstas

Posteriormente, procede-se à exponenciação da variável de interesse, para uma interpretação mais enriquecida da relação da Média_WRFM com a ocorrência de Churn, uma que vez os estimadores do modelo são apresentados na forma logarítmica.

Após a transformação obtém-se o valor de 1.7904 para a variável Média_WRFM, que pode ser explicado da seguinte forma: por cada variação unitária na média RFM a probabilidade de ocorrência de Churn aumenta 1.7904. Por outras palavras, por cada variação unitária na média RFM a probabilidade de ocorrência de churn aumenta em 79.04%.

De seguida construiu-se o intervalo de confiança a 95% para o coeficiente da variável Média_WRFM estimado. Tem-se que o valor obtido do coeficiente toma o valor de 1.7904, podendo variar de 1.7459 a 1.8365 a 95% de confiança.

Desta forma é possível utilizando os coeficientes do modelo de regressão logística prever a probabilidade de churn associada a uma determinada média RFM (variável Média_WRFM). Por exemplo para uma média RFM de 3.000146 a probabilidade de abandonar a companhia é de 78.67%.

Com o intuito de avaliar a qualidade do ajuste do modelo de regressão logística criou-se a matriz de confusão. Uma vez que a regressão logística proposta modela uma variável binária, tal como se referencia na secção 3.3, probabilidades inferiores a 0.5 serão classificadas como “0”, caso contrário serão classificadas como “1”.

A matriz de confusão é apresentada na tabela 6.13.

Tabela 6.13: Matriz de Confusão.

Valor Estimado/Valor Observado	$Y = 0$	$Y = 1$
$\hat{Y} = 0$	2183	2409
$\hat{Y} = 1$	3393	15047

Na tabela 6.14 apresenta-se alguns indicadores obtidos a partir dos resultados expostos na tabela anterior. Estes indicadores permitem retirar conclusões mais precisas sobre a qualidade do ajustamento do modelo.

Tabela 6.14: Indicadores obtidos a partir da Matriz de Confusão.

Indicadores	Proporção
Precisão	74.81%
Sensibilidade	86.20%
Especificidade	39.15%
Verdadeiro Preditivo Positivo	81.60%
Verdadeiro Preditivo Negativo	47.54%

A matriz de confusão retorna uma excelente precisão do modelo em 74.81%, que indica a proporção de previsões corretas. O modelo tem ainda uma boa capacidade em estimar o evento como 1 ($\hat{Y} = 1$) dado que o seu valor obervado é 1 ($Y = 1$), este indicador é dado pela sensibilidade. Já na proporção de verdadeiros negativos o modelo não tem um desempenho tão bom, acertos em 39.15%. Uma razão plausível poderá ser a discrepência entre valores observados como 1 e 0, na maior parte dos casos tem-se $Y = 1$. Como consequência o modelo consegue acertos de 81.60% na previsão de valores positivos ou de "eventos" e apenas 47.54% na de valores negativos ou de "não-eventos".

Finalmente, tem-se para cada cliente qual o valor que representa para a companhia, dado pela média RFM, qual o seu perfil e consequentemente quais as campanhas mais eficazes e também a probabilidade de abandonar a companhia.

Criou-se uma função que recebe como input o número de cliente e apresenta os resultados anteriormente referidos.

Tem-se, na tabela 6.15, um exemplo prático do que se obtém ao introduzir o número de um cliente aleatório:

Tabela 6.15: Output dado pela função que resume os resultados obtidos.

Nº de Cliente	1236673
Valor Cliente	3.30
Grupo a que pertence	Grupo de clientes 4
Classificação RFM de Grupo	R↓F↑M↑ (1.73)
Perfil de Cliente	"Potencias Clientes": Mantém-se ativos durante três anuidades Prémios comerciais elevados 52% de apólices ativas de 52% Idades jovens (31.74% entre os 25-45 anos) Género feminino Zona da Grande Lisboa
Top 3 campanhas mais eficazes	Oferta Mês, CROSS SELLING e BD FRIA - 25%
Probabilidade de Churn	81.46%

Capítulo

7 Conclusões e Trabalho Futuro

O estudo presente nesta dissertação focou-se na união da análise RFM ao algoritmo K-Means com o objetivo de caracterizar o cliente e posteriormente concluir sobre o seu perfil.

Verificou-se que embora a análise RFM, seja um bom ponto de partida para avaliar o valor que o cliente representa para a empresa, por si só não é suficiente para obter um conhecimento profundo e detalhado sobre o cliente. Assim, ao conjugá-la com a metodologia *K-Means* torna-se possível segmentar os clientes em grupos mais homogéneos, e a posteriori, analisar por exemplo quais as campanhas mais eficazes.

Finalmente, foi possível não só cumprir com o objetivo inicial, utilizar a análise RFM para a análise da utilização/desistência das campanhas em vigor, como também, acrescentar valor ao conjugar esta análise à metodologia K-Means e ainda ao calcular a probabilidade de *Churn*.

No geral o propósito do presente estudo foi alcançado, isto é melhorar a relação com o cliente que se atinge ao conhecer bem qual o cliente que procura, neste caso os serviços prestados pela Future Healhtcare.

Apesar de se ter adquirido um conhecimento muito mais completo sobre o cliente é ainda possível melhorar os resultados obtidos. Poder-se-à no futuro considerar, por exemplo, mais variáveis na análise RFM, que façam sentido no contexto dos seguros de saúde.

Na regressão logística existe ainda muito espaço para o melhoramento, dado que neste estudo se utilizou apenas uma variável independente e com certeza, existem muitas outras variáveis que influenciam a rotatividade, e por isso importantes para o cálculo da probabilidade e previsão de churn.

Referências Bibliográficas

ASF - Apoio ao Consumidor. (s.d.). Obtido de https://www.asf.com.pt/NR/exeres/E527B387-8D5D-4701-8143-EF1EF088CC56.htm. (Ver p. 1)

Birant, D. (2011). Data Mining Using RFM Analysis. Em K. Funatsu (Ed.), *Knowledge-Oriented Applications in Data Mining* (pp. 91–108). doi:10.5772/13683. (Ver pp. 8, 10)

Blattberg, R. C., Kim, B.-D. & Neslin, S. A. (2008). *Database Marketing* (1ª ed.). doi:10.1007/978-0-387-72579-6. (Ver p. 6)

Buckinx, W., Moons, E., Van den Poel, D. & Wets, G. (2004). Customer-adapted coupon targeting using feature selection. *Expert Systems with Applications*, *26*(4), 509–518. doi:10.1016/j.eswa.2003.10.009. (Ver p. 15)

Buckinx, W. & Van den Poel, D. (2005). Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *European Journal of Operational Research*, *164*(1), 252–268. doi:10.1016/j.ejor.2003.12.010. (Ver p. 15)

Cheng, C. H. & Chen, Y. S. (2009). Classifying the segmentation of customer value via RFM model and RS theory. *Expert Systems with Applications*, *36*(3), 4176–4184. doi:10.1016/j.eswa.2008.04.003. (Ver pp. 5, 6, 8–11)

Christy, A. J., Umamakeswari, A., Priyatharsini, L. & Neyaa, A. (2021). RFM ranking – An effective approach to customer segmentation. *Journal of King Saud University - Computer and Information Sciences*, *33*(10), 1251–1257. doi:10.1016/j.jksuci.2018.09.004. (Ver pp. 8–10)

Chuang, H. M. & Shen, C. C. (2008). A study on the applications of data mining techniques to enhance customer lifetime value - based on the department store industry. *2008 International Conference on Machine Learning and Cybernetics*. doi:10.1109/icmlc.2008.4620398. (Ver p. 6)

Colombo, R. & Jiang, W. (1999). A stochastic RFM model. *Journal of Interactive Marketing*, *13*(3), 2–12. doi:10.1002/(SICI)1520-6653(199922)13:3<2::AID-DIR1>3.0.CO;2-H. (Ver pp. 9, 10)

Dennis, C., Marsland, D. & Cockett, T. (2001). Data mining for shopping centres – customer knowledge-management framework. *Journal of Knowledge Management*, *5*(4), 368–374. doi:10.1108/13673270110411797. (Ver p. 15)

Dursun, A. & Caber, M. (2016). Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis. *Tourism Management Perspectives*, *18*, 153–160. doi:10.1016/j.tmp.2016.03.001. (Ver pp. 6, 8–10)

Future-Healthcare. (s.d.). Obtido de https://www.future-healthcare.pt. (Ver p. 2)

Guimarães, A. (2022). Número de portugueses com seguro de saúde disparou em 10 anos. Porquê? Obtido de https://cnnportugal.iol.pt/seguro-de-saude/privados/numero-de-portugueses-com-seguro-de-saude-disparou-em-10-anos-porque/20220923/6329cc9b0cf2ea367d4f1ef6. (Ver p. 2)

Gustriansyah, R., Suhandi, N. & Antony, F. (2020). Clustering optimization in RFM analysis Based on k-Means. *Indonesian Journal of Electrical Engineering and Computer Science*, *18*(1), 470–477. doi:10.11591/ijeecs.v18.i1.pp470-477. (Ver p. 16)

Hosseini, S. M. S., Maleki, A. & Gholamian, M. R. (2010). Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. *Expert Systems with Applications*, *37*(7), 5259–5264. doi:10.1016/j.eswa.2009.12.070. (Ver p. 11)

Huigevoort, C. (2015). *Customer churn prediction for an insurance company* (Dissertação de Mestrado). Universidade Tecnológica de Eindhoven. (Ver pp. 5, 18).

Kaymak, U. (2001). Fuzzy target selection using RFM variables. Em *Proceedings Joint 9th IFSA World Congress and 20th NAFIPS International Conference (Cat. No. 01TH8569)* (Vol. 2, pp. 1038–1043). doi:10.1109/NAFIPS.2001.944748. (Ver p. 8)

Khajvand, M., Zolfaghar, K., Ashoori, S. & Alizadeh, S. (2011). Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study. *Procedia Computer Science*, *3*, 57–63. doi:10.1016/j.procs.2010.12.011. (Ver p. 11)

Lourenço, J. M. (2021). *The NOVAthesis LaTeX Template User's Manual*. NOVA University Lisbon. Obtido de https://github.com/joaomlourenco/novathesis/raw/master/template.pdf. (Ver p. iii)

Martinho, A. R. C. (2014). *Evolução do seguro de saúde em Portugal nos últimos 15 anos* (Dissertação de Mestrado). Escola Superior de Gestão de Tomar do Instituto Politécnico de Tomar. (Ver p. 1).

Meena, P. & Sahu, P. (2021). Customer Relationship Management Research from 2000 to 2020: An Academic Literature Review and Classification. *Vision: The Journal of Business Perspective*, *25*(2), 136–158. doi:10.1177/0972262920984550. (Ver p. 7)

Ng, K. & Liu, H. (2001). Customer Retention via Data Mining. *Artificial Intelligence Review*, *14*, 569–590. doi:10.1023/A:1006676015154. (Ver p. 6)

Ngai, E., Xiu, L. & Chau, D. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, *36*(2), 2592–2602. doi:10.1016/j.eswa.2008.02.021. (Ver pp. 5, 6, 12–15)

Payne, A. & Frow, P. (2005). A Strategic Framework for Customer Relationship Management. *Journal of Marketing*, *69*(4), 167–176. doi:10.1509/jmkg.2005.69.4.167. (Ver p. 5)

Ravi, V. (2007). *Advances in Banking Technology and Management: Impacts of ICT and CRM*. Information Science Publishing. (Ver p. 5).

Silva de Araújo, C., Pedron, C. & Picoto, W. (2018). What's Behind CRM Research? A Bibliometric Analysis of Publications in the CRM Research Field. *Journal of Relationship Marketing*, *17*, 29–51. doi:10.1080/15332667.2018.1440139. (Ver p. 7)

Soltani, Z. & Navimipour, N. J. (2016). Customer relationship management mechanisms: A systematic review of the state of the art literature and recommendations for future research. *Computers in Human Behavior*, *61*, 667–688. doi:10.1016/j.chb.2016.03.008. (Ver p. 7)

Spiteri, M. & Azzopardi, G. (2018). Customer Churn Prediction for a Motor Insurance Company. *2018 Thirteenth International Conference on Digital Information Management (ICDIM)*. doi:10.1109/icdim.2018.8847066. (Ver pp. 17, 18)

Walters, M. & Bekker, J. (2017). CUSTOMER SUPER-PROFILING DEMONSTRATOR TO ENABLE EFFICIENT TARGETING IN MARKETING CAMPAIGNS. *South African Journal of Industrial Engineering*, *28*(3), 113–127. doi:10.7166/28-3-1846. (Ver pp. 16, 17)

Wei, J.-T., Lin, S.-Y. & Wu, H.-H. (2010). A review of the application of RFM model. *African Journal of Business Management*, *4*(19), 4199–4206. doi:10.5897/ajbm.9000026. (Ver pp. 8–11)

Wu, J., Shi, L., Lin, W.-P., Tsai, S.-B., Li, Y., Yang, L. & Xu, G. (2020). An Empirical Study on Customer Segmentation by Purchase Behaviors Using a RFM Model and K-Means Algorithm. *Mathematical Problems in Engineering*, *2020*, 1–7. doi:10.1155/2020/8884227. (Ver p. 17)

Anexo

I Metodologia AHP

Tabela I.1: Matriz de Julgamentos

Indicadores	R	F	M
R	1	1/7	1/3
F	7	1	5
M	3	1/5	1

Tabela I.2: Vetor das Prioridades

V
0.083
0.724
0.193